Technology for X-Informatics: MapReduce

June 19 2013

Geoffrey Fox

gcf@indiana.edu

http://www.infomall.org/X-InformaticsSpring2013/index.html

Associate Dean for Research, School of Informatics and Computing

Indiana University Bloomington

2013

Big Data Ecosystem in One Sentence

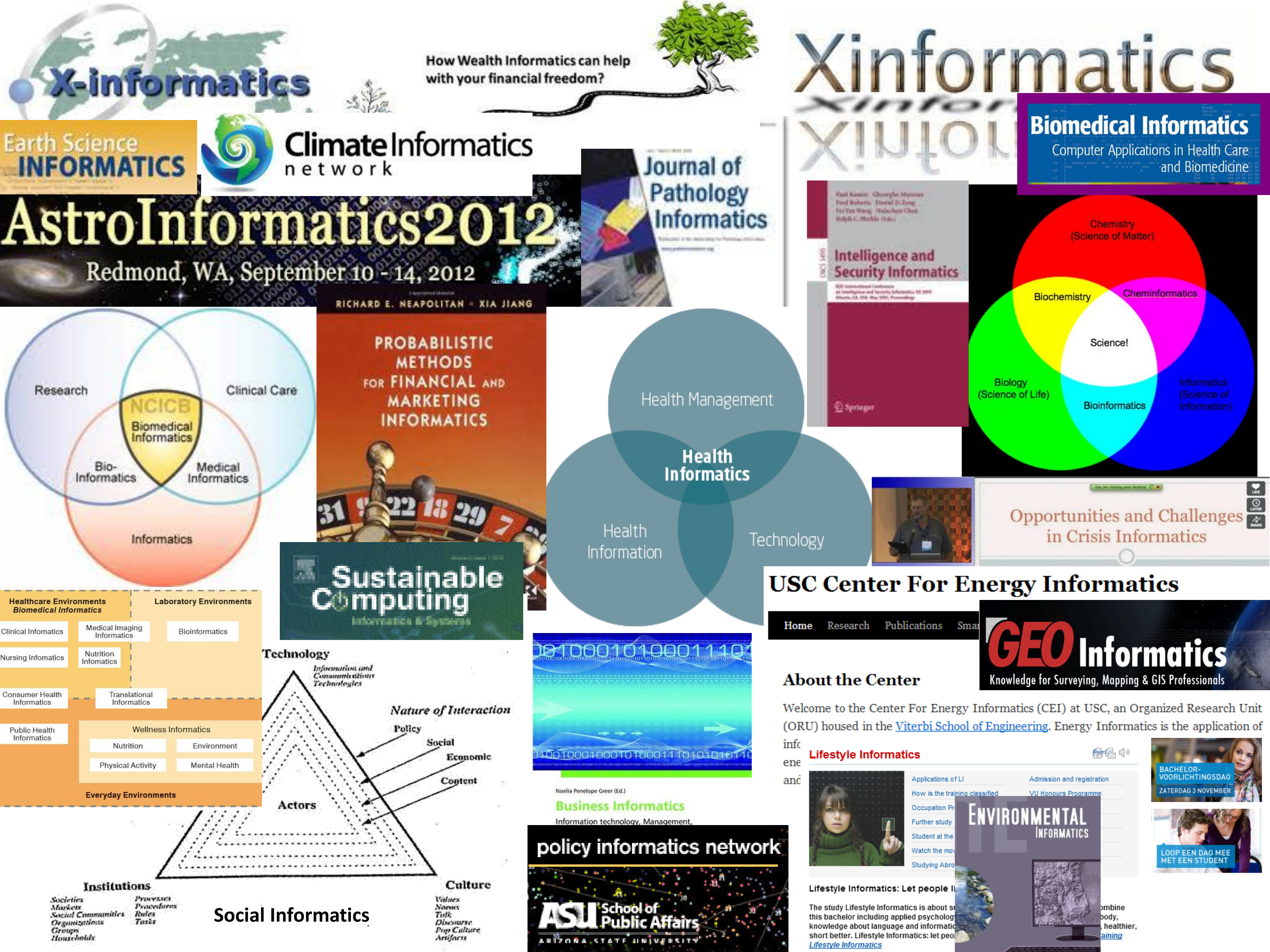
Use Clouds running Data Analytics Collaboratively processing Big Data to solve problems in X-Informatics (or e-X)

X = Astronomy, Biology, Biomedicine, Business, Chemistry, Climate, Crisis, Earth Science, Energy, Environment, Finance, Health, Intelligence, Lifestyle, Marketing, Medicine, Pathology, Policy, Radar, Security, Sensor, Social, Sustainability, Wealth and Wellness with more fields (physics) defined implicitly

Spans Industry and Science (research)

Education: Data Science see recent New York Times articles

http://datascience101.wordpress.com/2013/04/13/new-york-times-data-science-articles/

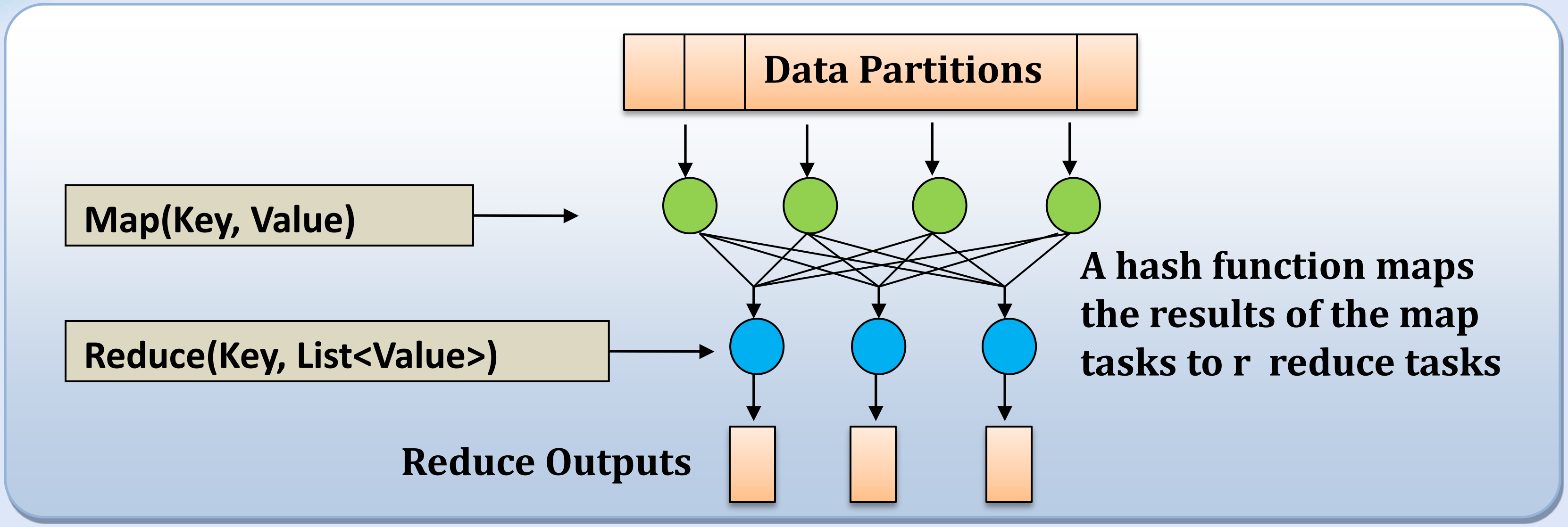

X-informatics
How Wealth Informatics can help with your financial freedom?
Xinformatics
Earth Science INFORMATICS
Climate Informatics network
Journal of Pathology Informatics
Biomedical Informatics
Computer Applications in Health Care and Biomedicine
AstroInformatics2012
Redmond, WA, September 10 - 14, 2012
Intelligence and Security Informatics
Springer
Chemistry (Science of Matter)
Biochemistry
Cheminformatics
Science!
Biology (Science of Life)
Bioinformatics
Informatics (Science of Information)
Research
Clinical Care
NCICB
Biomedical Informatics
Bio-Informatics
Medical Informatics
Informatics
RICHARD E. NEAPOLITAN • XIA JIANG
PROBABILISTIC METHODS FOR FINANCIAL AND MARKETING INFORMATICS
Health Management
Health Informatics
Health Information
Technology
Opportunities and Challenges in Crisis Informatics
Sustainable Computing
Informatics & Systems
USC Center For Energy Informatics
Home
Research
Publications
GEO Informatics
Knowledge for Surveying, Mapping & GIS Professionals
About the Center
Welcome to the Center For Energy Informatics (CEI) at USC, an Organized Research Unit (ORU) housed in the Viterbi School of Engineering. Energy Informatics is the application of
Healthcare Environments
Biomedical Informatics
Laboratory Environments
Clinical Infomatics
Medical Imaging Informatics
Bioinformatics
Nursing Infomatics
Nutrition Infomatics
Consumer Health Informatics
Translational Informatics
Public Health Informatics
Wellness Informatics
Nutrition
Environment
Physical Activity
Mental Health
Everyday Environments
Technology
Information and Communications Technologies
Nature of Interaction
Policy
Social
Economic
Content
Actors
Institutions
Societies
Markets
Social Communities
Organizations
Groups
Households
Processes
Procedures
Rules
Tasks
Culture
Values
Norms
Talk
Discourse
Pop Culture
Artifacts
Social Informatics
Noelia Penelope Greer (Ed.)
Business Informatics
Information technology, Management,
policy informatics network
ASU School of Public Affairs
ARIZONA STATE UNIVERSITY
Lifestyle Informatics
Applications of LI
How is the training classified
Further study
Admission and registration
VU Honours Programme
ENVIRONMENTAL INFORMATICS
BACHELOR-VOORLICHTINGSDAG
ZATERDAG 3 NOVEMBER
LOOP EEN DAG MEE MET EEN STUDENT
Lifestyle Informatics

MapReduce

Introduction

MapReduce

A parallel Runtime coming from Information Retrieval



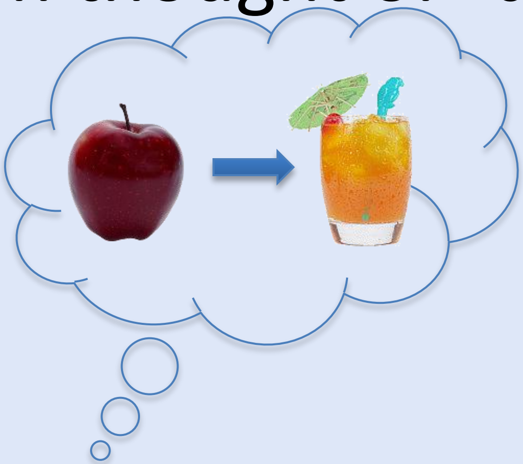
- Implementations support:
  - Splitting of data
  - Passing the output of map functions to reduce functions
  - Sorting the inputs to the reduce function based on the intermediate keys
  - Quality of services

SALSA

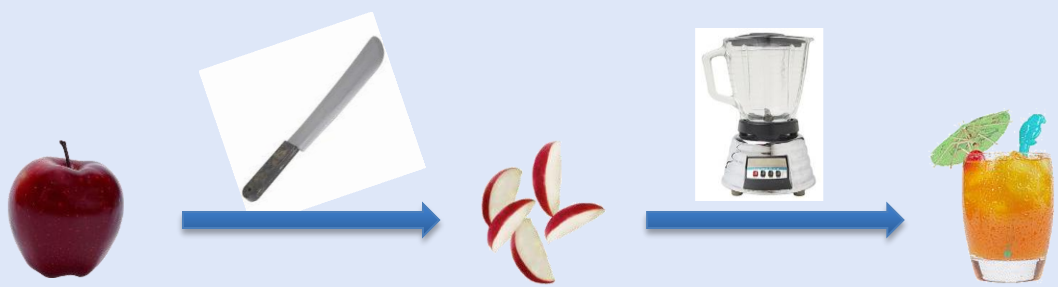
Sam’s Problem

http://www.slideshare.net/esaliya/mapreduce-in-simple-terms

- Sam thought of “drinking” the apple

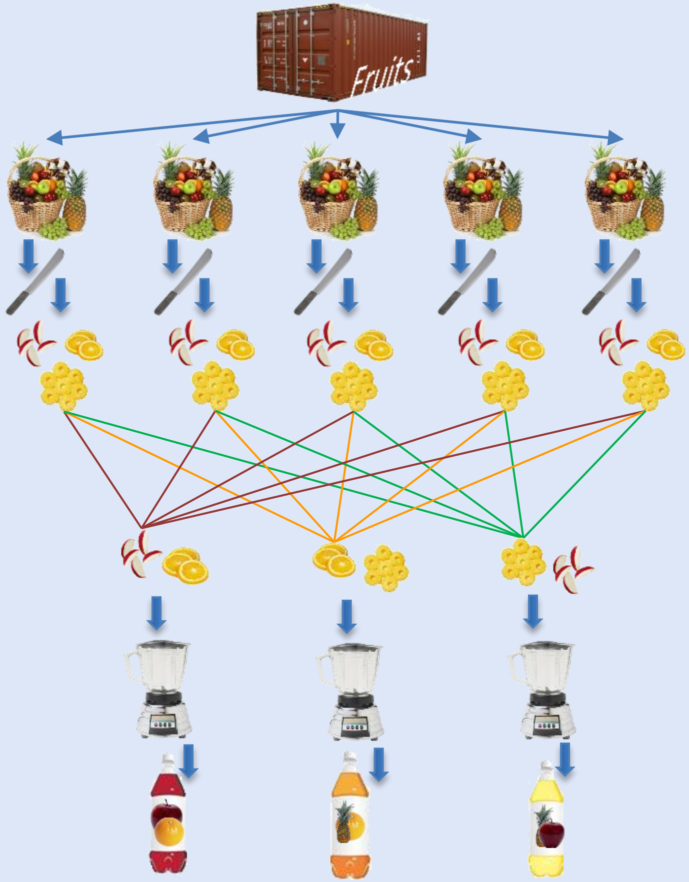


- He used a  to cut the  and a  to make juice.



Creative Sam

- Implemented a *parallel* version of his innovation



Each input to a map is a list of <key, value> pairs

(<a, > , <o, > , <p, > , …)

A *list of <key, value>* pairs mapped into another *list of <key, value>* pairs which gets grouped by the key and reduced into a *list of values*

Each output of slice is a list of <key, value> pairs

(<a’, > , <o’, > , <p’, >)

Grouped by key

The idea of Map Reduce in Data Intensive Computing

Ea… ossibly a li… ing mechanism)

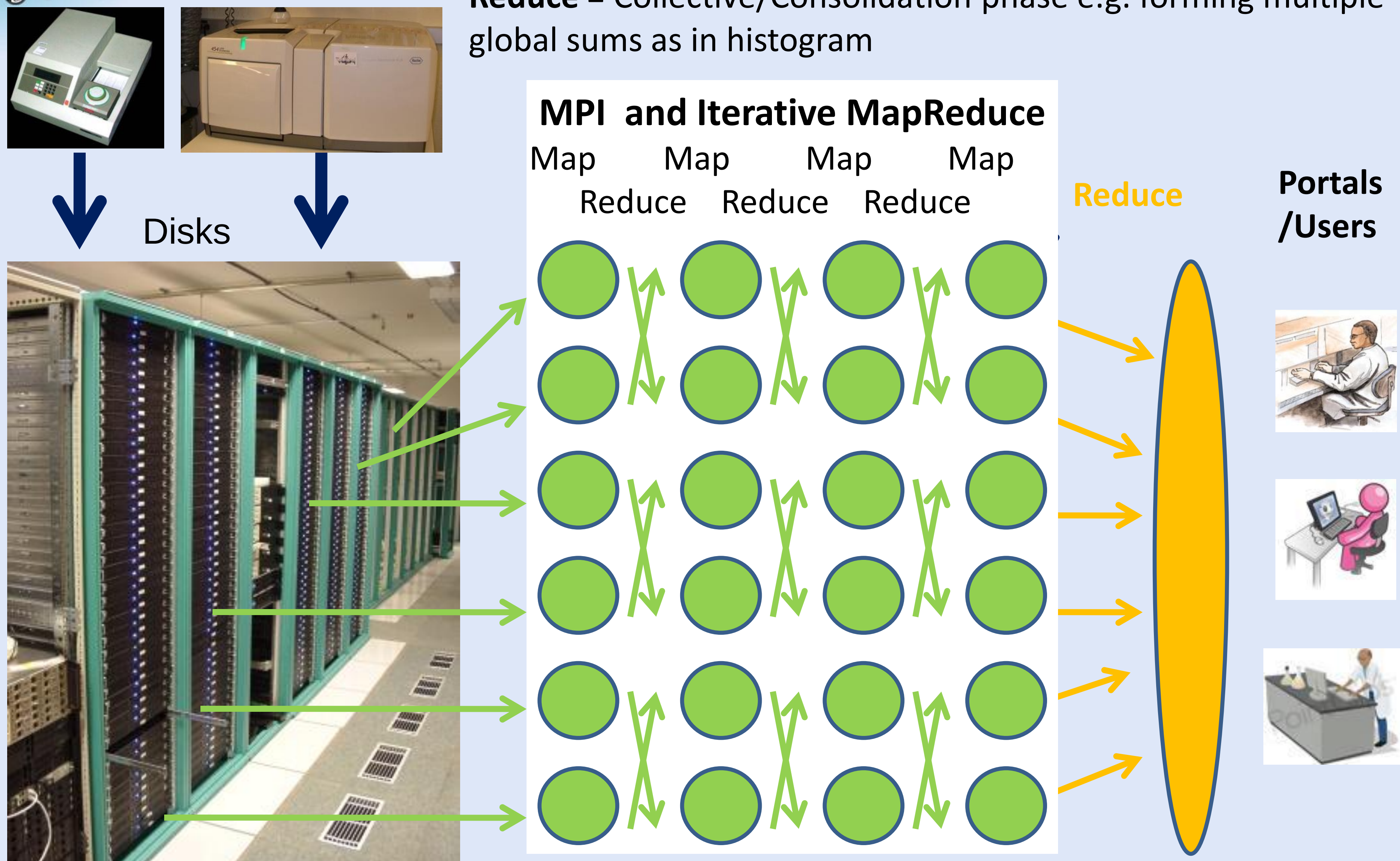
e.g. <ao, (…)>

Reduced into a list of values

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



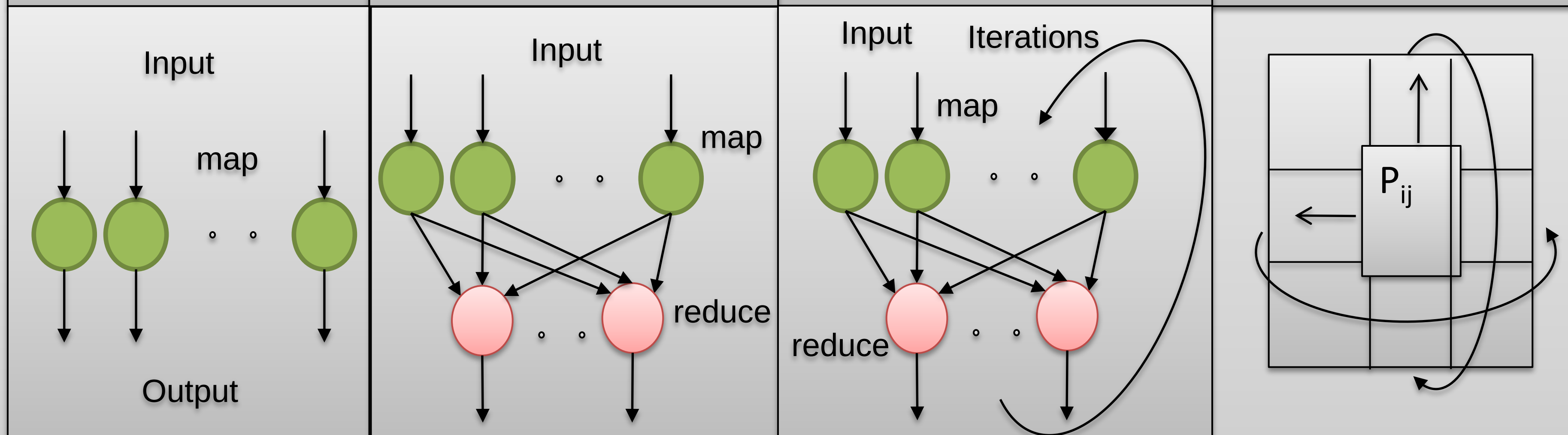
SALSA

MapReduce

Advanced Topics

4 Forms of MapReduce

(a) Map Only	(b) Classic MapReduce	(c) Iterative MapReduce	(d) Loosely Synchronous

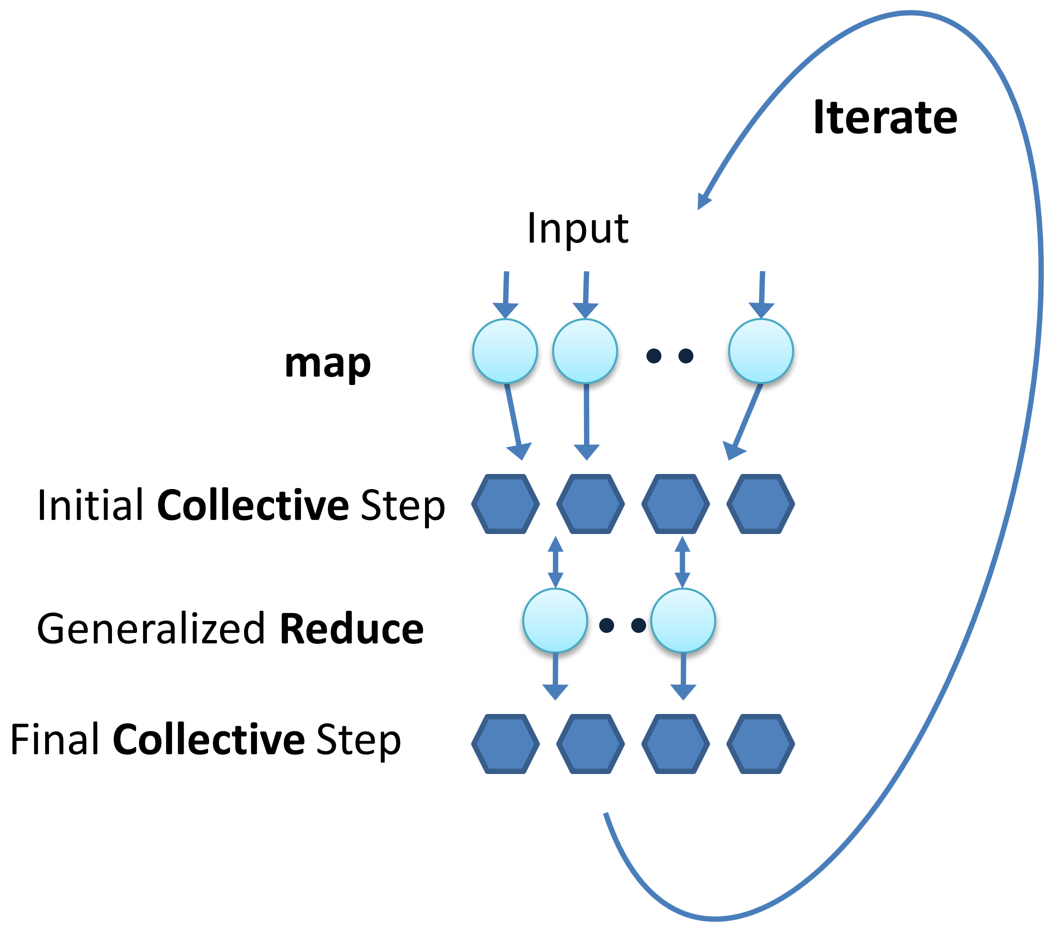


(a) Map Only	(b) Classic MapReduce	(c) Iterative MapReduce	(d) Loosely Synchronous
BLAST Analysis Parametric sweep Pleasingly Parallel	High Energy Physics (HEP) Histograms Distributed search	Expectation maximization Clustering e.g. Kmeans Linear Algebra, Page Rank	Classic MPI PDE Solvers and particle dynamics
← **Domain of MapReduce and Iterative Extensions** → **Science Clouds**			**MPI** **Exascale**

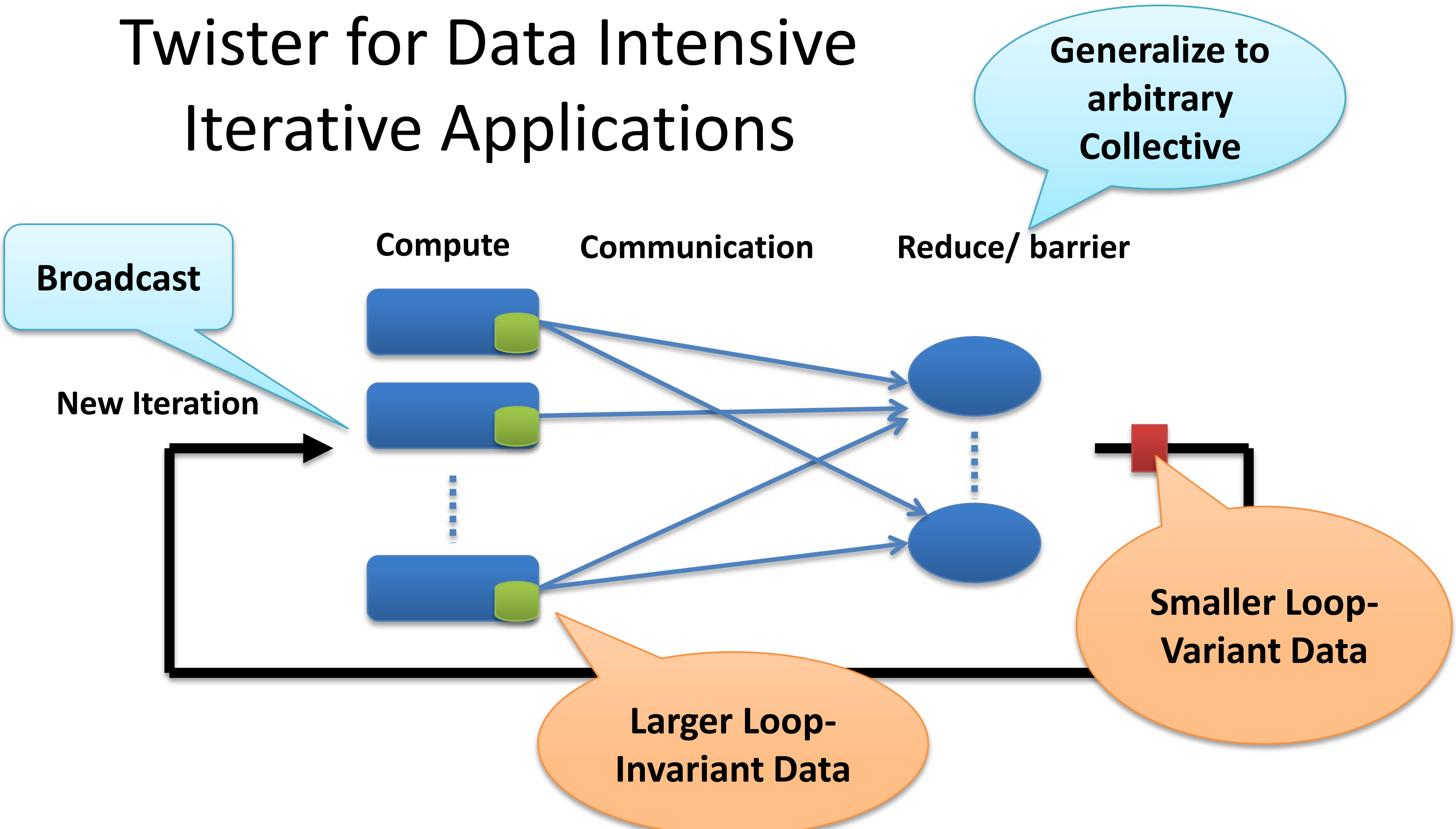
MPI is Map followed by Point to Point Communication – as in style d)

Map Collective Model (Judy Qiu)

- Combine MPI and MapReduce ideas
- Implement collectives optimally on Infiniband, Azure, Amazon ……
- Applies for general data; not necessarily key-value pairs



Twister for Data Intensive Iterative Applications

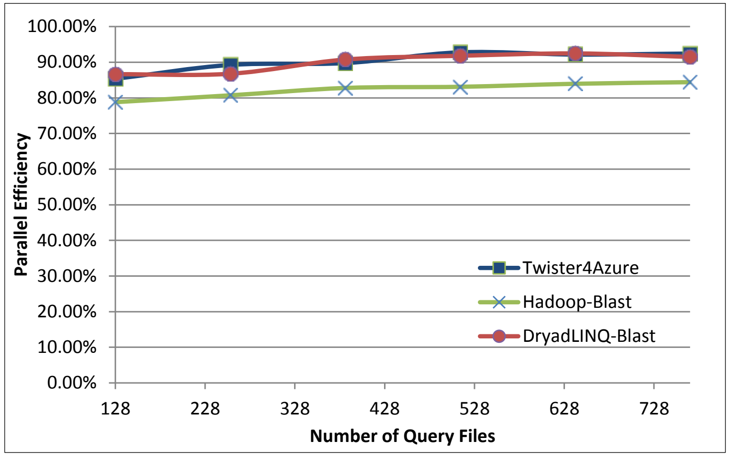


- (Iterative) MapReduce structure with Map-Collective is framework
- Twister runs on Linux or Azure
- Twister4Azure is built on top of Azure **tables**, **queues**, **storage**

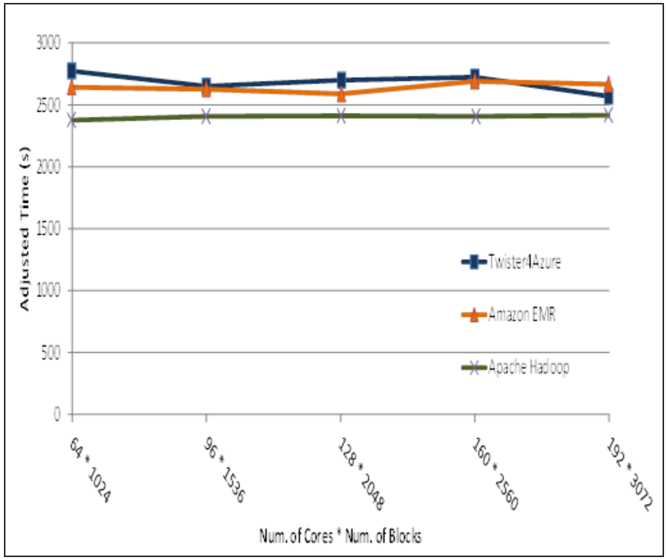
Qiu, Gunarathne

Pleasingly Parallel Performance Comparisons

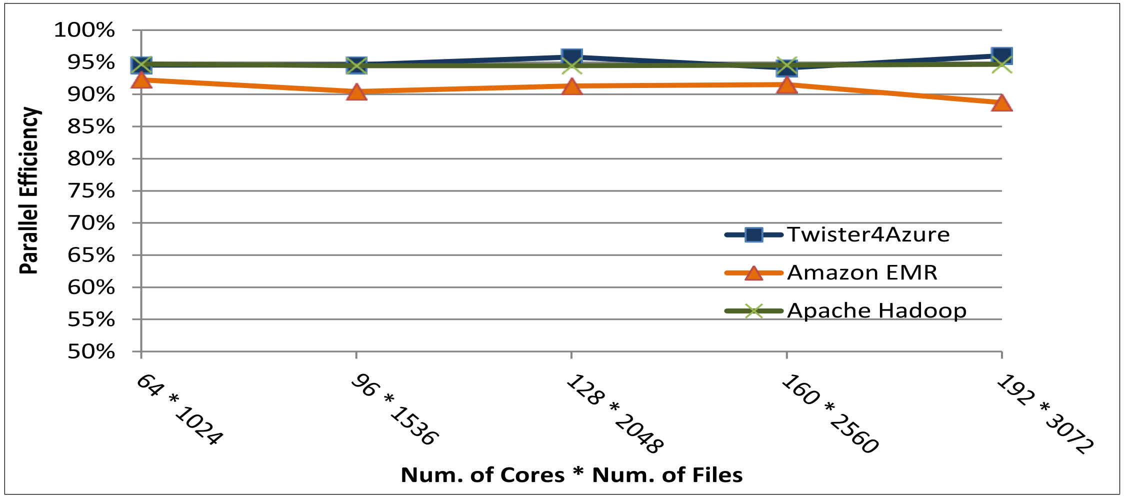
BLAST Sequence Search

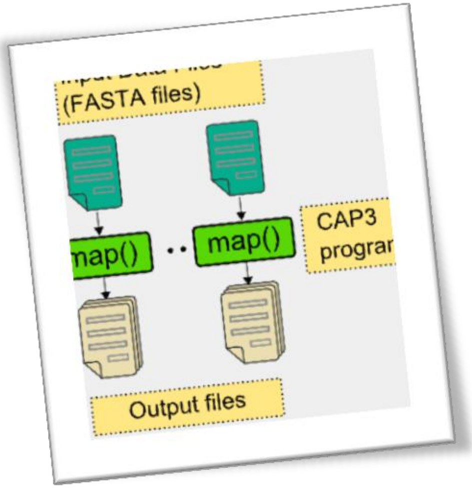


Smith Waterman Sequence Alignment

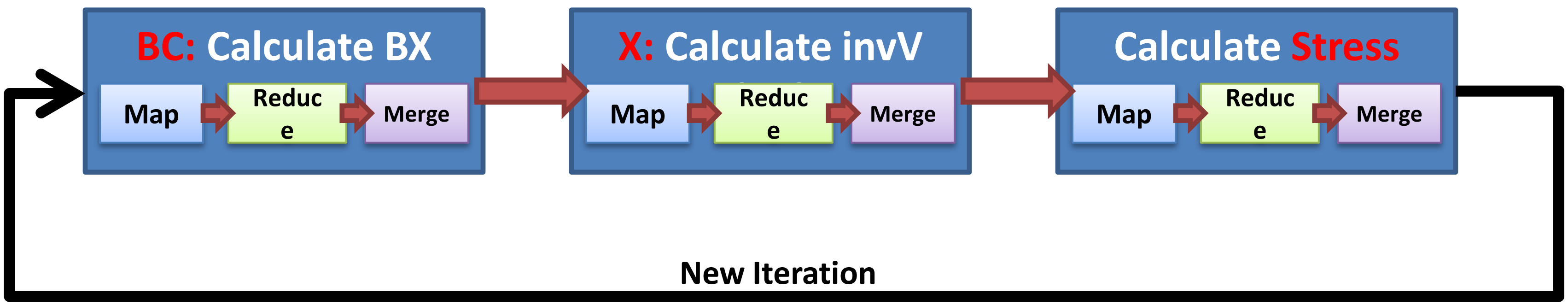


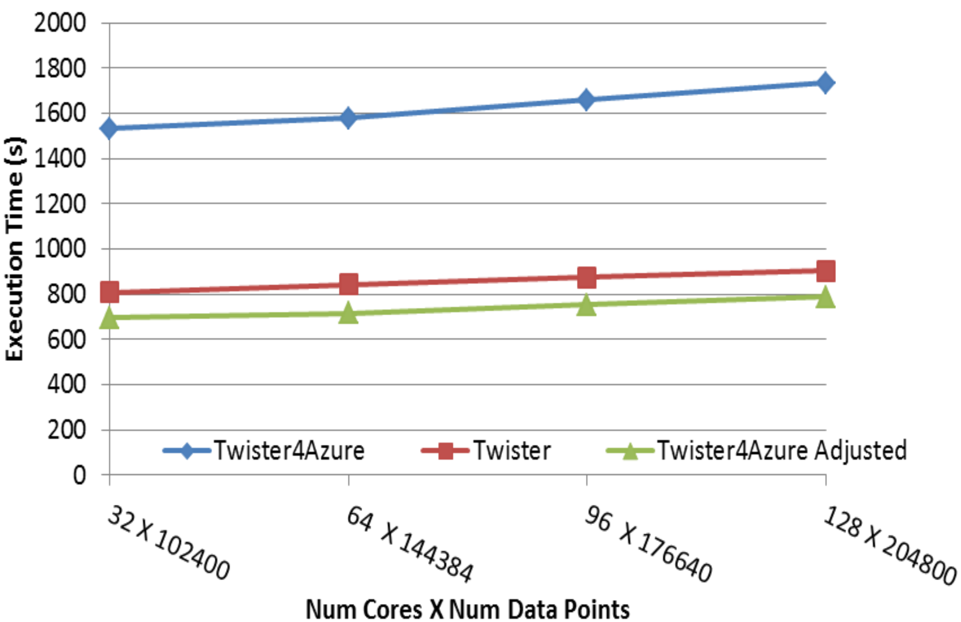
Cap3 Sequence Assembly



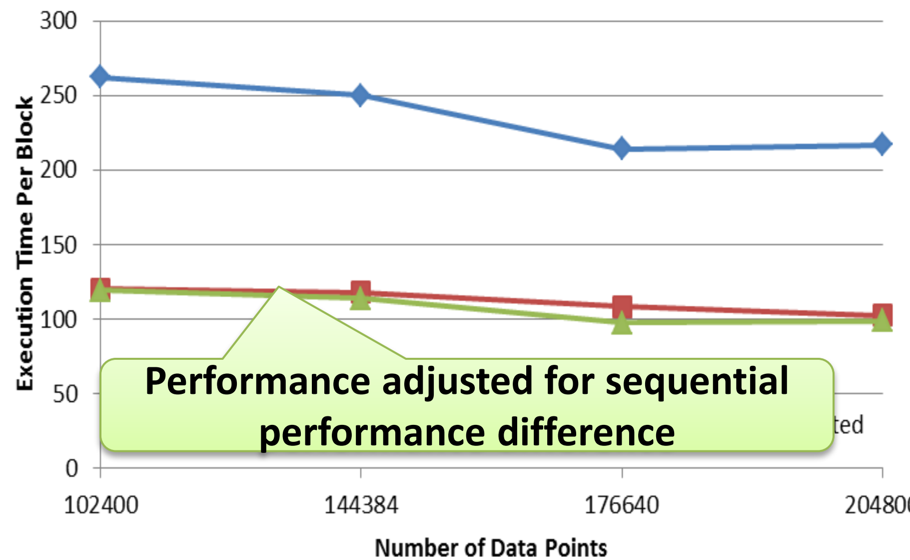


Multi Dimensional Scaling



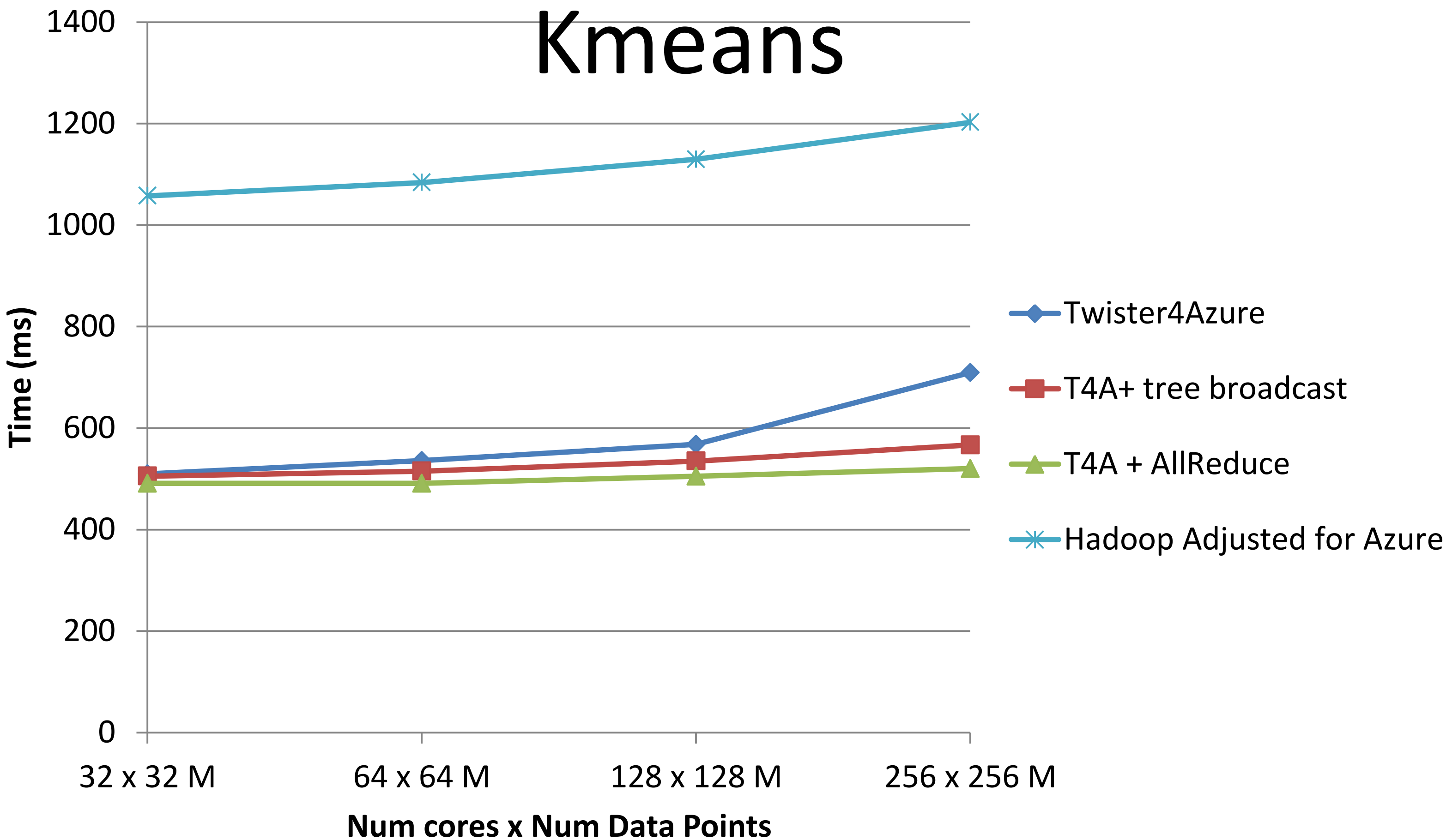


Weak Scaling



Data Size Scaling

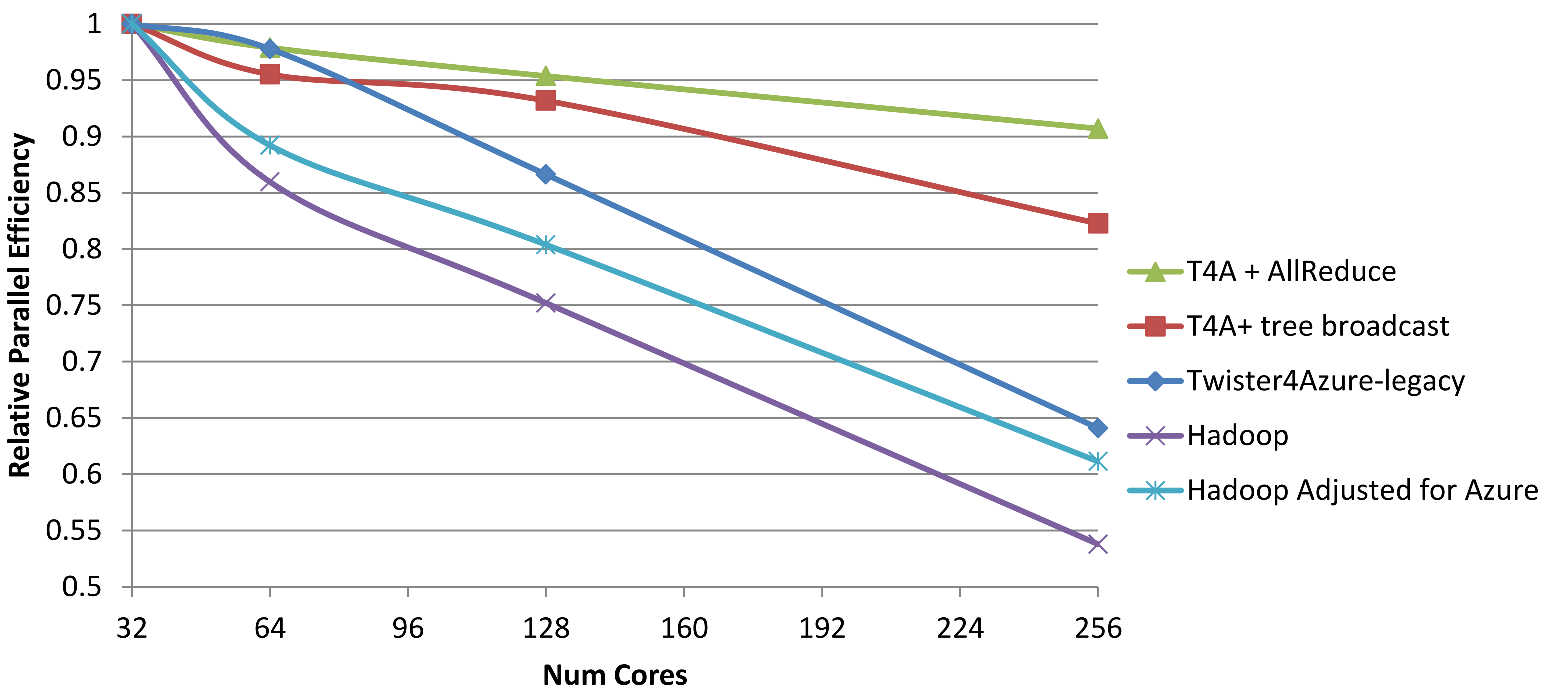
Scalable Parallel Scientific Computing Using Twister4Azure. Thilina Gunarathne, BingJing Zang, Tak-Lon Wu and Judy Qiu. Submitted to Journal of Future Generation Computer Systems. (Invited as one of the best 6 papers of UCC 2011)



Investigating Improved communication (collective) routines

Hadoop adjusted for Azure: implies Hadoop KMeans run time adjusted for the performance difference of iDataplex vs Azure

Kmeans Strong Scaling (with Hadoop Adjusted)



128 Million data points. 500 Centroids (clusters). 20 Dimensions. 10 iterations

Parallel efficiency relative to the 32 core run time.

Note Hadoop slower by factor of 2

Relative Efficiency = Time(32 cores)/Time(N cores) . 32/N