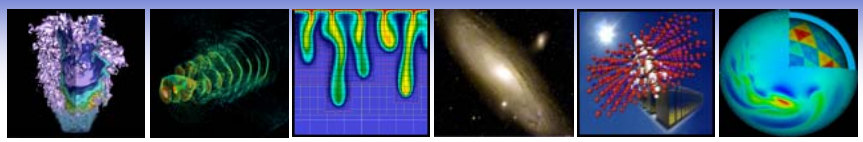


Magellan@NERSC

Jeff Broughton
System Department Head, NERSC

March 10, 2010

DOE Midrange Computing Report

"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."

From: Mid-range Computing in the Support of Science at Office of Science Laboratories. Report of a Workshop, October 2008

2

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

U.S. DEPARTMENT OF ENERGY | Office of Science

3

Issues in Midrange Computing

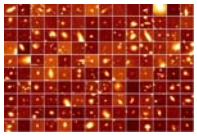
- Lack of dependable, multi-year funding
- Infrastructure limits
- Hidden costs
- Limited expertise
- Limited energy efficiency
- Unable to reach economies of scale
- Data management processes

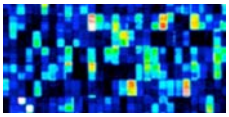
U.S. DEPARTMENT OF ENERGY | Office of Science

4

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





U.S. DEPARTMENT OF ENERGY | Office of Science

5

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?

U.S. DEPARTMENT OF ENERGY | Office of Science

6

Mid-range codes on Amazon EC2

Code	Slow down factor
FMMSpeed. Fast Multipole Method. Pthread parallel code with ½ GB IO	**1.3 to 2.1**
GASBOR. A Genetic algortihm ab initio reconstruction algorithm. Serial workload, minimal I/O (KB)	**1.12 to 3.67**
ABINT. DFT code that calculates the energy, charge density and electronic structure for molecules and periodic solids. Parallel MPI, minimal I/O.	**1.11 to 2.43**
HPCC. HPC Challenge Benchmark	**2.8 to 8.8**
VASP. Simulates property of systems at the atomic scale. MPI parallel application	**14.2 to 22.4**
IMB. Intel (formerly Pallas) MPI Benchmark . Alltoall among all MPI threads	**12.7 to 15.79**

- **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)**

U.S. DEPARTMENT OF ENERGY Office of Science

7

NERSC SSP on Amazon EC2

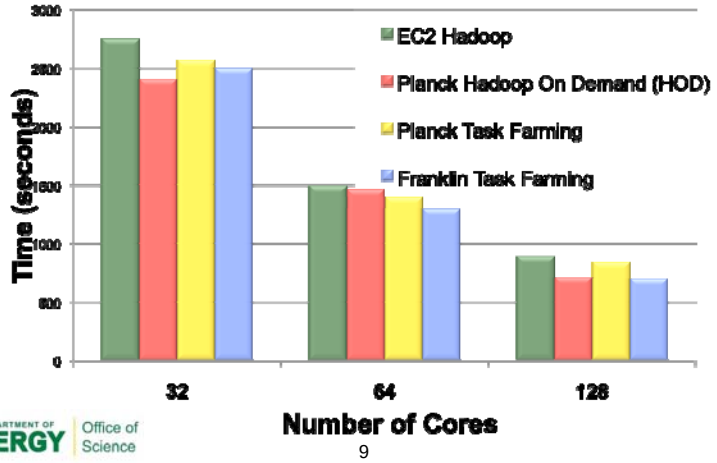
Codes	Science Area	Algorithm Space	Configuration	Slow-down	Reduction factor (SSP)	Comments
				Relative to Franklin		
CAM	Climate (BER)	Navier Stokes CFD	200 processors Standard IPCC5 D-Mesh resolution	3.05	0.33	Could not complete 240 proc run due to transient node failures. Some I/O and small messages
MILC	Lattice Gauge Physics (NP)	Conjugate gradient, sparse matrix; FFT	Weak scaled: 14^4 lattice on 8, 32, 64, 128, and 256processors	2.83	0.35	Erratic execution time
IMPACT-T	Accelerator Physics (HEP)	PIC, FFT component	64 processors, 64x128x128 grid and 4M particles	4.55	0.22	PIC portion performs well, but 3D FFT poor due to small message size
MAESTRO	Astrophysics (HEP)	Low Mach Hydro; block structured-grid multiphysics	128 processors for 128^3 computational mesh	5.75	0.17	Small messages and all-reduce for implicit solve.

U.S. DEPARTMENT OF ENERGY Office of Science

8

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



9

The Dark Side of Clouds

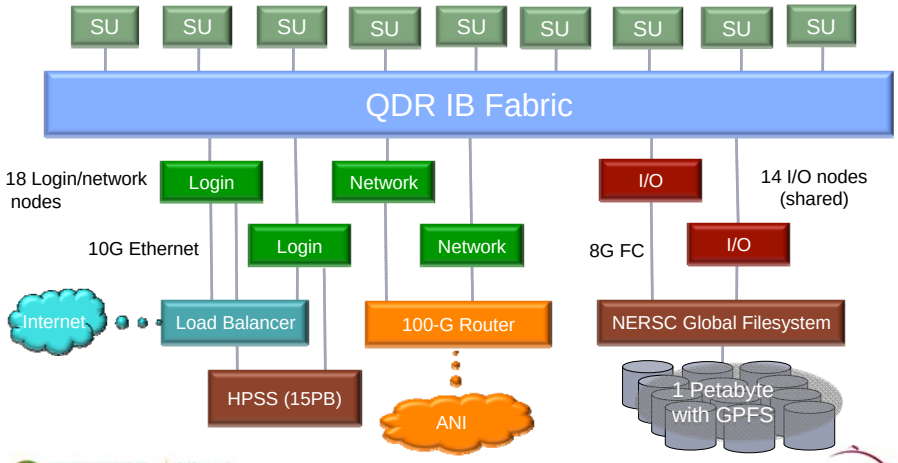
- Interconnect suitable only for loosely coupled applications
- Practical limits to the size of a cluster
- Non-uniform execution times (VM jitter)
- Poor shared disk I/O
- Substantial data storage and I/O costs
- Still self-supported

These issues are not intrinsic to clouds, only current implementations.

10

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



11

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

12

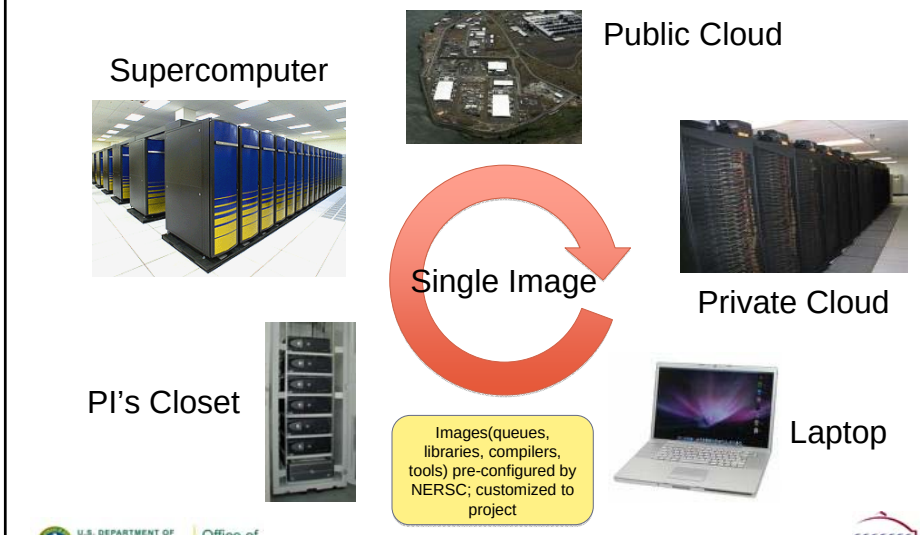
Key is flexible and dynamic scheduling of resources

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

U.S. DEPARTMENT OF ENERGY | Office of Science

13

Portable, Personalized Software Environments



U.S. DEPARTMENT OF ENERGY | Office of Science

14

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**

U.S. DEPARTMENT OF ENERGY | Office of Science

15

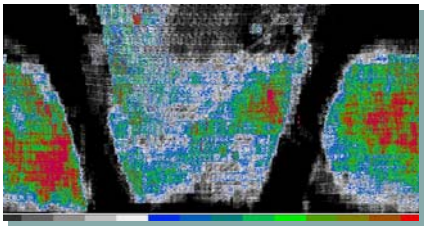
Deep Sky Science Gateway

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

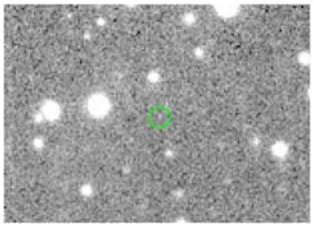


Map of the sky as viewed from Palomar Observatory; color shows the number of times an area was

U.S. DEPARTMENT OF ENERGY | Office of Science

16

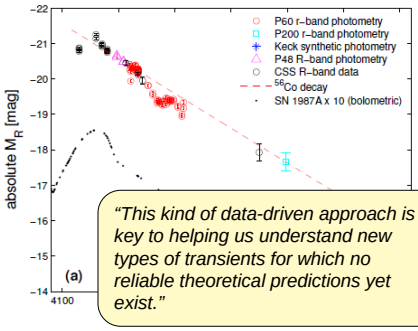
Scientific Impact of Deep Sky



GRB 071112C

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.

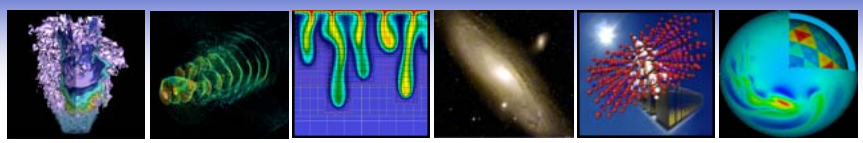
- First pair instability supernova (SN 2007bi)
- Published in Nature (Dec 2009)
- Result of super-massive star
- DeepSky data (black and triangles) was critical in the observations



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

U.S. DEPARTMENT OF ENERGY Office of Science

17



Accelerating Scientific Discovery

U.S. DEPARTMENT OF ENERGY Office of Science