Extreme Scale Data Intensive Computing at NERSC

Harvey Wasserman
NERSC User Services Group
Los Alamos Computer Science Symposium
Workshop on Performance Analysis of Extreme-Scale Systems and Applications

October 14, 2009
Outline

• The need: project examples
  – Current & potential future

• The response: architecture and methods

• Results

• N.B., Many additional projects at LBNL
Data Driven Science

- Ability to generate data is challenging our ability to store, analyze, & archive it.
  - Some observational devices grow in capability with Moore’s Law.
  - Data sets are growing exponentially.
- Petabyte (PB) data sets soon will be common:
  - Climate: next IPCC estimates 10s of PBs
  - Genome: JGI alone will have .5 PB this year and double each year
  - Particle physics: LHC projects 16 PB / yr
  - Astrophysics: LSST, others, estimate 5 PB / yr
- Redefine the way science is done?
  - One group generates data, different group analyzes
- Turning point: in 2003 NERSC changed from being a data source to a data sink
• Data mining: process of extracting hidden patterns from data
  – de novo genome assembly
  – Analysis of cosmological observations
  – Combine various DBs (protein/genome)

• Data-Intensive Predictive Science: simulations that generate lots of data

• Overarching need: fast I/O but not just BW

Nick Wright (SDSC/NERSC)
Intro to NERSC

• National Energy Research Scientific Computing Center
• Production computing for all DOE Office of Science (SC) research.
  • ~ 2,000 users

• DOE allocated ~225M hours for ~370 projects at NERSC for 2010
  – ~ 50% of what users requested
  – Plus ~ 56M from SC, NERSC reserve
  – Plus ~ 60M “Storage Resource Units”
<table>
<thead>
<tr>
<th>Project</th>
<th>Category</th>
<th>Compute Hours</th>
<th>Storage RUs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Supernovae Factory</td>
<td>HEP/ Astro</td>
<td>14k</td>
<td>1.8M</td>
</tr>
<tr>
<td>Palomar Transient Factory</td>
<td>HEP/ Astro</td>
<td>36k</td>
<td>1M</td>
</tr>
<tr>
<td>ALICE</td>
<td>NP/ Astro</td>
<td>10k</td>
<td>2.2M</td>
</tr>
<tr>
<td>CCSM</td>
<td>Climate</td>
<td>12M</td>
<td>2M</td>
</tr>
<tr>
<td>STAR</td>
<td>Nuclear Physics</td>
<td>-</td>
<td>8M</td>
</tr>
<tr>
<td>CMB: PLANCK +</td>
<td>HEP/ Astro</td>
<td>680k</td>
<td>500k</td>
</tr>
<tr>
<td>20th Century ReAnalysis</td>
<td>Climate</td>
<td>8M</td>
<td>4M</td>
</tr>
<tr>
<td>John Bell</td>
<td>Chem/Comb/Math</td>
<td>5.5M</td>
<td>7.5M</td>
</tr>
<tr>
<td>Lattice QCD</td>
<td>NP</td>
<td>1.4M</td>
<td>2M</td>
</tr>
<tr>
<td>PCMDI</td>
<td>Climate</td>
<td>20k</td>
<td>2M</td>
</tr>
<tr>
<td>KamLAND</td>
<td>NP / Astro</td>
<td>-</td>
<td>4M</td>
</tr>
<tr>
<td>JGI</td>
<td>Biological Science</td>
<td>10k</td>
<td>2M</td>
</tr>
</tbody>
</table>
Objective: Analyze data from the Planck satellite -- definitive Cosmic Microwave Background (CMB) data set.

Accomplishments: NERSC provides the components of the data pipeline for noise reduction, map-making, power spectrum analysis, and parameter estimation
- 2006 Nobel Prize in Physics
- 32 TB final data set size, ~400 users
- data sets analyzed as a whole because complex data correlations
- Extensive use of NGF / PDSF
- Launched May09, first “light” Sept09
- Also ~10k-core Cray XT4 MonteCarlo calibration runs, produce ~10X data
- Anticipate Moore’s law growth in data set size for 15 years

Implications: CMB: image of the universe at 400k years, relic radiation from Big Bang

PI: J. Borrill (LBNL)

https://c3.lbl.gov/
• Parallel Distributed Systems Facility
  – Heterogeneous commodity Linux cluster
    • GigE, I/B, several OSs, several CPUs
  – Open Science Grid
  – “Sub” clusters, data vaults for experiments
  – Funding comes from NERSC, NP and HEP
• NERSC Global File System (.45PB -> ~ 1PB)
  – Common global filesystem for all NERSC systems
  – GPFS
  – Extremely stable (zero unscheduled outages past two years)
Objective: Archive, analyze all stages of the US data from Kamioka Liquid Scintillator Anti-Neutrino Detector

Implications: Substantially increase our scientific knowledge of neutrinos

Accomplishments: Many significant physics milestones – neutrino oscillation, precise value for the neutrino oscillation parameter, etc.

- NERSC resources instrumental in reactor neutrino analysis and the preparations for the solar phase;
- Currently recording data at trigger rate of 100Hz, data rate of 200GB/day, 365 days/yr
- 0.6 PB of data stored from 6 years; plan to read large fraction of this in 2010

http://kamland.lbl.gov/
**Objective:** Data analysis and simulations for the ALICE heavy-ion detector experiment at the LHC.

**Implications:** Understanding of dense QCD matter.

**Notes:** Uses (primarily) NERSC’s PDSF cluster + LLNL + Grid resources;
- Expect ~600TB of data distributed over 1GB files, ~25% of USA obligation in 2010.
- Challenge of providing direct-charged resources for experimentation that might be delayed.
- Simulation resources to reconstruct and analyze detector events prior to the experiment.
- Longer term: Estimate 3.8 PB of disk space and 5.31 PB of HPSS in 2013, accessible by international community.

[ALICE](http://aliceinfo.cern.ch/Collaboration/)
Objective: Compare forecasts made by global climate models (GCMs) with varying initial conditions based on detailed observations to assess GCM accuracy.

Implications: Improved climate prediction; support for IPCC.

Accomplishments: Archived European observational data at NERSC mass storage facility;
- Extensive CAM runs on Franklin;
- New, very high resolution (~100 meters) Large Eddy Simulation (LES) model to be added in 2010.
- LES results in 2X storage increase
- Small time scale (~20 min) produces many files

PI: C. Covey (LLNL)

https://ccpp.llnl.gov/
**Objective:** Climate models that fully resolve key convective processes in clouds; ultimate goal is 1-km resolution.

**Implications:** Major transformation in climate/weather prediction, likely to be standard soon, just barely feasible now.

**Accomplishments:** Developed a coupled atmosphere-ocean-land model based on geodesic grids. 
- Multigrid solver scales perfectly on 20k cores of Franklin using grid with 167M elements.
- Invited lecture at SC09.

**NERSC:**
- 3-km 24-hr run, 30k cores = 10TB output
- NERSC/LBNL played key role in developing critical I/O code & Viz infrastructure to enable analysis of ensemble runs and icosahedral grid.
Objective: Archive all production and R&D data from three sequencing platforms at JGI

Implications: One of the world's largest public DNA sequencing facilities.

Accomplishments: NERSC, JGI staff collaborated to set up nightly back-up pipeline using ESnet’s new Bay Area MAN.
- Archiving sequencing data at NERSC allowed JGI to scale up infrastructure with minimal additional DOE investment.
- Data import expected to grow nearly exponentially in 2010; impossible to maintain data onsite at the JGI HQ.
- NERSC/DOE JGI collaboration to develop improved techniques for data access, handling.
- Note: additional Microbial Genome project

JGI is producing sequence data at increasing rate: 2 million files per month of trace data (25 to 100 KB each) plus 100 assembled projects per month (50 MB to 250 MB); total about 2 TB per month on average.
**Objective:** Process, analyze & make available data from Palomar Transient Sky survey (~300 GB / night) to expose rare and fleeting cosmic events.

**Implications:** First survey dedicated solely to finding transient events.

**Accomplishments:** Automated software for astrometric & photometric analysis and real-time classification of transients.

- Analysis at NERSC is fast enough to reveal transients as data are collected.
- Has already uncovered more than 40 supernovae explosions since Dec., 2008.
- Uncovering a new event about every 12 minutes.

- 40k MPP allocation + 1M HPSS in 2009; Stored on NERSC’s 450-TB NGF + gateway (other slide)

Two manuscripts submitted to Publications of the Astronomical Society of the Pacific
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 remote data/computing services
    • 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
  – Projects with mission-critical time constraints require guaranteed turn-around time.
  – Reservations for anticipated needs: Computational Beamlines
  – Friendly interfaces for applications and workflows
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?

PI: C. Aragon (NERSC)

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

See Peter Nugent’s NUG2009 Talk

- Other NERSC gateways: GCRM (climate), Planck (Astro), Gauge Connection (QCD)
**Objective**: Create & mine a database of molecular dynamics structures to identify similarities between native and unfolded states across all secondary and tertiary structure types and sequences.

**Implications**: Improved protein structure prediction algorithms by identifying patterns and general features of transition, intermediate and denatured states.

**Accomplishments**: To date, performed more than 6,000 simulations of nearly a 1,000 proteins for a combined simulation time of >140 microseconds.

- Continued data mining to identify similarities / differences between native and unfolded states across all 2\(^\circ\) and 3\(^\circ\) structure types and sequences.
- Expect repository similar to Protein DB, 100+ TB relational database.

The first 156 dynameomics simulation targets

http://www.dynameomics.org

**PI**: V. Daggett (U. Wash.)
Observations

• It’s not just about providing tapes / disk / fiber
  – It’s about organization & intelligent, secure, public access using modern tools
• Simulation output becomes too large to move “home.”
  – However, some science groups lack agreement on how much data needs to be available and where
• Kathy Yelick: Tape archives, vital to efficient science, use 2-3 orders of magnitude less power than disk
• Beyond “enormous growth,” precise requirements sometimes elusive
  – NERSC XT4 increased HPSS use 50% - why?
  – Data needs linked to machine reliability
  – Observational projects easier to characterize storage needs?
• Value of data varies: observations may be irreplaceable; rarely touched; processing raw data may result in 10X larger volume
• Manipulation and analysis of data is becoming a problem that can be addressed only by large HPC systems.
• Few projects are purely simulation or observation.
• Fast I/O is key
NERSC Response

• Upgrade I/O capability in NERSC-5*
• Increased I/O capability in NERSC-6
• Improved user access in NERSC-5/6
• User support for improved I/O

*NERSC-5 is “Franklin,” Cray XT4
But first…

- **SC09 Masterworks**: Data intensive computing and lots more!
  - [http://sc09.supercomputing.org/?pg=masterworks.html](http://sc09.supercomputing.org/?pg=masterworks.html)
- Talks from Google, Facebook
- Data Challenges in Genome Analysis
- Talks are Tu, Weds, Thurs; Portland Ballroom
• Difficulty is in finding tests that accurately capture the workload but are easy to use
• LBNL CRD research, using IOR to accurately capture I/O in applications (Oliker/Shalf/Borrill/Shan, SC07)
• NERSC-6 procurement approach:
  – One application writes checkpoint files
  – IOR and Metadata kernel tests
• Additional applications: GTC, Flash, S3D, POP
• Vary sizes, method (POSIX / MPI-IO / HDF / netCDF)
• Metrics: % of runtime / % of peak / % of relative peak
• Frequently asked to provide I/O “stress tests”
Congestion in the I/O subsystem had been a major cause of instability, poor achieved I/O rates.

<table>
<thead>
<tr>
<th></th>
<th>Before</th>
<th>After</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Nodes</td>
<td>9,660</td>
<td>9,572</td>
</tr>
<tr>
<td>Login Nodes</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>MOM Nodes</td>
<td>16 (also serve as login nodes)</td>
<td>6 (distinct MOM nodes)</td>
</tr>
<tr>
<td>Lustre OSS / OST</td>
<td>20 / 80 per filesystem</td>
<td>24 / 48 per 2 filesystems</td>
</tr>
<tr>
<td>DVS Nodes</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Filesystems</td>
<td>/scratch</td>
<td>/scratch /scratch2</td>
</tr>
<tr>
<td>Capacity</td>
<td>346 TB</td>
<td>420 TB (210 TB ea.)</td>
</tr>
<tr>
<td>I/O adaptors</td>
<td>PCI</td>
<td>PCI-e</td>
</tr>
<tr>
<td>Peak I/O perf.</td>
<td>~ 12.5 GB/s</td>
<td>~ 30 GB/s</td>
</tr>
</tbody>
</table>
NERSC XT4 I/O Improvement: HW

- 2\textsuperscript{nd} scratch filesystem added
  - reduce I/O congestion among simultaneous user jobs
- Disks better distributed, 2X # of controllers
- Service nodes redistributed.

\begin{itemize}
  \item Indicates that this cabinet contains service nodes
\end{itemize}
NERSC XT4 I/O Improvement: HW

<table>
<thead>
<tr>
<th></th>
<th>BlueGene</th>
<th>Jaguar</th>
<th>FranklinBefore</th>
<th>FranklinAfter</th>
</tr>
</thead>
<tbody>
<tr>
<td>%I/O 16000</td>
<td>24%</td>
<td>15%</td>
<td>29%</td>
<td>5%</td>
</tr>
<tr>
<td>%I/O 4096</td>
<td>7%</td>
<td>9%</td>
<td>7%</td>
<td>1%</td>
</tr>
<tr>
<td>%I/O 512</td>
<td>2%</td>
<td>&lt;1%</td>
<td>1%</td>
<td>&lt;1%</td>
</tr>
</tbody>
</table>

- 25-Mar Harmonic Mean = 4,972 MB/s
- 27-Feb Harmonic Mean = 955 MB/s

S3D aggregate performance, unpublished results

IOR benchmark aggregate read performance
NERSC XT4 I/O Improvement: SW

- Users report < 1 GB/s write bandwidth
- K. Antypas and A. Uselton (NERSC), CUG09
- Identify Sub-optimal MPI-IO implementation
- Study via IOR, Flash, MadBench
- Compare:
  - file-per-proc vs. shared file
  - Lustre block boundary alignment (1e6 vs. $2^{20}$ bytes)
• IPM
  – David Skinner (NERSC), Noel Keen (LBNL), Mark Howison (NERSC)
  – intercept libc open, close, read and write calls
  – http://www.nersc.gov/projects/ipm/

• Lustre Monitoring Tool (Andrew Uselton, NERSC) http://code.google.com/p/lmt
NERSC XT4 I/O Improvement: SW

• Adjust default stripe width to 4 MB (4x)
• Cray revised collective buffering algorithm to issue write calls that respect stripe boundaries
• Set # of writer nodes equal to the number of stripes (via trial & error using IOR)
  – led to an optimal OST assignment; performance on par with file-per-proc
• Result:
  – collective write bandwidths ~ 6.5 GB/s
NERSC XT4 I/O Improvement: SW

Antypas and Uselton, Proc. CUG 2009

FLASHCheckpointRead2048Procs
Improved XT4 Stability

- I/O improvements yield stability improvement

Franklin Outages Nov07 - Jul09 with six-month moving averages

HW 6-mo

SW 6-mo

% of Possible Usage in 24 hours

100

90

80

70

60

50

40

30

20

10

0

## NERSC-6 (Hopper) System

### Phase 1 – Cray XT5
- 668 nodes, 5,344 cores
- 2.4 GHz AMD Opteron (Shanghai, 4-core)
- 50 TF peak
- 5 TF SSP
- 11 TB DDR2 memory total
- Seastar2+ Interconnect
- 2 PB disk, 25 GB/s
- Air cooled

### Phase 2 Cray <??>
- > 6,000 nodes > 150,000 cores
- 12-core AMD Opteron (Magny-Cours)
- > 1 PF peak
- > 100 TF SSP
- > 200 TB DDR3 memory total
- Gemini Interconnect
- 2 PB disk, 80 GB/s*
- Liquid cooled

* measurable, sustained aggregate filesystem I/O bandwidth between the external parallel filesystem and the computational nodes using IOR.

<table>
<thead>
<tr>
<th>3Q09</th>
<th>4Q09</th>
<th>1Q10</th>
<th>2Q10</th>
<th>3Q10</th>
<th>4Q10</th>
</tr>
</thead>
</table>
- **GPU/Accelerator Testbed**
  - Large-memory (.5 TB RAM) with Nvidia Tesla (1 TF) GPU accelerators
  - Experiment with GPU accelerated sequence matching and OpenCL/CUDA programming model
  - Gain experience with administration of this kind of platform

- **Cloud Computing Testbed (NERSC/ANL: Magellan)**
  - Distributed, multi-institution dynamically expandable computing resource
  - Experiment with cost effectiveness of cloud computing paradigm, including Amazon EC2 evaluation

- **Solid State/FLASH Accelerated I/O**
  - Next slide

- **FPGA Accelerator Testbed (LBL Computing Research Division)**
  - Convey HC1 FPGA accelerator with 80GB/s vector memory subsystem: can be programmed with “custom personalities” for, e.g., bioinformatics applications
DOE Explores Cloud Computing

- **ASCR Magellan Project**
  - $32M project at NERSC and ALCF
  - ~100 TF/s compute cloud testbed (across sites)
  - Petabyte-scale storage cloud testbed

- **Cloud questions to explore on Magellan:**
  - Can a cloud serve DOE’s mid-range computing needs?
    → More efficient than cluster-per-PI model
  - What part of the workload can be served on a cloud?
  - What features (hardware and software) are needed of a “Science Cloud”? (Eucalyptus at ALCF; Linux at NERSC)
  - How does this differ, if at all, from commercial clouds?
Flash Storage Testbeds

- **~ 10TB in NERSC Global Filesystem (NGF)**
  - Metadata acceleration
  - High bandwidth, low-latency storage class

- **~ 16TB as local SSD in one ScalableUnit (~7 TF) of new “Magellan” cloud testbed**
  - Data analytics
  - Local read-only data
  - Local temp storage

- **~ 2TB in HPSS (metadata acceleration)**
Other NERSC Efforts

• Increase in I/O Bandwidth for GCRM project
  – Mark Howison (NERSC), PNNL
  – recently achieved aggregate write bandwidth of 5 GB/sec on XT4

• HDF5 I/O Library Performance Analysis, Optimization, and HDFPart API layer
  – Mark Howison, John Shalf (NERSC)
  – See NUG talk, Oct2009
Acknowledgments

• Shreyas Cholia, Katie Antypas, Andrew Uselton, John Shalf, Rei Lee, Mark Howison, Prabhat, Hongzhang Shan, Akbar Mokhtarani, Helen He, Janet Jacobsen
• Shane Canon, NERSC Data Systems Group Lead
• John Shalf, NERSC SDSA Group Lead
• Wes Bethel, NERSC Analytics Group Lead
• David Skinner, NERSC SW Integration Group Lead
• Brent Draney, NERSC Networking Group Lead

• http://www.nersc.gov
NERSC 2009 Configuration

Large-Scale Computing System
Franklin (NERSC-5): Cray XT4
  • 9,740 nodes; 38,288 Opteron cores,
  • 8 GB of memory per node
  • 26 Tflop/s sustained SSP (355 Tflops/s peak)
NERSC-6 (XT5) planned for 2010 production
  • 3-4x NERSC-5 in application performance

Clusters
  • Bassi IBM Power5 (888 cores)
  • Jacquard LNXI Opteron (712 cores)
  • New Nehalem / IB Cluster
  • PDSF (HEP/NP)
  • Linux cluster (~1K cores)

NERSC Global Filesystem (NGF)
  400 TB; 5.5 GB/s

HPSS Archival Storage
  • 60 PB capacity
  • 10 Sun robots
  • 130 TB disk cache

Analytics / Visualization
  • Davinci (SGI Altix)
• A facility-wide, high performance, parallel file system
  – Uses IBM’s GPFS technology for scalable high performance
  – Designed for user productivity
• A facility-wide, high performance, parallel file system
  – Uses IBM’s GPFS technology for scalable high performance
  – Designed for user productivity
**Objective:** Use an Ensemble Kalman filter to reconstruct global weather conditions in six-hour intervals from 1871 to the present.

**Implications:** Validate tools for future projections by successfully recreating – and explaining – climate anomalies of the past.

**Accomplishments:** First complete database of 3-D global weather maps for the 19th to 21st centuries.
- Provide missing information about the conditions in which extreme climate events occurred.
- Reproduced 1922 Knickerbocker storm, comprehensive description of 1918 El Niño
- Data can be used to validate climate and weather models

---

**PI:** G. Compo (U. Colorado)