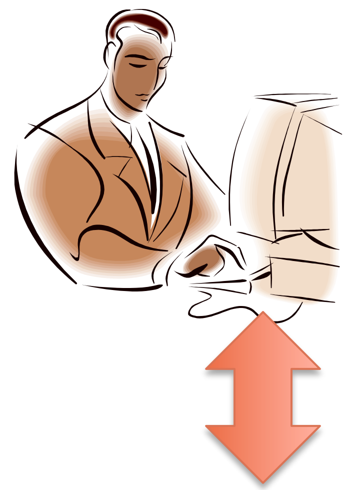
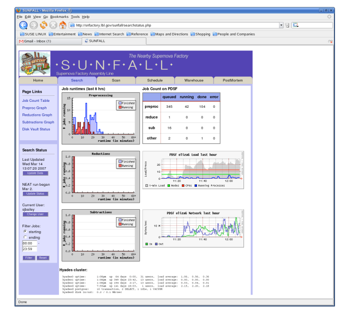
Science in the Clouds and Beyond

Lavanya Ramakrishnan

Computational Research Division (CRD) &
National Energy Research Scientific Computing Center (NERSC)
Lawrence Berkeley National Lab





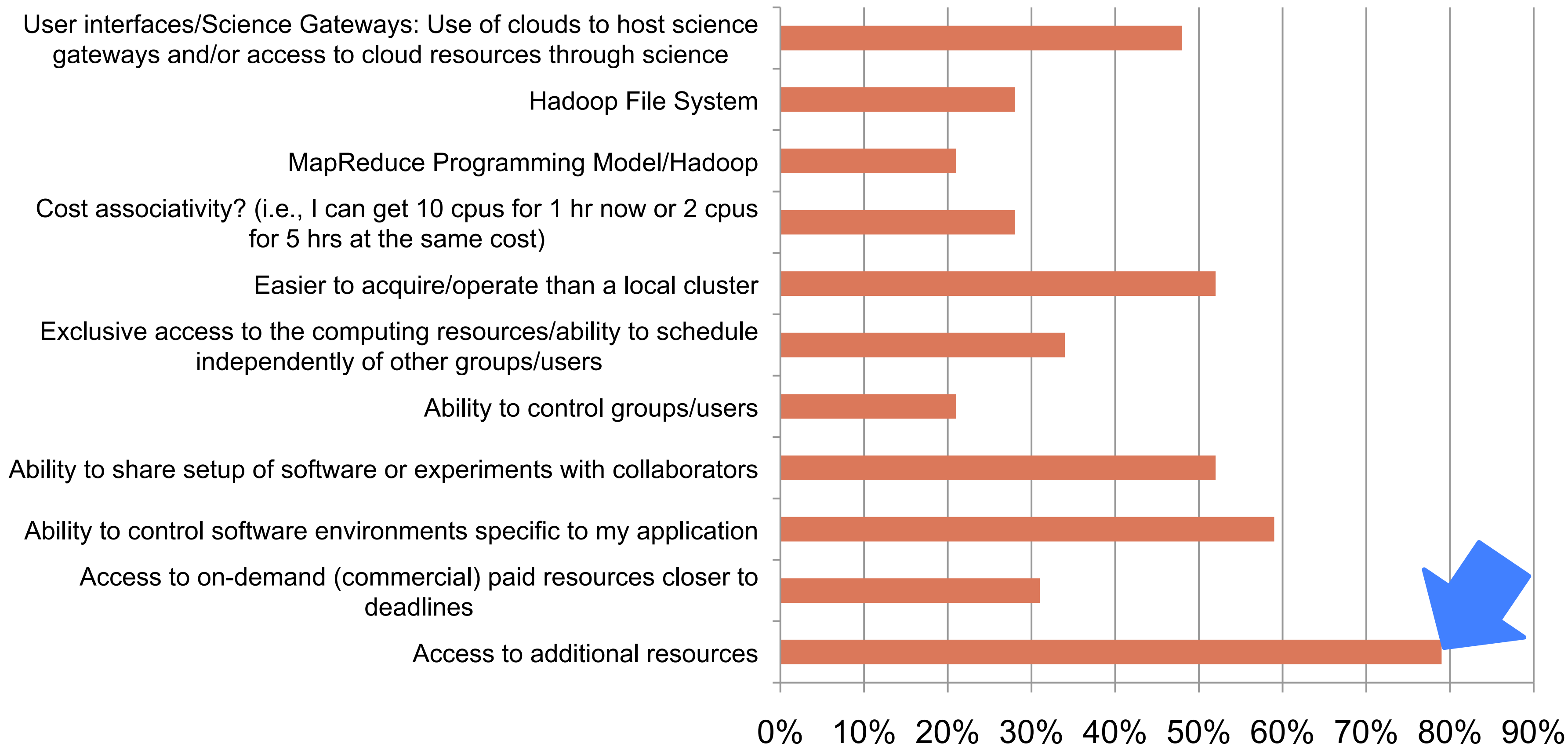
U.S. DEPARTMENT OF ENERGY | Office of Science

ESnet Energy Sciences Network

CRD computational research division

NERSC

The goal of Magellan was to determine the appropriate role for cloud computing for science



Program Office	
Advanced Scientific Computing Research	17%
Biological and Environmental Research	9%
Basic Energy Sciences -Chemical Sciences	10%
Fusion Energy Sciences	10%

Program Office	
High Energy Physics	20%
Nuclear Physics	13%
Advanced Networking Initiative (ANI) Project	3%
Other	14%

Magellan was architected for flexibility and to support research

Compute Servers

720Compute Servers
  Nehalem Dual quad-core 2.66GHz
  24GB RAM, 500GB Disk
Totals
  5760 Cores, 40TF Peak
  21TB Memory, 400 TB Disk

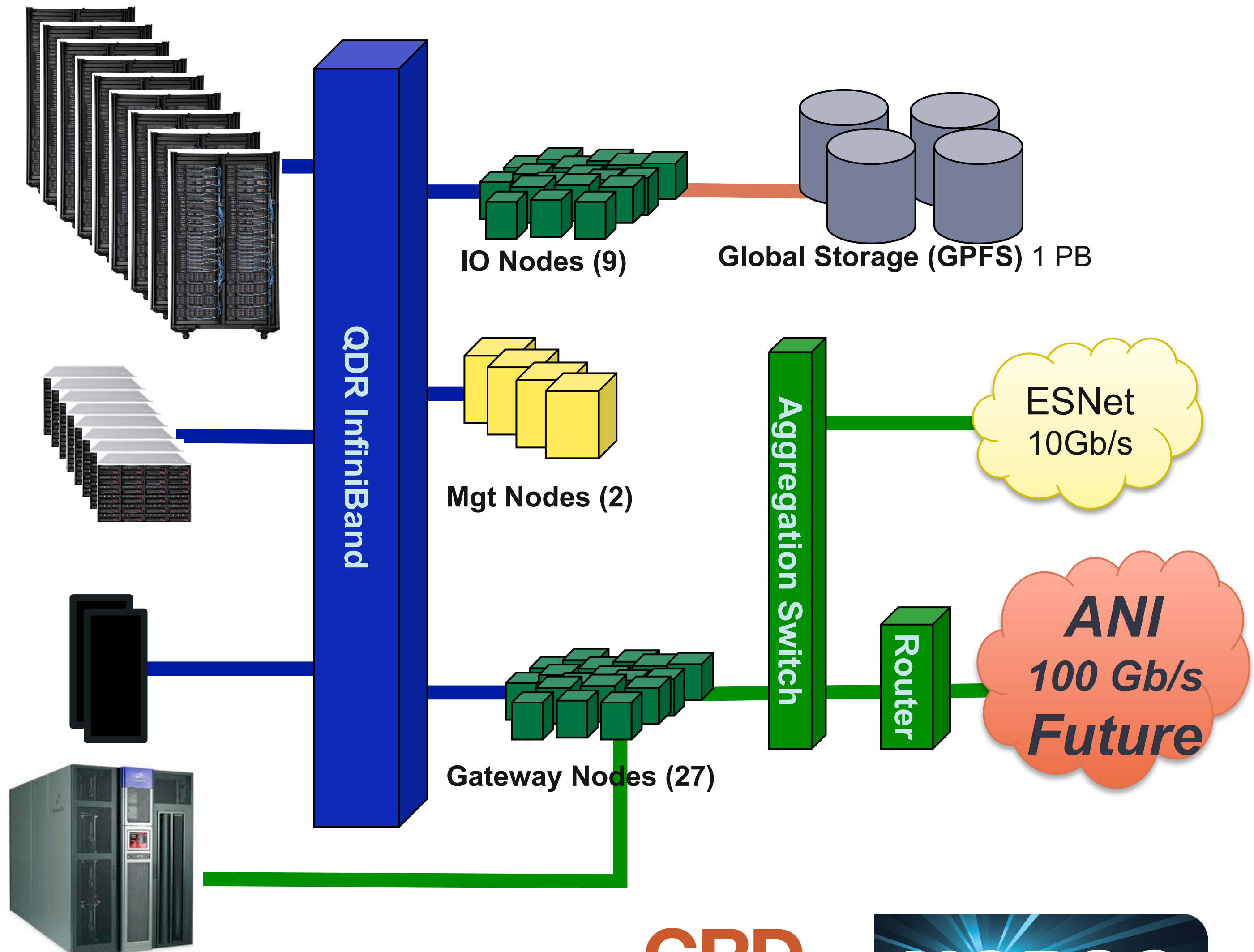
Flash Storage Servers

10 Compute/Storage Nodes
8TB High-Performance FLASH
20 GB/s Bandwidth

Big Memory Servers

2 Servers
  2TB Memory

Archival Storage



U.S. DEPARTMENT OF ENERGY | Office of Science

CRD computational research division

NERSC

Science + Clouds = ?

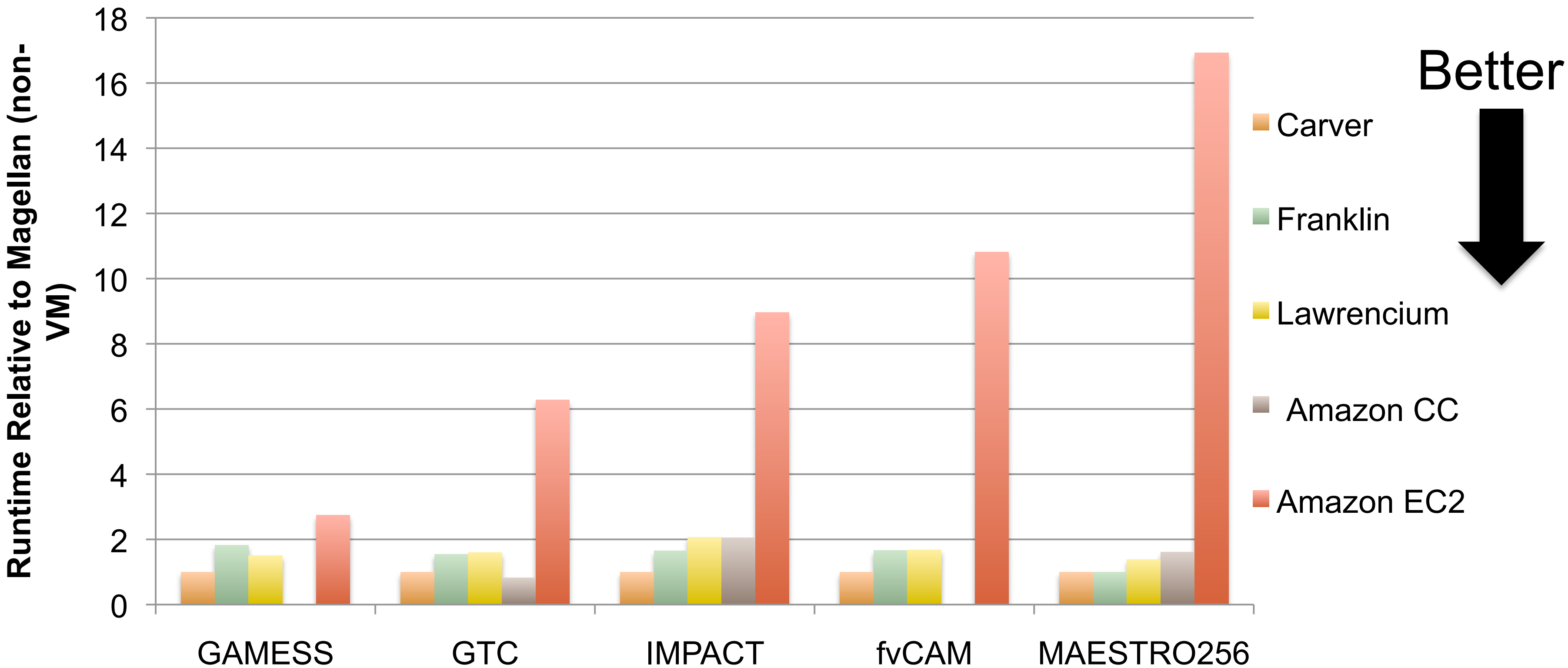
Business model for Science

Performance and Cost

Data Intensive Science

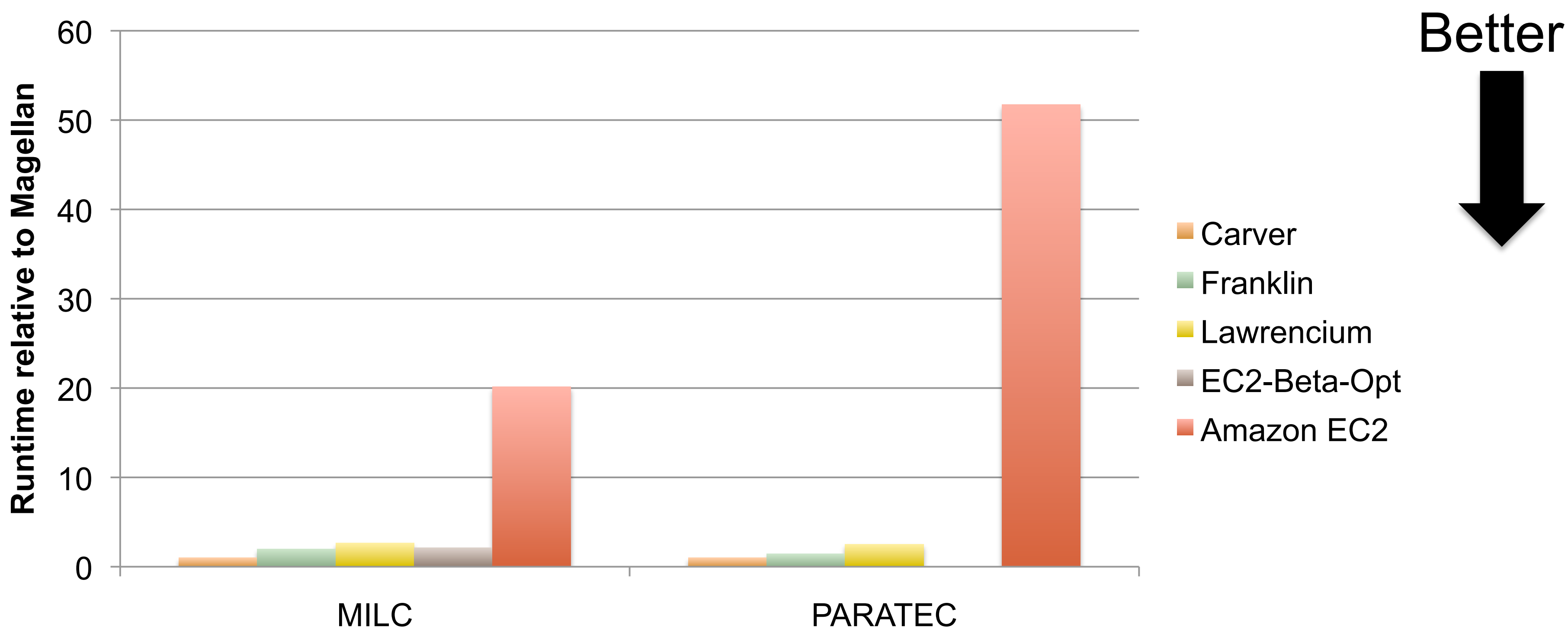
Technologies from Cloud

Scientific applications with minimal communication are best suited for clouds

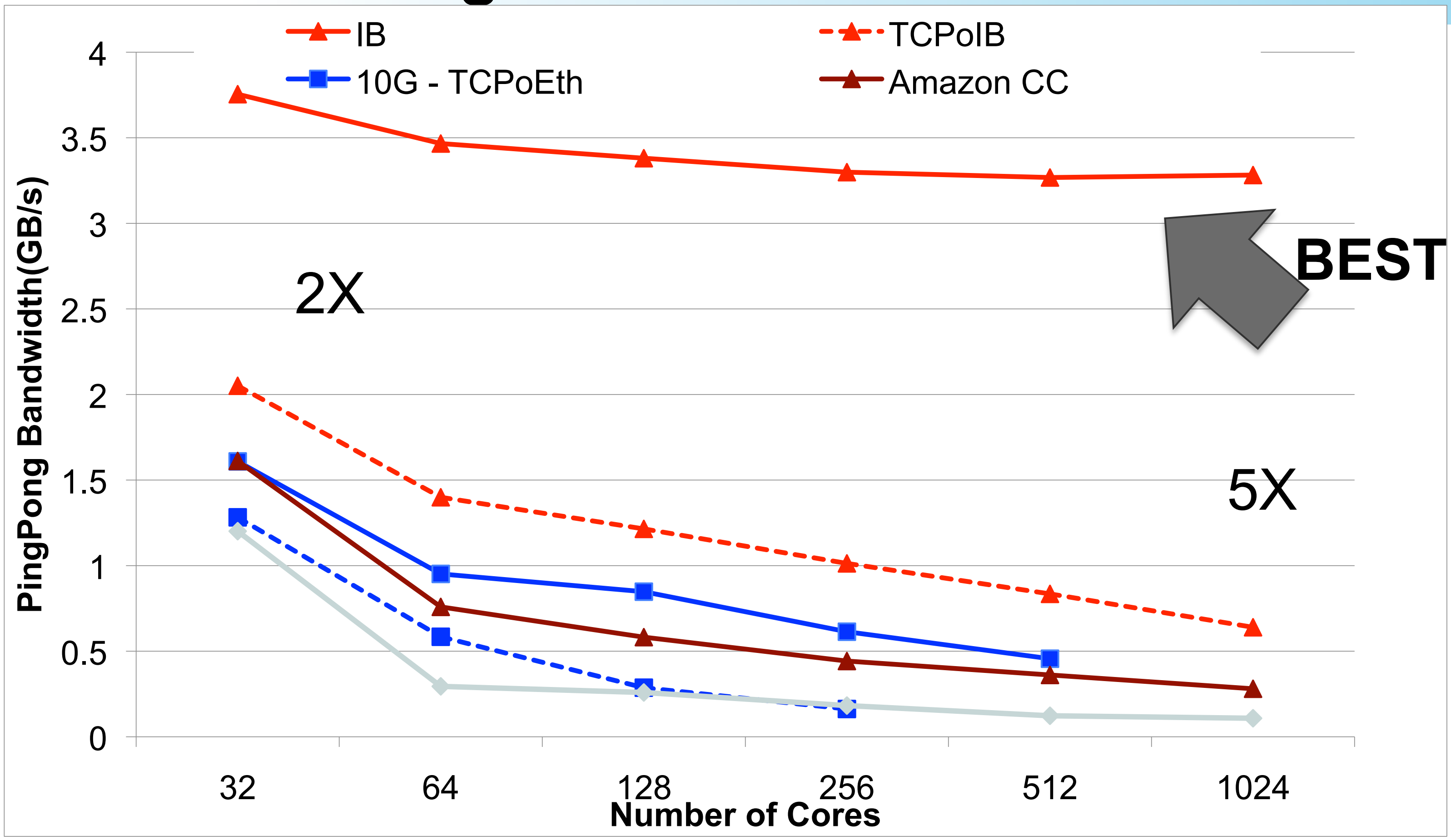


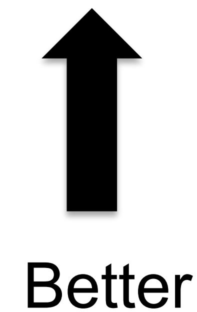
Performance Analysis of High Performance Computing Applications on the Amazon Web Services Cloud, CloudCom 2010

Scientific applications with minimal communication are best suited for clouds



The principle decrease in bandwidth occurs when switching to TCP over IB.

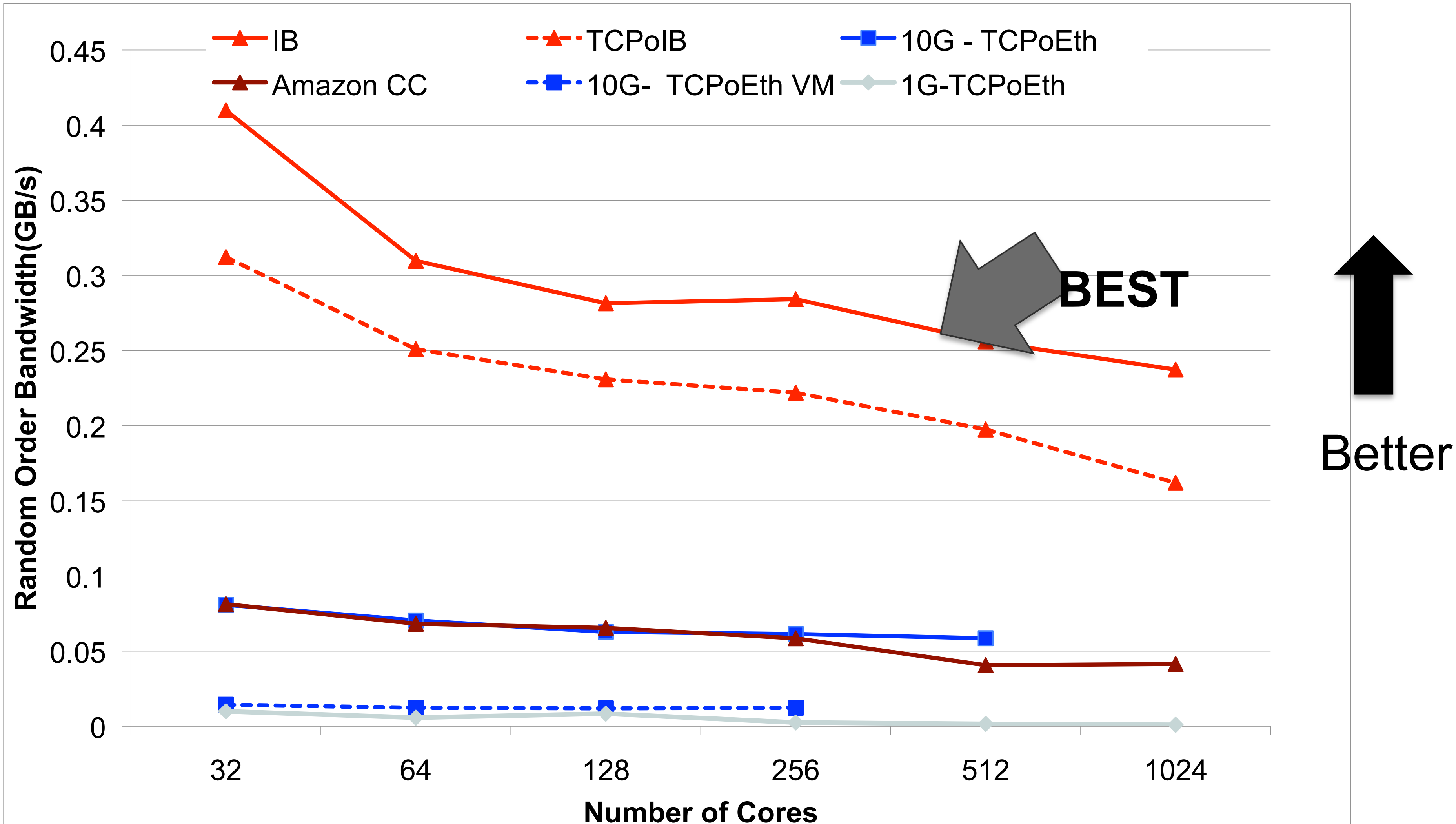




HPCC PingPong BW

Evaluating Interconnect and Virtualization Performance for High Performance Computing, ACM Perf Review 2012

Ethernet connections are unable to cope with significant amounts of network connection

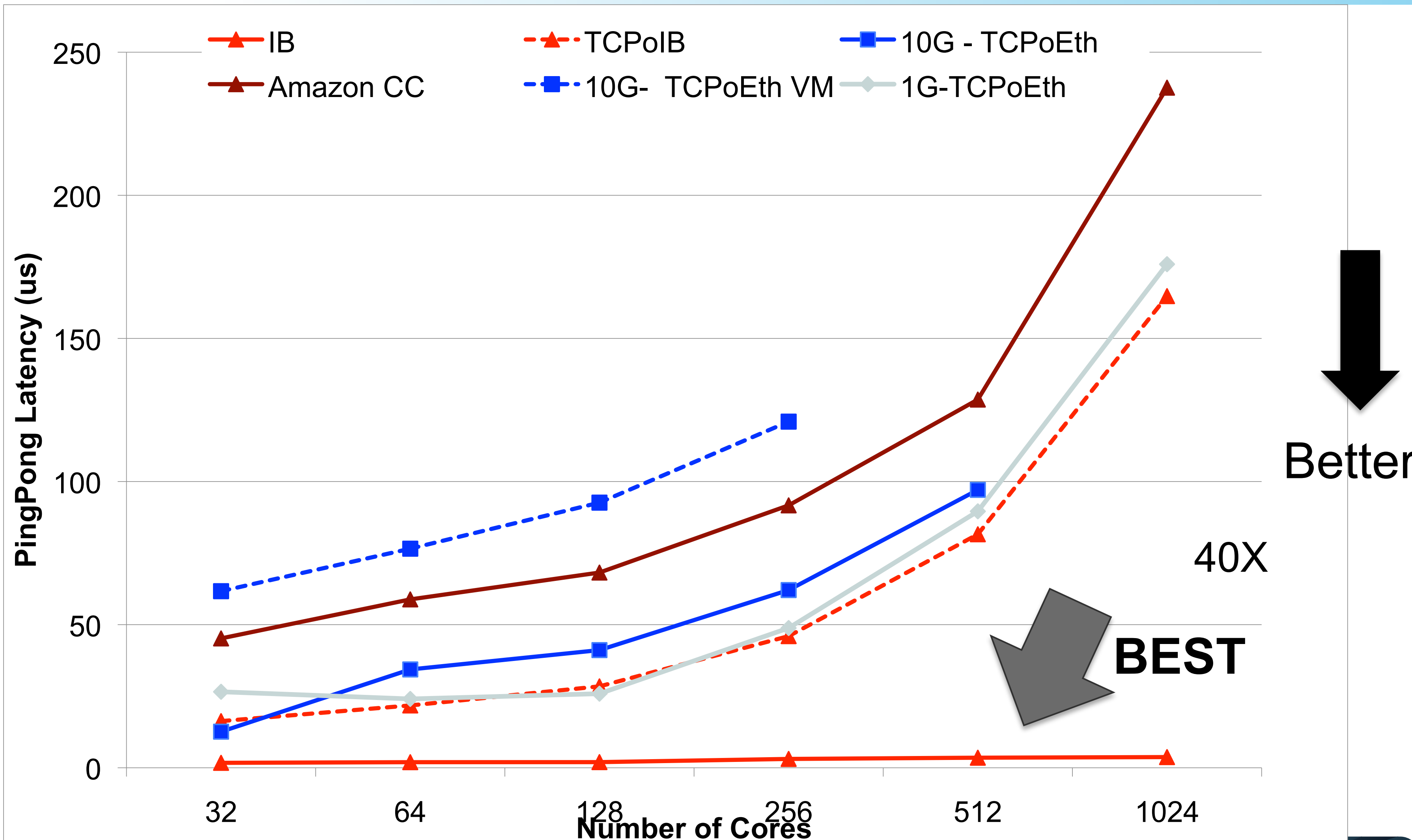


U.S. DEPARTMENT OF ENERGY | Office of Science

CRD computational research division

NERSC

The principle increase in latency occurs for TCP over IB even at mid-range concurrency



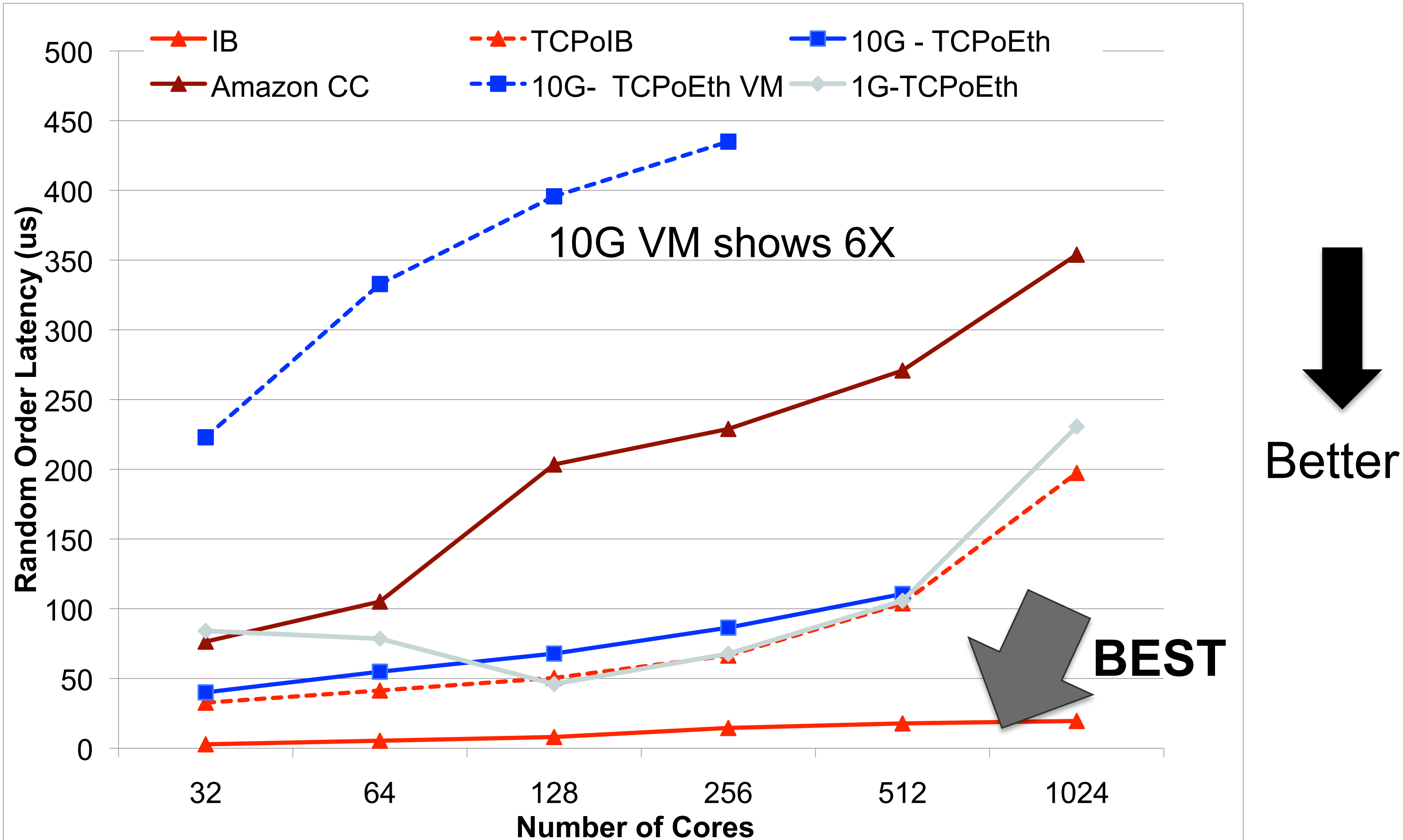
U.S. DEPARTMENT OF ENERGY | Office of Science

HPCC: PingPong Latency

CRD computational research division

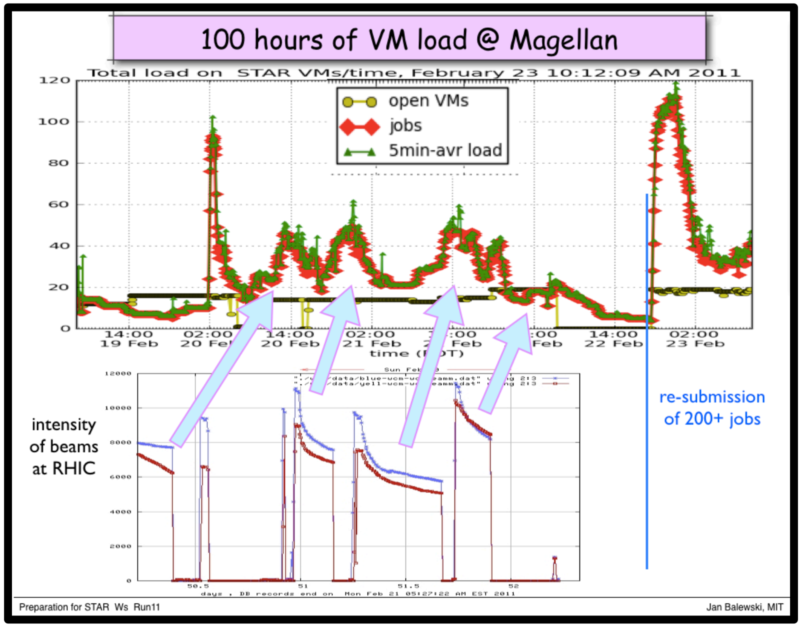
NERSC

Latency is affected by contention by a greater amount than the bandwidth

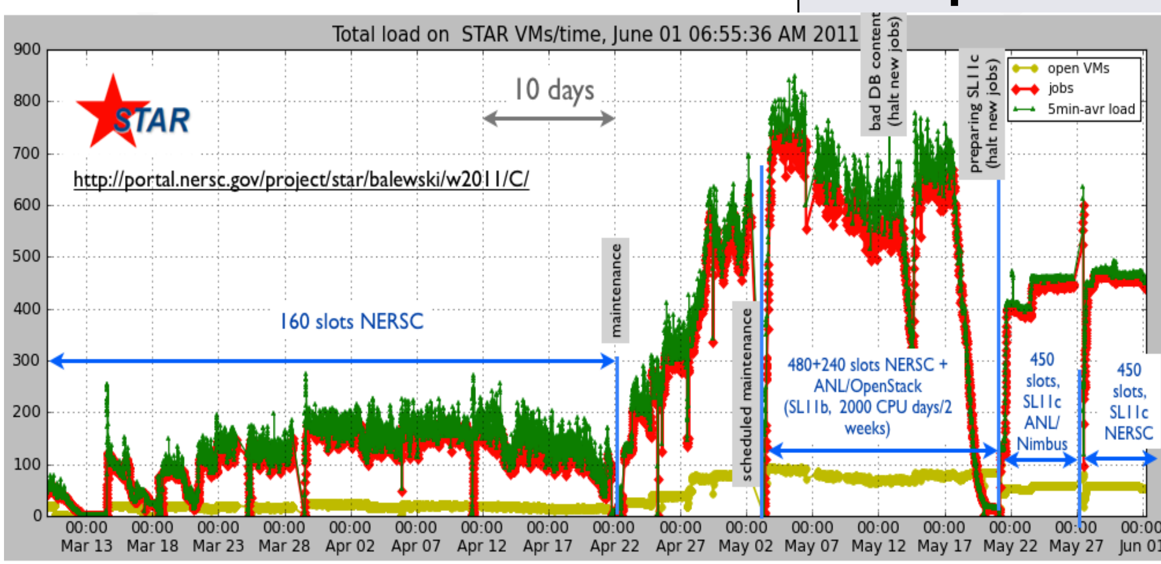


HPCC: RandomRing Latency

Clouds require significant programming and system administration support



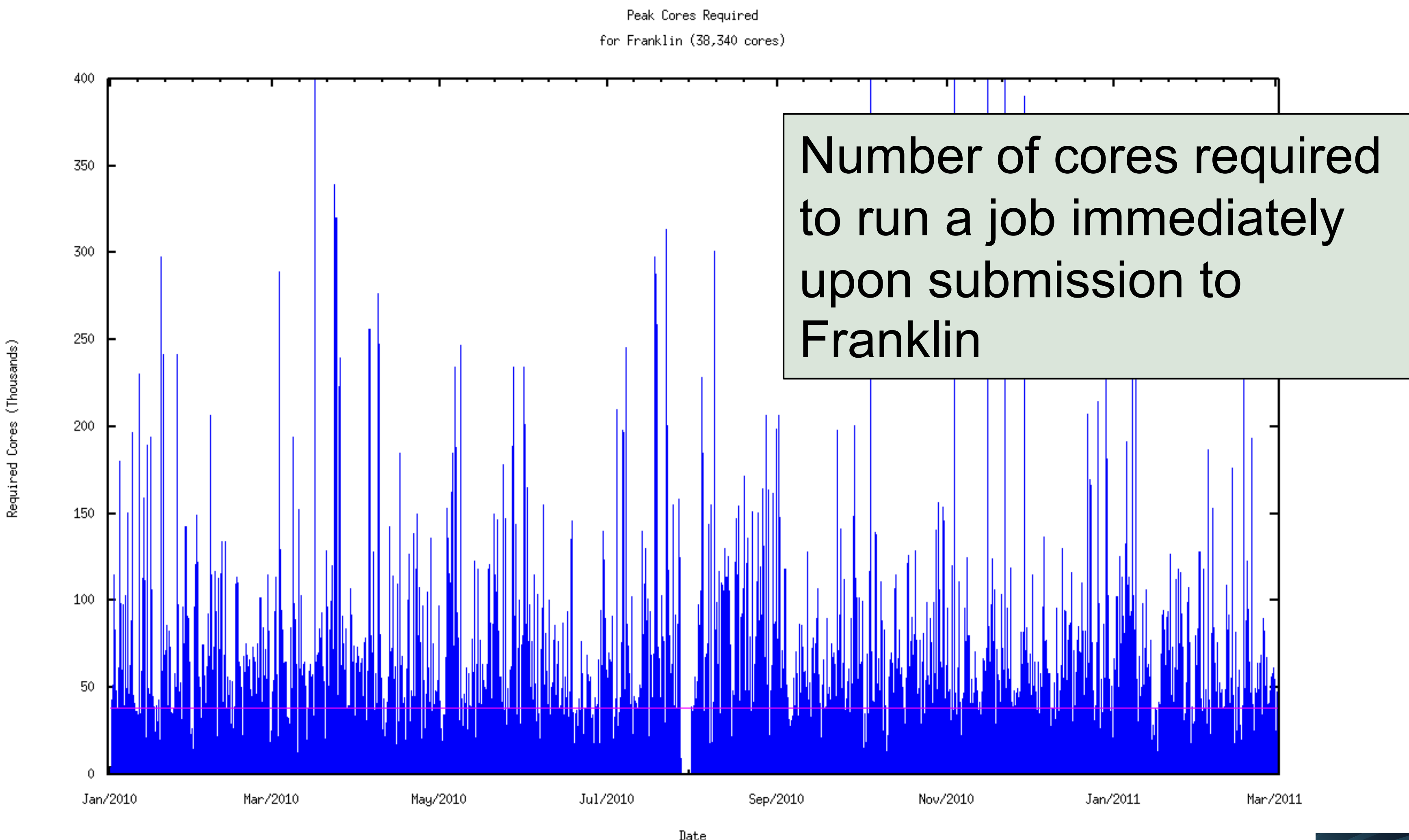
- STAR performed Real-time analysis of data coming from Brookhaven Nat. Lab
- First time data was analyzed in real-time to a high degree
- Leveraged existing OS image from NERSC system
- Started out with 20 VMs at NERSC and expanded to ANL.



CRD
computational research division

NERSC

On-demand access for scientific applications might be difficult if not impossible



Public clouds can be more expensive than in-house large systems

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**

Assumes 85% utilization and zero growth in HPSS and File System data.
Doesn't include the 2x-10x performance impact that has been measured.
This still only captures about 65% of NERSC's $55M annual budget.
No consulting staff, no administration, no support.

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

	Traditional Enterprise IT	HPC Centers
Typical Load Average	30% *	90%
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.

	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
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

Science + Clouds = ?

Business model for Science

Performance and Cost

Data Intensive Science

Technologies from Cloud

MapReduce shows promise but current implementations have gaps for scientific applications

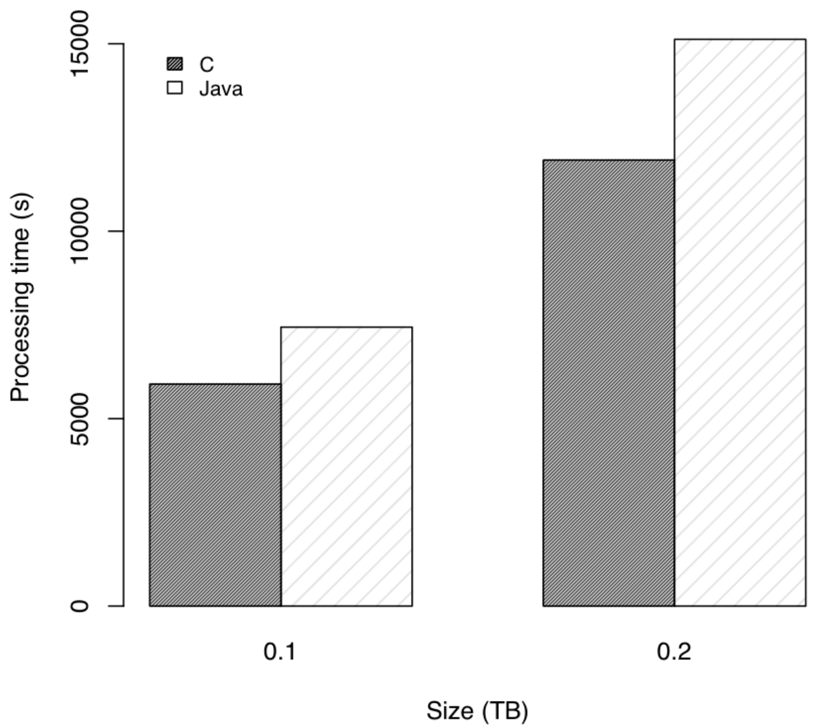
High throughput workflows

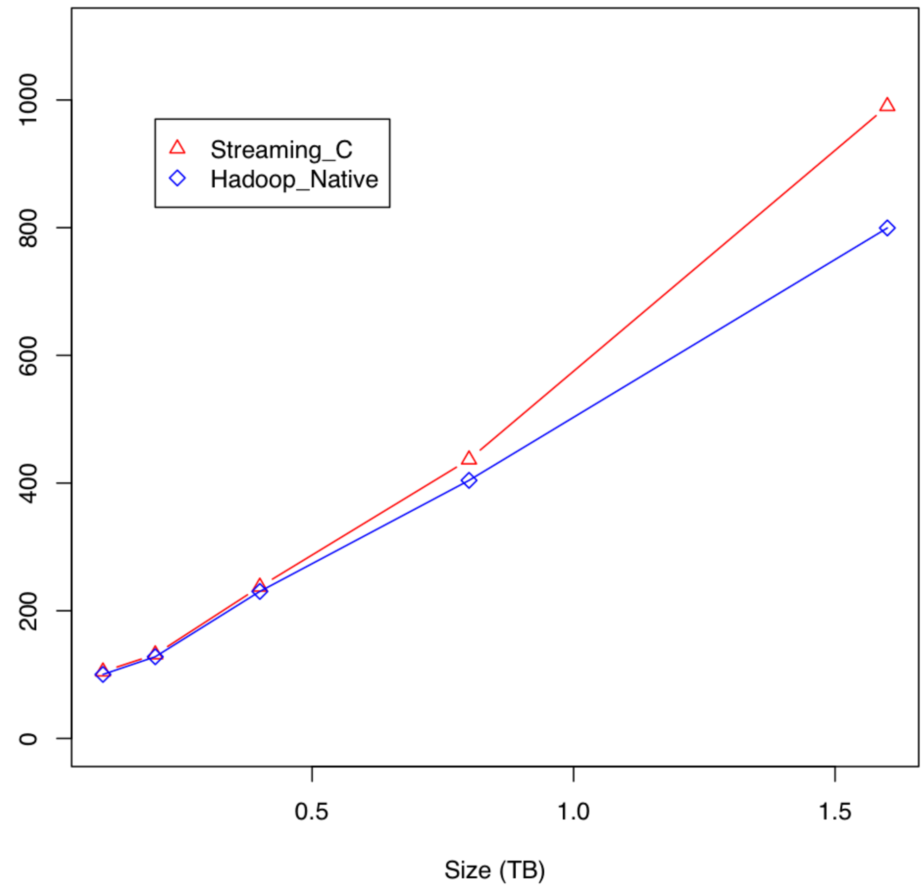
Scaling up from desktops

File system: non POSIX
Language: Java
Input and output formats: mostly line-oriented text
Streaming mode: restrictive i/p and o/p model
Data locality: what happens when multiple inputs?

Streaming adds a performance overhead

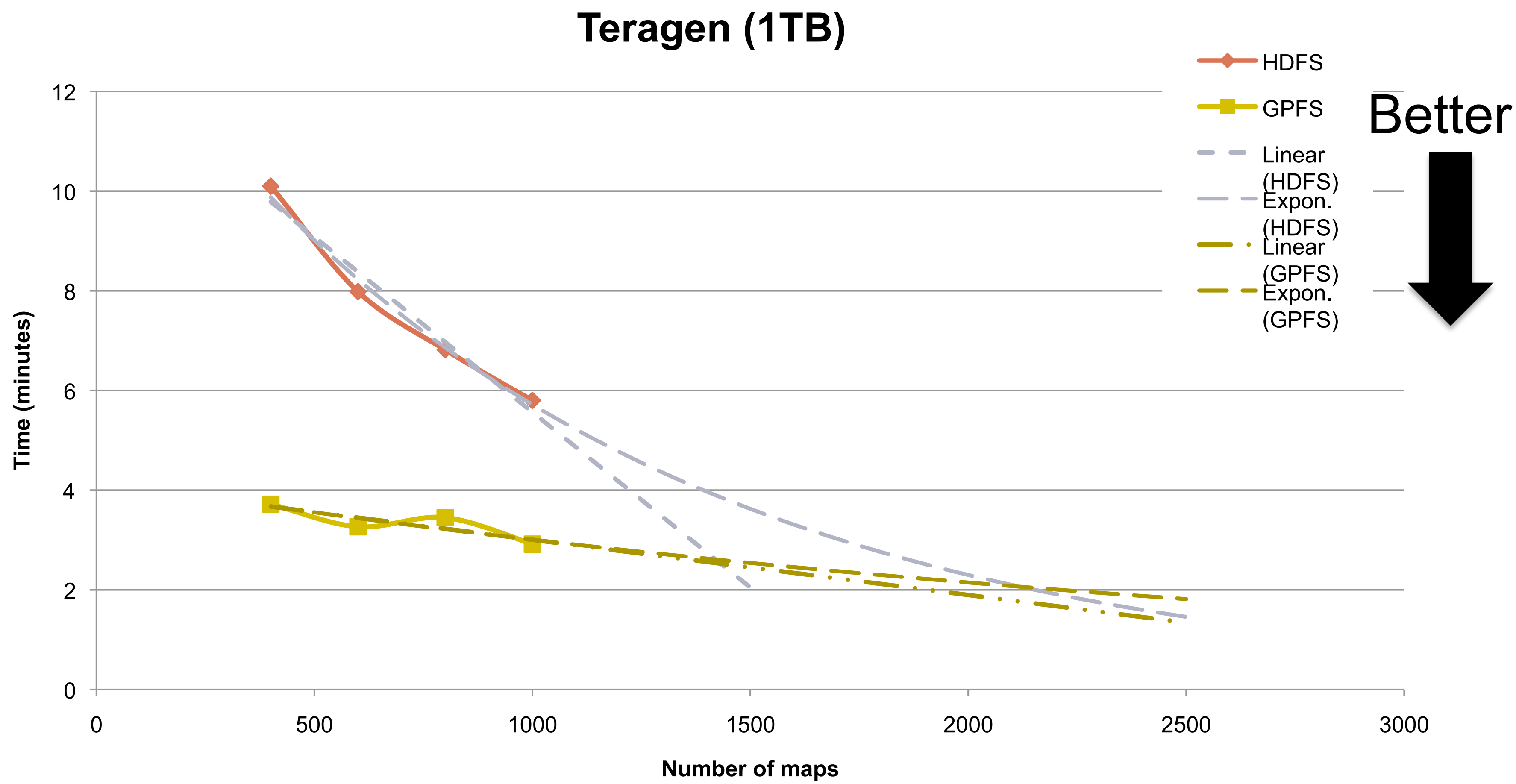
Better



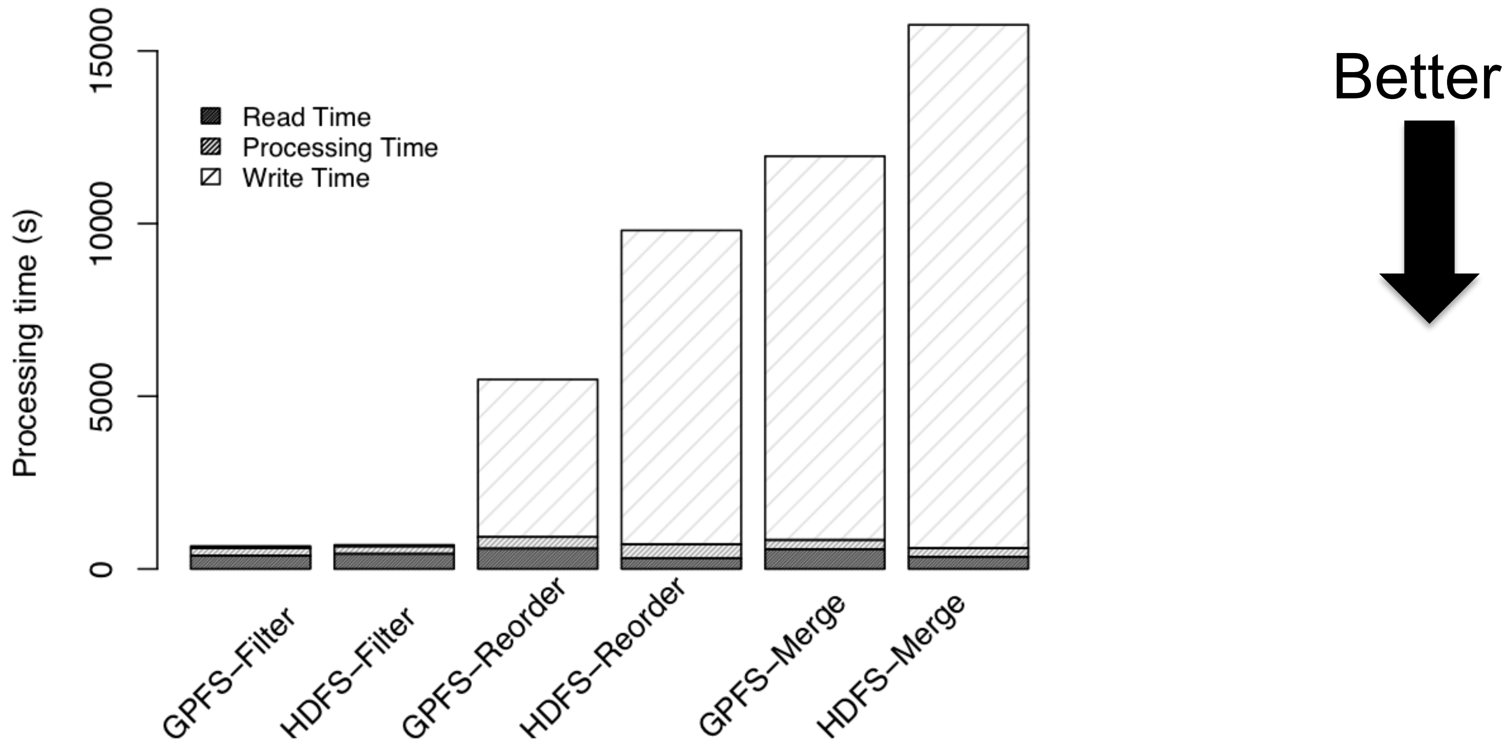


Evaluating Hadoop for Science, In submission

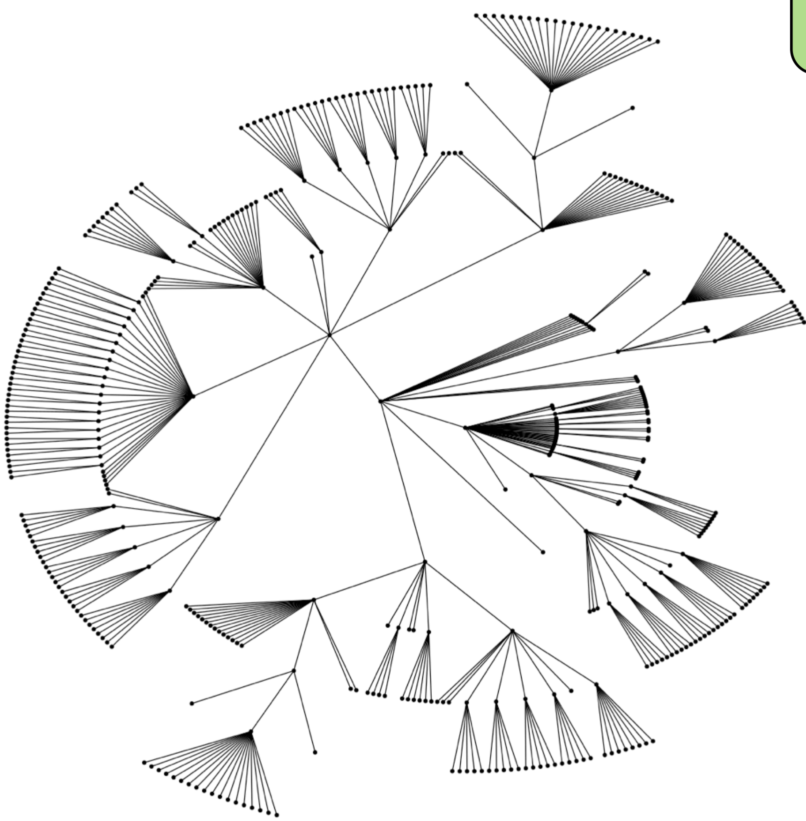
High performance file systems can be used with MapReduce at lower concurrency

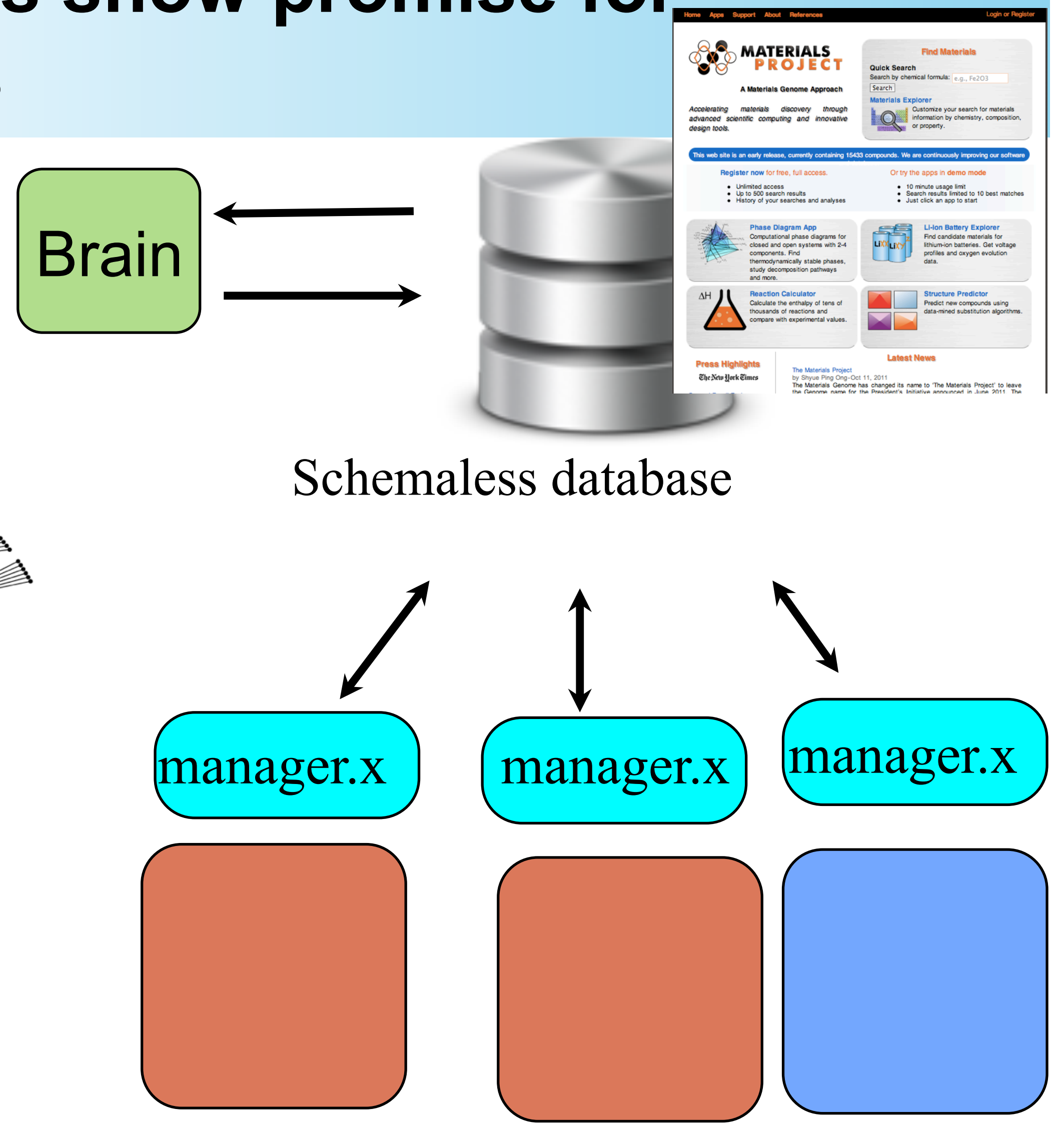


Data operations impacts the performance differences



Schemaless databases show promise for scientific applications

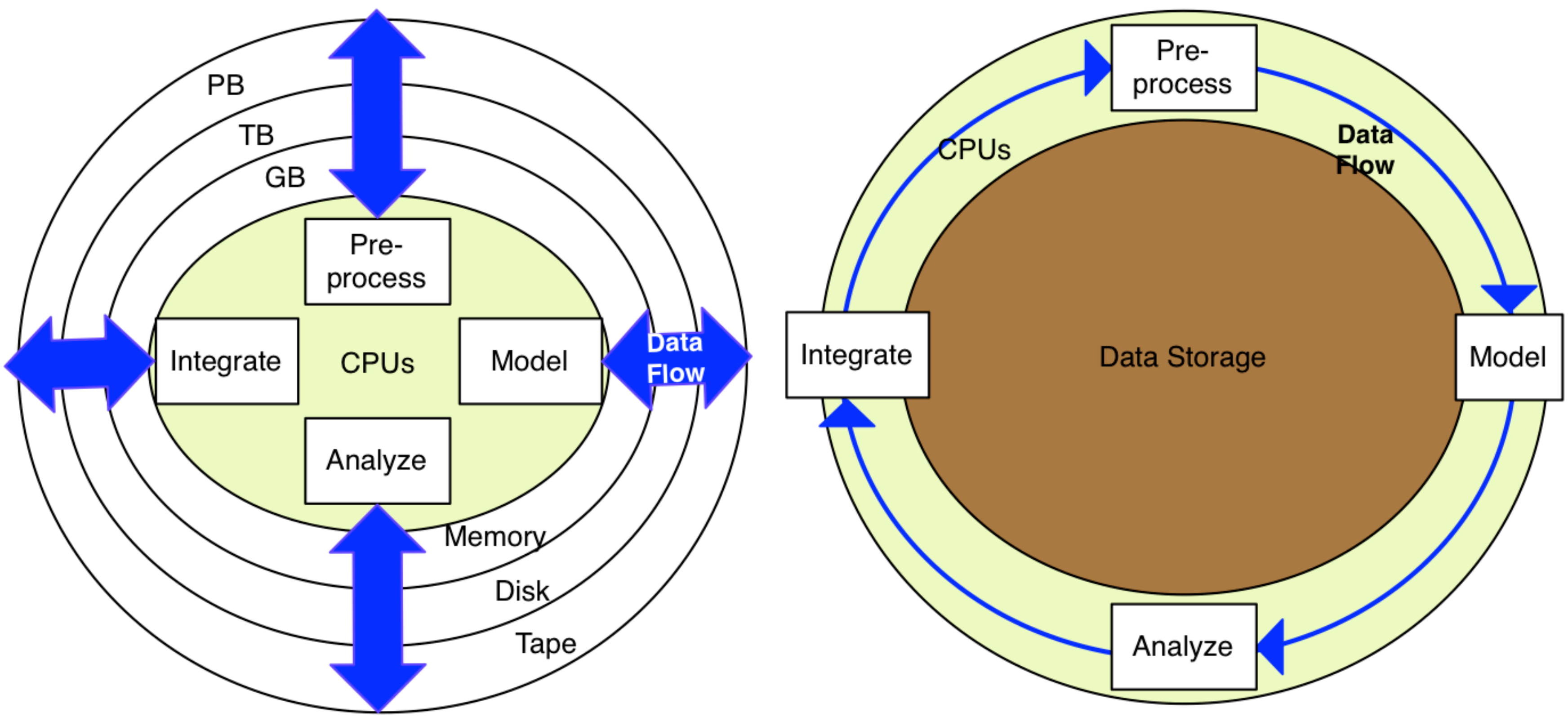




www.materialsproject.org

Source: Michael Kocher, Daniel Gunter

Data centric infrastructure will need to evolve to handle large scientific data volumes



Joint Genome Institute, Advance Light Source, etc are all facing a data tsunami

Cloud is a business model and can be applied at DOE supercomputing centers

- **Current day cloud computing solutions have gaps for science**
  - performance, reliability, stability
  - programming models are difficult for legacy apps
  - security mechanisms and policies
- **HPC centers can adopt some of the technologies and mechanisms**
  - support for data-intensive workloads
  - allow custom software environments
  - provide different levels of service

Acknowledgements

- US Department of Energy DE-AC02-05CH11232
- Magellan
  - Shane Canon, Tina Declerck, Iwona Sakrejda, Scott Campbell, Brent Draney
- Amazon Benchmarking
  - Krishna Muriki, Nick Wright, John Shalf, Keith Jackson, Harvey Wasserman, Shreyas Cholia
- Magellan/ANL
  - Susan Coghlan, Piotr T Zbiegiel, Narayan Desai, Rick Bradshaw, Anping Liu
- NERSC
  - Jeff Broughton, Kathy Yelick
- Applications
  - Jared Wilkening, Gabe West, Ed Holohan, Doug Olson, Jan Balewski, STAR collaboration, K. John Wu, Alex Sim, Prabhat, Suren Byna, Victor Markowitz