An Integrative, Cross-Foundation Cyberinfrastructure for Science & Engineering Research

"Multiple Authors in the Community[1] "

Concept: In an era where science and engineering communities are rapidly becoming digital and computational and data science is ever more critical to progress, the nation needs a strong, predictable path for supporting computation- and data-intensive research across a broad spectrum of research efforts (e.g., individuals, groups, communities, and projects, including MREFCs), requiring diverse capabilities (e.g., HPC, HTC, data and cloud) and magnitudes (current HPC facilities are saturated, while the nation lacks national data services in an exponentially growing big data era). With the need for computing and data resources outstripping current capacities and no clear long-term plan in place, research communities are both underserved and uncertain of the availability of the national resources needed to move their research forward.

To address the challenges facing our nation, research requires more deeply integrated approaches across theory, experiment and/or observation, data and simulation science, and disciplinary boundaries. But uncertainties in the availability of national computing and data resources lead to fragmentation, not integration, as individual major research projects prefer to develop their own solutions as they cannot plan on synergies with other national resources that may not be there in the future. Yet it is also substantially more expensive as research groups reinvent individual cyberinfrastructure (CI) solutions over and over again. What is needed is a long term commitment to an integrated national computing and data infrastructure that recognizes the importance of these resources for scientific and engineering progress as well as the importance of the technical culture, expertise and staff that can take years to develop around major computing and experimental facilities.

We propose a new approach that combines a modified version of NSF's MREFC process for developing and operating major, long-term research facilities, with a national commitment to provide the major computing and data resources as well as the critical services that connect them.[2] These are needed by all of NSF's Computational and Data-enabled Science and Engineering (CDS&E) communities, and numerous MREFC projects. This approach could include large-scale computing facilities to support the most compute- and data-intensive research, coupled with a small number of mid-range facilities serving the diverse needs of multiple research communities, as well as national-scale data repositories for multiple communities, high-speed networks to connect to them, and additional services that integrate these capabilities and services, including integration with the growing environments on university campuses.

[1] This already has input from about two dozen people across the community. If a broader group of converges on an improved document, we will of course put our names on it! This is just a starting point to argue over!

[2] Note the MREFC vehicle is used for multi-agency initiatives, e.g., between NSF and DOE, where MREFC and CD processes have to be intertwined. So this could in principle provide a mechanism for, say, NSF-NIH-DOE cooperation in providing a national computing and data infrastructure

Roughly speaking, the proposed National Compute and Data Infrastructure (NCDI) would combine some aspects of the current advanced computing (Track 1, Track 2, and XSEDE) investments, coupled with new data and networking services, into a major national investment that would integrate these capabilities distributed across a number of sites.

Discussion: The NCDI MREFC is envisioned as a critical outcome of earlier CIF21 programs that will unite four major thrusts into an integrated framework that would benefit researchers, ranging from those working within the long tail of science to those at the extreme range in fields all across NSF, from individual university campuses to the largest instruments that are being internationally developed and deployed. The four thrusts are CI programs that are part of CIF21 relating to (a) computational, (b) data, and (c) networking infrastructure, with (d) digital services that connect to those activities on campuses, in research communities and MREFC projects. The science-driven community efforts in each of these areas are maturing and are increasingly interdependent on CI.

The NCDI would provide unique opportunities for researchers to define and use a highly responsive full-scale CI that reaches from instruments to computational, data stewardship and analysis capabilities to publications for long-term highly innovative research programs. The development of community standards and approaches during the MREFC design process—an ongoing activity in the envisioned modified MREFC process—will accelerate research programs that will come to fruition during the construction and operation of the full scale infrastructure.

With an NCDI in place, or at least planned, individual MREFC projects, each of whom is currently developing their own (often duplicative and typically less capable) CI, would then be able to leverage long term, stable environments to deploy their needed computing, data stewardship and analysis facilities, more deeply integrating the services of NCDI with other MREFCs, enabling funding to be focused on scientific and engineering issues and better supporting complex problem solving for disciplinary, multidisciplinary and interdisciplinary research. Where appropriate, other MREFC projects would be able to plan longer-term investments that explicitly build on the NDCI. This would more naturally enable large-scale data services to be co-located with large scale computing facilities, which are needed in an era of big data, while better serving communities by leveraging commonly needed infrastructure, staff expertise, and services.

The MREFC process, as we know it, would need to be modified for such a project. As the component technologies evolve with characteristic times much shorter than the standard MREFC cycle, a set of inter-related design-build/integrate-deploy processes must be running concurrently with the operation of facilities and services to provide a robust, reliable and secure NCDI to provide the production needs of the nation's CDS&E community and upon which other MREFCs can also build. Long-term infrastructure is a continuing effort, longer than normal NSF Research and Related Activities (R&RA) awards. In order to ensure effectiveness of resources, NSF could consider options that would allow for a series of continuing awards, within the MREFC process, based on merit, successful operation, and ability to meet the various communities' needs. For example, previous NSF programs have used rolling 5-year awards, with "continue," "phase out" or "re-compete management" decisions through periodic reviews. Processes for assessing the NCDI MREFC and the evolution of community computing needs could be a part of these review processes. Additional activities, such as campus bridging, algorithm development, software institutes, related disciplinary research, and so on could be supported separately, as they are at present, but they would be more effective with a structure such as NCDI in place that could be leveraged.

In this model, NCDI construction funds would come from the MREFC budget, with a profile that is estimated to be in the range of $60-75M per year[3] (of order $750M over a decade). Assuming that the MREFC would be managed by ACI, the operating funds would come from CISE, potentially with some contributions from other directorates based on their specific needs. This would further allow ACI (and other NSF directorate programs) to better support its portfolio of services, software, development, CDS&E programs, and so on.

Advantages:

- The NCDI will provide the research community with a clear roadmap for planning a broad range of CDS&E research activities, from individual investigations to major facility projects;
- The NCDI will provide a means to coalesce community engagement around computing and data infrastructure;
- The NCDI will provide an integrated approach to data-intensive (Big Data) science. Big Data services, e.g., astronomy, environmental science, data pipelines, etc., will be more naturally and effectively supported and can be co-located with simulation services at major computing sites. This will drive a convergence of big data and big compute activities at common sites, which will be required as data volumes grow;
- The long-term stability of the NCDI will enable the retention and building of staff and community experience that is so critical to success in the rapidly evolving computation- and data-sciences. This is especially important for emerging data services, both at national and campus levels, that will need a framework that lives long past any individual grants;
- The NCDI will provide a national structure that other MREFC projects can depend on being in place, enabling them to plan for and take advantage of synergies provided by leveraging such an infrastructure leading to significant cost efficiencies;
- Exploiting such synergies, NCDI could reduce the cost of providing the major compute and data resources and services needed for scientific progress, resulting in savings to both CISE/ACI and NSF;
- The NCDI will enable NSF's MREFC projects to focus more on science and engineering issues since they will not have to completely reinvent their own CI and can leverage the investments in the NCDI for their own purposes;
- The NCDI will enable NSF projects to more fully exploit synergies that will be needed to better support interdisciplinary research trends (e.g., multi-messenger astronomy, which will require unified or coordinated computing and data services from multiple MREFCs and communities, such as LIGO, IceCube, LSST, DES, computational astrophysics, etc.);

[3] This is probably too low…a careful analysis needs to be done.

- The NCDI will support interagency cooperation around MREFC projects (e.g., with DOE and/or NIH) and can be leveraged for deeper cooperation.

Disadvantages:

- The MREFC process has been designed around a model of building a single instrument to be used by a single community for a decade or more, and has a very heavy, very long process of community engagement, conceptual design, preliminary design, final design, and NSB approval. This process would require revision to accommodate the fast paced evolution of computing technologies; a process already being developed by the XSEDE project.
- *Significant* urgent work will be needed among all of the stakeholders in the NCDI (research communities, NSF, OSTP and others). What often takes more than a decade would ideally be compressed into just a few years, and would have to be done at many levels. This would require dedication of a number of people in the community over a sustained period of time to define and implement the inter-related, concurrent processes necessary at all of these levels.

If this MREFC approach is to be developed and made relevant to any existing projects, a conceptual plan needs to be developed as soon as possible. This would need to involve numerous stakeholders and communities, including existing MREFC projects, to explore how such an environment would operate. It is likely that such an approach would lead to a significant restructuring of existing CI centers, university stakeholders, etc. Many assumptions above would need to be quantitatively validated, and additional issues would need to be considered to see if this is a viable approach. In particular, the need for and benefits of National CI Infrastructure would need to considered in depth, with not only its scope and cost considered, and its potential cost savings, but the science case, the impact on national competitiveness, the impact of training and workforce, the role of NSF with respect to other agencies, and more. This concept paper merely aims to raise awareness of the possibility of such an approach and to highlight issues that would need to be considered in exploring its viability.