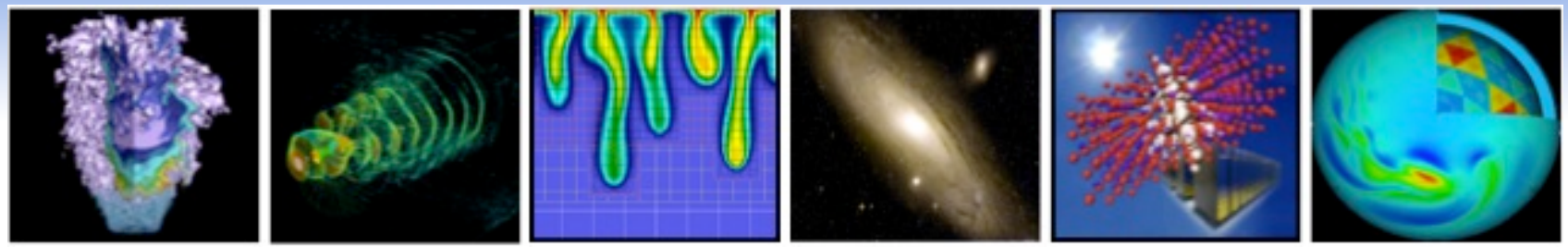


NERSC Role in Advanced Scientific Computing Research

Katherine Yelick
NERSC Director

Requirements Workshop

U.S. DEPARTMENT OF ENERGY | Office of Science

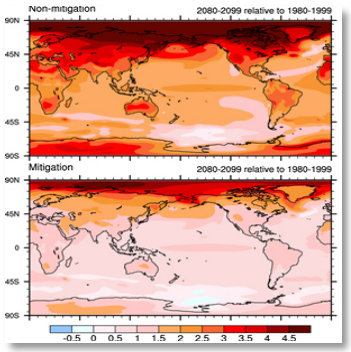
NERSC

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC Mission

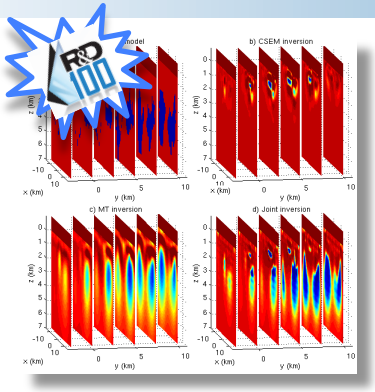
The mission of the National Energy Research Scientific Computing Center (NERSC) is to *accelerate the pace of scientific discovery* by providing high performance computing, information, data, and communications services for *all DOE Office of Science (SC) research.*

Sample Scientific Accomplishments at NERSC



Climate

Studies show that global warming can still be diminished if society cuts emissions of greenhouse gases.
(Warren Washington, NCAR)

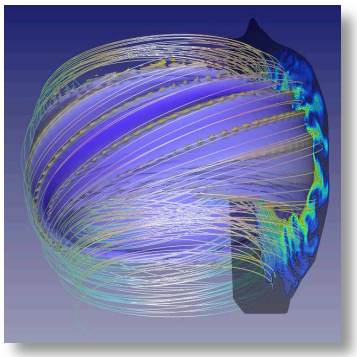


Energy Resources

Award-winning software uses massively-parallel supercomputing to map hydrocarbon reservoirs at unprecedented levels of detail.
(Greg Newman, LBNL)

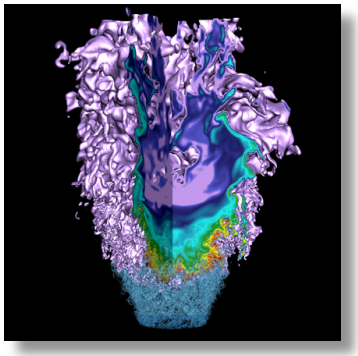
Fusion Energy

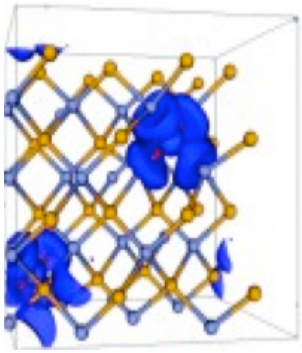
A new class of non-linear plasma instability has been discovered that may constrain design of the ITER device.
(Linda Sugiyama, MIT)



Combustion

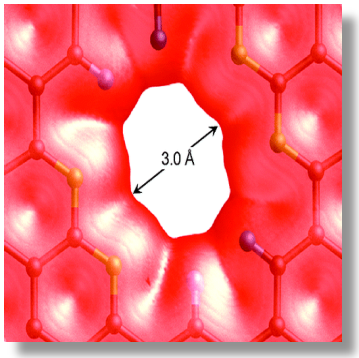
Adaptive Mesh Refinement allows simulation of a fuel-flexible low-swirl burner that is orders of magnitude larger & more detailed than traditional reacting flow simulations allow.
(John Bell, LBNL)





Materials

Electronic structure calculations suggest a range of inexpensive, abundant, non-toxic materials that can produce electricity from heat.
(Jeffrey Grossman, MIT)

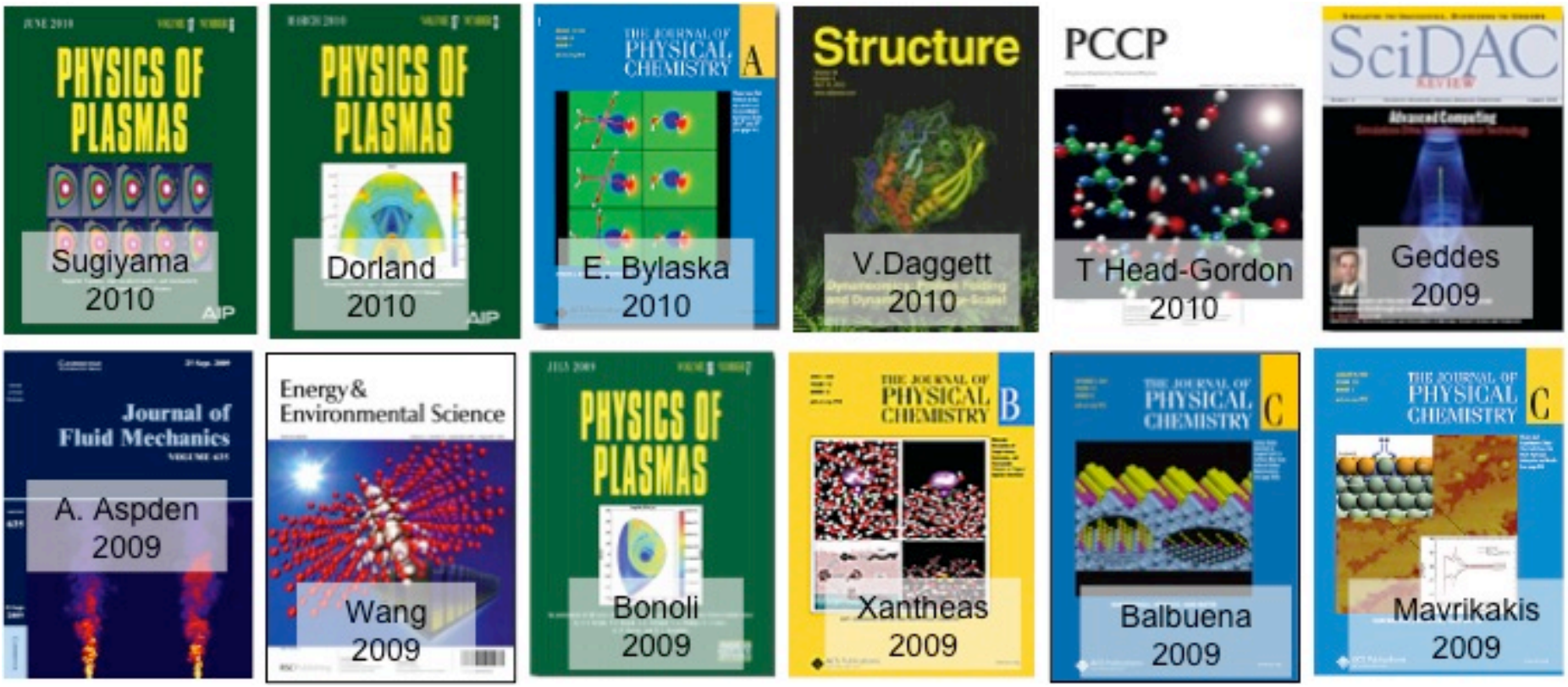


Nano Science

Using a NERSC NISE grant researchers discovered that Graphene may be the ultimate gas membrane, allowing inexpensive industrial gas production.
(De-en Jiang, ORNL)

U.S. DEPARTMENT OF ENERGY | Office of Science

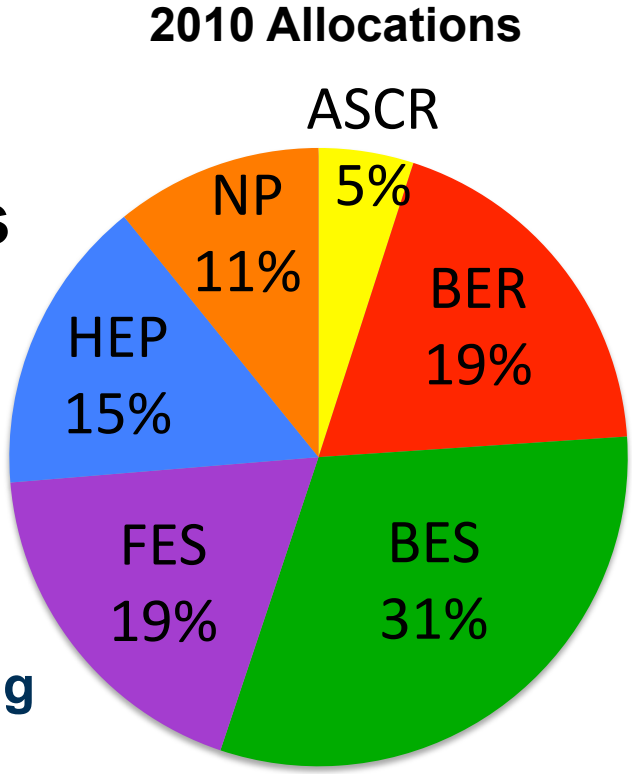
Recent Cover Stories from NERSC Research



NERSC is enabling new high quality science across disciplines, with over *1,600* refereed publications last year

NERSC is the Production Facility for DOE Office of Science

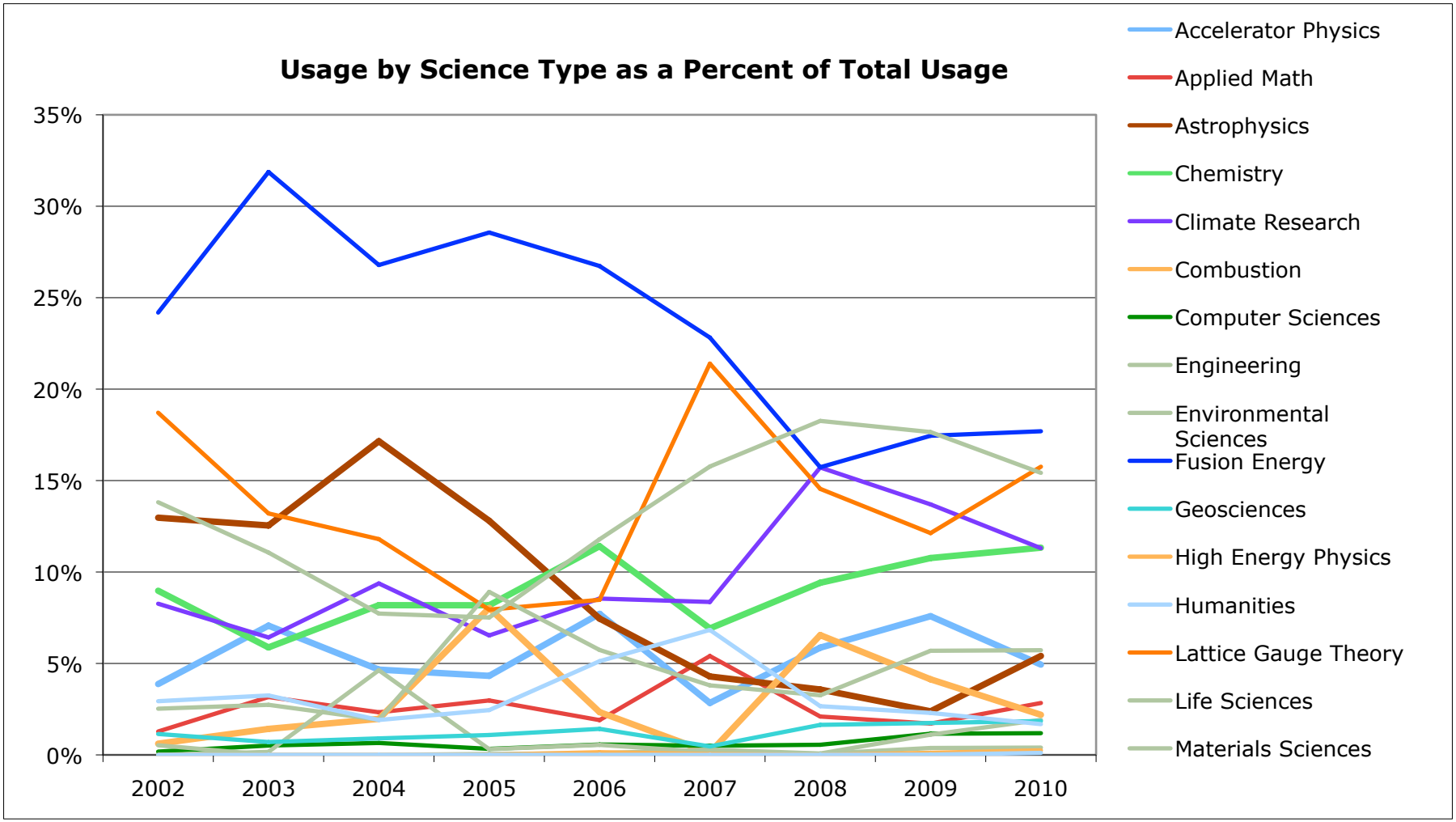
- **NERSC serves a large population**
  - Over 4000 users, 500 projects, 500 code instances
- **Focus on "unique" resources**
  - Expert consulting and other services
  - High end computing systems
  - High end storage systems
  - Interface to high speed networking
- **Science-driven**
  - Machines procured competitively using application benchmarks from DOE/SC
  - Allocations controlled by DOE/SC Program Offices to couple with funding decisions



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

DOE Priorities for NERSC Change Over Time



U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

ASCR's Computing Facilities

NERSC at LBNL

- 1000+ users,100+ projects
- Allocations:
  - 80% DOE program manager control
  - 10% ASCR Leadership Computing Challenge*
  - 10% NERSC reserve
- Science includes all of DOE Office of Science
- Machines procured for application performance

LCFs at ORNL and ANL

- 100+ users 10+ projects
- Allocations:
  - 60% ANL/ORNL managed INCITE process
  - 30% ACSR Leadership Computing Challenge*
  - 10% LCF reserve
- Science limited to largest scale; no limit to DOE/SC
- Machines procured for peak performance

NERSC Systems and Allocations

Large-Scale Computing Systems

Franklin (NERSC-5): Cray XT4

- 9,532 compute nodes; 38,128 cores
- ~25 Tflop/s on applications; 356 Tflop/s peak

Hopper (NERSC-6): Cray XE6

- Phase 1: Cray XT5, 668 nodes, 5344 cores
- Phase 2: > 1 Pflop/s peak (late 2010 delivery)

Clusters

105 Tflops total

Carver

- IBM iDataplex cluster

PDSF (HEP/NP)

- Linux cluster (~1K cores)

Magellan Cloud testbed

- IBM iDataplex cluster

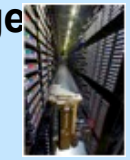
NERSC Global Filesystem (NGF)

Uses IBM's GPFS

1.5 PB; 5.5 GB/s

HPSS Archival Storage

- 40 PB capacity
- 4 Tape libraries



Analytics

Euclid (512 GB shared memory)

Dirac GPU testbed (48 nodes)

Allocated hours will increase rough 4x once Hopper is in full production

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC Uses Your Requirements for Strategic Planning and Prioritization

NERSC Strategy 2011

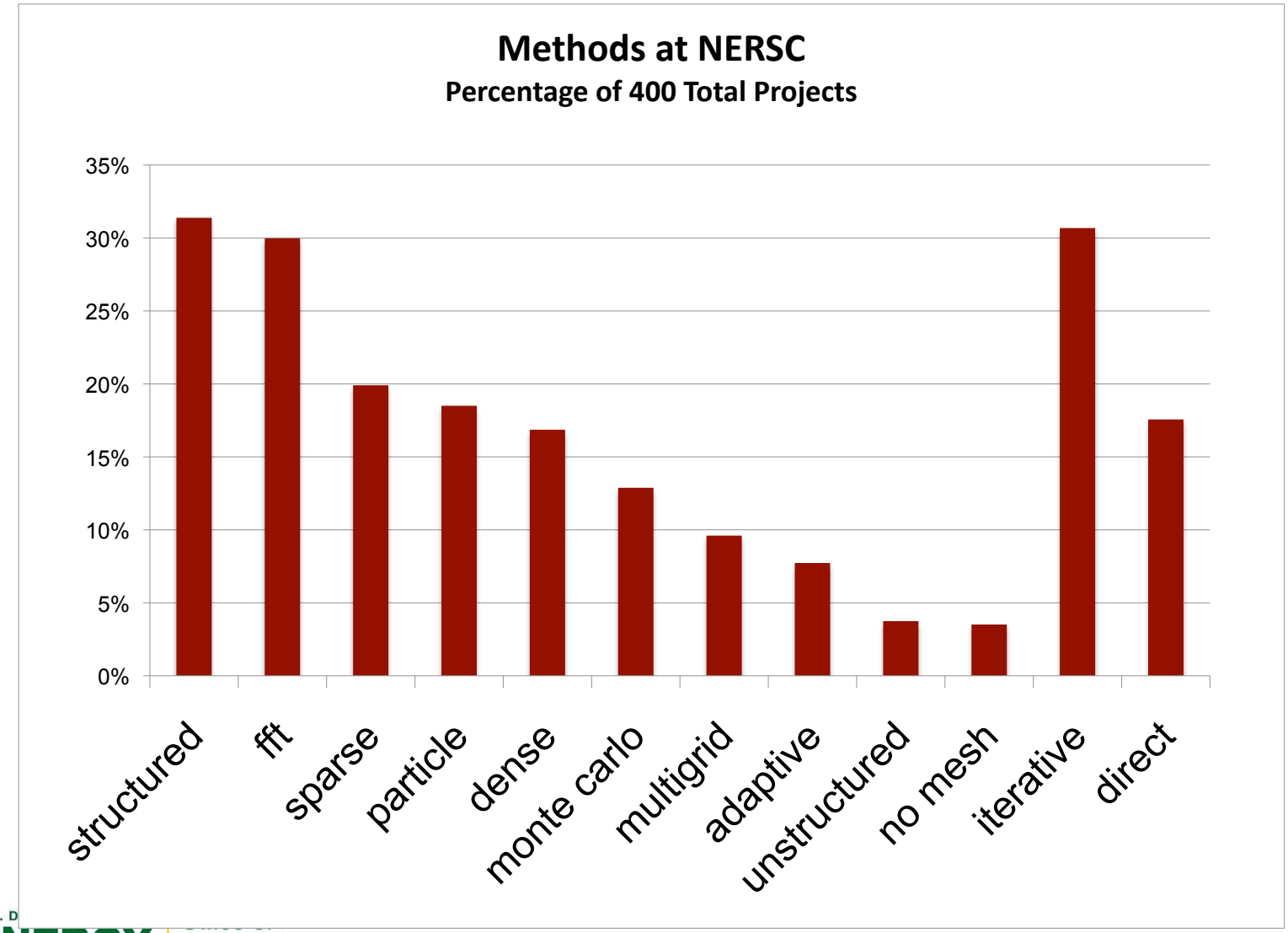
- **Observations**
  - End of single processor performance gains leading to *multicore, GPU*, and other hardware innovations
  - *Energy* costs a growing concern for facilities and possible barrier to exascale
  - New capacity computing available in *Clouds*
  - *Flood of data* is increasing from both simulations and experiments
- **NERSC future priorities are driven by science:**
  - Current user demand surpasses resources: Scientific need increases by at least 10x every 3 years
  - Growing interest in data services and systems

NERSC Response to Science Needs

- Installation of Hopper (NERSC-6) system
  - Phase 1 in production; Phase 2 production in 2011
- Replacement of Bassi and Jacquard by Carver
  - Saved money, space, and energy
- Replaced Davinci by Euclid for analytics
- Upgrading global filesystem (NGF) from 1.6 to 2.5 PB
  - Enabled NGF access from Franklin; available for Hopper
- Added external services to Franklin and Hopper
- Added testbeds for clouds (Magellan) and GPUs (Dirac) paid by non-program funds
- Facility upgrade from 6 MW to 9 MW

Numerical Methods at NERSC

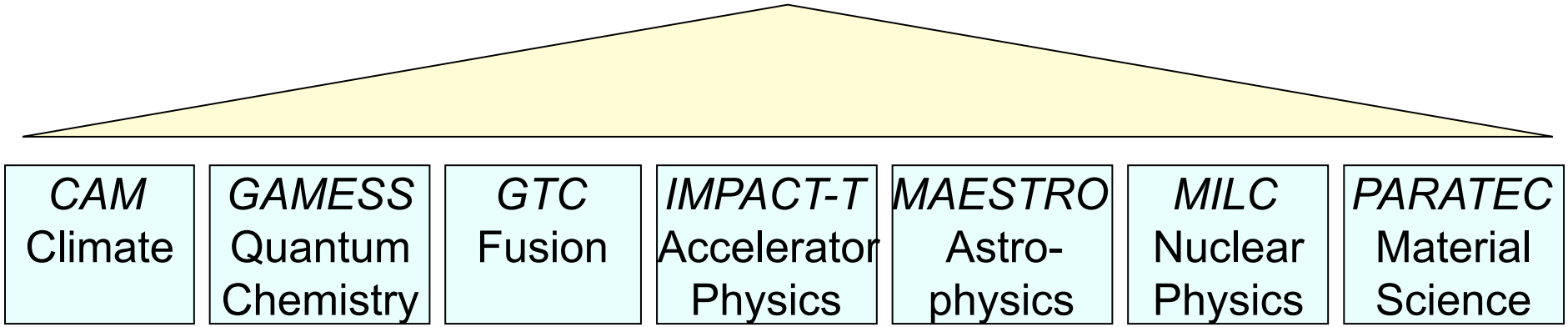
(Caveat: survey data from ERCAP requests)



Applications Drive NERSC Procurements

Because hardware peak performance does not necessarily reflect real application performance

NERSC-6 "SSP" Benchmarks



- Benchmarks reflect diversity of science and algorithms
- SSP = average performance (Tflops/sec) across machine
- Used before selection, during and after installation
- Question: What applications best reflect your workload?

Algorithm Diversity

Science areas	*Dense linear algebra*	*Sparse linear algebra*	*Spectral Methods (FFT)s*	*Particle Methods*	*Structured Grids*	*Unstructured or AMR Grids*
Accelerator Science		X	X	X	X	X
Astrophysics	X	X	X	X	X	X
Chemistry	X	X	X	X		
Climate			X		X	X
Combustion					X	X
Fusion	X	X		X	X	X
Lattice Gauge		X	X	X	X	
Material Science	X		X	X	X	

NERSC users require a system which performs well in all areas

NERSC-6 System "Hopper"

Grace Murray Hopper
(1906-1992)

- Cray system selected competitively:
  - Used application benchmarks from climate, chemistry, fusion, accelerator, astrophysics, QCD, and materials
  - Best application performance per dollar based
  - Best sustained application performance per MW
  - External Services for increased functionality and availability

Phase 1: Cray XT5

- *In production on 3/1/2010*
- 668 nodes, 5,344 cores
- 2.4 GHz AMD Opteron
- 2 PB disk, 25 GB/s
- Air cooled

Phase 2: Cray system

- > 1 Pflop/s peak
- ~ 150K cores, 6250 nodes
- 12 AMD Magny Cours chips, 2 per node (dual socket)
- Liquid cooled

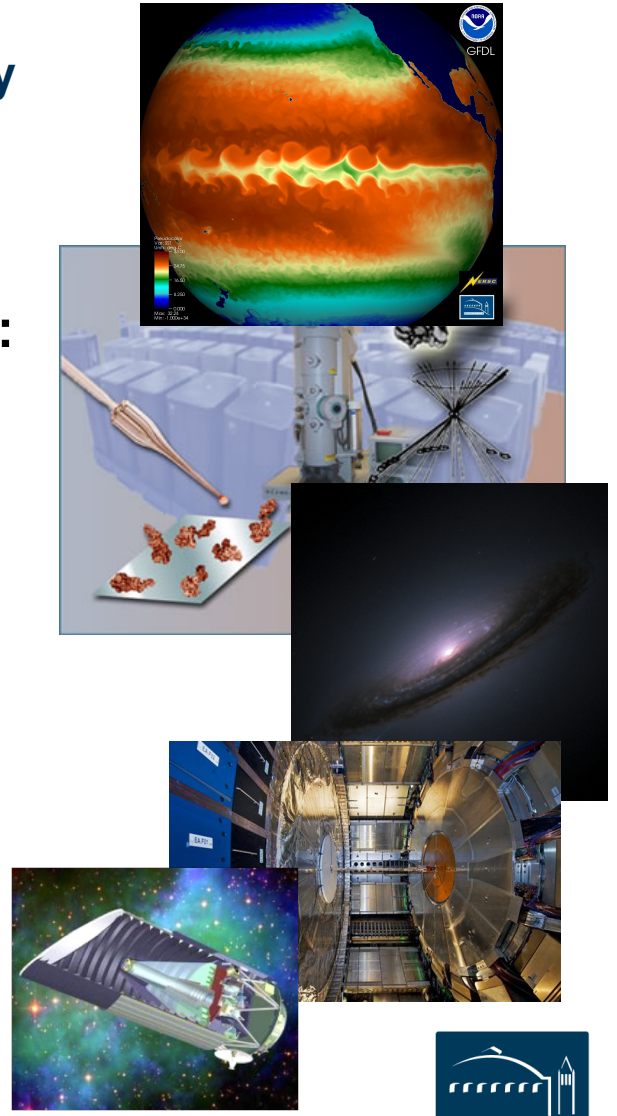
4Q09	1Q10	2Q10	3Q10	4Q10	1Q11

U.S. DEPARTMENT OF ENERGY | Office of Science

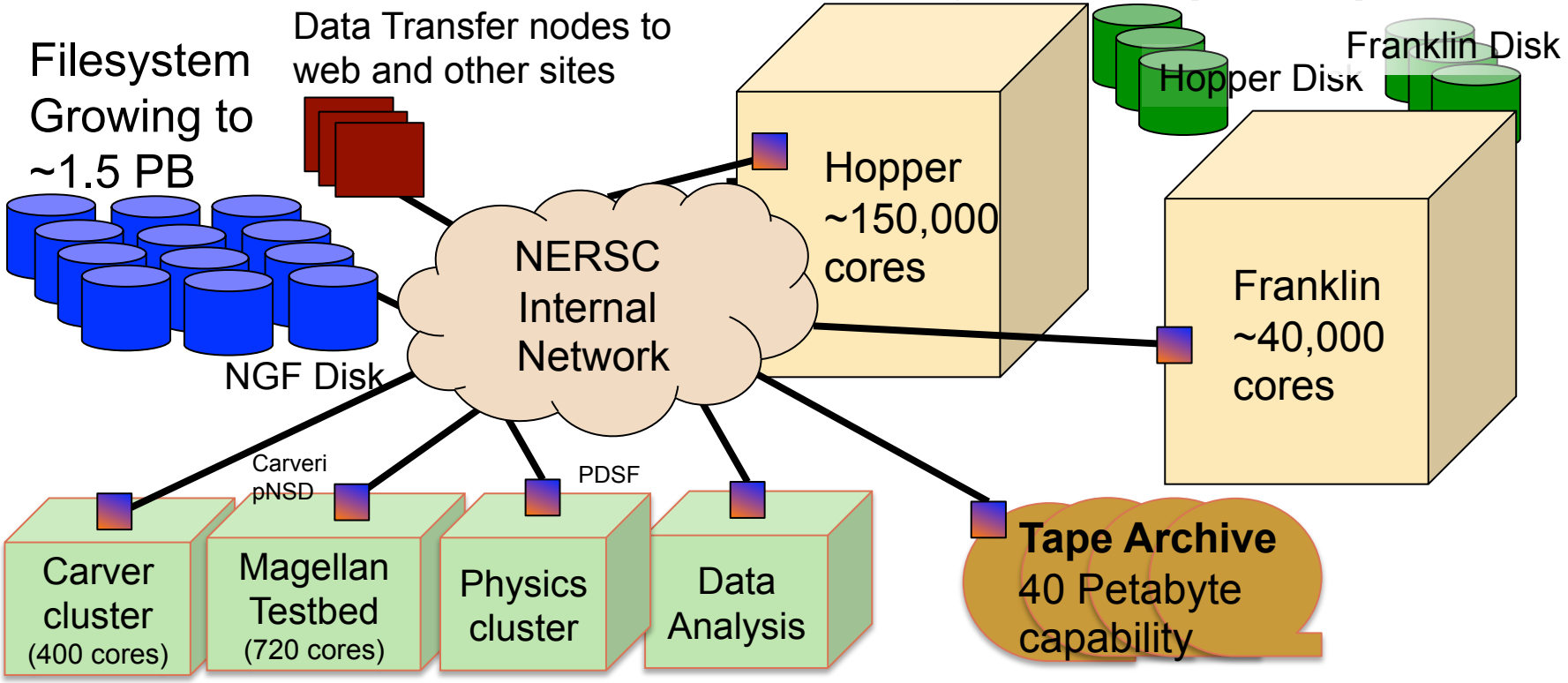
BERKELEY LAB
Lawrence Berkeley National Laboratory

Data Driven Science

- **Scientific data sets are growing exponentially**
  - **Ability to generate data is exceeding our ability to store and analyze**
  - **Simulation systems and some observational devices grow in capability with Moore's Law**
- **Petabyte (PB) data sets will soon be common:**
  - ***Climate modeling:* estimates of the next IPCC data is in 10s of petabytes**
  - ***Genome:* JGI alone will have .5 petabyte of data this year and double each year**
  - ***Particle physics*: LHC is projected to produce 16 petabytes of data per year**
  - ***Astrophysics*: LSST and others will produce 5 petabytes/year**
- **Create scientific communities with "Science Gateways" to data**



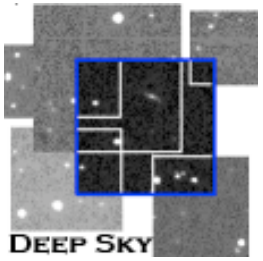
NERSC Architecture with NERSC Global Filesystem (NGF)



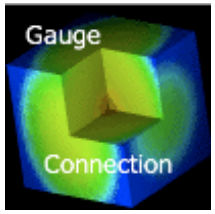
- NERSC invests annually in storage hardware, both filesystems and the HPSS Tape archive
- Innovate to make these more convenient for users

Science Gateways at NERSC

- **Create scientific communities around data sets**
  - Models for sharing vs. privacy differ across communities
  - Accessible by broad community for exploration, scientific discovery, and validation of results
  - Value of data also varies: observations may be irreplaceable



- **A *science gateway* is a set of hardware and software that provides data/services remotely**
  - Deep Sky – "Google-Maps" of astronomical image data
    - Discovered 140 supernovae in 60 nights (July-August 2009)
    - 1 of 15 international collaborators were accessing NGF data through the SG nodes 24/7 using both the web interface and the database.
  - Gauge Connection – Access QCD Lattice data sets
  - Planck Portal – Access to Planck Data



- **Building blocks for science on the web**
  - Remote data analysis, databases, job submission

Communication Services

Since 2007, NERSC is a net data importer. In support of our users, it is important that we take on a lead role in improving intersite data transfers.

- **Systems and software typically tuned only within a site**
- **Technical, social, and policy challenges abound:**
  - **High performance transfer software has too many options → hard to use.**
  - **Systems designed for computation can have bottlenecks in data transfers**
  - **Systems at different sites often often have incompatible versions of transfer software.**
  - **Trying to maintain security exceptions (firewall holes) for all the systems and software at each site was impossible.**
- **… and the list goes on.**
- **NERSC established Data Transfer Nodes (DTNs).**
  - **Reduced transfer time of 30 TB from 30+ days to 2 days**
  - **We formed a working group with experts at the three labs and ESNet**

http://www.nersc.gov/nusers/systems/DTN

http://fasterdata.es.net

Visualization Support

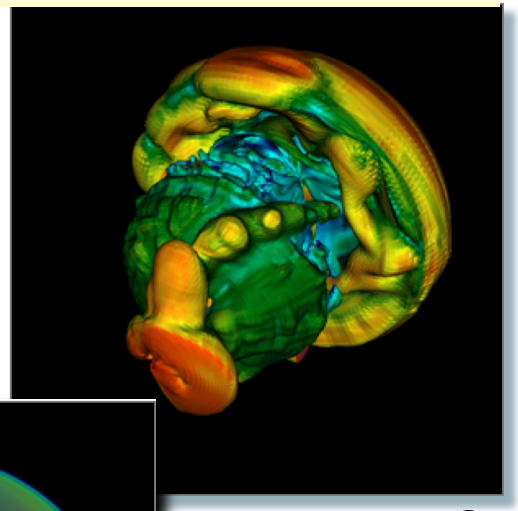
Petascale visualization: Demonstrate visualization scaling to unprecedented concurrency levels by ingesting and processing unprecedentedly large datasets.

Implications: Visualization and analysis of Petascale datasets requires the I/O, memory, compute, and interconnect speeds of Petascale systems.

Accomplishments: Ran VisIt SW on 16K and 32K cores of Franklin.

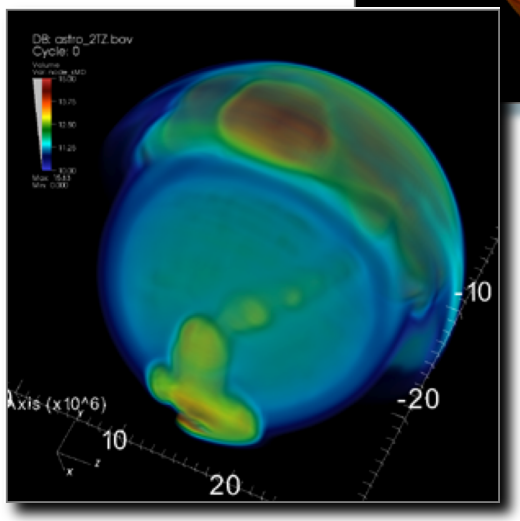
- First-ever visualization of two *trillion* zone problem (TBs per scalar); data loaded in parallel.
- Petascale visualization

Plots show 'inverse flux factor,' the ratio of neutrino intensity to neutrino flux, from an ORNL 3D supernova simulation using CHIMERA.



a

b



Isocontours (a) and volume rendering (b) of two trillion zones on 32K cores of Franklin.

NISE: NERSC Initiative for Scientific Exploration

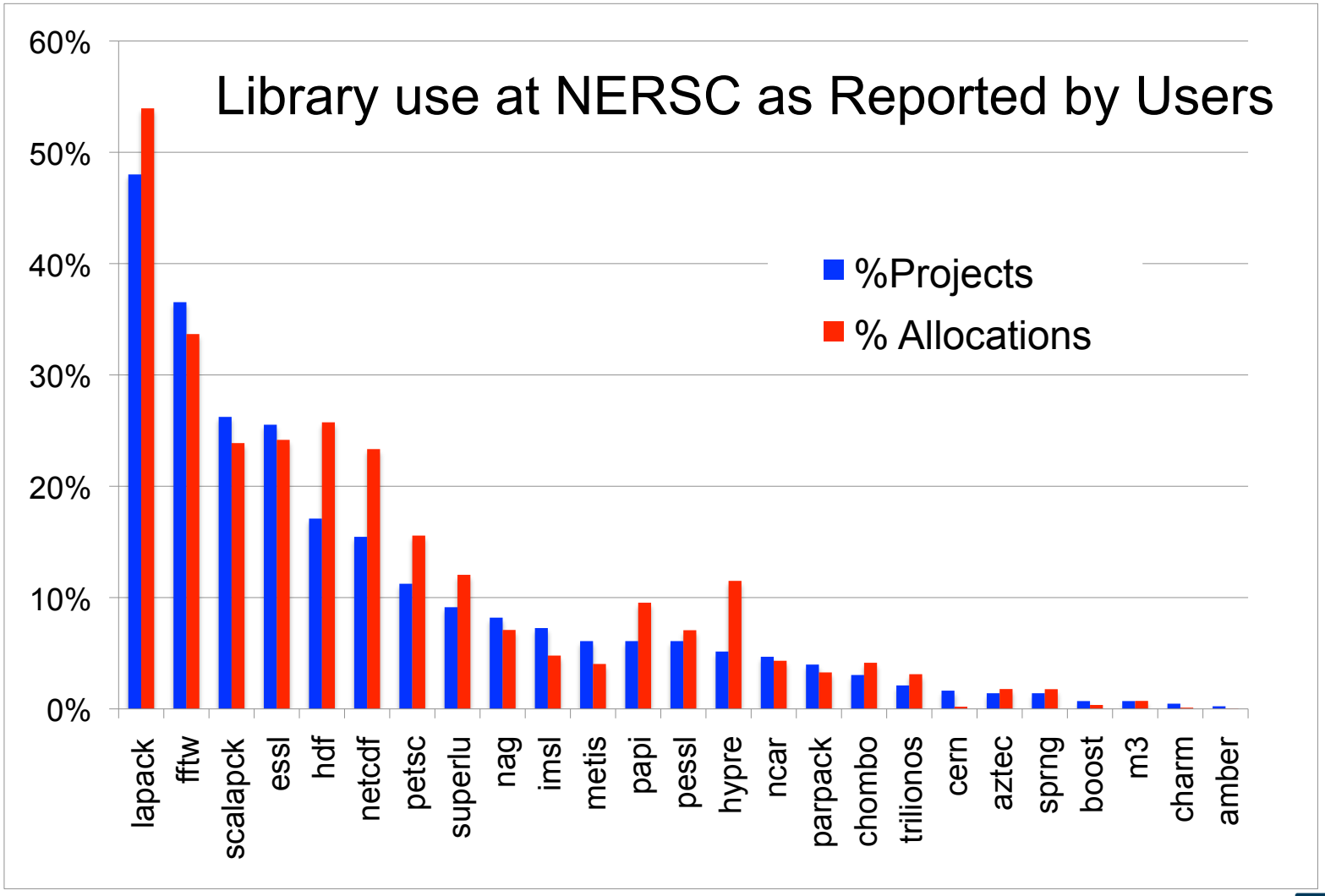
- NERSC Users: Open process for 10% NERSC time
- Modeled after original INCITE program from NERSC:
  - Focused computing and consulting resources
- NISE program at NERSC (started in 2009)
  - Programming techniques for multicore and scaling in general
  - Science problems near breakthrough (high risk/payoff)
  - http://www.nersc.gov/nusers/accounts/NISE.php
- ~30M hours made available to NISE projects in 2010
  - E.g., V. Izzo, GA, ITER rapid shut-down simulation
- ASCR's ALCC Program has a similar number of hours:

  http://www.er.doe.gov/ascr/Facilities/ALCC.html

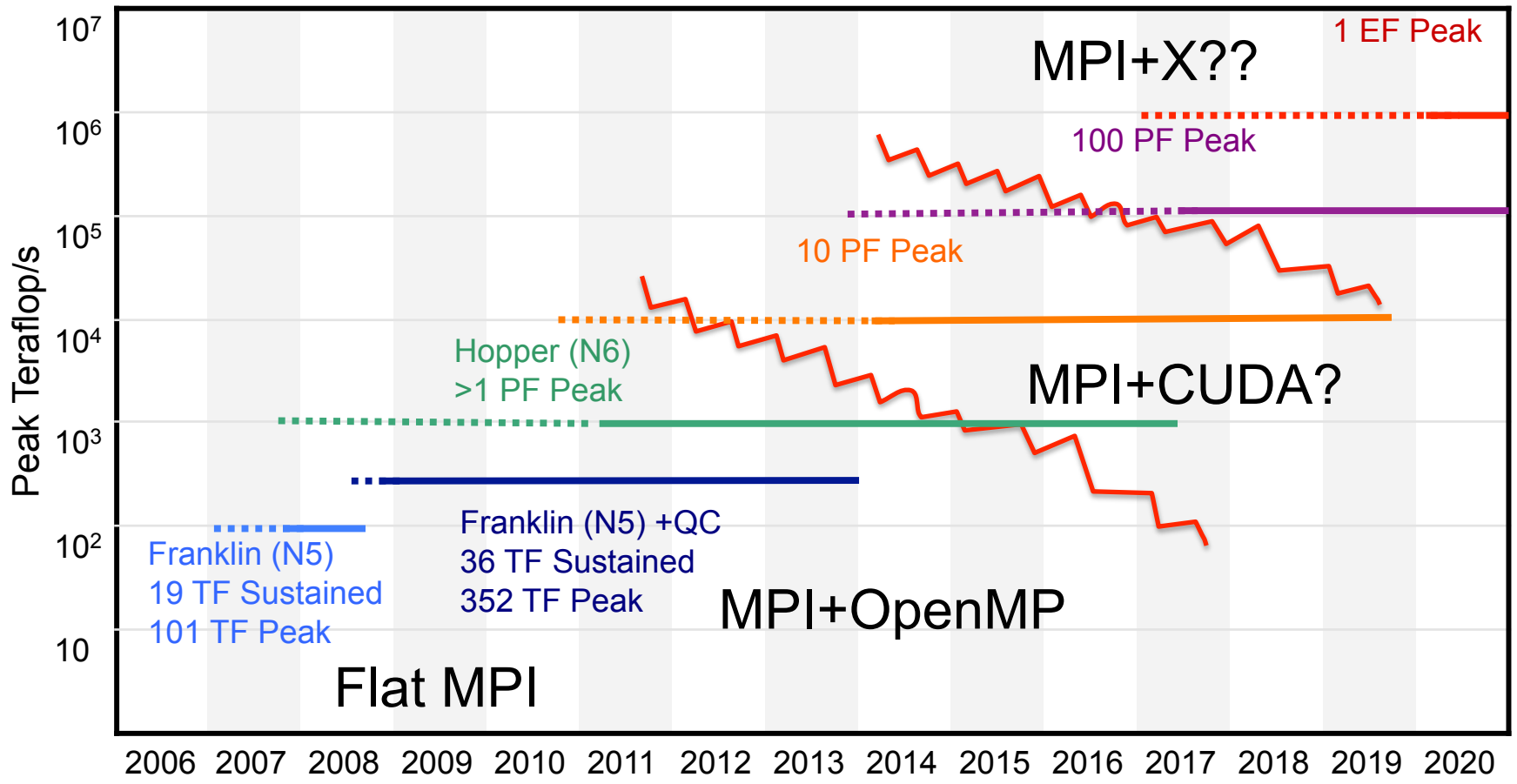
Reservations at NERSC

- **Reservation service being tested:**
  - **Reserve a certain date, time and duration**
    - **Debugging at scale**
    - **Real-time constraints in which need to analyze data before next run, e.g., daily target selection telescopes or genome sequencing pipelin**
  - **At least 24 hours advanced notice**
    - **https://www.nersc.gov/nusers/services/reservation.php**
  - **Successfully used for IMG run, Madcap, IO benchmarking, etc.**

Library Use at NERSC



How and When to Move Users



Want to avoid two paradigm disruptions on road to Exa-scale

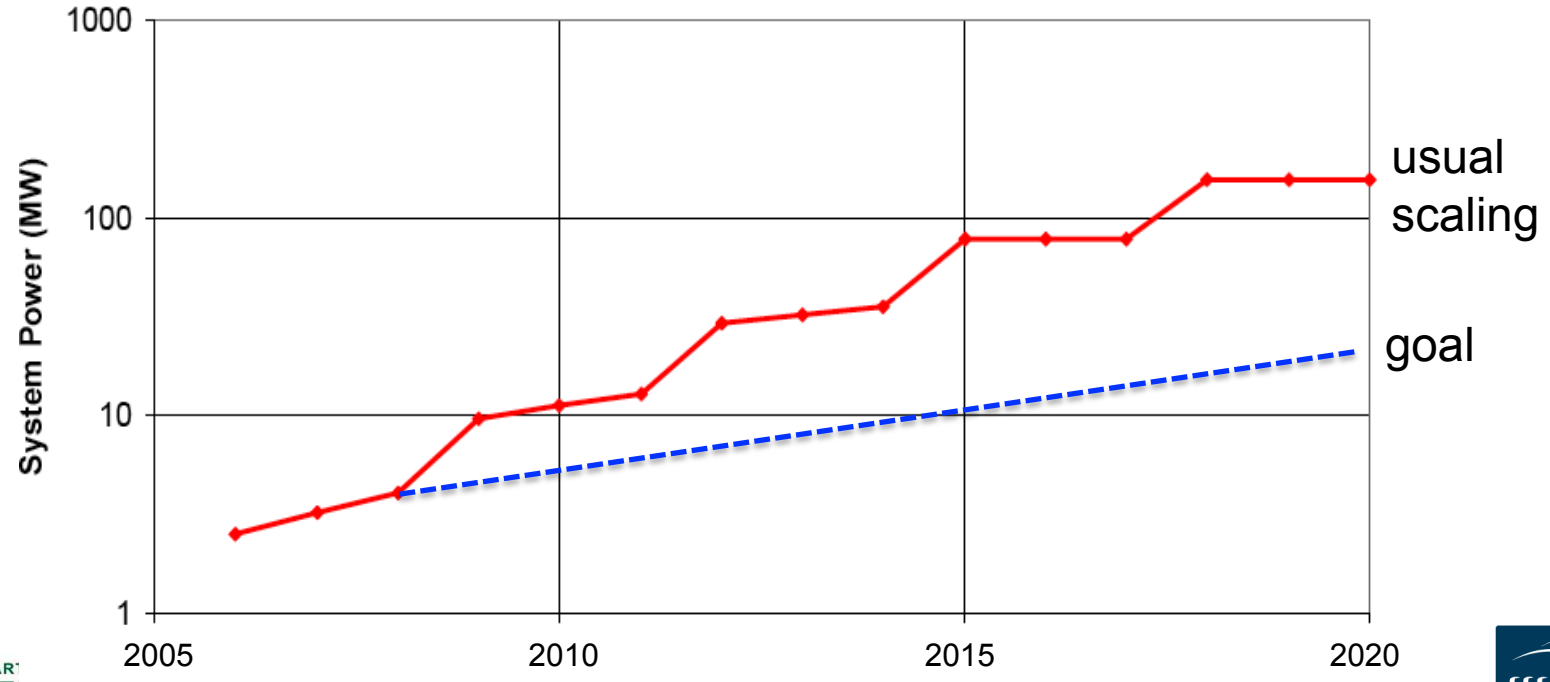
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Exascale is about Energy Efficient Computing

At $1M per MW, energy costs are substantial

- 1 petaflop in 2010 will use 3 MW
- 1 exaflop in 2018 at 200 MW with “usual” scaling
- 1 exaflop in 2018 at 20 MW is target

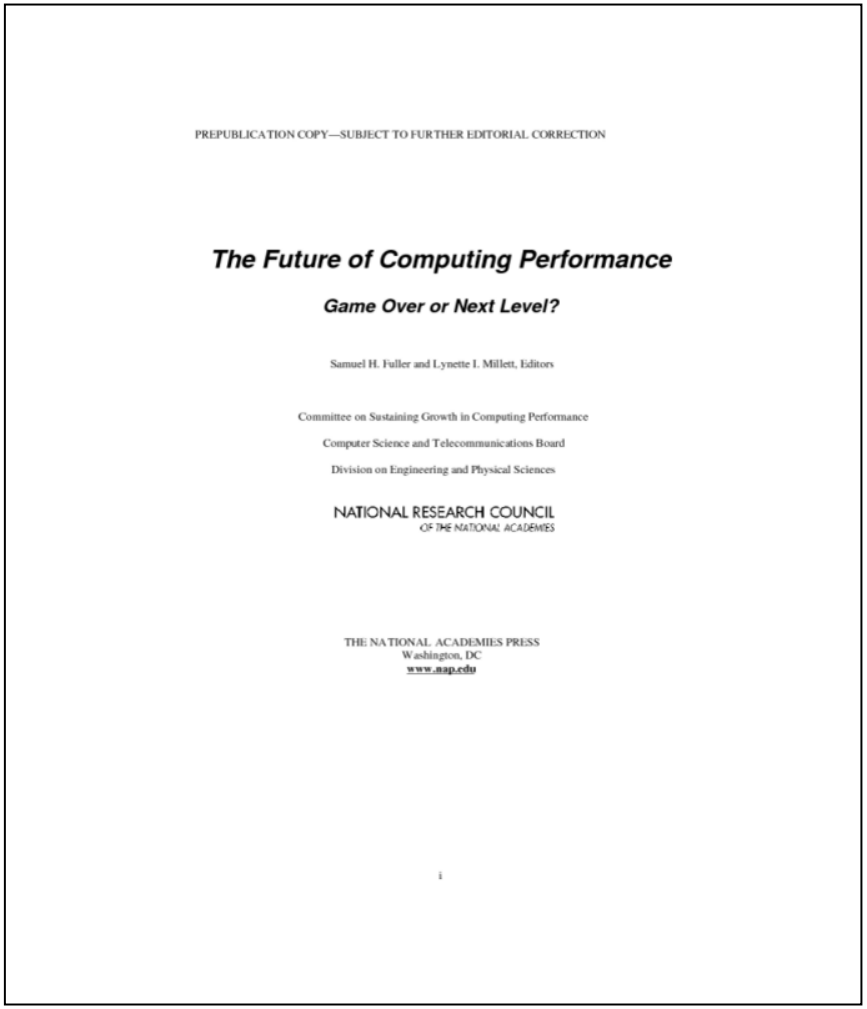


Challenges to ~~Exascale~~ Performance Growth

1) **System power** is the primary constraint
2) **Concurrency** (1000x today)
3) **Memory** bandwidth and capacity are not keeping pace
4) **Processor** architecture is an open question
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) **Reliability and resiliency** will be critical at this scale
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., 100 10-PF machines

National Academies Report on Computing Performance



PREPUBLICATION COPY—SUBJECT TO FURTHER EDITORIAL CORRECTION

The Future of Computing Performance

Game Over or Next Level?

Samuel H. Fuller and Lynette I. Millett, Editors

Committee on Sustaining Growth in Computing Performance

Computer Science and Telecommunications Board

Division on Engineering and Physical Sciences

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

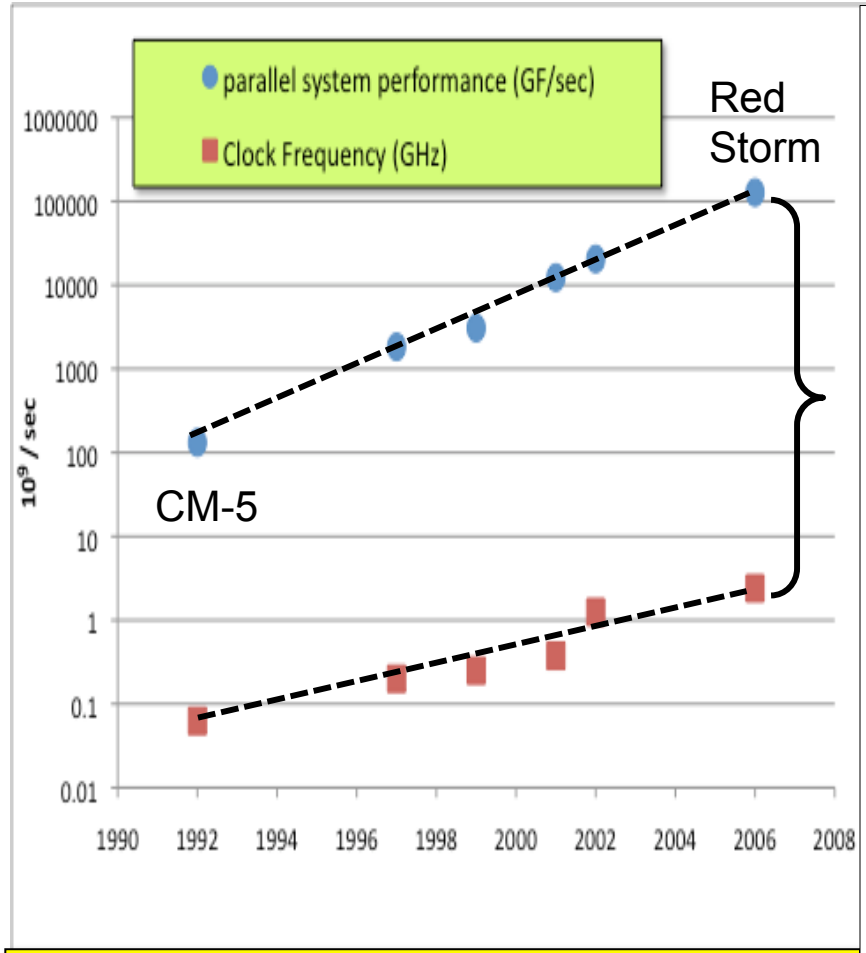
i

- **Report makes same points**
  - **Past performance increases have driven innovation**
  - **Processor speeds stalled**
  - **Energy is limitation**

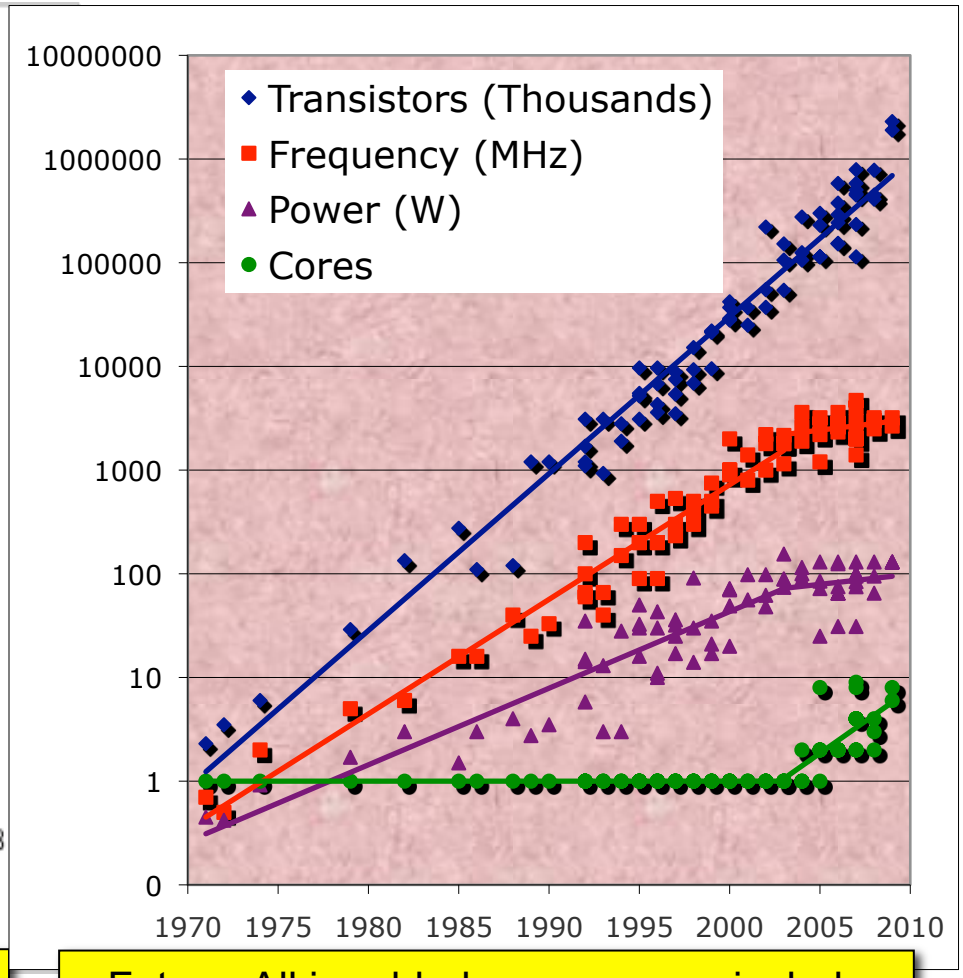
Revolutions Require a New Strategy

- **Preserve Application Performance as goal**
  - But, allow significant optimizations
  - Produced optimized versions of some codes
  - Understand performance early
- **E.g., for Fermi-based GPU system**
  - Fermi is nearly as expensive as host node
  - 48 nodes w/Fermi or 2x more nodes without
  - Minimum: Fermi must be 2x faster than host
- **Estimating value of acceleration for SSP**
  - Estimate "best possible performance"
  - Successive refinement of estimates
  - Stop if estimate < min threshold (2x host)

Moore's Law Continues, but Only with Added Concurrency

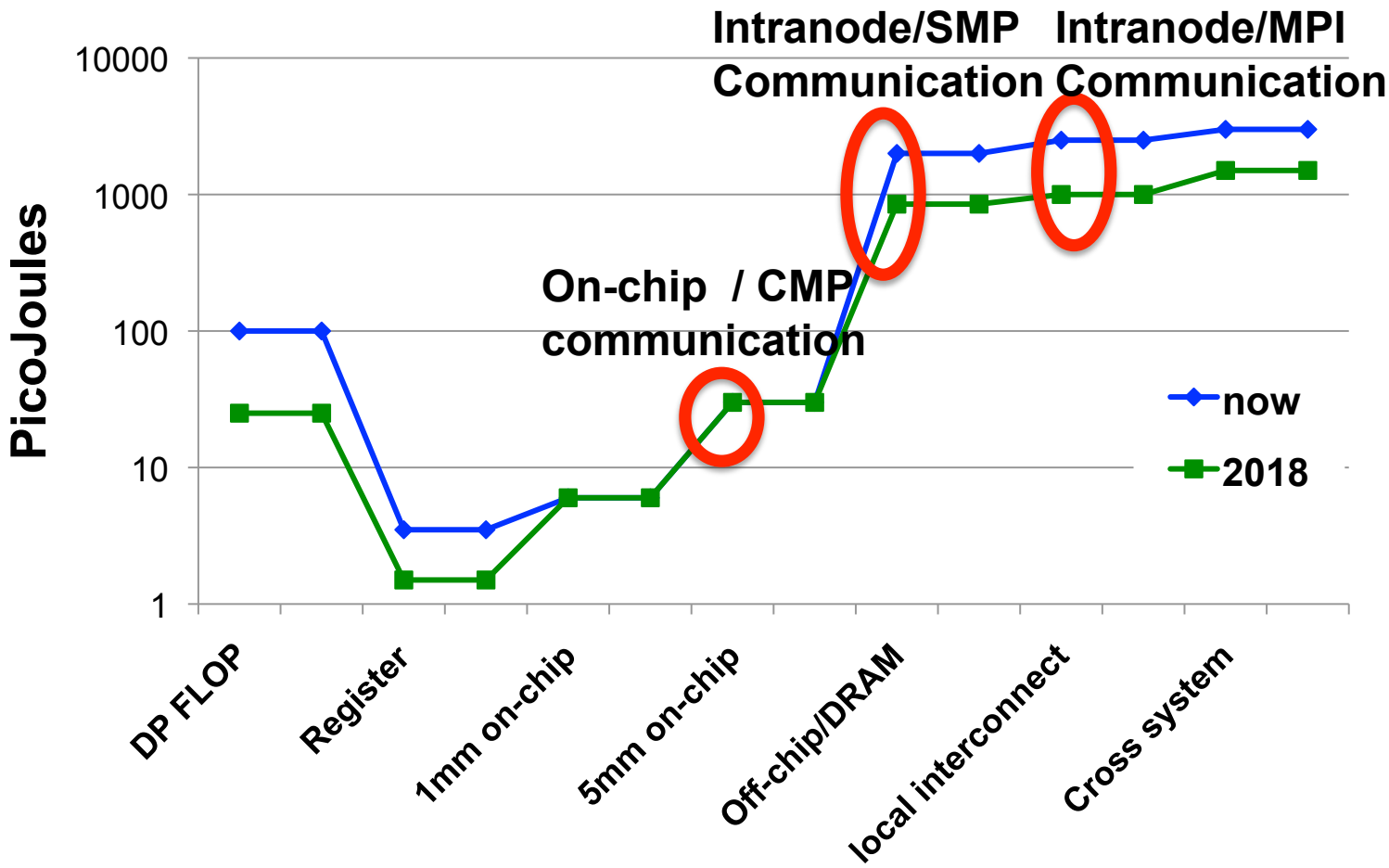


Past: 1000x performance increase was 40x clock speed and 25x concurrency



Future: All in added concurrency, include new on-chip concurrency

Where does the Energy (and Time) Go?

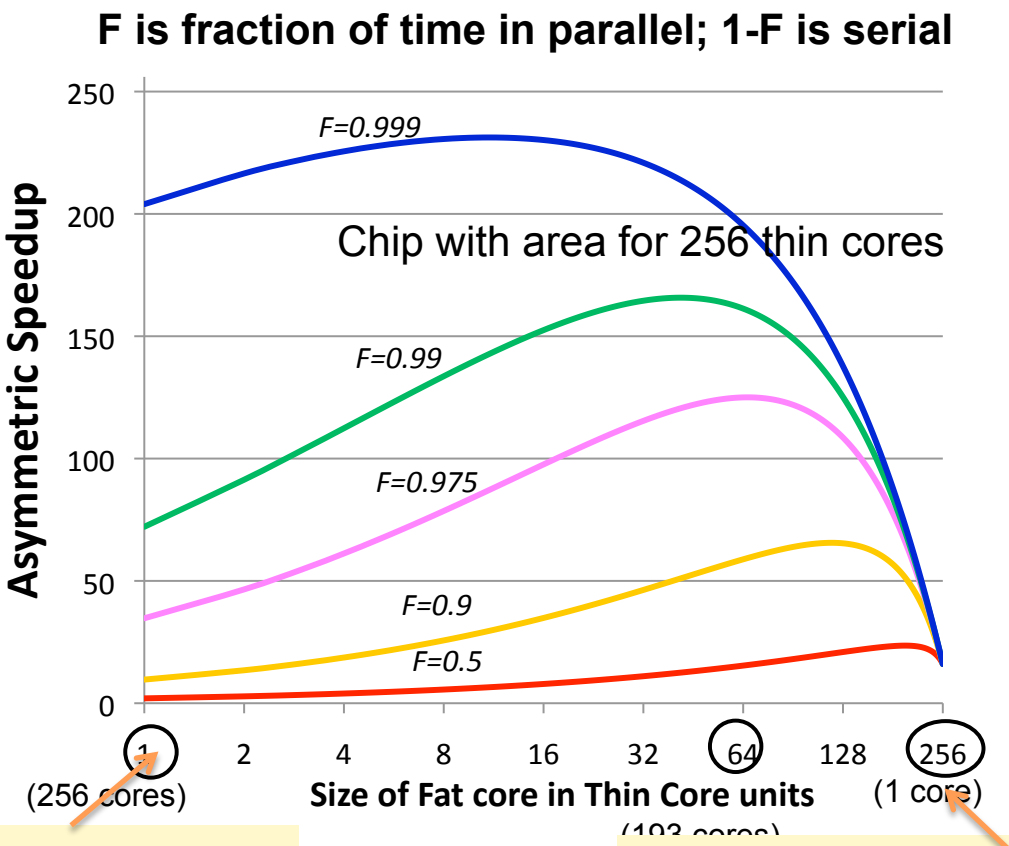


Counting flops is irrelevant, only data movement matters

Case for Lightweight Core and Heterogeneity

	Intel QC Nehalem	Tensil-ica	Overall Gain
Power (W)	100	.1	10^3
Area (mm^2)	240	2	10^2
DP flops	50	4	.1
Overall			10^4

Lightweight (thin) cores improve energy efficiency



256 small cores

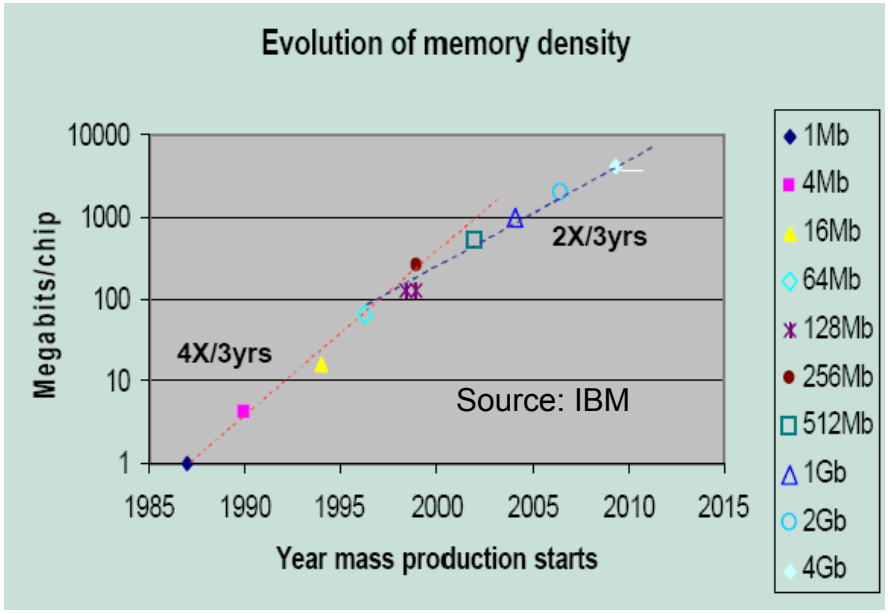
1 fat core

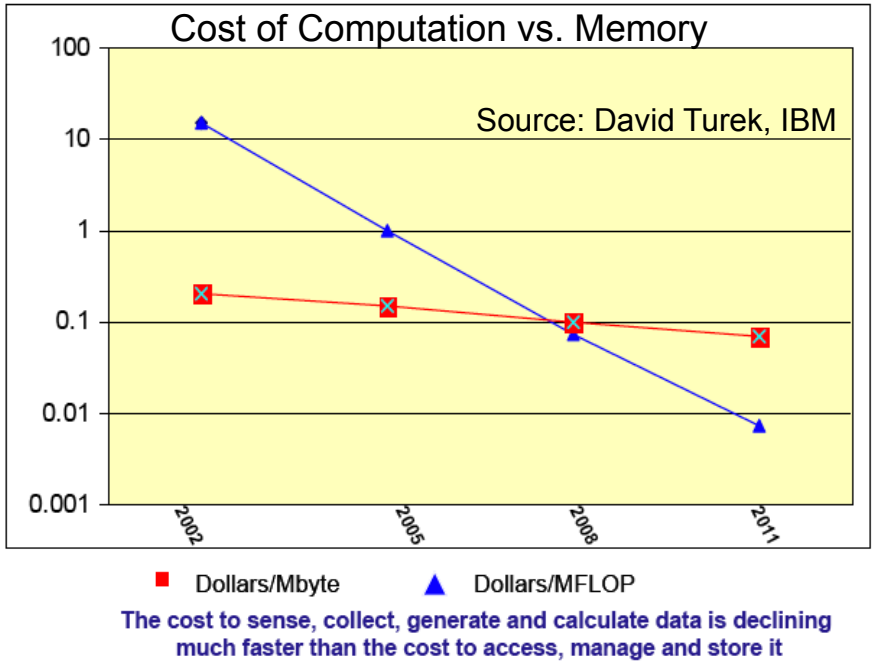
Ubiquitous programming model of today (MPI) will not work within a processor chip

Memory is Not Keeping Pace

Technology trends against a constant or increasing memory per core

- Memory density is doubling every three years; processor logic is every two
- Memory costs are dropping gradually compared to logic costs





Question: *Can you double concurrency without doubling memory?*

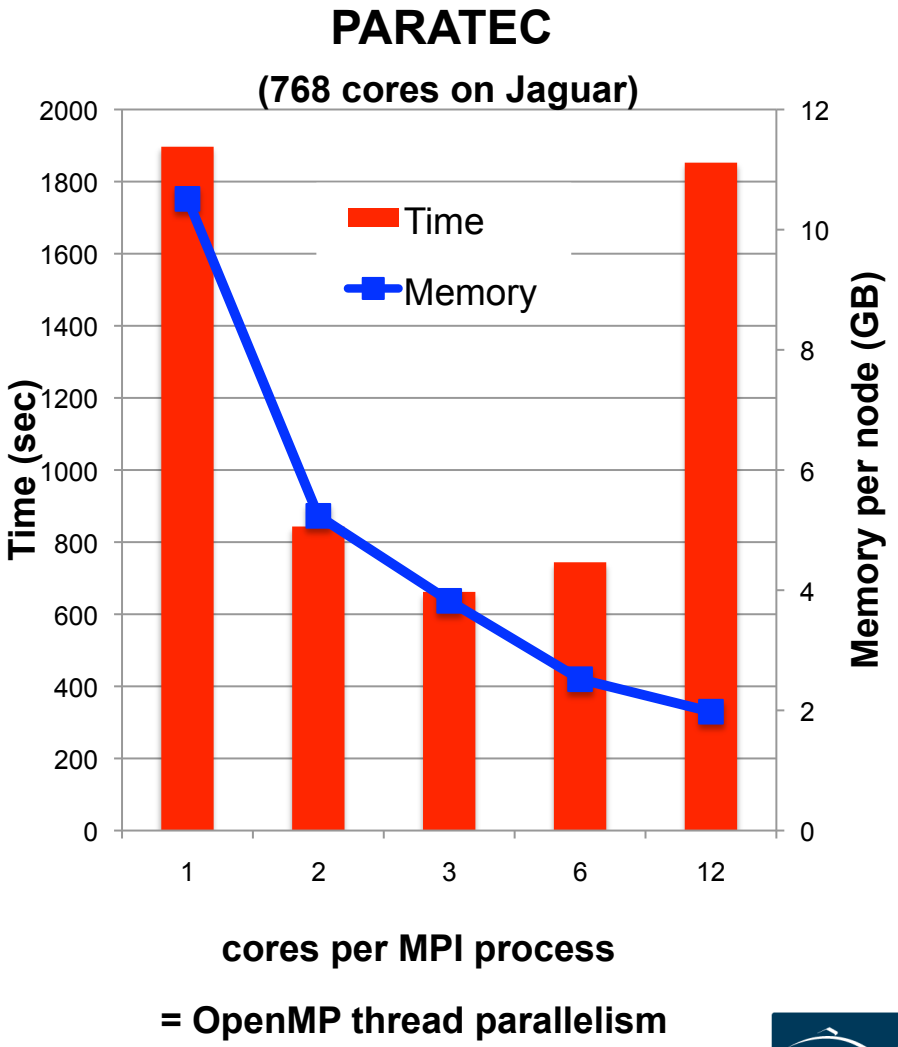
Develop Best Practices in Multicore Programming

Conclusions so far:

- Mixed OpenMP/MPI saves significant memory
- Running time impact varies with application
- 1 MPI process per socket is often good

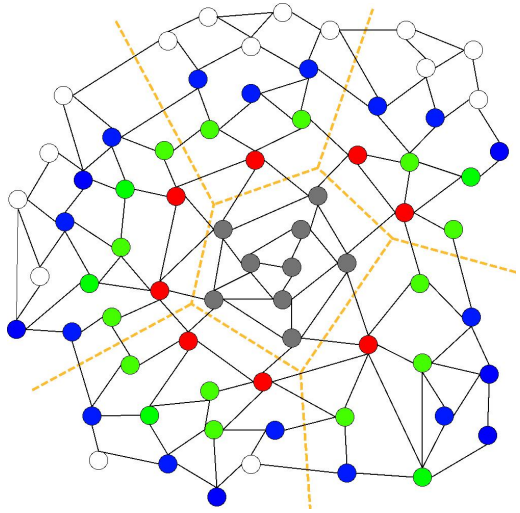
Run on Hopper next:

- 12 vs 6 cores per socket
- Gemini vs. Seastar



Communication-Avoiding Algorithms

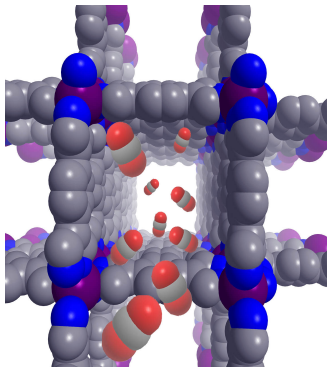
- **Sparse Iterative (Krylov Subpace) Methods**
  - Nearest neighbor communication on a mesh
  - Dominated by time to read matrix (edges) from DRAM
  - And (small) communication and global synchronization events at each step
- **Can we lower data movement costs?**
  - Take *k* steps with one matrix read from DRAM and one communication phase
    - Serial: O(1) moves of data moves vs. O(k)
    - Parallel: O(log p) messages vs. O(k log p)
- **Can we make communication provably optimal?**
  - Communication both to DRAM and between cores
  - Minimize independent accesses ('latency')
  - Minimize data volume ('bandwidth')



Joint work with Jim Demmel, Mark Hoemman, Marghoob Mohiyuddin

Provide GPU Testbed and Evaluation

- **Installed "Dirac" GPU testbed**
  - About100 users so far
  - Popular with SciDAC-E postdocs
- **Example: Q-Chem Routine**
  - Impressive single node speedups relative to 1 core on CPU
  - Highly variable with input structure

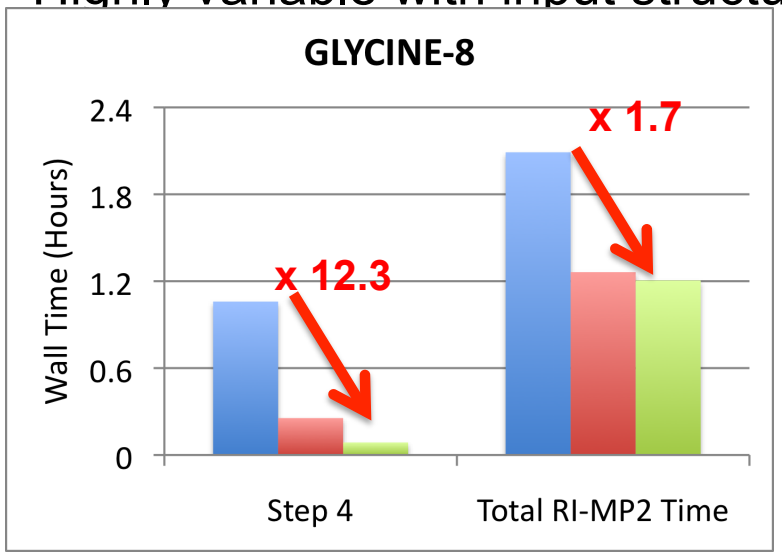


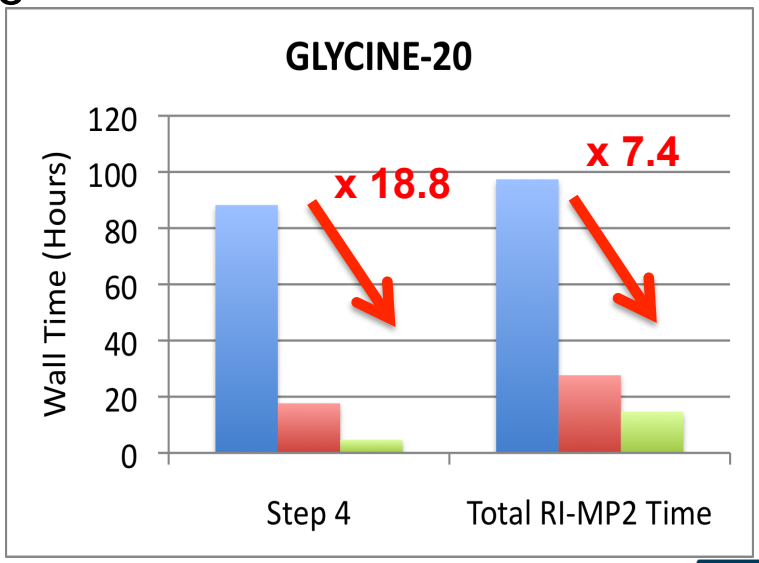
Fermi GPU Racks - NERSC

Blue:
CPU 1
thread

Red:
CPU 8
threads

Green:
GPU





Jihan Kim (SciDAC-E NERSC Postdoc)

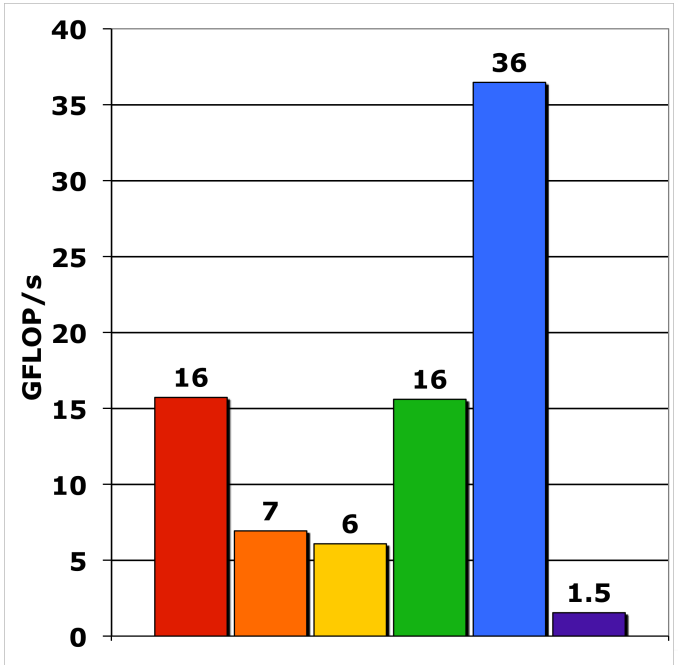
Are GPUs the Future?

(Includes Cell and GPU)

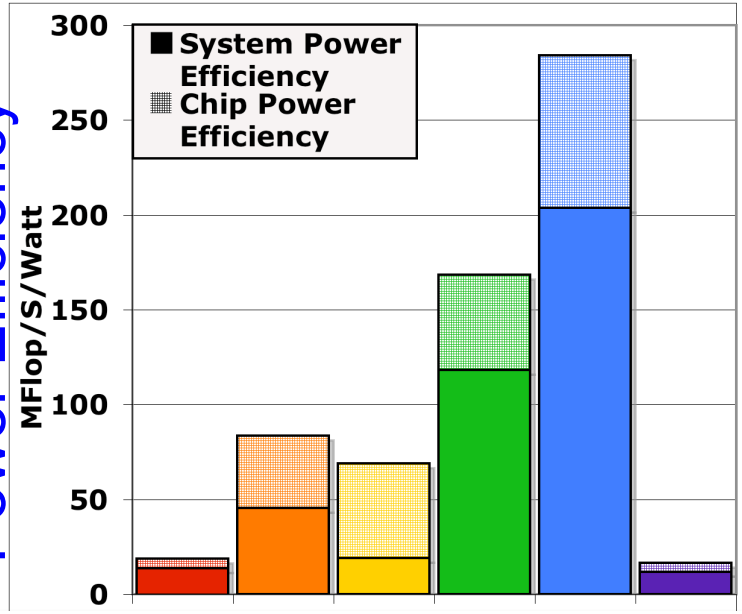
K. Datta, M. Murphy, V. Volkov, S. Williams , J. Carter, L. Oliker. D. Patterson, J. Shalf, K. Yelick, BDK11 book

1.7x speedup versus optimized Nehalem (C2050 w/ECC)

Performance



Power Efficiency

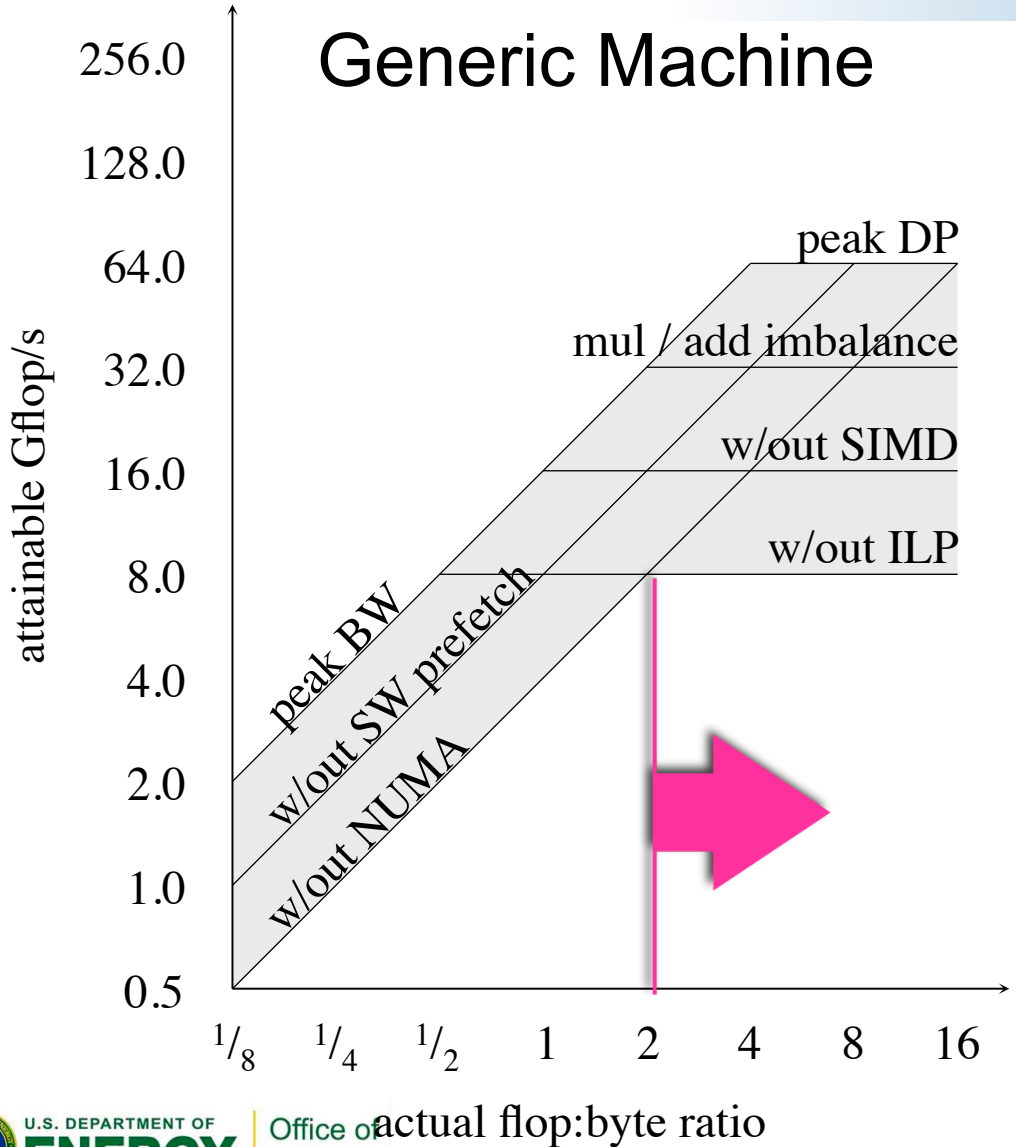


Cache-based	Local store-based
Gainestown	Cell Blade
Barcelona	GTX280
Victoria Falls	GTX280-Host

U.S. DEPARTMENT OF ENERGY | Office of Science

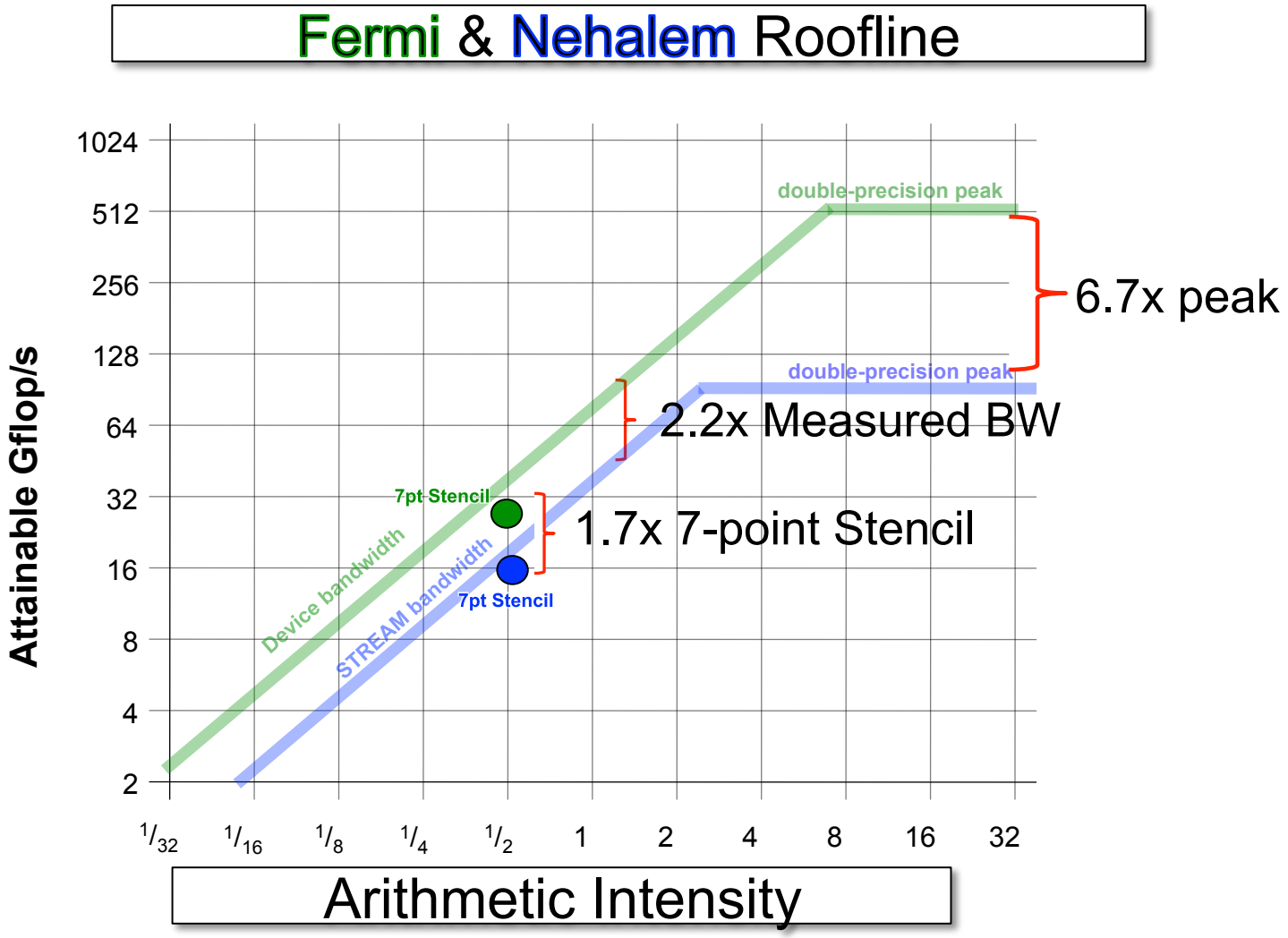
BERKELEY LAB
Lawrence Berkeley National Laboratory

The Roofline Performance Model

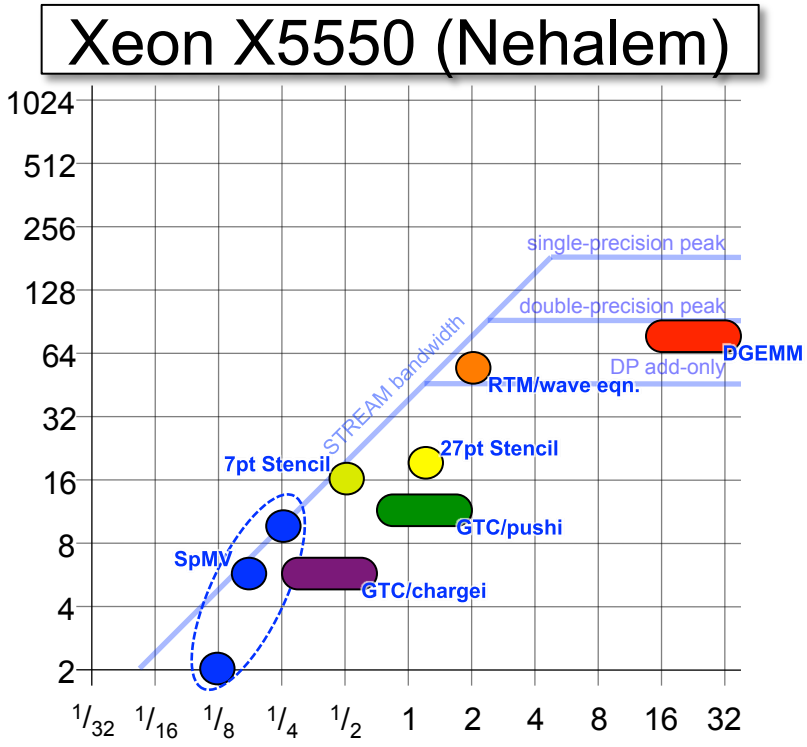


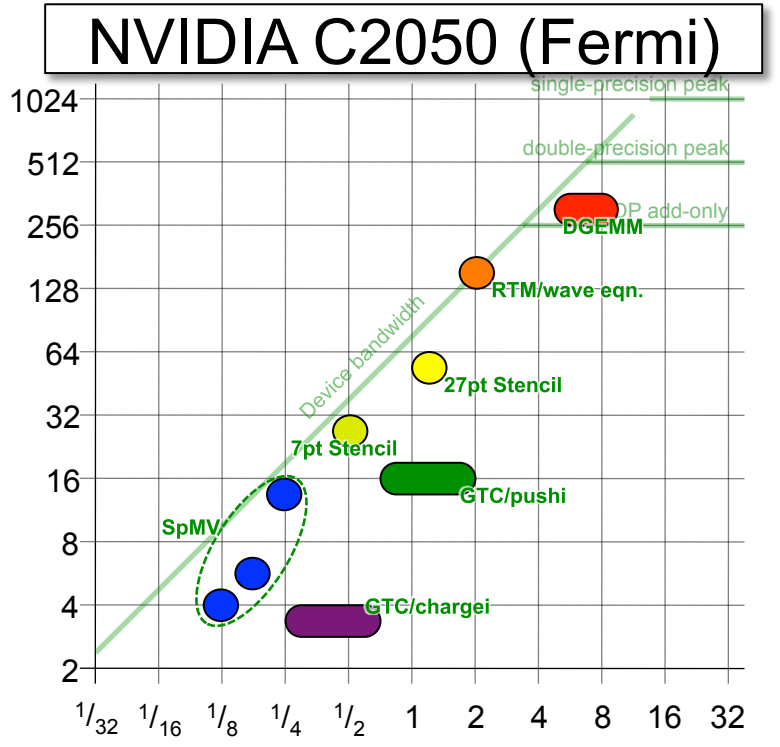
- The flat room is determined by arithmetic peak and instruction mix
- The sloped part of the roof is determined by peak DRAM bandwidth (STREAM)
- X-axis is the computational intensity of your computation

Relative Performance Expectations



Relative Performance Across Kernels

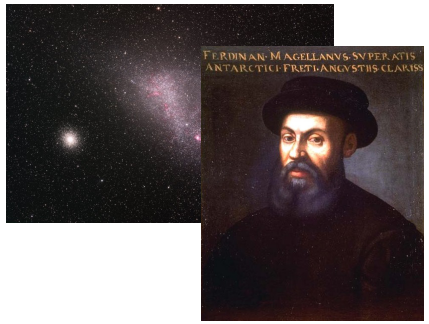




U.S. DEPARTMENT OF ENERGY | Office of Science

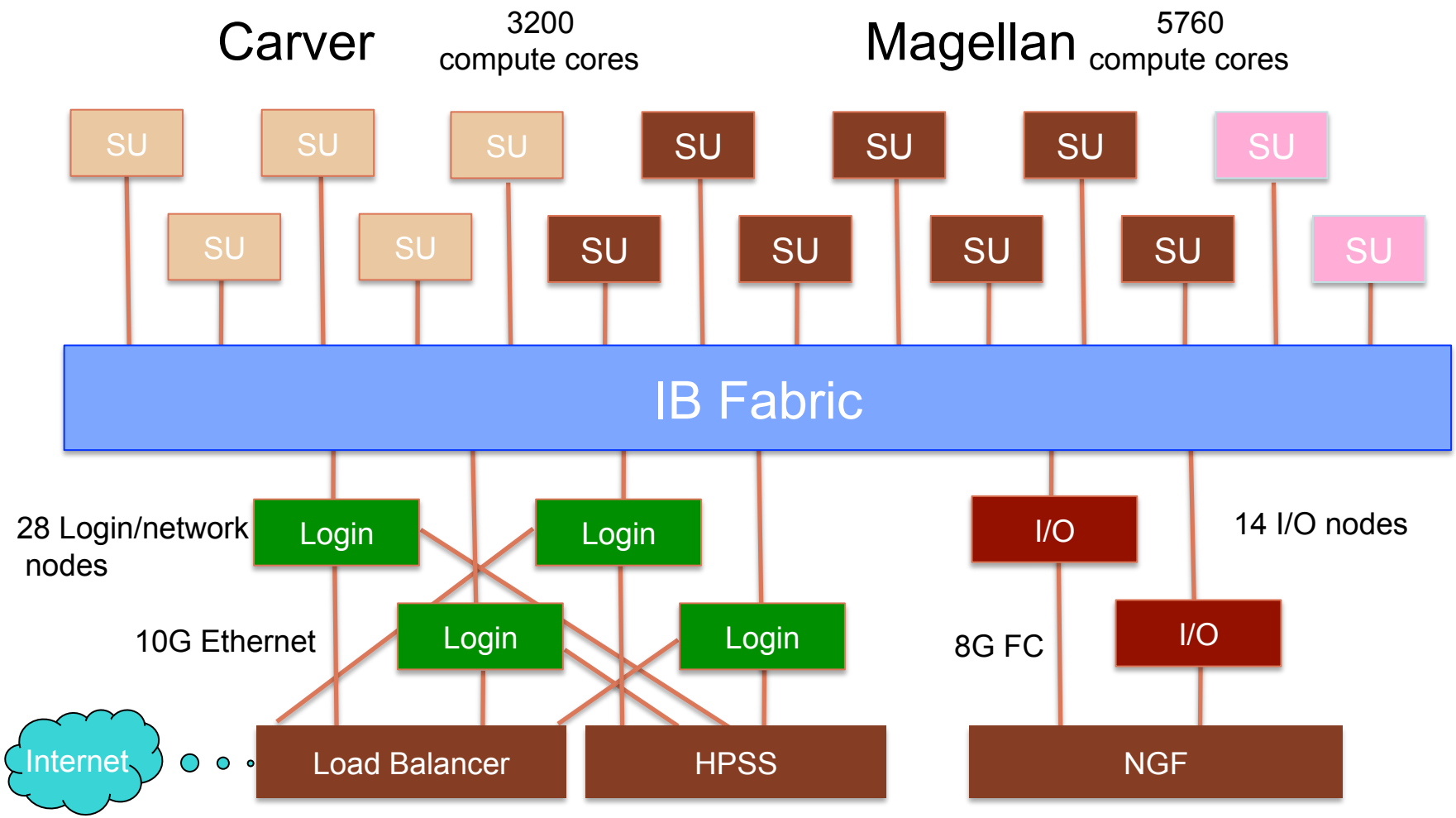
BERKELEY LAB
Lawrence Berkeley National Laboratory

DOE Explores Cloud Computing



- DOE's CS program focuses on HPC
  - No coordinated plan for clusters in SC
- DOE Magellan Cloud Testbed
- Cloud questions to explore:
  - Can a cloud serve DOE's mid-range computing needs?
  - What features (hardware and software) are needed of a "Science Cloud"? Commodity hardware?
  - What requirements do the jobs have (~100 cores, I/O,…)
  - How does this differ, if at all, from commercial clouds?
- What are main attractions?
  - Elastic computing: good business model for fixed computational problems, but what about science?
  - Control over software stack (virtualization)

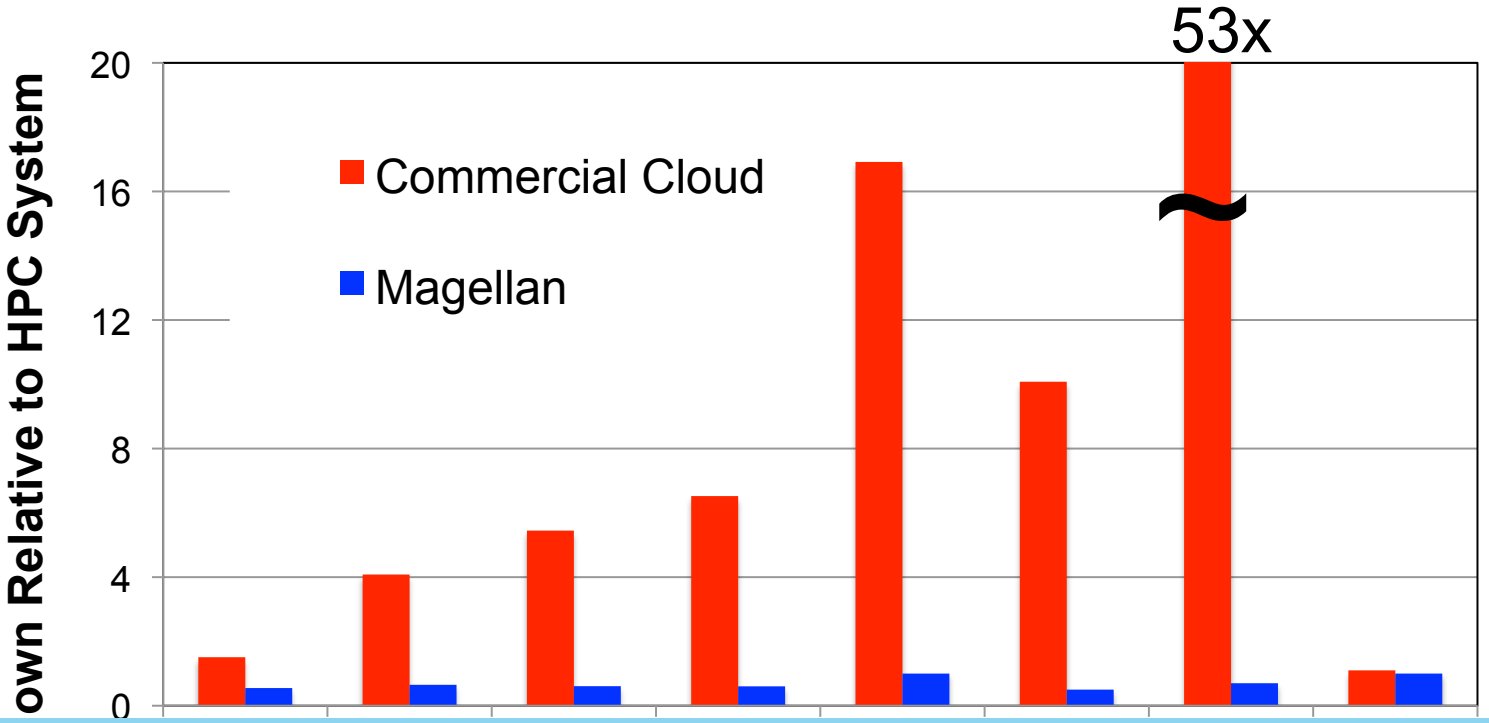
Cluster architecture



U.S. DEPARTMENT OF ENERGY | Office of Science

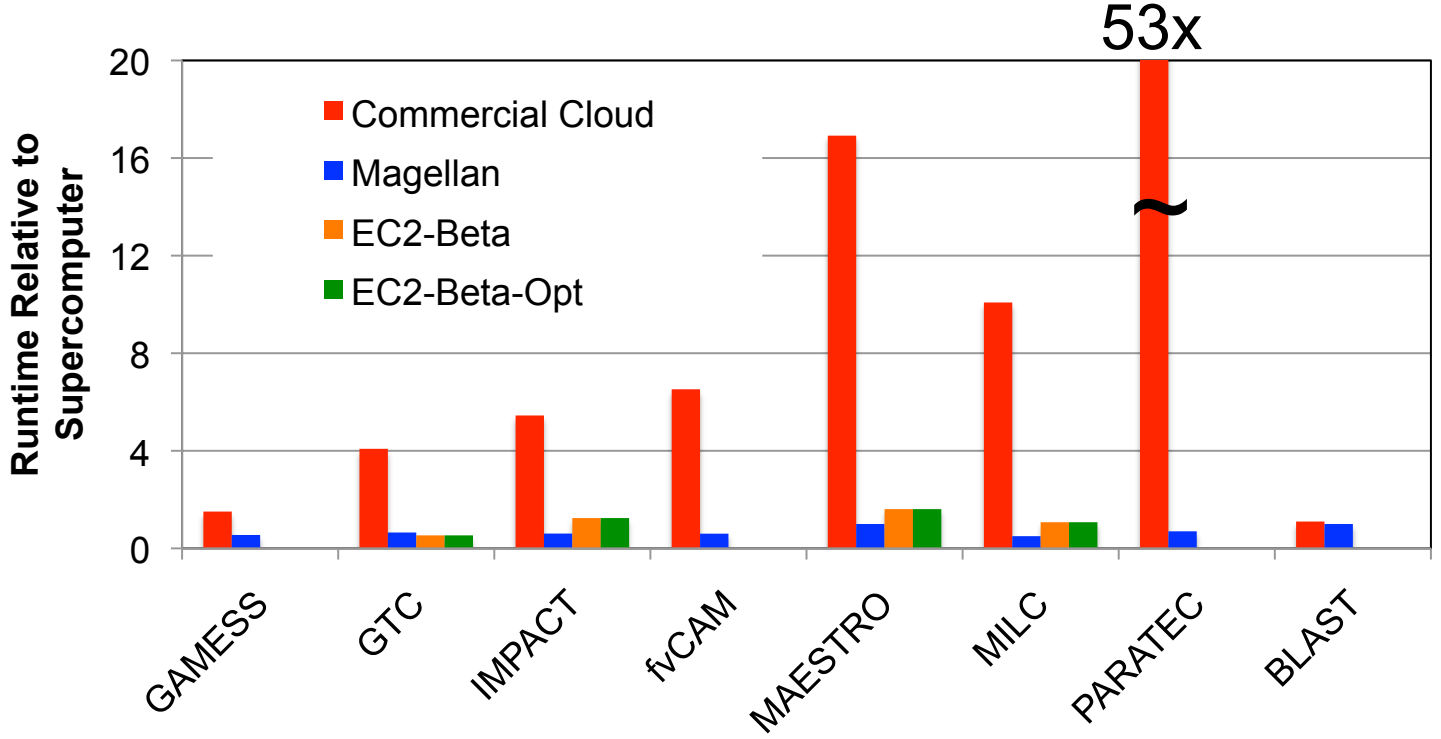
BERKELEY LAB
Lawrence Berkeley National Laboratory

Slowdown of Clouds Relative to an HPC System



- A cloud is a cluster. For modest (100s-way) parallel jobs, it is competitive with an HPC system.
- It needs a good network (e.g., Infiniband) and scheduler (batch) without virtualization

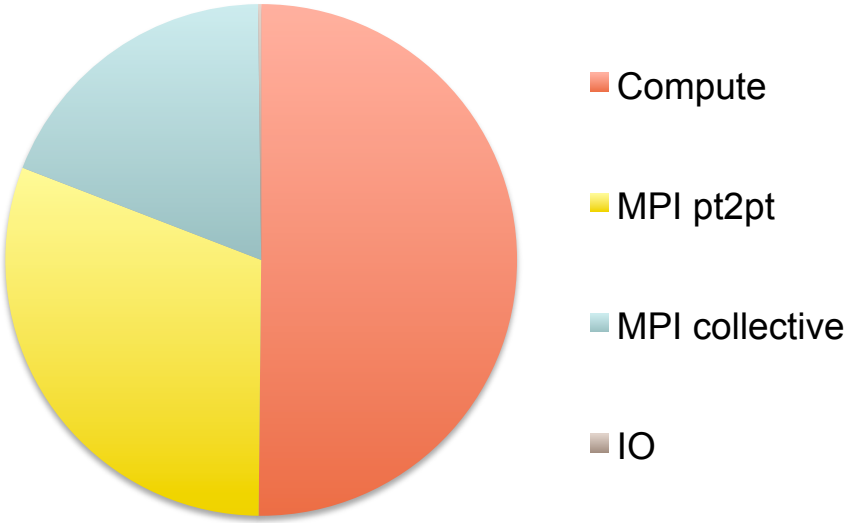
HPC Commercial Cloud Results



- **Commercial HPC clouds catch up with clusters if set up as shared cluster**
  - **High speed network (10GigE) and no over-subscription**

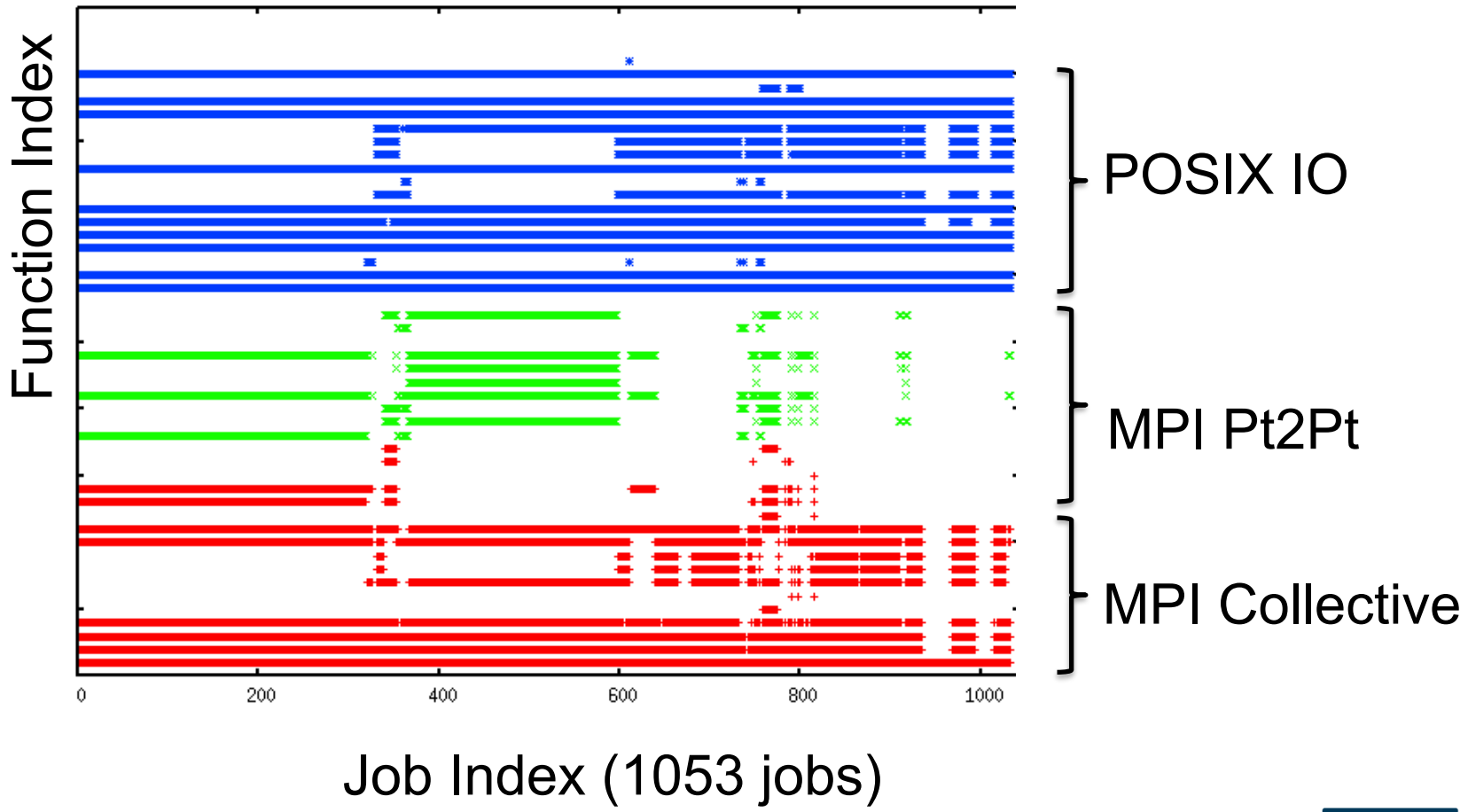
Workload Analysis on Magellan

- **Most HPC profiling is done on an opt-in basis. Users deploy tools to understand their code.**
- **Magellan profiling is system wide, passive, and automatic, a workload approach.**
- **October 4-27 :**
  - **1053 batch jobs**
  - **37 users**
  - **18 applications**
  - **4K cores**
  - **Preliminary results(*)**



(*) does not yet include non-MPI jobs

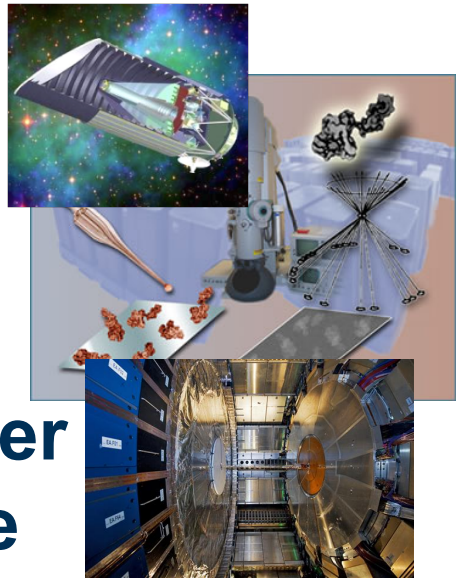
Workload Coverage: which jobs use which resources



IPM tool by David Skinner, Karl Fuerlinger, Nick Wright

What HPC Can Learn from Clouds

- Need to support surge computing
  - Predictable: monthly processing of genome data; nightly processing of telescope data
  - Unpredictable: computing for disaster recovery; response to facility outage



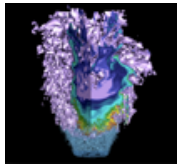
- Support for tailored software stack
- Different levels of service
  - Virtual private cluster: guaranteed service
  - Regular: low average wait time
  - Scavenger mode, including preemption

Conclusions

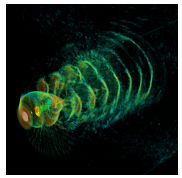
- **NERSC requirements**
  - Qualitative requirements shape NERSC functionality
  - Quantitative requirements set the performance

    "What gets measure gets improved"
- **Goals:**
  - Your goal is to make scientific discoveries
    - Articulate specific scientific goals and implications for broader community
  - Our goal is to enable you to do science
    - Specify resources (services, computers, storage, …) that NERSC could provide with quantities and dates

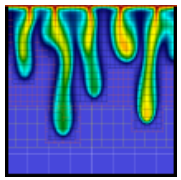
About the Cover



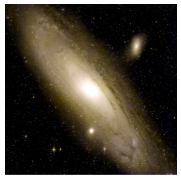
Low swirl burner combustion simulation. Image shows flame radical, OH (purple surface and cutaway) and volume rendering (gray) of vortical structures. Red indicates vigorous burning of lean hydrogen fuel; shows cellular burning characteristic of thermodiffusively unstable fuel. Simulated using an adaptive projection code. Image courtesy of John Bell, LBNL.



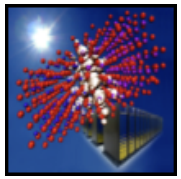
Hydrogen plasma density wake produced by an intense, right-to-left laser pulse. Volume rendering of current density and particles (colored by momentum orange - high, cyan - low) trapped in the plasma wake driven by laser pulse (marked by the white disk) radiation pressure. 3-D, 3,500 Franklin-core, 36-hour LOASIS experiment simulation using VORPAL by Cameron Geddes, LBNL. Visualization: Gunther Weber, NERSC Analytics.



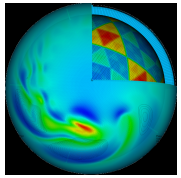
Numerical study of density driven flow for CO_2 storage in saline aquifers. Snapshot of CO_2 concentration after convection starts. Density-driven velocity field dynamics induces convective fingers that enhance the rate by which CO_2 is converted into negatively buoyant aqueous phase, thereby improving the security of CO_2 storage. Image courtesy of George Pau, LBNL



False-color image of the Andromeda Galaxy created by layering 400 individual images captured by the Palomar Transient Factory (PFT) camera in February 2009. NERSC systems analyzing the PTF data are capable of discovering cosmic transients in real time. Image courtesy of Peter Nugent, LBNL.



The exciton wave function (the white isosurface) at the interface of a ZnS/ZnO nanorod. Simulations performed on a Cray XT4 at NERSC, also shown. Image courtesy of Lin-Wang Wang, LBNL.



Simulation of a global cloud resolving model (GCRM). This image is a composite plot showing several variables: wind velocity (surface pseudocolor plot), pressure (b/w contour lines), and a cut-away view of the geodesic grid. Image courtesy of Professor David Randall, Colorado State University.