MAGIC Meeting
September 7, 2011, 2:00-4:00

Attendance:

Gabrielle Allen	NSF
Rachana Ananthakrishnan	ANL
Sam Angiuoli	Un of Md.
Shane Canon	NERSC
Rich Carlson	DOE
Susan Coughlin	ANL
Gary Crane	SURA
Ian Gable	Un of Victoria
Ollie Gray	
Dan Gunter	LBL
Chris Hill	MIT
Shantenu Jha	Rutgers University
Kate Keahy	Argonne National Lab
Ken Klingenstein	Internet2
Rob Knight	Un of Colorado
Lamal Krishnan	
Archit Kulshrestha	Indiana University
Stuart Martin	ANL
Don Middleton	UCAR
Grant Miller	NCO
Doug Olson	
Lavanya Ramakrishnan	LBNL/NERSC
Joel Replogle	OGF
Alan Sill	Texas Tech Un
Randall Sobie	Un of Victoria
Steve Tuecke	Un of Chicago
Jay Unger	vDesk Service Inc.
John Volmer	ANL
Von Welch	Indiana Un
Hal Warren	
Dean Williams	LL labs

Action Items

1. Please read, on the MAGIC WIKI, (http://connect.nitrd.gov/magicwiki/index.php?title=Special:ListFiles) the MAGIC tasking document and provide comments back so it can be finalized.

Proceedings

This meeting of MAGIC was chaired by Gabrielle Allen of the NSF and Rich Carlson of DOE. SC11 accepted a BOF which will be held on Wednesday at SC11.

Kate Keahy organized the discussion on Cloud Infrastructure (CI) for this MAGIC meeting. Discussants were asked to discuss several questions that include:

- Describe how your community uses infrastructure clouds
- What drew you to infrastructure clouds, i.e., what are the benefits for your community of using infrastructure clouds in order of significance?
- What are the challenges of using infrastructure clouds for your community, i.e., what makes them difficult to use in order of significance?
- How do infrastructure clouds compare to other options for outsourcing computation from the perspective of your community's needs?

Clouds in Bioinformatics: Rob Knight

New instruments in bioinformatics collect vast datasets. Investigators are interested in time-to-result. Demand is highly elastic. Workflow services (Galaxy, CIPRES, QIIME) deployed to the cloud enable rapid processing of sequence data for genome assembly and microbial community analysis. Cloud-enabled GUIs are also enabling biologists to run analyses themselves using vast computational resources.

Infrastructure Clouds (ICs) allow large resources (e.g. EC2, Amazon cloud) to be used on-demand at small cost relative to buying a resource to handle peak loads. Hardware support and systems administration is greatly reduced. Third parties, outside the project, using software can readily pay for those resources instead of drawing from project resources. Machine images can easily be used by end-users with no previous cloud experience.

It is difficult to get large data sets into the ICs. Currently mail drives are used. Larger network access is needed. EC2 has had random node and network failures requiring modifications of codes to effectively run applications... Existing clouds are suited only to a subset of bioinformatics tasks. Higher-memory configurations with more storage are needed.

EC2 and Magellan are vastly easier to use than native installations with arbitrarily configured clusters. ICs are easier to use than TeraGrid.

Distributed Clouds for Particle Physics and Astronomy Research Computing: Randall Sobie and Ian Gable

ICs are used in support of BaBar, Atlas Tier 3 computations, and CANFAR (Astronomical application). Virtual Machines (VMs) and clouds provide users with base-VMs to customize and store in image repositories. Access is provided to multiple computing centers configured as clouds. They seek to use these clouds as a single resource for batch jobs. Sysadmins do not have to be application specialists. X509 certificates are used for authentication. Users build their code within an application-specific VM and submit to a batch system. The system boots the user-customized application on one of the clouds and retrieves data from a local or remote source. These capabilities have been used for 1 year with tens of thousands of VMs booted and jobs processed. EC2 and Nimbus clouds are being used. The main challenges are security and authentication and data I/O.

Infrastructure Clouds: Chris Hill

ICs are being used to investigate the feasibility of running coupled Atmosphere-Ocean climate models on Amazon's EC2. Field experiments require ensembles of models run in real-time for a period of a few weeks per year. This bursty demand makes it more cost-effective to use cloud resources on-demand. The CCA model is a complex coupled model with MPI but modest scale of 10-100 cores. EC2 provides a system-in-a-box for non-experts to use resources as needed. This system is also useful for supporting teaching demonstrations. This application provides a simple interface between hardware and packager with a lot of flexibility. A VM can be provided for each simulation configuration customized to the need.

The performance is not as good as an HPC cluster. Persistent storage is particularly pricey. There is no standardized high-level interface for subscribing and launching an application... it works best for providing non-expert users access to research tools.

STAR Experiment: Doug Olson

The STAR Experiment is using many models and approaches including EC2, Nimbus, Condor outside, Condor inside, Virtual Organization Cluster, PBS, Kestrel, and IM mechanisms to control jobs. They also used the Magellan infrastructure at both NERSC and ANL. These resources were used to run Monte-Carlo simulations, running complex simulations spanning years long simulations compressed to human time scales for results delivery. They plan to use ICs for easy software provisioning at remote sites and for long-term software preservation.

Clouds provide rapid virtualization environments, providing a way to "can" software in a consistent and self-contained environment... They provide large savings of workforce. They enable long-term software preservation. Clouds work at scale.

Issues encountered include the lack or expense of storage solutions on commercial clouds. Operational support is sometimes problematic (who do you call?). Handling VM images is not always easy; stability in VM formats would be helpful. Interoperability and portability between cloud environments is needed. Consistent Global job monitoring and AA mechanisms are needed.

IC clouds provide true elasticity. They enable long-term software preservation.

Magellan: Lavanya Ramakrishnan

The Magellan Cloud infrastructure has a broad set of users. Magellan provides a wide range of resources and capabilities to users. Important capabilities for users included:

- Access to resources
- The ability to share setup of software and experiments with collaborators
- Ability to control software environments
- Easier to acquire and operate than a local cluster
- Use of clouds to host science gateways and to access cloud resources
- Many other capabilities

Challenges of using ICs include:

- Performance, scalability, reliability
- Security, allocation, and accounting

- Application design and development

Infrastructure Clouds, microbial genomics, and the cloud virtual resources project (CloVR): Sam Angiuoli

The microbial genomics community has developed a range of small to medium sized databases that are being shared in the community. CloVR integrates analysis software and automated pipelines into a portable VM that is run on local PCs and can access the cloud seamlessly on demand. EC2 has been used because it is reliable, flexible and enables root access and on-demand use. The community has developed higher throughput, improved control and the resources are easy to use (CloVR is prebuilt and local clusters do not have to be managed by using cloud resources).

Challenges to using cloud resources include poor reliability, poor availability (outside commercial resources), limited portability, need for authentication and authorization and big data needs increased throughput.

Discussion among the MAGIC members identified that there is a strong need for:

- The ability to port into the cloud, and to use, large data sets of 1-10 TB
- User interfaces that flexibly and simply accommodate to user requirements over a wide range of requirements
- Security, authentication, authorization
- Portability, interoperability among providers
- Balancing VM management and deployment time and efficiency

For the complete briefings on this topic, please see:
http://connect.nitrd.gov/magicwiki/index.php?title=Meeting_Minutes_and_Materials

MAGIC Input to the LSN Annual Planning Meeting

Rich Carlson provided, on the MAGIC WIKI, a statement and briefing on MAGIC accomplishments and potential taskings by the LSN to MAGIC.

AI: Please read, on the MAGIC WIKI (http://connect.nitrd.gov/magicwiki/index.php?title=Special:ListFiles) the MAGIC tasking document and provide comments back so it can be finalized.

Meetings of Interest

December: DOE workshop on extreme scale science collaboration. What environments do you need to support computing at exascale and to support tele-instrumentation requirements for large collaborative scientific instruments? Contact Thomas Ndousse of DOE if you wish to attend.

Next MAGIC Meetings

October 5, 2011, 2:00-4:00, NSF
November 2, 2011, 2:00-4:00, NSF