Office of Science
Notice DE-FG01-04ER04-13

Operating/Runtime Systems for Extreme Scale Scientific Computation

Department of Energy

Office of Science Financial Assistance Program Notice DE-FG01-04ER04-13; Operating/Runtime Systems for Extreme Scale Scientific Computation

AGENCY: U.S. Department of Energy

ACTION: Notice inviting grant applications.

SUMMARY: The Office of Advanced Scientific Computing Research (ASCR) of the Office of Science (SC), U.S. Department of Energy (DOE), hereby announce its interest in receiving applications for research grants in the area of operating and runtime systems for extreme scale scientific computation. Partnerships among universities, National Laboratories, and industry are encouraged. The full text of Program Notice 04-13, is available via the Internet using the following web site address: http://www.science.doe.gov/production/grants/grants.html.

DATES: Preapplications referencing Program Notice 04-13, should be received by March 26, 2004.

Formal applications in response to this notice should be received by 4:30 p.m., Eastern Time, May 4, 2004, to be accepted for merit review and funding in Fiscal Year 2004.

ADDRESSES: Preapplications referencing Program Notice 04-13, should be sent via e-mail using the following address: osruntime.preproposal@science.doe.gov with a copy to fjohnson@er.doe.gov.

Formal applications referencing Program Notice 04-13, must be sent electronically by an authorized institutional business official through DOE's Industry Interactive Procurement System (IIPS) at: http://e-center.doe.gov/. IIPS provides for the posting of solicitations and receipt of applications in a paperless environment via the Internet. In order to submit applications through IIPS your business official will need to register at the IIPS website. **IIPS offers the option of using multiple files, please limit submissions to one volume and one file if possible, with a maximum of no more than four PDF files.** The Office of Science will include attachments as part of this notice that provide the appropriate forms in PDF fillable format that are to be submitted through IIPS. Color images should be submitted in IIPS as a separate file in PDF format and identified as such. These images should be kept to a minimum due to the limitations of reproducing them. They should be numbered and referred to in the body of the technical scientific application as Color image 1, Color image 2, etc. Questions regarding the operation of

IIPS may be E-mailed to the IIPS Help Desk at: HelpDesk@pr.doe.gov or you may call the help desk at: (800) 683-0751. Further information on the use of IIPS by the Office of Science is available at: http://www.science.doe.gov/production/grants/grants.html.

If you are unable to submit the application through IIPS, please contact the Grants and Contracts Division, Office of Science at: (301) 903-5212 or (301) 903-3604, in order to gain assistance for submission through IIPS or to receive special approval and instruction on how to submit printed applications.

FOR FURTHER INFORMATION CONTACT: Dr. Frederick Johnson, U.S. Department of Energy, Office of Science, SC-31/Germantown Building, 1000 Independence Avenue, S.W., Washington, DC 20585-1290, telephone: (301) 903-3601, fax: (301) 903-7774, E-mail: fjohnson@er.doe.gov.

SUPPLEMENTARY INFORMATION:

The Forum to Address Scalable Technologies for Runtime and Operating Systems (FAST-OS) has conducted a series of workshops focused on issues associated with operating and runtime systems for very large computing systems used for high end scientific modeling and simulation. This workshop series was sponsored by the Office of Advanced Scientific Computing Research of the DOE Office of Science. The most recent workshop was held in July 2003, and the final report, together with other results of the workshop series may be found at: http://www.cs.unm.edu/~fastos. An interagency workshop, the Workshop on the Roadmap for the Revitalization of High-End Computing was held in June of 2003. Section 5 of the workshop report addresses runtime and operating systems. The charter of the researchers that produced this section was to establish baseline capabilities required in the operating systems for projected High-End Computing systems scaled to the end of this decade and determine the critical advances that must be undertaken to meet these goals. The report is available at: http://www.itrd.gov/hecrtf-outreach/20040112_cra_hecrtf_report.pdf.

Background: Operating and Runtime Systems (OS/R)

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 no fundamental mechanisms to support parallel systems.

Without reliable, robust operating systems and runtime environments the computational science research community will be unable to easily and completely employ future generations of extreme systems for scientific discovery. The application research community will miss important scientific opportunities in areas such as computational fusion, nanotechnology, and computational biology that are on the threshold of rapid advance through the innovative use of extreme-scale scientific computation. New investments in both basic and applied research are required to maintain the creative pace established by terascale computation for scientific discovery.

Background: High-End Computing Revitalization Task Force (HECRTF) and Academic Research

During the past summer, several federal agencies with interests in high performance computing participated in the HECRTF and developed a plan for future government investments in high-end computing. As part of this plan a renewed emphasis has been placed on coordination of federally-funded research in this area. As a major contributor to the HECRTF activity, the Office of Science is a leading participant in the coordination of research investments. The research activities described in this Notice have been coordinated with participating HECRTF research agencies, and this coordination will continue throughout the lifetime of the research activities. Additional information on the HECRTF may be found at: http://www.itrd.gov/hecrtf-outreach/index.html.

The Opportunity and the Challenge

By the end of this decade extreme scale systems will be available that are based on a variety of challenging architectures ranging from distributed memory clusters of unprecedented scale to the systems resulting from the DARPA High Productivity Computing Systems program that are likely to be based upon innovative architectural concepts, such as PIMs, FPGAs, and complex memory hierarchies that have no analog in today's terascale systems. Systems with tens to hundreds of thousands of processors and new architectural concepts will differ greatly in scale and complexity from today's systems, and this difference will place new and very difficult challenges on OS/R design and implementation.

There are many fundamental questions in operating system and runtime research that must be explored in order to enable scientific application developers and users to achieve maximum effectiveness and efficiency on this new generation of systems, including (but not limited to):

- **Ease of use.** Application users need a coherent, cohesive picture of these huge systems – they need to be able to look at jobs running on 100,000 processors in a meaningful way.

- **Support for architectural innovation.** Current operating systems often limit hardware innovation through the use of a hardware abstraction layer that cannot support innovative hardware paradigms.
- **Dynamic support for multiple management policies.** Current operating systems limit application development through the use of fixed resource management policies rather than dynamic policies responsive to changing application needs.
- **Leveraging mainstream technology.** Strategies are needed that enable OS/R systems developed to meet specialized needs of the HEC community to leverage the talents and technology development of the mainstream open source OS community.
- **Support for fault tolerance.** Extreme scale systems will require innovative new approaches to OS/R support for fault detection and management. Interrupts are likely to be the norm rather than the exception during any lengthy application run.
- **Rethinking the OS in terms of scalability and usability.** We need to determine how HPC requirements differ from those of general computing. HPC requirement differences will surely continue to dictate innovation in both OS structure and exported interfaces.
- **Scalability of operating systems** What should an operating system for a hundred thousand processor machine look like? Is a hierarchical approach best? How can the operating system make a fundamentally unreliable machine, in which some components are always broken, continue to effectively function?
- **Self awareness and optimization.** How can an extreme scale system (hardware and software) monitor and adapt to meet changing requirements of long running applications?

Technical challenges such as these represent an opportunity for basic and applied research to provide new insights into mechanisms for harnessing the potential of next generation extreme-scale systems.

Investment Plan of the Office of Science

The Secretary of Energy recently released a twenty year vision and plan for research facilities in the Office of Science in the document, Facilities for the Future of Science: A Twenty-Year Outlook. A copy of the plan may be found at: http://www.sc.doe.gov/Sub/Facilities_for_future/20-Year-Outlook-screen.pdf. The plan contains a prioritized list of new research facilities, and the number two priority is an UltraScale Scientific Computing Capability (USSCC), which will increase by at least a factor of 100 the computing capability available to support open scientific research and which will reduce from years to days the time required to simulate complex systems of interest to the Department. When fully realized, the computing capability of the USSCC will enable computation-based scientific advances that are unachievable by current large-scale computing systems. USSCC systems will place new and critical demands on operating systems and runtime environments to support complex applications and enable these systems to reach their full potential. The research supported by this notice is a critical step towards developing OS and runtime systems able to meet these needs.

Solicitation Emphasis

This notice is focused on research and development of operating and runtime systems which enable the effective management and use of extreme-scale systems (petascale and above) for scientific computation. The overall goal of this notice is to stimulate research and development related to operating and runtime systems for petascale systems the in 2010 timeframe. It is likely that these systems will include a combination of commodity and custom components, with different systems reflecting different degrees of customization. The research into runtime and operating systems must be driven from the needs of current and future applications. The primary focus is on supporting the needs of existing and anticipated SC and other DOE applications; however, the resulting systems should address issues related to the broader HEC code base. An ultimate and perhaps idealistic goal would be to develop a unified runtime and operating system that could fully support and exploit petascale and beyond systems and autonomously adapt for performance, upgrades, security, and fault tolerance. The activities supported by this notice may be a combination of basic and applied research, development, prototyping, testing and ultimately deployment.

Example Research Topics

Runtime and operating 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.** A key aspect of OS/R systems is that they provide "virtual devices." Virtualization must balance ease of use by detail hiding vs achieving scalability and performance by exploiting details.
- **Adaptation.** Traditionally, runtime and operating systems have been designed to provide a fixed set of services and to provide a single implementation for each of these services. Future runtime and operating systems will need to provide different sets of services and/or different implementations of these services based on the needs of applications and/or characteristics of the underlying system.
- **Usage models.** Large machines have typically been used in batch mode. Other modes of operation, including interactive usage for computational steering will also need to be supported in the future.
- **Metrics.** Metrics, benchmarks, and test suites are needed to evaluate progress and guide design. Challenges include determining what to measure and how to generate understandable analyses. Benchmarks and test suites must accurately reflect the needs of applications.
- **Support for fault handling in OS and run-time.** Many jobs will encounter an interrupt in service during their execution. Research is needed to address all aspects of fault tolerance, including fault detection, anticipation, management and tolerance. Research in checkpointing systems is also needed.
- **Memory hierarchy management.** It is clear that the memory hierarchy is going to become deeper and/or more complex. Applications will need significantly improved support for managing memory.

- **Security.** Scalable security mechanisms are needed to support new authorization, authentication and access control requirements.
- **Common API.** Research in common runtime/OS API's is required to greatly enhance application portability and ease the introduction of new systems. The current POSIX standard has been beneficial to the general community, but it is lacking in the support of high-end systems.
- **Scalable, single-system image.** In principle, the ability to treat a very large system as a single system has many advantages and provides significant simplifications from an end user perspective. However, it is not clear what the technical trade-offs are for single system image technology at extreme scale, and additional research is needed.
- **Parallel and Network I/O.** Some classes of future HEC systems will have specialized interconnect fabrics to provide communications and data movement among processors or groups of processors or storage devices. Operating systems and/or runtime systems will be required to share, schedule, and control these resources.
- **OS Support for efficient interprocessor communication.** Standard OS's do not recognize the concept of a parallel job. Support is needed for global operations which minimize local variations and avoid degradation of performance for the whole job.
- **Light-weight low-level communication paradigms.** Research in light-weight and low level communication mechanisms is needed to improve scalability and performance.

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 SciDAC.

Frameworks and Novel Approaches

Operating system and runtime research often requires a large overhead of supporting infrastructure code, such as device drivers, that must be developed before undertaking the core ideas of the research. This may be alleviated if an existing OS framework, such as Linux, K42, or Plan9, is chosen as a base of the research. Applications to this notice may choose to use an existing framework for their OS/Runtime research or they may propose to develop a new framework as part of the research activity. Any proposed new framework must be described and discussed at the community PI meetings. Smaller novel approaches are also encouraged.

Testbed strategy

Testbeds are essential to the future of the research sponsored by this notice, and the development of an effective testbed strategy is an important overall objective. Each proposal 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.

Operating system and runtime applications to the ASCR base programs through the Continuing Solicitation for all Office of Science Programs Notice 04-01, found at: http://www.science.doe.gov/production/grants/grants.html, which may have the potential for contributing to extreme scale systems, should so indicate.

Collaboration

Applicants are encouraged to collaborate with researchers in other institutions, such as universities, industry, non-profit organizations, federal laboratories and Federally Funded Research and Development Centers (FFRDCs), including the DOE National Laboratories, where appropriate, and to include cost sharing wherever feasible. Additional information on collaboration is available in the Application Guide for the Office of Science Financial Assistance Program that is available via the Internet at: http://www.sc.doe.gov/production/grants/Colab.html.

Program Funding

It is anticipated that up to $3 million annually will be available for multiple awards for this program. Initial awards will be made late in Fiscal Year 2004 or early Fiscal Year 2005, in the categories described above, 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 $1,500,000 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. Any proposed effort that exceeds the annual maximum in the out-years should be separately identified for potential award increases if additional funds become available. DOE is under no obligation to pay for any costs associated with the preparation or submission of applications if an award is not made.

Preapplications

Preapplications are strongly encouraged but not required prior to submission of a full application. However, notification of a successful preapplication is not an indication that an award will be made in response to the formal application. The preapplication should identify on the cover sheet the institution(s), Principal Investigator name(s), address(s), telephone, and fax number(s) and E-mail address(es), and the title of the project. A brief (one-page) vitae should be provided for each Principal Investigator. The preapplication should consist of a two to three page narrative describing the research project objectives, the approach to be taken, a description of any research partnerships, the duration, and an annual cost estimate.

Merit Review

Applications will be subjected to scientific merit review (peer review) and will be evaluated against the following evaluation criteria listed in descending order of importance as 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,
4. Reasonableness and Appropriateness of the Proposed Budget.

The evaluation of applications under item 1, Scientific and Technical Merit, will pay particular attention to:

a) The potential of the proposed project to make a significant impact in operating systems and runtime research.

b) The demonstrated capabilities of the applicants to perform basic research related to operating systems/runtime and transform these research results into software that can be widely deployed.

c) The likelihood that the methodologies and software components that result from this effort will have a substantial impact on the operating system research and vendor community outside of the projects.

The evaluation under item 2, Appropriateness of the Proposed Method or Approach, will also consider the following elements related to Quality of Planning:

a) Quality of the plan for effective coupling of operating system and runtime research, with application needs and transition to testbed environments.

b) Quality and clarity of proposed work schedule and deliverables.

c) Quality of the proposed approach to intellectual property management and open source licensing.

Note that external peer reviewers are selected with regard to both their scientific expertise and the absence of conflict-of-interest issues. 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. Reviewers will be selected to represent expertise in the technology areas proposed, applications groups that are potential users of the technology, and related programs in other Federal Agencies or parts of DOE, such as the Advanced Strategic Computing Initiative (ASCI) within DOE's National Nuclear Security Administration.

Information about the development and submission of applications, eligibility, limitations, evaluation, selection process, and other policies and procedures including detailed procedures for submitting applications from multi-institution partnerships may be found in 10 CFR Part 605, and in the Application Guide for the Office of Science Financial Assistance Program. Electronic access to the Guide and required forms is made available via the World Wide Web at:

http://www.science.doe.gov/production/grants/grants.html. The Project Description must be 20 pages or less, including tables and figures, but exclusive of attachments. The application must contain an abstract or project summary, letters of intent from collaborators, and short vitae.

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

Martin Rubinstein
Acting Director
Grants and Contracts Division
Office of Science

Published in the Federal Register March 17, 2004, Volume 69, Number 52, Pages 12648-12651.