NATIONAL ACADEMIES Sciences Engineering Medicine

NATIONAL ACADEMIES PRESS
Washington, DC

Charting a Path in a Shifting Technical and Geopolitical Landscape: Post-Exascale Computing for the National Nuclear Security Administration (2023)

DETAILS

120 pages | 8.5 x 11 | PAPERBACK
ISBN 978-0-309-70108-2 | DOI 10.17226/26916

CONTRIBUTORS

Committee on Post-Exascale Computing for the National Nuclear Security Administration; Computer Science and Telecommunications Board; Division on Engineering and Physical Sciences; National Academies of Sciences, Engineering, and Medicine

SUGGESTED CITATION

National Academies of Sciences, Engineering, and Medicine. 2023. *Charting a Path in a Shifting Technical and Geopolitical Landscape: Post-Exascale Computing for the National Nuclear Security Administration*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26916.

NATIONAL ACADEMIES Sciences Engineering Medicine

NATIONAL
ACADEMIES
PRESS
Washington, DC

Charting a Path in a Shifting Technical and Geopolitical Landscape

Post-Exascale Computing for the National Nuclear Security Administration

Committee on Post-Exascale Computing for the National Nuclear Security Administration

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

Consensus Study Report

NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by the National Nuclear Security Administration under award number DE-EP0000026/89233121FNA400371. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-70108-2
International Standard Book Number-10: 0-309-70108-2
Digital Object Identifier: https://doi.org/10.17226/26916

This publication is available from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America.

Suggested citation: National Academies of Sciences, Engineering, and Medicine. 2023. *Charting a Path in a Shifting Technical and Geopolitical Landscape: Post-Exascale Computing for the National Nuclear Security Administration*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26916.

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. John L. Anderson is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

Rapid Expert Consultations published by the National Academies of Sciences, Engineering, and Medicine are authored by subject-matter experts on narrowly focused topics that can be supported by a body of evidence. The discussions contained in rapid expert consultations are considered those of the authors and do not contain policy recommendations. Rapid expert consultations are reviewed by the institution before release.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

COMMITTEE ON POST-EXASCALE COMPUTING FOR THE NATIONAL NUCLEAR SECURITY ADMINISTRATION

KATHERINE A. YELICK (NAE), University of California, Berkeley, *Chair*
JOHN B. BELL (NAS), Lawrence Berkeley National Laboratory
WILLIAM W. CARLSON, IDA Center for Computing Sciences
FREDERIC T. CHONG, University of Chicago; Infleqtion
DONA L. CRAWFORD, Lawrence Livermore National Laboratory (retired)
MARK E. DEAN (NAE), University of Tennessee, Knoxville
JACK J. DONGARRA (NAE), University of Tennessee, Knoxville
IAN T. FOSTER, University of Chicago; Argonne National Laboratory
CHARLES F. McMILLAN, Los Alamos National Laboratory (retired)
DANIEL I. MEIRON, California Institute of Technology
DANIEL A. REED, The University of Utah
KAREN E. WILLCOX (NAE), The University of Texas at Austin

Staff

THƠ H. NGUYỄN, Senior Program Officer, *Study Director*
JON K. EISENBERG, Senior Board Director
GABRIELLE M. RISICA, Program Officer
SHENAE A. BRADLEY, Administrative Assistant

NOTE: See Appendix D, Disclosure of Unavoidable Conflicts of Interest.

COMPUTER SCIENCE AND TELECOMMUNICATIONS BOARD

LAURA HAAS (NAE), University of Massachusetts Amherst, *Chair*
DAVID CULLER (NAE), University of California, Berkeley
ERIC HORVITZ (NAE), Microsoft Research
CHARLES ISBELL, Georgia Institute of Technology
ELIZABETH MYNATT, Georgia Institute of Technology
CRAIG PARTRIDGE, Colorado State University
DANIELA RUS (NAE), Massachusetts Institute of Technology
MARGO SELTZER (NAE), University of British Columbia
NAMBIRAJAN SESHADRI (NAE), University of California, San Diego
MOSHE Y. VARDI (NAS/NAE), Rice University

Staff

JON K. EISENBERG, Senior Board Director
SHENAE A. BRADLEY, Administrative Assistant
RENEE HAWKINS, Finance Business Partner
THƠ H. NGUYỄN, Senior Program Officer
GABRIELLE M. RISICA, Program Officer
BRENDAN ROACH, Program Officer
NNEKA UDEAGBALA, Associate Program Officer

Reviewers

This Consensus Study Report was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published report as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the study charge. The review comments and draft manuscript remain confidential to protect the integrity of the deliberative process.

We thank the following individuals for their review of this report:

RANDAL BRYANT (NAE), Carnegie Mellon University
DAVID CULLER (NAE), Google
WILLIAM DALLY (NAE), NVIDIA Corporation
ALAN EDELMAN, Massachusetts Institute of Technology
DENNIS GANNON, University of Indiana
MARK HOROWITZ (NAE), Stanford University
DANIEL KATZ, University of Illinois at Urbana-Champaign
CHERRY MURRAY (NAS/NAE), University of Arizona
JIM RATHKOPF, Los Alamos National Laboratory; Lawrence Livermore National Laboratory
ROBERT ROSNER, University of Chicago
VALERIE TAYLOR, Argonne National Laboratory

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the conclusions or recommendations of this report, nor did they see the final draft before its release. The review of this report was overseen by WILLIAM GROPP (NAE), University of Illinois at Urbana-Champaign, and SUSAN GRAHAM (NAE), University of California, Berkeley. They were responsible for making certain that an independent examination of this report was carried out in accordance with the standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the authoring committee and the National Academies.

Contents

APPENDIXES

Preface

The National Nuclear Security Administration (NNSA) relies on advanced computing capabilities for modeling and simulation to deliver its stockpile stewardship mission. Underpinning NNSA's computing capabilities are leading-edge, high-performance computing (HPC) technologies, and a world-class scientific computing workforce. As the nuclear stockpile ages and evolves, the mission to accurately model and simulate weapons' behavior becomes significantly more complex. Concurrently, the computing technology and market landscape is rapidly shifting on all fronts, further challenging NNSA's ability to develop and deploy the kind of leadership computing capabilities needed to ensure the success of its mission. The committee believes that realizing post-exascale computing will require an integrated program that extends beyond hardware to include algorithms, software, and new operational models, as well as workforce development, among other considerations. This view informs the committee's approach to this study. As this report describes, meeting these challenges will require a roadmap, sustained support and investment from the policy community, and visionary leadership at both NNSA and its three national laboratories to ensure that the computing platforms and talent are in place to meet NNSA requirements in a post-exascale era.[1]

[1] This report uses "post-exascale era" as the 20-year period starting with the installation of the first DOE exascale system in 2022 and "post-exascale systems" as the leading-edge HPC systems that will follow the current exascale procurements. The committee has chosen not to describe these future systems as zettascale systems, because the focus is not on a particular floating-point rate but on time-to-solution for problems of the scale and accuracy needed for the future challenges associated with stockpile certification.

As mandated by Congress in the 2021 National Defense Authorization Act, a committee was established by the National Academies of Sciences, Engineering, and Medicine to review "the future of computing beyond exascale computing to meet national security needs at the National Nuclear Security Administration." In the context of the NNSA mission needs, the committee was asked to evaluate future technology trajectories as well as the U.S. industrial base required to meet those needs. (See Appendix A for the complete statement of task.)

The committee engaged with leading high-performance computing (HPC) developers, international HPC operators (especially for nuclear energy and technology), and the NNSA/Advanced Simulation and Computing (ASC) program itself. The committee received briefings and reference material, both classified and unclassified; reviewed the results of related government working groups and studies; and deliberated extensively to assess the current state of the NNSA/ASC mission, future mission scope and needs, current HPC capabilities and technology directions, and scientific computing workforce needs. In response to review feedback on this report, the committee also asked for written inputs from NNSA on computing demands, computational patterns, and the role of artificial intelligence. A classified annex was deemed unnecessary, as it would not provide any additional information that would affect the report's findings and recommendations. Examples of future mission drivers include aging of nuclear materials, high-resolution simulations of subcritical experiments, and understanding hypersonic flows in reentry environments, all with the ever-present need for assessment of margins and uncertainties. These will require advances not just in hardware capability but in mathematical modeling, algorithms, and software that go well beyond simple scaling of problem size or resolution. These models must be implementable in classified software that will run on these future machines and will depend on a wide range of software from operating systems to libraries developed by others.

High-performance computing supporting the nuclear deterrent has faced challenges in the past, most notably 30 years ago, when the United States entered a moratorium on full-scale underground nuclear testing. The challenges that NNSA and its laboratories face today are different from, but equally as daunting as, those at the beginning of the stockpile stewardship era. Thirty years ago, the technical path forward for computing was clear, and NNSA was able to leverage the growth path of the computing industry. Neither of these conditions holds today. Furthermore, the historical limits on timely solutions to relevant problems—floating-point operations per second—are rarely limiting today. As the laboratories articulated, memory access constraints often limit overall application performance. Thus, the committee advocates metrics such as solving hero calculations that today require a year of machine time in days. These types of metrics are far more relevant than the "scale" of the machine. Realization of these types of goals would dramatically increase the effectiveness of the scientific and engineering staff of the laboratories in supporting the deterrent.

The committee believes that bold and transformative actions will be required for NNSA to continue to succeed in its evolving mission. These actions include the following:

- Development of an aggressive computing roadmap with relevant metrics on compelling application requirements;
- New organizational models to support talents to focus on both short-term and long-term problems;
- Bold, sustained investments in research and engineering of hardware, software, and algorithms;
- Innovative partnership models with both traditional and nontraditional partners for acquisition and deployment;
- Organizational models to support focus and creativity; and
- Expanded government-wide collaborations.

Embracing these approaches will require key organizational leadership that is able to create vision, strategy, and advocacy to meet the post-exascale challenges. Simply put, NNSA needs to fundamentally rethink its advanced computing research, engineering, acquisition, deployment, and partnership strategy. As this report details, a simple extension of the strategy developed over the past 30 years will be insufficient for mission success; it must be reimagined to ensure success.

Summary

A core mission of the National Nuclear Security Administration (NNSA) is to ensure that the United States maintains a safe, secure, and reliable nuclear stockpile through the application of unparalleled science, technology, engineering, and manufacturing. The use of modeling and simulation is front and center in realizing science-based stockpile stewardship, especially since 1992, when the United States voluntarily ceased underground nuclear explosives testing. These simulations require leading-edge computer platforms, sophisticated physics and engineering application codes, and expertise in applied mathematics and computer science for the design, engineering qualification, surveillance, maintenance, and certification of the nuclear stockpile. NNSA's Advanced Simulation and Computing (ASC) program[1] has been successful in providing high-performance computing (HPC) systems, software, methods, and workforce that are the foundation for this computational work. The combination of high-end computing facilities and expertise also makes the NNSA laboratories a national resource that can be and has been deployed for other critical priorities, on both an ongoing and an emergency basis. The demand for increased computing capability will continue unabated, stemming from the need for more detailed simulations of the aging stockpile, higher confidence in those simulations, and, potentially, increased use of artificial intelligence (AI) methods, which may involve highly expensive training of large models.

In 2022, the United States installed its first exascale computing system for the Department of Energy (DOE) Office of Science, with an NNSA system scheduled for 2023. The DOE Exascale Computing Project (ECP)[2] has developed new applications

[1] Los Alamos National Laboratory, 2021, "Nuclear Weapon Simulation and Computing," Advanced Simulation and Computing (ASC) Program, https://www.lanl.gov/projects/advanced-simulation-computing.

[2] National Nuclear Security Administration, 2023, "Exascale Computing Project," https://www.exascaleproject.org.

capabilities, parallelization approaches, and software tools, while co-developing the computing systems in collaboration with vendor partners. NNSA is positioned to take full advantage of exascale computing, but demand for more computing will continue to grow beyond exascale, driven by both familiar applications and new mission drivers and new computational approaches that will use high-end computing. Visionary leaders and creativity will be needed to move existing codes to next-generation platforms, to reconsider the use of advanced computing for all of NNSA's current and emerging mission problems, and to envision new types of computing systems, algorithmic techniques implemented in software, partnerships, and models of system acquisition.

OVERARCHING FINDING: 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. *Business-as-usual will not be adequate.*

The approach used to reach petascale and now exascale capabilities is unlikely to be sufficient for the next two decades. Instead, NNSA will need to reevaluate how its mission problems, not limited to physics simulations, are best solved through advanced computing, and rethink what type of models, algorithms, and data analysis techniques are suited to each problem; what computing capabilities will be needed; and how it can best acquire those capabilities.

Owing to a confluence of technology, marketplace, and workforce challenges, NNSA's ASC program is at a critical crossroads. The program has for decades delivered impressive and state-of-the-art predictive simulation capabilities using in-house expertise in applied mathematics, computer science, and the physical sciences, along with research and development (R&D) investments in the computer vendor community. However, the current deployment model is not likely to be sufficient for future NNSA missions.

ROLE AND IMPORTANCE OF EXASCALE AND POST-EXASCALE COMPUTING FOR STOCKPILE STEWARDSHIP (CHAPTER 1)

Today's increasingly complex geopolitical landscape has refocused attention on nuclear security. Meanwhile, the nation's nuclear stockpile continues to age, requiring continual surveillance and the maintenance, redesign, and replacement of components. New demands may also arise, such as for alternative delivery mechanisms and to meet operational requirements for extreme environments. These challenges will require improved predictive capabilities and reduced uncertainties, more detailed models, and the inclusion of more detailed physics phenomena, whether obtained from first principles or learned from experimental data. In addition to more sophisticated simulations, advanced computing is essential for the analysis of data sets from NNSA's experimental facilities, for the use of new AI methods, and for the control of complex experimental systems. Therefore, it has become clear that mission requirements for advanced computing will continue to grow in both complexity and scale in the future.

FINDING 1: The demands for advanced computing continue to grow and will exceed the capabilities of planned upgrades across the NNSA laboratory complex, even accounting for the exascale system scheduled for 2023.

> **FINDING 1.1:** Future mission challenges, such as execution of integrated experiments, assessment of the effects of plutonium aging on the enduring stockpile, and facilitation of rapid design and development of new delivery modalities, will increase the importance of computation at and beyond the exascale level. Orders of magnitude improvement in application-level performance would allow for improved predictive capability, valuable exploration and iterative design processes, and improved confidence levels that will remain infeasible as long as a single hero calculation takes weeks to months to execute on an exascale system.
>
> **FINDING 1.2:** HPC has traditionally played an important role in support of weapons systems engineering. Some emerging challenges in this arena, such as qualifying future weapons systems for reentry environments, will require new approaches to mathematics, algorithms, software, and system design.
>
> **FINDING 1.3:** Assessments of margins and uncertainties for current weapons systems will require additional computational capability beyond exascale, a problem exacerbated by the aging of the stockpile. Enhanced computational capability will also be required in assessing margins and uncertainties should there emerge requirements for new military capabilities.
>
> **FINDING 1.4:** The rapidly evolving geopolitical situation reinforces the need for computing leadership as an important element of deterrence, and motivates increasing future computing capabilities.

DISRUPTIONS TO THE COMPUTING TECHNOLOGY ECOSYSTEM FOR STOCKPILE STEWARDSHIP (CHAPTER 2)

On the technology front, the single-thread performance of microprocessors continues to be relatively flat, and improvements in transistor density are slowing. Processing elements increasingly rely on more abundant, finer-grained parallelism and increasingly specialized hardware features that can improve performance by tailoring to a given computational domain. These trends are creating a significant disruption in available hardware components, with commercial interests focused largely on AI, embedded systems, and cloud services. There is considerable uncertainty as to whether the processors emerging in response to these trends, with their lower precision and limited high-speed memory, can be productively applied to NNSA applications.[3]

The market is also changing. The enormous scale of cloud-computing vendors now means that many cutting-edge hardware designs are being developed and deployed for in-house uses and are not available for direct purchase. Last, the number of HPC integrators has shrunk, placing the current system acquisition model at risk. In addition to the increased demand for HPC, technological and market shifts are challenging the vendor partnerships and other acquisition approaches used by the NNSA laboratories.

FINDING 2: The computing technology and commercial landscapes are shifting rapidly, requiring a change in NNSA's computing system procurement and deployment models.

> **FINDING 2.1:** Semiconductor manufacturing is now largely in the hands of offshore vendors who may experience supply-chain risk; U.S. sources are lagging.
>
> **FINDING 2.2:** All U.S. exascale systems are being produced by a single integrator, which introduces both technical and economic risks.
>
> **FINDING 2.3:** The joint ECP created a software stack for moving systems software and applications to exascale platforms, but although DOE has issued an initial call for proposals in 2023, there is not yet a plan to sustain it.

[3] In the early 1990s when the laboratories moved from parallel vector architectures such as those offered by Cray systems to massively parallel software based on MPI, the codes required massive rewrites. In the architectural transition anticipated by this report, it is likely that a similar investment in the software will be necessary to realize the potential of these new types of machines.

FINDING 2.4: Cloud providers are making significant investments in hardware and software innovation that are not aligned with NNSA requirements. The scale of these investments means that they have a much greater market influence than NNSA in terms of both technology and talent.

RECOMMENDATION 1: NNSA should develop and pursue new and aggressive comprehensive design, acquisition, and deployment strategies to yield computing systems matched to future mission needs. NNSA should document these strategies in a computing roadmap and have the roadmap reviewed by a blue-ribbon panel within a year after publication of this report and updated periodically thereafter.

RECOMMENDATION 1.1: The roadmap should lay out the case for future mission needs and associated computing requirements for both open and classified problems.

RECOMMENDATION 1.2: The roadmap should include any upfront research activities and how outcomes might affect later parts of the roadmap—for example, go/no-go decisions.

RECOMMENDATION 1.3: The roadmap should be explicit about traditional and nontraditional partnerships, including with commercial computing and cloud providers, and academia and government laboratories, and broader cross-government coordination, to ensure that NNSA has the influence and resources to develop and deploy the infrastructure needed to achieve mission success.

RECOMMENDATION 1.4: The roadmap should identify key government and laboratory leadership to develop and execute a unified organizational strategy.

RESEARCH AND DEVELOPMENT PRIORITIES (CHAPTER 3)

R&D activities have been a critical element of NNSA's Science-Based Stockpile Stewardship strategy. Historically, these activities have led to such achievements as the development of better mathematical models, numerical algorithms, parallel programming tools, and HPC operating systems.

Both AI and quantum computing have received significant and growing national attention and research investments in recent years. AI methods have revolutionized computational approaches in other disciplines, and NNSA has demonstrated their applicability in some limited domains while exploring the significant open questions limiting broader impact across NNSA's mission. At this time, NNSA can neither dismiss AI nor pivot entirely away from traditional modeling and simulation, and the most likely scenario is a complementary role for AI with simulation. Because of the uncertainty, AI research is critical, and the outcomes may influence post-exascale hardware roadmaps, industry partnerships, and applications capabilities. Quantum computing is also an exciting research area but is not sufficiently mature to be the basis for a computing strategy with a 20-year time horizon. Thus, while both are vital research areas, the future security of the nation's nuclear deterrent is too important to rely solely on either of these, as yet unproven, technologies for nuclear weapons development and assessment.

FINDING 3: Bold and sustained research and development investments in hardware, software, and algorithms—including higher-risk research activities to explore new approaches—are critical if NNSA is to meet its future mission needs.

> **FINDING 3.1:** Physics-based simulators will remain essential as the core of NNSA predictive simulation. However, given disruptions in computing technology and the HPC ecosystem combined with the end of the weak-scaling era, novel mathematical and computational science approaches will be needed to meet NNSA mission requirements.
>
> **FINDING 3.2:** Verification, validation, and uncertainty quantification (VVUQ) and trustworthiness remain of paramount importance to NNSA applications. VVUQ will become increasingly important as simulation methodology shifts toward more complex systems that incorporate models of different fidelity, including data-driven approaches.
>
> **FINDING 3.3:** Novel architectures can have a significant impact on NNSA computing; however, mathematical research will be needed to effectively exploit these new architectures. Involvement of applied mathematicians and computational scientists early in the development cycle for novel architectures will be important for reducing development time for these types of systems.
>
> **FINDING 3.4:** An end to transistor density scaling is likely to motivate industry to develop novel computer architectures for which today's numerical algorithms, software libraries, and programming models are ill suited.

FINDING 3.5: Recent advances in applied mathematics and computational science have the potential for impact on NNSA mission problems far beyond traditional roles in physics-based simulation.

FINDING 3.6: Co-design of hardware and systems for high-performance scientific computing applications has been a modest success to date and will be more important in the future and need to be deeper. Technological and market trends are likely to shift the balance of co-design to the laboratories, requiring more innovation and engineering in the areas of hardware design, system integration, and system software.

FINDING 3.7: Rapid innovation in AI methods, driven by advances in computing performance and growth in data sets, is producing frequent technological surprises that NNSA should continue to investigate and track. These advances may benefit the NNSA mission but will likely complement rather than replace traditional physics-based simulations in the post-exascale era.

FINDING 3.8: Quantum technology has the potential to improve the fundamental understanding of material properties needed by important NNSA applications. Analog quantum simulation or digital quantum simulation will likely be available before general quantum computers.

FINDING 3.9: Major breakthroughs in quantum algorithms and systems are needed to make quantum computing practical for multiphysics stockpile modeling. Quantum computing is more likely to serve as a special-purpose accelerator than to replace leading-edge computing.

RECOMMENDATION 2: NNSA should foster and pursue high-risk, high-reward research in applied mathematics, computer science, and computational science to cultivate radical innovation and ensure future intellectual leadership needed for its mission.

RECOMMENDATION 2.1: NNSA should strengthen efforts in applied mathematics and computational science R&D. Potential areas include using novel architectures, data-driven modeling, optimization, inverse problems, uncertainty quantification, reduced-order modeling, multiscale modeling, mathematical support for experiments, and digital twins.

RECOMMENDATION 2.2: NNSA should strengthen efforts in computer science research and development to build a substantial, sustained, and broad-based intramural research program that is positioned to address the technological challenges associated with post-exascale systems and co-design of those systems to ensure that the laboratories are positioned for leadership in computing breakthroughs relevant to NNSA mission problems.

RECOMMENDATION 2.3: NNSA should expand research in artificial intelligence to explore the use of these methods both for predictive science and for emerging applications, such as manufacturing and control of experiments, and develop machine learning techniques that provide the confidence in results required for NNSA applications.

RECOMMENDATION 2.4: NNSA should continue to invest in and track quantum computing research and development for future integration into its computational toolkit; these technologies should be considered an additional computational tool rather than a replacement for current approaches.

WORKFORCE NEEDS (CHAPTER 4)

Perhaps the most significant challenge facing the NNSA laboratory complex is in attracting and retaining top talent in areas that overlap and compete with the computing industry. Today's computing technology and services industry offer higher salaries, greater resources, more flexible work environments, and the ability to focus on compelling intellectual opportunities. Meanwhile, security concerns affect recruitment of foreign talent, both directly limiting NNSA hiring and indirectly discouraging international participation in the broader U.S. computing ecosystem.

Workforce areas of need include experts in computer hardware and performance optimization, algorithms and applied mathematics, physics modeling, numerical computations, and software development. Even more challenging are emerging areas of scientific computing such as machine learning, hardware co-design, and quantum information science. NNSA's ASC program and DOE's ECP program have been unique resources for addressing national priorities, with teams of world experts that deploy large-scale computing for complex analysis and prediction, but to continue in this role NNSA and the associated national laboratories need to attract and train talent from an increasingly diverse workforce, offer competitive compensation packages, and provide an intellectually exciting and stable environment in which to work on cutting-edge R&D problems.

FINDING 4: NNSA's laboratories face significant challenges in recruiting and retaining the highly creative workforce that NNSA needs, owing to competition from industry, a shrinking talent pipeline, and challenges in hiring diverse and international talent.

FINDING 4.1: The ASC program currently faces a challenge maintaining a competitive workforce. This challenge will continue to grow because of pipeline issues (small number of U.S. citizens going into graduate-level science, technology, engineering, and mathematics fields), industry competition, and emerging computing talent choosing not to focus on scientific computing.

FINDING 4.2: The U.S. national security enterprise has benefited enormously from the inclusion of global talent, but incorporating international scholars in the NNSA community is challenged by important concerns about protecting sensitive information. Failure to balance these risks with the risk of missing the best talent can result in not finding the best candidates for the job.

FINDING 4.3: Addressing the challenges laid out in this report will require a nurturing environment that reduces distractions, funding uncertainty, and administrative burdens, while providing employees the time and flexibility to explore areas of interest and do the creative thinking required to solve these problems.

RECOMMENDATION 3: NNSA should develop an aggressive national strategy through partnership across agencies and academia to address its workforce challenge.

RECOMMENDATION 3.1: NNSA should make concerted efforts to create an environment that nurtures and retains existing staff; more aggressively grow the pipeline; create an efficient and modern, yet secure environment; advertise and grow existing workforce programs (such as the Predictive Science Academic Alliance Program and the Computational Science Graduate Fellowship); and collaborate with other federal agencies to support ambitious talent development programs at all career stages.

RECOMMENDATION 3.2: NNSA should also develop a deliberate strategy to attract an international workforce and to provide them with a welcoming environment while thoughtfully managing the attendant national security risks.

1

Role and Importance of Exascale and Post-Exascale Computing for Stockpile Stewardship

This chapter describes the role of advanced computation in support of stewardship of the current stockpile. Computation has been an essential part of the nuclear weapons mission since its inception. For example, the Manhattan project relied on computational capabilities available at the time to determine that a weapon was feasible in terms of size and weight.

The sections below provide a description of some of the key phenomenology that must be simulated both for nuclear weapon operation and for some of the associated engineering requirements. Although it is impossible to be complete in this regard, the committee's objective is to provide some context for why such problems are so challenging and also to lay the groundwork for why continued growth in computational capability will be required even beyond the advent of exascale systems. A question that is often asked of the nuclear design laboratories is "What are the resolution and numerical throughput (also known as memory size and floating-point speed) requirements for accurate simulation of nuclear weapon performance?" A more colloquial way to ask this is "How much is enough?" The answer to this question depends on the end goal of the simulation. If one's objective is to simulate accurately all the relevant physical length and time scales (so-called direct numerical simulation), such a computation is currently out of reach. On the other hand, if one's objective is to explore the predictions of various modeling assumptions and then make comparisons with experiments, such calculations are performed routinely today.

The chapter begins by describing the role of computing in the early days of the weapons program, when nuclear explosive testing was possible and then describes how the role of computing has evolved. This discussion is followed by a more detailed description of the relevant physical processes. A discussion below on the modeling of high explosives used to initiate nuclear detonation is more detailed than the others for two reasons. First, many of the details of high-explosive modeling are unclassified and in the open literature. Second, the use of various levels of modeling to simulate high explosives mirrors some of the considerations that are relevant to present-day modeling of nuclear weapons—namely, the trade-off of modeling sophistication versus the need to explore efficiently a large design space.

A nuclear weapon is an energy amplifier that uses the chemical energy of high explosives to initiate nuclear fission. To give some idea of the level of amplification and to provide some perspective for the challenge of simulation, if completely exploded, a kilogram of conventional TNT chemical explosive provides about 4.6 megajoules of energy. In contrast, if completely fissioned, a kilogram of ^{239}Pu provides 88 terajoules of energy, a factor of roughly 20 million over TNT and chemical explosives in general. An overview from Lawrence Livermore National Laboratory (LLNL)[1] describes how energy is released by nuclear weapons. This level and rate of energy release makes overall system behavior sensitive to details of the component processes. This translates into computational requirements for simulation that are far more stringent in terms of resolution of length and time scales than those encountered for example in the simulation of conventional weapons.

Early in the history of nuclear weapons development, it was understood that the underlying physics of nuclear weapons were sufficiently complex that ab initio computation was not possible, for reasons further discussed below. Instead, computation could be used to provide feasibility estimates for proposed designs, with confirmatory experiments including underground nuclear testing essential to prove the workability of those designs. Phenomenological data from such experiments were used to develop calibrated models that improve the predictive capability of computation, provided consideration was limited to small excursions from the tested design. This approach of computation combined with underground testing was highly successful in developing and ensuring the robustness and reliability of today's modern stockpile.[2] Given that the current stockpile has been successfully developed and certified, it is reasonable to ask if there remains a need for additional computing capability. At present, however, the computational tools essential in designing the current stockpile continue to rely heavily on calibration with underground test data. The resulting methods are not predictive outside the range of the data in which they were calibrated and are viewed as possibly invalid if there is a change

[1] B.T. Goodwin, 2021, "Nuclear Weapons Technology 101 for Policy Wonks," Center for Global Security Research, Lawrence Livermore National Laboratory, https://cgsr.llnl.gov/content/assets/docs/CGSR_NW101_Policy_Wonks_WEB_210827.pdf.

[2] N. Lewis, 2021, "Trinity by the Numbers: The Computing Effort That Made Trinity Possible," *Nuclear Technology* 207(Sup 1):S176–S189, https://doi.org/10.1080/00295450.2021.1938487.

in design, manufacturing processes, or even the age of the weapon. In the absence of underground tests, high-fidelity simulation becomes a necessity in addressing significant changes in the stockpile.[3]

The entry of the United States into a voluntary moratorium on nuclear testing in 1992 led to the development of Science-Based Stockpile Stewardship (SBSS) and elevated the importance of high-performance computing (HPC) both in terms of ensuring the safety and reliability of the existing stockpile as well as providing scientific support should modifications be required. To understand the challenge presented by this HPC application, it should be noted that the cursory description of the operation of a nuclear weapon provided above does not do justice to the precision and timing required to achieve the staged amplification of energy. The length and time scales associated with the functioning of the weapon become increasingly short. As discussed below, the release of energy from explosives involves chemical reactions among molecules. In contrast, fission and fusion occur on length and time scales associated with the nucleus, which is more than 10,000 times smaller. Because of the enormous amplification of energy associated with nuclear reactions, small perturbations in any of the early stages can lead to large perturbations of the energy output or yield of the weapon. These considerations make it necessary to provide enhanced resolution of space and time scales to faithfully capture the dynamics via computation. While the relevant physical processes are all founded in established theory, the multiscale, coupled nature of the problem as well as the large range of relevant length and time scales have made computation—and, in particular, HPC—essential to making existing simulation codes more predictive even in the absence of supporting data from nuclear tests. This enhanced level of predictive capability is needed for several reasons:

- Early weapons designs employed computation, but these calculations assumed either simplified one-dimensional symmetry or two-dimensional axisymmetry, owing to limitations of computational speed and memory size at that time. Such calculations were not predictive and had to be calibrated via results from underground nuclear testing with the details of calibration dependent on the specifics of individual weapons designs. Once calibrated, these early weapons codes could be used to study the effect of small design changes, but they could not be used to explore changes considered significantly different from what was explored in previous nuclear testing. For this reason, future consideration of any new design concepts or changes to existing systems for which there is little or no testing history requires a more

[3] R.J. Hemley and D.I. Meiron, 2011, "Hydrodynamic and Nuclear Experiments," *JASON Defense Advisory Panel: Reports on Defense Science and Technology*, https://irp.fas.org/agency/dod/jason/hydro.pdf.

predictive capability, which in turn requires increased spatial and temporal resolution of the relevant phenomena and thus increased computing capability and capacity.

- It has long been appreciated that nuclear weapons are inherently three-dimensional both in geometric detail and in operation. Despite efforts to create as symmetric a geometric environment as possible, there are engineering features that are essential to the operation and delivery of the weapon, and these are inherently three-dimensional. In addition, the basic physical processes involved in the implosion of the primary and secondary possess intrinsic instabilities that can make the resulting weapon response diverge from the desired axisymmetry if not properly controlled.
- As part of the ongoing stewardship of the existing stockpile, every effort is made to ensure that refurbishment of the various systems—for example, those undergoing a life extension program (LEP)—use materials and components that are as close to identical as possible to those used in the original design. However, this has not always proven possible. First, some of the original materials are no longer manufactured. Second, manufacturing approaches evolve, and so it may not be practical to reestablish various component production lines. Last, the materials used often age with time (a notable example is plutonium, discussed in more detail below). Thus, the stockpile inevitably evolves with age, and both experiment and computation are required to make sure that a given weapons system has not unduly diverged from its original certified design.
- Although much progress has been made in understanding the basic physical processes that govern a nuclear weapon, there remain phenomena central to the operation of modern nuclear weapons for which understanding remains incomplete. The computational resources required for full numerical resolution of these physical processes typically exceed exascale capability. An example would be the detailed three-dimensional calculation of the implosion of an inertial confinement fusion capsule, taking into account the grain structure of the shell bounding the deuterium-tritium gas mixture. Such calculations can only be practicably performed today in an axisymmetric geometry.[4] To develop engineering models that facilitate rapid iteration of design calculations, high-fidelity simulation that captures as many physical scales as can be resolved given the limitations of computer speed and memory are used to generate data to inform reduced-order models.

[4] B.M. Haines, R.E. Olson, W. Sweet, S.A. Yi, A.B. Zylstra, P.A. Bradley, F. Elsner, et al., 2019, "Robustness to Hydrodynamic Instabilities in Indirectly Driven Layered Capsule Implosions," *Physics of Plasmas* 26:012707, https://doi.org/10.1063/1.5080262.

- In the absence of underground testing, it is necessary to acquire some relevant information via computation only. Understanding the relevant phenomena so that they can be predictively modeled requires a combination of focused experiments combined with HPC at the level of exascale and beyond.
- In addition to simulation of fundamental aspects of weapons operation, HPC plays an important role in simulation of nonnuclear components of a weapon in various environments that are encountered during the weapon's life cycle. This includes weapon response in normal, abnormal, and most challenging, hostile environments. This aspect of simulation also requires HPC capability.[5]

MODELING AND SIMULATION REQUIREMENTS FOR STOCKPILE STEWARDSHIP

Detonation of High Explosives

As indicated above, a nuclear weapon is effectively an energy amplifier wherein the energy released from detonation of a high-explosive charge is used to compress and initiate a fissile core and begin the nuclear chain reactions that drive further amplification stages. As an exemplar of the requirement for high-fidelity simulations needed to support the stockpile, the committee discusses here in detail some of the extant computational challenges of modeling and simulating the detonation of high explosives and why today such simulation requires state-of-the-art computational capability. In what follows, the committee provides a more detailed discussion of some of the issues as these are unclassified. In contrast, the committee provides more terse descriptions of later stages of weapons operation owing to classification restrictions.

Detonation refers to a particularly rapid form of combustion in which the transfer of energy occurs via a strong shock wave supported by an exothermic chemical reaction zone. The leading shock of a detonation wave compresses and heats the reactant material, thus initiating a chemical reaction that further drives the shock wave. Under the right conditions, this process is self-sustaining as a wave that propagates across the high-explosive fuel. The conversion of chemical to mechanical energy as the high explosive detonates is extremely rapid and intense. A solid explosive can convert energy at a rate of 10^{10} watts per square centimeter at the detonation front, or 1 MW per square meter. For comparison, insolation of Earth by the Sun provides 1 kW per square meter.[6]

The rapid release of chemical energy leads to strongly supersonic waves propagating at 8–10 kilometers/second that provide large pressures, on the order of tens to

[5] Briefings by LANL, LLNL, Sandia to National Academies panel, March 21, 2022.

[6] W. Fickett and C. Davis, 2001, "Detonation: Theory and Experiment," *Journal of Fluid Mechanics* 444:408–411, https://doi.org/10.1017/S0022112001265604.

hundreds of kilobars. Such pressures are well above the compressive strength of metals, and so high explosives are a natural candidate for rapid compression of the plutonium of a primary to a supercritical state.

High explosives are today used in mining, conventional weapons, and sometimes for materials applications like formation of metallic solids from metal powders. Precise characterization of the speed and energy released, or the detailed structure of the detonation wave, are not of top concern in such applications. For nuclear applications, however, understanding the structure and dynamics of the detonation, particularly its velocity and delivered energy, is crucial to characterizing the subsequent implosion of the nuclear material. The high-explosive charge of a modern nuclear weapon is carefully engineered to create a robust implosion that will rapidly assemble a supercritical mass.

For many applications other than nuclear weapons, it is appropriate to assume that the region wherein the reactants are compressed, react, and transition to products is adequately described by a plane steady reaction zone. This assumption is adequate for detonations where the radius of curvature of the detonation wave greatly exceeds the thickness of the reaction zone. If the reaction zone is assumed to be infinitesimally thin and the chemical reaction proceeds to completion instantaneously, the detonation wave can be treated as a discontinuity moving with a specified detonation velocity and the conservation laws of compressible fluid flow can be applied to determine its evolution. Once the detonation velocity is measured, the laws of conservation of mass, momentum, and energy determine the conditions behind the shock wave. In compressible flow without exothermic chemical reactions, shock waves of all strengths are possible, with the minimum velocity of the shock wave being the local sound velocity. When the heat release of an exothermic reaction is introduced, the shock wave must have a minimum velocity that is strongly supersonic relative to the ambient sound speed. This is the basis of the Chapman-Jouguet (CJ) theory of detonation. This minimum velocity is called the CJ velocity and the associated pressure, the CJ pressure. Unfortunately, the simplicity of the CJ theory is not adequate for the tight timing requirements of a nuclear weapon. The CJ theory when applied to detonation in gases is often correct to within 1–2 percent, but later experiments showed that other quantities such as the pressure at the CJ state were off by 10–15 percent; the assumption of a simple one-dimensional discontinuous solution is overly simplistic. The CJ theory is attractive owing to its simplicity; the computational cost to characterize the detonation wave is low, but it is not sufficient for the timing accuracy required, and more elaborate models are required.

A significant improvement in understanding was made independently by Zeldovich, von Neumann, and Döring (ZND).[7] They based their work on the one-dimensional,

[7] W. Döring, 1943, "Über Detonationsvorgang in Gasen" [On detonation processes in gases], *Annalen der Physik* (in German) 43(6–7):421–436.

inviscid, compressible Euler equations of hydrodynamics. The generally complex set of reacting species is modeled by one component and transitions from reactant to product at a finite rate. A detonation profile here still has a discontinuous shock wave, and so the conservation laws must still be used to connect the reactant state to the product state, but the profiles of pressure, density, velocity, and so on are now self-consistent solutions to one-dimensional partial differential equations. With the development of improved computational capabilities, it was found that the simple CJ state was inadequate in describing detonation dynamics. Instead, the detonation front had a complicated time-dependent fine structure that wandered about the CJ values.

As computational capabilities increased further, the one-dimensional theory was extended to include modeling of chemical reactions with multiple reaction rates, heat conduction, viscosity, and diffusion, making it possible to consider two-dimensional calculations and finite size effects such as curvature of the detonation wave and the effect of boundaries. One of the important conclusions of these studies was that with the addition of more physics, the deficiencies of the CJ and ZND models became more apparent, and the detonation velocity was now seen to be a complex function of the explosive material, including the reaction rates, which in many cases are not well known for condensed explosives at high pressure.

For gas-phase detonations, improved experimental diagnostics have shown that considerable three-dimensional fluid motion occurs behind the detonation front. For example, it is known that the leading shock wave of the detonation exhibits instability in the form of transverse shock waves running along the shock surface. Complex motions including spiral waves have also been observed, as well as galloping of the detonation surface. Over time, a qualitative view has emerged of the detonation front as a complex surface with propagating cellular perturbations proportional to the reaction time and so dependent on local conditions as the explosive reacts and detonates. Curvature of shocks leads to the generation of vorticity, the tendency of the flow to exhibit local differential rotation. In gas phase detonations, this fluid rotation leads to what appears to be turbulent dynamics.[8]

Much of the research that led to the results described above was based on detonation in gaseous media. For several reasons, nuclear weapons make use of solid explosives. First, such explosives have higher energy density and so can provide higher pressures for compression of the metal shell. Second, they can readily be molded around the nuclear core of the primary. These advantages, however, are offset by the fact that modeling and simulation of such explosives is significantly more complex than that required for gas-phase explosives. Solid explosives are complex materials consisting

[8] W. Fickett and C. Davis, 2001, "Detonation: Theory and Experiment," *Journal of Fluid Mechanics* 444:408–411, https://doi.org/10.1017/S0022112001265604.

of crystallites of high-explosive molecules that are held together by polymeric binder. Perhaps of more relevance to the requirements for computation, these materials, unlike gases, are opaque, and so it is more challenging to gather experimental data on the internal structure of the detonation to inform phenomenological models.

Numerical simulation of high explosives today remains challenging. First, there is a profusion of length and time scales that must be considered. For a conventional high explosive, the reaction zone thickness is typically on the order of tens of microns. This alone tells us that the ratio of the length scales of the gross dimension of the primary to the reaction zone is on the order of 10,000 to 1. Models for the chemical reactions of the combustion typically involve hundreds of species and thousands of possible reactions. In most cases, the reaction rates must be obtained from molecular dynamics calculations, but these too are challenging to compute because they involve simulation of chemical reactions at high pressures. In addition, the reactions must be understood across several phases: the solid phase of the explosive crystallites and the liquid and gaseous phase of the reactants and products. The time scales for many of these reactions is on the order of picoseconds. This places a severe restriction on the required spatial and temporal resolution needed to obtain resolved results. An example of the challenge is the pioneering work of Baer on the direct numerical simulation of solid explosives.[9] Using the Eulerian shock physics code CTH,[10] Baer performed simulations of the high-explosive cyclotetramethylene-tetranitramine (also known as HMX) on the state-of-the-art Advanced Simulation and Computing (ASC) platform at that time, the Accelerated Strategic Computing Initiative Red platform. The initial condition for this simulation was a small piece of explosive (about 1 mm^3) resolved at the mesoscale. That is, the explosive crystallites and the polymer binder are resolved at a scale of roughly 100–200 microns. The goal was to study the interaction of the crystallites as they were impacted by a strong shock wave ultimately leading to a detonation wave. This calculation, justifiably viewed as an important contribution to detonation science, required on the order of a billion grid cells and led to some important insights on the micromechanics of detonation.

Owing to memory restrictions, however, even this calculation did not include an accurate model of the crystalline deformation, friction, or void heating—effects that are known to play a role in the formation and propagation of detonations in solid explosives. In addition, only a relatively crude model of the combustion was used. Last, an important aspect of detonation science is the propagation of the detonation over curved geometry. To examine this, the calculation would need to be performed using samples

[9] M.R. Baer, 2001, "Computational Modeling of Heterogeneous Reactive Materials at the Mesoscale," *AIP Conference Proceedings* 505:27, https://doi.org/10.1063/1.1303415.

[10] E.S. Hertel, R.L. Bell, M.G. Elrick, A.V. Farnsworth, G.I. Kerley, J.M. McGlaun, S.V. Petney, S.A. Silling, P.A. Taylor, and L. Yarrington, 1995, "CTH: A Software Family for Multi-Dimensional Shock Physics Analysis," 377–382, Springer: Berlin/Heidelberg, https://doi.org/10.1007/978-3-642-78829-1_6.

of dimensions on the order of centimeters. If the resulting fluid mechanics behind the detonation wave is truly in the turbulent regime, it becomes necessary to resolve down to the micron scale in order to directly resolve the effects. Improving the quality of the solid mechanics and chemistry further increases the computational volume while also requiring a reduced time step to capture the details of wave propagation. Given the large number of degrees of freedom associated with fully resolved detonation simulation, detailed calculations at the scale of the weapon that include relevant geometrical effects could require resources exceeding those provided by exascale platforms. Indeed, this level of simulation of high-explosive detonation is not practical today even though it is agreed that understanding solid explosives at this level of detail would be of benefit to SBSS. In discussion of these issues with the design laboratories, it is acknowledged that

> Using high-fidelity reactive flow with chemical kinetics to model High Explosive (HE) burn requires extremely high resolution to resolve chemistry at the burn front and its inter-relationship with hydrodynamics. In general 3D [three-dimensional] would require billions to trillions of hydro cells/zones in the HE to be modeled predictively. AMR [adaptive mesh refinement] is routinely used to dynamically refine and de-refine the burn front thereby saving memory and other computational resources.[11]

Adaptive methods (in both the space and time dimensions) that focus resources on the regions about the detonation wave have been developed as indicated above, but efficient implementation of such methods on modern high-performance architectures is complicated by the fact that today's HPC platforms are not optimally suited to the irregular data access patterns that result from using such methods. If it should prove necessary to resolve fine-scale multispecies turbulence behind the detonation front, the resolution requirements become more demanding. As a result, simulation and experiment are used to create calibrated engineering models for high-explosive detonation, and it is these models that are used in simulation of full systems.

To improve nuclear safety, there are today increased efforts to understand the detonation of insensitive high explosives. Such explosives have lower CJ velocities and pressures so that more of the explosive is needed to implode the primary. On the other hand, these explosives are much more difficult to initiate (hence the name insensitive) and so are preferred for nuclear safety. There is a desire within the nuclear complex to eventually use only insensitive high explosives in the nuclear stockpile, but their lack of sensitivity paradoxically makes their simulation more complex, as the combustion processes are more complex.

[11] From summary of computational requirements provided to the committee by LANL, LLNL, and SNL ASC teams.

For engineering purposes, in which a designer wants to rapidly examine possible changes in the configuration of the explosive, it is not currently practical to use highly resolved computations to design and qualify high explosives. First, as discussed above, the material and chemical response of the initiation of solid explosives leading to detonation is still not completely understood. Second, the range of scales that must be resolved, as also discussed above, makes routine calculations infeasible. Instead, engineering models are developed that parameterize the results in various regions of interest. Such models can then be used to explore some of the design space. The development of such models continues to rely on experimental measurement and, to an increasing degree, detailed computation with the relevant computations, likely requiring exascale computational capability and beyond to properly capture the phenomena of interest.

Today, when a manufactured lot of high explosive is delivered to the nuclear weapons complex, it undergoes extensive testing to confirm that it meets the required specifications. Once an acceptable formulation is found, significant work is required to ensure that it can be manufactured reliably. HPC has been essential in this qualification process. But as it stands today, predictive simulation of high explosives is not feasible even with exascale computing capabilities.

The overall conclusion is that even for the initial phase of nuclear weapon operation, the detonation of solid high explosive, there is still no complete predictive understanding of detonation that would make it possible to specify material properties of the explosive such as crystallite volume fraction, binder composition, and so on that would allow for the ab initio design of a high explosive with a predictable detonation propagation speed that satisfies the very tight timing requirements associated with the initiation of a weapon. As it stands today, predictive simulation of high explosives is not feasible even with exascale computing capabilities. This remains an area of active research. As discussed below, the situation is quite similar as regards the other stages of nuclear weapon operation.

Dynamic Response of Materials Under Extreme Conditions

The second phase of the energy amplification process is the implosion of the fissile material by the high-explosive charge. As the plutonium shell is compressed, knowledge of how various thermodynamic properties, such as density and internal energy, vary with pressure and temperature is essential. This information is provided in the equation of state. While the equation of state has been determined analytically for some simple materials (e.g., gases at low to moderate pressures), for more complex materials, experiments and computation are combined to provide tabular results.

Plutonium is a very complex metal. At ambient pressure, as the temperature is raised toward its melting point, it undergoes six structural phase transitions corresponding to different crystalline symmetries.[12] At the very high pressures achieved in nuclear explosions, the equation of state has mostly been determined through experiment. More recently, HPC has been used to explore the equation of state at pressures and temperatures that can be achieved only in an underground nuclear explosion. Such calculations require the approximate quantum mechanical solution of the many-body problem of electronic structure under high pressure and temperature. For metals such as plutonium, this is particularly challenging because plutonium has 94 electrons, with significant correlations among the electrons occupying the various orbital shells. This is partly responsible for its rich crystalline morphologies.

The equation of state refers to the thermodynamic properties at equilibrium. Under conditions of rapid implosion, it is also necessary to consider the dynamic material strength of a metal that describes irreversible processes such as permanent deformation under applied stress. The physical basis for material strength and failure of metals has been understood for some time and is determined by the motion of imperfections (dislocations and so on) in the crystalline lattices of the crystallites comprising a metal. These propagate through the crystalline lattice under applied stress on length scales of microns. As in the case of high explosives, the dynamics have a multiscale character, but translating this basic physics into a model that can be applied at macroscopic scales has proven to be challenging. HPC has been applied productively here as well, and progress has been made in improving what have been in the past largely calibrated models of such phenomena, but a completely satisfactory model that can be efficiently used in design calculation is not yet available.[13]

A natural question to ask is whether it is necessary to know the details of the response of the plutonium under the high pressures of implosion with such accuracy. Again, because of the strong amplification of energy inherent in the nuclear explosion process, it turns out that the results are quite sensitive to both the equilibrium and dynamic properties of the plutonium. In particular, the rate of fissioning that governs the output of the weapon is very sensitive to these properties, and so accurate knowledge is essential to predict reliably the performance of the weapon.

Hydrodynamics

The ability to simulate hydrodynamics is essential to the understanding and prediction of nuclear weapon operation. By hydrodynamics, the committee means the transport of mass, momentum, energy, and all component species in the presence of internal and external stresses. The equations of hydrodynamics, properly modified for the operative

[12] S. Hecker, 2000, “Plutonium and Its Alloys: From Atoms to Microstructure,” *Los Alamos Science—Challenges in Plutonium Sciences* 26.

[13] Ibid.

physics, are used to model the detonation of the explosive, the implosion and explosion of the primary, and the subsequent dynamics of the secondary. The flow in all these cases is highly compressible, meaning that under the pressure forces experienced by the various media, significant density changes will occur. For example, as the primary implodes, the resulting flows exhibit velocities well in excess of the local speed of sound in the material. This is known as the hypervelocity regime of hydrodynamics, characterized by the formation of strong shock waves. Resolution of these waves using modern numerical codes requires tracking the evolution of a significant number of degrees of freedom, as such waves are marked by rapid variation of the fluid properties over very narrow and evolving spatial regions. It is not practical to directly simulate these narrow regions. Instead, modelers use various types of regularizations to compute large-scale motions while ignoring the smaller scales in the rapid transition regions. Flows in this high-velocity regime are also prone to development of turbulence in which the flow becomes chaotic. Again, it is currently not practical to numerically resolve such small scales, and so these too are modeled, often requiring significant calibration of the models via experiments and previous underground testing. While modern highly resolved calculations today track billions of degrees of freedom in simulating hydrodynamics, it has long been appreciated that developing a deeper understanding of such hydrodynamic flows requires resolution that requires exascale capabilities and beyond and that such computational capability is essential to developing models that are less reliant on phenomenological calibration.

Neutron and Radiation Transport

Simulation of both neutron and X-ray transport is essential to modeling of both fission and fusion processes. Typically, this is among the most computationally expensive parts of the simulation of weapon operation simply because of the number of degrees of freedom involved. The density, momentum, and energy of the imploding material are tracked computationally at each physical point as discussed above when simulating hydrodynamics. But in addition, it is necessary to track at each point the neutron angular distribution and energy spectrum. This adds an additional three degrees of freedom at each spatial position, leading to computation tracking six dimensions plus time and therefore a large factor to the operation count. Note too that such calculations require input to the transport equation such as the neutron cross sections for fission, absorption, and scattering, and these must be determined through theory or experiment.[14]

[14] A. Sood, R.A. Forster, B.J. Archer, and R.C. Little, 2021, "Neutronics Calculation Advances at Los Alamos: Manhattan Project to Monte Carlo," *Nuclear Technology* 207(Sup 1):S100–S133, https://doi.org/10.1080/00295450.2021.1956255.

A similar requirement arises for the secondary. The fluence and energy of X rays generated by the primary as it explodes are large enough to implode the secondary. Simulating the propagation and evolution of these X-ray photons requires the application of the equations of radiation hydrodynamics. The challenges here are similar in complexity to those encountered in neutron transport, although the details of the governing equations are quite different. Here, the relevant material properties are the opacities of materials as a function of X-ray energy, and these must be measured or computed using calculations of atomic transitions. Again, such transport calculations are significantly more computationally expensive than those associated with basic hydrodynamics.[15]

Quantification of Margins and Uncertainties

The National Nuclear Security Administration (NNSA) laboratories use quantification of margins and uncertainties (QMU) as a means of describing the overall reliability and robustness of weapons systems. As discussed above, the operation of a weapon proceeds in stages. Each stage is meant to create a set of environments that then make possible the next stage of operation. Detonation of the high explosive must create a compressive environment leading to the fission of the primary. The output of the primary must provide sufficient energy to drive the secondary. To ensure that the required conditions occur in a robust way, weapon designers engineer a certain amount of margin in the design so that even if uncertainty in the required processes leads to some degradation of the required environment, conditions are still sufficient to drive the next stage. To confirm that sufficient margins exist, the maximum and minimum of the operating ranges for the various stages of operation must be established. If there is insufficient margin, the weapon system may not behave in a predictable way, implying that there is a region of parameter space, known as a performance cliff, in which the next stage of operation cannot be achieved or becomes unreliable.

Assessments of where in parameter space such cliffs are located are by nature uncertain, and this uncertainty must also be estimated. The ratio of the assessed margin to uncertainty is a confidence factor, and the research done to determine margins and uncertainties is used to create an evidence file that is reviewed whenever the health of the stockpile is assessed. Margins and uncertainties can be assessed for all aspects of weapon function, and such assessments are performed not only for the nuclear explosive package (NEP) containing the primary and secondary, but also for all the supporting nonnuclear components that must function reliably and precisely. QMU is used today as a basis for evaluating all aspects of what is known as the stockpile to target sequence. Margins (and uncertainties) are not static. Components of the stockpile have undergone

[15] M. Dimitri and B. Weibel-Mihalas, 1984, *Foundations of Radiation Hydrodynamics*, New York: Oxford University Press.

refurbishment via LEPs. The materials comprising the stockpile age. To ensure that the weapons will operate as designed, ongoing assessments of margins and uncertainties are required. All recent assessments indicate that current stockpile systems have high-confidence factors, but, in the absence of underground testing, it is important to continually assess the stockpile using QMU.

To evaluate margins and uncertainties, ensemble analyses of weapon function as various key parameters are varied must be performed. Some of the required data are available from past nuclear tests, but such tests did not always comprehensively cover the variations of key parameters. Computation today plays an essential role in further assessing margins and uncertainties. It could be argued that such analyses require capacity computing to create the ensemble of calculations needed. But the challenge here is that the regions of parameter space where the various models of physical processes provide accurate results are not completely known. Underground testing and other experimental data explore some small regions of this parameter space, but the regions of validity of the various models (i.e., the failure boundaries) are presently not well understood. Exascale computing and beyond, combined with future basic experiments, is required to further explore this space and provide improved assessments of margins and, even more importantly, uncertainties.[16]

Engineering Challenges

The committee has up to now focused on the simulation of the NEP, indicating the computational challenges. There are also computational challenges associated with the nonnuclear components. These range, for example, from electronic components like the arming, fuzing, and firing system that controls the firing of the high explosives, to structural components that ensure the NEP survives the rigors of delivery. These must be qualified in normal delivery environments, and their response must also be characterized in abnormal environments such as those arising in case of an accident. While it might be assumed that modeling and simulation for these components should already be well in hand, there remain challenges for which exascale computing will aid future qualification.

First, the weapons system, be it a warhead or bomb, is subject to mechanical and thermal stresses during delivery. This is particularly true for reentry systems, as discussed below. Flight testing has been, and is still, used to confirm that the weapons systems can survive the reentry environment and accurately deliver the NEP to a specified target. However, such testing is expensive and so is not performed often. Simulation of the delivery environment is then required along with the use of QMU to provide additional

[16] National Research Council, 2009, *Evaluation of Quantification of Margins and Uncertainties Methodology for Assessing and Certifying the Reliability of the Nuclear Stockpile*, Washington, DC: The National Academies Press, https://doi.org/10.17226/12531.

input to qualification. Simulation of reentry environments remains an exascale challenge and will most likely require capability beyond exascale, as will be discussed later in this chapter. Second, a bomb or warhead is geometrically quite complex, with many internal parts that are used to provide mechanical and thermal stability for the NEP as the weapon encounters the stresses of the delivery environment. Simulation is required to understand the mechanical and thermal response. In addition, the dissipation of mechanical energy by joints, bolts, and so on undergoing vibration is today still largely modeled phenomenologically. Third, the weapon must fail safely in the event of an accident such as a fire or collision. Here, testing might provide confirmation of safety in only a few specific scenarios. Simulation is required to cover a range of possible thermal and mechanical loadings. The results of such simulations are often surprising, indicating unexpected failure scenarios. By simulating a wide range of accidents, QMU can then be used to quantify the likelihood that the weapon will fail gracefully. Last, perhaps the most challenging requirement is to quantify the survivability of a weapon in a hostile nuclear encounter. This requires assessing mechanical, thermal, as well as electrical insults to the weapon. Computation is essential here, as it is very difficult to fully replicate this environment experimentally.

COMPUTATIONAL REQUIREMENTS AND WORKLOAD FOR ASC APPLICATIONS

All three nuclear design laboratories, Los Alamos National Laboratory (LANL), Lawrence Livermore National Laboratory (LLNL), and Sandia National Laboratories (SNL), provided the committee with summaries of the computational requirements for their applications. It is important to note that the results provided below represent only a snapshot of typical workloads. For example, for the hydrodynamics calculations, the requirements depend on the number or materials in each computational cell, the level of refinement, the level of physics being described, and so on. For radiation transport using deterministic methods, the computational requirements will depend on the number of energy groups, angles, materials as well as material states, and so forth. LANL applications are characterized in Table 1-1.

As can be seen, the requirements per cell can vary significantly and will depend on the level of fidelity desired. The data access patterns also vary. Regular access patterns make it possible to benefit from memory caching strategies, whereas irregular patterns are harder to optimize. As will be seen in Chapter 2, efficient memory access can lead to performance enhancements. The corresponding characterization for LLNL is displayed in Table 1-2.

TABLE 1-1 Characteristics of Different Hydrodynamics and Transport Applications for Los Alamos National Laboratory

Typical Characteristics	Hydrodynamics (Hydro 1)	Hydrodynamics (Hydro 2)	Deterministic Transport	Monte Carlo Transportation
Memory needs	100–400 KB/cell	1–2 KB/side (sides is better metric of complexity)	128 KB/zone (highly dependent on number of groups and angles)	6–25 KB/zone
Access pattern	Irregular, with low to moderate spatial and temporal locality	Irregular, with low spatial and temporal locality	Regular, with high spatial and moderate temporal locality	Irregular, with low spatial and temporal locality
Communication pattern	Surface communication, point-to-point, and RMA for AMR and load balancing some collectives	Surface communication, point-to-point, some collectives	Point-to-point, with some global reductions	All reduce, point-to-point, some surface communication
MFLOPS	0.033/cell/cycle	0.0034/side/cycle	0.7/zone/cycle	0.083/zone/cycle 0.07671/particle/cycle

NOTE: AMR, adaptive mesh refinement; MFLOPS, million floating-point operations per second; RMA, random memory access.
SOURCE: From summary of computational requirements provided to the committee by LANL, LLNL, and SNL ASC teams.

TABLE 1-2 Characteristics of Different Hydrodynamics and Transport Applications for Lawrence Livermore National Laboratory (LLNL)

Typical Characteristics	Hydrodynamics	Deterministic Transport	Monte Carlo Transport	Diffusion
Memory needs	0.1–1 KB/zone	40–240 KB/zone	3–30 KB/zone	0.1–1 KB/zone
Access pattern	Regular, with modest spatial and temporal locality	Regular, low spatial but high temporal locality	Irregular, low spatial and temporal locality	Regular, good spatial and temporal locality
Communication pattern	Point-to-point, surface communication	Point-to-point, some volume	Point-to-point, some volume	Collective communications and point-to-point
MFLOPS/zone/cycle	0.02–0.1 (10× for iterative schemes)	2–12	0.03–0.07	0.1–3
I/O (startup data)	20–160 MB (EOS)[a]	0.3–12 MB (nuclear)	100–300 MB (nuclear)	0.1–1 KB/zone

NOTE: EOS, equation of state; I/O, input/output; MFLOPS, million floating-point operations per second.
SOURCE: From briefing to the committee by LLNL.

Again, memory requirements will vary depending on the nature of the computation, and LLNL and LANL use different approaches—for example, in their implementation of hydrodynamics or radiation transport. But a common theme for both laboratories is that the need to simulate radiation transport significantly increases memory requirements in either case. It is also apparent that flop rates per cell can be quite low. This will again be related to how well the memory subsystem can deliver data to the computational units and will be discussed further in Chapter 2.

Each laboratory provided sizes, memory requirements, wall-clock runtimes for both routine calculations versus larger hero calculations, which are executed less often but are needed occasionally for key stockpile decisions. Typical simulation characteristics representative of both LANL and LLNL are shown in Table 1-3.

Routine 1D, 2D, and even small 3D calculations are performed on capacity platforms (CTS-1). Larger 3D calculations require petascale platforms such as Trinity or Sierra. Hero calculations for LANL and LLNL are characterized in Table 1-4.

TABLE 1-3 Runtime and Resource Characteristics of "Typical" Simulations

Simulation Type	Number of Nodes	Memory Footprint	Wall-Clock Time
Routine 1D	1 (CTS-1)	All node memory	Minutes–hours
Routine 2D	4–32 (CTS-1)	128–2,048 GB	1 hour–1 week
Routine 3D (small)	10–100 (CTS-1)	1.3–12.8 TB	1 day–2 weeks
Semi-routine 3D (larger)	1,000–2,000 nodes (Trinity) 100–1,000 nodes (Sierra)	64–256 TB	10–100 hours

NOTES: Calculation of memory requirements are 64 GB × number of nodes. Also of note, there is a complicated relationship between simulation type, resolution, physics fidelity, and memory footprint. These are not comparable across codes or even applications of the same code across different simulation types. 1D, one-dimensional; 2D, two-dimensional; 3D, three-dimensional.
SOURCE: From summary of computational requirements provided to the committee by LANL, LLNL, and SNL ASC teams.

TABLE 1-4 Observed Runtime and Resource Characteristics of Prior Los Alamos National Laboratory (LANL) and Lawrence Livermore National Laboratory (LLNL) Hero Calculations

Simulation Type	Number of Nodes	Memory Footprint	Wall-Clock Time
LANL Class A simulation	2,400 (Trinity)	~300–400 TB	6 months
LANL Class B simulation	4,990 (Trinity)	~600 TB	3–4 months
LLNL Class C simulation	288 (CTS-1, ~25% of the machine)	~20 TB	1 month
LLNL Class D simulation	3,250 (Sierra, ~75%)	104 TB	5.8 days
LLNL Class E simulation	512 (Sierra, 12%)	32.8 TB	2 months

NOTES: Class A refers to a moderate-resolution and moderate-fidelity three-dimensional (3D) configuration. Class B refers to a high-resolution and moderate-fidelity 3D configuration for a different physical system. Class C refers to a moderate-resolution/high-fidelity 3D simulation; Class D to a high-resolution/moderate-fidelity 3D simulation; and Class E to a very-high-resolution/moderate-fidelity 3D simulation.
SOURCE: From summary of computational requirements provided to the committee by LANL, LLNL, and SNL ASC teams.

Calculations such as these are performed relatively rarely, depending on priorities and the potential contribution to programmatic decisions. They are disruptive in that they require the use a significant portion of the computational resource. One noteworthy aspect is that memory is at a premium on all these platforms, again reflecting that while floating-point capability has increased to the petascale and soon the exascale range, memory sizes have not kept pace. Apparently, even in the larger LANL hero calculations, approximations were made that degrade the fidelity below deemed desirable by experts. Running these calculations at the desired fidelity would require platforms with capabilities beyond exascale. They would exceed the available memory on an exascale system and require months or even years of wall-clock time. Memory footprints for such "post-exascale" computations are in the range of 5–10 petabytes, again reflecting the mismatch of computational capability and memory capacity and throughput. All laboratories concluded that future such hero simulations will be limited by the difficulty of strong scaling large-memory footprint calculations so that they meet practical time-to-solution requirements.

FUTURE CHALLENGES WILL REQUIRE COMPUTATION BEYOND EXASCALE

Having provided some context as to why HPC is crucial to existing efforts in stockpile stewardship, this section provides examples of some future challenges that will require computational capabilities at exascale and beyond.

Subcritical Experiments

Prior to 1992, the validation of various nuclear designs as well as the furtherance of understanding of nuclear weapon science was achieved through underground nuclear testing. In the era of stockpile stewardship, underground testing is no longer possible, and the emphasis has been on simulation, as well as a limited set of experiments. Experiments in which plutonium is driven by high explosives are allowed under the Comprehensive Test Ban Treaty,[17] but the resulting assembly cannot achieve criticality. Such experiments today must be performed at the Nevada National Security Site (NNSS, formerly the Nevada Test Site). These subcritical experiments have shown great value in resolving various stockpile issues and have also provided important validation data.

LANL and LLNL have recently proposed experiments in which scaled primaries are imploded. Because the primary is scaled to a fraction of its original size, there is no possibility of achieving criticality. The value of such experiments is that, using modern

[17] The United States has not ratified the Comprehensive Test Ban Treaty but currently does observe its restrictions.

diagnostics such as highly penetrating radiography as well as some innovative measurements of neutron output, it is possible to perform investigations of important physical aspects of nuclear performance. These experiments will provide important validation data. However, even modern diagnostics provide only a limited characterization of what happens in an experiment. Realizing the full value of these experiments will require high-resolution simulations that in turn will require significant computational capability. It may seem paradoxical that a higher level of resolution would be required for an implosion where criticality is not achieved. To make predictions about a system at full scale and get maximum utility from the experimental data, the simulations must resolve and be sensitive to a wide range of fission modes that are not important after a system becomes critical. These modes are very sensitive to the details of how the assembly implodes, and so enhanced resolution will be required. The facility for these subcritical experiments is now under construction at the U1A site at the NNSS. When complete and operational, this facility will be one of the few opportunities for future weapon designers to develop their skills. But, absent continued improvement in simulation capability, it will not be possible to realize the full value of these future subcritical experiments.[18]

Aging of Plutonium

At present, Los Alamos provides the only production capability for the plutonium shells that are used in modern primaries. Previously, the production of these shells was performed at the Rocky Flats Plant in Colorado, which is now closed. Additional production facilities are planned at the NNSA Savannah River site, but even after the various existing or planned production facilities are at full capability, it will be important to understand the impact of aging of plutonium on the existing stockpile primaries, as these new facilities will take time to fully impact the stockpile.

The principal decay mode of the isotope of plutonium used in primaries, ^{239}Pu, is the emission of an alpha particle (a helium nucleus) and a uranium atom. ^{239}Pu has a long half-life of 24,100 years, and so the transmutation of the metal via radioactive decay is not an immediate concern. However, when a decay occurs, the products are quite energetic, and so can produce damage in the surrounding crystal lattice of the metal. The damage ultimately manifests itself in the formation of helium bubbles in the lattice. The concern is that this damage will ultimately result in formation of voids that lead to swelling of the metal, something that does occur in other nuclear materials, and these consequences of aging could result in changes to the mechanical properties of the plutonium and affect the implosion characteristics of a primary.[19]

[18] S. Storar, 2021, "Shining a Bright Light on Plutonium," *Science and Technology Review*, Lawrence Livermore National Laboratory, https://str.llnl.gov/content/pages/2021-04/pdf/04.21.3.pdf.

[19] S. Hecker and J. Martz, 2000. "Aging of Plutonium and Its Alloys," *LANL Science* 26.

Careful analysis to date appears to indicate, rather surprisingly, that this lattice damage does not appear to manifest itself in measurable changes in material properties, at least on time scales of tens of years. However, given the potential impact of aging to the existing stockpile, continued experiments are being planned. Given the material complexity of plutonium, supporting computations using ab initio approaches will most likely require exascale capability and beyond to assess how aging affects nuclear performance.

Reentry Flows

As discussed earlier, the NNSA laboratories must qualify that the warheads deployed on the intercontinental ballistic missile (ICBM) leg of the strategic deterrent must survive the reentry environment. The qualification of the reentry body or reentry vehicle are the responsibilities of the Air Force and Navy, respectively, but the design laboratories must understand this environment to assess the thermal and mechanical loadings on the NEP as well as the supporting structures. Much of this assessment for the existing stockpile has been informed in the past through measurements performed during flight testing. But development of future systems or reentry trajectories will increasingly rely on simulation. Simulating these potentially new environments will require increased understanding of hypersonic flows, as well as turbulent transport in such flows. Current work to address these issues will benefit from exascale capability, but given the complexity of this problem there are requirements for capability beyond exascale.

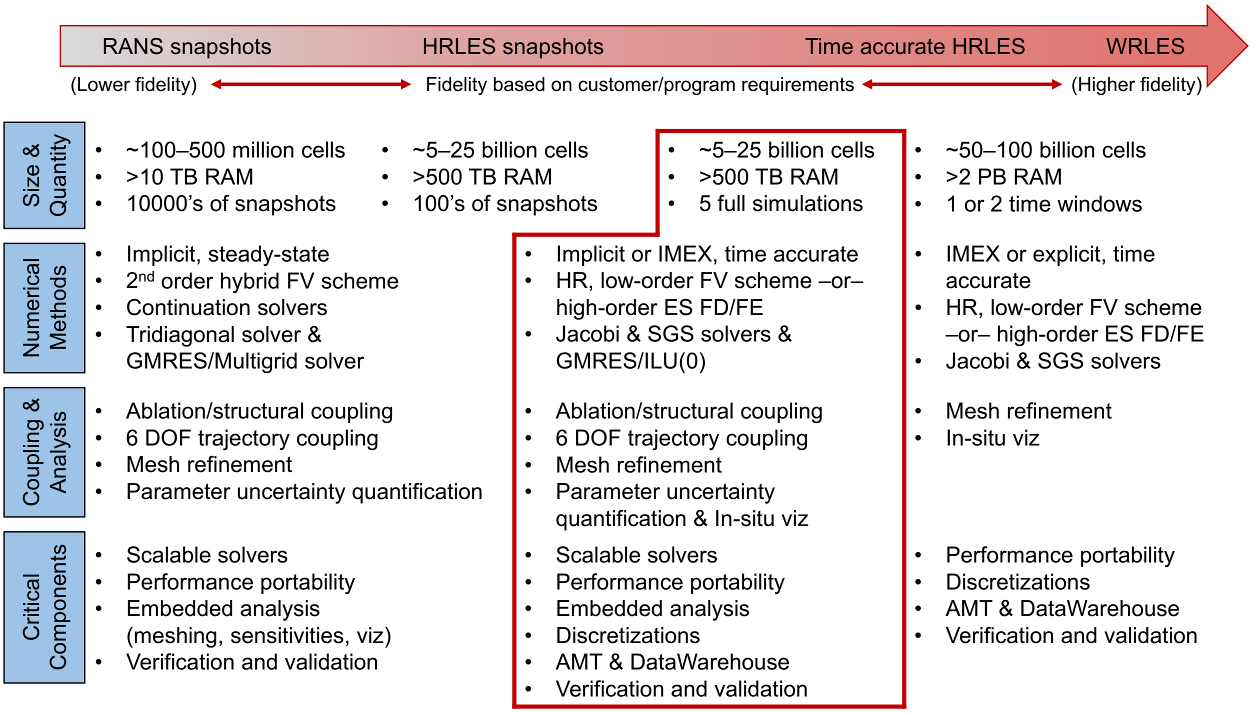
To give some idea of the challenge of surviving the reentry environment, a satellite in orbit at an altitude of 320 km possesses a specific kinetic energy of roughly 3×10^7 J/kg. Although reentry vehicles operate on suborbital trajectories, the specific energy of such a vehicle can approach that of an orbiting body. As a point of reference, carbon vaporizes at 6×10^7 J/kg, and for bodies descending from exoatmospheric altitudes, the rarefied atmosphere inhibits efficient diffusion of heat. As a result, the reentry process generates sufficient energy per unit mass to destroy the vehicle. The challenge then is to dissipate an energy of this magnitude without destroying the reentry vehicle. A conflicting requirement is that the warhead must have an aerodynamic shape to fly stably in the atmosphere. It turns out that a cone shape is not optimal for dissipating the heat of reentry, making the engineering problem more acute. As the reentry body enters the atmosphere, it undergoes significant mechanical and thermal loading, with decelerations on the order of 100–200 times the acceleration, owing to gravity and thermal loads on the order of 1/10 the initial reentry energy. The exact values are dependent on the velocity and angle of reentry. In current systems, the heat load is dealt with by using a sacrificial ablating layer at the tip of the reentry body. This changes the mass properties such as the center of pressure and center of mass, and so careful modeling is required to ensure that the reentry system does not lose aerodynamic stability as it descends through the atmosphere. Last, as the reentry vehicle descends, it experiences different regimes of fluid motion and thus time-varying mechanical and thermal loading. Above an altitude of 125 km, the body descends at 5–6 km/sec and encounters free molecular flow. At an altitude of 75 km, the flow transitions to the continuum limit, where the body experiences its maximum inertial and thermal loads. At these atmospheric densities, the body experiences noise from the wake at the rear of the body. As the body descends further, the boundary layer transitions from laminar to turbulent flow. This regime is particularly stressful because the high level of mixing caused by the turbulence results in higher velocity and thermal gradients near the body wall. Qualitative understanding of the flow dynamics is essential, as the increased thermal loading can result in uneven ablation and loss of stability of the reentry body. New codes are currently being designed to simulate this environment with computational requirements that will exceed exascale capabilities.[20] SNL has provided estimates of the computational requirements to develop a "virtual flight test capability" that would be used to study the performance of a reentry vehicle or body over a range of potential trajectories. The requirements at various levels of fidelity are shown in Figure 1-1.

The axis at the top of the figure refers to the level of fidelity for the simulation of the turbulent flow as the body reenters. The lowest fidelity (corresponding to Reynolds-averaged Navier-Stokes turbulence modeling) is computationally efficient but is known to be inaccurate in various key settings. In contrast, the use of more accurate large eddy simulation has memory requirements exceeding 2 petabytes. Not shown in this figure are the costs of performing a fully resolved direct numerical simulation of the transition to turbulence. SNL estimates that simulating 1 ms of time for transition on a full flight vehicle on a 50 exaflop machine would require approximately 3 days of computation.

HPC IN SUPPORT OF THE NNSA ENTERPRISE

In addition to studying weapons science, NNSA is also today effectively using HPC capabilities to support the entire weapons life cycle (known as the 6.x process), from concept through dismantlement. More recently, a particular need has arisen for HPC across the entire weapons complex to address and avoid issues in the current stockpile, including production, engineering, weapons surveillance, and assessment.

[20] F.J. Regan, 1984, *Re-Entry Vehicle Dynamics*, New York: American Institute of Aeronautics and Astronautics.



(Lower fidelity) ↔ Fidelity based on customer/program requirements ↔ (Higher fidelity)

	RANS snapshots	HRLES snapshots	Time accurate HRLES	WRLES
Size & Quantity	~100–500 million cells >10 TB RAM 10000's of snapshots	~5–25 billion cells >500 TB RAM 100's of snapshots	~5–25 billion cells >500 TB RAM 5 full simulations	~50–100 billion cells >2 PB RAM 1 or 2 time windows
Numerical Methods	Implicit, steady-state 2^{nd} order hybrid FV scheme Continuation solvers Tridiagonal solver & GMRES/Multigrid solver		Implicit or IMEX, time accurate HR, low-order FV scheme –or– high-order ES FD/FE Jacobi & SGS solvers & GMRES/ILU(0)	IMEX or explicit, time accurate HR, low-order FV scheme –or– high-order ES FD/FE Jacobi & SGS solvers
Coupling & Analysis	Ablation/structural coupling 6 DOF trajectory coupling Mesh refinement Parameter uncertainty quantification		Ablation/structural coupling 6 DOF trajectory coupling Mesh refinement Parameter uncertainty quantification & In-situ viz	Mesh refinement In-situ viz
Critical Components	Scalable solvers Performance portability Embedded analysis (meshing, sensitivities, viz) Verification and validation		Scalable solvers Performance portability Embedded analysis Discretizations AMT & DataWarehouse Verification and validation	Performance portability Discretizations AMT & DataWarehouse Verification and validation

FIGURE 1-1 Reentry flight simulation computing needs, numerical methods, analysis, and components for increasing fidelities of simulations.

NOTE: Acronyms are defined as follows: AMT, Analytics Middle Ter; DOF, degrees of freedom; ES, entropy stable; FD, finite difference; FE, finite element; FV, finite volume; GMRES, Generalized Minimal Residual method; HR, high resolution; HRLES, High Resolution Large-Eddy Simulations; ILU, lower-upper triangular factorization; IMEX, implicit-explicit time integration; RANS, Reynolds-Averaged Navier Stokes; SGS, Symmetric-Gauss-Seidel; WRLES, Wave-Resolved Large-Eddy Simulations.

SOURCE: From summary of computational requirements provided to the committee by LANL, LLNL, and SNL ASC teams.

COMPUTING AND EXPERIMENT—POSITIVE SYMBIOSIS

Although the committee has up to now focused on computational requirements for stewardship of the future stockpile, it is also important to point out that there is a symbiotic connection between computation and experiment that is essential to future success. At its origin, stockpile stewardship had two primary branches—strengthening both the nonnuclear experimental capabilities and the computing capabilities of NNSA. Progress on the first of these is evident through the success of facilities such as the National Ignition Facility (NIF), the Dual Axis Radiographic Hydrodynamic Test Facility, the Z Machine, and the U1A facility. Progress on the second is evident via the increased understanding of historic underground nuclear tests and continuing support for LEPs for systems in the stockpile, such as the W87, W76, and the B61.

It might appear that these two branches are independent; they are not. Rather, they are complementary. HPC is and will be a critical tool in *designing* the high-value experiments that are routinely conducted in these facilities and in *analyzing* the results. Without computing, these experimental facilities would be almost useless. Likewise, the results obtained from these experiments serve to highlight deficiencies in computational models, leading to advances in understanding, often at the extremes of temperature and pressure.

As these experimental facilities continue to advance in both diagnostic capabilities and control systems, the flood of data produced offers new opportunities for computing. These include looking for patterns that would be opaque to human observers as well as machine learning to assist in optimizing complex experimental controls. As the resolution of these experimental capabilities continues to increase, it will be important to continue to support these efforts with increasingly capable computation.

GEOPOLITICAL CONSIDERATIONS

The present geopolitical landscape also further reinforces the importance of HPC. As the committee writes this report, the world is experiencing the greatest changes in geopolitics since the end of the Cold War. There is a war of territorial aggression in Europe, where the threat of nuclear weapons is playing both a deterrent and a compellent role. The committee takes note of the development and deployment of new types of delivery systems by peer competitors—for example, the Avangard hypersonic system that Russia has deployed on the Sarmat heavy ICBM as well as China's test of a Fractional Orbital Bombardment System. These developments are occurring in an environment where extant nuclear arms control treaties are at their lowest level in many decades and the prospect of engaging in new treaty negotiations is poor.

While the United States has expressed a clear desire to reduce the role and salience of nuclear weapons in accordance with its commitments under Article VI of the Nonproliferation Treaty, the current geopolitical situation indicates that the challenges of stockpile stewardship will remain and will continue to stress the capabilities of the most advanced computers well beyond exascale computing. HPC capabilities must be responsive to new mission requirements, enabling rapid redesign, assessment, and deployment of new requirements for weapons systems and delivery vehicles. Last, HPC is also required to understand adversarial capabilities. If foreign systems differ in fundamental ways from U.S. systems, high-fidelity simulation will be required to characterize their behavior. Without test data to calibrate them, simplified or lower-dimensional models would not be applicable to alternative designs. HPC will be required to assess such systems to avoid with confidence the possibility of technical surprise.

SUMMARY OF FINDINGS

This chapter includes findings that are supported by the above discussion.

FINDING 1: The demands for advanced computing continue to grow and will exceed the capabilities of planned upgrades across the NNSA laboratory complex, even accounting for the exascale system scheduled for 2023.

> **FINDING 1.1:** Future mission challenges, such as execution of integrated experiments, assessment of the effects of plutonium aging on the enduring stockpile, and facilitation of rapid design and development of new delivery modalities will increase the importance of computation at and beyond the exascale level. Orders of magnitude improvement in application-level performance would allow for improved predictive capability, valuable exploration and iterative design processes, and improved confidence levels that will remain infeasible as long as a single hero calculation takes weeks to months to execute on an exascale system.
>
> **FINDING 1.2:** HPC has traditionally played an important role in support of weapons systems engineering. Some emerging challenges in this arena, such as qualifying future weapons systems for reentry environments, will require new approaches to mathematics, algorithms, software, and system design.
>
> **FINDING 1.3:** Assessments of margins and uncertainties for current weapons systems will require additional computational capability beyond exascale, a problem exacerbated by the aging of the stockpile. Enhanced computational capability will also be required in assessing margins and uncertainties should there emerge requirements for new military capabilities.
>
> **FINDING 1.4:** The rapidly evolving geopolitical situation reinforces the need for computing leadership as an important element of deterrence, and motivates increasing future computing capabilities.

2

Disruptions to the Computing Technology Ecosystem for Stockpile Stewardship

Via the Advanced Simulation and Computing (ASC) program, the National Nuclear Security Administration (NNSA) spearheaded the global development of terascale and then petascale computing, and its collaboration with the Department of Energy's (DOE's) Office of Science in the Exascale Computing Project (ECP) has led to the first U.S. exascale computers. Concurrently, fabrication of leading-edge semiconductors has shifted to offshore foundries such as the Taiwan Semiconductor Manufacturing Company (TSMC), with both supply-chain issues and potential national security implications, particularly for systems that must be deployed in secure environments. Meanwhile, there is credible evidence that China was the first country to deploy exascale computing systems, targeting China's own national security interests.

Moreover, technological shifts owing to the end of Dennard scaling and the slowing of Moore's law have raised questions about the technical and economic viability of continued reductions in transistor sizes and associated growth in computing performance, all at a time when the locus of semiconductor design is now being driven by artificial intelligence (AI) and cloud-computing (Box 2-1) workload needs, rather than predominantly by technical computing.

These shifts in the semiconductor ecosystem and associated market forces, together with rising costs and global competition, are indicative of the challenges now facing NNSA. Simply put, NNSA no longer has the same capability to drive the future of advanced computing as it did in the past. To meet mission needs, NNSA must partner in new and strategic ways to both meet its computing needs and respond to evolving geopolitical circumstances.

TECHNOLOGY DISRUPTIONS

The broad computing environment in which NNSA has acquired and deployed increasingly powerful computing systems as part of the ASC program is now in great flux. Not only are the underlying semiconductor ecosystem and computing hardware and computer system architectures shifting rapidly, so too is the software, along with rapid changes in computing business models and economics. All of these changes, together with shifting mission needs, have profound implications for how NNSA continues to co-design, configure, and deploy future computing systems.

Hardware and Architecture Disruptions

The use of commodity central processing units (CPUs), graphics processing units (GPUs), memory, and storage technologies dominates the computer system structures used today to support high-performance computing (HPC). Increasing the number of cores per chip, compute density, vector processing units, and memory capacity, as well as scaling through parallel processing, adding high-bandwidth memory (HBM), and widening interfaces to improve communications bandwidth—along with building larger systems—have been the primary mechanisms used to increase system performance over the past two to three generations of supercomputers.

Many of these traditional approaches are nearing either physical or practical economic limits. For example, energy dissipation limits on clock operating frequencies have led to only small increases in single-thread performance. As a result, ever larger chip core counts, along with data parallel accelerators (GPUs), have been needed to increase hardware performance, with associated pressure on application developers to increase parallelism and directly manage memory features. Likewise, continued increases in transistor density have become more technically challenging, and the rising cost of semiconductor fabrication facilities has shifted market economics.

Concurrently, the "hyperscalers"—the largest of the cloud service providers—have begun designing their own processors and accelerators, which are not available for purchase. And a growing market focus on improving the performance of machine learning is shifting the locus of hardware innovation. Taken together, the dominant economics of the x86-64 processor ecosystem are at risk, particularly with the rise of ARM, RISC-V, and other custom processor designs. ARM provides a large family of licensable hardware designs and RISC-V is an open architecture, the hardware analogy to open-source software; both support configuration of specialized chips from previously designed components. Last, system cooling capacity, power distribution, energy costs, and carbon impacts have become major considerations in the design and deployment of exascale-class computers. In the midst of these technical changes, the majority of state-of-the-art semiconductors

BOX 2-1 Cloud Computing

In the past decade, there has been tremendous growth in the availability of "cloud computing." This overloaded term has come to mean a variety of things to different people. The National Institute of Standards and Technology (NIST) has produced one set of definitions based on the level at which a cloud operator provides user access.[a]

Abstractly, the NIST definitions describe a cloud model as a "ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources ... that can be rapidly provisioned and released with minimal effort or service provider interaction." Importantly, the NIST definitions include three levels of service, from lowest to highest:

- **Infrastructure as a service:** Here, the user can provision almost all elements of the software stack, including the operating system and associated services. The cloud vendor provides the hardware and associated virtualization software.
- **Platform as a service:** Here, the system software and development tools are provided, and the user can configure and operate their own applications.
- **Software as a service:** Here, the applications are also provisioned, and the user can access and use those applications.

In addition to this level of software access, cloud services can be deployed in "private," "public," and "hybrid" models. As these terms suggest, a private cloud is limited to a single entity, often deployed "on premise" for use by that organization or at a single site for a geographically distributed organization. Similarly, public clouds usually denote those operated by vendors who sell access. Amazon Web Services, Microsoft Azure, and Google Cloud are the best known of these public clouds. Last, hybrid clouds are typically some mix of private and public, with some elements of the software able to spill from one model to the other, based on demand. A key feature of clouds is the appearance of elasticity—that is, from the user perspective, they can rapidly grow or shrink usage with near-immediate response and without new contracts or policy changes. For example, a private cloud with limited hardware capacity might shift work to a public cloud when demand exceeds local capacity.

While cloud computing began primarily a business model rather than a technology—that is, contracting for services rather than purchasing, installing, and maintaining computer systems—there are both hardware and software trends that emerged from the cloud ecosystem. Historically, public clouds were built for massive numbers of small, independent compute tasks but with high expectations of reliability. This focus led to a set of high-level programming frameworks that hide failures in the underlying hardware. More recently, the demand for systems to address large machine learning problems has led to cloud offerings with high-speed networks, graphics processing units, and most recently custom hardware.

When discussing the use of cloud computing for the National Nuclear Security Administration's Advanced Simulation and Computing program, it is important to use these more precise terms and areas of innovation, rather than "cloud computing" as a catch-all concept.

[a] P. Mell, and T. Grance, 2011, "The NIST Definition of Cloud Computing," *Recommendations of the National Institute of Standards and Technology,* Special Publication 800-145, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.

are now being fabricated offshore, with associated device provenance, national security, and economic competitiveness risks.

Looking forward, there are increasingly near and foreseeable limits on size, performance, and capacity for commonly used computer technologies, mediated by a combination of physical processes and feasible costs. The most significant of these is the slowing of Moore's law (i.e., the previous doubling of transistor density every 2 years with associated reductions in cost) owing to the technical and economic challenges described below. The second has been the end of Dennard scaling, which showed that transistor shrinkage resulted in other scaling factors leading to faster single-processor performance. Dennard scaling ended in roughly 2004, when power density and leakage current became limiting factors, and the continued increase in transistors were instead used for on-chip parallelism in the form of multicore and manycore (e.g., GPU) designs.[1] The following is a list of present technologies, and their associated constraints, that will limit the continued growth in performance and capacity of post-exascale supercomputers:

- **Silicon processing limitations** will start to limit the minimum practical size of a transistor, logic gate, wire, or device. Although there is still opportunity to continue shrinking semiconductor feature sizes from the 7-nm silicon process technology used in exascale systems—for example, Oak Ridge National Laboratory's (ORNL's) Frontier and Lawrence Livermore National Laboratory's (LLNL's) planned El Capitan—and the 4 nm in the future Los Alamos National Laboratory (LANL) Venado system, the limits on lithography technology, device isolation, and noise injection are increasingly near. The successful implementation of extreme ultraviolet (EUV) in manufacturing has removed a barrier in lithography resolution for logic and dynamic random access memory (DRAM), but noise, defects, overlay, and edge placement remain challenges. Another constraint for existing silicon processes is metal pitch feature scaling, presently 18 nm, which is projected to reach fundamental limits of around 7–8 nm by the end of this decade. The future will depend in part on a combination of technology advances and economic feasibility (i.e., when and where continued device shrinkage is made possible by markets that can amortize the associated fabrication costs).
- Semiconductor companies have primarily concentrated transistor design on reducing power consumption, rather than maximizing transistor speed, because **integrated circuit (IC) power dissipation** had become such a major design constraint. Along with the end of Dennard scaling, this

[1] National Research Council, 2011, *The Future of Computing Performance: Game Over or Next Level?*, Washington, DC: The National Academies Press.

constraint triggered the rise of multicore chips in the mid-2000s. However, further reductions in contacted gate pitch are essential to increasing transistor density.

- The combination of 3D scaling, both monolithically as a system-on-a-chip (SoC) and heterogeneously as a system-in-package (SiP) will enable compacting more and more devices in a well-defined area/volume, but physical limits of individual devices will eventually be reached owing to **topological or electrically related limitations**.[2]
- Heterogeneous multicore processors are being manufactured by major vendors, containing as many as 64 to 80 cores (so-called manycore) to support today's exascale systems. Multisocket configurations place more cores within the same motherboard. But packaging and **system cooling limits** will start to constrain how much integration can be achieved.
- **Lithographic reticle limits** and the yield of working chips per semiconductor wafer place a practical ceiling on how large silicon dies can be manufactured. The emergence of chiplets—integrating multiple chips, often from different vendors and fabrication processes, on a shared substrate—is both a technical and an economic consequence of chip yields and reticle limits. In turn, these limits pushed Cray/Hewlett Packard Enterprise (HPE) to use chiplets in the construction of processing nodes for ORNL's Frontier.
- A **chiplet** is an IC that provides a specific, unique, and/or optimized function to an integrated computer system. It is designed to be combined via an IC packaging technology with other chiplets, semiconductor components, and interconnect systems to produce a computing system.[3] Chiplets are smaller-scale circuits manufactured separately from the rest of the device, and then integrated into a component or "chip" using an advance packaging and interconnect technology. This allows for each module to be optimized individually, which means that the integrated system can be scaled without affecting performance. By breaking a complex SoC into smaller, modular chiplets and connecting them together, it becomes possible to continue scaling up the number of transistors and other components without hitting the physical limits of a single monolithic chip. Chiplets provide several advantages over traditional SoC designs, including higher performance, improved operating efficiencies, heterogeneous integration, higher yields, and reduced development time, among others.

[2] Institute for Electrical and Electronics Engineers, 2020, "International Roadmap for Devices and Systems," 2020 Edition, https://irds.ieee.org/images/files/pdf/2020/2020IRDS_ES.pdf.

[3] G. Kenyon, 2021, "Heterogeneous Integration and the Evolution of IC Packaging," *EE Times Europe,* April 6, https://www.eetimes.eu/heterogeneous-integration-and-the-evolution-of-ic-packaging.

- The use of chiplets, die stacking, and other advanced packaging technologies, versus monolithic SoCs increases the length of the interconnect among computing elements, **increasing signal delays, and associated nonuniform memory access (NUMA) effects**.
- The use of wider interfaces to increase communications bandwidth between computing components has increased the number of signals and the component pin count required to support these interfaces. **Component packaging limits** increasingly constrain the performance of chip-to-chip and board-to-board interfaces.
- Many of today's HPC applications depend on traditional CPUs and GPUs connected to coherent memory hierarchies using very wide interfaces (e.g., thousands of signals), location-specific signal delays (e.g., NUMA effects), and constrained signaling speed (e.g., chip-to-chip). These and other factors **limit system memory bandwidth and latency**. The increasing disparity between processing speeds and memory access times now challenges von Neumann designs and traditional approaches to parallelism.[4]
- CPUs and GPUs today implement fixed sets of operations—that is, **fixed Instruction Set Architectures** that support a broad range of applications. For a sufficiently narrow class of applications, there are proven energy and performance advantages from building hardware that is simplified and specialized to a specific application. This specialization may be done by using either reconfigurable hardware, such as field-programmable gate arrays (FPGAs), or with more dramatic improvements by using application-specific integrated circuits (ASICs). For either case, the challenges of designing the hardware and providing a software stack that can run on such hardware are enormous.
- Solid-state storage devices have dramatically pushed the **limits of storage systems performance** and become a more integral part of the memory system hierarchy. NAND Flash (a nonvolatile memory technology named for its relationship to the NOT-AND logic gate) will continue to dominate this category. While the number of bits/cell will probably not change (e.g., 3 bits/cell), the continued increase in the number of layers in three-dimensional (3D) NAND technology will provide ongoing increases in capacity. But it is unlikely that performance (latency or bandwidth) will improve at the same rate as capacity (if at all).[5]

[4] J. Dongarra, T. Sterling, H. Simon, and E. Strohmaier, 2005, "High-Performance Computing: Clusters, Constellations, MPPs, and Future Directions," *Computing in Science and Engineering* 7(2):51–59, https://doi.org/10.1109/MCSE.2005.34.

[5] C. Monzio Compagnoni, A. Goda, A.S. Spinelli, P. Feeley, A.L. Lacaita, and A. Visconti, 2017, "Reviewing the Evolution of the NAND Flash Technology," *Proceedings of the IEEE* 105(9):1609–1633, https://doi.org/10.1109/JPROC.2017.2665781.

These are just a few of the convolved technology and economic challenges faced by the computing industry and especially for the development of future HPCs. These technology shifts are now challenged by the rising cost of semiconductor fabrication facilities and practical yield limits on very large chips. Although there are further opportunities for semiconductor feature size reductions, the future is increasingly determined by a combination of economic feasibility (i.e., what is technically possible is not always economically feasible) and market demands.

Scaling Challenges

Any examination of leading-edge HPC systems over the past decade shows that system scale (i.e., the number of compute nodes) is growing rapidly. This growth is in part owing to a slower growth of individual node performance because of challenges associated with Dennard scaling and the slowing of Moore's law. Put another way, to increase performance, vendors have found it necessary to increase the absolute size of the systems while also integrating new technology. In turn, these developments have increased pressure on application developers to achieve higher performance by executing applications at larger scale. As noted below, this focus on weak scaling (Box 2-2) brings an additional set of challenges, given the nature of ASC workloads.

BOX 2-2 Weak Scaling

Weak scaling refers to a computational paradigm in which problem size and computational resources are increased simultaneously so that the problem size per processor element remains constant. Over the past several decades, researchers in applied mathematics and computational science have developed algorithms with good weak-scaling characteristics—that is, that minimize the growth in solution time as problem size increases. These algorithmic advances have enabled applications to harness the rapid growth in parallelism to solve increasingly larger problems. While weak scaling captures one aspect of how well an algorithm behaves and to what degree larger and larger computers will enable us to continue solving larger and larger problems, it neglects a critical aspect of time-dependent multiphysics simulations that characterize much of the National Nuclear Security Administration workload.

Multiphysics simulations often track the evolution of a physical system over time by sequentially advancing the solution over a discrete time interval, referred to as a time step. When referring to "problem size" in the context of weak scaling, we typically mean the "size" in a spatial sense; this is usually increased in one of two ways: either by increasing the size of the physical domain or by increasing the spatial resolution. For example, for a three-dimensional problem in fluid dynamics, the workload per time step would be increased by a factor of 1,000 by either enlarging the physical domain by a factor of 10 in each spatial direction or by keeping the domain constant and increasing the resolution by a factor of 10 in each spatial direction. However, for most algorithms used to evolve multiphysics systems, the number of time steps, whose size is constrained by accuracy and/or stability considerations, also scales with the resolution and/or length scales of the problem. In other words, while with perfect weak scaling we could advance one time step of an evolution equation at the same speed regardless of problem size (assuming the amount of computational resources is scaled with the spatial size of the problem), the actual time to solution will depend strongly on the number of time steps, which increases with system size. Consequently, state-of-the-art capability simulations that are projected to take weeks to months on an exascale machine would require years to complete in their scaled form, making them infeasible in practice.

Beyond the challenges of weak application scaling, increasing system scale brings other challenges, notably energy and cooling issues associated with the computing system's physical plant. In addition, the largest computing system installations, whether at DOE laboratories or at cloud computing vendors, have surfaced another, pernicious issue: the presence of silent errors. Given the rising complexity of chips, vendors have found it increasingly difficult to test new chips fully before shipment. The result is infrequent manifestations of bit errors, ones that occur due only to a rare combination of data and instructions. Although uncommon on a single computation node, when assembled in large numbers, as is the case in leading-edge HPC systems, the overall frequency of bit errors can rise to levels that lead to computational errors in the underlying computations.

Last, as the size and complexity of chips has continued to rise, so have their costs and the challenges of ensuring sufficient yield (i.e., the fraction of chips on each wafer that meet performance and correctness specifications). These challenges, plus the desire to integrate features from multiple sources, have driven the adoption of chiplets. The chiplet model brings both new design challenges and opportunities for integration of NNSA-specific workload accelerators.

Impact of Hardware and Architecture Disruptions on ASC Code Performance

Given the challenges enumerated above, Figure 2-1 shows the evolution of key performance indicators for leading HPC systems. This graph reflects hardware disruptions that have emerged over the years. First, peak compute node floating-point performance continues to increase as multicore architectures and other chip-level parallelization lead to increased floating-point capability. However, unless a calculation is completely local to a specific core of a processor, to make optimal use of this increased floating-point capability, other aspects of performance must also increase. Unfortunately, interconnect node bandwidth has increased, but not at a rate commensurate with the increase of floating-point capability. This disparity translates into a steep decrease in the byte per floating-point operation (FLOP) ratio for the overall interconnect. More serious is the decrease of the byte-per-FLOP ratios for overall memory and memory bandwidth. This decrease implies that a growing class of problems are dominated by memory access, and the actual achievable rate of such calculations has decreased over time.

Instructions for loading of data from memory, branching, and array indexing can easily dominate instructions associated with floating-point operations, especially for codes that have a linear amount of work on each data value. The use of optimized data structures and algorithms that have hierarchy or sparsity typically lead to irregular data access patterns and additional indexing data to compactly represent the structure. It may

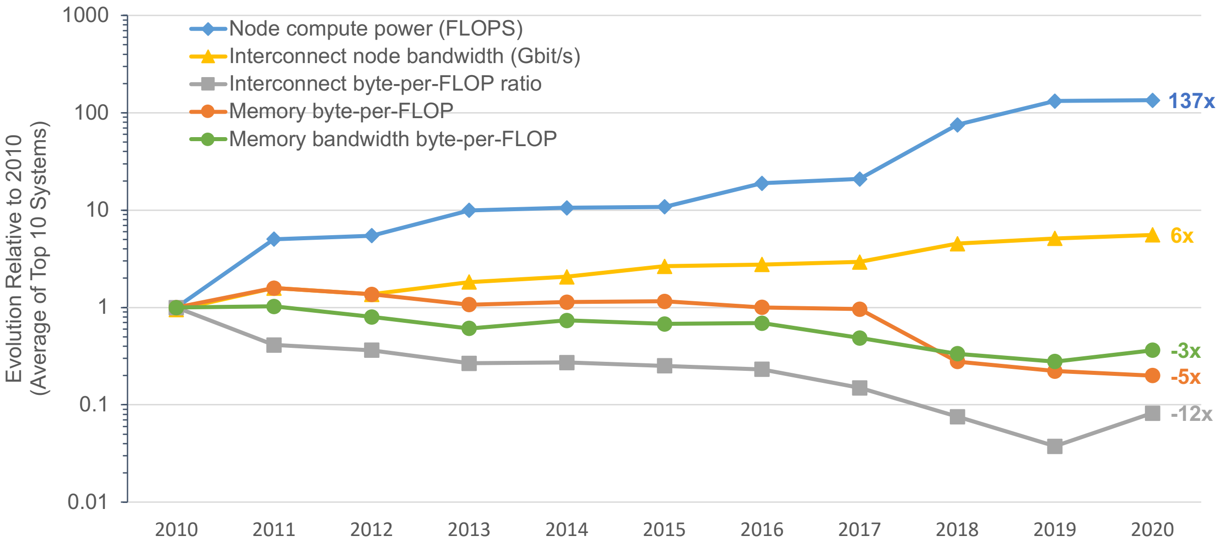


FIGURE 2-1 Evolution of key performance indicators for leading high-performance computing systems.
SOURCE: Provided to the committee by LANL. Created by LANL, which built on work by Keren Bergman (Columbia University).

also require significant amounts of data to update a computational cell because of the frequent need to track a large number of dynamic variables in each computational cell.

The overall impact is that the memory system, both the latency for irregular accesses and bandwidth for regular ones, becomes a bottleneck to achieving a substantial fraction of the peak floating-point capability. Various ingenious approaches such as multilevel caches, prefetching, and cache-friendly algorithms have been introduced to alleviate the issue, but cache or high-speed memory capacity also becomes a limiting factor at each level. This issue has been understood for quite some time, but growing memory capacity and bandwidth is often more expensive than growing arithmetic performance, both in terms of cost and energy consumption, so it has been increasingly challenging to keep machines balanced. Even for well-optimized ASC applications, current HPC systems have become increasingly unbalanced in terms of the overall byte-to-FLOP ratio.

Table 2-1 lists some of the common computational motifs used in LANL's ASC codes. For each motif, the type of parallelism, memory access patterns, and communication patterns, as well as observed bottlenecks are listed. What emerges is that memory bandwidth bottlenecks are apparent for many of the motifs especially if the data structures are dynamic. The yellow- and orange-shaded boxes emphasize that the use of sparse arrays or the need to branch will typically lead to memory bottlenecks. Of course, some of these issues can be addressed via the choice of algorithm, and there is significant ongoing work in reorganizing the data patterns for the various motifs so that they take maximum advantage of the capabilities of emerging HPC architectures such as the availability of GPU accelerators. This has also been a focus of research and development (R&D) with vendor partners.

TABLE 2-1 Common Computational Motifs and Bottlenecks in Los Alamos National Laboratory (LANL) Advanced Simulation and Computing Codes

Motif	Parallelism	Memory Access	Communications	Synchronization	Bottlenecks	Dynamic (Changing) Data Structures
Stencil operations on structured grids	Data parallel	Regular/dense	Neighboring boundary exchange	Point-to-point messages	Memory bandwidth bound	AMR
Stencil operations on unstructured grids	Data parallel	Irregular/dense	Neighboring boundary exchange	Point-to-point messages	Memory bandwidth bound	Sometimes, AMR
Particle methods	Data or thread parallel (divergent)	Irregular/sparse	Neighboring boundary exchange Global or subset collectives	Point-to-point messages Global or subset barriers	Memory latency, network latency	Yes
Sparse linear algebra and nonlinear solvers	Data parallel	Irregular/sparse	Global or subset collectives	Global or subset barriers	Communication bound	Sometimes, AMR
Dense linear algebra	Data parallel	Regular/dense	Local operations	N/A	FLOPS, cache	No, static
Monte Carlo methods	Data or thread parallel (divergent)	Irregular/sparse	Neighboring boundary exchange Global or subset collectives	Point-to-point messages Global or subset barriers	Memory latency, network latency	Generally static
Discrete ordinate methods	Data parallel	Irregular/dense	Neighboring boundary exchange	Point-to-point messages	Messaging rate (sweeps), FLOPS, cache	Generally static
Machine learning	Data parallel	Regular/dense	Local ops/neighbor coms	Global broadcast	FLOPS, memory bandwidth	No

NOTE: AMR, adaptive mesh refinement; FLOPS, floating-point operations per second; N/A, not applicable.
SOURCE: From briefing provided to the committee by LANL.

Nevertheless, the discussion above illustrates that for many ASC applications, performance is strongly tied to emerging hardware trends and that post-exascale strategies must be developed with these in mind. While this analysis is specific to the LANL ASC codes, many of the motifs are common across the laboratories, albeit with different emphases as regards their usage.

This issue is further reinforced in examining the percentage of floating points achieved in various applications that make use of the motifs listed in Table 2-1. Table 2-2 was provided to the committee as part of the presentations by the laboratories on code performance. The table lists several LANL codes that utilize important computational patterns relevant to weapons science. The Flag code is used to simulate hydrodynamic phenomena and uses an unstructured mesh in either a Lagrangian mode or in an Arbitrary Lagrangian-Eulerian (ALE) mode. In a fully Lagrangian mode, the mesh moves with the material velocity. In ALE mode, material can flow through the mesh. xRAGE is a 3D radiation hydrodynamics code that uses adaptive mesh refinement to refine important features such as shock waves while also solving for radiation transport. PartiSN computes radiation transport using a discrete ordinates Sn formulation. Last, Jayenne refers to the use of the Discrete Diffusion Monte Carlo method used to speed up radiation transport computations in optically thick media. Table 2-2 shows how well the memory subsystem is performing for each of the applications. The shading indicates whether the specific subsystem is a bottleneck for the computation, with red indicating a serious bottleneck and green indicating lack of a bottleneck. Floating-point performance is rarely a bottleneck, with the exception perhaps of the PartiSN application. In contrast, cache size, memory size, and memory bandwidth play a more significant role. Similarly, Figure 2-2

TABLE 2-2 Architectural Bottlenecks of Applications of Interest Shows That the Memory Subsystem (Caches, Memory Transaction Rate, Bandwidth) Is Often a Bottleneck in Los Alamos National Laboratory Applications

	Memory Subsystem						Floating Point (FP)		
	L1	L2	L3	DRAM	DRAM BW	Memory Latency	DP FLOPS	Vectorization	Non-FP
Flag 3D ALE AMR	🟨	🟨	🟩	🟨	🟧	🟩	3.70%	11.20%	96.30%
Flag 3D ALE Static	🟨	🟨	🟩	🟨	🟧	🟩	2.20%	7.20%	97.80%
xRAGE 3D AMR	🟥	🟩	🟨	🟨	🟥	🟩	6.50%	14.00%	93.50%
PartiSN 42 Groups	🟥	🟩	🟨	🟥	🟥	🟥	26.20%	90.40%	72.80%
Jayenne DDMC Holraum	🟨	🟩	🟩	🟥	🟩	🟥	15.50%	0.00%	84.50%

NOTES: Red indicates a hardware resource that is heavily utilized, orange indicates moderate utilization, and green indicates light utilization. 3D, three dimensional; ALE, Arbitrary Lagrangian-Eulerian; AMR, adaptive mesh refinement; BW, bandwidth; DDMC, Discrete Diffusion Monte Carlo; DP, double precision; DRAM, dynamic random access memory; FLOPS, floating-point operations per second.
SOURCE: From summary of computational requirements provided to the committee by LANL, LLNL, and SNL ASC teams.

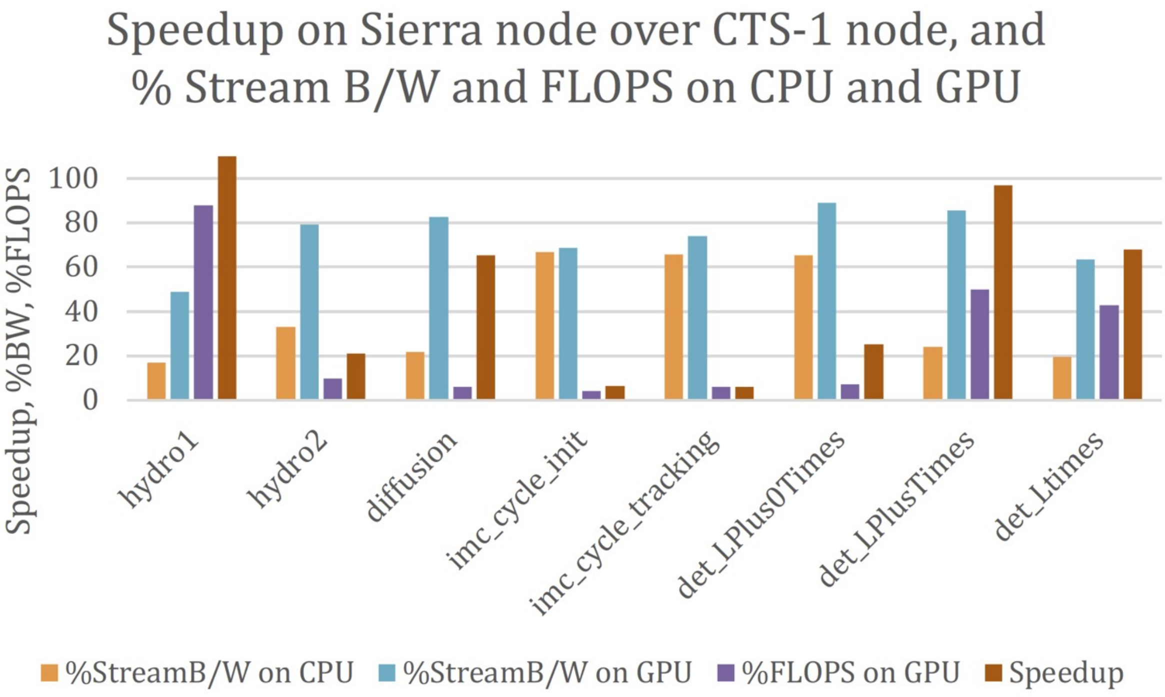


FIGURE 2-2 Measured bandwidth (B/W) and floating-point operations per second (FLOPS) on a subset of problems, kernels, and algorithms.
NOTE: CPU, central processing unit; GPU, graphics processing unit.
SOURCE: From summary of computational requirements provided to the committee by LANL, LLNL, and SNL ASC teams.

shows for LLNL codes that, while a number of codes get impressive speedup from using GPU systems, even those are more often limited by memory limits than by floating point.

Recently, LANL has investigated the use of HBM coupled with fourth-generation Intel processors (formerly codenamed Sapphire Rapids). These new processors feature a multi-die interconnect chiplet technology and are currently being deployed on the Crossroads ASC platform. In a recent paper,[6] LANL demonstrated that the use of these processors with HBM led to roughly an 8× decrease in time to solution on the xRAGE and Flag applications with no change to the codes themselves. This implies that there remain opportunities to further optimize future hardware for workloads relevant to ASC and NNSA.

Software Infrastructure Disruptions

For the past three decades, the global HPC community, including the NNSA and Office of Science laboratories, has leveraged a software and algorithm framework based on machines built as large collections of processing nodes connected via a message passing model called the Message Passing Interface (MPI). The laboratories have invested

[6] G. Shipman, G. Grider, J. Lujan, and R.J. Zerr, 2022, "Early Performance Results on 4th Gen Intel® Xeon® Scalable Processors with DDR and Intel® Xeon® Processors, Codenamed Sapphire Rapids with HBM," Arkiv:2211.05712v1 [cs.DC].

substantial effort in maintaining and extending MPI as a community standard, adapting to the needs of increasingly complex applications and computer systems while maintaining backward compatibility needed for portable software. They have also built reusable libraries, scientific software frameworks, and applications on this model and have leveraged international community efforts in open-source software for scientific computing and general-purpose software.

During that same period, as noted earlier, processing nodes have gone through multiple transformations, from single processors to multicore, manycore, and accelerators, with the introduction of complex nonuniform shared memory systems and software-managed memory. Thus, while the internode programming model has been relatively stable, the node-level algorithm and software model has gone through multiple disruptions, the most recent being the GPU-accelerated nodes in U.S. exascale systems. The laboratories have also developed new programming abstractions, influenced community standards such as OpenMP and C++, and worked with vendors to ensure that their applications would be well-supported on each platform generation. In addition, they created an environment for software development, curation, and distribution of exascale tools and applications.

The exascale software challenges offer a foreshadowing of the future, as each GPU vendor uses a different native programming language, and different hardware features are exposed to software even across generations of GPUs by the same vendor. This requires rewriting code and sometimes different fine-grained parallel algorithms. The DOE ECP (Box 2-3) addressed these challenges with R&D investments at all levels of the software stack, but without an established plan to sustain the DOE Office of Science investments even for software maintenance, much less for improvements required for future applications and systems. Moreover, these challenges pale in comparison to the software challenges associated with emerging specialized processors, which require tools to design and test the hardware and their own implementation of standard programming languages or their own language. Although large commercial entities have the resources and expertise to design and implement such programming systems for large markets, the resulting programming systems are sometimes not well-suited to the NNSA workloads. NNSA should attempt to team with these vendors and seek adaptation where needed.

The growth of infrastructure as a service (IaaS) and platform as a service (PaaS) over the past 20 years has enabled a large number of enterprises to develop large applications using new techniques for marshaling large amounts of computing resources and allowing significant communication between those resources. Starting with the goal of making systems easy to program and from the premise that hardware components fail, PaaS models provide a high-level programming interface similar to a small subset of collective operations in MPI, but with much more complex implementations to hide hardware failures, variable hardware speeds, and other system artifacts from

BOX 2-3 The Exascale Computing Project

The Exascale Computing Project (ECP) is a joint project between the Department of Energy (DOE) Office of Science and the National Nuclear Security Administration (NNSA) with responsibility for delivering a capable exascale ecosystem, including software, applications, and hardware technology to support the nation's exascale computing imperative. Formally managed under DOE Order 413.3b, ECP has a firm schedule, budget, and milestones, and while it is deploying state-of-the-art mathematics and computer science techniques, it is not designed to cover fundamental research. It builds on fundamental research from the past, but does not address the research that might be required for post-exascale systems and applications.

There are 24 ECP application projects covering a range of traditional modeling and simulation areas, such as materials science and climate modeling, as well as applied energy applications, such as the power grid and wind energy, and applications involving high-performance data analysis for DOE's experimental facilities and machine learning for health. Three nuclear security applications are fully funded by NNSA and developed entirely by the NNSA laboratories. The other 21 are open science and engineering applications and are funded by the Office of Science (the Advanced Scientific Computing Research Office [ASCR]). In addition, ECP includes co-design centers that are developing software that supports common computational motifs found in the applications and a broad set of software projects that address other components of the software stack. Applications, co-design, and software projects all involve teams from both NNSA and other DOE laboratories as well as university researchers. The NNSA applications are entirely funded by NNSA, but the other applications and co-design centers are entirely funded by ASCR, and software and hardware investments by both ASC and ASCR. The ASCR funding includes teams at the NNSA laboratories.

The total cost of the ECP effort is projected to be $1.8 billion over the 7 years of the project, which is aligned with the original plans. The software projects for both applications and systems software are generally expected to meet the original project goals on schedule. ECP includes hardware research and development but not the procurement costs for the first three exascale systems, which are part of the broader Exascale Computing Initiative: Frontier at Oak Ridge National Laboratory, Aurora at Argonne National Laboratory, and El Capitan at Lawrence Livermore National Laboratory. Each of those exascale systems have a very rough estimated cost of $600 million,[a] bringing the cost of ECP plus hardware procurement to around $4 billion.[b]

[a] O. Peckham, 2022, "The Final Frontier: US Has Its First Exascale Supercomputer," *HPC Wire*, May 30, https://www.hpcwire.com/2022/05/30/the-final-frontier-us-has-its-first-exascale-supercomputer.

[b] D. Reed, D. Gannon, and J. Dongarra, 2022, "Reinventing High Performance Computing: Challenges and Opportunities," https://arxiv.org/abs/2203.02544.

the programmer. PaaS programming systems such as MapReduce, Hadoop, and Spark are significantly more resilient to hardware and software faults than MPI, but also incur substantial runtime overheads. Over time, the PaaS models have evolved to be more efficient, and some of the latest models used for compute-intensive problems such as training deep neural networks have used MPI or similar message passing models that relax resilience requirements. There is both a significantly broader economic base of support for these PaaS models and a growing workforce that can be tapped when applying it. Therefore, while the value of these PaaS models for NNSA HPC application workloads is unclear, they may be useful for some mission applications, and in any case represent a growing disruptive force that will need to be either exploited or be in competition with by post-exascale computing programs of NNSA.

Last, tied in part to the commercial IaaS and PaaS ecosystems, there are entirely open-source software efforts to build higher-level programming models, such as Julia and Python, as well as new tools for collaborative science. This rich software ecosystem is evolving quickly and expansively, and it is largely, although not entirely, disjoint from that used by NNSA.

MARKET ECOSYSTEM DISRUPTIONS

When DOE's Accelerated Strategic Computing Initiative (ASCI)[7] and Advanced Scientific Computing Research (ASCR)[8] programs began, they were focused on leveraging the commodity computing ecosystem, itself driven by the market economics of complementary metal-oxide semiconductor (CMOS) microprocessors. At that time, several vendors developed and marketed high-end computing systems for both DOE needs and the broader scientific and engineering technical computing market.

The Sandia National Laboratories (SNL) ASCI Red system, built by Intel in late 1996, exemplified this approach by combining large numbers of x86-64 processors via a message passing network to become the first terascale computing system. Other ASC computing platforms from IBM (e.g., ASCI White, Purple, Roadrunner, and Sequoia) leveraged a combination of commercial microprocessors (IBM POWER) or custom, low-power processors (Cell and BlueGene), increasingly with GPU accelerators.

Although the ASCI and ASCR programs benefited immensely from leveraging the economically dominant computing ecosystem of the 1990s and early 2000s, that ecosystem has shifted markedly since the programs began. In the United States, HPE remains as the only U.S. high-end HPC integrator, with HPE having recently acquired Cray. In turn, IBM has largely exited the high-end HPC market, choosing to focus on more profitable market niches. As reflected in recent TOP500 lists, the leading-edge HPC market is now largely a monoculture, dominated by x86-64 processors and primarily by NVIDIA GPU accelerators.

Concurrent with vendor consolidation at the leading edge, extant and new hardware vendors have shifted their focus to IaaS cloud computing and AI. By market capitalization and annual infrastructure investments, Amazon Web Services, Microsoft (Azure), and Google (Cloud Platform), now have much greater economic influence on future computing system design (i.e., microprocessors, accelerators, network interfaces, and

[7] Department of Energy Defense Programs, 2000, "Accelerated Strategic Computing Initiative (ASCI) Program Plan," Office of Scientific and Technical Information (OSTI), https://dx.doi.org/10.2172/768266.

[8] Department of Energy, 2023, "Advanced Scientific Computing Research," Office of Science, https://www.energy.gov/science/ascr/advanced-scientific-computing-research.

storage) than do either the traditional HPC vendors or DOE. Equally important, these companies focus on selling value-added services, not hardware, although all of them develop custom hardware to support their software services.

Similarly, machine learning hardware and software are now a major focus of venture investments (e.g., Cerebras, GraphCore, Groq, Hailo, and SambaNova). Put another way, the locus of financial investment has shifted from traditional CPU vendors to that of cloud service providers and innovative machine learning hardware startup companies.

Simply put, the scale, scope, and rapidity of these cloud and AI investments now dwarf that of the DOE ASCR and ASC programs, including ECP. Tellingly, the market capitalization of a leading cloud services company now exceeds the sum of market capitalizations for traditional computing vendors. Collectively, this combination of ecosystem shifts increasingly suggests that NNSA's historical approach to system procurement and deployment—specifying and purchasing a leading-edge system roughly every 5–6 years—is unlikely to be successful in the future.

Looking forward, it is possible that no vendor will be willing to assume the financial risk, relative to other market opportunities, needed to develop and deploy a future system suitable for NNSA needs. Second, the hardware ecosystem evolution is accelerating, with billions of dollars being spent on cloud and AI hardware each year. NNSA's influence is limited by its relatively slow procurement cycles and, compared to the financial scale of other computing opportunities, is also limited in its financial leverage. Third, and even more importantly, the hardware ecosystem is increasingly focused on cloud services and machine learning, which overlap only in part with NNSA computing application requirements.

The implications are rather clear, especially when considering the mandates of the newly passed Creating Helpful Incentives to Produce Semiconductors (CHIPS) and Science Act (discussed below). NNSA and other federal agencies seeking to deploy next-generation HPC systems will need to develop new approaches to shaping the underlying technologies and partnering with vendors, other agencies, and academia. Potential examples include much deeper, earlier, and more robust co-design than has occurred to date, new partnerships with cloud vendors to jointly develop and build new computing technologies, and interagency collaborations and strategic government investment in promising technology areas and vendors.

Supply Chains, National Security, and the CHIPS and Science Act

Recent chip shortages have highlighted U.S. dependence on the global semiconductor ecosystem for critical semiconductor components, both for consumer products and for mission-critical national security interests. This shortage was triggered in part by the pandemic but also by technology shifts as U.S. semiconductor vendors focused on chip design and relied on offshore vendors for fabrication. Concurrently, many of these semiconductor foundries moved to EUV fabrication processes to continue increasing transistor density.

Intel's lag in adopting EUV has been cited as one of the underlying reasons for the current fabrication advantages of TSMC. Meanwhile, both NVIDIA and AMD, as fabless designers, depend on TSMC and GlobalFoundries for chip fabrication. Recognizing the severity of the chip shortage, DOE invoked national security priorities to ensure parts availability for its exascale systems.

As a result of these trends, the United States now finds itself significantly dependent on global supply chains for advanced semiconductors, with only roughly 10 percent of the global supply of chips and a limited fraction of the most advanced chips produced domestically.[9] To satisfy its need for the advanced semiconductors so critical for HPC, the U.S. government must ensure the viability of domestic suppliers—semiconductor design, chip fabrication, system design, and system integration.

Recognizing both these economic and national security risks, exacerbated by pandemic-related supply chain delays, the U.S. government responded by passing the 2022 CHIPS and Science Act.[10,11] Among its many provisions to strengthen U.S. competitiveness, the CHIPS and Science Act appropriates about $39 billion for semiconductor manufacturing facility incentives, about $2 billion for legacy chips used in automobiles and defense systems, and about $13.2 billion for R&D and workforce development. In response, Intel, Micron, Qualcomm, and GlobalFoundries, among others, have announced new plans for domestic semiconductor fabrication facilities.

The CHIPS and Science Act also requires recipients of U.S. financial assistance to join an agreement prohibiting certain material expansions of semiconductor manufacturing in China. In addition, the Department of Commerce recently imposed restrictions on exports to China, targeting chips key to deep learning and HPC.[12]

International Supercomputing Landscape

While there is general agreement that the LINPACK benchmark[13] is not the metric to use to determine appropriate computer system capabilities for NNSA, the TOP500 list based on LINPACK has been a useful tool to track trends globally and to gauge international

[9] Semiconductor Industry Association, 2021, "State of the U.S. Semiconductor Industry," https://www.semiconductors.org/wp-content/uploads/2021/09/2021-SIA-State-of-the-Industry-Report.pdf.

[10] CHIPS and Science Act, P.L. 117-167.

[11] "White House Fact Sheet on CHIPS and Science," 2022, https://www.whitehouse.gov/briefing-room/statements-releases/2022/08/09/fact-sheet-chips-and-science-act-will-lower-costs-create-jobs-strengthen-supply-chains-and-counter-china.

[12] Department of Commerce, 2022, "Commerce Implements New Export Controls on Advanced Computing and Semiconductor Manufacturing Items to the People's Republic of China (PRC)," Press Release, October 7, Bureau of Industry and Society, https://www.bis.doc.gov/index.php/documents/about-bis/newsroom.

[13] "The LINPACK Benchmark," 2022, TOP500, https://www.top500.org/project/linpack.

supercomputing progress starting in 1993. In June 2002, when the Japanese Earth-Simulator leapfrogged the previous fastest machine—a U.S. machine at LLNL, ASCI White—by 5× to become number one on the TOP500 list, the United States took notice, as did the world. Beyond the benchmark results, the machine was very well-suited to complex modeling and simulation problems and, for example, was sought after by the global climate modeling community. By November 2004, the United States had retaken the lead with a machine twice as fast, BlueGene/L at LLNL, which proved to be capable of solving several groundbreaking science and NNSA problems. The United States remained number one with that LLNL system, followed by Roadrunner at LANL and Jaguar at ORNL, until China took the number one spot in November 2010. Each of these major systems was an enormous research and engineering accomplishment, reflecting both a national commitment to scientific computing and the ability to assemble the team and technologies necessary for such a deployment.

China had been telegraphing its intent to enter the supercomputing race for a number of years, but even when it reached the number one spot in 2010 there was skepticism that China could compete more broadly with well-established HPC investments in the United States, Japan, and Europe. Since that time, it is apparent that China has been investing toward world leadership in supercomputing by building a substantial HPC infrastructure in terms of workforce and companies, developing indigenous hardware, software, and system capabilities and focusing on key applications. As a reflection of that broader program, in June 2022, China had 35 percent of the machines on the list, mostly displacing U.S. machines, which had only 25 percent of the systems but historically has averaged 50 percent. China continues its progress, published under the National High-Tech R&D Program also known as the 863 Program, which is implemented in successive 5-year plans. It is now widely rumored that China has two exascale computing platforms that have not been "officially" reported or entered into the TOP500 competition, and several AMC Gordon Bell Prize submissions (used to measure performance on applications) were run on one of those systems, OceanLight, with impressive results. Although there is limited information on how China is using its supercomputers, HPC could facilitate many of China's military modernization objectives, including upgrading and maintaining nuclear and conventional weapons.[14]

Japan focuses on a tiered computing infrastructure: one flagship computing system, 13 general-purpose HPC centers, and a number of other centers dedicated to specific user communities, all accessible to the science community. In 2002, the flagship computer was the Earth Simulator. Nine years later, in June 2011, the K computer

[14] U.S. House of Representatives, 1999, "High Performance Computers," Chapter 3 in *Report of the Select Committee on U.S. National Security and Military/Commercial Concerns with the People's Republic of China,* H. Rept. 105-851, January 2, https://www.govinfo.gov/content/pkg/GPO-CRPT-105hrpt851/html/ch3bod.html.

was number one on the TOP500 list and another 9 years later Fugaku was number one. Fugaku was developed by extensive co-design of custom processor hardware and system architecture, driven by development of target applications in several priority areas of benefit for society. As of June 2022, Fugaku continues to rank number one on the High Performance Conjugate Gradients benchmark list, a benchmark more reflective of memory bandwidth than the TOP500.

Europe is the other major player in supercomputing. There have been several projects to organize supercomputing applications within Europe. The Partnership for Advanced Computing in Europe provides coordinated HPC infrastructure for large-scale scientific and engineering applications, particularly for industry, across three tiers. The European High-Performance Computing Joint Undertaking is a joint initiative between the European Union, European countries, and private partners to further develop a supercomputing ecosystem that supports development and use of demand-oriented and user-driven applications across a large number of public and private users. In terms of hardware, Europe is pursuing a hybrid strategy, developing some custom hardware but also leveraging U.S. semiconductor designs.

Given restrictions related to U.S. technology, which have increased even as this report was being prepared, China has had little choice but to strengthen its independent research and technical innovation, striving to develop domestic replacements for all the necessary hardware, system software, and application codes for advanced computing. Hence, China’s recent exascale systems (OceanLight and Tianhe-3) have been based on domestically designed chips. The chips in these systems, like those in the U.S. exascale systems, depend on fabrication facilities in Taiwan.

Given the international interest and developments, maintaining U.S. computing leadership for national priorities in a globalized world will require increasing investments and attention.

RETHINKING INNOVATIONS, ACQUISITION, AND DEPLOYMENT

Given the dramatic changes in hardware—driven by a combination of semiconductor constraints, the cloud service provider market, and deep-learning workload demands; new software models arising from the explosive growth of infrastructure and platform services and deep learning; and computing ecosystem economics accruing from these hardware and software forces—it seems likely that NNSA will need new approaches. Among these is the need for better metrics, ones that emphasize reducing time to solution for real applications (e.g., moving the time to solution for important hero calculations from months to days). Today, NNSA applications are most often constrained by memory access bandwidth, not simply by floating-point speeds. An overemphasis on the latter is counterproductive. Instead, NNSA should emphasize time-to-solution and identify the memory access motifs core to key applications as part of an end-to-end, hardware-software, co-design strategy.

Hardware and Architectural Innovation and Diversity

It is possible that the next generation of HPC systems can be built using evolutionary variants of system architectures, component technologies, interfaces, and memory hierarchies, albeit likely with high acquisition costs and limits on the fraction of peak hardware performance delivered to applications. Given current technology challenges and market uncertainties, several alternative approaches are the subject of active research.

Unlike the transition from custom vector processing to commodity, CMOS-based massively parallel systems, there is less consensus about the most effective approaches for future computing systems. It is difficult to predict where the promise lies, whether it is in sustaining CMOS scaling, disruptive computing models (quantum, neuromorphic, or resistive computing), materials (graphene, nanotubes, or diamond transistors), or new structures (3D die stacking, photonics, or spintronics). Evolutionary technology advances will likely include incremental CMOS device scaling, a transition from fin-shaped field-effect transistors (FinFETs) to gate-all-around transistors, chip stacking and 3D chip technologies, targeted wafer-scale devices, and increased use of silicon photonics. As CMOS device scales approach one nanometer, the manufacturing challenges of existing CMOS scaling will continue to mount.

Concurrently, industry standards around packaged chiplets (e.g., via the Universal Chiplet Interconnect express [UCIe]) offer the opportunity for packaging heterogeneous components, including custom-designed accelerators that target elements of NNSA computational workloads. In this spirit, continued exploration of architectures specialized to common computational motifs as well as low-precision arithmetic and its interplay with algorithms and applications opens additional possibilities.

Last, more speculative approaches include alternative semiconductor device designs based on graphene or carbon nanotubes, coupled with novel computing structures and architectures. Other challenges lie in chip-to-chip signal delays, manufacturing yields, process control, and system cooling. As the emergence of chiplets shows, the ability to integrate different processes (e.g., memory, processors, and accelerators) as if they were on a single chip, while maintaining high capacities and density, opens new opportunities for architectural specialization and performance optimization, targeting key workloads.

Similarly, the increased use of HBM, a high-speed interface for 3D-stacked synchronous dynamic random-access memory, can and has reduced memory access latencies, increased data throughput, and reduced energy requirements for selected applications. Further exploration of new memory technologies, both DRAM and resilient, high-performance, nonvolatile memory will be key to ameliorating the large disparity between processor cycle times and memory access times.

Increasingly, processor and system architectures can no longer be developed independently. Software constructs like programming languages, tools, communications protocols, and operating environments will need to be developed in conjunction with new processor and system architectures. For example, the integration of FPGAs into processing units (CPUs, GPUs, etc.) could allow algorithm optimization using on-demand, integrated hardware/software compilation. The ability to compile hardware and software in on-demand or in real-time could significantly change the way future systems are constructed and operated. In any case, co-development of components to support post-exascale systems will be necessary to realize the operational efficiencies and performance needed for future applications of interest.

Software Innovation

The NNSA laboratories, as well as partners in the Office of Science, have been innovators in software development (Box 2-4), especially in the ECP. This project includes a large applications development effort focused on developing mission-critical applications that could effectively use exascale hardware and take advantage of state-of-the-art algorithms and software techniques. The software technologies and co-design elements of the project were valued based on their adoption by applications, yielding a vertically integrated software stack focused on meeting application requirements in which interoperability of the different components was a key feature. ECP also led to wider adoption of good software engineering practices such as regular regression testing and continuous integration development workflows within the HPC community. It is important that in the post-exascale era these innovations be continued and expanded.

Despite these important advances, there is an urgent need embrace state-of-the art industrial practices and toolchains, as well as higher-level abstractions and toolkits, to improve developer productivity, while also improving product quality, reducing development time and staffing resources, increasing software sustainability, and reducing the cost of maintaining, sustaining, and evolving software capabilities.[15] With the prospect of increasing hardware specialization to meet performance objectives and mission needs, abstractions that enable performance optimization while isolating hardware details and maximizing developer productivity are increasingly critical. In addition, the divergence of

[15] M.A. Heroux, L. McInnes, D.E. Bernholdt, A. Dubey, E. Gonsiorowski, O. Marques, J.D. Moulton, et al., 2020, "Advancing Scientific Productivity Through Better Scientific Software: Developer Productivity and Software Sustainability Report," Office of Scientific and Technical Information, https://dx.doi.org/10.2172/1606662.

BOX 2-4 Software

Software, in the context of high-performance computing, can refer to many different facets within the overall ecosystem. The following coarse taxonomy of the types of software found in a vertically integrated software stack is intended to facilitate a more precise discussion of software-related issues.

- **System software:** Software components of the operating system such as process management, memory management, network management, and system security.
- **Programming environments and tools:** Languages, compilers, debuggers, performance analysis tools, source code management tools, and so on.
- **Mathematical libraries and frameworks:** Software that encapsulates common mathematical operations that are used by applications.
- **Applications software:** End-use applications code used to model specific scientific problems.
- **Data management and analysis software:** Software used to manage, visualize, and analyze data from simulation (or experiments).

NNSA programming environments and tools from the mainstream creates both sustainability risks and workforce recruiting and retention challenges. While NNSA prioritizes performance and performance transparency with languages such as C++ and parallel extensions, much of industry and university education now focuses on languages like Java, Python, or Rust with their managed runtime systems, as well as machine learning frameworks that hide parallelism. Scientific productivity has been identified as one of the top 10 exascale research challenges,[16] and software productivity (the effort, time, and cost for software development, maintenance, and support) is a critical aspect of scientific productivity. A significant challenge is the need for high-quality, high-performance, reusable, sustainable scientific software, and programmer productivity tools so that scientists can collaborate more effectively across teams. Thus, although general and actionable metrics have proven difficult to define, an overarching focus on productivity and sustainability is essential to making clear decisions in the face of both highly disruptive architectural changes and demands for greater interaction across distinct teams and reliability of results.

In the post–Moore's law era, it will be increasingly important to use architectural innovations as well as software that can take advantage of that hardware, as parallelism at different levels and in new forms will only increase, leading to rapid changes in software and algorithms. Absent high-level, flexible toolkits and abstractions, the software burden for optimized applications, libraries, and programming tools will also grow as if hardware becomes more specialized to different workloads.[17]

[16] Department of Energy, 2014, *DOE Advanced Scientific Computing Advisory Subcommittee (ASCAC) Report: Top Ten Exascale Research Challenges*, February 10, https://www.osti.gov/servlets/purl/1222713.

[17] C.E. Leiserson, N.C. Thompson, J.S. Emer, B.C. Kuszmaul, B.W. Lampson, D. Sanchez, and T.B. Schardl, 2020, "There's Plenty of Room at the Top: What Will Drive Computer Performance After Moore's Law?" *Science* 368(6495):9744, https://www.science.org/doi/abs/10.1126/science.aam9744.

System Acquisition Innovation

As noted above, when describing ecosystem disruptions, it seems likely that NNSA will need to innovate in the mechanisms it uses to acquire systems in the post-exascale era. There are at least four potentially viable paths to consider:

- **Adapt the current model of procuring every few years a leading-edge system from the commercial HPC market.** This will require even tighter cooperation with the limited pool of potential vendors and, if government restrictions on the use of non-U.S. vendors continue, will likely result in very few competent bidders, perhaps zero or one. This will almost certainly result in increased acquisition cost and increased risk to timely delivery.
- **Become a self-integrator of systems.** The laboratories have already taken over a significant portion of the software stack in producing leading-edge systems. That trend could be increased to the point where an external integrator only integrated the hardware components of the systems. Carrying this further, it would be possible to have the laboratories shoulder the entire burden (potentially including the use of custom hardware), but that would require significant (and very difficult to achieve) increases in laboratory staff expertise. Arguably, the design and deployment of SNL Red Storm followed some elements of this path.
- **Cultivate a new relationship with the cloud vendors,** each of which do custom hardware design and significant self-integration. The benefit here is that one could attempt to leverage their workforce. For security reasons, NNSA will need to use private cloud deployments rather than a public cloud. Risks include vendors' potential lack of interest in the more niche HPC market.
- **Use the defense industrial base.** Generally, this would result in contractual arrangements based more on cost-plus models than fixed deliverables at a price, much as is done for military equipment. There are many companies that integrate very complex systems who could take on work of this sort, but the final net costs would be significantly higher than today. This is owing to both the lack of competitive contracting mechanisms and the lack of a commercial market over which to amortize the development costs.

Cloud Computing as an Alternative

For many organizations, cloud computing is a viable and cost-effective alternative to acquiring and operating computing systems, especially when compared to operating small clusters or individual servers at many sites across an organization, when workload varies widely over time—for example, owing to seasonal sales demand or when "burst demand" exceeds local capacity. NNSA's HPC computing is centralized at a small number of sites, operates in a classified environment, and requires highly integrated computing systems configured to extreme scale and other NNSA requirements—with a nearly unlimited set of computational problems, even their largest systems run at high utilization. Traditional pay-as-you-go cloud computing in public clouds is not an effective alternative. However, NNSA can consider private-cloud options,[18] consider various partnering opportunities with cloud providers, and examine opportunities to adopt ideas from cloud computing, which will be a powerful market force going forward, both in talent and in computing technology.

The cloud ecosystem offers the potential for significant innovation in NNSA's use of software, both because of the rate of innovation occurring in the cloud sector and also because of the models of resilience and management they offer. Commercial cloud data centers exceed the scale of DOE's facilities, and there is growing technical convergence in the need for high-speed networking and integration for large machine-learning workloads. Second, as previously described, the hyperscaler cloud providers are engaged in custom hardware development and will be more influential on the computing supply chain, including the semiconductor market, than NNSA alone or in partnership with the Office of Science. Moreover, as an increasing fraction of the workforce is being trained to use cloud services, there is also the potential to attract and leverage this experience, or conversely the inability to draw such talent if NNSA's environment is viewed as outdated or less productive than tools and systems used in the cloud.

ADAPTING TO A CHANGING COMMERCIAL ENVIRONMENT

The current NNSA procurement model for advanced computing systems is predicated on a vibrant commercial computing market whose interests and products align with the national laboratories' scientific computing needs. This is increasingly not the case.

The burgeoning PC market, which originally birthed the "attack of the killer micros," is increasingly stagnant, in counterpoint to the rapid growth in the smartphone, embedded devices, and data center markets. The former two trade off performance for low power and size, while the latter is increasingly dominated by the hardware demands of the cloud hyperscalers and targeted accelerators for deep learning, of which GPUs are but one instance. NNSA does not have sufficient market influence on processor and

[18] Department of Defense, 2022, "Department of Defense Announces Joint Warfighting Cloud Capability Procurement," December 7, https://www.defense.gov/News/Releases/Release/Article/3239378/department-of-defense-announces-joint-warfighting-cloud-capability-procurement.

memory vendors to compete with hyperscalers, and the reduction in the number of HPC system integrators raises other risks.

Reflecting the technical and financial challenges of a post–Moore's law environment, where it has become impractical to build ever-larger, higher-performance, monolithic chips, the semiconductor market is shifting rapidly to multiple chip packaging—chiplets that integrate multiple, heterogeneous chips via a high-bandwidth interconnect and package. Simply put, chiplets are a technical solution to a minimax problem—minimizing overall costs while maximizing chip yields and delivered performance. In turn, this has created innovation opportunities to develop targeted, ASICs that integrate with the extant ecosystem.

This is in marked contrast to the earlier "killer micro" world, where developers of custom processors faced daunting technical and economic challenges, needing to develop a complete software and hardware environment and keep pace with the relentless performance increases of the mainstream microprocessor market. In the past, many custom designs were tried and failed. Today, with slowing microprocessor performance increases and the chiplet disaggregation, it is both economically and technically viable to design and integrate such custom accelerators, and many startups are doing so, targeting the AI/deep learning market.

In this greatly changed marketspace, HPC customers and the HPC industry can no longer dictate product specifications. The leading-edge HPC market is too small, the procurements are too infrequent, and the financial risk to vendors is too high when compared to the size and scale of the hyperscaler and deep-learning markets. The message is clear. DOE must pursue several strategies directions concurrently: "join them," "beat them," and pursue a new procurement model.

First, NNSA must embrace the existing AI and hyperscaler hardware market, not as a risk-mitigation strategy, but as a mainstream approach to its next-generation hardware. This almost certainly means new types of vendor partnerships, where NNSA is the collaborative partner rather than the procurement driver. In so doing, NNSA may need to reconsider which mission problems can benefit from advanced computing and develop mathematical methods, algorithms, and software that match emerging commercial AI hardware attributes. This could be for components of simulation or data analysis problems by deploying AI methods. With a deeper bench of in-house AI methods and hardware experts, they may also find opportunities to influence commercial designs in ways that would benefit both commercial and NNSA workloads. This "join them" strategy leverages the economic trajectory of the current market, just as NNSA did during prior hardware ecosystem transitions. Without such engagement, NNSA may be missing the next computer revolution.

Second, NNSA must expand (and create where necessary) integrated teams that identify the key algorithmic and data access motifs in its applications and begin collaborative ab initio hardware development of supporting chiplet accelerators, working in concert with both startups and larger vendors. More than incremental code refactoring, this must be a first principles approach that considers alternative mathematical models to account for the limitations of weak scaling. This "beat them" strategy acknowledges that targeted, custom hardware specialization is required to meet NNSA's future HPC performance needs, something the mainstream market alone is increasingly unlikely to provide.

Last, the NNSA procurement model must change. Fixed-price contracts for future, not-yet-developed products have created undue technical and financial risks for NNSA and vendors alike, as the technical and delivery challenges surround early exascale systems illustrate. A better model embraces current market and technical realities, building on true, collaborative co-design involving algorithms, software, and architecture experts that harvests products and systems when the results warrant. However, the committee emphasizes that these vendor partnerships must be much deeper, more collaborative, more flexible, and more financially rewarding for vendors than past vendor R&D programs. This will require a cultural shift in DOE's business and procurement models.

OVERARCHING FINDING: The combination of increasing demands for computing with the technology and market challenges in HPC requires an intentional and thorough evaluation of ASC's approach to algorithms, software development, system design, computing platform acquisition, and workforce development. *Business-as-usual will not be adequate.*

FINDING 2: The computing technology and commercial landscapes are shifting rapidly, requiring a change in NNSA's computing system procurement and deployment models.

> **FINDING 2.1:** Semiconductor manufacturing is now largely in the hands of offshore vendors who may experience supply-chain risk; U.S. sources are lagging.
>
> **FINDING 2.2:** All U.S. exascale systems are being produced by a single integrator, which introduces both technical and economic risks.
>
> **FINDING 2.3:** The joint ECP created a software stack for moving systems software and applications to exascale platforms, but although DOE has issued an initial call for proposals in 2023, there is not yet a plan to sustain it.

FINDING 2.4: Cloud providers are making significant investments in hardware and software innovation that are not aligned with NNSA requirements. The scale of these investments means that they have a much greater market influence than NNSA in terms of both technology and talent.

Given the changing ecosystem and hardware markets, it is critical that NNSA rethink its approach to hardware and software acquisition; its market influence, already small, is declining rapidly. All of these shifts are convolved with global concerns about economic, political, and national security dependence on offshore semiconductor fabrication facilities and the slowing or end of Moore's law. As device feature sizes approach 1 nanometer, vendors have increasingly shifted to chiplet designs—mixing and matching multiple chips on a single substrate—to minimize costs and maximize yields. This creates both opportunities to integrate a custom chiplet suited for NNSA workloads and the consequent challenges of system heterogeneity and software support.

It would be a mistake to claim that NNSA's stockpile stewardship computational models require completely sui generis components, but neither are they fully compatible with the economic mainstream. Substantial customization, with the concomitant nonrecurring engineering costs, will be essential if NNSA is to raise application efficiency and elevate sustained performance. Much as NNSA once worked collaboratively with vendors such as IBM and Cray to design and develop custom computing systems matched to NNSA needs, NNSA must again embrace collaborative ab initio system design, rather than specification development and product procurement.

Such a model is likely to require more internal expertise in computer architecture, greater embrace of cloud software models, specification of novel and semi-custom architectures, end-to-end hardware prototyping at substantial scale for evaluation and testing, and partnership with nontraditional hardware and software vendors, notably AI and other hardware startups and cloud vendors. Most importantly, this new approach will almost certainly require (1) a new mix of laboratory staff skills, (2) continuous prototyping, and (3) substantially more investment than has been true in the past.

RECOMMENDATION 1: NNSA should develop and pursue new and aggressive comprehensive design, acquisition, and deployment strategies to yield computing systems matched to future mission needs. NNSA should document these strategies in a computing roadmap and have the roadmap reviewed by a blue-ribbon panel within a year after publication of this report and updated periodically thereafter.

RECOMMENDATION 1.1: The roadmap should lay out the case for future mission needs and associated computing requirements for both open and classified problems.

RECOMMENDATION 1.2: The roadmap should include any upfront research activities and how outcomes might affect later parts of the roadmap—for example, go/no-go decisions.

RECOMMENDATION 1.3: The roadmap should be explicit about traditional and nontraditional partnerships, including with commercial computing and cloud providers, and academia and government laboratories, and broader cross-government coordination, to ensure that NNSA has the influence and resources to develop and deploy the infrastructure needed to achieve mission success.

RECOMMENDATION 1.4: The roadmap should identify key government and laboratory leadership to develop and execute a unified organizational strategy.

3

Research and Development Priorities

Research and development (R&D) activities have been a critical piece of the National Nuclear Strategy Administration (NNSA) strategy in Science-Based Stockpile Stewardship, including the development of better mathematical models, numerical algorithms, parallel programming tools, high-performance computing (HPC) operating systems, and more. These investments include higher-risk research activities to explore new approaches, and are key to developing more sophisticated computation models, addressing computing technology challenges, and attracting top talent into the NNSA program.

MATHEMATICS AND COMPUTATIONAL SCIENCE R&D

As discussed in Chapter 1, NNSA has significant requirements for HPC beyond exascale to support efforts on all aspects of the weapons life cycle, including design, production, certification, and safety issues. The overall drivers for these requirements stem from pursuing future considerations of possible new designs, new manufacturing processes, aging of the stockpile, and potential needs for rapid response to new global threats. As discussed in Chapter 2, the requirements for increased HPC must be addressed against a backdrop of both technological and ecosystem disruptions. This section discusses the role of applied mathematics and computational science R&D in delivering the computational capabilities needed to meet NNSA mission requirements. Applied mathematics and computational science have played a significant role in the development of a wide range of simulation technologies that have been critical to the success of the Advanced

Simulation and Computing program,[1] a role that the committee anticipates will become increasingly important in the future. However, it is also important to appreciate that applied mathematics and computational science have advanced computational capabilities "beyond forward simulation," including large-scale optimization (for design and control), solution of inverse problems (for parameter estimation), and uncertainty quantification. Continued advances in these areas are required to support future mission needs across all aspects of the NNSA life cycle, from discovery, design exploration, and optimization, to manufacturing and certification, as well as deployment and surveillance.

Applied Mathematics and Computational Science to Enable Forward Simulations for NNSA Mission Problems

Simulation plays a critical role in scientific discovery and forms the backbone of engineering design and analysis. High-fidelity simulations are critical for the design of nuclear explosive packages (NEPs) as well as supporting other aspects of the weapons life cycle. Original designs of nuclear weapons were tightly coupled to tests. Early simulations were calibrated to test data with little predictive capability outside a narrow envelope defined by available data. Requirements to explore new design concepts and to ascertain reliability of the existing stockpile in the absence of testing has driven a need for improved predictive capability. For the past three decades, increasing computational capability, primarily in terms of number of processors, has driven a wave of weak-scaling-based advances where larger HPC systems allowed researchers to solve larger problems with higher-fidelity physics models. NNSA effectively exploited this trend, developing a sophisticated simulation methodology that significantly improved the quality of simulations. Applied mathematics and computational science contributed to this success, developing robust, accurate, and scalable discretizations for complex multiphysics applications to effectively utilize massively parallel HPC architecture and verification, validation, and uncertainty quantification (VVUQ) methodology targeted at assessing the fidelity of simulation results.

In spite of major advances in simulation capability, many problems remain that are beyond reach even with exascale computing. Examples of these types of problems include not only high-fidelity simulations of NEPs, but also simulations of more fundamental problems in weapons science that inform models in full-system simulations. Examples of the latter, as discussed earlier, include modeling of high explosives, behavior

[1] It is well documented that advances in algorithms over the past decades have led to computational speedups that have paralleled the exponential growth in computing power according to Moore's law. U. Rude, K. Willcox, L.C. McInnes, and H. De Sterck, 2018, "Research and Education in Computational Science and Engineering," *SIAM Review* 60(3):707–754.

of materials under extreme conditions, turbulence and turbulent mixing, and radiation/matter interaction.

A fundamental concern in many areas such as the examples mentioned above is that relying on a weak-scaling approach based on larger versions of architectures currently being deployed at the exascale is no longer a viable strategy. Increasing fidelity by increasing spatial resolution necessitates an increase in temporal resolution. Consequently, even if one assumes ideal scaling, this weak-scaling paradigm results in increasing time to solution. For example, increasing the resolution of a three-dimensional high-explosive detonation simulation by an order of magnitude will increase resources needed to store the solution by three orders of magnitude and increase the time to solution by an order of magnitude (or more if any aspect of the simulation fails to scale ideally). For simulations that require several weeks or more to complete now, the time-to-solution for a higher-resolution version of the same problem will be on the order of a year or more for a single simulation, making it, for all practical purposes, infeasible. Strong scaling, in which additional resources are used for a fixed-size problem, has been shown to have only limited success because performance quickly becomes limited by communication costs. Several of the problems mentioned above, as well as many of the engineering analyses relevant to the complete weapons life cycle, share this characterization.

This breakdown of the weak-scaling paradigm coupled with the disruptions in NNSA computing discussed in Chapter 2 will necessitate rethinking how simulations are performed. Advancing the state of the art in forward simulation will rely on significantly different algorithmic approaches and methods to exploit novel hardware advances effectively. Improved methodology for traditional simulation such as more accurate discretization, faster solvers, and better coupling approaches will undoubtedly play a role, but these need to be augmented with other types of approaches.

Exactly what other approaches will ultimately prove useful remains an open question. One potential area would be the incorporation of machine learning or other types of statistical data-driven approaches. The basic idea would be to replace some computationally expensive component of a simulation with a relatively inexpensive data-driven model based on either experiment or data computed from a separate simulation. Although potentially expensive to train, the resulting model could dramatically reduce simulation costs as discussed in the artificial intelligence (AI) R&D section later in this chapter. Another potential approach would be to develop multiscale techniques for different processes that are able to capture the effects of finer-scale behavior on the larger-scale dynamics, reducing the range of scales needed for simulations. In both cases, quantifying the fidelity of the new model and how it impacts the uncertainty of the overall simulation needs to be assessed.

Another issue that arises within the context of improved forward simulations is the need for improved models for key physical processes (discussed in Chapter 1). Data-driven multiscale approaches based on finer-scale simulations are needed to systematically define coarse-grained models for use in systems-level simulations with quantified fidelity. Of particular interest in this context are situations where there is important fine-scale behavior that cannot be readily captured by larger-scale models. Microstructure, small voids, and the presence of cracks have significant impact on the dynamic response of solids, which can alter the behavior of solid high explosives as well as other aspects of the weapons systems. Traditional models for fluid mechanics fail to accurately predict the internal structure of strong shock waves. Simulating systems of this type will require simulation methodology that uses different physical descriptions at different scales with systematic ways to identify what model is appropriate in a given part of the problem and to couple different types of representations dynamically.

VVUQ is also a critical element of NNSA mission problems and becomes even more important with the increased use of data-driven modeling and machine learning. Models must have quantified fidelity—with a clearly defined metric of what it means to be trusted—and the impact of an individual model's fidelity on overall simulation accuracy must be characterized. While the AI methods discussed in the section on AI R&D later in this chapter may provide new opportunities to augment, enhance, and accelerate physics-based simulators, their overall utility will be severely limited without rigorous VVUQ.

Meeting these challenges will require significant investment in applied mathematics and computational science. However, it is also important to recognize that this investment must support a broader range of mathematical sciences than in the past. There is a continued need for the mathematics that supports multiphysics, scalable algorithms, but also an increasing need for the mathematics that supports multiscale modeling, machine learning, AI, and statistical and data-driven modeling. There is also a need to address the challenges and opportunities of integrating data-driven models with traditional simulation methodologies to develop more effective predictive capabilities, recognizing the essential role that VVUQ must play in NNSA applications.

FINDING 3: Bold and sustained research and development investments in hardware, software, and algorithms—including higher-risk research activities to explore new approaches—are critical if NNSA is to meet its future mission needs.

> **FINDING 3.1:** Physics-based simulators will remain essential as the core of NNSA predictive simulation. However, given disruptions in computing technology and the HPC ecosystem combined with the end of the weak-scaling era, novel

mathematical and computational science approaches will be needed to meet NNSA mission requirements.

FINDING 3.2: VVUQ and trustworthiness remain of paramount importance to NNSA applications. VVUQ will become increasingly important as simulation methodology shifts toward more complex systems that incorporate models of different fidelity, including data-driven approaches.

Algorithms for Novel Architectures

Future computer architectures are expected to be considerably more heterogeneous than today's systems. Systems may incorporate a variety of different capabilities such as accelerators designed for machine learning. Applied mathematics research will be needed to develop new algorithmic approaches to effectively utilize these novel architectures and integrate those approaches into multiphysics/multiscale simulations. Custom-designed hardware targeted toward specific applications provides another novel approach to obtain improved performance. In this case, co-design of hardware and algorithms will be essential. Designing effective custom hardware will require a close partnership between hardware architects and applied mathematicians and computational science researchers. The emergence of highly heterogeneous architectures will also drive a need for theory and methods to achieve optimal management of heterogeneous models/data over hierarchical and distributed compute and network resources.

FINDING 3.3: Novel architectures can have a significant impact on NNSA computing; however, mathematical research will be needed to effectively exploit these new architectures. Involvement of applied mathematicians and computational scientists early in the development cycle for novel architectures will be important for reducing development time for these types of systems.

FINDING 3.4: An end to transistor density scaling is likely to motivate industry to develop novel computer architectures for which today's numerical algorithms, software libraries, and programming models are ill suited.

Applied Mathematics and Computational Science Beyond Forward Simulation

As noted above, computational science has long encompassed more than just forward simulation. The past decades have seen advances in large-scale optimization, inverse problems, and uncertainty quantification–sometimes referred to as "outer-loop" applications of computational science because they require many forward model evaluations[2,3,4]—with impact on NNSA design and VVUQ workflows. As the sophistication of predictive simulations continues to increase, there is a corresponding need to advance the mathematics of these outer-loop methods. Even as high-fidelity multiphysics simulations continue to mature, they are typically much too expensive to be used routinely in outer-loop applications such as design optimization, control, or autonomous experimentation. Surrogate and reduced-order modeling have received considerable attention in the past decades in the applied mathematics and computational science research communities, and provide a class of approaches that lead to increased simulation speed to address these challenges; however, it remains an outstanding mathematical challenge to quantify the limitations of these types of models and characterize their overall fidelity, especially for nonlinear multiphysics systems. As discussed in the section on AI R&D later in this chapter, AI-based approaches provide exciting opportunities to take computational technologies such as surrogates to a new level, but continued investment in applied mathematics and computational science that have physics-based modeling at their core remains essential.

Another nontraditional application of computation within NNSA is support for experimental facilities. As the capabilities of NNSA experimental facilities continue to advance, HPC can play a major role in experimental design, optimizing experimental controls, and analyzing the flood of data being generated by these facilities. Mathematics in support of facilities is still in its infancy and substantial development will be needed to realize this potential.

Integration of predictive simulation capabilities together with experimental data paves the way for digital twins, another key opportunity area for NNSA. A digital twin is a computational model or set of coupled models that evolves over time to persistently represent the structure, behavior, and context of a unique physical system or process.[5] Digital twins are characterized by a dynamic and continuous two-way flow of information between the computational model and the physical system. Data streams from the physical system are assimilated into the computational model to reduce uncertainties and improve its predictions, which in turn is used as a basis for controlling the physical system, optimizing data acquisition, and providing decision support. Digital twins have the potential to support the NNSA mission in a number of ways, including monitoring the condition of the stockpile, real-time monitoring, and adaptive control of

[2] D.E. Keyes, 2011, “Exaflop/s: The Why and the How,” *Comptes Rendus Mécanique* 339(2):70–77, https://doi.org/https://doi.org/10.1016/j.crme.2010.11.002.

[3] Department of Energy, 2014, *DOE Advanced Scientific Computing Advisory Subcommittee (ASCAC) Report: Top Ten Exascale Research Challenges*, February 10, https://www.osti.gov/servlets/purl/1222713.

[4] B. Peherstorfer, K. Willcox, and M. Gunzburger, 2018, “Survey of Multifidelity Methods in Uncertainty Propagation, Inference, and Optimization,” *SIAM Review* 60(3):550–591, https://doi.org/10.1137/16m1082469.

[5] AIAA Digital Engineering Integration Committee, 2020, “Digital Twin: Definition and Value,” AIAA and AIA position paper, American Institute of Aeronautics and Astronautics (AIAA) and Aerospace Industries Association (AIA), https://www.aiaa.org/advocacy/Policy-Papers/Institute-Position-Papers.

manufacturing processes, and improving control of hypersonic vehicles. While digital twins provide an exciting opportunity to drive improved decision making, realizing a digital twin at the scale, fidelity, and level of trustworthiness required for NNSA mission problems requires investment to address foundational applied mathematical and computational science challenges. Such challenges include managing and representing data, models, and decisions that cross multiple temporal and spatial scales; predictive modeling of complex systems that comprise multiple interacting subsystems; VVUQ for predictive digital twins; scalable algorithms for data assimilation, prediction, and control; and integrating complex data streams within the digital twin.[6]

> **FINDING 3.5:** Recent advances in applied mathematics and computational science have the potential for impact on NNSA mission problems far beyond traditional roles in physics-based simulation.

RECOMMENDATION 2: NNSA should foster and pursue high-risk, high-reward research in applied mathematics, computer science, and computational science to cultivate radical innovation and ensure future intellectual leadership needed for its mission.

> **RECOMMENDATION 2.1:** NNSA should strengthen efforts in applied mathematics and computational science research and development. Potential areas include using novel architectures, data-driven modeling, optimization, inverse problems, uncertainty quantification, reduced-order modeling, multiscale modeling, mathematical support for experiments, and digital twins.

COMPUTER SCIENCE R&D

In this section, the committee considers the computer science R&D questions, which are driven from below by the technology considerations and from above by the application and algorithm requirements. The first set of issues are based on the role of computer science research in co-designing future systems, methods, and applications with an even deeper understanding of the technology and ecosystem constraints than in previous generations. The following section looks at computer science research beyond traditional high-performance modeling and simulation. The last section acknowledges the

[6] S.A. Niederer, M.S. Sacks, M. Girolami, and K. Willcox, 2021, “Scaling Digital Twins from the Artisanal to the Industrial,” *Nature Computational Science* 1(5):313–320, https://doi.org/10.1038/s43588-021-00072-5.

important but separate problem of maintaining a robust software engineering program in the future.

Co-Design of Future High-Performance Computing Systems

More than 50 years of advances in HPC have taught us that each generation of HPC systems is accompanied by major new challenges of scale and complexity. Overcoming these challenges often necessitates changes in algorithmic approaches, systems software architecture, programming models, compilation techniques, and application development methods. For example, although the message passing interface remains an important and stable part of the HPC programming environment, changes in node architecture, memory systems, resilience characteristics, and usage models continue to require innovations in programming systems. Furthermore, if device-level performance benefits stall and architectural specialization becomes commonplace—as it has for machine learning—then NNSA will need to take a leadership role in the design and testing of scientific computing hardware and the development of a retargetable software stack.

Co-design has been a tenet of the HPC approach leading to exascale but has highly leveraged commodity components and placed much of the burden of hardware and systems software on vendors. This balance will shift toward the laboratories in the post-exascale era, likely requiring them to take on hardware accelerator design, system integration, and expanded system software development. Such roles are not new. In the past, NNSA has been a leader in transitioning supercomputing technologies and architectures from vector supercomputers to microprocessor-based parallel computers in the 1990s, multicore nodes in the 2000s, and graphics processing units (GPUs) in the 2010s. NNSA laboratories have also built HPC operating systems, developed compilers to support standard and novel programming languages, designed runtime systems to manage different types of parallelism, created communication libraries for both production and exploratory use, and built autotuners to intelligently search through possible implementations to find one best suited to a given piece of hardware. They have also deployed systems with low-power processors customized for HPC (e.g., the BlueGene line), with server processors adapted to add lightweight communication, and with processors from gaming or graphics markets (RoadRunner and GPU-based systems). The decisions behind each of these deployments may seem obvious in retrospect, but they required visionary leaders who were able to forge a path amid technology and business risks and achieve success with the support of creative research teams who solved anticipated and unanticipated challenges to effectively utilize each generation of systems.

Research into algorithms that exploit new levels or degrees of parallelism, avoid data movement at any level of the memory and communication hierarchy, tolerate various kinds of hardware failures, or dynamically balance computational or communication load are not stand-alone computer science problems but must often take into account higher-level application structures and mathematics. For example, optimized versions may preserve all of the original program dependencies and therefore leave numerical properties unchanged or may rely on properties like associativity of addition, which is true in exact arithmetic but not when floating-point roundoff is considered, or they may compute different but equally useful answers. Computations moved to hardware accelerators may not compute the same result as the main processor (central processing unit), especially if narrower data types are used (e.g., 8-bit floating-point formats are now being considered for machine-learning algorithms). Performance optimizations must also take into account hierarchical data structures, matrices that are so sparse (filled with mostly zeros) that only nonzero values and their locations are stored, or unstructured meshes representing the intricacies of a complex mechanical device.

Addressing foreseeable technology challenges requires a spectrum of high- and low-risk approaches and the technical expertise to design, build, and resolve challenges that arise. There are many open research questions. For example, can NNSA leverage AI hardware for non-AI workloads? Even if the AI architectures cannot be used, can some features, such as low-precision arithmetic, save memory and increase computational rates on NNSA problems? Will future semiconductor devices and packaging require new algorithms in response to changes in the relative cost of memory and computing operations—for example, devices that reduce energy consumption at the cost of higher latency. Are there new classes of parallel algorithms or cost models better suited to future machines? Can machine learning be used to produce performance portable software? How will detectable system failures and silent errors affect future post-exascale systems, and will they require new algorithms and software? Should new storage technology be integrated into hardware-controlled memory hierarchy or exposed to software control, and at what benefit to NNSA applications? Should NNSA computing infrastructure be configured to be resilient to natural and human-induced disasters? Are there ideas, tools, or lessons from cloud computing (specifically, PaaS [platform as a service] and SaaS [software as a service] models) that can aid in answering these questions?

The technology challenges facing the field of computing writ large will require advanced research on all levels of computer system design and use, from semiconductor device technology and computer architectures to the programming tools and abstract cost models used to design efficient algorithms. To ensure that NNSA has access to highly capable, world-leading computing systems that are suitable to their future workloads, they will need to consider much more aggressive models of co-design and strategically partner with industry, universities, and other laboratories.

FINDING 3.6: Co-design of hardware and systems for high-performance scientific computing applications has been a modest success to date and will be more important in the future and need to be deeper. Technological and market trends are likely to shift the balance of co-design to the laboratories, requiring more innovation and engineering in the areas of hardware design, system integration, and system software.

Computer Science Research Beyond High-Performance Computing

The DOE laboratories are among the world leaders in high-performance computing, especially as it pertains to modeling and simulation. Other areas of computer science research such as networking, distributed systems, computer architecture, cybersecurity, user interfaces, databases, graphics, and software engineering have relevance to NNSA problems but are not as well represented in the laboratories. For example, NNSA supports work on languages like FORTRAN and C++, but major innovations in high-productivity scientific computing for non-HPC problems, such as Python and Julia, as well as new models for collaborative science such as Jupyter notebooks, have come primarily from outside groups. These technology developments shape how students are trained and how data-intensive scientific research is conducted outside the laboratories, raising the possibility that future generations of weapons designers will demand higher-level programming interfaces and semi-automated tools. Methods for synthesizing code or hardware designs from higher-level specifications or generating test sets automatically using program verification techniques are largely absent in laboratory research, as are new models of wide area networking, hardware support for secure computing, and platforms for cleaning and analyzing large, messy data sets.

There are also many problems related to the management and analysis of large-scale data sets from experiments (such as the National Ignition Facility, the Dual Axis Radiographic Hydrodynamic Test Facility, the Z Machine, as well as other experiments) and simulations. The challenges of analyzing massive scientific data sets are compounded by data complexity that results from heterogeneous methods and devices for data generation and capture and the inherently multiscale, multiphysics nature of many sciences, resulting in data with hundreds of attributes or dimensions and spanning multiple spatial and temporal scales.[7] Research in the management and analysis of extreme-scale scientific data may overlap with HPC, but with different hardware requirements than modeling and simulation applications. For example, the need for high-speed input/output drives both hardware configuration, raises opportunities for new storage technologies, and requires the operating systems and system libraries to effectively use such a system. The

[7] Office of Science Financial Assistance, 2010, "Scientific Data Management and Analysis at Extreme Scale," https://science.osti.gov/ascr/Funding-Opportunities/-/media/grants/pdf/foas/2010/DE-FOA-0000256.pdf.

data rates from future experiments and simulations, as well as feedback between them, may require new tools and methods for data management and data analytics, including visual analysis, and scientific workflow tools for scientific discovery.

Even more striking, machine-learning algorithms and quantum algorithms described in later sections are not historical areas of strength, although both are growing within the laboratories. This delay emphasizes the need to have a vibrant research program, allowing for progress on known research challenges, but also allowing for the exploration of completely different approaches to the high-level mission problems in NNSA.

Software Development Is Not Computer Science Research

The practice of high-quality software engineering is essential to producing and maintaining computer applications and the underlying levels of software that can reliably make predictions about weapons systems and various component problems. There is a natural tendency to equate software engineering practice with computer science research. Software engineering practice is about reducing risk through use of known tools and techniques, defining clear interfaces, and adhering to standards, rigorously testing and documenting code and having a robust process for software management and releases. Computer science research, even in software engineering, is about exploring new ideas, testing hypotheses, and taking risks. NNSA needs a cadre of both software engineers and computer science researchers, as each plays a distinct role in meeting NNSA's mission.

Numerical libraries are important for many high-performance scientific applications and offer the potential to exploit the underlying computer systems without the application developer understanding the architectural details. Existing numerical libraries will need to be rewritten and new algorithms developed in light of emerging architectural changes, including increased concurrency, heterogeneous components, power management, and multiple types of memory. Because of the enhanced levels of concurrency on future systems, algorithms will need to embrace asynchrony to generate the number of required independent operations. New and evolving application requirements also require extensions to libraries, just as linear algebra libraries primarily created for simulation have been adapted for machine learning.

> **RECOMMENDATION 2.2:** NNSA should strengthen efforts in computer science R&D to build a substantial, sustained, and broad-based intramural research program that is positioned to address the technological challenges associated with post-exascale systems and co-design of those systems to ensure that the laboratories are positioned for leadership in computing breakthroughs relevant to NNSA mission problems.

ARTIFICIAL INTELLIGENCE RESEARCH AND DEVELOPMENT

The computational demands of machine-learning applications and the availability of optimized hardware for this workload should be considered when planning for post-exascale NNSA computing systems and activities. One important class of machine-learning methods, deep learning, has dominated recent success in AI (Box 3-1) perceptual tasks (e.g., image recognition, image classification, machine translation); demonstrated human-level, or better, skill in areas thought to require deep expertise (e.g., Go and chess); and produced intriguing results in scientific problems (e.g., protein structure prediction[8]) and other areas (e.g., automated generation of text and software using large language models[9]). A series of town halls led by the Department of Energy laboratories in 2019 on AI for Science (AI4Sci) and in 2022 on AI for Science, Energy, and Security (AI4SES) covered many of the opportunities and challenges of using AI in science, energy, and security applications, including a report produced for the earlier AI4Sci meetings.[10] The security aspects of these 2022 meetings covered problems of relevance to NNSA, although necessarily limited in scope owing to the unclassified nature of the meetings. The NNSA laboratories have also had internal discussion about the use of these methods.

The scale and speed of advances in AI applications have been remarkable, but equally important is a growing understanding that not all problems have sufficient observational data or the necessary constraints for automated training and that open problems remain when using AI in complex environments with multiple physics constraints or for safety-critical problems requiring strict confidence metrics. In these situations, AI methods may be used in concert with traditional simulations. For example, neural networks can be trained on data from simulations to produce surrogates to computational functions (or even entire simulations), achieving nonlinear improvements of multiple orders of magnitude in time-to-solution for HPC applications. Such surrogates can be used to accelerate design space exploration for problems such as materials.[11] As this example illustrates, AI methods can augment conventional computational simulation, enabling new approaches to old problems and providing paths to tackling previously intractable problems. From a workload perspective, both high-fidelity simulations and AI methods must be supported.

The recent success of deep-learning methods in the AI community can be attributed in large part to the growth in computing performance in the past few decades. The

[8] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, et al., 2021. "Highly Accurate Protein Structure Prediction with AlphaFold," *Nature* 596(7873):583–589.

[9] "Introducing ChatGPT," https://openai.com/blog/chatgpt.

[10] R. Stevens, V. Taylor, J. Nichols, A.B. MacCabe, K. Yelick, and D. Brown, 2020, "AI for Science," Argonne Scientific Publications, Argonne National Laboratory, https://www.anl.gov/ai-for-science-report.

[11] A. Agrawal and A. Choudhary, 2019, "Deep Materials Informatics: Applications of Deep Learning in Materials Science," *MRS Communications* 9(3):779–792.

BOX 3-1 Artificial Intelligence and Machine Learning Terminology

The term *artificial intelligence*, or *AI*, is used broadly to refer to the design and construction of intelligent agents that realize perceptual and goal-seeking activities, and includes computer vision, speech processing, and more.[a] Confusingly, AI has also become synonymous in many circles with *machine learning*, a class of methods often used in AI that "learn" via automated extraction of models, either from training data or within a constrained setting such as optimizing for game play. Yet more specifically, AI is sometimes used to refer to deep learning, the subclass of machine-learning methods based on multilayer neural networks, which have achieved success in AI problems such as computer vision, speech recognition, natural language translation, robotics, and playing games of strategy.

Applications of deep learning have also had tremendous commercial impact, resulting in a market for so-called AI hardware. While graphics processing units are commonly used for deep learning, these more specialized AI chips are optimized for matrix and tensor operations, often using low-precision arithmetic.

In a report for the AI4Sci town halls, the AI terminology is very broad: "In this report and in the Department of Energy laboratory community, we use the term 'AI for Science' to broadly represent the next generation of methods and scientific opportunities in computing, including the development and application of AI methods (e.g., machine learning, deep learning, statistical methods, data analytics, automated control, and related areas) to build models from data and to use these models alone or in conjunction with simulation and scalable computing to advance scientific research."[b]

[a] S.J. Russell and P. Norvig, 2021, *Artificial Intelligence: A Modern Approach*, 4th ed., Hoboken, NJ: Pearson.

[b] R. Stevens, V. Taylor, J. Nichols, A.B. MacCabe, K. Yelick, and D. Brown, 2020, "AI for Science," Argonne Scientific Publications, Argonne National Laboratory, https://www.anl.gov/ai-for-science-report.

demand for large-scale, highly optimized computing systems for deep-learning applications has already motivated substantial use of DOE HPC systems for model training on scientific problems, from the exploration of novel materials and cancer treatments to the identification of extreme climate events and rare astronomical phenomena. The NNSA laboratories are exploring the use of machine learning methods and hardware optimized for deep learning. This research needs to continue, and if AI proves to be broadly applicable, it should be a part of the workload used to design and select future computing systems.

Opportunities

There are several areas for exploration of AI methods in the NNSA mission, using the AI term broadly as in the AI4Sci report. In some cases, AI may provide a solution to replace manual processes, to augment traditional simulations, or to provide useful tools to aid human decision makers. Examples of AI-enabled capabilities that could advance the NNSA mission include the following:

- AI as surrogates for simulations of physical systems, ranging from practical problems such as optimized representations of neutron group structures, as already demonstrated by the laboratories, to building emulators for the three-dimensional evolution of engineered systems—a capability that is not realizable today and is estimated to require hundreds of exaflop years to train with current methods.

- AI in the loop for automated adaptive real-time control of experimental facilities and manufacturing processes that currently do not admit to real-time control or that require increasingly scarce human expertise owing to an aging workforce.
- AI to connect across the life cycle for end-to-end intelligent decisions, such as designing to explicitly account for manufacturing and aging, rather than suboptimal proxies for manufacturing/aging issues.
- AI to accelerate all stages of the complex physics cycle, encompassing hypothesis, design, execution/control, diagnosis, and analysis. For example, AI-based surrogates may be used to screen large collections of candidates (e.g., materials, structures) that cannot reasonably be evaluated via conventional simulation.
- AI to shorten the time to solution at all stages of nuclear weapon design and deployment: Discover, Design, Manufacture, Deploy with AI injection at all phases.
- AI for managing surveillance via automated analysis of multimodal data, leveraging, for example, self-supervised learning methods to detect unusual events.
- AI for enabling digital twins that integrate heterogeneous models and data from multiple sources, while leveraging edge computing and integration across edge/HPC.

A crosscutting theme is the use of AI methods to learn previously unknown relationships among entities (“serendipitous models”[12]) that can be evaluated far faster than by conventional means and/or without explicit programming—in the process, automating and accelerating previously manual steps to enable more rapid exploration of far larger design spaces.

FINDING 3.7: Rapid innovation in AI methods, driven by advances in computing performance and growth in data sets, is producing frequent technological surprises that NNSA should continue to investigate and track. These advances may

[12] K.E. Willcox, O. Ghattas, and P. Heimbach, 2021, “The Imperative of Physics-Based Modeling and Inverse Theory in Computational Science,” *Nature Computational Science* 1:166–168. https://doi.org/10.1038/s43588-021-00040-z.

benefit the NNSA mission but will likely complement rather than replace traditional physics-based simulations in the post-exascale era.

Challenges

Realizing these *potential* advances will be far from straightforward. The following factors, in particular, tend to make the application of AI methods to NNSA problems challenging:

- VVUQ is paramount in NNSA applications. Explainability, predictive capability, and trust are essential. For example, if AI-based surrogates are used directly to issue predictions and support decisions, they must use techniques that capture the underlying physics, including statistical properties, with known levels of confidence.
- Data are typically sparse and indirect in NNSA applications. Many problems are data-poor. Sensing technology is advancing, but many problems within the NNSA mission will never have abundant experimental data (e.g., previously archived nuclear tests). Simulation data may be explored for training, but the cost of generating sufficient simulation-based training data may be prohibitive with current methods and confidence levels considered carefully.
- NNSA applications often involve complex systems that engage multiple physics across multiple scales. Coupling among components and between physical phenomena can lead to nonlinear (emergent) behavior. Nonlinearities are often most severe in the most critical conditions (e.g., conditions approaching failure).
- Many NNSA applications are characterized by long life cycles, which from concept to design to manufacturing to deployment can span decades. The challenges of computational modeling of multiscale, multiphysics complex systems are exacerbated by the long time horizons over which predictions must remain accurate. For example, computational techniques that support surveillance applications must accurately characterize and predict system performance over decadal time scales.
- Rare events drive decision making in weapons design. Data around rare events (e.g., failures) are typically sparse, indirect, and expensive to acquire. Rare events also pose the largest challenges for predictive simulations. Acceptable probabilities of weapon system failure are typically orders of magnitude smaller than for mainstream AI applications.
- The classified nature of many elements of the NNSA mission is likely to hinder the large-scale sharing and aggregation of data across components and life-cycle elements that will be important for effective AI, especially when

combined with other institutional drivers for data silos (organizational, security, proprietary).

- The development and application of AI methods in the NNSA context will require new data infrastructure and AI-ready instrumentation that can interface with the rest of the AI ecosystem.[13] For example, manufacturing data availability, data resources, and data management processes must be advanced in order to realize the benefits of machine learning in enabling manufacturing for NNSA applications, especially as new sensors produce data at unprecedented rates.

Overcoming these challenges to realize capabilities such as those sketched earlier will require sustained investment in both foundational and applied AI R&D. Workshops, such as AI4SES, are an important opportunity within the post-exascale computing landscape, but they cannot be pursued in isolation. Future methods to advance AI4SES workshops must capitalize on advances in mainstream AI while at the same time deeply integrating physics and VVUQ via the computational science methods of the section on Mathematics and Computational Science R&D earlier in this chapter. Opportunities at the intersection of AI and computational science abound. The NNSA laboratories have been exploring the use of AI for certain mission problems, but the combination of enormous opportunities and enormous unanswered questions suggests that the current level of effort is insufficient.

RECOMMENDATION 2.3: NNSA should expand research in AI to explore the use of these methods both for predictive science and for emerging applications, such as manufacturing and control of experiments, and develop machine learning techniques that provide the confidence in results required for NNSA applications.

QUANTUM COMPUTING AND QUANTUM TECHNOLOGY R&D

In presentations to the study committee, the national laboratories asserted that quantum technology (Box 3-2) may play an important future role in enhancing the advancement of their mission-driven computational requirements, but that most of that impact was near the end timeframe of this study. Specifically, the Lawrence Livermore National Laboratory (LLNL) team suggested (see Figure 3-1) hybrid techniques that might be useful for calculation of physics model inputs such as equation of state, transport coefficients, and

[13] "Complex Physics Report-Out," from AI4SES Workshop, June 2022.

BOX 3-2 Quantum Computing

It is important to recognize that the phrase "quantum computing" is used to refer to a variety of technologies with markedly different computational abilities and expected dates of availability for use. Most familiar (with availability furthest in the future) is what is called error-corrected gate-level quantum computing. This technology offers the ability to execute algorithms, most famously Shor's algorithm for factoring numbers, and is called a fault-tolerant quantum computer. Because of the technical challenges in realizing such a computer, scientists are developing two other technologies that may have applicability sooner. One such technology was envisioned by Richard Feynman in 1981, and essentially uses a quantum computer to mimic the behavior of complex physical molecules and then use the quantum computer's ability to observe the quantum behavior of the molecule. One method to mimic such molecular behavior is analog quantum simulation, in which the analog quantum behavior of a machine is used to model a molecule. Last, there is a new model of computing being investigated by researchers called noisy intermediate-scale quantum computers, which tries to use hybrid combinations of non-error-corrected gate-level quantum computers and classical computers to achieve some interesting computations. One such interesting computation is an alternative method of emulating molecules in which the attributes of the molecular system are mapped in a more controlled way onto a digital quantum simulation. Mappings of more general optimization problems are also possible, leading to potential applications in areas such as logistics and finance. An accessible discussion of these concepts can be found in Preskill's 2021 paper, "Quantum Computing 40 Years Later."[a]

[a] J. Preskill, 2021, "Quantum Computing 40 Years Later," ArXiv 2106.10522v2.

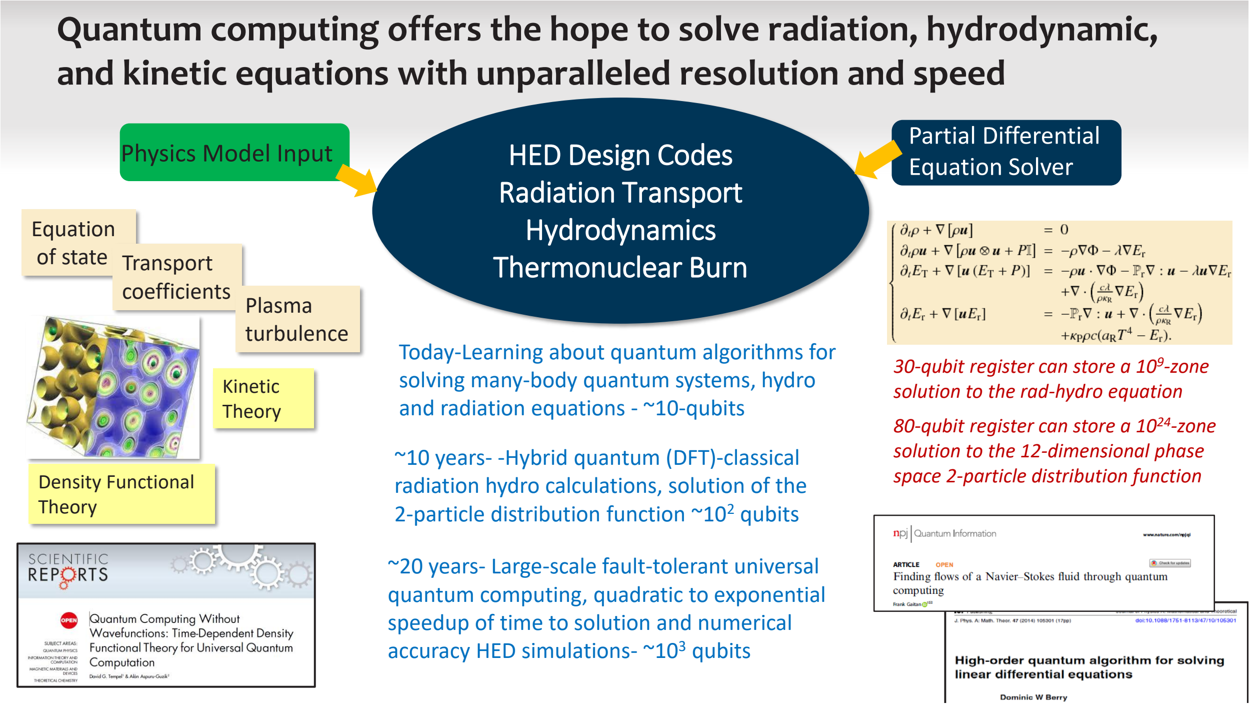
plasma turbulence in about 10 years from our study date. The LLNL team also suggested that partial differential equation solutions could be relevant on fully error-corrected machines in about 20 years (Figure 3-1). The quantum advantage for PDEs is still an open question, but may be in accuracy of solution rather than speed.[14] Los Alamos National Laboratory noted that while no practical speedups have been observed to date, some quantum simulation algorithms show promise. All laboratories highlighted an appropriate level of investment in various areas of quantum research.

Importantly, post-exascale classical HPC will still be required to solve applications in a hybrid classical-quantum computing model in which quantum hardware accelerates key core kernels, while classical computing provides the full solution by integrating and computing upon results from quantum solutions to many small subproblems.

Furthermore, practical quantum applications will likely emerge on quantum machines using a continuum of error-mitigation techniques ranging from partially fault-tolerant to fully fault-tolerant methods. NNSA DOE algorithms and software research should examine this continuum to bridge the gap between current NISQ machines and future fully error-corrected machines.[15]

[14] A.M. Childs, J.-P. Liu, and A. Ostrander, 2021, "High-Precision Quantum Algorithms for Partial Differential Equations," *Quantum* 5:574, https://doi.org/10.22331/q-2021-11-10-574.

[15] National Academies of Sciences, Engineering, and Medicine, 2019, *Quantum Computing: Progress and Prospects*, Washington, DC: The National Academies Press, https://doi.org/10.17226/25196.



Quantum computing offers the hope to solve radiation, hydrodynamic, and kinetic equations with unparalleled resolution and speed

Physics Model Input

HED Design Codes
Radiation Transport
Hydrodynamics
Thermonuclear Burn

Partial Differential Equation Solver

Equation of state

Transport coefficients

Plasma turbulence

Kinetic Theory

Density Functional Theory

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
QUANTUM PHYSICS
INFORMATION THEORY AND COMPUTATION
MAGNETIC MATERIALS AND DEVICES
THEORETICAL CHEMISTRY

Quantum Computing Without Wavefunctions: Time-Dependent Density Functional Theory for Universal Quantum Computation

David G. Tempel[1] & Alán Aspuru-Guzik[2]

Today-Learning about quantum algorithms for solving many-body quantum systems, hydro and radiation equations - ~10-qubits

~10 years- -Hybrid quantum (DFT)-classical radiation hydro calculations, solution of the 2-particle distribution function ~10^2 qubits

~20 years- Large-scale fault-tolerant universal quantum computing, quadratic to exponential speedup of time to solution and numerical accuracy HED simulations- ~10^3 qubits

$$\begin{cases} \partial_t\rho + \nabla\left[\rho\boldsymbol{u}\right] &= 0 \\ \partial_t\rho\boldsymbol{u} + \nabla\left[\rho\boldsymbol{u}\otimes\boldsymbol{u} + P\mathbb{I}\right] &= -\rho\nabla\Phi - \lambda\nabla E_{\mathrm{r}} \\ \partial_t E_{\mathrm{T}} + \nabla\left[\boldsymbol{u}\left(E_{\mathrm{T}}+P\right)\right] &= -\rho\boldsymbol{u}\cdot\nabla\Phi - \mathbb{P}_{\mathrm{r}}\nabla:\boldsymbol{u} - \lambda\boldsymbol{u}\nabla E_{\mathrm{r}} \\ & \quad +\nabla\cdot\left(\frac{c\lambda}{\rho\kappa_{\mathrm{R}}}\nabla E_{\mathrm{r}}\right) \\ \partial_t E_{\mathrm{r}} + \nabla\left[\boldsymbol{u}E_{\mathrm{r}}\right] &= -\mathbb{P}_{\mathrm{r}}\nabla:\boldsymbol{u} + \nabla\cdot\left(\frac{c\lambda}{\rho\kappa_{\mathrm{R}}}\nabla E_{\mathrm{r}}\right) \\ & \quad +\kappa_{\mathrm{P}}\rho c(a_{\mathrm{R}}T^4 - E_{\mathrm{r}}). \end{cases}$$

30-qubit register can store a 10^9-zone solution to the rad-hydro equation

80-qubit register can store a 10^{24}-zone solution to the 12-dimensional phase space 2-particle distribution function

npj | Quantum Information www.nature.com/npjqi

ARTICLE OPEN Check for updates

Finding flows of a Navier–Stokes fluid through quantum computing

Frank Gaitan

J. Phys. A: Math. Theor. 47 (2014) 105301 (17pp) doi:10.1088/1751-8113/47/10/105301

High-order quantum algorithm for solving linear differential equations

Dominic W Berry

FIGURE 3-1 Quantum computing hopes.
SOURCE: From a briefing provided to the committee by LLNL on April 8, 2022.

Sustained research in quantum algorithms, software, and hardware is needed to adapt NNSA applications to a hybrid classical-quantum computing model. In particular, classical optimization algorithms need to be separated into classical and quantum components that best leverage quantum hardware, and the classical component needs to be adapted to the output of the quantum hardware. For example, the classical component can both take advantage of quantum solutions and adjust for errors in the quantum computation.

Within this context, DOE quantum test beds are a key resource for this research, and a diversity of quantum technologies should be made available to scientists in order to future-proof algorithms and software as these technologies develop.

Given the technical and economic limits of scaling classical computing, quantum approaches should be explored to determine if specific problems or subproblems could be solved more efficiently or accurately. However, by the very nature of hybrid classical-quantum calculations, if a significant performance gain is expected, the vast majority of the work must be performed by the quantum kernels and achieving this will require substantial reengineering of software and algorithms.

Last, in order to justify large-scale deployment of quantum accelerators, algorithms and software research is needed to broaden the applications that can benefit from this technology.

FINDING 3.8: Quantum technology has the potential to improve the fundamental understanding of material properties needed by important NNSA applications. Analog quantum simulation or digital quantum simulation will likely be available before general quantum computers.

FINDING 3.9: Major breakthroughs in quantum algorithms and systems are needed to make quantum computing practical for multiphysics stockpile modeling. Quantum computing is more likely to serve as a special-purpose accelerator than to replace today's broadly applicable HPC systems.

RECOMMENDATION 2.4: NNSA should continue to invest in and track quantum computing research and development for future integration into its computational toolkit; these technologies should be considered an additional computational tool rather than a replacement for current approaches.

4

Workforce Needs

The increasingly central role of computing demands a capable and committed workforce. Although the National Nuclear Security Administration (NNSA) has several programs aimed at filling critical gaps in the current workforce, forward-looking investments must be made to envision and support the desirable Advanced Simulation and Computing (ASC) workforce of the future. Many workforce issues faced by NNSA will be faced by the U.S. computing industrial base, and therefore competition for appropriately skilled workers will be intense.

WORKFORCE CHALLENGES

Labor market volatility has been in the news for several reasons lately. In a post-COVID world, there are new perturbations in the job market, with workers reassessing their relationship to, and expectations for, work. At the same time, millennials (born between 1981 and 1996) are changing jobs every few years,[1] making it harder than ever to recruit and retain talent in general, but especially in computing.

NNSA laboratory positions require a substantial ramp-up, both owing to the clearance process and because they require cross-disciplinary expertise. The ASC workforce draws and will continue to draw on highly skilled talent across computer science, computational science, mathematics, engineering, and the physical sciences. While recruiting in all areas is a challenge, data for computer science is used as an exemplar.

[1] A. Adkins, 2022, "Millennials: The Job-Hopping Generation," *Gallup*, December 27, https://www.gallup.com/workplace/231587/millennials-job-hopping-generation.aspx.

There is a dramatic increase in the fraction of PhDs who specialize in artificial intelligence (AI) and machine learning (Figure 4-1). For example, data from the Computing Research Association indicates that between 2010 and 2019, AI-related PhDs have increased from 14.2 percent to 23 percent of the total, while software engineering, computer architecture, programming languages, and scientific computing all saw a drop. Furthermore, an increasing fraction of U.S. computer science PhDs are awarded to international students.[2]

This changing demographic of computer science expertise, combined with the fact that a large fraction of emerging talent is choosing to go to industry, provides a smaller talent pool for high-performance computing, scientific/numerical computing, software engineering, and hardware/architecture.

FINDING 4: NNSA's laboratories face significant challenges in recruiting and retaining the highly creative workforce that NNSA needs, owing to competition from industry, a shrinking science, technology, engineering, and mathematics (STEM) talent pipeline, and challenges in hiring diverse and international talent.

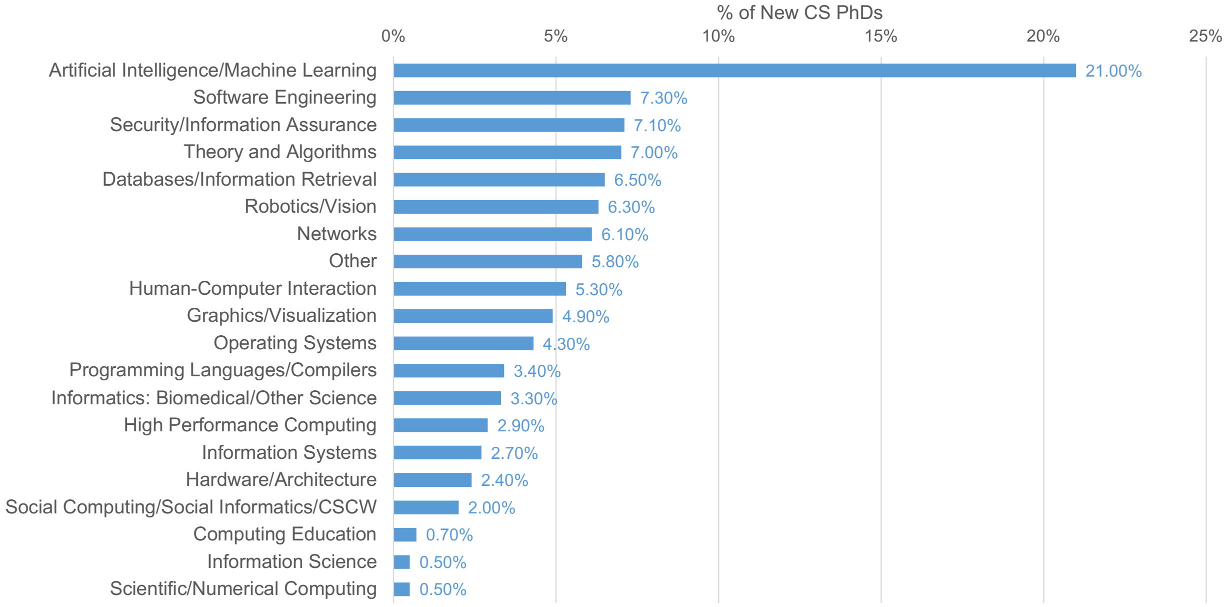


FIGURE 4-1 New computer science (CS) PhDs in the United States by specialty.
NOTE: CS, computer science; CSCW, computer-suported cooperative work.
SOURCE: Data from CRA Taulbee Survey, 2021.

[2] S. Mishra, 2021, "The AI Index: Emerging Trends in AI Education," Computing Research Association, May 21, https://cra.org/crn/2021/04/the-ai-index-emerging-trends-in-ai-education.

FINDING 4.1: The ASC program currently faces a challenge maintaining a competitive workforce; this challenge will continue to grow because of pipeline issues (small number of U.S. citizens going into graduate-level STEM fields), industry competition, and emerging computing talent choosing not to focus on scientific computing.

With a limited talent pool, recruiting and developing a *diverse* population is more and more important. Not only is it critical in terms of filling needed workforce positions, but also, teams composed of individuals from a wide range of backgrounds and experiences help facilitate progress in all areas.[3] According to the Bureau of Labor Statistics and Pew Research, the diversity in computing jobs is improving, albeit very slowly.[4] At current rates, it will take more than 100 years to see balanced representation.

In the past, workforce gaps have been filled by international talent, but that too is changing. From the National Science Board:

> The US has long been a magnet for top international STEM talent. This feature has been crucial for America's S&E [science and engineering] enterprise, both because international students and workers bring valuable knowledge and skills and because the US has failed to engage enough US citizens in STEM education and careers. Even as the US works to address its domestic talent shortfall, the country must continue to attract talent from around the world....
>
> While the US has tended to take foreign talent for granted, the world has changed. Other countries have learned from the US and are opening their doors to foreign talent, giving the world's top students an increasing number of options. Furthermore, as other countries invest in their own domestic R&D, internationally mobile S&E students and workers have access to more education, training, and job opportunities in other nations, including in their home countries.[5]

The NNSA laboratories are experiencing a growing number of barriers in engaging with foreign talent, including restrictions on international collaborations, foreign national visitors, and hiring foreign nationals.[6]

[3] See, for example, D. Rock and H. Grant, 2016, "Why Diverse Teams Are Smarter," *Harvard Business Review* 4(4):2–5, and A.W. Woolley, C.F. Chabris, A. Pentland, N. Hashmi, and T.W. Malone, 2010, "Evidence for a Collective Intelligence Factor in the Performance of Human Groups," *Science* 330(6004):686–688.

[4] K. Hendrickson, 2018, "Is Diversity in Computing Jobs Improving?" https://codeorg.medium.com/is-diversity-in-computing-jobs-improving-32f30068b7de; M. Froehlicher, L. K. Griek, A. Nematzadeh, L. Hall, and N. Stovall, 2021, "Gender Equality in the Workplace: Going Beyond Women on the Board," *The Sustainability Yearbook* 38–57, https://www.spglobal.com/esg/csa/yearbook/articles/gender-equality-workplace-going-beyond-women-on-the-board; World Economic Forum, 2020, "Global Gender Gap Report 2020" https://www.weforum.org/reports/gender-gap-2020-report-100-years-pay-equality.

[5] National Science Board, 2022, "International STEM Talent Is Crucial for a Robust U.S. Economy," https://www.nsf.gov/nsb/sei/one-pagers/NSB-International-STEM-Talent-2022.pdf.

[6] Department of Energy (DOE) Order 486.1A, DOE Policy 485.1A, DOE Order 142.3B.

FINDING 4.2: The U.S. national security enterprise has benefited enormously from the inclusion of global talent but incorporating international scholars in the NNSA community is challenged by important concerns about protecting sensitive information. Failure to balance these risks with the risk of missing the best talent can result in not finding the best candidate for the job.

As a counterpoint,

> In 2008, China's central government announced the Thousand Talents Plan: a scheme to bring leading Chinese scientists, academics and entrepreneurs living abroad back to China. In 2011, the scheme grew to encompass younger talent and foreign scientists, and a decade later, the Thousand Talents Plan has attracted more than 7,000 people overall. For Chinese scientists, the scheme has given them a strong financial incentive to return home. For foreigners, it's an opportunity to join the Chinese system.[7]

Last, while NNSA has entirely funded the mission-specific applications in the Exascale Computing Project (ECP) and can sustain those projects as needed, much of the funding for other science applications, co-design frameworks, and general software is from Advanced Scientific Computing Research (ASCR) and has an uncertain future. This includes ASCR funding for team members at the NNSA laboratories in areas such as molecular dynamics, combustion simulation, scientific libraries, and system software, as well as midcareer and senior leaders who have taken on project management roles within ECP. Even if these funding issues are eventually resolved through creation of new programs, the uncertainty itself creates a significant retention and recruiting risk.

WORKFORCE OPPORTUNITIES

Responding to these workforce challenges requires NNSA to proactively pursue several areas of opportunity that address workforce needs in both the shorter and longer terms. The following paragraphs highlight opportunities related to the nurturing and retention of existing staff, strategically growing partnerships, and proactively growing the pipeline.

The first area of opportunity is to increase efforts to nurture and retain existing staff. Given the external context described above, this is an area that requires urgent

[7] J. Hepeng, 2018, "China's Plan to Recruit Talented Researchers," *Nature* 553(7688):S8. https://doi.org/10.1038/d41586-018-00538-z.

attention because it has potential effects on the NNSA workforce in both the short and longer terms. As part of this retention effort, it must be recognized that total compensation for computing experts needs to be competitive. A recent DOE report speaks to these challenges:

> Demand for high performance computing, networking, algorithms, and mathematics is high across industries. A strong entry level graduate in computer science will regularly receive a compensation package of salary, bonus, and stock from a hyperscale internet company exceeding $300,000 per year; a principal engineer or research scientist's compensation can go to seven figures. These compensation rates are many times the amount this talent can earn at the national labs. The demand is driven by the impact that high performance computing has on the profitability and capabilities of these companies, in serving insatiable clickflows, data analytics, and particularly AI/ML computing demands.
>
> For entities that cannot compete for this talent with money, the strategy for successfully competing is offering a differentiated mission and culture, which ASCR can certainly do, particularly with mission: an opportunity to develop the essential tools for advancing science. Talent will be and is motivated by the opportunity to develop tools to understand our universe, our biology, the mind, our climate, and to develop new technologies applying that understanding. Attracting and retaining talented people requires that they feel valued as professionals, connected with their colleagues, engaged in contributing to the science mission goals of ASCR and DOE, and supported in their pursuit of career development opportunities.[8]

In other words, total compensation goes beyond competitive pay. There are numerous retention incentives that include job satisfaction, making a difference (mission and purpose), respectful treatment, feeling valued, working with talented people on hard problems, work-life balance, trust between employees and management, and opportunities for innovation. Funding fragmentation and being split between too many projects are often cited as a source of stress among laboratory researchers. Furthermore, the conclusion of ECP in 2024 will leave a void for a number of laboratory staff. ECP has provided more than just funding—it has also provided a connected community with a common sense of shared mission. NNSA must recognize these workforce stressors and be proactive in providing an environment that mitigates them.

[8] DOE Office of Science, Advanced Scientific Computing Research Program, 2020, "Transitioning ASCR After ECP," https://science.osti.gov/-/media/ascr/ascac/pdf/meetings/202004/Transition_Report_202004-ASCAC.pdf.

FINDING 4.3: Addressing the challenges laid out in this report will require a nurturing environment that reduces distractions, funding uncertainty, and administrative burdens, while providing employees the time and flexibility to explore areas of interest and do the creative thinking required to solve these problems.

Retention incentives also need to consider a multigenerational workforce. It is common for the laboratories to have employees of three generations working side by side. Accommodating diverse working styles and needs and leveraging the strengths of each generation can be important retention incentives.

A second important area of focus is to aggressively grow the workforce pipeline. The Department of Energy (DOE) has several workforce programs that are key elements of providing needed talent for the future. Continuing to grow and strengthen NNSA participation in these programs is critical for meeting workforce needs over the longer term.

One such program is the Nuclear Science and Security Consortium (NSSC),[9] a 5-year program to develop a new generation of laboratory-integrated nuclear experts. The NSSC enables a rich collaborative research environment between universities and the national laboratories, and fosters the development of science and technology underlying the nuclear security mission.

As part of its science and national security missions, NNSA supports students pursuing degrees in a spectrum of basic and applied research in science and engineering. In particular, NNSA seeks candidates who demonstrate the skills and potential to form the next generation of leaders in a number of fields via the DOE NNSA Laboratory Residency Graduate Fellowship program, including mathematics and computational science, especially multiscale and multiphysics theory, continuum numerical simulation, and particle-in-cell/fluid hybrid simulation, coupled with an emphasis on using data from advanced experiments to inform and validate simulations and methods.

The Predictive Science Academic Alliance Program (PSAAP) is a mechanism by which NNSA laboratories engage the U.S. academic community in advancing science-based modeling and simulation technologies:

> The PSAAP integrates modeling, simulation, and experiment in a manner that contributes to the nuclear security mission and prepares the next generation of scientists and engineers for careers in national security.
>
> —Dr. Mark Anderson, Assistant Deputy Administrator for Research, Development, Test and Evaluation in NNSA's Office of Defense Programs.[10]

[9] Nuclear Science and Security Consortium, https://nssc.berkeley.edu.

[10] National Nuclear Security Administration, 2020, "NNSA Announces Selection of Predictive Science Academic Alliance Program Centers of Excellence," Energy.gov, https://www.energy.gov/nnsa/articles/nnsa-announces-selection-predictive-science-academic-alliance-program-centers.

NNSA investments in academic institutions via basic research funding and academic alliance programs such as PSAAP are key drivers in creating the future workforce. Graduate students and postdoctoral fellows engaged in such research programs have immersive research experiences that hone their specific technical skills, cultivate their broader problem-solving skills, and expose them to the wider scientific and technological community. These graduate students work with collaborators from the national laboratories and from industry partners. They engage in internships. They are immersed in the notion of basic research that targets societal grand challenges together with a culture of rigorous mathematically grounded approaches and a culture of HPC at scale.[11] This immersion prepares them to contribute to NNSA's pressing scientific and technological challenges.

Established in 1991, the DOE's Computational Science Graduate Fellowship (CSGF)[12] provides outstanding benefits and opportunities to students pursuing doctoral degrees in fields that use high-performance computing to solve complex science and engineering problems. CSGF's specific objectives are:

- To help ensure an adequate supply of scientists and engineers appropriately trained to meet national workforce needs, including those of the DOE, in computational sciences.
- To raise the visibility of careers in the computational sciences and to encourage talented students to pursue such careers, thus building the next generation of leaders in the field.
- To provide practical work experiences for the fellows that allows them to encounter the cross-disciplinary, team-based, scientific research environment of the DOE national laboratories.
- To strengthen collaborative ties between the academic community and DOE national laboratories so that the fellowship's multidisciplinary nature builds the national community of scientists.

CSGF program fellows exemplify the type of students required to address future NNSA workforce needs: they are inspired by interdisciplinary topics at the interfaces of mathematics, computing and engineering, and scientific applications, and they are deeply committed to addressing the nation's scientific and technological challenges. The CSGF program thus plays a critical role in ensuring a strong, diverse pipeline of highly trained professionals who remain committed to scientific and engineering domains,

[11] K.E. Willcox, 202l, "Accelerating Discovery: The Future of Scientific Computing at the Department of Energy," https://republicans-science.house.gov/_cache/files/a/5/a5571d1b-2d8e-4d59-93cb-a9a5df4049b0/84E5CD8050480E2D757A77A6CAA3AD8D.2021-05-19-testimony-willcox.pdf.

[12] Department of Energy Computational Science Graduate Fellowship, https://www.krellinst.org/csgf.

rather than being lured away by more lucrative positions in commercial and business sectors. Furthermore, the CSGF program is unique among other prestigious national fellowships in the way it proactively shapes its trainees. This is achieved through requirements on graduate courses (and associated stringent oversight), through the required 12-week research traineeship at a national laboratory, and through concerted attention to mentoring and networking.

A consequence of the CSGF graduate course requirement is that the DOE has played an important role in incentivizing universities to create new interdisciplinary computational science graduate curricula that address DOE workforce needs. An expansion of the CSGF program would not only increase the supply of talented students trained at the interfaces of mathematics, computing, and science/engineering but also provide DOE with a key lever to encourage new university programs that respond to pressing scientific computing workforce needs. There is no doubt that demand for those trained through the CSGF program already outstrips supply, and this demand will only increase in the coming years. Further, the pool of highly qualified applicants far outstrips the availability of CSGF awards. DOE could double the size of the CSGF program—perhaps with a component of CSGF focused on NNSA mission needs or on innovative computing post-exascale—and both the demand for and quality of the fellows would remain extremely high.

NNSA laboratory efforts at the K–12 level also help secure the future workforce. Each laboratory works with its local and regional K–12 school districts in a variety of programs, training teachers and engaging students early in the benefits of pursuing a STEM education, including field trips, community science projects, fun with science projects, and internships that range from one day to several weeks.

A third important area of workforce opportunity is partnerships, which play a key role in multiple aspects of workforce needs. There are a number of entities where expanded NNSA partnerships (e.g., through undergraduate internships) could result in increased volume and diversity of the PhD pipeline. These entities include (but are not limited to) the National Society of Black Engineers; the Association for Computing Machinery's Richard Tapia Celebration of Diversity in Computing; the Society for Advancement of Chicanos/Hispanics and Native Americans; the Society of Women Engineers; the Society of Hispanic Professional Engineers; the Grace Hopper Celebration of Women in Computing; the Emerging Researchers National Conference in Science, Technology, Engineering, and Mathematics, aimed at underrepresented minorities and persons with disabilities; and career fairs and virtual information sessions at minority serving institutions.

Through these partnerships, the NNSA laboratories should be striving to build relationships with a diverse group of students early in their careers and to play a role in encouraging these students to pursue PhDs. A particularly powerful tool in this regard is the opportunity to retain employment at a national laboratory while pursuing a PhD. Such programs can be especially important for those from economically underserved backgrounds who cannot financially afford a more traditional PhD path. Expansion of such programs could thus be a valuable tool in helping to grow the diversity of the NNSA computing workforce.

In addition to normal hiring pathways, the laboratories should work to find and attract exceptional minds in nontraditional venues (e.g., review committee members, professors, industrial experts, high-school students) and bring them in with the conscious intent to engage them in mission problems—whether for a week, a summer, a sabbatical, an apprenticeship, an expert's camp, or a career. In all of these areas, the committee encourages the NNSA laboratories not to work individually or in competition, but rather to formulate strategic cross-laboratory and interagency efforts to collect data, coordinate recruiting, and work with the dimension of diversity and inclusion as ways to improve the workforce.

RECOMMENDATION 3: NNSA should develop an aggressive national strategy through partnership across agencies and academia to address its workforce challenge.

> **RECOMMENDATION 3.1:** NNSA should make concerted efforts to create an environment that nurtures and retains existing staff; more aggressively grow the pipeline; create an efficient and modern, yet secure environment; advertise and grow existing workforce programs (such as the Predictive Science Academic Alliance Program and the Computational Science Graduate Fellowship); and collaborate with other federal agencies to support ambitious talent development programs at all career stages.
>
> **RECOMMENDATION 3.2:** NNSA should also develop a deliberate strategy to attract an international workforce and to provide them with a welcoming environment while thoughtfully managing the attendant national security risks.

Appendixes

A

Statement of Task

As requested in section 3172 of the Fiscal Year (FY) 2021 National Defense Authorization Act (NDAA), an ad hoc committee of the National Academies of Sciences, Engineering, and Medicine will conduct a consensus study "reviewing the future of computing beyond exascale computing to meet national security needs at the National Nuclear Security Administration." (Exascale refers to a computer that performs near or above 10^{18} floating-point operations per second.)

The study will review:

1. NNSA's computing needs over the next 20 years that exascale computing will not support;
2. Future computing technologies for meeting those needs, including quantum computing and other novel hardware, computer architecture, and software;
3. The likely trajectory of promising hardware and software technologies and obstacles to their development and their deployment by NNSA; and
4. The ability of the U.S. industrial base, including personnel and microelectronics capabilities, to meet NNSA's needs.

The work will be carried out in parallel unclassified and classified tracks. The full committee will gather information, deliberate, and develop its report on an entirely unclassified basis.

In considering item (1) above, the committee will coordinate its work with the work of the committee for the soon-to-be-launched congressionally mandated and

NNSA-sponsored Assessment of High Energy Density Physics. This coordination will be performed by the respective project staff and, to the extent feasible, through overlapping committee memberships and/or designated committee liaisons.

A separately appointed and appropriately cleared small subset of the full study committee will commence its work at approximately the midpoint of the study. It will review pertinent classified information relating to NNSA's future computing needs. It will prepare an internal working document that will be submitted to NNSA for unclassified/public release and provided to the full study committee to inform its work.

The study committee will prepare an unclassified public report and the cleared subset of the committee will prepare a classified annex, as deemed appropriate.

B

Presentations to the Committee

Meeting 1—November 19, 2021—NNSA Background

Thuc Hoang, Advanced Simulation and Computing (ASC), National Nuclear Security Administration (NNSA)

Christopher Clouse, Lawrence Livermore National Laboratory (LLNL)

Jason Pruet, Los Alamos National Laboratory (LANL)

Scott Collis, Sandia National Laboratories (SNL)

Meeting 2—December 13, 2021—Industry Briefings

Brent Gorda, Senior Director, HPC Business, Arm

Alan Lee, Corporate Vice President and Head of AMD Research, AMD

Michael Schulte, Senior Fellow in AMD Research, AMD

Bill Dally, Chief Scientist, NVIDIA Corporation

Brijesh Tripathi, Vice President, Accelerated Computing Systems and Graphics Group; General Manager, Super Compute Platforms, Intel Corporation

Andrew Wheeler, HPE Fellow, Vice President and Director, Hewlett Packard Labs, Hewlett Packard Enterprise

James Sexton, IBM Fellow, Director of Future Computing Systems, IBM

Meeting 3—February 17, 2022—International/CLOSED

Satoshi Matsuoka, Riken

Daniel Verwaerde, Teratec

David Kahaner, Asian Technology Information Program

Meeting 4—March 21, 2022—LANL Meeting/CLOSED

Teresa Bailey, LLNL

Galen Shipman, LANL

Erik Strack, SNL

Thuc Hoang, ASC, NNSA

Meeting 5—April 6, 2022—Hyperion Research/CLOSED

Earl Joseph, Hyperion Research

Mike Thorp, Hyperion Research

Meeting 6—May 27, 2022—DOE Office of Science

Jeff Nichols, Oak Ridge National Laboratory

Jonathan Carter, Lawrence Berkeley National Laboratory

Rick Stevens, Argonne National Laboratory

Meeting 7—June 27, 2022—Post-Exascale Strategies

Mark Anderson, NNSA

Thuc Hoang, ASC, NNSA

Lynne Parker, Office of Science and Technology Policy

C

Committee Member Biographical Information

KATHERINE A. YELICK, *Chair*, is the vice chancellor for research and the Robert S. Pepper Distinguished Professor of Electrical Engineering and Computer Sciences at the University of California, Berkeley. Dr. Yelick is an internationally recognized expert in high-performance computing (HPC). Her research interests include parallel programming languages, automatic performance tuning, parallel algorithms, and computational genomics. Dr. Yelick concurrently holds a senior faculty scientist appointment at Lawrence Berkeley National Laboratory (LBNL), where she was an associate laboratory director of the Computing Sciences Area from 2010 through 2019. She was also the director of the National Energy Research Scientific Computing Center from 2008 through 2012. Dr. Yelick earned her PhD in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT).

JOHN B. BELL is a senior scientist at LBNL and the chief scientist of LBNL's Applied Mathematics and Computational Research Division. Dr. Bell's research focuses on the development and analysis of numerical methods for partial differential equations arising in science and engineering. He has made contributions in the areas of finite volume methods, numerical methods for low Mach number flows, adaptive mesh refinement, stochastic differential equations, interface tracking, and parallel computing. Dr. Bell is a fellow of the Society of Industrial and Applied Mathematics (SIAM) and a member of the National Academy of Sciences. He was recipient of the SIAM/Association for Computing Machinery (ACM) Prize in Computational Science and Engineering, the Sidney Fernbach Award, and the Berkeley Laboratory Lifetime Achievement Award. Dr. Bell received his PhD in mathematics from Cornell University in 1977.

WILLIAM W. CARLSON has been a member of the research staff at the IDA Center for Computing Sciences since 1990. Dr. Carlson's work has centered on innovations focused on computing at extreme scale and its applicability and relevance to a variety of scientific and mathematical applications. These include spearheading the design and implementation of the UPC programming language; work in major research programs such as HTMT, the Defense Advanced Research Projects Agency High-Productivity Computing Systems program, and Project 38; and current efforts to understand the role of languages such as Rust in large-scale computations. Dr. Carlson is a member of the Institute of Electrical and Electronics Engineers (IEEE) and the ACM. He received his PhD and MSEE from Purdue University and his BS from Worcester Polytechnic Institute.

FREDERIC T. CHONG is the Seymour Goodman Professor in the Department of Computer Science at the University of Chicago and the chief scientist for Quantum Software at Infleqtion. He is also the lead principal investigator for the EPiQC (Enabling Practical-scale Quantum Computing) Project, a National Science Foundation (NSF) Expedition in Computing. In 2020, he co-founded Super.tech, a quantum software company, which was acquired by Infleqtion (formerly ColdQuanta) in 2022. He is a fellow of the IEEE and a recipient of the NSF CAREER award, the Intel Outstanding Researcher Award, and 13 best paper awards. His research interests include emerging technologies for computing, quantum computing, multicore and embedded architectures, computer security, and sustainable computing. He received his PhD from MIT in 1996. Dr. Chong has published several articles on the subject of quantum computing, including "Emerging Technologies for Quantum Computing" in the *IEEE Micro* Special Issue on Quantum Computing in 2021; "Quantum Computer Systems for Scientific Discovery" in *Physical Review Research* in 2020; "Quantum Computer Systems: Research for Noisy Intermediate-Scale Quantum Computers" in *Synthesis Lectures in Computing* in 2020; "Quantum Computing for Enhancing Grid Security" in *IEEE Power Engineering Letters* in 2020; and "Greater Quantum Efficiency by Breaking Abstractions" in *Proceedings of the IEEE* in 2020. Quantum computing is a potential key technology to be considered for post-exascale computing.

DONA L. CRAWFORD retired as the associate director for computation from Lawrence Livermore National Laboratory (LLNL), where she led the laboratory's HPC efforts. In that capacity, Ms. Crawford was responsible for the development and deployment of an integrated computing environment for petascale simulations of complex physical phenomena. Prior to her LLNL appointment in 2001, Ms. Crawford was with Sandia National Laboratories since 1976, serving on many leadership projects, including the Accelerated Strategic Computing Initiative and the Nuclear Weapons Strategic Business Unit. She has served on a number of committees of the National Academies of Sciences, Engineering, and Medicine, including the Committee to Review Governance Reform in the National Nuclear Security Administration and the Committee to Evaluate the NSF's Vertically Integrated Grants for Research and Education (VIGRE) Program. Additionally, she served on the National Nuclear Security Administration Defense Programs Advisory Committee for High-Performance Computing (2018–2019). Ms. Crawford received her MS in operations research from Stanford University.

MARK E. DEAN is a professor emeritus at the University of Tennessee, Knoxville (UTK). His research focus was in advanced computer architecture (neuromorphic computing). Prior to joining UTK, Dr. Dean had a 34-year career in the computer industry working in various executive and research and development positions at IBM. He served as an IBM fellow, chief technology officer of the Middle East and Africa, and vice president of World Wide Strategy and Operations for IBM Research. Dr. Dean holds three of the nine patents for the original IBM PC and created the Industry Standard Architecture, which permitted add-on devices like the keyboard, disk drives, and printers to be connected to the motherboard, earning him election to the National Inventors Hall of Fame. Dr. Dean is currently a committee member for the Division on Engineering and Physical Sciences of the National Academies. Dr. Dean received a PhD in electrical engineering from Stanford University and is a member of the American Academy of Arts and Sciences and the National Academy of Engineering.

JACK J. DONGARRA is a distinguished professor of electrical engineering and computer science at the University of Tennessee, Knoxville, and a distinguished research staff member at Oak Ridge National Laboratory. Dr. Dongarra specializes in numerical algorithms in linear algebra, parallel computing, the use of advanced computer architectures, programming methodology, and tools for parallel computers. He was the recipient of the 2021 ACM A.M. Turing Award. Dr. Dongarra was the first recipient of the SIAM Special Interest Group on Supercomputing's award for Career Achievement in 2010; in 2019 he received the ACM/SIAM Computational Science and Engineering Prize, and in 2020 he received the IEEE Computer Society Computer Pioneer Award. He is a fellow of the American Association for the Advancement of Science (AAAS), ACM, IEEE, and SIAM and a foreign member of the Russian Academy of Science, a foreign member of the British Royal Society, and a U.S. National Academy of Engineering member. Dr. Dongarra served on the Committee to Study the Future of Supercomputing (2003) and the Committee for Technology Insight (2010). He received his PhD in computer science from the University of New Mexico.

IAN T. FOSTER is the Arthur Holly Compton Distinguished Service Professor of Computer Science at the University of Chicago and also a senior scientist, distinguished fellow, and director of the Data Science and Learning Division at Argonne National Laboratory. Dr. Foster's research deals with distributed, parallel, and data-intensive computing technologies and applications of those technologies to problems in such domains as materials science, climate change, and biomedicine. The Globus software that he co-invented is widely used in national and international cyberinfrastructures and science projects. Dr. Foster currently serves on the National Academies' U.S. National Committee for CODATA. Dr. Foster is a fellow of the AAAS, the ACM, the British Computer Society (BCS), and the IEEE, and is a Department of Energy (DOE) Office of Science Distinguished Scientists Fellow. He has received the BCS Lovelace Medal and the IEEE Babbage, Goode, and Kanai awards. Dr. Foster obtained a BSc from the University of Canterbury, New Zealand, and a PhD from Imperial College, United Kingdom, both in computer science.

CHARLES F. McMILLAN is a retired director of Los Alamos National Laboratory (LANL). Dr. McMillan has served in leadership roles in the nuclear weapons program at both LANL and LLNL, in which HPC has been essential to program success. During the 1990s, he helped to create the Accelerated Scientific Computing Initiative and led the software development group that produced the first three-dimensional, parallel simulations of a nuclear weapons primary. Dr. McMillan currently serves on the National Academies' Committee on Assessment of High Energy Density Physics. Dr. McMillan received a PhD in physics from MIT. Dr. McMillan provided classified Annual Assessment letters to the U.S. President, in which he has provided evaluations of the state of computing as one of the tools of stewardship as required by Congress. He has also provided congressional testimony on the U.S. nuclear deterrent that includes the results of modeling and simulation using HPC.

DANIEL I. MEIRON is currently a professor of aerospace and applied and computational math at the California Institute of Technology. Dr. Meiron's interests include computational fluid dynamics and computational materials science. Dr. Meiron previously served on the National Academies' NSF Graduate Panel on Applications of Mathematics. He received his ScD in applied mathematics at MIT in 1981.

DANIEL A. REED is the Presidential Professor at The University of Utah, where he is also a professor of computer science and electrical and computer engineering. Previously, Dr. Reed served as The University of Utah's senior vice president for academic affairs (provost), Microsoft's corporate vice president for technology policy and extreme computing, founding director of the Renaissance Computing Institute, and director of the

National Center for Supercomputing Applications. He was one of the principal investigators and chief architect for NSF's TeraGrid, which became NSF XSEDE. Dr. Reed currently chairs the National Science Board. Dr. Reed is serving as the chair of DOE's Advanced Scientific Computing Advisory Committee (ASCAC) (2016–present). ASCAC conducts studies and issues reports based on requests from DOE leadership. Dr. Reed is a fellow of the ACM, the IEEE, and the AAAS. He previously chaired the National Academies' Panel on Computational Sciences at the Army Research Laboratory. Other National Academies' committee memberships include the Committee on Future Directions of NSF Advanced Computing Infrastructures to Support U.S. Science in 2017–2020. Dr. Reed received his BS from the Missouri University of Science and Technology and his MS and PhD from Purdue University, all in computer science.

KAREN E. WILLCOX is the director of the Oden Institute for Computational Engineering and Sciences, associate vice president for research, and a professor of aerospace engineering and engineering mechanics at The University of Texas at Austin (UT). Dr. Willcox holds the W.A. "Tex" Moncrief Jr. Chair in Simulation-Based Engineering and Sciences and the Peter O'Donnell Jr. Centennial Chair in Computing Systems. Prior to UT, Dr. Willcox spent 17 years as a professor at MIT, where she served as the founding co-director of the MIT Center for Computational Engineering and the associate head of the MIT Department of Aeronautics and Astronautics. Dr. Willcox is currently a member of the National Academies' Board on Mathematical Sciences and Analytics. She previously served on the National Academies' Panel on Review of the Information Technology Laboratory at the National Institute of Standards and Technology. Dr. Willcox is a fellow of SIAM, a fellow of the American Institute of Aeronautics and Astronautics, and a member of the New Zealand Order of Merit for services to aerospace engineering and education. Dr. Willcox holds a bachelor of engineering degree from the University of Auckland, New Zealand, and master's degree and PhD in aeronautics and astronautics from MIT. Dr. Willcox submitted oral and written testimony to the Subcommittee on Energy of the House Committee on Science, Space, and Technology hearing on Accelerating Discovery. The statement highlights several key strategies for the future of scientific computing at DOE.

D

Disclosure of Unavoidable Conflicts of Interest

The conflict of interest policy of the National Academies of Sciences, Engineering, and Medicine (http://www.nationalacademies.org/coi) prohibits the appointment of an individual to a committee authoring a Consensus Study Report if the individual has a conflict of interest that is relevant to the task to be performed. An exception to this prohibition is permitted if the National Academies determine that the conflict is unavoidable and the conflict is publicly disclosed. A determination of a conflict of interest for an individual is not an assessment of that individual's actual behavior or character or ability to act objectively despite the conflicting interest.

Dr. Charles F. McMillan was determined to have a conflict of interest because of his compensated service on advisory boards to the National Nuclear Security Administration (NNSA) weapons laboratories—Los Alamos National Laboratory (LANL) and Sandia National Laboratories (SNL). At LANL, Dr. McMillan serves on the Weapons Capability Review Committee. At SNL, Dr. McMillan chairs the external review committee for Radiation Electronics and High Energy Density Science. The National Academies determined that the experience and expertise of the individual was needed for the committee to accomplish the task for which it was established. The National Academies could not find another available individual with the equivalent experience and expertise who did not have a conflict of interest. Therefore, the National Academies concluded that the conflict was unavoidable and publicly disclosed it on its website (www.nationalacademies.org).

Dr. Karen E. Willcox was determined to have a conflict of interest because of her compensated service on advisory boards to the NNSA weapons laboratories—LANL and SNL. At LANL, Dr. Willcox serves on the Advanced Simulation and Computing Advisory

Board. At SNL, Dr. Willcox serves on the external review board for the Computing and Information Sciences Research Foundation. The National Academies determined that the experience and expertise of the individual was needed for the committee to accomplish the task for which it was established. The National Academies could not find another available individual with the equivalent experience and expertise who did not have a conflict of interest. Therefore, the National Academies concluded that the conflict was unavoidable and publicly disclosed it on its website (www.nationalacademies.org).

E

Acronyms and Abbreviations

1D	one-dimensional
2D	two-dimensional
3D	three-dimensional
AI	artificial intelligence
AI4Sci	AI for Science
AI4SES	AI for Science, Energy, and Security
ALE	Arbitrary Lagrangian-Eulerian
AMR	adaptive mesh refinement
ANL	Argonne National Laboratory
ASC	Advanced Simulation and Computing
ASCI	Accelerated Strategic Computing Initiative
ASCR	Advanced Scientific Computing Research
ASIC	application-specific integrated circuit
CHIPS	Creating Helpful Incentives to Produce Semiconductors
CJ	Chapman-Jouguet
CMOS	complementary metal-oxide semiconductor
CSGF	Computational Science Graduate Fellowship
DARHT	Dual Axis Radiographic Hydrodynamic Test Facility
DOE	Department of Energy
DRAM	dynamic random access memory

ECP	Exascale Computing Project
EOS	equation of state
EUV	extreme ultraviolet
FLOP	floating-point operation
FLOPS	floating-point operations per second
FPGA	field-programmable gate array
HBM	high-bandwidth memory
HPC	high-performance computing
IaaS	infrastructure as a service
IC	integrated circuit
LANL	Los Alamos National Laboratory
LEP	life extension program
LES	large eddy simulation
LLNL	Lawrence Livermore National Laboratory
MFLOP	millions of floating-point operation
MPI	message passing interface
NDAA	National Defense Authorization Act
NEP	nuclear explosive package
NIF	National Ignition Facility
NISQ	noisy intermediate-scale quantum
NIST	National Institute of Standards and Technology
NNSA	National Nuclear Security Administration
NNSS	Nevada National Security Site
NSSC	Nuclear Science and Security Consortium
NUMA	nonuniform memory access
ORNL	Oak Ridge National Laboratory
PaaS	platform as a service
PDE	Partial Differential Equation
PSAAP	Predictive Science Academic Alliance Program

QMU	quantification of margins and uncertainties
R&D	research and development
RANS	Reynolds-averaged Navier-Stokes
SaaS	software as a service
SBSS	Science-Based Stockpile Stewardship
SNL	Sandia National Laboratories
SoC	system-on-a-chip
TSMC	Taiwan Semiconductor Manufacturing Company
VVUQ	verification, validation, and uncertainty quantification