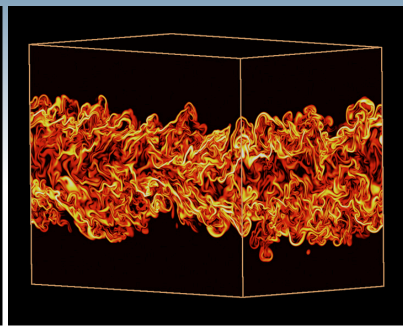
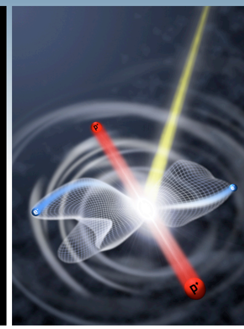
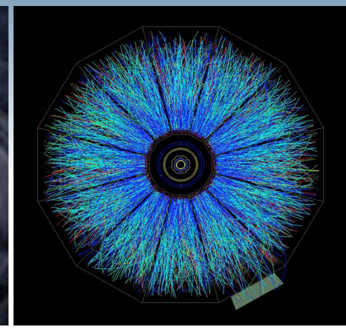
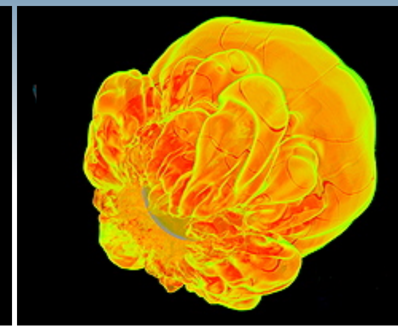






NERSC: National Energy Research Scientific Computing Center

HIPACC Workshop, Berkeley, CA, July 18th, 2011

Katie Antypas

Group Leader, NERSC User Services

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

NERSC Facility Leads DOE in Scientific Computing Productivity

NERSC computing for science

- 4000 users, 500 projects
- From 48 states; 65% from universities
- Hundreds of users each day
- ***1500 publications per year***

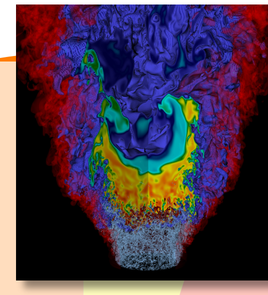
Systems designed for science

- 1.3PF Petaflop Cray system, Hopper
  - 2^{nd} Fastest computer in US
  - Fastest open Cray XE6 system
  - Additional .5 PF in Franklin system and smaller clusters

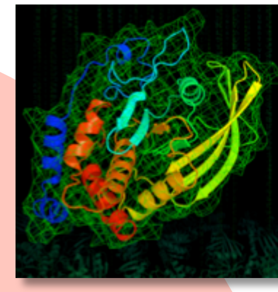
NERSC Serves the Computing and Data Needs of Science

- **NERSC provides computing, data, and consulting services for science**
- **Allocations managed by DOE based on mission priorities**

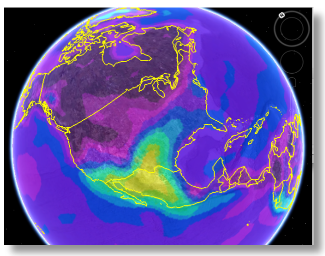
HEP 16%
NP 11%
BER 19%
FES 19%
BES 31%



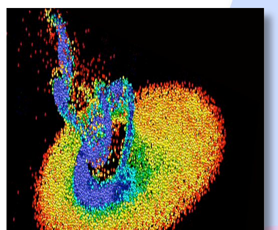
Burning structure in hydrogen leads to pockets of emissions



11,000 protein foldings, show common feature in amyloid development,

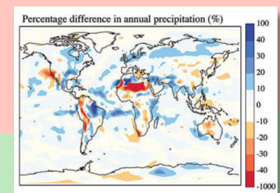


20th Century 3D climate maps reconstructed and in public database

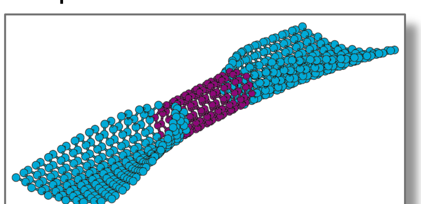


Location of dark companion to Milky Way found

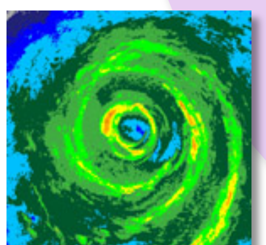
Supernova ignition depends on dimensionality of neutrino heating



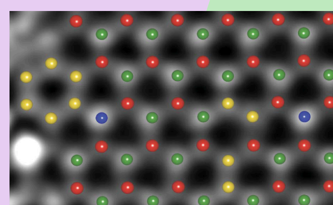
Flowering plants cool the earth



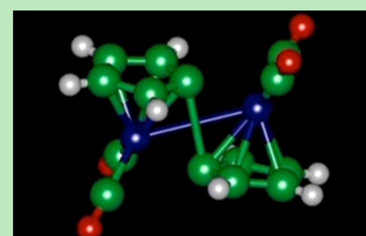
Carbon-based transistor junction created



Higher temperatures in Pliocene era linked to cyclones



Experiments+simulations "show" individual atoms of boron, carbon, & nitrogen.



Candidate molecule for reversible storage of solar energy identified

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

NERSC Systems

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

- 6,384 compute nodes, 153,216 cores
- 120 Tflop/s on applications; 1.3 Pflop/s peak

Clusters

140 Tflops total

Carver

- IBM iDataplex cluster

PDSF (HEP/NP)

- ~1K core cluster

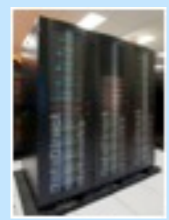
Magellan Cloud testbed

- IBM iDataplex cluster

GenePool (JGI)

- ~5K core cluster

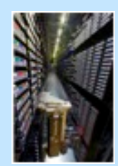
NERSC Global Filesystem (NGF)



Uses IBM's GPFS

- 1.5 PB capacity
- 5.5 GB/s of bandwidth

HPSS Archival Storage



- 40 PB capacity
- 4 Tape libraries
- 150 TB disk cache

Analytics

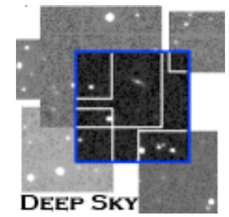
Euclid

(512 GB shared memory)

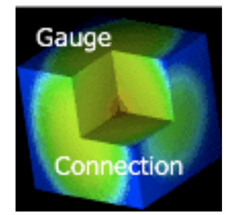
Dirac GPU testbed (48 nodes)

Develop and Provide Science Gateway Infrastructure

- **Goals of Science Gateways**
  - Allow sharing of data on NGF and HPSS
  - Make scientific computing easy
  - Broaden impact/quality of results from experiments and simulations
- **NEWT – NERSC Web Toolkit/API**
  - Building blocks for science on the web
  - Write a Gateway: HTML + Javascript
- **30+ projects use the NGF -> web**

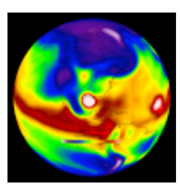


Deep Sky: 450+ Supernovae



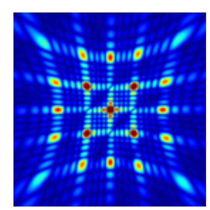
Gauge Connection: QCD

Daya Bay: Real-time processing and monitoring



20th Century Reanalysis

Earth Systems Grid



Coherent X-Ray Imaging Data Bank

U.S. DEPARTMENT OF ENERGY | Office of Science

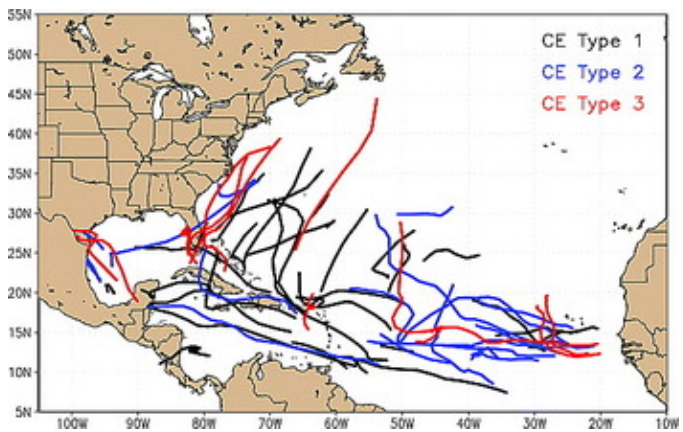
BERKELEY LAB

20th Center Climate Data Reconstructed

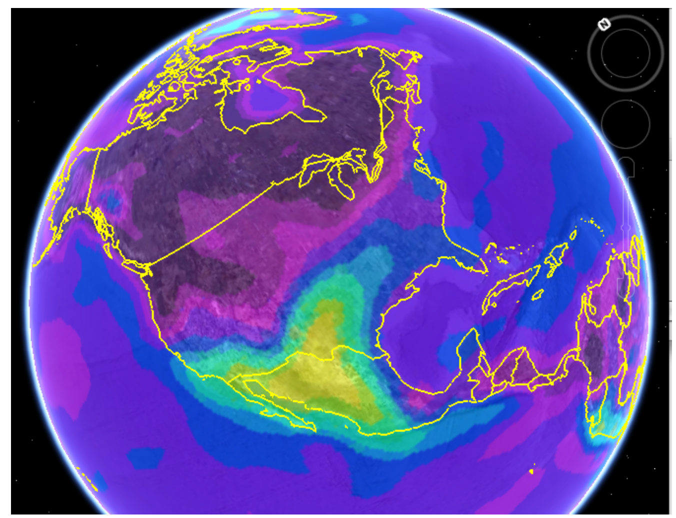
Reconstructed global weather conditions in 6-hour intervals from 1871-2010

- **Based on data from meteorologists, military, volunteers and ships' crews**
- **Over 10M hours at NERSC using reverse Kalman filter algorithms**
- **Data used in 16 papers to date: reproduced 1922 Knickerbocker storm, understand causes of the 1930 Dust Bowls, and determine whether recent extremes are sign of climate change**

NERSC has 2PB of online storage and up to 44 PB of archive for scientific data sets. New "Science Gateways" make it easy to make data accessible on the web



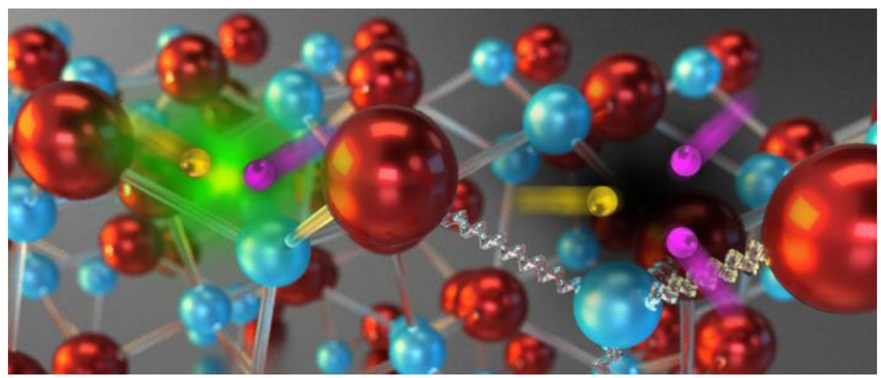
Previously undetected warm-core cyclones, *Geophys. Res. Letters*, 2011



Relative Humidity for 1920-1929
Gil Compo, PI (U. Colorado)

Material Science for Energy Efficient Lighting

- ***LEDs are up to 3x more energy efficient than fluorescent lights and last 10x longer***
  - ***"LED droop" makes them unusable for lighting rooms, since efficiency drops when current is scaled***
  - ***Cause? Auger recombination combined with carrier scattering.***
- **Science discovery explains cause of droop, allowing university and industry researchers to work on solutions.**



The illustration shows nitride-based LEDs. At left, an electron and electron hole recombine and release light. In Auger recombination (right) the electron and hole combine with a third carrier, releasing no photon. The energy loss is also assisted by indirect processes, vibrations in the crystal lattice shown as squiggles.

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

HPC Architecture

Why Do You Care About Architecture?

- **To use HPC systems well, you need to understand the basics and conceptual design**
  - Otherwise, too many things are mysterious
- **Programming for HPC systems is hard**
  - To get your code to work properly
  - To make it run efficiently (performance)
- **You want to efficiently configure the way your job runs**
- **The technology is cutting edge**

Definitions & Terminology

- **HPC**
  - High Performance Computing
  - Scientific computing at scale
- **CPU**
  - Central Processing Unit
  - Now ambiguous terminology
  - Generic for "some unit that computes"
  - Context-sensitive meaning
- **Core**
  - Hardware unit that performs arithmetic operations
  - A CPU may have more than one core
- **Die**
  - An integrated circuit manufactured as a unit
  - Many cores may be included on a die
- **Socket**
  - A physical package that connects to a computer board
  - A socket package may be composed of multiple dies

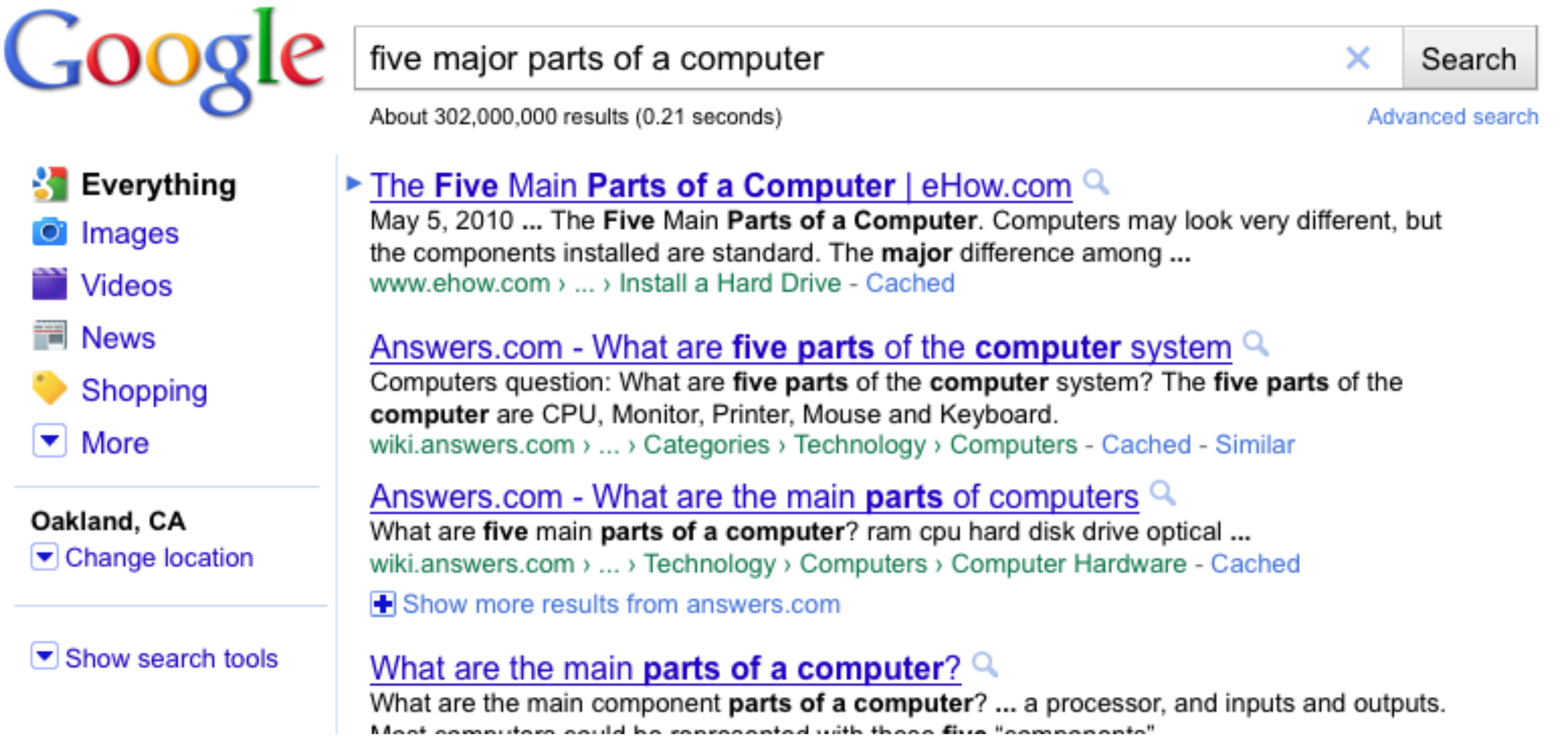
Definitions & Terminology

- **Memory**
  - Volatile storage of data or computer instructions
- **Bandwidth**
  - The rate at which data is transferred between destinations (typically GB/s)
- **Latency**
  - The time needed to initialize a data transfer (ranges from 10^{-9} to 10^{-6} secs or more)
- **FLOP: Floating Point Operation**
  - e.g., `a+b`, `a*b+c`
  - FLOPs/sec is a common performance metric
- **Interconnect**
  - A high-performance data network that connects nodes to each other and possibly other devices

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB | Lawrence Berkeley National Laboratory

What are the “5 major parts”?



Five Major Parts

<table>
<tr><th>eHow.com</th><th>Answers.com</th><th>Fluther.com</th><th>Yahoo!</th><th>Wikipedia</th></tr>
<tr><td>CPU</td><td>CPU</td><td>CPU</td><td>CPU</td><td>Motherboard</td></tr>
<tr><td>RAM</td><td>Monitor</td><td>RAM</td><td>RAM</td><td>Power Supply</td></tr>
<tr><td>Hard Drive</td><td>Printer</td><td>Storage</td><td>Power Supply</td><td>Removable Media</td></tr>
<tr><td rowspan="2">Video Card</td><td rowspan="2">Mouse</td><td>Keyboard/ Mouse</td><td rowspan="2">Video Card</td><td rowspan="2">Secondary Storage</td></tr>
<tr><td>Monitor</td></tr>
<tr><td rowspan="2">Motherboard</td><td rowspan="2">Keyboard</td><td>Motherboard</td><td rowspan="2">Motherboard</td><td>Sound Card</td></tr>
<tr><td>Case / Power Supply</td><td>IO Peripherals</td></tr>
</table>

It Depends on Your Perspective

- What is a computer?
  - It depends what you are interested in.
    - CPU, memory, video card, motherboard, ...
    - Monitor, mouse, keyboard, speakers, camera, …
- We'll take the perspective of an application programmer or a scientist running a code on an HPC system
- What features of an HPC system are important for you to know about?

5 Major Parts of an HPC System

1. CPUs
2. Memory (volatile)
3. Nodes
4. Inter-node network
5. Non-volatile storage (disks, tape)

Hopper

NERSC-6
Grace “Hopper”

Cray XE6

Performance

1.3 PF Peak

1.05 PF HPL (#8)

Processor

AMD MagnyCours

2.1 GHz 12-core

8.4 GFLOPs/core

24 cores/node

32-64 GB DDR3-1333 per node

System

Gemini Interconnect (3D torus)

6384 nodes

153,216 total cores

I/O

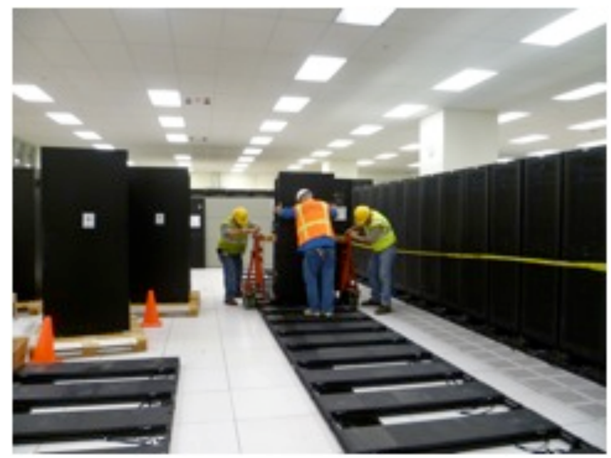
2PB disk space

70GB/s peak I/O Bandwidth

NERSC-6 Hopper

Hopper provides over 3 million computing hours per day to scientists

- 1.28 PFlop/s peak performance
- Over 1 billion annual core-hours facility wide
- Gemini high performance resilient interconnect
- Two 12-core AMD Magny-Cours chips per node
- Collaboration with NNSA ACES on testing



Hopper installation, August 2010

NERSC/Cray Center of Excellence

- Programming Models for Multicore systems
- Ensures effective use of new 24-core nodes

Evolution from Franklin (XT4) to Hopper (XE6)

Cray XT4: Franklin

Performance: 0.352 PF Peak
0.266 TF HPL (#27, debut@ #8)
Processor: AMD Budapest
4-core 2.3 GHz (9.2 GF/core)
4 cores/node
Memory: DDR2 667MHz
8 GB/node @ 21GB/s
2 GB/core
System
9,572 nodes (38,288 total cores)
Interconnect: SeaStar2 3D torus,
1.6GB/s measured @ 6-8usec
I/O
12GB/s peak I/O Bandwidth
0.436 PB disk space

Cray XE6: Hopper

Performance: 1.288 PF Peak
1.05 PF HPL (#8, debut@ #5)
Processor: AMD MagnyCours
12-core 2.1 GHz (8.4 GF/core)
24 cores/node
Memory: DDR3 1333MHz
32-64 GB/node @ 84GB/s
1.3 - 2.6 GB/core
System
6,384 nodes (153,216 total cores)
Interconnect: Gemini 3D torus,
8.3GB/s measured @ 2usec
I/O
70GB/s peak I/O Bandwidth
2PB disk space

Evolution from Franklin (XT4) to Hopper (XE6)

Cray XT4: Franklin

Performance: 0.352 PF Peak
0.266 TF HPL (#27, debut@ #8)
Processor: AMD Budapest
4-core 2.3 GHz (9.2 GF/core)
4 cores/node
Memory: DDR2 667MHz
8 GB/node @ 21GB/s
2 GB/core
System
9,572 nodes (38,288 total cores)
Interconnect: SeaStar2 3D torus,
1.6GB/s measured @ 6-8usec
I/O
12GB/s peak I/O Bandwidth
0.436 PB disk space

Cray XE6: Hopper

Performance: 1.288 PF Peak
1.05 PF HPL (#8, debut@ #5)
Processor: AMD MagnyCours
12-core 2.1 GHz (8.4 GF/core)
24 cores/node
Memory: DDR3 1333MHz
32-64 GB/node @ 84GB/s
1.3 - 2.6 GB/core
System
6,384 nodes (153,216 total cores)
Interconnect: Gemini 3D torus,
8.3GB/s measured @ 2usec
I/O
70GB/s peak I/O Bandwidth
2PB disk space

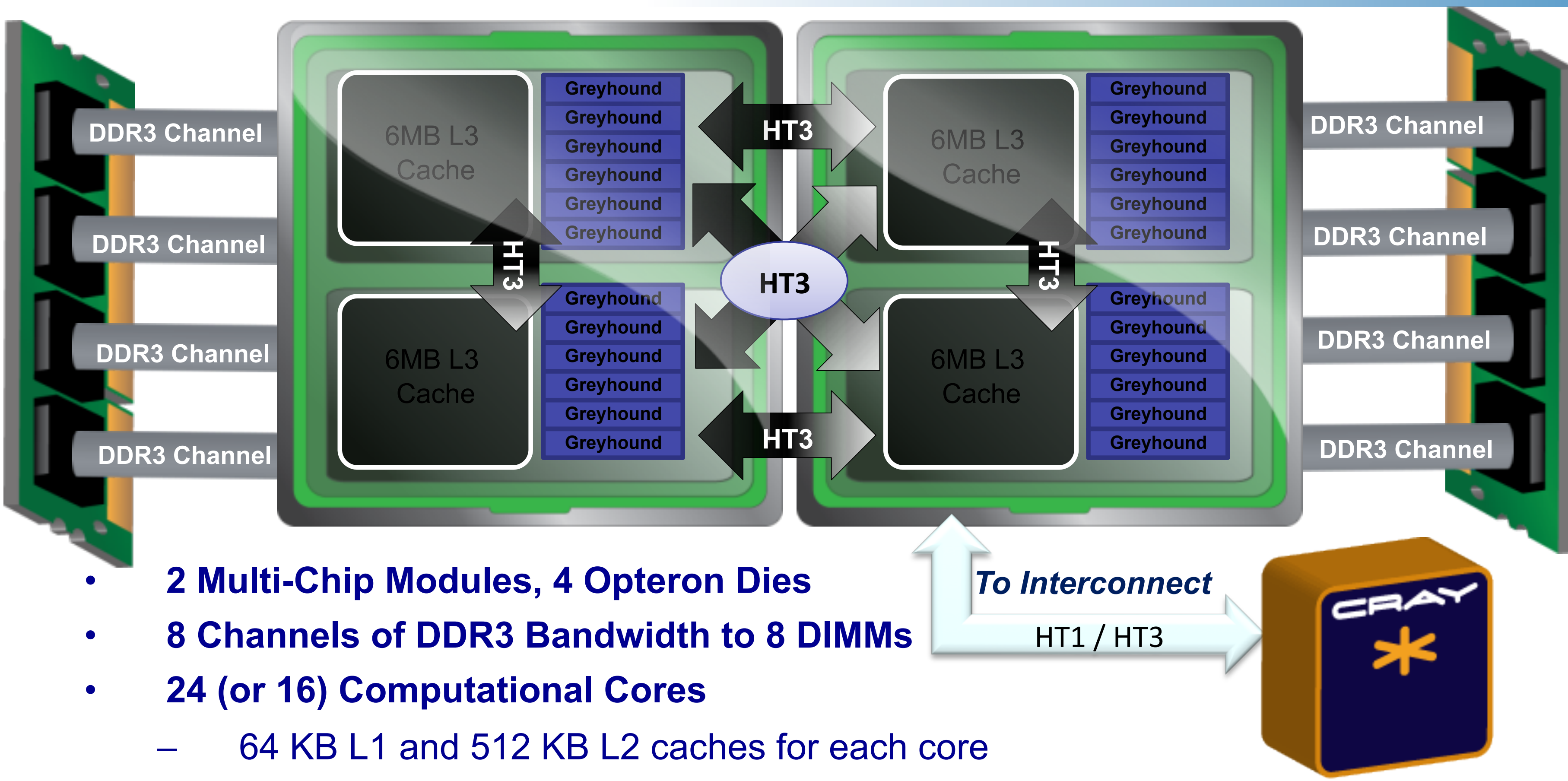
Preparing yourself for future hardware trends

- **CPU Clock rates are stalled (not getting faster)**
  - **# nodes is about the same, but # cores is growing exponentially**
  - **Think about parallelism from node level**
  - **Consider hybrid programming to tackle intra-node parallelism so you can focus on # of nodes rather than # of cores**

- **Memory capacity not growing as fast as FLOPs**
  - **Memory per node is still growing, but per core is diminishing**
  - **Threading (OpenMP) on node can help conserve memory**

- **Data locality becomes more essential for performance**
  - **NUMA effects (memory affinity: must always be sure to access data where it was first touched)**

XE6 Node Details: 24-core Magny Cours



- **2 Multi-Chip Modules, 4 Opteron Dies**
- **8 Channels of DDR3 Bandwidth to 8 DIMMs**
- **24 (or 16) Computational Cores**
  - 64 KB L1 and 512 KB L2 caches for each core
  - 6 MB of shared L3 cache on each die
- **Dies are fully connected with HT3**

Cray XE6 Compute Blade

- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor

U.S. DEPARTMENT OF ENERGY | Office of Science

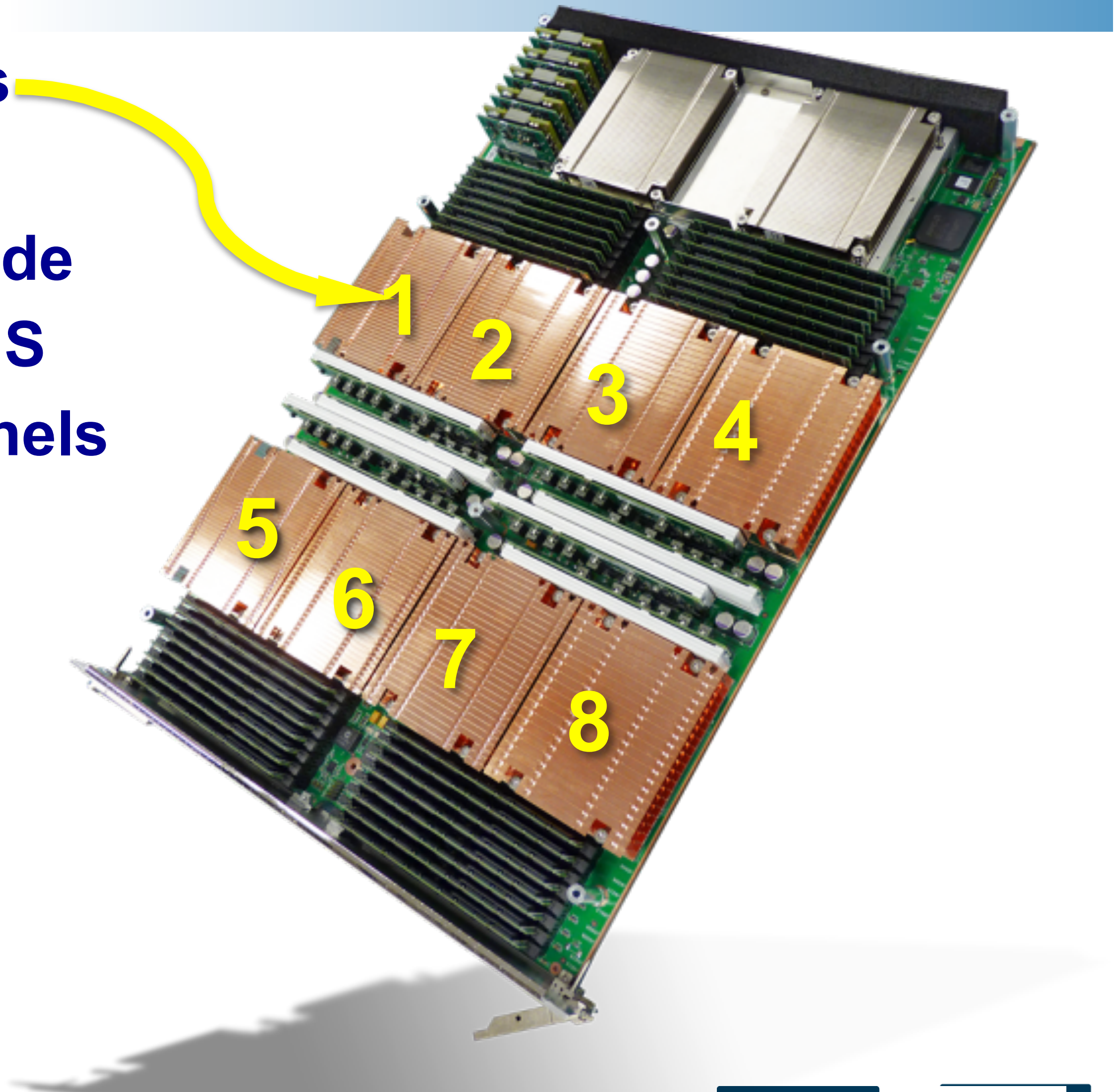
Images Courtesy of Cray Inc.

BERKELEY LAB

Lawrence Berkeley National Laboratory

Cray XE6 Compute Blade

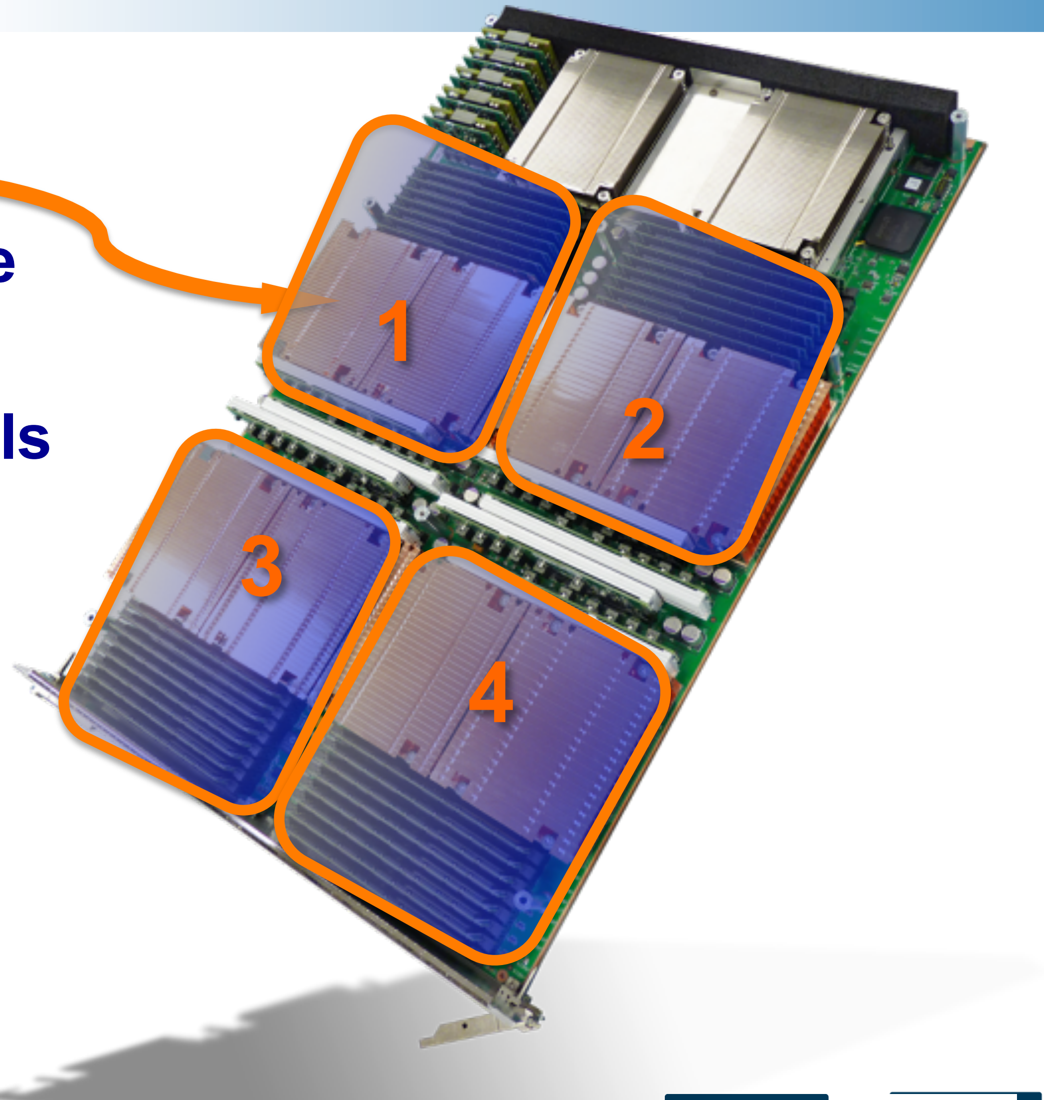
- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



Images Courtesy of Cray Inc.

Cray XE6 Compute Blade

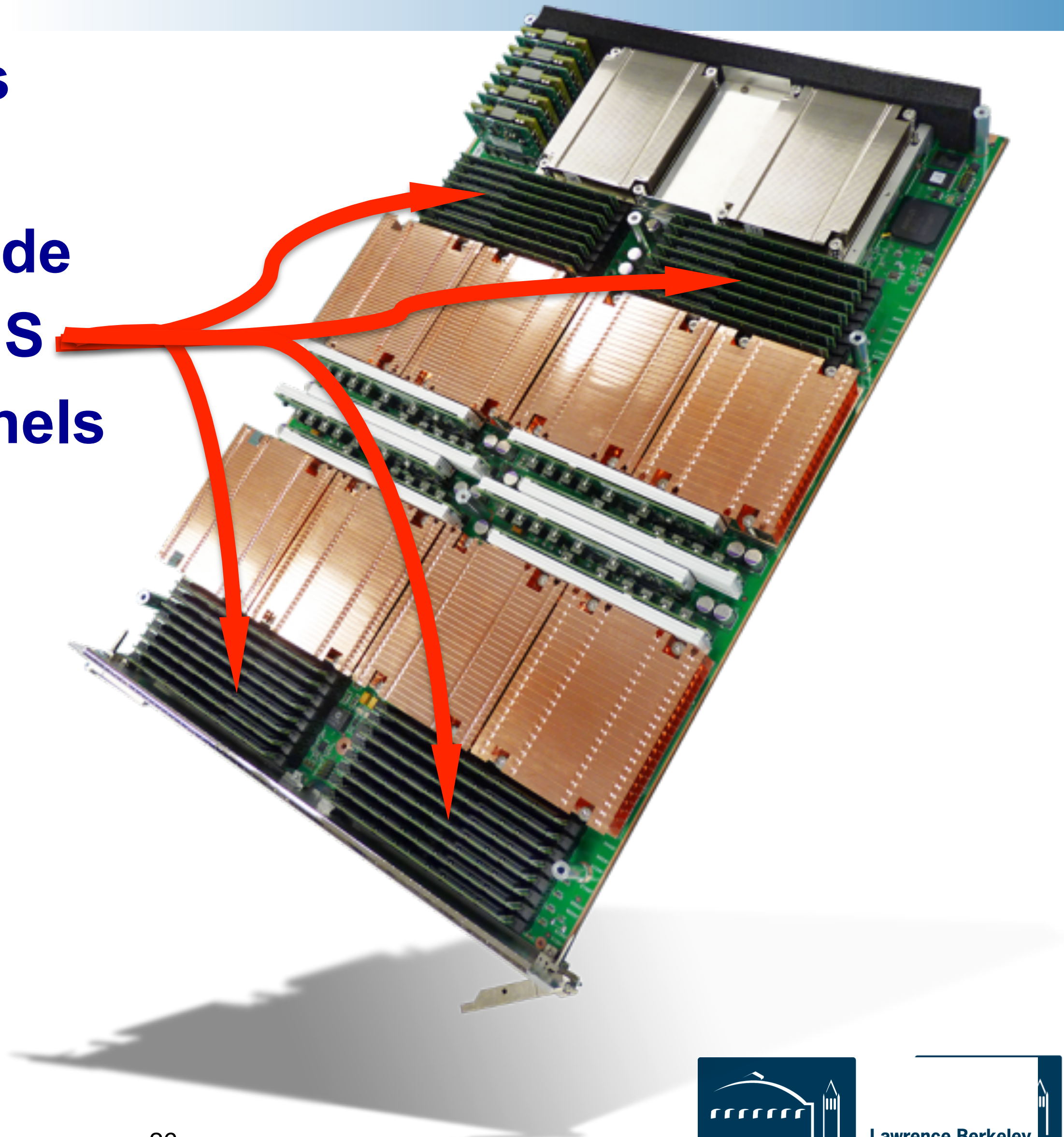
- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



Images Courtesy of Cray Inc.

Cray XE6 Compute Blade

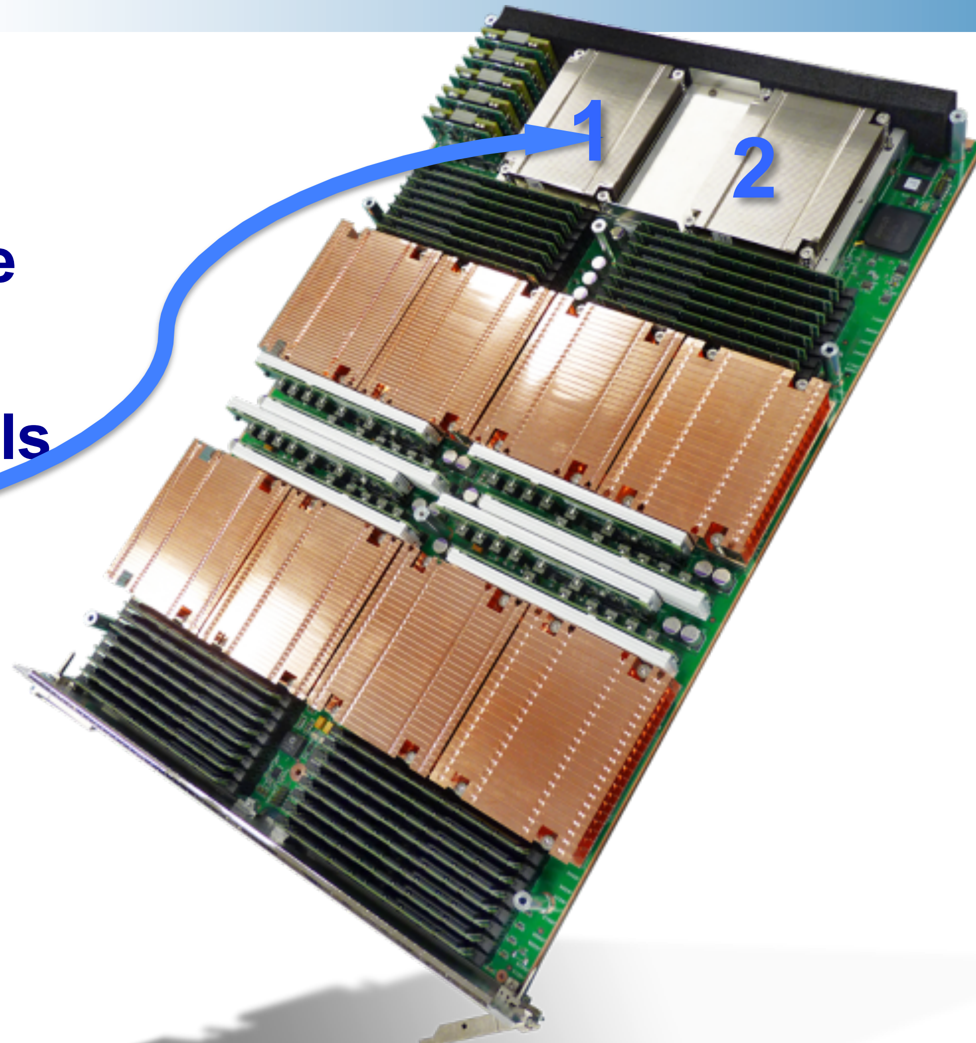
- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor



Images Courtesy of Cray Inc.

Cray XE6 Compute Blade

- 8 Magny Cours Sockets
- which == 4 Nodes
- 96 Compute Cores / blade
- 32 DDR3 Memory DIMMS
- 32 DDR3 Memory channels
- 2 Gemini ASICs
- L0 Blade management processor

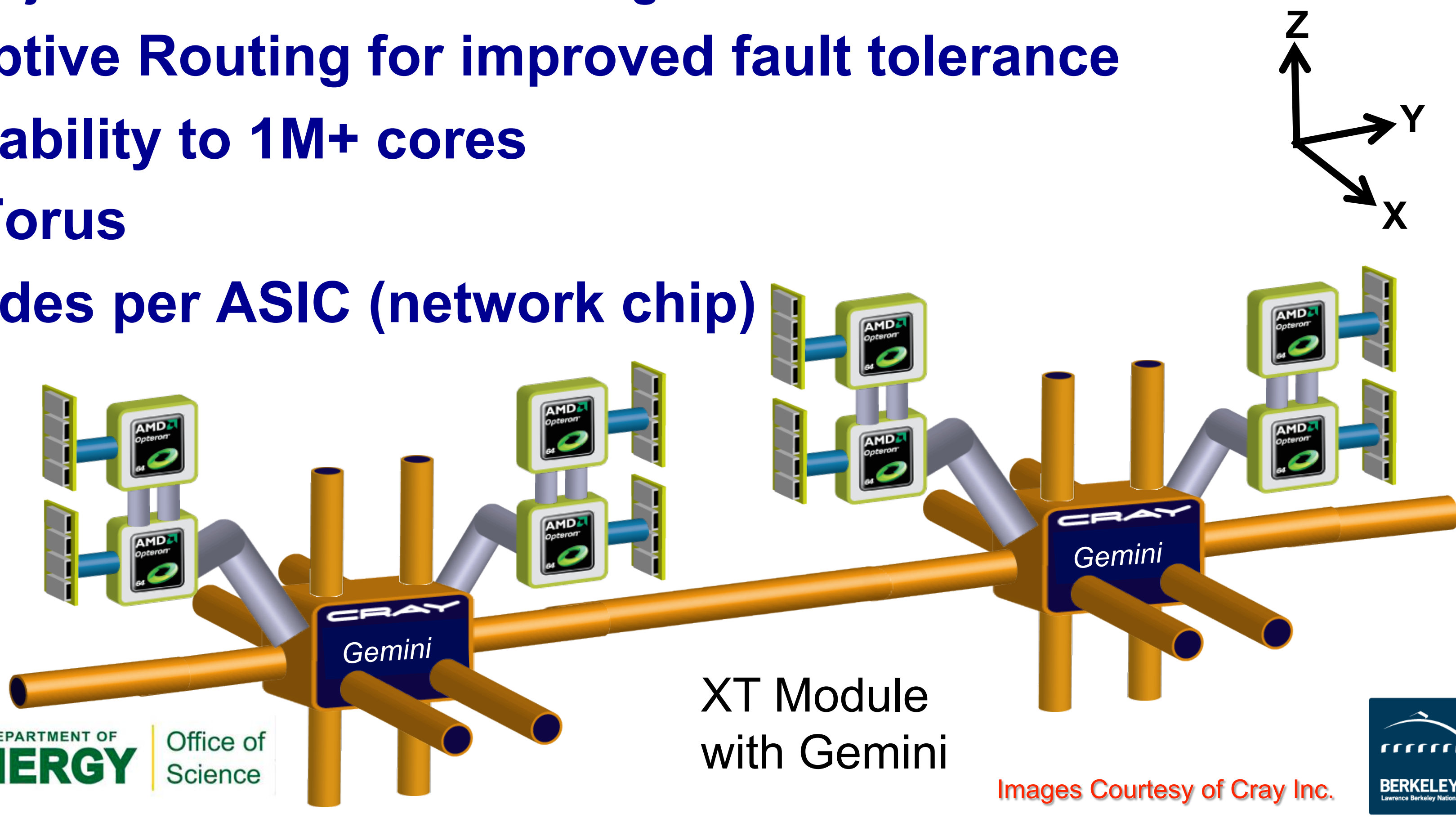


Images Courtesy of Cray Inc.

Interconnect

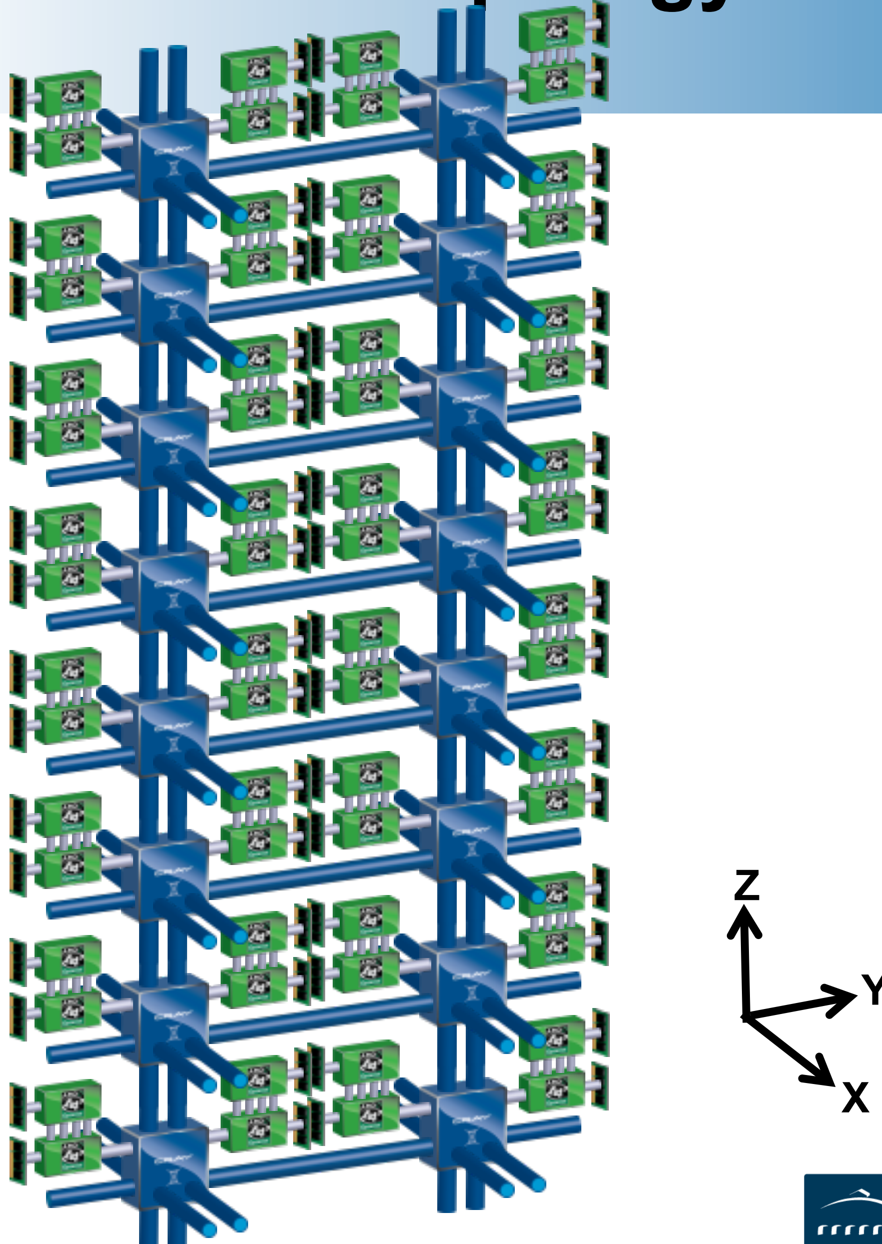
Hopper's Gemini Interconnect and Topology

- Performance
  - Latency 1-2 usec
  - Link bandwidth 9.3GB/s
  - Injection bandwidth from single node ~6GB/s
- Adaptive Routing for improved fault tolerance
- Scalability to 1M+ cores
- 3D Torus
- 2 nodes per ASIC (network chip)



Images Courtesy of Cray Inc.

Cray XE6 Chassis Topology



Images Courtesy of Cray Inc.

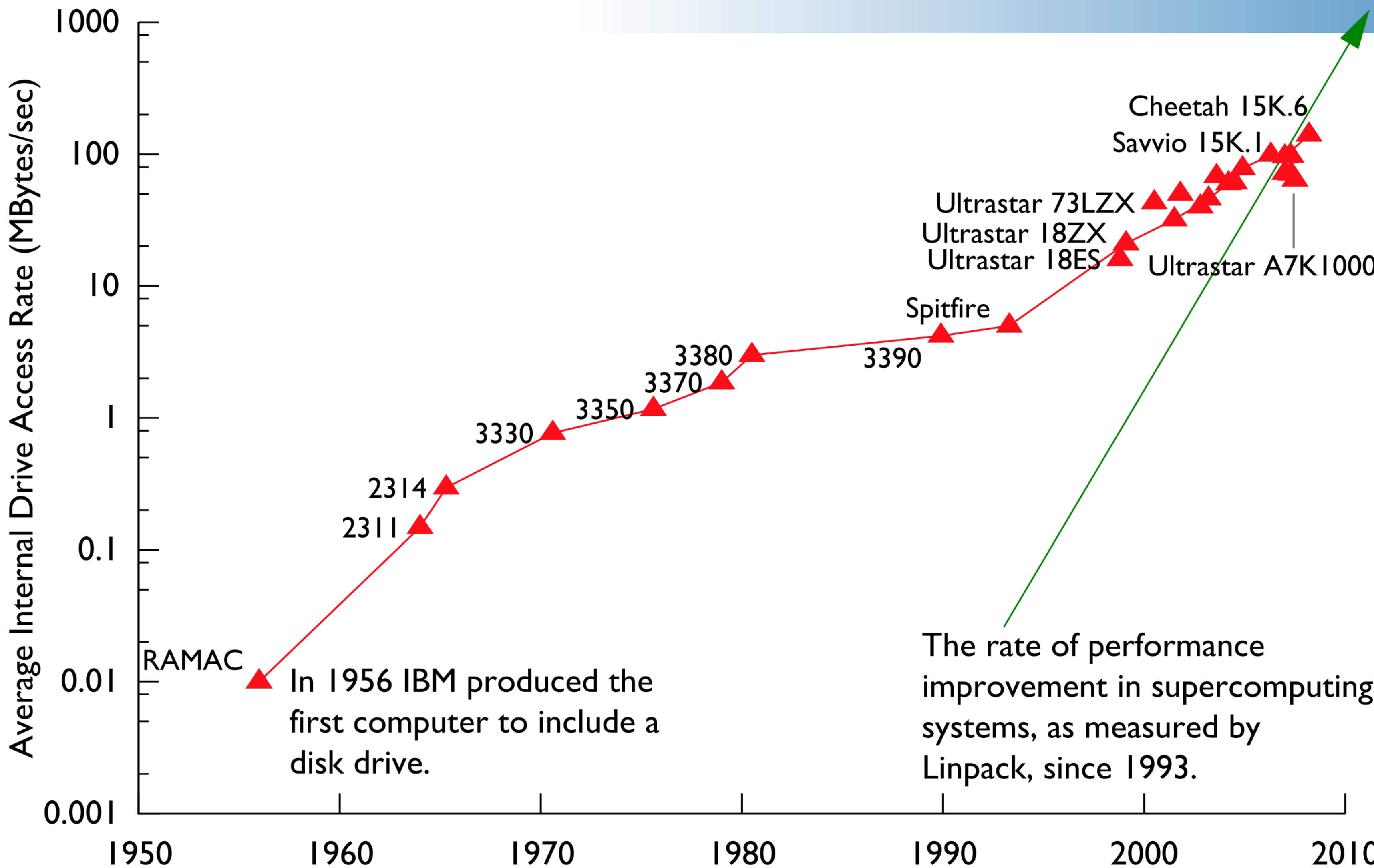
Wiring up the Cabinets

Images Courtesy of Cray Inc.

I/O

32

Disk Transfer Rates over Time

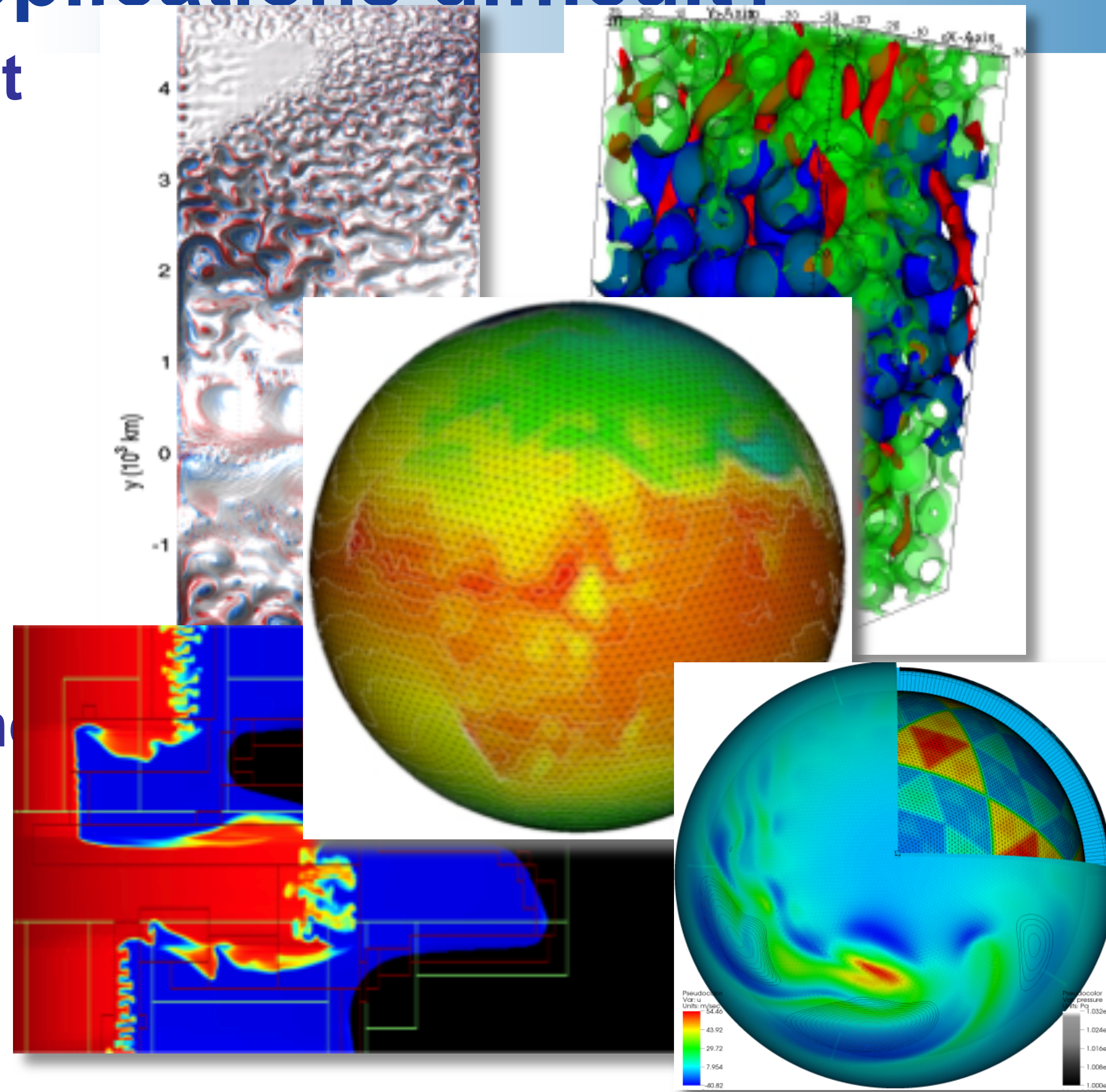


Thanks to R. Freitas of IBM Almaden Research Center for providing much of the data for this graph.

Slide from Rob Ross, Rob Latham at ANL

Why is Parallel I/O for science applications difficult?

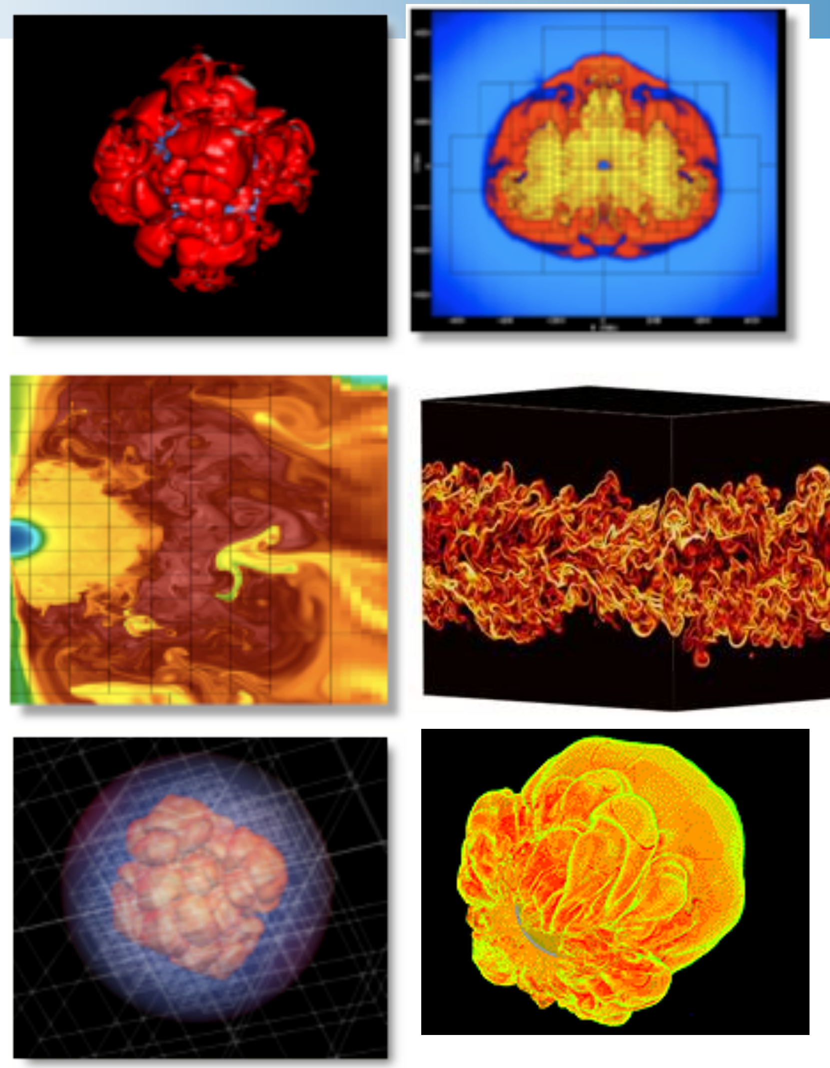
- Scientists think about data in terms of their science problem: molecules, atoms, grid cells, particles
- Ultimately, physical disks store bytes of data
- Layers in between, th application and physical disks are at various levels of sophistication



Images from David Randall, Paola Cessi, John Bell, T Scheibe

I/O in Astrophysics

- **What are the common characteristics of astrophysics applications?**
  - **Often have LOTS of data**
    - **Use all memory per core**
    - **Dump checkpoint and analysis files**
  - **Usually grid based**
    - **Structured/unstructure/adaptive grids**
    - **Can often collect data into large buffers and chunks**
    - **Regularly ordered, can be contiguous**
    - **Possible non-contiguous data with 3d decomposition**
  - **Particles data can be irregular**
  - **Some applications are out of out of core**



Images from Dr. Nordhaus, Prof Burrows, Prof. Lamb, Dr. Chen

Flash Center IO Nightmare…

- Large 32,000 processor run on LLNL BG/L
- Parallel IO libraries not yet available
- Intensive I/O application
  - checkpoint files .7 TB, dumped every 4 hours, 200 dumps
    - used for restarting the run
    - full resolution snapshots of entire grid
  - plotfiles - 20GB each, 700 dumps
    - coarsened by a factor of two averaging
    - single precision
    - subset of grid variables
  - particle files 1400 particle files 470MB each
- 154 TB of disk capacity
- 74 million files!
- Unix tool problems
- Took 2 years to sift though data, sew files together

Hopper Filesystems

- Home directories (GPFS)
  - Intended for storing source code, and small files
  - Mounted across all NERSC systems
  - Small quota – 40 GB
  - Low performance
- 2 scratch parallel file systems (Lustre)
  - Intended for high performance, production runs
  - 35 GB/sec each
  - 1 PB disk each
  - Local to the Hopper system
- global scratch and project file systems (GPFS)
  - ~10 GB/sec
  - Mounted across all NERSC systems

Generic Parallel File System Architecture

Compute Nodes

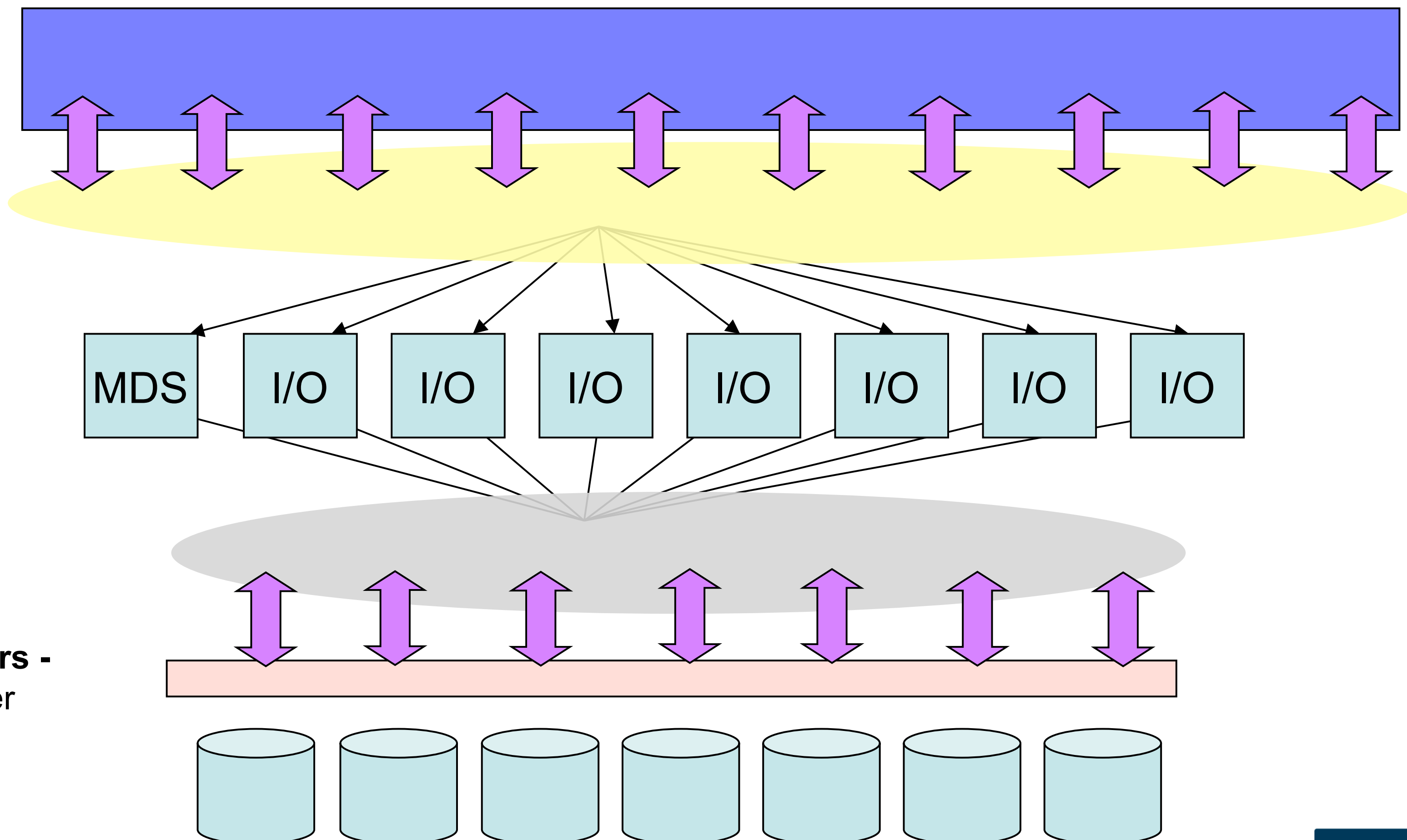
Internal Network

I/O Servers

External Network - (Likely FC)

Disk controllers - manage failover

Storage Hardware -- Disks

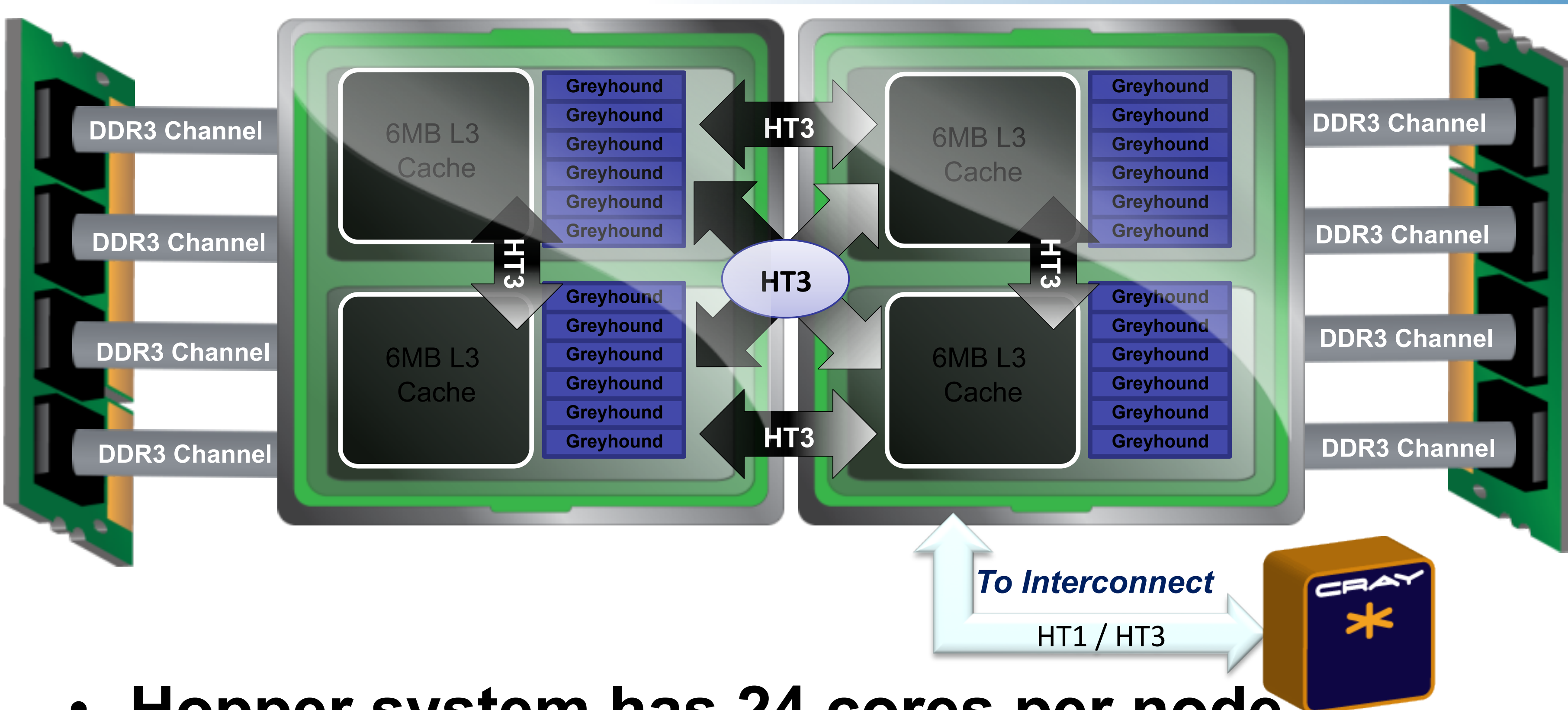


Don’t forget the Psychedelic Skins

Developing HPC Applications for Optimal Performance

40

What is Different About Hopper?



- Hopper system has 24 cores per node.
- The way that you use the new Hopper system may have to change as a result.

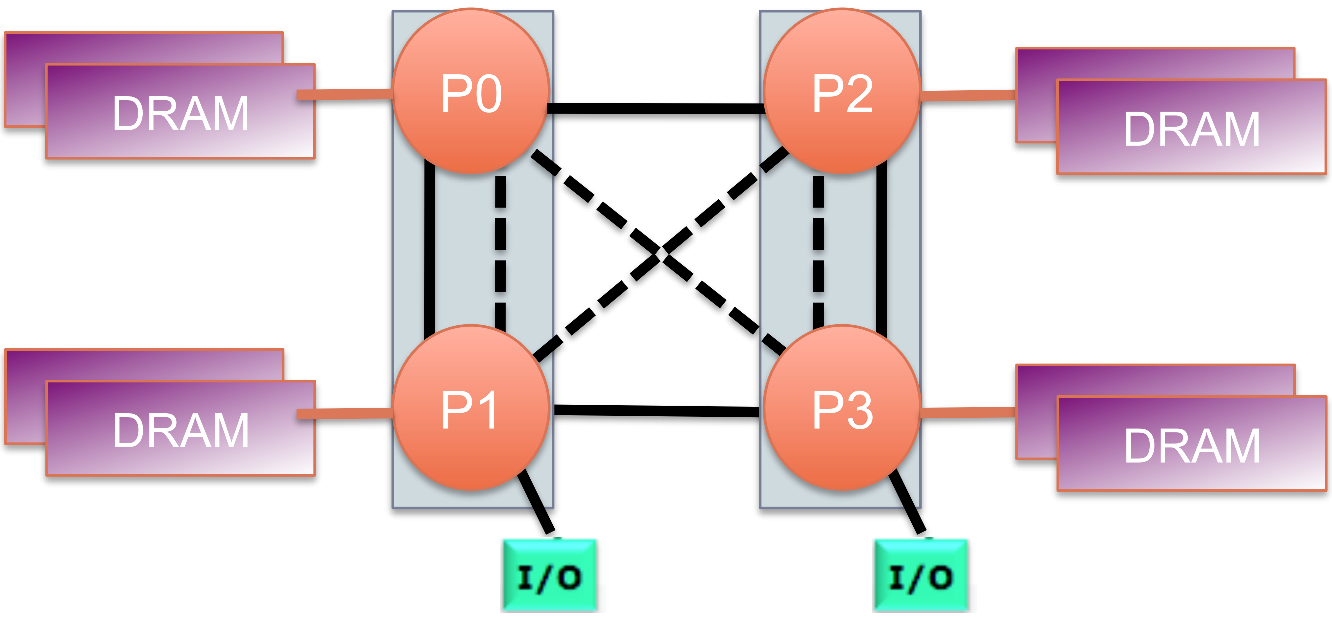
Hopper Node Topology
Understanding NUMA Effects

- **Heterogeneous Memory access between dies**
- **"First touch" assignment of pages to memory.**

2xDDR1333 channel
21.328 GB/s

3.2GHz x8 lane HT
6.4 GB/s bidirectional

3.2GHz x16 lane HT
12.8 GB/s bidirectional



- **Locality is key** *(just as per Exascale Report)*
- **Only *indirect* locality control with OpenMP**

U.S. DEPARTMENT OF ENERGY Office of Science
BERKELEY LAB Lawrence Berkeley National Laboratory

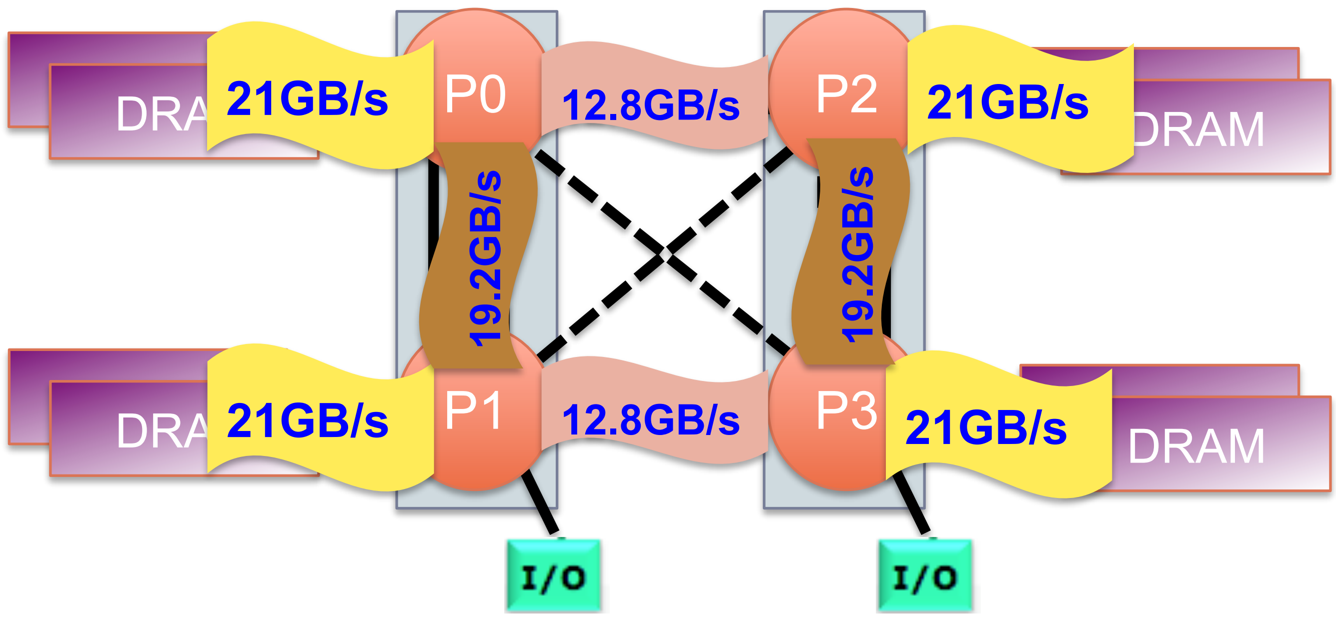
Hopper Node Topology
Understanding NUMA Effects

- **Heterogeneous Memory access between dies**
- **"First touch" assignment of pages to memory.**

2xDDR1333 channel
21.328 GB/s

3.2GHz x8 lane HT
6.4 GB/s bidirectional

3.2GHz x16 lane HT
12.8 GB/s bidirectional



- **Locality is key** *(just as per Exascale Report)*
- **Only *indirect* locality control with OpenMP**

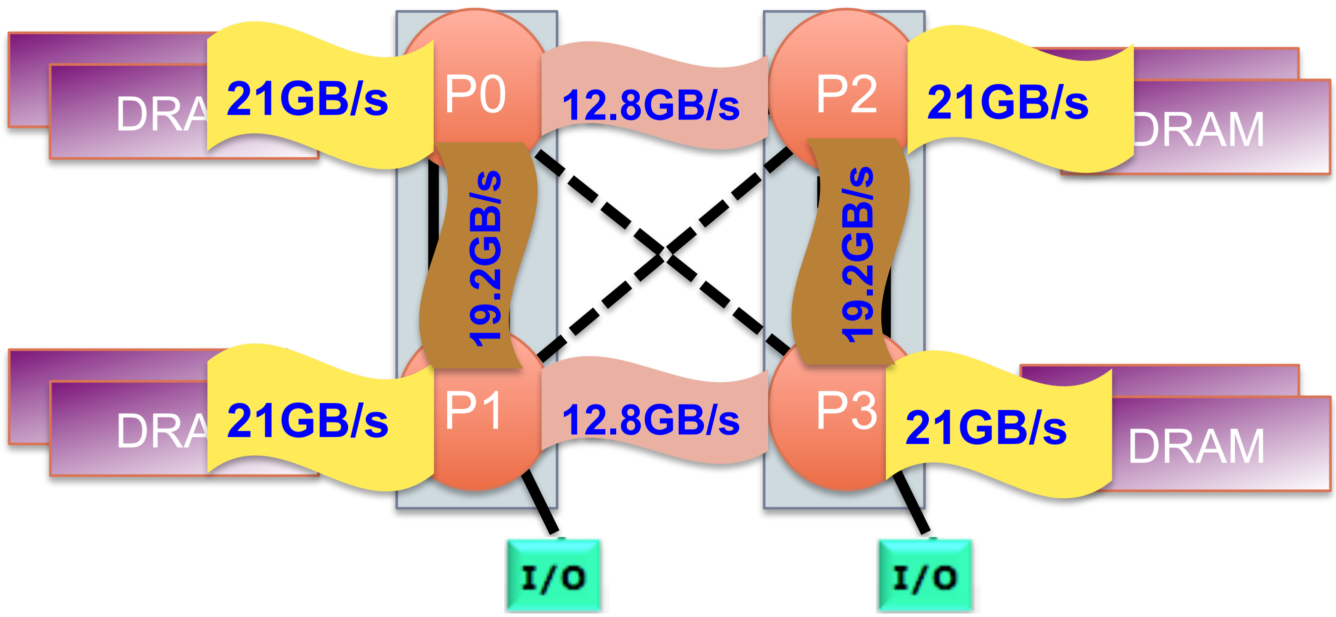
Hopper Node Topology
Understanding NUMA Effects

- **Heterogeneous Memory access between dies**
- **"First touch" assignment of pages to memory.**

2xDDR1333 channel
21.328 GB/s

3.2GHz x8 lane HT
6.4 GB/s bidirectional

3.2GHz x16 lane HT
12.8 GB/s bidirectional



- **Locality is key** *(just as per Exascale Report)*
- **Only *indirect* locality control with OpenMP**

What Else is Different ?

- **Less memory per core: 1.33 GB <u>*vs.*</u> 2.0 GB**
  - **8 GB per node (Franklin);**
  - **32 GB per node (Hopper, 6,008 nodes)**
- **“OOM killer terminated this process” error**

  OOM = Out of Memory
- **(Hopper has 384 larger-memory nodes 64 GB.)**

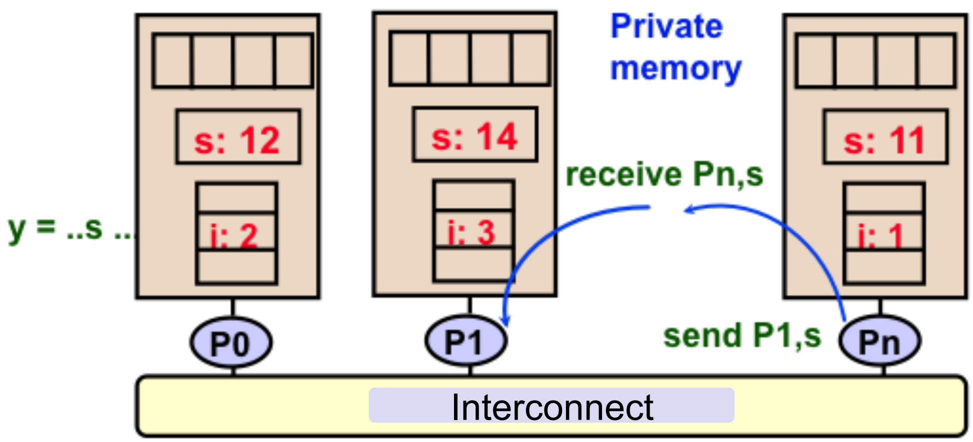
Will My Existing Pure MPI Code Run?

- **Probably, yes, your MPI code will run.** 
- **But the decrease in memory available per core may cause problems ...**
  - **May not be able to run the same problems.** 
  - **May be difficult to continue "weak" scaling (problem size grows in proportion to machine size).**
- **(and your MPI code might not use the machine most effectively.)**
- **Time to consider alternative programming models?**

What Does NERSC Recommend?

- **NERSC recognizes the huge investment in MPI.**
- **But given the technology trends…**
- **We suggest a move towards programming models other than pure MPI**
- **A good place to start: MPI + OpenMP ("Hybrid")**
  - **MPI for domain decomposition and OpenMP threads within a domain**
  - **Suggested primarily to help with memory capacity**

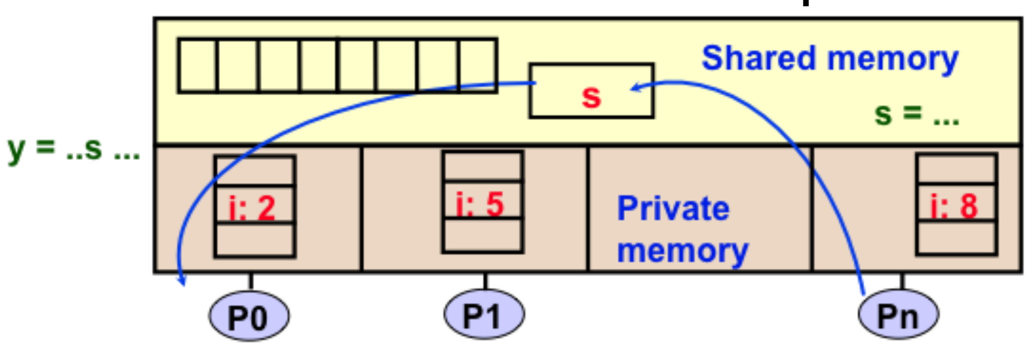
What are the Basic Differences Between MPI and OpenMP?



Message Passing Model

- Program is a collection of processes.
  - **Usually fixed at startup time**
- Single thread of control plus private address space -- NO shared data.
- Processes communicate by explicit send/ receive pairs
  - **Coordination is implicit in every communication event.**
- MPI is most important example.

Shared Address Space Model

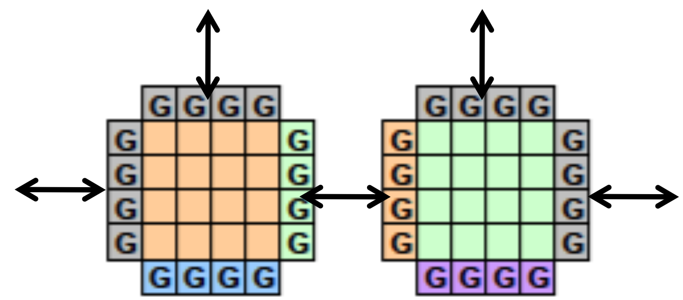


- Program is a collection of threads.
  - **Can be created dynamically.**
- Threads have private variables and shared variables
- Threads communicate implicitly by writing and reading shared variables.
  - **Threads coordinate by synchronizing on shared variables**
- OpenMP is an example

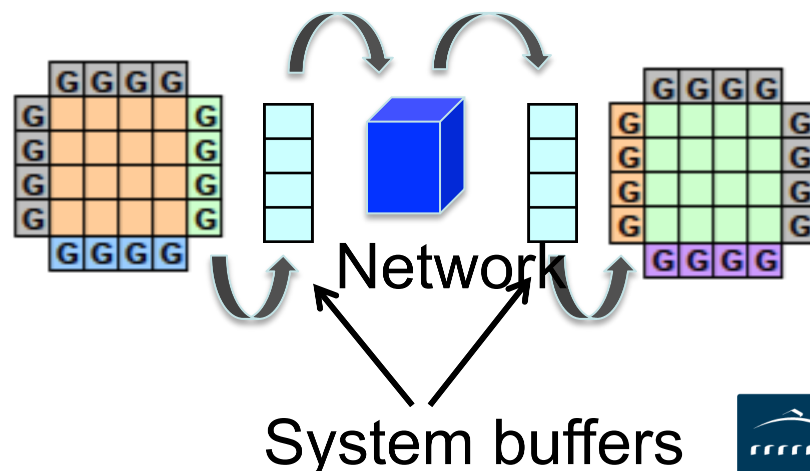
K.Yelick, CS267 UCB

Why are MPI-only Applications Memory Inefficient?

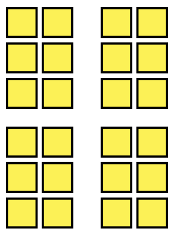
- **MPI codes consist of *n* copies of the program**
- **MPI codes require *application-level* memory for messages**
  - **Often called "ghost" cells**

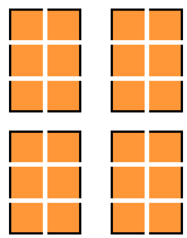


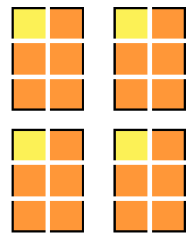
- **MPI codes require *system-level* memory for messages**
  - **Assuming the very common synchronous/blocking style**



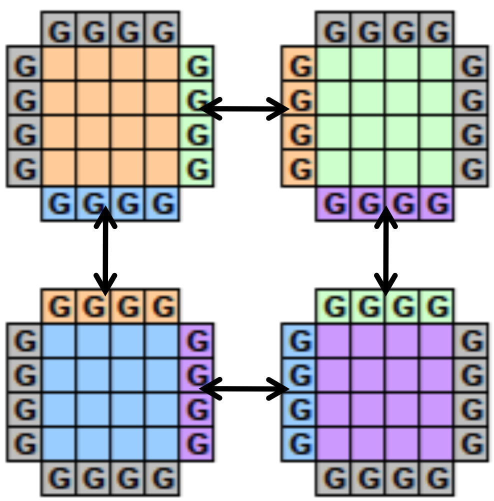
Why Does Hybrid/OpenMP Help?

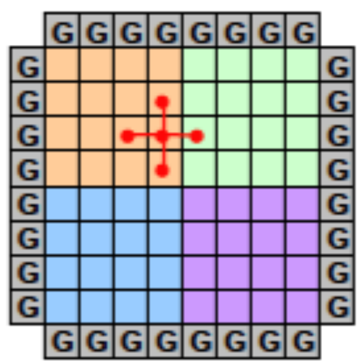

"Pure" MPI


"Pure" OpenMP


Hybrid: 4 MPI tasks, 6 threads per MPI

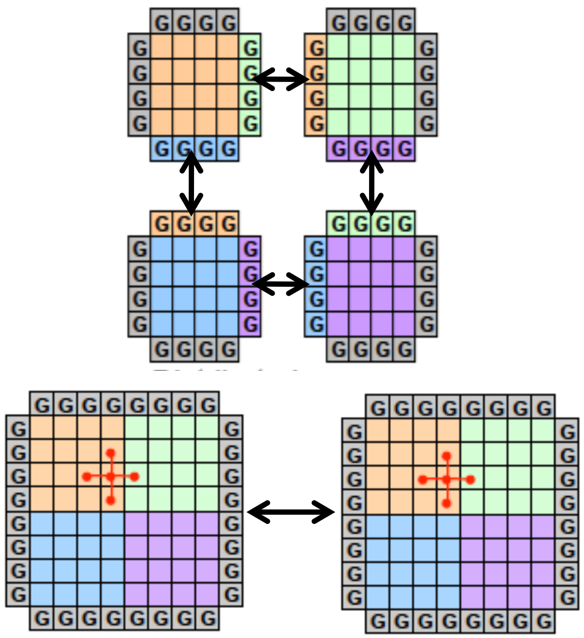
- **Reduced Memory Usage:**
  - **Fewer instances of your program on the node**
  - **Eliminate some ghost cell memory**

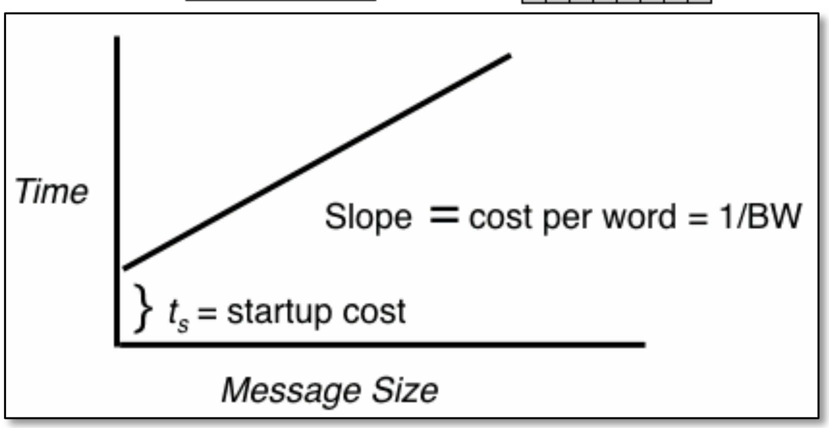

Distributed memory subgrid distribution


Shared memory subgrid distribution

Figures from Kaushik Datta, Ph.D. Dissertation, UC Berkeley, 2009

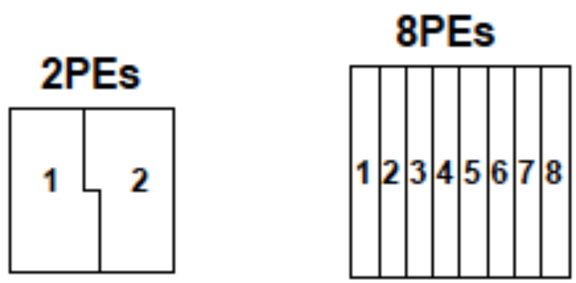
Why Does Hybrid/OpenMP Help?





- Send larger MPI messages
  - small messages are expensive
- No intra-node messages

Why Does Hybrid/OpenMP Help?



- There may be scalability limits to domain decomposition
- OpenMP adds fine granularity (larger message sizes) and allows flexibility of dynamic load balancing.
- Some problems have two levels of parallelism

What are the Benefits of OpenMP?

- Uses less memory per node
- Typically, at least equal performance
- Additional parallelization may fit algorithm well
  - especially for applications with limited domain parallelism
- Possible improved MPI performance and load balancing
  - Avoid MPI within node
- OpenMP is a standard so code is portable
- Some OpenMP code can be added incrementally
  - Can focus on performance-critical portions of code
- Better mapping to multicore architecture

What are the Disadvantages of OpenMP?

- **Additional programming complexity**
- **Can be difficult to debug race conditions**
- **Requires explicit synchronization**
- **Additional scalability bottlenecks:**
  - **thread creation overhead, critical sections, serial sections for MPI**
- **Cache coherence problems (false sharing) and data placement issues**
  - **Memory locality is key...**
  - **but OpenMP offers no direct control**

Are There Additional Solutions?

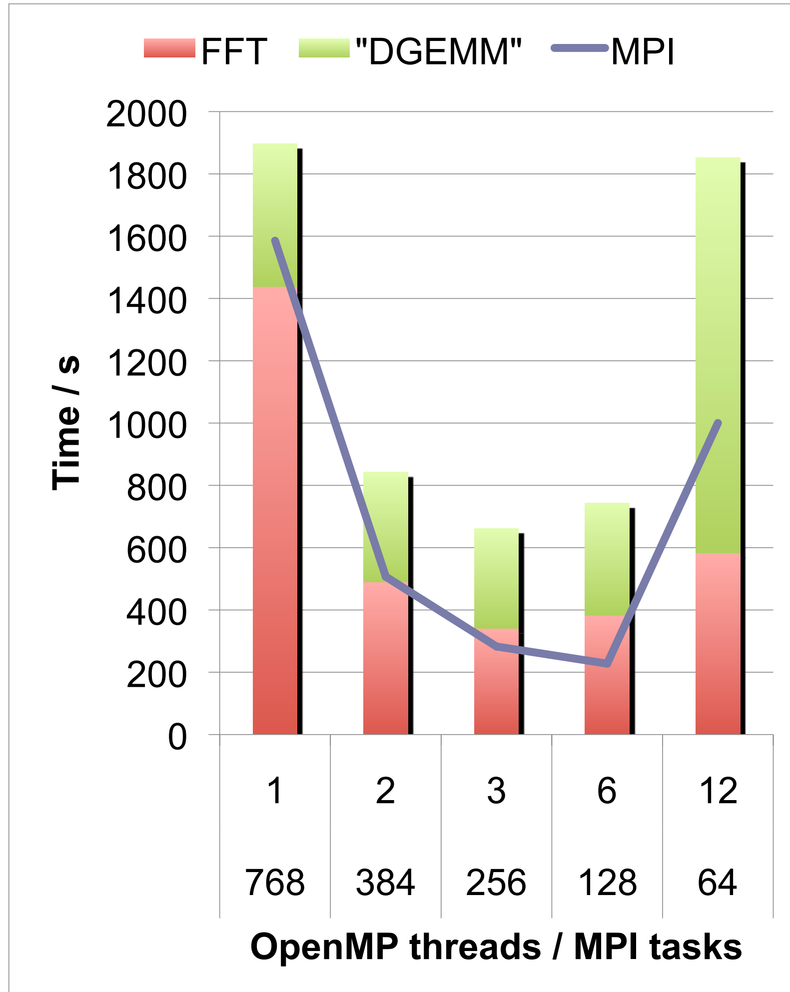
- Sometimes it may be better to leave cores idle
  - Improves memory capacity and bandwidth
  - Improves network bandwidth

- However, you are charged for all cores

Advice to NERSC Users

- **OpenMP + MPI can be faster than pure MPI and is often comparable in performance**
- **Mixed OpenMP/MPI saves significant memory**
- **Beware of NUMA ! – don't use more than 6 OpenMP threads unless you know how to first-touch memory perfectly.**

Paratec MPI+OpenMP Performance



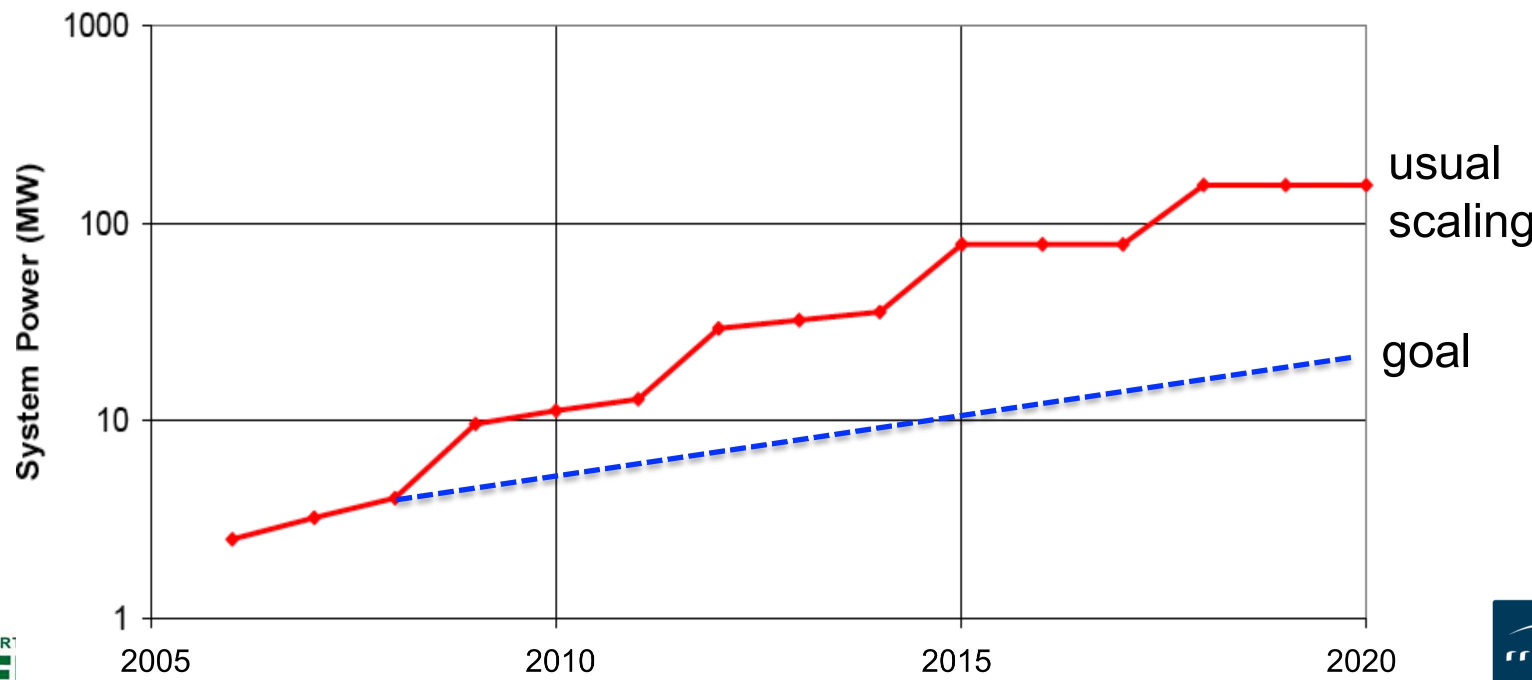
Challenges and Future Trends

57

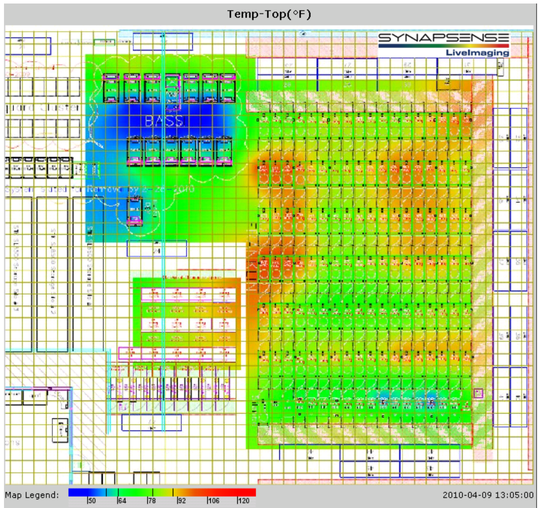
Energy Efficiency is Necessary for Computing

- Systems have gotten about 1000x faster over each 10 year period
- 1 petaflop (10^{15} ops) in 2010 will require 3MW

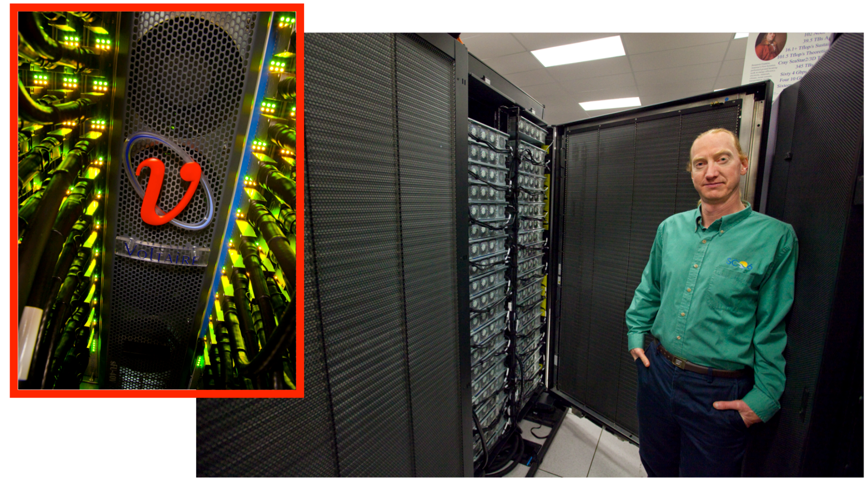
    → 3 GW for 1 Exaflop (10^{18} ops/sec)
- DARPA committee suggested 200 MW with “usual” scaling
- Target for DOE is 20 MW in 2018



Energy Efficiency Partnerships with Synapsense and IBM



600 Sensors for temperature, etc.



Rear door heat exchangers

- **Monitoring for energy efficiency (and reliability!)**
- **Liquid cooling on IBM system uses return water from another system, with modified CDU design**
  - **Reduces cooling costs to as much as ½**
  - **Reduces floor space requirements by 30%**

Air is colder coming out than going in!

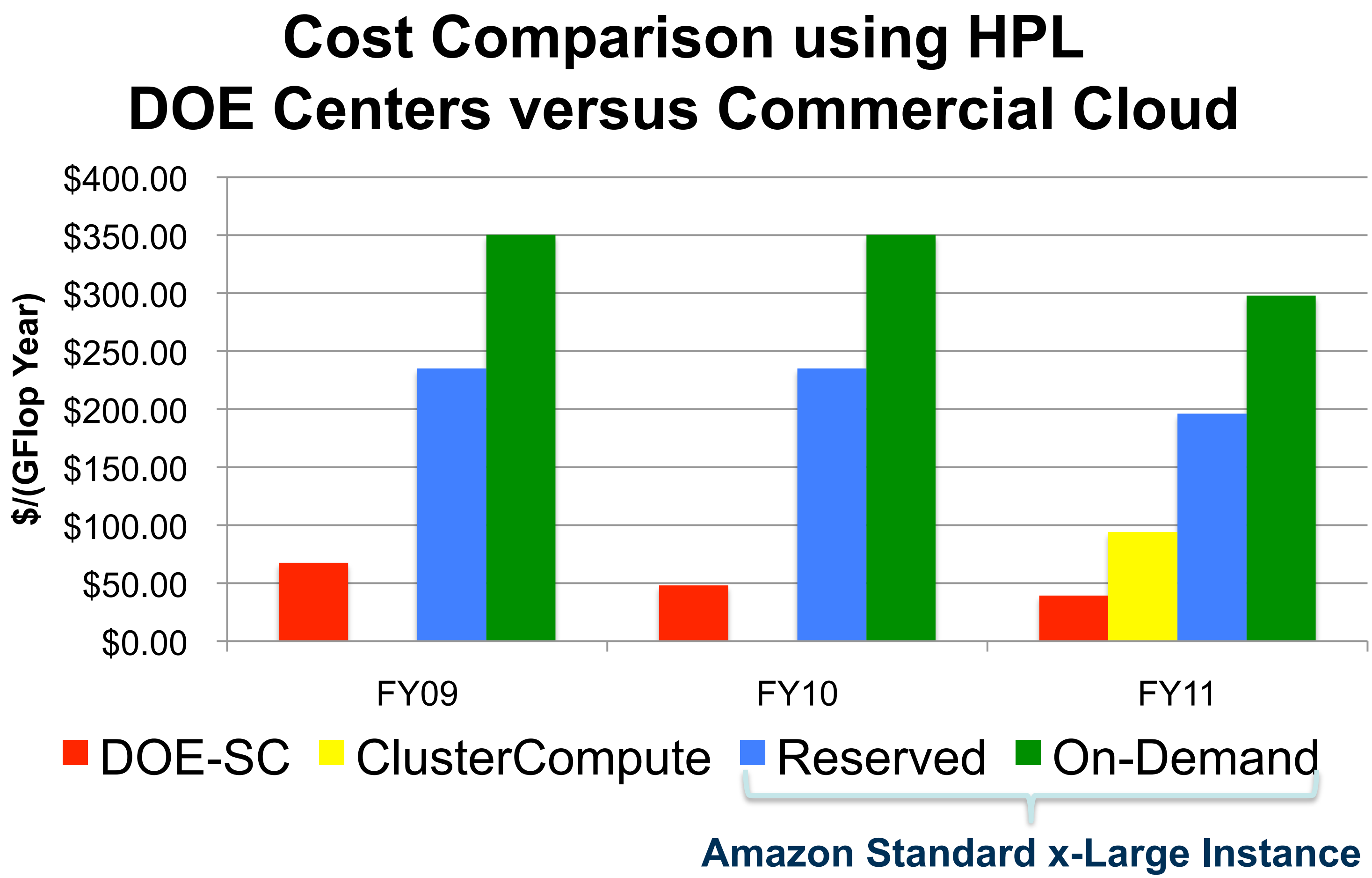
DOE Explores Cloud Computing

- In spite of NERSC and other DOE centers
  - Many scientists still by their own clusters
  - No coordinated plan for clusters in SC
- NERSC received funding for Magellan
  - $16M project at NERSC from Recovery Act
- Cloud questions to explore on Magellan:
  - Can a cloud serve DOE’s mid-range computing needs?
  - What features (hardware and software) are needed of a “Science Cloud”?
  - What requirements do the jobs have?
  - How does this differ, if at all, from commercial clouds which serve primarily independent serial jobs?
- Magellan testbed installed in early 2010

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

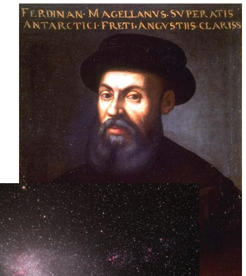
HPC Centers Are Cheaper than Clouds



Gap is higher for mid-range applications; grows with job size

NERSC

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**

Recent Cover Stories from NERSC Research

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

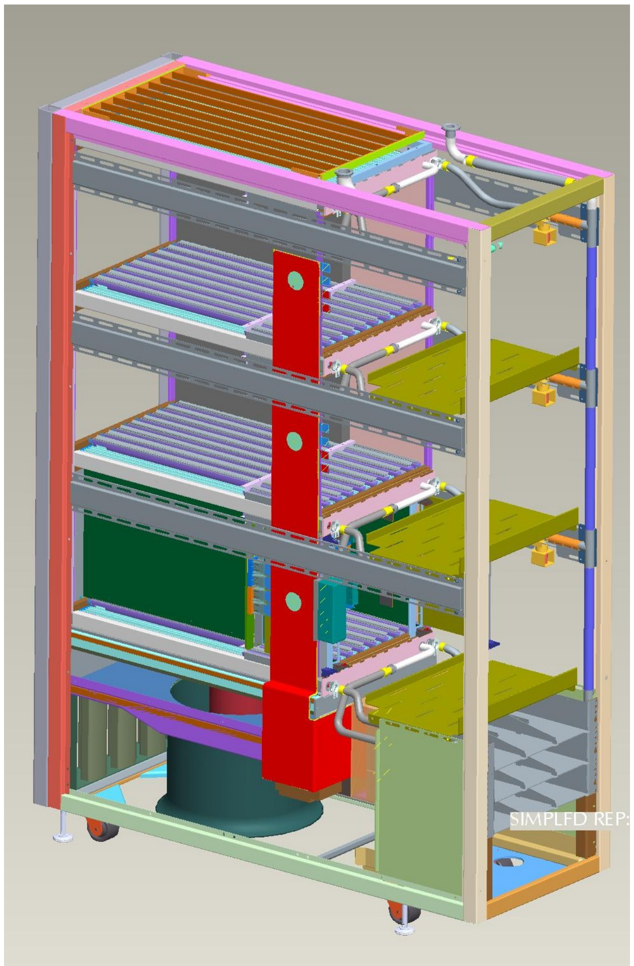
Extra Slides

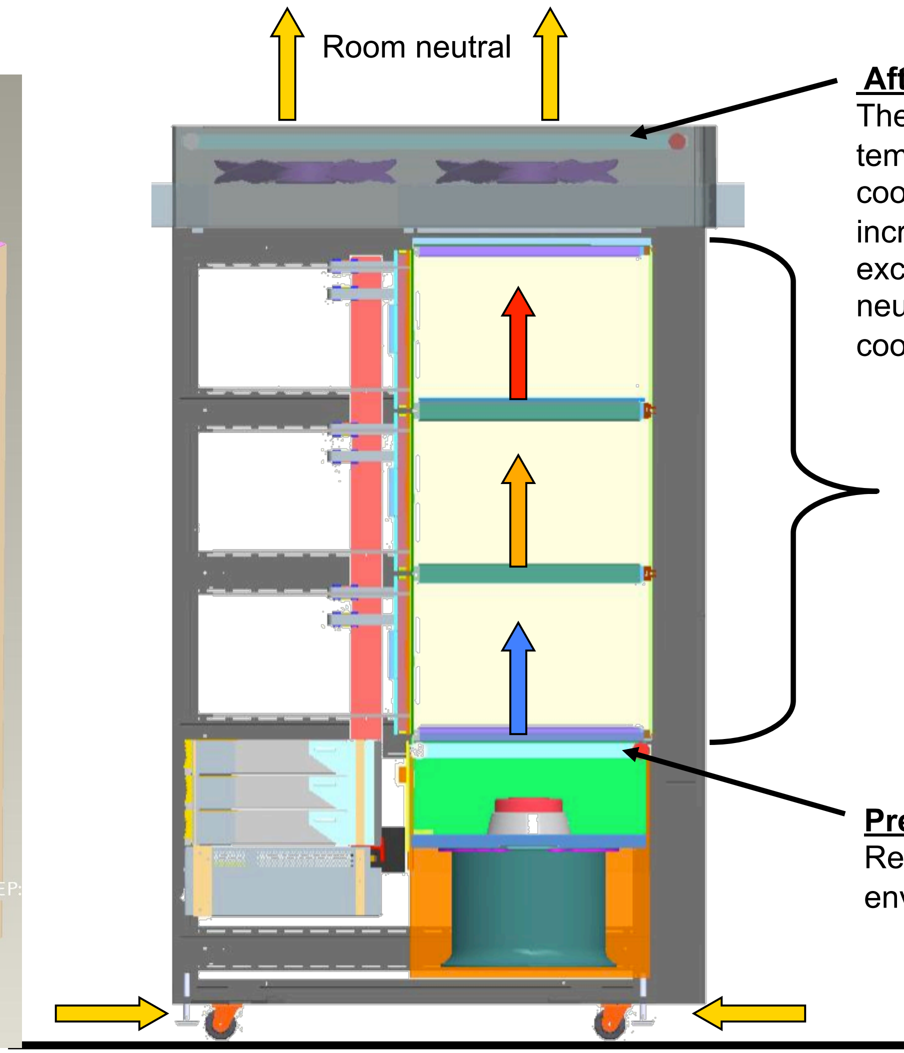
XE6 Cabinet Design

What's up with the hat??

Cray Cabinet Design

Energy Efficient Liquid Cooling





After-cooler assembly:
The extremely hot exhaust temperature of the HD air cooled chassis dramatically increases the capability of heat exchanger. This makes room neutral possible with single cooler assembly at exit.

HD Air Cooled Chassis:
Sandwich with R134a evaporators.

Pre-cooler assembly:
Required to operate in room environments over 20C.

Images Courtesy of Cray Inc.

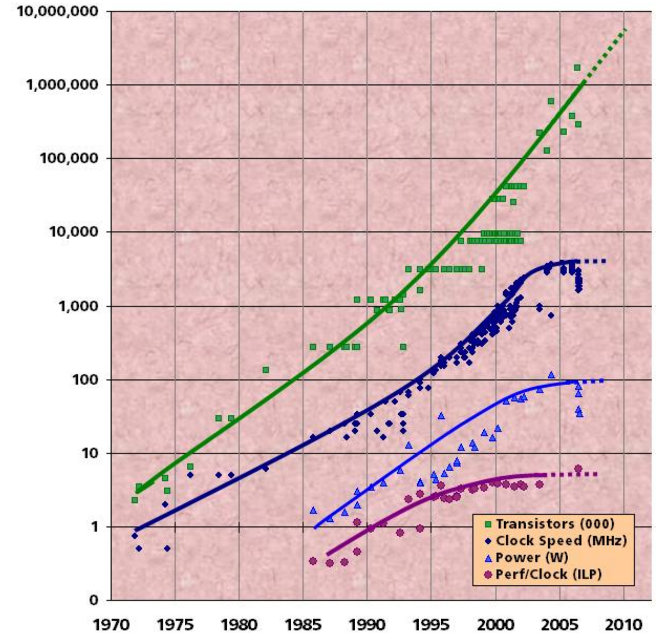
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB
Lawrence Berkeley National Laboratory

Hopper Cooling Apparatus

What About the Future?

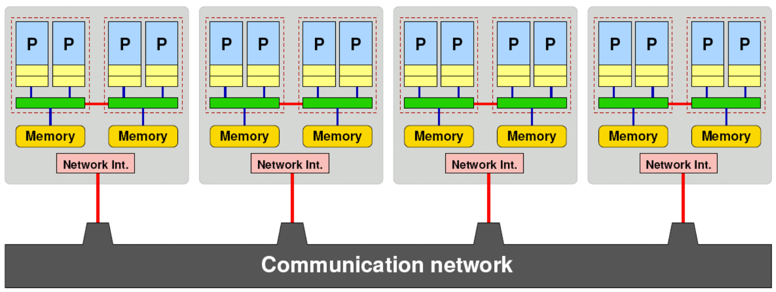
- **The technology trends point to**
  - **Little or no gain in clock speed or performance per core;**
  - **Rapidly increasing numbers of cores per node;**
  - **Decreased memory *capacity* per core (possible slight increase per node)**
  - **Decreased memory *bandwidth* per core**
  - **Decreased interconnect bandwidth per core**
  - **Deeper memory hierarchy**



- **Hopper is the first example at NERSC but surely not the last**

Isn't This the Same as Clusters of SMPs (.ca 2002)?

- **SMP: Symmetric Multiprocessor**
  - **aka clusters, Networks of Workstations, CLUMPS, ...**
  - **SGI Origin, ASCI Q/Blue Mountain, Berkeley NOW, IBM SP, ...**



- **In some ways the issues are the same:**
  - **Memory architecture is the key**
- **But chip multiprocessors have vastly improved inter-core latencies and bandwidth.**
- **<u>With today's trends we have no choice.</u>**

ASCR's Computing Facilities

NERSC at LBNL

- Thousands of users, hundreds 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 competitively

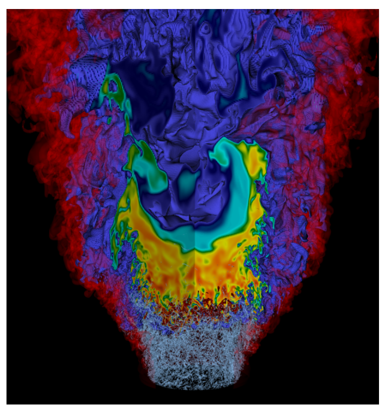
LCFs at ORNL and ANL

- Hundreds of users, tens of projects
- Allocations:
  - 60% ANL/ORNL managed INCITE process
  - 30% ACSR Leadership Computing Challenge*
  - 10% LCF reserve
- Science limited to largest scale; not just DOE/SC
- Machines procured through partnerships

Computation and Experiments at Berkeley Lab Improve Efficiency of Burners

- **Low Swirl Burners used by Solar Turbines (Caterpillar) and Maxon Corp. (Honeywell) to improve commercial burners**
  - **Efficient, low-emissions, Fuel-flexible (oil, gas, hydrogen-rich fuels)**
- **Simulations explain combustion process to improve designs**
  - **Modeled kinetics and chemical transport (15 species, 58 reactions)**
  - **Uses advanced math algorithms (AMR) equivalent to $4K^3$ mesh**
  - **Scales and runs in production at 20K cores**

Simulations show cellular burning in lean hydrogen leads to pockets of enhanced emissions, & increasing the turbulence enhances the effect.



Simulations reveal features not visible in lab (John Bell, PI, LBNL)

Experiments show feasibility: 50KW-50MW (Robert Cheng, PI, LBNL)

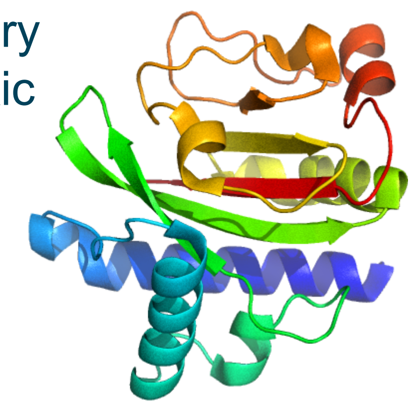
Low NOx technology licensed by industry

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

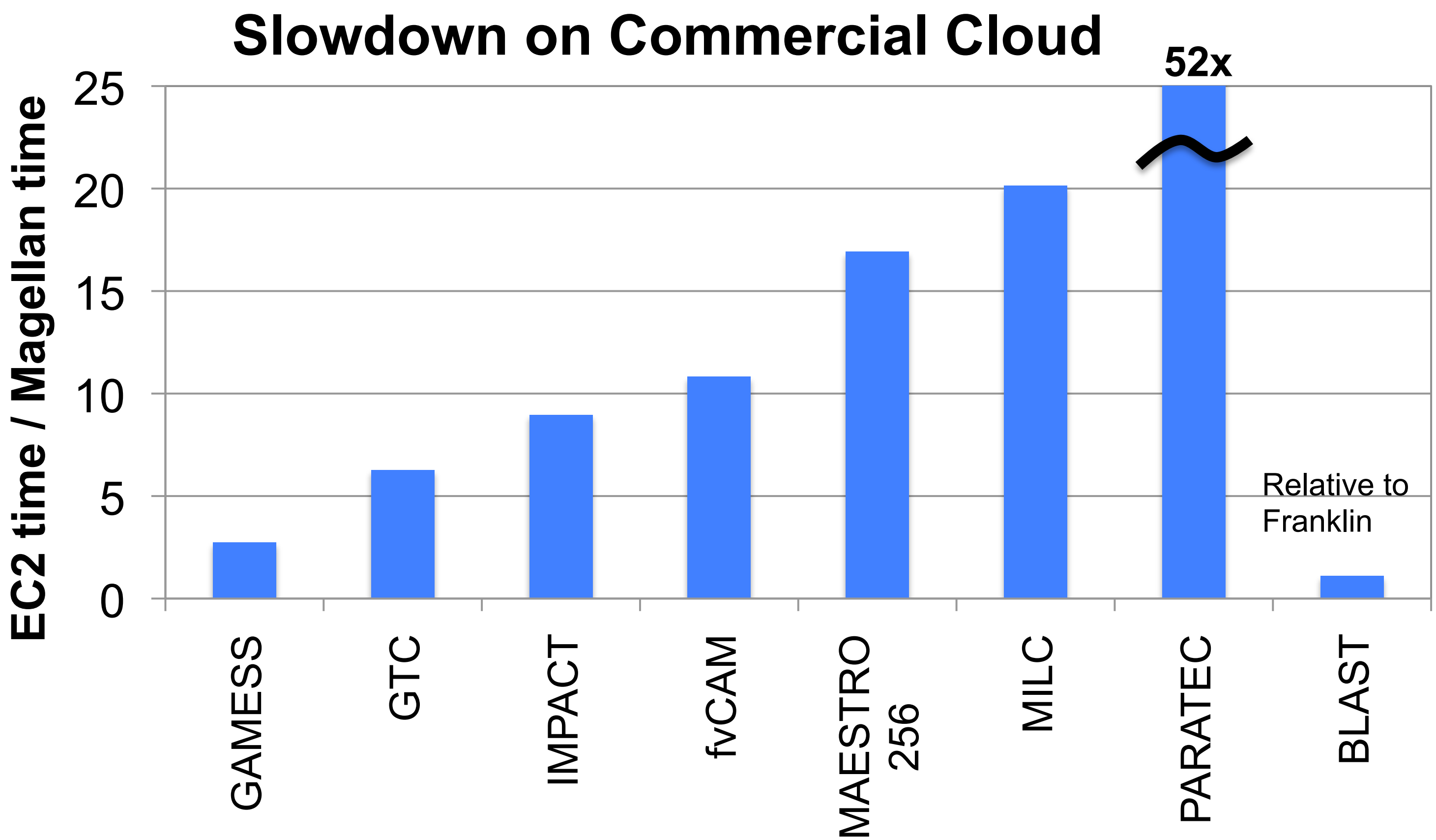
Simulations Populate a Database of Molecular Dynamics and Protein Folds

- **Produced public catalog of the unfolding dynamics of 11,000 proteins, covering all 807 self-contained autonomous folds**
- **Simulations used 12M hours of NERSC on custom code and help from NERSC on load balancing, optimizations, and workflow**
- **Mined amyloid producing proteins and found common structural feature between normal and toxic forms.**
  - Custom-designed complementary compounds, which bind with toxic forms of proteins that cause multiple diseases, including Alzheimer’s and mad cow.
  - Results suggest drug designs, screening for blood/food supply, and diagnostic tools for up to 25 amyloid diseases.



Valerie Daggett, PI, U. Washington

Provide Cloud Computing Testbed and Evaluation



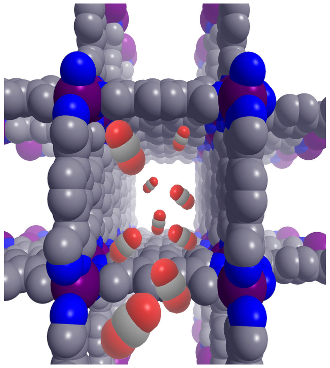
On traditional science workloads, standard cloud configurations see significant slowdown (up to 50x), but independent BLAST jobs run well

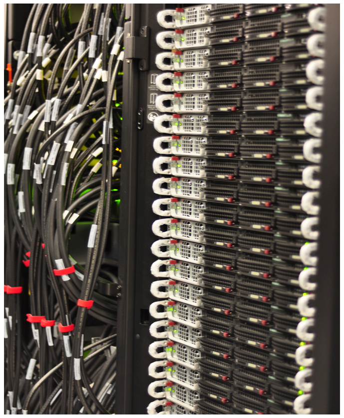
U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

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



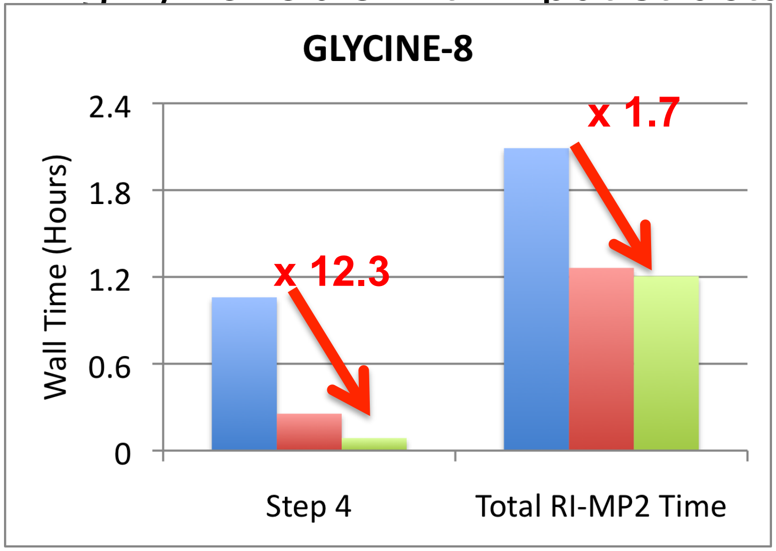


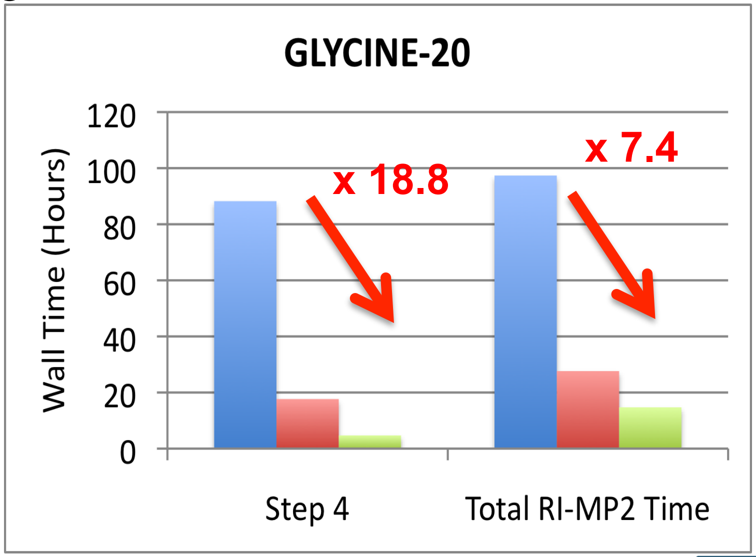
Fermi GPU Racks - NERSC

Blue: CPU 1 thread

Red: CPU 8 threads

Green: GPU





U.S. DEPARTMENT OF ENERGY | Office of Science

Jihan Kim (SciDAC-E NERSC Postdoc)

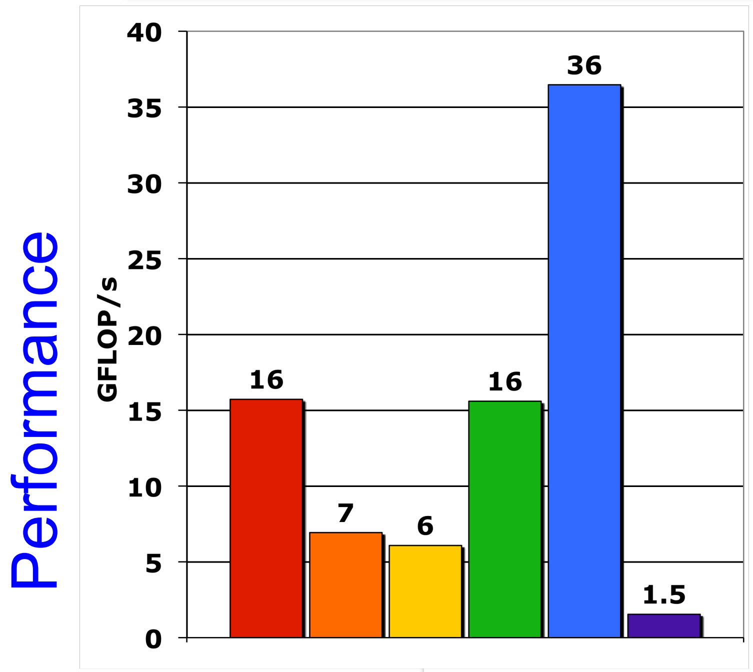
BERKELEY LAB

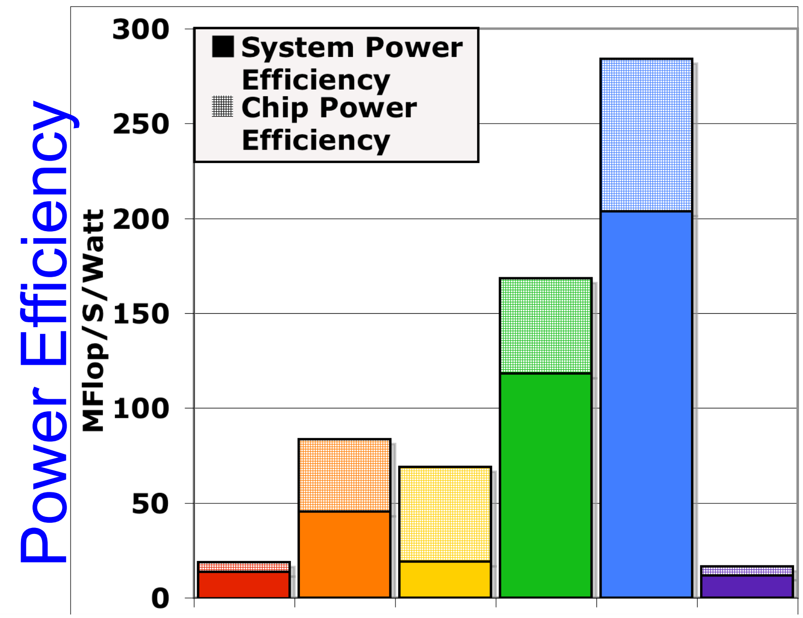
Don't Be Fooled by the Hype

(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)

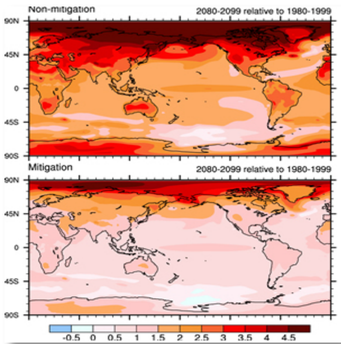




Cache-based
- Gainestown
- Barcelona
- Victoria Falls

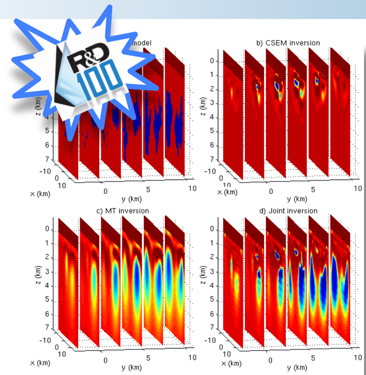
Local store-based
- Cell Blade
- GTX280
- GTX280-Host

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)

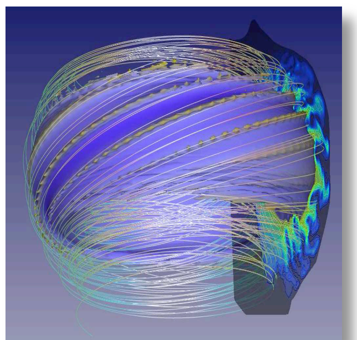


Energy Resources

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

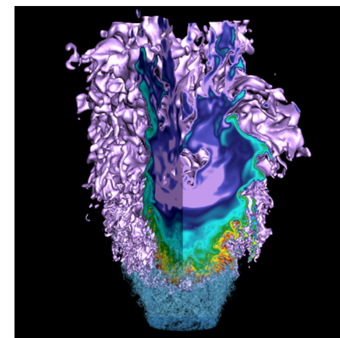
Fusion Energy

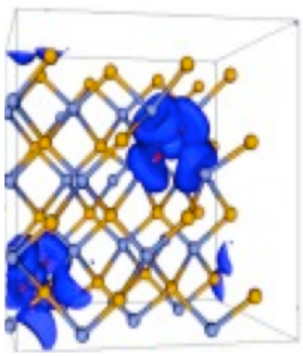
A new class of non-linear plasma instability has been discovered that may constrain design of the ITER device.
(Linda Sugiyama, MIT)



Combustion

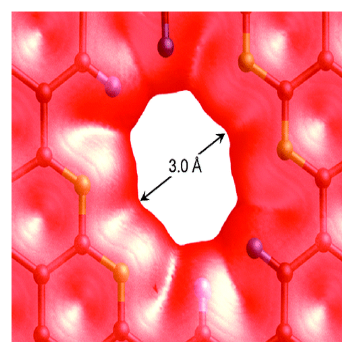
Adaptive Mesh Refinement allows simulation of a fuel-flexible low-swirl burner that is orders of magnitude larger & more detailed than traditional reacting flow simulations allow.
(John Bell, LBNL)





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)

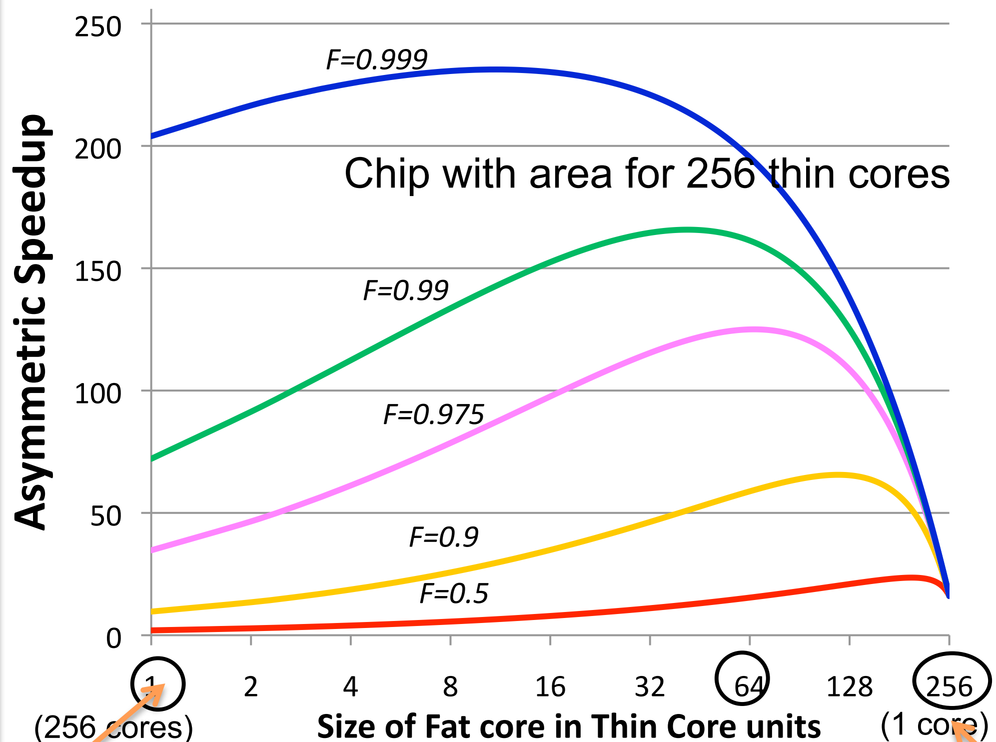
U.S. DEPARTMENT OF ENERGY | Office of Science

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

F is fraction of time in parallel; 1-F is serial



256 small cores

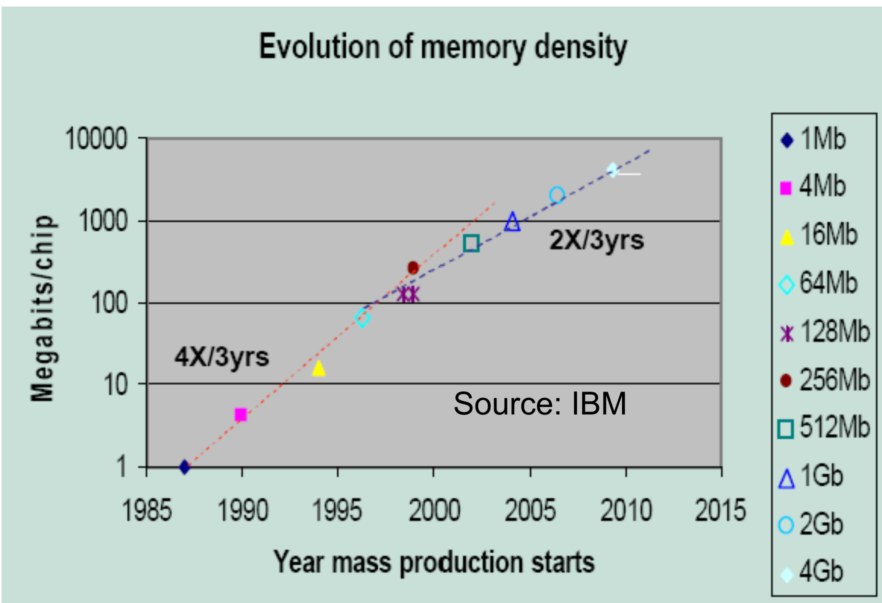
1 fat core

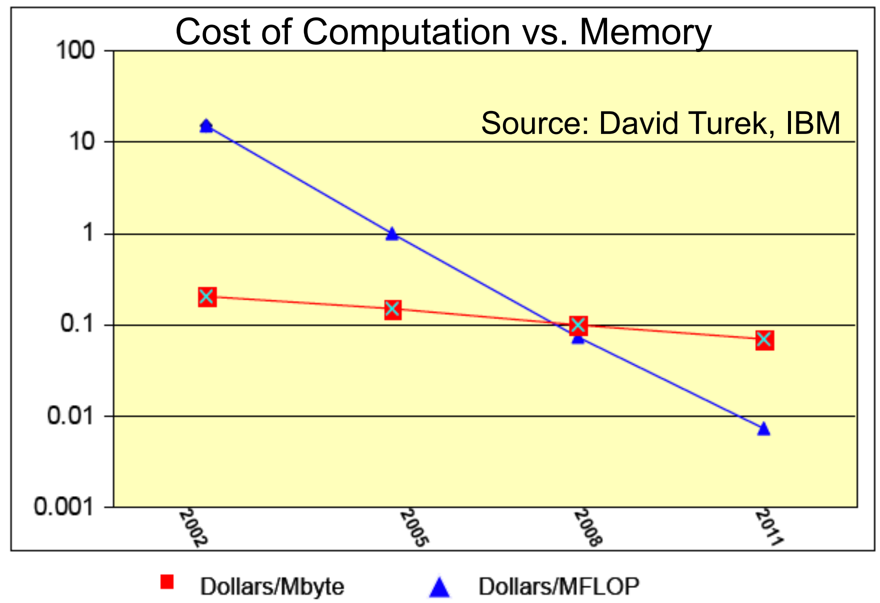
Ubiquitous programming model of today (MPI) will not work within a processor chip

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

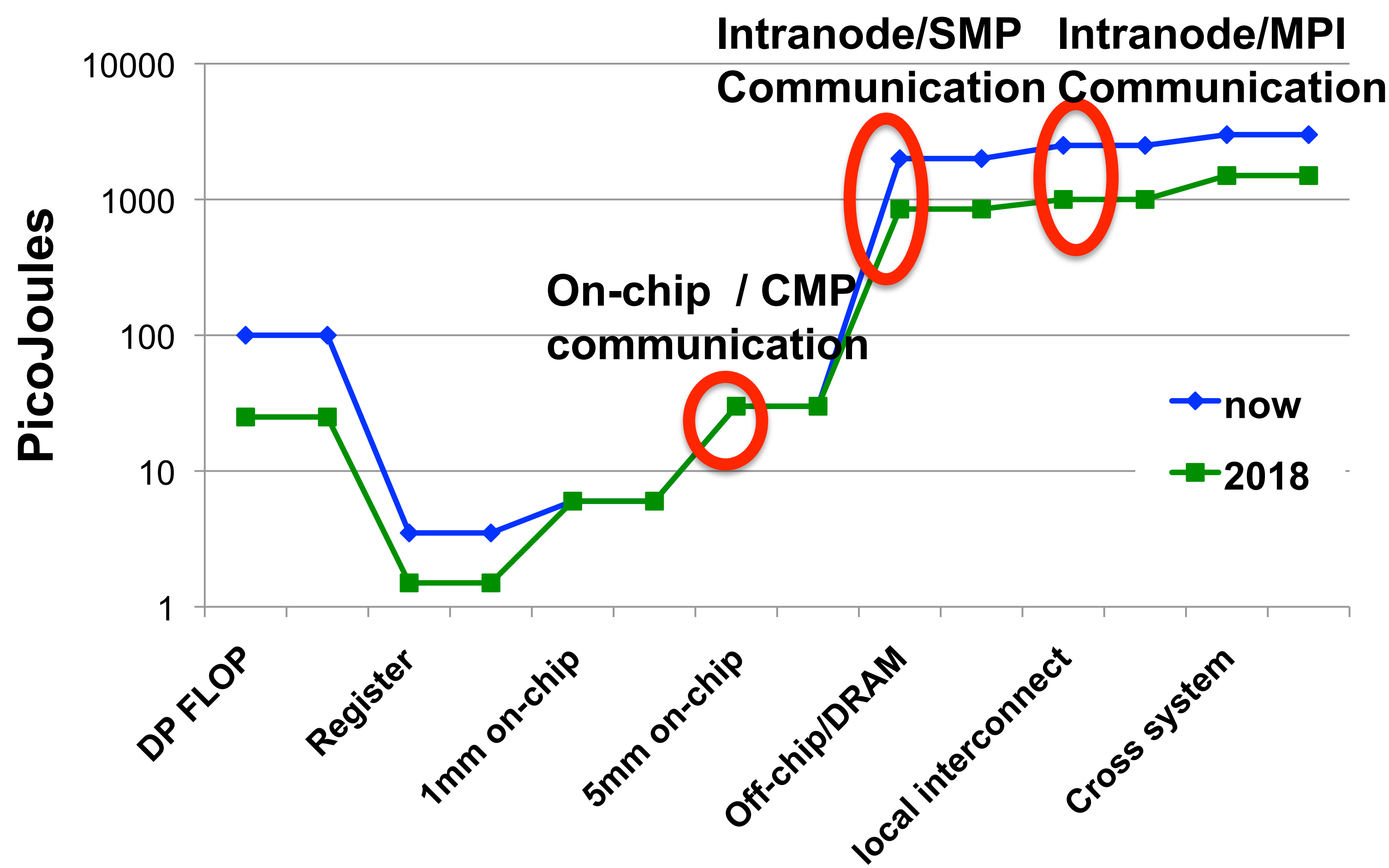




The cost to sense, collect, generate and calculate data is declining much faster than the cost to access, manage and store it

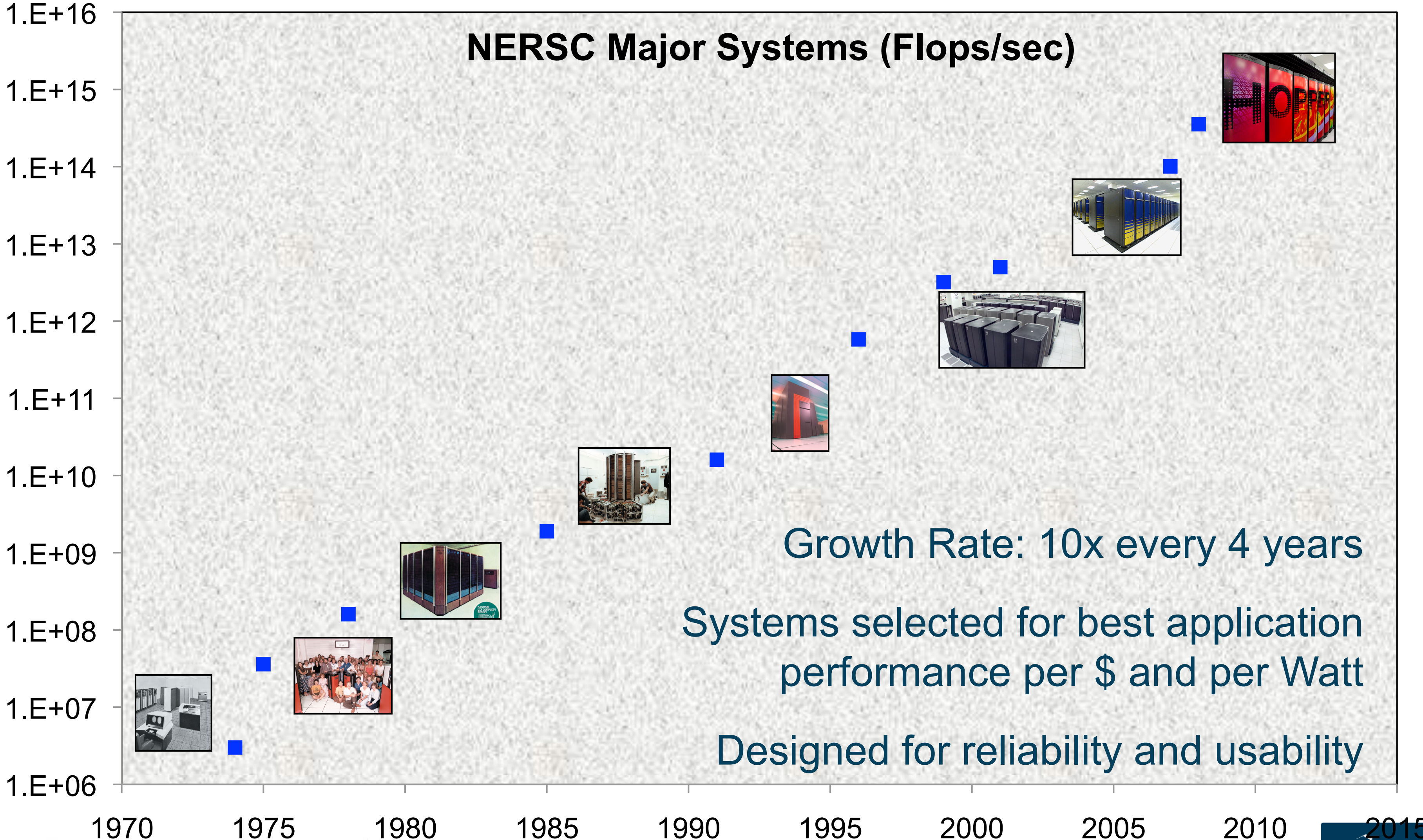
Question: *Can you double concurrency without doubling memory?*

Where does the Energy (and Time) Go?



Counting flops is irrelevant, only data movement matters

NERSC Responds to Scientific Demands for Computing and Services

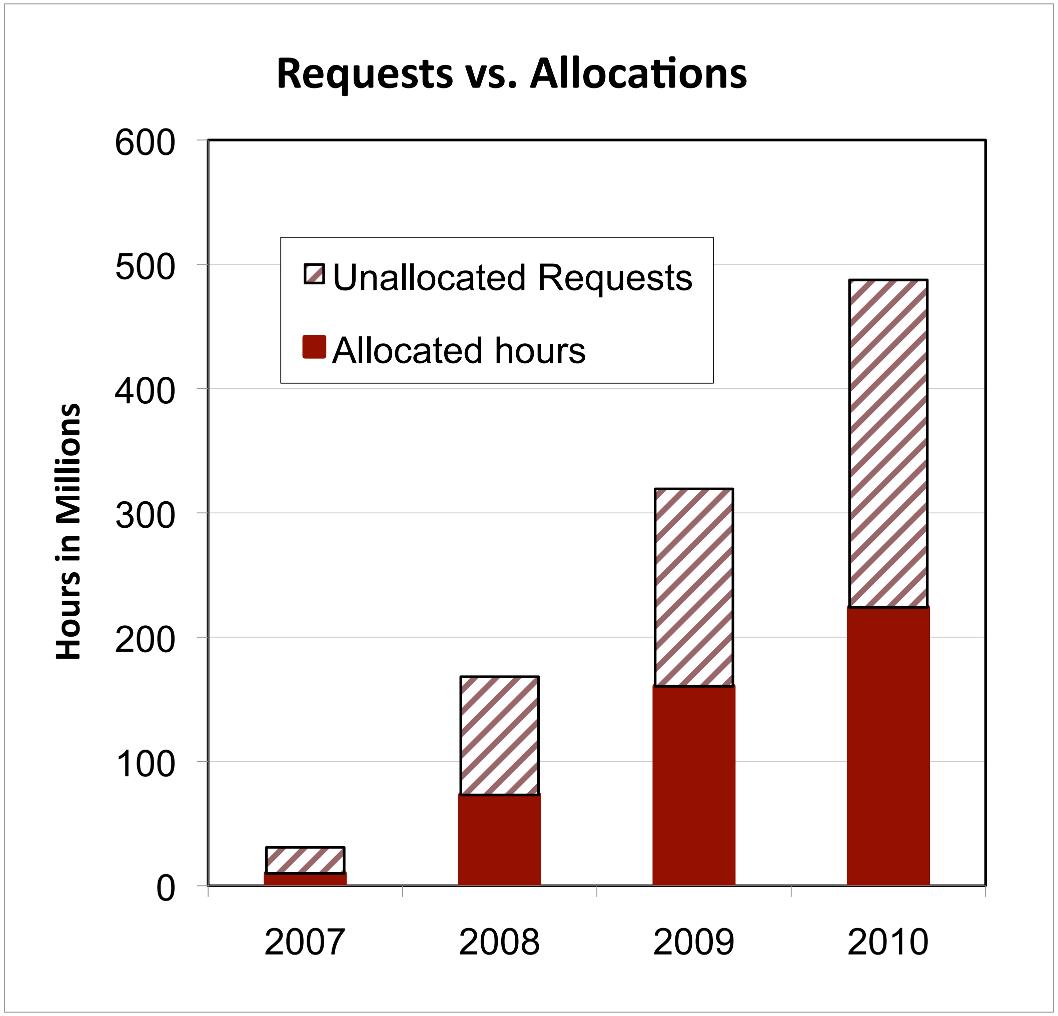


Challenges to ~~Exascale~~ Performance Growth

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

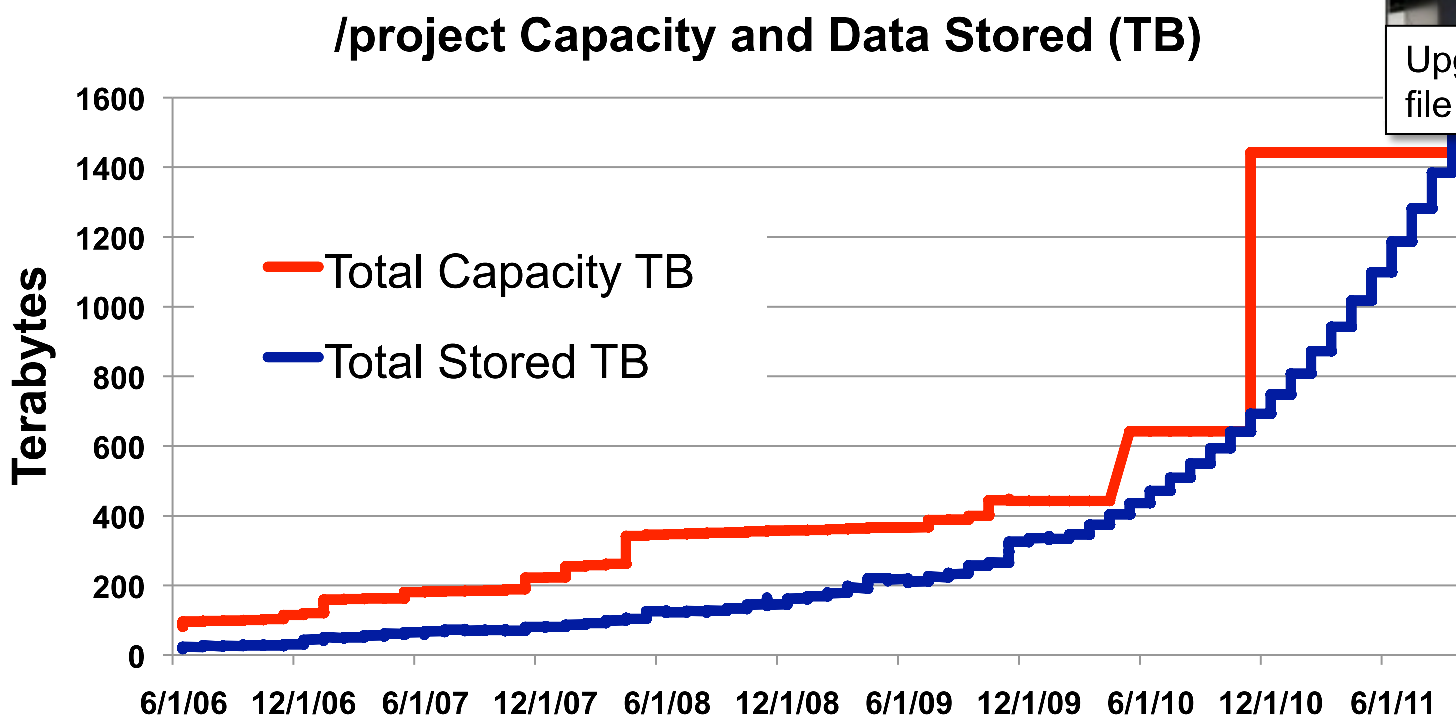
Demand for More Computing



- *Each year DOE users requests ~2x as many hours as can be allocated*
- *This 2x is artificially constrained by perceived availability*
- *Unfulfilled allocation requests amount to hundreds of millions of compute hours in 2010*

NERSC Global Filesystem Upgrades & Enhancements


Upgraded global file system



- Extended global filesystem from “project” to scratch and home directories for convenience
- Different service models for capacity (project), random access performance (home), temporary data (scratch)

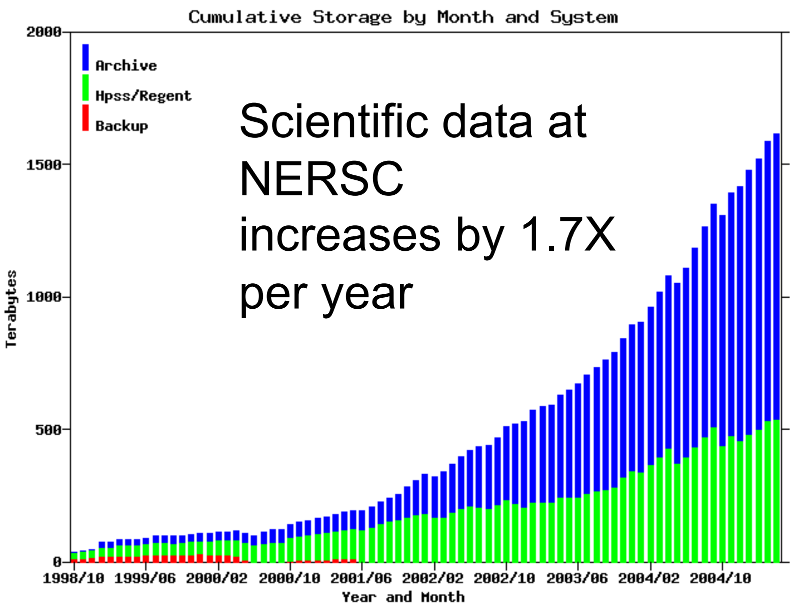
NERSC Strategy: *Science First*

- **Response to scientific needs**
  - Requirements setting activities
- **Support computational science:**
  - Provide effective machines that support fast algorithms
  - Deploy with flexible software
  - Help users with expert services
- **NERSC future priorities are driven by science:**
  - Increase application capability: “usable Exascale”
  - For simulation and data analysis

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

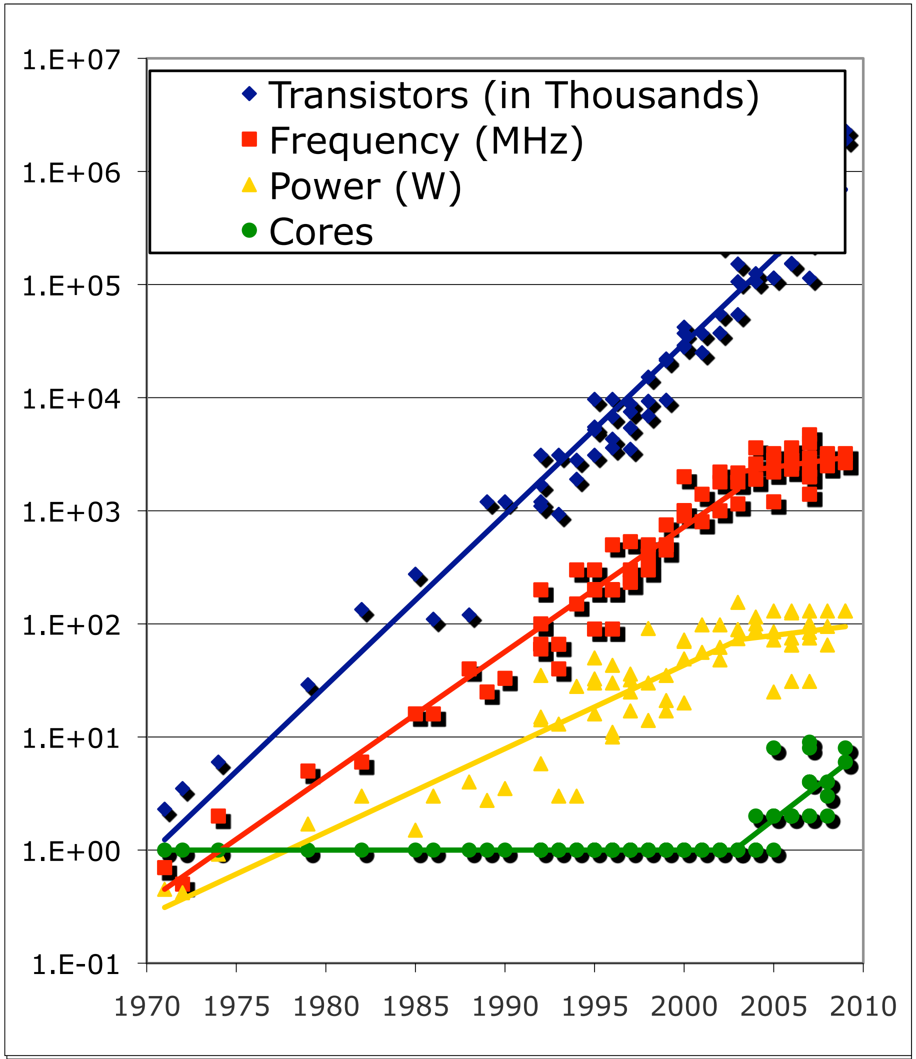
Tape Archives: Green Storage



- **Tape archives are important to efficient science**
  - **2-3 orders of magnitude less power than disk**
  - **Requires specialized staff and major capital investment**
  - **NERSC participates in development (HPSS consortium)**
- **Questions: What are your data sets sizes and growth rates?**

Moore's Law Continues, but Only with Added Concurrency

- **Power density limit single processor clock speeds**
- **Cores per chip is growing**
- **Simple doubling of cores is not enough to reach exascale**
  - **Also a problem in data centers, laptops, etc.**
- **Two paths to exascale:**
  - **Accelerators (GPUs)**
  - **Low power embedded cores**
  - **(Not x86 clusters)**

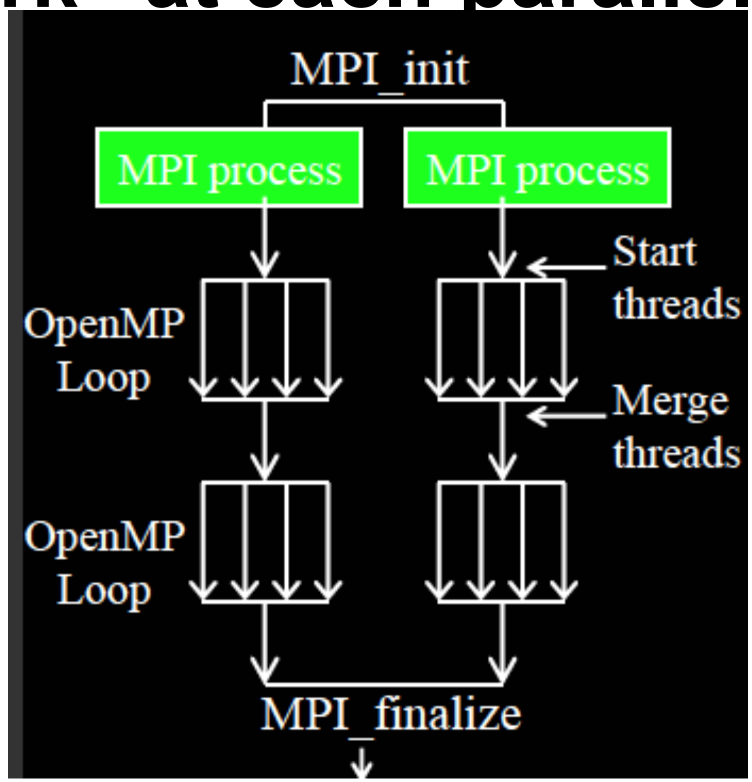


U.S. DEPARTMENT OF ENERGY | Office of Science

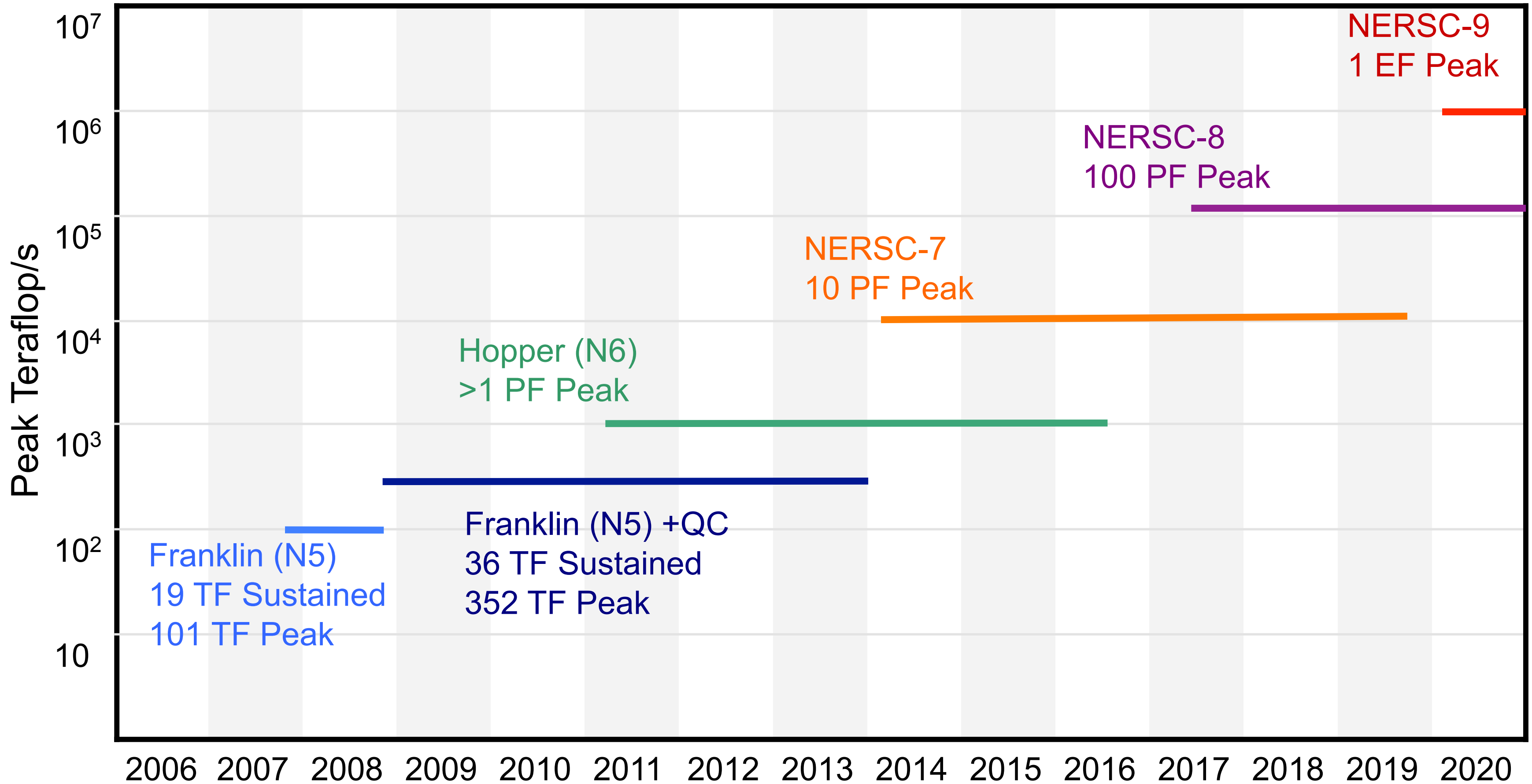
BERKELEY LAB

Typical OpenMP Program

- **Execution begins with a single "Master Thread"**
- **Threads "fork" at each parallel region, join at end**



NERSC Aggressive Roadmap



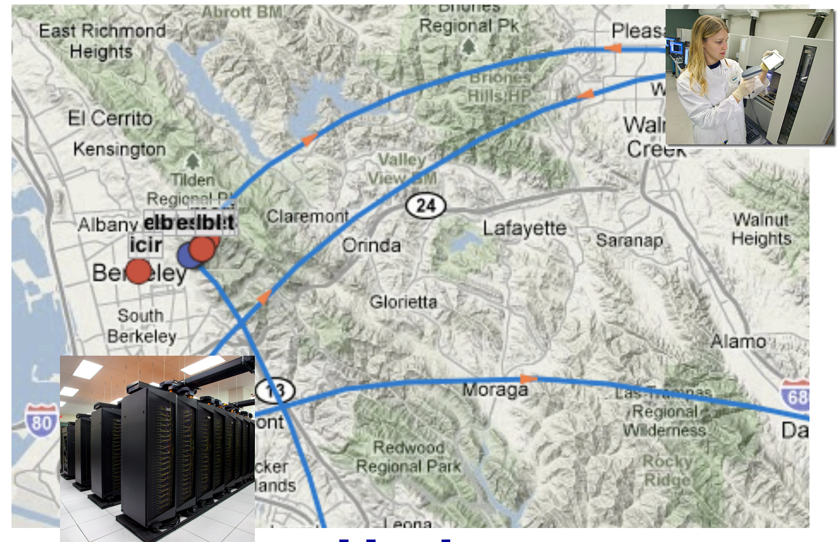
- NERSC goal is application performance (~10x every 3 years)
- Peak numbers assume (generous) 10% of peak for applications

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

Provide Cloud Computing Testbed and Evaluation

- **Demonstrated on-demand access to cycles for JGI**
  - **Esnet provisioned 9 GB Layer 2 circuit**
  - **NERSC configured 120-node cluster in Magellan**
  - **Data stayed at JGI**



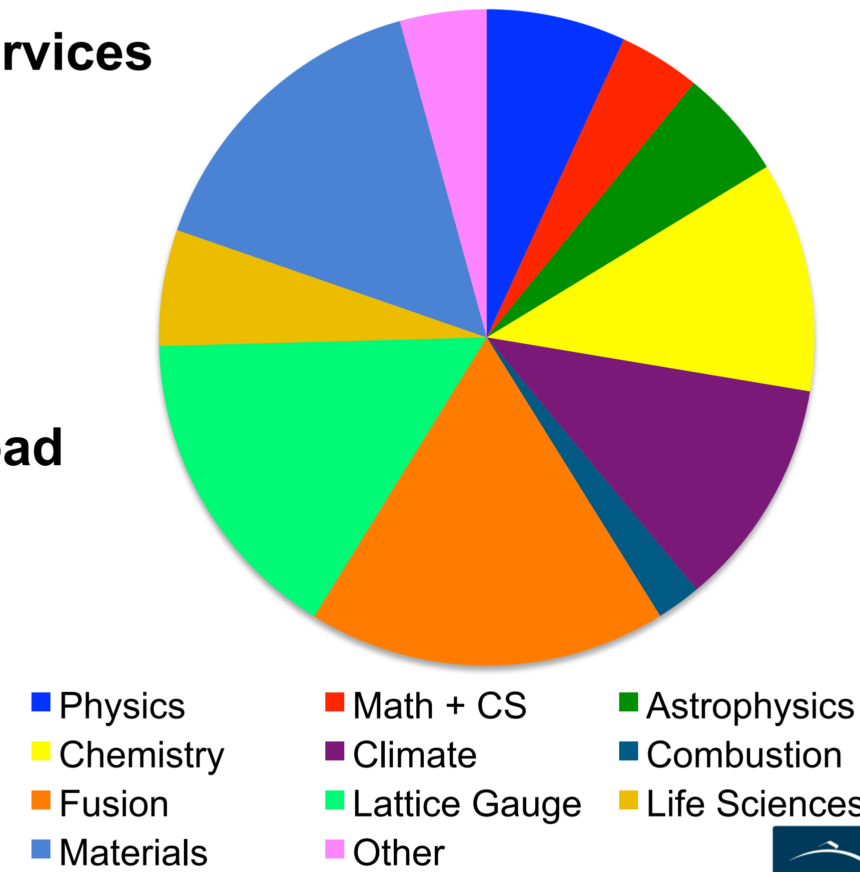
- **Deployed a MapReduce cluster running Hadoop**
  - **JGI removing errors from 5 billion reads of "next generation" sequence data (Rumen HiSeq dataset)**
  - **Next experiment will be Eucalyptus (virtualization)**
- **Demonstrated Hadoop model on Franklin**
- **Evaluating performance trade-offs of clouds**

NERSC Mission

NERSC's mission is to *accelerate the pace of scientific discovery* by providing high-performance computing, information, data, and communications services to the DOE Office of Science community.

NERSC is the Primary Computing Center for DOE Office of Science

- **NERSC serves a large population**
- **Focus on "unique" resources**
  - **Expert consulting and other services**
  - **High end computing systems**
  - **High end storage systems**
- **NERSC is known for:**
  - **Outstanding services**
  - **Large and diverse user workload**



"NERSC continues to be a gold standard of a scientific High Performance Computational Facility."
– HPCOA,Review August 2008