TUTORIAL III

Building Shared High Performance Computing Infrastructure for the Biomedical Sciences: Learnings from Biomed HPC 2007

Marcos Athanasoulis Dr.PH and Florence Reinisch MPH

Harvard Medical School
Boston, Massachusetts
USA

DESCRIPTION

In recent years high performance computing has moved from the sidelines to the mainstream of biomedical research. Increasingly researchers are employing computational methods to facilitate their wet lab research. Some emerging laboratories and approaches are based on a 100% computational framework. While there are many lessons to be learned from the computational infrastructure put into place for the physical and mechanical sciences, the character, nature and demands of biomedical computing differ from the needs of the other sciences. Biomedical computational problems, for example, tend to be less computationally intensive but more "bursty" in their needs. This creates both an opportunity (it is easier to meet capacity needs) and a challenge (job scheduling rules are more complicated to accommodate the bursts).

Harvard Medical School provides one of the most advanced shared high performance research computing centers at an academic medical center. In 2007, Harvard convened the first Biomedical High Performance Computing Leadership Summit to explore the issues in creating shared computing infrastructure for the biomedical sciences. We brought together over 100 leaders in the field to exchange ideas and approaches. Through special sessions and direct participant surveys a number of themes emerged around best practices in deploying shared computational infrastructure for the biomedical sciences. Based on prior experience and the summit findings, this workshop summarizes the approaches and ideas to providing a technical and process blueprint for organizations wishing to provide shared research computing research resources for groups small or large – from a few hundred CPUs and terabytes of data to thousands of CPUs and a petabyte or more of data.

TUTORIAL OUTLINE

The workshop includes the following topics:

- Summary of the current problems in Biomedical Sciences HPC
    - Image Processing
    - Simulation
    - 'Omics
    - Translational Research

- Data Centers and Hardware
  - Solving density problems
  - Power and cooling strategies
  - Blade servers and the multi-core machine
- Deployment Architectures
  - Approaches to system imaging
  - Supporting fault tolerant applications
  - Distributed storage, ready for production use?
  - Proprietary interconnects – cost/benefit analysis
  - Virtualization
- Job Scheduling
  - Approaches to time and resource based queues
  - Handling the challenges of parallel vs. distributed
  - Integrating "contributed" hardware
- Managing Storage Growth
  - SAN vs NAS
  - Distributed file systems
  - Archiving and near-line storage
  - New approaches to compression and de-duplication
- Organizational Challenges
  - How to ask for and get seed funding
  - Measuring performance and Return on Investment
  - Affiliating for group purchasing power
  - Workflow and support models
  - Working with the ego of the PI
  - Setting limits of services
- Putting it into Action
  - Online and offline resources
  - Communities and colleagues
  - Deployment planning tools

TARGET AUDIENCE

The target audience includes any researcher and or research IT core service provider who are interested in the challenges of providing shared high performance computing infrastructure to the biomedical sciences. From the postdoc who needs to set-up a modest compute cluster for their laboratory to the senior researcher who has been charged with providing world class infrastructure this tutorial will make them aware of the foundations and latest challenges of biomedical HPC.

REQUIRED BACKGROUND

While it is expected that tutorial attendees should be information technology professionals with a basic background in systems deployment and computer sciences, the session should prove valuable to anyone with an interest in the challenges and opportunities in creating high performance computing infrastructure for the biomedical sciences.

DURATION

The tutorial will take two hours. The bulk of the session will be devoted to the technical challenges and current issues in the field. In the final part of the session, participants will have the opportunity to present their plan for their home institution to the tutorial participants.

INSTRUCTOR BIOGRAPHIES

Dr. Athanasoulis is the chair of the Biomedical High Performance Computing Leadership Summit and Director of Client Services and Research Information Technology for Harvard Medical School where he oversees the IT service

operations for the school and leads the development of high performance computing infrastructure to support biomedical and healthcare research. During his career, Dr. Athanasoulis has worked in both the public and private sector to improve the quality and efficiency of healthcare and research through information systems. Prior to joining Harvard Medical School, Dr. Athanasoulis was the Vice President of Product Development at RelayHealth Corporation, Inc., where he oversaw the continuing development and implementation of an advanced patient-provider communication system. As Chief Technology Officer at HealthCentral.com, he led the development of health information systems for more than 100 hospitals and health plans as well as a consumer portal that served millions of consumers. Dr. Athanasoulis has consulted to a wide variety of health care organizations including, UC San Diego, the Koop Foundation, the California Department of Health Services, San Francisco General Hospital, Alta Bates Hospital, the National Community Pharmacists Association and the UC Berkeley Wellness Guide. He is also the chief technical advisor for Healia, Inc. and co-founder of the Healthy Communities Foundation. He holds a master's degree in epidemiology and biostatistics and a doctorate in health informatics, both from UC Berkeley where he was a University Fellow.

Ms. Reinisch, is the Program Director for the Biomedical High Performance Computing Leadership Summit. She has more than twelve years of experience designing information systems in the biomedical sector. As program officer for the Healthy Communities Foundation she leads the implementation of a web based indicators project deployed in California and Washing State. She has served as co-investigator on multiple NIOSH –funded projects, including a four year study to evaluate the 1994 CDC Guidelines for the control of nosocomial transmission of tuberculosis risk among health care workers. Ms Reinisch was previously Director for the California Sharps Injury Prevention Program, where she developed a dynamic web application for the program, and served as co-investigator in a three-year CDC grant. She has expertise conducting surveillance and epidemiologic studies in occupational and environmental health, designing data management systems, statistical analysis, project management and web application deployment. She holds a Masters degree in Epidemiology and Biostatistics from University of California, Berkeley.