Science in the Clouds and Beyond

Lavanya Ramakrishnan

Computational Research Division (CRD) & National Energy Research Scientific Computing Center (NERSC)

Lawrence Berkeley National Lab
The goal of Magellan was to determine the appropriate role for cloud computing for science.

- User interfaces/Science Gateways: Use of clouds to host science gateways and/or access to cloud resources through science gateways.
- Hadoop File System
- MapReduce Programming Model/Hadoop
- Cost associativity? (i.e., I can get 10 cpus for 1 hr now or 2 cpus for 5 hrs at the same cost)
- Easier to acquire/operate than a local cluster
- Exclusive access to the computing resources/ability to schedule independently of other groups/users
- Ability to control groups/users
- Ability to share setup of software or experiments with collaborators
- Ability to control software environments specific to my application
- Access to on-demand (commercial) paid resources closer to deadlines
- Access to additional resources

Program Office:

<table>
<thead>
<tr>
<th>Program Office</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced Scientific Computing Research</td>
<td>17%</td>
</tr>
<tr>
<td>Biological and Environmental Research</td>
<td>9%</td>
</tr>
<tr>
<td>Basic Energy Sciences -Chemical Sciences</td>
<td>10%</td>
</tr>
<tr>
<td>Fusion Energy Sciences</td>
<td>10%</td>
</tr>
<tr>
<td>High Energy Physics</td>
<td>20%</td>
</tr>
<tr>
<td>Nuclear Physics</td>
<td>13%</td>
</tr>
<tr>
<td>Advanced Networking Initiative (ANI) Project</td>
<td>3%</td>
</tr>
<tr>
<td>Other</td>
<td>14%</td>
</tr>
</tbody>
</table>
Magellan was architected for flexibility and to support research.

**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**

**Networks and Switches**
- QDR InfiniBand
- IO Nodes (9)
- Mgt Nodes (2)
- Gateway Nodes (27)
- Global Storage (GPFS) 1 PB
- ESNet 10Gb/s
- ANI 100 Gb/s Future
Science + Clouds = ?

Business model for Science
Performance and Cost

Data Intensive Science
Technologies from Cloud
Scientific applications with minimal communication are best suited for clouds

Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud, CloudCom 2010
Scientific applications with minimal communication are best suited for clouds.

Better

Runtime relative to Magellan

- Carver
- Franklin
- Lawrencium
- EC2-Beta-Opt
- Amazon EC2

MILC
PARATEC
The principle decrease in bandwidth occurs when switching to TCP over IB.
Ethernet connections are unable to cope with significant amounts of network connection.
The principle increase in latency occurs for TCP over IB even at mid-range concurrency.
Latency is affected by contention by a greater amount than the bandwidth

- IB
- TCPOIB
- 10G - TCPOEth
- Amazon CC
- 10G- TCPOEth VM
- 1G-TCPoEth

10G VM shows 6X better latency compared to other configurations.

HPCC: RandomRing Latency
Clouds require significant programming and system administration support

- STAR performed Real-time analysis of data coming from Brookhaven Nat. Lab
- First time data was analyzed in real-time to a high degree
- Leveraged existing OS image from NERSC system
- Started out with 20 VMs at NERSC and expanded to ANL.
On-demand access for scientific applications might be difficult if not impossible.

Number of cores required to run a job immediately upon submission to Franklin.
Public clouds can be more expensive than in-house large systems

<table>
<thead>
<tr>
<th>Component</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Systems (1.38B hours)</td>
<td>$180,900,000</td>
</tr>
<tr>
<td>HPSS (17 PB)</td>
<td>$12,200,000</td>
</tr>
<tr>
<td>File Systems (2 PB)</td>
<td>$2,500,000</td>
</tr>
<tr>
<td><strong>Total (Annual Cost)</strong></td>
<td><strong>$195,600,000</strong></td>
</tr>
</tbody>
</table>

Assumes 85% utilization and zero growth in HPSS and File System data. Doesn’t include the 2x-10x performance impact that has been measured. This still only captures about 65% of NERSC’s $55M annual budget.

No consulting staff, no administration, no support.
Cloud is a business model and can be applied to HPC centers

<table>
<thead>
<tr>
<th></th>
<th>Traditional Enterprise IT</th>
<th>HPC Centers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Typical Load Average</td>
<td>30% *</td>
<td>90%</td>
</tr>
<tr>
<td>Computational Needs</td>
<td>Bounded computing requirements – Sufficient to meet customer demand or transaction rates.</td>
<td>Virtually unbounded requirements – Scientist always have larger, more complicated problems to simulate or analyze.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Cloud</td>
<td>HPC Centers</td>
</tr>
<tr>
<td>NIST Definition</td>
<td>Resource Pooling, Broad network access, measured service, rapid elasticity, on-demand self service</td>
<td>Resource Pooling, Broad network access, measured service. Limited: rapid elasticity, on-demand self service</td>
</tr>
<tr>
<td>Workloads</td>
<td>High throughput modest data workloads</td>
<td>High Synchronous large concurrencies parallel codes with significant I/O and communication</td>
</tr>
<tr>
<td>Software Stack</td>
<td>Flexible user managed custom software stacks</td>
<td>Access to parallel file systems and low-latency high bandwidth interconnect. Preinstalled, pre-tuned application software stacks for performance</td>
</tr>
</tbody>
</table>
Science + Clouds = ?

Business model for Science
Performance and Cost

Data Intensive Science
Technologies from Cloud
MapReduce shows promise but current implementations have gaps for scientific applications.

- High throughput workflows
- Scaling up from desktops

**File system:** non POSIX
**Language:** Java
**Input and output formats:** mostly line-oriented text
**Streaming mode:** restrictive
**i/p and o/p model**
**Data locality:** what happens when multiple inputs?
Streaming adds a performance overhead

Better

Evaluating Hadoop for Science, In submission
High performance file systems can be used with MapReduce at lower concurrency

Teragen (1TB)

- HDFS
- GPFS
- Linear (HDFS)
- Expon. (HDFS)
- Linear (GPFS)
- Expon. (GPFS)

Better
Data operations impacts the performance differences
Schemaless databases show promise for scientific applications

www.materialsproject.org
Source: Michael Kocher, Daniel Gunter
Data centric infrastructure will need to evolve to handle large scientific data volumes

Joint Genome Institute, Advance Light Source, etc are all facing a data tsunami
Cloud is a business model and can be applied at DOE supercomputing centers

• Current day cloud computing solutions have gaps for science
  – performance, reliability, stability
  – programming models are difficult for legacy apps
  – security mechanisms and policies

• HPC centers can adopt some of the technologies and mechanisms
  – support for data-intensive workloads
  – allow custom software environments
  – provide different levels of service
Acknowledgements

• US Department of Energy DE-AC02-05CH11232
• Magellan
  – Shane Canon, Tina Declerck, Iwona Sakrejda, Scott Campbell, Brent Draney
• Amazon Benchmarking
  – Krishna Muriki, Nick Wright, John Shalf, Keith Jackson, Harvey Wasserman, Shreyas Cholia
• Magellan/ANL
  – Susan Coghlan, Piotr T Zbiegiel, Narayan Desai, Rick Bradshaw, Anping Liu
• NERSC
  – Jeff Broughton, Kathy Yelick
• Applications
  – Jared Wilkening, Gabe West, Ed Holohan, Doug Olson, Jan Balewski, STAR collaboration, K. John Wu, Alex Sim, Prabhat, Suren Byna, Victor Markowitz