Comments on Interim Report:
"*Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020*"

The NRC Committee on "*Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science and Engineering in 2017-2020*" issued an Interim report that raised several issues and questions. The committee requested comments on these issues and questions before preparing its final report. Following is a list of comments in response to the questions raised in the Interim Report.

Thom H. Dunning, Jr.
Co-director, Northwest Institute for Advanced Computing
Pacific Northwest National Laboratory & University of Washington

Professor, Department of Chemistry
University of Washington

Professor Emeritus, Department of Chemistry
University of Illinois at Urbana-Champaign

1. How can we create an advanced computing infrastructure that enables integrated discovery involving experiments, observations, analyses, theory and simulation?

To respond to this question, it is critical to recognize the increasingly vital role played by simulation- and data-enabled research in science and engineering. This is a direct result of the increasing scope and accuracy of simulations of natural and engineered systems as well as the growing volumes of data generated in simulations, experiments and observations. In addition, it is important to recognize that computational and data science is no longer the sole preserve of theoretical and computational scientists and engineers. Many experimental and observational scientists and engineers now use this approach to provide insights into the phenomena that they are studying, especially if the needed computational applications have user-friendly interfaces.

In planning an advanced computing infrastructure, NSF must also take into account the varied ways that simulation- and data-enabled science and engineering is used in research today. For example, activities such as the development of new mathematical approaches and computational techniques have different computational requirements than the development of applications to be used by a large community of researchers or the use of these applications for science and engineering research. To complicate matters further, the researchers in different disciplines often need different computing resources as well as different operational modes (batch, interactive, *etc.*) to address the problems in their fields. Thus, a computer system operated for large-scale applications may not be well suited for software development activities, nor a system configured for physics applications well suited for bioinformatics applications. These varied needs must all be considered to achieve an optimal solution.

Given the above challenges, it is clearly time for NSF to develop a comprehensive plan for the computing and data infrastructure needed by science and engineering communities to advance their research. This does not mean that NSF must provide all of the needed computing capabilities, research institutions should contribute their share while industry can contribute other capabilities, e.g., cloud computing services. What is needed now is an overall plan for a national computing and data infrastructure that meets the many varied needs of computational scientists and engineers. *The development of such a plan is beyond the charter of the current NRC Committee, but its creation could be one of its recommendations.*

Although the above focused on the use of computing in research, another critical endeavor is the infusion of simulation- and data-enabled science into the educational curriculum, from high school through undergraduate school to graduate school. Every young scientist and engineer should be exposed to computation- and data-enabled science and engineering as an integral part of the curriculum, just as experimental science and engineering has played a critical role in such courses in the past. NSF should explicitly encourage and support partnerships with universities to create the computational infrastructure needed at the campus level for educational purposes (which could, of course, be intimately linked with the research computing infrastructure). These partnerships could include regional consortium, which would provide economies of scale as well as foster collaborations in the development of computational modules for inclusion in existing courses as well as the development of new computational courses.

Much of the needed computing infrastructure can be funded through existing programs at NSF—the Major Research Instruments (MRI) and Advanced Cyberinfrastructure (ACI) programs. However, the cost of the most advanced computing systems is beyond the current budget of these programs. There are two means of approaching this problem. The first is to augment the existing MRI and ACI budgets. The second is to augment the MRI budget as before but explore the funding of an advanced cyberinfrastructure for the nation through the Major Research Equipment & Facilities Construction program. However, this program, as it is currently configured, is not designed to accommodate major instruments that are evolving as rapidly as computing systems. So, changes would need to be made to the MREFC program to ensure that there is a well thought-out plan for upgrading/replacing the computing systems on a regular basis. More will be said on this option later; however, it should be noted that both of these options require additional funding.

2. What are the technical challenges to building future, more capable advanced computing systems and how might NSF respond to them?

The challenge of designing and configuring advanced computing systems that advance science and engineering research will be even more challenging in the future than it is now. To date, few applications can take advantage of the highly parallel nature of modern computing systems and this problem will be further exacerbated by the inclusion in accelerators in future systems, a trend that has already begun. In addition, future science and engineering applications will have to contend with less memory per core, less memory bandwidth per flop, and less I/O bandwidth per flop. To date, the focus of many research organizations has been to ensure that the computer system they are procuring ranks high on the Top 500 list as a means of convincing their sponsors of the value of the investment they made. Placement on this list is determined by the performance of the system on the Linpack Benchmark. Although this benchmark was, with some exceptions, a useful benchmark a couple of decades ago, it does not represent the broad range of applications running on today's computers. Further, the desire for a high Linpack score often has undesirable side effects, *e.g.*, skimping on the amount memory or I/O bandwidth, neither of which contribute to the Linpack ranking, but which may be critical for solving the science or engineering problems of interest.

NSF broke with the Top 500 tradition when it supported the University of Illinois' decision to forgo the submission of the Linpack Benchmark results for the Blue Waters petascale computing system. NSF should forgo the Linpack (or other simple benchmark) "beauty contest" and maintain its focus on computing systems that advance computational science and engineering by acquiring computers that are configured to truly meet the needs of this community. However, computing technology continues to evolve. This requires NSF to simultaneously fund activities that promote the use of innovative computer technologies that promise to advance computational science and engineering. In addition, NSF could establish a research program in alternate compute architectures that are better suited to advancing science and engineering research.

Finally, there is one other significant gap in NSF's support for computation- and data-enabled science and engineering that exacerbates this problem. As noted above, few existing science and engineering applications can take full advantage of modern, highly parallel computing systems, especially those containing both standard microprocessors and accelerators. To make matters worse, we are not educating a new generation of computational scientists and engineers who can develop these applications. Although it is possible to fund this type of educational activity through the other NSF directorates, and some, in fact, do fund such activities, it would be appropriate for ACI program to also address this critical educational need, especially as it involves high-end computing.

3. How can the match between resources and demand for the full spectrum of systems, for both compute and data-intensive applications, be managed and

what are the impacts on the research if NSF can no longer provide state-of-the-art computing for the research community?

NSF has long been effective in using its "bully pulpit," namely, funding, to advocate for needed changes in science and engineering research. Rather than abrogating its responsibility to provide the computing resources needed by the science and engineering research community, NSF should develop a comprehensive plan for marshaling federal, state and institutional resources to meet this critical need. This plan can draw on resources currently available at NSF (MRI, ACI and Research & Related Activities) as well as at other federal agencies (especially those whose grantees make use of NSF computers). Many research universities and other research institutions are already investing in the computing resources needed by their faculty, either as a match for federal funds or as a discrete investment from university, state and other funds. Because of the growing importance of simulation- and data-enabled science and the potential that it has for advancing science and engineering, I believe it will be possible to develop a shared vision for a national research and education cyberinfrastructure that would be supported and contributed to by many stakeholders. But, we need to articulate a clear vision for this infrastructure and show how NSF will contribute to its creation.

As noted above, the cost of high-end computing systems, *i.e.*, capability systems, has risen to the point that it may no longer be possible to fund these systems within the ACI Division unless it receives a substantial increment in its budget (likely to be at least $100 million per year with an additional increment for the MRI program). One option for funding the construction of a national advanced cyberinfrastructure along with managing and operating this infrastructure is NSF's Major Research Equipment & Facilities Construction (MREFC) program. If this done, the program would also need additional funding and its processes adapted to the continuing evolution of computing technology.

Adaptation of the MREFC process to support the construction, management and operation of a National Advanced Cyberinfrastructure (NACI) has the advantage that:

- The NACI would be designed with input from scientists and engineers from all of the directorates at NSF, ensuring that it serves the entire research community—established computing communities as well as budding computing communities; life science and social, behavioral and economic sciences communities as well as physical science, geoscience and engineering communities; methods and software developers as well as disciplinary researchers.
- The issue of funding continuity would be addressed upfront. In the past, computer systems were awarded competitively every 4-5 years, often resulting in disruptions to the expertise assembled at any site that was not renewed with a corresponding loss of expertise and support for the national computational science and engineering community.

- Planning for the MREFC project would result in a more comprehensive understanding of the compute, data, networking and software requirements as well as the issues to be addressed in integrating laboratory, university and national cyberinfrastructures.

4. **What role can private industry and other federal agencies can play in providing advanced computing infrastructure, including the opportunities, costs, issues, and service models, as well as balancing the different costs and making tradeoffs in accessibly (e.g., guaranteeing on-demand access is more costly that provide best-effort access)?**

Over the past fifty years or so, industry has tried to play a role in providing computing services. By and large, these attempts have been failures. The one bright spot currently is cloud computing. Industry appears to be interested in providing cloud-computing services and, since this is a generic service, they may be able to do so successfully. The NACI should factor private industry into its plan for those applications that are well served by this computing approach.

Many other federal agencies, *e.g.*, the U.S. Department of Energy, the National Aeronautics and Space Administration, and the Department of Defense, fund high-end computing. The National Institutes of Health support software development for biomedical applications, although they tend to rely on the high-end computing systems supported by other federal agencies, including NSF, for research requiring those capabilities. The computing and data capabilities provided by these agencies should be included in a plan for an NACI. However, as mission agencies, it would be unwise for NSF to turn to these agencies to meet their advanced cyberinfrastructure needs. NSF’s charter is to support the broad range of science and engineering critical to the future of the nation. This places a set of constraints on the cyberinfrastructure that will likely not be satisfied by the cyberinfrastructures selected by the mission-oriented agencies, whose mission may be satisfied by a far more limited set of science and engineering applications. Further, security considerations may place additional constraints on access to the computer systems operated by these agencies.

The processes advocated in #1, #3 and #6 will lead to a more comprehensive understanding of the needs of simulation- and data-enabled science, which will naturally lead to identifying the research that requires different types of computing resources.

5. **What are the challenges facing researchers in obtaining allocations of computing resources and what suggestions would you make for improving the allocation and review processes for making advanced computing resources available to the research community?**

Currently, computational researchers needing large-scale computing resources are placed in a double bind. First, they must submit a proposal to NSF (or other agency) to fund their research project. Then, once the project is selected for funding, they must submit a request for an allocation of computer time to support the project. Although all parties attempt to make the process work smoothly, issues still arise, *e.g.*, the allocation of computer time may not be sufficient to complete the project. NSF should explore ways to better integrate these two processes.

NSF could establish an internal two-step process. The importance and innovativeness of the research outlined in a proposal would first be evaluated and then, if the proposal is deemed worthy of funding and requires an allocation of time on NSF-supported computer systems, NSF could ask a committee established by ACI to evaluate the computing request. This has the advantage of eliminating the double bind and, in the end, would enable the NSF program manager(s) and Principle Investigator to negotiate a scope of work that was consistent with the computer time allocation. This approach has the disadvantage that not all researchers would be covered, *e.g.*, those funded through other agencies, and it can be difficult to estimate the amount of computer time required for an given project (this is, after all, research). But, it would be worthwhile to establish a pilot project to test such a concept.

6. **Can the wide collection and more frequent updating of requirements for advanced computing be used to inform strategic planning, priority setting, and resource allocation; how might these requirements be used; and how might they best be developed, collected, aggregated, and analyzed?**

As noted above, it is time for NSF to begin systematically collecting and analyzing the computing needs of the research communities they support. The reason for this is two-fold: (*i*) the computing needs of most well established research communities, e.g., chemistry, physics and climate, are growing rapidly, driven by the need for more and more accurate predictions and the treatment of more and more complicated systems, and (*ii*) an increasing number of traditionally non-computing oriented research communities have a rapidly growing need for simulation- and/or data-enabled approaches to advance their research. This is not simply a matter of recording who is currently using NSF-supported computers; it involves a much deeper engagement with the various research communities. NSF needs to understand the current computing and data demand, but it also needs to understand how this demand will evolve in the near (5-year) future.

Given that (*i*) simulation- and data-enable science and engineering are still in a developmental stage in many disciplines and (*ii*) computing technology is continuing to evolve at a rather rapid pace, the collection, aggregation and analysis of the computing requirements is a rather daunting task. Nonetheless, it is imperative to begin this process and to develop ways to deal with these complexities. The collection, aggregation and analysis of the computing needs of the science and engineering research community could be done as a continuing part of the MREFC process, if this is the approach taken to develop a NACI, or it could be assigned to NSF's Advisory Committee on Cyberinfrastructure (with appropriate NSF staff support).

7. **What is the nature of the tension between the benefits of competition and the need for continuity as well as alternative models that might more clearly delineate the distinction between performance review and accountability and organizational continuity and service capabilities?**

Considerable expertise must be assembled at a site to select/configure, install, manage and operate an advanced computing system and support the community of scientists and engineers using that capability. Expertise is required in systems administration, file systems and software, storage systems and software, visualization systems and software as well as a broad suite of scientific applications. Competing the siting for these computing systems every four-five years, as is the current NSF practice, makes assembly of such expertise difficult and threatens its stability once it has been assembled. There are also many hidden, but nonetheless very real inefficiencies, associated with the recompetitions beyond the churn of staff. These include the time that must be devoted to preparing and defending proposals and the dampening effect that recompetitions have on collaboration between the sites.

The effectiveness of the recompetitions to "raise the bar" depends on the willingness of universities to invest a substantial amount of their funds to ensure the competitiveness of the proposal from their institution. For many institutions, these investments will be very difficult to maintain in the future. NSF should develop a long-term vision for siting its computing infrastructure that ensures the stability of the set of sites needed to resource the NACI. To ensure the success of this new approach, NSF could take the approach used by some other federal agencies, namely, to conduct regular performance reviews and terminate the cooperative agreement if performance deficiencies are found that cannot be corrected. This would address the continuity problem, while ensuring a high level of quality control.

8. **How can NSF best coordinate and set overall strategy for advanced computing-related activities and investments as well as the relative merits of both formal, top-down coordination and enhanced, bottom-up process.**

There are existing examples of the type of coordination and planning being discussed here in other fields of science and engineering. For example, the Particle Physics Project Prioritization Panel (P5) is responsible for drawing up a consensus plan for investments in high-energy physics based on input from the research community; see:

http://www.usparticlephysics.org/p5/

Although this is a much more homogeneous group than would be the case for simulation- and data-enabled research, it is an example of a scientific discipline gathering the data needed to make difficult recommendations about the investments needed to continue to advance the state of knowledge in their field. The astronomy and astrophysics community also regularly undertakes a decadal survey with the intent of identifying new scientific opportunities and making investment recommendations, see:

In both of these cases, the recommendations in the reports provide invaluable guidance for the funding requests made by the relevant federal agencies.

The diverse nature of the simulation- and data-enabled research communities is certainly a complication in developing an overall strategy for a national advanced computing infrastructure. Another complication is that progress in simulation- and data-enabled science and engineering requires the integration of several components: computer systems, data systems, visualization systems, networks, and many, many pieces of software, from computing systems software to science and engineering applications. The interplay between all of these components must be considered and this will vary depending on the research problem to be solved. The coordination and planning process must keep these complexities in mind. Nonetheless, it is time to begin.