Response to NSF 20-015, Dear Colleague Letter: Request for Information on Data-Focused Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

Reference ID: 11227075060_Chakravorty

Reference ID: 11227075060_Chakravorty
Submission Date and Time: 12/16/2019 4:56:17 PM

This contribution was submitted to the National Science Foundation in response to a Request for Information, https://www.nsf.gov/pubs/2020/nsf20015/nsf20015.jsp. Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).

Consent Statement: “I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publicly available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode).”

Consent answer: I consent to NSF’s use and display of the submitted information.

Author Names & Affiliations

Submitting author: Dhruva Chakravorty - High Performance Research Computing at Texas A&M University

Additional authors: High Performance Research Computing at Texas A&M University

Contact Email Address (for NSF use only): (hidden)

Research domain(s), discipline(s)/sub-discipline(s)

Several science domains

Title of Response

A CI Perspective on Challenges Faced by Data-Intensive Researchers Across Science and Engineering Domains

Abstract

Texas A&M University High Performance Research Computing (HPRC) serves the research computing needs of all Texas A&M system universities and state agencies. Texas A&M HPRC provides thousands of
researchers across several fields of science and engineering with access to cutting-edge compute environments, research facilitation, as well as commercial and freeware software packages. In addition, HPRC offers a number of informal training events on topics related to computing at Texas A&M university that are offered in in-person and online modalities. HPRC resources also support formal education at Texas A&M. Data-intensive research is a core pillar of a university’s research mission. It involves faculty, researchers, students, and administrators. In this response, we have included our vision for cross-cutting data enabled research environments that promote accelerated advances by integrating and optimizing technologies and approaches. As large scale computing providers, we view CI to include compute, data science technologies and the rich humans who provide the support and glue. This document summarizes challenges to data-intensive research, and urges us to re-imagine our support for this ecosystem by expanding continuing current efforts in computing, networking, storage, gateways and workforce development.

**Question 1 (maximum 400 words) – Data-Intensive Research Question(s) and Challenge(s).** Describe current or emerging data-intensive/data-driven S&E research challenge(s), providing context in terms of recent research activities and standing questions in the field. NSF is particularly interested in cross-disciplinary challenges that will drive requirements for cross-disciplinary and disciplinary-agnostic data-related CI.

Given the exponential growth of data, ensuring data-quality and transforming data into an effective aid for scientific decision making are major issues for research. Poor data-quality leads to lower utilization, lack of efficiency, higher costs, and occasionally might even lead to erroneous decisions. The application of data-driven analytics across multiple domains can significantly impact thinking in a cross-cutting manner. Some of the challenges we foresee from a CI-perspective are: 1. Various fields have taken independent approaches to adopting data science methods. Adoption of FAIR and FEAT standards has been encouraged in the life science and AI-rich domains, but are otherwise lagging in multi-disciplinary sciences. We need to identify ways for researchers to incorporate them via our CI resources. 2. Support for reproducibility in science, and adoption of open science practices is gaining ground. There are, however, times when research data can not be shared, particularly with partners in other nation states. Competing mandates across various Federal agencies could be restructured to improve S&E workflows. 3. Significant challenges in data-intensive research come from the enormous volume of data that needs to be rapidly verified. Indeed, ML algorithms are data hungry, and a reliable model requires an enormous amount of validated data for training. Research workflows rely on the data being accurate and may be compromised by a single instance of incorrectly entered data. There is an urgent need for data security, preservation, curation and sharing resources, including supporting services and capabilities. 4. Most HPC resources are not optimized for emerging technologies such as IoT, AI/ML and edge-computing. These workflows differ from traditional HPC workloads. Data needs to be persistent for AI training – this is not the case in current HPC workloads. We need to rethink how data is stored, transferred, and analyzed. They need access to large shared repositories, fast connections to the edge, storage elements that support unstructured data, an easily-accessible analytics layer, and a rapid publication mechanism. These workflows need burst capacity to larger resources when demand spikes.
5. As AI/ML and IoT become ubiquitous in research, availability and integration of CI capabilities via networking and software will be critical to continue the process of innovation and discovery. The libraries and other parts of the code development environment are not in place to support the various computing architectures in place. In addition, there are silos of expertise across Federal agencies that could pool resources.

Question 2 (maximum 600 words) – Data-Oriented CI Needed to Address the Research Question(s) and Challenge(s). Considering the end-to-end scientific data-to-discovery (workflow) challenges, describe any limitations or absence of existing data-related CI capabilities and services, and/or specific technical and capacity advancements needed in data-related and other CI (e.g., advanced computing, data services, software infrastructure, applications, networking, cybersecurity) that must be addressed to accomplish the research question(s) and challenge(s) identified in Question 1. If possible, please also consider the required end-to-end structural, functional and performance characteristics for such CI services and capabilities. For instance, how can they respond to high levels of data heterogeneity, data integration and interoperability? To what degree can/should they be cross-disciplinary and domain-agnostic? What is required to promote ease of data discovery, publishing and access and delivery?

There has been a steady uptick in the amount of data that is being collected, generated, and analyzed. This has been facilitated by the proliferation of a common set of research-enabling technologies including open source software, Jupyter Hub, R, and AI/ML across many areas of science and engineering. This in turn has led to complex workflows that now face challenges in working with different data standards. Questions remain how, when, and where data can be stored, retrieved and reused. Here, we discuss how CI can help address some of these challenges: 1. To facilitate the utilization of current computing resources, research workflows need to be resource agnostic. Software and workflows that are heavily used across research facilities should be standardized such that the end user does not need to know the hardware specifics of the computing facility to run their job. 2. CI is not a one-stop answer for data sciences. Researchers need additional facilities that generate or support data. These include libraries, data repositories, and instrumentation hubs, to name a few. We need networks and software to integrate these facilities with existing campus or regional computing resources. 3. With a focus on learning, the inclusion of education as a core component of the envisioned NSF AI institutes is a positive development. While AI/ML remain the technologies of interest at this point, history has shown us that these technologies will be supplanted by others. We need to identify the incremental steps toward a future when other technologies such as quantum computing replace AI/ML. 4. Technical considerations: a. Cybersecurity implementations and concerns are driven by a domain’s specific needs. An overarching cybersecurity framework that speeds-up workflows while addressing the various data-needs and data-types will strengthen multidisciplinary efforts. Identify key design factors to encourage wide-spread adoption of approaches for data-validation across multiple fields of science and engineering. b. Researchers need a balance of security and usability. There is a need for new mechanisms to rapidly approve new data management and transfer plans by Federal agencies, particularly for secure data. c. File systems that support data science (as opposed to HPC) across multiple facilities are needed. d. Investments in workflow optimization tools (ex. CWL,
containers) need to be prioritized. e. Language development for portability will be important. Choices will have to be made about the balance of optimization vs portability. This may not be the best, but sufficient for a pilot project. 5. Human and social considerations play an important role here: a. Access - who has access, who needs access and how do we give them better access? As funding and scope together force research projects to become more international, how do we simultaneously maintain control over our IP while promoting Open Science? b. Supporting Scientists and Research Staff: To support the needs of the community, we need a trained cadre of researchers who are familiar with both data science research and large scale computing technologies. This, however, presents considerable challenges. The twin forces of high-demand and lack of supply have made it hard to hire experienced data-science and computing personal. In addition to promoting learning and training programs, the NSF could incentivize universities to incorporate career-paths for non-tenure CI personnel. Such career paths exist for librarians. c. Promoting Regional Expertise: We find regional groups, like the NSF CC* Cyberteams to have a profound effect on institutions in a region. This powerful mechanism has boots on the ground, who understand the local researchers needs and constraints. It is a necessary complement to “Centers of Excellence” cater to the broader needs of a nationwide community.

Question 3 (maximum 300 words) – Other considerations. Please discuss any other relevant aspects, such as organization, processes, learning and workforce development, access and sustainability, that need to be addressed; or any other issues more generally that NSF should consider.

The demand for CI services is growing faster than the available resources. Disruptive technologies such as new accelerators and processors, together with the edge, cloud, and IoT challenge existing computing paradigm. Here, we list a few other considerations: 1. Multi-institutional and regional cooperation may be the way forward. For example, the recent NET+ agreement brokered by Internet 2 helps address some of the cost-related challenges that the public (commercial) cloud presents, helping reduce the gap between larger and smaller institutions. 2. While we realize the importance of networking and the cloud in research computing, we are unclear as to the extent they should be prioritized in an institution’s research portfolio. While the current discussion is centered on cost, we need to start considering how these resources can be used to augment the capabilities of all on-premise resources. For example, how do we configure networks such that local resources and the cloud work hand-in-hand to support instruments at the edge? 3. The current regulatory and computing landscape favor a “Bring your own Compute and Storage” model for researchers, much like what the cloud facilitates. While the cost-structure of the cloud makes it prohibitive for some researchers, this may well be the future. 4. There is considerable risk that CI may be as expensive as health care for employers. LWD steps need to be taken to ensure that there is a ready workforce. REUs and curricular support are needed. We need to find ways to get students to experience data-science research, and advance the participant’s disciplinary knowledge by working on an unsolved research problem. 5. Inclusivity: Smaller institutions and community colleges find it difficult to engage at the level of R1 schools on national efforts, making it difficult to participate in larger research efforts creating classes of haves and have-nots.
Response to NSF 20-015, Dear Colleague Letter: Request for Information on Data-Focused
Cyberinfrastructure Needed to Support Future Data-Intensive Science and Engineering Research

Reference ID: 11227075060_Chakravorty

-- End Submission --