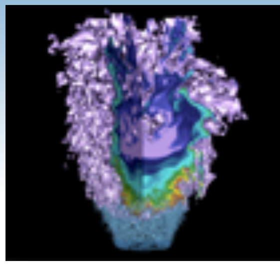
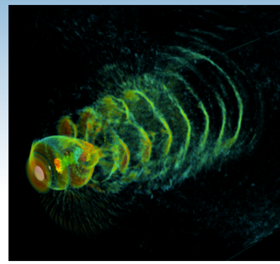
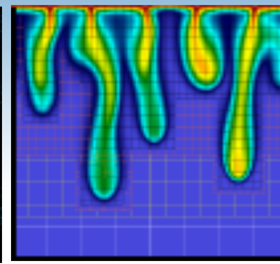
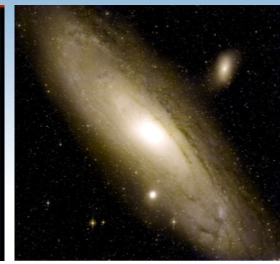
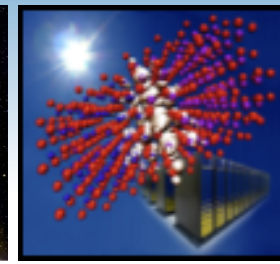
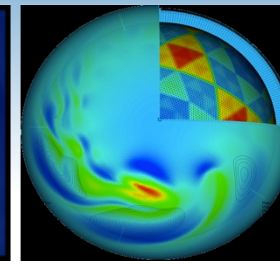







Final Report on Magellan and Update on Advanced Networking Initiative

Kathy Yelick
Associate Laboratory Director for Computing Sciences Lawrence Berkeley National Laboratory

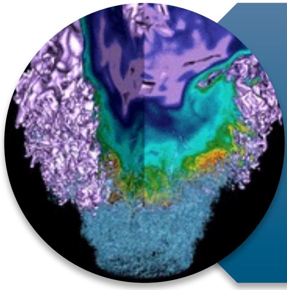
Professor of EECS, UC Berkeley

U.S. DEPARTMENT OF ENERGY | Office of Science

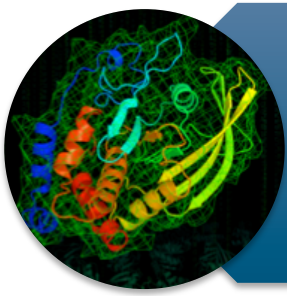
NERSC

BERKELEY LAB
Lawrence Berkeley National Laboratory

High Performance Computing in Science



Science at Scale



Science through Volume

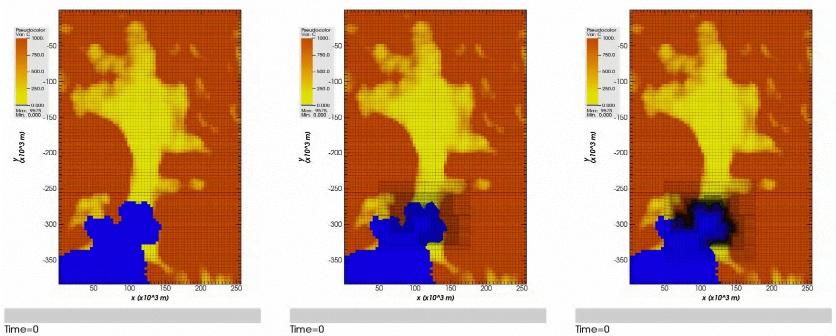
Science in Data

U.S. DEPARTMENT OF ENERGY | Office of Science

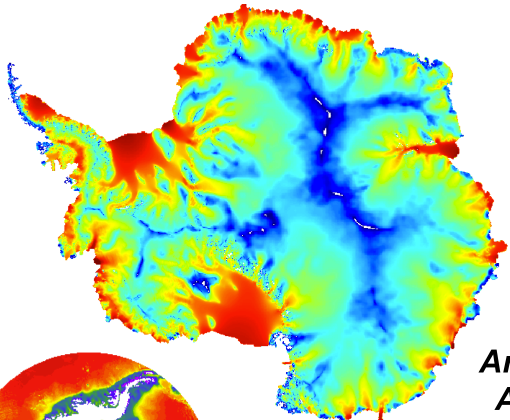
BERKELEY LAB
Lawrence Berkeley National Laboratory

Science at Scale: Simulations Aid in Understanding Climate Impacts

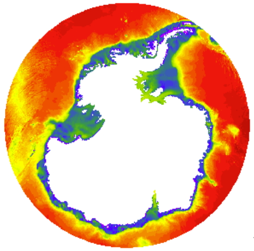
- **Warming ocean and Antarctic ice sheet key to sea level rise**
- **Previous models inadequate**
- **BISICLES ice sheet model built on FASTMath Chombo uses AMR to resolve ice-ocean interface.**
  - **Dynamics very fine resolution (AMR)**
  - **Antarctica still very large (scalability**
- **Ongoing collaboration among BISICLES and BER-sponsored IMPACTS, COSIM to couple ice sheet and ocean models**
  - **19M ALCC Hours at NERSC**



BISICLES Pine Island Glacier simulation – mesh resolution crucial for grounding line behavior.



Antarctic ice speed (left): AMR enables sub-1 km resolution (black, above)
(Using NERSC's Hopper)

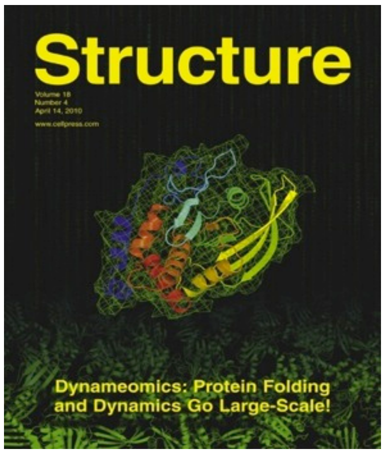


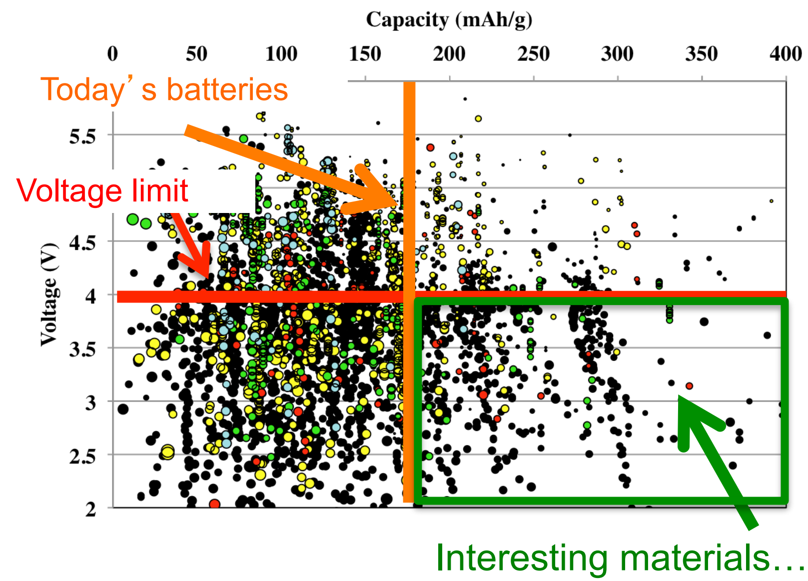
Enhanced POP ocean model solution for coupling to ice

U.S. DEPARTMENT OF ENERGY | Office of Science

Science through Volume: Screening Diseases to Batteries

- **Large number of simulations covering a variety of related materials, chemicals, proteins,…**


Structure
Dynameomics: Protein Folding and Dynamics Go Large-Scale!



Dynameomics Database
Improve understanding of disease and drug design, e.g., 11,000 protein unfolding simulations stored in a public database.

Materials Genome
Cut in half the 18 years from design to manufacturing, e.g., 20,000 potential battery materials stored in a database

Science in Data: From Simulation to Image Analysis

LBNL Computing key in 3 Nobel Prizes

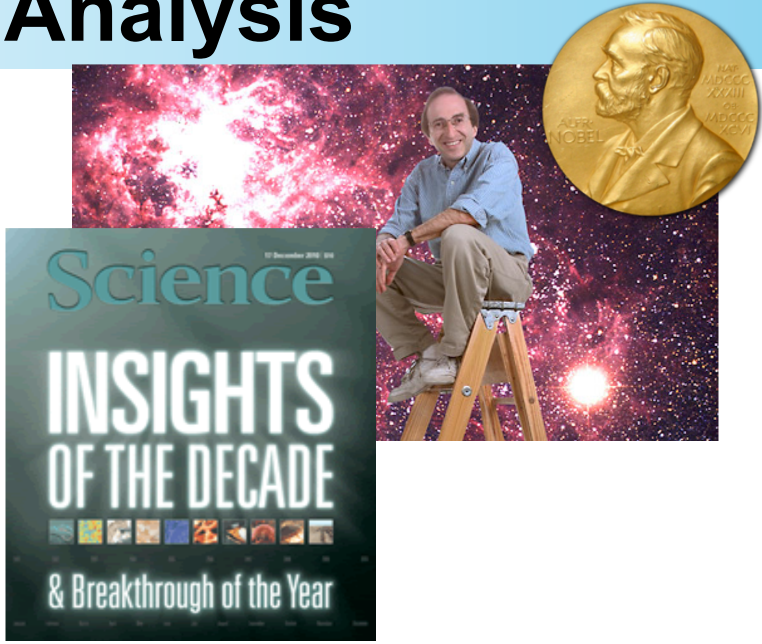
- Simulations at NERSC modeled the appearance of Supernovae.
- CMB data analysis done at CRD/NERSC
- IPCC simulations have used NERSC

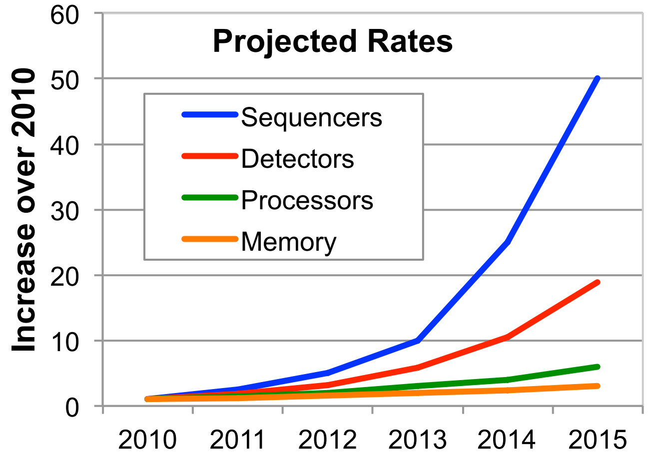
LBNL Computing key in 4 of 10 Science Breakthroughs of the decade

- 3 Genomics problems + CMB

Data rates from experimental devices will require exascale volume computing

- Cost of sequencing > Moore's Law
- Rate+Density of CCDs > Moore's Law
- Computing > Data, $O(n^2)$ common
- Computing performance < Moore Law





U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

DOE Facilities have Huge Science Data Challenges

Astronomy

Particle Physics

Chemistry and Materials

Genomics

Fusion

Petascale to Exascale

- Petabyte data sets today, many growing exponentially
- Processing grows super-linearly
- Exascale is both a driver and solution to Data challenges

Two ARRA Projects to Explore Advanced Technology for Science

- **ANI: Advanced Networking Initiative**

Science in Data

- **Magellan: Cloud testbed for science**

Science through Volume

ESnet is a Unique Capability for Science

ESnet designed for large data

- Connects 40 DOE sites to 140 other networks
- 72% annual traffic growth exceeds commercial networks
- 50% of traffic is from “big data”

First in performance:

- First 100G continental scale network
- Will transition to production this year
- ANI dark fiber can be leveraged to develop and deliver 1 terabit
- Services: Bandwidth reservations, monitoring, research testbeds

ESnet Policy Board

Policy Board highlights:

– Outstanding people/operations to be preserved

– Leverage unique dark fiber testbed for data-intensive science and basic networking research

Advanced Networking Initiative

- Goal: Accelerate 100 Gbps networking
- 100Gbps Prototype National Network
  - 4 sites (ALCF, OLCF, NERSC, and NY international exchange point)
- Network Research Testbed
  - Dark fiber
  - Research project support
- Starting point in 2009:
  - No 100Gbps standard; no carrier plans for 100G; little dark fiber due to consolidation

Advanced Networking Initiative

2009: "Table-top" testbed created;
Purchased Long Island dark fiber

2010: Transport RFP released;
Thirteen testbed projects started

2011: Partner with Internet2 (Level3 /
Cienna I / Alcatel-Lucent)
100Gb Prototype to 4 sites;

2012: Complete network buildout (Oct);
100G production "ESnet5" (Dec)

Unbundled Network
100Gbps Prototype Network

100Gbps Prototype Network

- Combines ANI funding with Internet2 stimulus funds to build full national footprint
- Internet2/Level3 Communications/Indiana Univ. manage the optical equipment and supporting infrastructure
- Uses Ciena Activeflex 6500 optical equipment
  - Backbone network: chassis and fiber owned by Internet2, but ESnet purchases and owns transponder cards
  - Metropolitan networks: All equipment and fiber owned by ESnet
  - Ability to provision wavelengths between any two add/drop or regeneration locations on network
- Uses Alcatel-Lucent 7750 routers
  - 14 chassis deployed with 33 100Gbps interfaces

Testbed: Monitoring And Visualization of Energy in Networks (MAVEN)

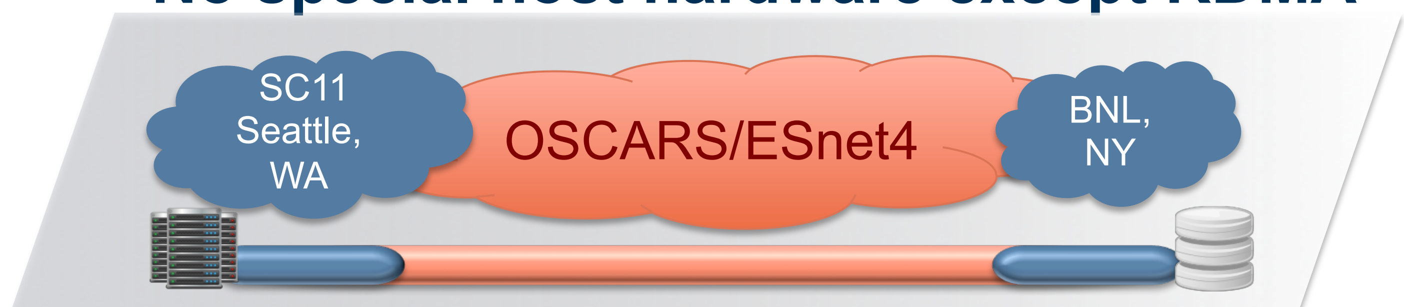
"what gets measured gets improved"

- **Establish energy baseline for end-to-end networking**
- **Provide real operational data to researchers**
- **Identify opportunities for improved efficiency**
- **Optimize globally (network of centers)**
- **First of kind in ESnet5**

Figure: Visualization of energy (alpha version, unreleased) consumed by ESnet's ANI prototype network.

Testbed: End-to-End Circuit Service with OpenFlow

- Dynamic "tunnels" across wide area
  - No manual configuration of virtual circuit
  - Automated discovery of circuit end-points
- High Performance RDMA-over-Ethernet (Remote Direct Memory Access)
  - 9.8 Gbps out of 10 Gbps NY to WA at SC11
  - Low overhead: 4% CPU vs. 80% with 1-stream TCP
  - No special host hardware except RDMA



Fully Automated, End to End, Dynamically Stitched, Virtual Connection

ANI Legacy

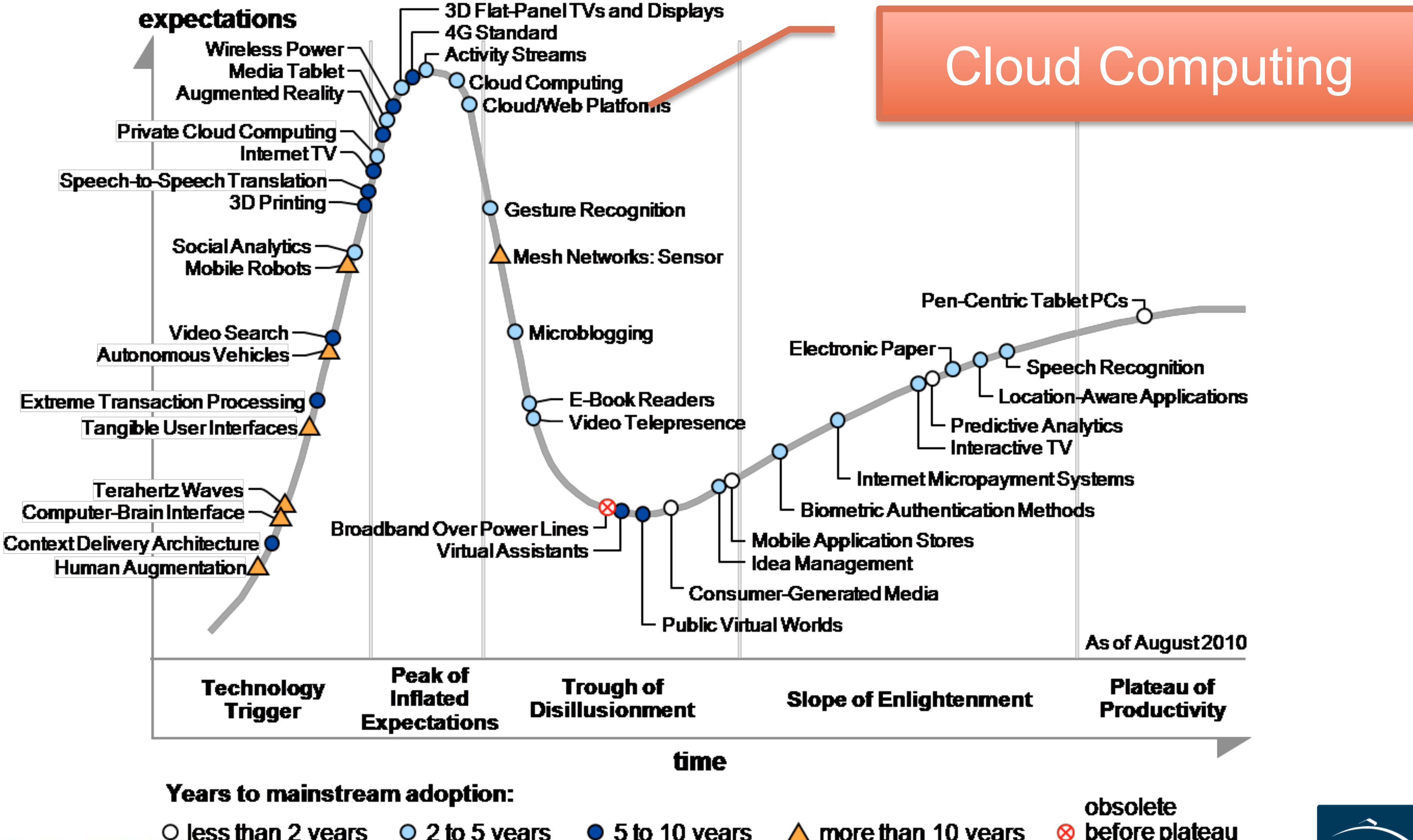
- **Unique 100G networking facility:**
  - Connects DOE facilities (experimental, computational)
- **Enables first-of-kind "Big Data" science**
  - Optimizations (OSCARS, perfSONAR, ScienceDMZ and Data Transfer Nodes)
- **Dark Fiber for future ESnet upgrades**
  - Future optical gear, routers, systems
- **Dark Fiber for networking research**
  - Enable previously-impossible wide area, high performance research for universities/companies

The Magellan Team

- **Magellan/NERSC**
  - **Shane Canon, Lavanya Ramakrishnan,** Tina Declerck, Iwona Sakrejda, Scott Campbell, Brent Draney, Jeff Broughton
- **Magellan/ANL**
  - **Susan Coghlan, Adam Scovel,** Piotr T Zbiegiel, Narayan Desai, Rick Bradshaw, Anping Liu
- **Amazon Benchmarking**
  - Krishna Muriki, Nick Wright, John Shalf, Keith Jackson, Harvey Wasserman, Shreyas Cholia
- **Applications**
  - Jared Wilkening, Gabe West, Ed Holohan, Doug Olson, Jan Balewski, STAR collaboration, K. John Wu, Alex Sim, Prabhat, Suren Byna, Victor Markowitz

Cloud Computing Hype

Gartner's 2010 Emerging Technologies Hype Cycle



What is a Cloud?

According to NIST…

- ***Resource pooling.*** **Resources are pooled across users for efficiency.**
- ***Broad network access.*** **Capabilities are available over the network.**
- ***Measured Service.*** **Usage is monitored and reported for transparency (pay-as-you-go).**
- ***Elasticity.*** **Capabilities can be rapidly scaled out and in.**
- ***Self-service.*** **Configuration without on-site system administration**

Why Clouds for Science?

- *Resource pooling.*
  - HPC Centers run at 90% utilization
  - Commercial clouds at 60% utilization
- *Measured Service (pay-as-you-go).*
  - HPC Centers charge in hours (not fungible with cash)
  - Commercial clouds charge in dollars
- *Elasticity.*
  - HPC Centers allow job scale-up but users wait in queues
  - Commercial clouds allow rapid growth in aggregate work
- *Self-service (control vs. ease-of-use).*
  - HPC Centers: fix some software (OS, compilers)
  - EC2 DIY administration; others fix entire software model

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

Magellan Research Agenda and Lines of Inquiry

- Are the *open source* cloud software stacks ready for DOE HPC science?
- Can DOE cyber security requirements be met within a cloud?
- Are the new cloud programming models useful for scientific computing?
- Can DOE HPC applications run efficiently in the cloud? What applications are suitable for clouds?
- How usable are cloud environments for scientific applications?
- When is it cost effective to run DOE HPC science in a cloud?

Magellan Testbed Architected for Flexibility

QDR Infiniband

+ 100 Gbps to ANI

Compute Servers

504 Nodes at ANL
720 Nodes at NERSC
Intel Nehalem
8 cores/node

Active Storage Servers

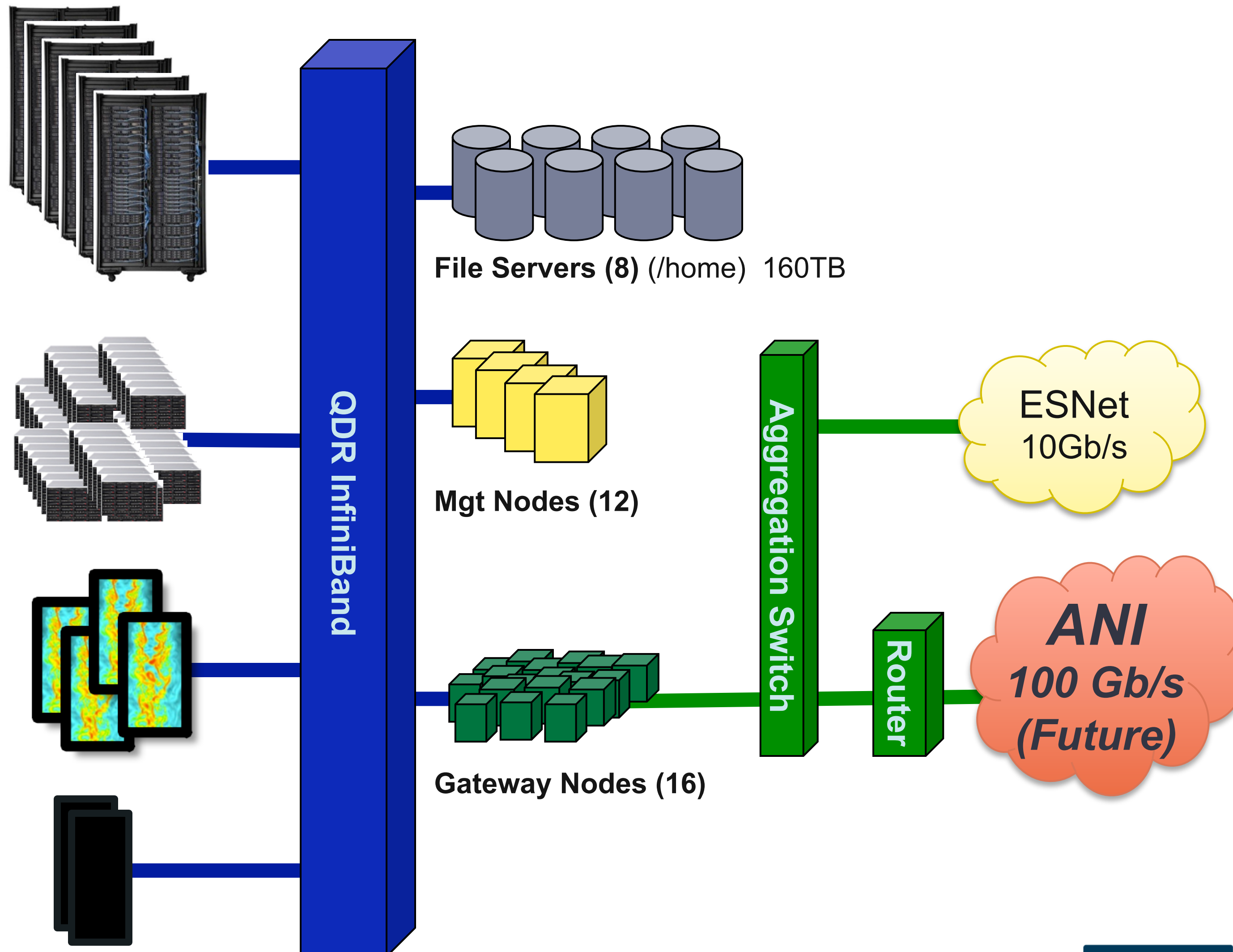
FLASH/SSD Storage

Big Memory Servers

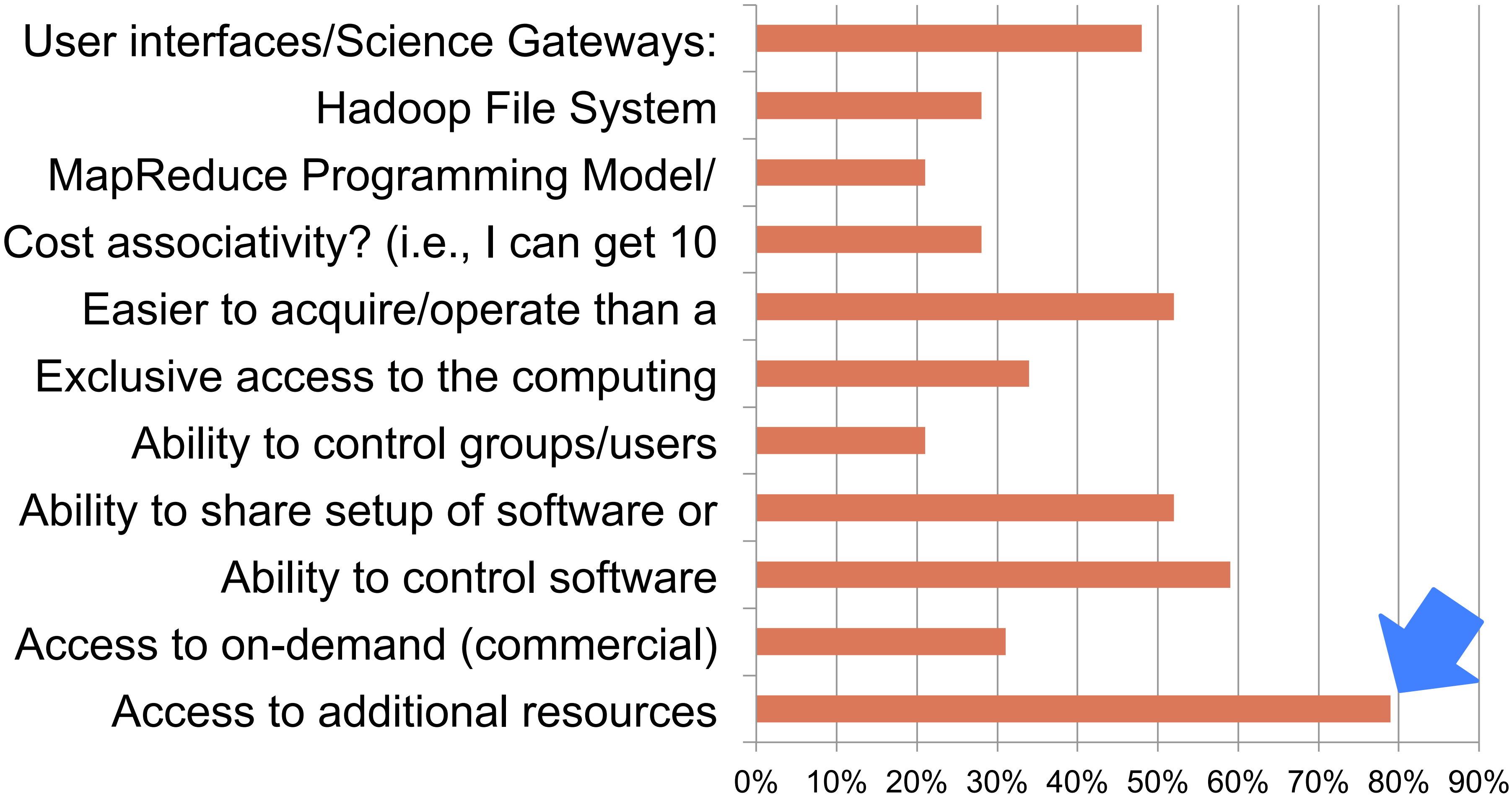
1 TB of Memory per node
15 at ANL / 2 at NERSC

GPU Servers

266 Nvidia cards at ANL

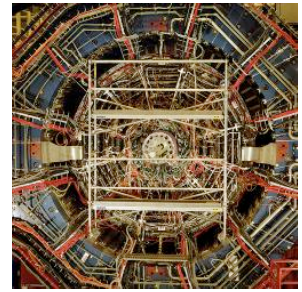


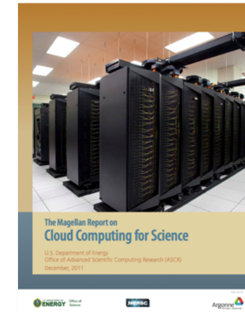
In 2009 Significant interest in cloud computing for science



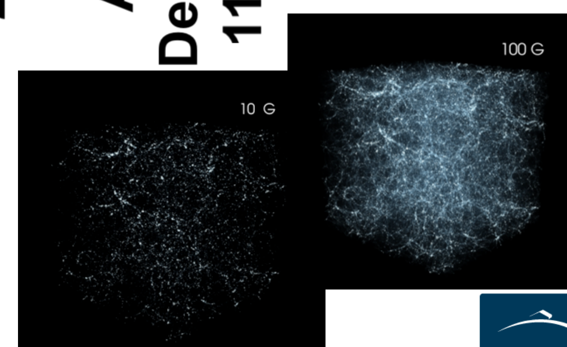
Demonstration of Cloud Technology for Science

Magellan Timeline

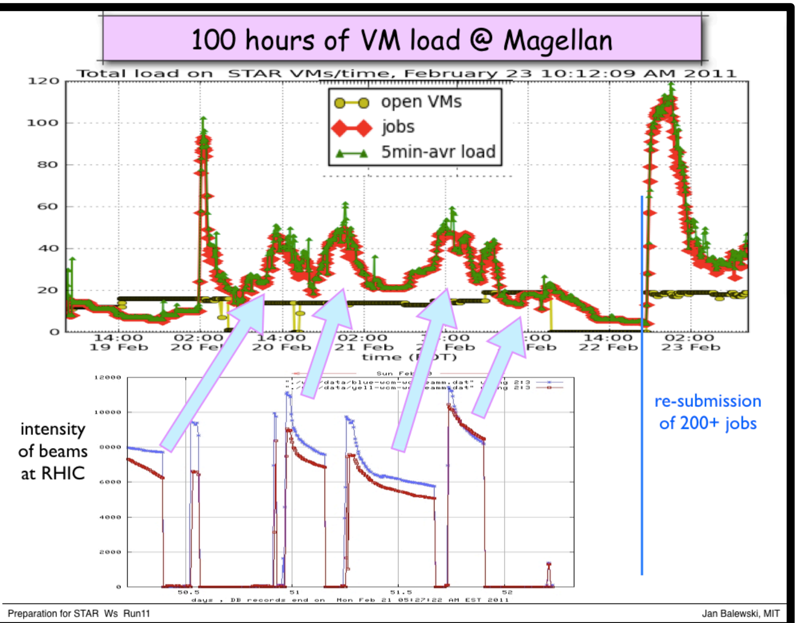




- Project Start 9/09
- Testbed Deployed 2/10
- Early Users 3/10
- JGI Demo 4/10
- Joint Demo 6/10
- Bench-marking 12/10-12/11
- OpenStack @ANL 10-11
- STAR Demo 2/11
- ASCAC Talk 3/11
- ANI Demo 11/11
- Final Report 12/11

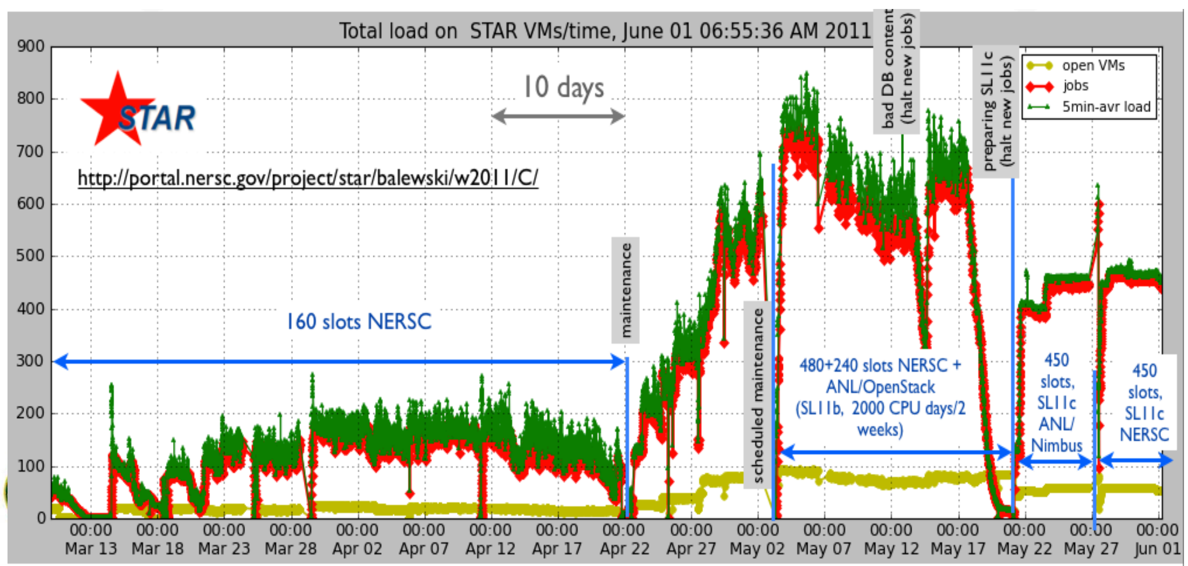


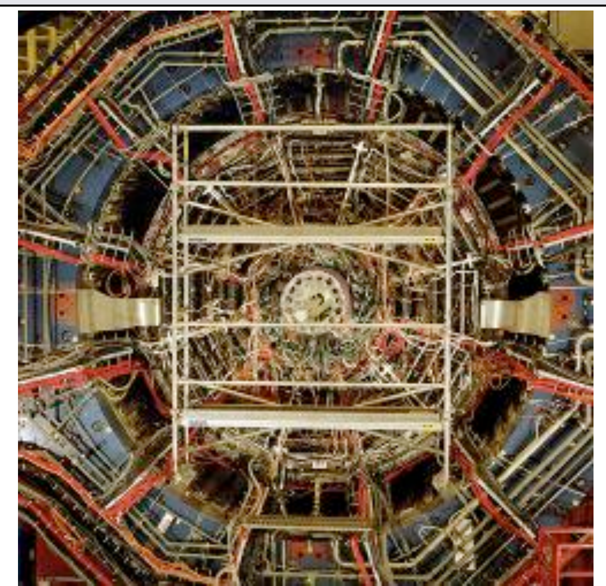
Federated Clouds provide elasticity, but with significant administrative support



STAR performed Real-time analysis of data coming from Brookhaven Nat. Lab

- First time data was analyzed in real-time to a high degree
- Leveraged existing OS image from NERSC system
- Started out with 20 VMs at NERSC and expanded to ANL.





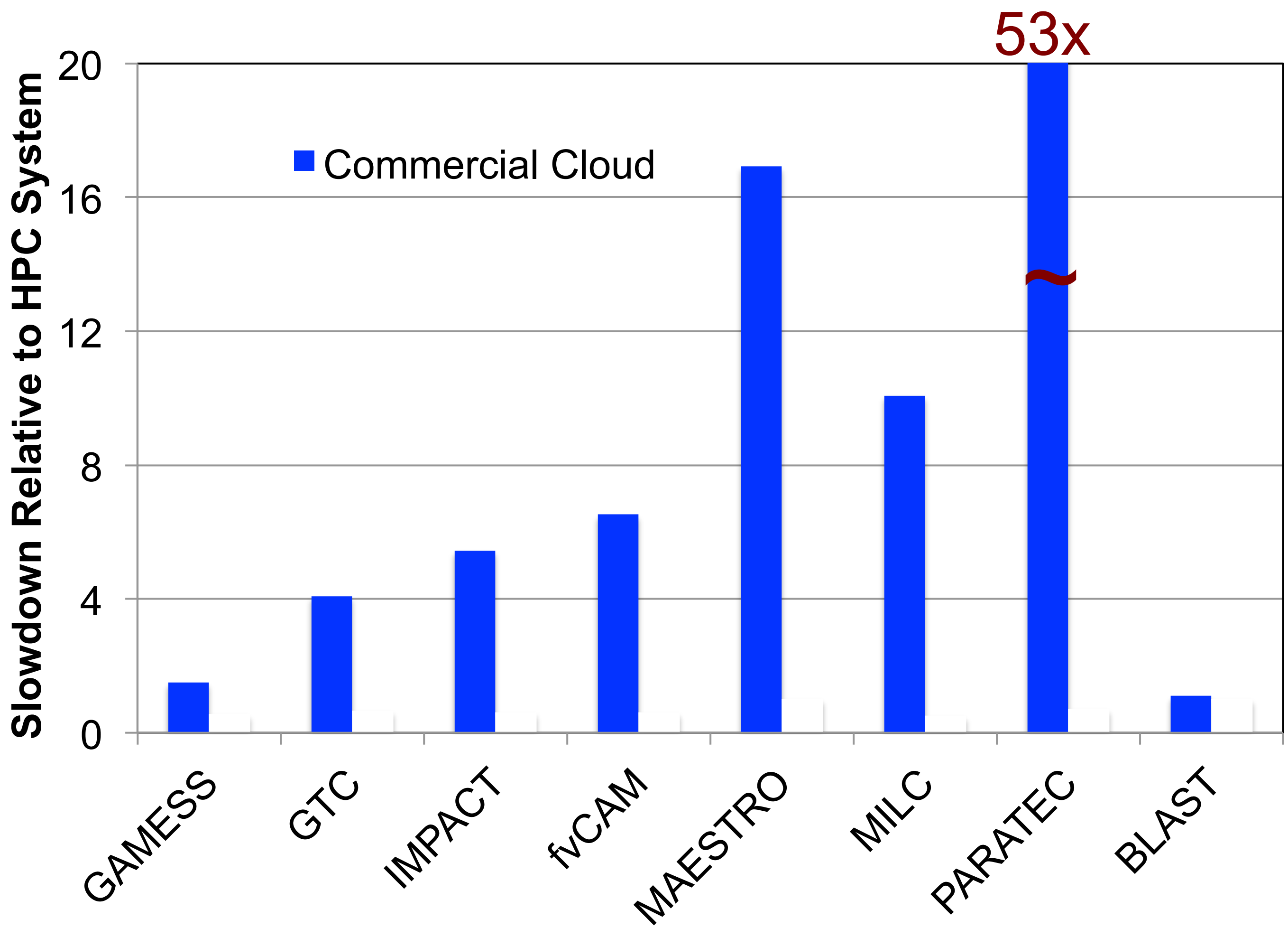
Performance of Clouds for Science

Applications Cover Algorithm and Science Space

Science areas	*Dense*	*Sparse*	*Spectral*	*Particles*	*Structured*	*Unstructured*	*Independent*
Accelerators		X	X **IMPACT**	X **IMPACT**	X **IMPACT**	X	
Fluids / Astro	X	X **MAESTRO**	X	X	X **MAESTRO**	X **(MAESTRO)**	
Chemistry	X **GAMESS**	X	X	X			
Climate			X **CAM**		X **CAM**	X	
Fusion	X	X		X **GTC**	X **GTC**	X	
Nuclear QCD		X **MILC**	X **MILC**	X **MILC**	X **MILC**		
Materials	X **PARATEC**		X **PARATEC**	X	X **PARATEC**		
Biology							X **BLAST**

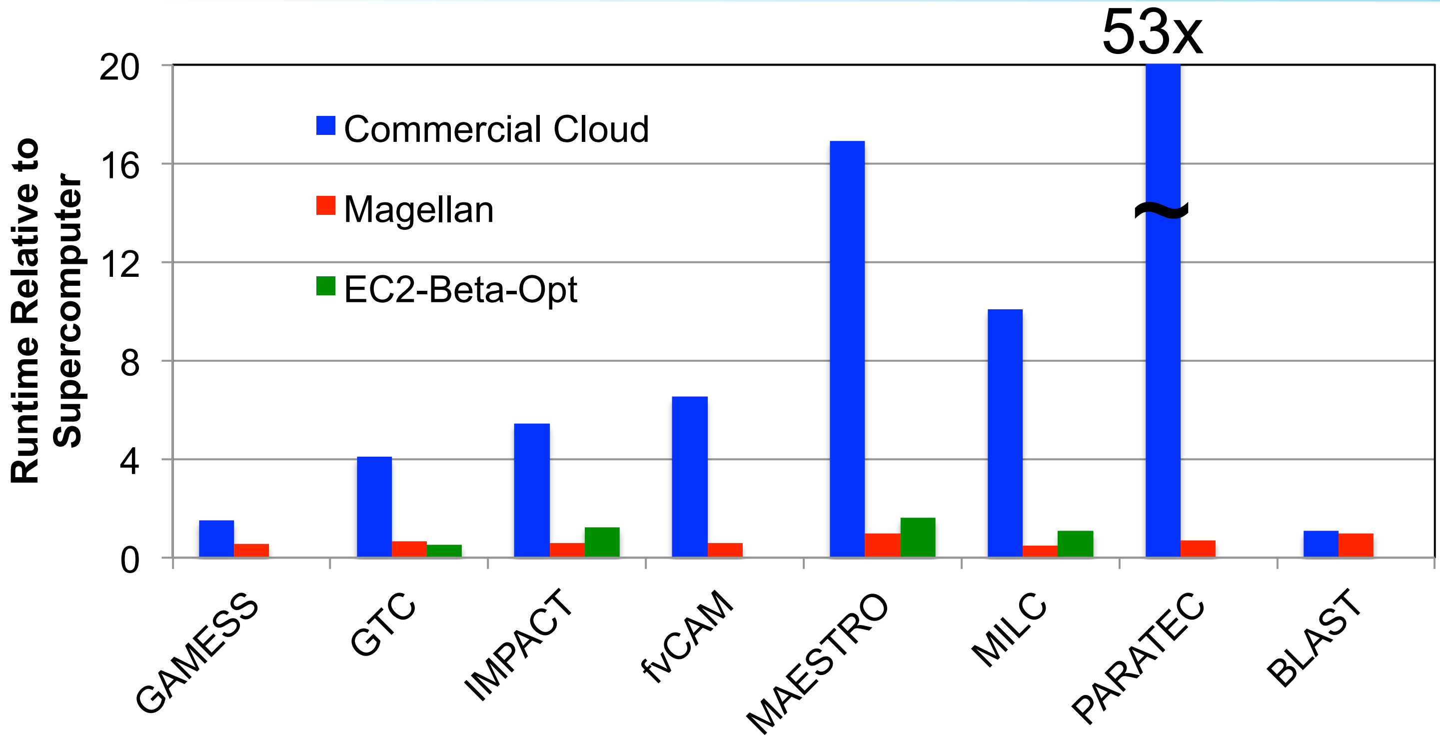
Parallel job size and input data drastically reduced for cloud benchmarking

Slowdown of Clouds Relative to an HPC System



Study by Jackson, Ramakrishnan, Muriki, Canon, Cholia, Shalf, Wasserman, Wright

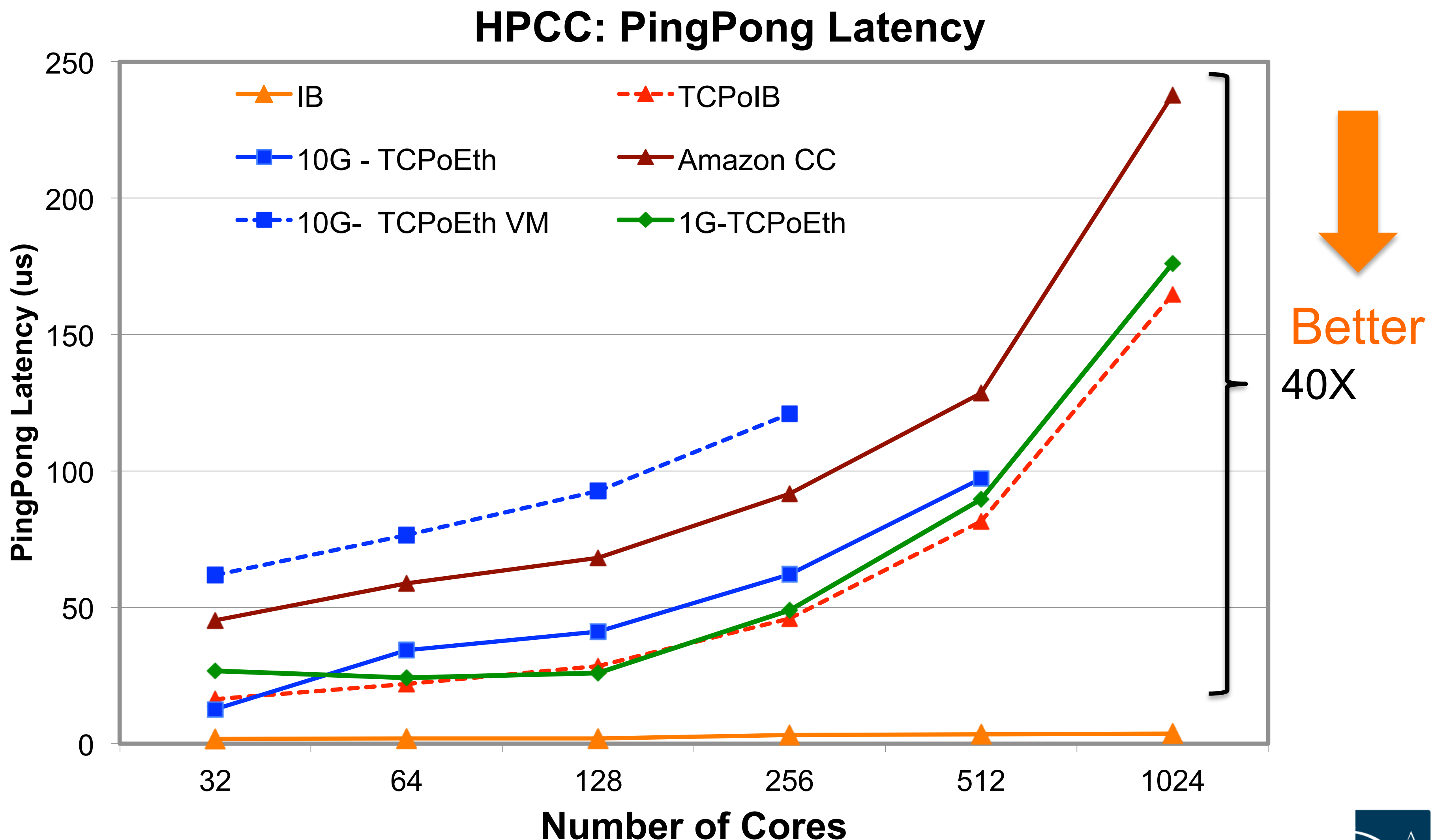
HPC Commercial Cloud Results



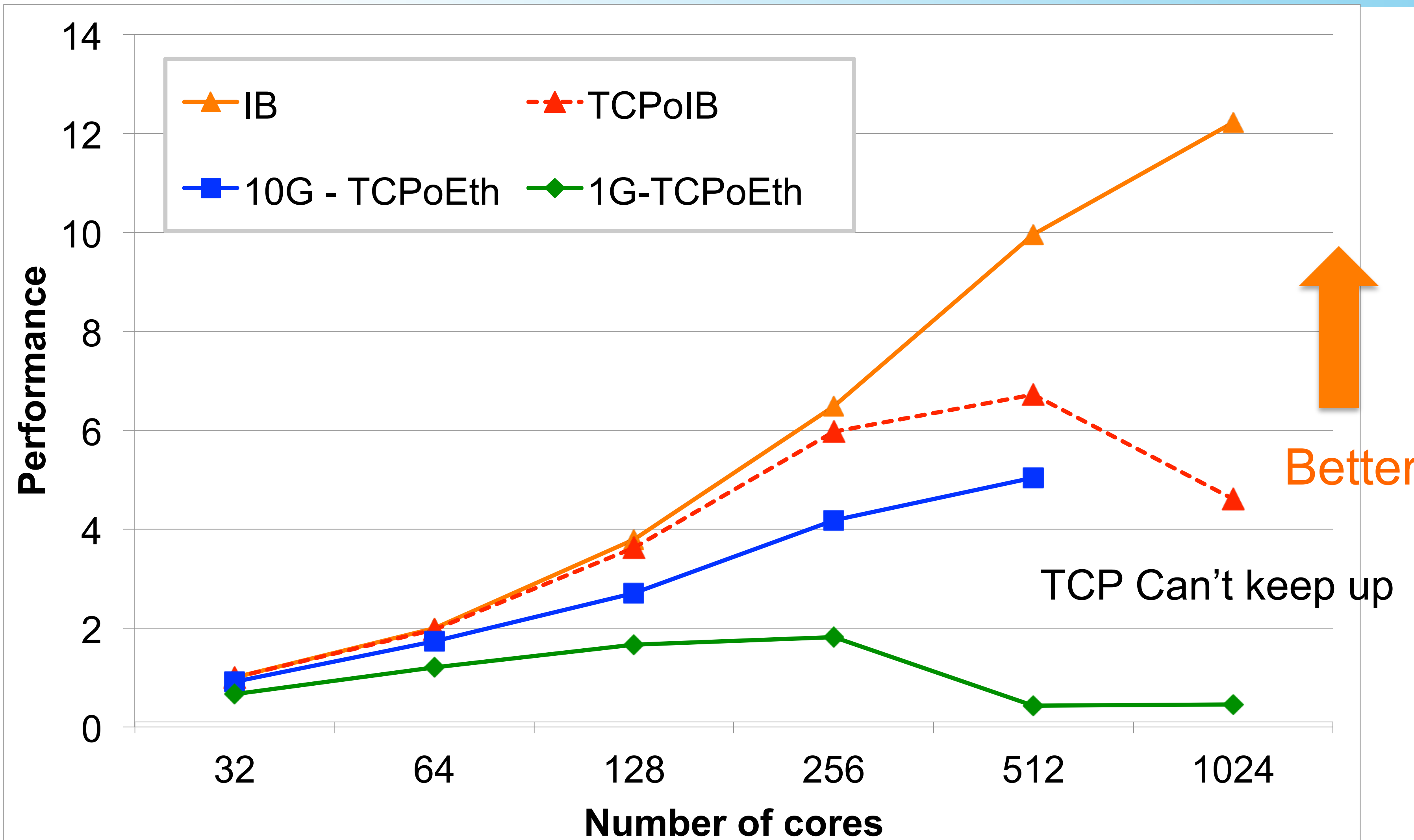
- **Commercial HPC clouds catch up with clusters if set up as shared cluster**
  - **High speed network (10GigE) and no over-subscription**
  - **Some slowdown from virtualization**

Keith Jackson, Lavanya Ramakrisha, John Shalf, Harvey Wasserman

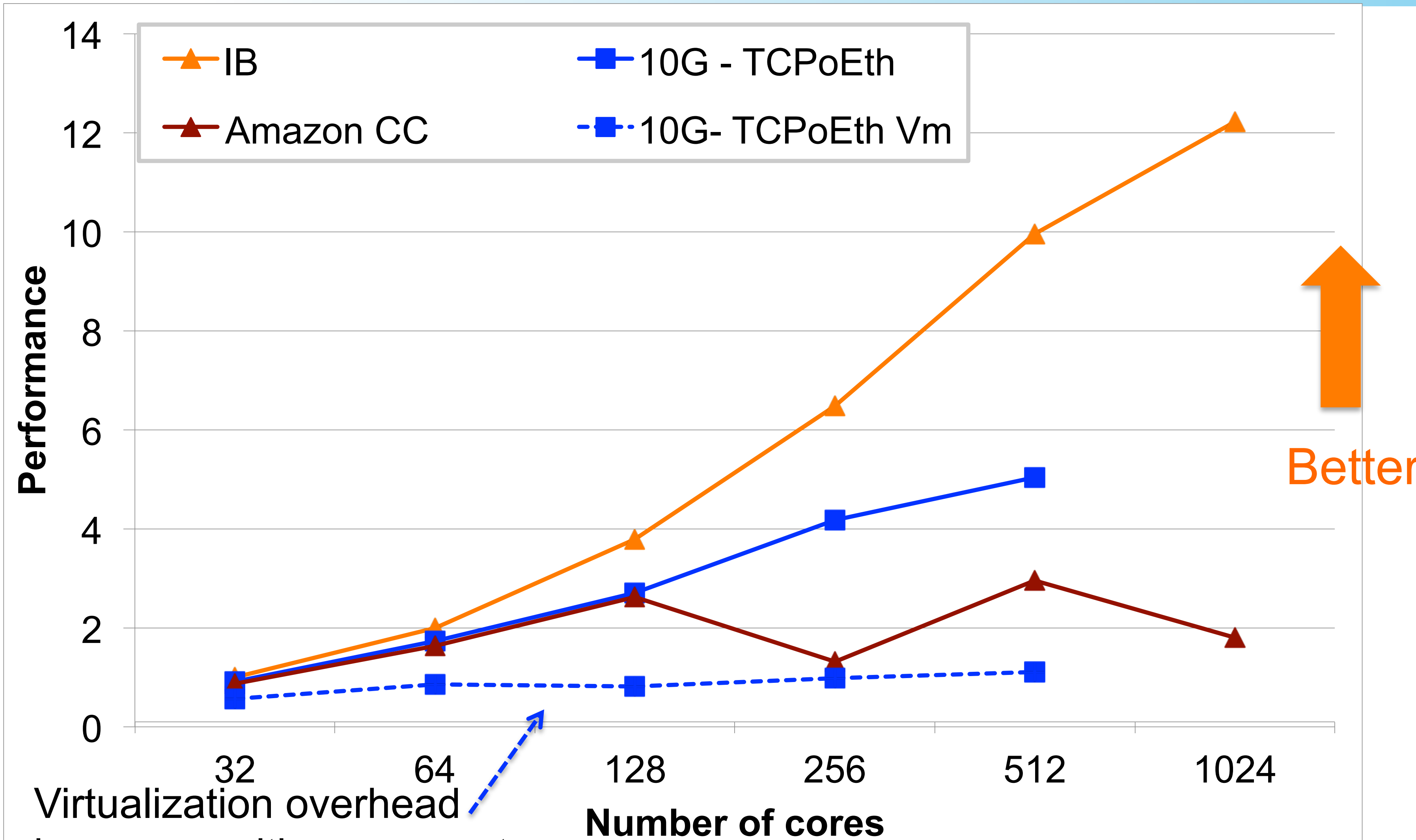
TCP is slower than IB even at modest concurrency



Network Hardware and Protocol Matter (PARATEC)



Virtualization Penalty is Substantial (PARATEC)



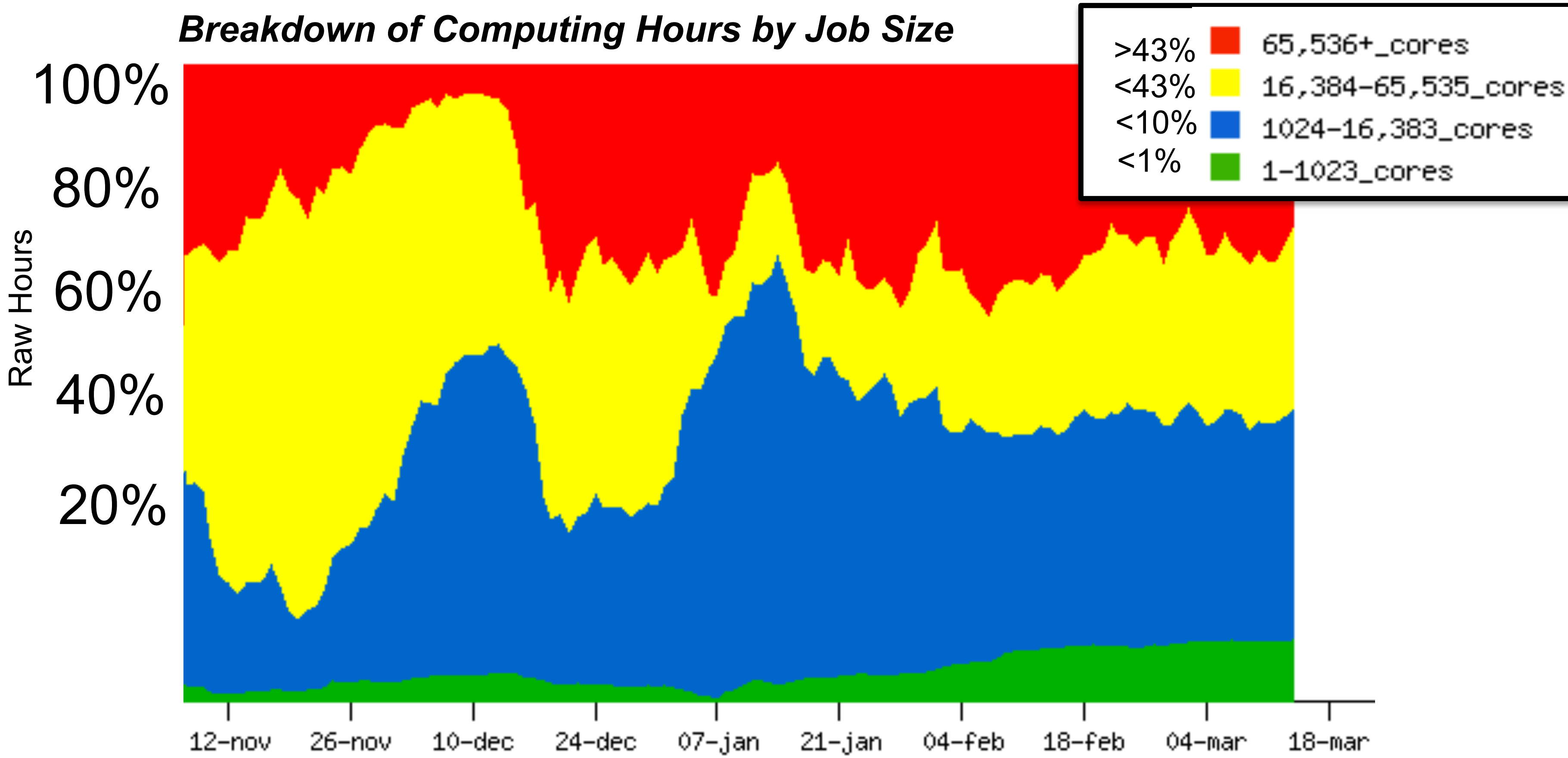
Virtualization overhead increases with core count

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB

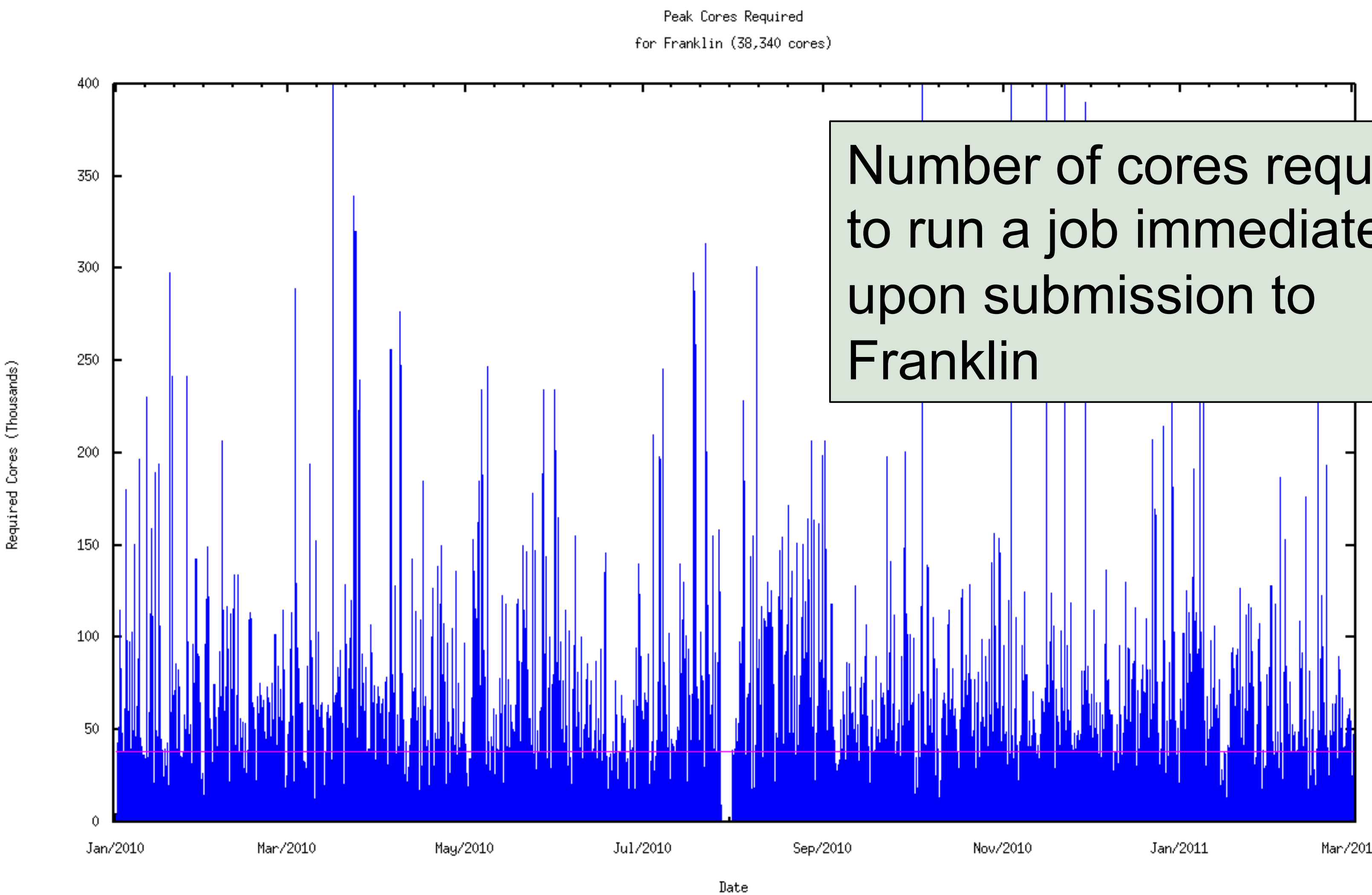
Elasticity Requirements for Science

Job Size Mix on Hopper "Unleashed"



- *Hopper is a 153,216 core system. During availability period, over 50% of hours were used for jobs larger than 16k cores.*

On-demand science access might be difficult if not impossible



Number of cores required to run a job immediately upon submission to Franklin

Costs of Clouds for Science

Cloud is a business model and can be applied to HPC centers

	Cloud	HPC Centers
NIST Definition	Resource Pooling, Broad network access, measured service, rapid elasticity, on-demand self service	Resource Pooling, Broad network access, measured service. Limited: rapid elasticity, on-demand self service
Computational Needs	Bounded computing requirements – Sufficient to meet customer demand or transaction rates.	Virtually unbounded requirements – Scientist always have larger, more complicated problems to simulate or analyze.
Scaling Approach	Scale-in. Emphasis on consolidating in a node using virtualization	Scale-Out Applications run in parallel across multiple nodes.
Workloads	High throughput modest data workloads	High Synchronous large concurrencies parallel codes with significant I/O and communication
Software Stack	Flexible user managed custom software stacks	Access to parallel file systems and low-latency high bandwidth interconnect. Preinstalled, pre-tuned application software stacks for performance

Public clouds compared to private HPC Centers

Component	Cost
Compute Systems (1.38B hours)	$180,900,000
HPSS (17 PB)	$12,200,000
File Systems (2 PB)	$2,500,000
Total (Annual Cost)	**$195,600,000**

Over estimate: These are “list” prices, but...
Underestimate:

- Doesn’t include the measured performance slowdown 2x-10x.
- This still only captures about 65% of NERSC’s $55M annual budget. **No consulting staff, no administration, no support.**

Factors in Price

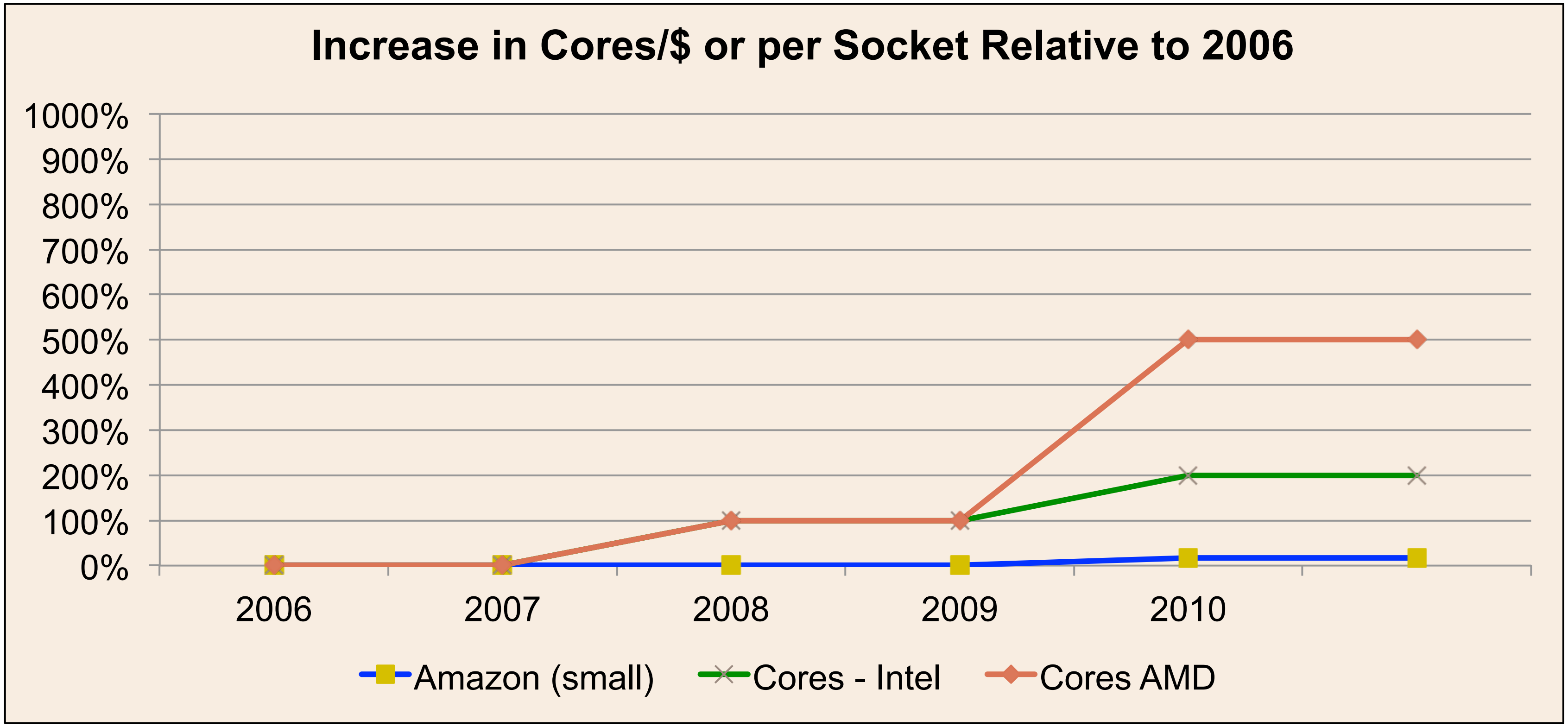
Factor	HPC Center	Public Cloud
Utilization (30% private, 90% HPC, 60%? Cloud); Note: trades off against wait times, elasticity		$$
Cost of people, largest machines lowest people costs/core	$	
Cost of power, advantage for placement of center, bulk	$$	
Energy efficiency (PUE, 1.1-1.3 is possible; 1.8 typical)		
Cost of specialized hardware (interconnect)	$	
Cost of commodity hardware	$	
Profit		$$$

$ means “cost disadvantage”

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

Where is Moore's Law (Cores/$) in Commercial Clouds?



- Cost of a small instance at Amazon dropped 18% over 5 years.
- Cores increased 2x-5x per socket; roughly constant cost.
- NERSC cost/core dropped by 10x (20K – 200K cores in 2007-2011)

Cloud Artifacts

- Lessons for HPC Centers from Clouds
  - Provide higher service level (for higher price) with guaranteed low wait
  - Allow users to control access (buy time)
  - Provide for configurable systems software
- Other features associated with Clouds
  - Virtualization for over-subscription of nodes
  - Map-Reduce programming model

Key Findings

- Cloud approaches provide many useful benefits such as customized environments and access to surge capacity.
- Cloud computing can require significant initial effort and skills in order to port applications to these new models.
- Significant gaps and challenges exist in the areas of managing virtual environments, workflows, data, cyber-security, etc.
- The key economic benefit of clouds comes from the consolidation of resources across a broad community, which results in higher utilization, economies of scale, and operational efficiency. DOE already achieves this with facilities like NERSC and the LCFs.
- Cost analysis shows that DOE centers are cost competitive, typically 3–7x less expensive, when compared to commercial cloud providers.

Magellan Legacy

- Magellan project is complete
- Hardware and infrastructure is still valuable
- DOE Systems Biology Knowledge Base
  - BER-funded
  - Hardware from Magellan
  - *Community-Driven Cyberinfrastructure for Sharing and Integrating Data and Analytical Tools to Accelerate Predictive Biology*

KBASE predictive biology

DOE Systems Biology Knowledgebase

- GPUs to become next ALCF vis/DA cluster
- Other Strategic Projects at NERSC
  - Data at large DOE facilities: Call for Proposals

Coming Soon!

- Use of private clouds at ANL

U.S. DEPARTMENT OF ENERGY | Office of Science

BERKELEY LAB Lawrence Berkeley National Laboratory

Magellan Final Report

- **Final Report released on ASCR website**
- **Joint ANL/NERSC**
- **Comprehensive**
  - 170 pages
  - User Experiences
  - Benchmarking
  - Programming
  - Security
  - Cost Analysis