Office of Science
Financial Assistance
Funding Opportunity Announcement
DE-FOA-0000256

Scientific Data Management and Analysis at Extreme Scale

SUMMARY:

The Office of Advanced Scientific Computing Research (ASCR) of the Office of Science (SC), U.S. Department of Energy (DOE), hereby announces its interest in receiving applications from interdisciplinary teams of Computer Science/Applied Mathematics/Statistics/Computational Science researchers in the areas of Scientific Data Management and Analysis at Extreme Scale. Multi Institutional proposals with cohesive emphasis on transformational discoveries that address key challenges in analysis and management of scientific data at extreme scale are encouraged. Partnerships among academic institutions, National Labs, and industry are strongly encouraged.

Science has shifted from data scarcity to an overwhelming abundance of data, as simulations and experiments generate many petabytes of data, with some sciences facing exabytes of data near term. For example, a recent report states that climate model data are growing faster than the data set size for any other scientific discipline, with collections of hundreds of exabytes expected by 2020 (***Challenges in Climate Change Science and the Role of Computing at the Extreme Scale***, http://extremecomputing.labworks.org/climate/report.stm, and the Large Hadron Collider (LHC) is expected to produce roughly 15 petabytes of data annually over its estimated 15 year lifespan. (http://public.web.cern.ch/Public/en/LHC/Computing-en.html)

The value of scientific data is realized only when data are effectively analyzed and results are presented to the science community, policy makers, and the public in an understandable way. The challenges of analyzing massive scientific data sets are compounded by data complexity that results from heterogeneous methods and devices for data generation and capture and the inherently multi-scale, multi-physics nature of many sciences, resulting in data with hundreds of attributes or dimensions and spanning multiple spatial and temporal scales. The combination of massive scale and complexity is such that high performance computers will be needed to analyze data, as well as to generate it through modeling and simulation.

This Funding Opportunity Announcement calls for innovative basic research in computer science for management and analysis of extreme-scale scientific data in the context of petascale computers and/or exascale computers with heterogeneous multi-core architectures. The activities supported by this Announcement may be a combination of basic research, creation of algorithms for advanced architectures, and development of usable data management and analysis tools for scientific discovery. Partnerships among universities, National Laboratories, and industry are strongly encouraged.

APPLICATION DUE DATE: March 18, 2010, 11:59 p.m. Eastern Time

Formal applications submitted in response to this FOA must be received by March 18, 2010, 11:59 p.m. Eastern time, to permit timely consideration of awards. **APPLICATIONS RECEIVED AFTER THE DEADLINE WILL NOT BE REVIEWED OR CONSIDERED FOR AWARD.**

IMPORTANT SUBMISSION INFORMATION:

The full text of the Funding Opportunity Announcement (FOA) is located on FedConnect. Instructions for completing the Grant Application Package are contained in the full text of the FOA which can be obtained at: https://www.fedconnect.net/FedConnect/?doc=DE-FOA-0000256&agency=DOE. To search for the FOA in FedConnect click on "Search Public Opportunities". Under "Search Criteria", select "Advanced Options", enter a portion of the title "Scientific Data Management and Analysis at Extreme Scale", then click on "Search". Once the screen comes up, locate the appropriate Announcement.

In order to be considered for award, Applicants must follow the instructions contained in the Funding Opportunity Announcement.

WHERE TO SUBMIT: Applications must be submitted through Grants.gov to be considered for award.

You cannot submit an application through Grants.gov unless you are registered. Please read the registration requirements carefully and start the process immediately. Remember you have to update your CCR registration annually. If you have any questions about your registration, you should contact the Grants.gov Helpdesk at 1-800-518-4726 to verify that you are still registered in Grants.gov.

Registration Requirements: There are several one-time actions you must complete in order to submit an application through Grants.gov (e.g., obtain a Dun and Bradstreet Data Universal Numbering System (DUNS) number, register with the Central Contract Registry (CCR), register with the credential provider, and register with Grants.gov). See http://www.grants.gov/GetStarted. Use the Grants.gov Organization Registration Checklist at http://www.grants.gov/assets/OrganizationRegCheck.pdf to guide you through the process. Designating an E-Business Point of Contact (EBiz POC) and obtaining a special password called an MPIN are important steps in the CCR registration process. Applicants, who are not registered with CCR and Grants.gov, should allow at least 21 days to complete these requirements. It is suggested that the process be started as soon as possible.

IMPORTANT NOTICE TO POTENTIAL APPLICANTS:
When you have completed the process, you should call the Grants.gov Helpdesk at 1-800-518-4726 to verify that you have completed the final step (i.e. Grants.gov registration).

Questions: Questions relating to the registration process, system requirements, how an application form works, or the submittal process must be directed to Grants.gov at

1-800-518-4726 or support@grants.gov. Part VII of the FOA explains how to submit other questions to the Department of Energy (DOE).

All applications should be in a single PDF file.

GENERAL INQUIRIES ABOUT THIS FOA SHOULD BE DIRECTED TO:

Technical/Scientific Program Contact:

Program Manager: Dr. Lucy Nowell
Office of Advanced Scientific Computing Research
Office of Science
U. S. Department of Energy
Phone: 301-903-3191
E-Mail: Lucy.Nowell@science.doe.gov

Program Manager: Dr. Daniel Hitchcock
Office of Advanced Scientific Computing Research
Office of Science
U. S. Department of Energy
Phone: 301-903-6767
E-Mail: Daniel.Hitchcock@science.doe.gov

SUPPLEMENTARY INFORMATION:

As the ***International Exascale Project Roadmap*** *(http://www.exascale.org) notes, "the potential impact of Exascale computing will be measured not just in the power it can provide for simulations but also in the capabilities it provides for managing and making sense of the data produced. ... Thus, managing scientific data has been identified by the scientific community as one of the most important emerging needs because of the sheer volume and increasing complexity of data."*

Exascale computer systems will be comprised of as many as a billion cores. Such systems will be capable of 10 billion-way concurrency in simultaneous operations. Industry reports indicate that data movement will be the limiting factor for exascale systems, rather than processors and computational operations, especially when power constraints are considered. At the same time, memory per core is expected to decline sharply for exaflop systems, and the performance of storage systems continues to lag far behind. Multi-level storage architectures that span multiple types of hardware are anticipated and will require new approaches to run-time data management and analysis. Research is needed to develop abstractions that reduce or eliminate the need for scientists to learn many details about computer hardware, operating systems, storage architectures and file management systems in order to create codes that are essential for their research.

Sharing, re-use, and re-purposing of scientific data and integration of data from multiple simulations and multiple disciplines are required to address mission-critical challenges in

complex systems. Analysis of massive heterogeneous data sets are required, for example, for understanding the impact of stockpile decay on containment materials over decades or understanding the causes and potential impacts of climate change. Such analyses may engage hundreds to thousands of scientists at multiple locations and from multiple disciplines.

Integration and/or comparison of data from simulations and observations are necessary for model validation, as well as requiring analysis in their own right. The scientific community is also calling for more interactivity with both simulations and data analysis applications to better support exploratory data analysis and so that effective steering of the computational process can occur, especially to assist with fault management. These requirements suggest the need for enhanced metadata generation/capture, including provenance representation and capture, and contextually informed semantic data integration and analysis, as well as enhanced support for collaborative or cooperative data analysis by teams of researchers who may or may not be co-located.

Different knowledge representation schemes enable different types of automated analysis, and integration of data from multiple sources and/or disciplines is facilitated or hindered by knowledge representation choices. Knowledge representation and machine reasoning research are needed for a variety of problems, especially including representation of and reasoning about uncertainty and sources of uncertainty in the capture/generation and analysis of scientific data. Also needed are knowledge representation methods that support automated analysis of large scientific data sets that include tensor flow fields, including vectors, such as electromagnetic fields, elastic and plastic strain in materials, viscosity, and velocity fields.

Visual analysis systems that enable interaction between the scientist users, the data analysis system, and the data are critical for supporting scientific discovery and understanding, as well as enhancing communication about science outcomes with the science community, policy makers, and the public. Current scientific visualization systems are characterized by limited display resolutions that do not allow even one pixel per datum (a challenge that is likely to persist), limited interaction with the analytic engine and data, difficulty of use such that scientists seldom create their own visualizations, and inadequate attention to human perceptual and cognitive characteristics that influence how information is extracted from the visualization. Research is needed to facilitate visual analysis of extreme scale, heterogeneous, high-dimensional scientific data, including support for multiple users who may not be co-located. In particular, research is needed to develop methods for visual representation of uncertainty, visual comparison of multiple data sets or outcomes, and visual representations for tensor flow fields and vectors. Also of interest are interactive visual representations of system performance that support fault management and/or user intervention in long-running simulations and/or analytic processes on extreme-scale systems with heterogeneous architectures.

This FOA seeks applications for innovative basic research in computer science for management and analysis of extreme-scale scientific data in the context of petascale computers and/or exascale computers with heterogeneous multi-core architectures. Topics of interest include, but are not limited to:

- Extreme-scale data storage and access systems that are functional at the granularity and massive scale of scientific endeavors, that support data models that are consistent with scientists' view of their data, and that minimize the need for scientists to have detailed knowledge of system hardware and operating systems;
- Scalable data triage, summarization, and analysis methods and tools, including adaptive, power-aware algorithms and software for in-situ data reduction and/or analysis of massive multivariate data sets;
- Semantic integration of heterogeneous scientific data, including support for integrated analysis of multi-scale, multi-physics data from multiple types of devices and experiments, and/or simulations and spanning multiple scientific disciplines;
- Knowledge representation and automated machine reasoning and/or data mining methods and tools that support automated analysis and integration of large scientific data sets, especially those that include tensor flow fields, and/or representation of and reasoning about uncertainty and sources of uncertainty; and
- Visual analysis of extreme-scale scientific data, including multi-user visual analysis methods and tools for scientist users who may or may not be co-located, and interactive visual steering of computational processes.

These example research topics represent only a portion of the research challenges in extreme-scale data management and analysis that are of interest to ASCR. All interested proposers are encouraged to study the following references for additional discussion and insight:

1. *International Exascale Software Project Roadmap*, http://www.exascale.org/mediawiki/images/a/a1/Iesp-roadmap-draft-0.93-complete.pdf
2. *The Office of Science Data-Management Challenge*, March-May, 2004, http://science.doe.gov/ascr/ProgramDocuments/Docs/Final-report-v26.pdf
3. *Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale*, October 2007, http://science.doe.gov/ascr/ProgramDocuments/Docs/DOE-Visualization-Report-2007.pdf
4. *Modeling and Simulation at the Exascale for Energy and the Environment*, http://science.doe.gov/ascr/ProgramDocuments/Docs/TownHall.pdf
5. *ExaScale Software Study: Software Challenges in Extreme Scale System*, DARPA Information Processing Techniques Office (IPTO), September 14, 2009: http://users.ece.gatech.edu/mrichard/ExascaleComputingStudyReports/ECSS%20report%20101909.pdf
6. *Mathematics for Analysis of Petascale Data Workshop Report*, http://www.sc.doe.gov/ascr/ProgramDocuments/Docs/PetascaleDataWorkshopReport.pdf

Interested proposers should also be aware that ASCR has a related program on Mathematics for Analysis of Petascale Data (http://science.doe.gov/ascr/Research/AM/09-MAPD.html). Awards for this program will be posted on the ASCR web site when the proposal review and award processes are completed.

Community Building

An important goal of this notice is to foster an active, integrated research community of versatile researchers who are committed to the common goal of extreme-scale data management and analysis for scientific discovery. Accordingly,

- Each research team should plan to send representatives to a kick-off meeting and an annual PI meeting, where they will give presentations on the status and promise of their research and engage in working sessions to address shared problems. Meeting attendees will include invited participants from other relevant research communities. The objectives of these meetings include fostering a sense of community and serving as a venue for exchange of information with complementary programs.

The application will need to include plans for the dissemination of research results, such as:

- Publications, conferences, and educational activities for the science user community: what mechanisms will the project employ to present its work to a broader community to ensure sustained activities in the research area and promote adoption by communities of scientists?
- Code release: how will the codes be released to allow other researchers to continue building and expanding on the knowledge gained?
- Testing at scale: Will the project perform software testing at scale? If so, what are the requirements for this testing (for example, hardware, specific architecture, specific test bed, etc)?

This program requires open source software development. Applications should identify the open source license to be used.

Program Funding

It is anticipated that up to a total of $5,000,000 will be available for multiple awards in Fiscal Year 2010, contingent on the availability of appropriated funds for this program. Funding will be available annually for up to three years, with out-year support contingent on the availability of funds, progress of the research, and programmatic needs. At this funding level, 10-15 awards are anticipated. Depending on the quality of the applications, this supports 2-3 awards for each of the five major topics, each large enough to engage multiple researchers. Noting the excellence of industry research on topics of interest, partnerships with industry are strongly encouraged.

DOE is under no obligation to pay for any costs associated with the preparation or submission of an application. DOE reserves the right to fund, in whole or in part, any, all, or none of the applications submitted in response to this FOA.

Merit Review

Applications will be subjected to scientific merit review (peer review) and will be evaluated against the following evaluation criteria which are listed in descending order of importance codified at 10 CFR 605.10(d):

1. Scientific and/or Technical Merit of the Project;
2. Appropriateness of the Proposed Method or Approach;
3. Competency of Applicant's Personnel and Adequacy of Proposed Resources; and
4. Reasonableness and Appropriateness of the Proposed Budget.

The evaluation process will include program policy factors such as the relevance of the proposed research to the terms of the announcement and the agency's programmatic needs. Note that external peer reviewers are selected with regard to both their scientific expertise and the absence of conflict-of-interest issues. Both Federal and non-Federal reviewers may be used, and submission of an application constitutes agreement that this is acceptable to the investigator(s) and the submitting institution.

The Catalog of Federal Domestic Assistance (CFDA) number for this program is 81.049, and the solicitation control number is ERFAP 10 CFR Part 605.

Posted on the Office of Science Grants and Contracts Web Site
January 29, 2010.