eScience:
Grids, Clouds and Parallel Computing

Challenges in eScience Workshop http://dexl.lncc.br/CIS/

Vale Real Hotel - Itaipava - Petrópolis RJ,

Brazil

October 29 2010

Geoffrey Fox

gcf@indiana.edu

http://www.infomall.org   http://www.futuregrid.org

Director, Digital Science Center, Pervasive Technology Institute
Associate Dean for Research and Graduate Studies, School of Informatics and Computing
Indiana University Bloomington

eScience:
Grids, Clouds and Parallel Computing

- We analyze the different tradeoffs and goals of Grid, Cloud and parallel (cluster/supercomputer) computing.
- They tradeoff performance, fault tolerance, ease of use (elasticity), cost, interoperability.
- Different application classes (characteristics) fit different architectures and we describe a hybrid model with Grids for data, traditional supercomputers for large scale simulations and clouds for broad based "capacity computing" including many data intensive problems.
- We discuss the impressive features of cloud computing platforms and compare MapReduce and MPI.
- We take most of our examples from the life science area.
- We conclude with a description of FutureGrid -- a TeraGrid system for prototyping new middleware and applications.

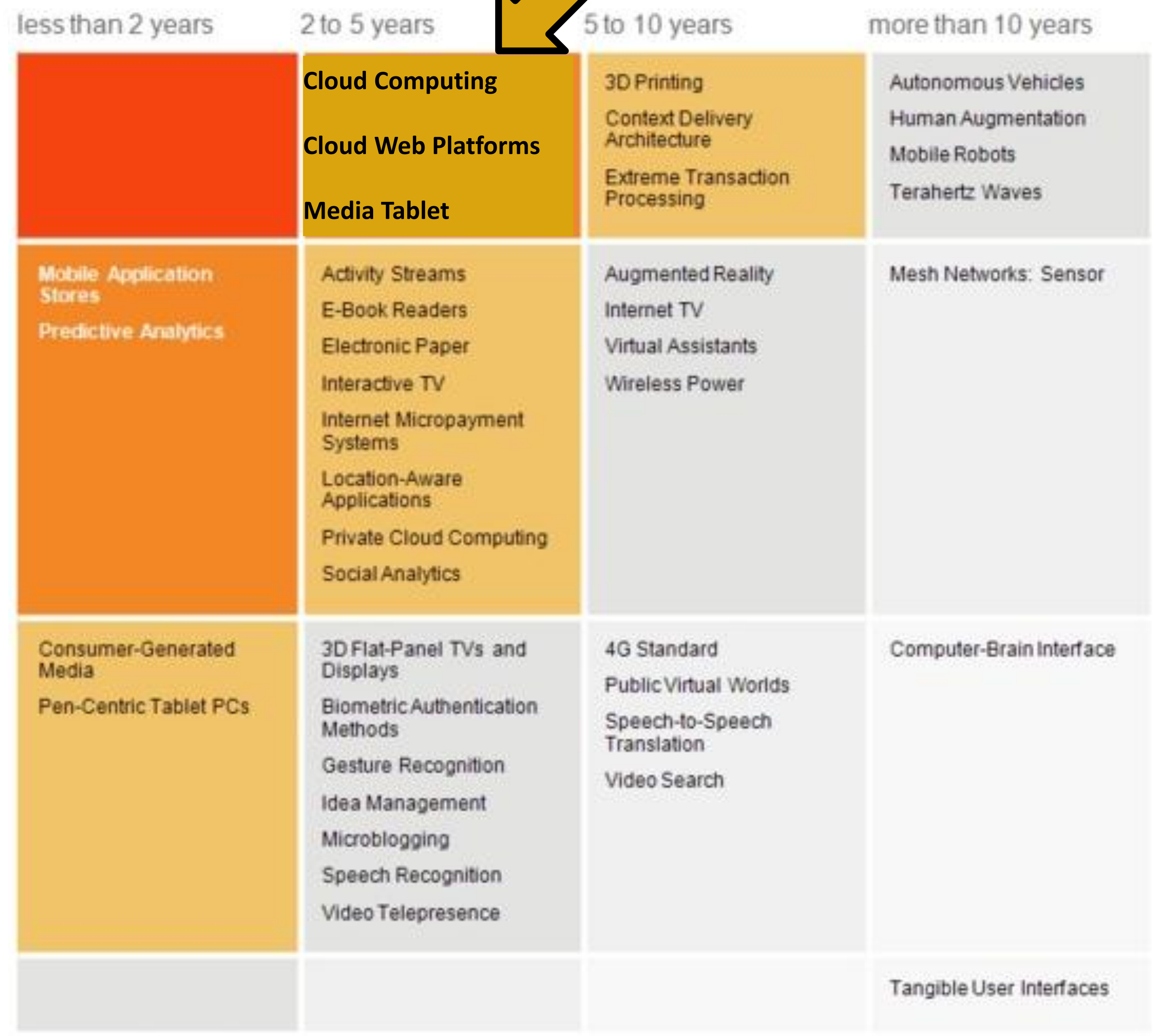
Talk Components

- Important Trends
- Clouds and Cloud Technologies
- Data Intensive Science Applications
  - Use of Clouds
- Summary
- FutureGrid

Important Trends

- Data Deluge in all fields of science
- Multicore implies parallel computing important again
  - Performance from extra cores – not extra clock speed
  - GPU enhanced systems can give big power boost
- Clouds – new commercially supported data center model replacing compute grids (and your general purpose computer center)
- Light weight clients: Sensors, Smartphones and tablets accessing and supported by backend services in cloud
- Commercial efforts moving much faster than academia in both innovation and deployment

2010 EMER



benefit	years to mainstream adoption			
	less than 2 years	2 to 5 years	5 to 10 years	more than 10 years
Transformational		**Cloud Computing** **Cloud Web Platforms** **Media Tablet**	3D Printing Context Delivery Architecture Extreme Transaction Processing	Autonomous Vehicles Human Augmentation Mobile Robots Terahertz Waves
High	Mobile Application Stores Predictive Analytics	Activity Streams E-Book Readers Electronic Paper Interactive TV Internet Micropayment Systems Location-Aware Applications Private Cloud Computing Social Analytics	Augmented Reality Internet TV Virtual Assistants Wireless Power	Mesh Networks: Sensor
Moderate	Consumer-Generated Media Pen-Centric Tablet PCs	3D Flat-Panel TVs and Displays Biometric Authentication Methods Gesture Recognition Idea Management Microblogging Speech Recognition Video Telepresence	4G Standard Public Virtual Worlds Speech-to-Speech Translation Video Search	Computer-Brain Interface
Low				Tangible User Interfaces

As of August 2010

Data Centers Clouds & Economies of Scale I

Range in size from “edge” facilities to megascale.

Economies of scale

Approximate costs for a small size center (1K servers) and a larger, 50K server center.

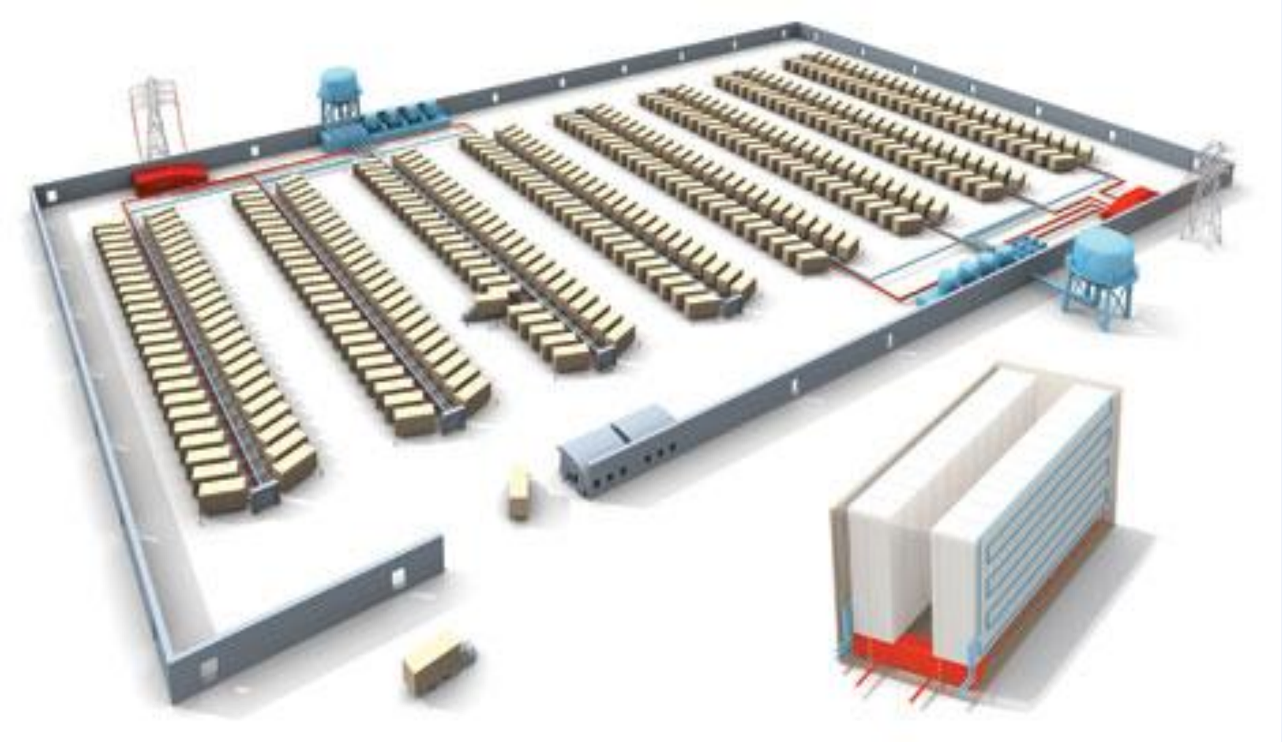
2 Google warehouses of computers on the banks of the Columbia River, in The Dalles, Oregon

Such centers use 20MW-200MW (Future) each with 150 watts per CPU

Save money from large size, positioning with cheap power and access with Internet

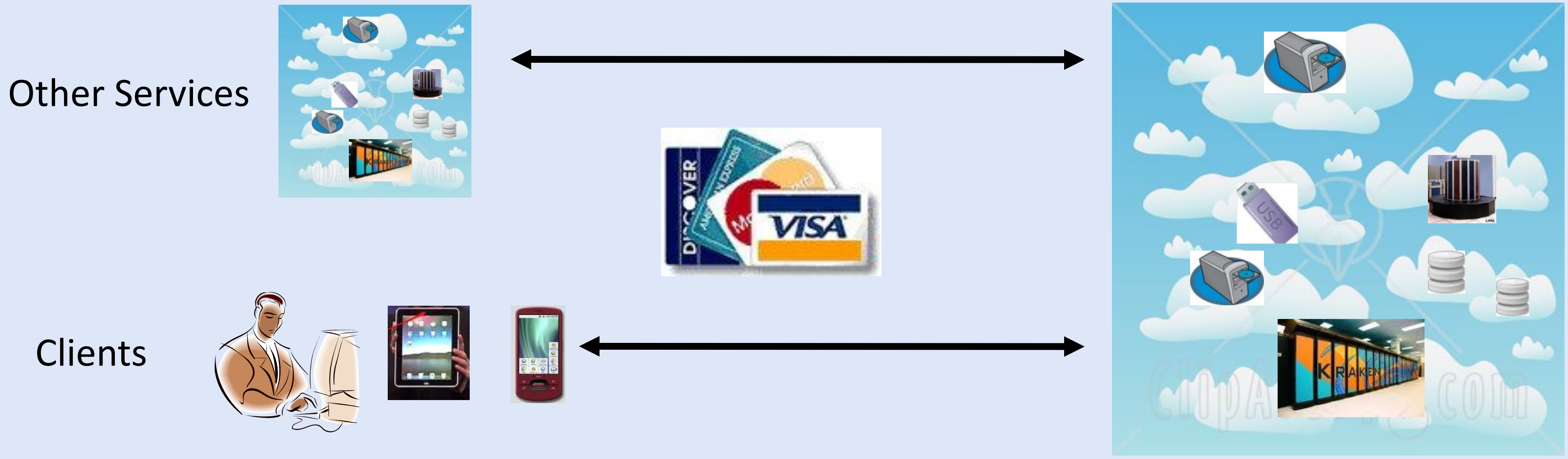
Data Centers, Clouds & Economies of Scale II

- Builds giant data centers with 100,000’s of computers; ~ 200-1000 to a shipping container with Internet access
- “Microsoft will cram between 150 and 220 shipping containers filled with data center gear into a new 500,000 square foot Chicago facility. This move marks the most significant, public use of the shipping container systems popularized by the likes of Sun Microsystems and Rackable Systems to date.”



X as a Service

- SaaS: Software as a Service imply software capabilities (programs) have a service (messaging) interface
  - Applying systematically reduces system complexity to being linear in number of components
  - Access via messaging rather than by installing in /usr/bin
- IaaS: Infrastructure as a Service or HaaS: Hardware as a Service – get your computer time with a credit card and with a Web interface
- PaaS: Platform as a Service is IaaS plus core software capabilities on which you build SaaS
- Cyberinfrastructure is "Research as a Service"



Other Services

Clients

Philosophy of Clouds and Grids

- Clouds are (by definition) commercially supported approach to large scale computing
  - So we should expect Clouds to replace Compute Grids
  - Current Grid technology involves “non-commercial” software solutions which are hard to evolve/sustain
  - Maybe Clouds ~4% IT expenditure 2008 growing to 14% in 2012 (IDC Estimate)
- Public Clouds are broadly accessible resources like Amazon and Microsoft Azure – powerful but not easy to customize and perhaps data trust/privacy issues
- Private Clouds run similar software and mechanisms but on “your own computers” (not clear if still elastic)
  - Platform features such as Queues, Tables, Databases currently limited
- Services still are correct architecture with either REST (Web 2.0) or Web Services
- Clusters are still critical concept for MPI or Cloud software

Grids MPI and Clouds

- Grids are useful for **managing distributed systems**
  - Pioneered service model for Science
  - Developed importance of Workflow
  - Performance issues – communication latency – intrinsic to distributed systems
  - Can never run large differential equation based simulations or datamining
- Clouds **can execute any job class that was good for Grids** plus
  - More attractive due to platform plus **elastic** on-demand model
  - **MapReduce easier to use than MPI for appropriate parallel jobs**
  - Currently have performance limitations due to poor affinity (locality) for compute-compute (MPI) and Compute-data
  - These limitations are not "inevitable" and should gradually improve as in July 13 Amazon Cluster announcement
  - Will probably never be best for most sophisticated parallel differential equation based simulations
- Classic Supercomputers (MPI Engines) run **communication demanding differential equation based simulations**
  - **MapReduce and Clouds replaces MPI** for other problems
  - Much more data processed today by MapReduce than MPI (Industry Informational Retrieval ~50 Petabytes per day)

Cloud Computing: Infrastructure and Runtimes

- Cloud infrastructure: outsourcing of servers, computing, data, file space, utility computing, etc.
  - Handled through Web services that control virtual machine lifecycles.
- Cloud runtimes or Platform: tools (for using clouds) to do data-parallel (and other) computations.
  - Apache Hadoop, Google MapReduce, Microsoft Dryad, Bigtable, Chubby and others
  - MapReduce designed for information retrieval but is excellent for a wide range of science data analysis applications
  - Can also do much traditional parallel computing for data-mining if extended to support iterative operations
  - MapReduce not usually on Virtual Machines

Authentication and Authorization: Provide single sign in to both FutureGrid and Commercial Clouds linked by workflow

Workflow: Support workflows that link job components between FutureGrid and Commercial Clouds. Trident from Microsoft Research is initial candidate

Data Transport: Transport data between job components on FutureGrid and Commercial Clouds respecting custom storage patterns

Program Library: Store Images and other Program material (basic FutureGrid facility)

Blob: Basic storage concept similar to Azure Blob or Amazon S3

DPFS Data Parallel File System: Support of file systems like Google (MapReduce), HDFS (Hadoop) or Cosmos (dryad) with compute-data affinity optimized for data processing

Table: Support of Table Data structures modeled on Apache Hbase/CouchDB or Amazon SimpleDB/Azure Table. There is “Big” and “Little” tables – generally NOSQL

SQL: Relational Database

Queues: Publish Subscribe based queuing system

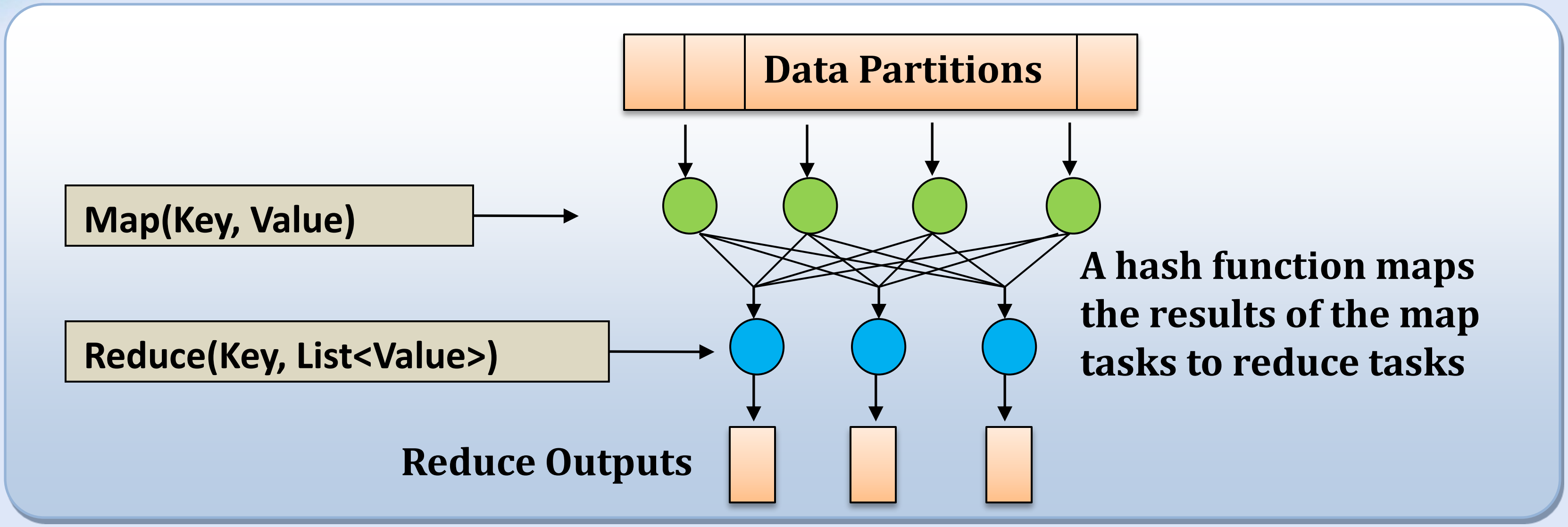
Worker Role: This concept is implicitly used in both Amazon and TeraGrid but was first introduced as a high level construct by Azure

MapReduce: Support MapReduce Programming model including Hadoop on Linux, Dryad on Windows HPCS and Twister on Windows and Linux

Software as a Service: This concept is shared between Clouds and Grids and can be supported without special attention

Web Role: This is used in Azure to describe important link to user and can be supported in FutureGrid with a Portal framework

MapReduce



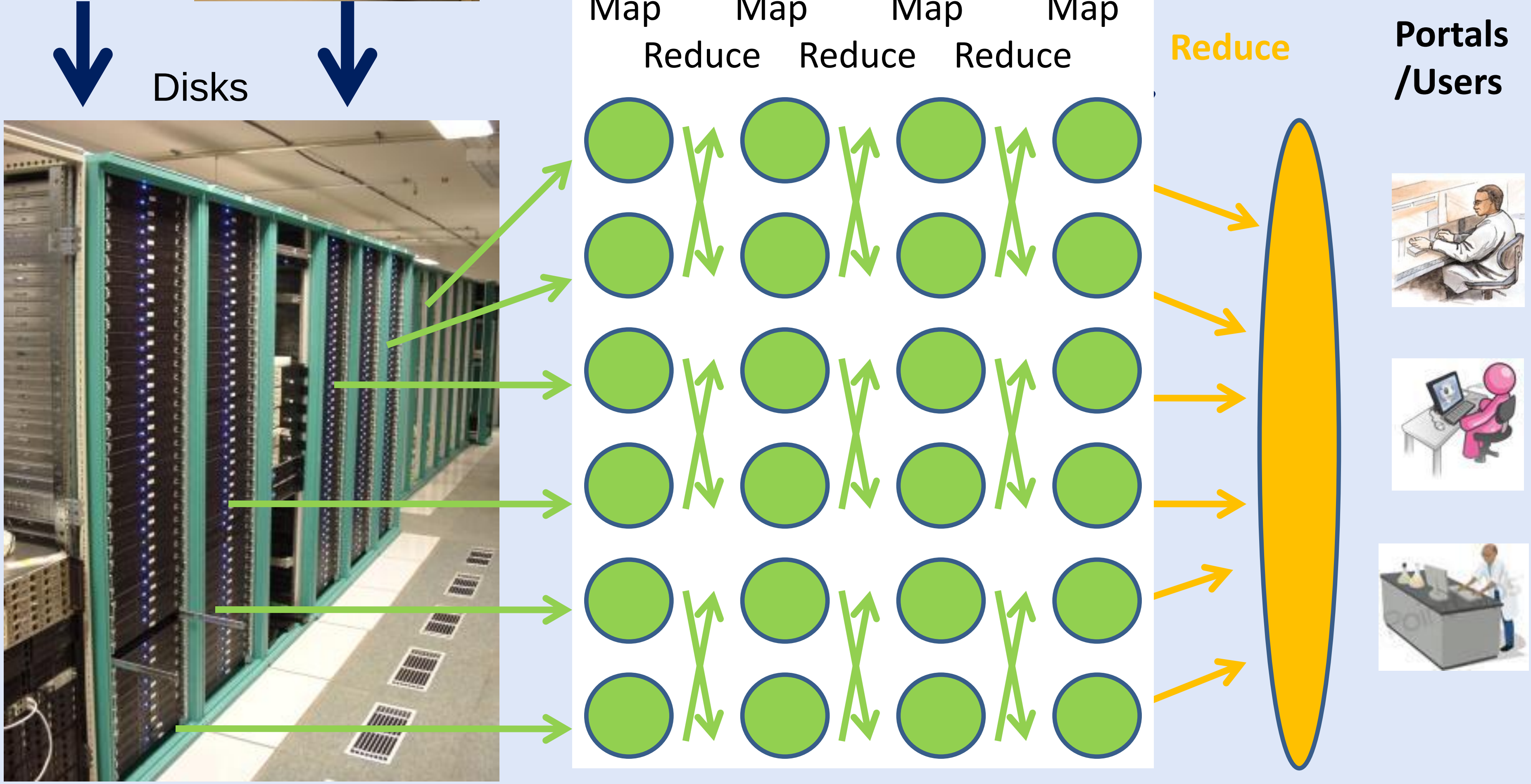
- Implementations (Hadoop – Java; Dryad – Windows) support:
  - Splitting of data
  - Passing the output of map functions to reduce functions
  - Sorting the inputs to the reduce function based on the intermediate keys
  - Quality of service

MapReduce “File/Data Repository” Parallelism

Map = (data parallel) computation reading and writing data

Reduce = Collective/Consolidation phase e.g. forming multiple global sums as in histogram

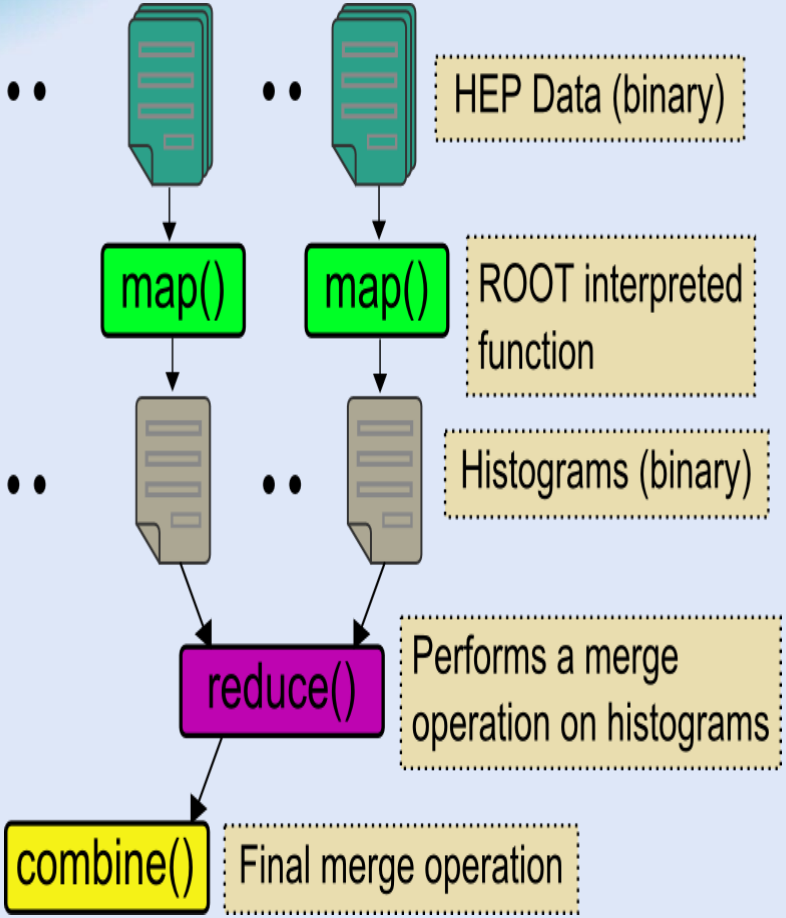
Instruments

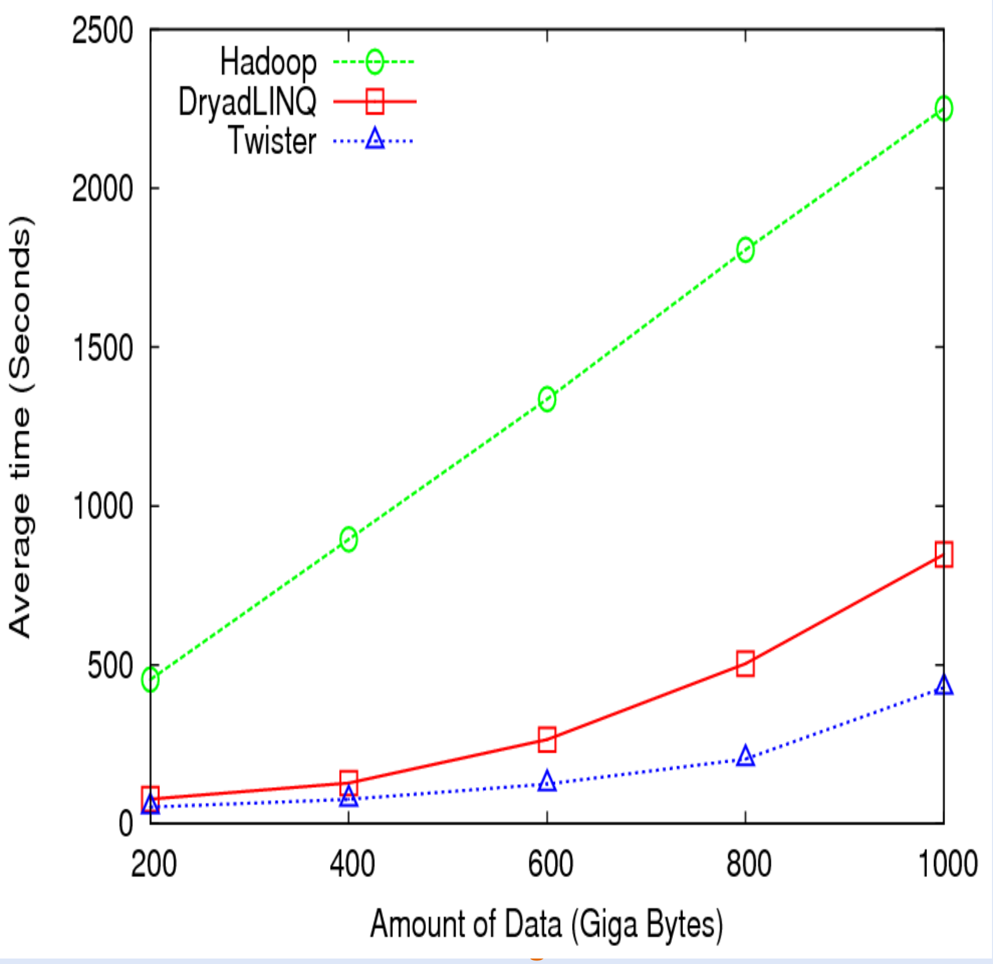


High Energy Physics Data Analysis

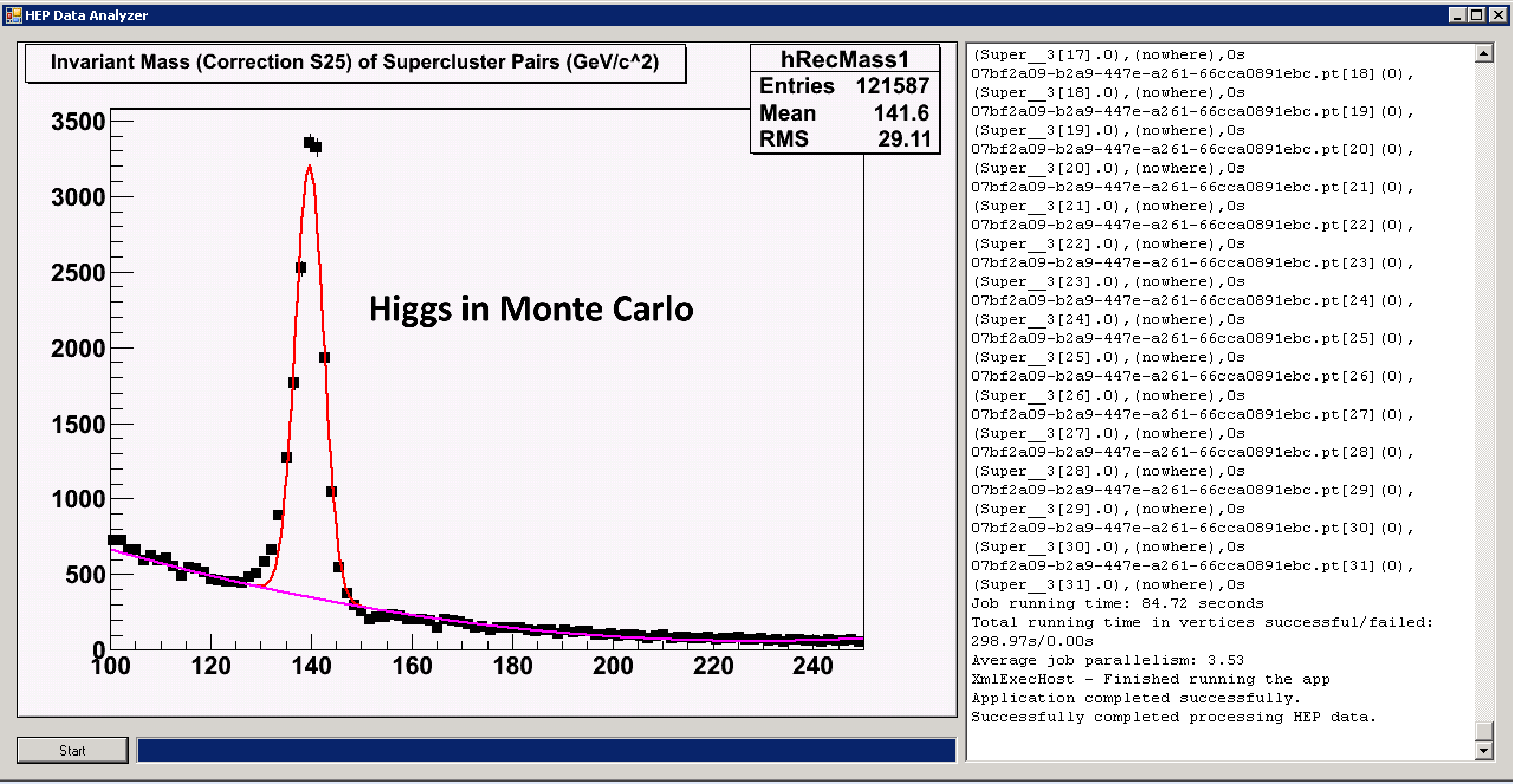
An application analyzing data from Large Hadron Collider

(1TB but 100 Petabytes eventually)





Reduce Phase of Particle Physics “Find the Higgs” using Dryad



```
(Super__3[17].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[18] (0),
(Super__3[18].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[19] (0),
(Super__3[19].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[20] (0),
(Super__3[20].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[21] (0),
(Super__3[21].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[22] (0),
(Super__3[22].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[23] (0),
(Super__3[23].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[24] (0),
(Super__3[24].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[25] (0),
(Super__3[25].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[26] (0),
(Super__3[26].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[27] (0),
(Super__3[27].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[28] (0),
(Super__3[28].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[29] (0),
(Super__3[29].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[30] (0),
(Super__3[30].0),(nowhere),0s
07bf2a09-b2a9-447e-a261-66cca0891ebc.pt[31] (0),
(Super__3[31].0),(nowhere),0s
Job running time: 84.72 seconds
Total running time in vertices successful/failed:
298.97s/0.00s
Average job parallelism: 3.53
XmlExecHost - Finished running the app
Application completed successfully.
Successfully completed processing HEP data.
```

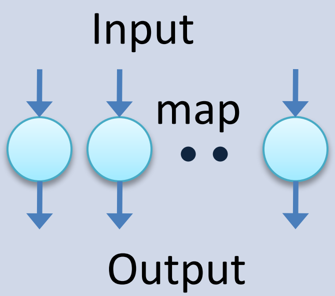
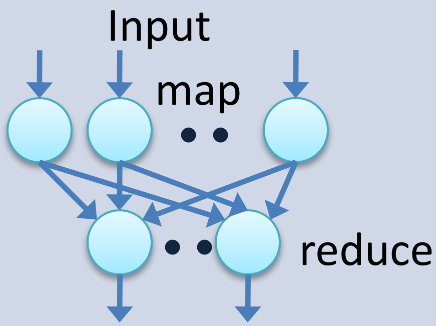
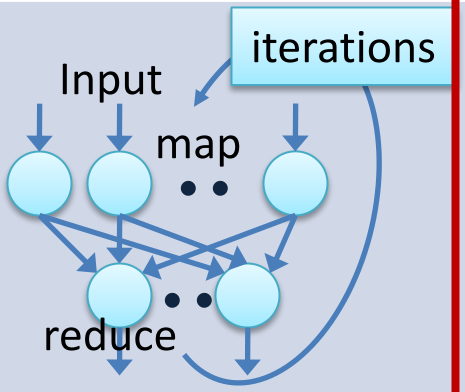
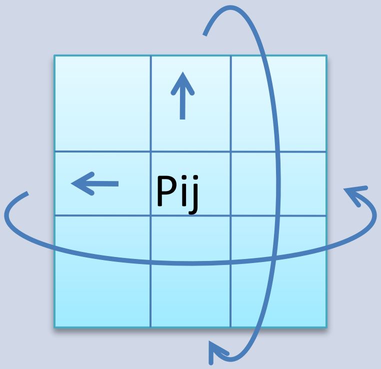
- Combine Histograms produced by separate Root “Maps” (of event data to partial histograms) into a single Histogram delivered to Client

Application Classes

Old classification of Parallel software/hardware in terms of 5 (becoming 6) “Application architecture” Structures)

1	Synchronous	Lockstep Operation as in SIMD architectures	**SIMD**
2	Loosely Synchronous	Iterative Compute-Communication stages with independent compute (map) operations for each CPU. Heart of most MPI jobs	**MPP**
3	Asynchronous	Computer Chess; Combinatorial Search often supported by dynamic threads	**MPP**
4	**Pleasingly Parallel**	Each component independent – in 1988, Fox estimated at 20% of total number of applications	**Grids**
5	**Metaproblems**	Coarse grain (asynchronous) combinations of classes 1)-4). The preserve of workflow.	**Grids**
6	**MapReduce++**	It describes file(database) to file(database) operations which has subcategories including. 1) Pleasingly Parallel Map Only 2) Map followed by reductions 3) Iterative “Map followed by reductions” – Extension of Current Technologies that supports much linear algebra and datamining	**Clouds** **Hadoop/ Dryad** **Twister**

Applications & Different Interconnection Patterns

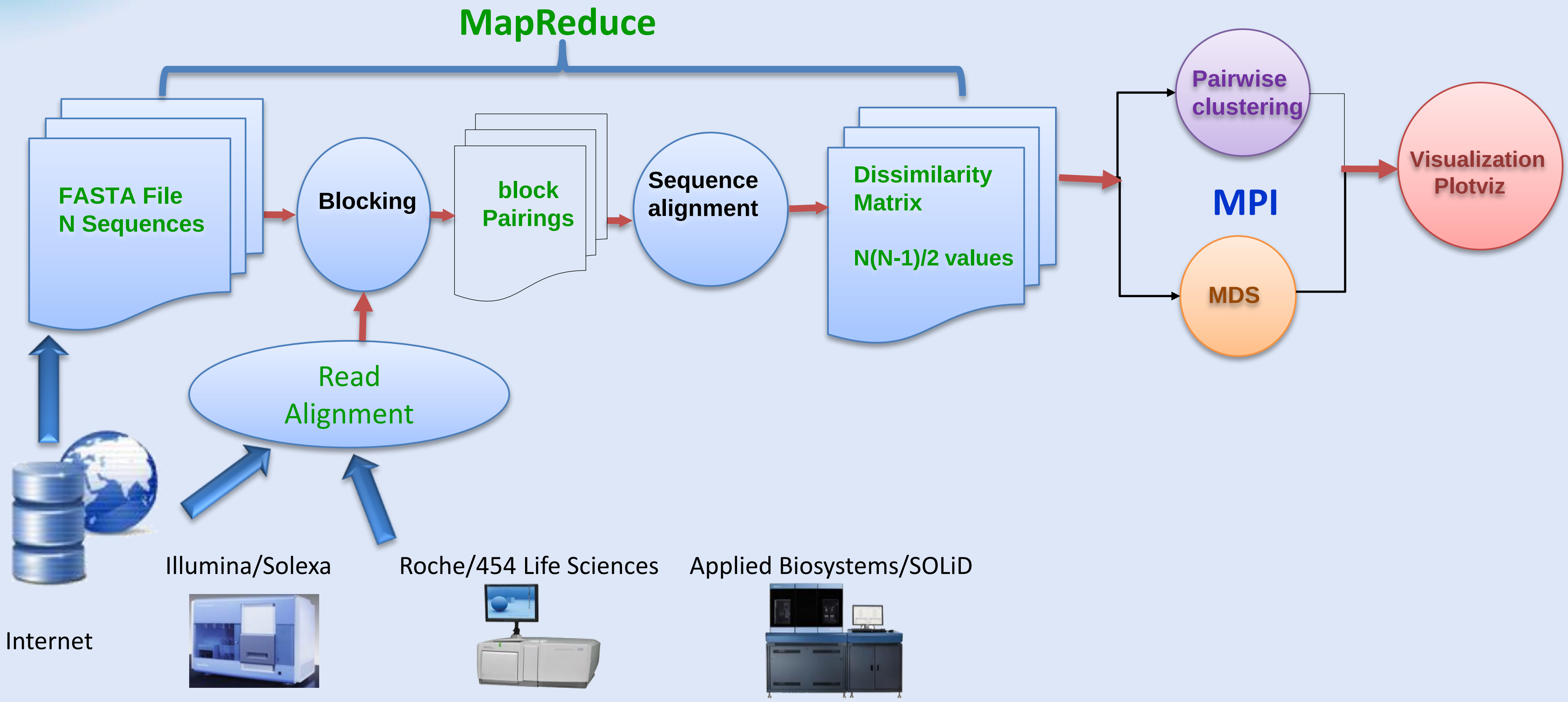
Map Only	Classic MapReduce	Iterative Reductions **MapReduce++**	Loosely Synchronous
			
CAP3 Analysis Document conversion (PDF -> HTML) Brute force searches in cryptography Parametric sweeps	High Energy Physics (**HEP**) Histograms **SWG** gene alignment Distributed search Distributed sorting Information retrieval	Expectation maximization algorithms Clustering Linear Algebra	Many MPI scientific applications utilizing wide variety of communication constructs including local interactions
- CAP3 Gene Assembly - PolarGrid Matlab data analysis	- Information Retrieval - HEP Data Analysis - Calculation of Pairwise Distances for ALU Sequences	- Kmeans - **Deterministic Annealing Clustering** - Multidimensional Scaling **MDS**	- Solving Differential Equations and - particle dynamics with short range forces

← Domain of MapReduce and Iterative Extensions → MPI

Fault Tolerance and MapReduce

- MPI does “maps” followed by “communication” including “reduce” but does this iteratively
- There must (for most communication patterns of interest) be a strict synchronization at end of each communication phase
  - Thus if a process fails then everything grinds to a halt
- In MapReduce, all Map processes and all reduce processes are independent and stateless and read and write to disks
  - As 1 or 2 (reduce+map) iterations, no difficult synchronization issues
- Thus failures can easily be recovered by rerunning process without other jobs hanging around waiting
- Re-examine MPI fault tolerance in light of MapReduce
  - Twister will interpolate between MPI and MapReduce

DNA Sequencing Pipeline



MapReduce

FASTA File
N Sequences

Blocking

block
Pairings

Sequence
alignment

Dissimilarity
Matrix

N(N-1)/2 values

Pairwise
clustering

MPI

MDS

Visualization
Plotviz

Read
Alignment

Internet

Illumina/Solexa

Roche/454 Life Sciences

Applied Biosystems/SOLiD

Modern Commercial Gene Sequencers

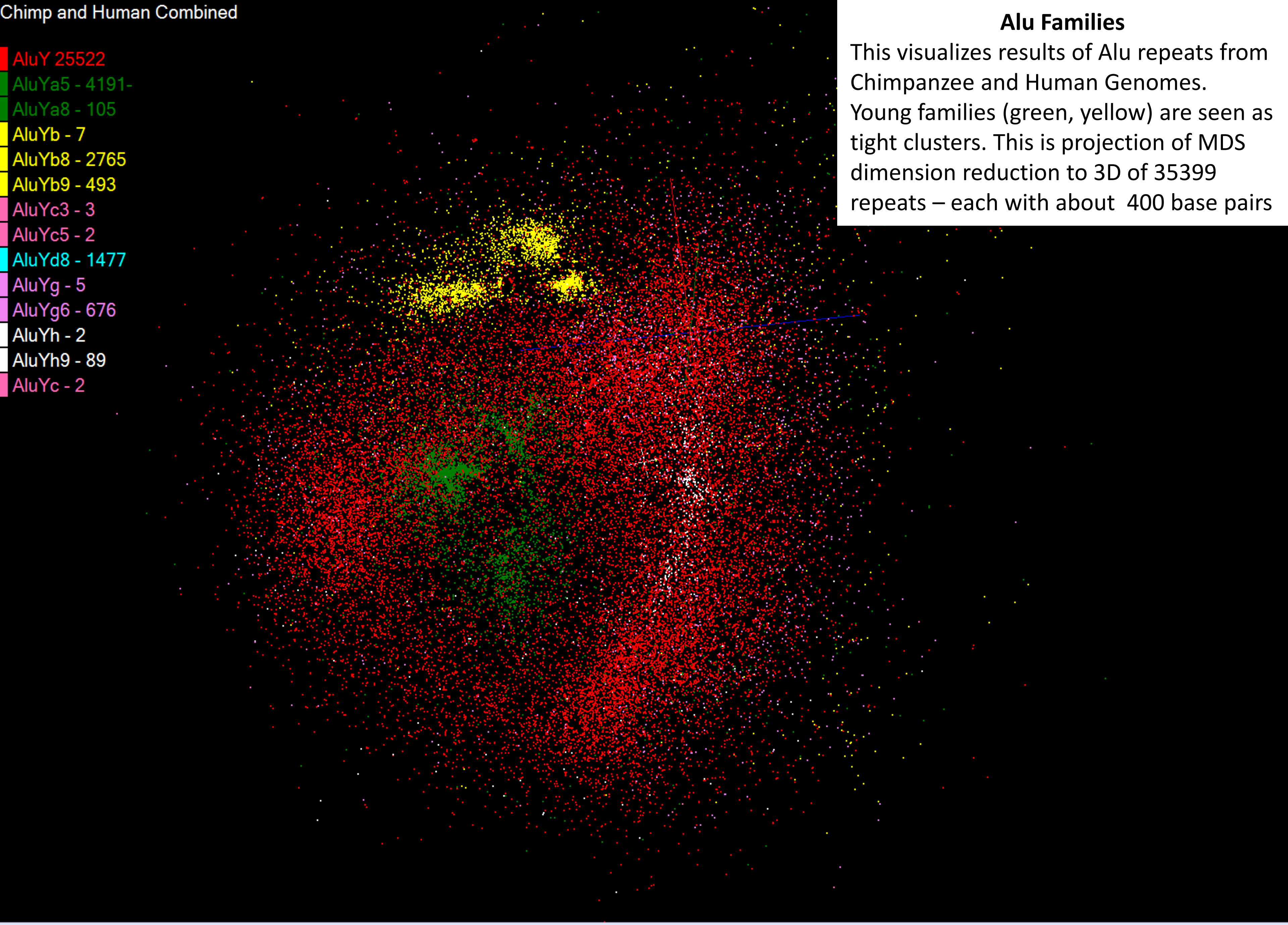
- This chart illustrate our research of a pipeline mode to provide services on demand (Software as a Service SaaS)
- User submit their jobs to the pipeline. The components are services and so is the whole pipeline.

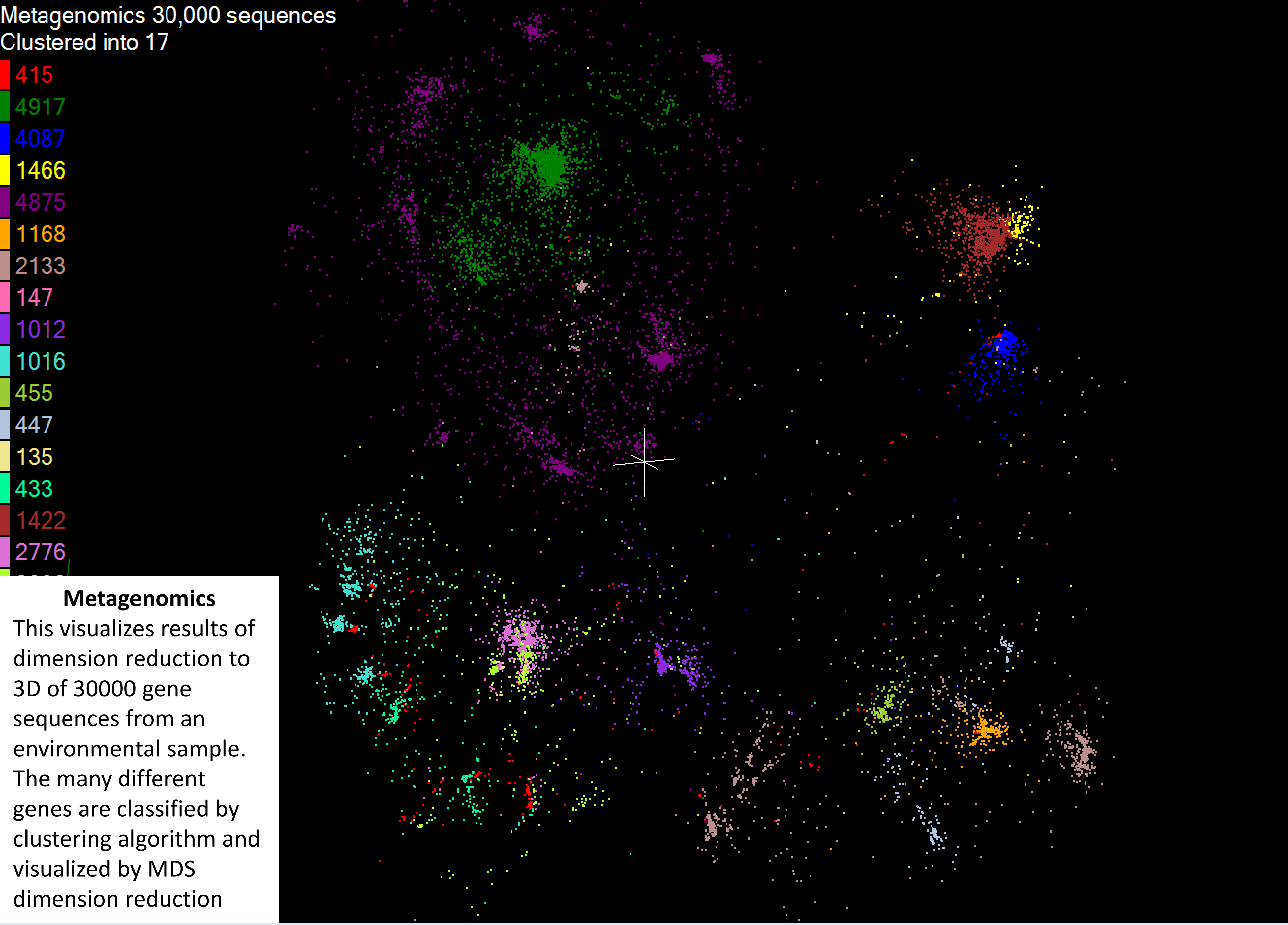
Alu and Metagenomics Workflow All Pairs

- Data is a collection of N sequences. Need to calculate N^2 dissimilarities (distances) between sequences.
  - These cannot be thought of as vectors because there are missing characters
- Step 1: Calculate N^2 dissimilarities (distances) between sequences
- Step 2: Find families by clustering (using much better methods than Kmeans). As no vectors, use vector free $O(N^2)$ methods
- Step 3: Map to 3D for visualization using Multidimensional Scaling (MDS) – also $O(N^2)$
- Note N = 50,000 runs in 10 hours (the complete pipeline above) on 768 cores
- Need to address millions of sequences; develop new O(NlogN) algorithms
- Currently using a mix of MapReduce (step 1) and MPI as steps 2,3 use classic matrix algorithms
- Twister could do all steps as MDS, Clustering just need MPI Broadcast/Reduce

Alu Families

This visualizes results of Alu repeats from Chimpanzee and Human Genomes. Young families (green, yellow) are seen as tight clusters. This is projection of MDS dimension reduction to 3D of 35399 repeats – each with about 400 base pairs

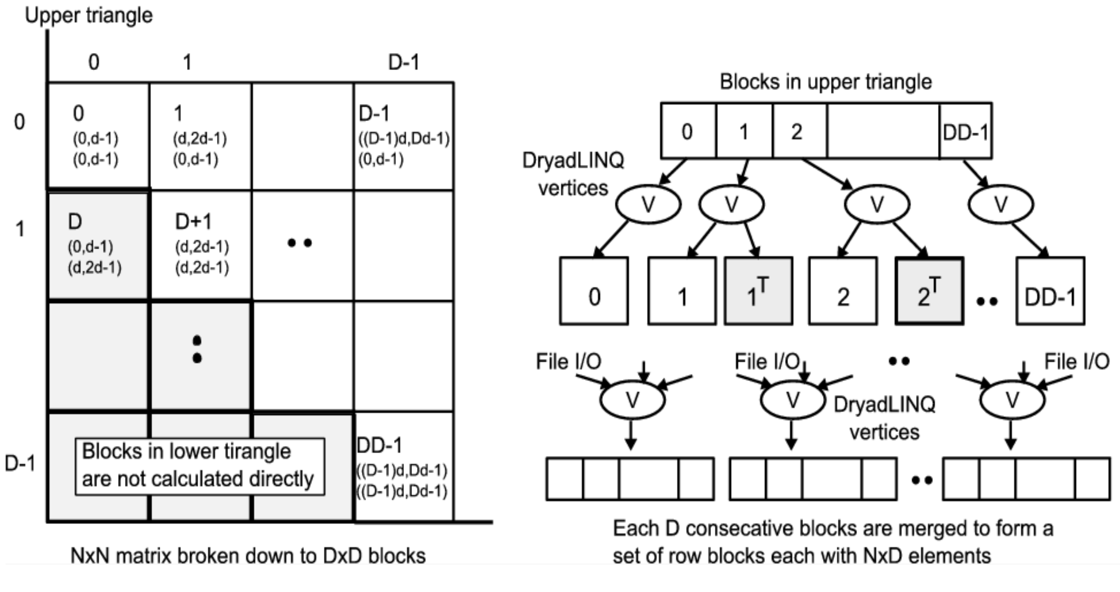




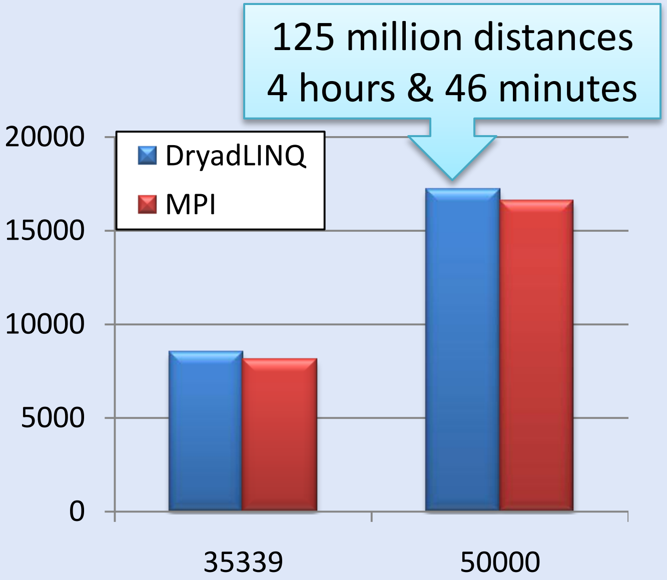
Metagenomics

This visualizes results of dimension reduction to 3D of 30000 gene sequences from an environmental sample. The many different genes are classified by clustering algorithm and visualized by MDS dimension reduction

All-Pairs Using DryadLINQ



Calculate Pairwise Distances (Smith Waterman Gotoh)

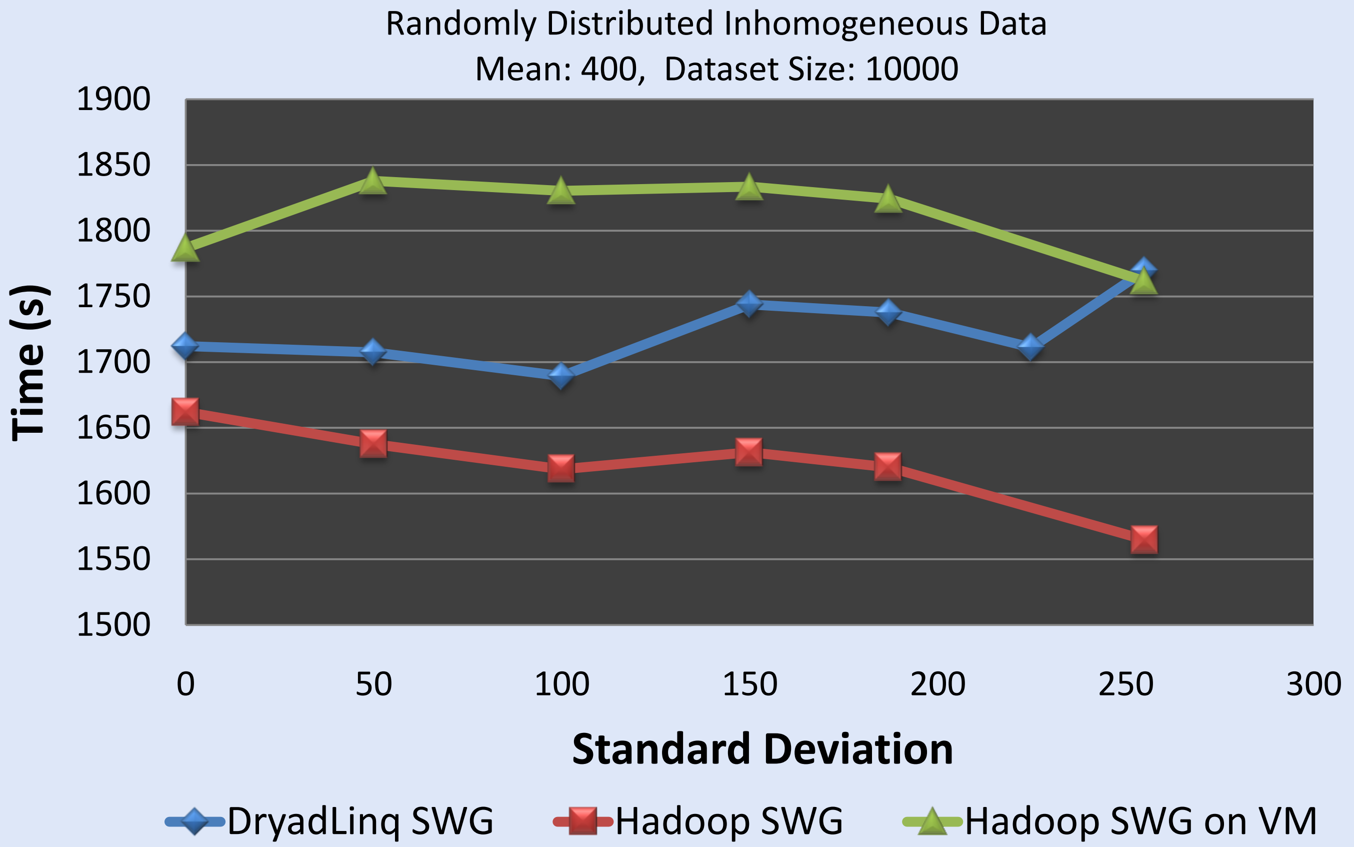

125 million distances
4 hours & 46 minutes

- Calculate pairwise distances for a collection of genes (used for clustering, MDS)
- Fine grained tasks in MPI
- Coarse grained tasks in DryadLINQ
- Performed on 768 cores (Tempest Cluster)

Moretti, C., Bui, H., Hollingsworth, K., Rich, B., Flynn, P., & Thain, D. (2009). All-Pairs: An Abstraction for Data Intensive Computing on Campus Grids. *IEEE Transactions on Parallel and Distributed Systems , 21*, 21-36.

Hadoop/Dryad Comparison

Inhomogeneous Data I

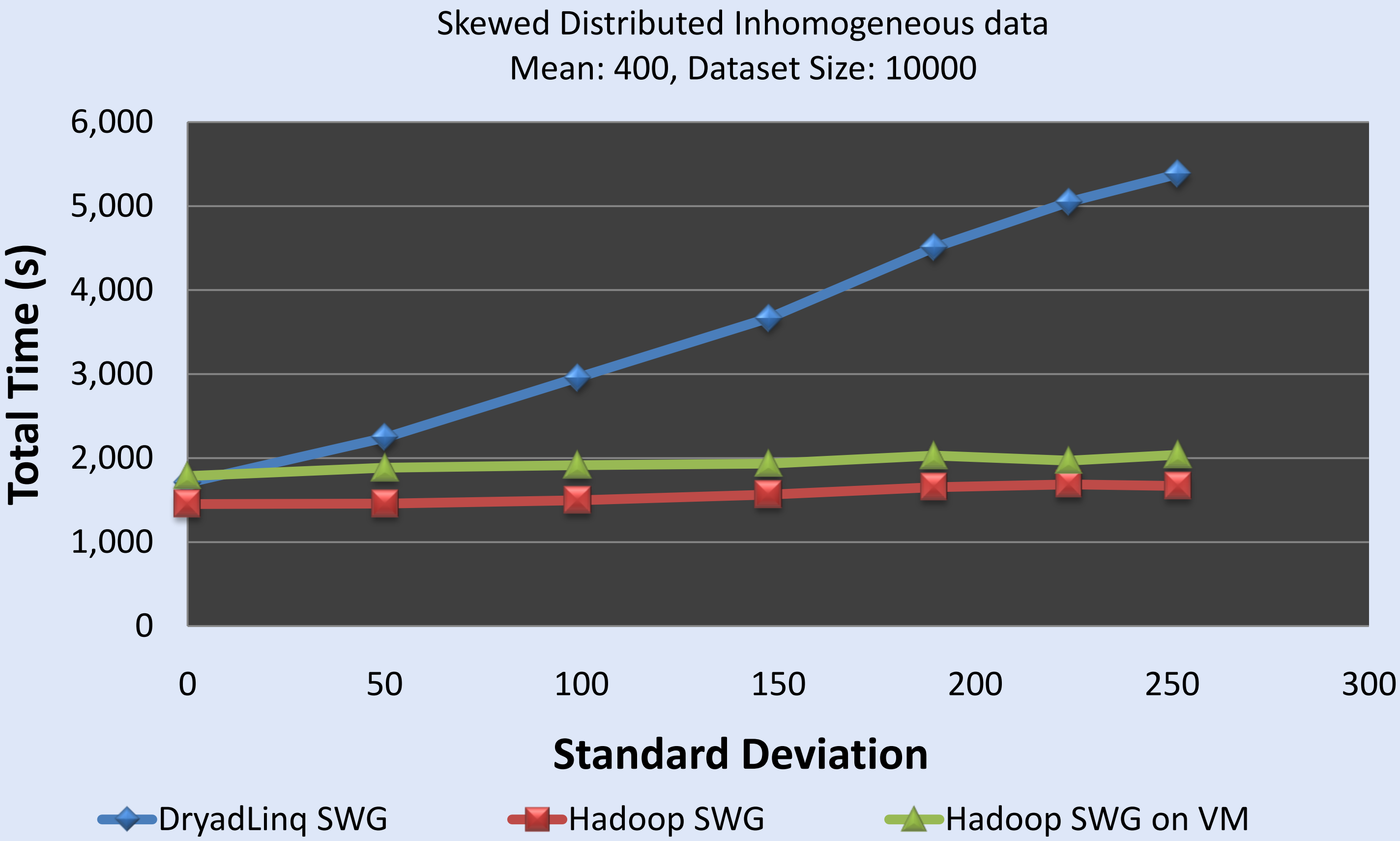


Inhomogeneity of data does not have a significant effect when the sequence lengths are randomly distributed

Dryad with Windows HPCS compared to Hadoop with Linux RHEL on Idataplex (32 nodes)

Hadoop/Dryad Comparison

Inhomogeneous Data II

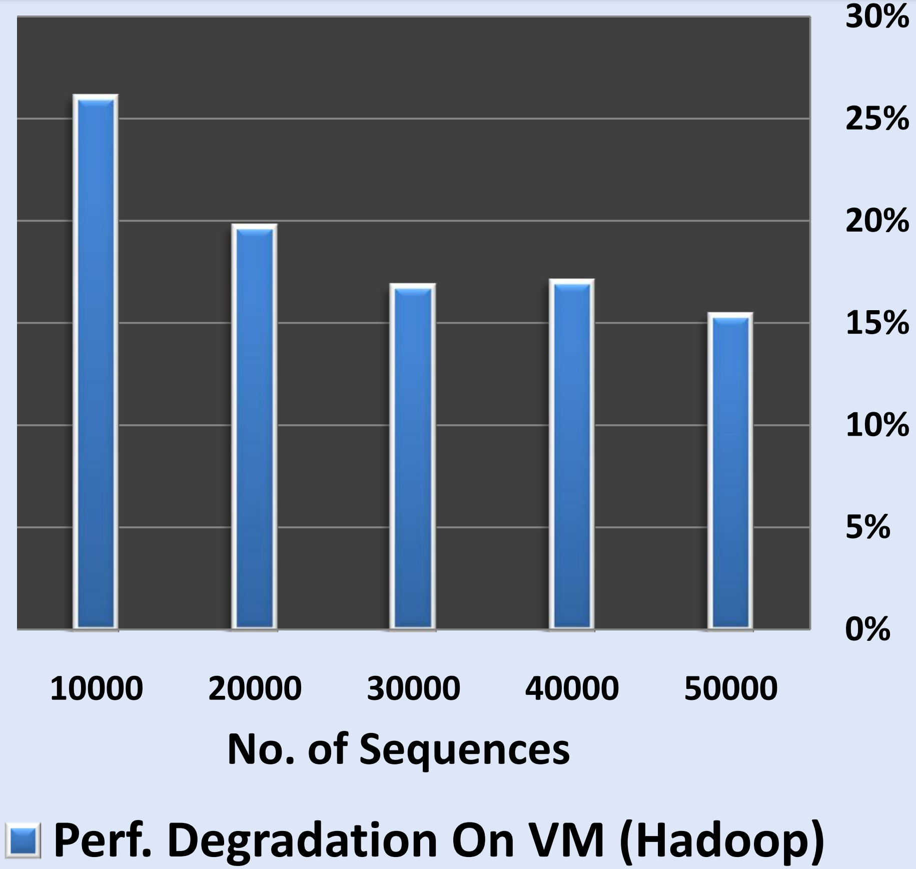


This shows the natural load balancing of Hadoop MR dynamic task assignment using a global pipe line in contrast to the DryadLinq static assignment

Dryad with Windows HPCS compared to Hadoop with Linux RHEL on Idataplex (32 nodes)

Hadoop VM Performance Degradation

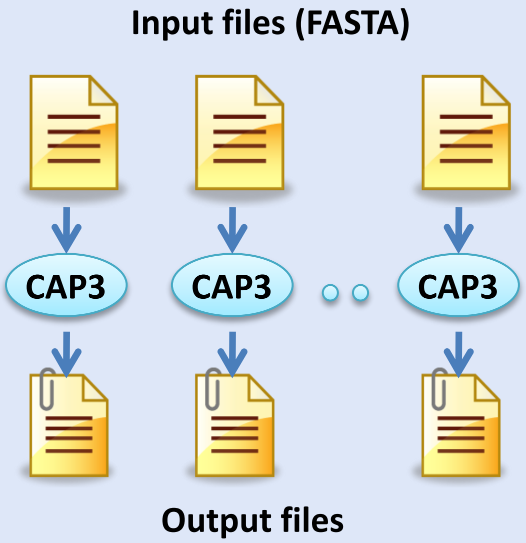
Perf. Degradation = $(T_{vm} - T_{baremetal})/T_{baremetal}$



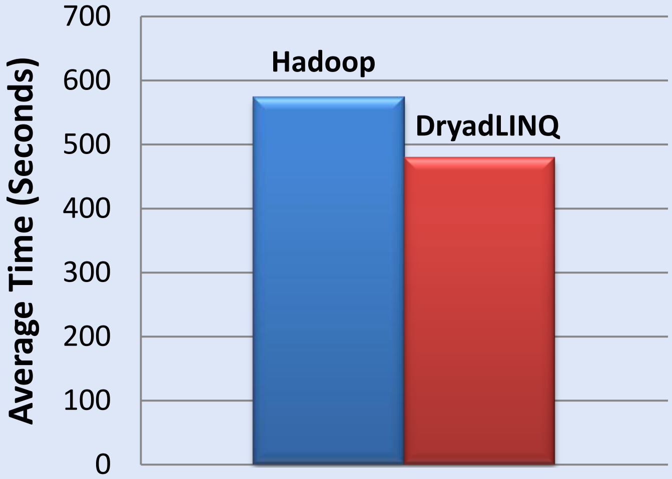
15.3% Degradation at largest data set size

Applications using Dryad & DryadLINQ

CAP3 - Expressed Sequence Tag assembly to re-construct full-length mRNA



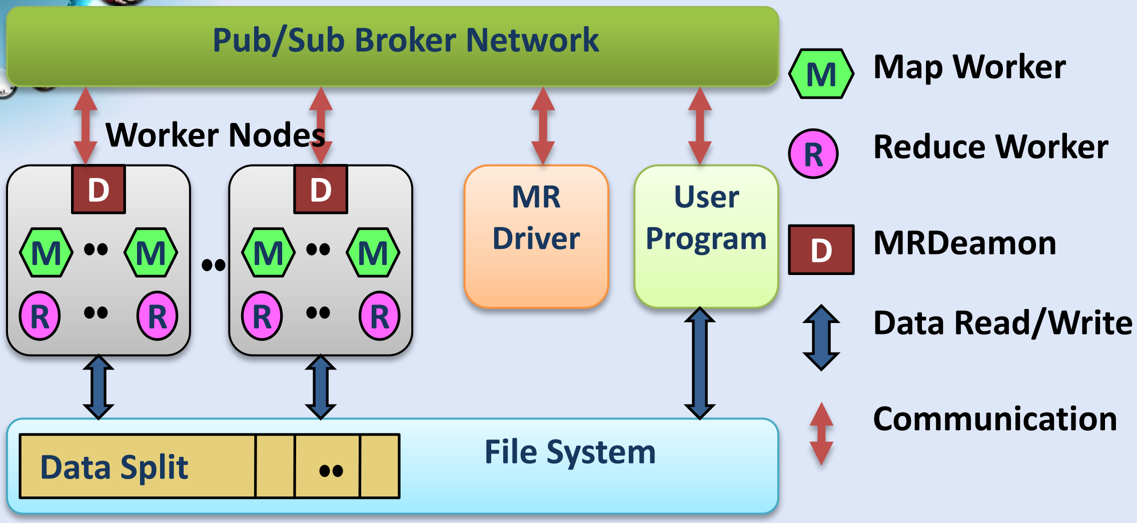
Time to process 1280 files each with ~375 sequences



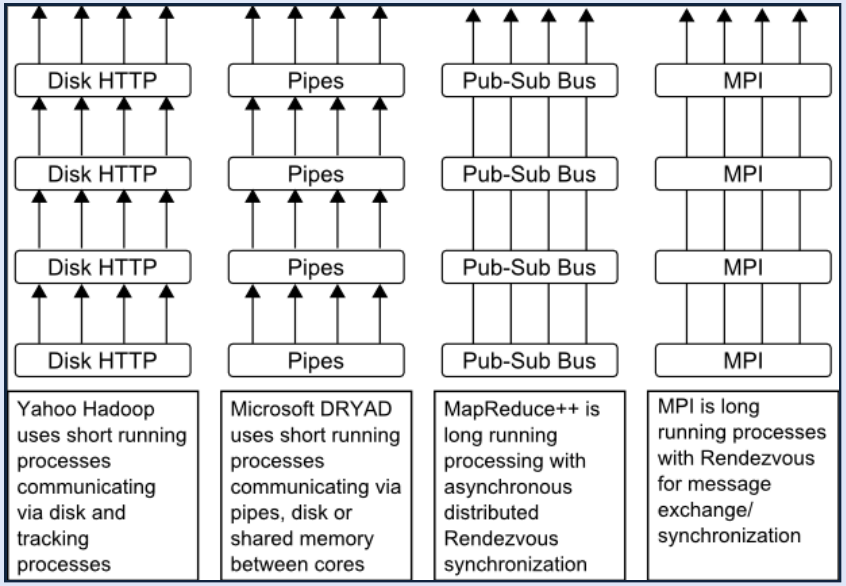
- Perform using DryadLINQ and Apache Hadoop implementations
- Single "Select" operation in DryadLINQ
- "Map only" operation in Hadoop

X. Huang, A. Madan, "CAP3: A DNA Sequence Assembly Program," Genome Research, vol. 9, no. 9, pp. 868-877, 1999.

Twister



- Streaming based communication
- Intermediate results are directly transferred from the map tasks to the reduce tasks – **eliminates local files**
- Cacheable map/reduce tasks
  - Static data remains in memory
- Combine phase to combine reductions
- User Program is the **composer** of MapReduce computations
- **Extends the MapReduce model to iterative computations**



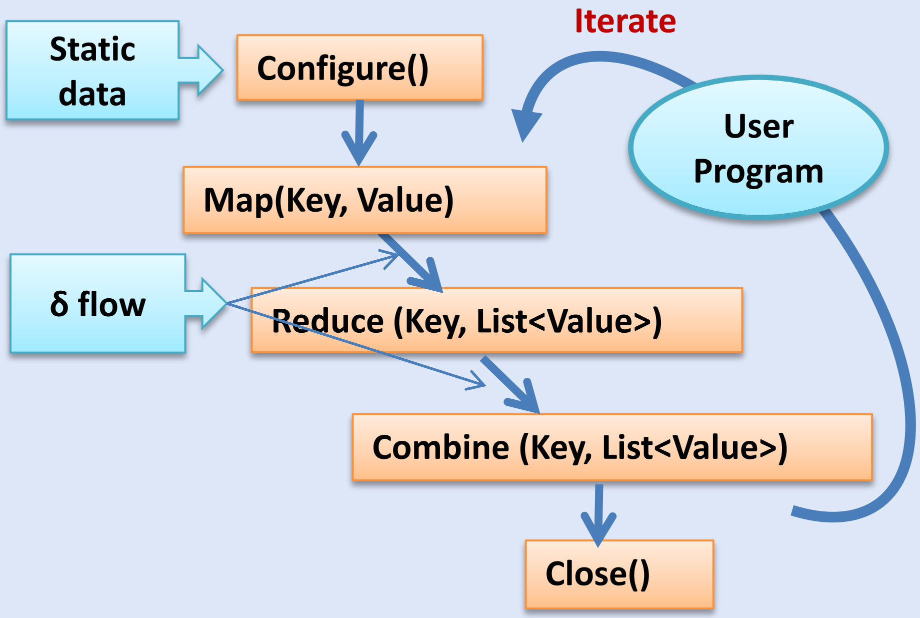
Yahoo Hadoop uses short running processes communicating via disk and tracking processes

Microsoft DRYAD uses short running processes communicating via pipes, disk or shared memory between cores

MapReduce++ is long running processing with asynchronous distributed Rendezvous synchronization

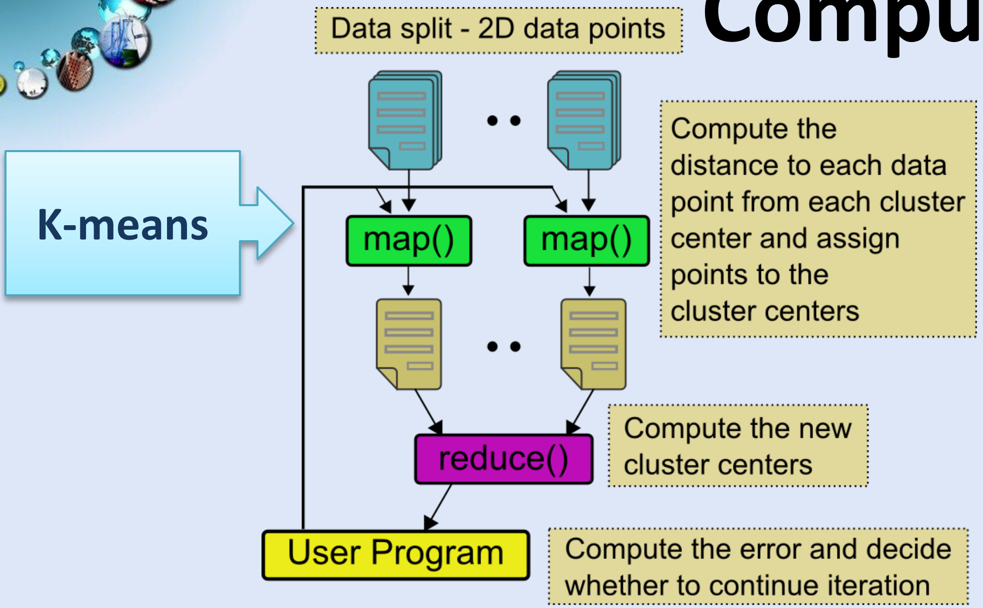
MPI is long running processes with Rendezvous for message exchange/ synchronization

Different synchronization and intercommunication mechanisms used by the parallel runtimes



Iterative and non-Iterative Computations

Future Grid



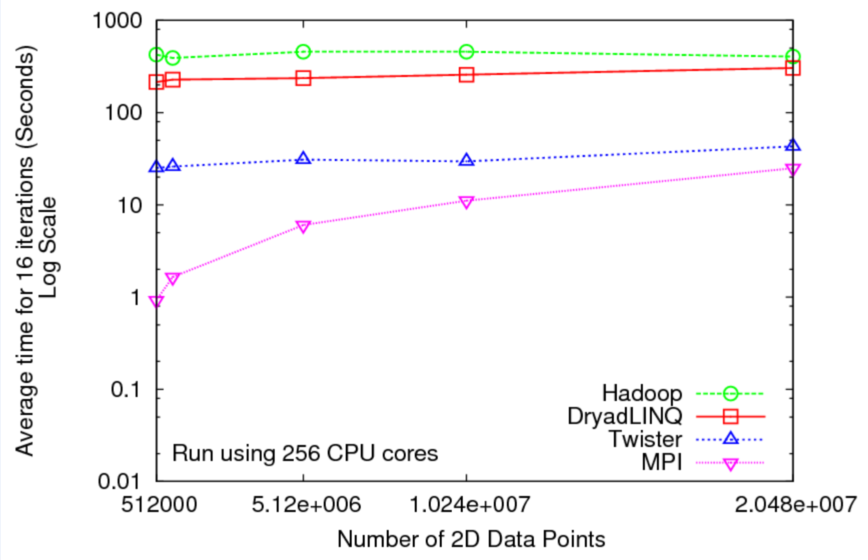
Compute the distance to each data point from each cluster center and assign points to the cluster centers

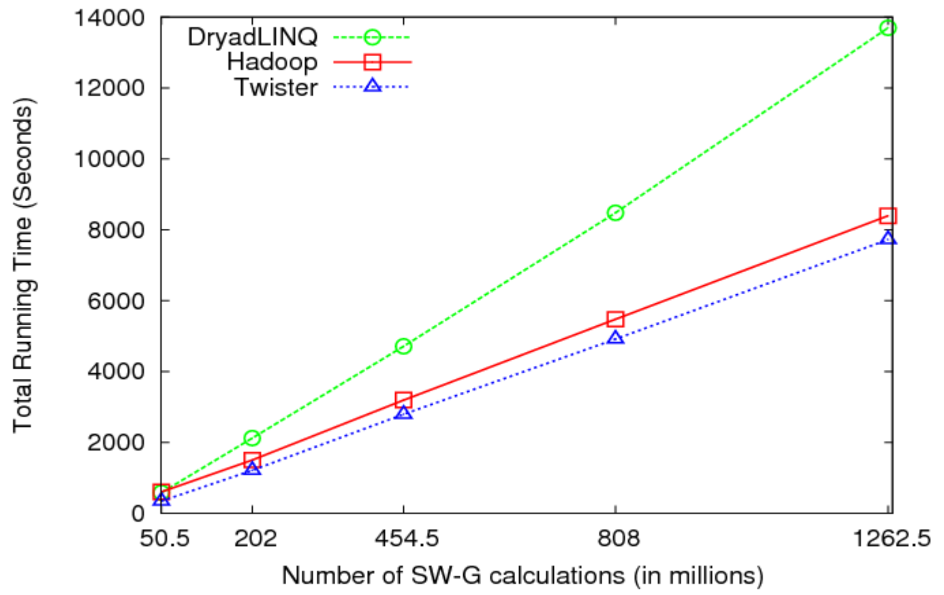
Compute the new cluster centers

Compute the error and decide whether to continue iteration

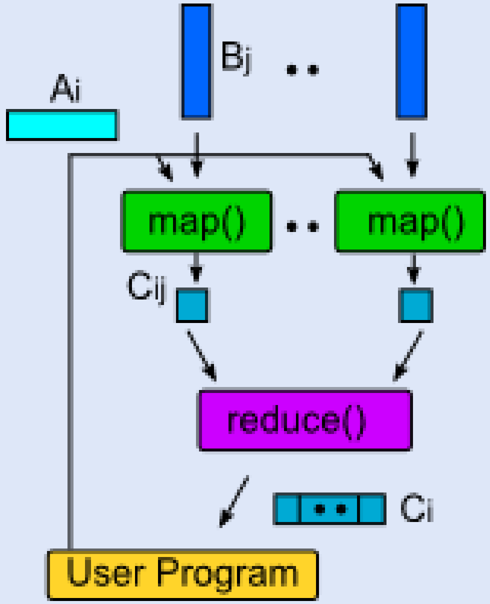
Smith Waterman is a non iterative case and of course runs fine

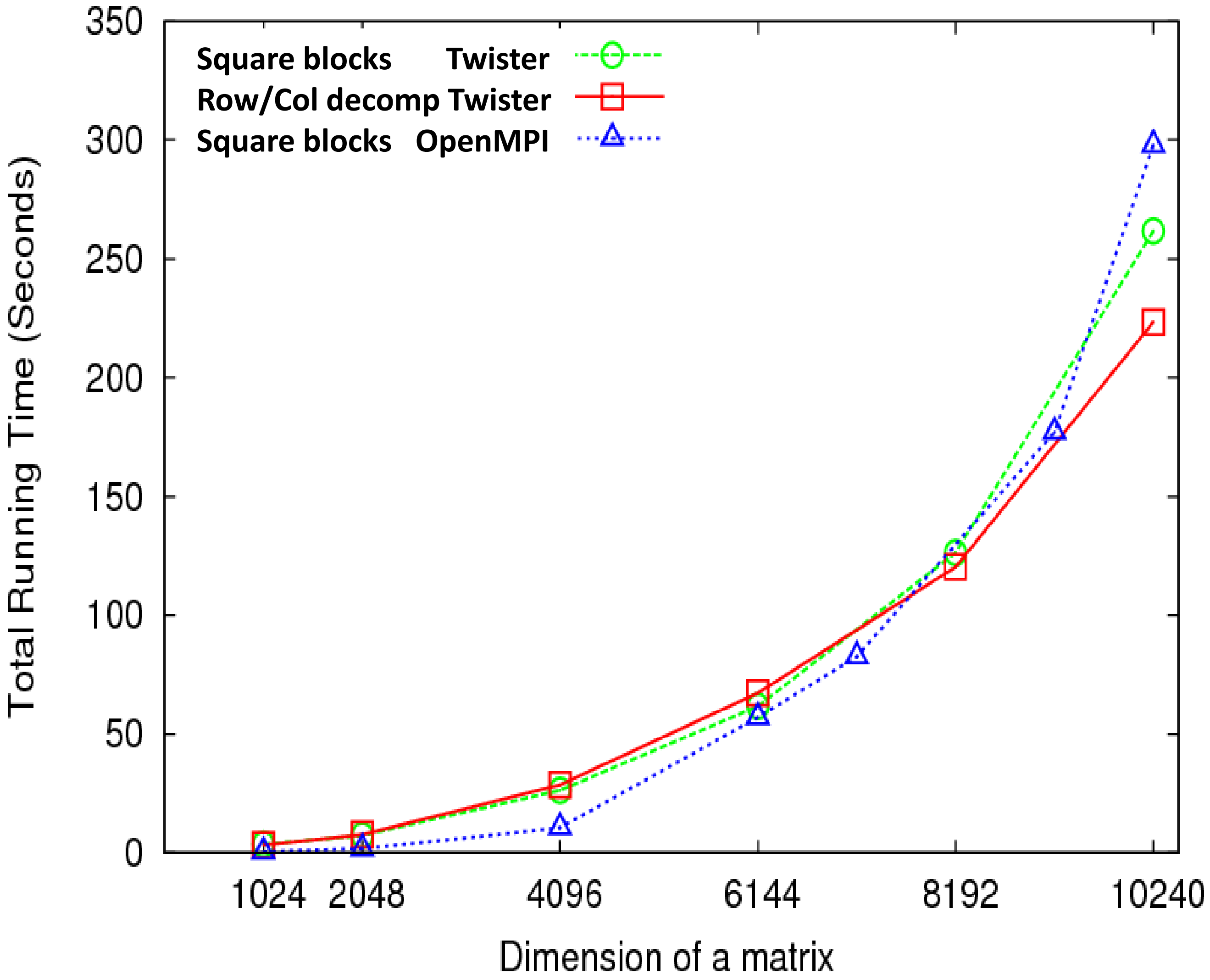
Performance of K-Means



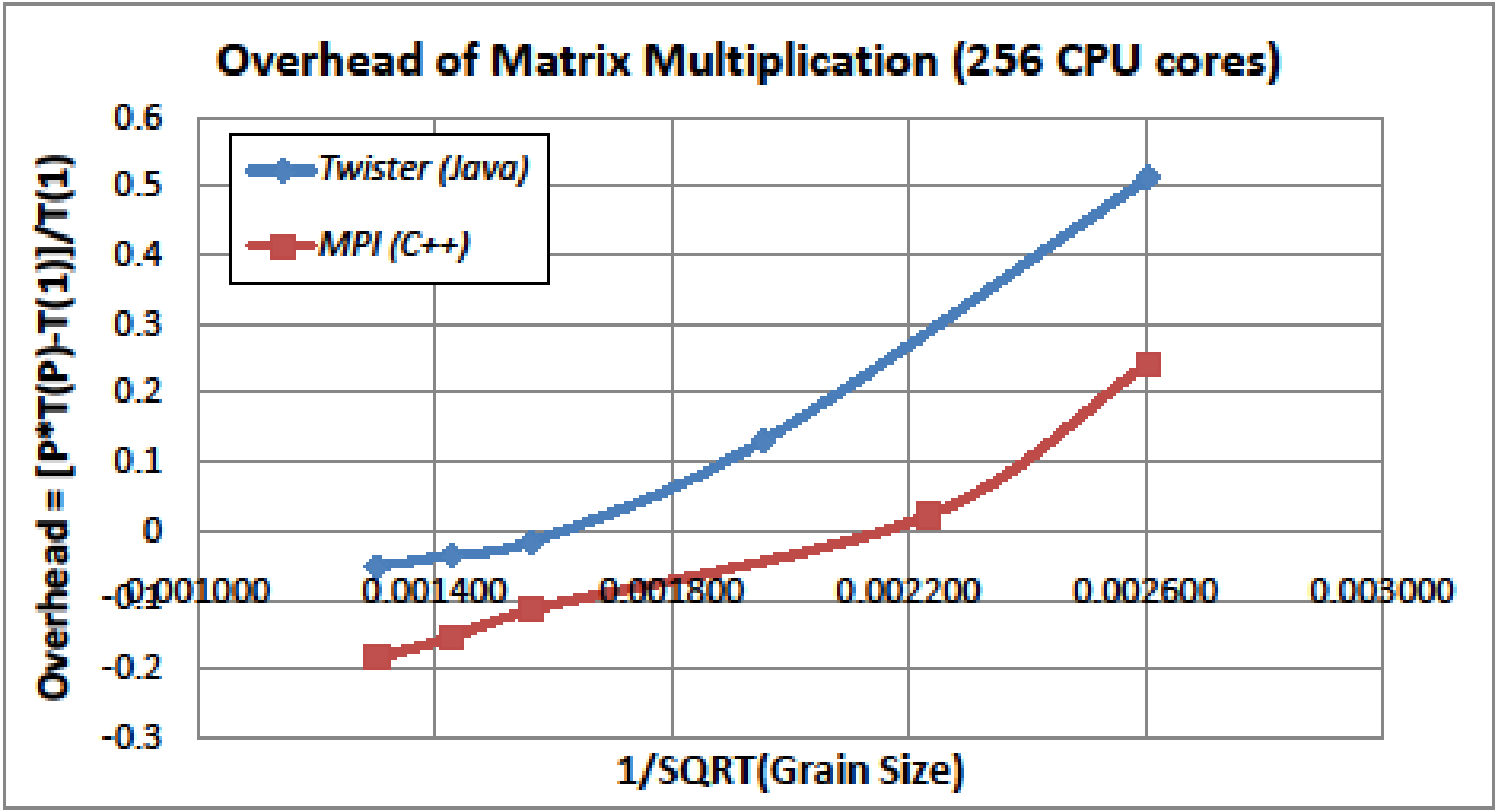


Matrix Multiplication 64 cores



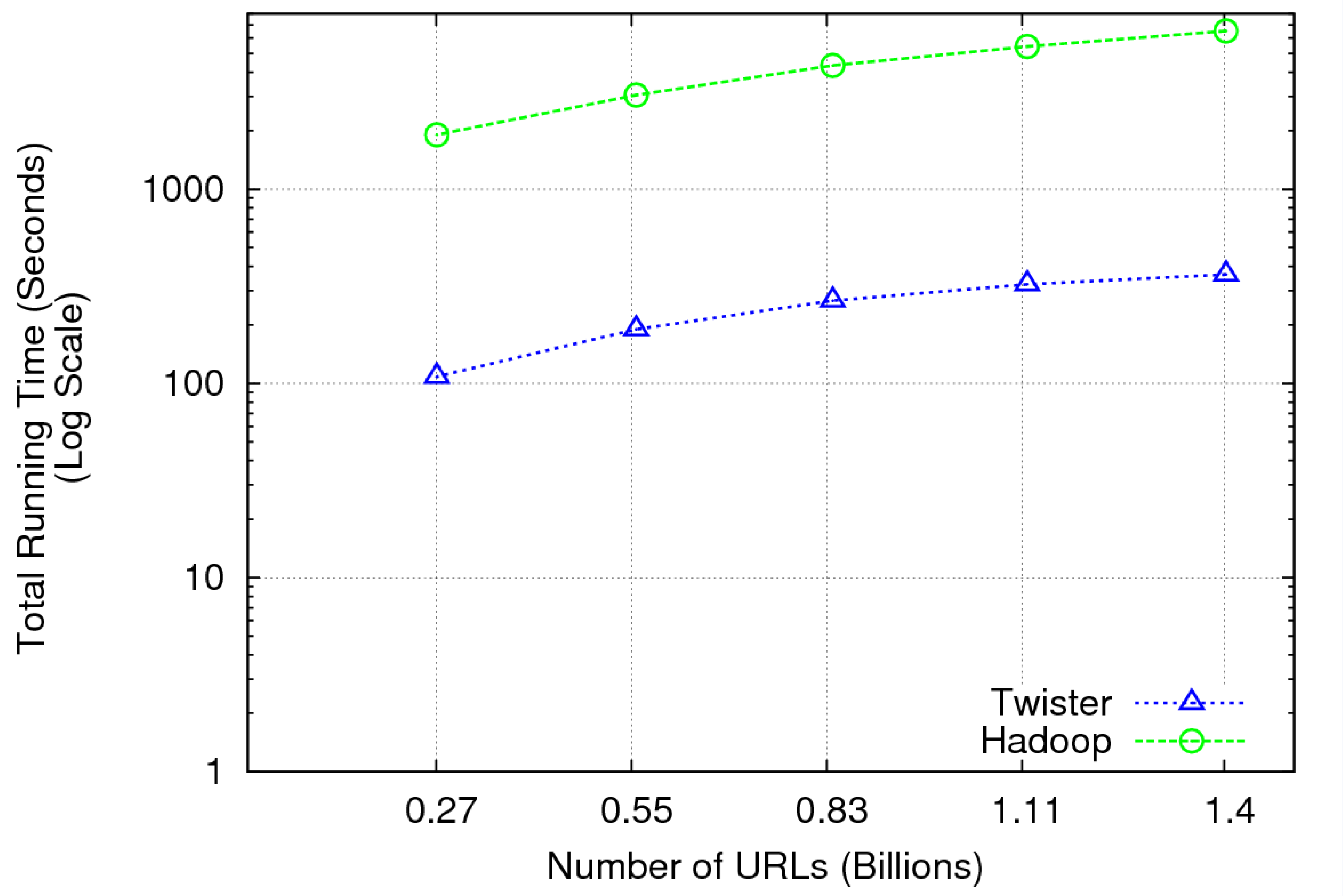


Overhead OpenMPI v Twister negative overhead due to cache

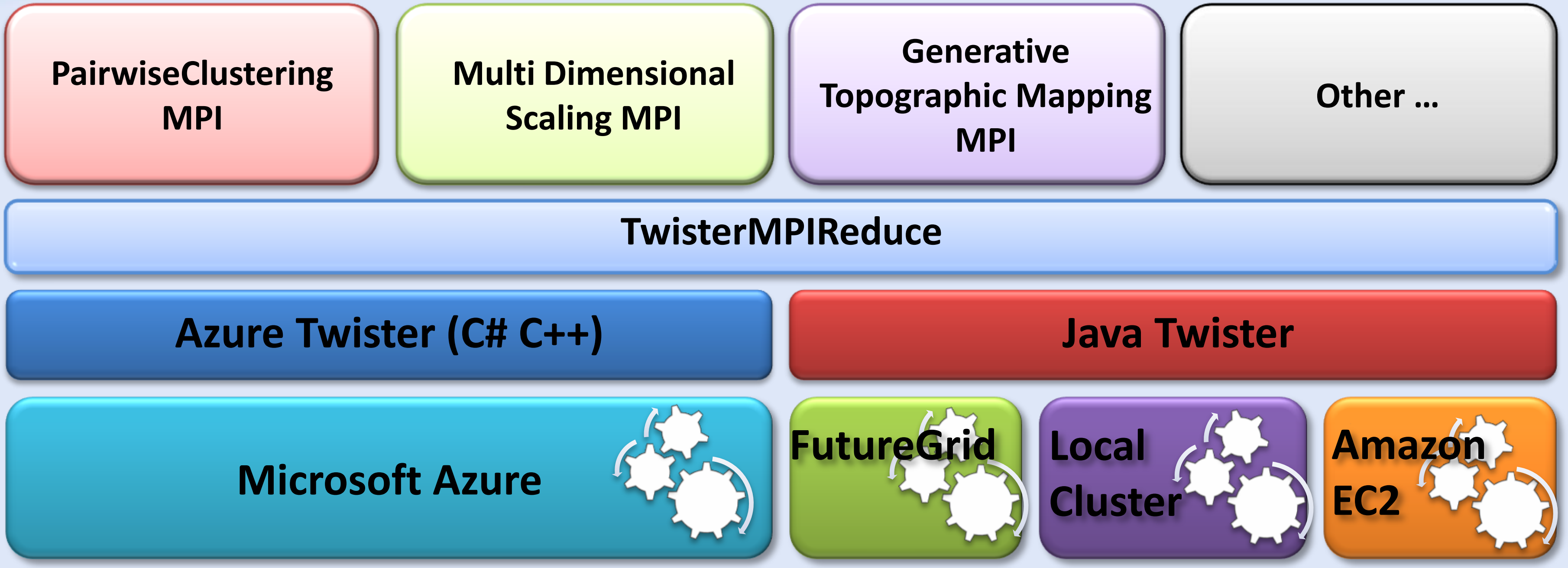


Performance of Pagerank using ClueWeb Data (Time for 20 iterations)

using 32 nodes (256 CPU cores) of Crevasse



TwisterMPIReduce



PairwiseClustering MPI

Multi Dimensional Scaling MPI

Generative Topographic Mapping MPI

Other …

TwisterMPIReduce

Azure Twister (C# C++)

Java Twister

Microsoft Azure

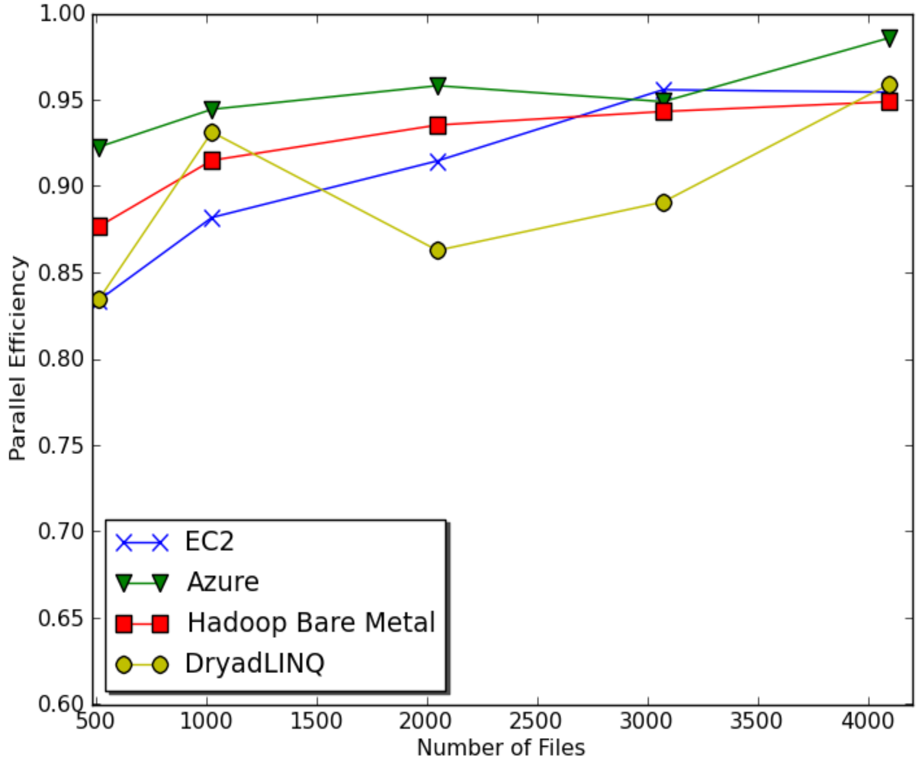
FutureGrid

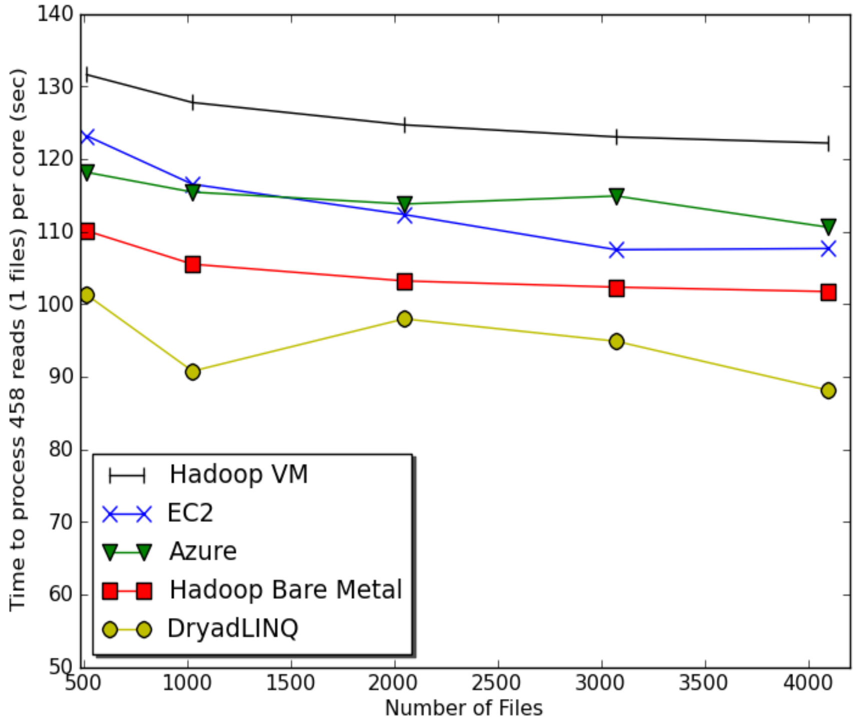
Local Cluster

Amazon EC2

- Runtime package supporting subset of MPI mapped to Twister
- Set-up, Barrier, Broadcast, Reduce

Sequence Assembly in the Clouds

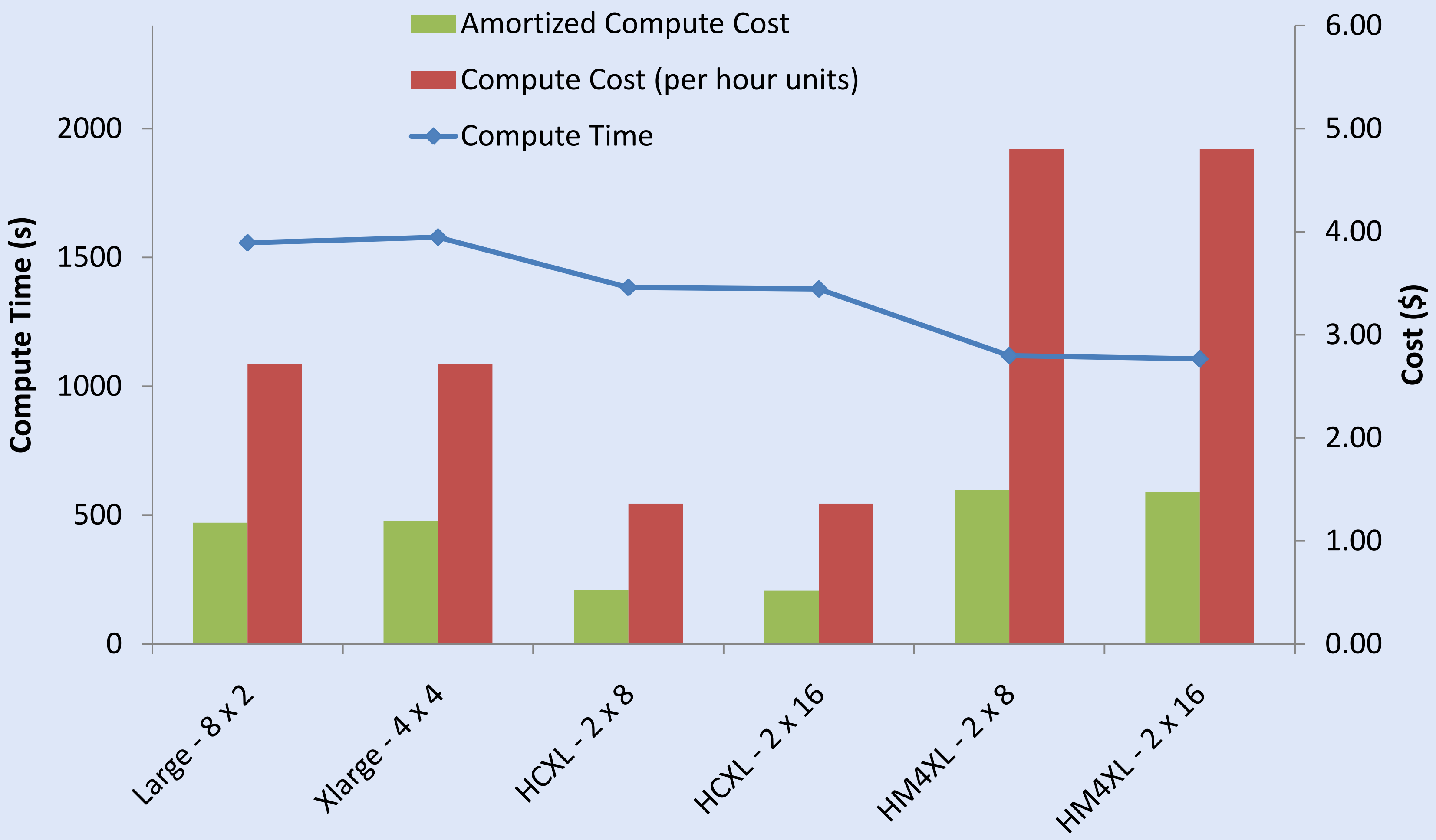




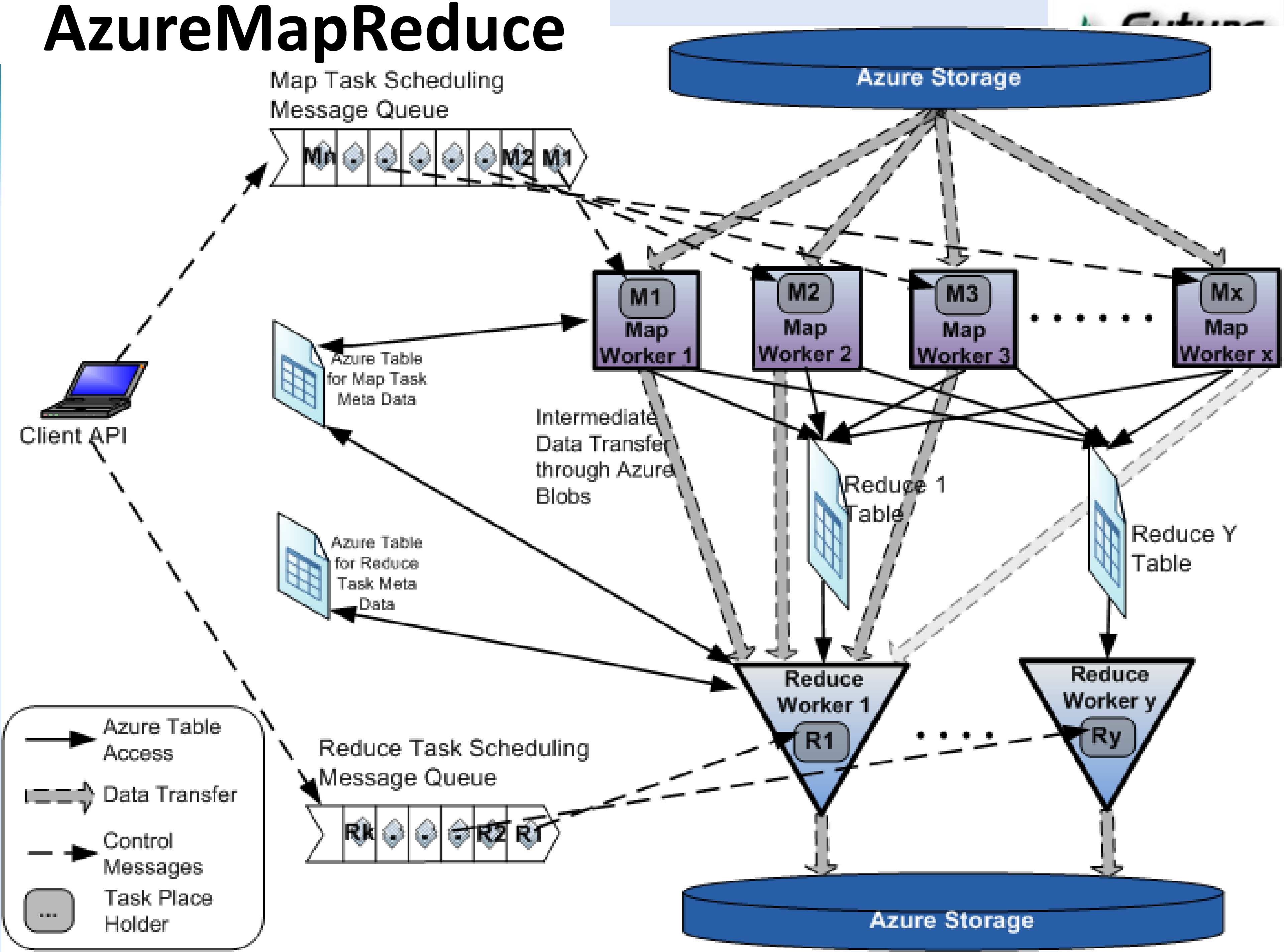
Cap3 Parallel **Efficiency**

Cap3 – Time Per core per file (458 reads in each file) to process sequences

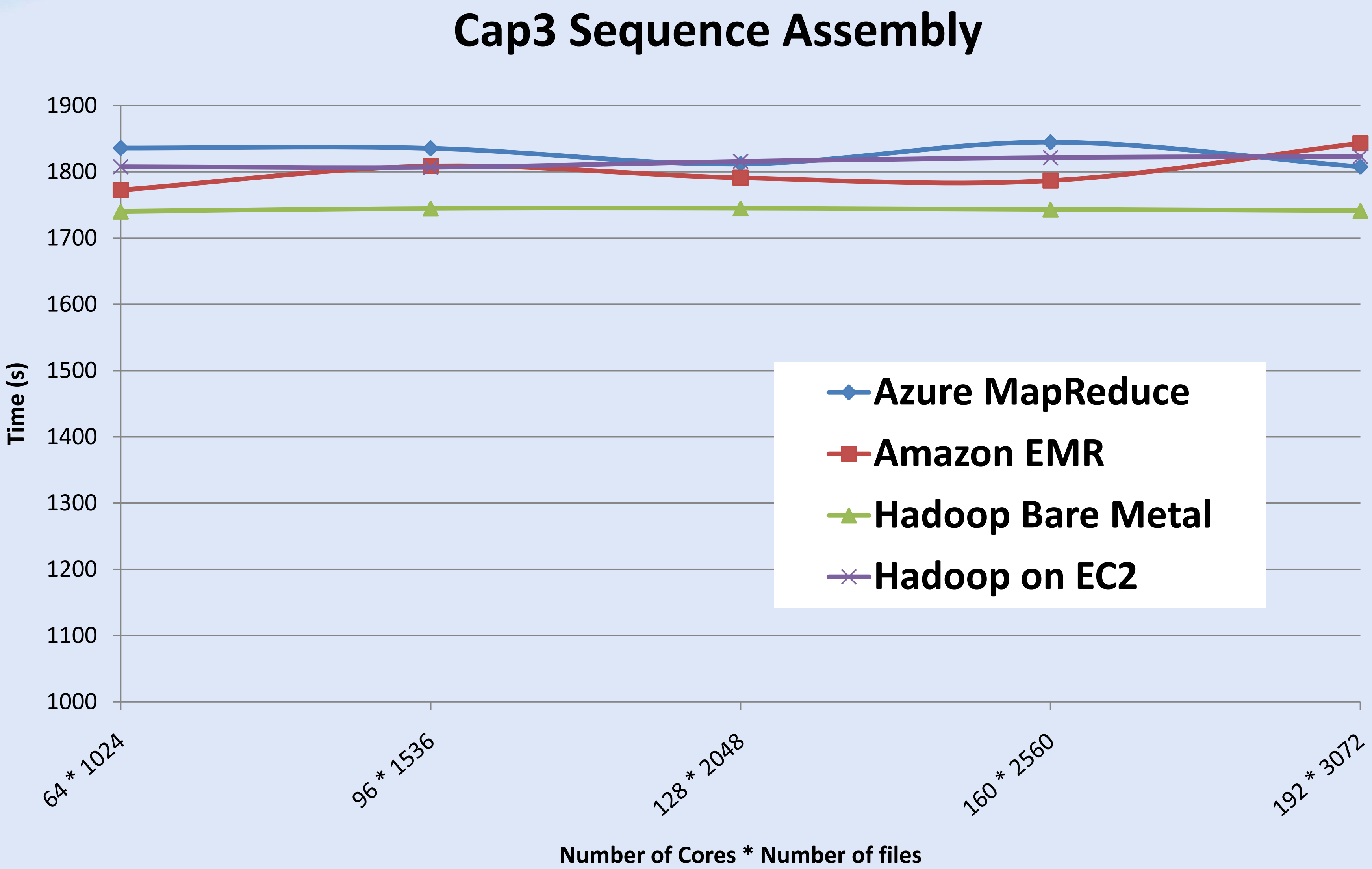
Cap3 Performance with Different EC2 Instance Types



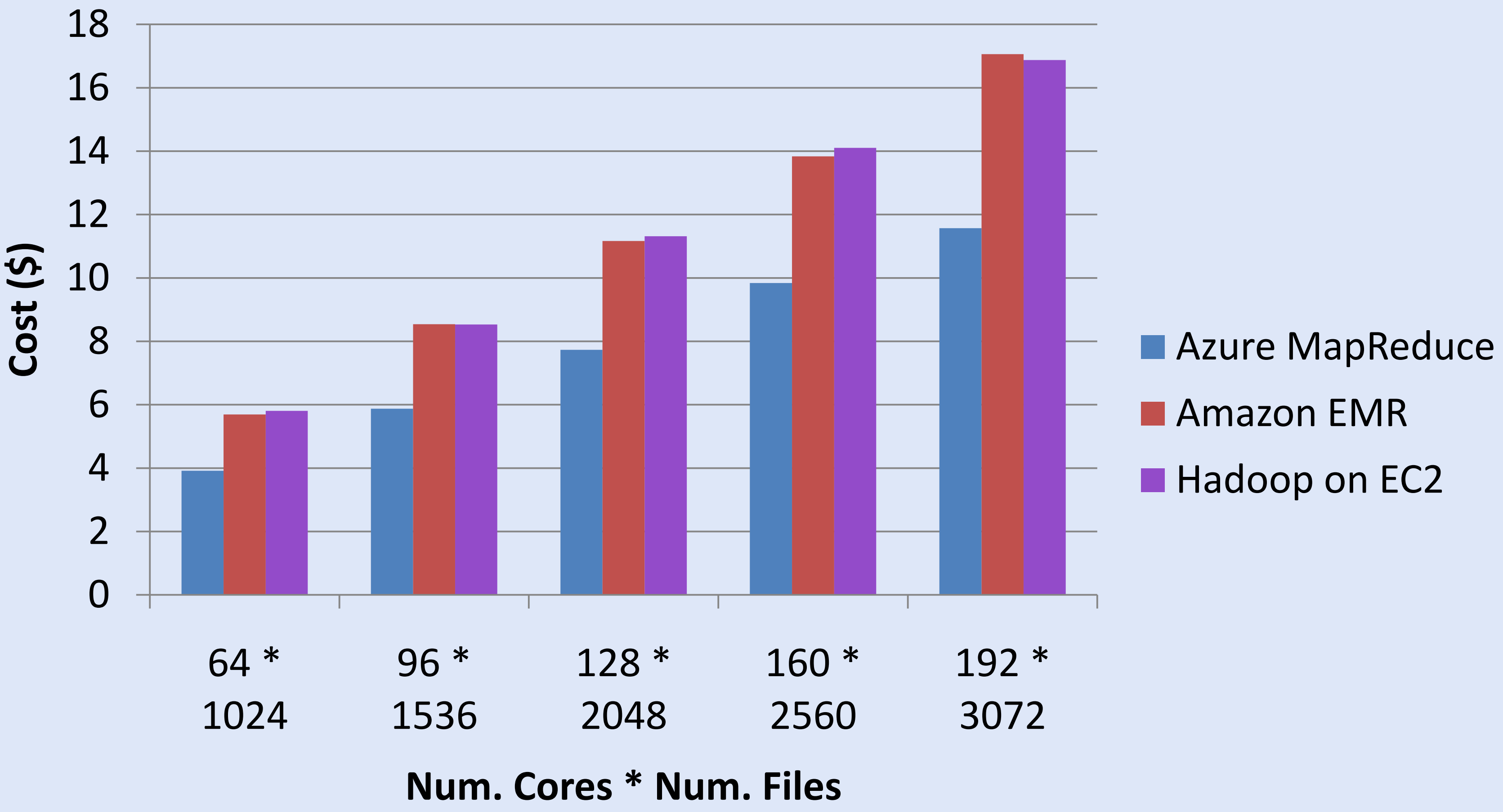
AzureMapReduce



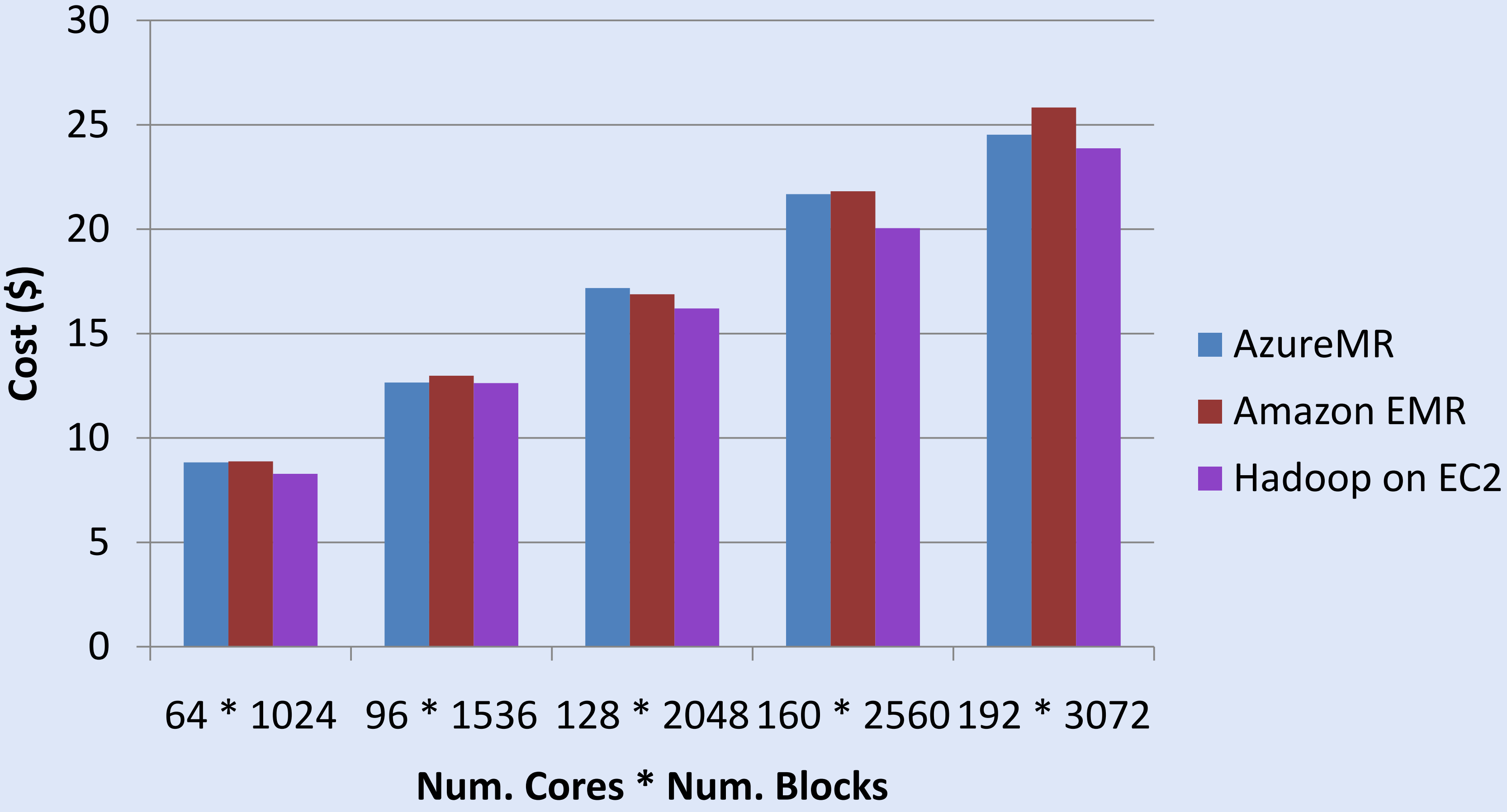
Early Results with Azure/Amazon MapReduce



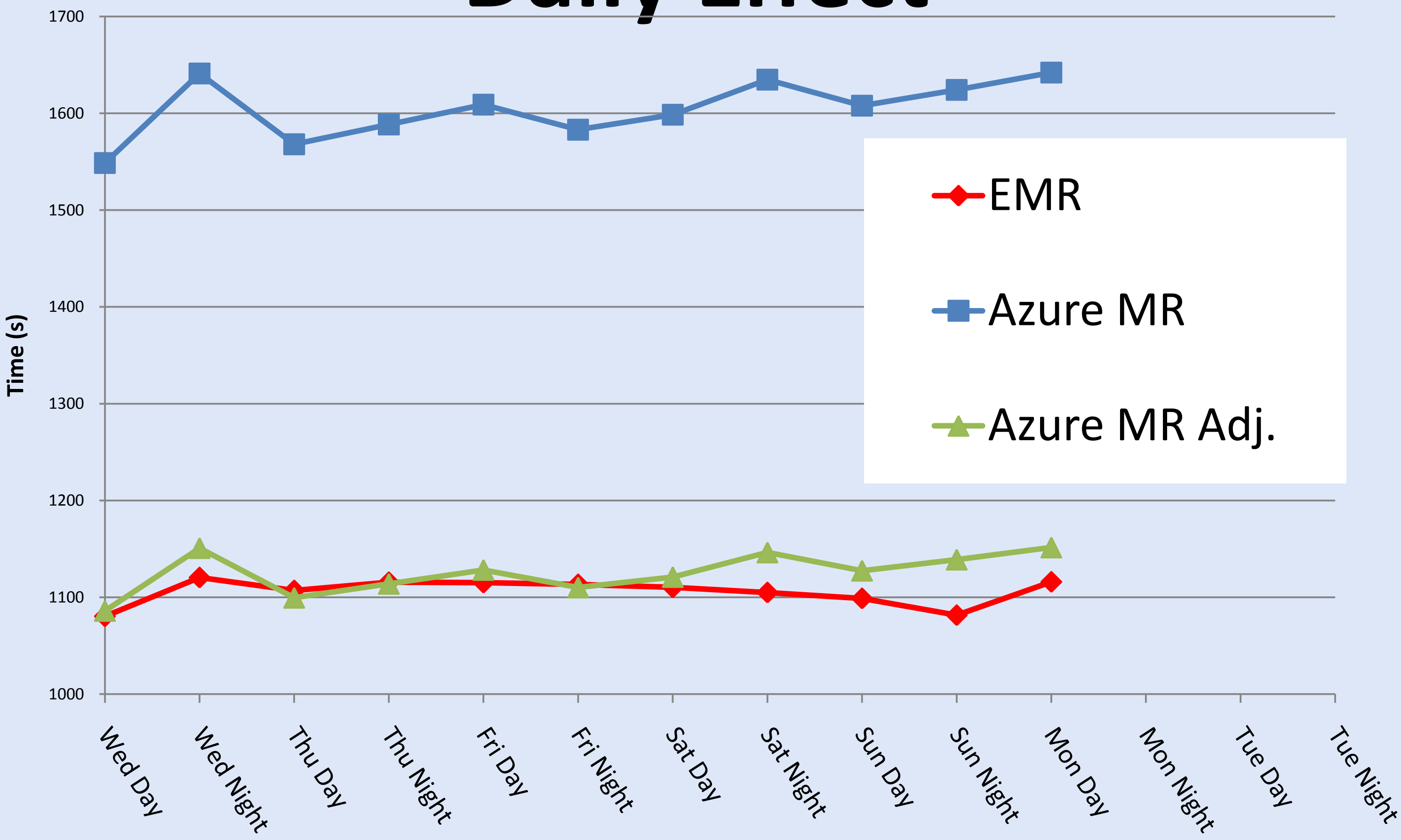
Cap3 Cost


18
16
14
12
10
8
6
4
2
0
Cost ($)
64 * 1024
96 * 1536
128 * 2048
160 * 2560
192 * 3072
Num. Cores * Num. Files
Azure MapReduce
Amazon EMR
Hadoop on EC2

SWG Cost


30
25
20
15
10
5
0
Cost ($)
64 * 1024
96 * 1536
128 * 2048
160 * 2560
192 * 3072
Num. Cores * Num. Blocks
AzureMR
Amazon EMR
Hadoop on EC2

Smith Waterman: Daily Effect



Some Issues with AzureTwister and AzureMapReduce

- Transporting data to Azure: Blobs (HTTP), Drives (GridFTP etc.), Fedex disks
- Intermediate data Transfer: Blobs (current choice) versus Drives (should be faster but don’t seem to be)
- Azure Table v Azure SQL: Handle all metadata
- Messaging Queues: Use real publish-subscribe system in place of Azure Queues
- Azure Affinity Groups: Could allow better data-compute and compute-compute affinity

Research and Clouds I

- Clouds are suitable for “Loosely coupled” data parallel applications
- Quantify “loosely coupled” and define appropriate programming model
- “Map Only” (really pleasingly parallel) certainly run well on clouds (subject to data affinity) with many programming paradigms
- Parallel FFT and adaptive mesh PDA solver probably pretty bad on clouds but suitable for classic MPI engines
- MapReduce and Twister are candidates for “appropriate programming model”
- 1 or 2 iterations (MapReduce) and Iterative with large messages (Twister) are “loosely coupled” applications
- How important is compute-data affinity and concepts like HDFS

Research and Clouds II

- Platforms: exploit Tables as in SHARD (Scalable, High-Performance, Robust and Distributed) Triple Store based on Hadoop
  - What are needed features of tables
- Platforms: exploit MapReduce and its generalizations: are there other extensions that preserve its robust and dynamic structure
  - How important is the loose coupling of MapReduce
  - Are there other paradigms supporting important application classes
- What are other platform features are useful
- Are “academic” private clouds interesting as they (currently) only have a few of Platform features of commercial clouds?
- Long history of search for latency tolerant algorithms for memory hierarchies
  - Are there successes? Are they useful in clouds?
  - In Twister, only support large complex messages
  - What algorithms only need TwisterMPIReduce

Research and Clouds III

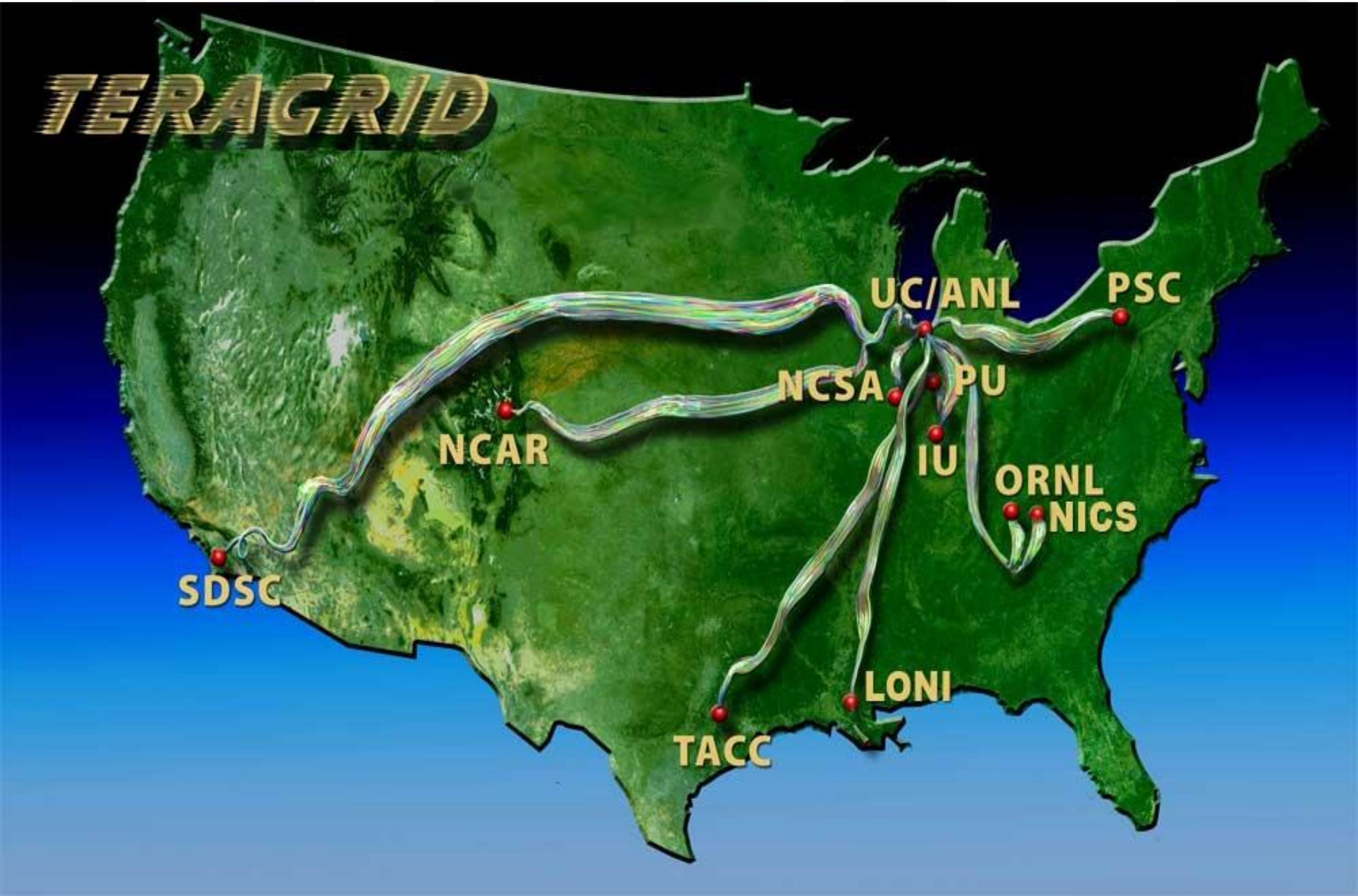
- Can cloud deployment algorithms be devised to support compute-compute and compute-data affinity
- What platform primitives are needed by datamining?
  - Clearer for partial differential equation solution?
- Note clouds have greater impact on programming paradigms than Grids
- Workflow came from Grids and will remain important
  - Workflow is coupling coarse grain functionally distinct components together while MapReduce is data parallel scalable parallelism
- Finding subsets of MPI and algorithms that can use them probably more important than making MPI more complicated
- Note MapReduce can use multicore directly – don’t need hybrid MPI OpenMP Programming models
- Develop Publish-Subscribe optimized for Twister communication

US Cyberinfrastructure Context

- There are a rich set of facilities
  - Production TeraGrid facilities with distributed and shared memory
  - Experimental "Track 2D" Awards
    - FutureGrid: Distributed Systems experiments cf. Grid5000
    - Keeneland: Powerful GPU Cluster
    - Gordon: Large (distributed) Shared memory system with SSD aimed at data analysis/visualization
  - Open Science Grid aimed at High Throughput computing and strong campus bridging

TeraGrid

- ~2 Petaflops; over 20 PetaBytes of storage (disk and tape), over 100 scientific data collections

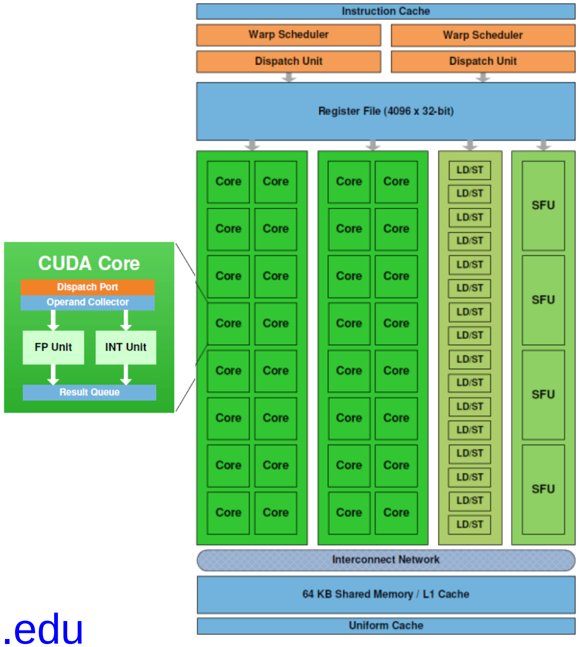


Keeneland – NSF-Funded Partnership to Enable Large-scale Computational Science on Heterogeneous Architectures

- NSF Track 2D System of Innovative Design
  - Georgia Tech
  - UTK NICS
  - ORNL
  - UTK
- Two GPU clusters
  - Initial delivery (~250 CPU, 250 GPU)
    - Being built now; Expected availability is November 2010
  - Full scale (> 500 GPU) – Spring 2012
  - NVIDIA, HP, Intel, Qlogic
- Operations, user support
- Education, Outreach, Training for scientists, students, industry
- Software tools, application development

- Exploit graphics processors to provide extreme performance and energy efficiency

NVIDIA's new Fermi GPU

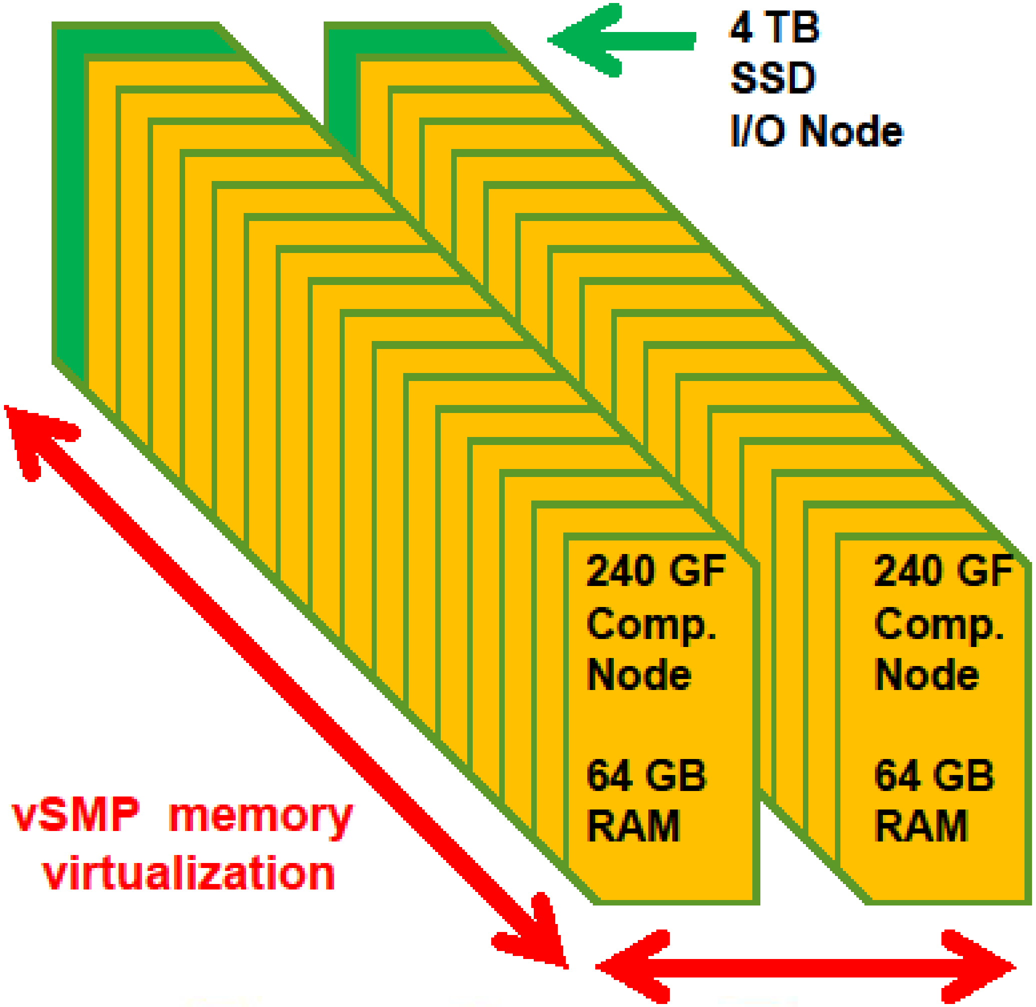


Prof. Jeffrey S. Vetter
Prof. Richard Fujimoto
Prof. Karsten Schwan
Prof. Jack Dongarra
Dr. Thomas Schulthess
Prof. Sudha Yalamanchili
Jeremy Meredith
Dr. Philip Roth
Richard Glassbrook
Patricia Kovatch
Stephen McNally
Dr. Bruce Loftis
Jim Ferguson
James Rogers
Kathlyn Boudwin
Arlene Washington
Many others…

http://keeneland.gatech.edu
http://ft.ornl.gov

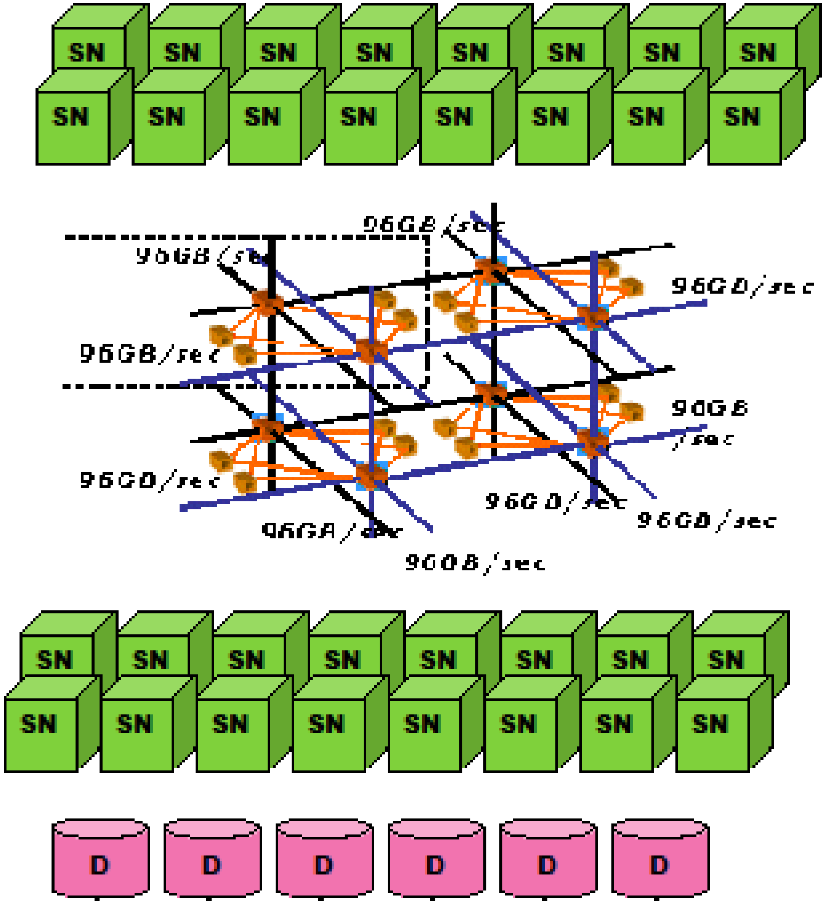
Gordon Architecture: "Supernode"

- **32 Appro Extreme-X compute nodes**
  - Dual processor Intel Sandy Bridge
    - 240 GFLOPS
    - 64 GB
- **2 Appro Extreme-X IO nodes**
  - Intel SSD drives
    - 4 TB ea.
    - 560,000 IOPS
- **ScaleMP vSMP virtual shared memory**
  - 2 TB RAM aggregate
  - 8 TB SSD aggregate



Gordon Architecture: Full Machine

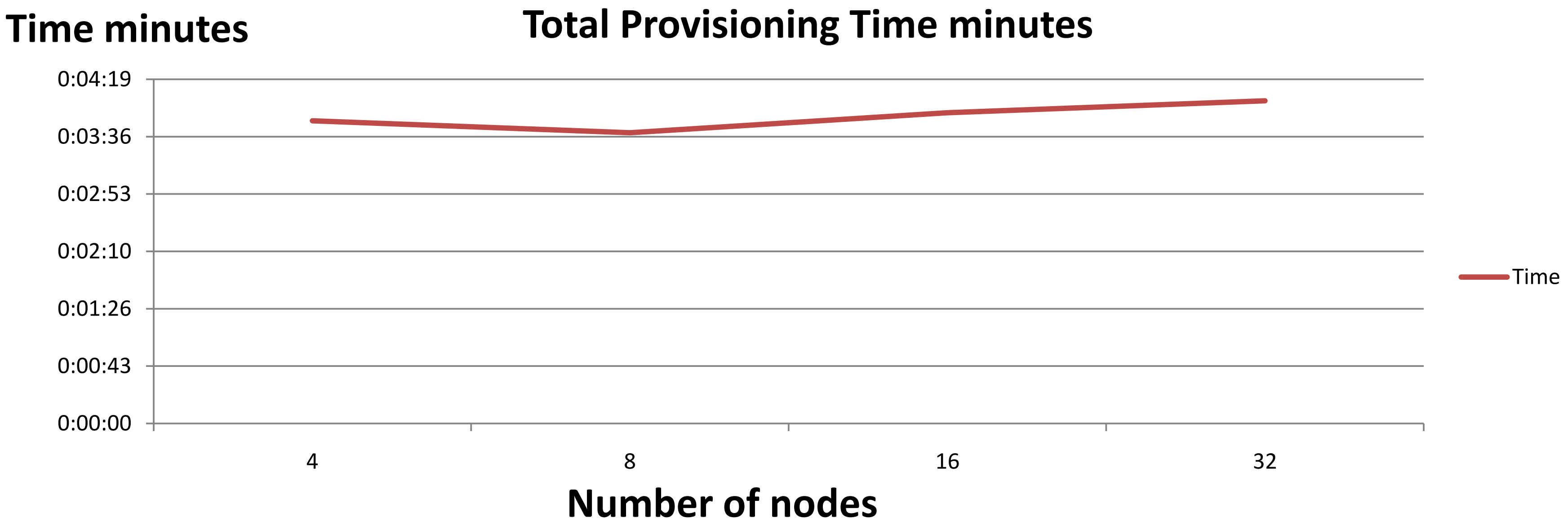
- **32 supernodes = 1024 compute nodes**
- **Dual rail QDR Infiniband network**
  - 3D torus (4x4x4)
- **4 PB rotating disk parallel file system**
  - >100 GB/s



FutureGrid key Concepts I

- FutureGrid is an international testbed modeled on Grid5000
- Rather than loading images onto VM's, FutureGrid supports Cloud, Grid and Parallel computing environments by dynamically provisioning software as needed onto "bare-metal" using Moab/xCAT
  - Image library for MPI, OpenMP, Hadoop, Dryad, gLite, Unicore, Globus, Xen, ScaleMP (distributed Shared Memory), Nimbus, Eucalyptus, OpenNebula, KVM, Windows …..
- The FutureGrid testbed provides to its users:
  - A flexible development and testing platform for middleware and application users looking at interoperability, functionality and performance
  - Each use of FutureGrid is an experiment that is reproducible
  - A rich education and teaching platform for advanced cyberinfrastructure classes
- Growth comes from users depositing novel images in library

Dynamic Provisioning Results



Time elapsed between requesting a job and the jobs reported start time on the provisioned node. The numbers here are an average of 2 sets of experiments.

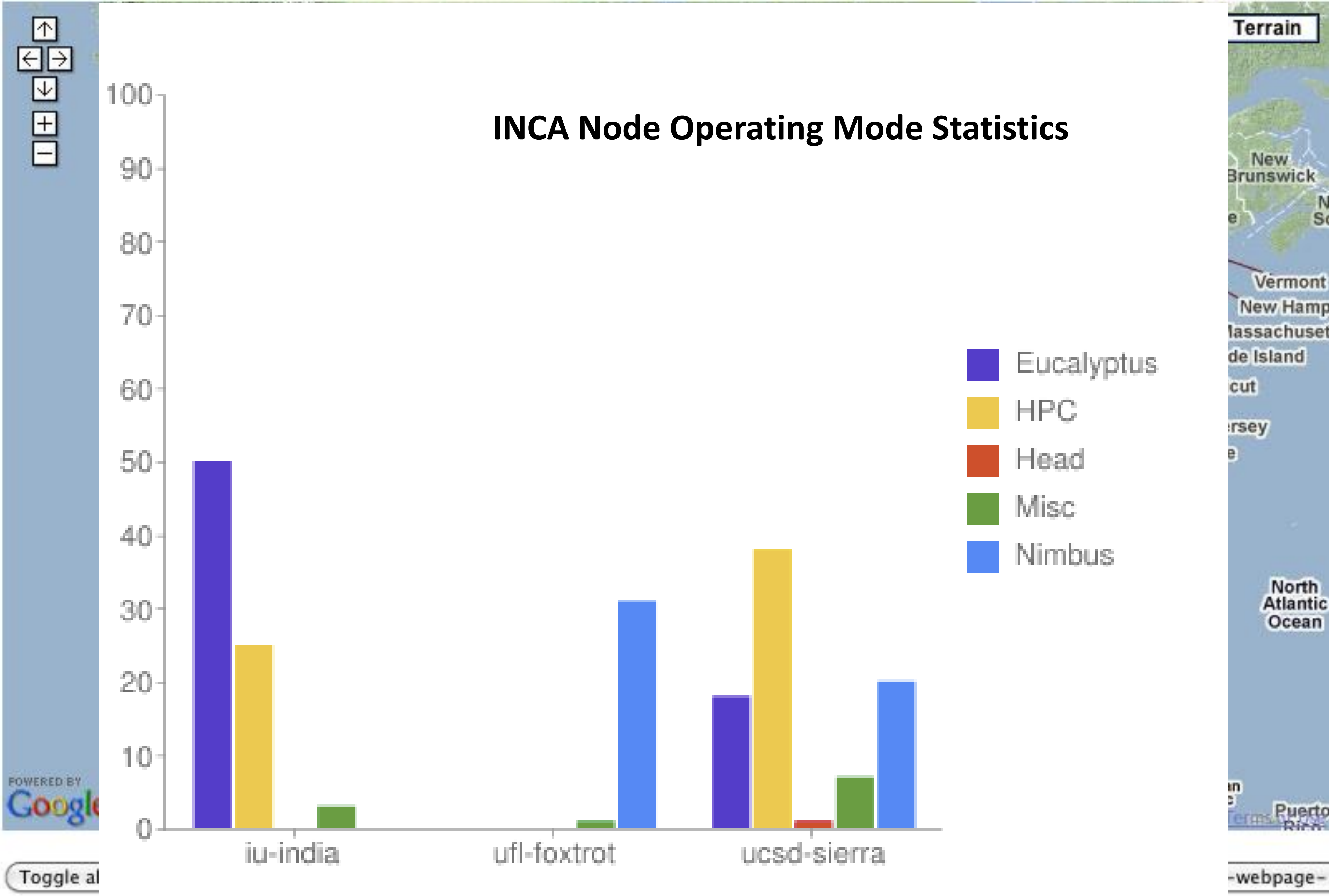
FutureGrid key Concepts II

- Support Computer Science and Computational Science
  - Industry and Academia
  - Asia, Europe and Americas
- FutureGrid has ~5000 distributed cores with a dedicated network and a Spirent XGEM network fault and delay generator
- **Key early user oriented milestones:**
  - **June 2010** Initial users
  - **November 2010-September 2011** Increasing number of users allocated by FutureGrid
  - **October 2011** FutureGrid allocatable via TeraGrid process
  - 3 classes using FutureGrid this fall
- Apply **now** to use FutureGrid on web site www.futuregrid.org

FutureGrid Partners

- Indiana University (Architecture, core software, Support)
  - Collaboration between research and infrastructure groups
- Purdue University (HTC Hardware)
- San Diego Supercomputer Center at University of California San Diego (INCA, Monitoring)
- University of Chicago/Argonne National Labs (Nimbus)
- University of Florida (ViNE, Education and Outreach)
- University of Southern California Information Sciences (Pegasus to manage experiments)
- University of Tennessee Knoxville (Benchmarking)
- University of Texas at Austin/Texas Advanced Computing Center (Portal)
- University of Virginia (OGF, Advisory Board and allocation)
- Center for Information Services and GWT-TUD from Technische Universtität Dresden. (VAMPIR)
- Red institutions have FutureGrid hardware

FutureGrid: a Grid/Cloud/HPC



INCA Node Operating Mode Statistics

Network & Internal Interconnects

- FutureGrid has dedicated network (except to TACC) and a network fault and delay generator
- Can isolate experiments on request; IU runs Network for NLR/Internet2
- (Many) additional partner machines will run FutureGrid software and be supported (but allocated in specialized ways)

Machine	Name	Internal Network
IU Cray	xray	Cray 2D Torus SeaStar
IU iDataPlex	india	DDR IB, QLogic switch with Mellanox ConnectX adapters Blade Network Technologies & Force10 Ethernet switches
SDSC iDataPlex	sierra	DDR IB, Cisco switch with Mellanox ConnectX adapters Juniper Ethernet switches
UC iDataPlex	hotel	DDR IB, QLogic switch with Mellanox ConnectX adapters Blade Network Technologies & Juniper switches
UF iDataPlex	foxtrot	Gigabit Ethernet only (Blade Network Technologies; Force10 switches)
TACC Dell	alamo	QDR IB, Mellanox switches and adapters Dell Ethernet switches

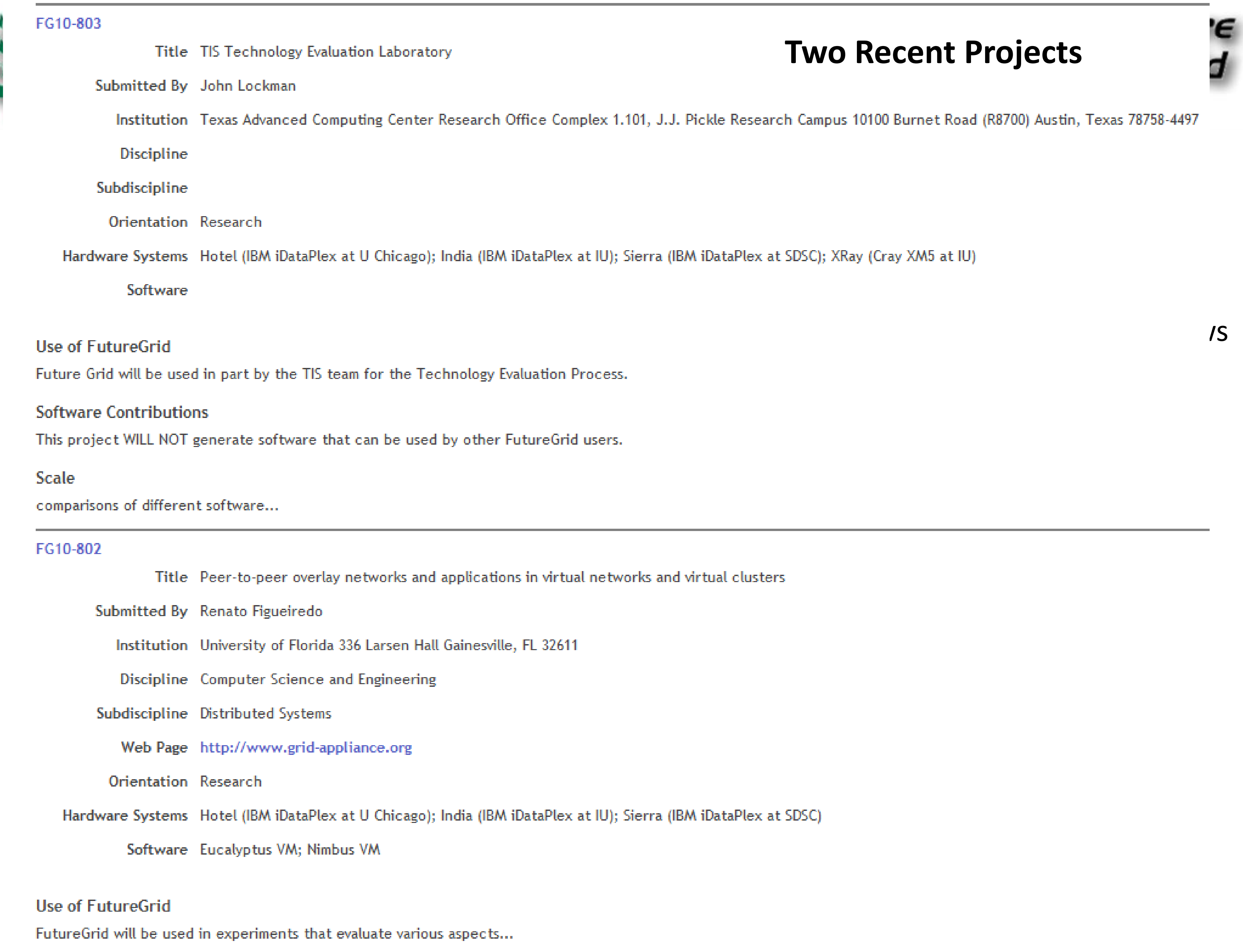
Network Impairment Device

- Spirent XGEM Network Impairments Simulator for jitter, errors, delay, etc
- Full Bidirectional 10G w/64 byte packets
- up to 15 seconds introduced delay (in 16ns increments)
- 0-100% introduced packet loss in .0001% increments
- Packet manipulation in first 2000 bytes
- up to 16k frame size
- TCL for scripting, HTML for manual configuration

FutureGrid Usage Model

- The goal of FutureGrid is to support the research on the future of distributed, grid, and cloud computing
- FutureGrid will build a robustly managed simulation environment and test-bed to support the development and early use in science of new technologies at all levels of the software stack: from networking to middleware to scientific applications
- The environment will mimic TeraGrid and/or general parallel and distributed systems – FutureGrid is part of TeraGrid (but not part of formal TeraGrid process for first two years)
  - Supports Grids, Clouds, and classic HPC
  - It will mimic commercial clouds (initially IaaS not PaaS)
  - Expect FutureGrid PaaS to grow in importance
- FutureGrid can be considered as a (small ~5000 core) Science/Computer Science Cloud but it is more accurately a virtual machine or bare-metal based simulation environment
- This test-bed will succeed if it enables major advances in science and engineering through collaborative development of science applications and related software

Two Recent Projects



FG10-803

Title	TIS Technology Evaluation Laboratory
Submitted By	John Lockman
Institution	Texas Advanced Computing Center Research Office Complex 1.101, J.J. Pickle Research Campus 10100 Burnet Road (R8700) Austin, Texas 78758-4497
Discipline	
Subdiscipline	
Orientation	Research
Hardware Systems	Hotel (IBM iDataPlex at U Chicago); India (IBM iDataPlex at IU); Sierra (IBM iDataPlex at SDSC); XRay (Cray XM5 at IU)
Software	

Use of FutureGrid

Future Grid will be used in part by the TIS team for the Technology Evaluation Process.

Software Contributions

This project WILL NOT generate software that can be used by other FutureGrid users.

Scale

comparisons of different software...

FG10-802

Title	Peer-to-peer overlay networks and applications in virtual networks and virtual clusters
Submitted By	Renato Figueiredo
Institution	University of Florida 336 Larsen Hall Gainesville, FL 32611
Discipline	Computer Science and Engineering
Subdiscipline	Distributed Systems
Web Page	http://www.grid-appliance.org
Orientation	Research
Hardware Systems	Hotel (IBM iDataPlex at U Chicago); India (IBM iDataPlex at IU); Sierra (IBM iDataPlex at SDSC)
Software	Eucalyptus VM; Nimbus VM

Use of FutureGrid

FutureGrid will be used in experiments that evaluate various aspects...

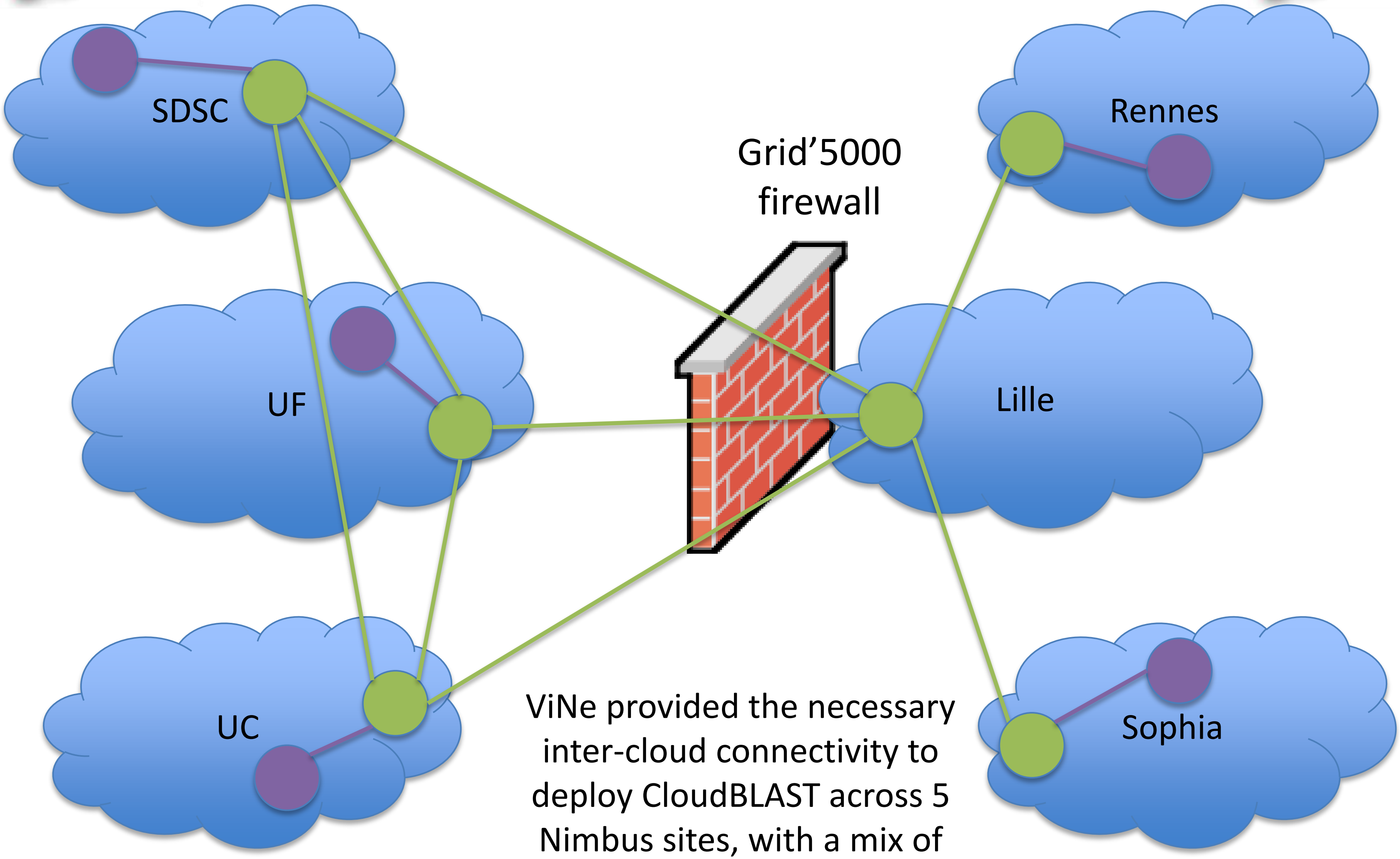
IS

Grid Interoperability from Andrew Grimshaw

- Colleagues,
- FutureGrid has as two of its many goals the creation of a Grid middleware testing and interoperability testbed as well as the maintenance of standards compliant endpoints against which experiments can be executed. We at the University of Virginia are tasked with bringing up three stacks as well as maintaining standard-endpoints against which these experiments can be run.
- We currently have UNICORE 6 and Genesis II endpoints functioning on X-Ray (a Cray). Over the next few weeks we expect to bring two additional resources, India and Sierra (essentially Linux clusters), on-line in a similar manner (Genesis II is already up on Sierra). As called for in the FutureGrid program execution plan, once those two stacks are operational we will begin to work on g-lite (with help we may be able to accelerate that). Other standards-compliant endpoints are welcome in the future , but not part of the current funding plan.

We RENKEI/NAREGI project are interested in the participation for interoperation demonstrations or projects for OGF and SC. We have a prototype middleware which can submit and receive jobs using the HPCBP specification. Can we have more detailed information of your endpoints(authentication, data staging method, and so on) and the participation conditions of the demonstrations/projects.

OGF'10 Demo



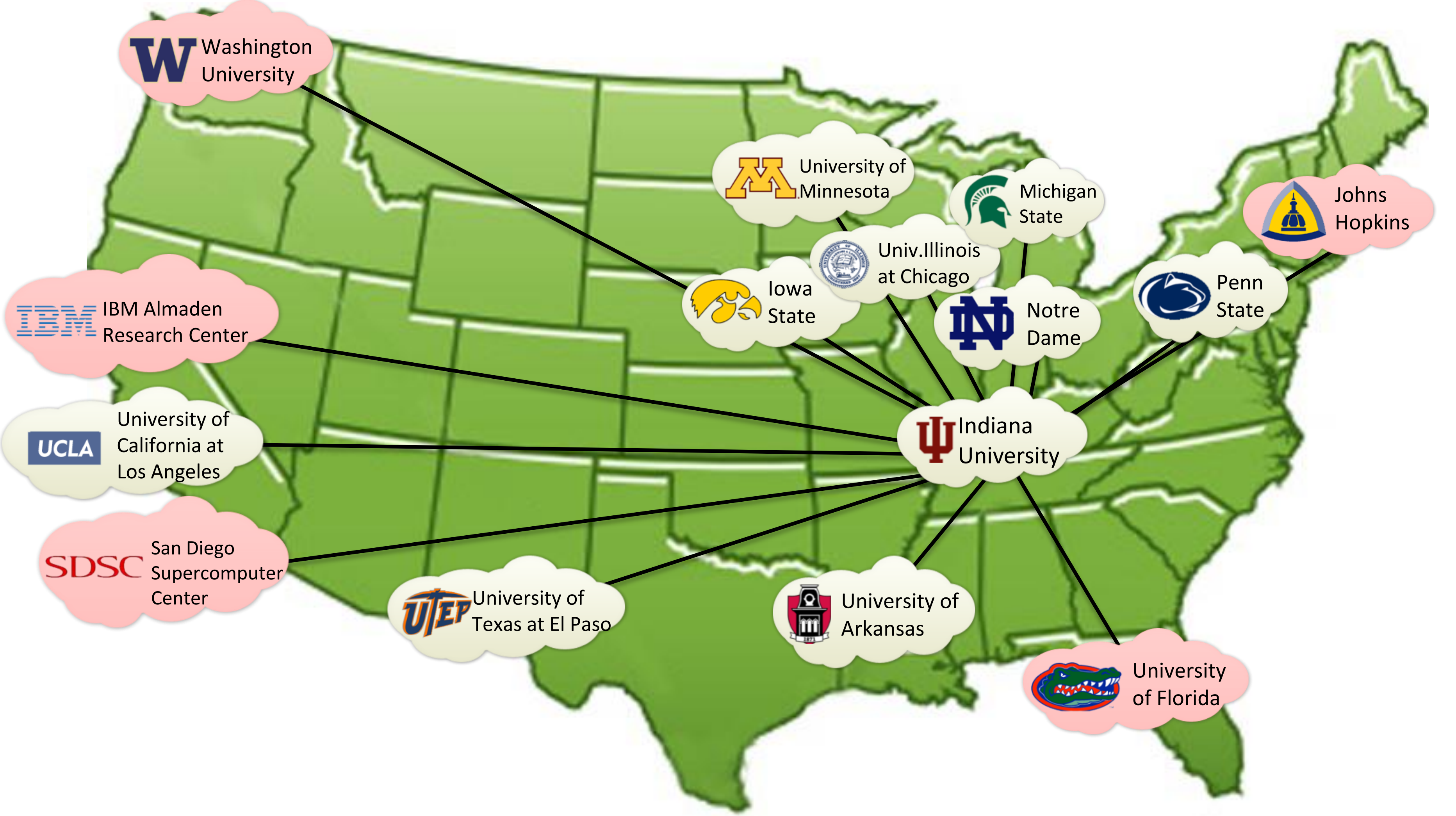
ViNe provided the necessary inter-cloud connectivity to deploy CloudBLAST across 5 Nimbus sites, with a mix of public and private subnets.

Big Data for Science

300+ Students learning about Twister & Hadoop MapReduce technologies, supported by FutureGrid.

July 26-30, 2010 NCSA Summer School Workshop

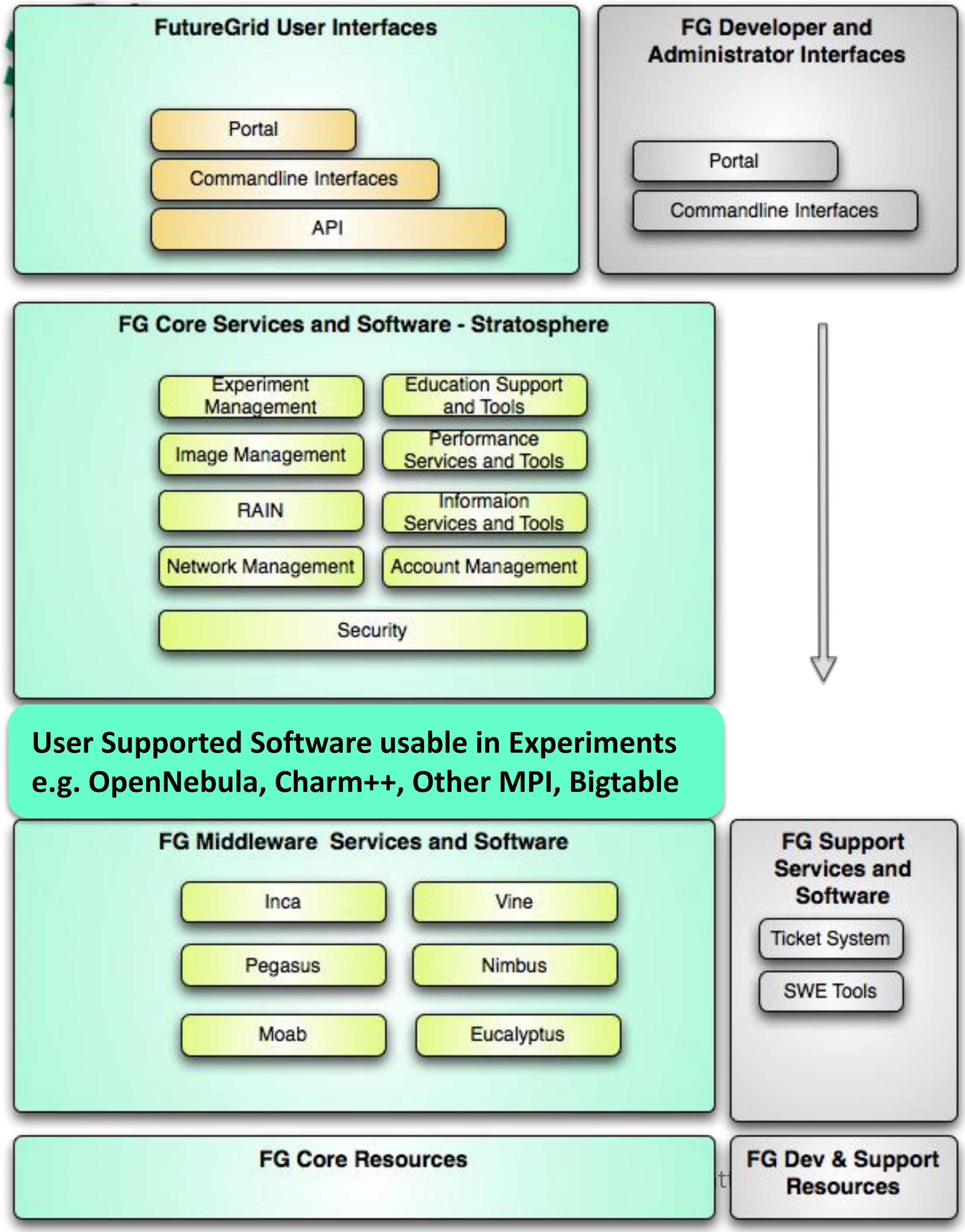
http://salsahpc.indiana.edu/tutorial



Software Components

- Portals including “Support” “use FutureGrid” “Outreach”
- Monitoring – INCA, Power (GreenIT)
- **Experiment Manager**: specify/workflow
- Image Generation and Repository
- Intercloud Networking ViNE
- Virtual Clusters built with virtual networks
- Performance library
- **Rain** or **R**untime **A**daptable **I**nsertio**N** Service: Schedule and Deploy images
- Security (including use of isolated network), Authentication, Authorization,

FutureGrid Layered Software Stack



FutureGrid User Interfaces

Portal

Commandline Interfaces

API

FG Developer and Administrator Interfaces

Portal

Commandline Interfaces

FG Core Services and Software - Stratosphere

Experiment Management

Education Support and Tools

Image Management

Performance Services and Tools

RAIN

Informaion Services and Tools

Network Management

Account Management

Security

User Supported Software usable in Experiments e.g. OpenNebula, Charm++, Other MPI, Bigtable

FG Middleware Services and Software

Inca

Vine

Pegasus

Nimbus

Moab

Eucalyptus

FG Support Services and Software

Ticket System

SWE Tools

FG Core Resources

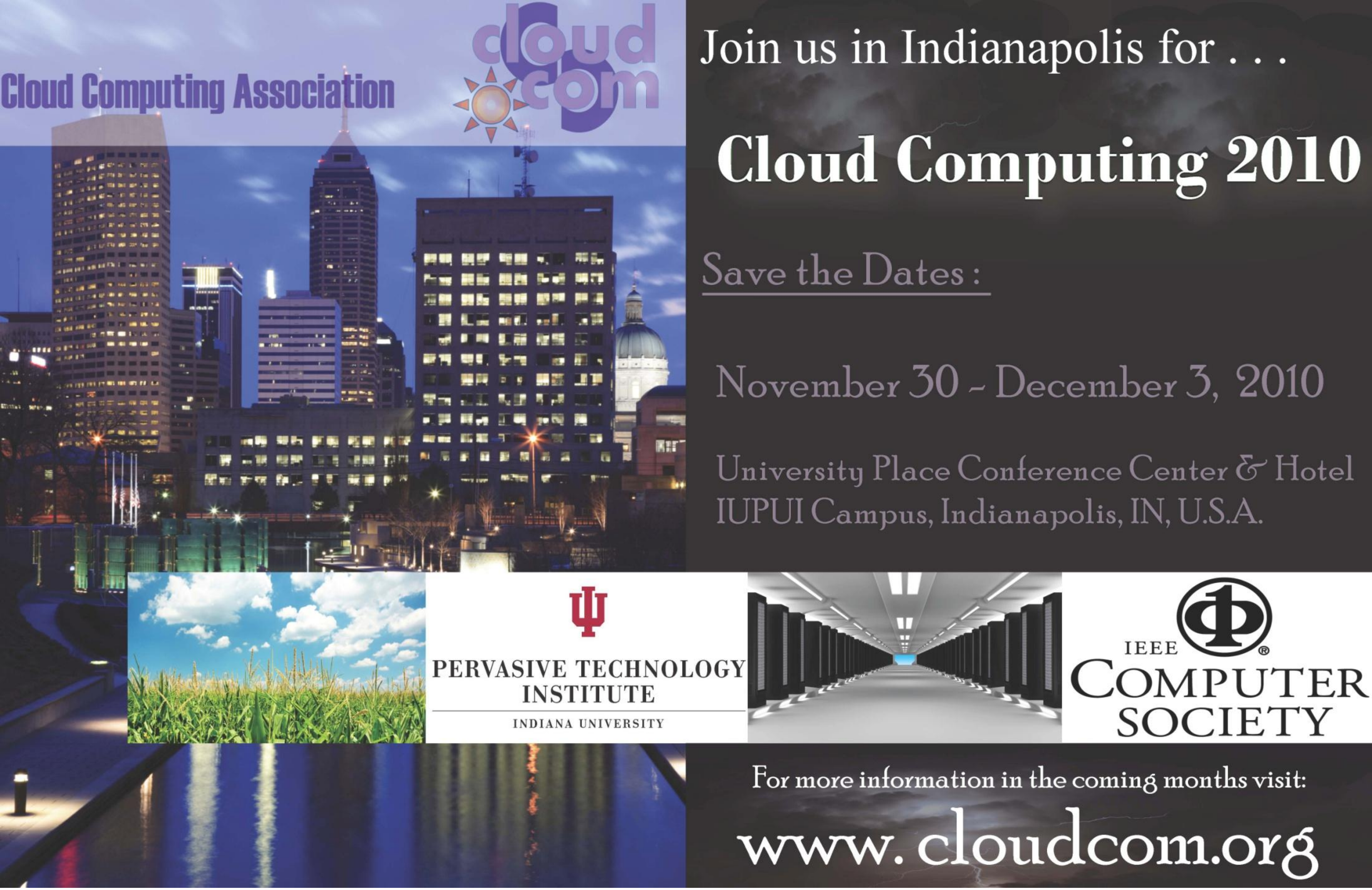
FG Dev & Support Resources

FutureGrid Interaction with Commercial Clouds

- We support experiments that link Commercial Clouds and FutureGrid with one or more workflow environments and portal technology installed to link components across these platforms
- We support environments on FutureGrid that are similar to Commercial Clouds and natural for performance and functionality comparisons
  - These can both be used to prepare for using Commercial Clouds and as the most likely starting point for porting to them
  - One example would be support of MapReduce-like environments on FutureGrid including Hadoop on Linux and Dryad on Windows HPCS which are already part of FutureGrid portfolio of supported software
- We develop expertise and support porting to Commercial Clouds from other Windows or Linux environments
- We support comparisons between and integration of multiple commercial Cloud environments – especially Amazon and Azure in the immediate future
- We develop tutorials and expertise to help users move to Commercial Clouds from other environments

FutureGrid Viral Growth Model

- Users apply for a project
- Users improve/develop some software in project
- This project leads to new images which are placed in FutureGrid repository
- Project report and other web pages document use of new images
- Images are used by other users
- And so on ad infinitum ………



Cloud Computing Association

cloud com

Join us in Indianapolis for . . .

Cloud Computing 2010

Save the Dates :

November 30 – December 3, 2010

University Place Conference Center & Hotel
IUPUI Campus, Indianapolis, IN, U.S.A.

PERVASIVE TECHNOLOGY INSTITUTE
INDIANA UNIVERSITY

IEEE COMPUTER SOCIETY

For more information in the coming months visit:

www.cloudcom.org

194 papers submitted to main track; 48 accepted; 4 days of tutorials