Office of Science
Financial Assistance
Funding Opportunity Announcement
DE-PS02-07ER07-23

Operating and Runtime Systems for Extreme Scale Scientific Computation

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 for research grants in Operating and Runtime Systems for Extreme Scale Scientific Computation (FASTOS). This announcement is focused on research and development of operating and runtime systems which enable the effective management and use of extreme-scale systems (petascale and beyond) for scientific computation. The overall goal of this announcement is to stimulate research and development related to operating and runtime systems for petascale systems the in 2010 to 2015 timeframe. It is likely that these systems will include a combination of commodity and custom components, with different systems reflecting different degrees of customization. Operating and runtime systems research must be driven from the needs of current and future applications, and the primary focus is on supporting the needs of existing and anticipated SC and other DOE applications. An ultimate goal would be the development of a unified operating and runtime system that could fully support and exploit petascale and beyond systems and autonomously adapt to meet specific application needs for performance, functionality, security, and fault tolerance. The activities supported by this notice may be a combination of basic research, development, prototyping, and testing. Partnerships among universities, National Laboratories, and industry are encouraged.

PREAPPLICATION DUE DATE: April 6, 2007, 4:30 pm, Eastern Time

Potential applicants are ***required*** to submit a two-page preapplication by email to fjohnsonr@ascr.doe.gov. Preapplications must be received by **April 6, 2007, 4:30 p.m., Eastern Time**. The subject line of the email should be: **"FASTOS Preapplication"**. The preapplication should be a Word file attached to the email, having 1 inch margins when printed. No FAX or mail submission of preapplications will be accepted.

Preapplications will be reviewed for conformance with the guidelines and technical areas specified in this announcement. A response to preapplications encouraging or discouraging formal applications will be communicated to the applicants by April 13, 2007. Applicants who have not received a response regarding the status of their preapplication by this date are responsible for contacting the program to confirm their status.

Preapplications should consist of no more than two pages total. This narrative should give the project title and describe the research objectives, the technical approach(s), and all proposed team members and their expertise. It should also include a rough estimate of the planned budget

request. The intent in requesting a preapplication is to save the time and effort of applicants in preparing and submitting a formal project application that may be inappropriate for the program. Preapplications also assist ASCR in planning the peer review process and the selection of potential reviewers for the application. **Formal applications will be accepted only from preapplicants encouraged to submit a formal application.**

APPLICATION DUE DATE: June 11, 2007, 8:00 pm, Eastern Time

Applications must be submitted using Grants.gov, the Funding Opportunity Announcement can be found using the CFDA Number, 81.049 or the Funding Opportunity Announcement Number,
DE-PS02-07ER07-23. Applicants must follow the instructions and use the forms provided on Grants.gov.

FOR FURTHER INFORMATION CONTACT:

For further information regarding this notice,

Dr. Frederick Johnson
Telephone: (301) 903-5800
Fax: (301) 903-7774
E-mail: fjohnson@er.doe.gov

SUPPLEMENTARY INFORMATION:

Operating and runtime systems provide mechanisms to manage system hardware and software resources for the efficient execution of large scale scientific applications. They are essential to the success of both large scale systems and complex applications. By the end of this decade petascale computers with thousands of times more computational power than any in current use will be vital tools for expanding the frontiers of science and for addressing vital National priorities. These systems will have tens to hundreds of thousands of processors, an unprecedented level of complexity, and will require significant new levels of scalability and fault management. The overwhelming size and complexity of such systems poses deep technical challenges that must be overcome to fully exploit their potential for scientific discovery. Applications require multiple services from OS/R layers, including: resource management and scheduling, fault-management (detection, prediction, recovery, and reconfiguration), configuration management, and file systems access and management. Current and future large scale parallel systems require that such services be implemented in a fast and scalable manner so that the OS/R does not become a performance bottleneck. The current trend in large scale scientific systems is to leverage operating systems developed for other areas of computing - operating systems that were not specifically designed for large scale, parallel computing platforms. Unix, Linux and other Unix derivatives are the most popular OS's in use for high end scientific computing, and these all reflect a technological heritage nearly 30-years old with few fundamental mechanisms to support parallel systems.

Example Research Topics

Operating and runtime systems provide the glue that bind running applications to hardware. The research activities supported by this activity need to bridge the gap between new languages and/or programming models and next-generation hardware, including interactions with novel architectures. Consequently, there are a wide variety of research topics that are appropriate for this effort. A brief listing of candidate topics is provided below, but research in other relevant areas and combinations of areas is encouraged:

Virtualization. Virtualization is expected to play an increasingly important role in the deployment of large scale systems, enabling multiple operating systems on a single platform and application specific operating systems. Virtualization includes the development and use of hypervisors, virtual machine monitors, and application/runtime virtualization for HPC systems. Specific topics of interest include: identification and quantification of problems with current hypervisors in HPC systems, novel uses of hypervisors in HPC systems (development, porting, etc), support for fault handling, better support for custom hardware, and lightweight mechanisms for virtual resources.

Fault Handling. As the number of components in a system increase from tens to hundreds of thousands, these systems will have significantly reduced mean time between interrupt (MTI). Mechanisms to support application resiliency in the face of hardware faults are needed to support long running applications. Specific topics of interest include: tradeoffs associated with handling failures at different layers (application, runtime, OS); understanding and identifying sources of faults; approaches to proactive fault handling; fault tolerance for alternate (non- MPI) programming models; languages/APIs for the bi-directional communication of fault information between layers (e.g., between the application and runtime layers); quantification of scalability issues; automatic, transparent, and efficient checkpoint/restart; and checkpointing when disks are far away.

OS Noise/Interference. Operating system interference or noise due to asynchronous overhead needed to implement system services, has been shown to have a significant impact on application performance on very large scale systems. Measurement and understanding the impact of OS interference on application performance at scale will be critical to the successful deployment of very large scale systems. Specific topics of interest include: OS design strategies for dealing with OS noise (e.g., implementations of critical services that minimize related noise and alternatives for timeouts and/or periodic service requirements); hardware features to control the impact of noise (e.g., hardware support for low overhead barriers); strategies to mitigate the impact of OS noise (e.g., exploiting asynchrony).

Exposing Resources. Bidirectional APIs to expose system information (performance counters) and to select implementations are critical for application level adaptability (need information about what is being used and may need to select alternate implementations). Specific topics of interest include: hooks for controlling resources; interfaces to allow code to query hardware characteristics; exposing communication related resources.

Resource Management. Managing the local and global resources provided by a computing system is a fundamental responsibility of any operating system, and exploration of policies and mechanisms for resource management is especially critical for petascale systems. Specific topics

of interest include: local resource management (memory management, processor scheduling (multi-core), and communication support); interfaces between local and external components (gang scheduling, virtual memory reservations and queries); support for alternate (non-MPI) programming models (e.g., UPC); OS service coordination (load balancing at scale, global memory management, topology aware mapping of work- and data-units); heterogeneous resource management (HW and SW); and power management.

Adaptability. The ability of operating and runtime systems to change their behaviors based on application needs to improve performance or tolerate faults needed to support the use of petascale systems. Specific topics of interest include: measurement and strategies to support adaptation; understanding and exploiting application phases; adapting collective communication components; and APIs to expose resource performance models and information.

Performance Measurement. Petascale systems will require models and tools to measure system performance, including hooks for application level performance monitoring; tools to measure runtime/OS performance; performance models (define what needs to be measured); and scalability.

System Management/Administration. Several issues related to overall system administration need to be addressed, including: usage models (space/time sharing); flexible space-sharing; changing processors allocated to running jobs; single system image issues to ease system management number of system administrators should not scale with the size of the system; node allocation; power management; software distribution; and RAS and RAS interfaces.

Parallel I/O: Efficient communication with external storage servers and parallel file systems is an essential component of a petascale system. Topics of interest include: support for high performance access to external servers, efficient, scalable I/O call forwarding, portable I/O models which support diverse storage instantiations, and parallel file systems.

Community building

An important goal of this notice is to foster the development of an active research community in operating systems and runtime environments for high end systems. In order to meet this goal the following are mandatory requirements for awardees:

- All developed code must be released under the most permissive open source license possible. This is to enable other researchers and vendors to build upon research successes with a minimum of intellectual property issues.
- Each research team should plan to send representatives to annual or semi-annual PI meetings and give presentations on the status and promise of their research. Meeting attendees will include invited participates from other relevant research communities, including the Linux community. Objectives of these meetings are to foster a sense of community and serve as a venue for exchange of information. These meetings will also serve as a means to exchange information on complementary programs including the DARPA HPCS program, NNSA ASC program and DOE/SC SciDAC program.

Testbed access

Applications should provide a plan for utilizing leadership class systems at Oak Ridge National Laboratory and Argonne National Laboratory and to systems at the National Energy Research Scientific Computing Center (NERSC) at Lawrence Berkeley National Laboratory for the purpose of software testing at scale. Each application should contain a section which discusses the characteristics of the test environments necessary for the research and identify the time frames in which specific testbed support will be required. Only a relatively limited amount of testing time will be available on these systems, and the individual testing plans will be used to develop an overall test plan for the FASTOS program.

Program Funding

It is anticipated that up to $3 million annually will be available for multiple awards for this program. Awards are planned to be made in Fiscal Year 2008, and applications may request project support for up to three years. All awards are contingent on the availability of funds and programmatic needs. Annual budgets for successful projects are expected to range from $500,000 to $1,000,000 per project although smaller projects of exceptional merit may be considered. Annual budgets may increase in the out-years but should remain within the overall annual maximum guidance.

References

FASTOS forum: http://www.cs.unm.edu/~fastos

Federal Plan for High-End Computing:
http://www.nitrd.gov/pubs/2004_hecrtf/20040702_hecrtf.pdf

Posted on the Office of Science Grants and Contracts Web Site
March 7, 2007.