How the United States Can Increase Access to Supercomputing

By Hodan Omaar | December 7, 2020

Academic researchers need large amounts of computing power to solve tough computational problems. As problems become more data-intensive and require larger amounts of power to be solved, the demand for access to systems with immense computing power, or high-performance computing (HPC), has also grown. Unfortunately, the supply of HPC resources has not kept up with growing demand.

Limited access to HPC is hampering the ability of AI researchers to develop new products and services that are vital in maintaining U.S. competitiveness.

Without sufficient access to HPC, researchers across a diverse range of fields, including engineering, Earth sciences, biology, and computer science, are not able to address important research challenges. Indeed, HPC has fueled various applications at the forefront of artificial intelligence (AI), including natural language processing and machine learning. But limited access to HPC is hampering the ability of AI researchers to develop new products and services that are vital in maintaining U.S. competitiveness, inhibiting AI practitioners from applying AI to defense innovation, and slowing innovation needed to address important societal challenges, including in health care and the environment.

Increasing access to HPC for researchers exploring new applications of AI will involve increasing access to different parts of HPC systems, including hardware, software, and expertise, for users with large computational needs and the "the long tail" of users with more modest HPC needs that represent, in aggregate, the majority of AI researchers. The Department of Energy (DOE) primarily invests in the former and has increased its investments in data-intensive, large-scale HPC resources over the last decade by almost 90 percent, from $277 million in 2010 to $538 million (in constant 2010 dollars) in 2019. In contrast, the National Science Foundation (NSF), which is the primary source of HPC investments for the

latter, has decreased its HPC funding by approximately 50 percent, from $325 million in 2010 to $167 million in 2019. This discrepancy has led to a U.S. HPC portfolio weighted toward very powerful systems that can only support a smaller number of researchers. However, both funding sources fail to meet current demand.

To increase access to HPC resources for more AI researchers, Congress should increase funding for supercomputing to $10 billion over the next 5 years. Specifically, Congress should increase total NSF funding in HPC infrastructure to at least $500 million per year to match the current demand for time on NSF's HPC resources, which is more than three times greater than current supply. In addition, Congress should increase total DOE funding in HPC infrastructure to at least $1.5 billion per year to match current demand for access to DOE's HPC resources, which is three times greater than what DOE is currently providing.

To determine where to allocate these funds, NSF should first measure how states are using HPC resources for AI research and then fund HPC systems in states that have low levels of HPC availability but whose institutions are conducting high levels of AI research. In states where HPC availability is high, federal investments in more HPC resources will not be the most effective way to close the national gap between HPC supply and demand because either institutions in the state already have funding for HPC-enabled AI research and are using it, or they do not have research funding, which means access to HPC is not the problem, but rather research funding is. Focusing on states with low HPC availability but high AI research potential will allow the government to address instances where the gap between HPC demand and supply is greatest. Additionally, DOE and NSF should diversify the portfolio of HPC resources they are making available to AI researchers, including by exploring cloud computing options.

Maximizing returns on investment in HPC will require careful resource management driven by an understanding of what system requirements AI researchers need and how existing grantees are using HPC systems. DOE and NSF should require those institutions that receive funding to adopt HPC auditing tools such as the XDMoD tool that reports on how optimally institutions are using HPC systems. NSF should also annually collect community requirements and publish roadmaps that allow it to better set HPC priorities and make more strategic decisions that reflect user requirements.

Maximizing returns on investment in AI research will require mechanisms to effectively translate basic AI research into products and services for the marketplace. To this end, NSF should foster more public-private partnerships by tripling the number of awards it grants through its Partnerships for Innovation aimed at accelerating the path to market for new technologies, from 50 to 150 grants. Further, as part of its recent

initiative to create AI research institutes across the country, NSF should support proposals that are of regional importance and foster collaboration and partnerships between universities, local businesses, and state and local governments.

Ensuring all individuals have equal opportunity to succeed in becoming the next generation of AI researchers will mean increasing access to HPC for groups that are traditionally underrepresented in science and engineering. NSF and DOE should support partnerships that coordinate the sharing of computing resources with Minority-Serving Institutions (MSIs) that include Historically Black Colleges and Universities, Hispanic-Serving Institutions (HSIs), and Tribal Colleges and Universities (TCUs), as well as re-establishing targeted grants that fund HPC resources at MSIs. In addition, NSF and DOE should lower the barrier women face in gaining access to the supercomputing resources by replicating the Blue Waters project at the National Center for Supercomputing Applications that created an HPC allocations category open to researchers at U.S. academic institutions who are women.

Finally, creating a well-prepared HPC workforce will require all students with computer science backgrounds to have clear, structured pathways into the HPC workforce. To ensure students with terminal two-year computer science degrees or who transfer from community colleges can seamlessly move into upper-division coursework at four-year colleges without having to spend time duplicating technical fundamentals, NSF should provide funding for consortiums of two-year colleges and four-year colleges to work together in developing structured HPC curriculums.

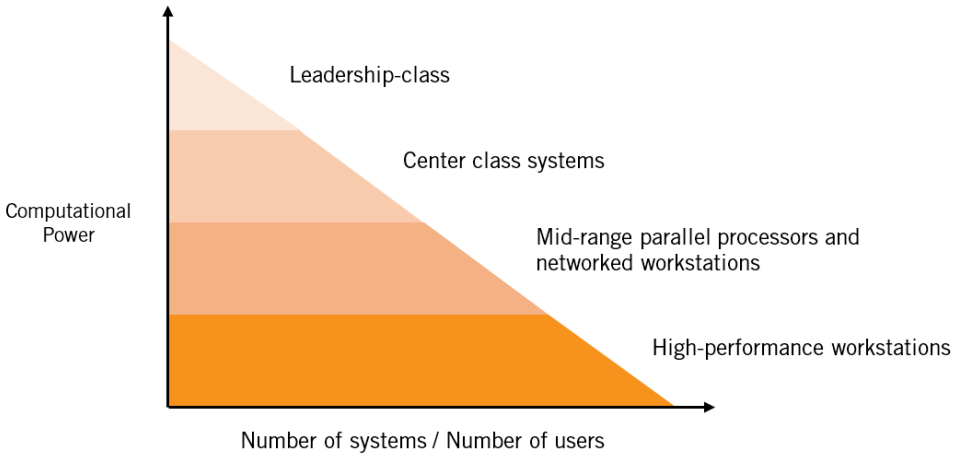
WHAT IS HIGH-PERFORMANCE COMPUTING?

HPC systems are those that have the computational power to solve difficult computational problems at any given time. Supercomputers are a subset of HPC systems, and describe "any computer that is one of the largest, fastest, and most powerful available at a given time."[1] HPC is more than just hardware, it is a combination of hardware, software, and computing expertise. Increasingly, HPC includes new modes of accessing computing resources, such as cloud computing, which makes it possible for a much larger pool of users to have access to HPC systems. Additionally, the performance of HPC systems depends on more than just a system's peak compute performance. For different applications, performance measures such as storage and memory size can be more relevant.

Figure 1 depicts the four main classes of computing systems: leadership-class systems that are the most powerful computing systems, such as those often hosted at national laboratories; center-class systems, such as the national HPC systems hosted on large campuses; mid-range HPC systems, such as the department-capacity systems hosted at smaller

universities and those small and medium-sized businesses might use; and personal computing systems, such as a personal desktop or workstation.

Figure 1: Branscomb pyramid describing the four tiers of computing



The graph shows that the more powerful an HPC system is, the fewer comparably powerful systems there are, and the fewer users and research projects they can support. This is because more capable systems are more expensive, meaning fewer institutions can afford to invest in them for the research projects they support. We use this framework to organize and talk about the HPC landscape throughout this report.

INCREASING ACCESS TO HPC REQUIRES INCREASING ACCESS TO HARDWARE, SOFTWARE, AND EXPERTISE

Access to state-of-the-art HPC requires considerable investments in state-of-the-art computing infrastructure, software that can effectively make the most of a system, and experts that can make both perform well.

HARDWARE

HPC systems consist of processors, memory, input-output (I/O) devices, and interconnection networks.

For processors, HPC systems usually use either central processing units (CPUs) or graphics processing units (GPUs). CPUs and GPUs have a lot in common, but have different architectures and are built for different purposes.[2] While CPUs can only have a small number of processing cores, they can focus those cores on getting individual tasks done quickly, making them well-suited to processing tasks wherein latency or per-core performance is important. GPUs are made up of many smaller and more specialized cores that work together to deliver massive performance on processing tasks that can be easily divided up and processed across many cores. This makes GPUs better suited to tasks wherein bandwidth rather

than speed is important. To see this, imagine a CPU is a sports car and a GPU is a semi-truck, and their task is to move a house full of boxes from one place to another. The sports car will move the boxes more quickly, but it will have to keep making the journey back and forth, whereas the semi-truck will carry a much greater load but will travel more slowly. An HPC system in this case is like a 100-lane highway, wherein many vehicles are working in parallel and the ratio of sports cars to semi-trucks depends on the size and nature of the boxes to be moved.

Vehicles need more than an engine to run though, and HPC systems require fundamental components other than processors to work. To move data in and out of the system, HPC systems use I/O devices to enable HPC systems to access and share data with other applications during simulations and analysis, which is important in highly data-intensive science applications.[3] To enable data to move between processors or between centralized data storage and the compute system, HPC systems use interconnection networks. These network connections help many processors collaborate on the solutions of single large tasks.

All high-performance computing platforms share these same core components—processors, memory, I/O devices, and interconnection networks—but how these components are combined to create an HPC system varies.

A custom supercomputer is a single system that has been custom built for specific applications. It uses specialized processors and interconnects, and typically has very high bandwidth. The NEC Earth Simulator in Japan that runs global climate models is an example of a custom supercomputer.[4] This system is composed of 5,120 CPUs that were developed by Japanese IT company NEC Corporation, and is housed in a specially designed building at Japan's Marine Science and Technology Center that is 70 meters long and 50 meters wide.[5] The operating system running on the Earth Simulator uses custom software built by NEC which means developers that wish to exploit the system's full capabilities must create applications tailored to the system.[6]

Custom supercomputers differ from other types of systems primarily in the memory bandwidth they can provide—that is, how quickly a processor can read or store data in memory—as well as the bandwidth they can offer between processors because of specialized interconnects.[7] For a small range of specific applications wherein bandwidth is important, such as Earth simulation, it can be worthwhile investing in a specialized custom supercomputer.

An HPC cluster is a collection of many separate servers, called nodes, connected via interconnects.[8] Commodity clusters typically use off-the-shelf processors, interconnects, and networks—and because they are

manufactured in high volume, enjoy economies of scale.[9] The Texas Advanced Computing Center (TACC) at the University of Texas, Austin, has several HPC clusters, including *Maverick2,* which is a dense cluster of GPUs designed to support machine learning and deep learning research.[10]

Importantly for governments, industry, and academic institutions wanting to invest in HPC systems, most systems only have a relatively short useful lifetime. A vibrant computing industry develops new technologies and products, meaning hardware depreciates over a three- to five-year lifetime. The perceived return on investment for a facility in the first five years therefore must be greater than the total cost of ownership because once hardware is depreciated and its performance and capability no longer competitive, only another infusion of capital will ensure service continuity.[11]

SOFTWARE

Researchers and businesses wanting to get the most out of HPC systems need access to leading-edge software that can efficiently exploit system capabilities. For example, within an individual computing core, there are individual sections of a processor which perform different tasks.[12] The software programs that get the most out of a processor are the ones that can execute instructions across multiple sections simultaneously, reducing processing time and improving efficiency.

Today's HPC software options are rich and diverse. The libraries used to develop software programs include a wide range of mathematical and statistical frameworks, from those for simple vector operations to ones for complex differential equation solvers, as well as software for visualization tools, graphics, simulation, data analysis, and program analysis that are necessary for scientific applications.[13] Whether used for climate modeling, fluid dynamics, astronomy, mechanical modeling, AI training, or something else, which software is the most appropriate to address researchers' needs depends on their computational needs.

These software tools, unlike hardware systems, can be useful for decades. Some of today's most popular software libraries are large community codes, such as Google's TensorFlow, a free open-source machine learning library that helps developers better train neural networks first launched in 2017; or Facebook's PyTorch, first launched in 2016, a machine learning library used for applications such as computer vision and natural language processing.[14] These frameworks, and others like them such as Caffe, PaddlePaddle, and Wolfram Language offer building blocks for designing, training, and validating AI models.[15]

However, the rapid growth in the diversity of hardware architectures makes it more difficult to predict what software developments will be needed to take advantage of future HPC systems, and how to create commercial

software that supports increasing hardware heterogeneity. In addition, as the Information Technology and Information Foundation's (ITIF) report *The Vital Importance of High-Performance Computing to U.S. Competitiveness* explains, the rate of improvement in hardware performance has slowed, shifting the burden of novelty software.[16] Enabling researchers to continue to work at frontiers of science, engineering, and AI therefore requires providing them with access to appropriate software resources. Further, to make effective use of their most valuable shared resources, governments, companies, and universities need to encourage developers to create more efficient software and research techniques for HPC systems.

EXPERTISE

Using an HPC system for any application requires the skills of domain experts and committed and well-trained advanced computing professionals. The larger and more complex the HPC system, the more expertise required to make them perform well. Think of the skills needed to drive a car, fly a plane, or operate a spacecraft. One can drive a car without having any significant knowledge of the science or engineering of how it works; to fly a plane, a person needs months of professional training, on-the-job experience, and a good understanding of how the system works; but to operate a space shuttle, an astronaut needs years of training and a strong understanding of the system and scientific principles. Similarly, low- and mid-range HPC systems are relatively straightforward to use with minimum training; users of center-class systems need more in-depth HPC training and practice; and using the extensive functionalities of leadership-class systems requires a firm understanding of the architecture and engineering involved, in addition to significant domain expertise.[17]

These include programming skills in computing language such as C++ and Python to understand and use existing HPC software and customize new software for novel applications; computer science skills to understand different hardware architectures, the functionalities they offer, and how to execute them; and extensive domain expertise to address challenging problems in science, engineering, business, social sciences, and the humanities.[18] Training and retaining people with a combination of these skills is key to enabling increasing usage of high-performance computing.

Despite their importance, these skills are often found in individuals who lack clear career paths and are dependent on an uncertain stream of funding for support in the public sector.[19] Although these individuals tend to gravitate toward research centers with HPC leadership, their salaries can be far higher in the private sector than in the academic or government research community.[20] This presents a challenge for universities and governments that need access to such expertise to stay competitive. As the National Security Commission on AI has noted, the growing divide in resources and opportunities between academia and the private sector is

"weighing the nation's research portfolio toward applied, market-driven endeavors." To ensure the growth of both public and private innovation activities, policymakers need to ensure they are not only training the next generation of researchers in HPC skills, but ensuring they have career opportunities that retain their talent within the academic community. For instance, offering reasonably secure, stable career paths and funding of research can incentivize skilled workers to stay in the public sector despite lower salaries.[21] In addition, increasing access to HPC system operators can support more research projects. For example, NSF's Extreme Science and Engineering Discovery Environment (XSEDE), a platform that coordinates the national sharing of supercomputing resources, has a practice that permits researchers to request an allocation of staff time along with computer time.[22] Scaling this practice across all national supercomputing resources, and encouraging universities to do the same, can help make HPC more accessible.

CLOUD COMPUTING IS BEST SUITED TO HIGHLY PARALLEL APPLICATIONS OR THOSE WITH VARIABLE DEMAND

HPC systems can either be implemented on-premises or in the cloud. Cloud computing provides virtual access to HPC resources, often over the Internet, and can be public or private. Private clouds provision exclusive access of computing resources to a single organization.[23] For instance, the U.S. National Aeronautics and Space Administration (NASA) has a center-wide, private cloud called the Goddard Private Cloud, which offers customizable operating environments to researchers so that they can process and analyze large datasets and securely share data.[24] Public clouds such as Amazon's Elastic Compute Cloud, on the other hand, provide open access to the general public.

The basic architecture of HPC systems used in cloud and on-premises computing are largely the same—that is, they both use single HPC systems or clusters of servers—but how users access these systems, and the services they can provide, differ. Cloud providers, such as Amazon Web Services (AWS), Microsoft Azure, and Google Cloud, host and maintain HPC systems and allow users to access computing power, storage, and databases, on an as-needed basis.[25] Such clouds give access to high-performance computing resources through a convenient network interface, making access easily available to those with an Internet connection, rather than restricting access only to those with access to a particular facility. Further, like an electrical grid, the resources available to a single job or project can vary from a single virtual CPU or GPU to a substantial fraction of the entire cloud, a feature often described as being "elastic."[26]

One benefit of cloud computing is it provides flexible, on-demand HPC resources to a wide range of scientific users in a relatively low-cost

environment, which is especially important for HPC users that are not wholly committed to, or technically capable of, justifying an in-house HPC system. For instance, an ongoing project at Los Alamos National Laboratory (LANL) involves searching for technical information in images that are openly available on the Internet.[27] Because images are being harvested from the web, there is a risk of inadvertently downloading objectionable or malicious content. To test whether AWS's image-recognition platform could prescreen large quantities of images quickly before researchers decided whether to bring them onto the LANL network, researchers put the cloud-based algorithm to task labelling pictures of bicycles. LLNL researchers found that the commercially available image-analysis algorithms were useful and cost effective in screening large quantities of bicycle images, although they were not fully tailored to identify all the types of bicycle images the researchers were looking for.

Another benefit of using cloud computing is improved resource utilization. Utilization is a common problem with on-premises HPC systems where computing capacity is fixed, leading to performance bottlenecks when demand is higher than supply, and creating wasted resources when demand is below supply. Cloud computing has an advantage because it matches computing supply to demand in an autonomic fashion, dynamically adding or removing resources.[28] This is particularly useful for applications that have sporadic and variable demand, such as electroencephalography (EEG) data analysis, which involves processing records of electrical activity in the brain.[29] Cloud computing is better suited for this application because it has the elasticity to fulfill the burst-like needs of processing EEG jobs without sitting idle for most of the time as an on-premises application would.

Cloud-based computing architectures are not well-suited to all applications, though. As researchers from Indiana University have shown, the applications that are best suited to cloud technologies are those that are highly parallelized, meaning those that have multiple processes that can be independently executed in different processing units.[30] Applications that have greater dependencies and require independent nodes working on separate problems to communicate with one another are better solved by on-premises systems because they have faster networks that enable nodes to communicate with each other better. For instance, predicting the weather in one country depends to a great extent on the weather in other places, which means making a forecast for one country needs to take into account forecasts elsewhere. Developing and operating a climate forecast model would therefore not be well-suited to parallel processing in the cloud. Similarly, data transfer costs on public clouds—though cheap and seemingly benign—can quickly eat into budgets for large simulations. It can be difficult to make clear comparisons between public cloud vendors, as

the pricing and fee structure for different solutions can change greatly from company to company.

MEASURING HPC PERFORMANCE ACCURATELY REQUIRES SUITABLE BENCHMARKS

HPC performance is typically measured by how many floating-point operations per second (flops) a system is capable of—a standard that evaluates binary and decimal arithmetic in computer programming environments.[31] There are two main benchmarks for these tests: The first is the High-Performance Linpack (HPL) benchmark, which tests a system's ability to perform large arithmetic operations using highly precise values. Since most non-AI HPC applications model phenomena in areas such as physics, chemistry, and biology, they need to be able to operate using highly precise numbers, such as Newton's gravitational constant (about 0.0000000000667) to accurately run simulations and process data. Representing these numbers using a standard computing format requires 64 bits of storage, which is why the HPL benchmark is typically referred to as testing 64-bit accuracy or "double-precision" math.[32] The other benchmark is the HPL-AI benchmark, which tests a supercomputer's ability to perform mixed-precision math.[33] Many applications that fuel advances in AI, such as machine learning, deep learning, and autonomous driving, achieve desired results at 32-bit and even lower precision formats because they do not need to use such highly precise values to accurately encode and manipulate the data needed to train and develop AI algorithms.

But research has shown that simple benchmarks such as these are, individually, rarely predictive of the performance of an application, and even collections of benchmarks give only a rough estimate. Other relevant measures of performance include memory size and bandwidth; data size and bandwidth; interconnect bandwidth and application sensitivity to interconnect latency; integer and floating-point performance; and long-term data storage requirements.[34] For instance, researchers from the University of California, San Diego and the U.S. Department of Defense (DOD) tested 10 HPC systems that span 9 distinct system architectures to see how well several simple and predictive metrics, including the HPL benchmark, could predict a range of different HPC applications.[35] They found that for some applications, metrics that are memory-oriented, such as the STREAM benchmark that measures sustainable memory bandwidth, were more predictive of system performance than the floating-point-oriented HPL benchmark. Further, their study shows that where application-specific computational requirements are understood, a few simple metrics can be combined and weighted appropriately to predict system performance with approximately 80 percent accuracy.[36] While this research focuses on applications and systems that use CPUs, researchers from the University of Virginia and semiconductor company AMD found similar results when

testing how accurate existing benchmarks are at predicting the performance of the applications that rely on GPUs.[37] Effectively predicting how well a system with CPUs or GPUs will support a particular application is best done by better understanding the computational requirements of an application, and testing a system's ability to run this application against a suitable set of key metrics.

Accurate measures of performance are important in guiding purchasers of HPC systems to the systems that best suit their goals, and for agencies such as DOE and NSF, which allocate valuable HPC resources, to more efficiently award computing resources. This will equip researchers with better resources, and enable them to leverage system capabilities to solve problems more quickly.

INCREASING ACCESS TO HPC IS IMPORTANT FOR MAINTAINING U.S. LEADERSHIP IN AI

The ability for public and private institutions to use HPC systems to leverage AI in solving problems and building new capabilities will determine, in part, the future of the United States' economy, national security, and society.

ECONOMY

From an economic perspective, access to computing resources is important in maintaining U.S. market share in AI products and services. AI is poised to make a significant impact on the global economy, adding $15.7 trillion to the GDP by 2030, but success in the AI economy depends on how effectively firms can leverage data to generate insights and unlock value.[38]

One area in which HPC-enabled AI is having significant impacts on the economy is the automotive industry, wherein advances in automotive and technology sectors have led to the emergence of connected and autonomous vehicles (CAVs). Autonomous and semi-autonomous vehicles are those in which at least some aspect of a safety-critical control function, such as steering, throttle, or braking, occurs without direct driver input.[39] Connected vehicles are those that identify threats and hazards on the roadway and communicate this information over wireless networks to give drivers alerts and warnings using a combination of technologies, such as advanced wireless communications, advanced vehicle-sensors, and GPS navigation.[40] Developing and training CAVs and simulating more efficient traffic flows at scale requires the power and performance only HPC can bring.[41] For example, the University of Michigan has invested in a supercomputer that supports machine learning applications to enable researchers in its MCity program, which develops intelligent transportation systems, to perform more complex simulations and better train deep learning models to recognize signs, pedestrians, and hazards.[42] While the

private sector is financially incentivized to support research into how to make autonomous vehicles more accurate and sensitive, public support for advanced research in areas that cover key public policy goals is still needed because many public policy goals related to autonomous vehicles, such as improving safety, achieving carbon neutrality, and enhancing military CAV capabilities, represent public goods that the private sector cannot or will not adequately provide.

Given automotive exports are a significant segment of all national exports, it is important the United States maintain its ability to compete globally in this industry. The United States exported $142 billion worth of vehicles and parts in 2018—more than any other U.S. industrial sector.[43] If the United States cedes industry leadership in the research, development, and integration of CAV innovations, it stands to lose both the direct economic benefits that come from automotive exports, and the indirect effects CAVs bring to other industries.[44] For instance, the global trucking and ground-shipping industry could experience a significant boost from the development of driverless vehicles, which industry analysts estimate could bring economic gains ranging from $100—500 billion per year by 2025.[45] CAVs may also shift personal transport toward shared autonomous vehicle fleet use, reducing demand for the construction of parking lots, and enabling more efficient utilization of land.

More broadly, data-driven AI research and development in areas such as transportation, manufacturing, and home automation need data-intensive computing capabilities to make products and services. Firms such as Microsoft, Google, Facebook, and Amazon have already driven the growth of the commercial AI ecosystem by establishing advanced, well-resourced research labs that, as Yann LeCun, chief AI scientist at Facebook pointed out, are a major distinction between the United States and its competitors.[46]

NATIONAL SECURITY

Greater access to HPC systems is critical in employing AI for defense innovation. The nature of warfare is rapidly changing to one wherein weapons systems and warfighters need to have an algorithmic and informational advantage to outmaneuver adversaries.[47] To maintain its strategic advantage by combining new capabilities with new concepts of operation—what has been termed the "third offset"—the federal government needs to ensure it is equipping its researchers with the HPC tools they need to drive defense innovation.[48]

For example, DOD scientists and engineers are using HPC systems at the Navy DOD Supercomputing Resource Center (DSRC) in Mississippi to improve the Navy's ability to forecast ocean environments.[49] Predicting an accurate model of the oceanic climate requires researchers to monitor

changes that include the movement of waters, heat and carbon content, freshwater, biogeochemistry, and sea levels, as well as studying the interactions of the ocean with the atmosphere, land, and ecosystems.[50] The complexity and level of analysis such models need to effectively inform Naval forces securing the seas can only be done using HPC systems—and to make these models even better, DOD should be looking to AI-enabled systems that can better weigh the relative importance of data regarding atmospheric and ocean dynamics.[51] Fortunately, DOD already recognizes this, having recently invested in a new AI supercomputer that is to be implemented at Navy DSRC, alongside two similar systems at the U.S. Army research laboratory and one new system at the U.S. Air Force research laboratory.[52]

AI and HPC can also improve the effectiveness of traditional military campaigns. For instance, U.S. nuclear weapons have not been tested since 1992, which was before the United States signed a treaty that banned all nuclear tests. However, researchers from the Lawrence Livermore National Laboratory (LLNL) and the University of California, Berkeley, have shown that machine learning can offer new ways to monitor the effectiveness, yield, and explosive capability of nuclear weapons without breaking this treaty.[53] Their research uses high-performance computing to model the various possible seismic events that might occur from deploying U.S. nuclear weapons, and machine learning to predict which of these is most likely to happen. In a similar example, researchers from the University of Tokyo launched a tool in 2018 that can predict the direction of radioactive material dispersion.[54] Radioactive materials are generally concentrated downwind of their origins when wind blows continuously in one direction. The researchers were able to show that machine learning can estimate dispersion directions using models of the atmosphere with an average success rate of 85 percent.

SOCIETY

From a social welfare perspective, broadening access to HPC systems could enable more researchers to use AI in tackling problems for the public good in areas such as health care and the environment.[55]

For instance, HPC has long been used in computational drug discovery and design, wherein techniques such as molecular simulation can help model a biological target associated with a disease, and identify drugs that might effectively bind to those targets and achieve a desired therapeutic outcome.[56] But when the range of possible drug compounds is large, this process can be very long and the costs of running many simulations extremely high, slowing down the creation of life-saving drugs. Machine learning can complement this process by initially screening the known range of drug candidates to focus testing and simulation only on those with the right features to be successful. According to a 2019 report from the

U.S. Government Accountability Office, machine learning can save research and development (R&D) costs of between $300 million and $400 million per successful drug by accelerating drug discovery.[57] These sentiments are echoed by scientists at LLNL who are combining AI, bioinformatics, and traditional supercomputing to help discover candidates for new antibodies and pharmaceutical drugs to combat COVID-19. Using two high-performance computing systems and AI algorithms, researchers at LLNL screened over 10^{40} antibodies capable of binding to the virus that causes COVID-19, and narrowed the potential antibody candidates to an initial set of just 20. In essence, the broad advantage of converging AI and HPC is, as Jim Brase, LLNL's deputy associate director for data science explained, "Now, we're not just searching blindly. We're actually creating structures that we think are in the proper part of the design space, then we do our evaluations on those."[58]

Similarly, HPC and data-intensive algorithms are important in advancing climate research. In order to fully understand climate change, research needs to be focused not only at global scales, but also on regional and local scales, using high-resolution global models, regional models, and observational datasets. For example, the World Climate Research Program has for over a decade coordinated tens of modeling groups in as many countries, running the same prescribed set of climate change scenarios on the most advanced supercomputers to produce petabytes of standardized output containing hundreds of physical variables spanning tens and hundreds of years.[59] These climate model simulations of the past, current, and future climate have become one of the foundational elements of climate science, and have had direct impact on climate policy.[60] The European Union has also recognized the opportunity HPC presents for accelerating carbon neutrality, as evidenced by the €2 million it contributed between 2015 and 2017 to a project in Spain that was applying HPC techniques to energy industry simulations, including using HPC to design more efficient wind turbines, develop more efficient combustion systems for biogas, and explore geophysics for hydrocarbon reservoirs.[61]

ACADEMIC HPC DEMAND FAR OUTWEIGHS SUPPLY

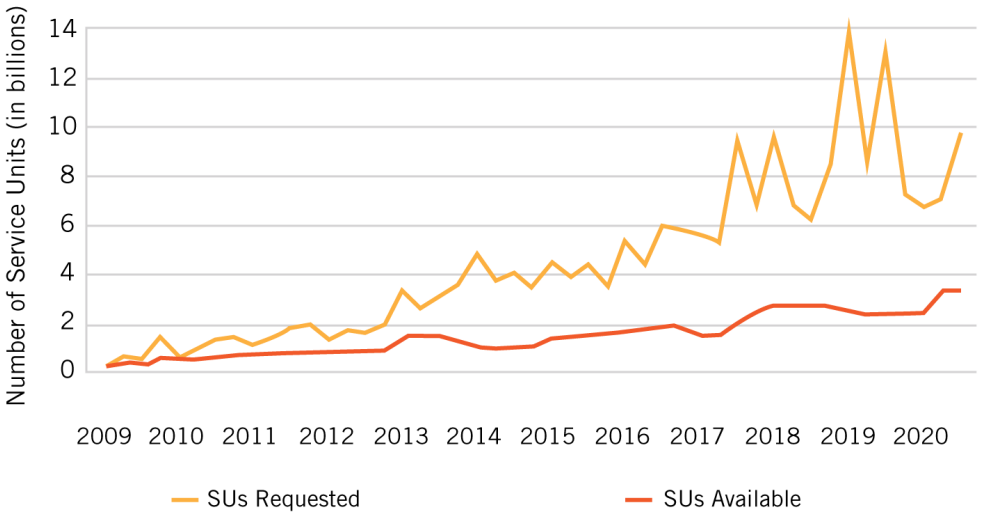
Publicly funded academic researchers requiring access to HPC capabilities for AI can either use systems that are hosted at their academic institutions or at national HPC centers. Allocations for computing time on HPC systems at the national level are made principally through competitive processes managed by DOE and NSF respectively. DOE's allocation program, the Innovative and Novel Computational Impact on Theory and Experiment (INCITE) program, is in its 15th year and awards 60 percent of supercomputing time at DOE's *Theta*, hosted at the Argonne National Laboratory, and *Summit*, the United States' most powerful supercomputer located at the Oak Ridge National Laboratory (ORNL).[62] Over its 15 years,

INCITE has allocated time on its leadership-class supercomputers to over 700 projects, and has been the chief method for researchers accessing the most powerful systems in the United States.[63] While there is not much available data on the number of computing hours applicants have requested through INCITE, the program has noted that the total number of node hours applicants requested in 2020 was more than three times what the program plans to award this year.

NSF's resource allocation program, the Extreme Science and Engineering Discovery Environment (XSEDE) program, is more wide-reaching, and coordinates the sharing of eight HPC systems hosted at four sites: Indiana University; the Texas Advanced Computing Center (TACC); the Pittsburgh Supercomputing Center (PSC); and the San Diego Supercomputer Center (SDSC).[64] XSEDE allocates resources in service units (SUs), which are equal to hours of processor core time but are more useful for allocators and users of compute time to compare allocations across HPC systems that may differ widely in both architecture and time of deployment.[65] SUs, however, are only based on the result of the High Performance LINPACK benchmark run on each system.[66] As explained earlier, excluding other relevant system parameters that may be more important to applications, such as memory or storage use, makes it difficult to ensure the most suitable systems are being allocated to researchers. This is particularly important as the demand for HPC systems is growing more quickly than the available supply, as shown in figure 2.

It is difficult to know exactly how much time on HPC systems the nation's researchers require, but one available metric between HPC supply and demand is the amount of computer time requested on XSEDE resources. As figure 2 shows, that gap has grown consistently over the last decade.

Figure 2: Comparing requested XSEDE service units (SUs) to available SUs[67]

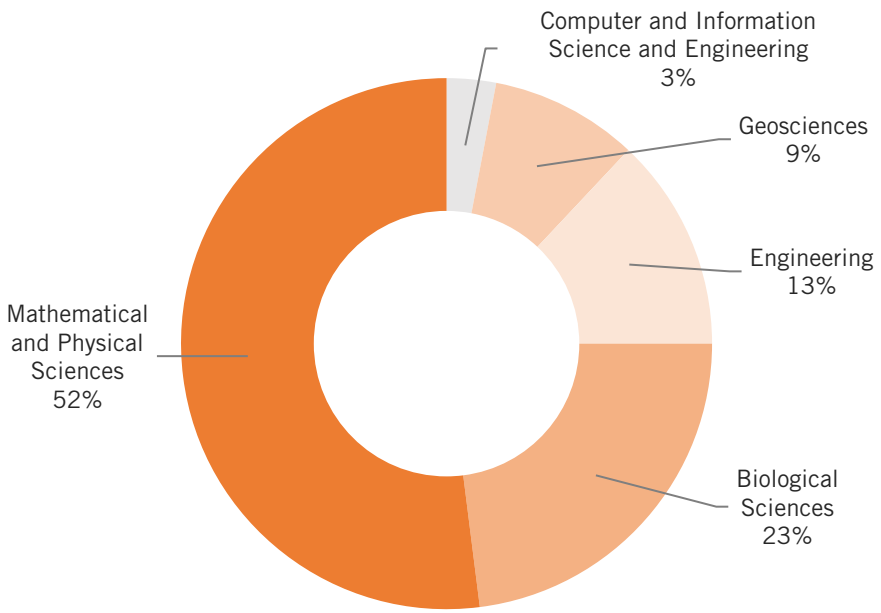


There is a growing gap between the number of service units requested for XSEDE resources and the number being made available, with demands for resources growing more quickly than supply. This implies insufficient computing resources are inhibiting the effective allocation of resources and hampering valuable work. The merit of this work is usually already established prior to the allocation process as most researchers applying for compute time have peer-reviewed funding awards and just need time on a HPC system. Those that do not already have funding are subject to an additional pre-review by the XSEDE resource allocation committee. In sum, a lack of resources is constraining already funded or merited research.[68]

Some XSEDE resources are optimized for AI, such as PSC's *Bridges-AI*, which is composed of 88 GPUs, but since all of XSEDE's resources are currently used for AI applications, we can infer that demand for HPC resources to conduct both AI and non-AI research outstrips the supply of resources.[69]

To get a sense of what areas of research are being funded, it is useful to look at a breakdown of XSEDE awards (in SUs) by research area. Figure 3 indicates that a wide range of areas covering a majority of NSF directorates are represented, including mathematical and physical sciences, geosciences, engineering, biological sciences, and computer and information science and engineering (CISE).

Figure 3: Allocated XSEDE service units by research area, in 2019[70]



Again, it is difficult to know exactly what percentage of allocations in each research area was for AI and non-AI applications as XSEDE does not

aggregate this data. However, looking at the list of active allocations provided on the XSEDE portal, it is easy to come across AI applications across a range of research areas. For example, one project allocated time at PSC led by a researcher from the University of Whitworth in Washington is investigating how AI can reduce the time to simulate molecular proteins.[71] Another, led by a researcher from ORNL, is using the systems at PSC and TACC to train AI algorithms to make causal inferences on large COVID-19 observational datasets.[72] A third, led by a researcher from the University of Pittsburgh, is generating metadata about historical letters dating from 1889 to 1940, and using AI to recognize patterns among them.[73] It is therefore likely that allocations to AI applications follow a similar breakdown to the total allocations shown in figure 3, with the majority of allocations given to those in AI applications in mathematical and physical sciences, biological sciences, and geosciences. Given how tight resources are and demand is continuing to grow, it is commendable to see a predominant share of allocations are focused on science and engineering, where the opportunities for production capabilities are large.

Further, to maximize the return on investment from AI research, NSF and DOE should explore more public-private partnerships. For example, one current XSEDE allocation led by a researcher at the University of California is investigating precipitation variability using machine learning.[74] If NSF were to facilitate a partnership between this project and, for example, a commercial weather forecasting service such as Accuweather, it could ensure research outcomes could be more easily used to create useful products and services, such as a tool to help make better forecasts. NSF already has a program called Partnerships for Innovation aimed at accelerating the path to market for new technologies by providing funding for NSF-backed projects to work with industry on R&D, but funding is only available for up to approximately 50 projects each year.[75] Similarly, NSF announced in August 2020 that it had partnered with the Department of Transportation, the Department of Homeland Security, and the Department of Agriculture to establish seven AI research institutes in six different states.[76] The new AI institutes have specific target application areas, such as trustworthy AI in weather, climate, and coastal oceanography, and are designed to be hubs of AI innovation that emphasize use-inspired research. So far, NSF has announced it has partnered with four companies—Accenture, Amazon, Google, and Intel—all of which are committed to solving AI problems of national importance.

Making more efficient use of resources will also require more accurately allocating resources to researchers. As discussed earlier, poor measures of system performance present an obstacle to matching users with the resources that best meet their particular application needs; however, there is also a gap in understanding what the computational requirements of a user's application are. Currently, proposals focus heavily on extracting

information on what peak speeds a project requires.[77] The proposal mechanism at DOE and NSF should be updated to extract more useful information about what capabilities an application needs from a system, such as interconnect bandwidth and long-term data storage requirements. However, agencies should not make the application process for researchers too cumbersome, as this creates an obstacle to accessibility.

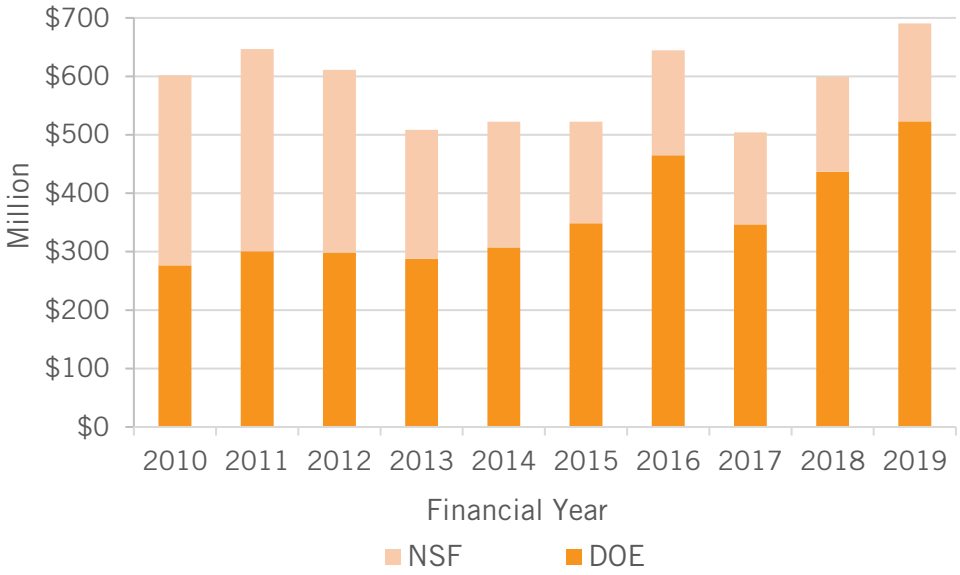
One method to better understand user requirements and engage the research community is to establish roadmaps. An HPC roadmap is a long-term plan that articulates what future investments an agency will make. Importantly, roadmaps do not suggest a single path to a destination, but rather multiple routes to a variety of goals. Such roadmaps could help make AI requirements concrete and relate them to future computing capabilities, facilitating planning by researchers, program directors, and facility operators at centers and on campuses over a longer time horizon. By capturing anticipated technology trends, the roadmaps could also provide guidance to those responsible for scientific software projects. In general, creating a strategy for acquiring the next generation of computing facilities could ensure researchers have access to the state-of-the-art systems they need to be productive and innovate at a higher rate than their competitors.

NSF NEEDS TO INCREASE ITS HPC INVESTMENT

NSF's investments in HPC have fallen considerably over the last decade, even as the gap between demand and available computing resources has grown.[78]

The Networking and Information Technology Research and Development (NITRD) program's National Coordination Office records trends in overall HPC investment by NSF and DOE. Figure 4 shows the total federal investment in high-performance computing infrastructure, which includes hardware, directly associated software, communications, storage, data management infrastructure, and other resources supporting HPC, from 2010 to 2019. This data was adjusted for inflation by using consumer price indices from the U.S. Bureau of Labor Statistics normalized to 2010.[79] The graph illustrates that while DOE's investments increased by approximately 90 percent from $276 million to $523 million (in constant 2010 dollars) during this period, NSF's investments fell by approximately 50 percent, from roughly $352 million to $167 million.

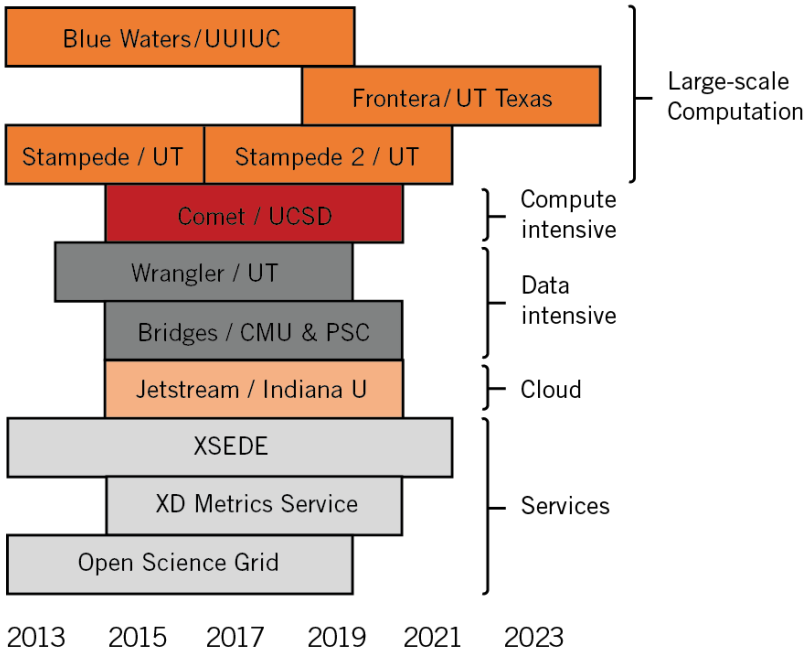
Figure 4: NSF and DOE investment in high-end computing infrastructure and applications by Networking and Information Technology Research and Development from 2010 to 2019[80]



It is important that NSF maintains high levels of HPC funding because, while both NSF and DOE invest in research to advance scientific understanding, NSF is primarily responsible for supporting the long tail of users that represent the majority of researchers and a significant proportion of research advances.[81] NSF's user base cuts across all federal agencies and academic fields, which means its investments are important in complementing those of DOE, DOD, and other agencies by promoting the scale and scope of impacts from state-of-the-art HPC. Because NSF investments in HPC systems support the majority of HPC users, insufficient NSF funding will not only cause the gap between HPC supply and demand to grow, but will accelerate the speed at which the gap grows.

Figure 5 shows that NSF's HPC portfolio includes a diverse set of resources but lacks a sufficient quantity of systems that can support increasing demand. As of 2019, NSF supported several leadership facilities, including *Blue Waters* at the University of Illinois, Urbana-Champaign and *Stampede* at the University of Texas, Austin; a high-speed, compute-intensive system, *Comet,* at the University of California, San Diego; high-bandwidth, data-intensive resources in *Wrangler* at University of Texas, Austin and *Bridges-AI* at the Pittsburgh Supercomputing Center; and a cloud resource, *Jetstream,* at Indiana University.

Figure 5: An overview of NSF's HPC systems and services, as of 2019



NSF's portfolio is well-suited to meet the diversity of future needs, but not the quantity of future needs. Most of the systems in NSF's portfolio have been decommissioned recently or will be soon, which means NSF has an opportunity now to consider which systems to invest in and how many, in order to best improve the nation's HPC capacity and capability for AI. To do this, NSF should invest in more moderate-sized, campus-based HPC systems. This makes sense for a number of reasons: First, most AI research projects that need HPC only require tens or hundreds of processors but need computations to be run many times.[82] Using national leadership facilities that are designed for the most demanding computational problems, and fitted with tens of thousands of cores and expensive internode networks for large I/O rates, would not be a cost-effective use of these resources. Second, the operating costs of campus systems, including power, cooling, and staffing, are typically borne by the host institution, which means NSF can better maximize their investment.[83]

Other nations are investing substantially in increasing access to HPC that will significantly buoy their ability to make AI advances, such as the European Union, which recently announced it plans to provide €8 billion over the next decade to develop, deploy, and extend the EU's HPC infrastructure.[84] It is time for the United States to equip its own researchers.

THE GOVERNMENT SHOULD INVEST IN STATES WHERE HPC USAGE IS LOW BUT AI RESEARCH IS HIGH

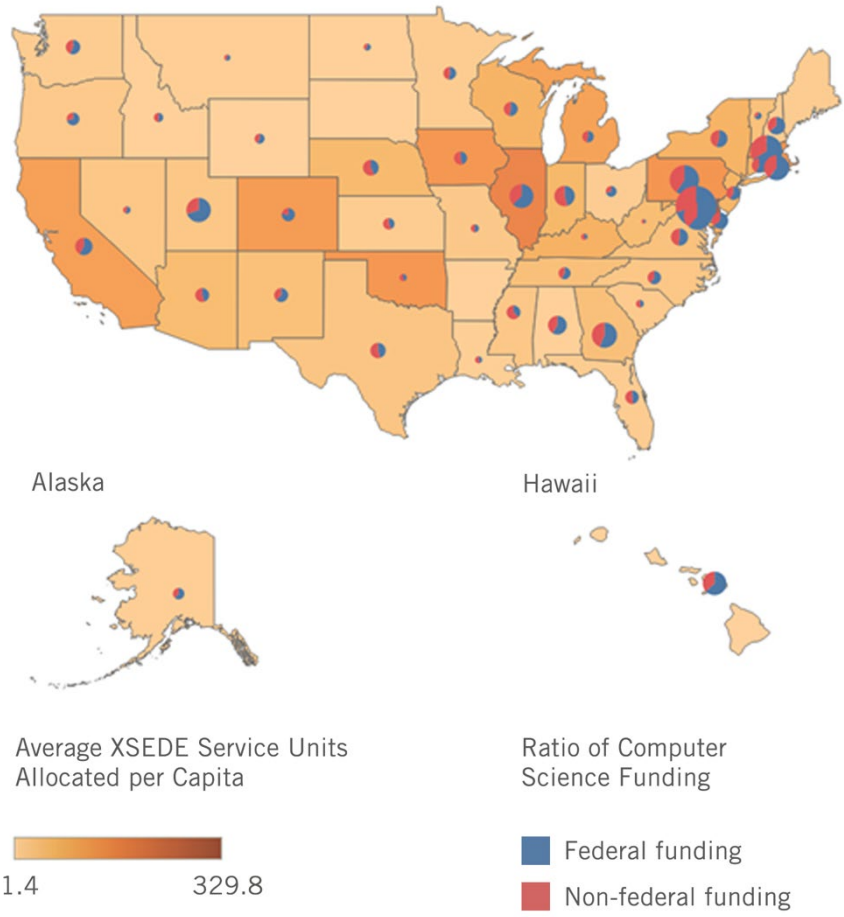
While the federal government should provide more access to center-class and mid-range systems for AI research, the questions are: Which states should the government invest in, and which institutions should it partner with? Since institutions themselves gain a competitive advantage in attracting and retaining faculty from providing HPC resources, another question is how much computing should individual institutions be responsible for providing? To demarcate institutional and NSF responsibility, policymakers should first measure how states are using current HPC resources for AI research.

The government should prioritize investing in cases wherein states have low levels of HPC availability but there is demonstrable evidence that the institutions within them are conducting high levels of AI research. This will allow the government to address instances wherein the gap between demand and supply is greatest. In cases wherein HPC availability is already high, federal investments in more HPC resources will not be the most effective way to close the gap because either institutions already have funding for HPC-enabled AI research and are using it, or they don't have research funding which means access to HPC is not the problem, research funding is. In cases wherein HPC availability and AI research is low, the government should require institutions to first increase funding for AI research or prove that they have sought partnerships with industry, as there is a risk that investments may not return increases in AI research.

To show the distribution of center-class HPC resources per capita, figure 6 uses XSEDE data on the service units researchers used in 2017, 2018, and 2019, as well as data on the researcher's organization and the state in which the organization is located.[85] We assume the more service units researchers in a given state are using, the more access they have to HPC resources.

To see the relationship between how much local access researchers in each state have to HPC resources and how much funding for AI research there is, figure 6 also uses data on how much funding the top institutions (defined by whether they are ranked among the top 500 research institutions) in each state are making in computer science, demarcating federal research expenditures from non-federal research expenditures.[86] This data is limited to R1 (very high research activity) and R2 (high research activity) universities, including only the 161 institutions that have annual federal research expenditures of $40 million or more overall. Given cloud computing is best suited to highly parallelized applications or those with variable demand, having local access is important in supporting the full range of AI applications.

Figure 6: Proportion of XSEDE service units allocated per capita and size of research funding from high-research institutions in computer science per capita in each state



The key insight from this figure is more access to center-class HPC resources is found in states that have leading academic institutions, which can either stand up their own HPC centers or partner with other leading research institutions in their state to create multi-institutional centers. Of these states, those with high levels of funding in computer science are the ones attracting and retaining the talent needed to foster HPC-enabled AI research. Little access to center-class HPC resources is found in states with few research academic institutions. But of these states, those with high levels of research funding for computer science are the ones attracting the partnerships, funding, and talent needed to invest in better systems.

Looking first to the states with high HPC usage and computer science research funding, indicates Massachusetts, Pennsylvania, and Illinois as some of the best performing states. Each used more than 120 service units per capita at their institutions between 2017 and 2019, and together

were the states with the second, third, and eighth highest funding in computer science research respectively. The key to these states' success is they have a number of top-ranked institutions that have partnered together to create supercomputing centers of excellence, thereby giving them a competitive advantage in attracting and retaining researchers, faculty, and federal research grants. In Pennsylvania, for instance, Carnegie Mellon University and the University of Pittsburgh partnered to create the Pittsburgh Supercomputing Center, which is one of the nation's leading supercomputing centers, particularly for AI applications.[87] PSC recently won a $5 million NSF award to build a new AI supercomputer which, in addition to its *Bridges-AI* supercomputer that exclusively uses GPUs, will further cement the center as a hub for HPC-enabled AI research.[88] Similarly, in Massachusetts, Boston University, Harvard University, the Massachusetts Institute of Technology (MIT), Northeastern University, and the University of Massachusetts have partnered to operate the Massachusetts Green High Performance Computing Center (MGHPCC), a world-class supercomputing center with an emphasis on fostering research collaborations in energy, climate, and the environment.[89] MGHPCC's resources are available to all partner university faculty, their students, and their collaborators for research and educational use in courses related to computational science.[90] This helps retain staff and attract students to the computer science programs at Massachusetts' universities, which are already some of the best in the world. MIT also hosts the Lincoln Laboratory Supercomputing Center (LLSC), a DOD-funded supercomputing facility focused on research to meet national security needs.[91] Through collaborations, student internship programs, and seminar series, LLSC and MIT share talents, facilities, and resources. For example, LLSC recently acquired an AI supercomputer optimized for training machine learning algorithms and performing deep neural network operations, providing MIT researchers with access to the 20th-best supercomputer in the United States, according to June's *Top 500* list.[92] In Illinois, the flagship institution of the University of Illinois hosts the National Center for Supercomputing Applications, one of the original national supercomputing centers NSF developed and deployed under its supercomputing program.[93] With support from NSF, the state of Illinois, industry partners, and other federal agencies, Illinois is able to provide researchers with lots of access to HPC resources, and is attracting AI researchers with large amounts of funding for computer science.

States with high levels of HPC usage and lower levels of research in computer science include California, Colorado, and Oklahoma. All of these states have large, well-funded HPC centers, but funding for computer science is either not highly prioritized or is limited by the tight budgets public institutions in these states face. In California, for example, the San Diego Supercomputing Center is a leading HPC center with four supercomputers, one of which is specifically designed to serve the "long

tail" of science.[94] This supercomputer, *Comet*, has 288 GPUs, making it well-suited to AI applications and computer science researchers. However, unlike Massachusetts, Illinois, or Pennsylvania, the majority of research institutions in California, Colorado, and Oklahoma are public ones that rely on federal and state funding, meaning they are subject to funding changes that come with changing policy priorities and economic conditions.[95] For example, Colorado State University, Fort Collins cut more than 355 faculty and staff positions between 2009 and 2013 in response to state budget cuts. Further, despite all having large-scale supercomputing centers, the top research facilities in these states are publicly funded and have less than the optimal amount of research in computer science. In Colorado, the National Center for Atmospheric Research (NCAR) provides the university community with world-class facilities to better predict severe weather, model flooding with decision-making tools, develop more detailed air quality forecasts, and simulate the impacts of climate change on the planet.[96] Machine learning models could help in each of these goals by more intelligently drawing useful information from data; making forecasts, models, and simulations more accurate; and reducing the budget, time, and resources needed to run simulations. But Colorado is trailing in AI research funding, ranking 28th in the United States in providing computer science research at top institutes. Similarly, Oklahoma has the OU Supercomputing Center for Education and Research, which supports 23 institutions across Oklahoma, including Langston University, the only Historically Black College and University (HBCU) in the state. But Oklahoma is 42nd in the United States in providing computer science research at top institutes. A lack of sufficient funding for computer science in states that already have significant HPC resources inhibits an easy opportunity to foster AI research.

States with low levels of HPC usage and high levels of research in computer science include Alabama, Indiana, Utah, and Georgia. These states have top research institutions that have shown they recognize the importance of R&D in AI and HPC, but lack access to powerful local HPC resources. At Indiana University, AI research has become a major focus after an alumnus and founder of cloud computing company ServiceNow, Fred Luddy, gifted the university $60 million to establish a multidisciplinary AI initiative at its school of computing and engineering.[97] This gift, the second largest in the school's history, has allowed the school to create a comprehensive research program covering several branches of AI, including machine learning, robotics, computer vision, and deep learning.[98] But until this year, the university did not have the HPC resources to support its research program. In January 2020, Indiana University unveiled its new supercomputer, *Big Red 200*, which has GPUs and memory components that specifically gear it toward AI applications.[99] When *Big Red 200* goes into production, it will help provide Indiana's AI researchers with the HPC access they need to compete with other states. Similarly, at the Georgia

Institute of Technology, AI and machine learning represent a large part of faculty and research interests.[100] The university's research efforts have been significant enough to attract a partnership from Sandia National Laboratories, a national R&D laboratory focused on testing nuclear weapons, and the Pacific Northwest National Laboratory, a national R&D laboratory focused on energy, national security, and the environment.[101] The plan is to launch a new research center focused on AI, and to support its continued competitive AI research. Georgia Tech has acquired a new supercomputer, *Hive*, through a $3.7 million NSF grant and a $1.6 million contribution by the university. The university has reserved about 20 percent of *Hive's* capacity to support the research activities of regional partners and HBCUs through the XSEDE program, with Morehouse College, Spelman College, and Clark Atlanta University all currently having time allotted on the system. This partnership and investment in HPC resources will provide Georgia's researchers with the access they need to push Georgia into being an HPC and AI competitor. In Utah, the Scientific Computing and Imaging (SCI) Institute is a research institute that focuses on conducting application-driven research in new scientific computing and visualization techniques and tools.[102] The SCI Institute's faculty and alumni are recognized around the world for their contributions to scientific computing, helping explain why Utah attracts large amounts of computer science funding.[103] But Utah's academic supercomputers are neither particularly large nor particularly powerful. The University of Utah, for example, has several HPC clusters, but none of them would be considered center-class systems.[104] Given Salt Lake City and Ogden have been identified among the top 10 regions in the United States whose workforces AI will impact the most, it is important for Utah to prepare its industries for the future by providing resources for AI experimentation.[105]

States with low levels of HPC usage and research in computer science include Arkansas, South Dakota, and Wyoming. These states lack the computer science funding needed to attract researchers, which in turn can attract funding for better facilities. In Arkansas, the high-performance computing center at the University of Arkansas does have some HPC resources, including *Trestles*, which is geared toward a diverse set of non-AI applications; and, as of 2019, *Pinnacle*, which has 20 GPUs. While these resources are important for creating an HPC ecosystem, as explained earlier, high-performance computing requires expertise.[106] The University of Arkansas has no significant funding in computer science, which impedes it from attracting skilled HPC facility staff and training the next generation of AI researchers. Similarly, the University of South Dakota has two modest HPC systems, *Lawrence* and *Legacy*, that it makes available to faculty and students, but South Dakota, together with Maine, are the only states to have no research facilities among the 161 institutions identified as conducting high-level research in any field.[107] Likewise, while the University of Montana recently won a $400,000 NSF grant to acquire a modest HPC

cluster, the new system, as the executive director of cyber infrastructure there described, will be close to what other universities had "15 years ago."[108]

In each of these scenarios, the actions needed to increase HPC-enabled AI research varies, from increasing funding in computer science to funding the acquisition of more HPC systems to doing both. Some institutions such as Indiana University, and more recently the University of Florida, have significant gifts from alumni that propel them into AI research hubs that independently increase access for the whole state and attract partnerships, talent, and more funding. But other states, such as Arkansas, may have never obtained an HPC system for AI without a $400,000 grant from NSF. The complexities involved in national HPC needs and the unpredictability of institutional HPC funding means federal policymakers need to have a coherent, detailed understanding of current HPC requirements at a sufficiently granular level in order to make informed, strategic decisions.

RECOMMENDATIONS

There are eleven key steps DOE, NSF, and Congress should be taking to ensure the United States can increase access to HPC resources for AI researchers.

1. CONGRESS SHOULD PROVIDE A TOTAL OF $10 BILLION IN HPC FUNDING OVER THE NEXT FIVE YEARS TO NSF AND DOE TO MATCH SUPPLY TO DEMAND.

Congress should significantly increase funding for HPC to both NSF and DOE to match current demand.

NSF funding in HPC fell by approximately 50 percent from $352 million in 2010 to $167 million (in constant 2010 dollars) in 2019. This level of funding supported less than a third of the demand for access to NSF's HPC systems in 2019 (as shown in figure 2). To meet the current demand for time on NSF's HPC resources, which is more than ten times what it was in 2010, Congress should increase NSF funding in HPC infrastructure to at least $500 million per year for the next five years.

DOE funding in HPC, on the other hand, has increased by approximately 90 percent from $276 million in 2010 to $523 million (in constant 2010 dollars) in 2019. However, current demand for access to DOE's HPC resources is still three times greater than what DOE is providing. Congress should also authorize DOE to increase funding in HPC infrastructure to at least $1.5 billion per year.

2. NSF SHOULD SUPPORT THE LONG TAIL OF POTENTIAL HPC USERS WHO REPRESENT THE MAJORITY OF RESEARCHERS.

States such as Alabama, Indiana, Utah, and Georgia have top research institutions that are conducting high levels of AI research but lack access to powerful local HPC resources for AI research. NSF should focus funding for mid-range and center-class systems in states that have low HPC availability but are conducting high levels of AI research. NSF should not focus HPC funding in states that already have high levels of HPC resource availability. Instead, for institutions in those states that have low levels of AI research, NSF should increase funding opportunities for AI research.

3. DOE AND NSF SHOULD ALLOCATE HPC COMPUTE TIME MORE EFFICIENTLY.

To accurately allocate HPC resources to researchers, DOE and NSF need to understand what users' computational requirements are. Currently, allocation proposal processes focus heavily on extracting information only on what peak HPC system speeds a given project requires. DOE and NSF should update their allocation proposal processes to extract additional information on what capabilities users need, beyond system speed. This should include details on the interconnect bandwidth and long-term data storage requirements researchers need.

4. DOE AND NSF SHOULD PROVIDE ACCESS TO HPC EXPERTS TO IMPROVE RESEARCHER PRODUCTIVITY.

Using an HPC system for any application requires the skills of well-trained advanced computing professionals. The larger and more complex the HPC system, the more expertise required to make them perform well. DOE and NSF should explore ways to provision HPC expertise in more effective and scalable ways to improve researcher productivity. For instance, NSF should make more widespread the XSEDE practice that permits researchers to request an allocation of staff time along with HPC time across all NSF-funded HPC systems. These staff experts are from the XSEDE partner sites whose time is devoted to working with allocated projects to accelerate progress toward research objectives. DOE should similarly expand supporting HPC staff resources for researchers at the Argonne Leadership Computing Facility and the Oak Ridge Leadership Computing Facility.

5. NSF AND DOE SHOULD INCREASE ACCESS TO HPC FOR MINORITY-SERVING INSTITUTIONS.

From 1997 to 2004, NSF supported the Education, Outreach, and Training Partnership for Advanced Computational Infrastructure (EOT-PACI), a partnership of dozens of institutions and organizations throughout the nation that coordinated the sharing of computing resources with Minority-Serving Institutions (MSIs). EOT-PACI was noted as broadening participation to over 50,000 underrepresented researchers.[109] NSF should re-establish and support such partnerships that coordinate the sharing of

computing resources with MSIs that include HBCUs, HSIs, and TCUs. For example, NSF has already established an initiative to enhance the participation of underserved communities in scientific research called the National Inclusion across the Nation of Communities of Learners of Underrepresented Discoverers in Engineering and Science (INCLUDES).[110] Since 2018, INCLUDES has provided more than $7 million in funding for the Computing Alliance of HSIs, a program working to help Hispanic Americans participate in, contribute to, and become leaders in computing. But this is currently the only initiative INCLUDES is funding that is aimed at advancing minority participation in computing.[111] NSF should provide similar funding for computing networks and collaborations that target HBCUs, such as the three-year initiative called the Alliance between Historically Black Universities and Research Universities for Collaborative Education and Research in Computing Disciplines, which NSF funded in 2006.[112] NSF should also provide similar funding for computing alliances targeting TCUs, such as the Tribal Computational Science Program that was once a part of EOT-PACI.[113] Additionally, NSF should re-establish targeted grants that fund HPC resources specifically at MSIs, such as the Minority Institutions Infrastructure grant. DOE, on the other hand, should re-establish projects that provide funding for MSIs with the aim of training minority students in HPC for eventual employment with the agency, such as the three-year project funded at Alabama A&M University from August 2006 to August 2009.[114]

6. NSF AND DOE SHOULD INCREASE ACCESS TO HPC FOR WOMEN.

Women are highly underrepresented in HPC, with research estimating that women make up less than 17 percent of the HPC workforce.[115] An analysis of participant data from nine HPC and HPC-related peer-reviewed conferences from 2017 also found that only 10 percent of all HPC paper authors were women, with representation of women being particularly low among industry researchers and at higher experience levels.[116] This is due in part to the gender imbalance of those with science and engineering backgrounds that ultimately make up the HPC community. To support the pipeline of qualified candidates in HPC, NSF and DOE should encourage more women to enroll and persist in engineering and science-based computing degrees by providing funding initiatives across the country that improve recruitment and retention rates of women in computing majors, such as the grants NSF has provided to the Mississippi Alliance for Women in Computing.[117] To lower the barrier for women gaining access to supercomputing resources, NSF and DOE should replicate the Blue Waters project at the National Center for Supercomputing Applications that created an allocations category open to researchers at U.S. academic institutions who are women.[118]

7. NSF SHOULD PROVIDE FUNDING TO DEVELOP HPC CURRICULA AT TWO-YEAR COLLEGES THAT ENABLE SEAMLESS TRANSFER INTO FOUR-YEAR COLLEGES.

The demand for a well-prepared HPC workforce is growing, but there is a gap between the technical skill sets needed, particularly at the entry level, and the number of individuals with adequate skills and training. This is partly because there is a leaky pipeline from the computer science pre-baccalaureate awards two-year colleges provide to the computer science bachelor's degrees four-year colleges provide, hindering many students from following a computer science workforce pathway.[119] To facilitate transfer pathways for students, NSF should provide funding for consortiums of both two-year and four-year colleges to work together in developing HPC curricula that ensure students with terminal two-year computer science degrees or who transfer from a community college have covered all of the lower division coursework they need to in order to seamlessly move into upper division coursework with little duplication of technical fundamentals or need for remediation. Western Oregon University, for example, has worked with two community colleges to revise their own information systems curriculum to ensure students with two-year computer systems and information technology degrees can transfer directly into upper division courses. NSF should re-establish the funds it has provided in the past to encourage other institutions to do the same.[120]

8. NSF SHOULD DIVERSIFY THE PORTFOLIO OF HPC RESOURCES IT MAKES AVAILABLE TO AI RESEARCHERS.

Cloud computing gives access to high-performance computing resources through a convenient network interface, making access easily available to those with an Internet connection, rather than restricting access only to those with access to a particular facility. Congress should authorize the National AI Research Resource Task Force Act of 2020 introduced by Rep. Eshoo (D-CA) that directs NSF to establish a task force focused on developing a public national cloud computing resource for AI research.[121] In addition, NSF currently only awards computing time on one private cloud environment, *Jetstream*, hosted by Indiana University through its XSEDE program. NSF should expand its HPC portfolio to enable awards of more cloud services by investing in more private clouds and partnering with multiple public clouds. For example, approved users could receive a budget to spend with their chosen commercial cloud provider.

9. NSF SHOULD ESTABLISH AND PUBLISH ROADMAPS THAT ARTICULATE WHAT FUTURE INVESTMENTS IT WILL MAKE.

Creating a long-term plan for acquiring the next generation of computing facilities ensures researchers have access to the state-of-the-art systems they need to be productive and innovate at a higher rate than their competitors, as well as providing guidance to those responsible for scientific software projects. DOE has already established roadmaps as part of its Exascale Computing Initiative, and NSF should follow suit. NSF should

annually collect community requirements and publish roadmaps that allow it to better set HPC priorities and make more strategic decisions that reflect user requirements. This should be led by a national AI research resource task force, such as the one proposed in the National AI Research Resource Task Force Act of 2020.

10. NSF SHOULD FOSTER MORE PUBLIC-PRIVATE PARTNERSHIPS.

Maximizing returns on investment in AI research will require mechanisms to effectively translate basic AI research into products and services for the marketplace. NSF already has a program called Partnerships for Innovation aimed at accelerating the path to market for new technologies by providing funding for NSF-backed projects to work with industry on R&D, but funding is only available for up to approximately 50 projects each year. NSF should at least triple the number of awards it grants through this program from 50 to 150. Similarly, as part of its National AI Research Institutes initiative, NSF has partnered with four companies—Accenture, Amazon, Google, and Intel—all of which are committed to solving AI problems of national importance. NSF should also support proposals that are of regional importance, including those that expand regional research capabilities and align with regional economic development goals. To that end, NSF should encourage proposals that foster collaboration and partnerships between universities, local businesses, and state and local governments.

11. DOE AND NSF SHOULD ADOPT NEW TOOLS AND PROCESSES TO ENSURE GRANTEES ARE USING HPC RESOURCES WISELY AND EFFICIENTLY.

Maximizing returns on investment in HPC will require careful resource management informed by an understanding of how existing grantees are using HPC systems. DOE and NSF should require all institutions that receive funding for HPC resources to adopt auditing tools, such as the XDMoD tool, that report on how optimally they are using HPC systems. NSF should then also establish a regular process for reviewing center-class institutions and create operational follow-ups with grantees to obtain feedback it can use to update its HPC strategy.

REFERENCES

1 . Academic Press, Dictionary of Science and Technology, (London: Academic Press Inc., 1992), 2135.

2. "CPU vs. GPU: Making the Most of Both," accessed September 15, 2020, https://www.intel.com/content/www/us/en/products/docs/processors/cpu-vs-gpu.html.

3. Rob Latham and Rob Ross et al., "HPC I/O for Computational Scientists" (presented at Argonne National Laboratory, January, 2014), https://press3.mcs.anl.gov//atpesc/files/2017/08/ATPESC_2017_Track-3_07_8-4_130pm_Carns-Understanding_IO.pdf.

4. "NEC Delivers SX-ACE Vector Supercomputers for use as the Earth Simulator," last modified May 26, 2015, https://www.nec.com/en/press/201505/global_20150526_02.html.

5. "Earth Simulator System Overview, Hardware," accessed September 18, 2020, http://www.jamstec.go.jp/es/en/system/hardware.html.

6. Ibid.

7. Shane Cook, "Memory Handling with CUDA" in *CUDA Programming: A Developer's Guide to Parallel Computing with GPUs* (Massachusetts: Elsevier, 2013), 107–202, https://doi.org/10.1016/B978-0-12-415933-4.01001-2.

8. "What is an HPC cluster," accessed September 20, 2020, https://www.hpc.iastate.edu/guides/introduction-to-hpc-clusters/what-is-an-hpc-cluster.

9. National Research Council, *Getting Up to Speed: The Future of Supercomputing* (Washington, D.C.: The National Academies Press, 2005), 23, https://doi.org/10.17226/11148.

10. "Maverick2 User Guide," last modified October 21, 2020, https://portal.tacc.utexas.edu/user-guides/maverick2.

11. "Considerations for Managing HPC Resources," last modified July 31, 2019, https://insidehpc.com/2019/07/considerations-for-managing-hpc-resources.

12. Stephen J. Ezell and Robert D. Atkinson, "The Vital Importance of High-Performance Computing to U.S. Competitiveness" (ITIF, April, 2016), http://www2.itif.org/2016-high-performance-computing.pdf?_ga=2.129077516.1271053775.1598903059-679502762.1578912342.

13. "Math Libraries and Interactive Tools," accessed October 2, 2020, https://hpc.llnl.gov/manuals/mathematical-software/math-libraries-and-interactive-tools; "Livermore Computing Resources and Environment";https://computing.llnl.gov/tutorials/lc_resources/#SoftwareLists.

14. "Tensorflow Release 1.0.0," last modified February 11, 2017, https://github.com/tensorflow/tensorflow/blob/07bb8ea2379bd459832b23951fb20ec47f3fdbd4/RELEASE.md; "Pytorch alpha-1 release," last modified September 1, 2016, https://github.com/pytorch/pytorch/releases/tag/v0.1.1.

15. "Deep Learning Software," accessed October 1, 2020, https://developer.nvidia.com/deep-learning-software.

16. Stephen J. Ezell and Robert D. Atkinson, "The Vital Importance of High-Performance Computing to U.S. Competitiveness" (ITIF, April, 2016), http://www2.itif.org/2016-high-performance-computing.pdf?_ga=2.129077516.1271053775.1598903059-679502762.1578912342.

17. Andrew Jones, "Should programming supercomputers be hard?" *ZDNet*, October 1, 2019, https://www.zdnet.com/article/should-programming-supercomputers-be-hard/.

18. "Graduate Level Educational Competencies for Computational Science Overview," accessed October 2, 2020, http://hpcuniversity.org/educators/gradCompetencies.

19. National Academies of Sciences, Engineering, and Medicine, *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017–2020* (Washington, D.C.: The National Academies Press), 72, https://doi.org/10.17226/21886.

20. National Security Commission on Artificial Intelligence (NSCAI), First Quarter Recommendations (Washington, DC: NSCAI, 2020), 21, https://www.nscai.gov/reports.

21. Marie Gottschalk, "If You Want Engaged Employees, Offer Them Stability," *Harvard Business Review,* August 15, 2019, https://hbr.org/2019/08/if-you-want-engaged-employees-offer-them-stability.

22. "XSEDE Allocations Info & Policies," last modified June 21, 2020, https://portal.xsede.org/allocations/policies.

23. National Institute of Standards and Technology (NIST), The NIST Definition of Cloud Computing (Washington, D.C.: NIST, 2011), 2, https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-145.pdf.

24. "Cloud Computing," accessed October 3, 2020, https://www.nccs.nasa.gov/services/cloud-computing.

25. "Amazon EC2 Overview," accessed October 3, 2020, https://aws.amazon.com/ec2/.

26. National Academies of Sciences, Engineering, and Medicine, *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017–2020* (Washington, D.C.: The National Academies Press), 112, https://doi.org/10.17226/21886.

27. Los Alamos National Laboratory, *The ISTI Rapid Response on Exploring Cloud Computing 2018* (Washington D.C.: LANL, 2018), 38, https://permalink.lanl.gov/object/tr?what=info:lanl-repo/lareport/LA-UR-18-31581.

28. Gary Andrew McGilvary et al., "Ad-hoc Cloud Computing: From Concept to Realization," *IEEE International Conference on Cloud Computing*, (2015): 1063–1068, doi: 10.1109/CLOUD.2015.153.

29. Zoltan Juhasz, "Quantitative Cost Comparison of On-premise and Cloud Infrastructure Based EEG Data Processing," *Cluster Computing* (2020), https://doi.org/10.1007/s10586-020-03141-y.

30. Jaliya Ekanayake et al., “High Performance Parallel Computing with Cloud and Cloud Technologies,” *Cloud Computing and Software Services* (2010), doi: 10.1201/EBK1439803158-c12.

31. Institute of Electrical and Electronics Engineers (IEEE), *IEEE 754-2019 – IEEE Standard for Floating-Point Arithmetic* (New York: 2019),”https://standards.ieee.org/standard/754-2019.html.

32. David Greaves, “Floating Point Computation” (presented at the University of Cambridge February, 2010), https://www.cl.cam.ac.uk/teaching/1011/FPComp/fpcomp10slides.pdf.

33. “HPL-AI Mixed-Precision Benchmark,” last modified November 13, 2019, https://icl.bitbucket.io/hpl-ai/.

34. National Academies of Sciences, Engineering, and Medicine, *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017–2020* (Washington, D.C.: The National Academies Press), 81, https://doi.org/10.17226/21886.

35. L. Carrington et al., “How Well Can Simple Metrics Represent the Performance of HPC Applications?” *SC '05: Proceedings of the 2005 ACM/IEEE Conference on Supercomputing* (2005), doi: 10.1109/SC.2005.33.

36. Ibid.

37. Shuai Che and Kevin Skadron, “BenchFriend: Correlating the performance of GPU benchmarks,” *International Journal of High Performance Computing Applications* vol.28 (2013), 238–250, doi:10.1177/1094342013507960.

38. PricewaterhouseCoopers, “Global Artificial Intelligence Study: Exploiting the AI Revolution” (2017), https://www.pwc.com/gx/en/issues/data-and-analytics/publications/artificial-intelligence-study.html.

39. “Connected/Automated Vehicles,” accessed October 10, 2020, https://www.ite.org/technical-resources/topics/connected-automated-vehicles.

40. Ibid.

41. “Autonomous & Connected Vehicles,” accessed October 10, 2020, https://robotics.umich.edu/research/focus-areas/autonomous-connected-vehicles.

42. “How HPC, AI, and IoT Drive the Future of Smarter Vehicles,” *The Next Platform,* January 7, 2020, https://www.nextplatform.com/2020/01/07/how-hpc-ai-and-iot-drive-the-future-of-smarter-vehicles.

43. American Automotive Policy Council, *State of the U.S. Automotive Industry* (2020), 2, http://www.americanautocouncil.org/sites/aapc2016/files/AAPC%20ECR%20Q3%202020.pdf.

44. National Science & Technology Council and the United States Department of Transportation (U.S. DOT), “Ensuring American Leadership in Automated Vehicle Technologies” (Washington D.C.: U.S. DOT, 2020), https://www.transportation.gov/sites/dot.gov/files/docs/policy-initiatives/automated-vehicles/360956/ensuringamericanleadershipav4.pdf.

45. McKinsey, “Disruptive technologies: Advances that will transform life, business, and the global economy” (2013), 83 https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Disruptive%20technologies/MGI_Disruptive_technologies_Full_report_May2013.pdf.

46. Cai Yiwen, “Facebook AI Exec: China Lags Behind US Due to Lack of Top Labs,” *Sixth Tone,* March 24, 2017, http://www.sixthtone.com/news/2109/facebook-ai-exec3A-china-lags-behind-us-due-to-lack-of-top-labs.

47. Elsa B. Kania, “Chinese Military Innovation in Artificial Intelligence” (Center for New American Security, June 2019), https://www.cnas.org/publications/congressional-testimony/chinese-military-innovation-in-artificial-intelligence; Alina Polyakova, “Weapons of the weak: Russia and AI-driven asymmetric warfare” (Brookings Institute, November 2018), https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/.

48. Mackenzie Eaglen, “What is the Third Offset Strategy?” *Real Clear Defense*, February 16, 2019, https://www.realcleardefense.com/articles/2016/02/16/what_is_the_third_offset_strategy_109034.html.

49. “About the Navy DSRC,” last modified August 13, 2020, https://www.navydsrc.hpc.mil/about/index.html.

50. Detlef Stammer et al., “Ocean Climate Observing Requirements in Support of Climate Research and Climate Information,” *Frontiers in Marine Science* vol.6 (2019), https://doi.org/10.3389/fmars.2019.00444.

51. Renee Cho, “Artificial Intelligence—A Game Changer for Climate Change and the Environment,” *Columbia University State of the Planet Blog,* June 5, 2018, https://blogs.ei.columbia.edu/2018/06/05/artificial-intelligence-climate-environment/.

52. Tiffany Trader, “DOD Orders Two AI-Focused Supercomputers from Liqid,” *HPCWire*, August 24, 2020, https://www.hpcwire.com/2020/08/24/dod-orders-two-ai-focused-supercomputers-from-liqid/.

53. “23 September 1992 - Last U.S. Nuclear Test,” accessed November 2, 2020, https://www.ctbto.org/specials/testing-times/23-september-1992-last-us-nuclear-test.

54. Yoshikane, T. and Yoshimura, K., “Dispersion characteristics of radioactive materials estimated by wind patterns,” *Scientific Reports*, vol. 8 (2018), article no. 9926, July 2, 2018, https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1038%2Fs41598-018-27955-4.

55. Robert D. Atkinson, Mark Muro, and Jacob Whiton, “The Case for Growth Centers” (ITIF and Brookings 2019), http://www2.itif.org/2019-growth-centers.pdf.

56. Stephen J. Ezell and Robert D. Atkinson, “The Vital Importance of High-Performance Computing to U.S. Competitiveness” (ITIF, April 2016), http://www2.itif.org/2016-high-performance-computing.pdf?_ga=2.129077516.1271053775.1598903059-679502762.1578912342.

57. United States Government Accountability Office (GAO), *Artificial Intelligence in Health Care* (Washington D.C.: GAO, December 2019), https://www.gao.gov/assets/710/703558.pdf.

58. Jeremy Thomas, "Lab antibody, anti-viral research aids COVID-19 response," *Lawrence Livermore National Laboratory Blog,* March 26, 2020, https://www.llnl.gov/news/lab-antibody-anti-viral-research-aids-covid-19-response.

59. Veronika Eyring et al., "Towards improved and more routine Earth system model evaluation in CMIP," *Earth Systems Dynamics* vol.7 (2016), 813-830, https://esd.copernicus.org/articles/7/813/2016/esd-7-813-2016.html.

60. Veronika Eyring et al., "Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization," *Geoscientific Model Development* vol.9 (2015), 1937–1958, https://gmd.copernicus.org/articles/9/1937/2016/gmd-9-1937-2016.html.

61. "European Commission: HPC for Energy," last modified July 3, 2017, https://cordis.europa.eu/project/id/689772.

62. Katie Bathea, "In its 15th year, INCITE advances open science with supercomputer grants to 47 projects," Argonne Leadership Computing Facility Blog, November 18, 2019, https://www.alcf.anl.gov/news/its-15th-year-incite-advances-open-science-supercomputer-grants-47-projects.

63. Ibid.

64. "XSEDE Resources," last modified January 17, 2019, https://www.xsede.org/ecosystem/resourceshttps://www.xsede.org/ecosystem/resources.

65. "SU Converter," accessed October 20, 2020, https://www.xsede.org/su-converter.

66. Ibid.

67. Data from Open XDMoD, University at Buffalo (J.T. Palmer et al., Open XDMoD: A tool for the comprehensive management of high-performance computing resources, Computing in Science and Engineering 17.4(2015): 52–62, 2015). Custom query by Robert L. DeLeon.

68. "XSEDE Allocations Info & Policies," last modified June 21, 2020, https://portal.xsede.org/allocations/policies#71.

69. "Bridges User Guide," last modified June 19,2020, https://portal.xsede.org/psc-bridges.

70. Data obtained by querying XDMoD database for XSEDE SUs charged by NSF Directorate between 01-01-2019 to 12-31-2019. Jeffrey T. Palmer et al., "Open XDMoD: A Tool for the Comprehensive Management of High-Performance Computing Resources," Computing in Science & Engineering, Vol 17, Issue 4, 2015, 52–62. 10.1109/MCSE.2015.68.

71. Kent Jones et al., "Supporting Whitworth Science faculty with High Performance Molecular Modelling and AI Applications, Whitworth University Molecular Biosciences" (presented on XSEDE User Portal, accessed October 2020), https://portal.xsede.org/allocations/current.

72. Yan Liu et al., "High-Performance Causal Inference for COVID-19 Mitigation and Response, Oak Ridge National Laboratory" (presented on

XSEDE User Portal, accessed October 2020), https://portal.xsede.org/allocations/current.

73. Raja Adal et al., "Applications of Computer Vision and Machine Learning to Historical Documents, University of Pittsburgh" (presented on XSEDE User Portal, accessed October 2020), https://portal.xsede.org/allocations/current.

74. Fiaz Ahmed "Investigating precipitation variability using Machine Learning Methods" (presented on XSEDE User Portal, accessed October 2020), https://portal.xsede.org/allocations/current.

75. "Partnerships for Innovation (PFI)" accessed October 21, 2020, https://www.nsf.gov/pubs/2019/nsf19506/nsf19506.htm

76 . "NSF advances artificial intelligence research with new nationwide institutes," last modified August 26. 2020, https://www.nsf.gov/news/special_reports/announcements/082620.jsp.

77. National Academies of Sciences, Engineering, and Medicine, *Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017–2020* (Washington, D.C.: The National Academies Press), 81, https://doi.org/10.17226/21886.

78. Networking & Information Technology Research & Development (NITRD), *Supplement to the President's FY2021 Budget* (Washington D.C.: NITRD, 2020), 23, https://www.nitrd.gov/pubs/FY2021-NITRD-Supplement.pdf.

79. Bureau of Labor Statistics, CPI for All Urban Consumers (CPI-U), https://www.bls.gov/cpi/data.htm.

80. National Coordination Office for the Networking and Information Technology Research and Development program, Supplement to the President's Budget FY2010 – FY2019 (actual nominal investment in High-Capability Computing Infrastructure and Applications coverted to real dollar amount in 2010), https://www.nitrd.gov/Publications/SupplementsAll.aspx.

81. "The NSF Mission NSF Act of 1950," accessed October 30, 2020, https://www.nsf.gov/pubs/1995/nsf9524/mission.htm.

82. Thomas R. Furlani et al., "Using XDMoD to facilitate XSEDE operations, planning and analysis," *XSEDE '13: Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery*, vol.46 (2013), 1–8, https://doi.org/10.1145/2484762.2484763.

83. Thomas R. Furlani et al., "Submission in Response to NSF CI 2030 Request for Information," *National Science Foundation*, March 26, 2017, https://www.nsf.gov/cise/oac/ci2030/pdf/RFI-Furlani-195_with_Attachment.pdf.

84. "European Commission: High Performance Computing," accessed October 17, 2020, https://ec.europa.eu/digital-single-market/en/high-performance-computing.

85. Data obtained by querying XDMoD database for XSEDE service units by state based on principal investigator location between 01-01-2017 and 12-31-2017. Jeffrey T. Palmer et al,, "Open XDMoD: A Tool for the Comprehensive Management of High-Performance Computing Resources," Computing in Science & Engineering, Vol 17, Issue 4, 2015, 52–62. 10.1109/MCSE.2015.68.

86. John V. Lombardi, Craig W. Abbey, and Diane D. Craig, "The Top American Research Universities," *The Center for Measuring University Performance (CMUP) 2018 Annual Report*, (CMUP, 2019), 13–16, https://mup.umass.edu/sites/default/files/mup-2018-top-american-research-universities-annual-report.pdf.

87. "About PSC," accessed October 14, 2020, https://www.psc.edu/homepage/about-psc.

88. "Category II: Unlocking Interactive AI Development for Rapidly Evolving Research," accessed October 25, 2020, https://www.nsf.gov/awardsearch/showAward?AWD_ID=2005597&HistoricalAwards=false.

89. "About the MGHPCC," accessed October, 22, 2020, https://www.mghpcc.org/about/about-the-mghpcc/.

90. "Resources," accessed October, 22, 2020, https://www.mghpcc.org/resources/.

91. "About Lincoln Laboratory," accessed October, 22, 2020, https://www.ll.mit.edu/about.

92. Kylie Foy, "Lincoln Laboratory's new AI supercomputer is the most powerful at a university," *Lincoln Laboratory Newspage,* September 26, 2019, https://www.ll.mit.edu/news/lincoln-laboratorys-new-ai-supercomputer-most-powerful-university.

93. "About NCSA," accessed October, 23, 2020, http://www.ncsa.illinois.edu/about.

94. "Comet User Guide," accessed October, 23, 2020, https://www.sdsc.edu/support/user_guides/comet.html.

95. Susan K. Urahn et al., "Two Decades of Change in Federal and State Higher Education Funding" (Pew Trusts, October 2019), https://www.pewtrusts.org/en/research-and-analysis/issue-briefs/2019/10/two-decades-of-change-in-federal-and-state-higher-education-funding.

96. "NCAR: Who We Are," accessed October 24, 2020, https://ncar.ucar.edu/who-we-are.

97. "Landmark $60M gift to establish major initiative in artificial intelligence at Indiana University," *Indiana University Bloomington Newspage,* October 18, 2019, https://luddy.indiana.edu/research/research-areas/artificial-intelligence.html.

98. Ibid.

99. "About Big Red 200 at IU," last modified November 2. 2020, https://kb.iu.edu/d/brcc.

100. "Artificial Intelligence & Machine Learning," accessed November 4, 2020, https://ic.gatech.edu/content/artificial-intelligence-machine-learning.

101. John Toon, "National Labs, Georgia Tech, Collaborate on AI Research," *Georgia Tech News Center*, November 5, 2019, https://news.gatech.edu/2019/11/05/national-labs-georgia-tech-collaborate-ai-research.

102. "Scientific Computing and Imaging History," accessed November 15, 2020, https://www.sci.utah.edu/home.html.

103. Ben Shneiderman, *The New ABCs of Research: Achieving Breakthrough Collaborations* (Oxford: Oxford University Press, 2016), 320.

104. "Resources available at CHPC, HPC Clusters," accessed November 18, https://www.chpc.utah.edu/resources/HPC_Clusters.php.

105. Mark Muro, Jacob Whiton, and Robert Maxim, "What jobs are affected by AI? Better-paid, better-educated workers face the most exposure," (Brookings, November 2019), 19, https://www.brookings.edu/research/what-jobs-are-affected-by-ai-better-paid-better-educated-workers-face-the-most-exposure/.

106. "Arkansas High Performance Computing Center: Resources," accessed November 23, 2020, https://hpc.uark.edu/hpc-resources/index.php.

107. "University of South Dakota: High Performance Computing," accessed November 23, 2020, https://www.usd.edu/technology/research/high-performance-computing.

108. Addie Slanger and Montana Kaimin, "Gadgets galore: UM is upgrading to new supercomputer," *Montana Kaimin College News Blog,* September 12, 2019, http://www.montanakaimin.com/news/gadgets-galore-um-is-upgrading-to-new-supercomputer/article_ea53fe1a-d583-11e9-ac5c-4f8109aa0009.html.

109. "NSF EPIC Press Release," last modified April 2005, http://mvhs.shodor.org/epic/pressrelease.html.

110. "About Us: NSF INCLUDES," accessed November 6, 2020, https://www.includesnetwork.org/new-a/about-us.

111. "NSF INCLUDES Alliance: Computing Alliance of Hispanic-Serving Institutions," accessed November 6, 2020, https://www.nsf.gov/awardsearch/showAward?AWD_ID=1834620&HistoricalAwards=false.

112. "BPC-A: Collaborative Research: Alliance between Historically Black Universities and Research Universities for Collaborative Education and Research in Computing Disciplines," accessed November 6, 2020, "https://www.nsf.gov/awardsearch/showAward?AWD_ID=0540561.

113. "A Strong but Sensitive Computing Initiative for Native American Communities," accessed November 6, 2020, https://www.sdsc.edu/pub/envision/v16.2/native-americans.html.

114. Qian, Xiaoqing and Deng, Z. T., "Alliance for Computational Science Collaboration: HBCU Partnership at Alabama A&M University Continuing High Performance Computing Research and Education at AAMU," November 2009, doi:10.2172/967143.

115. Athina Frantzana, "Women's representation and experiences in the high performance computing community," PhD thesis, The University of Edinburgh, 2019, https://era.ed.ac.uk/handle/1842/36127.

116. Eitan Frachtenberg and Rhody Kaner, "Representation of Women in High-Performance Computing Conferences," EasyChair no. 2799, February 28, 2020, https://easychair.org/publications/preprint_open/2nVv.

117. "NSF INCLUDES: Mississippi Alliance for Women in Computing (MAWC)," accessed November 6, 2020, https://www.nsf.gov/awardsearch/showAward?AWD_ID=1649312&HistoricalAwards=false

118. “Blue Waters Awards 21 Broadening Participation Allocations,” *HPCWire*, April 26, 2018, https://www.hpcwire.com/off-the-wire/blue-waters-awards-21-broadening-participation-allocations.

119. Louise A. Lyon and Jeffrey Forbes, “Lighting the Path: From Community College to Computing Careers” (Association for Computing Machinery, October 2018), https://www.acm.org/binaries/content/assets/education/lighting-the-path-from-community-college-to-computing-careers.pdf.

120. “National Center of Excellence for High Performance Computing Technology,” accessed November 6, 2020, https://www.nsf.gov/awardsearch/showAward?AWD_ID=0202452.

121. National AI Research Resource Task Force Act of 2020, H.R.7096, 116th Cong. (2020).

ABOUT THE AUTHOR

Hodan Omaar is a policy analyst at the Center for Data Innovation. Previously, she worked as a senior consultant on technology and risk management in London and as a crypto-economist in Berlin. She has an MA in Economics and Mathematics from the University of Edinburgh.

ABOUT THE CENTER FOR DATA INNOVATION

The Center for Data Innovation is the leading global think tank studying the intersection of data, technology, and public policy. With staff in Washington, D.C. and Brussels, the center formulates and promotes pragmatic public policies designed to maximize the benefits of data-driven innovation in the public and private sectors. It educates policymakers and the public about the opportunities and challenges associated with data, as well as technology trends such as predictive analytics, open data, cloud computing, and the Internet of Things. The center is a nonprofit, nonpartisan research institute proudly affiliated with the Information Technology and Innovation Foundation.

contact: info@datainnovation.org

datainnovation.org