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

## MEETING MINUTES

**June 12-13, 2023**

**HYBRID MEETING**

## ADVANCED SCIENTIFIC COMPUTING ADVISORY COMMITTEE

The U.S. Department of Energy (DOE) Advanced Scientific Computing Advisory Committee (ASCAC) convened a hybrid on Monday and Tuesday, June 12-13, 2023 at the DoubleTree Hotel, 300 Army-Navy Drive, Arlington, VA 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 in person or in virtual attendance**

Daniel Reed (Chairperson)
Richard Arthur
Keren Bergman
Martin Berzins
Tina Brower-Thomas
Jacqueline Chen
Silvia Crivelli
Jack Dongarra
Mark Dean
Jack Dongarra
Timothy Germann
Roscoe Giles
Susan Gregurick
Bruce Hendrickson
Gilbert Herrera
Anthony Hey
Alexandra Landsberg
Mary Ann Leung
Satoshi Matsuoka
John Negele
Edward Seidel
Krysta Svore
Valerie Taylor

**ASCAC members absent**

Vinton Cerf
John Dolbow
Jill Mesirov
Vivek Sarkar

**Also attending in person or virtually**

Jim Ang, Pacific Northwest National Laboratory (PNNL)
Asmeret Berhe, DOE SC
Amber Boehnlein, Thomas Jefferson National Acceleratory Facility (JLAB)
Ben Brown, Advanced Scientific Computing Research (ASCR)
Christine Chalk, ASCAC Designated Federal Officer, Oak Ridge Leadership Computing Facility (OLCF) and (ASCR)
Susan Coghland, Argonne National Laboratory (ANL)
Leland Cogliani, Lewis-Burke Associates
Lori Diachin, DOE
Gariela Negrete Garcia, Scripps Institution of Oceanography
Shuan Gleason, ORNL
Jeffrey Hittinger, LLNL
Vanessa Lopez Marrero, Brookhaven National Laboratory (BNL)
Bronson Messer, ORNL
Karl Mueller, PNNL
Robert Rallo, PNNL
Sameer Shende, University of Oregon and ParaTools
Jim Stewart, Sandia National Laboratory (SNL)
Rick Stevens, ANL
Ceren Susut, ASCR
Charlie Tahan, National Quantum Coordination Office
Cristina Thomas, 3M, retired
Juan Torres, National Renewable Energy Laboratory (NREL)

*Attendance continued from previous page*
Georgia Tourassi, ORNL
Theresa Windus, Iowa State University
Kathy Yelick, University of California, Berkeley

There were approximately 180 individuals present virtually for all or part of the meeting.

## Monday, June 12, 2023

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

Issues of U.S. global competitiveness in scientific computing cut across national security, the economy, research prowess, and Science, Technology, Engineering, and Math (STEM) workforce needs.

**VIEW FROM GERMANTOWN**, Ceren Susut, Acting Associate Director of the Office of Science for Advanced Scientific Computing Research

Susut reviewed organizational and personnel changes within DOE SC and ASCR as well as recent community awards and recognitions.

The FY24 President's Budget Request (PBR) of ~ $1.13B represents an ~5.4% increase over the FY23 Enacted Budget. A priority of the budget is to effectively transition Exascale Computing Project (ECP) resources, personnel, and capabilities. Research funds collectively increase by ~$73M, with funding changes of approximately +$15M for Applied Mathematics; +$25M for Computer Sciences; -$8M for Computational Partnerships (decreased funds reflect the move of the Quantum Computing Research Portfolio to Computer Science); and +$41M for Advanced Computing Research. The Energy Earthshot Research Centers' (EERCs') budget is flat, and the ECP budget declines (-$63M) as efforts shift to project closeout and documentation. Funding for the High Performance Production Computing (HPPC) and Network Facilities Division increases by ~$47M, with funding growth of approximately +$10M for NERSC; +$37M for the Leadership Computing Facilities; and +$0.2M for High Performance Network Facilities and Testbeds. The latter increases support planning activities for the Integrated Research Infrastructure (IRI).

The Creating Helpful Incentives to Produce Semiconductors (CHIPS) and Science Act authorized DOE to establish a crosscutting microelectronics research, development, and deployment (RD&D) program, including the establishment of up to four new Microelectronics Science Research Centers (MSRCs). Across DOE SC, the FY24 budget requests $60M for the foundation of these centers, with $25M of these funds allocated for ASCR. MSRCs will complement existing SC microelectronics awards and focus on fundamental science and early-stage research driven by the DOE mission space. Proposed DOE activities will complement those of the Department of Commerce (DOC) National Semiconductor Technology Center and the Department of Defense, respectively, which focus on later-stage prototyping and applied RD&D and DOD capabilities.

FY23 ASCR-issued solicitations include Distributed Resilient Systems; Funding for Accelerated Inclusive Research (FAIR); Reaching a New Energy Sciences Workforce (RENEW); EERCs; Scientific Machine Learning for Complex Systems; Biopreparedness Research Virtual Environment (BRaVE); Exploratory Research for Extreme-Scale Science (EXPRESS); Accelerate Innovation in Emerging Technologies (ACCELERATE); Quantum Testbed Pathfinder; Scientific Discovery Through Advanced Computing (SciDAC) – Office of Fusion Energy Science (FES) Partnerships, Science Foundations for Energy Earthshots, and Scientific Enablers of Scalable Quantum Communications. Additionally, the High Performance Data Facility (HPDF) solicitation ($300M funding cap) supports a new facility specializing in advanced infrastructure for data-intensive science, and the Advanced Scientific Computing Research for DOE Facilities ($27M funding cap) supports development of advanced algorithms

and software stacks for on-the-fly data analysis and autonomous experimentation at light and neutron facilities. All DOE SC proposals must include a Promoting Inclusive and Equitable Research (PIER) Plan for peer review.

The community's response to ASCR's October 2021 Request for Information regarding software sustainability and community vibrancy yielded feedback leading to a solicitation for seed grants. Six collaborations are now engaging with the high performance computing (HPC) and scientific software communities to gather requirements, build coalitions, and plan for the future. The ASCR community is encouraged to work with these collaborations.

DOE honored the Energy Sciences Network (ESnet) for completion under budget and two years ahead of schedule.

Early science on the NERSC-9 Perlmutter system has been productive. Draft technical specifications for NERSC-10 have been released with plans to obtain Critical Decision 1 (CD-1) in 2024. The Oak Ridge Leadership Computing Facility's (OLCF's) Frontier system remains #1 on the Top500 ranking; all allocation programs were enabled on Frontier in April 2023. The Argonne Leadership Computing Facility (ALCF) deployed Aurora's Sunspot Test and Development System (TDS), expanded the Artificial Intelligence (AI) Testbed program to include a new Graphcore system, and upgraded the Cerebras and SambaNova machines.

Reviews for the National Quantum Information Science Research Centers (NQISRCs) and the ECP Independent Project Review returned positive feedback; no major issues were identified.

ASCR celebrated World Quantum Day on April 14, 2023 (the date reflects Planck's constant), by holding a public webinar with the Office of Basic Energy Sciences (BES) to showcase the DOE's capabilities in quantum information science (QIS).

ASCR's Basic Research Needs (BRN) Workshop in Quantum Computing and Networking is set for July 11-13, 2023.

**DISCUSSION**

**Dongarra** highlighted anxiety at universities and DOE labs regarding ECP's conclusion in six months. When will affected personnel be notified to seek new jobs, and does ASCR have plans to estimate workforce loss? How will the debt ceiling impact future budgets? **Susut** relayed ASCR is working to obtain workforce numbers. Full-time employee (FTE) positions are at stake. ASCR is increasing research investments as reflected in the FY24 PBR. New FY23 solicitations that will continue in FY24 offer opportunities to fold ECP participants into core programs and new initiatives, including the Earthshots and the MSRCs. Although ECP's ecosystem is unique, personnel have developed capabilities that will enable transition to different research areas. ASCR cannot speculate on debt ceiling impacts.

**Bergman** asked about the MSRC timeline. **Susut** explained the MSRC initiative is part of the FY24 request. If funds are conferred, the MSRC program will launch in FY24.

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

An ASCAC charge to assesses how the U.S. can stay at the leading edge of scientific computing has vital implications for economic and national security. A second charge to evaluate outcomes from a partnership between DOE and the National Cancer Institute (NCI) is witnessing fruition of exascale capabilities applied to the Biden Administration Cancer Moonshot initiative. During recent House testimony, the Deputy Director for Science programs highlighted DOE SC accomplishments through interagency partnerships.

ECP collaborations are delivering solutions to previously unsolvable challenges. Arrival of the exascale era has set the stage for the next grand challenge: to develop and harness explainable and trustworthy AI for science, engineering, and national security. The innovative power of DOE's national labs and workforce are exemplified by Frontier, which is the most energy efficient, most powerful AI, and fastest hypothesis generating system in the world, making DOE SC and ASCR important players in future AI efforts. The Senate Majority leader is currently advancing a second competitiveness package for AI that highlight's DOE's potential; this package follows the CHIPS and Science Act which provided a strong DOE authorization.

Scientific data is being generated at unmatched volumes and velocities. The IRI initiative sets a vision for an open, accessible, and connected STEM ecosystem and infrastructure.

DOE SC quantum centers and research efforts are indispensable assets that deliver regular advances in this rapidly changing field. Congress is considering the reauthorization of the National Quantum Initiative (NQI) Act.

The ambitious FY24 budget request contrasts with the Fiscal Responsibility Act of 2023, which caps discretionary spending far below the request. Constraints are expected, though DOE SC is hopeful for a strong budget. If realized, the FY24 request will double investments in RENEW with inclusion of a new graduate research fellowship program; address operational, staffing, maintenance and supply chain needs for user facilities; increase Earthshot investments; and designate MSRC funds.

RENEW, FAIR and PIER are notable examples of the focus on increasing opportunities for those historically minoritized and underserved in all STEM fields. All new DOE SC proposals require PIER plans, and feedback on this program is welcomed.

**DISCUSSION**

**Taylor** sought information about DOE's engagement with the National AI Research Resource (NAIRR) Taskforce. **Berhe** explained NAIRR's goal is to democratize and expand access to advanced AI computational and data resources while federally raising the bar for ethical AI development and application. This is essential for national security and public good. Given infrastructural and talent resources, DOE participated in generating a Taskforce report and is engaging in leadership roles, conversing with the National Science Foundation (NSF) and the White House about the proposed governance structure, and seeking a win-win approach for all.

**Berzins** asked about the goals of the new graduate research fellowship program. **Berhe** explained the RENEW program has an associated fellowship similar to the Computational Science Graduate Fellowship (CSGF). To improve the diversity of applicant pools, increasing outreach and awareness are necessary along with use of inclusive selection criteria.

**Hey** observed the acronym for the Funding Accelerated Inclusive Research is already used globally to designate Findable, Accessible, Interoperable, Reusable Data (FAIR). Dual use could lead to confusion. **Berhe** agreed. This comment is fair.

**Giles** remarked the 2010 Exascale Report highlighted concerns and unique opportunities, but it took until 2017 to see budget changes. Currently, needs and opportunities have arisen for AI, QIS, and building a diverse workforce. How can concerns effectively be translated into actions? **Berhe** observed levers move slowly in Washington. DOE SC recognizes these concerns and invites feedback if any advocacy opportunities are being missed. Advocating for the CHIPS and Science Act authorization is a priority, as funds would solve many issues.

**Berzins** suggested if national labs received more base funding, the greater job security would enable competition for the very best people. **Berhe** agreed. Conversations, including with

lab leadership, to address these points are ongoing. Facilities have some stability but still struggle to support operations. An example of mitigation is a new mid-career funding opportunity.

**Matsuoka** asked about forging an international agreement for sharing AI science data. The U.S. will be left behind if every sharing instance must be signed. Japan is already working on such international agreements and hopes to include the U.S. **Berhe** answered DOE SC is very supportive of interagency partnerships and broader agreements whenever they make sense and can be accommodated. Agreements have to be on a case-by-case basis. If the open science, open data dialogue being led out of the White House Office of Science Technology and Policy (OSTP) continues, some of these concerns will be addressed. **Susut** agreed with the importance of such agreements. New initiatives like AI and quantum have given opportunities to work with other countries at the interagency level. Conversations are ongoing with the possibility of a whole-of-government approach. **Herrera** warned if data policies do not keep up, including the leveraging commercial datasets for AI, the U.S. will be left ridiculously far behind.

**Reed** reminded there are two ways to get funding: through the 302B process which is a zero-sum game or through emergency appropriation. These have different politics. Issues being discussed are whole-of-government, interagency challenges. All must work together for funding.

**NATIONAL ACADEMIES REPORT ON ADVANCED COMPUTING**, Kathy Yelick, University of California, Berkeley

Responding to the FY21 National Defense Authorization Act (NDAA), the National Academies of Science evaluated the 1) National Nuclear Security Administration's (NNSA's) computing needs over the next 20 years not supported by Exascale; 2) future of computing technologies; 3) trajectory of promising hardware and software technologies and development obstacles; and 4) ability of the U.S. industrial base to meet NNSA's needs. Input was derived from government agencies, national labs, OSTP, industry, and academia. The resulting 2023 consensus report, *Charting a Path in a Shifting Technical and Geopolitical Landscape: Post-Exascale Computing for the National Nuclear Security Administration,* recommends development of a roadmap. Business-as-usual will not be sufficient and NNSA/Advanced Simulation and Computing program's approach to algorithms, software development, system design, computing platform acquisition, and workforce development must be re-evaluated.

Findings addressed 1) growing computing demands surpassing planned NNSA upgrades; 2) rapid shifts in technology and commercial landscapes, especially as cloud hyperscalers dominate computing and largely off-shore manufacture of semiconductors and other hardware may not be tailored to NNSA needs; 3) a need for bold and sustained research and development (R&D) investments, including support for higher-risk activities; and 4) significant challenges in workforce recruitment and retention.

Recommendations called for 1) new, aggressive, and comprehensive design, acquisition, and deployment strategies to meet future systems needs; 2) high-risk, high-reward research in applied mathematics, computer science, and computational science; and 3) development of a national strategy, partnering with agencies and academia, to address workforce challenges.

**DISCUSSION**

**Reed** summarized major changes in the global computational landscape from 30 years ago. The U.S. no longer globally controls the semiconductor ecosystems, there are workforce issues, and solutions will require investment of billions. Issues are complex and traverse all of

government with implications for national security and economic preparedness. System approaches will require rethinking prototyping and evaluation and the testing of new ideas.

**Landsberg** questioned whether AI workforce growth will require retraining Ph.D.s or new expertise. **Yelick** indicated both retraining and new expertise will be needed. Researchers across domains are already incorporating AI methods. Utilizing novel hardware will require new approaches. Traditional computer scientists will be needed to evaluate AI methods.

**Gregurick** inquired about quantum-HPC hybridization for accelerated applications. **Yelick** explained the investigation of integration was limited. Hybrid approaches are relevant to the materials multi-physics workload, but quantum technology is unlikely to replace HPC systems in 10-20 years. Exciting work is happening at the HPC-quantum boundary, but quantum accelerators, for example, may not necessarily be used for accelerated weapons design.

**Svore** (via chat) clarified quantum computers should be viewed as accelerators; they will provide more accurate chemistry and materials models than can be efficiently achieved classically. How are software being prepared for rapid adoption of quantum, AI, and HPC hybrid capabilities? **Yelick** agreed quantum capabilities may be revolutionary for particular applications; the report considered a much broader application space.

**Arthur** asked about common infrastructure and standards for NNSA and industry data and model interoperability. **Yelick** conceded the report did not go into the specifics of this topic; there is a general statement about data movement between applications.

**Chen** wondered about leveraging overlap among ECP, SC, and NNSA data and multi-physics applications. **Yelick** explained the report identified knowledge transfer from those working on multi-physics simulations in the labs to those working on classified NNSA simulations. However, a combined strategy looking beyond ECP is not evident. An approach is emerging with the AI4SES efforts. The report comments on the need for agencies such as the Department of Defense (DOD), National Aeronautics and Space Administration (NASA), and National Oceanic and Atmospheric Administration (NOAA), to work together through a whole-of-government strategy going forward for HPC modeling and simulation systems. AI will require the labs to work with hyperscalers.

**Matsuoka** questioned what evidence indicates hyperscaler products will diverge from NNSA needs. Applications have performed well when utilizing Amazon's Graviton processor on Virtual Fugaku. Processors are unlikely to be designed with low memory bandwidth. **Yelick** reiterated hyperscalars are optimizing for a workload that is not the traditional NNSA workload. What hyperscalers design for AI will likely benefit the AI part of NNSA's workload. Hopefully, NNSA and DOE SC can influence products, as was done for GPUS. Memory bandwidth is important, but more so for certain sparse types of methods than for dense linear algebra. Some processes are designed with limited, indirect addressing, or irregular memory access patterns, which leads to concerns about future support. The AI space may be over optimizing for a small set of deep learning (DL) methods and not thinking broadly about future methods. **Reed** pointed out hyperscalers sell services, not hardware. This differs from the past when chips could be purchased from a silicon vendor and integrated. Relationships must shift because power players and the money have shifted; this is a socio-economic issue separate from the technology. **Gil** suggested there are lessons to be learned from the intelligence community, which has shifted to a model of buying from cloud providers and has utilized some of their unique chips. Co-design is a very different practice when one does not help with or specify the design.

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

**INDUSTRY PERSPECTIVE ON COLLABORATING WITH SC USER FACILITIES,** Cristina Thomas, 3M (retired)

3M's technology platforms, products, and innovation capabilities are company pillars that enable customer service. Establishing Cooperative Research and Development Agreements (CRADAs) with national labs has been pivotal in the development of new products. 3M has benefited from DOE expertise and supercomputers to 1) simulate and reduce energy required for the melt-blown fiber manufacturing process to generate filters in N95 masks (with use of the ALCF and Laboratory Computing Resource Center, LCRC); and 2) create a machine learning (ML) reverse-image search process for 3M's atomic force microscopy library for materials design (in partnership with ANL and the University of Chicago); and 3) develop multi-physics approaches for the design and manufacture of materials, such as metamaterial films for passive solar cooling (with SNL and Advanced Research Projects Agency-Energy, ARPA-E).

DOE-industry collaborations require significant time and investment from both parties. Mechanisms enabling partnerships are vital, but it may take years to secure agreements. The expertise and resources provided by DOE are unique.

**DISCUSSION**

**Reed** wondered how collaborations begin. **Thomas** shared with ALCF, for example, work began with mutual visits. ALCF was eager to learn from and adapt codes for commercial data while 3M benefited from materials development that improved competitiveness and economic outlook. Willingness to reach out and respond to requests is important.

**ECP UPDATE,** Lori Diachin, Lawrence Livermore National Laboratory

ECP is focused on meeting Key Performance Parameters (KPPs) for application and software technology. Since Frontier access was granted in April 2023, seven of 11 applications and five out of ten applications have tentatively met KPP-1 (achieving a performance figure of merit) and KPP-2 (addressing a base challenge problem) criteria, respectively. Tracking KPP-3 (client use of a significant capability) is complex, with weighted points awarded when a team demonstrates integration. At present, 27.5 integration points have been awarded. KPP-4 is complete, with all 267 vendor PathForward milestones delivered.

After achieving KPP-1 and KPP-2, ECP teams transition to early science. Six teams were awarded Innovative and Novel Computational Impact on Theory and Experiment (INCITE) allocations. Overall, ECP team node usage is >2.1M node-hours to date. Some Frontier challenges remain, including challenges with the software stack (e.g., immaturity of OpenMP with target offload); limitations in user knowledge; node hardware failures; and performance variability at scale. Moving forward Aurora will provide an important portability test.

Science highlights showcased progress from ExaSky and MFIX-Exa.

The May 2023 Independent Project Review (IPR) found ECP 1) has made progress in addressing 2022 IPR points; 2) is on track to meet KPP thresholds; 3) is prepared for project closeout; 4) is managing risk and contingency; and 5) is being properly managed overall. Comments address communication; staffing and succession planning; the KPP verification process; code reliability and robustness; and stretch science problems.

ECP's communication strategy engages several platforms (e.g., podcasts, Twitter, Youtube). ECP has increased staff resources for communication, and additional efforts are planned for 2023. ECP outreach to the Industry and Agency Council (IAC) has been active, and the ECP Broadening Participation Initiative has continued to expand the workforce pipeline

through internships, career awards, HPC educational materials, and the HPC Workforce Development and Retention Action Group.

Towards closeout, ECP is extending one quarter though December 31, 2023. Activities are focused on completing technical work, satisfying 413.3b documentation and formal review processes, return of uncommitted/unspent funds, transition of project tools, and continued outreach and stakeholder engagement. Leadership will remain engaged through FY24. Diachin reviewed select ECP leadership transitions; ECP has enjoyed stability at the L2 and L3 levels. ASCR is targeting possibilities for follow-on funding from DOE and other agencies.

The Post-ECP Software Sustainability Organization (PESO) was awarded ASCR seedling funds to deploy a hub-and-spoke model to support a wide array of software product communities and enable crosscutting communities of practice. Past and upcoming community engagement opportunities include workshops and meetings.

**DISCUSSION**

**Seidel** encouraged future presentation of how ECP has enabled qualitatively new and different research and also reiterated prior concerns about ECP personnel as the project closes out. **Diachin** offered examples of novel earthquake and turbine simulation capabilities and acknowledged ECP closeout concerns.

**Herrera** posed a question about optimizing GPUs for AI/ ML applications. **Diachin** called attention to ExaLearn, a co-design project center focused on AI/ ML and the CANcer Distributed Learning Environment (CANDLE) application which supports ML-based techniques. AI/ ML techniques have gained attention and are incorporated in many projects' stretch goals. **Siegel** (chat) added ML has organically entered many projects as a key component. For example, ML potentials for molecular dynamics, subgrid modeling, etc. **Heroux** (chat) noted many of the libraries and tools in the ECP Software Technology portfolio are integrating ML-related capabilities to provide support for low-precision data types and algorithms using low-precision features. E4S contains all of these libraries and supports the most important AI/ ML libraries like TensorFlow and PyTorch; there is portably across all three GPU architectures and in cloud environments. Though there is a lot of AI/ ML work ahead, DOE is not starting from scratch.

**Arthur** asked if KPP lags on Frontier are also seen on Aurora. **Diachin** said a few applications will likely perform better on Aurora and are prioritized for early access. **Siegel** recalled issues with OpenMP on Frontier; early indications suggest better application performance with respect to OpenMP on Aurora.

**Diachin** encouraged engagement with PESO and other seedling projects through Leadership Scientific Software (LSSW) townhalls and other platforms.

**EXTENDING THE IMPACT OF THE ECP SOFTWARE ECOSYSTEM,** Sameer Shende, University of Oregon and ParaTools, Inc.

With increasing software complexity, it is becoming increasingly difficult to measure the performance of and correctly install tools and libraries in an integrated and interoperable software stack for HPC application deployment to the cloud. ParaTools, founded in 2004, supports technology transition by offering custom HPC performance tools and engineering; HPC training; and support for parallel runtime systems and cloud platforms.

ParaTool's Tau Performance System® comprises a portable profiling and tracing toolkit for performance analysis of all HPC parallel programs. The Tau Commander improves usability of the Tau Performance System®.

Interfacing with the ECP, ParaTools supports the Extreme-Scale Scientific Software Stack (E4S) which contains ~71 unique software products that enable application developers to create highly parallel applications for diverse exascale architectures from different vendors. E4S products, with a community generated collection of software tools and libraries external to the ECP, are released quarterly via the flexible package manager, Spack. Notably, the E4S-Intel agreement, made possible through the ECP-ParaTools partnership, makes Intel compilers and MPI libraries available through E4S containers. e4S-alc is a new tool that enables customization of container images.

Deploying HPC/ AI workloads to the cloud includes consideration of MPI communication, inter-node network adapters, and intra-node communication. ParaTools is building E4S for optimization with MVAPICH-Plus and using Adaptive Computing's On-Demand Data Center (ODDC) interface for launch of E4S jobs on multiple cloud providers. Additional commercial support services for E4S offered by ParaTools include issue tracking and resolution, E4S installation and maintenance, and ECP Application Development (AD) engagement.

The 50 E4S Electronic Design Automation (EDA) products, accessible on commercial cloud platforms, may support proposed CHIPS and Science Act efforts.

**DISCUSSION**

**Berzins** asked about long-term growth for DOE software efforts and commercial sustainability. **Shende** stated ParaTools intends to support E4S into the future, including the quarterly release, installation, and proper running of >100 products at user facilities. ParaTools will continue work with AD teams, has connected with PESO, and is seeking relationships with international organizations, industry, other labs, and government agencies.

**Seidel** inquired about international cooperation and joint funding efforts. Exascale systems are rare, and more are needed for robust software development. **Shende** referred to collaborations with supercomputing centers in Europe and Australia, and NSF partnerships. There is growing international interest in E4S. Regarding funding, some supercomputing centers have contracts with ParaTools for training and outreach activities.

Noting potential for architectural diversity to explode, **Reed** drew attention to future software sustainability. **Shende** indicated ParaTools is interested in advanced hardware, such as accelerated processing units (APUs), MI300 GPUs, or Grace Hopper GPUs. E4S will need to be optimized for such architecture. **Heroux** (chat) added a software ecosystem with performance portability layers is one of the most effective and efficient ways to respond to increasing architectural diversity. **Windus** (chat) emphasized the importance of software sustainability projects to engage and include the applications in their efforts since applications are now more dependent on other software.

**Matsuoka** observed leveraging E4S on Fugaku and sister systems and sharing tools will be easy as software is already delivered via Spack. Are all ECP products required to be ECP compatible and Spack compliant? What is the ballpark figure for maintaining E4S delivery features but not the software contained therein? Figures should be publicly available for collaborators. **Shende** looks forward to collaborating with Japan. There is an ECP mandate for E4S and Spack compliance. **Heroux** advised sustaining the E4S ecosystem is highly variable in scope, with estimated costs ranging from >$5M to <$20M per year. An analysis for a variety of scenarios is available and has been provided to sponsors.

**SOFTWARE DEPLOYMENT AT FACILITIES**, Ryan Prout, Oak Ridge National Laboratory

Cohesive ECP software deployment hinges on relationships among responsible parties and common standards and infrastructure for packaging, testing, delivering, and integrating software across facilities. Together, the Software Integration (SI) and Continuous Integration (CI) teams provide the infrastructure and support for ECP software deployment while working with ParaTools and facility staff.

E4S uses Spack to distribute HPC software packages. E4S community policies govern the build, validation, documentation, and public accessibility of Spack packages. Subsequently, high-quality Spack recipes are regularly tested at the OLCF, ALCF and NERSC with integration support from ParaTools. Facility staff at each site provide E4S team contacts and a CI-based infrastructure for automated testing to ultimately deliver an operational software stack, driven by unique user requests and facility requirements. Of note, security hurdles may prevent "true" CI directly on production, multitenant, and facility HPC systems. Following each Spack release and integration into facility systems, support requests are relayed to developers as needed. Collectively, these efforts and relationships enable a sustainable ecosystem for teams to package, distribute, and manage software at facilities through the E4S vehicle.

**DISCUSSION**

**Giles** inquired about staff turnover and ease of training new employees in SI and CI activities for E4S. **Prout** explained many working in deployment have been there a long time. Spack is becoming widely used, so many who join are already familiar with the tools. Each site has documentation which enables training new employees. Standardizing CI infrastructure security requirements has presented the biggest hurdle. **Diachin** added an IAC member described a three- to four-month learning process to implement E4S after making the deliberate decision to forego expert help.

**UPDATE ON AURORA,** Susan Coghlan, Argonne National Laboratory

The ALCF-3 Aurora system boasts a theoretical peak performance of ≥2 exaflops with double precision (DP). Aurora uses an HPE Cray-Ex platform and contains 10,624 nodes housed in 166 compute racks. Each node comprises two Intel Xeon Max Series CPUs with high-bandwidth memory (HBM), six Intel Data Center Max Series 1550 GPUs, 1 terabyte (TB) of DDR5 ram, and 1 TB of HBM. Nodes have unified memory architecture (UMA) and eight fabric endpoints. Aggregate system memory comprises 10.9 petabytes (PBs) of DDR5, 1.36 PBs of HBM CPUs, and 8.16 PB of HBM GPUs. Aurora uses the HPE Slingshot 11 interconnect with dragonfly topology and adaptive routing to deliver a peak injection bandwidth of 2.12 PBs/ second (s) and a peak bisection bandwidth of 0.69 PB/ s. The 220-PB high-performance storage houses 1,024 distributed asynchronous object storage (DAOS) nodes with a 31 TB/ s DAOS bandwidth. Intel has released information about hardware to the public, though some nuances are reserved. Aurora supports several programming models as well as ML and DL capabilities.

Sunspot, Aurora's TDS, contains the same hardware but in smaller quantities (2 racks with 128 compute nodes and a small DAOS). Aurora's Software Development Kit (SDK) and programming environment are installed, and Sunspot opened to all ECP and Aurora Early Science Project (ESP) teams in December 2022.

Following resolution of supply chain issues, 99% of Aurora's compute blades have been installed and powered on. Hardware upgrades were completed in June 2022. Current Aurora

efforts are focused on hardware and fabric testing and system stabilization. Scaling and optimization work are beginning. Though not open to external users, internal team testing is underway, and early user access is expected in the next few months. Across Aurora ESP and ECP teams, 44 codes will be tested, with 3 codes intended for CPU use only. Early application results from Aurora or Sunspot have been presented at various professional gatherings.

**DISCUSSION**

**Berzins** asked if Aurora's new architecture and software stack have affected porting of codes. Will other systems with similar architectures be made available? **Coghlan** shared the Aurora system is not completely unique but is currently acquiring all available Intel Data Center Max Series 1550 GPUs. If and when smaller subsystems will become available are unknowns. The porting of codes is going well; SYCL and HIP are both available on Aurora.

**Giles** pivoted to the timeline for Intel to disclose additional system information. **Coghlan** commented much information has already been disclosed, but Intel may never release some information. The process for AD teams has been lengthy, and many have signed Restricted Secret Non-Disclosure Agreements (RSNDAs).

**UPDATE FROM WORKING GROUP ON DOE-NCI COLLABORATION,** Tony Hey, ASCAC

ASCAC's DOE-NCI subcommittee met at the Frederick National Laboratory in June 2023 to review the progress of joint projects.

Findings for Modeling Outcomes using Surveillance data and Scalable Artificial Intelligence for Cancer (MOSSAIC) addressed clinical information extraction and abstraction with high positive predictive values and uncertainty quantification (UQ); a recurrence and metastasis predictive model; registry deployment; validation; rapid case ascertainment (RCA); the open-source Framework for Exploring Scalable Computational Oncology (FrESCO) pipeline; and federated learning. Comments touched on the scalable Natural Language Processing (NLP) system; development of a transformer-based foundation model; application (API) deployment; leveraging of RCA; a privacy-preserving API; Health Insurance Portability and Accountability Act- (HIPAA-) compliant HPC tools; and a successful run on Frontier. Recommendations centered on leveraging foundation models; data quality improvement; deployment in translational settings; continued adaptation for RCA; engagement of new partners; and comparative assessment of pipelines.

Findings for AI-Driven Multi-scale Investigation of RAS/RAF Activation Lifecycle (ADMIRRAL) highlighted improvements to the domain decomposition molecular dynamics (ddcMD) multi-physics particle dynamics code and incorporation of Martini-3 lipid potentials; Multiscale Machine-learned Modeling Infrastructure (MuMMI) upgrades; and experimental validation. Comments addressed performance portability; experimental validation; and model translation. Outstanding recommendations from 2022 include assessment of opportunities and priorities due to AI advances and the effort required to harmonize tools with E4S. Recommendations from 2023 advised on experiments, including those for validation; community engagement; and general applicability of the methodology.

Findings for Innovative Methodologies and New Data for Predictive Oncology Model Evolution (IMPROVE) addressed collaborations; the Scientific Advisory Committee; curation, hyperparameter optimization, and benchmarking of framework models; understanding predictive accuracy and ramifications for drug design; curated and generated benchmark data; and an alpha

release of the codebase and documentation. Comments centered on hackathons; progress determination; potential use cases for patient digital twins and precision medicine; possible data markets; and incentives for community data contribution before publication. Outstanding 2022 recommendations addressed publication of methodology; partnership expansion; and the productivity/ usability of models. New 2023 recommendations seek compelling results stories; clarification of data features; formal characterization of model competence regions; consideration of industry formalisms; consideration of phantom results; and results publications.

Overall, CANDLE is a significantly successful ECP research and software project. All milestones were delivered on time. CANDLE has run successfully on Summit and Frontier and exceeded its KPP-1 metric by a factor of 5x. The framework is used for 30+ cancer DL models. CANDLE additionally delivered COVID-19 research, demonstrating potential for technology transfer. Additional findings and comments address code availability, usage, and software sustainability; hands-on workshops; hyperparameter optimization and ensemble learning; and tool adoption by IMPROVE. An additional, ongoing stretch goal is development of a DOE transformer-based modeling framework that can run at scale. Cross-ECP efforts and dialogue with AI4SES is ongoing. An international consortium is being formed to develop a 1-trillion parameter model using NVIDIA Megatron and Microsoft DeepSpeed codebases for training.

**DISCUSSION**

**Siedel** wondered about long-term DOE-NCI outcomes. **Hey** remarked some in the NCI community were initially was skeptical of the utility of HPC for cancer. Funded by the Cancer Moonshot, projects like MOSSAIC have since demonstrated significant impact. Foundation models will further increase outcomes. **Tourrasi** elaborated on the value proposition of the DOE-NCI partnership. For the NCI, MOSSAIC is already accelerating reportability of cancer data in the most cost-effective way. Regarding the DOE, no one thought NLP was a supercomputing problem, but the model has stretched HPC. Further, OLCF is now enabling scientific collaboration through creation of Citadel to protect proprietary data. There is ongoing work on privacy preserving computing with federated learning. In the future, such approaches could enable international collaboration for training lab-scale transformer models. This is the true value of open AI — where issues of bias, transparency and reproducibility can be addressed.

**Herrera** asked about lessons learned from the failure of the MD Health Anderson partnership. **Hey** stated NCI-DOE projects have made excellent progress and are demonstrating the role of HPC in cancer.

**Crivelli** posed a question about explainability and bias. **Tourassi** said MOSSAIC has developed explainability models and is exploring bias in partnership with clinicians and registrars to build confidence. The fact that models are in production across 16 registries illustrates user community trust. **Berzins** noted CANDLE has UQ built in; though it is early days, it is important to address issues. **Hey** agreed verification, validation, and UQ are important.

**Seidel** commented that a senator expressed concerns about DOE using Chat GPT. **Hey** clarified DOE is not using Chat GPT. AI4SES efforts are proposing to train foundation models with scientific data to solve scientific problems.

**TOWARDS AND INTEGRATED RESEARCH INFRASTRUCTURE**, Ben Brown, ASCR Facilities Division Director

The IRI's vision, "to empower researchers to meld DOE's world-class research tools, infrastructure, and user facilities seamlessly and securely in novel ways to radically accelerate

discovery and innovation," is driven by exascale science, a deluge of experimental and observational data, and AI4SES. Meanwhile, enabling technologies like Esnet6 have been completed. The FY21 PBR called for an Integrated Computation and Data Infrastructure Initiative, leading to creation of the ASCR IRI taskforce and IRI Blueprint Activity in 2021 and 2022, respectively. IRI Program Development commenced at the end of 2022, and the FY24 PBR advances the IRI and HPDF.

The IRI Blueprint Activity created a framework for IRI implementation around three Science Patterns (Time-sensitive; Data integration-intensive; and Long-term campaign) and six Practice Areas (User experience; Resource co-operations; Cybersecurity and federated access; Workflows, interfaces, and automation; Scientific data life; and Portable/ scalable solutions).

IRI implementation is ongoing through four steps: 1) Invest in IRI foundational infrastructure; 2) Bring existing IRI projects into formal coordination; 3) Deploy an IRI pathfinding testbed across the four ASCR facilities; and 4) Stand up an IRI program structure at both headquarters and in the field. Corresponding early focus projects for each step include: 1) reporting IRI science pattern requirements across all SC programs and assessing IRI readiness; and advancing the National Energy Research Scientific Computing-10 (NERSC-10), HPDF, and OLCF-6 projects to CD-1; 2) moving towards a common user experience across ASCR HPC facilities; and creating an IRI HPC allocation; 3) initiating projects for the Time-sensitive and Data-integration intensive patterns; and engaging ASCR facilities with new IRI oriented research projects; and 4) developing a DOE Authentication/ authorization standard; and devising and revising a governance structure with the community.

Release of the final IRI Blueprint Activity report is anticipated in 2023 as well as the ESNet Requirements Reviews IRI Meta-analysis and the IRI Pathfinding Testbed white paper. An IRI convening event is also anticipated.

**DISCUSSION**

**Arthur** articulated the need to manage data provenance and pedigree and UQ. For example, removing affected data from models will be important in case of a bad sensor. From an engineering perspective, there is an idealized pace for science. Balancing system demand, costs, and confidence thresholds are important. There are examples of these dynamics from the Arnold Air Force Base, the James Webb Telescope, and the security community. **Brown** recognized the importance of both data provenance and a living body politic. The IRI Blueprint Activity records the importance of user experience and how relevance may be impacted by technological advances during product development. IRI service should probably have a three-nines uptime, but as a first articles endeavor, services may not always be up.

**Hey** asked whether the HPDF will include all Office of Basic Energy Sciences (BES) data. Astronomy data from select United Kingdom projects is managed by the community and data capture methods offer interesting examples for IRI consideration. How can we ensure data continues to be available, especially given storage costs? The Office of Science and Technical Information (OSTI) was never mandated to make data available, but there are exciting opportunities going forward for open science. **Brown** cannot comment on the HPDF because of an ongoing competition. More information is available in the lab call. Community ownership of data is sacrosanct and is governed by social, technical, and political factors. But, when asked for help, ASCR thinks about building a flexible infrastructure to be a rising tide for all ships. Data storage is costly and will need to be allocated in the future. DOE SC leadership and staff are

working on an open science effort. After the OSTP memo to make federally funded research freely and publicly accessible, agencies are producing open access plans.

**Dongarra** called for greater clarity in the competition of facility research projects to allow wider engagement. Research funding is a zero-sum game. **Brown** noted ongoing conversations to increase visibility of research efforts in the community; this should be a primary IRI goal. Laboratory Directed Research and Development (LDRD) and facilities research to meet user needs have been the primary sources used to support activities. **Susut** commented the HPDF is currently the only open competition. The IRI is a new initiative.

Going forward, **Taylor** raised connections with the research community in the formation of the governance board and program structure. **Brown** invited community feedback throughout the IRI process. Having a variety of venues through which all in the community can engage is paramount and part of ongoing discussions.

**Gregurick** asked about extending DOE resources to non-DOE facilities as a longer-term goal. **Brown** recognized potential for DOE to act as leading voice in the research community. Partnerships, such as the DOE-NCI effort, offer an excellent opportunity for transformational learning on both sides. Projecting expertise into other fields by leveraging sociological lessons learned, such as through E4S, can extend impact along with supporting communication.

**Matsuoka** asked about division of labor and management costs between DOE and communities for individual computing centers and storage facilities, instruments with their own data processing capabilities, and the cloud. **Brown** explained ASCR has clearly demarcated mission boundaries surrounding foundational infrastructure and services. However, workflows are end-to-end problems and partnerships create social, technical, and policy interfaces. Communication is key. ASCR may present certain design patterns where risk has been bought down. Those that independently adopt these software and hardware designs can then tap into national computing resources. The community will need to reconcile with such choices. For example, it took two years to frame and stand up ESnet.

**Public Comment**

**Jim Ang** (PNNL) serves on the DOC Industrial Advisory Committee for the CHIPS and Science Act. A subcommittee is identifying R&D gaps for DOC and the National Institute of Standards and Technology (NIST) investments related to the formation of the National Semiconductor Technology Center. Investments will establish hardware and integration capabilities for advanced packaging for the U.S. and allies. For example, investments will lead to creation of heterogeneous processors. Though DOC will build testing and prototyping fabrication capabilities, DOC must rely on mission partners — such as the DOD, DOE, NASA, and other agencies — to test system software and evaluate performance of prototypes to drive co-design for system and application challenges. DOE should have a seat at the table through appropriations to participate in these CHIPS and Science Act efforts.

*Reed dismissed the meeting for the day at 5:18 pm.*

**Tuesday, June 13, 2023**

*Reed convened the meeting at 10:00 a.m. Eastern Time.*

**AI FOR SCIENCE AND SECURITY WORKSHOPS AND REPORT,** Rick Stevens, Argonne National Laboratory

In 2020, the ASCAC *AI for Science* report recommended a major DOE AI for Science (AI4Science) program. Subsequent major advances in AI led to the need for another report. Following workshops and solicitation of community-wide input in 2022, the SC and NNSA released a report in 2023 titled *AI for Science, Energy, and Security*. Workshops addressed six crosscutting themes: 1) AI for advanced properties, inference, and inverse design; 2) AI and robotics for autonomous discovery; 3) AI-based surrogates for HPC; 4) AI for software engineering and programming; 5) AI for prediction and control of complex engineered systems; and 6) Foundation, Assured AI for scientific knowledge. Importantly, advances in foundation model capabilities have progressed rapidly since 2019, and DOE has the opportunity to develop foundation models to assist in research. Future foundation model tasks may include summarization and synthesis of knowledge and development of research plans, including hypothesis generation.

The report discusses SES application spaces where AI could play a transformative role. As AI poses risks to society and to global security, the development of responsible and trustworthy AI capabilities that are aligned with human values is imperative. DOE/ NNSA are well positioned be a U.S. leader in the development of such AI capabilities. Driven by their mission space, DOE/ NNSA have a long history of relevant world-leading R&D expertise (e.g., ECP), the world's most capable user facilities, and strong ties with private sector technology and energy organizations. Thus, the report posits "only DOE/ NNSA can advance responsible co-design of AI R&D with a strong focus on science, energy, and national security via simultaneously tying R&D to their mission, thereby creating and implementing solutions." Though not an implementation plan, the report envisions a new initiative — Secure, Trustworthy, Reliable artificial Intelligence for Discovery, Energy, and Security (STRIDES) — with efforts addressing integrated science R&D on alignment, ethics, and responsibility; transformational hub-scale centers for key AI4SES themes; crosscutting AI technologies; and dedicated access to computing and experimental facilities. The IRI is essential to the latter effort.

**DISCUSSION**

**Giles** raised the potential for science-based foundation models to hallucinate. **Stevens** acknowledged R&D is needed in this area, but there is no reason to believe this problem cannot be solved. Building on protein modeling work that previously won a Gordon Bell award, ongoing research indicates models hallucinate because of training, especially due to oddities in training data. Reinforcement and human feedback combined with highly curated data may reduce hallucinations. However, it is unlikely models will ever be 100% accurate or truthful. A mechanism to recognize when models make things up is needed. Related, Perplexity is a system built on top of Chat GPT-4 that utilizes augmented retrieval to pull citations for materials delivered in response to a prompt. The rate of false declarations is low. Models like Claude have been trained using constitutional AI techniques to be better aligned with human requests.

**Landsberg** applauded the strength of the presentation's case for DOE to lead national efforts in AI4SES but requested additional information about partnerships with other countries and government agencies, like NSF. **Stevens** sees significant potential for agency and international partnerships. The National AI Strategy encourages partnerships with like-minded allies, and discussions are already ongoing with Japan about collaborations for large language models and other infrastructure. Related efforts are currently identifying additional partners.

DOE's work in this space will complement NSF's broad mission, and conversations are considering how best to engage with NAIRR, NIST, the Food and Drug Administration (FDA), and the National Institutes of Health (NIH).

**Crivelli** underscored the importance of being able to access data from partners. Efforts must move rapidly because of rapid AI developments. **Stevens** agreed. DOE and broader organizations understand the urgency of this topic. Work is ongoing to get this initiative off the ground. Staffers on both sides of the Hill have been briefed.

**Herrera** asked if, in the context of scientific research, the report explores the concepts of reliability, quantification of margins and uncertainty, and explainability. **Stevens** confirmed these topics are discussed in the 2019 and 2023 reports. Today's slide deck was used in presentations on the Hill and was shared with ASCAC to show how messaging is being framed. The ability to obtain predictions and the error around predictions is vital. ASCR and NNSA are currently investing in this research.

**Thomas** commented on future opportunities to use AI in combination with domain knowledge; such tools may assist but not necessarily drive research. How can ASCAC help? **Stevens** requested ASCAC be supportive of the initiative that emerges from this work. AI4SES has momentum and national imperative. DOE needs to be part of this process for its own mission and act as a neutral broker of this technology. This initiative could be several times the size of ECP and has great potential to lift computing at the labs and for domain scientists.

**Taylor** wondered how the ASCR community can help. **Stevens** called for uniting the community behind this effort as it advances through internal agency and budget processes.

**Seidel** asked what is AI's current capability to yield discoveries by synthesizing information across scientific literature. **Stevens** remarked an AI tool, such as a foundation model, that could read all the scientific literature and make suggestions or generate hypotheses would be a game changer. Models are getting there, but will require capabilities that are much more grounded and powerful than Chat GPT-4.

**Dean** wondered about preventing AI from causing harm. Will the government set guidelines? **Stevens** stated the science and math behind model alignment with human interests will be a vital component of this initiative. There is strong interest within DOE. Current strategies to align models include placing guardrails or using constitutional AI-based training. However, there is no single magic bullet, and many techniques will need to be developed. Regulation will likely occur in the future; major AI companies are already asking for guidance. The challenge will be determining whether a system is performing as it should and monitoring activity to catch bad actors. DOE can be a major player in the broader dialogue and is likely to have a role in monitoring for future AI security threats, just as DOE currently plays a role in national and international cybersecurity networks. This initiative's scale is likely to reflect the magnitude of the impact AI will have on the nation's future.

**Svore** sought additional input on using AI as an accelerator for applications and how to leverage domain expertise at the national labs to augment human feedback models. **Stevens** explained AI surrogates enable creation of accelerated application versions via kernel approximation that operates within an acceptable level of error. A dozen published examples show speedups on the order 10K to 100K. This approach fundamentally changes what can be accomplished and will likely be hardwired into simulation over the next five to ten years. AI surrogates may also drive evolution of computer architecture. Scientists have a handle on such techniques but do not yet have a well-defined program. With regards to domain experts, bigger teams and partnerships are needed across science domains, not just within computer science and

math. A scale grander than the Beyond the Imitation Game Benchmark (BIG-bench), which has 400 people building test cases to evaluate AI, is needed. DOE may involve 1K people in a multi-year process that will be a once in a lifetime transition in modeling.

**Matsuoka** asked if DOE can be an innovation leader in fundamentally changing how fields collect data. Many fields do not collect data in a continuous, high-throughput, systematic, multimodal, and high-resolution fashion. New equipment is needed. For example, the replication scale for biological experiments frequently falls far below the level of data needed for foundation models. **Stevens** agreed innovations in data collection will be important going forward. Changes will technical, social, and educational components; people must be willing to explore new approaches to obtaining and managing data. DOE does not yet have a coherent way of handling data across program offices. Thus, the IRI is essential. Interagency partnerships will also be important. For example, DOE-NCI is jointly addressing data-related processes.

**Hey** questioned the initiative's timeline; this is an urgent need that should have been addressed earlier. How will the budget ceiling impact funding? **Stevens** said predicting how rapidly Congress will provide funding is difficult. Optimistically, follow-on funding to the CHIPS and Science Act could be delivered this year or next. DOE will need to dramatically increase the workforce focused on AI. There is urgency behind this initiative and with momentum, there is the opportunity to remap individuals from ECP and to hire new talent.

**Gregurick** (chat) commented obtaining large-scale data collection in biology that is AI-ready is a challenge, though data size may not always be as important as having representative data that is AI capable. There is potential for synergistic activities between the IRI and the DOE-NCI partnership or with the NIH Bridge2AI initiative.

**Reed** reiterated sentiments that a tsunami is approaching the future of computing. DOE will need to build new approaches through interagency collaborations to weather changes.

**REPORT FROM THE SUBCOMMITTEE ON INTERNATIONAL COMPETITIVENESS**, Jack Dongarra, ASCAC

In response to a DOC SC charge received in March 2022, ASCAC formed a subcommittee and generated a report titled *Can the U.S. Maintain its Leadership in High Performance Computing?* Charge elements addressed 1) critical areas for ASCR leadership; 2) advanced research tools; 3) building and maintaining strategic industry and international partnerships; and 4) strategies for workforce success, recruitment, and retention.

Key findings of U.S. strengths touched on 1) science and engineering applications of national importance and future needs for increasingly capable advanced computing systems; 2) U.S. leadership in applied mathematics and computational science; 3) synergy among HPC, big data, simulation, and AI/ ML; 4) ECP as an exemplar of U.S. leadership; and 5) a history of close partnerships between DOE and industry. Notably, the end of the 6) ECP heralds a success and risk as DOE is highly vulnerable to losing knowledgeable staff. Key findings addressing U.S. challenges noted 7) critical areas under threat due to global competitiveness, funding and budgetary constraints, and brain drain; 8) fundamental changes in the technology landscape; 9) a horizontal and international HPC supply chain; 10) level or declining funds for ASCR research in real terms; and 11) declining prestige and attractiveness of national lab careers. Key findings relating to outlook foresee 12) increasing need for international collaborations; and 13) an interdisciplinary approach requiring co-design.

The report's four recommendations called for 1) building on existing strengths in high-end modeling and simulation, AI, leading edge computing architectures and systems, and

advanced networks and future internet architectures; 2) a decadal-plus post-exascale vision and strategy with a focus on providing sustained investments; 3) a vision, associated goals, and milestones for international collaboration focused on post-exascale computing; and 4) a strategy in long-term, forward-looking co-design research in architecture, hardware and system concepts.

The report concluded the U.S. is losing its historical leadership position in the field. Further, it is no longer the case that the U.S. is a highly desirable destination for the career development of scientists or that the national laboratories host the most talented researchers. ASCR should revive stable funding, maintain its stewardship of state-of-the-art facilities, and develop a long-term visionary research program for advanced scientific computing.

**DISCUSSION**

**Reed** resonated with the message that business as usual is no longer sufficient. Development and execution of a strategy addressing STEM and workforce issues are crucial to secure the future of U.S. competitive capabilities.

**Seidel** appreciated the results and grave tone of the study. What is the envisioned role of NSF as a collaborating organization? **Dongarra** explained DOE requires larger HPC resources than NSF. NSF research is generally more basic, while DOE's application work dovetails with and practically advances applied activities. There are future grounds for collaboration, and NSF may continue providing the field's long-term historic computational underpinnings.

**Taylor** praised the report and emphasized select findings and recommendations. Sustained, long-term funding enables the discovery process by allowing exploration of different research pathways. In addition to competing with industry, the national labs are losing staff to academia; universities are recruiting now that ECP has ended. The stability of the nine-month salary plus the ability to conduct research or apply for grants over the summer is appealing. Co-design for microelectronics across the full stack is vital. **Dongarra** acknowledged comments, noting the DOE also does not afford the luxury of tenure.

**Bergman** suggested inclusion of more quantitative information in the report. How many people does ECP employ and what fraction will be lost with the conclusion of ECP and the onset of an at least six-month funding gap? Starting a new AI initiative will be difficult with a leaky workforce faucet. Long-term research funding is important. Nine-month university salaries cover teaching. Grants are needed for research, and computational research is expensive. Funding usually lasts only 12-18 months. Including figures in the report may increase impact. **Diachin** said DOE has been tracking ECP numbers. ECP touches a little more than 800 people at national labs; this figure excludes university subcontracts. Of these 800 people, 400 are FTEs. **Dongarra** appreciated these suggestions. **Reed** commented on the possibility of diverting facility funds to support people during the six-month gap. **Heroux** (chat) noted recruitment challenges due to a lack of guaranteed funding past December 31, 2023.

**Windus** reiterated the importance of long-term funding and the ability to explore multiple research paths. ASCR built great strength in high-fidelity, physics-based models; not losing these capabilities is important. **Dongarra** recognized these remarks.

**Matsuoka** commented ECP brought about an important and novel business model shift from asynchronous hardware and software design to co-design. Additionally, there was front-end planning to ensure applications were ready for operation on ECP machines as soon as they became available. Fugaku's approach has similarities, but the allocation for applications, 40% of funds, were delivered separately. Going forward, DOE does not appear to have committed resources for application teams, who have had to compete from day one for resources. This

strategy appears suboptimal. If the ultimate goal is to enable applications, committed, sustained funds are needed with planning for infrastructure timelines. **Dongarra** noted these comments.

**Berzins** wondered why there is no exit survey for the ECP to document both the impact of funding and current uneasiness at labs as funding ends. There is variety in how DOE funding is dispersed, and if everyone had to write five or more grants a year, the system might change quickly. The notion that short-term funding keeps one sharp is outdated. Writing lots of grants tires people out and does not align with the competition that currently exists in academia and the technology sector. There must be a hard look at realities if DOE's workforce is to be retained. It's all about the people. **Dongarra** agreed.

**Arthur** recognized funding issues extend to industry. Industrial lab have also seen declines in their ability to do fundamental research. There is focus on short-term objectives and funding must be chased on an annual basis, with possible exceptions in the financial sector, biotech, social media, and ad delivery. Declines in capabilities of GE, Bell Labs, and others affects research competitiveness. Thus, there is even more reason for resources like the national labs. Other countries make national resources available to programs like Fraunhofer-Gesellschaft (Germany) and Science and Technologies Facilities Council (STFC, United Kingdom-Research and Innovation). **Dongarra** concurred.

**Thomas** advised quantitatively highlighting not only talent loss but also loss of projects, applications, and scientific outcomes due to the end of ECP funding. Conveying a failed return-on-investment from the developed capabilities due to short-sighted fiscal planning is impactful.

**Giles** reminded the subcommittee of the 2020 report about the future of ECP. A third of the report addressed the impact of the program on the workforce and the country. How can HPC needs be communicated more effectively to those making appropriations? **Reed** remarked inflation is acting as a de facto budget cut and the debt ceiling limits budget growth. Therefore, the only viable strategy is emergency appropriations. Further funding would have to be folded into the base budget to address sustainability issues. This is a story of who will own the future of information technology and HPC, and perhaps it can be told in the context of politics that seem to have traction. i.e., competitiveness with China and the domestic workforce.

*All subcommittee members in attendance voted in favor of accepting the report.*

**EARLY SCIENCE ON FRONTIER EXASCALE SYSTEM**, Bronson Messer, Oak Ridge National Laboratory

Frontier is open for early production. All allocation programs were enabled in April 2023. More than 10M node hours have been delivered to >1,280 users representing 169 projects with funding from INCITE, ECP AD, the Argonne Leadership Computing Challenge (ALCC), and laboratory director's discretionary funds.

While Frontier was being developed, eight applications were prepared for early science through the Center for Accelerated Application Readiness (CAAR) program. Select applications spanned research domains and were developed at national labs, user facilities and/ or universities. Twelve additional ECP AD programs, also spanning research domains, made early use of Frontier.

Science highlights from CAAR applications featured: large-scale density functional theory calculations using LSMS and correlation analysis via CoMet. Science highlights from ECP AD featured the 2023 Gordon Bell Award Winner WarpX, hierarchical parallelism (ParSplice) from EXAALT, fully coupled modeling from ExaSMR. Additional science highlights showcased HAAC's contributions in computational astrophysics and GE's use of

Frontier to perform first-ever 3D eddy simulations for Revolutionary Innovation for Sustainable Engines (RISE) fan architecture.

From Titan to Frontier, the ECP has advanced the state of the art in computational capabilities. Advances now enabling resolution and timescales that confer 1) quantitative understanding and prediction; 2) the ability to move beyond constraints of periodic boundary conditions; and 3) the ability to perform fully coupled multi-physics calculations at each spatial point at new orders of magnitude.

**DISCUSSION**

None.

**NQIAC REPORT: RENEWING THE NATIONAL QUANTUM INITIATIVE: RECOMMENDATOINS FOR SUSTAINING AMERICAN LEADERSHIP IN QUANTUM INFORMATION SCIENCE,** Charlie Tahan, National Quantum Coordination Office

The NQI, passed in 2018, outlines the U.S. strategy for QIS. The NQI takes a science-first approach; provides infrastructure; builds a workforce; nurtures industry; balances economic and national security; and continues developing international collaboration. Additional technical and workforce strategic areas have been added since. Passage of the NQI resulted in formation of the five NSF Quantum Leap Institutes and the five DOE QIS Research Centers. NQI coordinating bodies include the NSTC Subcommittee on Quantum Information Science (SCQIS); National Quantum Coordination Office (NQCO), and NQI Advisory Committee (NQIAC). Industry additionally formed the Quantum Economic Development Consortium. After the National Defense Authorization of Act in FY22, the National Science and Technology Council (NSTC) Subcommittee Economic and Security Implications of Quantum Science (ESIX) was also created. NQCO coordinates daily implementation of the NQI, supports other subcommittees, and oversees interagency NQI coordination. NQIAC advises the president, SCQIS, and ESIX and independently assess the NQI's progress.

In preparation for the 2023 reauthorization of the NQI, NQIAC issued a report in June 2023 titled *Renewing the National Quantum Initiative: Recommendations for Sustaining American Leadership in Quantum Information Science.* Report findings address: 1) the U.S.'s capacity in quantum information science and technology (QIST); 2) QIST's role in U.S. economic and national security; and 3) outstanding scientific, engineering, and integration challenges. Overarching recommendations centered on 1) reauthorization and expansion of the NQI; 2) the U.S. QIST workforce; 3) U.S. QIST leadership; and 4) the pace of technology development, research programs, and maturation and scaling of systems to application. Nine detailed recommendations advised on 1) reauthorization of and appropriation for the NQI; 2) research expansion; 3) funding for industry-led partnerships; 4) equipment and infrastructure investments; 5) promotion of international cooperation; 6) promotion and protection of U.S. QIST R&D; 7) strengthening supply chains; 8) domestic talent development; and 9) attraction and retention of foreign talent.

On June 7, 2023, the NQCO offered testimony to the House Science Committee. Shared recommendations addressed: 1) reauthorization of the NSF and DOE QIS research and education centers; 2) expansion and broadening of QIS participation; 3) expanding core agencies with support from a dedicated international fund; 4) translation of QIS discoveries to commercial utility and agency missions; and 5) funding for infrastructure and facility upgrades.

**DISCUSSION**

**Reed** asked what is the biggest obstacle faced by NQI. **Tahan** identified the need for talent. The QIST ecosystem is at an early stage; development comes down to people. Following the 2020 Presidential proclamation limiting graduate student visas from China, a report on the role of international talent in QIS was released. Obtaining talent, not stealing blueprints, that will get countries ahead.

**Berzins** inquired about industry's current position on quantum. **Tahan** stated industry is still enthusiastic, but there is a growing realization of the expensive to develop quantum technology. There are questions about sustaining costs potentially for another 25 years, especially if the global economy continues to decline. Time, however, is needed to realize quantum computing. Quantum sensing is underfunded, though there are near-term applications in timekeeping, radiometry, and biomedicine.

**Brower-Thomas** asked about educational partnerships. Growing the number of U.S. QIS programs is a step, but more radical solutions are needed. **Tahan** relayed the CHIPS and Science Act authorized $8M/ year in sustained funding for K-12$^{th}$ grade education. The number of universities with PhD programs also need to be expanded beyond the current ~15 programs. The resulting workforce cannot scale to future industry needs. Growing the number of university programs in this space requires significant investments to build new labs, supply equipment, and train people. The NSF Expanding Capacity in Quantum Information Science and Engineering (ExpandQISE) program is an example of delivering such infrastructure and capabilities. NQCO also plans to hold an Uncomfortable Education Workshop to consider solutions that may upend normal agency practices to expand the workforce. **Reed** agreed alternative solutions are needed to broaden participation in STEM and increase the number of university programs. The U.S. is short millions of STEM workers. **Seidel** noted the University of Washington's Physics Department meets regularly with community college presidents to encourage them to invest in QIS programs for welding students. Many startups are limited by not having workers capable of building circuit boards. There are sociological challenges to advancing the field.

**PUBLIC COMMENT**

**None.**

*Reed adjourned the meeting at 12:45 p.m.*

*Respectfully submitted on July 7, 2023,*
*by Holly Holt, Ph.D. and Patrick Cosme, Ph.D.,*
*Science Writers, Oakridge Institute for Science and Education.*