“Midrange computing, and the associated data management play a vital and growing role in advancing science in disciplines where capacity is as important as capability.”

“Demand seems to be limited only by the availability of computational resources.”

“The number of alternative ways for providing these capabilities is increasing.”

Cloud Computing and Midrange

NERSC Policy Board

March 2, 2010

Midrange Computing Sweet Spots

- Serial or scalability-challenged codes
- Science that does not require tight coupling
  - Trivially parallel app, Parameter sweeps, Monte Carlo methods
- Science that can run at low-concurrency
  - 2D v. 3D, different scales for different steps, parameter validation
- On-ramp to the large centers
  - Training, code development, staging
- Data-intensive science
  - Includes Real-time, Visualization

Issues in Midrange Computing

- Lack of dependable, multi-year funding
- Infrastructure limits
- Hidden costs
- Limited expertise
- Limited energy efficiency
- Unable to reach economies of scale
- Data management processes
Why Clouds for Science?

• More than just “cheap” cycles…
• On-demand access to compute resources
  – e.g. Cycles from a credit card. Avoid batch
    wait times., Bypass allocations process.
• Overflow capacity to supplement existing
  systems
  – e.g., Berkeley Water Center has analysis
    that far exceeds the capacity of desktops
• Customized and controlled environments
  – e.g. Supernova Factory codes have
    sensitivity to OS/compiler version
• Parallel programming models for data
  intensive science
  – e.g., BLAST on Hadoop
• Create scientific communities around data sets
  – e.g. DeepSky provides a “Google Maps” for
    astronomical data

Magellan Research Agenda

• What part of DOE’s midrange computing
  workload can be served economically by a
  commercial or private-to-DOE cloud?
• What are the necessary hardware and
  software features of a science-focused
  cloud and how does this differ from
  commercial clouds or supercomputers?
• Do emerging cloud computing models
  (e.g. map-reduce, distribution of virtual
  system images, software-as-a-service)
  offer new approaches to the conduct of
  midrange computational science?
• Can clouds at different DOE-SC facilities
  be federated to provide backup or
  overflow capacity?
## Mid-range codes on Amazon EC2

<table>
<thead>
<tr>
<th>Code</th>
<th>Slow down factor</th>
</tr>
</thead>
<tbody>
<tr>
<td>FMMSpeed</td>
<td>1.3 to 2.1</td>
</tr>
<tr>
<td>GASBOR</td>
<td>1.12 to 3.67</td>
</tr>
<tr>
<td>ABINT</td>
<td>1.11 to 2.43</td>
</tr>
<tr>
<td>HPCC</td>
<td>2.8 to 8.8</td>
</tr>
<tr>
<td>VASP</td>
<td>14.2 to 22.4</td>
</tr>
<tr>
<td>IMB</td>
<td>12.7 to 15.79</td>
</tr>
</tbody>
</table>

- Lawrencium Cluster – 64 bit/Dual sockets per node/8 cores per node/16GB memory, Infiniband interconnect
- EC2 – 64 bit/2 cores per node/75GB, 15GB and 7GB memory, Laboratory Research Computing (LRC)

## NERSC SSP on Amazon EC2

<table>
<thead>
<tr>
<th>Codes</th>
<th>Science Area</th>
<th>Algorithm Space</th>
<th>Configuration</th>
<th>Slow-down</th>
<th>Reduction factor (SSP)</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAM</td>
<td>Climate (BER)</td>
<td>Navier Stokes CFD</td>
<td>200 processors Standard IPCC5 D-Mesh resolution</td>
<td>3.65</td>
<td>0.33</td>
<td>Could not complete 240 proc run due to transient node failures. Some I/O and small messages</td>
</tr>
<tr>
<td>MILC</td>
<td>Lattice Gauge Physics (NP)</td>
<td>Conjugate gradient, sparse matrix; FFT</td>
<td>Weak scaled: 14^d lattice on 8, 32, 64, 128, and 256 processors</td>
<td>2.83</td>
<td>0.35</td>
<td>Erratic execution time</td>
</tr>
<tr>
<td>IMPACT-T</td>
<td>Accelerator Physics (HEP)</td>
<td>PIC, FFT component</td>
<td>64 processors, 64x128x128 grid and 4M particles</td>
<td>4.55</td>
<td>0.22</td>
<td>PIC portion performs well, but 3D FFT poor due to small message size</td>
</tr>
<tr>
<td>MAESTRO</td>
<td>Astrophysics (HEP)</td>
<td>Low Mach Hydro; block structured-grid multiphysics</td>
<td>128 processors for 128^3 computational mesh</td>
<td>5.75</td>
<td>0.17</td>
<td>Small messages and all-reduce for implicit solve.</td>
</tr>
</tbody>
</table>
Performance Comparison of Hadoop and Task Farming

- Evaluated small-scale BLAST problem (2500 sequences) on multiple platforms (Limited by access and costs)
- Similar per-core performance across platforms

The Dark Side of Clouds

- Interconnect suitable only for loosely coupled applications
- Practical limits to the size of a cluster
- Non-uniform execution times (VM jitter)
- Poor shared disk I/O
- Substantial data storage and I/O costs
- Still self-supported

These issues are not intrinsic to clouds, only current implementations.
Cloud Computing and Midrange

NERSC Policy Board

March 2, 2010

Magellan Cloud
Purpose-built for Science Applications

720 nodes, 5760 cores in 9 Scalable Units (SUs) ➔ 61.9 Teraflops
SU = IBM iDataplex rack with 640 Intel Nehalem cores

- SU
- SU
- SU
- SU
- SU
- SU
- SU
- SU
- SU

QDR IB Fabric

- 18 Login/network nodes
- 10G Ethernet
- Network
- Login
- I/O
- 8G FC
- I/O
- 14 I/O nodes (shared)
- HPSS (15PB)
- ANI
- 100-G Router
- NERSC Global Filesystem
- Internet
- Load Balancer

Science-oriented Features of Magellan

- Node aggregation into virtual clusters (as opposed to node virtualization into independent systems)
- Provisioning of full, virtual private clusters for individual research projects
- Dynamic provisioning of multiple software environments
- High bandwidth, low-latency interconnect (InfiniBand QDR)
- Global file system, shared with other NERSC systems
- Access NERSC’s large tape archive for bulk storage of scientific data
Cloud Computing and Midrange

Key is flexible and dynamic scheduling of resources

- Runtime provisioning of software images
- Rolling upgrades can improve availability
- Ability to schedule to local or remote cloud for most cost effective cycles

Magellan Cluster

Portable, Personalized Software Environments

Images (queues, libraries, compilers, tools) pre-configured by NERSC; customized to project

NERSC Policy Board
Science Gateways

- Create scientific communities around data sets
  - NERSC HPSS, NGF accessible by broad community for exploration, scientific discovery, and validation of results
  - Increase value of existing data

- Science gateway: custom hardware, software to provide remotely data/computing services
  - Deep Sky – “Google-Maps” for astronomical image data
    - Discovered 36 supernovae in 6 nights during the PTF Survey
    - 15 collaborators worldwide worked for 24 hours non-stop
  - GCRM – Interactive subselection of climate data (pilot)
  - Gauge Connection – Access QCD Lattice data sets
  - Planck Portal – Access to Planck Data

- New models of computational access
  - Work with large data remotely. Just in time sub-selection from unwieldy data sets.
  - Manipulating streams of jobs, data and HPC workflows through canned interfaces
  - Outreach - Gateways bring HPC apps to those familiar with the web but not the command line

Deep Sky Science Gateway

**Objective:** Pilot project to create a richer set of compute- and data-resource interfaces for next-generation astrophysics image data, making it easier for scientists to use NERSC and creating world-wide collaborative opportunities.

**Implications:** Efficient, streamlined access to massive amounts of data – some archival, some new – for broad user communities.

**Accomplishments:** Open-source Postgres DBMS customized to create Deep Sky DB and interface: www.deepskyproject.org
- 90TB of 6-MB images stored in HPSS / NGF (biggest NGF project now)
  - images + calibr. data, ref. images, more
  - special storage pool focused on capacity not bandwidth
- Like “Google Earth” for astronomers

Map of the sky as viewed from Palomar Observatory; color shows the number of times an area was observed.
Scientific Impact of Deep Sky

We have published several results in the Gamma Ray Bursts Coordinates Network Circulars and in the Astronomer’s Telegrams on the discovery (or limiting brightness) for many host galaxies of GRB’s and/or supernovae.

- First pair instability supernova (SN 2007bi)
- Published in Nature (Dec 2009)
- Result of super-massive star
- DeepSky data (black and triangles) was critical in the observations

“This kind of data-driven approach is key to helping us understand new types of transients for which no reliable theoretical predictions yet exist.”