Magellan Status Report
A Test Bed to Explore Cloud Computing for Science

Susan Coghlan (Argonne)
Shane Canon (Berkeley Lab)

May 17, 2011
Outline

• Overview of the Magellan Project
• Overview of Cloud Computing
• Overview of the distributed Testbed
• Lines of Inquiry (early findings)
• Conclusions
Co-located at two DOE-SC Facilities

- Argonne Leadership Computing Facility (ALCF)
- National Energy Research Scientific Computing Center (NERSC)
- Funded by DOE under the American Recovery and Reinvestment Act (ARRA)
Magellan Scope

• Mission
  – Determine the appropriate role for private cloud computing for DOE/SC midrange workloads

• Approach
  – Deploy a test bed to investigate the use of cloud computing for mid-range scientific computing
  – Evaluate the effectiveness of cloud computing models for a wide spectrum of DOE/SC applications
## Magellan Timeline

<table>
<thead>
<tr>
<th>Activity</th>
<th>Argonne</th>
<th>NERSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Project Start</td>
<td></td>
<td>Sep 2009</td>
</tr>
<tr>
<td>User Access</td>
<td>Mar 2010 (Cloud)</td>
<td>April 2010 (Cluster)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Oct 2010 (Cloud)</td>
</tr>
<tr>
<td>Acceptance</td>
<td>Feb 2010</td>
<td>May 2010</td>
</tr>
<tr>
<td>Hadoop User Access</td>
<td>Dec 2010</td>
<td>May 2010</td>
</tr>
<tr>
<td>Joint Demo (MG-RAST)</td>
<td></td>
<td>June 2010</td>
</tr>
<tr>
<td>Nimbus Deployed</td>
<td>Jun 2010</td>
<td>N/A</td>
</tr>
<tr>
<td>OpenStack Deployed</td>
<td>Dec 2010</td>
<td>N/A</td>
</tr>
<tr>
<td>Eucalyptus 2.0 Deployed</td>
<td>Jan 2011</td>
<td>Feb 2011</td>
</tr>
<tr>
<td>ANI research projects on</td>
<td></td>
<td>Apr 2011 – Dec 2011</td>
</tr>
<tr>
<td>Magellan cloud ends</td>
<td></td>
<td>Sep 2011</td>
</tr>
<tr>
<td>ANI 100G active</td>
<td></td>
<td>Oct 2011</td>
</tr>
<tr>
<td>Magellan ANI ends</td>
<td></td>
<td>Dec 2011</td>
</tr>
</tbody>
</table>
What is a Cloud?
Definition

According to the National Institute of Standards & Technology (NIST)...

• **Resource pooling.** Computing resources are pooled to serve multiple consumers.
• **Broad network access.** Capabilities are available over the network.
• **Measured Service.** Resource usage is monitored and reported for transparency.
• **Rapid elasticity.** Capabilities can be rapidly scaled out and in (pay-as-you-go)
• **On-demand self-service.** Consumers can provision capabilities automatically.
What is a cloud?

Cloud Models

<table>
<thead>
<tr>
<th>Hardware focus</th>
<th>Software focus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Infrastructure as a Service (IaaS)</strong></td>
<td><strong>Platform as a Service (PaaS)</strong></td>
</tr>
<tr>
<td>Provisions processing, storage, networks, and other fundamental computing resources. Consumer can deploy and run arbitrary software, including OS.</td>
<td>Provides programming languages and tools. Consumer applications created with provider’s tools.</td>
</tr>
<tr>
<td>• Amazon EC2</td>
<td>• Microsoft Azure</td>
</tr>
<tr>
<td>• RackSpace</td>
<td>• Google AppEngine</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- Opaque infrastructure
- Capacity >> Demand
- Available for rent
- Self-service
Magellan Distributed Testbed
Distributed Testbed Summary

• Compute
  – IBM iDataPlex: 504 nodes at Argonne and 720 nodes at NERSC

• Storage
  – Mix of disk storage, archival storage, and two classes of flash storage

• Architected for flexibility and to support research
  – Similar to high-end hardware in HPC clusters
  – Suitable for scientific applications
  – Included some specialized hardware such as GPUs
Argonne Magellan Hardware

**Compute Servers**
504 Compute Servers
  - Nehalem Dual quad-core 2.66GHz
  - 24GB RAM, 500GB Disk
Totals
  - 4032 Cores, 40TF Peak
  - 12TB Memory, 250TB Disk

**Active Storage Servers**
200 Compute/Storage Nodes
  - 40TB SSD Storage
  - 9.6TB Memory
  - 1.6PB SATA Storage

**Big Memory Servers**
15 Servers
  - 15TB Memory, 15TB Disk

**GPU Servers**
133 GPU Servers
  - 8.5TB Memory, 133TB Disk
  - 266 Nvidia 2070 GPU cards

**File Servers**
  - (8) (/home) 160TB

**Gateway Nodes**
  - (16)

**ESNet**
  - 10Gb/s

**ANI**
  - 100 Gb/s Future

**QDR InfiniBand**

**Router**

**Aggregation Switch**
NERSC Magellan Hardware

**Compute Servers**
- 720 Compute Servers
  - Nehalem Dual quad-core 2.66GHz
  - 24GB RAM, 500GB Disk
- Totals
  - 5760 Cores, 40TF Peak
  - 21TB Memory, 400 TB Disk

**Flash Storage Servers**
- 10 Compute/Storage Nodes
  - 8TB High-Performance FLASH
  - 20 GB/s Bandwidth

**Big Memory Servers**
- 2 Servers
  - 2TB Memory

**Archival Storage**

---

**Global Storage (GPFS)**
- 1 PB

**ESNet**
- 10Gb/s

**ANI**
- 100 Gb/s

**Future**

---

**QDR InfiniBand**

**IO Nodes (9)**

**Mgt Nodes (2)**

**Gateway Nodes (27)**

**Aggregation Switch**

**Router**

---

**NERSC**

---

**U.S. DEPARTMENT OF ENERGY**

**Office of Science**

---
Early Findings
Based on progress to date
Magellan Research Agenda and Lines of Inquiry

- Are the open source cloud software stacks ready for DOE HPC science?
- Can DOE cyber security requirements be met within a cloud?
- Are the new cloud programming models useful for scientific computing?
- Can DOE HPC applications run efficiently in the cloud? What applications are suitable for clouds?
- How usable are cloud environments for scientific applications?
- When is it cost effective to run DOE HPC science in a cloud?
- What are the ramifications for data intensive computing?
Cloud Software Stacks

Are the *open source* cloud software stacks ready for DOE HPC science?

• **DOE HPC cluster software stacks**
  – Mature
  – Stable
  – Scalable
  – Depth and breath in tool availability
  – Integrated I/O
  – High performance

• **What about the IaaS cloud software stacks?**
Cloud Software Stacks
Evaluation Process

- Evaluated the top *open source* cloud software stacks
  - All but one were deployed on Magellan
    - OpenNebula evaluation was based on staff code analysis and documentation review as well as evaluations run at CERN and Fermi
  - Evaluation done by staff + special users
    - Test suite with stress tests, scaling tests, etc.
    - Code analysis, documentation review
    - Scientific users running regular workloads and stress test workloads
Cloud Software Stacks
Evaluation Criteria

• Evaluation criteria included
  – Feature Set
  – Stability
  – Infrastructure Scalability
  – Usability
  – Manageability
  – Sustainability

• Evaluation did not include performance
  – Except to note I/O performance challenges
<table>
<thead>
<tr>
<th>Evaluation Area</th>
<th>Eucalyptus 1.6.2</th>
<th>Eucalyptus 2.0</th>
<th>OpenStack</th>
<th>Nimbus</th>
<th>OpenNebula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature Set</td>
<td>Red</td>
<td>Green</td>
<td>External</td>
<td>External</td>
<td>External</td>
</tr>
<tr>
<td>Stability</td>
<td>Red</td>
<td>Green</td>
<td>External</td>
<td>External</td>
<td>External</td>
</tr>
<tr>
<td>Infrastructure Scalability</td>
<td>Red</td>
<td>Yellow</td>
<td>Green</td>
<td>External</td>
<td>External</td>
</tr>
<tr>
<td>Usability</td>
<td>Yellow</td>
<td>Green</td>
<td>External</td>
<td>External</td>
<td>External</td>
</tr>
<tr>
<td>Manageability</td>
<td>Red</td>
<td>Yellow</td>
<td>Green</td>
<td>External</td>
<td>External</td>
</tr>
<tr>
<td>Sustainability</td>
<td>Red</td>
<td>Yellow</td>
<td>Green</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Early Findings:

- Significant improvements in stability and scaling in past year
  - Not production ready yet
- Accounting, monitoring, logging, debugging not at necessary levels
- Networking is complicated and challenging to get right
  - Current architecture bottlenecks performance and scalability

Next Steps:

- Scalability – implement highly distributed infrastructure, integrate new data storage and retrieval module
- Performance – utilize Infiniband for I/O and distributed infrastructure
- Features – provide Infiniband access to users
DOE Cyber Security in the Cloud

Can DOE cyber security requirements be met within a cloud?

- Current cyber security frameworks, architectures and mitigating controls were developed for onsite traditional HPC cluster installations
- Some parallels between clusters and clouds
- But cloud systems provide unique challenges beyond the traditional HPC clusters
  - These require new approaches
- Biggest cyber security risks are with the IaaS cloud model
  - Much of this work was required to deploy the testbeds
## IaaS Cyber Security Overview

**DOE Private Cloud**

<table>
<thead>
<tr>
<th>Defined Risk Areas</th>
<th>Defined Threats</th>
<th>Defined Mitigations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Machine Definition and Management</strong></td>
<td>• User owned, managed, shared Virtual Machine Images (VMI).</td>
<td>• DOE provides secured and approved machine images as a base for user customization.</td>
</tr>
<tr>
<td></td>
<td>• Malicious images shared with users.</td>
<td>• DOE audits user supplied images</td>
</tr>
<tr>
<td></td>
<td>• Encrypted VMIs are opaque to sites</td>
<td></td>
</tr>
<tr>
<td><strong>System Instance Configuration Management</strong></td>
<td>• Users with no system administration experience with full root privileges.</td>
<td>• User education for cyber sec and system administration best practices</td>
</tr>
<tr>
<td></td>
<td>• Relying on users to comply with cyber security best practices and DOE cyber security requirements</td>
<td>• Limit root access for users</td>
</tr>
<tr>
<td></td>
<td>• System level audit data disappears with exit of instance</td>
<td>• Limited system and network based auditing for intrusion and anomaly detection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Develop forensic analysis tools</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• Develop auditing tools for VMs</td>
</tr>
<tr>
<td><strong>Network Authorization and Management</strong></td>
<td>• Users manage the firewall conduits for their machines.</td>
<td>• File and system integrity tools and network access controls implemented to prevent virtual machine cross-talk</td>
</tr>
<tr>
<td></td>
<td>• Potential malicious network activity generated by/from virtual machine instances.</td>
<td>• Constant scanning for bad accounts, bad passwords, open ports</td>
</tr>
</tbody>
</table>
Cyber Security
Early Findings and Next Steps

Early Findings:
• Trust issues
  • User provided VMIs uploaded and shared
  • Root privileges by untrained users opens the door for mistakes
• Network separation is complicated
  • Due to the ephemeral nature of virtual machine instances, an effective Intrusion Detection System (IDS) strategy challenging
• Fundamental threats are the same, security controls are different

Next Steps:
• Can hypervisors play new roles in security monitoring and auditing?
• What sort of forensic analysis could be done on virtual machine instances?
Are the new cloud programming models useful for scientific computing?

- Platform as a Service models have appeared that provide their own Model
  - Parallel processing of large data sets
  - Examples include Hadoop and Azure

- Common constructs
  - MapReduce: map and reduce functions
  - Queues, Tabular Storage, Blob storage
• Bioinformatics using MapReduce
  – Researchers at the Joint Genome Institute have developed over 12 applications written in Hadoop and Pig
  – Constructing end-to-end pipeline to perform gene-centric data analysis of large metagenome data sets
  – Complex operations that generate parallel execution can be described in a few dozen lines of Pig
Programming Models Evaluating Hadoop for Science

- Benchmarks such as Teragen and Terasort
  - Evaluation of different file systems and storage options
- Ported applications to use Hadoop Streaming
  - Bioinformatics, Climate100 data analysis
Programming Models
Early Findings and Next Steps

Early Findings:
• New models are useful for addressing data intensive computing
• Hides complexity of fault tolerance
• High-level languages can improve productivity
• Challenge in casting algorithms and data formats into the new model

Next Steps:
• Evaluate scaling of Hadoop and HDFS
• Evaluate Hadoop with alternate file systems
• Identify other applications that can benefit from these programming models
Can DOE HPC applications run efficiently in the cloud? What applications are suitable for clouds?

• Can parallel applications run effectively in virtualized environments?
• How critical are high-performance interconnects that are available in current HPC systems?
• Are some applications better suited than others?
Application Performance
Application Benchmarks

Runtime relative to Carver

MILC  PARATEC

Carver
Franklin
Lawrencium
EC2-Beta-Opt
Amazon EC2
Application Performance Application Scaling

**PARATEC**

- TCPoIB
- TCPoEth
- AmazonCC

**MILC**

- TCPoIB
- TCPoEth
- AmazonCC

Performance relative to native (IB) vs. Number of cores
Application Performance
Early Findings and Next Steps

Early Findings:

- Benchmarking efforts demonstrate the importance of high-performance networks to tightly coupled applications
- Commercial offerings optimized for web applications are poorly suited for even small (64 core) MPI applications

Next Steps:

- Analyze price-performance in the cloud compared with traditional HPC centers
- Analyze workload characteristics for applications running on various mid-range systems
- Examine how performance compares at larger scales
- Gathering additional data running in commercial clouds
User Experience

How usable are cloud environments for scientific applications?

• How difficult is it to port applications to Cloud environments?
• How should users manage their data and workflow?
User Experience

User Community

- Magellan has a broad set of users
  - Various domains and projects (MG-RAST, JGI, STAR, LIGO, ATLAS, Energy+)
  - Various workflow styles (serial, parallel) and requirements
  - Recruiting new projects to run on cloud environments
- Three use cases discussed today
  - MG-RAST - Deep Soil sequencing
  - STAR – Streamed real-time data analysis
  - Joint Genome Institute
Background: Genome sequencing of two soil samples pulled from two plots at the Rothamsted Research Center in the UK.

Goal: Understand impact of long-term plant influence (rhizosphere) on microbial community composition and function.

Used: 150 nodes for one week to perform one run (1/30 of work planned)

Observations: MG-RAST application is well suited to clouds. User was already familiar with the Cloud
Details

- STAR performed Real-time analysis of data coming from RHIC at BNL
- First time data was analyzed in real-time to a high degree
- Leveraged existing OS image from NERSC system
- Used 20 8-core instances to keep pace with data from the detector
- STAR is pleased with the results
User Experience
JGI on Magellan

• Magellan resources made available to JGI to facilitate disaster recovery efforts
  – Used up to 120 nodes
  – Linked sites over layer-2 bridge across ESnet SDN link
  – Manual provisioning took ~1 week including learning curve
  – Operation was transparent to JGI users
• Practical demonstration of HaaS
  – Reserve capacity can be quickly provisioned (but automation is highly desirable)
  – Magellan + ESnet were able to support remote departmental mission computing
User Experience

Early Findings and Next Steps

Early Findings:
• IaaS clouds can require significant system administration expertise and can be difficult to debug due to lack of tools.
• Image creation and management are a challenge
• I/O performance is poor
• Workflow and data management are problematic and time consuming
• Projects were eventually successful, simplifying further use of cloud computing

Next Steps:
• Gather additional use cases
• Deploy fully configured virtual clusters
• Explore other models to deliver customized environments
• Improve tools to simplify deploying private virtual clusters
Conclusions
Cloud Potential

• Enables rapid prototyping at a larger scale than the desktop without the time consuming requirement for an allocation and account
  – DOE cyber security requirements may block this benefit
• Supports tailored software stacks
• Supports different levels of service
• Supports surge computing
• Facilitates resource pooling
  – But DOE HPC clusters are frequently saturated
Conclusions

Cloud Challenges

• Open source cloud software stacks are still immature, but evolving rapidly
• Current MPI-based application performance can be poor even at small scales due to interconnect
• Cloud programming models can be difficult to apply to legacy applications
• New security mechanisms and potentially policies are required for ensuring security in the cloud
Conclusions

Next Steps

• Characterize mid-range applications for suitability to cloud model
• Cost analysis of cloud computing for different workloads
• Finish performance analysis including IO performance in cloud environments
• Support the Advanced Networking Initiative (ANI) research projects
• Final Magellan Project report
Thank you!

Contact Info:
Shane Canon
Scanon@lbl.gov
magellan.nersc.gov

Susan Coghlan
smc@alcf.anl.gov
magellancloud.org