

Reference ID: 11226867586_Peterka

Reference ID: 11226867586_Peterka

Submission Date and Time: 12/16/2019 3:33:08 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: Tom Peterka - Argonne National Laboratory (ANL)

Additional authors: Bogdan Nicolae (ANL), Franck Cappello (ANL), Sheng Di (ANL), Hanqi Guo (ANL)

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

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

high-performance computing; in situ data management; storage and I/O; data reduction; data preservation; feature tracking; uncertainty quantification; scientific workflows

Title of Response

Data Cyberinfrastructure for Scientific Discovery

Abstract

Scientific discovery in the 21st century requires a data cyberinfrastructure (DCI)---an ecosystem of computing, networking, software, data, and people in the pursuit of scientific discovery---comprising a wide range of data services across a continuum of computing and data management. This continuum

will span a variety of platforms and will support computational modeling, scientific observations, experiments, and data analytics. Numerous challenges need to be solved in order to make the DCI a reality. High-level data characteristics---semantics, intent, and uncertainty---must be supported. Machine efficiency and human productivity must be streamlined. Scientific validity must also be facilitated: guaranteed reproducibility and quantified uncertainty are essential if results are to be trusted by peers, policy makers, and the public. Exciting opportunities for new computing research include (i) dynamic data management in the DCI, (ii) intelligent data storage and reduction, (iii) high-performance metadata preservation and manipulation, and (iv) understanding of data uncertainty. Potential benefits include accelerating machine utilization and human productivity; focusing on science goals rather than mundane data management tasks; training and educating the next generation of scientists and engineers; and regaining trust in scientific results, the cornerstone of open science.

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.

Data from measurements (sensors in the field, engineering probes, physics instruments, social data) integrated with data from computational models can enable understanding, refine designs, enhance precision, and improve validation. The ever increasing volume, velocity, and diversity of these data, however, could reduce the potential for scientific and engineering discoveries without advances in the DCI that preserve data provenance, dynamically adapt to user's needs, intelligently mitigate data volume and velocity, and bound the uncertainty propagated in science workflows. The following desired cross-disciplinary capabilities are agnostic to any one discipline. Datasets are becoming increasingly heterogeneous from every perspective: data source, size, content, and access pattern. The abundance of such datasets leads to an explosion of metadata needed to describe them. Under such circumstances, scalable and flexible data management approaches are needed to simultaneously handle both data and metadata and provide advanced capabilities such as long-term preservation, tracking, search, and reuse. Managing data and metadata will require a high degree of autonomous operation, while still affording human interaction for exploratory investigation. Unpredictable cyberinfrastructure behavior can be caused by system variability, transient data generation, and dynamic data-driven processing requirements. In spite of these variabilities, users still require predictable, resilient, and repeatable performance of their workflows, especially low-latency and time-critical ones. Elastic and adaptive bridging of the various parts of the data infrastructure---experimental facilities, HPC facilities, instruments and sensors, cloud providers, and networks---is needed. One example of adaptivity is autotuning of data storage and reduction. While particular settings for specific applications can be found, such static settings cannot adapt to diverse simulations, complex ensembles, or dynamic data changes. The realization that data management needs to adapt to the data, users, and computing environment raises several open questions: How to develop data reduction techniques while the diversity of these data sources is increasing dramatically? How to reduce data from multiple sources in

such a way that these reduced data can be combined to enable scientific and engineering discoveries? How to understand the uncertainty and potential bias that may be introduced as a result? Quantifying uncertainties in data and in high-level features must also be addressed. The analysis and understanding of data may be biased because of the inherent uncertainties hidden in the data or introduced by processing, such as compression. Uncertainty and bias hinder our ability to validate scientific data and to make sound data-driven decisions.

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?

Infrastructure is needed to create novel data management solutions that facilitate streaming and online processing of multirun, multiscale, and multisource data in support of science. Potential research directions include dynamic control of the DCI and the applications running in it, data storage and reduction approaches, metadata preservation methods, and understanding of the uncertainty and error introduced by the above methodologies. **Dynamic data management:** Research is needed to develop adaptive, autonomous control systems for the DCI and for data-intensive tasks and workflows supported by it. How to combine autonomic control with human exploration is an open question. Required research includes (i) capturing system state (system, data, and algorithmic parameters), (ii) modeling system dynamics (using reduced-order approximations to find the most critical or sensitive parameters), and (iii) designing low-latency controllers (using linear approximations or machine-learned genetic or deep neural networks). **Intelligent data reduction and storage:** Storage solutions must support researchers' conceptual models of scientific data while optimizing data performance and productivity. One example is intelligent data reduction. The DCI will need to provide generalizable and consistent capabilities to compose, deploy, operate, and control data reduction to enable the combination of extreme volume, velocity, and diversity from multiple data sources in the computing continuum. Developing such capabilities requires research in (i) new adaptable lossy reduction techniques responsive to a large variety of needs and constraints (adapting to dynamic data changes, diverse parameter settings, various user requirements), (ii) new capabilities to deploy data reduction from the edge to the users, (iii) new capability to coordinate data reduction, data movement, and storage to enable time-coherent combinations, checkpoint/restart, data analysis, and data curation, and (iv) new understanding of the composability of data from multiple sources with traceability to inform, record, and verify how data have been managed. **Data and metadata preservation:** For decades, data services available on supercomputers have seen a relatively static evolution: they are dominated by parallel file

systems and object stores that offer low-level abstractions operating with loosely coupled sequences of bytes (files and key-value pairs) and basic operations (read/write and put/get). However, such low-level abstractions are inherently unable to convey enough information about the nature of data, metadata, and relationships between them. Lacking such information, current techniques treat all data and metadata equally, which is a fundamental limitation in achieving scalability and flexibility considering the irregular and unstructured access patterns triggered by heterogeneity. To address these gaps, research in several directions is necessary: (i) capturing dependencies between data and metadata with minimal performance and space overhead, (ii) using such dependencies to optimize the data layout on heterogeneous storage, and (iii) exposing such dependencies to users in order to improve their understanding of the data and their ability to organize and archive them. Understanding data uncertainty: Uncertainty quantification has the potential to considerably reduce the amount of data to store by keeping only the most salient and robust features. To achieve this goal, the DCI will need to address three research challenges. (i) Infrastructure is needed to create novel data management, analysis, and visualization paradigms to enable streaming and online processing of multirun, multiscale, and multisource data in scientific workflows. (ii) Techniques, mathematical methods, and surrogate models are needed to quantify, model, and represent uncertainties of features for future scientific applications. (iii) The ability to reduce scientific data based on uncertainties of features needs to be developed, and quantifying feature uncertainty is needed for overall understanding of robustness, confidence, and importance of scientific conclusions derived from data.

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.

Solving grand challenges demands a multidisciplinary effort associating researchers and engineers from domain disciplines, computer science, and applied mathematics as well as operators of the DCI and the facilities connected to them. A co-design effort with application groups is needed to define high-level data requirements, emphasize the limitations of existing data services from a system-level perspective, and explore the opportunities of new data models and methods. In particular, collaborations between DOE-funded national laboratories and NSF-funded academia are critical to the success of the future DCI and more broadly to U.S. leadership in international scientific research. National laboratories sponsored by DOE operate national-scale facilities and instruments and have some of the world's most advanced data-intensive applications. No less important are universities sponsored by NSF: in addition to supporting large-scale infrastructure and experiments, universities develop cutting-edge research and are critical for R&D workforce development. Only by working in partnership can we solve the nation's most pressing data cyberinfrastructure challenges and realize an ecosystem of computing, networking, software, data, and people in the pursuit of scientific discovery.

-- End Submission --

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: «Respondent_ID»_«Primary_Last»
