NERSC Role in HEP and Research and Emerging Technologies

Sudip Dosanjh
Director

November 27, 2012
Career History

• **1980:** Summer Intern at LBL
• **1977-1986:** U.C. Berkeley student
• **1986-2012:** Sandia National Labs
  - Modeled Three Mile Island on Cray YMPs
  - Massively parallel computing (chemically reacting flows, material science, computational science, algorithms)
  - Computational Science and Applications
  - Extreme-scale Computing
    - Exascale
    - Co-design
    - Computer architectures
    - Algorithms

---

**THE INFLUENCE OF TURBULENCE ON EROSION BY A PARTICLE-LADEN FLUID JET**

SUDIP DOSANJH and JOSEPH A. C. HUMPHREY
Materials and Molecular Research Division, Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720 (U.S.A.)
(Received July 2, 1984; accepted February 7, 1985)

Wear, 102 (1985) 309 - 330
NERSC Provides Computing for Science

Broad user community
• 4844 users, 663 projects
• 48 states; 65% from universities
• Hundreds of users each day
• ~1500 publications per year

Systems for science
• 1.3PF Hopper + .5 PF clusters
• Services for consulting, data analysis and more
NERSC Has a Broad Range of Computational Problems

Science at Scale
Petaflops to Exaflops

Science through Volume
Thousands to Millions of Simulations

Science in Data
Petabytes to Exabytes of Data
Computational Modeling and Big Data

Large-Scale discovery of Events
- Petascale simulations produce data too large for manual analysis
- Data analysis using new algorithms (FastBit, Machine Learning) discover events

Materials Project
- Tens of thousands of simulations screen materials
- Goal: cut in half the 18 year from design to manufacturing
- Advance machine learning and data systems

Today’s batteries
Voltage limit
Interesting materials…
DOE has Unique Data Challenges

- DOE provides many of the large scale user facilities
- Some are producing Petabytes of data today
- NERSC has about 4 Petabytes of disk and 40 of tape
Vision: Accelerate scientific discovery across a broad community through advanced computing

- **Energy efficient computing**: Improve application performance per Watt by 100x necessary for exascale
- **High throughput computing**: Provide tools and infrastructure for ensemble runs and deliver database of results to science community
- **Data driven computing**: Improve insight through access to and analysis of data from experimental facilities
The Production Facility for DOE SC

• NERSC Focus on unique resources
  – High end computing systems
    • Configured for both large-scale jobs and large numbers of jobs
  – High end storage systems
    • Large shared file system
    • Tape archive
  – Interface to high speed network
    • ESnet 100 Gb/s

• Allocate time / storage
  – Current processor hours and tape storage

% Use in 2011

- BES 33%
- FES 19%
- BER 16%
- HEP 15%
- NP 13%
- ASCR 4%
DOE’s Changing Computing Priorities

Usage by Science Type as a Percent of Total Usage
### Production Computing at NERSC / LBNL
- **100s of Projects**
- **Allocations**
  - 80% divided and allocated by each Science Office
  - 10% ASCR Leadership Computing Challenge
  - 10% Directors’ reserve
- **Limited to DOE-relevant science**
- **Includes storage and computing allocations**

### Leadership Computing at ANL and ORNL
- **10s of projects**
- **Allocations**
  - 60% by INCITE program managed by ANL/ORNL
  - 30% ASCR Leadership Computing Challenge
  - 10% Director’s reserve
- **Includes industry and non-DOE applications**
- **Focused on applications at scale**
NERSC is Very Cost Effective Relative to Clouds

<table>
<thead>
<tr>
<th>Component</th>
<th>Annual Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute Systems (1.38B hours)</td>
<td>$181M</td>
</tr>
<tr>
<td>HPSS (17 PB)</td>
<td>$12M</td>
</tr>
<tr>
<td>File Systems (2 PB)</td>
<td>$3M</td>
</tr>
<tr>
<td><strong>Total (Annual Cost)</strong></td>
<td><strong>~$200M</strong></td>
</tr>
</tbody>
</table>

NERSC cost/core hours dropped 10x (1000%) from 2007 to 2011
Amazon pricing dropped 15% in the same period

These are “list” prices, which overestimate cloud costs, but several factors underestimate the cost:
- Doesn’t include the measured performance slowdown 2x-50x.
- Only accounts for about 65% of NERSC’s $57M annual budget.

No consulting staff, no account management, no software support.
Current NERSC Systems

Large-Scale Computing Systems
Hopper (NERSC-6): Cray XE6
• 6,384 compute nodes, 153,216 cores
• 144 Tflop/s on applications; 1.3 Pflop/s peak

Edison (NERSC-7): Cray Cascade
• To be delivered in 2013
• Over 200 Tflop/s on applications, 2 Pflop/s peak

Midrange
140 Tflops total
Carver
• IBM iDataplex cluster
• 9884 cores; 106TF
PDSF (HEP/NP)
• ~1K core cluster
GenePool (JGI)
• ~5K core cluster
• 2.1 PB Isilon File System

NERSC Global Filesystem (NGF)
Uses IBM’s GPFS
• 8.5 PB capacity
• 15GB/s of bandwidth

HPSS Archival Storage
• 240 PB capacity
• 5 Tape libraries
• 200 TB disk cache

Analytics & Testbeds
Euclid
(512 GB shared memory)
Dirac 48 Fermi GPU nodes
Magellan Hadoop
Limitations of Existing Programming Models

- We can run 1 MPI process per core, but there are problems with 6-12+ cores/socket:
  - Insufficient memory: user level data and internal buffers
  - Runtime overheads: copying and synchronization
- OpenMP, Pthreads, or other shared memory models
  - No control over locality, e.g., Non-Uniform Memory Access
  - No explicit memory movement, e.g., accelerators or NVRAM
- Even on petascale systems, tuning is non-obvious

Nick Wright, John Shalf et al, NERSC/Cray Center of Excellence
• NERSC performance has traditionally grown at 10x every 3-4 years
NERSC-7 Coming Soon

• NERSC will install a Cray “Cascade” system in 2013
  – First all new Cray design since Red Storm; developed for the DARPA HPCS program (including >$70M from DOE)
  – Intel Processors with >2PF peak performance
  – New “Aries” interconnect using a “dragonfly” topology
  – 6.5PB storage using Cray Sonexion Lustre appliances

• Good match for diverse NERSC user needs
  – Both High-throughput and high-concurrency workloads.

• Excellent energy efficiency
  – Allows chiller-less “free cooling” with only 10% “overhead”

• Will deliver ~1B Hopper-equivalent core hours
  – 18
UC’s Computational Research and Theory (CRT) Facility

- Unique energy efficient design from weather / hillside
- Collaborative space for 300
- $124M UC Project (up $12M)
- $20M DOE Project
- 100 MW at Berkeley Lab and space for 2 exascale systems
NERSC Plan Will Help Take Science through Technology Transition

NERSC criteria in procurements is application performance, not peak

NERSC will transition users (once) to manycore architectures

NERSC transitioned users from vector supercomputers to massive parallelism
NERSC-8 Plans

Goals:

- 10x-50x increase in application performance over Hopper
- Transition to energy-efficient architectures
- High applications performance per watt
- Most energy efficient machine in most energy efficient facility

Plans:

- Production HPC resources for 2015/2016.
- Transition to new energy-efficient architectures on road to exascale
- Collaborate with Trinity/ACES to share expertise, reduce risk, and strengthen SC/NNSA alliance on road to exascale
Technology Challenges and Strategies
Power Limits Computing Performance Growth

- Power density limits single processor performance
- Strategy: Redesign architecture, memory, software, algorithms for low power and (implied need) resilience

Processor industry running at "maneuvering speed"
- David Liddle

The Expectation Gap
Energy Efficient Computing is Key to Performance Growth

At $1M per MW, energy costs are substantial

- 1 petaflop in 2010 used 3 MW
- 1 exaflop in 2018 would use 130 MW with “Moore’s Law” scaling

This problem doesn’t change if we were to build 1000 1-Petaflop machines instead of 1 Exaflop machine. It affects every university department cluster and cloud data center.
Measuring Efficiency

• One important factor in computing efficiency is utilization

• If we measure productivity by publications...
  – NERSC in 2010 ran at 450 publications per MW-year

• Application performance per Watt
New Processor Designs are Needed to Save Energy

- Cell phone processor (0.1 Watt, 4 Gflop/s)
- Server processor (100 Watts, 50 Gflop/s)

- Server processors have been designed for performance, not energy
  - Graphics processors are 10-100x more efficient
  - Embedded processors are 100-1000x
  - Need manycore chips with thousands of cores
Where does the Power Go?

Intranode/SMP Communication

Intranode/MPI Communication

On-chip / CMP communication

Pic Joules

DP FLOP  Register  1mm on-chip  5mm on-chip  Off-chip/DRAM  local interconnect  Cross system

[Graph showing power consumption across different components and time periods (2018 vs. now).]
The Roofline Performance Model: Understanding Communication Limits

The flat room is determined by arithmetic peak and instruction mix.

The sloped part of the roof is determined by peak DRAM bandwidth (STREAM).

X-axis is the computational intensity of your computation.
**Exascale Programming: Memory System Structure**

**Known:** Communication wall will get worse;
- Optimizing for memory/network more important than ever
- Automatic data movement (caches, VM) can be wasteful
- Autotuning (search) helps reach bandwidth limits

**Unknown:**
- How much explicit memory be management?
What is Manycore?

• NVIDIA, AMD/ATI, Intel MIC, are all Manycore processors

• Case for manycore
  – Many small cores are needed for energy efficiency and power density; could have their own PC or use a wide SIMD
  – May need at least one fat core (heterogeneity) for running the OS, etc.

• Local store, explicitly managed memory hierarchy
  – More efficient (get only what you need) and simpler to implement in hardware

• Co-Processor interface and PCI between CPU and Accelerator
  – Market: GPUs are separate chips for specific domains
  – Hoping this will go away

• Transition at NERSC-8, not NERSC-7
NERSC’s Computing Strategy

• Two major systems on the floor in steady state
  – Maximize stability and usability rather than peak flops

• Optimization for application performance not peak
  – Procurements done using application benchmarks

• Balance computing with growth in data services
  – Disk, tape, network, data transfer nodes, gateways

• Provide for large jobs and large numbers of jobs
  – Both full OS support and lightweight OS

• Minimize number of technology transitions
  – Need to move to manycore is necessary
  – Transition programming model once and choose carefully
Requirements Gathering Ensures NERSC Meets DOE Needs

How we use your input

- Communicate science needs and impact with case studies
- Direct input into Mission Need for NERSC-9 and 10
- Inform priorities for computing, storage, infrastructure
- Inform priorities for staffing and services
- Set clear, quantitative needs
Conclusions

• **NERSC requirements**
  – Qualitative requirements shape NERSC functionality
  – Quantitative requirements set the performance
    “What gets measure gets improved”

• **Goals:**
  – Your goal is to make scientific discoveries
  – Our goal is to enable you to do science
Backup Slides
Challenges to Exascale

1) **System power** is the primary constraint
2) **Concurrency** (1000x today)
3) **Memory** bandwidth and capacity are not keeping pace
4) **Processor** architecture is open, but likely heterogeneous
5) **Programming model** heroic compilers will not hide this
6) **Algorithms** need to minimize data movement, not flops
7) **I/O bandwidth** unlikely to keep pace with machine speed
8) **Resiliency** critical at large scale (in time or processors)
9) **Bisection bandwidth** limited by cost and energy

*Unlike the last 20 years most of these (1-7) are equally important across scales, e.g., 1000 1-PF machines*
Accelerating Remote Display

• Problem: remote display operations are very slow due to network latency.
• Solution: deploy new technology at NERSC that hides network latency in remote display operations to improve user productivity.
• Deployed Summer 2008 to entire NERSC user community.
• Results: improves remote display by a factor of about 10x.
Berkeley Lab’s Big-Data Activities in Biology and Environment

JGI @ NERSC, Genomics pipelines (IMG), Knowledge Base (KBase)

Bioimaging

End-to-end solutions for data management, curation and analysis

Results of analysis = ideally

Sanitized data

Sanitize

Desanitize (adversary)

Raw data

Analyze (analyst)

privacy

Medical record sanitation and analysis

SDAV: Scalability Data Analytics and Visualization at LBNL

Scalability Data Management, Analysis, and Visualization

Results of analysis

Analyze (analyst)
Science in Data: From Simulation to Image Analysis

LBNL Computing on Data key in 4 of 10 Breakthroughs of the decade
• 3 Genomics problems + CMB

Data rates from experimental devices will require exascale volume computing
• Cost of sequencing > Moore’s Law
• Rate from CCDs > Moore’s Law
• Computing needs > Data size
• Computer performan < Moore Law

![Projected Rates Graph]

- Sequencers
- Detectors
- Processors
- Memory
Template Info

• **Fonts**
  – Title: Helvetica Neue Bold Condensed
  – Body: Calibri; bold level 1, regular 2+

• **Title**
  – Single line at 32pt.
  – Autofit.
  – Wraps with proper second line that fits in title box
## Theme Colors & Variants

<table>
<thead>
<tr>
<th></th>
<th>Background 1</th>
<th>Text 1</th>
<th>Background 2</th>
<th>Text 2</th>
<th>Accent 1</th>
<th>Accent 2</th>
<th>Accent 3</th>
<th>Accent 4</th>
<th>Accent 5</th>
<th>Accent 6</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Theme Color</strong></td>
<td>Black</td>
<td></td>
<td>Yellow</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lighter 80%</td>
<td>Light gray</td>
<td></td>
<td>Light yellow</td>
<td></td>
<td>Light blue</td>
<td>Light yellow</td>
<td>Light green</td>
<td>Light red</td>
<td>Light blue</td>
<td>Light yellow</td>
</tr>
<tr>
<td>Lighter 60%</td>
<td>Light gray</td>
<td></td>
<td>Light yellow</td>
<td></td>
<td>Light blue</td>
<td>Light yellow</td>
<td>Light green</td>
<td>Light red</td>
<td>Light blue</td>
<td>Light yellow</td>
</tr>
<tr>
<td>Lighter 40%</td>
<td>Light gray</td>
<td></td>
<td>Light yellow</td>
<td></td>
<td>Light blue</td>
<td>Light yellow</td>
<td>Light green</td>
<td>Light red</td>
<td>Light blue</td>
<td>Light yellow</td>
</tr>
<tr>
<td>Darker 25%</td>
<td>Black</td>
<td></td>
<td>Black</td>
<td></td>
<td>Black</td>
<td>Black</td>
<td>Black</td>
<td>Black</td>
<td>Black</td>
<td>Black</td>
</tr>
<tr>
<td>Darker 50%</td>
<td>Black</td>
<td></td>
<td>Black</td>
<td></td>
<td>Black</td>
<td>Black</td>
<td>Black</td>
<td>Black</td>
<td>Black</td>
<td>Black</td>
</tr>
</tbody>
</table>
### Sample Tables

#### Light Style - Accent 5

<table>
<thead>
<tr>
<th>Light Style - Accent 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

#### Medium Style - Accent 1

<table>
<thead>
<tr>
<th>Medium Style - Accent 1</th>
<th>Medium Style - Accent 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sample Chart

Category 1
Category 2

Series 1
Series 2
Series 3
Series 4
Series 5
Series 6
Headers, Footers and Dates

- All controlled by “View/Headers and Footers” Menu
- The date appears on the title and handout pages
  - Use “fixed” for a known presentation date.
  - Use “update automatically” to track the current date.
- Footer information appears to the right of the Lab logo
  - Optional. Use for name of presentation, copyright info or other usage designations.
- Notes and handout pages have separate header, footer and date information.
  - Need to set this redundantly
Importing from Existing Presentations

• Works OK if the source presentation used a well-formed template
  – May need to reapply the slide template one or two times.
  – Then correct text size directly or with autofit.

• Doesn’t work well if it was manually formatted.
  – Observe text frames after importing
  – May need to cut and paste to the text boxes generated from the master slides.
Backup Slides Follow Big Logo

• Never have a slide that says “Backup”
  – Especially if the backup slides address issues that you would rather not cover.
  – It will only invite discussion.
Web Color Palette

Primary Color Palette

- Slate
  - R11 G33 B57

- Mute Turquoise
  - R0 G143 B184

- Turquoise
  - R35 G171 B227

- Light Grey Blue
  - R210 G227 B235

Secondary Color Palette

- Dark Teal
  - R25 G73 B99

- Teal
  - R35 G108 B144

- Orange
  - R248 G150 B29

- Green
  - R34 G146 707
Earlier Web Color Palette

Primary Color Palette

- Slate
  - R 12 G 21 B 39
  - HEX 0c1527
- Mute Turquoise
  - R 0 G 144 B 189
  - HEX 0090bd
- Turquoise
  - R 0 G 175 B 229
  - HEX 00afe5

Secondary Color Palette

- Gold
  - R 255 G 185 B 0
  - HEX ff9000
- Dark Teal
  - R 17 G 71 B 102
  - HEX 1147bd
- Bright Cyan
  - R 107 G 212 B 251
  - HEX 6bd4fb
- Orange
  - R 255 G 101 B 0
  - HEX ff6500

Neutral Color Palette

- Cream
  - R 255 G 247 B 220
  - HEX ff7dc
- Light Warm Gray
  - R 218 G 218 B 218
  - HEX daddda
- Medium Warm Grey
  - R 137 G 148 B 150
  - HEX 899496
- Dark Teal
  - R 17 G 71 B 102
  - HEX 1147bd