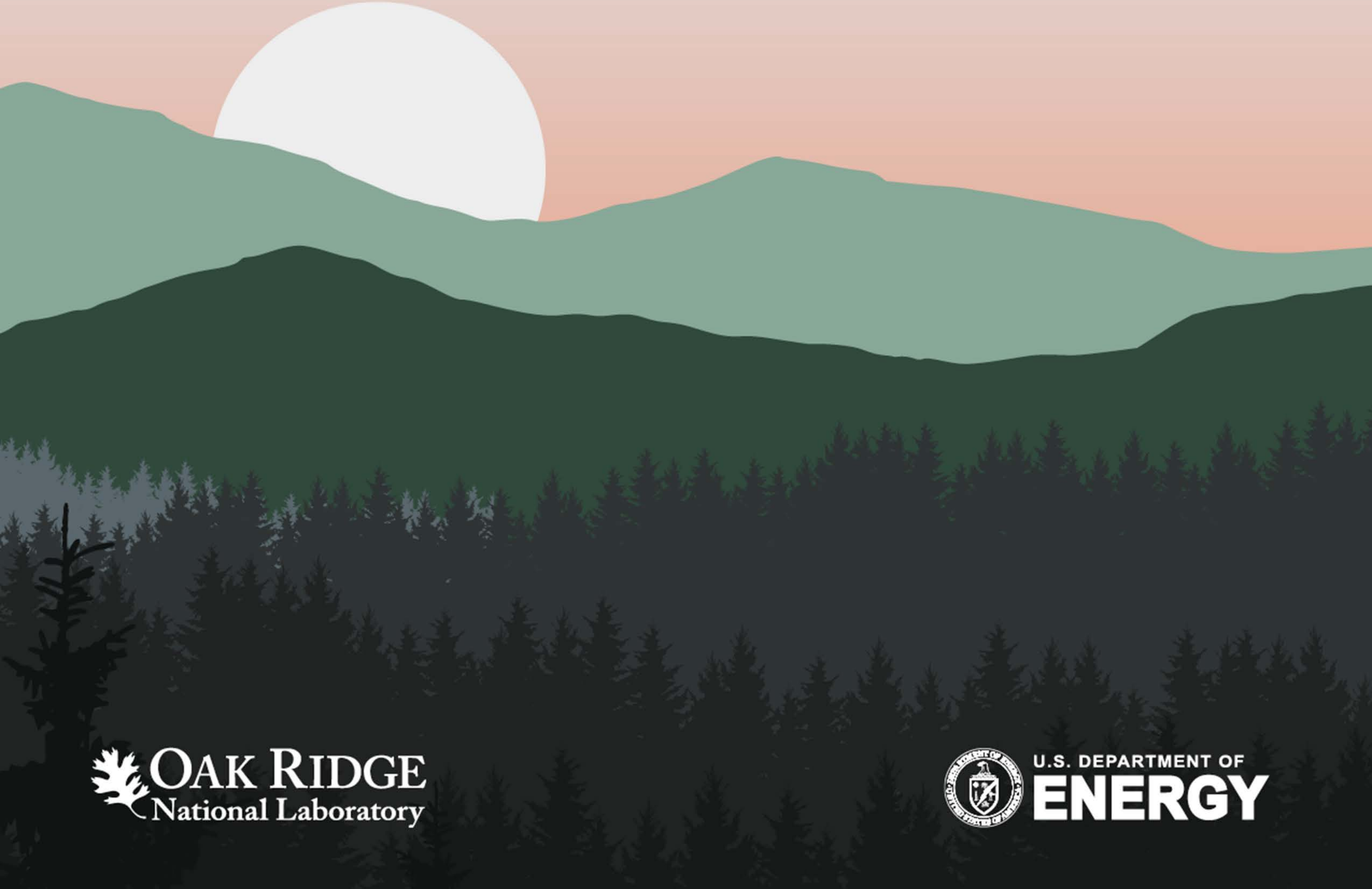
SMOKY MOUNTAINS

Computational Sciences and Engineering Conference

2018



OAK RIDGE
National Laboratory

U.S. DEPARTMENT OF
ENERGY

Contents

A great conference for a great year!

2018 was a remarkable year for computing at Oak Ridge National Laboratory.

Our Leadership Computing Facility celebrated its 25th anniversary and launched Summit, an IBM system that is currently the world's fastest and most AI-compatible computer, marking the third time that ORNL has claimed the number one spot on the coveted TOP500 list; we acquired a vendor for our first exascale machine, Frontier, to be delivered in 2021; and we once again hosted the best and brightest from the computing and data worlds at the 2018 Smoky Mountains Conference in Gatlinburg, Tennessee.

2018's theme, "the integration of experiment, data analytics, and modeling and simulation into instruments for discoveries in science and engineering," proved a fertile one for discussion as our esteemed lineup of speakers brought us up to speed on the states of the art in both computing and big data.

With more than 180 people registered, 15 great sponsors, and more Data Challenge participants than we have ever had, 2018 marked the biggest SMC on record. The conference has come a long way since its humble beginnings in 2003, when 45 participants descended on Fall Creek Falls in Pikeville, Tennessee.

And as we approach the exascale and technologies such as artificial intelligence continue to revolutionize our understanding of data, I strongly suspect that the SMC will only continue to grow. I hope you'll come along for the ride.

I would like to extend a very special thanks to Joe Citeno of GE Power, who delivered an extraordinary keynote and set the stage for an entertaining and productive three days.

And of course, none of this would be possible without the support of our generous sponsors, who have once again ensured that we can come together in our shared mission to solve some of the most complex problems in energy science and computing.

To those who joined us in Gatlinburg in August, we hope you had an enjoyable and informative experience. And to those who didn't, we hope to see you in 2019 for what is sure to be another outstanding conference.

– Jeff Nichols, Associate Laboratory Director for Computing and Computational Sciences, Oak Ridge National Laboratory

Thanks to our sponsors

DELL

AMD

intel®

CRAY

Micron®

NVIDIA®

arm

IBM®

D:WAVE
The Quantum Computing Company™

PENGUIN
COMPUTING

DataDirect™
NETWORKS

PGI®

Hewlett Packard
Enterprise

Google Cloud

Mellanox®
TECHNOLOGIES

A conference with a history

Believe it or not, the Smoky Mountains Conference is almost old enough to drive.

2018 represented the fifteenth time ORNL has hosted what has come to be known as the laboratory's foremost conference on computing and data. And much like a child, the conference has grown by the year to the point that its unrecognizable from its humble beginnings.

These first invitation-only conferences, originally held in Fall Creek Falls State Park in Pikeville, Tennessee, just west of ORNL, focused primarily on particular science domains – the inaugural 2003 conference, for instance, highlighted climate and materials research.

"A lot of people don't realize this," said Jeff Nichols, an original attendee and now the associate laboratory director for computing and computational sciences at ORNL, "but the first SMC predates ORNL's Leadership Computing Facility. Back then it was called 'Fall Creek Falls,' and it was pretty rough. There was no cellphone or wireless service and the 'country-style' food wasn't exactly a hit with the California crowd."

It wasn't long before the conference outgrew the Pikeville accommodations, and Nichols and fellow organizers decided to take their show on the road and allow future attendees to see more of the great state of Tennessee. From the Opryland Hotel in Nashville in 2007 to Montgomery Bell State Park in 2008 to the Chattanoogan Hotel in 2009 to the Memphis Health Science Center in 2010, the conference roamed a bit before settling down in its latest location of Gatlinburg, where it's been held since 2011.

"We were sort of looking for a home," said Nichols. But the conference's zigzagging path to Gatlinburg was also a time of great growth, as it expanded to include ever more science domains and encompass all four

discovery paradigms, namely theory, experiment, modeling and simulation, and data.

Such broad scope has required consistent collaboration across an international community from academia, industry, and the national labs, an evolution that has dovetailed nicely with ORNL's strengths in accelerated node computing and integrated instruments for science.

"Creating instruments for science has long been a primary theme of the conference," said Nichols. "And with the emergence of big data, this has increasingly come to include industry. By mingling with computing and domain experts, our industry partners can really move the needle, so the conference pays off especially well for them."

But everyone seems to get something from the conference. Approximately 45 people attended the initial conference; 2018, on the other hand, saw more than 180 speakers and attendees, a measure of growth that corresponds with the increasing interests in computing and data as a means to solve some of society's greatest scientific challenges.

With the launch of Summit and the rapid growth in AI and quantum information, 2019 may well prove to be the biggest SMC yet.

"It's just been amazing to watch it grow, and to witness the connections across the different landscapes," said Nichols. "It really is a one-of-a kind conference."

Joe Citeno,
General Electric Power

Keynote

In 2017, US-based General Electric delivered its newest heavy-duty gas turbine, the 7HA.02, to two power plants in Texas. The installations marked a milestone in natural gas–derived electricity generation, setting new marks in efficiency and emissions for utility-scale turbomachinery.

A key ingredient in GE's successful recipe for this breakthrough technology was the ability to conduct critical modeling and simulation on the Titan supercomputer, a leadership-class system managed by the Oak Ridge Leadership Computing Facility, a US Department of Energy Office of Science User Facility at DOE's Oak Ridge National Laboratory.

In 2015, GE researchers—in collaboration with software company Cascade Technologies—used Titan to investigate an unexpected combustion instability that had been revealed during tests of a GE gas turbine 1 year earlier. Although GE had determined that instability would not affect gas turbine performance, the company wanted to understand its cause and whether it would manifest in a new gas turbine design.

A concern was that the instability might be excited and increase amplitude with axial fuel staging. Developed with DOE support, AFS enabled turbine operation at higher temperatures with lower emissions. It accomplished this by introducing a second fuel injection stage into the turbine's combustion system at a point later than the current industry standard.

"To help meet our deadlines, we wanted to see if we could reproduce the instability numerically," said Joe Citeno, GE Power combustion engineering manager. "But we didn't have a predictive model or the internal computational horsepower needed to run it."

The Cascade team, in partnership with GE, created a groundbreaking high-fidelity model on Titan to examine combustion physics at unprecedented resolution. The team successfully reproduced the instability numerically on Titan and then predicted it would not affect performance in the new 7HA.02 gas turbine design. Physical experiments later confirmed the prediction, validating that the combined GE–

Cascade team had created a powerful advanced modeling approach that can predict a range of combustion performance variables for GE's gas turbine combustor designs, including air-fuel mixing, flame stability, emissions, and exit temperatures.

"The availability of the Titan supercomputer enabled a breakthrough that otherwise would have remained unrealized," Citeno said. The new 7HA.02 gas turbine was put into production on schedule, and the first systems were successfully installed in Texas in the spring of 2017.

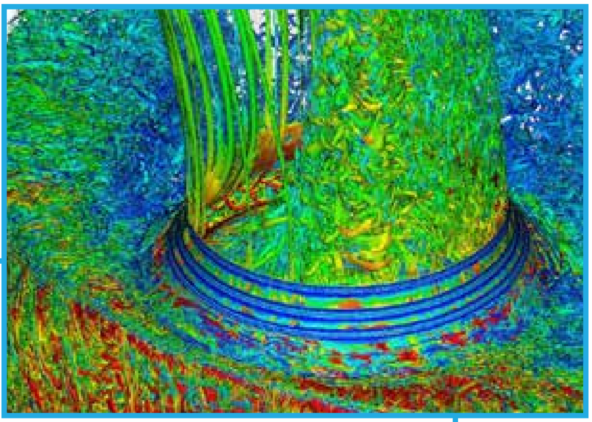
Now, GE is busy fulfilling orders for this new turbine, which can achieve an overall efficiency of more than 62 percent in a combined-cycle plant; it is projected to exceed world-record efficiency marks set by GE's 9HA turbine model. "For our customers, that means lower fuel consumption and reduced cost per megawatt," Citeno said.

The 7HA.02 installations are more than 2 percent higher than previous 7F-class gas turbines in combined-cycle efficiency. This improvement may seem small, but even a small change can produce an enormous impact. For example, just a 1 percent efficiency gain applied across the entire US combined-cycle fleet would save an estimated $11 billion or more in fuel over the next 20 years.

At the Texas plants, the new turbines can respond quickly to fluctuating power demands, filling a critical role in the state's energy portfolio. During periods of high demand, the plants can provide power for more than 2 million homes.

Building on its simulation success and still working with Cascade, GE continues to use advanced software and its own high-performance computers to evaluate new product designs and test the limits of modern engineering—meaning high-performance computing will likely play a major role in the next leap forward in combustion systems.

"It's enabling us to be more confident in the moves we make to increase efficiency and product flexibility," Citeno said.



SESSION 1:
ENABLING TECHNOLOGIES AND SCALABLE APPLICATIONS

Moderators: Stuart Slattery and Judy Hill, Oak Ridge National Laboratory

The purpose of this session was to expose the relationship between physics codes and libraries designed for exascale computing. The DOE Exascale Computing Project is investing heavily in scientific applications and their associated enabling technologies, such as libraries. The relationships between applications and libraries can be complex, however, because they can and often do interact in numerous ways. The inaugural session represented three ways in which libraries and applications interact:

- General libraries provide numerical capabilities for HPC through ECP.
- ECP applications use external libraries extensively to achieve functionality and performance.
- ECP applications have limited production use of external libraries.

The moderators further asked that participants consider the following questions throughout the session:

- Are enabling technologies a viable path for application performance at exascale?
- What challenges are posed to library developers by supporting multiple application stakeholders?
- What challenges are posed to application developers by using one or many enabling technologies?
- What do library and application developers need from each other to make an impact at exascale?
- How will the diversity of potential exascale hardware platforms affect the future interaction between libraries and applications?

AMReX: an overview

Ann Almgren, Lawrence Berkeley National Laboratory

AMReX is an ECP co-design center with a mandate to support the development of adaptive mesh refinement applications, evaluate new software technologies, and interact with hardware vendors. The AMReX software framework, whose development has a long history starting in the DOE mathematics program, is being actively used by five of the ECP application projects (accelerator design, astrophysics, combustion, cosmology, and multiphase flow) and numerous non-ECP projects. The AMReX framework is designed to support the agile development of new algorithms and the performance at scale of more established codes. It supports multilevel mesh, particle, and particle-mesh operations with a variety of strategies to enable hybrid parallelism, effective load balancing, and asynchronous task parallelism.

Almgren stressed that enabling technologies such as AMReX and other well-designed software components are important, and in fact necessary, components of application performance at the exascale. She further stressed that library and application developers need to foster "communication and trust" in order to be able to leverage the progress made by others. The era of perpetually reinventing the "numerical wheel" needs to come to an end. Almgren also embraced the open-source model for software development; all branches of AMReX and the entire AMReX development history are publicly available on GitHub.

CEED: Center for Efficient Exascale Discretizations

Stan Tomov, The University of Tennessee–Knoxville

The Center for Efficient Exascale Discretizations works to develop software that maximizes data movement through exascale systems and minimizes computation time, facilitating efficient simulations for wide-ranging science applications. CEED is a collaboration of more than 30 researchers at two national laboratories and five universities. Its R&D efforts address the needs of users in government and industry and commercial developers, manufacturers, and vendors of emerging exascale hardware and software technologies. The research group aims to bolster the use of HPC in fields as diverse as wind energy, additive manufacturing, urban systems, subsurface geology, nuclear reactors, and fusion. Stan Tomov, a research assistant professor at the University of Tennessee–Knoxville, said CEED supports ECP applications by developing simulation code, discretization libraries, miniapps, and standards as well as designing new architectures that harness arithmetic intensity more effectively and improve performance, portability, and scalability.

Discretization involves applying finite element methods to translate continuous data, which can have an unlimited number of possible values, into discrete integers to ready them for computation. Parallelism drives HPC by analyzing small packets of data simultaneously, speeding up computation time. CEED develops kernels—what CEED calls "ceedlings"—that manage memory and CPU operations during simulations and other forms of data analysis. Tomov said that working closely with stakeholders eases deployment of the center's simulation tools and ensures that they are relevant to end users.

The group released CEED 1.0, its first software distribution, this year. In his presentation, Tomov highlighted efforts to improve existing HPC products and the anticipated release of CEED 2.0 during the group's next two program years.

"CEED is co-designing high-order discretization algorithms, software technologies, and optimized discretization libraries to ensure that a wide variety of ECP applications run efficiently on exascale hardware," said Tomov. "The CEED developments can increase performance by orders of magnitude over traditional methods."

Building a scalable plasma physics capability from components

Roger Pawlowski, Sandia National Laboratories

Sandia's Roger Pawlowski recounted his team's effort to construct a hybrid fusion code that included the five-moment plasma system model (including variables such as energy and density), with the electromagnetics coupled to particle-in-cell. Fusion researchers seek to solve systems that span multiple time and length scales, which presents special challenges in the code design process.

The team utilized a broad range of tools, including 35 packages from Trilinos. Critical components included Kokkos for performance portability, Sacado for embedded automatic differentiation, Phalanx for assembly, Tempus for time integration, and Teko for handling the blocking scheme for the linear system. The code performed well overall, showing good weak scalability on a number of test problems.

Pawlowski presented several lessons learned from the team's experience:

- Every component dependency is a risk to applications.
- Components really help getting started but can cause problems at the finish.
- Components provide general solutions that improve productivity but are usually not specific enough; exposing low-level building blocks can help.
- Components can make it difficult for new team members to orient themselves with an application code; allowing for simple implementations side-by-side can help.
- Heavy adoption of a component can negatively impact agility.

Multiscale challenges of blade-resolved wind turbine simulations

Michael Sprague, National Renewable Energy Laboratory

ExaWind, an ECP endeavor, is creating predictive simulation capabilities for wind energy. The goal is to enable new understanding of the complex flow dynamics in wind plants and to expose new pathways to optimize wind plant performance. The project is a collaboration of the National Renewable Energy Laboratory (NREL), Sandia National Laboratories, ORNL, University of Texas at Austin, and Parallel Geometric Algorithms, LLC. Researchers use simulations to examine poorly understood turbine–wake interactions, a key element to scaling the technology and implementing active turbine control in industrial-size wind plants. Michael Sprague, a principal scientist at NREL, said efforts to simulate wind plants face challenges due to the huge range of temporal and spatial scales that are interacting. Additionally, these simulations are extremely complex and require multiple moving meshes to capture rotating blades and yawing turbine nacelles that move to track changes in wind speed and direction. The ultimate goal of the ExaWind project is to simulate 100 multimegawatt turbines in a large wind farm with fully resolved flow and turbine dynamics.

Effective turbulence simulations allow wind energy researchers to better predict how turbines interact with wind forces in large-scale settings. ExaWind collaborators have developed open-source application codes based on C/C++ and Fortran programming languages and Trilinos and STK software libraries to target all major HPC platforms on which simulations could be run. The modeling and algorithmic pathways of ExaWind include unstructured-grid finite volume spatial discretization and pressure-projection methods for incompressible flow. Sprague stressed that, while great strides have been accomplished with the ExaWind software stack, time to solution remains too large. Efforts to reduce that time are focused on improving strong scaling of the linear system solver stack, optimizing time-update algorithms, and better enabling of next-generation hardware.

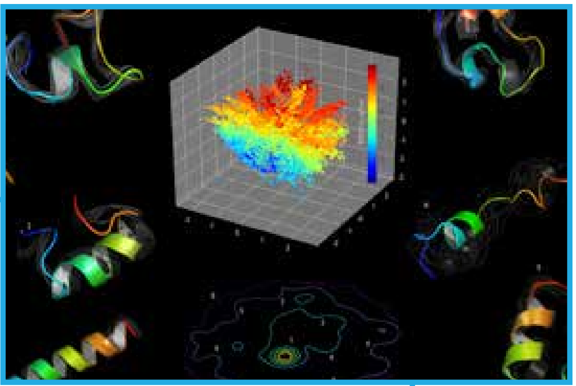
“ExaWind is enabling a predictive wind energy simulation capability that will enable researchers to better predict and understand the complex flows in large wind farms,” said Sprague. “That new understanding will expose new technology pathways and better tools for optimizing turbines and farms.”

Exascale cosmological simulations: the HACC story

Salman Habib, Argonne National Laboratory

Multiwavelength cosmological surveys, from the cosmic microwave background to the optical bands, and observations of large areas of the sky, have revolutionized cosmology. These measurements have driven the need for development of the Hardware/Hybrid Accelerated Cosmology Code, an extreme-scale, particle-based framework for computational cosmology. HACC was designed for maximal performance, achieving 50 percent of the peak possible on all systems, to be fully scalable, to be performance portable, and to run on predeployment machines as a benchmark.

Habib stressed that enabling technologies provide a viable path for application performance at the exascale but cautioned that it could be difficult to support numerous applications with diverse requirements. Possible solutions include lightweight toolkits and well-designed application programming interfaces. He also predicted that combining multiple enabling technologies was bound to be problematic, particularly as architecture evolution remains mysterious and full of risks. To this end Habib's team is currently working on a lightweight development branch for HACC to investigate next-generation architectures.



SESSION 2:
DATA ANALYTIC COMPUTING AND AI: IS BIGGER BETTER?

Moderators: Laura Pullum and David Womble, Oak Ridge National Laboratory

The second session explored how very large and exascale-class computing systems could best be used for data analytic computing and AI while considering whether bigger machines are necessarily better for these applications. The moderators looked both within and outside the national lab system for a variety of opinions on the best use of leading systems. These systems have the ability to collect and store massive amounts of data, and modern architectures increasingly feature components, such as GPUs and reduced-precision tensor cores, that lend themselves well to applications in machine learning and AI.

Session speakers explored the issues surrounding the use of the ever-increasing scale of compute power and how this evolution enables advances in data analytics, machine learning, and AI.

The moderators asked attendees to consider the following questions:

- What have you done, or will you do, on Summit that you could not have done otherwise?
- What can we accomplish on leadership computing that we cannot accomplish on cloud-based computing?
- How should we measure the return on investment for AI on any of these machines?
- How should leadership-class machines be used if only DOE labs have them?
- Should the DOE labs maintain their own software stack, recognizing that it is very expensive to do so and that the field of AI is changing so rapidly?
- Summit has an order of magnitude more compute power than Titan. How much better can you expect your results to be?
- What algorithmic challenges do you expect?
- What are the unique AI challenges in the DOE and other application spaces?
- In the AI space, what would your top investment priority be? In machine/hardware? In software? In research?

Experiences in developing scalable machine learning and artificial intelligence approaches for biological and healt hcare applications

Arvind Ramanathan, Oak Ridge National Laboratory

Arvind Ramanathan pointed out that while DOE has done a good job in the arena of large-scale numerical simulations, the emergence of scalable data analytics, deep learning, and AI are revolutionizing this once-familiar paradigm, and new architectures are allowing for their merger.

Ramanathan recounted his experiences with the CANDLE project, a joint DOE and National Cancer Institute effort that seeks to harness the power of DOE's world-leading supercomputers to tackle the most challenging problems in cancer research. Specifically, Ramanathan and his colleagues took the lead on the project's third pilot, which focuses on using deep learning to extract information from large cancer datasets but which also supports the deep-learning needs of the other pilots. These foci require collaborating with DOE computing centers, HPC vendors, and ECP co-design and software projects. A key goal, he explained, is to improve interaction with ECP and the DOE co-design centers. The team has optimized numerous machine-learning and deep-learning approaches and has scaled across supercomputers, including up to 50 percent on Summit, to see how models perform under unique hyperparameter configurations.

Multiscale simulations pose unique challenges at the exascale (e.g., the need for in situ analytics, the need for faster and more efficient training in deep-learning and AI), and they consume 45 to 60 percent of computing time. Ramanathan elaborated on how best to use AI to drive multiscale simulations. He stressed the need to interleave data analytics and simulation and used protein-folding simulations to illustrate how enabling deep-learning and reinforcement-learning approaches can achieve near-real-time prediction.

Summit cooling intelligence: event-driven machine learning for cooling efficiency

Ryan Quick, Providentia Worldwide

Computers have long been designed to reduce systemic heat with a network of water-cooled pipes. Providentia Worldwide brings the power of machine learning, AI, and data aggregation to efforts to improve Summit's cooling system to provide real-time feedback to the logic circuits controlling water flow to the supercomputer's nodes. Providentia's focus is on the middle of the cooling process—where, Ryan Quick said, machine learning and AI reside—because you have to start a process before you can get information to feed back into the process to make it smarter.

Providentia examines data analysis models that either downplay or emphasize the availability, performance, and governance of data, and numerous kinds of data—command and control, event, aggregated outcome, and monitoring/alerting data—are collected and analyzed by the cooling system. Prometheus, Kafka, and OpenBMC provide the data-processing power behind Summit's "smart cooling" capabilities. Providentia leverages machine learning and AI against the data, so recommendations to Summit's cooling systems result in improved workload management, more efficient cooling, and reduced total cost of ownership. Quick stressed that using machine learning to drive a supercomputer's cooling system is a new R&D area, and there is much to be learned about providing real-time feedback in hyperscale.

Learning-based predictive models: a new approach to integrating large-scale simulations and experiments

Brian Van Essen, Lawrence Livermore National Laboratory

Predictive capabilities can be advanced, according to Brian Van Essen, by challenging our simulations with experimental data, but new techniques are also needed to truly improve prediction. He pointed to the potential of machine learning as a means to improve predictive modeling. According to Van Essen, improved prediction essentially requires improved models, software tools to develop and guide models, and computational platforms to better support predictive tools. For now, however, the latent spaces inherent in machine-learning and deep-learning networks can be crafted to improve predictability.

To truly improve models, researchers need three things:

- massive datasets;
- scalable, unsupervised learning methods; and
- neural network training software that can scale up to meet both the complexity of a model and the size of a dataset.

He described the Livermore Big Artificial Neural Network Toolkit, a project that is optimized for multiple levels of parallelism and that provides in situ learning on large datasets. With his colleagues Van Essen is seeking to generate a billion-sample dataset for Sierra via an inertial confinement fusion simulation. They will share the results with the community.

Van Essen went on to stress that even at very large scales, carefully choosing which simulations to execute is critical for success, and that researchers must always ask whether the simulation is evolving as predicted. He concluded that, in the end, machine learning will ultimately tie simulation and experiment together.

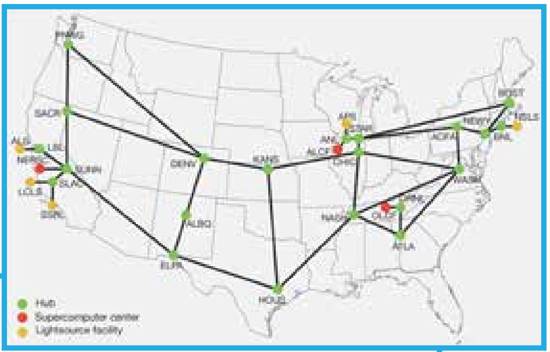
The power of abstraction in computational exploration seismology

Felix Herrmann, Georgia Institute of Technology

Imaging of three-dimensional seismic maps requires extreme compute and input/output capabilities because it is based on wave equations, which are computationally expensive to solve, and because it involves a large number of wavelengths, more than 10,000 time steps, and up to a billion unknowns. Computational abstraction and machine learning are critical to reducing development time and to generating code that minimizes computational costs. To this end, Felix Herrmann and his colleagues employ DEVITO, a just-in-time compiler that generates codes based on input that is specific to the problem being solved.

With DEVITO, abstraction takes place at every level that expresses physics, allowing the user to manage the complexity of very difficult codes. Herrmann also touched on JUDI, an additional abstraction layer in Julia for linear algebra. These tools have helped Herrmann and his colleagues demonstrate the power of abstractions to simplify data- and compute-intensive tasks and to merge ideas from machine learning. To be successful, however, researchers will need access to machines such as Summit for high-fidelity seismic simulations and to scaled-up convolutional neural networks to carry out the computation needed to generate low- to high-fidelity seismic maps.

“The right abstractions hold the key to managing complexity of data and compute-intensive imaging problems merging ideas from CSE & ML,” said Herrmann.



SESSION 3:
ARCHITECTURE OF FEDERATED INSTRUMENTS FOR SCIENCE

Moderators: Barney Maccabe and Nageswara Rao, Oak Ridge National Laboratory

Today, large-scale science is conducted by teams of scientists, using independent instruments at different facilities. To use multiple instruments, these teams must address the challenges inherent in merging independent observations into a coherent, meaningful result. The goal of federating multiple instruments, possibly at different facilities, to create the appearance of a seamlessly integrated environment has been a long-standing goal. Recent advances in "softwarized" infrastructure (including software-defined networks), rapid growth in sensing technologies and the subsequent growth in data, new technologies for reducing data, and continued reduction in computing costs mean that this vision is closer than ever. The moderators asked the speakers and attendees to consider the following questions:

- How do you characterize the computational infrastructure needed to support federated instruments? What kinds of computational capabilities are needed, and how are they distributed?
- Which policies need to be adapted to support federation? What are the minimal changes required, and what are the incremental benefits associated with additional adaptation? (Remember that many applications will not need federation.)
- With regard to challenges in automated management and optimization of a complex, federated infrastructure, how well can we predict utilization and availability of critical resources? How important is it that we do this with a high degree of accuracy and a high degree of confidence?
- Will programmability of the underlying infrastructure enhance or undermine our confidence in our ability to ensure the needed isolation and protection of critical operations and information?

Federating large-scale instruments for accelerated fusion research

C-S Chang, Princeton Plasma Physics Laboratory

Interpreting fusion data from tokamaks requires a high degree of precision and a large-scale organized data-science effort, a necessity especially for ITER, which will be the world's largest tokamak. Millions of magnetic sensors on the device will produce data with all five "big-V" properties: volume, value, velocity, variety, and veracity. To better understand data from ITER and other reactors, Chang said that federating diagnostic instruments and computational resources remotely by hundreds of collaborators is the next logical step. Before they can be analyzed and monitored accurately, the datasets currently produced by tokamak projects must be reduced, a step that adds time and cost.

Chang concluded that shared usage of distributed resources such as leadership-class HPC systems and cloud computing could help accelerate the development of next-generation fusion reactors. Federation, he said, would make researchers better equipped to predict performance, execute workflows, and quantify uncertainties, among other benefits. For example, if the scientists who manage various instruments with deep learning and AI applications were part of a common federated framework, they could help interpret data, plan the next experiment, compare research results, and investigate data inconsistencies in real time. Using automated research tools in combination with HPC systems, cloud-computing resources, or local clusters could help researchers accelerate fusion research, he explained. Additionally, if new ideas could be tested via supercomputer simulations, scientists could more efficiently determine how various phenomena would manifest on ITER and other reactors and could accelerate their development.

Federation, Chang said, would lead to increased trust among researchers, and having their results verified by various methods would produce a higher degree of accuracy and would instill greater confidence in the research. He emphasized that while usability is important for any federated system, security safeguards such as authorizing tiered control of the data and a limited number of programmers would also be essential, as would having a unified system across the national laboratories.

Smart infrastructure, smart science

Steve Oberlin, NVIDIA

Moore's Law is now being outrun by GPU scaling, and datasets generated by the likes of Google and Facebook are increasing by an order of magnitude annually in some cases. These developments mean that AI's role in HPC is likewise increasing.

For example, deep-learning techniques applied in simulations of fusion stability for the ITER tokamak increased the accuracy of the simulations from 80 percent (using CNNs) to 95 percent. In adaptive optics, deep learning is enabling a sharper image than ever thought possible from the world's largest ground-based telescope, and in electromicroscopy, accelerated computing has made possible atomic-level resolution of biological imaging, analysis, and modeling.

Oberlin said that today's data science users expect an interactive user model that features:

- virtualization,
- isolation,
- persistent services and storage,
- elastic resource management,
- data security,
- on-demand "pay by the sip,"
- service-level agreements, and
- abstracted workflows.

He concluded by pointing out that convergence of computational science and data science is an enabler for federated systems, and the infrastructure required to support data and HPC in a single center lays the groundwork for cooperation between centers and how federated instruments should be constructed.

Experiences and recommendations from analyzing extreme-scale scientific workflows

Erich Strohmaier, Lawrence Berkeley National Laboratory

The Extreme Scale Scientific Workflow Analysis and Prediction project aims to understand the performance, behavior, and requirements of scientific workflows from various fields. According to Strohmaier, working with scientists and observing the software and hardware capabilities of different research facilities are key steps toward the eventual development of "superfacilities" optimized to improve scientific discovery and elevate data analysis techniques.

The X-SWAP project, he explained, continuously gathers and analyzes data to thoroughly understand how a workflow functions from the beginning to the end of any given research endeavor. The project team has studied workflows involving light sources, astrophysics, and genomics; identified inefficiencies; and designed and simulated new workflows that address those challenges. Overcoming bottlenecks and improving scientific workflows could have a significant impact on vital performance factors such as speed and predictability, he added.

Strohmaier also noted that simplified workflows, not complex models, lead to a better understanding of scientific concepts. Eventually, the X-SWAP team anticipates that superfacilities will be the ideal platforms for implementing these new workflows to advance data analysis techniques and monitor research projects from beginning to end, improvements that may prove to be valuable as HPC continues to grow more diverse.

System software issues in federating instruments for science

Raj Kettimuthu, Argonne National Laboratory

Today's most sophisticated science instruments are producing more data than ever before, and researchers are increasingly seeking real-time analysis of experimental data to enable "smart" experimentation. Although the instruments possess local compute resources, those resources cannot keep up with modern datasets. The analysis codes must be scaled up to take advantage of HPC resources. Executing a code once it has been scaled up will almost certainly require special arrangements. For example, HPC centers will be able to take advantage of maintenance and shutdown periods and share resources with other workflows to accommodate these users.

In studying the implications of supporting real-time jobs on big machines (in this case, the Mira supercomputer at Argonne), Kettimuthu and his colleagues observed that some batch jobs, especially the ones that run for a short period of time and/or use a small number of nodes, benefitted. Furthermore, they found that:

- users who submit real-time jobs might be willing to compromise in one or more dimensions (e.g., size, duration), and
- many jobs (both batch and real-time) are moldable (i.e., allocation can be adjusted at job start time).

Some of these accommodations can be made up for in pricing, said Kettimuthu. For instance, real-time jobs should be charged at a higher rate, and mechanisms are needed to compensate batch jobs.

"Changing the status quo is always hard. Socializing the idea of real-time jobs preempting batch jobs and appropriate compensation for batch jobs with the users is very important," said Kettimuthu. "Enabling users to set limits on delays they are willing to tolerate subject to a fair compensation will be helpful. Providing accurate pricing estimates for real-time jobs at submission time will be a challenge."

The challenge of data heterogeneity

Edmon Begoli, Oak Ridge National Laboratory

Because data in computer science, applied engineering research, and related science domains continue to become more diverse and complex, Begoli discussed the potential of some approaches for dealing with the increasing diversity and influx of information. Although members of the data management community have been managing heterogeneous data for many years, it has become apparent that data heterogeneity has become one of the most pressing issues. Begoli also contested the established notion that a significant amount of time in data science must be devoted to routine tasks such as cleaning data. It is a symptom of a deeper problem, and a call for action. Identifying and solving those problems is a challenge to be addressed, and solving this challenge through more effective and sophisticated data management could lead to more productive data analytics processes for many data-intensive organizations.

Finding solutions to these challenges is not simple, but necessary. The challenges that need to be addressed are related to comprehensive record linkage, effective data quality, and scalable methods for managing heterogeneous data, both in type, volume, and structure. In addition, modern data management requires new data production workflows, involving data engineers and data scientists in both production of the data products and their analyses.

Begoli supported these views with case studies rooted in ORNL's work with the Department of Veterans Affairs, where almost all of the observed challenges are present. Begoli presented his team's work that involved scalable methods for data cleaning and information extraction, and data production workflows that connect data science and data engineering teams, aimed at creation of datasets for research from both structured and unstructured medical data.

Mathematical modeling, design, and optimal control of integrated infrastructure

Sven Leyffer, Argonne National Laboratory

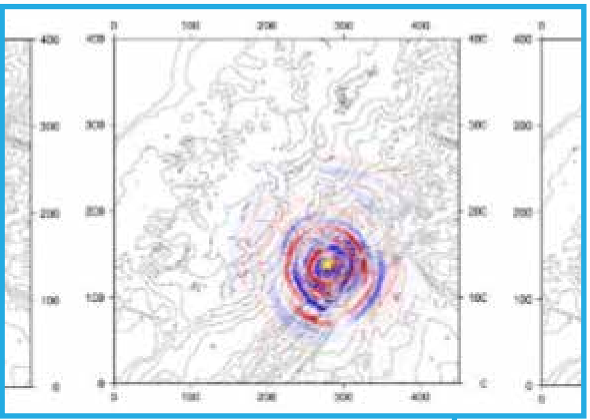
The speed of data acquisition and the size of datasets at light source facilities are outstripping Moore's Law, and at some point in the near future will almost certainly overwhelm our available computing capacity. In fact, projections show a tenfold increase in data rates and sizes over the next four years. This data tsunami, when coupled with heterogeneous experiments and novel experimental designs, presents an enormous challenge for researchers, but it also presents an excellent opportunity for machine learning and the design, control, and operation of a federal HPC superfacility.

Leyffer described a recent thought exercise in which he and his colleagues built a model and simulation tool for the planning of an eventual integrated cyber infrastructure with the goal of avoiding overprovisioning and bottlenecks.

The team looked at data from Argonne's Advanced Photon Source, Lawrence Berkeley's Advanced Light Source, Brookhaven's National Synchrotron Light Source, SLAC's Linac Coherent Light Source, and the Stanford Synchotron Radiation Light Source and measured and/or estimated resource requirements.

They found that achieving utilization above 34 percent is impossible unless users accept a slowdown, and they urged future studies to incorporate uncertainty for robust design and scheduling; explore design and scheduling alternatives, including incentives for slowdowns; and investigate where to build new links or locate facilities.

"The design, management, and operation of a DOE superfacility poses a range of challenging and interesting engineering, scientific, computer science, and mathematical problems," said Leyffer. "I am excited to be working on a problem with such broad scope and fundamental implications for the future of science within the DOE."



SESSION 4:
PRE-EXASCALE ARCHITECTURES AND EARLY RESULTS

Moderators: Jack Wells and Tjerk Straatsma, Oak Ridge National Laboratory

Session 4 presented perspectives from HPC vendors and centers on selected pre-exascale systems, namely new systems with architectures possessing promise to scale to the exascale and for which early performance results were available to discuss. The session was divided between vendor introductions of selected architectures and the experiences of HPC facilities in deploying those systems. The retrospective wasn't intended to be comprehensive, but rather to serve as reference for the path forward toward an eventual exascale system(s). The moderators, OLCF Director of Science Jack Wells and ORNL Senior Scientist and Director of the OLCF's Center for Accelerated Application Readiness Tjerk Straatsma, asked attendees to consider the following questions:

- What have we learned from these pre-exascale projects that will be important for the co-design of exascale supercomputers and applications?
- What are best practices to improve software readiness early in the life of new supercomputers?
- How is the convergence of mod/sim, data analytics, and AI/machine learning being realized on pre-exascale HPC systems?
- Are we on track to meet our energy-efficiency goals?

Architecture for the Sierra and Summit computers

James Sexton, IBM

The Sierra and Summit supercomputers, while wholly different systems, feature numerous similarities that enable them to accomplish world-class science. Both are designed to accommodate AI and feature:

- system scalabilty,
- heterogeneous compute and memory elements,
- enhanced Mellanox interconnect,
- an open-source compiler, and
- water- and air-cooling capabilities.

Sexton elaborated on the finer points of programming the hybrid CPU–GPU architecture, pointing out that managing memory, defining the execution space, and managing the data were are all critical to success, as were internode and intranode communications. Fortunately, OpenMP addresses most of these requirements.

The systems are able to handle integrated applications that involve modeling and simulation, machine learning, and steering, and they have performed better than expected. Sexton concluded by pointing out that both systems were very close to acceptance, are exceeding committed performance, and are emerging as integrated HPC, AI, and analytics platforms.

Early application results on Summit

Tjerk Straatsma, Oak Ridge National Laboratory

ORNL launched Summit, the world's most powerful and smartest scientific supercomputer, in June 2018. Located at ORNL's Oak Ridge Leadership Computing Facility, the new system will enable research teams using computational methods to gain insights into major scientific challenges in astrophysics, chemistry, fusion, biophysics, nuclear physics, materials science, combustion, and many other domain sciences.

The OLCF and the IBM/NVIDIA Center of Excellence are partnering with developers to prepare scientific codes for Summit through the Center for Accelerated Application Readiness. These teams, as well as a number of other researchers with Summit-ready scientific applications, will receive access to Summit through the Early Science Program. This program, Straatsma said, will demonstrate the system's capabilities with science at scale, be used to further harden the system's hardware and software stack, and ensure that the facility is fully prepared when time on Summit becomes available to the user programs. Because programming for GPUs and optimizing a nonaccelerated code can be an intensive, long-term commitment, Straatsma stressed that the results from CAAR and the Early Science Program will help provide best practices that inform other scientific software developers and users of accelerated applications to make most effective use of this world-class capability. Many of the CAAR codes exhibit excellent scalability and performance, and five development teams in the Early Science Program who used Summit were selected as finalists for the prestigious Gordon Bell Prize, including the two winning teams.

"The excitement about Summit in the computational science community is evident from the enormous interest in the Early Science Program," said Straatsma. "The applications developed in the CAAR program will allow many of these teams to make a running start on Summit."

The arm ecosystem approach to design: current products and future outlook

Dan Ernst, Cray Inc.

Cray Inc., has collaborated with many individuals and organizations in the HPC community to jointly develop necessary technologies for HPC systems based on the Arm ecosystem. Ernst cited these developments as essential for creating opportunities, removing barriers, and improving the outcomes of developing technologies needed to improve the performance, efficiency, and productivity of science applications.

The task of bringing these technologies to market is increasingly enabled by the availability of open standards. To improve efficiency, Cray and collaborators have brought initial products to market based on Arm hardware with Cray software, achieving high levels of performance while also remaining productive for scientists. Ernst emphasized that these results would not have been possible without support from the HPC community. Going forward, additional efforts in this vein will likely focus on continued exploration of architectures that more effectively address HPC applications, including new hardware, software, and integration opportunities.

Architecture of the Tokyo Tech HPE SGI 8600

Mark R. Fernandez, Hewlett Packard Enterprise

According to Fernandez, many wondered how the SGI product line would be affected when Hewlett Packard Enterprise purchased SGI in 2016. This was particularly true of the ICE hardware line, which was in its fourth generation with the 8400 model. However, the 8400 would soon be expanded to the 8600 via a collaboration between HPE and the Tokyo Institute of Technology.

The HPE SGI 8600, with its e-cell–based architecture, was constructed for performance, density/scale, and efficiency. The new system, dubbed the Tsubame 3, features closed-loop liquid cooling with large, unified cooling racks, but also possesses an air-cooled capability. The Tsubame 3 exploits a co-designed SGI ICE-XA server node and uses a massively BYTES-centric architecture to converge AI and big data with HPC. The new system has worked so well that Hewlett Packard Enterprise has introduced it into its commercial product line.

"HPE is pleased with our collaboration with Tokyo Tech to create a server node exploiting our capabilities while addressing their needs," said Fernandez.

Early application results on Tsubame 3

Rio Yokota, Tokyo Institute of Technology

TSUBAME3, the fourth most powerful supercomputer in Japan and optimized for AI and big data applications, began operations at Tokyo Tech's Global Scientific Information and Computing Center in 2017. The center, along with several other supercomputing facilities, make up the Joint Usage/Research Center for Interdisciplinary Large-Scale Information Infrastructures. Through the High-Performance Computer Infrastructure program, the center provides facilities where teams of scientists from academia and industry pursue a variety of research projects.

For example, researchers have used Tsubame 3 to study molecular dynamics, fusion plasma turbulence, seismic wave propagation and tsunami simulation, in situ particle-based volume rendering, neural networks, and more. Researchers can also earn time on Tsubame 3 and other systems through the Tsubame Grand Challenge Program. Yokota emphasized the importance of these widespread collaborations, noting that running the same code on multiple pre-exascale systems throughout Japan and around the world would produce substantially more accurate and credible results.

"TSUBAME2 was the first production supercomputer to adopt GPUs back in 2010," said Yokota. "Since then, we have accumulated know-how and techniques to help our users transition to GPUs and learned how to manage such heterogenous systems. With TSUBAME3 we are finding increasing use in the machine learning and big data area, which benefit greatly from our experience with large-scale GPU systems."

Early application results on Isambard

Simon McIntosh-Smith, University of Bristol

Isambard is the world's first production 64-bit Arm supercomputer and is hosted by the UK's Met office. The system is a "tier 2" in the UK's tiered model of HPC provision, meaning that it is designed to experiment with promising technologies. Isambard features:

- 10,752 Armv8 cores,
- a high-speed Aries interconnect,
- a Cray software stack, and
- Marvell Thunder X2 CPUs.

The ThunderX2 has higher memory bandwidth than most mainstream CPUs, and while it has a lower peak floating point rate, its competitive pricing makes it very attractive from a price/performance point of view.

Isambard's performance on bandwidth-bound mini-apps such as STREAM, CloverLeaf, and TeaLeaf was above the competition. The system is intended to run real codes and achieve real science results, so the team optimized the ten most heavily used codes on the UK's national supercomputer system, Archer, such as GROMACS, NAMD, and NEMO. So far, the ThunderX2 CPU outperforms all of its competitors on OpenFOAM and comes a close second on CP2K, NAMD, NEMO, and OpenSBLI, a stunning achievement for a such competitive price tag.

The experience has taught the Isambard team that:

- performance of the Thunder X2 is competitive with high-end server CPUs,
- it enables a positive software ecosystem experience,
- Arm-based systems are now real alternatives to HPC and can bring much-needed competition to the market, and
- Arm ecosystems represent the best opportunity for real co-design.

ORNL staff form first-of-its-kind regional computing association

ORNL has a reputation as a computing trailblazer.

DOE's largest science and energy laboratory has stood up three of the world's fastest computers, and its latest champion, Summit, is hailed as the "smartest," or most AI-compatible, supercomputer ever designed.

It's this same pioneering spirit that led a group of East Tennessee computing researchers, including several ORNL staff, to announce the formation of the first regional chapter of the Association for Computing Machinery's Special Interest Group on High-Performance Computing at SMC18.

SIGHPC describes itself as "the first international group within a major professional society that is devoted exclusively to the needs of students, faculty, researchers, and practitioners in high-performance computing . . . to help spread the use of HPC, help raise the standards of the profession, and help ensure a rich and rewarding career for people involved in the field."

Appalachian HPC, as the area chapter will be known, is composed of member institutions across the region including the University of Tennessee–Knoxville, the University of Tennessee–Chattanooga, Vanderbilt University, Tennessee Tech University, and others.

It was this unusual concentration of computing expertise in a relatively small area that inspired Jeff Nichols, ORNL's Associate Laboratory Director for Computing and Computational Sciences who serves on the advisory board of SIGHPC, to suggest the chapter.

ORNL's Matthew Wolf is working with Terry Moore, associate director of the University of Tennessee's Innovative Computing Laboratory, and other stakeholders to draft a charter proposal for the regional SIGHPC chapter. Once ACM approves the charter document, the new group will be official, making Appalachian HPC the first organization of its kind to formally link the computing expertise of Knoxville, Chattanooga, Nashville, and Cookeville.

"An alliance with an ACM professional society opens up enormous professional development opportunities for the region's immense talent," said Nichols. "At the same time, these development opportunities benefit professional societies, and computing as a whole, by growing future society fellows and leaders throughout the high-performance computing community."

The chapter charter contains three overarching goals:

- Knowledge Exchange—sharing knowledge and technical context among members and activities such as flash talks, regional conferences, and vendor presentations.
- Workforce Development—sharing HPC expertise with students as well as professionals in related spheres with activities such as hackathons, training sessions, and mentorships.
- Community Building—developing professional and personal rapport and activities such as an awards committee, social meet-up-style networking, and a shared web presence.

"We thought these three goals served as a nice summary of the value people saw in this group," said Wolf, adding that if all goes well the chapter should officially launch in the fall with around 100 initial members, adding yet another "first" for a region with a knack for breaking new ground.

Dinner and a show

When you have so many smart people in a single location, you take advantage.

And what better way to siphon all of the knowledge from SMC18's speakers and attendees than a dinner panel? After all, people come to SMC to learn, even while they chew.

The panel's theme, "Artificial Intelligence Ethics In-Action," focused on behavior and the actions people and their organizations should or should not take in regards to AI research. Moderators included representatives from ORNL (Gina Tourassi and James Peery), industry (Providentia Worldwide's Ryan Quick), and academia (Georgia Tech's Justin Romberg).

"We wanted people from different organizations, to get a range of different perspectives," said ORNL's Amy Wolfe, an anthropologist and group leader for the laboratory's Society, Energy, and Environment Group who co-organized the panel with ORNL's AI Program Director David Womble and Laura Pullum in the Computational Data Analytics Group. "After all, this is a topic that affects all of us and, eventually, will likely affect everyone."

Few topics are hotter than AI, and Wolfe, Womble, and Pullum knew it would lead to plenty of conversations. "We felt like, after two full days of talks, attendees may be a little tired of being talked 'at.' The purpose of the dinner panel was to elicit more of a back-and-forth discussion," said Wolfe.

While the term "artificial intelligence" can mean many different things, attendant to every form of AI is a set of similar issues including data quality, perpetuation of implicit bias, and replicability and explainability. As AI becomes an increasing player in scientific inquiry, addressing these issues is critical for organizations tasked with conducting world-class science ethically and responsibly.

The theme transcended ethics, said Wolfe, and was intended to explore how individuals should behave and perform when faced with the complex issues inherent in AI.

The panel and audience members were presented with three scenarios.

The first, known as the "trolley scenario," is a classic case in which a trolley on a track will inevitably crash and kill one or many people. Participants are asked to decide who to kill and are expected to justify their decision. ORNL being a DOE lab, however, the trolley was replaced with an autonomous vehicle, but the moral conundrum was the same.

The second scenario asked the panel and audience to imagine an AI project intended to aid logistics in responses to natural disasters around the world, with the original goal being to save as many lives as possible. However, the scenario continues, those same data then are used by the government to identify and persecute dissidents. The panel and audience were then asked to consider questions such as "What is your professional responsibility?" "What is your organization's responsibility?" and "Who's accountable?" among others.

The third and final scenario presented the conference with a hypothetical survey in which national lab and university scientists say they want to "do good" and make the world a better place, but fail to point to a robust set of procedures they use to vet AI data sources, models, and dimensions for data training, the extent to which the models used are suited to the data at hand, and a process for guarding against the weaponization of the data.

"Each scenario was intended to make the panelists and audience members consider their own responsibility" as research professionals, said Wolfe. "Furthermore, to consider their responsibility in terms of ensuring that their institutions are using AI responsibly in their research. There was plenty of discussion, and we hope that the discussions, and the thoughts they provoke, will extend beyond the dinner and permeate peoples' professional lives."

The fact that ORNL is a multifaceted research organization that both uses and creates AI in its research operations makes it an ideal candidate to lead in terms of AI and ethics.

Data Challenge brings joy of discovery to participants, new insights to ORNL researchers

Participants in the 2018 SMC Data Challenge might have had different reasons for entering the annual data analysis contest, but they all came away with a joy of discovery. In its second year, the competition was organized by ORNL's Computing and Computational Sciences Directorate, and teams presented their results at SMC18.

The Data Challenge aims to connect ORNL domain researchers with data analysis enthusiasts to generate insights about data collected during basic and applied experiments. Researchers with the OLCF, Manufacturing Demonstration Facility, Center for Nanophase Materials Sciences, Spallation Neutron Source, and CCSD donated datasets and developed the Data Challenges. Teams chose from six challenges that included a series of tasks or questions and a dataset with which to explore the tasks. Datsets were related to supercomputing languages, materials discovery, quality control for 3-D printing, energy consumption, and machine learning for publications mining.

Teams had approximately 2 months to analyze and mine discoveries from the datasets. They then created a poster and video. Participants ranged from graduate student first-timers and young ORNL researchers—some newcomers, some returning—to consultants who work professionally on projects similar to the Data Challenge. One team's entry resulted from a yearlong collaboration with researchers at SNS.

The ORNL researchers, known as data sponsors, asked participants to describe general trends in the data and relationships between the variables captured by their datasets. Sponsors also generated challenge-specific tasks for teams to complete, including creating 2-D maps of materials components, determining the types of buildings whose energy use is most impacted by weather, and identifying historical shifts and geographic distribution in research topics. Participants were encouraged to produce data visualizations and consider issues of scalability.

All teams started by paring down their datasets to workable chunks, and, in some cases, participants revised the way data were configured to make them suitable for the analysis tools they wanted to use. A couple of the challenges required first discarding data related to human behavior. For example, teams looking at environment–building impacts on energy consumption noted that their dataset showed a large spike in energy use at the beginning of each day, perhaps from people turning on their computers before starting to work.

Ketan Maheshwari, a systems engineer with CADES at ORNL and winner of the Overall Professional Winner and Most Novel Solution awards, said his project turned out to be "super interesting" as he made progress. One of his first discoveries was that many of the publications contained in his dataset were written in languages other than English, a wrinkle he had not considered. (He decided to weed those out.) He also had to work, trial-and-error style, to find workable column separators for the data.

"At first I thought I should use an ampersand, but there are millions of ampersands in the dataset!" Maheshwari remembered. "I knew from experience I couldn't use colons or dashes, so I started using two-letter combinations and still struck out." He finally settled on a four-character separator.

Maheshwari had fun exploring the dataset itself. He found his own papers, discovering how they are organized within the dataset, and found hundreds of papers written by Albert Einstein and thousands by mathematician Ramanujan. "I enjoyed having all that interesting information at my fingertips," he said. "I spent hours and hours on weekends just trying to figure out what all I could do with the dataset."

Data sponsors and participants alike recognized that scientific investigations not traditionally intersecting with computer science can benefit from the application of big data methods. "These premier scientists are really trying hard to solve hard problems, but they need help from outside [their own ranks]," Yawei Hui, an astrophysicist with ORNL's Computer Science and Mathematics Division, said. "By combining the power of both sides, we can do more."

Kristin Tippey, a postdoctoral fellow with the Neutron Sciences Directorate at ORNL, summed up her experience analyzing a materials-related dataset as a back-and-forth exercise with the domain scientists—examining the data and finding something, then passing it back to the materials experts, saying "I've got something, but I'm not sure what I've got." In her view, the Data Challenge and similar exercises are "definitely a two-way street."

An exciting outcome of the Data Challenge is the real-world knowledge participants generate. Analyzing data for insights on how computer applications operate at the coding level could help optimize supercomputing operations; using data analysis to improve in situ quality control processes could help spur adoption of 3-D-printing technologies among manufacturers. William Gurecky, a University of Texas at Austin graduate student whose team won the Overall Student Winner award, pointed out that targeted data analysis could be used to produce forecasts of energy consumption, which are used in some cities to set fair market values for the next day's pricing.

Awards also included Best Data Story, given to Max Grossman and Nav Ravindranath of 7pod Technologies for their exploration of weather–building impacts on energy use, and Best Video Presentation, given to Cameron Kuchta and Thy Pham of ORNL for their use of machine learning with neutron scattering data.

Data sponsors for this year's competition included Melissa Allen, Alex Belianinov, Reuben Budiardja, Garrett Granroth, Christina Hoffmann, Stephen Jesse, Graham Lopez, Joshua New, Vincent Paquit, Robert Patton, Pete Peterson, Thomas Proffen, Jibo Sanyal, Vivek Sarkar (Georgia Institute of Technology), Jack Wells, Ross Whitfield, and Jisheng Zhao (Georgia Institute of Technology).

Data Challenge participants also included Supriya Chinthavali, Cooper Colglazier, Sajal Dash, J. Austin Ellis, Shubhankar Gahlot, Ethan Hicks, Philip Hicks, Olumide Kayode, Yaohua Liu, Tommy Moore, Byung H. Park, Zack Taylor, Ali Saman Tosun, and Zhonghua Zheng.

Tiffany Mintz organized the 2018 Data Challenge, with assistance from Folami Alamudun and Dasha Herrmannova.

SMC presenter bios

Ann Almgren
Lawrence Berkeley National Laboratory

Ann Almgren is a senior scientist in the Computational Research Division of Lawrence Berkeley National Laboratory and the Group Lead of the Center for Computational Sciences and Engineering. Her primary research interest is in computational algorithms for solving PDE's in a variety of application areas. Her current projects include the development and implementation of new multiphysics algorithms in high-resolution adaptive mesh codes that are designed for the latest multicore architectures. She is a SIAM Fellow and the Deputy Director of the ECP AMR Co-Design Center, and serves on the editorial boards of CAMCoS, IJHPCA and SIREV.

Stan Tomov
University of Tennessee, Knoxville

Stanimire (Stan) Tomov is a Research Director in the Innovative Computing Laboratory (ICL) and Research Assistant Professor in the Electrical Engineering and Computer Science Department at the University of Tennessee, Knoxville. He specializes in parallel algorithms, numerical analysis, and high-performance scientific computing (HPC). He is leading the development of the MAGMA libraries, targeting to provide a modernized LAPACK/ScaLAPACK on the next-generation of architectures, including multi/many-core CPUs, GPUs, and large scale heterogeneous systems.

Roger Pawlowski
Sandia National Laboratories

Roger P Pawlowski is a Principal Member of Technical Staff in the Computational Science department at Sandia National Laboratories in Albuquerque, NM. He joined Sandia in 2000 after earning a Ph.D. in Chemical Engineering from the State University of New York at Buffalo. His research interests include numerical algorithm development, high performance computing, and software design for next-generation architectures. He develops HPC applications for computational fluid dynamics, magnetohydrodynamics and plasma physics. He leads the Nonlinear Analysis product area for Trilinos, an effort to develop algorithms and enabling technologies within an object-oriented software framework for the solution of large-scale, complex multi-physics engineering and scientific problems. Dr. Pawlowski is currently the PI of the Software Components effort for the ECP SNL ATDM Math Libraries project.

Michael Sprague
National Renewable Energy Laboratory

Michael Sprague is a Principal Scientist at the National Renewable Energy Laboratory (NREL). Before joining NREL in 2010, he was an Assistant Professor and Founding Faculty of Applied Mathematics at the University of California, Merced. He earned his PhD in Mechanical Engineering at the University of Colorado. He is currently leading several computational science projects in wind energy, including a Department of Energy (DOE) Exascale Computing Project (https://exascaleproject.org/) called ExaWind (https://www.exawind.org/). Mike chaired a 2015 DOE Office of Science workshop, "Turbulent Flow Simulation at the Exascale: Opportunities and Challenges." His research interests include computational mechanics of fluids and structures (and their interaction), and high-performance computing.

Salman Habib
Argonne National Laboratory

Salman Habib is a member of the High Energy Physics (Group Leader for the Cosmological Physics and Advanced Computing (CPAC) group) and Mathematics and Computer Science Divisions at Argonne National Laboratory, a Senior Member of the Kavli Institute for Cosmological Physics at the University of Chicago, and a Senior Institute Fellow in the Northwestern Argonne Institute of Science and Engineering. His research interests and contributions cover the broad sweep of classical and quantum dynamical systems, from field theories to particles, and from the largest scales to the smallest. Since the early 1990's, he has worked on algorithm and code development for parallel supercomputers in a variety of fields -- ranging from particle and atomic physics to astrophysics and cosmology -- and on many platforms.

Habib initiated the development of the program that led to the HACC framework originally at Los Alamos National Laboratory (2008). He is the PI for the ExaSky Applications Development project under the DOE ECP, which currently supports HACC development; he is also the PI for a joint DOE ASCR-HEP SciDAC project on data-intensive computing with a focus on advanced statistical methods and machine learning.

Arvind Ramanathan
Oak Ridge National Laboratory

Arvind Ramanathan is a staff scientist in the Computational Science and Engineering Division and the Health Data Sciences Institute at Oak Ridge National Laboratory. His research interests are at the intersection of data science, high performance computing and biological/healthcare science. He builds data analytic tools to gain insights into the structure-dynamics- function relationships of bio-molecules. In conjunction with biophysical/biochemical experiments and long time- scale computational simulations, his group investigates bio-molecular systems that have implications for human health. In addition, his group has also developed novel data analytic tools for public health surveillance. He has published over 30 papers, and his work has been highlighted in the popular media, including NPR and NBC News.

Within ORNL, he is the technical lead for the Exascale Computing Project on Cancer Deep Learning Environment (CANDLE) and the Joint Discovery of Advanced Computing Solutions for Cancer (JDACS4C) between the DOE and NCI on the RAS project. His team develops advanced machine learning and artificial intelligence approaches for large-scale molecular simulation datasets. More information about his group and research interests can be found at http://ramanathanlab.org.

Ryan Quick
Providentia Worldwide

Ryan Quick received degrees in English and Philosophy from Vanderbilt University and went on to study American Christian Ethics at Yale University. He has been active in the Internet and Linux communities since the early 1990s. He focused on distributed systems for the last 25 years, with special attention to the interaction between applications, operating systems, and the hardware and networks underlying them. Ryan holds patents for messaging middleware systems, and pioneers bridging High-Performance Computing technologies with enterprise best-practices. His HPC work for real-time analytics garnered provisional patents and awards. He is recognized for innovation in hardware and application design, messaging ontology, and event-driven systems. Currently, he brings machine learning, real-time streaming, set-selection, and digital signal processing technologies to predictive analytics for self-healing in command and control systems.

Brian van Essen
Lawrence Livermore National Laboratory

Brian Van Essen is the Informatics Group leader and a Computer Scientist at the Center for Applied Scientific Computing at Lawrence Livermore National Laboratory (LLNL). He is actively pursuing research in large-scale deep learning for scientific domains and training deep neural networks using high-performance computing systems. He is the project leader for the Livermore Big Artificial Neural Network (LBANN) open-source deep learning toolkit. Additionally, he co-leads an effort to mapping these scientific, data-intensive, and machine learning applications to Neuromorphic architectures. His research interests also include developing new Operating Systems and Runtimes (OS/R) that exploit persistent memory architectures, including distributed and multi-level non-volatile memory hierarchies, for high-performance, data-intensive computing.

Felix Herrmann
Georgia Institute of Technology

Dr. Felix J. Herrmann is Georgia Research Alliance Eminent Scholar in Energy and a professor at the Georgia Institute of Technology with appointments in the Schools of Earth and Atmospheric Sciences, Computational Science and Engineering, and Electrical and Computer Engineering. Dr. Herrmann will be the 2019 Distinguished Lecturer of the Society of Exploration Geophysicists (SEG).

During his career, Dr. Herrmann has worked on the development of the next-generation of industrial seismic data acquisition and imaging technologies. His research group is known to drive innovations by leveraging recent developments in the mathematical and computational sciences. Dr. Herrmann's group is widely recognized for breakthroughs in seismic data acquisition with compressive sensing and wave-equation inversion with stochastic and constrained optimization. More recently, Dr. Herrmann's group has been involved in the development of Devito—a Domain-specific Language (DSL) and automatic code generation framework for highly optimized finite differences for use in inversion methods and JUDI, an abstract framework for large-scale seismic modeling and inversion.

C-S Chang
Princeton Plasma Physics Laboratory

C-S Chang has been leading several large-scale, multi-institutional, multi-disciplinary projects whose members are composed of fusion energy scientists, applied mathematicians and computer scientists. For example, presently, he is the director of the SciDAC-4 Partnership Center for High-fidelity Boundary Plasma Simulation (HBPS), Science Co-Director of the ECP-Application Whole Device Modeling Center (WDM), Institutional PI of the ECP Co-design Center for Particle Methods (COPA). In the past, he directed the SciDAC-2 CPES and SciDAC-3 EPSI fusion application projects, and led the FES/ASCR Exascale Requirement Review Workshop activities. C-S Chang's home institution is Princeton Plasma Physics Laboratory as Managing Principal Physicist. He is a Fellow of the American Physical Society.

Steve Oberlin
NVIDIA

Steve Oberlin is the Chief Technology Officer for Accelerated Computing at NVIDIA. His large-scale computing technology career has spanned over 30 years, launched in 1980 at Cray Research bringing up CRAY-1 supercomputer systems. Starting in 1981, he worked for Seymour Cray as a designer and project engineer on the CRAY-2 and CRAY-3 supercomputers. In 1988, he led early massively parallel processing research at Cray that ultimately led to his role as the chief architect of the CRAY T3D MPP and its successor, the CRAY T3E. He holds 15 architecture and design patents for the T3D and T3E.

Steve was VP of Hardware Engineering in Chippewa Falls, Wisconsin, for Cray/SGI from 1996 until early 1999, responsible for hardware development and support of all Cray products and their follow-ons. He left SGI to found Unlimited Scale, Inc., in July of 2000, and spent the next 13 years creating new cloud computing infrastructure management and intelligent resource optimization technologies for start-up Cassatt and CA Technologies.

Steve returned to HPC at NVIDIA in November 2013. As CTO for Accelerated Computing, he is responsible NVIDIA's Tesla roadmap and architecture.

Erich Strohmaier
Lawrence Berkeley National Laboratory

Strohmaier joined Berkeley Lab in 2001 and currently leads the Computational Research Division's Performance and Algorithms Research Group. His current research focuses on performance characterization, evaluation, modeling, and prediction for HPC systems; analysis of advanced computer architectures and parallel programming paradigms; classification of and programming patterns for scientific computational kernels; and analysis and optimization of data-intensive large scale scientific workflows. He was a member of the team awarded the ACM Gordon Bell Prize in 2008 for parallel processing research in the special category for algorithmic innovation.

Rajkumar Kettimuthu
Argonne National Laboratory

Rajkumar Kettimuthu received the B.E. degree from Anna University, Chennai, India, and an M.S. and Ph. D. from the Ohio State University, Columbus, OH, USA, all in Computer Science and Engineering. Since 2003, he has been working at Argonne National Laboratory, where he is currently a Computer Scientist in the Data Science and Learning Division. He has co-authored more than 100 articles in the areas of high performance computing, distributed computing, and high-performance networking. He is a recipient of R&D 100 award.

Edmund Begoli
Oak Ridge National Laboratory

Edmon Begoli, PhD, is the Chief Data Architect with Computational Sciences and Engineering Division at Oak Ridge National Laboratory (ORNL), and is a Principal Investigator (PI) for the joint DOE and VA program (MVP CHAMPION) in precision medicine. During his tenure at ORNL, Edmon led several major national projects in healthcare and defense, including Knowledge Discovery Initiative (KDI) for Centers for Medicare and Medicaid Services (CMS).

Edmon holds undergraduate, graduate, and doctoral degrees in Computer Science, and an adjunct professor of computer science at the University of Tennessee-Knoxville.

Sven Leyffer
Argonne National Laboratory

Sven is a co-editor of Mathematical Programming and served as an editor-in-chief of Mathematical Methods of Operations Research. He serves on the editorial board of Computational Optimization and Applications and Mathematics of Computation. In addition, Sven has served as the INFORMS Optimization Vice-Chair for nonlinear programming and as the Program Director of the SIAM activity group on optimization. Sven was SIAM Vice President for Programs from 2010-2013. Currently, Sven serves as the Secretary of ICIAM. For other appointments, see this list of other appointments.

Together with Roger Fletcher and Philippe L. Toint, Sven was awarded the Lagrange prize in optimization in 2006. In 2009, Sven became a SIAM Fellow. In 2013, Sven won the 2012 COAP Best Paper Prize with Chungen Chen and Roger Fletcher for their paper on nonmonotone filter methods. In 2016, Sven was awarded the Farkas Prize for Mid-career Researchers by INFORMS

James Sexton
IBM

Dr. James Sexton is an IBM Fellow and Director of the Data Centric Systems department at IBM T. J. Watson Research Center in New York. Dr. Sexton received his Ph.D. in Theoretical Physics from Columbia University, NY. His areas of interest lie in High Performance Computing, Computational Science, Applied Mathematics and Analytics. Prior to joining IBM, Dr. Sexton held appointments as Lecturer then Professor at Trinity College Dublin, and as postdoctoral fellow at IBM T. J. Watson Research Center, at the Institute for Advanced Study at Princeton and at Fermi National Accelerator Laboratory.

Tjerk Straatsma
Oak Ridge National Laboratory

Dr. Tjerk P. Straatsma is an internationally recognized scientist with more than 30 years of experience in the development, efficient implementation and application of advanced modeling and simulation methods as key scientific tools in the study of chemical and biomolecular systems, complementing analytical theories and experimental studies. Dr. Straatsma joined Oak Ridge National Laboratory in 2013, where he currently leads the Center for Accelerated Application Readiness in the National Center for Computational Sciences.

Dan Ernst
Cray

Dr. Daniel Ernst is currently a Principal Engineer in Cray's Advanced Technology Development team, where he leads multiple investigations into future HPC architectures. His focus is on high-performance memory systems, application-optimized architectures, future HPC and HPDA node architectures, and system simulation. Dan was the Principal Investigator for Cray's Department of Energy Fast Forward 2program, which focused on bringing Arm technologies into HPC.

Dan is Cray's primary representative to the JEDEC memory standards committee, the CCIX Consortium, and serves on the Board of Directors of the Gen-Z Consortium. He received his Ph.D. in Computer Science and Engineering from the University of Michigan in 2005, where he studied high-performance, low-power, and fault-tolerant microarchitectures.

Mark Fernandez
HPE

Dr. Mark Fernandez has 20+ years of experience in creating, configuring, and assisting with multiple types of technical, scientific and high performance computing (HPC) solutions. His current position with SGI/HPE is Americas HPC Technology Officer. Specifically, he is the Spaceborne Computer Payload Developer (PD) for software and the project's Co-Investigator. Dr. Fernandez and the SGI/HPE Team created the Spaceborne Computer experiment to study the practicality of running and managing a COTS high performance computer (HPC) system in orbit aboard the International Space Station (ISS).

Dr. Fernandez is also responsible for working to capture requirements and incorporating them into future SGI/HPE HPC systems and solutions. He also works closely with Engineering throughout the product development cycle to insure not only that the products are relevant and meet requirements, but also are candidates for inclusion into larger, more complete solutions.

Rio Yokota
Tokyo Institute of Technology

Rio Yokota is an Associate Professor at the Global Scientific Information and Computing Center at Tokyo Institute of Technology. Dr. Yokota's research interests include developing scalable hierarchical algorithms for scientific computing. Dr. Yokota is the main developer fo the exaFMM code, and is a co-developer of the HiCMA code. Dr. Yokota is a recipient of the ACM Gordon Bell prize (price performance) in 2009. His most recent research interests are at the intersection of high performance computing and deep learning.

Simon McIntosh-Smith
University of Bristol

Simon McIntosh-Smith is a full Professor of High Performance Computing at the University of Bristol in the UK. He began his career as a microprocessor architect at Inmos and STMicroelectronics in the early 1990s, before co-designing the world's first fully programmable GPU at Pixelfusion in 1999. In 2002 he co-founded ClearSpeed Technology where, as Director of Architecture and Applications, he co-developed the first modern many-core HPC accelerators. He now leads the High Performance Computing Research Group at the University of Bristol, where his research focuses on performance portability and application based fault tolerance. He plays a key role in designing and procuring HPC services at the local, regional and national level, including the UK's national HPC server, Archer. In 2016 he led the successful bid by the GW4 Alliance along with the UK's Met Office and Cray, to design and build 'Isambard', the world's first production, Armv8-based supercomputer.

SMC data challenges

Challenge 1: Discovering Features in $Sr_{14}Cu_{24}O_{41}$

Neutron single crystal diffraction is a powerful way to examine the atomic structure of technologically interesting materials. The particular material represented in this data is $Sr_{14}Cu_{24}O_{41}$i, ii, a semiconducting solid with industrial application potential for optoelectronic and thermoelectric devices. $Sr_{14}Cu_{24}O_{41}$ has two intertwined structures. One is a Cu_2O_3 two-leg ladders, the other is CuO_2 chains. The ladder and chain structures are incommensurate. That means there is not a simple ratio of the number of chain links to the number of ladder rungs.

Neutron diffraction is highly sensitive to the internal structure or atomic arrangement of the material to be studied, which is built from a periodic arrangement of closely connected atoms as structural units, repeating on a 3-D translational grid in (h, k, l) units.

This data set has 4 dimensions: Intensity (I), and three spatial dimensions (h, k, l) which describe three directions within the material. If there is a large, sharp intensity, I, at a given (h, k, l) coordinate, that means there are many pairs of atoms spaced at that characteristic length scale. If there is a broad distribution of atom spacing broader features associated with disorder in the system appear.

The data itself has already been corrected for experimental effects, normalized, and put onto a regular three-dimensional grid in (h, k, l). However, not all of the regular grid has been measured. At these points the intensity (I) has been set to not-a-number (NAN). The data is in an HDF5 file.

H5py is a useful set of libraries for accessing HDF5 files from python.

If you are not familiar with hdf5 and want to browse the data structure want to browse the data structure we recommend hdfview or nexpy. However, the developed algorithms should run independently of these codes.

Do these:

1. Segment the data – There are features that vary in sharpness and intensity as well as background. Separate these into various categories
2. There are multiple sets of sharp features (known as Bragg peaks) that are 3-D arrays. Pick out these arrays and give the minimum distance between peaks for each array in each direction.

Do at least two of the following:

3. Characterize the sharp features and relationships between them
    a. Identify regular intensity variations, directionality, widths etc.
    b. Identify how many collections of these features are in the data
4. Characterize the broad features and relationships between them
    a. Identify regular intensity variations, directionality, etc.
    b. Identify how many collections of these features are in the data

5. Are there relationships between the sharp and broad features

6. Visualize the above results

Our preference is for the algorithms to be implemented in Python with use of the numpy library, other necessary libraries are welcome. If the participant prefers a compiled code, C/C++, is our preferred solution with the output being in a python or numpy data structure.

This data set is moderately sized. Larger data sets are possible so a design with an eye toward scalability is desired.

[i] J. Etrillarda, M. Braden, A. Gukasov, U. Ammerahl, A. Revcolevschi; Physica C 403 (2004) 290–296; DOI: 10.1016/j.physc.2004.01.003

[ii] X. Chen, D. Bansal, S. Sullivan, D.L. Abernathy, A.A. Aczel, J. Zhou, O. Delaire, L. Shi; Phys. Rev. B 94, 134309 (2016); DOI: 10.1103/PhysRevB.94.134309

Challenge 2: Impact of Urban Weather on Energy Use

Recent advances in multiscale coupling of models have started to provide unique insights into how interdependent processes affect one another. The effect of these processes is uniquely observable in urban environments.

This data set comprises of three elements:

a. High resolution, 90-meter simulated weather data for one month at 15-minute intervals (with known gaps towards the end of each month). These files are in netcdf file format and about 45 GB in size.

b. A mapping of individual buildings with individual IDs, their lat/lon location, their 2-D footprint, and height. (Excel file)

c. Energy simulation output of these individual buildings, at 15-minute intervals for a whole year.

The questions that are of interest for this challenge are:

1. Are there interesting variations in the weather and building energy use data for the geographic area?

2. Which buildings in the study have their energy use impacted the most by external factors including, including the weather?

3. Are there any interesting visualizations that illustrate the changing dynamics of the simulated urban environment?

Participants are welcome to bring in additional datasets and fuse with the provided data to create meaningful insights.

Challenge 3: Fly Away with Me

This challenge is driven by efforts to expedite materials data analysis and generate insight into physics and chemistry of industry relevant materials. Ferroelectric lithium niobate (LiNbO3) is widely used in integrated and waveguides due to its optical, piezoelectric, electro-optic, elastic, photoelastic, and photorefractive properties.1 This is a human-made dielectric and does not exist in nature; with its feoelectric properties first shown in 1949.2 It is now extensively used in the telecoms market, for mobile telephones, optical modulators, and of surface acoustic wave devices.3 Although lithium niobate is important in numerous broad areas of technological significance, the details, and more importantly the origins of its physical as well as chemical properties remain hotly debated. The purpose of this data challenge is to piece together the chemical behavior of LiNbO3 from Time of Flight Mass Secondary Mass Spectrometry data as a function of applied electrical bias.

Time of Flight Mass Secondary Mass Spectrometry (ToF-SIMS) is a destructive analysis technique designed to reveal the chemical composition of the sample's topmost layer.4Ions from a primary ion source strike the surface breaking bonds and releasing some material particles and their associated fragments. Fragments produced in the top 2-3 monolayers of the sample will have enough energy to overcome the surface binding energy and leave the sample. A small portion of those will be charged, either positively or negatively depending on their electron configuration. The mass to charge ratio (m/z) of the species are analyzed and yield positive and negative secondary ion mass spectra consisting of the ion m/z versus the number of ions detected at each m/z.

Data generated by ToF-SIMS contains millions of points per spectra. The ToF detector effectively counts every single event, (that is every charged atom or molecule!) as a function of time. This generates sparse data sets containing many single or zero events, with robust peaks containing thousands to million counts. A spectrum is collected at each spatial pixel. There are typically 256 × 256 or 128 × 128 pixels per single chemical image. In this data series we contain multiple chemical images as a function of depth into material, the applied electrical bias, and the distance between two biased electrodes. This combinatorial approach is already a challenge in sample preparation and data collection; which is further exacerbated by a wealth of extracted information at both global and local scales necessitating a drastic improvement in capability to transfer, store and analyze multidimensional data sets.

In this challenge, contestants will be supplied with a series of 3-D chemical image data sets as a function of spatial position, applied electrical bias, and distance between two biased electrodes. Each 3-D data set consists of a mass spectrum mass 1 – 500 at each pixel. The challenge is to spatially co-register datasets taken in the same location; offer insight into which elements, or complexes, are most affected by bias; how this effect changes with electrode spacing; and visualize the distribution of key chemical players and their change as a function of bias and electrode spacing.

Challenge Questions

1. Co-register 3-D chemical image sets taken in the same location.

Deliverable: Co-registered data and the transformation matrix.

2. Identify key peaks (m/z values) that show the largest response to:

    a. Applied bias

    b. Distance between biased electrodes.

Deliverable: (a) 2-D maps of key components and their evolution as a function of bias and distance between biased electrodes, quantification and visualization of this difference.

3. (Advanced) Identify inter-relationship between the key components (m/z values). Which elements or fragments are co-dependent? Which are mutually exclusive? How does this behavior change as a function of bias, distance between biased electrodes? Deliverable: Visualization of key m/z value behavior as a function of each other and the experimentally varying conditions: bias, distance between biased electrodes.

References

1. Weis, R.; Gaylord, T., Lithium niobate: summary of physical properties and crystal structure. Applied Physics A 1985, 37 (4), 191-203.
2. Matthias, B.; Remeika, J., Ferroelectricity in the ilmenite structure. Physical Review 1949, 76 (12), 1886.
3. Toney, J. E., Lithium Niobate Photonics. Artech House: 2015.
4. Belu, A. M.; Graham, D. J.; Castner, D. G., Time-of-flight secondary ion mass spectrometry: techniques and applications for the characterization of biomaterial surfaces. Biomaterials 2003, 24 (21), 3635-3653.

Challenge 4: Scientific Publication Mining

Scientific research continues to expand both human understanding of our world and solve societal problems through technical progress. One way that this progress is documented is through scientific publications. However, there are now millions of publications available for researchers from all science and technology domains. Consequently, it is nearly impossible for humans to thoroughly research across these millions of publications. The goal of this challenge is to develop and apply machine learning and statistical techniques to mine these publications and identify key characteristics and patterns that can be used by human researchers to develop useful knowledge and further enhance scientific discovery.

The dataset available for this task consist of scientific publication records. The metadata for each publication include title, abstract, author list and the publication venue and date. For a portion of the publications the full-text of the paper will also be available. The participants are welcome to use external data in their approaches as long as that data is publicly accessible. All participants will be asked to document all external data sources, and detail how the data was used.

Challenge Questions

1. Identify the individual or group of individuals who appear to be the expert in a particular field or sub-field.

Experts are people with high level of knowledge in a certain area. Recognizing experts can be beneficial to students familiarizing themselves with a new area or to scientists looking for collaborators. The goal of this task is to employ different methods, for example modelling or graph-based algorithms, and apply them on the dataset to discover people with high level of expertise. The response to this task should include example output, such as the model or graph developed with highlighted important nodes or a list of names, and a description of tools and methods used to produce the output.

2. Identify topics that have been researched across all publications.

Given a collection of documents, the goal of this task is to extract topics that recur in the collection so that a person not familiar with the collection can quickly explore its contents. The aim is to assist human understanding, so a good solution should identify topics in a way that makes sense to a person. This task could explore for example graph or text clustering methods. The solution should also include a description of methods used for the task.

3. Visualize the geographic distribution of the topics in the publications.

Researchers are associated with different institutions across the globe. Following up on the previous task, the goal of this task is to visualize the identified topics with respect to their geographical distribution. Are there certain locations which focus on specific topics? The solution should again contain a description of how was the output produced.

4. Identify how topics have shifted over time.

The goal of this task is understanding popularity evolution of topics over time, or in other words how the knowledge base is changing over time with the influx of new topics, growth or decay of older topics. Understanding the popularity of topics is important because it helps in identifying trending topics. Same as in case of the previous tasks, the solution should include example output and a description of methods used to produce the output.

5. Given a research proposal, determine whether the proposed work has been accomplished previously.

Choosing which proposals to fund is a complicated task, because the evaluators needs to be aware of the research area and whether the proposed research is novel. The goal of this task is to identify whether there are any publications which have previously tackled the proposed research.

Challenge 5: Automated in-situ Defects Detection in Powder Bed Metal Additive Manufacturing Parts

Additive manufacturing (AM) is the ability to deposit materials layer-by-layer or point-by-point to fabricate complex components directly from computer-aided design models. Although AM technologies have demonstrated the ability to fabricate complex geometries capable of achieving improved performance characteristics, few AM components are currently being used in production environments, mainly due to the challenges and costs associated with the certification and qualification of components. The current state of the industry is to certify components by using expensive methods such as computed tomography or mechanical testing, but their cost is working against the business case for AM components. An alternative method is to take a data driven approach to fully understand how the series of interconnected material deposition/melting events results in specific spatial material properties and/or defects. This imposes to create first a digital twin of the additive part as we built it using in-situ measurements and then to use data analytics techniques to learn from such data. This concept is the foundation of the Data Analytics Framework for Manufacturing that the ORNL Manufacturing Demonstration Facility (MDF) is actively developing to address the certification and qualification problem.

As part of this framework, the proposed data challenge focuses on the detection of specific defects in parts manufactured using an electron beam powder bed system, the ARCAM Q10 machine, (http://www.arcam.com/technology/products/arcam-q10/). To understand how the powder bed melting process works, please refer to this video: https://www.youtube.com/watch?v=M_qSnjKN7f8). For this challenge we are mainly interested to quantitatively assess the geometric accuracy of the part and the presence of failure points such as porosity, swelling, cracks, delamination, and lack of fusion. The importance of detecting defects in-situ is twofold: (1) detected early they can eventually be corrected on the fly with a feedback loop control mechanism, hence insuring a higher manufacturing success rate; and (2) these defects can be used as criteria to discard or to accept a part if the intended use of such is or not compromised. Either option will help circumvent the need for expensive testing. On the Q10 system, hundreds of heterogeneous sensing modalities are monitored to ensure the machine operation. Amongst them, for in-situ quality control, the ARCAM Q10 machine is equipped with a near-infrared sensitive camera capturing an emissivity map of the powder bed once a layer is completed. Each image (see Figure 1) shows variations in pixel intensities as a function of temperature, variations indicative of the presence of a feature of interest.

The dataset provided was created using the Dream3D open source platform. It includes an HDF5 file with the extension ".dream3d" and ".xmdf" files that can be used in Paraview to visualize the data. The dataset contains one data container per additive part, and each data container contains multiple attribute matrices, one for each modality of the digitized version of the build. For this challenge, we have only included two image modalities:

- STL slices images: additive parts are printed by stacking up slices extracted from the source CAD file. We have recorded for each layer the corresponding slice, extracted at the desired height as a black and white image, where white regions correspond to the intended printed regions and black region should not be printed. Note: there is an antialiasing effect around the contours (at the transition black/white).
- Near infrared images: they represent the emissivity measurement at the end of the print for each layer. In Figure 1, porosity appears as bright dots of various sizes, the printed contours (white curvilinear shapes) are delineating homogeneous grey regions corresponding to the infill melt areas, form the unmelted black region. As a rule of thumb, any disturbance of the grey region corresponds to a defect. Going through the entire stack of NIR images one will notice that the grey value within a region is almost never the same throughout the height of a part. This is caused by the scan strategy optimization for each layer which make the electron beam visit the same area at different times when building up. As a result, the thermal emissivity varies, hence the change in measurement

Each attribute matrix holds a stack of thousands of images registered in space. The dimensions of the 3-D stack, its resolution and its position in space are recorded in the dream3d file.

Challenge Questions: we are proposing five challenge questions, ranked by complexity:

1. Delineate the inside contour of each part: for each part delineate the interior region with subpixel accuracy
2. Detect and map all defects present in each part: identify non-uniform pattern in the melted region, without necessarily labeling them one of the aforementioned defects of interest.

3. Detect and map porosity: porosity is one of the most critical defect to identify. Building upon question 2, implement a classification mechanism to distinguish between pores and the other defects

4. Delineate the outside contour of each part: delineating the outside contour can be more challenging when two objects are close to each other. Your solution from question 1 will most likely have to be adapted to achieve subpixel location of the outside contour.

5. Implement a solution that can delineate the outside of each contour and map porosity for each layer with the computing time under one second: there is approximately 5 seconds between the capture of the NIR image and the beginning of the next layer. In the scheme of a feedback loop control implementation, the detection of major defects should be completed before the next layer start in order to implement corrective actions. Therefore, the geometric accuracy assessment and porosity map should be completed in maximum one second, to leave time for the system reconfiguration.

6. (optional) we will offer to benchmark the algorithm against at least one dataset from a similar build for which we have high resolution CT of the parts showing the exact location of pores. The results will be provided to the participant to include in their final submission.

There are non-constraints on the type of technique to use to process the data, anything ranging from image processing, statistical analysis, machine learning, etc. is welcome.

Challenge 6: Massive Visualization of Application Codes

HPC Application developers are hero programmers because writing parallel programs is much harder than writing sequential ones. They have to understand the intricate details of the target architecture (e.g. GPUs, etc), and the programming models to exploit this parallelism and express this in the code structure of the application. Using performance data, they need to go through thousands or millions of lines of code so that they can devise a strategy to port to the target architecture. Typically, most codes start by parallelizing the application across nodes and then by adding in-node parallelisms incrementally to exploit the multicores or accelerators available on the node until the target performance is met.

The data set for this challenge contains metadata of the program information for the E3SM (Energy Exascale Earth System Model) application. It contains information about the usage of Fortran features, programming models, subroutine calls, numbers of statements that are parallelized, type of statements, Fortran module usage, source and object file location of subroutines, etc.

Challenge Questions

We need a scalable way to visualize this information. The challenges are:

1. Build a call graph of the application (in the order of 7,000 nodes and 22,000 edges) based on the JSON files.
2. Visualize the call graph in a scalable way using different algorithms to show the distance between the nodes.
3. Classify the nodes of the call graph, for example, based on
    a. directory location of the source files or alternative way(s) to indicate information about the physical model (sea, ice, land, atmosphere)
    b. OpenMP and OpenACC
    c. code usage from a given Fortran module
    d. number of executable statements and/or variables
    e. relationships between function callers/callees
    f. library call invocations and/or invocation frequencies
    g. (any others)
4. Demonstrate the distribution of the different libraries' usage in application.
5. Show the relative similarities between nodes based on code features or programming techniques such as, for example:
    a. Number of loops or looping structures
    b. Parallelization
    c. Module usage
    d. Call site sequences
    e. Other classifications from #3
    f. (any others)

See you next year!