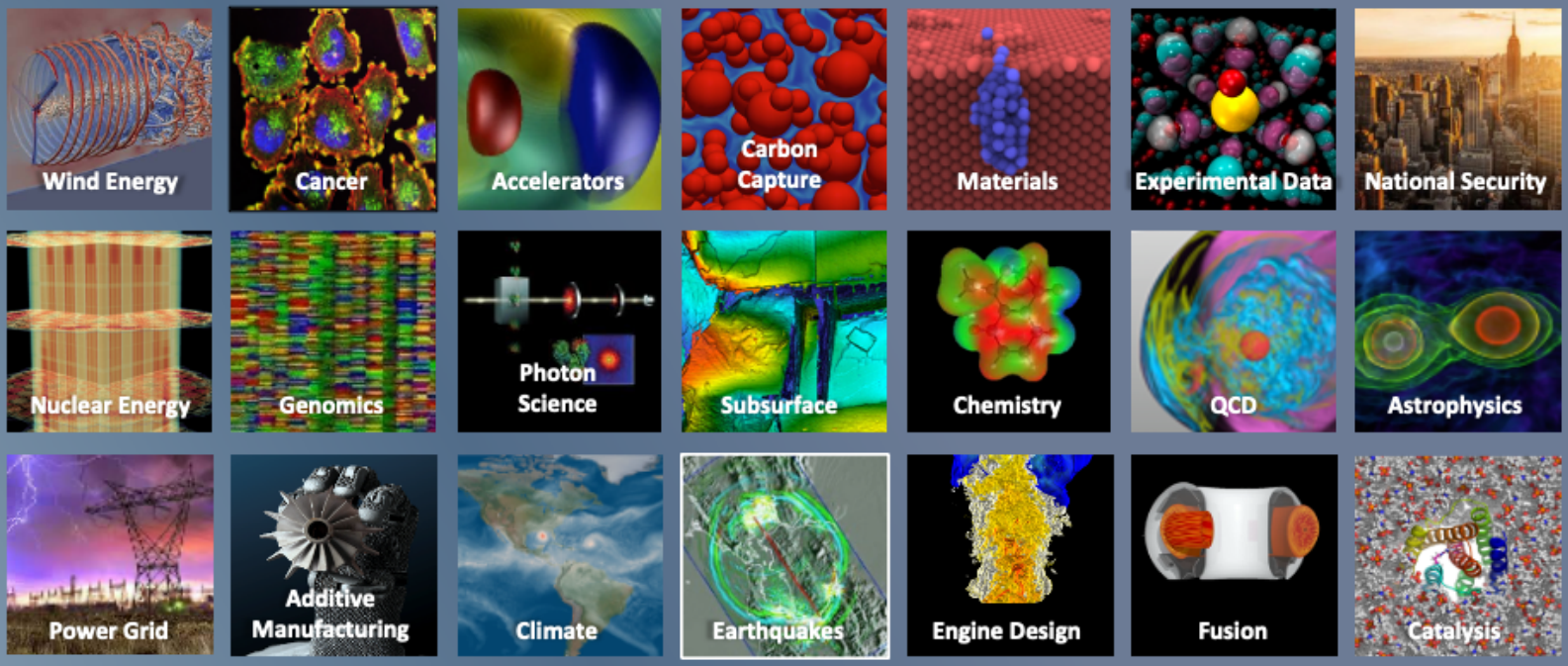
Can the United States Maintain Its Leadership in High-Performance Computing?

A report from the ASCAC Subcommittee on American Competitiveness and Innovation to the ASCR office



Chair
Jack Dongarra, University of Tennessee, Knoxville & Oak Ridge National Laboratory

Vice Chair
Ewa Deelman, University of Southern California

Subcommittee Members
Tony Hey, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell
Satoshi Matsuoka, RIKEN & Tokyo Institute of Technology
Vivek Sarkar, Georgia Institute of Technology
Greg Bell, Corelight
Ian Foster, Argonne National Laboratory & University of Chicago
David Keyes, King Abdullah University of Science and Technology
Dieter Kranzlmueller, Leibniz Supercomputing Centre & Ludwig Maximilian University of Munich
Bob Lucas, Ansys
Lynne Parker, University of Tennessee, Knoxville
John Shalf, Lawrence Berkeley National Laboratory
Dan Stanzione, Texas Advanced Computing Center
Rick Stevens, Argonne National Laboratory & University of Chicago
Katherine Yelick, University of California, Berkeley & Lawrence Berkeley National Laboratory

Department of Energy
Office of Science
Washington, DC 20585

Office of the Director

Professor Daniel A. Reed, Chair of the ASCAC
Senior Vice President for Academic Affairs
Professor of Computer Science and Electrical & Computer Engineering
The University of Utah
201 Presidents Circle, Room 205
Salt Lake City, Utah 84112-9007

Dear Professor Reed:

Thank you for your work as Committee Chair on the Advanced Scientific Computing Advisory Committee (ASCAC's) and for the ongoing review of the collaboration with the National Cancer Institute. The ASCAC recommendations will help us to improve the management of this important program.

As you know, the Administration and Congress have been keenly interested in the recent issues with the supply chain and U.S. competitiveness and innovation. Looking to the future, we want to ensure that the U.S. continues to be a leader in advanced computing, high end computational science and engineering, advanced scientific networks, and the fields and workforce that underpin these efforts.

To that end, we must develop and maintain world-leading capabilities in key technologies, especially microelectronics, high performance computer architectures and software, computer science, applied mathematics Artificial Intelligence, Quantum Information Science; and also provide compelling, inclusive, and equitable opportunities for all those who want to work in this fast paced and ever-changing area of research.

Therefore, I request that ASCAC develop a report to address the following questions:

- How can the Department maintain critical international cooperation in an increasingly competitive environment for both talent and resources? In areas where the U.S. is leading, how can we sustain our roles and attract the best industry and international partners? In other areas, how can the Department build and maintain its reputation as a "partner of choice"? In general, are there barriers that can hinder our ability to form effective and enduring international and industry partnerships?

- Identify key areas where the U.S. currently has, or could aspire to, leadership roles in advanced computing and high-end computational science and engineering, including unique or world-leading capabilities (i.e., advanced scientific facilities, testbeds and networks) or leading scientific and technical resources, such as highly trained personnel and supporting infrastructure. This may include emerging areas or opportunities that offer significant promise for leadership.

- To preserve and foster U.S. leadership roles within reasonable resource constraints, are there particular technical areas or capabilities that could be emphasized? Are there

other technical resources and capabilities that could be leveraged in to achieve these goals, possibly through collaborations within and beyond the ASCR community?

- How can programs and facilities be structured to attract and retain talented people? What are the barriers to successfully advancing careers of scientific and technical personnel in advanced computing, computational science and engineering, and related fields and how can the Department address those barriers? A complete answer to these questions should address how we can ensure that we are recruiting, training, mentoring, and retaining the best talent from all over the world, including among traditionally underrepresented groups within the U.S.

We would appreciate receiving a written report by the Spring meeting in 2023.

If you or the subcommittee chair have any questions, please contact Christine Chalk, Designated Federal Official for ASCAC at 301-903-5152 or by e-mail at christine.chalk@science.doe.gov.

I appreciate ASCAC's willingness to undertake this important activity.

Sincerely,

J. Stephen Binkley
Acting Director
Office of Science

ASCAC SUBCOMMITTEE MEMBERS

Chair

Jack Dongarra, University of Tennessee, Knoxville & Oak Ridge National Laboratory

Vice Chair

Ewa Deelman, University of Southern California

Subcommittee Members

Tony Hey, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell
Satoshi Matsuoka, RIKEN & Tokyo Institute of Technology
Vivek Sarkar, Georgia Institute of Technology
Greg Bell, Corelight
Ian Foster, Argonne National Laboratory & University of Chicago
David Keyes, King Abdullah University of Science and Technology
Dieter Kranzlmueller, Leibniz Supercomputing Centre & Ludwig Maximilian University of Munich
Bob Lucas, Ansys
Lynne Parker, University of Tennessee, Knoxville
John Shalf, Lawrence Berkeley National Laboratory
Dan Stanzione, Texas Advanced Computing Center
Rick Stevens, Argonne National Laboratory & University of Chicago
Katherine Yelick, University of California, Berkeley & Lawrence Berkeley National Laboratory

Table of Content

ASCAC SUBCOMMITTEE MEMBERS....3
Executive Summary....6
Key Findings....7
Key Recommendations....9
1. Introduction....12
2. Current Challenges to ASCR's Leadership....16
Technical Changes....17
Market Pressures....17
Geopolitical Changes....18
Changing International Research Landscape....19
DOE ASCR Research Funding....21
Findings....22
3. Building and Maintaining Strategic Industry and International Partnerships....23
Successful Collaborations in Standards and Software....23
Industry Collaborations....25
United States' Collaboration in the Global Context....26
Findings....27
Recommendations....28
4. Critical Scientific Areas for Leadership in ASCR....29
High-End Computational Science and Engineering....29
Artificial Intelligence for Science and Engineering....32
Leading-Edge Computing Architectures....35
Advanced Networking and Future Internet Architectures....37
5. Cultivating and Sustaining Advanced Research Capabilities....39
Facilities....39
Findings....42
Recommendations....42
Careers in National Laboratories....45
Societal Relevance....46
Engagement with Academia....46
International Perspective....48
Increasing the Domestic Talent Pool....49

Findings 49
Recommendations 50
Conclusions 52
References 53
Appendix I: Historical Perspectives on ASCR's Leadership in Computational Science and Engineering and HPC Systems 57
Applied Mathematics 57
Computer Science 59
Supercomputing Facilities, and the Complementary Relationship between Research and Facilities 60
Computational Partnerships 61
DOE Graduate Fellowships 61
Industrial Partnerships 62
Appendix II: Exascale Projects in Europe and Japan 63
EuroHPC Joint Undertaking 63
The Japanese FugakuNEXT Project 64
Abbreviations and Acronyms 65

Executive Summary

The United States (U.S.) is no longer the unambiguous leader in the vitally important field of high-performance computing (HPC). Japan, the European Union (EU), and China have fielded systems that are on par with our fastest supercomputers. The supply chain for everything from semiconductors to scientific software is globally distributed. Yet our economic future and security depend critically on our ability to innovate faster than our competitors, and the speed of innovation depends increasingly on large-scale computational science and engineering and thus HPC. How should the United States respond to this challenge? This report seeks to initiate a new and potentially transformative national discussion on this vital question.

The Department of Energy's (DOE) Advanced Scientific Computing Research (ASCR) program is well-positioned to make informed, targeted decisions about where the United States should cooperate and where it should compete in the global market for scientific exploration and discovery. By setting its sights on problems critical to our nation and the world, by establishing productive new collaborations, and by making strategic investments, ASCR can restore and maintain U.S. scientific leadership in the critical areas described in this report while strengthening our research infrastructure and training a large, diverse cohort of scientists. In doing so, ASCR and its scientists will pave the way for a secure and prosperous future for America.

For more than 30 years, the ASCR program has provided the HPC and networking capabilities and expertise needed to support DOE's mission to advance the national, economic, and energy security of the United States. The program now faces the challenge of developing and deploying the next generation of HPC systems and technologies, as well as supporting the application of HPC and artificial intelligence (AI) technologies to a wide range of scientific and engineering research problems. Through its research and development efforts, the ASCR program must also advance the state of the art in HPC and accelerate the pace of scientific discovery and technological innovation.

Fulfilling this promise will require significantly increased investments, as well as innovative policies and programs. This subcommittee is aware that we are making recommendations and calls for action at a time when federal resources are limited. We understand that a wide range of competing priorities must be balanced by the nation's leaders and that there is a need to leverage resources in new ways and seek efficiencies in facilities and operations. However, we must not let these realities limit our imagination or silence our advocacy. The ASCR program is a key part of the U.S. research infrastructure and an important component of economic growth and U.S. competitiveness. ASCR has a responsibility to pursue its mission, including advanced scientific computing, applications of AI technologies, and the required advanced research facilities, with determination and enthusiasm.

To fulfill the scientific enterprise's responsibility to the nation, the ASCR program must not only develop and publish a clear vision with an associated list of goals, priorities, and recommendations but also demonstrate scientific leadership by consistently securing long-term funding. This will allow the program to build on its achievements to date, to realize its ambitious vision, and to make lasting contributions to the field.

The report makes the following high-level findings and recommendations.

Key Findings

1) **Science and engineering applications of national importance will continue to require increasingly more capable advanced computing systems** to model complex phenomena, process, analyze, and manage vast amounts of data, and support cutting-edge experiments. Meeting those requirements so as to maintain international leadership will require major and sustained advances in computing, networking, mathematical, and AI technologies. The national labs and their university partners are uniquely qualified to produce those advances, but only if supported appropriately in terms of leadership, vision, and predictable and sustained funding.

2) **Led in large part by DOE, the United States has been an international leader in applied mathematics and in computational science and engineering (CS&E) research** and has used that expertise to develop and deliver unique modeling and simulation capabilities for national priorities in science, energy, and nuclear security. The United States has also been a leader in computer science, with DOE's role focused on those areas related to HPC (e.g., programming, parallel algorithms, and performance optimization techniques) as well as networking and data science (methods and tools for scientific discovery).

3) **Big data and HPC are both important to scientific discovery and are synergistic**. Experimental facilities across DOE's Office of Science are increasing the demands for leading-edge computing and networking facilities, methods, and services. These demands include the ability to move, analyze, share, and manage exponentially growing datasets from observational sensors and increasingly powerful scientific instruments and to use AI technologies to integrate that data with physics-based and data-driven models, which may themselves produce enormous datasets and require massive computing for model training and inference.

4) **The Exascale Computing Initiative (ECI) is an exemplar of U.S. leadership in high-performance computing**, incorporating the latest mathematical and computational innovations into scientific applications, creating a comprehensive exascale software stack, and advancing the capabilities at the leadership-class computing facilities to enable future scientific breakthroughs.

5) **DOE has a history of working closely with industry partners** to develop, deploy, and apply advanced technology, particularly in the context of leadership-class computer systems and cutting-edge network services. DOE laboratories work closely with end users from industry and have achieved numerous high-impact results that extend the capabilities of participating companies.

6) **The end of the Exascale Computing Project (ECP) is both a success and a huge risk.** The project delivered great capabilities, both human and technical. Now, however, DOE is highly vulnerable to losing the knowledge and skills of trained staff as future funding is unclear.

7) **U.S., DOE, and ASCR leadership in key areas is under threat**. This situation is due to increased international competition (e.g., it is reported that China may deploy ten exascale machines by 2025) and geopolitical changes (e.g., a less cooperative and more competitive relationship with China), as well as increased market pressures in the United States that draw talent, capital, and attention toward near-term commercial objectives.

8) **The technology landscape has fundamentally changed:**
   a) Dennard scaling ended a decade ago and the effect of Moore's law is now waning.
   b) Huge investments in computing by hyperscalers (e.g., cloud and social networking companies) are shaping the marketplace toward their specific needs.
   c) Artificial intelligence-related computation is now a major performance driver for high-end HPC systems in industry and within hyperscaler data centers.
   d) The rise of custom/ semi-custom silicon (wafer-scale AI chips, chiplets, extensible or even open instruction set architectures, etc.) creates new possibilities to leverage commodity technologies.
   e) There is now investment in potentially disruptive technologies, such as quantum computing and networking (devices, architectures, models); however, these technologies may take decades to refine and fully mature.

9) **Unlike in the past, today's scientific research landscape and HPC supply chain is horizontal and international**, including hardware/ software/ networking components and talent. Leadership in HPC requires proactive, long-term, and sustained engagement with this broad international ecosystem, as in other Office of Science disciplines (e.g., High Energy Physics (HEP), Fusion Energy Sciences (FES)). Industry partnerships are essential and merit attention and improvement. Particularly in co-design, there are lessons to be learned from ECP and other international efforts so that the process can be improved in the future.

10) **ASCR funding levels for research are declining in real terms** (and are spread more thinly across new research directions such as quantum information science (QIS) and AI/ machine learning (ML)). At the same time, funding for facilities is on the increase to meet the outyear requirements of supporting the exascale platforms deployed in ECI; existing research funding is increasingly organized in short-term competitive tracks, with reduced allocations to "stable" base funding; and ECP is ending with no follow-on program to allow the cadre of well-trained, skilled, and talented researchers — precisely in the most competitive domains of HPC and AI — to remain in the DOE labs. The resulting uncertainty is generating much anxiety among lab staff, in particular for junior researchers, and does so at a time when talent competition from industry is increasing. Therefore, there is a significant risk to ASCR's leadership in the research and development of innovative technologies and solutions.

11) **The attractiveness and prestige of careers in national labs have been on the decline** because of internal and external factors, including lack of long-term program vision and stable funding from within the labs, and increased competition from industries such as HPC, AI, and big data from outside. Autonomy and flexibility in lab careers have also decreased. The COVID-19 pandemic resulted in many companies offering more flexible

work arrangements than did the national labs, including joint appointments with academia and industry as well as joint open-source projects, which are opportunities to increase research impact and reduce the compensation gap.

12) In the current resource-constrained climate, **big science and advanced scientific computing and networking increasingly require international collaborations** and bring benefit to the parties involved in these efforts. We see examples of fruitful exchanges of personnel, ideas, software, and technologies at a global scale. Significant opportunities for international collaboration exist in areas such as standard interfaces and libraries, which are not closely linked to large commercial markets or national security interests.

13) As has been shown in exascale and in previous advances in computing, achieving the breakthroughs in science needed by DOE and the nation requires innovation in both hardware systems and software infrastructure. Moving the **ASCR facilities forward will continue to require an interdisciplinary approach anchored in co-design, rather than a reliance solely on the vendor marketplace**. ASCR will see success from continuing to encourage collaboration across science teams, computing researchers, facilities staff, and vendors.

Key Recommendations

1) Building on its existing strengths in CS&E advanced computing and unique user facilities, ASCR must focus its future efforts on achieving and sustaining leadership in four key areas:
    a) High-end modeling and simulation for science and engineering (e.g., applied math, software, advanced applications);
    b) Artificial intelligence for science and engineering (e.g., AI methods, software, data sets, advanced applications);
    c) Leading-edge computing architectures and systems on the path beyond exascale (e.g., hardware architecture, software, deployed infrastructure);
    d) Advanced networks and future internet architectures for an integrated research infrastructure (e.g., architecture, software, deployed infrastructure).

    Note that all four of these areas align with the White House list of critical and emerging technologies [1].

    Each of these four areas has long-term research challenges that should be pursued through a combination of base program funding (promotes career development) and opportunistic calls (provides flexibility). Each area also demands the development and deployment of infrastructure (e.g., codes, libraries, models, HPC, AI, data and edge hardware facilities, national facilities) that supports the broader research enterprise.

2) ASCR leadership should work with the DOE labs to develop a decadal-plus post-exascale vision and strategy that builds on ASCR's strengths in mathematics and computing

research working together with DOE's world-class facilities. The focus should be on providing sustained investments to preserve and extend ASCR's current leadership in CS&E research and multidisciplinary team science while also establishing new application areas in emerging topics such as digital twins and AI for science, energy, and security, together with addressing daunting computing challenges as Moore's law fades.

The strategy should include development of an associated ASCR technology and investment roadmap that includes the following:

a) A plan for key technology investments post-exascale, including:
   (i) a multicycle decadal facilities roadmap to meet the increasing needs for computation in modeling, simulation, and AI;
   (ii) power and cooling considerations needed to field globally leading systems, as well as aggressive research to improve hardware energy efficiency; and
   (iii) emerging and unconventional architecture considerations, with pathfinding activities for alternatives that arrive externally.
b) Reinvestment in areas where ASCR has already established a leadership position, lest the U.S. lose that lead and the intellectual resources that underpin that capability (e.g., software tools and numerical libraries for scientific computing).
c) Emphasis on proactive, forward-looking investments in emerging areas where DOE is well-positioned to establish leadership (e.g., large-scale AI methods focused on world-leading scientific problems in the DOE mission space).
d) Focus on maintaining and developing human capability. Key to this focus is a compelling, long-term vision supported by stable funding models for long-term research, to recruit and retain top talent in advanced scientific computing, with a special emphasis on developing traditionally underrepresented groups.
e) An explicit role for industrial partners to help with retention. Particular attention should be given to the advantageous role that joint appointments and other types of collaborations can have on enhancing both the human capabilities of the labs and the DOE technology footprint.

3) ASCR needs to articulate a vision, associated goals, and milestones for international collaboration focused on post-exascale computing and networking. ASCR should work with the labs to identify critical research and facilities opportunities that may require international partnership to create and sustain international leadership, either because of the scale of investments needed or because of the unique capabilities that international partnerships can provide. ASCR should work to establish trust relationships with strategic partners, evangelize and socialize these efforts, define agreement structures (perhaps beyond the traditional memorandum of understanding (MOU)), and provide resources to develop flexible multiparty collaborations.

4) ASCR needs to invest in long-term forward-looking co-design research in advanced computer architecture and system concepts to identify potential solutions for sustaining

continued scientific productivity increases for future scientific computing systems. Such a co-design effort will require substantially increased government investment in basic research and development. In addition, DOE should fund the building of real hardware and software prototypes at scale to test new ideas using custom silicon and associated software.

1. Introduction

Scientific discovery has played a vital role in the prosperity of America. It is difficult to think of any aspect of our lives or work that has not been influenced, shaped, or improved by science and technology. For over 75 years, our national investment in American science and scientists has brought immeasurable benefits to our country. Basic scientific research, which seeks to understand fundamental principles, often leads to unexpected discoveries that serve as the foundation for innovation and technological advancements. Many technologies that we rely on today originated from basic research conducted in the United States using computational science and engineering.

CS&E is an interdisciplinary field, combining elements of computer science, mathematics, and domain-specific knowledge to create computational tools and methods for solving problems that cannot be solved analytically or by experiments alone. CS&E methods are employed across numerous areas, including physics, chemistry, biology, engineering, and finance. They allow scientists to analyze and understand phenomena that would be too complex, too time-consuming, or even too dangerous to study by using traditional methods, and to make predictions and design experiments based on these analyses. CS&E methods play an important role in data analysis and visualization, providing scientists with the tools they need to make sense of the vast amounts of data generated by modern experiments, observations, and simulations. They allow scientists to identify patterns and trends and to test hypotheses and theories. Overall, CS&E is an essential part of modern scientific discovery.

CS&E relies heavily on advanced scientific computing (ASC), including leadership-class computing facilities, to perform increasingly complex and data-intensive simulations and analyses. Increased computation power allows scientists to solve bigger and more complex problems and simulate more realistic models, leading to more accurate results and opening new research opportunities. It also allows for more efficient use of resources, faster completion of calculations, and greater flexibility in the types of problems that can be tackled.

Led in large part by DOE, the United States has been an international leader in applied mathematics and in computational science and engineering research **and has used that expertise to develop and deliver unique modeling and simulation capabilities for national priorities in science, energy, and nuclear security. The United States has been a leader in computer science, with DOE's role focused on those areas related to HPC (e.g., programming, parallel algorithms, and performance optimization techniques) as well as networking and data science (methods and tools for scientific discovery).**

For many years, DOE laboratories led the world in the delivery and application of extreme-scale computing systems. In 1997, the ASCI Red system at Sandia National Laboratories broke the teraflop barrier; in 2008, the Roadrunner supercomputer at Los Alamos National Laboratory reached the petascale barrier; and in 2022, Oak Ridge National Laboratory's Frontier passed the exascale mark. These advances were achieved by highly productive partnerships between computer scientists at DOE laboratories, in U.S. universities, and in U.S. industry.

Yet in the past three decades, U.S. leadership in HPC has eroded (Fig. 1). In terms of HPC systems, the first serious challenge came from Japan, with vector mainframes in the 1990s and then a series of leadership-class systems in this century (i.e., the Earth Simulator, K, and now Fugaku). The situation today is somewhat murky, as Chinese researchers are presenting exascale scientific results but not disclosing publicly the capabilities of the machines that generated those results. Still, it seems clear that China has at least matched U.S. HPC capabilities. The proliferation of HPC technology overseas has been followed by development of expertise in CS&E and first-rate applications. As a result, in many aspects of CS&E, the United States now finds itself in a position where it is one of several peers.

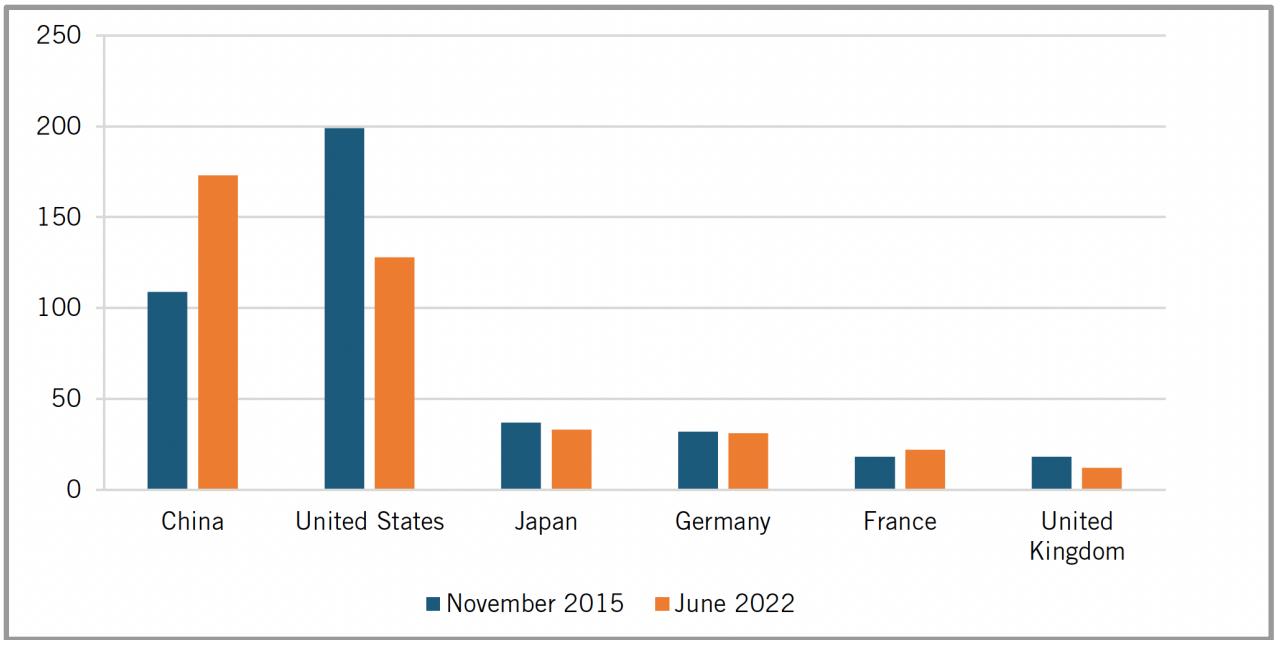


Fig. 1: Comparison of the number of leading supercomputers per country based on TOP500 data (http://www.top500.org)

The United States has historically been the global leader in computer science research, as illustrated by metrics such as the number of scientific publications [2]. Over the past decade, China has increased its investments in computer science education and research and has incentivized publications in high-quality venues. Today, China tops the authorship ranking in computer science conferences and journals, as illustrated by Table 1 from Wikipedia [3].

Table 1: Top 10 Countries in Terms of Scientific Publications [3]

Rank	Country	Number of scientific publications (2020)	Scientific publications per capita (in ppm)
1	China	744,042	527
2	United States	624,554	1875

3	United Kingdom	198,500	2959
4	India	191,590	138
5	Germany	174,524	2097
6	Italy	127,502	2159
7	Japan	127,408	1016
8	Canada	121,111	3184
9	Russia	119,195	819
10	France	112,838	1664

Without ongoing U.S. investment in basic and computational science and high-performance computing, future discoveries and technological innovation by U.S. scientists and institutions will be hindered. There is a legitimate concern that the United States is falling behind other countries in its investments in scientific research and development. Evidence shows that the United States is not keeping pace with other countries in funding these critical areas. For example, China is on track to surpass the United States in spending on research and development, which would be the first time in a century that the United States is not in the top position. Other countries are also investing in advanced research tools and providing long-term support for programs of all sizes. A 2022 report [4] noted that the United States ranked sixth in total research and development (R&D) intensity, 13th in government R&D, and tenth in basic science intensity. R&D intensity measures R&D investments as a share of a country's gross domestic product, commonly viewed as an indicator of a nation's innovative capacity.

Given these concerns, the Advanced Scientific Computing Advisory Committee (ASCAC) has been tasked by the U.S. Department of Energy Office of Science to (1) identify key research areas in advanced scientific computing; (2) evaluate U.S. competitiveness in these areas, from the perspective of research outputs, major research facilities, tools, and funding mechanisms; and (3) recommend strategies to improve the U.S. position compared with its global competitors.

ASCAC selected its member Professor Jack Dongarra (University of Tennessee, Knoxville, and Oak Ridge National Laboratory) to form and lead an international subcommittee on International Competitiveness. He formed this subcommittee to encompass diversity in terms of expertise (high-performance and distributed computing, artificial intelligence, computing architecture, quantum computing, networking) and experience and to include members from DOE national laboratories, academia, industry, and international partners. In addition to Jack Dongarra, the subcommittee comprises vice-chair Ewa Deelman (University of Southern California) and members Tony Hey (Rutherford Appleton Laboratory, Science and Technology Facilities Council), Satoshi Matsuoka (RIKEN and Tokyo Institute of Technology), Vivek Sarkar (Georgia Institute of Technology), Greg Bell (Corelight), Ian Foster (Argonne National Laboratory and University of Chicago), David Keyes (King Abdullah University of Science and Technology), Dieter Kranzlmueller (Leibniz Suprcomputing Centre and Ludwig Maximilian University of Munich), Bob Lucas (Ansys), Lynne

Parker (University of Tennessee, Knoxville), John Shalf (Lawrence Berkeley National Laboratory), Dan Stanzione (Texas Advanced Computing Center), Rick Stevens (Argonne National Laboratory and University of Chicago), and Katherine Yelick (University of California, Berkeley and Lawrence Berkeley National Laboratory (LBNL)).

The subcommittee identified four broad areas while preparing this report: (1) creation and maintenance of strategic industry and international partnerships, (2) critical scientific areas for leadership in ASCR, (3) needed advanced research capabilities, and (4) talent recruitment and retention. These areas were chosen as representative examples of the competitiveness challenges facing many activities within the jurisdiction of the ASCR program. To provide evidence-based conclusions about U.S. competitiveness in these areas, the subcommittee gathered data from various sources, including scientific literature and presentations at major scientific conferences. The report includes examples of the relevant data. The subcommittee also consulted with many leading scientists to identify issues related to U.S. competitiveness, such as funding mechanisms, global competition for scientific talent, and limitations on access to major research facilities, and potential solutions to these issues.

The subcommittee searched for specific examples of research in the identified critical areas to demonstrate both the social and the economic benefits of such research and some of the constraints and challenges faced by U.S. science and scientists. The report includes nontechnical descriptions of these examples, many of which provide a personal perspective on the science.

The results of these investigations reveal a complex but clear picture: U.S. scientific leadership, in the identified critical areas and beyond, now faces significant challenges, and consequently the country's ability to compete internationally is at risk.

In this highly competitive world and with limited financial and human resources, the United States and ASCR need to invest in CS&E and AI to maintain and enhance strategic international partnerships in order to enable the next generation of scientific advances. We already see that the EU, Japan, and China are investing heavily in HPC and AI. ASCR needs a vision and associated strategy for the post-exascale and post-Moore's law era. It needs to invest in research, development, collaboration, and talent in key areas such as high-end modeling and simulation for science and engineering, AI for science and engineering, leading-edge computing architectures and systems, advanced networks, and future internet architectures.

It is also crucial to attract and retain scientific talent. In the past, gaining research experience in the United States was seen as essential for any aspiring young scientist, and thus the United States was able to attract thousands of highly skilled individuals worldwide with little effort. Many of these highly talented individuals ultimately decided to stay and build their careers in the United States. However, this is no longer the case. The nation's declining ability to attract and retain international talent is reflected in the decreasing number of foreign students, postdocs, and early career scientists who choose to study and work in U.S. universities and laboratories [5], [6]. This is a significant problem because it limits the talent necessary to drive scientific discovery and innovation in the country. ASCR needs to find a way to reverse this trend.

2. Current Challenges to ASCR's Leadership

The computing facilities at the DOE labs, together with ESnet – the high-performance network connecting supercomputers, large-scale facilities, researchers, and instruments at the DOE labs – are a unique and world-leading resource for U.S. science and innovation. Historically, DOE ASCR has been a leader in the areas of applied mathematics, computer science, supercomputing and advanced networking facilities, and computational partnerships and cross-disciplinary technology translation. Appendix I provides an overview of many of the accomplishments in these areas. However, as we will describe in more detail in this section, the geopolitical and economic landscape is changing and negatively affecting the U.S. position in the world of science and engineering.

A recent Special Competitive Studies Project (SCSP) report titled "Mid-Decade Challenges to National Competitiveness" made it clear that U.S. leadership in science and innovation is under threat from the rapid rise of China as a major technological and military power [7]. The SCSP report identified three areas of technology in which the United States cannot afford to lose: microelectronics, fifth-generation wireless technology (5G), and AI.

In the area of microelectronics, 92% of the U.S. supply of leading-edge semiconductor chips are produced in Taiwan (where "leading edge" is defined as 7 nm process or better), underlining the vulnerability of the U.S. supply chain. The Creating Helpful Incentives to Produce Semiconductors (CHIPS) Act, together with European investments in semiconductor fabrication, should begin to address some of this vulnerability by providing access to facilities that can manufacture high-performance integrated circuits. In addition, the U.S. government recently announced new limits on the sale of semiconductor technologies to China.

The United States and other Western countries clearly missed out on the development of competitive, market-ready equipment to support the implementation of 5G networks. As a result, Chinese companies were well on their way to controlling much of the future network hardware required for both the global internet and for mobile communications. However, U.S. export controls, combined with a diplomatic campaign to persuade other countries not to allow Chinese 5G technologies into the network core, have slowed Chinese progress toward total dominance in the 5G arena. This comes at a time when the race to create 5G applications in autonomous systems, advanced manufacturing, and the Internet of Things is just beginning. Moreover, it is imperative that the United States be a leader in the development of 6G technologies and the design and implementation of future internet architecture standards and technologies.

In the AI space, the SCSP report says that "intelligent systems and applications driven by computing power, algorithms, and data will connect a constellation of technologies to transform entire industries.'' For DOE, AI technologies have the potential to transform whole areas of science and engineering. The success of Google's DeepMind United Kingdom (UK) subsidiary in solving the "protein folding Grand Challenge'' is one such example [8]. With leading-edge HPC simulation and modeling, together with huge datasets from DOE's world-leading experimental facilities, ASCR has the potential to be a world leader in a whole range of AI for Science applications.

In this section we examine some of the technical shifts, market pressures, and geopolitical changes that present significant challenges to ASCR's leadership in CS&E.

Technical Changes

Approaches for achieving increased computational capabilities are changing rapidly, as scientists and engineers tackle more ambitious questions with existing computational methods and tools and pursue entirely new methods and applications. For example, computational methods used previously to study how increased greenhouse gas concentrations may affect global climate are now being applied to study possible outcomes at regional and even local scales. Computational methods in materials science design are being applied not just for predicting the properties of individual materials but also for inverse design to identify new materials with desired properties. These and many other important applications require orders of magnitude more computational power than is available today just to run the codes that scientists already have; at the same time, researchers are hard at work developing new applications that, for example, leverage various forms of deep neural networks to guide computational campaigns and to extract structure from complex datasets in new ways. These methods are proving effective in many areas but are highly computationally demanding and arguably may consume as much as or even more computing power than the simulation applications that have long dominated facility workloads.

Another challenge is the increasing difficulty of building more powerful computers. As the rate of growth in microprocessor performance declines, the need for innovation in all areas of the advanced computing stack increases. Energy costs make continued scaling of existing technologies increasingly impractical. Ideas such as specialized accelerators, reduced-precision arithmetic, new computational methods, and integration of machine learning are being explored, but none offers a clear path to the orders-of-magnitude improvements required to meet new needs. Understanding how to leverage such advances will require substantial coordinated effort involving facilities, software and methods, and applications, similar to that undertaken in recent decades but likely at a larger scale because of the anticipated greater complexity of future architectures.

Market Pressures

The computing industry is now dominated by a small set of cloud companies. Facebook (now Meta), Amazon Web Services (AWS), Apple, Netflix, and Google (now Alphabet) (collectively referred to as FAANG), together with Microsoft and their Chinese counterparts Baidu, Alibaba, and Tencent (BAT), are called *hyperscalers*. Their market capitalization is an order of magnitude greater than that of the companies that supply the components with which both HPC and cloud computing systems are built. Naturally, the computing marketplace focuses on the hyperscalers' needs, a situation that does not bode well for the much smaller science and engineering communities that have historically driven HPC developments. For example, increasing emphasis is given to low-precision arithmetic operations suitable for AI computations, rather than to the higher precision generally needed for science and engineering.

Hyperscale data centers dwarf today's scientific HPC centers by multiple quantitative measures, including footprint, power consumption, and the number of servers installed. Yet while many HPC jobs can be run effectively on cloud computers, important differences exist between cloud and

HPC systems. Because cloud data centers typically do not have to be installed in secure facilities where space and power may be constrained, they can be built in a location that is advantageous from the point of view of power usage effectiveness (PUE). The hyperscalers design their systems to maximize capacity and overall throughput for a broad range of applications; therefore, unlike today's exascale systems, they have a heterogeneous mix of compute nodes, with different processors, memories, and accelerators (AWS offered 33 different instances in December 2022). Cloud data centers deliver computing services that grow and shrink elastically based on aggregate customer demand and their networks (both local-area and wide-area) are optimized for a variety of ever-changing commercial workloads. In contrast, local networks within scientific computing centers and wide-area networks that interconnect them are designed to minimize latency and maximize throughput, so that HPC performance and scalability can be optimized. HPC applications typically have less stringent resilience requirements than do interactive cloud services that need to meet stringent service-level agreements.

These differences mean that, in general, it is not feasible for DOE to either outsource its HPC workload to the cloud or order a cloud data center instead of an HPC machine. This is not to imply that the HPC community cannot benefit from collaborating with the hyperscalers and adopting technology from them. Organizations such as Open Compute are developing standards for common functions so that those functions can be commoditized to reduce costs.

Hyperscalers are a major driver of innovation in software, and, in particular, software for managing large quantities of computation and data. As with hardware, these efforts tackle somewhat different problems from comparable software efforts in the HPC community but present opportunities for productive collaboration. For example, increasingly interactive use of HPC systems is spurring investigations of Kubernetes as an alternative, in some environments, to the batch schedulers normally used in HPC.

The foundation of the exponential growth experienced by HPC over the past half-century has been the continuous improvements in the semiconductor technology with which systems are constructed. Dennard scaling has long since ended and, arguably, Moore’s law as well. Nevertheless, with the deployment of extreme ultraviolet (EUV) lithography, transistor density continues to grow. Manufacturers are increasingly specialized, with foundries such as Taiwan Semiconductor Manufacturing Company (TSMC), Samsung, and GlobalFoundries producing devices designed by fabless semiconductor companies such as Advanced Micro Devices (AMD), Apple, and Nvidia. Increased design automation and open-source hardware such as RISC-V are lowering the cost of designing application specific integrated circuits (ASICs), making it possible for a new generation of ML startups such as Cerebras, Groc, and Samba Nova to create specialized devices. However, these advancements are made to support the hyperscalers and the needs of their applications rather than the needs of DOE CS&E.

Geopolitical Changes

Foreign competitors are rapidly learning how to build and deploy supercomputers with effectiveness comparable to that of supercomputers deployed in the United States. (Indeed, in terms of raw numbers of the fastest computers, China appears to be well ahead of the United States

and has announced plans to deploy as many as ten exascale systems by 2025.[1]) This development is problematic for several reasons. First, foreign leadership in supercomputing is likely to translate into leadership in many other areas of importance to the United States, from materials design to defense technologies. Second, if foreign vendors start selling supercomputers that are cheaper and/or more effective than those of U.S. vendors, the U.S. advanced computing industry will suffer, making it harder for DOE labs and other U.S. entities to acquire the most powerful systems. Third, if the fastest computers are overseas rather than in the United States, the best scientists are likely to direct their efforts to developing their applications for those computers, with the result that vital expertise will spread more rapidly to our competitors and that the best codes will run less well, or not at all, on U.S. supercomputers. *(*As a historical example, we note that the Japanese Earth Simulator, the fastest supercomputer in the world from 2002 to 2004, attracted many U.S. teams, who developed there rather than on U.S. systems and, furthermore, were required to provide their code to the Japanese in return.) Fourth, a decline in the relative performance of U.S. systems will make retention and recruiting of top talent more difficult.

Changing International Research Landscape

A recent United Nations Educational, Scientific and Cultural Organization (UNESCO) report [9] found that while the United States invests significant funding in research, increased investments by countries such as China mean that the United States is not keeping pace (Fig. 2). Additionally, the global share of researchers in the United States decreased between 2014 and 2018 as the number of researchers in the EU, China, and the Republic of Korea increased. This downward trend can also be seen in the decrease in the percentage of U.S. publications and patents, while the biggest growth is seen in China. The report concludes: "The USA faces increasing competition in science, technology, and innovation from Asian players particularly, China, the Republic of Korea, and India. This competition is likely to intensify" [9].

Figure 2 shows that it is important to combine U.S. resources and talent with other strategic partners to address global problems. We have seen such successful global collaborations, particularly during the COVID-19 pandemic, when the world came together to share data, methods, treatments, and vaccines to combat the virus.

[1] https://www.datacenterdynamics.com/en/news/china-may-be-planning-10-exascale-supercomputers-by-2025/

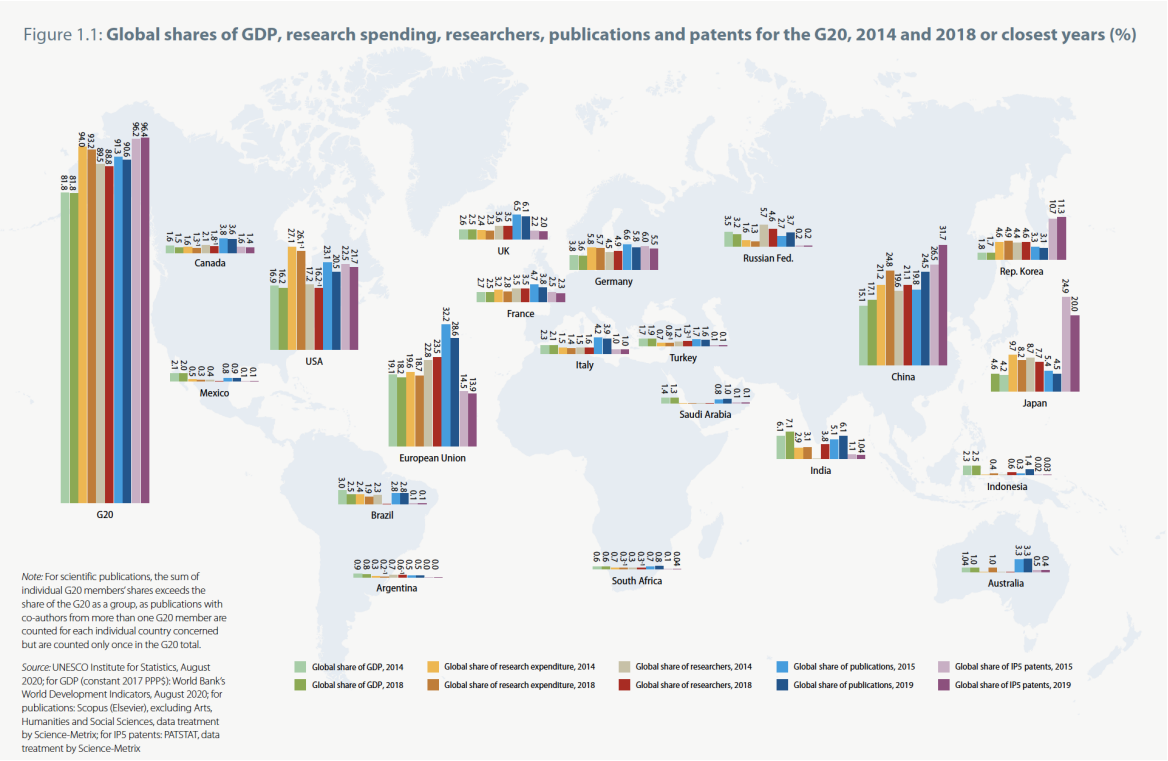


Fig. 2: Global shares of GDP, research spending, researchers, publications, and patents for the G20, 2014 and 2018 or closest years (%). Source: [9]

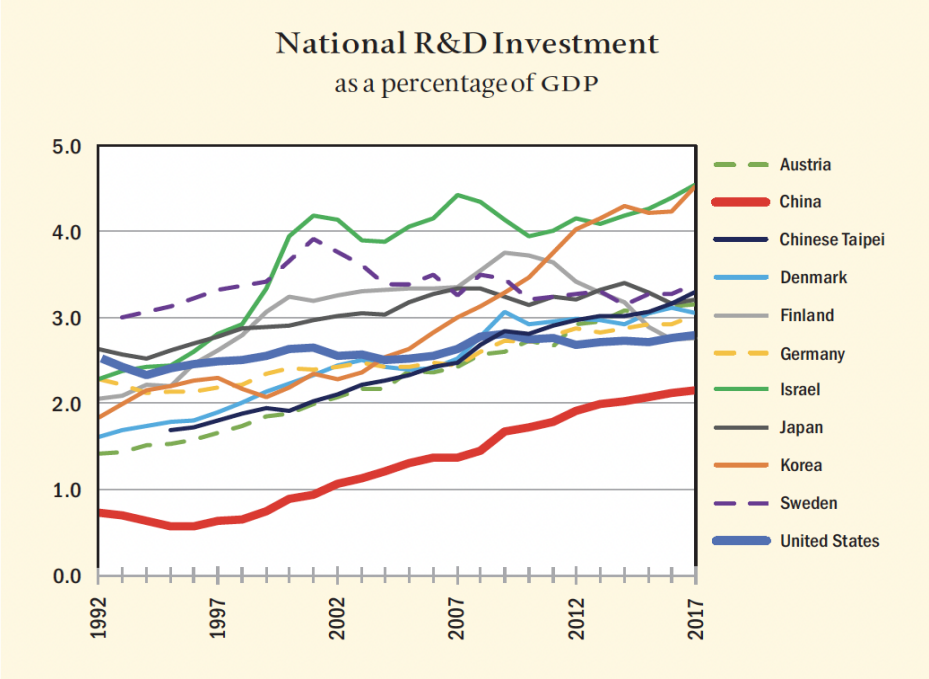


Fig. 3: National R&D investment as a percentage of GDP. Source: [10]

The American Association for the Advancement of Science (AAAS) report [10] discussed the R&D conducted in industry but noted that companies focus primarily on short-term returns on investment rather than long-term research funding. The report made the case that "most transformative research … is now much more dependent on government and other (nonbusiness) sources of funding such as private philanthropy."

We also see increased funding specifically in HPC across the globe, particularly in China, the EU, and Japan. While some efforts, like those in China, are in direct competition with the United States, others in the EU and Japan offer opportunities for strategic collaboration. In the EU, the European High Performance Computing Joint Undertaking (EuroHPC JU) is funding large-scale HPC efforts that recently resulted in the Large Unified Modern Infrastructure (LUMI) and Leonardo systems being placed as numbers three and four, respectively, on the TOP500. In Japan, significant effort is being spent on a ten-year roadmap for post-exascale research infrastructures (FugakuNEXT). Appendix II details these projects.

DOE ASCR Research Funding

Figure 4 shows the absolute funding within ASCR, divided between research and facilities. Funding for research has remained relatively flat with most of ASCR's investment directed to the facilities.

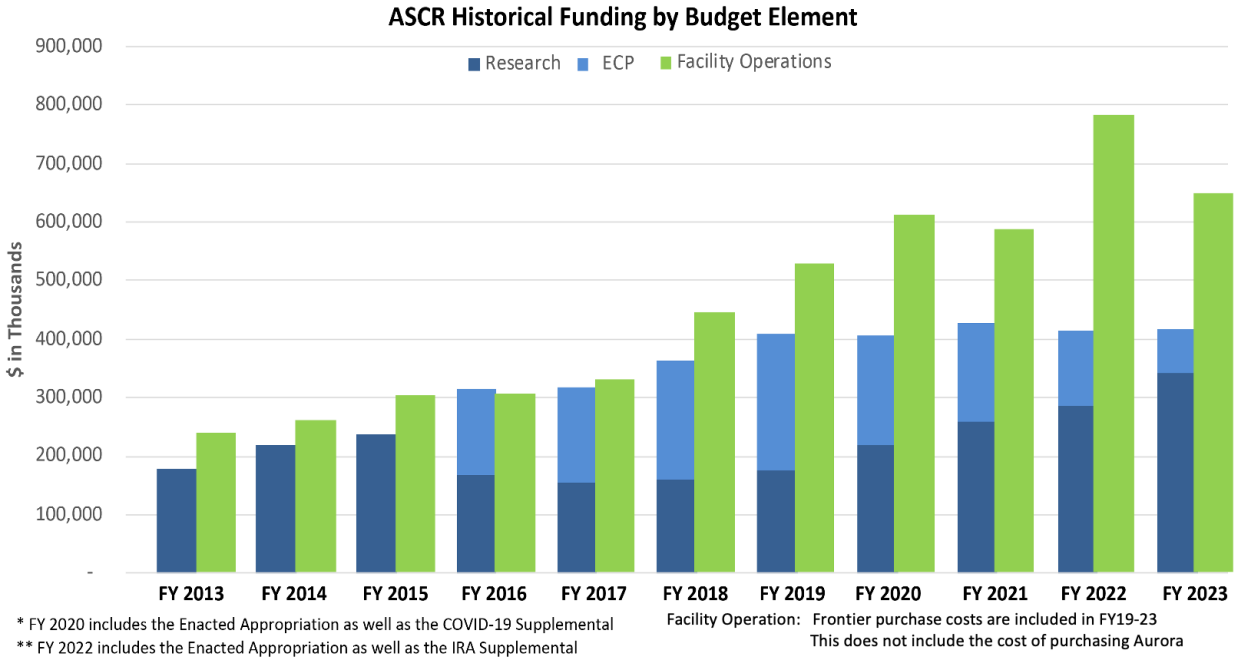


Fig. 4: ASCR funding between 2013 and 2023 (in thousands of dollars)

In 2016, DOE launched the Exascale Computing Initiative (ECI). ECI aims to be a broad-based research and development initiative that encompasses a range of activities aimed at advancing the state-of-the-art in high-performance computing. The initiative included efforts to develop new hardware and software technologies, explore new programming models, and improve energy efficiency. It incorporated the latest mathematical and computational innovations into scientific

applications, creating a comprehensive exascale software stack and advancing the capabilities at the leadership-class computing facilities to enable future scientific breakthroughs.

The Exascale Computing Program (ECP), which is part of ECI, is a focused program aimed specifically at developing exascale computing systems by the early 2020s. The program is led by DOE and involves collaborations with industry, academia, and national laboratories. The ECP focuses on addressing key technical challenges in the development of exascale systems, such as hardware design, software development, and application optimization.

Since 2016 (and up to its expected wind-down in 2023), the ECP program averaged 43% of the total ASCR research funding (and in the three years 2017-2019, ECP received over half of ASCR's research budget).

The ECP is the latest DOE program that has followed, to a significant extent, the Scientific Discovery Through Advanced Computing (SciDAC) model of supporting applications and enabling technology groups to migrate *together* to the early exascale systems. The program is widely regarded as successfully preparing teams for the exascale systems now coming online at the leadership computing facilities. As it approaches a funding ramp-down that is already delaying new hiring, it is important to maintain the global competitiveness of the enabling technologies that have been created.

Findings

1) The United States retains its leadership in HPC and AI at this time, but the gap has closed substantially, and leadership will likely be lost without further actions. Historically, ASCR has been the leader in this field – not only in research but also in deploying, operating, and applying large-scale computing systems to science. This leadership is under threat, not just because of increased competition internationally, but also because of domestic competition for talent from hyperscale operators, who are making significantly larger investments than those of DOE and the U.S. government.

2) The Exascale Computing Initiative is an exemplar of U.S. leadership in high-performance computing, incorporating the latest mathematical and computational innovations into scientific applications, creating a comprehensive exascale software stack, and advancing the capabilities at the leadership-class computing facilities to enable future scientific breakthroughs.

3) The hyperscale information technology (IT) companies require similar skills and expertise from their employees as the DOE national laboratories, but industry's focus is not on scientific or national security challenges. Moreover, both corporate and laboratory R&D in this space has typically been reliant on a large contingent of foreign-born workers. Recent changes in geopolitics have made this a questionable strategy from both a security and a competitive perspective. In addition, past investments from these companies have helped jumpstart the capacity of the foreign powers who now threaten U.S. leadership.

3. Building and Maintaining Strategic Industry and International Partnerships

Charge question:

How can the Department maintain critical international cooperation in an increasingly competitive environment for both talent and resources? In areas where the U.S. is leading, how can we sustain our roles and attract the best industry and international partners? In other areas, how can the Department build and maintain its reputation as a "partner of choice"? In general, are there barriers that can hinder our ability to form effective and enduring international and industry partnerships?

In the current geopolitical and economic climate, it is critical for the United States to form and maintain partnerships with strategic international and industrial partners. As we illustrate below, ASCR has a history of successful collaborations to build on.

Successful Collaborations in Standards and Software

Research has always benefited from the open exchange of ideas and the opportunity to build on the achievements of others. We see examples of fruitful exchanges of personnel, ideas, software, and technologies at a global scale. Technological and scientific progress is often made when researchers can build solutions on top of software with stable interfaces and well-defined functionalities, which are built in global collaborations of volunteers. Among such successful community-driven standardization efforts are Message Passing Interface (MPI) [11], which supports message passing within HPC applications; OpenMP [12], which supports shared-memory programming; and, more recently, LLVM [13], a toolkit for the construction of highly optimized compilers, optimizers, and runtime environments. These libraries and toolkits have enabled the development of countless science applications and system software.

Several tools from international sources are being used by DOE applications. U.S. researchers, specifically DOE scientists, have had historically strong collaborations with Europe and Japan in science and research. Many software products developed abroad have had an impact on DOE science. Here we list some examples of such open-source software in use by DOE scientists:

- The Vienna Ab initio Simulation Package (VASP), developed in Austria with contributions from DOE, is used for atomic-scale materials modeling [14].
- The Cactus computational toolkit is a problem-solving environment that can be customized to support several scientific communities, including gravitational wave physics. Cactus is a joint development of the Max Planck Institute for Gravitational physics (the Albert Einstein Institute), the National Center for Supercomputing Applications (NCSA) at the University of Illinois, and Louisiana State University [15].
- GROMACS is a molecular dynamics package widely used for simulations of proteins, lipids, and nucleic acids [16]. GROMACS was developed primarily in the Netherlands, then Sweden, with contributions from around the world.

- Geant4. a toolkit developed as part of global collaboration, is used in high-energy and nuclear physics to simulate the passage of particles through matter [17] and is used by U.S. and DOE scientists.
- WARP-X code for beam plasma simulation at exascale [18] was developed collaboratively between DOE scientists and the French Alternative Energies and Atomic Energy Commission (CEA).
- SeisSol is a software package for simulating wave propagation and dynamic rupture based on the arbitrary high-order accurate derivative discontinuous Galerkin method (ADER-DG). Developed at LMU Munich and Technical University of Munich (TUM), Germany, it is used by scientists to simulate earthquake dynamics as a key component in physics-based approaches to strong motion prediction for seismic hazard assessment and in physically constrained inversion approaches to earthquake source imaging from seismological and geodetic observations.
- Vampir is a software performance visualizer focused on highly parallel applications. Originally developed at the Technical University of Dresden, Germany, it presents a unified view on an application run including the use of programming paradigms such as MPI, OpenMP, PThreads, Compute Unified Device Architecture (CUDA), OpenCL, and OpenACC. It also incorporates file input/output (I/O), hardware performance counters, and other performance data sources.
- Dristhi, a visualization software package developed at National Computing Infrastructure (NCI) of the Australian National University (ANU), Canberra, Australia, is used in centers in the United States.
- Spack is an open-source software package manager developed mainly by Lawrence Livermore National Laboratory (LLNL) as a part of ECP, and much of the software package scripts for Arm processors were contributed by RIKEN Center for Computational Science (R-CCS) as a part of the Fugaku project. Spack has a growing community internationally and will be key to controlling the software complexity across different architectures for supercomputers worldwide as well as other large infrastructures such as clouds.
- VeloC is now probably the most widely used checkpoint-restart system for fault tolerance for large systems. Currently being developed at Argonne National Laboratory (ANL), VeloC has its origin both in Fault Tolerant Interface (FTI), which originated in Japan in 2011 and was further developed for production-level use in European centers, and in the Scalable Checkpoint/ Restart (SCR) framework, which originated from and was developed at Berkeley. It is truly a product of years of international collaboration.
- LAPACK and ScaLAPACK are software libraries for mathematical subroutines for solving problems in linear algebra, including matrix factorizations, linear systems, eigenvalue problems, and singular value decomposition. Both packages were developed by an international collaboration of researchers from the University of Tennessee, the University of California, Berkeley, and the Numerical Algorithms Group in the UK. The software is widely used in both academia and industry.

In addition, exchanges of benchmarks for procurement and evaluation and even concurrent development of future exascale systems are performed in collaboration with centers such as ANL and LBNL and European exascale centers, as well as RIKEN R-CCS in Japan through the DOE-

Ministry of Education, Culture, Sports, Science and Technology of Japan (MEXT) collaboration. This exchange concerns not only computational performance but also data storage and archives.

Another example of international activities is the Joint Laboratory for Extreme Scale Computing (JLESC), a collaborative research initiative focused on advancing high-performance computing technologies and applications. JLESC is a partnership of six major HPC research institutions: Argonne National Laboratory, the University of Illinois at Urbana-Champaign, and the University of Tennessee in the United States; the Barcelona Supercomputing Center in Spain; National Institute for Research in Digital Science and Technology (INRIA) in France; and RIKEN's Center for Computational Science in Japan. JLESC research activities, which started in 2010, include work on HPC system architecture and design, software development, algorithm optimization, and applications in areas such as climate modeling, materials science, and genomics. The initiative includes efforts to advance HPC education and training, with the goal of developing the next generation of HPC researchers and practitioners.

Such international collaborations may prove key to the rapidly evolving field of quantum computing, a field that is critical to national security. The United States is facing challenges in maintaining its leadership position in this area. While maintaining leadership may be difficult, if not impossible, the United States can remain competitive by leveraging the expertise and resources of key strategic partners. Building on existing collaborations between U.S. DOE labs and European and Japanese initiatives, such as the Munich Quantum Valley and RIKEN's Quantum Computing Center, the United States can position itself as a leader in quantum computing and ensure that it remains at the forefront of technological advancements in this crucial field. By fostering these closer collaborations, the United States can pool its resources, share knowledge, and jointly develop new technologies, enabling it to maintain its strategic advantage in quantum computing.

Industry Collaborations

DOE has a long tradition of collaborating with industry on leadership-class systems. Collaborations include system design, integration, and deployment. Additionally, large-scale programs such as ECP seek the advice of industry and other national agencies while conducting their activities [19]. The ECP Industry and Agency Council includes users and developers of HPC software and applications.

- In the past several years, roughly 45% to 50% of the TOP500 systems (number of machines) have been installed in industry, compared with about 30% in 1993.
- In industry, CS&E provides a competitive edge by transforming business and engineering practices. CS&E researchers are at the forefront of developing new technologies, such as artificial intelligence, machine learning, and blockchain. These technologies have the potential to disrupt entire industries and create new business models.
- In the area of software development, codes developed within DOE, such as CHEMKIN [20] for chemically reacting flows, were commercialized by Reaction Design and later by ANSYS.

Closer collaboration with industry may bring more innovation to the U.S. computing landscape. As leadership-class and edge systems become more complex and heterogeneous, not just in their architecture, but also in their usage, opportunities arise to involve industry at all levels of co-design, from hardware to the system stack, to develop more interoperable, scalable, and robust capabilities. Of critical importance are extensive and standard sets of interfaces across the system software stack that can create a reusable software ecosystem.

Furthermore, as the competition for talent continues to increase globally, pooling resources across borders and providing unique and exciting opportunities for research and growth may be beneficial. More discussion on talent is in Section 5.

United States' Collaboration in the Global Context

We note that although there is much international collaboration in the United States with external partners, this collaboration is not at the level seen in other countries. Figure 5 shows the shares of internationally co-authored publications in recent years by selected countries. The United States is in sixth place behind the UK, France, Canada, Germany, and Italy.

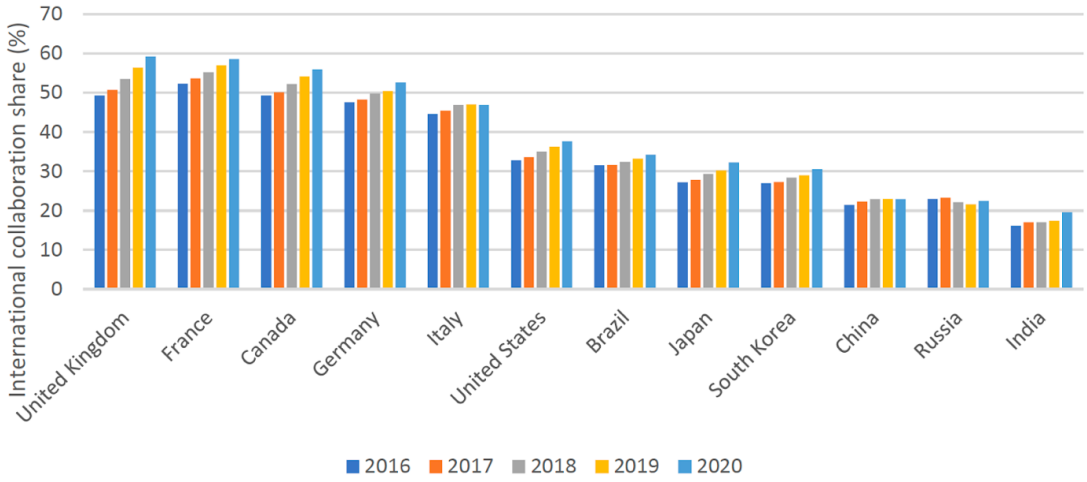


Fig. 5: Annual shares of internationally co-authored publications from 2016 to 2020 by selected countries. Source: [21]

At the same time that we see an increase in publications from international collaborations, the same report [21] shows a worrying trend in the decline in the total share of publications along with the decline of highly cited U.S. publications (Fig. 6).

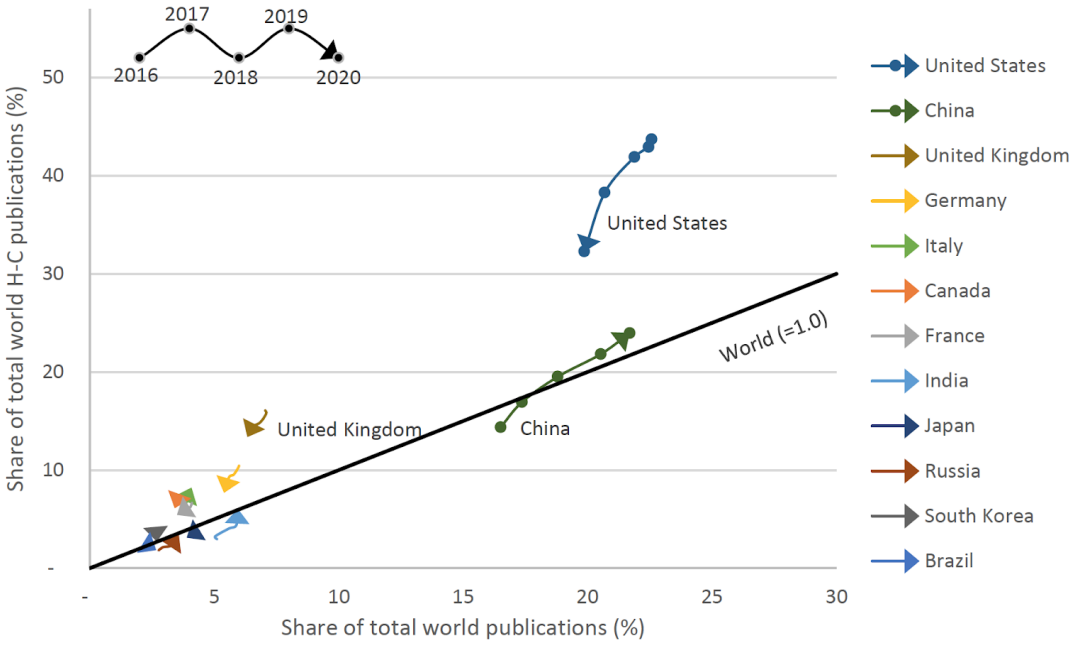


Fig. 6: "Share of world's highly cited publications versus share of world's publications … in the period 2016 to 2020. The lines represent transition from 2016 to 2020, with vertical reading representing a change in share of total world highly-cited publications and the horizontal line representing a change in share of total world publications. The arrow represents the direction of change." Source: [21].

Findings

1) High-end computational science and engineering is an area in which the United States, DOE, and ASCR have historically had a global leadership role. This leadership is now under threat from technological and geopolitical changes. Talent recruitment and retention in this area will be a challenge.

2) U.S., DOE, and ASCR leadership in key areas is under threat. This situation is due to increased international competition (e.g., it is reported that China may deploy ten exascale machines by 2025) and geopolitical changes (e.g., a less cooperative and more competitive relationship with China), as well as increased market pressures in the United States that draw talent, capital, and attention toward near-term commercial objectives.

3) In some areas, like HPC, AI, and quantum computing, it may be impossible to be the only leader globally, but it is even more important not to fall behind.

4) ASCR funding levels for research are declining in real terms (and are spread thinner across new research directions such as QIS and AI/ ML). At the same time, funding for facilities is on the increase to meet the outyear requirements of supporting the exascale platforms deployed in ECI.

5) In the current resource-constrained climate, big science and advanced scientific computing increasingly require international collaborations and bring benefit to the many parties involved in these efforts. In light of the relatively small community of supercomputing researchers, international collaborations are particularly beneficial.

6) DOE has a history of working closely with industry partners to develop, deploy, and apply advanced technology, particularly in leadership-class computer systems. DOE laboratories work closely with end users from industry and have achieved numerous high-impact results that extend the capabilities of participating companies.

7) Based on experiences in ECP, there are opportunities to improve partnerships with industry, particularly in the area of co-design, taking into account lessons learned from ECP and other international HPC efforts.

Recommendations

1) ASCR needs to articulate a vision, associated goals, and milestones for international collaboration focused on post-exascale computing, establish trust relationships with strategic partners, evangelize and socialize these efforts, define structures, and provide resources that support flexible multiparty collaborations.

2) ASCR should develop a more intentional and organized program of international collaborations. The collaborations should include both technical and social planes, advancing basic and applied research as well as enhancing the talent pool. International collaborations can result in foreign talent coming to the United States for short-term exchanges or long-term appointments. These types of opportunities exist today but are ad hoc.

3) International collaboration needs support structures. MOUs, and in some cases more formal agreements, and funding are needed to sustain existing and foster new collaborations. To obtain the necessary resources, it is important to advocate for these efforts among policymakers.

4) ASCR needs to collaborate with strategic partners to create its own research post-exascale roadmap. This is a necessary step toward the U.S. retaining its leadership position in high-end computing into the 2030s. Japan has already officially initiated its national feasibility study toward post-exascale research infrastructures (FugakuNEXT) and, in Europe, EuroHPC is conducting similar activities.

5) ASCR should work with industry on technical and social planes. For example, opportunities exist in the area of co-design. It would be beneficial to examine collaborations with industry within the sunsetting of the ECP and derive the lessons learned to inform future partnerships with industry. Additionally, ASCR should work with industry to formulate creative ways to develop talent that can bridge academia and industry.

6) Intellectual property policies should be developed to clarify how discoveries and resources are managed in international and industrial collaborations, taking into account the need for open science while simultaneously protecting sensitive information and technologies.

7) To remain the partner of choice for other countries and for industry, ASCR must have a leadership role in a number of key areas such as high-end modeling and simulation for science and engineering, artificial intelligence for science and engineering, leading-edge computing architectures and systems on the path beyond exascale, and advanced networks and future internet architectures for an integrated research infrastructure.

4. Critical Scientific Areas for Leadership in ASCR

Charge Question:

Identify key areas where the U.S. currently has, or could aspire to, leadership roles in advanced computing and high-end computational science and engineering, including unique or world-leading capabilities (i.e., advanced scientific facilities, testbeds, and networks) or leading scientific and technical resources, such as highly trained personnel and supporting infrastructure. These may include emerging areas or opportunities that offer significant promise for leadership.

Previously, we identified four major CS&E areas where ASCR has historically been a leader: applied mathematics, computer science, supercomputing, and advanced networking. In Section 2, we laid out the landscape of geopolitical and technical changes that impact ASCR's leadership in these areas. In addition to the traditional areas of leadership, ASCR can play a leading role in important emerging fields such as AI for science, energy, and security; post-Moore microelectronics (together with Basic Energy Sciences (BES) and HEP, ASCR leads the modeling, architecture, and computer science aspects); and quantum information science.

In this section we discuss four critical areas on which ASCR should focus to maintain and extend its leadership:

- High-end computational science and engineering;
- Artificial intelligence for science and engineering;
- Leading-edge computing architectures and systems on the path beyond exascale;
- Advanced networks and future internet architectures for an integrated research infrastructure.

High-End Computational Science and Engineering

DOE and the United States hold a pre-eminent place in the high-end computational science and engineering applications developed using high-end computing systems. While the full scope of this work extends beyond the ASCR portfolio, ASCR and its leadership-class HPC systems play a key role in maintaining its leadership in high-end computational science and engineering. Critical to that success is ASCR's research in applied mathematics and computer science.

Applied Mathematics

The need for fundamental algorithmic advances is particularly apparent today in scientific machine learning, where the training of neural networks with up to trillions of parameters on huge data sets stresses energy budgets as much as it does memory and processing limits. While commodity-scale machine learning proceeds routinely in single graphics processing units (GPUs), algorithmic advances in machine learning are required to address issues of scale, data privacy, and the efficient exploitation of the computing continuum, minimizing expensive and slow data traffic. Scale requires that the training data or the network parameters, or both, be distributed. DOE *must* lead internationally in this burgeoning mode of scientific discovery and engineering design.

Data privacy puts a premium on leaving the training data in place with the owner, where it can be accessed by federated learning algorithms without first being centrally gathered. The computing continuum, with data coming from and control being exercised at the edge, likewise puts a premium on distributed approaches to machine learning and inference.

Whereas the vast majority of training is done with algorithms possessing only first-order convergence, essentially derivatives of stochastic gradient descent (SGD), the slow asymptotic convergence of SGD invites the importation of second-order methods from traditional optimization and root finding in simulation-based computation, where ASCR has expertise. Because of the massive scale of training applications, Hessians and gradients require compression by means of hierarchically low-rank methods. Optimization of the network will benefit from nonlinear preconditioning because of the highly irregular loss function landscape. Both of these ideas have been well developed for simulation over the past two decades. Furthermore, communication in distributed training often requires data compression, which can be lossy.

In an era in which digital data doubles approximately every year, annually adding an amount comparable to the sum of all digital data ever stored previously, data compression is another domain in which it is essential that DOE own a competitive position. I/O is the most stringent bottleneck in some DOE applications, such as snapshotted weather and climate simulations, and processing capacity increases faster than new disk capacity. Numerous different means exist for compressing data arising from images, continuous fields, or symbolic streams – binary, algebraic, functional transforms, etc. – with various knobs for tuning loss to requirements. This is another domain ripe for near-term, high-impact mathematical inventiveness.

Indeed, the many synergies anticipated from the convergence of traditional cyberinfrastructure (CI) with artificial intelligence – both “CI for AI” and “AI for CI” – will require close integration of applied mathematicians with computer scientists and experts in the applications. The DOE can be highly competitive in particular because, through programs like SciDAC, DOE has created a culture where such synergism is solicited and rewarded.

Computer Science

Within the United States, DOE is the largest funder of high-performance computing research, with the DOE labs and DOE-funded university researchers often in leadership roles at major conferences such as the International Conference for High Performance Computing, Networking, Storage, and Analysis (SC). Much of this research addresses the software systems for HPC environments, including programming models, compilers, runtime systems, performance modeling, automatic performance tuning, scientific libraries, and operating systems.

For the past 30 years, HPC performance has improved on average by three orders of magnitude every decade. However, if we project forward from the past eight years of performance improvements, that rate of improvement has dropped significantly to only one order of magnitude every decade. Thus the scale of a system can no longer be a primary measure for the success of a facilities program; instead, measures focused on delivering more *effective* HPC to scientific users will need to be developed. Improving the synergy between pathfinding research and facilities for scale-up and deployment is essential to overcome these challenges.

ASCR has a significant research effort in data-intensive science, where the data come from simulations or from experimental and observational facilities. Topics addressed include algorithms and software for scientific data analysis and visualization, methods for ultra-high-speed data movement, and tools for managing complex scientific workflow and pipelines. Methods and tools for data cleaning, integrity, provenance, compression, and indexing have also been critical. A frequent theme in this work is problems of scaling, including in dataset size, the speed with which data are produced, and the parallelism on the machines where the techniques run.

Rapid advances in instrumentation and the emergence of new AI methods raise new opportunities and challenges for ASCR and the labs. New edge computing, data filtering, analysis, storage, and sharing methods are needed to cope with data rates that are increasing by many orders of magnitude. 5G, and soon 6G networks, and potentially also free-space optics, that provide ubiquitous connectivity both for wireless laboratories and in field experiments will pose new challenges. The automated operation of experimental apparatus in so-called self-driving laboratories is a related area in which rapid changes are expected. ASCR will need to work hard to keep up with, and ideally lead, advances in these areas in order to preserve U.S. leadership in many areas of scientific research.

DOE, through its National Quantum Information Science Research Centers [22] is investing in research and development of quantum computing technologies, which have the potential to revolutionize computing and solve problems that are currently intractable using classical computing methods. The investments focus on building quantum computers, quantum networks, quantum sensing technologies, and using quantum science for the discovery and design of new materials with unprecedented properties.

Findings

1) Science and engineering applications of national importance will continue to require increasingly more capable advanced computing systems to model complex phenomena,

process and manage vast amounts of data, and support cutting-edge experiments. Meeting those requirements requires major and sustained advances in applied mathematics and computer sciences.

2) DOE has a breadth of expertise in applied mathematics including the development of new numerical methods, the creation of mathematical models for physical systems, the development of algorithms for large-scale data analysis and visualization, and the creation of tools for high-performance computing.

3) Fundamental advances in HPC require innovation in systems software architecture, programming models, compilation techniques, and application development methods. Operating systems, programming libraries, and applications must be updated, reverse-engineered, or rewritten from scratch to incorporate these advances.

Recommendations

1) ASCR needs to increase its investment in basic and applied mathematics, computational science, and engineering focused not only on traditional HPC applications, but also on new applications of AI and machine learning technologies. This investment should be used to build better connections between DOE laboratories and academia by fostering collaborations and spurring innovation.

2) To facilitate a synergistic relationship between the research community and both the leadership computing and the experimental DOE laboratory facilities, ASCR should create a substantial and sustained broad-based research program in computer science and software engineering. This program should be focused around a clear HPC and AI technology roadmap and involve strong national and international collaboration. Such a program will be essential for maintaining U.S. leadership in the effective use of post-exascale systems for computational science and engineering. The roadmap should include quantum information science in the list of topics potentially relevant to high-end computational engineering and science.

Artificial Intelligence for Science and Engineering

Deep learning (DL) neural networks are now a key technology for the IT industry and are used for a wide variety of commercially important applications such as image classification, facial recognition, handwriting transcription, machine translation, speech recognition, text-to-speech conversion, autonomous driving, and targeted advertising. More recently, Google's UK subsidiary DeepMind has used DL neural networks to develop the world's best Go playing systems with their AlphaGo variants [23]. However, of particular interest for DOE's "AI for Science" agenda is DeepMind's AlphaFold protein-folding prediction system [24]. The latest version of AlphaFold convincingly won the most recent Critical Assessment of Protein Structure Prediction (CASP) protein-folding [25]. As Nobel Prize winner Venki Ramakrishnan, has said [26]: "This computational work represents a stunning advance on the protein folding problem, a 50-year-old

grand challenge in biology. It has occurred decades before many people in the field would have predicted. It will be exciting to see the many ways in which it will fundamentally change biological research."

Huge quantities of experimental data now come from many sources, including satellites, gene sequencers, powerful telescopes, X-ray synchrotrons, neutron sources, and electron microscopes. These sources already generate many petabytes of data per year and planned upgrades of these facilities will create at least an order of magnitude more data. Extracting meaningful scientific insights from these ever-increasing mountains of data will be a major challenge for scientists. The premise of AI for Science is that such "big scientific data" represents an exciting opportunity for the application of new AI technologies in ways that could be truly transformative for many areas of science.

In 2019, the DOE laboratories organized a series of townhall meetings, attended by hundreds of scientists, computer scientists, and participants from industry, academia, and government, to examine the opportunities for AI to accelerate and potentially transform the scientific research fields under the domain of the DOE's Office of Science [16]. The stated goal of this endeavor was as follows: "To examine scientific opportunities in the areas of artificial intelligence (AI), Big Data, and high-performance computing (HPC) in the next decade, and to capture the big ideas, grand challenges, and next steps to realizing these opportunities."

The townhall meetings used the term "AI for Science" to broadly represent the next generation of methods and scientific opportunities in computing and data analysis. This includes the development and application of AI methods, for example, machine learning, deep learning, statistical methods, data analytics, and automated control, to build models from data and to use these models alone or in conjunction with simulation data to advance scientific research. The meetings concluded that the use of AI methods in science has the potential to transform many areas of scientific research over the next decade:

- Accelerate the design, discovery, and evaluation of new materials;
- Advance the development of new hardware and software systems, instruments, and simulation data streams;
- Identify new science and theories enabled by high-bandwidth instrument data streams;
- Improve experiments by inserting inference capabilities in control and analysis loops;
- Enable the design, evaluation, autonomous operation, and optimization of complex systems from light sources and accelerators to instrumented detectors and HPC data centers;
- Advance the development of self-driving laboratories and scientific workflows;
- Dramatically increase the capabilities of exascale and future supercomputers by capitalizing on AI surrogate models;
- Automate the large-scale creation of FAIR – Findable, Accessible, Interoperable, and Reusable – data.

In 2022, ASCR organized a new set of workshops on AI for Science, Energy, and Security (AI4SES). These were again well attended and a report is in press. From the range of applications of AI technologies and the enthusiasm of the participants it seems clear that the Office of Science

should consider funding a major initiative in such AI applications [27]. Such an initiative could capture the imagination of staff at the DOE laboratories and assist in recruiting well-qualified staff as a follow-on to the ECP.

Findings

The following observations are distilled from the Artificial Intelligence Index Report 2022 [28]. This report contains more data and details of global AI trends.

1) Big data and HPC are both important to scientific discovery and are synergistic. Experimental facilities across DOE's Office of Science are increasing the demands for leading-edge computing facilities, methods, and services. These demands include the ability to move, analyze, share, and manage exponentially growing datasets from observational sensors and increasingly powerful scientific instruments and to integrate that data with physics-based and data-driven models, which may themselves produce enormous datasets and require massive computing for model training and inference.

2) Large language models (LLMs), such as Chat-GPT developed by OpenAI, are an exciting recent development. However, the efficacy, impact, and safety implications of LLMs for ASCR science are still uncertain, in part because dramatic progress in these models has been quite recent and yet to be fully understood. LLMs may hold the potential to accelerate scientific workflows (by automating the process of assimilating and summarizing technical material, generating code, and computational pipelines, for example), but many issues around replicability, open access, and safety remain to be tackled. The National AI Research Resource (NAIRR) Task Force recently published a report describing how NAIRR [29] seeks to give academic researchers access to competitive AI resources to enable the democratization of such technology.

3) AI applications have become significantly more affordable and higher performing with lower training costs and faster training times.

Recommendations

1) DOE urgently needs to make a significant long-term investment in R&D in AI technologies and their application to the huge datasets now generated by the experimental facilities at the national laboratories. As a follow-on to the Exascale Computing Project a major DOE initiative in AI for Science, Energy, and Security could place DOE as the world leader in that space.

2) DOE should assist in increasing U.S. competitiveness by democratizing access to the leadership-class computational resources needed for AI R&D. This could include working

with other federal agencies to make the NAIRR platform a reality by providing increased access to high-performance and AI computing infrastructure at its national laboratories.

3) New advanced computing capabilities to be developed by ASCR should include the needs of AI research in the requirements and include the co-design of networks and computational hardware, as directed by the National AI Initiative Act of 2020 [30].

Leading-Edge Computing Architectures

The situation in large-scale computing architectures, for both HPC and AI, is well summarized in the paper "HPC Forecast: Cloudy and Uncertain" by Reed, Gannon, and Dongarra [31]. The authors discuss the implications of both the end of Dennard scaling and the increasing influence of the cloud hyperscaler companies over the traditional HPC hardware and software vendors. Although we use the term "HPC" here, increasingly we see architectures for high-end HPC and AI systems as being similar and potentially converging. We, therefore, consider these two areas to be linked when considering leadership in "leading-edge computing architecture." We also note that the use of AI in traditional scientific computing is on the rise and will become more tightly integrated into scientific codes, but in this section, we limit the discussion to computer architectures. HPC architectures are at an important inflection point, being reshaped by a combination of technical challenges and market ecosystem shifts.

In the near term, conventional architectures will likely continue to dominate but with more use of custom silicon solutions for particular classes of problems. In the longer term, however, the possibility of quantum, neuromorphic, DNA computing, or more unconventional architectures playing a larger role must be considered. Some of these technologies may be more useful in the medium term as domain-specific accelerators, rather than a general-purpose system.

Building the next generation of leading-edge HPC systems will require rethinking many fundamentals and historical approaches by embracing end-to-end co-design; custom hardware configurations and packaging; large-scale prototyping, as was common 30 years ago; and collaborative partnerships with the dominant computing ecosystem companies.

Looking forward, it seems increas-ingly unlikely that future high-end HPC systems will be procured and assembled solely by commercial in-tegrators from only commodity com-ponents. Rather, future advances will require embracing end-to-end design, testing, evaluating advanced prototypes, and partnering strategi-cally with not only traditional chip and HPC vendors but with the new cloud ecosystem vendors. These are likely to involve collaborative part-nerships among academia, govern-ment laboratories, chip vendors, and cloud providers; increasingly bespoke systems designed and built collabora-tively to support key scientific and en-gineering workload needs; or a com-bination of these two.

Findings

1) Much of the innovation in computer architecture will likely be developed by, or for, the hyperscale/cloud vendors. In addition, increased use of custom and semi-custom silicon is likely with the introduction of chiplet technology. This will allow combinations of processing cores with multiple special-purpose accelerators on a single die. HPC and AI architecture research funded by ASCR will need to adapt to these trends.

2) The end of Dennard scaling has opened the door for new competition in the processor space (e.g., AI accelerators and proliferation of ARM variants). However, an aging software base makes agile adoption of new technologies difficult within the scientific computing enterprise. Changing computer architectures always requires a concomitant investment in algorithms and system software. Adapting to this new technology ecosystem with the development and adoption of open standards will promote continued progress.

Recommendations

1) DOE needs to invest in long-term forward-looking research in advanced computer architecture and system concepts to identify potential solutions for sustaining continued scientific productivity increases for future scientific computing systems.

2) DOE should enhance its technological leadership by investing in state-of-the-art technologies for the electronic design and reimagination of future leadership-class systems. This will require a parallel investment in interoperable system software and programming systems that can enhance productivity by leveraging the capabilities of heterogeneous computational, data, and network resources.

3) Given semiconductor constraints, substantially increased system performance will require end-to-end co-design from device physics to applications. In addition to developing partnerships with hardware vendors and cloud ecosystem operators, such a co-design effort will require substantially increased government investment in basic research and development. To escape the present HPC monoculture and build systems better suited to current and emerging scientific workloads at the leading edge, DOE should fund the building of hardware and software prototypes at a scale that can test such new ideas using custom silicon with chiplet technology and the development of associated software.

4) New collaborative models of partnership and funding are needed that recognize these ecosystem changes and their implications, both in the use of cloud services where appropriate and in the collaborative development of new system architectures. Such architectures will need to take into account the increased importance of energy consumption in large-scale systems, from both a cost and an environmental perspective.

5) ASCR should consider the timeframe in which neuromorphic, quantum, or other unconventional architectures may become relevant, either as a domain-specific accelerator or as a replacement technology.

6) Building large-scale prototypes, whether in industry, national laboratories, or academia, depends on recruiting and sustaining integrated research teams consisting of chip designers, packaging engineers, system software developers, programming environment developers, mathematicians, and application domain experts. This will require coordinated funding, not only for such workforce recruitment but also for the basic research and applied R&D needed to develop and test prototype systems. DOE should create attractive career paths with exposure to cutting-edge research technologies at the national laboratories.

Advanced Networking and Future Internet Architectures

The DOE Energy Sciences Network, ESnet, is recognized as a global leader within the ecosystem of National Research and Education Networks (NRENs) by virtue of its advanced production capabilities, research initiatives and testbeds, and science engagement programs [32]. Although poorly understood outside of the NREN space, this global ecosystem is vital for the productivity of the Office of Science mission. ESnet provides an interoperable network substrate for data acquisition and placement; a diverse array of testbeds for architecture and applications research; a pool of early technology adopters; and some degree of market power, especially for the acquisition of advanced optical technologies.

ESnet's recent upgrade provides transformative capabilities including its immense network capacity, available on a just-in-time basis, as well as features for programmability and orchestration. In addition, it provides improved resilience against failure as well as against cyberattacks.

We also observe that the computing continuum is making new demands on networks. New sensors that generate data at ultra-high rates (up to terabit/s) need new approaches for linking instruments with computing, both within and across DOE laboratories. This has implications for ASCR research programs and for its computing and networking facilities.

We see an increase in collaborative and distributed scientific applications that require new services to enable reliable, high-performance, and secure workflows that span many resources and institutions. Developing and applying these services will require a new level of collaboration between application scientists, computer scientists, and network architects. For example, emerging science workflows in Earth systems and other fields will require dense monitoring and measurement via a new generation of 5G (and beyond) connected instruments, combined with flexible edge computing resources to support AI, data reduction, and other functions.

Finally, scientific drivers for quantum networking include the interconnection of quantum computers that are physically separated and the integration of quantum sensors (i.e., sensors that use quantum phenomena to measure signals) into experimental architectures.

Findings

1) DOE's network infrastructure needs to evolve to accommodate processing within the computing continuum (from the edge to the cloud or HPC center) as well as support a more collaborative mode of work and distributed science workflows.

2) Emerging scientific workflows in Earth systems and other fields will require dense monitoring and measurement via a new generation of 5G connected instruments, combined with flexible edge computing resources to support AI data analytics and data reduction. Given the loss of U.S. leadership in 5G infrastructure technologies, there is a clear follow-on risk that the United States will fail to lead in the development of such 5G applications. There is also a danger that the United States will not lead in the necessary R&D for 6G network infrastructure and next-generation internet technologies.

3) Failure to be competitive in the networking space would have adverse consequences across all Office of Science (and indeed all U.S. federal) missions, including increased cybersecurity risk; increased risk that the internet will fragment along geopolitical and ideological lines; and risk that the special requirements of data-intensive science are not well-supported by commercial networking components.

Recommendations

1) Funding should be provided to support ESnet's advanced testbed services and collaborations, especially in the areas of programmable networking and 5G application development. ESnet testbeds should be made available to support R&D performed by commercial networking vendors, including startups focused on developing new 5G-connected data acquisition/ edge computing technologies.

2) Regional optical networks and global NRENs should be coordinated to advance the new, multinational declaration on "Principles for the Future of the Internet," especially with regard to the goal that "infrastructure is designed to be secure, interoperable, reliable, and sustainable" [33].

3) Funding should be provided for R&D into quantum networking investigations including the interconnection of quantum computers that are physically separated and the integration of quantum sensors (i.e., sensors that use quantum phenomena to measure signals) into experimental network architectures.

5. Cultivating and Sustaining Advanced Research Capabilities

Charge question:

To preserve and foster U.S. leadership roles within reasonable resource constraints, are there particular technical areas or capabilities that could be emphasized? Are there other technical resources and capabilities that could be leveraged in to achieve these goals, possibly through collaborations within and beyond the ASCR community?

DOE ASCR has been a leader in the design, integration, and deployment of high-performance computing systems and in supporting their use across DOE-relevant science and engineering applications and beyond. As the report discussed in Section 2, this leadership is being challenged. A recent report [34] examining the National Nuclear Security Administration (NNSA)'s Advanced Simulation and Computing (ASC) program found similar challenges and advocated for new approaches to technical and human resource management. It stated that "*The combination of increasing demands for computing with the technology and market challenges in HPC requires an intentional and thorough reevaluation of ASC's approach to algorithms, software development, system design, computing platform acquisition, and workforce development"* to support NNSA's core mission of ensuring "that the United States maintains a safe, secure, and reliable nuclear stockpile through the application of unparalleled science, technology, engineering, and manufacturing." In this section, we focus on ASCR's technical resources followed by Section 6 discussing needs and opportunities related to cultivating human talent.

Facilities

DOE ASCR's computing facilities have had a significant impact on computational science, networking, and engineering. They provide researchers with access to state-of-the-art supercomputers and other high-performance computing resources, enabling them to conduct simulations and analyses at ever-increasing scales of complexity and size.

As a result, scientists have used the computing power of ASCR's facilities to make advancements in material science in order to understand the fundamental properties of materials and predict their behavior under different conditions; in climate modeling, to improve understanding of climate change and its impacts; in energy research, where researchers simulate complex energy systems and optimize their performance to develop more efficient and cost-effective energy production; and in engineering design and analysis of complex systems in a way that optimizes their performance and reliability [35].

From the computing perspective, ASCR facilities play a unique role as a trusted partner for industry when it comes to the development of HPC hardware and software technologies and advanced networking technology. Their role ranges from providing expert advice and quantitative evidence to improved system design all the way to deep co-design partnerships with industry to create new HPC concepts. DOE is trusted because there is no concern about DOE competing against industry. DOE's capabilities in groundbreaking pathfinding research into advanced computing architectures and systems have been successfully translated into innovative systems deployed at scale by DOE facilities. Thus, advanced computing technology research has a

synergistic relationship with supercomputing and networking facilities. However, as we described in Section 2, the technological landscape is changing.

Beyond exascale, numerous opportunities exist to refill DOE's innovation pipeline with new technology directions. Ever since the "attack of the killer micros" in the 1990s, DOE's long-standing economic model for stoking the innovation pipeline for HPC has been founded on leveraging commercial off-the-shelf (COTS) technologies and innovating around them to deliver HPC systems (Fig. 7). But Moore's law is fading rapidly [36], and architectural specialization is coming back into force again to deliver continued performance improvements for industry. At the same time, as the COTS market has become dominated by the needs of the hyperscaler platforms for commercial artificial intelligence (Apple, Google, AWS, Meta, etc.), COTS technology specializations are moving farther away from HPC requirements. Hyperscaler datacenters are actively creating supply chains for scalable system components and even open/ multi-vendor chiplets ecosystems that enable agile and cost-effective specialization to serve the needs of their workloads. The cost of reversing those trends is unaffordable for the HPC community [39]. Just as with the attack of the killer micros, DOE should embrace the marketplace changes and focus on learning how to leverage the technology supply chain that is supporting the hyperscaler need for specialized systems.

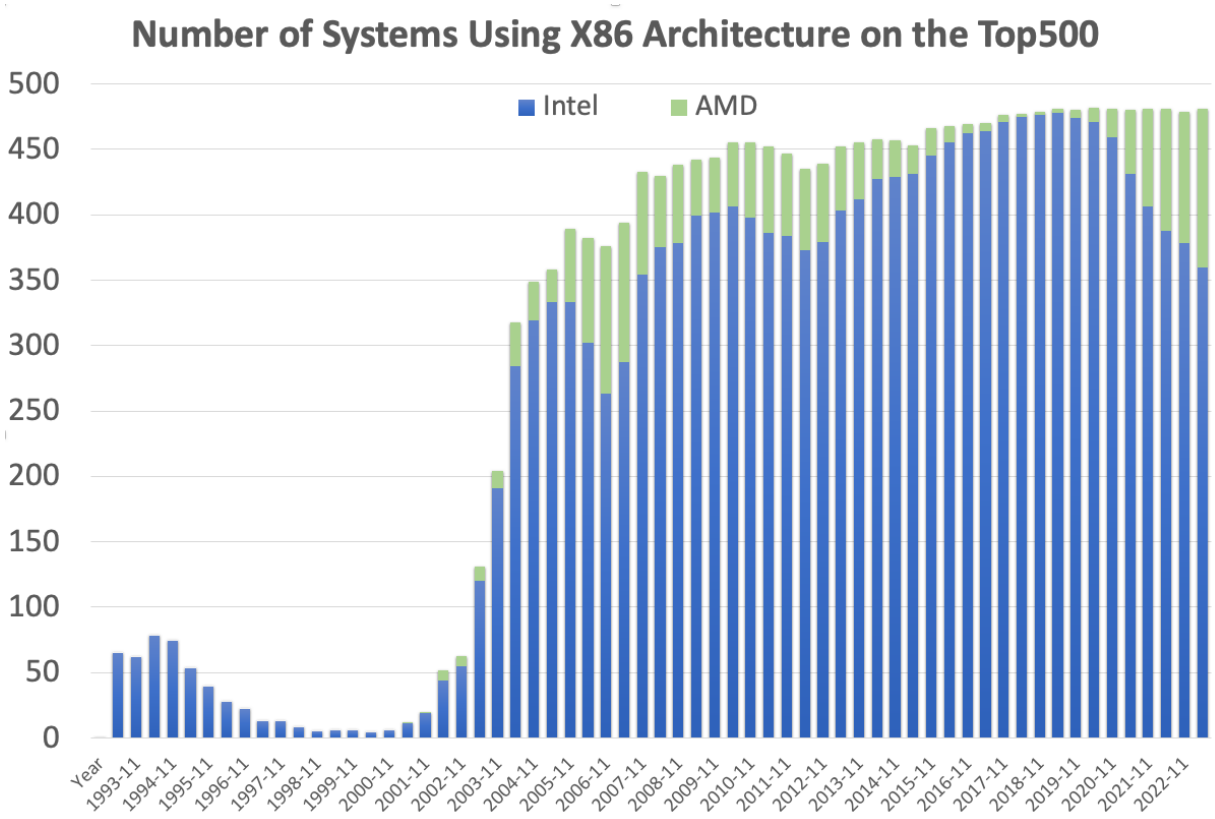


Fig. 7: In the early 2000's, attack of the killer micros led to adoption of HPC systems based on COTS microprocessor technologies supported by the broader market. The economic context in 2023 has changed dramatically where hyperscalers are dominant in the market, see Fig 8.

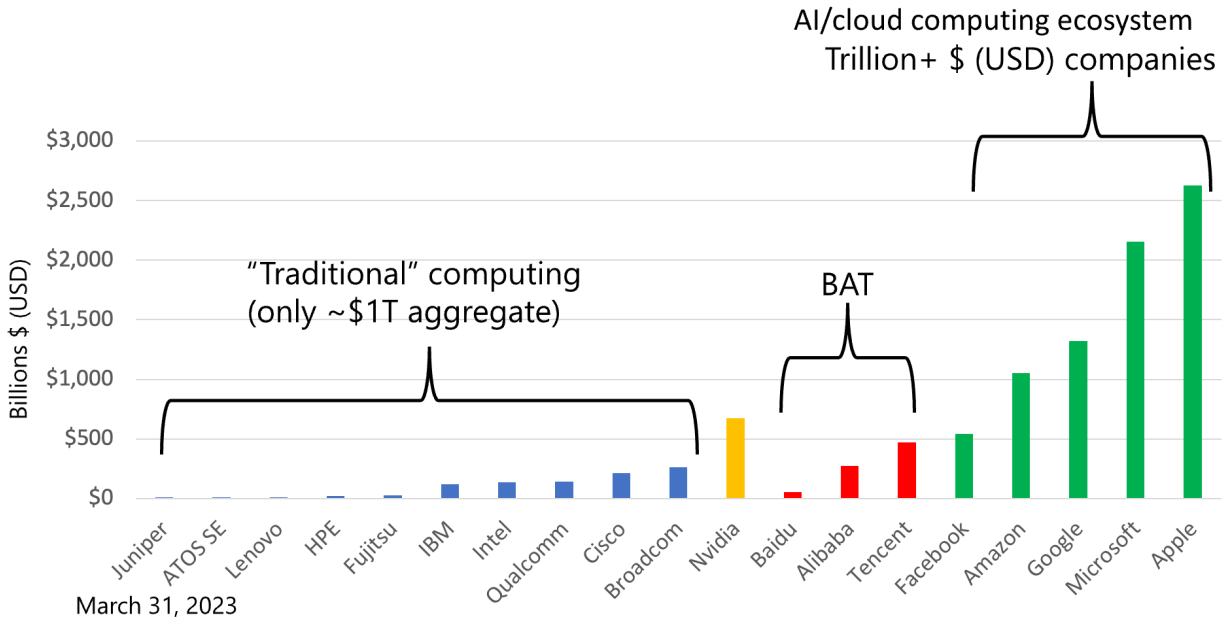


Fig. 8: The market capitalizations of the smartphone, social media, and hyperscalers now dwarf that of those companies that manufacture microprocessors, networking equipment, and other computing components [37].

DOE must develop the understanding and skills necessary to leverage the supply chains and methodologies being developed to serve this much larger market so that they can serve the needs of science. This is fundamental research and not solvable through a procurement process.

In addition to pushing digital processing beyond exascale, it is incumbent on ASCR to enable DOE to capitalize on the emerging commercial market of quantum computing. This marketplace is driven primarily by cryptography and security, but the potential scientific applications of a supercomputer with a quantum-attached processor are too substantial to be left to competitors in CS&E.

To retain and capture new competitive positions and to share the plentiful room at the top will require doubling down on the co-design processes that led to previous-generation HPC system design success. DOE will need to dive deeper into a co-design process that spans algorithms, architecture, and software. This is a grand challenge that can be addressed nowhere else as well as by the unique capabilities that DOE brings in multidisciplinary collaboration to complex systems-scale problems.

Prior experiences in designing, integrating, and deployment of HPC resources show that long-term sustained relationships spanning five years to a decade between DOE researchers and our industry partners are needed to ready emerging ground-breaking technologies and make them practical to use for HPC. Likewise, sustained relationships are needed between facilities and industry to be able to scale up those technologies. Moving to shorter-term relationships (~3-year engagements) or failure to cultivate these complementary/ symbiotic roles between industry, DOE research, and

DOE HPC facilities could dry up a productive innovation pipeline and ultimately cause DOE and the United States to lose their unmatched leadership role in HPC.

Findings

1) DOE's long history of repeatedly acquiring, deploying, and operating the world's most powerful supercomputers and networks has long benefited U.S. CS&E by ensuring access to state-of-the-art computing facilities and it has contributed to sustaining a strong U.S. advanced computing industry. However, this leadership is under threat.

2) Significant investments in technology are being made by industry that may prove relevant for ASCR's computing facilities.

3) The technology landscape has fundamentally changed:
    a) Dennard scaling ended a decade ago and the effect of Moore's law is now waning.
    b) Huge investments in computing by hyperscalers (e.g., cloud and social networking companies) are shaping the marketplace toward their specific needs.
    c) Artificial intelligence (AI)-related computation is now a major performance driver for high-end HPC systems in industry and within hyperscaler data centers.
    d) The rise of custom/semi-custom silicon (AI chips, chiplets, extensible or even open Instruction Set Architectures (ISAs), etc.) creates new possibilities to leverage commodity technologies.
    e) There is now investment in potentially disruptive technologies, such as quantum computing and networking (devices, architectures, models); however, these may take decades to mature.

4) Achieving the breakthroughs in science needed by DOE and the nation requires innovation in both hardware systems and software infrastructure. Moving the ASCR facilities forward will continue to require an interdisciplinary approach anchored in co-design, rather than a reliance solely on the vendor marketplace. ASCR will see success from continuing to encourage collaboration across science teams, computing researchers, facilities staff, and vendors.

Recommendations

1) DOE should establish a post-exascale roadmap that looks beyond our current approaches and considers alternative economic models in addition to technology solutions.

2) DOE should re-establish and strengthen the complementary relationship between the forward-looking research activities and the facilities' strengths for scale-up and deployment of viable solutions that emerge from such research. These constitute two different programs, each of which has substantially different metrics for success, hence the complementary and synergistic relationship. Forward-looking research identifies what options could be available for the future (potentially many options) and the facilities program selects from those options to deploy the best-value solution to the scientific community. They should be managed as distinct but complementary activities.

3) DOE must develop the understanding and skills necessary to leverage the supply chains and methodologies being developed by industry so that it can serve the needs of science.

6. Strategies for Success in Recruitment, Retention, and Career Advancement

Charge question:

How can programs and facilities be structured to attract and retain talented people? What are the barriers to successfully advancing careers of scientific and technical personnel in advanced computing, computational science and engineering, and related fields and how can the Department address those barriers? A complete answer to these questions should address how we can ensure that we are recruiting, training, mentoring, and retaining the best talent from all over the world, including among traditionally underrepresented groups within the U.S.

Before we explore issues of recruitment and retention in general, we must address the urgent need to capture ECP talent. In December 2023, the ECP program will be completed, and currently there is no plan to fund the researchers working on the program beyond that time frame.

It is critical that DOE ASCR and the DOE laboratories devise a strategy to retain ECP personnel and provide them with long-term career opportunities. The Exascale Computing Project (ECP) in FY22 had roughly 424 FTE's and over 1190 distinct individuals, who diligently worked at the various DOE laboratories. Additionally, there are roughly 200 professionals contributing to the project's success at universities. However, as the ECP program approaches its anticipated conclusion on December 31, 2023, there is a growing concern within the DOE ASCR program. Specifically, there is a lack of available funding to facilitate a seamless transition from the ECP to the next ambitious large-scale project.

While the ASCR program has presented a few limited funding opportunities throughout 2023, these opportunities pale in comparison to the magnitude of resources required to bridge the gap effectively between the ECP and its imminent successor. Consequently, the Department of Energy finds itself in a precarious position, as it faces the imminent risk of losing the invaluable knowledge and exceptional skills possessed by the trained staff involved in the ECP. The absence of a clear and definitive message regarding post-ECP plans only exacerbates the prevailing uncertainty and anxiety among the workforce.

The magnitude of this challenge cannot be overstated. It is imperative to establish a sustained program that goes beyond the realm of exascale computing. Without adequate support and a concrete plan for the future, the DOE risks not only losing highly trained personnel but also hindering the progress and advancement of cutting-edge computing initiatives.

To shed light on the comprehensive nature of the ECP program, the accompanying figure (see below) provides a detailed overview of its key components. These include program management, application development, software technologies, and hardware and integration. Each of these aspects plays a crucial role in the overall success of the ECP, underscoring the interdisciplinary nature and collaborative efforts required to push the boundaries of computational capabilities.

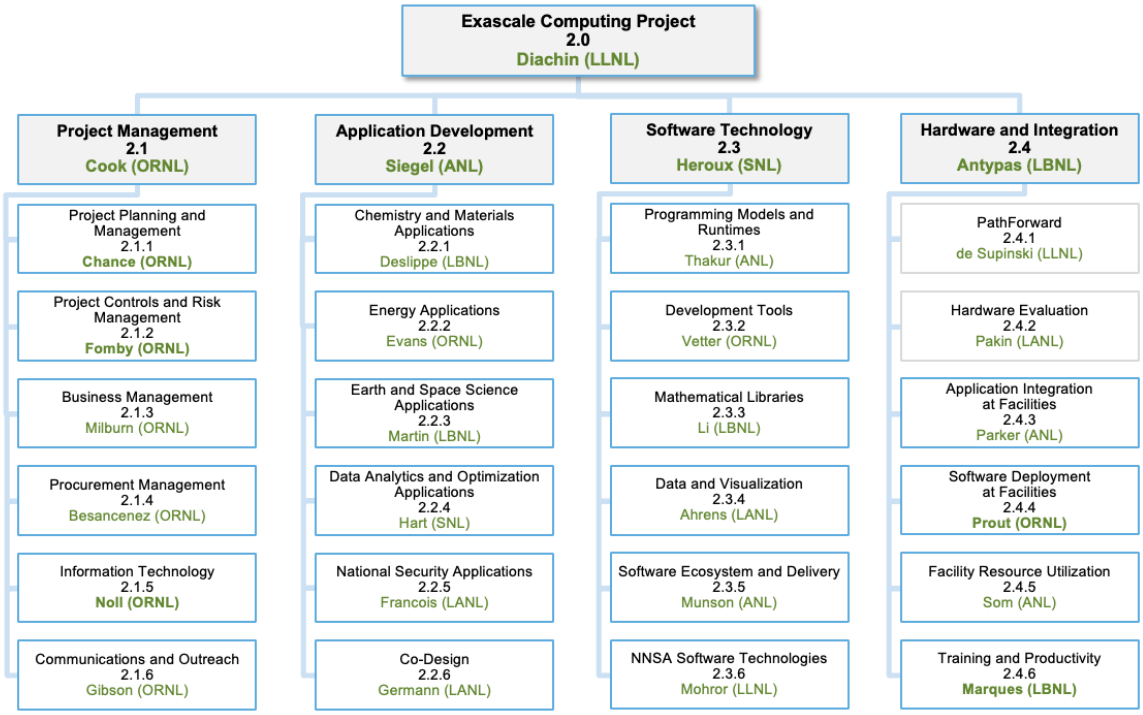


Fig. 9: Work breakdown structure of the ECP with individual responsibilities for each area indicated.

The upcoming end of the multiyear ECP in 2023 has become a major source of uncertainty and anxiety since no clear message has been communicated as to what the post-ECP plans are to cover the personnel who are currently working on ECP. Despite current financial challenges faced by many companies, the demand for computing-related talent is sufficiently high that there will be job opportunities in the industry for individuals with skills related to advanced computing who are currently present in national labs.

The danger is that this cadre of scientists, who are well-educated and well-trained, will find positions outside of DOE and the lab talent drain will be significant and irreparable. Moreover, since ECP software is now being utilized by the strategic partners of the United States, for example, the EU and Japan, the software must be maintained and improved over the coming years to retain the leadership qualities of the infrastructure. If significant talent is lost quickly, such quality could be compromised and will result in the loss of competitiveness, not just for the United States, but, holistically, for its strategic partners. ASCR needs to follow through with capitalizing on the ECP advances, but it also needs to establish a sustained program to look beyond exascale, as our peers in Japan and the EU are already doing.

Careers in National Laboratories

Funding

An urgent need exists to establish programs that will enable the United States to continue attracting and retaining top talent from across the world. One suggestion is to substantially increase the mean value of DOE's Early Career awards. Historically, the average award for a DOE National Laboratory has been $2,500,000 for five years. Another suggestion is to earmark a portion of research funding as "base funds" to enable the establishment of a stable base research environment, and only compete for funding launching from this stable base, modeled on the Bell Labs example [38]. This would give scientists the intellectual freedom and mental comfort to pursue long-term, impactful research that cannot be done within a short time horizon. It would help to ensure ASCR's ability to recruit and retain the most talented individuals from industry and academia to tackle problems in fundamental research that would have the broadest societal impact.

Flexibility

National labs have traditionally been viewed as inspiring places to work where researchers and scientists contribute to large-team projects that simultaneously advance science and contribute to vital missions for the nation. A career in a national lab often offered more autonomy, flexibility, job stability, and an overall *esprit de corps*, relative to careers in industry and academia. However, this picture has changed dramatically in recent years for personnel with advanced computing and computational science skills that are vital for ASCR. In terms of flexibility, recent events with the COVID-19 pandemic have shown that many companies offer more flexible arrangements for part-time or full-time remote work relative to national labs. One area in which national labs still often retain an advantage over industry is the independence and autonomy with respect to intellectual freedom and the technical content of the job. While industry employers make their best effort to offer job flexibility to their top talent, business priorities ultimately take precedence when evaluating which technical contributions are recognized more highly.

Compensation Gap between National Labs and Industry

The compensation gap between careers in national labs and industry in computing-related areas has increased significantly in recent years. While every attempt should be made to reduce the gap in base pay between industry and national labs (anecdotally, the base pay gap is under 50%), it is unrealistic to expect this gap to disappear.

However, many leading scientists choose to pursue a career in academia because of intellectual freedom, job security, and benefits such as sabbaticals. National labs should aspire to a similar goal with respect to recruiting and retention.

Furthermore, a national lab career can offer more attractive retirement benefits than an industry career can, especially with respect to pensions, and this tradeoff could be made more explicit in

recruiting and retention discussions. For example, each national lab employee should be made aware of the total dollar amount of savings required for someone retiring from industry to generate an annuity that is comparable to the pension that one would receive after retiring from a national lab.

Societal Relevance

A notable positive trend can be seen in many fresh graduates and early-career personnel (especially those belonging to Gen Z — those born between 1997 and 2012) in their search for meaning and societal relevance in their careers. National labs have an opportunity to present their careers as being more meaningful than industry careers, which are often viewed as strictly transactional.

In addition, early career personnel have expressed a strong preference for being in a workplace that clearly embraces diversity, equity, and inclusion. This applies to personnel belonging to all groups and especially those from underrepresented groups who benefit from clear signs that national lab careers are welcoming for them. It is notable that industry is investing heavily in continuing to make the workplace more inclusive and highlights these investments and changes as part of their recruiting messages.

Engagement with Academia

Visibility of National Lab Careers in Academia

The 2021 Taulbee Survey [39] provides information about careers pursued by recent computing-related Ph.D.s from over 170 North American institutions who participated in the survey.

The survey found that 56.3% of these Ph.D.s went on to industry careers, 10.7% obtained tenure-track positions in academia, and 11.4% pursued postdocs in academia. In contrast, only 1.6% (22 out of 1,358 computing Ph.D.s) pursued a career in government labs/ agencies[2]. Since graduating Ph.D. students also influence career choices of future Ph.D. students, national labs seem to be caught in a circle whereby only a small fraction of computing-related students have exposure to and/ or understanding of careers in national labs and there does not seem to be much momentum currently for this fraction to grow. Current postdocs in academia could be a promising target of opportunity for national lab careers because they have signaled that an industry career is not a dominant first choice for them.

Opportunities to Increase Engagement with Academia

[2] The percentage can be higher in specialties that have been traditionally tied to national labs. For example, 21 of the 1,358 computing Ph.D.s self-identified as being in the HPC speciality and four among these 21 pursued a career in a government lab or agency. On the other hand, only two of 298 Ph.D.s in AI/ML went to a government lab or agency.

Partnerships between national labs and universities have enjoyed a long history of significant mutual benefits, however, the level of academic engagement has declined relative to the past. This decline in engagement may have, in part, contributed to the reduced levels of visibility discussed in the preceding section.

Joint appointments can be one way for national labs to increase their engagement levels with academia. Joint appointments between academia and the national labs have traditionally been a win-win for all institutions involved and have helped amplify research impact in both directions. For scientists at national labs, a joint or an adjunct appointment at a university can often contribute to their job satisfaction and retention. In addition, any time that scientists spend teaching or mentoring students will help increase the visibility of national labs among the students they engage with and could lead to direct recruiting opportunities.

Moreover, joint appointments for university researchers at national labs can further increase the level of vibrancy for the national lab groups that they engage with. This is especially true when they help bring in students to contribute to the partnership, which in turn can motivate researchers at national labs to invest in mentoring and partnering with these students.

Opportunities to Increase Engagement with Industry

There is an emerging trend, currently in its infancy, of joint appointments between industry and academia. This is new territory where important topics, including intellectual property and conflicts of interest, are being actively discussed and managed by both employers for personnel participating in such joint appointments.

In light of these developments, an opportunity exists to explore the possibility of joint appointments between national labs and industry, in cases where conflicts of interest can be suitably managed. Recent technology trends (e.g., AI, post-Moore computing, quantum computing, cloud computing, edge computing) have increased the synergies between advanced computing research and development in industry and national labs. There are notable cases of open-source projects with contributors from national labs, academia, and industry (for example, MPI and LAPACK). In such cases, joint appointments might further increase the success of an open-source project because the person involved will have greater insight on win-win opportunities for national labs and industry. While the joint appointment path would likely be feasible for only a small number of personnel, it could help retain key personnel engaged in critical projects for ASCR by reducing the compensation gap relative to a 100% national lab appointment.

Opportunities to Tap into Academic Recruiting Pipelines

Many candidates who apply to academia are also excellent candidates for national lab careers. At the same time, the number of openings in academia is usually far more limited than those in national labs. There is an opportunity for national labs to engage with university partners to tap into their recruiting pipelines and explore ways to get referrals for candidates who do not receive offers from academia. In many cases, these candidates are comparable in technical talent to those

who do receive offers from academia, but were not considered because of specific needs and hiring priorities in academic departments.

To recruit personnel deciding between a career in academia and a national lab, the national labs can emphasize team success, as opposed to single-principal investigator (PI) success that is often a feature of academic careers. The potential to work with a team and contribute to something significant can be inspiring and national labs need to find ways to reinforce that message in their recruiting and retention messages. Anecdotally, the same motivations observed in Gen Z students and early career researchers to have societal impact are correlated with a desire to contribute to team success.

National labs, academia, and industry are all strongly committed to promoting diversity, equity, and inclusion in the workforce. In addition to the general recruiting opportunities mentioned earlier, there appears to be a targeted opportunity for national labs to partner with academia in recruiting students from underrepresented groups. This can be accomplished through focused efforts such as in recruiting students for lab internships and mentoring students by lab personnel.

International Perspective

Historically, national labs have been successful in attracting international talent, because of their prestige and world-class facilities and research groups, as well as comparatively attractive personal benefits relative to careers in Europe and Japan (for example). As in many areas of science, international talent has contributed to our nation's historical leadership in HPC over the years. Thus, it is important to continue leveraging these international partnerships as recruiting channels for the future. In general, international labs offer pathways for recruiting, in addition to joint research. As a simple example, international partner labs can help advertise internship and postdoc opportunities at U.S. labs.

Recruiting from a worldwide talent pool to bring in top researchers expands DOE's opportunities and is another path to diversifying staff and strengthening the ASCR program. ASCR could model a program after successful peer national laboratory programs in other countries such as the Max Planck Institutes in Germany, INRIA in France, and RIKEN in Japan, which actively recruit the best and brightest researchers from around the world (in addition to domestically) to spend time at their institutions and build life-long connections. Implementing such a program can take the form of personnel exchanges and visiting positions, providing significant funding over a multiyear period for established researchers to come to the United States to develop research programs within the DOE laboratories. Similar to the recommendations in the 2021 Basic Energy Sciences Advisory Committee (BESAC) International Benchmarking Study [40], we recommend that such funding be at the level of $5M over a period of five years per person for top computer scientists [41].

A recent report by BESAC [40] found that "the investment in science in other countries, including resources available for research and the freedom to travel and exchange ideas, has changed the landscape so that the U.S. is no longer automatically the preferred destination for career development." As a result, fewer students and researchers are coming to study and work in the

United States. Thus, the available talent pool is shrinking and the diversity of that pool is being diminished.

Increasing the Domestic Talent Pool

Investments by Other Agencies in Increasing the Domestic Talent Pool

One of the priorities to address the insufficient number of domestic students must be increased investment in growing the talent pool. This need has been recognized by the National Science Foundation (NSF) in its CSGrad4US Graduate Fellowship program, see https://www.nsf.gov/cise/CSGrad4US/. There may be opportunities for DOE to partner with NSF in broadening this program to further increase the talent pool for national labs.

Retention of Senior Talent

While attention is usually focused on attracting and retaining the younger members of the domestic talent pool, retention of senior members can also play an important role. Since HPC has become a center of attention for recruitment by industry with the aggressive rise of AI and other related high-performance requirements in the datacenter, non-traditional programs for talent recruitment and retention should be pursued. One possible strategy would be to recruit experienced workers who have decided to retire from industry, academia, or a national lab. These experts could contribute to HPC R&D and operations, but on a part-time basis. The experienced personnel could take on the responsibility of transferring their knowledge and expertise to those beginning their DOE careers" This approach is often practiced in Japan so as to retain expertise even after the legal retirement age of 65. Such "dialing down" of occupational responsibility will allow talented individuals to be associated with the labs while respecting their changing work-life balance.

Findings

1) The attractiveness and prestige of careers in national labs have been on the decline because of internal and external factors such as lack of long-term program vision and stable funding from within the labs and increased competition from industries such as HPC, AI, and big data from outside. Autonomy and flexibility in lab careers have decreased. The COVID-19 pandemic resulted in many companies offering more flexible work arrangements than did the national labs, including joint appointments between academia and industry, as well as joint open-source projects, both of which are opportunities to increase research impact and reduce the compensation gap.

2) DOE is highly vulnerable to losing the knowledge and skills of trained ECP staff as future funding is unclear.

Recommendations

Retention

1) Focus on enabling long-term research agendas and stable funding for lab personnel, especially early/ mid-career employees, as well as recognition and career growth opportunities for them.

2) Build success in retention by coaching all managers (group leads, section heads, directors, etc.) and team leaders (PIs) in best practices for retention ("Don't eat the seed corn") and require them to document in their annual evaluations what measures they have taken in support of retention in the past year.

3) To convey job security, have senior management buffer individual researchers from low-level funding pressures.

4) Explore more ways to recognize and reward accomplishments in national labs at all levels (especially early- and mid-career levels), as is done in industry.

5) Encourage increased retirement ages for those who choose to work longer, while finding ways to "dial down" commitments to support part-time employment. Also, explore emeritus positions as in academia.

6) Document and socialize best practices for engagement between national labs and industry, academia, and international partners. The Computing Research Association (CRA)-Industry committee can help convene and facilitate some of these discussions.

7) To further strengthen retention, ensure that managers and team leads take responsibility for exploring career growth opportunities for team members.

8) Invest in employee growth:
    a) Encourage joint research with partners in other domestic/ international labs.
    b) Establish periodic sabbatical-like rotations/ assignments with domestic/ international partners in academia and national labs.
    c) Increase leadership opportunities for those with interest and potential (e.g., rotation through DOE headquarters or other labs).

Recruiting

1) Invest in a refurbished communications program related to careers in national labs targeted to candidates in advanced computing, emphasizing attractive retirement benefits.
    a) Shared excitement of working on large-team research projects.
    b) Organize "National Lab Day" events at university and K-12 levels with a special focus on traditionally underrepresented groups. Identify early career "brand

ambassadors" in the national labs who can help amplify the communications messaging.

c) Invest in programs that help build future recruiting pipelines, for example, internships, fellowships, campus outreach, and academic/ industry collaborations.
d) Improve and broaden communications related to national lab careers (e.g., create compelling content (videos, interviews) to highlight as role models diverse national lab employees to generate excitement for national lab careers.
e) Convey what is possible at a national lab, for example, how AI can be used to advance science and society (relative to how it is used in industry).
f) Highlight unique aspects of a national lab career, including diversity.
g) Make extensive use of social media for all of the above.

2) Increase investment in internship programs at national labs.

3) Leverage international partners for recruiting talent. Engage with organizations that have direct connections to relevant academic departments for recruiting (for example, CRA for access to CV database, mentoring events).

4) Coordinate recruiting across national labs, for example, joint advertising with pointers to individual lab sites and sharing of applicant resumes.

Best Practices in International Partnerships

1) Invest in enabling select national lab personnel to travel and engage in international research collaborations with the goal of advancing their core research goals and contributing to their retention.

2) Explore joint funding models for international collaborations for mutual benefit, including research advances and recruiting pipelines.

Conclusions

The overall conclusion of this report is that the United States is losing its historical leadership position in advanced scientific computing research that is of interest to the Department of Energy's Office of Advanced Scientific Computing Research. This trend is likely to lead to the United States producing a smaller overall share of technological innovations. A major factor contributing to this decline is the significant investment in fundamental research by other countries: the European Union, Japan, and particularly China in recent years. This investment is already yielding dividends, as research based in these regions has attracted talent and has had a significant impact in the areas studied. The United States has remained competitive in advanced research facilities, such as high-performance computing, because of strategic planning and investment by ASCR and NNSA, but reduced investment in intellectual underpinnings raise concerns about the future success of DOE's scientific endeavors.

The United States was long considered a highly desirable destination for the career development of scientists and its national laboratories were respected worldwide as hosting the most talented researchers. These views are increasingly untenable today. Funding uncertainties and a move away from sustained funding intended to produce foundational innovations to short-term research contracts focused on near-term goals have made DOE laboratories a far less attractive location for both junior and established researchers. Advanced computing research has become the "seed corn" of scientific innovation. Yet ASCR is at risk of losing its current leadership in this vital area because it no longer provides the long-term stable support that has enabled scientists to take a strategic, visionary approach to research in computing science.

To re-establish the DOE ASCR laboratories as a vibrant and exciting place to conduct research in mathematics and computer science, 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.

References

[1] Fast Track Action Subcommittee on Critical and Emerging Technologies of the National Science and Technology Council, "Critical and Emerging Technologies List Update." Feb. 2022 [Online]. Available: https://www.whitehouse.gov/wp-content/uploads/2022/02/02-2022-Critical-and-Emerging-Technologies-List-Update.pdf

[2] D. Fiala, "Bibliometric analysis of CiteSeer data for countries," *Inf. Process. Manag.*, vol. 48, no. 2, pp. 242–253, Mar. 2012.

[3] Wikipedia contributors, "List of countries by number of scientific and technical journal articles," *Wikipedia, The Free Encyclopedia*, Dec. 23, 2022. [Online]. Available: https://en.wikipedia.org/w/index.php?title=List_of_countries_by_number_of_scientific_and_technical_journal_articles&oldid=1128986635

[4] American Association for the Advancement of Science, "U.S. R&D and Innovation in a Global Context: 2022 Data Update." May 2022 [Online]. Available: https://www.aaas.org/sites/default/files/2022-05/AAAS%20Global%20R%26D%20Update%20May%202022.pdf

[5] V. Gewin, "Are divisive US politics repelling international early-career scientists?," *Nature*, Nov. 2022, doi: 10.1038/d41586-022-03604-9. [Online]. Available: http://dx.doi.org/10.1038/d41586-022-03604-9

[6] K. Fischer and S. Aslanian, "The U.S. may never regain its dominance as a destination for international students. Here's why that matters," *APM Reports*, Aug. 03, 2021. [Online]. Available: https://www.apmreports.org/episode/2021/08/03/fading-beacon-why-america-is-losing-international-students. [Accessed: Apr. 05, 2023]

[7] "Mid-decade challenges for national competitiveness archives," *SCSP*. [Online]. Available: https://www.scsp.ai/reports/mid-decade-challenges-for-national-competitiveness/. [Accessed: Jan. 29, 2023]

[8] A. W. Senior *et al.*, "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, Jan. 2020.

[9] UNESCO, *UNESCO Science Report: The race against time for smarter development*. UNESCO Publishing, 2021.

[10] American Academy of Arts and Sciences, "The Perils of Complacency: America at a Tipping Point in Science & Engineering." Sep. 2020 [Online]. Available: https://www.amacad.org/publication/perils-of-complacency

[11] W. Gropp, E. Lusk, N. Doss, and A. Skjellum, "A high-performance, portable implementation of the MPI message passing interface standard," *Parallel Comput.*, vol. 22, no. 6, pp. 789–828, Sep. 1996.

[12] L. Dagum and R. Menon, "OpenMP: an industry standard API for shared-memory programming," *IEEE Computational Science and Engineering*, vol. 5, no. 1, pp. 46–55, Jan. 1998.

[13] C. Lattner and V. Adve, "LLVM: a compilation framework for lifelong program analysis & transformation," in *International Symposium on Code Generation and Optimization, 2004. CGO 2004.*, Mar. 2004, pp. 75–86.

[14] "VASP - Vienna Ab initio Simulation Package." [Online]. Available: https://www.vasp.at/. [Accessed: Sep. 24, 2022]

[15] "Welcome." [Online]. Available: https://www.cactuscode.org/index.html. [Accessed: Sep. 24, 2022]

[16] "Welcome to GROMACS — GROMACS webpage https://www.gromacs.org documentation." [Online]. Available: https://www.gromacs.org/. [Accessed: Sep. 24, 2022]
[17] "Overview." [Online]. Available: https://geant4.web.cern.ch/. [Accessed: Sep. 24, 2022]
[18] J.-L. Vay *et al.*, "Warp-X: a new exascale computing platform for beam-plasma simulations," *arXiv [physics.acc-ph]*, Jan. 08, 2018 [Online]. Available: http://arxiv.org/abs/1801.02568
[19] "Industry and agency council," *Exascale Computing Project*, Jul. 25, 2018. [Online]. Available: https://www.exascaleproject.org/industry-and-agency-council/. [Accessed: Sep. 24, 2022]
[20] R. J. Kee, J. A. Miller, and T. H. Jefferson, "CHEMKIN: a general-purpose, problem-independent, transportable, FORTRAN chemical kinetics code package," Sandia Labs., SAND--80-8003, 1980 [Online]. Available: https://inis.iaea.org/search/search.aspx?orig_q=RN:11548068. [Accessed: Sep. 24, 2022]
[21] UK Department for Business, Energy & Industrial Strategy, "International comparison of the UK research base, 2022." 2022.
[22] "National QIS Research Centers," Apr. 06, 2022. [Online]. Available: https://science.osti.gov/Initiatives/QIS/QIS-Centers. [Accessed: Apr. 17, 2023]
[23] "AlphaGo." [Online]. Available: https://www.deepmind.com/research/highlighted-research/alphago. [Accessed: Apr. 17, 2023]
[24] E. Callaway, "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures," *Nature*, vol. 588, no. 7837, pp. 203–204, Dec. 2020.
[25] E. Callaway, "After AlphaFold: protein-folding contest seeks next big breakthrough," *Nature*, vol. 613, no. 7942, pp. 13–14, Jan. 2023.
[26] R. F. Service, "'The game has changed.' AI triumphs at protein folding," *Science*, vol. 370, no. 6521, pp. 1144–1145, Dec. 2020.
[27] R. Stevens, V. Taylor, J. Nichols, A. Maccabe, K. Yelick, and D. Brown, "AI for science: Report on the department of energy (DOE) town halls on artificial intelligence (AI) for science," Argonne National Laboratory (ANL), Feb. 2020 [Online]. Available: https://www.osti.gov/biblio/1604756
[28] The AI Index 2022 Annual Report, "Daniel Zhang, Nestor Maslej, Erik Brynjolfsson, John Etchemendy, Terah Lyons, James Manyika, Helen Ngo, Juan Carlos Niebles, Michael Sellitto, Ellie Sakhaee, Yoav Shoham, Jack Clark, and Raymond Perrault." [Online]. Available: https://aiindex.stanford.edu/wp-content/uploads/2022/03/2022-AI-Index-Report_Master.pdf
[29] National Artificial Intelligence Research Resource Task Force, "Strengthening and Democratizing the U.S. Artificial Intelligence Innovation Ecosystem: An Implementation Plan for a National Artificial Intelligence Research Resource." Jan. 2023 [Online]. Available: https://www.ai.gov/wp-content/uploads/2023/01/NAIRR-TF-Final-Report-2023.pdf
[30] Johnson and E. B. [d-Tx-30], *National Artificial Intelligence Initiative Act of 2020*. 2020 [Online]. Available: http://www.congress.gov/. [Accessed: Apr. 17, 2023]
[31] D. Reed, D. Gannon, and J. Dongarra, "HPC Forecast: Cloudy and Uncertain," *Commun. ACM*, vol. 66, no. 2, pp. 82–90, Jan. 2023.
[32] P. Beckman *et al.*, "5G enabled energy innovation: Advanced wireless networks for science (workshop report)," USDOE Office of Science (SC) (United States), Mar. 2020 [Online]. Available: https://www.osti.gov/servlets/purl/1606538/

[33] "Declaration for the future of the Internet," *United States Department of State*, Apr. 28, 2022. [Online]. Available: https://www.state.gov/declaration-for-the-future-of-the-internet. [Accessed: Apr. 17, 2023]
[34] Committee on Post-Exascale Computing for the National Nuclear Security Administration, Computer Science and Telecommunications Board, Division on Engineering and Physical Sciences, and National Academies of Sciences, Engineering, and Medicine, "Charting a path in a shifting technical and geopolitical landscape." National Academies Press, Washington, D.C., 2023 [Online]. Available: https://www.nap.edu/catalog/26916
[35] J. Ahrens *et al.*, "Envisioning Science in 2050," Office of Scientific and Technical Information (OSTI), Jun. 2022 [Online]. Available: https://www.osti.gov/servlets/purl/1871683/
[36] J. Shalf, "The future of computing beyond Moore's Law," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 378, no. 2166, p. 20190061, Jan. 2020.
[37] D. Reed, D. Gannon, and J. Dongarra, "Reinventing High Performance Computing: Challenges and Opportunities," *arXiv [cs.DC]*, Mar. 04, 2022 [Online]. Available: http://arxiv.org/abs/2203.02544
[38] I. Georgescu, "Bringing back the golden days of Bell Labs," *Nat Rev Phys*, vol. 4, no. 2, pp. 76–78, Jan. 2022.
[39] S. Z. A. Bizot, "2021 Taulbee Survey CS Enrollment Grows at All Degree Levels, With Increased Gender Diversity." 2021 [Online]. Available: https://cra.org/wp-content/uploads/2022/05/2021-Taulbee-Survey.pdf
[40] BESAC Subcommittee on International Benchmarking, "International Benchmarking Study Report on Critical Research Frontiers and Strategies: Can The U.S. Compete In Basic Energy Sciences?" 2021 [Online]. Available: https://science.osti.gov/-/media/bes/pdf/reports/2021/International_Benchmarking-Report.pdf
[41] M. Riordan, "The end of AT&T," *IEEE Spectrum*, Jul. 01, 2005. [Online]. Available: https://spectrum.ieee.org/the-end-of-att. [Accessed: May 09, 2023]
[42] "Department of Energy Mission," *Energy.gov*. [Online]. Available: https://www.energy.gov/mission. [Accessed: Jan. 29, 2023]
[43] "DOE Office of Science Mission," *Energy.gov*. [Online]. Available: https://www.energy.gov/science/mission. [Accessed: Jan. 29, 2023]
[44] S. Williams, A. Waterman, and D. Patterson, "Roofline: an insightful visual performance model for multicore architectures," *Commun. ACM*, vol. 52, no. 4, pp. 65–76, Apr. 2009.
[45] "Publications resulting from the use of NERSC resources," *NERSC*. [Online]. Available: https://www.nersc.gov/news-publications/publications-reports/nersc-user-publications/. [Accessed: May 09, 2023]
[46] "Publications." [Online]. Available: https://www.alcf.anl.gov/science/publications. [Accessed: May 09, 2023]
[47] "Publications," *Oak Ridge Leadership Computing Facility*. [Online]. Available: https://www.olcf.ornl.gov/publications/. [Accessed: May 09, 2023]
[48] "Discover EuroHPC JU," *The European High Performance Computing Joint Undertaking (EuroHPC JU)*. [Online]. Available: https://eurohpc-ju.europa.eu/about/discover-eurohpc-ju_en. [Accessed: Jan. 29, 2023]
[49] "Key enabling technologies for Europe's technological sovereignty." [Online]. Available: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2021)697184.

[Accessed: Jan. 29, 2023]

Appendix I: Historical Perspectives on ASCR's Leadership in Computational Science and Engineering and HPC Systems

The United States is currently among the global leaders in advanced computing and high-end CS&E in terms of algorithmic invention, theoretical understanding, computing facilities and infrastructure, and translation into applications. It can remain so by continuing to apply the recipes and sustaining the "hardware-software-wetware ecosystem" that have made the United States an attractive destination for computational mathematicians since the dawn of digital computing. DOE's missions are a natural attractor within this ecosystem, as exemplified by von Neumann and Metropolis at Los Alamos in the 1940s and early 1950s.

The mission of the Advanced Scientific Computing Research program is to discover, develop, and deploy computational and networking capabilities to analyze, model, simulate, and predict complex phenomena important to the Department of Energy [42], [43]. ASCR has world-class capabilities in the following areas that have been well established over the history of the program:

- Applied mathematics
- Computer science
- Supercomputing and advanced networking facilities (the complementary role of research and facilities)
- Computational partnerships and cross-disciplinary technology translation

Applied Mathematics

Within the U.S. computational ecosystem, DOE (ASCR and to some extent NNSA) is the largest funder of applied mathematics research, exceeding that of NSF, any other federal agency, and any identifiable industrial lab. The ASCR Applied Mathematics program once made block grants to the labs to build up and retain human capacity using the research model pioneered by Bell Labs. In more recent decades, applied mathematicians have consistently demonstrated the value of their presence in multidisciplinary programs across DOE, such as SciDAC. Advances in applied mathematics have also been fostered by top-down targeted topical initiatives carried out through competitive awards, such as multiscale mathematics, uncertainty quantification, and scientific machine learning.

The tremendous advances in hardware fueled by Moore's law within a single core, the concurrent use of many cores, and the integration of accelerators hardwired for specialized operations within a single cycle have been complemented and even dwarfed by advances in algorithms over a sustained period of decades. For example, for one of the most common kernels in CS&E, namely, solving the multidimensional Poisson (or potential, equilibrium, or diffusion) equation, performance improvements due to algorithmic innovation have outpaced those due to Moore's law. A multigrid Poisson solver on a 1024^3 cube that takes one second today would have required six months on hardware from 36 years ago but would require nearly 32 thousand years by the algorithm known to von Neumann on *today's* hardware. Fortunately, despite the challenges of mapping a mathematical hierarchy of scales onto an architectural hierarchy of memory and

processing elements, algorithm designers have consistently proved able to implement the enviable catalog of today's algorithms on *today's* hardware, thus multiplying together the improvement factors available from each for users.

Multigrid is just one in a series of advances from mathematicians that appear regularly and shave powers off of the previously best-known computational complexity for an important computational kernel. Following the fast Fourier transform in the 1960s and multigrid in the 1970s, the 1980s saw the emergence of the fast multipole method, the 1990s brought sparse grids, the 2000s hierarchically low-rank matrices, and the 2010s randomized algorithms of linear algebra. Of these major advances in complexity reduction, paying increasing benefits at increasing problem scales, four of the six originated or co-originated from researchers working in the United States. All six are now part of the computational infrastructure supporting DOE's CS&E missions.

With the exception of randomized algorithms, which open up a new paradigm that is in its infancy of applicability, the earlier five breakthroughs mentioned are all hierarchical, making use of a recursion of scales in an intuitively motivated way. Intuition often leads to theoretical understanding in algorithmic invention, and, in turn, theoretical understanding often suggests new ways to apply an existing algorithm. In both cases, a rich interplay of theory and practice improves the ability to tune algorithmic parameters in order to maximize the accuracy in outputs of interest within a computational budget. Historically, DOE lab-based researchers have contributed more by creating and stretching algorithmic techniques, and academic researchers have contributed more by supplying theoretical insight; but contributions flow in both directions to advance DOE's computational missions.

Applied mathematicians have adapted the solvers required by computational scientists and engineers to each stage of evolution in high-end architecture that comes their way, such as the transformation of dense linear algebra solvers from Level 1 Basic Linear Algebra Subprograms (BLAS)-oriented LINPACK for vector hardware, to Level 3 BLAS-oriented LAPACK for cache-based architectures, followed by ScaLAPACK for distributed-memory architectures, PLASMA/MAGMA for many-core and GPU processors, and SLATE for today's heterogeneous processors. Moreover, algorithm developers have sometimes led architectural developments by demonstrating that computational tasks can be reduced to readily implemented special processing units, such as batched small-matrix convolutions on GPUs today.

Linear algebra is just one area in which applied mathematicians regularly infuse DOE CS&E with new computational techniques. High-order discretizations possess a higher arithmetic intensity relative to conventional second-order discretizations and are well-matched to increasingly bandwidth-limited many-core processors. Mimetic methods preserve conserved quantities such as energy in a discretized system and can thus rule out instabilities. Reduced-order modeling has become a valuable source of surrogates, particularly for optimization, where the quality of the surrogate can be low at first and ramped up as the optimum is approached. Statistical emulation directly from data for systems with high variability has become an important alternative to Monte Carlo on ensembles of first-principles simulations. Machine learning can deliver reliable constitutive properties for simulations that are expensive to compute from first principles. The catalog of mathematical contributions that began as "blue sky" research and migrated into quotidian use goes on and on.

Computer Science

DOE's computer science research program has not spanned the full breadth of areas covered by most academic computer science departments, focusing instead on areas related to parallel computing, modeling and simulation, and to scientific data management and analysis.

There is a vibrant international HPC community. Strong research programs around the world, but especially in China, Japan, and several European countries, are notable for their innovations and impact in HPC architectures, system software, and applications. Much HPC systems software research benefits enormously from international collaborations, even when there are differences in the underlying hardware vendors and architectures and in applications representing different national priorities.

As an example of strong international collaboration, the Message Passing Interface (MPI) dominates HPC programming worldwide as the internode parallelism model. First published in 1994, MPI has seen three major revisions and many minor ones, the most recent (MPI 4.0) in 2021. The standard has expanded and adapted over the years based on input from application developers and computer architects, changing node architectures, the need to integrate parallel I/O, fault tolerance, and adding and then adapting one-sided communication. The basic abstractions in MPI have persisted over close to 30 years, namely, a set of processes that communicate with each other as if they have a direct connection, despite major changes in network topologies and features such as offload.

Significant research has informed both the MPI interface and its implementations. The interface was developed in an open community, the MPI Forum, with international participation of researchers and software developers from national laboratories, universities, and vendors. This broad participation has been essential to the widespread acceptance of MPI, making it a de facto requirement for any HPC system used for modeling and simulation.

As will be discussed in more detail below, node architectures have changed significantly over the decades from a single core to shared memory (SMP modes), multicore, and many-core, and with the addition of accelerators. These developments have led to disagreement on the proper programming approaches for use within nodes and more tightly controlled standards efforts for interfaces such as OpenMP have sometimes slowed innovation. This situation has spurred much research on thread management, automatic load balancing, and other runtime issues, as well as language and compiler support for parallelization and synchronization. Some of this runtime work required an understanding of and ability to augment or work around operating system features. The lack of a widely accepted portable programming model for GPUs has led to several efforts within the DOE community to provide both portability and good performance across GPUs via various combinations of libraries, code generators, and runtime systems.

A major node-level performance challenge has been the ever-deepening memory hierarchy with increasing idiosyncratic memory spaces in the form of scratch pads, non-uniform costs for shared memory, and various types of accelerator memory. DOE's work in performance analysis and modeling, such as the Roofline model [44], and various detective-like studies of application

performance have provided important insights that have influenced production hardware and software. In addition, techniques for automatically generating and searching for well-optimized code have become common for routines such as matrix operations (i.e., the BLAS), fast Fourier transforms (FFTs), stencils, sparse matrix kernels, higher-level libraries, and application frameworks.

Much of DOE's science requires distributed computing with large national and international collaborations that share data, major instruments such as telescopes and light sources that support remote access, and myriad sensors, both within laboratories and embedded in the environment. ASCR has a long record of leadership in the development of wide area networking technologies and distributed computing technologies to connect researchers with science facilities and with each other. Already in the 1980s, 10 kbps Internet links connected many laboratories and universities. Today, ESnet, established in 1986, is deploying 400 Gbps links, a stunning eight orders of magnitude performance improvement in 35 years. Underpinning these developments are a long series of research innovations in network protocols and applications, from early work on network congestion control to more recent work on instrumentation, bandwidth reservation, and high-speed transport. ASCR has led the way in exploiting the increasingly ubiquitous high-speed connectivity provided by ESnet and other science networks to deliver new capabilities, such as federated identity, automated data movement and replication, and distributed science workflows. ASCR research continues to tackle emerging new research challenges in distributed computing, such as managing access and identity; architecting the distributed system to minimize data movement and maximize the use of shared resources; and supporting real-time workflows, including addressing failures in the network, computing systems, or instruments.

Supercomputing Facilities, and the Complementary Relationship between Research and Facilities

Since the 1970s, DOE HPC facilities have operated, and made accessible to approved users, the most advanced computing systems available for science and engineering research. Today, for example, National Energy Research Scientific Computing Center (NERSC) operates Perlmutter and Cori, Oak Ridge Leadership Computing Facility (OLCF) operates Summit and Frontier, and Argonne Leadership Computing Facility (ALCF) operates Polaris and Theta (and, in 2023, Aurora). These and predecessor systems have enabled tens of thousands of scientists and engineers to conduct impactful research in many areas, as documented in numerous publications [45]–[47].

A key reason for the sustained success of DOE HPC facilities has been the close partnerships that they have established and sustained over many years: with the computer industry on supercomputer design; with HPC software and algorithms experts, within and outside DOE labs, on the design and implementation of new methods and tools needed to exploit new supercomputer architectures; and with application scientists on the design and implementation of the increasingly sophisticated application codes needed to take advantage of ever-more complex supercomputer architectures. These partnerships have allowed scientists and engineers to make ever-more-effective use of supercomputers as they increased in speed from 160 megaflop/s (1.6×10^7 flop/s)

in 1980 to close to 1 exaflop/s (10^{18} flop/s) in 2022: an increase of 10 orders of magnitude in little more than 40 years.

The partnerships have been necessary because of the rapid evolution in computing technologies. From the 1980s onwards, DOE labs led the way in harnessing the rapidly evolving capabilities of microprocessor technologies, linking first tens, then thousands, and today millions of conventional microprocessors and accelerators. As the power of these systems increased, the architectures became more and more complex with new deep memory hierarchies, communications technologies, and parallel data storage technologies, and other changes in both their constituent components and how those components were assembled. These developments in turn required major changes in the design and implementation of scientific codes, the algorithms on which they were based, and the software used to operate computing systems and computer facilities.

Computational Partnerships

ASCR has excelled in fostering computational partnerships between its own primary researchers in applied mathematics and computer science and those of the other offices in the Office of Science and in NNSA – a highly competitive *organizational-level success story*.

The flagship of all such programs is the Scientific Discovery through Advanced Computing program (SciDAC), now in its twenty-second year. SciDAC PIs from the applications offices were not allocated full-time equivalents (FTEs) to develop their own solver, discretization, I/O, visualization, or other infrastructure, but were encouraged to adopt it from ASCR partners. ASCR enabling technologists were not allowed to claim success for their work on model problems but were required to demonstrate it on DOE applications.

Many anecdotes have been told by application domain computational scientists in SciDAC about their positive interactions with ASCR's enabling technologists. A favorite from the first SciDAC PI meeting in 2001 occurred at the poster of the HEP lattice quantum chromodynamics team when visited by an ASCR algebraic multigrid team (AMG). The latter recognized immediately to their amazement that the HEP LQCD team had discovered a form of algebraic multigrid a decade earlier but it was missing some ingredients, causing it to stall. When the ASCR AMG team returned a few weeks later with their results on a model LQCD system, the HEP team declared to their amazement that the slowly decaying fine space modes proposed as coarse space basis functions discovered by the ASCR AMG team were essentially their "instantons." This collaboration proceeded in many dimensions, literally and figuratively, as did many others when researchers from different disciplines met at the "watering hole" of the supercomputer where insights are shared.

DOE Graduate Fellowships

The DOE Computational Science Graduate Fellowship (CSGF) program is a jewel in the crown of ASCR and NNSA, accepting into lucrative, lab-oriented Ph.D. fellowships each year 15 to 30 U.S. citizen or permanent resident students. The students are circulated to the labs and are required

to participate in at least one DOE lab internship. Their selectivity is higher than the labs themselves, typically about 5% of applicants. The lab mentors nearly universally report strong progress and often importation into computational practice of a recently invented technique from academia as a result of these internships. CSGF fellows are not inexpensive, since the DOE offers their host universities full tuition for four years of doctoral candidacy; however, their contributions are sometimes priceless. The CSGF program provides a barometer for disciplines that will be of interest to future DOE computing. Computational biology, machine learning, and quantum computing are among the subjects that began to swell in the ranks of CSGF applicants before the labs were hiring as high a percentage of employees in these categories.

A major benefit to DOE from these programs[8] shaped and supported by ASCR programs is that they lead to the creation and maintenance of progressive[9] software environments at the leadership-class facilities, which is an important contributor to productivity and hence competitiveness.

Industrial Partnerships

There are complementary roles played by DOE research in computer architectures and systems to advance cutting-edge technology developments and the DOE facilities that must ensure reliable and usable deployment of systems at scale for our user community. For example, the long path from GPUs being special-purpose graphics accelerators to becoming the primary engine for petascale and exascale computing was a decades-long process involving both research and facilities at different stages of technology development. Recognition by DOE computer architecture researchers of the opportunities in the computational throughput and memory performance of GPUs led to the establishment of small-scale demonstration systems that made these special-purpose systems more accessible and programmable through targeted generalization and experimental programming systems at small scale. As the idea of "generalizing" the GPU (the GPGPU) began to take hold, the commercial interest in adding those features – including hardware features such as double-precision floating point and advanced language technologies such as CUDA – grew, leading to the emergence of truly flexible GPGPU platforms that really could deliver for science. This research endeavor involved hardware experts in DOE who could work closely with industry and with DOE's applied mathematicians. The broad-based cross-disciplinary nature of these collaborations is a unique capability that cannot be matched by industry research laboratories or by academia.

With Titan, OLCF took on the herculean task of ensuring that the emerging GPGPU technology could scale-up to leadership class and be reliable and programmable at scale. Titan became the template for a series of follow-up systems that presaged the mass movement of HPC from a central processing unit (CPU)-dominated market to systems dominated by GPU acceleration at exascale. But Titan would not have been a practical option to field were it not for the decade or more of small-scale experimental systems and collaborations between the HPC hardware experts in research and their industry partners to get GPUs to the point that they were a credible option for HPC at scale.[10]

These kind of long-term sustained research partnerships to bring new technologies into the mainstream and symbiotic efforts in scaling up and productizing that capability have been essential to the innovation pipeline over the decades of DOE leadership of HPC system design and are essential to DOE's continued leadership well into the future. A partnership of LLNL, ANL, and IBM co-designed the highly influential Blue Gene system series, which received a National Medal of Technology and Innovation. The co-design effort between Cray and Sandia for the development of Red Storm became the basis for the long-lasting and highly successful Cray XT machine series that underpinned Jaguar and Jaguar-PF at OLCF and NERSC's Franklin system.

Appendix II: Exascale Projects in Europe and Japan

EuroHPC Joint Undertaking

In 2018, Europe initiated the European High Performance Computing Joint Undertaking (EuroHPC JU), a legal entity to coordinate European efforts for pooling resources into supercomputing [48].

The mission of EuroHPC JU is threefold: (1) to develop, deploy, extend, and maintain a world-leading federated, secure, and hyperconnected supercomputing, quantum computing, services, and data infrastructure ecosystem; (2) to support the development and uptake of demand-oriented and user-driven innovative and competitive supercomputing systems based on a supply chain that will ensure components, technologies, and knowledge, limiting the risk of disruptions, and the development of a wide range of applications optimized for these systems; and (3) to widen the use of that supercomputing infrastructure to a large number of public and private users and support the development of key HPC skills for European science and industry.

The current funding cycle of EuroHPC JU is running from 2021 to 2027, with a multiannual financial framework from the European Union, the EU member states, and industry. Furthermore, EuroHPC JU is currently calling for proposals for a Framework Partnership Agreement (FPA) for developing a large-scale European initiative for an HPC ecosystem based on the RISC-V processor architecture.

Several countries in Europe (e.g., Germany, United Kingdom) are investing in their own national exascale and quantum computing capabilities in addition to the EuroHPC JU undertaking. Some of these initiatives have timelines reaching to 2031/ 2032.

The EU and several national parliaments are discussing technological sovereignty for key enabling technologies, including advanced manufacturing and materials, life sciences, micro/ nanoelectronics and photonics, artificial intelligence, and security and connectivity technologies[49].

The Japanese FugakuNEXT Project

The current (ca. 2023) flagship supercomputer in Japan is Fugaku, which was successfully developed and deployed with the Flagship2020 project, sponsored by MEXT (the Japanese Ministry of Education, Sports, Culture, Science and Technology) and executed by RIKEN R-CCS along with its industrial partner Fujitsu. The project spanned more than ten years, from mid-2010 to early 2021, with proactive co-design activities that involved the entire high-end scientific computing community of Japan centered around R-CCS. The resulting processor A64FX was a server-grade HPC chip with high bandwidth memory integration for the first time for a general-purpose CPU and sported an over three time increase in memory bandwidth compared with mainstream HPC CPUs, while being completely compatible with the ARMrm ecosystem. Fugaku became the fastest machine in the world in multiple benchmarks including High-Performance Linpack (HPL)/TOP500 and High Performance Conjugate Gradients (HPCG) in June 2020 and achieved speedups of nearly two orders of magnitude in target applications. It remains competitive with U.S. exascale machines, ranking second to Frontier in the TOP500 November 2022 tests, and shows impressive results especially in real applications.

As a follow-on to Fugaku, Japan/MEXT has already officially initiated its national feasibility study endeavor toward post-exascale research infrastructures (FugakuNEXT), with a 10-year roadmap report to be formulated within a few years. The project is expected to largely track the timeline and the higher-level R&D co-design methodologies that had been conducted for Fugaku, but with several improvements. One is to institute international collaboration with friendly partners from the inception of the project, especially those within the United States, including the DOE labs and U.S. industrial vendors. A second improvement is not only to focus on hardware development and applications but to place more emphasis on system software and machine operations. A third improvement is to accommodate broader computational capabilities, including possible adoption of new computing models such as AI and quantum computing (or simulations thereof) at extreme capabilities, possibly well above industrial standards. Overall, the performance target for FugakuNEXT would be more than an order of magnitude performance increase over Fugaku, while staying within a similar power and financial budget.

The official first phase of the project, the Feasibility Study (FS) for FugakuNEXT started in the late summer of 2022. The overall project consists of multiple teams involving researchers and engineers from RIKEN and its domestic research/ academic partners, as well as multiple major IT vendors, especially those from the United States. It is expected that FS will be extended to 2.5 years and will persist until the end of Japanese fiscal year 2024 (end of March 2025). Related programs run in parallel by MEXT and other ministries such as Ministry of Economics, Trade and Industry (METI) in advanced IT and semiconductors could have a strong relationship to FS. If the FS is successful in coming up with multiple, credible system candidates as the successor to Fugaku, the next phase of the project, basic design development, will commence immediately following FS in April 2025. Again, potential partnership with international entities such as those in the United States, including DOE, will be extensively investigated for the subsequent phases. If all goes well, manufacturing of the machine will start in 2029, with full installation and deployment in 2030.

Abbreviations and Acronyms

AI	Artificial Intelligence
AMR	Adaptive Mesh Refinement
ANL	Argonne National Laboratory
ASC	Advanced Simulation and Computing
ASCI	Accelerated Strategic Computing Initiative
ASCAC	Advanced Scientific Computing Advisory Committee
ASCR	Advanced Scientific Computing Research
CHIPS	Creating Helpful Incentives to Produce Semiconductors
CPU	Central Processing Unit
CSGF	Computer Science Graduate Fellowship
DOE	Department of Energy
DQS	Digital Quantum Simulation
ECI	Exascale Computing Initiative
ECP	Exascale Computing Project
EUV	Extreme Ultraviolet
FLOP	Floating-Point Operation
FPGA	Field Programmable Gate Arrays
GPUs	Graphics Processing Units
HBM	High-Bandwidth Memory
HPC	High-Performance Computing
IaaS	Infrastructure as a Service
ISA	Instruction Set Architecture
LANL	Los Alamos National Laboratory
LLNL	Lawrence Livermore National Laboratory
MPI	Message Passing Interface
NIST	National Institute of Standards and Technology
NNSA	National Nuclear Security Administration
NRE	Non-Recurring Engineering
NUMA	Non-Uniform Memory Access
NVM	Non-Volatile Memory
ORNL	Oak Ridge National Laboratory
PaaS	Platform as a Service
PRACE	Partnership for Advanced Computing in Europe

SaaS	Software as a Service
SDRAM	Synchronous Dynamic Random-Access Memory
SiP	System-in-Package
SNL	Sandia National Laboratories
SoC	System-on-a-Chip
TSMC	Taiwan Semiconductor Manufacturing Company