Scientific Workflows and Cloud Computing

Gideon Juve
Ewa Deelman

University of Southern California
Information Sciences Institute

This work is funded by NSF

Computational challenges faced by science applications

- Be able to compose complex applications from smaller components
- Execute the computations reliably and efficiently
- Take advantage of any number/types of resources
- Cost is an issue
  - Cluster, Cyberinfrastructure, Cloud

Possible solution

somewhat subjective

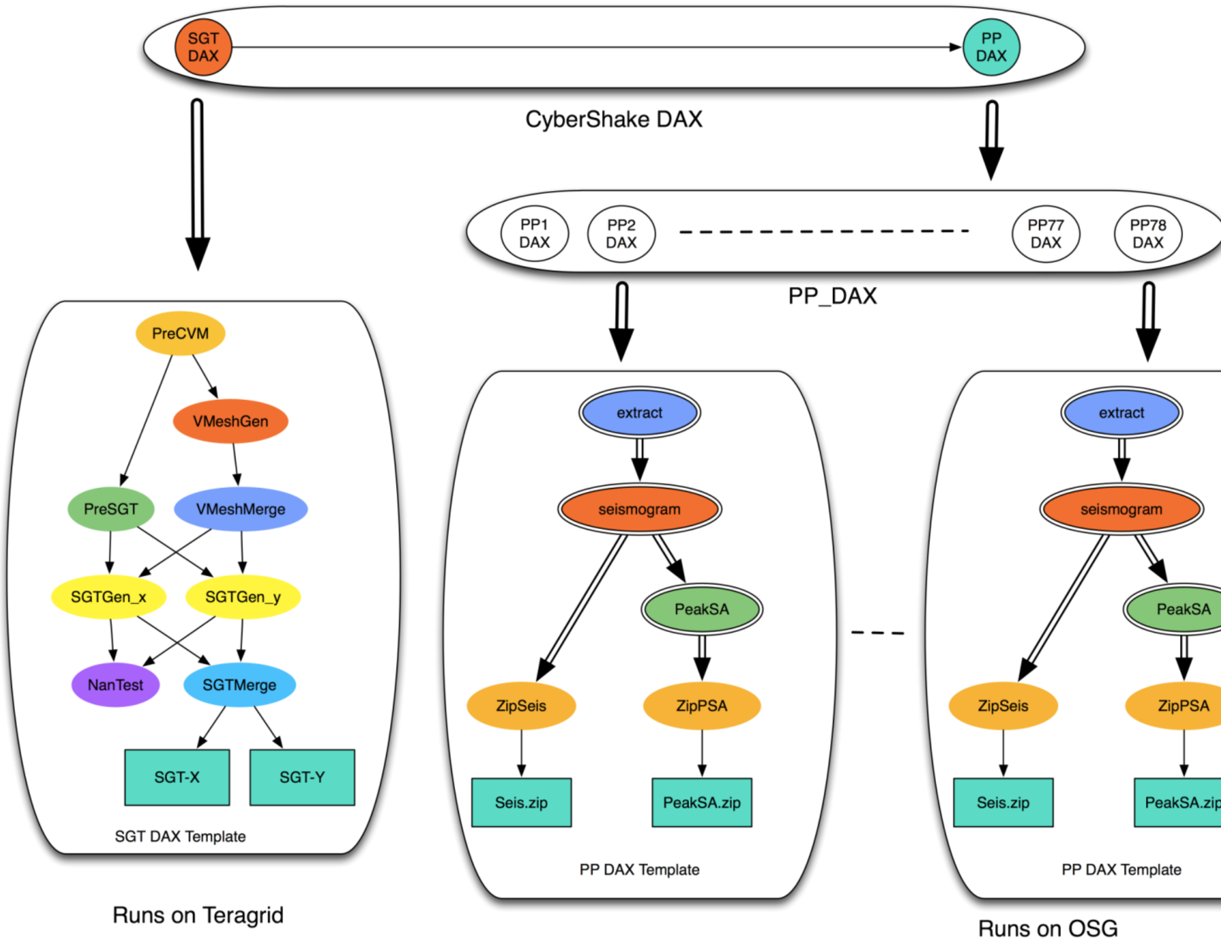
- Structure an application as a workflow (task graph)
  - Describe data and components in logical terms (resource independent)
  - Use a Workflow Management System to map it onto a number of execution environments
  - Optimize it and repair if faults occur--the WMS can recover
  - Use a WMS (Pegasus-WMS) to manage the application on a number of resources

Pegasus-Workflow Management System (est. 2001)

- Leverages abstraction for workflow description to obtain ease of use, scalability, and portability
- Provides a compiler to map from high-level descriptions to executable workflows
  - Correct mapping
  - Performance enhanced mapping
- Provides a runtime engine to carry out the instructions (Condor DAGMan)
  - Scalable manner
  - Reliable manner
- Can execute on a number of resources: local machine, campus cluster, Grid, Cloud

Ewa Deelman, deelman@isi.edu

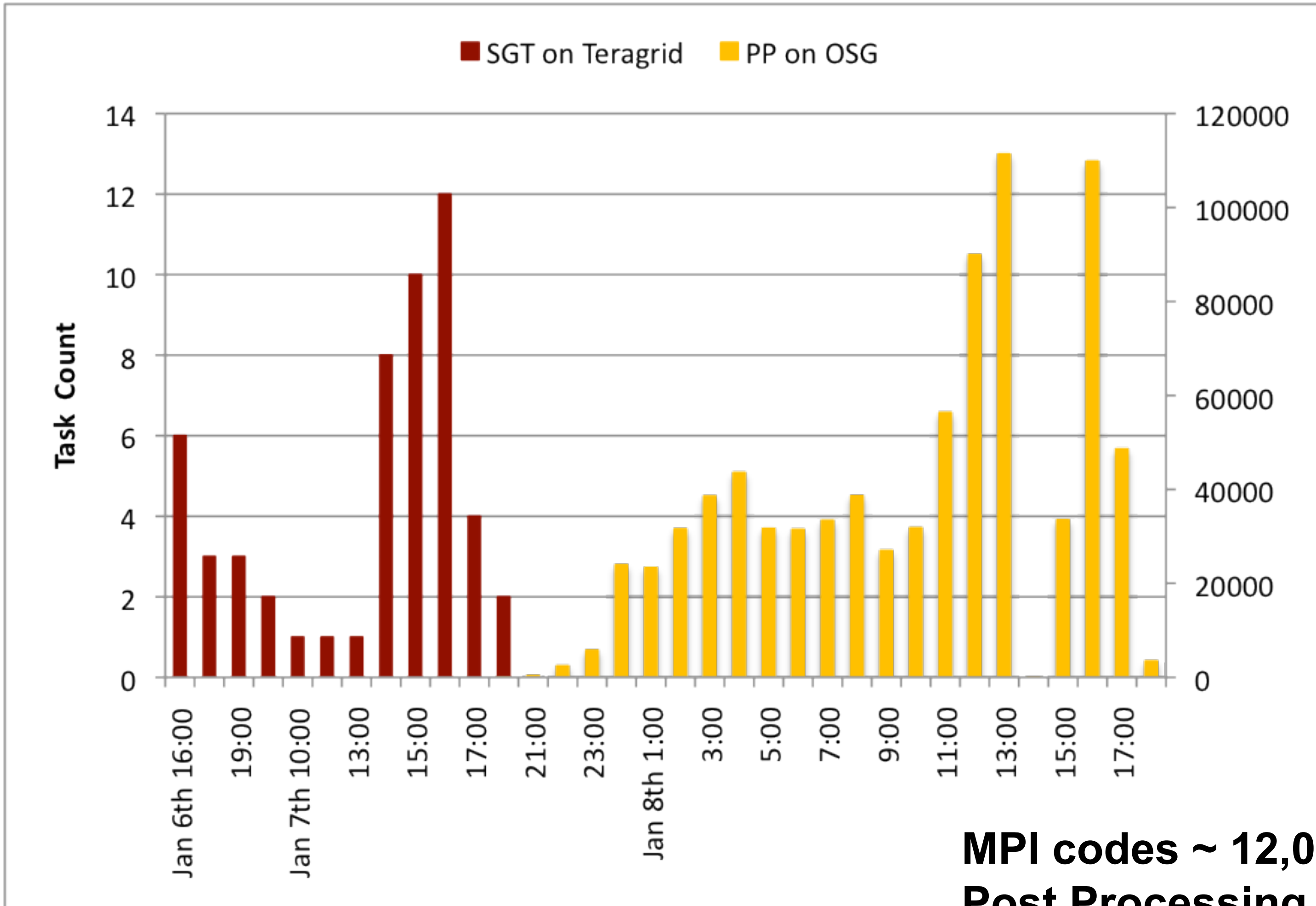
So far applications have been running on local/campus clusters or grids



SCEC CyberShake

- Uses physics-based approach
  - 3-D ground motion simulation with anelastic wave propagation
  - Considers ~415,000 earthquakes per site
    - <200 km from site of interest
    - Magnitude >6.5

Applications can leverage different Grids: SCEC across the TeraGrid and OSG with Pegasus



SoCal Map needs 239 of those

MPI codes ~ 12,000 CPU hours,
Post Processing 2,000 CPU hours
Data footprint ~ 800GB

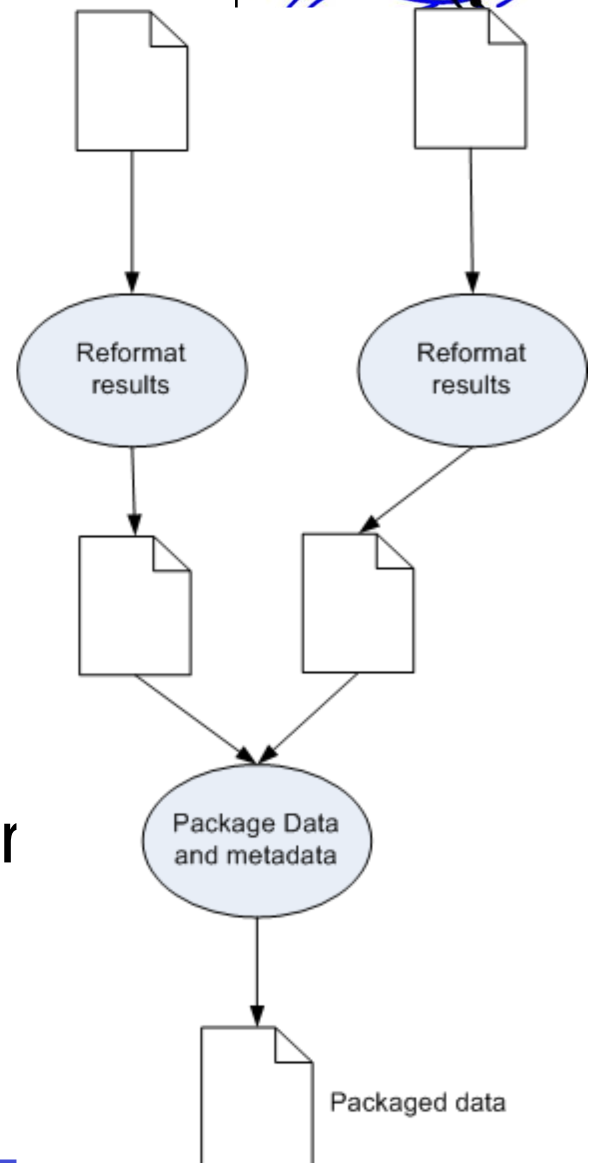
Peak # of cores on OSG 1,600
Walltime on OSG 20 hours, could be done in 4 hours on 800 cores

Some applications want science done "now"

- Looking towards the Cloud—they like the ability to provision computing and storage
- They don't know how to best leverage the infrastructure, how to configure it
- They often don't want to modify the application codes
- They are concerned about costs

One approach: Build Virtual Cluster on the Cloud

- Clouds provide resources, but the software is up to the user
- Running on multiple nodes may require cluster services (e.g. scheduler)
- Dynamically configuring such systems is not trivial
- Some tools are available (Nimbus Context Broker– now Amazon cluster with mapreduce)
- Workflows need to communicate data—often through files

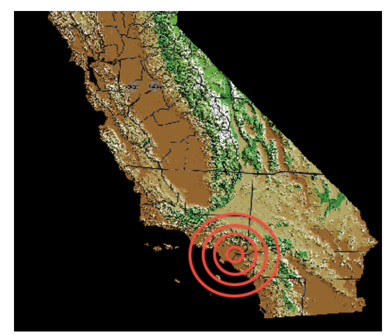


Experiments

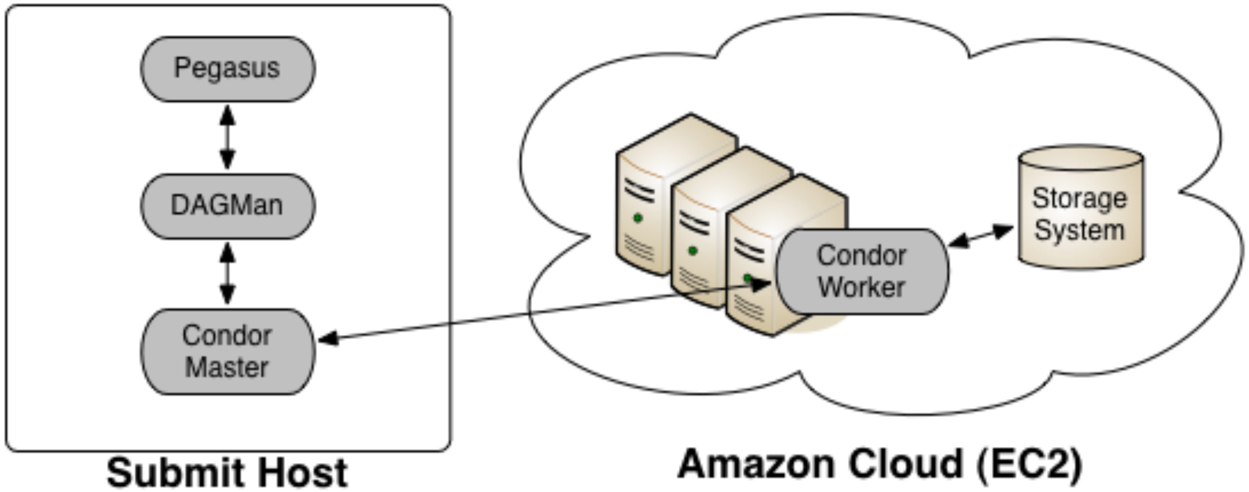
- Goal: Evaluate different file systems for VC
- Take a few applications with different characteristics
  - Evaluate them on a Cloud—single virtual instance (Amazon)
  - Compare the performance to that of a TG cluster
- Take a few well-known file systems, deploy on a virtual cluster
  - Compare their performance
- Quantify monetary costs

Applications

- Not CyberShake SoCal map (PP) could cost at least $60K for computing and $29K for data storage (for a month) on Amazon (one workflow ~$300)
- Montage (astronomy, provided by IPAC)
  - 10,429 tasks, 4.2GB input, 7.9GB of output
  - I/O: High (95% of time waiting on I/O)
  - Memory: Low, CPU: Low
- Epigenome (bioinformatics, USC Genomics Center)
  - 81 tasks 1.8GB input, 300 MB output
  - I/O: Low, Memory: Medium
  - CPU: High (99% time of time)
- Broadband (earthquake science, SCEC)
  - 320 tasks, 6GB of input, 160 MB output
  - I/O: Medium
  - Memory: High (75% of task time requires > 1GB mem)
  - CPU: Medium

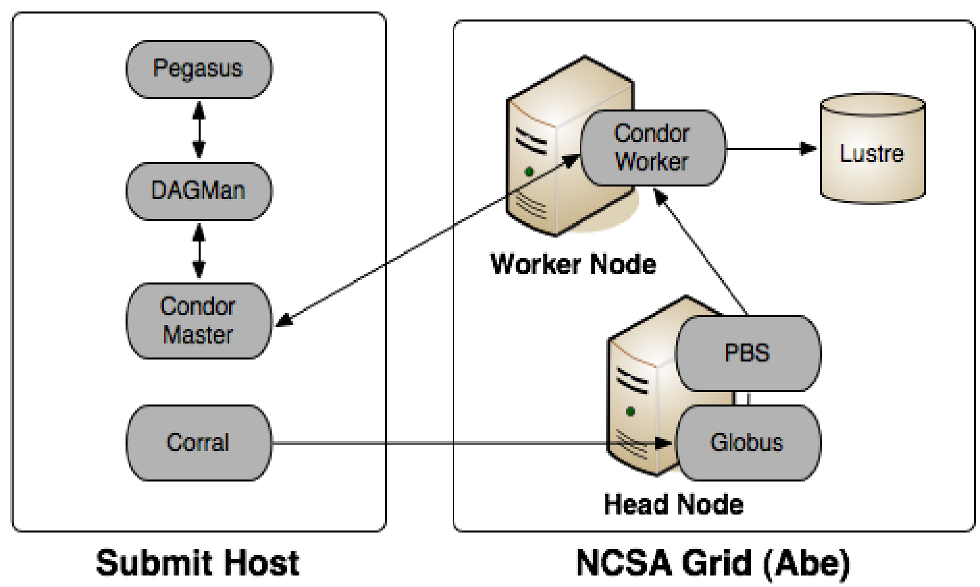


Experimental Setup



Cloud

Grid (TeraGrid)



Resource Type Experiments

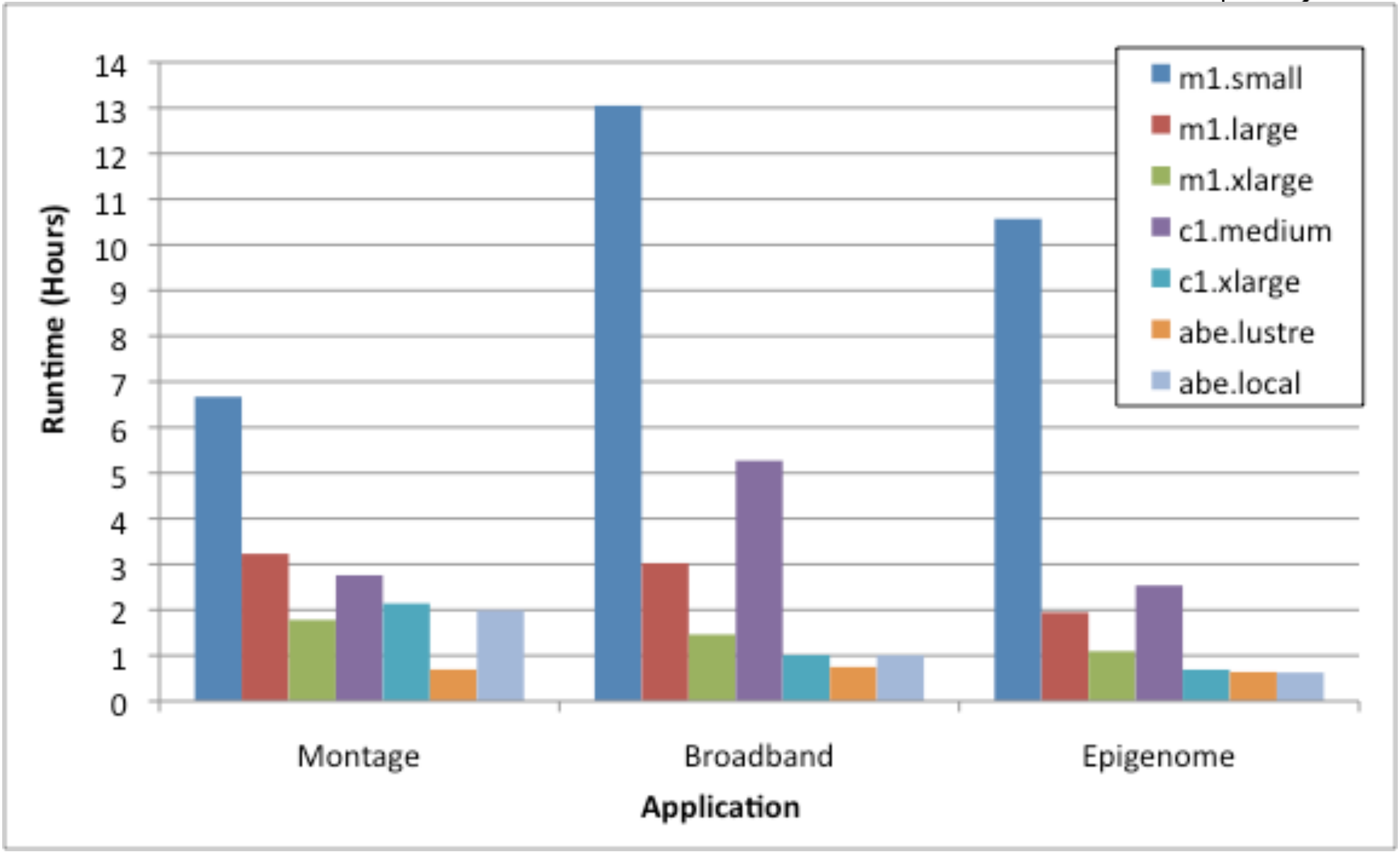
- Resource Types Tested

Type	Arch.	CPU	Cores	Memory	Network	Storage	Price
m1.small	32-bit	2.0-2.6 GHz Opteron	1/2	1.7 GB	1-Gbps Ethernet	Local disk	$0.085/hr
m1.large	64-bit	2.0-2.6 GHz Opteron	2	7.5 GB	1-Gbps Ethernet	Local disk	$0.12/hr
m1.xlarge	64-bit	2.0-2.6 GHz Opteron	4	15 GB	1-Gbps Ethernet	Local disk	$0.68/hr
c1.medium	32-bit	2.33-2.66 GHz Xeon	2	1.7 GB	1-Gbps Ethernet	Local disk	$0.17/hr
c1.xlarge	64-bit	2.33-2.66 GHz Xeon	8	7.5 GB	1-Gbps Ethernet	Local disk	$0.68/hr
abe.local	64-bit	2.33 GHz Xeon	8	8 GB	10-Gbps InfiniBand	Local disk	N/A
abe.lustre	64-bit	2.33 GHz Xeon	8	8 GB	10-Gbps InfiniBand	Lustre	N/A

Amazon S3

- $0.15 per GB-Month for storage resources on S3
- $0.10 per GB for transferring data into its storage system
- $0.15 per GB for transferring data out of its storage system
- $0.01 per 1,000 I/O Requests

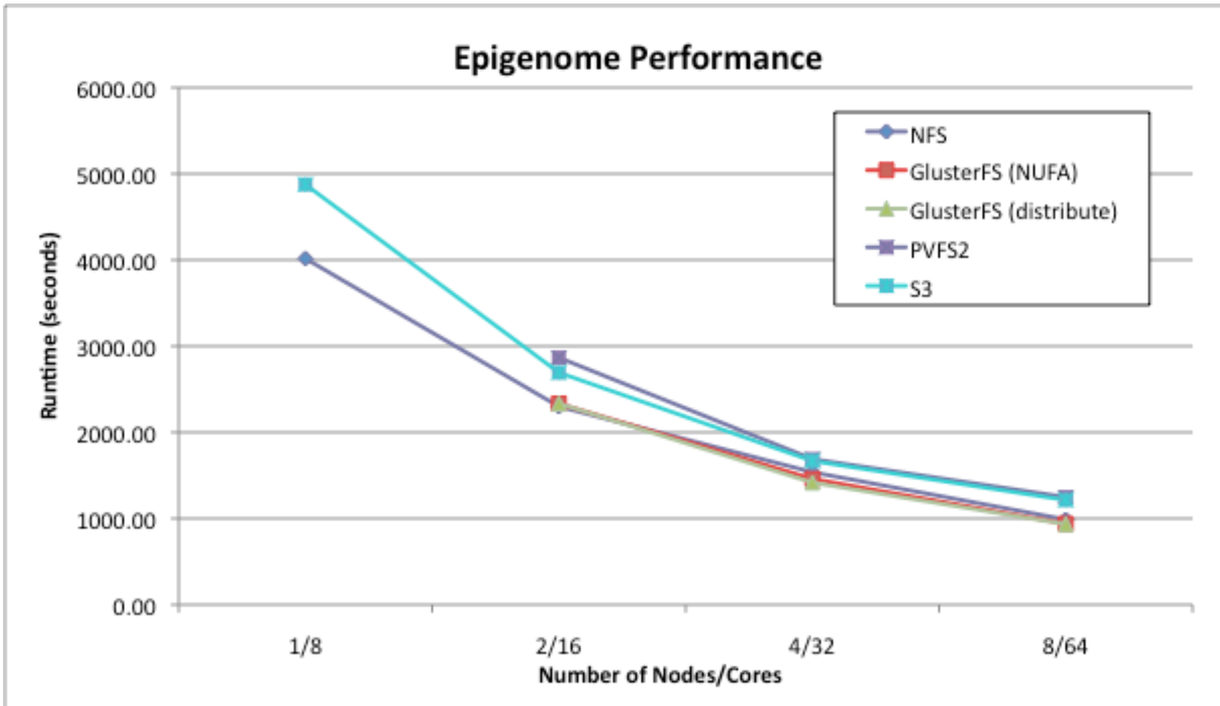
Resource Type Performance, one instance



Storage System Experiments

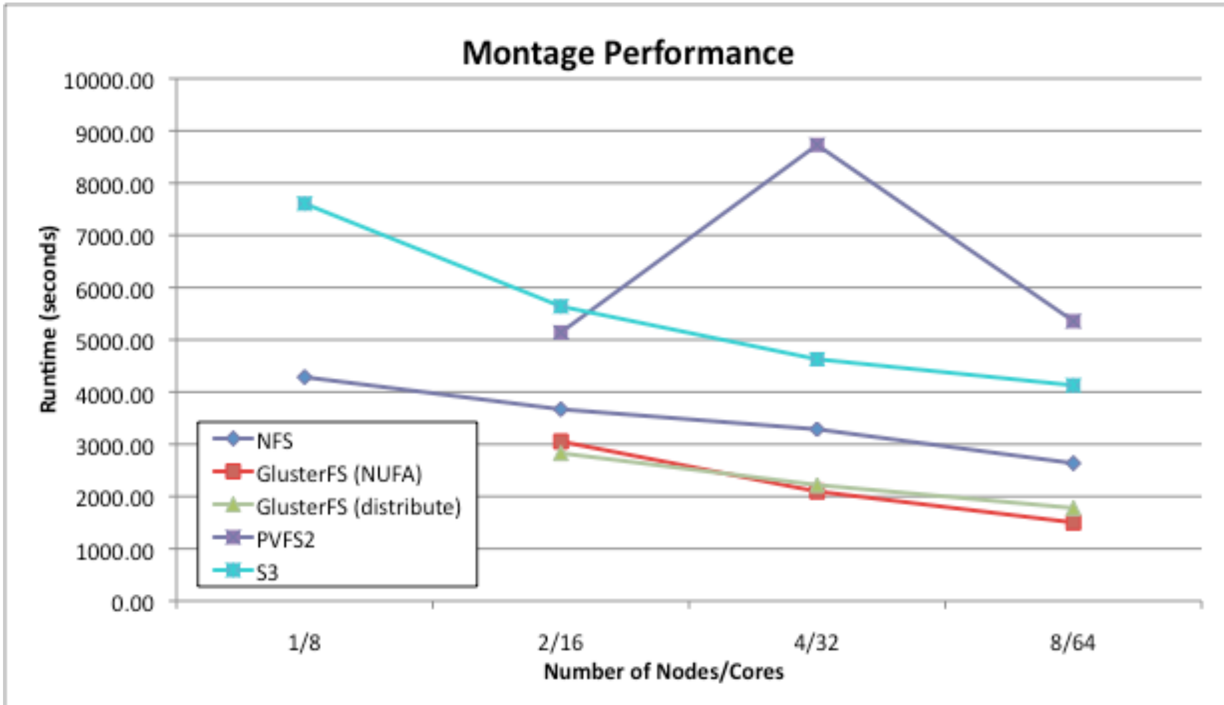
- Investigate different options for storing intermediate data
- Storage Systems
  - Local Disk
  - NFS: Network file system
  - PVFS: Parallel, striped cluster file system
  - GlusterFS: Distributed file system
  - Amazon S3: Object-based storage system
- Amazon Issues
  - Some systems don’t work on EC2 (Lustre, Ceph, etc.)

Storage System Performance

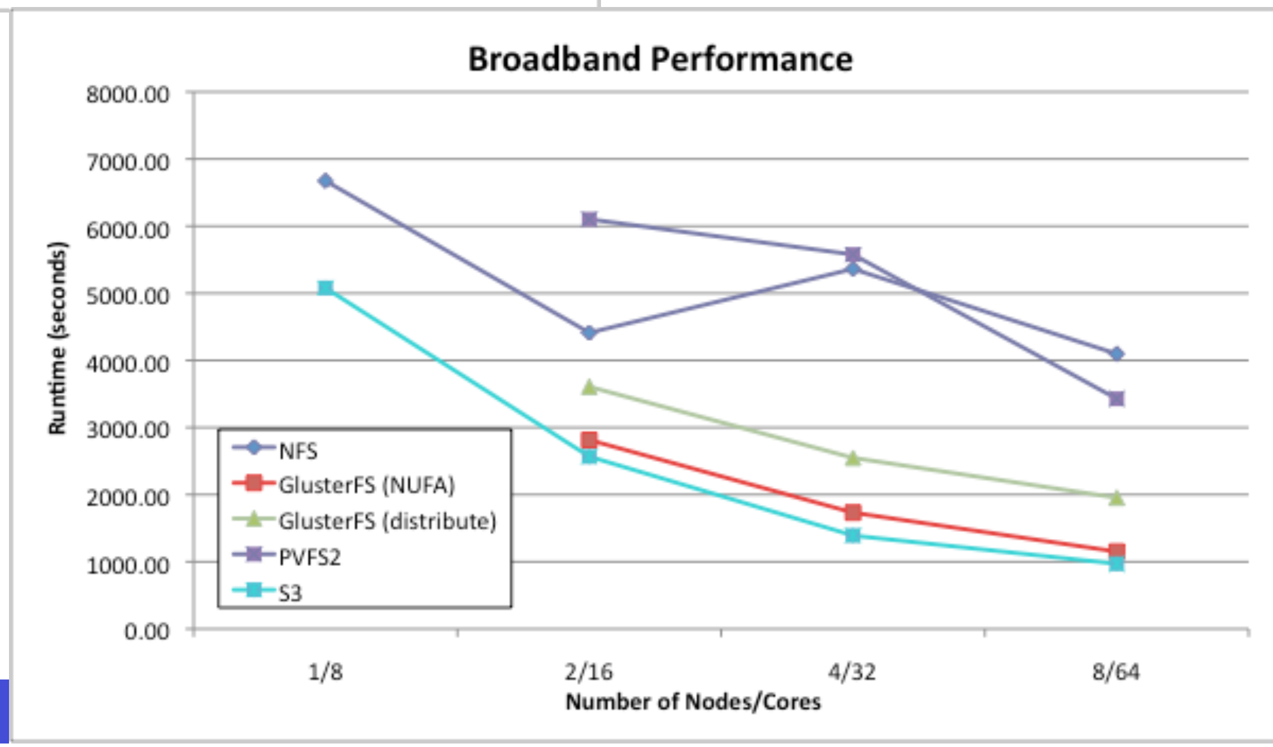


- NFS uses an extra node
- PVFS, GlusterFS use workers to store data, S3 does not
- PVFS, GlusterFS use 2 or more nodes
- We implemented whole file caching for S3

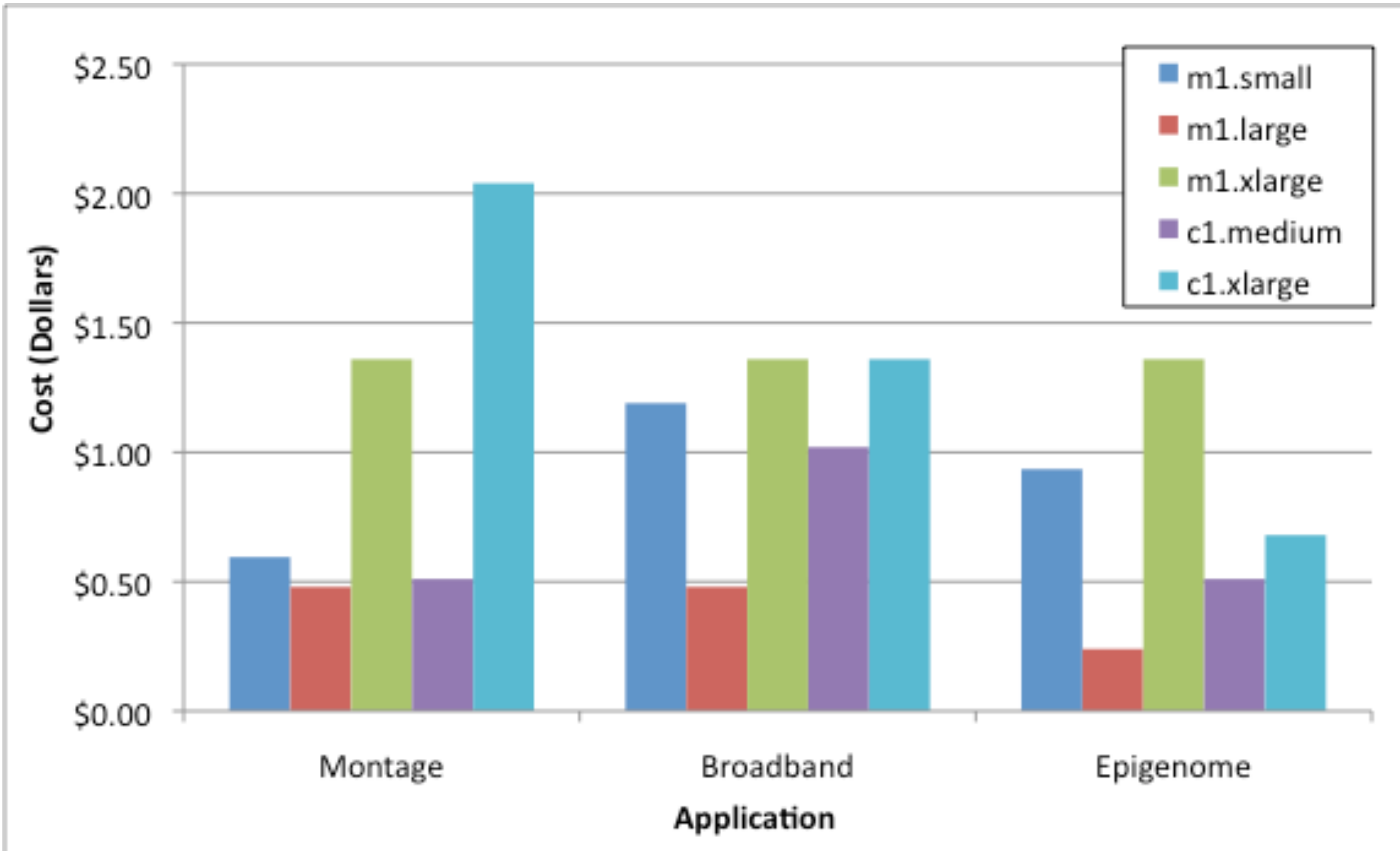
Lots of small files



Re-reading the same file

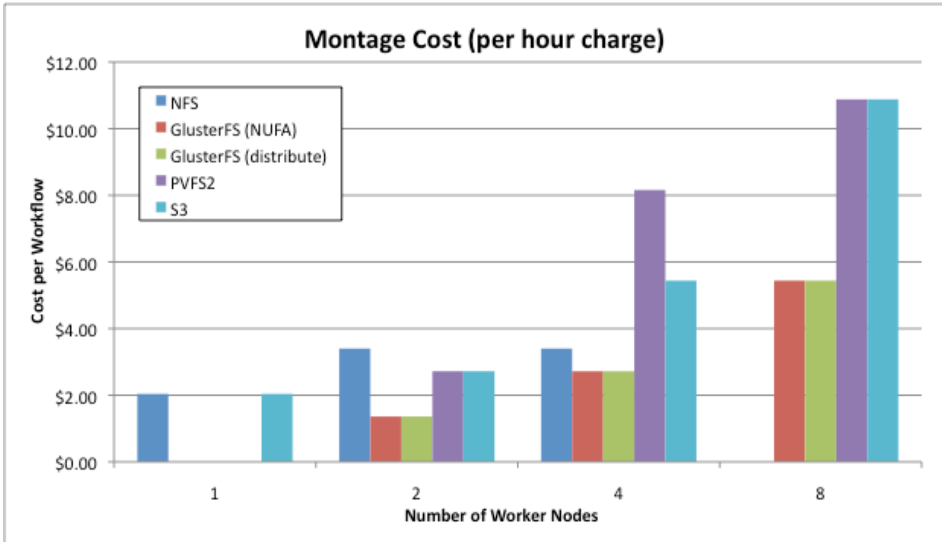


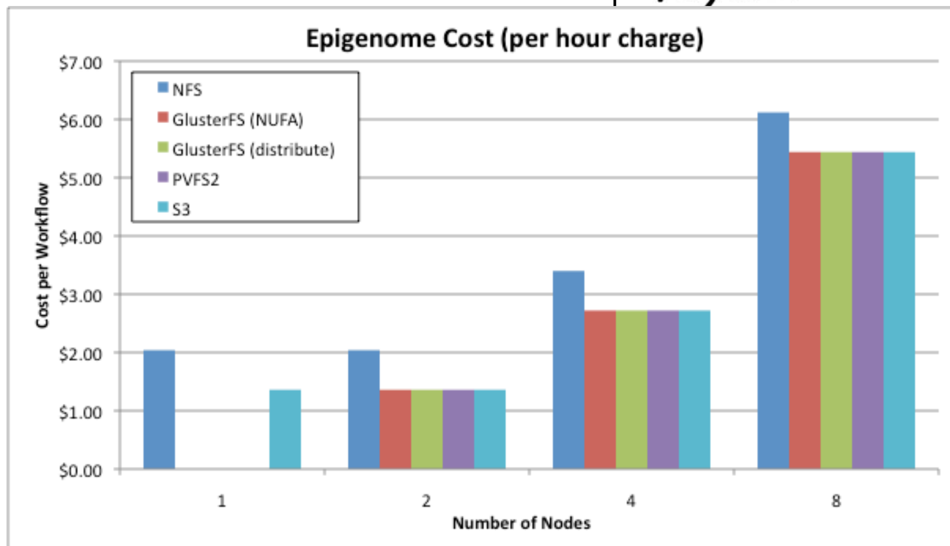
Resource Cost (by Resource Type)



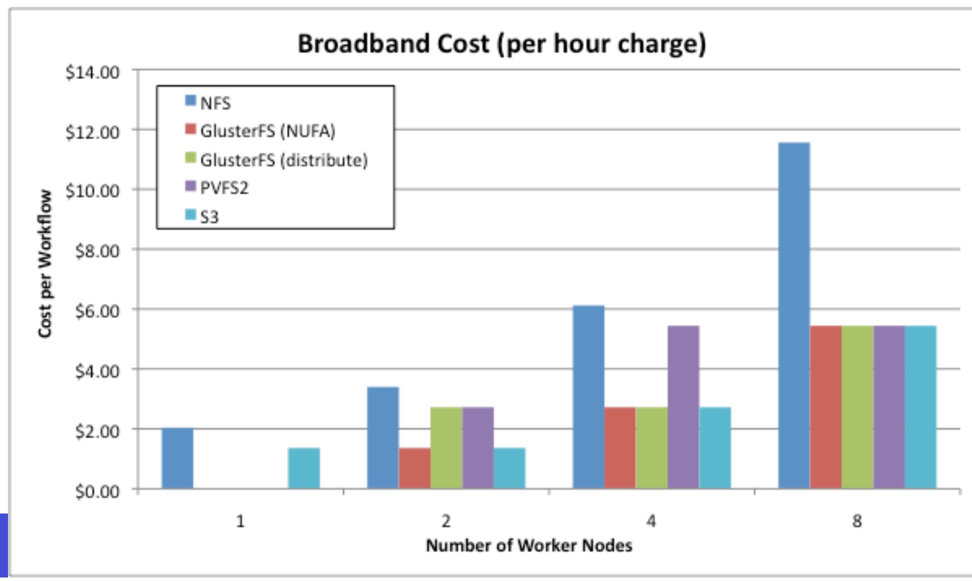
Important: Amazon charges per hour

Resource Cost
(by Storage System)





- Cost tracks performance
- Price not unreasonable
- Adding resources does not usually reduce cost



Transfer and Storage Costs

Application	Input	Output	Logs
Montage	4291 MB	7970 MB	40 MB
Broadband	4109 MB	159 MB	5.5 MB
Epigenome	1843 MB	299 MB	3.3 MB

Transfer Sizes

Application	Input	Output	Logs	Total
Montage	$0.42	$1.32	< $0.01	$1.75
Broadband	$0.40	$0.03	< $0.01	$0.43
Epigenome	$0.18	$0.05	< $0.01	$0.23

Transfer Costs

- Transfer costs are a relatively large fraction of total cost
- Costs can be reduced by storing input data in the cloud and using it for multiple runs

Input data stored in EBS

Application	Volume Size	Monthly Cost
Montage	5GB	$0.66
Broadband	5GB	$0.60
Epigenome	2GB	$0.26

VMs stored in S3

Image	Size	Monthly Cost
32-bit	773 MB	$0.11
64-bit	729 MB	$0.11

Summary

- Commercial clouds are usually a reasonable alternative to grids for a number of workflow applications
  - Performance is good
  - Costs are OK for small workflows
  - Data transfer can be costly
  - Storage costs can become high over time
- Clouds require additional configurations to get desired performance
  - In our experiments GlusterFS did well overall
- Need tools to help evaluate costs for entire computational problems, not just one workflows
- Need tools to help manage the costs
  - Or use science clouds like FutureGrid

Acknowledgements

- SCEC: Scott Callaghan, Phil Maechling, Tom Jordan, and others (USC)
- Montage: Bruce Berriman and John Good (Caltech)
- Epigenomics: Ben Berman (USC Epigenomic Center)
- Corral: Gideon Juve, Mats Rynge (USC/ISI)
- Pegasus: Gaurang Mehta, Karan Vahi (USC /ISI)