SC18

Dallas, TX | hpc inspires.

final program

Program November 11-16, 2018
Exhibits November 12-15, 2018

KAY BAILEY HUTCHISON CONVENTION CENTER
DALLAS, TEXAS, USA

The International Conference for High Performance Computing, Networking, Storage, and Analysis

Sponsored by: IEEE computer society | TCHPC | acm sighpc

MICHAEL S. RAWLINGS
Mayor of Dallas

November 11, 2018

Supercomputing – SC18 Conference:

As Mayor of Dallas, it is my pleasure to welcome SC18 – the ultimate international conference for high performance computing, networking, storage, and analysis – to the Kay Bailey Hutchison Convention Center!

I applaud SC18 for bringing together high-performance computing (HPC) professionals and students from across the globe to share ideas, attend and participate in technical presentations, papers and workshops, all while inspiring the next generation of HPC professionals.

Dallas is the No. 1 visitor destination in Texas and one of the top cities in the nation for meetings and conventions. In addition to having the resources to host major events and conventions, Dallas is a greatly diverse American city and a melting pot of cultures. This important convergence of uniqueness and differences is reflected throughout the sights and sounds of our city. Dallas' authentic arts, music, food, historic landmarks and urban lifestyle all contribute to the city's rich cultural makeup.

Let the bold, vibrant spirit of Dallas, with a touch of Texas charm, guide the way to your BIG Dallas moments. Our city is filled with the unique experiences one can only have in Dallas – from cuisine by celebrity chefs, to the cultural landscape in one of the nation's largest arts districts, to first-class shopping - make the most of your time in our great city! From transportation to hotels, restaurants, retail and entertainment venues, Dallas is ready to roll out our red carpet to host SC18.

Again, welcome back to BIG D! Please accept my warmest wishes for an unforgettable event in our city. I know you will enjoy your time here and leave with Dallas in your heart.

Thank you,

Michael S. Rawlings
Mayor

OFFICE OF THE MAYOR CITY HALL 1500 MARILLA ST., 5EN DALLAS, TEXAS 75201

General Information

On-Site Registration and Conference Store

Navigating SC

Information Booths

ATM/Banks/Currency Exchange

Business Centers

Coat and Bag Check

Exhibits

Family Resources

Hotel Shuttle Schedule

Lost and Found

Quiet Room

Receptions

SC19 Information

Security

Visit Dallas Booth

Registration Pass Access

On-Site Registration and Conference Store

Level 2, Lobby D
Saturday, November 10, 1–6 pm
Sunday, November 11, 7 am–6 pm
Monday, November 12, 7 am–9 pm
Tuesday, November 13, 7:30 am–6 pm
Wednesday, November 14, 7:30 am–6 pm
Thursday, November 15, 7:30 am–5 pm
Friday, November 16, 8–11 am

Navigating SC

The SC Conference extends a warm, enthusiastic welcome to all first-time attendees and returning attendees. There will be two Navigating SC18 sessions specifically designed to help you organize your conference experience.

Level 2, Room D227
Monday, November 12, 4:45–5:30 pm
Tuesday, November 13, 7:30–8:15 am

Information Booths

Main Information Booth

Level 2, Between Lobby C and Lobby D
Saturday, November 10, 1–5 pm
Sunday, November 11, 8 am–6 pm
Monday, November 12, 8 am–7 pm
Tuesday, November 13, 8 am–6 pm
Wednesday, November 14, 8 am–6 pm
Thursday, November 15, 8 am–6 pm
Friday, November 16, 8:30 am–noon

Satellite Booth

Level 2, Lobby E
Sunday, November 11, 8 am–5 pm
Monday, November 12, 8 am–5 pm
Tuesday, November 13, 10 am–6 pm
Wednesday, November 14, 8 am–4 pm

ATM/Banks/Currency Exchange

U.S. Bank ATMs
Level 2, Exhibit Halls A/C/D
Kay Bailey Hutchison Convention Center
650 South Griffin St, C Hall, Dallas, TX 75202

Chase Bank
1700 Pacific Ave #125, Dallas, TX 7520

Comerica Bank
1717 Main St, Dallas, TX 75201

Wells Fargo Bank
1445 Ross Ave #100, Dallas, TX 75202

Business Centers

UPS
Omni Dallas Hotel
(connected to the Kay Bailey Hutchison Convention Center)
2nd Floor
555 S Lamar St, Dallas, TX 75202
(214) 652-4256

Monday–Friday: 7 am–6 pm
Saturday: 9 am–4 pm
Sunday: 9 am–4 pm

FedEx
Hyatt Regency Dallas Hotel
300 Reunion Blvd E, Dallas, TX 75207
(214) 741-2763

Monday–Friday: 7 am–7 pm
Saturday: 10 am–5 pm
Sunday: noon–5 pm

Coat and Bag Check
Level 1, Lower Lobby D

Saturday, November 10, 11 am–8 pm
Sunday, November 11, 7 am–6 pm
Monday, November 12, 7 am–10 pm
Tuesday, November 13, 7 am–9 pm
Wednesday, November 14, 7 am–8 pm
Thursday, November 15, 7 am–11 pm
Friday, November 16, 7 am–1 pm

Exhibits
Level 2, Exhibit Halls C/D/E/F

Tuesday, November 13, 10 am–6 pm
Wednesday, November 14, 10 am–6 pm
Thursday, November 15, 10 am–3 pm

Family Resources

Parents Room
Sometimes you just need a quiet place to be with your child. At the convention center, we will have a Parents Room with private feeding areas, refrigerators, changing tables, lockers to store your personal items, and plush chairs. We hope this room provides you a comfortable place to spend time with your child while you are away from home. The Parents Room is intended for parents with infants. This will ensure peace and quiet for parents caring for their babies.

Level 1, Room D160
Sunday, November 11, 8 am–6 pm
Monday, November 12, 8 am–9 pm
Tuesday, November 13, 8 am–7 pm
Wednesday, November 14, 8 am–7 pm
Thursday, November 15, 8 am–6 pm
Friday, November 16, 8 am–noon

On-Site Child Care

Provided by KiddieCorp, child care services are available in the convention center to registered attendees and exhibitors. For a small fee, children between the ages of 6 months and 12 years can participate.

We recommend registering children for childcare in advance as space may fill up.

On-Site Child Care Registration

There may be limited availability for drop-ins depending on the day, time, and age. Attendees are welcome to drop by the Child Care room and check the schedule.

Level 1, Room D162/164
Sunday, November 11, 8 am–6 pm
Monday, November 12, 8 am–9 pm
Tuesday, November 13, 8 am–7 pm
Wednesday, November 14, 8 am–7 pm
Thursday, November 15, 8 am–6 pm

Family Day

Level 2, Exhibits Hall
Wednesday, November 14, 4–6 pm

During this time, registered attendees can invite family members to join them for a walk through the Exhibits Hall. Attendees must check in their family members at Assisted Registration in the registration area located on Level 2, Lobby D.

Hotel Shuttle Schedule

SC provides shuttle bus service between hotels and the Kay Bailey Hutchison Convention Center.

Sunday, November 11, 7 am–6 pm
Monday, November 12, 7 am–9 pm
Tuesday–Thursday, November 13–15, 7 am–6 pm
Friday, November 16, 8 am–11 am

Route 1
Hyatt Regency

Route 2
Adolphus Hotel
Magnolia Hotel
Hilton Garden Inn

Route 3
Hampton Inn & Suites
Residence Inn
AC Marriott

Route 4
Sheraton Dallas Hotel
Marriott City Center

Route 5
Springhill Suites W. End
Westin Dallas
Crowne Plaza Hotel
Homewood Suites

Route 6
Indigo Hotel
La Quinta Hotel

Route 7
Fairmont Dallas Hotel

Route 8
Lorenze Hotel

Lost and Found

Located at the Exhibitor Approved Contractor (EAC) desk across from Registration on *Level 2, Lobby D.*

Quiet Room

The quiet room is a physical space where conversation and interaction are not allowed, where attendees can go for quiet, prayer, meditation, or similar activities.

Level 3, Room A301
Saturday, November 10, 9 am–6 pm
Sunday, November 11, 8 am–6 pm
Monday, November 12, 8 am–9 pm
Tuesday, November 13, 8 am–7 pm
Wednesday, November 14, 8 am–7 pm
Thursday, November 15, 8 am–7 pm

Receptions

At SC18, you'll have the opportunity to attend several receptions associated with key conference elements. Enjoy these times to meet and socialize with presenters and fellow attendees.

Exhibitor Reception

Gilley's
1135 S. Lamar St, Dallas, TX 75215
Sunday, November 11, 6–9 pm

Food, beverages, and entertainment for registered SC18 exhibitors. An Exhibitor badge, event ticket, and government-issued ID are required.

Grand Opening Gala Reception

Level 2, Exhibit Halls C/D/E/F
Monday, November 12, 7–9 pm

This is your first opportunity to see the latest high performance computing, networking, storage, analysis, and research products, services, and innovations. Open to all Technical Program, Exhibitor, and Students@SC registrants.

Posters Reception

Level 2, Ballroom C2/C3/C4
Tuesday, November 13, 5:15–7 pm

Celebrate the opening of the SC18 Posters. Meet this year's poster authors and enjoy complimentary refreshments and appetizers. Open to attendees with a Technical Program badge.

Technical Program Reception

Perot Museum of Nature and Science
2201 N. Field St., Dallas, TX 75201
Thursday, November 15, 7–10 pm

SC18 hosts this reception in thanks to our attendees with food, drink, and socializing. Here's to 30 more years of the SC Conference Series! A Technical Program badge, event ticket, and government-issued ID are required.

SC19 Information

Members of next year's SC committee will be available in the SC19 preview booth to offer information and discuss next year's SC conference in Denver, Colorado. Stop by for a copy of next year's Call for Participation and pick up some free gifts!

SC19 Preview Booth

Level 2, Lobby E
Tuesday, November 13, 10 am–5 pm
Wednesday, November 14, 10 am–5 pm
Thursday, November 15, 10 am–3 pm

SC19 Preview

Level 2, Exhibit Hall B
Thursday, November 15, 8:30–8:35 am

Security

The SC18 security office is located on *Level 2, Room A201.*

Visit Dallas Booth

Information specialists can assist with providing information on attractions, dining, transportation options, maps, brochures, and a myriad of other helpful hints for navigating Dallas.

Level 2, Lobby D
Sunday, November 11, 8 am–6 pm
Monday, November 12, 8 am–6 pm
Tuesday, November 13, 8 am–6 pm
Wednesday, November 14, 8 am–6 pm
Thursday, November 15, 8:30 am–5 pm
Friday, November 16, 8 am–noon

Registration Pass Access

Technical Program (TP)

- HPC Inspires Plenary (Mon)
- Keynote (Tue)
- Early Career (Mon)
- Grand Opening Gala Reception (Mon)
- Papers (Tue–Thu)
- Panels (Tue–Fri)
- Posters (Tue–Thu)
- Posters Reception (Tue)
- HPC Impact Showcase (Wed)
- Emerging Technologies Showcase (Wed)
- Doctoral Showcase (Wed)
- Scientific Visualization & Data Analytics Showcase (Wed)
- Student Cluster Competition (Mon–Wed)
- Invited Talks (Tue–Thu)
- Birds of a Feather (Tue–Thu)
- Awards (Tue–Thu)
- Awards Ceremony (Thu)
- Exhibits (Tue–Thu)
- Exhibitor Forum (Tue–Thu)
- Technical Program Reception (Thu, additional ticket required)
- One copy of the SC18 Proceedings

Tutorials (TUT)

- HPC Inspires Plenary (Mon)
- Keynote (Tue)
- Panels (Friday only)
- Tutorials (Sun and/or Mon)
- Tutorial Lunch (Sun and/or Mon)
- Electronic access to all tutorial notes (Tutorial notes are also available on a USB Flash Memory device for an additional $5)

Workshops (W)

- HPC Inspires Plenary (Mon)
- Keynote (Tue)
- Panels (Friday only)
- Workshops (Sun, Mon, Fri)

Exhibitor 24-hour Access (EX)

- Exhibitor Reception (Sun)
- Grand Opening Gala Reception (Mon)
- HPC Inspires Plenary (Mon)
- Keynote (Tue)
- Panels (Fri only)
- Posters (Tue–Thu)
- HPC Impact Showcase (Wed)
- Student Cluster Competition (Mon–Wed)
- Invited Talks (Tue–Thu)
- Birds of a Feather (Tue–Thu)
- Awards (Tue–Thu)
- Awards Ceremony (Thu)
- Exhibits (Tue–Thu)
- Exhibitor Forum (Tue–Thu)

Exhibits Hall Only (EXH)

- HPC Inspires Plenary (Mon)
- Keynote (Tue)
- Panels (Fri)
- HPC Impact Showcase (Wed)
- Student Cluster Competition (Mon–Wed)
- Birds of a Feather (Tue–Thu)
- Awards (Tue–Thu)
- Awards Ceremony (Thu)
- Exhibits (Tue–Thu)
- Exhibitor Forum (Tue–Thu)
- Additional: Grand Opening Gala Reception (Mon, tickets available on-site at the conference store: $100)

SC18 Schedule

ACM Gordon Bell Finalist

ACM Student Research Competition

Awards Presentation

Birds of a Feather

Doctoral Showcase

Early Career Program

Emerging Technologies

Exhibitor Forum

Experiencing HPC for Undergraduates

Exhibits

HPC Impact Showcase

HPC Inspires Plenary

Invited Talk

Keynote

Navigating SC18

Panel

Paper

Poster

Reception

SC19

Scientific Visualization & Data Analytics Showcase

Student Job/Opportunity Fair

Student Cluster Competition

Students@SC Event

Test of Time

Tutorial

Workshop

ACM Gordon Bell Finalist

(back to top)

Wednesday, November 14th

Room: A2 Ballroom
3:30 pm - 5:00 pm

Gordon Bell Prize Finalist Session 1

A Fast Scalable Implicit Solver for Nonlinear Time-Evolution Earthquake City Problem on Low-Ordered Unstructured Finite Elements with Artificial Intelligence and Transprecision Computing
Tsuyoshi Ichimura (University of Tokyo)

To address problems that occur due to earthquakes in urban areas, we propose a method that utilizes artificial intelligence (AI) and transprecision computing to accelerate a nonlinear dynamic low-order unstructured finite-element solver. The AI is used to improve the convergence of iterative solver leading to 5.56-fold reduction in arithmetic count from a standard solver, and FP16-FP21-FP32-FP64 computing is used to accelerate the sparse matrix-vector product kernel, which demonstrated 71.4% peak FP64 performance on Summit. This is 25.3 times faster than a standard solver and 3.99 times faster than the state-of-the-art SC14 Gordon Bell Finalist solver. Furthermore, the proposed solver demonstrated high scalability (88.8% on the K computer and 89.5% on Piz Daint), leading to 14.7% peak FP64 performance on 4096 nodes of Summit. The proposed approach utilizing AI and FP16 arithmetic has implications for accelerating other implicit solvers used for earthquake city simulations as well as various fields.

167-PFlops Deep Learning for Electron Microscopy: From Learning Physics to Atomic Manipulation
J. Travis Johnston (Oak Ridge National Laboratory)

An artificial intelligence system called MENNDL, which used 25,200 Nvidia Volta GPUs on Oak Ridge National Laboratory's Summit machine, automatically designed an optimal deep learning network in order to extract structural information from raw atomic-resolution microscopy data. In a few hours, MENNDL creates and evaluates millions of networks using a scalable, parallel, asynchronous genetic algorithm augmented with a support vector machine to automatically find a superior deep learning network topology and hyper-parameter set than a human expert can find in months. For the application of electron microscopy, the system furthers the goal of improving our understanding of the electron-beam-matter interactions and real-time image-based feedback, which enables a huge step beyond human capacity toward nanofabricating materials automatically. MENNDL has been scaled to the 4,200 available nodes of Summit achieving a measured 152.5 PFlops, with an estimated sustained performance of 167 PFlops when the entire machine is available.

Exascale Deep Learning for Climate Analytics
Thorsten Kurth (Lawrence Berkeley National Laboratory), Sean Treichler (Nvidia Corporation)

We extract pixel-level masks of extreme weather patterns using variants of Tiramisu and DeepLabv3+ neural networks. We describe improvements to the software frameworks, input pipeline, and the network training algorithms necessary to efficiently scale deep learning on the Piz Daint and Summit systems. The Tiramisu network scales to 5300 P100 GPUs with a sustained throughput of 21.0 PF/s and parallel efficiency of 79.0%. DeepLabv3+ scales up to 27360 V100 GPUs with a sustained throughput of 325.8 PF/s and a parallel efficiency of 90.7% in single precision. By taking advantage of the FP16 Tensor Cores, a half-precision version of the DeepLabv3+ network achieves a peak and sustained throughput of 1.13 EF/s and 999.0 PF/s respectively.

Thursday, November 15th

Room: A2 Ballroom
10:30 am - 12:00 pm

Gordon Bell Prize Finalist Session 2

Simulating the Weak Death of the Neutron in a Femtoscale Universe with Near-Exascale Computing
Ken McElvain (University of California, Berkeley; Lawrence Berkeley National Laboratory)

The fundamental particle theory called Quantum Chromodynamics (QCD) dictates everything about protons and neutrons, from their intrinsic properties to interactions that bind them into atomic nuclei. Quantities that cannot be fully resolved through experiment, such as the neutron lifetime (whose precise value is important for the existence of light-atomic elements that make the sun shine and life possible), may be understood through numerical solutions to QCD. We directly solve QCD using Lattice Gauge Theory and calculate nuclear observables such as neutron lifetime. We have developed an improved algorithm that exponentially decreases the time-to-solution and applied it on the new CORAL supercomputers, Sierra and Summit. We use run-time autotuning to distribute GPU resources, achieving 20% performance at low node count. We also developed optimal application mapping through a job manager, which allows CPU and GPU jobs to be interleaved, yielding 15% of peak performance when deployed across large fractions of CORAL.

ShenTu: Processing Multi-Trillion Edge Graphs on Millions of Cores in Seconds
Heng Lin (Tsinghua University, Fma Technology), Xiaowei Zhu (Tsinghua University, Qatar Computing Research Institute), Bowen Yu (Tsinghua University), Xiongchao Tang (Tsinghua University, Qatar Computing Research Institute), Wei Xue (Tsinghua University), Wenguang Chen (Tsinghua University), Lufei Zhang (State Key Laboratory of Mathematical Engineering and Advanced Computing), Torsten Hoefler (ETH Zurich), Xiaosong Ma (Qatar Computing Research Institute)

Graphs are an important abstraction used in many scientific fields. With the magnitude of graph-structured data constantly increasing, effective data analytics requires efficient and scalable graph

processing systems. Although HPC systems have long been used for scientific computing, people have only recently started to assess their potential for graph processing, a workload with inherent load imbalance, lack of locality, and access irregularity. We propose ShenTu, the first general-purpose graph processing framework that can efficiently utilize an entire petascale system to process multi-trillion edge graphs in seconds. ShenTu embodies four key innovations: hardware specializing, supernode routing, on-chip sorting, and degree-aware messaging, which together enable its unprecedented performance and scalability. It can traverse an unprecedented 70-trillion-edge graph in seconds. Furthermore, ShenTu enables the processing of a spam detection problem on a 12-trillion edge Internet graph, making it possible to identify trustworthy and spam web pages directly at the fine-grained page level.

Attacking the Opioid Epidemic: Determining the Epistatic and Pleiotropic Genetic Architectures for Chronic Pain and Opioid Addiction

Wayne Joubert (Oak Ridge National Laboratory), Daniel Jacobson (Oak Ridge National Laboratory)

We describe the CoMet application for large-scale epistatic Genome-Wide Association Studies (eGWAS) and pleiotropy studies. High performance is attained by transforming the underlying vector comparison methods into highly performant generalized distributed dense linear algebra operations. The 2-way and 3-way Proportional Similarity metric and Custom Correlation Coefficient are implemented using native or adapted GEMM kernels optimized for GPU architectures. By aggressive overlapping of communications, transfers and computations, high efficiency with respect to single GPU kernel performance is maintained up to the full Titan and Summit systems. Nearly 300 quadrillion element comparisons per second and over 2.3 mixed precision ExaOps are reached on Summit by use of Tensor Core hardware on the Nvidia Volta GPUs. Performance is four to five orders of magnitude beyond comparable state of the art. CoMet is currently being used in projects ranging from bioenergy to clinical genomics, including for the genetics of chronic pain and opioid addiction.

ACM Student Research Competition

(back to top)

Tuesday, November 13th

Room: C2/3/4 Ballroom
8:30 am - 5:00 pm

ACM Student Research Competition Posters

Session Description: SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom. The schedule of the ACM Student Research Competition session will be made available Wednesday evening form the results of the semi-finalists selection that will happen after the poster session.

SC18 Research Posters
SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Precomputing Outputs of Hidden Layers to Speed Up Deep Neural Network Training
Sohil Lal Shrestha (University of Texas, Arlington)

Deep learning has recently emerged as a powerful technique for many tasks including image classification. A key bottleneck of deep learning is that the training phase takes a lot of time, since state-of-the-art deep neural networks have millions of parameters and hundreds of hidden layers. The early layers of these deep neural networks have the fewest parameters but take up the most computation.

In this work, we reduce training time by progressively freezing hidden layers, pre-computing their output and excluding them from training in both forward and backward paths in subsequent iterations. We compare this technique to the most closely related approach for speeding up the training process of neural network.

Through experiments on two widely used datasets for image classification, we empirically demonstrate that our approach can yield savings of up to 25% wall-clock time during training with no loss in accuracy.

Designing Shared Address Space MPI Libraries in Many-Core Era
Jahanzeb Maqbool Hashmi (Ohio State University)

The emergence of modern multi-/many-cores has put more emphasis on optimizing intra-node communication. Existing designs in MPI libraries that work on the concept of distributed address spaces incur the overhead of intermediate memory copies to stage the data between processes. This

can lead to severe performance degradation especially on emerging many-core architectures like Intel Skylake and IBM OpenPOWER. This work proposes a high-performance "shared address-space"-based MPI point-to-point and collective communication designs using XPMEM. We first characterize the bottlenecks associated with XPMEM based communication and propose new designs for efficient MPI large message communication. Then we propose novel collective designs that are contention-free and offer true zero-copy reduction operations. The proposed designs are evaluated on different multi-/many-core architectures using various micro-benchmarks and application kernels such as MiniAMR and AlexNet DNN training on CNTK. The proposed designs have shown significant performance improvement over state-of-the-art available in MPI libraries.

Modeling Single-Source Shortest Path Algorithm Dynamics to Control Performance and Power Tradeoffs

Sara Karamati (Georgia Institute of Technology), Jeffrey Young (Georgia Institute of Technology), Rich Vuduc (Georgia Institute of Technology)

This work presents a new methodology to improve the performance of parallel algorithms by tuning the amount of available parallelism for execution throughout the runtime. As such, we expose key parameters controlling the performance and parallelism of the algorithm and build a software-based controller with the objective of maintaining the optimal performance. Our controller allows for tuning the level of parallelism executed in each time epoch to optimize for performance while preserving power usage. More specifically, our experimental evaluation focuses on a tunable variation of a GPU-based delta-stepping algorithm for computing the single-source shortest path (SSSP); As the available parallelism for the delta-stepping SSSP is highly irregular and strongly input-dependent, our extensive experiments show that average power can be reduced while average parallelism is increased. This increase in average parallelism provides substantial energy savings, independent of the hardware.

Accelerating DNA Long Read Mapping with Emerging Technologies

Roman Kaplan (Israel Institute of Technology)

DNA sequencing technologies output only short fragments of a genome, called reads. New single-molecule real-time sequencing technologies can produce long reads, up to tens of thousands base pairs, within minutes. However, these long reads may contain up to 15% errors.

To construct a genome from DNA reads, a computationally expensive bioinformatics task, read mapping, is required. Read mapping finds the best-fitting location for each DNA read on a long reference sequence. The length and error rate of long reads poses a challenge for existing read mapping hardware solutions, designed for short reads with low error rates. This work presents a novel DNA read mapping hardware architecture, RASSA. RASSA is a Resistive Approximate Similarity Search Accelerator that exploits charge distribution and parallel in-memory processing to reflect a mismatch count between DNA sequences. RASSA implementation of long read DNA mapping outperforms state-of-the-art long read mapping solution by 16-77x with comparable accuracy.

SimFS: A Simulation Data Virtualizing File System Interface

Salvatore Di Girolamo (ETH Zurich)

In the big (simulation) data era, simulations often produce petabytes of data to be stored in parallel filesystems or large-scale databases. This data is accessed, often by thousands of analysts and scientists, over the course of decades. However, storing these volumes of data for long time periods of time is not cost effective and, in some cases, practically impossible.

SimFS transparently virtualizes the simulation output, relaxing the storage requirements and re-simulating missing data on-demand. SimFS monitors the analysis access pattern in order to decide (1) which data to store and (2) apply prefetching techniques to improve the analysis performance. SimFS enables a trade-off between on-disk solutions, where all the simulation data is stored on disk, and in-situ, where no data is stored and analyses are always coupled with simulations. Overall, by exploiting the growing computing power and relaxing the storage capacity requirements, SimFS offers a viable path towards exa-scale simulations.

Holistic Root Cause Analysis of Node Failures in Production HPC

Anwesha Das (North Carolina State University)

Production HPC clusters endure failures incurring computation and resource wastage. Despite the presence of various failure detection and prediction schemes, a comprehensive understanding of how nodes fail considering various components and layers of the system is required for sustained resilience. This work performs a holistic root cause diagnosis of node failures using a measurement-driven approach on contemporary system logs that can help vendors and system administrators support exascale resilience.

Our work shows that lead times can be increased by at least 5 times if external subsystem correlations are considered as opposed to considering the events of a specific node in isolation. Moreover, when detecting sensor measurement outliers and interconnect related failures, triggering automated recovery events can exacerbate the situation if recovery is unsuccessful.

Geomancy: Automated Data Placement Optimization

Oceane Bel (University of California, Santa Cruz)

Exascale cloud storage and High-Performance Computing Systems (HPC) deliver unprecedented storage capacity and levels of computing power, though the full potential of these systems remain untapped because of inefficient data placement. Changes in data access patterns can cause a system's performance to suffer. To mitigate performance losses, system designers implement strategies to preemptively place popular data on higher performance nodes. However, these strategies fail to address a diverse userbase whose users individually demand the highest performance, and they must be carefully constructed by an expert of the system.

We propose Geomancy, a tool that reorganizes data to increase I/O throughput. In systems where heuristic-based improvements may become resource intensive, Geomancy determines new placement policies by training a deep neural network with past workload and system traces. With real workload traces, Geomancy calculated an example placement policy that demonstrated a 49% increase in average throughput compared to the default data layout.

Numerical Simulation of a Flue Instrument with Finite-Difference Lattice Boltzmann Method using GPGPU

Ryoya Tabata (Kyushu Institute of Technology)

In this work, we discuss the possibility of using GPGPU techniques for Aeroacoustic Simulation (especially for flue instruments) with the finite-difference lattice Boltzmann method (FDLBM). Compressible flow simulation has been used in direct aeroacoustic simulation; however, the computational cost is huge due to the requirement of high computational mesh resolution, with small time steps. The lattice Boltzmann method (LBM) has been used as an efficient method for fluid simulation using GPGPU. However, LBM is not accurate enough when applied to some aeroacoustic problems. On the other hand, FDLBM is able to use high-order finite-difference schemes and it has a high arithmetic intensity compared to LBM. We present a performance evaluation of the LBM and FDLBM with several finite-difference schemes on GPU with the roofline model.

Recursive Algebraic Coloring Engine

Christie Louis Alappat (University of Erlangen-Nuremberg)

Many iterative numerical methods for sparse systems and building blocks of sparse linear algebra are difficult to parallelize due to data dependencies. These may be loop-carried dependencies as they occur in solvers like Gauss-Seidel or write conflicts as in symmetric sparse matrix vector. Most of the existing parallelization strategies suffer from low performance on modern hardware, are matrix specific, or require tailored storage formats.

In this poster, we introduce a novel recursive level based algorithm called Recursive Algebraic Coloring (RAC), which achieves high hardware efficiency on modern multi-core architectures and works with simple data formats like compressed row storage. Method is implemented in a library called Recursive Algebraic Coloring Engine (RACE). Thorough performance analysis shows that RACE outperforms traditional multicoloring methods and Intel-MKL implementations with a factor of 2–2.5×. We are on par with Algebraic Block Multicoloring for small matrices, while for large matrices we gain a factor of 1.5–2×.

Accelerating Microscope Data Analysis Using Parallel Computing

John Ravi (North Carolina State University)

Single-Molecule Localization Microscopy (SMLM) techniques deal with the diffraction limit of fluorescent microscopy by localizing single molecules with high precision by stochastically switching molecules on and off. Thousands of camera frames containing subsets of blinking molecules are recorded to obtain a single super-resolution image. Each blinking molecule in each frame is subjected to localization protocols that fit the shape of the blink, assess the quality of the blink and then estimate their center. The algorithm implemented originally in MATLAB and compiled CUDA C, to compute a 'Super Resolution' image took around 6 minutes to process 256x256 pixel images of a moderately dense dataset. I ported the algorithm to C++ and parallelized it using OpenMP to compute multiple frames in parallel.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures

Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

PotC: Many-Body Potential Implementations à La Carte

Markus Höhnerbach (RWTH Aachen University)

Molecular dynamics is a valuable investigation tool for simulations in computational chemistry and materials science. In these simulations, atoms move according to so-called potentials, functions that typically describe the distance-dependent interactions between pairs of atoms. For some applications more complex "many-body potentials" are required. From a HPC perspective, such potentials pose challenges: small neighborhoods hinder vectorization; redundant force expressions are tedious and error prone to derive; the implementations are large, runtime-critical, and can not be broken into simpler "kernels". Consequently, only selected many-body potentials receive optimizations---with major effort.

PotC is a DSL for MD potentials and corresponding compiler to generate high-performance implementations. The compiler centralizes optimization knowledge for many-body potentials, and unburdens users from manual optimization and force derivation. The performance optimization of the generated code is still work-in-progress. However, independent of performance, the tool can also be applied when prototyping, testing, or possibly symbolically manipulating potentials.

OoO Instruction Benchmarking Framework on the Back of Dragons

Julian Hammer (University of Erlangen-Nuremberg, RRZE)

In order to construct an accurate instruction execution model for modern out-of-order micro architectures, an accurate description of instruction latency, throughput and concurrency is indispensable. Already existing resources and vendor provided information is neither complete nor detailed enough and sometimes incorrect. We therefore proclaim to deduct this information through runtime instruction benchmarking and present a framework to support such investigations based on LLVM's just-in-time and cross-platform compilation capabilities.

pyasmjit abstracts instructions, operands, and dependency chains, to easily construct the necessary benchmarks. The synthesized code is interactively compiled and executed using the llvmlite library, which in turn is based on the stable LLVM C-API. pyasmjit offers a command line as well as a programming interface. Unlike other approaches, we do not rely on model specific performance counters and focus on interoperability and automation to support quick modeling of future microarchitectures.

Studying the Impact of Power Capping on MapReduce-Based, Data-Intensive Mini-Applications on Intel KNL and KNM Architectures

Joshua H. Davis (University of Delaware)

In this poster, we quantitatively measure the impacts of data movement on performance in MapReduce-based applications when executed on HPC systems. We leverage the PAPI 'powercap' component to identify ideal conditions for execution of our applications in terms of (1) dataset characteristics (i.e., unique words); (2) HPC system (i.e., KNL and KNM); and (3) implementation of the MapReduce programming model (i.e., with or without combiner optimizations). Results confirm the high energy and runtime costs of data movement, and the benefits of the combiner optimization on these costs.

Accelerating Big Data Processing in the Cloud with Scalable Communication and I/O Schemes

Shashank Gugnani (Ohio State University)

With the advent of cloud computing, the field of Big Data has seen rapid growth. Most cloud providers provide hardware resources such as NVMe SSDs, large memory nodes, and SR-IOV. This opens up the possibility of large-scale high-performance data analytics and provides opportunities to use these resources to develop new designs. Cloud computing provides flexibility, security, and reliability, which are important requirements for Big Data frameworks. However, several important requirements are missing, such as performance, scalability, consistency, and quality of service (QoS). The focus of this work revolves around developing communication and I/O designs and concepts which can provide these requirements to Big Data frameworks. Specifically, we explore new ways to provide QoS and consistency in cloud storage systems, and provide scalable and high-performance communication frameworks.

Accelerating 2D FFT: Exploit GPU Tensor Cores through Mixed-Precision

Xiaohe Cheng (Hong Kong University of Science and Technology), Anumeena Sorna (National Institute of Technology, Tiruchirappalli)

The two-dimensional Fourier Transform is a widely-used computational kernel in many HPC applications. The popular NVIDIA cuFFT library provides a simple interface to compute 2D FFT on GPUs, but it's yet to utilize the recent hardware advancement in half-precision floating-point arithmetic. In this poster, we propose a mixed-precision method to accelerate 2D FFT by exploiting the FP16 matrix-multiply-and-accumulate units on the newest GPU architecture, known as tensor cores. We achieve a balance between speed and accuracy by dynamically splitting the single-precision input data into two half-precision operands and performing FFT separately. We present a CUDA-based implementation that achieves 3-digit more accuracy than half-precision cuFFT. We also demonstrate the stability and scalability of our approach and conclude that it attains high accuracy with tolerable splitting overhead.

Monitoring Parsl Workflows

Connor Pigg (University of Illinois)

As a Python library that enables workflows, Parsl gives users the ability to define complex workflows

in Python and run them in parallel on any computer system. This poster describe the process of adding monitoring to Parsl. Simple and comprehensive monitoring of a workflow's state and resource usage lets users audit, debug, and confirm workflow execution. The poster discusses how Parsl monitors workflow components, what data it captures (task status and resource usage), and the tools it used to do so (Elasticsearch) and to display the information (Kibana). A Kibana dashboard visualizes the collected logs in a real time, with an interactive user interface. This enhanced Parsl allows users the option to monitor the status and resource usage of their workflows via an Elasticsearch database and Kibana dashboard.

Identifying Network Data Transfer Bottlenecks in HPC Systems

Karen Tu (Lawrence Berkeley National Laboratory; University of California, Berkeley)

Improving network data transfer performance is a major factor for improving high performance computing systems. Most studies analyze data transfer and file system IO performance separately, but understanding the relationship between the two is essential for optimizing scheduling and resource management. Intuitively, if data is being transferred to a busy file system the transfer rate would be slower than a file system at regular activity levels.

This study analyzes patterns between file system activity and network throughput for several use cases of file writing and data transfers using a parallel file system. The parameters changed among the use cases were file striping for the file system, and buffer size and parallelism for data transfer. The main bottleneck for network data transfer rate was the number of OSTs the data was striped across. For a large number of OSTs (16 or greater), writing to the file system was the bottleneck.

Dendro-GR: Massively Parallel Simulations of Binary Black Hole Intermediate-Mass-Ratio Inspirals

Milinda Fernando (University of Utah)

We present a portable and highly-scalable algorithm and framework that targets problems in the astrophysics and numerical relativity communities. This framework combines together a parallel octree-refined adaptive mesh with wavelet adaptive multiresolution and a physics module to solve the Einstein equations of general relativity in the BSSN~formulation. The goal of this work is to perform advanced, massively parallel numerical simulations of Intermediate Mass Ratio Inspirals (IMRIs) of binary black holes with mass ratios on the order of 100:1. These studies will be used to generate waveforms for use in LIGO data analysis and to calibrate semi-analytical approximate methods. This advanced framework is designed to easily accommodate many existing algorithms in astrophysics for plasma dynamics and radiation hydrodynamics. We have designed novel algorithms to enable efficient simulations for such experiments and demonstrate excellent weak scalability up to 131K cores on ORNL's Titan for binary mergers for mass ratios up to 100.

Hardware Transactional Persistent Memory

Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM

by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories

Luann C. Jung (Massachusetts Institute of Technology, University of Chicago), Brendan T. Whitaker (Ohio State University, University of Chicago)

As scientific data repositories and filesystems grow in size and complexity, they become increasingly disorganized. The coupling of massive quantities of data with poor organization makes it challenging for scientists to locate and utilize relevant data, thus slowing the process of analyzing data of interest. To address these issues, we explore an automated clustering approach for quantifying the organization of data repositories. Our parallel pipeline processes heterogeneous filetypes (e.g., text and tabular data), automatically clusters files based on content and metadata similarities, and computes a novel "cleanliness" score from the resulting clustering. We demonstrate the generation and accuracy of our cleanliness measure using both synthetic and real datasets, and conclude that it is more consistent than other potential cleanliness measures.

Supercomputing for the Multi-Driver Routing

Zeyang Ye (Stony Brook University)

Supercomputing is essential for routing traffic by providing drivers the optimal routes with minimal traveling distances or time. The unique challenges that require supercomputers to overcome are of multiple folds: numerous drivers, massive simultaneous requests, multiple locations, and needs of instant gratifications, etc. We developed two parallel methods, PSAD and PSAD-M, by using domain decomposition and state-mixing techniques. On the same computing platform with 96 cores, for the same problem, our PSAD methods outperform all published benchmarks by over a hundred times, while improving the solution quality. For the same routing problem on 384 cores, our PSAD-M reduced the elapsed time from the unbearable ten minutes to the reasonable 5 seconds, achieving a record-breaking speedup of 170. By providing instant routing solutions that enable online recommendations, our methods break the bottleneck of the widely adopted offline approaches.

NautDB: Toward a Hybrid Runtime for Processing Compiled Queries

Samuel Grayson (University of Texas, Dallas)

General purpose operating and database system suffer under the load of their generality which makes achieving optimal performance extremely hard, especially on modern hardware. The goal of this research is to integrate, for the first time, specialization techniques from the OS community (hybrid runtimes) and DB community (compiled queries) for high-performance query processing on modern

hardware. We envision a system called NautDB, a hybrid dataflow runtime for executing compiled queries. As a first step toward our goal, we evaluate the performance of compiled queries on Linux and run as a Nautilus hybrid runtime using a simple prototype. Our results demonstrate that combining these specialization techniques has transformative potential for building the next generation (distributed) high-performance query processing systems and big data platforms.

Mitigating Performance and Progress Variability in Iterative Asynchronous Algorithms
Justs Zarins (University of Edinburgh)

Large HPC machines are susceptible to irregular performance. Factors like chip manufacturing differences, heat management, and network congestion combine to result in varying execution time for the same code and input sets. Asynchronous algorithms offer a partial solution. In these algorithms, fast workers are not forced to synchronize with slow ones. Instead they continue computing updates, and moving toward the solution, using the latest data available to them, which may have become stale (i.e. a number of iterations out of date compared to the most recent data). While this allows for high computational efficiency, the convergence rate of asynchronous algorithms tends to be lower.

To address this problem, we are using the unique properties of asynchronous algorithms to develop load balancing strategies for iterative asynchronous algorithms in both shared and distributed memory. Our poster shows how our solution attenuates noise, resulting in significant reduction progress imbalance and time-to-solution variability.

Eulerian Algorithms for the Discretization of Plasma Kinetic Equations
James L. Juno (University of Maryland)

While fluid models are common tools in the study of plasmas, many of these systems, whether in astrophysics or the lab, are only weakly collisional and far from equilibrium, making them more accurately described by kinetic equations. Kinetic equations can be computationally demanding due to the need to solve for the distribution function of the particles in a higher dimensional phase space, with position and velocity coordinates. Despite this challenge, the motivation for solving the plasma kinetic equation is large as there remains a vast array of questions concerning collisionless dynamics in real plasma systems. Here we present algorithms in an Eulerian framework for the discretization of the plasma kinetic equation, using a high-order discontinuous Galerkin finite element method due to its arithmetic intensity and parallelizability. Scaling and performance of the algorithm are discussed, and benchmarks of the algorithm are presented as well.

Room: C2/3/4 Ballroom
5:15 pm - 7:00 pm

Poster Reception

Session Description: The Posters Reception is an opportunity for attendees to interact with poster presenters and includes research and ACM Student Research Competition posters, Doctoral Showcase posters as well as the Scientific Visualization & Data Analytics Showcase.

Wednesday, November 14th

Room: C2/3/4 Ballroom
8:30 am - 5:00 pm

ACM Student Research Competition Posters

Session Description: SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom. The schedule of the ACM Student Research Competition session will be made available Wednesday evening form the results of the semi-finalists selection that will happen after the poster session.

SC18 Research Posters
SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Precomputing Outputs of Hidden Layers to Speed Up Deep Neural Network Training
Sohil Lal Shrestha (University of Texas, Arlington)

Deep learning has recently emerged as a powerful technique for many tasks including image classification. A key bottleneck of deep learning is that the training phase takes a lot of time, since state-of-the-art deep neural networks have millions of parameters and hundreds of hidden layers. The early layers of these deep neural networks have the fewest parameters but take up the most computation.

In this work, we reduce training time by progressively freezing hidden layers, pre-computing their output and excluding them from training in both forward and backward paths in subsequent iterations. We compare this technique to the most closely related approach for speeding up the training process of neural network.

Through experiments on two widely used datasets for image classification, we empirically demonstrate that our approach can yield savings of up to 25% wall-clock time during training with no loss in accuracy.

Designing Shared Address Space MPI Libraries in Many-Core Era
Jahanzeb Maqbool Hashmi (Ohio State University)

The emergence of modern multi-/many-cores has put more emphasis on optimizing intra-node communication. Existing designs in MPI libraries that work on the concept of distributed address spaces incur the overhead of intermediate memory copies to stage the data between processes. This can lead to severe performance degradation especially on emerging many-core architectures like Intel Skylake and IBM OpenPOWER. This work proposes a high-performance "shared address-space"-

based MPI point-to-point and collective communication designs using XPMEM. We first characterize the bottlenecks associated with XPMEM based communication and propose new designs for efficient MPI large message communication. Then we propose novel collective designs that are contention-free and offer true zero-copy reduction operations. The proposed designs are evaluated on different multi-/many-core architectures using various micro-benchmarks and application kernels such as MiniAMR and AlexNet DNN training on CNTK. The proposed designs have shown significant performance improvement over state-of-the-art available in MPI libraries.

Modeling Single-Source Shortest Path Algorithm Dynamics to Control Performance and Power Tradeoffs

Sara Karamati (Georgia Institute of Technology), Jeffrey Young (Georgia Institute of Technology), Rich Vuduc (Georgia Institute of Technology)

This work presents a new methodology to improve the performance of parallel algorithms by tuning the amount of available parallelism for execution throughout the runtime. As such, we expose key parameters controlling the performance and parallelism of the algorithm and build a software-based controller with the objective of maintaining the optimal performance. Our controller allows for tuning the level of parallelism executed in each time epoch to optimize for performance while preserving power usage. More specifically, our experimental evaluation focuses on a tunable variation of a GPU-based delta-stepping algorithm for computing the single-source shortest path (SSSP); As the available parallelism for the delta-stepping SSSP is highly irregular and strongly input-dependent, our extensive experiments show that average power can be reduced while average parallelism is increased. This increase in average parallelism provides substantial energy savings, independent of the hardware.

Accelerating DNA Long Read Mapping with Emerging Technologies

Roman Kaplan (Israel Institute of Technology)

DNA sequencing technologies output only short fragments of a genome, called reads. New single-molecule real-time sequencing technologies can produce long reads, up to tens of thousands base pairs, within minutes. However, these long reads may contain up to 15% errors.

To construct a genome from DNA reads, a computationally expensive bioinformatics task, read mapping, is required. Read mapping finds the best-fitting location for each DNA read on a long reference sequence. The length and error rate of long reads poses a challenge for existing read mapping hardware solutions, designed for short reads with low error rates. This work presents a novel DNA read mapping hardware architecture, RASSA. RASSA is a Resistive Approximate Similarity Search Accelerator that exploits charge distribution and parallel in-memory processing to reflect a mismatch count between DNA sequences. RASSA implementation of long read DNA mapping outperforms state-of-the-art long read mapping solution by 16-77x with comparable accuracy.

SimFS: A Simulation Data Virtualizing File System Interface

Salvatore Di Girolamo (ETH Zurich)

In the big (simulation) data era, simulations often produce petabytes of data to be stored in parallel filesystems or large-scale databases. This data is accessed, often by thousands of analysts and

scientists, over the course of decades. However, storing these volumes of data for long time periods of time is not cost effective and, in some cases, practically impossible.

SimFS transparently virtualizes the simulation output, relaxing the storage requirements and re-simulating missing data on-demand. SimFS monitors the analysis access pattern in order to decide (1) which data to store and (2) apply prefetching techniques to improve the analysis performance. SimFS enables a trade-off between on-disk solutions, where all the simulation data is stored on disk, and in-situ, where no data is stored and analyses are always coupled with simulations. Overall, by exploiting the growing computing power and relaxing the storage capacity requirements, SimFS offers a viable path towards exa-scale simulations.

Holistic Root Cause Analysis of Node Failures in Production HPC

Anwesha Das (North Carolina State University)

Production HPC clusters endure failures incurring computation and resource wastage. Despite the presence of various failure detection and prediction schemes, a comprehensive understanding of how nodes fail considering various components and layers of the system is required for sustained resilience. This work performs a holistic root cause diagnosis of node failures using a measurement-driven approach on contemporary system logs that can help vendors and system administrators support exascale resilience.

Our work shows that lead times can be increased by at least 5 times if external subsystem correlations are considered as opposed to considering the events of a specific node in isolation. Moreover, when detecting sensor measurement outliers and interconnect related failures, triggering automated recovery events can exacerbate the situation if recovery is unsuccessful.

Geomancy: Automated Data Placement Optimization

Oceane Bel (University of California, Santa Cruz)

Exascale cloud storage and High-Performance Computing Systems (HPC) deliver unprecedented storage capacity and levels of computing power, though the full potential of these systems remain untapped because of inefficient data placement. Changes in data access patterns can cause a system's performance to suffer. To mitigate performance losses, system designers implement strategies to preemptively place popular data on higher performance nodes. However, these strategies fail to address a diverse userbase whose users individually demand the highest performance, and they must be carefully constructed by an expert of the system.

We propose Geomancy, a tool that reorganizes data to increase I/O throughput. In systems where heuristic-based improvements may become resource intensive, Geomancy determines new placement policies by training a deep neural network with past workload and system traces. With real workload traces, Geomancy calculated an example placement policy that demonstrated a 49% increase in average throughput compared to the default data layout.

Numerical Simulation of a Flue Instrument with Finite-Difference Lattice Boltzmann Method using GPGPU

Ryoya Tabata (Kyushu Institute of Technology)

In this work, we discuss the possibility of using GPGPU techniques for Aeroacoustic Simulation (especially for flue instruments) with the finite-difference lattice Boltzmann method (FDLBM). Compressible flow simulation has been used in direct aeroacoustic simulation; however, the computational cost is huge due to the requirement of high computational mesh resolution, with small time steps. The lattice Boltzmann method (LBM) has been used as an efficient method for fluid simulation using GPGPU. However, LBM is not accurate enough when applied to some aeroacoustic problems. On the other hand, FDLBM is able to use high-order finite-difference schemes and it has a high arithmetic intensity compared to LBM. We present a performance evaluation of the LBM and FDLBM with several finite-difference schemes on GPU with the roofline model.

Recursive Algebraic Coloring Engine

Christie Louis Alappat (University of Erlangen-Nuremberg)

Many iterative numerical methods for sparse systems and building blocks of sparse linear algebra are difficult to parallelize due to data dependencies. These may be loop-carried dependencies as they occur in solvers like Gauss-Seidel or write conflicts as in symmetric sparse matrix vector. Most of the existing parallelization strategies suffer from low performance on modern hardware, are matrix specific, or require tailored storage formats.

In this poster, we introduce a novel recursive level based algorithm called Recursive Algebraic Coloring (RAC), which achieves high hardware efficiency on modern multi-core architectures and works with simple data formats like compressed row storage. Method is implemented in a library called Recursive Algebraic Coloring Engine (RACE). Thorough performance analysis shows that RACE outperforms traditional multicoloring methods and Intel-MKL implementations with a factor of 2–2.5×. We are on par with Algebraic Block Multicoloring for small matrices, while for large matrices we gain a factor of 1.5–2×.

Accelerating Microscope Data Analysis Using Parallel Computing

John Ravi (North Carolina State University)

Single-Molecule Localization Microscopy (SMLM) techniques deal with the diffraction limit of fluorescent microscopy by localizing single molecules with high precision by stochastically switching molecules on and off. Thousands of camera frames containing subsets of blinking molecules are recorded to obtain a single super-resolution image. Each blinking molecule in each frame is subjected to localization protocols that fit the shape of the blink, assess the quality of the blink and then estimate their center. The algorithm implemented originally in MATLAB and compiled CUDA C, to compute a 'Super Resolution' image took around 6 minutes to process 256x256 pixel images of a moderately dense dataset. I ported the algorithm to C++ and parallelized it using OpenMP to compute multiple frames in parallel.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures

Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new

programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

PotC: Many-Body Potential Implementations à La Carte

Markus Höhnerbach (RWTH Aachen University)

Molecular dynamics is a valuable investigation tool for simulations in computational chemistry and materials science. In these simulations, atoms move according to so-called potentials, functions that typically describe the distance-dependent interactions between pairs of atoms. For some applications more complex "many-body potentials" are required. From a HPC perspective, such potentials pose challenges: small neighborhoods hinder vectorization; redundant force expressions are tedious and error prone to derive; the implementations are large, runtime-critical, and can not be broken into simpler "kernels". Consequently, only selected many-body potentials receive optimizations---with major effort.

PotC is a DSL for MD potentials and corresponding compiler to generate high-performance implementations. The compiler centralizes optimization knowledge for many-body potentials, and unburdens users from manual optimization and force derivation. The performance optimization of the generated code is still work-in-progress. However, independent of performance, the tool can also be applied when prototyping, testing, or possibly symbolically manipulating potentials.

OoO Instruction Benchmarking Framework on the Back of Dragons

Julian Hammer (University of Erlangen-Nuremberg, RRZE)

In order to construct an accurate instruction execution model for modern out-of-order micro architectures, an accurate description of instruction latency, throughput and concurrency is indispensable. Already existing resources and vendor provided information is neither complete nor detailed enough and sometimes incorrect. We therefore proclaim to deduct this information through runtime instruction benchmarking and present a framework to support such investigations based on LLVM's just-in-time and cross-platform compilation capabilities.

pyasmjit abstracts instructions, operands, and dependency chains, to easily construct the necessary benchmarks. The synthesized code is interactively compiled and executed using the llvmlite library, which in turn is based on the stable LLVM C-API. pyasmjit offers a command line as well as a programming interface. Unlike other approaches, we do not rely on model specific performance counters and focus on interoperability and automation to support quick modeling of future microarchitectures.

Studying the Impact of Power Capping on MapReduce-Based, Data-Intensive Mini-Applications on

Intel KNL and KNM Architectures
Joshua H. Davis (University of Delaware)

In this poster, we quantitatively measure the impacts of data movement on performance in MapReduce-based applications when executed on HPC systems. We leverage the PAPI 'powercap' component to identify ideal conditions for execution of our applications in terms of (1) dataset characteristics (i.e., unique words); (2) HPC system (i.e., KNL and KNM); and (3) implementation of the MapReduce programming model (i.e., with or without combiner optimizations). Results confirm the high energy and runtime costs of data movement, and the benefits of the combiner optimization on these costs.

Accelerating Big Data Processing in the Cloud with Scalable Communication and I/O Schemes
Shashank Gugnani (Ohio State University)

With the advent of cloud computing, the field of Big Data has seen rapid growth. Most cloud providers provide hardware resources such as NVMe SSDs, large memory nodes, and SR-IOV. This opens up the possibility of large-scale high-performance data analytics and provides opportunities to use these resources to develop new designs. Cloud computing provides flexibility, security, and reliability, which are important requirements for Big Data frameworks. However, several important requirements are missing, such as performance, scalability, consistency, and quality of service (QoS). The focus of this work revolves around developing communication and I/O designs and concepts which can provide these requirements to Big Data frameworks. Specifically, we explore new ways to provide QoS and consistency in cloud storage systems, and provide scalable and high-performance communication frameworks.

Accelerating 2D FFT: Exploit GPU Tensor Cores through Mixed-Precision
Xiaohe Cheng (Hong Kong University of Science and Technology), Anumeena Sorna (National Institute of Technology, Tiruchirappalli)

The two-dimensional Fourier Transform is a widely-used computational kernel in many HPC applications. The popular NVIDIA cuFFT library provides a simple interface to compute 2D FFT on GPUs, but it's yet to utilize the recent hardware advancement in half-precision floating-point arithmetic. In this poster, we propose a mixed-precision method to accelerate 2D FFT by exploiting the FP16 matrix-multiply-and-accumulate units on the newest GPU architecture, known as tensor cores. We achieve a balance between speed and accuracy by dynamically splitting the single-precision input data into two half-precision operands and performing FFT separately. We present a CUDA-based implementation that achieves 3-digit more accuracy than half-precision cuFFT. We also demonstrate the stability and scalability of our approach and conclude that it attains high accuracy with tolerable splitting overhead.

Monitoring Parsl Workflows
Connor Pigg (University of Illinois)

As a Python library that enables workflows, Parsl gives users the ability to define complex workflows in Python and run them in parallel on any computer system. This poster describe the process of adding

monitoring to Parsl. Simple and comprehensive monitoring of a workflow's state and resource usage lets users audit, debug, and confirm workflow execution. The poster discusses how Parsl monitors workflow components, what data it captures (task status and resource usage), and the tools it used to do so (Elasticsearch) and to display the information (Kibana). A Kibana dashboard visualizes the collected logs in a real time, with an interactive user interface. This enhanced Parsl allows users the option to monitor the status and resource usage of their workflows via an Elasticsearch database and Kibana dashboard.

Identifying Network Data Transfer Bottlenecks in HPC Systems

Karen Tu (Lawrence Berkeley National Laboratory; University of California, Berkeley)

Improving network data transfer performance is a major factor for improving high performance computing systems. Most studies analyze data transfer and file system IO performance separately, but understanding the relationship between the two is essential for optimizing scheduling and resource management. Intuitively, if data is being transferred to a busy file system the transfer rate would be slower than a file system at regular activity levels.

This study analyzes patterns between file system activity and network throughput for several use cases of file writing and data transfers using a parallel file system. The parameters changed among the use cases were file striping for the file system, and buffer size and parallelism for data transfer. The main bottleneck for network data transfer rate was the number of OSTs the data was striped across. For a large number of OSTs (16 or greater), writing to the file system was the bottleneck.

Dendro-GR: Massively Parallel Simulations of Binary Black Hole Intermediate-Mass-Ratio Inspirals

Milinda Fernando (University of Utah)

We present a portable and highly-scalable algorithm and framework that targets problems in the astrophysics and numerical relativity communities. This framework combines together a parallel octree-refined adaptive mesh with wavelet adaptive multiresolution and a physics module to solve the Einstein equations of general relativity in the BSSN~formulation. The goal of this work is to perform advanced, massively parallel numerical simulations of Intermediate Mass Ratio Inspirals (IMRIs) of binary black holes with mass ratios on the order of 100:1. These studies will be used to generate waveforms for use in LIGO data analysis and to calibrate semi-analytical approximate methods. This advanced framework is designed to easily accommodate many existing algorithms in astrophysics for plasma dynamics and radiation hydrodynamics. We have designed novel algorithms to enable efficient simulations for such experiments and demonstrate excellent weak scalability up to 131K cores on ORNL's Titan for binary mergers for mass ratios up to 100.

Hardware Transactional Persistent Memory

Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel

persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories

Luann C. Jung (Massachusetts Institute of Technology, University of Chicago), Brendan T. Whitaker (Ohio State University, University of Chicago)

As scientific data repositories and filesystems grow in size and complexity, they become increasingly disorganized. The coupling of massive quantities of data with poor organization makes it challenging for scientists to locate and utilize relevant data, thus slowing the process of analyzing data of interest. To address these issues, we explore an automated clustering approach for quantifying the organization of data repositories. Our parallel pipeline processes heterogeneous filetypes (e.g., text and tabular data), automatically clusters files based on content and metadata similarities, and computes a novel "cleanliness" score from the resulting clustering. We demonstrate the generation and accuracy of our cleanliness measure using both synthetic and real datasets, and conclude that it is more consistent than other potential cleanliness measures.

Supercomputing for the Multi-Driver Routing

Zeyang Ye (Stony Brook University)

Supercomputing is essential for routing traffic by providing drivers the optimal routes with minimal traveling distances or time. The unique challenges that require supercomputers to overcome are of multiple folds: numerous drivers, massive simultaneous requests, multiple locations, and needs of instant gratifications, etc. We developed two parallel methods, PSAD and PSAD-M, by using domain decomposition and state-mixing techniques. On the same computing platform with 96 cores, for the same problem, our PSAD methods outperform all published benchmarks by over a hundred times, while improving the solution quality. For the same routing problem on 384 cores, our PSAD-M reduced the elapsed time from the unbearable ten minutes to the reasonable 5 seconds, achieving a record-breaking speedup of 170. By providing instant routing solutions that enable online recommendations, our methods break the bottleneck of the widely adopted offline approaches.

NautDB: Toward a Hybrid Runtime for Processing Compiled Queries

Samuel Grayson (University of Texas, Dallas)

General purpose operating and database system suffer under the load of their generality which makes achieving optimal performance extremely hard, especially on modern hardware. The goal of this research is to integrate, for the first time, specialization techniques from the OS community (hybrid runtimes) and DB community (compiled queries) for high-performance query processing on modern hardware. We envision a system called NautDB, a hybrid dataflow runtime for executing compiled

queries. As a first step toward our goal, we evaluate the performance of compiled queries on Linux and run as a Nautilus hybrid runtime using a simple prototype. Our results demonstrate that combining these specialization techniques has transformative potential for building the next generation (distributed) high-performance query processing systems and big data platforms.

Mitigating Performance and Progress Variability in Iterative Asynchronous Algorithms
Justs Zarins (University of Edinburgh)

Large HPC machines are susceptible to irregular performance. Factors like chip manufacturing differences, heat management, and network congestion combine to result in varying execution time for the same code and input sets. Asynchronous algorithms offer a partial solution. In these algorithms, fast workers are not forced to synchronize with slow ones. Instead they continue computing updates, and moving toward the solution, using the latest data available to them, which may have become stale (i.e. a number of iterations out of date compared to the most recent data). While this allows for high computational efficiency, the convergence rate of asynchronous algorithms tends to be lower.

To address this problem, we are using the unique properties of asynchronous algorithms to develop load balancing strategies for iterative asynchronous algorithms in both shared and distributed memory. Our poster shows how our solution attenuates noise, resulting in significant reduction progress imbalance and time-to-solution variability.

Eulerian Algorithms for the Discretization of Plasma Kinetic Equations
James L. Juno (University of Maryland)

While fluid models are common tools in the study of plasmas, many of these systems, whether in astrophysics or the lab, are only weakly collisional and far from equilibrium, making them more accurately described by kinetic equations. Kinetic equations can be computationally demanding due to the need to solve for the distribution function of the particles in a higher dimensional phase space, with position and velocity coordinates. Despite this challenge, the motivation for solving the plasma kinetic equation is large as there remains a vast array of questions concerning collisionless dynamics in real plasma systems. Here we present algorithms in an Eulerian framework for the discretization of the plasma kinetic equation, using a high-order discontinuous Galerkin finite element method due to its arithmetic intensity and parallelizability. Scaling and performance of the algorithm are discussed, and benchmarks of the algorithm are presented as well.

Room: C145
3:00 pm - 5:00 pm

ACM Student Research Competition

Precomputing Outputs of Hidden Layers to Speed Up Deep Neural Network Training
Sohil Lal Shrestha (University of Texas, Arlington)

Deep learning has recently emerged as a powerful technique for many tasks including image classification. A key bottleneck of deep learning is that the training phase takes a lot of time, since state-of-the-art deep neural networks have millions of parameters and hundreds of hidden layers. The early layers of these deep neural networks have the fewest parameters but take up the most computation.

In this work, we reduce training time by progressively freezing hidden layers, pre-computing their output and excluding them from training in both forward and backward paths in subsequent iterations. We compare this technique to the most closely related approach for speeding up the training process of neural network.

Through experiments on two widely used datasets for image classification, we empirically demonstrate that our approach can yield savings of up to 25% wall-clock time during training with no loss in accuracy.

Designing Shared Address Space MPI Libraries in Many-Core Era

Jahanzeb Maqbool Hashmi (Ohio State University)

The emergence of modern multi-/many-cores has put more emphasis on optimizing intra-node communication. Existing designs in MPI libraries that work on the concept of distributed address spaces incur the overhead of intermediate memory copies to stage the data between processes. This can lead to severe performance degradation especially on emerging many-core architectures like Intel Skylake and IBM OpenPOWER. This work proposes a high-performance "shared address-space"-based MPI point-to-point and collective communication designs using XPMEM. We first characterize the bottlenecks associated with XPMEM based communication and propose new designs for efficient MPI large message communication. Then we propose novel collective designs that are contention-free and offer true zero-copy reduction operations. The proposed designs are evaluated on different multi-/many-core architectures using various micro-benchmarks and application kernels such as MiniAMR and AlexNet DNN training on CNTK. The proposed designs have shown significant performance improvement over state-of-the-art available in MPI libraries.

Modeling Single-Source Shortest Path Algorithm Dynamics to Control Performance and Power Tradeoffs

Sara Karamati (Georgia Institute of Technology), Jeffrey Young (Georgia Institute of Technology), Rich Vuduc (Georgia Institute of Technology)

This work presents a new methodology to improve the performance of parallel algorithms by tuning the amount of available parallelism for execution throughout the runtime. As such, we expose key parameters controlling the performance and parallelism of the algorithm and build a software-based controller with the objective of maintaining the optimal performance. Our controller allows for tuning the level of parallelism executed in each time epoch to optimize for performance while preserving power usage. More specifically, our experimental evaluation focuses on a tunable variation of a GPU-based delta-stepping algorithm for computing the single-source shortest path (SSSP); As the available parallelism for the delta-stepping SSSP is highly irregular and strongly input-dependent, our extensive experiments show that average power can be reduced while average parallelism is increased. This increase in average parallelism provides substantial energy savings, independent of the hardware.

Accelerating DNA Long Read Mapping with Emerging Technologies
Roman Kaplan (Israel Institute of Technology)

DNA sequencing technologies output only short fragments of a genome, called reads. New single-molecule real-time sequencing technologies can produce long reads, up to tens of thousands base pairs, within minutes. However, these long reads may contain up to 15% errors.

To construct a genome from DNA reads, a computationally expensive bioinformatics task, read mapping, is required. Read mapping finds the best-fitting location for each DNA read on a long reference sequence. The length and error rate of long reads poses a challenge for existing read mapping hardware solutions, designed for short reads with low error rates. This work presents a novel DNA read mapping hardware architecture, RASSA. RASSA is a Resistive Approximate Similarity Search Accelerator that exploits charge distribution and parallel in-memory processing to reflect a mismatch count between DNA sequences. RASSA implementation of long read DNA mapping outperforms state-of-the-art long read mapping solution by 16-77x with comparable accuracy.

SimFS: A Simulation Data Virtualizing File System Interface
Salvatore Di Girolamo (ETH Zurich)

In the big (simulation) data era, simulations often produce petabytes of data to be stored in parallel filesystems or large-scale databases. This data is accessed, often by thousands of analysts and scientists, over the course of decades. However, storing these volumes of data for long time periods of time is not cost effective and, in some cases, practically impossible.

SimFS transparently virtualizes the simulation output, relaxing the storage requirements and re-simulating missing data on-demand. SimFS monitors the analysis access pattern in order to decide (1) which data to store and (2) apply prefetching techniques to improve the analysis performance. SimFS enables a trade-off between on-disk solutions, where all the simulation data is stored on disk, and in-situ, where no data is stored and analyses are always coupled with simulations. Overall, by exploiting the growing computing power and relaxing the storage capacity requirements, SimFS offers a viable path towards exa-scale simulations.

Holistic Root Cause Analysis of Node Failures in Production HPC
Anwesha Das (North Carolina State University)

Production HPC clusters endure failures incurring computation and resource wastage. Despite the presence of various failure detection and prediction schemes, a comprehensive understanding of how nodes fail considering various components and layers of the system is required for sustained resilience. This work performs a holistic root cause diagnosis of node failures using a measurement-driven approach on contemporary system logs that can help vendors and system administrators support exascale resilience.

Our work shows that lead times can be increased by at least 5 times if external subsystem correlations are considered as opposed to considering the events of a specific node in isolation. Moreover, when detecting sensor measurement outliers and interconnect related failures, triggering automated

recovery events can exacerbate the situation if recovery is unsuccessful.

Geomancy: Automated Data Placement Optimization

Oceane Bel (University of California, Santa Cruz)

Exascale cloud storage and High-Performance Computing Systems (HPC) deliver unprecedented storage capacity and levels of computing power, though the full potential of these systems remain untapped because of inefficient data placement. Changes in data access patterns can cause a system's performance to suffer. To mitigate performance losses, system designers implement strategies to preemptively place popular data on higher performance nodes. However, these strategies fail to address a diverse userbase whose users individually demand the highest performance, and they must be carefully constructed by an expert of the system.

We propose Geomancy, a tool that reorganizes data to increase I/O throughput. In systems where heuristic-based improvements may become resource intensive, Geomancy determines new placement policies by training a deep neural network with past workload and system traces. With real workload traces, Geomancy calculated an example placement policy that demonstrated a 49% increase in average throughput compared to the default data layout.

Numerical Simulation of a Flue Instrument with Finite-Difference Lattice Boltzmann Method using GPGPU

Ryoya Tabata (Kyushu Institute of Technology)

In this work, we discuss the possibility of using GPGPU techniques for Aeroacoustic Simulation (especially for flue instruments) with the finite-difference lattice Boltzmann method (FDLBM). Compressible flow simulation has been used in direct aeroacoustic simulation; however, the computational cost is huge due to the requirement of high computational mesh resolution, with small time steps. The lattice Boltzmann method (LBM) has been used as an efficient method for fluid simulation using GPGPU. However, LBM is not accurate enough when applied to some aeroacoustic problems. On the other hand, FDLBM is able to use high-order finite-difference schemes and it has a high arithmetic intensity compared to LBM. We present a performance evaluation of the LBM and FDLBM with several finite-difference schemes on GPU with the roofline model.

Recursive Algebraic Coloring Engine

Christie Louis Alappat (University of Erlangen-Nuremberg)

Many iterative numerical methods for sparse systems and building blocks of sparse linear algebra are difficult to parallelize due to data dependencies. These may be loop-carried dependencies as they occur in solvers like Gauss-Seidel or write conflicts as in symmetric sparse matrix vector. Most of the existing parallelization strategies suffer from low performance on modern hardware, are matrix specific, or require tailored storage formats.

In this poster, we introduce a novel recursive level based algorithm called Recursive Algebraic Coloring (RAC), which achieves high hardware efficiency on modern multi-core architectures and works with simple data formats like compressed row storage. Method is implemented in a library called Recursive Algebraic

Coloring Engine (RACE). Thorough performance analysis shows that RACE outperforms traditional multicoloring methods and Intel-MKL implementations with a factor of 2–2.5×. We are on par with Algebraic Block Multicoloring for small matrices, while for large matrices we gain a factor of 1.5–2×.

Accelerating Microscope Data Analysis Using Parallel Computing
John Ravi (North Carolina State University)

Single-Molecule Localization Microscopy (SMLM) techniques deal with the diffraction limit of fluorescent microscopy by localizing single molecules with high precision by stochastically switching molecules on and off. Thousands of camera frames containing subsets of blinking molecules are recorded to obtain a single super-resolution image. Each blinking molecule in each frame is subjected to localization protocols that fit the shape of the blink, assess the quality of the blink and then estimate their center. The algorithm implemented originally in MATLAB and compiled CUDA C, to compute a 'Super Resolution' image took around 6 minutes to process 256x256 pixel images of a moderately dense dataset. I ported the algorithm to C++ and parallelized it using OpenMP to compute multiple frames in parallel.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

PotC: Many-Body Potential Implementations à La Carte
Markus Höhnerbach (RWTH Aachen University)

Molecular dynamics is a valuable investigation tool for simulations in computational chemistry and materials science. In these simulations, atoms move according to so-called potentials, functions that typically describe the distance-dependent interactions between pairs of atoms. For some applications more complex "many-body potentials" are required. From a HPC perspective, such potentials pose challenges: small neighborhoods hinder vectorization; redundant force expressions are tedious and error prone to derive; the implementations are large, runtime-critical, and can not be broken into simpler "kernels". Consequently, only selected many-body potentials receive optimizations---with major effort.

PotC is a DSL for MD potentials and corresponding compiler to generate high-performance implementations. The compiler centralizes optimization knowledge for many-body potentials, and unburdens users from manual optimization and force derivation. The performance optimization of the

generated code is still work-in-progress. However, independent of performance, the tool can also be applied when prototyping, testing, or possibly symbolically manipulating potentials.

OoO Instruction Benchmarking Framework on the Back of Dragons

Julian Hammer (University of Erlangen-Nuremberg, RRZE)

In order to construct an accurate instruction execution model for modern out-of-order micro architectures, an accurate description of instruction latency, throughput and concurrency is indispensable. Already existing resources and vendor provided information is neither complete nor detailed enough and sometimes incorrect. We therefore proclaim to deduct this information through runtime instruction benchmarking and present a framework to support such investigations based on LLVM's just-in-time and cross-platform compilation capabilities.

pyasmjit abstracts instructions, operands, and dependency chains, to easily construct the necessary benchmarks. The synthesized code is interactively compiled and executed using the llvmlite library, which in turn is based on the stable LLVM C-API. pyasmjit offers a command line as well as a programming interface. Unlike other approaches, we do not rely on model specific performance counters and focus on interoperability and automation to support quick modeling of future microarchitectures.

Studying the Impact of Power Capping on MapReduce-Based, Data-Intensive Mini-Applications on Intel KNL and KNM Architectures

Joshua H. Davis (University of Delaware)

In this poster, we quantitatively measure the impacts of data movement on performance in MapReduce-based applications when executed on HPC systems. We leverage the PAPI 'powercap' component to identify ideal conditions for execution of our applications in terms of (1) dataset characteristics (i.e., unique words); (2) HPC system (i.e., KNL and KNM); and (3) implementation of the MapReduce programming model (i.e., with or without combiner optimizations). Results confirm the high energy and runtime costs of data movement, and the benefits of the combiner optimization on these costs.

Accelerating Big Data Processing in the Cloud with Scalable Communication and I/O Schemes

Shashank Gugnani (Ohio State University)

With the advent of cloud computing, the field of Big Data has seen rapid growth. Most cloud providers provide hardware resources such as NVMe SSDs, large memory nodes, and SR-IOV. This opens up the possibility of large-scale high-performance data analytics and provides opportunities to use these resources to develop new designs. Cloud computing provides flexibility, security, and reliability, which are important requirements for Big Data frameworks. However, several important requirements are missing, such as performance, scalability, consistency, and quality of service (QoS). The focus of this work revolves around developing communication and I/O designs and concepts which can provide these requirements to Big Data frameworks. Specifically, we explore new ways to provide QoS and consistency in cloud storage systems, and provide scalable and high-performance communication frameworks.

Accelerating 2D FFT: Exploit GPU Tensor Cores through Mixed-Precision

Xiaohe Cheng (Hong Kong University of Science and Technology), Anumeena Sorna (National Institute of Technology, Tiruchirappalli)

The two-dimensional Fourier Transform is a widely-used computational kernel in many HPC applications. The popular NVIDIA cuFFT library provides a simple interface to compute 2D FFT on GPUs, but it's yet to utilize the recent hardware advancement in half-precision floating-point arithmetic. In this poster, we propose a mixed-precision method to accelerate 2D FFT by exploiting the FP16 matrix-multiply-and-accumulate units on the newest GPU architecture, known as tensor cores. We achieve a balance between speed and accuracy by dynamically splitting the single-precision input data into two half-precision operands and performing FFT separately. We present a CUDA-based implementation that achieves 3-digit more accuracy than half-precision cuFFT. We also demonstrate the stability and scalability of our approach and conclude that it attains high accuracy with tolerable splitting overhead.

Monitoring Parsl Workflows

Connor Pigg (University of Illinois)

As a Python library that enables workflows, Parsl gives users the ability to define complex workflows in Python and run them in parallel on any computer system. This poster describe the process of adding monitoring to Parsl. Simple and comprehensive monitoring of a workflow's state and resource usage lets users audit, debug, and confirm workflow execution. The poster discusses how Parsl monitors workflow components, what data it captures (task status and resource usage), and the tools it used to do so (Elasticsearch) and to display the information (Kibana). A Kibana dashboard visualizes the collected logs in a real time, with an interactive user interface. This enhanced Parsl allows users the option to monitor the status and resource usage of their workflows via an Elasticsearch database and Kibana dashboard.

Identifying Network Data Transfer Bottlenecks in HPC Systems

Karen Tu (Lawrence Berkeley National Laboratory; University of California, Berkeley)

Improving network data transfer performance is a major factor for improving high performance computing systems. Most studies analyze data transfer and file system IO performance separately, but understanding the relationship between the two is essential for optimizing scheduling and resource management. Intuitively, if data is being transferred to a busy file system the transfer rate would be slower than a file system at regular activity levels.

This study analyzes patterns between file system activity and network throughput for several use cases of file writing and data transfers using a parallel file system. The parameters changed among the use cases were file striping for the file system, and buffer size and parallelism for data transfer. The main bottleneck for network data transfer rate was the number of OSTs the data was striped across. For a large number of OSTs (16 or greater), writing to the file system was the bottleneck.

Dendro-GR: Massively Parallel Simulations of Binary Black Hole Intermediate-Mass-Ratio Inspirals

Milinda Fernando (University of Utah)

We present a portable and highly-scalable algorithm and framework that targets problems in the astrophysics and numerical relativity communities. This framework combines together a parallel octree-refined adaptive mesh with wavelet adaptive multiresolution and a physics module to solve the Einstein equations of general relativity in the BSSN~formulation. The goal of this work is to perform advanced, massively parallel numerical simulations of Intermediate Mass Ratio Inspirals (IMRIs) of binary black holes with mass ratios on the order of 100:1. These studies will be used to generate waveforms for use in LIGO data analysis and to calibrate semi-analytical approximate methods. This advanced framework is designed to easily accommodate many existing algorithms in astrophysics for plasma dynamics and radiation hydrodynamics. We have designed novel algorithms to enable efficient simulations for such experiments and demonstrate excellent weak scalability up to 131K cores on ORNL's Titan for binary mergers for mass ratios up to 100.

Hardware Transactional Persistent Memory

Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories

Luann C. Jung (Massachusetts Institute of Technology, University of Chicago), Brendan T. Whitaker (Ohio State University, University of Chicago)

As scientific data repositories and filesystems grow in size and complexity, they become increasingly disorganized. The coupling of massive quantities of data with poor organization makes it challenging for scientists to locate and utilize relevant data, thus slowing the process of analyzing data of interest. To address these issues, we explore an automated clustering approach for quantifying the organization of data repositories. Our parallel pipeline processes heterogeneous filetypes (e.g., text and tabular data), automatically clusters files based on content and metadata similarities, and computes a novel "cleanliness" score from the resulting clustering. We demonstrate the generation and accuracy of our cleanliness measure using both synthetic and real datasets, and conclude that it is more consistent than other potential cleanliness measures.

Supercomputing for the Multi-Driver Routing

Zeyang Ye (Stony Brook University)

Supercomputing is essential for routing traffic by providing drivers the optimal routes with minimal traveling distances or time. The unique challenges that require supercomputers to overcome are of multiple folds: numerous drivers, massive simultaneous requests, multiple locations, and needs of instant gratifications, etc. We developed two parallel methods, PSAD and PSAD-M, by using domain decomposition and state-mixing techniques. On the same computing platform with 96 cores, for the same problem, our PSAD methods outperform all published benchmarks by over a hundred times, while improving the solution quality. For the same routing problem on 384 cores, our PSAD-M reduced the elapsed time from the unbearable ten minutes to the reasonable 5 seconds, achieving a record-breaking speedup of 170. By providing instant routing solutions that enable online recommendations, our methods break the bottleneck of the widely adopted offline approaches.

NautDB: Toward a Hybrid Runtime for Processing Compiled Queries

Samuel Grayson (University of Texas, Dallas)

General purpose operating and database system suffer under the load of their generality which makes achieving optimal performance extremely hard, especially on modern hardware. The goal of this research is to integrate, for the first time, specialization techniques from the OS community (hybrid runtimes) and DB community (compiled queries) for high-performance query processing on modern hardware. We envision a system called NautDB, a hybrid dataflow runtime for executing compiled queries. As a first step toward our goal, we evaluate the performance of compiled queries on Linux and run as a Nautilus hybrid runtime using a simple prototype. Our results demonstrate that combining these specialization techniques has transformative potential for building the next generation (distributed) high-performance query processing systems and big data platforms.

Mitigating Performance and Progress Variability in Iterative Asynchronous Algorithms

Justs Zarins (University of Edinburgh)

Large HPC machines are susceptible to irregular performance. Factors like chip manufacturing differences, heat management, and network congestion combine to result in varying execution time for the same code and input sets. Asynchronous algorithms offer a partial solution. In these algorithms, fast workers are not forced to synchronize with slow ones. Instead they continue computing updates, and moving toward the solution, using the latest data available to them, which may have become stale (i.e. a number of iterations out of date compared to the most recent data). While this allows for high computational efficiency, the convergence rate of asynchronous algorithms tends to be lower.

To address this problem, we are using the unique properties of asynchronous algorithms to develop load balancing strategies for iterative asynchronous algorithms in both shared and distributed memory. Our poster shows how our solution attenuates noise, resulting in significant reduction progress imbalance and time-to-solution variability.

Eulerian Algorithms for the Discretization of Plasma Kinetic Equations

James L. Juno (University of Maryland)

While fluid models are common tools in the study of plasmas, many of these systems, whether in astrophysics or the lab, are only weakly collisional and far from equilibrium, making them more accurately described by kinetic equations. Kinetic equations can be computationally demanding due to the need to solve for the distribution function of the particles in a higher dimensional phase space, with position and velocity coordinates. Despite this challenge, the motivation for solving the plasma kinetic equation is large as there remains a vast array of questions concerning collisionless dynamics in real plasma systems. Here we present algorithms in an Eulerian framework for the discretization of the plasma kinetic equation, using a high-order discontinuous Galerkin finite element method due to its arithmetic intensity and parallelizability. Scaling and performance of the algorithm are discussed, and benchmarks of the algorithm are presented as well.

Thursday, November 15th

Room: C2/3/4 Ballroom
8:30 am - 5:00 pm

ACM Student Research Competition Posters

Session Description: SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

SC18 Research Posters
SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Precomputing Outputs of Hidden Layers to Speed Up Deep Neural Network Training
Sohil Lal Shrestha (University of Texas, Arlington)

Deep learning has recently emerged as a powerful technique for many tasks including image classification. A key bottleneck of deep learning is that the training phase takes a lot of time, since state-of-the-art deep neural networks have millions of parameters and hundreds of hidden layers. The early layers of these deep neural networks have the fewest parameters but take up the most computation.

In this work, we reduce training time by progressively freezing hidden layers, pre-computing their output and excluding them from training in both forward and backward paths in subsequent iterations. We compare this technique to the most closely related approach for speeding up the training process of neural network.

Through experiments on two widely used datasets for image classification, we empirically demonstrate that our approach can yield savings of up to 25% wall-clock time during training with no loss in accuracy.

Designing Shared Address Space MPI Libraries in Many-Core Era

Jahanzeb Maqbool Hashmi (Ohio State University)

The emergence of modern multi-/many-cores has put more emphasis on optimizing intra-node communication. Existing designs in MPI libraries that work on the concept of distributed address spaces incur the overhead of intermediate memory copies to stage the data between processes. This can lead to severe performance degradation especially on emerging many-core architectures like Intel Skylake and IBM OpenPOWER. This work proposes a high-performance "shared address-space"-based MPI point-to-point and collective communication designs using XPMEM. We first characterize the bottlenecks associated with XPMEM based communication and propose new designs for efficient MPI large message communication. Then we propose novel collective designs that are contention-free and offer true zero-copy reduction operations. The proposed designs are evaluated on different multi-/many-core architectures using various micro-benchmarks and application kernels such as MiniAMR and AlexNet DNN training on CNTK. The proposed designs have shown significant performance improvement over state-of-the-art available in MPI libraries.

Modeling Single-Source Shortest Path Algorithm Dynamics to Control Performance and Power Tradeoffs

Sara Karamati (Georgia Institute of Technology), Jeffrey Young (Georgia Institute of Technology), Rich Vuduc (Georgia Institute of Technology)

This work presents a new methodology to improve the performance of parallel algorithms by tuning the amount of available parallelism for execution throughout the runtime. As such, we expose key parameters controlling the performance and parallelism of the algorithm and build a software-based controller with the objective of maintaining the optimal performance. Our controller allows for tuning the level of parallelism executed in each time epoch to optimize for performance while preserving power usage. More specifically, our experimental evaluation focuses on a tunable variation of a GPU-based delta-stepping algorithm for computing the single-source shortest path (SSSP); As the available parallelism for the delta-stepping SSSP is highly irregular and strongly input-dependent, our extensive experiments show that average power can be reduced while average parallelism is increased. This increase in average parallelism provides substantial energy savings, independent of the hardware.

Accelerating DNA Long Read Mapping with Emerging Technologies

Roman Kaplan (Israel Institute of Technology)

DNA sequencing technologies output only short fragments of a genome, called reads. New single-molecule real-time sequencing technologies can produce long reads, up to tens of thousands base pairs, within minutes. However, these long reads may contain up to 15% errors.

To construct a genome from DNA reads, a computationally expensive bioinformatics task, read mapping, is required. Read mapping finds the best-fitting location for each DNA read on a long reference sequence. The length and error rate of long reads poses a challenge for existing read mapping hardware solutions, designed for short reads with low error rates. This work presents a novel DNA read mapping hardware architecture, RASSA. RASSA is a Resistive Approximate Similarity Search Accelerator that exploits charge distribution and parallel in-memory processing to reflect a mismatch count between DNA sequences. RASSA implementation of long read DNA mapping

outperforms state-of-the-art long read mapping solution by 16-77x with comparable accuracy.

SimFS: A Simulation Data Virtualizing File System Interface
Salvatore Di Girolamo (ETH Zurich)

In the big (simulation) data era, simulations often produce petabytes of data to be stored in parallel filesystems or large-scale databases. This data is accessed, often by thousands of analysts and scientists, over the course of decades. However, storing these volumes of data for long time periods of time is not cost effective and, in some cases, practically impossible.

SimFS transparently virtualizes the simulation output, relaxing the storage requirements and re-simulating missing data on-demand. SimFS monitors the analysis access pattern in order to decide (1) which data to store and (2) apply prefetching techniques to improve the analysis performance. SimFS enables a trade-off between on-disk solutions, where all the simulation data is stored on disk, and in-situ, where no data is stored and analyses are always coupled with simulations. Overall, by exploiting the growing computing power and relaxing the storage capacity requirements, SimFS offers a viable path towards exa-scale simulations.

Holistic Root Cause Analysis of Node Failures in Production HPC
Anwesha Das (North Carolina State University)

Production HPC clusters endure failures incurring computation and resource wastage. Despite the presence of various failure detection and prediction schemes, a comprehensive understanding of how nodes fail considering various components and layers of the system is required for sustained resilience. This work performs a holistic root cause diagnosis of node failures using a measurement-driven approach on contemporary system logs that can help vendors and system administrators support exascale resilience.

Our work shows that lead times can be increased by at least 5 times if external subsystem correlations are considered as opposed to considering the events of a specific node in isolation. Moreover, when detecting sensor measurement outliers and interconnect related failures, triggering automated recovery events can exacerbate the situation if recovery is unsuccessful.

Geomancy: Automated Data Placement Optimization
Oceane Bel (University of California, Santa Cruz)

Exascale cloud storage and High-Performance Computing Systems (HPC) deliver unprecedented storage capacity and levels of computing power, though the full potential of these systems remain untapped because of inefficient data placement. Changes in data access patterns can cause a system's performance to suffer. To mitigate performance losses, system designers implement strategies to preemptively place popular data on higher performance nodes. However, these strategies fail to address a diverse userbase whose users individually demand the highest performance, and they must be carefully constructed by an expert of the system.

We propose Geomancy, a tool that reorganizes data to increase I/O throughput. In systems where

heuristic-based improvements may become resource intensive, Geomancy determines new placement policies by training a deep neural network with past workload and system traces. With real workload traces, Geomancy calculated an example placement policy that demonstrated a 49% increase in average throughput compared to the default data layout.

Numerical Simulation of a Flue Instrument with Finite-Difference Lattice Boltzmann Method using GPGPU

Ryoya Tabata (Kyushu Institute of Technology)

In this work, we discuss the possibility of using GPGPU techniques for Aeroacoustic Simulation (especially for flue instruments) with the finite-difference lattice Boltzmann method (FDLBM). Compressible flow simulation has been used in direct aeroacoustic simulation; however, the computational cost is huge due to the requirement of high computational mesh resolution, with small time steps. The lattice Boltzmann method (LBM) has been used as an efficient method for fluid simulation using GPGPU. However, LBM is not accurate enough when applied to some aeroacoustic problems. On the other hand, FDLBM is able to use high-order finite-difference schemes and it has a high arithmetic intensity compared to LBM. We present a performance evaluation of the LBM and FDLBM with several finite-difference schemes on GPU with the roofline model.

Recursive Algebraic Coloring Engine

Christie Louis Alappat (University of Erlangen-Nuremberg)

Many iterative numerical methods for sparse systems and building blocks of sparse linear algebra are difficult to parallelize due to data dependencies. These may be loop-carried dependencies as they occur in solvers like Gauss-Seidel or write conflicts as in symmetric sparse matrix vector. Most of the existing parallelization strategies suffer from low performance on modern hardware, are matrix specific, or require tailored storage formats.

In this poster, we introduce a novel recursive level based algorithm called Recursive Algebraic Coloring (RAC), which achieves high hardware efficiency on modern multi-core architectures and works with simple data formats like compressed row storage. Method is implemented in a library called Recursive Algebraic Coloring Engine (RACE). Thorough performance analysis shows that RACE outperforms traditional multicoloring methods and Intel-MKL implementations with a factor of 2–2.5×. We are on par with Algebraic Block Multicoloring for small matrices, while for large matrices we gain a factor of 1.5–2×.

Accelerating Microscope Data Analysis Using Parallel Computing

John Ravi (North Carolina State University)

Single-Molecule Localization Microscopy (SMLM) techniques deal with the diffraction limit of fluorescent microscopy by localizing single molecules with high precision by stochastically switching molecules on and off. Thousands of camera frames containing subsets of blinking molecules are recorded to obtain a single super-resolution image. Each blinking molecule in each frame is subjected to localization protocols that fit the shape of the blink, assess the quality of the blink and then estimate their center. The algorithm implemented originally in MATLAB and compiled CUDA C, to compute a 'Super Resolution' image took around 6 minutes to process 256x256 pixel images of a moderately

dense dataset. I ported the algorithm to C++ and parallelized it using OpenMP to compute multiple frames in parallel.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

PotC: Many-Body Potential Implementations à La Carte
Markus Höhnerbach (RWTH Aachen University)

Molecular dynamics is a valuable investigation tool for simulations in computational chemistry and materials science. In these simulations, atoms move according to so-called potentials, functions that typically describe the distance-dependent interactions between pairs of atoms. For some applications more complex "many-body potentials" are required. From a HPC perspective, such potentials pose challenges: small neighborhoods hinder vectorization; redundant force expressions are tedious and error prone to derive; the implementations are large, runtime-critical, and can not be broken into simpler "kernels". Consequently, only selected many-body potentials receive optimizations---with major effort.

PotC is a DSL for MD potentials and corresponding compiler to generate high-performance implementations. The compiler centralizes optimization knowledge for many-body potentials, and unburdens users from manual optimization and force derivation. The performance optimization of the generated code is still work-in-progress. However, independent of performance, the tool can also be applied when prototyping, testing, or possibly symbolically manipulating potentials.

OoO Instruction Benchmarking Framework on the Back of Dragons
Julian Hammer (University of Erlangen-Nuremberg, RRZE)

In order to construct an accurate instruction execution model for modern out-of-order micro architectures, an accurate description of instruction latency, throughput and concurrency is indispensable. Already existing resources and vendor provided information is neither complete nor detailed enough and sometimes incorrect. We therefore proclaim to deduct this information through runtime instruction benchmarking and present a framework to support such investigations based on LLVM's just-in-time and cross-platform compilation capabilities.

pyasmjit abstracts instructions, operands, and dependency chains, to easily construct the necessary

benchmarks. The synthesized code is interactively compiled and executed using the llvmlite library, which in turn is based on the stable LLVM C-API. pyasmjit offers a command line as well as a programming interface. Unlike other approaches, we do not rely on model specific performance counters and focus on interoperability and automation to support quick modeling of future microarchitectures.

Studying the Impact of Power Capping on MapReduce-Based, Data-Intensive Mini-Applications on Intel KNL and KNM Architectures

Joshua H. Davis (University of Delaware)

In this poster, we quantitatively measure the impacts of data movement on performance in MapReduce-based applications when executed on HPC systems. We leverage the PAPI 'powercap' component to identify ideal conditions for execution of our applications in terms of (1) dataset characteristics (i.e., unique words); (2) HPC system (i.e., KNL and KNM); and (3) implementation of the MapReduce programming model (i.e., with or without combiner optimizations). Results confirm the high energy and runtime costs of data movement, and the benefits of the combiner optimization on these costs.

Accelerating Big Data Processing in the Cloud with Scalable Communication and I/O Schemes

Shashank Gugnani (Ohio State University)

With the advent of cloud computing, the field of Big Data has seen rapid growth. Most cloud providers provide hardware resources such as NVMe SSDs, large memory nodes, and SR-IOV. This opens up the possibility of large-scale high-performance data analytics and provides opportunities to use these resources to develop new designs. Cloud computing provides flexibility, security, and reliability, which are important requirements for Big Data frameworks. However, several important requirements are missing, such as performance, scalability, consistency, and quality of service (QoS). The focus of this work revolves around developing communication and I/O designs and concepts which can provide these requirements to Big Data frameworks. Specifically, we explore new ways to provide QoS and consistency in cloud storage systems, and provide scalable and high-performance communication frameworks.

Accelerating 2D FFT: Exploit GPU Tensor Cores through Mixed-Precision

Xiaohe Cheng (Hong Kong University of Science and Technology), Anumeena Sorna (National Institute of Technology, Tiruchirappalli)

The two-dimensional Fourier Transform is a widely-used computational kernel in many HPC applications. The popular NVIDIA cuFFT library provides a simple interface to compute 2D FFT on GPUs, but it's yet to utilize the recent hardware advancement in half-precision floating-point arithmetic. In this poster, we propose a mixed-precision method to accelerate 2D FFT by exploiting the FP16 matrix-multiply-and-accumulate units on the newest GPU architecture, known as tensor cores. We achieve a balance between speed and accuracy by dynamically splitting the single-precision input data into two half-precision operands and performing FFT separately. We present a CUDA-based implementation that achieves 3-digit more accuracy than half-precision cuFFT. We also demonstrate the stability and scalability of our approach and conclude that it attains high accuracy with tolerable

splitting overhead.

Monitoring Parsl Workflows

Connor Pigg (University of Illinois)

As a Python library that enables workflows, Parsl gives users the ability to define complex workflows in Python and run them in parallel on any computer system. This poster describe the process of adding monitoring to Parsl. Simple and comprehensive monitoring of a workflow's state and resource usage lets users audit, debug, and confirm workflow execution. The poster discusses how Parsl monitors workflow components, what data it captures (task status and resource usage), and the tools it used to do so (Elasticsearch) and to display the information (Kibana). A Kibana dashboard visualizes the collected logs in a real time, with an interactive user interface. This enhanced Parsl allows users the option to monitor the status and resource usage of their workflows via an Elasticsearch database and Kibana dashboard.

Identifying Network Data Transfer Bottlenecks in HPC Systems

Karen Tu (Lawrence Berkeley National Laboratory; University of California, Berkeley)

Improving network data transfer performance is a major factor for improving high performance computing systems. Most studies analyze data transfer and file system IO performance separately, but understanding the relationship between the two is essential for optimizing scheduling and resource management. Intuitively, if data is being transferred to a busy file system the transfer rate would be slower than a file system at regular activity levels.

This study analyzes patterns between file system activity and network throughput for several use cases of file writing and data transfers using a parallel file system. The parameters changed among the use cases were file striping for the file system, and buffer size and parallelism for data transfer. The main bottleneck for network data transfer rate was the number of OSTs the data was striped across. For a large number of OSTs (16 or greater), writing to the file system was the bottleneck.

Dendro-GR: Massively Parallel Simulations of Binary Black Hole Intermediate-Mass-Ratio Inspirals

Milinda Fernando (University of Utah)

We present a portable and highly-scalable algorithm and framework that targets problems in the astrophysics and numerical relativity communities. This framework combines together a parallel octree-refined adaptive mesh with wavelet adaptive multiresolution and a physics module to solve the Einstein equations of general relativity in the BSSN~formulation. The goal of this work is to perform advanced, massively parallel numerical simulations of Intermediate Mass Ratio Inspirals (IMRIs) of binary black holes with mass ratios on the order of 100:1. These studies will be used to generate waveforms for use in LIGO data analysis and to calibrate semi-analytical approximate methods. This advanced framework is designed to easily accommodate many existing algorithms in astrophysics for plasma dynamics and radiation hydrodynamics. We have designed novel algorithms to enable efficient simulations for such experiments and demonstrate excellent weak scalability up to 131K cores on ORNL's Titan for binary mergers for mass ratios up to 100.

Hardware Transactional Persistent Memory

Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories

Luann C. Jung (Massachusetts Institute of Technology, University of Chicago), Brendan T. Whitaker (Ohio State University, University of Chicago)

As scientific data repositories and filesystems grow in size and complexity, they become increasingly disorganized. The coupling of massive quantities of data with poor organization makes it challenging for scientists to locate and utilize relevant data, thus slowing the process of analyzing data of interest. To address these issues, we explore an automated clustering approach for quantifying the organization of data repositories. Our parallel pipeline processes heterogeneous filetypes (e.g., text and tabular data), automatically clusters files based on content and metadata similarities, and computes a novel "cleanliness" score from the resulting clustering. We demonstrate the generation and accuracy of our cleanliness measure using both synthetic and real datasets, and conclude that it is more consistent than other potential cleanliness measures.

Supercomputing for the Multi-Driver Routing

Zeyang Ye (Stony Brook University)

Supercomputing is essential for routing traffic by providing drivers the optimal routes with minimal traveling distances or time. The unique challenges that require supercomputers to overcome are of multiple folds: numerous drivers, massive simultaneous requests, multiple locations, and needs of instant gratifications, etc. We developed two parallel methods, PSAD and PSAD-M, by using domain decomposition and state-mixing techniques. On the same computing platform with 96 cores, for the same problem, our PSAD methods outperform all published benchmarks by over a hundred times, while improving the solution quality. For the same routing problem on 384 cores, our PSAD-M reduced the elapsed time from the unbearable ten minutes to the reasonable 5 seconds, achieving a record-breaking speedup of 170. By providing instant routing solutions that enable online recommendations, our methods break the bottleneck of the widely adopted offline approaches.

NautDB: Toward a Hybrid Runtime for Processing Compiled Queries

Samuel Grayson (University of Texas, Dallas)

General purpose operating and database system suffer under the load of their generality which makes achieving optimal performance extremely hard, especially on modern hardware. The goal of this research is to integrate, for the first time, specialization techniques from the OS community (hybrid runtimes) and DB community (compiled queries) for high-performance query processing on modern hardware. We envision a system called NautDB, a hybrid dataflow runtime for executing compiled queries. As a first step toward our goal, we evaluate the performance of compiled queries on Linux and run as a Nautilus hybrid runtime using a simple prototype. Our results demonstrate that combining these specialization techniques has transformative potential for building the next generation (distributed) high-performance query processing systems and big data platforms.

Mitigating Performance and Progress Variability in Iterative Asynchronous Algorithms

Justs Zarins (University of Edinburgh)

Large HPC machines are susceptible to irregular performance. Factors like chip manufacturing differences, heat management, and network congestion combine to result in varying execution time for the same code and input sets. Asynchronous algorithms offer a partial solution. In these algorithms, fast workers are not forced to synchronize with slow ones. Instead they continue computing updates, and moving toward the solution, using the latest data available to them, which may have become stale (i.e. a number of iterations out of date compared to the most recent data). While this allows for high computational efficiency, the convergence rate of asynchronous algorithms tends to be lower.

To address this problem, we are using the unique properties of asynchronous algorithms to develop load balancing strategies for iterative asynchronous algorithms in both shared and distributed memory. Our poster shows how our solution attenuates noise, resulting in significant reduction progress imbalance and time-to-solution variability.

Eulerian Algorithms for the Discretization of Plasma Kinetic Equations

James L. Juno (University of Maryland)

While fluid models are common tools in the study of plasmas, many of these systems, whether in astrophysics or the lab, are only weakly collisional and far from equilibrium, making them more accurately described by kinetic equations. Kinetic equations can be computationally demanding due to the need to solve for the distribution function of the particles in a higher dimensional phase space, with position and velocity coordinates. Despite this challenge, the motivation for solving the plasma kinetic equation is large as there remains a vast array of questions concerning collisionless dynamics in real plasma systems. Here we present algorithms in an Eulerian framework for the discretization of the plasma kinetic equation, using a high-order discontinuous Galerkin finite element method due to its arithmetic intensity and parallelizability. Scaling and performance of the algorithm are discussed, and benchmarks of the algorithm are presented as well.

Awards Presentation

(back to top)

Tuesday, November 13th

Room: Exhibit Hall B
1:30 pm - 3:00 pm

Test of Time Award Presentation

The Parallel Hashed Oct-Tree Algorithm Revisited
Mike Warren (Descartes Labs Inc), John Salmon (D.E. Shaw Research)

Algorithms such as fast N-body methods, coupled with the extraordinary increase in computational power over the past three decades, have played a major part in the process of understanding complex physical systems. In this talk, we provide some context and history of gravitational N-body simulations and discuss the evolution of the parallel hashed oct-tree N-body algorithm and the problems it has been applied to over the years since its publication. We touch on concurrent evolution in hardware, such as Beowulf clusters and the Cloud as well as physical applications in cosmology, galaxy formation, supernovae, hydrodynamics and planet formation.

Wednesday, November 14th

Room: Exhibit Hall B
8:30 am - 10:00 am

ACM and IEEE-CS Award Presentations

Fernbach Award Presentation - The Roles of Computing in the Advancement of Science: A Case Study *Linda Petzold (University of California, Santa Barbara)*
Stochasticity plays an important role in many biological processes. At the same time, algorithms and software for discrete stochastic simulation have advanced to the point where not only simulation of well-mixed systems, but spatial stochastic simulation on 3D complex geometries can be performed with accuracy and reliability. A few years ago we embarked on a quest to build a unified software environment, StochSS, to enable biologists to easily harness the power of these tools. We envisioned that users might build an ODE model or discrete stochastic model on a laptop, and scale it up to increasing levels of complexity, accessing tools such as those mentioned above, and deploying computing resources from the cloud with the push of a button when they are needed. As the capabilities of StochSS have grown, so has our vision of the roles that computing can play in the advancement of science.

Kennedy Award Presentation - Memory Consistency Models: They Are Broken and Why We Should Care *Sarita Adve (University of Illinois)*
The memory consistency model for a shared address space specifies the value a load can return,

affecting programmability and performance. For such a fundamental property, it, unfortunately, still routinely causes heads to spin. I will first briefly trace the evolution of memory models over three decades. The 1990s saw an explosion in memory models from hardware vendors and researchers. The 2000s drove a convergence centered on the more software-centric view of the data-race-free model. The last decade has struggled with mind-twisting implications of "out-of-thin-air" values and relaxed atomics, pointing to a fundamentally broken paradigm for hardware and software.

The end of Moore's law is driving transformational change in hardware with specialization and heterogeneity within and across chips, including application-specialized and heterogeneous parallelism, coherence, and communication. How does this affect the memory model, or more broadly, the hardware-software interface? From our recent research in the DeNovo project, I will show examples of how we are again in danger of repeating the mistakes of the hardware-centric 1990s to create another memory model mess. On the other hand, there is a golden opportunity for hardware-software cooperation to redefine our interface from the ground up and find a fundamental resolution to the problem. I believe this will require rethinking how we represent parallelism, communication, and correctness in software, how we provide coherence and communication in hardware, and that the HPC community's expertise in how to explicitly manage communication will have a key role to play.

Room: A2 Ballroom
3:30 pm - 5:00 pm

Gordon Bell Prize Finalist Session 1

A Fast Scalable Implicit Solver for Nonlinear Time-Evolution Earthquake City Problem on Low-Ordered Unstructured Finite Elements with Artificial Intelligence and Transprecision Computing
Tsuyoshi Ichimura (University of Tokyo)

To address problems that occur due to earthquakes in urban areas, we propose a method that utilizes artificial intelligence (AI) and transprecision computing to accelerate a nonlinear dynamic low-order unstructured finite-element solver. The AI is used to improve the convergence of iterative solver leading to 5.56-fold reduction in arithmetic count from a standard solver, and FP16-FP21-FP32-FP64 computing is used to accelerate the sparse matrix-vector product kernel, which demonstrated 71.4% peak FP64 performance on Summit. This is 25.3 times faster than a standard solver and 3.99 times faster than the state-of-the-art SC14 Gordon Bell Finalist solver. Furthermore, the proposed solver demonstrated high scalability (88.8% on the K computer and 89.5% on Piz Daint), leading to 14.7% peak FP64 performance on 4096 nodes of Summit. The proposed approach utilizing AI and FP16 arithmetic has implications for accelerating other implicit solvers used for earthquake city simulations as well as various fields.

167-PFlops Deep Learning for Electron Microscopy: From Learning Physics to Atomic Manipulation
J. Travis Johnston (Oak Ridge National Laboratory)

An artificial intelligence system called MENNDL, which used 25,200 Nvidia Volta GPUs on Oak Ridge National Laboratory's Summit machine, automatically designed an optimal deep learning network in

order to extract structural information from raw atomic-resolution microscopy data. In a few hours, MENNDL creates and evaluates millions of networks using a scalable, parallel, asynchronous genetic algorithm augmented with a support vector machine to automatically find a superior deep learning network topology and hyper-parameter set than a human expert can find in months. For the application of electron microscopy, the system furthers the goal of improving our understanding of the electron-beam-matter interactions and real-time image-based feedback, which enables a huge step beyond human capacity toward nanofabricating materials automatically. MENNDL has been scaled to the 4,200 available nodes of Summit achieving a measured 152.5 PFlops, with an estimated sustained performance of 167 PFlops when the entire machine is available.

Exascale Deep Learning for Climate Analytics
Thorsten Kurth (Lawrence Berkeley National Laboratory), Sean Treichler (Nvidia Corporation)

We extract pixel-level masks of extreme weather patterns using variants of Tiramisu and DeepLabv3+ neural networks. We describe improvements to the software frameworks, input pipeline, and the network training algorithms necessary to efficiently scale deep learning on the Piz Daint and Summit systems. The Tiramisu network scales to 5300 P100 GPUs with a sustained throughput of 21.0 PF/s and parallel efficiency of 79.0%. DeepLabv3+ scales up to 27360 V100 GPUs with a sustained throughput of 325.8 PF/s and a parallel efficiency of 90.7% in single precision. By taking advantage of the FP16 Tensor Cores, a half-precision version of the DeepLabv3+ network achieves a peak and sustained throughput of 1.13 EF/s and 999.0 PF/s respectively.

Thursday, November 15th

Room: A2 Ballroom
10:30 am - 12:00 pm

Gordon Bell Prize Finalist Session 2

Simulating the Weak Death of the Neutron in a Femtoscale Universe with Near-Exascale Computing
Ken McElvain (University of California, Berkeley; Lawrence Berkeley National Laboratory)

The fundamental particle theory called Quantum Chromodynamics (QCD) dictates everything about protons and neutrons, from their intrinsic properties to interactions that bind them into atomic nuclei. Quantities that cannot be fully resolved through experiment, such as the neutron lifetime (whose precise value is important for the existence of light-atomic elements that make the sun shine and life possible), may be understood through numerical solutions to QCD. We directly solve QCD using Lattice Gauge Theory and calculate nuclear observables such as neutron lifetime. We have developed an improved algorithm that exponentially decreases the time-to-solution and applied it on the new CORAL supercomputers, Sierra and Summit. We use run-time autotuning to distribute GPU resources, achieving 20% performance at low node count. We also developed optimal application mapping through a job manager, which allows CPU and GPU jobs to be interleaved, yielding 15% of peak performance when deployed across large fractions of CORAL.

ShenTu: Processing Multi-Trillion Edge Graphs on Millions of Cores in Seconds
Heng Lin (Tsinghua University, Fma Technology), Xiaowei Zhu (Tsinghua University, Qatar Computing Research Institute), Bowen Yu (Tsinghua University), Xiongchao Tang (Tsinghua University, Qatar Computing Research Institute), Wei Xue (Tsinghua University), Wenguang Chen (Tsinghua University), Lufei Zhang (State Key Laboratory of Mathematical Engineering and Advanced Computing), Torsten Hoefler (ETH Zurich), Xiaosong Ma (Qatar Computing Research Institute)

Graphs are an important abstraction used in many scientific fields. With the magnitude of graph-structured data constantly increasing, effective data analytics requires efficient and scalable graph processing systems. Although HPC systems have long been used for scientific computing, people have only recently started to assess their potential for graph processing, a workload with inherent load imbalance, lack of locality, and access irregularity. We propose ShenTu, the first general-purpose graph processing framework that can efficiently utilize an entire petascale system to process multi-trillion edge graphs in seconds. ShenTu embodies four key innovations: hardware specializing, supernode routing, on-chip sorting, and degree-aware messaging, which together enable its unprecedented performance and scalability. It can traverse an unprecedented 70-trillion-edge graph in seconds. Furthermore, ShenTu enables the processing of a spam detection problem on a 12-trillion edge Internet graph, making it possible to identify trustworthy and spam web pages directly at the fine-grained page level.

Attacking the Opioid Epidemic: Determining the Epistatic and Pleiotropic Genetic Architectures for Chronic Pain and Opioid Addiction
Wayne Joubert (Oak Ridge National Laboratory), Daniel Jacobson (Oak Ridge National Laboratory)

We describe the CoMet application for large-scale epistatic Genome-Wide Association Studies (eGWAS) and pleiotropy studies. High performance is attained by transforming the underlying vector comparison methods into highly performant generalized distributed dense linear algebra operations. The 2-way and 3-way Proportional Similarity metric and Custom Correlation Coefficient are implemented using native or adapted GEMM kernels optimized for GPU architectures. By aggressive overlapping of communications, transfers and computations, high efficiency with respect to single GPU kernel performance is maintained up to the full Titan and Summit systems. Nearly 300 quadrillion element comparisons per second and over 2.3 mixed precision ExaOps are reached on Summit by use of Tensor Core hardware on the Nvidia Volta GPUs. Performance is four to five orders of magnitude beyond comparable state of the art. CoMet is currently being used in projects ranging from bioenergy to clinical genomics, including for the genetics of chronic pain and opioid addiction.

Room: Exhibit Hall B
12:45 pm - 1:30 pm

Awards Ceremony

SC18 Awards Session
The SC18 conference awards, as well as selected ACM and IEEE awards, will be presented. The awards include: Best Paper, Best Student Paper, Best Poster, Best Scientific Visualization, ACM Student Research Competition, ACM Gordon Bell Prize, ACM/IEEE-CS George Michael Memorial HPC

Fellowship, ACM SIGHPC / Intel Computational & Data Science Fellowships, ACM SIGHPC Emerging Woman Leader in Technical Computing Award, IEEE TCHPC Award for Excellence for Early Career Researchers in High Performance Computing, and Student Cluster Competition.

Birds of a Feather

(back to top)

Tuesday, November 13th

Room: C140/142
12:15 pm - 1:15 pm

LUSTRE Community BOF: Lustre in HPC and Emerging Data Markets: Roadmap, Features and Challenges

Sarp Oral (OpenSFS Inc, Oak Ridge National Laboratory)

Lustre is the leading open source file system for HPC. Over 70% of the top 100 supercomputers use Lustre. It is a community developed technology with contributors from around the world. Lustre currently supports many HPC infrastructures beyond scientific research, sch as financial services, oil and gas and advanced manufacturing. Lustre clients are available for variety of technologies, such as x86, POWER, and ARM. At this BOF, Lustre developers, administrators, and solution providers will gather to discuss recent Lustre developments, challenges, and corresponding opportunities for the next 5 years of Lustre, including the role of Lustre in future exascale environments.

Room: C145
12:15 pm - 1:15 pm

Energy Efficiency Considerations for HPC Procurements

Jason Hick (Los Alamos National Laboratory)

The goal for procurement of HPC systems is to identify optimal solutions to both technical and financial targets that maximize the contribution of the system to the organization's mission. It is important to consider total cost of ownership, cooling and power requirements, the needs for interfacing with the HPC facility, and power management and control. In this BoF, we present energy efficiency highlights in recent procurements by four sites and encourage audience participation. Presenting sites are from Japan on power efficiency; Europe on cooling enhancements; the United States on future efficiency requirements as well as on facility integration and efficiency.

Room: C146
12:15 pm - 1:15 pm

Open MPI State of the Union 2018

Jeffrey Squyres (Cisco Systems)

Open MPI continues to drive the start of the art in HPC. This year, we've added new features, fixed bugs, improved performance, and collaborated with many across the HPC community. We'll discuss what Open MPI has accomplished over the past year and present a roadmap for the next year.

One of Open MPI's strength lies in its diversity: we represent many different viewpoints across the HPC ecosystem. To that end, many developers from the community will be present to discuss and answer your questions both during and after the BOF.

Room: C147/148/154
12:15 pm - 1:15 pm

Meeting HPC Container Challenges as a Community

CJ Newburn (Nvidia Corporation)

Containers increase the likelihood users can run the latest version of HPC applications in their data center, because it's easier for administrators to install them and developers to deploy them. But we need to build and leverage an active community to solve remaining challenges.

The BoF presents, prioritizes, and gathers further input on top issues and budding solutions around containerization of HPC applications. Lightning talks nominated in ongoing community discussions describe the community's top issues, frame challenges, highlight progress and plans. Led discussions expand on the seeded issues, enumerate additional problems, and draw the audience into our ongoing community-wide forum.

Room: C155/156
12:15 pm - 1:15 pm

SIGHPC Annual Member Meeting

SIGHPC Annual Member Meeting
The annual business meeting of SIGHPC is your opportunity to hear about and discuss the status of SIGHPC and its chapters. We will also be discussing upcoming plans for the year. All of the elected officers and many of the other volunteers will be present to answer your questions about SIGHPC. Representatives from our chapters: Education, Big Data, Resource Constrained Environments (RCE), and System Administration will also be available. Come tell us how SIGHPC can meet your needs.

Room: D163
12:15 pm - 1:15 pm

SAGE2 10th Annual International SC BOF: Scalable Amplified Group Environment for Global Collaboration

Jason Leigh (University of Hawaii at Manoa)

SAGE2 (Scalable Amplified Group Environment) is an open-source, web-based, user-centered platform for local and/or distributed small or large groups of researchers, educators and students to communicate information and ideas as if they were in the same room. Think of it as WebEx or Skype for cyber-enabled, data-intensive science, engineering and education. SAGE2, like its predecessor SAGE (Scalable Adaptive Graphics Environment), is the defacto operating system for managing Big Data on tiled walls, providing the scientific community with persistent visualization and collaboration services for global cyberinfrastructure. This year's BoF will focus on eliciting community input on use cases and cybersecurity enhancements.

Room: D165
12:15 pm - 1:15 pm

Impacting Cancer with HPC: Opportunities and Challenges

Eric Stahlberg (Frederick National Laboratory for Cancer Research, Leidos Biomedical Research)

High-Performance Computing is experiencing a dramatic increase in interest as new frontiers in deep learning are explored, volumes and variety of data are exploding, cloud-based HPC provides ready-access and extreme scale computing leads to new questions and approaches. Changes are occurring globally, providing new avenues for collaborations, identifying new insights to benefit all. Key topics including the role of machine learning, access to quality data, sharing and accessing predictive models, and insights from the patient perspectives. Following the successful format of past years, the BoF will be informative and highly-interactive with presentations to inspire and encourage significant discussion among attendees.

Room: D166
12:15 pm - 1:15 pm

Unified Communication X (UCX) Community

Gilad Shainer (Mellanox Technologies)

In order to exploit the capabilities of HPC systems, communication software needs to scale on millions of cores and support applications with adequate functionality to express their parallelism. UCX is a collaboration between industry, labs and academia that consolidates multiple technologies to provide a unified open-source framework. The UCX project is managed by the UCF consortium (http://www.ucfconsortium.org) and includes members from LANL, ORNL, ANL, Ohio State University, AMD, ARM, IBM, Mellanox, NVIDIA and more. The session will serve as the UCX community meeting and will introduce the latest development and specification to HPC developers and the broader user community.

Room: D167
12:15 pm - 1:15 pm

HPCG Benchmark Update

Michael Heroux (Sandia National Laboratories, St. John's University)

The High Performance Conjugate Gradients (HPCG) Benchmark is a community metric for ranking high performance computing systems, officially part of the TOP500 and a companion to the LINPACK benchmark.

In this BOF we first present an update of HPCG policies and opportunities for optimizing performance and follow with presentations from vendors who have participated in HPCG optimization efforts, in particular efforts for ARM-based systems, Nvidia, Intel and IBM. We spend the remaining time in open discussion about the future of HPCG design and implementation strategies for further improvements.

Room: D174
12:15 pm - 1:15 pm

Pros and Cons of HPCx benchmarks

Jun Makino (Kobe University, RIKEN)

HPL has been and still is the single most widely used benchmark for the HPC systems, even though there have been many criticisms. The most important criticism is that HPL measures only the peak floating point performance and its result has little correlation with real application performance. HPCG (and also HPGMG) have been proposed as either alternative or complimentary benchmarks. HPCG measures mainly the main memory bandwidth. In this BoF, we want to exchange opinions on what aspects of machines should be measured and how by these benchmarks, in particular, when used as the requirement for new machines.

Room: D168
12:15 pm - 1:15 pm

What the Heck Is HEC?

Sandy Landsberg (US Department of Defense HPC Modernization Program)

Federal High-End Computing (HEC) is evolving in response to the needs of the HEC community and advancing technological landscape, including rapidly changing architecture, anticipated end to Moore's Law, facing the challenges associated with the data tsunami, and scaling software to newer and more complex platforms. Consequently, the HEC environment is more complex than ever and these factors stress the HEC community in every way. This BoF will provide an overview of the current and emerging federal HEC efforts and seeks to have a candid discussion with the HEC community on how it can be collaboratively cultivated to serve tomorrow's challenges.

Room: D171
12:15 pm - 1:15 pm

LLVM in HPC: What's New?

Jim Cownie (Intel Corporation)

LLVM is the prime environment for developing new compilers and language-processing tools. In this BoF, a group of LLVM experts who are driving the development of LLVM will give short presentations of their work, then answer your questions on LLVM and its future in HPC.

As well as Hal Finkel, we expect our experts to include Doru Bercea (IBM/OpenMP). Chandler Carruth (clang/Google), and Keno Fischer (Julia/Julia).

Our goal is to connect you with LLVM experts so that you understand some of the uses of LLVM, and they understand what tools and LLVM enhancements you want.

Room: D173
12:15 pm - 1:15 pm

High Level Programming Languages for Quantum Computation

Torsten Hoefler (ETH Zurich)

Several independent groups have demonstrated small-scale quantum computers containing several dozens of qubits with a limited programmability. Today's machines are programmed at the bit level with application of single quantum gates. New developments of high-level languages such as Quipper, Q#, OpenQASM promise to support the orchestration of tens-of-thousands of qubits at an abstraction level similar to modern classical programming. Yet, quantum algorithms are fundamentally different and require complex schemes such as gate synthesis and quantum error correction. We propose to bring the community of industry and academic researchers and users together to discuss recent results and design a way forward.

Room: D175
12:15 pm - 1:15 pm

17th Graph500 List

Richard Murphy (Micron Technology Inc)

Data intensive supercomputer applications are increasingly important workloads, especially for “Big Data” problems, but are ill suited for most of today’s computing platforms (at any scale!). The Graph500 list has grown to over 239 entries and has demonstrated the challenges of even simple

analytics. The new SSSP kernel introduced at SC17 has increased the benchmark's overall difficulty. This BOF will unveil the Graph500 lists, discuss the new streaming kernel, and enhance the new energy metrics the Green Graph500. It will offer a forum for community and provide a rallying point for data intensive supercomputing problems.

Room: D220
12:15 pm - 1:15 pm

Best Practices for Scaling-Up and Sustaining HPC Education, Outreach and Training

Nitin Sukhija (Slippery Rock University of Pennsylvania)

The HPC community is delivering HPC educational content and training to address workforce preparation at all levels - from undergraduate students to career professionals. Flexible and scalable pedagogical and andragogical approaches are needed to address the broad workforce needs from the basics of computational thinking to extreme scale performance. Through brief presentations followed by open discussions, this ACM SIGHPC Education Chapter coordinated BOF will engage the audience in discussions of the challenges and solutions for scaling-up and sustaining successful HPC teaching and learning. The activities include formal courses, MOOCs, extracurricular activities, webcast training, and Carpentries models of teaching and learning.

Room: D227
12:15 pm - 1:15 pm

Spack Community BoF

Todd Gamblin (Lawrence Livermore National Laboratory)

Spack is a package manager for scientific software, with a rapidly growing open source community. Spack has over 300 contributors from academia, industry, and laboratories across the world. It is used to manage software releases for the U.S. Exascale Computing Project and the Summit supercomputer. At this BoF, Spack developers will give updates on the community and on new features such as environments and binary distribution. Sites using Spack will describe their experiences, and we will open the floor for discussion. All are invited to provide feedback, request features, and discuss future directions. Help us make installing HPC software simple!

Room: C140/142
5:15 pm - 6:45 pm

Containers in HPC

Andrew Younge (Sandia National Laboratories)

Container computing has revolutionized how many industries and enterprises develop and deploy software and services. Recently, this model has gained traction in the HPC community through enabling technologies like Charliecloud, Docker, Kubernetes, Shifter, and Singularity. While containers look to provide greater software flexibility, reliability, ease of deployment, and portability for users, there are still open questions that need to be addressed as we race toward adoption in the exascale era. In this BoF, we will provide an opportunity for the HPC community to engage with the leaders in the field who can provide real-world experiences to containerization in HPC.

Room: C141/143/149
5:15 pm - 6:45 pm

Enabling Data Services for HPC

Jerome Soumagne (HDF Group)

Distributed data services can enhance HPC productivity by providing storage, analysis, and visualization capabilities not otherwise present in conventional parallel file systems. Such services are difficult to develop, maintain, and deploy in a scientific workflow, however, due to the complexities of specialized HPC networks, RDMA data transfers, protocol encoding, and fault tolerance.

This BoF will bring together a growing community of researchers, developers, vendors, and facility operators who are either using or developing HPC data services. Participants will share and discuss practical experiences, implementation examples, new features, and best practices to construct and deploy production-quality, high-performance distributed services.

Room: C144
5:15 pm - 6:45 pm

The Facility Perspective on Liquid Cooling: Experiences and Proposed Open Specification

Dale Sartor (Lawrence Berkeley National Laboratory)

As compute densities increase, there is growing demand to more effectively cool power-dense equipment and improve energy efficiency with compressor-less cooling. This BoF will explore the steps necessary to take advantage of warm liquid-cooling in the data-center and introduce an open-specification for a secondary fluid warm liquid-cooled rack.. Lawrence Berkeley National Laboratory and China Institute of Electronics steer this initiative and seek input from the HPC community. This BoF will feature a panel of seasoned operations managers from major supercomputing centers to talk about strategies for effectively enabling warm-water cooling, including a discussion on the need for industry standards.

Room: C145

5:15 pm - 6:45 pm

HPC in the DoD

Laura Ulibarri (Maui High Performance Computing Center)

The DoD has invested significant time and funding to support a large base of users on a variety of HPC-backed projects. This BoF will use lightning talks about current research, plans, and thrusts to illustrate DoD goals and opportunities for engagement. These lightning talks are intended to help external organizations and researchers connect with DoD users and sites to encourage partnerships and help solve problems. External engagement will help DoD users and sites grow expertise and connect to the larger HPC community.

Room: C146
5:15 pm - 6:45 pm

Purpose-Built HPC: Last Hope for Earth System Prediction?

Dave McCarren (US Navy, Oceanographer of the Navy)

As earth system prediction models continue to increase in complexity, so do the high-performance computing (HPC) requirements for running these models. Unfortunately, current HPC platforms are designed, built and optimized for a broad array of computationally intensive applications which results in them being inefficient for earth system prediction applications. Specifically, recent top HPC platforms are mostly optimized for machine learning, artificial intelligence, and n-body simulations. They are not designed with the specific purpose of solving partial differential equations (PDEs), such as the Navier-Stokes equations, which are the foundation of earth system modeling.

Room: C147/148/154
5:15 pm - 6:45 pm

Multi-Level Memory and Storage for HPC and Data Analytics

Hans-Christian Hoppe (Intel Corporation)

Recent progress in storage-class memory (SCM) technologies combined with emerging workloads mixing HPC simulations with big data processing or machine learning do create a perfect opportunity for innovative HPC systems embracing SCM. These will substantially improve delivered performance, scalability and energy efficiency for data-oriented HPC codes as well as “mixed” applications, and can play a large role in scaling the “I/O wall” when moving to exascale-class systems.

This BoF brings together technology providers, application and system SW developers, and system operators to discuss the evolving SCM landscape in the context of specific use cases, emerging technologies, and actual success stories.

Room: C155/156
5:15 pm - 6:45 pm

Analyzing Parallel I/O

Philip Carns (Argonne National Laboratory)

Parallel application I/O performance often fails to meet user expectations. In addition, subtle changes in access patterns may lead to significant changes in performance due to complex interactions between hardware and software. These challenges call for sophisticated tools to capture, analyze, understand, and tune application I/O.

In this BoF, we will highlight recent advances in monitoring and characterization tools to help address this problem. We will also encourage community discussion to compare best practices, identify gaps in measurement and analysis, and find ways to translate parallel I/O analysis into actionable outcomes for users, facility operators, and researchers.

Room: D163
5:15 pm - 6:45 pm

Strategies for Inclusive and Scalable HPC Outreach and Education

Julie Mullen (Massachusetts Institute of Technology)

The HPC community has consistently identified public outreach and education of new and existing community members as vital to the growth of the field. New strategies and alternative teaching methods are needed to improve access, promote diversity, attract new HPC practitioners, and encourage the next generation of scientists and gain the support of the general public. To share effective strategies, practices, and spawn new ideas to meet the challenge of providing accessible, inclusive and scalable HPC outreach, education and training, we'll use a combination of discussion and demonstrations of existing training and outreach activities. Come and try them for yourself!

Room: D166
5:15 pm - 6:45 pm

HPC Graph Toolkits and the GraphBLAS Forum

José Moreira (IBM)

Government agencies, industry and academia are demanding a new generation of tools to efficiently solve large scale analytics problems in a variety of business, scientific, and national security applications. This BoF aims at gathering the community of people interested in frameworks and workflows for large scale graph analytics, surveying the current approaches, identifying new challenges and opportunities, and laying a path toward future interoperable infrastructures. As in

previous editions, we will invite the GraphBLAS community to participate in a discussion of the current state and evolution of GraphBLAS, with the goal of developing requirements and recommendations for future tools.

Room: D174
5:15 pm - 6:45 pm

Ceph Applications in HPC Environments

Benjamin Lynch (University of Minnesota)

Ceph is an open-source distributed object store with an associated file system widely used in cloud and distributed computing. In addition, both the object store and file system components are seeing increasing deployments as primary data storage for traditional HPC. Ceph is backed by a robust, worldwide open source community effort with broad participation from major HPC and storage vendors. This BOF session will bring together Ceph implementers to share their deployment experiences, as well as provide feedback to the developer community on needed features and enhancements specific to the HPC community.

Room: D168
5:15 pm - 6:45 pm

Distributed and Heterogeneous Programming in C++ for HPC 2018

Hal Finkel (Argonne National Laboratory)

After last year's successful Heterogeneous and Distributed Computing in C++ for HPC BoF, there was popular demand for continuing updates on the progress of adding these capabilities into ISO C++. We are close to finalizing C++20, and this BoF will provide updates from active participants in the standardization process on what is possible for C++20 and for C++23.

There are a number of C++ frameworks for parallel programming, including HPX, KoKKos, Raja, C++AMP, HCC, Boost.Compute, CUDA, and more. SYCL, from Khronos, provides heterogeneous computing built on OpenCL and C++, and Codeplay has released ComputeCpp Community Edition.

Room: D171
5:15 pm - 6:45 pm

OpenACC API User Experience, Vendor Reaction, Relevance, and Roadmap

Duncan Poole (Nvidia Corporation)

OpenACC, a directive-based high-level parallel programming model, has gained rapid momentum among scientific application users - the key drivers of the specification. The user-friendly programming

model has facilitated acceleration of over 130 applications including CAM, ANSYS Fluent, Gaussian, VASP, Synopsys on multiple platforms and is also seen as an entry-level programming model for the top supercomputers (Top500 list) such as Summit, Sunway Taihulight, and Piz Daint. As in previous years, this BoF invites scientists, programmers, and researchers to discuss their experiences in adopting OpenACC for scientific applications, learn about the roadmaps from implementers and the latest developments in the specification.

Room: D173
5:15 pm - 6:45 pm

Interactivity in HPC

Peter Messmer (Nvidia Corporation)

Interactive exploration of data, preparation and debugging of large-scale scientific simulations, in-situ visualization, and application steering are compelling scenarios for exploratory science, design optimizations, and signal processing. However, various technical, organizational and sociological challenges must be overcome to make these interactive workflows mainstream in HPC centers: What problem domains and workflows can benefit most from interactivity? How can we simplify the tool-chain? How can centers support highly interactive workflows? The goal of this BoF is bring together domain scientists, tool developers, and HPC center administrators to identify the scientific impact and technical challenges of highly interactive access to HPC resources.

Room: D175
5:15 pm - 6:45 pm

The New NSF-Funded Resource: Frontera - Towards a Leadership Class Computing Facility

Dan Stanzione (Texas Advanced Computing Center, University of Texas)

On July 24th 2018, the National Science Foundation announced the intent to negotiate an award in it's "Towards a Leadership Class Computing Facility" competition for a replacement system for the Blue Waters Supercomputer. This award will be made to the Texas Advanced Computing Center. While details of this award are still embargoed at submission time, they will be made public by SC18. This BoF will be to brief the community on this new resource, and encourage participation.

Room: D220
5:15 pm - 6:45 pm

Advanced Architecture Testbeds: A Catalyst for Co-design Collaborations

Kevin Barker (Pacific Northwest National Laboratory)

Our supercomputing community is in a renaissance of computer and system architectures. The explosion in architectural diversity is supported by the use of testbeds to analyze application performance and energy efficiency on real system software stacks. Reasons for active participation in this BoF session include alignment of efforts, increased coverage of diverse architectures, sharing of lessons learned with advanced architectures, support for establishing and useing common proxy application and benchmarks. This BoF will bring together advanced architecture testbed efforts: at CENATE – PNNL, HAAPS – SNL, Rogues Gallery – GA-Tech, HPCMP – testbeds at several sites, and ExCL at ORNL.

Room: D227
5:15 pm - 6:45 pm

Designing and Building Next-Generation Computer Systems for Deep Learning

Volodymyr Kindratenko (University of Illinois, National Center for Supercomputing Applications)

Deep Learning (DL) heavily relies on fast hardware and parallel algorithms to train complex neural networks. his BoF will bring together researchers and developers working on the design of next-g eneration computer systems for DL and parallel DL algorithms that can exploit the potential of these new systems. Research teams working on major deep learning systems deployed in the field are invited to discuss latest hardware and software trends and to exchange views on the role Artificial Intelligence (AI) in general and DL in particular will play in the near future for big data analytics and HPC applications.

Room: Exhibit Hall B
5:15 pm - 7:00 pm

TOP500 Supercomputers

Erich Strohmaier (Lawrence Berkeley National Laboratory)

The TOP500 list of supercomputers serves as a “Who’s Who” in the field of High Performance Computing (HPC). It started as a list of the most powerful supercomputers in the world and has evolved to a major source of information about trends in HPC. The 52nd TOP500 list will be published in November 2018 just in time for SC18.

This BoF will present detailed analyses of the TOP500 and discuss the changes in the HPC marketplace during the past years. The BoF is meant as an open forum for discussion and feedback between the TOP500 authors and the user community.

Wednesday, November 14th

Room: C140/142
12:15 pm - 1:15 pm

OpenHPC Community BoF

Karl Schulz (University of Texas)

Over the last several years, OpenHPC has emerged as a community-driven stack providing a variety of common, pre-built ingredients to deploy and manage an HPC Linux cluster. Formed initially in November 2015 and formalized as a Linux Foundation project in June 2016, OpenHPC has been adding new software components and now supports multiple OSes/architectures. At this BoF, speakers from the OpenHPC Technical Steering Committee will provide technical updates from the project and near-term roadmaps. We then invite open discussion giving attendees an opportunity to provide feedback on current conventions, packaging, request additional components and configurations, and discuss general future trends.

Room: C141/143/149
12:15 pm - 1:15 pm

The Message Passing Interface (MPI): Version 4.0 and Beyond

Martin Schulz (Technical University Munich)

The Message Passing Interface (MPI) API is the most dominant programming approach for HPC environments. Its specification is driven by the MPI forum, an open forum consisting of MPI developers, vendors and users. This BoF Meeting will provide insights into the current topics discussed in the forum, plans and timelines for upcoming versions of MPI, as well as the process of how features are added to the standard. It is intended to keep the larger HPC community informed about current activities and long-term directions, as well as encourage larger community participation in this crucial standard for the supercomputing community.

Room: C144
12:15 pm - 1:15 pm

A Look Ahead: Energy and Power Aware Job Scheduling and Resource Management

Kevin Pedretti (Sandia National Laboratories)

Energy and power aware job scheduling and resource management (EPAJSRM) capabilities are implemented or planned for large-scale HPC systems in ~10 sites worldwide. Some of the sites are interested in using these capabilities to allow an application to provide hints and other relevant information to an EPAJSRM job scheduler. Another important capability is to notify applications of power management decisions, such as changes in power usage targets and providing awareness of what is going on in the machine that might have made a job run slower. This BoF explores the these capabilities from the perspective of three different sites.

Room: C145
12:15 pm - 1:15 pm

Women in HPC: the Importance of Male Allies

Rebecca Hartman-Baker (National Energy Research Scientific Computing Center (NERSC))

In a male-dominated field such as HPC, male allies are particularly vital to ensure that women advance equitably in their careers. Research shows that men who work to support, mentor, and sponsor their women colleagues are viewed favorably, while women advocating for equity are viewed unfavorably. Organizations with effective male allies exhibit improved career satisfaction and better retention of women. In this BoF, panelists will explore how male allies have made a difference in their careers and contrast their allyship with examples of exclusionary experiences.

Room: C146
12:15 pm - 1:15 pm

HDF5: I/O Middleware and Ecosystem for HPC and Experimental and Observational Sciences

Elena Pourmal (HDF Group, Lawrence Berkeley National Laboratory)

We will provide a forum for the HDF5 user community to share ideas and discuss initiatives in the areas of HPC and Experimental and Observational Sciences. Elena Pourmal will present HDF5 features under development, the HDF5 roadmap, including upcoming releases and solicit input on the future roadmap. Quincey Koziol will moderate a panel with representatives from research, commercial, and government organizations who will present case studies on how they leverage HDF technologies in the fields of Experimental and Observational Sciences to solve big data problems as well as discuss the challenges of using HDF5 in the HPC environment.

Room: C147/148/154
12:15 pm - 1:15 pm

Big Data Challenge - How to Engage with Large Scale Facilities?

John Towns (National Center for Supercomputing Applications)

In recent years there has been a rapid growth in the number of large scale facilities (LSF) – that produce large amounts of data. The classical examples are the Large Hadron Collider and SKA. In the context of a set of concrete examples, this BoF will discuss both current and projected requirements for processing, storing, sharing, and securing of data as well as data quantities, processing needs, and expected data transfer needs with the goal of collecting an aggregate set of requirements from the LSFs allowing for gap analysis of these requirements against the roadmaps of international HPC data

centers.

Room: C155/156
12:15 pm - 1:15 pm

Batched, Reproducible, and Reduced Precision BLAS

Piotr Luszczek (University of Tennessee, Innovative Computing Laboratory)

This BoF will bring together the community focused on extending the Basic Linear Algebra Software (BLAS). The existing BLAS proved to be effective in assisting portable, efficient software for sequential and the current class of high-performance computers. We'd like to investigate the possibility of extending the currently accepted standards to provide greater parallelism for small size operations, reproducibility, and reduced precision support. This is an open forum to discuss and formalize details. The agenda and talks from past workshops can be found here:

http://bit.ly/Batch-BLAS-2017

http://bit.ly/Batch-BLAS-2016

http://icl.utk.edu/bblas/

A standard interface will be considered for the Batched, Reproducible, and Reduced Precision BLAS.

Room: D166
12:15 pm - 1:15 pm

InfiniBand In-Network Computing Technology and Roadmap

Gilad Shainer (Mellanox Technologies)

The latest revolution in HPC is the effort around the co-design approach, a collaborative effort to reach exascale performance by taking a holistic system-level approach to fundamental performance improvements, is In-Network Computing. The CPU-centric approach has reached the limits of its scalability in several aspects, and In-Network Computing acting as "distributed co-processor" can handle and accelerates performance of various data algorithms, such as reductions and more. The session will cover the latest development of the InfiniBand roadmap and In-Network Computing technologies, and will include discussions and presentations from the first HDR 200G InfiniBand sites.

Room: D167
12:15 pm - 1:15 pm

PBS Pro Open Source Project Community BoF

Bill Nitzberg (Altair Engineering)

PBS Pro schedules jobs and manages workloads for HPC clusters and clouds. The software was born at NASA in the 1990s, and, by the 2010s, became one of the top tools of its kind. In mid-2016, PBS Pro became a fully open source solution, with a growing community of developers and users around the globe. Join the community (www.pbspro.org) -- users, developers, contributors, and open source enthusiasts -- to learn what's new and to drive the future of PBS Pro.

Room: D174
12:15 pm - 1:15 pm

HPC in Space: An Update on Spaceborne Computer after 1+ Year on the ISS

Mark Fernandez (Hewlett Packard Enterprise)

Spaceborne Computer, a stock, Linux-based, commercial off-the-shelf (COTS) HPC system was launched aboard the SpaceX-12 Dragon spaceship on August 14, 2017. After a month in cold stowage, it was installed and powered up in the International Space Station (ISS) on September 14, 2017. Despite many predictions, Spaceborne Computer continues to run in the harsh environment of space without expensive, time-consuming, or bulky protective shielding. The use of "HARDening with SOFTware" will be shared; the "anomalies" (both expected and unexpected) that were encountered in space will be covered; and the experiment results to date will be shared.

Room: D168
12:15 pm - 1:15 pm

Special Interest Group on HPC in Resource Constrained Environments (SIGHPC-RCE)

Hensley Omorodion (University of Benin)

This SC18 BoF will be the fourth meeting of the Special Interest Group on HPC in Resource Constrained Environments (SIGHPC-RCE). This SIGHPC chapter was chartered in 2015, and its first BoF was held during SC15. SIGHPC-RCE membership is open to professional staff who manage resources in research institutions and college campuses (systems administrators, network engineers, data scientists, etc). Topics for discussion include: open-source software solutions, cloud-enabled efficiencies, cybersecurity challenges, network topology and NREN peering, online training resources, administrative challenges of shared cyber-ecosystems, outreach, and creative ways that decommissioned hardware can be employed for light research and workforce development.

Room: D169
12:15 pm - 1:15 pm

Third Annual Meeting of the SIGHPC - Big Data Chapter

Stratos Efstathiadis (New York University)

The goal of the BoF is to gather, for the third time, members and non-members of the SIGHPC-BigData Virtual Chapter who are interested in learning about innovative solutions for challenges of converging Big Data and HPC. The BoF will give attendees the opportunity to hear about existing challenges and openly discuss solutions, tools, and new approaches on how to best utilize available Big Data and HPC resources.

Room: D171
12:15 pm - 1:15 pm

Software Engineers: Careers in Research

Benjamin Pritchard (Molecular Sciences Software Institute, Virginia Tech)

Research organizations are increasingly performing some sort of software development as part of their research. Such software is often developed by researchers with a combined expertise in both programming and a deep understanding of the research being performed. While such developer/scientists are indispensable, they often do not have formal career paths, and often lack recognition within academia.

Researchers in these roles are often designated Research Software Engineers in the UK, US, and Europe, with the term also gaining traction around the world. This BoF aims to raise awareness of the role and bring the community together to address challenges.

Room: D173
12:15 pm - 1:15 pm

Benchmarking Scientific Reconfigurable / FPGA Computing

Franck Cappello (Argonne National Laboratory)

In the past three years, FPGAs have gone from niche components to a central part of many data centers worldwide. But while many use cases match those of traditional accelerators (e.g., AWS), most of these millions of newly deployed FPGAs are in other configurations such as smart NICs and storage architectures. A central problem in advancing HPC with FPGAs is therefore developing new benchmarking methodologies. In this BoF, leading researchers from industry, government, and academia will showcase cutting-edge FPGAs applications related to HPC and propose evaluation metrics. Discussion will aim at creating a new benchmark. Audience participation is highly encouraged.

Room: D175
12:15 pm - 1:15 pm

Getting Scientific Software Installed

Davide Vanzo (Vanderbilt University)

We intend to provide a platform for presenting and discussing tools to deal with the ubiquitous problems that come forward when building and installing scientific software, which is known to be a tedious and time consuming task. Several user support tools for allowing scientific software to be installed and used will briefly be presented, for example (but not limited to) EasyBuild, Lmod, Spack, Singularity, etc. We will bring various experienced members of HPC user support teams and system administrators as well as users together for an open discussion on tools and best practices.

Room: D220
12:15 pm - 1:15 pm

Charm++ and AMPI: Adaptive and Asynchronous Parallel Programming

Sam White (University of Illinois, Charmworks Inc)

A community gathering about parallel programming using Charm++, Adaptive MPI, the many applications built on them, and associated tools. This session will cover recent advances in Charm++ and the experiences of application developers with Charm++. There will also be a discussion on the future directions of Charm++/AMPI and opportunities to learn more and form collaborations.

Charm++ is a production-grade many-tasking programming framework and runtime system for modern HPC systems. It offers high productivity and performance portability through features such as multicore and accelerator support, dynamic load balancing, fault tolerance, latency hiding, interoperability with MPI and OpenMP, and online job-resizing.

Room: D221
12:15 pm - 1:15 pm

Exascale Archiving - Challenges and Opportunities

Meghan McClelland (Versity Software Inc)

Archives are changing. With exascale rapidly approaching and new workloads such as machine learning and IOT/instruments shifting the use paradigm to read more than write, HPC organizations need archival storage systems that can support the rapidly increasing scale and changing workloads while meeting the capacity and financial requirements of the organization.

In this interactive BoF, an expert panel of leaders in the archiving industry will discuss and debate key archive requirements, how archives need to change to meet modern workflows, cost considerations for archives, and what long term data storage in future data centers will look like.

Room: D222
12:15 pm - 1:15 pm

IEEE CS TCHPC Meeting

IEEE CS TCHPC Meeting
The annual IEEE Computer Society Technical Consortium on High Performance Computing (TCHPC) Birds of a Feather session at SC provides a forum to discuss the consortium and its mission, activities, and initiatives with members of the community, as well as to get input from the community. It gives members of the community a chance to interact with the TCHPC volunteers and representatives from its member technical committees and to explore opportunities to become involved.

Room: D227
12:15 pm - 1:15 pm

Power API and Redfish: Standardizing Power Measurement and Control for HPC

Ryan Grant (Sandia National Laboratories)

The HPC community faces considerable constraints on power and energy of HPC installations. A standardized, vendor-neutral API for power measurement and control is needed for portable solutions to these issues at the various layers of the software stack. In this BoF, we discuss the Power API and Redfish; APIs for measurement and control of power/energy on large systems. The BoF will introduce newcomers to these efforts, differentiate the goals of the two APIs and discuss inter-operability. An interactive panel discussion with experts from involved organizations will facilitate discussions between both API communities with ample time for audience questions and comments.

Room: C140/142
5:15 pm - 6:45 pm

OpenMP® 5.0 Is Here: Find Out All the Things You Need to Know About It!

Jim Cownie (Intel Corporation)

OpenMP directives are the primary way in which shared-memory parallelism is expressed in HPC programs. Although OpenMP is 21 this year, it is not standing still. The fifth major release of the standard (OpenMP 5.0) was recently completed, and provides significant enhancements to OpenMP's capabilities in many areas, including improved support for offload accelerators.

In this BoF, we'll provide a short introduction to the new features of OpenMP 5.0 (from the people who wrote the standard), then give you plenty of time to ask them questions. We'll also give you time to grill members of the OpenMP Architecture Review Board.

Room: C141/143/149
5:15 pm - 6:45 pm

Cloud Infrastructure Solutions To Run HPC Workloads

Martial Michel (Data Machines Corporation)

Virtualization and containers have grown to see more prominent use within the realm of HPC. Adoption of these tools has enabled IT Organizations to reduce costs while making it easier to manage large pools of compute, storage and networking resources. However, performance overheads, networking integrations, and system complexity pose daunting architectural challenges.

OpenStack, Docker, Charliecloud, Singularity, Kubernetes, and Mesos all pose their own set of unique benefits and challenges. This Birds of a Feather is aimed at architects, administrators, software engineers, and scientists interested in designing and deploying cloud infrastructure solutions to run HPC workloads.

Room: C144
5:15 pm - 6:45 pm

PMIx: Enabling Workflow Orchestration

Ralph Castain (Intel Corporation)

The PMI Exascale (PMIx) community will be concluding its fourth year that included release of the PMIx v3 Standard. We will briefly review the last year's accomplishments, but the majority of the meeting will focus on how community members are using PMIx at the application level. Examples are expected to include population modeling and swarm intelligence; coordination of resource utilization between programming libraries; orchestration of dynamic workflows; and recovery from failures.

Please join the discussion as we explore the expanding uses of PMIx beyond the traditional MPI/OpenSHMEM programming models.

Room: C145
5:15 pm - 6:45 pm

OpenSHMEM in the Era of Exascale

Steve Poole (Los Alamos National Laboratory)

OpenSHMEM is a PGAS API for single-sided asynchronous scalable communications in HPC applications. OpenSHMEM is a community driven standard for the SHMEM API across multiple architectures/implementations. This BoF brings together the OpenSHMEM community to present the latest accomplishments since release of the 1.4 specification, and discuss future directions for the OpenSHMEM community as we develop version 1.5. The BoF will consist of talks from end-users,

implementers, middleware and tool developers to discuss their experiences and plans for using OpenSHMEM. We will then open the floor for discussion of the specification and our mid to long term goals.

Room: C146
5:15 pm - 6:45 pm

MPICH: A High Performance Open-Source MPI Implementation

Ken Raffenetti (Argonne National Laboratory)

MPICH is a widely used, opensource implementation of the MPI message passing standard. It has been ported to many platforms and used by several vendors and research groups as the basis for their own MPI implementations. This BoF session will provide a forum for users of MPICH as well as developers of MPI implementations derived from MPICH to discuss experiences and issues in using and porting MPICH. Future plans for MPICH will be discussed. Representatives from MPICH-derived implementations will provide brief updates on the status of their efforts. MPICH developers will also be present for an open forum discussion.

Room: C147/148/154
5:15 pm - 6:45 pm

Applications of Deep Learning in Industry and Research

Lucas Wilson (Dell EMC)

The inaugural Applications of Deep Learning in Industry and Research BoF will explore in-production and pre-production uses of deep learning (DL) across industry segments and research domains for various customer/scientist-facing or back-of-office/lab applications. We invite all data scientists, decision makers, researchers and analysts to discuss the need or desire to apply DL in their businesses or research projects, the technical challenges faced when implementing DL-based solutions, the potential value DL-based solutions have provided or may provide, lessons learned in adding deep learning to existing data center operations and applications workflows, and desires for addressing current gaps and future needs.

Room: C155/156
5:15 pm - 6:45 pm

Exascale Machine Learning

Naoya Maruyama (Lawrence Livermore National Laboratory)

HPC has been making a profound impact in recent success of data-driven machine learning. Training

deep learning models such as convolutional and recurrent neural networks with big data is an extremely compute intensive task that routinely benefits from parallel and distributed processing. Upcoming exascale machines are expected to further accelerate the advance of data-oriented machine learning. This BoF starts with short talks by experts from academia, national labs and industry about the current state of the art in HPC-based machine learning, and concludes with a panel session that provides a forum to discuss opportunities and challenges with future exascale systems.

Room: D166
5:15 pm - 6:45 pm

Consolidating the European Exascale Effort

Marcin Ostasz (European Technology Platform for High Performance Computing (ETP4HPC), Barcelona Supercomputing Center)

Europe will present the results of its Exascale programme to date (covering the entire HPC system stack and application expertise), the ambitions of the new round of basic technology, application, co-design (DEEP-EST/EuroExa), prototype and processor design (EPI) projects, its post-Exascale plans, collaboration with Big Data, IoT and other areas, and a selection of best-performing HPC technology SMEs, emphasising Europe's global contributions.

EuroHPC, an initiative of EC and Member States, will provide the resources and mechanisms needed to integrate these efforts and build world-class Exascale machines.

A summary of European HPC and its projects ('European HPC 2018') is available at www.etp4hpc.eu/euexascale.

Room: D167
5:15 pm - 6:45 pm

The Green 500: Trends in Energy Efficient Supercomputing

Wu Feng (Virginia Tech)

With power becoming a first-order design constraint on-par with performance, it is important to measure and analyze energy-efficiency trends in supercomputing. To raise the awareness of greenness as a first-order design constraint, the Green500 seeks to characterize the energy-efficiency of supercomputers for different metrics, workloads, and methodologies. This BoF discusses trends across the Green500 and highlights from the current Green500 list. In addition, the Green500, Top500, and Energy-Efficient HPC Working Group have been working together on improving power-measurement methodology, and this BoF presents case studies from sites that have made power submissions that meet the highest quality of measurement methodology.

Room: D174
5:15 pm - 6:45 pm

Big Data and Exascale Computing (BDEC2) Application Roundtable

Geoffrey Fox (Indiana University)

The emergence of large scale data analytics in science and engineering, and the explosive growth of data generated by new instruments and IoT in edge environments, has disrupted the landscape for exascale computing. The international Big Data and Extreme-Scale Computing (BDEC) initiative is starting a new workshop series, BDEC2, beginning with a comprehensive study of application requirements. This BoF will review the Pathways to Convergence already identified and preview the application requirements to be discussed at the first BDEC2 meeting in November 2018. We will seek input from the SC community on the importance of HPC to such data-intensive applications.

Room: D168
5:15 pm - 6:45 pm

Software Engineering and Reuse in Computational Science and Engineering

David Bernholdt (Oak Ridge National Laboratory)

Software engineering (SWE) for computational science and engineering (CSE) is challenging, with ever-more sophisticated and higher fidelity simulation of ever-larger and more complex problems involving larger data volumes, more domains, and more researchers. Targeting both commodity and custom high-end computers multiplies these challenges. We invest a great deal in creating these codes, but rarely talk about that experience; we just focus on the results.

Our goal is to raise awareness of SWE for CSE on supercomputers as a major challenge, and to develop an international "community of practice" to continue these important discussions outside of workshops and other "traditional" venues.

Room: D169
5:15 pm - 6:45 pm

The ARM HPC Experience: From Testbeds to Exascale

Mitsuhisa Sato (RIKEN)

The ARM architecture has gained substantial traction in the HPC community as evidenced by several ARM-based projects including the Japanese Post-K, European Mont-Blanc, U.S. DOE Vanguard, and UK GW4/EPSRC efforts, as well as the commercial availability of ARM nodes from HPC vendors including Cray, Gigabyte, HPE, and Ingrasys. We will bring together experts to share their experiences investigating ARM HPC, from testbeds with current ARM processors to plans for future exascale systems. The BoF includes a panel session where members and the audience together can discuss

progress and define the future state of ARM in the HPC community.

Room: D171
5:15 pm - 6:45 pm

Revisiting the 2008 ExaScale Computing Study and Venturing Predictions for 2028

Jeffrey Vetter (Oak Ridge National Laboratory)

On this 10-year anniversary of the 'ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems,' we will reexamine the report's predictions, and use this information to crowdsource projections for HPC architectures in 2028. First, members from the original team, including the study leader, Peter Kogge, will briefly highlight the report's hits and misses. Second, the panelists will present initial thoughts about important architectural trends for 2028, guided by the outcome from the earlier retrospective. Throughout the session, we will electronically survey the audience on these hits and misses, their predictions for 2028, and then ask the panelists to address them.

Room: D173
5:15 pm - 6:45 pm

The IO-500 and the Virtual Institute of I/O

Julian Kunkel (University of Reading)

The IO500 is quickly becoming the de facto benchmarking standard for HPC storage. Developed two years ago, the IO500 has released two official lists so far. A BoF highlight is the presentation of the third IO-500 list.

The general purpose of this BoF is to foster the IO500 and VI4IO communities to ensure forward progress towards the common goals of creating, sharing, and benefiting from a large corpus of shared storage data. We also serve as a repository of detailed information about production storage system architectures over time as a knowledge base for other researchers and system designers to use.

Room: D175
5:15 pm - 6:45 pm

Monitoring Large-Scale HPC Systems: Extracting and Presenting Meaningful System and Application Insights

Ann Gentile (Sandia National Laboratories)

We explore opportunities and challenges in extracting and presenting meaningful insights into HPC

System and Application behavior via monitoring. Panelists from large data/platform sites will interact with the audience on: Data stores to support performant analyses, Exploration of data to discover meaningful relationships, Machine learning, and more. Further, we discuss results from two multi-site reports on the "state of the practice" in HPC monitoring.

We invite system administrators, analysis and visualization developers, and application users and developers, to facilitate community progress by identifying tools, techniques, gaps, and requirements for exploratory and production scenarios. Results will be posted at: https://sites.google.com/site/monitoringlargescalehpcsystems.

Room: D220
5:15 pm - 6:45 pm

Personalized Medicine and HPC

Cristin Merritt (Alces Flight Limited)

As HPC becomes more inclusive through hardware and cloud advancements, we are now at the point where personalized medicine can take the steps necessary to move from concept to reality. But how? By combining the work of clinical researchers and hospitals with the biomedical community, we can embrace upcoming technologies to better diagnose and treat the patient. CompBioMed, the Centre of Excellence working towards the advancement of computational modeling and simulation within biomedicine presents their findings, gives some antidotes on dealing with complexity, and raises the question - How can we make personalized medicine a reality?

Room: D221
5:15 pm - 6:45 pm

Deep500: An HPC Deep Learning Benchmark and Competition

Tal Ben-Nun (ETH Zurich)

This BoF discusses the creation of a new deep learning HPC benchmark and competition focused on scientific computing applications. The panel members of the BoF are experts on various aspects of the field, coming from leading universities around the world and top hardware manufacturers in the industry. The goal of the BoF is to standardize a new benchmark, which necessitates a collaborative brainstorming of the entire HPC community.

Room: D227
5:15 pm - 6:45 pm

Workloads and Benchmarks for System Acquisition

Neil Bright (Georgia Institute of Technology)

New HPC systems must balance storage, networking, and computational power to meet the different needs of its user base and intended usage. Government agencies, universities, and companies invest significant time, effort, and money when evaluating and acquiring new HPC systems. Unfortunately, many of these evaluations are duplicated from site-to-site. This BoF aims to start a best-practices collaboration between HPC sites to reduce the effort required to make good system evaluations. This session's leaders will engage with the audience to discuss practical solutions to system benchmarking and evaluation for acquisition purposes.

Thursday, November 15th

Room: C144
12:15 pm - 1:15 pm

Volunteer Opportunities for SC Conference Planning

Christine Cuicchi (SC20, US Department of Defense HPC Modernization Program)

The SC Conference desires to provide opportunities to the HPC community to volunteer with the SC Conference Planning Committee. Attendees will meet SC Committee members and learn which areas of SC Committee planning are best suited to their interests should they wish to participate in planning future SC conferences.

SC Conference Planning Committee members will highlight the entry-level opportunities available for new volunteers and the professional development from which volunteers and their supporting institutions will benefit.

Room: C145
12:15 pm - 1:15 pm

Spectral Analysis: Building an LGBTQIA+ Community in Scientific Computing

Daniel Gens (Lawrence Berkeley National Laboratory, National Energy Research Scientific Computing Center (NERSC))

Spectral Analysis will be the first Birds of a Feather session dedicated entirely to welcoming members of the Supercomputing community who self-identify with a minority gender identity or sexual orientation. This BoF will discuss the state of the community within high performance computing (HPC), form personal and professional ties with fellow community members, and share different visions of diversity and inclusion efforts across scientific computing. Most importantly, this event will provide a safe space for sharing personal experiences and developing a set of community goals and resources.

Room: C146
12:15 pm - 1:15 pm

HPC Meets Real-Time Data: Interactive Supercomputing for Urgent Decision Making

Nick Brown (University of Edinburgh)

Big data and digitalization are creating exciting new opportunities that have the potential to move HPC well beyond traditional computational workloads. Bearing in mind the availability of fast growing social and sensor networks, the combination of HPC simulation with high velocity data and live analytics is likely to be of great importance in the future for solving societal, economic and environmental problems. The idea of this BoF is to bring together those interested in the challenges and opportunities of enhancing HPC with real-time data flows and interactive exploration capacities to support urgent decision making.

Room: C155/156
12:15 pm - 1:15 pm

SLURM User Group Meeting

Morris Jette (SchedMD LLC)

Slurm is an open source workload manager used many on TOP500 systems and provides a rich set of features including topology aware optimized resource allocation, the ability to expand and shrink jobs on demand, the ability to power down idle nodes and restart them as needed, hierarchical bank accounts with fair-share job prioritization and many resource limits. The meeting will consist of three parts: The Slurm development team will present details about changes in the new version 18.08, describe the Slurm roadmap, and solicit user feedback. Everyone interested in Slurm use and/or development is encouraged to attend.

Room: D161
12:15 pm - 1:15 pm

TCHPC Career Panel

Manish Parashar (Rutgers University)

The career panel will consist of representatives with a HPC background in industry and higher education. The panel will share advice on what different careers in the field look like after graduation. The primary audience for this event is current undergraduates, graduate students and post-docs. The panel will include representatives from industry and a broad array of higher education institutions. The format will feature a brief introduction by each speaker and then a moderated discussion with questions from the audience.

Room: D166
12:15 pm - 1:15 pm

The Power of Storytelling: Exposing User Experiences and Lessons Learned to Inspire and Instruct Technology Adoption

Mary Bass (University of Chicago, Globus)

Public relations, marketing, and communications professionals are invited to join this BoF to share experiences with finding, extracting, crafting, and publishing HPC user stories. Our goal is to define best practices for developing instructive stories that help organizations make better decisions about adopting HPC technologies and solutions. Topics include identifying and crafting a good story; winning permission to tell it; getting your story seen; and lessons learned about what works and doesn't. Attendees will leave armed with practical advice they can use to get more, better user stories in the hands of those who will benefit from hearing them.

Room: D167
12:15 pm - 1:15 pm

Collaboration Toward a Software Stack for System Power Optimization: The HPC PowerStack

Martin Schulz (Technical University Munich)

This interactive BoF brings together vendors, labs, and academic researchers to discuss an emerging community effort to develop a software stack for system-wide power optimization. The HPC PowerStack effort is the first to identify what power optimization software actors are needed; how they interoperate to achieve stable, synchronized optimization; and how to glue together existing open source projects to engineer a cost-effective but cohesive, cross-platform power stack implementation.

This BoF disseminates key insights acquired in the project, provides prototyping status updates, invites audience feedback on current directions, brainstorms solutions to open questions, and issues a call-to-action to join the collaboration.

Room: D174
12:15 pm - 1:15 pm

On Launching Ask.CI, a Q&A Platform for Research Computing, Using StackExchange and Discourse

Julie Ma (Massachusetts Green High Performance Computing Center)

In September, 2017, the Northeast Cyberteam Initiative began a project to build Ask.Cyberinfrastructure.org, aka Ask.CI, a Q&A site which will allow the research computing

community to achieve better/faster research results by making it easier to leverage/share experience and knowledge.

Establishing a Q&A site of this nature requires some tenacity. In partnership with the Campus Champions, we have gained some traction, and hope to engage the broader community to firmly establish this platform as a tool for the global research computing community. At this BoF, we will describe the process to-date, and interactively encourage the audience to join the effort.

Room: D168
12:15 pm - 1:15 pm

Navigating the SC Conference Technical Program Submission Process

Michael Heroux (Sandia National Laboratories, St. John's University)

The Supercomputing Conference series has introduced a variety of changes to the technical program in the past few year, to increase inclusivity and quality. Double-blind reviews, a two-phase review process and reproducibility appendices are most notable, along with the schedule adjustments needed to accommodate these additions. In this BoF, four SC19 technical program members will give an overview of the anticipated process for SC19, describing strategies and challenges of authors in previous years and suggesting approaches for future authors. We will also spend a substantial portion of the BoF taking feedback and input from the community for improvements.

Room: D171
12:15 pm - 1:15 pm

Data Analytics for System and Facility Energy Management

Ghaleb Abdulla (Lawrence Livermore National Laboratory)

Several leading edge supercomputing centers across the globe have been working on developing or acquiring data acquisition and management systems to support energy efficiency studies and reporting, to understand rare and important events, and to help energy providers with energy scheduling and management. This BoF presents experiences from HPC centers regarding data source integration and analytics and seeks those interested in hearing about or sharing experiences with dynamic power and energy management. It is important to get the community together to share solutions and educate each other about the challenges, success stories, and use cases.

Room: D173
12:15 pm - 1:15 pm

National Research Infrastructure: Collaborative Session

Kenton McHenry (University of Illinois)

This BoF will cover national research infrastructures including the Open Storage Network (OSN), the National Research Platform (NRP) growing out of the PRP, and the Open Science Grid (OSG). The OSN will leverage existing high-speed connections at universities and research entities across the nation. The OSN will distribute a cumulative ~5PB of storage at four of the NSF's bdhubs. The storage is intended to promote the sharing of active datasets in the research community. The NRP is committed to social engineering among a diverse group of science, R&E network, and IT leaders, as well as to provide proven end-to-end networking

Room: D175
12:15 pm - 1:15 pm

International HPC Certification Program

Julian Kunkel (University of Reading)

There is a generally accepted set of skills and competencies necessary to efficiently use HPC resources. Making clear what skills are required or recommended for a competent HPC user would benefit both the HPC service providers and practitioners. This BoF will present the status of the recently founded International HPC Certification Forum aiming to categorize, define, and examine these skills to create an ecosystem. To ensure this is an independent community-wide effort, we invite anyone interested in HPC teaching/ training to participate in the discussion and get involved. Inputs from academia, industry and HPC centers will steer the future effort.

Room: D220
12:15 pm - 1:15 pm

Achieving Performance on Large-Scale Intel Xeon-Based Systems

David Martin (Argonne National Laboratory)

This BoF, organized by the Intel Extreme Computing Users Group (IXPUG), will focus on achieving performance at scale on large Xeon-based systems. The community has gained experience in deploying generations of Xeon Phi and more recently the Xeon Scalable family of processors and is moving to achieve sustained performance. Through invited talks and open discussion, the BoF will provide a forum for tool developers, applications programmers, HPC center staff and industry experts to share tips and techniques gained through deployments on large-scale systems. Special emphasis will be given to performance across different Xeon architectures and in heterogeneous Xeon systems.

Room: D221
12:15 pm - 1:15 pm

Data Science and HPC Education and Outreach

Linda McIver (Australian Data Science Education Institute)

There are many people in the SC community doing great work in Education and Outreach in the Data Science and HPC space. It's time to build community so that we can share resources and ideas, warn each other of pitfalls, and make our combined efforts more effective. This BoF is a chance to share what you do, connect with others, and help plan the inaugural International Conference on Data Science and HPC Education and Outreach. Connecting with schools is a crucial way to increase diversity in Supercomputing.

Room: D227
12:15 pm - 1:15 pm

The Future of NSF Supported Advanced Cyberinfrastructure

Manish Parashar (National Science Foundation)

The National Science Foundation's vision and investment plans for cyberinfrastructure (CI) are designed to address the evolving needs of the science and engineering research community. Program Directors from NSF's Office of Advanced Cyberinfrastructure (OAC) will update attendees on new NSF cyberinfrastructure strategies and activities, and discuss the latest funding opportunities in advanced computing, software infrastructure, data infrastructure, networking, cybersecurity, and learning and workforce development. Presentations will also discuss Leadership Class Computing, the Research Core program, innovative capacity focused HPC, and NSF's Big Ideas. Substantial time will be devoted to Q&A between attendees and NSF staff.

Break

(back to top)

Sunday, November 11th

Room: D200s Pre-function
10:00 am - 10:30 am

Morning Break

Room: Lower Level C & D Corridors
10:00 am - 10:30 am

Morning Break

Room: D200s Pre-function
3:00 pm - 3:30 pm

Afternoon Break

Room: Lower Level C & D Corridors
3:00 pm - 3:30 pm

Afternoon Break

Monday, November 12th

Room: D200s Pre-function
10:00 am - 10:30 am

Morning Break

Room: Lower Level C & D Corridors
10:00 am - 10:30 am

Morning Break

Room: D200s Pre-function
3:00 pm - 3:30 pm

Afternoon Break

Room: Lower Level C & D Corridors
3:00 pm - 3:30 pm

Afternoon Break

Tuesday, November 13th

Room: D200s Pre-function
10:00 am - 10:30 am

Tech Program Morning Break

Room: Lower Level C & D Corridors
10:00 am - 10:30 am

Tech Program Morning Break

Room: D200s Pre-function
3:00 pm - 3:30 pm

Tech Program Afternoon Break

Room: Lower Level C & D Corridors
3:00 pm - 3:30 pm

Tech Program Afternoon Break

Wednesday, November 14th

Room: D200s Pre-function
10:00 am - 10:30 am

Tech Program Morning Break

Room: Lower Level C & D Corridors
10:00 am - 10:30 am

Tech Program Morning Break

Room: D200s Pre-function
3:00 pm - 3:30 pm

Tech Program Afternoon Break

Room: Lower Level C & D Corridors
3:00 pm - 3:30 pm

Tech Program Afternoon Break

Thursday, November 15th

Room: D200s Pre-function
10:00 am - 10:30 am

Tech Program Morning Break

Room: Lower Level C & D Corridors
10:00 am - 10:30 am

Tech Program Morning Break

Room: D200s Pre-function
3:00 pm - 3:30 pm

Tech Program Afternoon Break

Room: Lower Level C & D Corridors
3:00 pm - 3:30 pm

Tech Program Afternoon Break

Friday, November 16th

Room: Lower Level D Corridors
10:00 am - 10:30 am

Tech Program Morning Break

Doctoral Showcase

(back to top)

Tuesday, November 13th

Room: C2/3/4 Ballroom
8:30 am - 5:00 pm

Exhibition of Doctoral Showcase Posters

Compiler and Runtime Based Parallelization and Optimization for GPUs
Guray Ozen (Barcelona Supercomputing Center, Polytechnic University of Catalonia)

This thesis targets directive-based programming models to enhance their capability for GPU programming. It introduces a new dialect model, which is a combination of OpenMP and OmpSs. The new model allows the use of multiple GPUs in conjunction with the heavily multithreaded capabilities in multicore processors automatically. The thesis also focuses on code transformation techniques and proposes the LazyNP method to support nested parallelism for irregular applications such as sparse matrix operations, graph and graphics algorithms. Finally, the thesis conducts a thorough exploration of loop scheduling methods on GPUs to find the advantage and disadvantages of them. It then proposes the concept of optimized dynamic loop scheduling.

The contributions of this thesis improve the programmability of GPUs. This has had an outstanding impact on the whole OpenMP and OpenACC language committee. Additionally, it includes contributions to widely used compilers such as Mercurium, Clang and PGI compilers.

Efficient Deployment of Irregular Computations on Multi- and Many-Core Architectures
Hancheng Wu (North Carolina State University)

Multi- and manycore processors have been advancing High Performance Computing with their high throughput and power efficiency. There has been an increasing interest in accelerating irregular computations on these devices that offer massive parallelism. My thesis focuses on compiler techniques and code transformations that facilitate the deployment of irregular computations on multi- and many-core processors, aiming to achieve higher performance and better programmability. My contributions are below. We propose a compiler-based consolidation framework to improve the efficiency of irregular graph and tree computations written with Dynamic Parallelism on GPUs. We analyze and categorize parallel recursive tree traversal patterns, then provide insights on how to select the platform and code template based on identified traversal patterns. We propose compiler techniques to support a SIMT programming model on Intel multi- and many-core architectures with wide vector units, and point out the main challenges in supporting the SIMT model, especially for irregular computations.

Linear Algebra Is the Right Way to Think About Graphs
Carl Yang (University of California, Davis; Lawrence Berkeley National Laboratory)

Graph algorithms are challenging to implement on new accelerators such as GPUs. To address this

problem, GraphBLAS is an innovative on-going effort by the graph analytics community to formulate graph algorithms as sparse linear algebra, so that they can be expressed in a performant, succinct and in a backend-agnostic manner. Initial research efforts in implementing GraphBLAS on GPUs for graph processing and analytics have been promising, but challenges such as feature-incompleteness and poor performance still exist compared to their vertex-centric ("think like a vertex") graph framework counterparts. For our thesis, we propose a multi-language graph framework aiming to simplify the development of graph algorithms, which 1) provides a multi-language GraphBLAS interface for the end-users to express, develop, and refine graph algorithms more succinctly than existing distributed graph frameworks; 2) abstracts away from the end-users performance-tuning decisions; 3) utilizes the advantages of existing low-level GPU computing primitives to maintain high performance.

Designing High-Performance, Resilient, and Heterogeneity-Aware Key-Value Storage for Modern HPC Clusters

Dipti Shankar (Ohio State University)

Distributed key-value stores are being increasingly used to accelerate Big Data workloads on modern HPC clusters. The advances in HPC technologies (e.g., RDMA, SSDs) has directed several efforts towards employing hybrid storage with RDMA, for designing high- performance key-value stores. With this as basis, in my research, I take a holistic approach to designing a high-performance key-value storage system for HPC clusters that can maximize end-to-end performance while ensuring data resilience, that encompasses: (1) RDMA-enabled networking, (2) high-speed NVMs, and, (3) heterogeneous compute capabilities, available on current HPC systems. Towards this, I introduce RDMA-aware designs to enable: (1) non-blocking API semantics for designing high-performance client-side read/write pipelines, (2) fast online erasure coding for memory-efficient resilience, and, (3) SIMD-aware server-side accelerations; to enable Big Data applications to optimally leverage hybrid key-value stores in HPC environments.

Enabling Efficient Data Infrastructure and Analytics on HPC Systems

Huansong Fu (Florida State University)

We propose to leverage PGAS and one-sided communication for building data infrastructure and analytics frameworks on HPC systems. Specifically, we have developed SHMEMCache, a distributed in-memory key-value store and SHMEMGraph, a balanced graph processing framework. We have tackled unique challenges in data consistency, load balancing, etc. with novel design features. The experimental results show that SHMEMCache achieves significant performance improvements over other KV stores such as Memcached in terms of latency and throughput. In addition, SHMEMGraph achieves an average improvement of 35.5% in terms of job completion time compared to a state-of-the-art graph processing system called Gemini.

Hardware Transactional Persistent Memory

Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Parallel and Scalable Combinatorial String and Graph Algorithms on Distributed Memory Systems
Patrick Flick (Georgia Institute of Technology)

Methods for processing and analyzing DNA and genomic data are built upon combinatorial graph and string algorithms. The advent of high-throughput DNA sequencing is enabling the generation of billions of reads per experiment. Classical and sequential algorithms can no longer deal with these growing data sizes, which for the last 10 years have greatly out-paced advances in processor speeds. To process and analyze state-of-the-art genomic data sets require the design of scalable and efficient parallel algorithms and the use of large computing clusters. Here, we present our distributed-memory parallel algorithms for indexing large genomic datasets, including algorithms for construction of suffix- and LCP arrays, solving the All-Nearest-Smaller-Values problem and its application to the construction of suffix trees. Our parallel algorithms exhibit superior runtime complexity and practical performance compared to the state-of-the-art. Furthermore, we present distributed-memory algorithms for clustering de-Bruijn graphs and its application to solving a grand challenge metagenomic dataset.

Scalable Non-Blocking Krylov Solvers for Extreme-Scale Computing
Paul R. Eller (University of Illinois)

This study investigates preconditioned conjugate gradient method variations designed to reduce communication costs by decreasing the number of allreduces and overlapping communication with computation using a non-blocking allreduce. Experiments show scalable PCG methods can outperform standard PCG at scale and demonstrate the robustness of these methods.

To develop the most optimal Krylov methods we need a clear understanding of the factors limiting performance at scale. Detailed timings and network counters are used to more thoroughly measure the performance of these methods. Performance models with penalty terms are developed that provide reasonable explanations of observed performance and guide development of optimizations. The effectiveness of scalable PCG methods and these performance analysis tools is demonstrated using Quda and Nek5000, two HPC applications seeking improved performance at scale.

High Performance Middlewares for Next Generation Architectures: Challenges and Solutions
Sourav Chakraborty (Ohio State University)

The emergence of modern multi-/many-core architectures and high-performance interconnects have fueled the growth of large-scale supercomputing clusters. Due to this unprecedented growth in scale and compute density, high performance computing (HPC) middlewares now face a plethora of new challenges to solve in order to extract the best performance from such systems. In this work, we study four such challenges - a) launching and bootstrapping jobs on very large scale clusters, b) contention in collective communication, c) point-to-point communication protocols, and d) scalable fault-tolerance and recovery and propose efficient solutions for them. The proposed solutions have been implemented on MVAPICH2, a popular MPI and PGAS runtime used by scientists and HPC clusters around the

world.

In-Memory Accelerator Architectures for Machine Learning and Bioinformatics
Roman Kaplan (Israel Institute of Technology)

Most contemporary accelerators are von Neumann machines. With the increasing sizes of gathered and then processed data, memory bandwidth is the main limiting of performance. One approach to mitigate the bandwidth constraint is to bring the processing units closer to the data. This approach is known as *near-data processing* (NDP). However, NDP architecture (e.g., Hybrid Memory Cube) are still inherently limited because they are based on replicating the von Neumann architecture in memory.
My research proposes two new processing-in-storage architectures, where each bitcell can both store information and perform computation. The main building block of the architectures are memristors, an emerging memory technolgy.
The first architecture I propose, PRinS, was applied to accelerate machine learning and large-scale DNA sequence alignment. Using Associative Processing, PRinS achieves massive parallelism. The second, RASSA, accelerates DNA long read mapping, with approximate Hamming distance computation and quantifying mismatch of a pattern to voltage level.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

Productive Data Locality Optimizations in Distributed Memory
Engin Kayraklioglu (George Washington University)

With deepening memory hierarchies in HPC systems, the challenge of managing data locality gains more importance. Coincidentally, increasing ubiquity of HPC systems and wider range of disciplines utilizing HPC introduce more programmers to the HPC community. Given these two trends, it is imperative to have scalable and productive ways to manage data locality.

In this research, we address the problem in multiple ways. We propose a novel language feature that programmers can use to transform shared memory applications to distributed memory applications easily. We introduce a high-level profiling tool to help understand how distributed arrays are used in an application. As next steps, we are designing a model to describe the implementation of data locality optimizations as an engineering process, which can lend itself to combinatorial optimization. We are also implementing a profile-based automatic optimization framework that utilizes AI to replace the programmer completely in implementing optimizations for distributed memory.

The Algorithm and Framework Designs and Optimizations for Scalable Automata Processing on

HPC Platforms
Xiaodong Yu (Virginia Tech)

Automata processing could perform as the core of many applications in the areas such as network security, text mining, and bioinformatics. Achieving high-speed and scalable automata processing is exceptionally challenging. For one thing, the classic DFA representation is memory-bandwidth efficient but suffer from the state explosion problem in the presence of large datasets with complex patterns. And for another, the automata processing is inherently difficult to be parallelized due to the strong dependencies and unstructured memory-access pattern.

In this thesis, we provide a comprehensive scheme for improving automata processing efficiency. At the algorithm level, we propose JFA that uses state variables to avoid state explosion. We also propose O3FA to handle out-of-order packets in NIDS. At the implementation level, we propose a comprehensive GPU-based automata processing framework and a code-generation framework for Automata Processors. Moreover, we provide a toolchain to conduct the apple-to-apple automata processing accelerators comparison.

Pattern Matching on Massive Metadata Graphs at Scale
Tahsin Reza (University of British Columbia)

Pattern matching is a powerful graph analysis tool. Unfortunately, existing solutions have limited scalability, support only a limited set of patterns, and/or focus on only a subset of the real-world problems associated with pattern matching. First, we present a new algorithmic pipeline based on graph pruning that: (i) enables highly scalable exact pattern matching on labeled graphs, (ii) supports arbitrary patterns, (iii) enables trade-offs between precision and time-to-solution, and (iv) supports a set of popular analytics scenarios. We implement our approach on top of HavoqGT and demonstrate its advantages through strong and weak scaling experiments on massive-scale real-world (up to 257B edges) and synthetic (up to 4.4T edges) graphs, respectively, and at scales (1,024 nodes / 36,864 cores) orders of magnitude larger than used in the past for similar problems. Furthermore, we explore avenues to enable approximate matching within the graph pruning model, targeting contemporary and emerging high-impact, real-world applications.

Scalable Methods for Genome Assembly
Priyanka Ghosh (Washington State University)

Genome assembly is a fundamental problem in the field of bioinformatics wherein the goal lies in the reconstruction of an unknown genome from short DNA fragments obtained from it. With the advent of high-throughput sequencing technologies, billions of reads can be generated in a few hours. My reseach deals with addressing the challenges of conducting genome assembly at scale and developing new methods for conducting extreme-scale genome assembly suited for microbial genomes, and complex eukaryotic genomes. Our approach to the problem is two-fold, wherein we target fast and space-efficient assemblies of moderate sized genomes on a single node with our shared-memory based implementation (FastEtch). Secondly we aim to tackle the assembly of very large genomes with our distributed-memory based approach (PaKman), in an attempt to minimize the turnaround time such that it could potentially pave the way to conducting the assembly process in real-time.

Fast and Generic Concurrent Message-Passing
Hoang-Vu Dang (University of Illinois)

Communication hardware and software have a significant impact on the performance of clusters and supercomputers. Message-passing model and the Message-Passing Interface (MPI) is a widely used model of communications in the High-Performance
Computing (HPC) community. However, MPI has recently faced new challenges due to the emergence of many-core architecture and of programming models with dynamic task parallelism, assuming a large number of concurrent threads. These applications come from important classes of applications such as graph and data analytics.

In this thesis, we studied MPI under the new assumptions. We identified several factors in the standard which were inherently problematic for scalability and performance. Next, we analyzed graph, threading and data-flow frameworks to understand the communication. We then proposed a communication engine (LCI) targetting these frameworks. Our thesis benefits MPI by developing several patches in production MPI. Furthermore, LCI presents a simple and ground-up design which benefits various frameworks of study.

Room: D221
10:30 am - 12:00 pm

Doctoral Showcase I

Pattern Matching on Massive Metadata Graphs at Scale
Tahsin Reza (University of British Columbia)

Pattern matching is a powerful graph analysis tool. Unfortunately, existing solutions have limited scalability, support only a limited set of patterns, and/or focus on only a subset of the real-world problems associated with pattern matching. First, we present a new algorithmic pipeline based on graph pruning that: (i) enables highly scalable exact pattern matching on labeled graphs, (ii) supports arbitrary patterns, (iii) enables trade-offs between precision and time-to-solution, and (iv) supports a set of popular analytics scenarios. We implement our approach on top of HavoqGT and demonstrate its advantages through strong and weak scaling experiments on massive-scale real-world (up to 257B edges) and synthetic (up to 4.4T edges) graphs, respectively, and at scales (1,024 nodes / 36,864 cores) orders of magnitude larger than used in the past for similar problems. Furthermore, we explore avenues to enable approximate matching within the graph pruning model, targeting contemporary and emerging high-impact, real-world applications.

Parallel and Scalable Combinatorial String and Graph Algorithms on Distributed Memory Systems
Patrick Flick (Georgia Institute of Technology)

Methods for processing and analyzing DNA and genomic data are built upon combinatorial graph and string algorithms. The advent of high-throughput DNA sequencing is enabling the generation of billions of reads per experiment. Classical and sequential algorithms can no longer deal with these growing data sizes, which for the last 10 years have greatly out-paced advances in processor speeds. To process and analyze state-of-the-art genomic data sets require the design of scalable and efficient parallel algorithms and the use of large computing clusters. Here, we present our distributed-memory parallel algorithms for indexing large genomic datasets, including algorithms for construction of suffix- and LCP arrays, solving the All-Nearest-Smaller-Values problem and its application to the

construction of suffix trees. Our parallel algorithms exhibit superior runtime complexity and practical performance compared to the state-of-the-art. Furthermore, we present distributed-memory algorithms for clustering de-Bruijn graphs and its application to solving a grand challenge metagenomic dataset.

Linear Algebra Is the Right Way to Think About Graphs
Carl Yang (University of California, Davis; Lawrence Berkeley National Laboratory)

Graph algorithms are challenging to implement on new accelerators such as GPUs. To address this problem, GraphBLAS is an innovative on-going effort by the graph analytics community to formulate graph algorithms as sparse linear algebra, so that they can be expressed in a performant, succinct and in a backend-agnostic manner. Initial research efforts in implementing GraphBLAS on GPUs for graph processing and analytics have been promising, but challenges such as feature-incompleteness and poor performance still exist compared to their vertex-centric ("think like a vertex") graph framework counterparts. For our thesis, we propose a multi-language graph framework aiming to simplify the development of graph algorithms, which 1) provides a multi-language GraphBLAS interface for the end-users to express, develop, and refine graph algorithms more succinctly than existing distributed graph frameworks; 2) abstracts away from the end-users performance-tuning decisions; 3) utilizes the advantages of existing low-level GPU computing primitives to maintain high performance.

Scalable Methods for Genome Assembly
Priyanka Ghosh (Washington State University)

Genome assembly is a fundamental problem in the field of bioinformatics wherein the goal lies in the reconstruction of an unknown genome from short DNA fragments obtained from it. With the advent of high-throughput sequencing technologies, billions of reads can be generated in a few hours. My reseach deals with addressing the challenges of conducting genome assembly at scale and developing new methods for conducting extreme-scale genome assembly suited for microbial genomes, and complex eukaryotic genomes. Our approach to the problem is two-fold, wherein we target fast and space-efficient assemblies of moderate sized genomes on a single node with our shared-memory based implementation (FastEtch). Secondly we aim to tackle the assembly of very large genomes with our distributed-memory based approach (PaKman), in an attempt to minimize the turnaround time such that it could potentially pave the way to conducting the assembly process in real-time.

In-Memory Accelerator Architectures for Machine Learning and Bioinformatics
Roman Kaplan (Israel Institute of Technology)

Most contemporary accelerators are von Neumann machines. With the increasing sizes of gathered and then processed data, memory bandwidth is the main limiting of performance. One approach to mitigate the bandwidth constraint is to bring the processing units closer to the data. This approach is known as *near-data processing* (NDP). However, NDP architecture (e.g., Hybrid Memory Cube) are still inherently limited because they are based on replicating the von Neumann architecture in memory.
My research proposes two new processing-in-storage architectures, where each bitcell can both store information and perform computation. The main building block of the architectures are memristors, an emerging memory technolgy.
The first architecture I propose, PRinS, was applied to accelerate machine learning and large-scale DNA sequence alignment. Using Associative Processing, PRinS achieves massive parallelism. The second, RASSA, accelerates DNA long read mapping, with approximate Hamming distance

computation and quantifying mismatch of a pattern to voltage level.

Scalable Non-Blocking Krylov Solvers for Extreme-Scale Computing
Paul R. Eller (University of Illinois)

This study investigates preconditioned conjugate gradient method variations designed to reduce communication costs by decreasing the number of allreduces and overlapping communication with computation using a non-blocking allreduce. Experiments show scalable PCG methods can outperform standard PCG at scale and demonstrate the robustness of these methods.

To develop the most optimal Krylov methods we need a clear understanding of the factors limiting performance at scale. Detailed timings and network counters are used to more thoroughly measure the performance of these methods. Performance models with penalty terms are developed that provide reasonable explanations of observed performance and guide development of optimizations. The effectiveness of scalable PCG methods and these performance analysis tools is demonstrated using Quda and Nek5000, two HPC applications seeking improved performance at scale.

Room: D221
1:30 pm - 2:45 pm

Doctoral Showcase II

Productive Data Locality Optimizations in Distributed Memory
Engin Kayraklioglu (George Washington University)

With deepening memory hierarchies in HPC systems, the challenge of managing data locality gains more importance. Coincidentally, increasing ubiquity of HPC systems and wider range of disciplines utilizing HPC introduce more programmers to the HPC community. Given these two trends, it is imperative to have scalable and productive ways to manage data locality.

In this research, we address the problem in multiple ways. We propose a novel language feature that programmers can use to transform shared memory applications to distributed memory applications easily. We introduce a high-level profiling tool to help understand how distributed arrays are used in an application. As next steps, we are designing a model to describe the implementation of data locality optimizations as an engineering process, which can lend itself to combinatorial optimization. We are also implementing a profile-based automatic optimization framework that utilizes AI to replace the programmer completely in implementing optimizations for distributed memory.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best

of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

Enabling Efficient Data Infrastructure and Analytics on HPC Systems
Huansong Fu (Florida State University)

We propose to leverage PGAS and one-sided communication for building data infrastructure and analytics frameworks on HPC systems. Specifically, we have developed SHMEMCache, a distributed in-memory key-value store and SHMEMGraph, a balanced graph processing framework. We have tackled unique challenges in data consistency, load balancing, etc. with novel design features. The experimental results show that SHMEMCache achieves significant performance improvements over other KV stores such as Memcached in terms of latency and throughput. In addition, SHMEMGraph achieves an average improvement of 35.5% in terms of job completion time compared to a state-of-the-art graph processing system called Gemini.

Fast and Generic Concurrent Message-Passing
Hoang-Vu Dang (University of Illinois)

Communication hardware and software have a significant impact on the performance of clusters and supercomputers. Message-passing model and the Message-Passing Interface (MPI) is a widely used model of communications in the High-Performance
Computing (HPC) community. However, MPI has recently faced new challenges due to the emergence of many-core architecture and of programming models with dynamic task parallelism, assuming a large number of concurrent threads. These applications come from important classes of applications such as graph and data analytics.

In this thesis, we studied MPI under the new assumptions. We identified several factors in the standard which were inherently problematic for scalability and performance. Next, we analyzed graph, threading and data-flow frameworks to understand the communication. We then proposed a communication engine (LCI) targetting these frameworks. Our thesis benefits MPI by developing several patches in production MPI. Furthermore, LCI presents a simple and ground-up design which benefits various frameworks of study.

Compiler and Runtime Based Parallelization and Optimization for GPUs
Guray Ozen (Barcelona Supercomputing Center, Polytechnic University of Catalonia)

This thesis targets directive-based programming models to enhance their capability for GPU programming. It introduces a new dialect model, which is a combination of OpenMP and OmpSs. The new model allows the use of multiple GPUs in conjunction with the heavily multithreaded capabilities in multicore processors automatically. The thesis also focuses on code transformation techniques and proposes the LazyNP method to support nested parallelism for irregular applications such as sparse matrix operations, graph and graphics algorithms. Finally, the thesis conducts a thorough exploration of loop scheduling methods on GPUs to find the advantage and disadvantages of them. It then proposes the concept of optimized dynamic loop scheduling.

The contributions of this thesis improve the programmability of GPUs. This has had an outstanding

impact on the whole OpenMP and OpenACC language committee. Additionally, it includes contributions to widely used compilers such as Mercurium, Clang and PGI compilers.

Room: D221
3:30 pm - 4:45 pm

Doctoral Showcase III

The Algorithm and Framework Designs and Optimizations for Scalable Automata Processing on HPC Platforms

Xiaodong Yu (Virginia Tech)

Automata processing could perform as the core of many applications in the areas such as network security, text mining, and bioinformatics. Achieving high-speed and scalable automata processing is exceptionally challenging. For one thing, the classic DFA representation is memory-bandwidth efficient but suffer from the state explosion problem in the presence of large datasets with complex patterns. And for another, the automata processing is inherently difficult to be parallelized due to the strong dependencies and unstructured memory-access pattern.

In this thesis, we provide a comprehensive scheme for improving automata processing efficiency. At the algorithm level, we propose JFA that uses state variables to avoid state explosion. We also propose O3FA to handle out-of-order packets in NIDS. At the implementation level, we propose a comprehensive GPU-based automata processing framework and a code-generation framework for Automata Processors. Moreover, we provide a toolchain to conduct the apple-to-apple automata processing accelerators comparison.

High Performance Middlewares for Next Generation Architectures: Challenges and Solutions

Sourav Chakraborty (Ohio State University)

The emergence of modern multi-/many-core architectures and high-performance interconnects have fueled the growth of large-scale supercomputing clusters. Due to this unprecedented growth in scale and compute density, high performance computing (HPC) middlewares now face a plethora of new challenges to solve in order to extract the best performance from such systems. In this work, we study four such challenges - a) launching and bootstrapping jobs on very large scale clusters, b) contention in collective communication, c) point-to-point communication protocols, and d) scalable fault-tolerance and recovery and propose efficient solutions for them. The proposed solutions have been implemented on MVAPICH2, a popular MPI and PGAS runtime used by scientists and HPC clusters around the world.

Designing High-Performance, Resilient, and Heterogeneity-Aware Key-Value Storage for Modern HPC Clusters

Dipti Shankar (Ohio State University)

Distributed key-value stores are being increasingly used to accelerate Big Data workloads on modern HPC clusters. The advances in HPC technologies (e.g., RDMA, SSDs) has directed several efforts towards employing hybrid storage with RDMA, for designing high- performance key-value stores.

With this as basis, in my research, I take a holistic approach to designing a high-performance key-value storage system for HPC clusters that can maximize end-to-end performance while ensuring data resilience, that encompasses: (1) RDMA-enabled networking, (2) high-speed NVMs, and, (3) heterogeneous compute capabilities, available on current HPC systems. Towards this, I introduce RDMA-aware designs to enable: (1) non-blocking API semantics for designing high-performance client-side read/write pipelines, (2) fast online erasure coding for memory-efficient resilience, and, (3) SIMD-aware server-side accelerations; to enable Big Data applications to optimally leverage hybrid key-value stores in HPC environments.

Hardware Transactional Persistent Memory
Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Efficient Deployment of Irregular Computations on Multi- and Many-Core Architectures
Hancheng Wu (North Carolina State University)

Multi- and manycore processors have been advancing High Performance Computing with their high throughput and power efficiency. There has been an increasing interest in accelerating irregular computations on these devices that offer massive parallelism. My thesis focuses on compiler techniques and code transformations that facilitate the deployment of irregular computations on multi- and many-core processors, aiming to achieve higher performance and better programmability. My contributions are below. We propose a compiler-based consolidation framework to improve the efficiency of irregular graph and tree computations written with Dynamic Parallelism on GPUs. We analyze and categorize parallel recursive tree traversal patterns, then provide insights on how to select the platform and code template based on identified traversal patterns. We propose compiler techniques to support a SIMT programming model on Intel multi- and many-core architectures with wide vector units, and point out the main challenges in supporting the SIMT model, especially for irregular computations.

Room: C2/3/4 Ballroom
5:15 pm - 7:00 pm

Doctoral Showcase Posters Reception

Session Description: Reception catering is available for Technical Program registrants only.

Compiler and Runtime Based Parallelization and Optimization for GPUs
Guray Ozen (Barcelona Supercomputing Center, Polytechnic University of Catalonia)

This thesis targets directive-based programming models to enhance their capability for GPU programming. It introduces a new dialect model, which is a combination of OpenMP and OmpSs. The new model allows the use of multiple GPUs in conjunction with the heavily multithreaded capabilities in multicore processors automatically. The thesis also focuses on code transformation techniques and proposes the LazyNP method to support nested parallelism for irregular applications such as sparse matrix operations, graph and graphics algorithms. Finally, the thesis conducts a thorough exploration of loop scheduling methods on GPUs to find the advantage and disadvantages of them. It then proposes the concept of optimized dynamic loop scheduling.

The contributions of this thesis improve the programmability of GPUs. This has had an outstanding impact on the whole OpenMP and OpenACC language committee. Additionally, it includes contributions to widely used compilers such as Mercurium, Clang and PGI compilers.

Efficient Deployment of Irregular Computations on Multi- and Many-Core Architectures
Hancheng Wu (North Carolina State University)

Multi- and manycore processors have been advancing High Performance Computing with their high throughput and power efficiency. There has been an increasing interest in accelerating irregular computations on these devices that offer massive parallelism. My thesis focuses on compiler techniques and code transformations that facilitate the deployment of irregular computations on multi- and many-core processors, aiming to achieve higher performance and better programmability. My contributions are below. We propose a compiler-based consolidation framework to improve the efficiency of irregular graph and tree computations written with Dynamic Parallelism on GPUs. We analyze and categorize parallel recursive tree traversal patterns, then provide insights on how to select the platform and code template based on identified traversal patterns. We propose compiler techniques to support a SIMT programming model on Intel multi- and many-core architectures with wide vector units, and point out the main challenges in supporting the SIMT model, especially for irregular computations.

Linear Algebra Is the Right Way to Think About Graphs
Carl Yang (University of California, Davis; Lawrence Berkeley National Laboratory)

Graph algorithms are challenging to implement on new accelerators such as GPUs. To address this problem, GraphBLAS is an innovative on-going effort by the graph analytics community to formulate graph algorithms as sparse linear algebra, so that they can be expressed in a performant, succinct and in a backend-agnostic manner. Initial research efforts in implementing GraphBLAS on GPUs for graph processing and analytics have been promising, but challenges such as feature-incompleteness and poor performance still exist compared to their vertex-centric ("think like a vertex") graph framework counterparts. For our thesis, we propose a multi-language graph framework aiming to simplify the development of graph algorithms, which 1) provides a multi-language GraphBLAS interface for the end-users to express, develop, and refine graph algorithms more succinctly than existing distributed graph frameworks; 2) abstracts away from the end-users performance-tuning decisions; 3) utilizes the

advantages of existing low-level GPU computing primitives to maintain high performance.

Designing High-Performance, Resilient, and Heterogeneity-Aware Key-Value Storage for Modern HPC Clusters
Dipti Shankar (Ohio State University)

Distributed key-value stores are being increasingly used to accelerate Big Data workloads on modern HPC clusters. The advances in HPC technologies (e.g., RDMA, SSDs) has directed several efforts towards employing hybrid storage with RDMA, for designing high- performance key-value stores. With this as basis, in my research, I take a holistic approach to designing a high-performance key-value storage system for HPC clusters that can maximize end-to-end performance while ensuring data resilience, that encompasses: (1) RDMA-enabled networking, (2) high-speed NVMs, and, (3) heterogeneous compute capabilities, available on current HPC systems. Towards this, I introduce RDMA-aware designs to enable: (1) non-blocking API semantics for designing high-performance client-side read/write pipelines, (2) fast online erasure coding for memory-efficient resilience, and, (3) SIMD-aware server-side accelerations; to enable Big Data applications to optimally leverage hybrid key-value stores in HPC environments.

Enabling Efficient Data Infrastructure and Analytics on HPC Systems
Huansong Fu (Florida State University)

We propose to leverage PGAS and one-sided communication for building data infrastructure and analytics frameworks on HPC systems. Specifically, we have developed SHMEMCache, a distributed in-memory key-value store and SHMEMGraph, a balanced graph processing framework. We have tackled unique challenges in data consistency, load balancing, etc. with novel design features. The experimental results show that SHMEMCache achieves significant performance improvements over other KV stores such as Memcached in terms of latency and throughput. In addition, SHMEMGraph achieves an average improvement of 35.5% in terms of job completion time compared to a state-of-the-art graph processing system called Gemini.

Hardware Transactional Persistent Memory
Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Parallel and Scalable Combinatorial String and Graph Algorithms on Distributed Memory Systems
Patrick Flick (Georgia Institute of Technology)

Methods for processing and analyzing DNA and genomic data are built upon combinatorial graph and string algorithms. The advent of high-throughput DNA sequencing is enabling the generation of billions of reads per experiment. Classical and sequential algorithms can no longer deal with these growing data sizes, which for the last 10 years have greatly out-paced advances in processor speeds. To process and analyze state-of-the-art genomic data sets require the design of scalable and efficient parallel algorithms and the use of large computing clusters. Here, we present our distributed-memory parallel algorithms for indexing large genomic datasets, including algorithms for construction of suffix- and LCP arrays, solving the All-Nearest-Smaller-Values problem and its application to the construction of suffix trees. Our parallel algorithms exhibit superior runtime complexity and practical performance compared to the state-of-the-art. Furthermore, we present distributed-memory algorithms for clustering de-Bruijn graphs and its application to solving a grand challenge metagenomic dataset.

Scalable Non-Blocking Krylov Solvers for Extreme-Scale Computing
Paul R. Eller (University of Illinois)

This study investigates preconditioned conjugate gradient method variations designed to reduce communication costs by decreasing the number of allreduces and overlapping communication with computation using a non-blocking allreduce. Experiments show scalable PCG methods can outperform standard PCG at scale and demonstrate the robustness of these methods.

To develop the most optimal Krylov methods we need a clear understanding of the factors limiting performance at scale. Detailed timings and network counters are used to more thoroughly measure the performance of these methods. Performance models with penalty terms are developed that provide reasonable explanations of observed performance and guide development of optimizations. The effectiveness of scalable PCG methods and these performance analysis tools is demonstrated using Quda and Nek5000, two HPC applications seeking improved performance at scale.

High Performance Middlewares for Next Generation Architectures: Challenges and Solutions
Sourav Chakraborty (Ohio State University)

The emergence of modern multi-/many-core architectures and high-performance interconnects have fueled the growth of large-scale supercomputing clusters. Due to this unprecedented growth in scale and compute density, high performance computing (HPC) middlewares now face a plethora of new challenges to solve in order to extract the best performance from such systems. In this work, we study four such challenges - a) launching and bootstrapping jobs on very large scale clusters, b) contention in collective communication, c) point-to-point communication protocols, and d) scalable fault-tolerance and recovery and propose efficient solutions for them. The proposed solutions have been implemented on MVAPICH2, a popular MPI and PGAS runtime used by scientists and HPC clusters around the world.

In-Memory Accelerator Architectures for Machine Learning and Bioinformatics
Roman Kaplan (Israel Institute of Technology)

Most contemporary accelerators are von Neumann machines. With the increasing sizes of gathered and then processed data, memory bandwidth is the main limiting of performance. One approach to mitigate the bandwidth constraint is to bring the processing units closer to the data. This approach is known as *near-data processing* (NDP). However, NDP architecture (e.g., Hybrid Memory Cube) are still

inherently limited because they are based on replicating the von Neumann architecture in memory. My research proposes two new processing-in-storage architectures, where each bitcell can both store information and perform computation. The main building block of the architectures are memristors, an emerging memory technolgy.
The first architecture I propose, PRinS, was applied to accelerate machine learning and large-scale DNA sequence alignment. Using Associative Processing, PRinS achieves massive parallelism. The second, RASSA, accelerates DNA long read mapping, with approximate Hamming distance computation and quantifying mismatch of a pattern to voltage level.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

Productive Data Locality Optimizations in Distributed Memory
Engin Kayraklioglu (George Washington University)

With deepening memory hierarchies in HPC systems, the challenge of managing data locality gains more importance. Coincidentally, increasing ubiquity of HPC systems and wider range of disciplines utilizing HPC introduce more programmers to the HPC community. Given these two trends, it is imperative to have scalable and productive ways to manage data locality.

In this research, we address the problem in multiple ways. We propose a novel language feature that programmers can use to transform shared memory applications to distributed memory applications easily. We introduce a high-level profiling tool to help understand how distributed arrays are used in an application. As next steps, we are designing a model to describe the implementation of data locality optimizations as an engineering process, which can lend itself to combinatorial optimization. We are also implementing a profile-based automatic optimization framework that utilizes AI to replace the programmer completely in implementing optimizations for distributed memory.

The Algorithm and Framework Designs and Optimizations for Scalable Automata Processing on HPC Platforms
Xiaodong Yu (Virginia Tech)

Automata processing could perform as the core of many applications in the areas such as network security, text mining, and bioinformatics. Achieving high-speed and scalable automata processing is exceptionally challenging. For one thing, the classic DFA representation is memory-bandwidth efficient but suffer from the state explosion problem in the presence of large datasets with complex patterns. And for another, the automata processing is inherently difficult to be parallelized due to the strong dependencies and unstructured memory-access pattern.

In this thesis, we provide a comprehensive scheme for improving automata processing efficiency. At the algorithm level, we propose JFA that uses state variables to avoid state explosion. We also propose O3FA to handle out-of-order packets in NIDS. At the implementation level, we propose a comprehensive GPU-based automata processing framework and a code-generation framework for Automata Processors. Moreover, we provide a toolchain to conduct the apple-to-apple automata processing accelerators comparison.

Pattern Matching on Massive Metadata Graphs at Scale
Tahsin Reza (University of British Columbia)

Pattern matching is a powerful graph analysis tool. Unfortunately, existing solutions have limited scalability, support only a limited set of patterns, and/or focus on only a subset of the real-world problems associated with pattern matching. First, we present a new algorithmic pipeline based on graph pruning that: (i) enables highly scalable exact pattern matching on labeled graphs, (ii) supports arbitrary patterns, (iii) enables trade-offs between precision and time-to-solution, and (iv) supports a set of popular analytics scenarios. We implement our approach on top of HavoqGT and demonstrate its advantages through strong and weak scaling experiments on massive-scale real-world (up to 257B edges) and synthetic (up to 4.4T edges) graphs, respectively, and at scales (1,024 nodes / 36,864 cores) orders of magnitude larger than used in the past for similar problems. Furthermore, we explore avenues to enable approximate matching within the graph pruning model, targeting contemporary and emerging high-impact, real-world applications.

Scalable Methods for Genome Assembly
Priyanka Ghosh (Washington State University)

Genome assembly is a fundamental problem in the field of bioinformatics wherein the goal lies in the reconstruction of an unknown genome from short DNA fragments obtained from it. With the advent of high-throughput sequencing technologies, billions of reads can be generated in a few hours. My reseach deals with addressing the challenges of conducting genome assembly at scale and developing new methods for conducting extreme-scale genome assembly suited for microbial genomes, and complex eukaryotic genomes. Our approach to the problem is two-fold, wherein we target fast and space-efficient assemblies of moderate sized genomes on a single node with our shared-memory based implementation (FastEtch). Secondly we aim to tackle the assembly of very large genomes with our distributed-memory based approach (PaKman), in an attempt to minimize the turnaround time such that it could potentially pave the way to conducting the assembly process in real-time.

Fast and Generic Concurrent Message-Passing
Hoang-Vu Dang (University of Illinois)

Communication hardware and software have a significant impact on the performance of clusters and supercomputers. Message-passing model and the Message-Passing Interface (MPI) is a widely used model of communications in the High-Performance
Computing (HPC) community. However, MPI has recently faced new challenges due to the emergence of many-core architecture and of programming models with dynamic task parallelism, assuming a large number of concurrent threads. These applications come from important classes of applications such as graph and data analytics.

In this thesis, we studied MPI under the new assumptions. We identified several factors in the

standard which were inherently problematic for scalability and performance. Next, we analyzed graph, threading and data-flow frameworks to understand the communication. We then proposed a communication engine (LCI) targetting these frameworks. Our thesis benefits MPI by developing several patches in production MPI. Furthermore, LCI presents a simple and ground-up design which benefits various frameworks of study.

Wednesday, November 14th

Room: C2/3/4 Ballroom
8:30 am - 5:00 pm

Exhibition of Doctoral Showcase Posters

Compiler and Runtime Based Parallelization and Optimization for GPUs
Guray Ozen (Barcelona Supercomputing Center, Polytechnic University of Catalonia)

This thesis targets directive-based programming models to enhance their capability for GPU programming. It introduces a new dialect model, which is a combination of OpenMP and OmpSs. The new model allows the use of multiple GPUs in conjunction with the heavily multithreaded capabilities in multicore processors automatically. The thesis also focuses on code transformation techniques and proposes the LazyNP method to support nested parallelism for irregular applications such as sparse matrix operations, graph and graphics algorithms. Finally, the thesis conducts a thorough exploration of loop scheduling methods on GPUs to find the advantage and disadvantages of them. It then proposes the concept of optimized dynamic loop scheduling.

The contributions of this thesis improve the programmability of GPUs. This has had an outstanding impact on the whole OpenMP and OpenACC language committee. Additionally, it includes contributions to widely used compilers such as Mercurium, Clang and PGI compilers.

Efficient Deployment of Irregular Computations on Multi- and Many-Core Architectures
Hancheng Wu (North Carolina State University)

Multi- and manycore processors have been advancing High Performance Computing with their high throughput and power efficiency. There has been an increasing interest in accelerating irregular computations on these devices that offer massive parallelism. My thesis focuses on compiler techniques and code transformations that facilitate the deployment of irregular computations on multi- and many-core processors, aiming to achieve higher performance and better programmability. My contributions are below. We propose a compiler-based consolidation framework to improve the efficiency of irregular graph and tree computations written with Dynamic Parallelism on GPUs. We analyze and categorize parallel recursive tree traversal patterns, then provide insights on how to select the platform and code template based on identified traversal patterns. We propose compiler techniques to support a SIMT programming model on Intel multi- and many-core architectures with wide vector units, and point out the main challenges in supporting the SIMT model, especially for irregular computations.

Linear Algebra Is the Right Way to Think About Graphs

Carl Yang (University of California, Davis; Lawrence Berkeley National Laboratory)

Graph algorithms are challenging to implement on new accelerators such as GPUs. To address this problem, GraphBLAS is an innovative on-going effort by the graph analytics community to formulate graph algorithms as sparse linear algebra, so that they can be expressed in a performant, succinct and in a backend-agnostic manner. Initial research efforts in implementing GraphBLAS on GPUs for graph processing and analytics have been promising, but challenges such as feature-incompleteness and poor performance still exist compared to their vertex-centric ("think like a vertex") graph framework counterparts. For our thesis, we propose a multi-language graph framework aiming to simplify the development of graph algorithms, which 1) provides a multi-language GraphBLAS interface for the end-users to express, develop, and refine graph algorithms more succinctly than existing distributed graph frameworks; 2) abstracts away from the end-users performance-tuning decisions; 3) utilizes the advantages of existing low-level GPU computing primitives to maintain high performance.

Designing High-Performance, Resilient, and Heterogeneity-Aware Key-Value Storage for Modern HPC Clusters

Dipti Shankar (Ohio State University)

Distributed key-value stores are being increasingly used to accelerate Big Data workloads on modern HPC clusters. The advances in HPC technologies (e.g., RDMA, SSDs) has directed several efforts towards employing hybrid storage with RDMA, for designing high- performance key-value stores. With this as basis, in my research, I take a holistic approach to designing a high-performance key-value storage system for HPC clusters that can maximize end-to-end performance while ensuring data resilience, that encompasses: (1) RDMA-enabled networking, (2) high-speed NVMs, and, (3) heterogeneous compute capabilities, available on current HPC systems. Towards this, I introduce RDMA-aware designs to enable: (1) non-blocking API semantics for designing high-performance client-side read/write pipelines, (2) fast online erasure coding for memory-efficient resilience, and, (3) SIMD-aware server-side accelerations; to enable Big Data applications to optimally leverage hybrid key-value stores in HPC environments.

Enabling Efficient Data Infrastructure and Analytics on HPC Systems

Huansong Fu (Florida State University)

We propose to leverage PGAS and one-sided communication for building data infrastructure and analytics frameworks on HPC systems. Specifically, we have developed SHMEMCache, a distributed in-memory key-value store and SHMEMGraph, a balanced graph processing framework. We have tackled unique challenges in data consistency, load balancing, etc. with novel design features. The experimental results show that SHMEMCache achieves significant performance improvements over other KV stores such as Memcached in terms of latency and throughput. In addition, SHMEMGraph achieves an average improvement of 35.5% in terms of job completion time compared to a state-of-the-art graph processing system called Gemini.

Hardware Transactional Persistent Memory

Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM

by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Parallel and Scalable Combinatorial String and Graph Algorithms on Distributed Memory Systems
Patrick Flick (Georgia Institute of Technology)

Methods for processing and analyzing DNA and genomic data are built upon combinatorial graph and string algorithms. The advent of high-throughput DNA sequencing is enabling the generation of billions of reads per experiment. Classical and sequential algorithms can no longer deal with these growing data sizes, which for the last 10 years have greatly out-paced advances in processor speeds. To process and analyze state-of-the-art genomic data sets require the design of scalable and efficient parallel algorithms and the use of large computing clusters. Here, we present our distributed-memory parallel algorithms for indexing large genomic datasets, including algorithms for construction of suffix- and LCP arrays, solving the All-Nearest-Smaller-Values problem and its application to the construction of suffix trees. Our parallel algorithms exhibit superior runtime complexity and practical performance compared to the state-of-the-art. Furthermore, we present distributed-memory algorithms for clustering de-Bruijn graphs and its application to solving a grand challenge metagenomic dataset.

Scalable Non-Blocking Krylov Solvers for Extreme-Scale Computing
Paul R. Eller (University of Illinois)

This study investigates preconditioned conjugate gradient method variations designed to reduce communication costs by decreasing the number of allreduces and overlapping communication with computation using a non-blocking allreduce. Experiments show scalable PCG methods can outperform standard PCG at scale and demonstrate the robustness of these methods.

To develop the most optimal Krylov methods we need a clear understanding of the factors limiting performance at scale. Detailed timings and network counters are used to more thoroughly measure the performance of these methods. Performance models with penalty terms are developed that provide reasonable explanations of observed performance and guide development of optimizations. The effectiveness of scalable PCG methods and these performance analysis tools is demonstrated using Quda and Nek5000, two HPC applications seeking improved performance at scale.

High Performance Middlewares for Next Generation Architectures: Challenges and Solutions
Sourav Chakraborty (Ohio State University)

The emergence of modern multi-/many-core architectures and high-performance interconnects have fueled the growth of large-scale supercomputing clusters. Due to this unprecedented growth in scale and compute density, high performance computing (HPC) middlewares now face a plethora of new challenges to solve in order to extract the best performance from such systems. In this work, we study

four such challenges - a) launching and bootstrapping jobs on very large scale clusters, b) contention in collective communication, c) point-to-point communication protocols, and d) scalable fault-tolerance and recovery and propose efficient solutions for them. The proposed solutions have been implemented on MVAPICH2, a popular MPI and PGAS runtime used by scientists and HPC clusters around the world.

In-Memory Accelerator Architectures for Machine Learning and Bioinformatics
Roman Kaplan (Israel Institute of Technology)

Most contemporary accelerators are von Neumann machines. With the increasing sizes of gathered and then processed data, memory bandwidth is the main limiting of performance. One approach to mitigate the bandwidth constraint is to bring the processing units closer to the data. This approach is known as *near-data processing* (NDP). However, NDP architecture (e.g., Hybrid Memory Cube) are still inherently limited because they are based on replicating the von Neumann architecture in memory.
My research proposes two new processing-in-storage architectures, where each bitcell can both store information and perform computation. The main building block of the architectures are memristors, an emerging memory technolgy.
The first architecture I propose, PRinS, was applied to accelerate machine learning and large-scale DNA sequence alignment. Using Associative Processing, PRinS achieves massive parallelism. The second, RASSA, accelerates DNA long read mapping, with approximate Hamming distance computation and quantifying mismatch of a pattern to voltage level.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

Productive Data Locality Optimizations in Distributed Memory
Engin Kayraklioglu (George Washington University)

With deepening memory hierarchies in HPC systems, the challenge of managing data locality gains more importance. Coincidentally, increasing ubiquity of HPC systems and wider range of disciplines utilizing HPC introduce more programmers to the HPC community. Given these two trends, it is imperative to have scalable and productive ways to manage data locality.

In this research, we address the problem in multiple ways. We propose a novel language feature that programmers can use to transform shared memory applications to distributed memory applications easily. We introduce a high-level profiling tool to help understand how distributed arrays are used in an application. As next steps, we are designing a model to describe the implementation of data locality optimizations as an engineering process, which can lend itself to combinatorial optimization. We are

also implementing a profile-based automatic optimization framework that utilizes AI to replace the programmer completely in implementing optimizations for distributed memory.

The Algorithm and Framework Designs and Optimizations for Scalable Automata Processing on HPC Platforms
Xiaodong Yu (Virginia Tech)

Automata processing could perform as the core of many applications in the areas such as network security, text mining, and bioinformatics. Achieving high-speed and scalable automata processing is exceptionally challenging. For one thing, the classic DFA representation is memory-bandwidth efficient but suffer from the state explosion problem in the presence of large datasets with complex patterns. And for another, the automata processing is inherently difficult to be parallelized due to the strong dependencies and unstructured memory-access pattern.

In this thesis, we provide a comprehensive scheme for improving automata processing efficiency. At the algorithm level, we propose JFA that uses state variables to avoid state explosion. We also propose O3FA to handle out-of-order packets in NIDS. At the implementation level, we propose a comprehensive GPU-based automata processing framework and a code-generation framework for Automata Processors. Moreover, we provide a toolchain to conduct the apple-to-apple automata processing accelerators comparison.

Pattern Matching on Massive Metadata Graphs at Scale
Tahsin Reza (University of British Columbia)

Pattern matching is a powerful graph analysis tool. Unfortunately, existing solutions have limited scalability, support only a limited set of patterns, and/or focus on only a subset of the real-world problems associated with pattern matching. First, we present a new algorithmic pipeline based on graph pruning that: (i) enables highly scalable exact pattern matching on labeled graphs, (ii) supports arbitrary patterns, (iii) enables trade-offs between precision and time-to-solution, and (iv) supports a set of popular analytics scenarios. We implement our approach on top of HavoqGT and demonstrate its advantages through strong and weak scaling experiments on massive-scale real-world (up to 257B edges) and synthetic (up to 4.4T edges) graphs, respectively, and at scales (1,024 nodes / 36,864 cores) orders of magnitude larger than used in the past for similar problems. Furthermore, we explore avenues to enable approximate matching within the graph pruning model, targeting contemporary and emerging high-impact, real-world applications.

Scalable Methods for Genome Assembly
Priyanka Ghosh (Washington State University)

Genome assembly is a fundamental problem in the field of bioinformatics wherein the goal lies in the reconstruction of an unknown genome from short DNA fragments obtained from it. With the advent of high-throughput sequencing technologies, billions of reads can be generated in a few hours. My reseach deals with addressing the challenges of conducting genome assembly at scale and developing new methods for conducting extreme-scale genome assembly suited for microbial genomes, and complex eukaryotic genomes. Our approach to the problem is two-fold, wherein we target fast and space-efficient assemblies of moderate sized genomes on a single node with our shared-memory based implementation (FastEtch). Secondly we aim to tackle the assembly of very large genomes with our distributed-memory based approach (PaKman), in an attempt to minimize the turnaround time

such that it could potentially pave the way to conducting the assembly process in real-time.

Fast and Generic Concurrent Message-Passing
Hoang-Vu Dang (University of Illinois)

Communication hardware and software have a significant impact on the performance of clusters and supercomputers. Message-passing model and the Message-Passing Interface (MPI) is a widely used model of communications in the High-Performance
Computing (HPC) community. However, MPI has recently faced new challenges due to the emergence of many-core architecture and of programming models with dynamic task parallelism, assuming a large number of concurrent threads. These applications come from important classes of applications such as graph and data analytics.

In this thesis, we studied MPI under the new assumptions. We identified several factors in the standard which were inherently problematic for scalability and performance. Next, we analyzed graph, threading and data-flow frameworks to understand the communication. We then proposed a communication engine (LCI) targetting these frameworks. Our thesis benefits MPI by developing several patches in production MPI. Furthermore, LCI presents a simple and ground-up design which benefits various frameworks of study.

Thursday, November 15th

Room: C2/3/4 Ballroom
8:30 am - 5:00 pm

Exhibition of Doctoral Showcase Posters

Compiler and Runtime Based Parallelization and Optimization for GPUs
Guray Ozen (Barcelona Supercomputing Center, Polytechnic University of Catalonia)

This thesis targets directive-based programming models to enhance their capability for GPU programming. It introduces a new dialect model, which is a combination of OpenMP and OmpSs. The new model allows the use of multiple GPUs in conjunction with the heavily multithreaded capabilities in multicore processors automatically. The thesis also focuses on code transformation techniques and proposes the LazyNP method to support nested parallelism for irregular applications such as sparse matrix operations, graph and graphics algorithms. Finally, the thesis conducts a thorough exploration of loop scheduling methods on GPUs to find the advantage and disadvantages of them. It then proposes the concept of optimized dynamic loop scheduling.

The contributions of this thesis improve the programmability of GPUs. This has had an outstanding impact on the whole OpenMP and OpenACC language committee. Additionally, it includes contributions to widely used compilers such as Mercurium, Clang and PGI compilers.

Efficient Deployment of Irregular Computations on Multi- and Many-Core Architectures
Hancheng Wu (North Carolina State University)

Multi- and manycore processors have been advancing High Performance Computing with their high throughput and power efficiency. There has been an increasing interest in accelerating irregular computations on these devices that offer massive parallelism. My thesis focuses on compiler techniques and code transformations that facilitate the deployment of irregular computations on multi- and many-core processors, aiming to achieve higher performance and better programmability. My contributions are below. We propose a compiler-based consolidation framework to improve the efficiency of irregular graph and tree computations written with Dynamic Parallelism on GPUs. We analyze and categorize parallel recursive tree traversal patterns, then provide insights on how to select the platform and code template based on identified traversal patterns. We propose compiler techniques to support a SIMT programming model on Intel multi- and many-core architectures with wide vector units, and point out the main challenges in supporting the SIMT model, especially for irregular computations.

Linear Algebra Is the Right Way to Think About Graphs

Carl Yang (University of California, Davis; Lawrence Berkeley National Laboratory)

Graph algorithms are challenging to implement on new accelerators such as GPUs. To address this problem, GraphBLAS is an innovative on-going effort by the graph analytics community to formulate graph algorithms as sparse linear algebra, so that they can be expressed in a performant, succinct and in a backend-agnostic manner. Initial research efforts in implementing GraphBLAS on GPUs for graph processing and analytics have been promising, but challenges such as feature-incompleteness and poor performance still exist compared to their vertex-centric ("think like a vertex") graph framework counterparts. For our thesis, we propose a multi-language graph framework aiming to simplify the development of graph algorithms, which 1) provides a multi-language GraphBLAS interface for the end-users to express, develop, and refine graph algorithms more succinctly than existing distributed graph frameworks; 2) abstracts away from the end-users performance-tuning decisions; 3) utilizes the advantages of existing low-level GPU computing primitives to maintain high performance.

Designing High-Performance, Resilient, and Heterogeneity-Aware Key-Value Storage for Modern HPC Clusters

Dipti Shankar (Ohio State University)

Distributed key-value stores are being increasingly used to accelerate Big Data workloads on modern HPC clusters. The advances in HPC technologies (e.g., RDMA, SSDs) has directed several efforts towards employing hybrid storage with RDMA, for designing high- performance key-value stores. With this as basis, in my research, I take a holistic approach to designing a high-performance key-value storage system for HPC clusters that can maximize end-to-end performance while ensuring data resilience, that encompasses: (1) RDMA-enabled networking, (2) high-speed NVMs, and, (3) heterogeneous compute capabilities, available on current HPC systems. Towards this, I introduce RDMA-aware designs to enable: (1) non-blocking API semantics for designing high-performance client-side read/write pipelines, (2) fast online erasure coding for memory-efficient resilience, and, (3) SIMD-aware server-side accelerations; to enable Big Data applications to optimally leverage hybrid key-value stores in HPC environments.

Enabling Efficient Data Infrastructure and Analytics on HPC Systems

Huansong Fu (Florida State University)

We propose to leverage PGAS and one-sided communication for building data infrastructure and

analytics frameworks on HPC systems. Specifically, we have developed SHMEMCache, a distributed in-memory key-value store and SHMEMGraph, a balanced graph processing framework. We have tackled unique challenges in data consistency, load balancing, etc. with novel design features. The experimental results show that SHMEMCache achieves significant performance improvements over other KV stores such as Memcached in terms of latency and throughput. In addition, SHMEMGraph achieves an average improvement of 35.5% in terms of job completion time compared to a state-of-the-art graph processing system called Gemini.

Hardware Transactional Persistent Memory
Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Parallel and Scalable Combinatorial String and Graph Algorithms on Distributed Memory Systems
Patrick Flick (Georgia Institute of Technology)

Methods for processing and analyzing DNA and genomic data are built upon combinatorial graph and string algorithms. The advent of high-throughput DNA sequencing is enabling the generation of billions of reads per experiment. Classical and sequential algorithms can no longer deal with these growing data sizes, which for the last 10 years have greatly out-paced advances in processor speeds. To process and analyze state-of-the-art genomic data sets require the design of scalable and efficient parallel algorithms and the use of large computing clusters. Here, we present our distributed-memory parallel algorithms for indexing large genomic datasets, including algorithms for construction of suffix- and LCP arrays, solving the All-Nearest-Smaller-Values problem and its application to the construction of suffix trees. Our parallel algorithms exhibit superior runtime complexity and practical performance compared to the state-of-the-art. Furthermore, we present distributed-memory algorithms for clustering de-Bruijn graphs and its application to solving a grand challenge metagenomic dataset.

Scalable Non-Blocking Krylov Solvers for Extreme-Scale Computing
Paul R. Eller (University of Illinois)

This study investigates preconditioned conjugate gradient method variations designed to reduce communication costs by decreasing the number of allreduces and overlapping communication with computation using a non-blocking allreduce. Experiments show scalable PCG methods can outperform standard PCG at scale and demonstrate the robustness of these methods.

To develop the most optimal Krylov methods we need a clear understanding of the factors limiting performance at scale. Detailed timings and network counters are used to more thoroughly measure the performance of these methods. Performance models with penalty terms are developed that provide reasonable explanations of observed performance and guide development of optimizations. The effectiveness of scalable PCG methods and these performance analysis tools is demonstrated using Quda and Nek5000, two HPC applications seeking improved performance at scale.

High Performance Middlewares for Next Generation Architectures: Challenges and Solutions
Sourav Chakraborty (Ohio State University)

The emergence of modern multi-/many-core architectures and high-performance interconnects have fueled the growth of large-scale supercomputing clusters. Due to this unprecedented growth in scale and compute density, high performance computing (HPC) middlewares now face a plethora of new challenges to solve in order to extract the best performance from such systems. In this work, we study four such challenges - a) launching and bootstrapping jobs on very large scale clusters, b) contention in collective communication, c) point-to-point communication protocols, and d) scalable fault-tolerance and recovery and propose efficient solutions for them. The proposed solutions have been implemented on MVAPICH2, a popular MPI and PGAS runtime used by scientists and HPC clusters around the world.

In-Memory Accelerator Architectures for Machine Learning and Bioinformatics
Roman Kaplan (Israel Institute of Technology)

Most contemporary accelerators are von Neumann machines. With the increasing sizes of gathered and then processed data, memory bandwidth is the main limiting of performance. One approach to mitigate the bandwidth constraint is to bring the processing units closer to the data. This approach is known as *near-data processing* (NDP). However, NDP architecture (e.g., Hybrid Memory Cube) are still inherently limited because they are based on replicating the von Neumann architecture in memory.
My research proposes two new processing-in-storage architectures, where each bitcell can both store information and perform computation. The main building block of the architectures are memristors, an emerging memory technolgy.
The first architecture I propose, PRinS, was applied to accelerate machine learning and large-scale DNA sequence alignment. Using Associative Processing, PRinS achieves massive parallelism. The second, RASSA, accelerates DNA long read mapping, with approximate Hamming distance computation and quantifying mismatch of a pattern to voltage level.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

Productive Data Locality Optimizations in Distributed Memory
Engin Kayraklioglu (George Washington University)

With deepening memory hierarchies in HPC systems, the challenge of managing data locality gains more importance. Coincidentally, increasing ubiquity of HPC systems and wider range of disciplines utilizing HPC introduce more programmers to the HPC community. Given these two trends, it is imperative to have scalable and productive ways to manage data locality.

In this research, we address the problem in multiple ways. We propose a novel language feature that programmers can use to transform shared memory applications to distributed memory applications easily. We introduce a high-level profiling tool to help understand how distributed arrays are used in an application. As next steps, we are designing a model to describe the implementation of data locality optimizations as an engineering process, which can lend itself to combinatorial optimization. We are also implementing a profile-based automatic optimization framework that utilizes AI to replace the programmer completely in implementing optimizations for distributed memory.

The Algorithm and Framework Designs and Optimizations for Scalable Automata Processing on HPC Platforms
Xiaodong Yu (Virginia Tech)

Automata processing could perform as the core of many applications in the areas such as network security, text mining, and bioinformatics. Achieving high-speed and scalable automata processing is exceptionally challenging. For one thing, the classic DFA representation is memory-bandwidth efficient but suffer from the state explosion problem in the presence of large datasets with complex patterns. And for another, the automata processing is inherently difficult to be parallelized due to the strong dependencies and unstructured memory-access pattern.

In this thesis, we provide a comprehensive scheme for improving automata processing efficiency. At the algorithm level, we propose JFA that uses state variables to avoid state explosion. We also propose O3FA to handle out-of-order packets in NIDS. At the implementation level, we propose a comprehensive GPU-based automata processing framework and a code-generation framework for Automata Processors. Moreover, we provide a toolchain to conduct the apple-to-apple automata processing accelerators comparison.

Pattern Matching on Massive Metadata Graphs at Scale
Tahsin Reza (University of British Columbia)

Pattern matching is a powerful graph analysis tool. Unfortunately, existing solutions have limited scalability, support only a limited set of patterns, and/or focus on only a subset of the real-world problems associated with pattern matching. First, we present a new algorithmic pipeline based on graph pruning that: (i) enables highly scalable exact pattern matching on labeled graphs, (ii) supports arbitrary patterns, (iii) enables trade-offs between precision and time-to-solution, and (iv) supports a set of popular analytics scenarios. We implement our approach on top of HavoqGT and demonstrate its advantages through strong and weak scaling experiments on massive-scale real-world (up to 257B edges) and synthetic (up to 4.4T edges) graphs, respectively, and at scales (1,024 nodes / 36,864 cores) orders of magnitude larger than used in the past for similar problems. Furthermore, we explore avenues to enable approximate matching within the graph pruning model, targeting contemporary and emerging high-impact, real-world applications.

Scalable Methods for Genome Assembly

Priyanka Ghosh (Washington State University)

Genome assembly is a fundamental problem in the field of bioinformatics wherein the goal lies in the reconstruction of an unknown genome from short DNA fragments obtained from it. With the advent of high-throughput sequencing technologies, billions of reads can be generated in a few hours. My reseach deals with addressing the challenges of conducting genome assembly at scale and developing new methods for conducting extreme-scale genome assembly suited for microbial genomes, and complex eukaryotic genomes. Our approach to the problem is two-fold, wherein we target fast and space-efficient assemblies of moderate sized genomes on a single node with our shared-memory based implementation (FastEtch). Secondly we aim to tackle the assembly of very large genomes with our distributed-memory based approach (PaKman), in an attempt to minimize the turnaround time such that it could potentially pave the way to conducting the assembly process in real-time.

Fast and Generic Concurrent Message-Passing

Hoang-Vu Dang (University of Illinois)

Communication hardware and software have a significant impact on the performance of clusters and supercomputers. Message-passing model and the Message-Passing Interface (MPI) is a widely used model of communications in the High-Performance
Computing (HPC) community. However, MPI has recently faced new challenges due to the emergence of many-core architecture and of programming models with dynamic task parallelism, assuming a large number of concurrent threads. These applications come from important classes of applications such as graph and data analytics.

In this thesis, we studied MPI under the new assumptions. We identified several factors in the standard which were inherently problematic for scalability and performance. Next, we analyzed graph, threading and data-flow frameworks to understand the communication. We then proposed a communication engine (LCI) targetting these frameworks. Our thesis benefits MPI by developing several patches in production MPI. Furthermore, LCI presents a simple and ground-up design which benefits various frameworks of study.

Early Career Program

(back to top)

Monday, November 12th

Room: A302
8:30 am - 10:00 am

Connecting and Thinking Strategically through Your Strengths

Session Description: Developing a career in HPC requires a variety of skills and activities including forming important connections with peers, sponsors, collaborators, students, and others. This session will provide a jump start to helping you form connections and think strategically about your career. The session will include group activities aimed at helping you develop your communication skills and understand your strengths as well as an open and interactive "fireside chat" with leaders in the HPC and SC conference communities about how to think strategically about your career and get involved in the SC conference community and committees.

Welcome and Introduction *Mary Ann Leung (Sustainable Horizons Institute)*
This program is designed to help guide early-career researchers within the first five years of a permanent position in navigating a successful career path. Topics include strategic career planning, career development, publishing, and presenting research. The program also includes mentor-protégé matching, a discussion on the importance of mentoring, speed matching, and mentor-protégé activities during the session and throughout the week. This session presents an overview of the day's activities and a fun activity for participants to get to know each other and build their communication skills.

Building a Career on Your Strengths *Mary Ann Leung (Sustainable Horizons Institute)*
Building a career in high performance computing requires a variety of skills including mastering the art of working with people and building teams. In this session, participants will explore the idea of using your strengths to build your career, teams, and projects. Funded participants will be provided access to the Gallup Poll Strengthsfinder Assessment prior to the conference and will use the results in an interactive session following a brief discussion.

Learning to Lead in HPC - Strategies to Start Your Leadership Journey *Ralph McEldowney (US Department of Defense HPC Modernization Program, Air Force Research Laboratory), Jeff Hammond (Intel Corporation), Christopher Johnson (University of Utah), Damian Rouson (Sourcery Institute), Katherine Yelick (Lawrence Berkeley National Laboratory)*
The Learning to Lead panel brings together trusted leaders in the HPC community with a broad set of personal experiences. The panel will explore what is leadership, why it is important to consider leadership early in your career, how to find leadership learning opportunities, and what are their core leadership values. We'll then open this session up to Q/A.

Room: A302
10:30 am - 12:00 pm

CV Review and Career Development Panel

Session Description: Developing a vibrant research career consists of a few interrelated components. These components are publications, collaborations, and funding. A track record of publications attracts collaborations and funding, which would in turn facilitate research and lead to more publications. As running records of publications and other achievements, CVs attracts further opportunities for funding and collaborations. And, in academia, successfully developing these components leads to tenure. The goal of this session is to bring together experts who have excelled in developing these pieces to share their experience and advice with ECP participants in the format of panel presentations. Prior to the event, participants may provide their CVs for review and receive feedback from experienced researchers. After panel presentations, panelists and participants will further the discussions of presented topics through group discussion.

CV Review *Yipkei Kwok (Missouri Western State University)*
Developing a vibrant research career consists of a few interrelated components. These components are publications, collaborations, and funding. A track record of publications attracts collaborations and funding, which would in turn facilitate research and lead to more publications. As running records of publications and other achievements, CVs attract further opportunities for funding and collaborations. And, in academia, successfully developing these components leads to tenure. The goal of this session is to bring together experts who have excelled in developing these pieces to share their experience and advice with ECP participants in the format of panel presentations. Prior to the event, participants may provide their CVs for review and receive feedback from experienced researchers.

Career Development Panel *Jeffrey K. Hollingsworth (University of Maryland), Jaydeep Bardhan (GlaxoSmithKline), Kerstin Kleese (Brookhaven National Laboratory)*
A group of HPC-veteran panelists will present their insights on topics including the Academic Tenure Process and How to Stay Ahead, Establishing Connections with Industry, and Fostering Research Collaborations. To facilitate interactions between panelists and participants, panelists will lead group discussions based on the guiding questions posed in their presentations.

Room: A302
1:30 pm - 3:00 pm

Early Career Survival Guide for Successful Communication: Preparing Effective Grant Proposals, Publications, and Presentations

Session Description: While technical excellence is a necessary condition for a successful scientific career, it is often not a sufficient condition. Unless we know how to convey our technical advances and exciting research ideas to a broader audience, they will not generate the interest and impact needed to propel our careers. The second important component of successful research is the ability to understand audience (e.g. proposal and paper reviewers, readers) and the strength to take constructively critics. This session will touch upon communication skills for successful grant proposal writing, publications, technical presentations, and networking. Panel discussions, and hands-on

practice with your peers will provide you with the opportunity to learn and enhance your communication skills.

Understanding the Reader *Lucy Nowell (US Department of Energy Office of Advanced Scientific Computing Research)*
This session discusses how to maximize the success of your proposal and increase the probability of a positive review. Particularly, you will learn about the important role of a Program Manager (PM) in this process. Dr. Lucy Nowell, a seasoned DOE Program Manager, will share her insights on how to decide to submit to a call, how to avoid common mistakes in proposal writing, and how to position a proposal to get the best reviews.

Communication with the Reader *Torsten Hoefler (ETH Zurich)*
This session addresses the question of how to engage with the readers of our scientific publications, and more in particular, how to address peer review critiques of our scientific journal and proceedings papers. While peer review critiques can be uncomfortable to digest, they provide a wealth of information about how our intended message is received by the audience, giving us the importunity to fine tune our message for maximum impact. After a brief presentation on how to make the best use of reviewer critiques, there will be room for questions and answers on this topic.

Presenting / Communication *Bert Debusschere (Sandia National Laboratories), Olga Scrivner (Indiana University)*
This session will cover the general relevance of presentation skills, use of ignite talks, and elevator pitches. It will include a short hands-on workshop: create your own elevator pitch / ignite talk on your research topic.

Room: A2 Ballroom
3:30 pm - 4:10 pm

Building Lasting and Effective Mentoring Relationships

Session Description: Building a successful career in HPC requires developing and maintaining important connections with others in the field. One of these important connections is the Mentor/Protege relationship. Mentoring relationships can be invaluable to an individual's personal and professional growth; providing coaching, counseling and exposure to the challenges in the field. This joint session between the ECP and Students@SC programs will provide guidance on fostering these mentoring relationships. Specifically, this session will aid in: outlining the benefits of the mentor/protege relationship, identifying a lasting mentor, maintaining a mentoring relationship, and finding opportunities to give back. This sessions will also provide an overview of the week-long mentoring initiatives and plans for continued engagement.

Mentor-Protégé Informational Session *Kurt Ferreira (Sandia National Laboratories)*
This session will include a panel of speakers with a history of mentoring. Topics will include the following: identifying a lasting mentor and areas in which you need mentoring, initiative needed by a protégé to maintain a mentoring relationship, finding opportunities to give back and start mentoring now. It will also

include an overview of our conference week long mentoring initiatives and plans for continued engagement.

3:30 pm - 4:10 pm

Students@SC Mentoring Panel

Session Description: This joint informational session between the SC Early Career and Students@SC programs will provide practical guidance on fostering and maintaining mentoring relationships. Specifically, this session will aid in: outlining the benefits of the mentor/protege relationship, identifying a lasting mentor, maintaining a mentoring relationship, and finding opportunities to give back by becoming a mentor. This session will also provide attendees the opportunity to ask mentoring related questions from an experienced and diverse panel of invited guests.

Emerging Technologies

(back to top)

Monday, November 12th

Room: Booth 619
7:00 pm - 9:00 pm

Emerging Technologies Showcase (Opening Gala)

Session Description: The Emerging Technologies track will examine innovative solutions that may significantly improve and extend the world of HPC in the next five to fifteen years. Consisting of carefully selected exhibits, Emerging Technologies will showcase innovative technologies from industry, government labs, or academia. Typically these technologies will be very early-stage and so not yet available from industry as products.

Energy Efficient Computer in the Exascale Age
Matthew T. Ziegler (Lenovo)

Achieving exascale-level computing requires innovations in all aspects of computing from novel chipset designs to datacenter construction. Exascale debates typically revolve around novel architectures that would be capable of exascale computing. Importantly though, faster typically is synonymous with more power. It's estimated an exascale system would need somewhere in the 20MW range of input power to run. Most datacenters do not have that amount of power to dedicate solely to computing. Therefore, a holistic datacenter design approach is just as critical as to the underlying technology inside the datacenter. Lenovo has partnered with Leibniz Supercomputing Center (LRZ) in Munich to attack the datacenter challenges starting with an efficient liquid-cooling approach that removes internal fans. The next generation of innovation involves utilizing absorption chilling to produce chilled water using the heat exhaust from computers. In essence, the heat that the computer system generates is responsible for creating the chilled water needed to cool other systems in the datacenter. By implementing absorption chiller technology in conjunction with liquid cooling, Lenovo has moved closer to a solution that will ensure the power required in exascale computing is delivered as efficiently as possible.

DAQDB - a Distributed Key-Value Store for Petascale Hot Storage
Grzegorz Jereczek (Intel Corporation), Maciej Maciejewski (Intel Corporation), Piotr Pelplinski (Intel Corporation), Jakub Radtke (Intel Corporation), Pawel Makowski (Intel Corporation), Aleksandra Wisz (Intel Corporation)

High-capacity persistent memory, fast SSD-based storage, and platform-integrated RDMA network transport open new opportunities in the design of high-bandwidth data acquisition systems. Petascale hot storage solutions for data incoming at rate of multi TB/s over long periods of time are now becoming feasible.

In this project we design and implement DAQDB - a distributed key-value store for high-bandwidth data acquisition systems like the LHC experiments at CERN. After high luminosity upgrades of the LHC, data will be produced at a rate of 6 TB/s. This poses interesting challenges in terms of bandwidth and petascale capacity requirements, if hot storage solutions can be even considered viable options. The DAQDB project is to address them and thus should open completely new alternatives in the design of those systems.

Parallelware Analyzer: Speeding Up the Parallel Software Development Lifecycle.

Toni Collis (Appentra Solutions, Collis-Holmes Innovations Limited), Manuel Arenaz (Appentra Solutions), Javier Novo Rodríguez (Appentra Solutions), Daniel Otero Oubiña (Appentra Solutions), Jaime González Cuevas (Appentra Solutions), Belén Torrente Torrente (Appentra Solutions)

Parallelware Analyzer is the second product in the Parallelware Tools Suite and is being designed around the needs of the user. arallelware Analyzer is a command line reporting tool designed to help he HPC application developer speed up their development process, and also provide the HPC service rovider a detailed understanding of what parallelization software solutions are being used on their ystems. The tool enables the user to 'zoom in' on what they need to know, providing key information or understanding hybrid MPI+X applications from the point of view of computations, memory and ontrol flow. The reports will cover code features such as understanding memory access patterns, data ayout and data scoping; identifying functions free of side-effects beyond procedure boundaries; nalyzing dependencies beyond loop boundaries to identify regions that are already parallelized and omputational patterns that offer parallelization opportunities; annotating the source code with parallelization primitives including vectorization, SIMD, multi-threading, offloading to accelerators etc. This is all done by static analysis of the source code using LLVM-based compiler technology, without the need to 'instrument' or benchmark your code and can even be run on your laptop, irrespective of how many cores your code is designed to run on.

NVXL Acceleration Platform for Polymorphic Acceleration

Jeffery Bunting (NVXL Technology Inc), Samit Chaudhuri (NVXL Technology Inc), Michael Chang (NVXL Technology Inc)

NVXL has engineered a scalable approach to address the ever-increasing need for compute performance using a new technology: polymorphic acceleration. Polymorphic acceleration is the only technology that is built on the general premise that a multitude of compute-intensive applications needs a multitude of computing architectures including CPU, FPGA, GPU, or a new breed of ASICs that are currently being invented. Hence, instead of building silos of specific computing architectures, which is the current state of the art, data centers can use polymorphic acceleration to offer software-defined acceleration platforms in real time using a flexible accelerator pool that can easily be expanded and upgraded. Application performance is improved at a lower cost without the risk of betting on one particular device architecture.

FlexLION: Scalable and Reconfigurable All-to-All Photonic Interconnects

Roberto Proietti (University of California, Davis), Xian Xiao (University of California, Davis), Sebastian Werner (University of California, Davis), Pouya Fotouhi (University of California, Davis), Jeff Messig (Enablence Technologies Inc), S.J. Ben Yoo (University of California, Davis)

We propose FlexLION (Flexible Low-Latency Interconnect Optical Network) to pursue energy-efficient, scalable, and reconfigurable all-to-all silicon photonic (SiPh) interconnects. FlexLION offers contention-less and arbitration-free all-to-all communication among selected nodes (processors or electronic switches) while providing bandwidth-adaptive interconnection between the other nodes. The proposed technology consists of a scalable all-to-all Thin-CLOS architecture [1] based on arrayed waveguide grating routers (AWGRs) combined with micro-ring-based flexible wavelength selective filters and optical MEMS (Microelectromechanical systems) [2]. Benchmarking simulation studies indicate ≥ 3× energy efficiency improvements compared to state-of-the-art switching solutions for typical workloads. At the SC14 Emerging Technologies track, we presented the first all-to-all optically-interconnected 8-node testbed demo with 32-port AWGRs and FPGA-based nodes, and at SC17, we reported technology advancements toward close 2.5D integration between electronic ICs and silicon photonic transceivers interconnected by AWGRs. For the SC18 Emerging Technology showcase, we will present the following innovations: (1) design, fabrication and testing of a 64-port All-to-All Thin-CLOS AWGR prototype; (2) architecture and simulation studies for the proposed FlexLION; (3) current research activities toward SiPh integration of FlexLION.

A Cost-Effective Flexible System Optimized for DNN and ML

I-Hsin Chung (IBM), Yomi Yeh (H3 Platform Inc), Andy Yang (H3 Platform Inc), Brian Pan (H3 Platform Inc)

Hardware accelerators (e.g., GPU) are increasingly used for compute-intensive tasks (e.g., AI and HPC). When multiple accelerator and storage devices are present, direct data paths between the devices bypassing the host memory may be used (P2P). Current P2P provided by NVIDIA CUDA driver is limited to the NVIDIA GPUs under the same PCIe root complex and only up to 9 GPUs allowed in the P2P communication.

In our design, we used a simplified architecture as the basic building block. The new PCIe switch allows PCIe ID translation between different PCIe domains and customized routing. Together with the PCIe Gen 4, the blocks can stack together to scale out. This design is especially desired for the collective communications in DNN/ML and many HPC applications. Compared to other PCIe expansion enclosures, our design allows a CPU card installed to make the system self-sufficient/operational.

On the system software side, our solution breaks the 9-GPU under the same PCIe root complex limit and is not limited to NVIDIA GPUs. For example, the data can be transferred between NVMe storage and GPU memory directly.

Overall the new design provides a more cost-effective, robust and flexible solution that is optimized for DNN/ML and HPC applications.

The Data-Centric Future and Gen-Z's Next Generation Interconnect

Kurtis Bowman (Gen-Z Consortium)

Current computer architectures allow for network and storage transfers to occur at much lower rates than memory transfers, so they must have separate buses, control signals, and command structures. Processors must wait endlessly for these transfers to finish or must find other work to do. A great deal

of time is spent moving data between buffers to allow components working at highly different speeds to communicate effectively. Extra hardware is often needed to create DMA channels that perform transfers outside the normal flow of instructions and data. Resuming or cancelling partially completed transfers is difficult and error-prone. Gen-Z technology is different: a high-bandwidth, low-latency fabric with separate media and memory controllers that can be realized inside or beyond traditional chassis limits. Gen-Z enables much higher throughput and much lower complexity for big data solutions in such applications as data analytics, deep packet inspection, artificial intelligence, machine learning, and video and image processing.

The Gen-Z Consortium, an industry consortium comprised of over 50 leading technology companies, recently released its Core Specification 1.0 to the public.

Tuesday, November 13th

Room: Booth 619
10:00 am - 6:00 pm

Emerging Technologies Showcase (Day 1)

Session Description: The Emerging Technologies track will examine innovative solutions that may significantly improve and extend the world of HPC in the next five to fifteen years. Consisting of carefully selected exhibits, Emerging Technologies will showcase innovative technologies from industry, government labs, or academia. Typically these technologies will be very early-stage and so not yet available from industry as products.

Energy Efficient Computer in the Exascale Age
Matthew T. Ziegler (Lenovo)

Achieving exascale-level computing requires innovations in all aspects of computing from novel chipset designs to datacenter construction. Exascale debates typically revolve around novel architectures that would be capable of exascale computing. Importantly though, faster typically is synonymous with more power. It's estimated an exascale system would need somewhere in the 20MW range of input power to run. Most datacenters do not have that amount of power to dedicate solely to computing. Therefore, a holistic datacenter design approach is just as critical as to the underlying technology inside the datacenter. Lenovo has partnered with Leibniz Supercomputing Center (LRZ) in Munich to attack the datacenter challenges starting with an efficient liquid-cooling approach that removes internal fans. The next generation of innovation involves utilizing absorption chilling to produce chilled water using the heat exhaust from computers. In essence, the heat that the computer system generates is responsible for creating the chilled water needed to cool other systems in the datacenter. By implementing absorption chiller technology in conjunction with liquid cooling, Lenovo has moved closer to a solution that will ensure the power required in exascale computing is delivered as efficiently as possible.

DAQDB - a Distributed Key-Value Store for Petascale Hot Storage

Grzegorz Jereczek (Intel Corporation), Maciej Maciejewski (Intel Corporation), Piotr Pelplinski (Intel Corporation), Jakub Radtke (Intel Corporation), Pawel Makowski (Intel Corporation), Aleksandra Wisz (Intel Corporation)

High-capacity persistent memory, fast SSD-based storage, and platform-integrated RDMA network transport open new opportunities in the design of high-bandwidth data acquisition systems. Petascale hot storage solutions for data incoming at rate of multi TB/s over long periods of time are now becoming feasible.

In this project we design and implement DAQDB - a distributed key-value store for high-bandwidth data acquisition systems like the LHC experiments at CERN. After high luminosity upgrades of the LHC, data will be produced at a rate of 6 TB/s. This poses interesting challenges in terms of bandwidth and petascale capacity requirements, if hot storage solutions can be even considered viable options. The DAQDB project is to address them and thus should open completely new alternatives in the design of those systems.

Parallelware Analyzer: Speeding Up the Parallel Software Development Lifecycle.

Toni Collis (Appentra Solutions, Collis-Holmes Innovations Limited), Manuel Arenaz (Appentra Solutions), Javier Novo Rodríguez (Appentra Solutions), Daniel Otero Oubiña (Appentra Solutions), Jaime González Cuevas (Appentra Solutions), Belén Torrente Torrente (Appentra Solutions)

Parallelware Analyzer is the second product in the Parallelware Tools Suite and is being designed around the needs of the user. Parallelware Analyzer is a command line reporting tool designed to help the HPC application developer speed up their development process, and also provide the HPC service provider a detailed understanding of what parallelization software solutions are being used on their systems. The tool enables the user to 'zoom in' on what they need to know, providing key information for understanding hybrid MPI+X applications from the point of view of computations, memory and control flow. The reports will cover code features such as understanding memory access patterns, data layout and data scoping; identifying functions free of side-effects beyond procedure boundaries; analyzing dependencies beyond loop boundaries to identify regions that are already parallelized and computational patterns that offer parallelization opportunities; annotating the source code with parallelization primitives including vectorization, SIMD, multi-threading, offloading to accelerators etc. This is all done by static analysis of the source code using LLVM-based compiler technology, without the need to 'instrument' or benchmark your code and can even be run on your laptop, irrespective of how many cores your code is designed to run on.

NVXL Acceleration Platform for Polymorphic Acceleration

Jeffery Bunting (NVXL Technology Inc), Samit Chaudhuri (NVXL Technology Inc), Michael Chang (NVXL Technology Inc)

NVXL has engineered a scalable approach to address the ever-increasing need for compute performance using a new technology: polymorphic acceleration. Polymorphic acceleration is the only technology that is built on the general premise that a multitude of compute-intensive applications needs a multitude of computing architectures including CPU, FPGA, GPU, or a new breed of ASICs that are currently being invented. Hence, instead of building silos of specific computing architectures,

which is the current state of the art, data centers can use polymorphic acceleration to offer software-defined acceleration platforms in real time using a flexible accelerator pool that can easily be expanded and upgraded. Application performance is improved at a lower cost without the risk of betting on one particular device architecture.

FlexLION: Scalable and Reconfigurable All-to-All Photonic Interconnects

Roberto Proietti (University of California, Davis), Xian Xiao (University of California, Davis), Sebastian Werner (University of California, Davis), Pouya Fotouhi (University of California, Davis), Jeff Messig (Enablence Technologies Inc), S.J. Ben Yoo (University of California, Davis)

We propose FlexLION (Flexible Low-Latency Interconnect Optical Network) to pursue energy-efficient, scalable, and reconfigurable all-to-all silicon photonic (SiPh) interconnects. FlexLION offers contention-less and arbitration-free all-to-all communication among selected nodes (processors or electronic switches) while providing bandwidth-adaptive interconnection between the other nodes. The proposed technology consists of a scalable all-to-all Thin-CLOS architecture [1] based on arrayed waveguide grating routers (AWGRs) combined with micro-ring-based flexible wavelength selective filters and optical MEMS (Microelectromechanical systems) [2]. Benchmarking simulation studies indicate ≥ 3× energy efficiency improvements compared to state-of-the-art switching solutions for typical workloads. At the SC14 Emerging Technologies track, we presented the first all-to-all optically-interconnected 8-node testbed demo with 32-port AWGRs and FPGA-based nodes, and at SC17, we reported technology advancements toward close 2.5D integration between electronic ICs and silicon photonic transceivers interconnected by AWGRs. For the SC18 Emerging Technology showcase, we will present the following innovations: (1) design, fabrication and testing of a 64-port All-to-All Thin-CLOS AWGR prototype; (2) architecture and simulation studies for the proposed FlexLION; (3) current research activities toward SiPh integration of FlexLION.

A Cost-Effective Flexible System Optimized for DNN and ML

I-Hsin Chung (IBM), Yomi Yeh (H3 Platform Inc), Andy Yang (H3 Platform Inc), Brian Pan (H3 Platform Inc)

Hardware accelerators (e.g., GPU) are increasingly used for compute-intensive tasks (e.g., AI and HPC). When multiple accelerator and storage devices are present, direct data paths between the devices bypassing the host memory may be used (P2P). Current P2P provided by NVIDIA CUDA driver is limited to the NVIDIA GPUs under the same PCIe root complex and only up to 9 GPUs allowed in the P2P communication.

In our design, we used a simplified architecture as the basic building block. The new PCIe switch allows PCIe ID translation between different PCIe domains and customized routing. Together with the PCIe Gen 4, the blocks can stack together to scale out. This design is especially desired for the collective communications in DNN/ML and many HPC applications. Compared to other PCIe expansion enclosures, our design allows a CPU card installed to make the system self-sufficient/operational.

On the system software side, our solution breaks the 9-GPU under the same PCIe root complex limit and is not limited to NVIDIA GPUs. For example, the data can be transferred between NVMe storage and GPU memory directly.

Overall the new design provides a more cost-effective, robust and flexible solution that is optimized for DNN/ML and HPC applications.

The Data-Centric Future and Gen-Z's Next Generation Interconnect
Kurtis Bowman (Gen-Z Consortium)

Current computer architectures allow for network and storage transfers to occur at much lower rates than memory transfers, so they must have separate buses, control signals, and command structures. Processors must wait endlessly for these transfers to finish or must find other work to do. A great deal of time is spent moving data between buffers to allow components working at highly different speeds to communicate effectively. Extra hardware is often needed to create DMA channels that perform transfers outside the normal flow of instructions and data. Resuming or cancelling partially completed transfers is difficult and error-prone. Gen-Z technology is different: a high-bandwidth, low-latency fabric with separate media and memory controllers that can be realized inside or beyond traditional chassis limits. Gen-Z enables much higher throughput and much lower complexity for big data solutions in such applications as data analytics, deep packet inspection, artificial intelligence, machine learning, and video and image processing.

The Gen-Z Consortium, an industry consortium comprised of over 50 leading technology companies, recently released its Core Specification 1.0 to the public.

Wednesday, November 14th

Room: Booth 619
10:00 am - 6:00 pm

Emerging Technologies Showcase (Day 2)

Session Description: The Emerging Technologies track will examine innovative solutions that may significantly improve and extend the world of HPC in the next five to fifteen years. Consisting of carefully selected exhibits, Emerging Technologies will showcase innovative technologies from industry, government labs, or academia. Typically these technologies will be very early-stage and so not yet available from industry as products.

Energy Efficient Computer in the Exascale Age
Matthew T. Ziegler (Lenovo)

Achieving exascale-level computing requires innovations in all aspects of computing from novel chipset designs to datacenter construction. Exascale debates typically revolve around novel architectures that would be capable of exascale computing. Importantly though, faster typically is synonymous with more power. It's estimated an exascale system would need somewhere in the 20MW range of input power to run. Most datacenters do not have that amount of power to dedicate solely to computing. Therefore, a holistic datacenter design approach is just as critical as to the

underlying technology inside the datacenter. Lenovo has partnered with Leibniz Supercomputing Center (LRZ) in Munich to attack the datacenter challenges starting with an efficient liquid-cooling approach that removes internal fans. The next generation of innovation involves utilizing absorption chilling to produce chilled water using the heat exhaust from computers. In essence, the heat that the computer system generates is responsible for creating the chilled water needed to cool other systems in the datacenter. By implementing absorption chiller technology in conjunction with liquid cooling, Lenovo has moved closer to a solution that will ensure the power required in exascale computing is delivered as efficiently as possible.

DAQDB - a Distributed Key-Value Store for Petascale Hot Storage

Grzegorz Jereczek (Intel Corporation), Maciej Maciejewski (Intel Corporation), Piotr Pelplinski (Intel Corporation), Jakub Radtke (Intel Corporation), Pawel Makowski (Intel Corporation), Aleksandra Wisz (Intel Corporation)

High-capacity persistent memory, fast SSD-based storage, and platform-integrated RDMA network transport open new opportunities in the design of high-bandwidth data acquisition systems. Petascale hot storage solutions for data incoming at rate of multi TB/s over long periods of time are now becoming feasible.

In this project we design and implement DAQDB - a distributed key-value store for high-bandwidth data acquisition systems like the LHC experiments at CERN. After high luminosity upgrades of the LHC, data will be produced at a rate of 6 TB/s. This poses interesting challenges in terms of bandwidth and petascale capacity requirements, if hot storage solutions can be even considered viable options. The DAQDB project is to address them and thus should open completely new alternatives in the design of those systems.

Parallelware Analyzer: Speeding Up the Parallel Software Development Lifecycle.

Toni Collis (Appentra Solutions, Collis-Holmes Innovations Limited), Manuel Arenaz (Appentra Solutions), Javier Novo Rodríguez (Appentra Solutions), Daniel Otero Oubiña (Appentra Solutions), Jaime González Cuevas (Appentra Solutions), Belén Torrente Torrente (Appentra Solutions)

Parallelware Analyzer is the second product in the Parallelware Tools Suite and is being designed around the needs of the user. Parallelware Analyzer is a command line reporting tool designed to help the HPC application developer speed up their development process, and also provide the HPC service provider a detailed understanding of what parallelization software solutions are being used on their systems. The tool enables the user to 'zoom in' on what they need to know, providing key information for understanding hybrid MPI+X applications from the point of view of computations, memory and control flow. The reports will cover code features such as understanding memory access patterns, data layout and data scoping; identifying functions free of side-effects beyond procedure boundaries; analyzing dependencies beyond loop boundaries to identify regions that are already parallelized and computational patterns that offer parallelization opportunities; annotating the source code with parallelization primitives including vectorization, SIMD, multi-threading, offloading to accelerators etc. This is all done by static analysis of the source code using LLVM-based compiler technology, without the need to 'instrument' or benchmark your code and can even be run on your laptop, irrespective of how many cores your code is designed to run on.

NVXL Acceleration Platform for Polymorphic Acceleration

Jeffery Bunting (NVXL Technology Inc), Samit Chaudhuri (NVXL Technology Inc), Michael Chang (NVXL Technology Inc)

NVXL has engineered a scalable approach to address the ever-increasing need for compute performance using a new technology: polymorphic acceleration. Polymorphic acceleration is the only technology that is built on the general premise that a multitude of compute-intensive applications needs a multitude of computing architectures including CPU, FPGA, GPU, or a new breed of ASICs that are currently being invented. Hence, instead of building silos of specific computing architectures, which is the current state of the art, data centers can use polymorphic acceleration to offer software-defined acceleration platforms in real time using a flexible accelerator pool that can easily be expanded and upgraded. Application performance is improved at a lower cost without the risk of betting on one particular device architecture.

FlexLION: Scalable and Reconfigurable All-to-All Photonic Interconnects

Roberto Proietti (University of California, Davis), Xian Xiao (University of California, Davis), Sebastian Werner (University of California, Davis), Pouya Fotouhi (University of California, Davis), Jeff Messig (Enablence Technologies Inc), S.J. Ben Yoo (University of California, Davis)

We propose FlexLION (Flexible Low-Latency Interconnect Optical Network) to pursue energy-efficient, scalable, and reconfigurable all-to-all silicon photonic (SiPh) interconnects. FlexLION offers contention-less and arbitration-free all-to-all communication among selected nodes (processors or electronic switches) while providing bandwidth-adaptive interconnection between the other nodes. The proposed technology consists of a scalable all-to-all Thin-CLOS architecture [1] based on arrayed waveguide grating routers (AWGRs) combined with micro-ring-based flexible wavelength selective filters and optical MEMS (Microelectromechanical systems) [2]. Benchmarking simulation studies indicate ≥ 3× energy efficiency improvements compared to state-of-the-art switching solutions for typical workloads. At the SC14 Emerging Technologies track, we presented the first all-to-all optically-interconnected 8-node testbed demo with 32-port AWGRs and FPGA-based nodes, and at SC17, we reported technology advancements toward close 2.5D integration between electronic ICs and silicon photonic transceivers interconnected by AWGRs. For the SC18 Emerging Technology showcase, we will present the following innovations: (1) design, fabrication and testing of a 64-port All-to-All Thin-CLOS AWGR prototype; (2) architecture and simulation studies for the proposed FlexLION; (3) current research activities toward SiPh integration of FlexLION.

A Cost-Effective Flexible System Optimized for DNN and ML

I-Hsin Chung (IBM), Yomi Yeh (H3 Platform Inc), Andy Yang (H3 Platform Inc), Brian Pan (H3 Platform Inc)

Hardware accelerators (e.g., GPU) are increasingly used for compute-intensive tasks (e.g., AI and HPC). When multiple accelerator and storage devices are present, direct data paths between the devices bypassing the host memory may be used (P2P). Current P2P provided by NVIDIA CUDA driver is limited to the NVIDIA GPUs under the same PCIe root complex and only up to 9 GPUs allowed in the P2P communication.

In our design, we used a simplified architecture as the basic building block. The new PCIe switch

allows PCIe ID translation between different PCIe domains and customized routing. Together with the PCIe Gen 4, the blocks can stack together to scale out. This design is especially desired for the collective communications in DNN/ML and many HPC applications. Compared to other PCIe expansion enclosures, our design allows a CPU card installed to make the system self-sufficient/operational.

On the system software side, our solution breaks the 9-GPU under the same PCIe root complex limit and is not limited to NVIDIA GPUs. For example, the data can be transferred between NVMe storage and GPU memory directly.

Overall the new design provides a more cost-effective, robust and flexible solution that is optimized for DNN/ML and HPC applications.

The Data-Centric Future and Gen-Z's Next Generation Interconnect
Kurtis Bowman (Gen-Z Consortium)

Current computer architectures allow for network and storage transfers to occur at much lower rates than memory transfers, so they must have separate buses, control signals, and command structures. Processors must wait endlessly for these transfers to finish or must find other work to do. A great deal of time is spent moving data between buffers to allow components working at highly different speeds to communicate effectively. Extra hardware is often needed to create DMA channels that perform transfers outside the normal flow of instructions and data. Resuming or cancelling partially completed transfers is difficult and error-prone. Gen-Z technology is different: a high-bandwidth, low-latency fabric with separate media and memory controllers that can be realized inside or beyond traditional chassis limits. Gen-Z enables much higher throughput and much lower complexity for big data solutions in such applications as data analytics, deep packet inspection, artificial intelligence, machine learning, and video and image processing.

The Gen-Z Consortium, an industry consortium comprised of over 50 leading technology companies, recently released its Core Specification 1.0 to the public.

Thursday, November 15th

Room: Booth 619
10:00 am - 3:00 pm

Emerging Technologies Showcase (Day 3)

Session Description: The Emerging Technologies track will examine innovative solutions that may significantly improve and extend the world of HPC in the next five to fifteen years. Consisting of carefully selected exhibits, Emerging Technologies will showcase innovative technologies from industry, government labs, or academia. Typically these technologies will be very early-stage and so not yet available from industry as products.

Energy Efficient Computer in the Exascale Age

Matthew T. Ziegler (Lenovo)

Achieving exascale-level computing requires innovations in all aspects of computing from novel chipset designs to datacenter construction. xascale debates typically revolve around novel rchitectures that would be capable of exascale computing. Importantly though, faster typically is ynonymous with more power. I 's estimated an exascale system would need somewhere in the 2 MW range of input power to run. Most datacenters do not have that amount of power to dedicate s lely to computing. Th refore, a holistic datacenter design approach is just as critical as to the un erlying technology inside the datacenter. Lenovo has partnered with Leibniz Supercomputing Ce ter (LRZ) in Munich to attack the datacenter challenges starting with an efficient liquid-cooling ap roach that removes internal fans. The next generation of innovation involves utilizing absorption ch illing to produce chilled water using the heat exhaust from computers. In essence, the heat that the computer system generates is responsible for creating the chilled water needed to cool other systems in the datacenter. By implementing absorption chiller technology in conjunction with liquid cooling, Lenovo has moved closer to a solution that will ensure the power required in exascale computing is delivered as efficiently as possible.

DAQDB - a Distributed Key-Value Store for Petascale Hot Storage

Grzegorz Jereczek (Intel Corporation), Maciej Maciejewski (Intel Corporation), Piotr Pelplinski (Intel Corporation), Jakub Radtke (Intel Corporation), Pawel Makowski (Intel Corporation), Aleksandra Wisz (Intel Corporation)

High-capacity persistent memory, fast SSD-based storage, and platform-integrated RDMA network transport open new opportunities in the design of high-bandwidth data acquisition systems. Petascale hot storage solutions for data incoming at rate of multi TB/s over long periods of time are now becoming feasible.

In this project we design and implement DAQDB - a distributed key-value store for high-bandwidth data acquisition systems like the LHC experiments at CERN. After high luminosity upgrades of the LHC, data will be produced at a rate of 6 TB/s. This poses interesting challenges in terms of bandwidth and petascale capacity requirements, if hot storage solutions can be even considered viable options. The DAQDB project is to address them and thus should open completely new alternatives in the design of those systems.

Parallelware Analyzer: Speeding Up the Parallel Software Development Lifecycle.

Toni Collis (Appentra Solutions, Collis-Holmes Innovations Limited), Manuel Arenaz (Appentra Solutions), Javier Novo Rodríguez (Appentra Solutions), Daniel Otero Oubiña (Appentra Solutions), Jaime González Cuevas (Appentra Solutions), Belén Torrente Torrente (Appentra Solutions)

Parallelware Analyzer is the second product in the Parallelware Tools Suite and is being designed around the needs of the user. Parallelware Analyzer is a command line reporting tool designed to help the HPC application developer speed up their development process, and also provide the HPC service provider a detailed understanding of what parallelization software solutions are being used on their systems. The tool enables the user to 'zoom in' on what they need to know, providing key information for understanding hybrid MPI+X applications from the point of view of computations, memory and

control flow. The reports will cover code features such as understanding memory access patterns, data layout and data scoping; identifying functions free of side-effects beyond procedure boundaries; analyzing dependencies beyond loop boundaries to identify regions that are already parallelized and computational patterns that offer parallelization opportunities; annotating the source code with parallelization primitives including vectorization, SIMD, multi-threading, offloading to accelerators etc. This is all done by static analysis of the source code using LLVM-based compiler technology, without the need to 'instrument' or benchmark your code and can even be run on your laptop, irrespective of how many cores your code is designed to run on.

NVXL Acceleration Platform for Polymorphic Acceleration

Jeffery Bunting (NVXL Technology Inc), Samit Chaudhuri (NVXL Technology Inc), Michael Chang (NVXL Technology Inc)

NVXL has engineered a scalable approach to address the ever-increasing need for compute performance using a new technology: polymorphic acceleration. Polymorphic acceleration is the only technology that is built on the general premise that a multitude of compute-intensive applications needs a multitude of computing architectures including CPU, FPGA, GPU, or a new breed of ASICs that are currently being invented. Hence, instead of building silos of specific computing architectures, which is the current state of the art, data centers can use polymorphic acceleration to offer software-defined acceleration platforms in real time using a flexible accelerator pool that can easily be expanded and upgraded. Application performance is improved at a lower cost without the risk of betting on one particular device architecture.

FlexLION: Scalable and Reconfigurable All-to-All Photonic Interconnects

Roberto Proietti (University of California, Davis), Xian Xiao (University of California, Davis), Sebastian Werner (University of California, Davis), Pouya Fotouhi (University of California, Davis), Jeff Messig (Enablence Technologies Inc), S.J. Ben Yoo (University of California, Davis)

We propose FlexLION (Flexible Low-Latency Interconnect Optical Network) to pursue energy-efficient, scalable, and reconfigurable all-to-all silicon photonic (SiPh) interconnects. FlexLION offers contention-less and arbitration-free all-to-all communication among selected nodes (processors or electronic switches) while providing bandwidth-adaptive interconnection between the other nodes. The proposed technology consists of a scalable all-to-all Thin-CLOS architecture [1] based on arrayed waveguide grating routers (AWGRs) combined with micro-ring-based flexible wavelength selective filters and optical MEMS (Microelectromechanical systems) [2]. Benchmarking simulation studies indicate ≥ 3× energy efficiency improvements compared to state-of-the-art switching solutions for typical workloads. At the SC14 Emerging Technologies track, we presented the first all-to-all optically-interconnected 8-node testbed demo with 32-port AWGRs and FPGA-based nodes, and at SC17, we reported technology advancements toward close 2.5D integration between electronic ICs and silicon photonic transceivers interconnected by AWGRs. For the SC18 Emerging Technology showcase, we will present the following innovations: (1) design, fabrication and testing of a 64-port All-to-All Thin-CLOS AWGR prototype; (2) architecture and simulation studies for the proposed FlexLION; (3) current research activities toward SiPh integration of FlexLION.

A Cost-Effective Flexible System Optimized for DNN and ML

I-Hsin Chung (IBM), Yomi Yeh (H3 Platform Inc), Andy Yang (H3 Platform Inc), Brian Pan (H3 Platform Inc)

Hardware accelerators (e.g., GPU) are increasingly used for compute-intensive tasks (e.g., AI and HPC). When multiple accelerator and storage devices are present, direct data paths between the devices bypassing the host memory may be used (P2P). Current P2P provided by NVIDIA CUDA driver is limited to the NVIDIA GPUs under the same PCIe root complex and only up to 9 GPUs allowed in the P2P communication.

In our design, we used a simplified architecture as the basic building block. The new PCIe switch allows PCIe ID translation between different PCIe domains and customized routing. Together with the PCIe Gen 4, the blocks can stack together to scale out. This design is especially desired for the collective communications in DNN/ML and many HPC applications. Compared to other PCIe expansion enclosures, our design allows a CPU card installed to make the system self-sufficient/operational.

On the system software side, our solution breaks the 9-GPU under the same PCIe root complex limit and is not limited to NVIDIA GPUs. For example, the data can be transferred between NVMe storage and GPU memory directly.

Overall the new design provides a more cost-effective, robust and flexible solution that is optimized for DNN/ML and HPC applications.

The Data-Centric Future and Gen-Z's Next Generation Interconnect

Kurtis Bowman (Gen-Z Consortium)

Current computer architectures allow for network and storage transfers to occur at much lower rates than memory transfers, so they must have separate buses, control signals, and command structures. Processors must wait endlessly for these transfers to finish or must find other work to do. A great deal of time is spent moving data between buffers to allow components working at highly different speeds to communicate effectively. Extra hardware is often needed to create DMA channels that perform transfers outside the normal flow of instructions and data. Resuming or cancelling partially completed transfers is difficult and error-prone. Gen-Z technology is different: a high-bandwidth, low-latency fabric with separate media and memory controllers that can be realized inside or beyond traditional chassis limits. Gen-Z enables much higher throughput and much lower complexity for big data solutions in such applications as data analytics, deep packet inspection, artificial intelligence, machine learning, and video and image processing.

The Gen-Z Consortium, an industry consortium comprised of over 50 leading technology companies, recently released its Core Specification 1.0 to the public.

Exhibitor Forum

(back to top)

Tuesday, November 13th

Room: D173
10:30 am - 12:00 pm

Networking

Cybersecurity Considerations and Best Practices for Supercomputers
Brennen Wright (Keysight Technologies Inc)

How do you know if your supercomputer has been compromised? Are you relying on firewalls and network perimeter devices for protection? Maybe it's not connected to the internet, but a supercomputer is vulnerable the same way nuclear reactors are vulnerable. Have you tested your job scheduler application for vulnerabilities? What pro-active measures can you take to better defend your supercomputer? What if you could see the communication that happens "inside" your supercomputer cluster? Let Ixia, an industry leader in security testing and monitoring, explain how.

Plug and Play IP Fabrics
Jigar Shah (Extreme Networks)

Next generation datacenter technologies that make use of BGP EVPN based IP fabrics could be cumbersome to provision, especially for large scale deployments. It takes a lot of time to manually configure fabrics using CLI. There are automation tools available that would setup the IP fabric within a few seconds based on the underlying physical topology. These tools could be part of the switch firmware or standalone and help with Day 0 to Day N provisioning, troubleshooting and potential remediation of underlay and overlay networks.

Interconnect Your Future with Mellanox "Smart" Interconnect
Scot A. Schultz (Mellanox Technologies), Gilad Shainer (Mellanox Technologies)

The next generation of intelligent interconnect devices, such as Mellanox/s Quantum™ HDR 200Gbps InfiniBand Switching with SHARP (Scalable Hierarchical Aggregation and Reduction Protocol) and ConnectX-6 200Gb/s VPI Host Channel Adapters with in-network co-processing, enables communication framework processing and user algorithm processing on the interconnect fabric increasing system performance by an order of magnitude or more. We will explore the features and benefits of the next generation capabilities of these smart-interconnect devices, off-load architecture and in-network co-processing as we prepare to move towards the next milestones and even beyond exascale computing.

Room: D171
1:30 pm - 3:00 pm

HPC in the Cloud

Extending On-Premise HPC to the Cloud

Gabriel Broner (Rescale)

For many years we have used HPC on premises. A large system is shared by multiple users who take advantage of a powerful system. But what happens when HPC in the cloud becomes feasible? What applications can move to the cloud? What advantages and disadvantages does cloud offer? Can the flexibility of cloud be realized for HPC users? How can challenges be overcome? In this talk, Gabriel Broner will cover HPC on premises, HPC in the cloud, and show significant use cases that today are taking advantage of cloud resources. The talk will answer key benefits and challenges of using the cloud for HPC, and will help understand the benefits of a hybrid world, where on premise and cloud will coexist for the foreseeable future.

Enabling HPC and Deep Learning Workloads at Extreme Scale in the Cloud

Bill Bryce (Univa Corporation)

Independent research (Reuther et al., J. Parallel Distrib. Comput., 111, 2018, 76–92) underscores the importance of efficient workload management: "For both supercomputers and big data systems, the efficiency of the job scheduler represents a fundamental limit on the efficiency of the system." However enabling efficiency at extreme scale in the cloud, for workload management or other purposes, requires sophisticated integration and automation that also scales. By deeply integrating with AWS-specific APIs, the capabilities of this public-cloud provider are fully leveraged via Navops Launch in a highly automated fashion. As a compelling proof point, Navops Launch makes routine the scaling of a compute cluster to more than 1,000,000 cores, across 55,000 heterogeneous spot instances spanning three availability zones. As a consequence, in demanding policy-based launching of cloud instances, heroics are no longer required to scale HPC and Deep Learning workloads to the extreme.

What Would You Do with a Million Cores of Compute Capacity?

Ian Colle (Amazon Web Services)

High Performance Computing as always been about solving the most complex problems, but for far oo long HPC applications and workloads have been constrained by infrastructure capacity to allow for n unfettered imagination of problems could be solved if infrastructure capacity were virtually nlimited. I this session, the sp aker will highlight how virtually unlimited capacity and scale, ac cessed instantly on the cloud, can create a paradigm shift in the way researchers and engineers approach innovation. AWS has helped set up up massive HPC clusters that scale to over a million vCPUs, multiple times. In this session, we will highlight the process of setting up extreme-scale HPC clusters in the cloud and the lessons we learned along the way. We will lay out a broad vision of how the steadily increasing interest in running HPC workloads on the cloud, combined with the advances in AI/ML is a catalyst for sustained innovation.

Room: D173
1:30 pm - 3:00 pm

Networking

Faucet: SDN made Easy

Richard Nelson (Faucet Foundation, University of Waikato)

Faucet is a an open-source enterprise style SDN controller. It aims to be as simple to deploy, operate and support with a (reasonably) easy to understand code base. Faucet contains no driver code for specific switches so it enables multi-vendor networking. Faucet currently works on six vendors switches, including Cisco, Allied Telesis and Noviflow. Faucet includes a comprehensive test switch that can be used to confirm both the software and the hardware it is to be deployed to. In combination with standard automation tools this allows dev-ops style continuous integration and deployment.

100G SSL/TLS Decryption Is Indeed Possible for High Capacity Links

Yuri Kolomiyets (Corsa Technology Inc)

Currently it's not possible to decrypt traffic on 100G network links in a single appliance in order to analyze traffic for security exposure. This is problematic for high capacity networks, especially in HPC environments where large data transfers occur. I will introduce a new architectural concept of using a Corsa appliance to horizontally scale out traffic, and effectively load balance 100G traffic, into physical or virtual appliances of lower capacity for SSL/TLS decryption. Traffic is then programmatically service chained through relevant security functions as required before being returned for SSL encryption and continued transit. This is an important use case as more than 75% of internet traffic is encrypted. This talk will dig into the architecture and dissect the various elements required to successfully deploy 100G SSL. It is intended to be interactive, and participants are invited to bring their network security challenge to the session for discussion.

Make Sure the Network Isn't the Problem! 400GE Considerations and Best Practices for Testing the Cluster Fabric

James Low (Keysight Technologies Inc)

400GE by its very nature, introduces new challenges—increased bit error rate, mandatory FEC, and the performance of the PCS in the presence of errors that 56Gb/s SERDES and PAM-4 encoding inherently provide. 400GE is destined to be adopted by the industry at least 2x faster than was 100GE, and is right around the corner for supercomputing clusters. With all new optics and cables, switches, and servers driven by 400/200/100/50/25GE speeds, the need for testing has never been greater. Testing not just by chip and equipment makers, but by independent benchmark and performance test organizations.

The end users—data centers, supercomputing environments, service providers, and network operators—need empirical data on which equipment will satisfy their particular network needs. This is evident in the use of traffic generation to verify SCinet circuits. It's a critical time to renew efforts to compare and benchmark the performance of these new network products and devices.

Room: D171
3:30 pm - 5:00 pm

HPC Workflow

Simplifying HPC Data Management at Scale

John Leonardini (Komprise)

The rapid growth of HPC data necessitates a multi-layer storage strategy to get the best performance and data protection while cutting costs. Since a bulk of data becomes inactive or cold soon after its created, Komprise Intelligent Data Management enables HPC institutions to analyze data across multi-vendor storage and by policy move cold data to lower cost storage and data protection models through its lifecycle, without any changes to user or application access. This talk describes how Komprise uses a distributed scale-out architecture built on open protocols to work across a variety of on premise and cloud storage and transparently preserve the original namespace of data without fronting all data or metadata access, so there is no user disruption yet no lock-in. HPC use cases of how research institutes have leveraged data lifecycle management to cut 70%+ of costs will also be presented.

Puppet in HPC: Building on 10 Years of Practice

Paul Anderson (Puppet)

HPC administrators are a versatile bunch, frequently applying novel solutions to balance the seemingly contradictory requirements of maximizing computational output and minimizing downtime, while servicing the ever more complex requests for configuration customization from their user community, security team, and application developers. With practical examples from Puppet's ten years of working with the HPC community, this presentation suggests both tried-and-true and new, DevOps-inspired approaches that can keep supercomputing manageable as systems grow in scale, complexity, and throughput.

CCIX: Seamless Data Movement for Accelerated Applications

Gaurav Singh (CCIX Consortium, Xilinx Inc)

CCIX or Cache Coherent Interface for Accelerators is a chip-to-chip interconnect architecture specification created to solve the performance and efficiency challenges of emerging acceleration applications such as machine learning, network processing, storage/memory expansion, and analytics that combine processing and acceleration.

The CCIX Consortium has released the production revision of the CCIX Base Specification 1.0 that enables seamless data sharing in heterogeneous compute systems between CPUs, Accelerators and Memory expansion devices. The CCIX specification leverages the PCI Express™ 4.0 architecture and ecosystem while increasing the throughput to 25GT/s per lane.

In this Exhibitor Forum, CCIX Consortium members will discuss the advancements CCIX brings to the system hardware and software architecture, and the use cases that benefit from the cache coherent shared virtual memory paradigm and driverless data movement between processors and accelerators including FPGAs, GPUs, network/storage adapters, intelligent networks and custom ASICs.

Room: D173

3:30 pm - 5:00 pm

HPC Workflow

OpenMP API Version 5.0 - Getting Ready for Exascale

Michael Klemm (OpenMP Architecture Review Board)

Since its creation in 1997, the OpenMP(r) API has become the standard programming model for multi-threading in HPC applications and has enabled many scientific discoveries by making it easy for scientists to exploit the power of modern computers. The OpenMP API uses directives to augment code written in C/C++ and Fortran with parallelization, vectorization, and offload information for the compiler.

OpenMP has become a programming model that supports modern task-based programming as well as heterogeneous programming for offload devices such as GPUs, or, even, FPGAs. In the week before SC18, the OpenMP Architecture Review Board will release Version 5.0 of the OpenMP API Specification. This is a major improvement with more powerful parallelization features for modern multi-threaded applications. Here we will review the new features of version 5.0 of the OpenMP API and describe how they extend the current OpenMP API.

Analytic Based Monitoring of High Performance Computing Applications

William Leinberger (General Dynamics Mission Systems)

The complexity of High Performance Computing (HPC) systems and the innate premium on system efficiency necessitate the use of automated tools to monitor not only system-level health and status, but also job performance. Current vendor-provided and third party monitoring tools, such as Nagios or Ganglia, enable system-level monitoring using features that reflect the state of system resources. None of those tools, however, are designed to determine the health and status of a user's application, or job, as it executes.

This presentation introduces the concept of job-level, analytics-based monitoring using system features external to the job, like those reported by Ganglia. Preliminary results show these features contain sufficient information content to characterize key behaviors of an executing job when incorporated into a job-specific, application-neutral analytic model; transitions between computational phases, onset of a load imbalance, and anomalous activity on a compute node may each be detected using this approach.

Advanced Technologies and Techniques for Debugging HPC Applications

Bill Burns (Rogue Wave Software Inc)

Debugging and analyzing today's HPC applications requires a tool with capabilities and features to support the demands of today's complex HPC applications. Debugging tools must be able to handle the extensive use of C++ templates and the STL, use of many shared libraries, optimized code, code leveraging GPU accelerators and applications constructed with multiple languages.

This interactive session walks through the different advanced technologies provided by the debugger, TotalView for HPC, and shows how they can be used to easily understand complex code and quickly solve difficult problems. Showcasing TotalView's new user interface, you will learn how to leverage

the amazing technology of reverse debugging to replay how your program ran. You will also see how TotalView provides a unified view across applications that utilize Python and C++, debug CUDA applications, find memory leaks in your HPC codes and other powerful techniques for improving the quality of your code.

Wednesday, November 14th

Room: D171
10:30 am - 12:00 pm

Machine Learning and AI

Productive and Performant AI Platforms of the Future
Rangan Sukumar (Cray Inc)

Thus far, contributions to hardware and software tools for advanced analytics and artificial intelligence has come from the commodity/cloud computing community. In this talk, we share exciting results from efforts that ported software frameworks such as Apache Spark and TensorFlow onto high performance computing (HPC) hardware to make the case that HPC-approaches future-proof AI investments. We will demonstrate performance gains from combining a HPC interconnect with algorithmic cleverness using communication collectives for graph analytics, deep learning and matrix methods – all components of the modern data science and enterprise AI workflows. Based on experience from several use-cases, we will argue how HPC futureproofs investments for AI journey – particularly around emerging requirements around (I) non-traditional data (graphs, medical imagery, genomic sequences); (ii) building custom domain-specific models with hyper-parameter learning; (iii) the need for ensemble and model parallelism (iv) latency on edge-devices and training cadence with custom processors.

Experience New Records for Speed and Scale: High Performance Genomics and Imaging
Frank Lee (IBM)

Through real use cases and live demo, Frank Lee, PhD, Global Industry Leader for Healthcare & Life Sciences, and his extended team, will illustrate the reference architecture and solution for high performance data and AI platforms (HPDA), deployed for cloud-scale data management, multi-cloud workload orchestration and converged high performance computing with deep learning.

This presentation will demonstrate the four key values of an HPDA architecture - high performance, low cost, ease of use, and collaboration - based on the following use cases of: the fastest and fully-automated clinical genomics pipeline and the world's largest genomics database that can ingest and annotate 10 million variants per hour. Also, high-performance deep learning and AI for cancer and medical imaging, data ocean for genomics that spans multiple public clouds, metadata and provenance tracking for complex analytical workflow, and much more …

Accelerate Machine Learning with High Performance Memory
Mark Hur (Micron Technology Inc)

Machine learning is driving the next industrial revolution, through the use of easy to program accelerator boards, you can be in the driver seat. In this talk, Micron will discuss a new novel inference engine architecture that provides the benefits of hardware acceleration, that supports a wide range of ML frameworks and networks; and is software programmable. This new architecture/platform is a flexible and programmable neural network accelerator co-processor is designed for 1) maximum hardware utilization, 2) efficient memory bandwidth usage, 3) ultra-low power operation. This architecture coupled with a ML complier allows the user to take advantage of hardware acceleration while maintaining software programmability.

Room: D173
10:30 am - 12:00 pm

HPC System Architectures

Bringing Innovation to the HPC Market with Marvell's ThunderX2 Processors
Larry Wikelius (Marvell Technology Group LTD)

Marvell has established a clear leadership position in bringing Arm® -based compute solutions to the High Performance Computing and data center market. This year at SC18, Marvell will provide exciting updates on our product strategy and in particular we will provide more details on our ThunderX2 processor family.

OpenCAPI: High Performance, Host-Agnostic, Coherent Accelerator Architecture and Ecosystem
Brian Allison (OpenCAPI Consortium)

The Open Coherent Accelerator Processor Interface (OpenCAPI) is an industry standard architecture targeted for emerging accelerator solutions and workloads. OpenCAPI enables the development of host-agnostic devices which can coherently connect to any CPU which supports the OpenCAPI standard, providing the device with the capability to coherently cache host memory to facilitate faster accelerator execution. OpenCAPI utilizes high-frequency differential signaling technology while providing high bandwidth in addition to low latency needed by advanced accelerators. OpenCAPI encapsulates the serializing cache access and address translation constructs in high speed host silicon technology to minimize SW path length and design complexity in attached acceleration silicon such as FPGAs and foundry ASICs. OpenCAPI architecturally ties together Transaction Layer, Data Link Layer, and Physical Layer attributes to enable high bandwidth, highly parallel exploitation of attached silicon. Finally, this session will also cover all the OpenCAPI enabled servers as well as the ecosystem that surrounds OpenCAPI.

High Performance OpenMP for GPUs
Michael Wolfe (Nvidia Corporation, PGI Compilers and Tools)

OpenMP has a 20 year history in HPC. It was the parallel programming model of choice for large SMP servers, and later proved effective for hybrid MPI+OpenMP programming on scalable systems. As core counts rose, best practices for OpenMP implementations and programming have evolved. The early focus was on maximizing efficiency of loop-level parallelism. More recently a global, outer-parallel approach has been used to improve scalability on nodes with large numbers of CPU cores. GPUs are

optimized to deliver high performance and throughput on massively parallel, regular, loop-structured algorithms. In this talk, we'll survey the features of the latest OpenMP standards, describe how OpenMP programs can exploit the performance potential of today's GPUs, and summarize the implications for a high-performance OpenMP implementation that supports both multicore CPUs and GPU accelerators.

Room: D173
1:30 pm - 3:00 pm

Machine Learning and AI

Managing the Convergence of HPC and AI

J.J. Falkanger (Lenovo)

HPC environments are the most significant source of processing capacity in many organizations, and more users want to leverage the power of the "SuperComputer" for their workloads to get performance beyond the single box. These "new customers" for your HPC cluster may have little knowledge on how to access, configure and deploy workloads where typical open-source cluster management solutions are used, driving a significant amount of handholding for administrators. In particular, users wanting to experiment with or deploy AI training on the cluster may simply have data and a desire to train, without the technical expertise to configure scripts, resources, frameworks, libraries, etc. to run. Bringing these new users into the HPC environment is a significant opportunity to grow and expand the value of your infrastructure – but only if it is easy to use and easy to manage, and consistent for both HPC and AI workloads.

Data Protection Solutions for ML/AI

Allison Armstrong (Igneous Systems Inc), David Clements (Igneous Systems Inc)

Igneous has customers building artificial intelligence models that revolutionize the diagnosis and treatment of cancer and other life-threatening diseases. Their machine learning workflows utilize a proprietary storage solution as an archive tier, working in conjunction with high performance storage and computing.

Igneous acts as the central repository for petabytes of imaging data. Small subsets of data (< 2% of the entire dataset) are active at any one time for high performance compute, which acts as a "hot edge" for the data to be processed by image processing software running on a high-performance deep learning platform. In addition, Igneous acts as the central repository to archive all computational results, enabling the "hot edge" to be cleared for subsequent workloads. Delivered as-a-Service and remotely managed, Igneous enables organizations to keep their IT departments lean so that they can focus on groundbreaking research instead.

Cassandra in Dockers Deployment Using an NVMe Fabric

Kais Belgaied (Newisys), David Paulsen (Newisys)

The Cassandra distributed database achieves high throughput and large capacity at an unparalleled resiliency by scaling linearly to a large number of nodes, while keeping multiple copies of the data. This

comes at the price of increased latency and inefficient use of CPU and storage. Newisys has developed an approach whereby storage is decoupled from the nodes that run the database by deploying a Cassandra Cluster over dockers that is connected to NVMeoF and served by the Newisys NSS-2248 NVME target appliance. We achieve significantly lower latency by using NVMe and reducing the replication factors. The decoupling of the volumes from the database nodes enables much more efficient use of CPU resources by allowing multiple instances to share CPU and memory resources on the same servers, while improving the overall flexibility and resiliency of the system.

Room: D173
3:30 pm - 5:00 pm

Datacenter and Cooling Technologies

The Case for Thermoplastic Quick Disconnects in Liquid Cooling

Kristin Anderson (Colder Products Company)

Liquid cooling system components must meet requirements for chemical compatibility, flow rates, temperature/pressure exposures. Ease of use and reliability over long periods of time are additional desirable attributes. To avoid potential damage to expensive electronic equipment due to leaks, secure drip-free connections are essential for any liquid cooling system.

Quick disconnects (QDs) designed and built specifically for HPC applications are now available, simplifying connector selection. However, questions surface around component materials – metal or thermoplastics and their ability to meet HPC thermal requirements. Intended use, ease of installation and maintenance, operating conditions, durability, and potential for leakage have bearing on the selection process.

This presentation describes the performance and reliability factors of thermoplastics for fluid handling. The differences between thermoplastic and metal fluid connectors are discussed. Special focus is given to sealability, integrity over time, condensation/corrosion, chemical compatibility and weight. The presentation highlights CPC's just-announced thermoplastic quick disconnect products.

Specifying Rack Level High Density Liquid Cooling Solutions

Michael Wood (nVent)

The higher the density of heat generating equipment, the higher the risk of any failure resulting from inadequate rack, row and room level cooling. Presenting rack level cooling design considerations and methods for managing high heat loads in HPC applications. Passive or active rear door solutions offer an efficient way of cooling up to 45kW per rack. Whereas, the hot exhaust air from the rack-mounted equipment passes through a passive door mounted cooling coil and water circuit – an integrated fan assembly delivers additional CFM in active cooling configurations. When optimizing the heat transfer, the coil layout and optimized air/water circuit temperatures can improve cooling performance. Some HPC configurations can generate even higher heat loads and may require high-density liquid-to-liquid cooling as an ideal alternative – both chip and system level. Additionally, it is essential to deploy reliable monitoring of working parameters as well as leak detection along the pipework.

Taming Datacenter Thermodynamics with Lenovo Neptune Technology
Matthew Ziegler (Lenovo), Vinod Kamath (Lenovo)

Lenovo has been delivering solutions based on liquid-cooling since 2012 and installed the first x86-based warm-water (50°C intake) cooled cluster at Leibniz-Rechenzentrum (LRZ) in Munich, Germany. The latest generation of technology based on direct to node (DTN) liquid cooling, the Lenovo ThinkSystem SD650 that directly cool the CPUs (up to 240W), IO, RAM, HDDs and voltage regulators. Another liquid-cooling based approach Lenovo uses is a novel heat-sink which uses water to move heat away from a processor via a hermetically sealed heat pipe located inside a wide rear heat spreader. This Thermal Transfer Module (TTM) allows for servers to run processors up to 205W in traditional dense form-factor. Finally, Rear Door Heat Exchangers (RHDX) extract server heat from the exhaust air before entering the datacenter through liquid passing through the rear rack door. These three technologies make up the Lenovo Neptune family of liquid-cooled solutions.

Thursday, November 15th

Room: D171
10:30 am - 12:00 pm

HPC System Architectures

Post-K Supercomputer with Fujitsu's Original CPU, Powered by Arm ISA
Toshiyuki Shimizu (Fujitsu Laboratories Ltd)

Fujitsu has provided supercomputers for over forty years; recently the K computer in 2011, winning 1st place twice in the TOP500, the Fujitsu-original PRIMEHPC FX10 in 2012 and the PRIMEHPC FX100 in 2015. Fujitsu also delivers x86 clusters for wide application areas and customer requirements.

RIKEN and Fujitsu are now developing the “post-K” supercomputer as Japan’s new national supercomputer project. The post-K inherits the K computer’s architecture and targets up to 100 times higher application performance with superior power efficiency. The post-K employs the newly developed Fujitsu original CPU which recently announced as FUJITSU A64FX featuring the Armv8-A and SVE ISA and widening application opportunities. The post-K contributes to the Arm ecosystem for HPC applications as well as science and society.

At SC18, Fujitsu will provide updates on post-K development and some performance evaluation results.

Next Generation Vector Supercomputer
Masashi Ikuta (NEC Corporation)

The NEC SX-Aurora TSUBASA is the newest in the line of NEC SX Vector Processors with the world's highest memory bandwidth. The Processor that is implemented in a PCI-e formfactor can be configured in many flexible configurations together with a standard x86 cluster. No special framework is needed to build the applications, but standard programming paradigms in C/C++ and Fortran, for

which the NEC compiler will automatically carry out vectorization and parallelization.

The target is a higher sustained performance with minimum effort by utilizing the Vector Engine hardware.

Seeking Quantum Supremacy with Numerical Simulation
Cyril ALLOUCHE (Atos, Bull)

Quantum Computing promises to tackle computations which are intractable for today's supercomputers. Among the expected beneficiaries: numerical analysis, artificial intelligence, chemistry… Though, most of the known quantum algorithms underlying these applications require *perfect* qubits. That is fully connected, unlimited coherence time, perfect fidelity. Designing such ideal qubits is a tremendous challenge for Science and Technology, and could require a decade or more. In the meantime, the "NISQ" era (Noisy-Intermediate-Scale-Quantum) is coming, with devices with 100+ *imperfect* qubits. To take the most of these intermediate hardwares, brand new algorithms must be invented, as noise-tolerant and resource-efficient as possible. Numerical modeling and simulation of these imperfect qubits can provide an invaluable tool to take on this NISQ challenge. The Atos Quantum Learning Machine (QLM) embeds the state of the art of these techniques. I will introduce, and live-demo them during the talk, on a real and self-illustrating use case.

Room: D173
10:30 am - 12:00 pm

HPC Storage and Memory Architectures

Next Generation NVMe-Native Parallel File System Accelerating AI Workloads
Liran Zvivbel (WekaIO Ltd)

Any IT infrastructure that supports an AI and analytics workflow needs to be able to handle a large number of files, and large amounts of data storage with high throughput access to all the data. Legacy file systems can't supply high throughput and high file IOPS, as they were designed for HDD and are not suitable for low latency, small file IO and metadata heavy workloads that are common in AI and analytics. This results in I/O starvation, a major problem for an AI system. We'll introduce a next-gen massively parallel shared file system that is NAND FLASH and NVMe-optimized, which is solving the I/O starvation problem. What you will learn: - Case study: how a large autonomous driving car manufacturer maximizes their investment in GPUs by ensuring they are saturated with data - How to make I/O compute bound again - How to maximize data accessibility for your cluster

Enabling High Performance Memory for the Broad HPC markets
Andreas Schlapka (Micron Technology Inc)

High performance compute demands advanced memory solutions. Designing for maximum performance has traditionally carried the burden of substantial investment in cost and time to implement "in package memory on silicon substrates". Early in 2018 Micron has introduced new technology being adopted across markets for performance solutions. Memory performance required for applications like networking routing and switching, autonomous driving, Crypto mining, Gaming

Consoles, Graphics cards, and HPC acceleration. In this discussion, Micron will outline packaged memory solutions, with bandwidth required for HPC applications with the ecosystem enablement and adoption deployed in production. For the HPC community, this is provide an easy implementation of high performance solutions.

Architecture of a Next-Generation Object Storage Device in the Panasas Filesystem
Ellis H. Wilson (Panasas Inc.)

The fundamental storage component of the scale-out Panasas Filesystem (PanFS) is an Object Storage Device (OSD). After 17 years of incremental improvements to the hardware and software of their OSD, Panasas unveils its completely re-engineered, next-generation OSD and associated OSD Filesystem (OSDFS). It is designed to handle the extreme performance, reliability, and cost-efficiency requirements of exascale HPC and the parallel storage systems that must back it, and lays the foundation for future innovations.

In this presentation, we discuss the unique requirements of a parallel filesystem OSDFS that drove our original design, how evolving workload characteristics necessitated changes to hardware and software, the need for portable software, and key details of the new OSD architecture that is being announced and demonstrated at SC'18. We conclude with real-world performance analysis of our next-generation OSD and discuss its deep contributions in the near- and long-term to the overall Panasas performance and features roadmaps.

Exhibits

(back to top)

Monday, November 12th

Room: Exhibit Hall C, D, E and F
7:00 pm - 9:00 pm

Gala Opening Reception

Gala Opening Reception
SC18 will host its annual Grand Opening Gala in the Exhibit Hall. This will be your first opportunity to see the latest high performance computing, networking, storage, analysis, and research products, services, and innovations. This event is open to all Technical Program, Exhibitors and Students@SC registrants.

Tuesday, November 13th

Room: Exhibit Hall C, D, E and F
10:00 am - 6:00 pm

Exhibits

Exhibits
Come visit with over 300 exhibitors that range from industry to academia and research. There is sure to be something for everyone.

Wednesday, November 14th

Room: Exhibit Hall C, D, E and F
10:00 am - 6:00 pm

Exhibits

Exhibits
Come visit with over 300 exhibitors that range from industry to academia and research. There is sure to be something for everyone.

Room: Exhibit Hall C, D, E and F
4:00 pm - 6:00 pm

Family Day

Family Day
On Wednesday, November 14, from 4:00pm to 6:00pm, families are invited to walk through the Exhibit Hall. Everyone will be required to have a badge and children (12 years and older) must be accompanied by an adult at all times. If you need a badge for a family member, go to the Attendee Registration area.

Thursday, November 15th

Room: Exhibit Hall C, D, E and F
10:00 am - 3:00 pm

Exhibits

Exhibits
Come visit with over 300 exhibitors that range from industry to academia and research. There is sure to be something for everyone.

Experiencing HPC for Undergraduates

(back to top)

Monday, November 12th

Room: D169
2:00 pm - 3:00 pm

Experiencing HPC for Undergraduates Opening Reception

Session Description: This is a required event for all students attending the Experiencing HPC for Undergraduates program. The opening reception provides an opportunity to meet the other student and will provide an overview to the program and conference in general.

Room: D169
3:00 pm - 5:00 pm

Experiencing HPC for Undergraduates Orientation

Session Description: This is a required event for all students attending the Experiencing HPC for Undergraduates program. The orientation will provide an overview of the topics covered at SC, events at the conference, and general program.

Tuesday, November 13th

Room: D169
10:30 am - 12:00 pm

Experiencing HPC for Undergraduates: Introduction to HPC Research

Session Description: This panel will provide an overview of HPC research from parallel computing from a computer science perspective to applications to data science and visualization. Bill Gropp from UIUC, Aydin Buluc from LBNL, and Venkatram Vishwanath from ANL will present shortly and serve on a panel for interactive discussion after. This is a required event for all students attending the Experiencing HPC for Undergraduates program.

Wednesday, November 14th

Room: D169
10:30 am - 12:00 pm

Experiencing HPC for Undergraduates: Graduate Student Perspective

Session Description: This panel will provide an introduction to graduate school from the student perspective. A panel of current students will discuss the application process, how best to prepare during the undergraduate program, how to pick a program, and other related topics to graduate school. This is a required event for all students attending the Experiencing HPC for Undergraduates program.

Thursday, November 15th

Room: D169
10:30 am - 12:00 pm

Experiencing HPC for Undergraduates: Careers in HPC

Session Description: This will be a panel discussing different career paths in parallel computing. Rick Arthur from GE will be discussing roles in industry, Tanzima Islam from Western Washington University will talk about academia, and Judith Hill from ORNL to talk about national laboratory options. This is a required event for all students attending the Experiencing HPC for Undergraduates program.

HPC Impact Showcase

(back to top)

Wednesday, November 14th

Room: D171
1:30 pm - 3:00 pm

HPC Impact at TAE Technologies and Pratt & Whitney

Data Fusion for Nuclear Fusion – Using HPC To Put a Star in a Bottle *Sean Dettrick (TAE Technologies)*
Fusion energy offers the prospect of a carbon-neutral, environmentally responsible, and inexhaustible energy source. TAE Technologies Inc. is trying to greatly accelerate fusion research to develop the world's first commercially viable fusion-powered generator for electricity production.

To this end, TAE has invested $100 million of capital expenditure in the construction of its latest magnetically confined fusion experiment, "Norman," an advanced beam-driven field-reversed configuration (FRC) plasma device. A central challenge of understanding the physics in a fusion plasma experiment is that most of the experimental diagnostics are indirect in nature and require inverse problems, such as tomographic inversion, and that there are many interacting degrees of freedom each of which requires its own diagnostic and inversion process. To understand the "what" of the complete plasma state measured in this way requires data science, in particular, sensor fusion and Bayesian Inference. To understand the "why" of the plasma state requires theory and advanced computation, which is made challenging by multiple time and space scales and multi-physics interactions. The central problem of keeping the plasma hot enough for long enough to achieve fusion cannot be addressed without using HPC to understand non-linear wave-particle interactions, which can bring both great benefit in terms of kinetic stabilization of modes at the macroscale, and detriments in the form of heat losses due to kinetic microturbulence.

Analysis of existing experiments, and prediction of performance in new parameter regimes, requires a fusion of the "what" and the "why" with a combination of data science and HPC modeling. TAE is partnering with data science heavyweights – Google and others – and big iron heavyweights such as the Department of Energy Leadership Computing Facilities – to bring the most advanced data science algorithms and the fastest computers to bear on these problems.

High Performance Computing in the Cloud at United Technologies *Pete Bradley (United Technologies Corporation - Pratt & Whitney Division)*

Room: D171
3:30 pm - 5:00 pm

HPC Impact at Procter & Gamble, Boeing and GE

HPC Enables Simulation-Led Innovation in Places You Would Not Expect *Miguel Palacios (Procter and Gamble Company)*
When people think of High Performance Computing, visions of simulating the motion of thousands of stars and planets or jet engines come to mind. While these are indeed HPC applications, simulation-led innovation is now occurring in a wide range of industrial and consumer product applications. These range from predicting phase chemistry of a detergent to strength of a package, to how skin moisturizers actually work to make skin beautiful. Computationally intensive simulation is the perfect link between incomplete theory and hard to run experiments and leads to important insights that in turn lead to innovations that billions of people use every day

The Enterprise HPC Service at Boeing *Joerg Gablonsky (Boeing)*
HPC and simulations are a critical enablers to the design of Boeing's products. In this presentation we will talk about the characteristics of a successful Enterprise HPC Service that provides HPC capabilities to a diverse range of users and organizations. We will also discuss some of the business reasons behind the need for ever increasing HPC capabilities and how HPC can be linked directly to significant savings for the company.

Computationally-Accelerated Engineering at GE: Physics + Deep Learning *Richard B. Arthur (General Electric Company)*
In support of a global team of engineers working on designing, building and servicing complex machines such as jet engines and gas turbines, GE Global Research hosts a centralized HPC environment for computational modeling, simulation and analytics. GE's Digital Thread for Design (DT4D) framework revolutionizes product engineering by harnessing data services, automation and machine learning to advance multi-disciplinary modeling and simulation productivity and capability. Engineers can offload previously labor-intensive, manual, and ad-hoc tasks to focus on improving product performance and reliability. Through collaborations between domain experts and data scientists, we introduce "Deep Physics" – blending traditional physics models and knowledge with the emerging power of deep learning to enable deeper exploration of parametric trade-off spaces and reshape the state of the art.

Thursday, November 15th

Room: D171
1:30 pm - 3:00 pm

HPC Impact at BP and Lockheed Martin

The BP Data Science Sandbox *Max Grossman (BP), Anar Yusifov (BP)*
Recent years have seen major advances in the state-of-the-art of machine learning, particularly in fields such as natural language processing and 2D computer vision.

These advances have naturally spurred interest in the application of similar techniques to new fields in medicine, science, and engineering. However, the problems in these fields are differentiated from previous machine learning successes by the level of domain expertise required. While classifying an image as a cat, dog, horse, etc is a task that anyone can understand, automatic identification of malignant tumors, subsurface faults, or financial fraud (for example) often requires far more

background in the specific domain. Unfortunately, it is rare today for people to have both the skills of a data scientist/statistician and a domain expert (e.g. an oncologist or petroleum engineer). This problem can generally be solved in two ways: (1) through education (of your data scientists and/or domain experts), or (2) through co-location of these two groups of people such that they can work closely together.

This talk will introduce the BP Data Science Sandbox (DSS) – an internal environment at BP that supports both of the above solutions. The sandbox is a platform made up of hardware, software, and people. On the hardware front, the sandbox includes everything from big memory machines to GPU machines to compute clusters, enabling users of the sandbox to pick and choose the platform that meets their resource requirements. On the software front, the sandbox is built on entirely free and open source software, including common tools such as Jupyter, JupyterHub, Spark, Dask, Tensorflow, and other packages in the Conda ecosystem. On the people front, the sandbox is supported by a team of dedicated data scientists and infrastructure engineers who support users and internal customers of the sandbox.

Challenges and Solutions in the Industrial Application of HPC *Douglas D. Cline (Lockheed Martin Aeronautics Company)*
Industrial support for High Performance Computing in the development of engineered products presents unique challenges when compared to government laboratories and universities. Some of these challenges include supporting a diverse user community, security compliance, and coordination of HPC tasks across multiple organizational boundaries. In this presentation, we will describe some of Lockheed Martin Aeronautics' unique solutions to these challenges including HPC system architecture for security, unique discipline-specific architectures and overall HPC requirements in an industrial environment.

Room: D171
3:30 pm - 5:00 pm

HPC Impact at GM and John Deere

HPC Drives GM *Ajay Deshpande (General Motors Company)*
Over the past decade, Computer Aided Engineering (CAE) has increased its role in vehicle development. CAE is not just a set of tools to provide physical understanding; CAE-based designs and optimizations continue to provide integral input into product development. Accurate modelling and characteristics in CAE is a key step to establish robust math-based simulations for design and development. For example, in internal combustion engines, the mixture preparation process is highly coupled with combustion performance and emissions behavior. The presentation will outline how HPC is enabling engineers and researchers for experimental and modelling work, undertaken at General Motors to support the corporate Road-to-Lab-to-Math (RLM) strategy.

Emergence of Tools - a Competitive Advantage at John Deere *Mohamad S. El-Zein (Deere & Company)*
The trends in the past 10-15 years are to drive bigger models and brute-force computing, hence, HPC flourished internally and externally in many enterprises. As the technology keeps advancing and business cases become clearer, a new trend is becoming more lucrative: combining HPC, data science,

machining learning, and IoT. A model of how to use all of the aforementioned will be presented for an industrial setting.

HPC Inspires Plenary

(back to top)

Monday, November 12th

Room: Exhibit Hall B
5:30 pm - 6:45 pm

HPC Inspires Plenary: HPC and AI: Helping to Solve Humanity's Grand Challenges

Evan Fraser (Arrell Food Institute, Global Food Security), Robert S. Hart (Global Good, Institute for Disease Modeling at Intellectual Ventures), Marguerite Nyhan (United Nations Global Pulse)

From solving the global food security crisis, to preventing epidemics and understanding the impact of environmental health on our urban centers; high performance computing is revolutionizing how we address and manage global crises.

Simulation and modeling along with AI are being applied to some of our most challenging global threats and humanitarian crises. The SC18 plenary session will explore the capacity of HPC to help mitigate human suffering and advance our capacity to protect the most vulnerable, identify methods to produce enough food globally and ensure the safety of our planet and natural resources.

We encourage all HPC experts to join the discussion and be inspired to apply their expertise to these real world grand challenges:

* Over the next two generations, we face an enormous human security challenge. We must adapt to rapid economic and climate change by creating a food system that provides adequate and appropriate nutrition for 9 billion people in a way that does not compromise the environment.

* Infectious disease poses an ongoing threat to human populations, exacerbated by migration, urbanization and globalization dynamics. Many of the world's epidemic diseases are known to be sensitive to climate. It is critical we have the ability to build accurate disease forecasting models for prevention.

* Urbanization and population growth dynamics pose significant challenges to human health and safety. Analytical models of the interrelationship between transportation, air quality and other environmental conditions enable us to better understand risk but more importantly advance policies to mitigate impacts on the most vulnerable populations.

Our panelists include:

* Dr. Evan Fraser, Director of Arrell Food Institute and Canada Research Chair in Global Food Security. He leads a $100M program geared at developing the tools to promote the "digital agricultural revolution." This involves work not only with technologists, but also policy makers, economists, politicians and private industry.

* Robert S. Hart, Ph.D., Vice President of Global Good and General Manager of the Institute for Disease Modeling at Intellectual Ventures. IDM's mission is to guide global efforts towards the eradication and control of infectious disease through the use and promotion of quantitative analysis. IDM has grown from a small project focused on malaria eradication to an independent center shaping strategies and building innovative tools to accelerate the eradication of infectious disease by the global health community.

* Dr. Marguerite Nyhan, Research Scientist at United Nations Global Pulse. Her work is focused on data science for humanitarian and sustainable development efforts, particularly in the area of environmental health.

Invited Talk

(back to top)

Tuesday, November 13th

Room: Exhibit Hall B
10:30 am - 12:00 pm

Invited Talk Session 1

The Age of Data - Visualizing the Revolution
Chris Johnson (University of Utah)

We live in the Age of Data. 90% of all data in the world has been created in the past two years at a rate of exabytes per day. New data is growing exponentially in every way: structured, unstructured, quantitative, qualitative, spatial, temporal. One of our greatest scientific challenges is to effectively understand and make use of the vast amount of data being produced. Visual data analysis will be among our most important tools to understand such large-scale complex data. Visualization is useful for detecting patterns, assessing situations, and prioritizing tasks. Visualization facilitates the reasoning process by supporting the human capacity to perceive, understand, and reason about complex large-scale data and enables researchers to derive knowledge from data. In this talk, I will present state-of-the-art visual analysis techniques, insights and examples of how visualization can enable understanding in the Age of Data.

Brain-Inspired Massively-Parallel Computing
Steve Furber (University of Manchester)

Neuromorphic computing, that is, computing based upon brain-like principles - can be traced back to the pioneering work of Carver Mead in the 1980s. Academic research into neuromorphic systems has continued since then in various forms, including analog, digital and hybrid systems, primarily with the objective of improving understanding of information processing in the brain. More recently, industrial neuromorphic systems have emerged - first the IBM TrueNorth, and then the Intel Loihi - with a greater focus on practical applications. In parallel, the last decade has seen an explosion of interest in less brain-like, though still brain-inspired, artificial neural networks in machine learning applications that have, for example, placed high-quality speech recognition systems into everyday consumer use. However, these artificial neural networks consume significant computer and electrical power, particularly during training, and there is strong interest in bringing these requirements down and in enabling continuous on-line learning to take place in self-contained, mobile configurations. There is a growing expectation, so far unsubstantiated by compelling evidence, that neuromorphic technologies will have a role to play in delivering these efficiency gains. The SpiNNaker (Spiking Neural Network Architecture) platform is an example of a highly flexible digital neuromorphic platform, based upon a massively-parallel configuration of small processors with a bespoke interconnect fabric designed to support the very high connectivity of biological neural nets in real-time models. Although designed primarily to support brain science, it can also be used to explore more applications-oriented research.

Room: Exhibit Hall B
3:30 pm - 5:00 pm

Invited Talk Session 2

What Is the Role of Architecture and Software Researchers in Making Quantum Computing Practical?
Margaret Martonosi (Princeton University)

In the past 3-5 years, Quantum Computing (QC) has reached an interesting and important inflection point. For decades, quantum computing research was comprised of very abstract mathematical algorithms development "up high" that demonstrated some potential for future impact, and physics device demonstrations "down low" that were modest in size but that offered some hope for eventual implementations. However, with prominent QC algorithms like Shor's factoring algorithm needing roughly a million times more physical quantum bits (qubits) than successful implementations currently provide, there has been a cavernous gap between algorithm and implementation. What is needed now are computer scientists to develop the crucial intermediate tool flows, abstraction layers, and programming languages that will help quantum computing scale through the current so-called NISQ (noisy, intermediate-scale quantum) era. My talk will both (i) give details about our new approaches for optimal and near-optimal spatial-temporal placement of QC algorithms onto real systems, and (ii) more broadly advocate for the role that computer architecture, compiler and programming languages researchers must play in order for QC to reach its full potential.

Applying Deep Learning
Bryan Catanzaro (Nvidia Corporation)

I will discuss how we go about applying deep learning to our work at NVIDIA, solving problems in a variety of domains from graphics and vision to text and speech. I'll discuss the properties of successful machine learning applications, the criteria that we use to choose projects, and things to watch out for while creating new deep learning applications. I'll discuss the role of HPC in helping us conduct our research, and I'll show some examples of projects that benefit from scale.

Wednesday, November 14th

Room: Exhibit Hall B
10:30 am - 12:00 pm

Invited Talk Session 3

Delivering on the Exascale Computing Project Mission for the US Department of Energy
Doug Kothe (Oak Ridge National Laboratory)

The vision of the US Department of Energy (DOE) Exascale Computing Project (ECP), initiated in 2016 as a formal DOE project executing through 2022, is to accelerate innovation with exascale simulation and data science solutions that enhance U.S. economic competitiveness, strengthen our national security, and change our quality of life. ECP's mission is to deliver exascale-ready applications and

solutions that address currently intractable problems of strategic importance and national interest; create and deploy an expanded and vertically integrated software stack on DOE HPC exascale and pre-exascale systems, thereby defining the enduring US exascale ecosystem; and leverage U.S. HPC vendor research activities and products into DOE HPC exascale systems. The project is a joint effort of two DOE programs: the Office of Science Advanced Scientific Computing Research Program and the National Nuclear Security Administration Advanced Simulation and Computing Program. ECP's RD&D activities, which encompass the development of applications, software technologies, and hardware technologies and architectures, is carried out by over 100 small teams of scientists and engineers from the DOE national laboratories, universities, and industry.

China's Effort on Exascale Computing: Current Status and Perspectives
Depei Qian (Beihang University, Sun Yat-sen University)

After a brief introduction to the key project on HPC under the national key R&D program, this talk focuses on the progress made by the key project, including the exascale prototype system development, the HPC applications and enabling platforms and technologies, and the national HPC environment together with the services provided. Major issues on exascale system development and the HPC application ecosystem will be discussed. The perspectives of HPC in China, especially connection of HPC with BD and AI and the possible converged development of HPC, BD and AI are discussed toward the end of the talk.

Room: Exhibit Hall B
3:30 pm - 5:00 pm

Invited Talk Session 4

ABCI – Envisioning High Performance Human-Centered AI for Industry
Satoshi Sekiguchi (National Institute of Advanced Industrial Science and Technology (AIST))

Artificial intelligence (AI) is leading the world's industrial revolution in relation to IoT, Big Data, Robot, advanced biotechnology, new manufacturing industry, and so on. Although there is an ongoing discussion of the AI singularity, there is a more immediate need to focus on practical applications of AI. In these applications, AI can be defined as sensors coupled with deep learning, which can substitute for some human sensory functions.

Thus, AI will relieve workers from the burden of some physically or mentally intensive tasks. AI algorithmic techniques have developed dramatically with a high dependence on big data sources as training data. More importantly, at a a practical level, I will highlight contributions from HPC technology that are indispensable for learning. Until now, we have supported AI as a system, but cooperation between HPC applications and machine learning will intensify and become a very important factor in the future of AI.

AIST is constructing one of the world's largest AI focused, open, and public computing infrastructure called ABCI (AI Bridging Cloud Infrastructure). It will support not only industrial AI applications to be deployed for use in society, but also academic AI research. To facilitate this, we are connecting ABCI to a global, high-speed network research platform to enhance AI research with a broad number of

institutions. One of the first grand challenge projects we are supporting is mapping whole brain-scale signal integration events in zebrafish. This will open the possibility of creating novel, neural deep learning architectures of the fish brain to be used in a wide variety of applications including robotics, sensor design, and behavioral sciences.

ABCI will be operational shortly in 2018 and will serve as a reference architecture for data centers combining HPC and AI to discover new innovations in the future.

Urban Energy Science and High Performance Computing
Mary Ann Piette (Lawrence Berkeley National Laboratory)

This presentation will cover current research to evaluate how to reduce energy use in cities and report on progress of exascale computing funded by the US Department of Energy. Several national laboratories are conducting this work. Cities in the US consume 70% of total energy use and drive economic activities. Research to model energy in cities includes the use of common data standards to describe the physical systems in cities as well as building energy models and how buildings influence the local climate. The presentation will review the status of energy models for city energy use and how cities are using these models. The talk will describe how these models are being integrated to include transportation, buildings, human activity, and weather. Progress on task grouping to reduce computing resources will also be reported.

Thursday, November 15th

Room: Exhibit Hall B
8:35 am - 10:05 am

Invited Talk Session 5

A Quantum Future of Computation
Matthias Troyer (Microsoft Corporation)

Still in early development, quantum computing is already overturning our contemporary notions of computational methods and devices. Using new concepts of computing based in quantum physics, quantum computers will be able to solve certain problems that are completely intractable on any imaginable classical computer, such as accurate simulations of molecules and materials, or breaking public key encryption. While this potential is real, quantum computers are best viewed as special purpose accelerators for specific problem classes, but they also have their limitations. Attempting to bring clarity to the fast growing field of quantum computing, I will describe the hardware and software architecture of quantum computers and discuss how they differ from conventional classical high performance computers. Based on this, I will attempt to dispel myths and hype surrounding the field and present a realistic assessment of the potential of these devices, and the specific application areas on which they are expected to have a large impact.

The Human Side of Data Science
Cecilia Aragon (University of Washington)

Extraordinary advances in our ability to acquire and generate data are transforming the fundamental nature of discovery across domains. Much of the research in the field of data science has focused on automated methods of analyzing data such as machine learning and new database techniques. However, the human aspects of data science, including how to maximize scientific creativity and human insight, how to address ethical concerns, and the consideration of societal impacts, are vital to the future of data science. Human-centered data science is a necessary part of the success of 21st century discovery. I will discuss promising research in this area, describe ongoing initiatives at the University of Washington in Seattle, and speculate upon future directions for data science.

Room: Exhibit Hall B
10:30 am - 12:00 pm

Invited Talk Session 6

Artificial Intelligence at the Edge: How the Internet of Things and HPC Connect in the Computing Continuum
Pete Beckman (Argonne National Laboratory)

The number of network-connected devices (sensors, actuators, instruments, computers, and data stores) now substantially exceeds the number of humans on this planet. Billions of things that sense, think, and act are connected to a planet-spanning network of cloud and high-performance computing centers that contain more computers than the entire Internet did just a few years ago. Parallel computation and machine learning are providing the basis for this new computing continuum that analyses data in-situ, and uses HPC to model, predict, and learn. This new paradigm is giving rise to intelligent cities, smart agriculture, and advanced manufacturing. The Amazon Inc. DeepLens system is an example of this new model that links machine learning at the edge to the cloud. Another example is the Waggle Platform, developed at Argonne National Laboratory. The Array of Things project at the University of Chicago is deploying hundreds of Waggle-based nodes with advanced wireless sensors in Chicago and other cities. Each of the nodes support parallel edge computing. This presentation will explore the computing continuum, and how artificial intelligence at the edge is now firmly connected to supercomputing.

Superscaling Performance through Energy-Efficient Supercomputing
Padma Raghavan (Vanderbilt University)

In June 2008, Roadrunner sped past the 1 petaflop/s milestone, clocking in at 1.026 petaflops/s. Ten short years later in June 2018, Summit has now taken over the as the world's fastest supercomputer spot by achieving 122.3 petaflops/s. What is remarkable is that the over 100x scaling up of performance was achieved with less than 4x scaling of hardware power. In this talk, I will highlight the key advances in energy-efficiency that made possible this superscaling of performance with a particular focus on the influence of linear algebra and graph algorithms. I will also speculate on trajectories for their continued development to meet the demands of data centric computing as we face slow-downs in reducing the power consumption of transistors.

Keynote

(back to top)

Tuesday, November 13th

Room: Exhibit Hall B
8:30 am - 10:00 am

Keynote: Explore How to Deploy the Unruly Power of Machine, Platform, and Crowd

Erik Brynjolfsson (Massachusetts Institute of Technology)

Erik Brynjolfsson of MIT will join us to explore the seismic changes being brought about in societies, economies, and organizations by three primary elements of the machine age: Machines are now transforming the role of human decision-making, digital platforms allow products and services of others to be sold and brokered, and there's a proliferation of an almost-magical effectiveness for obtaining ideas from the general public–the crowd–rather than from the experts at the core of the business.

Navigating SC18

(back to top)

Monday, November 12th

Room: D227
4:45 pm - 5:30 pm

Navigating SC

Navigating SC *Elsa Gonsiorowski (Lawrence Livermore National Laboratory)*
Wondering how to make the most of your busy SC week or what your admissions badge entitles you to participate in?

The SC18 Conference welcomes first-time attendees. We hope to attract new and diverse groups of HPC professionals and students to the conference each year with the goal of sparking new conversations, new connections, and new ideas. However, we understand that it can be difficult to navigate the conference and experience all that SC has to offer.

We are hosting sessions on Monday afternoon and Tuesday morning to help first time attendees navigate the conference.

Tuesday, November 13th

Room: D227
7:30 am - 8:15 am

Navigating SC

Navigating SC *Elsa Gonsiorowski (Lawrence Livermore National Laboratory)*
Wondering how to make the most of your busy SC week or what your admissions badge entitles you to participate in?

The SC18 Conference welcomes first-time attendees. We hope to attract new and diverse groups of HPC professionals and students to the conference each year with the goal of sparking new conversations, new connections, and new ideas. However, we understand that it can be difficult to navigate the conference and experience all that SC has to offer.

We are hosting sessions on Monday afternoon and Tuesday morning to help first time attendees navigate the conference.

Panel

(back to top)

Tuesday, November 13th

Room: C147/148/154
10:30 am - 12:00 pm

Quantum Communication Networks and Technologies

Moderator: *Leonie Mueck (Public Library of Science)*
Panelist: *Joseph Lykken (Fermi National Accelerator Laboratory), Maria Spiropulu (California Institute of Technology), Christoph Simon (University of Calgary), Mercedes Gimeno-Segovia (PsiQuantum), Jake Taylor (Office of Science and Technology Policy (OSTP))*

Quantum Information Science (QIS) and technology has the potential to revolutionize computation and communication and to help solve fundamental questions in physics. Communication and interconnection of quantum devices are important for the success of this emergent field and is an active topic of research. Quantum channels are fundamental to QIS since they allow the communication of quantum information without the loss of quantum coherence. Fundamental research on quantum communication not only promises to bring transformational advances in long-distance communication and in high performance computation but may result in groundbreaking results on long-standing fundamental physics questions including the elucidation of the deep connection between entanglement and space time. In this panel, we interrogate the status and prospects for quantum networks and relevant quantum technologies and discuss potential breakthroughs.

Room: C155/156
10:30 am - 12:00 pm

HPC in Cloud or Cloud in HPC: Myths, Misconceptions and Misinformation

Moderator: *Sadaf R. Alam (Swiss National Supercomputing Centre)*
Panelist: *Victoria Stodden (University of Illinois), Eli Dart (Energy Sciences Network (ESnet)), Dirk Pleiter (Juelich Supercomputing Centre), David Hancock (Indiana University), Stephen Poole (Los Alamos National Laboratory)*

Scientific workflows today, especially those involving large-scale data sources, require an ecosystem of HPC and cloud computing, storage and networking technologies. Thanks to these technologies, it becomes possible to address bigger challenges collectively through federated IT infrastructures. Several initiatives have emerged as part of nationally funded research infrastructure and public cloud providers to facilitate ever-increasing needs of computing and storage capabilities alongside accessibility and quality of service requirements such as interactivity and security. This panel brings together a diverse group of experts and practitioners to unravel myths, misconceptions, and misinformation around growing service portfolios involving HPC and X-as-a-service technologies. In particular, the panelists will reflect on solutions that provide cost-to-performance efficiencies for

different quantitative and qualitative metrics, such as need for exascale while maintaining information security together with isolation and customization of services by adopting cloud technologies in HPC and vice-versa, and opportunities to increase computational reproducibility for scientific workflows.

Room: C147/148/154
1:30 pm - 3:00 pm

Swiss Army Programming: Performance and Portability from Modern Tools

Moderator: *Jason D. Sewall (Intel Corporation)*
Panelist: *Barbara Chapman (Stony Brook University), Olivier Giroux (Nvidia Corporation), Gihan Mudalige (University of Warwick), Sunita Chandrasekaran (University of Delaware), Christian Trott (Sandia National Laboratories), Andrew Richards (Codeplay Software Ltd)*

Hardware is evolving in complexity and diversity, enabling us to reach ever-higher peaks of performance. However, this comes at a cost to software development: some hardware requires special consideration that can hinder portability, and even portable code that runs well on one device may not run well on another. Given the great expense of software development and validation, we continually hunt for new tools---and augmentations to old ones---that mitigate these costs. Tools for developing single-source, cross-platform codes---like Kokkos, RAJA, OpenACC and OpenMP---are popular, while libraries like LIBXSMM use new code-generation techniques to improve performance.

Can we build a do-it-all compiler? Are prescriptive or descriptive approaches better? How do you prepare for the Next Great Thing? This panel discussion features experts in popular libraries and languages for portable, high-performance computing, each with a different philosophy, a distinct approach to these problems, and a willingness to stand up and defend their perspective.

Room: C155/156
3:30 pm - 5:00 pm

How Can Lessons Learned in the Past Forty Years Guide Future HPC Research Strategies?

Moderator: *Paul Messina (Argonne National Laboratory)*
Panelist: *Arthur Bland (Oak Ridge National Laboratory), Jackie Chen (Sandia National Laboratories), Phil Colella (Lawrence Berkeley National Laboratory), Jack Dongarra (University of Tennessee, Oak Ridge National Laboratory), Thom Dunning (University of Washington, Pacific Northwest National Laboratory), Wendy Huntoon (Keystone Initiative for Network Based Education and Research), Dan Reed (University of Iowa)*

The HPC, computer science, and mathematics communities face disruptive changes due to the end of Dennard scaling and the emergence of quantum, neuromorphic, and other technologies. Having weathered previous disruptions, are there insights into effective investment strategies and guidance in research investments in the coming decades? Drawing on the experience of the Advanced Scientific Computing Research program of the U.S. Department of Energy, as well as other agencies, the

panelists will address questions including:

Which programmatic techniques, including partnering with other agencies or industry, were most successful and might translate to the future?

What are examples of investment failures that might have been avoided?

What lessons have been learned to help adapt to the end of Moore's Law and the likely shift in software architecture that will be needed?

What has been the impact of supercomputing facilities and computer networks and will such facilities continue to be relevant?

Wednesday, November 14th

Room: C147/148/154
10:30 am - 12:00 pm

SC: The Conference

Moderator: *Beverly Clayton (Pittsburgh Supercomputing Center)*
Panelist: *Joanne Martin (Hartman Executive Advisors), Roscoe Giles (Boston University), Becky Verastegui (Oak Ridge National Laboratory), William Gropp (University of Illinois), Bernd Mohr (Forschungszentrum Juelich), Michela Taufer (Association for Computing Machinery (ACM); University of Delaware, University of Tennessee)*

SC: The Conference

Started as the Supercomputing Conference in 1988, The Conference on High Performance Computing, Networking, Storage and Analysis (SC) will be examined by a panel of Past and Future Conference Chairs with audience interaction. The Chairs will discuss SC's impact and role in advancing HPC, how the field has evolved in 30 years, and projected future impacts. Topics include:

—What impact has SC had on industry? Science? Society? What is the future of SC? What should it be?

—Has SC broadened participation in the field by students, underrepresented groups, and international collaborations? How can it do more?

—Has SC played a part or demonstrated how the community addressed challenges such as network growth and use, interoperability, high productivity software, benchmarking, and power consumption?

—What has been SC's impact on the relationships among industry, academia, and government?

—How can we ensure that the conference remains relevant?

Room: C155/156
10:30 am - 12:00 pm

Innovative Approaches for Developing Accessible, Productive, Scalable HPC Training

Moderator: *Toni Collis (Women in High Performance Computing, Appentra Solutions)*
Panelist: *Ashley Barker (Oak Ridge National Laboratory), Osni Marques (Lawrence Berkeley National Laboratory), Manuel Arenaz (Appentra Solutions), Weronika Filinger (University of Edinburgh), Elaine Raybourn (Sandia National Laboratories)*

As the provision of HPC rapidly increases, an increasing number of application domains and users are becoming interested in the potential that HPC can unlock for their field. This is creating an urgent need for scalable, accessible training to ensure effective use of resources. The majority of current training for HPC is still provided to domain-experts by computers scientists in a classroom environment. This necessarily limits both the scalability of the training and the tailoring of the training to the needs of domains new to HPC. This panel will focus on innovative approaches to improve the learning experience for domain-experts and how to reach an ever increasingly diverse and dispersed audience.

Topics covered will include using new tools and teaching methods, whether hackathons are useful for the non-specialists, and the need for providing training that recognizes the needs of a diverse population, including how to improve access for underrepresented groups.

Room: C147/148/154
1:30 pm - 3:00 pm

Reconfigurable Computing for HPC: Will It Make It this Time?

Moderator: *Franck Cappello (Argonne National Laboratory)*
Panelist: *Torsten Hoefler (ETH Zurich), Kentaro Sano (RIKEN), Maya Gokhale (Lawrence Livermore National Laboratory), Andrew Putnam (Microsoft Corporation), Kazutomo Yoshii (Argonne National Laboratory), Ronan Keryell (Xilinx Inc)*

Reconfigurable computing has been adopted in many domains requiring fast computing with a relatively low power budget. It has been recently introduced in large-scale systems such as the Cloud and extreme-scale data centers. The end of Moore's law and the introduction of devices with exceptional floating point computing capabilities make reconfigurable computing more attractive than ever for HPC. Many questions are open. How to program efficiently reconfigurable devices at high level? Where to put the reconfigurable devices (compute node, network, storage)? What algorithms/data representation provide advantage for reconfigurable devices over CPUs and GPUs? What metrics should be used to measure performance of reconfigurable devices and to compare with CPUs and GPUs? The panel will explore these questions and discuss why reconfigurable computing did not succeed in HPC in the past and why it could make it this time. The panelists will be reconfigurable computing experts from industry and academia.

Room: C155/156
1:30 pm - 3:00 pm

Runtime for Exascale and Beyond: Convergence or Divergence?

Moderator: *Marc Snir (University of Illinois)*
Panelist: *Pavan Balaji (Argonne National Laboratory), Laxmikant Kale (University of Illinois), Vivek Sarkar (Georgia Institute of Technology), Sean Treichler (Nvidia Corporation), Hartmut Kaiser (Louisiana State University), Raymond Namyst (University of Bordeaux)*

We propose to discuss in the panel the following three assertions:

a) The coming decade is likely to see an accelerated evolution in programming models for HPC, for several reasons:

-- As Moore's law plateaus, increased performance requires increasingly heterogeneous platforms and asynchronous algorithms.

-- As platforms become more dynamic (e.g., due to dynamic power management) and algorithms become more adaptive, runtime adaptation is increasingly important. Asynchronous Task Models (ATM), such as Charm++, Legion, PaRSEC, etc., seem better suited to handle this evolution than the prevalent model of MPI+OpenMP.

b) The potential benefits of new programming models cannot fully realized without a more powerful runtime that better integrates internode and intranode synchronization and communication with local resource scheduling and memory management.

c) The development and deployment of new programming models, as well as more efficient support for current models can be accelerated by standardizing new interfaces that enable such integration

Room: C147/148/154
3:30 pm - 5:00 pm

Sustaining Research Software

Moderator: *Daniel S. Katz (University of Illinois, National Center for Supercomputing Applications)*
Panelist: *Sandra Gesing (University of Notre Dame), Patrick Aerts (Netherlands eScience Center, Data Archiving and Networked Services (DANS)), Anshu Dubey (Argonne National Laboratory), David E. Pearah (HDF Group), Neil P. Chue Hong (Software Sustainability Institute, University of Edinburgh), Henry J. Neeman (University of Oklahoma)*

Many science advances have been possible thanks to the use of research software, which has become essential to advancing virtually every Science, Technology, Engineering and Mathematics (STEM) discipline (and many non-STEM disciplines as well). And while much of it is made available under open source licenses, work is needed to develop, support, and sustain it, as underlying systems and software as well as user needs evolve. This panel will discuss challenges and solutions related to sustaining research software and will collect examples of successes and failures from the audience

that can and will be widely shared. This will enable advanced compute and data infrastructures to positively impact future research software and future research.

Room: C155/156
3:30 pm - 5:00 pm

The Difference Between HPC on Premises and in the Cloud

Moderator: *Lara Kisielewska (Xand Marketing)*
Panelist: *Naoki Shibata (XTREME-D Inc), Martin Hull (Arista Networks), Ian Colle (Amazon Web Services), Gregory Kurtzer (Sylabs), Brett Newman (Microway Inc)*

Cloud computing is gaining popularity. And with this is the need to evaluate the factors that determine when an HPC user can or should use the cloud. Our panelists represent a spectrum of perspectives by entities that provide services for high-performance workloads. They will discuss different cloud computing options, infrastructure choices for cloud versus on-premise computing (including containers), and how to evaluate options based on factors such as performance, availability, data security, cost, and time-to-solution.

Thursday, November 15th

Room: C147/148/154
10:30 am - 12:00 pm

"If you can't measure it, you can't improve it" -- Software Improvements from Power/Energy Measurement Capabilities

Moderator: *Daniel Reed (University of Iowa)*
Panelist: *Satoshi Matsuoka (Tokyo Institute of Technology, RIKEN), Sadaf Alam (Swiss National Supercomputing Centre), Jack Deslippe (Lawrence Berkeley National Laboratory), Tapasya Patki (Lawrence Livermore National Laboratory)*

"If you can't measure it, you can't improve it." We now have some sites that are extremely well instrumented for measuring power and energy. Lawrence Berkeley National Laboratory's supercomputing center is a prime example, with sensors, data collection and analysis capabilities that span the facility and computing equipment. We have made major gains in improving the energy efficiency of the facility as well as computing hardware, but there are still large gains to be had with software- particularly application software. Just tuning code for performance isn't enough; the same time to solution can have very different power profiles. We are at the point where measurement capabilities are allowing us to see cross-cutting issues- such as the cost of spin waits. These new measurement capabilities should provide a wealth of information to see the tall poles in the tent. This panel will explore how we can identify these emerging tall poles.

Room: C155/156
10:30 am - 12:00 pm

Federated Cloud: An Evolutionary Path from Grid Computing

Moderator: *Khalil Yazdi (Open Research Cloud Alliance)*
Panelist: *Robert Bohn (National Institute of Standards and Technology), Martial Michel (Data Machines Corporation), Craig Lee (Aerospace Corporation), Andrew Grimshaw (University of Virginia), Frederica Darema (United States Air Force)*

The current cloud ecosystem is one in which cloud providers do not interoperate. This hinders the service reach, resources and scalability that can be offered to the customer. This has led to a growing recognition that the lack of cloud federation in a landscape of multiple independent cloud providers is a technical and economic challenge. The federation of research clouds is essential and necessary enabling technology for the development and translation of innovations in IoT, high-performance computing, distributed big-data analytics, and global scientific collaboration. Previous studies and implementations of grid computing have encountered and solved similar problems.

-- What are some of these unique challenges that impede progress to federated cloud?

-- Can previous work in grid computing lead to an evolved set of solutions to maximize the return on investment for providers and users of cloud computing?

-- What has grid computing been accomplished that cloud federation can leverage?

Room: C155/156
1:30 pm - 3:00 pm

Approximating for Faster, Better and Cheaper Scientific Computing

Moderator: *Hatem Ltaief (King Abdullah University of Science and Technology)*
Panelist: *George Biros (University of Texas), Jack Dongarra (University of Tennessee; ORNL, University of Manchester), Fernanda Foertter (Nvidia Corporation), Sherry Li (Lawrence Berkeley National Laboratory), Satoshi Matsuoka (Tokyo Institute of Technology, RIKEN), William Tang (Princeton University)*

HPC simulations and big data applications face similar challenges when solving extreme-scale scientific problems: high algorithmic complexity and large memory footprint. CMOS and memory technology scaling have continuously mitigated these challenges with an exponential growth in processor performance and a constant increase in speed and capacity, respectively. The free lunch is perhaps over but now artificial intelligence comes to the rescue, with the sheer volume of data to be crunched.

This has caused the community to explore disruptive hardware and software solutions to overcome both challenges. One of the trends consists of exploring numerical approximations, e.g., mixed/low precision and hierarchical low-rank matrix computations, with hardware support for a variety of precisions.

The panelists investigate alternatives to today's state-of-the-art chips and numerical algorithms. The main idea consists in trading-off accuracy for performance throughout the hardware and software ecosystems in the context of exascale scientific computing.

Room: C147/148/154
3:30 pm - 5:00 pm

The Next Wave of HPC in the Datacenter

Moderator: *Timothy Prickett Morgan (The Next Platform)*
Panelist: *Greg Stoner (Advanced Micro Devices Inc), Dr. Rene Meyer (AMAX), Dhabaleswar K. "DK" Panda (Ohio State University), Joseph George (Cray Inc)*

What's the driving force behind the HPC revolution? Using deep learning (DL), machine learning (ML), and artificial intelligence (AI) to create actionable insights from your data. This is where the intersection between CPU, GPU, and software makes all the difference. The race is on to power today's insight-driven world in the most impactful ways possible. Join this discussion to learn how organizations are using HPC to unlock the value of their data. You'll walk away with the key tenants for how to best deploy infrastructure to meet today's demanding AI workloads, insight into how emerging technologies can best support HPC and AI, and real-world examples of how top industry players are already putting these best practices to work.

Friday, November 16th

Room: C155/156
10:30 am - 12:00 pm

Convergence between HPC and Big Data: The Day After Tomorrow

Moderator: *Bilel Hadri (King Abdullah University of Science and Technology)*
Panelist: *Sadaf Alam (Swiss National Supercomputing Centre), Katie Antypas (Lawrence Berkeley National Laboratory), David Bader (Georgia Institute of Technology), Dan Reed (University of Iowa), Rio Yokota (Tokyo Institute of Technology)*

The relationship between HPC and Big-Data Analytics (BDA) is a recurring theme in the community. There is no doubt that the convergence between the third and fourth paradigms is on the march. The multiple combinations of various techniques from both ends, ranging from virtualization, containerization, and cloud to help researchers in their simulations, represent only an aspect of it. As HPC and BDA become increasingly ubiquitous and intertwined, this synergism will deliver opportunities and challenges that neither disciplines can produce alone.

Join the expert panelists, to learn from their experiences and discuss with them on:

-- What are the challenges and the opportunities that the community will face out of this synergistic

relationship?

-- Will users need to make changes to their practices to use upcoming modern resources to solve their workflow more efficiently?

-- How to develop sustainable competencies around implementing and managing new technologies to support diverse workloads?

Paper

(back to top)

Tuesday, November 13th

Room: C140/142
10:30 am - 12:00 pm

Next-Generation Networking

Exploiting Idle Resources in a High-Radix Switch for Supplemental Storage
Matthias A. Blumrich (Nvidia Corporation), Nan Jiang (Nvidia Corporation), Larry R. Dennison (Nvidia Corporation)

A general-purpose switch for a high-performance network is usually designed with symmetric ports providing credit-based flow control and error recovery via link-level retransmission. Because port buffers must be sized for the longest links and modern asymmetric network topologies have a wide range of link lengths, we observe that there can be a significant amount of unused buffer memory, particularly in edge switches. We also observe that the tiled architecture used in many high-radix switches contains an abundance of internal bandwidth. We combine these observations to create a new switch architecture that allows ports to stash packets in unused buffers on other ports, accessible via excess internal bandwidth in the tiled switch. We explore this architecture through two use cases: end-to-end resilience and congestion mitigation. We find that stashing is highly effective and does not negatively impact network performance.

Best Paper Finalist: yes
Best Student Paper Finalist: no

Fine-Grained, Multi-Domain Network Resource Abstraction as a Fundamental Primitive to Enable High-Performance, Collaborative Data Sciences
Qiao Xiang (Yale University), J. Jensen Zhang (Tongji University), X. Tony Wang (Tongji University), Y. Jace Liu (Tongji University), Chin Guok (Lawrence Berkeley National Laboratory), Franck Le (IBM), John MacAuley (Lawrence Berkeley National Laboratory), Harvey Newman (California Institute of Technology), Y. Richard Yang (Yale University)

Multi-domain network resource reservation systems are being deployed, driven by the demand and substantial benefits of providing predictable network resources. However, a major lack of existing systems is their coarse granularity, due to the participating networks' concern of revealing sensitive information, which can result in substantial inefficiencies. This paper presents Mercator, a novel multi-domain network resource discovery system to provide fine-grained, global network resource information, for collaborative sciences. The foundation of Mercator is a resource abstraction through algebraic-expression enumeration (i.e., linear inequalities/equations), as a compact representation of the available bandwidth in multi-domain networks. In addition, we develop an obfuscating protocol, to address the privacy concerns by ensuring that no participant can associate the algebraic expressions with the corresponding member networks. We also introduce a superset projection technique to increase Mercator's scalability. Finally, we implement Mercator and demonstrate both its efficiency and

efficacy through extensive experiments using real topologies and traces.

Best Paper Finalist: no
Best Student Paper Finalist: no

Light-Weight Protocols for Wire-Speed Ordering
Hans Eberle (Nvidia Corporation), Larry Dennison (Nvidia Corporation)

We describe light-weight protocols for selective packet ordering in out-of-order networks that carry memory traffic. The protocols are designed for heterogeneous high-performance systems, in particular, accelerated systems with endpoints that have few resources available for interfacing the network.

The protocols preserve the semantics of a relaxed memory ordering model as adopted by highly-threaded many-core processors and accelerators.

The protocols achieve link-rate performance through the following techniques: (1) speculative connection setup avoids round-trip delays found in protocols with little knowledge about endpoint resources, (2) target-side ordering avoids round-trip delays found in source-side ordering mechanisms, (3) fine-grained ordering removes dependencies unwarranted by program code avoiding cumulative ordering dependencies caused by coarse-grained ordering, (4) ordering relaxations and optimizations for producer/consumer communication patterns.

We describe two ordering protocols that provide (1) strict sequential ordering and (2) relaxed ordering for multi-packet transfers. The protocols impose no restrictions on routing, including multipath routing.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C141/143/149
10:30 am - 12:00 pm

Resilience

GPU Age-Aware Scheduling to Improve the Reliability of Leadership Jobs on Titan
Christopher Zimmer (Oak Ridge National Laboratory), Don Maxwell (Oak Ridge National Laboratory), Stephen McNally (Oak Ridge National Laboratory), Scott Atchley (Oak Ridge National Laboratory), Sudharshan S. Vazhkudai (Oak Ridge National Laboratory)

The increasing rate of failures on the Oak Ridge Leadership Computing Facility's (OLCF) Titan supercomputer, resulted in the replacement of 50% of its GPUs between 2015 and 2017. The largest jobs, also known as "leadership jobs", continued to experience increased application failures. These jobs contained significant amounts of low-failure rate and high-failure rate GPUs. The impacts of these failures were felt more by leadership jobs due to longer wait times, runtimes, and higher charge rates. In this work, we have designed techniques to increase the use of low-failure GPUs in leadership jobs through targeted resource allocation. This employed two complementary techniques, updating both

the system ordering and the allocation mechanisms. In simulation, the application of these techniques resulted in a 33% increase in low-failure GPU hours being assigned to leadership jobs. Our GPU Age-Aware Scheduling has been used in production on Titan since July of 2017.

Best Paper Finalist: no
Best Student Paper Finalist: no

FlipTracker: Understanding Natural Error Resilience in HPC Applications
Luanzheng Guo (University of California, Merced), Dong Li (University of California, Merced), Ignacio Laguna (Lawrence Livermore National Laboratory), Martin Schulz (Technical University Munich)

As high-performance computing systems scale in size and computational power, the danger of silent errors, i.e., errors that can bypass hardware detection mechanisms and impact application state, grows dramatically. Consequently, applications running on HPC systems need to exhibit resilience to such errors. Previous work has found that, for certain codes, this resilience can come for free, i.e., some applications are naturally resilient, but few works have shown the code patterns—combinations or sequences of computations—that make an application naturally resilient. In this paper, we present FlipTracker, a framework designed to extract these patterns using fine-grained tracking of error propagation and resilience properties, and we use it to present a set of computation patterns that are responsible for making representative HPC applications naturally resilient to errors. This not only enables a deeper understanding of resilience properties of these codes, but also can guide future application designs toward patterns with natural resilience.

Best Paper Finalist: no
Best Student Paper Finalist: no

Doomsday: Predicting Which Node Will Fail When on Supercomputers
Anwesha Das (North Carolina State University), Frank Mueller (North Carolina State University), Paul Hargrove (Lawrence Berkeley National Laboratory), Eric Roman (Lawrence Berkeley National Laboratory), Scott Baden (Lawrence Berkeley National Laboratory)

Predicting which node will fail and how soon remains a challenge for HPC resilience, yet may pave the way to exploiting proactive remedies before jobs fail. Not only for increasing scalability up to exascale systems, but even for contemporary supercomputer architectures does it require substantial efforts to distill anomalous events from noisy raw logs. To this end, we propose a novel phrase extraction mechanism called TBP (time-based phrases) to pin-point node failures, which is unprecedented. Our study, based on real system data and statistical machine learning, demonstrates the feasibility to predict which specific node will fail in Cray systems. TBP achieves no less than 83% recall rates with lead times as high as 2 minutes. This opens up the door for enhancing prediction lead times for supercomputing systems in general, thereby facilitating efficient usage of both computing capacity and power in large scale production systems.

Best Paper Finalist: no
Best Student Paper Finalist: yes

Room: C146

10:30 am - 12:00 pm

Data and Storage

SP-Cache: Load-Balanced, Redundancy-Free Cluster Caching with Selective Partition

Yinghao Yu (Hong Kong University of Science and Technology), Renfei Huang (Hong Kong University of Science and Technology), Wei Wang (Hong Kong University of Science and Technology), Jun Zhang (Hong Kong University of Science and Technology), Khaled Ben Letaief (Hong Kong University of Science and Technology)

Data-intensive clusters increasingly employ in-memory solutions to improve I/O performance. However, the routinely observed file popularity skew and load imbalance create hotspots, which significantly degrades the benefits of in-memory solutions. Common approaches to tame load imbalance include copying multiple replicas of hot files and creating parity chunks using storage codes. Yet, these techniques either suffer from high memory redundancy or incur non-trivial encoding/decoding overhead. In this paper, we propose a different approach to achieve load balancing without memory redundancy or encoding/decoding overhead. Our solution, termed SP-Cache, selectively partitions files based on their popularity and evenly caches those partitions across the cluster. We develop an efficient algorithm to determine the optimal number of partitions for hot files—too few partitions are incapable of mitigating hotspots, while too many are susceptible to stragglers. EC2 deployment and trace-driven simulations show that, compared with existing solutions, SP-Cache reduces the read latencies by up to 40%.

Best Paper Finalist: no
Best Student Paper Finalist: no

BESPOKV: Application Tailored Scale-Out Key-Value Stores

Ali Anwar (IBM), Yue Cheng (George Mason University), Hai Huang (IBM), Jingoo Han (Virginia Tech), Hyogi Sim (Oak Ridge National Laboratory), Dongyoon Lee (Virginia Tech), Fred Douglis (Perspecta Labs), Ali R. Butt (Virginia Tech)

Enterprise KV stores are not well suited for HPC applications, and entail customization and cumbersome end-to-end KV design to extract the HPC application needs. In this paper we present BESPOKV, an adaptive, extensible, and scale-out KV store framework. BESPOKV decouples the KV store design into the control plane for distributed management and the data plane for local data store. BESPOKV takes as input a single-server KV store, called a datalet, and transparently enables a scalable and fault-tolerant distributed KV store service. The resulting distributed stores are also adaptive to consistency or topology requirement changes and can be easily extended for new types of services. Experiments show that BESPOKV-enabled distributed KV stores scale horizontally to a large number of nodes, and performs comparably and sometimes better than the state-of-the-art systems.

Best Paper Finalist: no
Best Student Paper Finalist: no

Scaling Embedded In Situ Indexing with DeltaFS

Qing Zheng (Carnegie Mellon University), Charles D. Cranor (Carnegie Mellon University), Danhao Guo (Carnegie Mellon University), Gregory R. Ganger (Carnegie Mellon University), George Amvrosiadis (Carnegie Mellon University), Garth A. Gibson (Carnegie Mellon University), Bradley W. Settlemyer

(Los Alamos National Laboratory), Gary Grider (Los Alamos National Laboratory), Fan Guo (Los Alamos National Laboratory)

Analysis of large-scale simulation output is a core element of scientific inquiry, but analysis queries may experience significant I/O overhead when the data is not structured for efficient retrieval. While in-situ processing allows for improved time-to-insight for many applications, scaling in-situ frameworks to hundreds of thousands of cores can be difficult in practice. The DeltaFS in-situ indexing is a new approach for in-situ processing of massive amounts of data to achieve efficient point and small-range queries. This paper describes the challenges and lessons learned when scaling this in-situ processing function to hundreds of thousands of cores. We propose techniques for scalable all-to-all communication that is memory and bandwidth efficient, concurrent indexing, and specialized LSM-Tree formats. Combining these techniques allows DeltaFS to control the cost of in-situ processing while maintaining 3 orders of magnitude query speedup when scaling alongside the popular VPIC particle-in-cell code to 131,072 cores.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C140/142
1:30 pm - 3:00 pm

Biology Applications

Extreme Scale De Novo Metagenome Assembly

Evangelos Georganas (Intel Corporation), Rob Egan (Lawrence Berkeley National Laboratory), Steven Hofmeyr (Lawrence Berkeley National Laboratory), Eugene Goltsman (Lawrence Berkeley National Laboratory), Bill Arndt (Lawrence Berkeley National Laboratory), Andrew Tritt (Lawrence Berkeley National Laboratory), Aydin Buluc (Lawrence Berkeley National Laboratory), Leonid Oliker (Lawrence Berkeley National Laboratory), Katherine Yelick (Lawrence Berkeley National Laboratory)

Metagenome assembly is the process of transforming a set of short, overlapping, and potentially erroneous DNA segments from environmental samples into the accurate representation of the underlying microbiomes's genomes. State-of-the-art tools require large shared memory machines and cannot handle contemporary metagenome datasets that exceed terabytes in size. In this paper, we introduce the metaHipMer pipeline, a high-quality and high-performance metagenome assembler that employs an iterative de Bruijn graph approach. MetaHipMer leverages a specialized scaffolding algorithm that produces long scaffolds and accommodates the idiosyncrasies of metagenomes. MetaHipMer is end-to-end parallelized using the Unified Parallel C language and therefore can run seamlessly on shared and distributed-memory systems. Experimental results show that metaHipMer matches or outperforms the state-of-the-art tools in terms of accuracy. Moreover, metaHipMer scales efficiently to large concurrencies and is able to assemble previously intractable grand challenge metagenomes.

Best Paper Finalist: yes
Best Student Paper Finalist: no

Optimizing High Performance Distributed Memory Parallel Hash Tables for DNA k-mer Counting
Tony C. Pan (Georgia Institute of Technology, School of Computational Science and Engineering), Sanchit Misra (Intel Corporation, Parallel Computing Lab), Srinivas Aluru (Georgia Institute of Technology, School of Computational Science and Engineering)

High-throughput DNA sequencing is the mainstay of modern genomics research. A common operation used in bioinformatic analysis for many applications of high-throughput sequencing is the counting and indexing of fixed length substrings of DNA sequences called k-mers. Counting k-mers is often accomplished via hashing, and distributed memory k-mer counting algorithms for large data sets are memory access and network communication bound. In this work, we present two optimized distributed parallel hash table techniques that utilize cache friendly algorithms for local hashing, overlapped communication and computation to hide communication costs, and vectorized hash functions that are specialized for k-mer and other short key indices. On 4096 cores of the NERSC Cori supercomputer, our implementation completed index construction and query on an approximately 1 TB human genome dataset in just 11.8 seconds and 5.8 seconds, demonstrating speedups of 2.06x and 3.7x, respectively, over the previous state-of-the-art distributed memory k-mer counter.

Best Paper Finalist: no
Best Student Paper Finalist: no

Redesigning LAMMPS for Petascale and Hundred-Billion-Atom Simulation on Sunway TaihuLight
Xiaohui Duan (Shandong University), Ping Gao (Shandong University), Tingjian Zhang (Shandong University), Meng Zhang (Shandong University), Weiguo Liu (Shandong University), Wusheng Zhang (Tsinghua University), Wei Xue (Tsinghua University), Haohuan Fu (Tsinghua University), Lin Gan (Tsinghua University), Dexun Chen (Tsinghua University), Xiangxu Meng (Shandong University), Guangwen Yang (Tsinghua University)

Large-scale molecular dynamics (MD) simulations on supercomputers play an increasingly important role in many research areas. In this paper, we present our efforts on redesigning the widely used LAMMPS MD simulator for Sunway TaihuLight supercomputer and its ShenWei many-core architecture (SW26010). The memory constraints of SW26010 bring a number of new challenges for achieving efficient MD implementation on it. In order to overcome these constraints, we employ four levels of optimization: (1) a hybrid memory update strategy; (2) a software cache strategy; (3) customized transcendental math functions; and (4) a full pipeline acceleration. Furthermore, we redesign the code to enable all possible vectorization. Experiments show that our redesigned software on a single SW26010 processor can outperform over 100 E5-2650 cores for running the latest stable release (11Aug17) of LAMMPS. We also achieve a performance of over 2.43 PFlops for a Tersoff simulation when using 16,384 nodes on Sunway TaihuLight.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C141/143/149
1:30 pm - 3:00 pm

Performance and Energy Analysis

A Parallelism Profiler with What-If Analyses for OpenMP Programs

Nader Boushehrinejadmoradi (Rutgers University), Adarsh Yoga (Rutgers University), Santosh Nagarakatte (Rutgers University)

This paper proposes OMP-WHIP, a profiler that measures inherent parallelism in the program for a given input and provides what-if analyses to estimate improvements in parallelism. We propose a novel OpenMP series parallel graph representation (OSPG) that precisely captures series-parallel relations induced by various directives between different fragments of dynamic execution. OMP-WHIP constructs the OSPG and measures the computation performed by each dynamic fragment using hardware performance counters. This series-parallel representation along with the fine-grained measurement of computation is a performance model of the program for a given input, which enables computation of inherent parallelism. This novel performance model also enables what-if analyses where a programmer can estimate improvements in parallelism when bottlenecks are parallelized. We have used OMP-WHIP to identify parallelism bottlenecks in more than forty applications and then designed strategies to improve the speedup in seven applications.

Best Paper Finalist: no
Best Student Paper Finalist: no

Energy Efficiency Modeling of Parallel Applications

Mark Endrei (University of Queensland), Chao Jin (University of Queensland), Minh Ngoc Dinh (University of Queensland), David Abramson (University of Queensland), Heidi Poxon (Cray Inc), Luiz DeRose (Cray Inc), Bronis R. de Supinski (Lawrence Livermore National Laboratory)

Energy efficiency has become increasingly important in high performance computing (HPC), as power constraints and costs escalate. Workload and system characteristics form a complex optimization search space in which optimal settings for energy efficiency and performance often diverge. Thus, we must identify trade-off options to find the desired balance. We present an innovative statistical model that accurately predicts the Pareto optimal trade-off options using only user-controllable parameters. Our approach can also tolerate both measurement and model errors. We study model training and validation using several HPC kernels, then with more complex workloads, including AMG and LAMMPS. We can calibrate an accurate model from as few as 12 runs, with prediction error of less than 10%. Our results identify trade-off options allowing up to 40% energy efficiency improvement at the cost of under 20% performance loss. For AMG, we reduce the required sample measurement time from 13 hours to 74 minutes.

Best Paper Finalist: no
Best Student Paper Finalist: no

HPL and DGEMM Performance Variability on the Xeon Platinum 8160 Processor

John D. McCalpin (University of Texas, Texas Advanced Computing Center)

During initial testing of a large cluster equipped with Xeon Platinum 8160 processors, we observed infrequent, but significant, performance drops in HPL benchmark results. The variability was seen in both single node and multi-node runs, with approximately 0.4% of results more than 10% slower than the median. We were able to reproduce this behavior with a single-socket (24-core) DGEMM benchmark. Performance counter analysis of several thousand DGEMM runs showed that increased DRAM read traffic is the primary driver of increased execution time. Increased DRAM traffic in this

benchmark is primarily generated by dramatically elevated snoop filter evictions, which arise due to the interaction of high-order (physical) address bits with the hash used to map addresses across the 24 coherence agents on the processor. These conflicts (and the associated performance variability) were effectively eliminated (for both DGEMM and HPL) by using 1 GiB large pages.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C146
1:30 pm - 3:00 pm

Large-Scale Algorithms

Large-Scale Hierarchical K-Means for Heterogeneous Many-Core Supercomputers
Liandeng Li (Tsinghua University; National Supercomputing Center, Wuxi), Teng Yu (University of St Andrews), Wenlai Zhao (Tsinghua University; National Supercomputing Center, Wuxi), Haohuan Fu (Tsinghua University; National Supercomputing Center, Wuxi), Chenyu Wang (University of St Andrews; National Supercomputing Center, Wuxi), Li Tan (Beijing Technology and Business University), Guangwen Yang (Tsinghua University; National Supercomputing Center, Wuxi), John Thomson (University of St Andrews)

This paper presents a novel design and implementation of k-means clustering algorithm targeting the Sunway TaihuLight supercomputer. We introduce a multi-level parallel partition approach that not only partitions by dataflow and centroid, but also by dimension. Our multi-level (nkd) approach unlocks the potential of the hierarchical parallelism in the SW26010 heterogeneous many-core processor and the system architecture of the supercomputer.

Our design is able to process large-scale clustering problems with up to 196,608 dimensions and over 160,000 targeting centroids, while maintaining high performance and high scalability, significantly improving the capability of k-means over previous approaches. The evaluation shows our implementation achieves performance of less than 18 seconds per iteration for a large-scale clustering case with 196,608 data dimensions and 2,000 centroids by applying 4,096 nodes (1,064,496 cores) in parallel, making k-means a more feasible solution for complex scenarios.

Best Paper Finalist: no
Best Student Paper Finalist: no

TriCore: Parallel Triangle Counting on GPUs
Yang Hu (George Washington University), Hang Liu (University of Massachusetts, Lowell), H. Howie Huang (George Washington University)

Triangle counting algorithm enumerates the triangles in a graph by identifying the common neighbors between two vertices of every edge. In this work, we present TriCore, a new GPU-based high-performance and scalable triangle counting system that consists of three main techniques. First, we design a binary search based counting algorithm that tremendously increases both thread parallelism and memory performance. Second, TriCore exploits a 2-D partition method to distribute the CSR representation across

multiple GPUs, combined with a new streaming buffer to load the edge list from outside of GPUs. Third, we develop a dynamic workload management technique to balance the workload across multiple GPUs. Our evaluation demonstrates TriCore is 22× faster than the state-of-the-art parallel triangle counting projects. In addition, TriCore can not only process big graphs that are significant larger than the memory size of one GPU but also achieve 24× speedup when scaling to 32 GPUs.

Best Paper Finalist: no
Best Student Paper Finalist: no

Distributed-Memory Hierarchical Compression of Dense SPD Matrices
Chenhan D. Yu (University of Texas), Severin Reiz (Technical University Munich), George Biros (University of Texas)

We present a distributed-memory algorithm for the hierarchical compression of SPD matrices. Our method is based on GOFMM, an algorithm that appeared in doi:10.1145/3126908.3126921.

For many SPD matrices, GOFMM enables compression and approximate matrix-vector multiplication in NlogN time---as opposed to quadratic work required for a dense matrix. But GOFMM supports only shared memory parallelism. In this paper, we use the message passing interface, extending the ideas of GOFMM to the distributed memory setting. We also introduce an asynchronous algorithm for faster multiplication. We present different usage scenarios of SPD matrices that are related to graphs, neural-networks, and covariance operators. We also compare with STRUMPACK, which, to our knowledge, is the only other parallel software that can compress arbitrary SPD matrices. In our largest run, we were able to compress a 67M-by-67M matrix within three minutes and perform a multiplication with 512 vectors within 5 seconds on 6,144 Intel Skylake cores.

Best Paper Finalist: no
Best Student Paper Finalist: yes

Room: C140/142
3:30 pm - 5:00 pm

Resource Management and Interference

RM-Replay: A High-Fidelity Tuning, Optimization and Exploration Tool for Resource Management
Maxime Martinasso (Swiss National Supercomputing Centre), Miguel Gila (Swiss National Supercomputing Centre), Mauro Bianco (Swiss National Supercomputing Centre), Sadaf R. Alam (Swiss National Supercomputing Centre), Colin McMurtrie (Swiss National Supercomputing Centre), Thomas C. Schulthess (Swiss National Supercomputing Centre)

Leading hybrid and heterogeneous supercomputing systems process hundreds of thousands of jobs using complex scheduling algorithms and parameters. The centers operating these systems aim to achieve higher levels of resource utilization while being restricted by compliance with policy constraints. There is a critical need for a high-fidelity, high-performance tool with familiar interfaces that allows not only tuning and optimization of the operational job scheduler but also enables exploration of new resource management algorithms. We propose a new methodology and a tool

called RM-Replay which is not a simulator but instead a fast replay engine for production workloads. Slurm is used as a platform to demonstrate the capabilities of our replay engine.

The tool accuracy is discussed and our investigation shows that, by providing better job runtime estimation or using topology-aware allocation, scheduling metric values vary. The presented methodology to create fast replay engines can be extended to other complex systems.

Best Paper Finalist: no
Best Student Paper Finalist: no

Evaluation of an Interference-Free Node Allocation Policy on Fat-Tree Clusters
Samuel D. Pollard (University of Oregon), Nikhil Jain (Lawrence Livermore National Laboratory), Stephen Herbein (Lawrence Livermore National Laboratory), Abhinav Bhatele (Lawrence Livermore National Laboratory)

Interference between jobs competing for network bandwidth on a fat-tree cluster can cause significant variability and degradation in performance. These performance issues can be mitigated or completely eliminated if the resource allocation policy takes the network topology into account when allocating nodes to jobs. We implement a fat-tree network topology aware node allocation policy that allocates isolated partitions to jobs in order to eliminate inter-job interference. We compare the impact of this node allocation policy to a topology-oblivious policy with respect to the execution time of individual jobs with different communication patterns. We also evaluate the cluster's quality of service using metrics such as system utilization, schedule makespan, and job wait time for both policies. The results obtained for production workloads indicate that a topology-aware node allocation can provide interference-free execution without negatively impacting the cluster's quality of service.

Best Paper Finalist: no
Best Student Paper Finalist: no

Mitigating Inter-Job Interference Using Adaptive Flow-Aware Routing
Staci A. Smith (University of Arizona), Clara E. Cromey (University of Arizona), David K. Lowenthal (University of Arizona), Jens Domke (Tokyo Institute of Technology), Nikhil Jain (Lawrence Livermore National Laboratory), Jayaraman J. Thiagarajan (Lawrence Livermore National Laboratory), Abhinav Bhatele (Lawrence Livermore National Laboratory)

On most high performance computing platforms, applications share network resources with other jobs running concurrently on the system. Inter-job network interference can have a significant impact on the performance of communication-intensive applications, and no satisfactory solutions yet exist for mitigating this degradation.

In this paper, we analyze network congestion caused by multi-job workloads on two production systems that use popular network topologies---fat-tree and dragonfly. For each system, we establish a regression model to relate network hotspots to application performance degradation, showing that current routing strategies are insufficient to load-balance network traffic and mitigate interference on production systems. We then propose an alternative type of adaptive routing strategy, which we call adaptive flow-aware routing. We implement a prototype of our strategy, and tests on the fat-tree system show up to a 46% improvement in job run time when compared to the default routing.

Best Paper Finalist: no
Best Student Paper Finalist: yes

Room: C141/143/149
3:30 pm - 5:00 pm

Algorithms on Sparse Data

HiCOO: Hierarchical Storage of Sparse Tensors

Jiajia Li (Georgia Institute of Technology), Jimeng Sun (Georgia Institute of Technology), Richard Vuduc (Georgia Institute of Technology)

This paper proposes a new storage format for sparse tensors, called Hierarchical COOrdinate (HiCOO; pronounced: "haiku"). It derives from coordinate (COO) format, arguably the de facto standard for general sparse tensor storage. HiCOO improves upon COO by compressing the indices in units of sparse tensor blocks, with the goals of preserving the "mode-agnostic" simplicity of COO while reducing the bytes needed to represent the tensor and promoting data locality. We evaluate HiCOO by implementing a single-node, multicore-parallel version of the matricized tensor-times-Khatri-Rao product (MTTKRP) operation, which is the most expensive computational core in the widely used CANDECOMP/PARAFAC decomposition(CPD) algorithm. This MTTKRP implementation achieves up to 23.0× (6.8× on average) speedup over COO format and up to 15.6× (3.1× on average) speedup over another state-of-the-art format, compressed sparse fiber (CSF), by using less or comparable storage of them. When used within CPD, we also observe speedups against COO- and CSF-based implementations.

Best Paper Finalist: no
Best Student Paper Finalist: yes

Distributed Memory Sparse Inverse Covariance Matrix Estimation on High-Performance Computing Architectures

Aryan Eftekhari (University of Lugano), Matthias Bollhöfer (Braunschweig University of Technology), Olaf Schenk (University of Lugano)

We consider the problem of estimating sparse inverse covariance matrices for high-dimensional datasets using the l1-regularized Gaussian maximum likelihood method. This task is particularly challenging as the required computational resources increase superlinearly with the dimensionality of the dataset. We introduce a performant and scalable algorithm which builds on the current advancements of second-order, maximum likelihood methods. The routine leverages the intrinsic parallelism in the linear algebra operations and exploits the underlying sparsity of the problem. The computational bottlenecks are identified and the respective subroutines are parallelized using an MPI-OpenMP approach. Experiments conducted on a Cray XC50 system at the Swiss National Supercomputing Center show that, in comparison to the state-of-the-art algorithms, the proposed routine provides significant strong scaling speedup with ideal scalability up to 128 nodes. The developed framework is used to estimate the sparse inverse covariance matrix of both synthetic and real-world datasets with up to 10 million dimensions.

Best Paper Finalist: no

Best Student Paper Finalist: no

PruneJuice: Pruning Trillion-Edge Graphs to a Precise Pattern-Matching Solution

Tahsin Reza (University of British Columbia), Matei Ripeanu (University of British Columbia), Nicolas Tripoul (University of British Columbia), Geoffrey Sanders (Lawrence Livermore National Laboratory), Roger Pearce (Lawrence Livermore National Laboratory)

Pattern matching is a powerful graph analysis tool. Unfortunately, existing solutions have limited scalability, support only a limited set of search patterns, and/or focus on only a subset of the real-world problems associated with pattern matching. This paper presents a new algorithmic pipeline that: (i) enables highly scalable pattern matching on labeled graphs, (ii) supports arbitrary patterns, (iii) enables trade-offs between precision and time-to-solution (while always selecting all vertices and edges that participate in matches, thus offering 100% recall), and (iv) supports a set of popular data analytics scenarios. We implement our approach on top of HavoqGT and demonstrate its advantages through strong and weak scaling experiments on massive-scale real-world (up to 257 billion edges) and synthetic (up to 4.4 trillion edges) graphs, respectively, and at scales (1,024 nodes / 36,864 cores) orders of magnitude larger than used in the past for similar problems.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C146
3:30 pm - 5:00 pm

Performance Optimization Studies

Many-Core Graph Workload Analysis

Stijn Eyerman (Intel Corporation), Wim Heirman (Intel Corporation), Kristof Du Bois (Intel Corporation), Joshua B. Fryman (Intel Corporation), Ibrahim Hur (Intel Corporation)

Graph applications have specific characteristics that are not common in other application domains. In this paper, we analyze multiple graph applications on current multi- and many-core processors and provide conclusions and recommendations for future designs. We provide new insights on executing graph applications on many-core processors.

Our main novel observations are (i) some memory streams do show locality, while others show no locality, (ii) thread imbalance becomes a major problem with many threads, and (iii) many threads are required to saturate high-bandwidth memories. We recommend a selective memory access policy, where accesses with locality are cached and prefetched, while accesses without locality can remain uncached to save cache capacity. Additionally, more threads are needed, but they are not used efficiently due to thread imbalance. Our recommendation is to revise the graph analysis algorithms to provide more parallelism, and to provide a few high-performance cores that speedup sections with low parallelism.

Best Paper Finalist: no
Best Student Paper Finalist: no

Lessons Learned from Analyzing Dynamic Promotion for User-Level Threading
Shintaro Iwasaki (University of Tokyo), Abdelhalim Amer (Argonne National Laboratory), Kenjiro Taura (University of Tokyo), Pavan Balaji (Argonne National Laboratory)

A performance vs. practicality trade-off exists between user-level threading techniques. The community has settled mostly on a black-and-white perspective; fully fledged threads assume that suspension is imminent and incur overheads when suspension does not take place, and run-to-completion threads are more lightweight but less practical since they cannot suspend. Gray areas exist, however, whereby threads can start with minimal capabilities and then can be dynamically promoted to acquire additional capabilities when needed. This paper investigates the full spectrum of threading techniques from a performance vs. practicality trade-off perspective on modern multicore and many-core systems. Our results indicate that achieving the best trade-off highly depends on the suspension likelihood; dynamic promotion is more appropriate when suspension is unlikely and represents a solid replacement for run to completion, thanks to its lower programming constraints, while fully fledged threads remain the technique of choice when suspension likelihood is high.

Best Paper Finalist: no
Best Student Paper Finalist: no

Topology-Aware Space-Shared Co-Analysis of Large-Scale Molecular Dynamics Simulations
Preeti Malakar (Indian Institute of Technology Kanpur), Todd Munson (Argonne National Laboratory), Christopher Knight (Argonne National Laboratory), Venkatram Vishwanath (Argonne National Laboratory), Michael E. Papka (Argonne National Laboratory, Northern Illinois University)

Analysis of scientific simulation data can be concurrently executed with simulation either in time- or space-shared mode. This mitigates the I/O bottleneck. However it results in either stalling the simulation for performing the analysis or transferring data for analysis. In this paper, we improve the throughput of space-shared in situ analysis of large-scale simulations by topology-aware mapping and optimal process decomposition. We propose node interconnect topology-aware process placement for simulation and analysis to reduce the data movement time. We also present an integer linear program for optimal 3D decompositions of simulation and analysis processes. We demonstrate our approach using molecular dynamics simulation on Mira, Cori and Theta supercomputers. Our mapping schemes, combined with optimal 3D process decomposition and code optimizations resulted in up to 30% lower execution times for space-shared in situ analysis than the default approach. Our mappings also reduce MPI collective I/O times by 10-40%.

Best Paper Finalist: no
Best Student Paper Finalist: no

Wednesday, November 14th

Room: C140/142
10:30 am - 12:00 pm

MPI Optimization and Characterization

Cooperative Rendezvous Protocols for Improved Performance and Overlap

S. Chakraborty (Ohio State University), M. Bayatpour (Ohio State University), J. Hashmi (Ohio State University), H. Subramoni (Ohio State University), D. K. Panda (Ohio State University)

With the emergence of larger multi-/many-core clusters, performance of large message communication is becoming more important. MPI libraries use different Rendezvous protocols to perform large message communication. However, existing Rendezvous protocols do not consider the overall communication pattern and make optimal use of the Sender and the Receiver CPUs. In this work, we propose a cooperative Rendezvous protocol that can provide up to 2x improvement in intra-node bandwidth and latency for large messages. We also propose a scheme to dynamically choose the best Rendezvous protocol for each message based on the communication pattern. Finally, we show how these improvements can increase the overlap of computation with intra-node and inter-node communication, and lead to application level benefits. We evaluate proposed designs on three different architectures including Intel Xeon, Knights Landing, and OpenPOWER with different HPC applications and obtain benefits up to 19% with Graph500, 16% with CoMD, and 10% with MiniGhost.

Best Paper Finalist: no
Best Student Paper Finalist: yes

Framework for Scalable Intra-Node Collective Operations Using Shared Memory

Surabhi Jain (Intel Corporation), Rashid Kaleem (Intel Corporation), Marc Gamell Balmana (Intel Corporation), Akhil Langer (Intel Corporation), Dmitry Durnov (Intel Corporation), Alexander Sannikov (Intel Corporation), Maria Garzaran (Intel Corporation)

Collective operations are used in MPI programs to express common communication patterns, collective computations, or synchronizations. In many collectives, such as barrier or allreduce, the intra-node component of the collective is in the critical path, as the inter-node communication cannot start until the intra-node component has been executed. Thus, with increasing number of core counts in each node, intra-node optimizations that leverage the intra-node shared memory become increasingly important.

In this paper, we focus on the performance benefit of optimizing intra-node collectives using shared memory. We optimize several collectives using the primitives in broadcast and reduce as building blocks for other collectives. A comparison of our implementation on top of MPICH shows significant performance speedups with respect to the original MPICH implementation, MVAPICH, and OpenMPI, among others.

Best Paper Finalist: no
Best Student Paper Finalist: no

Characterization of MPI Usage on a Production Supercomputer

Sudheer Chunduri (Argonne National Laboratory), Scott Parker (Argonne National Laboratory), Pavan Balaji (Argonne National Laboratory), Kevin Harms (Argonne National Laboratory), Kalyan Kumaran (Argonne National Laboratory)

MPI is the most prominent programming model used in scientific computing today. espite it's importance, however, how scientific applications use it in production is not very well understood due

to the lack of low overhead profiling tools. We used a lightweight profiling tool, called autoperf, to log the MPI usage characteristics of production applications on a large supercomputing system (Mira) and its corresponding development system (Cetus). Autoperf limits the amount of information that it records in order to keep the overhead to a minimum while still storing enough data to derive useful insights. MPI usage statistics have been collected for over 100K jobs that were run within a 2-year period and are analyzed. The analysis of this data is intended as a mechanism to provide useful insights for MPI developers and network hardware developers for their next generation of improvements, and for supercomputing center operators for their next system procurements.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C141/143/149
10:30 am - 12:00 pm

Non-Volatile Memory

Runtime Data Management on Non-Volatile Memory-Based Heterogeneous Memory for Task-Parallel Programs

Kai Wu (University of California, Merced), Jie Ren (University of California, Merced), Dong Li (University of California, Merced)

Non-volatile memory (NVM) provides a scalable solution to replace DRAM as main memory. Because of relatively high latency and low bandwidth of NVM (comparing with DRAM), NVM often pairs with DRAM to build a heterogeneous main memory system (HMS). Deciding data placement on NVM-based HMS is critical to enable future NVM-based HPC. In this paper, we study task-parallel programs and introduce a runtime system to address the data placement problem on NVM-based HMS. Leveraging semantics and execution mode of task-parallel programs, we efficiently characterize memory access patterns of tasks and reduce data movement overhead. We also introduce a performance model to predict performance for tasks with various data placements on HMS. Evaluating with a set of HPC benchmarks, we show that our runtime system achieves higher performance than a conventional HMS-oblivious runtime (24% improvement on average) and two state-of-the-art HMS-aware solutions (16% and 11% improvement on average, respectively).

Best Paper Finalist: no
Best Student Paper Finalist: no

DRAGON: Breaking GPU Memory Capacity Limits with Direct NVM Access

Pak Markthub (Tokyo Institute of Technology), Mehmet E. Belviranli (Oak Ridge National Laboratory), Seyong Lee (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory), Satoshi Matsuoka (RIKEN, Tokyo Institute of Technology)

Heterogeneous computing with accelerators is growing in importance in high performance computing (HPC). Recently, application datasets have expanded beyond the memory capacity of these accelerators, and often beyond the capacity of their hosts. Meanwhile, nonvolatile memory (NVM) storage has emerged as a pervasive component in HPC systems because NVM provides massive

amounts of memory capacity at affordable cost. Currently, for accelerator applications to use NVM, they must manually orchestrate data movement across multiple memories and this approach only performs well for applications with simple access behaviors. To address this issue, we developed DRAGON, a solution that enables all classes of GP-GPU applications to transparently compute on terabyte datasets residing in NVM. DRAGON leverages the page-faulting mechanism on the recent NVIDIA GPUs by extending capabilities of CUDA Unified Memory (UM). Our experimental results show that DRAGON transparently expands memory capacity and obtain additional speedups via automated I/O and data transfer overlapping.

Best Paper Finalist: no
Best Student Paper Finalist: no

Siena: Exploring the Design Space of Heterogeneous Memory Systems
Ivy B. Peng (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory)

Memory systems are crucial to the performance, power, and cost of high-performance computing systems. Recently, multiple factors are driving the need for more complex, deep memory hierarchies. However, architects and customers are struggling to design memory systems that effectively balance multiple, often competing, factors in this large, multidimensional, and fast-moving design space. In this paper, we systematically explore the organization of heterogeneous memory systems on a framework, called Siena. Siena facilitates quick exploration of memory architectures with flexible configurations of memory systems and realistic memory workloads. We perform a design space exploration on 22 proposed memory systems using eight relevant workloads. Our results show that horizontal organizations of memories can achieve higher performance than that of vertical organizations when the distribution of memory traffic balances the performance gap between memories. However, the coupling effects through shared resources and application behaviors could negate the advantage of high-performance memory in horizontal organizations.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C146
10:30 am - 12:00 pm

Task-Based Programming

Dynamic Tracing: Memoization of Task Graphs for Dynamic Task-Based Runtimes
Wonchan Lee (Stanford University), Elliott Slaughter (SLAC National Accelerator Laboratory), Michael Bauer (Nvidia Corporation), Sean Treichler (Nvidia Corporation), Todd Warszawski (Stanford University), Michael Garland (Nvidia Corporation), Alex Aiken (Stanford University)

Many recent programming systems for both supercomputing and data center workloads generate task graphs to express computations that run on parallel and distributed machines. Due to the overhead associated with constructing these graphs the dependence analysis that generates them is often statically computed and memoized, and the resulting graph executed repeatedly at runtime. However, many applications require a dynamic dependence analysis due to data dependent behavior, but there

are new challenges in capturing and re-executing task graphs at runtime. In this work, we introduce dynamic tracing, a technique to capture a dynamic dependence analysis of a trace that generates a task graph, and replay it. We show that an implementation of dynamic tracing improves strong scaling by an average of 4.9X and up to 7.0X on a suite of already optimized benchmarks.

Best Paper Finalist: no
Best Student Paper Finalist: no

Runtime-Assisted Cache Coherence Deactivation in Task Parallel Programs

Paul Caheny (Barcelona Supercomputing Center, Polytechnic University of Catalonia), Lluc Alvarez (Barcelona Supercomputing Center), Mateo Valero (Barcelona Supercomputing Center, Polytechnic University of Catalonia), Miquel Moretó (Barcelona Supercomputing Center, Polytechnic University of Catalonia), Marc Casas (Barcelona Supercomputing Center)

With increasing core counts, the scalability of directory-based cache coherence has become a challenging problem. To reduce the area and power needs of the directory, recent proposals reduce its size by classifying data as private or shared, and disable coherence for private data. However, existing classification methods suffer from inaccuracies and require complex hardware support with limited scalability.

This paper proposes a hardware/software co-designed approach: the runtime system identifies data that is guaranteed by the programming model semantics to not require coherence and notifies the microarchitecture. The microarchitecture deactivates coherence for this private data and powers off unused directory capacity. Our proposal reduces directory accesses to just 26% of the baseline system and supports a 64× smaller directory with only 2.8% performance degradation. By dynamically calibrating the directory size, our proposal saves 86% of dynamic energy consumption in the directory without harming performance.

Best Paper Finalist: no
Best Student Paper Finalist: no

A Divide and Conquer Algorithm for DAG Scheduling Under Power Constraints

Gökalp Demirci (University of Chicago), Ivana Marincic (University of Chicago), Henry Hoffmann (University of Chicago)

We consider the problem of scheduling a parallel computation–represented as a directed acyclic graph (DAG)–on a distributed parallel system with a global resource constraint–specifically a global power budget–and configurable resources, allowing a range of different power/performance tradeoffs. There is a rich body of literature on the independent problems of (1) scheduling DAGs and (2) scheduling independent applications under resource constraints. Very little, however, is known about the combined problem of scheduling DAGs under resource constraints. We present a novel approximation algorithm using a divide-and-conquer method for minimizing application execution time. We prove that the length of the schedule returned by our algorithm is always within O(log n)-factor of the optimum that can be achieved with selection of configurations for the tasks. We implement and test our algorithm on simulations of real application DAGs. We find that our divide-and-conquer method improves performance by up to 75% compared to greedy scheduling algorithms.

Best Paper Finalist: no

Best Student Paper Finalist: no

Room: C140/142
1:30 pm - 3:00 pm

Physics and Tensor Applications

Simulating the Wenchuan Earthquake with Accurate Surface Topography on Sunway TaihuLight
Bingwei Chen (Tsinghua University; National Supercomputing Center, Wuxi), Haohuan Fu (Tsinghua University; National Supercomputing Center, Wuxi), Yanwen Wei (Tsinghua University; National Supercomputing Center, Wuxi), Conghui He (Tsinghua University; National Supercomputing Center, Wuxi), Wenqiang Zhang (University of Science and Technology of China), Yuxuan Li (Tsinghua University; National Supercomputing Center, Wuxi), Wubin Wan (National Supercomputing Center, Wuxi), Wei Zhang (National Supercomputing Center, Wuxi), Lin Gan (Tsinghua University; National Supercomputing Center, Wuxi), Wei Zhang (Southern University of Science and Technology, China), Zhenguo Zhang (Southern University of Science and Technology, China), Guangwen Yang (Tsinghua University; National Supercomputing Center, Wuxi), Xiaofei Chen (Southern University of Science and Technology, China)

This paper reports our efforts on performing 50-m resolution earthquake simulation of the Wenchuan Earthquake (Ms 8.0, China) on Sunway TaihuLight. To accurately capture the surface topography, we adopt a curvilinear grid finite-difference method with a traction image free surface implementation and redesign the algorithm to reduce memory access costs for heterogeneous many-core architectures. We then derive a performance model of our algorithm to guide and drive the further optimization and tuning of various parameters using a genetic algorithm. A data layout transformation is also proposed to improve the direct memory access (DMA) efficiency further. Our efforts improve the simulation efficiency from 0.05% to 7.6%, with a sustained performance of 9.07 Pflops using the entire machine of the Sunway TaihuLight (over 10 million cores), and a large-scale simulation of the Wenchuan earthquake with accurate surface topography and improved coda wave effects.

Best Paper Finalist: no
Best Student Paper Finalist: no

Accelerating Quantum Chemistry with Vectorized and Batched Integrals
Hua Huang (Georgia Institute of Technology), Edmond Chow (Georgia Institute of Technology)

This paper presents the first quantum chemistry calculations using a recently developed vectorized library for computing electron repulsion integrals. To lengthen the SIMD loop and thus improve SIMD utilization, the approach used in this paper is to batch together the computation of multiple integrals that have the same code path. The standard approach is to compute integrals one at a time, and thus a batching procedure had to be developed. This paper shows proof-of-concept and demonstrates the performance gains possible when the batched approach is used. Batching also enables certain optimizations when the integrals are used to compute the Fock matrix. We further describe several other optimizations that were needed to obtain up to a 270% speedup over the no batching version of the code, making a compelling case for adopting the presented techniques in quantum chemistry software.

Best Paper Finalist: no
Best Student Paper Finalist: no

High-Performance Dense Tucker Decomposition on GPU Clusters
Jee Choi (IBM), Xing Liu (Intel Corporation), Venkatesan Chakaravarthy (IBM)

The Tucker decomposition method is one of the most popular algorithms for analyzing and compressing data with multi-way relationship. Its execution time is typically dominated by dense matrix multiplication, which makes it well-suited for GPU acceleration. State-of-the-art distributed dense Tucker implementations for CPU clusters adopt multi-dimensional partitioning that optimizes for storage and communication. This, however, leads to smaller matrix dimensions that result in under-utilizing the GPU.

In this paper, we present our optimized implementation and performance analysis of dense Tucker decomposition on a multi-GPU cluster. We propose three optimizations: a new partitioning strategy that improves GPU performance, a new tensor matricization layout that halves the number of communication/matricization steps, and a variation of the randomized SVD algorithm to overcome the eigenvalue bottleneck that arises from the high speedups gained from GPU acceleration. Our GPU implementation employing all three optimizations achieves up to 11.8x speedup on 64 nodes over state-of-the-art TuckerMPI.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C141/143/149
1:30 pm - 3:00 pm

Clouds and Distributed Computing

A Reference Architecture for Datacenter Scheduling: Design, Validation, and Experiments
Georgios Andreadis (Delft University of Technology, Vrije University Amsterdam), Laurens Versluis (Vrije University Amsterdam), Fabian Mastenbroek (Delft University of Technology), Alexandru Iosup (Vrije University Amsterdam, Delft University of Technology)

Datacenters act as cloud-infrastructure to stakeholders across industry, government, and academia. To meet growing demand yet operate efficiently, datacenter operators employ increasingly more sophisticated scheduling systems, mechanisms, and policies. Although many scheduling techniques already exist, relatively little research has gone into the abstraction of the scheduling process itself, hampering design, tuning, and comparison of existing techniques. In this work, we propose a reference architecture for datacenter schedulers. The architecture follows five design principles: components with clearly distinct responsibilities, grouping of related components where possible, separation of mechanism from policy, scheduling as complex workflow, and hierarchical multi-scheduler structure. To demonstrate the validity of the reference architecture, we map to it state-of-the-art datacenter schedulers. We find scheduler-stages are commonly underspecified in peer-reviewed publications. Through trace-based simulation and real-world experiments, we show underspecification of scheduler-stages can lead to significant variations in performance.

Best Paper Finalist: no
Best Student Paper Finalist: no

Dynamically Negotiating Capacity Between On-Demand and Batch Clusters
Feng Liu (University of Minnesota), Kate Keahey (Argonne National Laboratory), Pierre Riteau (University of Chicago), Jon Weissman (University of Minnesota)

In the era of rapid experimental expansion data analysis needs are rapidly outpacing the capabilities of small institutional clusters and looking to integrate HPC resources into their workflow. We propose one way of reconciling on-demand needs of experimental analytics with the batch managed HPC resources within a system that dynamically moves nodes between an on-demand cluster configured with cloud technology (OpenStack) and a traditional HPC cluster managed by a batch scheduler (Torque). We evaluate this system experimentally both in the context of real-life traces representing two years of a specific institutional need, and via experiments in the context of synthetic traces that capture generalized characteristics of potential batch and on-demand workloads. Our results for the real-life scenario show that our approach could reduce the current investment in on-demand infrastructure by 82% while at the same time improving the mean batch wait time almost by an order of magnitude (8x).

Best Paper Finalist: no
Best Student Paper Finalist: no

A Lightweight Model for Right-Sizing Master-Worker Applications
Nathaniel Kremer-Herman (University of Notre Dame), Benjamin Tovar (University of Notre Dame), Douglas Thain (University of Notre Dame)

When running a parallel application at scale, a resource provisioning policy should minimize over-commitment (idle resources) and under-commitment (resource contention). However, users seldom know the quantity of resources to appropriately execute their application. Even with such knowledge, over- and under-commitment of resources may still occur because the application does not run in isolation. It shares resources such as network and filesystems.

We formally define the capacity of a parallel application as the quantity of resources that may effectively be provisioned for the best execution time in an environment. We present a model to compute an estimate of the capacity of master-worker applications as they run based on execution and data-transfer times. We demonstrate this model with two bioinformatics workflows, a machine learning application, and one synthetic application. Our results show the model correctly tracks the known value of capacity in scaling, dynamic task behavior, and with improvements in task throughput.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C146
1:30 pm - 3:00 pm

Resilience 2

Lessons Learned from Memory Errors Observed Over the Lifetime of Cielo

Scott Levy (Sandia National Laboratories), Kurt B. Ferreira (Sandia National Laboratories), Nathan DeBardeleben (Los Alamos National Laboratory), Taniya Siddiqua (Advanced Micro Devices Inc), Vilas Sridharan (Advanced Micro Devices Inc), Elisabeth Baseman (Los Alamos National Laboratory)

Maintaining the performance of high-performance computing (HPC) applications as failures increase is a major challenge for next-generation extreme-scale systems. Recent research demonstrates that hardware failures are expected to become more common due to increased component counts, reduced device-feature sizes, and tightly-constrained power budgets. Few existing studies, however, have examined failures in the context of the entire lifetime of a single platform. In this paper, we analyze failure data collected over the entire lifetime of Cielo, a leadership-class HPC system. Our analysis reveals several key findings, including: (i) Cielo's memory (DRAM and SRAM) exhibited no discernible aging effects; (ii) correctable memory faults are not predictive of future uncorrectable memory faults; (iii) developing more comprehensive logging facilities will improve failure analysis on future machines; (iv) continued advances will be required to ensure current failure mitigation techniques remain a viable option for future platforms.

Best Paper Finalist: no
Best Student Paper Finalist: no

Partial Redundancy in HPC Systems with Non-Uniform Node Reliabilities

Zaeem Hussain (University of Pittsburgh), Taieb Znati (University of Pittsburgh), Rami Melhem (University of Pittsburgh)

We study the usefulness of partial redundancy in HPC message passing systems where individual node failure distributions are not identical. Prior research works on fault tolerance have generally assumed identical failure distributions for the nodes of the system. In such settings, partial replication has never been shown to outperform the two extremes (full and no-replication) for any significant range of node counts. We argue that partial redundancy may provide the best performance under the more realistic assumption of non-identical node failure distributions. We provide theoretical results on arranging nodes with different reliability values among replicas such that system reliability is maximized. Moreover, using system reliability to compute MTTI (mean-time-to-interrupt) and expected completion time of a partially replicated system, we numerically determine the optimal partial replication degree. Our results indicate that partial replication can be a more efficient alternative to full replication at system scales where Checkpoint/Restart alone is not sufficient.

Best Paper Finalist: no
Best Student Paper Finalist: no

Evaluating and Accelerating High-Fidelity Error Injection for HPC

Chun-Kai Chang (University of Texas), Sangkug Lym (University of Texas), Nicholas Kelly (University of Texas), Michael B. Sullivan (Nvidia Corporation), Mattan Erez (University of Texas)

We address two important concerns in the analysis of the behavior of applications in the presence of hardware errors: (1) when is it important to model how hardware faults lead to erroneous values (instruction-level errors) with high fidelity, as opposed to using simple bit-flipping models, and (2) how to enable fast high-fidelity error injection campaigns, in particular when error detectors are employed. We present and verify a new nested Monte Carlo methodology for evaluating high-fidelity gate-level

fault models and error-detector coverage, which is orders of magnitude faster than current approaches. We use that methodology to demonstrate that, without detectors, simple error models suffice for evaluating errors in 9 HPC benchmarks.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C140/142
3:30 pm - 5:00 pm

Large Scale System Deployments

The Design, Deployment, and Evaluation of the CORAL Pre-Exascale Systems
Sudharshan S. Vazhkudai (Oak Ridge National Laboratory), Bronis R. de Supinski (Lawrence Livermore National Laboratory), Arthur S. Bland (Oak Ridge National Laboratory), Al Geist (Oak Ridge National Laboratory), James Sexton (IBM), Jim Kahle (IBM), Christopher J. Zimmer (Oak Ridge National Laboratory), Scott Atchley (Oak Ridge National Laboratory), Sarp H. Oral (Oak Ridge National Laboratory), Don E. Maxwell (Oak Ridge National Laboratory), Veronica G. Vergara Larrea (Oak Ridge National Laboratory), Adam Bertsch (Lawrence Livermore National Laboratory), Robin Goldstone (Lawrence Livermore National Laboratory), Wayne Joubert (Oak Ridge National Laboratory), Chris Chambreau (Lawrence Livermore National Laboratory), David Appelhans (IBM), Robert Blackmore (IBM), Ben Casses (Lawrence Livermore National Laboratory), George Chochia (IBM), Gene Davison (IBM), Matthew A. Ezell (Oak Ridge National Laboratory), Tom Gooding (IBM), Elsa Gonsiorowski (Lawrence Livermore National Laboratory), Leopold Grinberg (IBM), Bill Hanson (IBM), Bill Hartner (IBM), Ian Karlin (Lawrence Livermore National Laboratory), Matthew L. Leininger (Lawrence Livermore National Laboratory), Dustin Leverman (Oak Ridge National Laboratory), Chris Marroquin (IBM), Adam Moody (Lawrence Livermore National Laboratory), Martin Ohmacht (IBM), Ramesh Pankajakshan (Lawrence Livermore National Laboratory), Fernando Pizzano (IBM), James H. Rogers (Oak Ridge National Laboratory), Bryan Rosenburg (IBM), Drew Schmidt (Oak Ridge National Laboratory), Mallikarjun Shankar (Oak Ridge National Laboratory), Feiyi Wang (Oak Ridge National Laboratory), Py Watson (Lawrence Livermore National Laboratory), Bob Walkup (IBM), Lance D. Weems (Lawrence Livermore National Laboratory), Junqi Yin (Oak Ridge National Laboratory)

CORAL, the Collaboration of Oak Ridge, Argonne and Livermore, is fielding two similar IBM systems, Summit and Sierra, with NVIDIA GPUs that will replace the existing Titan and Sequoia systems. Summit and Sierra are currently ranked No. 1 and No. 3, respectively, on the Top500 list. We discuss the design and key differences of the systems. Our evaluation of the systems highlights the following. Applications that fit in HBM see the most benefit and may prefer more GPUs; however, for some applications, the CPU-GPU bandwidth is more important than the number of GPUs. The node-local burst buffer scales linearly, and can achieve a 4X improvement over the parallel file system for large jobs; smaller jobs, however, may benefit from writing directly to the PFS. Finally, several CPU, network and memory bound analytics and GPU-bound deep learning codes achieve up to a 11X and 79X speedup/node, respectively over Titan.

Best Paper Finalist: no
Best Student Paper Finalist: no

Best Practices and Lessons from Deploying and Operating a Sustained-Petascale System: The Blue Waters Experience

Gregory H. Bauer (University of Illinois, National Center for Supercomputing Applications), Brett Bode (University of Illinois, National Center for Supercomputing Applications), Jeremy Enos (University of Illinois, National Center for Supercomputing Applications), William T. Kramer (University of Illinois, National Center for Supercomputing Applications), Scott Lathrop (University of Illinois, National Center for Supercomputing Applications), Celso L. Mendes (University of Illinois, National Center for Supercomputing Applications), Roberto R. Sisneros (University of Illinois, National Center for Supercomputing Applications)

Building and operating versatile extreme-scale computing systems that work productively for a range of frontier research domains present many challenges and opportunities. Solutions created, experiences acquired, and lessons learned, while rarely published, could drive the development of new methods and practices and raise the bar for all organizations supporting research, scholarship, and education. This paper describes the methods and procedures developed for deploying, supporting, and continuously improving the Blue Waters system and its services during the last five years. Being the first US sustained-petascale computing platform available to the open-science community, the Blue Waters project pioneered various unique practices that we are sharing to be adopted and further improved by the community. We present our support and service methodologies, and the leadership practices employed for ensuring that the system stays highly efficient and productive. We also provide the return on investment summaries related to deploying and operating the system.

Best Paper Finalist: no
Best Student Paper Finalist: no

Performance Evaluation of a Vector Supercomputer SX-Aurora TSUBASA

Kazuhiko Komatsu (Tohoku University), Shintaro Momose (Tohoku University, NEC Corporation), Yoko Isobe (Tohoku University, NEC Corporation), Osamu Watanabe (Tohoku University, NEC Corporation), Akihiro Musa (Tohoku University, NEC Corporation), Mitsuo Yokokawa (Kobe University), Toshikazu Aoyama (NEC Corporation), Masayuki Sato (Tohoku University), Hiroaki Kobayashi (Tohoku University)

A new SX-Aurora TSUBASA vector supercomputer has been released with a new system architecture and a new execution model to achieve high sustained performance, especially for memory-intensive applications. In SX-Aurora TSUBASA, the vector host (VH) of a standard x86 Linux node is attached to the vector engine (VE) of a newly developed vector processor. An application is executed on the VE, and only system calls are offloaded to the VH. This new execution model can avoid redundant data transfers between a VH and a VE that can easily become a bottleneck in the conventional execution model. This paper examines the potential of SX-Aurora TSUBASA. First, the basic performance of SX-Aurora TSUBASA is clarified by evaluating benchmark programs. Then, the effectiveness of the new execution model is examined by using a microbenchmark. Finally, the high potential of SX-Aurora TSUBASA is clarified through evaluations of practical applications.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C141/143/149
3:30 pm - 5:00 pm

Arithmetic and Optimization

Associative Instruction Reordering to Alleviate Register Pressure

Prashant Singh Rawat (Ohio State University), Aravind Sukumaran-Rajam (Ohio State University), Atanas Rountev (Ohio State University), Fabrice Rastello (French Institute for Research in Computer Science and Automation (INRIA)), Louis-Noel Pouchet (Colorado State University), P. Sadayappan (Ohio State University)

Register allocation is generally considered a practically solved problem. For most applications, the register allocation strategies in production compilers are very effective in controlling the number of loads/stores and register spills. However, existing register allocation strategies are not effective and result in excessive register spilling for computation patterns with a high degree of many-to-many data reuse, e.g., high-order stencils and tensor contractions. We develop a source-to-source instruction reordering strategy that exploits the flexibility of reordering associative operations to alleviate register pressure. The developed transformation module implements an adaptable strategy that can appropriately control the degree of instruction-level parallelism, while relieving register pressure. The effectiveness of the approach is demonstrated through experimental results using multiple production compilers (GCC, Clang/LLVM) and target platforms (Intel Xeon Phi, and Intel x86 multi-core).

Best Paper Finalist: no
Best Student Paper Finalist: no

Harnessing GPU's Tensor Cores Fast FP16 Arithmetic to Speedup Mixed-Precision Iterative Refinement Solvers

Azzam Haidar (University of Tennessee, Innovative Computing Laboratory), Stan Tomov (University of Tennessee), Jack Dongarra (University of Tennessee), Nicholas Higham (University of Manchester, School of Mathematics)

The use of low-precision arithmetic in computing methods has been a powerful tool to accelerate numerous scientific computing applications including Artificial Intelligence. We present an investigation showing that other HPC applications can harness this power too, and in particular, the general HPC problem of solving Ax = b, where A is a large dense matrix, and the solution is needed in FP64 accuracy. Our approach is based on the mixed-precision (FP16->FP64) iterative refinement technique – we generalize and extend prior advances into a framework, for which we develop architecture-specific algorithms and highly-tuned implementations where we show how the use of FP16-TC (tensor cores) arithmetic can provide up to 4X speedup and improve the energy consumption by a factor of 5 achieving 74 Gflop/Watt. This is due to the performance boost that the FP16 (Tensor Cores) provide and to its better accuracy that outperforms the classical FP16.

Best Paper Finalist: no
Best Student Paper Finalist: no

ADAPT: Algorithmic Differentiation Applied to Floating-Point Precision Tuning

Harshitha Menon (Lawrence Livermore National Laboratory), Michael O. Lam (James Madison University, Lawrence Livermore National Laboratory), Daniel Osei-Kuffuor (Lawrence Livermore

National Laboratory), Markus Schordan (Lawrence Livermore National Laboratory), Scott Lloyd (Lawrence Livermore National Laboratory), Kathryn Mohror (Lawrence Livermore National Laboratory), Jeffrey Hittinger (Lawrence Livermore National Laboratory)

HPC applications extensively use floating point arithmetic operations to solve computational problems in various domains. Mixed precision computing, use of lowest precision data type sufficient to achieve a desired accuracy, have been explored to improve performance, reduce power consumption and data movement. Manually optimizing the program to use mixed precision is challenging. In this work, we present ADAPT, an approach for mixed precision analysis on HPC workloads while providing guarantees about the final output error. Our approach uses algorithmic differentiation to accurately estimate the output error for mixed precision configuration. ADAPT provides floating-point precision sensitivity of programs, which highlights regions of the code that that can potentially be converted to lower precision, is used to make algorithmic choices and develop mixed precision configurations. We evaluate ADAPT on six benchmarks and a proxy application and show that we are able to achieve a speedup of 1.2x on the proxy application, LULESH.

Best Paper Finalist: no
Best Student Paper Finalist: no

Thursday, November 15th

Room: C140/142
10:30 am - 12:00 pm

Graph Algorithms and Systems

iSpan: Parallel Identification of Strongly Connected Components with Spanning Trees
Yuede Ji (George Washington University), Hang Liu (University of Massachusetts, Lowell), H. Howie Huang (George Washington University)

Detecting strongly connected components (SCCs) in a directed graph is crucial for understanding the structure of graphs. Most real-world graphs have one large SCC that contains the majority of the vertices, and many small SCCs whose sizes are reversely proportional to the frequency of their occurrence. For both types of SCCs, current approaches that rely on depth or breadth first search (DFS or BFS) face the challenges of strict synchronization requirement and high computation cost. In this paper, we advocate a new paradigm of identifying SCCs with simple spanning trees, since SCC detection requires only the knowledge of connectivity among the vertices. We have developed a prototype called iSpan which consists of parallel, relaxed synchronization construction of spanning trees for detecting the large and small SCCs. The evaluations show that iSpan is able to significantly outperform current state-of-the-art DFS and BFS-based methods by average 18× and 4×, respectively.

Best Paper Finalist: no
Best Student Paper Finalist: no

Adaptive Anonymization of Data with b-Edge Covers
Arif Khan (Pacific Northwest National Laboratory), Krzysztof Choromanski (Google LLC), Alex Pothen

(Purdue University), S M Ferdous (Purdue University), Mahantesh Halappanavar (Pacific Northwest National Laboratory), Antonino Tumeo (Pacific Northwest National Laboratory)

We explore the problem of sharing data that pertains to individuals with anonymity guarantees, where each user requires a desired level of privacy. We propose the first shared-memory as well as distributed memory parallel algorithms for the adaptive anonymity problem that achieves this goal, and produces high quality anonymized datasets.

The new algorithm is based on an optimization procedure that iteratively computes weights on the edges of a dissimilarity matrix, and at each iteration computes a minimum weighted b-Edge cover in the graph. We are able to solve adaptive anonymity problems with hundreds of thousands of instances and hundreds of features on a leadership-class supercomputer in under five minutes. Our algorithm scales up to 4K cores on a distributed memory supercomputer, while also providing good speedups on shared memory multiprocessors. On smaller problems, where an algorithm based on Belief Propagation is feasible, our algorithm is two orders of magnitude faster.

Best Paper Finalist: no
Best Student Paper Finalist: no

faimGraph: High Performance Management of Fully-Dynamic Graphs Under Tight Memory Constraints on the GPU

Martin Winter (Graz University of Technology), Daniel Mlakar (Graz University of Technology), Rhaleb Zayer (Max Planck Institute for Informatics), Hans-Peter Seidel (Max Planck Institute for Informatics), Markus Steinberger (Graz University of Technology, Max Planck Institute for Informatics)

In this paper, we present a fully-dynamic graph data structure for the Graphics Processing Unit (GPU). It delivers high update rates while keeping a low memory footprint using autonomous memory management directly on the GPU. The data structure is fully-dynamic, allowing not only for edge but also vertex updates. Performing the memory management on the GPU allows for fast initialization times and efficient update procedures without additional intervention or reallocation procedures from the host. faimGraph is the first GPU graph framework that fully reclaims unused memory, permitting long time application with highly changing graph structures. Performance evaluations show that our approach outperforms that previous state-of-the-art in for all types of graph updates. Furthermore, evaluate algorithmic performance using a PageRank and a Static Triangle Counting (STC) implementation, demonstrating the suitability of the framework even for memory access intensive algorithms.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C141/143/149
10:30 am - 12:00 pm

Programming Systems Tools

Dynamic Data Race Detection for OpenMP Programs

Yizi Gu (Rice University), John Mellor-Crummey (Rice University)

Two concurrent accesses to a shared variable that are unordered by synchronization are said to be a data race if at least one access is a write. Data races cause shared memory parallel programs to behave unpredictably. This paper describes ROMP -- a tool for detecting data races in executions of scalable parallel applications that employ OpenMP for node-level parallelism. The complexity of OpenMP, which includes primitives for managing data environments, SPMD and SIMD parallelism, work sharing, tasking, mutual exclusion, and ordering, presents a formidable challenge for data race detection. ROMP is a hybrid data race detector that tracks accesses, access orderings, and mutual exclusion. Unlike other OpenMP race detectors, ROMP detects races with respect to logical parallelism rather than implementation threads. Experiments show that ROMP yields precise race reports for a broader set of OpenMP constructs than prior state-of-the-art race detectors.

Best Paper Finalist: no
Best Student Paper Finalist: no

ParSy: Inspection and Transformation of Sparse Matrix Computations for Parallelism
Kazem Cheshmi (University of Toronto), Shoaib Kamil (Adobe Research), Michelle Mills Strout (University of Arizona), Maryam Mehri Dehnavi (University of Toronto)

In this work, we describe ParSy, a framework that uses a novel inspection strategy along with a simple code transformation to optimize parallel sparse algorithms for shared memory processors. Unlike existing approaches that can suffer from load imbalance and excessive synchronization, ParSy uses a novel task coarsening strategy to create well-balanced tasks that can execute in parallel, while maintaining locality of memory accesses. Code using the ParSy inspector and transformation outperforms existing highly-optimized sparse matrix algorithms such as Cholesky factorization on multi-core processors with speedups of 2.8× and 3.1× over the MKL Pardiso and PaStiX libraries respectively.

Best Paper Finalist: no
Best Student Paper Finalist: no

Detecting MPI Usage Anomalies via Partial Program Symbolic Execution
Fangke Ye (Georgia Institute of Technology), Jisheng Zhao (Georgia Institute of Technology), Vivek Sarkar (Georgia Institute of Technology)

MPI is a message passing based programming model for distributed-memory parallelism that has been had been widely used for programming supercomputers for over 25 years. However, debugging and verification of MPI programs is widely recognized to be a deep technical challenge. This challenge is further exacerbated by a recent increase in the use of nonblocking MPI operations that bring new classes of bugs related to data races.

In this paper, we introduce a new MPI program debugging approach based on partial symbolic execution so as to avoid the false alarms inherent in the static analysis based methodology. Compared with the dynamic approach, our approach can be applied to incomplete programs and explore multiple execution paths, thereby bringing more flexibility and precision. By comparing with well known static/dynamic tools on real-world MPI applications, our approach shows same precision as the dynamic tool and avoids false positive produced by the static tool.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C140/142
1:30 pm - 3:00 pm

Deep Learning

Exploring Flexible Communications for Streamlining DNN Ensemble Training Pipelines
Randall Pittman (North Carolina State University), Hui Guan (North Carolina State University), Xipeng Shen (North Carolina State University), Seung-Hwan Lim (Oak Ridge National Laboratory), Robert M. Patton (Oak Ridge National Laboratory)

Parallel training of a Deep Neural Network (DNN) ensemble on a cluster of nodes is a common practice to train multiple models in order to construct a model with a higher prediction accuracy. Existing ensemble training pipelines can perform a great deal of redundant operations, resulting in unnecessary CPU usage, or even poor pipeline performance. In order to remove these redundancies, we need pipelines with more communication flexibility than existing DNN frameworks provide.

This project investigates a series of designs to improve pipeline flexibility and adaptivity, while also increasing performance. We implement our designs using Tensorflow with Horovod, and test it using several large DNNs. Our results show that the CPU time spent during training is reduced by 2-11X. Furthermore, our implementation can achieve up to 10X speedups when CPU core limits are imposed. Our best pipeline also reduces the average power draw of the ensemble training process by 5-16%.

Best Paper Finalist: no
Best Student Paper Finalist: no

CosmoFlow: Using Deep Learning to Learn the Universe at Scale
Amrita Mathuriya (Intel Corporation), Deborah Bard (National Energy Research Scientific Computing Center (NERSC), Lawrence Berkeley National Laboratory), Pete Mendygral (Cray Inc), Lawrence Meadows (Intel Corporation), James Arnemann (University of California, Berkeley), Lei Shao (Intel Corporation), Siyu He (Carnegie Mellon University), Tuomas Karna (Intel Corporation), Diana Moise (Cray Inc), Simon J. Pennycook (Intel Corporation), Kristyn Maschhoff (Cray Inc), Jason Sewall (Intel Corporation), Nalini Kumar (Intel Corporation), Shirley Ho (Lawrence Berkeley National Laboratory, Carnegie Mellon University), Michael F. Ringenburg (Cray Inc), Mr Prabhat (Lawrence Berkeley National Laboratory, National Energy Research Scientific Computing Center (NERSC)), Victor Lee (Intel Corporation)

Deep learning is a promising tool to determine the physical model that describes our universe. To handle the considerable computational cost of this problem, we present CosmoFlow: a highly scalable deep learning application built on top of the TensorFlow framework.

CosmoFlow uses efficient implementations of 3D convolution and pooling primitives, together with improvements in threading for many element-wise operations, to improve training performance on Intel Xeon Phi processors. We also utilize the Cray PE Machine Learning Plugin for efficient scaling to

multiple nodes. We demonstrate fully synchronous data-parallel training on 8192 nodes of Cori with 77% parallel efficiency, achieving 3.5 Pflop/s sustained performance.

To our knowledge, this is the first large-scale science application of the TensorFlow framework at supercomputer scale with fully-synchronous training. These enhancements enable us to process large 3D dark matter distribution and predict the cosmological parameters Omega_M, sigma_8 and N_s with unprecedented accuracy.

Best Paper Finalist: no
Best Student Paper Finalist: no

Anatomy of High-Performance Deep Learning Convolutions on SIMD Architectures
Evangelos Georganas (Intel Corporation), Sasikanth Avancha (Intel Corporation), Kunal Banerjee (Intel Corporation), Dhiraj Kalamkar (Intel Corporation), Greg Henry (Intel Corporation), Hans Pabst (Intel Corporation), Alexander Heinecke (Intel Corporation)

Convolution layers are prevalent in many classes of deep neural networks, including Convolutional Neural Networks (CNNs) which provide state-of-the-art results for tasks like image recognition, neural machine translation, and speech recognition. The computationally expensive nature of a convolution operation has led to the proliferation of implementations including matrix-matrix multiplication formulation, and direct convolution primarily targeting GPUs. In this paper, we introduce direct convolution kernels for x86 architectures, in particular for Xeon and Xeon Phi systems, which are implemented via a dynamic compilation approach. Our JIT-based implementation shows close to theoretical peak performance, depending on the setting and the CPU architecture at hand. We additionally demonstrate how these JIT-optimized kernels can be integrated into a light-weight multi-node graph execution model. This illustrates that single- and multi-node runs yield high efficiencies and high image-throughputs when executing state of the art image recognition tasks on CPUs.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C141/143/149
1:30 pm - 3:00 pm

Resilience 3: GPUs

Optimizing Software-Directed Instruction Replication for GPU Error Detection
Abdulrahman Mahmoud (University of Illinois), Siva Kumar Sastry Hari (Nvidia Corporation), Michael B. Sullivan (Nvidia Corporation), Timothy Tsai (Nvidia Corporation), Stephen W. Keckler (Nvidia Corporation)

Application execution on safety-critical and high-performance computer systems must be resilient to transient errors. As GPUs become more pervasive in such systems, they must supplement ECC/parity for major storage structures with reliability techniques that cover more of the GPU hardware logic. Instruction duplication has been explored for CPU resilience; however, it has never been studied in the context of GPUs, and it is unclear whether the performance and design choices it presents makes it a

feasible GPU solution. This paper describes a practical methodology to employ instruction duplication for GPUs and identifies implementation challenges that can incur high overheads (69% on average). It explores GPU-specific software optimizations that trade fine-grained recoverability for performance. It also proposes simple ISA extensions with limited hardware changes and area costs to further improve performance, cutting the runtime overheads by more than half to an average of 30%.

Best Paper Finalist: no
Best Student Paper Finalist: no

Fault Tolerant One-Sided Matrix Decompositions on Heterogeneous Systems with GPUs

Jieyang Chen (University of California, Riverside), Hongbo Li (University of California, Riverside), Sihuan Li (University of California, Riverside), Xin Liang (University of California, Riverside), Panruo Wu (University of Houston), Dingwen Tao (University of Alabama), Kaiming Ouyang (University of California, Riverside), Yuanlai Liu (University of California, Riverside), Kai Zhao (University of California, Riverside), Qiang Guan (Kent State University), Zizhong Chen (University of California, Riverside)

Current algorithm-based fault tolerance (ABFT) approach for one-sided matrix decomposition on heterogeneous systems with GPUs have following limitations: (1) they do not provide sufficient protection as most of them only maintain checksum in one dimension; (2) their checking scheme is not efficient due to redundant checksum verifications; (3) they fail to protect PCIe communication; (4) the checksum calculation based on a special type of matrix multiplication is far from efficient. By overcoming the above limitations, we design an efficient ABFT approach providing stronger protection for one-sided matrix decomposition methods on heterogeneous systems. First, we provide full matrix protection by using checksums in two dimensions. Second, our checking scheme is more efficient by prioritizing the checksum verification according to the sensitivity of matrix operations to soft errors. Third, we protect PCIe communication by reordering checksum verifications and decomposition steps. Fourth, we accelerate the checksum calculation by 1.7x via better utilizing GPUs.

Best Paper Finalist: no
Best Student Paper Finalist: no

PRISM: Predicting Resilience of GPU Applications Using Statistical Methods

Charu Kalra (Northeastern University), Fritz Previlon (Northeastern University), Xiangyu Li (Northeastern University), Norman Rubin (Nvidia Corporation), David Kaeli (Northeastern University)

As Graphics Processing Units (GPUs) become more pervasive in HPC and safety-critical domains, ensuring that GPU applications can be protected from data corruption grows in importance. Despite prior efforts to mitigate errors, we still lack a clear understanding of how resilient these applications are in the presence of transient faults. Due to the random nature of these faults, predicting whether they will alter the program output is a challenging problem. In this paper, we build a framework named PRISM, which uses a systematic approach to predict failures in GPU programs. PRISM extracts micro-architecture agnostic features to characterize program resiliency, which serve as predictors in our statistical model. PRISM enables us to predict failures in applications without running exhaustive fault-injection campaigns on a GPU, thereby reducing the error estimation effort. PRISM can also be used to gain insight into potential architectural support required to improve the reliability of GPU applications.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C140/142
3:30 pm - 5:00 pm

Astrophysics Applications

Phase Asynchronous AMR Execution for Productive and Performant Astrophysical Flows
Muhammad Nufail Farooqi (Koc University), Tan Nguyen (Lawrence Berkeley National Laboratory), Weiqun Zhang (Lawrence Berkeley National Laboratory), Ann S. Almgren (Lawrence Berkeley National Laboratory), John Shalf (Lawrence Berkeley National Laboratory), Didem Unat (Koc University)

Adaptive Mesh Refinement (AMR) is an approach to solving PDEs that reduces the computational and memory requirements at the expense of increased communication. Although adopting asynchronous execution can overcome communication issues, manually restructuring an AMR application to realize asynchrony is extremely complicated and hinders readability and long-term maintainability. To balance performance against productivity, we design a user-friendly API and adopt phase asynchronous execution model where all subgrids at an AMR level can be computed asynchronously.

We apply the phase asynchrony to transform a real-world AMR application, CASTRO, which solves multicomponent compressible hydrodynamic equations for astrophysical flows. We evaluate the performance and programming effort required to use our carefully designed API and execution model for transitioning large legacy codes from synchronous to asynchronous execution up to 278,528 Intel-KNL cores. CASTRO is about 100K lines of code but less than 0.2% code changes are required to achieve significant performance improvement.

Best Paper Finalist: no
Best Student Paper Finalist: no

Computing Planetary Interior Normal Modes with a Highly Parallel Polynomial Filtering Eigensolver
Jia Shi (Rice University), Ruipeng Li (Lawrence Livermore National Laboratory), Yuanzhe Xi (University of Minnesota), Yousef Saad (University of Minnesota), Maarten V. de Hoop (Rice University)

A highly parallel algorithm has been developed and exploited to compute the planetary normal modes of the elastic-gravitational system, which is approximated via the mixed finite element method on unstructured tetrahedral meshes. The eigenmodes of the relevant generalized eigenvalue problem were extracted by a Lanczos approach combined with polynomial filtering. In contrast with the standard shift-and-invert and the full-mode coupling algorithms, the polynomial filtering technique is ideally suited for solving large-scale 3-D interior eigenvalue problems since it significantly enhances the memory and computational efficiency without loss of accuracy. The parallel efficiency and scalability of this approach are demonstrated on Stampede2 at the Texas Advanced Computing Center. To our knowledge, this is the first time that the direct calculation of the normal modes of 3-D strongly heterogeneous planets, in particular, Earth and Mars, is made feasible via a combination of multiple matrix-free methods and a separation of the essential spectra.

Best Paper Finalist: no
Best Student Paper Finalist: no

Room: C141/143/149
3:30 pm - 5:00 pm

File Systems: Data Movement and Provenance

Dac-Man: Data Change Management for Scientific Datasets on HPC Systems

Devarshi Ghoshal (Lawrence Berkeley National Laboratory), Lavanya Ramakrishnan (Lawrence Berkeley National Laboratory), Deborah Agarwal (Lawrence Berkeley National Laboratory)

Scientific data is growing rapidly and often changes due to instrument configurations, software updates, or quality assessments. These changes in datasets can result in significant waste of compute and storage resources on HPC systems as downstream pipelines are reprocessed. Data changes need to be detected, tracked, and analyzed for understanding the impact of data change, managing data provenance, and making efficient and effective decisions about reprocessing and use of HPC resources. Existing methods for identifying and capturing change are often manual, domain-specific, and error-prone and do not scale to large scientific datasets. In this paper, we describe the design and implementation of Dac-Man framework, which identifies, captures, and manages change in large scientific datasets, and enables plug-in of domain-specific change analysis with minimal user effort. Our evaluations show that it can retrieve file changes from directories containing millions of files and terabytes of data in less than a minute.

Best Paper Finalist: no
Best Student Paper Finalist: no

Stacker: An Autonomic Data Movement Engine for Extreme-Scale Data Staging-Based In Situ Workflows

Pradeep Subedi (Rutgers University), Philip Davis (Rutgers University), Shaohua Duan (Rutgers University), Scott Klasky (Oak Ridge National Laboratory), Hemanth Kolla (Sandia National Laboratories), Manish Parashar (Rutgers University)

Data staging and in situ workflows are being explored extensively as an approach to address data-related costs at very large scales. However, the impact of emerging storage architectures (e.g., deep memory hierarchies and burst buffers) upon data staging solutions remains a challenge. In this paper, we investigate how burst buffers can be effectively used by data staging solutions, for example, as a persistence storage tier of the memory hierarchy. Furthermore, we use machine learning based prefetching techniques to move data between the storage levels in an autonomous manner. We also present Stacker, a prototype of the proposed solutions implemented within the Data\-Spaces data staging service, and experimentally evaluate its performance and scalability using the S3D combustion workflow on current leadership class platforms. Our experiments demonstrate that Stacker achieves low latency, high volume data-staging with low overhead as compared to in-memory staging services for production scientific workflows.

Best Paper Finalist: no
Best Student Paper Finalist: no

A Year in the Life of a Parallel File System

Glenn K. Lockwood (Lawrence Berkeley National Laboratory), Shane Snyder (Argonne National Laboratory), Teng Wang (Lawrence Berkeley National Laboratory), Suren Byna (Lawrence Berkeley National Laboratory), Philip Carns (Argonne National Laboratory), Nicholas J. Wright (Lawrence Berkeley National Laboratory)

I/O performance is a critical aspect of data-intensive scientific computing. We seek to advance the state of the practice in understanding and diagnosing I/O performance issues through investigation of a comprehensive I/O performance data set that captures a full year of production storage activity at two leadership-scale computing facilities. We demonstrate techniques to identify regions of interest, perform focused investigations of both long-term trends and transient anomalies, and uncover the contributing factors that lead to performance fluctuation.

We find that a year in the life of a parallel file system is comprised of distinct regions of long-term performance variation in addition to short-term performance transients. We demonstrate how systematic identification of these performance regions, combined with comprehensive analysis, allows us to isolate the factors contributing to different performance maladies at different time scales. From this, we present specific lessons learned and important considerations for HPC storage practitioners.

Best Paper Finalist: no
Best Student Paper Finalist: no

Poster

(back to top)

Tuesday, November 13th

Room: C2/3/4 Ballroom
8:30 am - 5:00 pm

Research Posters

Session Description: SC18 Research Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

SC18 Research Posters
SC18 Research Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Exploring Application Performance on Fat-Tree Networks in the Presence of Congestion
Philip A. Taffet (Rice University, Lawrence Livermore National Laboratory), Sanil Rao (University of Virginia, Lawrence Livermore National Laboratory), Ian Karlin (Lawrence Livermore National Laboratory)

Network congestion, which occurs when multiple applications simultaneously use shared links in cluster network, can cause poor communication performance, decreasing the performance and scalability of parallel applications. Many studies are performed while clusters also run other production workloads, which makes it harder for them to isolate causes and their effects. To look at congestion in a more controlled setting we used dedicated access time on an HPC cluster and measured the performance of three HPC applications with different communication patterns run with varying amounts and types of background traffic. This enables us to assess the relative sensitivity of the applications to congestion caused by different traffic patterns. Our tests show that the applications were not significantly impacted by even the most aggressive neighboring patterns, with all the performance degradation being 7% or less, pointing to the resiliency of the fat-tree topology.

Best Poster Finalist: no

GPU-Accelerated Interpolation for 3D Image Registration
Naveen Himthani (University of Texas, Institute for Computational Engineering and Sciences), Andreas Mang (University of Houston), Amir Gholami (University of California, Berkeley), George Biros (University of Texas, Institute for Computational Engineering and Sciences)

Image registration is a key technology in image computing with numerous applications in medical imaging. Our overarching goal is the design of a consistent and unbiased computational framework for the integration of medical imaging data with simulation and optimization to support clinical decision making for glioma brain tumor patients. A major issue in 3D image registration is the time to solution, which poses the demand for effective numerical schemes and the utilization of high performance

computing resources.

In this poster, we extend present a GPU-accelerated implementation of the Lagrange interpolation kernel using hardware texture filtering feature of modern hardware. Ongoing work involves implementing a unified single-GPU code for 3D image registration along with other computational kernels such as FFT. I will present my work by briefly explaining image registration followed by the explanation of the interpolation algorithm and its features and then demonstrate the results obtained.

Best Poster Finalist: no

Energy Efficiency of Reconfigurable Caches on FPGAs

Tianqi Wang (Boston University), Ang Li (Pacific Northwest National Laboratory), Tong Geng (Boston University), Martin Herbordt (Boston University)

The performance of a given cache architecture depends largely on the applications that run on it. Even though each application has its best-suited cache configuration, vendors of fixed HPC systems must provide compromise designs. Reconfigurable caches can adjust cache configuration dynamically to get best-suited cache parameters in runtime and notably reduce energy consumption. For example, when it is possible to deploy a low capacity low associativity design without increasing the miss rate substantially. For modern multi-core processors, each core's memory access behavior can be influenced by other cores. So it is more complicated to design reconfigurable caches for them.

In this paper, a design for a reconfigurable cache on FPGAs is presented that can run in modes with different capacities, associativity. We demonstrate that better performance and energy efficiency can be achieved by tuning these cache parameters at runtime.

Best Poster Finalist: no

RGB (Redfish Green500 Benchmarker): A Green500 Benchmarking Tool Using Redfish

Elham Hojati (Texas Tech University), Yong Chen (Texas Tech University), Alan Sill (Texas Tech University), Jon Hass (Dell Inc)

Performance and energy are important factors for supercomputers and data-centers with a trade-off between them. Energy efficiency metric considers both of these properties. The Green500 is a branch of Top500 project which provides a list of supercomputers based on energy efficiency. It has a manual methodology for this calculation.

Redfish is a new generation of management technologies for the hardware layer of data-centers. Our project focuses on designing and developing an automated Green500 benchmark tool using Redfish, called Redfish Green500 Benchmarker, or RGB in short. It offers the integration of Redfish and Green500, and automates Green500 benchmarking process with leveraging the internal capability of Redfish enabled equipment. It also enhances the Redfish standard to make sure it addresses the requirements of Green500 calculations. This research will also conduct validation and evaluation of RGB on real-world clusters for small-scale to medium-scale tests, and on the data-center simulator we have developed.

Best Poster Finalist: no

Optimization of Ultrasound Simulations on Multi-GPU Servers

Filip Vaverka (Brno University of Technology, Faculty of Information Technology), Matej Spetko (Brno University of Technology, Faculty of Information Technology), Bradley E. Treeby (University College London, Biomedical Ultrasound Group), Jiri Jaros (Brno University of Technology, Faculty of Information Technology)

Realistic ultrasound simulations have found a broad area of applications in preoperative photoacoustic screening and non-invasive ultrasound treatment planing. However, the domains are typically thousands of wavelengths in size, leading to large-scale numerical models with billions of unknowns. The current trend in accelerated computing is towards the use of fat nodes with multiple GPUs per node. The multi-GPU version of our k-Wave acoustic toolbox is based on the local Fourier basis domain decomposition where 3D simulation domain is partitioned into rectangular cuboid blocks assigned to particular GPUs. This paper investigates the benefits of using the CUDA-Aware MPI and CUDA peer-to-peer transfers on an 8-GPU server equipped with Nvidia P40 GPUs. The server has a total GPU memory of 192 GB and a single-precision performance of 96 Tflops. These techniques reduces the overall simulation time a factor of 2-3.6.

Best Poster Finalist: no

GPGPU Performance Estimation with Core and Memory Frequency Scaling

Qiang Wang (Hong Kong Baptist University), Xiaowen Chu (Hong Kong Baptist University)

Graphics processing units (GPUs) support dynamic voltage and frequency scaling to balance computational performance and energy consumption. However, simple and accurate performance estimation for a given GPU kernel under different frequency settings is still lacking for real hardware, which is important to decide the best frequency configuration for energy saving. We reveal a fine-grained analytical model to estimate the execution time of GPU kernels with both core and memory frequency scaling. Over a wide scaling range of both core and memory frequencies among 20 GPU kernels, our model achieves accurate results (4.83% error on average) with real hardware. Compared to the cycle-level simulators, our model only needs simple micro-benchmarks to extract a set of hardware parameters and kernel performance counters to produce such high accuracy.

Best Poster Finalist: no

Making Sense of Scientific Simulation Ensembles

Mai Dahshan (Virginia Tech), Nicholas Polys (Virginia Tech)

Scientists run many simulations with varying initial conditions, known as "ensembles", to understand the influence and relationships among multiple parameters or ensemble members. Most of the ensemble visualization and analysis approaches and techniques focus on analyzing the relationships between either the ensemble members or output parameter space while neglecting the effect of input parameters and humans in the analysis loop. Therefore, we developed an approach to the visual analysis of scientific data that merges human expertise and intuition with machine learning and statistics allowing scientists to explore, search, filter, and make sense of their high dimensional ensemble. Our tool, "GLEE" (Graphically-Linked Ensemble Explorer), is an interactive visualization tool that consists of three visual views: Ensemble View, Parameter View, and Statistical View. Each view offers different functionality for exploration and interoperation of the relations and correlations between different runs, a subset of runs, and input and output parameters.

Best Poster Finalist: no

Which Architecture Is Better Suited for Matrix-Free Finite-Element Algorithms: Intel Skylake or Nvidia Volta?

Martin Kronbichler (Technical University Munich), Momme Allalen (Leibniz Supercomputing Centre), Martin Ohlerich (Leibniz Supercomputing Centre), Wolfgang A. Wall (Technical University Munich)

This work presents a performance comparison of highly tuned matrix-free finite element kernels from the finite element library on different contemporary computer architectures, NVIDIA V100 and P100 GPUs, an Intel Knights Landing Xeon Phi, and two multi-core Intel CPUs (Broadwell and Skylake). The algorithms are based on fast integration on hexahedra using sum factorization techniques. For small problem sizes, when all data fits into CPU caches, Skylake is very competitive with Volta. For larger sizes, however, the GPU holds an advantage of approximately a factor of three over Skylake, because all architectures operate in the memory-bandwidth limited regime. A detailed performance analysis contrasts the throughput-oriented character of GPUs versus the more latency-optimized CPUs for the scenario of high-order finite element computations.

Best Poster Finalist: no

SpotSDC: an Information Visualization System to Analyze Silent Data Corruption

Zhimin Li (University of Utah), Harshitha Menon (Lawrence Livermore National Laboratory), Yarden Livnat (University of Utah), Kathryn Mohror (Lawrence Livermore National Laboratory), Valerio Pascucci (University of Utah)

Aggressive technology scaling trends are expected to make the hardware of HPC systems more susceptible to transient faults. Transient faults in hardware may be masked without affecting the program output, cause a program to crash, or lead to silent data corruptions (SDC). While fault injection studies can give an overall resiliency profile for an application, without a good visualization tool it is difficult to summarize and highlight critical information obtained. In this work, we design SpotSDC, a visualization system to analyze a program's resilience characteristics to SDC. SpotSDC provides an overview of the SDC impact on an application by highlighting regions of code that are most susceptible to SDC and will have a high impact on the program's output. SpotSDC also enables users to visualize the propagation of error through an application execution.

Best Poster Finalist: no

High-Accuracy Scalable Solutions to the Dynamic Facility Layout Problem

Apan Qasem (Texas State University), Clara Novoa (Texas State University), Chandra Kolla (Texas State University), Samantha Coyle (Texas State University)

The dynamic facility layout problem (DFLP) is concerned with finding arrangements of facilities within plant locations that minimize the sum of material handling and relocation costs over a planning horizon. DFLP is relevant in manufacturing engineering; accurate solutions can reduce operational costs by as much as 30%. We present a new scalable solution that formulates the task of finding the optimal arrangement as a shortest-path (SP) problem. The new parallel algorithm to find the SP employs a problem-specific heuristic to substantially cut down the search space. Compiler-level optimizations improve the performance across different execution platforms, including an auto-tuning strategy to derive the optimal SMT configuration on a POWER8 system. Results show a factor of 13

speedup over existing methods. For the six-facilities problems the best known solution is reached and for sets with 15 and 30 facilities the solution is within 2.83% and 5.66% of the best solution, respectively.

Best Poster Finalist: no

Large Scale MPI-Parallelization of LBM and DEM Systems: Accelerating Research by Using HPC

Bohumir Jelinek (Mississippi State University), George Mason (Mississippi State University), John Peters (Mississippi State University), Daniel Johnson (Mississippi State University), Marcus Brumfield (Mississippi State University), Alex Carrillo (US Army Engineer Research and Development Center), Clay Goodman (Mississippi State University), Farshid Vahedifard (Mississippi State University)

Casting, solidification, and the behavior of dry, saturated, and partially saturated granular media are examples of interesting and important problems in multiple areas of civil, mechanical, and chemical engineering. For interacting particle-fluid systems, the Discrete Element Method (DEM) and Lattice-Boltzmann Method (LBM) provide valuable high-resolution numerical models. Their main issue is high computational demand, which can be addressed by use of HPC resources. This work demonstrates the use of MPI-parallelized LBM and DEM models to accelerate research in solidification and macroscopic behavior of dry and saturated granular media. Large scale parallel simulations of dendritic growth, the calibration-chamber cone penetration test, and a parametric study of shear thickening in granular suspension were performed. Use of HPC dramatically reduced the computational time for these studies and provided high-resolution representation of physical experiments.

Best Poster Finalist: no

SciGaP: Apache Airavata Hosted Science Gateways

Marlon Pierce (Indiana University), Suresh Marru (Indiana University), Eroma Abeysinghe (Indiana University), Sudhakar Pamidighantam (Indiana University), Marcus Christie (Indiana University), Dimuthu Upeksha (Indiana University)

The goal of the Science Gateways Platform as a service (SciGaP.org) project is to provide core services for building and hosting science gateways. Over the last two years, SciGaP services have been used to build and host over twenty-five science gateways. SciGaP services support these gateways through a single hosted version of the Apache Airavata software system that supports multiple tenants. Apache Airavata services include scientific application execution management on HPC and cloud environments, input and output data staging, and provenance tracking for user-created computational experiments.

The poster presents highlights of some selected SciGaP-hosted gateways. Clients interested in SciGaP services can request a tenant through https://scigap.org/. Clients who need extensive support for building user interfaces and integrating unconventional resources can request support through the Science Gateways Community Institute's Extended Developer Support program. To integrate with XSEDE resources, clients can request support through XSEDE's Extended Collaborative Support Services.

Best Poster Finalist: no

Reproducibility as Side Effect

Shu Wang (University of Chicago), Zhuo Zhen (University of Chicago), Jason Anderson (University of Chicago), Kate Keahey (Argonne National Laboratory, University of Chicago)

The ability to keep records and reproduce experiments is a critical element of the scientific method for any discipline. However, the recording and publishing of research artifacts that allow to reproduce and directly compare against existing research continue to be a challenge. In this paper, we propose an experiment précis framework that helps the experiment repeatability. Guided by the framework, we implement a prototype tool called ReGen which generates repeatable experiment scripts that can be used or shared along with a detailed experiment description automatically. The evaluation shows that ReGen is effective in reducing the researcher's efforts of creating a repeatable experiment in a real setting.

Best Poster Finalist: no

Using Darshan and CODES to Evaluate Application I/O Performance

Harsh Khetawat (North Carolina State University), Christopher Zimmer (Oak Ridge National Laboratory), Frank Mueller (North Carolina State University), Sudharshan Vazhkudai (Oak Ridge National Laboratory), Scott Atchley (Oak Ridge National Laboratory)

Burst buffers have become increasingly popular in HPC systems, allowing bursty I/O traffic to be serviced faster without slowing down application execution. The ubiquity of burst buffers creates opportunities for studying their ideal placement in the HPC topology. Furthermore, the topology of the network interconnect can also affect the performance of the storage hierarchy for different burst buffer placement schemes. To that end, we create a reproducible framework that allows individual centers to develop their own models and evaluate performance based on their workload characteristics. We use CODES to create models that simulate the network and storage layers of an HPC system and Darshan traces for I/O replay. We augment the Darshan traces with synchronization primitives, and allow multi-dimensional scaling of traces to represent future workload characteristics. Finally, we evaluate the effect of network topology, storage architecture, and application I/O patterns on overall I/O performance.

Best Poster Finalist: no

Multi-Client DeepIO for Large-Scale Deep Learning on HPC Systems

Yue Zhu (Florida State University), Fahim Chowdhury (Florida State University), Huansong Fu (Florida State University), Adam Moody (Lawrence Livermore National Laboratory), Kathryn Mohror (Lawrence Livermore National Laboratory), Kento Sato (Lawrence Livermore National Laboratory), Weikuan Yu (Florida State University)

With the growth of computation power, leadership High-Performance Computing (HPC) systems can train larger datasets for Deep neural networks (DNNs) more efficiently. On HPC systems, a training dataset is on a parallel file system or node-local storage devices. However, not all HPC clusters have node-local storage, and large mini-batch sizes stress the read performance of parallel systems since the large datasets cannot fit in file system caches. Thus, it is a challenge for training DNNs with large datasets on HPC systems.

In prior work, we proposed DeepIO to mitigate the I/O pressure. DeepIO is designed to assist the mini-batch generation of TensorFlow. However, DeepIO does not support multiple training workers on a

single compute node. We address this gap with modification on DeepIO framework, and evaluate multi-client DeepIO performance against state-of-the-art in-memory file systems, compare DeepIO and TensorFlow data loading API, and explore the potential of DeepIO in DNN training.

Best Poster Finalist: no

HIVE: A Cross-Platform, Modular Visualization Ecosystem for Heterogeneous Computational Environments

Jorji Nonaka (Riken Center for Computational Science), Kenji Ono (Kyushu University, RIKEN), Naohisa Sakamoto (Kobe University, RIKEN), Kengo Hayashi (Kobe University, RIKEN), Tomohiro Kawanabe (Riken Center for Computational Science), Fumiyoshi Shoji (Riken Center for Computational Science), Masahiro Fujita (LTE Inc), Kentaro Oku (Kashika Inc), Kazuma Hatta (Imagica Digitalscape)

HPC operational environments usually have supporting computational systems for assisting pre- and post-processing activities such as the visualization and analysis of simulation results. A wide variety of hardware systems can be found at different HPC sites, and in our case, we have a CPU-only (x86) large memory server, a planned OpenStack-based CPU/GPU Cluster, SPARC64 fx CPU based HPC system (K computer), and an ARM based HPC system in the future. Therefore, heterogeneity and scalability are needed to be tackled to efficiently use these heterogeneous computational resources for large-scale data visualization on both post-hoc and in-situ contexts. In this poster we present HIVE (Heterogeneously Integrated Visual-analytics Environment), a cross-platform and modular ecosystem for providing visualization service building blocks in such heterogeneous computational environments. Lightweight Lua scripting language is used to glue necessary visualization pipeline related modules, and this loosely coupled modular approach facilitates long-term development and maintenance.

Best Poster Finalist: no

Improving the I/O Performance and Memory Usage of the Xolotl Cluster Dynamics Simulator

Philip C. Roth (Oak Ridge National Laboratory), Sophie Blondel (University of Tennessee), David E. Bernholdt (Oak Ridge National Laboratory), Brian D. Wirth (University of Tennessee)

Xolotl is a cluster dynamics simulator used to predict gas bubble evolution in solids. It is currently being used to simulate bubble formation in the plasma-facing surface within fusion reactors and the nuclear fuel used in fission reactors. After observing performance problems in coupled-code simulations of fusion reactors, we used Xolotl's built-in performance data collection infrastructure and an external profiling tool to identify inefficiencies when writing Xolotl's two types of checkpoint files. We changed the code to use true parallel writes via the HDF5 data management library, resulting in a code that is approximately 57x faster when writing the program's main checkpoint file at the scale used in the coupled-code simulations, and that exhibits less performance variability due to external system activity. We also identified and addressed a memory usage problem that reduced Xolotl peak memory usage by approximately 88% per compute node.

Best Poster Finalist: no

Performance Evaluation of the Shifted Cholesky QR Algorithm for Ill-Conditioned Matrices

Takeshi Fukaya (Hokkaido University), Ramaseshan Kannan (Arup UK), Yuji Nakatsukasa (National Institute of Informatics, Japan), Yusaku Yamamoto (University of Electro-Communications, Japan), Yuka Yanagisawa (Waseda University)

The Cholesky QR algorithm, which computes the QR factorization of a matrix, is a simple yet efficient algorithm for high-performance computing. However it suffers from numerical instability. In a recent work, this instability has been remedied by repeating Cholesky QR twice (CholeskyQR2). ChokeskyQR2, however, is still prone to numerical breakdown when applied to ill-conditioned matrices. To overcome this limitation, we introduce a shifting technique to Cholesky QR and use it as a preconditioning step before CholeskyQR2. The key idea is that Cholesky QR with shift reduces the condition number of the input matrix. We call the resulting algorithm shifted CholeskyQR3, which is still simple and only requires double precision arithmetic. In this poster, we present the results of our performance evaluation of shifted CholeskyQR3. We demonstrate that shifted CholeskyQR3 accurately computes the QR factorization of ill-conditioned matrices and that it outperforms other conventional algorithms in execution time.

Best Poster Finalist: no

HPC-as-a-Service for Life Sciences

Vaclav Svaton (Technical University of Ostrava, Czech Republic), Jan Martinovic (Technical University of Ostrava, Czech Republic), Nina Jeliazkova (IDEAconsult Ltd, Bulgaria), Vladimir Chupakhin (Janssen Pharmaceutika NV), Pavel Tomancak (Max Planck Institute of Molecular Cell Biology and Genetics), Petr Vojta (Palacký University Olomouc, Czech Republic)

HPC-as-a-Service is a well-known term in the area of high performance computing. It enables users to access an HPC infrastructure without a need to buy and manage their own infrastructure. Through this service, academia and industry can take advantage of the technology without an upfront investment in the hardware. This approach further lowers the entry barrier for users who are interested in utilizing massive parallel computers but often do not have the necessary level of expertise in the area of parallel computing. To provide this simple and intuitive access to the supercomputing infrastructure, an in-house application framework called HEAppE has been developed. HEAppE's universally designed software architecture enables unified access to different HPC systems through a simple object-oriented API. Thus providing HPC capabilities to the users but without the necessity to manage the running jobs from the command-line interface of the HPC scheduler directly on the cluster.

Best Poster Finalist: no

Hermes: a Multi-Tiered Distributed I/O Buffering System for HDF5

Hariharan Devarajan (Illinois Institute of Technology, HDF Group)

High-Performance Computing (HPC) systems' increasing ability to run data-intensive problems at larger scale and resolution has driven the evolution of modern storage technologies. In addition, extreme amounts of data are collected by large scientific instruments and sensor network is resulting in a push for more capable storage systems. Hierarchical Data Format (HDF) technologies address the problems of how to organize, access, analyze, and preserve data in the face of enormous growth in size and complexity. To mitigate this I/O performance bottleneck, modern storage subsystems are going through extensive changes, by adding additional levels of memory and storage in a hierarchy. In this work, we present Hermes: a new, heterogeneous-aware, multi-tiered, dynamic, and distributed I/O buffering system for HDF5. Hermes enables, manages, and supervises I/O buffering in DMSH. We tested our solution on Cori and show Hermes can accelerate applications by 62x than the state of the buffering platform.

Best Poster Finalist: no

Workflow for Parallel Processing of Sequential Mesh Databases

Ondřej Meca (Technical University of Ostrava, Czech Republic), Lubomír Říha (Technical University of Ostrava, Czech Republic), Tomáš Brzobohatý (Technical University of Ostrava, Czech Republic)

This poster presents a workflow for parallel loading of sequentially stored mesh databases. It can be used as a connection between tools for the creation of complex engineering models along with parallel solvers to allow broader usage of HPC by the engineering community. Scalability tests show that the proposed algorithm is able to prepare a mesh with hundreds of millions of nodes/elements in several seconds.

Best Poster Finalist: yes

The NAStJA Framework: Non-Collective Scalable Global Communications

Marco Berghoff (Karlsruhe Institute of Technology), Ivan Kondov (Karlsruhe Institute of Technology)

In recent years, simulations in various areas of science and engineering have proven to be very useful. To efficiently deploy simulation codes on current and future high-performance computer systems, high node level performance, scalable communication and the exclusion of unnecessary calculations are an absolute must when developing new solvers.

We have introduced the NAStJA framework, a block-based MPI parallel solver for algorithms, based on regular grid methods, i.e., stencil codes. NAStJA has a dynamic block adaptation, which modifies the calculation domain around the region in which the calculation is currently taking place. The creation and deletion of blocks are autonomously managed within local neighborhoods. Collective all-gather communication is avoided by using a multi-hop network to distribute information across the entire domain that greatly improves the application scaling. In this contribution, we present applications that can benefit from this adaptive method and scaling tests demonstrating the excellent scalability.

Best Poster Finalist: no

Hardware Acceleration of CNNs with Coherent FPGAs

Md Syadus Sefat (Texas State University), Semih Aslan (Texas State University), Apan Qasem (Texas State University)

This paper describes a new flexible approach to implementing energy-efficient CNNs on FPGAs. Our design leverages the Coherent Accelerator Processor Interface (CAPI) which provides a cache-coherent view of system memory to attached accelerators. Convolution layers are formulated as matrix multiplication kernels and then accelerated on CAPI-supported Kintex FPGA board. Our implementation bypasses the need for device driver code and significantly reduces the communication and I/O transfer overhead. To improve the performance of the entire application, not just the convolution layers, we propose a collaborative model of execution in which the control of the data flow within the accelerator is kept independent, freeing-up CPU cores to work on other parts of the application. For further performance enhancements, we propose a technique to exploit data locality in the cache, situated in the CAPI Power Service Layer (PSL). Finally, we develop a resource-conscious implementation for more efficient utilization of resources and improved scalability.

Best Poster Finalist: no

Distributed Fast Boundary Element Methods

Michal Merta (Technical University of Ostrava, Czech Republic), Jan Zapletal (Technical University of Ostrava, Czech Republic), Michal Kravcenko (Technical University of Ostrava, Czech Republic)

We present a parallel implementation of the fast boundary element method (BEM) for the Helmholtz equation. After a brief description of BEM, vectorization of the computationally most demanding kernels, and shared memory parallelization, we focus on the distributed memory parallelization using a new approach for distribution of the fast BEM system matrices among computational nodes. Moreover, we present a modification of the adaptive cross approximation (ACA) method capable of dealing with BEM matrices containing zero blocks, which usually lead to problems for the original ACA algorithm. Numerical experiments demonstrating the performance of the implementation were carried out using several Intel architectures.

Best Poster Finalist: no

Development of Numerical Coupled Analysis Method by Air Flow Analysis and Snow Accretion Analysis

Kohei Murotani (Railway Technical Research Institute, Japan), Koji Nakade (Railway Technical Research Institute, Japan), Yasushi Kamata (Railway Technical Research Institute, Japan), Daisuke Takahashi (Railway Technical Research Institute, Japan)

In this research, to take countermeasures for the snow accretion damage, we developed a simulator of realizing the snow accretion process in the following steps. Firstly, air flow analysis is performed by “Airflow simulator” developed by RTRI (Railway Technical Research Institute). Secondly, trajectory of flying snow is calculated by the equation of motion for gravity and drag using distribution of velocity of air flow. Thirdly, snow accretion analysis is performed by “Particle simulator” developed by RTRI. The shape modified by snow accretion is reflected onto the boundary conditions of the air flow analysis. In this year, we show the results of the snow accretion analysis for simple cubic shapes in order to aim at system development and validation and discuss the result of the snow accretion analysis for a train bogie model.

Best Poster Finalist: no

Portable Parallel Performance via Multi-Dimensional Homomorphisms

Ari Rasch (University of Münster), Richard Schulze (University of Münster), Sergei Gorlatch (University of Münster)

Achieving portable performance over different parallel architectures and varying problem sizes is hard: e.g., a program optimized for multi-core CPUs on large input sizes can significantly differ from the same program optimized for Graphics Processing Units (GPUs) on small sizes.

We propose an approach to ensuring portability of performance by relying on multi-dimensional homomorphisms (MDHs) -- a class of parallelizable functions that cover important application areas including linear algebra routines (BLAS) and stencil computations. We develop an extended OpenCL implementation schema for MDHs that is generic in the performance-critical parameters of the OpenCL model, and we enable portability of performance by being automatically optimized for different target architectures and input sizes using the auto-tuning approach.

Our results demonstrate competitive and often even significantly better performance than state-of-the-art

approaches for BLAS and Stencil as used in the important application area of deep learning.

Best Poster Finalist: no

Performance Evaluation of the NVIDIA Tesla V100: Block Level Pipelining vs. Kernel Level Pipelining
Xuewen Cui (Virginia Tech), Thomas R. W. Scogland (Lawrence Livermore National Laboratory), Bronis R. de Supinski (Lawrence Livermore National Laboratory), Wu Feng (Virginia Tech)

As accelerators become more common, expressive and performant, interfaces for them become ever more important. Programming models like OpenMP offer simple-to-use but powerful directive-based offload mechanisms. By default, these models naively copy data to or from the device without overlapping computation. Achieving performance can require extensive hand-tuning to apply optimizations such as pipelining. To pipeline a task, users must manually partition the task into multiple chunks then launch multiple sub-kernels. This approach can suffer from high kernel launch overhead. Also, the hyper parameters must be carefully tuned to achieve optimal performance. To ameliorate this issue, we propose a block-level pipeline approach that overlaps data transfers and computation in one kernel handled by different streaming multiprocessors on GPUs. Our results show that, without exhaustive tuning, our approach can provide 95% to 108% stable performance compared to the best tuned results with traditional kernel-level pipelining on NVIDIA V100 GPUs.

Best Poster Finalist: no

Enabling Data Analytics Workflows Using Node-Local Storage
Tu Mai Anh Do (University of Southern California, Information Sciences Institute), Ming Jiang (Lawrence Livermore National Laboratory), Brian Gallagher (Lawrence Livermore National Laboratory), Albert Chu (Lawrence Livermore National Laboratory), Cyrus Harrison (Lawrence Livermore National Laboratory), Karan Vahi (University of Southern California, Information Sciences Institute), Ewa Deelman (University of Southern California, Information Sciences Institute)

The convergence of high-performance computing (HPC) and Big Data is a necessity with the push toward extreme-scale computing. As HPC simulations become more complex, the analytics need to process larger amounts of data, which poses significant challenges for coupling HPC simulations with Big Data analytics. This poster presents a novel node-local approach that uses a workflow management system (WMS) to enable the coupling between the simulations and the analytics in scientific workflows by leveraging node-local non-volatile random-access memory (NVRAM).

Best Poster Finalist: no

OpeNNdd: Open Neural Networks for Drug Discovery: Creating Free and Easy Methods for Designing Medicine
Bryce Kroencke (American River College), Shawn Shacterman (University of California, Berkeley), Nicholas Pavini (American River College), Benjamin Samudio (American River College, Sierra College), Silvia Crivelli (Lawrence Berkeley National Laboratory)

Bringing new medicines to patients can be prohibitively expensive in terms of time, cost, and resources. This leaves many diseases without therapeutic interventions. In addition, new and reemerging diseases are increasing in prevalence across the globe at an alarming rate. The speed and scale of medicine discovery must be increased to effectively meet this challenge. OpeNNdd is a neural network platform bringing

together people, machine learning, and supercomputing to solve the challenge of creating medicines. We have developed a novel neural network which quickly and accurately models candidate medicines interacting with a disease target, a metric to delineate its domain of applicability, and a process that communicates neural network results to participants in a readily interpretable way. OpeNNdd leverages the scale of supercomputing, the power and speed of neural networks, and the creativity of people across the globe in an open and collaborative way to protect and improve global health.

Best Poster Finalist: no

FeatherCNN: Fast Inference Computation with TensorGEMM on ARM Architectures
Haidong Lan (Shandong University), Jintao Meng (Tencent Holdings Ltd), Christian Hundt (Johannes Gutenberg University Mainz), Bertil Schmidt (Johannes Gutenberg University Mainz), Minwen Deng (Tencent Holdings Ltd), Weiguo Liu (Shandong University), Yanjie Wei (Shenzhen Institutes of Advanced Technology), Shengzhong Feng (Shenzhen Institutes of Advanced Technology)

This poster presents a fast inference computation library for ARM architecture named as CNNForward. CNNForward is trying to improve the efficiency of inference computation for convolutional neural networks on ARM-based multi-core and many-core architectures using both mathematical formula reconstruction/simplification and in-depth NEON instruction optimization. Experimental results reveal that, forward computation for VGG-16 on a server with 64 ARM A72 cores, CNNForward can scale up to 32 cores with an parallel efficiency of 33%, and achieve 35.4x, 8.7x and 10.6x speedup over Caffe+OpenBlas, Caffe2+Eigen and Caffe2+NNPACK, respectively.

Best Poster Finalist: no

Boosting the Scalability of Car-Parrinello Molecular Dynamics Simulations for Multi- and Manycore Architectures
Tobias Klöffel (University of Erlangen-Nuremberg), Bernd Meyer (University of Erlangen-Nuremberg), Gerald Mathias (Leibniz Supercomputing Centre)

We present our recent optimizations of the ultra-soft pseudo-potential (USPP) code path of the ab inito molecular dynamics program CPMD (www.cpmd.org). Following the internal instrumentation of CPMD, all relevant USPP routines have been revised to fully support hybrid MPI+OpenMP parallelization. For two time-critical routines, namely the multiple distributed 3D FFTs of the electronic states and a key distributed matrix-matrix multiplication, we have implemented hybrid parallel algorithms with overlapping computation and communication. The achievements in performance and scalability are demonstrated on a small reference system of 128 water molecules and further systems of increasing size. Performance evaluation shows gains of up to one order of magnitude and around 50% peak performance for simulation systems readily used in production.

Best Poster Finalist: no

Characterizing Declustered Software RAID for Enhancing Storage Reliability and Performance
Zhi Qiao (University of North Texas), Song Fu (University of North Texas), Hsing-bung Chen (Los Alamos National Laboratory), Bradley Settlemyer (Los Alamos National Laboratory)

Redundant array of independent disks (RAID) has been widely used to address the reliability issue in storage systems. As the scale of modern storage systems continues growing, disk failure becomes the

norm. With ever-increasing disk capacity, RAID recovery based on disk rebuild becomes more costly, which causes significant performance degradation and even unavailability of storage systems. Declustered parity and data placement in RAID aims to enhance the recovery performance by shuffling data among all disks in a RAID group. All disks in the RAID group participate in data reconstruction, which leads to reduction of the RAID rebuild time. In this work, we extensively evaluate declustered RAID in terms of the performance of application I/O and recovery time. Our experimental results in ZFS show that the speedup of declustered RAID over traditional RAID is sub-linear to the number of disks in the storage pool.

Best Poster Finalist: no

Parallel Implementation of Machine Learning-Based Many-Body Potentials on CPU and GPU
Yaoguang Zhai (University of California, San Diego), Nathaniel Danandeh (University of California, San Diego), Zhenye Tan (University of California, San Diego; Tongji University), Sicun Gao (University of California, San Diego), Francesco Paesani (University of California, San Diego), Andreas W. Goetz (San Diego Supercomputer Center)

Machine learning models can be used to develop highly accurate and efficient many-body potentials for molecular simulations based on the many-body expansion of the total energy. A prominent example is the MB-pol water model that employs permutationally invariant polynomials (PIPs) to represent the 2-body and 3-body short-range energy terms.

We have recently shown that the PIPs can be replaced by Behler-Parinello neural networks (BP-NN). We present OpenMP parallel implementations of both PIP and BP-NN models as well as a CUDA implementation of the BP-NN model for GPUs. The OpenMP implementations achieve linear speedup with respect to the optimized single threaded code. The BP-NN GPU implementation outperforms the CPU implementation by a factor of almost 8. This opens the door for routine molecular dynamics simulations with highly accurate many-body potentials on a diverse set of hardware.

Best Poster Finalist: no

Implementing Efficient Data Compression and Encryption in a Persistent Key-Value Store for HPC
Jungwon Kim (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory)

Recently, persistent data structures, like key-value stores (KVSs), which are stored in an HPC system's nonvolatile memory, provide an attractive solution for a number of emerging challenges like limited I/O performance. This paper investigates how to efficiently integrate data compression and encryption into persistent KVSs for HPC with the ultimate goal of hiding their costs and complexity in terms of performance and ease of use. We implement the proposed techniques on top of a distributed embedded KVS to evaluate the benefits and costs of incorporating these capabilities along different points in the dataflow path, illustrating differences in effective bandwidth, latency, and additional computational expense.

Best Poster Finalist: yes

A Parallel-Efficient GPU Package for Multiphase Flow in Realistic Nano-Pore Networks
Yidong Xia (Idaho National Laboratory), Ansel Blumers (Brown University, Idaho National Laboratory), Zhen Li (Brown University), Lixiang Luo (IBM), Jan Goral (University of Utah), Matthew Andrew (Carl Zeiss X-ray Microscopy Inc), Joshua Kane (Idaho National Laboratory), Yu-Hang Tang (Lawrence Berkeley National Laboratory)

Simulations of fluid flow in oil/gas shale rocks are challenging in part due to the heterogeneous pore sizes ranging from a few nanometers to a few micrometers. Additionally, the complex fluid-solid interaction occurring physically and chemically must be captured with high resolution. To address these challenges while minimizing computational cost, we present a GPU code that has implemented a many-body dissipative particle dynamics (mDPD) model for multiphase flow in shale. Realistic nano- to micro-pore channels in shale are constructed from 3D high-resolution stack images. In our benchmark tests, the code delivers nearly perfect weak and strong scalings on up to 512 K20X GPUs on Oak Ridge National Laboratory (ORNL) Titan supercomputer. Moreover, single-GPU bench-marks on the DGX-1 (V100/no NVLink), ORNL's SummitDev (P100/NVLink 1.0) and Summit (V100/NVLink 2.0) suggest that the latest Host-to-Device NVLink can significantly boost overall performance, in addition to the Device-to-Device NVLink.

Best Poster Finalist: no

Processing-in-Storage Architecture for Machine Learning and Bioinformatics

Roman Kaplan (Israel Institute of Technology), Leonid Yavits (Israel Institute of Technology), Ran Ginosar (Israel Institute of Technology)

User-generated and bioinformatics database volumes has been increasing exponentially for more than a decade. With the slowdown and approaching end of Moore's law, traditional technologies cannot satisfy the increasing demands for processing power. This work presents PRINS, a highly-parallel in-storage processing architecture. PRINS combines non-volatile memory with processing capabilities on every bitcell. An emerging technology, memristors, form the basis for the design.

Implementations of three data-intensive and massively parallel algorithms are demonstrated: (1) Smith-Waterman DNA local sequence alignment (bioinformatics), (3) K-means clustering (machine learning) and (3) data deduplication. Performance and energy efficiency of PRINS compared to other published solutions is presented for each algorithm. PRINS is shown to achieve orders-of-magnitude improvement in performance and power efficiency over existing solutions, from large-scale bioinformatics and machine-learning to single-GPU or FPGA implementations.

Best Poster Finalist: no

Kernel-Based and Total Performance Analysis of CGYRO on 4 Leadership Systems

Igor Sfiligoi (General Atomics), Jeff Candy (General Atomics), Emily Belli (General Atomics)

We present the results of an exhaustive performance analysis of the CGYRO code on 4 leadership systems spanning 5 different configurations (2 KNL-based, 1 Skylake-based, and 2 hybrid CPU-GPU architectures). CGYRO is an Eulerian gyrokinetic solver designed and optimized for collisional, electromagnetic, multiscale fusion plasma simulation. It is based on the well-known GYRO code, but redesigned from the ground up to operate efficiently on multicore and GPU-accelerated systems. The gyrokinetic equations specify a 5-dimensional distribution function for each species, with species coupled through both the Maxwell equations and collision operator. For the cross-machine performance analysis, we report and compare timings for 4 computational and 4 communication kernels. This kernel-based breakdown illustrates the strengths and weaknesses of the floating-point and communication architectures of the respective systems. An overview of the physical equations solved, the scalable numerical methods used, and data communication patterns required by each kernel are also given.

Best Poster Finalist: no

Redesigning The Absorbing Boundary Algorithm for Asynchronous High Performance Acoustic Wave Propagation

Rached Abdelkhalak (King Abdullah University of Science and Technology), Kadir Akbudak (King Abdullah University of Science and Technology), Vincent Etienne (Saudi Aramco), Thierry Tonellot (Saudi Aramco)

Exploiting high concurrency, relaxing the synchrony of existing algorithms, and increasing data reuse have immense effect in performance. We integrate the Multicore-optimized Wavefront Diamond (MWD) tiling approach by Malas et al. [SIAM SISC, 2015, ACM Trans. Parallel Comput. 2017], which takes into account the three aforementioned ingredients, into the industrial project codenamed ExaWave framework beside the traditional spatial blocking (SB) technique for stencil computations. However, the fine-grained asynchronous handling of the Convolution Perfectly Matched Layer (CPML) for absorbing boundary conditions turns out to be a challenging open research problem, due to severe inherent data dependencies constraints, which impedes MWD performance impact. We propose techniques of loop fusion to reduce memory traffic and sliding windows to cut down the engendered extra flops, in order to consolidate CPML integration with the overall asynchronous MWD technique. The experimental results on Intel's latest processors show the effectiveness of the proposed techniques.

Best Poster Finalist: no

Capsule Networks for Protein Structure Classification

Dan A. Rosa de Jesus (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Julian Cuevas Paniagua (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Wilson Rivera (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Silvia Crivelli (Lawrence Berkeley National Laboratory)

Capsule Networks have great potential to tackle problems in structural biology because of their attention to hierarchical relationships. This work describes the implementation and application of a capsule network architecture to the classification of RAS protein family structures on GPU-based computational resources. Our results show that the proposed capsule network trained on 2D and 3D structural encodings can successfully classify HRAS and KRAS structures. The capsule network can also classify a protein-based dataset derived from a PSI-BLAST search on sequences of KRAS and HRAS mutations. Experimental results show an accuracy improvement compared to traditional convolutional networks.

Best Poster Finalist: no

Cross-Layer Group Regularization for Deep Neural Network Pruning

Shuang Gao (Nvidia Corporation), Xin Liu (Nvidia Corporation)

Improving weights sparsity is a common strategy for deep neural network pruning. Most existing methods use regularizations that only consider structural sparsity within an individual layer. In this paper, we propose a cross-layer group regularization taking into account the statistics from multiple layers. For residual networks, we use this approach to align kernel sparsity across layers that are tied to each other through element-wise operations: the ith kernel of these layers are put into one regularization group, they either stay or be removed simultaneously during pruning. In this way, the computational and parameter storage cost could be significantly reduced. Experimental results show that this method does not only improve weights sparsity but also align kernel weights sparsity across related layers. Our method is able to prune ResNet up

to 90.4% of parameters and improve runtime by 1.5x speedup, without loss of accuracy.

Best Poster Finalist: yes

Machine Learning for Adaptive Discretization in Massive Multiscale Biomedical Modeling
Changnian Han (Stony Brook University), Prachi Gupta (Stony Brook University), Peng Zhang (Stony Brook University), Danny Bluestein (Stony Brook University), Yuefan Deng (Stony Brook University)

For multiscale problems, traditional time stepping algorithms use a single smallest time stepsize in order to capture the finest details; using this scale leads to a significant waste of computing resources for simulating coarse-grained portion of the problem. To improve computing efficiency for multiscale modeling, we propose a novel state-driven adaptive time stepping (ATS) algorithm to automatically adapt the time stepsizes to the underlying biophysical phenomena at multiple scales. In this, we use a machine-learning based solution framework to classify and label these states for regulating the time stepsizes. We demonstrate the values of our ATS algorithm by assessing the accuracy and efficiency of a multiscale two-platelet aggregation simulation. By comparing with traditional algorithm for this simulation, our ATS algorithm significantly improves the efficiency while maintaining accuracy. Our novel ATS algorithm presents a more efficient framework for solving massive multiscale biomedical problems.

Best Poster Finalist: no

Multi-GPU Accelerated Non-Hydrostatic Numerical Ocean Model with GPUDirect RDMA Transfers
Takateru Yamagishi (Research Organization for Information Science and Technology, Japan), Yoshimasa Matsumura (University of Tokyo), Hiroyasu Hasumi (University of Tokyo)

We have implemented our “kinaco” numerical ocean model on Tokyo University’s Reedbush supercomputer, which utilizes the latest Nvidia Pascal P100 GPUs with GPUDirect technology. We have also optimized the model’s Poisson/Helmholtz solver by adjusting the global memory alignment and thread block configuration, introducing shuffle functions to accelerate the creation of coarse grids and merging small kernels in the multigrid preconditioner. We also utilize GPUDirect RDMA transfers to improve MPI communication efficiency. By exploiting the GPUs’ capabilities, the GPU implementation is now twice as fast as the CPU version, and it shows good weak scalability to multiple GPUs. Most of the GPU kernels are accelerated, and the velocity diagnosis functions in particular are now approximately seven times faster. The performance of inter-node data transfers using a CUDA-aware MPI library with GPUDirect RDMA transfers is comparable to that on CPUs.

Best Poster Finalist: no

A Locality and Memory Congestion-Aware Thread Mapping Method for Modern NUMA Systems
Mulya Agung (Tohoku University), Muhammad Alfian Amrizal (Tohoku University), Ryusuke Egawa (Tohoku University), Hiroyuki Takizawa (Tohoku University)

On modern NUMA systems, the memory congestion problem could degrade performance more than the memory access locality problem because a large number of processor cores in the systems can cause heavy congestion on memory controllers. In this work, we propose a thread mapping method that considers the spatio-temporal communication behavior of multi-threaded applications to improve the locality and to reduce the memory congestion on modern NUMA systems. We evaluate the proposed method using NPB applications on a NUMA system. Experiments show that our proposed method can achieve up to 20%

performance improvement compared with locality-based and balance-based methods.

Best Poster Finalist: no

Tuning CFD Applications for Intel Xeon Phi with TAU Commander and ParaTools ThreadSpotter

Izaak B. Beekman (ParaTools Inc), Nicholas Chaimov (ParaTools Inc), Sameer Shende (ParaTools Inc, University of Oregon), Allen D. Malony (ParaTools Inc, University of Oregon), Nicholas Bisek (US Air Force Research Laboratory), Ryan Gosse (US Air Force Research Laboratory), Andrew Wissink (Create AV, US Army)

Tuning and understanding the performance characteristics of computational fluid dynamics (CFD) codes on many-core, NUMA architectures is challenging. One must determine how programming choices impact algorithm performance and how best to utilize the available memory caches, high-bandwidth memory, and inter---and intra---node communication. Once collected, performance data must be translated into actionable code improvements. In addition, performance engineering experiments must be organized and tracked to quantify the benefit of any attempted tuning.

In the poster we present, examine and tune two CFD applications running on the Intel® Xeon Phi™ partition of a Cray® XC 40/50 using TAU Commander and ParaTools ThreadSpotter. TAU Commander implements a streamlined, managed performance engineering workflow and highlights source regions limiting scalability through profiling and aggregate summary statistics. ParaTools ThreadSpotter analyzes an application as it is running and ranks individual performance problems. It provides a report annotating source code lines with actionable recommendations and quantifying performance metrics.

Best Poster Finalist: no

Massively Parallel Stress Chain Characterization for Billion Particle DEM Simulation of Accretionary Prism Formation

Mikito Furuichi (Japan Agency for Marine-Earth Science and Technology), Daisuke Nishiura (Japan Agency for Marine-Earth Science and Technology), Takane Hori (Japan Agency for Marine-Earth Science and Technology)

Herein, a novel algorithm for characterizing stress chains using a large parallel computer system is presented. Stress chains are important for analyzing the results of large-scale discrete element method (DEM) simulations. However, the general algorithm is difficult to parallelize especially when selecting networks longer than several particles. Therefore, we propose a new parallel algorithm to count the number of particles that are tightly connected, based on iterative operations with nearest-neighbor computations and communications. The new algorithm is examined via a real-scale numerical sandbox experiment using 2.4 billion particles. We successfully compute the stress chains with a reasonable computational cost comparable to the single-step DEM computation time. The visualization of the stress chains from the large-scale DEM simulation result reveals the existence of arcuate stress structures that may control accretionary prism formation, which is an important scientific discovery.

Best Poster Finalist: no

Toward Smoothing Data Movement Between RAM and Storage

Tariq Alturkestani (King Abdullah University of Science and Technology), Thierry Tonellot (Saudi Aramco), Vincent Etienne (Saudi Aramco), Hatem Ltaief (King Abdullah University of Science and Technology)

We propose to design and implement a software framework, which provides a Multilayer Buffer System (MBS) to cache in/out datasets into CPU main memory from/to slower storage media, such as parallel file systems (e.g., Lustre), solid-state drive (e.g., Burst Buffer) or non-volatile RAM. Although MBS scope may be broad in terms of scientific applications, we focus on the RTM application as a proxy for I/O intensive workloads, since reading and writing are ubiquitous operations during the dumping phase of the source wavefield (forward propagation) as well as its retrieval phase (backward propagation) for the image condition calculations.

Best Poster Finalist: no

Interactive HPC Deep Learning with Jupyter Notebooks

Wahid Bhimji (Lawrence Berkeley National Laboratory), Steven Farrell (Lawrence Berkeley National Laboratory), Oliver Evans (Lawrence Berkeley National Laboratory), Matthew Henderson (Lawrence Berkeley National Laboratory), Shreyas Cholia (Lawrence Berkeley National Laboratory), Aaron Vose (Cray Inc), Mr Prabhat (Lawrence Berkeley National Laboratory), Rollin Thomas (Lawrence Berkeley National Laboratory), Richard Shane Canon (Lawrence Berkeley National Laboratory)

Deep learning researchers are increasingly using Jupyter notebooks to implement interactive, reproducible workflows. Such solutions are typically deployed on small-scale (e.g. single server) computing systems. However, as the sizes and complexities of datasets and associated neural network models increase, distributed systems become important for training and evaluating models in a feasible amount of time. In this poster, we describe our work on Jupyter notebook solutions for distributed training and hyper-parameter optimization of deep neural networks on high-performance computing systems.

Best Poster Finalist: no

Fast and Accurate Training of an AI Radiologist

Lucas A. Wilson (Dell EMC), Vineet Gundecha (Dell EMC), Srinivas Varadharajan (Dell EMC), Alex Filby (Dell EMC), Pei Yang (Dell EMC), Quy Ta (Dell EMC), Valeriu Codreanu (SURFsara), Damian Podareanu (SURFsara), Vikram Saletore (Intel Corporation)

The health care industry is expected to be an early adopter of AI and deep learning to improve patient outcomes, reduce costs, and speed up diagnosis. We have developed models for using AI to diagnose pneumonia, emphysema, and other thoracic pathologies from chest x-rays. Using the Stanford University CheXNet model as inspiration, we explore ways of developing accurate models for this problem with fast parallel training on Zenith, the Intel Xeon-based supercomputer at Dell EMC's HPC and AI Innovation Lab. We explore various network topologies to gain insight into what types of neural networks scale well in parallel and improve training time from days to hours. We then explore transferring this learned knowledge to other radiology subdomains, such as mammography, and whether this leads to better models than developing subdomain models independently.

Best Poster Finalist: no

Full State Quantum Circuit Simulation by Using Lossy Data Compression

Xin-Chuan Wu (University of Chicago, Argonne National Laboratory), Sheng Di (Argonne National Laboratory), Franck Cappello (Argonne National Laboratory), Hal Finkel (Argonne National Laboratory), Yuri Alexeev (Argonne National Laboratory), Frederic T. Chong (University of Chicago)

In order to evaluate, validate, and refine the design of a new quantum algorithm or a quantum computer, researchers and developers need methods to assess their correctness and fidelity. This requires the capabilities of simulation for full quantum state amplitudes. However, the number of quantum state amplitudes increases exponentially with the number of qubits, leading to the exponential growth of the memory requirement. In this work, we present our technique to simulate more qubits than previously reported by using lossy data compression. Our empirical data suggests that we can simulate full state quantum circuits up to 63 qubits with 0.8 petabytes memory.

Best Poster Finalist: no

Enabling Reproducible Microbiome Science through Decentralized Provenance Tracking in QIIME 2

Ahmad Turan Naimey (Northern Arizona University, Pathogen and Microbiome Institute), Christopher Keefe (Northern Arizona University, Pathogen and Microbiome Institute)

In this poster, we demonstrate the ways in which automatic, integrated, decentralized provenance tracking in QIIME 2, a leading microbiome bioinformatics platform, enables reproducible microbiome science. We use sample data from a recent study of arid soil microbiomes (Significant Impacts of Increasing Aridity on the Arid Soil Microbiome; Neilson et al, 2017), to illustrate specific analyses that QIIME 2 supports, and to frame our discussion of the QIIME 2 platform.

QIIME 2 actions yield as outputs artifacts integrating the requested data or visualization with comprehensive data provenance that describes the computational history of that data or visualization, including all methods and parameters involved in its creation. This approach gives users, reviewers, and readers powerful tools for understanding, reproducing, and extending studies. The benefits this approach provides to both the researcher and the scientific community are significant and provide a useful model for research software developers across disciplines.

Best Poster Finalist: no

Optimizing Next Generation Hydrodynamics Code for Exascale Systems

Dana Akhmetova (KTH Royal Institute of Technology), Sumathi Lakshmiranganatha (University of Wyoming), Diptajyoti Mukherjee (Allegheny College), Frederick Oullet (University of Florida), Patrick Payne (Los Alamos National Laboratory), Nicholas Stegmeier (South Dakota State University), Christoph Junghans (Los Alamos National Laboratory), Robert Pavel (Los Alamos National Laboratory), Vinay Ramakrishnaiah (Los Alamos National Laboratory)

Studying continuum dynamics problems computationally can illuminate complex physical phenomena where experimentation is too costly. However, the models used in studying these phenomena usually require intensive calculations, some of which are beyond even the largest supercomputers to date. Emerging high performance computing (HPC) platforms will likely have varied levels of heterogeneity, making hybrid programming with MPI+X essential for achieving optimal performance. This research investigates hybrid programming and unconventional approaches like machine learning for a next generation hydrodynamics code, FleCSALE, in the context of tabular equation of state (EOS). We demonstrate an overall 5x speedup to the code, the use of GPUs to accelerate EOS tabular interpolation, and a proof of concept machine learning approach to EOS.

Best Poster Finalist: no

MPI/OpenMP parallelization of the Fragment Molecular Orbitals Method in GAMESS

Vladimir Mironov (Lomonosov Moscow State University), Yuri Alexeev (Argonne National Laboratory), Dmitri Fedorov (National Institute of Advanced Industrial Science and Technology (AIST))

In this work, we present a novel parallelization strategy for the Fragment Molecular Orbital (FMO) method in the quantum chemistry package GAMESS. The original FMO code has been parallelized only with MPI, which limits scalability of the code on multi-core massively parallel machines. To address this problem, we parallelized FMO with a new hybrid MPI-OpenMP scheme that shows excellent scaling up to 2,048 Intel Xeon Phi nodes (131,072 cores) on Theta supercomputer. MPI-OpenMP code not only scales better compared to MPI code, but also performs up to two times faster and has significantly smaller memory footprint.

Best Poster Finalist: no

Automatic Generation of Mixed-Precision Programs

Logan Moody (Lawrence Livermore National Laboratory, James Madison University), Nathan Pinnow (Lawrence Livermore National Laboratory, Western Washington University), Michael O. Lam (James Madison University, Lawrence Livermore National Laboratory), Harshitha Menon (Lawrence Livermore National Laboratory), Markus Schordan (Lawrence Livermore National Laboratory), G. Scott Lloyd (Lawrence Livermore National Laboratory), Tanzima Islam (Western Washington University)

Floating-point arithmetic is foundational to scientific computing in HPC, and choices about floating-point precision can have a significant effect on the accuracy and speed of HPC codes. Unfortunately, current precision optimization tools require significant user interaction, and few work on the scale of HPC codes due to significant analysis overhead. We propose an automatic search and replacement system that finds the maximum speedup using mixed precision given a required level of accuracy. To achieve this, we integrated three existing analysis tools into a system that requires minimal input from the user. If a speedup is found, our system can provide a ready-to-compile mixed-precision version of the original program.

Best Poster Finalist: no

UPC++ and GASNet-EX: PGAS Support for Exascale Applications and Runtimes

Scott B. Baden (Lawrence Berkeley National Laboratory), Paul H. Hargrove (Lawrence Berkeley National Laboratory), Hadia Ahmed (Lawrence Berkeley National Laboratory), John Bachan (Lawrence Berkeley National Laboratory), Dan Bonachea (Lawrence Berkeley National Laboratory), Steven Hofmeyr (Lawrence Berkeley National Laboratory), Mathias Jacquelin (Lawrence Berkeley National Laboratory), Amir Kamil (Lawrence Berkeley National Laboratory), Brian van Straalen (Lawrence Berkeley National Laboratory)

Lawrence Berkeley National Lab is developing a programming system to support HPC application development using the Partitioned Global Address Space (PGAS) model. This work is driven by the emerging need for adaptive, lightweight communication in irregular applications at exascale. We present an overview of UPC++ and GASNet-EX, including examples and performance results.

GASNet-EX is a portable, high-performance communication library, leveraging hardware support to efficiently implement Active Messages and Remote Memory Access (RMA). UPC++ provides higher-level abstractions appropriate for PGAS programming such as: one-sided communication (RMA), remote procedure call, locality-aware APIs for user-defined distributed objects, and robust support for asynchronous execution to hide latency. Both libraries have been redesigned relative to their predecessors

to meet the needs of exascale computing. While both libraries continue to evolve, the system already demonstrates improvements in microbenchmarks and application proxies.

Best Poster Finalist: no

An Efficient SIMD Implementation of Pseudo-Verlet Lists for Neighbor Interactions in Particle-Based Codes

James Willis (Durham University, Institute for Computational Cosmology), Matthieu Schaller (Leiden Observatory), Pedro Gonnet (Google LLC)

In particle-based simulations, neighbour finding (i.e. finding pairs of particles to interact within a given range) is the most time consuming part of the computation. One of the best such algorithms, which can be used for both Molecular Dynamics (MD) and Smoothed Particle Hydrodynamics (SPH) simulations is the pseudo-Verlet list algorithm. The algorithm improves the neighbour finding by reducing the number of spurious pair-wise distance calculations. This algorithm, however, does not vectorize trivially, and hence makes it difficult to exploit SIMD-parallel architectures. On this poster, we present several novel modifications as well as a vectorization strategy for the algorithm which lead to overall speed-ups over the scalar version of the algorithm of 2.21x for the AVX instruction set (SIMD width of 8), 2.41x for AVX2, and 3.65x for AVX-512 (SIMD width of 16).

Best Poster Finalist: no

Understanding Potential Performance Issues Using Resource-Based alongside Time Models

Nan ding (Lawrence Berkeley National Laboratory), Victor W. Lee (Intel Corporation), Wei Xue (Tsinghua University), Weimin Zheng (Tsinghua University)

Numerous challenges and opportunities are introduced by the complexity and enormous code legacy of HPC applications, the diversity of HPC architectures, and the nonlinearity of interactions between applications and HPC systems. To address these issues, we propose the Resource-based Alongside Time (RAT) modeling method to help to understand the application run-time performance efficiently. First, we use hardware counter-assisted profiling to identify the key kernels and non-scalable kernels in the application. Second, we show how to apply the resource-based profiling into performance models to understand the potential performance issues and predict performance in the regimes of interest to developers and performance analysts. Third, we propose an easy-to-use performance modeling tool for scientists and performance analytics. Our evaluations demonstrate that by only performing a few small-scale profilings, RAT is able to keep the average model error rate around 15% with average performance overheads of 3% in multiple scenarios.

Best Poster Finalist: no

MGRIT Preconditioned Krylov Subspace Method

Ryo Yoda (Kogakuin University), Akihiro Fujii (Kogakuin University), Teruo Tanaka (Kogakuin University)

MGRIT re-discretize the problem with larger time-step width at the coarse-levels, which often cause unstable convergence. We propose a Krylov subspace method with MGRIT preconditioning as a more stable solver. For unstable problems, MGRIT preconditioned Krylov subspace method performed better than MGRIT in terms of the number of iterations. The contributions of the paper are organized as follows. We showed the matrix form of MGRIT operations, and the improvement of eigenvalue or singular-value

distribution. We exemplified MGRIT with Krylov subspace method reaching convergence faster than MGRIT.

Best Poster Finalist: no

Enabling Neutrino and Antineutrino Appearance Observation Measurements with HPC Facilities

Norm Buchanan (Colorado State University), Steven Calvez (Colorado State University), Pengfei Ding (Fermi National Accelerator Laboratory), Derek Doyle (Colorado State University), Alex Himmel (Fermi National Accelerator Laboratory), Burt Holzman (Fermi National Accelerator Laboratory), Jim Kowalkowski (Fermi National Accelerator Laboratory), Andrew Norman (Fermi National Accelerator Laboratory), Alex Sousa (University of Cincinnati), Marc Paterno (Fermi National Accelerator Laboratory), Saba Sehrish (Fermi National Accelerator Laboratory), Brandon White (Fermi National Accelerator Laboratory), Christopher Green (Fermi National Accelerator Laboratory)

When fitting to data with low statistics and near physical boundaries, extra measures need to be taken to ensure proper statistical coverage. The method NOvA uses is called the Feldman-Cousins procedure, which entails fitting thousands of independent pseudoexperiments to generate acceptance intervals that are then used to correct our fits. The scale required by the Feldman-Cousins procedure makes it extremely computationally intensive. In past analyses, it has taken up to several weeks to complete, bottlenecking our final results. Here, I present recent work by members of the NOvA experiment and the SciDAC4 collaboration to enable the use of the supercomputing facilities at NERSC to process our Feldman-Cousins corrections over 50x faster, allowing us to perform more studies, increase the precision of our fits, and produce results quickly.

Best Poster Finalist: no

Large Scale Computation of Quantiles Using MELISSA

Alejandro Ribes (EDF Research and Development), Théophile Terraz (French Institute for Research in Computer Science and Automation (INRIA)), Yvan Fournier (EDF Research and Development), Bertrand Iooss (EDF Research and Development), Bruno Raffin (French Institute for Research in Computer Science and Automation (INRIA))

Quantiles being order statistics, the classical approach for their computation requires availability of the full sample before ranking it. This approach is not suitable at exascale. Large ensembles would need to gather a prohibitively large amount of data. We propose an iterative approach based on the stochastic quantile algorithm of Robbins-Monro. We rely on the Melissa framework, a file avoiding, adaptive, fault tolerant and elastic framework in order to compute in transit ubiquitous quantiles. Quantiles are updated on-the-fly as soon as the in transit parallel server receives results from one of the running simulations. We run 80,000 fluid dynamics parallel simulations of 6M hexahedra and 100 timespteps. They were executed on up to 4800 cores, avoiding 288 TB of file storage. We produce ubiquitous spatio-temporal maps of quantiles and inter-quantile based intervals.

Best Poster Finalist: no

FlowOS-RM: Disaggregated Resource Management System

Ryousei Takano (National Institute of Advanced Industrial Science and Technology (AIST)), Kuniyasu Suzaki (National Institute of Advanced Industrial Science and Technology (AIST)), Hidetaka Koie (National Institute of Advanced Industrial Science and Technology (AIST))

A traditional data center consists of monolithic-servers is confronted with limitations including lack of operational flexibility, low resource utilization, low maintainability, etc. Resource disaggregation is a promising solution to address the above issues. We propose a concept of disaggregated data center architecture called Flow-in-Cloud (FiC) that enables an existing cluster computer to expand an accelerator pool through a high-speed network. FiC is a shared pool of heterogeneous accelerators such as GPU and FPGA, which are directly connected by a circuit-switched network. From the pool of accelerators, a slice is dynamically configured and provided according to a user request. FlowOS-RM manages the entire FiC resources, and supports execution of a user job on provided slices. This poster demonstrates effective resource sharing on the prototype system using a distributed deep learning application.

Best Poster Finalist: no

Programming the EMU Architecture: Algorithm Design Considerations for Migratory-Threads-Based Systems

Mehmet E. Belviranli (Oak Ridge National Laboratory), Seyong Lee (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory)

The decades-old memory bottleneck problem for data-intensive applications is getting worse as the processor core counts continue to increase. Workloads with sparse memory access characteristics only achieve a fraction of a system's total memory bandwidth. EMU architecture provides a radical approach to the issue by migrating the computational threads to the location where the data resides.

In EMU architecture, data distribution and thread creation strategies play a crucial role in achieving optimal performance in the EMU platform. In this work, we identify several design considerations that need to be taken care of while developing applications for the new architecture and we evaluate their performance effects on the EMU-chick hardware.

Best Poster Finalist: no

OpenACC to FPGA: A Directive-Based High-Level Programming Framework for High-Performance Reconfigurable Computing

Seyong Lee (Oak Ridge National Laboratory), Jacob Lambert (University of Oregon), Jungwon Kim (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory), Allen D. Malony (University of Oregon)

Accelerator-based heterogeneous computing has become popular solutions for power-efficient high performance computing (HPC). Along these lines, Field Programmable Gate Arrays (FPGAs) have offered more advantages in terms of performance and energy efficiency for specific workloads than other accelerators. Nevertheless, FPGAs have traditionally suffered several disadvantages limiting their deployment in HPC systems, mostly due to the challenges of programmability and portability. We present a directive-based, high-level programming framework for high-performance reconfigurable computing. It takes a standard, portable OpenACC C program as input and generates a hardware configuration file for execution on FPGAs. We implemented this prototype system in our open-source OpenARC compiler, which uses the Altera OpenCL compiler as its backend. Preliminary evaluation of the proposed framework on an Intel Stratix~V with five OpenACC benchmarks demonstrates that our proposed FPGA-specific compiler optimizations and novel OpenACC pragma extensions assist the compiler in generating more efficient FPGA programs.

Best Poster Finalist: no

Tensor-Optimized Hardware Accelerates Fused Discontinuous Galerkin Simulations

Alexander Breuer (University of California, San Diego), Alexander Heinecke (Intel Corporation), Yifeng Cui (San Diego Supercomputer Center)

In recent years the compute/memory balance of processors has been continuously shifting towards compute. The rise of Deep Learning, based on matrix multiplications, accelerated this path, especially in terms of single precision and lower precision compute. An important research question is if this development can be leveraged for traditional HPC. We demonstrate that a high-order discontinuous Galerkin solver for seismic wave propagation can execute in single precision without loss of modeling accuracy. Additionally, we extended its kernels to support the Intel Knights Mill CPU with 14 TFLOPS of single precision deep-learning performance. This allows us to harvest the hardware's special compute capabilities, even in an application with sparse linear algebra kernels. On cluster-level, Knights Mill can obtain the same application performance as the latest top-bin dual-socket Intel Xeon Platinum nodes, while consuming lower power. Compared to the HPC-focused Knights Landing processor, scenario-dependent speed-ups of up to 1.6× are possible.

Best Poster Finalist: no

AI Matrix – Synthetic Benchmarks for DNN

Wei Wei (Alibaba Inc), Lingjie Xu (Alibaba Inc), Lingling Jin (Alibaba Inc), Wei Zhang (Alibaba Inc), Tianjun Zhang (University of California, Berkeley)

The current AI benchmarks suffer from a number of drawbacks. First, they cannot adapt to the emerging changes of deep learning (DL) algorithms and are fixed once selected. Second, they contain tens to hundreds of applications and have very long running time. Third, they are mainly selected from open sources, which are restricted by copyright and not representable of the proprietary applications. To address these drawbacks, this work firstly proposes a synthetic benchmark framework that generates a small number of benchmarks that best represent a broad range of applications using their profiled workload characteristics. The synthetic benchmarks can adapt to new DL algorithms by re-profiling new applications and updating itself, greatly reduce number of benchmark tests and running time, and strongly represent DL applications of interests. The framework is validated by using log data profiled from DL models running on Alibaba AI platform, and is representable of real workload characteristics.

Best Poster Finalist: no

Applying the Execution-Cache-Memory Model: Current State of Practice

Georg Hager (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Jan Eitzinger (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Julian Hornich (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Francesco Cremonesi (Swiss Federal Institute of Technology in Lausanne), Christie L. Alappat (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Thoams Roehl (University of Erlangen-Nuremberg), Gerhard Wellein (University of Erlangen-Nuremberg, Erlangen Regional Computing Center)

The ECM (Execution-Cache-Memory) model is an analytic, resource-based performance model for steady-state loop code running on multicore processors. Starting from a machine model, which describes the interaction between the code and the hardware, and static code analysis, it allows an accurate prediction of

the runtime of sequential loop code. Together with a scaling assumption, it also gives a performance scaling prediction. This poster summarizes the current state of practice in constructing and applying the ECM model, points out problems and open questions, and applies the model to three new and nontrivial use cases. For the first time, overlap assumptions for all relevant CPU architectures in high performance computing are presented.

Best Poster Finalist: yes

WarpX: Toward Exascale Modeling of Plasma Particle Accelerators

Maxence Thevenet (Lawrence Berkeley National Laboratory), Jean-Luc Vay (Lawrence Berkeley National Laboratory), Ann Almgren (Lawrence Berkeley National Laboratory), John Bell (Lawrence Berkeley National Laboratory), Remi Lehe (Lawrence Berkeley National Laboratory), Andrew Myers (Lawrence Berkeley National Laboratory), Jaehong Park (Lawrence Berkeley National Laboratory), Olga Shapoval (Lawrence Berkeley National Laboratory), Weiqun Zhang (Lawrence Berkeley National Laboratory), Lixin Ge (SLAC National Accelerator Laboratory), Mark Hogan (SLAC National Accelerator Laboratory), Cho Ng (SLAC National Accelerator Laboratory), Dave Grote (Lawrence Livermore National Laboratory)

Turning the current experimental plasma accelerator state-of-the-art from a promising technology into mainstream scientific tools depends critically on high-performance, high-fidelity modeling of complex processes that develop over a wide range of space and time scales. As part of the U.S. Department of Energy's Exascale Computing Project, a team composed of LBNL, SLAC and LLNL researchers is developing a new plasma accelerator simulation tool: WarpX. We will present the code structure and how it articulates around its main components: the new Particle-In-Cell Scalable Application Resource (PICSAR) and the adaptive mesh refinement library AMReX, which are combined with redesigned elements of the Warp code, in the new WarpX software. The status, examples of convergence, scaling and applications will be presented.

Best Poster Finalist: no

Job Simulation for Large-Scale PBS-Based Clusters with the Maui Scheduler

Georg Zitzlsberer (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic), Branislav Jansik (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic), Jan Martinovic (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic)

For large-scale High Performance Computing centers with a wide range of different projects and heterogeneous infrastructures, efficiency is an important consideration. Understanding how compute jobs are scheduled is necessary for improving the job scheduling strategies in order to optimize cluster utilization and job wait times. This increases the importance of a reliable simulation capability, which in turn requires accuracy and comparability with historic workloads from the cluster. Not all job schedulers have a simulation capability, including the Portable Batch System (PBS) resource manager. Hence, PBS based centers have no direct way to simulate changes and optimizations before they are applied to the production system. We propose and discuss how to run job simulations for large-scale PBS based clusters with the Maui Scheduler. For validation purposes, we use historic workloads collected at the IT4Innovations supercomputing center, and demonstrate the viability of our approach.

Best Poster Finalist: no

Script of Scripts Polyglot Notebook and Workflow System

Gao Wang (University of Chicago), Man Chong Leong (Rice University), Bo Peng (University of Texas, MD Anderson Cancer Center)

Computationally intensive disciplines such as computational biology often use tools implemented in different languages and analyze data on high-performance computing systems. Although scientific workflow systems can powerfully execute large-scale data-processing, they are not suitable for ad hoc data analysis. Interactive tools such as Jupyter Notebook can be used for data exploration, but it remains difficult to work with multiple scripting languages and to streamline analysis for batch data processing. To bridge the gap between interactive and workflow systems we developed Script of Scripts (SoS), which consists of a polyglot notebook that supports multiple languages in a single notebook and a workflow engine that provides an intuitive syntax for multi-style workflows and a unified interface for executing tasks on a variety of computing platforms. By allowing the use of SoS workflow engine in a polyglot notebook environment, SoS provides a unified environment for both interactive data analysis and batch data processing.

Best Poster Finalist: no

Enabling High-Level Graph Processing via Dynamic Tasking

Maurizio Drocco (Pacific Northwest National Laboratory), Vito Giovanni Castellana (Pacific Northwest National Laboratory), Marco Minutoli (Pacific Northwest National Laboratory), Antonino Tumeo (Pacific Northwest National Laboratory), John Feo (Pacific Northwest National Laboratory)

Data-intensive computing yields irregular and unbalanced workloads, in particular on large-scale problems running on distributed systems. Task-based runtime systems are commonly exploited to implement higher-level data-centric programming models, promoting multithreading and asynchronous coordination for performance. However, coping with dynamic workloads (e.g., those yielded by large-scale graph processing) is challenging.

In this work, we took an exploratory approach to overcome some typical bottlenecks in tasking systems. In particular, we propose 1. a novel task allocator based on dynamic per-thread allocation and all-to-all recycling networks, and 2. a reservation-free remote spawning schema, based on receiver-side buffering and back-pressure feedback/sensing to avoid overflows.

As a proof of concept, we implemented the proposed techniques underneath a high-level library of distributed C++ containers. Preliminary experimental evaluation shows consistent scalability, a neat improvement in performance (e.g., 1.5x speedup with respect to the original code over an 8M-nodes graph), and less sensitiveness to parameter tuning.

Best Poster Finalist: no

An Alternative Approach to Teaching Bigdata and Cloud Computing Topics at CS Undergraduate Level

Debzani Deb (Winston-Salem State University), Muztaba Fuad (Winston-Salem State University), Keith Irwin (Winston-Salem State University)

Big data and cloud computing collectively offer a paradigm shift in the way businesses are now acquiring, using and managing information technology. This creates the need for every CS student to be equipped with foundation knowledge in this collective paradigm and to possess some hands-on-experience in deploying and managing big data applications in the cloud. We argue that, for substantial coverage of big

data and cloud computing concepts and skills, the relevant topics need to be integrated into multiple core courses of undergraduate CS curriculum rather than creating additional standalone core or elective courses. Our approach to including these topics is to develop learning modules for specific core courses in which their coverage might find an appropriate context. In this poster, three such modules are presented and our classroom experiences during these interventions are documented. Our objective is to share our experience and to receive feedback about our approach.

Best Poster Finalist: no

Binarized ImageNet Inference in 29us

Tong Geng (Boston University, Pacific Northwest National Laboratory), Ang Li (Pacific Northwest National Laboratory), Tianqi Wang (Boston University), Shuaiwen Leon Song (Pacific Northwest National Laboratory), Martin Herbordt (Boston University)

We propose a single-FPGA-based accelerator for ultra-low-latency inference of ImageNet in this work. The design can complete the inference of Binarized AlexNet within 29us with accuracy comparable to other BNN implementations. We achieve this performance with the following contributions: 1. We completely remove floating-point from NL through layer fusion. 2. By using model parallelism rather than data parallelism, we can simultaneously configure all layers and the control flow graphs. Also, the design is flexible enough to achieve nearly perfect load balancing, leading to extremely high resource utilization. 3. All convolution layers are fused and processed in parallel through inter-layer pipelining. Therefore, in case the pipeline is full, latency is just the delay of a single convolution layer plus the FC layers. Note that the dependency pattern of the FC layer prevents it from being integrated into the current pipeline.

Best Poster Finalist: no

Refactoring and Optimizing Multiphysics Combustion Models for Data Parallelism

Christopher Stone (US Department of Defense HPC Modernization Program, Engility Corporation), Alexei Poludnenko (Texas A&M University), Brian Taylor (US Air Force Research Laboratory)

High-fidelity combustion simulations combine high-resolution computational fluid dynamics numerical methods with multi-physics models to capture chemical kinetics and transport processes. These multi-physics models can dominate the computation cost of the simulation. Due to the high cost of combustion simulations and the important role simulations play in propulsion and power research, acceleration methods are needed to reduce the computational time and cost. Multi-physics models within each mesh cell are often independent leading to significant parallelism. However, the iterative algorithms often impede efficient SIMD data parallelism, a key performance feature on modern HPC systems. Refactoring methods for multi-physics models (e.g., kinetics, equation-of-state, diffusion) with nonuniform workloads are demonstrated and benchmarked on a range of platforms (AVX2, KNL, AVX-512). Realized speed-ups over 6x were achieved on KNL and 4x on Skylake (SKX) for complex chemical kinetics models and over 3x on SKX for iterative EOS computations.

Best Poster Finalist: no

Tensorfolding: Improving Convolutional Neural Network Performance with Fused Microkernels

Michael Anderson (Intel Corporation), Evangelos Georganas (Intel Corporation), Sasikanth Avancha (Intel Corporation), Alexander Heinecke (Intel Corporation)

Convolution layers are prevalent in many classes of deep neural networks, including Convolutional Neural Networks (CNNs) which provide state-of-the-art results for tasks like image recognition, neural machine translation and speech recognition. In the recent past, several techniques to improve generalization capabilities of neural networks have been developed; the most prominent and successful is batch normalization. In deep neural network training, the batch normalization layer consists of a memory-bandwidth bound kernel. On the latest Intel Skylake based Xeon processors, a significant portion of execution time is spent in this kernel. By leveraging the CPU's large caches and its latency-optimized execution model, we are able to reduce this kernel's time to a bare minimum while allowing to improve forward pass layer runtimes by 21% compared to an unfused implementation and by 2% compared to a fused implementation.

Best Poster Finalist: no

MATEDOR: MAtrix, TEnsor, and Deep-Learning Optimized Routines

Ahmad Abdelfattah (University of Tennessee), Jack Dongarra (University of Tennessee), Stanimire Tomov (University of Tennessee), Ichitaro Yamazaki (University of Tennessee), Azzam Haidar (Nvidia Corporation)

The MAtrix, TEnsor, and Deep-learning Optimized Routines (MATEDOR) project develops software technologies and standard APIs, along with a sustainable and portable library, for large-scale computations that can be broken down into very small matrix or tensor computations. The main target of MATEDOR is to accelerate applications from important fields that fit this profile, including deep learning, data mining, astrophysics, image and signal processing, hydrodynamics, and more.

MATEDOR is a high-performance numerical library for batched linear algebra subroutines autotuned for modern processor architectures and system designs. The MATEDOR library includes LAPACK-compliant routines that target many small dense problems, tensor, and application-specific operations, e.g., for deep-learning. These routines are constructed as much as possible out of calls to batch BLAS routines and their look-alikes required in sparse computation context.

Best Poster Finalist: no

Distributed Adaptive Radix Tree for Efficient Metadata Search on HPC Systems

Wei Zhang (Texas Tech University), Houjun Tang (Lawrence Berkeley National Laboratory), Suren Byna (Lawrence Berkeley National Laboratory), Yong Chen (Texas Tech University)

Affix-based search allows users to retrieve data without the need to remember all relevant information precisely. While building an inverted index to facilitate efficient affix-based search is a common practice for standalone databases and desktop file systems, they are often insufficient for high-performance computing (HPC) systems due to the massive amount of data and the distributed nature of the storage. In this poster, we present Distributed Adaptive Radix Tree (DART) which enables scalable and efficient affix-based search. DART maintains a balanced keyword distribution and optimizes for excessive keyword requests dynamically at scale. Our evaluation shows that compared with the “full string hashing” used by the commonly adopted DHT approach, DART achieves up to 55x throughput speedup for prefix and suffix search, and has a comparable throughput for exact and infix search. Also, DART maintains balanced keyword distribution and alleviates excessive query workload on popular keywords.

Best Poster Finalist: no

Improving Error-Bounded Lossy Compression for Cosmological N-Body Simulation

Sihuan Li (University of California, Riverside), Sheng Di (Argonne National Laboratory), Xin Liang (University of California, Riverside), Zizhong Chen (University of California, Riverside), Franck Cappello (Argonne National Laboratory)

Cosmological simulations may produce extremely large amount of data, such that its successful run depends on large storage capacity and huge I/O bandwidth, especially in the exascale computing scale. Effective error-bounded lossy compressors with both high compression ratios and low data distortion can significantly reduce the total data size while guaranteeing the data valid for post-analysis. In this poster, we propose a novel, efficient compression model for cosmological N-body simulation framework, by combining the advantages of both space-based compression and time-based compression. The evaluation with a well-known cosmological simulation code shows that our proposed solution can get much higher compression quality than other existing state-of-the-art compressors, with comparable compression/decompression rates.

Best Poster Finalist: no

VeloC: Very Low Overhead Checkpointing System

Bogdan Nicolae (Argonne National Laboratory), Franck Cappello (Argonne National Laboratory), Adam Moody (Lawrence Livermore National Laboratory), Elsa Gonsiorowski (Lawrence Livermore National Laboratory), Kathryn Mohror (Lawrence Livermore National Laboratory)

Checkpointing large amounts of related data concurrently to stable storage is a common I/O pattern of many HPC applications. However, such a pattern frequently leads to I/O bottlenecks that lead to poor scalability and performance. As modern HPC infrastructures continue to evolve, there is a growing gap between compute capacity vs. I/O capabilities. Furthermore, the storage hierarchy is becoming increasingly heterogeneous: in addition to parallel file systems, it comprises burst buffers, key-value stores, deep memory hierarchies at node level, etc. In this context, state of art is insufficient to deal with the diversity of vendor APIs, performance and persistency characteristics. This poster proposes VeloC, a low-overhead checkpointing system specifically designed to address the checkpointing needs of future exascale HPC systems. VeloC offers a simple API at user level, while employing an advanced multi-level resilience strategy that transparently optimizes the performance and scalability of checkpointing by leveraging heterogeneous storage.

Best Poster Finalist: no

Estimating Molecular Dynamics Chemical Shift with GPUs

Eric F. Wright (University of Delaware), Mauricio H. Ferrato (University of Delaware)

Experimental chemical shifts (CS) from solution and solid state magic-angle-spinning nuclear magnetic resonance spectra provide atomic level data for each amino acid within a protein or complex. However, structure determination of large complexes and assemblies based on NMR data alone remains challenging due to the complexity of the calculations. Here, we present a hardware accelerated strategy for the estimation of NMR chemical-shifts of large macromolecular complexes. We demonstrate the feasibility of our approach in systems of increasing complexity ranging from 2,000 to 11,000,000 atoms.

Best Poster Finalist: no

Using Thrill to Process Scientific Data on HPC

Mariia Karabin (Clemson University, Los Alamos National Laboratory), Xinyu Chen (University of New Mexico), Supreeth Suresh (University of Wyoming), Ivo Jimenez (University of California, Santa Cruz), Li-Ta Lo (Los Alamos National Laboratory), Pascal Grosset (Los Alamos National Laboratory)

With ongoing improvement of computational power and memory capacity, the volume of scientific data keeps growing. To gain insights from vast amounts of data, scientists are starting to look at Big Data processing and analytics tools such as Apache Spark. In this poster, we explore Thrill, a framework for big data computation on HPC clusters that provides an interface similar to systems like Apache Spark but delivers higher performance since it is built on C++ and MPI. Using Thrill, we implemented several analytics operations to post-process and analyze data from plasma physics and molecular dynamics simulations. Those operations were implemented with less programming effort than hand-crafted data processing programs would require and obtained preliminary results which were verified by scientists at LANL.

Best Poster Finalist: no

GPU Acceleration at Scale with OpenPower Platforms in Code_Saturne

Samuel Antao (IBM), Charles Moulinec (Science and Technology Facilities Council, UK), Yvan Fournier (EDF Research and Development), Robert Sawko (IBM), Malgorzata Zimon (IBM), Christopher Thompson (IBM), Alex Skillen (Science and Technology Facilities Council, UK), Juan Uribe (EDF Research and Development), David Emerson (Science and Technology Facilities Council, UK)

Code_Saturne is a widely used computational fluid dynamics software package that uses finite-volume methods to simulate different kinds of flows tailored to tackle multi-bilion-cell unstructured mesh simulations. This class of codes has shown to be challenging to accelerate on GPUs as they consist of many kernels and regular inter-process communication in between. In this poster we show how template pack expansion with CUDA can combine multiple kernels into a single one reducing launching latencies and along with the specification of data environments help reduce host-device communication. We tested these techniques on ORNL Summit Supercomputer based on OpenPOWER platform delivering almost 3x speedup over CPU-only runs on 256 nodes. We also show how the latest generation NVLINK(TM) interconnect available in POWER9(TM)improves scaling efficiency, enabling consistent GPU acceleration with just 100K-cells per process.

Best Poster Finalist: yes

Large-Message Size Allreduce at Wire Speed for Distributed Deep Learning

Kenji Tanaka (Japan Telegraph and Telephone Corporation), Yuki Arikawa (Japan Telegraph and Telephone Corporation), Kenji Kawai (Japan Telegraph and Telephone Corporation), Junichi Kato (Japan Telegraph and Telephone Corporation), Tsuyoshi Ito (Japan Telegraph and Telephone Corporation), Huy Cu Ngo (Japan Telegraph and Telephone Corporation), Kazutaka Morita (Japan Telegraph and Telephone Corporation), Fumiaki Miura (Japan Telegraph and Telephone Corporation), Takeshi Sakamoto (Japan Telegraph and Telephone Corporation), Satoshi Shigematsu (Japan Telegraph and Telephone Corporation)

In large-scale distributed deep learning, the Allreduce operation for large messages (100 KB or more) is critical for gathering gradients from multiple worker nodes and broadcasting the sum of the gradients to them. When the message is large, the latency in Allreduce operation would make it difficult to take advantage of large-scale distributed deep learning. To reduce the latency, we devised a dataflow architecture with an Allreduce-specific hardware accelerator that performs data aggregation and reduction

while data is being transferred. The accelerator is designed to immediately start Allreduce operation before an entire message is recived. Furthermore, Allreduce can be operated at wire speed by vectorizing the gradients and summing them in parallel. Experimental results reveal that the proposed architecture performs Allreduce at 96% of wire speed for a large message. Moreover, the latency of Allreduce is reduced by 65% compared with a state-of-the-art Allreduce method when applied for ResNet-50.

Best Poster Finalist: no

Sol: Transparent Neural Network Acceleration Platform

Nicolas Weber (NEC Laboratories Europe, NEC Corporation)

With the usage of neural networks in a wide range of application fields, the necessity to execute these efficiently on high performance hardware is one of the key problems for artificial intelligence (AI) framework providers. More and more new specialized hardware types and corresponding libraries appear from various manufacturers. The biggest problem arising is that these libraries usually are only supported by a very limited set of AI frameworks and interoperability can become an issue. In this extended abstract we present Sol, a transparent middleware for neural network acceleration. Sol comes with an optimizing compiler engine, allowing to use device specific libraries and to implement own optimizations, that can be leveraged on all target devices. In contrast to other projects Sol explicitly aims at optimizing prediction and training of neural networks.

Best Poster Finalist: no

Detection of Silent Data Corruptions in Smooth Particle Hydrodynamics Simulations

Aurélien Cavelan (University of Basel), Florina M. Ciorba (University of Basel), Ruben M. Cabezón (University of Basel)

Soft errors, such as silent data corruptions (SDCs) hinder the correctness of large-scale scientific applications. Ghost replication (GR) is proposed herein as the first SDCs detector relying on the fast error propagation inherent to applications that employ the smooth particle hydrodynamics (SPH) method. GR follows a two-steps selective replication scheme. First, an algorithm selects which particles to replicate on a different process. Then, a different algorithm detects SDCs by comparing the data of the selected particles with the data of their ghost. The overhead and scalability of the proposed approach are assessed through a set of strong-scaling experiments conducted on a large HPC system under error-free conditions, using upwards of 3, 000 cores. The results show that GR achieves a recall and precision similar to that of full replication methods, at only a fraction of the cost, with detection rates of 91–99.9%, no false-positives, and an overhead of 1–10%.

Best Poster Finalist: no

DeepSim-HiPAC: Deep Learning High Performance Approximate Calculation for Interactive Design and Prototyping

Ahmed Al-Jarro (Fujitsu Laboratories Ltd), Serban Georgescu (Fujitsu Laboratories Ltd), Yasumoto Tomita (Fujitsu Laboratories Ltd), Kouta Nakashima (Fujitsu Laboratories Ltd)

We present a data-driven technique that can learn from physical-based simulations for the instant prediction of field distribution for 3D objects. Such techniques are extremely useful when considering, for example, computer aided engineering (CAE), where computationally expensive simulations are

often required. To accelerate this process, we propose a deep learning framework that can predict the principal field distribution given a 3D object. This work allows us to learn a system's response using simulation data of arbitrarily shaped objects and an auto-encoder inspired deep neural network that maps the input of the 3D object shape to its principal 3D field distribution. We show that our engine, DeepSim-HiPAC, can estimate field distribution for two distinctive applications: micro-magnetics design in computational electromagnetics (CEM) and interactive cooling systems design in computational fluid dynamics (CFD), several orders of magnitude faster, up to 250000X, than the native calculations and at a cost of low error rate.

Best Poster Finalist: no

Top-Down Performance Analysis of Workflow Applications

Christian Herold (Technical University Dresden), Bill Williams (Technical University Dresden)

Scientific simulation frameworks are common to use on HPC systems. They contain parallelized algorithms and provide various solvers for a specific application domain. Usually, engineers execute multiple steps to solve a particular problem which are often distributed over multiple jobs. Finding performance bottlenecks and the causing step in such a complex system is very difficult. herefore in this work, we present a top-down approach that provides summarized performance metrics for the workflow, jobs and job steps. These summaries guides the user to identify inefficiencies and determine the causing job step. Finally, Vampir can be used for a detailed analysis of the regarding execution in order to resolve the issue.

Best Poster Finalist: no

Convolutional Neural Networks for Coronary Plaque Classification in Intravascular Optical Coherence Tomography (IVOCT) Images

Chaitanya Kolluru (Case Western Reserve University), David Prabhu (Case Western Reserve University), Yanzan Gharaibeh (Case Western Reserve University), David Wilson (Case Western Reserve University), Sanjaya Gajurel (Case Western Reserve University)

Currently, IVOCT is the only imaging technique with the resolution necessary to identify vulnerable thin cap fibro-atheromas (TCFAs). IVOCT also has greater penetration depth in calcified plaques as compared to Intravascular Ultrasound (IVUS). Despite its advantages, IVOCT image interpretation is challenging and time consuming with over 500 images generated in a single pullback. In this poster, we propose a method to automatically classify A-lines in IVOCT images using a convolutional neural network. Conditional random fields were used to clean network predictions across frames. The neural network was trained using a dataset of nearly 4,500 image frames across 48 IVOCT pullbacks. Ten-fold cross validation with held-out pullbacks resulted in a classification accuracy of roughly 76% for fibrocalcific, 84% for fibrolipidic, and 85% for other. Classification results across frames displayed in en face view matched closely to annotated counterparts.

Best Poster Finalist: no

Compiling SIMT Programs on Multi- and Many-Core Processors with Wide Vector Units: A Case Study with CUDA

Hancheng Wu (North Carolina State University), John Ravi (North Carolina State University), Michela Becchi (North Carolina State University)

There has been an increasing interest in SIMT programming tools for multi- and manycore (co)processors with wide vector extensions. In this work, we study the effective implementation of a SIMT programming model (a subset of CUDA C) on Intel platforms with 512-bit vector extensions (hybrid MIMD/SIMD architectures). We first propose a set of compiler techniques to transform programs written using a SIMT programming model into code that leverages both the x86 cores and the vector units of a hybrid MIMD/SIMD architecture, thus providing programmability, high system utilization and portability. We then evaluate the proposed techniques on various hybrid systems using microbenchmarks and real-world applications. Finally, we point out the main challenges in supporting the SIMT model on hybrid systems.

Best Poster Finalist: no

A Massively Parallel Evolutionary Markov Chain Monte Carlo Algorithm for Sampling Complicated Multimodal State SpacesState

Wendy K. Tam Cho (University of Illinois), Yan Liu (University of Illinois)

We develop an Evolutionary Markov Chain Monte Carlo (EMCMC) algorithm for sampling from large multi-modal state spaces. Our algorithm combines the advantages of evolutionary algorithms (EAs) as optimization heuristics and the theoretical convergence properties of Markov Chain Monte Carlo (MCMC) algorithms for sampling from unknown distributions. We harness massive computational power with a parallel EA framework that guides a large set of Markov chains. Our algorithm has applications in many different fields of science. We demonstrate its effectiveness with an application to political redistricting.

Best Poster Finalist: no

MLModelScope: Evaluate and Measure Machine Learning Models within AI Pipelines

Abdul Dakkak (University of Illinois), Cheng Li (University of Illinois), Wen-mei Hwu (University of Illinois), Jinjun Xiong (IBM)

The current landscape of Machine Learning (ML) and Deep Learning (DL) is rife with non-uniform frameworks, models, and system stacks but lacks standard tools to facilitate the evaluation and measurement of models. Due to the absence of such tools, the current practice for evaluating and comparing the benefits of proposed AI innovations (be it hardware or software) on end-to-end AI pipelines is both arduous and error prone — stifling the adoption of the innovations. We propose MLModelScope— a hardware/software agnostic platform to facilitate the evaluation, measurement, and introspection of ML models within AI pipelines. MLModelScope aids application developers in discovering and experimenting with models, data scientists developers in replicating and evaluating for publishing models, and system architects in understanding the performance of AI workloads.

Best Poster Finalist: no

A Compiler Framework for Fixed-Topology Non-Deterministic Finite Automata on SIMD Platforms

Marziyeh Nourian (North Carolina State University), Hancheng Wu (North Carolina State University), Michela Becchi (North Carolina State University)

Automata traversal acceleration has been studied on various parallel platforms. Many existing acceleration methods store finite automata states and transitions in memory. For these designs

memory size and bandwidth are the main limiting factors to performance and power efficiency. Many applications, however, require processing several fixed-topology automata that differ only in the symbols associated to the transitions. This property enables the design of alternative, memory-efficient solutions. We target fixed-topology non-deterministic finite automata (NFAs) and propose a memory-efficient design, suitable to SIMD architectures, that embeds the automata topology in code and stores only the transition symbols in memory. We design a compiler that automates deployment of this design on SIMD platforms for a set of fixed-topology NFAs. Our compiler framework performs a combination of platform-agnostic and platform-specific design decisions and optimizations. This poster describes the compiler toolchain and shows the achieved throughput on GPU and Intel SIMD devices.

Best Poster Finalist: no

A Low-Communicaton Method to Solve Poisson's Equation on Locally-Structured Grids

Brian Van Straalen (Lawrence Berkeley National Laboratory), Peter McCorquodale (Lawrence Berkeley National Laboratory), Phil Colella (Lawrence Berkeley National Laboratory), Christos Kavouklis (Lawrence Livermore National Laboratory)

This poster describes a new algorithm, Method of Local Corrections (MLC), and a high-performance implementation for solving Poisson's equation with infinite-domain boundary conditions, on locally-refined nested rectangular grids. The data motion is comparable to that of only a single V-cycle of multigrid, and hence is an order of magnitude smaller than traditional multigrid iteration. The computational kernels are 3D FFTs on small domains. Strong scaling tests on 64 to 4096 cores on NERSC Cori I (Haswell) show over 60% efficiency, and weak scaling by replication tests over 64 to 32768 cores show 92% efficiency on the same platform. We find comparable solve times between HPGMG on a uniform grid with one billion grid points, and MLC on the same number of grid points adaptively distributed. MLC is designed for AMR, able to solve problems with much higher resolution at the finest level than an algorithm on a uniform grid.

Best Poster Finalist: no

Floating-Point Autotuner for CPU-Based Mixed-Precision Applications

Ruidong Gu (North Carolina State University), Paul A. Beata (North Carolina State University), Michela Becchi (North Carolina State University)

In this poster, we present the design and development of an autotuning tool for floating-point code. The goal is to balance accuracy and performance in order to produce an efficient and accurate mixed-precision program. The tuner starts by maximizing accuracy through the use of a high-precision library called CAMPARY and then achieves performance gains under a given error bound by tuning down groups of variables and operations from the higher precision down to double precision. We tested our tuning strategy on a computational fluid dynamics benchmark where we show a 4x speedup relative to the fully high-precision version during the iterative tuning process and achieve an average absolute error of 2.8E-16 compared with the reference solution computed using the 256-bit GNU MPFR extended precision library.

Best Poster Finalist: no

8:30 am - 5:00 pm

ACM Student Research Competition Posters

Session Description: SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom. The schedule of the ACM Student Research Competition session will be made available Wednesday evening form the results of the semi-finalists selection that will happen after the poster session.

SC18 Research Posters

SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Precomputing Outputs of Hidden Layers to Speed Up Deep Neural Network Training

Sohil Lal Shrestha (University of Texas, Arlington)

Deep learning has recently emerged as a powerful technique for many tasks including image classification. A key bottleneck of deep learning is that the training phase takes a lot of time, since state-of-the-art deep neural networks have millions of parameters and hundreds of hidden layers. The early layers of these deep neural networks have the fewest parameters but take up the most computation.

In this work, we reduce training time by progressively freezing hidden layers, pre-computing their output and excluding them from training in both forward and backward paths in subsequent iterations. We compare this technique to the most closely related approach for speeding up the training process of neural network.

Through experiments on two widely used datasets for image classification, we empirically demonstrate that our approach can yield savings of up to 25% wall-clock time during training with no loss in accuracy.

Designing Shared Address Space MPI Libraries in Many-Core Era

Jahanzeb Maqbool Hashmi (Ohio State University)

The emergence of modern multi-/many-cores has put more emphasis on optimizing intra-node communication. Existing designs in MPI libraries that work on the concept of distributed address spaces incur the overhead of intermediate memory copies to stage the data between processes. This can lead to severe performance degradation especially on emerging many-core architectures like Intel Skylake and IBM OpenPOWER. This work proposes a high-performance "shared address-space"-based MPI point-to-point and collective communication designs using XPMEM. We first characterize the bottlenecks associated with XPMEM based communication and propose new designs for efficient MPI large message communication. Then we propose novel collective designs that are contention-free and offer true zero-copy reduction operations. The proposed designs are evaluated on different multi-/many-core architectures using various micro-benchmarks and application kernels such as MiniAMR and AlexNet DNN training on CNTK. The proposed designs have shown significant performance

improvement over state-of-the-art available in MPI libraries.

Modeling Single-Source Shortest Path Algorithm Dynamics to Control Performance and Power Tradeoffs

Sara Karamati (Georgia Institute of Technology), Jeffrey Young (Georgia Institute of Technology), Rich Vuduc (Georgia Institute of Technology)

This work presents a new methodology to improve the performance of parallel algorithms by tuning the amount of available parallelism for execution throughout the runtime. As such, we expose key parameters controlling the performance and parallelism of the algorithm and build a software-based controller with the objective of maintaining the optimal performance. Our controller allows for tuning the level of parallelism executed in each time epoch to optimize for performance while preserving power usage. More specifically, our experimental evaluation focuses on a tunable variation of a GPU-based delta-stepping algorithm for computing the single-source shortest path (SSSP); As the available parallelism for the delta-stepping SSSP is highly irregular and strongly input-dependent, our extensive experiments show that average power can be reduced while average parallelism is increased. This increase in average parallelism provides substantial energy savings, independent of the hardware.

Accelerating DNA Long Read Mapping with Emerging Technologies

Roman Kaplan (Israel Institute of Technology)

DNA sequencing technologies output only short fragments of a genome, called reads. New single-molecule real-time sequencing technologies can produce long reads, up to tens of thousands base pairs, within minutes. However, these long reads may contain up to 15% errors.

To construct a genome from DNA reads, a computationally expensive bioinformatics task, read mapping, is required. Read mapping finds the best-fitting location for each DNA read on a long reference sequence. The length and error rate of long reads poses a challenge for existing read mapping hardware solutions, designed for short reads with low error rates. This work presents a novel DNA read mapping hardware architecture, RASSA. RASSA is a Resistive Approximate Similarity Search Accelerator that exploits charge distribution and parallel in-memory processing to reflect a mismatch count between DNA sequences. RASSA implementation of long read DNA mapping outperforms state-of-the-art long read mapping solution by 16-77x with comparable accuracy.

SimFS: A Simulation Data Virtualizing File System Interface

Salvatore Di Girolamo (ETH Zurich)

In the big (simulation) data era, simulations often produce petabytes of data to be stored in parallel filesystems or large-scale databases. This data is accessed, often by thousands of analysts and scientists, over the course of decades. However, storing these volumes of data for long time periods of time is not cost effective and, in some cases, practically impossible.

SimFS transparently virtualizes the simulation output, relaxing the storage requirements and re-simulating missing data on-demand. SimFS monitors the analysis access pattern in order to decide (1) which data to store and (2) apply prefetching techniques to improve the analysis performance. SimFS

enables a trade-off between on-disk solutions, where all the simulation data is stored on disk, and in-situ, where no data is stored and analyses are always coupled with simulations. Overall, by exploiting the growing computing power and relaxing the storage capacity requirements, SimFS offers a viable path towards exa-scale simulations.

Holistic Root Cause Analysis of Node Failures in Production HPC

Anwesha Das (North Carolina State University)

Production HPC clusters endure failures incurring computation and resource wastage. Despite the presence of various failure detection and prediction schemes, a comprehensive understanding of how nodes fail considering various components and layers of the system is required for sustained resilience. This work performs a holistic root cause diagnosis of node failures using a measurement-driven approach on contemporary system logs that can help vendors and system administrators support exascale resilience.

Our work shows that lead times can be increased by at least 5 times if external subsystem correlations are considered as opposed to considering the events of a specific node in isolation. Moreover, when detecting sensor measurement outliers and interconnect related failures, triggering automated recovery events can exacerbate the situation if recovery is unsuccessful.

Geomancy: Automated Data Placement Optimization

Oceane Bel (University of California, Santa Cruz)

Exascale cloud storage and High-Performance Computing Systems (HPC) deliver unprecedented storage capacity and levels of computing power, though the full potential of these systems remain untapped because of inefficient data placement. Changes in data access patterns can cause a system's performance to suffer. To mitigate performance losses, system designers implement strategies to preemptively place popular data on higher performance nodes. However, these strategies fail to address a diverse userbase whose users individually demand the highest performance, and they must be carefully constructed by an expert of the system.

We propose Geomancy, a tool that reorganizes data to increase I/O throughput. In systems where heuristic-based improvements may become resource intensive, Geomancy determines new placement policies by training a deep neural network with past workload and system traces. With real workload traces, Geomancy calculated an example placement policy that demonstrated a 49% increase in average throughput compared to the default data layout.

Numerical Simulation of a Flue Instrument with Finite-Difference Lattice Boltzmann Method using GPGPU

Ryoya Tabata (Kyushu Institute of Technology)

In this work, we discuss the possibility of using GPGPU techniques for Aeroacoustic Simulation (especially for flue instruments) with the finite-difference lattice Boltzmann method (FDLBM). Compressible flow simulation has been used in direct aeroacoustic simulation; however, the computational cost is huge due to the requirement of high computational mesh resolution, with small

time steps. The lattice Boltzmann method (LBM) has been used as an efficient method for fluid simulation using GPGPU. However, LBM is not accurate enough when applied to some aeroacoustic problems. On the other hand, FDLBM is able to use high-order finite-difference schemes and it has a high arithmetic intensity compared to LBM. We present a performance evaluation of the LBM and FDLBM with several finite-difference schemes on GPU with the roofline model.

Recursive Algebraic Coloring Engine

Christie Louis Alappat (University of Erlangen-Nuremberg)

Many iterative numerical methods for sparse systems and building blocks of sparse linear algebra are difficult to parallelize due to data dependencies. These may be loop-carried dependencies as they occur in solvers like Gauss-Seidel or write conflicts as in symmetric sparse matrix vector. Most of the existing parallelization strategies suffer from low performance on modern hardware, are matrix specific, or require tailored storage formats.

In this poster, we introduce a novel recursive level based algorithm called Recursive Algebraic Coloring (RAC), which achieves high hardware efficiency on modern multi-core architectures and works with simple data formats like compressed row storage. Method is implemented in a library called Recursive Algebraic Coloring Engine (RACE). Thorough performance analysis shows that RACE outperforms traditional multicoloring methods and Intel-MKL implementations with a factor of 2–2.5×. We are on par with Algebraic Block Multicoloring for small matrices, while for large matrices we gain a factor of 1.5–2×.

Accelerating Microscope Data Analysis Using Parallel Computing

John Ravi (North Carolina State University)

Single-Molecule Localization Microscopy (SMLM) techniques deal with the diffraction limit of fluorescent microscopy by localizing single molecules with high precision by stochastically switching molecules on and off. Thousands of camera frames containing subsets of blinking molecules are recorded to obtain a single super-resolution image. Each blinking molecule in each frame is subjected to localization protocols that fit the shape of the blink, assess the quality of the blink and then estimate their center. The algorithm implemented originally in MATLAB and compiled CUDA C, to compute a 'Super Resolution' image took around 6 minutes to process 256x256 pixel images of a moderately dense dataset. I ported the algorithm to C++ and parallelized it using OpenMP to compute multiple frames in parallel.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures

Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and

11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

PotC: Many-Body Potential Implementations à La Carte

Markus Höhnerbach (RWTH Aachen University)

Molecular dynamics is a valuable investigation tool for simulations in computational chemistry and materials science. In these simulations, atoms move according to so-called potentials, functions that typically describe the distance-dependent interactions between pairs of atoms. For some applications more complex "many-body potentials" are required. From a HPC perspective, such potentials pose challenges: small neighborhoods hinder vectorization; redundant force expressions are tedious and error prone to derive; the implementations are large, runtime-critical, and can not be broken into simpler "kernels". Consequently, only selected many-body potentials receive optimizations---with major effort.

PotC is a DSL for MD potentials and corresponding compiler to generate high-performance implementations. The compiler centralizes optimization knowledge for many-body potentials, and unburdens users from manual optimization and force derivation. The performance optimization of the generated code is still work-in-progress. However, independent of performance, the tool can also be applied when prototyping, testing, or possibly symbolically manipulating potentials.

OoO Instruction Benchmarking Framework on the Back of Dragons

Julian Hammer (University of Erlangen-Nuremberg, RRZE)

In order to construct an accurate instruction execution model for modern out-of-order micro architectures, an accurate description of instruction latency, throughput and concurrency is indispensable. Already existing resources and vendor provided information is neither complete nor detailed enough and sometimes incorrect. We therefore proclaim to deduct this information through runtime instruction benchmarking and present a framework to support such investigations based on LLVM's just-in-time and cross-platform compilation capabilities.

pyasmjit abstracts instructions, operands, and dependency chains, to easily construct the necessary benchmarks. The synthesized code is interactively compiled and executed using the llvmlite library, which in turn is based on the stable LLVM C-API. pyasmjit offers a command line as well as a programming interface. Unlike other approaches, we do not rely on model specific performance counters and focus on interoperability and automation to support quick modeling of future microarchitectures.

Studying the Impact of Power Capping on MapReduce-Based, Data-Intensive Mini-Applications on Intel KNL and KNM Architectures

Joshua H. Davis (University of Delaware)

In this poster, we quantitatively measure the impacts of data movement on performance in MapReduce-based applications when executed on HPC systems. We leverage the PAPI ‘powercap’ component to identify ideal conditions for execution of our applications in terms of (1) dataset

characteristics (i.e., unique words); (2) HPC system (i.e., KNL and KNM); and (3) implementation of the MapReduce programming model (i.e., with or without combiner optimizations). Results confirm the high energy and runtime costs of data movement, and the benefits of the combiner optimization on these costs.

Accelerating Big Data Processing in the Cloud with Scalable Communication and I/O Schemes
Shashank Gugnani (Ohio State University)

With the advent of cloud computing, the field of Big Data has seen rapid growth. Most cloud providers provide hardware resources such as NVMe SSDs, large memory nodes, and SR-IOV. This opens up the possibility of large-scale high-performance data analytics and provides opportunities to use these resources to develop new designs. Cloud computing provides flexibility, security, and reliability, which are important requirements for Big Data frameworks. However, several important requirements are missing, such as performance, scalability, consistency, and quality of service (QoS). The focus of this work revolves around developing communication and I/O designs and concepts which can provide these requirements to Big Data frameworks. Specifically, we explore new ways to provide QoS and consistency in cloud storage systems, and provide scalable and high-performance communication frameworks.

Accelerating 2D FFT: Exploit GPU Tensor Cores through Mixed-Precision
Xiaohe Cheng (Hong Kong University of Science and Technology), Anumeena Sorna (National Institute of Technology, Tiruchirappalli)

The two-dimensional Fourier Transform is a widely-used computational kernel in many HPC applications. The popular NVIDIA cuFFT library provides a simple interface to compute 2D FFT on GPUs, but it's yet to utilize the recent hardware advancement in half-precision floating-point arithmetic. In this poster, we propose a mixed-precision method to accelerate 2D FFT by exploiting the FP16 matrix-multiply-and-accumulate units on the newest GPU architecture, known as tensor cores. We achieve a balance between speed and accuracy by dynamically splitting the single-precision input data into two half-precision operands and performing FFT separately. We present a CUDA-based implementation that achieves 3-digit more accuracy than half-precision cuFFT. We also demonstrate the stability and scalability of our approach and conclude that it attains high accuracy with tolerable splitting overhead.

Monitoring Parsl Workflows
Connor Pigg (University of Illinois)

As a Python library that enables workflows, Parsl gives users the ability to define complex workflows in Python and run them in parallel on any computer system. This poster describe the process of adding monitoring to Parsl. Simple and comprehensive monitoring of a workflow's state and resource usage lets users audit, debug, and confirm workflow execution. The poster discusses how Parsl monitors workflow components, what data it captures (task status and resource usage), and the tools it used to do so (Elasticsearch) and to display the information (Kibana). A Kibana dashboard visualizes the collected logs in a real time, with an interactive user interface. This enhanced Parsl allows users the option to monitor the status and resource usage of their workflows via an Elasticsearch database and

Kibana dashboard.

Identifying Network Data Transfer Bottlenecks in HPC Systems

Karen Tu (Lawrence Berkeley National Laboratory; University of California, Berkeley)

Improving network data transfer performance is a major factor for improving high performance computing systems. Most studies analyze data transfer and file system IO performance separately, but understanding the relationship between the two is essential for optimizing scheduling and resource management. Intuitively, if data is being transferred to a busy file system the transfer rate would be slower than a file system at regular activity levels.

This study analyzes patterns between file system activity and network throughput for several use cases of file writing and data transfers using a parallel file system. The parameters changed among the use cases were file striping for the file system, and buffer size and parallelism for data transfer. The main bottleneck for network data transfer rate was the number of OSTs the data was striped across. For a large number of OSTs (16 or greater), writing to the file system was the bottleneck.

Dendro-GR: Massively Parallel Simulations of Binary Black Hole Intermediate-Mass-Ratio Inspirals

Milinda Fernando (University of Utah)

We present a portable and highly-scalable algorithm and framework that targets problems in the astrophysics and numerical relativity communities. This framework combines together a parallel octree-refined adaptive mesh with wavelet adaptive multiresolution and a physics module to solve the Einstein equations of general relativity in the BSSN~formulation. The goal of this work is to perform advanced, massively parallel numerical simulations of Intermediate Mass Ratio Inspirals (IMRIs) of binary black holes with mass ratios on the order of 100:1. These studies will be used to generate waveforms for use in LIGO data analysis and to calibrate semi-analytical approximate methods. This advanced framework is designed to easily accommodate many existing algorithms in astrophysics for plasma dynamics and radiation hydrodynamics. We have designed novel algorithms to enable efficient simulations for such experiments and demonstrate excellent weak scalability up to 131K cores on ORNL's Titan for binary mergers for mass ratios up to 100.

Hardware Transactional Persistent Memory

Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares

well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories

Luann C. Jung (Massachusetts Institute of Technology, University of Chicago), Brendan T. Whitaker (Ohio State University, University of Chicago)

As scientific data repositories and filesystems grow in size and complexity, they become increasingly disorganized. The coupling of massive quantities of data with poor organization makes it challenging for scientists to locate and utilize relevant data, thus slowing the process of analyzing data of interest. To address these issues, we explore an automated clustering approach for quantifying the organization of data repositories. Our parallel pipeline processes heterogeneous filetypes (e.g., text and tabular data), automatically clusters files based on content and metadata similarities, and computes a novel "cleanliness" score from the resulting clustering. We demonstrate the generation and accuracy of our cleanliness measure using both synthetic and real datasets, and conclude that it is more consistent than other potential cleanliness measures.

Supercomputing for the Multi-Driver Routing

Zeyang Ye (Stony Brook University)

Supercomputing is essential for routing traffic by providing drivers the optimal routes with minimal traveling distances or time. The unique challenges that require supercomputers to overcome are of multiple folds: numerous drivers, massive simultaneous requests, multiple locations, and needs of instant gratifications, etc. We developed two parallel methods, PSAD and PSAD-M, by using domain decomposition and state-mixing techniques. On the same computing platform with 96 cores, for the same problem, our PSAD methods outperform all published benchmarks by over a hundred times, while improving the solution quality. For the same routing problem on 384 cores, our PSAD-M reduced the elapsed time from the unbearable ten minutes to the reasonable 5 seconds, achieving a record-breaking speedup of 170. By providing instant routing solutions that enable online recommendations, our methods break the bottleneck of the widely adopted offline approaches.

NautDB: Toward a Hybrid Runtime for Processing Compiled Queries

Samuel Grayson (University of Texas, Dallas)

General purpose operating and database system suffer under the load of their generality which makes achieving optimal performance extremely hard, especially on modern hardware. The goal of this research is to integrate, for the first time, specialization techniques from the OS community (hybrid runtimes) and DB community (compiled queries) for high-performance query processing on modern hardware. We envision a system called NautDB, a hybrid dataflow runtime for executing compiled queries. As a first step toward our goal, we evaluate the performance of compiled queries on Linux and run as a Nautilus hybrid runtime using a simple prototype. Our results demonstrate that combining these specialization techniques has transformative potential for building the next generation (distributed) high-performance query processing systems and big data platforms.

Mitigating Performance and Progress Variability in Iterative Asynchronous Algorithms
Justs Zarins (University of Edinburgh)

Large HPC machines are susceptible to irregular performance. Factors like chip manufacturing differences, heat management, and network congestion combine to result in varying execution time for the same code and input sets. Asynchronous algorithms offer a partial solution. In these algorithms, fast workers are not forced to synchronize with slow ones. Instead they continue computing updates, and moving toward the solution, using the latest data available to them, which may have become stale (i.e. a number of iterations out of date compared to the most recent data). While this allows for high computational efficiency, the convergence rate of asynchronous algorithms tends to be lower.

To address this problem, we are using the unique properties of asynchronous algorithms to develop load balancing strategies for iterative asynchronous algorithms in both shared and distributed memory. Our poster shows how our solution attenuates noise, resulting in significant reduction progress imbalance and time-to-solution variability.

Eulerian Algorithms for the Discretization of Plasma Kinetic Equations
James L. Juno (University of Maryland)

While fluid models are common tools in the study of plasmas, many of these systems, whether in astrophysics or the lab, are only weakly collisional and far from equilibrium, making them more accurately described by kinetic equations. Kinetic equations can be computationally demanding due to the need to solve for the distribution function of the particles in a higher dimensional phase space, with position and velocity coordinates. Despite this challenge, the motivation for solving the plasma kinetic equation is large as there remains a vast array of questions concerning collisionless dynamics in real plasma systems. Here we present algorithms in an Eulerian framework for the discretization of the plasma kinetic equation, using a high-order discontinuous Galerkin finite element method due to its arithmetic intensity and parallelizability. Scaling and performance of the algorithm are discussed, and benchmarks of the algorithm are presented as well.

Room: C2/3/4 Ballroom
5:15 pm - 7:00 pm

Poster Reception

Session Description: The Posters Reception is an opportunity for attendees to interact with poster presenters and includes research and ACM Student Research Competition posters, Doctoral Showcase posters as well as the Scientific Visualization & Data Analytics Showcase.

Wednesday, November 14th

Room: C2/3/4 Ballroom

8:30 am - 5:00 pm

Research Posters

Session Description: SC18 Research Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

SC18 Research Posters

SC18 Research Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Exploring Application Performance on Fat-Tree Networks in the Presence of Congestion

Philip A. Taffet (Rice University, Lawrence Livermore National Laboratory), Sanil Rao (University of Virginia, Lawrence Livermore National Laboratory), Ian Karlin (Lawrence Livermore National Laboratory)

Network congestion, which occurs when multiple applications simultaneously use shared links in cluster network, can cause poor communication performance, decreasing the performance and scalability of parallel applications. Many studies are performed while clusters also run other production workloads, which makes it harder for them to isolate causes and their effects. To look at congestion in a more controlled setting we used dedicated access time on an HPC cluster and measured the performance of three HPC applications with different communication patterns run with varying amounts and types of background traffic. This enables us to assess the relative sensitivity of the applications to congestion caused by different traffic patterns. Our tests show that the applications were not significantly impacted by even the most aggressive neighboring patterns, with all the performance degradation being 7% or less, pointing to the resiliency of the fat-tree topology.

Best Poster Finalist: no

GPU-Accelerated Interpolation for 3D Image Registration

Naveen Himthani (University of Texas, Institute for Computational Engineering and Sciences), Andreas Mang (University of Houston), Amir Gholami (University of California, Berkeley), George Biros (University of Texas, Institute for Computational Engineering and Sciences)

Image registration is a key technology in image computing with numerous applications in medical imaging. Our overarching goal is the design of a consistent and unbiased computational framework for the integration of medical imaging data with simulation and optimization to support clinical decision making for glioma brain tumor patients. A major issue in 3D image registration is the time to solution, which poses the demand for effective numerical schemes and the utilization of high performance computing resources.

In this poster, we extend present a GPU-accelerated implementation of the Lagrange interpolation kernel using hardware texture filtering feature of modern hardware. Ongoing work involves implementing a unified single-GPU code for 3D image registration along with other computational kernels such as FFT. I will present my work by briefly explaining image registration followed by the explanation of the interpolation algorithm and its features and then demonstrate the results obtained.

Best Poster Finalist: no

Energy Efficiency of Reconfigurable Caches on FPGAs

Tianqi Wang (Boston University), Ang Li (Pacific Northwest National Laboratory), Tong Geng (Boston University), Martin Herbordt (Boston University)

The performance of a given cache architecture depends largely on the applications that run on it. Even though each application has its best-suited cache configuration, vendors of fixed HPC systems must provide compromise designs. Reconfigurable caches can adjust cache configuration dynamically to get best-suited cache parameters in runtime and notably reduce energy consumption. For example, when it is possible to deploy a low capacity low associativity design without increasing the miss rate substantially. For modern multi-core processors, each core's memory access behavior can be influenced by other cores. So it is more complicated to design reconfigurable caches for them.

In this paper, a design for a reconfigurable cache on FPGAs is presented that can run in modes with different capacities, associativity. We demonstrate that better performance and energy efficiency can be achieved by tuning these cache parameters at runtime.

Best Poster Finalist: no

RGB (Redfish Green500 Benchmarker): A Green500 Benchmarking Tool Using Redfish

Elham Hojati (Texas Tech University), Yong Chen (Texas Tech University), Alan Sill (Texas Tech University), Jon Hass (Dell Inc)

Performance and energy are important factors for supercomputers and data-centers with a trade-off between them. Energy efficiency metric considers both of these properties. The Green500 is a branch of Top500 project which provides a list of supercomputers based on energy efficiency. It has a manual methodology for this calculation.

Redfish is a new generation of management technologies for the hardware layer of data-centers. ur roject focuses on designing and developing an automated Green500 benchmark tool using Redfish, alled Redfish Green500 Benchmarker, or RGB in short. It offers the integration of Redfish and Green500, and automates Green500 benchmarking process with leveraging the internal capability of Redfish enabled equipment. It also enhances the Redfish standard to make sure it addresses the requirements of Green500 calculations. This research will also conduct validation and evaluation of RGB on real-world clusters for small-scale to medium-scale tests, and on the data-center simulator we have developed.

Best Poster Finalist: no

Optimization of Ultrasound Simulations on Multi-GPU Servers

Filip Vaverka (Brno University of Technology, Faculty of Information Technology), Matej Spetko (Brno University of Technology, Faculty of Information Technology), Bradley E. Treeby (University College London, Biomedical Ultrasound Group), Jiri Jaros (Brno University of Technology, Faculty of Information Technology)

Realistic ultrasound simulations have found a broad area of applications in preoperative photoacoustic screening and non-invasive ultrasound treatment planing. However, the domains are typically

thousands of wavelengths in size, leading to large-scale numerical models with billions of unknowns. The current trend in accelerated computing is towards the use of fat nodes with multiple GPUs per node. The multi-GPU version of our k-Wave acoustic toolbox is based on the local Fourier basis domain decomposition where 3D simulation domain is partitioned into rectangular cuboid blocks assigned to particular GPUs. This paper investigates the benefits of using the CUDA-Aware MPI and CUDA peer-to-peer transfers on an 8-GPU server equipped with Nvidia P40 GPUs. The server has a total GPU memory of 192 GB and a single-precision performance of 96 Tflops. These techniques reduces the overall simulation time a factor of 2-3.6.

Best Poster Finalist: no

GPGPU Performance Estimation with Core and Memory Frequency Scaling

Qiang Wang (Hong Kong Baptist University), Xiaowen Chu (Hong Kong Baptist University)

Graphics processing units (GPUs) support dynamic voltage and frequency scaling to balance computational performance and energy consumption. However, simple and accurate performance estimation for a given GPU kernel under different frequency settings is still lacking for real hardware, which is important to decide the best frequency configuration for energy saving. We reveal a fine-grained analytical model to estimate the execution time of GPU kernels with both core and memory frequency scaling. Over a wide scaling range of both core and memory frequencies among 20 GPU kernels, our model achieves accurate results (4.83% error on average) with real hardware. Compared to the cycle-level simulators, our model only needs simple micro-benchmarks to extract a set of hardware parameters and kernel performance counters to produce such high accuracy.

Best Poster Finalist: no

Making Sense of Scientific Simulation Ensembles

Mai Dahshan (Virginia Tech), Nicholas Polys (Virginia Tech)

Scientists run many simulations with varying initial conditions, known as "ensembles", to understand the influence and relationships among multiple parameters or ensemble members. Most of the ensemble visualization and analysis approaches and techniques focus on analyzing the relationships between either the ensemble members or output parameter space while neglecting the effect of input parameters and humans in the analysis loop. Therefore, we developed an approach to the visual analysis of scientific data that merges human expertise and intuition with machine learning and statistics allowing scientists to explore, search, filter, and make sense of their high dimensional ensemble. Our tool, "GLEE" (Graphically-Linked Ensemble Explorer), is an interactive visualization tool that consists of three visual views: Ensemble View, Parameter View, and Statistical View. Each view offers different functionality for exploration and interoperation of the relations and correlations between different runs, a subset of runs, and input and output parameters.

Best Poster Finalist: no

Which Architecture Is Better Suited for Matrix-Free Finite-Element Algorithms: Intel Skylake or Nvidia Volta?

Martin Kronbichler (Technical University Munich), Momme Allalen (Leibniz Supercomputing Centre), Martin Ohlerich (Leibniz Supercomputing Centre), Wolfgang A. Wall (Technical University Munich)

This work presents a performance comparison of highly tuned matrix-free finite element kernels from the finite element library on different contemporary computer architectures, NVIDIA V100 and P100 GPUs, an Intel Knights Landing Xeon Phi, and two multi-core Intel CPUs (Broadwell and Skylake). The algorithms are based on fast integration on hexahedra using sum factorization techniques. For small problem sizes, when all data fits into CPU caches, Skylake is very competitive with Volta. For larger sizes, however, the GPU holds an advantage of approximately a factor of three over Skylake, because all architectures operate in the memory-bandwidth limited regime. A detailed performance analysis contrasts the throughput-oriented character of GPUs versus the more latency-optimized CPUs for the scenario of high-order finite element computations.

Best Poster Finalist: no

SpotSDC: an Information Visualization System to Analyze Silent Data Corruption

Zhimin Li (University of Utah), Harshitha Menon (Lawrence Livermore National Laboratory), Yarden Livnat (University of Utah), Kathryn Mohror (Lawrence Livermore National Laboratory), Valerio Pascucci (University of Utah)

Aggressive technology scaling trends are expected to make the hardware of HPC systems more susceptible to transient faults. Transient faults in hardware may be masked without affecting the program output, cause a program to crash, or lead to silent data corruptions (SDC). While fault injection studies can give an overall resiliency profile for an application, without a good visualization tool it is difficult to summarize and highlight critical information obtained. In this work, we design SpotSDC, a visualization system to analyze a program's resilience characteristics to SDC. SpotSDC provides an overview of the SDC impact on an application by highlighting regions of code that are most susceptible to SDC and will have a high impact on the program's output. SpotSDC also enables users to visualize the propagation of error through an application execution.

Best Poster Finalist: no

High-Accuracy Scalable Solutions to the Dynamic Facility Layout Problem

Apan Qasem (Texas State University), Clara Novoa (Texas State University), Chandra Kolla (Texas State University), Samantha Coyle (Texas State University)

The dynamic facility layout problem (DFLP) is concerned with finding arrangements of facilities within plant locations that minimize the sum of material handling and relocation costs over a planning horizon. DFLP is relevant in manufacturing engineering; accurate solutions can reduce operational costs by as much as 30%. We present a new scalable solution that formulates the task of finding the optimal arrangement as a shortest-path (SP) problem. The new parallel algorithm to find the SP employs a problem-specific heuristic to substantially cut down the search space. Compiler-level optimizations improve the performance across different execution platforms, including an auto-tuning strategy to derive the optimal SMT configuration on a POWER8 system. Results show a factor of 13 speedup over existing methods. For the six-facilities problems the best known solution is reached and for sets with 15 and 30 facilities the solution is within 2.83% and 5.66% of the best solution, respectively.

Best Poster Finalist: no

Large Scale MPI-Parallelization of LBM and DEM Systems: Accelerating Research by Using HPC

Bohumir Jelinek (Mississippi State University), George Mason (Mississippi State University), John Peters (Mississippi State University), Daniel Johnson (Mississippi State University), Marcus Brumfield (Mississippi State University), Alex Carrillo (US Army Engineer Research and Development Center), Clay Goodman (Mississippi State University), Farshid Vahedifard (Mississippi State University)

Casting, solidification, and the behavior of dry, saturated, and partially saturated granular media are examples of interesting and important problems in multiple areas of civil, mechanical, and chemical engineering. For interacting particle-fluid systems, the Discrete Element Method (DEM) and Lattice-Boltzmann Method (LBM) provide valuable high-resolution numerical models. Their main issue is high computational demand, which can be addressed by use of HPC resources. This work demonstrates the use of MPI-parallelized LBM and DEM models to accelerate research in solidification and macroscopic behavior of dry and saturated granular media. Large scale parallel simulations of dendritic growth, the calibration-chamber cone penetration test, and a parametric study of shear thickening in granular suspension were performed. Use of HPC dramatically reduced the computational time for these studies and provided high-resolution representation of physical experiments.

Best Poster Finalist: no

SciGaP: Apache Airavata Hosted Science Gateways

Marlon Pierce (Indiana University), Suresh Marru (Indiana University), Eroma Abeysinghe (Indiana University), Sudhakar Pamidighantam (Indiana University), Marcus Christie (Indiana University), Dimuthu Upeksha (Indiana University)

The goal of the Science Gateways Platform as a service (SciGaP.org) project is to provide core services for building and hosting science gateways. Over the last two years, SciGaP services have been used to build and host over twenty-five science gateways. SciGaP services support these gateways through a single hosted version of the Apache Airavata software system that supports multiple tenants. Apache Airavata services include scientific application execution management on HPC and cloud environments, input and output data staging, and provenance tracking for user-created computational experiments.

The poster presents highlights of some selected SciGaP-hosted gateways. Clients interested in SciGaP services can request a tenant through https://scigap.org/. Clients who need extensive support for building user interfaces and integrating unconventional resources can request support through the Science Gateways Community Institute's Extended Developer Support program. To integrate with XSEDE resources, clients can request support through XSEDE's Extended Collaborative Support Services.

Best Poster Finalist: no

Reproducibility as Side Effect

Shu Wang (University of Chicago), Zhuo Zhen (University of Chicago), Jason Anderson (University of Chicago), Kate Keahey (Argonne National Laboratory, University of Chicago)

The ability to keep records and reproduce experiments is a critical element of the scientific method for any discipline. However, the recording and publishing of research artifacts that allow to reproduce and directly compare against existing research continue to be a challenge. In this paper, we propose an experiment précis framework that helps the experiment repeatability. Guided by the framework, we

implement a prototype tool called ReGen which generates repeatable experiment scripts that can be used or shared along with a detailed experiment description automatically. The evaluation shows that ReGen is effective in reducing the researcher's efforts of creating a repeatable experiment in a real setting.

Best Poster Finalist: no

Using Darshan and CODES to Evaluate Application I/O Performance

Harsh Khetawat (North Carolina State University), Christopher Zimmer (Oak Ridge National Laboratory), Frank Mueller (North Carolina State University), Sudharshan Vazhkudai (Oak Ridge National Laboratory), Scott Atchley (Oak Ridge National Laboratory)

Burst buffers have become increasingly popular in HPC systems, allowing bursty I/O traffic to be serviced faster without slowing down application execution. The ubiquity of burst buffers creates opportunities for studying their ideal placement in the HPC topology. Furthermore, the topology of the network interconnect can also affect the performance of the storage hierarchy for different burst buffer placement schemes. To that end, we create a reproducible framework that allows individual centers to develop their own models and evaluate performance based on their workload characteristics. We use CODES to create models that simulate the network and storage layers of an HPC system and Darshan traces for I/O replay. We augment the Darshan traces with synchronization primitives, and allow multi-dimensional scaling of traces to represent future workload characteristics. Finally, we evaluate the effect of network topology, storage architecture, and application I/O patterns on overall I/O performance.

Best Poster Finalist: no

Multi-Client DeepIO for Large-Scale Deep Learning on HPC Systems

Yue Zhu (Florida State University), Fahim Chowdhury (Florida State University), Huansong Fu (Florida State University), Adam Moody (Lawrence Livermore National Laboratory), Kathryn Mohror (Lawrence Livermore National Laboratory), Kento Sato (Lawrence Livermore National Laboratory), Weikuan Yu (Florida State University)

With the growth of computation power, leadership High-Performance Computing (HPC) systems can train larger datasets for Deep neural networks (DNNs) more efficiently. On HPC systems, a training dataset is on a parallel file system or node-local storage devices. However, not all HPC clusters have node-local storage, and large mini-batch sizes stress the read performance of parallel systems since the large datasets cannot fit in file system caches. Thus, it is a challenge for training DNNs with large datasets on HPC systems.

In prior work, we proposed DeepIO to mitigate the I/O pressure. DeepIO is designed to assist the mini-batch generation of TensorFlow. However, DeepIO does not support multiple training workers on a single compute node. We address this gap with modification on DeepIO framework, and evaluate multi-client DeepIO performance against state-of-the-art in-memory file systems, compare DeepIO and TensorFlow data loading API, and explore the potential of DeepIO in DNN training.

Best Poster Finalist: no

HIVE: A Cross-Platform, Modular Visualization Ecosystem for Heterogeneous Computational

Environments

Jorji Nonaka (Riken Center for Computational Science), Kenji Ono (Kyushu University, RIKEN), Naohisa Sakamoto (Kobe University, RIKEN), Kengo Hayashi (Kobe University, RIKEN), Tomohiro Kawanabe (Riken Center for Computational Science), Fumiyoshi Shoji (Riken Center for Computational Science), Masahiro Fujita (LTE Inc), Kentaro Oku (Kashika Inc), Kazuma Hatta (Imagica Digitalscape)

HPC operational environments usually have supporting computational systems for assisting pre- and post-processing activities such as the visualization and analysis of simulation results. A wide variety of hardware systems can be found at different HPC sites, and in our case, we have a CPU-only (x86) large memory server, a planned OpenStack-based CPU/GPU Cluster, SPARC64 fx CPU based HPC system (K computer), and an ARM based HPC system in the future. Therefore, heterogeneity and scalability are needed to be tackled to efficiently use these heterogeneous computational resources for large-scale data visualization on both post-hoc and in-situ contexts. In this poster we present HIVE (Heterogeneously Integrated Visual-analytics Environment), a cross-platform and modular ecosystem for providing visualization service building blocks in such heterogeneous computational environments. Lightweight Lua scripting language is used to glue necessary visualization pipeline related modules, and this loosely coupled modular approach facilitates long-term development and maintenance.

Best Poster Finalist: no

Improving the I/O Performance and Memory Usage of the Xolotl Cluster Dynamics Simulator

Philip C. Roth (Oak Ridge National Laboratory), Sophie Blondel (University of Tennessee), David E. Bernholdt (Oak Ridge National Laboratory), Brian D. Wirth (University of Tennessee)

Xolotl is a cluster dynamics simulator used to predict gas bubble evolution in solids. It is currently being used to simulate bubble formation in the plasma-facing surface within fusion reactors and the nuclear fuel used in fission reactors. After observing performance problems in coupled-code simulations of fusion reactors, we used Xolotl's built-in performance data collection infrastructure and an external profiling tool to identify inefficiencies when writing Xolotl's two types of checkpoint files. We changed the code to use true parallel writes via the HDF5 data management library, resulting in a code that is approximately 57x faster when writing the program's main checkpoint file at the scale used in the coupled-code simulations, and that exhibits less performance variability due to external system activity. We also identified and addressed a memory usage problem that reduced Xolotl peak memory usage by approximately 88% per compute node.

Best Poster Finalist: no

Performance Evaluation of the Shifted Cholesky QR Algorithm for Ill-Conditioned Matrices

Takeshi Fukaya (Hokkaido University), Ramaseshan Kannan (Arup UK), Yuji Nakatsukasa (National Institute of Informatics, Japan), Yusaku Yamamoto (University of Electro-Communications, Japan), Yuka Yanagisawa (Waseda University)

The Cholesky QR algorithm, which computes the QR factorization of a matrix, is a simple yet efficient algorithm for high-performance computing. However it suffers from numerical instability. In a recent work, this instability has been remedied by repeating Cholesky QR twice (CholeskyQR2). ChokeskyQR2, however, is still prone to numerical breakdown when applied to ill-conditioned matrices. To overcome this limitation, we introduce a shifting technique to Cholesky QR and use it as a preconditioning step before CholeskyQR2. The key idea is that Cholesky QR with shift reduces the

condition number of the input matrix. We call the resulting algorithm shifted CholeskyQR3, which is still simple and only requires double precision arithmetic. In this poster, we present the results of our performance evaluation of shifted CholeskyQR3. We demonstrate that shifted CholeskyQR3 accurately computes the QR factorization of ill-conditioned matrices and that it outperforms other conventional algorithms in execution time.

Best Poster Finalist: no

HPC-as-a-Service for Life Sciences

Vaclav Svaton (Technical University of Ostrava, Czech Republic), Jan Martinovic (Technical University of Ostrava, Czech Republic), Nina Jeliazkova (IDEAconsult Ltd, Bulgaria), Vladimir Chupakhin (Janssen Pharmaceutika NV), Pavel Tomancak (Max Planck Institute of Molecular Cell Biology and Genetics), Petr Vojta (Palacký University Olomouc, Czech Republic)

HPC-as-a-Service is a well-known term in the area of high performance computing. It enables users to access an HPC infrastructure without a need to buy and manage their own infrastructure. Through this service, academia and industry can take advantage of the technology without an upfront investment in the hardware. This approach further lowers the entry barrier for users who are interested in utilizing massive parallel computers but often do not have the necessary level of expertise in the area of parallel computing. To provide this simple and intuitive access to the supercomputing infrastructure, an in-house application framework called HEAppE has been developed. HEAppE's universally designed software architecture enables unified access to different HPC systems through a simple object-oriented API. Thus providing HPC capabilities to the users but without the necessity to manage the running jobs from the command-line interface of the HPC scheduler directly on the cluster.

Best Poster Finalist: no

Hermes: a Multi-Tiered Distributed I/O Buffering System for HDF5

Hariharan Devarajan (Illinois Institute of Technology, HDF Group)

High-Performance Computing (HPC) systems' increasing ability to run data-intensive problems at larger scale and resolution has driven the evolution of modern storage technologies. In addition, extreme amounts of data are collected by large scientific instruments and sensor network is resulting in a push for more capable storage systems. Hierarchical Data Format (HDF) technologies address the problems of how to organize, access, analyze, and preserve data in the face of enormous growth in size and complexity. To mitigate this I/O performance bottleneck, modern storage subsystems are going through extensive changes, by adding additional levels of memory and storage in a hierarchy. In this work, we present Hermes: a new, heterogeneous-aware, multi-tiered, dynamic, and distributed I/O buffering system for HDF5. Hermes enables, manages, and supervises I/O buffering in DMSH. We tested our solution on Cori and show Hermes can accelerate applications by 62x than the state of the buffering platform.

Best Poster Finalist: no

Workflow for Parallel Processing of Sequential Mesh Databases

Ondřej Meca (Technical University of Ostrava, Czech Republic), Lubomír Říha (Technical University of Ostrava, Czech Republic), Tomáš Brzobohatý (Technical University of Ostrava, Czech Republic)

This poster presents a workflow for parallel loading of sequentially stored mesh databases. It can be used

as a connection between tools for the creation of complex engineering models along with parallel solvers to allow broader usage of HPC by the engineering community. Scalability tests show that the proposed algorithm is able to prepare a mesh with hundreds of millions of nodes/elements in several seconds.

Best Poster Finalist: yes

The NAStJA Framework: Non-Collective Scalable Global Communications

Marco Berghoff (Karlsruhe Institute of Technology), Ivan Kondov (Karlsruhe Institute of Technology)

In recent years, simulations in various areas of science and engineering have proven to be very useful. To efficiently deploy simulation codes on current and future high-performance computer systems, high node level performance, scalable communication and the exclusion of unnecessary calculations are an absolute must when developing new solvers.

We have introduced the NAStJA framework, a block-based MPI parallel solver for algorithms, based on regular grid methods, i.e., stencil codes. NAStJA has a dynamic block adaptation, which modifies the calculation domain around the region in which the calculation is currently taking place. The creation and deletion of blocks are autonomously managed within local neighborhoods. Collective all-gather communication is avoided by using a multi-hop network to distribute information across the entire domain that greatly improves the application scaling. In this contribution, we present applications that can benefit from this adaptive method and scaling tests demonstrating the excellent scalability.

Best Poster Finalist: no

Hardware Acceleration of CNNs with Coherent FPGAs

Md Syadus Sefat (Texas State University), Semih Aslan (Texas State University), Apan Qasem (Texas State University)

This paper describes a new flexible approach to implementing energy-efficient CNNs on FPGAs. Our design leverages the Coherent Accelerator Processor Interface (CAPI) which provides a cache-coherent view of system memory to attached accelerators. Convolution layers are formulated as matrix multiplication kernels and then accelerated on CAPI-supported Kintex FPGA board. Our implementation bypasses the need for device driver code and significantly reduces the communication and I/O transfer overhead. To improve the performance of the entire application, not just the convolution layers, we propose a collaborative model of execution in which the control of the data flow within the accelerator is kept independent, freeing-up CPU cores to work on other parts of the application. For further performance enhancements, we propose a technique to exploit data locality in the cache, situated in the CAPI Power Service Layer (PSL). Finally, we develop a resource-conscious implementation for more efficient utilization of resources and improved scalability.

Best Poster Finalist: no

Distributed Fast Boundary Element Methods

Michal Merta (Technical University of Ostrava, Czech Republic), Jan Zapletal (Technical University of Ostrava, Czech Republic), Michal Kravcenko (Technical University of Ostrava, Czech Republic)

We present a parallel implementation of the fast boundary element method (BEM) for the Helmholtz equation. After a brief description of BEM, vectorization of the computationally most demanding kernels,

and shared memory parallelization, we focus on the distributed memory parallelization using a new approach for distribution of the fast BEM system matrices among computational nodes. Moreover, we present a modification of the adaptive cross approximation (ACA) method capable of dealing with BEM matrices containing zero blocks, which usually lead to problems for the original ACA algorithm. Numerical experiments demonstrating the performance of the implementation were carried out using several Intel architectures.

Best Poster Finalist: no

Development of Numerical Coupled Analysis Method by Air Flow Analysis and Snow Accretion Analysis

Kohei Murotani (Railway Technical Research Institute, Japan), Koji Nakade (Railway Technical Research Institute, Japan), Yasushi Kamata (Railway Technical Research Institute, Japan), Daisuke Takahashi (Railway Technical Research Institute, Japan)

In this research, to take countermeasures for the snow accretion damage, we developed a simulator of realizing the snow accretion process in the following steps. Firstly, air flow analysis is performed by "Airflow simulator" developed by RTRI (Railway Technical Research Institute). Secondly, trajectory of flying snow is calculated by the equation of motion for gravity and drag using distribution of velocity of air flow. Thirdly, snow accretion analysis is performed by "Particle simulator" developed by RTRI. The shape modified by snow accretion is reflected onto the boundary conditions of the air flow analysis. In this year, we show the results of the snow accretion analysis for simple cubic shapes in order to aim at system development and validation and discuss the result of the snow accretion analysis for a train bogie model.

Best Poster Finalist: no

Portable Parallel Performance via Multi-Dimensional Homomorphisms

Ari Rasch (University of Münster), Richard Schulze (University of Münster), Sergei Gorlatch (University of Münster)

Achieving portable performance over different parallel architectures and varying problem sizes is hard: e.g., a program optimized for multi-core CPUs on large input sizes can significantly differ from the same program optimized for Graphics Processing Units (GPUs) on small sizes.

We propose an approach to ensuring portability of performance by relying on multi-dimensional homomorphisms (MDHs) -- a class of parallelizable functions that cover important application areas including linear algebra routines (BLAS) and stencil computations. We develop an extended OpenCL implementation schema for MDHs that is generic in the performance-critical parameters of the OpenCL model, and we enable portability of performance by being automatically optimized for different target architectures and input sizes using the auto-tuning approach.

Our results demonstrate competitive and often even significantly better performance than state-of-the-art approaches for BLAS and Stencil as used in the important application area of deep learning.

Best Poster Finalist: no

Performance Evaluation of the NVIDIA Tesla V100: Block Level Pipelining vs. Kernel Level Pipelining

Xuewen Cui (Virginia Tech), Thomas R. W. Scogland (Lawrence Livermore National Laboratory), Bronis R.

de Supinski (Lawrence Livermore National Laboratory), Wu Feng (Virginia Tech)

As accelerators become more common, expressive and performant, interfaces for them become ever more important. Programming models like OpenMP offer simple-to-use but powerful directive-based offload mechanisms. By default, these models naively copy data to or from the device without overlapping computation. Achieving performance can require extensive hand-tuning to apply optimizations such as pipelining. To pipeline a task, users must manually partition the task into multiple chunks then launch multiple sub-kernels. This approach can suffer from high kernel launch overhead. Also, the hyper parameters must be carefully tuned to achieve optimal performance. To ameliorate this issue, we propose a block-level pipeline approach that overlaps data transfers and computation in one kernel handled by different streaming multiprocessors on GPUs. Our results show that, without exhaustive tuning, our approach can provide 95% to 108% stable performance compared to the best tuned results with traditional kernel-level pipelining on NVIDIA V100 GPUs.

Best Poster Finalist: no

Enabling Data Analytics Workflows Using Node-Local Storage

Tu Mai Anh Do (University of Southern California, Information Sciences Institute), Ming Jiang (Lawrence Livermore National Laboratory), Brian Gallagher (Lawrence Livermore National Laboratory), Albert Chu (Lawrence Livermore National Laboratory), Cyrus Harrison (Lawrence Livermore National Laboratory), Karan Vahi (University of Southern California, Information Sciences Institute), Ewa Deelman (University of Southern California, Information Sciences Institute)

The convergence of high-performance computing (HPC) and Big Data is a necessity with the push toward extreme-scale computing. As HPC simulations become more complex, the analytics need to process larger amounts of data, which poses significant challenges for coupling HPC simulations with Big Data analytics. This poster presents a novel node-local approach that uses a workflow management system (WMS) to enable the coupling between the simulations and the analytics in scientific workflows by leveraging node-local non-volatile random-access memory (NVRAM).

Best Poster Finalist: no

OpeNNdd: Open Neural Networks for Drug Discovery: Creating Free and Easy Methods for Designing Medicine

Bryce Kroencke (American River College), Shawn Shacterman (University of California, Berkeley), Nicholas Pavini (American River College), Benjamin Samudio (American River College, Sierra College), Silvia Crivelli (Lawrence Berkeley National Laboratory)

Bringing new medicines to patients can be prohibitively expensive in terms of time, cost, and resources. This leaves many diseases without therapeutic interventions. In addition, new and reemerging diseases are increasing in prevalence across the globe at an alarming rate. The speed and scale of medicine discovery must be increased to effectively meet this challenge. OpeNNdd is a neural network platform bringing together people, machine learning, and supercomputing to solve the challenge of creating medicines. We have developed a novel neural network which quickly and accurately models candidate medicines interacting with a disease target, a metric to delineate its domain of applicability, and a process that communicates neural network results to participants in a readily interpretable way. OpeNNdd leverages the scale of supercomputing, the power and speed of neural networks, and the creativity of people across the globe in an open and collaborative way to protect and improve global health.

Best Poster Finalist: no

FeatherCNN: Fast Inference Computation with TensorGEMM on ARM Architectures

Haidong Lan (Shandong University), Jintao Meng (Tencent Holdings Ltd), Christian Hundt (Johannes Gutenberg University Mainz), Bertil Schmidt (Johannes Gutenberg University Mainz), Minwen Deng (Tencent Holdings Ltd), Weiguo Liu (Shandong University), Yanjie Wei (Shenzhen Institutes of Advanced Technology), Shengzhong Feng (Shenzhen Institutes of Advanced Technology)

This poster presents a fast inference computation library for ARM architecture named as CNNForward. CNNForward is trying to improve the efficiency of inference computation for convolutional neural networks on ARM-based multi-core and many-core architectures using both mathematical formula reconstruction/simplification and in-depth NEON instruction optimization. Experimental results reveal that, forward computation for VGG-16 on a server with 64 ARM A72 cores, CNNForward can scale up to 32 cores with an parallel efficiency of 33%, and achieve 35.4x, 8.7x and 10.6x speedup over Caffe+OpenBlas, Caffe2+Eigen and Caffe2+NNPACK, respectively.

Best Poster Finalist: no

Boosting the Scalability of Car-Parrinello Molecular Dynamics Simulations for Multi- and Manycore Architectures

Tobias Klöffel (University of Erlangen-Nuremberg), Bernd Meyer (University of Erlangen-Nuremberg), Gerald Mathias (Leibniz Supercomputing Centre)

We present our recent optimizations of the ultra-soft pseudo-potential (USPP) code path of the ab inito molecular dynamics program CPMD (www.cpmd.org). Following the internal instrumentation of CPMD, all relevant USPP routines have been revised to fully support hybrid MPI+OpenMP parallelization. For two time-critical routines, namely the multiple distributed 3D FFTs of the electronic states and a key distributed matrix-matrix multiplication, we have implemented hybrid parallel algorithms with overlapping computation and communication. The achievements in performance and scalability are demonstrated on a small reference system of 128 water molecules and further systems of increasing size. Performance evaluation shows gains of up to one order of magnitude and around 50% peak performance for simulation systems readily used in production.

Best Poster Finalist: no

Characterizing Declustered Software RAID for Enhancing Storage Reliability and Performance

Zhi Qiao (University of North Texas), Song Fu (University of North Texas), Hsing-bung Chen (Los Alamos National Laboratory), Bradley Settlemyer (Los Alamos National Laboratory)

Redundant array of independent disks (RAID) has been widely used to address the reliability issue in storage systems. As the scale of modern storage systems continues growing, disk failure becomes the norm. With ever-increasing disk capacity, RAID recovery based on disk rebuild becomes more costly, which causes significant performance degradation and even unavailability of storage systems. Declustered parity and data placement in RAID aims to enhance the recovery performance by shuffling data among all disks in a RAID group. All disks in the RAID group participate in data reconstruction, which leads to reduction of the RAID rebuild time. In this work, we extensively evaluate declustered RAID in terms of the performance of application I/O and recovery time. Our experimental results in ZFS show that the speedup of declustered RAID over traditional RAID is sub-linear to the number of disks in the storage pool.

Best Poster Finalist: no

Parallel Implementation of Machine Learning-Based Many-Body Potentials on CPU and GPU
Yaoguang Zhai (University of California, San Diego), Nathaniel Danandeh (University of California, San Diego), Zhenye Tan (University of California, San Diego; Tongji University), Sicun Gao (University of California, San Diego), Francesco Paesani (University of California, San Diego), Andreas W. Goetz (San Diego Supercomputer Center)

Machine learning models can be used to develop highly accurate and efficient many-body potentials for molecular simulations based on the many-body expansion of the total energy. A prominent example is the MB-pol water model that employs permutationally invariant polynomials (PIPs) to represent the 2-body and 3-body short-range energy terms.

We have recently shown that the PIPs can be replaced by Behler-Parinello neural networks (BP-NN). We present OpenMP parallel implementations of both PIP and BP-NN models as well as a CUDA implementation of the BP-NN model for GPUs. The OpenMP implementations achieve linear speedup with respect to the optimized single threaded code. The BP-NN GPU implementation outperforms the CPU implementation by a factor of almost 8. This opens the door for routine molecular dynamics simulations with highly accurate many-body potentials on a diverse set of hardware.

Best Poster Finalist: no

Implementing Efficient Data Compression and Encryption in a Persistent Key-Value Store for HPC
Jungwon Kim (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory)

Recently, persistent data structures, like key-value stores (KVSs), which are stored in an HPC system's nonvolatile memory, provide an attractive solution for a number of emerging challenges like limited I/O performance. This paper investigates how to efficiently integrate data compression and encryption into persistent KVSs for HPC with the ultimate goal of hiding their costs and complexity in terms of performance and ease of use. We implement the proposed techniques on top of a distributed embedded KVS to evaluate the benefits and costs of incorporating these capabilities along different points in the dataflow path, illustrating differences in effective bandwidth, latency, and additional computational expense.

Best Poster Finalist: yes

A Parallel-Efficient GPU Package for Multiphase Flow in Realistic Nano-Pore Networks
Yidong Xia (Idaho National Laboratory), Ansel Blumers (Brown University, Idaho National Laboratory), Zhen Li (Brown University), Lixiang Luo (IBM), Jan Goral (University of Utah), Matthew Andrew (Carl Zeiss X-ray Microscopy Inc), Joshua Kane (Idaho National Laboratory), Yu-Hang Tang (Lawrence Berkeley National Laboratory)

Simulations of fluid flow in oil/gas shale rocks are challenging in part due to the heterogeneous pore sizes ranging from a few nanometers to a few micrometers. Additionally, the complex fluid-solid interaction occurring physically and chemically must be captured with high resolution. To address these challenges while minimizing computational cost, we present a GPU code that has implemented a many-body dissipative particle dynamics (mDPD) model for multiphase flow in shale. Realistic nano- to micro-pore channels in shale are constructed from 3D high-resolution stack images. In our benchmark tests, the code

delivers nearly perfect weak and strong scalings on up to 512 K20X GPUs on Oak Ridge National Laboratory (ORNL) Titan supercomputer. Moreover, single-GPU bench-marks on the DGX-1 (V100/no NVLink), ORNL's SummitDev (P100/NVLink 1.0) and Summit (V100/NVLink 2.0) suggest that the latest Host-to-Device NVLink can significantly boost overall performance, in addition to the Device-to-Device NVLink.

Best Poster Finalist: no

Processing-in-Storage Architecture for Machine Learning and Bioinformatics

Roman Kaplan (Israel Institute of Technology), Leonid Yavits (Israel Institute of Technology), Ran Ginosar (Israel Institute of Technology)

User-generated and bioinformatics database volumes has been increasing exponentially for more than a decade. With the slowdown and approaching end of Moore's law, traditional technologies cannot satisfy the increasing demands for processing power. This work presents PRINS, a highly-parallel in-storage processing architecture. PRINS combines non-volatile memory with processing capabilities on every bitcell. An emerging technology, memristors, form the basis for the design.

Implementations of three data-intensive and massively parallel algorithms are demonstrated: (1) Smith-Waterman DNA local sequence alignment (bioinformatics), (3) K-means clustering (machine learning) and (3) data deduplication. Performance and energy efficiency of PRINS compared to other published solutions is presented for each algorithm. PRINS is shown to achieve orders-of-magnitude improvement in performance and power efficiency over existing solutions, from large-scale bioinformatics and machine-learning to single-GPU or FPGA implementations.

Best Poster Finalist: no

Kernel-Based and Total Performance Analysis of CGYRO on 4 Leadership Systems

Igor Sfiligoi (General Atomics), Jeff Candy (General Atomics), Emily Belli (General Atomics)

We present the results of an exhaustive performance analysis of the CGYRO code on 4 leadership systems spanning 5 different configurations (2 KNL-based, 1 Skylake-based, and 2 hybrid CPU-GPU architectures). CGYRO is an Eulerian gyrokinetic solver designed and optimized for collisional, electromagnetic, multiscale fusion plasma simulation. It is based on the well-known GYRO code, but redesigned from the ground up to operate efficiently on multicore and GPU-accelerated systems. The gyrokinetic equations specify a 5-dimensional distribution function for each species, with species coupled through both the Maxwell equations and collision operator. For the cross-machine performance analysis, we report and compare timings for 4 computational and 4 communication kernels. This kernel-based breakdown illustrates the strengths and weaknesses of the floating-point and communication architectures of the respective systems. An overview of the physical equations solved, the scalable numerical methods used, and data communication patterns required by each kernel are also given.

Best Poster Finalist: no

Redesigning The Absorbing Boundary Algorithm for Asynchronous High Performance Acoustic Wave Propagation

Rached Abdelkhalak (King Abdullah University of Science and Technology), Kadir Akbudak (King Abdullah University of Science and Technology), Vincent Etienne (Saudi Aramco), Thierry Tonellot (Saudi Aramco)

Exploiting high concurrency, relaxing the synchrony of existing algorithms, and increasing data reuse have immense effect in performance. We integrate the Multicore-optimized Wavefront Diamond (MWD) tiling approach by Malas et al. [SIAM SISC, 2015, ACM Trans. Parallel Comput. 2017], which takes into account the three aforementioned ingredients, into the industrial project codenamed ExaWave framework beside the traditional spatial blocking (SB) technique for stencil computations. However, the fine-grained asynchronous handling of the Convolution Perfectly Matched Layer (CPML) for absorbing boundary conditions turns out to be a challenging open research problem, due to severe inherent data dependencies constraints, which impedes MWD performance impact. We propose techniques of loop fusion to reduce memory traffic and sliding windows to cut down the engendered extra flops, in order to consolidate CPML integration with the overall asynchronous MWD technique. The experimental results on Intel's latest processors show the effectiveness of the proposed techniques.

Best Poster Finalist: no

Capsule Networks for Protein Structure Classification

Dan A. Rosa de Jesus (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Julian Cuevas Paniagua (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Wilson Rivera (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Silvia Crivelli (Lawrence Berkeley National Laboratory)

Capsule Networks have great potential to tackle problems in structural biology because of their attention to hierarchical relationships. This work describes the implementation and application of a capsule network architecture to the classification of RAS protein family structures on GPU-based computational resources. Our results show that the proposed capsule network trained on 2D and 3D structural encodings can successfully classify HRAS and KRAS structures. The capsule network can also classify a protein-based dataset derived from a PSI-BLAST search on sequences of KRAS and HRAS mutations. Experimental results show an accuracy improvement compared to traditional convolutional networks.

Best Poster Finalist: no

Cross-Layer Group Regularization for Deep Neural Network Pruning

Shuang Gao (Nvidia Corporation), Xin Liu (Nvidia Corporation)

Improving weights sparsity is a common strategy for deep neural network pruning. Most existing methods use regularizations that only consider structural sparsity within an individual layer. In this paper, we propose a cross-layer group regularization taking into account the statistics from multiple layers. For residual networks, we use this approach to align kernel sparsity across layers that are tied to each other through element-wise operations: the ith kernel of these layers are put into one regularization group, they either stay or be removed simultaneously during pruning. In this way, the computational and parameter storage cost could be significantly reduced. Experimental results show that this method does not only improve weights sparsity but also align kernel weights sparsity across related layers. Our method is able to prune ResNet up to 90.4% of parameters and improve runtime by 1.5x speedup, without loss of accuracy.

Best Poster Finalist: yes

Machine Learning for Adaptive Discretization in Massive Multiscale Biomedical Modeling

Changnian Han (Stony Brook University), Prachi Gupta (Stony Brook University), Peng Zhang (Stony Brook University), Danny Bluestein (Stony Brook University), Yuefan Deng (Stony Brook University)

For multiscale problems, traditional time stepping algorithms use a single smallest time stepsize in order to capture the finest details; using this scale leads to a significant waste of computing resources for simulating coarse-grained portion of the problem. To improve computing efficiency for multiscale modeling, we propose a novel state-driven adaptive time stepping (ATS) algorithm to automatically adapt the time stepsizes to the underlying biophysical phenomena at multiple scales. In this, we use a machine-learning based solution framework to classify and label these states for regulating the time stepsizes. We demonstrate the values of our ATS algorithm by assessing the accuracy and efficiency of a multiscale two-platelet aggregation simulation. By comparing with traditional algorithm for this simulation, our ATS algorithm significantly improves the efficiency while maintaining accuracy. Our novel ATS algorithm presents a more efficient framework for solving massive multiscale biomedical problems.

Best Poster Finalist: no

Multi-GPU Accelerated Non-Hydrostatic Numerical Ocean Model with GPUDirect RDMA Transfers
Takateru Yamagishi (Research Organization for Information Science and Technology, Japan), Yoshimasa Matsumura (University of Tokyo), Hiroyasu Hasumi (University of Tokyo)

We have implemented our “kinaco” numerical ocean model on Tokyo University’s Reedbush supercomputer, which utilizes the latest Nvidia Pascal P100 GPUs with GPUDirect technology. We have also optimized the model’s Poisson/Helmholtz solver by adjusting the global memory alignment and thread block configuration, introducing shuffle functions to accelerate the creation of coarse grids and merging small kernels in the multigrid preconditioner. We also utilize GPUDirect RDMA transfers to improve MPI communication efficiency. By exploiting the GPUs’ capabilities, the GPU implementation is now twice as fast as the CPU version, and it shows good weak scalability to multiple GPUs. Most of the GPU kernels are accelerated, and the velocity diagnosis functions in particular are now approximately seven times faster. The performance of inter-node data transfers using a CUDA-aware MPI library with GPUDirect RDMA transfers is comparable to that on CPUs.

Best Poster Finalist: no

A Locality and Memory Congestion-Aware Thread Mapping Method for Modern NUMA Systems
Mulya Agung (Tohoku University), Muhammad Alfian Amrizal (Tohoku University), Ryusuke Egawa (Tohoku University), Hiroyuki Takizawa (Tohoku University)

On modern NUMA systems, the memory congestion problem could degrade performance more than the memory access locality problem because a large number of processor cores in the systems can cause heavy congestion on memory controllers. In this work, we propose a thread mapping method that considers the spatio-temporal communication behavior of multi-threaded applications to improve the locality and to reduce the memory congestion on modern NUMA systems. We evaluate the proposed method using NPB applications on a NUMA system. Experiments show that our proposed method can achieve up to 20% performance improvement compared with locality-based and balance-based methods.

Best Poster Finalist: no

Tuning CFD Applications for Intel Xeon Phi with TAU Commander and ParaTools ThreadSpotter
Izaak B. Beekman (ParaTools Inc), Nicholas Chaimov (ParaTools Inc), Sameer Shende (ParaTools Inc, University of Oregon), Allen D. Malony (ParaTools Inc, University of Oregon), Nicholas Bisek (US Air Force

Research Laboratory), Ryan Gosse (US Air Force Research Laboratory), Andrew Wissink (Create AV, US Army)

Tuning and understanding the performance characteristics of computational fluid dynamics (CFD) codes on many-core, NUMA architectures is challenging. One must determine how programming choices impact algorithm performance and how best to utilize the available memory caches, high-bandwidth memory, and inter---and intra---node communication. Once collected, performance data must be translated into actionable code improvements. In addition, performance engineering experiments must be organized and tracked to quantify the benefit of any attempted tuning.

In the poster we present, examine and tune two CFD applications running on the Intel® Xeon Phi™ partition of a Cray® XC 40/50 using TAU Commander and ParaTools ThreadSpotter. TAU Commander implements a streamlined, managed performance engineering workflow and highlights source regions limiting scalability through profiling and aggregate summary statistics. ParaTools ThreadSpotter analyzes an application as it is running and ranks individual performance problems. It provides a report annotating source code lines with actionable recommendations and quantifying performance metrics.

Best Poster Finalist: no

Massively Parallel Stress Chain Characterization for Billion Particle DEM Simulation of Accretionary Prism Formation

Mikito Furuichi (Japan Agency for Marine-Earth Science and Technology), Daisuke Nishiura (Japan Agency for Marine-Earth Science and Technology), Takane Hori (Japan Agency for Marine-Earth Science and Technology)

Herein, a novel algorithm for characterizing stress chains using a large parallel computer system is presented. Stress chains are important for analyzing the results of large-scale discrete element method (DEM) simulations. However, the general algorithm is difficult to parallelize especially when selecting networks longer than several particles. Therefore, we propose a new parallel algorithm to count the number of particles that are tightly connected, based on iterative operations with nearest-neighbor computations and communications. The new algorithm is examined via a real-scale numerical sandbox experiment using 2.4 billion particles. We successfully compute the stress chains with a reasonable computational cost comparable to the single-step DEM computation time. The visualization of the stress chains from the large-scale DEM simulation result reveals the existence of arcuate stress structures that may control accretionary prism formation, which is an important scientific discovery.

Best Poster Finalist: no

Toward Smoothing Data Movement Between RAM and Storage

Tariq Alturkestani (King Abdullah University of Science and Technology), Thierry Tonellot (Saudi Aramco), Vincent Etienne (Saudi Aramco), Hatem Ltaief (King Abdullah University of Science and Technology)

We propose to design and implement a software framework, which provides a Multilayer Buffer System (MBS) to cache in/out datasets into CPU main memory from/to slower storage media, such as parallel file systems (e.g., Lustre), solid-state drive (e.g., Burst Buffer) or non-volatile RAM. Although MBS scope may be broad in terms of scientific applications, we focus on the RTM application as a proxy for I/O intensive workloads, since reading and writing are ubiquitous operations during the dumping phase of the source wavefield (forward propagation) as well as its retrieval phase (backward propagation) for the image

condition calculations.

Best Poster Finalist: no

Interactive HPC Deep Learning with Jupyter Notebooks

Wahid Bhimji (Lawrence Berkeley National Laboratory), Steven Farrell (Lawrence Berkeley National Laboratory), Oliver Evans (Lawrence Berkeley National Laboratory), Matthew Henderson (Lawrence Berkeley National Laboratory), Shreyas Cholia (Lawrence Berkeley National Laboratory), Aaron Vose (Cray Inc), Mr Prabhat (Lawrence Berkeley National Laboratory), Rollin Thomas (Lawrence Berkeley National Laboratory), Richard Shane Canon (Lawrence Berkeley National Laboratory)

Deep learning researchers are increasingly using Jupyter notebooks to implement interactive, reproducible workflows. Such solutions are typically deployed on small-scale (e.g. single server) computing systems. However, as the sizes and complexities of datasets and associated neural network models increase, distributed systems become important for training and evaluating models in a feasible amount of time. In this poster, we describe our work on Jupyter notebook solutions for distributed training and hyper-parameter optimization of deep neural networks on high-performance computing systems.

Best Poster Finalist: no

Fast and Accurate Training of an AI Radiologist

Lucas A. Wilson (Dell EMC), Vineet Gundecha (Dell EMC), Srinivas Varadharajan (Dell EMC), Alex Filby (Dell EMC), Pei Yang (Dell EMC), Quy Ta (Dell EMC), Valeriu Codreanu (SURFsara), Damian Podareanu (SURFsara), Vikram Saletore (Intel Corporation)

The health care industry is expected to be an early adopter of AI and deep learning to improve patient outcomes, reduce costs, and speed up diagnosis. We have developed models for using AI to diagnose pneumonia, emphysema, and other thoracic pathologies from chest x-rays. Using the Stanford University CheXNet model as inspiration, we explore ways of developing accurate models for this problem with fast parallel training on Zenith, the Intel Xeon-based supercomputer at Dell EMC's HPC and AI Innovation Lab. We explore various network topologies to gain insight into what types of neural networks scale well in parallel and improve training time from days to hours. We then explore transferring this learned knowledge to other radiology subdomains, such as mammography, and whether this leads to better models than developing subdomain models independently.

Best Poster Finalist: no

Full State Quantum Circuit Simulation by Using Lossy Data Compression

Xin-Chuan Wu (University of Chicago, Argonne National Laboratory), Sheng Di (Argonne National Laboratory), Franck Cappello (Argonne National Laboratory), Hal Finkel (Argonne National Laboratory), Yuri Alexeev (Argonne National Laboratory), Frederic T. Chong (University of Chicago)

In order to evaluate, validate, and refine the design of a new quantum algorithm or a quantum computer, researchers and developers need methods to assess their correctness and fidelity. This requires the capabilities of simulation for full quantum state amplitudes. However, the number of quantum state amplitudes increases exponentially with the number of qubits, leading to the exponential growth of the memory requirement. In this work, we present our technique to simulate more qubits than previously reported by using lossy data compression. Our empirical data suggests that we can simulate full state

quantum circuits up to 63 qubits with 0.8 petabytes memory.

Best Poster Finalist: no

Enabling Reproducible Microbiome Science through Decentralized Provenance Tracking in QIIME 2
Ahmad Turan Naimey (Northern Arizona University, Pathogen and Microbiome Institute), Christopher Keefe (Northern Arizona University, Pathogen and Microbiome Institute)

In this poster, we demonstrate the ways in which automatic, integrated, decentralized provenance tracking in QIIME 2, a leading microbiome bioinformatics platform, enables reproducible microbiome science. We use sample data from a recent study of arid soil microbiomes (Significant Impacts of Increasing Aridity on the Arid Soil Microbiome; Neilson et al, 2017), to illustrate specific analyses that QIIME 2 supports, and to frame our discussion of the QIIME 2 platform.

QIIME 2 actions yield as outputs artifacts integrating the requested data or visualization with comprehensive data provenance that describes the computational history of that data or visualization, including all methods and parameters involved in its creation. This approach gives users, reviewers, and readers powerful tools for understanding, reproducing, and extending studies. The benefits this approach provides to both the researcher and the scientific community are significant and provide a useful model for research software developers across disciplines.

Best Poster Finalist: no

Optimizing Next Generation Hydrodynamics Code for Exascale Systems
Dana Akhmetova (KTH Royal Institute of Technology), Sumathi Lakshmiranganatha (University of Wyoming), Diptajyoti Mukherjee (Allegheny College), Frederick Oullet (University of Florida), Patrick Payne (Los Alamos National Laboratory), Nicholas Stegmeier (South Dakota State University), Christoph Junghans (Los Alamos National Laboratory), Robert Pavel (Los Alamos National Laboratory), Vinay Ramakrishnaiah (Los Alamos National Laboratory)

Studying continuum dynamics problems computationally can illuminate complex physical phenomena where experimentation is too costly. However, the models used in studying these phenomena usually require intensive calculations, some of which are beyond even the largest supercomputers to date. Emerging high performance computing (HPC) platforms will likely have varied levels of heterogeneity, making hybrid programming with MPI+X essential for achieving optimal performance. This research investigates hybrid programming and unconventional approaches like machine learning for a next generation hydrodynamics code, FleCSALE, in the context of tabular equation of state (EOS). We demonstrate an overall 5x speedup to the code, the use of GPUs to accelerate EOS tabular interpolation, and a proof of concept machine learning approach to EOS.

Best Poster Finalist: no

MPI/OpenMP parallelization of the Fragment Molecular Orbitals Method in GAMESS
Vladimir Mironov (Lomonosov Moscow State University), Yuri Alexeev (Argonne National Laboratory), Dmitri Fedorov (National Institute of Advanced Industrial Science and Technology (AIST))

In this work, we present a novel parallelization strategy for the Fragment Molecular Orbital (FMO) method in the quantum chemistry package GAMESS. The original FMO code has been parallelized only with MPI,

which limits scalability of the code on multi-core massively parallel machines. To address this problem, we parallelized FMO with a new hybrid MPI-OpenMP scheme that shows excellent scaling up to 2,048 Intel Xeon Phi nodes (131,072 cores) on Theta supercomputer. MPI-OpenMP code not only scales better compared to MPI code, but also performs up to two times faster and has significantly smaller memory footprint.

Best Poster Finalist: no

Automatic Generation of Mixed-Precision Programs

Logan Moody (Lawrence Livermore National Laboratory, James Madison University), Nathan Pinnow (Lawrence Livermore National Laboratory, Western Washington University), Michael O. Lam (James Madison University, Lawrence Livermore National Laboratory), Harshitha Menon (Lawrence Livermore National Laboratory), Markus Schordan (Lawrence Livermore National Laboratory), G. Scott Lloyd (Lawrence Livermore National Laboratory), Tanzima Islam (Western Washington University)

Floating-point arithmetic is foundational to scientific computing in HPC, and choices about floating-point precision can have a significant effect on the accuracy and speed of HPC codes. Unfortunately, current precision optimization tools require significant user interaction, and few work on the scale of HPC codes due to significant analysis overhead. We propose an automatic search and replacement system that finds the maximum speedup using mixed precision given a required level of accuracy. To achieve this, we integrated three existing analysis tools into a system that requires minimal input from the user. If a speedup is found, our system can provide a ready-to-compile mixed-precision version of the original program.

Best Poster Finalist: no

UPC++ and GASNet-EX: PGAS Support for Exascale Applications and Runtimes

Scott B. Baden (Lawrence Berkeley National Laboratory), Paul H. Hargrove (Lawrence Berkeley National Laboratory), Hadia Ahmed (Lawrence Berkeley National Laboratory), John Bachan (Lawrence Berkeley National Laboratory), Dan Bonachea (Lawrence Berkeley National Laboratory), Steven Hofmeyr (Lawrence Berkeley National Laboratory), Mathias Jacquelin (Lawrence Berkeley National Laboratory), Amir Kamil (Lawrence Berkeley National Laboratory), Brian van Straalen (Lawrence Berkeley National Laboratory)

Lawrence Berkeley National Lab is developing a programming system to support HPC application development using the Partitioned Global Address Space (PGAS) model. This work is driven by the emerging need for adaptive, lightweight communication in irregular applications at exascale. We present an overview of UPC++ and GASNet-EX, including examples and performance results.

GASNet-EX is a portable, high-performance communication library, leveraging hardware support to efficiently implement Active Messages and Remote Memory Access (RMA). UPC++ provides higher-level abstractions appropriate for PGAS programming such as: one-sided communication (RMA), remote procedure call, locality-aware APIs for user-defined distributed objects, and robust support for asynchronous execution to hide latency. Both libraries have been redesigned relative to their predecessors to meet the needs of exascale computing. While both libraries continue to evolve, the system already demonstrates improvements in microbenchmarks and application proxies.

Best Poster Finalist: no

An Efficient SIMD Implementation of Pseudo-Verlet Lists for Neighbor Interactions in Particle-Based

Codes

James Willis (Durham University, Institute for Computational Cosmology), Matthieu Schaller (Leiden Observatory), Pedro Gonnet (Google LLC)

In particle-based simulations, neighbour finding (i.e. finding pairs of particles to interact within a given range) is the most time consuming part of the computation. One of the best such algorithms, which can be used for both Molecular Dynamics (MD) and Smoothed Particle Hydrodynamics (SPH) simulations is the pseudo-Verlet list algorithm. The algorithm improves the neighbour finding by reducing the number of spurious pair-wise distance calculations. This algorithm, however, does not vectorize trivially, and hence makes it difficult to exploit SIMD-parallel architectures. On this poster, we present several novel modifications as well as a vectorization strategy for the algorithm which lead to overall speed-ups over the scalar version of the algorithm of 2.21x for the AVX instruction set (SIMD width of 8), 2.41x for AVX2, and 3.65x for AVX-512 (SIMD width of 16).

Best Poster Finalist: no

Understanding Potential Performance Issues Using Resource-Based alongside Time Models

Nan ding (Lawrence Berkeley National Laboratory), Victor W. Lee (Intel Corporation), Wei Xue (Tsinghua University), Weimin Zheng (Tsinghua University)

Numerous challenges and opportunities are introduced by the complexity and enormous code legacy of HPC applications, the diversity of HPC architectures, and the nonlinearity of interactions between applications and HPC systems. To address these issues, we propose the Resource-based Alongside Time (RAT) modeling method to help to understand the application run-time performance efficiently. First, we use hardware counter-assisted profiling to identify the key kernels and non-scalable kernels in the application. Second, we show how to apply the resource-based profiling into performance models to understand the potential performance issues and predict performance in the regimes of interest to developers and performance analysts. Third, we propose an easy-to-use performance modeling tool for scientists and performance analytics. Our evaluations demonstrate that by only performing a few small-scale profilings, RAT is able to keep the average model error rate around 15% with average performance overheads of 3% in multiple scenarios.

Best Poster Finalist: no

MGRIT Preconditioned Krylov Subspace Method

Ryo Yoda (Kogakuin University), Akihiro Fujii (Kogakuin University), Teruo Tanaka (Kogakuin University)

MGRIT re-discretize the problem with larger time-step width at the coarse-levels, which often cause unstable convergence. We propose a Krylov subspace method with MGRIT preconditioning as a more stable solver. For unstable problems, MGRIT preconditioned Krylov subspace method performed better than MGRIT in terms of the number of iterations. The contributions of the paper are organized as follows. We showed the matrix form of MGRIT operations, and the improvement of eigenvalue or singular-value distribution. We exemplified MGRIT with Krylov subspace method reaching convergence faster than MGRIT.

Best Poster Finalist: no

Enabling Neutrino and Antineutrino Appearance Observation Measurements with HPC Facilities

Norm Buchanan (Colorado State University), Steven Calvez (Colorado State University), Pengfei Ding (Fermi National Accelerator Laboratory), Derek Doyle (Colorado State University), Alex Himmel (Fermi National Accelerator Laboratory), Burt Holzman (Fermi National Accelerator Laboratory), Jim Kowalkowski (Fermi National Accelerator Laboratory), Andrew Norman (Fermi National Accelerator Laboratory), Alex Sousa (University of Cincinnati), Marc Paterno (Fermi National Accelerator Laboratory), Saba Sehrish (Fermi National Accelerator Laboratory), Brandon White (Fermi National Accelerator Laboratory), Christopher Green (Fermi National Accelerator Laboratory)

When fitting to data with low statistics and near physical boundaries, extra measures need to be taken to ensure proper statistical coverage. The method NOvA uses is called the Feldman-Cousins procedure, which entails fitting thousands of independent pseudoexperiments to generate acceptance intervals that are then used to correct our fits. The scale required by the Feldman-Cousins procedure makes it extremely computationally intensive. In past analyses, it has taken up to several weeks to complete, bottlenecking our final results. Here, I present recent work by members of the NOvA experiment and the SciDAC4 collaboration to enable the use of the supercomputing facilities at NERSC to process our Feldman-Cousins corrections over 50x faster, allowing us to perform more studies, increase the precision of our fits, and produce results quickly.

Best Poster Finalist: no

Large Scale Computation of Quantiles Using MELISSA

Alejandro Ribes (EDF Research and Development), Théophile Terraz (French Institute for Research in Computer Science and Automation (INRIA)), Yvan Fournier (EDF Research and Development), Bertrand Iooss (EDF Research and Development), Bruno Raffin (French Institute for Research in Computer Science and Automation (INRIA))

Quantiles being order statistics, the classical approach for their computation requires availability of the full sample before ranking it. This approach is not suitable at exascale. Large ensembles would need to gather a prohibitively large amount of data. We propose an iterative approach based on the stochastic quantile algorithm of Robbins-Monro. We rely on the Melissa framework, a file avoiding, adaptive, fault tolerant and elastic framework in order to compute in transit ubiquitous quantiles. Quantiles are updated on-the-fly as soon as the in transit parallel server receives results from one of the running simulations. We run 80,000 fluid dynamics parallel simulations of 6M hexahedra and 100 timespteps. They were executed on up to 4800 cores, avoiding 288 TB of file storage. We produce ubiquitous spatio-temporal maps of quantiles and inter-quantile based intervals.

Best Poster Finalist: no

FlowOS-RM: Disaggregated Resource Management System

Ryousei Takano (National Institute of Advanced Industrial Science and Technology (AIST)), Kuniyasu Suzaki (National Institute of Advanced Industrial Science and Technology (AIST)), Hidetaka Koie (National Institute of Advanced Industrial Science and Technology (AIST))

A traditional data center consists of monolithic-servers is confronted with limitations including lack of operational flexibility, low resource utilization, low maintainability, etc. Resource disaggregation is a promising solution to address the above issues. We propose a concept of disaggregated data center architecture called Flow-in-Cloud (FiC) that enables an existing cluster computer to expand an accelerator pool through a high-speed network. FiC is a shared pool of heterogeneous accelerators such as GPU and

FPGA, which are directly connected by a circuit-switched network. From the pool of accelerators, a slice is dynamically configured and provided according to a user request. FlowOS-RM manages the entire FiC resources, and supports execution of a user job on provided slices. This poster demonstrates effective resource sharing on the prototype system using a distributed deep learning application.

Best Poster Finalist: no

Programming the EMU Architecture: Algorithm Design Considerations for Migratory-Threads-Based Systems

Mehmet E. Belviranli (Oak Ridge National Laboratory), Seyong Lee (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory)

The decades-old memory bottleneck problem for data-intensive applications is getting worse as the processor core counts continue to increase. Workloads with sparse memory access characteristics only achieve a fraction of a system's total memory bandwidth. EMU architecture provides a radical approach to the issue by migrating the computational threads to the location where the data resides.

In EMU architecture, data distribution and thread creation strategies play a crucial role in achieving optimal performance in the EMU platform. In this work, we identify several design considerations that need to be taken care of while developing applications for the new architecture and we evaluate their performance effects on the EMU-chick hardware.

Best Poster Finalist: no

OpenACC to FPGA: A Directive-Based High-Level Programming Framework for High-Performance Reconfigurable Computing

Seyong Lee (Oak Ridge National Laboratory), Jacob Lambert (University of Oregon), Jungwon Kim (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory), Allen D. Malony (University of Oregon)

Accelerator-based heterogeneous computing has become popular solutions for power-efficient high performance computing (HPC). Along these lines, Field Programmable Gate Arrays (FPGAs) have offered more advantages in terms of performance and energy efficiency for specific workloads than other accelerators. Nevertheless, FPGAs have traditionally suffered several disadvantages limiting their deployment in HPC systems, mostly due to the challenges of programmability and portability. We present a directive-based, high-level programming framework for high-performance reconfigurable computing. It takes a standard, portable OpenACC C program as input and generates a hardware configuration file for execution on FPGAs. We implemented this prototype system in our open-source OpenARC compiler, which uses the Altera OpenCL compiler as its backend. Preliminary evaluation of the proposed framework on an Intel Stratix~V with five OpenACC benchmarks demonstrates that our proposed FPGA-specific compiler optimizations and novel OpenACC pragma extensions assist the compiler in generating more efficient FPGA programs.

Best Poster Finalist: no

Tensor-Optimized Hardware Accelerates Fused Discontinuous Galerkin Simulations

Alexander Breuer (University of California, San Diego), Alexander Heinecke (Intel Corporation), Yifeng Cui (San Diego Supercomputer Center)

In recent years the compute/memory balance of processors has been continuously shifting towards compute. The rise of Deep Learning, based on matrix multiplications, accelerated this path, especially in terms of single precision and lower precision compute. An important research question is if this development can be leveraged for traditional HPC. We demonstrate that a high-order discontinuous Galerkin solver for seismic wave propagation can execute in single precision without loss of modeling accuracy. Additionally, we extended its kernels to support the Intel Knights Mill CPU with 14 TFLOPS of single precision deep-learning performance. This allows us to harvest the hardware's special compute capabilities, even in an application with sparse linear algebra kernels. On cluster-level, Knights Mill can obtain the same application performance as the latest top-bin dual-socket Intel Xeon Platinum nodes, while consuming lower power. Compared to the HPC-focused Knights Landing processor, scenario-dependent speed-ups of up to 1.6× are possible.

Best Poster Finalist: no

AI Matrix – Synthetic Benchmarks for DNN

Wei Wei (Alibaba Inc), Lingjie Xu (Alibaba Inc), Lingling Jin (Alibaba Inc), Wei Zhang (Alibaba Inc), Tianjun Zhang (University of California, Berkeley)

The current AI benchmarks suffer from a number of drawbacks. First, they cannot adapt to the emerging changes of deep learning (DL) algorithms and are fixed once selected. Second, they contain tens to hundreds of applications and have very long running time. Third, they are mainly selected from open sources, which are restricted by copyright and not representable of the proprietary applications. To address these drawbacks, this work firstly proposes a synthetic benchmark framework that generates a small number of benchmarks that best represent a broad range of applications using their profiled workload characteristics. The synthetic benchmarks can adapt to new DL algorithms by re-profiling new applications and updating itself, greatly reduce number of benchmark tests and running time, and strongly represent DL applications of interests. The framework is validated by using log data profiled from DL models running on Alibaba AI platform, and is representable of real workload characteristics.

Best Poster Finalist: no

Applying the Execution-Cache-Memory Model: Current State of Practice

Georg Hager (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Jan Eitzinger (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Julian Hornich (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Francesco Cremonesi (Swiss Federal Institute of Technology in Lausanne), Christie L. Alappat (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Thoams Roehl (University of Erlangen-Nuremberg), Gerhard Wellein (University of Erlangen-Nuremberg, Erlangen Regional Computing Center)

The ECM (Execution-Cache-Memory) model is an analytic, resource-based performance model for steady-state loop code running on multicore processors. Starting from a machine model, which describes the interaction between the code and the hardware, and static code analysis, it allows an accurate prediction of the runtime of sequential loop code. Together with a scaling assumption, it also gives a performance scaling prediction. This poster summarizes the current state of practice in constructing and applying the ECM model, points out problems and open questions, and applies the model to three new and nontrivial use cases. For the first time, overlap assumptions for all relevant CPU architectures in high performance computing are presented.

Best Poster Finalist: yes

WarpX: Toward Exascale Modeling of Plasma Particle Accelerators

Maxence Thevenet (Lawrence Berkeley National Laboratory), Jean-Luc Vay (Lawrence Berkeley National Laboratory), Ann Almgren (Lawrence Berkeley National Laboratory), John Bell (Lawrence Berkeley National Laboratory), Remi Lehe (Lawrence Berkeley National Laboratory), Andrew Myers (Lawrence Berkeley National Laboratory), Jaehong Park (Lawrence Berkeley National Laboratory), Olga Shapoval (Lawrence Berkeley National Laboratory), Weiqun Zhang (Lawrence Berkeley National Laboratory), Lixin Ge (SLAC National Accelerator Laboratory), Mark Hogan (SLAC National Accelerator Laboratory), Cho Ng (SLAC National Accelerator Laboratory), Dave Grote (Lawrence Livermore National Laboratory)

Turning the current experimental plasma accelerator state-of-the-art from a promising technology into mainstream scientific tools depends critically on high-performance, high-fidelity modeling of complex processes that develop over a wide range of space and time scales. As part of the U.S. Department of Energy's Exascale Computing Project, a team composed of LBNL, SLAC and LLNL researchers is developing a new plasma accelerator simulation tool: WarpX. We will present the code structure and how it articulates around its main components: the new Particle-In-Cell Scalable Application Resource (PICSAR) and the adaptive mesh refinement library AMReX, which are combined with redesigned elements of the Warp code, in the new WarpX software. The status, examples of convergence, scaling and applications will be presented.

Best Poster Finalist: no

Job Simulation for Large-Scale PBS-Based Clusters with the Maui Scheduler

Georg Zitzlsberer (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic), Branislav Jansik (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic), Jan Martinovic (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic)

For large-scale High Performance Computing centers with a wide range of different projects and heterogeneous infrastructures, efficiency is an important consideration. Understanding how compute jobs are scheduled is necessary for improving the job scheduling strategies in order to optimize cluster utilization and job wait times. This increases the importance of a reliable simulation capability, which in turn requires accuracy and comparability with historic workloads from the cluster. Not all job schedulers have a simulation capability, including the Portable Batch System (PBS) resource manager. Hence, PBS based centers have no direct way to simulate changes and optimizations before they are applied to the production system. We propose and discuss how to run job simulations for large-scale PBS based clusters with the Maui Scheduler. For validation purposes, we use historic workloads collected at the IT4Innovations supercomputing center, and demonstrate the viability of our approach.

Best Poster Finalist: no

Script of Scripts Polyglot Notebook and Workflow System

Gao Wang (University of Chicago), Man Chong Leong (Rice University), Bo Peng (University of Texas, MD Anderson Cancer Center)

Computationally intensive disciplines such as computational biology often use tools implemented in different languages and analyze data on high-performance computing systems. Although scientific workflow systems can powerfully execute large-scale data-processing, they are not suitable for ad hoc data

analysis. Interactive tools such as Jupyter Notebook can be used for data exploration, but it remains difficult to work with multiple scripting languages and to streamline analysis for batch data processing. To bridge the gap between interactive and workflow systems we developed Script of Scripts (SoS), which consists of a polyglot notebook that supports multiple languages in a single notebook and a workflow engine that provides an intuitive syntax for multi-style workflows and a unified interface for executing tasks on a variety of computing platforms. By allowing the use of SoS workflow engine in a polyglot notebook environment, SoS provides a unified environment for both interactive data analysis and batch data processing.

Best Poster Finalist: no

Enabling High-Level Graph Processing via Dynamic Tasking

Maurizio Drocco (Pacific Northwest National Laboratory), Vito Giovanni Castellana (Pacific Northwest National Laboratory), Marco Minutoli (Pacific Northwest National Laboratory), Antonino Tumeo (Pacific Northwest National Laboratory), John Feo (Pacific Northwest National Laboratory)

Data-intensive computing yields irregular and unbalanced workloads, in particular on large-scale problems running on distributed systems. Task-based runtime systems are commonly exploited to implement higher-level data-centric programming models, promoting multithreading and asynchronous coordination for performance. However, coping with dynamic workloads (e.g., those yielded by large-scale graph processing) is challenging.

In this work, we took an exploratory approach to overcome some typical bottlenecks in tasking systems. In particular, we propose 1. a novel task allocator based on dynamic per-thread allocation and all-to-all recycling networks, and 2. a reservation-free remote spawning schema, based on receiver-side buffering and back-pressure feedback/sensing to avoid overflows.

As a proof of concept, we implemented the proposed techniques underneath a high-level library of distributed C++ containers. Preliminary experimental evaluation shows consistent scalability, a neat improvement in performance (e.g., 1.5x speedup with respect to the original code over an 8M-nodes graph), and less sensitiveness to parameter tuning.

Best Poster Finalist: no

An Alternative Approach to Teaching Bigdata and Cloud Computing Topics at CS Undergraduate Level

Debzani Deb (Winston-Salem State University), Muztaba Fuad (Winston-Salem State University), Keith Irwin (Winston-Salem State University)

Big data and cloud computing collectively offer a paradigm shift in the way businesses are now acquiring, using and managing information technology. This creates the need for every CS student to be equipped with foundation knowledge in this collective paradigm and to possess some hands-on-experience in deploying and managing big data applications in the cloud. We argue that, for substantial coverage of big data and cloud computing concepts and skills, the relevant topics need to be integrated into multiple core courses of undergraduate CS curriculum rather than creating additional standalone core or elective courses. Our approach to including these topics is to develop learning modules for specific core courses in which their coverage might find an appropriate context. In this poster, three such modules are presented and our classroom experiences during these interventions are documented. Our objective is to share our experience and to receive feedback about our approach.

Best Poster Finalist: no

Binarized ImageNet Inference in 29us

Tong Geng (Boston University, Pacific Northwest National Laboratory), Ang Li (Pacific Northwest National Laboratory), Tianqi Wang (Boston University), Shuaiwen Leon Song (Pacific Northwest National Laboratory), Martin Herbordt (Boston University)

We propose a single-FPGA-based accelerator for ultra-low-latency inference of ImageNet in this work. The design can complete the inference of Binarized AlexNet within 29us with accuracy comparable to other BNN implementations. We achieve this performance with the following contributions: 1. We completely remove floating-point from NL through layer fusion. 2. By using model parallelism rather than data parallelism, we can simultaneously configure all layers and the control flow graphs. Also, the design is flexible enough to achieve nearly perfect load balancing, leading to extremely high resource utilization. 3. All convolution layers are fused and processed in parallel through inter-layer pipelining. Therefore, in case the pipeline is full, latency is just the delay of a single convolution layer plus the FC layers. Note that the dependency pattern of the FC layer prevents it from being integrated into the current pipeline.

Best Poster Finalist: no

Refactoring and Optimizing Multiphysics Combustion Models for Data Parallelism

Christopher Stone (US Department of Defense HPC Modernization Program, Engility Corporation), Alexei Poludnenko (Texas A&M University), Brian Taylor (US Air Force Research Laboratory)

High-fidelity combustion simulations combine high-resolution computational fluid dynamics numerical methods with multi-physics models to capture chemical kinetics and transport processes. These multi-physics models can dominate the computation cost of the simulation. Due to the high cost of combustion simulations and the important role simulations play in propulsion and power research, acceleration methods are needed to reduce the computational time and cost. Multi-physics models within each mesh cell are often independent leading to significant parallelism. However, the iterative algorithms often impede efficient SIMD data parallelism, a key performance feature on modern HPC systems. Refactoring methods for multi-physics models (e.g., kinetics, equation-of-state, diffusion) with nonuniform workloads are demonstrated and benchmarked on a range of platforms (AVX2, KNL, AVX-512). Realized speed-ups over 6x were achieved on KNL and 4x on Skylake (SKX) for complex chemical kinetics models and over 3x on SKX for iterative EOS computations.

Best Poster Finalist: no

Tensorfolding: Improving Convolutional Neural Network Performance with Fused Microkernels

Michael Anderson (Intel Corporation), Evangelos Georganas (Intel Corporation), Sasikanth Avancha (Intel Corporation), Alexander Heinecke (Intel Corporation)

Convolution layers are prevalent in many classes of deep neural networks, including Convolutional Neural Networks (CNNs) which provide state-of-the-art results for tasks like image recognition, neural machine translation and speech recognition. In the recent past, several techniques to improve generalization capabilities of neural networks have been developed; the most prominent and successful is batch normalization. In deep neural network training, the batch normalization layer consists of a memory-bandwidth bound kernel. On the latest Intel Skylake based Xeon processors, a significant portion of execution time is spent in this kernel. By leveraging the CPU's large caches and its latency-optimized

execution model, we are able to reduce this kernel's time to a bare minimum while allowing to improve forward pass layer runtimes by 21% compared to an unfused implementation and by 2% compared to a fused implementation.

Best Poster Finalist: no

MATEDOR: MAtrix, TEnsor, and Deep-Learning Optimized Routines

Ahmad Abdelfattah (University of Tennessee), Jack Dongarra (University of Tennessee), Stanimire Tomov (University of Tennessee), Ichitaro Yamazaki (University of Tennessee), Azzam Haidar (Nvidia Corporation)

The MAtrix, TEnsor, and Deep-learning Optimized Routines (MATEDOR) project develops software technologies and standard APIs, along with a sustainable and portable library, for large-scale computations that can be broken down into very small matrix or tensor computations. The main target of MATEDOR is to accelerate applications from important fields that fit this profile, including deep learning, data mining, astrophysics, image and signal processing, hydrodynamics, and more.

MATEDOR is a high-performance numerical library for batched linear algebra subroutines autotuned for modern processor architectures and system designs. The MATEDOR library includes LAPACK-compliant routines that target many small dense problems, tensor, and application-specific operations, e.g., for deep-learning. These routines are constructed as much as possible out of calls to batch BLAS routines and their look-alikes required in sparse computation context.

Best Poster Finalist: no

Distributed Adaptive Radix Tree for Efficient Metadata Search on HPC Systems

Wei Zhang (Texas Tech University), Houjun Tang (Lawrence Berkeley National Laboratory), Suren Byna (Lawrence Berkeley National Laboratory), Yong Chen (Texas Tech University)

Affix-based search allows users to retrieve data without the need to remember all relevant information precisely. While building an inverted index to facilitate efficient affix-based search is a common practice for standalone databases and desktop file systems, they are often insufficient for high-performance computing (HPC) systems due to the massive amount of data and the distributed nature of the storage. In this poster, we present Distributed Adaptive Radix Tree (DART) which enables scalable and efficient affix-based search. DART maintains a balanced keyword distribution and optimizes for excessive keyword requests dynamically at scale. Our evaluation shows that compared with the “full string hashing” used by the commonly adopted DHT approach, DART achieves up to 55x throughput speedup for prefix and suffix search, and has a comparable throughput for exact and infix search. Also, DART maintains balanced keyword distribution and alleviates excessive query workload on popular keywords.

Best Poster Finalist: no

Improving Error-Bounded Lossy Compression for Cosmological N-Body Simulation

Sihuan Li (University of California, Riverside), Sheng Di (Argonne National Laboratory), Xin Liang (University of California, Riverside), Zizhong Chen (University of California, Riverside), Franck Cappello (Argonne National Laboratory)

Cosmological simulations may produce extremely large amount of data, such that its successful run depends on large storage capacity and huge I/O bandwidth, especially in the exascale computing scale.

Effective error-bounded lossy compressors with both high compression ratios and low data distortion can significantly reduce the total data size while guaranteeing the data valid for post-analysis. In this poster, we propose a novel, efficient compression model for cosmological N-body simulation framework, by combining the advantages of both space-based compression and time-based compression. The evaluation with a well-known cosmological simulation code shows that our proposed solution can get much higher compression quality than other existing state-of-the-art compressors, with comparable compression/decompression rates.

Best Poster Finalist: no

VeloC: Very Low Overhead Checkpointing System

Bogdan Nicolae (Argonne National Laboratory), Franck Cappello (Argonne National Laboratory), Adam Moody (Lawrence Livermore National Laboratory), Elsa Gonsiorowski (Lawrence Livermore National Laboratory), Kathryn Mohror (Lawrence Livermore National Laboratory)

Checkpointing large amounts of related data concurrently to stable storage is a common I/O pattern of many HPC applications. However, such a pattern frequently leads to I/O bottlenecks that lead to poor scalability and performance. As modern HPC infrastructures continue to evolve, there is a growing gap between compute capacity vs. I/O capabilities. Furthermore, the storage hierarchy is becoming increasingly heterogeneous: in addition to parallel file systems, it comprises burst buffers, key-value stores, deep memory hierarchies at node level, etc. In this context, state of art is insufficient to deal with the diversity of vendor APIs, performance and persistency characteristics. This poster proposes VeloC, a low-overhead checkpointing system specifically designed to address the checkpointing needs of future exascale HPC systems. VeloC offers a simple API at user level, while employing an advanced multi-level resilience strategy that transparently optimizes the performance and scalability of checkpointing by leveraging heterogeneous storage.

Best Poster Finalist: no

Estimating Molecular Dynamics Chemical Shift with GPUs

Eric F. Wright (University of Delaware), Mauricio H. Ferrato (University of Delaware)

Experimental chemical shifts (CS) from solution and solid state magic-angle-spinning nuclear magnetic resonance spectra provide atomic level data for each amino acid within a protein or complex. However, structure determination of large complexes and assemblies based on NMR data alone remains challenging due to the complexity of the calculations. Here, we present a hardware accelerated strategy for the estimation of NMR chemical-shifts of large macromolecular complexes. We demonstrate the feasibility of our approach in systems of increasing complexity ranging from 2,000 to 11,000,000 atoms.

Best Poster Finalist: no

Using Thrill to Process Scientific Data on HPC

Mariia Karabin (Clemson University, Los Alamos National Laboratory), Xinyu Chen (University of New Mexico), Supreeth Suresh (University of Wyoming), Ivo Jimenez (University of California, Santa Cruz), Li-Ta Lo (Los Alamos National Laboratory), Pascal Grosset (Los Alamos National Laboratory)

With ongoing improvement of computational power and memory capacity, the volume of scientific data keeps growing. To gain insights from vast amounts of data, scientists are starting to look at Big Data

processing and analytics tools such as Apache Spark. In this poster, we explore Thrill, a framework for big data computation on HPC clusters that provides an interface similar to systems like Apache Spark but delivers higher performance since it is built on C++ and MPI. Using Thrill, we implemented several analytics operations to post-process and analyze data from plasma physics and molecular dynamics simulations. Those operations were implemented with less programming effort than hand-crafted data processing programs would require and obtained preliminary results which were verified by scientists at LANL.

Best Poster Finalist: no

GPU Acceleration at Scale with OpenPower Platforms in Code_Saturne

Samuel Antao (IBM), Charles Moulinec (Science and Technology Facilities Council, UK), Yvan Fournier (EDF Research and Development), Robert Sawko (IBM), Malgorzata Zimon (IBM), Christopher Thompson (IBM), Alex Skillen (Science and Technology Facilities Council, UK), Juan Uribe (EDF Research and Development), David Emerson (Science and Technology Facilities Council, UK)

Code_Saturne is a widely used computational fluid dynamics software package that uses finite-volume methods to simulate different kinds of flows tailored to tackle multi-bilion-cell unstructured mesh simulations. This class of codes has shown to be challenging to accelerate on GPUs as they consist of many kernels and regular inter-process communication in between. In this poster we show how template pack expansion with CUDA can combine multiple kernels into a single one reducing launching latencies and along with the specification of data environments help reduce host-device communication. We tested these techniques on ORNL Summit Supercomputer based on OpenPOWER platform delivering almost 3x speedup over CPU-only runs on 256 nodes. We also show how the latest generation NVLINK(TM) interconnect available in POWER9(TM)improves scaling efficiency, enabling consistent GPU acceleration with just 100K-cells per process.

Best Poster Finalist: yes

Large-Message Size Allreduce at Wire Speed for Distributed Deep Learning

Kenji Tanaka (Japan Telegraph and Telephone Corporation), Yuki Arikawa (Japan Telegraph and Telephone Corporation), Kenji Kawai (Japan Telegraph and Telephone Corporation), Junichi Kato (Japan Telegraph and Telephone Corporation), Tsuyoshi Ito (Japan Telegraph and Telephone Corporation), Huy Cu Ngo (Japan Telegraph and Telephone Corporation), Kazutaka Morita (Japan Telegraph and Telephone Corporation), Fumiaki Miura (Japan Telegraph and Telephone Corporation), Takeshi Sakamoto (Japan Telegraph and Telephone Corporation), Satoshi Shigematsu (Japan Telegraph and Telephone Corporation)

In large-scale distributed deep learning, the Allreduce operation for large messages (100 KB or more) is critical for gathering gradients from multiple worker nodes and broadcasting the sum of the gradients to them. When the message is large, the latency in Allreduce operation would make it difficult to take advantage of large-scale distributed deep learning. To reduce the latency, we devised a dataflow architecture with an Allreduce-specific hardware accelerator that performs data aggregation and reduction while data is being transferred. The accelerator is designed to immediately start Allreduce operation before an entire message is recived. Furthermore, Allreduce can be operated at wire speed by vectorizing the gradients and summing them in parallel. Experimental results reveal that the proposed architecture performs Allreduce at 96% of wire speed for a large message. Moreover, the latency of Allreduce is reduced by 65% compared with a state-of-the-art Allreduce method when applied for ResNet-50.

Best Poster Finalist: no

Sol: Transparent Neural Network Acceleration Platform

Nicolas Weber (NEC Laboratories Europe, NEC Corporation)

With the usage of neural networks in a wide range of application fields, the necessity to execute these efficiently on high performance hardware is one of the key problems for artificial intelligence (AI) framework providers. More and more new specialized hardware types and corresponding libraries appear from various manufacturers. The biggest problem arising is that these libraries usually are only supported by a very limited set of AI frameworks and interoperability can become an issue. In this extended abstract we present Sol, a transparent middleware for neural network acceleration. Sol comes with an optimizing compiler engine, allowing to use device specific libraries and to implement own optimizations, that can be leveraged on all target devices. In contrast to other projects Sol explicitly aims at optimizing prediction and training of neural networks.

Best Poster Finalist: no

Detection of Silent Data Corruptions in Smooth Particle Hydrodynamics Simulations

Aurélien Cavelan (University of Basel), Florina M. Ciorba (University of Basel), Ruben M. Cabezón (University of Basel)

Soft errors, such as silent data corruptions (SDCs) hinder the correctness of large-scale scientific applications. Ghost replication (GR) is proposed herein as the first SDCs detector relying on the fast error propagation inherent to applications that employ the smooth particle hydrodynamics (SPH) method. GR follows a two-steps selective replication scheme. First, an algorithm selects which particles to replicate on a different process. Then, a different algorithm detects SDCs by comparing the data of the selected particles with the data of their ghost. The overhead and scalability of the proposed approach are assessed through a set of strong-scaling experiments conducted on a large HPC system under error-free conditions, using upwards of 3, 000 cores. The results show that GR achieves a recall and precision similar to that of full replication methods, at only a fraction of the cost, with detection rates of 91–99.9%, no false-positives, and an overhead of 1–10%.

Best Poster Finalist: no

DeepSim-HiPAC: Deep Learning High Performance Approximate Calculation for Interactive Design and Prototyping

Ahmed Al-Jarro (Fujitsu Laboratories Ltd), Serban Georgescu (Fujitsu Laboratories Ltd), Yasumoto Tomita (Fujitsu Laboratories Ltd), Kouta Nakashima (Fujitsu Laboratories Ltd)

We present a data-driven technique that can learn from physical-based simulations for the instant prediction of field distribution for 3D objects. Such techniques are extremely useful when considering, for example, computer aided engineering (CAE), where computationally expensive simulations are often required. To accelerate this process, we propose a deep learning framework that can predict the principal field distribution given a 3D object. This work allows us to learn a system's response using simulation data of arbitrarily shaped objects and an auto-encoder inspired deep neural network that maps the input of the 3D object shape to its principal 3D field distribution. We show that our engine, DeepSim-HiPAC, can estimate field distribution for two distinctive applications: micro-magnetics design in computational electromagnetics (CEM) and interactive cooling systems design in computational fluid dynamics (CFD), several orders of magnitude faster, up to 250000X, than the native calculations and at a cost of low error rate.

Best Poster Finalist: no

Top-Down Performance Analysis of Workflow Applications

Christian Herold (Technical University Dresden), Bill Williams (Technical University Dresden)

Scientific simulation frameworks are common to use on HPC systems. They contain parallelized algorithms and provide various solvers for a specific application domain. Usually, engineers execute multiple steps to solve a particular problem which are often distributed over multiple jobs. Finding performance bottlenecks and the causing step in such a complex system is very difficult. Therefore in this work, we present a top-down approach that provides summarized performance metrics for the workflow, jobs and job steps. These summaries guides the user to identify inefficiencies and determine the causing job step. Finally, Vampir can be used for a detailed analysis of the regarding execution in order to resolve the issue.

Best Poster Finalist: no

Convolutional Neural Networks for Coronary Plaque Classification in Intravascular Optical Coherence Tomography (IVOCT) Images

Chaitanya Kolluru (Case Western Reserve University), David Prabhu (Case Western Reserve University), Yanzan Gharaibeh (Case Western Reserve University), David Wilson (Case Western Reserve University), Sanjaya Gajurel (Case Western Reserve University)

Currently, IVOCT is the only imaging technique with the resolution necessary to identify vulnerable thin cap fibro-atheromas (TCFAs). IVOCT also has greater penetration depth in calcified plaques as compared to Intravascular Ultrasound (IVUS). Despite its advantages, IVOCT image interpretation is challenging and time consuming with over 500 images generated in a single pullback. In this poster, we propose a method to automatically classify A-lines in IVOCT images using a convolutional neural network. Conditional random fields were used to clean network predictions across frames. The neural network was trained using a dataset of nearly 4,500 image frames across 48 IVOCT pullbacks. Ten-fold cross validation with held-out pullbacks resulted in a classification accuracy of roughly 76% for fibrocalcific, 84% for fibrolipidic, and 85% for other. Classification results across frames displayed in en face view matched closely to annotated counterparts.

Best Poster Finalist: no

Compiling SIMT Programs on Multi- and Many-Core Processors with Wide Vector Units: A Case Study with CUDA

Hancheng Wu (North Carolina State University), John Ravi (North Carolina State University), Michela Becchi (North Carolina State University)

There has been an increasing interest in SIMT programming tools for multi- and manycore (co)processors with wide vector extensions. In this work, we study the effective implementation of a SIMT programming model (a subset of CUDA C) on Intel platforms with 512-bit vector extensions (hybrid MIMD/SIMD architectures). We first propose a set of compiler techniques to transform programs written using a SIMT programming model into code that leverages both the x86 cores and the vector units of a hybrid MIMD/SIMD architecture, thus providing programmability, high system utilization and portability. We then evaluate the proposed techniques on various hybrid systems using

microbenchmarks and real-world applications. Finally, we point out the main challenges in supporting the SIMT model on hybrid systems.

Best Poster Finalist: no

A Massively Parallel Evolutionary Markov Chain Monte Carlo Algorithm for Sampling Complicated Multimodal State SpacesState

Wendy K. Tam Cho (University of Illinois), Yan Liu (University of Illinois)

We develop an Evolutionary Markov Chain Monte Carlo (EMCMC) algorithm for sampling from large multi-modal state spaces. Our algorithm combines the advantages of evolutionary algorithms (EAs) as optimization heuristics and the theoretical convergence properties of Markov Chain Monte Carlo (MCMC) algorithms for sampling from unknown distributions. We harness massive computational power with a parallel EA framework that guides a large set of Markov chains. Our algorithm has applications in many different fields of science. We demonstrate its effectiveness with an application to political redistricting.

Best Poster Finalist: no

MLModelScope: Evaluate and Measure Machine Learning Models within AI Pipelines

Abdul Dakkak (University of Illinois), Cheng Li (University of Illinois), Wen-mei Hwu (University of Illinois), Jinjun Xiong (IBM)

The current landscape of Machine Learning (ML) and Deep Learning (DL) is rife with non-uniform frameworks, models, and system stacks but lacks standard tools to facilitate the evaluation and measurement of models. Due to the absence of such tools, the current practice for evaluating and comparing the benefits of proposed AI innovations (be it hardware or software) on end-to-end AI pipelines is both arduous and error prone — stifling the adoption of the innovations. We propose MLModelScope— a hardware/software agnostic platform to facilitate the evaluation, measurement, and introspection of ML models within AI pipelines. MLModelScope aids application developers in discovering and experimenting with models, data scientists developers in replicating and evaluating for publishing models, and system architects in understanding the performance of AI workloads.

Best Poster Finalist: no

A Compiler Framework for Fixed-Topology Non-Deterministic Finite Automata on SIMD Platforms

Marziyeh Nourian (North Carolina State University), Hancheng Wu (North Carolina State University), Michela Becchi (North Carolina State University)

Automata traversal acceleration has been studied on various parallel platforms. Many existing acceleration methods store finite automata states and transitions in memory. For these designs memory size and bandwidth are the main limiting factors to performance and power efficiency. Many applications, however, require processing several fixed-topology automata that differ only in the symbols associated to the transitions. This property enables the design of alternative, memory-efficient solutions. We target fixed-topology non-deterministic finite automata (NFAs) and propose a memory-efficient design, suitable to SIMD architectures, that embeds the automata topology in code and stores only the transition symbols in memory. We design a compiler that automates deployment of this design on SIMD platforms for a set of fixed-topology NFAs. Our compiler framework performs a

combination of platform-agnostic and platform-specific design decisions and optimizations. This poster describes the compiler toolchain and shows the achieved throughput on GPU and Intel SIMD devices.

Best Poster Finalist: no

A Low-Communicaton Method to Solve Poisson's Equation on Locally-Structured Grids
Brian Van Straalen (Lawrence Berkeley National Laboratory), Peter McCorquodale (Lawrence Berkeley National Laboratory), Phil Colella (Lawrence Berkeley National Laboratory), Christos Kavouklis (Lawrence Livermore National Laboratory)

This poster describes a new algorithm, Method of Local Corrections (MLC), and a high-performance implementation for solving Poisson's equation with infinite-domain boundary conditions, on locally-refined nested rectangular grids. The data motion is comparable to that of only a single V-cycle of multigrid, and hence is an order of magnitude smaller than traditional multigrid iteration. The computational kernels are 3D FFTs on small domains. Strong scaling tests on 64 to 4096 cores on NERSC Cori I (Haswell) show over 60% efficiency, and weak scaling by replication tests over 64 to 32768 cores show 92% efficiency on the same platform. We find comparable solve times between HPGMG on a uniform grid with one billion grid points, and MLC on the same number of grid points adaptively distributed. MLC is designed for AMR, able to solve problems with much higher resolution at the finest level than an algorithm on a uniform grid.

Best Poster Finalist: no

Floating-Point Autotuner for CPU-Based Mixed-Precision Applications
Ruidong Gu (North Carolina State University), Paul A. Beata (North Carolina State University), Michela Becchi (North Carolina State University)

In this poster, we present the design and development of an autotuning tool for floating-point code. The goal is to balance accuracy and performance in order to produce an efficient and accurate mixed-precision program. The tuner starts by maximizing accuracy through the use of a high-precision library called CAMPARY and then achieves performance gains under a given error bound by tuning down groups of variables and operations from the higher precision down to double precision. We tested our tuning strategy on a computational fluid dynamics benchmark where we show a 4x speedup relative to the fully high-precision version during the iterative tuning process and achieve an average absolute error of 2.8E-16 compared with the reference solution computed using the 256-bit GNU MPFR extended precision library.

Best Poster Finalist: no

8:30 am - 5:00 pm

ACM Student Research Competition Posters

Session Description: SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom. The schedule of the ACM Student Research Competition session will be made available Wednesday evening form the results of the

semi-finalists selection that will happen after the poster session.

SC18 Research Posters

SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Precomputing Outputs of Hidden Layers to Speed Up Deep Neural Network Training

Sohil Lal Shrestha (University of Texas, Arlington)

Deep learning has recently emerged as a powerful technique for many tasks including image classification. A key bottleneck of deep learning is that the training phase takes a lot of time, since state-of-the-art deep neural networks have millions of parameters and hundreds of hidden layers. The early layers of these deep neural networks have the fewest parameters but take up the most computation.

In this work, we reduce training time by progressively freezing hidden layers, pre-computing their output and excluding them from training in both forward and backward paths in subsequent iterations. We compare this technique to the most closely related approach for speeding up the training process of neural network.

Through experiments on two widely used datasets for image classification, we empirically demonstrate that our approach can yield savings of up to 25% wall-clock time during training with no loss in accuracy.

Designing Shared Address Space MPI Libraries in Many-Core Era

Jahanzeb Maqbool Hashmi (Ohio State University)

The emergence of modern multi-/many-cores has put more emphasis on optimizing intra-node communication. Existing designs in MPI libraries that work on the concept of distributed address spaces incur the overhead of intermediate memory copies to stage the data between processes. This can lead to severe performance degradation especially on emerging many-core architectures like Intel Skylake and IBM OpenPOWER. This work proposes a high-performance "shared address-space"-based MPI point-to-point and collective communication designs using XPMEM. We first characterize the bottlenecks associated with XPMEM based communication and propose new designs for efficient MPI large message communication. Then we propose novel collective designs that are contention-free and offer true zero-copy reduction operations. The proposed designs are evaluated on different multi-/many-core architectures using various micro-benchmarks and application kernels such as MiniAMR and AlexNet DNN training on CNTK. The proposed designs have shown significant performance improvement over state-of-the-art available in MPI libraries.

Modeling Single-Source Shortest Path Algorithm Dynamics to Control Performance and Power Tradeoffs

Sara Karamati (Georgia Institute of Technology), Jeffrey Young (Georgia Institute of Technology), Rich Vuduc (Georgia Institute of Technology)

This work presents a new methodology to improve the performance of parallel algorithms by tuning the amount of available parallelism for execution throughout the runtime. As such, we expose key parameters controlling the performance and parallelism of the algorithm and build a software-based controller with the objective of maintaining the optimal performance. Our controller allows for tuning the level of parallelism executed in each time epoch to optimize for performance while preserving power usage. More specifically, our experimental evaluation focuses on a tunable variation of a GPU-based delta-stepping algorithm for computing the single-source shortest path (SSSP); As the available parallelism for the delta-stepping SSSP is highly irregular and strongly input-dependent, our extensive experiments show that average power can be reduced while average parallelism is increased. This increase in average parallelism provides substantial energy savings, independent of the hardware.

Accelerating DNA Long Read Mapping with Emerging Technologies

Roman Kaplan (Israel Institute of Technology)

DNA sequencing technologies output only short fragments of a genome, called reads. New single-molecule real-time sequencing technologies can produce long reads, up to tens of thousands base pairs, within minutes. However, these long reads may contain up to 15% errors.

To construct a genome from DNA reads, a computationally expensive bioinformatics task, read mapping, is required. Read mapping finds the best-fitting location for each DNA read on a long reference sequence. The length and error rate of long reads poses a challenge for existing read mapping hardware solutions, designed for short reads with low error rates. This work presents a novel DNA read mapping hardware architecture, RASSA. RASSA is a Resistive Approximate Similarity Search Accelerator that exploits charge distribution and parallel in-memory processing to reflect a mismatch count between DNA sequences. RASSA implementation of long read DNA mapping outperforms state-of-the-art long read mapping solution by 16-77x with comparable accuracy.

SimFS: A Simulation Data Virtualizing File System Interface

Salvatore Di Girolamo (ETH Zurich)

In the big (simulation) data era, simulations often produce petabytes of data to be stored in parallel filesystems or large-scale databases. This data is accessed, often by thousands of analysts and scientists, over the course of decades. However, storing these volumes of data for long time periods of time is not cost effective and, in some cases, practically impossible.

SimFS transparently virtualizes the simulation output, relaxing the storage requirements and re-simulating missing data on-demand. SimFS monitors the analysis access pattern in order to decide (1) which data to store and (2) apply prefetching techniques to improve the analysis performance. SimFS enables a trade-off between on-disk solutions, where all the simulation data is stored on disk, and in-situ, where no data is stored and analyses are always coupled with simulations. Overall, by exploiting the growing computing power and relaxing the storage capacity requirements, SimFS offers a viable path towards exa-scale simulations.

Holistic Root Cause Analysis of Node Failures in Production HPC

Anwesha Das (North Carolina State University)

Production HPC clusters endure failures incurring computation and resource wastage. Despite the presence of various failure detection and prediction schemes, a comprehensive understanding of how nodes fail considering various components and layers of the system is required for sustained resilience. This work performs a holistic root cause diagnosis of node failures using a measurement-driven approach on contemporary system logs that can help vendors and system administrators support exascale resilience.

Our work shows that lead times can be increased by at least 5 times if external subsystem correlations are considered as opposed to considering the events of a specific node in isolation. Moreover, when detecting sensor measurement outliers and interconnect related failures, triggering automated recovery events can exacerbate the situation if recovery is unsuccessful.

Geomancy: Automated Data Placement Optimization

Oceane Bel (University of California, Santa Cruz)

Exascale cloud storage and High-Performance Computing Systems (HPC) deliver unprecedented storage capacity and levels of computing power, though the full potential of these systems remain untapped because of inefficient data placement. Changes in data access patterns can cause a system's performance to suffer. To mitigate performance losses, system designers implement strategies to preemptively place popular data on higher performance nodes. However, these strategies fail to address a diverse userbase whose users individually demand the highest performance, and they must be carefully constructed by an expert of the system.

We propose Geomancy, a tool that reorganizes data to increase I/O throughput. In systems where heuristic-based improvements may become resource intensive, Geomancy determines new placement policies by training a deep neural network with past workload and system traces. With real workload traces, Geomancy calculated an example placement policy that demonstrated a 49% increase in average throughput compared to the default data layout.

Numerical Simulation of a Flue Instrument with Finite-Difference Lattice Boltzmann Method using GPGPU

Ryoya Tabata (Kyushu Institute of Technology)

In this work, we discuss the possibility of using GPGPU techniques for Aeroacoustic Simulation (especially for flue instruments) with the finite-difference lattice Boltzmann method (FDLBM). Compressible flow simulation has been used in direct aeroacoustic simulation; however, the computational cost is huge due to the requirement of high computational mesh resolution, with small time steps. The lattice Boltzmann method (LBM) has been used as an efficient method for fluid simulation using GPGPU. However, LBM is not accurate enough when applied to some aeroacoustic problems. On the other hand, FDLBM is able to use high-order finite-difference schemes and it has a high arithmetic intensity compared to LBM. We present a performance evaluation of the LBM and FDLBM with several finite-difference schemes on GPU with the roofline model.

Recursive Algebraic Coloring Engine

Christie Louis Alappat (University of Erlangen-Nuremberg)

Many iterative numerical methods for sparse systems and building blocks of sparse linear algebra are difficult to parallelize due to data dependencies. These may be loop-carried dependencies as they occur in solvers like Gauss-Seidel or write conflicts as in symmetric sparse matrix vector. Most of the existing parallelization strategies suffer from low performance on modern hardware, are matrix specific, or require tailored storage formats.

In this poster, we introduce a novel recursive level based algorithm called Recursive Algebraic Coloring (RAC), which achieves high hardware efficiency on modern multi-core architectures and works with simple data formats like compressed row storage. Method is implemented in a library called Recursive Algebraic Coloring Engine (RACE). Thorough performance analysis shows that RACE outperforms traditional multicoloring methods and Intel-MKL implementations with a factor of 2–2.5×. We are on par with Algebraic Block Multicoloring for small matrices, while for large matrices we gain a factor of 1.5–2×.

Accelerating Microscope Data Analysis Using Parallel Computing
John Ravi (North Carolina State University)

Single-Molecule Localization Microscopy (SMLM) techniques deal with the diffraction limit of fluorescent microscopy by localizing single molecules with high precision by stochastically switching molecules on and off. Thousands of camera frames containing subsets of blinking molecules are recorded to obtain a single super-resolution image. Each blinking molecule in each frame is subjected to localization protocols that fit the shape of the blink, assess the quality of the blink and then estimate their center. The algorithm implemented originally in MATLAB and compiled CUDA C, to compute a 'Super Resolution' image took around 6 minutes to process 256x256 pixel images of a moderately dense dataset. I ported the algorithm to C++ and parallelized it using OpenMP to compute multiple frames in parallel.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

PotC: Many-Body Potential Implementations à La Carte
Markus Höhnerbach (RWTH Aachen University)

Molecular dynamics is a valuable investigation tool for simulations in computational chemistry and

materials science. In these simulations, atoms move according to so-called potentials, functions that typically describe the distance-dependent interactions between pairs of atoms. For some applications more complex "many-body potentials" are required. From a HPC perspective, such potentials pose challenges: small neighborhoods hinder vectorization; redundant force expressions are tedious and error prone to derive; the implementations are large, runtime-critical, and can not be broken into simpler "kernels". Consequently, only selected many-body potentials receive optimizations---with major effort.

PotC is a DSL for MD potentials and corresponding compiler to generate high-performance implementations. The compiler centralizes optimization knowledge for many-body potentials, and unburdens users from manual optimization and force derivation. The performance optimization of the generated code is still work-in-progress. However, independent of performance, the tool can also be applied when prototyping, testing, or possibly symbolically manipulating potentials.

OoO Instruction Benchmarking Framework on the Back of Dragons

Julian Hammer (University of Erlangen-Nuremberg, RRZE)

In order to construct an accurate instruction execution model for modern out-of-order micro architectures, an accurate description of instruction latency, throughput and concurrency is indispensable. Already existing resources and vendor provided information is neither complete nor detailed enough and sometimes incorrect. We therefore proclaim to deduct this information through runtime instruction benchmarking and present a framework to support such investigations based on LLVM's just-in-time and cross-platform compilation capabilities.

pyasmjit abstracts instructions, operands, and dependency chains, to easily construct the necessary benchmarks. The synthesized code is interactively compiled and executed using the llvmlite library, which in turn is based on the stable LLVM C-API. pyasmjit offers a command line as well as a programming interface. Unlike other approaches, we do not rely on model specific performance counters and focus on interoperability and automation to support quick modeling of future microarchitectures.

Studying the Impact of Power Capping on MapReduce-Based, Data-Intensive Mini-Applications on Intel KNL and KNM Architectures

Joshua H. Davis (University of Delaware)

In this poster, we quantitatively measure the impacts of data movement on performance in MapReduce-based applications when executed on HPC systems. We leverage the PAPI 'powercap' component to identify ideal conditions for execution of our applications in terms of (1) dataset characteristics (i.e., unique words); (2) HPC system (i.e., KNL and KNM); and (3) implementation of the MapReduce programming model (i.e., with or without combiner optimizations). Results confirm the high energy and runtime costs of data movement, and the benefits of the combiner optimization on these costs.

Accelerating Big Data Processing in the Cloud with Scalable Communication and I/O Schemes

Shashank Gugnani (Ohio State University)

With the advent of cloud computing, the field of Big Data has seen rapid growth. Most cloud providers provide hardware resources such as NVMe SSDs, large memory nodes, and SR-IOV. This opens up the possibility of large-scale high-performance data analytics and provides opportunities to use these resources to develop new designs. Cloud computing provides flexibility, security, and reliability, which are important requirements for Big Data frameworks. However, several important requirements are missing, such as performance, scalability, consistency, and quality of service (QoS). The focus of this work revolves around developing communication and I/O designs and concepts which can provide these requirements to Big Data frameworks. Specifically, we explore new ways to provide QoS and consistency in cloud storage systems, and provide scalable and high-performance communication frameworks.

Accelerating 2D FFT: Exploit GPU Tensor Cores through Mixed-Precision

Xiaohe Cheng (Hong Kong University of Science and Technology), Anumeena Sorna (National Institute of Technology, Tiruchirappalli)

The two-dimensional Fourier Transform is a widely-used computational kernel in many HPC applications. The popular NVIDIA cuFFT library provides a simple interface to compute 2D FFT on GPUs, but it's yet to utilize the recent hardware advancement in half-precision floating-point arithmetic. In this poster, we propose a mixed-precision method to accelerate 2D FFT by exploiting the FP16 matrix-multiply-and-accumulate units on the newest GPU architecture, known as tensor cores. We achieve a balance between speed and accuracy by dynamically splitting the single-precision input data into two half-precision operands and performing FFT separately. We present a CUDA-based implementation that achieves 3-digit more accuracy than half-precision cuFFT. We also demonstrate the stability and scalability of our approach and conclude that it attains high accuracy with tolerable splitting overhead.

Monitoring Parsl Workflows

Connor Pigg (University of Illinois)

As a Python library that enables workflows, Parsl gives users the ability to define complex workflows in Python and run them in parallel on any computer system. This poster describe the process of adding monitoring to Parsl. Simple and comprehensive monitoring of a workflow's state and resource usage lets users audit, debug, and confirm workflow execution. The poster discusses how Parsl monitors workflow components, what data it captures (task status and resource usage), and the tools it used to do so (Elasticsearch) and to display the information (Kibana). A Kibana dashboard visualizes the collected logs in a real time, with an interactive user interface. This enhanced Parsl allows users the option to monitor the status and resource usage of their workflows via an Elasticsearch database and Kibana dashboard.

Identifying Network Data Transfer Bottlenecks in HPC Systems

Karen Tu (Lawrence Berkeley National Laboratory; University of California, Berkeley)

Improving network data transfer performance is a major factor for improving high performance computing systems. Most studies analyze data transfer and file system IO performance separately, but understanding the relationship between the two is essential for optimizing scheduling and resource

management. Intuitively, if data is being transferred to a busy file system the transfer rate would be slower than a file system at regular activity levels.

This study analyzes patterns between file system activity and network throughput for several use cases of file writing and data transfers using a parallel file system. The parameters changed among the use cases were file striping for the file system, and buffer size and parallelism for data transfer. The main bottleneck for network data transfer rate was the number of OSTs the data was striped across. For a large number of OSTs (16 or greater), writing to the file system was the bottleneck.

Dendro-GR: Massively Parallel Simulations of Binary Black Hole Intermediate-Mass-Ratio Inspirals
Milinda Fernando (University of Utah)

We present a portable and highly-scalable algorithm and framework that targets problems in the astrophysics and numerical relativity communities. This framework combines together a parallel octree-refined adaptive mesh with wavelet adaptive multiresolution and a physics module to solve the Einstein equations of general relativity in the BSSN~formulation. The goal of this work is to perform advanced, massively parallel numerical simulations of Intermediate Mass Ratio Inspirals (IMRIs) of binary black holes with mass ratios on the order of 100:1. These studies will be used to generate waveforms for use in LIGO data analysis and to calibrate semi-analytical approximate methods. This advanced framework is designed to easily accommodate many existing algorithms in astrophysics for plasma dynamics and radiation hydrodynamics. We have designed novel algorithms to enable efficient simulations for such experiments and demonstrate excellent weak scalability up to 131K cores on ORNL's Titan for binary mergers for mass ratios up to 100.

Hardware Transactional Persistent Memory
Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories
Luann C. Jung (Massachusetts Institute of Technology, University of Chicago), Brendan T. Whitaker (Ohio State University, University of Chicago)

As scientific data repositories and filesystems grow in size and complexity, they become increasingly

disorganized. The coupling of massive quantities of data with poor organization makes it challenging for scientists to locate and utilize relevant data, thus slowing the process of analyzing data of interest. To address these issues, we explore an automated clustering approach for quantifying the organization of data repositories. Our parallel pipeline processes heterogeneous filetypes (e.g., text and tabular data), automatically clusters files based on content and metadata similarities, and computes a novel "cleanliness" score from the resulting clustering. We demonstrate the generation and accuracy of our cleanliness measure using both synthetic and real datasets, and conclude that it is more consistent than other potential cleanliness measures.

Supercomputing for the Multi-Driver Routing

Zeyang Ye (Stony Brook University)

Supercomputing is essential for routing traffic by providing drivers the optimal routes with minimal traveling distances or time. The unique challenges that require supercomputers to overcome are of multiple folds: numerous drivers, massive simultaneous requests, multiple locations, and needs of instant gratifications, etc. We developed two parallel methods, PSAD and PSAD-M, by using domain decomposition and state-mixing techniques. On the same computing platform with 96 cores, for the same problem, our PSAD methods outperform all published benchmarks by over a hundred times, while improving the solution quality. For the same routing problem on 384 cores, our PSAD-M reduced the elapsed time from the unbearable ten minutes to the reasonable 5 seconds, achieving a record-breaking speedup of 170. By providing instant routing solutions that enable online recommendations, our methods break the bottleneck of the widely adopted offline approaches.

NautDB: Toward a Hybrid Runtime for Processing Compiled Queries

Samuel Grayson (University of Texas, Dallas)

General purpose operating and database system suffer under the load of their generality which makes achieving optimal performance extremely hard, especially on modern hardware. The goal of this research is to integrate, for the first time, specialization techniques from the OS community (hybrid runtimes) and DB community (compiled queries) for high-performance query processing on modern hardware. We envision a system called NautDB, a hybrid dataflow runtime for executing compiled queries. As a first step toward our goal, we evaluate the performance of compiled queries on Linux and run as a Nautilus hybrid runtime using a simple prototype. Our results demonstrate that combining these specialization techniques has transformative potential for building the next generation (distributed) high-performance query processing systems and big data platforms.

Mitigating Performance and Progress Variability in Iterative Asynchronous Algorithms

Justs Zarins (University of Edinburgh)

Large HPC machines are susceptible to irregular performance. Factors like chip manufacturing differences, heat management, and network congestion combine to result in varying execution time for the same code and input sets. Asynchronous algorithms offer a partial solution. In these algorithms, fast workers are not forced to synchronize with slow ones. Instead they continue computing updates, and moving toward the solution, using the latest data available to them, which may have become stale (i.e. a number of iterations out of date compared to the most recent data). While this allows for high

computational efficiency, the convergence rate of asynchronous algorithms tends to be lower.

To address this problem, we are using the unique properties of asynchronous algorithms to develop load balancing strategies for iterative asynchronous algorithms in both shared and distributed memory. Our poster shows how our solution attenuates noise, resulting in significant reduction progress imbalance and time-to-solution variability.

Eulerian Algorithms for the Discretization of Plasma Kinetic Equations
James L. Juno (University of Maryland)

While fluid models are common tools in the study of plasmas, many of these systems, whether in astrophysics or the lab, are only weakly collisional and far from equilibrium, making them more accurately described by kinetic equations. Kinetic equations can be computationally demanding due to the need to solve for the distribution function of the particles in a higher dimensional phase space, with position and velocity coordinates. Despite this challenge, the motivation for solving the plasma kinetic equation is large as there remains a vast array of questions concerning collisionless dynamics in real plasma systems. Here we present algorithms in an Eulerian framework for the discretization of the plasma kinetic equation, using a high-order discontinuous Galerkin finite element method due to its arithmetic intensity and parallelizability. Scaling and performance of the algorithm are discussed, and benchmarks of the algorithm are presented as well.

Thursday, November 15th

Room: C2/3/4 Ballroom
8:30 am - 5:00 pm

Research Posters

Session Description: SC18 Research Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

SC18 Research Posters
SC18 Research Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Exploring Application Performance on Fat-Tree Networks in the Presence of Congestion
Philip A. Taffet (Rice University, Lawrence Livermore National Laboratory), Sanil Rao (University of Virginia, Lawrence Livermore National Laboratory), Ian Karlin (Lawrence Livermore National Laboratory)

Network congestion, which occurs when multiple applications simultaneously use shared links in cluster network, can cause poor communication performance, decreasing the performance and scalability of parallel applications. Many studies are performed while clusters also run other production workloads, which makes it harder for them to isolate causes and their effects. To look at congestion in

a more controlled setting we used dedicated access time on an HPC cluster and measured the performance of three HPC applications with different communication patterns run with varying amounts and types of background traffic. This enables us to assess the relative sensitivity of the applications to congestion caused by different traffic patterns. Our tests show that the applications were not significantly impacted by even the most aggressive neighboring patterns, with all the performance degradation being 7% or less, pointing to the resiliency of the fat-tree topology.

Best Poster Finalist: no

GPU-Accelerated Interpolation for 3D Image Registration

Naveen Himthani (University of Texas, Institute for Computational Engineering and Sciences), Andreas Mang (University of Houston), Amir Gholami (University of California, Berkeley), George Biros (University of Texas, Institute for Computational Engineering and Sciences)

Image registration is a key technology in image computing with numerous applications in medical imaging. Our overarching goal is the design of a consistent and unbiased computational framework for the integration of medical imaging data with simulation and optimization to support clinical decision making for glioma brain tumor patients. A major issue in 3D image registration is the time to solution, which poses the demand for effective numerical schemes and the utilization of high performance computing resources.

In this poster, we extend present a GPU-accelerated implementation of the Lagrange interpolation kernel using hardware texture filtering feature of modern hardware. Ongoing work involves implementing a unified single-GPU code for 3D image registration along with other computational kernels such as FFT. I will present my work by briefly explaining image registration followed by the explanation of the interpolation algorithm and its features and then demonstrate the results obtained.

Best Poster Finalist: no

Energy Efficiency of Reconfigurable Caches on FPGAs

Tianqi Wang (Boston University), Ang Li (Pacific Northwest National Laboratory), Tong Geng (Boston University), Martin Herbordt (Boston University)

The performance of a given cache architecture depends largely on the applications that run on it. Even though each application has its best-suited cache configuration, vendors of fixed HPC systems must provide compromise designs. Reconfigurable caches can adjust cache configuration dynamically to get best-suited cache parameters in runtime and notably reduce energy consumption. For example, when it is possible to deploy a low capacity low associativity design without increasing the miss rate substantially. For modern multi-core processors, each core's memory access behavior can be influenced by other cores. So it is more complicated to design reconfigurable caches for them.

In this paper, a design for a reconfigurable cache on FPGAs is presented that can run in modes with different capacities, associativity. We demonstrate that better performance and energy efficiency can be achieved by tuning these cache parameters at runtime.

Best Poster Finalist: no

RGB (Redfish Green500 Benchmarker): A Green500 Benchmarking Tool Using Redfish

Elham Hojati (Texas Tech University), Yong Chen (Texas Tech University), Alan Sill (Texas Tech University), Jon Hass (Dell Inc)

Performance and energy are important factors for supercomputers and data-centers with a trade-off between them. Energy efficiency metric considers both of these properties. The Green500 is a branch of Top500 project which provides a list of supercomputers based on energy efficiency. It has a manual methodology for this calculation.

Redfish is a new generation of management technologies for the hardware layer of data-centers. Our project focuses on designing and developing an automated Green500 benchmark tool using Redfish, called Redfish Green500 Benchmarker, or RGB in short. It offers the integration of Redfish and Green500, and automates Green500 benchmarking process with leveraging the internal capability of Redfish enabled equipment. It also enhances the Redfish standard to make sure it addresses the requirements of Green500 calculations. This research will also conduct validation and evaluation of RGB on real-world clusters for small-scale to medium-scale tests, and on the data-center simulator we have developed.

Best Poster Finalist: no

Optimization of Ultrasound Simulations on Multi-GPU Servers

Filip Vaverka (Brno University of Technology, Faculty of Information Technology), Matej Spetko (Brno University of Technology, Faculty of Information Technology), Bradley E. Treeby (University College London, Biomedical Ultrasound Group), Jiri Jaros (Brno University of Technology, Faculty of Information Technology)

Realistic ultrasound simulations have found a broad area of applications in preoperative photoacoustic screening and non-invasive ultrasound treatment planing. However, the domains are typically thousands of wavelengths in size, leading to large-scale numerical models with billions of unknowns. The current trend in accelerated computing is towards the use of fat nodes with multiple GPUs per node. The multi-GPU version of our k-Wave acoustic toolbox is based on the local Fourier basis domain decomposition where 3D simulation domain is partitioned into rectangular cuboid blocks assigned to particular GPUs. This paper investigates the benefits of using the CUDA-Aware MPI and CUDA peer-to-peer transfers on an 8-GPU server equipped with Nvidia P40 GPUs. The server has a total GPU memory of 192 GB and a single-precision performance of 96 Tflops. These techniques reduces the overall simulation time a factor of 2-3.6.

Best Poster Finalist: no

GPGPU Performance Estimation with Core and Memory Frequency Scaling

Qiang Wang (Hong Kong Baptist University), Xiaowen Chu (Hong Kong Baptist University)

Graphics processing units (GPUs) support dynamic voltage and frequency scaling to balance computational performance and energy consumption. However, simple and accurate performance estimation for a given GPU kernel under different frequency settings is still lacking for real hardware, which is important to decide the best frequency configuration for energy saving. We reveal a fine-grained analytical model to estimate the execution time of GPU kernels with both core and memory frequency scaling. Over a wide scaling range of both core and memory frequencies among 20 GPU kernels, our model achieves accurate results (4.83% error on average) with real hardware. Compared

to the cycle-level simulators, our model only needs simple micro-benchmarks to extract a set of hardware parameters and kernel performance counters to produce such high accuracy.

Best Poster Finalist: no

Making Sense of Scientific Simulation Ensembles

Mai Dahshan (Virginia Tech), Nicholas Polys (Virginia Tech)

Scientists run many simulations with varying initial conditions, known as "ensembles", to understand the influence and relationships among multiple parameters or ensemble members. Most of the ensemble visualization and analysis approaches and techniques focus on analyzing the relationships between either the ensemble members or output parameter space while neglecting the effect of input parameters and humans in the analysis loop. Therefore, we developed an approach to the visual analysis of scientific data that merges human expertise and intuition with machine learning and statistics allowing scientists to explore, search, filter, and make sense of their high dimensional ensemble. Our tool, "GLEE" (Graphically-Linked Ensemble Explorer), is an interactive visualization tool that consists of three visual views: Ensemble View, Parameter View, and Statistical View. Each view offers different functionality for exploration and interoperation of the relations and correlations between different runs, a subset of runs, and input and output parameters.

Best Poster Finalist: no

Which Architecture Is Better Suited for Matrix-Free Finite-Element Algorithms: Intel Skylake or Nvidia Volta?

Martin Kronbichler (Technical University Munich), Momme Allalen (Leibniz Supercomputing Centre), Martin Ohlerich (Leibniz Supercomputing Centre), Wolfgang A. Wall (Technical University Munich)

This work presents a performance comparison of highly tuned matrix-free finite element kernels from the finite element library on different contemporary computer architectures, NVIDIA V100 and P100 GPUs, an Intel Knights Landing Xeon Phi, and two multi-core Intel CPUs (Broadwell and Skylake). The algorithms are based on fast integration on hexahedra using sum factorization techniques. For small problem sizes, when all data fits into CPU caches, Skylake is very competitive with Volta. For larger sizes, however, the GPU holds an advantage of approximately a factor of three over Skylake, because all architectures operate in the memory-bandwidth limited regime. A detailed performance analysis contrasts the throughput-oriented character of GPUs versus the more latency-optimized CPUs for the scenario of high-order finite element computations.

Best Poster Finalist: no

SpotSDC: an Information Visualization System to Analyze Silent Data Corruption

Zhimin Li (University of Utah), Harshitha Menon (Lawrence Livermore National Laboratory), Yarden Livnat (University of Utah), Kathryn Mohror (Lawrence Livermore National Laboratory), Valerio Pascucci (University of Utah)

Aggressive technology scaling trends are expected to make the hardware of HPC systems more susceptible to transient faults. Transient faults in hardware may be masked without affecting the program output, cause a program to crash, or lead to silent data corruptions (SDC). While fault injection studies can give an overall resiliency profile for an application, without a good visualization

tool it is difficult to summarize and highlight critical information obtained. In this work, we design SpotSDC, a visualization system to analyze a program's resilience characteristics to SDC. SpotSDC provides an overview of the SDC impact on an application by highlighting regions of code that are most susceptible to SDC and will have a high impact on the program's output. SpotSDC also enables users to visualize the propagation of error through an application execution.

Best Poster Finalist: no

High-Accuracy Scalable Solutions to the Dynamic Facility Layout Problem

Apan Qasem (Texas State University), Clara Novoa (Texas State University), Chandra Kolla (Texas State University), Samantha Coyle (Texas State University)

The dynamic facility layout problem (DFLP) is concerned with finding arrangements of facilities within plant locations that minimize the sum of material handling and relocation costs over a planning horizon. DFLP is relevant in manufacturing engineering; accurate solutions can reduce operational costs by as much as 30%. We present a new scalable solution that formulates the task of finding the optimal arrangement as a shortest-path (SP) problem. The new parallel algorithm to find the SP employs a problem-specific heuristic to substantially cut down the search space. Compiler-level optimizations improve the performance across different execution platforms, including an auto-tuning strategy to derive the optimal SMT configuration on a POWER8 system. Results show a factor of 13 speedup over existing methods. For the six-facilities problems the best known solution is reached and for sets with 15 and 30 facilities the solution is within 2.83% and 5.66% of the best solution, respectively.

Best Poster Finalist: no

Large Scale MPI-Parallelization of LBM and DEM Systems: Accelerating Research by Using HPC

Bohumir Jelinek (Mississippi State University), George Mason (Mississippi State University), John Peters (Mississippi State University), Daniel Johnson (Mississippi State University), Marcus Brumfield (Mississippi State University), Alex Carrillo (US Army Engineer Research and Development Center), Clay Goodman (Mississippi State University), Farshid Vahedifard (Mississippi State University)

Casting, solidification, and the behavior of dry, saturated, and partially saturated granular media are examples of interesting and important problems in multiple areas of civil, mechanical, and chemical engineering. For interacting particle-fluid systems, the Discrete Element Method (DEM) and Lattice-Boltzmann Method (LBM) provide valuable high-resolution numerical models. Their main issue is high computational demand, which can be addressed by use of HPC resources. This work demonstrates the use of MPI-parallelized LBM and DEM models to accelerate research in solidification and macroscopic behavior of dry and saturated granular media. Large scale parallel simulations of dendritic growth, the calibration-chamber cone penetration test, and a parametric study of shear thickening in granular suspension were performed. Use of HPC dramatically reduced the computational time for these studies and provided high-resolution representation of physical experiments.

Best Poster Finalist: no

SciGaP: Apache Airavata Hosted Science Gateways

Marlon Pierce (Indiana University), Suresh Marru (Indiana University), Eroma Abeysinghe (Indiana University), Sudhakar Pamidighantam (Indiana University), Marcus Christie (Indiana University),

Dimuthu Upeksha (Indiana University)

The goal of the Science Gateways Platform as a service (SciGaP.org) project is to provide core services for building and hosting science gateways. Over the last two years, SciGaP services have been used to build and host over twenty-five science gateways. SciGaP services support these gateways through a single hosted version of the Apache Airavata software system that supports multiple tenants. Apache Airavata services include scientific application execution management on HPC and cloud environments, input and output data staging, and provenance tracking for user-created computational experiments.

The poster presents highlights of some selected SciGaP-hosted gateways. Clients interested in SciGaP services can request a tenant through https://scigap.org/. Clients who need extensive support for building user interfaces and integrating unconventional resources can request support through the Science Gateways Community Institute's Extended Developer Support program. To integrate with XSEDE resources, clients can request support through XSEDE's Extended Collaborative Support Services.

Best Poster Finalist: no

Reproducibility as Side Effect

Shu Wang (University of Chicago), Zhuo Zhen (University of Chicago), Jason Anderson (University of Chicago), Kate Keahey (Argonne National Laboratory, University of Chicago)

The ability to keep records and reproduce experiments is a critical element of the scientific method for any discipline. However, the recording and publishing of research artifacts that allow to reproduce and directly compare against existing research continue to be a challenge. In this paper, we propose an experiment précis framework that helps the experiment repeatability. Guided by the framework, we implement a prototype tool called ReGen which generates repeatable experiment scripts that can be used or shared along with a detailed experiment description automatically. The evaluation shows that ReGen is effective in reducing the researcher's efforts of creating a repeatable experiment in a real setting.

Best Poster Finalist: no

Using Darshan and CODES to Evaluate Application I/O Performance

Harsh Khetawat (North Carolina State University), Christopher Zimmer (Oak Ridge National Laboratory), Frank Mueller (North Carolina State University), Sudharshan Vazhkudai (Oak Ridge National Laboratory), Scott Atchley (Oak Ridge National Laboratory)

Burst buffers have become increasingly popular in HPC systems, allowing bursty I/O traffic to be serviced faster without slowing down application execution. The ubiquity of burst buffers creates opportunities for studying their ideal placement in the HPC topology. Furthermore, the topology of the network interconnect can also affect the performance of the storage hierarchy for different burst buffer placement schemes. To that end, we create a reproducible framework that allows individual centers to develop their own models and evaluate performance based on their workload characteristics. We use CODES to create models that simulate the network and storage layers of an HPC system and Darshan traces for I/O replay. We augment the Darshan traces with synchronization primitives, and allow multi-dimensional scaling of traces to represent future workload characteristics.

Finally, we evaluate the effect of network topology, storage architecture, and application I/O patterns on overall I/O performance.

Best Poster Finalist: no

Multi-Client DeepIO for Large-Scale Deep Learning on HPC Systems

Yue Zhu (Florida State University), Fahim Chowdhury (Florida State University), Huansong Fu (Florida State University), Adam Moody (Lawrence Livermore National Laboratory), Kathryn Mohror (Lawrence Livermore National Laboratory), Kento Sato (Lawrence Livermore National Laboratory), Weikuan Yu (Florida State University)

With the growth of computation power, leadership High-Performance Computing (HPC) systems can train larger datasets for Deep neural networks (DNNs) more efficiently. On HPC systems, a training dataset is on a parallel file system or node-local storage devices. However, not all HPC clusters have node-local storage, and large mini-batch sizes stress the read performance of parallel systems since the large datasets cannot fit in file system caches. Thus, it is a challenge for training DNNs with large datasets on HPC systems.

In prior work, we proposed DeepIO to mitigate the I/O pressure. DeepIO is designed to assist the mini-batch generation of TensorFlow. However, DeepIO does not support multiple training workers on a single compute node. We address this gap with modification on DeepIO framework, and evaluate multi-client DeepIO performance against state-of-the-art in-memory file systems, compare DeepIO and TensorFlow data loading API, and explore the potential of DeepIO in DNN training.

Best Poster Finalist: no

HIVE: A Cross-Platform, Modular Visualization Ecosystem for Heterogeneous Computational Environments

Jorji Nonaka (Riken Center for Computational Science), Kenji Ono (Kyushu University, RIKEN), Naohisa Sakamoto (Kobe University, RIKEN), Kengo Hayashi (Kobe University, RIKEN), Tomohiro Kawanabe (Riken Center for Computational Science), Fumiyoshi Shoji (Riken Center for Computational Science), Masahiro Fujita (LTE Inc), Kentaro Oku (Kashika Inc), Kazuma Hatta (Imagica Digitalscape)

HPC operational environments usually have supporting computational systems for assisting pre- and post-processing activities such as the visualization and analysis of simulation results. A wide variety of hardware systems can be found at different HPC sites, and in our case, we have a CPU-only (x86) large memory server, a planned OpenStack-based CPU/GPU Cluster, SPARC64 fx CPU based HPC system (K computer), and an ARM based HPC system in the future. Therefore, heterogeneity and scalability are needed to be tackled to efficiently use these heterogeneous computational resources for large-scale data visualization on both post-hoc and in-situ contexts. In this poster we present HIVE (Heterogeneously Integrated Visual-analytics Environment), a cross-platform and modular ecosystem for providing visualization service building blocks in such heterogeneous computational environments. Lightweight Lua scripting language is used to glue necessary visualization pipeline related modules, and this loosely coupled modular approach facilitates long-term development and maintenance.

Best Poster Finalist: no

Improving the I/O Performance and Memory Usage of the Xolotl Cluster Dynamics Simulator

Philip C. Roth (Oak Ridge National Laboratory), Sophie Blondel (University of Tennessee), David E. Bernholdt (Oak Ridge National Laboratory), Brian D. Wirth (University of Tennessee)

Xolotl is a cluster dynamics simulator used to predict gas bubble evolution in solids. It is currently being used to simulate bubble formation in the plasma-facing surface within fusion reactors and the nuclear fuel used in fission reactors. After observing performance problems in coupled-code simulations of fusion reactors, we used Xolotl's built-in performance data collection infrastructure and an external profiling tool to identify inefficiencies when writing Xolotl's two types of checkpoint files. We changed the code to use true parallel writes via the HDF5 data management library, resulting in a code that is approximately 57x faster when writing the program's main checkpoint file at the scale used in the coupled-code simulations, and that exhibits less performance variability due to external system activity. We also identified and addressed a memory usage problem that reduced Xolotl peak memory usage by approximately 88% per compute node.

Best Poster Finalist: no

Performance Evaluation of the Shifted Cholesky QR Algorithm for Ill-Conditioned Matrices

Takeshi Fukaya (Hokkaido University), Ramaseshan Kannan (Arup UK), Yuji Nakatsukasa (National Institute of Informatics, Japan), Yusaku Yamamoto (University of Electro-Communications, Japan), Yuka Yanagisawa (Waseda University)

The Cholesky QR algorithm, which computes the QR factorization of a matrix, is a simple yet efficient algorithm for high-performance computing. However it suffers from numerical instability. In a recent work, this instability has been remedied by repeating Cholesky QR twice (CholeskyQR2). ChokeskyQR2, however, is still prone to numerical breakdown when applied to ill-conditioned matrices. To overcome this limitation, we introduce a shifting technique to Cholesky QR and use it as a preconditioning step before CholeskyQR2. The key idea is that Cholesky QR with shift reduces the condition number of the input matrix. We call the resulting algorithm shifted CholeskyQR3, which is still simple and only requires double precision arithmetic. In this poster, we present the results of our performance evaluation of shifted CholeskyQR3. We demonstrate that shifted CholeskyQR3 accurately computes the QR factorization of ill-conditioned matrices and that it outperforms other conventional algorithms in execution time.

Best Poster Finalist: no

HPC-as-a-Service for Life Sciences

Vaclav Svaton (Technical University of Ostrava, Czech Republic), Jan Martinovic (Technical University of Ostrava, Czech Republic), Nina Jeliazkova (IDEAconsult Ltd, Bulgaria), Vladimir Chupakhin (Janssen Pharmaceutika NV), Pavel Tomancak (Max Planck Institute of Molecular Cell Biology and Genetics), Petr Vojta (Palacký University Olomouc, Czech Republic)

HPC-as-a-Service is a well-known term in the area of high performance computing. It enables users to access an HPC infrastructure without a need to buy and manage their own infrastructure. Through this service, academia and industry can take advantage of the technology without an upfront investment in the hardware. This approach further lowers the entry barrier for users who are interested in utilizing massive parallel computers but often do not have the necessary level of expertise in the area of parallel computing. To provide this simple and intuitive access to the supercomputing infrastructure, an in-house application framework called HEAppE has been developed. HEAppE's universally designed software architecture

enables unified access to different HPC systems through a simple object-oriented API. Thus providing HPC capabilities to the users but without the necessity to manage the running jobs from the command-line interface of the HPC scheduler directly on the cluster.

Best Poster Finalist: no

Hermes: a Multi-Tiered Distributed I/O Buffering System for HDF5

Hariharan Devarajan (Illinois Institute of Technology, HDF Group)

High-Performance Computing (HPC) systems' increasing ability to run data-intensive problems at larger scale and resolution has driven the evolution of modern storage technologies. In addition, extreme amounts of data are collected by large scientific instruments and sensor network is resulting in a push for more capable storage systems. Hierarchical Data Format (HDF) technologies address the problems of how to organize, access, analyze, and preserve data in the face of enormous growth in size and complexity. To mitigate this I/O performance bottleneck, modern storage subsystems are going through extensive changes, by adding additional levels of memory and storage in a hierarchy. In this work, we present Hermes: a new, heterogeneous-aware, multi-tiered, dynamic, and distributed I/O buffering system for HDF5. Hermes enables, manages, and supervises I/O buffering in DMSH. We tested our solution on Cori and show Hermes can accelerate applications by 62x than the state of the buffering platform.

Best Poster Finalist: no

Workflow for Parallel Processing of Sequential Mesh Databases

Ondřej Meca (Technical University of Ostrava, Czech Republic), Lubomír Říha (Technical University of Ostrava, Czech Republic), Tomáš Brzobohatý (Technical University of Ostrava, Czech Republic)

This poster presents a workflow for parallel loading of sequentially stored mesh databases. It can be used as a connection between tools for the creation of complex engineering models along with parallel solvers to allow broader usage of HPC by the engineering community. Scalability tests show that the proposed algorithm is able to prepare a mesh with hundreds of millions of nodes/elements in several seconds.

Best Poster Finalist: yes

The NAStJA Framework: Non-Collective Scalable Global Communications

Marco Berghoff (Karlsruhe Institute of Technology), Ivan Kondov (Karlsruhe Institute of Technology)

In recent years, simulations in various areas of science and engineering have proven to be very useful. To efficiently deploy simulation codes on current and future high-performance computer systems, high node level performance, scalable communication and the exclusion of unnecessary calculations are an absolute must when developing new solvers.

We have introduced the NAStJA framework, a block-based MPI parallel solver for algorithms, based on regular grid methods, i.e., stencil codes. NAStJA has a dynamic block adaptation, which modifies the calculation domain around the region in which the calculation is currently taking place. The creation and deletion of blocks are autonomously managed within local neighborhoods. Collective all-gather communication is avoided by using a multi-hop network to distribute information across the entire domain that greatly improves the application scaling. In this contribution, we present applications that can benefit from this adaptive method and scaling tests demonstrating the excellent scalability.

Best Poster Finalist: no

Hardware Acceleration of CNNs with Coherent FPGAs

Md Syadus Sefat (Texas State University), Semih Aslan (Texas State University), Apan Qasem (Texas State University)

This paper describes a new flexible approach to implementing energy-efficient CNNs on FPGAs. Our design leverages the Coherent Accelerator Processor Interface (CAPI) which provides a cache-coherent view of system memory to attached accelerators. Convolution layers are formulated as matrix multiplication kernels and then accelerated on CAPI-supported Kintex FPGA board. Our implementation bypasses the need for device driver code and significantly reduces the communication and I/O transfer overhead. To improve the performance of the entire application, not just the convolution layers, we propose a collaborative model of execution in which the control of the data flow within the accelerator is kept independent, freeing-up CPU cores to work on other parts of the application. For further performance enhancements, we propose a technique to exploit data locality in the cache, situated in the CAPI Power Service Layer (PSL). Finally, we develop a resource-conscious implementation for more efficient utilization of resources and improved scalability.

Best Poster Finalist: no

Distributed Fast Boundary Element Methods

Michal Merta (Technical University of Ostrava, Czech Republic), Jan Zapletal (Technical University of Ostrava, Czech Republic), Michal Kravcenko (Technical University of Ostrava, Czech Republic)

We present a parallel implementation of the fast boundary element method (BEM) for the Helmholtz equation. After a brief description of BEM, vectorization of the computationally most demanding kernels, and shared memory parallelization, we focus on the distributed memory parallelization using a new approach for distribution of the fast BEM system matrices among computational nodes. Moreover, we present a modification of the adaptive cross approximation (ACA) method capable of dealing with BEM matrices containing zero blocks, which usually lead to problems for the original ACA algorithm. Numerical experiments demonstrating the performance of the implementation were carried out using several Intel architectures.

Best Poster Finalist: no

Development of Numerical Coupled Analysis Method by Air Flow Analysis and Snow Accretion Analysis

Kohei Murotani (Railway Technical Research Institute, Japan), Koji Nakade (Railway Technical Research Institute, Japan), Yasushi Kamata (Railway Technical Research Institute, Japan), Daisuke Takahashi (Railway Technical Research Institute, Japan)

In this research, to take countermeasures for the snow accretion damage, we developed a simulator of realizing the snow accretion process in the following steps. Firstly, air flow analysis is performed by “Airflow simulator” developed by RTRI (Railway Technical Research Institute). Secondly, trajectory of flying snow is calculated by the equation of motion for gravity and drag using distribution of velocity of air flow. Thirdly, snow accretion analysis is performed by “Particle simulator” developed by RTRI. The shape modified by snow accretion is reflected onto the boundary conditions of the air flow analysis. In this year, we show the results of the snow accretion analysis for simple cubic shapes in order to aim at system development and

validation and discuss the result of the snow accretion analysis for a train bogie model.

Best Poster Finalist: no

Portable Parallel Performance via Multi-Dimensional Homomorphisms
Ari Rasch (University of Münster), Richard Schulze (University of Münster), Sergei Gorlatch (University of Münster)

Achieving portable performance over different parallel architectures and varying problem sizes is hard: e.g., a program optimized for multi-core CPUs on large input sizes can significantly differ from the same program optimized for Graphics Processing Units (GPUs) on small sizes.

We propose an approach to ensuring portability of performance by relying on multi-dimensional homomorphisms (MDHs) -- a class of parallelizable functions that cover important application areas including linear algebra routines (BLAS) and stencil computations. We develop an extended OpenCL implementation schema for MDHs that is generic in the performance-critical parameters of the OpenCL model, and we enable portability of performance by being automatically optimized for different target architectures and input sizes using the auto-tuning approach.

Our results demonstrate competitive and often even significantly better performance than state-of-the-art approaches for BLAS and Stencil as used in the important application area of deep learning.

Best Poster Finalist: no

Performance Evaluation of the NVIDIA Tesla V100: Block Level Pipelining vs. Kernel Level Pipelining
Xuewen Cui (Virginia Tech), Thomas R. W. Scogland (Lawrence Livermore National Laboratory), Bronis R. de Supinski (Lawrence Livermore National Laboratory), Wu Feng (Virginia Tech)

As accelerators become more common, expressive and performant, interfaces for them become ever more important. Programming models like OpenMP offer simple-to-use but powerful directive-based offload mechanisms. By default, these models naively copy data to or from the device without overlapping computation. Achieving performance can require extensive hand-tuning to apply optimizations such as pipelining. To pipeline a task, users must manually partition the task into multiple chunks then launch multiple sub-kernels. This approach can suffer from high kernel launch overhead. Also, the hyper parameters must be carefully tuned to achieve optimal performance. To ameliorate this issue, we propose a block-level pipeline approach that overlaps data transfers and computation in one kernel handled by different streaming multiprocessors on GPUs. Our results show that, without exhaustive tuning, our approach can provide 95% to 108% stable performance compared to the best tuned results with traditional kernel-level pipelining on NVIDIA V100 GPUs.

Best Poster Finalist: no

Enabling Data Analytics Workflows Using Node-Local Storage
Tu Mai Anh Do (University of Southern California, Information Sciences Institute), Ming Jiang (Lawrence Livermore National Laboratory), Brian Gallagher (Lawrence Livermore National Laboratory), Albert Chu (Lawrence Livermore National Laboratory), Cyrus Harrison (Lawrence Livermore National Laboratory), Karan Vahi (University of Southern California, Information Sciences Institute), Ewa Deelman (University of Southern California, Information Sciences Institute)

The convergence of high-performance computing (HPC) and Big Data is a necessity with the push toward extreme-scale computing. As HPC simulations become more complex, the analytics need to process larger amounts of data, which poses significant challenges for coupling HPC simulations with Big Data analytics. This poster presents a novel node-local approach that uses a workflow management system (WMS) to enable the coupling between the simulations and the analytics in scientific workflows by leveraging node-local non-volatile random-access memory (NVRAM).

Best Poster Finalist: no

OpeNNdd: Open Neural Networks for Drug Discovery: Creating Free and Easy Methods for Designing Medicine

Bryce Kroencke (American River College), Shawn Shacterman (University of California, Berkeley), Nicholas Pavini (American River College), Benjamin Samudio (American River College, Sierra College), Silvia Crivelli (Lawrence Berkeley National Laboratory)

Bringing new medicines to patients can be prohibitively expensive in terms of time, cost, and resources. This leaves many diseases without therapeutic interventions. In addition, new and reemerging diseases are increasing in prevalence across the globe at an alarming rate. The speed and scale of medicine discovery must be increased to effectively meet this challenge. OpeNNdd is a neural network platform bringing together people, machine learning, and supercomputing to solve the challenge of creating medicines. We have developed a novel neural network which quickly and accurately models candidate medicines interacting with a disease target, a metric to delineate its domain of applicability, and a process that communicates neural network results to participants in a readily interpretable way. OpeNNdd leverages the scale of supercomputing, the power and speed of neural networks, and the creativity of people across the globe in an open and collaborative way to protect and improve global health.

Best Poster Finalist: no

FeatherCNN: Fast Inference Computation with TensorGEMM on ARM Architectures

Haidong Lan (Shandong University), Jintao Meng (Tencent Holdings Ltd), Christian Hundt (Johannes Gutenberg University Mainz), Bertil Schmidt (Johannes Gutenberg University Mainz), Minwen Deng (Tencent Holdings Ltd), Weiguo Liu (Shandong University), Yanjie Wei (Shenzhen Institutes of Advanced Technology), Shengzhong Feng (Shenzhen Institutes of Advanced Technology)

This poster presents a fast inference computation library for ARM architecture named as CNNForward. CNNForward is trying to improve the efficiency of inference computation for convolutional neural networks on ARM-based multi-core and many-core architectures using both mathematical formula reconstruction/simplification and in-depth NEON instruction optimization. Experimental results reveal that, forward computation for VGG-16 on a server with 64 ARM A72 cores, CNNForward can scale up to 32 cores with an parallel efficiency of 33%, and achieve 35.4x, 8.7x and 10.6x speedup over Caffe+OpenBlas, Caffe2+Eigen and Caffe2+NNPACK, respectively.

Best Poster Finalist: no

Boosting the Scalability of Car-Parrinello Molecular Dynamics Simulations for Multi- and Manycore Architectures

Tobias Klöffel (University of Erlangen-Nuremberg), Bernd Meyer (University of Erlangen-Nuremberg), Gerald Mathias (Leibniz Supercomputing Centre)

We present our recent optimizations of the ultra-soft pseudo-potential (USPP) code path of the ab inito molecular dynamics program CPMD (www.cpmd.org). Following the internal instrumentation of CPMD, all relevant USPP routines have been revised to fully support hybrid MPI+OpenMP parallelization. For two time-critical routines, namely the multiple distributed 3D FFTs of the electronic states and a key distributed matrix-matrix multiplication, we have implemented hybrid parallel algorithms with overlapping computation and communication. The achievements in performance and scalability are demonstrated on a small reference system of 128 water molecules and further systems of increasing size. Performance evaluation shows gains of up to one order of magnitude and around 50% peak performance for simulation systems readily used in production.

Best Poster Finalist: no

Characterizing Declustered Software RAID for Enhancing Storage Reliability and Performance

Zhi Qiao (University of North Texas), Song Fu (University of North Texas), Hsing-bung Chen (Los Alamos National Laboratory), Bradley Settlemyer (Los Alamos National Laboratory)

Redundant array of independent disks (RAID) has been widely used to address the reliability issue in storage systems. As the scale of modern storage systems continues growing, disk failure becomes the norm. With ever-increasing disk capacity, RAID recovery based on disk rebuild becomes more costly, which causes significant performance degradation and even unavailability of storage systems. Declustered parity and data placement in RAID aims to enhance the recovery performance by shuffling data among all disks in a RAID group. All disks in the RAID group participate in data reconstruction, which leads to reduction of the RAID rebuild time. In this work, we extensively evaluate declustered RAID in terms of the performance of application I/O and recovery time. Our experimental results in ZFS show that the speedup of declustered RAID over traditional RAID is sub-linear to the number of disks in the storage pool.

Best Poster Finalist: no

Parallel Implementation of Machine Learning-Based Many-Body Potentials on CPU and GPU

Yaoguang Zhai (University of California, San Diego), Nathaniel Danandeh (University of California, San Diego), Zhenye Tan (University of California, San Diego; Tongji University), Sicun Gao (University of California, San Diego), Francesco Paesani (University of California, San Diego), Andreas W. Goetz (San Diego Supercomputer Center)

Machine learning models can be used to develop highly accurate and efficient many-body potentials for molecular simulations based on the many-body expansion of the total energy. A prominent example is the MB-pol water model that employs permutationally invariant polynomials (PIPs) to represent the 2-body and 3-body short-range energy terms.

We have recently shown that the PIPs can be replaced by Behler-Parinello neural networks (BP-NN). We present OpenMP parallel implementations of both PIP and BP-NN models as well as a CUDA implementation of the BP-NN model for GPUs. The OpenMP implementations achieve linear speedup with respect to the optimized single threaded code. The BP-NN GPU implementation outperforms the CPU implementation by a factor of almost 8. This opens the door for routine molecular dynamics simulations with highly accurate many-body potentials on a diverse set of hardware.

Best Poster Finalist: no

Implementing Efficient Data Compression and Encryption in a Persistent Key-Value Store for HPC
Jungwon Kim (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory)

Recently, persistent data structures, like key-value stores (KVSs), which are stored in an HPC system's nonvolatile memory, provide an attractive solution for a number of emerging challenges like limited I/O performance. This paper investigates how to efficiently integrate data compression and encryption into persistent KVSs for HPC with the ultimate goal of hiding their costs and complexity in terms of performance and ease of use. We implement the proposed techniques on top of a distributed embedded KVS to evaluate the benefits and costs of incorporating these capabilities along different points in the dataflow path, illustrating differences in effective bandwidth, latency, and additional computational expense.

Best Poster Finalist: yes

A Parallel-Efficient GPU Package for Multiphase Flow in Realistic Nano-Pore Networks
Yidong Xia (Idaho National Laboratory), Ansel Blumers (Brown University, Idaho National Laboratory), Zhen Li (Brown University), Lixiang Luo (IBM), Jan Goral (University of Utah), Matthew Andrew (Carl Zeiss X-ray Microscopy Inc), Joshua Kane (Idaho National Laboratory), Yu-Hang Tang (Lawrence Berkeley National Laboratory)

Simulations of fluid flow in oil/gas shale rocks are challenging in part due to the heterogeneous pore sizes ranging from a few nanometers to a few micrometers. Additionally, the complex fluid-solid interaction occurring physically and chemically must be captured with high resolution. To address these challenges while minimizing computational cost, we present a GPU code that has implemented a many-body dissipative particle dynamics (mDPD) model for multiphase flow in shale. Realistic nano- to micro-pore channels in shale are constructed from 3D high-resolution stack images. In our benchmark tests, the code delivers nearly perfect weak and strong scalings on up to 512 K20X GPUs on Oak Ridge National Laboratory (ORNL) Titan supercomputer. Moreover, single-GPU bench-marks on the DGX-1 (V100/no NVLink), ORNL's SummitDev (P100/NVLink 1.0) and Summit (V100/NVLink 2.0) suggest that the latest Host-to-Device NVLink can significantly boost overall performance, in addition to the Device-to-Device NVLink.

Best Poster Finalist: no

Processing-in-Storage Architecture for Machine Learning and Bioinformatics
Roman Kaplan (Israel Institute of Technology), Leonid Yavits (Israel Institute of Technology), Ran Ginosar (Israel Institute of Technology)

User-generated and bioinformatics database volumes has been increasing exponentially for more than a decade. With the slowdown and approaching end of Moore's law, traditional technologies cannot satisfy the increasing demands for processing power. This work presents PRINS, a highly-parallel in-storage processing architecture. PRINS combines non-volatile memory with processing capabilities on every bitcell. An emerging technology, memristors, form the basis for the design.

Implementations of three data-intensive and massively parallel algorithms are demonstrated: (1) Smith-Waterman DNA local sequence alignment (bioinformatics), (3) K-means clustering (machine learning) and (3) data deduplication. Performance and energy efficiency of PRINS compared to other published solutions is presented for each algorithm. PRINS is shown to achieve orders-of-magnitude improvement in performance and power efficiency over existing solutions, from large-scale bioinformatics and machine-

learning to single-GPU or FPGA implementations.

Best Poster Finalist: no

Kernel-Based and Total Performance Analysis of CGYRO on 4 Leadership Systems

Igor Sfiligoi (General Atomics), Jeff Candy (General Atomics), Emily Belli (General Atomics)

We present the results of an exhaustive performance analysis of the CGYRO code on 4 leadership systems spanning 5 different configurations (2 KNL-based, 1 Skylake-based, and 2 hybrid CPU-GPU architectures). CGYRO is an Eulerian gyrokinetic solver designed and optimized for collisional, electromagnetic, multiscale fusion plasma simulation. It is based on the well-known GYRO code, but redesigned from the ground up to operate efficiently on multicore and GPU-accelerated systems. The gyrokinetic equations specify a 5-dimensional distribution function for each species, with species coupled through both the Maxwell equations and collision operator. For the cross-machine performance analysis, we report and compare timings for 4 computational and 4 communication kernels. This kernel-based breakdown illustrates the strengths and weaknesses of the floating-point and communication architectures of the respective systems. An overview of the physical equations solved, the scalable numerical methods used, and data communication patterns required by each kernel are also given.

Best Poster Finalist: no

Redesigning The Absorbing Boundary Algorithm for Asynchronous High Performance Acoustic Wave Propagation

Rached Abdelkhalak (King Abdullah University of Science and Technology), Kadir Akbudak (King Abdullah University of Science and Technology), Vincent Etienne (Saudi Aramco), Thierry Tonellot (Saudi Aramco)

Exploiting high concurrency, relaxing the synchrony of existing algorithms, and increasing data reuse have immense effect in performance. We integrate the Multicore-optimized Wavefront Diamond (MWD) tiling approach by Malas et al. [SIAM SISC, 2015, ACM Trans. Parallel Comput. 2017], which takes into account the three aforementioned ingredients, into the industrial project codenamed ExaWave framework beside the traditional spatial blocking (SB) technique for stencil computations. However, the fine-grained asynchronous handling of the Convolution Perfectly Matched Layer (CPML) for absorbing boundary conditions turns out to be a challenging open research problem, due to severe inherent data dependencies constraints, which impedes MWD performance impact. We propose techniques of loop fusion to reduce memory traffic and sliding windows to cut down the engendered extra flops, in order to consolidate CPML integration with the overall asynchronous MWD technique. The experimental results on Intel's latest processors show the effectiveness of the proposed techniques.

Best Poster Finalist: no

Capsule Networks for Protein Structure Classification

Dan A. Rosa de Jesus (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Julian Cuevas Paniagua (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Wilson Rivera (Lawrence Berkeley National Laboratory, University of Puerto Rico at Mayaguez), Silvia Crivelli (Lawrence Berkeley National Laboratory)

Capsule Networks have great potential to tackle problems in structural biology because of their attention to hierarchical relationships. This work describes the implementation and application of a capsule network

architecture to the classification of RAS protein family structures on GPU-based computational resources. Our results show that the proposed capsule network trained on 2D and 3D structural encodings can successfully classify HRAS and KRAS structures. The capsule network can also classify a protein-based dataset derived from a PSI-BLAST search on sequences of KRAS and HRAS mutations. Experimental results show an accuracy improvement compared to traditional convolutional networks.

Best Poster Finalist: no

Cross-Layer Group Regularization for Deep Neural Network Pruning
Shuang Gao (Nvidia Corporation), Xin Liu (Nvidia Corporation)

Improving weights sparsity is a common strategy for deep neural network pruning. Most existing methods use regularizations that only consider structural sparsity within an individual layer. In this paper, we propose a cross-layer group regularization taking into account the statistics from multiple layers. For residual networks, we use this approach to align kernel sparsity across layers that are tied to each other through element-wise operations: the ith kernel of these layers are put into one regularization group, they either stay or be removed simultaneously during pruning. In this way, the computational and parameter storage cost could be significantly reduced. Experimental results show that this method does not only improve weights sparsity but also align kernel weights sparsity across related layers. Our method is able to prune ResNet up to 90.4% of parameters and improve runtime by 1.5x speedup, without loss of accuracy.

Best Poster Finalist: yes

Machine Learning for Adaptive Discretization in Massive Multiscale Biomedical Modeling
Changnian Han (Stony Brook University), Prachi Gupta (Stony Brook University), Peng Zhang (Stony Brook University), Danny Bluestein (Stony Brook University), Yuefan Deng (Stony Brook University)

For multiscale problems, traditional time stepping algorithms use a single smallest time stepsize in order to capture the finest details; using this scale leads to a significant waste of computing resources for simulating coarse-grained portion of the problem. To improve computing efficiency for multiscale modeling, we propose a novel state-driven adaptive time stepping (ATS) algorithm to automatically adapt the time stepsizes to the underlying biophysical phenomena at multiple scales. In this, we use a machine-learning based solution framework to classify and label these states for regulating the time stepsizes. We demonstrate the values of our ATS algorithm by assessing the accuracy and efficiency of a multiscale two-platelet aggregation simulation. By comparing with traditional algorithm for this simulation, our ATS algorithm significantly improves the efficiency while maintaining accuracy. Our novel ATS algorithm presents a more efficient framework for solving massive multiscale biomedical problems.

Best Poster Finalist: no

Multi-GPU Accelerated Non-Hydrostatic Numerical Ocean Model with GPUDirect RDMA Transfers
Takateru Yamagishi (Research Organization for Information Science and Technology, Japan), Yoshimasa Matsumura (University of Tokyo), Hiroyasu Hasumi (University of Tokyo)

We have implemented our “kinaco” numerical ocean model on Tokyo University’s Reedbush supercomputer, which utilizes the latest Nvidia Pascal P100 GPUs with GPUDirect technology. We have also optimized the model’s Poisson/Helmholtz solver by adjusting the global memory alignment and thread block configuration, introducing shuffle functions to accelerate the creation of coarse grids and merging

small kernels in the multigrid preconditioner. We also utilize GPUDirect RDMA transfers to improve MPI communication efficiency. By exploiting the GPUs' capabilities, the GPU implementation is now twice as fast as the CPU version, and it shows good weak scalability to multiple GPUs. Most of the GPU kernels are accelerated, and the velocity diagnosis functions in particular are now approximately seven times faster. The performance of inter-node data transfers using a CUDA-aware MPI library with GPUDirect RDMA transfers is comparable to that on CPUs.

Best Poster Finalist: no

A Locality and Memory Congestion-Aware Thread Mapping Method for Modern NUMA Systems
Mulya Agung (Tohoku University), Muhammad Alfian Amrizal (Tohoku University), Ryusuke Egawa (Tohoku University), Hiroyuki Takizawa (Tohoku University)

On modern NUMA systems, the memory congestion problem could degrade performance more than the memory access locality problem because a large number of processor cores in the systems can cause heavy congestion on memory controllers. In this work, we propose a thread mapping method that considers the spatio-temporal communication behavior of multi-threaded applications to improve the locality and to reduce the memory congestion on modern NUMA systems. We evaluate the proposed method using NPB applications on a NUMA system. Experiments show that our proposed method can achieve up to 20% performance improvement compared with locality-based and balance-based methods.

Best Poster Finalist: no

Tuning CFD Applications for Intel Xeon Phi with TAU Commander and ParaTools ThreadSpotter
Izaak B. Beekman (ParaTools Inc), Nicholas Chaimov (ParaTools Inc), Sameer Shende (ParaTools Inc, University of Oregon), Allen D. Malony (ParaTools Inc, University of Oregon), Nicholas Bisek (US Air Force Research Laboratory), Ryan Gosse (US Air Force Research Laboratory), Andrew Wissink (Create AV, US Army)

Tuning and understanding the performance characteristics of computational fluid dynamics (CFD) codes on many-core, NUMA architectures is challenging. One must determine how programming choices impact algorithm performance and how best to utilize the available memory caches, high-bandwidth memory, and inter---and intra---node communication. Once collected, performance data must be translated into actionable code improvements. In addition, performance engineering experiments must be organized and tracked to quantify the benefit of any attempted tuning.

In the poster we present, examine and tune two CFD applications running on the Intel® Xeon Phi™ partition of a Cray® XC 40/50 using TAU Commander and ParaTools ThreadSpotter. TAU Commander implements a streamlined, managed performance engineering workflow and highlights source regions limiting scalability through profiling and aggregate summary statistics. ParaTools ThreadSpotter analyzes an application as it is running and ranks individual performance problems. It provides a report annotating source code lines with actionable recommendations and quantifying performance metrics.

Best Poster Finalist: no

Massively Parallel Stress Chain Characterization for Billion Particle DEM Simulation of Accretionary Prism Formation
Mikito Furuichi (Japan Agency for Marine-Earth Science and Technology), Daisuke Nishiura (Japan Agency

for Marine-Earth Science and Technology), Takane Hori (Japan Agency for Marine-Earth Science and Technology)

Herein, a novel algorithm for characterizing stress chains using a large parallel computer system is presented. Stress chains are important for analyzing the results of large-scale discrete element method (DEM) simulations. However, the general algorithm is difficult to parallelize especially when selecting networks longer than several particles. Therefore, we propose a new parallel algorithm to count the number of particles that are tightly connected, based on iterative operations with nearest-neighbor computations and communications. The new algorithm is examined via a real-scale numerical sandbox experiment using 2.4 billion particles. We successfully compute the stress chains with a reasonable computational cost comparable to the single-step DEM computation time. The visualization of the stress chains from the large-scale DEM simulation result reveals the existence of arcuate stress structures that may control accretionary prism formation, which is an important scientific discovery.

Best Poster Finalist: no

Toward Smoothing Data Movement Between RAM and Storage

Tariq Alturkestani (King Abdullah University of Science and Technology), Thierry Tonellot (Saudi Aramco), Vincent Etienne (Saudi Aramco), Hatem Ltaief (King Abdullah University of Science and Technology)

We propose to design and implement a software framework, which provides a Multilayer Buffer System (MBS) to cache in/out datasets into CPU main memory from/to slower storage media, such as parallel file systems (e.g., Lustre), solid-state drive (e.g., Burst Buffer) or non-volatile RAM. Although MBS scope may be broad in terms of scientific applications, we focus on the RTM application as a proxy for I/O intensive workloads, since reading and writing are ubiquitous operations during the dumping phase of the source wavefield (forward propagation) as well as its retrieval phase (backward propagation) for the image condition calculations.

Best Poster Finalist: no

Interactive HPC Deep Learning with Jupyter Notebooks

Wahid Bhimji (Lawrence Berkeley National Laboratory), Steven Farrell (Lawrence Berkeley National Laboratory), Oliver Evans (Lawrence Berkeley National Laboratory), Matthew Henderson (Lawrence Berkeley National Laboratory), Shreyas Cholia (Lawrence Berkeley National Laboratory), Aaron Vose (Cray Inc), Mr Prabhat (Lawrence Berkeley National Laboratory), Rollin Thomas (Lawrence Berkeley National Laboratory), Richard Shane Canon (Lawrence Berkeley National Laboratory)

Deep learning researchers are increasingly using Jupyter notebooks to implement interactive, reproducible workflows. Such solutions are typically deployed on small-scale (e.g. single server) computing systems. However, as the sizes and complexities of datasets and associated neural network models increase, distributed systems become important for training and evaluating models in a feasible amount of time. In this poster, we describe our work on Jupyter notebook solutions for distributed training and hyper-parameter optimization of deep neural networks on high-performance computing systems.

Best Poster Finalist: no

Fast and Accurate Training of an AI Radiologist

Lucas A. Wilson (Dell EMC), Vineet Gundecha (Dell EMC), Srinivas Varadharajan (Dell EMC), Alex Filby (Dell

EMC), Pei Yang (Dell EMC), Quy Ta (Dell EMC), Valeriu Codreanu (SURFsara), Damian Podareanu (SURFsara), Vikram Saletore (Intel Corporation)

The health care industry is expected to be an early adopter of AI and deep learning to improve patient outcomes, reduce costs, and speed up diagnosis. We have developed models for using AI to diagnose pneumonia, emphysema, and other thoracic pathologies from chest x-rays. Using the Stanford University CheXNet model as inspiration, we explore ways of developing accurate models for this problem with fast parallel training on Zenith, the Intel Xeon-based supercomputer at Dell EMC's HPC and AI Innovation Lab. We explore various network topologies to gain insight into what types of neural networks scale well in parallel and improve training time from days to hours. We then explore transferring this learned knowledge to other radiology subdomains, such as mammography, and whether this leads to better models than developing subdomain models independently.

Best Poster Finalist: no

Full State Quantum Circuit Simulation by Using Lossy Data Compression

Xin-Chuan Wu (University of Chicago, Argonne National Laboratory), Sheng Di (Argonne National Laboratory), Franck Cappello (Argonne National Laboratory), Hal Finkel (Argonne National Laboratory), Yuri Alexeev (Argonne National Laboratory), Frederic T. Chong (University of Chicago)

In order to evaluate, validate, and refine the design of a new quantum algorithm or a quantum computer, researchers and developers need methods to assess their correctness and fidelity. This requires the capabilities of simulation for full quantum state amplitudes. However, the number of quantum state amplitudes increases exponentially with the number of qubits, leading to the exponential growth of the memory requirement. In this work, we present our technique to simulate more qubits than previously reported by using lossy data compression. Our empirical data suggests that we can simulate full state quantum circuits up to 63 qubits with 0.8 petabytes memory.

Best Poster Finalist: no

Enabling Reproducible Microbiome Science through Decentralized Provenance Tracking in QIIME 2

Ahmad Turan Naimey (Northern Arizona University, Pathogen and Microbiome Institute), Christopher Keefe (Northern Arizona University, Pathogen and Microbiome Institute)

In this poster, we demonstrate the ways in which automatic, integrated, decentralized provenance tracking in QIIME 2, a leading microbiome bioinformatics platform, enables reproducible microbiome science. We use sample data from a recent study of arid soil microbiomes (Significant Impacts of Increasing Aridity on the Arid Soil Microbiome; Neilson et al, 2017), to illustrate specific analyses that QIIME 2 supports, and to frame our discussion of the QIIME 2 platform.

QIIME 2 actions yield as outputs artifacts integrating the requested data or visualization with comprehensive data provenance that describes the computational history of that data or visualization, including all methods and parameters involved in its creation. This approach gives users, reviewers, and readers powerful tools for understanding, reproducing, and extending studies. The benefits this approach provides to both the researcher and the scientific community are significant and provide a useful model for research software developers across disciplines.

Best Poster Finalist: no

Optimizing Next Generation Hydrodynamics Code for Exascale Systems

Dana Akhmetova (KTH Royal Institute of Technology), Sumathi Lakshmiranganatha (University of Wyoming), Diptajyoti Mukherjee (Allegheny College), Frederick Oullet (University of Florida), Patrick Payne (Los Alamos National Laboratory), Nicholas Stegmeier (South Dakota State University), Christoph Junghans (Los Alamos National Laboratory), Robert Pavel (Los Alamos National Laboratory), Vinay Ramakrishnaiah (Los Alamos National Laboratory)

Studying continuum dynamics problems computationally can illuminate complex physical phenomena where experimentation is too costly. However, the models used in studying these phenomena usually require intensive calculations, some of which are beyond even the largest supercomputers to date. Emerging high performance computing (HPC) platforms will likely have varied levels of heterogeneity, making hybrid programming with MPI+X essential for achieving optimal performance. This research investigates hybrid programming and unconventional approaches like machine learning for a next generation hydrodynamics code, FleCSALE, in the context of tabular equation of state (EOS). We demonstrate an overall 5x speedup to the code, the use of GPUs to accelerate EOS tabular interpolation, and a proof of concept machine learning approach to EOS.

Best Poster Finalist: no

MPI/OpenMP parallelization of the Fragment Molecular Orbitals Method in GAMESS

Vladimir Mironov (Lomonosov Moscow State University), Yuri Alexeev (Argonne National Laboratory), Dmitri Fedorov (National Institute of Advanced Industrial Science and Technology (AIST))

In this work, we present a novel parallelization strategy for the Fragment Molecular Orbital (FMO) method in the quantum chemistry package GAMESS. The original FMO code has been parallelized only with MPI, which limits scalability of the code on multi-core massively parallel machines. To address this problem, we parallelized FMO with a new hybrid MPI-OpenMP scheme that shows excellent scaling up to 2,048 Intel Xeon Phi nodes (131,072 cores) on Theta supercomputer. MPI-OpenMP code not only scales better compared to MPI code, but also performs up to two times faster and has significantly smaller memory footprint.

Best Poster Finalist: no

Automatic Generation of Mixed-Precision Programs

Logan Moody (Lawrence Livermore National Laboratory, James Madison University), Nathan Pinnow (Lawrence Livermore National Laboratory, Western Washington University), Michael O. Lam (James Madison University, Lawrence Livermore National Laboratory), Harshitha Menon (Lawrence Livermore National Laboratory), Markus Schordan (Lawrence Livermore National Laboratory), G. Scott Lloyd (Lawrence Livermore National Laboratory), Tanzima Islam (Western Washington University)

Floating-point arithmetic is foundational to scientific computing in HPC, and choices about floating-point precision can have a significant effect on the accuracy and speed of HPC codes. Unfortunately, current precision optimization tools require significant user interaction, and few work on the scale of HPC codes due to significant analysis overhead. We propose an automatic search and replacement system that finds the maximum speedup using mixed precision given a required level of accuracy. To achieve this, we integrated three existing analysis tools into a system that requires minimal input from the user. If a speedup is found, our system can provide a ready-to-compile mixed-precision version of the original program.

Best Poster Finalist: no

UPC++ and GASNet-EX: PGAS Support for Exascale Applications and Runtimes

Scott B. Baden (Lawrence Berkeley National Laboratory), Paul H. Hargrove (Lawrence Berkeley National Laboratory), Hadia Ahmed (Lawrence Berkeley National Laboratory), John Bachan (Lawrence Berkeley National Laboratory), Dan Bonachea (Lawrence Berkeley National Laboratory), Steven Hofmeyr (Lawrence Berkeley National Laboratory), Mathias Jacquelin (Lawrence Berkeley National Laboratory), Amir Kamil (Lawrence Berkeley National Laboratory), Brian van Straalen (Lawrence Berkeley National Laboratory)

Lawrence Berkeley National Lab is developing a programming system to support HPC application development using the Partitioned Global Address Space (PGAS) model. This work is driven by the emerging need for adaptive, lightweight communication in irregular applications at exascale. We present an overview of UPC++ and GASNet-EX, including examples and performance results.

GASNet-EX is a portable, high-performance communication library, leveraging hardware support to efficiently implement Active Messages and Remote Memory Access (RMA). UPC++ provides higher-level abstractions appropriate for PGAS programming such as: one-sided communication (RMA), remote procedure call, locality-aware APIs for user-defined distributed objects, and robust support for asynchronous execution to hide latency. Both libraries have been redesigned relative to their predecessors to meet the needs of exascale computing. While both libraries continue to evolve, the system already demonstrates improvements in microbenchmarks and application proxies.

Best Poster Finalist: no

An Efficient SIMD Implementation of Pseudo-Verlet Lists for Neighbor Interactions in Particle-Based Codes

James Willis (Durham University, Institute for Computational Cosmology), Matthieu Schaller (Leiden Observatory), Pedro Gonnet (Google LLC)

In particle-based simulations, neighbour finding (i.e. finding pairs of particles to interact within a given range) is the most time consuming part of the computation. One of the best such algorithms, which can be used for both Molecular Dynamics (MD) and Smoothed Particle Hydrodynamics (SPH) simulations is the pseudo-Verlet list algorithm. The algorithm improves the neighbour finding by reducing the number of spurious pair-wise distance calculations. This algorithm, however, does not vectorize trivially, and hence makes it difficult to exploit SIMD-parallel architectures. On this poster, we present several novel modifications as well as a vectorization strategy for the algorithm which lead to overall speed-ups over the scalar version of the algorithm of 2.21x for the AVX instruction set (SIMD width of 8), 2.41x for AVX2, and 3.65x for AVX-512 (SIMD width of 16).

Best Poster Finalist: no

Understanding Potential Performance Issues Using Resource-Based alongside Time Models

Nan ding (Lawrence Berkeley National Laboratory), Victor W. Lee (Intel Corporation), Wei Xue (Tsinghua University), Weimin Zheng (Tsinghua University)

Numerous challenges and opportunities are introduced by the complexity and enormous code legacy of HPC applications, the diversity of HPC architectures, and the nonlinearity of interactions between applications and HPC systems. To address these issues, we propose the Resource-based Alongside Time

(RAT) modeling method to help to understand the application run-time performance efficiently. First, we use hardware counter-assisted profiling to identify the key kernels and non-scalable kernels in the application. Second, we show how to apply the resource-based profiling into performance models to understand the potential performance issues and predict performance in the regimes of interest to developers and performance analysts. Third, we propose an easy-to-use performance modeling tool for scientists and performance analytics. Our evaluations demonstrate that by only performing a few small-scale profilings, RAT is able to keep the average model error rate around 15% with average performance overheads of 3% in multiple scenarios.

Best Poster Finalist: no

MGRIT Preconditioned Krylov Subspace Method

Ryo Yoda (Kogakuin University), Akihiro Fujii (Kogakuin University), Teruo Tanaka (Kogakuin University)

MGRIT re-discretize the problem with larger time-step width at the coarse-levels, which often cause unstable convergence. We propose a Krylov subspace method with MGRIT preconditioning as a more stable solver. For unstable problems, MGRIT preconditioned Krylov subspace method performed better than MGRIT in terms of the number of iterations. The contributions of the paper are organized as follows. We showed the matrix form of MGRIT operations, and the improvement of eigenvalue or singular-value distribution. We exemplified MGRIT with Krylov subspace method reaching convergence faster than MGRIT.

Best Poster Finalist: no

Enabling Neutrino and Antineutrino Appearance Observation Measurements with HPC Facilities

Norm Buchanan (Colorado State University), Steven Calvez (Colorado State University), Pengfei Ding (Fermi National Accelerator Laboratory), Derek Doyle (Colorado State University), Alex Himmel (Fermi National Accelerator Laboratory), Burt Holzman (Fermi National Accelerator Laboratory), Jim Kowalkowski (Fermi National Accelerator Laboratory), Andrew Norman (Fermi National Accelerator Laboratory), Alex Sousa (University of Cincinnati), Marc Paterno (Fermi National Accelerator Laboratory), Saba Sehrish (Fermi National Accelerator Laboratory), Brandon White (Fermi National Accelerator Laboratory), Christopher Green (Fermi National Accelerator Laboratory)

When fitting to data with low statistics and near physical boundaries, extra measures need to be taken to ensure proper statistical coverage. The method NOvA uses is called the Feldman-Cousins procedure, which entails fitting thousands of independent pseudoexperiments to generate acceptance intervals that are then used to correct our fits. The scale required by the Feldman-Cousins procedure makes it extremely computationally intensive. In past analyses, it has taken up to several weeks to complete, bottlenecking our final results. Here, I present recent work by members of the NOvA experiment and the SciDAC4 collaboration to enable the use of the supercomputing facilities at NERSC to process our Feldman-Cousins corrections over 50x faster, allowing us to perform more studies, increase the precision of our fits, and produce results quickly.

Best Poster Finalist: no

Large Scale Computation of Quantiles Using MELISSA

Alejandro Ribes (EDF Research and Development), Théophile Terraz (French Institute for Research in Computer Science and Automation (INRIA)), Yvan Fournier (EDF Research and Development), Bertrand

looss (EDF Research and Development), Bruno Raffin (French Institute for Research in Computer Science and Automation (INRIA))

Quantiles being order statistics, the classical approach for their computation requires availability of the full sample before ranking it. This approach is not suitable at exascale. Large ensembles would need to gather a prohibitively large amount of data. We propose an iterative approach based on the stochastic quantile algorithm of Robbins-Monro. We rely on the Melissa framework, a file avoiding, adaptive, fault tolerant and elastic framework in order to compute in transit ubiquitous quantiles. Quantiles are updated on-the-fly as soon as the in transit parallel server receives results from one of the running simulations. We run 80,000 fluid dynamics parallel simulations of 6M hexahedra and 100 timespteps. They were executed on up to 4800 cores, avoiding 288 TB of file storage. We produce ubiquitous spatio-temporal maps of quantiles and inter-quantile based intervals.

Best Poster Finalist: no

FlowOS-RM: Disaggregated Resource Management System

Ryousei Takano (National Institute of Advanced Industrial Science and Technology (AIST)), Kuniyasu Suzaki (National Institute of Advanced Industrial Science and Technology (AIST)), Hidetaka Koie (National Institute of Advanced Industrial Science and Technology (AIST))

A traditional data center consists of monolithic-servers is confronted with limitations including lack of operational flexibility, low resource utilization, low maintainability, etc. Resource disaggregation is a promising solution to address the above issues. We propose a concept of disaggregated data center architecture called Flow-in-Cloud (FiC) that enables an existing cluster computer to expand an accelerator pool through a high-speed network. FiC is a shared pool of heterogeneous accelerators such as GPU and FPGA, which are directly connected by a circuit-switched network. From the pool of accelerators, a slice is dynamically configured and provided according to a user request. FlowOS-RM manages the entire FiC resources, and supports execution of a user job on provided slices. This poster demonstrates effective resource sharing on the prototype system using a distributed deep learning application.

Best Poster Finalist: no

Programming the EMU Architecture: Algorithm Design Considerations for Migratory-Threads-Based Systems

Mehmet E. Belviranli (Oak Ridge National Laboratory), Seyong Lee (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory)

The decades-old memory bottleneck problem for data-intensive applications is getting worse as the processor core counts continue to increase. Workloads with sparse memory access characteristics only achieve a fraction of a system's total memory bandwidth. EMU architecture provides a radical approach to the issue by migrating the computational threads to the location where the data resides.

In EMU architecture, data distribution and thread creation strategies play a crucial role in achieving optimal performance in the EMU platform. In this work, we identify several design considerations that need to be taken care of while developing applications for the new architecture and we evaluate their performance effects on the EMU-chick hardware.

Best Poster Finalist: no

OpenACC to FPGA: A Directive-Based High-Level Programming Framework for High-Performance Reconfigurable Computing

Seyong Lee (Oak Ridge National Laboratory), Jacob Lambert (University of Oregon), Jungwon Kim (Oak Ridge National Laboratory), Jeffrey S. Vetter (Oak Ridge National Laboratory), Allen D. Malony (University of Oregon)

Accelerator-based heterogeneous computing has become popular solutions for power-efficient high performance computing (HPC). Along these lines, Field Programmable Gate Arrays (FPGAs) have offered more advantages in terms of performance and energy efficiency for specific workloads than other accelerators. Nevertheless, FPGAs have traditionally suffered several disadvantages limiting their deployment in HPC systems, mostly due to the challenges of programmability and portability. We present a directive-based, high-level programming framework for high-performance reconfigurable computing. It takes a standard, portable OpenACC C program as input and generates a hardware configuration file for execution on FPGAs. We implemented this prototype system in our open-source OpenARC compiler, which uses the Altera OpenCL compiler as its backend. Preliminary evaluation of the proposed framework on an Intel Stratix~V with five OpenACC benchmarks demonstrates that our proposed FPGA-specific compiler optimizations and novel OpenACC pragma extensions assist the compiler in generating more efficient FPGA programs.

Best Poster Finalist: no

Tensor-Optimized Hardware Accelerates Fused Discontinuous Galerkin Simulations

Alexander Breuer (University of California, San Diego), Alexander Heinecke (Intel Corporation), Yifeng Cui (San Diego Supercomputer Center)

In recent years the compute/memory balance of processors has been continuously shifting towards compute. The rise of Deep Learning, based on matrix multiplications, accelerated this path, especially in terms of single precision and lower precision compute. An important research question is if this development can be leveraged for traditional HPC. We demonstrate that a high-order discontinuous Galerkin solver for seismic wave propagation can execute in single precision without loss of modeling accuracy. Additionally, we extended its kernels to support the Intel Knights Mill CPU with 14 TFLOPS of single precision deep-learning performance. This allows us to harvest the hardware's special compute capabilities, even in an application with sparse linear algebra kernels. On cluster-level, Knights Mill can obtain the same application performance as the latest top-bin dual-socket Intel Xeon Platinum nodes, while consuming lower power. Compared to the HPC-focused Knights Landing processor, scenario-dependent speed-ups of up to 1.6× are possible.

Best Poster Finalist: no

AI Matrix – Synthetic Benchmarks for DNN

Wei Wei (Alibaba Inc), Lingjie Xu (Alibaba Inc), Lingling Jin (Alibaba Inc), Wei Zhang (Alibaba Inc), Tianjun Zhang (University of California, Berkeley)

The current AI benchmarks suffer from a number of drawbacks. First, they cannot adapt to the emerging changes of deep learning (DL) algorithms and are fixed once selected. Second, they contain tens to hundreds of applications and have very long running time. Third, they are mainly selected from open sources, which are restricted by copyright and not representable of the proprietary applications. To address these drawbacks, this work firstly proposes a synthetic benchmark framework that generates a small

number of benchmarks that best represent a broad range of applications using their profiled workload characteristics. The synthetic benchmarks can adapt to new DL algorithms by re-profiling new applications and updating itself, greatly reduce number of benchmark tests and running time, and strongly represent DL applications of interests. The framework is validated by using log data profiled from DL models running on Alibaba AI platform, and is representable of real workload characteristics.

Best Poster Finalist: no

Applying the Execution-Cache-Memory Model: Current State of Practice

Georg Hager (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Jan Eitzinger (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Julian Hornich (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Francesco Cremonesi (Swiss Federal Institute of Technology in Lausanne), Christie L. Alappat (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Thoams Roehl (University of Erlangen-Nuremberg), Gerhard Wellein (University of Erlangen-Nuremberg, Erlangen Regional Computing Center)

The ECM (Execution-Cache-Memory) model is an analytic, resource-based performance model for steady-state loop code running on multicore processors. Starting from a machine model, which describes the interaction between the code and the hardware, and static code analysis, it allows an accurate prediction of the runtime of sequential loop code. Together with a scaling assumption, it also gives a performance scaling prediction. This poster summarizes the current state of practice in constructing and applying the ECM model, points out problems and open questions, and applies the model to three new and nontrivial use cases. For the first time, overlap assumptions for all relevant CPU architectures in high performance computing are presented.

Best Poster Finalist: yes

WarpX: Toward Exascale Modeling of Plasma Particle Accelerators

Maxence Thevenet (Lawrence Berkeley National Laboratory), Jean-Luc Vay (Lawrence Berkeley National Laboratory), Ann Almgren (Lawrence Berkeley National Laboratory), John Bell (Lawrence Berkeley National Laboratory), Remi Lehe (Lawrence Berkeley National Laboratory), Andrew Myers (Lawrence Berkeley National Laboratory), Jaehong Park (Lawrence Berkeley National Laboratory), Olga Shapoval (Lawrence Berkeley National Laboratory), Weiqun Zhang (Lawrence Berkeley National Laboratory), Lixin Ge (SLAC National Accelerator Laboratory), Mark Hogan (SLAC National Accelerator Laboratory), Cho Ng (SLAC National Accelerator Laboratory), Dave Grote (Lawrence Livermore National Laboratory)

Turning the current experimental plasma accelerator state-of-the-art from a promising technology into mainstream scientific tools depends critically on high-performance, high-fidelity modeling of complex processes that develop over a wide range of space and time scales. As part of the U.S. Department of Energy's Exascale Computing Project, a team composed of LBNL, SLAC and LLNL researchers is developing a new plasma accelerator simulation tool: WarpX. We will present the code structure and how it articulates around its main components: the new Particle-In-Cell Scalable Application Resource (PICSAR) and the adaptive mesh refinement library AMReX, which are combined with redesigned elements of the Warp code, in the new WarpX software. The status, examples of convergence, scaling and applications will be presented.

Best Poster Finalist: no

Job Simulation for Large-Scale PBS-Based Clusters with the Maui Scheduler

Georg Zitzlsberer (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic), Branislav Jansik (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic), Jan Martinovic (IT4Innovations, Czech Republic; Technical University of Ostrava, Czech Republic)

For large-scale High Performance Computing centers with a wide range of different projects and heterogeneous infrastructures, efficiency is an important consideration. Understanding how compute jobs are scheduled is necessary for improving the job scheduling strategies in order to optimize cluster utilization and job wait times. This increases the importance of a reliable simulation capability, which in turn requires accuracy and comparability with historic workloads from the cluster. Not all job schedulers have a simulation capability, including the Portable Batch System (PBS) resource manager. Hence, PBS based centers have no direct way to simulate changes and optimizations before they are applied to the production system. We propose and discuss how to run job simulations for large-scale PBS based clusters with the Maui Scheduler. For validation purposes, we use historic workloads collected at the IT4Innovations supercomputing center, and demonstrate the viability of our approach.

Best Poster Finalist: no

Script of Scripts Polyglot Notebook and Workflow System

Gao Wang (University of Chicago), Man Chong Leong (Rice University), Bo Peng (University of Texas, MD Anderson Cancer Center)

Computationally intensive disciplines such as computational biology often use tools implemented in different languages and analyze data on high-performance computing systems. Although scientific workflow systems can powerfully execute large-scale data-processing, they are not suitable for ad hoc data analysis. Interactive tools such as Jupyter Notebook can be used for data exploration, but it remains difficult to work with multiple scripting languages and to streamline analysis for batch data processing. To bridge the gap between interactive and workflow systems we developed Script of Scripts (SoS), which consists of a polyglot notebook that supports multiple languages in a single notebook and a workflow engine that provides an intuitive syntax for multi-style workflows and a unified interface for executing tasks on a variety of computing platforms. By allowing the use of SoS workflow engine in a polyglot notebook environment, SoS provides a unified environment for both interactive data analysis and batch data processing.

Best Poster Finalist: no

Enabling High-Level Graph Processing via Dynamic Tasking

Maurizio Drocco (Pacific Northwest National Laboratory), Vito Giovanni Castellana (Pacific Northwest National Laboratory), Marco Minutoli (Pacific Northwest National Laboratory), Antonino Tumeo (Pacific Northwest National Laboratory), John Feo (Pacific Northwest National Laboratory)

Data-intensive computing yields irregular and unbalanced workloads, in particular on large-scale problems running on distributed systems. Task-based runtime systems are commonly exploited to implement higher-level data-centric programming models, promoting multithreading and asynchronous coordination for performance. However, coping with dynamic workloads (e.g., those yielded by large-scale graph processing) is challenging.

In this work, we took an exploratory approach to overcome some typical bottlenecks in tasking systems. In particular, we propose 1. a novel task allocator based on dynamic per-thread allocation and all-to-all

recycling networks, and 2. a reservation-free remote spawning schema, based on receiver-side buffering and back-pressure feedback/sensing to avoid overflows.

As a proof of concept, we implemented the proposed techniques underneath a high-level library of distributed C++ containers. Preliminary experimental evaluation shows consistent scalability, a neat improvement in performance (e.g., 1.5x speedup with respect to the original code over an 8M-nodes graph), and less sensitiveness to parameter tuning.

Best Poster Finalist: no

An Alternative Approach to Teaching Bigdata and Cloud Computing Topics at CS Undergraduate Level

Debzani Deb (Winston-Salem State University), Muztaba Fuad (Winston-Salem State University), Keith Irwin (Winston-Salem State University)

Big data and cloud computing collectively offer a paradigm shift in the way businesses are now acquiring, using and managing information technology. This creates the need for every CS student to be equipped with foundation knowledge in this collective paradigm and to possess some hands-on-experience in deploying and managing big data applications in the cloud. We argue that, for substantial coverage of big data and cloud computing concepts and skills, the relevant topics need to be integrated into multiple core courses of undergraduate CS curriculum rather than creating additional standalone core or elective courses. Our approach to including these topics is to develop learning modules for specific core courses in which their coverage might find an appropriate context. In this poster, three such modules are presented and our classroom experiences during these interventions are documented. Our objective is to share our experience and to receive feedback about our approach.

Best Poster Finalist: no

Binarized ImageNet Inference in 29us

Tong Geng (Boston University, Pacific Northwest National Laboratory), Ang Li (Pacific Northwest National Laboratory), Tianqi Wang (Boston University), Shuaiwen Leon Song (Pacific Northwest National Laboratory), Martin Herbordt (Boston University)

We propose a single-FPGA-based accelerator for ultra-low-latency inference of ImageNet in this work. The design can complete the inference of Binarized AlexNet within 29us with accuracy comparable to other BNN implementations. We achieve this performance with the following contributions: 1. We completely remove floating-point from NL through layer fusion. 2. By using model parallelism rather than data parallelism, we can simultaneously configure all layers and the control flow graphs. Also, the design is flexible enough to achieve nearly perfect load balancing, leading to extremely high resource utilization. 3. All convolution layers are fused and processed in parallel through inter-layer pipelining. Therefore, in case the pipeline is full, latency is just the delay of a single convolution layer plus the FC layers. Note that the dependency pattern of the FC layer prevents it from being integrated into the current pipeline.

Best Poster Finalist: no

Refactoring and Optimizing Multiphysics Combustion Models for Data Parallelism

Christopher Stone (US Department of Defense HPC Modernization Program, Engility Corporation), Alexei Poludnenko (Texas A&M University), Brian Taylor (US Air Force Research Laboratory)

High-fidelity combustion simulations combine high-resolution computational fluid dynamics numerical methods with multi-physics models to capture chemical kinetics and transport processes. These multi-physics models can dominate the computation cost of the simulation. Due to the high cost of combustion simulations and the important role simulations play in propulsion and power research, acceleration methods are needed to reduce the computational time and cost. Multi-physics models within each mesh cell are often independent leading to significant parallelism. However, the iterative algorithms often impede efficient SIMD data parallelism, a key performance feature on modern HPC systems. Refactoring methods for multi-physics models (e.g., kinetics, equation-of-state, diffusion) with nonuniform workloads are demonstrated and benchmarked on a range of platforms (AVX2, KNL, AVX-512). Realized speed-ups over 6x were achieved on KNL and 4x on Skylake (SKX) for complex chemical kinetics models and over 3x on SKX for iterative EOS computations.

Best Poster Finalist: no

Tensorfolding: Improving Convolutional Neural Network Performance with Fused Microkernels
Michael Anderson (Intel Corporation), Evangelos Georganas (Intel Corporation), Sasikanth Avancha (Intel Corporation), Alexander Heinecke (Intel Corporation)

Convolution layers are prevalent in many classes of deep neural networks, including Convolutional Neural Networks (CNNs) which provide state-of-the-art results for tasks like image recognition, neural machine translation and speech recognition. In the recent past, several techniques to improve generalization capabilities of neural networks have been developed; the most prominent and successful is batch normalization. In deep neural network training, the batch normalization layer consists of a memory-bandwidth bound kernel. On the latest Intel Skylake based Xeon processors, a significant portion of execution time is spent in this kernel. By leveraging the CPU's large caches and its latency-optimized execution model, we are able to reduce this kernel's time to a bare minimum while allowing to improve forward pass layer runtimes by 21% compared to an unfused implementation and by 2% compared to a fused implementation.

Best Poster Finalist: no

MATEDOR: MAtrix, TEnsor, and Deep-Learning Optimized Routines
Ahmad Abdelfattah (University of Tennessee), Jack Dongarra (University of Tennessee), Stanimire Tomov (University of Tennessee), Ichitaro Yamazaki (University of Tennessee), Azzam Haidar (Nvidia Corporation)

The MAtrix, TEnsor, and Deep-learning Optimized Routines (MATEDOR) project develops software technologies and standard APIs, along with a sustainable and portable library, for large-scale computations that can be broken down into very small matrix or tensor computations. The main target of MATEDOR is to accelerate applications from important fields that fit this profile, including deep learning, data mining, astrophysics, image and signal processing, hydrodynamics, and more.

MATEDOR is a high-performance numerical library for batched linear algebra subroutines autotuned for modern processor architectures and system designs. The MATEDOR library includes LAPACK-compliant routines that target many small dense problems, tensor, and application-specific operations, e.g., for deep-learning. These routines are constructed as much as possible out of calls to batch BLAS routines and their look-alikes required in sparse computation context.

Best Poster Finalist: no

Distributed Adaptive Radix Tree for Efficient Metadata Search on HPC Systems

Wei Zhang (Texas Tech University), Houjun Tang (Lawrence Berkeley National Laboratory), Suren Byna (Lawrence Berkeley National Laboratory), Yong Chen (Texas Tech University)

Affix-based search allows users to retrieve data without the need to remember all relevant information precisely. While building an inverted index to facilitate efficient affix-based search is a common practice for standalone databases and desktop file systems, they are often insufficient for high-performance computing (HPC) systems due to the massive amount of data and the distributed nature of the storage. In this poster, we present Distributed Adaptive Radix Tree (DART) which enables scalable and efficient affix-based search. DART maintains a balanced keyword distribution and optimizes for excessive keyword requests dynamically at scale. Our evaluation shows that compared with the "full string hashing" used by the commonly adopted DHT approach, DART achieves up to 55x throughput speedup for prefix and suffix search, and has a comparable throughput for exact and infix search. Also, DART maintains balanced keyword distribution and alleviates excessive query workload on popular keywords.

Best Poster Finalist: no

Improving Error-Bounded Lossy Compression for Cosmological N-Body Simulation

Sihuan Li (University of California, Riverside), Sheng Di (Argonne National Laboratory), Xin Liang (University of California, Riverside), Zizhong Chen (University of California, Riverside), Franck Cappello (Argonne National Laboratory)

Cosmological simulations may produce extremely large amount of data, such that its successful run depends on large storage capacity and huge I/O bandwidth, especially in the exascale computing scale. Effective error-bounded lossy compressors with both high compression ratios and low data distortion can significantly reduce the total data size while guaranteeing the data valid for post-analysis. In this poster, we propose a novel, efficient compression model for cosmological N-body simulation framework, by combining the advantages of both space-based compression and time-based compression. The evaluation with a well-known cosmological simulation code shows that our proposed solution can get much higher compression quality than other existing state-of-the-art compressors, with comparable compression/decompression rates.

Best Poster Finalist: no

VeloC: Very Low Overhead Checkpointing System

Bogdan Nicolae (Argonne National Laboratory), Franck Cappello (Argonne National Laboratory), Adam Moody (Lawrence Livermore National Laboratory), Elsa Gonsiorowski (Lawrence Livermore National Laboratory), Kathryn Mohror (Lawrence Livermore National Laboratory)

Checkpointing large amounts of related data concurrently to stable storage is a common I/O pattern of many HPC applications. However, such a pattern frequently leads to I/O bottlenecks that lead to poor scalability and performance. As modern HPC infrastructures continue to evolve, there is a growing gap between compute capacity vs. I/O capabilities. Furthermore, the storage hierarchy is becoming increasingly heterogeneous: in addition to parallel file systems, it comprises burst buffers, key-value stores, deep memory hierarchies at node level, etc. In this context, state of art is insufficient to deal with the diversity of vendor APIs, performance and persistency characteristics. This poster proposes VeloC, a low-overhead checkpointing system specifically designed to address the checkpointing needs of future exascale HPC

systems. VeloC offers a simple API at user level, while employing an advanced multi-level resilience strategy that transparently optimizes the performance and scalability of checkpointing by leveraging heterogeneous storage.

Best Poster Finalist: no

Estimating Molecular Dynamics Chemical Shift with GPUs

Eric F. Wright (University of Delaware), Mauricio H. Ferrato (University of Delaware)

Experimental chemical shifts (CS) from solution and solid state magic-angle-spinning nuclear magnetic resonance spectra provide atomic level data for each amino acid within a protein or complex. However, structure determination of large complexes and assemblies based on NMR data alone remains challenging due to the complexity of the calculations. Here, we present a hardware accelerated strategy for the estimation of NMR chemical-shifts of large macromolecular complexes. We demonstrate the feasibility of our approach in systems of increasing complexity ranging from 2,000 to 11,000,000 atoms.

Best Poster Finalist: no

Using Thrill to Process Scientific Data on HPC

Mariia Karabin (Clemson University, Los Alamos National Laboratory), Xinyu Chen (University of New Mexico), Supreeth Suresh (University of Wyoming), Ivo Jimenez (University of California, Santa Cruz), Li-Ta Lo (Los Alamos National Laboratory), Pascal Grosset (Los Alamos National Laboratory)

With ongoing improvement of computational power and memory capacity, the volume of scientific data keeps growing. To gain insights from vast amounts of data, scientists are starting to look at Big Data processing and analytics tools such as Apache Spark. In this poster, we explore Thrill, a framework for big data computation on HPC clusters that provides an interface similar to systems like Apache Spark but delivers higher performance since it is built on C++ and MPI. Using Thrill, we implemented several analytics operations to post-process and analyze data from plasma physics and molecular dynamics simulations. Those operations were implemented with less programming effort than hand-crafted data processing programs would require and obtained preliminary results which were verified by scientists at LANL.

Best Poster Finalist: no

GPU Acceleration at Scale with OpenPower Platforms in Code_Saturne

Samuel Antao (IBM), Charles Moulinec (Science and Technology Facilities Council, UK), Yvan Fournier (EDF Research and Development), Robert Sawko (IBM), Malgorzata Zimon (IBM), Christopher Thompson (IBM), Alex Skillen (Science and Technology Facilities Council, UK), Juan Uribe (EDF Research and Development), David Emerson (Science and Technology Facilities Council, UK)

Code_Saturne is a widely used computational fluid dynamics software package that uses finite-volume methods to simulate different kinds of flows tailored to tackle multi-bilion-cell unstructured mesh simulations. This class of codes has shown to be challenging to accelerate on GPUs as they consist of many kernels and regular inter-process communication in between. In this poster we show how template pack expansion with CUDA can combine multiple kernels into a single one reducing launching latencies and along with the specification of data environments help reduce host-device communication. We tested these techniques on ORNL Summit Supercomputer based on OpenPOWER platform delivering almost 3x speedup over CPU-only runs on 256 nodes. We also show how the latest generation NVLINK(TM)

interconnect available in POWER9(TM)improves scaling efficiency, enabling consistent GPU acceleration with just 100K-cells per process.

Best Poster Finalist: yes

Large-Message Size Allreduce at Wire Speed for Distributed Deep Learning

Kenji Tanaka (Japan Telegraph and Telephone Corporation), Yuki Arikawa (Japan Telegraph and Telephone Corporation), Kenji Kawai (Japan Telegraph and Telephone Corporation), Junichi Kato (Japan Telegraph and Telephone Corporation), Tsuyoshi Ito (Japan Telegraph and Telephone Corporation), Huy Cu Ngo (Japan Telegraph and Telephone Corporation), Kazutaka Morita (Japan Telegraph and Telephone Corporation), Fumiaki Miura (Japan Telegraph and Telephone Corporation), Takeshi Sakamoto (Japan Telegraph and Telephone Corporation), Satoshi Shigematsu (Japan Telegraph and Telephone Corporation)

In large-scale distributed deep learning, the Allreduce operation for large messages (100 KB or more) is critical for gathering gradients from multiple worker nodes and broadcasting the sum of the gradients to them. When the message is large, the latency in Allreduce operation would make it difficult to take advantage of large-scale distributed deep learning. To reduce the latency, we devised a dataflow architecture with an Allreduce-specific hardware accelerator that performs data aggregation and reduction while data is being transferred. The accelerator is designed to immediately start Allreduce operation before an entire message is recived. Furthermore, Allreduce can be operated at wire speed by vectorizing the gradients and summing them in parallel. Experimental results reveal that the proposed architecture performs Allreduce at 96% of wire speed for a large message. Moreover, the latency of Allreduce is reduced by 65% compared with a state-of-the-art Allreduce method when applied for ResNet-50.

Best Poster Finalist: no

Sol: Transparent Neural Network Acceleration Platform

Nicolas Weber (NEC Laboratories Europe, NEC Corporation)

With the usage of neural networks in a wide range of application fields, the necessity to execute these efficiently on high performance hardware is one of the key problems for artificial intelligence (AI) framework providers. More and more new specialized hardware types and corresponding libraries appear from various manufacturers. The biggest problem arising is that these libraries usually are only supported by a very limited set of AI frameworks and interoperability can become an issue. In this extended abstract we present Sol, a transparent middleware for neural network acceleration. Sol comes with an optimizing compiler engine, allowing to use device specific libraries and to implement own optimizations, that can be leveraged on all target devices. In contrast to other projects Sol explicitly aims at optimizing prediction and training of neural networks.

Best Poster Finalist: no

Detection of Silent Data Corruptions in Smooth Particle Hydrodynamics Simulations

Aurélien Cavelan (University of Basel), Florina M. Ciorba (University of Basel), Ruben M. Cabezón (University of Basel)

Soft errors, such as silent data corruptions (SDCs) hinder the correctness of large-scale scientific applications. Ghost replication (GR) is proposed herein as the first SDCs detector relying on the fast error propagation inherent to applications that employ the smooth particle hydrodynamics (SPH) method. GR

follows a two-steps selective replication scheme. First, an algorithm selects which particles to replicate on a different process. Then, a different algorithm detects SDCs by comparing the data of the selected particles with the data of their ghost. The overhead and scalability of the proposed approach are assessed through a set of strong-scaling experiments conducted on a large HPC system under error-free conditions, using upwards of 3, 000 cores. The results show that GR achieves a recall and precision similar to that of full replication methods, at only a fraction of the cost, with detection rates of 91–99.9%, no false-positives, and an overhead of 1–10%.

Best Poster Finalist: no

DeepSim-HiPAC: Deep Learning High Performance Approximate Calculation for Interactive Design and Prototyping

Ahmed Al-Jarro (Fujitsu Laboratories Ltd), Serban Georgescu (Fujitsu Laboratories Ltd), Yasumoto Tomita (Fujitsu Laboratories Ltd), Kouta Nakashima (Fujitsu Laboratories Ltd)

We present a data-driven technique that can learn from physical-based simulations for the instant prediction of field distribution for 3D objects. Such techniques are extremely useful when considering, for example, computer aided engineering (CAE), where computationally expensive simulations are often required. To accelerate this process, we propose a deep learning framework that can predict the principal field distribution given a 3D object. This work allows us to learn a system's response using simulation data of arbitrarily shaped objects and an auto-encoder inspired deep neural network that maps the input of the 3D object shape to its principal 3D field distribution. We show that our engine, DeepSim-HiPAC, can estimate field distribution for two distinctive applications: micro-magnetics design in computational electromagnetics (CEM) and interactive cooling systems design in computational fluid dynamics (CFD), several orders of magnitude faster, up to 250000X, than the native calculations and at a cost of low error rate.

Best Poster Finalist: no

Top-Down Performance Analysis of Workflow Applications

Christian Herold (Technical University Dresden), Bill Williams (Technical University Dresden)

Scientific simulation frameworks are common to use on HPC systems. They contain parallelized algorithms and provide various solvers for a specific application domain. Usually, engineers execute multiple steps to solve a particular problem which are often distributed over multiple jobs. Finding performance bottlenecks and the causing step in such a complex system is very difficult. Therefore in this work, we present a top-down approach that provides summarized performance metrics for the workflow, jobs and job steps. These summaries guides the user to identify inefficiencies and determine the causing job step. Finally, Vampir can be used for a detailed analysis of the regarding execution in order to resolve the issue.

Best Poster Finalist: no

Convolutional Neural Networks for Coronary Plaque Classification in Intravascular Optical Coherence Tomography (IVOCT) Images

Chaitanya Kolluru (Case Western Reserve University), David Prabhu (Case Western Reserve University), Yanzan Gharaibeh (Case Western Reserve University), David Wilson (Case Western Reserve University), Sanjaya Gajurel (Case Western Reserve University)

Currently, IVOCT is the only imaging technique with the resolution necessary to identify vulnerable thin cap fibro-atheromas (TCFAs). IVOCT also has greater penetration depth in calcified plaques as compared to Intravascular Ultrasound (IVUS). Despite its advantages, IVOCT image interpretation is challenging and time consuming with over 500 images generated in a single pullback. In this poster, we propose a method to automatically classify A-lines in IVOCT images using a convolutional neural network. Conditional random fields were used to clean network predictions across frames. The neural network was trained using a dataset of nearly 4,500 image frames across 48 IVOCT pullbacks. Ten-fold cross validation with held-out pullbacks resulted in a classification accuracy of roughly 76% for fibrocalcific, 84% for fibrolipidic, and 85% for other. Classification results across frames displayed in en face view matched closely to annotated counterparts.

Best Poster Finalist: no

Compiling SIMT Programs on Multi- and Many-Core Processors with Wide Vector Units: A Case Study with CUDA

Hancheng Wu (North Carolina State University), John Ravi (North Carolina State University), Michela Becchi (North Carolina State University)

There has been an increasing interest in SIMT programming tools for multi- and manycore (co)processors with wide vector extensions. In this work, we study the effective implementation of a SIMT programming model (a subset of CUDA C) on Intel platforms with 512-bit vector extensions (hybrid MIMD/SIMD architectures). We first propose a set of compiler techniques to transform programs written using a SIMT programming model into code that leverages both the x86 cores and the vector units of a hybrid MIMD/SIMD architecture, thus providing programmability, high system utilization and portability. We then evaluate the proposed techniques on various hybrid systems using microbenchmarks and real-world applications. Finally, we point out the main challenges in supporting the SIMT model on hybrid systems.

Best Poster Finalist: no

A Massively Parallel Evolutionary Markov Chain Monte Carlo Algorithm for Sampling Complicated Multimodal State SpacesState

Wendy K. Tam Cho (University of Illinois), Yan Liu (University of Illinois)

We develop an Evolutionary Markov Chain Monte Carlo (EMCMC) algorithm for sampling from large multi-modal state spaces. Our algorithm combines the advantages of evolutionary algorithms (EAs) as optimization heuristics and the theoretical convergence properties of Markov Chain Monte Carlo (MCMC) algorithms for sampling from unknown distributions. We harness massive computational power with a parallel EA framework that guides a large set of Markov chains. Our algorithm has applications in many different fields of science. We demonstrate its effectiveness with an application to political redistricting.

Best Poster Finalist: no

MLModelScope: Evaluate and Measure Machine Learning Models within AI Pipelines

Abdul Dakkak (University of Illinois), Cheng Li (University of Illinois), Wen-mei Hwu (University of Illinois), Jinjun Xiong (IBM)

The current landscape of Machine Learning (ML) and Deep Learning (DL) is rife with non-uniform frameworks, models, and system stacks but lacks standard tools to facilitate the evaluation and measurement of models. Due to the absence of such tools, the current practice for evaluating and comparing the benefits of proposed AI innovations (be it hardware or software) on end-to-end AI pipelines is both arduous and error prone — stifling the adoption of the innovations. We propose MLModelScope— a hardware/software agnostic platform to facilitate the evaluation, measurement, and introspection of ML models within AI pipelines. MLModelScope aids application developers in discovering and experimenting with models, data scientists developers in replicating and evaluating for publishing models, and system architects in understanding the performance of AI workloads.

Best Poster Finalist: no

A Compiler Framework for Fixed-Topology Non-Deterministic Finite Automata on SIMD Platforms

Marziyeh Nourian (North Carolina State University), Hancheng Wu (North Carolina State University), Michela Becchi (North Carolina State University)

Automata traversal acceleration has been studied on various parallel platforms. Many existing acceleration methods store finite automata states and transitions in memory. For these designs memory size and bandwidth are the main limiting factors to performance and power efficiency. Many applications, however, require processing several fixed-topology automata that differ only in the symbols associated to the transitions. This property enables the design of alternative, memory-efficient solutions. We target fixed-topology non-deterministic finite automata (NFAs) and propose a memory-efficient design, suitable to SIMD architectures, that embeds the automata topology in code and stores only the transition symbols in memory. We design a compiler that automates deployment of this design on SIMD platforms for a set of fixed-topology NFAs. Our compiler framework performs a combination of platform-agnostic and platform-specific design decisions and optimizations. This poster describes the compiler toolchain and shows the achieved throughput on GPU and Intel SIMD devices.

Best Poster Finalist: no

A Low-Communicaton Method to Solve Poisson's Equation on Locally-Structured Grids

Brian Van Straalen (Lawrence Berkeley National Laboratory), Peter McCorquodale (Lawrence Berkeley National Laboratory), Phil Colella (Lawrence Berkeley National Laboratory), Christos Kavouklis (Lawrence Livermore National Laboratory)

This poster describes a new algorithm, Method of Local Corrections (MLC), and a high-performance implementation for solving Poisson's equation with infinite-domain boundary conditions, on locally-refined nested rectangular grids. The data motion is comparable to that of only a single V-cycle of multigrid, and hence is an order of magnitude smaller than traditional multigrid iteration. The computational kernels are 3D FFTs on small domains. Strong scaling tests on 64 to 4096 cores on NERSC Cori I (Haswell) show over 60% efficiency, and weak scaling by replication tests over 64 to 32768 cores show 92% efficiency on the same platform. We find comparable solve times between HPGMG on a uniform grid with one billion grid points, and MLC on the same number of grid points adaptively distributed. MLC is designed for AMR, able to solve problems with much higher resolution at the finest level than an algorithm on a uniform grid.

Best Poster Finalist: no

Floating-Point Autotuner for CPU-Based Mixed-Precision Applications

Ruidong Gu (North Carolina State University), Paul A. Beata (North Carolina State University), Michela Becchi (North Carolina State University)

In this poster, we present the design and development of an autotuning tool for floating-point code. The goal is to balance accuracy and performance in order to produce an efficient and accurate mixed-precision program. The tuner starts by maximizing accuracy through the use of a high-precision library called CAMPARY and then achieves performance gains under a given error bound by tuning down groups of variables and operations from the higher precision down to double precision. We tested our tuning strategy on a computational fluid dynamics benchmark where we show a 4x speedup relative to the fully high-precision version during the iterative tuning process and achieve an average absolute error of 2.8E-16 compared with the reference solution computed using the 256-bit GNU MPFR extended precision library.

Best Poster Finalist: no

8:30 am - 5:00 pm

ACM Student Research Competition Posters

Session Description: SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

SC18 Research Posters

SC18 ACM Student Research Competition Posters will be on display on Tuesday, Wednesday, Thursday from 8:30am to 5pm in the C2/3/4 Ballroom.

Precomputing Outputs of Hidden Layers to Speed Up Deep Neural Network Training

Sohil Lal Shrestha (University of Texas, Arlington)

Deep learning has recently emerged as a powerful technique for many tasks including image classification. A key bottleneck of deep learning is that the training phase takes a lot of time, since state-of-the-art deep neural networks have millions of parameters and hundreds of hidden layers. The early layers of these deep neural networks have the fewest parameters but take up the most computation.

In this work, we reduce training time by progressively freezing hidden layers, pre-computing their output and excluding them from training in both forward and backward paths in subsequent iterations. We compare this technique to the most closely related approach for speeding up the training process of neural network.

Through experiments on two widely used datasets for image classification, we empirically demonstrate that our approach can yield savings of up to 25% wall-clock time during training with no loss in accuracy.

Designing Shared Address Space MPI Libraries in Many-Core Era

Jahanzeb Maqbool Hashmi (Ohio State University)

The emergence of modern multi-/many-cores has put more emphasis on optimizing intra-node communication. Existing designs in MPI libraries that work on the concept of distributed address spaces incur the overhead of intermediate memory copies to stage the data between processes. This can lead to severe performance degradation especially on emerging many-core architectures like Intel Skylake and IBM OpenPOWER. This work proposes a high-performance "shared address-space"-based MPI point-to-point and collective communication designs using XPMEM. We first characterize the bottlenecks associated with XPMEM based communication and propose new designs for efficient MPI large message communication. Then we propose novel collective designs that are contention-free and offer true zero-copy reduction operations. The proposed designs are evaluated on different multi-/many-core architectures using various micro-benchmarks and application kernels such as MiniAMR and AlexNet DNN training on CNTK. The proposed designs have shown significant performance improvement over state-of-the-art available in MPI libraries.

Modeling Single-Source Shortest Path Algorithm Dynamics to Control Performance and Power Tradeoffs

Sara Karamati (Georgia Institute of Technology), Jeffrey Young (Georgia Institute of Technology), Rich Vuduc (Georgia Institute of Technology)

This work presents a new methodology to improve the performance of parallel algorithms by tuning the amount of available parallelism for execution throughout the runtime. As such, we expose key parameters controlling the performance and parallelism of the algorithm and build a software-based controller with the objective of maintaining the optimal performance. Our controller allows for tuning the level of parallelism executed in each time epoch to optimize for performance while preserving power usage. More specifically, our experimental evaluation focuses on a tunable variation of a GPU-based delta-stepping algorithm for computing the single-source shortest path (SSSP); As the available parallelism for the delta-stepping SSSP is highly irregular and strongly input-dependent, our extensive experiments show that average power can be reduced while average parallelism is increased. This increase in average parallelism provides substantial energy savings, independent of the hardware.

Accelerating DNA Long Read Mapping with Emerging Technologies

Roman Kaplan (Israel Institute of Technology)

DNA sequencing technologies output only short fragments of a genome, called reads. New single-molecule real-time sequencing technologies can produce long reads, up to tens of thousands base pairs, within minutes. However, these long reads may contain up to 15% errors.

To construct a genome from DNA reads, a computationally expensive bioinformatics task, read mapping, is required. Read mapping finds the best-fitting location for each DNA read on a long reference sequence. The length and error rate of long reads poses a challenge for existing read mapping hardware solutions, designed for short reads with low error rates. This work presents a novel DNA read mapping hardware architecture, RASSA. RASSA is a Resistive Approximate Similarity Search Accelerator that exploits charge distribution and parallel in-memory processing to reflect a mismatch count between DNA sequences. RASSA implementation of long read DNA mapping

outperforms state-of-the-art long read mapping solution by 16-77x with comparable accuracy.

SimFS: A Simulation Data Virtualizing File System Interface
Salvatore Di Girolamo (ETH Zurich)

In the big (simulation) data era, simulations often produce petabytes of data to be stored in parallel filesystems or large-scale databases. This data is accessed, often by thousands of analysts and scientists, over the course of decades. However, storing these volumes of data for long time periods of time is not cost effective and, in some cases, practically impossible.

SimFS transparently virtualizes the simulation output, relaxing the storage requirements and re-simulating missing data on-demand. SimFS monitors the analysis access pattern in order to decide (1) which data to store and (2) apply prefetching techniques to improve the analysis performance. SimFS enables a trade-off between on-disk solutions, where all the simulation data is stored on disk, and in-situ, where no data is stored and analyses are always coupled with simulations. Overall, by exploiting the growing computing power and relaxing the storage capacity requirements, SimFS offers a viable path towards exa-scale simulations.

Holistic Root Cause Analysis of Node Failures in Production HPC
Anwesha Das (North Carolina State University)

Production HPC clusters endure failures incurring computation and resource wastage. Despite the presence of various failure detection and prediction schemes, a comprehensive understanding of how nodes fail considering various components and layers of the system is required for sustained resilience. This work performs a holistic root cause diagnosis of node failures using a measurement-driven approach on contemporary system logs that can help vendors and system administrators support exascale resilience.

Our work shows that lead times can be increased by at least 5 times if external subsystem correlations are considered as opposed to considering the events of a specific node in isolation. Moreover, when detecting sensor measurement outliers and interconnect related failures, triggering automated recovery events can exacerbate the situation if recovery is unsuccessful.

Geomancy: Automated Data Placement Optimization
Oceane Bel (University of California, Santa Cruz)

Exascale cloud storage and High-Performance Computing Systems (HPC) deliver unprecedented storage capacity and levels of computing power, though the full potential of these systems remain untapped because of inefficient data placement. Changes in data access patterns can cause a system's performance to suffer. To mitigate performance losses, system designers implement strategies to preemptively place popular data on higher performance nodes. However, these strategies fail to address a diverse userbase whose users individually demand the highest performance, and they must be carefully constructed by an expert of the system.

We propose Geomancy, a tool that reorganizes data to increase I/O throughput. In systems where

heuristic-based improvements may become resource intensive, Geomancy determines new placement policies by training a deep neural network with past workload and system traces. With real workload traces, Geomancy calculated an example placement policy that demonstrated a 49% increase in average throughput compared to the default data layout.

Numerical Simulation of a Flue Instrument with Finite-Difference Lattice Boltzmann Method using GPGPU

Ryoya Tabata (Kyushu Institute of Technology)

In this work, we discuss the possibility of using GPGPU techniques for Aeroacoustic Simulation (especially for flue instruments) with the finite-difference lattice Boltzmann method (FDLBM). Compressible flow simulation has been used in direct aeroacoustic simulation; however, the computational cost is huge due to the requirement of high computational mesh resolution, with small time steps. The lattice Boltzmann method (LBM) has been used as an efficient method for fluid simulation using GPGPU. However, LBM is not accurate enough when applied to some aeroacoustic problems. On the other hand, FDLBM is able to use high-order finite-difference schemes and it has a high arithmetic intensity compared to LBM. We present a performance evaluation of the LBM and FDLBM with several finite-difference schemes on GPU with the roofline model.

Recursive Algebraic Coloring Engine

Christie Louis Alappat (University of Erlangen-Nuremberg)

Many iterative numerical methods for sparse systems and building blocks of sparse linear algebra are difficult to parallelize due to data dependencies. These may be loop-carried dependencies as they occur in solvers like Gauss-Seidel or write conflicts as in symmetric sparse matrix vector. Most of the existing parallelization strategies suffer from low performance on modern hardware, are matrix specific, or require tailored storage formats.

In this poster, we introduce a novel recursive level based algorithm called Recursive Algebraic Coloring (RAC), which achieves high hardware efficiency on modern multi-core architectures and works with simple data formats like compressed row storage. Method is implemented in a library called Recursive Algebraic Coloring Engine (RACE). Thorough performance analysis shows that RACE outperforms traditional multicoloring methods and Intel-MKL implementations with a factor of 2–2.5×. We are on par with Algebraic Block Multicoloring for small matrices, while for large matrices we gain a factor of 1.5–2×.

Accelerating Microscope Data Analysis Using Parallel Computing

John Ravi (North Carolina State University)

Single-Molecule Localization Microscopy (SMLM) techniques deal with the diffraction limit of fluorescent microscopy by localizing single molecules with high precision by stochastically switching molecules on and off. Thousands of camera frames containing subsets of blinking molecules are recorded to obtain a single super-resolution image. Each blinking molecule in each frame is subjected to localization protocols that fit the shape of the blink, assess the quality of the blink and then estimate their center. The algorithm implemented originally in MATLAB and compiled CUDA C, to compute a 'Super Resolution' image took around 6 minutes to process 256x256 pixel images of a moderately

dense dataset. I ported the algorithm to C++ and parallelized it using OpenMP to compute multiple frames in parallel.

Using Integrated Processor Graphics to Accelerate Concurrent Data and Index Structures
Joel Fuentes (University of California, Irvine)

With the advent of computing systems with on-die integrated processor graphics (iGPU), new programming challenges have emerged from these heterogeneous systems. We proposed different data and index structure algorithms that can benefit from the Intel's iGPU architecture and the C for Media (CM) programming model. We aim that certain data structures can run on the iGPU more efficiently than the CPU cores, achieving important performance gains and energy savings. To the best of our knowledge, this is the first attempt to use iGPU for running workloads on concurrent data and index structures. Experimental results show speedups of up to 4x on concurrent data structures and 11x on index structures when comparing with state-of-the-art CPU implementations. Energy savings of up to 300% are also obtained when running these algorithms on iGPU.

PotC: Many-Body Potential Implementations à La Carte
Markus Höhnerbach (RWTH Aachen University)

Molecular dynamics is a valuable investigation tool for simulations in computational chemistry and materials science. In these simulations, atoms move according to so-called potentials, functions that typically describe the distance-dependent interactions between pairs of atoms. For some applications more complex "many-body potentials" are required. From a HPC perspective, such potentials pose challenges: small neighborhoods hinder vectorization; redundant force expressions are tedious and error prone to derive; the implementations are large, runtime-critical, and can not be broken into simpler "kernels". Consequently, only selected many-body potentials receive optimizations---with major effort.

PotC is a DSL for MD potentials and corresponding compiler to generate high-performance implementations. The compiler centralizes optimization knowledge for many-body potentials, and unburdens users from manual optimization and force derivation. The performance optimization of the generated code is still work-in-progress. However, independent of performance, the tool can also be applied when prototyping, testing, or possibly symbolically manipulating potentials.

OoO Instruction Benchmarking Framework on the Back of Dragons
Julian Hammer (University of Erlangen-Nuremberg, RRZE)

In order to construct an accurate instruction execution model for modern out-of-order micro architectures, an accurate description of instruction latency, throughput and concurrency is indispensable. Already existing resources and vendor provided information is neither complete nor detailed enough and sometimes incorrect. We therefore proclaim to deduct this information through runtime instruction benchmarking and present a framework to support such investigations based on LLVM's just-in-time and cross-platform compilation capabilities.

pyasmjit abstracts instructions, operands, and dependency chains, to easily construct the necessary

benchmarks. The synthesized code is interactively compiled and executed using the llvmlite library, which in turn is based on the stable LLVM C-API. pyasmjit offers a command line as well as a programming interface. Unlike other approaches, we do not rely on model specific performance counters and focus on interoperability and automation to support quick modeling of future microarchitectures.

Studying the Impact of Power Capping on MapReduce-Based, Data-Intensive Mini-Applications on Intel KNL and KNM Architectures

Joshua H. Davis (University of Delaware)

In this poster, we quantitatively measure the impacts of data movement on performance in MapReduce-based applications when executed on HPC systems. We leverage the PAPI 'powercap' component to identify ideal conditions for execution of our applications in terms of (1) dataset characteristics (i.e., unique words); (2) HPC system (i.e., KNL and KNM); and (3) implementation of the MapReduce programming model (i.e., with or without combiner optimizations). Results confirm the high energy and runtime costs of data movement, and the benefits of the combiner optimization on these costs.

Accelerating Big Data Processing in the Cloud with Scalable Communication and I/O Schemes

Shashank Gugnani (Ohio State University)

With the advent of cloud computing, the field of Big Data has seen rapid growth. Most cloud providers provide hardware resources such as NVMe SSDs, large memory nodes, and SR-IOV. This opens up the possibility of large-scale high-performance data analytics and provides opportunities to use these resources to develop new designs. Cloud computing provides flexibility, security, and reliability, which are important requirements for Big Data frameworks. However, several important requirements are missing, such as performance, scalability, consistency, and quality of service (QoS). The focus of this work revolves around developing communication and I/O designs and concepts which can provide these requirements to Big Data frameworks. Specifically, we explore new ways to provide QoS and consistency in cloud storage systems, and provide scalable and high-performance communication frameworks.

Accelerating 2D FFT: Exploit GPU Tensor Cores through Mixed-Precision

Xiaohe Cheng (Hong Kong University of Science and Technology), Anumeena Sorna (National Institute of Technology, Tiruchirappalli)

The two-dimensional Fourier Transform is a widely-used computational kernel in many HPC applications. The popular NVIDIA cuFFT library provides a simple interface to compute 2D FFT on GPUs, but it's yet to utilize the recent hardware advancement in half-precision floating-point arithmetic. In this poster, we propose a mixed-precision method to accelerate 2D FFT by exploiting the FP16 matrix-multiply-and-accumulate units on the newest GPU architecture, known as tensor cores. We achieve a balance between speed and accuracy by dynamically splitting the single-precision input data into two half-precision operands and performing FFT separately. We present a CUDA-based implementation that achieves 3-digit more accuracy than half-precision cuFFT. We also demonstrate the stability and scalability of our approach and conclude that it attains high accuracy with tolerable

splitting overhead.

Monitoring Parsl Workflows

Connor Pigg (University of Illinois)

As a Python library that enables workflows, Parsl gives users the ability to define complex workflows in Python and run them in parallel on any computer system. This poster describe the process of adding monitoring to Parsl. Simple and comprehensive monitoring of a workflow's state and resource usage lets users audit, debug, and confirm workflow execution. The poster discusses how Parsl monitors workflow components, what data it captures (task status and resource usage), and the tools it used to do so (Elasticsearch) and to display the information (Kibana). A Kibana dashboard visualizes the collected logs in a real time, with an interactive user interface. This enhanced Parsl allows users the option to monitor the status and resource usage of their workflows via an Elasticsearch database and Kibana dashboard.

Identifying Network Data Transfer Bottlenecks in HPC Systems

Karen Tu (Lawrence Berkeley National Laboratory; University of California, Berkeley)

Improving network data transfer performance is a major factor for improving high performance computing systems. Most studies analyze data transfer and file system IO performance separately, but understanding the relationship between the two is essential for optimizing scheduling and resource management. Intuitively, if data is being transferred to a busy file system the transfer rate would be slower than a file system at regular activity levels.

This study analyzes patterns between file system activity and network throughput for several use cases of file writing and data transfers using a parallel file system. The parameters changed among the use cases were file striping for the file system, and buffer size and parallelism for data transfer. The main bottleneck for network data transfer rate was the number of OSTs the data was striped across. For a large number of OSTs (16 or greater), writing to the file system was the bottleneck.

Dendro-GR: Massively Parallel Simulations of Binary Black Hole Intermediate-Mass-Ratio Inspirals

Milinda Fernando (University of Utah)

We present a portable and highly-scalable algorithm and framework that targets problems in the astrophysics and numerical relativity communities. This framework combines together a parallel octree-refined adaptive mesh with wavelet adaptive multiresolution and a physics module to solve the Einstein equations of general relativity in the BSSN~formulation. The goal of this work is to perform advanced, massively parallel numerical simulations of Intermediate Mass Ratio Inspirals (IMRIs) of binary black holes with mass ratios on the order of 100:1. These studies will be used to generate waveforms for use in LIGO data analysis and to calibrate semi-analytical approximate methods. This advanced framework is designed to easily accommodate many existing algorithms in astrophysics for plasma dynamics and radiation hydrodynamics. We have designed novel algorithms to enable efficient simulations for such experiments and demonstrate excellent weak scalability up to 131K cores on ORNL's Titan for binary mergers for mass ratios up to 100.

Hardware Transactional Persistent Memory

Ellis Giles (Rice University)

This research solves the problem of creating durable transactions in byte-addressable Non-Volatile Memory or Persistent Memory (PM) when using Hardware Transactional Memory (HTM)-based concurrency control. It shows how HTM transactions can be ordered correctly and atomically into PM by the use of a novel software protocol. We exploit the ordering mechanism to design a novel persistence method that decouples HTM concurrency from back-end PM operations. Failure atomicity is achieved using redo logging coupled with aliasing to guard against mistimed cache evictions.

The algorithm uses efficient lock-free mechanisms with bounded static memory requirements and executes on existing Intel based processors. A back-end distributed memory controller alternative provides a hardware implementation choice for catching PM cache evictions. Our approach compares well with standard (volatile) HTM transactions and yields significant gains in latency and throughput over other persistence methods.

Measuring Swampiness: Quantifying Chaos in Large Heterogeneous Data Repositories

Luann C. Jung (Massachusetts Institute of Technology, University of Chicago), Brendan T. Whitaker (Ohio State University, University of Chicago)

As scientific data repositories and filesystems grow in size and complexity, they become increasingly disorganized. The coupling of massive quantities of data with poor organization makes it challenging for scientists to locate and utilize relevant data, thus slowing the process of analyzing data of interest. To address these issues, we explore an automated clustering approach for quantifying the organization of data repositories. Our parallel pipeline processes heterogeneous filetypes (e.g., text and tabular data), automatically clusters files based on content and metadata similarities, and computes a novel "cleanliness" score from the resulting clustering. We demonstrate the generation and accuracy of our cleanliness measure using both synthetic and real datasets, and conclude that it is more consistent than other potential cleanliness measures.

Supercomputing for the Multi-Driver Routing

Zeyang Ye (Stony Brook University)

Supercomputing is essential for routing traffic by providing drivers the optimal routes with minimal traveling distances or time. The unique challenges that require supercomputers to overcome are of multiple folds: numerous drivers, massive simultaneous requests, multiple locations, and needs of instant gratifications, etc. We developed two parallel methods, PSAD and PSAD-M, by using domain decomposition and state-mixing techniques. On the same computing platform with 96 cores, for the same problem, our PSAD methods outperform all published benchmarks by over a hundred times, while improving the solution quality. For the same routing problem on 384 cores, our PSAD-M reduced the elapsed time from the unbearable ten minutes to the reasonable 5 seconds, achieving a record-breaking speedup of 170. By providing instant routing solutions that enable online recommendations, our methods break the bottleneck of the widely adopted offline approaches.

NautDB: Toward a Hybrid Runtime for Processing Compiled Queries

Samuel Grayson (University of Texas, Dallas)

General purpose operating and database system suffer under the load of their generality which makes achieving optimal performance extremely hard, especially on modern hardware. The goal of this research is to integrate, for the first time, specialization techniques from the OS community (hybrid runtimes) and DB community (compiled queries) for high-performance query processing on modern hardware. We envision a system called NautDB, a hybrid dataflow runtime for executing compiled queries. As a first step toward our goal, we evaluate the performance of compiled queries on Linux and run as a Nautilus hybrid runtime using a simple prototype. Our results demonstrate that combining these specialization techniques has transformative potential for building the next generation (distributed) high-performance query processing systems and big data platforms.

Mitigating Performance and Progress Variability in Iterative Asynchronous Algorithms

Justs Zarins (University of Edinburgh)

Large HPC machines are susceptible to irregular performance. Factors like chip manufacturing differences, heat management, and network congestion combine to result in varying execution time for the same code and input sets. Asynchronous algorithms offer a partial solution. In these algorithms, fast workers are not forced to synchronize with slow ones. Instead they continue computing updates, and moving toward the solution, using the latest data available to them, which may have become stale (i.e. a number of iterations out of date compared to the most recent data). While this allows for high computational efficiency, the convergence rate of asynchronous algorithms tends to be lower.

To address this problem, we are using the unique properties of asynchronous algorithms to develop load balancing strategies for iterative asynchronous algorithms in both shared and distributed memory. Our poster shows how our solution attenuates noise, resulting in significant reduction progress imbalance and time-to-solution variability.

Eulerian Algorithms for the Discretization of Plasma Kinetic Equations

James L. Juno (University of Maryland)

While fluid models are common tools in the study of plasmas, many of these systems, whether in astrophysics or the lab, are only weakly collisional and far from equilibrium, making them more accurately described by kinetic equations. Kinetic equations can be computationally demanding due to the need to solve for the distribution function of the particles in a higher dimensional phase space, with position and velocity coordinates. Despite this challenge, the motivation for solving the plasma kinetic equation is large as there remains a vast array of questions concerning collisionless dynamics in real plasma systems. Here we present algorithms in an Eulerian framework for the discretization of the plasma kinetic equation, using a high-order discontinuous Galerkin finite element method due to its arithmetic intensity and parallelizability. Scaling and performance of the algorithm are discussed, and benchmarks of the algorithm are presented as well.

Reception

(back to top)

Sunday, November 11th

Room: Gilley's
6:00 pm - 9:00 pm

Exhibitors' Reception

Session Description: SC18 will host an Exhibitor Party at Gilley's Dallas (1135 S Lamar St, Dallas, TX 75215) for registered exhibitors. The party is SC18's way of thanking exhibitors for their participation and support of the conference. The event will include entertainment along with plentiful food and drinks. Gilley's is a Dallas icon entertainment location with a large dance floor and the famous mechanical bull. It is within walking distance (@ .6 mile) from the Kay Bailey Hutchison Convention Center Dallas. Limited busing will be provided from the Convention Center Lobby Entrance A/B to Gilley's starting at 5:45 pm and back to hotels until 9:00 pm. An Exhibitor badge is required to attend this event. Guest tickets may be purchased in advance at the Registration desk.

Monday, November 12th

Room: Exhibit Hall C, D, E and F
7:00 pm - 9:00 pm

Gala Opening Reception

Gala Opening Reception
SC18 will host its annual Grand Opening Gala in the Exhibit Hall. This will be your first opportunity to see the latest high performance computing, networking, storage, analysis, and research products, services, and innovations. This event is open to all Technical Program, Exhibitors and Students@SC registrants.

Tuesday, November 13th

Room: C2/3/4 Ballroom
5:15 pm - 7:00 pm

Poster Reception

Session Description: The Posters Reception is an opportunity for attendees to interact with poster presenters and includes research and ACM Student Research Competition posters, Doctoral Showcase

posters as well as the Scientific Visualization & Data Analytics Showcase.

Thursday, November 15th

Room: The Perot Museum
7:00 pm - 10:00 pm

Technical Program Reception

Session Description: To thank our Technical Program attendees and to encourage continued interactions, SC18 is hosting a conference reception for all Technical Program attendees at the Perot Museum of Nature and Science (2201 N. Field Street, Dallas, TX 75202), five floors packed with fun, interactive and interesting learning activities. The event will include museum activities along with food and drinks provided by Wolfgang Puck catering. **Please NOTE: This event starts at 7:00 pm.** This year's event features the wonderful exhibits in the Perot museum. With 180,000 square feet of space, standing 170 feet tall (equivalent to an average 14-story building) the revolutionary Perot Museum extends beyond the typical "museum" perception. The extraordinary building serves as a living science lesson, offering provocative illustrations of engineering, technology and conservation. A Tech Program badge, event ticket, and government-issued photo ID are required to attend this event. It is about a mile from the Kay Bailey Hutchison Convention Center Dallas. Busing will be provided from the Convention Center Lobby Entrance A/B to the Perot Museum starting at 6:45pm and back to hotels until 10:00pm. Note, you will need your badge and your event ticket to get on the shuttle.

SC19

(back to top)

Tuesday, November 13th

Room: Lobby E
10:00 am - 5:00 pm

SC19 Preview Booth

Wednesday, November 14th

Room: Lobby E
10:00 am - 5:00 pm

SC19 Preview Booth

Thursday, November 15th

Room: Exhibit Hall B
8:30 am - 8:35 am

SC19 Conference Preview

Session Description: A preview of SC19 by the Conference Chair Michela Taufer

Room: Lobby E
10:00 am - 3:00 pm

SC19 Preview Booth

Scientific Visualization & Data Analytics Showcase

(back to top)

Tuesday, November 13th

Room: Lower Lobby C
8:30 am - 5:00 pm

Scientific Visualization & Data Analytics Showcase Posters

Room: Lower Lobby C
5:15 pm - 7:00 pm

Scientific Visualization & Data Analytics Showcase Reception

Wednesday, November 14th

Room: Lower Lobby C
8:30 am - 5:00 pm

Scientific Visualization & Data Analytics Showcase Posters

Room: D220
10:30 am - 12:00 pm

Scientific Visualization & Data Analytics Showcase

Visualizing Outbursts of Massive Stars
Joseph A. Insley (Argonne National Laboratory, Northern Illinois University), Yan-Fei Jiang (University of California, Santa Barbara), Matteo Cantiello (Flatiron Institute), Lars Bildsten (University of California, Santa Barbara), Eliot Quataert (University of California, Berkeley), Omer Blaes (University of California, Santa Barbara), James Stone (Princeton University)
Massive stars play an important role in many astrophysical environments such as star formation and the structure of the interstellar medium in galaxies. However, the structures and mass loss of massive stars, which are crucial to understand the evolution and fate of massive stars, are still mysteries. Global radiation hydrodynamic simulations of an 80 solar mass star envelope were performed on the Argonne Leadership Computing Facility's supercomputer, Mira, to find the answers. In this work, we present visualizations of the data produced in these simulations, which show the important role that helium opacity plays in outbursts from massive stars.

Visualization of Droplet Dynamics in Cloud Turbulence

Bipin Kumar (Indian Institute of Tropical Meteorology), Matt Rehme (National Center for Atmospheric Research), Neethi Suresh (Indian Institute of Tropical Meteorology)

The study of droplet dynamics is a crucial part of cloud physics and involves investigation of each and every droplet's properties. It is difficult to carry out such investigations using field observations. Small scale simulation is one method to study such phenomena, and visualization of these processes provides a deep and quick understanding. This work depicts data obtained from Direct Numerical Simulation (DNS) of entrainment and mixing processes at a cloud's edge, which affect the droplet dynamics due to evaporation and condensation. This simulation contains coupled Eulerian and Lagrangian frames. Animations are created for both Eulerian grid data and Lagrangian droplet movement. Scientific visualization provides a way to examine these turbulent properties in a particular part of a cloud and learn about droplet evolution and mixing phenomena in such highly turbulent areas.

Arctic Ocean-Sea Ice Interactions

Greg Foss (University of Texas, Texas Advanced Computing Center), An Nguyen (University of Texas, Institute for Computational Engineering and Sciences), Arash Bigdeli (University of Texas, Institute for Computational Engineering and Sciences), Victor Ocaña (University of Texas, Institute for Computational Engineering and Sciences), Briana Bradshaw (University of Texas, Texas Advanced Computing Center), Patrick Heimbach (University of Texas, Institute for Computational Engineering and Sciences)

The Arctic Ocean, the smallest and shallowest of the five major oceans, is a unique physical and ecological system. With the recent shifts in global climate, which are amplified in the Arctic, this system is undergoing profound changes. Scientists are working to document these changes to provide a broad assessment of their local and global impact. The hostile environment makes comprehensive measurements challenging, calling for simulation-based science to support quantitative understanding. A critical element of simulation is visualizing the complex time-evolving three-dimensional ocean state. The showcased animation, created at the Texas Advanced Computing Center, University of Texas at Austin (UT), visualizes results of a high-resolution data-constrained numerical simulation of the circulation of the Arctic Ocean. Different view angles and zooms highlight emergent features, key to understanding some of the Arctic Ocean's most important processes.

The visualization serves as a public-outreach component of an NSF-funded project aimed at understanding and quantifying the Arctic ocean-sea ice mean state and its changes in response to the Earth's recent warming. The research is carried out at The Institute for Computational Engineering and Sciences, the Institute for Geophysics, and the Jackson School of Geosciences, UT. This paper describes briefly the science behind the simulation, the HPC requirements for running the high-resolution model simulation, and the iterative and evolving process of creating the animation. The animation is currently being shown at the exhibition "Exploring the Arctic Ocean", which runs at the UT Visual Arts Center in Austin through the fall 2018 semester.

The First Water in the Universe

Brandon Wiggins (Southern Utah University), Francesca Samsel (University of Texas), Kristin Hoch (Los Alamos National Laboratory), Greg Abram (University of Texas, Texas Advanced Computing Center), Alex Gagliano (University of Illinois), Joseph Smidt (Los Alamos National Laboratory)

Recent work in the chemistry of molecular clouds in the early universe has found reactions that may create ancient water molecules that predate our Sun. For these to take place, however, particular conditions of molecular density and temperature must exist. Water molecules will be destroyed by

either high temperatures or when densities are insufficient to shield them from UV rays. Water molecules will only be created and persist in areas of the universe in which these properties are suitable, and will do so at different rates depending on local conditions.

We present a combined a high resolution hydrodynamics simulation of the early universe closely coupled with this hydroxyl and water-producing chemistry model to determine how water molecules would be created and distributed in space and time in the early universe. By comparing these simulation results to astronomical observations we can verify both the hydrodynamic and chemical model of ancient water formation.

This work is enabled by the computational power of today's supercomputers and simulation technology. The complexity of the chemistry model is significantly higher than that of simple hydrodynamics, making this a computationally intensive model. Vast difference in scale of the physics involved, from the cosmological scale of the universe through the stellar scale of stars and novae to the molecular scale of chemical reactions requires that adaptive mesh refinement (AMR) techniques be used to provide resolution that varies as demanded by the physics. The visualizations presented herein will show the dynamics of the simulation as it evolves over time.

Volume Renderings of Sheared Thermal Convection

Jean M. Favre (Swiss National Supercomputing Centre), Alexander Blass (University of Twente)

Oceans play a big role in the nature of our planet. About 70% of our earth is covered by water. Strong currents are transporting warm water around the world and therefore not only make life possible, but also allow us to harvest its power producing energy. But humanity tends to easily forget that oceans also yield a much more deadly side. Floods and tsunamis can easily annihilate whole cities and destroy life in seconds. The earth's climate system is also very much linked to the circulation in the ocean due to its large coverage of the earth's surface. Deep ocean currents can be simulated by means of wall-bounded turbulent flow simulations. We present visualizations of sheared thermal convection in very large scale numerical simulations. The visualizations are based on volume renderings of the temperature field. To address the needs of supercomputer users with different hardware and software resources, we evaluate different implementations supported in the ParaView environment: two GPU-based solutions with Kitware's native volume mapper or NVIDIA's IndeX library, and a software-only OSPRay-based implementation.

Programmable Interactive Visualization of a Core-Collapse Supernova Simulation

Roland Haas (University of Illinois, National Center for Supercomputing Applications), Philipp Mösta (University of California, Berkeley), Mahendra Roopa (Nvidia Corporation), Alexander Kuhn (Nvidia Corporation), Marc Nienhaus (Nvidia Corporation)

Core-collapse supernovae (CCSN) are the birth places of neutron stars and black holes. They liberate the ashes of stellar evolution, seeding the interstellar gas with the elements from which planets form and life is made. Despite their importance for much of astrophysics, our understanding of the supernova explosion mechanism and its dependence on progenitor star properties, is still incomplete. In addition, CCSN simulations generate terabytes of output posing additional challenges regarding the effective analysis of the simulated physical processes. To address those challenges, we present Nvidia IndeX as a scalable framework to visualize the simulation output. The framework features a streaming-based architecture to interactively explore simulation results in distributed multi-GPU environments. We demonstrate how to customize specialized sampling programs for volume and surface rendering to cover specific analysis questions of CCSN simulations. This provides an extensive level of control over the visualization while efficiently using available resources to achieve high-levels of performance

and visual accuracy.

Thursday, November 15th

Room: Lower Lobby C
8:30 am - 5:00 pm

Scientific Visualization & Data Analytics Showcase Posters

Student Cluster Competition

(back to top)

Monday, November 12th

Room: Exhibit Hall F
7:30 pm - 8:00 pm

Student Cluster Competition Kick-Off

Student Cluster Competition Kick-Off
Come to the back of Exhibit Hall F to help kick off the Student Cluster Competition and cheer on 16 student teams competing in this real-time, non-stop, 48-hour challenge to administer a computational cluster and demonstrate the greatest sustained performance across a series of applications. Teams selected for this year's competition come from universities in the United States, Germany, Taiwan, Poland, China, Singapore, Indonesia, and Australia. The teams are using big iron hardware solutions limited by a 3 KW power cap to execute a sampling of HPC applications.

Tuesday, November 13th

Room: Exhibit Hall F
10:00 am - 6:00 pm

Student Cluster Competition

Student Cluster Competition
Come to the back of Exhibit Hall F to see how 16 teams from around the world are progressing in the Student Cluster Competition.

Wednesday, November 14th

Room: Exhibit Hall F
10:00 am - 6:00 pm

Student Cluster Competition

Student Cluster Competition
Come to the back of Exhibit Hall F to see how 16 teams from around the world are progressing in the Student Cluster Competition.

Thursday, November 15th

Room: Exhibit Hall F
10:00 am - 3:00 pm

Student Cluster Competition

Student Cluster Competition
Come to the back of Exhibit Hall F to see how 16 teams from around the world are progressing in the Student Cluster Competition.

Student Job/Opportunity Fair

(back to top)

Wednesday, November 14th

Room: D163/165/170/172
10:00 am - 3:00 pm

Student/Postdoc Job Fair

Student/Postdoc Job Fair
All students attending SC18 are invited to attend the Student/Postdoc Job Fair. Hiring organizations will be present from 10am-3pm. Students may stop by at any time.

Students@SC Event

(back to top)

Sunday, November 11th

Room: D227
8:30 am - 9:45 am

Students@SC Keynote: Livin' on the Edge: Thoughts on Careers in High Performance Computing

Session Description: Bruce Hendrickson, the associate director for Computation at Lawrence Livermore National Laboratory, will give his thoughts on careers in HPC.

Students@SC Keynote: Livin' on the Edge: Thoughts on Careers in High Performance Computing
Bruce Hendrickson (Lawrence Livermore National Laboratory)
High performance computing has always been a discipline at the leading edge of the computing ecosystem. By its very nature, HPC involves pushing technology to its limits and making new things possible. While exciting and motivating, these opportunities also involve stress and risk. Sometimes it's hard to tell the difference between being on the "cutting edge" and being on the "edge of the cliff"! Edgy even in the calmest of times, HPC is on the cusp of major technology and application disruptions that will require fresh approaches to nearly everything. In this talk I will share thoughts and advice on career paths in this highly dynamic discipline.

Room: D227
10:30 am - 12:00 pm

Students@SC: Careers in Industry, Research Labs, and Academia

Session Description: This panel discussion will highlight a variety of career options and career paths that are open to students. Panelists will share their experiences in developing their careers in several different areas of HPC.

Students@SC: Careers in Industry, Research Labs, and Academia *Allison Baker (National Center for Atmospheric Research), Ron Minnich (Google LLC), Kathryn Mohror (Lawrence Livermore National Laboratory), Philip C. Roth (Oak Ridge National Laboratory), Tyler Simon (University of Maryland, Baltimore County), Sameer Shende (University of Oregon), Michelle Strout (University of Arizona)*

Room: A3 Ballroom
1:30 pm - 3:00 pm

Students@SC: Resume Workshop

Session Description: Students will have an opportunity to consult with professionals from industry, national labs, and academia to get feedback on their résumés and curriculum vitae in advance of the student job fair.

Room: D227
3:30 pm - 5:00 pm

Students@SC: Making the Best of Your HPC Education

Session Description: This panel discussion will highlight a variety of skills students need to gain through formal and informal education to prepare for their future careers in HPC. Panelists from different areas of HPC and types of employment will share their experiences.

Students@SC: Making the Best of Your HPC Education *Rebecca Hartman-Baker (Lawrence Berkeley National Laboratory), Elisa Heymann (University of Wisconsin, Autonomous University of Barcelona), Jesmin Jahan Tithi (Intel Corporation), Richard (Rich) Vuduc (Georgia Institute of Technology), Kenneth Weiss (Lawrence Livermore National Laboratory), Xinghui Zhao (Washington State University)*

Monday, November 12th

Room: D227
8:30 am - 9:45 am

Students@SC Keynote: The Computing Hidden in Everyday Things

Session Description: Thomas Lange will talk about the role of HPC in manufacturing things we use every day.

Students@SC Keynote: The Computing Hidden in Everyday Things *Thomas Lange (retired)*
Prosperity in commerce has always leveraged the travel, transport, and information flow of that era. Transitions from Water (rivers and seas) to Railroads to Interstate Highways to Airplanes and the Internet have resulted in the rise of cities like Chicago and the rise and decline of enterprises from Sears to Walmart to Amazon. Along the way, inventions from catalogues, to a standard time led the way for computing based logistics, and internet ordering. Today, behind our commerce transactions is computing that many suspect. However, what is less well known is the computing behind the manufacturing of the stuff we order.

We will go through examples of how engineering and science new third leg (added to theory and

experiment) are making innovation in everyday products possible. These examples will include Computational Chemistry, Finite Element Analysis, Computational Fluid Dynamics, and even some very large 'non-physics … e.g. time, probability and money' problems that have outgrown laptops and desktops. Problems that were worked on exclusively by empirical correlation to experiments are being solved by direct solution using HPC.

Room: A3 Ballroom
10:30 am - 12:15 pm

Students@SC: Resume Workshop

Session Description: Students will have an opportunity to consult with professionals from industry, national labs, and academia to get feedback on their résumés and curriculum vitae in advance of the student job fair.

Room: D227
1:30 pm - 3:00 pm

Students@SC: HPC Research

Session Description: Four talks on current research in HPC. Compilers, tools, programming models, and extreme heterogeneity.

Parallel Programming Models for the Extreme Scale Era *Laxmikant (Sanjay) Kale (University of Illinois)*

Parallel programming has always been challenging. In addition to the traditional challenges of (sequential) programming, parallel programming has to deal with complex issues of performance such as load imbalances, long critical paths, and communication overheads, as well as correctness issues arising from race conditions. But with the current and upcoming era in parallel computing the challenges have further increased. There are hardware issues of heterogeneity as well as static and dynamic performance variability, and application-induced issues of multi-module multi-physics programming, adaptive refinements, and strong-scaling. I will explain the programming challenges these issues create, and describe some programming concepts and abstractions that are useful in dealing with them. I will also briefly discuss some of the non-traditional parallel programming approaches, in addition to the dominant MPI-OpenMP paradigm. The objective of this brief talk will be to convey some of the excitement of this research area that I am passionate about, and to provide pointers to research that you may follow later.

Software Development Tools for HPC and at Scale *Martin Schulz (Technical University Munich)*
Debugging and performance optimization are critical steps in the software development process for HPC applications and need to be supported by appropriate tools. Compared to comparable standard tools, however, HPC tools need to fulfill one additional requirement: they need to scale themselves, i.e., work well with applications running at large scale. In this talk, I will highlight the challenges in scaling

tools, both for debugging and performance analysis, and will discuss several examples. In particular, I will focus on the use of intuitive visualizations for performance analysis and how it can help to expose and eliminate bottlenecks.

A Renaissance for Domain-Specific Languages, Compilers and Code Generators for HPC and Big Data *Mary Hall (University of Utah)*
Today's HPC architectures are diverse and complex. Achieving high performance often requires low-level, architecture-specific code that is difficult to derive and does not port across architectures. It seems that writing such code has reached a pain threshold for programmers, and this has motivated a true Renaissance in the role of programming system technology (languages, compilers and code generators) for mapping high-level code specifications to high-performance, architecture-specific implementations. This talk will highlight the role programming system technology in achieving performance portability from high-level code specifications.

Opportunities for Extreme Heterogeneity in High Performance Architectures *Maya Gokhale (Lawrence Livermore National Laboratory)*
HPC architectures now provide special-purpose accelerated processing units to augment general purpose compute cores. In the future it is expected that architectural diversity will greatly increase, with heterogeneous components appearing at many levels of granularity. In this talk, I will discuss opportunities for including heterogeneous elements in the CPU, memory, and interconnect, ultimately resulting in HPC systems as diverse in in architectural components as the modern cell phone.

Room: A2 Ballroom
3:30 pm - 4:10 pm

Building Lasting and Effective Mentoring Relationships

Session Description: Building a successful career in HPC requires developing and maintaining important connections with others in the field. One of these important connections is the Mentor/Protege relationship. Mentoring relationships can be invaluable to an individual's personal and professional growth; providing coaching, counseling and exposure to the challenges in the field. This joint session between the ECP and Students@SC programs will provide guidance on fostering these mentoring relationships. Specifically, this session will aid in: outlining the benefits of the mentor/protege relationship, identifying a lasting mentor, maintaining a mentoring relationship, and finding opportunities to give back. This sessions will also provide an overview of the week-long mentoring initiatives and plans for continued engagement.

Mentor-Protégé Informational Session *Kurt Ferreira (Sandia National Laboratories)*
This session will include a panel of speakers with a history of mentoring. Topics will include the following: identifying a lasting mentor and areas in which you need mentoring, initiative needed by a protégé to maintain a mentoring relationship, finding opportunities to give back and start mentoring now. It will also include an overview of our conference week long mentoring initiatives and plans for continued engagement.

3:30 pm - 4:10 pm

Students@SC Mentoring Panel

Session Description: This joint informational session between the SC Early Career and Students@SC programs will provide practical guidance on fostering and maintaining mentoring relationships. Specifically, this session will aid in: outlining the benefits of the mentor/protege relationship, identifying a lasting mentor, maintaining a mentoring relationship, and finding opportunities to give back by becoming a mentor. This session will also provide attendees the opportunity to ask mentoring related questions from an experienced and diverse panel of invited guests.

Wednesday, November 14th

Room: D163/165/170/172
10:00 am - 3:00 pm

Student/Postdoc Job Fair

Student/Postdoc Job Fair
All students attending SC18 are invited to attend the Student/Postdoc Job Fair. Hiring organizations will be present from 10am-3pm. Students may stop by at any time.

Test of Time

(back to top)

Tuesday, November 13th

Room: Exhibit Hall B
1:30 pm - 3:00 pm

Test of Time Award Presentation

The Parallel Hashed Oct-Tree Algorithm Revisited
Mike Warren (Descartes Labs Inc), John Salmon (D.E. Shaw Research)

Algorithms such as fast N-body methods, coupled with the extraordinary increase in computational power over the past three decades, have played a major part in the process of understanding complex physical systems. In this talk, we provide some context and history of gravitational N-body simulations and discuss the evolution of the parallel hashed oct-tree N-body algorithm and the problems it has been applied to over the years since its publication. We touch on concurrent evolution in hardware, such as Beowulf clusters and the Cloud as well as physical applications in cosmology, galaxy formation, supernovae, hydrodynamics and planet formation.

Tutorial

(back to top)

Sunday, November 11th

Room: C140
8:30 am - 12:00 pm

High Performance Computing (HPC) Data Center Planning and TCO: A Case Study and Roadmap

Anna Maria Bailey (Lawrence Livermore National Laboratory, Energy Efficient HPC Working Group), Michael Thomas (Syska Hennessy Group), David Lambiaso (Environmental Systems Design (ESD))

Leveraging the "Data Center Planning and Design for HPC Folks" Tutorials presented at both SC16 and SC17, this track not only provides best practices and lessons learned gleaned from multiple HPC data center infrastructure capacity planning, total cost of ownership (TCO) analysis and business case justification initiatives over the last 15 years but, significantly, provides a deeper-dive case study and roadmap around Lawrence Livermore National Laboratory's (LLNL) methodology to accommodate future HPC systems. The presenters understand that the HPC data center mechanical, electrical, and structural infrastructure can be a significant enabler or roadblock to the timely deployment and optimized performance of HPC systems. We are encouraged to see more TCO/business case dialogue at SC17 and in the HPC community than in years past and will share firsthand experience about how participants can develop/enhance their organization's HPC facility infrastructure to match their deployment methodology. Topics covered include mission statement and business case development, key stakeholder engagement, existing conditions survey, future-state requirements development, gap analysis, alternative options analysis, budgeting, reporting, and approvals. This tutorial will improve both your technical knowledge and financial confidence to navigate this process and drive successful outcomes.

Room: C143/149
8:30 am - 12:00 pm

Secure Coding Practices and Automated Assessment Tools

Elisa Heymann (University of Wisconsin), Barton Miller (University of Wisconsin)

High performance computing increasingly involves the development and deployment of services to access resources for computation, communication, data, and analytics. Unique to the HPC field is the large amount of software that we develop to drive these services. These services must assure data integrity and availability, while providing access to a global scientific and engineering community.

Securing your network is not enough. Every service that you deploy is a window into your data center from the outside world, and a window that could be exploited by an attacker.

This tutorial is relevant to anyone wanting to learn about minimizing security flaws in the software they develop or manage. You will learn skills critical for software developers and analysts concerned with security.

Software assurance tools – tools that scan the source or binary code of a program to find weaknesses – are the first line of defense in assessing the security of a software project. These tools can catch flaws in a program that affect both the correctness and safety of the code. This tutorial is also relevant to anyone wanting to learn how to use these tools.

To prepare for this tutorial attendees should follow the instructions available at http://www.cs.wisc.edu/mist/Tools.pdf

Room: C144
8:30 am - 12:00 pm

Introduction to Kubernetes

Robert T. Killen (University of Michigan), Scott Paschke (University of Michigan)

Containers have shifted the way applications are packaged and delivered. Their use in data science and machine learning is skyrocketing with the beneficial side effect of enabling reproducible research. This rise in use has necessitated the need to explore and adopt better container-centric orchestration tools. Of these tools, Kubernetes - an open-source container platform born within Google -- has become the de facto standard.

Kubernetes' API-driven, highly extensible design has lead to its adoption by numerous vendors and projects. IBM, JupyterHub, Spark, Pachyderm, kubeflow, and many other tools boast native methods of integration.

The aim of this half-day tutorial is to introduce those researchers and sys admins who may already be familiar with container concepts to the architecture and fundamental concepts of Kubernetes. Attendees will explore these concepts through a series of hands-on exercises and leave with the leg-up in continuing their container education, and gain a better understanding of how Kubernetes may be used for research applications.

Attendees should come prepared with a system capable of running a 2 core, 2GB RAM Virtual Machine, if possible come with the following software pre-installed: git, kubectl, minikube, virtualbox (or other minikube compatible hypervisor). For install instructions, please see the guide: https://goo.gl/Nuvb17

Room: C146
8:30 am - 12:00 pm

InfiniBand, Omni-Path, and High-Speed Ethernet for Beginners

Dhabaleswar Panda (Ohio State University), Hari Subramoni (Ohio State University), Sourav Chakraborty (Ohio State University)

InfiniBand (IB), Omni-Path, and High-Speed Ethernet (HSE) technologies are generating a lot of excitement toward building next generation High-End Computing (HEC) systems including clusters, datacenters, file systems, storage, cloud computing, and Big Data (Hadoop, Spark, HBase, and Memcached) environments. RDMA over Converged Enhanced Ethernet (RoCE) technology is also being widely deployed. This tutorial will provide an overview of these emerging technologies, their offered architectural features, their current market standing, and their suitability for designing HEC systems. It will start with a brief overview of IB, Omni-Path, and HSE. An in-depth overview of the architectural features of IB, Omni-Path, and HSE (including iWARP and RoCE), their similarities and differences, and the associated protocols will be presented. Next, an overview of the OpenFabrics stack which encapsulates IB, HSE, and RoCE (v1/v2) in a unified manner will be presented. An overview of the Libfabrics stack will also be provided. Hardware/software solutions and the market trends behind IB, Omni-Path, HSE, and RoCE will be highlighted. Finally, sample performance numbers of these technologies and protocols for different environments will be presented.

Room: C147
8:30 am - 12:00 pm

Introduction to Quantum Computing

Scott Pakin (Los Alamos National Laboratory), Eleanor G. Rieffel (NASA Ames Research Center)

Quantum computing offers the potential to revolutionize high-performance computing by providing a means to solve certain computational problems asymptotically faster than any classical computer. Relatively recently, quantum computing has advanced from merely a theoretical possibility to engineered reality, including commercial entities offering early prototype quantum processors, both special-purpose quantum annealers and general-purpose gate-model processors. The media has been showcasing each new development and implicitly conveying the message that quantum-computing ubiquity is nigh. Here, we will respond to this hype and provide an overview of the exciting but still early state of the field.

In this tutorial, we introduce participants to the computational model that gives quantum computing its immense computational power. We examine the thought processes that programmers need to map problems onto the two dominant quantum-computing architectures. And we discuss the current hardware and algorithmic challenges that must be overcome before quantum computing becomes a component of the HPC developer's repertoire.

Room: C142
8:30 am - 5:00 pm

Parallel Computing 101

Quentin F. Stout (University of Michigan), Christiane Jablonowski (University of Michigan)

This tutorial provides a comprehensive overview of parallel computing, emphasizing those aspects most relevant to the user. It is suitable for new users, managers, students, and anyone seeking an overview of parallel computing. It discusses software and hardware/software interaction, with an emphasis on standards, portability, and systems that are widely available.

The tutorial surveys basic parallel computing concepts, using examples selected from multiple engineering, scientific, and data analysis problems. These examples illustrate using MPI on distributed memory systems; OpenMP on shared memory systems; MPI+OpenMP on hybrid systems; and CUDA and compiler directives on GPUs and accelerators. It discusses numerous parallelization and load balancing approaches, and software engineering and performance improvement aspects, including the use of state-of-the-art tools.

The tutorial helps attendees make intelligent decisions by covering the primary options that are available, explaining how they are used and what they are most suitable for. Extensive pointers to web-based resources are provided to facilitate follow-up studies.

Room: C141
8:30 am - 5:00 pm

Advanced OpenMP: Host Performance and 5.0 Features

Christian Terboven (RWTH Aachen University), Michael Klemm (Intel Corporation), Ruud van der Pas (Oracle), Bronis R. de Supinski (Lawrence Livermore National Laboratory)

With the increasing prevalence of multicore processors, shared-memory programming models are essential. OpenMP is a popular, portable, widely supported and easy-to-use shared-memory model. Developers usually find OpenMP easy to learn. However, they are often disappointed with the performance and scalability of the resulting code. This disappointment stems not from shortcomings of OpenMP but rather with the lack of depth with which it is employed. Our “Advanced OpenMP Programming” tutorial addresses this critical need by exploring the implications of possible OpenMP parallelization strategies, both in terms of correctness and performance.

We assume attendees understand basic parallelization concepts and know the fundamentals of OpenMP. We focus on performance aspects, such as data and thread locality on NUMA architectures, false sharing, and exploitation of vector units. All topics are accompanied with extensive case studies and we discuss the corresponding language features in-depth. Based on the feedback from previous years, we have taken out the part on directives for attached compute accelerators, to focus solely on performance programming for multi-core architectures. Throughout all topics, we present the recent additions of OpenMP 4.5 and extensions that have been subsequently adopted by the OpenMP Language Committee.

Room: C145
8:30 am - 5:00 pm

OpenMP Common Core: a "Hands-On" Exploration

Tim Mattson (Intel Corporation), Alice Koniges (University of Hawaii, Maui High Performance Computing Center), Yun (Helen) He (Lawrence Berkeley National Laboratory), David Eder (Maui High Performance Computing Center)

OpenMP is the de facto standard for writing parallel applications for shared memory computers. Born 20 years ago in 1997, it runs on just about every shared memory platform in the market. It's also very complicated. We created OpenMP to be the "easy API" for the general application programmer. With a specification running to over 300 pages, OpenMP has grown into an intimidating API viewed by many as for "experts only".

Most OpenMP programmers, however, use around 19 items from the specification. We call these 19 items the "OpenMP Common Core". By focusing on the common core, we make OpenMP what it was always meant to be; an easy API for parallel application programmers.

In this hands-on tutorial, we explore the common core of OpenMP. Students should come with an internet enabled laptop and be comfortable with basic commands to compile code. We utilize active learning through a carefully selected set of exercises, so students will master the common core and learn to apply it to their own problems. We will provide access to many-core systems that support OpenMP (such as Knights Landing). Alternatively, students can load an OpenMP compiler onto their laptops before the tutorial for the exercises.

Room: C148
8:30 am - 5:00 pm

Fault-Tolerance for High Performance and Distributed Computing: Theory and Practice

George Bosilca (University of Tennessee), Aurelien Bouteiller (University of Tennessee), Thomas Herault (University of Tennessee), Yves Robert (ENS Lyon, University of Tennessee)

Reliability is one of the major concerns when envisioning future exascale platforms. The International Exascale Software Project forecasts an increase in node performance and concurrency by one or two orders of magnitude, which translates, even under the most optimistic perspectives, in a mechanical decrease of the mean time to interruption of at least one order of magnitude. Because of this trend, platform providers, software implementors, and high-performance application users who target capability runs on such machines cannot regard the occurrence of interruption due to a failure as a rare dramatic event, but must consider faults inevitable, and therefore design and develop software components that have some form of fault-tolerance integrated at their core.

In this tutorial, we present a comprehensive survey on the techniques proposed to deal with failures in high performance and distributed systems. At the end of the tutorial, each attendee will have a better understanding of the fault tolerance premises and constraints, will know some of the available techniques, and will be able to determine, integrate, and adapt the technique which best suits their applications. In addition, the participants will learn how to employ existing fault tolerant infrastructure software to support more productive application development and deployment.

Room: C154
8:30 am - 5:00 pm

How to Analyze the Performance of Parallel Codes 101

Martin Schulz (Technical University Munich, Chair for Computer Architecture and Parallel Systems), Jim Galarowicz (Krell Institute), Don Maghrak (Krell Institute), Jennifer Green (Los Alamos National Laboratory), David Montoya (Los Alamos National Laboratory)

Performance analysis is an essential step in the development of HPC codes. It will even gain in importance with the rising complexity of machines and applications that we are seeing today. Many tools exist to help with this analysis, but the user is too often left alone with interpreting the results.

We will provide a practical road map for the performance analysis of HPC codes and will provide users step-by-step advice on how to detect and optimize common performance problems, covering both on-node performance and communication optimization as well as issues on threaded and accelerator-based architectures. In addition, we will review current approached in ubiquitous monitoring of performance utilizing lightweight tools. Throughout this tutorial, we will show live demos using Open|SpeedShop, a comprehensive and easy to use tool set. Additionally, at the end of each section we will provide hands-on exercises for attendees to try out the new techniques learned with performance comparison across multiple architectures. All techniques will, however, apply broadly to any tool, and we will point out alternative tools where useful.

Room: C155
8:30 am - 5:00 pm

Parallel-IO in Practice

Robert Latham (Argonne National Laboratory), Robert Ross (Argonne National Laboratory), Brent Welch (Google LLC), Glenn Lockwood (Lawrence Berkeley National Laboratory)

I/O on HPC systems is a black art. This tutorial sheds light on the state-of-the-art in parallel I/O and provides the knowledge necessary for attendees to best leverage I/O resources available to them. We cover the entire I/O software stack including storage and parallel file systems at the lowest layer, the role of burst buffers (NVRAM), intermediate layers (such as MPI-IO), and high-level I/O libraries (such as HDF-5). We emphasize ways to use these interfaces that result in high performance and tools for generating insight into these stacks. Benchmarks on real systems are used throughout to show real-world results.

In the first third of the tutorial we cover the fundamentals of parallel I/O. We discuss storage technologies, both present and near-future. Our parallel file systems material covers general concepts and gives examples from Lustre, GPFS, PanFS, HDFS, Ceph, and Cloud Storage.

Our second third takes a more application-oriented focus. We examine the upper library layers of the I/O stack, covering MPI-IO, Parallel netCDF, and HDF5. We discuss interface features, show code

examples, and describe how application calls translate into PFS operations.

Finally, we discuss tools for capturing and understanding I/O behavior.

Room: C156
8:30 am - 5:00 pm

Node-Level Performance Engineering

Georg Hager (University of Erlangen-Nuremberg, Erlangen Regional Computing Center), Gerhard Wellein (University of Erlangen-Nuremberg, Department of Computer Science)

The advent of multi- and manycore chips has led to a further opening of the gap between peak and application performance for many scientific codes. This trend is accelerating as we move from petascale to exascale. Paradoxically, bad node-level performance helps to "efficiently" scale to massive parallelism, but at the price of increased overall time to solution. If the user cares about time to solution on any scale, optimal performance on the node level is often the key factor. We convey the architectural features of current processor chips, multiprocessor nodes, and accelerators, as far as they are relevant for the practitioner. Peculiarities like SIMD vectorization, shared vs. separate caches, bandwidth bottlenecks, and ccNUMA characteristics are introduced, and the influence of system topology and affinity on the performance of typical parallel programming constructs is demonstrated. Performance engineering and performance patterns are suggested as powerful tools that help the user understand the bottlenecks at hand and to assess the impact of possible code optimizations. A cornerstone of these concepts is the roofline model, which is described in detail, including useful case studies, limits of its applicability, and possible refinements.

Room: C1/2/3/4 Ballroom
12:00 pm - 1:30 pm

Tutorials Lunch

Room: C140
1:30 pm - 5:00 pm

Container Computing for HPC and Scientific Workflows

Richard Shane Canon (Lawrence Berkeley National Laboratory), Rollin C. Thomas (Lawrence Berkeley National Laboratory)

Container computing has rapidly gained adoption in HPC and is revolutionizing the way applications are developed and delivered. It offers opportunities that never existed before for significantly improving efficiency of scientific workflows and easily moving these workflows from the laptop to the supercomputer. Tools like Docker and Shifter enable a new paradigm for scientific and technical

computing. However, to fully unlock its potential, users and administrators need to understand how to utilize these new approaches.

This tutorial will introduce attendees to the basics of creating container images, explain best practices, and cover more advanced topics such as creating images to be run on HPC platforms using Shifter. The tutorial will also explain how research scientists can utilize container-based computing to accelerate their research and how these tools can boost the impact of their research by enabling better reproducibility and sharing of their scientific process without compromising security. This is an extension of the highly successful tutorial presented at SC17 that was attended by more than 100 people. The SC17 tutorial was very highly rated with 2.8 / 3 stars for "would recommend" and 4.1 / 5 stars for overall quality.

Room: C143/149
1:30 pm - 5:00 pm

Tools and Best Practices for Distributed Deep Learning with Supercomputers

Weijia Xu (Texas Advanced Computing Center), Zhao Zhang (University of Texas), David Walling (University of Texas)

This tutorial is a practical guide on how to run distributed deep learning over distributed compute nodes effectively. Deep Learning (DL) has emerged as an effective analysis method and has been adapted quickly across many scientific domains in recent years. Domain scientists are embracing DL as both a standalone data science method, as well as an effective approach to reducing dimensionality in the traditional simulation. However, due to its inherent high computational requirement, application of DL is limited by the available computational resources.

Recently, we have seen the fusion of DL and HPC: supercomputers show an unparalleled capacity to reduce DL training time from days to minutes; HPC techniques have been used to speed up parallel DL training. Therefore distributed deep learning has great potential to augment DL applications by leveraging existing high performance computing cluster. This tutorial consists of three sessions. First, we will give an overview of the state-of-art approaches to enabling deep learning at scale. The second session is an interactive hands-on session to help attendees running distributed deep learning with resources at the Texas Advanced Computing Center. In the last session, we will focus on the best practices to evaluate and tune up performance.

Room: C144
1:30 pm - 5:00 pm

Productive Parallel Programming for FPGA with High-Level Synthesis

Johannes de Fine Licht (ETH Zurich), Torsten Hoefler (ETH Zurich)

As the scale of large high performance computing systems increases, so does their power consumption, making energy efficiency a first class citizen in their design. While GPUs and custom

processors have improved this situation significantly, reconfigurable architectures, such as FPGAs, promise another major step in energy efficiency, constituting a middle ground between fixed hardware architectures and custom-built ASICs.

Programming FPGAs has traditionally been done in hardware description languages, requiring extensive hardware knowledge and significant engineering effort. This tutorial shows how high-level synthesis (HLS) can be harnessed to productively achieve scalable pipeline parallelism on FPGAs. Attendees will learn how to target FPGA resources from high-level C++ or OpenCL code, guiding the mapping from imperative code to hardware, enabling them to develop massively parallel designs with real performance benefits. We treat concrete examples well known from the software world, relating traditional code optimizations to both corresponding and new transformations for hardware, building on existing knowledge when introducing new topics. By bridging the gap between software and hardware optimization, our tutorial aims to enable developers from a larger set of backgrounds to start tapping into the potential of FPGAs with real high performance codes.

Room: C146
1:30 pm - 5:00 pm

InfiniBand, Omni-Path, and High-Speed Ethernet: Advanced Features, Challenges in Designing HEC Systems, and Usage

Dhabaleswar Panda (Ohio State University), Hari Subramoni (Ohio State University), Sourav Chakraborty (Ohio State University)

As InfiniBand (IB), Omni-Path, and High-Speed Ethernet (HSE) technologies mature, they are being used to design and deploy various High-End Computing (HEC) systems: HPC clusters with GPGPUs supporting MPI, Storage and Parallel File Systems, Cloud Computing systems with SR-IOV Virtualization, Grid Computing systems, and Deep Learning systems. These systems are bringing new challenges in terms of performance, scalability, portability, reliability and network congestion. Many scientists, engineers, researchers, managers and system administrators are becoming interested in learning about these challenges, approaches being used to solve these challenges, and the associated impact on performance and scalability.

This tutorial will start with an overview of these systems. Advanced hardware and software features of IB, Omni-Path, HSE, and RoCE and their capabilities to address these challenges will be emphasized. Next, we will focus on Open Fabrics RDMA and Libfabrics programming, and network management infrastructure and tools to effectively use these systems. A common set of challenges being faced while designing these systems will be presented. Case studies focusing on domain-specific challenges in designing these systems, their solutions and sample performance numbers will be presented. Finally, hands-on exercises will be carried out with Open Fabrics and Libfabrics software stacks and Network Management tools.

Room: C147
1:30 pm - 5:00 pm

SENSEI Cross-Platform View of In Situ Analytics

E. Wes Bethel (Lawrence Berkeley National Laboratory), David Thompson (Kitware Inc), Burlen Loring (Lawrence Berkeley National Laboratory), Silvio Rizzi (Argonne National Laboratory), Brad Whitlock (Intelligent Light), Matthew Wolf (Oak Ridge National Laboratory), Patrick O'Leary (Kitware Inc)

This tutorial covers the design and use of SENSEI, a platform for in situ visualization and analysis. For simulation developers, you can instrument your code with the SENSEI simulation API and gain access to a wide variety of I/O, analysis, and visualization tools. For tool developers, you can add support for the SENSEI analysis API and support a wide range of simulations.

Attendees will learn the basics of in situ analysis and visualization — which eliminates the need for writing large simulation state files or other data that prevents scaling to large machines — while being exposed to advanced analysis such as autocorrelation, interactive monitoring, and computational steering. We'll demonstrate how simulations can couple to ADIOS, Henson, ParaView Catalyst, and VisIt Libsim through the SENSEI platform in order to to gain insight from simulation data even when it is increasingly impractical to save it to persistent storage.

The tutorial will include a hands-on session that uses a virtual machine to demonstrate SENSEI features. Please download it before SC18 from https://www.sensei-insitu.org/ ; it will be available in early November.

Monday, November 12th

Room: C140
8:30 am - 12:00 pm

The Business of HPC: TCO, Funding Models, Metrics, Value, and More

Andrew Jones (Numerical Algorithms Group), Owen G. M. Thomas (Red Oak Consulting), Ingrid Barcena Roig (KU Leuven, Belgium), Caleb Hamilton (Numerical Algorithms Group)

The tutorial provides an impartial, practical, non-sales focused guide to the business aspects of HPC facilities and services. It presents a rare and invaluable opportunity for HPC managers, practitioners, and stakeholders to learn about using TCO models; the pros and cons of different cost recovery and funding models; appropriate metrics for HPC services, and assessing the value of HPC.

Well-managed TCO, return on investment, and cost recovery models can be hugely beneficial to HPC managers and operators by demonstrating the value of HPC to the organization, driving the continuation and growth of HPC investment. They can also help uncover practical improvements to deliver better services to users.

After the tutorial, attendees will be in a stronger position to calculate and use TCO within their organizations, to design and use cost-recovery models, to select and use appropriate metrics, and articulate the value of HPC.

The tutorial is based on experience across a diverse set of real world cases in various countries, in both private and public sectors, with projects of all sizes and shapes.

The material is based on our successful SC17 TCO and Business Case tutorials, with selected elements of those tutorials brought together with additional new content.

Room: C141
8:30 am - 12:00 pm

Mastering Tasking with OpenMP

Christian Terboven (RWTH Aachen University), Michael Klemm (Intel Corporation), Sergi Mateo Bellido (Barcelona Supercomputing Center), Xavier Teruel (Barcelona Supercomputing Center), Bronis R. de Supinski (Lawrence Livermore National Laboratory)

With the increasing prevalence of multi-core processors, shared-memory programming models are essential. OpenMP is a popular, portable, widely supported and easy-to-use shared-memory model. Since version 3.0 released in 2008, OpenMP offers tasking to support the creation of composable parallel software blocks and the parallelization of irregular algorithms. Developers usually find OpenMP easy to learn. However, mastering the tasking concept of OpenMP requires a change in the way developers reason about the structure of their code and how to expose the parallelism of it. Our tutorial addresses this critical aspect by examining the tasking concept in detail and presenting patterns as solutions to many common problems.

We assume attendees understand basic parallelization concepts and know the fundamentals of OpenMP. We present the OpenMP tasking language features in detail and focus on performance aspects, such as introducing cut-off mechanisms, exploiting task dependencies, and preserving locality. Extensive case studies accompany all aspects. Throughout all topics, we present the features of OpenMP 4.5 and the additions that will appear in OpenMP 5.0, to be released during SC18.

Room: C146
8:30 am - 12:00 pm

Containers, Collaboration, and Community: Hands-On Building a Data Science Environment for Users and Admins

Rion Dooley (University of Texas), Steve Brandt (Louisiana State University), Zebula Sampedro (University of Colorado), Kathryn Traxler (Louisiana State University), Aaron Holt (University of Colorado)

In this tutorial we combine best practices and lessons learned in evolving traditional HPC data centers at TACC, CU, and LSU into more integrated data science environments. In the first session, participants will learn about best practices in data science and software engineering and apply them while containerizing and scaling an MPI application using multiple container technologies across clouds, clusters, and a sandbox environment we will provide. They will then leverage continuous

integration and delivery to increase the portability, visibility, and availability of their individual application codes. The first session will conclude with participants learning how the same approaches can be used by sysadmins to improve update and release velocities of their entire application catalogs while better balancing security and regression support concerns.

In the second session, participants will learn how to leverage automation and cloud services to better handle version and metadata management. From there, they will customize their own data science environment and gain hands-on experience using cloud services and the Agave Platform by extending their environment to publish and share applications, manage data, orchestrate simulations across both HPC and cloud resources, capture provenance information, and foster collaboration.

Room: C154
8:30 am - 12:00 pm

Introduction of Practical Approaches to Data Analytics for HPC with Spark

Michela Taufer (University of Tennessee), Travis Johnston (Oak Ridge National Laboratory), Stephen Herbein (Lawrence Livermore National Laboratory, University of Tennessee), Danny Rorabaugh (University of Tennessee), Michael Wyatt (University of Delaware, University of Tennessee), Dylan Chapp (University of Delaware, University of Tennessee)

This tutorial provides a practical introduction to big data analytics, blending theory (e.g., of clustering algorithms and techniques for dealing with noisy data) and practice (e.g., using Apache Spark, Jupyter Notebooks, and Github). Over the course of five modules, participants will become familiar with modern data science methods, gain comfort with the tools of the trade, explore real-world data sets, and leverage the power of HPC resources to extract insights from data. Upon completing the tutorial, participants will have: used Jupyter notebooks to create reproducible, explanatory data science workflows; learned a modern MapReduce implementation, Apache Spark; implemented parallel clustering methods in Spark; studied strategies for overcoming the common imperfections in real-world datasets, and applied their new skills to extract insights from a high-dimensional medical dataset.

Room: C142
8:30 am - 5:00 pm

Programming Your GPU with OpenMP: A Hands-On Introduction

Tim Mattson (Intel Corporation), Simon McIntosh-Smith (University of Bristol), Eric Stotzer (Texas Instruments)

OpenMP 1.0 was released in 1997 when the primary concern was symmetric multiprocessors. Over time, hardware has evolved with more complex memory hierarchies forcing us to embrace NUMA machines and work to understand how OpenMP fits in with distributed memory systems.

Current trends in hardware bring co-processors such as GPUs into the fold. A modern platform is

often a heterogeneous system with CPU cores, GPU cores, and other specialized accelerators. OpenMP has responded by adding directives that map code and data onto a device. We refer to this family of directives as the target directives.

In this hands-on tutorial, we will explore these directives as they apply to programming GPUs. We assume people know the fundamentals of OpenMP (perhaps by taking the OpenMP Common Core tutorial) so we can focus on deeply understanding the target directives and their use in complex application programs. We expect students to use their own laptops (with Windows, Linux, or OS/X) to connect to remote servers with GPUs, but the best option is for students to load an OpenMP compiler onto their laptops before the tutorial. Information about OpenMP compilers is available at www.openmp.org.

Room: C143/149
8:30 am - 5:00 pm

Advanced MPI Programming

Pavan Balaji (Argonne National Laboratory), William Gropp (University of Illinois), Torsten Hoefler (ETH Zurich), Rajeev Thakur (Argonne National Laboratory)

The vast majority of production parallel scientific applications today use MPI and run successfully on the largest systems in the world. At the same time, the MPI standard itself is evolving to address the needs and challenges of future extreme-scale platforms as well as applications. This tutorial will cover several advanced features of MPI, including new MPI-3 features, that can help users program modern systems effectively. Using code examples based on scenarios found in real applications, we will cover several topics including efficient ways of doing 2D and 3D stencil computation, derived datatypes, one-sided communication, hybrid (MPI + shared memory) programming, topologies and topology mapping, and neighborhood and nonblocking collectives. Attendees will leave the tutorial with an understanding of how to use these advanced features of MPI and guidelines on how they might perform on different platforms and architectures.

Room: C144
8:30 am - 5:00 pm

Deep Learning at Scale

Steven A. Farrell (Lawrence Berkeley National Laboratory), Deborah Bard (Lawrence Berkeley National Laboratory), Michael F. Ringenburg (Cray Inc), Thorsten Kurth (Lawrence Berkeley National Laboratory), Mr Prabhat (Lawrence Berkeley National Laboratory)

Deep learning is rapidly and fundamentally transforming the way science and industry use data to solve problems. Deep neural network models have been shown to be powerful tools for extracting insights from data across a large number of domains. As these models grow in complexity to solve increasingly challenging problems with larger and larger datasets, the need for scalable methods and software to train them grows accordingly.

The Deep Learning at Scale tutorial aims to provide attendees with a working knowledge of deep learning on HPC class systems, including core concepts, scientific applications, and techniques for scaling. We will provide training accounts and example Jupyter notebook-based exercises, as well as datasets, to allow attendees to experiment hands-on with training, inference, and scaling of deep neural network machine learning models.

Room: C145
8:30 am - 5:00 pm

Application Porting and Optimization on GPU-Accelerated POWER Architectures

Dirk Pleiter (Juelich Supercomputing Centre), Christoph Hagleitner (IBM Zurich Research Laboratory), Andreas Herten (Juelich Supercomputing Centre), Tom Papatheodore (Oak Ridge National Laboratory), Archana Ravindar (IBM India), Mathias Wagner (Nvidia Corporation)

The POWER processor has re-emerged as a technology for supercomputer architectures. One major reason is the tight integration of processor and GPU accelerator through the NVLink technology. Two major sites in the US, ORNL and LLNL, have already decided to have their pre-exascale systems being based on this new architecture (Summit and Sierra, respectively). This tutorial will give an opportunity to obtain in-depth knowledge and experience with GPU-accelerated POWER nodes. It focuses on porting applications to a single node and covers the topics architecture, compilers, performance analysis and tuning, and multi-GPU programming. The tutorial will include an overview of the NVLink-based node architectures, lectures on first-hand experience in porting to this architecture, and exercises using tools to focus on performance.

Room: C147
8:30 am - 5:00 pm

High Performance I/O Frameworks 101

Scott Klasky (Oak Ridge National Laboratory), Qing Liu (New Jersey Institute of Technology), Manish Parashar (Rutgers University), Norbert Podhorszki (Oak Ridge National Laboratory), David Pugmire (Oak Ridge National Laboratory), Kesheng Wu (Lawrence Berkeley National Laboratory)

As concurrency and complexity continue to increase on high-end machines, I/O performance is rapidly becoming a fundamental challenge to achieving exascale computing. Wider adoption of higher-level I/O abstractions will be critically important to address this challenge. Modern I/O libraries provide data models, portable APIs, storage abstractions, and self-describing data containers. They achieve high performance and scalability, allow data to be managed more effectively throughout the data lifecycle, and enable reproducible science.

Part I of this tutorial will provide an overview of parallel I/O systems and summarize the key techniques for obtaining high performance I/O on high-performance computing (HPC) resources at scale. Part II introduces ADIOS and HDF5 libraries, delving through their usage models and examples,

showing how to achieve high performance scalable I/O. Part III explains data compression. Part IV covers techniques for creating in situ analytics and teaches how to generate visualization services using VTK-M. Finally, Part V will explain data indexing/querying and how to use the libraries to query data both in situ and on files. Over one half of this tutorial will be hands-on sessions, where we provide access to the software and go through live examples.

Instructions for installing software for the tutorial: https://users.nccs.gov/~pnorbert/SC18Notes.pdf

Room: C148
8:30 am - 5:00 pm

Managing HPC Software Complexity with Spack

Gregory B. Becker (Lawrence Livermore National Laboratory), Massimiliano Culpo (Swiss Federal Institute of Technology in Lausanne), Matthew P. LeGendre (Lawrence Livermore National Laboratory), Mario Melara (Lawrence Berkeley National Laboratory), Peter Scheibel (Lawrence Livermore National Laboratory), Adam J. Stewart (University of Illinois), Todd Gamblin (Lawrence Livermore National Laboratory)

HPC software is becoming increasingly complex. The largest applications require over 100 dependency libraries, and they combine interpreted languages like Python with lower-level C, C++, and Fortran libraries. To achieve good performance, developers must tune for multiple compilers, build options, and libraries like MPI, BLAS, and LAPACK. The space of possible build configurations is combinatorial, and developers waste countless hours rebuilding software instead of producing new scientific results.

Spack is an open-source tool for HPC package management, used on some of the fastest supercomputers in the world. It allows developers to write simple recipes in Python to automate builds with arbitrary combinations of compilers, MPI versions, build options, and dependency libraries. With Spack, users can install over 2,700 community-maintained packages without knowing how to build them, developers can automate builds of tens or hundreds of dependency libraries, and staff at HPC centers can deploy many versions of software for thousands of users. We provide a thorough introduction to Spack's capabilities: basic software installation, creating new packages, development workflows using Spack, and advanced multi-user deployment.

Most sessions involve hands-on demonstrations, so attendees should bring a laptop computer.

Room: C155
8:30 am - 5:00 pm

Better Scientific Software

David E. Bernholdt (Oak Ridge National Laboratory), Anshu Dubey (Argonne National Laboratory, University of Chicago), Michael A. Heroux (Sandia National Laboratories), Jared O'Neal (Argonne National Laboratory)

The computational science and engineering (CSE) community is in the midst of an extremely challenging period created by the confluence of disruptive changes in computing architectures, demand for greater scientific reproducibility, and new opportunities for greatly improved simulation capabilities, especially through coupling physics and scales. Computer architecture changes require new software design and implementation strategies, including significant refactoring of existing code. Reproducibility demands require more rigor across the entire software endeavor. Code coupling requires aggregate team interactions including integration of software processes and practices. These challenges demand large investments in scientific software development and improved practices. Focusing on improved developer productivity and software sustainability is both urgent and essential.

This tutorial will provide information and hands-on experience with software practices, processes, and tools explicitly tailored for CSE. Goals are improving the productivity of those who develop CSE software and increasing the sustainability of software artifacts. We discuss practices that are relevant for projects of all sizes, with emphasis on small teams, and on aggregate teams composed of small teams. Topics include software licensing, effective models, tools, and processes for small teams (including agile workflow management), reproducibility, and scientific software testing (including automated testing and continuous integration).

Room: C156
8:30 am - 5:00 pm

Quantum Computing for Scientific Applications

Costin Iancu (Lawrence Berkeley National Laboratory), Wibe de Jong (Lawrence Berkeley National Laboratory), Pavel Lougovski (Oak Ridge National Laboratory), Ojas Parekh (Sandia National Laboratories), Jarrod McClean (Google LLC), Nicholas Rubin (Rigetti Computing), Ali Javadi (IBM), Davide Venturelli (NASA)

Quantum computing is an emerging technology which promises to revolutionize many computational tasks. However, for a non-specialist it may appear that it is surrounded by a shroud of hype and misconception. The main goal of this tutorial is to de-mystify practical quantum computing and all its vital algorithmic aspects to a general audience of computer scientists with little to none prior knowledge of the subject. We plan to achieve this through a combination of lecture materials, demonstrations, and hands on exercises delivered by quantum computing experts with extensive experience of public speaking and teaching from the government research laboratories (DOE, NASA), industry (IBM, Google, Rigetti), and academia. In particular, we aim to elucidate quantum computing use for scientific applications, covering the following areas: 1) quantum algorithm design; 2) quantum programming toolkits; and 3) practical error mitigation for quantum algorithms. We will focus on the design and implementation of hybrid quantum-classical computational strategies including variational quantum eigensolver (VQE) and quantum approximate optimization algorithms in the context of quantum chemistry, nuclear structure, and quantum field theory problems. We will discuss multiple practical ways to mitigate systematic coherent errors in the nascent quantum hardware, including general techniques such as randomized compilation.

Room: C1/2/3/4 Ballroom
12:00 pm - 1:30 pm

Tutorials Lunch

Room: C140
1:30 pm - 5:00 pm

Procurement and Commissioning of HPC Systems

Andrew Jones (Numerical Algorithms Group), Ingrid Barcena Roig (KU Leuven, Belgium), Owen G. M. Thomas (Red Oak Consulting), Caleb Hamilton (Numerical Algorithms Group), Vicki Lockhart (Red Oak Consulting)

This tutorial will walk attendees through the whole process of purchasing and deploying a HPC system. It will cover the whole process from engaging with stakeholders in securing funding, requirements capture, market survey, specification of the tender/request for proposal documents, engaging with suppliers, evaluating proposals, and managing the installation.

Attendees will learn how to specify what they want, yet enable the suppliers to provide innovative solutions beyond their specification both in technology and in the price; how to demonstrate to stakeholders that the solution selected is best value for money; and the common risks, pitfalls and mitigation strategies essential to achieve an on-time and on-quality installation process. The tutorial has 4 parts: procurement process including RFP; benchmarks, bid evaluation, clarifications; contracting, project management, commissioning, acceptance testing, and transition to user service.

The presenters have been involved in numerous major HPC procurements for several decades, as service managers, bidders to funding agencies, as customers and as advisors. The presenters are from the UK and Europe but the tutorial will be applicable to HPC procurements anywhere. The tutorial is based on experience across a diverse set of real world cases in various countries, in private and public sectors.

Room: C141
1:30 pm - 5:00 pm

Compression for Scientific Data

Franck Cappello (Argonne National Laboratory), Peter Lindstrom (Lawrence Livermore National Laboratory)

Large-scale numerical simulations and experiments generate very large datasets that are difficult to analyze, store, and transfer. This problem will be exacerbated for future generations of systems. Data reduction becomes a necessity in order to reduce as much as possible the time lost in data transfer and storage. Data compression is an attractive and efficient reduction technique that is rather agnostic to the application. This tutorial will introduce motivating examples, basic compression techniques, state

of the art data transformation, prediction, quantization, and coding techniques; discuss in detail the SZ and ZFP lossy compressors (including their latest developments); introduce compression error assessment metrics; and show how lossy compression impacts visualization and other data analytics. The tutorial will also cover Z-checker, a tool to characterize data sets and assess compression error. The tutorial will use examples of real world compressors and datasets coming from simulations and instruments to illustrate the different compression techniques, their performance and their impact. From a user perspective, the tutorial will detail how to use ZFP, SZ and Z-checker as stand alone software/libraries and as modules integrated in parallel I/O libraries (ADIOS, HDF5).

Room: C146
1:30 pm - 5:00 pm

Exploiting HPC Technologies for Accelerating Big Data Processing and Associated Deep Learning

Dhabaleswar K. Panda (Ohio State University), Xiaoyi Lu (Ohio State University), Shashank Gugnani (Ohio State University)

The convergence of HPC, Big Data, and Deep Learning is the next game-changing business opportunity. Apache Hadoop, Spark, gRPC/TensorFlow, and Memcached are becoming standard building blocks for Big Data processing. Recent studies have shown that default designs of these components cannot efficiently leverage the features of modern HPC clusters, like RDMA-enabled high-performance interconnects, high-throughput parallel storage systems (e.g. Lustre), Non-Volatile Memory (NVM), NVMe/NVMe-over-Fabric. This tutorial will provide an in-depth overview of the architecture of Hadoop, Spark, gRPC/TensorFlow, and Memcached. We will examine the challenges in re-designing networking and I/O components of these middleware with modern interconnects and storage architectures. Using the publicly available software packages in the High-Performance Big Data project (HiBD, http://hibd.cse.ohio-state.edu), we will provide case studies of the new designs for several Hadoop/Spark/gRPC/TensorFlow/Memcached components and their associated benefits. Through these, we will also examine the interplay between high-performance interconnects, storage, and multi-core platforms to achieve the best solutions for these components and applications on modern HPC clusters. We also present in-depth case-studies with modern Deep Learning tools (e.g., Caffe, TensorFlow, CNTK, BigDL) with RDMA-enabled Hadoop, Spark, and gRPC. Finally, hands-on exercises will be carried out with RDMA-Hadoop and RDMA-Spark software stacks over a cutting-edge HPC cluster.

Room: C154
1:30 pm - 5:00 pm

Performance Tuning of Scientific Codes with the Roofline Model

Samuel W. Williams (Lawrence Berkeley National Laboratory), Aleksandar Ilic (INESC-ID, Portugal), Zakhar Matveev (Intel Corporation), Charlene Yang (Lawrence Berkeley National Laboratory)

The Roofline performance model offers an insightful and intuitive method for extracting the key execution

characteristics of HPC applications and comparing them against the performance bounds of modern processors. Its capability to abstract the complexity of modern non-uniform memory hierarchies and identify the most profitable optimization techniques has made Roofline-based analysis increasingly popular in the HPC community. Although different flavors of the Roofline model have been developed to deal with various definitions of memory data movement, there is still a need for a more systematic methodology when applying them to analyze the efficiency of applications running on multi- and manycore systems. The tutorial aims to bridge this gap by exposing the fundamental aspects behind different Roofline modeling principles and providing several practical use case scenarios to highlight their efficacy for application optimization. This tutorial presents a unique and solid combination of novel methodologies applied to optimize a representative set of open science use cases, while practice-oriented, hands-on topics and labs are given by the lead methodology researchers and the main designer of Intel's Roofline automation tools. The tutorial presenters have a long history of working with the Roofline model and have presented several Roofline-based tutorials.

Workshop

(back to top)

Sunday, November 11th

Room: D172
9:00 am - 12:30 pm

5th International Workshop on HPC User Support Tools: HUST-18

Supercomputing centers exist to drive scientific discovery by supporting researchers in computational science fields. To make users more productive in the complex HPC environment, HPC centers employ user support teams. These teams serve many roles, from setting up accounts, to consulting on math libraries and code optimization, to managing HPC software stacks. Often, support teams struggle to adequately support scientists. HPC environments are extremely complex, and combined with the complexity of multi-user installations, exotic hardware, and maintaining research software, supporting HPC users can be extremely demanding.

With the 5th HUST workshop, we will continue to provide a necessary forum for system administrators, user support team members, tool developers, policy makers, and end users. We will provide a forum to discuss support issues and we will provide a publication venue for current support developments. Best practices, user support tools, and any ideas to streamline user support at supercomputing centers are in scope.

OOOPS: An Innovative Tool for IO Workload Management on Supercomputers *Lei Huang (University of Texas), Si Liu (University of Texas)*
Modern supercomputer applications are demanding powerful storage resources in addition to fast computing resources. However, these storage resources, especially parallel shared filesystems, have become the Achilles' heel of many powerful supercomputers. A single user's IO-intensive work can result in global filesystem performance degradation and even unresponsiveness. In this project, we developed an innovative IO workload managing tool that controls the IO workload from user applications' side. This tool is capable of automatically detecting and throttling intensive IO workload caused by supercomputer users to protect parallel shared filesystems.

CView and NWPerf for Supercomputer Performance Collection and Display. *Evan Felix (Pacific Northwest National Laboratory)*
CView is a full 3D interactive graphics tool for displaying collected System and Application performance data. Data is collected using various open source tools common on large clusters such as ganglia, collectl, or collectd. When collecting data on large scale systems with thousands of nodes to tens of thousands of nodes, this data set becomes quite large over time. For example, on The Cascade System at the Environmental Molecular Sciences Laboratory at PNNL one days collection of performance data stored as raw floats is about 2.5 Gigabytes of data. The CView Graphics environment allows a user to easily visualize such data-sets. This allows System Administrators to

'see' how the system is performing in near-time data streams. System Administrators and System Users use it for visualizing the performance of the whole systems or just the nodes involved in a specific job run. The Systems at PNNL use the NWPerf software, backed by a Ceph Object store to route and store the performance data our clusters. By storing the data this method a user can display data from any point in time and investigate various metrics at the system or job level.

HUST – Workshop Morning Break

ReFrame: A Regression Testing and Continuous Integration Framework for HPC systems *Vasileios Karakasis (Swiss National Supercomputing Centre)*
Regression testing of HPC systems is of crucial importance when it comes to ensure the quality of service offered to the end users. At the same time, it poses a great challenge to the systems and application engineers to continuously maintain regression tests that cover as many aspects as possible of the user experience. In this presentation, we introduce ReFrame, a new framework for writing regression tests for HPC systems. ReFrame is designed to abstract away the complexity of the interactions with the system and separate the logic of a regression test from the low-level details, which pertain to the system configuration and setup. Regression tests in ReFrame are simple Python classes that specify the basic parameters of the test plus any additional logic. The framework will load the test and send it down a well-defined pipeline which will take care of its execution. All the system interaction details, such as programming environment switching, compilation, job submission, job status query, sanity checking and performance assessment, are performed by the different pipeline stages. Thanks to its high-level abstractions and modular design, ReFrame can also serve as a tool for continuous integration (CI) of scientific software, complementary to other well-known CI solutions. Finally, we present the use cases of two large HPC centers that have adopted or are now adopting ReFrame for regression testing of their computing facilities.

A Compiler and Profiler Based Tool for Querying HPC Application Characteristics *Aaron Welch (Oak Ridge National Laboratory)*
Emerging HPC platforms are becoming extraordinarily difficult to program as a result of complex, deep and heterogeneous memory hierarchies, heterogeneous cores, and the need to divide work and data among them. New programming models and libraries are being developed to aid in porting efforts of application developers, but substantial code restructuring is still necessary to fully make use of these new technologies. To do this effectively, these developers need information about their source code characteristics, including static and dynamic (e.g. performance) information to direct their optimisation efforts and make key decisions. On the other hand, system administrators need to understand how users are using the software stack and resources on their systems to understand what software they need to provide to improve the productivity of the users on a platform.

In this presentation, we describe a tool that combines compiler and profiler information to query performance characteristics within an application or across multiple applications on a given platform. Static and dynamic data about applications are collected and stored together in an SQL database that can be queried by either a developer or a system administrator. We will demonstrate the capabilities of this tool via a real-world example from application-driven case studies that aims at understanding the use of scientific libraries on a routine from the molecular simulation application CP2K.

ColdFront: An Open Source HPC Resource Allocation System *Mohammad Zia (State University of New York at Buffalo)*

ColdFront is an open source resource allocation system designed to provide a central portal for administering HPC resources and collecting return on investment (ROI) metrics. ColdFront was created to help HPC centers manage access to center resources across large groups of users and provide a rich set of data for comprehensive reporting of ROI metrics such as user publications and external funding. ColdFront is written in Python and released under the GPLv3 license. This presentation will include an overview of ColdFront and include a live demo of its installation and use.

Concluding Remarks

Room: D171
9:00 am - 12:30 pm

ResCuE-HPC: 1st Workshop on Reproducible, Customizable, and Portable Workflows for HPC

Experiment reproducibility and artifact sharing is gradually becoming a norm for publications at HPC conferences. However, our recent experience to validate experimental results during artifact evaluation (AE) at PPoPP, CGO, PACT and SC also highlighted multiple problems. Ad-hoc experimental workflows, lack of common experimental methodology and tools, and the ever changing software/hardware stack all place a heavy burden on evaluators when installing, running and analyzing complex experiments. Worse still, even when artifacts (e.g. benchmarks, data sets, models) are shared, they are typically difficult to customize, port, reuse, and build upon. Establishing common experimental workflow frameworks with common formats/APIs and portable package managers can considerably simplify validation and reuse of experimental results.

This workshop will bring together HPC researchers and practitioners interested in developing common experimental methodologies, workflow frameworks and package managers for HPC (in particular, participants in CLUSTER competitions and the recent SC reproducibility initiative). We will discuss the best practices of sharing benchmarks, data sets, tools and experimental results in a customizable and reusable way. We will also cover state-of-the-art in frameworks for running workloads on ever changing systems, their current drawbacks and future improvements. Finally, we will compile a report on the current state-of-the art techniques and tools which we will share with artifact evaluation committees at SC and other top-tier conferences, and the ACM taskforce on reproducibility where we are founding members. We believe such a practical approach will help to improve reproducibility initiative and artifact exchange at SC while accelerating hardware/software co-design of efficient HPC systems.

Keynote *Michael A. Heroux (Sandia National Laboratories)*

ResCuE-HPC – Workshop Morning Break

Assessing Reproducibility: An Astrophysical Example of Computational Uncertainty in the HPC Context

Victoria Stodden (University of Illinois)

We present an experiment using the Enzo simulation code on NCSA's Blue Waters system to highlight the importance of computational and numerical uncertainty in scientific computing on HPC systems. We quantify the (surprising) variability of outputs from 200 identical simulation runs. We make two recommendations to improve the assessment of reproducibility of computational results: the inclusion of computational variability measures in standard reporting practices; and modular software designs that permit localized assessments of the source of computational uncertainty.

Considering the Development Workflow to Achieve Reproducibility with Variation

Michael Mercier (Atos), Adrien Faure (Atos)

The ability to reproduce an experiment is fundamental in computer science. Existing approaches focus on repeatability, but this is only the first step to reproducibility: continuing a scientific work from a previous experiment requires being able to modify it. This ability is called reproducibility with variation.

In this contribution, we show that capturing the environment of execution is necessary but not sufficient; we also need the environment of development. The variation also implies that those environments are subject to evolution, so the whole software development lifecycle needs to be considered. To take into account these evolutions, software environments need to be clearly defined, reconstructible with variation, and easy to share. We propose to leverage functional package managers to achieve this goal.

Spotting Black Swans With Ease: The Case for a Practical Reproducibility Platform

Ivo Jimenez (University of California, Santa Cruz)

Advances in agile software delivery methodologies and tools (commonly referred to as _DevOps_) have not yet materialized in academic scenarios such as university, industry and government laboratories. In this position paper, we make the case for _Black Swan_, a platform for the agile implementation, maintenance, and curation of experimentation pipelines by embracing a DevOps approach.

Supporting Thorough Artifact Evaluation with Occam

Bruce Childers (University of Pittsburgh)

Efforts such as Artifact Evaluation (AE) have been growing, gradually making software evaluation an integral part of scientific publication. In this paper, we describe how Occam can help to mitigate some of the challenges faced by both authors and reviewers. For authors, Occam provides the means to package their artifacts with enough detail to be used within experiments that can be easily repeated. For the reviewers, Occam provides the means to thoroughly evaluate artifacts by: allowing them to repeat the author's experiments; providing the ability to modify inputs, parameters, and software to run different experiments.

Semantically Organized Containers for Reproducible Research
Tanu Malik (DePaul University)

Experiments are a key component in systems and HPC-related research. They help validate new ideas and concepts. Sharing and reproducing experiments, however, is a challenge, especially when computational experiments reside in multiple computing environments, are disorganized into multiple directories, are disconnected from each other, or lack sufficient documentation.

In this paper, we show how sharing, porting, and reproducing distributive and iterative experiments can be simplified by using an automatic containerization tool for capturing/repeating an experiment and a convention for organizing repeated runs of an experiment. Using a simulation-analysis workflow, we show how semantically organized containers can help a reviewer find all experiments for a given result and re-execute all experiments with fail-proof guarantee. We discuss outstanding challenges of adopting this method as an artifact evaluation mechanism.

Open Panel: Automating Artifact Sharing, Evaluation, and Reuse *Grigori Fursin (Dividiti Ltd, cTuning Foundation), Todd Gamblin (Lawrence Livermore National Laboratory), Michela Taufer (University of Delaware), Michael A. Heroux (Sandia National Laboratories), Stephen Lien Harrell (Purdue University)*

Room: D175
9:00 am - 12:30 pm

Fifth Workshop on Accelerator Programming Using Directives (WACCPD)

Current hardware trends lead to ever more complex compute node architectures offering multiple, heterogeneous levels of massive parallelism. As a result, the 'X' in MPI+X demands more focus. In order to exploit the maximum available parallelism out of such systems, we are in dire need of sophisticated programming approaches that can provide scalable as well as portable solutions without compromising on performance. The expectation from the scientific community is that such solutions should allow programmers to maintain a single code base whenever possible and to avoid requiring maintaining and debug multiple versions of the same code.

Raising the abstraction of the code is one of the effective methodologies to reduce the burden on the programmer. At the same time such a methodology will require a heroic compiler to be designed. Software abstraction-based programming models such as OpenMP and OpenACC have been serving this purpose over the past several years. These programming models address the 'X' component by providing programmers high-level directive-based approaches to accelerate and port scientific applications to heterogeneous platforms.

The focus of this workshop is to explore this 'X' component in a hybrid MPI+X programming approach.

We present technical papers discussing innovative high-level language features and their (early prototype) implementations needed to address hierarchical heterogeneous systems, stories and lessons learned while using directives to migrate scientific legacy code to parallel processors, state-of-the-art compilation and runtime scheduling techniques, techniques to optimize performance, mechanisms to keep communication and synchronization efficient.

Session1: WACCPD Keynote: Experiences in Using Directive-Based Programming for Accelerated Computing Architectures *Jack Wells (Oak Ridge National Laboratory)*
Accelerated computing architectures have grown in their application within scientific computing since their introduction approximately ten-years ago. From the earliest days, there has been a focus on the programmability of these systems. A variety of desired outcomes have driven the development of directive-based programming approaches for accelerated computing, including improvements in developer productivity and application portability, APIs that are non-proprietary or vendor non specific, and that support incremental acceleration of application codes. The first specification was OpenACC 1.0 introduced in November 2011. With major enhancements, OpenACC has evolved to version 2.5, and is providing constructive inputs to the OpenMP specification. In this talk, we discuss how the use of compiler directives have evolved over time and their implementation status on Titan and Summit. The talk will also discuss which applications on Titan are using directives and how their usage has been changing over time. To end, we will discuss the challenges that need to be solved and how new emerging frameworks are changing the way applications are using directives (e.g. as backends for Kokkos, etc) for C++.

Session 2: Porting Scientific Applications Using Directives *Sandra Wienke (RWTH Aachen University)*

Heterogeneous Programming and Optimization of Gyrokinetic Toroidal Code Using Directives
Wenlu Zhang (Institute of Physics, Chinese Academy of Sciences; University of California, Irvine)

The latest production version of the fusion particle simulation code, Gyrokinetic Toroidal Code (GTC), has been ported to and optimized for the next generation exascale GPU supercomputing platform. Heterogeneous programming using directives has been utilized to fuse and thus balance the continuously implemented physical capabilities and rapidly evolving software/hardware systems. The original code has been refactored to a set of unified functions/calls to enable the acceleration for all the species of particles. Binning and GPU texture caching technique have also been used to boost the performance of the particle push and shift operations. In order to identify the hotspots, the GPU version of the GTC code was the first benchmarked on up to 8000 nodes of the Titan supercomputer, which shows about 2–3 times overall speedup comparing NVidia M2050 GPUs to Intel Xeon X5670 CPUs. This Phase I optimization was followed by further optimizations in Phase II, where single-node tests show an overall speedup of about 34 times on SummitDev and 7.9 times on Titan. The real physics tests on Summit machine showed impressive scaling properties that reaches roughly 50% efficiency on 928 nodes of Summit. The GPU+CPU speed up from purely CPU is over 20 times, leading to an unparalleled speed.

WACCPD – Workshop Morning Break

Using Compiler Directives for Performance Portability in Scientific Computing: Kernels from Molecular Simulation

Ada Sedova (Oak Ridge National Laboratory)

Achieving performance portability for high-performance computing (HPC) applications in scientific fields has become an increasingly important initiative due to large differences in emerging supercomputer architectures. Here we test some key kernels from molecular dynamics (MD) to determine whether the use of the OpenACC directive-based programming model when applied to these kernels can result in performance within an acceptable range for these types of programs in the HPC setting. We find that for easily parallelizable kernels, performance on the GPU remains within this range. On the CPU, OpenACC-parallelized pairwise distance kernels would not meet the performance standards required, when using AMD Opteron "Interlagos" processors, but with IBM Power 9 processors, performance remains within an acceptable range for small batch sizes. These kernels provide a test for achieving performance portability with compiler directives for problems with memory-intensive components as are often found in scientific applications.

Session 3: Using OpenMP *Jeff Larkin (Nvidia Corporation)*

OpenMP Target Offloading: Splitting GPU Kernels, Pipelining Communication and Computation, and Selecting Better Grid Geometries

Jose N. Amaral (University of Alberta)

This paper presents three ideas that focus on improving the execution of high-level parallel code in GPUs. The first addresses programs that include multiple parallel blocks within a single region of GPU code. A proposed compiler transformation can split such regions into multiple, leading to the launching of multiple kernels, one for each parallel region. Advantages include the opportunity to tailor grid geometry of each kernel to the parallel region that it executes and the elimination of the overheads imposed by a code-generation scheme meant to handle multiple nested parallel regions. Second, is a code transformation that sets up a pipeline of kernel execution and asynchronous data transfer. This transformation enables the overlap of communication and computation. Intricate technical details that are required for this transformation are described. The third idea is that the selection of a grid geometry for the execution of a parallel region must balance the GPU occupancy with the potential saturation of the memory throughput in the GPU. Adding this additional parameter to the geometry selection heuristic can often yield better performance at lower occupancy levels.

A Case Study for Performance Portability Using OpenMP 4.5

Rahulkumar Gayatri (Lawrence Berkeley National Laboratory), Charlene Yang (Lawrence Berkeley National Laboratory)

In recent years, the HPC landscape has shifted away from traditional CPU based systems to energy efficient architectures, relying on many-core CPUs or accelerators to achieve high performance. The goal of performance portability is to enable developers to rapidly produce applications which run efficiently on a variety of these architectures and require little to no architecture specific code adoptions. Directive based programming models (OpenMP and OpenACC) are attractive in this regard as they do not require a major code restructuring and they support incremental portability.

Session 4: Using OpenACC *Randy Allen (Mentor, a Siemens Business)*

OpenACC Routine Directive Propagation Using Interprocedural Analysis
Aniket Shivam (University of California, Irvine)

Accelerator programming today requires the programmer to specify what data to place in device memory, and what code to run on the accelerator device. When programming with OpenACC, directives and clauses are used to tell the compiler what data to copy to and from the device, and what code to compile for and run on the device. In particular, the programmer inserts directives around code regions, typically loops, to identify compute constructs to be compiled for and run on the device. If the compute construct calls a procedure, that procedure also needs to marked for device compilation, as does any routine called in that procedure, and so on transitively. In addition, the marking needs to include the kind of parallelism that is exploited within the procedure, or within routines called by the procedure. When using separate compilation, the marking where the procedure is defined must be replicated in any file where it is called. This causes much frustration when first porting existing programs to GPU programming using OpenACC.

This paper presents an approach to partially automate this process. The approach relies on interprocedural analysis (IPA) to analyze OpenACC regions and procedure definitions, and to propagate the necessary information forward and backward across procedure calls spanning all the linked files, generating the required accelerator code through recompilation at link time. This approach can also perform correctness checks to prevent compilation or runtime errors. This method is implemented in the PGI OpenACC compiler.

OpenACC-Based GPU Parallelization of Plane Sweep Algorithm for Geometric Intersection
Anmol Paudel (Marquette University), Satish Puri (Marquette University)

Line segment intersection is one of the elementary operations in computational geometry. Complex problems in Geographic Information Systems (GIS) like finding map overlays or spatial joins using polygonal data require solving segment intersections. Plane sweep paradigm is used for finding geometric intersection in an efficient manner. However, it is difficult to parallelize due to its in-order processing of spatial events. We present a new fine-grained parallel algorithm for geometric intersection and its CPU and GPU implementation using OpenMP and OpenACC. To the best of our knowledge, this is the first work demonstrating an effective parallelization of plane sweep on GPUs. We chose compiler directive based approach for implementation because of its simplicity to parallelize sequential code. Using Nvidia Tesla P100 GPU, our implementation achieves around 40X speedup for line segment intersection problem on 40K and 80K data sets compared to sequential CGAL library.

5th Workshop on Accelerator Programming Using Directives (WACCPD): Closing Remarks *Sunita Chandrasekaran (University of Delaware), Sandra Wienke (RWTH Aachen University), Guido Juckeland (Helmholtz-Zentrum Dresden-Rossendorf)*
WACCPD Closing Remarks

Room: D167/174
9:00 am - 12:35 pm

Fourth International Workshop on Heterogeneous High-Performance Reconfigurable Computing (H2RC'18)

As in the previous three years, this workshop will bring together application experts, software developers, and hardware engineers, both from industry and academia, to share experiences and best practices to leverage the practical application of reconfigurable logic to Scientific Computing, Machine/Deep Learning, and “Big Data” applications. In particular, the workshop will focus on sharing experiences and techniques for accelerating applications and/or improving energy efficiency with FPGAs using OpenCL, OpenMP, OpenACC, SYCL, C, C++, and other high-level design flows, which enable and improve cross-platform functional and performance portability while also improving productivity. Particular emphasis is given to cross-platform comparisons and combinations that foster a better understanding within the industry and research community on what are the best mappings of applications to a diverse range of hardware architectures that are available today (e.g., FPGA, GPU, Many-cores and hybrid devices, ASICs), and on how to most effectively achieve cross-platform compatibility.

Preserving Privacy through Processing Encrypted Data *Miriam Leeser (Northeastern University)*
Secure Function Evaluation (SFE) allows an interested party to evaluate a function over private data without learning anything about the inputs other than the outcome of this computation. This offers a strong privacy guarantee: SFE enables, e.g., a medical researcher, a statistician, or a data analyst, to conduct a study over private, sensitive data, without jeopardizing the privacy of the study's participants (patients, online users, etc.). Nevertheless, applying SFE to “big data” poses several challenges, most significantly in the excessive processing time for applications.

In this talk, I describe Garbled Circuits (GCs), a technique for implementing SFE that can be applied to any problem that can be described as a Boolean circuit. GC is a particularly good application to accelerate with FPGAs due to the good match between GC implementations and FPGA circuits. As our goal is to use GC for extremely large problems, including machine learning algorithms, we propose to address these problems by running GCs on clusters of machines equipped with FPGAs in the datacenter to accelerate the processing. In this talk, I will present our progress and challenges with this approach.

Bringing FPGAs to HPC Production Systems and Codes *Christian Plessl (Paderborn University)*
FPGA architectures and development tools have made great strides towards a platform for high-performant and energy-efficient computing, competing head to head with other processor and accelerator technologies. While we have seen the first large-scale deployments of FPGAs in public and private clouds, FPGAs still have to make inroads in general purpose HPC systems. At the Paderborn Center for Parallel Computing, we are at the forefront of this development and have recently put "Noctua" our first HPC cluster with FPGAs into production.

In this talk, I will share some of the experiences we made on our journey from the planning, to the procurement to the installation of the Noctua cluster and highlight critical aspects for FPGAs and how

we addressed them. Further, I will present results from on-going work to port libraries and MPI-parallel HPC codes to the 32 Intel Stratix 10 FPGA boards in our cluster.

H2RC'18– Workshop Morning Break *Jason Bakos (University of South Carolina)*

SimBSP: Enabling RTL Simulation for Intel FPGA OpenCL Kernels
Martin C. Herbordt (Boston University)

RTL simulation is an integral step in FPGA development since it provides cycle accurate information regarding the behavior and performance of custom architectures, without having to compile the design to actual hardware. Despite its advantages, however, RTL simulation is not currently supported by a number of commercial FPGA OpenCL toolflows, including Intel OpenCL SDK for FPGAs (IOCLF). Obtaining reliable performance values for OpenCL kernels requires a full compilation to hardware, while emulation can only provide functional verification of the C code. Thus, development and optimization time-frames for IOCLF designs can be on the order of days, even for simple applications. In this work, we present our custom Board Support Package for IOCLF, called SimBSP, which enables OpenCL kernels to be compiled for RTL simulation.

We provide details regarding the standard kernel ports created by the IOCLF compiler, which can be used by testbenches to interface the generated design. We also list the addresses and descriptions of configuration registers that are used to set kernel parameters and provide a start trigger. Finally, we present details of SimBSP toolflow, which is integrated into the standard IOCLF and automates the process of generating kernel HDL and testbenches, and setting up the simulation environment. Our work on SimBSP will be made available Open Source to drive a community effort towards further improving the toolflow.

Scalable FPGA Deployments for HPC and DC Applications *Christoph Hagleitner (IBM)*
FPGAs have recently found their way and niche in large-scale data-center (DC) applications, eg, for endpoint encryption/compression, video transcoding, and genomics applications. We present two research projects that address two remaining roadblocks on the way to scalable performance and energy-efficiency gains: The cloudFPGA project proposes a disagregated FPGA architecture for scale-out applications and near-memory computing project uses openCAPI-attached FPGAs to tear down the "memory wall" for HPC and DC applications.

First Steps in Porting the LFRic Weather and Climate Model to the FPGAs of the EuroExa Architecture
Mike Ashworth (University of Manchester)

The EuroExa project proposes a High-Performance Computing (HPC) architecture which is both scalable to exascale performance levels and delivers world-leading power efficiency. This is achieved through the use of low-power ARM processors accelerated by closely-coupled FPGA programmable components. In order to demonstrate the efficacy of the design, the EuroExa project includes application porting work across a rich set of applications. One such application is the new weather and climate model, LFRic (named in honor of Lewis Fry Richardson), which is being developed by the UK Met Office and its partners for operational deployment in the middle of the next decade.

Much of the run-time of the LFRic model consists of compute intensive operations which are suitable for acceleration using FPGAs. Programming methods for such high-performance numerical workloads are still immature for FPGAs compared with traditional HPC architectures. The paper describes the porting of a matrix-vector kernel using the Xilinx Vivado toolset, including High-Level Synthesis (HLS), discusses the benefits of a range of optimizations and reports performance achieved on the Xilinx UltraScale+ SoC.

Performance results are reported for the FPGA code and compared with single socket OpenMP performance on an Intel Broadwell CPU. We find the performance of the FPGA to be competitive when taking into account price and power consumption.

Integrating Network-Attached FPGAs into the Cloud Using Partial Reconfiguration
Alexander Ditter (University of Erlangen-Nuremberg)

Accelerating Intelligence *John Davis (Bigstream Networks)*
Massive amounts of data are being consumed and processed to drive business. The exponential increase in data has not been matched by the computational power of processors. This has led to the rise of accelerators. However, big data algorithms for ETL, ML, AI, and DL are evolving rapidly and/or have significant diversity. These moving targets are poor candidates for ASICs, but match the capabilities and flexibility of FPGAs. Furthermore, FPGAs provide a platform to move computation to the data, away from the CPU, by providing computation at line rate at the network and/or storage. Bigstream is bridging the gap between high-level big data frameworks and accelerators using our Hyper-acceleration Layer built on top of SDAccel. In this talk, we will describe the Bigstream Hyper-acceleration that automatically provides computational acceleration, by the CPU, storage, or network, for big data platforms with zero code change.

The MANGO Process for Designing and Programming Multi-Accelerator Multi-FPGA Systems
Rafael Tornero (Technical University of Valencia)

This paper describes the approach followed in the European FETHPC MANGO project to design and program systems made of multiple FPGAs interconnected. The MANGO approach relies on the instantiation and management of multiple generic and custom-made accelerators which can be programmed to communicate each other via shared memory and through synchronization registers. The paper introduces the low level architecture including the multi-FPGA interconnect deployed, the communication protocol and the architectural template-based approach to simplify the design process.

Stream Computing of Lattice-Boltzmann Method on Intel Programmable Accelerator Card
Takaaki Miyajima (RIKEN)

Intel Programmable Accelerator Card (Intel-PAC) and Open Programmable Acceleration Engine (OPAE) aim at saving developers time and enabling code re-use across multiple FPGA platforms. We implemented a Lattice-Boltzmann Method (LBM) computing core, a computational fluid dynamics application, on Intel-PAC. We evaluated four designs of LBM core. In our evaluation of four LBM cores,

sustained performance of each design is the same as the theoretical peak performance.

Room: D161
9:00 am - 5:30 pm

The 3rd International Workshop on Post-Moore Era Supercomputing (PMES)

Session Description: Session 1

The 3rd International Workshop on Post Moore's Era Supercomputing (PMES) follows the very successful PMES workshops at SC16 and SC17. This interdisciplinary workshop is organized to explore the scientific issues, challenges, and opportunities for supercomputing beyond the scaling limits of Moore's Law, with the ultimate goal of keeping supercomputing at the forefront of computing technologies beyond the physical and conceptual limits of current systems. Continuing progress of supercomputing beyond the scaling limits of Moore's Law is likely to require a comprehensive re-thinking of technologies, ranging from innovative materials and devices, circuits, system architectures, programming systems, system software, and applications.

The workshop is designed to foster interdisciplinary dialog across the necessary spectrum of stakeholders: applications, algorithms, software, and hardware. Motivating workshop questions will include the following. "What technologies might prevail in the Post Moore's Era?" "How can applications effectively prepare for these changes through co-design?" "What architectural abstractions should be in place to represent the traditional concepts like hierarchical parallelism, multi-tier data locality, and new concepts like variable precision, approximate solutions, and resource tradeoff directives?" "What programming models might insulate applications from these changes?"

Tranquility Amidst Turbulence: A Vision for Advancing Scientific Discovery in the Era of Extreme Heterogeneity *Lucy Nowell (US Department of Energy Office of Advanced Scientific Computing Research)*
Rapid changes in computing technology are driving ever-faster changes in the way scientific discoveries are made. Dr. Lucy Nowell will discuss the dimensions of heterogeneity that are expected to impact the software stack for computational and data-intensive in the 2025-2035 timeframe. In particular, she will discuss her vision for the research that is needed to keep technological turbulence from disrupting scientific progress.

PMES – Workshop Morning Break

Hybrid Quantum-Classical Computing Architectures
Martin Suchara (Argonne National Laboratory)

We describe how classical supercomputing can aid unreliable quantum processors of intermediate size to solve large problem instances reliably. We advocate using a hybrid quantum-classical architecture where larger quantum circuits are broken into smaller sub-circuits that are evaluated separately, either using a quantum processor or a quantum simulator running on a classical supercomputer. Circuit compilation techniques that determine which qubits are simulated classically will greatly impact the system performance as well as provide a tradeoff between circuit reliability and runtime.

Stochastic Computing on Quantum Gates

Yoshito Kanamori (University of Alaska, Anchorage)

The concept of Stochastic Computing is applied to Quantum Computing. A qubit in a superposition state is used to represent a probability as an input for Stochastic Computation, instead of a random bit stream. Measuring the output qubit from the quantum circuit produces zero or one with the probability of the computation result. This process is repeated N times to generate the N-bit stochastic number, where N is determined by the required accuracy.

Memory-Efficient Quantum Circuit Simulation by Using Lossy Data Compression

Xin-Chuan Wu (University of Chicago, Argonne National Laboratory)

In order to evaluate, validate, and refine the design of new quantum algorithms or quantum computers, researchers and developers need methods to assess their correctness and fidelity. This requires the capabilities of quantum circuit simulations. However, the number of quantum state amplitudes increases exponentially with the number of qubits, leading to the exponential growth of the memory requirement for the simulations.

In this work, we present our memory-efficient quantum circuit simulation by using lossy data compression. Our empirical data shows that we reduce the memory requirement to 16.5% and 2.24E-06 of the original requirement for QFT and Grover's search, respectively. This finding further suggests that we can simulate deep quantum circuits up to 63 qubits with 0.8 petabytes memory.

Community Detection Across Emerging Quantum Architectures

Ruslan Shaydulin (Clemson University)

One of the roadmap plans for quantum computers is an integration within HPC ecosystems assigning them a role of accelerators for a variety of computationally hard tasks. However, in the near term, quantum hardware will be in a constant state of change. Heading towards solving real-world problems, we advocate development of portable, architecture-agnostic hybrid quantum-classical frameworks and demonstrate one for the community detection problem evaluated using quantum annealing and gate-based universal quantum computation paradigms.

Comparing Deep Learning with Quantum Inference on The D-Wave 2X

Nga Nguyen (Los Alamos National Laboratory)

We used a quantum annealing D-Wave 2X computer to obtain solutions to NP-hard sparse coding

problems for inferring representation of reduced dimensional MNIST images. For comparison, we implemented two deep neural network architectures. The first (AlexNet-like) approximately matched the architecture of the sparse coding model. The second was state-of-the-art (RESNET). Classification based on the D-Wave 2X was superior to matching pursuit and AlexNet and nearly equivalent to RESNET.

Speaker Panel 1

PMES – Workshop Lunch (on your own)

Doing Moore with Less – Leapfrogging Moore's Law with Inexactness for Supercomputing
Sven Leyffer (Argonne National Laboratory)

We investigate the use of inexactness, to trade solution quality with energy savings to reduce power consumption in HPC systems, using commercial off-the-shelf processors for lower-precision arithmetic. We implement an inexact Newton algorithm for solving nonlinear equations to show how to reinvest energy savings to improve solution quality. We provide an analytical model to explain our findings, and provide a justification for using nonstandard half-precision arithmetic for ill-conditioned problems.

A Task-Based Abstraction Layer for User Productivity and Performance Portability in Post-Moore's Era Supercomputing
Steve Petruzza (University of Utah)

The proliferation of heterogeneous computing architectures in current and future supercomputing systems dramatically increases the complexity of software development and exacerbates the divergence of software stacks. Currently, task-based runtimes attempt to alleviate these impediments, however their effective use requires expertise and deep integration that does not facilitate reuse and portability. We propose to introduce a task-based abstraction layer that separates the definition of the algorithm from the runtime-specific implementation, while maintaining performance portability.

Non-Neural Network Applications for Spiking Neuromorphic Hardware
Catherine D. Schuman (Oak Ridge National Laboratory)

Increasing power costs for large-scale computing in a post-Moore's Law system have forced the high-performance computing community to explore heterogeneous systems. Neuromorphic architectures, inspired by biological neural systems, have so far been relegated to auxiliary machine learning applications. Here, we discuss growing research showing the viability of ultra-low-power neural accelerators as co-processors for classic compute algorithms, such as random walk simulations and graph analytics.

PMES – Workshop Afternoon Break

Shortest Path and Neighborhood Subgraph Extraction on a Spiking Memristive Neuromorphic Implementation
Catherine Schuman (Oak Ridge National Laboratory)

Spiking neuromorphic computers (SNCs) are promising as a post Moore's law technology because of their potential for very low power computation. SNCs have primarily been demonstrated on machine learning applications, but they can also be used for applications beyond machine learning. Here, we demonstrate two graph problems (shortest path and neighborhood subgraph extraction) that can be solved using SNCs. We estimate the performance of a memristive SNC for these applications on three real-world graphs.

Speaker Panel 2

Domain-Specific System on Chips (DSSoC) *Tom Rondeau (Defense Advanced Research Projects Agency (DARPA), Domain-Specific System on Chip (DSSoC) Program)*
The invention of the digital computer came about as a proof of computable numbers, and so numerical processing was solved by the Turing Machine that has led to the general purpose computer. These computers are good at solving multiple types of problems with a single machine. However, within the scope of computable numbers are specific computational problems where purpose-built machines can solve them faster while using less energy. An example of this is the digital signal processor (DSP) that performs the multiply and accumulate operation important to signal processing, or the more recent development of the tensor processing unit (TPU) to compute dense matrix multiplies that are at the core of deep neural networks. These specialized machines solve a single problem but with optimized efficiency.

As a result, there exists a tension between the flexibility in general purpose processors and the efficiency of specialized processors. Domain-Specific System on Chip (DSSoC) intends to demonstrate that the tradeoff between flexibility and efficiency is not fundamental. The program will develop a method for determining the right amount and type of specialization while making a system as programmable and flexible as possible. It will take a vertical view of integrating today's programming environment with a goal of full-stack integration. DSSoC will de-couple the programmer from the underlying hardware with enough abstraction but still be able to utilize the hardware optimally through intelligent scheduling. DSSoC specifically targets embedded systems where the domain of applications sits at the edge and near the sensor. Workloads consist of small chunks of data but often with a large number of algorithms required in the processing, meaning that high compute power and low latency at low power are required.

This presentation will explore the concepts being developed in the DSSoC program, including an overview of the projects being funded to study the research questions and produce technology that furthers industry's ability to effectively use specialization within our SoCs.

Workshop Closing

Room: D163
9:00 am - 5:30 pm

Innovating the Network for Data Intensive Science (INDIS)

Wide area networks are now an integral and essential part of this data-driven supercomputing ecosystem connecting information sources, data stores, processing, simulation, visualization and user communities together. Networks for data-intensive science have more extreme requirements than general-purpose networks. These requirements not only closely impact the design of processor interconnects in supercomputers and cluster computers, but they also impact campus networks, regional networks and national backbone networks. This workshop brings together the network researchers and innovators to present challenges and novel ideas that stretch network research and SC's own innovative network, SCinet. We invite papers that propose new and novel techniques to present solutions for meeting these networking needs; and developments that are essential in the information systems infrastructure for the scientific discovery process.

INDIS Invited Talk: Introduction to SCinet *Jason Zurawski (Energy Sciences Network (ESnet))*

INDIS Showcases Panel: NRE and XNET and Architecture

INDIS – Workshop Morning Break

INDIS Morning Keynote *Josh Bailey (Google LLC)*
Josh Bailey will give a keynote on FAUCET. FAUCET is an open source, production SDN controller, supporting multiple network vendors, and is deployed as part of SC18's SCinet network among other networks. This talk will be an update on FAUCET's current status and features and an informal Q&A on the state of OpenFlow implementations in practical, production settings.

Flowzilla: A Methodology for Detecting Data Transfer Anomalies in Research Networks
Anna Giannakou (Lawrence Berkeley National Laboratory)

Research networks are designed to support high volume scientific data transfers that span multiple network links. Like any other network, research networks experience anomalies. Anomalies are deviations from profiles of normality in a science network's traffic levels. Diagnosing anomalies is critical both for network operators and scientists. In this paper, we present Flowzilla, a general framework for detecting and quantifying anomalies on scientific data transfers of arbitrary size. Flowzilla incorporates Random Forest Regression for predicting the size of data transfers and utilizes an adaptive threshold mechanism for detecting outliers. Our results demonstrate that our framework achieves up to 92.5% detection accuracy. Furthermore, we are able to predict data transfer sizes up to 10 weeks after training with accuracy above 90%.

Fast Detection of Elephant Flows with Dirichlet-Categorical Inference

Aditya Gudibanda (Reservoir Labs Inc), Jordi Ros-Giralt (Reservoir Labs Inc)

The problem of elephant flow detection is a longstanding research area with the goal of quickly identifying flows in a network that are large enough to affect the quality of service of smaller flows. Past work in this field has largely been either domain-specific, based on thresholds for a specific flow size metric, or required several hyperparameters, reducing their ease of adaptation to the great variety of traffic distributions present in real-world networks. In this paper, we present an approach to elephant flow detection that avoids these limitations, utilizing the rigorous framework of Bayesian inference. By observing packets sampled from the network, we use Dirichlet-Categorical inference to calculate a posterior distribution that explicitly captures our uncertainty about the sizes of each flow. We then use this posterior distribution to find the most likely subset of elephant flows under this probabilistic model. Our algorithm rapidly converges to the optimal sampling rate at a speed O(1/n), where n is the number of packet samples received, and the only hyperparameter required is the targeted detection likelihood, defined as the probability of correctly inferring all the elephant flows. Compared to the state-of-the-art based on static sampling rate, we show a reduction in error rate by a factor of 20 times. The proposed method of Dirichlet-Categorical inference provides a novel, powerful framework to elephant flow detection that is both highly accurate and probabilistically meaningful.

Tracking Network Flows with P4

Joseph Hill (University of Amsterdam)

Tracking flows within a single device, as well as tracking the full path a flow takes in a network, are core components in securing networks. Malicious traffic can be easily identified and its source blocked. Traditional methods have performance and precision shortcomings, while new programmable devices open up new possibilities. In this paper we present methods based on the P4 programming language that allow to track flows in a device, as well methods toward full path tracking. A core component of this work are Bloom filters, which we have implemented fully in P4. To validate our approach and implementation we have carried a study in a specific use case, namely the detection of SYN attacks.

INDIS – Workshop Lunch (on your own)

SDN for End-to-End Networked Science at the Exascale (SENSE)

Inder Monga (Lawrence Berkeley National Laboratory, Energy Sciences Network (ESnet)), Tom Lehman (University of Maryland, Mid-Atlantic Crossroads)

The Software-defined network for End-to-end Networked Science at Exascale (SENSE) research project is building smart network services to accelerate scientific discovery in the era of 'big data' driven by Exascale, cloud computing, machine learning and AI. The project's architecture, models, and demonstrated prototype define the mechanisms needed to dynamically build end-to-end virtual guaranteed networks across administrative domains, with no manual intervention. In addition, a highly intuitive 'intent' based interface, as defined by the project, allows applications to express their high-level service requirements, and an intelligent, scalable model-based software orchestrator converts that intent into appropriate network services, configured across multiple types of devices. The significance of these capabilities is the ability for science applications to manage the network as a

first-class schedulable resource akin to instruments, compute, and storage, to enable well defined and highly tuned complex workflows that require close coupling of resources spread across a vast geographic footprint such as those used in science domains like high-energy physics and basic energy sciences.

Social Computational Trust Model (SCTM): A Framework to Facilitate Selection of Partners
Ameneh Deljoo (University of Amsterdam), Cees de Laat (University of Amsterdam)

Creating a cyber security alliance among network domain owners, as a means to minimize security incidents, has gained the interest of practitioners and academics in the last few years. A cyber security alliance, like any membership organization, requires the creation and maintenance of trust among its members, in this case the network domain owners. To promote the disclosure and sharing of cyber security information among the network domain owners, a trust framework is needed.

This paper discusses a social computational trust model (SCTM), that helps alliance members to select the right partner to collaborate with and perform collective tasks, and encourages the sharing of incident data and intelligence. The social computational trust model combines benevolence and competence to estimate the risk of interaction. Benevolence is computed from personal experiences gained through direct interactions and competence is assessed on the base of the received feedback from the other members. An agent based model case study is presented to demonstrate our approach. The practicability of the proposed risk estimation is validated with a detailed experiment.

INDIS Afternoon Keynote *Kate Keahey (Argonne National Laboratory)*
Kate Keahey will give the second keynote on "Chameleon: New Capabilities for Experimental Computer Science". Chameleon is a large-scale, deeply reconfigurable testbed built specifically to explore a broad range of different state-of-the-art hardware options, assess scalability of systems, and provide conditions that allow deep reconfigurability and isolation so that one user does not impact the experiments of another. It currently consists of almost 20,000 cores, over 5PBs of total disk space hosted at the University of Chicago and TACC, and leverages 100 Gbps connection between the sites. The hardware includes a large-scale homogenous compute partition, as well a diversity of configurations and architectures including Infiniband, GPUs, FPGAs, storage hierarchies with a mix of HDDs, SDDs, NVRAM, and high memory as well as non-x86 architectures such as ARMs and Atoms. To support systems experiments, Chameleon provides a configuration system giving users full control of the software stack including root privileges, kernel customization, and console access. To date, Chameleon has supported 2,700+ users working on over 350+ projects.

This talk will describe the evolution of the testbed as well as the current work towards broadening the range of supported experiments. In particular, I will discuss recently deployed hardware and new networking capabilities allowing experimenters to deploy their own switch controllers and experiment with Software Defined Networking (SDN). I will also describe new capabilities targeted at improving experiment management, monitoring, and analysis as well as tying together testbed features to improve experiment repeatability. Finally, I will outline our plans for packaging the Chameleon infrastructure to allow others to reproduce its configuration easily and thereby making the process of configuring a CS testbed more sustainable.

Bandwidth Scheduling for Big Data Transfer with Deadline Constraint between Data Centers

Aiqin Hou (Northwest University, China), Chase Q. Wu (New Jersey Institute of Technology)

An increasing number of applications in scientific and other domains have moved or are in active transition to clouds, and the demand for the movement of big data between geographically distributed cloud-based data centers is rapidly growing. Many modern backbone networks leverage logically centralized controllers based on software-defined networking (SDN) to provide advance bandwidth reservation for data transfer requests. How to fully utilize the bandwidth resources of the links connecting data centers with guaranteed QoS for each user request is an important problem for cloud service providers. Most existing work focuses on bandwidth scheduling for a single request for data transfer or multiple requests using the same service model. In this work, we construct rigorous cost models to quantify user satisfaction degree and formulate a generic problem of bandwidth scheduling for multiple deadline-constrained data transfer requests of different types to maximize the request scheduling success ratio while minimizing the data transfer completion time of each request. We prove this problem to be NP-complete and design a heuristic solution. Extensive simulation results show that our scheduling scheme significantly outperforms existing methods in terms of user satisfaction degree and scheduling success ratio.

INDIS – Workshop Afternoon Break

Analysis of CPU Pinning and Storage Configuration in 100 Gbps Network Data Transfer

Se-young Yu (International Center for Advanced Internet Research (iCAIR), Northwestern University)

A common bottleneck for high-speed network data transfers is lack of CPU resources. A number of techniques and solutions have been proposed to reduce CPU load for data transfer. One can optimize the core affinity settings in their Non-Uniform Memory Access (NUMA) system and use NVMe over Fabrics to avoid CPU bottlenecks in high-speed network data transfers. Our assumption is that binding processes to the local processor improves the overall performance of the high-speed network data transfers compared to binding the processes to actual cores or leaving them unbounded. Furthermore, using NVMe over Fabrics reduces the CPU utilization more with a lower number of processors. To evaluate these assumptions, we performed a series of experiments with different core affinity and storage settings. We found evidence that binding processes to the local processor instead of the cores improve the file transfer performance for most of the use-cases and NVMe over Fabrics is more efficient in transferring files compared to traditional file transfers in Local Area Networks (LANs). We were able to achieve the maximum SSD performance threshold using 32 transfer processes with traditional file transfers while using 8 processes with NVMe over Fabrics and reduced CPU utilization.

BigData Express: Toward Schedulable, Predictable, and High-Performance Data Transfer

Qiming Lu (Fermi National Accelerator Laboratory), Wenji Wu (Fermi National Accelerator Laboratory), Phil Demar (Fermi National Accelerator Laboratory), Chin Guok (Energy Sciences Network (ESnet)), Se-young Yu (Northwestern University), Jim-hao Chen (Northwestern University), Joe Mambretti (Northwestern University), Jim Kim (Korea Advanced Institute of Science and Technology), Xi Yang (University of Maryland), Tom Lehman (University of Maryland), Gary Liu (New Jersey Institute of Technology)

Big Data has emerged as a driving force for scientific discoveries. Large scientific instruments (e.g.,

colliders, and telescopes) generate exponentially increasing volumes of data. To enable scientific discovery, science data must be collected, indexed, archived, shared, and analyzed, typically in a widely distributed, highly collaborative manner. Data transfer is now an essential function for science discoveries, particularly within big data environments. Although significant improvements have been made in the area of bulk data transfer, the currently available data transfer tools and services can not successfully address the high-performance and time-constraint challenges of data transfer required by extreme-scale science applications for the following reasons: disjoint end-to-end data transfer loops, cross-interference between data transfers, and existing data transfer tools and services are oblivious to user requirements (deadline and QoS requirements). Fermilab has been working on the BigData Express project to address these problems. BigData Express seeks to provide a schedulable, predictable, and high-performance data transfer service for big data science. The BigData Express software is being deployed and evaluated at multiple research institutions, which include UMD, StarLight, FNAL, KISTI, KSTAR, SURFnet, Ciena, and other sites. Meanwhile, the BigData Express research team is collaborating with the StarLight International/National Communications Exchange Facility to deploy BigData Express at various research platforms, including Pacific Research Platform, National Research Platform, and Global Research Platform. It is envisioned that we are working toward building a high-performance data transfer federation for big data science.

Room: D165
9:00 am - 5:30 pm

Fourth Computational Approaches for Cancer Workshop (CAFCW18)

As the drive towards precision medicine has accelerated, the opportunities and challenges in using computational approaches in cancer research and clinical application are rapidly growing. The expanding development of new approaches are reshaping the way computation is being applied in cancer applications. Recent national level initiatives all underscore the importance of a workshop that brings together experts and insights across the spectrum of computational approaches for cancer.

In the workshop, we bring together the computational community exploring and using high-performance computing, analytics, predictive modeling, and large datasets in cancer research and clinical applications. The workshop is inherently inter-disciplinary, with the common interest in cancer and computation the unifying theme. Maintaining a perspective of accelerating scientific insights and translation of insights to clinical application for improved patient outcomes, the workshop brings together many interests from across the technology, cancer research and clinical domains.

In addition to broad topic submissions for collaborative submissions involving computation and cancer domain expertise, the CAFCW workshop annually identifies a special workshop focus for the year. With the critical importance on both sensitive and large amounts of data in cancer applications, the rapidly evolving use of new technologies such as machine learning, and the simultaneous drive to improve patient outcomes, the interest in approaches that improve portability and repeatability to foster rapid clinical translation is significant within the cancer community. This special topic session will include innovative solutions in portability and repeatability of computational approaches applied to

cancer, with emphasis on clinical translation.

Morning Keynote – Computational Approaches in Clinical Applications *Ketan Paranjape (Roche - Diagnostic Information Solutions)*

CAFCW18 – Workshop Morning Break

Developing a Reproducible WDL-Based Workflow for RNASeq Data Using Modular, Software Engineering-Based Approaches
Steven Eschrich (Moffitt Cancer Center)

Computational workflows have become standard in many disciplines, including bioinformatics and genomics. Workflow languages, such as the Workflow Description Language (WDL) and Common Workflow Language (CWL) have been developed to express workflow processing syntax. These languages can be highly expressive and customizable however this can result in perpetuating the complex tangle of code that can be difficult to maintain and comprehend. The Moffitt Cancer Center participates in the ORIEN Avatar project, a multi-center project that has generated molecular profiles (DNASeq, RNASeq) on ~1,000 tissues to date. To support reproducibility in the analysis of RNASeq data for this project, we have implemented an RNA Sequencing Genomics analysis pipeline using Cromwell, a WDL-based workflow engine, in our HPC environment. Constraining the language to specific structural conventions and emphasizing modularity, we have built a pipeline suitable for operational purposes and maintainability. We implemented individual tasks with built-in unit testing and nested levels of workflow integration for successively complex integration testing. This pipeline has been successfully used by bioinformatics staff at Moffitt Cancer Center with minimal training.

Safety, Reproducibility, Performance: Accelerating Cancer Drug Discovery with Cloud, ML, and HPC Technologies
Amanda J. Minnich (Lawrence Livermore National Laboratory)

New computational opportunities and challenges have emerged within the cancer research and clinical application areas as the size, number, variety and complexity of cancer datasets have grown in recent years. Simultaneously, advances in computational capabilities have grown and are expected to continue to reach unprecedented scales. Such opportunities to impact cancer computationally are underscored in the 2016 Twenty-first Century Cures Act. The workshop focuses on bringing together interested individuals ranging from clinicians, mathematicians, data scientists, computational scientists, hardware experts, engineers, developers, leaders and others with an interest in advancing the use of computation at all levels to better understand, diagnose, treat and prevent cancer. With an interdisciplinary focus, the workshop provides opportunities for participants to learn how computation is employed across multiple areas including imaging, genomics, analytics, modeling, pathology and drug discovery with a focus on impacting cancer. As an interdisciplinary workshop, the cross-disciplinary sharing of insight and challenges fosters collaborations and future innovations to accelerate the progress in computationally and data driven cancer research and clinical applications. The forward focus of the workshop looks at challenges and opportunities for large scale HPC, including exascale applications involving cancer.

The Gen3 Approach to Portability and Repeatability for Cancer Genomics Projects
Robert L. Grossman (University of Chicago)

The Gen3 software stack is a open-source platform for managing, analyzing, and sharing petabyte-scale research data. In this note, we describe the approach that we have used with Gen3 to support portability and repeatibility for cancer genomics projects. Data in a Gen3 data commons is divided into projects. Project data is of two types: large files, such as BAM files and image files, that are managed as data objects and stored in one or more private and public clouds, and all of the other data associated with a project, including all of the the clinical phenotype data and biospecimen data. We call this other data "core data" and have developed data serialization format for it, which includes versioning and schema information. Data objects are available across multiple data commons, while core data can be exported and imported using the serialization format. In this way, we support portability for data projects. We support repeatibility by representing workflows using the Common Workflow Language (CWL) and managing the CWL files as data objects. With this approach, we simply need to manage and version the data objects, core data, and CWL files associated with a project.

Scalable Deep Ensemble Learning for Cancer Drug Discovery
Samson Jacobs (Lawrence Livermore National Laboratory)

In this work, we demonstrate how the Livermore Tournament Fast Batch (LTFB) ensemble algorithm is able to efficiently tune hyperparameters and accelerate the time to solution for several cancer drug discovery networks. Drawn from the DOE-NCI Pilot 1 and ECP CANDLE projects we show significantly improved training quality for the "Uno" data set and associated network and a dramatic reduction in the wall-clock time for training the "Combo" network to a fixed level of convergence. LTFB is an ensemble method that creates a set of neural network models and trains each instance of these models in parallel. Periodically, each model selects another model to pair with, exchanges models, and then run a local tournament against held-out tournament datasets. The winning model will continue training on the local training datasets. LTFB is implemented in the Livermore Big Artificial Neural Network toolkit (LBANN), a toolkit optimized for composing multiple levels of parallelism on HPC architectures.

Panel Discussion: Reproducibility and Accessibility - Challenges and Opportunities
Interactive panel discussion on frontiers and perspectives for computation in cancer research and clinical applications

CAFCW18 – Workshop Lunch (on your own)

Afternoon Keynote – Genomic Profiling of Normal, Premalignant, and Heterogeneous Tissues in Cancer Patients *Paul Scheet (MD Anderson Cancer Center)*
Normal tissues adjacent to tumor and premalignant lesions present an opportunity for in vivo human models of early disease pathology. Genomic studies of such "at risk" tissues may identify molecular pathways involved in a transition to malignant phenotypes and/or targets for personalized prevention or precision medicine. Yet, challenges to this objective include: 1) the small size of lesions or limited available tissue, often presenting "either or" choices for molecular technologies (e.g. DNA or RNA,

NGS or arrays); and 2) low mutant cell fractions due to heterogeneous tissues and their corresponding early stages of disease. To address these, we have 1) conducted targeted next-generation sequencing of DNA and RNA and, when possible, genome-wide DNA SNP arrays, 2) considered various ensemble strategies for off-the-shelf single-nucleotide variant calling algorithms to determine mutations, and 3) developed sensitive haplotype-based techniques (hapLOH) to determine megabase-scale regions of allelic imbalance that reflect chromosomal deletions, duplications and copy-neutral loss-of-heterozygosity. We have applied combinations of these strategies to normal appearing epithelial airway samples adjacent to non-small cell lung cancers, premalignant tissues, and to public data from paired normal and tumor samples from 11,000 patients of The Cancer Genome Atlas (TCGA). In mutational analyses of tissues annotated by alterations discovered in paired tumors, we identify key drivers, document two-hit models of tumorigenesis, highlight immune-related expression phenotypes in premalignant lesions and construct phylogenetic trees of intra-patient samples. We also give examples of systematic errors in copy number changes in TCGA that can be corrected by hapLOH.

CAFCW18 – Workshop Afternoon Break

HPC-Based Hyperparameter Search of MT-CNN for Information Extraction from Cancer Pathology Reports

Hong-Jun Yoon (Oak Ridge National Laboratory)

Finding optimal hyperparameters is necessary to identify the best performing deep learning models, but the process is costly. In this paper, we applied model-based optimization, also known as Bayesian optimization, using the CANDLE framework implemented on a High-Performance Computing environment. As a use case, we selected information extraction from cancer pathology reports using a multi-task convolutional neural network, a model with 10 hyperparameters to be optimized. We utilized a synthesized text corpus of 8,000 training cases and 2,000 validation cases with four types of clinical task labels including primary cancer site, laterality, behavior, and histological grade. We conducted the experiments on the Titan supercomputer at the Oak Ridge Leadership Computing Facility (OLCF), reported the optimal hyperparameters found, and demonstrated that hyperparameter optimization using the CANDLE framework is a feasible approach with respect to both scalability and clinical task performance.

Toward a Computational Simulation of Circulating Tumor Cell Transport in Vascular Geometries

John Gounley (Duke University)

Computational models can provide much needed insight into the mechanisms driving cancer cell trajectory. However, capabilities must be expanded to enable simulations in larger sections of micro- and meso-vasculature and account for the more complex fluid dynamic patterns that occur in patient-derived vascular geometries. The increased size and complexity of these simulations demands, in turn, the development of highly flexible and scalable computational frameworks. In this work, we extend the massively parallel hemodynamics solver HARVEY to perform the necessary fluid-structure interaction for CTC transport in high-resolution blood flow. We couple lattice Boltzmann and finite element methods, for the fluid and cells, respectively, with the immersed boundary method for the fluid-structure interaction. Parallelized with MPI, HARVEY is designed to handle the sparse and tortuous blood vessels encountered in patient-derived vascular geometries while maintaining computational

efficiency. HARVEY can be scaled to simulate vasculature geometries containing hundreds of millions of blood cells, equivalent to blood volumes on the order of tens of milliliters. In sum, the resulting framework has the potential to significantly improve the model fidelity of CTC simulations with respect to both the complexity and size of vascular geometries being considered.

Hummingbird: Efficient Performance Prediction for Executing Genomics Applications in the Cloud
Utsab Ray (North Carolina State University)

A major drawback of executing existing genomics pipelines on cloud computing facilities is that the onus of efficiently executing it on the best configuration lies on the user. Lack of knowledge regarding which cloud configuration is best to execute a pipeline often results in an unnecessary increase in cost due to selecting a more expensive cloud tier than needed. Resources in the cloud are expensive, so determining the best configuration before actually running the pipeline saves money and time. To this end, we introduce Hummingbird, a framework that predicts the best configuration to execute genomics pipelines on Google cloud.

Toward a Pre-Cancer Image Atlas through Crowdsourcing and Machine Learning
Ashish Mahabal (California Institute of Technology)

We describe how crowdsourcing can be combined with advanced machine learning for early cancer detection. We demonstrate our system for lung cancer (using data from the National Lung cancer Screen Trial), but in such a fashion that it can easily be replicated for other organs. Thus this becomes a step towards the pre-cancer image atlas, a focus area of the National Cancer Institute. Additionally, by keeping explainability and interpretability at the core of our deep learning methods, we make the endeavor reproducible.

Extending Frontiers for Computing in Cancer – Special Session
Provide a glimpse into the latest developments at the frontiers of computing in cancer. Highlight efforts underway with Joint Design of Advanced Computing Solutions for Cancer and other efforts at the cutting-edge of cancer research and high performance computing.

Room: D170
9:00 am - 5:30 pm

MCHPC'18: Workshop on Memory Centric High Performance Computing

The growing disparity between CPU speed and memory speed, known as the memory wall problem, has been one of the most critical and long-standing challenges in the computing industry. The situation is further complicated by the recent expansion of the memory hierarchy, which is becoming deeper and more diversified with the adoption of new memory technologies and architectures including 3D-stacked memory, non-volatile random-access memory (NVRAM), memristor, hybrid

software and hardware caches, etc. Computer architecture and hardware system, operating systems, storage and file systems, programming stack, performance model and tools are being enhanced, augmented, or even redesigned to address the performance, programmability and energy efficiency challenges of the increasingly complex and heterogeneous memory systems for HPC and data-intensive applications.

The MCHPC workshop aims to bring together computer and computational science researchers, from industry, government labs and academia, concerned with the challenges of efficiently using existing and emerging memory systems.

MCHPC'18 Morning Keynote: Converging Storage and Memory *Frank Hady (Intel Corporation)*
Order of magnitude advances in non-volatile memory density and performance are upon us bringing significant systems level architecture opportunities. The NAND Memory transition to 3D and the introduction of QLC have recently increased NAND SSD storage density at a very rapid pace. Products featuring one terabit per die are available from Intel® Corporation allowing dense storage, for example one PByte in 1U. This large improvement in density brings great value to systems, but also increases the performance/capacity/cost gap between DRAM and storage within the long evolving memory and storage hierarchy. Intel® 3D XPoint™ Memory, with much higher performance than NAND and greater density than DRAM has entered the platform to address this gap - first as SSDs. These Intel® Optane™ SSDs are in use within client and data center platforms as both fast storage volumes and as paged extensions to system memory delivering significant application performance improvements. With low latency and fine grained addressability, this new memory can be accessed as Persistent Memory (PM), avoiding the 4kByte block size and multiple microsecond storage stack that accompany system storage. This Intel® Optane Data Center Persistent Memory is made possible through a series of hardware and software advances. The resulting high capacity, high performance, persistent memory creates opportunities for rethinking algorithms to deliver much higher performance applications. This presentation will explain these new memory technologies, explore their impact on the computing system at the architecture and solution level, and suggest areas of platform exploration relevant to the HPC community.

MCHPC'18 – Workshop Morning Break

Challenges of High-Capacity DRAM Stacks and Potential Directions

With rapid growth in data volumes and an increase in number of CPU/GPU cores per chip, the capacity and bandwidth of main memory can be scaled up to accommodate performance requirements of data-intensive applications. Recent 3D-stacked in-package memory devices such as high-bandwidth memory (HBM) and similar technologies can provide high amounts of memory bandwidth at low access energy. However, 3D-stacked in-package memory have limited memory capacity. In this paper, we study and present challenges of scaling the capacity of 3D-stacked memory devices by stacking more DRAM dies within a device and building taller memory stacks. We also present potential directions and mitigations to building tall HBM stacks of DRAM dies. Although taller stacks are a potentially interesting approach to increase HBM capacity, we show that more research is needed to enable high-capacity memory stacks while simultaneously scaling up their memory bandwidth. Specifically, alternative bonding and stacking technologies can be investigated as a

potentially major enabler of tall HBM stacks.

Evaluation of Intel Memory Drive Technology Performance for Scientific Applications

In this paper, we present benchmark data for Intel Memory Drive Technology (IMDT), which is a new generation of Software-defined Memory (SDM) based on Intel ScaleMP collaboration and using 3D XPoint TM based Intel Solid-State Drives (SSDs) called Optane. We studied IMDT performance for synthetic benchmarks, scientific kernels, and applications. We chose these benchmarks to represent different patterns for computation and accessing data on disks and memory. To put performance of IMDT in comparison, we used two memory configurations: hybrid IMDT DDR4/Optane and DDR4 only systems. The performance was measured as a percentage of used memory and analyzed in detail. We found that for some applications DDR4/Optane hybrid configuration outperforms DDR4 setup by up to 20%

xBGAS: Toward a RISC-V ISA Extension for Global, Scalable, Shared Memory

Given the switch from monolithic architectures to integrated systems of commodity components, scalable high performance computing architectures often suffer from unwanted latencies when operations depart an individual device domain. Transferring control and/or data across loosely coupled commodity devices implies a certain degree of cooperating in the form of complex system software. The end result being a total system architecture the operates with low degrees of efficiency.

This work presents initial research into creating micro architecture extensions to the RISC-V instruction set that provide tightly coupled support for common high performance computing operations. This xBGAS micro architecture extension provides applications the ability to access globally shared memory blocks directly from rudimentary instructions. The end result is a highly efficient micro architecture for scalable shared memory programming environments.

Understanding Application Recomputability without Crash Consistency in Non-Volatile Memory

Emerging non-volatile memory (NVM) is promising to be used as main memory, because of its good performance, density, and energy efficiency. Leveraging the non-volatility of NVM as main memory, we can recover data objects and resume application computation (recomputation) after application crashes. The existing work studies how to ensure that data objects stored in NVM can be recovered to a consistent version during system recovery, a property referred to as crash consistency. However, enabling crash consistency often requires program modification and brings large runtime overhead.

In this paper, we use a different view to examine application recomputation in NVM. Without taking care of consistency of data objects, we aim to understand if the application can be recomputable, given possible inconsistent data objects in NVM. We introduce a PIN-based simulation tool, NVC, to study application recomputability in NVM without crash consistency. The tool allows the user to randomly trigger application crash and then perform postmortem analysis (i.e., the analysis on data

consistency) on data values in caches and memory. We use NVC to study a set of applications. We reveal that some applications are inherently tolerant to crash consistency. We perform a detailed analysis of the reasons. We study an optimization technique to accelerate the simulation performance of NVC. The technique allows us to use NVC to study data-intensive applications with large data sets.

A Preliminary Study of Compiler Transformations for Graph Applications on the Emu System

Unlike dense linear algebra applications, graph applications typically suffer from poor performance because of 1) inefficient utilization of memory systems through random memory accesses to graph data, and 2) overhead of executing atomic operations. Hence, there is a rapid growth in improving both software and hardware platforms to address the above challenges. One such improvement in the hardware platform is a realization of the Emu system, a thread migratory and near-memory processor. In the Emu system, a thread responsible for computation on a datum is automatically migrated over to a node where the data resides without any intervention from the programmer. The idea of thread migrations is very well suited to graph applications as memory accesses of the applications are irregular. However, thread migrations can hurt the performance of graph applications if overhead from the migrations dominates benefits achieved through the migrations.

In this preliminary study, we explore two high-level compiler optimizations, i.e., loop fusion and edge flipping, and one low-level compiler transformation leveraging hardware support for remote atomic updates to address overheads arising from thread migration, creation, synchronization, and atomic operations. We performed a preliminary evaluation of these compiler transformations by manually applying them on three graph applications over a set of RMAT graphs from Graph500.—Conductance, Bellman-Ford's algorithm for the single-source shortest path problem, and Triangle Counting. Our evaluation targeted a single node of the Emu hardware prototype, and has shown an overall geometric mean reduction of 22.08% in thread migrations.

Data Placement Optimization in GPU Memory Hierarchy Using Predictive Modeling

Modern supercomputers often use Graphic Processing Units (or GPUs) to meet the ever-growing demands for high performance computing. GPUs typically have a complex memory architecture with various types of memories and caches, such as global memory, shared memory, constant memory, and texture memory.The placement of data on these memories has a tremendous impact on the performance of the HPC applications and identifying the optimal placement location is non-trivial.

In this paper, we propose a machine learning-based approach to determine the best class of GPU memory that will minimize GPU kernel execution time. The machine learning process utilizes a set of performance counters obtained from profiling runs and combines with relevant hardware features to generate trained models. We evaluate our approach on several generations of NVIDIA GPUs, including Kepler, Maxwell, Pascal, and Volta on a set of benchmarks. The results show that the trained models achieve prediction accuracies over 90%.

MCHPC'18 – Workshop Lunch (on your own)

MCHPC'18 Afternoon Keynote: All Tomorrow's Memory Systems *Bruce Jacob (University of Maryland)*
Memory and communication are the primary reasons that our time-to-solution is no better than it currently is … the memory system is slow; the communication overhead is high; and yet a significant amount of research is still focused on increasing processor performance, rather than decreasing (the cost of) data movement. I will discuss recent and near-term memory-system technologies including high-bandwidth DRAMs and nonvolatile main memories, as well as the impact of tomorrow's memory technologies on tomorrow's applications and operating systems.

MCHPC'18 – Workshop Afternoon Break

On the Applicability of PEBS-Based Online Memory Access Tracking for Heterogeneous Memory Management at Scale

Operating systems have historically had to manage only a single type of memory device. The imminent availability of heterogeneous memory devices based on emerging memory technologies confronts the classic single memory model and opens a new spectrum of possibilities for memory management. Transparent data movement based on access patterns of applications is a desired feature to hide the complexity of memory management to end users. However, capturing memory access patterns of an application at runtime comes at a cost, which is particularly challenging for large scale parallel applications that may be sensitive to system noise.In this work, we focus on the access pattern profiling. We study the feasibility of using Intel's Processor Event Based Sampling (PEBS) feature to record memory accesses by sampling at runtime and study the overhead at scale. We have implemented a custom PEBS driver in the IHK/McKernel lightweight multi-kernel operating system, one of whose advantages is minimal system interference due to the lightweight kernel's simple design compared to other OS kernels such as Linux. We present the PEBS overhead of a set of scientific applications and show the access patterns identified in noise sensitive HPC applications. Our results show that clear access patterns can be captured with 10% overhead in the worst case when running on up to 128k CPU cores (2,048 Intel Xeon Phi Knights Landing nodes). We conclude that online memory access profiling using PEBS at large scale is promising for memory management in heterogeneous memory environments.

Exploring Allocation Policies in Disaggregated Non-Volatile Memories

Many modern applications have memory footprints that are increasingly large, driving system memory capacities higher and higher. However, due to the diversity of applications that run on High-Performance Computing (HPC) systems, the memory utilization can fluctuate widely from one application to another, which results in underutilization issues when there are many jobs with small memory footprints. Since memory chips are collocated with the compute nodes, this necessitates the need for message passing APIs to be able to share information between nodes.

To address some of these issues, vendors are exploring disaggregated memory-centric systems. In this type of organization, there are discrete nodes, reserved solely for memory, which are shared

across many compute nodes. Due to their capacity, low-power, and non-volatility, Non-Volatile Memories (NVMs) are ideal candidates for these memory nodes. Moreover, larger memory capacities open the door to different programming models (more shared memory style approaches) which are now being added to the C++ and Fortran language specifications. This paper proposes a simulation model for studying disaggregated memory architectures using a publicly available simulator, SST Simulator, and investigates various memory allocation policies.

Heterogeneous Memory and Arena-Based Heap Allocation

Nonuniform Memory Access (NUMA) will likely continue to be the chief abstraction used to expose heterogeneous memory. One major problem with using NUMA in this way is, the assignment of memory to devices, mediated by the hardware and Linux OS, is only resolved to page granularity. That is, pages, not allocations, are explicitly assigned to memory devices. This is particularly troublesome if one wants to migrate data between devices:\ since only pages can be migrated, other data allocated on the same pages will be migrated as well, and it isn't easy to tell what data will be swept along to the target device. We propose a solution to this problem based on repurposing arena-based heap management to keep locality among related data structures that are used together, and discuss our work on such a heap manager.

MCHPC'18 Panel: Research Challenges in Memory-Centric Computing *Maya Gokhale (Lawrence Livermore National Laboratory)*

Room: D166
9:00 am - 5:30 pm

Workshop on Exascale MPI (ExaMPI)

The aim of workshop is to bring together researchers and developers to present and discuss innovative algorithms and concepts in the Message Passing programming model and to create a forum for open and potentially controversial discussions on the future of MPI in the Exascale era. Possible workshop topics include innovative algorithms for collective operations, extensions to MPI, including datacentric models, scheduling/routing to avoid network congestion, “fault-tolerant” communication, interoperability of MPI and PGAS models, integration of task-parallel models in MPI, and use of MPI in large scale simulations.

ExaMPI Keynote *Torsten Hoefler (ETH Zurich)*

ExaMPI – Workshop Morning Break
Morning Break for Workshops

10:00 AM - 10:30 AM

Food provided by SC

Tree-Based Fault-Tolerant Collective Operations for MPI
Alexander Margolin (Hebrew University of Jerusalem)

With the increase in size and complexity of high-performance computing systems, the probability of failures and the cost of recovery grow. Parallel applications running on these systems should be able to continue running in spite of node failures at arbitrary times. Collective operations are essential for many parallel MPI applications, and are often the first to detect such failures. This work presents tree-based fault-tolerant collective operations, which combine fault detection and recovery as an integral part each operation. We do this by extending existing tree-based algorithms, to allow for a collective operation to succeed despite failing nodes before or during its run. This differs from other approaches, where recovery takes place after a failure of such operations have failed. The paper includes a comparison between the performance of the proposed algorithm and other approaches, as well as a simulator-based analysis of performance at scale.

Understanding the Usage of MPI in Exascale Proxy Applications
Nawrin Sultana (Auburn University)

The Exascale Computing Project (ECP) focuses on the development of future exascale-capable applications. Most ECP applications use the Message Passing Interface (MPI) as their parallel programming model and create mini-apps to serve as proxies. This paper explores the explicit usage of MPI in ECP proxy applications. For our study, we empirically analyze fourteen proxy applications from the ECP Proxy Apps Suite. Our result shows that a small subset of features from MPI is commonly used in the proxies of exascale capable applications, even when they reference third-party libraries. Our study contributes to a better understanding of the use of MPI in current exascale applications. This finding can help focus software investments made for exascale systems in the MPI middleware including optimization, fault-tolerance, tuning, and hardware-offload.

Optimal Algorithms for Half-Duplex Inter-Group All-to-All Broadcast on Fully Connected and Ring Topologies
Qiao Kang (Northwestern University)

Half-duplex inter-group collective communications are bipartite message transfer patterns such that the processes in a sender group pass messages to the processes in a receiver group. These communication patterns serve as basic operations for scientific application workflows. In this paper, we present optimal parallel algorithms for half-duplex inter-group all-to-all broadcast under bidirectional communication constraint on fully connected and ring topologies. We implement the algorithms using MPI communication functions and perform experiments on Cori. For the fully connected topology case, we compare our algorithms with production MPI libraries. For the ring topology case, we implement our proposed algorithms using MPI_Sendrecv function to emulate a ring topology environment. The proposed algorithms are compared with the intra-group Allgather algorithm emulated under the same environment. Message sizes ranging from 32KB to 4MB are used

for evaluations. The proposed algorithms for fully connected topology are up to 5 times faster than the root gathering algorithm adopted by MPICH. The proposed algorithms for the ring topology are up to 1.4 times faster than the intra-group Allgather algorithm.

ExaMPI – Workshop Lunch (on your own)

AITuning: Machine Learning-Based Tuning Tool for Run-Time Communication Libraries
Alessandro Fanfarillo (National Center for Atmospheric Research)

In this work, we address the problem of tuning communication libraries by using a deep reinforcement learning approach. Reinforcement learning is a machine learning technique incredibly effective in solving game-like situations. In fact, tuning a set of parameters in a communication library in order to get better performance in a parallel application can be expressed as a game: find the right combination/path that provides the best reward. Even though AITuning has been designed to be utilized with different run-time libraries, we focused this work on applying it to the OpenCoarrays run-time communication library, built on top of MPI-3. This work not only shows the potential of using a reinforcement learning algorithm for tuning communication libraries, but also demonstrates how the MPI Tool Information Interface, introduced by the MPI-3 standard, can be used effectively by run-time libraries to improve the performance without human intervention.

A Novel Approach to Supporting Communicators for In-Switch Processing of MPI Collectives
Martin Herbordt (Boston University)

MPI collective operations can often be performance killers in HPC applications, and we seek to solve this bottleneck by offloading them to switch hardware. We've seen from previous works including our own that moving collectives into the network offers significant performance benefits. However, there has been little advancement in providing support for sub-communicator collectives in an efficient manner. Using FPGAs, which provide the ability to couple communication and computation, we have designed an in-switch hardware accelerator to implement support for MPI communicators and full offload of MPI collectives. With our design, preliminary results show that we can achieve up to a 10x speedup over conventional clusters for short message collectives over irregular intra-communicators.

ExaMPI – Workshop Afternoon Break

ExaMPI Invited Talk *Ron Brightwell (Sandia National Laboratories)*

Anycast: Rootless Broadcasting with MPI
Tonglin Li (Lawrence Berkeley National Laboratory)

Flexible communication routines that enable high-performance computing (HPC) applications to operate with less synchronization are a great benefit to application algorithm creation and developer productivity. Further, communication operations that reduce synchronization in HPC applications while

continuing to scale well are critical to application performance in the exascale era. We present an algorithm that allows an application to scalably broadcast a message from any MPI rank without specifying the origin of the message in advance, demonstrating application development flexibility, highly scalable operation, and reduced application synchronization.

ExaMPI Panel *Ryan E. Grant (Sandia National Laboratories)*

Room: D168
9:00 am - 5:30 pm

Workshop on Education for High Performance Computing (EduHPC)

The EduHPC Workshop is devoted to the development and assessment of educational resources for undergraduate and graduate education in High Performance Computing (HPC) and Parallel and Distributed Computing (PDC). Data science curriculum (e.g. for new degree programs and within data science centers) and topics related to Internet of Things are also in the workshop scope. PDC, HPC, and data science now permeate the world of computing to a degree that makes it imperative for even entry-level computer professionals to incorporate these computing modalities into their computing toolboxes, no matter what type of computing problems they work on.

This workshop focuses on the state of the art in HPC and PDC education via contributed and invited papers from academia, industry, government laboratories and other educational and research institutions. Topics of interest include all topics pertaining to the teaching of PDC and HPC within Computer Science and Engineering, Computational Science, and Domain Science and Engineering curricula. The emphasis of the workshop is undergraduate education, but fundamental issues related to graduate education are also welcome. The target audience will broadly include SC18 attendees from academia, industry, and research laboratories. This includes both researchers and educators, as well as the early adopters of the NSF/TCPP curriculum guidelines on teaching PDC (http://www.cs.gsu.edu/~tcpp/curriculum/index.php). The workshop is coordinated by the NSF-supported Center for Parallel and Distributed Computing Curriculum Development and Educational Resources (CDER). EduHPC has been an SC workshop since 2013 with attendance of 90 in 2015 and 75 in both 2016 and 2017.

Keynote Talk: Student Engagement: View from the Trenches *Neena Thota (University of Massachusetts)*
We are well aware of the importance of imparting PDC and HPC knowledge and skills to our students to prepare them for careers in Computer Science (CS), Computer Engineering (CE), Data Science, and other related careers. In my talk, I focus on the challenges of keeping students engaged, the pedagogical issues in incorporating PDC and HPC topics in our curricula, and the rigors of systematically collecting evidence to validate innovative approaches to teaching and learning.

EduHPC Workshop Coffee Break

Teaching HPC Systems and Parallel Programming with Small Scale Clusters of Embedded SoCs
Filippo Mantovani (Barcelona Supercomputing Center)

In the last decades, the continuous proliferation of High-Performance Computing (HPC) systems and data centers has augmented the demand for expert HPC system designers, administrators and programmers. For this reason, most universities have introduced courses on HPC systems and parallel programming in their degrees. However, the laboratory assignments of these courses generally use clusters that are owned, managed, and administrated by the university. This methodology has been shown effective to teach parallel programming, but using a remote cluster prevents the students from experimenting with the design, set up, and administration of such systems.

This paper presents a methodology and framework to teach HPC systems and parallel programming using a small-scale cluster of embedded System-on-Chip (SoC) boards. These SoCs are very cheap, their processors are fundamentally very similar to the ones found in HPC, and they are ready to execute Linux out of the box, so they provide a great opportunity to be used in laboratory assignments for the students to experience with assembling a cluster, setting it up, and configuring all the software ecosystem. In addition, this paper shows that the small-scale cluster can be used as the evaluation platform for parallel programming assignments.

Jupyter Notebooks and User-Friendly HPC Access
Ben Glick (Lewis & Clark College)

High performance computing systems can be hard to use, and can require advanced knowledge of the command line, resource managers, and other details that scientists without a systems administration background may find distracting. In this project, we describe our extensible gateway of tools which has allowed students and researchers with minimal HPC backgrounds to use our system with virtually no training. Performance measurements show that our gateway incurs marginal overhead. With only a small performance cost, the gateway provides an easily accessible HPC environment. This environment is specially tailored for education. We provide a means for teachers to easily provide students access to information and assignments, and give students their own personal HPC sandbox, including access to a fully-functional, personalized Jupyter Notebook server.

SMPI Courseware: Teaching Distributed-Memory Computing with MPI in Simulation
Henri Casanova (University of Hawaii at Manoa)

It is typical in High Performance Computing (HPC) courses to give students access to HPC platforms so that they can benefit from hands-on learning opportunities. Using such platforms, however, comes with logistical and pedagogical challenges. For instance, a logistical challenge is that access to representative platforms must be granted to students, which can be difficult for some institutions or course modalities; and a pedagogical challenge is that hands-on learning opportunities are constrained by the configurations of these platforms.

A way to address these challenges is to instead simulate program executions on arbitrary HPC platform configurations. In this work, we focus on simulation in the specific context of distributed-

memory computing and MPI programming education. While using simulation in this context has been explored in previous works, our approach offers two crucial advantages. First, students write standard MPI programs and can both debug and analyze the performance of their programs in simulation mode. Second, large-scale executions can be simulated in short amounts of time on a single standard laptop computer. This is possible thanks to SMPI, an MPI simulator provided as part of SimGrid. After detailing the challenges involved when using HPC platforms for HPC education and providing background information about SMPI, we present SMPI Courseware. SMPI Courseware is a set of in-simulation assignments that can be incorporated into HPC courses to provide students with hands-on experience for distributed-memory computing and MPI programming learning objectives. We describe some of these assignments, highlighting how simulation with SMPI enhances the student learning experience.

Employing Student Retention Strategies for an Introductory GPU Programming Course
Julian Gutierrez (Northeastern University)

Graphics Processing Units (GPUs) have become a vital hardware resource for the industry and research community due to their high computing capabilities. Despite this, GPUs have not been introduced into the undergraduate curriculum of Computer Engineering and are barely covered in graduate courses.

Breaching the gap between university curriculum and industry requirements for GPU expertise is ongoing, but this process takes time. Offering an immediate opportunity for students to learn GPU programming is key for their professional growth.

The Northeastern University Computer Architecture Research Lab offers a free GPU programming course to incentivize students from all disciplines to learn how to efficiently program a GPU. In this paper, we discuss the methods used to keep students engaged in a course with no academic obligations. By applying these strategies, we have been able to retain 80% of the students throughout the whole course and obtained positive feedback from students for these strategies.

Paper Discussion

PDC Curriculum Update *Charles Weems (University of Massachusetts), Ramachandran Vaidyanathan (Louisiana State University), Alan Sussman (University of Maryland), Sushil Prasad (Georgia State University), Trilce Estrada (University of New Mexico), Krishna Kant (Temple University), Arnold Rosenberg (University of Massachusetts), Anshul Gupta (IBM), Sheikh Ghafoor (Tennessee Technological University)*

EduHPC – Workshop Lunch (on your own)

Filling the Gap between Education and Industry: Evidence-Based Methods for Introducing Undergraduate Students to HPC
Fabio Francisco Banchelli Gracia (Barcelona Supercomputing Center), Filippo Mantovani (Barcelona

Supercomputing Center)

Educational institutions provide, in most cases, basic theoretical background covering several computational science topics. However High Performance Computing (HPC) and Parallel and Distributed Computing (PDC) markets require specialized technical profiles. Even the most skilled students are often not prepared to face production HPC applications of thousand of lines nor complex computational frameworks from other disciplines nor heterogeneous multinode machines accessed by hundreds of users. In this paper, we present an educational package for filling this gap. Leveraging the 4-year experience of the Student Cluster Competition, we present our educational journey together with the lessons learned and the outcomes of our methodology. We present how, in a time span of a semester and an affordable budget, a university can implement an educational package preparing pupils for starting competitive professional careers. Our findings also highlight that 78% of the students exposed to our methods remain within the HPC high-education, research or industry.

Evaluating the Wide Area Classroom after 10,500 HPC Students
John Urbanic (Pittsburgh Supercomputing Center), Thomas Maiden (Pittsburgh Supercomputing Center)

As of mid-2018 we have taught over 10,500 students in the course of 58 HPC events using the Wide Area Classroom, a novel distributed teaching platform. This has been a successful effort gauged by several important metrics. We describe both the technical and logistical structure of these events as well as the specific HPC curriculums which have proven to be most popular.

A One Year Retrospective on a MOOC in Parallel, Concurrent, and Distributed Programming in Java
Vivek Sarkar (Georgia Institute of Technology), Max Grossman (Rice University)

Much progress has been made on integrating parallel programming into the core Computer Science curriculum of top-tier universities in the United States. For example, "COMP 322: Introduction to Parallel Programming" at Rice University is a required course for all undergraduate students pursuing a bachelors degree. It teaches a wide range of parallel programming paradigms, from task-parallel to SPMD to actor-based programming.

However, courses like COMP 322 do little to support members of the Computer Science community that need to develop these skills but who are not currently enrolled in a four-year program with parallel programming in the curriculum. This group includes (1) working professionals, (2) students at USA universities without parallel programming courses, or (3) students in countries other than the USA without access to a parallel programming course.

To serve these groups, Rice University launched the "Parallel, Concurrent, and Distributed Programming in Java" Coursera specialization on July 31, 2017. In 2017, the authors of that specialization also wrote an experiences paper about launching the specialization.

In this paper, the sequel to our previous publication, we look back at the first year of the Coursera specialization. In particular, we ask the following questions: (1) how did our assumptions about the student body for this course hold up?, (2) how has the course changed since launch?, and (3) what can we learn about how students are progressing through the specialization from Coursera's built-in

analytics?

Paper Discussion

EduHPC Workshop Coffee Break

Toward Developing a Repository of Logical Errors Observed in Parallel Code for Teaching Code Correctness

Trung Nguyen (University of Massachusetts), Ritu Arora (Texas Advanced Computing Center, University of Texas)

Debugging parallel programs can be a challenging task, especially for the beginners. While parallel debuggers like DDT and TotalView can be extremely useful in tracking down the program statements that are connected to the bugs, often the onus is on the programmers to reason about the logic of the program statements in order to fix the bugs in them. These debuggers may neither be able to precisely indicate the logical errors in the parallel programs nor they may provide information on fixing those errors. Therefore, there is a need for developing tools and educational content on teaching the pitfalls in parallel programming and writing correct code. Such content can help in guiding the beginners in avoiding commonly observed logical errors and in verifying the correctness of their parallel programs. In this paper, we 1) enumerate some of the logical errors that we have seen in the parallel programs that were written by the beginners working with us, and 2) discuss the ways to fix those errors. The documentation on these logical errors can contribute in enhancing the productivity of the beginners, and can potentially help them in their debugging efforts. We have added the code samples containing the logical errors and their solutions in a Github repository so that the others in the community can reproduce the errors on their systems and learn from them. The content presented in this paper may also be useful for those developing high-level tools for detecting and removing logical errors in parallel programs.

Paper Discussion

Peachy Introduction *David Bunde (Knox College)*

Computing a Movie of Zooming into a Fractal

Martin Burtscher (Texas State University)

none

Optimization of an Image Processing Algorithm: Histogram Equalization

Julian Gutierrez (Northeastern University)

Many textbooks rely on classical linear algebra examples to illustrate best practices in parallel

programming (e.g., matrix multiplication and vector add). Despite their common use in class, these examples lack sophistication of a complete application. We have found that students seem to be more motivated to work with imaging processing algorithms, where the student can view the before and after image, visually inspecting the results of their processing.

This assignment focuses on improving the performance of the histogram equalization algorithm applied to an image. Histogram equalization is a popular image processing algorithm used to increase the contrast of an image to better highlight its features. It is a common algorithm used in many scientific applications such as x-ray imaging, thermal imaging and as a pre-processing task for multiple computer vision/deep learning algorithms.

OpenMP: What's Inside the Black Box?

This paper presents the "Implementing a minimal OpenMP runtime using Pthreads" assignment that is offered to students of Parallel Programming and Architectures (PAP), a third-year subject in the Bachelor Degree in Informatics Engineering at the Barcelona School of Informatics (FIB) of the Universitat Politecnica de Catalunya (UPC) - BarcelonaTech. We believe this is a high-quality assignment, previously-tested (and evolved) in the past 3 academic years, that is readily adoptable by other educators with the objectives of 1) understanding how a high-level parallel programming model such as OpenMP is implemented and 2) learning Pthreads in a very interesting and applied way. This assignment contributes to open the black box behind the compilation and execution command lines for OpenMP programs.

The Wave Equation as a Motivating Example for High Performance Computing
David Joiner (Kean University)

A series of activities based on the solution of the wave equation in 1 and 3 dimensions spanning multiple concepts in a traditional high performance computing class are presented.

Storms of High-Energy Particles: An assignment for OpenMP, MPI, and CUDA/OpenCL
Arturo Gonzalez-Escribano (University of Valladolid)

We present an assignment used in a Parallel Computing course to teach the approaches to the same problem in different parallel programming models. It targets basic concepts of shared-memory programming with OpenMP, distributed-memory programming with MPI, and GPU programming with CUDA or OpenCL. This assignment is based on a highly simplified simulation of the impact of high-energy particle storms in an exposed surface. The idea is inspired by simulations to test the robustness of space-vessels material. The program is designed from scratch to be simply, easy to understand by students, and to include specific parallel structures and optimization opportunities. A simple parallelization in the three models considered is quite direct. But the program is plenty of opportunities for further improvements, optimizations, and more sophisticated techniques usage. It has been successfully used in parallel programming contests during our course, using the performance obtained by the students code as a measure of success.

Presentation of Updates from the Community

Experience Report: 4 Years of Teaching Cloud Computing and Big Data at the University Level *Paul Talaga (University of Indianapolis)*
Valuable insights can be gained from teaching a combined Cloud Computing/Big Data course over many years between differing sized institutions and hundreds of students. We report and reflect on the content of the course, schedule, assignments, student feedback, and hardware/software used starting at the University of Cincinnati in 2014 at the undergraduate/graduate level and culminating at the University of Indianapolis in 2017 with an undergraduate-only course.

Building a Low Budget Cluster Through Hardware Reuse *Daniel Ellsworth (Colorado College), Andrew Watson (Colorado College), Keith Conger (Colorado College)*
Exploring parallel programming at the undergraduate level is difficult due to the limited availability of hardware exposing sufficient parallelism. While increasing core counts in consumer hardware is helping to address this, core counts are still low, and system noise from desktop operating systems and background processes make performance measurement on student owned systems less than ideal. For short duration classes, the time needed on production clusters to establish student accounts and queue waits per job submission are significant barriers to use. This poster discusses MCSCN, a small educational cluster, built using effectively free hardware to address these challenges at a small liberal arts university.

Introducing Three Basic Concepts in Parallel Computation to 1st Year Computer Science Students in a Simple and Effective Way
Haklin Kimm (East Stroudsburg University of Pennsylvania)

What are some simple computation problems, with non-trivial applications, by which we can introduce the three basic concepts in parallel computation to 1st year Computer Science students who have little programming experience and knowledge of algorithms. Those three basic concepts are: (1) How parallel computation can reduce an algorithms completion-time, including methods to determine the completion-time for a given number of computing agents (people or CPUs, for simplicity). (2) How to identify parallelizable parts of an algorithm given in a pseudocode form. (3) How to balance the computation-load of the agents, with perhaps some modification of the algorithm, to further reduce the completion time. We present one such "ideal" computation problem, which has several simple intuitive solution algorithms and multiple ways of parallelization. The results of a short quiz following 2 one hour lectures covering one parallel algorithm of the ideal problem showed the students achieved a good understanding of the above three basic concepts. One could cover a bit more material in one or two additional lectures. A full discussion of different parallelizations of our ideal problem and their performance comparisons is fit for 2nd year and more advanced students.

Engaging Students in Parallel and Distributed Computing Learning by Games Design Using Unity
Michelle Zhu (Montclair State University)

Parallel and distributed computing (PDC) is increasingly important in science and engineering discipline, yet many computer science curriculum especially in the early college years either do not provide students with adequate PDC components, or there are pedagogical challenges for students to

learn PDC concepts. Meanwhile, educational games has become an important teaching tool in a wide variety of disciplines. We plan to integrate more PDC concepts in our game design courses. Our experiences indicates that (1) students tirelessly strive to understand the PDC concepts in order to design game challenges, and (2) the virtual reality immersion experience is enormously engaging and fun.

Lessons from Integrating Parallelism into Undergraduate Curriculum at UMKC

Mohammad Amin Kuhail (University of Missouri, Kansas City)

At the University of Missouri-Kansas City (UMKC), we successfully integrated parallelism into the undergraduate Data Structures course. In this paper, we share lessons learned from the teaching strategies that have successfully worked and the strategies that unfortunately did not. Our results are based on three years of continuous work on integrating parallelism into the course. We collected the students' feedback using surveys. We refined our approach over several semesters to improve our teaching techniques.

A Comprehensive Parallel Computing Curriculum: From Second Year to Professionals

The Australian National University has had a long tradition of teaching all aspects of Parallel Computing. It begins with introductory elements of parallelism in a second year computer organization course and is followed up by a later second year course in Concurrent and Distributed Systems. The curriculum progresses with a High Performance Scientific Computation course in 3rd year, and is capped with a Parallel Systems course in 4th year. A specialty Multicore Computing course was also taught at this level. Recently, we also began delivering one-week intensive courses to external professionals; these have been in Shared and Distributed Memory HPC.

Eight Years Analysis of Adopting PDC in Data Structures at UCA

University of Central Arkansas (UCA) is located at the center of Arkansas. The student population of UCA is between 10,000 and 12,000. The student body of UCA is around 20% minority and 60% female. UCA was established (in 1907) over a hundred years ago. It is the third largest public school as well as the third fastest enrollment growing public school in Arkansas.

Since 2011 fall semester, the Department of Computer Science at UCA has been selected as one of early adaptors for NSF/TCPP Curriculum Initiative on Parallel and Distributed Computing (PDC). Data Structures has adopted the parallel and distributed concepts for the past eight years. Our goal is to revise our adapted courses, share our successful experiences through this paper and attract more adopters.

Introducing PDC Concepts with Spatial Computing Curriculum

Satish Puri (Marquette University)

From online maps, such as Google Maps, to consumer GPS devices, we immensely benefit from spatial computing. As a new early adopter (started in 2018) from Marquette University, I present my project in which introduce parallel and distributed computing (PDC) concepts using topics from spatial computing domain. The curriculum development tasks in the project are the creation of instructional materials at the intersection of PDC and spatial computing. The broader community under the umbrella of spatial computing includes instructors who teach algorithms, computational geometry, computer graphics, spatial databases, geographic information system (GIS), etc. Instructional materials developed so far including sequential implementations using C/C++ and Java and a few parallel implementations using threads will be presented. Line segment intersection reporting is chosen as an exemplar, and the overall material is built around it.

Discussions on Peachy Assignment and Community Update

Best Paper Announcement *Erik Saule (University of North Carolina, Charlotte)*

Closing Remarks *Erik Saule (University of North Carolina, Charlotte)*

Room: D173
9:00 am - 5:30 pm

WORKS 2018: 13th Workshop on Workflows in Support of Large-Scale

Data Intensive Workflows (aka scientific workflows) are routinely used in most scientific disciplines today, especially in the context of parallel and distributed computing. Workflows provide a systematic way of describing the analysis and rely on workflow management systems to execute the complex analyses on a variety of distributed resources. This workshop focuses on the many facets of data-intensive workflow management systems, ranging from job execution to service management and the coordination of data, service and job dependencies. The workshop therefore covers a broad range of issues in the scientific workflow lifecycle that include: data intensive workflows representation and enactment; designing workflow composition interfaces; workflow mapping techniques that may optimize the execution of the workflow; workflow enactment engines that need to deal with failures in the application and execution environment; and a number of computer science problems related to scientific workflows such as semantic technologies, compiler methods, fault detection, and tolerance.

Keynote *Ilkay Altintas (San Diego Supercomputer Center)*

WORKS – Workshop Morning Break

Reduction of Workflow Resource Consumption Using a Density-based Clustering Model
Douglas Thain (University of Notre Dame)

Often times, a researcher running a scientific workflow will ask for orders of magnitude too few or too many resources to run their workflow. If the resource requisition is too small, the job may fail due to resource exhaustion; if it is too large, resources will be wasted though job may succeed. It would be ideal to achieve a near-optimal number of resources the workflow runs to ensure all jobs succeed and minimize resource waste. We present a strategy for solving the resource allocation problem: (1) resources consumed by each job are recorded by a resource monitor tool; (2) a density-based clustering model is proposed for discovering clusters in all jobs; (3) a maximal resource requisition is calculated as the ideal number of each cluster. We ran experiments with a synthetic workflow of homogeneous tasks as well as the bioinformatics tools Lifemapper, SHRIMP, BWA and BWA-GATK to capture the inherent nature of resource consumption of a workflow, the clustering allowed by the model, and its usefulness in real workflows. In Lifemapper, the least time saving, cores saving, memory saving, and disk saving are 13.82%, 16.62%, 49.15%, and 93.89%, respectively. In SHRIMP, BWA, and BWA-GATK, the least cores saving, memory saving and disk saving are 50%, 90.14%, and 51.82%, respectively. Compared with fixed resource allocation strategy, our approach provide a noticeable reduction of workflow resource consumption.

Flux: Overcoming Scheduling Challenges for Exascale Workflows
Dong H. Ahn (Lawrence Livermore National Laboratory)

Many emerging scientific workflows that target high-end HPC systems require complex interplay with the resource and job management software~(RJMS). However, portable, efficient and easy-to-use scheduling and execution of these workflows is still an unsolved problem. We present Flux, a novel, hierarchical RJMS infrastructure that addresses the key scheduling challenges of modern workflows in a scalable, easy-to-use, and portable manner. At the heart of Flux lies its ability to be nested seamlessly within batch allocations created by other schedulers as well as itself. Once a hierarchy of Flux instance is created within each allocation, its consistent and rich set of well-defined APIs portably and efficiently support those workflows that can often feature non-traditional execution patterns such as requirements for complex co-scheduling, massive ensembles of small jobs and coordination among jobs in an ensemble.

LOS: Level Order Sampling for Task Graph Scheduling on Heterogeneous Resources
Carl Witt (Humboldt-Universität zu Berlin)

List scheduling is an approach to task graph scheduling that has been shown to work well for scheduling tasks with data dependencies on heterogeneous resources. Key to the performance of a list scheduling heuristic is its method to prioritize the tasks, and various ranking schemes have been proposed in the literature. We propose a method that combines multiple random rankings instead of a using a deterministic ranking scheme.

We introduce L-Orders, which are a subset of all topological orders of a directed acyclic graph. L-Orders can be used to explore targeted regions of the space of all topological orders. Using the observation that the makespans in one such region are often approximately normal distributed, we estimate the expected time to solution improvement in certain regions of the search space. We

combine targeted search and improvement time estimations into a time budgeted search algorithm that balances exploration and exploitation of the search space. In 40,500 experiments, our schedules are 5% shorter on average and up to 40% shorter in extreme cases than schedules produced by HEFT.

Energy-Aware Workflow Scheduling and Optimization in Clouds Using Bat Algorithm

With the ever-increasing deployment of data centers and computer networks around the world, cloud computing has emerged as one of the most important paradigms for large-scale data-intensive applications. However, these cloud environments face many challenges including energy consumption, execution time, heat and CO2 emission, and operation cost. Due to the extremely large scale of these applications and a huge amount of resource consumption, even a small portion of the improvements in any of the above fields can yield huge ecological and financial rewards. Efficient and effective workflow scheduling in cloud environments is one of the most significant ways to confront the above problems and achieve optimal resource utilization. We propose an Energy Aware, Time and Throughput Optimization heuristic (EATTO) based on bat algorithm. Our goal is to minimize energy consumption and execution time of computation-intensive workflows while maximizing throughput, without imposing any significant loss on the Quality of Service (QoS) guarantee.

A Practical Roadmap for Provenance Capture and Data Analysis in Spark-based Scientific Workflows
Daniel Oliveira (UFF)

Whenever high-performance computing applications meet data-intensive scalable systems, an attractive approach is the use of Apache Spark for the management of scientific workflows. Spark provides several advantages such as being widely supported and granting efficient in-memory data management for large-scale applications. However, Spark still lacks support for data tracking and workflow provenance. Additionally, Spark's memory management requires accessing all data movements between the workflow activities. Therefore, the running of legacy programs on Spark is interpreted as a "black-box" activity, which prevents the capture and analysis of implicit data movements. Here, we present SAMbA, an Apache Spark extension for the gathering of prospective and retrospective provenance and domain data within distributed scientific workflows. Our approach relies on enveloping both RDD structure and data contents at runtime so that (i) RDD-enclosure consumed and produced data are captured and registered by SAMbA in a structured way, and (ii) provenance data can be queried during and after the execution of scientific workflows. By following the W3C PROV representation, we model the roles of RDD regarding prospective and retrospective provenance data. Our solution provides mechanisms for the capture and storage of provenance data without jeopardizing Spark's performance. The provenance retrieval capabilities of our proposal are evaluated in a practical case study, in which data analytics are provided by several SAMbA parameterizations.

End-to-End Online Performance Data Capture and Analysis for Scientific Workflows
George Papadimitriou (University of Southern California)

With the increased prevalence of employing workflows for scientific computing and a push toward

exascale computing, it has become paramount that we are able to analyze characteristics of scientific applications to better understand the impact on the underlying infrastructure and vice-versa. Such analysis can help drive the design, development, and optimization of these next generation systems and solutions. In this paper, we present the architecture, integration with existing well-established and newly developed tools, to collect online performance statistics of workflow executions from various, heterogeneous sources and publish them in a distributed database (Elasticsearch). Using this architecture, we are able to correlate online workflow performance data with data from the underlying infrastructure, and present them in a useful and intuitive way via an online dashboard. We have validated our approach by executing two classes of real-world workflows, both under normal and anomalous conditions. The first is an I/O-intensive genome analysis workflow; the second, a CPU- and memory-intensive material science workflow. Based on the data collected in Elasticsearch, we are able to demonstrate that we can correctly identify anomalies that we injected. We identify this end-to-end data collection of workflow performance data as an important resource of training data for automated machine learning analysis.

WORKS – Workshop Lunch (on your own)

Planner: Cost-efficient Execution Plans Placement for Uniform Stream Analytics on Edge and Cloud
Laurent Prosperi (ENS Paris - Saclay)

Stream processing applications handle unbounded and continuous flows of data items which are generated from multiple geographically distributed sources. Two approaches are commonly used for processing: cloud-based analytics and edge analytics. The first one routes the whole data set to the Cloud, incurring significant costs and late results from the high latency networks that are traversed. The latter can give timely results but forces users to manually define which part of the computation should be executed on Edge and to interconnect it with the remaining part executed in the Cloud, leading to sub-optimal placements. In this paper, we introduce Planner, a middleware for uniform and transparent stream processing across Edge and Cloud. Planner automatically selects which parts of the execution graph will be executed at the Edge in order to minimize the network cost. Real-world micro-benchmarks show that Planner reduces the network usage by 40% and the makespan (end-to-end processing time) by 15% compared to state-of-the-art.

Dynamic Distributed Orchestration of Node-RED IOT Workflows Using a Vector Symbolic Architecture
Christopher Simpkin (Cardiff University)

There are a large number of workflow systems designed to work in various scientific domains, including support for the Internet of Things (IoT). One such workflow system is Node-RED, which is designed to bring workflow-based programming to IoT. However, the majority of scientific workflow systems, and specifically systems like Node-RED, are designed to operate in a fixed networked environment, which rely on a central point of coordination in order to manage the workflow. The main focus of the work described in this paper is to investigate means whereby we can migrate Node-RED workflows into a decentralized execution environment, so that such workflows can run on Edge networks, where nodes are extremely transient in nature. In this work, we demonstrate the feasibility of such an approach by showing how we can migrate a Node-RED based traffic congestion workflow

into a decentralized environment. The traffic congestion algorithm is implemented as a set of Web services within Node-RED and we have architected and implemented a system that proxies the centralized Node-RED services using cognitively-aware wrapper services, designed to operate in a decentralized environment. Our cognitive services use a Vector Symbolic Architecture to semantically represent service descriptions and workflows in a way that can be unraveled on the fly without any central point of control. The VSA-based system is capable of parsing Node-RED workflows and migrating them to a decentralized environment for execution; providing a way to use Node-RED as a front-end graphical composition tool for decentralized workflows.

Optimizing the Throughput of Storm-Based Stream Processing in Clouds
Huiyan Cao (New Jersey Institute of Technology)

There is a rapidly growing need for processing large volumes of streaming data in real time in various big data applications. As one of the most commonly used systems for streaming data processing, Apache Storm provides a workflow-based mechanism to execute directed acyclic graph (DAG)-structured topologies. With the expansion of cloud infrastructures around the globe and the economic benefits of cloud-based computing and storage services, many such Storm workflows have been shifted or are in active transition to clouds. However, modeling the behavior of streaming data processing and improving its performance in clouds still remain largely unexplored. We construct rigorous cost models to analyze the throughput dynamics of Storm workflows and formulate a budget-constrained topology mapping problem to maximize Storm workflow throughput in clouds. We show this problem to be NP-complete and design a heuristic solution that takes into consideration not only the selection of virtual machine type but also the degree of parallelism for each task (spout/bolt) in the topology. The performance superiority of the proposed mapping solution is illustrated through extensive simulations and further verified by real-life workflow experiments deployed in public clouds in comparison with the default Storm and other existing methods.

WORKS – Workshop Afternoon Break

DagOn*: Executing direct acyclic graphs as parallel jobs on anything
Raffaele Montella (University of Naples Parthenope)

The democratization of computational resources, thanks to the advent of public, private, and hybrid clouds, changed the rules in many science fields. For decades, one of the effort of computer scientists and computer engineers was the development of tools able to simplify access to high-end computational resources by computational scientists. However, nowadays any science field can be considered “computational” as the availability of powerful, but easy to manage workflow engines, is crucial. In this work, we present DagOn* (Direct acyclic graph On anything), a lightweight Python library implementing a workflow engine able to execute parallel jobs represented by direct acyclic graphs on any combination of local machines, on-premise high performance computing clusters, containers, and cloud-based virtual infrastructures. We use a real-world production-level application for weather and marine forecasts to illustrate the use of this new workflow engine.

WRENCH: A Framework for Simulating Workflow Management Systems

Henri Casanova (University of Hawaii at Manoa), Rafael Ferreira da Silva (University of Southern California)

Scientific workflows are used routinely in numerous scientific domains, and Workflow Management Systems (WMSs) have been developed to orchestrate and optimize workflow executions on distributed platforms. WMSs are complex software systems that interact with complex software infrastructures. Most WMS research and development activities rely on empirical experiments conducted with full-fledged software stacks on actual hardware platforms. Such experiments, however, are limited to hardware and software infrastructures at hand and can be labor- and/or time-intensive. As a result, relying solely on real-world experiments impedes WMS research and development. An alternative is to conduct experiments in simulation.

In this work we present WRENCH, a WMS simulation framework, whose objectives are (i) accurate and scalable simulations; and (ii) easy simulation software development. WRENCH achieves its first objective by building on the SimGrid framework. While SimGrid is recognized for the accuracy and scalability of its simulation models, it only provides low-level simulation abstractions and thus large software development efforts are required when implementing simulators of complex systems. WRENCH thus achieves its second objective by providing high-level and directly re-usable simulation abstractions on top of SimGrid. After describing and giving rationales for WRENCH's software architecture and APIs, we present a case study in which we apply WRENCH to simulate the Pegasus production WMS. We report on ease of implementation, simulation accuracy, and simulation scalabiliy so as to determine to which extent WRENCH achieves its two above objectives. We also draw both qualitative and quantitative comparisons with a previously proposed workflow simulator.

WORKS 2018 Panel

Room: D220
9:00 am - 5:30 pm

Women in HPC: Diversifying the HPC Community

The ninth international Women in HPC workshop will be held at SC18, Dallas, USA. Following the overwhelming success of the WHPC workshop at SC17, we will focus on the following topics this year: - Building workplace resilience and maintaining well-being, while managing work stress. - Being part of the solution: instructions for advocates and allies. - Best practices from organizations on improving workplace diversity. - Managing the two body problem and achieving effective work-life balance.

We will also provide opportunities aimed at promoting and providing women with the skills to thrive in HPC including: - Short talks by women working in HPC - Pointers on how to handle workplace conflicts and effectively respond to discrimination - Short talks on: best practices from outside the HPC community for developing equity frameworks, suggestions and challenges facing women in

venture capital, effectively dealing with workplace discrimination, building a successful career and excelling in the workforce.

A Black Woman's Sojourn in High Performance Computing: Recovering Lost History *Ruby Mendenhall (University of Illinois)*
In 1797, Isabella Bomfree was born a slave. She escaped to freedom in 1827 and changed her name to Sojourner Truth in 1843 to indicate that she would travel far and wide to tell people what was right. She gave her famous "Ain't I a Woman?" speech in 1851 to highlight the strengths and vulnerabilities of being Black and a woman. Throughout history, Black women's lived experiences have often been invisible and erased. Therefore, it is important to combat the erasure of Black women and move toward a correction and claiming of their space within the digitized record. This presentation will discuss a study that employs latent dirichlet allocation (LDA) algorithms and comparative text mining to search 800,000 periodicals in JSTOR (Journal Storage) and HathiTrust from 1746 to 2014 to identify the types of conversations that emerge about Black women's shared experience over time and the resulting knowledge that developed. This presentation will also discuss what attracted Mendenhall to HPC, what she sees as the strengths of HPC and her plans for future research which involves developing a data base with a cohort of 100,000 Black women citizen scientists who will help to conduct and analyze longitudinal research based on their lived experiences.

Women in HPC – Workshop Morning Break

Hot Topics Discussion I: Thriving at Work *Linda Akli (Southeastern Universities Research Association (SURA)_), Maytal Dahan (University of Texas), Laura Schulz (Leibniz Supercomputing Centre)*
The session will be focused on hot topic presentations including the following topics: - Being a diversity ally—what it means and how you can help by Linda Akli, Southeastern Universities Research Association - Telecommuting boss - being a successful remote employee by Maytal Dahan, TACC - Leadership in HPC by Laura Schulz, Leibniz Supercomputing Center

Hot Topics Discussion II: Thriving at Work *Lorna Rivera (Georgia Institute of Technology), Lucy Nowell (US Department of Energy Office of Advanced Scientific Computing Research), Carissa Holohan (Argonne National Laboratory)*
The session will resume hot topic presentations on the following topics: - Effectively responding to discrimination -- Lorna Rivera, Georgia Tech - Effective workplace communication -- Lucy Nowell, DoE Office of Science - Coming out as a transgender woman at a major HPC center -- Carissa Holohan, Argonne National Laboratory

Panel Discussion – Best Practices from Organizations on Improving Workplace Diversity. *Christine Cuicchi (US Department of Defense)*
The panel discussion will address strategies to improve diversity and inclusion at workplaces.

Women in HPC – Workshop Lunch (on your own)

High Performance Computing in Dynamic Traffic Simulation

Dynamic traffic simulation enables to emulate the traffic congestion, which is a necessary aspect of traffic evolution. However, dynamic traffic simulations must compute in seconds if they are to be used in real-time traffic management systems. High performance computing (HPC) resources provide the power and computation to significantly speed up these simulations, thus enabling their use in instantaneous traffic control.

Macroscopic dynamic simulation uses continuum traffic-flow models that are based on traffic volume, density and speed. Though there have been many works that apply parallel computation in HPC for agent-based microscopic dynamic simulation, parallel macroscopic simulation has not been studied adequately. Hence, we devised a parallel strategy for the Berkeley Advanced Traffic Simulator (BeATS), which a simulation framework for macroscopic dynamic simulation. Given n cores, the parallel simulation begins by having the root core (processor 0) partitioning the network into n minimum-cut partitions using the METIS program. Core i loads the ith partition and computes the corresponding traffic states. The cores use graph-based MPI interface to communicate boundary information to other cores that have adjacent network partitions.

We implemented the parallel BeATS simulator on Cori supercomputer at NERSC (nersc.gov). The parallel BeATS simulator was tested on a synthetic grid network with 2500 nodes, 10000 links, and over 600 origin-destination pairs. Results showed linear speed-up as the number of compute cores grew from 1 up to 8 cores. The simulation time was reduced from 28 minutes to 25 seconds with 256 cores.

Best Practices from Organizations on Improving Workplace Diversity

Elizabett Hillery (Purdue University), Claire Stirm (Purdue University), Carolyn Ellis (Purdue University), Laura Theademan (Purdue University)

Purdue Research Computing has been a proven leader in cultivating future STEM professionals for over a decade and creating a diverse workforce. We have built a pro-diversity and inclusive workplace culture that embraces a variety of academic degrees, generations, demographics, expertise, and experience. We are fortunate to be part of a broader diversity and recruiting initiative at Purdue University. As we celebrate our 150th anniversary as a land-grant institute, we are proud of our university's rich culture. 2018 fall semester, Purdue ranks 3rd in international student enrollment among all public United States colleges and universities and 2nd in the Big 10. 22% of incoming freshman are a part of the minority domestic student population with 12% of the class enrolling in the Polytechnic Institute, (https://www.admissions.purdue.edu/academics/enrollment.php).

Our diversity outreach efforts have organically evolved to mentoring and developing these diverse students and young professionals through initiatives like the establishment of the Purdue Women in HPC organization, mentoring Purdue's first high school team to the 2017 Supercomputing Student Cluster Competition, and bringing to the 2018 Supercomputing Student Cluster Competition Purdue University's first all-female team. Our workplace diversity has developed over time through small incremental changes. Each successful connection brings new avenues into future possibilities of additional variety.

We aspire to cultivate future leaders who will be engaged and empowered to share new ideas through an environment that champions transparency, communication, and inclusion. We recognize that diversity is not an "end-state", and it must remain as a metric we always grade ourselves against.

Efficient Application of Low Mach Number Hydrodynamics Code to Stellar Flows
Duoming Fan (Lawrence Berkeley National Laboratory)

Astrophysical phenomena that occur in the low Mach number regime are often computationally expensive because the time step is constrained by both the characteristic fluid velocity and, by comparison, the much larger speed of sound. In addition, astrophysical flows are generally highly turbulent, and considerable computational cost is required to resolve the local flow in smaller regions of interest. We introduce MAESTROeX, a low Mach number hydrodynamics code for computing stellar flows that uses the AMReX C++/F90 libraries and software structure. By filtering out the acoustic waves from the governing equations, MAESTROeX uses a time step constraint that is based only on the fluid velocity rather than the sound speed. It also incorporates adaptive mesh refinement (AMR) to efficiently increase spatial resolutions by locally refining the grid in the regions of interest. Additionally, to allow the proper capture of the effects of an expanding star, MAESTROeX uses a novel one-dimensional radial base state whose evolution is coupled to the evolution of the full state. In our latest attempt to improve the local hydrodynamic equilibrium errors caused by the radial base states, the base states are now computed at uneven intervals as to directly map to the Cartesian grid. The performance and scalability of MAESTROeX are evaluated on NERSC and OLCF systems scalable to over 100K cores. Future work includes improving the current algorithm and offloading simple subroutines to GPUs to increase computational efficiency, and incorporating new physics into the system such as those needed for rotating stars.

The Movement toward HPC Inclusivity: Achieving On-Demand Accessibility of High Performance Computing (HPC) through Ephemeral Projects Utilizing the Alces Gridware Project
Cristin Merritt (Alces Flight Limited)

This presentation covers off the adoption of public cloud over the past two years. Using five public case studies, we will discuss how moving into cloud is not an 'all or nothing' option and demonstrate how adding cloud to an HPC solution invites in more opportunities for HPC to grow. But jumping into cloud isn't always an easy! Our presentation will note key differences between hardware and cloud, how those differences play out in the real world, and how open source options are starting to tie these platforms together.

Special thanks to Across the Cloud, Institute of Cancer Research, Sterling Geo, RSE Sheffield, University of Liverpool and Amazon Web Services for supporting this research.

Optimizing Python Data Processing for the DESI Experiment on the NERSC Cori Supercomputer
Laurie A. Stephey (Lawrence Berkeley National Laboratory, National Energy Research Scientific Computing Center (NERSC))

The goal of the Dark Energy Spectroscopic Instrument (DESI) experiment is to better understand dark energy by making the most detailed 3D map of the universe to date. The images obtained each night over a period of 5 years starting in 2019 will be sent to the NERSC Cori supercomputer for processing and scientific analysis.

The DESI spectroscopic pipeline for processing these data is written exclusively in Python. Writing in

Python allows the DESI scientists to write very readable scientific code in a relatively short amount of time. However, the drawback is that Python can be substantially slower than more traditional HPC languages like C, C++, and Fortran.

The goal of this work is to increase the efficiency of the DESI spectroscopic data processing at NERSC while satisfying their requirement that the software remain in Python. As of this writing we have obtained speedups of over 6x and 7x on the Cori Haswell and KNL partitions, respectively. Several profiling techniques were used to determine potential areas for improvement including Python's cProfile, line_profiler, Intel Vtune, and Tau. Once we identified expensive kernels, we used the following techniques: 1) JIT-compiling hotspots using Numba (the most successful strategy so far), 2) reducing MPI data transfer where possible (i.e. replacing broadcast operations with scatter), and 3) re-structuring the code to compute and store important data rather than repeatedly calling expensive functions. We will continue using these strategies and also explore the requirements for future architectures (for example, transitioning the DESI workload to GPUs).

Study of Performance Variability on Dragonfly Systems

Xin Wang (Illinois Institute of Technology)

Dragonfly networks are being widely adopted in high-performance computing systems. On these networks, however, interference caused by resource sharing can lead to significant network congestion and performance variability. On a shared network, different job placement policies lead to different traffic distributions. Contiguous job placement policy achieves localized communication by assigning adjacent compute nodes to the same job. Random job placement policy, on the other hand, achieves balanced network traffic by placing application processes sparsely across the network to uniformly distribute the message load. Localized communication and balanced network traffic have opposite advantages and drawbacks. Localizing communication reduces the number of hops for message transfers at the cost of potential network congestion, while balancing network traffic reduces potential local congestion at the cost of increased message transfer hops.

In this study, we first present a comparative analysis exploring the trade-off between localizing communication and balancing network traffic using trace-based simulations, and demonstrate the effect of external network interference by introducing background traffic and show that localized communication can help reduce the application performance variation caused by network sharing. We then introduce an online simulation framework that improves performance and scalability, and discuss the validation of the simulation observations to a production Dragonfly system in respect of performance variability.

A Deferred Correction Coupling Strategy for Cosmological Simulations

Jean M. Sexton (Lawrence Berkeley National Laboratory)

We present a strategy for coupling the equations of baryonic gas dynamics to those for radiative heating and cooling in the context of Nyx, an N-body + gas dynamics code for large-scale cosmological simulations. The radiative and advective processes have differing time scales. To avoid Strang splitting errors, we couple the processes with a simplified spectral deferred correction strategy. Preliminary results indicate that the improved coupling strategy improves stability and reduces the cost of integrating the heating and cooling terms for our Lyman-alpha forest test case.

Deep Learning: Extrapolation Tool for Computational Nuclear Physics

Gianina Alina Negoita (Iowa State University)

The goal of nuclear theory is to understand how nuclei arise from interacting nucleons based on the underlying theory of the strong interactions, quantum chromodynamics (QCD). The interactions among the nucleons inside a nucleus are dominated by the strong interaction, which is non-perturbative in the low-energy regime relevant for nuclear physics. With access to powerful High Performance Computing (HPC) systems, several ab initio approaches have been developed to study nuclear structure and reactions, such as the No-Core Shell Model (NCSM). The NCSM and other approaches require an extrapolation of the results obtained in a finite basis space to the infinite basis space limit and assessment of the uncertainty of those extrapolations. Each observable requires a separate extrapolation and most observables have no proven extrapolation method at the present time. We propose a feed-forward artificial neural network (ANN) method as an extrapolation tool to obtain the ground state energy and the ground state point proton root-mean-square (rms) radius and their extrapolation uncertainties. We have generated theoretical data for 6Li by performing ab initio NCSM calculations using basis spaces up through the largest computationally feasible basis space. The designed ANNs are sufficient to produce results for these two very different observables in ^6Li from the ab initio NCSM results in small basis spaces that satisfy the following theoretical physics condition: independence of basis space parameters in the limit of extremely large matrices. Comparisons of the ANN results with other extrapolation methods are also provided.

Kinetic Simulations of Plasma Turbulence Using the Discontinuous Galerkin Finite Element Method

Tess Bernard (University of Texas)

We use an advanced computational framework called Gkeyll to simulate plasma turbulence in the edge of magnetic confinement fusion experiments. In these experiments, charged particles gyrate quickly around magnetic field lines. When time scales of interest are much greater than the gyration period and the wavelengths of disturbances parallel to field lines are much larger than the gyro-radius, one can average over the gyro-period of the plasma particles and effectively model them as rings. This gyrokinetic formulation reduces the particles' probability density function from six dimensions (three spatial and three velocity) to five (three spatial and three velocity) and is the model we use in our research. Gkeyll uses the discontinuous Galerkin (DG) finite element method, which combines the benefits of finite element methods, such as high-order accuracy, with those of finite volume methods, including locality of data. DG also allows for conservative schemes. We have benchmarked the code by modeling the Texas Heliamk, a basic plasma physics experiment, and comparing with experimental data. We have also developed a new and faster version of the code with improved conservation properties. This research demonstrates Gkeyll's progress toward 5D simulations of the edge region of fusion devices.

FLAME GPU: Complex System Simulation Framework

Mozhgan Kabiri Chimeh (University of Sheffield)

Agent based modeling can be used to represent complex systems at differing biological scales

FLAME GPU (Flexible Large Scale Agent Based Modelling Environment) : • Aim: High performance

agent based simulation without writing lots of GPU code ● Technique: Agents are state machines with memory (communicating X-Machines) o Uses state based agents to automatically generate CUDA GPU simulation code ● Optimisations: Reduce divergence and use efficient data structures/algorithms for common agent based behaviours ● Scope: Large numbers of relatively simple discrete time step agents ● Examples: o Swarm simulations o Cellular level simulation of epithelium tissue growth o Pedestrian simulation

Error Analysis in HPC Applications Using Algorithmic Differentiation

Harshitha Menon (Lawrence Livermore National Laboratory)

Computer applications running on supercomputers are used to solve critical problems. These systems are expected to perform tasks not just quickly, but also correctly. Various factors that can affect correctness of programs include faults, reduced precision, lossy data reduction, iteration and truncation. In the presence of these errors, how do we know whether our program is producing correct results? I have developed a method to understand the impact of these errors on a computer program. The method employs algorithmic differentiation (AD) to analyze the sensitivity of the simulation output to errors in program variables. A tool that we developed based on this method evaluates a given computer program and identifies vulnerable regions that need to be protected from errors. We use this to selectively protect variables against Silent Data Corruptions (SDC). We also use this method to study floating point sensitivity of the code and develop mixed-precision configurations to achieve performance improvement without affecting accuracy. Using this tool we can ensure that the computer simulation applications give us the correct results in the presence of these errors, so that scientists and policy makers relying on these results can make accurate predictions that can have lasting impact.

Power Aware Heterogeneous Node Assembly

Bilge Acun (IBM Research)

To meet ever increasing computational requirements, supercomputers and data centers are beginning to utilize fat compute nodes with multiple hardware components such as manycore CPUs and accelerators. These components have intrinsic power variations even among same model components from same manufacturer. In this paper, we argue that node assembly techniques that consider these intrinsic power variations can achieve better power efficiency without any performance trade off on large scale supercomputing facilities and data centers. We propose three different node assembly techniques: (1) Sorted Assembly, (2) Balanced Power Assembly, and (3) Application-Aware Assembly. In Sorted Assembly, node components are categorized (or sorted) into groups according to their power efficiency, and components from the same group are assembled into a node. In Balanced Power Assembly, components are assembled to minimize node-to-node power variations. In Application-Aware Assembly, the most heavily used components by the application are selected based on the highest power efficiency. We evaluate the effectiveness and cost savings of the three techniques compared to the standard random assembly under different node counts and variability scenarios.

Using a Robust Metadata Management System to Accelerate Scientific Discovery at Extreme Scales

Margaret Lawson (University of Illinois, Sandia National Laboratories)

Large-scale scientific simulations are an important tool for scientific discovery. In recent years, there has been a rapid growth in the amount of data output by these simulations. Extended runs of simulations such as XGC edge plasma fusion can easily generate datasets in the terabyte to petabyte range. With such large datasets, it is no longer feasible for scientists to load entire simulation outputs in search of features of interest. Scientists need an efficient, low-memory usage way of identifying which simulations produce a phenomenon, when and where the phenomenon appears, and how the phenomenon changes over time. However, current I/O systems such as HDF, NetCDF, and ADIOS do not provide these metadata capabilities. While some alternative tools have been developed that are optimized for a single type of analysis (global, spatial or temporal), no system provides an efficient way to perform all of these types of analysis. To fill this need, I have created EMPRESS, an RDBMS-based metadata management system for extreme scale scientific simulations. EMPRESS offers users the ability to efficiently tag and search features of interest without having to read in the associated datasets. Users can then use this metadata to perform spatial, temporal or global analysis and make discoveries. EMPRESS has been tested using several of Sandia's capacity clusters. Testing has primarily involved 1000, 2000, and 4000 cores, but several 8000 core tests were performed as well. Testing has proved that EMPRESS offers vastly better performance on these vital metadata functions than HDF5.

From Message Passing to PGAS

Hadia Ahmed (Lawrence Berkeley National Laboratory)

The continuous increase in the number of nodes across the whole system in supercomputers makes Remote-memory-access (RMA) a good alternative for improving applications' performance on these machines. However, in order for existing applications to benefit from it, they have to change their code to introduce the RMA functionalities. Doing so manually is error-prone.

In our research, we are working to develop a translator tool to generate UPC++ code that has RMA functionality. UPC++ is a C++11 library providing classes and functions that includes RMA communication operations.

In our first prototype, the translator uses annotations to guide transformation. First, it abstracts the message as a global pointer and enables ranks to distribute their local copies of their messages to other ranks. Then, it translates the MPI two-sided communication to corresponding UPC++ (RMA-rput) and a local synchronization. As a proof of concept, we applied this translator to stencil-method solver applications. The translator successfully generated a correct UPC++ code that achieves similar performance to the MPI version.

With stencil applications we do not expect better performance from corresponding MPI version. However, we were able to demonstrate that we can produce correct UPC++ implementation. Our next step is to apply the translator on applications that demonstrates irregular communication and/or fine-grained messaged. These are expected to show better performance with UPC++ under distributed memory architectures. The overall goal of this work is to help scientist in adapting their code to use PGAS effortlessly. Hence, they benefit from HPC resources and achieve better performance.

In Situ Performance Analysis of Event-driven Simulations to Support the Codesign of Extreme-Scale Systems

Caitlin J. Ross (Rensselaer Polytechnic Institute)

Parallel discrete event simulation (PDES) is a productive and cost-effective tool in exploring the design space of high performance computing (HPC) system architectures. In discrete event simulation, the modeled entities (e.g., network router) interact through the exchange of timestamped messages. For parallel simulations, the synchronization protocol ensures causal correctness of the simulation. In PDES, synchronization protocols are classified as conservative or optimistic. Conservative methods ensure that events are processed only when it is safe to do so, while optimistic methods allow for speculative processing of events and provide out-of-order event detection and recovery mechanisms. Because optimistic protocols can better exploit the parallelism inherent in models, optimistic simulations tend to be more scalable than conservative simulations. However, optimizing optimistic simulations to minimize the time spent performing unproductive work (i.e., rolling back the simulation state to fix causality errors) is not a trivial task. The factors that affect optimistic performance exist at multiple levels of the simulation, from the physical hardware running the simulation, communication of MPI ranks, as well as characteristics of the simulated model itself. The interplay of these factors is difficult to understand, making it difficult for optimistic PDES to be used efficiently, especially by simulation users, such as network architects.

Characterization of the Impact of Soft Errors on Iterative Methods

Burcu Mutlu (Pacific Northwest National Laboratory, Polytechnic University of Catalonia)

Soft errors caused by transient bit flips can lead to silent data corruption which can significantly impact an application's behavior. When a transient bit flip affects a hardware component, the application is said to be impacted by a soft error. When a soft error escapes hardware detection and impacts the application state, it can impact execution by leading to incorrect results or significantly impacting application execution times. Recent architectural trends, such as near-threshold voltage operation and constrained power budgets, exacerbate the frequency and impact of soft errors. This has motivated design of numerous soft error detection and correction techniques, which focuses localizing the application vulnerabilities, and then correcting these errors using micro-architectural, architectural, compilation-based, or application-level techniques.

A broad array of techniques has been designed to understand application behavior under soft errors and to detect, isolate, and correct soft-error-impacted application state. The first step toward tolerating soft errors involves understanding an application's behavior under soft errors. This can help elucidate the need for error detection/correction techniques.

In this study, we designed a deterministic application level error injection method that allows us explore the error injection space over several iterative solvers. As iterative methods are a crucial component of scientific applications, and consume a significant fraction of supercomputing time, we performed this study on the iterative methods. We used real life datasets to evaluate and characterize vulnerabilities and strengths of each solver under identical settings.

Use Cases of Neuromorphic Co-Processors in Future HPC Environments

Catherine Schuman (Oak Ridge National Laboratory)

With the looming end of Moore's law and the end of Dennard scaling, the HPC community is exploring

the use of specialized hardware as accelerators for certain tasks. Neuromorphic computing is a field in which neural networks are implemented in hardware to achieve intelligent computation with lower power and on a smaller footprint than traditional von Neumann architectures. Neuromorphic systems are compelling candidates for inclusion as co-processors in future HPCs, and they are suitable as co-processors for multiple types of applications. Here, we discuss neuromorphic systems as machine learning and graph algorithm accelerators. As more users are exposed to neuromorphic systems, we anticipate that even more use cases will arise.

Optimization of a Lattice Boltzmann Program

Muhong Zhou (Numerical Algorithms Group, BP Center For High Performance Computing)

In this project, I optimized a Lattice Boltzmann (LB) program developed internally by the digital rock team at BP to simulate single-phase flow and predict permeability of porous rocks. LB algorithm is simple to formulate and implement, and it is easily parallelizable on computer clusters. My work in this project includes adding parallel input/output, building solid side-wall boundary condition, and optimizing code runtime through improving input load balance and automatic loop vectorization. These runtime optimizations lead to a 3x speedup over the current (now previous) production LB code.

Plasma Meets Portability: A Journey to Performance Portability in a Particle-in-Cell Code

Joy Kitson (University of Delaware, Los Alamos National Laboratory)

In our journey to exascale, it is imperative that applications not only run efficiently, but do so on a wide variety of platforms at a reasonable cost. Codes must thus deal with a explosion of diversity among emerging architectures while avoiding a corresponding explosion in code size. This requires that HPC developers make their codes performance portable, that is, able to efficiently utilise the resources of each targeted platform, while simultaneously building a codebase in which they can work productively. For our work, we focus primarily on the Vector Particle in Cell (VPIC) application. VPIC is a plasma physics code that is highly performant, but past efforts to make it more portable have come at a heavy maintenance cost. In order to alleviate this cost, we leverage a portability framework called Kokkos to maximize both the code shared between platforms and the performance of the application. Throughout the port, we apply a logging tool to gather data on the development process.

We also analyze a VPIC Kernel which was successfully ported to Kokkos. The ported version can now run on GPUs, where the original cannot, and on CPUs, where their performance is on par with the original. This is accomplished with a single codepath. Comparing performance across all tested platforms shows that the port increased the application's performance efficiency. We couple this performance assessment with analysis of the port itself, based on data from the logging tool. Preliminary analysis of the data shows three stages in the development workflow.

Title: Distributed Memory Fast Fourier Transforms in the Exascale Era

Samar Aseeri (King Abdullah University of Science and Technology)

In this project, we plan to improve the implementation of the Fast Fourier Transform (FFT) on emerging high-performance computing architectures and galvanize the international community of experts to benchmark our and their efforts, leading the faster and more efficient adoption of the best

FFT software Technology.

FFT is an accurate low computational cost algorithm that is widely used for high-performance computing and engineering applications and it will still be needed on exascale computers. The implementations of the FFT (and other fast transforms) require many long-range interprocessor communications, which are a parallelization bottleneck. Therefore, the identification of alternative algorithms to the FFT along with comparisons of efficiency will lead to optimal use of high-performance computers. In this lightning talk, I will highlight the project objective, approach, and expected outcome.

Challenges of Performance Portability for Fortran Unstructured Mesh Codes
Abigail Hsu (Stony Brook University, Los Alamos National Laboratory)

What pathways exist for Fortran performance portability to exascale? Fortran-based codes present different challenges for performance portability and productivity. Ideally, we want to develop one codebase that can run on many different HPC architectures. In reality, each architecture has its own idiosyncrasies, requiring architecture-specific code. Therefore, we strive to write code that is as portable as possible to minimize the amount of development and maintenance effort. This project investigates how different approaches to parallel optimization impact the performance portability for unstructured mesh Fortran codes. In addition, it explores the productivity challenges due to the software tool and compiler support limitations unique to Fortran. For this study, we use the Truchas software, a casting manufacturing simulation code, and develop initial ports for OpenMP CPU, OpenMP offload GPU. and CUDA for computational kernels. There is no CUDA Fortran compiler compatible with Truchas, it must rewrite kernel in CUDA C and have the interface linked for C function calls in Fortran. Meanwhile, only the IBM xlf compiler is supported for OpenMP offload GPU at this moment and it is still immature. In additional of the difficulty that Fortran brings, the unstructured mesh uses more complex data access patterns. From the analysis of the Truchas gradient calculation computational kernel, we show some success for performance and portability along with some issues unique to Fortran using unstructured mesh. Through this study, we hope to encourage users and venders to focus on the productive pathways to developing Fortran applications for exascale architectures.

Large-Scale PDE-Constrained Optimization
Oana Marin (Argonne National Laboratory)

Optimization of time-dependent PDE-constrained optimization problems is extremely challenging from a computational perspective. Presume one forward simulation of a differential equation with N degrees of freedom, advancing in time for M timesteps requires a time to solution T. In this case optimizing for a certain parameter constrained by the PDE takes at least 2kT, where k is the number of iterations up to the convergence of the optimization scheme. We hereby explore strategies for achieving maximum speedup under controlled incurred errors. It is noteworthy that high errors in the computation of the gradient increase the number of iterations required to achieve convergence of the optimization algorithm, which is, in essence, more damaging than any gains made in the computation of the PDE.

Women in HPC – Workshop Afternoon Break

Developing Workplace Resilience and Managing Stress *Toni Collis (Appentra Solutions)*
The first part of this session "strategies to build resilience" (1 hour) will provide a hands on discussion session working through the resilience toolkit. The second half will be a "Panel Discussion and Interactive session"— identifying your personal triggers, how to manage triggers and resilience.

Room: D222
10:30 am - 5:30 pm

2nd ATIP Workshop on International Next-Generation Computing Programs and Workforce Development

Session Description: SC18 is the 30th anniversary of the SC conference series. In the spirit of highlighting that accomplishment, ATIP's 2018 Workshop emphasizes country-by-country achievements that have been made towards exascale and next-generation computing as well as plans and strategies for the future, including mechanisms to expand the pool of qualified HPC practitioners. The workshop will conclude with a panel composed of representatives from the Student Cluster Competition teams competing this year. Each team representative will introduce themselves and their team to the larger international HPC community and speak briefly about their work.

After ATIP's very successful Workshop at SC17 on International Exascale and Next-Generation Computing Programs, ATIP proposes a full-day sequel workshop at SC18 presenting and generating discussions on International Next-Generation Computing Programs and Workforce Development. SC18 will be the 30th anniversary of the SC conference series. In the spirit of highlighting that accomplishment our proposed 2018 Workshop will be more inclusive and incorporate more younger scientists. It will emphasize achievements that have been made and plans for the future, including mechanisms to expand the pool of qualified HPC practitioners. Speakers will be selected from leaders in Asia, Europe, and possibly elsewhere.

Welcome and Introduction *David Kahaner (Asian Technology Information Program), Taisuke Boku (University of Tsukuba)*

Panel 1: Role of Federated Polish HPC Centers in Polish AI Initiatives and EuroHPC Program *Marek Michalewicz (Interdisciplinary Center for Mathematical and Computational Modeling, University of Warsaw)*

Panel 1: A Site-local View of Creating a Pan-European Federated Research Infrastructure *Colin McMurtrie (Swiss National Supercomputing Centre)*

Panel 1: Service-Oriented HPC and Data Infrastructures for Science in Germany *Dirk Pleiter (Forschungszentrum Juelich)*
Addressing grand challenges in sciences requires HPC and data infrastructures that facilitate collaboration and data sharing in combination with the capability to store vast amounts of data as well as to access and process them with high performance. In this talk we report on efforts of several European supercomputing centres towards Fenix, a federated high-performance compute and data infrastructure. This is currently being implemented by the ICEI project as part of the Human Brain Project. As such, the brain research community is the main driver for designing this infrastructure. It targets, however, also other science communities like materials science.

Panel 1: European Region Q&A / Discussion Moderated by Prof. Taisuke Boku

Panel 2: United States National Science Foundation (NSF) Office of Advanced Cyberinfrastructure Programs and Workforce Development *Sushil Prasad (National Science Foundation)*

Panel 2: Arabia's Leap into the Cyber Era *David Keyes (King Abdullah University of Science and Technology)*

Panel 2: HPC in India
Sanjay Wandhekar (Centre for Development of Advanced Computing, India)

The talk will summarize various HPC projects in the country including the National Supercomputing Mission (NSM). The Mission focus is on Development and creation of HPC Facilities and Infrastructure in the country, HPC Applications Development, HPC proficient manpower development and advanced R&D for next generation supercomputing technologies. The talk will elaborate more on HPC manpower development efforts in the country.

Panel 2: Russian HPC Trends: a View From a Local Vendor Trench *Alexander Moskovsky (RSC Group)*
In Russia, high-performance computing (HPC) is considered as a technological priority, as stated in many official documents and statements on national scientific, research, and development strategies. A vivid scientist and engineer community emerged, including all major areas of HPC-related research, numerical simulation scientists, and engineering support from both domestic and international vendors. This talk is dedicated to the analysis of past and future trends in Russian HPC market based on data from both the worldwide Top500 and the Russian Top50 ratings.

Panel 2: US, Saudi Arabia, India and Russia Q&A / Discussion Moderated by Prof. Taisuke Boku

ATIP – Workshop Lunch (on your own)

Keynote: The Post-K for General-Purpose, Energy-Efficient, and Sustained Application Performance
Mitsuhisa Sato (Riken Center for Computational Science, RIKEN Advanced Institute for Computational

Science (AICS))

Panel 3: Challenge and Chance for Supercomputing Center in China *Yutong Lu (National Supercomputer Center Guangzhou)*
Nowadays, advanced computing and visualizing tools are facilitating scientists and engineers to perform virtual experiments and analyze large-scale datasets. Computing-driven and BigData-driven scientific discovery has become a necessary approach in global environment change, life science, nano-materials, high energy physics, and other fields. Furthermore, the increased computing requirements from economic and social development also call for the birth of the exascale system. This talk will discuss the challenge and chance for Supercomputing Center in China, and our efforts for the convergence of the HPC, BigData, and AI to embrace the new era.

Panel 3: Japan's HPC Program for System Development and Deployment toward Exascale *Taisuke Boku (University of Tsukuba, Center for Computational Sciences)*
In this talk, activities in Japan toward national flagship machine development and other supporting Tier-2 system deployment are introduced. Currently, Post-K Computer is under development toward its full operation starting on 2021 (Flagship2020 Project) and many of key issues and performance values have been shared. On the other hand, under HPCI (HPC Infrastructure) program, 9 national universities and several national institutes provide their supercomputer resources for the advanced computational science/engineering as Tier-2 supporting systems. In this talk, the deployment plan and their role especially for bridging time from K Computer to Post-K Computer are introduced.

Panel 3: HPC Status in Thailand *Putchong Uthayopas (Kasetsart University, Thailand)*

Panel 3: The Path from HPC to AI in Taiwan's NCHC *Ce-Keun Shieh (National Center for High-Performance Computing, Taiwan)*
Artificial Intelligence (AI) has been regarded as one of the competitive edges of economy. The acceleration of compute cycles, which is often regarded as high-performance computing, is traditionally considered as the strength of scientific advance in research. It is now turned into the powerhouse of AI. NCHC was commissioned by Taiwan government to develop a new generation of acceleration platform, namely Taiwan Cloud Computing (TWCC). TWCC cleverly integrates 2016 Tesla V100 GPUs into a HPC system and, by the end of this year, the system will be ready with capacity of more than 8 PF performance and 50 PB storage. The energy awareness of the system built will lead to an exceptional power usage effectiveness of 1.2. This significant development will support and bridge Taiwan's academics and industry as the compute backend of AI innovation. By plug-in into the cloud infrastructure, NCHC is able to upscale AI with HPC and to benefit the two worlds of science and economics and satisfy both scientific query and real life needs. In this presentation, I will give an overview of Taiwan AI strategy and the role of NCHC play in the strategy. Further, I will give you how we develop TWCC to support the Taiwan's eco-system of AI, which includes stakeholders of general users, domain experts, system developers and operators and conclude in our future plan for the next 3-year horizon.

Panel 3: Asian Region Q&A / Discussion Moderated by Dr. David Kahaner

ATIP – Workshop Afternoon Break

Introduction to Student Cluster Competitions *Dan Olds (Gabriel Consulting Group)*

Panel 4: Asia Supercomputing Community: Profound Inspiration through Strong Competition *Vangel Bojaxhi (Inspur, Asia Supercomputer Community (ASC))*

Panel 4: Student Spotlight Presentation *Julia Bazińska (University of Warsaw)*

Student Cluster Competition Team Panel Presentation *Adam Sobecki (University of Warsaw)*

Student Cluster Competition Team Panel Presentation *Michael Lau (Texas A&M University)*

Student Cluster Competition Team Panel Presentation *Evan Donato (University of Massachusetts, Boston)*

Student Cluster Competition Team Panel Presentation *Eva Dengler (University of Erlangen-Nuremberg), David Sauerwein (University of Erlangen-Nuremberg)*

Student Cluster Competition Team Panel Presentation *Yu-Hsuan Cheng (National Tsing Hua University, Taiwan)*

Student Cluster Competition Team Panel Presentation *Meiro Hao (Nanyang Technological University, Singapore)*

Student Cluster Competition Team Panel Presentation *P. Nigel Brown (Laney College)*

Student Cluster Competition Team Panel Presentation *Jason Booth (Northeastern University)*

Student Cluster Competition Team Panel Presentation *Jiping Yu (Tsinghua University)*

Student Cluster Competition Team Panel Presentation *Rafif Akila Hakim (Telkom University, Indonesia), M. Garda Khadafi (Telkom University, Indonesia)*

Student Cluster Competition Team Panel Presentation *Roshan Rajan (University of Illinois)*

Student Cluster Competition Team Panel Presentation *Jia Lin Cheoh (Purdue University)*

Student Cluster Competition Team Panel Presentation *Alex Balmer (Illinois Institute of Technology)*

Student Cluster Competition Team Panel Presentation *Paul Kefer (Wake Forest University)*

Student Cluster Competition Team Panel Presentation *Grace Han (Monash University)*

Panel 4: Student Competition and Workforce Development Q&A / Discussion Moderated by Dan Olds

Networking / Independent Discussion
After all questions have been answered, the remaining time can be used for networking and one-on-one discussion.

Room: D172
2:00 pm - 5:30 pm

HPC Systems Professionals Workshop (HPCSYSPROS18)

In order to meet the demands of high performance computing (HPC) researchers, large-scale computational and storage machines require many staff members who design, install, and maintain these systems. These HPC systems professionals include system engineers, system administrators, network administrators, storage administrators and operations staff who face problems that are unique to high performance computing systems. While many conferences exist for the HPC field and the system administration field, none exist that focus on the needs of HPC systems professionals. Support resources can be difficult to find to help with the issues encountered in this specialized field. Often times, systems staff turn to the community as a support resource and opportunities to strengthen and grow those relationships are highly beneficial. This workshop is designed to share solutions to common problems, provide a platform to discuss upcoming technologies, and to present state of the practice techniques so that HPC centers will get a better return on their investment, increase performance and reliability of systems, and researchers will be more productive.

HPCSYSPROS18: Keynote *Bill Anderson (National Center for Atmospheric Research)*
This talk will share knowledge and insights about best system engineering practices in managing massive amounts of data. Drawing on decades of experience at the National Center for Atmospheric Research, the presentation will cover topics such as data integrity, performance, and issues when scaling storage systems to very large sizes. The presentation will also share examples of what's

worked well and what hasn't over the years.

Message from the SIGHPC SYSPROS Virtual Chapter President *Jenett Tillotson (Purdue University)*

HPCSYSPROS18 – Workshop Afternoon Break

Studying Effects of Meltdown and Spectre Patches on the Performance of HPC Applications Using Application Kernel Module of XDMoD
Nikolay A. Simakov (State University of New York at Buffalo)

In this work we examine how the updates addressing Meltdown and Spectre vulnerabilities impact the performance of HPC applications. To study this we use the application kernel module of XDMoD to test the performance before and after the application of the vulnerability patches. The application kernel module is designed for continuous performance monitoring of HPC systems. Here, we tested the performance difference for multiple applications and benchmarks including: NWChem, NAMD, GAMESS, ENZO, HPCC, IOR, MDTest and IMB. The results show that although some specific functions can have execution times decreased by as much as 74%, the majority of individual metrics indicate little to no decrease in performance. The real-world applications show a 2-3% decrease in performance for single node jobs and a 5-11% decrease for two node jobs. For node-counts up to 8 the degradation continues to increase reaching 27% in some cases.

Compliant Cloud+Campus Hybrid HPC Infrastructure
Matt Vander Werf (University of Notre Dame)

We present a hybrid cloud+campus model for deploying secure research cyberinfrastructure in line with a growing number of federally funded scientific research security requirements. While the security mechanisms are inline with many regulations, we specifically focus on the NIST 800-171 CUI compliance that is now required by many US DARPA funded contracts. We discuss our architectural framework and rationale for leveraging shared ND services in AWS GovCloud in concert with CUI compliant HPC systems on our campus in a hybrid fashion; allowing an individual CUI regulated research project to bridge two connected but distinct infrastructures.

cgroups py : Using Linux Control Groups and Systemd to Manage CPU Time and Memory
Curtis Maves (Purdue University)

Cgroups provide a mechanism to limit user andprocess resource consumption on Linux systems. This paper discusses cgroups_py, a Python script that runs as a systemd service that dynamically throttles users on shared resource systems, such as HPC cluster front-ends. This is done using the cgroups kernel API and systemd.

Making Container Easier with HPC Container Maker
Scott McMillan (Nvidia Corporation)

Containers make it possible for scientists and engineers to easily use portable and reproducible software environments. However, if the desired application is not already available from a container registry, generating a custom container image from scratch can be challenging for users accustomed to a bare metal software environment. HPC Container Maker (HPCCM) is an open source project to address these challenges. HPCCM provides a high level, configurable Python recipe format for deploying HPC components into container images according to best practices. HPCCM recipes are more portable and powerful than native container specification formats, easier to create, and produce smaller, well-constructed container images.

Rapid Deployment of Bare-Metal and In-Container HPC Clusters Using OpenHPC playbooks
Violeta Holmes (University of Huddersfield)

In this paper, we present a toolbox of reusable Ansible roles and playbooks in order to configure a cluster software environment described by the freely available OpenHPC recipes. They can be composed in order to deploy a robust and reliable cluster environment, instilled with the best practice offered by the OpenHPC packaging, and the repeatability and integrity guarantees of the configuration managed approach. With container virtualization setting a new trend in scientific software management, we focus this effort on supporting the deployment of such environments on both bare-metal and container-based targets.

xCAT and Masterless Puppet: Aiming for Ideal Configuration Management
Jason St. John (Purdue University)

Configuration management is a major factor in the design of an HPC system. This paper provides an evaluation of the new architecture of the HPC systems operated by Purdue University, focusing on the configuration management system.

Stateless Provisioning: Modern Practice in HPC
John Blaas (University of Colorado), John Roberts (Argonne National Laboratory)

We outline a model for creating a continuous integration and continuous delivery work flow targeted at provisioning CPIO based initramfs images that are used to run computational work nodes in a bare metal cluster running RHEL or CentOS.

Next-Generation Cluster Management Software
Paul Peltz Jr (Los Alamos National Laboratory)

Over the last six decades, Los Alamos National Laboratory (LANL) has acquired, accepted, and integrated over 100 new HPC systems, from MANIAC in 1952 to Trinity in 2017. These systems range from small clusters to large supercomputers. The high performance computing (HPC) system architecture has progressively changed over this time as well; from single system images to complex, interdependent service infrastructures within a large HPC system. The authors are proposing a redesign of the current HPC system architecture to help reduce downtime and provide a more resilient architectural design.

HPCSYSPROS18: Author Panel
Questions for presenters from the HPCSYSPROS18 workshop.

Upcoming Events in the HPC Systems Professionals Community *Neil Bright (Georgia Institute of Technology), Rich Knepper (Cornell University), Alana Romanella (Virginia Tech), Marisa Brazil (Purdue University)*
Discussion of upcoming events in the HPC Systems Professionals community

Room: D167/174
2:00 pm - 5:30 pm

Machine Learning in HPC Environments

The intent of this workshop is to bring together researchers, practitioners, and scientific communities to discuss methods that utilize extreme scale systems for machine learning. This workshop will focus on the greatest challenges in utilizing HPC for machine learning and methods for exploiting data parallelism, model parallelism, ensembles, and parameter search. We invite researchers and practitioners to participate in this workshop to discuss the challenges in using HPC for machine learning and to share the wide range of applications that would benefit from HPC powered machine learning.

Deep Learning Evolutionary Optimization for Regression of Rotorcraft Vibrational Spectra
Daniel A. Martinez-Gonzalez (US Army Engineer Research and Development Center), Wesley Brewer (US Department of Defense HPC Modernization Program)

A method for Deep Neural Network (DNN) hyperparameter search using evolutionary optimization is proposed for nonlinear high-dimensional multivariate regression problems. Deep networks often lead to extensive hyperparameter searches which can become an ambiguous process due to network complexity. Therefore, we propose a user-friendly method that integrates Dakota optimization library, TensorFlow, and Galaxy HPC workflow management tool to deploy massively parallel function evaluations in a Genetic Algorithm (GA). Deep Learning Evolutionary Optimization (DLEO) is the current GA implementation being presented. Compared with random generated and hand-tuned models, DLEO proved to be significantly faster and better searching for optimal architecture hyperparameter configurations. Implementing DLEO allowed us to find models with higher validation accuracies at lower computational costs in less than 72 hours, as compared with weeks of manual and random search. Moreover, parallel coordinate plots provided valuable insights about network architecture designs and their regression capabilities

Machine Learning – Workshop Morning Break

Ramifications of Evolving Misbehaving Convolutional Neural Network Kernel and Batch Sizes
Mark Coletti (Oak Ridge National Laboratory)

Deep-learners have many hyper-parameters including learning rate, batch size, kernel size --- all playing a significant role toward estimating high quality models. Discovering useful hyper-parameter guidelines is an active area of research, though the state of the art generally uses a brute force, uniform grid approach or random search for finding ideal settings. We share the preliminary results of using an alternative approach to deep learner hyper-parameter tuning that uses an evolutionary algorithm to improve the accuracy of a deep-learner models used in satellite imagery building footprint detection. We found that the kernel and batch size hyper-parameters surprisingly differed from sizes arrived at via a brute force uniform grid approach. These differences suggest a novel role for evolutionary algorithms in determining the number of convolution layers, as well as smaller batch sizes in improving deep-learner models.

Auto-Tuning TensorFlow Threading Model for CPU Backend
Niranjan Hasabnis (Intel Corporation)

TensorFlow is a popular deep learning framework used to solve machine learning and deep learning problems such as image classification and speech recognition. It also allows users to train neural network models or deploy them for inference using GPUs, CPUs, and custom-designed hardware such as TPUs. Even though TensorFlow supports a variety of optimized backends, realizing the best performance using a backend requires additional efforts. Getting the best performance from a CPU backend requires tuning of its threading model. Unfortunately, the best tuning approach used today is manual, tedious, time-consuming, and, more importantly, may not guarantee the best performance.

In this paper, we develop an automatic approach, called TENSORTUNER, to search for optimal parameter settings of TensorFlow's threading model for CPU backends. We evaluate TENSORTUNER on both Eigen and Intel's MKL CPU backends using a set of neural networks from TensorFlow's benchmarking suite. Our evaluation results demonstrate that the parameter settings found by TENSORTUNER produce 2% to 123% performance improvement for the Eigen CPU backend and 1.5% to 28% performance improvement for the MKL CPU backend over the performance obtained using their best-known parameter settings. This highlights the fact that the default parameter settings in Eigen CPU backend are not the ideal settings; and even for a carefully hand-tuned MKL backend, the settings are sub-optimal. Our evaluations also revealed that TENSORTUNER is efficient at finding the optimal settings — it is able to converge to the optimal settings quickly by pruning more than 90% of the parameter search space.

Automated Parallel Data Processing Engine with Application to Large-Scale Feature Extraction
Kesheng Wu (Lawrence Berkeley National Laboratory)

As new scientific instruments generate ever more data, we need to parallelize advanced data analysis algorithms such as machine learning to harness the available computing power. The success of commercial Big Data systems demonstrated that it is possible to automatically parallelize these algorithms. However, these Big Data tools have trouble handling the complex analysis operations from scientific applications. To overcome this difficulty, we have started to build an automated parallel data processing engine for science, known as SystemA1. This paper provides an overview of this data

processing engine, and a use case involving a complex feature extraction task from a large-scale seismic recording technology, called distributed acoustic sensing (DAS). The key challenge associated with DAS is that it produces a vast amount of noisy data. The existing methods used by the DAS team for extracting useful signals like traveling seismic waves from this data are very time-consuming. Our parallel data processing engine reduces the job execution time from 100s of hours to 10s of seconds, and achieves 95% parallelization efficiency. We are implementing more advanced techniques including machine learning using SystemA, and plan to work with more scientific applications.

Training Speech Recognition Models on HPC Infrastructure
Deepthi Karkada (Intel Corporation)

Automatic speech recognition is used extensively in speech interfaces and spoken dialogue systems. To accelerate the development of new speech recognition models and techniques, developers at Mozilla have open sourced a deep learning based Speech-To-Text engine known as project DeepSpeech based on Baidu's DeepSpeech research. In order to make model training time quicker on CPUs for DeepSpeech distributed training, we have developed optimizations on the Mozilla DeepSpeech code to scale the model training to a large number of Intel® CPU system, including Horovod integration into DeepSpeech. We have also implemented a novel dataset partitioning scheme to mitigate compute imbalance across multiple nodes of an HPC cluster. We demonstrate that we are able to train the DeepSpeech model using the LibriSpeech clean dataset to its state-of-the-art accuracy in 6.45Hrs on 16-Node Intel® Xeon® based HPC cluster.

Room: D171
2:00 pm - 5:30 pm

Fifth International Workshop on Visual Performance Analysis (VPA 18)

Over the last decades an incredible amount of resources has been devoted to building ever more powerful supercomputers. However, exploiting the full capabilities of these machines is becoming exponentially more difficult with each new generation of hardware. To help understand and optimize the behavior of massively parallel simulations the performance analysis community has created a wide range of tools and APIs to collect performance data, such as flop counts, network traffic or cache behavior at the largest scale. However, this success has created a new challenge, as the resulting data is far too large and too complex to be analyzed in a straightforward manner. Therefore, new automatic analysis and visualization approaches must be developed to allow application developers to intuitively understand the multiple, interdependent effects that their algorithmic choices have on the final performance.

This workshop will bring together researchers from the fields of performance analysis and visualization to discuss new approaches of applying visualization and visual analytics techniques to large scale applications.

VPA18 Keynote

VPA18 – Workshop Afternoon Break

Visual Analytics Challenges in Analyzing Calling Context Trees

Performance analysis is an integral part of developing and optimizing parallel applications for high-performance computing (HPC) platforms. Hierarchical data from different sources is typically available to identify performance issues or anomalies. Some hierarchical data such as the calling context can be very large in terms of breadth and depth of the hierarchy. Classic tree visualizations quickly reach their limits in analyzing such hierarchies with the abundance of information to display. In this position paper, we identify the challenges commonly faced by the HPC community in visualizing hierarchical performance data, with a focus on calling context trees. Furthermore, we motivate and lay out the bases of a visualization that addresses some of these challenges.

Using Deep Learning for Automated Communication Pattern Characterization: Little Steps and Big Challenges

Philip C. Roth (Oak Ridge National Laboratory)

Characterization of a parallel application's communication patterns can be useful for performance analysis, debugging, and system design. However, obtaining and interpreting a characterization can be difficult. AChax implements an approach that uses search and a library of known communication patterns to automatically characterize communication patterns. Our approach has some limitations that reduce its effectiveness for the patterns and pattern combinations used by some real-world applications. By viewing AChax's pattern recognition problem as an image recognition problem, it may be possible to use deep learning to address these limitations. In this position paper, we present our current ideas regarding the benefits and challenges of integrating deep learning into AChax and our conclusion that a hybrid approach combining deep learning classification, regression, and the existing AChax approach may be the best long-term solution to the problem of parameterizing recognized communication patterns.

Visualizing Multidimensional Health Status of Data Centers

Tommy Dang (Texas Tech University)

Monitoring data centers is challenging due to their size, complexity, and dynamic nature. This project proposes a visual approach for situational awareness and health monitoring of high-performance computing systems. The visualization requirements are expanded on the following dimensions: 1) High performance computing spatial layout, 2) Temporal domain (historical vs. real-time tracking), and 3) System health services such as temperature, CPU load, memory usage, fan speed, and power consumption. To show the effectiveness of our design, we demonstrate the developed prototype on a medium-scale high-performance computing system of 10 racks and 467 hosts. The work was developed using feedback from both industrial and academic domain experts.

PaScal Viewer: A Tool for the Visualization of Parallel Scalability Trends
Samuel Xavier de Souza (Federal University of Rio Grande do Norte)

Taking advantage of the growing number of cores in supercomputers to increase the scalabilty of parallel programs is an increasing challenge. Many advanced profiling tools have been developed to assist programmers in the process of analyzing data related to the execution of their program. Programmers can act upon the information generated by these data and make their programs reach higher performance levels. However, the information provided by profiling tools are generally designed to optimize the program for a specific execution environment, with a target number of cores and a target problem size. A code optimization driven towards scalability rather than specific performance requires the analysis of many distinct execution environments instead of details about a single environment. With the goal of providing more useful information for the analysis and optimization of code for parallel scalability, this work introduces the PaScal Viewer tool. It presents an novel and productive way to visualize scalability trends of parallel programs. It consists of four diagrams that offers visual support to identify parallel efficiency trends of the whole program, or parts of it, when running on scaling parallel environments with scaling problem sizes.

Room: D175
2:00 pm - 5:30 pm

The 4th International Workshop on Data Reduction for Big Scientific Data (DRBSD-4)

As the speed gap between compute and storage continues to exist and widen, the increasing data volume and velocity pose major challenges for big data applications in terms of storage and analysis. This demands new research and software tools that can further reduce data by several orders of magnitude, taking advantage of new architectures and hardware available on next generation systems. This international workshop on data reduction is a response to this renewed research direction and will provide a focused venue for researchers in this area to present their research results, exchange ideas, identify new research directions, and foster new collaborations within the community. Topics of interest include but are not limited to: ● Application use-cases which can drive the community to develop MiniApps ● Data reduction methods for scientific data including: ● Data deduplication methods ● Motif-specific methods (structured and unstructured meshes, particles, tensors, ...) ● Optimal design of data reduction methods ● Methods with accuracy guarantees ● Metrics to measure reduction quality and provide feedback ● Data analysis and visualization techniques that take advantage of the reduced data ● Hardware and data co-design ● Accuracy and performance trade-offs on current and emerging hardware ● New programming models for managing reduced data ● Runtime systems for data reduction

Perspectives on Data Reduction from ASCR *Laura Biven (US Department of Energy Office of Advanced Scientific Computing Research), Lucy Nowell (US Department of Energy Office of Advanced Scientific Computing Research)*

DRBSD-4 – Workshop Afternoon Break *Qing Liu (New Jersey Institute of Technology)*

Feature-Relevant Data Reduction for In Situ Workflows
Will Fox (Massachusetts Institute of Technology)

As the amount of data produced by HPC simulations continues to grow and I/O throughput fails to keep up, in situ data reduction is becoming an increasingly necessary component of HPC workflows. Application scientists, however, prefer to avoid reduction in order to preserve data fidelity for post-hoc analysis. In an attempt to compromise between data quality and data quantity, this work introduces the concept of feature-relevant compression. We explore two scientific datasets in an attempt to quantify the impacts of compression on features of interest by identifying such features and analyzing changes in their properties after compression. We find that it is indeed possible to compress simulation data in a lossy manner while preserving desired properties within a predetermined error rate. Additionally, we suggest that this error quantification could be applied as part of an in situ workflow to dynamically tune compression parameters during simulation, compressing aggressively when features are simple but preserving structure where data complexity increases. Future work should focus on implementation, extension to additional compression algorithms, and analysis of these techniques on quantities which are derived from original simulation data.

A Statistical Analysis of Compressed Climate Model Data
Dorit Hammerling (National Center for Atmospheric Research)

The data storage burden resulting from large climate model simulations continues to grow. While lossy data compression methods can alleviate this burden, they introduce the possibility that key climate variables could be altered to the point of affecting scientific conclusions. Therefore, developing a detailed understanding of how compressed model output differs from the original is important. Here, we evaluate the effects of two leading compression algorithms, SZ and ZFP, on daily surface temperature and precipitation rate data from a popular climate model. While both algorithms show promising fidelity with the original output, detectable artifacts are introduced even at relatively low error tolerances. This study highlights the need for evaluation methods that are sensitive to errors at different spatiotemporal scales and specific to the particular climate variable of interest, with the ultimate goal to improve lossy compression collaboratively with the algorithm development teams.

Data Reduction Challenges in Coordinated Simulation and Experimental Fusion Science *Sean Dettrick (TAE Technologies)*

Exploring Best Lossy Compression Strategy By Combining SZ with Spatiotemporal Decimation
Xin Liang (University of California, Riverside)

In today's extreme-scale scientific simulations, vast volumes of data are being produced such that the data cannot be accommodated by the parallel file system or the data writing/reading performance will be fairly low because of limited I/O bandwidth. In the past decade, many snapshot-based (or space-based) lossy compressors have been developed, most of which rely on the smoothness of the data in

space. However, the simulation data may get more and more complicated in space over time steps, such that the compression ratios decrease significantly. In this paper, we propose a novel, hybrid lossy compression method by leveraging spatiotemporal decimation under the SZ compression model. The contribution is twofold. (1) We explore several strategies of combining the decimation method with the SZ lossy compression model in both the space dimension and time dimension during the simulation. (2) We investigate the best-fit combined strategy upon different demands based on a couple of typical real-world simulations with multiple fields. Experiments show that either the space-based SZ or time-based SZ leads to the best rate distortion. Decimation methods have very high compression rate with low rate distortion though, and SZ combined with temporal decimation is a good tradeoff.

Synthetic Data Generation for Evaluating Parallel I/O Compression Performance and Scalability

Sean B. Ziegeler (US Department of Defense HPC Modernization Program, Engility Corporation)

Compression is one of the most common forms of data reduction and is typically the least invasive. As compute capability continues to outpace I/O bandwidths, compression becomes that much more attractive. This paper explores the scalable performance of parallel compression and presents an in-depth analysis of a coherent noise algorithm to generate synthetic data that can be used to easily evaluate parallel compression. The algorithm favors simplicity, ease-of-use, and scalability over high-fidelity reconstruction of real data, so we go to lengths to show that the synthetic data generated is suitable as a proxy for evaluating compression, especially in benchmarks and mini-apps.

Amplitude-Aware Lossy Compression for Quantum Circuit Simulation

Classical simulation of quantum circuits is crucial for evaluating and validating the design of new quantum algorithms. However, the number of quantum state amplitudes increases exponentially with the number of qubits, leading to the exponential growth of the memory requirement for the simulations. In this paper, we present a new data reduction technique to reduce the memory requirement of quantum circuit simulations. We apply our amplitude-aware lossy compression technique to the quantum state amplitude vector to trade the computation time and fidelity for memory space. The experimental results show that our simulator only needs 1/16 of the original memory requirement to simulate Quantum Fourier Transform circuits with 99.95% fidelity. The reduction amount of memory requirement suggests that we could increase 4 qubits in the quantum circuit simulation comparing to the simulation without our technique. Additionally, for some specific circuits, like Grover's search, we could increase the simulation size by 18 qubits.

A Study on Checkpoints Compression for Adjoint Computation

Kai-Yuan Hou (Northwestern University)

When we want to understand the sensitivity of a simulation model with respect to an input value or to optimize an objective function, the gradient usually provides a good hint. The adjoint state method is a widely used numerical method to compute the gradient of a function. It decomposes functions into a sequence of basic operations. It performs a forward sweep to evaluate the function, followed by a backward sweep to calculate the gradient using the chain rule iteratively. One limitation of the adjoint

state method is that all intermediate values from the forward sweep are needed by the backward sweep. Usually, we keep only a portion of those values, called checkpoints, in the memory because of limited space. The remaining values are either stored on the hard disk or recomputed from the nearest checkpoint whenever needed. In this work, we seek to compress the intermediate values in order to better utilize limited space in the memory and to speed the I/O when checkpointing to the hard disk.

Room: D221
2:00 pm - 5:30 pm

Computational Reproducibility at Exascale 2018 (CRE2018)

Reproducibility is an important concern in all areas of computation. As such, computational reproducibility is receiving increasing interest from a variety of parties who are concerned with different aspects of computational reproducibility. Computational reproducibility encompasses several concerns including the sharing of code and data, as well as reproducible numerical results which may depend on operating system, tools, levels of parallelism, and numerical effects. In addition, the publication of reproducible computational results motivates a host of computational reproducibility concerns that arise from the fundamental notion of reproducibility of scientific results that has normally been restricted to experimental science.

The workshop addresses issues in reproducibility that arise when computing at exascale. It will include issues of numerical reproducibility as well as approaches and best practices to sharing and running code and the reproducible dissemination of computational results.

The workshop is meant to address the scope of the problems of computational reproducibility in HPC in general, and those anticipated as we scale up to exascale machines in the next decade. The participants of this workshop will include government, academic, and industry stakeholders; the goals of this workshop are to understand the current state of the problems that arise, what work is being done to deal with this issues, and what the community thinks the possible approaches to these problem are.

CRE2018 – Plenary I *Victoria Stodden (University of Illinois)*

Debugging and Optimization of HPC Programs in Mixed Precision with the Verrou Tool
François Févotte (EDF Research and Development)

Floating-Point (FP) arithmetic is becoming a hotter and hotter topic in High-Performance Computing (HPC). First, high computational performance is often achieved at the expense of a loss of control over the order in which FP operations are executed; second, optimizing the use of FP precision is often key to achieving high performance.

In this paper, we present how the Verrou tool can help deal with these issues in the context of large, industrial, high-performance scientific computing codes such as the ones developed and used by leading actors in the industry. In particular, we detail the various new features recently implemented in the Verrou tool, which allow performing the complete analysis of FP-related issues in mixed precision, high-performance codes. The historical stochastic arithmetic back-end of Verrou has been extended to handle specific issues related to the use of mixed precision in simulation codes, by instrumenting type conversions leading to round-off errors (such as double to float cast). A new arithmetic back-end has also been introduced in Verrou, allowing to emulate the use of a smaller precision in (part of) the code.

Together with the existing debugging and localization techniques proposed by Verrou, these new features make it possible to use the tool not only for diagnostics, but also as an help during the mixed-precision optimization process.

High Performance Implementation of Reproducible BLAS Routines with Tunable Accuracy Using Ozaki Scheme

Daichi Mukunoki (Tokyo Woman's Christian University)

This study presents a high performance implementation of Basic Linear Algebra Subprograms (BLAS) routines supporting reproducibility and tunable accuracy using the Ozaki scheme, which is an accurate matrix-multiplication method proposed by Ozaki et al. in 2011. The method ensures reproducibility and realizes tunable accuracy, including correct-rounding, by eliminating the effect of rounding-error in the computation. The most advantage of the method is that the method can be constructed based on level-3 BLAS. In this study, we show the implementation of three routines from level 1-3 BLAS: inner-product (DOT), matrix-vector multiplication (GEMV), and matrix-matrix multiplication (GEMM), with several optimization techniques for reducing the memory consumption and improving the performance. The performance evaluation on Titan V GPU demonstrates that our implementation achieves more than 73% of the expected peak performance.

Reproducibility for Streaming Analysis

Christopher J. Wright (Columbia University)

The natural and physical sciences increasingly need streaming data processing for live data analysis and autonomous experimentation. Furthermore, data provenance and replicability are important to assure the veracity of scientific results. Here we describe a software system that combines high performance computing, streaming data processing, and automatic data provenance capturing to address this need. Data provenance and streaming data processing share a common data structure, the directed acyclic graph (DAG), which describes the order of each computational step. Data processing requires the DAG to specify what computations to run in what order, and the execution can be recreated from the graph, reproducing the analyzed data and capturing provenance. In our framework the description and ordering of the analysis steps (the pipeline) are separated from their execution (the streaming analysis) and the DAG created for the streaming data processing is captured during data analysis. Streaming data can have high throughputs and our system allows users to choose among multiple parallel processing backends, including Dask. To guarantee reproducibility, unique links to the incoming data, and their timestamps are captured alongside the DAG. Analyzed data, along with provenance metadata, are stored in a database, which can re-run analysis from raw data, enabling verification of results, exploring how parameters change outcomes, and data processing

reuse. This system is running in production at the National Synchrotron Light Source-II (NSLS-II) x-ray powder diffraction beamlines.

CRE2018 – Workshop Afternoon Break

CRE218 – Plenary II *Bert Debusschere (Sandia National Laboratories)*

Understanding Simultaneous Impact of Network QoS and Power on HPC Application Performance

With the growing complexity of high performance computing (HPC) systems, application performance variation has increased enough to disrupt the overall throughput of the systems. Such performance variation is expected to worsen in the future, when job schedulers will have to manage flow resources such as power, network and I/O in addition to traditional resources such as nodes and memory. In this work, we study the simultaneous impact of inter-job interference, Infiniband service levels, and power capping on different applications in a controlled experimental setup, with the goal of understanding the range of performance variation as well as potential mitigation strategies.

Panel Discussion

Monday, November 12th

Room: D165
8:59 am - 5:30 pm

The 9th International Workshop on Performance Modeling, Benchmarking, and Simulation of High-Performance Computer Systems (PMBS18)

Session Description: The PMBS18 workshop is concerned with the comparison of high-performance computer systems through performance modeling, benchmarking or through the use of tools such as simulators. The aim of this workshop is to bring together researchers, from industry and academia, concerned with the qualitative and quantitative evaluation and modeling of high-performance computing systems. Authors are invited to submit novel research in all areas of performance modeling, benchmarking and simulation, and we welcome research that brings together current theory and practice. We recognize that the coverage of the term performance has broadened to include power consumption and reliability, and that performance modeling is practiced through analytical methods and approaches based on software tools and simulators.

The PMBS18 workshop is concerned with the comparison of high-performance computing systems through performance modeling, benchmarking or through the use of tools such as simulators. We are particularly interested in research which reports the ability to measure and make tradeoffs in software/hardware co-design to improve sustained application performance. We are also keen to capture the assessment of future systems, for example through work that ensures continued application scalability through peta- and exa-scale systems.

The aim of this workshop is to bring together researchers, from industry and academia, concerned with the qualitative and quantitative evaluation and modeling of high-performance computing systems. Authors are invited to submit novel research in all areas of performance modeling, benchmarking and simulation, and we welcome research that brings together current theory and practice. We recognize that the coverage of the term 'performance' has broadened to include power consumption and reliability, and that performance modeling is practiced through analytical methods and approaches based on software tools and simulators.

Performance modeling, benchmarking, and simulation will underpin software and hardware design choices as we advance towards the exascale era. This workshop continues to attract high quality input from industry, government, and academia.

Improving MPI Reduction Performance for Manycore Architectures with OpenMP and Data Compression

Hongzhang Shan (Lawrence Berkeley National Laboratory)

MPI reductions are widely used in many scientific applications and often become the scaling performance bottleneck. When performing reductions on vectors, different algorithms have been developed to balance messaging overhead and bandwidth. However, most implementations have ignored the effect of single-thread performance not scaling as fast as aggregate network bandwidth. In this work, we propose, implement, and evaluate two approaches (threading and exploitation of sparsity) to accelerate MPI reductions on large vectors when running on manycore-based supercomputers. Our benchmark results show that our new techniques improve the MPI_Reduce performance up to 4x and improve BIGSTICK application performance by up to 2.6x.

Exploring and Quantifying How Communication Behaviors in Proxies Relate to Real Applications

Omar Aaziz (Sandia National Laboratories), Jeanine Cook (Sandia National Laboratories)

Proxy applications, or proxies, are simple applications meant to exercise systems in a way that mimics real applications (their parents). However, characterizing the relationship between the behavior of parent and proxy applications is not an easy task. In prior work, we presented a data-driven methodology to characterize the relationship between parent and proxy applications based on collecting runtime data from both and then using data analytics to find their correspondence or divergence. We showed that it worked well for hardware counter data, but our initial attempt using MPI function data was less satisfactory. In this paper, we present an exploratory effort at making an improved quantification of the correspondence of communication behavior for proxies and their respective parent applications. We present experimental evidence of positive results using four proxy applications from the current ECP Proxy Application Suite and their corresponding parent applications (in the ECP application portfolio). Results show that each proxy analyzed is representative of its parent

with respect to communication data. In conjunction with our method presented earlier (correspondence between computation and memory behavior), we get a strong understanding of how well a proxy predicts the comprehensive performance of its parent.

PMBS18 – Workshop Morning Break

Deep Learning at Scale on Nvidia V100 Accelerators
Rengan Xu (Dell EMC)

The recent explosion in the popularity of Deep Learning (DL) is due to a combination of improved algorithms, access to large datasets and increased computational power. This had led to a plethora of open-source DL frameworks, each with varying characteristics and capabilities. End users are then left with the difficult task of determining software and hardware configurations to get optimal performance from each framework.

We share our experiences and develop best practices for DL training with TensorFlow, MXNet, and Caffe2. The paper also looks at DL inferencing with TensorRT on Nvidia V100 "Volta" GPUs. It focuses on one of the more prominent neural network architectures, Resnet50, combined with Imagenet dataset. We quantify the impact of hardware attributes on DL workloads such as the usage of PCIe vs NVLink GPUs, performance past a single worker node, effect of high speed interconnect such as InfiniBand EDR on training, and the implication of utilizing a network attached storage and its advantages.

Benchmarking Machine Learning Methods for Performance Modeling of Scientific Applications
Prasanna Balaprakash (Argonne National Laboratory)

Performance modeling is an important and active area of research in high-performance computing (HPC). It helps in better job scheduling and also improves overall performance of coupled applications. Sufficiently rich analytical models are challenging to develop, however, because of interactions between different node components, network topologies, job interference, and application complexity. When analytical performance models become restrictive because of application dynamics and/or multicomponent interactions, machine-learning-based performance models can be helpful. While machine learning (ML) methods do not require underlying system or application knowledge, they are efficient in learning the unknown interactions of the application and system parameters empirically using application runs. We present a benchmark study in which we evaluate eleven machine learning methods for modeling the performance of four representative scientific applications that are irregular and with skewed domain configurations on four leadership-class HPC platforms. We assess the impact of feature engineering, size of training set, modern hardware platforms, transfer learning, extrapolation on the prediction accuracy, and training and inference times. We find that bagging, boosting, and deep neural network ML methods are promising approaches with median R^2 values greater than 0.95, and these methods do not require feature engineering. We demonstrate that cross-platform performance prediction can be improved significantly using transfer learning with deep neural networks.

Algorithm Selection of MPI Collectives Using Machine Learning Techniques

Sascha Hunold (Technical University Wien)

Autotuning is a well established method to improve software performance for a given system, and it is especially important in High Performance Computing. The goal of autotuning is to find the best possible algorithm and its best parameter settings for a given instance. Autotuning can also be applied to MPI libraries, such as OpenMPI or IntelMPI. These MPI libraries provide numerous parameters that allow users to adapt them to a given system. Some of these tunable parameters enable users to select a specific algorithm that should be used internally by an MPI collective operation. For the purpose of automatically tuning MPI collectives on a given system, the Intel MPI library is shipped with mpitune. The drawback of tools like mpitune is that results can only be applied to cases (e.g., number of processes, message size) for which the tool has performed the optimization.

To overcome this limitation, we present a first step towards tuning MPI libraries also for unseen instances by applying machine learning techniques. Our goal is to create a classifier that takes the collective operation, the message size and communicator characteristics (number of compute nodes, number of processes per node) as an input and gives the predicted best algorithm for this problem as an output. We show how such a model can be constructed and what pitfalls should be avoided. We demonstrate by thorough experimentation that our proposed prediction model is able to outperform the default configuration

miniVite: A Graph Analytics Benchmarking Tool for Massively Parallel Systems

Mahantesh Halappanavar (Pacific Northwest National Laboratory)

Benchmarking of high performance computing systems can help provide critical insights for efficient design of computing systems and software applications. Although a large number of tools for benchmarking exist, there is a lack of representative benchmarks for the class of irregular computations as exemplified by graph analytics. We therefore propose miniVite as a representative graph analytics benchmark tool to test a variety of distributed-memory systems. Graph clustering, popularly known as community detection, is a prototypical graph operation used in numerous scientific computing and analytics applications. The goal of clustering is to partition a graph into clusters (or communities) such that each cluster consists of vertices that are densely connected within the cluster and sparsely connected to the rest of the graph. Modularity optimization is a popular technique for identifying clusters in a graph. Efficient parallelization of modularity optimization-based algorithms is challenging. One successful approach was conceived in Vite, a distributed-memory implementation of the Louvain method that incorporates several heuristics. We develop miniVite as a representative but simplified variant of Vite, to serve as a prototypical graph analytics benchmarking tool. Similar to many graph algorithms, miniVite is characterized by irregular communication patterns, high communication to computation ratios, and load imbalances among participating processes, thus making it a representative benchmarking tool.

Unlike some graph-based methods such as breadth-first search and betweenness centrality, miniVite represents highly complex computational patterns stressing a variety of system features. This can in turn help provide crucial insight for codesign of future computing systems. We believe that miniVite will serve as a platform for benchmarking systems and design communication primitives that will be applicable to a broad set of irregular computing applications as well as a platform for the design of efficient graph algorithms.

Unified Cross-Platform Profiling of Parallel C++ Applications
Vladyslav Kucher (University of Münster)

To address the great variety of available parallel hardware architectures (CPUs, GPUs, etc.), high-performance applications increasingly demand cross-platform portability. While unified programming models like OpenCL or SYCL provide the ultimate portability of code, the profiling of applications in the development process is still done by using different platform-specific tools of the corresponding hardware vendors. We design and implement a unified, cross-platform profiling interface by extending the PACXX framework for unified programming in C++. With our profiling interface, a single tool is used to profile parallel C++ applications across different target platforms. We illustrate and evaluate our uniform profiler using an example application of matrix multiplication for CPU and GPU architectures.

PMBS18 – Workshop Lunch (on your own)

A Metric for Evaluating Supercomputer Performance in the Era of Extreme Heterogeneity
Brian Austin (Lawrence Berkeley National Laboratory)

When acquiring a supercomputer, it is desirable to specify its performance using a single number. For many procurements, this is usually stated as a performance increase over a current generation platform, for example machine A provides 10 times greater performance than machine B. The determination of such a single number is not necessarily a simple process; there is no universal agreement on how this calculation is performed, and each facility usually uses their own method. In the future, the landscape will be further complicated because systems will contain a heterogeneous mix of node types, and, by design, every application will not run on every node type. For example, at the National Energy Research Scientific Computing Center (NERSC) the Cori supercomputer contains two node types, nodes based on dual-socket Intel Xeon (Haswell) processors and nodes based on Intel Xeon Phi (Knights Landing) processors. However, NERSC evaluated these two partitions separately, without utilizing a single, combined performance metric. NERSC will be deploying its next-generation machine, NERSC-9, in the year 2020 and anticipates that it too will be a heterogeneous mix of node types. The purpose of this paper is to describe a single performance metric for a heterogeneous system.

Evaluating SLURM Simulator with Real-Machine SLURM and Vice Versa
Marco D'Amico (Barcelona Supercomputing Center)

Having a precise and a fast job scheduler model that resembles the real-machine job scheduling software behavior is extremely important in the field of job scheduling. The idea behind SLURM simulator is preserving the original code of the core SLURM functions while allowing for all the advantages of a simulator. Since 2011, SLURM simulator has passed through several iterations of improvements in different research centers. In this work, we present our latest improvements of SLURM simulator and perform the first-ever validation of the simulator on the real machine. In particular, we improved the simulator's performance for about 2.6 times, made the simulator deterministic across several same set-up runs, and improved the simulator's accuracy. Its deviation from the real-machine is lowered from previous 12% to at most 1.7%. Finally, we illustrate with

several use cases the value of the simulator for job scheduling researchers, SLURM-system administrators, and SLURM developers.

PMBS18 – Workshop Afternoon Break

Is Data Placement Optimization Still Relevant on Newer GPUs?

Md Abdullah Shahneous Bari (Stony Brook University)

Modern supercomputers often use Graphic Processing Units (or GPUs) to meet the evergrowing demands for energy efficient high performance computing. GPUs have a complex memory architecture with various types of memories and caches, such as global memory, shared memory, constant memory, and texture memory. Data placement optimization, i.e. optimizing the placement of data among these different memories, has a significant impact on the performance of HPC applications running on early generations of GPUs. However, newer generations of GPUs have new memory features. They also implement the same high-level memory hierarchy differently.

In this paper, we design a set of experiments to explore the relevance of data placement optimizations on several generations of NVIDIA GPUs, including Kepler, Maxwell, Pascal, and Volta. Our experiments include a set of memory microbenchmarks, CUDA kernels and a proxy application. The experiments are configured to include different CUDA thread blocks, data input sizes, and data placement choices. The results show that newer generations of GPUs are less sensitive to data placement optimization compared to older ones, mostly due to improvements to caches of the global memory.

Approximating a Multi-Grid Solver

Leonardo Bautista-Gomez (Barcelona Supercomputing Center), Osman Unsal (Barcelona Supercomputing Center)

Multi-grid methods are numerical algorithms used in parallel and distributed processing. The main idea of multi-grid solvers is to speed up the convergence of an iterative method by reducing the problem to a coarser grid a number of times. Multi-grid methods are widely exploited in many application domains, thus it is important to improve their performance and energy efficiency. This paper aims to reach this objective based on the following observation: Given that the intermediary steps do not require full accuracy, it is possible to save time and energy by reducing precision during some steps while keeping the final result within the targeted accuracy.

To achieve this goal, we first introduce a cycle shape different from the classic V-cycle used in multi-grid solvers. Then, we propose to dynamically change the floating-point precision used during runtime according to the accuracy needed for each intermediary step. Our evaluation considering a state-of-the-art multi-grid solver implementation demonstrates that it is possible to trade temporary precision for time to completion without hurting the quality if the final result. In particular, we are able to reach the same accuracy results as with full double-precision while gaining between 15% and 30% execution time improvement.

Evaluating the Impact of Spiking Neural Network Traffic on Extreme-Scale Hybrid Systems

Noah Wolfe (Rensselaer Polytechnic Institute)

As we approach the limits of Moore's law, there is increasing interest in non-Von Neuman architectures such as neuromorphic computing to take advantage of improved compute and low power capabilities. Spiking neural network (SNN) applications have so far shown very promising results running on a number of processors, motivating the desire to scale to even larger systems having hundreds and even thousands of neuromorphic processors. Since these architectures currently do not exist in large configurations, we use simulation to scale real neuromorphic applications running on a single neuromorphic chip, to thousands of chips in an HPC class system. Furthermore, we use a novel simulation workflow to perform a full scale systems analysis of network performance and the interaction of neuromorphic workloads with traditional CPU workloads in a hybrid supercomputer environment. On average, we find Slim Fly, Fat-Tree, Dragonfly-1D, and Dragonfly-2D are 45%, 46%, 76%, and 83% respectively faster than the worst case performing topology for both convolutional and Hopfield NN workloads running alongside CPU workloads. Running in parallel with CPU workloads translates to an average slowdown of 21% for a Hopfield type workload and 184% for convolutional NN workloads across all HPC network topologies.

Automated Instruction Stream Throughput Prediction for Intel and AMD Microarchitectures
Jan Laukemann (University of Erlangen-Nuremberg)

An accurate prediction of scheduling and execution of instruction streams is a necessary prerequisite for predicting the in-core performance behavior of throughput-bound loop kernels on out-of-order processor architectures. Such predictions are an indispensable component of analytical performance models, such as the Roofline and the Execution-Cache-Memory (ECM) model, and allow a deep understanding of the performance-relevant interactions between hardware architecture and loop code.

We present the Open Source Architecture Code Analyzer (OSACA), a static analysis tool for predicting the execution time of sequential loops comprising x86 instructions under the assumption of an infinite first-level cache and perfect out-of-order scheduling. We show the process of building a machine model from available documentation and semi-automatic benchmarking, and carry it out for the latest Intel Skylake and AMD Zen micro-architectures.

To validate the constructed models, we apply them to several assembly kernels and compare runtime predictions with actual measurements. Finally we give an outlook on how the method may be generalized to new architectures.

Room: D161
9:00 am - 5:30 pm

9th Workshop on Latest Advances in Scalable Algorithms for Large-Scale Systems

Session Description: Novel scalable scientific algorithms are needed in order to enable key science applications to exploit the computational power of large-scale systems. This is especially true for the current tier of leading petascale machines and the road to exascale computing as HPC systems

continue to scale up in compute node and processor core count. These extreme-scale systems require novel scientific algorithms to hide network and memory latency, have very high computation/communication overlap, have minimal communication, and have no synchronization points. With the advent of Big Data in the past few years the need of such scalable mathematical methods and algorithms able to handle data and compute intensive applications at scale becomes even more important. Scientific algorithms for multi-petaflop and exaflop systems also need to be fault tolerant and fault resilient, since the probability of faults increases with scale. Resilience at the system software and at the algorithmic level is needed as a crosscutting effort. Finally, with the advent of heterogeneous compute nodes that employ standard processors as well as GPGPUs, scientific algorithms need to match these architectures to extract the most performance. This includes different system-specific levels of parallelism as well as co-scheduling of computation. Key science applications require novel mathematics and mathematical models and system software that address the scalability and resilience challenges of current- and future-generation extreme-scale HPC systems.

Novel scalable scientific algorithms are needed in order to enable key science applications to exploit the computational power of large-scale systems. This is especially true for the current tier of leading petascale machines and the road to exascale computing as HPC systems continue to scale up in compute node and processor core count. These extreme-scale systems require novel scientific algorithms to hide network and memory latency, have very high computation/communication overlap, have minimal communication, and have no synchronization points. With the advent of Big Data in the past few years the need of such scalable mathematical methods and algorithms able to handle data and compute intensive applications at scale becomes even more important.

Scientific algorithms for multi-petaflop and exaflop systems also need to be fault tolerant and fault resilient, since the probability of faults increases with scale. Resilience at the system software and at the algorithmic level is needed as a crosscutting effort. Finally, with the advent of heterogeneous compute nodes that employ standard processors as well as GPGPUs, scientific algorithms need to match these architectures to extract the most performance. This includes different system-specific levels of parallelism as well as co-scheduling of computation. Key science applications require novel mathematics and mathematical models and system software that address the scalability and resilience challenges of current- and future-generation extreme-scale HPC systems.

Introduction *Vassil Alexandrov (Barcelona Supercomputing Center)*

Keynote 1: A Few Scheduling Problems for Resilience at Scale *Yves Robert (ENS Lyon)*
The talk will address scheduling problems related to multi-level checkpointing, IO interference, silent errors detection and correction, and workflows.

Latest Advances – Workshop Morning Break

Keynote 2: HPC and AI as Drivers for Industrial Engagement *Alison Kennedy (Hartree Centre)*
The Hartree Centre is transforming UK industry through high performance computing, big data and

cognitive technologies. Backed by over £170 million of government funding and significant strategic partnerships with organisations such as IBM and Atos, the Hartree Centre is home to some of the most technically advanced high performance computing, data analytics, machine learning technologies and experts in the UK. From early stage SMEs to international corporations, we work with industry and the research community to address real life challenges and accelerate the adoption of high performance technologies, delivering transformative gains in performance, productivity and time to market. The Hartree Centre works to maintain the UK's position at the forefront of industrial innovation through a range of computational services and collaborative partnerships. This talk gives examples of the work being undertaken at the Hartree Centre, applying emerging technologies and scalable algorithms to industry challenge-led projects.

Event-Triggered Communication in Parallel Computing

Soumyadip Ghosh (University of Notre Dame)

Communication overhead in parallel systems can be a significant bottleneck in scaling up parallel computation. In this paper, we propose event-triggered communication methods to reduce such communication overhead for numerical simulation of partial differential equations. As opposed to traditional communication in which processing elements exchange data at every iteration of the numerical algorithm, the main idea behind event-triggered communication is to exchange data only when necessary as dictated by a suitably defined event. We show through numerical experiments that these methods have potential to reduce simulation time. Strong scaling plots show that the approach may be effective on large machines as well.

Communication Reduced Multi-Timestep Algorithm for Real-Time Wind Simulation on GPU-Based Supercomputers

Naoyuki Onodera (Japan Atomic Energy Agency)

We develop a communication reduced multi-time-step (CRMT) algorithm for a Lattice Boltzmann method (LBM) based on a block-structured adaptive mesh refinement (AMR). This algorithm is based on the temporal blocking method, and can improve computational efficiency by replacing a communication bottleneck with additional computation. The proposed method is implemented on an extreme scale airflow simulation code CityLBM, and its impact on the scalability is tested on GPU based supercomputers, TSUBAME and Reedbush. Thanks to the CRMT algorithm, the communication cost is reduced by ~ 64%, and weak and strong scalings are improved up to ~ 200 GPUs. The obtained performance indicates that real time airflow simulations for about 2km square area with the wind speed of ~ 5m/s is feasible using 1m resolution.

Communication Avoiding Multigrid Preconditioned Conjugate Gradient Method for Extreme Scale Multiphase CFD Simulations

Yasuhiro Idomura (Japan Atomic Energy Agency)

A communication avoiding (CA) multigrid preconditioned conjugate gradient method (CAMGCG) is applied to the pressure Poisson equation in a multiphase CFD code JUPITER, and its computational performance and convergence property are compared against CA Krylov methods. A new geometric multigrid preconditioner is developed using a preconditioned Chebyshev iteration smoother, in which

no global reduction communication is needed, halo data communication is reduced by a mixed precision approach, and eigenvalues are computed using the CA Lanczos method. The CAMGCG solver has robust convergence properties regardless of the problem size, and shows both communication reduction and convergence improvement, leading to higher performance gain than CA Krylov solvers, which achieve only the former. The CAMGCG solver is applied to extreme scale multiphase CFD simulations with ~90 billion DOFs, and it is shown that compared with a preconditioned CG solver, the number of iterations, and thus, All_Reduce is reduced to ~1/800, and ~11.6x speedup is achieved with keeping excellent strong scaling up to 8,000 KNLs on the Oakforest-PACS.

Non-Collective Scalable Global Network Based on Local Communications
Marco Berghoff (Karlsruhe Institute of Technology)

To efficiently perform collective communications in current high-performance computing systems is a time-consuming task. With future exascale systems, this communication time will be increased further. However, global information is frequently required in various physical models. By exploiting domain knowledge of the model behaviors globally needed information can be distributed more efficiently, using only peer-to-peer communication which spread the information to all processes asynchronous during multiple communication steps. In this article, we introduce a multi-hop based Manhattan Street Network (MSN) for global information exchange and show the conditions under which a local neighbor exchange is sufficient for exchanging distributed information. Besides the MSN, in various models, global information is only needed in a spatially limited region inside the simulation domain. Therefore, a second network is introduced, the local exchange network, to exploit this spatial assumption.

Both non-collective global exchange networks are implemented in the massively parallel NAStJA framework. Based on two models, a phase-field model for droplet simulations and the cellular Potts model for biological tissue simulations, we exemplary demonstrate the wide applicability of these networks. Scaling tests of the networks demonstrate a nearly ideal scaling behavior with an efficiency of over 90%. Theoretical prediction of the communication time on future exascale systems shows an enormous advantage of the presented exchange methods of O(1) by exploiting the domain knowledge.

Latest Advances – Workshop Lunch (on your own)

Iterative Randomized Algorithms for Low Rank Approximation of Terascale Matrices with Small Spectral Gaps
Chander J. Iyer (Rensselaer Polytechnic Institute, Yahoo! Research)

Randomized approaches for low rank matrix approximations have become popular in recent years and often offer significant advantages over classical algorithms because of their scalability and numerical robustness on distributed memory platforms. We present a distributed implementation of randomized block iterative methods to compute low rank matrix approximations for dense tera-scale matrices. We are particularly interested in the behavior of randomized block iterative methods on matrices with small spectral gaps. Our distributed implementation is based on four iterative algorithms: block subspace iteration, the block Lanczos method, the block Lanczos method with explicit restarts, and the

thick-restarted block Lanczos method. We analyze the scalability and numerical stability of the four block iterative methods and demonstrate the performance of these methods for various choices of the spectral gap. Performance studies demonstrate superior runtimes of the block Lanczos algorithms over the subspace power iteration approach on (up to) 16,384 cores of AMOS, Rensselaer's IBM Blue Gene/Q supercomputer.

Shift-Collapse Acceleration of Generalized Polarizable Reactive Molecular Dynamics for Machine Learning-Assisted Computational Synthesis of Layered Materials
Kuang Liu (University of Southern California), Subodh Tiwari (University of Southern California)

Reactive molecular dynamics is a powerful simulation method for describing chemical reactions. Here, we introduce a new generalized polarizable reactive force-field (ReaxPQ+) model to significantly improve the accuracy by accommodating the reorganization of surrounding media. The increased computation is accelerated by (1) extended Lagrangian approach to eliminate the speed-limiting charge iteration, (2) shiftcollapse computation of many-body renormalized n-tuples, which provably minimizes data transfer, (3) multithreading with roundrobin data privatization, and (4) data reordering to reduce computation and allow vectorization. The new code achieves (1) weak-scaling parallel efficiency of 0.989 for 131,072 cores, and (2) eight-fold reduction of time-to-solution (T2S) compared with the original code, on an Intel Knights Landing-based computer. The reduced T2S has for the first time allowed purely computational synthesis of atomically-thin transition metal dichalcogenide layers assisted by machine learning to discover a novel synthetic pathway.

Machine Learning-Aided Numerical Linear Algebra: Convolutional Neural Networks for the Efficient Preconditioner Generation

Markus Götz received his Bachelors and Masters degree in Software Engineering from the University of Potsdam in 2010 and 2014 respectively. Afterwards, he has been with the Research Center Jülich and the University of Iceland, from which Markus obtained his PhD degree in Computational Engineering for his works on parallel data-analysis algorithms on high-performance computing (HPC) systems. Since the beginning of 2018 Markus is with the Steinbuch Centre for Computing (SCC) at the Karlsruhe Institute of Technology (KIT). There, he manages the Helmholtz Analytics Framework project, a german-wide initiative with the aim of developing the data sciences in the Helmholtz Association. His research topics include applied machine learning, scalable data analysis frameworks and parallel algorithms.

Latest Advances – Workshop Afternoon Break

Keynote 3: Hierarchical Algorithms on Hierarchical Architectures *David Keyes (King Abdullah University of Science and Technology)*
Some algorithms achieve optimal arithmetic complexity with low arithmetic intensity (flops/Byte), or possess high arithmetic intensity but lack optimal complexity, while some hierarchical algorithms, such as Fast Multipole and its H-matrix algebraic generalizations, realize a combination of optimal complexity and high intensity. Implemented with task-based dynamic runtime systems, such methods also have potential for relaxed synchrony, which is important for future energy-austere architectures,

since there may be significant nonuniformity in processing rates of different cores even if task sizes can be controlled. We describe modules of KAUST's Hierarchical Computations on Manycore Architectures (HiCMA) software toolkit that illustrate these features and are intended as building blocks of more sophisticated applications, such as matrix-free higher-order methods in optimization. HiCMA's target is hierarchical algorithms on emerging architectures, which have hierarchies of their own that generally do not align well with those of the algorithm. Some modules of this open source project have been adopted in the software libraries of major vendors. We describe what is currently available and some motivating applications.

Low Thread-Count Gustavson: A Multithreaded Algorithm for Sparse Matrix-Matrix Multiplication Using Perfect Hashing

James Elliott (Sandia National Laboratories)

Sparse matrix-matrix multiplication is a critical kernel for several scientific computing applications, especially the setup phase of algebraic multigrid. The MPI+X programming model, which is growing in popularity, requires that such kernels be implemented in a way that exploits on-node parallelism. We present a single-pass OpenMP variant of Gustavson's sparse matrix-matrix multiplication algorithm designed for architectures (e.g. CPU or Intel Xeon Phi) with reasonably large memory and modest thread counts (tens of threads, not thousands). These assumptions allow us to exploit perfect hashing and dynamic memory allocation achieve performance improvements of up to 2x over third-party kernels for matrices derived from algebraic multigrid setup.

A General-Purpose Hierarchical Mesh Partitioning Method with Node Balancing Strategies for Large-Scale Numerical Simulations

Fande Kong (Idaho National Laboratory)

Large-scale parallel numerical simulations are essential for a wide range of engineering problems that involve complex, coupled physical processes interacting across a broad range of spatial and temporal scales. The data structures involved in such simulations (meshes, sparse matrices, etc.) are frequently represented as graphs, and these graphs must be optimally partitioned across the available computational resources in order for the underlying calculations to scale efficiently. Partitions which minimize the number of graph edges that are cut (edge-cuts) while simultaneously maintaining a balance in the amount of work (i.e. graph nodes) assigned to each processor core are desirable, and the performance of most existing partitioning software begins to degrade in this metric for partitions with more than than $O(10^3)$ processor cores. In this work, we consider a general-purpose hierarchical partitioner which takes into account the existence of multiple processor cores and shared memory in a compute node while partitioning a graph into an arbitrary number of subgraphs. We demonstrate that our algorithms significantly improve the preconditioning efficiency and overall performance of realistic numerical simulations running on up to 32,768 processor cores with nearly 10^9 unknowns.

Dynamic Load Balancing of Plasma and Flow Simulations

Gerrett Diamond (Rensselaer Polytechnic Institute)

Extracting performance from simulations with complex information dependencies on massively parallel

computers requires the computational work to be evenly distributed across the processing resources while maintaining low communication costs. Plasma simulations using a particle-in-cell method and computational fluid dynamics using unstructured mesh-based finite element and volume methods present three distinct distribution requirements. To meet these needs, we present EnGPar's diffusive partition improvement method. An initial demonstration of EnGPar's particle distribution improvement is provided along with fluid dynamics mesh partition improvement results on up to 512Ki processes on an IBM BlueGene/Q.

On Advanced Monte Carlo Methods for Linear Algebra on Advanced Accelerator Architectures
Vassil Alexandrov (ICREA, Barcelona Supercomputing Center)

In this paper we present computational experiments performed using the Markov Chain Monte Carlo Matrix Inversion (MCMCMI) on several architectures of NVIDIA accelerators and two iterations of the Intel x86 architecture and investigate their impact on performance and scalability of the method. The method is used as a preconditioner and iterative methods, such as generalized minimal residuals (GMRES) or bi-conjugate gradient stabilized (BICGstab), are used for solving the corresponding system of linear equations. Numerical experiments are carried out to highlight the benefits and deficiencies of both architecture types and to assess their overall usefulness in light of the scalability of the method.

Room: D163
9:00 am - 5:30 pm

PDSW-DISCS: Joint International Workshop on Parallel Data Storage and Data Intensive Scalable Computing Systems

We are pleased to announce that the third Joint International Workshop on Parallel Data Storage and Data Intensive Scalable Computing Systems (PDSW-DISCS'18) will be hosted at SC18: The International Conference for High Performance Computing, Networking, Storage, and Analysis. The objective of this one day joint workshop is to combine two overlapping communities, HPC storage and data intensive computing, and to better promote and stimulate researchers' interactions to address some of the most critical challenges for scientific data storage, management, devices, and processing infrastructure for both traditional compute intensive simulations and data-intensive high performance computing solutions. Special attention will be given to issues in which community collaboration can be crucial for problem identification, experiment reproducibility, workload capture, solution interoperability, standards with community buy-in, and shared tools.

Welcome and Introduction

Keynote Address *Rangan Sukumar (Cray Inc)*

Integration of Burst Buffer in High-Level Parallel I/O Library for Exascale Computing Era

While the computing power of supercomputers continues to improve at an astonishing rate, companion I/O systems are struggling to keep up in performance. To mitigate the performance gap, several supercomputing systems have been configured to incorporate burst buffers into their I/O stack; the exact role of which, however, still remains unclear. In this paper, we examine the features of burst buffers and study their impact on application I/O performance. Our goal is to demonstrate that burst buffers can be utilized by parallel I/O libraries to significantly improve performance. To this end, we developed an I/O driver in PnetCDF that uses a log-based format to store individual I/O requests on the burst buffer – later to be flushed to the parallel file system as one request. We evaluated our implementation by running standard I/O benchmarks on Cori, a Cray XC40 supercomputer at NERSC with a centralized burst buffer system, and Theta, a Cray XC40 supercomputer at ALCF with locally available SSDs. Our results show that IO aggregation is a promising role for burst buffers in high-level I/O libraries.

Using a Robust Metadata Management System to Accelerate Scientific Discovery at Extreme Scales

Our previous work, which can be referred to as EMPRESS 1.0, showed that rich metadata management provides a relatively low-overhead approach to facilitating insight from scale-up scientific applications. However, this system did not provide the functionality needed for a viable production system or address whether such a system could scale. Therefore, we have extended our previous work to create EMPRESS 2.0, which incorporates the features required for a useful production system.

Through a discussion of EMPRESS 2.0, this paper explores how to incorporate rich query functionality, fault tolerance, and atomic operations into a scalable, storage system independent metadata management system that is easy to use. This paper demonstrates that such a system offers significant performance advantages over HDF5, providing metadata querying that is 150X to 650X faster, and can greatly accelerate post-processing. Finally, since the current implementation of EMPRESS 2.0 relies on an RDBMS, this paper demonstrates that an RDBMS is a viable technology for managing data-oriented metadata.

Evaluation of HPC Application I/O on Object Storage Systems
Jialin Liu (Lawrence Berkeley National Laboratory)

POSIX-based parallel file systems provide strong consistency semantics, which many modern HPC applications do not need and do not want. Object store technologies avoid POSIX consistency and are designed to be extremely scalable, for use in cloud computing and similar commercial environments. In this work, we evaluate three object store systems: Intel DAOS, Ceph RADOS, and Openstack Swift, and evaluate them with three HPC applications: VPIC, H5Boss, and BDCATS. We have developed

virtual object layer (VOL) plugins for HDF5 that can redirect the applications' HDF5 calls to the underlying object storage systems' APIs, with minimum application code change. Through our evaluation, we found that object stores have better scalability in many cases than POSIX file systems, but are not optimized for common HPC use cases, such as collective I/O. Understanding current object store I/O details and limitations will enable us to better design object stores for future HPC systems.

WIP Session 1

PDSW-DISCS - Workshop Lunch (on your own)

Pufferbench: Evaluating and Optimizing Malleability of Distributed Storage
Nathanael Cheriere (IRISA, ENS Rennes)

Malleability is the property of an application to be dynamically rescaled at run time. It requires the possibility to dynamically add or remove resources to the infrastructure without interruption. Yet, many Big Data applications cannot benefit from their inherent malleability, since their colocated distributed storage system is not malleable in practice. Commissioning or decommissioning storage nodes is generally assumed to be slow, as such operations have typically been designed for maintenance only. New technologies, however, enable faster data transfers. Still, evaluating the performance of rescaling operations on a given platform is a challenge in itself: no tool currently exists for this purpose.

We introduce Pufferbench, a benchmark for evaluating how fast one can scale up and down a distributed storage system on a given infrastructure and, thereby, how viably can one implement storage malleability on it. Besides, it can serve to quickly prototype and evaluate mechanisms for malleability in existing distributed storage systems. We validate Pufferbench against theoretical lower bounds for commission and decommission: it can achieve performance within 16% of them. We use Pufferbench to evaluate in practice these operations in HDFS: commission in HDFS could be accelerated by as much as 14 times! Our results show that: (1) the lower bounds for commission and decommission times we previously established are sound and can be approached in practice; (2) HDFS could handle these operations much more efficiently; most importantly, (3) malleability in distributed storage systems is viable and should be further leveraged for Big Data applications.

Understanding SSD Reliability in Large-Scale Cloud Systems

Modern datacenters increasingly use flash-based solid state drives (SSDs) for high performance and low energy cost. However, SSDs introduce more complex failure modes compared to traditional hard disks. While great efforts have been made to understand the reliability of SSDs itself, it remains unclear how the device-level errors may affect upper layers, or how the services running on top of the storage stack may affect the SSDs.

In this paper, we take a holistic view to examine the reliability of SSD-based storage systems in Alibaba's datacenters, which covers about half-million SSDs under representative cloud services over

three years. By vertically analyzing the error events across three layers (i.e., SSDs, OS, and the distributed file system), we discover a number of interesting correlations. For example, SSDs with UltraDMA CRC errors, while seems benign at the device level, are nearly 3 times more likely to lead to OS-level error events. As another example, different cloud services may lead to different usage patterns of SSDs, some of which are detrimental from the devices perspective.

WIP Session 2

PDSW-DISCS - Workshop Afternoon Break

Characterizing Deep-Learning I/O Workloads in TensorFlow

The performance of Deep-Learning (DL) computing frameworks rely on the performance of data ingestion and checkpointing. In fact, during the training, a considerable high number of relatively small files are first loaded and pre-processed on CPUs and then moved to accelerator for computation. In addition, checkpointing and restart operations are carried out to allow DL computing frameworks to restart quickly from a checkpoint. Because of this, I/O affects the performance of DL applications. In this work, we characterize the I/O performance and scaling of TensorFlow, an open-source programming framework developed by Google and specifically designed for solving DL problems. To measure TensorFlow I/O performance, we first design a micro-benchmark to measure TensorFlow reads, and then use a TensorFlow mini-application based on AlexNet to measure the performance cost of I/O and checkpointing in TensorFlow. To improve the checkpointing performance, we design and implement a burst buffer. We find that increasing the number of threads increases TensorFlow bandwidth by a maximum of 2.3× and 7.8× on our benchmark environments. The use of the tensorFlow prefetcher results in a complete overlap of computation on accelerator and input pipeline on CPU eliminating the effective cost of I/O on the overall performance. The use of a burst buffer to checkpoint to a fast small capacity storage and copy asynchronously the checkpoints to a slower large capacity storage resulted in a performance improvement of 2.6× with respect to checkpointing directly to slower storage on our benchmark environment.

Toward Understanding I/O Behavior in HPC Workflows

Jakob Luettgau (German Climate Computing Center, Argonne National Laboratory)

Scientific discovery increasingly depends on complex workflows consisting of multiple phases and sometimes millions of parallelizable tasks or pipelines. These workflows access storage resources for a variety of purposes, including preprocessing, simulation output, and postprocessing steps. Unfortunately, most workflow models focus on the scheduling and allocation of computational resources for tasks while the impact on storage systems remains a secondary objective and an open research question. I/O performance is not usually accounted for in workflow telemetry reported to users.

In this paper, we present an approach to augment the I/O efficiency of the individual tasks of workflows by combining workflow description frameworks with system I/O telemetry data. A conceptual architecture and a prototype implementation for HPC data center deployments are

introduced. We also identify and discuss challenges that will need to be addressed by workflow management and monitoring systems for HPC in the future. We demonstrate how real-world applications and workflows could benefit from the approach, and we show how the approach helps communicate performance-tuning guidance to users.

Methodology for the Rapid Development of Scalable HPC Data Services
Matthieu Dorier (Argonne National Laboratory)

Growing evidence in the scientific computing community indicates that parallel file systems are not sufficient for all HPC storage workloads. This realization has motivated extensive research in new storage system designs. The question of which design we should turn to implies that there could be a single answer satisfying a wide range of diverse applications. We argue that such a generic solution does not exist. Instead, custom data services should be designed and tailored to the needs of specific applications on specific hardware. Furthermore, custom data services should be designed in close collaboration with users. In this paper, we present a methodology for the rapid development of such data services. This methodology promotes the design of reusable building blocks that can be composed together efficiently through a runtime based on high-performance threading, tasking, and remote procedure calls. We illustrate the success of our methodology by showcasing three examples of data services designed from the same building blocks, yet targeting entirely different applications.

WIP Session 3

Room: D170
9:00 am - 5:30 pm

The 2nd Industry/University Joint International Workshop on Data Center Automation, Analytics, and Control (DAAC)

Session Description: The DAAC workshop series is focused on fostering discussion among industry, academic, and national laboratory participants and promoting collaboration on solutions for automated data centers and associated data center analytics and control issues. The objective is to promote and stimulate community's interactions to address some of most critical challenges in automation, analytics, and control specifically aimed for the needs of large-scale data centers in high-performance and other forms of highly scaled computing. This year's workshop features an exciting mix of ideas from industry and university contributors with experience in diverse settings for large-scale data center deployments.

Our plan is for an Industry/University Joint International Workshop on Data-Center Automation, Analytics, and Control (DAAC) to be hosted at SC18. This workshop plan is an outcome of intensive discussions from academia, industry, and national laboratory researchers that led to a successful

previous-year instance hosted at the 10th IEEE/ACM International Conference on Utility and Cloud Computing (UCC'17). Looking at the last year's attendance at UCC'17, DAAC has attracted around 50 attendees, even though UCC is a smaller conference compared to SC18. DAAC'17 featured three invited speakers from industry and a panel of five experts from national labs, academia, and industry to discuss data center automation, analytics, and control solutions, in addition to presentations from peer-reviewed papers.

This new instance of the DAAC workshop at SC is also jointly organized by academic and industry researchers. The objective is to promote and stimulate community interactions to address some of most critical challenges in automation, analytics, and control specifically aimed for the needs of large-scale data centers in high-performance/high-end computing. DAAC'18 will be a unique workshop at SC18 providing a valuable addition to main conference programs. Taking advantage of the opportune match to the SC18 audience, DAAC'18 expects to attract a larger number of attendees from academia, industry, and government national labs who are interested in data center automated management, operation, and maintenance.

Welcome, Workshop Goals, and Opening Remarks *Alan Sill (Texas Tech University), Yong Chen (Texas Tech University)*

Contention-Aware Container Placement Strategy for Docker Swarm

Containerization technology utilizes operating system level virtualization to package applications so they can run with required libraries and are isolated from other processes on the same host. Lightweight and quick deployment make containers popular in many data centers. Running distributed applications in data centers usually involves multiple clusters of machines. Docker Swarm is a container orchestration tool for managing a cluster of Docker containers and their hosts. However, Docker Swarm's scheduler does not consider resource utilization when placing containers in a cluster. This paper first investigated performance interference in container clusters. Our experimental study showed that distributed applications' performance can be degraded by about 15% when co-located with other containers which aggressively consume resources. We then proposed a new scheduler to improve performance while keeping high resource utilization. The experimental results demonstrated that the proposed prototype can effectively improve distributed applications' performance by up to 3.13%.

Workload Time Series Prediction in Storage Systems: A Deep Learning Based Approach

TACC's Cloud Deployer: Automating the Management of Distributed Software Systems

The Cloud and Interactive Computing (CIC) Group at the Texas Advanced Computing Center develops,

deploys and administers a growing catalog of cloud-based, distributed software systems for national-scale cyberinfrastructure projects. Additionally, these systems are being deployed off-site at prominent partner institutions such as the University of Hawai'i Manoa and the Centers for Disease Control (CDC). To automate operations, both internally and at partner sites, the CIC group has developed the TACC Cloud Deployer, a command line tool and software system combining Docker, Ansible and custom Python code. Initially released in March of 2016, CIC recently completed design and initial development for Deployer "V2" based on a number of lessons learned over the prior two years. In this paper, we describe the evolution of TACC's Cloud Deployer, including the architecture and major design decisions that went into its most recent incarnation.

DAAC – Workshop Morning Break

HPCViz: Monitoring Health Status of High Performance Computing Systems

This paper introduces HPCViz, a visual analytic tool for tracking and monitoring system events through a RESTful interface. The goals of this tool are: 1) to monitor a set of system events from multiple hosts and racks in real time statistics, 2) to support system administrators in alarming and detecting unusual signature-based patterns exhibited by health records of hosts in a complex system, and 3) to help in performing system debugging with a visual layout for both computing resource allocations and health monitoring map that mimics the actual system. A case study was conducted in a Redfish environment with a sample of 10 racks and 467 hosts. The result of the case study shows that the visualization tool offers excellent support for system analysists to profile and observe system behavior and further identify the traces of issues occurred.

Tivan: A Scalable Data Collection and Analytics Cluster

Log analysis is a critical part of every data center to diagnose system issues and determine performance problems. As data centers increase in size, the need for a dedicated, scalable log analysis system becomes necessary. At Los Alamos National Laboratory, our clusters and support infrastructure are currently producing over seventy million logs a day. The large amount of log data makes it extremely difficult for system administrators, application developers, and researchers to find relevant events and perform detailed analysis. In order to manage this data and to run scalable analytics, we designed and deployed a data analytics cluster based on open source software. In this paper, we describe the design and trade-offs, the current uses of this system, and analytics utilizing the system that are aiding in the discovery of system issues. We have found that this data analytics cluster has helped us to proactively identify cluster issues more efficiently.

DiG: Enabling Out-of-Band Scalable High-Resolution Monitoring for Data-Center Analytics, Automation, and Control

Data centers are increasing in size and complexity, and we need scalable approaches to support their

automated analysis and control. Performance, power consumption and reliability are their key "vital signs". State-of-the-Art monitoring systems provide built-in tools to collect performance measurements, and custom solutions to get insight on their power consumption. However, with the increase in measurement resolution (in time and space) and the ensuing huge amount of measurement data to handle, new challenges arise, such as bottlenecks on the network bandwidth, storage and software overhead on the monitoring units. To face these challenges we propose a novel monitoring platform for data centers, which enables real-time high-resolution profiling (i.e., all available performance counters and the entire signal bandwidth of the power consumption at the plug - sampling up to 20us) and analytics, both on the edge (node-level analysis) and on a centralized unit (cluster-level analysis). The monitoring infrastructure is completely out-of-band, scalable, technology agnostic and low cost.

Out-of-Band (BMC based) Data Center Monitoring DMTF Redfish API Integration with Nagios

Nagios is an industry standard for HPC infrastructure monitoring including hosts and associated hardware components, networks, storages, services, and applications. However, there are significant issues with traditional Nagios including 1) Nagios requires human intervention for the definition and maintenance of remote hosts configurations in Nagios Core. 2) It requires Nagios Remote Plugin Executor (NRPE) on Nagios Server and each monitored remote host. 3) It also mandates Nagios Service Check Acceptor (NSCA) on each monitored remote host. 4) And also requires Check specific agents (e.g. SNMP) on each monitored remote host. In order to address these issues, we have integrated Distributed Management Task Force (DMTF)'s Redfish API with Nagios core. DMTF's Redfish API is an open industry standard specification and schema designed to meet the expectations of end users for simple, modern and secure management of scalable platform hardware. Redfish API is essentially out-of-band protocol which is implemented in baseboard management controller (BMC) of the system. Redfish API supports network-based auto-discovery which is quite instrumental in automating configuration of remote hosts. Nagios plugins will be directly communicating to BMC so it eliminates the requirement of any agent and configuration on remote hosts. Redfish API integration with Nagios is potentially a huge paradigm shift in Nagios based monitoring in terms of: 1) simplifying communication between Nagios server and monitored hosts; and 2) eliminating computational cost and complexity of running Nagios native protocols (NRPE and NSCA) and other agents (SNMP) on the monitored hosts.

DAAC – Workshop Lunch (on your own)

Invited Talk: Resource Control at Facebook *Tejun Heo (Facebook)*
Facebook has been actively experimenting with cgroup2 resource control for years. In the process, we developed several kernel and userland mechanisms, fixed numerous isolation issues, and discovered a number of surprising interactions. We finally have work-conserving full-OS resource control working and are now in the process of refining and deploying for various applications including workload protection.

Let's take a look at the mistakes, the lessons, and the result.

Invited Talk: The Campus Compute Cooperative Project as an Alternative to Commercial Clouds

Andrew Grimshaw (University of Virginia)
Wide-area, federated, compute-sharing systems (such as Condor, gLite, Globus, and Legion) have been around for over twenty years. Outside of particular domains such as physics, these systems have not been widely adopted. Recently, however, universities are starting to propose and join resource-sharing platforms. Why this sudden change?

Mostly, this change has come in response to cost concerns. HPC managers are under new pressure from university administrators who demand that infrastructure outlays be economically justified. "Why not just put it all on Amazon?" goes the administration's refrain. In response, HPC managers have begun to document the true cost of university-, department-, and research-group-owned infrastructure, thus enabling a legitimate cost comparison with Amazon or Azure. Additionally, it may be noted, this pressure to consider outsourcing computing infrastructure has legitimized both remote computing and paying for computation.

In this talk, I will briefly describe the Campus Compute Cooperative's (CCC). I will then detail both the results of our market simulations and the take-aways from interviews with stakeholders. By both of these measures, the CCC is valuable and viable: first, the simulation results clearly show the gains in institutional value; second, stakeholders indicated that many institutions are open to trading resources. Most promising, some institutions expressed interest in selling resources and others expressed willingness to pay.

Invited Talk: Current Status of the OpenHPC Project *Karl Schulz (University of Texas; Institute for Computational Engineering and Sciences, Dell Medical School)*
Over the last several years, OpenHPC has emerged as a community-driven stack providing a variety of common, pre-built ingredients to deploy and manage an HPC Linux cluster including provisioning tools, resource management, I/O clients, runtimes, development tools, and a variety of scientific libraries. Formed initially in November 2015 and formalized as a Linux Foundation project in June 2016, OpenHPC has been adding new software components and now supports multiple OSes and architectures. This presentation will present an overview of the project, currently available software, the general build and test architecture employed to deliver a flexible development environment, and highlight general project updates and future plans.

DAAC – Workshop Afternoon Break

Industry Panel: Data-Center Automation, Analytics, and Control from an Industry Perspective *Jon Hass (Dell Inc), Kevin Hughes (Cray Inc), Samer El Haj Mahmoud (Lenovo), Jeff Autor (Hewlett Packard Enterprise)*
Data center automation, analytics, and control have a long history in industry with a wide variety of approaches having been pursued. Recently, efforts have emerged to unify approaches to these topics with standards-based efforts. This panel will examine the challenges, history, and present status of software, protocols, and API development including reasons that the Redfish standard and related protocol efforts have emerged to play a strong role in creating the basis for common approaches to these challenges. The participants in this panel are well versed in these topics and will bring their collective expertise together to present each company's perspective on the issues.

Simulating Data Centers with Redfish-Enabled Equipment

Analytic data, such as temperature and power consumption of equipment, are becoming growingly critical in high-end enterprise and also in high-performance computing platforms, especially for current petascale and future exascale systems on the horizon, consisting of tens of thousands or even hundreds of thousands of servers. Also, these data are important for accurate and real-time system monitoring. Examining new ideas for data gathering, data analysis and systems management by running different tests against production data center could negatively affect the production environment and also could have unintentional consequences. For these reasons, a simulated data center can enable researchers, developers, and administrators to perform thorough tests before carrying out on a real data center or even to eliminate the need for physical examination. This research aims to discuss a new methodology and a framework we have developed via using container technology to simulate a scalable data center with Redfish-enabled equipment to achieve the objectives.

Energy and Power Aware Job Scheduling and Resource Management: Global Survey --- An In-Depth Analysis

This paper presents a detailed analysis of a first-of-kind global survey of high-performance computing centers that actively employ techniques within job scheduling and resource management middleware layers for managing energy and power on production supercomputers.

Our group conducted a series of comprehensive interviews of leading-edge supercomputing centers during 2016 and 2017. The group presented the motivation of the survey center selection, questionnaire details, and a preliminary analysis of the survey results in a previous publication. This paper presents a more detailed analysis of the survey results. The goal is to find commonalities, approaches that are to be developed, and hints and guidelines for other centers to move toward a more energy efficient and power aware management of their compute resources.

Dynamic and Portable Vulnerability Assessment Testbed with Linux Containers to Ensure the Security of MongoDB in Singularity LXCs

To find the available vulnerabilities against any system, it is mandatory to conduct vulnerability assessments as scheduled tasks in a regular manner. Thus, an easily deployable, easily maintainable, accurate vulnerability assessment testbed or a model is helpful as facilitated by Linux containers. Nowadays Linux containers (LXCs) which have operating system level virtualization, are very popular over virtual machines (VMs) which have hypervisor or kernel level virtualization in high performance computing (HPC) due to reasons, such as high portability, high performance, efficiency and high security. Hence, LXCs can make an efficient and scalable vulnerability assessment testbed or a model by using already developed analyzing tools such as OpenVas, Dagda, PortSpider, OWASP Zed Attack Proxy, and OpenSCAP, to assure the required security level of a given system very easily. To verify the overall security of any given software system, this paper first introduces a virtual, portable and easily deployable vulnerability assessment general testbed within the Linux container network. Next, the paper presents, how to conduct experiments using this testbed on a MongoDB database implemented in Singularity Linux containers to find the available vulnerabilities in images accompanied by

containers, host, and network by integrating three tools; OpenVas, Dagda, and PortSpider to the container-based testbed. Finally, it discusses how to use generated results to improve the security level of the given system.

Closing Remarks

Room: D172
9:00 am - 5:30 pm

IA^3 2018: 8th Workshop on Irregular Applications: Architectures and Algorithms

Due to the heterogeneous data sets they process, data intensive applications employ a diverse set of methods and data structures, exhibiting irregular memory accesses, control flows, and communication patterns. Current supercomputing systems are organized around components optimized for data locality and bulk synchronous computations. Managing any form of irregularity on them demands substantial programming effort, and often leads to poor performance. Holistic solutions to these challenges emerge only by considering the problem from multiple perspectives: from micro- to system-architectures, from compilers to languages, from libraries to runtimes, and from algorithm design to data characteristics. Only strong collaborative efforts among researchers with different expertise, including domain experts and end users, can lead to significant breakthroughs. This workshop brings together scientists with different backgrounds to discuss methods and technologies for efficiently supporting irregular applications on current and future architectures.

Photonic Interconnects for Extreme Scale Computing *Madeleine Glick (Columbia University)*
The capabilities of large-scale high performance computing systems, either as supercomputers or warehouse scale data centers, are increasingly pervasive in different areas of modern life, from weather predictions to film and fashion recommendations. New applications using data intensive computations are putting more stress on the interconnection network at the same time that Moore's Law is slowing. advances in transistor density. As a consequence, the high bandwidth interconnects, essential for maintaining computation performance, are representing an increasing portion of the total energy and cost budgets.

It is widely accepted that new approaches are required to meet the new challenges. Photonic interconnection networks, with recent advances in integrated silicon photonics offer an opportunity to overcome limitations of conventional electrical wires to solve bottlenecks and improve interconnect performance. Beyond alleviating the bandwidth and energy bottlenecks, embedded photonics can enable new disaggregated architectures that leverage the distance independence of optical transmission. In this talk, I will present an overview of recent photonic advances and discuss how we can use emerging photonics technologies to increase resource utilization and reduce energy consumption.

IA^3 2018 – Workshop Morning Break

Software Prefetching for Unstructured Mesh Applications

Ioan Hadade (Oxford Thermofluids Institute, University of Oxford)

Applications that exhibit regular memory access patterns usually benefit transparently from hardware prefetchers that bring data into the fast on-chip cache just before it is required, thereby avoiding expensive cache misses. In contrast, unstructured mesh applications contain irregular access patterns that are often more difficult to identify in hardware. An alternative for such workloads is software prefetching, where special non-blocking instructions load data into the cache hierarchy. However, there are currently few examples in the literature on how to incorporate such software prefetches into existing applications with positive results.

This paper addresses these issues by demonstrating the utility and implementation of software prefetching in an unstructured finite volume CFD code of representative size and complexity to an industrial application and across a number of processors. We present the benefits of auto-tuning for finding the optimal prefetch distance values across different computational kernels and architectures and demonstrate the importance of choosing the right prefetch destination across the available cache levels for best performance. We discuss the impact of the data layout on the number of prefetch instructions required in kernels with indirect-access patterns and show how to integrate them on top of existing optimizations such as vectorization. Through this we show significant full application speed-ups on a range of processors, such as the Intel Xeon Skylake CPU (15%) as well as on the in-order Intel Xeon Phi Knights Corner (1.99X) architecture and the out-of-order Knights Landing (33%) many-core processor.

A Fast and Simple Approach to Merge and Merge Sorting Using Wide Vector Instructions

Merging and sorting algorithms are the backbone of many modern computer applications. As such, efficient implementations are desired. Recent architectural advancements in CPUs (Central Processing Units), such as wider and more powerful vector instructions, allow for algorithmic improvements. This paper presents a new approach to merge sort using vector instructions. Traditional approaches to vectorized sorting typically utilize a bitonic sorting network (Batcher's Algorithm) which adds significant overhead. Our approach eliminates the overhead from this approach. We start with a branch-avoiding merge algorithm and then use the Merge Path algorithm to split up merging between the different SIMD lanes. Testing demonstrates that the algorithm not only surpasses the SIMD based bitonic counterpart, but that it is over 2.94X faster than a traditional merge, merging over 300M keys per second in one thread and over 16B keys per second in parallel. Our new sort reaches is over 5X faster than quicksort and 2X faster than Intel's IPP library sort, sorting over 5.3M keys per second for a single processor and in parallel over 500M keys per second and a speedup of over 2X from a traditional merge sort.

Impact of Traditional Sparse Optimizations on a Migratory Thread Architecture

Thomas B. Rolinger (University of Maryland, Laboratory for Physical Sciences at University of Maryland)

Achieving high performance for sparse applications is challenging due to irregular access patterns and weak locality. These properties preclude many static optimizations and degrade cache performance on traditional systems. To address these challenges, novel systems such as the Emu architecture have been proposed. The Emu design uses light-weight migratory threads, narrow memory, and near-memory processing capabilities to address weak locality and reduce the total load on the memory system. Because the Emu architecture is fundamentally different than cache based hierarchical memory systems, it is crucial to understand the cost-benefit tradeoffs of standard sparse algorithm optimizations on Emu hardware. In this work, we explore sparse matrix-vector multiplication (SpMV) on the Emu architecture. We investigate the effects of different sparse optimizations such as dense vector data layouts, work distributions, and matrix reorderings. Our study finds that initially distributing work evenly across the system is inadequate to maintain load balancing over time due to the migratory nature of Emu threads. In severe cases, matrix sparsity patterns produce hot-spots as many migratory threads converge on a single resource. We demonstrate that known matrix reordering techniques can improve SpMV performance on the Emu architecture by as much as 70% by encouraging more consistent load balancing. This can be compared with a performance gain of no more than 16% on a cache-memory based system.

There Are Trillions of Little Forks in the Road: Choose Wisely! -- Estimating the Cost and Likelihood of Success of Constrained Walks to Optimize a Graph Pruning Pipeline

Nicolas Tripoul (University of British Columbia), Tahsin Reza (University of British Columbia), Geoffrey Sanders (Lawrence Livermore National Laboratory), Roger Pearce (Lawrence Livermore National Laboratory), Matei Ripeanu (University of British Columbia)

We have developed [Reza et al. SC18] a highly scalable algorithmic pipeline for pattern matching in labeled graphs and demonstrated it on trillion-edge graphs. This pipeline: (i) supports arbitrary search patterns, (ii) identifies all the vertices and edges that participate in matches - offering 100% precision and recall, and (iii) supports realistic data analytics scenarios. This pipeline is based on graph pruning: it decomposes the search template into individual constraints and uses them to repeatedly prune the graph to a final solution.

Our current solution, however, makes a number of ad-hoc intuition-based decisions with impact on performance. In a nutshell these relate to (i) constraint selection - which constraints to generate? (ii) constraint ordering - in which order to use them? and (iii) individual constraint generation - how to best verify them? This position paper makes the observation that by estimating the runtime cost and likelihood of success of a constrained walk in a labeled graph one can inform these optimization decisions. We propose a preliminary solution to make these estimates, and demonstrate - using a prototype shared-memory implementation - that this: (i) is feasible with low overheads, and (ii) offers accurate enough information to optimize our pruning pipeline by a significant margin.

Scale-Free Graph Processing on a NUMA Machine

Tanuj K. Aasawat (University of British Columbia), Tahsin Reza (University of British Columbia), Matei Ripeanu (University of British Columbia)

Modern shared-memory systems embrace the NUMA architecture which has proven to be more scalable than the SMP architecture. In many ways, a NUMA system resembles a shared-nothing distributed system: physically distinct processing units and memory regions. Memory accesses to

remote NUMA domains are more expensive than local accesses. This poses the opportunity to transfer the know-how and design of distributed graph processing to develop shared-memory graph processing solutions optimized for NUMA systems. To this end, we explore if a distributed-memory like middleware that makes graph partitioning and communication between partitions explicit, can improve the performance on a NUMA system. We design and implement a NUMA aware graph processing framework that embraces design philosophies of distributed graph processing system: in particular explicit partitioning and inter-partition communication, and at the same time exploits optimization opportunities specific to single-node systems. We demonstrate up to 13.9x speedup over a state-of-the-art NUMA-aware framework, Polymer and up to 3.7x scalability on a four-socket NUMA machine using graphs with tens of billions of edges.

IA^3 2018 – Workshop Lunch (on your own)

Versal: The New Xilinx Adaptive Compute Acceleration Platforms (ACAP) *Kees Vissers (Xilinx Inc)*
In this presentation, I will present the new Adaptive Compute Acceleration Platform. I will show the overall system architecture of the family of devices including the Arm cores (scalar engines), the programmable logic (Adaptable Engines) and the new vector processor cores (AI engines). I will focus on the new AI engines in more detail and show the concepts for the programming environment, the architecture, the integration in the total device, and some application domains, including Machine Learning and 5G wireless applications. I will illustrate the initial design rationale and the architecture trade-offs. These platforms extend the concept of tuning the memory hierarchy to the problem.

IA^3 2018 – Workshop Afternoon Break

Mix-and-Match: A Model-Driven Runtime Optimization Strategy for BFS on GPUs
Merijn Elwin Verstraaten (University of Amsterdam)

It is universally accepted that the performance of graph algorithms is heavily dependent on the algorithm, the execution platform, and the structure of the input graph. This variability remains difficult to predict and hinders the choice of the right algorithm for a given problem.

In this work, we focus on a case study: breadth-first search (BFS), a level-based graph traversal algorithm, running on GPUs. We first demonstrate the severity of this variability by presenting 32 variations of 5 implementation strategies for GPU-enabled BFS, and showing how selecting one single algorithm for the entire traversal can significantly limit performance. To alleviate these performance losses, we propose to mix-and-match, at runtime, different algorithms to compose the best performing BFS traversal. Our approach is based on two novel elements: a predictive model, based on a decision tree, which is able to dynamically select the best performing algorithm for each BFS level, and a quick context switch between algorithms, which limits the overhead of the combined BFS.

We demonstrate empirically that our dynamic switching BFS achieves better performance, outperforming our non-switching implementations by 2x and existing state-of-the-art GPU BFS implementations by 3x. We conclude that mix-and-match BFS is a competitive approach for doing fast graph traversal, while being easily extended to include more BFS implementations and easily adaptable to other types of processors or specific types of graphs.

A Block-Oriented, Parallel, and Collective Approach to Sparse Indefinite Preconditioning on GPUs
Daniel Thuerck (Technical University Darmstadt)

Large sparse symmetric indefinite matrices are notoriously hard to precondition. They often lack diagonal dominance and exhibit Schur-complements that render zero fill-in factorization preconditioning ineffective. Pivoting, a necessity for stable LDLt factorizations, complicates parallel approaches that can take advantage of the latest massively-parallel HPC hardware such as GPUs. We present an approach based on ad-hoc blocking and reordering strategies that allows local, independent collective-oriented processing of small dense blocks. A hybrid block-memory layout compensates for irregular memory access patterns found in sparse matrices. Our method allows restricted fill-in, supernodal pivoting and a dual threshold dropping strategy at little additional cost. It delivers robust preconditioners that in our experiments obtain an average speedup of ~6x even for tough matrices from optimization problems.

High-Performance GPU Implementation of PageRank with Reduced Precision Based on Mantissa Segmentation
Hartwig Anzt (Karlsruhe Institute of Technology)

We address the acceleration of the PageRank algorithm for web information retrieval on graphics processing units (GPUs) via a modular precision framework that adapts the input data format in memory to the numerical requirements as the iteration converges. In detail, we abandon the ieee 754 single- and double-precision number representation formats, employed in the standard implementation of PageRank, to instead store the data in memory in some specialized formats. Furthermore, we avoid the data duplication by leveraging a data layout based on mantissa segmentation. Our evaluation on a V100 graphics card from NVIDIA shows acceleration factors of up to 30% with respect to the standard algorithm operating in double-precision.

IA^3 Debate *John Feo (Pacific Northwest National Laboratory)*
The IA^3 oxfordian debate!

Room: D166
9:00 am - 5:30 pm

ESPM2 2018: Fourth International Workshop on Extreme Scale Programming Models and Middleware

Next generation architectures and systems being deployed are characterized by high concurrency, low memory per-core, and multiple levels of hierarchy and heterogeneity. These characteristics bring out new challenges in energy efficiency, fault-tolerance, and scalability. It is commonly believed that software has the biggest share of the responsibility to tackle these challenges. In other words, this responsibility is delegated to the next generation programming models and their associated

middleware/runtimes.

This workshop focuses on different aspects of programming models such as task-based parallelism (Charm++, OCR, X10, HPX, etc), PGAS (OpenSHMEM, UPC, CAF, Chapel, UPC++, etc.), BigData (Hadoop, Spark, etc), Deep Learning (Caffe, Microsoft CNTK, Google TensorFlow), directive-based languages (OpenMP, OpenACC) and hybrid MPI+X, etc. It also focuses on their associated middleware (unified runtimes, interoperability for hybrid programming, tight integration of MPI+X, and support for accelerators and FPGAs) for next generation systems and architectures. The ultimate objective of the ESPM2 workshop is to serve as a forum that brings together researchers from academia and industry working in the areas of programming models, runtime systems, compilation and languages, and application developers.

Exascale Challenges in Across-Node Parallelism for Languages and Runtimes *Laxmikant Kale (University of Illinois)*
Machines with peak performance exceeding one exaflop/s are just around the corner, and promises of sustained exaflop/s machines abound. Are there significant challenges in runtime frameworks and languages that need to be met to harness the power of these machines? We will examine this question and associated issues.

There are some architectural trends that are becoming clear and some hazily appearing. Individual nodes are getting "fatter" computationally. Accelerators such as GPGPUs and possibly FPGAs are likely parts of the exascale landscape. High bandwidth memory, and non-coherent caches (such as the caches in GPGPUs used typically for constant memory), NVRAMS, and resultant deeper and more complex memory hierarchies will also have to be dealt with.

There is an argument going around in the community, that we have already figured out how to deal with tens of thousands of nodes (100,000 with BG/Q), and now since the number of nodes is not likely to increase, we (the extreme-scale HPC community) have to focus research almost entirely on within-node issues. I believe this is not quite a well-founded argument. I will explain why issues of power/energy/temperature, whole machine (multi-job) optimizations, across node issues like communication optimization, load balancing and fault tolerance are still worthy of significant attention of the exascale runtime and language community. At the same time, there exist issues in handling within-node parallelism that arise mainly or only in the context large multi-node runs.

I will also address the question of how our community should approach research if a large segment of funding sources and application community have started thinking some of the above issues are irrelevant. What should the courage of our convictions lead us to?

ESPM2 – Workshop Morning Break

Distributed Memory Futures for Compile-Time, Deterministic-by-Default Concurrency in Distributed C++ Applications
Jeremiah Wilke (Sandia National Laboratories)

Futures are a widely-used abstraction for enabling deferred execution in imperative programs. Deferred execution enqueues tasks rather than explicitly blocking and waiting for them to execute.

Many task-based programming models with some form of deferred execution rely on explicit parallelism that is the responsibility of the programmer. Deterministic-by-default (implicitly parallel) models instead use data effects to derive concurrency automatically, alleviating the burden of concurrency management. Both implicitly and explicitly parallel models are particularly challenging for imperative object-oriented programming. Fine-granularity parallelism across member functions or amongst data members may exist, but is often ignored. In this work, we define a general permissions model that leverages the C++ type system and move semantics to define an asynchronous programming model embedded in the C++ type system. Although a default distributed memory semantic is provided, the concurrent semantics are entirely configurable through C++ constexpr integers. Correct use of the defined semantic is verified at compile-time, allowing deterministic- by-default concurrency to be safely added to applications. Here we demonstrate the use of these "extended futures" for distributed memory asynchronous communication and load balancing. An MPI particle-in-cell application is modified with the wrapper class using this task model, with results presented for a Haswell system up to 64 nodes.

Design of Data Management for Multi-SPMD Workflow Programming Model

Miwako Tsuji (Riken Center for Computational Science)

As both the complexity of algorithms and architecture increase, development of scientific software becomes a challenge. In order to exploit future architecture, we consider a Multi-SPMD workflow programing model. Then, data transfer between tasks during computation highly depends on the architecture and middleware used. In this study, we design an adaptive system for data management in a parallel programming environment which can express two level of parallelism. We show how the consideration of multiple strategies based on I/O and direct message passing can improve performances and fault tolerance in the YML-XMP environment. On a real application with a sufficiently large amount of local data, speedup of 1.36 for a mixed strategy to 1.73 for a direct message passing method are obtained compared to our original design.

Integration of CUDA Processing within the C++ Library for Parallelism and Concurrency (HPX)

Patrick Diehl (Louisiana State University)

Experience shows that on today's high performance systems, the utilization of different acceleration cards in conjunction with a high utilization of all other parts of the system is difficult. Future architectures, like exascale clusters, are expected to aggravate this issue as the number of cores are expected to increase and memory hierarchies are expected to become deeper. One big aspect for distributed applications is to guarantee high utilization of all available resources, including local or remote acceleration cards on a cluster while fully using all the available CPU resources and the integration of the GPU work into the overall programming model.

For the integration of CUDA code we extended HPX and enabled asynchronous data transfers from and to the GPU device and the asynchronous invocation of CUDA kernels on this data. Both operations are well integrated into the general programming model of HPX which allows to seamlessly overlap any GPU operation with work on the main cores. Any user-defined CUDA kernel can be launched.

We present asynchronous implementations for the data transfers and kernel launches for CUDA code

as part of a HPX asynchronous execution graph. Using this approach we can combine all remotely and locally available acceleration cards on a cluster to utilize its full performance capabilities. Overhead measurements show, that the integration of the asynchronous operations as part of the HPX execution graph imposes no additional computational overhead and significantly eases orchestrating coordinated and concurrent work on the main cores and the used GPU devices.

Automatic Generation of High-Order Finite-Difference Code with Temporal Blocking for Extreme-Scale Many-Core Systems

Hideyuki Tanaka (ExaScaler Inc), Jun Makino (Kobe University)

In this paper we describe the basic idea, implementation and achieved performance of our DSL for stencil computation, Formura, on systems based on PEZY-SC2 many-core processor. Formura generates, from high-level description of the differential equation and simple description of finite-difference stencil, the entire simulation code with MPI parallelization with overlapped communication and calculation, advanced temporal blocking and parallelization for many-core processors. Achieved performance is 4.78 PF, or 21.5% of the theoretical peak performance for an explicit scheme for compressive CFD, with the accuracy of fourth-order in space and third-order in time. For a slightly modified implementation of the same scheme, efficiency was slightly lower (17.5%) but actual calculation time per one timestep was faster by 25%. Temporal blocking improved the performance by up to 70%. Even though the B/F number of PEZY-SC2 is low, around 0.02, we have achieved the efficiency comparable to those of highly optimized CFD codes on machines with much higher memory bandwidth such as K computer. We have demonstrated that automatic generation of the code with temporal blocking is a quite effective way to make use of very large-scale machines with low memory bandwidth for large-scale CFD calculations.

ESPM2 – Workshop Lunch (on your own)

Asynchronous Execution of Python Code on Task Based Runtime Systems

Mohammad Tohid (Louisiana State University)

Despite advancements in the areas of parallel and distributed computing, the complexity of programming on High Performance Computing (HPC) resources has deterred many domain experts, especially in the areas of machine learning and artificial intelligence (AI), from utilizing performance benefits of such systems. Researchers and scientists favor high-productivity languages to avoid the inconvenience of programming in low-level languages and costs of acquiring the necessary skills required for programming at this level. In recent years, Python, with the support of linear algebra libraries like NumPy, has gained popularity despite facing limitations which prevent this code from distributed runs. Here we present a solution which maintains both high level programming extractions as well as parallel and distributed efficiency. Phylanx, is an asynchronous array processing toolkit which transforms Python and NumPy operations into code which can be executed in parallel on HPC resources by mapping Python and NumPy functions and variables into a dependency tree executed by HPX, a general purpose, parallel, task-based runtime system written in C++. Phylanx additionally provides introspection and visualization capabilities for debugging and performance analysis. We have tested foundations of our approach by comparing our implementation of widely used machine learning algorithms to accepted NumPy standards.

A Unified Runtime for PGAS and Event-Driven Programming
Sri Raj Paul (Rice University)

A well-recognized characteristic of extreme scale systems is that their computation bandwidths far exceed their communication bandwidths. PGAS runtimes have proven to be effective in enabling efficient use of communication bandwidth, due to their efficient support for short nonblocking one-sided messages. However, they were not designed for exploiting the massive levels of intra-node parallelism found in extreme scale systems.

In this paper, we explore the premise that event-driven intra-node runtimes could be promising candidates for integration with PGAS runtimes, due to their ability to overlap computation with long-latency operations. For this, we use OpenSHMEM as an exemplar of PGAS runtimes, and Node.js as an exemplar of event-driven runtimes. While Node.js may seem an unusual choice for high-performance computing, its prominent role as an event-driven runtime for server-side Javascript is a good match for what we need for optimizing the use of communication bandwidth. Availability of excess computation bandwidth in extreme scale systems can help mitigate the local computation overheads of Javascript. Further, since the Node.js environment is single-threaded, we get an automatic guarantee that no data races will occur on Javascript objects, a guarantee that we cannot get with C++.

Our integration of OpenSHMEM and Node.js makes it possible to expose nonblocking PGAS operations as JavaScript constructs that can seamlessly be used with the JavaScript asynchronous mechanisms, such as those based on promises. We believe that the exploration and preliminary results in this paper offer a new direction for future research on building high productivity runtimes for extreme scale systems.

ESPM2 – Workshop Afternoon Break
Workshop Afternoon Break

3:00 PM to 3:30 PM

Food provided by SC

Portable and Reusable Deep Learning Infrastructure with Containers to Accelerate Cancer Studies
George Zaki (Frederick National Laboratory for Cancer Research)

Advanced programming models, domain specific languages, and scripting toolkits have the potential to greatly accelerate the adoption of high performance computing. These complex software systems, however, are often difficult to install and maintain, especially on exotic high-end systems. We consider deep learning workflows used on petascale systems and redeployment on research clusters using containers. Containers are used to deploy the MPI-based infrastructure, but challenges in efficiency, usability, and complexity must be overcome. In this work, we address these challenges through enhancements to a unified workflow system that manages interaction with the container abstraction, the cluster scheduler, and the programming tools. We also report results from running the application on our system, harnessing 298~TFLOPS (single precision).

Analysis of Explicit vs. Implicit Tasking in OpenMP Using Kripke
Charles C. Jin (Reservoir Labs Inc)

Dynamic task-based parallelism has become a widely-accepted paradigm in the quest for exascale computing. In this work, we deliver a non-trivial demonstration of the advantages of explicit over implicit tasking in OpenMP 4.5 in terms of both expressiveness and performance. We target the Kripke benchmark, a mini-application used to test the performance of discrete particle codes, and find that the dependence structure of the core "sweep" kernel is well-suited for dynamic task-based systems. Our results show that explicit tasking delivers a 31.7% and 8.1% speedup over a pure implicit implementation for a small and large problem, respectively, while a hybrid variant also underperforms the explicit variant by 13.1% and 5.8%, respectively.

Heterogeneous Systems and the Road to Exascale for HPC and AI *Daniel Holmes (University of Edinburgh)*

ESPM2 2018: Closing Remarks *Dhabaleswar Panda (Ohio State University), Hari Subramoni (Ohio State University), Karl Schulz (University of Texas)*

Room: D167/174
9:00 am - 5:30 pm

Machine Learning in HPC Environments

The intent of this workshop is to bring together researchers, practitioners, and scientific communities to discuss methods that utilize extreme scale systems for machine learning. This workshop will focus on the greatest challenges in utilizing HPC for machine learning and methods for exploiting data parallelism, model parallelism, ensembles, and parameter search. We invite researchers and practitioners to participate in this workshop to discuss the challenges in using HPC for machine learning and to share the wide range of applications that would benefit from HPC powered machine learning.

Workshop Overview

Morning Keynote – Azalia Mirhoseini (Google) *Azalia Mirhoseini (Google LLC)*
Advances in computer systems have been key to the success of Machine Learning (ML) in recent years. With the ubiquitous success of ML, it is now time for a new era where we can transform the way computer systems are built -- with learning. This talk highlights some of the challenges that modern computer systems are facing, and how we can use machine learning to address them. Specifically, it will cover various combinatorial optimization problems that appear in computational graph optimizations and then delve into some of our recent efforts at Google in addressing these problems

with deep Reinforcement Learning (RL). Our results show that we can use RL-based techniques to optimize such problems without the need to characterize details of the target hardware or the computational graph. Instead, RL finds a solution by only incorporating the reward function of interest such as runtime or memory. Using the reward function, RL learns the implicit trade-offs in the underlying hardware and can achieve results that outperform traditional optimization techniques that rely on heuristics.

Machine Learning – Workshop Morning Break

Communication-Efficient Parallelization Strategy for Deep Convolutional Neural Network Training
Sunwoo Lee (Northwestern University)

Training modern Convolutional Neural Network (CNN) models is extremely time-consuming, and the efficiency of its parallelization plays a key role in finishing the training in a reasonable amount of time. The well-known parallel synchronous Stochastic Gradient Descent (SGD) algorithm suffers from high costs of inter-process communication and synchronization. To address such problems, the asynchronous SGD algorithm employs a master-slave model for parameter update. However, it can result in a poor convergence rate due to the staleness of gradient. In addition, the master-slave model is not scalable when running on a large number of compute nodes. In this paper, we present a communication-efficient gradient averaging algorithm for synchronous SGD, which adopts a few design strategies to maximize the degree of overlap between computation and communication. The time complexity analysis shows our algorithm outperforms the traditional algorithms that use MPI allreduce-based communication. By training the two popular deep CNN models, VGG-16 and ResNet-50, on ImageNet dataset, our experiments performed on Cori Phase-I, a Cray XC40 supercomputer at NERSC show that our algorithm can achieve up to 2516.36x speedup for VGG-16 and 2734.25x speedup for ResNet-50 when running on up to 8192 cores.

Large-Scale Clustering Using MPI-Based Canopy
Thomas Heinis (Imperial College, London)

Analyzing massive amounts of data and extracting value from it has become key across different disciplines. Many approaches have been developed to extract insight from the plethora of data available. As the amount of data grow rapidly, however, current approaches for analysis struggle to scale. This is particularly true for clustering algorithms which try to find patterns in the data.

A wide range of clustering approaches has been developed in recent years. What they all share is that they require parameters (number of clusters, size of clusters etc.) to be set a priori. Typically these parameters are determined through trial and error in several iterations or through pre-clustering algorithms. Several pre-clustering algorithms have been developed, but similarly to clustering algorithms, they do not scale well for the rapidly growing amounts of data.

In this paper, we thus take one such pre-clustering algorithm, Canopy, and develop a parallel version based on MPI. As we show, doing so is not straightforward and without optimization, a considerable amount of time is spent waiting for synchronization, severely limiting scalability. We thus optimize our approach to spend as little time as possible with idle cores and synchronization barriers. As our

experiments show, our approach scales near linear with increasing dataset size.

Automated Labeling of Electron Microscopy Images Using Deep Learning

Gunther Weber (Lawrence Berkeley National Laboratory; University of California, Davis)

Searching for scientific data requires metadata providing a relevant context. Today, generating metadata is a time and labor intensive manual process that is often neglected, and important datasets are not accessible through search. We investigate the use of machine learning to generalize metadata from a subset of labeled data, thus increasing the availability of meaningful metadata for search. Specifically, we consider electron microscopy images collected at the National Center for Electron Microscopy at the Lawrence Berkeley National Laboratory and use of deep learning to discern characteristics from a small subset of labeled images and transfer labels to the entire image corpus.

Relatively small training set sizes and a minimum resolution of 512x512 pixels required by the application domain pose unique challenges. We overcome these challenges by using a simple yet powerful convolutional network architecture that limits the number of free parameters to lower the required amount of computational power and reduce the risk of overfitting. We achieve a classification accuracy of approximately 80% in discerning between images recorded in two operating modes of the electron microscope---transmission electron microscopy (TEM) and scanning transmission electron microscopy (STEM). We use transfer learning–i.e., re-using the pre-trained convolution layers from the TEM vs. STEM classification problem–to generalize labels and achieve an accuracy of approximately 70% despite current experiments being limited to small training data sets. We present these predictions as suggestions to domain scientists to accelerate the labeling process with the goal of further validating our approach and improving the accuracy of automatically created labels.

Scaling Deep Learning for Cancer with Advanced Workflow Storage Integration

Justin Wozniak (Argonne National Laboratory)

Cancer Deep Learning Environment (CANDLE) benchmarks and workflows will combine the power of exascale computing with neural network-based machine learning to address a range of loosely connected problems in cancer research. This application area poses unique challenges to the exascale computing environment. Here, we identify one challenge in CANDLE workflows, namely, saving neural network model representations to persistent storage. In this paper, we provide background on this problem, describe our solution, the Model Cache, and present performance results from running the system on a test cluster, ANL/LCRC Blues, and the petascale supercomputer NERSC Cori. We also sketch next steps for this promising workflow storage solution.

Machine Learning – Workshop Lunch (on your own)

On Adam-Trained Models and a Parallel Method to Improve the Generalization Performance

Adam is a popular stochastic optimizer that uses adaptive estimates of lower-order moments to update weights and requires little hyper-parameter tuning. Some recent studies have called the

generalization and out-of-sample behavior of such adaptive gradient methods into question, and argued that such methods are of only marginal value. Notably for many of the well-known image classification tasks such as CIFAR-10 and ImageNet-1K, current models with best validation performance are still trained with SGD with a manual schedule of learning rate reduction.

We analyze Adam and SGD trained models for 7 popular neural network architectures for image classification tasks using the CIFAR-10 dataset. Visualization shows that for classification Adam trained models frequently "focus" on areas of the images not occupied by the objects to be classified. Weight statistics reveal that Adam trained models have larger weights and L2 norms than SGD trained ones. Our experiments show that weight decay and reducing the initial learning rates improves generalization performance of Adam, but there still remains a gap between Adam and SGD trained models.

To bridge the generalization gap, we adopt a K-step model averaging parallel algorithm with the Adam optimizer. With very sparse communication, the algorithm achieves high parallel efficiency. For the 7 models on average the improvement in validation accuracy over SGD is 0.72%, and the average parallel speedup is 2.5 times with 6 GPUs.

Aluminum: An Asynchronous, GPU-Aware Communication Library Optimized for Large-Scale Training of Deep Neural Networks on HPC Systems
Nikoli Dryden (University of Illinois, Lawrence Livermore National Laboratory)

We identify communication as a major bottleneck for training deep neural networks on large-scale GPU clusters, taking over 10x as long as computation. To reduce this overhead, we discuss techniques to overlap communication and computation as much as possible. This leads to much of the communication being latency-bound instead of bandwidth-bound, and we find that using a combination of latency- and bandwidth-optimized allreduce algorithms significantly reduces communication costs. We also discuss a semantic mismatch between MPI and CUDA that increases overheads and limits asynchrony, and propose a solution that enables communication to be aware of CUDA streams. We implement these optimizations in the open-source Aluminum communication library, enabling optimized, asynchronous, GPU-aware communication. Aluminum demonstrates improved performance in benchmarks and end-to-end training of deep networks, for both strong and weak scaling.

Machine Learning – Workshop Afternoon Break

Workshop Overview

Afternoon Keynote - Robinson Pino (DOE ASCR) *Robinson Pino (US Department of Energy Office of Advanced Scientific Computing Research)*

Optimizing Machine Learning on Apache Spark in HPC Environments
Zhenyu Li (University of Warwick)

Machine learning has established itself as a powerful tool for the construction of decision making models and algorithms through the use of statistical techniques on training data. However, a significant impediment to its progress is the time spent training and improving the accuracy of these models. A common approach to accelerate this process is to employ the use of multiple machines simultaneously, a trait shared with the field of High Performance Computing (HPC) and its clusters. However, existing distributed frameworks for data analytics and machine learning are designed for commodity servers, which do not realize the full potential of a HPC cluster.

In this work, we adapt the application of Apache Spark, a distributed data-flow framework, to support the use of machine learning in HPC environments for the purposes of machine learning. There are inherent challenges to using Spark in this context; memory management, communication costs and synchronization overheads all pose challenges to its efficiency. To this end we introduce: (i) the application of MapRDD, a fine grained distributed data representation; (ii) a task-based all-reduce implementation; and (iii) a new asynchronous Stochastic Gradient Descent (SGD) algorithm using non-blocking all-reduce. We demonstrate up to a 2.6x overall speedup (or a 11.2x theoretical speedup with a Nvidia K80 graphics card), when training the GoogLeNet model to classify 10% of the ImageNet dataset on a 32-node cluster. We also demonstrate a comparable convergence rate using the new asynchronous SGD with respect to the synchronous method.

Large Minibatch Training on Supercomputers with Improved Accuracy and Reduced Time to Train
Valeriu Codreanu (SURFsara)

For the past 6 years, the ILSVRC competition and the ImageNet dataset have attracted a lot of interest from the Computer Vision community, allowing for state-of-the-art accuracy to grow tremendously. This should be credited to the use of deep artificial neural network designs. As these became more complex, the storage, bandwidth, and compute requirements increased. This means that with a non-distributed approach, even when using the most high-density server available, the training process may take weeks, making it prohibitive. Furthermore, as datasets grow, the representation learning potential of deep networks grows as well by using more complex models. This synchronicity triggers a sharp increase in the computational requirements and motivates us to explore the scaling behaviour on petaflop scale supercomputers. In this paper we describe the challenges and novel solutions needed in order to train ResNet-50 in a large scale environment. We demonstrate above 90 percent scaling efficiency and a training time of 28 minutes using up to 104K x86 cores. This is supported by software tools from Intel's ecosystem. Moreover, we show that with regular 90 - 120 epoch train runs we can achieve a top-1 accuracy as high as 77 percent for the unmodified ResNet-50 topology. We also introduce the novel Collapsed Ensemble technique that allows us to obtain a 77.5 percent top-1 accuracy, similar to that of a ResNet-152, while training a unmodified ResNet-50 topology for the same fixed training budget.

Room: D168
9:00 am - 5:30 pm

ISAV 2018: In Situ Infrastructures for Enabling Extreme-Scale Analysis and Visualization

The considerable interest in the HPC community regarding in situ analysis and visualization is due to several factors. First is an I/O cost savings, where data is analyzed/visualized while being generated, without first storing to a filesystem. Second is the potential for increased accuracy, where fine temporal sampling of transient analysis might expose some complex behavior missed in coarse temporal sampling. Third is the ability to use all available resources, CPU's and accelerators, in the computation of analysis products.

The workshop brings together researchers, developers and practitioners from industry, academia, and government laboratories developing, applying, and deploying in situ methods in extreme-scale, high performance computing. The goal is to present research findings, lessons learned, and insights related to developing and applying in situ methods and infrastructure across a range of science and engineering applications in HPC environments; to discuss topics like opportunities presented by new architectures, existing infrastructure needs, requirements, and gaps, and experiences to foster and enable in situ analysis and visualization; to serve as a "center of gravity" for researchers, practitioners, and users/consumers of in situ methods and infrastructure in the HPC space.

Keynote: Perspectives on In Situ *Laura Biven (US Department of Energy Office of Advanced Scientific Computing Research)*

ISAV 2018 – Workshop Morning Break

A Flexible System For In Situ Triggers
Matthew Larsen (Lawrence Livermore National Laboratory)

Triggers are an important mechanism for adapting I/O and visualization actions as a simulation runs. We describe the system for triggers in the Ascent in situ infrastructure. This system splits a trigger into two components, when to perform an action and what actions to perform. The decision for when to perform an action can be based on different types of factors, such as mesh topology, scalar fields, or performance data. The actions to perform are also varied, ranging from the traditional action of saving simulation state to disk to performing arbitrary visualizations and analyses. We also include details on the implementation and a short example demonstrating how it can be used.

PaDaWAn: a Python Infrastructure for Loosely Coupled In Situ Workflows
Marc Perache (Atomic Energy and Alternative Energies Commission (CEA))

This paper presents PaDaWAn, an infrastructure written in Python to provide loosely coupled in situ capabilities to accelerate file-based simulation workflows. It provides services for in-memory data exchange between applications and a simple configuration model to switch from a file-based workflow to a loosely coupled in situ workflow. The infrastructure is currently based on CEA-DAM Hercule parallel I/O library by providing an ABI-compatible library to intercept simulation data in a transparent way and to facilitate integration into existing simulation codes and tools. PaDaWAn implements a producer-consumer pattern with buffering of data in an in-memory staging service with

automatic memory management and running on dedicated resources. We describe the key design decisions and main architectural features, and share the lessons learned from the development of the infrastructure and from test runs on two production-like workflow cases. We conclude on the perspectives for our infrastructure.

In Situ Data-Driven Adaptive Sampling for Large-scale Simulation Data Summarization
Ayan Biswas (Los Alamos National Laboratory), James Ahrens (Los Alamos National Laboratory)

Recent advancements in the high-performance computing have enabled the scientists to model various scientific phenomena in great detail. However, the analysis and visualization of the output data from such large-scale simulations are posing significant challenges due to the excessive size of output data and disk I/O bottleneck. One viable solution to this problem is to create a sub-sampled dataset which is able to preserve the important information of the data and also is significantly smaller in size compared to the raw data. Creating an in situ workflow for generating such intelligently sub-sampled datasets is of prime importance for such simulations. In this work, we propose an information-driven data sampling technique and compare it with two well-known sampling methods to demonstrate the superiority of the proposed method. The in situ performance of the proposed method is evaluated by applying the sampling techniques to the Nyx Cosmology simulation. We compare and contrast the performances of these various sampling algorithms and provide a holistic view of all the methods so that the scientists can choose appropriate sampling schemes based on their analysis requirements.

Python-based In Situ Analysis and Visualization
Burlen Loring (Lawrence Berkeley National Laboratory)

This work focuses on enabling the use of Python-based methods for the purpose of performing in situ analysis and visualization. This approach facilitates access to and use of a rapidly growing collection of Python-based, third-party libraries for analysis and visualization, as well as lowering the barrier to entry for user-written Python analysis codes. Beginning with a simulation code that is instrumented to use the SENSEI in situ interface, we present how to couple it with a Python-based data consumer, which may be run in situ, and in parallel at the same concurrency as the simulation. We present two examples that demonstrate the new capability. One is an analysis of the reaction rate in a proxy simulation of a chemical reaction on a 2D substrate, while the other is a coupling of an AMR simulation to Yt, a parallel visualization and analysis library written in Python. In the examples, both the simulation and Python in situ method run in parallel on a large-scale HPC platform.

Leveraging Scalable Event Distribution to Enable Data-driven In-situ Scientific Workflows
Zhe Wang (Rutgers University), Anthony Simonet (Rutgers University)

Novel event-driven workflow systems have been effectively used to increase the performance of large-scale scientific applications by removing most of the implicit synchronization required to orchestrate distributed tasks. However, these event-driven workflow systems, by focusing only on events related to the completion of tasks and data transfers, fail to address the dynamic and irregular workflows that require fine adaptation of the execution to the environment, faults, and to partial results from the application itself. In this article, we explore the idea of a programming model for irregular and dynamic workflows that is not only based on task-related events, but also on the

intermediate data produced the tasks. We contend that compared to traditional workflow execution systems this technique will ease development, increase flexibility and performance by removing implicit synchronization and automating previously tedious tasks related to workflow steering. We identify the classes of workflows that will benefit the most from this model and discuss design considerations for future implementations. In particular, we discuss how novel in-situ analysis techniques can be leveraged to implement a workflow system based on events of various natures and origins, from the infrastructure to the intermediate data while a workflow is running.

Scheduling for In-machine Analytics: Data Size is Important
Valentin Honore (LaBRI - Univ. Bordeaux, INRIA)

With the goal of performing exascale computing, the importance of I/O management becomes increasingly critical to maintain system performance. While the computing capacities of machines are getting higher, the I/O capabilities of systems do not follow the same trend.

To address this issue, the HPC community proposed new solutions such as online in-machine analysis to overcome the limitations of basic post-mortem data analysis where the data have to be stored on the Parallel File System (PFS) first to be processed later.

In this paper, we propose to study different scheduling strategies for in-machine analytics. Our goal is to extract the most important features of analytics that directly determine the efficiency of scheduling strategies. To do so, we propose a memory-constraint modelization for in-machine analysis. It automatically determines hardware resource partitioning and proposes scheduling policies for simulation and analysis.

We evaluate our model through simulations and observe that it is critical to base scheduling decisions on the memory needs of each analytics. We also note unexpected behaviors from which we deduce that modeling the in-machine paradigm for HPC applications requires deep understanding of task placement, data movement and hardware partitioning.

Lightning Round Questions

ISAV 2018 – Workshop Lunch (on your own)

In-Transit Molecular Dynamics Analysis with Apache Flink
Emilio Padrón (Universidade da Coruña)

In this paper, an on-line parallel analytics framework is proposed to process and store in transit all the data being generated by a Molecular Dynamics (MD) simulation run using staging nodes in the same cluster executing the simulation. The implementation and deployment of such a parallel workflow with standard HPC tools, managing problems such as data partitioning and load balancing can be a hard task for scientists. In this paper we propose to leverage Apache Flink, a scalable stream processing engine from the Big Data domain, in this HPC context. Flink enables to program analyses within a simple window based map/reduce model, while the runtime takes care of the deployment, load

balancing and fault tolerance. We build a complete in transit analytics workflow, connecting an MD simulation to Apache Flink and to a distributed database, Apache HBase, to persist all the desired data. To demonstrate the expressivity of this programming model and its suitability for HPC scientific environments, two common analytics in the MD field have been implemented. We assessed the performance of this framework, concluding that it can handle simulations of sizes used in the literature while providing an effective and versatile tool for scientists to easily incorporate on-line parallel analytics in their current workflows.

libIS: A Lightweight Library for Flexible In Transit Visualization
Will Usher (SCI Insitute, University of Utah; Intel Corporation)

As simulations grow in scale, the need for in situ analysis methods to handle the large data produced grows correspondingly. One desirable approach to in situ visualization is in transit visualization. By decoupling the simulation and visualization code, in transit approaches alleviate common difficulties with regard to the scalability of the analysis, ease of integration, usability, and impact on the simulation. We present libIS, a lightweight, flexible library which lowers the bar for using in transit visualization. Our library works on the concept of abstract regions of space containing data, which are transferred from the simulation to the visualization clients using a client-server model. We also provide a SENSEI analysis adaptor, which allows for transparent deployment of in transit visualization. We demonstrate the flexibility of our approach on batch analysis and interactive visualization use cases on different HPC resources.

Community Discussion

ISAV 2018 – Workshop Afternoon Break

Invited Talk: Data Science Meets CFD *Steve Legensky (Intelligent Light)*

UnPanel on the State of the In Situ Community
Come join our moderators as we engage in a community discussion about gaps, fundamental research opportunities, and the maturity of the community infrastructure.

ISAV 2018 Wrap Up

Room: D171
9:00 am - 5:30 pm

2nd International Workshop on Software Correctness for HPC Applications (Correctness 2018)

Ensuring the correctness of high-performance computing (HPC) applications is one of the fundamental challenges that developers and users of these applications face today. An application is correct when it performs what a user expects with respect to a specification. Given today's complex HPC software stack, correctness is very difficult to achieve: the use of combined parallel programing models (e.g., MPI+OpenMP), complex compiler optimizations/transformations, floating-point precision issues, and unanticipated scale-dependent behavior, are some of the challenges to achieve correctness. As emerging programing models and heterogeneous architectures become more predominant in HPC, the level of nondeterminism in applications increase, which makes the isolation of software bugs much harder. The aim of this workshop is to bring together researchers and developers to present and discuss novel ideas to address the problem of correctness in HPC.

Making Formal Methods for HPC Disappear *Ganesh L. Gopalakrishnan (University of Utah)*
Formal methods include rigorous specification methods that can render language standards reliable and unambiguous. They also include rigorous testing methods that target well-specified coverage criteria, and formal concepts that help guide debugging tool implementations. Those who say formal methods don't apply to HPC probably misunderstand formal methods to be some esoteric diversion, and not as a software productivity booster in the sense we describe.

Undoubtedly, HPC correctness is far too complex: there are the accidentally flipping bits, unpredictable floating point rounding, threads that incur a data race, and capricious compilers whose optimizations change results. All these can severely impact overall productivity. Formal approaches are possible for many of these pursuits, while for others they may emerge if one persists, and if there is a community invested in developing them in the long run. A worthwhile direction is to invest in formal methods based pedagogy: not only does this help buy us some time to develop useful formal methods for HPC, but it also gives some hope to save future generations from today's debugging drudgery. Today's parallel computing education still only gives lip service to correctness - let alone formal.

My talk will try and present examples of all of this. We will present examples where formal ideas did transition to production-level data race checking tools. I will also present examples where we finished production-level tools for floating-point non-reproducibility in the field, and hope to backfill the formalism eventually.

Eventually, as Rushby says, Formal Methods must "disappear" - be incorporated into standard practice and we don't see them.

Correctness – Workshop Morning Break

Hybrid Theorem Proving as a Lightweight Method for Verifying Numerical Software
Alper Altuntas (National Center for Atmospheric Research)

Large-scale numerical software requires substantial computer resources that complicate testing and debugging. single run of a climate model may require many millions of core-hours and terabytes of isk space, making trial-and-error experiments burdensome and time consuming. I this study, we a ply hybrid theorem proving from the field of cyber-physical systems to problems in scientific c mputation, and show how to verify the correctness of discrete updates that appear in the simulation o

continuous physical systems. By viewing numerical software as a hybrid system that combines di crete and continuous behavior, test coverage and confidence in findings can be increased. We es ribe abstraction approaches for modeling numerical software and demonstrate the applicability of the approach in a case study that reproduces undesirable behavior encountered in a parameterization sch me, called the K-profile parameterization, widely used in ocean components of large-scale climate mod els. We then identify and model a fix in the configuration of the scheme, and verify that the undesired behavior is eliminated for all possible execution sequences. We conclude that hybrid theorem proving is an effective and efficient approach that can be used to verify and reason about properties of large-scale numerical software.

HPC Software Verification in Action: A Case Study with Tensor Transposition
Erdal Mutlu (Pacific Northwest National Laboratory)

As HPC platforms get increasingly complex, the complexity of software optimized for these platforms has also increased. There is a pressing need to ensure correctness of scientific applications to enhance our confidence in the results they produce. In this paper, we focus on checking the correctness of libraries providing a small but important functionality---tensor transposition---used in computational chemistry applications. While several correctness tools have been developed and deployed, there are several practical challenges in using them to verify production HPC software. We present our experiences using two tools---CIVL and CodeThorn---in checking the correctness of two index permutation libraries. We observe that, with some effort, the tools we evaluated can handle kernels from production codes. We present observations that will aid library writers to write code that can be checked with these tools.

Correctness of Dynamic Dependence Analysis for Implicitly Parallel Tasking Systems
Wonchan Lee (Stanford University)

In this paper, we formally verify the correctness of dynamic dependence analysis, a key algorithm for parallelizing programs in implicitly parallel tasking systems. A dynamic dependence analysis of a program results in a task graph, a DAG of tasks constraining the order of task execution. Because a program is automatically parallelized based on its task graph, the analysis algorithm must generate a graph with all the dependencies that are necessary to preserve the program's original semantics for any non-deterministic parallel execution of tasks. However, this correctness is not straightforward to verify as implicitly parallel tasking systems often use an optimized dependence analysis algorithm. To study the correctness of dynamic dependence analysis in a realistic setting, we design a model algorithm that captures the essence of realistic analysis algorithms. We prove that this algorithm constructs task graphs that soundly and completely express correct parallel executions of programs. We also show that the generated task graph is the most succinct one for a program when the program satisfies certain conditions.

Verifying Qthreads: Is Model Checking Viable for User Level Tasking Runtimes?

This paper describes a formal specification and verification of an HPC runtime through the design, implemen- tation and evaluation of a model checked implementation of the Qthreads many task

runtime. We implement our model in Promela by doing a function to function translation of Qthreads' C implementation to Promela code. This translation works around the differences in modeling and implementation languages by translating C's rich pointer semantics, functions and non-local gotos to Promela's comparatively simple semantics. We then evaluate our implementation to show that it is both tractable and useful, exhaustively searching the state-space for counterexamples in reasonable time on modern architectures and use it to a lingering concurrency error in the Qthreads runtime.

Incremental Static Race Detection in OpenMP Programs
Bradley Swain (Texas A&M University)

OpenMP is a high level API that allows programmers to write concurrent programs on multi-core systems. OpenMP provides an interface for easily managing teams of threads and concurrent tasks allowing the programmer to focus on modeling a problem concurrently rather than dealing with low level thread management details for each system on which the code may run. Although OpenMP can automatically handle many parts of writing parallel programs (thread management, work distribution, scheduling, etc.), OpenMP offers no protections against data races. The programmer is left with the difficult task of ensuring that OpenMP programs are free of data races. Many concurrent programs, especially those related to high performance computing, can be difficult for programmers to reason through and are particularly prone to data races due to programmer error.

This paper presents an extensible and incremental static analysis technique for detecting data races in OpenMP programs at compile time. Our technique is comprised of two primary components:

(i): Array Index Analysis, which is used to detect when two array accesses made from separate threads may overlap resulting in multiple accesses to the same location in memory.

(ii): An Incremental Phase Graph is used to build a May Happen in Parallel (MHP) model that is utilized in determining if overlapping array accesses are capable of occurring concurrently.

Using Polyhedral Analysis to Verify OpenMP Applications Are Data Race Free

Among the most common and hardest to debug types of bugs in concurrent systems are data races. In this paper, we present an approach for verifying that an OpenMP program is data race free. We use polyhedral analysis to verify those parts of the program where we detect parallel affine loop nests. We show the applicability of polyhedral analysis for data race detection in HPC applications by evaluating our approach with the dedicated data race benchmark suite DataRaceBench and the LLNL Proxy Application AMG2013. Our evaluation shows that polyhedral analysis can classify 40% of the DataRaceBench 1.2.0 benchmarks as either data race free or having data races, and verify that 41 of the 114 (36%) loop nests of AMG2013 are data race free. AMG2013 consists of 75,000 LOC.

Correctness – Workshop Lunch (on your own)
Lunch on your own

Correctness of Floating Point Programs - Exception Handling and Reproducibility *James Demmel (University of California, Berkeley)*

We consider two related aspects of analyzing and guaranteeing correctness of floating point programs: exception handling and reproducibility. Exception handling refers to reliable and consistent propagation of errors due to overflow, invalid operations (like sqrt(-1)), convergence failures, etc. Reproducibility refers to getting bitwise reproducible results from multiple runs of the same program, e.g., despite parallelism causing floating point sums to be evaluated in different order with different roundoff errors. We describe the efforts of two standards committees, the Basic Linear Algebra Subprograms (BLAS) Standard, and the IEEE 754 Floating Point Standard, to address these issues, and how these efforts should make it easier to accomplish these goals for higher level applications, such as linear algebra libraries.

Correctness – Workshop Afternoon Break

Compiler-Aided Type Tracking for Correctness Checking of MPI Applications
Jan-Patrick Lehr (Technical University Darmstadt, Scientific Computing)

MUST, a dynamic MPI correctness checker, is extended with a type and memory allocation tracking sanitizer called TypeART for C/C++ codes based on the LLVM compiler framework. The sanitizer extracts type information and inserts instrumentation to track memory allocations and deallocations relevant to MPI communication. This allows MUST to check for type compatibility between the type-less communication buffer and the declared MPI datatype at all phases of the MPI communication, namely message assembly, message transfer and message disassembly into the receiving buffer. We evaluate our approach on benchmarks taken from SPEC MPI 2007 and two CORAL mini applications. The results show that our approach typically exhibits acceptable runtime and memory overheads. In particular, the memory overhead was below 20% in all cases.

Toward Deductive Verification of Message-Passing Parallel Programs
Ziqing Luo (University of Delaware)

Program verification techniques based on deductive reasoning can provide a very high level of assurance of correctness. These techniques are capable of proving correctness without placing artificial bounds on program parameters or on the sizes of inputs. While there are a number of mature frameworks for deductive verification of sequential programs, there is much less for parallel programs, and very little for message-passing. We propose a method for the deductive verification of message-passing programs that involves transforming the program into an annotated sequential program that can be verified with off-the-shelf deductive tools, such as Frama-C. The method can prove user-specified correctness properties without any bounds on the number of processes or other parameters. We illustrate this method on a toy example, and analyze its strengths and weaknesses.

PARCOACH Extension for a Full-Interprocedural Collectives Verification
Emmanuelle Saillard (French Institute for Research in Computer Science and Automation (INRIA)), Pierre Huchant (French Institute for Research in Computer Science and Automation (INRIA))

The advent to exascale requires more scalable and efficient techniques to help developers to locate, analyze and correct errors in parallel applications. PARallel COntrol flow Anomaly CHecker

(PARCOACH) is a framework that detects the origin of collective errors in applications using MPI and/or OpenMP. In MPI, such errors include collective operations mismatches. In OpenMP, a collective error can be a barrier not called by all tasks in a team. In this paper, we present an extension of PARCOACH which improves its collective errors detection. We show our analysis is more precise and accurate than the previous one on different benchmarks and real applications.

Facilitating the Adoption of Correctness Tools in HPC Applications
This open panel will discuss Facilitating the Adoption of Correctness Tools in HPC Applications.

Room: D173
9:00 am - 5:30 pm

4th Workshop for Open Source Supercomputing (OpenSuCo)

As we approach the end of lithographic/Dennard scaling, the HPC community needs a way to continue performance scaling. One way of providing that scaling is an increase in the number and diversity of specialized architectures tailored for specific applications. To accelerate the architecture specification and verification of these new architectures, more rapid prototyping methods are needed. At the same time, these new architectures need software stacks and programming models to be able to actually use these new designs.

In concert with this increase in architecture heterogeneity, there has been a consistent march toward development of open source based hardware and software solutions for each of these components to be used in lieu of existing closed source solutions.

We present the proposed third installment of OpenSuCo: Workshop for Open Source Supercomputing. We aim to continue the previous two workshops' focus on exploring and collaborating on building HPC systems using open-source hardware and system software IP. The goal is to engage the HPC community and explore open source solutions across the system stack – from silicon to applications.

Opening Remarks *David Donofrio (Lawrence Berkeley National Laboratory)*

Performance and Communication Modeling for Exascale Proxy Architecture in Aspen
Mariam Umar (Virginia Tech)

The performance capability of current high-performance computing cores are already at their threshold, leading to an unprecedented increase in number and complexity of exascale computational processors. Anticipating the performance and communication modeling required for millions of exascale cores necessitates profiling and modeling of proxy exascale applications on current supercomputers to help us pin-point to the application's hardware requirements to perform

corresponding what-if analysis. This has to be done while considering both applications and hardware, leading to efficient co-design for exascale proxy architectures.

In this paper, we study and profile the computation and communication of one of the exascale proxy applications -- Co-design for Molecular Dynamics (CoMD). We model one of the proposed exascale homogeneous architecture in the Aspen machine modeling language and develop the model into a refined proxy architecture. We model the CoMD's computational and communication requirements on Exascale proxy architecture in the Aspen application modeling language to study the scalability analysis. We also present an analysis about communication profiling for CoMD in Aspen, which guides the efficient communication mapping of the application on the proxy architecture.

FireSim: FPGA-Accelerated Cycle-Exact Scale-Out System Simulation in the Public Cloud *Sagar Karandikar (University of California, Berkeley)*

OpenSuCo – Workshop Morning Break

RV128 Instruction Set Architecture *Rashmi Agrawal (Boston University)*

PULP - An Open Parallel Ultra-Low-Power Processing-Platform *Luca Carloni (Columbia University)*

Panel: Open-Source Hardware

OpenSuCo – Workshop Lunch (on your own)

Keynote: Full Stack Open Source Supercomputing *Nicholas Malaya (Advanced Micro Devices Inc)*

HPC PowerStack: a community-wide open collaboration for enabling system-wide power efficiency
Siddhartha Jana (Intel Corporation)
One of the challenges to Open Supercomputing in the exascale era is the design of software solutions that drive Power Management across the system stack. In order to tackle the power challenges, the HPC community has designed various open source and vendor-specific solutions to meet power management goals at various levels of granularity - from system level to node-level. These components are built under certain assumptions of the system behavior and remain completely agnostic of each other's techniques. For example, an open source workload manager attempting to reduce the system-level power consumption may remain completely oblivious of the power demands by the application-level runtime and node-level hardware components. Attempting to integrate them into a unified system stack often lead to them overstepping on each other's solutions which, in turn leads to unreliable system performance. To avoid this, system integrators end up designing their stack with tightly-coupled cherry-picked vendor-specific solutions. This lack of coordination between the vendors and the open solution community leads to underutilization of system Watts and FLOPS. As a result, there is an urgent need for the HPC community to (A) identify the key software actors needed

in a system power stack: job-schedulers, application-level runtime, hardware knobs (B) arrive at a consensus on the roles and responsibilities of these actors, (C) design communication protocols for bidirectional control and feedback signals among the actors to enable scalable closed-loop coordination at different granularities, and (E) study and combine existing standalone engineering and development prototypes and build a community that actively participates in open development and engineering efforts.

This realization led to the formation of the PowerStack Community, in 2016. The participants of this community include global-wide members from academia, government labs, and vendors working on different layers of the system software stack. This talk is intended to spread awareness among the workshop attendees and solicit participation. The hope is to facilitate integrating open power management solutions into future PowerStack-compatible systems.

Open-Source Supercomputing *John Gustafson (National University of Singapore)*

OpenSuCo – Workshop Afternoon Break

GraphBLAS Forum and Its Relevant Software Zoo *Aydin Buluç (Lawrence Berkeley National Laboratory)*

Driving Asynchronous Distributed Tasks with Events
Nick Brown (University of Edinburgh)

Open-source matters, not just to the current cohort of HPC users, but also to potential new HPC communities, such as machine learning, themselves often rooted in open-source. Many of these potential new workloads are, by their very nature, far more asynchronous and unpredictable than traditional HPC codes and open-source solutions must be found to enable new communities of developers to easily take advantage of large scale parallel machines. Task-based models have the potential to help here, but many of these either entirely abstract the user from the distributed nature of their code, placing emphasis on the runtime to make important decisions concerning scheduling and locality, or require the programmer to explicitly combine their task-based code with a distributed memory technology such as MPI, which adds considerable complexity. In this paper we describe a new approach where the programmer still splits their code up into distinct tasks, but is explicitly aware of the distributed nature of the machine and drives interactions between tasks via events. This provides the best of both worlds; the programmer is able to direct important aspects of parallelism whilst still being abstracted from the low level mechanism of how this parallelism is achieved. We demonstrate our approach via two use-cases, the Graph500 BFS benchmark and in-situ data analytics of MONC, an atmospheric model. For both applications, we demonstrate considerably improved performance at large core counts and the result of this work is an approach and open-source library which is readily applicable to a wide range of codes.

Opportunities for Open-Source Development for D-Wave Systems *Steve Reinhardt (D-Wave Systems Inc)*
Classical supercomputers have open-source projects ranging through the technology stack from

applications, tools, compilers, operating systems, firmware, and even hardware. Quantum computers are less mature and have converged to common forms much less than classical computers. In this talk we examine the opportunities for open-source contributions at the various levels of the technology stack for D-Wave's annealing-based quantum computers. Many of the best opportunities are in mapping various problem types to the quadratic unconstrained binary optimization (QUBO) form optimized by the quantum processing unit.

Open-Source Modeling and Simulation *Gwendolyn Renae Voskuilen (Sandia National Laboratories)*

Panel: Open-Source Software

Closing Remarks *David Donofrio (Lawrence Berkeley National Laboratory)*

Room: D175
9:00 am - 5:30 pm

LLVM-HPC2018: The Fifth Workshop on the LLVM Compiler Infrastructure in HPC

LLVM, winner of the 2012 ACM Software System Award, has become an integral part of the software-development ecosystem for optimizing compilers, dynamic-language execution engines, source-code analysis and transformation tools, debuggers and linkers, and a whole host of programming-language and toolchain-related components. Now heavily used in both academia and industry, where it allows for rapid development of production-quality tools, LLVM is increasingly used in work targeted at high-performance computing. Research in, and implementation of, program analysis, compilation, execution, and profiling has clearly benefited from the availability of a high-quality, freely-available infrastructure on which to build. This workshop will focus on recent developments, from both academia and industry, that build on LLVM to advance the state of the art in high-performance computing.

Keynote: Glow: An Optimizing Compiler for High-Performance Machine Learning *Bert Maher (Facebook)*
Machine learning is an increasingly large fraction of datacenter workloads, making efficient execution of ML models a priority for industry. At the same time, the slow down of Moore's Law has created space for a plethora of innovative hardware designs to wring maximum performance from each transistor. To bridge the gap between software and hardware, we need compilers that understand both the characteristics of ML workloads and the nuances of the hardware. In this talk, I will describe how Facebook's Glow compiler leverages LLVM infrastructure to build a high-performance software stack for machine learning, by combining high-level domain-specific optimizations with customized low-level code generation strategies.

OpenMP GPU Offload in Flang and LLVM
Guray Ozen (Nvidia Corporation)

Graphics Processing Units (GPUs) have been widely adopted to accelerate the execution of High Performance Computing (HPC) workloads due to their enormous computational throughput, ability to execute a large number of threads inside SIMD groups in parallel, and their use of multithreaded hardware to hide long pipelining and memory access latency. However, developing applications able to exploit the high performance of GPUs requires proper code tuning. As a consequence, computer scientists proposed different approaches to simplify GPU programming, including directive-based programming models such as OpenMP and OpenACC. Their intention is to solve the aforementioned programming challenges with a directive-based approach which allows the users to insert non-executable pragma constructs that guide the compiler to handle the low-level complexities of the system. Flang, a Fortran front end for the LLVM Compiler Infrastructure, has drawn attention from the HPC community. Although Flang supports OpenMP for multicore architectures, it has no capability of offloading parallel regions to accelerator devices. In this paper, we present OpenMP Offload support in Flang targeting NVIDIA GPUs. Our goal is to investigate possible implementation strategies of OpenMP GPU offloading into Flang. The experimental results show that our approach is able to achieve performance similar to existing compilers with OpenMP GPU offload support.

Pointers Inside Lambda Closure Objects in OpenMP Target Offload Regions
David Truby (University of Warwick)

With the diversification of HPC architectures beyond traditional CPU-based clusters, a number of new frameworks for performance portability across architectures have arisen. One way of implementing such frameworks is to use C++ templates and lambda expressions to design loop-like functions. However, lower level programming APIs that these implementations must use are often designed with C in mind and do not specify how they interact with C++ features such as lambda expressions.

This paper discusses a change to the behavior of the OpenMP specification with respect to lambda expressions such that when functions generated by lambda expressions are called inside GPU regions, any pointers used in the lambda expression correctly refer to device pointers. This change has been implemented in a branch of the Clang C++ compiler and demonstrated with two representative codes. This change has also been accepted into the draft OpenMP specification for inclusion in OpenMP 5. Our results show that the implicit mapping of lambda expressions always exhibits identical performance to an explicit mapping but without breaking the abstraction provided by the high level frameworks.

Clacc: Translating OpenACC to OpenMP in Clang
Joel Denny (Oak Ridge National Laboratory)

OpenACC was launched in 2010 as a portable programming model for heterogeneous accelerators. Although various implementations already exist, no extensible, open-source, production-quality

compiler support is available to the community. This deficiency poses a serious risk for HPC application developers targeting GPUs and other accelerators, and it limits experimentation and progress for the OpenACC specification. To address this deficiency, Clacc is a recent effort funded by the US Exascale Computing Project to develop production OpenACC compiler support for Clang and LLVM. A key feature of the Clacc design is to translate OpenACC to OpenMP to build on Clang's existing OpenMP compiler and runtime support. In this paper, we describe the Clacc goals and design. We also describe the challenges that we have encountered so far in our prototyping efforts, and we present some early performance results.

LLVM and the Automatic Vectorization of Loops Invoking Math Routines: -fsimdmath
Francesco Petrogalli (ARM Ltd)

The vectorization of loops invoking math function is an important optimization that is available in most commercial compilers. This paper describes a new command line option, -fsimdmath, available in Arm Compiler for HPC, that enables auto-vectorization of math functions in C and C++ code, and that will also be applicable to Fortran code in a future version.

The design of -fsimdath is based on open standards and public architectural specifications. The library that provides the vector implementation of the math routines, libsimdmath.so, is shipped with the compiler and based on the SLEEF library libsleefgnuabi.so. SLEEF is a project that aims to provide a vector implementation of all C99 math functions, for a wide variety of vector extensions and architectures, across multiple platforms.

This feature is very important for HPC programmers, because the vector units of new CPUs are getting wider. Whether you are targeting Intel architectures with the AVX512 vector extension, or Arm architectures with the Scalable Vector Extension, good quality auto-vectorization is of increasing importance.

Although -fsimdmath has been implemented in a commercial compiler, it has been designed with portability and compatibility in mind, so that its use is not limited only to the vector extensions of the Arm architectures, but can be easily introduced as a major optimization for all the vector extensions that LLVM supports.

If accepted upstream, this new feature will enlarge the set of loops that LLVM will be able to auto-vectorize.

LLVM – Workshop Lunch (on your own) *Hal Finkel (Argonne National Laboratory)*

Function/Kernel Vectorization via Loop Vectorizer

Currently, there are three vectorizers in the LLVM trunk: Loop Vectorizer, SLP Vectorizer, and Load-Store Vectorizer. There is a need for vectorizing functions/kernels: 1) Function calls are an integral part of programming real world application code and we cannot always rely on fully inlining them. When a function call is made from a vectorized context such as vectorized loop or vectorized function, if there

are no vectorized callees available, the call has to be made to a scalar callee, one vector element at a time. At the programming model level, OpenMP declare simd is a standardized syntax to address this problem. LLVM needs a vectorizer to properly vectorize OpenMP declare simd functions. 2) Also, in the GPGPU programming model, such as OpenCL, work-item (thread) parallelism is not expressed with a loop; it is implicit in the execution of the kernels. In order to exploit SIMD parallelism at this top-level (thread-level), we need to start from vectorizing the kernel.

One of the obvious ways to vectorize functions/kernels is to add a fourth vectorizer that specifically deals with function vectorization. In this paper, we argue that such a naive approach will lead us to sub-optimal performance and/or higher maintenance burden. Instead, we present a technique to take advantages of the current functionalities and future improvements of Loop Vectorizer in order to vectorize functions and kernels.

User-Directed Loop-Transformations in Clang
Michael Kruse (Argonne National Laboratory, Argonne Leadership Computing Facility)

Directives for the compiler such as pragmas can help programmers to separate an algorithm's semantics from its optimization. This keeps the code understandable and easier to optimize for different platforms. Simple transformations such as loop unrolling are already implemented in most mainstream compilers.

We recently submitted a proposal to add generalized loop transformations to the OpenMP standard. We are also working on an implementation in LLVM/Clang/Polly to show its feasibility and usefulness. The current prototype allows applying patterns common to matrix-matrix multiplication optimizations.

LLVM – Workshop Afternoon Break *Hal Finkel (Argonne National Laboratory)*

OP2-Clang: A Source-to-Source Translator Using Clang/LLVM LibTooling
Gábor Dániel Balogh (Pázmány Péter Catholic University, Hungary)

Domain Specific Languages or Active Library frameworks have recently emerged as an important method for gaining performance portability, where an application can be efficiently executed on a wide range of HPC architectures without significant manual modifications. Embedded DSLs such as OP2, provides an API embedded in general purpose languages such as C/C++/Fortran. They rely on source-to-source translation and code refactorization to translate the higher-level API calls to platform specific parallel implementations. OP2 targets the solution of unstructured-mesh computations, where it can generate a variety of parallel implementations for execution on architectures such as CPUs, GPUs, distributed memory clusters and heterogeneous processors making use of a wide range of platform specific optimizations. Compiler tool-chains supporting source-to-source translation of code written in mainstream languages currently lack the capabilities to carry out such wide-ranging code transformations. Clang/LLVM's Tooling library (LibTooling) has long been touted as having such capabilities but have only demonstrated its use in simple source refactoring tasks.

In this paper, we introduce OP2-Clang, a source-to-source translator based on LibTooling, for OP2's C/C++ API, capable of generating target parallel code based on SIMD, OpenMP, CUDA and their

combinations with MPI. OP2-Clang is designed to significantly reduce maintenance, particularly making it easy to be extended to generate new parallelizations and optimizations for hardware platforms. In this research, we demonstrate its capabilities including (1) the use of LibTooling's AST matchers together with a simple strategy that use parallelization templates or skeletons to significantly reduce the complexity of generating radically different and transformed target code and (2) chart the challenges and solution to generating optimized parallelizations for OpenMP, SIMD and CUDA. Results indicate that OP2-Clang produces near-identical parallel code to that of OP2's current source-to-source translator. We believe that the lessons learned in OP2-Clang can be readily applied to developing other similar source-to-source translators, particularly for DSLs.

PInT: Pattern Instrumentation Tool for Analyzing and Classifying HPC Applications
Fabian Schlebusch (RWTH Aachen University), Sandra Wienke (RWTH Aachen University)

The relationship of application performance to its required development effort plays an important role in today's budget-oriented HPC environment. This effort-performance relationship is especially affected by the structure and characterization of an HPC application. We aim at a classification of HPC applications using (design) patterns for parallel programming. For an efficient analysis of parallel patterns and applicable pattern definitions, we introduce our tool PInT that is based on source code instrumentation and Clang LibTooling. Furthermore, we propose metrics to examine occurrences and compositions of patterns that can be automatically evaluated by PInT. In two case studies, we show the applicability and functionality of PInT.

AIWC: OpenCL-Based Architecture Independent Workload Characterization
Beau Johnston (Australian National University)

Measuring performance-critical characteristics of application workloads is important both for developers, who must understand and optimize the performance of codes, as well as designers and integrators of HPC systems, who must ensure that compute architectures are suitable for the intended workloads. However, if these workload characteristics are tied to architectural features that are specific to a particular system, they may not generalize well to alternative or future systems. An architecture-independent method ensures an accurate characterization of inherent program behaviour, without bias due to architecture-dependent features that vary widely between different types of accelerators. This work presents the first architecture-independent workload characterization framework for heterogeneous compute platforms, proposing a set of metrics determining the suitability and performance of an application on any parallel HPC architecture. The tool, AIWC, is a plugin for the open-source Oclgrind simulator. It supports parallel workloads and is capable of characterizing OpenCL codes currently in use in the supercomputing setting. AIWC simulates an OpenCL device by directly interpreting LLVM instructions, and the resulting metrics may be used for performance prediction and developer feedback to guide device-specific optimizations. An evaluation of the metrics collected over a subset of the Extended OpenDwarfs Benchmark Suite is also presented.

Compiler Optimization for Heterogeneous Locality and Homogeneous Parallelism in OpenCL and LLVM
Dorit Nuzman (Intel Corporation)

Heterogeneous platforms may include accelerators such as Digital Signal Processors (DSP's) that employ SW-controlled scratch-pad memories instead of, or in addition to standard HW-cached memory. Controlling scratch-pads efficiently typically requires tiling and pipelining loops, thereby optimizing for memory locality rather than parallelism as a primary objective. On the other hand, achieving high performance on CPU's and GPU's typically requires optimizing for data-level parallelism as a primary objective, compromising locality. In this lightning talk, we show how OpenCL and LLVM can be used to achieve both target-dependent locality and target-independent parallelism. Such an approach facilitates the development of optimized software for DSP accelerators while enabling its efficient execution on standard servers. Following the work of Tian et al., our approach leverages automatic compiler optimization and relies purely on OpenCL, including its device-side enqueue capability and SPIR-V format.

A Study of OpenMP Device Offloading in LLVM: Correctness and Consistency *Lechen Yu (Georgia Institute of Technology)*
To leverage widely available accelerators, OpenMP has introduced device constructs. Device constructs simplify the development of heterogeneous parallel programs and improve the performance. Many compilers including Clang already have support for device constructs, but there exist few documentations about the implementation details of device constructs. Lacking implementation details makes it cumbersome to understand the root cause of concurrency bugs and performance issues encountered on accelerators. In this paper, we conduct a study on Clang to analyze the implementation of device constructs for GPUs. We manually analyze the generated Parallel Thread Execution (PTX) code for each OpenMP construct to determine the relationship between the construct and PTX instructions. Based on the analysis, we evaluate the correctness of these constructs and discuss potential concurrency bugs incurred by incorrect usage of device constructs, for instance, data races, stale data and atomicity violation. Furthermore, we also talk about three observed inconsistencies in Clang, which may misinform programmers while writing an OpenMP program. Our work can help programmers gain a better understanding of device offloading and avoid hidden pitfalls when using Clang and OpenMP.

Challenges of C++ Heterogeneous Programming Using SYCL Implementation Experience: the Four Horsemen of the Apocalypse *Victor Lomuller (Codeplay Software Ltd)*
The C++ Direction Group has set a future direction for C++ and includes a guidance towards Heterogeneous C++. The introduction of the executors TS means for the first time in C++ there will be a standard platform for writing applications which can execute across a wide range of architectures including multi-core and many-core CPUs, GPUs, DSPs, and FPGAs.

The SYCL standard from the Khronos Group is a strong candidate to implement this upcoming C++ standard as are many other C++ frameworks from DOE, and HPX for the distributed case. One of SYCL's main strength is the capability to support constraint accelerator systems as it only requires OpenCL 1.2. One of the core ideas of the standard is that everything must be standard C++, the only exception being that some feature of C++ cannot be used in places that can be executed on an OpenCL device, often due to hardware limitation.

This paper presents some of the challenges and solutions to implement a Heterogeneous C++ standard in clang based on our implementation of Khrono's SYCL language with Codeplay's ComputeCpp compiler, with the fast growth of C++ and clang being a platform of choice to prototype many of the new C++ features.

We describe the major issues with ABI for separate compilation tool chain that comes from non-standard layout type of lambdas, as well as the issues of data addressing that comes from non-flat and possibly non-coherent address space.

We also describe various papers which are being proposed to ISO C++ to move towards standardizing heterogeneous and distributed computing in C++. The introduction of a unified interface for execution across a wide range of different hardware, extensions to this to support concurrent exception handling and affinity queries, and an approach to improve the capability of the parallel algorithms through composability. All of this adds up to a future C++ which is much more aware of heterogeneity and capable of taking advantage of it to improve parallelism and performance.

LLVM-HPC2018: Final Discussion *Hal Finkel (Argonne National Laboratory)*

Room: D220
9:00 am - 5:30 pm

8th Workshop on Python for High-Performance and Scientific Computing

Python is an established, high-level programming language with a large community in academia and industry. Scientists, engineers, and educators use Python for data science, high-performance computing, and distributed computing. Since Python is extremely easy to learn with a very clean syntax, it is well-suited for education in scientific computing. Programmers are much more productive by using Python.

The workshop will bring together researchers and practitioners using Python in all aspects of data science and high performance computing. The goal is to present Python applications, to discuss general topics regarding the use of Python, and to share experiences using Python in scientific computing. While Python is extremely strong in supporting human productivity as well reproducible science, it still lacks in computational performance compared to classical HPC languages such as Fortran or C. We especially encourage authors to submit novel research in improving performance of Python applications as well as research on productivity of development with Python.

Since Python itself can be seen as a ‘traditional’, established language for computational sciences, we like to move the focus of the workshop toward the role of Python within today's and future computing ecosystems. On the one hand, we foster discussion by seeking submissions that compare Python to other - dynamics and emerging - languages, which are widely used for HPC and scientific computing. On the other hand, we seek contributions on the use of Python on novel computing architectures, such as quantum computers or neuromorphic systems.

More information: http://www.dlr.de/sc/pyhpc2018

Keynote: Better Scientific Software (BSSw) *Anshu Dubey (Argonne National Laboratory), Stephen*

Hudson (Argonne National Laboratory)

PyHPC – Workshop Morning Break

AutoParallel: A Python Module for Automatic Parallelization and Distributed Execution of Affine Loop Nests
Cristian Ramon-Cortes (Barcelona Supercomputing Center)

The latest improvements in programming languages, programming models, and frameworks have focused on abstracting the users from many programming issues. Among others, recent programming frameworks include simpler syntax, automatic memory management and garbage collection, simplifies code re-usage through library packages, and easily configurable tools for deployment. For instance, Python has raised to the top of the list of the programming languages due to the simplicity of its syntax, while still achieving a good performance even being an interpreted language. Moreover, the community has helped to develop a large number of libraries and modules, tuning the most commonly used to obtain great performance.

However, there is still room for improvement when preventing users from dealing directly with distributed and parallel issues. This paper proposes and evaluates AutoParallel, a Python module to automatically find an appropriate task-based parallelization of affine loop nests to execute them in parallel in a distributed computing infrastructure. This parallelization can also include the building of data blocks to increase task granularity in order to achieve a good execution performance. Moreover, AutoParallel is based on sequential programming and only contains a small annotation in the form of a Python decorator so that anyone with little programming skills can scale up an application to hundreds of cores.

Managing Python in HPC Environments
Daniel Gall (Engility Corporation)

Python has seen a rapid adoption in the weather and climate modeling science communities. This swift rise has taken HPC system administrators by surprise, leading to inadequate support. These trends, like those in other sciences, led to the development and widespread adoption of user managed binary distributions. An example being Anaconda in 2012, which comes with security risks. We present a system for mirroring Anaconda Python that embeds PyRats, a dependency analyzer and logger descended from Blais' Snakefood into the Anaconda installers. We show that Anaconda performance, reliability, security, and availability can be improved dramatically and enable timely integration into automated test environments. The dependency logging yields insights into which packages users rely on most. It can help prioritize optimization efforts such as building packages for the CPU families used in a given HPC environment. We also discuss related work, including a complimentary automated Python provisioning effort by Oak Ridge National Laboratory that they call PythonEnv-noaa.

Panel: Interactivity in Supercomputing *William Scullin (Argonne National Laboratory), Rollin Thomas (Lawrence Berkeley National Laboratory)*

PyHPC – Workshop Lunch (on your own)

Accelerating the Signal Alignment Process in Time-Evolving Geometries Using Python

Vinay B. Ramakrishnaiah (Los Alamos National Laboratory)

This paper addresses the computational challenges involved in postprocessing of signals received using multiple collectors (satellites). Multiple low cost, small sized satellites can be used as dynamic beamforming arrays (DBA) in remote sensing satellites. This usually requires precise metrology and synchronized clocks. In order to mitigate this requirement, correlation searches can be performed across time and frequency offset values to align the signal. However, this process can take considerable time on traditional CPUs. We explore the use of heterogeneous parallel architectures to expedite the computation process, while trying to maintain the flexibility and ease of development using Python.

The Cross-Ambiguity Function (CAF) is used to perform correlation searches across a range of all possible frequency differences of arrival and time differences of arrival for a given emitter-collector geometry, followed by a phase alignment search. For evolving geometries, maintaining the signal alignment over long time periods require time evolving CAF searches, which is computationally expensive. Consequently, we explore the use of massively parallel architectures using both distributed and shared memory parallelism, and show performance results. We also propose a simple load balancing scheme for efficient use of heterogenous architectures.

We show that the NumPy implementation provides the same performance as the compiled Armadillo C++ code. Using different optimization techniques, the results show a performance improvement of 150x on a GPU compared to the naive implementation on a CPU.

Performance, Power, and Scalability Analysis of the Horovod Implementation of the CANDLE NT3 Benchmark on the Cray XC40 Theta

Xingfu Wu (Argonne National Laboratory, University of Chicago)

Training scientific deep learning models requires the large amount of computing power provided by HPC systems. In this paper, we use the distributed deep learning framework Horovod to parallelize NT3, a Python benchmark from the exploratory research project CANDLE (Cancer Distributed Learning Environment). We analyze NT3's scalability, performance, and power characteristics with different batch sizes and learning rates under two memory modes, cache and flat, on the DOE pre-exascale production system Cray XC40 Theta at Argonne National Laboratory. Our experimental results indicate that the power profiles for the node, CPU, and memory are useful in showing how the Horovod NT3 benchmark behaves on the underlying system. Using the communication timeline of this benchmark, we found that the Horovod communication overhead in NT3 increases significantly with the number of nodes although Horovod has the ability to scale up.

The benchmark leads to smaller runtime and lower power consumption for the node and CPU under the cache mode than under the flat mode. Furthermore, increasing the batch size leads to a runtime decrease and slightly impacts the power. Increasing the learning rate results in a slight decrease in

runtime and node power and an increase in accuracy. Several issues raised by the Horovod NT3 benchmark results are discussed, and suggestions are proposed for further work.

PyHPC – Workshop Afternoon Break

Data-Parallel Python for High Energy Physics Analyses
Marc Paterno (Fermi National Accelerator Laboratory)

In this paper, we explore features available in Python which are useful for data reduction tasks in High Energy Physics (HEP). High-level abstractions in Python are convenient for implementing data reduction tasks. However, in order for such abstractions to be practical, the efficiency of their performance must also be high. Because the data sets we process are typically large, we care about both I/O performance and in-memory processing speed. In particular, we evaluate the use of data-parallel programming, using MPI and numpy, to process a large experimental data set (42 TiB) stored in an HDF5 file. We measure the speed of processing of the data, distinguishing between the time spent reading data and the time spent processing the data in memory, and demonstrate the scalability of both, using up to 1200 KNL nodes (76800 cores) on Cori at NERSC.

Balsam: Automated Scheduling and Execution of Dynamic, Data-Intensive HPC Workflows
Michael A. Salim (Argonne National Laboratory)

We introduce the Balsam service to manage high-throughput task scheduling and execution on supercomputing systems. Balsam allows users to populate a task database with a variety of tasks ranging from simple independent tasks to dynamic multi-task workflows. With abstractions for the local resource scheduler and MPI environment, Balsam dynamically packages tasks into ensemble jobs and manages their scheduling lifecycle. The ensembles execute in a pilot "launcher" which (i) ensures concurrent, load-balanced execution of arbitrary serial and parallel programs with heterogeneous processor requirements, (ii) requires no modification of user applications, (iii) is tolerant of task-level faults and provides several options for error recovery, (iv) stores provenance data (e.g task history, error logs) in the database, (v) supports dynamic workflows, in which tasks are created or killed at runtime. Here, we present the design and Python implementation of the Balsam service and launcher. The efficacy of this system is illustrated using two case studies: hyperparameter optimization of deep neural networks, and high-throughput single-point quantum chemistry calculations. We find that the unique combination of flexible job-packing and automated scheduling with dynamic (pilot-managed) execution facilitates excellent resource utilization. The scripting overheads typically needed to manage resources and launch workflows on supercomputers are substantially reduced, accelerating workflow development and execution.

PyHPC Lightning Talks *William Spotz (Sandia National Laboratories)*

Room: D221
9:00 am - 5:30 pm

Ninth Annual Workshop for the Energy Efficient HPC Working Group (EE HPC WG)

Session Description: This annual workshop is organized by the Energy Efficient HPC Working Group (http://eehpcwg.llnl.gov/). This workshop closes the gap between facility and IT system with regards to energy efficiency analysis and improvements. For sustainable exascale computing, power and energy are a main concern, which can only be addressed by taking a holistic view combining the HPC facility, HPC system, HPC system software, and the HPC application needs. The EE HPC WG, which is a group with over 700 members from ~25 different countries, provides this cross-sectional perspective. This workshop is unique in that it provides a forum for sharing power and energy related information and research from supercomputing centers from around the world. Discussion and audience participation is encouraged. There are presentations, panels and discussions. Presenters are mostly from major governmental and academic supercomputing centers. The panels encourage discussion around more controversial topics and include panelists from supercomputing centers, academic institutions as well as the vendor community. SC17 topics included case studies of energy efficient operational lessons learned; the power grid- or "what you need to know about the power grid before adding a 10 MW step-function load generator"; the United States Department of Energy's Path Forward and other Exascale programs; and the software stack's implications for energy efficiency. The keynote speaker was Buddy Bland from ORNL. Buddy has seen more than 30 years of HPC deployment at ORNL and his keynote provided insight into operations and energy efficiency for some of the largest supercomputers.

This annual workshop is organized by the Energy Efficient HPC Working Group (http://eehpcwg.llnl.gov/). This workshop closes the gap between facility and IT system with regards to energy efficiency analysis and improvements. For sustainable exascale computing, power and energy are a main concern, which can only be addressed by taking a holistic view combining the HPC facility, HPC system, HPC system software, and the HPC application needs. The EE HPC WG, which is a group with over 700 members from ~25 different countries, provides this cross-sectional perspective.

This workshop is unique in that it provides a forum for sharing power and energy related information and research from supercomputing centers from around the world. Discussion and audience participation is encouraged. There are presentations, panels and discussions. Presenters are mostly from major governmental and academic supercomputing centers. The panels encourage discussion around more controversial topics and include panelists from supercomputing centers, academic institutions as well as the vendor community.

SC17 topics included case studies of energy efficient operational lessons learned; the power grid- or "what you need to know about the power grid before adding a 10 MW step-function load generator"; the United States Department of Energy's Path Forward and other Exascale programs; and the software stack's implications for energy efficiency. The keynote speaker was Buddy Bland from ORNL. Buddy has seen more than 30 years of HPC deployment at ORNL and his keynote provided insight into operations and energy efficiency for some of the largest supercomputers.

Keynote *Satoshi Matsuoka (Riken Center for Computational Science)*

EE HPC WG – Workshop Morning Break

State of the Working Group

The Power Grid
Talk titles:

1. Josip Loncaric: HPC Power Increasingly Challenges the Power Grid

2. Gary New: Electrical Utility Interface - Protections and Challenges

3. Ghaleb Abdulla: Forecasting Extreme Site Power Fluctuations Using Fast Fourier Transformation

Thermosyphon

Workshop Lunch (on your own)
Facilitated Discussion – (1) Liquid Cooling; (2) Software Stack; (3) Measurement, Monitoring, Management (location TBD)

Machine Installations at ORNL, LRZ, Sandia

EE HPC WG – Workshop Afternoon Break

Quantum Computing and Impact on Facility

Challenges in Holistic Monitoring and Data Integration (LLNL, LBNL, UniBo/Cineca)

Panel Discussion – Software for Energy Efficiency (PowerStack, PowerAPI, READEX, GEOPM)

Closing Remarks

Room: D222
9:00 am - 5:30 pm

Fifth SC Workshop on Best Practices for HPC Training and Education

High-Performance Computing (HPC) has become central for empowering progress in scientific, technical, social, business and medical domains by facilitating the solution of large-scale complex scientific and big data applications in a range of diverse fields. However, the inherent wide distribution, heterogeneity, and dynamism of current and future computing and software environments increasingly challenge new and existing practitioners along with the cyberinfrastructure facilitators, trainers and educators who are charged with developing and delivering HPC educational content and training. Furthermore, recent advances in educational technology and andragogical approaches create both challenges and opportunities for reaching a larger and more diverse community. The commonality of challenges faced by facilitators, educators and trainers necessitates community effort and involvement to develop and provide reproducible, reliable, and reusable educational materials.

Since 2014, the Best Practices for HPC Training workshop series at SC has provided a global forum for addressing common challenges and solutions for enhancing HPC training and education, for sharing information through large and small group discussions, and for fostering collaboration opportunities. The Fifth workshop, an ACM SIGHPC Education Chapter coordinated effort, is aimed at, and will provide opportunities for, extending collaborations among practitioners from traditional and emerging fields, exploring the challenges in developing and deploying HPC training and education, and identifying new challenges and opportunities for the latest HPC platforms. The workshop will also be a platform for disseminating results and lessons learned in these areas and will be captured in a Special Edition of the Journal of Computational Science Education.

Best Practices – Workshop Morning Break

HPC Education and Training: An Australian Perspective

Karina Nunez (Pawsey Supercomputing Centre), Maciej Cytowski (Pawsey Supercomputing Centre)

The Pawsey Supercomputing Centre has been running a variety of education, training and outreach activities addressed to all Australian researchers for a number of years. Based on experience and user feedback we have developed a mix of on-site and on-line training, roadshows, user forums and hackathon-type events. We have also developed an open repository of materials covering different aspects of HPC systems usage, parallel programming techniques as well as cloud and data resources usage. In this talk we will share our experience in using different learning methods and tools to address specific educational and training purposes. The overall goal will be to emphasise that there is no universal learning solution, instead various solutions and platforms need to be carefully selected for different groups of interest.

Trends in Demand, Growth, and Breadth in Scientific Computing Training Delivered by a High-Performance Computing Center

Ramses van Zon (SciNet HPC Consortium, University of Toronto)

We analyze the changes in the training and educational efforts of the SciNet HPC Consortium, a Canadian academic High Performance Computing Center, in the areas of Scientific Computing and

High-Performance Computing, over the last six years. Initially, SciNet offered isolated training events on how to use HPC systems and write parallel code, but the training program now consists of a broad range of workshops and courses that users can take toward certificates in scientific computing, data science, or high-performance computing. Using data on enrollment, attendance, and certificate numbers from SciNet's education website, used by almost 1800 users so far, we extract trends on the growth, demand, and breadth of SciNet's training program. Among the results are a steady overall growth, a sharp and steady increase in the demand for data science training, and a wider participation of 'non-traditional' computing disciplines, which has motivated an increasingly broad spectrum of training offerings. Of interest is also that many of the training initiatives have evolved into courses that can be taken as part of the graduate curriculum at the University of Toronto.

Evaluating Active Learning Approaches for Teaching Intermediate Programing at an Early Undergraduate Level

Dhruva Chakravorty (Texas A&M University)

There is a growing need to provide intermediate programing classes to STEM students early in their undergraduate careers. These efforts face significant challenges owing to the varied computing skill-sets of learners, requirements of degree programs and the absence of a common programing standard. Instructional scaffolding and active learning methods using Python offer avenues to support these students with varied needs. Here, we report on quantitative and qualitative outcomes from three distinct models of programing education that (i) connect coding to hands-on “maker” activities; (ii) incremental learning of computational thinking elements through guided exercises using Jupyter Notebooks; and (iii) problem-based learning with step-wise code fragments leading to algorithmic implementation. Performance in in-class activities, capstone projects, in-person interviews and extensive surveys informed us about the effectiveness of these approaches on various aspects of student learning. Students with previous coding experience were able to rely on broader skills and grasp concepts faster than students who recently attended an introductory programing session. We find that while maker-space activities were engaging and explained basic programing concepts, they lost their appeal in complex programing scenarios. Students grasped coding concepts fastest using the Jupyter notebooks, while the problem-based learning approach was best at having students understand the core problem and create inventive means to address them.

The Impact of MOOC Methodology on the Scalability, Accessibility and Development of HPC Education and Training

Julia Mullen (Massachusetts Institute of Technology)

This work explores the applicability of Massively Open Online Courses (MOOCs) for scaling High Performance Computing (HPC) training and education. Most HPC centers recognize the need to provide their users with HPC training; however, the current educational structure and accessibility prevents many scientists and engineers needing HPC knowledge and skills from becoming HPC practitioners. To provide more accessible and scalable learning paths toward HPC expertise, the authors explore MOOCs and their related technologies and teaching approaches. In this paper the authors outline how MOOC courses differ from face-to-face training, video-capturing of live events, webinars, and other established teaching methods with respect to pedagogical design, development issues and deployment concerns. The work proceeds to explore two MOOC case studies, including the design decisions, pedagogy and delivery. The MOOC development methods discussed are universal

and easily replicated by educators and trainers in any field; however, HPC has specific technical needs and concerns not encountered in other online courses. Strategies for addressing these HPC concerns are discussed throughout the work.

Training Computational Scientists to Build and Package Code
Prentice Bisbal (Princeton Plasma Physics Laboratory)

High performance computing training and education typically emphasizes the first-principles of scientific programming, such as numerical algorithms and parallel programming techniques. However, as applications and libraries proliferate on the Internet and mature, the need for computational scientists to write their own libraries and applications is decreasing. Instead, many computational scientists need to know how to compile and link to applications built by others. Likewise, those who create the libraries and applications need to understand how to organize their code to make it as portable as possible and package it so that it is straight-forward for others to use. These topics are not currently addressed by the current HPC training curriculum and users are typically left to develop their own approaches. This work will discuss observations made by the author over the last 20 years regarding the common problems encountered in the scientific community when developing their own codes and building codes written by other computational scientists. Recommendations will be provided for a training curriculum to address these shortcomings.

CiSE-ProS - Using Virtual Reality to Enforce Principles of Physical Cybersecurity
Jinsil Hwaryoung Seo (Texas A&M University), Dhruva K. Chakravorty (Texas A&M University)

The Cyberinfrastructure Security Education for Professionals and Students (CiSE-ProS) virtual reality environment is an exploratory project that uses evidence-generating approaches to evaluate the impact of learning environments produced by augmented reality (AR) and virtual reality (VR) technologies for teaching cybersecurity concepts. The program is steeped in well-reviewed pedagogy; the refinement of the educational methods based on constant assessment is a critical factor that has contributed to its success. In its current implementation, the program supports undergraduate student education. The overarching goal is to develop the CiSE-ProS VR program for implementation at institutions with low cyberinfrastructure (CI) adoption to where students may not have access to a physical data center to learn about the physical aspects of cybersecurity.

Best Practices – Workshop Lunch (on your own)

Toward a HPC Certification Program
Julian Kunkel (University of Reading)

The HPC community has always considered the training of new and existing HPC practitioners to be of high importance to its growth. This diversification of HPC practitioners challenges the traditional training approaches, which are not able to satisfy the specific needs of users, often coming from non-traditionally HPC disciplines, and only interested in learning a particular set of competences. Challenges for HPC centers are to identify and overcome the gaps in users' knowledge, while users struggle to identify relevant skills.

We have developed a first version of an HPC certification program that would clearly categorize, define, and examine competences. Making clear what skills are required of or recommended for a competent HPC user would benefit both the HPC service providers and practitioners. Moreover, it would allow centers to bundle together skills that are most beneficial for specific user roles and scientific domains. From the perspective of content providers, existing training material can be mapped to competences allowing users to quickly identify and learn the skills they require. Finally, the certificates recognized by the whole HPC community simplify inter-comparison of independently offered courses and provide additional incentive for participation.

Potential Influence of Prior Experience in an Undergraduate-Graduate Level HPC Course
Chris Fietkiewicz (Case Western Reserve University)

A course on High Performance Computing at Case Western Reserve University included students with a range of technical and academic experience. We consider these experiential differences with regard to student performance and perceptions. The course relied heavily on C programming and multithreading, but one third of the students had no prior experience with these techniques. Academic experience also varied, as the class included 3rd and 4th year undergraduates, master's students, PhD students, and non-degree students. Results indicate that student performance did not depend on technical experience. However, average performance was slightly higher for graduate students. Additionally, we report on students' perceptions of the course, including opinions on the level of theoretical coverage and expectations for report writing.

Deep Learning by Doing: Nvidia Deep Learning Institute
Xi Chen (University of Kentucky), Joe Bungo (Nvidia Corporation)

Over the past two decades, High-Performance Computing (HPC) communities have developed many models for delivering education aiming to help students understand and harness the power of parallel and distributed computing. Most of these courses either lack hands-on experience or heavily focus on theoretical characterization behind complex algorithms. To bridge the gap between application and scientific theory, NVIDIA Deep Learning Institute (DLI) has designed an online education and training platform (https://courses.nvidia.com/) that helps students, developers, and engineers solve real-world problems in a wide range of domains using deep learning and accelerated computing. Accelerated computing course content starts with the fundamentals of accelerating applications with CUDA and OpenACC in addition to other courses in training and deploying neural networks for deep learning. Advanced and domain-specific courses in deep learning are also available. The online platform enables students to use the latest AI frameworks, SDKs, and GPU-accelerated technologies on fully-configured GPU servers in the cloud so the focus is more on learning and less on environment setup. Students are offered project-based assessment and certification at the end of each course. To support academics and university researchers teaching accelerated computing and deep learning, the DLI University Ambassador Program enables educators to teach free DLI courses to university students, faculty, and researchers.

Using CloudLab as a Scalable Platform for Teaching Cluster Computing
Jeff Denton (Clemson University)

A significant challenge in teaching cluster computing, an advanced topic in the parallel and distributed computing body of knowledge, is to provide students with an adequate environment where they can become familiar with real-world infrastructures that embody the conceptual principles taught in lectures. In this paper, we describe our experience setting up such an environment by leveraging CloudLab, a national experimentation platform for advanced computing research. We explored two approaches in using CloudLab to teach advanced concepts in cluster computing: direct deployment of virtual machines (VMs) on bare-metal nodes and indirect deployment of VMs inside a CloudLab-based cloud.

Programmable Education Infrastructure: Cloud Resources as HPC Education Environments
J. Eric Coulter (Indiana University, XSEDE)

Cloud computing is a growing area for educating students and performing meaningful scientific research. The challenge for many educators and researchers is knowing how to use some of the unique aspects of computing in the cloud. One key feature is elastic computing - resources on demand. The elasticity and programmability of cloud resources make them an excellent tool for educators who require access to a wide range of computing environments. In the field of HPC education, such environments are an absolute necessity, and getting access to them can create a large burden on educators above and beyond designing content.

While cloud resources won't replace traditional HPC environments for large research projects, they are an excellent option for providing both user and administrator education on HPC environments. The highly configurable nature of cloud environments allows educators to tailor the educational resource to the needs of their attendees, and provide a range of hands-on experiences. In this demo, we'll show how the Jetstream cloud environment can be used to provide training for both new HPC administrators and users, by showing a ground-up build of a simple HPC system. While this approach uses the Jetstream cloud, it is generalizable across any cloud provider. We will show how this allows an educator to tackle everything from basic command-line concepts and scheduler use to advanced cluster-management concepts such as elasticity and management of scientific software. We will also discuss lessons learned from providing this content as a tutorial at several CI conferences.

The HPC Best Practices Webinar Series
Osni Marques (Lawrence Berkeley National Laboratory)

The Best Practices for HPC Software Developers (HPC-BP) webinar series is a major component of the outreach efforts of the IDEAS (Interoperable Design of Extreme-scale Application Software) Productivity Project, funded by the US Department of Energy's (DOE) Exascale Computing Project (ECP). Since its inception, the IDEAS Project has been addressing the merging of trends in hardware and increasing demands for predictive multiscale, multiphysics simulations. It has been also responding to trends for continuous refactoring with efficient agile software engineering methodologies and improved software design. The webinar series originated when the IDEAS Project was particularly focused on working with the terrestrial ecosystem modeling community within DOE, and has been moderately adapted as the focus (and sponsorship) of the IDEAS Project shifted to ECP, wherein the focus is on helping both application and software tools teams be more productive in their software development efforts and produce more sustainable results. In this contribution, we discuss the process we have adopted for HPC-BP, and give a sample of the webinars that we have organized

and delivered (22 webinars at the time of this writing). We provide an overview of the process we follow for the selection of topics, how the webinars are executed, unique features of the series and the future we foresee for it.

Best Practices – Workshop Afternoon Break

Discussion Session

Breakout Groups

Friday, November 16th

Room: D161
8:30 am - 12:00 pm

Deep Learning on Supercomputers

The 1st Deep Learning (DL) on Supercomputers workshop provides a forum for practitioners working on any and all aspects of DL in the supercomputing context to present their latest research results and development, deployment, and application experiences. The workshop's theme is the intersection of DL and supercomputing: novel uses of supercomputers to accelerate deep learning training and inference, and innovative applications of deep learning in traditional numerical simulation. Its scope encompasses application development in industrial and scientific scenarios using supercomputers; DL methods applied to numerical simulation; fundamental algorithms, enhanced procedures, and software development methods to enable scalable training and inference; hardware changes with impact on future supercomputer design; and machine deployment, performance evaluation, and reproducibility practices for DL applications on supercomputers. This workshop presents a series of invited talks from researchers who are working on the crossroads of deep learning and supercomputing.

Keynote: Learning-Based Predictive Models: a New Approach to Integrating Large-Scale Simulations and Experiments

CANDLE Framework for Large Scale Deep Learning

Fast and Accurate Deep Neural Networks Training on Distributed Systems

Deep Learning – Workshop Morning Break

Deep Learning at NERSC: Usability, Capability, and Everything in Between

Artificial Intelligence Enabled Multiscale Molecular Simulations

Scalable and Distributed DNN Training on Modern HPC Systems

High-Performance AI: A View from Systems and Frameworks

Large Scale Deep Learning in PFN: from 15-Min Imagenet to PFDet

Enabling Scalable and Efficient Deep Learning on Supercomputers

Room: D167
8:30 am - 12:00 pm

International Workshop on Performance, Portability, and Productivity in HPC (P3HPC)

The ability for applications to achieve both portability and high performance across computer architectures remains an open challenge. It is often unrealistic or undesirable for developers to maintain separate implementations for each target architecture, yet in many cases, achieving high performance and fully utilizing an architecture's underlying features requires the use of specialized language constructs and libraries. Likewise, abstractions and standards that promise portability cannot necessarily deliver high performance without additional algorithmic considerations, and performance compromises are often made to remain portable. Application developers, who strive to work productively while balancing these concerns, often find the goal to be elusive.

There is a clear need to develop ways of managing the complexity that arises from system diversity that balance the need for performant specializations with the economy of appropriate and efficient abstractions. Despite growth in the number of available architectures, there are similarities that represent general trends in current and emerging HPC hardware: increased thread parallelism; wider vector units; and deep, complex, memory hierarchies. This in turn offers some hope for common programming techniques and language support as community experience matures.

The purpose of this workshop is to provide an opportunity for attendees to share ideas, practical experiences, and methodologies for tackling the challenge of achieving performance portability and developer productivity across current and future homogeneous and heterogeneous computer architectures.

High-Performance Molecular Dynamics Simulation for Biological and Materials Sciences: Challenges of Performance Portability
Ada Sedova (Oak Ridge National Laboratory)

Highly-optimized parallel molecular dynamics programs have, in recent years, allowed researchers to achieve ground-breaking results in biological and materials sciences. This type of performance has come at the expense of portability: a significant effort is required for performance optimization on each new architecture. Using a metric that emphasizes speedup and time-to-solution, we analyze the code-bases and performance portabilities of four different high-performing molecular dynamics programs--GROMACS, NAMD, LAMMPS and CP2K-- each having a particular scope of application. We find that for all four programs, the contributions of the non-portable components to speed are essential to the programs' performances; without them we see a reduction in time-to-solution of a magnitude that is insufferable to domain scientists. We discuss possible solutions to this difficult problem, which must come from developers, industry and funding institutions, and possibly new developments in programming languages.

An Empirical Roofline Methodology for Quantitatively Assessing Performance Portability
Charlene Yang (Lawrence Berkeley National Laboratory)

System and node architectures continue to diversify to better balance on-node computation, memory capacity, memory bandwidth, interconnect bandwidth, power, and cost for specific computational workloads. For many applications developers, however, achieving performance portability (effectively exploiting the capabilities of multiple architectures) is a desired goal. Unfortunately, dramatically different per-node performance coupled with differences in machine balance can lead to developers being unable to determine whether they have attained performance portability or simply written portable code. The Roofline model provides a means of quantitatively assessing how well a given application makes use of a target platform's computational capabilities. In this paper, we extend the Roofline model so that it 1) empirically captures a more realistic set of performance bounds for CPUs and GPUs, 2) factors in the true cost of different floating-point instructions when counting FLOPs, 3) incorporates the effects of different memory access patterns, and 4) with appropriate pairing of code performance and Roofline ceiling, facilitates the performance portability analysis.

Effective Performance Portability
Stephen Lien Harrell (Purdue University), Robert Robey (Los Alamos National Laboratory)

Exascale computing brings with it diverse machine architectures and programming approaches which challenge application developers. Applications need to perform well on a wide range of architectures while simultaneously minimizing development and maintenance overheads. In order to alleviate these costs, developers have begun leveraging portability frameworks to maximize both the code shared between platforms and the performance of the application. We explore the effectiveness of several such frameworks through applying them to small production codes. Throughout the process, we apply a logging tool to gather data on the development process. We use this information to develop metrics of application development productivity, which can be used to holistically assess how productively a performance-portable application was developed.

Evaluating the Impact of Proposed OpenMP 5.0 Features on Performance, Portability, and Productivity
Simon J. Pennycook (Intel Corporation)

We investigate how specialization mechanisms proposed for OpenMP 5.0 -- specifically, the metadirective and declare variant directives -- may be deployed in a real-life code, using the miniMD benchmark from the Mantevo suite.

Additionally, we develop an OpenMP 4.5 implementation of miniMD that achieves a performance portability of 59.35% across contemporary CPU and GPU hardware, discuss the processes of porting and enabling this code, and show that the use of specialization would enable our code to be expressed in a significantly more compact form, with implications for productivity.

P3HPC Session 1 Panel Discussion *Doug Doerfler (Lawrence Berkeley National Laboratory)*
Moderated panel discussion

P3HPC – Workshop Morning Break *Kathryn O'Brien (IBM)*

Performance Portability of an Unstructured Hydrodynamics Mini-Application
Timothy R. Law (University of Warwick)

In this work we study the parallel performance portability of BookLeaf: a recent 2D unstructured hydrodynamics mini-application. The aim of BookLeaf is to provide a self-contained and representative testbed for exploration of the modern hydrodynamics application design-space.

We present a previously unpublished reference C++11 implementation of BookLeaf parallelised with MPI, alongside hybrid MPI+OpenMP and MPI+CUDA versions, and two implementations using C++11 performance portability frameworks: Kokkos and RAJA, which both target a variety of parallel back-ends. We assess the scalability of our implementations on the ARCHER Cray XC30 up to 4096 nodes (98,304 cores) and on the Ray EA system at Lawrence Livermore National Laboratory up to 16 nodes (64 Tesla P100 GPUs), with a particular focus on the overheads introduced by Kokkos and RAJA relative to our handwritten OpenMP and CUDA implementations. We quantify the performance portability achieved by our Kokkos and RAJA implementations across five modern architectures using a metric previously introduced by Pennycook et al.

We find that our BookLeaf implementations all scale well, in particular the hybrid configurations (the MPI+OpenMP variant achieves a parallel efficiency above 0.8 running on 49,152 cores). The Kokkos and RAJA variants exhibit competitive performance in all experiments, however their CPU performance is best in memory-bound situations where the overhead introduced by the frameworks is partially shadowed by the need to wait for data. The overheads seen in the GPU experiments are extremely low. We observe overall performance portability scores of 0.928 for Kokkos and 0.876 for RAJA.

Performance Portability Challenges for Fortran Applications
Abigail Hsu (Stony Brook University, Los Alamos National Laboratory), David Howard Neill (Grinnell College, Los Alamos National Laboratory)

This project investigates how different approaches to parallel optimization impact the performance portability for Fortran codes. In addition, we explore the productivity challenges due to the software tool-chain limitations unique to Fortran. For this study, we build upon the Truchas software, a metal casting manufacturing simulation code based on unstructured mesh methods and our initial efforts for accelerating two key routines, the gradient and mimetic finite difference calculations. The acceleration methods include OpenMP, for CPU multi-threading and GPU offloading, and CUDA for GPU offloading. Through this study, we find that the best optimization approach is dependent on the priorities of performance versus effort and the architectures that are targeted. CUDA is the most attractive where performance is the main priority, whereas the OpenMP on CPU and GPU approaches are preferable when emphasizing productivity. Furthermore, OpenMP for the CPU is the most portable across architectures. OpenMP for CPU multi-threading yields 3%-5% of achievable performance, whereas the GPU offloading generally results in roughly 74%-90% of achievable performance. However, GPU offloading with OpenMP 4.5 results in roughly 5% peak performance for the mimetic finite difference algorithm, suggesting further serial code optimization to tune this kernel. In general,

these results imply low performance portability, below 10% as estimated by the Pennycook metric. Though these specific results are particular to this application, we argue that this is typical of many current scientific HPC applications and highlights the hurdles we will need to overcome on the path to exascale.

Delivering Performance-Portable Stencil Computations on CPUs and GPUs Using Bricks

Tuowen Zhao (University of Utah)

Achieving high performance on stencil computations poses a number of challenges on modern architectures. The optimization strategy varies significantly across architectures, types of stencils, and types of applications. The standard approach to adapting stencil computations to different architectures, used by both compilers and application programmers, is through the use of iteration space tiling, whereby the data footprint of the computation and its computation partitioning are adjusted to match the memory hierarchy and available parallelism of different platforms. In this paper, we explore an alternative performance portability strategy for stencils, a data layout library for stencils called bricks, that adapts data footprint and parallelism through fine-grained data blocking. Bricks are designed to exploit the inherent multi-dimensional spatial locality of stencils, facilitating improved code generation that can adapt to CPUs or GPUs, and reducing pressure on the memory system. We demonstrate that bricks are performance-portable across CPU and GPU architectures and afford performance advantages over various tiling strategies, particularly for modern multi-stencil and high-order stencil computations. For a range of stencil computations, we achieve high performance on both the Intel Knights Landing (Xeon Phi) and Skylake (Xeon) CPUs as well as the Nvidia P100 (Pascal) GPU delivering up to a 5x speedup against tiled code.

Heterogeneous CPU-GPU Execution of Stencil Applications

Balint Siklosi (Pázmány Péter Catholic University, Hungary)

Heterogeneous computer architectures are now ubiquitous in high performance computing; the top 7 supercomputers are all built with CPUs and accelerators. Portability across different CPUs and GPUs is becoming paramount, and heterogeneous scheduling of computations is also of increasing interest to make full use of these systems. In this paper, we present research on the hybrid CPU-GPU execution of an important class of applications: structured mesh stencil codes. Our work broadens the performance portability capabilities of the Oxford Parallel library for Structured meshes (OPS), which allows a science code written once at a high level to be automatically parallelised for a range of different architectures. We explore the traditional per-loop load balancing approach used by others, and highlighting its shortcomings, we develop an algorithm that relies on polyhedral analysis and transformations in OPS to allow load balancing on the level of larger computational stages, reducing data transfer requirements and synchronisation points.

We evaluate our algorithms on a simple heat equation benchmark, as well as a substantially more complex code, the CloverLeaf hydrodynamics mini-app. To demonstrate performance portability, we study Intel and IBM systems equipped with NVIDIA Kepler, Pascal, and Volta GPUs, evaluating CPU-only, GPU-only and hybrid CPU-GPU performance. We demonstrate a 1.05-1.2x speedup on CloverLeaf. Our results highlight the ability of the OPS domain specific language to deliver effortless performance portability for its users across a number of platforms.

P3HPC Session 2 Panel Discussion *John Pennycook (Intel Corporation)*

P3HPC Community Discussion and Next Steps *Rob Neely (Lawrence Livermore National Laboratory)*

Room: D174
8:30 am - 12:00 pm

Workshop on Fault-Tolerance for HPC at Extreme Scale (FTXS)

Addressing failures in extreme-scale systems remains a significant challenge to reaching exascale. Current projections suggest that at the scale necessary to sustain exaflops of computation, systems could experience failures as frequently as once per hour. As a result, robust and efficient fault tolerance techniques are critical to obtaining reasonable application performance. Additionally, it is also imperative that we develop and understanding of trends in hardware devices may affect the reliability of future systems. The emergence of high-bandwidth memory devices, the continued deployment of burst buffers, and the development of near-threshold devices to address power concerns will all impact fault tolerance on new systems. These design trends coupled with increases in the number, variety, and complexity of components required to compose an extreme-scale system means that systems will experience significant increases in aggregate fault rates, fault diversity, and the complexity of isolating root causes.

Due to the continued need for research on fault tolerance in extreme-scale systems, the 8th Workshop on Fault-Tolerance for HPC at Extreme Scale (FTXS 2018) will present an opportunity for innovative research ideas to be shared, discussed, and evaluated by researchers in fault-tolerance, resilience, dependability, and reliability from academic, government, and industrial institutions. Building on the success of the previous editions of the FTXS workshop, the organizers will assemble quality publications, invited talks, and panels to facilitate a lively and thought-provoking group discussion.

Towards Ad Hoc Recovery For Soft Errors
Leonardo Bautista-Gomez (Barcelona Supercomputing Center)

The coming exascale era is a great opportunity for high performance computing (HPC) applications. However, high failure rates on these systems will hazard the successful completion of their execution. Bit-flip errors in dynamic random access memory (DRAM) account for a noticeable share of the failures in supercomputers. Hardware mechanisms, such as error correcting code (ECC), can detect and correct single-bit errors and can detect some multi-bit errors while others can go undiscovered. Unfortunately, detected multi-bit errors will most of the time force the termination of the application and lead to a global restart. Thus, other strategies at the software level are needed to tolerate these type of faults more efficiently and to avoid a global restart. In this work, we extend the FTI checkpointing library to facilitate the implementation of custom recovery strategies for MPI applications, minimizing the

overhead introduced when coping with soft errors. The new functionalities are evaluated by implementing local forward recovery on three HPC benchmarks with different reliability requirements. Our results demonstrate a reduction on the recovery times by up to 14%.

Fault Tolerant Cholesky Factorization on GPUs

Parameswaran Ramanathan (University of Wisconsin-Madison)

Direct Cholesky-based solvers are typically used to solve large linear systems where the coefficient matrix is symmetric positive definite. These solvers offer faster performance in solving such linear systems, compared to other more general solvers such as LU and QR solvers. In recent days, graphics processing units (GPUs) have become a popular platform for scientific computing applications, and are increasingly being used as major computational units in supercomputers. However, GPUs are susceptible to transient faults caused by events such as alpha particle strikes and power fluctuations. As a result, the possibility of an error increases as more and more GPU computing nodes are used. In this paper, we introduce two efficient fault tolerance schemes for the Cholesky factorization method, and study their performance using a direct Cholesky solver in the presence of faults. We utilize a transient fault injection mechanism for NVIDIA GPUs and compare our schemes with a traditional checksum fault tolerance technique, and show that our proposed schemes have superior performance, good error coverage and low overhead.

Improving Application Resilience by Extending Error Correction with Contextual Information

Alexandra L. Poulos (Coastal Carolina University), William M. Jones (Coastal Carolina University)

Extreme-scale systems are growing in scope and complexity as we approach exascale. Uncorrectable faults in such systems are also increasing, so resilience efforts addressing these are of great importance. In this paper, we extend a method that augments hardware error detection and correction (EDAC) contextually, and show an application-based approach that takes detectable uncorrectable (DUE) data errors and corrects them.

We applied this application-based method successfully to data errors found using common EDAC, and discuss operating system changes that will make this possible on existing systems. We show that even when there are many acceptable correction choices (which may be seen in floating point), a large percentage of DUEs are corrected, and even the miscorrected data are very close to correct. We developed two different contextual criteria for this application: local averaging and global conservation of mass. Both did well in terms of closeness, but conservation of mass outperformed averaging in terms of actual correctness.

The contributions of this paper are: 1) the idea of application-specific EDAC-based contextual correction, 2) its demonstration with great success on a real application, 3) the development of two different contextual criteria, and 4) a discussion of attainable changes to the OS kernel that make this possible on a real system.

A Comprehensive Informative Metric for Analyzing HPC System Status Using the LogSCAN Platform

Yawei Hui (Oak Ridge National Laboratory)

Log processing by Spark and Cassandra-based ANalytics (LogSCAN) is a newly developed analytical platform that provides flexible and scalable data gathering, transformation and computation. One major challenge is to effectively summarize the status of a complex computer system, such as the Titan supercomputer at the Oak Ridge Leadership Computing Facility (OLCF). Although there is plenty of operational and maintenance information collected and stored in real time, which may yield insights about short- and long-term system status, it is difficult to present this information in a comprehensive form. In this work, we present system information entropy (SIE), a newly developed metric that leverages the powers of traditional machine learning techniques and information theory. By compressing the multi-variant multi-dimensional event information recorded during the operation of the targeted system into a single time series of SIE, we demonstrate that the historical system status can be sensitively represented concisely and comprehensively. Given a sharp indicator as SIE, we argue that follow-up analytics based on SIE will reveal in-depth knowledge about system status using other sophisticated approaches, such as pattern recognition in the temporal domain or causality analysis incorporating extra independent metrics of the system.

FTXS 2018 - Workshop Morning Break *Scott Levy (Sandia National Laboratories)*

Analyzing the Impact of System Reliability Events on Applications in the Titan Supercomputer
Rizwan A. Ashraf (Oak Ridge National Laboratory)

Extreme-scale computing systems employ Reliability, Availability and Serviceability (RAS) mechanisms and infrastructure to log events from multiple system components. In this paper, we analyze RAS logs in conjunction with the application placement and scheduling database, in order to understand the impact of common RAS events on application performance. This study conducted on the records of about 2 million applications executed on Titan supercomputer provides important insights for system users, operators and computer science researchers. We investigate the impact of RAS events on application performance and its variability by comparing cases where events are recorded with corresponding cases where no events are recorded. Such a statistical investigation is possible since we observed that system users tend to execute their applications multiple times. Our analysis reveals that most RAS events do impact application performance, although not always. We also find that different system components affect application performance differently. In particular, our investigation includes the following components: parallel file system, processor, memory, graphics processing units, system and user software issues. Our work establishes the importance of providing feedback to system users for increasing operational efficiency of extreme-scale systems.

Extending and Evaluating Fault-Tolerant Preconditioned Conjugate Gradient Methods
Carlos Pachajoa (University of Vienna)

We compare and refine exact and heuristic fault-tolerance extensions for the \emph{preconditioned conjugate gradient} (PCG) and the \emph{split preconditioner conjugate gradient} (SPCG) methods for recovering from failures of compute nodes of large-scale parallel computers. In the \emph{exact state reconstruction} (ESR) approach, which is based on a method proposed by Chen (2011), the solver keeps extra information from previous search directions of the (S)PCG solver, so that its state can be fully reconstructed if a node fails unexpectedly. ESR does not make use of checkpointing or external storage for saving dynamic solver data and has only negligible computation and communication

overhead compared to the failure-free situation. In exact arithmetic, the reconstruction is exact, but in finite-precision computations, the number of iterations until convergence can differ slightly from the failure-free case due to rounding effects. We perform experiments to investigate the behavior of ESR in floating-point arithmetic and compare it to the heuristic \emph{linear interpolation} (LI) approach by Langou et al. (2007) and Agullo et al. (2016), which does not have to keep extra information and thus has lower memory requirements. Our experiments illustrate that ESR, on average, has essentially zero overhead in terms of additional iterations until convergence, whereas the LI approach incurs much larger overheads.

CPU Overheating Characterization in HPC Systems: a Case Study
Marc PLATINI (University of Grenoble, Atos)

With the increase in size of supercomputers, also increases the number of abnormal events. Some of these events might lead to an application failure. Others might simply impact the system efficiency. CPU overheating is one such event that decreases the system efficiency: when a CPU overheats, it reduces its frequency. This paper studies the problem of CPU overheating in supercomputers. In a first part, we analyze data collected over one year on a supercomputer of the top500 list to understand under which conditions CPU overheating occurs. Our analysis show that overheating events are due to some specific applications. In a second part, we evaluate the impact of such overheating events on the performance of MPI applications. Using 6 representative HPC benchmarks, we show that for a majority of the applications, a frequency drop on one CPU impacts the execution time of distributed runs proportionally to the duration and to the extent of the frequency drop.

SaNSA - the Supercomputer and Node State Architecture

In this work, we present SaNSA, the Supercomputer and Node State Architecture, a software infrastructure for historical analysis and anomaly detection. SaNSA consumes data from multiple sources including system logs, the resource manager, scheduler, and job logs. Furthermore, additional context such as scheduled maintenance events or dedicated application run times for specific science teams can be overlaid. We discuss how this contextual information allows for more nuanced analysis. SaNSA allows the user to apply arbitrary attributes, for instance, positional information where nodes are located in a data center. We show how using this information we identify anomalous behavior of one rack of a 1,500 node cluster. We explain the design of SaNSA and then test it on four open compute clusters at LANL. We ingest over 1.1 billion lines of system logs in our study of 190 days in 2018. Using SaNSA, we perform a number of different anomaly detection methods and explain their findings in the context of a production supercomputing data center. For example, we report on instances of misconfigured nodes which receive no scheduled jobs for a period of time as well as examples of correlated rack failures which cause jobs to crash.

Influence of A-Posteriori Subcell Limiting on Fault Frequency in Higher-Order DG Schemes
Anne Reinarz (Technical University Munich)

Soft error rates are increasing as modern architectures require increasingly small features at low voltages. Due to the large number of components used in HPC architectures, these are particularly

vulnerable to soft errors. Hence, when designing applications that run for long time periods on large machines, algorithmic resilience must be taken into account. In this paper we analyse the inherent resiliency of a-posteriori limiting procedures in the context of the explicit ADER DG hyperbolic PDE solver ExaHyPE. The a-posteriori limiter checks element-local high-order DG solutions for physical admissibility, and can thus be expected to also detect hardware-induced errors. Algorithmically, it can be interpreted as element-local checkpointing and restarting of the solver with a more robust finite volume scheme on a fine subgrid. We show that the limiter indeed increases the resilience of the DG algorithm, detecting and correcting particularly those faults which would otherwise lead to a fatal failure.

Room: D168
8:30 am - 12:00 pm

PAW-ATM: Parallel Applications Workshop - Alternatives to MPI

The increasing complexity in heterogeneous and hierarchical parallel architectures and technologies has put a stronger emphasis on the need for more effective parallel programming techniques. Traditional low-level approaches place a greater burden on application developers who must use a mix of distinct programming models (MPI, CUDA, OpenMP, etc.) in order to fully exploit the performance of a particular machine. The lack of a unifying parallel programming model that can fully leverage all the available hardware technologies affects not only the portability and scalability of applications but also the overall productivity of software developers and the maintenance costs of HPC applications. In contrast, high-level parallel programming models have been developed to abstract implementation details away from the programmer, delegating them to the compiler, runtime system, and OS. Such alternatives to traditional MPI+X programming include parallel programming languages (Chapel, Fortran, UPC, Julia), systems for large-scale data processing and analytics (Spark, Tensorflow, Dask), and frameworks and libraries that extend existing languages (Charm++, Unified Parallel C++ (UPC++), Coarray C++, HPX, Legion, Global Arrays). While there are tremendous differences between these approaches, all strive to support better programmer abstractions for concerns such as data parallelism, task parallelism, dynamic load balancing, and data placement across the memory hierarchy.

This workshop will bring together applications experts who will present concrete practical examples of using such alternatives to MPI in order to illustrate the benefits of high-level approaches to scalable programming.

An Application Perspective on Programming Models for the Future *Anshu Dubey (Argonne National Laboratory)*

Development and Performance Comparison of MPI and Fortran Coarrays within an Atmospheric Research Model
Soren C. Rasmussen (Cranfield University)

A mini-application of the Intermediate Complexity Research (ICAR) Model offers an opportunity to compare the costs and performance of the Message Passing Interface (MPI) versus coarray Fortran, two methods of communication across processes. The application requires repeated communication of halo regions, which is performed with either MPI or coarrays. The MPI communication is done using non-blocking two-sided communication, while the coarray library is implemented using a one-sided MPI or OpenSHMEM communication backend. We examine the development cost in addition to strong and weak scalability analysis to understand the performance costs.

Efficient Algorithms for Collective Operations with Notified Communication in Shared Windows
Christian Simmendinger (T-System Solutions for Research), Roman Iakymchuk (KTH Royal Institute of Technology)

Collective operations are commonly used in various parts of scientific applications. Especially in strong scaling scenarios, collective operations can negatively impact the overall applications performance: while the load per rank decreases with increasing core counts, time spent in e.g. barrier operations will increase logarithmically with the core count.

In this article, we develop novel algorithmic solutions for collective operations -- such as Allreduce and Allgather(V) -- by leveraging notified communication in shared windows. To this end, we have developed an extension of GASPI which enables all ranks participating in a shared window to observe the entire notified communication targeted at the window. By exploring benefits of this extension, we deliver high performing implementations of Allreduce and Allgather(V) on Intel and Cray clusters. These implementations clearly achieve 2x-4x performance improvements compared to the best performing MPI implementations for various data

Comparison of the HPC and Big Data Java Libraries Spark, PCJ and APGAS
Jonas Posner (University of Kassel)

Although Java is rarely used in HPC, there are a few notable libraries. Use of Java may help to bridge the gap between HPC and big data processing.

This paper compares the big data library Spark, and the HPC libraries PCJ and APGAS, regarding productivity and performance. We refer to Java versions for all libraries. For APGAS, we include both the original version and an own extension by locality-flexible tasks. We consider three benchmarks: Calculation of pi from HPC, Unbalanced Tree Search (UTS) from HPC, and WordCount from the big data domain.

In performance measurements with up to 144~workers, the extended APGAS library was the clear winner. With 144 workers, APGAS programs were up to a factor of more than two faster than Spark programs, and up to about 30% faster than PCJ programs. Regarding productivity, the extended APGAS programs consistently needed the lowest number of different library constructs. Spark ranged second in productivity, and PCJ third.

Computational Cosmology and Astrophysics on Adaptive Meshes Using Charm++
James Bordner (San Diego Supercomputer Center; University of California, San Diego)

Astrophysical and cosmological phenomena involve a large variety of physical processes, and can encompass an enormous range of scales. To effectively investigate these phenomena computationally, applications must similarly support modeling these phenomena on enormous ranges of scales; furthermore, they must do so efficiently on high-performance computing platforms of ever increasing parallelism and complexity. We describe Enzo-P, a Petascale redesign of the ENZO adaptive mesh refinement astrophysics and cosmology application, along with Cello, a reusable and scalable adaptive mesh refinement framework, on which Enzo-P is based. Cello's scalability is enabled by the Charm++ Parallel Programming System, whose data-driven asynchronous execution model is ideal for taking advantage of the available parallelism in adaptive mesh refinement-based applications. We present weak scaling results on the NSF Blue Waters supercomputer, and outline our future plans to bring Enzo-P to the exascale era by targeting highly-heterogeneous accelerator-based platforms.

PAW-ATM – Workshop Morning Break

GASNet-EX Performance Improvements Due to Specialization for the Cray Aries Network
Paul H. Hargrove (Lawrence Berkeley National Laboratory)

GASNet-EX is a portable, open-source, high-performance communication library designed to efficiently support the networking requirements of PGAS runtime systems and other alternative models in future exascale machines. This paper reports on the improvements in performance observed on Cray XC-series systems due to enhancements made to the GASNet-EX software. These enhancements, known as "specializations", primarily consist of replacing network-independent implementations of several recently added features with implementations tailored to the Cray Aries network. Performance gains from specialization include (1) Negotiated-Payload Active Messages improve bandwidth of a ping-pong test by up to 14%, (2) Immediate Operations reduce running time of a synthetic benchmark by up to 93%, (3) non-bulk RMA Put bandwidth is increased by up to 32%, (4) Remote Atomic performance is 70% faster than the reference on a point-to-point test and allows a hot-spot test to scale robustly, and (5) non-contiguous RMA interfaces see up to 8.6x speedups for an intra-node benchmark and 26% for inter-node. These improvements are available in the GASNet-EX 2018.3.0 release.

Chapel Aggregation Library (CAL)
Louis Jenkins (Pacific Northwest National Laboratory)

Fine-grained communication is a fundamental principle of the Partitioned Global Address Space (PGAS), which serves to simplify creating and reasoning about programs in the distributed context. However, per-message overheads of communication rapidly accumulate in programs that generate a high volume of small messages, limiting the effective bandwidth and potentially increasing latency if the messages are generated at a much higher rate than the effective network bandwidth. One way to reduce such fine-grained communication is by coarsening the granularity by aggregating data, or by buffering the smaller communications together in a way that they can be dispatched in bulk. Once these communications are buffered, an additional optimization called coalescing can be performed to make processing of the data more efficient for the receiver by combining multiple units of data. The Chapel Aggregation Library (CAL) provides a straightforward approach to handling both aggregation and coalescing of data in Chapel and aims to be as generic and minimal as possible to maximize code

reuse and minimize its increase in complexity on user applications. CAL provides a high-performance, distributed, and parallel-safe solution that is entirely written as a Chapel module. In addition to being easy to use, CAL has been shown to improve performance of some benchmarks by one to two orders of magnitude over naive implementations at 32 compute-nodes on a Cray XC50.

Semi-Static and Dynamic Load Balancing for Asynchronous Hurricane Storm Surge Simulations
Maximilian H. Bremer (University of Texas at Austin)

The performance of hurricane storm surge simulations is critical to forecast and mitigate the deadly effects of hurricane landfall. Supercomputers play a key role to run these simulations quickly; however, disruptive changes in future computer architectures will require adapting simulators to maintain high performance, such as increasing asynchrony and improving load balance.

We introduce two new multi-constraint, fully asynchronous load balancers and a new discrete-event simulator (DGSim) that is able to natively model the execution of task-based hurricane simulations based on efficient one-sided, active message-based communication protocols. We calibrate and validate DGSim using it to compare the algorithms' load balancing capabilities and task migration costs under many parameterizations, saving over 5,000x core-hours compared to running the application code directly. Our load balancing algorithms achieve a performance improvement of up to 56 percent over the original static balancer and up to 97 percent of the optimal speed-up.

Distributed L-Shaped Algorithms in Julia
Martin Biel (KTH Royal Institute of Technology)

We present L-Shaped Solvers, a suite of scalable stochastic programming solvers implemented in the Julia programming language. The solvers, which are based on the L-shaped algorithm, run efficiently in parallel, exploit problem structure, and operate on distributed data. The implementation introduces several flexible high-level abstractions that result in a modular design and simplify the development of algorithm variants. In addition, we demonstrate how the abstractions available in the Julia module for distributed computing are exploited to simplify the implementation of the parallel algorithms. The performance of the solvers is evaluated on large-scale problems for finding optimal orders on the Nordic day-ahead electricity market. With 16 worker cores, the fastest algorithm solves a distributed problem with 2.5 million variables and 1.5 million linear constraints about 19 times faster than Gurobi is able to solve the extended form directly.

Panel Discussion *Benjamin Robbins (Cray Inc), Abhinav Bhatele (Lawrence Livermore National Laboratory), Bradford L. Chamberlain (Cray Inc), Anshu Dubey (Argonne National Laboratory), Salvatore Filippone (Cranfield University), Kimberly Keeton (Hewlett Packard Enterprise)*

Room: D171
8:30 am - 12:00 pm

Computational Phenomics @Scale: From Supercomputers to Bedside

Our workshop will bring together various stakeholders from academia, medical centers, industry, federal agencies, and national laboratories to present state-of-the-art phenotypic technologies and their computational and data challenges. We will engage in discussions on the problems and solutions encountered in the translational practice of deep phenotyping at the intersection of computational and data science to support and improve healthcare delivery. The workshop will include presentations on large-scale phenotypic technologies such as natural language processing, multi-scale image analytics, etc. as well as the emergence of advanced machine learning in this space. During the panel discussion, the panelists will address clinical practice challenges such as integrative phenomic data analysis, the role of causal inference in genomic-phenomic associations, and the importance of reproducibility of phenotypic discoveries.

Precision and Personalized Medicine: The Time Has Arrived *Aristides Patrinos (Novim Group)*

The Role of Computing in Predictive and Precision Oncology *Warren Kibbe (Duke University School of Medicine)*

Complex Phenomics in the MVP *J. Michael Gaziano (Harvard Medical School)*

Computational Phenomics @Scale – Workshop Morning Break

Population Genetics and Computation in the Area of Precision Medicine *Eimear Kenny (Icahn School of Medicine at Mount Sinai)*

The Impact of Deep Learning and Artificial Intelligence in Radiology *Ronald Summers (National Institutes of Health)*

AI-Enabled Disease Phenotyping: Opportunities and Computational Challenges *Georgia Tourassi (Oak Ridge National Laboratory)*

Panel Discussion on Currents Trends, Needs, and Bottlenecks in Computational Human Phenomics
Patricia Kovatch (Icahn School of Medicine at Mount Sinai)

Room: D173
8:30 am - 12:00 pm

7th Workshop on Extreme-Scale Programming Tools (ESPT)

The path to extreme computing keeps broadening: large scale systems toward exascale and beyond, growing many core systems with deep memory hierarchies and massively parallel accelerators are just a few of the platforms we can expect. This trend will challenge HPC application developers in their quest to achieve the maximum potential that their systems have to offer, both on and across nodes. Factors such as limited power budgets, heterogeneity, hierarchical memories, shrinking I/O bandwidths, and performance variability will make it increasingly difficult to create productive applications on future platforms. To address these challenges, we need tools for debugging, performance measurement and analysis, and tuning to overcome the architectural, system, and programming complexities expected in these environments.

At the same time, research and development progress for HPC tools themselves faces equally difficult challenges: adaptive systems with an increased emphasis on autotuning, dynamic monitoring and adaptation, heterogeneous analysis and new metrics such as power, energy and temperature require new methodologies, techniques, and engagement with application teams. This workshop will serve as a forum for HPC application developers, system designers and tool researchers to discuss the requirements for tools assisting developers in identifying, investigating and handling the challenges in future extreme scale environments, both for highly parallel nodes and in large-scale HPC systems.

The workshop is the seventh in a series of SC conference workshops organized by the Virtual Institute - High Productivity Supercomputing (VI-HPS), an international initiative of HPC researchers and developers focused on programming and performance tools for parallel systems.

Understanding Software Sustainability: Learning from Parsl and Other Projects *Daniel S. Katz (National Center for Supercomputing Applications, University of Illinois)*

Understanding the Scalability of Molecular Simulation Using Empirical Performance Modeling
Sergei Shudler (Technical University Darmstadt)

Molecular dynamics (MD) simulation allows for the study of static and dynamic properties of molecular ensembles at various molecular scales, from monatomics to macromolecules such as proteins and nucleic acids. It has applications in biology, materials science, biochemistry, and biophysics. Recent developments in simulation techniques spurred the emergence of the computational molecular engineering (CME) field, which focuses specifically on the needs of industrial users in engineering. Within CME, the simulation code ms2 allows users to calculate thermodynamic properties of bulk fluids. It is a parallel code that aims to scale the temporal range of the simulation while keeping the execution time minimal. In this paper, we use empirical performance modeling to study the impact of simulation parameters on the execution time. Our approach is a systematic workflow that can be used as a blue-print in other fields that aim to scale their simulation codes. We show that the generated models can help users better understand how to scale the simulation with minimal increase in execution time.

Advanced Event Sampling Support for PAPI
Forrest Smith (University of Maine)

The PAPI performance library is a widely used tool for gathering performance data from running applications. Modern processors support advanced sampling interfaces, such as Intel's Precise Event Based Sampling (PEBS) and AMD's Instruction Based Sampling (IBS). The current PAPI sampling interface predates the existence of these interfaces and only provides simple instruction-pointer based samples.

We propose a new, improved, sampling interface that provides support for the extended sampling information available on modern hardware. We extend the PAPI interface to add a new PAPI_sample_init call that uses the Linux perf_event interface to access the extra sampled information. A pointer to these samples is returned to the user, who can either decode them on the fly, or write them to disk for later analysis.

By providing extended sampling information, this new PAPI interface allows advanced performance analysis and optimization that was previously not possible. This will enhance the ability to optimize software in modern extreme-scale programming environments.

ESPT – Workshop Morning Break

PARLOT: Efficient Whole-Program Call Tracing for HPC Applications
Saeed Taheri (University of Utah)

The complexity of HPC software and hardware is quickly increasing. As a consequence, the need for efficient execution tracing to gain insight into HPC application behavior is steadily growing. Unfortunately, available tools either do not produce traces with enough detail or incur large overheads. An efficient tracing method that overcomes the tradeoff between maximum information and minimum overhead is therefore urgently needed. This paper presents such a method and tool, called ParLoT, with the following key features. (1) It describes a technique that makes low-overhead on-the-fly compression of whole-program call traces feasible. (2) It presents a new, highly efficient, incremental trace-compression approach that reduces the trace volume dynamically, which lowers not only the needed bandwidth but also the tracing overhead. (3) It collects all caller/callee relations, call frequencies, call stacks, as well as the full trace of all calls and returns executed by each thread, including in library code. (4) It works on top of existing dynamic binary instrumentation tools, thus requiring neither source-code modifications nor recompilation. (5) It supports program analysis and debugging at the thread, thread-group, and program level.

This paper establishes that comparable capabilities are currently unavailable. Our experiments with the NAS parallel benchmarks running on the Comet supercomputer with up to 1,024 cores show that ParLoT can collect whole-program function-call traces at an average tracing bandwidth of just 56 kB/s per core.

Gotcha: A Function-Wrapping Interface for HPC Tools
David Poliakoff (Lawrence Livermore National Laboratory)

This paper introduces Gotcha, a function wrapping interface and library for HPC tools. Many HPC tools, and performance analysis tools in particular, rely on function wrapping to integrate with applications. But existing mechanisms, such as LD_PRELOAD on Linux, have limitations that lead to tool instability and complexity. Gotcha addresses the limitations in existing mechanisms, provides a programmable interface for HPC tools to manage function wrapping, and supports function wrapping across multiple tools. In addition, this paper introduces the idea of interface-independent function wrapping, which makes it possible for tools to wrap arbitrary application functions.

HPC Software Infrastructures at German Aerospace Center *Achim Basermann (German Aerospace Center)*

ESPT Closing Remarks *Felix Wolf (Technical University Darmstadt)*