Reflections on an Observation of the Interim Report of the Committee on Future Directions for NSF Advanced Computing Infrastructure to Support U.S. Science in 2017-2020.

Texas Advanced Computing Center

The interim report observes that the NSF may be unable to invest in appropriate advanced computing infrastructure in the future, *"But it is unclear, given their likely cost, whether NSF will be able to invest in future highest- tier systems in the same class as those being pursued by the Department of Energy, Department of Defense, and other federal mission agencies and overseas".* The committee suggests that a possible solution for the future involves acquiring such infrastructure from external (i.e., outside NSF) sources: *"Options for providing highest-tier capabilities that merit further exploration include purchasing computing services from federal agencies (thus increasing access beyond that driven by direct mission interests) or by making arrangements with commercial services (rather than more expensive purchases by individual researchers)."*

As a solution this sort of outsourcing presents four substantial risks to the accomplishment of the NSF's objectives as a national funding agency for basic research.

Risk 1. Other agencies have mandates that are not aligned with the mission of the NSF and may actually be at odds with it. Over time it will be difficult to prevent the erosion of resources available to the national science community in response to the mission pressures experienced by the hosting agency, and it will be difficult to correct this erosion once it has begun.

The NSF is the only U.S. federal agency with a mandate to support *all* non-medical fields of research. Over time the agency or agencies that serve as computing infrastructure providers for the NSF's community of researchers may begin to favor their own internal needs over the needs of the external community of users. This may happen without intent or malice over decades as memories of past commitments and the spirit of government cooperation fades. It may happen in response to a perceived crisis (for example, a national security event) that begins as an exceptional redirection of resources that, through repetition over time, becomes routine. Or it may happen as the mismatch between the mission of the sourcing organization and the mission of the NSF research community becomes too difficult to overcome. Both the Department of Energy and the Department of Defense support missions that are vital to the security of the nation, and for which there is a huge gap between computational requirements and resources deployed. In FY15, for example, the DoD is expected to satisfy less than 1/3 of its validated requirements for access to supercomputers and HPC expertise, and both agencies report full utilization of their existing complement of supercomputers.

Whatever the cause, it is difficult to imagine the capacity available to the NSF community of scientists not eroding over time.

Further, both agencies have substantially heightened security postures, and the willingness to tolerate security risk is decreasing in each over time in each. The NSF research population is very diverse, supporting citizens of a variety of nationalities in highly dynamic environments. Existing security requirements in the DoD and DoE require careful assessment when foreign nationals must interact with advanced computing resources. This assessment takes resources and the accompanying agency checks and investigations take time, both of which drain resources that would otherwise fund research.

Risk 2. Current commercial cloud business models are not well-aligned with the needs of either the data- or compute-intensive communities.

Cloud computing is mentioned several times in the interim report. The report correctly points out that the common cloud architecture is not an appropriate replacement for purpose-built supercomputers with high bandwidth, low latency interconnects that are required to enable modern, complex, multi-physics applications. The report also posits that the cloud architecture and service delivery model may be a good match to the needs of the data-intensive computing community. That this is the case is less clear; more research is needed to establish an appropriate software and hardware architecture for data-intensive computing and to establish that the objectives of compute- and data-intensive computing cannot be

satisfied in a merged architecture. If it can, then there is likely something to be gained from adopting **components** of the cloud-delivery model for use in centers that support data-intensive computing.

But there is a deeper question around the costs of servicing data-intensive computing needs in the commercial cloud ecosystem.

A common feature of cloud business models is that the computation is much less than the costs of moving data in and out for processing, or storing with the cloud provider. In fact, a study completed in 2013 by the DoD HPC Modernization Program showed that that program could provide cycles at less raw cost than Amazon's Elastic Compute Cloud solution (based on pricing at that time) when comparing costs **only** on a per FLOPs basis. This despite the fact that DoD HPCMP machines are designed with more expensive special purpose interconnects than available on ECC. When the actual costs of data storage and movement were included, the value proposition tipped overwhelmingly in the government's favor.

The point here is that commercial cloud providers are optimized to provide a good value for entities that have computational demands insufficient to fully utilize an in-house computing resource. The NSF's current advanced computing infrastructure is fully utilized and, as the report identifies, requirements exceed capacity in the system. Engaging a commercial provider would result in that provider having to acquire new systems and dedicate them to NSF needs, which is precisely the model the NSF uses now with universities playing the role of commercial providers. Thus there is no economic advantage to this approach, and the NSF would at the same time invite the simultaneous disadvantages that would accompany separation of computing from the university environment in which it is conducted (with its common cultural basis) and place the integrity of the advanced computing workforce at risk (see below).

Risk 3. Reducing the size and diversity of the advanced computing provider ecosystem may impair the Nation's ability to sustain a critical mass of experienced computing designers and operators, ultimately reducing the competitiveness of US research on the global stage.

As the study committee itself points out, advanced computing capability is key to future growth and stability of the US economy and our national security. Advanced computing by itself is necessary but not sufficient to meet our national needs — supercomputers won't turn on and run by themselves, full service architectures that support complex scientific workflows don't develop spontaneously, and crafting applications that use these systems effectively is a highly specialized activity. Given that effective utilization of advanced computing is critical to our global strategic position, it necessarily follows that identifying, developing, and sustaining a broad-based advanced computing workforce must be a driver in US plans for a future advanced computing infrastructure.

It is also clear that the majority of the burden for developing the advanced computing workforce must fall to federal and academic institutions. At present there is no large scale commercial environment in which a future workforce can develop that is capable of fielding compute-intensive infrastructure needed to support national security and research priorities, and the commercial data-intensive computing community is still developing. Further, much of the commercial data-intensive computing community is driven by requirements that are fundamentally different in terms of time horizons for results, application space, and (possibly) algorithms themselves than the research and national security data-intensive computing communities, resulting in a workforce that will have a general background that is not directly relevant to national science needs.

Similarly, the architectures, algorithms, security environment, and cost drivers within DoD and DoE — all tailored to the unique requirements of those environments — will produce a workforce with different skills than those needed to design and operate advanced computing infrastructure suitable to support large-scale basic science research.

The NSF is the only organization currently positioned to provide for the development of a workforce with the necessary skills and understanding of the basic research context to design, operate, and maintain the specific advanced computing infrastructure needed to most effectively support science in the US. The NSF must maintain its investment in advanced computing infrastructure in order to satisfy its obligation

to ensure the development of this workforce.

Risk 4. Outsourcing places NSF at risk for abandoning responsibility for complete support of national science agenda.

Implicit in the interim report released by the committee is that the agency will face budget shortfalls or reductions that will drive it to reduce its investment in computing, presumably to maintain a threshold level of funding for science research.To present a choice between investing limited funds in research or investing in advanced computing to support research is to present a **false choice**.

Computing is a peer in the research enterprise with theory and experiment, and the study committee posits the emergence of a fourth pillar in that enterprise with the emergence of data-driven science. It is not credible to propose that a broadly-based science research portfolio supporting national needs can be successful without a sufficient investment in advanced computing to enable and underpin that research. The NSF must fund advanced computing in proportion with its more direct research investments in order for that research to succeed.

To propose otherwise is analogous to proposing to fund the purchase of bullets for an Army and not funding the purchase of guns to fire them.