ADVANCED SCIENTIFIC COMPUTING ADVISORY COMMITTEE
to the
U.S. DEPARTMENT OF ENERGY

MEETING MINUTES

July 21-22, 2022

HYBRID MEETING

ADVANCED SCIENTIFIC COMPUTING ADVISORY COMMITTEE

The U.S. Department of Energy (DOE) Advanced Scientific Computing Advisory Committee (ASCAC) convened a hybrid meeting on Thursday and Friday, July 21-22, 2022 at the Westin Crystal City Reagan National Airport Marriott Hotel (1800 Richmond Highway, Arlington, Virginia) and via Zoom. The meeting was open to the public and conducted in accordance with the requirements of the Federal Advisory Committee Act (FACA). Information about ASCAC and this meeting can be found at http://science.osti.gov/ascr/ascac.

ASCAC Members Present

Daniel Reed (Chairperson)
Richard Arthur
Keren Bergman (remote)
Martin Berzins
Tina Brower-Thomas (remote)
Vinton Cerf
Barbara Chapman (remote)
Jacqueline Chen (remote)
Silvia Crivelli
Mark Dean (remote)
Jack Dongarra (remote)
Timothy Germann
Roscoe Giles
Susan Gregurick (remote)
Bruce Hendrickson (remote)
Anthony Hey (remote)
Richard Lethin
Mary Ann Leung
Satoshi Matsouka
Jill Mesirov (remote)
John Negele (remote)
Edward Seidel (remote)
Krysta Svore (remote)
Valerie Taylor

ASCAC Members Absent

John Dolbow
Gilbert Herrera
Alexandra Landsberg
Vivek Sarkar

Also Participating

Scott Atchley, Oak Ridge National Laboratory (ORNL)
Jim Ang, Pacific Northwest National Laboratory (PNNL)
Christine Chalk, ASCAC Designated Federal Officer, Oak Ridge Leadership Computing Facility (OLCF), Advanced Scientific Computing Research (ASCR)
Asmeret Berhe, Director, Office of Science, Department of Energy (DOE SC)
Cynthia Friend, Kavli Foundation and Harvard University
Barbara Helland, ASCR
Mike Heroux, Sandia National Laboratories (SNL)
Morgan Kelley, Dell Technologies
Ceren Susut, ASCR
Noah Mandell, Massachusetts Institute of Technology
George Michelogiannakis, Lawrence Berkeley National Laboratory (LBNL)
Todd Munson, Argonne National Laboratory (ANL)
Jeff Miller, Harvard University
Jordan Thomas, ASCR
Venkat Vishwanath, ANL
Justin Whitt, ORNL

Attending

In person:

Tom Beck, ORNL
Perrin Chalk, student
Leland Cogliani, Lewis-Burke Associates
Jody Crisp, Oak Ridge Institute for Science Education (ORISE)
Lori Diachin, Lawrence Livermore National Laboratory (LLNL)
Erik Draeger, LLNL
Paul Hovland, ANL
Jeff Hittinger, LLNL
Jim Malone, ORISE
Dave Martin, ANL
Kathryn Mohror, LLNL
Griffin Reinecke, Lewis-Burke Associates
Suzy Tichenor, ORNL

On Zoom:

There were approximately 250 individuals present in total for all or part of the meeting.

OPENING REMARKS, **Reed**, ASCAC Chair, convened the meeting at 10:00 a.m. Eastern Time and welcomed attendees.

Computational Science Graduate Fellowship (CSGF) posters presented yesterday represented the breadth and depth of fellows' research. Completion of a National Academy of Sciences (NAS) study examining post-exascale computing is projected for the end of 2022. U.S. Innovation and Competition Act (USICA) reconciliation is ongoing. Congressional decisions regarding authorizations and appropriations for the Creating Helpful Incentives to Produce Semiconductors (CHIPS) Act will cut across all DOE science areas.

VIEW FROM GERMANTOWN, Barbara Helland, Associate Director for Advanced Scientific Computing Research

Helland reviewed new ASCAC members, changes to ASCR personnel, and open ASCR positions. Select DOE programs have been reorganized under the Office of the Under Secretary for Science and Innovation (S4) led by Dr. Geraldine Richmond.

The President's Budget Request (PBR) for fiscal year 2023 (FY23) of ~$1.07B for ASCR represents an ~3% increase over the FY22 Enacted Appropriations. The FY23 PBR allocates ~$72M for Applied Mathematics Research; ~$70M for Computer Sciences Research; ~$98M for Computational Partnerships; ~$114M for Advanced Computing Research; ~$25M for the Energy Earthshot Research Centers (EERCs); ~$115M for High Performance Production Computing; ~$408M for the Leadership Computing Facilities; $77M for the Exascale Computing Project (ECP), and ~$90M for High Performance Network Facilities and Testbeds. Within these funds, ~$159 is designated for the Argonne Leadership Computing Facility (ALCF); ~$249 for the Oak Ridge Leadership Computing Facility (OLCF); ~$115 for the National Energy Research Computing (NERSC) Center; ~$90M for the Energy Sciences Network (ESnet); and the remaining ~$379M for research. Among other programs and initiatives, the PBR continues support for ASCR's participation in the Biopreparedness Research Virtual Environment (BRaVE); collaboration with the National Institutes of Health (NIH); operation of the National Quantum Information Science Research Centers (NQISCs); and basic research in quantum information sciences (QIS), and artificial intelligence and machine learning (AI/ ML). The PBR also supports optimal facility operations and seeks funding for software sustainability.

The FY23 House Mark advises spending no less than ~$1.05B for ASCR. Guidance specifies ≥$170M, ≥$250M, ≥$120M, and ≥$90M for the ALCF, OLCF, NERSC, and ESnet, respectively. ASCR is instructed to spend ≥$300M on Mathematical, Computational, and Computer Sciences Research; and between $15M and $45M for the development of advanced memory technologies by a U.S.-based manufacturer of memory systems and memory semantic storage. ASCR and DOE SC are instructed to continue the planning and design for the High-Performance Data Facility (HPDF); to support creation of a cross-cutting research program to deliver AI research, development, and deployment to increase user facility productivity via the Center for Advanced Mathematics for Energy Research Applications (CAMERA); to explore the viability of photonic quantum computing in coordination with other federal agencies; and to consider mechanisms to provide access to ion trap quantum computing resources. Across DOE SC, the House Mark allocates ≥$60M for the Reaching a New Energy Sciences Workforce (RENEW) and Funding to Accelerate Inclusive Research (FAIR) initiatives; $100M across SC for the Energy Earthshots, with $25M from ASCR; and ≥$35M for the Establish Program to Stimulate Competitive Research (EPSCoR). The House Appropriations Committee expressed

disappointment with DOE SC's lack of support for robust user facility operations in the PBR and directed DOE to prioritize user facility stewardship in FY23 and future budget requests.

Thus far, ASCR has released FY22 solicitations for the following programs: Randomized Algorithms for Combinatorial Scientific Computing; Mathematical Multifaceted Integrated Capability Centers (MMICS); Data Visualization for Scientific Discovery, Decision-Making, and Communication; Management and Storage of Scientific Data; RENEW; Advancing Computer Modeling and Epidemiology for Biopreparedness and Response; Advancing Computer Modeling; and Scientific Discovery through Advanced Computing-5 (SciDAC-5) Partnerships in Nuclear Energy (NE), Earth System Science (Biological and Environmental Research [BER]), High Energy Physics (HEP), and Nuclear Physics (NP). ASCR also combined funding across several programs to provide up to $20M for the Exploratory Research for Extreme-Scale Science (EXPRESS) program in FY22.

Recipients of the Society for Industrial and Applied Mathematics (SIAM) Fellowship, Ernest Orlando Lawrence Award, and the ASCR Early Career Research Program (ECRP) Award have been announced.

In relation to USICA, ASCR community members briefed staffers in key emergent technology areas, including AI, QIS, materials and chemistry for clean energy, microelectronics, and the bioeconomy in 2022.

ASCR Facilities Division is leading the planning process for a DOE SC integrated research ecosystem. Efforts during this first year have centered on collecting input and are transitioning to a design phase towards one or more Integrated Research Infrastructure (IRI) Architecture Blueprints.

The DOE laboratories issued a joint request for information (RFI) in June 2022 soliciting computing technology vendors' input concerning future technologies, Advanced Computing Ecosystem components, non-recurring engineering, and other factors that will help inform DOE's approach to the next generation of supercomputing systems in the 2025-2030 timeframe.

ASCR's next steps related to software stewardship include: 1) finalizing the targeted scope for potential FY23 activities; 2) defining the relationship between those activities and synergistic activities in the Facilities, Research, and Advanced Computing Technologies (ACT) Divisions; 3) Developing a funding opportunity announcement (FOA); 4) developing a dear colleague letter (DCL) in the case of a continuing resolution; and 5) working with the ECP, ASCR facilities, and other stakeholders to enable a common understanding of how all will contribute to the overall process.

Science Highlights addressed application-oriented performance benchmarks for quantum computing; co-design for energy-efficient embedded neuromorphic computing; and the Argonne privacy preserving federated learning framework (APPFL).

ASCR honors the many contributions from Dr. Ewing Lusk, who passed away in 2022.

DISCUSSION

Lethin inquired about the status of the IRI Architecture Blueprint. **Helland** replied the document is not finished. Efforts will include National Science Foundation (NSF) facilities.

Cerf raised challenges associated with federated learning and data access. Combining different models without the original raw data, which is locally collected, can lead to incorrect results as opposed to aggregating data and then building the models. Bias may evolve from having incomplete data. **Helland** replied all federal agencies, including DOE, are focused on

addressing bias. Foundational research to explain AI results is important, especially when AI is being used to generate predictions and assist in decision-making. DOE has issued related calls.

Reed remarked the EPSCoR program is the center of discussions regarding the geography of innovation, equity in talent cultivation, and distribution of federal dollars. **Helland** stated EPSCoR funding resides in the Basic Energy Sciences (BES) program, but all programs are focused on funding more projects in EPSCoR states. If ASCR funds are distributed to an EPSCoR state, BES may add funding. There are pockets of innovation throughout the country, and DOE is working to broaden participation across the board.

Crivelli raised data integration across agencies. **Helland** agreed this is an important issue requiring interagency agreements. The COVID-19 Consortium taught all parties that data access is vital. The Office Science and Technology Policy (OSTP) is managing agreement documents regarding the National Strategic Computing Research (NSCR) in the event of another pandemic.

Germann sought clarification on the House's comment related to DOE facilities. **Helland** said the comment was not directed at ASCR, which operates its facilities continuously. Other SC facilities are operated for a certain number of weeks per year. The House felt funding requests for these facilities was not high enough.

PROGRAM RESPONSE TO REPORT FROM THE COMMITTEE OF VISITORS, Ceren Susut, Research Division Director, Advanced Scientific Computing Research

The ASCR Research Committee of Visitors (COV) met virtually in August 2021 to review the Applied Mathematics, Computer Science, Computational Partnerships, and Research and Evaluation Prototypes programs for FY16-FY19. This review period was characterized by the launch of the ECP with continued investments in core basic research; consideration of post-ECP capabilities like QIS and AI; expanded partnerships; and increased investments in ECRP and the CSGF programs to grow the workforce.

The COV issued 21 recommendations, with some recommendations repeated across programs and/ or ASCR as a whole. As this was a retrospective review, ASCR has already begun implementing some recommendations.

Recommendations and ASCR responses regarding solicitation, review, recommendation, and documentation of proposal activities addressed pre-application review processes; access to Portfolio Analysis and Management System (PAMS) statistics; diversification of principal investigators (PIs); and communication strategies to keep the community informed of ASCR opportunities and directions.

Recommendations and responses concerning the monitoring of active projects and programs centered on support for promising early career investigators and establishment of metrics to evaluate long-term projects and math centers.

Recommendations and responses focused on the breadth and depth of portfolio elements related to communication of programmatic shifts; re-establishment of university-based small group and single-PI programs; clear articulation of SciDAC goals and technical shifts; and processes to encourage experimentation in quantum testbeds.

Recommendations and responses about emerging challenges in high performance computing (HPC) and DOE missions addressed ASCR's North Star research vision and metrics of success; exploration of new and emerging research areas; experimentation in quantum test beds; and defining five- and ten-year success targets for program outcomes.

Recommendations and responses about national and international standing raised holistic SciDAC documentation; presentation of program stories to the COV; and continued emphasis on expanding CSGF program diversity.

DISCUSSION

Mesirov inquired about ASCR considerations when scoring and funding proposals from early career investigators. **Susut** replied ASCR supports the SC-wide ECRP. Early career status is also considered as a program policy factor, among many others, for ASCR FOAs.

Berzins asked whether the COV discussed community input regarding the size of programs, including the Applied Mathematics program. **Susut** did not recall a specific discussion centered on the math program. However, challenges related to the size of the core research program were raised with respect to the launch of the ECP and other initiatives. The funding size of the core research program increased during the time of this retrospective study.

George Michelogiannakis (LBNL, via chat) asked about the ECRP's historical acceptance rates. **Susut** will provide information from ASCR's website.

Lethin requested more information about the COV's North Star recommendation in relation to success metrics and ASCR's charter. **Susut** interprets this recommendation as the need for better communication of ASCR's research priorities to the community. ASCR is now providing regular public updates on research priorities, accomplishments, and future plans at least on a yearly basis at ASCAC meetings. Future efforts will build on this activity. Metrics are meant to assess progress towards priorities and articulate accomplishments. ASCR is using NAS indicators of excellence, relevance, and leadership, but more work is needed in defining metrics for national and international standing.

James Ang (PNNL), questioned strategies for reconciling COV recommendations for five- and ten-year progress and performance goals with the fact that many FOAs are for shorter award periods. The DOE Computational Research Leadership Council (CRLC) offered recommendations addressing integrated research areas with the potential to span many laboratories. **Susut** appreciated these comments. ASCR has supported AI, QIS, and topic areas in applied mathematics for long periods of time during which several related FOAs were issued. There are opportunities to review program outcomes for these and other ASCR topics.

UPDATE ON EXASCALE SYSTEMS – FRONTIER, Justin Whitt, Oak Ridge National Laboratory

Frontier is ORNL's newest supercomputer and the latest in a line-up of accelerated-node computing models. In 2022, Frontier ranked first on the Top500 for a 1.1-exaflop (EF) performance and first on the High Performance LINPACK for Accelerator Introspection (HPL-AI) benchmark after achieving 6.88-EF. A single Frontier cabinet secured the first place on the Green500 list for a 62.04 gigaflops/watt (GF/W) power efficiency followed by the full system in second place operating at 52.23 GF/W.

Frontier's 74 Olympus computer racks collectively occupy 4K square feet and contain 9,408 nodes. Each rack holds 128 nodes, each comprising one AMD Trento central processing unit (CPU) with 512 Gibibytes (GiB) of double-data rate fourth-generation synchronous dynamic random-access memory **(**DDR4) memory, four AMD MI250X graphics processing units (GPUs) with 128 GiB of high bandwidth memory (HBM) per GPU, and four Cassini network interface cards (NICs). Overall, Frontier features 9.2 petabytes (PB) of memory with 37 PB of local node storage and 716 PB of center-wide storage. The system's Cray Slingshot network has dragonfly

topology. Compared to that of the Jaguar system circa 2009, Frontier's 15-megawatt (MW) /EF operations has realized a >200x reduction in energy requirements due to targeted ASCR investments in reducing chip power consumption and use of a warm water cooling system.

Facility preparations for Frontier were mostly completed by June 2021. Approximately 30 offices, eight laboratories, and 20K square feet of Titan's old data center were repurposed. Sturdier floors were installed to accommodate Frontier's 8K-pound cabinets. Additional cooling towers and power supply lines were added to achieve 50 MW of cooling capabilities and 40MW of power, respectively.

During Frontier's build, chip shortages and supply chain disruption increased order lead times by six to 12 months. Heroic efforts on the part of AMD and HPE reduced the anticipated year delay for parts delivery to two months. The last of the 60M parts required for Frontier arrived on October 18, 2021, the same day the last cabinet was assembled. Wiring between all nodes ensued, requiring 81K cables. System debugging and tuning followed at a rapid pace. Completion of Frontier testing is expected in September 2022.

To prepare for Frontier access, early science teams in the Center for Accelerated Application Readiness (CAAR) and ECP received access to Frontier's Test and Development System (TDS), Crusher, in November 2021. To date, the majority of ECP Key Performance Paramater-1 (KPP-1) and KPP-2 applications are already running on Crusher, with several ready to operate on Frontier. Select applications have also run on Frontier, with highlights showcasing performance of the CoMet, LSMS, and Cholla applications. Following final testing, application teams will have full system access for final application readiness testing. Production is scheduled and user science programs are scheduled for January 2023.

DISCUSSION

ASCAC Members appreciated the impressive work put into Frontier.

Cerf commented Google reduced cooling-associated power demands for its data centers by transitioning from manual to AI-based management of the cooling system. **Whitt** expressed interest in this approach. Frontier's cooling system incorporates many sensors along with distribution points, presenting opportunities for optimization.

Berzins asked for more information regarding Frontier's bandwidth performance between the HBM and GPU architecture. Online information suggested performances of up to 1,600 gigabytes/second (GB/s). How does this value compare to leading-edge CPUs; this figure is important for comparison when considering lower intensity arithmetic applications and is indicative of application speedups. **Whitt** observed memory bandwidth is a core design constraint. As much bandwidth is provided as possible, but there are always applications that will be bottlenecked. **Scott Atchley** (ORNL, chat) noted Summit streams 800 GB/s of the peak 900 GB/s bandwidth. Frontier is expected to achieve ~80-90% of the peak bandwidth, which is 3.27 terabytes (TB)/s per GPU and ~1.3 TB/s per chiplet.

Arthur (chat) requested more information about the Summit-node to Frontier-node speed-up factor used to evaluate application readiness. **Atchley** (chat) stated the per-node floating point operations per second (FLOP) increase from Summit to Frontier is 3.9x. The aggregate GPU memory bandwidth increase is 2.42x.

Dean inquired about Frontier's life expectancy and current error rates. **Whitt** explained Frontier's life expectancy is approximately five to six years. This timeframe allows for delivery of a new project around the five-year mark and one year of operational overlap to support

transition. Memory errors have been observed, and AMD is tracking these down. Some are difficult to find because they are hard to reproduce.

Seidel raised interagency cooperation and voiced support for joint DOE and NSF efforts in hardware and software funding. **Helland** stated NSF has participated in ECP reviews since project inception. DOE and NSF sit on many of the same committees, and ECP has briefed NSF on ECP software as a part of sustainability efforts. NSF is implementing ECP software. Going forward, DOE would like as many people as possible to utilize ECP software. An even closer working relationship between the two agencies in the future is likely. **Gregurick** added the ECP Cancer Distributed Learning Environment (CANDLE) project represents a DOE-National Cancer Institute (NCI) collaboration. The project seeks to characterize molecular dynamics of the RAS signaling protein, which is associated with 30% of human cancers; repurpose drugs for cancer treatments; and evaluate patient trajectories using a compendium of medical information. **Whitt** added one of the ASCR Leadership Computing Challenge (ALCC) projects is connected to LLNL's Cancer Moonshot initiative.

Matsouka remarked Fugaku's memory is the least reliable system component; there have been non-standard fixes to mitigate problems over the last two years. If permitted, Fugaku may share this information to support Frontier. What are the acceptance criteria for Frontier and application performance? **Whitt** replied the bigger systems get, the more components they have, and the more challenging acceptance is. Frontier's acceptance testing will begin soon and will consider three traditional components. Functionality is generally a straightforward ticking of boxes. Performance will address the Collaboration of Oak Ridge, Argonne, and Livermore (CORAL) benchmarks. Vendors are contracted to deliver on benchmark figures of merit (FOMs) related to scalable science, throughput, and data science and AI/ ML. System stability will subsequently be assessed through a rigorous multi-week test.

Taylor inquired about power and energy efficiency FOMs for applications. **Whitt** said this is a burgeoning effort. With the current level of instrumentation, power-aware scheduling and programming environments are possible, but the project is not there yet.

Cerf asked about evaluating and updating ECP algorithms for parallelism. **Whitt** confirmed applications have been re-examination. **Helland** elaborated ASCR has funded 24 ECP applications so experts can focus on refactoring applications as opposed to producing science. This is part of the ECP's legacy. **Michael Heroux** (chat) advised much of the ECP investment in applications, libraries, and tools was in new math formulations, new algorithms, and new implementations of existing algorithms. There is still much to do, but ECP accelerated progress.

Bergman inquired about future needs in technology innovations. Were there high-risk areas in building Frontier, and how did these pan out? **Whitt** highlighted energy efficiency as an important area. Algorithmic approaches to energy savings will become increasingly important as hardware curves flatten with Moore's Law. Frontier's use of a 32°C water cooling system was risky, but the savings were too good to pass on. A recent hot day in Oak Ridge, Tennessee offers evidence for proof of concept.

Chen revisited interagency collaborations. Some of the ECP applications have participated in NSF calls for Science and Technology Centers with integrated partnerships engaging universities and industries. Methods to enable more DOE laboratories to participate in large, integrative teams will assist in transferring software to NSF. **Whitt** appreciated this comment. ASCR facilities personnel have participated in NSF Major Research Equipment and Facilities Construction (MREFC) reviews. Reciprocal efforts are leading to collaboration.

ASCAC DISCUSSION ON THE FUTURE OF ADVANCED COMPUTING
Cancelled.

Reed dismissed the meeting for lunch at 12:15 p.m. and reconvened at 1:45 p.m.

GENERAL ELECTRIC (GE) COLLABORATIONS WITH DOE AT THE EXASCALE,
Richard Arthur, General Electric Research and ASCAC

GE makes consequential, regulated products that support critical infrastructure and require a long field life. Research activities require collaboration with government agencies and other strategic partners. GE has employed modeling and simulation to see, understand, and to predict outcomes across diverse products over the last two decades. As of 2022, GE has received ~14 Innovative and Novel Computational Impact on Theory and Experiment (INCITE), ~22 ALCC, and about six HPC4Manufacturing awards, engaged in multiple pan-lab collaborations, and produced numerous post-grant publications. DOE's production of state-of-the-art hardware, the accompanying software environment and benchmarks, and feasibility studies pave the way for GE to tune hardware and software for specific needs, conduct validation studies at scale, and tackle higher Technology Readiness Level (TRL) problems for proprietary cases.

Examples of GE using models "to see" included identification of a wake caused by strut placement in an aircraft engine on Jaguar; aeroacoustics on Mira; and ice formation physics on Titan. A 2016 GE poster presented showcased elucidation of airplane engine combustion, high- and low-pressure turbine dynamics, and exhaust in collaboration with ORNL, LLNL, and ANL.

Selected case studies of harnessing models "to understand" included identifying and remedying the cause of thermo-acoustic instability in never-before simulated multi-combustor interactions with ORNL and ANL; optimizing wind farm design in collaboration with ANL; understanding the impact of low-level wind jets on offshore wind farm performance and reliability with ORNL; and evaluating additive manufacturing defects through the HPC4EnergyInnovation program.

Highlights of deploying models "to predict" included forecasting the impact of farm-scale wakes on down-flow wind farms with ECP; and anticipating detailed flow physics on airplane turbine blades by understanding how behaviors vary with Reynolds number. The latter example is applicable to a multi-partner effort to develop Revolutionary Innovation for Sustainable Engines (RISE) open-rotor airplane engines, which present a pathway to hydrogen fuel. Due to limitations in wind tunnel size, it is not possible to test the full-size product. Building on previous INCITE 2021 work and early Perlmutter science, Frontier will enable product-scale flight testing through an ALCC award.

With increased data storage capabilities, there are also opportunities to save targeted, high-fidelity simulation results for ML applications. Re-running simulations to fill gaps, and then retraining models offers a pathway to creating a bespoke surrogate model factory.

DISCUSSION

Cerf posed a question about anisotropic dynamics and zero gravity. **Arthur** recalled a 2010 instance of a spiral effect that did not manifest in a 1/8 engine symmetry model because there was not enough problem geometry. Hypersonic and zero gravity conditions require rethinking machinery. Measurement technologies go hand-in-hand with generating models.

Reed referenced a quote from the statistician George Box. **Arthur** appreciated the quote's applicability to understanding how all articulated or a data-driven machine-learned models are wrong but also useful.

Seidel inquired about using modelling and simulation approaches in lieu of physical tests for regulatory processes. **Arthur** considered a U.S. Council on Competitiveness focus group ten years ago that addressed this topic with regulators. The Food and Drug Administration (FDA) has advanced thinking in this area and utilizes *in silico* regulation. There are opportunities for the Federal Aviation Administration (FAA) to learn from and leverage such practices. Currently, audits and tests are high consequence and can be adversarial. Modeling has the potential to reshape the entire posture of regulatory science.

Berzins invited comments on the next generation of software required for next-generation machines. **Arthur** said the discussed simulations use a higher-order large eddy simulation (LES) code built in collaboration with the University of Kansas. Making this code scalable for efficient runs on large machines required a lot of effort. Ongoing work is integrating other ECP codes to confer additional capabilities. Similarly, there are adjacent scalability problems in finite element analysis and structural mechanics. There are obstacles when vendors perceive only a small portion of the market requires at-scale and high fidelity simulations, and GE must pull together its own codes or rely on the national laboratories. A dream would be to have a multi-scale, multi-physics toolbox that would enable pulling together different system parts, generating co-simulations, and running analyses in a consistent manner.

Ang asked about wind farm simulations and pointed to potentially interested community members in Norway. Many ECP applications are looking for new use cases and new sponsors, and GE's work may present opportunities to integrate ExaWind with ExaLearn. A related PNNL project is seeks to use 5G advanced wireless communication to manage a wind farm in real time. This may be an opportunity to collect data on operations and implement low-latency training and learning local to the farm. **Arthur** explained the relevant wind farm example used AMR-Wind and Nalu-Wind. Beyond wind farms in Norway and the Baltic, there are examples off the coast of Scotland and Long Island, New York. There are opportunities to consider novel physics-inspired neural networks for industrial applications to explain anomalies in complex time series data.

Chen cited an AMReX code that has also been coupled closely with ExaLearn for development of anomaly detection methods. Acquiring either *in situ* machine learning and reduced order modeling or higher moment statistics to guide surrogate models in instances where there are transient events or chaos, such as turbulence, would be great. Capturing anomalies and transients through forward simulation is an alternative. **Arthur** concurred. Surrogate models must be aware of what lies within data. If there is insufficient data for eigenvalues, phase transitions, and other factors to be discovered through learning, they will be invisible. Embedding sensors to provide feedback into parameterizable models will allow for continual learning via the digital twin concept.

BESAC ASSESSMENT ON INTERNATIONAL STANDING, Cynthia Friend, President of the Kavli Foundation and Jeffrey Miller, Harvard University

The BESAC International Benchmarking Subcommittee evaluated the status of BES's research, capabilities, and workforce prospects in the context of intensifying globalization in a report titled *Can the U.S. Compete in Basic Energy Sciences?*

The Subcommittee leveraged benchmarking methods recommended by the NAS and the American Academy of Arts and Sciences. A Scientific Areas subteam reviewed BES Basic Research Needs (BRN) studies dating from 2010 and identified five BES priority areas to delineate the report's topical scope. To evaluate the selected areas, the subteam engaged BRN chairs and other experts in deep-dive discussions to capture interviewee perceptions of each area's global status. International publication and conference metrics were gathered and analyzed to complement expert opinion. Finally, this information was compiled with award records, webinars promoted by professional societies, and other metrics. A Strategies subteam used a recursive interview process to identify U.S. strategies. The subteam generated hypotheses by consulting over 50 early career scientists and individuals representing leadership from U.S. national laboratories; NSF, private foundations; universities; U.S. and international industries; and international research facilities. Hypotheses were tested via additional consultations. The Subcommittee additionally selected nine sidebar stories to emphasize important findings, adding human interest to the report and better speaking to non-technical audiences. Before assembling the report, an overview of the methodology and findings was shared through several town halls to elicit community feedback. In addition to Subcommittee member contributions, the report benefited from science writing, data and analysis, library science, and graphics support.

Major findings indicate an overall downward trend in U.S. competitiveness in all research areas. Research in Asia is surging, with this trend primarily driven by investments in China. U.S. advanced research facilities are no longer unique. Obtaining support for mid- and small-scale instrumentation is difficult. Finally, global competition for scientific talent is fierce. Possible strategies for maintaining U.S. leadership include 1) increasing investment in BES research; 2) augmenting investment in computation, data analysis methods, and computer hardware and architecture; 3) boosting support for early-career and mid-career scientists to enhance U.S. competitiveness for talent; 4) balancing the need for new facilities with support for existing facilities; and 5) better integrating research across the basic-to-applied-to-industrial spectrum.

DISCUSSION

Reed invited recommendations for transitioning from acknowledgement of issues identified by benchmarking reports to action supporting funding for basic research. **Friend** emphasized the importance of telling stories when engaging with lawmakers and the public. There are many examples of how fundamental science, followed by technological breakthroughs, has led to economic gain. For example, transistors and QIS are results of government-funded research. The latter is now contributing to a number of products like new TV screens. Everyone argues for more money. Instead, it is important to argue for talent; without it, the U.S. will not be able to compete internationally. For example, following World War II, emigration to the U.S. had a huge impact on the country's scientific and technological prowess. Issues of international competition, visas, and immigration need to be part of the discussion. There is no easy answer. Outreach and having data help, but stories help people understand why basic research is relevant to them and their constituencies.

Reed agreed with comments regarding international talent and segued to challenges in attracting domestic talent, which entail diversity and inclusion issues. In many disciplines, the majority of graduate students are international. On some level, this is good because it means the U.S. is attracting talent. However, it indicates there is not enough domestic talent. Pointing to pressure on talent from international sources, especially in computing, **Friend** opined it is important to highlight international talent as a good thing; without it, the U.S. would be in big

trouble. Inspiring domestic talent will require exposing people to the possibilities of scientific careers early on. Exposure mechanisms do not have to be complicated and can be as simple as conversations. This vast issue was beyond the scope of the BESAC benchmarking report.

Cerf observed numbers are against the U.S.; there are 7B people in the world, 330M people in the U.S., and talent is distributed uniformly. The U.S. is a place where people can exercise talent, and it makes sense to try to retain talent. Even if talent cannot be retained, the U.S. can leverage the work international individuals produce during their stay. Beyond research, the NSF's new directorate offers another pathway, aside from the Small Business Innovation Research (SBIR) program, to move research from the labs to the market. Money and support are needed to address this challenge across agencies.

Seidel pointed to report findings that the U.S. is potentially falling behind in computational and data science. Finding support for this key area challenging. Rethinking support may be important to the U.S.'s future. **Friend** finds the joint DOE programs and the availability of computation time to be very effective. For those without coding expertise or experience using computational facilities, support is important. Finding more ways to bridge across programs will help both science and to expand the talent pipeline. DOE generally does a good job of bridging, but there is room for improvements.

Taylor asked about forums and advertising for community input. **Friend** said the subcommittee could not conduct a survey. Instead, community input was collected through advertised sessions at well-known meetings. All sessions were virtual, which may have increased participation. Those unable to attend were able to provide input online. Widening this approach to other professional societies may help capture greater diversity.

OPTIMIZING THE PERFORMANCE OF FUSION REACTORS AT EXASCALE, Noah Mandell, Massachusetts Institute of Technology and DOE CSGF Alumnus and Howes Scholar

Viable commercial development of fusion, a clean and virtually limitless energy source, will be a game-changer for the health of the planet. The Joint European Torus (JET) tokamak, which uses magnetic confinement fusion, recently set a record of 59 mega joules (MJ) of fusion power over five seconds. The National Ignition Facility (NIF) at LLNL uses inertial confinement fusion and produced 14 kilojoules of energy, which was less than the 1.8MJ laser used to implode pellets, but more than the x-ray energy absorbed by the pellets. The international ITER tokamak experiment, scheduled to begin operations in ~2025, will produce 500 MW of fusion power from 50 MW of heating power with 400+ second pulses. Commonwealth Fusion Systems (CFS) will begin operating the SPARC tokamak in 2025, and though smaller than ITER, SPARC's stronger magnetic fields are anticipated to yield a similar energy performance.

Tokamak fusion challenges include loss of core heat due to turbulence and potential damage to device walls from boundary plasma heat exhaust. To make problems tractable for computational resources, multi-scale numerical algorithms and theory are needed. The GX code models core turbulence with spectral methods on GPUs, and couples to a transport solver like the Trinity code to form a multi-scale core transport model. The Gkeyll code models boundary turbulence with discontinuous Galerkin methods and a first-of-a-kind kinetic scheme that includes magnetic fluctuations. Collectively, these codes present a whole-device transport model that will enable direct study of interactions between core confinement and boundary exhaust under different conditions. Adding more physics to models will allow a true predict-first modeling capability. Optimization for exascale will require additional work to ensure whole-device model calculations are sub-exascale and can compute FOMs. Approaches may consider

simultaneous variation of ~20 shaping parameters, potentially via a parallel optimization algorithm. There is potential to use a hierarchy of models of varying speed and accuracy and incorporating ML to narrow the design space. Models may also build in economic and environmental safety factors.

DISCUSSION

Cerf raised possible connections to TriAlpha ML algorithms which can detect and respond to plasma instability before a human can. **Mandell** discussed potential for feeding turbulence knowledge into models to improve stability control and enhance confinement. No one is currently working on this.

Lethin asked about porting codes to Frontier. **Mandell** said efforts are currently targeting Perlmutter. The project is considering use of the HIPIFY tool, which translates CUDA sources to HIP and makes code portable to both NVIDIA and AMD devices. Hopefully, HIPIFY will also make codes' performance portable.

ASCR LEADERSHIP COMPUTING CHALLENGE PORTFOLIO FOR THE 22-23 ALLOCATION YEAR, Jordan Thomas, ASCR

The ASCR HPC and Networking Facilities offer world-leading capabilities spanning supercomputing, data analysis, data transport, and testbeds. Access to the ALCF and OLCF is managed through the INCITE, ALCC, and Director's Discretionary (DD) programs, with 60%, 30%, and 10% of resources allocated to each program, respectively. Access to NERSC is managed through the Energy Research Computing Allocations Process (ERCAP), ALCC and DD programs, with 80%, 10%, and 10% of resources dedicated to each program, respectively.

The annual ALCC solicitation focuses on DOE priorities, including SC priorities, national emergencies, interagency partnerships, and industry. The program seeks to broaden the community of researchers capable of using HPC resources. Awards target small-to-medium allocations, and proposals may request no more than 25% of any resource. The cross-Department ALCC Working Group receives preproposals and conducts a peer review of full proposals. Considerations for the next ALCC cycle include increasing Working Group representation across the DOE; incorporating panel groups to address review of cross-cutting initiatives; requesting applicant and reviewer demographic data to track and improve diversity; tracking new users and renewal applicants; supporting out-of-cycle ALCC proposal processes; better defining the ALCC within the ASCR allocation space; and offering applicants more resources.

In FY22, the ALCC received 100 pre-proposals. Of these, 98 preproposals were encouraged, and 87 full proposals were submitted. A total of 45 awards were announced in July 2022, with recipients at national laboratories (52%), universities (42%), industry (4%), and other entities (2%). The total numbers of FY22 preproposals, full proposals, and awards were slightly higher than those from FY21. FY22 awards were distributed across research domains with ten conferred in Nuclear Physics; nine in Physics; seven in Earth Science; six in Biology; five in Materials Science; four in Fluid Dynamics; and three in Chemistry. Across these projects, a total of 7M node hours were awarded on Summit (OLCF); 785K node hours on Polaris (ALCF); 6M node hours on Theta (ALCF); and 2.65M node hours on Perlmutter (NERSC). Additional hours were also granted to select projects on Cori (NERSC) and Frontier (OLCF).

Featured projects addressed seismic hazard modeling; particle-in-cell simulations of beam-driven, field-reversed configuration plasmas; climate change mitigation through zero

carbon fuels; and privacy-preserving transformer models for clinical natural language processing.

DISCUSSION

Taylor inquired about the ALCC's acceptance of educational proposals, including those for classroom training. **Chalk** responded the ALCC is the correct venue for medium-sized educational proposals. Smaller proposals targeting training would be better served by the DD program. DOE is working on a secure container to circumvent issues related to export control and foreign students. The magnitude of future requests must be understood; ASCR cannot promise every American institution class time on Frontier or Summit. Any ASCR-funded RENEW project would have access to an ERCAP allocation.

Giles requested more information about project flow across different allocation programs. **Thomas** indicated funding trajectories vary by project. Many ALCC projects previously received DD or ERCAP funds to scale up their codes. ALCC does not provide individualized support for scaling codes. This limitation for new users is being critically examined. Some users previously had INCITE allocations. Others receive multiple ALCC allocations. ERCAP is only open to SC projects, and SC PIs interested in leadership computing facility resources eventually transition to ALCC or INCITE.

Matsouka asked about tracking research evolution with allocations. On Fugaku, experienced groups that receive the greatest allocations are typically the most innovative. **Chalk** explained this was motivation for forming the cross-DOE Working Group. Members are asked to list the number of years programs have invested in codes and to provide internal knowledge about project status. Other than retrieving and evaluating all ALCC records, there is no obvious mechanism for capturing how investments and allocations have impacted project trajectory. Questions may also be added to the proposal to allow applicants to share project-specific considerations.

Leung appreciated ALCC plans to collect demographic data. Once baselines are established, how will ALCC attract both diverse users and reviewers? **Thomas** stated the ALCC will first focus on collecting baseline data about the current user base. Collected information will address demographics, new HPC user status, and repeat submissions. Once a foundation has been laid for needed data, the ALCC will craft a plan to broaden the HPC community, which is a focus of the SC and ASCR.

Matsouka questioned how increased computing capacity driven by Frontier will affect allocations. **Thomas** advised with increased capacity, there is the potential to support more ALCC projects and broaden the user community. However, future conversations with facilities will help find the balance between allocating hours to projects and not overloading facilities with too many projects.

PUBLIC COMMENT

None.

Reed dismissed the meeting for the day at 4:21 p.m.

FRIDAY, JULY 22, 2022

OPENING REMARKS, **Reed** convened the meeting at 10:00 a.m.

DYNAMIC MODELING AND OPTIMAL SCHEDULING OF CHEMICAL PROCESSES PARTICIPATING IN FAST-CHANGING ELECTRICITY MARKETS, Morgan Kelley, Dell Technologies and DOE CSGF Alumna and Howes Scholar

The gap between the availability of renewable energy and grid demand drives electricity prices. Encouraging industrial consumers to overproduce and store products on a regular day when renewable sources are high can shift the energy demand curve. Developing computational models to guide industrial behavior is challenging because many time scales, ranging from seconds to months, must be considered. Scale-bridging models enable combining longer scheduling horizons with the shorter control times needed for the fast and frequent changes to maximize industrial profits while parallel computing enables rapid calculations. Approaches may also incorporate Autoregressive with Extra Input (ARX) models and must account for data errors due to faulty plant sensors.

Presented examples demonstrated successful demand-response (DR) model applications to a small-scale case study of cryogenic air separation and a large-scale case study of an industrial gas production plant. There is potential to transfer current models to new plants using Kalman filters. A further study of grid-based emissions in California indicates DR-driven models consistently lower emissions even though these models aim to minimize operating costs. Models targeting lower emissions can increase operating costs during the summer months.

Changes to industry operating habits involves little to no capital expenditure, and thus DR models have huge potential to mitigate grid instability and reduce emissions while saving companies electricity costs. Beyond industrial plants, DR models may have select applications in time-of-use pricing for residential and commercial entities. DR also may play a role in remote computing tasks, such as flexibly scheduling the run times and locations of large problems based on grid conditions in different places. Advances in computer technology, models, and algorithms will further promote efficient solution of large-scale DR problems.

DISCUSSION

Arthur asked about integrating incentive models to change consumer behavior. **Kelley** replied time-of-use pricing incorporates incentives because such models can save a significant amount of money. Minimizing costs may inherently lower emissions for industries. Models focused on minimizing emissions, however, may cause greater electricity prices that are higher than carbon taxes.

Cerf recalled a scenario from Hawaii where a power company became economically unviable due to a solar reward program. How can incentive reward models be chosen to ensure system stability? **Kelley** observed Hawaii presents an interesting problem because it is not connected to other grids. Most of the U.S. grids are connected, with the exception of those in California and Texas. For such exceptions, if solar energy is overproduced, the cost of electricity will dip negative, and the company will be paying people to use electricity. Also, if many companies begin implementing rewards programs, prices and demand will shift, creating a game theory problem. **Reed** observed utility company business models, currently predicated on growth in energy demand, are encountering new dynamics.

Svore expressed interest in other industrial application scenarios. **Kelley** cited application of models to ammonia and steel plants as well as commercial and residential scenarios. However, it makes the most sense for industrial sectors to deploy these models.

Crivelli inquired about combining data samples across models and then applying results to new plants. **Kelley** explained the team had only one plant's data, which was used to generate

related data presuming operation by the same company. Common filtering was used to see if models could be updated. This is likely a continued direction for this project, which is an ongoing, multi-university effort with Lindy Electronics.

VIEW FROM WASHINGTON, Asmeret Berhe, Director of the Office of Science

Sustained support for research and innovation across DOE SC's broad physical sciences portfolio, including facilities and infrastructure, is key to advancing scientific discoveries and technology development through the lens of inclusive excellence and economic growth. ASCR is essential to DOE's core mission, and Dr. Berhe is committed to advancing all SC programs as well as emerging technologies such as AI, QIS, and microelectronics that enable progress in all areas of science. Only through community-driven strategic planning can the nation's and world's scientific priorities be reached.

ASCR's recent accomplishment of producing the first exascale computer placed first by the Top500 list promises great advances in all areas of science. Crucially, AI and advanced computing will support the Administration's goals of addressing climate change and developing clean energy technologies. Higher resolution and more advanced climate models, including the SC's groundbreaking Energy Exascale Earth System Model (E3SM), are one of the most important applications for new exascale capabilities. Realizing basic and use-inspired research and technologies in priority areas requires strengthened partnerships with the Applied Energy Offices and leveraging connections with the scientific community as a whole.

Healthy stewardship of the DOE national laboratories and user facilities is necessary to expand their roles as regional hubs for economic opportunities and community benefit in partnership with federal, state, and local governments; universities; and the private sector. Partnerships through the National Virtual Biotechnology Laboratory (NVBL) delivered life-saving COVID-19 breakthroughs. Ongoing efforts are developing capabilities for a biopreparedness program across the SC portfolio to which ASCR's contributions will be central.

Dr. Berhe looks forward to results from the RFI to vendors soliciting feedback on post-exascale computation. ASCR has a rich history of partnering with industry to benefit U.S. research while making U.S. companies more competitive. As Congress considers legislation for strategic investments in this area, the SC is optimistic ASCR and the national labs will continue to play a leading role.

Beyond internal DOE collaborations, advancing cross-agency collaborations to maximize federal research and development (R&D) investments will benefit the broader U.S. science ecosystem. For example, DOE looks to build upon ASCR's leadership in the National Quantum Initiative, involving collaboration with the NSF, Defense, National Institute of Standards and Technology, and other federal agencies. Expanding international collaborations in a responsible manner will advance Administration priorities through partnerships that maximize scientific access while maintaining research security.

Central to all these efforts is the continuing and vital priority of increasing the accessibility of DOE SC-funded efforts through the principles of belonging, accessibility, justice, equity, diversity, and inclusion (BA JEDI). The RENEW initiative will significantly expand training opportunities for undergraduate and graduate students from underrepresented and underserved groups. The CSGF has been a crucial to ASCR's workforce development pipeline and has been recruiting more diverse cohorts in recent years, with great strides made in recruiting women. Justice40 is ensuring SC is meeting the needs of communities most at risk. Broadening participation efforts must be inclusive of all groups and institutions currently underrepresented

and underserved in the SC portfolio, including Minority Serving Institutions (MSIs), Historically Black Colleges and Universities (HBCUs), as well as institutions and states that have not attracted significant federal research funds. Broadening participation to tap into all of America's talent will require invigoration of existing and development of new communication strategies to better share scientific successes within DOE, with Congress, and with the public. All are encouraged to take meaningful steps in broadening participation and giving back to the public that has supported the scientific careers of many.

DISCUSSION

Reed appreciated Dr. Berhe's remarks and asked how ASCAC can best support SC's vision. **Berhe** said SC looks to FACA committees to advise on science priorities. While all would like more funds to support science and technology ideas, SC must work within budget guidelines. SC also welcomes innovative ideas for broadening participation. Climate change cannot be ignored and must be considered in the future of the field. There is funding directed towards each of these efforts within and across agencies. SC also looks forward to hearing the results of ASCAC's RFI.

Cerf asked how DOE can ensure HPC is applied to topics that may not be flashy, but are important. **Berhe** underscored the importance of this question and how computational sciences as a tool are advancing many other SC areas. **Helland** remarked the CSGF helps bring innovative ideas to the forefront, especially when fellows take applied math and computer science courses along with other domains. The facilities run outreach programs that have supported activities ranging from modeling truck modifications for energy savings to modeling flooding in relation to cement and insurance. SBIR considers how to transfer software developed through SciDAC, ECP, and other DOE programs to small businesses.

Citing the exascale initiative as an example, **Giles** commented SC manages to accomplish large projects over a sustained period. Doing so, however, is challenging and requires commitment of enough resources and partners to succeed. Is there a way to make such processes routine and sustainable? These considerations similarly apply to workforce diversification. BA JEDI efforts as goals must be sustained over a long period. **Berhe** agreed. There are times when budgets are tight, but right now is not one of those times. SC has incredible support for its science mission and to enable facilities, setting the stage for years to come. It is crucial all work together to take advantage of the opportunities currently available so all can continue to obtain the resources needed to push scientific frontiers forward.

ARTIFICIAL INTELLIGENCE TESTBEDS AT ARGONNE, Venkat Vishwanath, Argonne National Laboratory

The ≥2 exaflop (EF) Aurora system is currently being installed at the ALCF. Aurora uses an HPE Cray-Ex platform and will contain >9K nodes, each comprising two Sapphire Rapids with HBM Intel Xeon CPUs and six Ponte Vecchio Intel GPUs with unified memory architecture (UMA) and eight fabric endpoints. The tile-based chiplet GPU architecture employs HBM and Foveros 3D Integration. Aurora will use HPE Slingshot 11 with dragonfly topology and adaptive routing. The network switches will deliver 25.6 terabytes per second (TB/s) per switch. The system will have ≥10 PB of aggregate memory and 220 PB of high-performance storage with ≥25 TB/s of decentralized autonomous organizations (DAOs). The system will support several programming models as well as ML and deep learning frameworks.

Aurora's Early Science Program supports applications spanning several disciplines and diverse simulation, learning, and data computational challenges. To support these challenges, ALCF is evaluating effective coupling of AI systems with exascale computers and experimental facilities from both hardware and software perspectives. Future architectures may couple AI accelerators with nodes or incorporate AI accelerators as disaggregated systems. AI accelerators may also be embedded in facility infrastructure or positioned at the edge. Users participate in evaluating AI accelerators supplied by vendors. Successful AI accelerators are gradually scaled in size. Current AI testbeds include Cerebras (CS-2); SambaNova (SN); Graphcore; Habana; and Groq. These AI accelerators leverage data flow architecture through diverse hardware designs and have varying software stack requirements. CS-2 and SN are available for user allocations.

Science highlights illustrated AI accelerator impact on training a conditional variational autoencoder (CVAE) model for scaling from COVID-19 cryogenic electron microscopy data to atomistic fluctuations; image segmentation for the Liquid Argon Time Projection Chamber; forecasting plasma instability in tokamak reactors; and screening candidate drug molecules for COVID-19 treatment. Other studies compared performance across AI accelerator testbeds. Due to differences in AI accelerator architecture, testbed performance varied across the measured metrics, including input/ output (I/ O) and pre-processing time, training time, and throughput and communication when scaling device number.

Ongoing efforts include upgrade plans for select testbeds; working with AI vendors to support large-language models; evaluating new AI accelerator options; integrating AI testbeds with the schedule system to improve user experience; evaluating traditional HPC on AI accelerators; and understanding how to integrate AI accelerators with ALCF's existing and upcoming supercomputers.

To engage the community, ALCF has hosted testbed training workshops with additional training sessions planned for the fall hosted by ALCF or at professional conferences. This fall's ALCF AI for Science training series for students will also include AI testbed materials.

Users may apply for allocations through the ALCF DD or the Argonne Laboratory Directed Research and Development (LDRD) programs.

DISCUSSION

Reed applauded new opportunities for exploring AI architecture and software, recalling the impact of parallel computing testbed facilities at ANL 30 years ago.

Berzins requested insights to future pricing for AI accelerators noting high costs for the first Cerebras technology generation. AI accelerators will be competing against mass-market GPUs. **Vishwanath** replied ALCF is currently evaluating AI accelerator systems for their impact on science. Cost is an important question for future consideration.

Hey asked how ALCF results related to protein folding and tokamak fusion control compared to findings from Google's DeepMind. **Vishwanath** said ongoing work is currently making these comparisons.

Lethin inquired about disaggregated architecture for AI accelerators. **Vishwanath** replied different AI accelerators have different structures, allowing a variety of architectures, ranging from on-node to rack-scale. Interconnects also differ. It is unlikely a one-size-fits-all solution will be found for all applications. ALCF is working with both users and vendors to establish benchmarks

Matsuoka speculated advantages provided by SN and Graphcore arise from their local, large memory footprint. These accelerators might compare to 30 of the A100 GPUs or 10 of the

upcoming H100 GPUs. CS-2, due to its wafer-scale configuration, is better compared to a multi-node GPU configuration. Have studies pinpointed the origins of performance advantages? Disadvantages arise from the software environment. In HPC, code performance can be distilled to the performance of kernels. Is it appropriate to characterize performance in terms of DeepBench-like kernels or microkernels, and then conjoin kernels to arrive at a performance estimate that will be consistent across different types of hardware? Of concern, increases to AI accelerator internal memory will be key to performance gains for large models. Once done, the advantages of these customized chips is negated. **Vishwanath** stated ongoing work is comparing AI accelerator systems across a variety of kernels; evaluating memory capacities is a part of this effort. Vendors are supplying tools to evaluate performance based on graph layout and effective use of units. Static random-access memory (SRAM) and staying on chip improves performance. Graph configuration is also key. Former HPC evaluations have helped ALCF understand how to make apples-to-apples comparisons in data flow and system utilization. Notionally, the suggested approach to evaluating kernels is correct, but DeepBench caters more to enterprise workloads that may not reflect scientific workloads. Identifying a set of kernels representative of scientific workloads will allow performance comparisons across hardware. Vendors may have innovative architectural solutions.

Bergman requested more information about testbed energy consumption and future steps to increase the energy efficiency of big training model workloads. This is an area where ALCF can lead. **Vishwanath** said ALCF is pushing vendors to open their applications to deliver fine-grain profiling information. Data can currently be obtained at the rack or node level. It is important chips have enough sensors to provide the needed monitoring capabilities.

Cerf remarked Google has discovered the learning phase for ML hardware requires more precision than the operational phase. Also, the hidden layers of multi-layer neural networks and their interconnections affect performance. Are there automatic ways to explore interconnection design space? Finally, related to federated learning, how can results from independent models be combined? **Vishwanath** responded most vendors support 32-bit and 16-bit systems or lower. Some vendors can support up to 90-bits precision. ALCF monitors model training accuracy. Models are first run on a CPU or GPU and accuracy metrics are compared with those from new hardware. ALCF has made progress in training and inference in a variety of scientific domains. However, one team requires 64-bit and 128-bit precision for training and would require different systems than those used today. Vendors may add position support in the future. Scaling is a challenge. The compilers supported by these architectures enable accessing multiple chips on a node and take care of data movement between the chips. Moving from chips to the network or scaling to multiple systems is an active area of research. Some vendors have solutions in this space. One approach is to have separate interconnects for AI accelerators and other parts of the computation. ALCF is working with interconnect vendors to explore the possibility of creating virtual lanes within the same interconnect.

EXASCALE COMPUTING PROJECT UPDATE, Mike Heroux, Sandia National Laboratories and Todd Munson, Argonne National Laboratory

The ECP Software Technology (ST) Team has developed a software stack that enables performance-portable application development on leadership platforms. ECP applications and others have multiple dependencies on this stack.

The ECP's KPP-3 measures ST and Co-Design (CD) project integration and creation of a productive and sustainable environment for clients. The ECP ST Advisory & Review Team

(START) is evaluating parameters related to clients, tool usage, facilities development, community ecosystem, and vendor development. ST's Dashboard, JIRA, tracks KPP-3 progress The 2021 *Application Results on Early Exascale Hardware* report highlights several ST and co-design (CD) integration success stories.

ECP has provided two new platforms to foster collaboration and cooperation: the open source Extreme-Scale Scientific Software Stack (E4S) is a comprehensive portfolio of HPC products and dependencies, and the Software Development Kits (SDKs) offer domain-specific collaborative and aggregate product suites for thematic areas, including math libraries, visualization, and programming models. SDKs are integrated into regular releases of E4S via software packaging (Spack) technologies. Spacks deploy large software collections to facilities and mediate the interoperability of container technologies for exascale computing environments. Exaworks is creating an SDK for community curated, portable, scalable, interoperable, sustainable, and trusted workflows. The most recent E4S release in May 2022 included >100 full release products. Discussion of select E4S activities addressed: the E4S build cache; testing and validation; community policies; and the E4S DocPortal for all supported software technologies.

The E4S user support model has evolved to provide a single point of contact for planning and support, as well as to deliver an integrated set of libraries and tools. The recent addition of a commercial E4S team allows facilities, industries, agencies, and other entities to acquire support with universal shared costs and benefits.

To further software sustainability activities, ST has held a series of Leadership Scientific Software (LSSw) town hall meetings on Zoom. The final meeting this year on July 28, 2022 is titled *Expanding Laboratory, University, and Industry Collaborations*. The brochure from the December 2021 BRN on *The Science of Scientific Software Development and Use* is now available. The workshop report is in progress.

ST is exploring creation of a Software Sustainability Organization (SSO) for future SDK and E4S portfolio aggregation and management via a hub and spoke model. Roles and responsibilities for SSO stakeholders are under discussion. Likely activities include regular identification of emerging needs in scientific libraries and tools; selection of new products or new functionalities within existing products; retention of products to meet ongoing needs; and the transition of products to new environments or trimming of products as appropriate to make room for new efforts. The SSO may employ a tiered approach to the continued development, delivery, deployment, and support of ECP libraries and tools across four layers pertaining to the ecosystem; hardening and delivery; porting and optimization; and development of capabilities. Price points for these layers are under discussion and will vary with the novelty of each system. Finally, ST is investigating scenarios for how the SSO will integrate within the DOE ecosystem and fill a long-term gap in TRLs.

DISCUSSION

Cerf asked about Spacks. **Munson** clarified Spack recipes describe and explain how to build software and relay all the software's dependencies.

Giles invited more information about how software created by Application Development teams that is used by a larger community but not across the DOE will fit into the proposed ecosystem. Commenting CD projects are good proxies, **Heroux** remarked one of the LSSw town halls focused on how to improve the accessibility of application-specific reusable capabilities that are not broadly used. These capabilities will benefit from hardening and documentation among other efforts to broaden their audience. This is an important topic for future conversation.

PUBLIC COMMENT

None.

Reed adjourned the meeting at 1:11 p.m.

Respectfully submitted on August 26, 2022,
Holly Holt, PhD
Science Writer, ORISE