Twister4Azure
Iterative MapReduce for Azure Cloud

Twister4Azure: Parallel Data Analytics on Azure

Judy Qiu Thilina Gunarathne

SALSA HPC Group

http://salsahpc.indiana.edu

School of Informatics and Computing

Indiana University

CAREER Award

Microsoft

Outline

- Iterative Mapreduce Programming Model
- Interoperability
- Reproducibility

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

Big Data for Science

Home Tutorials Contact

Cloud Computing

Keynote: Distributed Data-Parallel Computing
- Powerpoint Link
- Sector/Sphere Tutorial
- Downloadable Link

Overview of FutureGrid
- Powerpoint Link
- Downloadable Link

Plug-and-play virtual appliance clusters running Hadoop
- Powerpoint Link
- Downloadable Link

Overview of Cloud Computing Platforms
- Powerpoint Link
- Downloadable Link

Introduction to Azure
- Powerpoint Link
- Downloadable Link

AzureMapReduce
- Powerpoint Link
- Downloadable Link

An Introduction to DryadLINQ
- Powerpoint Link
- Downloadable Link

Introduction to Amazon EC2
- Powerpoint Link
- Downloadable Link

Data

Opening Keynote: Data-intensive Computing
- Powerpoint Link
- Downloadable Link

Making the most of the I/O Software Stack
- Powerpoint Link
- Downloadable Link

Data movement & Storage (Data Capacitor WAN Filesystem)
- Powerpoint Link
- Downloadable Link

Data Transport (With Specific TG Examples) and File Systems
- Powerpoint Link
- Downloadable Link

Scalable and Distributed Visualization using Paraview
- Powerpoint Link
- Downloadable Link

Science

Studying Science from Large-Scale Usage Data
- Powerpoint Link
- Downloadable Link

Big Data in Drug Discovery
- Powerpoint Link
- Downloadable Link

Cancer epigenomics study using the next generation sequencing data
- Powerpoint Link
- Downloadable Link

Virtual Observatory Technologies
- Powerpoint Link
- Downloadable Link

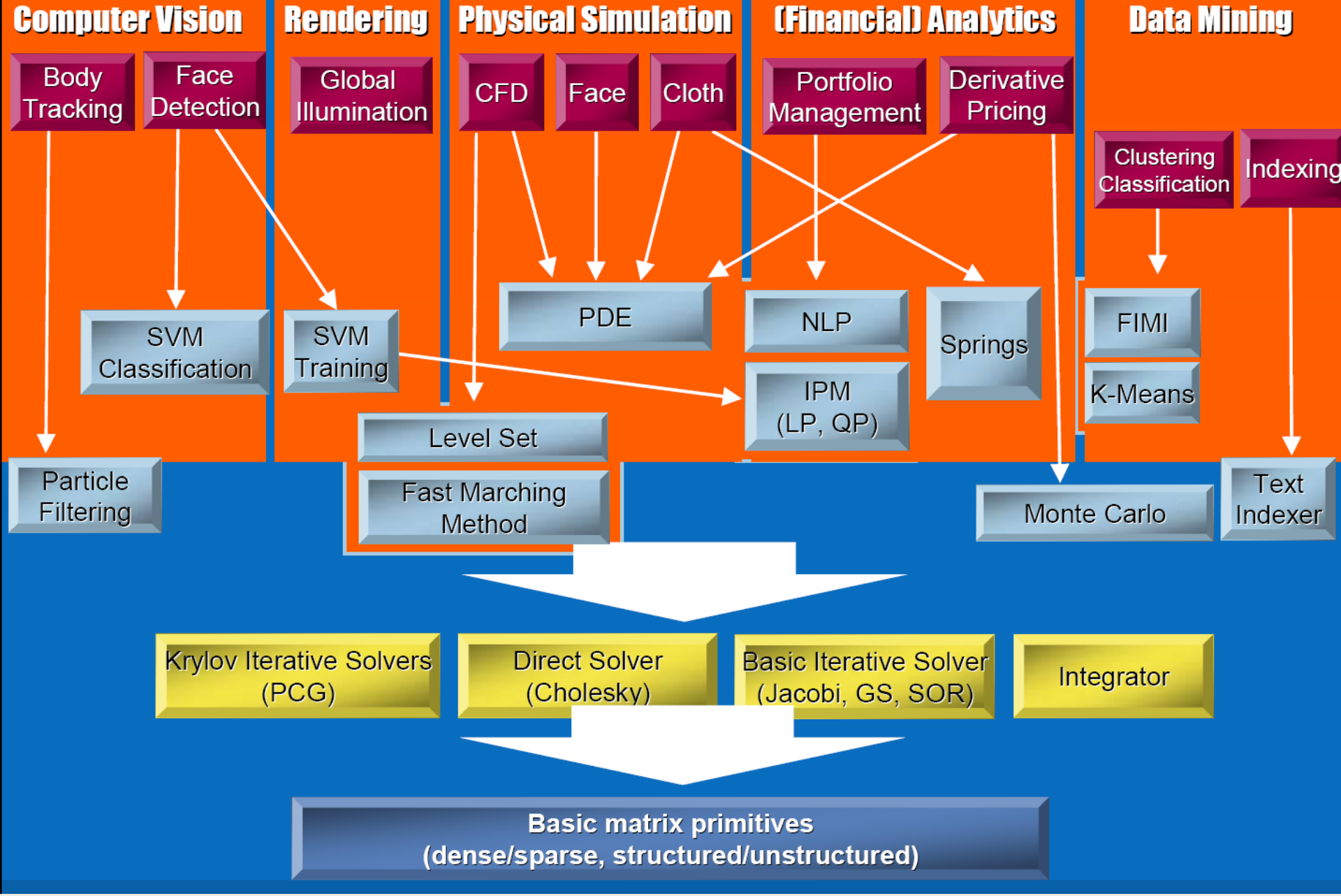
Hands-On

Tutorial on using FutureGrid
- Powerpoint Link
- FutureGrid Machine Access

Introductary Tutorial on MapReduce and Hadoop
- Powerpoint Link
- Hadoop
- Prerequisites & Resources

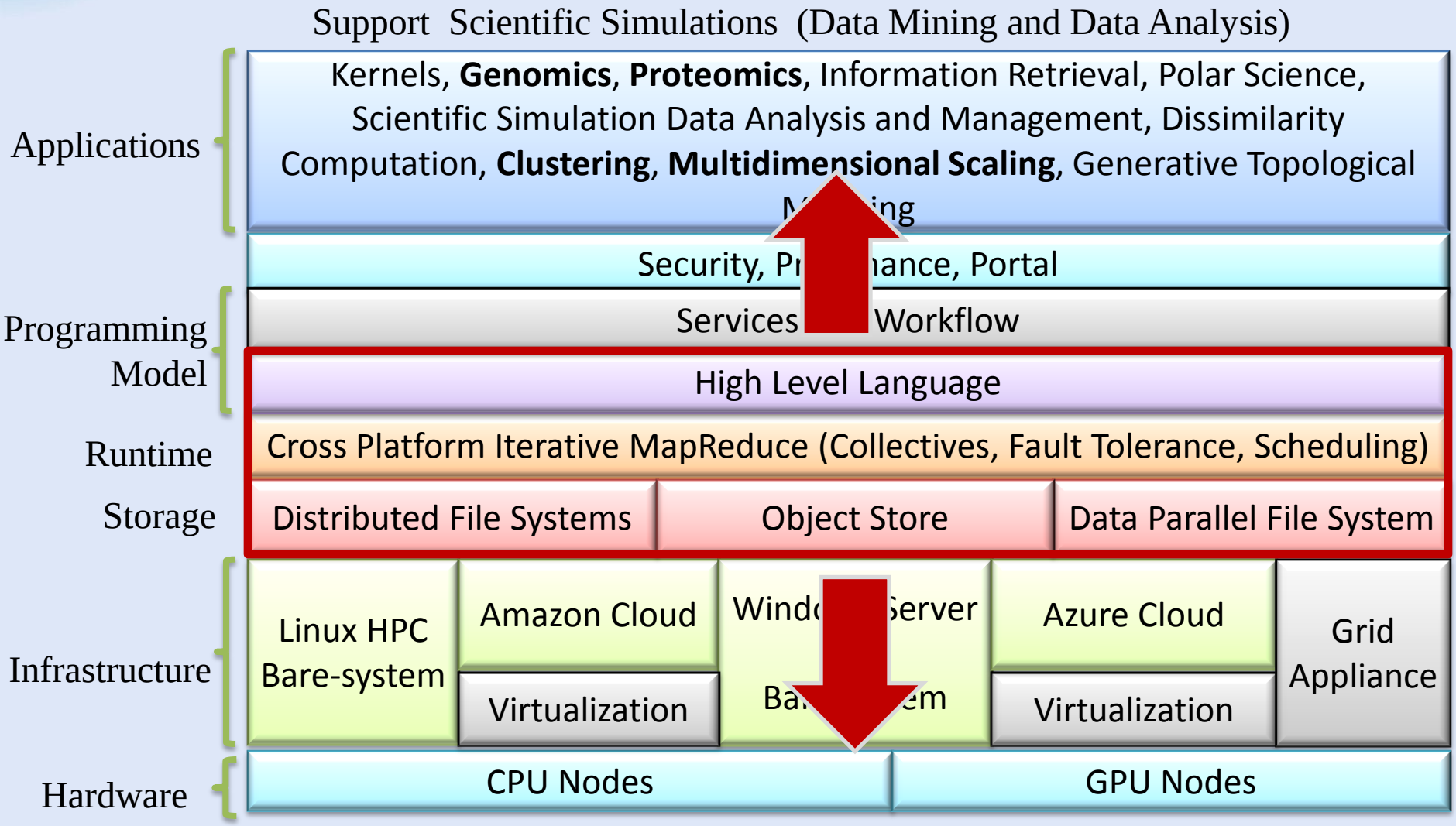
Tutorial on Iterative MapReduce
- Powerpoint Link
- Twister Tutorials
- Prerequisites & Resources

Tutorial on DryadLINQ
- Powerpoint Link
- DryadLINQ Tutorials
- Download

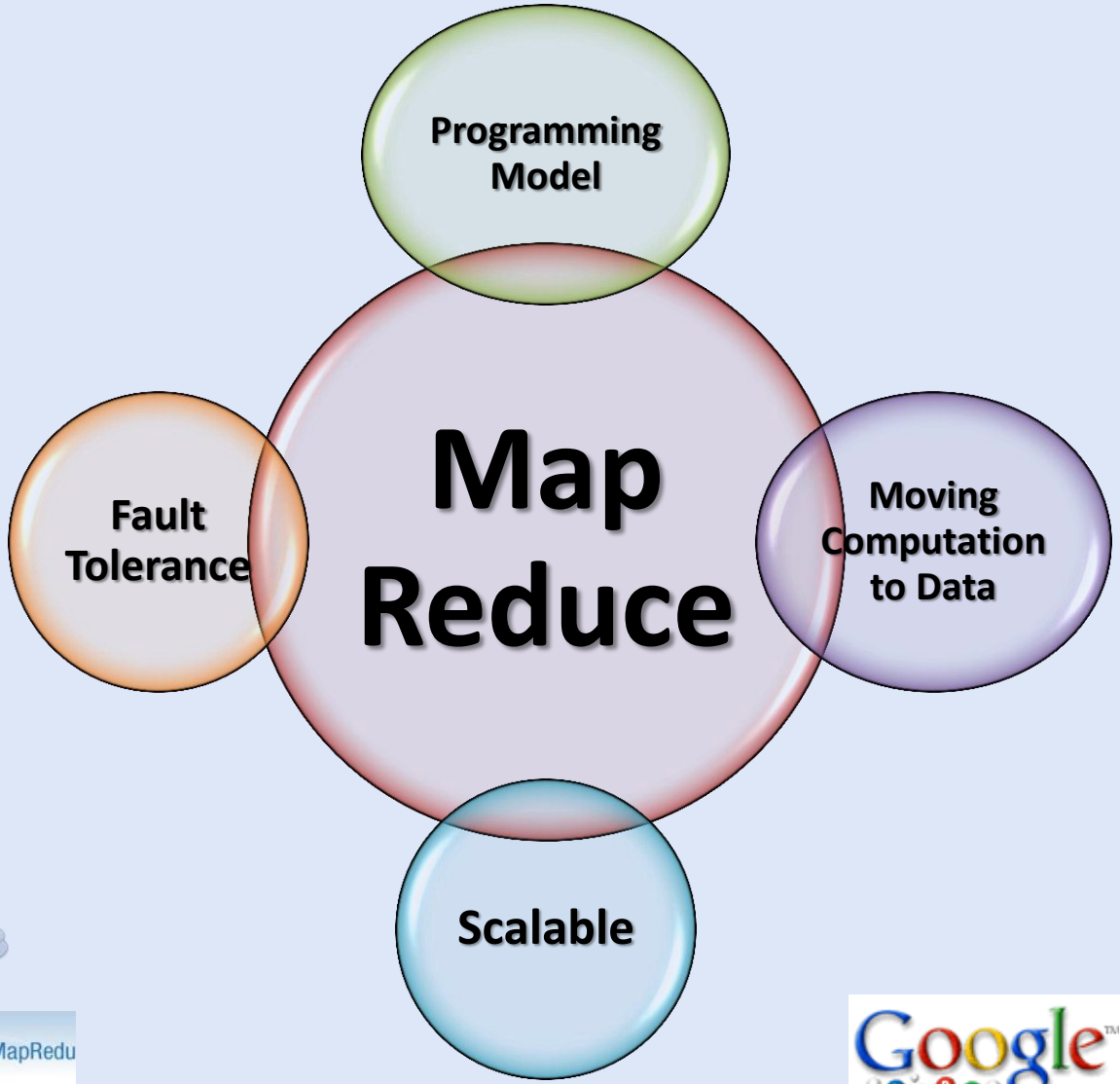


Intel's Application Stack

(Iterative) MapReduce in Context

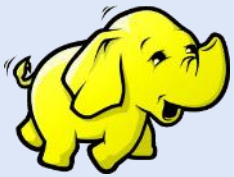


SALSA



Twister4Azure

amazon web services Amazon Elastic MapRedu

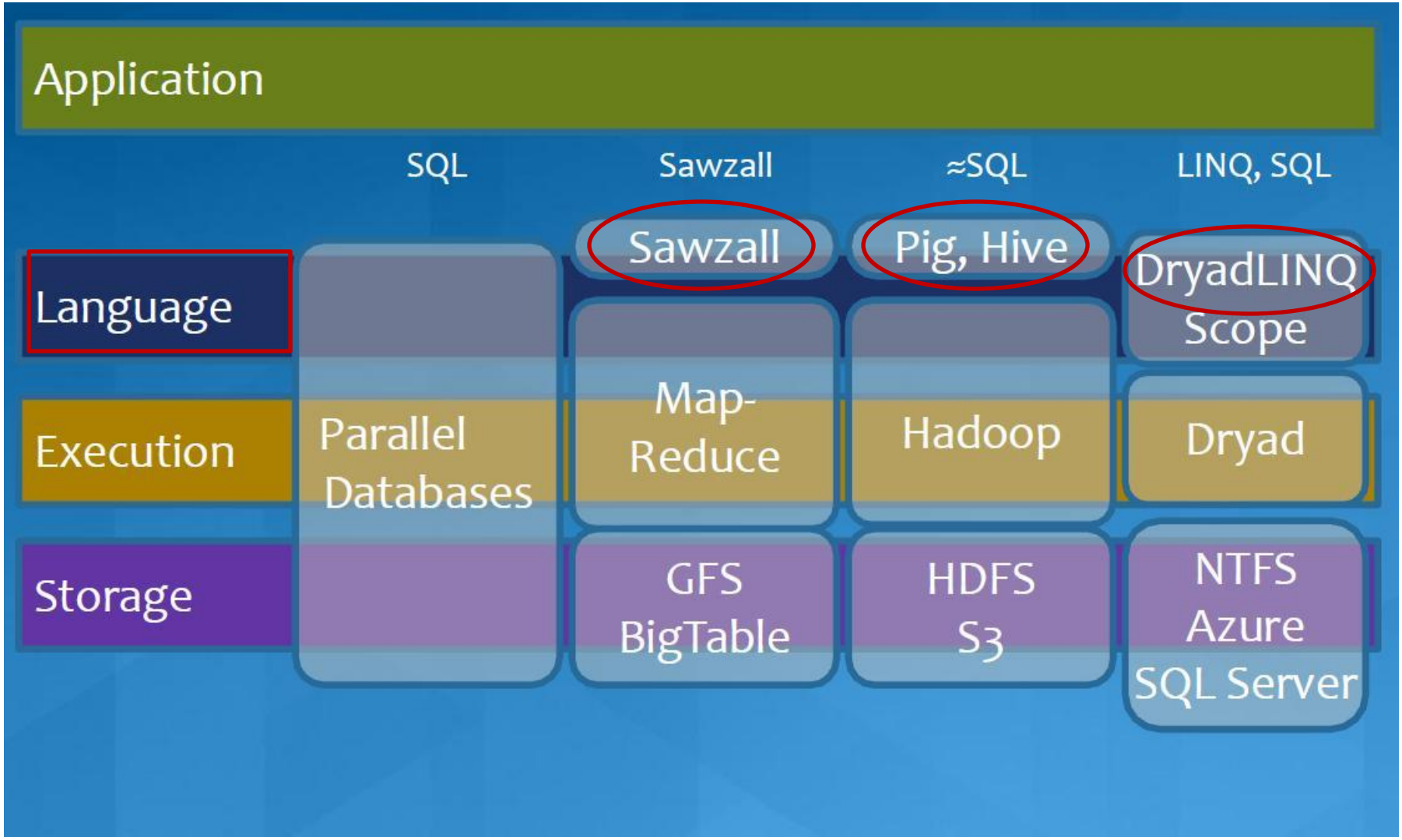


Google Twister Iterative MapReduce

Ideal for data intensive pleasingly parallel applications

SALSA

MapReduce in Heterogeneous Environment



Iterative MapReduce Frameworks

- Twister[1]
  - Map->Reduce->Combine->Broadcast
  - Long running map tasks (data in memory)
  - Centralized driver based, statically scheduled.
- Daytona[3]
  - Iterative MapReduce on Azure using cloud services
  - Architecture similar to Twister
- Haloop[4]
  - On disk caching, Map/reduce input caching, reduce output caching
- Spark[5]
  - Distributed querying with working sets

Others

- Mate-EC2[6]
  - Local reduction object
- Network Levitated Merge[7]
  - RDMA/infiniband based shuffle & merge
- Asynchronous Algorithms in MapReduce[8]
  - Local & global reduce
- MapReduce online[9]
  - online aggregation, and continuous queries
  - Push data from Map to Reduce
- Orchestra[10]
  - Data transfer improvements for MR
- iMapReduce[11]
  - Async iterations, One to one map & reduce mapping, automatically joins loop-variant and invariant data
- CloudMapReduce[12] & Google AppEngine MapReduce[13]
  - MapReduce frameworks utilizing cloud infrastructure services

Twister4Azure

Azure Cloud Services

- Highly-available and scalable
- Utilize eventually-consistent , high-latency cloud services effectively

Decentralized

- Avoids Single Point of Failure
- Global queue based dynamic scheduling
- Dynamically scale up/down

MapReduce

- Iterative MapReduce for Azure
- Fault tolerance

Applications of Twister4Azure

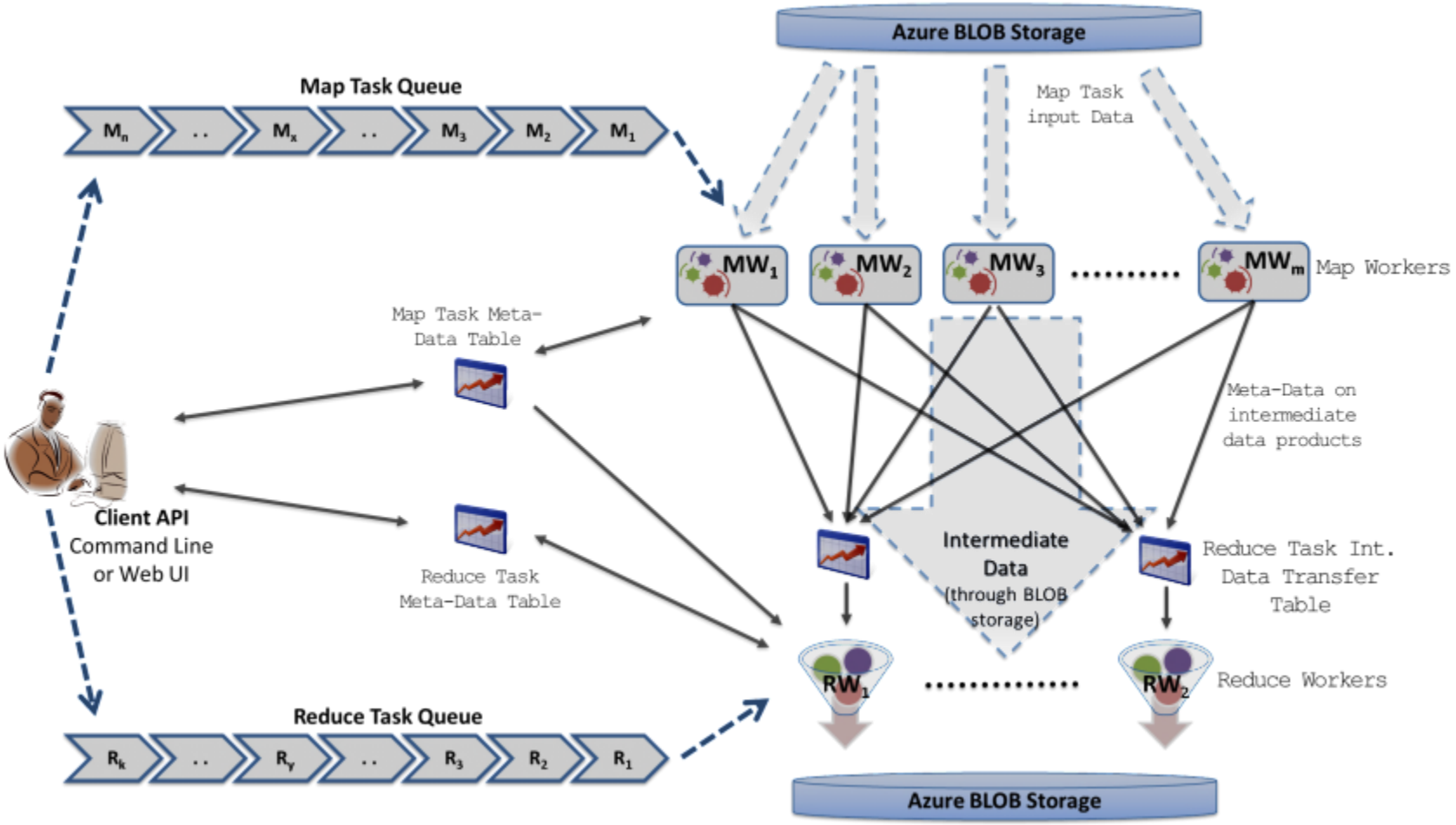
- Implemented
  - **Multi Dimensional Scaling**
  - **KMeans Clustering**
  - PageRank
  - SmithWatermann-GOTOH sequence alignment
  - WordCount
  - Cap3 sequence assembly
  - Blast sequence search
  - GTM & MDS interpolation
- Under Development
  - **Latent Dirichlet Allocation**
  - **Descendent Query**

Twister4Azure – Iterative MapReduce

- **Extends MapReduce programming model**
- **Decentralized iterative MR architecture for clouds**
  - **Utilize highly available and scalable Cloud services**
- **Multi-level data caching**
  - **Cache aware hybrid scheduling**
- **Multiple MR applications per job**
- **Collective communication primitives**
  - **Outperforms Hadoop in local cluster by 2 to 4 times**
- **Sustain features**
  - **dynamic scheduling, load balancing, fault tolerance, monitoring, local testing/debugging**

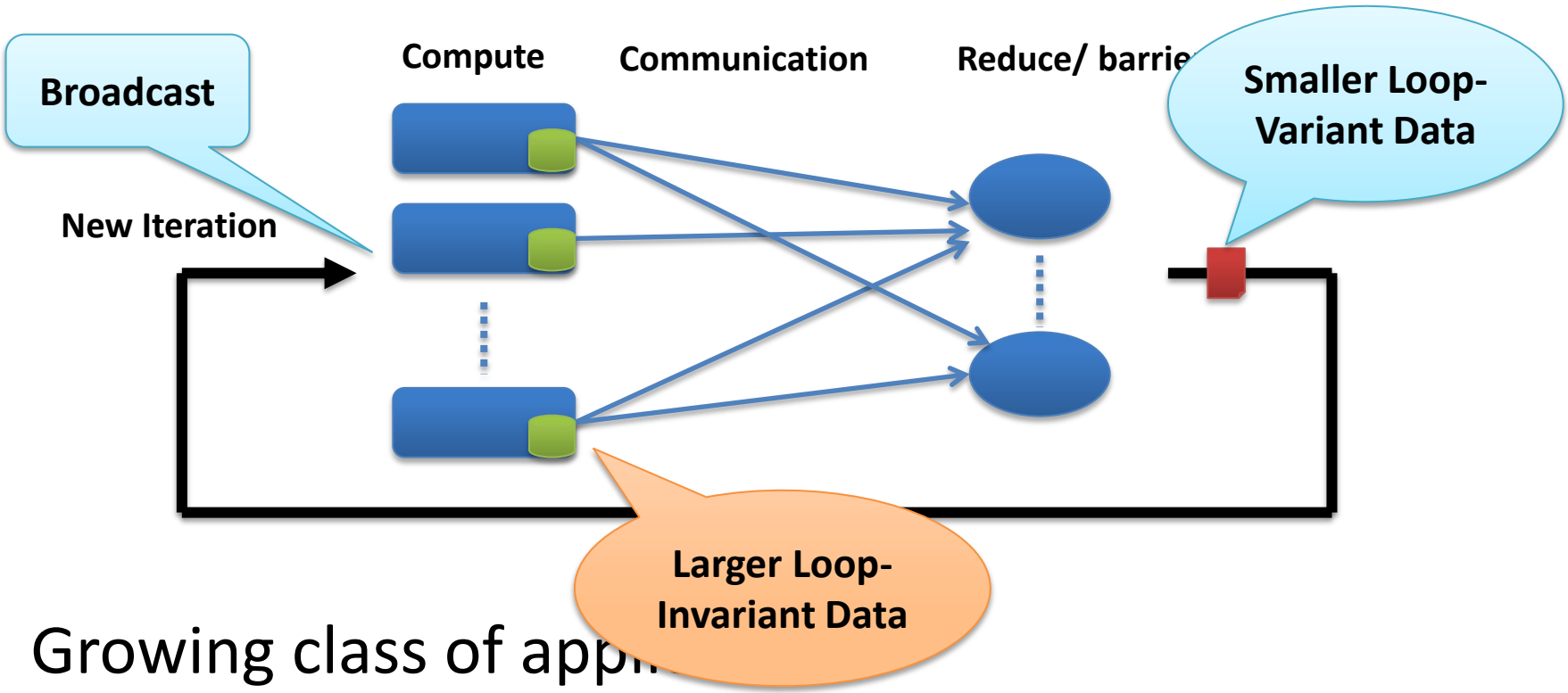
http://salsahpc.indiana.edu/twister4azure/

Twister4Azure Architecture



Azure Queues for scheduling, Tables to store meta-data and monitoring data, Blobs for input/output/intermediate data storage.

Data Intensive Iterative Applications

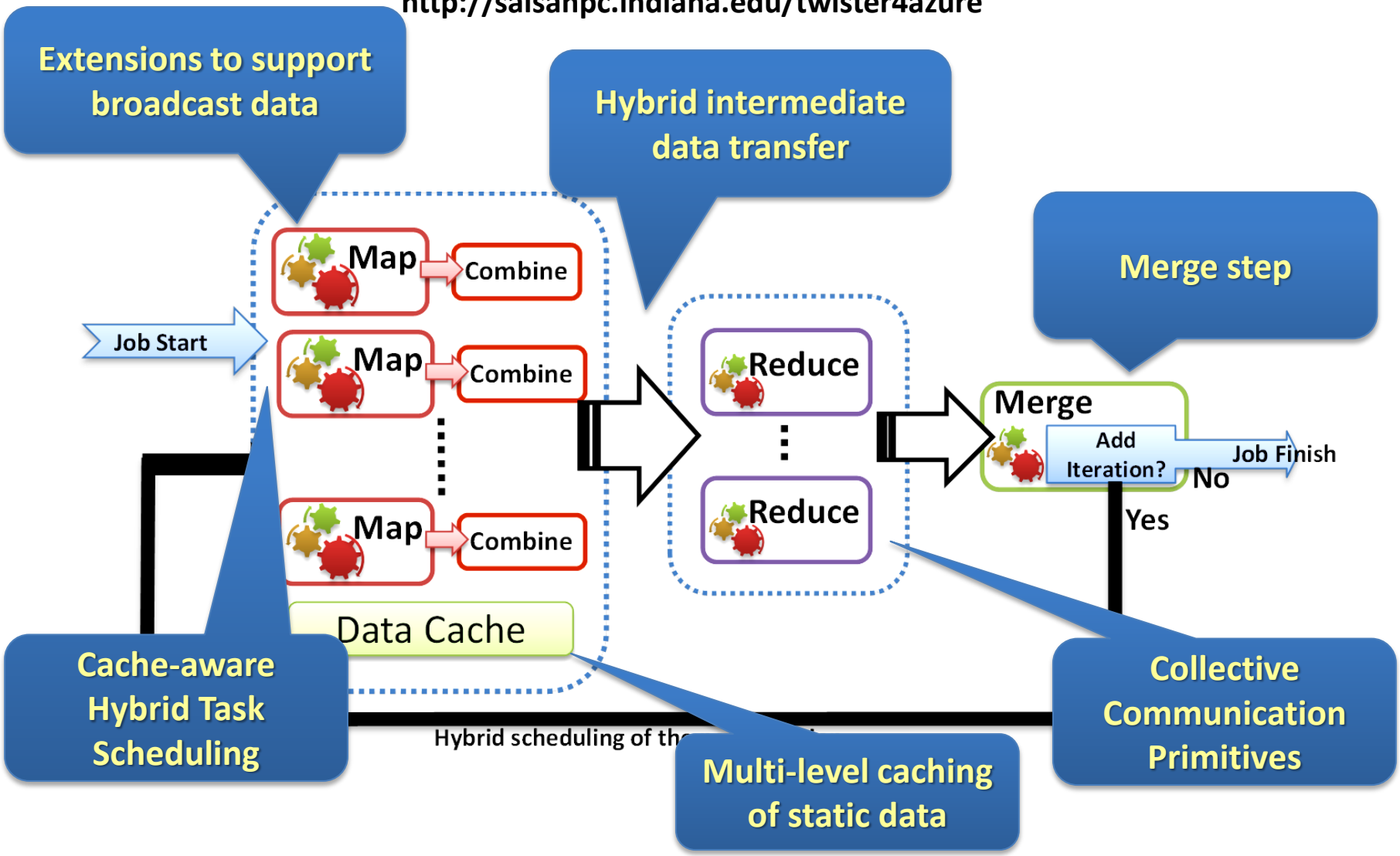


- Growing class of appl
  - Clustering, data mining, machine learning & dimension reduction applications
  - Driven by data deluge & emerging computation fields

Twister4Azure

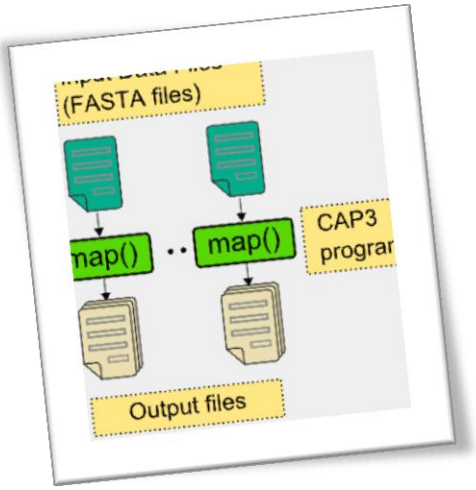
Iterative MapReduce for Azure Cloud

http://salsahpc.indiana.edu/twister4azure

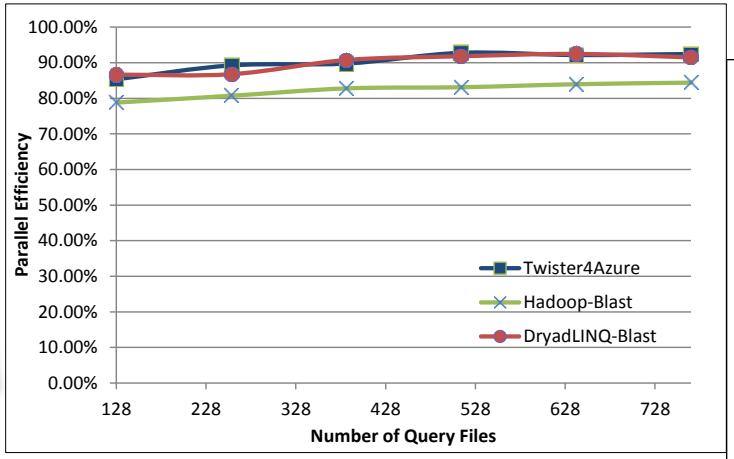


Portable Parallel Programming on Cloud and HPC: Scientific Applications of Twister4Azure, Thilina Gunarathne, Bingjing Zang, Tak-Lon Wu and Judy Qiu, (UCC 2011) , Melbourne, Australia.

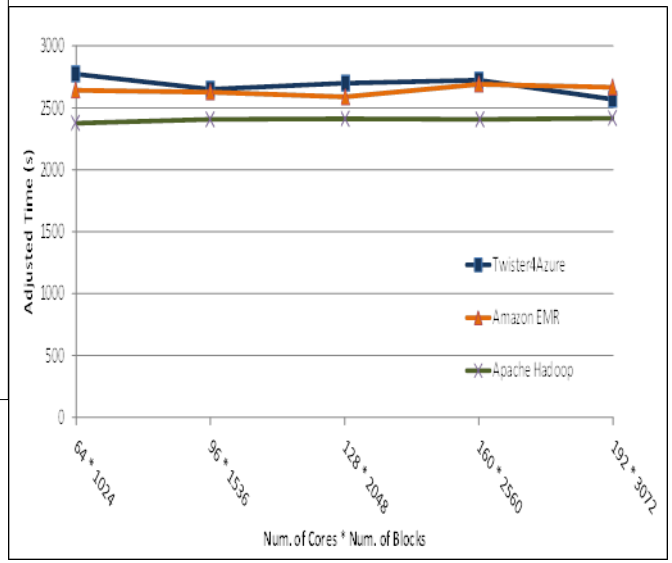
Performance of Pleasingly Parallel Applications on Azure



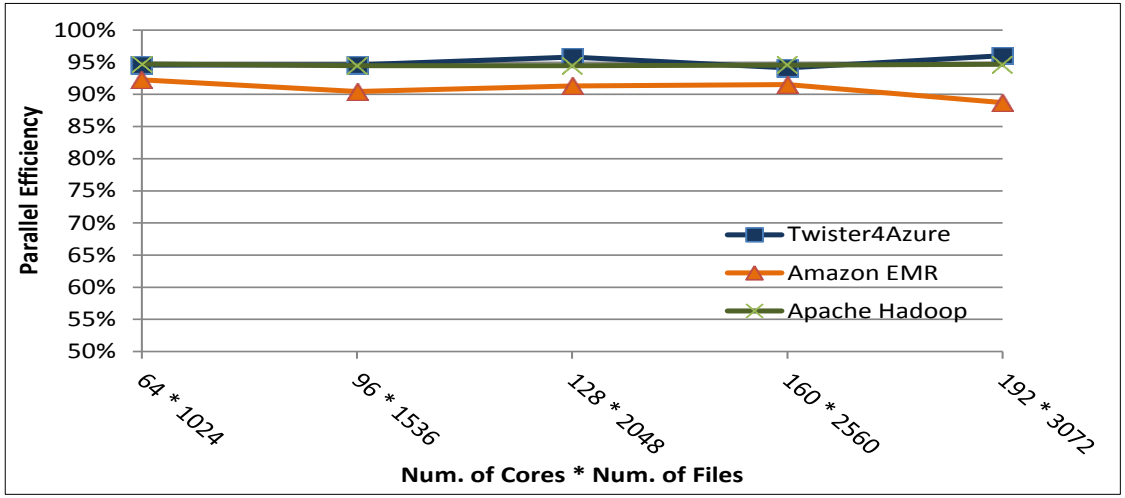
BLAST Sequence Search

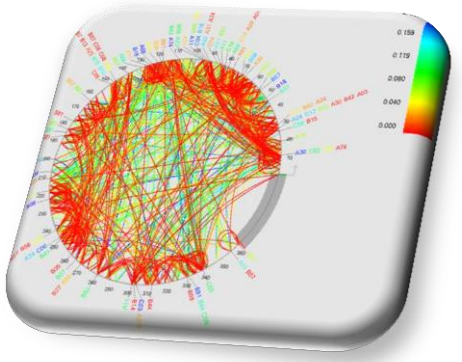


Smith Watermann Sequence Alignment



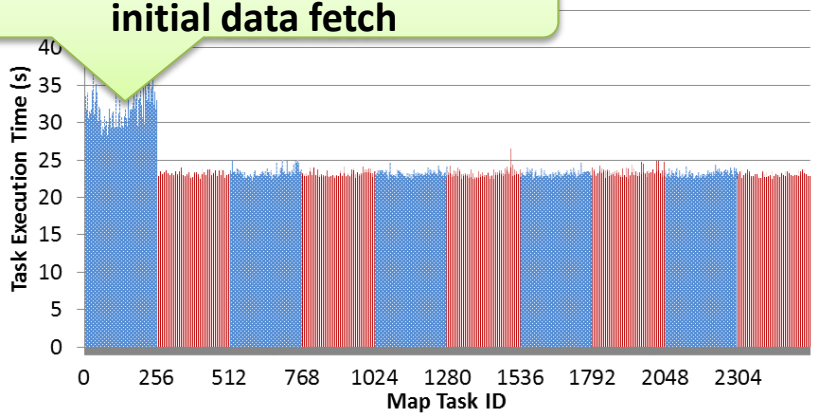
Cap3 Sequence Assembly





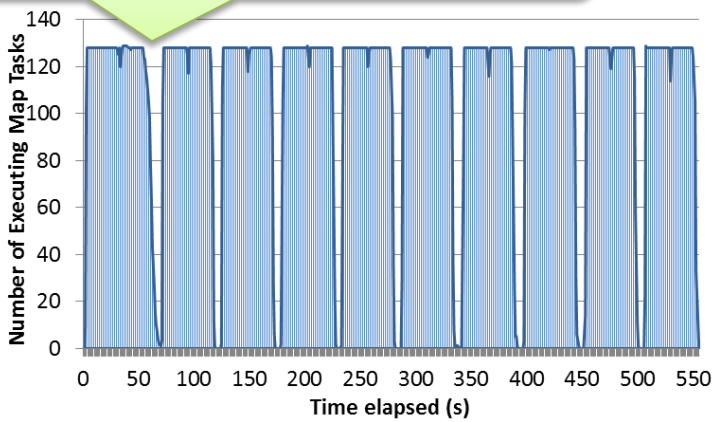
Performance – Kmeans Clustering

First iteration performs the initial data fetch

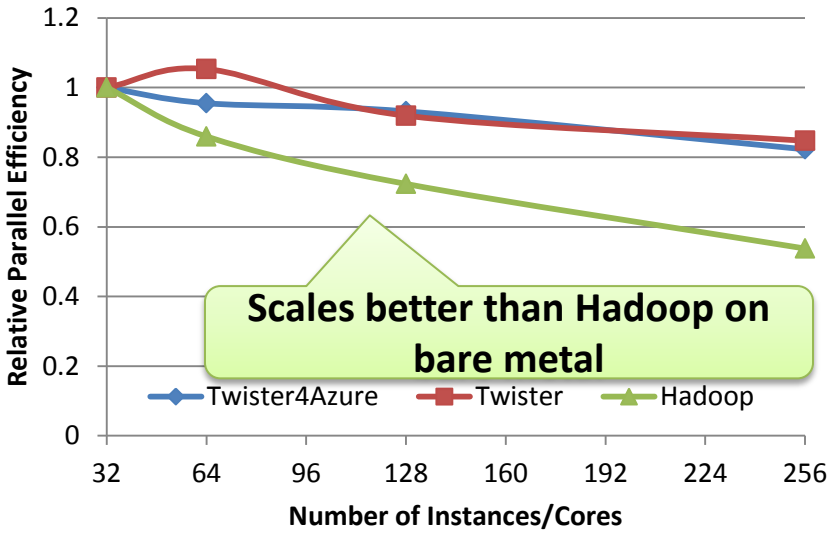


Task Execution Time Histogram

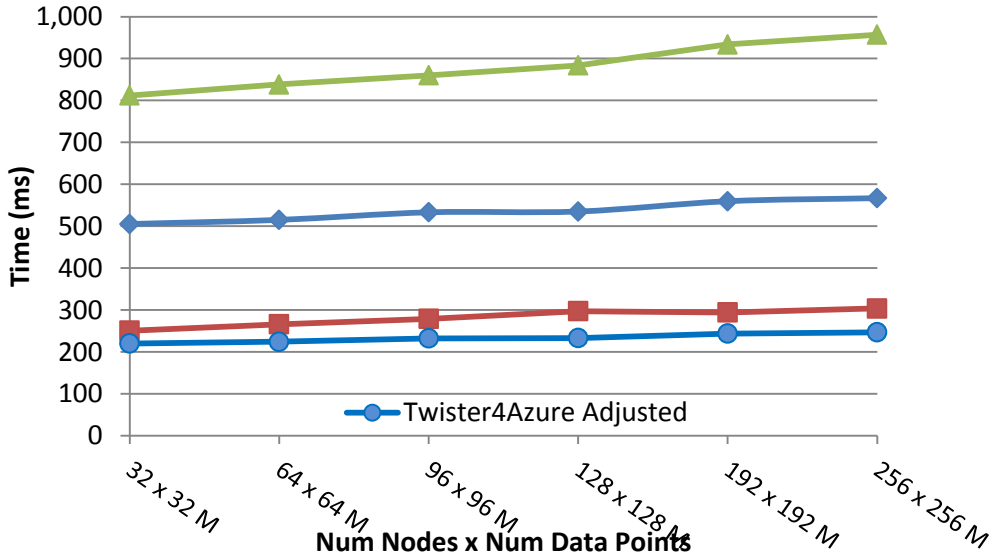
Overhead between iterations



Number of Executing Map Task Histogram



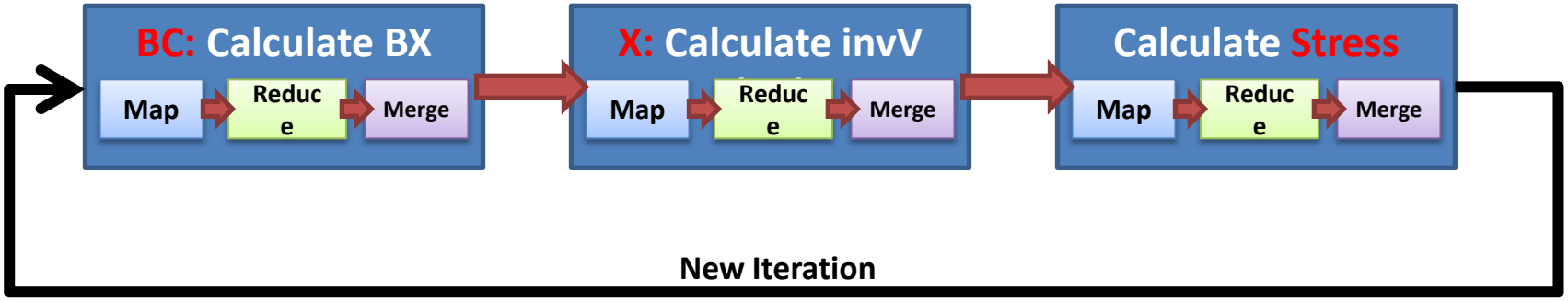
Strong Scaling with 128M Data Points

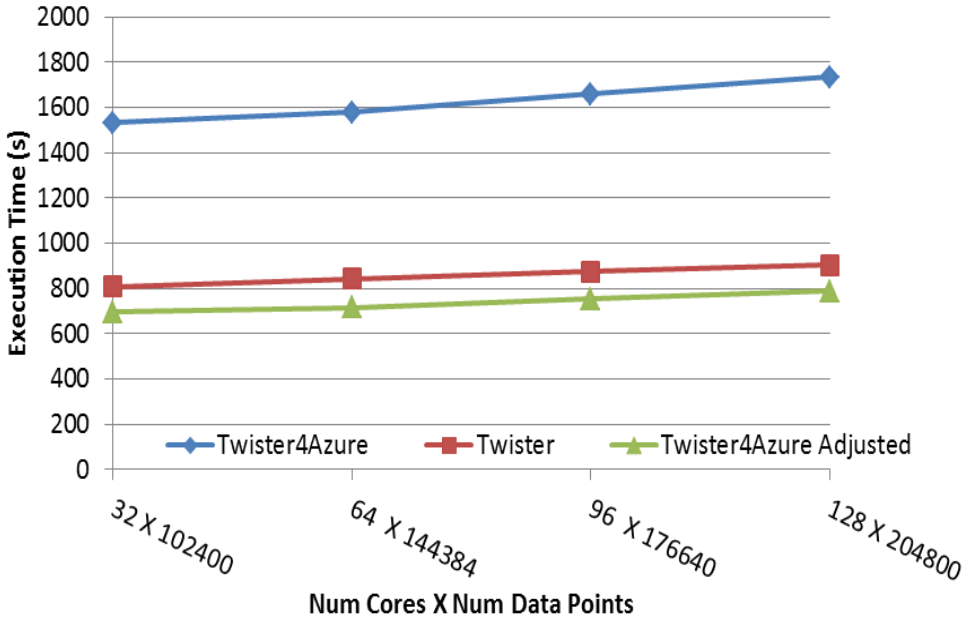


Weak Scaling

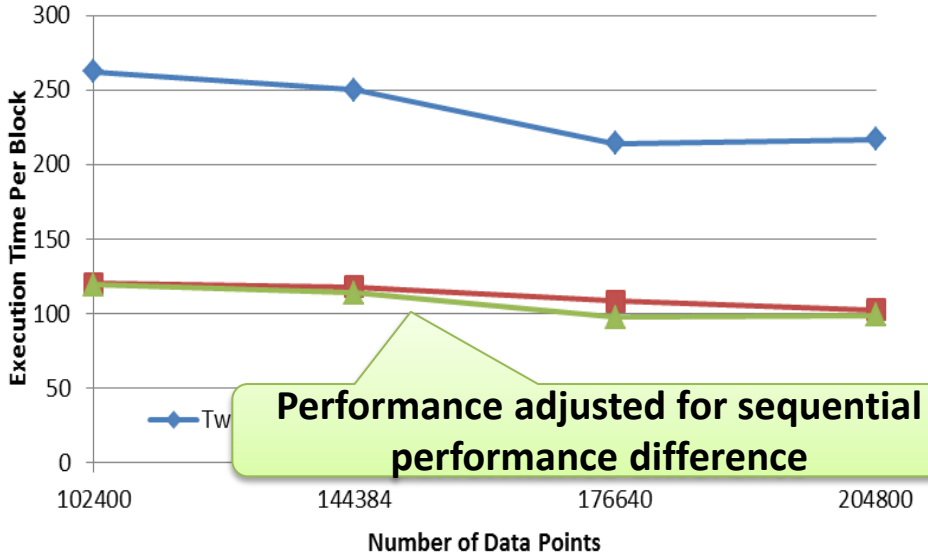
SALSA HPC

Performance – Multi Dimensional Scaling





Weak Scaling



Data Size Scaling

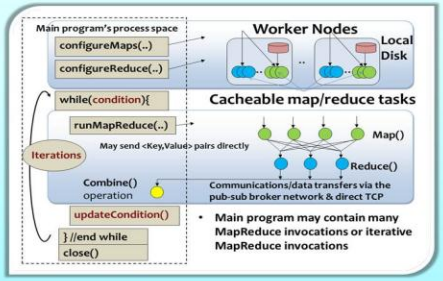
Scalable Parallel Scientific Computing Using Twister4Azure. Thilina Gunarathne, BingJing Zang, Tak-Lon Wu and Judy Qiu. Submitted to Journal of Future Generation Computer Systems. (Invited as one of the best 6 papers of UCC 2011)

Iterative MapReduce Enabling HPC-Cloud Interoperability

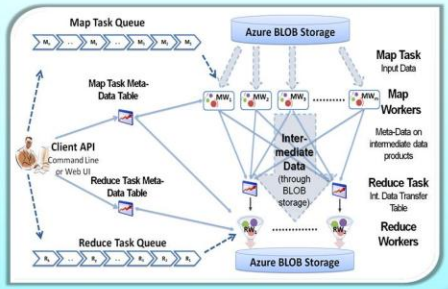
Scientific Applications

Iterative MapReduce Runtime

Twister
Cross Platform Iterative MapReduce



Twister4Azure
Iterative MapReduce for Azure Cloud



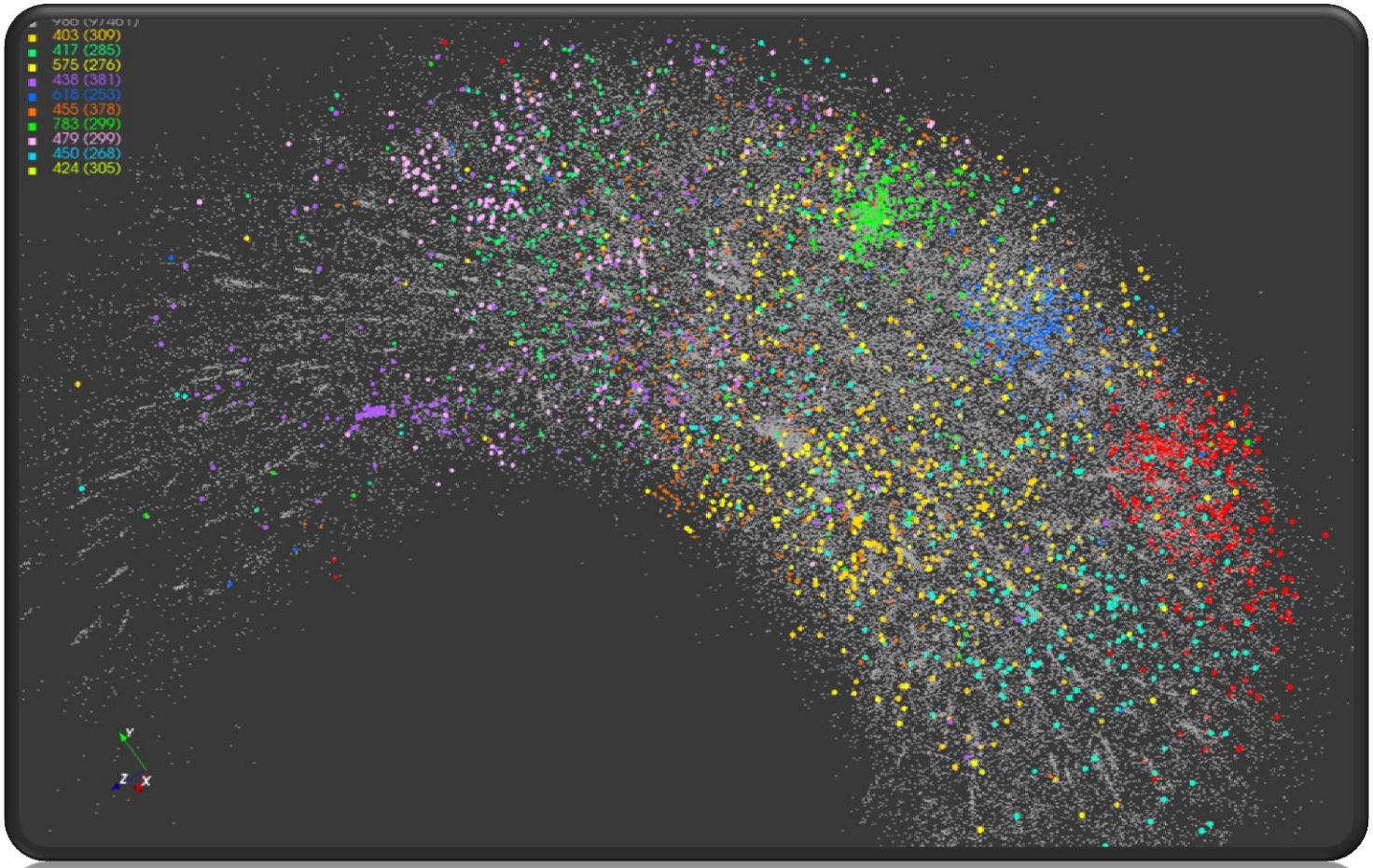
Large Scale Infrastructure

HPC Cluster

Windows Azure

amazon web services

Twister-MDS Output



MDS projection of 100,000 protein sequences showing a few experimentally identified clusters in preliminary work with Seattle Children's Research Institute

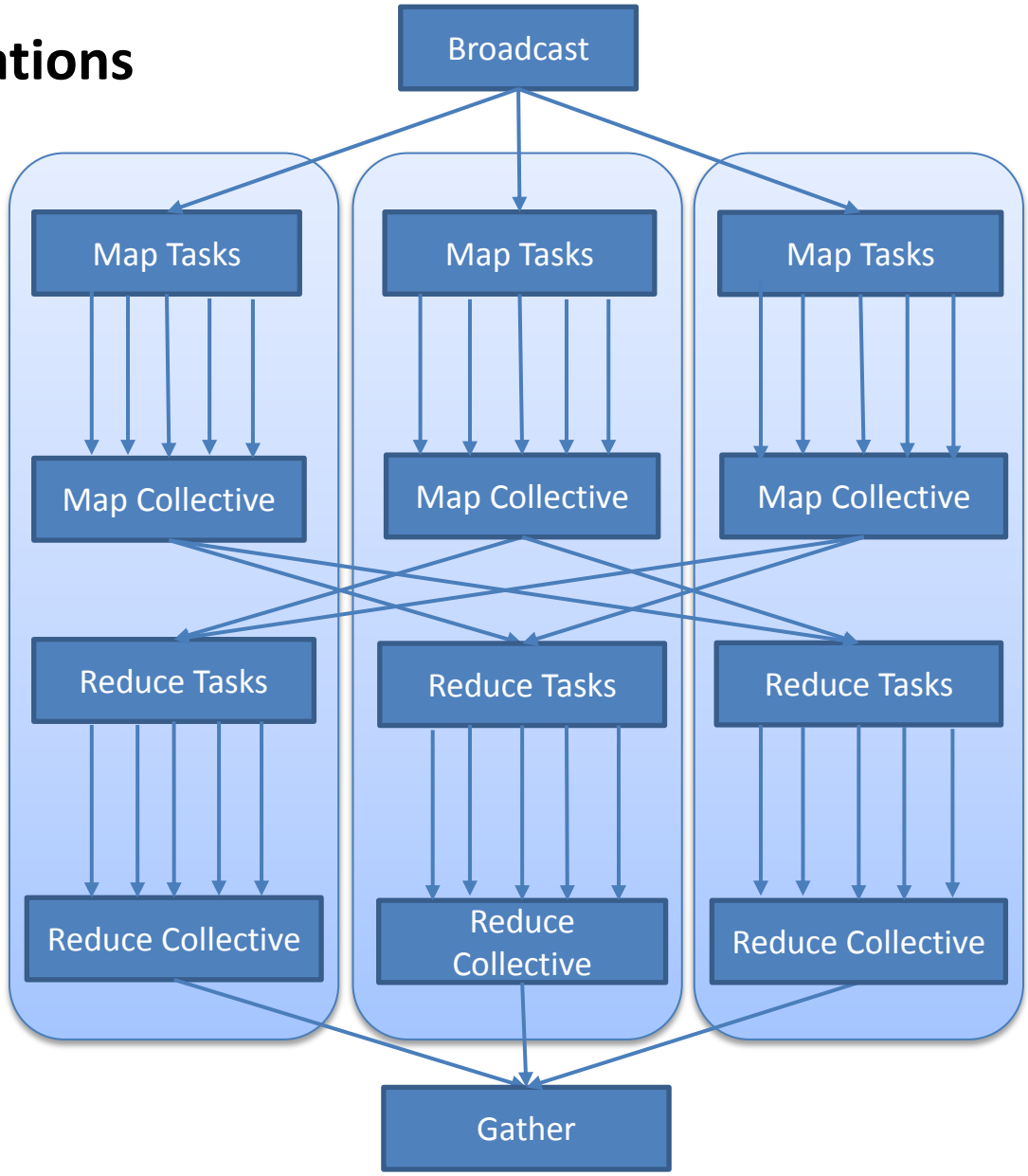
Twister v0.9

New Infrastructure for Iterative MapReduce Programming

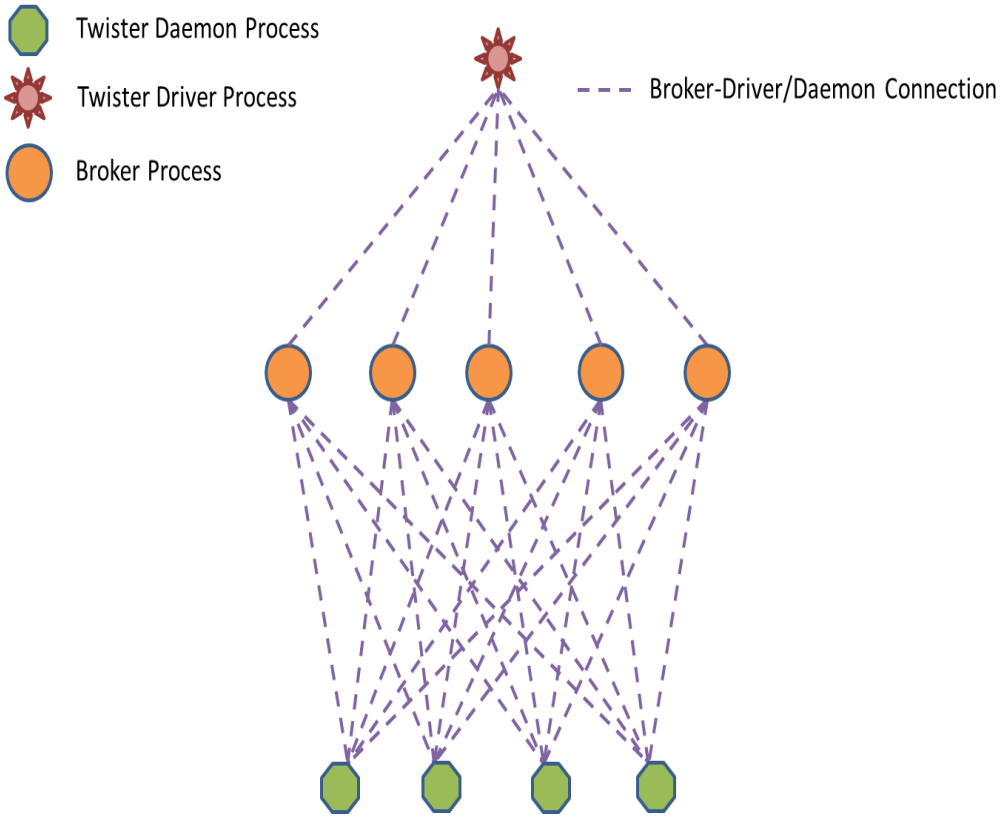
- *Configuration Program to setup Twister environment automatically on a cluster*
- *Full mesh network of brokers for facilitating communication*
- *New messaging interface for reducing the message serialization overhead*
- *Memory Cache to share data between tasks and jobs*

Twister4Azure Communications

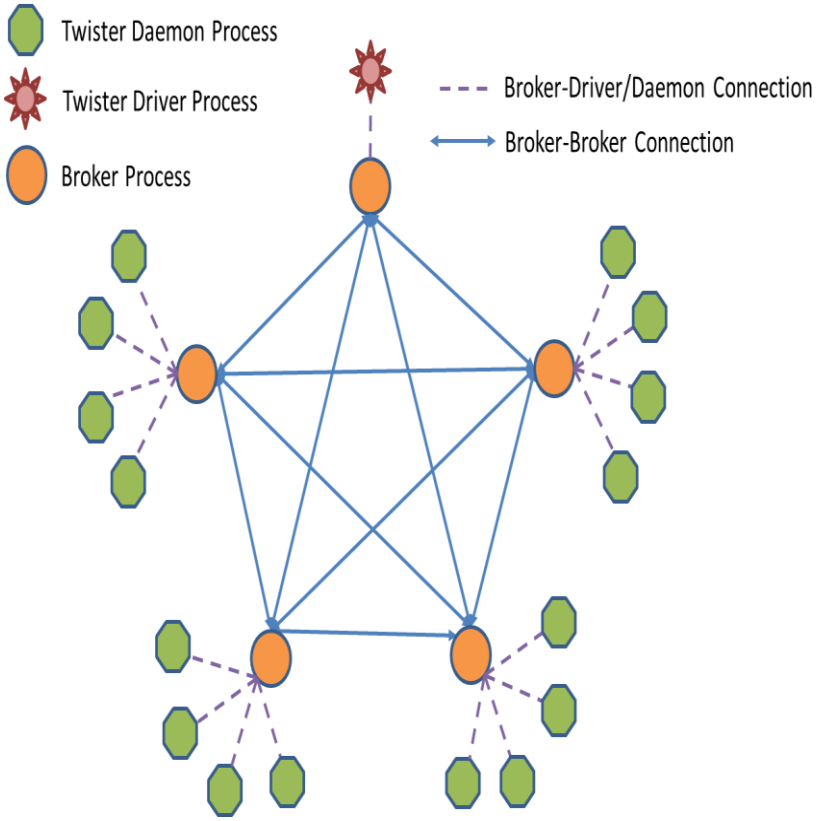
- Broadcasting
  - Data could be large
  - Chain & MST
- Map Collectives
  - Local merge
- Reduce Collectives
  - Collect but no merge
- Combine
  - Direct download or Gather



Improving Performance of Map Collectives



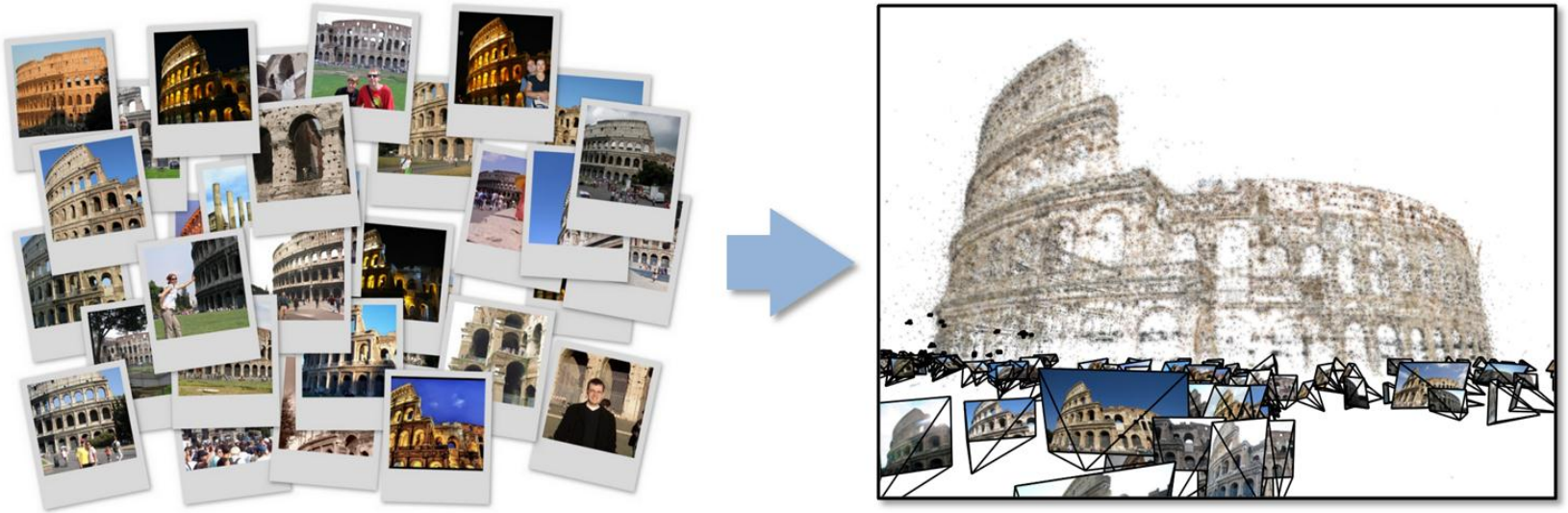
Full Mesh Broker Network



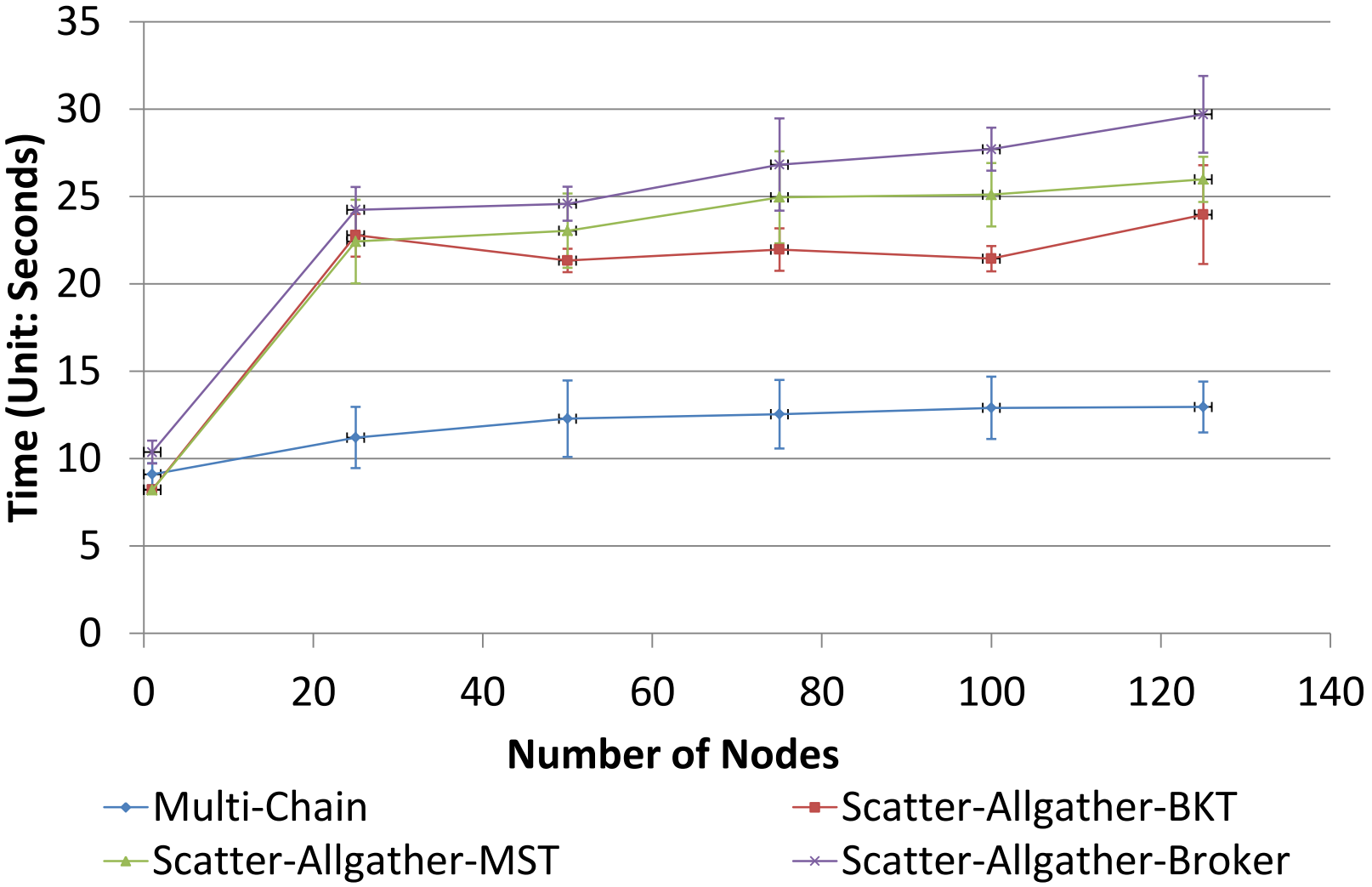
Scatter and Allgather

Data Intensive Kmeans Clustering

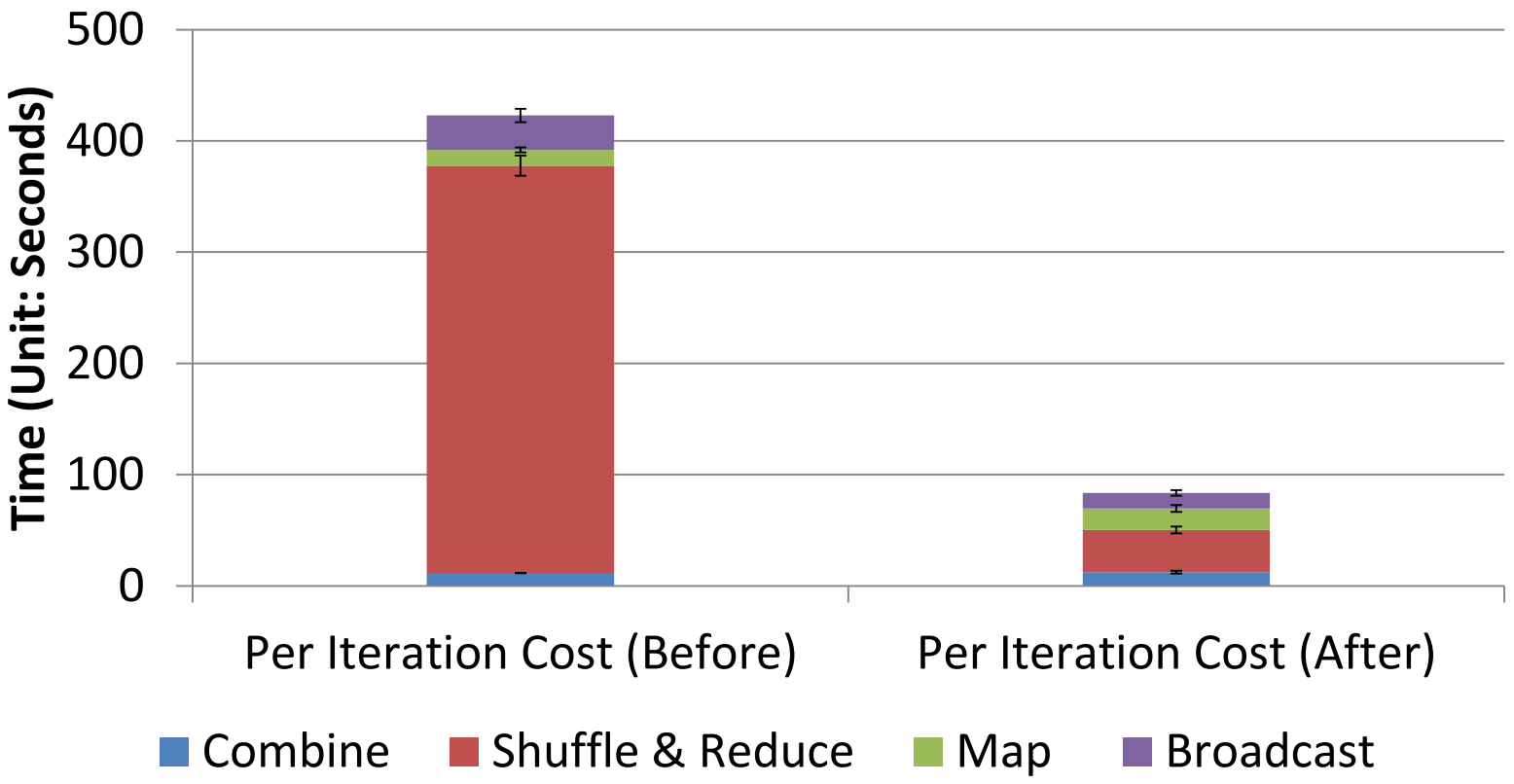
— *Image Classification: 1.5 TB;* 1.5 TB; 500 features per image;10k clusters 1000 Map tasks; 1GB data transfer per Map task



Polymorphic Scatter-Allgather in Twister



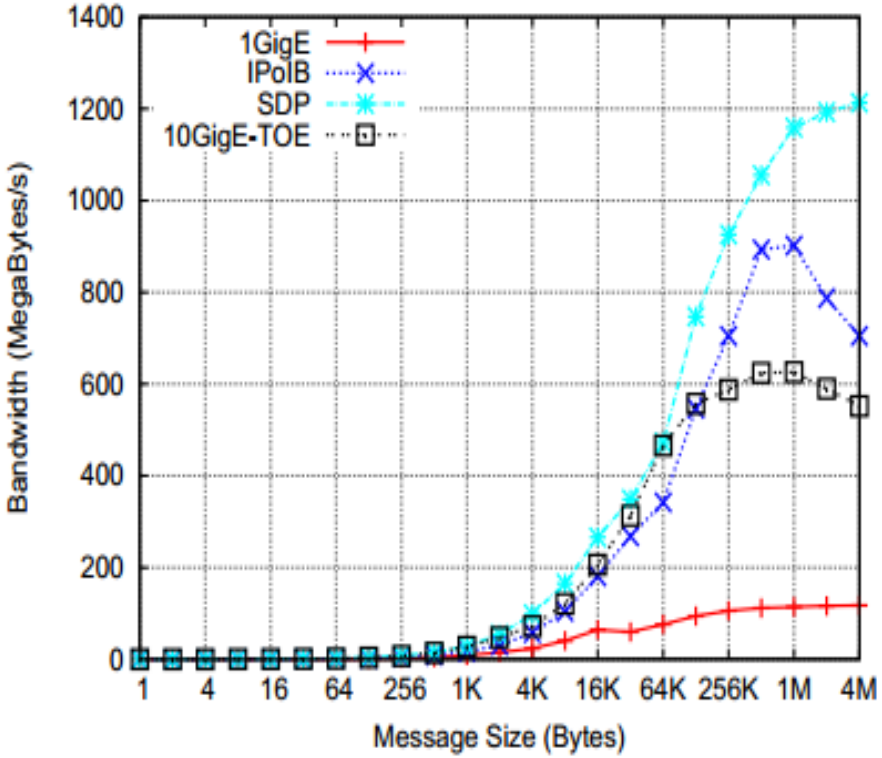
Twister Performance on Kmeans Clustering



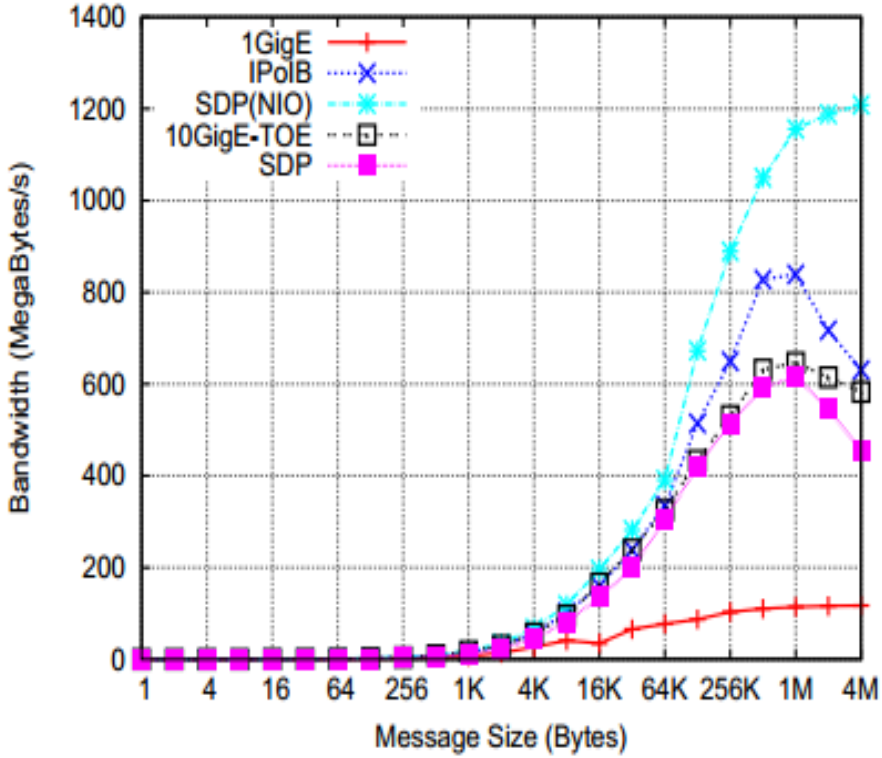
Twister on InfiniBand

- InfiniBand successes in HPC community
  - More than 42% of Top500 clusters use InfiniBand
  - Extremely high throughput and low latency
    - Up to 40Gb/s between servers and 1µsec latency
  - Reduce CPU overhead up to 90%
- Cloud community can benefit from InfiniBand
  - Accelerated Hadoop (sc11)
  - HDFS benchmark tests
- RDMA can make Twister faster
  - Accelerate static data distribution
  - Accelerate data shuffling between mappers and reducer
- In collaboration with ORNL on a large InfiniBand cluster

Bandwidth comparison of HDFS on various network technologies

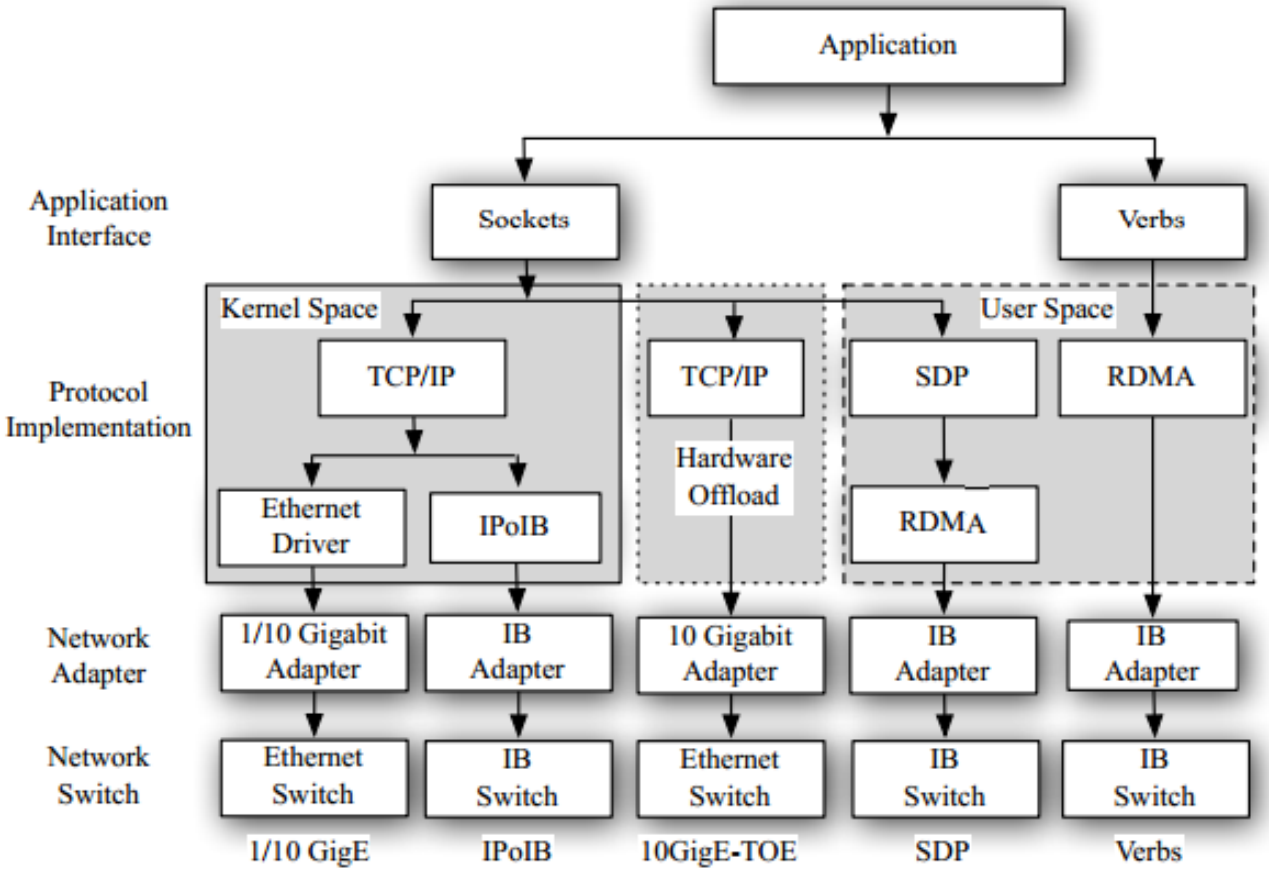


(a) Bandwidth with C



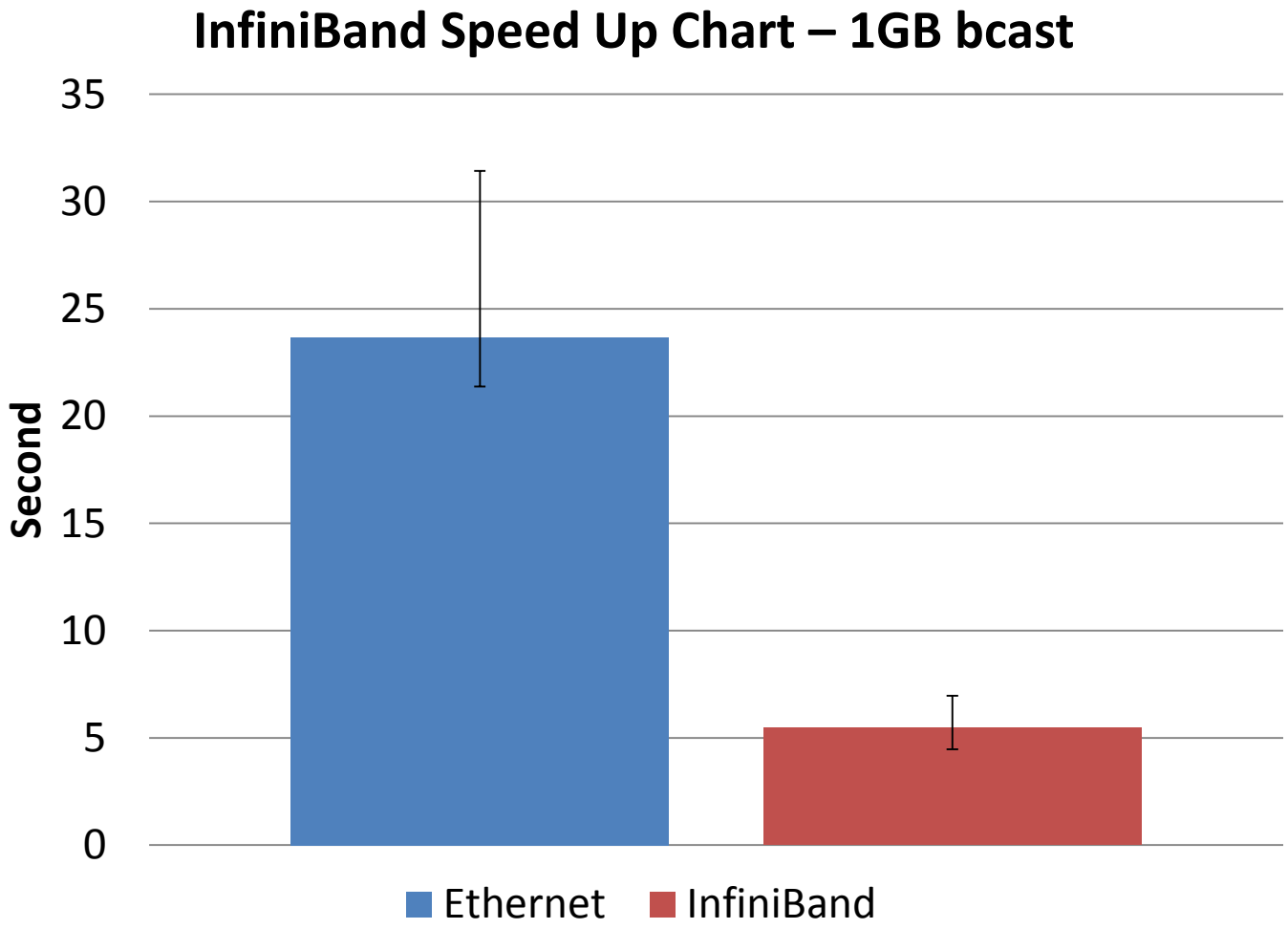
(b) Bandwidth with Java

Using RDMA for Twister on InfiniBand



(a) Networking Layers, OS-Bypass and Hardware Offload

Twister Broadcast Comparison: Ethernet vs. InfiniBand



Building Virtual Clusters

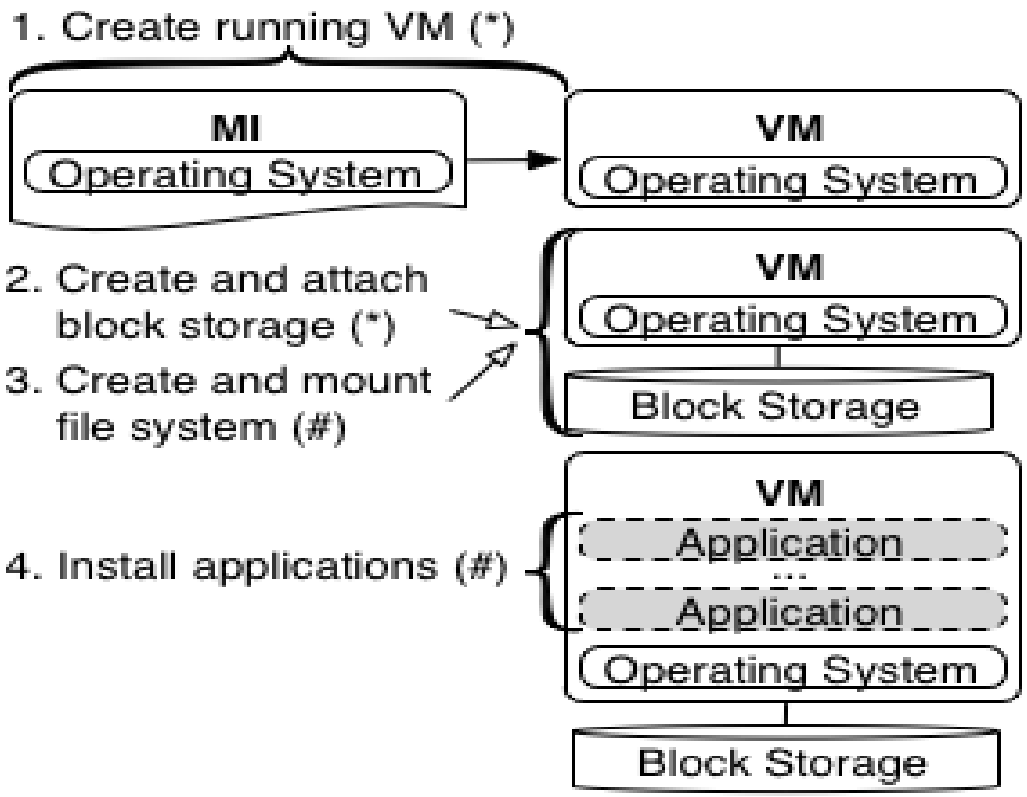
Towards Reproducible eScience in the Cloud

Separation of concerns between two layers

- **Infrastructure Layer** – interactions with the Cloud API
- **Software Layer** – interactions with the running VM

Separation Leads to Reuse

Infrastructure Layer = (*) **Software Layer = (#)**

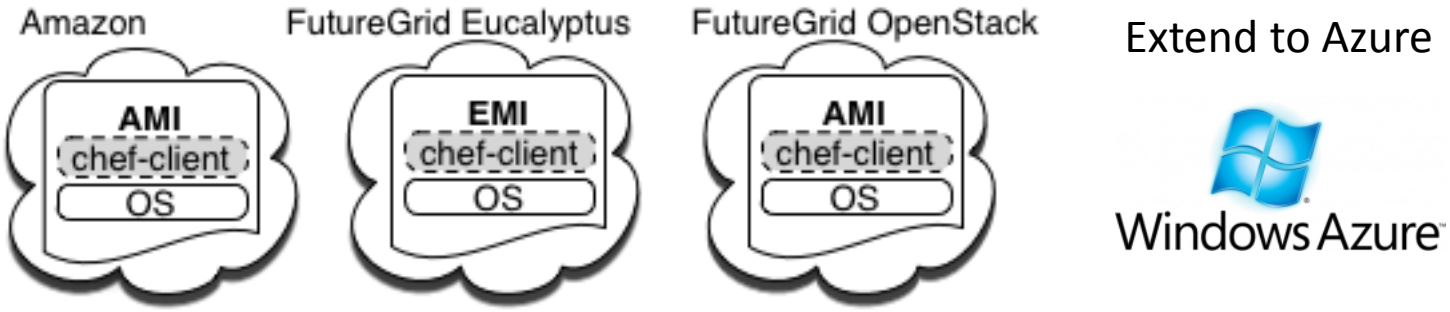


By separating layers, one can reuse software layer artifacts in separate clouds

Design and Implementation

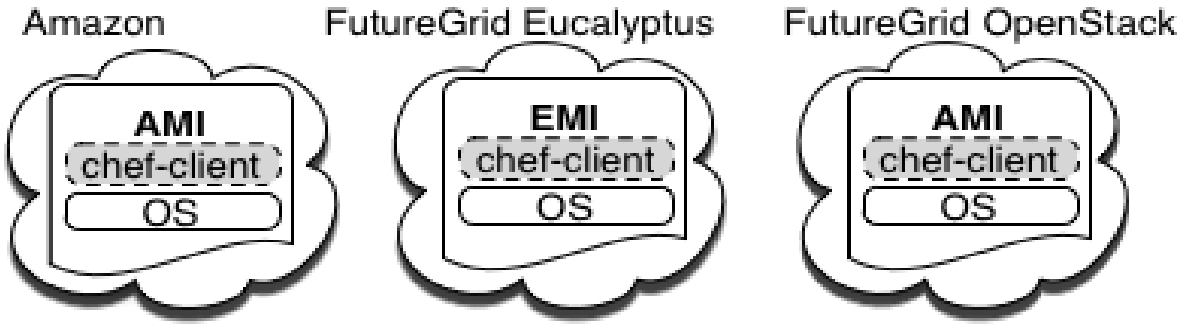
Equivalent machine images (MI) built in separate clouds

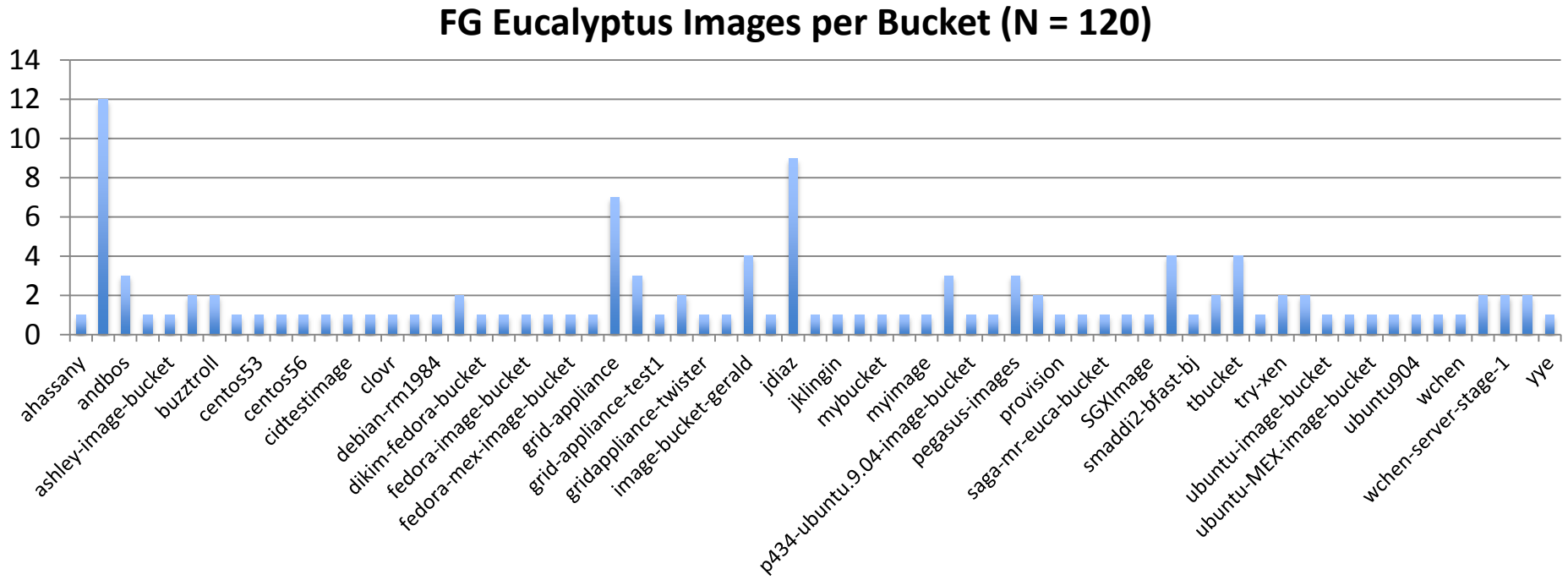
- Common underpinning in separate clouds for software installations and configurations



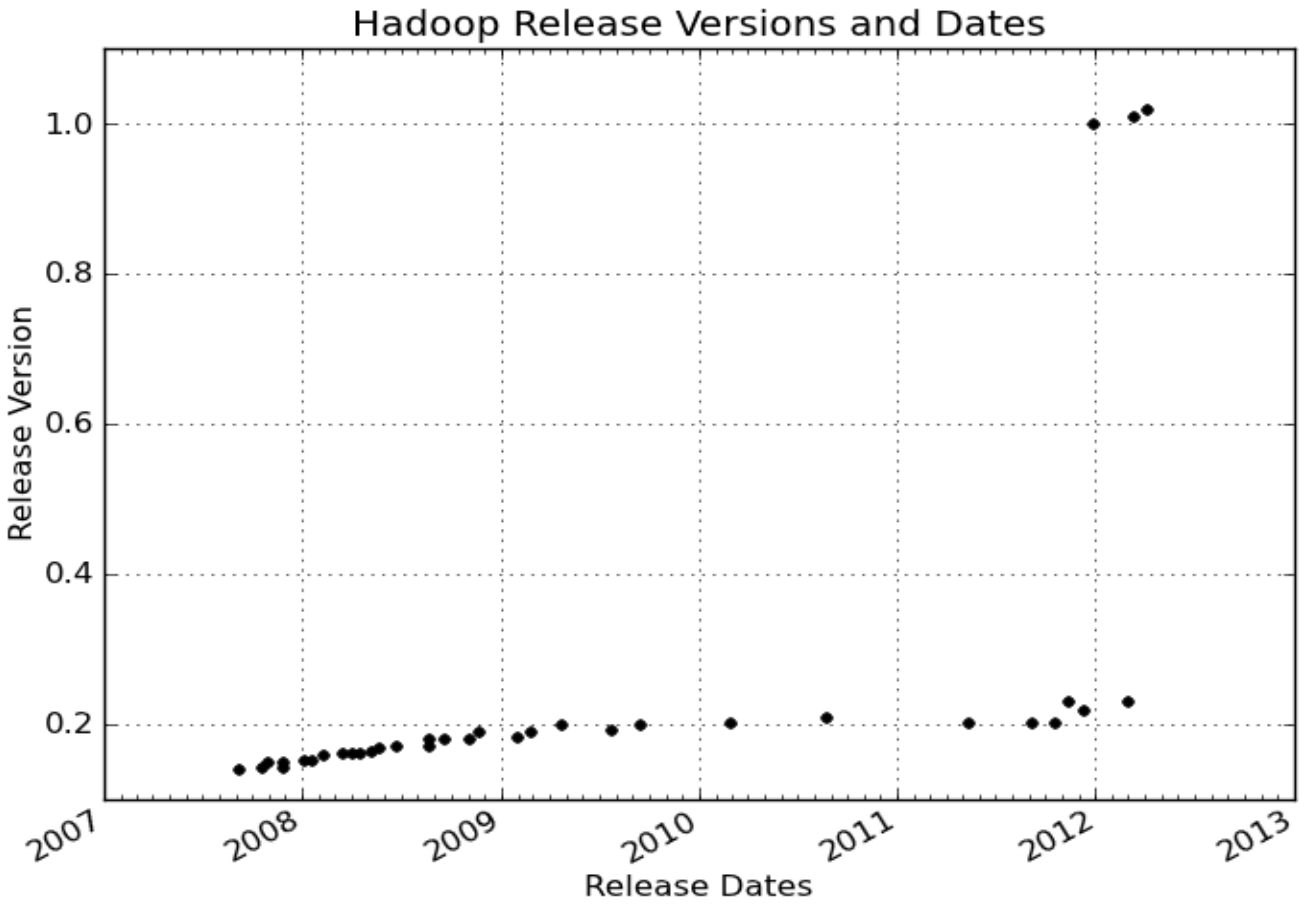
- Configuration management used for software automation

Cloud Image Proliferation





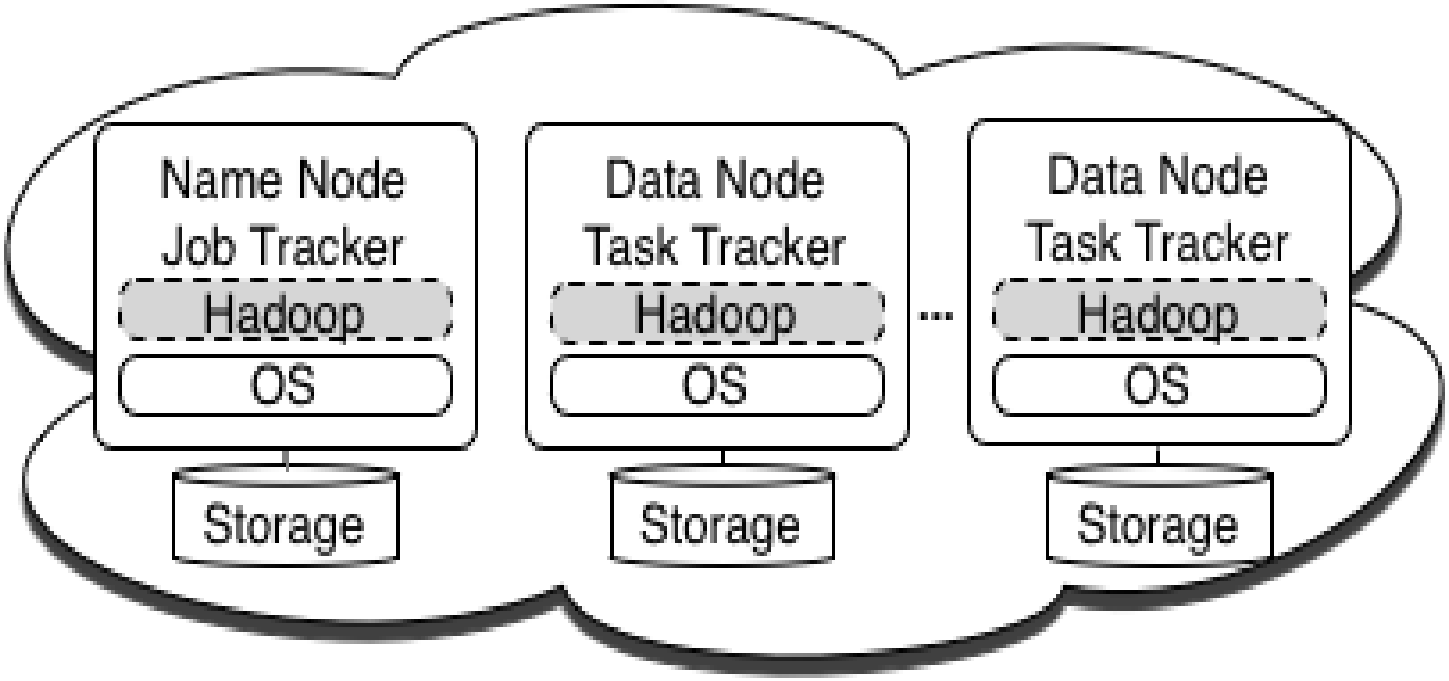
Changes of Hadoop Versions



Implementation - Hadoop Cluster

Hadoop cluster commands

- knife hadoop launch {name} {slave count}
- knife hadoop terminate {name}

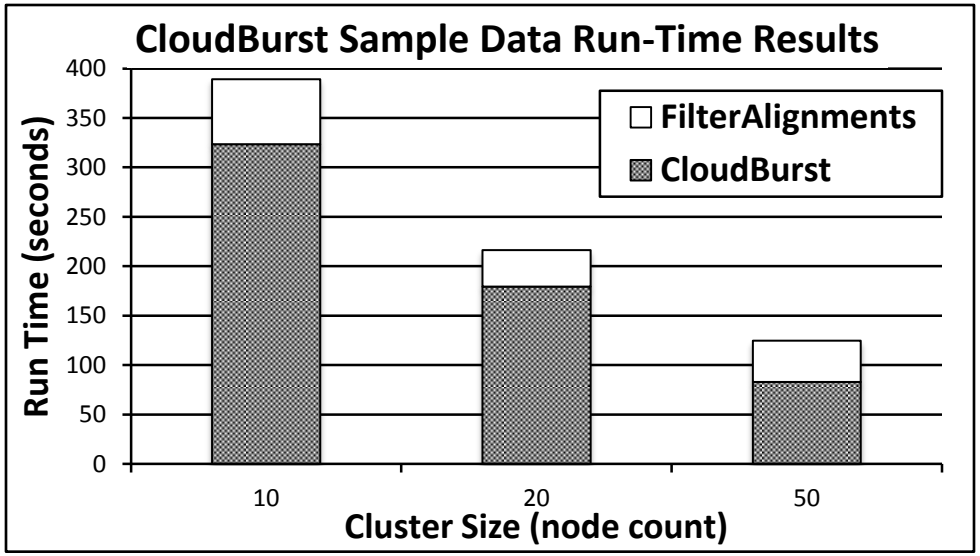


Running CloudBurst on Hadoop

Running CloudBurst on a 10 node Hadoop Cluster

- knife hadoop launch cloudburst 9
- echo '{"run list": "recipe[cloudburst]"}' > cloudburst.json
- chef-client -j cloudburst.json

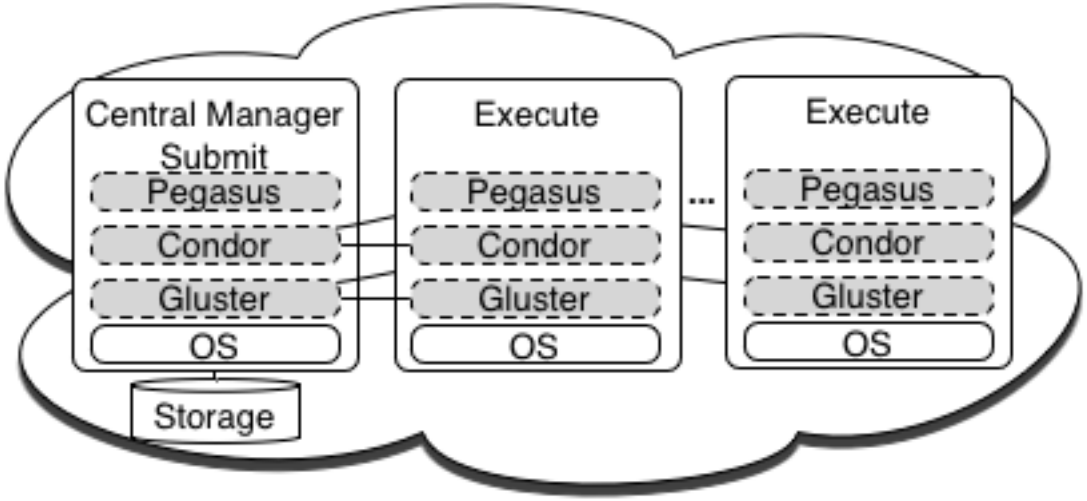
CloudBurst on a 10, 20, and 50 node Hadoop Cluster



Implementation - Condor Pool

Condor Pool commands

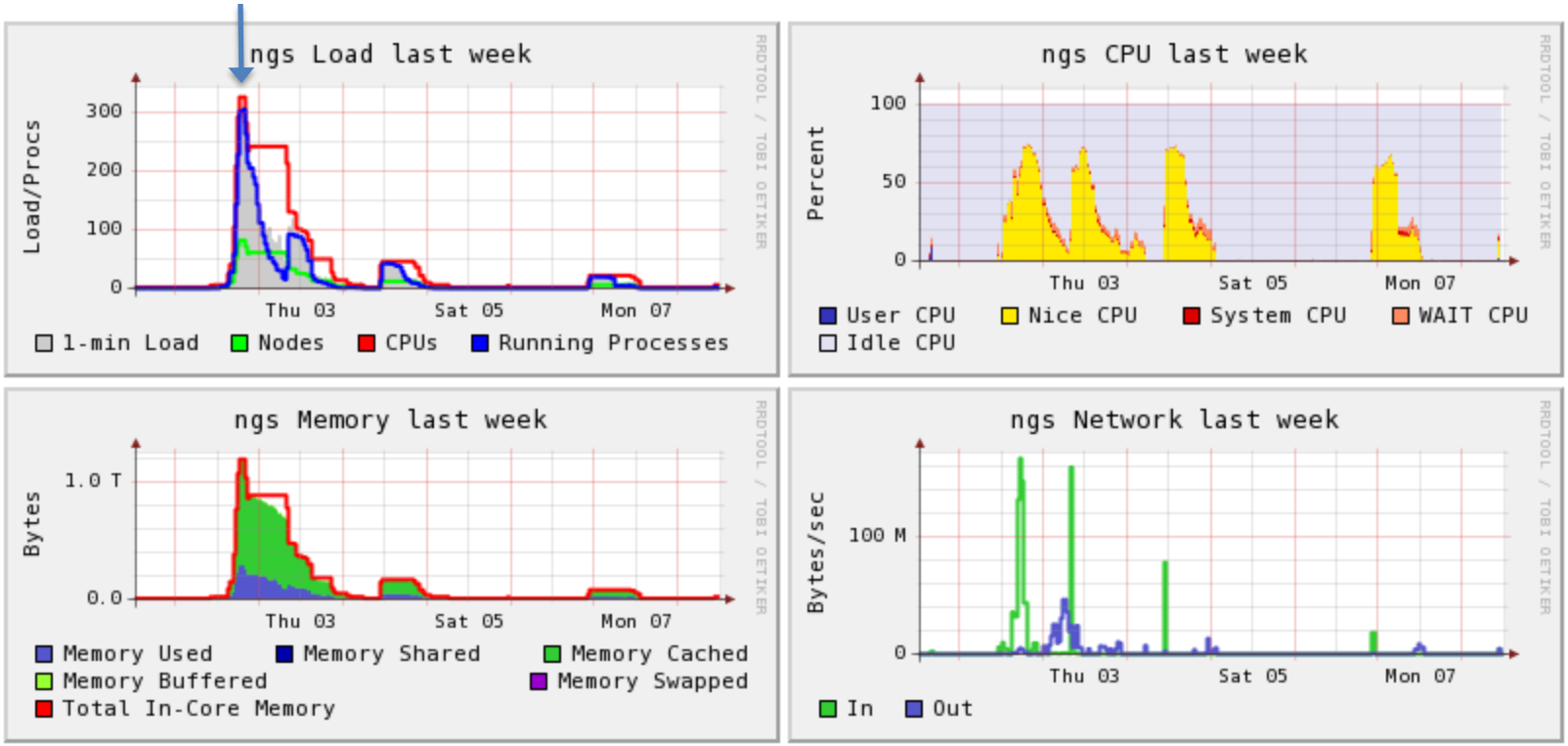
- knife cluster launch {name} {exec. host count}
- knife cluster terminate {name}
- knife cluster node add {name} {node count}



Implementation - Condor Pool

Ganglia screen shot of a Condor pool in Amazon EC2

80 node – (320 core) at this point in time



Ackowledgements

SALSA HPC Group

http://salsahpc.indiana.edu

School of Informatics and Computing

Indiana University

Data Enabled Science
















SALSA HPC

Tutorials

View | Edit | Outline | Revisions

Note: This page is maintained by Renato Figueiredo from UF.

On this page you will find a number of links to tutorials. Tutorials are broadly organized into topics, and each tutorial is classified based on the user's target level of expertise with FutureGrid (novice, intermediate, advanced). (If you are a tutorial developer, for instructions on how to add a tutorial to this list, please refer to the TEOS page).

If you have any corrections or suggestions to our tutorial content, please fill out a help request at https://portal.futuregrid.org/help .

Tutorial Topic 1: Cloud Provisioning Platforms

- Using Nimbus on FutureGrid [novice]
- Nimbus One-click Cluster Guide [intermediate]
- Using OpenStack Nova on FutureGrid [novice]
- Using Eucalyptus on FutureGrid [novice]
- Connecting private network VMs across Nimbus clusters using ViNe [novice]
- Using the Grid Appliance to run FutureGrid Cloud Clients [novice]

Tutorial Topic 2: Cloud Run-time Map/Reduce Platforms

- Running Hadoop as a batch job using MyHadoop [novice]
- Running SalsaHadoop (one-click Hadoop) on HPC environment [beginner]
- Running Twister on HPC environment [beginner]
- Running SalsaHadoop on Eucalyptus [intermediate]
- Running FG-Twister on Eucalyptus [intermediate]
- Running One-click Hadoop WordCount on Eucalyptus [beginner]
- Running One-click Twister K-means on Eucalyptus [beginner]

Tutorial Topic 3: Grid Appliances for Training, Education and Outreach

- Running a Grid Appliance on your desktop [novice]
- Running a Grid Appliance on FutureGrid [novice]
- Running an OpenStack virtual appliance on FutureGrid [novice]
- Running Condor tasks on the Grid Appliance [novice]
- Running MPI tasks on the Grid Appliance [novice]
- Running Hadoop tasks on the Grid Appliance [novice]
- Deploying virtual private Grid Appliance clusters using Nimbus [intermediate]
- Building an educational appliance from Ubuntu 10.04 [intermediate]
- Customizing and registering Grid Appliance images using Eucalyptus [intermediate]

Tutorial Topic 4: High Performance Computing

- Basic High Performance Computing [novice]
- Running Hadoop as a batch job using MyHadoop [novice]
- Performance Analysis with Vampir [advanced]
- Instrumentation and tracing with VampirTrace [advanced]

Tutorial Topic 5: Experiment Management

- Running interactive experiments [novice]
- Running workflow experiments using Pegasus
  - Pegasus 4.0 on FutureGrid Walkthrough [novice]
  - Pegasus 4.0 on FutureGrid Tutorial [intermediary]

- Pegasus 4.0 on FutureGrid Virtual Cluster [advanced]

Tutorial Topic 6: Image Management and Rain

- Using Image Management and Rain [novice]

Tutorial Topic 7: Storage

- Using HPSS from FutureGrid [novice]

Other Tutorials and Educational Materials

- Additional tutorials on FutureGrid-related technologies
- FutureGrid community educational materials
- CI Tutor performance tutorials (requires a brief registration process to view content)

- FutureGrid Grid Appliance for Nimbus and Eucalyptus
- One-click Hadoop WordCount on Eucalyptus FutureGrid
- test

< Image Management and Rain in FutureGrid up FutureGrid Grid Appliance for Nimbus and Eucalyptus >

Add child page Printer-friendly version Bookmark this Subscribe to: This post PDF version

The FutureGrid project is funded by the National Science Foundation (NSF) and is led by **Indiana University** with **University of Chicago**, **University of Florida**, **San Diego Supercomputing Center**, **Texas Advanced Computing Center**, **University of Virginia**, **University of Tennessee**, **University of Southern California**, **Dresden**, **Purdue University**, and **Grid 5000** as partner sites. This material is based upon work supported in part by the National Science Foundation under Grant No. 0910812.

References

1. M. Isard, M. Budiu, Y. Yu, A. Birrell, D. Fetterly, Dryad: Distributed data-parallel programs from sequential building blocks, in: ACM SIGOPS Operating Systems Review, ACM Press, 2007, pp. 59-72
2. J.Ekanayake, H.Li, B.Zhang, T.Gunarathne, S.Bae, J.Qiu, G.Fox, Twister: A Runtime for iterative MapReduce, in: Proceedings of the First International Workshop on MapReduce and its Applications of ACM HPDC 2010 conference June 20-25, 2010, ACM, Chicago, Illinois, 2010.
3. Daytona iterative map-reduce framework. http://research.microsoft.com/en-us/projects/daytona/.
4. Y. Bu, B. Howe, M. Balazinska, M.D. Ernst, HaLoop: Efficient Iterative Data Processing on Large Clusters, in: The 36th International Conference on Very Large Data Bases, VLDB Endowment, Singapore, 2010.
5. Yanfeng Zhang , Qinxin Gao , Lixin Gao , Cuirong Wang, iMapReduce: A Distributed Computing Framework for Iterative Computation, Proceedings of the 2011 IEEE International Symposium on Parallel and Distributed Processing Workshops and PhD Forum, p.1112-1121, May 16-20, 2011
6. Tekin Bicer, David Chiu, and Gagan Agrawal. 2011. MATE-EC2: a middleware for processing data with AWS. In *Proceedings of the 2011 ACM international workshop on Many task computing on grids and supercomputers* (MTAGS '11). ACM, New York, NY, USA, 59-68.
7. Yandong Wang, Xinyu Que, Weikuan Yu, Dror Goldenberg, and Dhiraj Sehgal. 2011. Hadoop acceleration through network levitated merge. In *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis* (SC '11). ACM, New York, NY, USA, , Article 57 , 10 pages.
8. Karthik Kambatla, Naresh Rapolu, Suresh Jagannathan, and Ananth Grama. Asynchronous Algorithms in MapReduce. In *IEEE International Conference on Cluster Computing (CLUSTER)*, 2010.
9. T. Condie, N. Conway, P. Alvaro, J. M. Hellerstein, K. Elmleegy, and R. Sears. Mapreduce online. In NSDI, 2010.
10. M. Chowdhury, M. Zaharia, J. Ma, M.I. Jordan and I. Stoica, Managing Data Transfers in Computer Clusters with Orchestra *SIGCOMM 2011*, August 2011
11. M. Zaharia, M. Chowdhury, M.J. Franklin, S. Shenker and I. Stoica. Spark: Cluster Computing with Working Sets, *HotCloud 2010*, June 2010.
12. Huan Liu and Dan Orban. Cloud MapReduce: a MapReduce Implementation on top of a Cloud Operating System. In 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, pages 464–474, 2011
13. AppEngine MapReduce, July 25th 2011; http://code.google.com/p/appengine-mapreduce.
14. J. Dean, S. Ghemawat, MapReduce: simplified data processing on large clusters, Commun. ACM, 51 (2008) 107-113.