White House National Strategic Computing Initiative Workshop Proceedings

Summary of a Workshop

October 20-21, 2015
Hilton McLean Tysons Corner
McLean, VA

The *White House National Strategic Computing Initiative Workshop Proceedings* is a summary of the workshop held at the Hilton McLean Tysons Corner in McLean, Virginia on October 20-21, 2015. The Networking and Information Technology Research and Development (NITRD) program National Coordination Office (NCO) has edited and released this report in support of the White House National Strategic Computing Initiative (NSCI).

This workshop was made possible by the financial and organizational support of the U.S. Department of Energy on behalf of the NSCI Executive Council.

Table of Contents

Preface

On July 29, 2015, President Obama issued an Executive Order establishing the National Strategic Computing Initiative (NSCI) to ensure the United States continues leading in this field over the coming decades[1]. This document summarizes the *White House National Strategic Computing Initiative Workshop,* that was held, at the request of the NSCI Executive Council, on October 20 and 21, 2015, at the Hilton McLean Tysons Corner in McLean, Virginia. The goal of the workshop was to capture the full breadth of the participants' individual opinions and ideas on four NSCI themes:

- Convergence of data analytics and computationally-intensive computing
- Future computing technology (beyond Moore's Law)
- Improving productivity in HPC application development and deployment
- Workforce development

The two-day workshop brought together approximately 250 participants representing industry, government, academia, and other organizations. The workshop was the first public forum where stakeholders from different sectors could share opinions on the challenges and opportunities of increasing computing demands, the increasing role of big data, and the evolving technological landscape in high performance computing (HPC).

An organizing committee (Appendix A: Organizing Committee) was tasked to develop a workshop program that would foster discussion on ways the NSCI could address increasing computing demands in the era of big data and in light of a rapidly evolving technological landscape. The organizing committee was responsible for organizing the workshop, identifying the plenary speakers and panelists, and developing the breakout topics to stimulate discussion and exchange of ideas. While the workshop speakers and panelists were recommended by the organizing committee, workshop attendance was open to all interested parties.

During the workshop, topics were structured to follow the three NSCI themes, using a mix of keynote speeches, panel sessions, and breakout sessions. The breakout discussions were an essential element of the workshop, allowing the broad range of subject matter experts in attendance to expand and debate the workshop themes and the various concepts espoused by the panelists and keynote speakers. To avoid conflicts with the Federal Advisory Committee Act, facilitators sought presentation and discussion of individual opinions, and did not poll the group or ask for consensus on particular topics. Workshop attendees were randomly assigned to five breakout groups. The five groups were asked to discuss topics one and two on the first day and topics three and four on the second day of the workshop. More detail on the breakout topics and prompting questions can be found in Appendix C: Breakout Topics.

[1] https://www.whitehouse.gov/the-press-office/2015/07/29/executive-order-creating-national-strategic-computing-initiative

This workshop proceedings report is based on the workshop agenda (see Appendix B: Workshop Agenda) and includes summaries of the keynote presentations, panel discussions, and the breakout sessions. While all agenda items were covered during the workshop, the schedule was modified significantly on the second day to accommodate a multi-hour power outage. Presentations are included in Appendix D: Presentations; the subsequent sections supplement the slides with notes from the discussions in the plenary sessions and breakout sessions respectively.

Context and Discussion Topics

The NSCI vision is to establish public and private sector collaborations for developing and broadly deploying next generation HPC paradigms that will drive economic growth, enable scientific discovery, and foster innovation. Within that context, the Workshop Planning Committee developed an agenda that reflected four of the five overarching themes central to the NSCI[2].

- The convergence of data-intensive and numerically intensive computing;
- Hardware technology for future HPC systems; and
- Improving productivity in HPC application development and deployment
- Creating a trained HPC workforce.

The historic focus on HPC has been on computers designed to simulate models of physical systems. In the last 10 years, a new class of HPC system has emerged to collect, manage, and analyze vast quantities of data arising from diverse sources, such as Internet web pages and scientific instruments. By combining the computing power and the data capacity of these two classes of HPC systems, deeper insights can be gained through new approaches that combine simulation with actual data. This combination of capabilities will also enable data analytic methods that require large degrees of numeric processing. Achieving this combination will require finding a convergence between the hardware and software technology for these two classes of systems. The NSCI seeks to drive the convergence of compute-intensive and data-intensive systems, while also increasing overall performance. Workshop topics in the convergence of data-intensive and compute-intensive systems included:

- Example applications that can benefit from combining numerical modeling with data analytics;
- Appropriate types of computing, communication, and storage hardware for different applications;
- Appropriate programming models and software environments; and
- Dependence of hardware and software frameworks on real-time response requirements.

[2] https://www.whitehouse.gov/sites/default/files/microsites/ostp/nsci_fact_sheet.pdf

While the performance of computer systems has increased at a steady rate over the past 70 years largely through improvements in the underlying hardware technology, semiconductor technology is reaching the limits of Moore's Law scaling behavior. There are a number of candidate technologies that may one day succeed current CMOS technology--however, none is close to being ready for deployment. Workshop participants were asked to consider:

- The driving forces that must be addressed by future technology;
- The ultimate capability that can be achieved by continued improvements to CMOS-based technologies;
- Promising technologies for digital computing beyond the CMOS computing era;
- The prospects for radically new computing models, such as quantum- and neuromorphic-based computing; and
- The interplay between new hardware structures and system architectures, programming models, and software requirements.

Current HPC systems are very difficult to program, requiring careful measurement and tuning to obtain maximum performance on the targeted machine. Migrating a program to a new machine can require repeating much of this process, and it also requires validating the new code by careful comparison of results with the old system. The level of expertise and effort required today to develop HPC applications poses a major barrier to widespread adoption. Government agencies will support research on new approaches to building and programming HPC systems that make it possible to express programs at more abstract levels and then automatically map them onto specific machines.

Currently there are many companies and research projects that could benefit from HPC technology, but they lack expertise and access. Many scientists and engineers also lack training in the concepts and tools for modeling and simulation and data analytics.

Workshop participants were asked to address the following subjects:

- Approaches to eliminate the need for low-level tuning of application code to achieve required performance;
- Approaches that could be taken to increase software sharing and reuse;
- Deployment models that enable sustainable and widespread access to HPC resources; and
- Building and sustaining a workforce capable of creating HPC applications and of applying HPC to develop capabilities for economic competitiveness and scientific discovery.

October 20 (Day 1)

Plenary Session

The workshop opened with welcoming remarks from Tom Kalil, Deputy Director for Technology and Innovation at the Office of Science and Technology Policy. Kalil highlighted the workshop's critical roles for the NSCI:

- Fostering public/private collaboration;
- Informing long-term planning;
- Exploring the underlying challenges of convergence of data-intensive and numerically intensive computing;
- Identifying hardware technology development paths for future HPC systems; and
- Improving productivity in the application of HPC and deployment through a well-trained workforce.

Kalil also announced the Administration's first nanotechnology grand challenge.[3] In June 2015, the Office of Science and Technology Policy issued a Request for Information seeking suggestions for *Nanotechnology-Inspired Grand Challenges for the Next Decade*. After considering over 100 responses, OSTP selected a challenge that supported three Administration priorities: the National Nanotechnology Initiative, the National Strategic Computing Initiative (NSCI), and the BRAIN initiative. The Challenge statement is:

> *Create a new type of computer that can proactively interpret and learn from data, solve unfamiliar problems using what it has learned, and operate with the energy efficiency of the human brain.*

This announcement set the tone for two days of forward-looking and ambitious discussion of research and development opportunities and priorities.

The technical sessions began with the two panels on Convergence of Data Analytics and Computationally-Intensive Computing.

The opening panel was moderated by Rob Leland (Sandia National Laboratories) and featured William Burnett (US Navy), Patricia McBride (Fermilab), Barry Chen (Lawrence Livermore National Laboratory), Ryan Quick (Paypal), and William Lapenta (NOAA) as panelists. To seed discussion, the moderator asked the panel to consider the following framing questions:

- What is meant by convergence?
- What are common underlying design principles?
- What are effective approaches for conversion?

The access and transfer of data was highlighted as a topic central to data analytics and computationally intensive computing. HPC activities can generate massive amounts of data, which leads to issues in storage of the produced data. The Large Hadron Collider can generate massive amounts of data, which has prompted intense investment in development of

[3] https://www.whitehouse.gov/blog/2015/10/15/nanotechnology-inspired-grand-challenge-future-computing

techniques to filter extraneous data. NOAA also generates petabytes of data, and is exploring avenues for extramural researchers to access and leverage these datasets. One audience member expressed the need for the management of large sets of unannotated data. Barry Chen responded that deep learning was promising in that it has been successfully used to annotate unstructured datasets. Supervised learning processes have been developed and implemented, and further investment in unsupervised feature discovery is absolute necessity. The audience member also noted that his use of distributed computer networks requires the movement of large amounts of data in real time. Ryan Quick commented that most data in existence is currently “in-the-wire”. Though photonics and other emerging technologies may hold promise in improving reliability of data transfer, Quick asserts that the underlying issues arise because the sender may not own the wire.

Several questions were directed to panelists about the potential for “spin on” (or government adoption of commodity solutions) in government computing. William Burnett stated that Department of Defense prerogatives entail stringent cybersecurity standards, which may limit the degree of adoption of privately developed applications. Quick replied that he believed that industry could meet security requirements if the Department of Defense were willing to communicate those stipulations.

Future trajectories of convergence seem to be largely dictated by application and field. Certain private industries may be able to manage higher degrees of risk than public institutions, like NOAA, which was cited to be more risk averse. However, this does not preclude private-public partnerships in the development of new technologies and methods.

The second panel was moderated by Dona Crawford (Lawrence Livermore National Laboratory) and featured Chris Johnson (University of Utah), David Bader (Georgia Tech), Peter Koenig (Procter and Gamble), Anna Michalak (Carnegie Institution for Science), and Fred Streitz (Lawrence Livermore National Laboratory) as panelists.

Dona Crawford initiated the discussion by asking how to mitigate risk involved in “bad analytics” or “bad data.” David Bader remarked that traditionally, HPC is concerned more about accuracy, while data analytics is more focused on making decisions, which entails embracing some degree of noise and risk. Michalak commented that in environmental modeling research, all data is imperfect data, as surrogate data is often employed. When informational content is limited, it is tempting to incorporate all that was collected – the central challenge is deciding what should be excluded from the analysis.

One audience member noted the need to document the computational practices at every step of analysis (algorithms employed, metadata, custody chain, etc.) as reproducibility issues may magnify at exascale. Chris Johnson agreed with this sentiment, and noted that there were packages in development for maintaining provenance. Michalak commented that open access to data and models is essential for reproducibility, as reconstructing models and datasets is infeasible. With true open access to data, capacity for community leverage grows by orders of magnitude and also improves reproducibility. Audience members reinforced that in addition to provenance, there must be true independence in reproduction of results.

An audience member asked for elaboration on Streitz's proposed algorithmic hypothesis machine, which could generate and test hypotheses autonomously. Streitz responded that the scientific method requires rigorous formalization before such a machine is actionable.

After lunch, the workshop shifted to the future computing technology theme. Thomas Theis delivered the first keynote of the workshop, *Emerging Devices for Computing: A Still-unexplored Landscape*.

Theis covered three broad topics leading to his conclusion. First, he reviewed the state of the art to establish both the need and opportunity for post-CMOS device research. The second topic was the importance and inherent challenges of growing the research investment in post-CMOS device technologies. The third broad topic focused on the challenges for benchmarking post-CMOS device performance; comparing and evaluating these disparate technologies will be a nontrival and perhaps controversial problem. His key conclusion was that a vast landscape of promising research possibilities remains unexplored, and that the convergence of new devices and new architectures has an exciting future.

Research and development in nontraditional architectures is generally materials agnostic. Though novel materials can enable new devices, it typically requires years of research to break barriers associated with developing and characterizing these new materials. When asked about his assessment on the affordability of printing of small structures, Theis stated that as long as there were economic drivers, there would be size reductions until physical constraints were reached.

The third panel was moderated by Irene Qualters (NSF) and featured Shekhar Borkar (Intel Corp), Wilfried Haensch (IBM Research), Steve Oberlin (Nvidia), Doug Burger (Microsoft), and John Martinis (UC Santa Barbara and Google) as panelists.

An audience member asked when non-CMOS technologies would stabilize comparably to that of conventional technologies. Doug Burger believed that non-CMOS tech development would prompt a verticalization within the industry. John Martinis said assembling large interdisciplinary teams was a major issue, a problem not observed in the private sector. Another audience member inquired if computer engineers were truly exploiting the discoveries of neural architecture from the BRAIN initiative. Panelists agreed that neuroscience would require very significant advances before biomimicry would become feasible. Doug Burger noted that, in launching massive commercial, high-value enterprises, introduction of new computing paradigms would prompt the hiring of specialized programmers within that space.

After the panel, workshop participants moved to five breakout sessions where they explored the Day 1 themes in additional detail. Summaries of the breakout sessions—ordered by topic and aggregated by groups-- appear in the next section.

Breakout Session

Convergence of data and computationally intensive computing

Group 1: Many in the group noted that the detailed points of the breakout questions seemed to imply that "data analytics" equated to a cloud paradigm while computationally intensive computing was equivalent to tightly coupled HPC. The group thought this simplification led to a path of false choices. It is much too early to attempt "convergence" within the dynamic space of data analytics. Platforms are numerous and experimentation of diverse approaches is essential. It is important to take an expansive view of "data analytics" including visualization, image analysis, assimilation, machine learning as well as graph analytics. Moreover, for both types of computationally intensive computing, a system view including I/O and storage as well as application investment needs greater attention in the discussion.

Group 2: The group started with a discussion on tightly coupled, HPC-style computing in data analytics. Tight coupling is especially beneficial when intensive data communication is needed. The attendees also considered the specific problems that need to be resolved for better convergence. Among those are real-time needs of data-oriented applications, which current HPC systems don't support well, differences in techniques for threading and latency tolerance suitable for data analysis (DA) versus HPC problems, and the likely increase in heterogeneity of DA-capable HPC systems. Unlike some DA problems, HPC applications generally cannot easily take advantage of cloud systems. Among the obstacles are issues of cost efficiency, security, and scalability. There is an important opportunity to learn from decades of HPC algorithm experience in order to create better data analytics for which the development of new mathematics is needed. Machine learning was mentioned as an example where HPC technology should be applied to gain efficiency. The discussion also touched on issues of Open Data; concerns were expressed when data reside in HPC systems, which may not be readily open for access. Open data also creates opportunities for accelerating innovation and "democratizing" research. A point was made that not all DA applications need HPC, thus care should be taken when choosing the right system balance. Approximate computing was discussed as an emerging technology that is especially interesting in the big-data area. "Hardware-in-the-loop", or hybrid simulations, were mentioned as interesting applications combining data and HPC. Another important class of such problems is decision support systems. Finally, the point was made that data is often messy, heterogeneous and complex; HPC systems need to be built with this understanding.

Group 3: Applications must drive the technology. The NSCI is putting heavy focus on hardware, but with globalization, it will always be challenging to maintain a U.S. advantage. The key advantage that we must always maintain is in the applications space.

Challenges in the software domain are abundant and are as critical as hardware for the success of the NSCI. Business requirements, particularly in big data management and computation, will increasingly rely upon a growing family of diverse software stacks. Convergence (in terms of one software stack to rule them all) is not likely to be achieved due to the broad spectrum of applications needs, and this will have a correspondingly big impact on workforce. We will need a workforce that is fluent in a more diverse set of languages. Data is also important, both in how it is managed, as well as how it is structured and analyzed. Part of new computing paradigms will require new data models.

Group 4: Overall, in current and foreseeable HPC, the cost of data movement relative to floating point arithmetic will remain a factor in implementing efficient algorithms. Strategies, such as reduced precision arithmetic or non-repeatable computation, show promise but may have limited impact if they do not factor in the cost of data movement.

More attention to developing effective workflows and to matching workloads to architectures will remain important. More effort should be focused on identifying methods for scheduling specific types of computational work to the most optimal environments is needed. Workloads exist that could function well enough on high-latency, asynchronous, low precision, or quantum systems. More broadly, when addressing the cost-effectiveness of the various methods for addressing "the full range of computing needs" it is imperative that the specific needs of the planning workloads drive a sufficiently variable solution rather than trying to cram highly disparate workload types onto a monolithic platform.

Cloud computing can play a role in HPC, but is unlikely to fit all needs. Current and projected "bleeding-edge" HPC architectures will not reach broad use in commercial clouds until there is an adequate market. In the meantime, scientific institutions (laboratories, universities, industries) that operate HPCs at high duty cycles (greater than 90%) will continue to find it cost effective to have dedicated systems under local control.

Group 5: There was broad agreement that some applications will benefit if systems representing this convergence of technology emerge, but with the recognition that achieving this convergence will require a significant effort. There were different opinions regarding the return on investment. Some of the applications that would benefit from a converged system could be implemented by two loosely coupled systems designed for data and computationally intensive computing, respectively.

The discussion on competing HPC paradigms was followed by a discussion of business models. The business models that drive cloud computing are an important component of their success, and need to be understood and applied to traditional computationally intensive computing and to converged systems to ensure economic viability. The group then discussed the role of public data, and a comparison was drawn to the Library of Congress and its procedures for physical publications. To ensure that data is maintained for its useful life, better planning is needed, including a plan to transition from public to private sector funding.

Future Computing Technology

Group 1: CMOS is the technology for the foreseeable future. While support is needed from both the federal and public sectors for future technology, including quantum, consideration of medium term technologies such as cryogenics merit renewed attention. More than one technology may be part of a future, more heterogeneous, computing environment. Future technologies for memory, storage, interconnects, and software should also be considered.

Group 2: The group spent substantial time discussing the relevance of a number-one United States position as technology leader. It is essential to better articulate the need for R&D investments, which have slipped in recent years, and to attract top young people to these fields. Security and economic considerations motivate the need for technology leadership. It was also mentioned that the government should ensure that the next “Silicon Valley” remain in the U.S. Other countries increasingly emulate U.S. success models, making it more difficult to maintain a leadership position.

The growing reliance on foreign technology is a significant concern, as some control of manufacturing is essential for national security concerns. On the question of roles of government and industry, it was pointed out that the two parties enable different aspects, so both are essential. Furthermore, there was agreement that public-private partnerships are key for maintaining and growing the economy and for increasing our knowledge base. Smart funding of startups can lead to leaps in innovation. Research needs long-term thinking and investment, as the development of future technologies may take decades. Government push is considered essential.

In addition to the technologies discussed in the workshop panels and talks, the development of optics for communication and switching (the latter being in early research state) and cryogenics were mentioned as promising future technologies.

Group 3: In terms of future computing technology, we are already behind the power curve and need to accelerate development of alternatives, both in digital and alternative computing paradigms. However, it is too soon to start picking winners and we need to invest in assessing a broad set of possible technology paths. The development strategies for government technology must roadmap through commercialization, not just to commercialization. A key area of investment must be how HPC technologies are integrated with standard business/production environments. “HPC in the Cloud” is a catchphrase, but it’s not sufficient to describe the full breadth of work needed.

Group 4: Future computing technologies, such as quantum computing, show promise and are an important investment area for the United States. Work remains to be done to understand these future architectures and the roles they may play in HPC may be different from and could augment current uses of HPC. Trained workforce needs in this area will be acute due to the inherent multidisciplinary nature of quantum computing.

Many in the group were optimistic regarding the possibility that future technology advances will continue to track Moore's Law and that, "where there is a will, there is a way." Given the potential for high profits, major companies will continue to push the frontiers of Moore's Law and it is not wise to doubt the fruits of innovation but, rather, to develop contingencies for if/when the pace of innovation slows or stalls. Specifically, "don't count silicon-based microelectronics out yet," given that steady advances over past three decades have occurred and it is therefore not unreasonable to expect further advances. However, the group expressed concerns that achieving feature sizes at and below 5 nanometers looks very difficult. In addition, the cost of constructing each successive generation of fabrication facilities is increasing and may become a limiting factor in the future.

Investments in ancillary technologies, such as communication and storage technology, will continue to underpin sustained progress in HPC. Examples the group identified include embedding communication technologies into CPUs themselves; software defined networks to optimize overall network efficiency; use of non-volatile memory, which would provide higher capacities and persistent storage; and next-generation photonic technologies, both on- and off-chip.

Group 5: While the participants noted the growing limitations of U.S. government influence on the private sector, the Federal Government still has opportunities to play a significant role in HPC evolution. Development of new technologies to support Post-Moore's Law computing paradigms, benchmarks for systems and applications, workforce, software libraries, and public-private partnerships were all identified as worthy targets of opportunity.

Highlights of breakout session

The importance of *convergence of computationally and data intensive computing* for emerging applications was acknowledged in all breakout sessions, but participants expressed a range of opinions on the importance of supporting these applications in a single system. For some classes of applications, loosely coupled systems (i.e., traditional supercomputer and a data analytic-focused system connected by a fast network) may provide an acceptable solution.

The groups also agreed that a focus on *future computing technologies* was timely, and that it was too soon to forecast the "winners". There was also broad agreement that our current CMOS infrastructure was likely to form the foundation for delivering any of these technologies. Ancillary computing technologies, such as networking, storage, and memory also have tremendous potential to transform the HPC landscape.

October 21 (Day 2)

Plenary Session

Day 2 began in plenary session with a summary of the Day 1 breakout sessions. Although the power failed during the first report, the moderators continued with their reports. Kathy Yelick proceeded with her keynote presentation, "More Data, More Science and… Moore's Law" without the benefit of a projector. Yelick's presentation covered three broad topics, and then offered observations on the implications for the NSCI. The three core topics were:

- Science is poised for transformation;
- Computing performance growth is slowing; and
- More productivity through better tools.

Yelick gave examples that evidence the transformation that science is undergoing today. Scientific computing, scientific globalization, and the possibility of outsourcing have fundamentally changed the scientific process and the nature of what it means to be a scientist. Finding, filtering, and de-noising data are some of the new challenges that scientists face.

The end of clock speed scaling and the impending end of transistor density scaling are the primary culprits for the disappointing trends in computer performance. When juxtaposed against a generic architecture for future HPC systems, the challenges to supporting the new science were clear. The community needs to develop an endgame for Moore's law, with an emphasis on exploratory work on how to take advantage of new technologies such as neuromorphic or quantum computing. New HPC systems with lightweight cores, explicitly managed memory, and increasing specialization of computing resources will bridge the gap in the interim.

Finding the right machine or modality and the right software can have a significant impact on productivity. Cloud computing is a business model and not a technology. The HPC community should look to cloud models for efficiency improvements. High bandwidth networks and libraries with efficient algorithms are essential since data movement has become the most costly component of HPC applications.

Achieving the high-level goals of the NSCI demands addressing the full range of computing, rather than just integrating enough chips to reach an exaflop. It will require reducing the cost of computing, improving efficiency with respect to electrical power requirements, develop technologies that decrease the size and increase the speed of computers, and most importantly, focus on solving real problems that require more computing.

One audience member commented that genome assembly is not a problem that requires more than tens of thousands of cores, and it is possible that some life science tasks do not need exascale systems. Yelick believed that the exascale computers would provide biologists a valuable resource for sequencing of metagenomes at the community level. For community library needs, applied mathematicians were the key for translating mathematical tools for scientific needs. Many branches of science will experience some aspect of convergence, which implies that it is necessary to identify the convergent elements and determine their solvability.

As in Day 1, workshop participants moved to five breakout sessions where they explored Day 2 themes in detail. Summaries of the breakout sessions—ordered by topic-- appear in the next section.

Breakout Session

Improving Productivity in HPC Application Development

Group 1: Productivity can have several meanings and should be defined more precisely, focusing on the user perspective. Additional layers of software abstraction are needed along with promotion of libraries and other forms of reuse. The market is changing and data analytics will be package-based. Innovation is difficult to direct from the top. There is value in collectively engaging and fostering exchange of ideas among diverse experts representing all layers of the platform stack (e.g. application, algorithm, and hardware).

Group 2: The group discussed a range of issues related to productivity, including: the tradeoff between usage, portability and performance; the need for flexibility and extensibility; as well as problems related to reusability and reproducibility. The large number of computer languages available today allows flexibility in expressing diverse problems but creates excessive complexity for users. The diversity of language environments raises further issues; for example, different compilers may produce different code and thus inconsistent results.

Productivity has improved only slowly over time. The urgency of the issue is generally not well understood and expressed. A key question posed was, how much performance are we willing to give up for productivity? The answer depends on many factors, such as user expertise, needed time to solution, and desired precision of the result. Among the new aspects of productivity in the present context are the likely emergence of specialized architectures and the need for data-oriented environments, requiring support for synchronous, streaming, and real-time processing. All these aspects further complicate the issue of productivity.

Abstraction was discussed as key for many solution approaches, as it hides complexity. Abstraction is needed at all levels of the software stack. Domain-specific languages are at the highest level of abstraction and show promise. Many attendees agreed that a spectrum of users and programmers must be supported, from “hero” programmers to domain users who are not computer scientists; different tradeoffs between abstraction and performance will be needed. Translators were mentioned as a possible solution to the issue of language variety. Integrating research and operations may help improve productivity. Finally, community portals have the potential to increase productivity significantly for classes of domain users.

Group 3: In addition to scientists and engineers, enhancing HPC “literacy” is essential in all business sectors and decision makers. The value of HPC should be evident to CEOs, not just CTOs. The business impact of HPC should be part of the MBA curriculum. The community should find ways to tell a compelling HPC story to a broad set of leaders in the public policy sector (Congress, Administration) and also in industry and academia. Messaging will be critical to the success of the NSCI.

Group 4: Greater and easier access to HPC systems would benefit the broader scientific community and would quicken the pace of scientific discovery. Improving the discoverability of HPC software for users, by using mechanisms such as a national library of HPC software/algorithms/data formats that are used across the US agencies, would improve broad usability of HPC systems.

HPC applications and large-data analytics would benefit by using rigorous software engineering principles throughout, for both current and anticipated future computing technologies. The development of appropriate standards and more consistent adherence to software best practices is important in the move from petascale to exascale computing. It is important to gain a more robust understanding of the uncertainties inherent in the computing environments. Inadequate documentation and a lack of good coding practices limit code reuse and sharing. Opportunities exist for improving collaboration between government agencies to share code. Domain specific languages (DSLs) could play a useful role in lowering the barrier of entry for scientists to use HPC systems. A significant challenge in this area is to ensure the abstractions created by DLSs do not significantly reduce overall performance (i.e. time to solution). Software frameworks have the potential to provide a basis upon which domain experts could build domain specific languages.

Group 5: Challenges for software development range from funding to the expenses and opportunity cost associated with maintaining HPC code. Developing an open source community would be a start, but architecture-independent abstractions are needed to preserve the benefits across computing advances. Workforce development is an expected additional benefit from open source software investments.

To expand HPC access, HPC will need to become more user friendly. Web portals, such as the NSF Science Gateways, are a promising mechanism for researchers. Emerging virtual machine models, such as the NSF Comet system, are promising mechanism for reaching small businesses. Increasing usability of workflows was also highlighted as a critical need. The need for a cloud-like business model was revisited, with a discussion of "HPC in the cloud".

Achieving Broad Deployment and Creating a Trained Workforce

Group 1: There is a dearth of students in the pipeline for HPC; a lack of computer science teaching in K-12. Neither computational science nor data science have a pedagogical home inhibiting curriculum development and adoption. This is further aggravated by academic turf issues and the changing view of the legitimacy of computer science as a science discipline. Workforce development and broad deployment necessarily encompass and confront the issue of poor diversity. Curriculum delivery vehicles which have low barriers to entry such as video on demand will be needed to attract high school students. Both broad deployment and workforce development also require consideration of career paths not based on current work practices. A discussion of privacy in the context of the 21st century is needed.

Group 2: The discussion focused mainly on the issue of workforce development. Strengthening the workforce in HPC and data processing is a key concern. The lack of student interest in general STEM disciplines is well known and is at the basis of the problem with building an HPC (and data) workforce. One solution approach is to show how STEM disciplines relate to end products, especially products that young people can relate to. The success of robotics teams was mentioned in this context. Other “grand challenge” teams or competitions could be organized, such as the Cluster Challenge at SC conferences. The underrepresentation of women in the HPC workforce was of special concern. Early intervention in grade and middle school is as important as for all STEM disciplines. Teaching HPC from the “bottom up” was also mentioned as an interesting and successful approach – students learn from the transistor up what is inside an HPC system. We must also make efforts to understand the problem of workforce development better and more quantitatively. To this end, defining and measuring metrics was mentioned is important. Not all scientists using HPC need to become computer scientists. Interdisciplinary teams are considered essential for combining domain and computer science expertise. Funding considerations should keep this balance in mind. In general, robust programs in diversity for broadening participation in HPC are needed.

A short amount of time was spent discussing issues of deployment. It was mentioned that small- and medium-sized enterprises may not understand the potential for return-on-investment from HPC. Investing in efforts that demonstrate this potential may be worthwhile. It was also pointed out that mobile devices aren't properly leveraged yet and that decision-support systems could profit from HPC. Both technologies could bring economic advantages.

Group 3: This group had considerable concern regarding the shortage of STEM workers. There remains a need to find ways to engage the American public and grow an HPC workforce, starting in K-12. Some suggested means of engaging the public included utilizing grand challenges, “computing/coding bees”, etc. to inspire and capture the imagination. Like math, coding and computer science should be a baseline skill for all technical workforce, not an esoteric discipline in itself. Additionally, some level of understanding and familiarity is necessary for the non-technical workforce as well, especially those in management.

Group 4: Availability of competent, trained workforce in HPC applications and large-data analytics should remain a priority, both for developing the tools needed in these areas and for producing people trained to use the tools.

Cyber security remains a challenge in HPC at multiple levels, including supply chain dependencies, verifiability of microcircuit functionality, and assurance that software does not contain unsecure or malicious code. Further research is warranted to alleviate and mitigate the overheads imposed by cyber-security.

Group 5: Workforce development was another popular subject. The participants felt that HPC should be introduced to students at a much earlier age, perhaps 7th grade. To expand HPC utilization to new communities of researchers, new efforts in the graduate school stage will be required. Graduate students would benefit from a dual mentor model, with mentors in their domain science and computer science.

Highlights of breakout session

A common refrain for *improving productivity in HPC application development* was abstracting the complexities of architecture dependencies away from the HPC user. Moreover, the abstraction layer should be configurable so that it can be tailored to the user's level of expertise. The advantages of this approach will have to be balanced against the loss in application performance. While some thought that the heterogeneity in programming languages today were a barrier to productivity, other suggested that domain specific languages could accelerate adoption and productivity for some.

All groups agreed that for *achieving broad deployment and creating a trained work force*, there was a thin pipeline of HPC students. Some suggested increasing the exposure of K-12 communities to HPC and computer science while also increasing the number of women and underrepresented groups in the field. Grand challenges, prize competitions, and "coding bees" were identified as effective tools for HPC outreach to new communities. A cybersecure HPC ecosystem is essential for broad deployment and further R&D is warranted to ensure minimal impact on HPC performance.

Workshop chair Randy Bryant (OSTP) provided closing remarks. He began by thanking the speakers, scribes and staff for their contributions, DOE for hosting the workshop, and the attendees for making the workshop such a success. Computing technologies will continue to shift and evolve and effective policy must be as flexible as the technology and practices it guides.

Panel 4 was moderated by Jerry Blazey (Northern Illinois University), and featured Ewa Deelman (University of Southern California), Keshav Pingali (University of Texas, Austin), Michael Franklin (UC Berkley), Patrick McCormick (Los Alamos National Laboratory), Barbara Chapman (University of Houston and Stony Brook), and Todd Gagorik (Amazon) as panelists. In addition to the return of power during the second presentation, this panel was distinguished by lively discussion of the challenges entailed by the NSCI's goal of broad deployment.

Deelman envisions mobile apps that initiate hybrid workflows where HPC resources are supplemented by small computer post processing. The obstacles to such a future include resistance to change by scientists, funding and marketing software advances, and the analysis and interpretation of HPC results. Pingali opened with a broad discussion of parallel programming and lessons learned from successful application in various domains. He proposed data-centric (rather than control-centric) paradigms for future systems. Franklin discussed the dual roles of student driven code, where learning opportunities can also lead to important innovations. The HPC community needs to harness the power of the open source community to achieve these goals for a broader community. He closed by noting that aiming at productivity enhancements does not preclude scientific innovation. McCormick focused on the importance of data abstraction, the performance costs and benefits of enhanced scheduling algorithms, and the importance of forcing cross-discipline interactions. Chapman focused her remarks on the implementation of APIs and community standards, and the key role multidisciplinary experts play in the deployment and application of these tools. Training in domain science and the skill sets needed for high end computing development is ideal, but current educational environments do not provide this combination. Gagorik focused on Amazon's experience with large architectures and the hyperscaling of workloads on cloud platforms. The Amazon cloud has proven its utility for computational research on a variety of workloads. He closed by noting that focusing on computational speed without understanding developer time use can result in overall inefficiency.

Audience members commented that cloud computing represented a massive commercial investment, and the displacement caused by cloud resources may not be salutary, as tools have already been developed for other system. Gagorik acknowledged this concern, but believed that democratization of compute power could prompt the scientific community to resolve these issues. Laptop-accessible cloud tools enable the community to tinker and build on a wide scale. Deelman remarked that though commercial vendors provide a tremendous resource, true computational democratization would require a trained user base. One audience member cautioned overselling the cloud to scientific communities. Cloud resources may have hidden costs as researchers need to tweak the system when moving data to and from the cloud in real time. This may require changes in cost models. Gagorik replied that Amazon is currently working on transport fee models for the scientific community. Deelman identified workflow tools as a potential solution to some of these issues, as abstraction provides scientists a "higher level" solution and improves reliability. Franklin stated that there was a general overemphasis in the need for computation speed and performance. Current benchmarks favor absolute speed while neglecting other cost benefits. As systems become more complex, the need for abstraction increases.

Conclusion

During the workshop, several ideas emerged that provide context to the NSCI and inform its implementation. Some of the ideas are:

- Data analytic computing spans a wide range of computing styles, calling for different machine organizations and programming approaches. The challenges mount when real-time processing is required and when complex patterns must be discerned across large volumes of data. This is a field that is rapidly evolving at both the application and implementation levels, giving uncertainty in how HPC for data analytics will develop.

- Coupling data-analytic computing with simulation could greatly advance many fields of science and engineering. Currently, these tasks are typically performed in sequence, often on different machines. Coupling them together in a loop on a single machine would enable richer interactions, for example in driving an iterative optimization process with simulations providing validations for hypotheses generated through data analysis.
- Improvements in hardware technology for HPC can proceed along two different fronts. First, there are many opportunities to squeeze more performance out of CMOS, by migrating application function into hardware and by reducing the amount of data movement. Second, fundamentally new approaches, including neuromorphic, superconducting, and quantum technologies must be explored to ensure the long-term health of HPC. The NSCI must accommodate a breadth of possibilities and avoid any premature down select of future technology.
- There are many different deployment models for HPC, including today's batch-oriented HPC centers as well as cloud computing resources that support demand scheduling through resource virtualization. Furthermore, users access these resources at different levels, from direct programming of machines to application-level services. The NSCI can support the exploration of different delivery models to determine the most effective way to meet user needs while keeping an eye on resource utilization and cost.
- Although much of the discussion of HPC focuses on hardware, software presents a much bigger set of challenges. The NSCI must support efforts to increase the productivity of HPC software development through higher degrees of automation and through increased code reuse and sharing.
- Ensuring a trained HPC workforce will continue to pose challenges. Shortages will be especially acute for government and academic institutions when competing for talent with industry. Deploying HPC across a wider range of organizations will require both a larger, trained workforce that includes more women and underrepresented groups, as well as ways for less experienced users to tap into the expertise of seasoned developers. Doing this will require deeper engagement with the industrial (non-computing) sector.

Finally, it is necessary for NSCI agencies to continue to engage with industry and academia to build a whole-of-Nation effort. Government, industry, and academia will benefit from a broader understanding of the potential scientific and economic impacts, emerging technologies, and shared challenges.

Appendix A: Organizing Committee

Vikrum Aiyer	National Economic Council, Executive Office of the President
Steve Binkley	Office of Science, U.S. Department of Energy
Rupak Biswas	National Aerospace and Aeronautics Administration
Randy Bryant[4]	Office of Science and Technology Policy, Executive Office of the President
Saul Gonzalez Martirena	Office of Science and Technology Policy, Executive Office of the President
Susan Gregurick	National Institutes of Health
Brian Gross	National Oceanic and Atmospheric Administration
Barbara Helland	Office of Science, U.S. Department of Energy
Peter Highnam	Intelligence Advanced Research Projects Activity
Rudi Eigenmann	National Science Foundation
Dimitri Kusnezov	National Nuclear Security Administration, Department of Energy
Brad Martin	National Security Agency
Glen McWright	National Security Agency
Bob Meisner	National Nuclear Security Administration, U.S. Department of Energy
David Michaud	National Oceanic and Atmospheric Administration
Reed Mosher	Department of Defense
Tim Polk	Office of Science and Technology Policy, Executive Office of the President
Irene M. Qualters	National Science Foundation
Erin Szulman	Office of Science and Technology Policy, Executive Office of the President
Doug Wade	National Nuclear Security Administration, U.S. Department of Energy

[4] Chair of the Organizing Committee

Appendix B: Workshop Agenda

THE WHITE HOUSE
WASHINGTON

White House National Strategic Computing Initiative Workshop

Themes: convergence of data-intensive and numerically intensive computing; hardware technology for future HPC (beyond Moore's law); and improving productivity in HPC application development and deployment, and workforce development.

Day 1 | October 20, 2015 | 8:45 AM – 5:00 PM

Registration (8:15-8:45)

Welcoming Remarks (8:45-9:15)

Logistics (9:15-9:30)

Panel 1: Convergence of Data Analytics and Computationally-Intensive Computing (9:30-10:30)

Moderator: Rob Leland
Sandia National Laboratories

William H. Burnett
U.S. Navy

Barry Chen
Lawrence Livermore National Laboratory

Patricia McBride
Fermilab

Ryan Quick
PayPal

Bill Lapenta
National Oceanic and Atmospheric Administration

Break (10:30-10:45)

Panel 2: Convergence of Data Analytics and Computationally-Intensive Computing (Further Discussion) (10:45-11:45)

Moderator: Dona Crawford
Lawrence Livermore National Laboratory

David Bader
Georgia Institute of Technology

Chris Johnson
University of Utah

Peter Koenig
Procter & Gamble

Anna M. Michalak
Carnegie Science & Stanford University

Fred Streitz
Lawrence Livermore National Laboratory

Lunch (11:45-12:45)

Keynote (12:45-1:30)

Thomas Theis
Semiconductor Research Corporation

Panel 3: Future Computing Technology (1:30-2:30)

Moderator: Irene Qualters
National Science Foundation

Wilfried Haensch
IBM Research

Steve Oberlin
NVIDIA

Shekhar Borkar
Intel

Doug Burger
Microsoft

John Martinis
University of California, Santa Barbara & Google

Break (2:30-3:00)

Breakout Sessions (3:00-5:00)

THE WHITE HOUSE
WASHINGTON

Day 2 | October 21, 2015 | 9:00 AM – 3:30 PM

Summary of Breakout Groups from Day 1 (9:00-9:30)

Keynote (9:30-10:15)

Kathy Yelick
Lawrence Berkeley National Laboratory
University of California, Berkeley

Break (10:15-10:30)

Panel 4: Improving productivity in HPC application development and deployment, and workforce development (10:30-11:45)

Moderator: Gerald Blazey
Northern Illinois University

Ewa Deelman
University of Southern California

Patrick McCormick
Los Alamos National Laboratory

Keshav Pingali
University of Texas

Barbara Chapman
University of Houston & Stony Brook University

Michael Franklin
University of California, Berkeley

Todd Gagorik
Amazon

Lunch (11:45-12:45)

Breakout Groups (12:45-2:30)

Summary of Breakout Groups from Day 2 (2:30-3:00)

Closing Remarks (3:00-3:30)

Appendix C: Breakout Topics

This section of the report begins by documenting the instructions to facilitators and questions presented to the workshop participants, and follows with a summary of discussions in each of the five respective breakout sessions.

Facilitators were asked to guide their group's discussion of two topics per breakout session. The topics and a set of prompts to encourage discussion were provided to moderators and participants. Facilitators were instructed to view the prompts as a starting point rather than limitations for the discussion; if the group was interested in exploring other aspects for a particular topic, they were free to discuss them as well. The facilitators and scribes for the five breakout groups are listed in the following table.

Group #	Facilitator	Scribes
1	Irene M. Qualters (NSF)	Kevin Newmeyer, Marian Nemec
2	Rudi Eigenmann (NSF)	Gabe Perez-Giz, Claire Schulkey
3	Will Koella (DOD)	Alejandro Suarez, Carolyn Lauzon
4	Steve Binkley (DOE)	Alex O'Bannon, Garrison Vaughan
5	Tim Polk (OSTP)	John Russell, Eric Stahlberg

The topics and prompting questions were the following:
Topic 1: Convergence of data and computationally-intensive computing
(As covered by Panels #1 & #2)

- Is there a role for tightly-coupled, HPC-style computing in data analytics?
- What is the most cost-effective way for an organization to support its full range of computing needs, and to allow multiple types of computing (e.g., simulation + data analytics) to interoperate?
- Can modeling & simulation algorithms be devised that can operate in a more asynchronous, high-latency computing environment?
- Is there a role in HPC applications for approximate computation, either using lower precision or allowing nonrepeatable results (e.g., due to hardware errors or data races)?
- What is the long-term role for large, shared-memory machines in terms of scalability and application demand?
- What should be the priorities for federally-funded research investment in these areas?

Topic 2: Future computing technology
(As covered by Dr. Theis' keynote and Panel #3)

- Can we rely on industry to keep CMOS going for the foreseeable future?
- If progress in core technology slows down, how much could be gained with better circuit designs, better architectures, and better programming approaches?
- What is the potential role of programmable hardware in HPC?

- How much should the U.S. be supporting research in quantum computing?
- What are the prospects for technology that will not be suitable for consumer use (e.g., ones that require cryogenic operation)?
- In addition to computing technology, which aspects of communication and storage technology must be addressed to enable sustained progress in HPC.
- How can the U.S. play a role in technology, given that much of the industry is not based in the U.S.?
- Given the huge industry investment in technology, what roles should Federal research funding play?

Topic 3: Improving productivity in HPC application development
(As covered by Dr. Yelick's keynote and Panel #4)

- What factors limit code reuse and sharing in HPC?
- What steps can be taken to reverse the trend that application programs must increasingly be tuned to the computing, memory, and communication structures of specific machines? To what extent can HPC programming be made target independent?
- How can future programming models capture the concerns of communication cost and the benefits of locality?
- Can domain-specific languages and programming environments play a useful role in HPC? What are some example application domains?
- What should be the priorities for federally-funded research investment in these areas?

Topic 4: Achieving broad deployment and creating a trained workforce
(As covered by Dr. Yelick's keynote and Panel #4)

- What lessons can we learn from cloud services and other resource models to promote deployment and commercialization of converged HPC systems?
- What strategies would best support development of a trained HPC application developer workforce? What factors are impeding workforce development?
- How important will educating nonspecialists (i.e., subject matter experts in other fields) in the use of HPC an important factor for broad deployment?
- How can we promote HPC as an effective approach to solving problems of societal importance by the current and next generation of scientists and engineers?
- How do we develop deployment strategies that recognize and promote the benefits of a diverse and changing landscape of technology innovation?
- Given increasing requirements to support large dynamic data flows and interoperability with national and international instruments, networks, and communities, how can we ensure that robust, secure, and high-performing network fabrics remain a strength of the evolving HPC ecosystem?
- What other barriers must be addressed in achieving broad deployment? (E.g., dealing with export controls, cybersecurity, etc.)

Appendix D: Presentations

White House NSCI Workshop Presentations – October 2015

Appendix E: Abbreviations and Acronyms

CMOS	complementary metal oxide semiconductor
DA	data analytics
DOD	Department of Defense
DOE	Department of Energy
EC	Executive Council
HPC	high performance computing
NCO	National Coordination Office
NITRD	Networking and Information Technology Research and Development
NOAA	National Oceanic and Atmospheric Administration
NSCI	National Strategic Computing Initiative
NSF	National Science Foundation
OSTP	Office of Science and Technology Policy
STEM	Science, Technology, Engineering and Mathematics