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

PUBLIC MEETING MINUTES

May 29th, 2024

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 Wednesday, May 29, 2024 at The Pitch at the Wharf, 800 Maine Ave S.W., Washington, DC 20024 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.

DFO for ASCAC
Ceren Susut, Associate Director, ASCR

ASCAC members present in-person
Richard Arthur, General Electric (GE)
Sunita Chandrasekaran, University of Delaware
Roscoe Giles (Vice Chair), Boston University
Susan Gregurick, National Institutes of Health (NIH)
Mary Ann Leung, Sustainable Horizons Institute
Vanessa Lopez-Marrero, Brookhaven National Laboratory (BNL)
Satoshi Matsuoka, RIKEN Center for Computational Science
Irene Qualters, Los Alamos National Laboratory (LANL)
Edward Seidel, University of Wyoming
Valerie Taylor, Argonne National Laboratory (ANL)

ASCAC members present virtually
Martin Berzins (Chair), University of Utah
Jacqueline (Jackie) Chen, Sandia National Laboratories (SNL)
Mark Dean, University of Tennessee
Gilbert Herrera, Department of Defense
Anthony Hey, University of Washington
Alice Koniges, University of Hawaii
Alexandra (Sandy) Landsberg, Office of Naval Research
Jill Mesirov, University of California, San Diego
John Negele, Massachusetts Institute of Technology (MIT)
Sameer Shende, ParaTools Inc.
Krysta Svore, Microsoft
David Torres, Northern New Mexico College
James Whitfield, Dartmouth College
Theresa Windus, Iowa State University

ASCAC members absent
Keren Bergman, Columbia University
Tina Brower-Thomas, Howard University
Vinton G. Cerf, Google Inc.
Timothy Germann, LANL
Vivek Sarkar, Georgia Institute of Technology
Cristina Thomas, 3M Company (retired)
Juan Torres, National Renewable Energy Laboratory

Presenters

Ryan Adamson, Oak Ridge National Laboratory (ORNL) (virtual)
Debbie Bard. Lawrence Berkeley National Laboratory (LBNL) (virtual)
Amber Boehnlein, Thomas Jefferson National Accelerator Facility (JLAB)
Ben Brown, DOE Advanced Scientific Computing Research (ASCR)
Bill Carlson, IDA
Tiffany Connors, LBNL (virtual)
Lori Diachin, Lawrence Livermore National Laboratory (LLNL)
Sudip Dosanjh, LBNL (virtual)
Rafael Ferreira da Silva, ORNL (virtual)
Chin Guok, LBNL (virtual)
Harriet Kung, Acting Director, DOE Office of Science (virtual)
Inder Monga, LBNL (virtual)
Mike Papka, Argonne National Laboratory (ANL)
Amanda Randles, Duke University (virtual)
Dan Reed, University of Utah (virtual)
Geraldine Richmond, Under Secretary for Science and Innovation, DOE (prerecorded)
Arjun Shankar, ORNL (virtual)
Rick Stevens, ANL
Nhan Tran, Fermi National Accelerator Laboratory (Fermilab) (virtual)
Tom Uram, ANL (virtual)

Attending in-person

James Ang
Richard Carlson
Christine Chalk
Hal Finkel
Carol Hawk
John Hill
Jeff Hittinger
Barb Helland
Bronson Messer
Tzvetan Metodi
Michael Parks
Ashley Predith
Griffin Reinecke
Juan Restrepo
Damian Rouson
Suzy Tichenor

Attending Virtually

There were approximately 200 individuals present for all or part of the meeting.

Wednesday, May 29, 2024

OPENING REMARKS, Ceren Susut, Associate Director of the Office of Science for Advanced Scientific Computing Research

Susut convened the meeting at 10:08 a.m. Eastern Time (ET) and welcomed attendees, new members of ASCAC, and new Chair Martin Berzins. Susut is now Designated Federal Office (DFO). Thanks were extended to departing members.

Reed, departing Chair, recently testified to Congress on the future of the science and technology (S&T) enterprise. The U.S. is in a global competition with China and S&T is critical to soft power and ability to project influence globally. China has passed U.S. in numerous areas of global S&T. U.S. federal funding is flat and could be cut again. Internal division must be put

aside to create a coherent vision and strategy across government and industry. A coherent national strategy is the primary goal; business as usual is not going to allow the U.S. to realize the future.

Berzins, Chair, thanked Reed for his thoughts and expressed hope the group could assist the Office of Science (SC) in moving forward to address the challenges.

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

Susut reviewed organizational and personnel changes within DOE SC and ASCR and highlighted the Advanced Computing Technology (ACT) Division Director position open through June 17.

The ASCR FY25 budget request of ~$1.15B represents a ~13.5% increase over the FY24 enacted budget. Mathematical, computational and computer sciences research funds collectively increase ~ 35.9%, to ~$418.4M, with the following funding requested: Applied Mathematics ~$77.5M (~ +48.6%), Computer Science ~$86.7M (+30.0%), Computational Partnerships ~$93.4M (+24.3%), Advanced Computing ~$148M (~ +36.1%), and Energy Earthshot Research Centers (EERCs) $12.5M (+150.0%). The overall high performance computing (HPC) and network facilities request of ~$718.2M represents a small increase (~ +1.6%) from FY24. HPC and facilities funding requested includes: High Performance Production Computing $146.5M (~ +3.2%), Leadership Computing Facilities ~ $475.1M (~+0.25%), High Performance Network Facilities and Testbeds ~ $93.5M (~ +2.8%), and the Integrated Research Infrastructure (IRI) $3M (new). The High Performance Data Facility (HPDF) Construction request was $16M (up from $1M). No further funding for the Exascale Computing Project (ECP) is requested. Budget priorities include: advancing the exascale and AI-enabled science era; extending frontiers in AI for science, security, and technology; next-generation facilities; leveraging energy-efficient heterogeneous architectures; continued implementation efforts for the IRI initiative; maintaining strategic partnerships within DOE and between agencies; and broadening ASCR's impact through Reaching a New Energy Sciences Workforce (RENEW), Funding for Accelerated, Inclusive Research (FAIR), Computational Science Graduate Fellowships (CSGF), and the Established Program to Stimulate Competitive Research (EPSCoR).

The enacted FY24 budget is ~$1.02B, which represents a decrease of $52M/ 4.7% relative to the enacted FY23 budget. Desired outcomes in FY24 include: successful completion of the Exascale Computing Project (ECP); maintaining facility operations and existing upgrade projects, and continued planning of activities for National Energy Research Scientific Computing Facilities (NERSC)-10 and Leadership Computing Facility (LCF) upgrades; continued emphasis on foundational applied mathematics and computer science research to support artificial intelligence (AI), machine learning (ML), and quantum computing; initiation of Microelectronics Science Research Centers (MSRCs); continued efforts in broadening participation and retention of under-represented groups in ASCR's programs and workforce through RENEW, CSGF, FAIR, and EPSCoR; continued planning for the HPDF; maintaining strategic computational partnerships; and the continued investment in the SC Energy Earthshots initiative. Efforts are ongoing to reinvigorate ASCR research to respond to critical technology trends.

FY24 solicitation highlights include: Building a Core Research Base: Competitive portfolios for fundamental research, including a recompete of ASCR's "base math" portfolio, broadened to computer science; the SC Energy Earthshots Initiative, for which SC announced 29 awards in FY 2023 and ASCR supports eight EERCs; and the $160M lab call, MSRC Projects for Energy Efficiency and Extreme Environments, with proposals due July 25, 2024. Funding partially implements Section 10731, Microelectronics Research for Energy Innovation, from the

CHIPS and Science Act. Other FY24 funding opportunities (FO) were released, all of which reflect efforts in responding to critical technology trends. ASCR participation in solicitations led by other offices and agencies involves: DOE Office of Energy Efficiency & Renewable Energy (EERE), Advanced Materials & Manufacturing Technologies Office (AMMTO) for projects related to "Codes for High Performance Computing for Manufacturing"; the National Science Foundation (NSF), NIH and international partners for Collaborative Research in Computational Neuroscience (CRCNS); and with NSF for Correctness for Scientific Computing Systems (CS^2). Proposals for FY24 solicitations are under review.

The Consortium for the Advancement of Scientific Software (CASS) is a newly formed organization, sponsored by ASCR and established by Software Stewardship Organizations (SSOs), to advance the scientific software ecosystem. SSOs include the Collaboration of ORNL, LBNL, and ANL for Better Software (COLABS), the Center for Open-Source Research Software Stewardship and Advancement (CORSA), Frameworks, Algorithms and Scalable Technologies for Mathematics (FASTMath), the Post ECP Software Sustainability Organization (PESO), Scientific Discovery through Advanced Computing (SciDAC) Resource and Application Productivity through computation, Information, and Data Science (RAPIDS), Sustainability for Node Level Programming Systems and Tools (S4PST), The Software Tools Ecosystem Project (STEP), and The Center for Sustaining Workflows and Application Services (SWAS). ASCR provided Phase I funding to six pilot SSOs in FY23 after an open, competitive review process. ASCR provided Phase II funding to six SSOs for January 2024 after an additional round of proposal reviews.

DOE stewards the leading-edge high-performance computing capability for the nation: Aurora (ANL) debuted as second-fastest supercomputer in the world; Frontier (ORNL) remains world's fastest; El Capitan (LLNL, supported by the National Nuclear Security Administration (NNSA)) will be the third exascale computer. Aurora took the top spot in AI performance benchmark using only 89% of the system. The HPDF project, which was allocated $8M by FY24 appropriations, will have a hub and spoke distributed operations model. Primary infrastructure will be at JLAB with a resilience site at LBNL. The mission of HPDF is to enable and accelerate scientific discovery by delivering state-of-the-art data management infrastructure, capabilities, and tools.

ECP Critical Decision (CD)-4 Independent Project review (IPR) was completed in April and is closing out in July. The project resulted in many accomplishments for the community and the nation. The community is experiencing a new science era, brought about by exascale capabilities which enables the AI of tomorrow. Science highlights included: Exascale AI on Frontier Advancing Science, with the Oak Ridge Base Foundation Model for Earth System Predictability (ORBIT); and The National Aeronautics and Space Administration (NASA) use of Frontier to simulate a human-scale landing on Mars.

An executive order for the Safe, Secure, and Trustworthy Development and Use of AI was issued October 30, 2023, with directives to over 20 federal agencies, and deadlines ranging from 30 to 365 days. DOE has critical roles to play in developing guidelines, standards, and best practices for AI safety and security; promoting innovation and competition; protecting privacy; strengthening leadership abroad; and the federal government's use of AI.

DOE contributions to the National Artificial Intelligence Research Resource (NAIRR) pilot program include: extended operations of the ORNL's Summit through October 2024; access to the Argonne Leadership Computing Facility (ALCF) AI testbed; AI software and user outreach activities; participation in the NAIRR Pilot Program Management Office; and co-leading the NAIRR Secure Pilot with NIH to inform design.

ASCR's end-to-end approach to advance trustworthy an energy-efficient Frontier AI

involves four major thrusts, each with research, and facilities aspects. Hardware innovation, in conjunction with vendor partners, will utilize research for co-design to improve energy efficiency, and facilities to provide computing and automated labs testbeds. Breakthrough tools for trustworthy AI will utilize research for digital twins, software, and highly curated data, and facilities to provide prototypes for ultra energy-efficient data centers. Deep understanding of AI models requires research and development (R&D) in applied math and computer science with highly curated data, and the world's most capable LCFs to train models. Finally, AI-driven high-precision science R&D requires research on edge AI and robotics and knowledge extraction, and IRI, the Energy Sciences Network (ESnet), and AI-enabled real-time control.

Regarding community news, DOE and Japan's Ministry of Education, Culture, Sports, Science and Technology (MEXT) has renewed a Memorandum of Understanding (MOU), enabling continued collaboration in HPC, AI, and related areas. A series of research workshops are planned for the Summer of 2024. Several Principal Investigator (PI) meetings have been held, and outreach to, and participation from, surrounding universities, including minority-serving and emerging research institutions, have included poster sessions and other meeting activities.

ASCR holds virtual office hours on the second Tuesday of each month at 2 pm ET to broaden awareness of programs. No prior DOE funding is required, and researchers, educators, and leaders within research administration from all institutional types are encouraged to join. The ASCR website has more information, including slides and recordings of past events.

DISCUSSION

Chen asked for more information about SciDAC, whether the embedding model, which was successful for ECP will be used, and whether the MOU will involve technology exchanges only, or include co-development of new research. **Susut** explained SciDAC details are scarce, but partnerships will focus on exascale science and AI applications, and planning of the program's future cycles is ongoing. Successful models will continue to be utilized and applied to new research. The MOU will involve both technology exchanges and research collaborations. **Matsuoka** emphasized the MOU's focus on collaborations for new projects and activities, and the goal of generating new funding.

Herrera asked whether formal methods will be included in the expansion of research and mathematics. **Susut** confirmed, and mentioned the expansion strategy involves taking a broad approach and utilizing community feedback to identify areas of importance. Additional details should be available during next ASCAC meeting.

REPORT FROM THE FACILITIES OF THE FUTURE SUBCOMMITTEE, Ed Seidel, University of Wyoming, and Amanda Randles, Duke University

Seidel identified subcommittee members and explained work on the Facilities Construction Projects (FCP) charge, issued by past SC Director Asmeret Berhe, has been ongoing since late December 2023. The charge requires ranking facilities according to potential of positioning SC at the forefront of scientific discovery and includes a list of five specific ASCR facilities: ALCF, OLCF, NERSC, HPDF, and ESnet. Facilities were evaluated in terms of readiness for construction; sufficiency of R&D performed to ensure technical feasibility; extent to which the cost to build and operate the facility is understood; and site infrastructure readiness.

The subcommittee made three principal recommendations: support and develop the five ASCR facilities specified, all of which were rated as absolutely essential but had various states of construction readiness; recognize and manage ASCR facilities into an integrated computational Ecosystem with a science-driven imperative, to support national priorities and serve U.S

industry, and which will require new ways of governing and potentially funding; and launch a comprehensive R&D program, including a five-year timescale that informs pathways to future systems and prototyping, which in turn would support decadal systems, and is governed both within DOE and across agencies. Facilities rated as ready to initiate construction were ALCF-4, OLCF-6, NERSC-10, and ESnet-7. ALCF-5, OLCF-7, NERSC-11, ESnet-8 and HPDF Hub were rated as having significant scientific and engineering challenges to resolve, and HPDF Spokes 1 and 2 were rated as having mission and technical requirements not yet fully defined.

Recommendations were informed by four key findings: ASCR advanced computing systems continue to be critical for SC to remain at the forefront of scientific discovery; a combination of complementary facilities is needed to support DOE mission science, and through integration, the facilities should be considered as a single overarching ecosystem with multiple components; ASCR facilities are also important to other organizations, including but not limited to NNSA, NSF, NIH, the National Institute of Standards and Technology (NIST), NASA, National Oceanic and Atmospheric Administration (NOAA), DoD, and to U.S. industrial competitiveness; and the continued success of DOE mission science requires an "all hands on deck" approach to developing next-generation computing infrastructure.

Failure to follow recommendations risks the Nation's advanced computing ecosystem, including loss of U.S. global leadership in advanced computing; further destabilization of the computing hardware vendor ecosystem due to premature technology choices; inability to achieve DOE's science objectives, as well as collateral science effects at other agencies that depend on DOE; and new generations of systems with even lower efficiency, with concomitant scientific, technical, and political risks. Failure to adopt a long-term, integrated R&D program may lead to erosion or loss of program funding. Deep changes in research are anticipated for all SC offices due to the increasing integration of research disciplines, and the entire set of facilities should be seen as an integrated Ecosystem to support changing needs. An all-of government approach is required for success.

Arthur commented the cross-agency aspect is not just about research, but all requirements of success, such as data standardization strategies.

Seidel thanked the subcommittee and emphasized Randles's contributions as co-chair.

Susut thanked the group and checked for recusals. Seven recusals were noted (Mary Ann Leung, Sunita Chandrasekaran, Irene Qualters, Alice Koniges, Sameer Shende, Krysta Svore, and David Torres)

DISCUSSION AND VOTE ON REPORT

Gregurick asked how all 28 SC facilities would be connected and mentioned the Federal Government's difficulties in performing procurement processes. The vision of an ecosystem is powerful but there will be some limitations moving forward. **Seidel** explained a meeting was not held with all 28 facilities, but with other subcommittee chair/co-chairs, who corroborated the report's recommendations. ASCR is working on an all facility meeting as part of the IRI. The difficulties with the current generation of procurements were acknowledged. Next generation procurements should be informed by prototype systems, although coordinated, global procurements may be impossible.

Hey requested details on upgrades for Aurora and Frontier. **Seidel** replied R&D is required, and developments of the systems are on the next five year timescale. **Carlson** added the committee was given plans for future systems and did not address upgrades for current systems.

Matsuoka explained similar efforts are underway in Japan and DOE could use the subcommittee's plan for global coordination. The difficulty of building prototypes on a five-year

timescale was discussed and it was asked whether the prototypes mentioned would be design or physical prototypes, whether prototypes are worth the required hundreds of millions of dollars, and how governance will be organized at the national, federal, and international levels. The European High-Performance Computing Joint Undertaking (EuroHPC JU) was mentioned as a model of governance. **Seidel** explained several pathways need development and while the difficulty is acknowledged, prototypes are preferred if possible. **Reed** explained the need to avoid the constant building of machines and allow time to step back and ask questions, "bend metal", and try different things. Answers will require real computer architecture R&D and could be done with tens, not hundreds of millions of dollars. **Seidel** noted while discussions tend to get wrapped up in the "next machine", the report is about the ecosystem and the support of science. **Arthur** added co-design could involve rethinking algorithms and memory hierarchies. **Seidel** mentioned ASCR has a primary role in governance, but will need help from the coupling of DOE, the National Security Agency (NSA), and other agencies, both domestic and foreign, who are both users of the facilities and have capacity to conduct additional R&D. More discussion is needed to determine the most effective form of governance.

Berzins noted the report's lack of discussion of quantum technologies and thought an opportunity was missed to be more imaginative, move beyond current topics, and discuss the tremendous potential of rethinking standard architectures, AI, physics-based simulations, and quantum technologies. **Seidel** conceded the topics mentioned could have had a stronger presence in the report, but quantum technologies were implied within the scope of comprehensive R&D programs. Incrementalism should not be the approach. **Randles** agreed the report could have placed more emphasis on quantum, but surprisingly, quantum technologies were not heavily discussed during facilities meetings. The point made in the report was that tomaximize impact money should not be spent on the next incremental generation of the same hardware. An R&D program involving new hardware prototypes, new system software, and all other necessary components is required. **Seidel** added caution was taken against giving an unfunded mandate to SC/ASCR. The problems are too big to be done without interagency cooperation.

Chen asked for the vision of serving a variety of communities with the HPDF hub and spoke system, whether polling will be conducted to determine the use cases of different offices, and whether the public repository aspects could be picked up by industry, or ML and data science communities like Kaggle, having data sets available for training. **Seidel** confirmed that thought is going into all aspects mentioned. HPDF site selection completed in October, and spokes, and hub integration are yet to be defined, so there is still a lot of work to be done. **Reed** added defining spoke interaction, as well as how spokes are funded, is critical to build long term support for the user community. The committee's role was to raise and present motivation to answer HPDF questions, not directly provide all the answers.

Giles requested details on the timescale for ecosystem development, and ASCR's leadership role in solving related challenges, as outreach must start immediately. **Seidel** mentioned several aspects are already moving towards an integrated ecosystem; therefore, outreach would not start from scratch. However, linking procurements to what is needed will require planning to start immediately.

Windus expressed concern that governance would reduce system agility, possibly hindering centers from meeting missions, and asked how the various methods of user input would be consolidated. **Seidel** replied the report is not prescriptive. SC programs and lab directors will be involved in defining the governance system, but the current system is not sufficient and must be redefined. Managing input will be left to SC, but adviser committees must consider each facility's place in the overall ecosystem. The importance of user communities in R&D was emphasized.

Germann (written comment read by **Susut**) agreed with the need of an integrated ecosystem but expressed concern about the viability of a super facility and sought clarification on whether the ALCF and OLCF were considered as a single facility with two sites. **Seidel** shared concern over a super facility and explained the facilities would be linked, but maintain a differentiated set of capabilities. **Susut** confirmed ALCF and OLCF is a single facility with two sites.

Berzins called for a vote to accept the report. The report was accepted unanimously.

REMARKS FROM THE UNDER SECRETARY FOR SCIENCE AND INNOVATION, Geraldine Richmond, Under Secretary for Science and Innovation at the Department of Energy (prerecorded)

Richmond congratulated the DOE ASCR community for holding the top two positions in the world for most powerful supercomputers in both fidelity simulations and artificial intelligence. ORNL Frontier retains its top spot and Argonne's Aurora has officially broken the exascale barrier. As a result, climate researchers were able to simulate an Earth-year length cloud-resolving model in less than one compute day, earning them the first ever Gordon Bell Special Prize for climate research.

The just-completed ECP delivered on all its promises on-time and within budget, including improving energy efficiency, making the software stack available across the country, and providing a flexible platform that can be leveraged for applications that span the department's mission challenges. ASCR is advancing AI for Science in ways to meet the nation's future needs. The Facility Subcommittee report provides excellent suggestions for continued progress and addressing future problems. Dr. Berhe left SC well-positioned to meet present and future challenges.

VIEW FROM WASHINGTON, Harriet Kung, Acting Director of the Office of Science

Kung expressed gratitude to Dr. Berhe, whose accomplishments include driving efforts related to Urban Integrated Field Labs, fusion energy sciences public-private partnerships, exascale computing, and SC Energy Earthshots. Dr. Berhe served as the Head of Delegation to the International Thermonuclear Experimental Reactor (ITER), deepened relationships with international partners, broadened community outreach efforts, strengthened participation in inclusive research and capacity building programs, and brought increased rigor and a robust diversity, equity, and inclusion (DEI) effort to SC.

Dr. Berhe led SC realignment efforts and streamlined leadership structure to two Deputy Directorships: the Deputy Director for Science Programs (DDSP) and the Deputy Director for Operations (DDO). Leadership changes include: Dr. Linda Horton, Associate DDSP and Acting Director for Nuclear Physics (NP); Dr. Timothy Hallman, Senior Advisor on Equity, Inclusion, and Accessibility; and Dr. Andrew Schwartz, Acting Director for Basic Energy Sciences (BES).

Two new SC divisions were created effective April 12, 2024. The Fusion Energy Sciences (FES) Enabling Science and Partnerships (ESP) Division will address the expanded FES mission, established in the Energy Act of 2020, to develop a competitive U.S. fusion power industry, support enabling science programs, and grow partnerships with the private sector and international fusion ecosystem. The High Energy Physics (HEP) Accelerator and Technology (AT) Division will consolidate capabilities and expertise in accelerator research and deployment, steward the Accelerator Research and Development and Production (ARDAP) subprogram, and improve effectiveness of SC investments in critical technologies, such as quantum information science (QIS), AI, and microelectronics.

Dr. Berhe was a proponent of making SC accessible to the general public. To make its

mission more publicly accessible, SC reorganized its mission statement into three pillars: Driving Discovery Science for the Nation, Fostering Great Minds and Great Ideas, and Providing Unique World-Class Facilities. SC connects people with tools to unleash discovery and advance scientific innovation to drive energy and national security priorities.

The FY25 SC budget request totals ~$8.6B. Highlights include: $259M for AI research, an increase of $93.1M from FY24; $94.7M for microelectronics, an increase of $22M, which includes $45M for MSRCs; an $18.8M increase for U.S. Fusion Acceleration, which includes the Fusion Innovation Research Engine (FIRE) collaboratives; $20M for research on climate change and clean energy; $115M for SC Energy Earthshots, representing a $95M increase; $120M for RENEW, a $68.6M increase; and $64M for FAIR, an increase of $31.6M. Roughly $190M is allocated for scientific user facility operations, $50M for upgrading core laboratory infrastructure, an increase of $31.7M; and $5M for the Laboratory Operations Apprentice Program, an increase of $2M. $259M was requested for AI/ML, representing an increase of $93.1M. The Frontiers in Artificial Intelligence for Science, Security, and Technology (FASST) initiative was highlighted as a mechanism to recognize the value and mitigate the threats posed by AI/ML. The five areas of focus are: AI for Science, including Scientific AI Foundation Models (FMs); AI Hardware Innovation; AI for User Facilities and Advanced Instrumentation/ Technology; AI Tools for Design and Evaluation of Trustworthy AI Systems; and a diverse AI workforce.

The SC portfolio includes 28 user facilities, which support nearly 40,000 users, serving a broad-based community.

DISCUSSION

Giles expressed excitement over Kung's reflection on the facilities report. **Kung** noted the report broadened the lens to encompass R&D and the stewardship of relationships with other programs. Congratulations were given for adopting an excellent report.

Seidel felt the role of ASCR has been elevated to a "first-class citizen," and completely essential to DOE's goals. Thanks were extended to all the committee members. **Kung** agreed Seidel and Randles showed great leadership, particularly in reaching out to the other five FACA groups.

DOE EXASCALE COMPUTING PROJECT – THE FINAL UPDATE

Lori Diachin, Project Director, Lawrence Livermore National Laboratory

Diachin explained ECP successfully delivered on-time, under-budget, and far exceeded expectations. Four mission-needs statements were determined at the inception of ECP in 2016: maintain international leadership in HPC; promote the health of the U.S. HPC industry; deliver a long-term, sustainable software ecosystem that can be used and maintained for years to come; and ensure that exascale systems can be used to deliver mission-critical applications. Results were delivered for each statement. Close collaboration was critical to ECP's success: applications required work with software technology developers and performance engineers embedded at ASCR HPC facilities, especially Perlmutter, Frontier, and Aurora. HPC has advanced a great deal since ECP began and pre-ECP Design Forward and Fast Forward 1 and 2 programs had several payoffs in exascale systems, including Slingshot network architecture, distributed asynchronous object storage (DAOS) object store architecture, and AMD node architecture.

Supported by ~$1.8B over seven years, ECP was divided into three technical focus areas: Application Development, Software Technology, and Hardware and Integration. Success for ECP was measured using four Key Performance Parameters (KPPs). These were to demonstrate

performance improvement for mission-critical problems using 11 selected applications; broaden the reach of exascale science and mission capability in 14 selected applications; meet an aggregate capability integration score in 76 software products; and deliver 267 vendor baselined milestones in the PathForward element.

KPP-1 was both a science challenge and performance problem and was for applications doing well at the inception of ECP on petascale machines. Ten out of 11 applications met the objective. KPP-2 examined applications not at petascale performance at ECP inception. Users had to define a challenge problem and demonstrate the capability and effective use of the GPU/exascale system. Nine out of 10 SC applications and three out of four NNSA Advanced Technology Development and Mitigation (ATDM) applications were successful.
KPP-3 focused on the sustained integration of significant capabilities into client environments (e.g., application codes, co-design code, facilities/vendors, software standards, etc.). Each product needed to hit either four or eight integrations to earn a weighted point with 68 weighted points possible. As of April 2024, 66.5 integrations are complete with several more under review, well above the threshold score of 34. KPP-4 centered on PathForward program milestones and met 100%.

The KPP verification workflow involved many external subject matter experts (SMEs) and was a detailed and months-long process that included iterative SME review and PI response. Early skepticism of the ECP program waned as positive results accumulated. WarpX is an example of co-design that resulted in virtual design through simulation to dramatically cut accelerator size and cost, making its use in scientific and medical applications more practicable.

In early 2024, four ECP teams were selected to demonstrate the promise of AI/ML on Frontier. ExaWind used extremely high-fidelity models to develop surrogate models for unsteady loads on wind turbine blades for use in engineering design codes. ExaLearn worked with fusion scientists at Princeton Plasma Physics Laboratory (PPPL) and General Atomics to use AI/ML trained on experimental data to predict when disruptions may happen in tokamaks and apply real-time controls. The Exascale Atomistics for Accuracy, Length, and Time (EXAALT) project used high-precision quantum simulations to develop machine learned potentials to predict the survivability of plasma facing materials in ITER. Climate involved ML techniques applied to high-precision Simple Cloud Resolving Energy Exascale Earth System Model (E3SM) Atmosphere Model (SCREAM) calculations to automatically tune subgrid scale models related to clouds.

The KPP-2 application, Optimizing Stochastic Grid Dynamics at Exascale (ExaSGD) enables, “what if” evaluation of complex damage by optimizing weather scenarios and complex disruptions. The nature of data available required new solver strategies for GPUs for large sparse systems of equations. Strategies were developed through collaboration between Supernodal Lower Upper (SuperLU), Gingko, and ExaSGD teams.

The Extreme-scale Scientific Software Stack (E4S) HPC software ecosystem, a curated software portfolio built on software development toolkits (SDKs) will likely be a legacy of ECP. There are well over 100 packages all built using Spack-based distribution of software tested for interoperability and portability to multiple architectures. Software is available from source, containers, cloud, and binary caches and is an open resource for all supported by DOE and commercial entities.

Exascale applications are designed to be flexible and adaptive. The distribution of ECP programming languages and models has shifted over time. Several different programming models are used to achieve performance portability, such as GPU-specific kernels, Loop pragma models, C++ abstractions, and co-design frameworks. Performance portability was a key aspect of code design across teams.

ECP engaged in significant outreach and stakeholder activities in its final two years, to engage stakeholders in industry and other agencies to share new capabilities and broaden the community of potential users of exascale-ready applications and technologies. Notable collaborations occurred with NOAA, DOD, and TAE Technologies.

A successful CD-4 review was completed in April 2024. Key themes from the report were the idea of software as a facility, prioritizing software sustainability, encouragement of the various *-Forward programs as a bridge, and recognition of the culture change and strong collaborations created among applications, software technologies, facilities , NNSA and SC.

Aggressive use of project change requests was crucial to success, with most requests related to risk mitigation, scope adjustment, and management changes. Notable successes of ECP were integration across multiple disciplines, agility in project planning, and formal project management practices. Improvement opportunities included the impact of differences in funding, staffing, and sponsors; impact of using DOE Order 413.3B (acquisition of capital assets) for a software research development and demonstration (RD&D) project; and project planning for major transitions including post-ECP activities.

A mix of funding sources for ECP applications exists moving forward. SC application offices base programs, SciDAC, DOE Earthshots, NNSA, applied energy offices, and Laboratory Directed Research and Development (LDRD) are a few examples. About half of the teams have 50%+ of their ECP-level of funding, although the scope of research may have shifted. DOE is funding CASS to address the stewardship needs of the ASCR scientific software ecosystem. The PESO project is specifically focused on stewardship of the scientific software ecosystem.

DISCUSSION

Berzins applauded the work of the ECP.

Seidel encouraged heavy promotion of ECP-related science to generate interest from the next generation. The lessons learned can be integrated into recommendations. Concern was expressed for the personnel and talent base developed going forward. **Diachin** recognized some personnel left but very few who wanted to stay did not find funding. Transitions are still happening, perhaps into AI initiatives, and there is some concern about the FY25 budget.

Giles commended the degree to which ECP-developed software and technologies carried over to many applications, and asked about tracking the progress of ECP personnel over time, such as a “where are they now” in 10 years. **Diachin** explained 3,000 people were employed by ECP over the course of the project, from all of the DOE labs and numerous universities, although choosing a smaller population might be feasible and certainly interesting.

ADVANCING DOE’S INTEGRATED RESEARCH INFRASTRUCTURE (IRI): UPDATE ON PROGRESS, Ben Brown and Hal Finkel, ASCR; Mike Papka, and Tom Uram, Argonne National Laboratory; Debbie Bard, Tiffany Connors, Sudip Dosanjh, Chin Guok, and Inder Monga, Lawrence Berkeley National Laboratory; Ryan Adamson, Rafael Ferreira da Silva, and Arjun Shankar, Oak Ridge National Laboratory; Amber Boehnlein and Graham Hayes, Thomas Jefferson National Accelerator Facility; and Nhan Tran, Fermi National Accelerator Laboratory

The FY21 President’s Budget Request (PBR) included the Integrated Computation and Data Infrastructure Initiative, and an ASCR task force was launched the same year. Standup of the IRI Program is a DOE FY 2024-25 Agency Priority Goal and is now moving out of the initial stages of development. IRI will provide a foundational infrastructure for next-generational experimental science, post-exascale computing, AI, and other initiatives.

Governance structure for IRI was developed by SC Headquarters IRI Coordination Group

in 2022-23. In late 2023, ASCR directed the ASCR Facilities group to jumpstart the IRI Management Council, comprised of an executive committee, leadership group, and various subcommittees. The executive committee consists of ASCR facility directors, who play a key role overall, and DOE leadership, and is responsible for program oversight, strategic resource allocation, risk identification and management, and annual reviews. The leadership group prioritizes and organizes work, translates strategic goals into plans, provides outreach to DOE Program Offices, and works closely with the subcommittees. The IRI Program seeks to expand membership through a call, expected in August 2024, for participation in the technical subcommittees and user groups.

ASCR is implementing IRI through four major elements: investment in foundational infrastructure; formal coordination of projects; standup of program governance and technical activities; and deployment of a testbed across ASCR facilities. IRI requires an exponential amount of collaboration across ASCR facilities and multi-facility workflows are an integral part of infrastructure design and strategic planning. Even prior to IRI, researchers commonly engaged in multi-site workflows and resource use, notably through NERSC, ALCF, ESnet, and OLCF.

The IRI Architecture Blueprint Activity (ABA) gathered requirements from 24 science teams across the span of SC, collating and categorizing challenges scientists face in building workflows integrated across DOE resources. The IRI Framework comprises three Science Patterns and six Practice Areas. Science Patterns represent use cases across science domains: time-sensitive, data integration, and long campaigns. Practice Areas represent critical topics that require close coordination to realize and sustain an IRI ecosystem across institutions: workflows, interfaces, and automation; scientific data lifecycle; user experience; portable/scalable solutions; cybersecurity and federated access; and resource co-operations. A requirements analysis based on pathfinder and testbed projects and engagement activities is projected in the next year.

Pathfinder projects were nominated by program managers as ready to work on in the first stages of developing IRI technologies. These initial projects were light source projects in BES (Advanced Light Source (ALS), Advanced Photon Source (APS), Linac Coherent Light Source (LCLS), National Synchrotron Light Source II (NSLS-II)); DIII-D National Fusion Facility (Fusion Energy Science, FES); and Earth System Grid Federation (ESGF) at LLNL (Biological and Environmental Research, BER). Projects goals range from automated, resilient, fast turnaround for large-scale data analysis coupled with simulation (DIII-D); added services and increased accessibility to the world's largest climate model data archive (ESGF); and access to a multi-tiered architecture for full lifecycle data management (light sources).

IRI technical work is managed by subcommittees that coordinate work, identify goals, and estimate required resources. Membership is open to DOE labs and user facilities. Subcommittees report to the IRI Leadership Group and have the ability to stand-up working groups for specific matters. Working group members may include non-DOE consultants. Three program area subcommittees have begun work toward FY 24-25 goals: Trusted IRI Designs (TRUSTID), Outreach/Engagement, and Interfaces. TRUSTID was established to assist in enabling cross-facility workflows within the IRI ecosystem. An example is the establishment of a federated identity access management system that can be used across participating facilities. Outreach and engagement are necessary to build momentum and identify and develop a requirements strategy for IRI. Human contact to bridge gaps in culture and expectations is essential in this work. Future work includes attendance at conferences and PI and Collaboration meetings, and a proposed IRI Community Meeting in early 2025. The Interfaces group will work on remote access to machines. Areas still under development include IRI Allocations Program, Scheduling/Preemption Technical Subcommittee, Data Movement Technical Subcommittee, and Software Deployment and Portability.

IRI is driving major infrastructure investments as seen in the Request for Proposals (RFPs) for NERSC-10, OLCF-6, and ALCF4. HDPF was conceived as a data focused component of the IRI ecosystem and ESnet 6/7 is designed with IRI in mind. The IRI Testbed is in-process with key FY24-25 activities like cataloging existing testbeds at ASCR facilities; co-designing and implementing a plan to use the ESnet testbed network to connect existing testbeds; developing a best current practice (BCP) connection service template that can be used to connect non-ASCR test environments to the IRI testbed; publish facility-specific testbed access procedures, and agreed upon procedures (AUPs); and identify and execute early test cases to validate initial IRI testbed design.

ASCR research is strongly aligned with IRI objectives, particularly in recent portfolios like the FY21 Integrated Computational and Data Infrastructure (ICDI), which included Experimental/Computational/Computer Science collaborations and Intelligent Distributed Infrastructure Simulation Capabilities, and the FY23 Distributed Resilient Systems, Scalable System Modeling and Adaptive Management and Partitioning of Resources. Under the FY21 Data Reduction Science FOA, one team examined fast ML for science and tools and techniques for the extreme edge. The team has determined robust and efficient methods, tools, workflows, collaborations, and support for co-design. Edge capabilities and interface/synergy with HPC are critical to develop within IRI objectives.

DISCUSSION

Seidel conveyed excitement for the vision described, and drew parallels between the Facilities report findings and elements in IRI, especially a broader governance model and building an overall ecosystem. A budget line for the IRI effort, and ways in which the community could help was requested. **Brown** remarked the FY25 PBR contained a modest beginning request for IRI but stressed a lot of support will come from a coalition of the willing through the power of collective voices, but funding will be needed.

Gregurick wondered if, considering pending proposals, each facility was partitioning off capabilities for IRI work when use is dependent on the proposal being accepted. **Brown** acknowledged resource allocation can be difficult, especially between multiple facilities, and emphasized the idea of an ecosystem must be an exercised goal, not just a philosophical statement.

Taylor requested additional information about the pathfinder projects. **Brown** explained the origin of the pathfinder projects is in existing partnerships in which researchers are willing to take the next step and engage in a little risk. The name is an expression of this moment and how the effort can start, rather than a term of art. The portfolio should be broadened as soon as resources and human capacity allow. **Dart** noted the intent going forward is for projects to run with the benefits of the added capabilities.

Giles reminded the group, relying on emergence from existing efforts tends to work slowly and a formal effort to support IRI and the facilities ecosystem may be needed. ECP is a model of what a highly organized, well-funded effort can look like.

UPDATE ON FRONTIERS IN AI FOR SCIENCE, SECURITY, AND TECHNOLOGY (FASST), Rick Stevens, Argonne National Laboratory

Stevens revealed AI engagement across all fields is exponential and becoming ubiquitous. AI is becoming the dominant driver/signal of techno-economic progress and competition, and provides the potential for non-linear progress in technological, economic, and national security domains, with competition and positioning between Western democracies, semi-aligned petro-states, and adversarial Sino-Russian players. The U.S. is the clear leader in

commercial AI, but the U.S. government's focus has been on mitigating AI risks, and it is under-investing in non-commercial, non-defense frontier AI systems development and adoption.

DOE is in a unique position for AI leadership due to operation of the most capable scientific computing systems and the world's largest collection of advanced experimental facilities; being the largest producer of classified and unclassified scientific data in the world; supporting the strongest foundation combining physical, biological, environmental, energy, mathematical and computing sciences; employing the largest scientific workforce in the free world; and maintaining strong ties with private sector technology and energy organizations and stakeholders.

DOE is the best-placed agency to lead, having the technical ability to lead and partner with other agencies to advance the use of transformational AI across the government. Government efforts to regulate AI will require independent expertise in AI technologies and risks. DOE is well-positioned as a trusted neutral party because effective execution of the DOE mission requires experience with the best and most powerful scientific tools. AI systems in the hands of bad actors pose a new type of asymmetric threat that will need new ideas for risk management, and private sector AI efforts alone will not address the deep scientific and national security requirements of DOE use cases.

DOE has gathered input from over 1,300 researchers since 2019 and a 2020 ASCR report recommended a major DOE AI for Science (AI4S) program. On October 30, 2023, the White House issued an executive order on AI including tasking DOE with expanding partnerships with industry, academia, other agencies, and international partners to utilize DOE's computing capabilities and AI testbeds to build FMs that support new applications in science, energy, and national security. DOE is also tasked as the lead agency through NNSA to reduce the risks at the intersection of AI and chemical, biological, radiological, and nuclear (CBRN) threats. On May 8, 2024, DOE announced the FASST initiative. The Bipartisan Senate AI Working Group released a roadmap on May 14 that recommended $32B/year investment for non-defense AI across the government.

Current AI is trending toward FMs and a smaller number of universal models, characterized by emergent behavior and homogenization. Notably, models are integrating knowledge without human supervision, leading to the Platonic Representation Hypothesis: that as models get bigger and trained on more data, they appear to converge. This trend will tend to continue and get stronger. Assuming this to be true, the Spring 2022 conference AI for Science, Energy, and Security (AI4SES) was organized around six clusters: AI for advanced properties inference and inverse design; AI and robotics for autonomous discovery; AI-based surrogates for HPC; AI for software engineering and programming; AI for prediction and control of complex engineered systems; and FMs for scientific knowledge tasks.

FMs are large scale models trained on sizeable datasets from many sources (text, papers, datasets, code, molecules, etc.) that may undergo additional training to improve the human interaction experience. FMs are remarkably flexible and exhibit emergent behaviors, i.e., are capable of tasks not trained to do, and applications are built on top of them. There are multiple early efforts underway in DOE labs to create FMs explicitly targeting scientific use cases. FMs are trained on pieces of information referred to as tokens. Opportunities for FM use in science are based on their ability to summarize and synthesize information, generate plans and solve logic problems, generate hypotheses and eventually new theories for exploration. FMs may be natively multi-modal, with input tokens of varying types, that are integrated and transformed to provide results in the format requested. Scientific FMs, sometimes called domain FMs, may have inputs and outputs that are not natural language but, for example, genome sequence or crystal structure.

FASST is currently organized around four pillars: data, compute, models, and applications. It is unlikely that one foundation model would be powerful enough for all science needs, although combining multiple domain models using a natural language model to allow communication with humans would be a possibility. Creating domain models that map to DOE mission areas, such as high-energy physics, materials chemistry, or biology, is a first step. Frontier AI, the biggest and most powerful AI, is large-scale science and large teams (hundreds of people) and high amounts of computing power are needed to build a single model. DOE must decide whether to use outside models and fine tune them with the appropriate data or to create proprietary models that are trained *de novo*.

FASST's broad goals are to ensure U.S. (DOE) global leadership in technical capability for its missions in Science, Energy, and National Security; to create, deploy, and sustain world-leading frontier AI systems and applications for DOE mission areas to provide advantage to the U.S. and its partners; to increase productivity and capabilities of DOE laboratories and academic, agency, and international partners; and to develop an AI-forward workforce for DOE. Outcomes include acceleration and improved effectiveness in discovery science; acceleration, reduced risk, and improved translation in energy transition; and anticipated and mitigated risks and accelerated mission in national security.

Each of the four pillars has high level objectives. DOE data, estimated at ~1,000 trillion tokens, must be organized and readied for training as quickly as possible. Investments in multiple frontier AI training platforms and AI inference optimized systems are needed, and energy efficiency is a key factor. A pipeline for continuous training of domain-oriented models will be necessary. Hundreds to thousands of applications relying on these models are anticipated.

Target production of 100 trillion tokens in the first few years is necessary as data preparation rate-limits the effort. To train ~10 frontier FMs per year, AI training resources greater than 10x current exascale AI flops must be built. Thousands of inference servers are necessary as is an increase of approximately 2000 Full-Time Equivalents (FTEs). Application development should begin now, with the ability to swap FMs as better models become available. Although DOE has a great deal of raw data, very little has been translated into high quality data tokens.

DISCUSSION

Chen asked about the possibility of using hybrid systems, such as domain-specific models that inform FMs, to make training more efficient and more accurate. **Stevens** agreed, noting the need for simulations to generate synthetic training data. Challenge problems and hard sets of evaluations are needed to integrate simulations with AI.

Mesirov shared experiences using commercial general FMs and supplemental training resulting in surprisingly positive results. **Stevens** acknowledged fine-tuning existing models could be more efficient and cautioned uses in national security or life-critical decision making might require more transparency in training. Some specialized domain models will need to be built from scratch.

Qualters wondered if industry partnerships had been considered, noting a faster pace in development and better compensation than at labs. NSF is an example of efficient training and DOE could use this model to develop efficient algorithms. **Stevens** confirmed partnerships with major players were possible, although there was concern about the volume of data DOE would share, and the increase in compute costs. Methods optimization is a topic of research.

Matsuoka worried simulations are being overlooked and emphasized the importance of communicating to funders the efforts are not just about AI. **Stevens** noted traditional simulation will evolve and AI will be used on top of resulting work. A rhetorical construct to allow for better

understanding and discussion would be helpful, so people understand the continuing need for simulation even with AI.

PUBLIC COMMENT

None.

Giles adjourned the meeting at 5:05 p.m.

Respectfully submitted on June 26, 2024
By Ann B. Gonzalez, J.D., M.S.I. and Patrick Cosme, Ph.D.
Science Writers, Oak Ridge Institute for Science and Education.