U.S. Leadership in High Performance Computing (HPC)

A Report from the NSA-DOE Technical Meeting on High Performance Computing

December 1, 2016

1. Executive Summary

In June of 2016, China announced benchmark results for its new TaihuLight system that were nearly triple the previous record-setting numbers of another Chinese supercomputer, Tianhe-2. These results were achieved using a Chinese designed and fabricated 28 nm System-on-Chip (SOC) architecture. This announcment was coupled with impressive advancements in indigenously developed system software and high performance computing (HPC) applications, as shown by having three Chinese research teams as finalists for the 2016 Gordon Bell prize. These results indicate that China has attained a near-peer status with the U.S. in HPC. The U.S. asserted its intention to maintain a leadership position in HPC in the July 2015 Executive Order establishing the National Strategic Computing Initiative (NSCI). It is now clear that future U.S. leadership will be challenged by the Chinese. The *2012 Net Assessment of Foreign HPC* noted the aggressive development of Chinese HPC capabilities, and, in particular, the accelerated rate of investment that China was making in these areas. In September 2016, experts and leaders from the U.S. HPC community convened to update that assessment in light of the recent Chinese results. **Meeting participants expressed significant concern that – absent aggressive action by the U.S. – the U.S. will lose leadership and not control its own future in HPC.** There was broad support for maintaining HPC leadership, and a common belief in the importance of that leadership, as well as wide agreement on key steps to maintain this leadership position.

National security requires the best computing available, and loss of leadership in HPC will severely compromise our national security. HPC plays a vital role in the design, development or analysis of many – perhaps almost all – modern weapon systems and national security systems: e.g., nuclear weapons, cyber, ships, aircraft, encryption, missile defense, precision-strike capability, and hypersonics. Loss of leadership in HPC could significantly reduce the U.S. nuclear deterrence and the sophistication of our future weapons systems. Conversely, if China fields a weapons system with new capabilities based on superior HPC, and the U.S. cannot accurately estimate its true capabilities, there is a serious possibility of over- or under-estimating the threat. Either possibility leads to unwelcome situations such as distortions in the allocation of R&D resources and strategic planning for defense, uncertainty in national policy-making, and incorrect responses to world events. The U.S. enjoys relatively cost effective HPC due to our indigenous ecosystem and trusted access**. Loss of leadership could easily require the USG to acquire HPC capabilities in much the same way it acquires aircraft carriers, at vastly increased cost**.

HPC leadership has important economic benefits because of HPC's role as an enabling technology. A loss of leadership will be felt beyond the HPC vendor community, which would be significantly impaired: HPC resources are required for the development of a variety of military, scientific, and industrial capabilities. Loss of a U.S. leading position would threaten our ability to compete internationally in all of these fields. The effects of losing leadership can be expected to be long-lived; HPC helps drive the development of talent in math, science, and engineering. The economic and social benefits of the "Post Moore's Law" era may very well be concentrated where HPC leadership exists. For industrial applications, reliance on foreign HPC resources could threaten the loss of intellectual property and competitive edge. Personal email and private information, social networks, and the emerging Internet of Things are all subject to even greater privacy risks if offshore entities have superior HPC analytics or control the data / information markets.

Leadership positions, once lost, are expensive to regain. **To maintain U.S. leadership in HPC, a surge of USG investment and action is needed to address HPC priorities.** Many of these priorities have been outlined by the NSCI and some are clearly under way, in particular, DOE efforts in accelerating delivery of a capable exascale computing system (Objective #1), and some aspects of establishing "post-Moore's Law" computing (Objective #3), led by IARPA (Intelligence Advanced Research Projects Activity) and other agencies. Other NSCI objectives need more concrete plans and milestones to launch. The following recommendations are meant to guide these plans.

1. It is critical to lead the exploration and development of innovative computing architectures that will unleash the creativity of the HPC community. This recommendation directly supports Objectives #2, #3, and #4 of the NSCI. Meeting participants agreed that this work is essential and that it would be a significant part of a revitalized HPC community. The end result will be a broader, more diverse HPC ecosystem for the U.S. and the world.

2. Workforce development is a major concern in HPC and a priority for supporting NSCI Objectives #4 and #5. We must inspire a new generation of students to master the skills required for HPC, and we must develop "public-private" relationships between the USG and industry to insure that there are rewarding careers for people with these skills.

3. NSCI leadership must work to modernize export control practices to account for the new reality of Chinese technological capability and business practices, and develop more efficient contracting regulations to improve the public-private partnership in HPC science and technology development. This would also directly support NSCI Objectives #4 and #5.

2. Background on the 2016 HPC Technical Meeting

In 2012, an interagency group of high performance computing (HPC) experts and science and technology analysts, plus national security and defense leaders, reviewed the state of HPC and technical supercomputing in the U.S. and the world. That meeting informed the *2012 Net Assessment of Foreign HPC*, which noted the development of Chinese HPC capabilities, and, in particular, the increased rate of investment by China. Some of the recommendations of this report were incorporated into the National Strategic Computing Initiative (NSCI).

The recent world-record performance announcement and the technical details of the Sunway TaihuLight supercomputer indicate that China is rapidly transitioning from a position of "aspiring toward HPC leadership" to one of "peer" with the U.S. Accordingly, senior leaders within NSA and DOE agreed that assessing the current situation and its implications for the future of HPC in the U.S. was prudent. A meeting to facilitate this discussion was held on September 28-29, 2016, hosted by the University of Maryland, Baltimore County (UMBC) at one of its research park facilities.

Approximately 60 attendees, representing all aspects of the HPC community, participated in this meeting:

- 40 representatives from USG agencies
  - Senior leaders from DOE National Nuclear Security Administration (NNSA), DOE Office of Science, and NSA
  - Technical experts and leadership from 7 DOE National Laboratories
  - SMEs (Subject Matter Experts) in HPC operational offices, HPC systems, and HPC research from the National Security Agency
  - High level attendees from IARPA, NSF, and other agencies
- 10 representatives from industry, representing HPC vendors, HPC technology developers, and HPC users
- 10 SMEs from academia and other organizations with strong backgrounds in HPC

The group was specifically asked to address the following questions:

- Has the state of HPC leadership changed since 2012? How?
- What does this mean for U.S. leadership in HPC, which is recognized as a key component of national and economic security?
- What should we do?

The discussions and recommendations by the meeting participants are synthesized in this report.

The report appendices include the following information:

- Appendix A provides a brief overview of the Sunway TaihuLight system
- Appendix B presents the full text of the specific charge given to the meeting participants
- Appendix C describes the agenda for the two-day meeting
- Appendix D is the list of background information provided to participants, including both required and optional reading material
- Appendix E has overview information on the NSCI
- Appendix F is a table of acronyms

3. Has the State of HPC Leadership Changed Since 2012? How?

TaihuLight as a Marker of China's Progress in HPC

In the spring of 2016, China's National Supercomputing Center in Wuxi stood up the Sunway TaihuLight system (system details and applications can be found in Appendix A). This system demonstrates that China's indigenous capabilities in HPC are quite advanced and have increased impressively since 2012. The TaihuLight system represents both a serious, innovative "capability" machine and is evidence of the sea change forecasted in the *2012 Net Assessment of Foreign HPC*.

Meeting participants were very impressed with the TaihuLight system:

- It is homegrown. The architecture includes specialized processors that were designed and fabricated in China. TaihuLight is under-resourced from the viewpoint of a traditional, general-purpose HPC system in terms of memory capability (bandwidth and capacity). Nevertheless, it appears to be effective for many classes of use given certain programming approaches. The performance is attracting the interest of independent software vendors (ISVs), both in China and the U.S. The system was brought online six months ahead of schedule. Overall, TaihuLight represents a major step for China towards an "independent and controllable" HPC technology base, where the entire stack of software and hardware is locally controlled.
- It is innovative. The chip design includes architectural innovations, incorporating both heterogeneous processors and system-on-chip (SOC) features. The system software was developed using a true co-design approach, with the applications tuned to take maximum advantage of unique architectural features in the design. Despite the architectural specialization, the system provides a relatively familiar development environment. This is clearly evident in the compiler design that provides some abstraction from system-specific implementation details.
- It is not a stunt. TaihuLight is a significant step up in performance for China; indeed, its 93 petaflop/s is significantly greater than the aggregate flops available to DOE today – Titan, Sequoia, Mira, Trinity (Haswell), etc. More importantly, where previous Chinese HPC systems were unimpressive except for running benchmarks (i.e., LINPACK tests), TaihuLight is being used for cutting-edge research, with real-world applications running well at scale. This year, three of six finalists for the Gordon Bell competition (see Appendix A) are Chinese efforts; China has never been a finalist before.

Chinese Goals and Motivation in HPC

Meeting participants, especially those from industry, noted that it can be easy for Americans to draw the wrong conclusions about what HPC investments by China mean – without considering China's motivations. These participants stressed that their personal interactions with Chinese researchers and at supercomputing centers showed a mindset where computing is first and foremost a strategic capability for improving the country: for pulling a billion people out of poverty; for supporting companies that are looking to build better products, or bridges, or rail networks; for transitioning away from a role as a low-cost manufacturer for the world; for enabling the economy to move from "Made in China" to "Made by China." Having said that, their focus on using HPC systems and codes to build more advanced nuclear reactors and jet engines suggests an aggressive plan to achieve leadership in high-tech manufacturing, which would undermine profitable parts of the U.S. economy. And such codes, together with their scientific endeavors, are good proxies for the tools needed to design many different weapons systems.

Outlook for HPC Leadership

The *2012 Net Assessment of Foreign HPC* noted that China's investments in HPC created a trajectory where surpassing the U.S. in leadership was quite possible. The capabilities demonstrated by the TaihuLight system, along with other HPC advancements, indicate that China is now a near peer to the U.S. Assuming status quo conditions, the meeting participants believe that a change in HPC leadership was extremely likely, with only minor disagreement on the timescale; many suggested that China would be leading the U.S. as early as 2020.

China has stated (most recently in their 13th 5-year plan) that indigenous HPC capability is a strategic goal. The country's progress since 2012, as demonstrated by the TaihuLight system, indicates it is on a trajectory to meet that goal. When it comes to HPC, China controls its future.

Meeting participants expressed significant concern that – absent aggressive action by the U.S. – the U.S. will not control its own future in HPC.

4. What Does This Mean for U.S. Leadership in HPC?

The meeting identified several potential implications of the scenario where the U.S. does not assert control of its HPC future and instead becomes heavily reliant on capabilities developed by China. Participants were not in a position to quantitatively assess the likelihood of these outcomes, but did comment on which seemed most relevant.

Implications: National Security

Modeling and simulation using HPC play a vital role in the design, development, and analysis of many – perhaps almost all – modern weapons systems and national security systems: e.g., nuclear weapons, cyber, ships, aircraft, encryption, missile defense, precision-strike capability, and hypersonics. For these and many other systems, the last 20 years have provided ample evidence that additional compute capacity is rapidly and productively used for greater fidelity and resultant system capability. Simply put, leading-edge HPC is now instrumental to getting a world-class, large-scale engineering system out the door. In this light, a scenario where the U.S. loses access to the most current HPC technology can be clearly tied to detrimental outcomes. What is the cost of a significant delay in the roll-out of a new generation of fighters / bombers / missiles / tanks / destroyers / submarines? What is the effect on our national defense if another country can dictate the timing with which the U.S. fields a strategic capability?

From the opposite perspective, if China fields a weapons system with new capabilities based on superior HPC, and the U.S. cannot accurately estimate its true capabilities, there is a serious possibility of over- or under-estimating the threat. Either possibility leads to unwelcome situations such as distortions in the allocation of R&D resources and strategic planning for defense, uncertainty in national policy-making, and incorrect responses to world events.

Loss of leadership in HPC means loss of a trusted supply chain, or increased cost from building a supply chain that does not spread costs across commercial and government users.

The shipbuilding industry was noted by participants as an informative analog. At one point, the U.S. had a leadership position in shipbuilding, but the entire commercial sector has moved offshore (albeit not so much for loss of technical leadership as for unfavorable economics). Shipbuilding requires enormous capital costs and non-recurring engineering, and the U.S. Navy now has to support a dedicated collection of suppliers who cannot spread those costs across both military and commercial production. Similarly, the capital costs of a modern fab are enormous. The cost model challenges resulting from dedicating a cutting-edge fab to government customers are difficult to quantify, but the magnitude of the issue can clearly be seen in the experience of trusted fabs in the last 20 years. In short, maintaining a USG-dedicated computing supply chain could drive up costs precipitously, but it might be the only sound option. We do not expect the Air Force to fly commercial off-the-shelf (COTS) aircraft, and the country has chosen to build its own warships, so why would we expose similarly important computing capabilities to foreign control?

Choosing to accept an insecure supply chain avoids astronomical costs but carries obvious downsides. How much information can be gleaned about USG priorities / capabilities / strategy from what key agencies purchase (in terms of configuration or quantity)? What is the information security and trust risk for operational systems, either in terms of access or information leakage? What infrastructure systems (e.g., energy grids) are exposed to these risks?

Meeting participants concluded that national security requires the best computing available, and loss of leadership in HPC will severely compromise our national security.

Implications: U.S. HPC Industry Disrupted

HPC systems offer astonishing performance but do so with sustained funding commitments and for some very specific needs. The need for sophisticated users and a complex programming environment limit adoption to only a small set of customers where the value proposition is so strong that very high costs can be justified. This model tends to create a monoculture – architectures, databases, file systems, application libraries, and even open-source software that lack diversity and portability. With a relatively small market and limited growth for HPC, the U.S. vendor community struggles to make the case for increased private investment in the current model; self-funding, an alternative approach with lower margins, is not seen as feasible when it supports such a small set of users. Thus, the U.S. vendor community is effectively locked into the current high-cost model and is not well-positioned to respond to a disruptive alternative. Much of the community is in agreement that current customers for HPC represent a small fraction of a bigger potential customer base. Any significant change that creates and captures this larger potential customer base would be highly disruptive; this represents both an opportunity and a threat to the U.S. HPC industry.

Many at the meeting believe that emerging HPC innovations in China (as exemplified by TaihuLight) may represent the start of a disruptive threat. China is willing to tolerate the financial costs and technical risks associated with investing in a variety of novel approaches (for example, they have three competing HPC architectural teams at their national universities and defense labs). Using a lightweight kernel and a lightweight core and letting go of compatibility with legacy codes is in keeping with the general approach of China's HPC program. Extrapolating further, it could end up being a structurally lower-cost option than that which is provided by U.S. industry. This makes for a worrisome asymmetry and suggests that if the disruption plays out, U.S. providers would have a severely diminished role.

HPC leadership can provide a "first mover" advantage; the technologies and skills needed to design, develop, and deploy leadership class systems often lead requirements for other computing systems by many years. Understanding and driving innovations at the leading edge can enable a valuable learning curve for the leaders; reacting to these innovations can be expensive and can lead to loss of markets by the incumbents to the innovators.

Implications: Economic Security Weakened

Loss of U.S. leadership would have clear economic impact because of HPC's role as an enabling technology. Indeed, some participants posited that a loss in HPC could be "the Innovator's Dilemma" [1] on an international scale. HPC is a tool to amplify human intelligence, to allow one to think about, and solve, bigger problems, much faster. A loss in HPC leadership potentially threatens U.S. companies in industries that are already heavy users of HPC: automotive, aerospace, advanced manufacturing, oil discovery, pharmaceutical research, finance, etc. These sectors tend to already be hyper competitive globally, and thus a modest disadvantage to U.S. competitiveness (e.g., if China used HPC leadership to provide compute capability to domestic industry at reduced, or even no cost) could translate into much larger economic effects. There are also significant intellectual property risks: If a pharma firm's drug discovery is performed overseas, to what extent is IP leakage an unseen drag on U.S. competitiveness?

A related theme is the potential for reduced access to the most critical technologies. Outright denial (embargo restrictions) is only one avenue, and a heavy-handed one. Delaying access to the latest systems or software libraries, or "slow rolling" their availability by causing unpredictable delays, or other more nuanced tactics, can be used to increase risk and cost to U.S. businesses.

Implications: Loss of Opportunity

The skills needed to effectively employ HPC systems (math, computer science, engineering) provide a knowledge base for society that can also be used in a wide variety of other important ways. People with those skills, or people who wish to learn those skills, tend to gravitate towards research centers with HPC leadership (whether government, industry or academia). The U.S. currently takes advantage of its leadership position in HPC to attract significant amounts of U.S. and foreign talent; this situation could change quickly.

HPC also has an important role in several of the most intriguing, and likely to be lucrative, emerging industries and opportunities. Meeting participants noted potential business opportunities such as autonomous vehicles, international logistics, energy storage, space travel, new digital business capabilities (e.g., distributed contracts, digital currency), adapting to climate change, and biosciences.

Down the road, investment and leadership in HPC will likely determine the region where the economic and social benefits of the "Post Moore's Law future" are concentrated.

[1] Christensen, Clayton M., The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail. Boston, MA: Harvard Business School Press, 1997.

Additional Implications

Participants identified other possibilities that had some relationship to the loss of HPC leadership. During the meeting, attendees did not reach a clear consensus on the importance of losing U.S. leadership in HPC to the likelihood of these scenarios.

In many ways, control of computing equals control of information. What if the personal email and social networks of a sizable portion of U.S. citizens are hosted overseas? Looking forward, many new types of information will exist online. The emerging Internet of Things has the potential for providing a variety of innovations that could increase quality of life, but, in doing so, exponentially increase the amount of sensitive digital information: medical conditions from wearable diagnostic devices, audio from always-listening artificial intelligence (AI) assistants, activity information from an array of connected sensors in homes and in businesses. The use of HPC by foreign entities to analyze the data acquired by these systems is a potential threat to individual and societal privacy.

Internet companies (e.g., Google, Facebook, and Amazon) have emerged as extraordinarily heavy users of computing cycles, and thus there are potential competitiveness consequences for companies in this sector. If Baidu, Alibaba, or Tencent can operate infrastructure for 30% less than a U.S. company and devote those resources to R&D, how much of a disadvantage does this create for companies such as Google, Facebook, and Amazon?

5. What Should We Do?

The opening statement in the introduction for the NSCI Strategic Plan states, "HPC is essential to the Nation's global economic competitiveness and scientific discovery." There was a broad consensus among the meeting participants in both the importance of the NSCI itself, as well as the specific objectives upon which it focuses (see Appendix E for a listing of NSCI Guiding Principles and Objectives). Some of the objectives of the NSCI are clearly under way, in particular, the DOE Exascale Computing Program (ECP) for developing an effective exascale computing capability in the next decade, as well as some aspects of the post-Moore's Law challenge, led by IARPA and other agencies. Meeting participants made the following recommendations that either amplify or elucidate specific parts of the Initiative, or are suggestions for additional options that should be pursued.

Provide an Investment Surge to Ensure U.S. Leadership in HPC

The *2012 Net Assessment of Foreign HPC* noted a divergence in R&D investment between the U.S. (slowing) and China (accelerating). Objectives #1, #2, and #3 of the NSCI can be seen as identifying the USG investments necessary to support a healthy HPC ecosystem for the next 30+ years. A notional timeline for the impact of these investments is the following:

- Today to 2025 – HPC ecosystem nurtured by USG investments to reach a capable Exascale system
- 2025 to 2035 – HPC ecosystem takes advantage of USG leadership in architectural innovations (described below)
- 2035 and beyond – HPC ecosystem endures because of USG investments in "Post Moore's Law" era

If these investments are not made, the U.S. can expect an HPC capability gap to emerge and widen in less than a decade. Given the long lead time (a decade or more) for research in HPC to have a noticeable impact on the HPC ecosystem, the time to make these investments is now. The slowing of USG investment in recent years, the emergence of China as a serious competitor to HPC leadership, and the transition from the dominance of the traditional von Neumann architecture and semiconductor technology have combined to create a new, unstable environment for HPC leadership. To maintain U.S. leadership in HPC, a surge of USG investment is needed. This surge would be required for approximately a decade. After the surge, investments in HPC R&D would need to remain at a level more typical of historical norms and clearly reverse recent trends. Absent these investments, a loss of HPC leadership by the U.S. appears nearly inevitable. Once lost, HPC leadership would be extremely expensive to regain.

As China's economy modernizes, its standard of living improves, and salaries rise, and certain lower cost advantages will abate. Purchasing parity for R&D expenditure is difficult to forecast or even precisely assess, but industry representatives at the session noted that the gap was large but shrinking quickly, and estimated 5-10 years to near-parity.[2] The Department of Commerce calculates that, for manufacturing labor, the wage differential is on track to close from 20x in 2008 to 4x in 2020. A study by the Boston Consulting Group found evidence of significant growth in reshoring.[3] Although manufacturing costs are not a perfect proxy for R&D, and additional research on the topic may be warranted, the trend line seems clear: what was only recently an insurmountable gap is heading towards something that can be overcome with innovation advantages. This indicates that a temporary surge, approximately 10 years in duration, but well above historical investments in HPC, is the appropriate action. This will need to be followed by investment levels clearly above the recent trend and more typical of historical norms for HPC R&D.

Lead the Generally-Specialized Computing Revolution

Making a serious investment in architectural exploration was by far the most common recommendation that came from the working group deliberations. This recommendation clearly supports Objectives #2, #3, and #4 of the NSCI and requires an investment commensurate with the challenges posed by these objectives. We are at an inflection point in the development of computer architectures. Most of the history of computing has focused on advances in general-purpose architectures that were not targeted at a specific problem, but could address a broad spectrum of scientific and data science problems. As the rate of progress has departed from the exponential scaling described by Moore's law, specialized architectures have

[2] Department of Commerce, Labor Costs, website, accessed Oct 11, 2016.

[3] 17.5% of respondents in 2015 indicated they were actively reshoring production, 2.5x the number from a similar survey in 2012. See a brief article here and a 2012 report on "Made in America, Again," here.

emerged as one of the most attractive paths to achieving further performance gains without shifting to a radically different computing model (i.e., to promising, but not-yet-generally-realizable approaches such as quantum and neuromorphic computing). Examples are the Anton architecture that targets molecular dynamics, the GRAPE architecture for gravity / astronomic simulation, and the Google Tensorflow Processing Unit for machine learning applications. The use of graphical processing units (GPUs) in HPC systems could be considered an initial exploration into these "generally-specialized" architectures. In many ways, this trend extends the co-design efforts at the DOE and elsewhere, in which algorithms, application software, architectures, system software, and hardware technologies are iteratively developed and combined to create highly capable computing systems. We expect that future computing architectures will continue this trend of having a combination of general and specialized processors.

To maintain and extend U.S. leadership in HPC, it is critical to lead the exploration and development of innovative computing architectures that will unleash the creativity of the HPC community. **The lessons of the last several decades of disruptive innovation theory suggest that the only way to fight a disruptive opponent is to become disruptive oneself. This is why participants recommend the U.S. engage in innovative architectural exploration and development, tied to another recommendation for making public-private partnerships significantly easier through changes in both policy and law. We need to overcome our adversaries with even more innovation and disruption. This will require a strong and sustained commitment to a variety of architectural approaches, where each is clearly focused on a reasonable subset of clearly important applications** (e.g., modeling and simulation, large-scale analytics, cybersecurity, etc.). Multiple teams, each capable of developing novel computing architectures using a complete co-design approach, will be required. This will ensure a variety of explorations, increasing the opportunity for major innovations. Fully functional system prototypes from these explorations will need to be developed and evaluated for their specific application subset. Technology advancements common to many (or all) architectures will also need to be supported (i.e., new memories, silicon photonics, programming tools and languages, etc.). There will be technical risks; there will be some explorations that turn out not to be viable. Meeting participants agreed that this work is essential and that it would be a significant part of a revitalized HPC community. The end result will be a broader, more diverse HPC ecosystem for the U.S. and the world. Leadership from the NSCI will be needed in formulating an architectural innovation strategy.

The development of toolchains for the design and validation of large custom circuits required by such architectures will be a critical enabling technology for reducing the non-recurring engineering expenses of these approaches. Workshop participants were enthusiastic about the potential for innovations here. Currently, the process requires a multitude of specialized and expensive skillsets as well as enormous time and effort. Some participants envisioned an end state where lowering barriers to generally-specialized systems unleashes U.S. competitiveness in a way analogous to what has occurred in product design via 3D printers and where better tools allow vastly smaller teams to compete very effectively on the strength of innovation and design. The culture in the U.S. has fostered creativity in ways that will be difficult to replicate overseas. Thus, work in this area could help emphasize a very sustainable advantage for the U.S.

Innovations in architecture, especially if they lead to diversity, will increase the need for software portability. The DOE, in particular, has emphasized the use of open-source software in future systems; this is intended to support portability needs and reduce reliance on proprietary solutions. Meeting participants felt that an open-source software strategy, perhaps initiated by a DOE-led workshop, is an important component of architectural leadership.

Improve HPC Education and Workforce Development

Workforce development was identified as a major concern in HPC, and a priority. The U.S. needs a modern, well-trained, diverse workforce to continue the innovations that will enable the U.S. to maintain its leadership position in HPC. There are many challenges in doing so, and the meeting participants suggested conducting a workshop that assembles HPC, education, and policy experts to more broadly explore how to proceed. Issues such as K-12 engagement, access to HPC, and web-based training for HPC would be the type of topics addressed at the workshop. This is an area that requires modest investments relative to architectural innovation; coordination with the broader STEM community is the key.

HPC interests, technologies, and priorities often are lost in the face of modern computer science curricula. In part, this is due to the explosion of data-centric jobs and technologies in companies such as Google and Facebook, and, in part, because of the difficulty of accessing HPC resources. One current solution to this problem is the DOE Computational Science Graduate Fellowship (CSGF) program, which funds graduate students in a program that has requirements for both a curriculum that specifies HPC programming skills and technologies, as well as practicums at the DOE National Laboratories to introduce these students to laboratory scientists and application areas. This program should be used as a template for further development of an HPC workforce.

The U.S. still educates a large portion of the world's best students in computer and computational science, many of whom are foreign national citizens. We recommend consideration of programs that incentivize these students to stay in the U.S. after graduation so that they contribute to the development of U.S. HPC capabilities rather than returning to their own countries; these programs would take the form of Visa or Green Card programs that target Ph.D. graduates in HPC.

Adopting the recommendation for a major investment in architectural innovation, as described above, would also energize students, faculty, and the general workforce to develop stronger skills in HPC.

Improve Public-Private Partnerships

A major U.S. strategic advantage is the strength and diversity of its HPC industry, and public-private partnerships between the Government and the vendor community will be critical toward advancing HPC for further scientific discovery, economic competitiveness, and national security (NSCI Objective #5). A key role for the U.S. government is to make strategic, long-term investments and avoid *The Innovator's Dilemma* that would otherwise limit the exploration and development of new architectures. However, partnerships with industry have become onerous in recent years, primarily due to (i) regulations that classify HPC R&D as IT procurements that hugely complicate coordination of research activities, and (ii) export control requirements on HPC components that hurt international competitiveness of U.S. technologies and do little to retard the development of foreign competition. Multiple participants provided anecdotal evidence that export control decisions from the 1980s still affect working relationships today. It is critical that we repair these relationships and update policies so that the U.S. may leverage its industrial strength. Given the complexity of U.S. procurement and trade laws, understanding the proper steps to take will likely require a meeting that convenes people with appropriate expertise and interests. Meeting participants strongly recommend that the NSCI leadership tackle the challenges of modernizing export control processes to account for the new reality of Chinese technological capability and business practices, and update contracting regulations to improve HPC science and technology development. They recommend the leadership explore obtaining a change in ITAR (International Traffic in Arms Regulations) to not classify USG supported HPC innovations as restricted by ITAR. This is an area that will require focus, energy, and perseverance, not funding.

Participants also suggested stronger engagement with industry in the development of HPC grand challenges and in having industry participation in the review of research projects funded by the USG. The OSTP Nanotechnology-Inspired Grand Challenge for Future Computing[4], is an example of constructive industry engagement.

Summary

Since the earliest days of electronic computing, the U.S. has taken advantage of a "virtuous cycle" in HPC. Significant USG investments in people, applications, systems, and technology created a comprehensive HPC ecosystem. The USG was able to partner with a wide variety of entities in this robust ecosystem to develop truly exceptional HPC capabilities at relatively low cost. This ecosystem then created significant economic benefits for the HPC industry and the broader computing market; U.S. society has also clearly benefited from the computing revolution of the past half century. The emergence of China's indigenous HPC capability presents a potentially disruptive threat to this virtuous cycle, with serious implications for national and economic security in the U.S. Meeting participants have developed a set of specific recommendations that should be followed to enable the U.S. to remain in control of their HPC future and continue this virtuous cycle. These recommendations are meant to amplify or elucidate specific parts of the NSCI or are suggestions for additional options that should be pursued.

[4] Whitehouse.gov, issued October 2015, see the Nanotechnology-Inspired Grand Challenge for Future Computing website, accessed Oct 24, 2016.

Appendix A: Summary of the Sunway TaihuLight System and the Submissions for the Gordon Bell Prize

The Sunway TaihuLight System was developed by the National Research Center of Parallel Computer Engineering and Technology and installed at the National Supercomputing Center in Wuxi, a public supercomputing center. The complete system has a theoretical peak performance of 125.4 Pflop/s with 10,649,600 cores, 1.31 PB of primary memory, and 20 PB of storage. It is based on the SW26010 processor developed by the Shanghai High Performance IC Design Center, designed and built in China using 28nm fabrication technology. It features the Shenwei-64 instruction set, a RISC (Reduced Instruction Set Computing) architecture that was also developed indigenously. This is the first Chinese supercomputer based upon an indigenous design and using indigenous manufacturing.

The building block of the machine is a Core Group that is comprised of a cluster of 64 Computing Processing Elements (CPEs) arranged in an 8x8 grid, along with a Management Processing Element (MPE), and a Memory Controller (MC). The MPE consists of a 64-bit RISC core that supports user and system modes, 256-bit vector instructions, 32 KB L1 instruction and data caches, and 256KB L2 cache. The CPEs have a similar architecture, but support only user mode, with 256-bit vector instructions, 16 KB L1 instruction cache, and 64 KB Scratch Pad Memory.

A node of the TaihuLight system consists of 4 MPE/CPE clusters with a Network on Chip (NoC) interface connected to the System Interface (SI), with 8 GB of DDR3 memory. Each node has a peak performance of 3.06 Tflop/s. The SI is a standard PCIe interface with a bidirectional bandwidth of 16 GB/s with a latency of 1 us.

The TaihuLight system is relatively light on memory capacity and bandwidth. The management processing elements (heavyweight cores) have 32 kB of Level 1 instruction and data caches, and a 256kB L2 cache. The computing processing elements (lightweight cores) have 16kB L1 caches and a 64kB scratchpad memory. Each chip has only 32 GB of external (or main) memory. Memory bandwidth is 136.5 GBps, with a network interface of 16 GBps; the system bisection bandwidth is 70 TBps. The ratio of memory and bandwidth to Flops are low compared to other leading-edge processor designs and systems.

The TaihuLight system uses Sunway Raise OS 2.0.5, based on Linux. The software stack includes basic compiler components such as C/C++ and Fortran compilers, an automatic vectorization tool, and basic math libraries. The system also includes Sunway OpenACC, a customized parallel compilation tool that extends OpenACC to unique characteristics of the SW26010 processor.

In June, 2016, the TaihuLight reported a High Performance Linpack (HPL) benchmark result of 93 Pflop/s out of a theoretical peak of 125 Pflop/s, for an efficiency of 74.15%. This result is ranked in the TOP500 list as the fastest supercomputer in the world and nearly triples the previous best result of the Tianhe-2 computer of 34 Pflop/s. The power consumption average during the benchmark was 15.371 MW, or 6 Gflop/W.

In contrast, the benchmark results for the High Performance Conjugate Gradient (HPCG) benchmark reached only 0.3% of peak performance, which shows weaknesses of the architecture with slow memory and modest interconnect performance.

However, despite the slow memory, the system has notably implemented register-level communications, that, for problems that can make use of it, such as problems on structured meshes, avoid the memory bottleneck within a Core Group.

Nonetheless, there are sizeable applications already fielded on the machine, including three finalists for the 2016 Gordon Bell Prize, demonstrating that the system is capable of running real applications and is not merely a stunt machine. The Gordon Bell Prize is presented by the Association for Computer Machinery each year in conjunction with the Supercomputing Conference for outstanding achievement in high-performance computing applications and is intended for achievements that demonstrate algorithmic innovation and improvements over the state of the art in scalability, time to solution, efficiency with respect to bottleneck resources, and peak performance. The technical submissions from China for the Gordon Bell prize are the following:

- A fully implicit nonhydrostatic dynamic solver for cloud-resolving atmospheric simulations. They demonstrated scaling that utilized 80% of the machine and an application performance of 1.5 PFlops (1.2% of peak).

- A global surface wave numerical simulation with ultra-high resolution. They demonstrated scaling to 80% of the machine size and 30.7 PFlops (25% of peak).
- A large-scale phase-field simulation of coarsening dynamics based on the Cahn-Hilliard equation with degenerated mobility. 85% scaling and 39.7 PFlops (32% of peak).

References

- ORNL / Univ of Tennessee: **Report on the Sunway TaihuLight System**, J. Dongarra, Tech Report UT-EECS-16-742, (June 2016), full text here.
- State Key Lab of Mathematical Engineering and Advanced Computing (China): **Cooperative computing techniques for a deeply fused and heterogeneous many-core processor architecture**, F. Zheng et al., *Journal of Computer Science and Technology,* 30(1):145-162 (2015), abstract here; full text here.

Appendix B: Charge to Attendees of HPC Technical Exchange

In 2012, an interagency group of HPC experts and science and technology analysts, plus national security and defense leaders, reviewed the state of HPC and technical supercomputing in the U.S. and the world. Some of the recommendations of the subsequent report were incorporated into the National Strategic Computing Initiative. China's desire to achieve HPC leadership was noted in the report, where HPC leadership can be defined in terms of people, applications, systems, and technology. The recent announcement and technical details of the TaihuLight supercomputer indicates China is rapidly transitioning from aspirations of HPC leadership to peer status with the U.S. in HPC. This meeting, bringing together HPC experts across a wide spectrum of interests, will enable the USG to assess the current and future HPC landscape.

We expect this group to answer the following questions:

- Has the state of HPC leadership changed since 2012? How?
- What does this mean for U.S. leadership in HPC, which is recognized as a key component of national and economic security?
- What should we do?

The meeting report, to be finished by October 31, 2016, will be delivered to senior leaders within DOE, NSA, and OSTP, as well as other senior USG officials. This report will provide an assessment of China's HPC capabilities and potential impact on economic and national security. It will be used by these senior leaders to define priorities and resources necessary for continued HPC leadership by the U.S.

Appendix C: HPC Technical Exchange Meeting Agenda

The agenda was structured to maximize discussions and deliberations among the participants. The first half day was devoted to presentations by technical and HPC experts; the remaining one and a half days were devoted to small group discussions. These discussions were focused on specific topical questions, had co-facilitators to help guide the conversation, and were introduced using contextual information for helpful guidance of the discussions.

Wednesday, September 28, 2016

0800-0900	Registration and networking
0900-0910	Welcome and charge to the meeting participants
0910-0920	Logistics
0920-0930	Overview of the agenda
0930-0945	TaihuLight capabilities - System Software
0945-1000	TaihuLight capabilities - Application development
1000-1015	TaihuLight capabilities – Hardware
1015-1030	Break
1030-1100	China HPC trends - ATIP and IDC
1100-1130	DOE HPC plans
1130-1200	Additional HPC leadership info or Q&A
1200-1300	Lunch and poster presentations
1300-1345	Topic 1 intro. Topic 1: Current/future HPC status and concerns for loss of U.S. leadership (scenario development)
1345-1600	Topic 1 discussion group breakouts (including a break if desired)
1600-1630	Reconvene in full group for Day 1 wrap-up and quick agenda check
1630-1730	Executive session (limited attendance). Networking or continue small group discussions for others

Thursday, September 29, 2016

0830-0900	Networking
0900-0915	Welcome
0915-1045	Topic 1 small group discussion outbriefs (Red, White, Blue)
1045-1100	Break
1100-1130	Topic 1 small group discussion outbrief (Purple)
1130-1200	Topic 2 intro. Topic 2: What is needed for enduring HPC leadership? (Identify NSCI alignments where appropriate)
1200-1300	Lunch and poster presentations
1300-1430	Topic 2 discussion group breakouts
1430-1500	Break
1500-1700	Topic 2 small group discussion outbriefs (Purple, Blue, White, Red)
1700-1730	Wrap-up and open Q&A
1730-1830	Executive session (limited attendance). Networking or continue small group discussions for others

Appendix D: Additional Background Information

The following items were distributed to workshop participants as background reading. Items not linked are available upon request.

Required reading for participants:

- 2012 Net Assessment of Foreign HPC, no link available.
- NSCI: **National Strategic Computing Initiative Strategic Plan**, prepared by the NSCI Council, John Holdren, Shaun Donovan, co-chairs, (July 2016), full text here.
- ITIF: **The Vital Importance of High-Performance Computing to U.S. Competitiveness**, Stephen J. Ezell and Robert D. Atkinson, (April 2016), pp. 1-58, access report here. *See, in particular, policy recommendations on pp. 44-48.*
- ORNL / Univ of Tennessee: **Report on the Sunway TaihuLight System**, J. Dongarra, Tech Report UT-EECS-16-742, (June 2016), full text here.

Optional reading for participants:

- Council on Competitiveness: **Solve.** The Exascale Effect: The Benefits of Supercomputing Investment for U.S. Industry, (October 2014), pp. 1-76, report here.
- Nat'l Nanotechnology Initiative / OSTP: **A Federal Vision for Future Computing: A Nanotechnology-Inspired Grand Challenge**, collaborating agencies: DoE, NSF, DoD, NIST, IC, (July 2016), access whitepaper here.
- Sandia National Labs: **SOC for HPC: An Agile Approach to Building HPC Systems from Commodity Components**, John Shalf and Jim Ang, (July 2016), report here.
- DoD HPC Modernization Program: **HPCMP Technology Outlook Architectural Trends**, presentation by Roy Campbell, (August 2016), no link available.
- Sandia National Labs: **Early Evaluation of the Sunway Architecture**, SAND2016-9681E, presentation by Rob Hoekstra, no link available.
- State Key Lab of Mathematical Engineering and Advanced Computing (China): **Cooperative computing techniques for a deeply fused and heterogeneous many-core processor architecture**, F. Zheng et al., *Journal of Computer Science and Technology,* 30(1):145-162 (2015), abstract here; full text here.
- Tsinghua Univ, Nat'l Supercomputing Center in Wuxi, Chinese Acad of Sci, First Inst. of Oceanography (all of PR China): **The Sunway TaihuLight Supercomputer: System and applications**, H. Fu, et al. *Science China Information Sciences*, 59(7):072001 (2016), access full text here.

Appendix E: NSCI Excerpts

The 2016 Strategic Plan for the National Strategic Computing Initiative and the text of Executive Order 13702 is available here. Two key elements – the guiding principles and the objectives – are excerpted below.

NSCI Guiding Principles

1) The United States must deploy and apply new HPC technologies broadly for economic competitiveness and scientific discovery.
2) The United States must foster public-private collaboration, relying on the respective strengths of government, industry, and academia to maximize the benefits of HPC.
3) The United States must adopt a whole-of-government approach that draws upon the strengths of and seeks cooperation among all executive departments and agencies with significant expertise or equities in HPC while also collaborating with industry and academia.
4) The United States must develop a comprehensive technical and scientific approach to transition HPC research on hardware, system software, development tools, and applications efficiently into development and, ultimately, operations.

NSCI Objectives

1) Accelerating delivery of a capable exascale computing system that integrates hardware and software capability to deliver approximately 100 times the performance of current 10 petaflop systems across a range of applications representing government needs.
2) Increasing coherence between the technology base used for modeling and simulation and that used for data analytic computing.
3) Establishing, over the next 15 years, a viable path forward for future HPC systems even after the limits of current semiconductor technology are reached (the "post- Moore's Law era").
4) Increasing the capacity and capability of an enduring national HPC ecosystem by employing a holistic approach that addresses relevant factors such as networking technology, workflow, downward scaling, foundational algorithms and software, accessibility, and workforce development.
5) Developing an enduring public-private collaboration to ensure that the benefits of the research and development advances are, to the greatest extent, shared between the United States Government and industrial and academic sectors.

Appendix F: Table of Acronyms

COTS	Commercial Off-the-Shelf
CPE	Computing Processing Elements
CSGF	Computational Science Graduate Fellowship (DOE)
DOE	Department of Energy
ECP	Exascale Computing Project (DOE)
GPU	Graphical Processing Unit
GRAPE	Gamma-Ray detector Array with Position and Energy (Riken special-purpose computer)
HPC	High Performance Computing
HPL	High Performance Linpack
HPCG	High Performance Conjugate Gradient
IARPA	Intelligence Advanced Research Projects Activity
ISV	Independent Software Vendor
ITAR	International Traffic in Arms Regulations
ITIF	Information Technology & Innovation Foundation
MC	Memory Controller
MPE	Management Processing Element
NNSA	National Nuclear Security Administration
NOC	Network on Chip
NSA	National Security Agency
NSF	National Science Foundation
NSCI	National Strategic Computing Initiative
OSTP	Office of Science & Technology Policy (Executive office of the President)
R&D	Research and Development
RISC	Reduced Instruction Set Computing
SMEs	Subject Matter Experts
SOC	System-on-Chip
STEM	Science, Technology, Engineering and Mathematics (academic disciplines)
USG	United States Government