

A photograph showing an underwater view of a blue wave tunnel, with water swirling and creating a circular opening in the center.

Data Intensive Computing

PASIG
June , 2009

Simon CW See
HPC & Cloud Computing
Technology Center
Sun Microsystems, Inc.

Notice of Confidentiality

This presentation contains information related to projects that are currently in planning and/or development stages. This information represents the current intentions of Sun Microsystems, Inc. *However, all aspects of these projects, including, but not limited to, funding, availability, shipping dates, configurations, capacities, performance, and all other characteristics are subject to change and/or cancellation without notice.*

This material is confidential to Sun Microsystems and should be disclosed to non-employees only under the terms of an executed confidential-disclosure agreement.

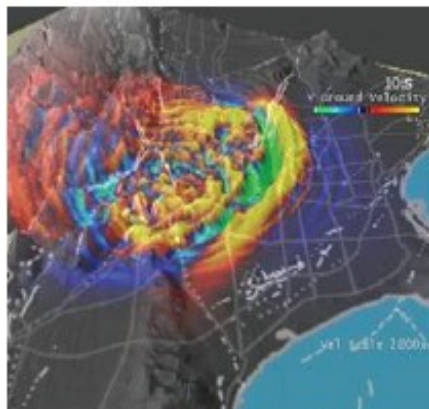


Steve Schlosser,
Michael Ryan, Dave
O'Hallaron(IRP)
*~1 PB of data
uncompressed*_{Image}



Sloan Digital Sky Survey

- New Mexico telescope captures 200 GB image data / day
- Latest dataset release: 10 TB, 287 million celestial objects
- SkyServer provides SQL access



TerashakeSims
~1 PB for LA basin

Cardi CT
4GB per 3D scan,
1000s of scans/year



Trends in Astronomy

CMB Surveys (pixels)

- 1990 **COBE** 1000
- 2000 Boomerang 10,000
- 2002 CBI 50,000
- 2003 **WMAP** 1 Million
- 2008 Planck 10 Million

Angular Galaxy Surveys (obj)

- 1970 Lick 1M
- 1990 APM 2M
- 2005 **SDSS** 200M
- 2008 VISTA 1000M
- 2012 **LSST** 3000M

Time Domain

- QUEST
- **SDSS Extension survey**
- Dark Energy Camera
- **PanStarrs**
- SNAP...
- **LSST...**

Galaxy Redshift Surveys (obj)

- 1986 CfA 3500
- 1996 LCRS 23000
- 2003 **2dF** 250000
- 2005 **SDSS** 750000

Petabytes/year by the end of the decade...

Source : Alex Szalay JHU, Scientific Applications of Large Database

The Big Picture of Genomic Data

- The **Imminent** Data “deluge”
 - Exponential growth of sequence data
 - Unstoppable growth of microarray data
 - New petabytes of data set from Cell imaging technology
 - “I am terrified by terabytes” -- *Anonymous*
 - “I am petrified by petabytes” -- *Jim Gray*
- Technology Innovation
 - “Moore’s Law”– computing capacity doubles every 18 months
 - True for the past 30 years
 - Will hold true for the next 10 years (hopefully)
 - No more or limit frequency increase but many more cores
- Moore’s Law outpaced by growth of genomic data!

Growth of GenBank

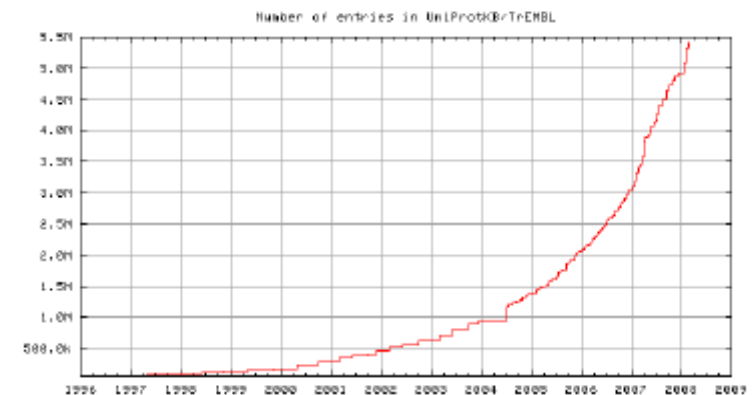
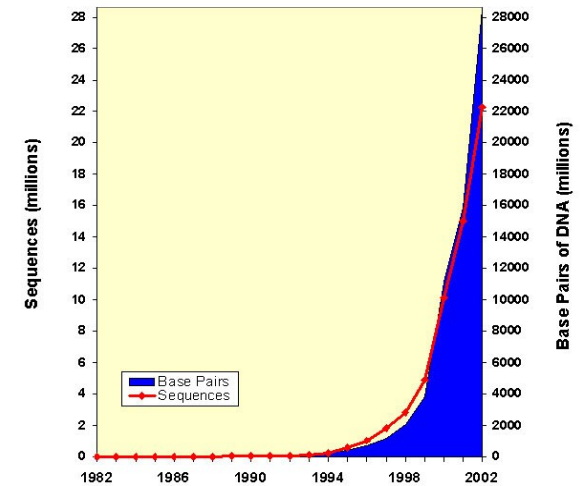
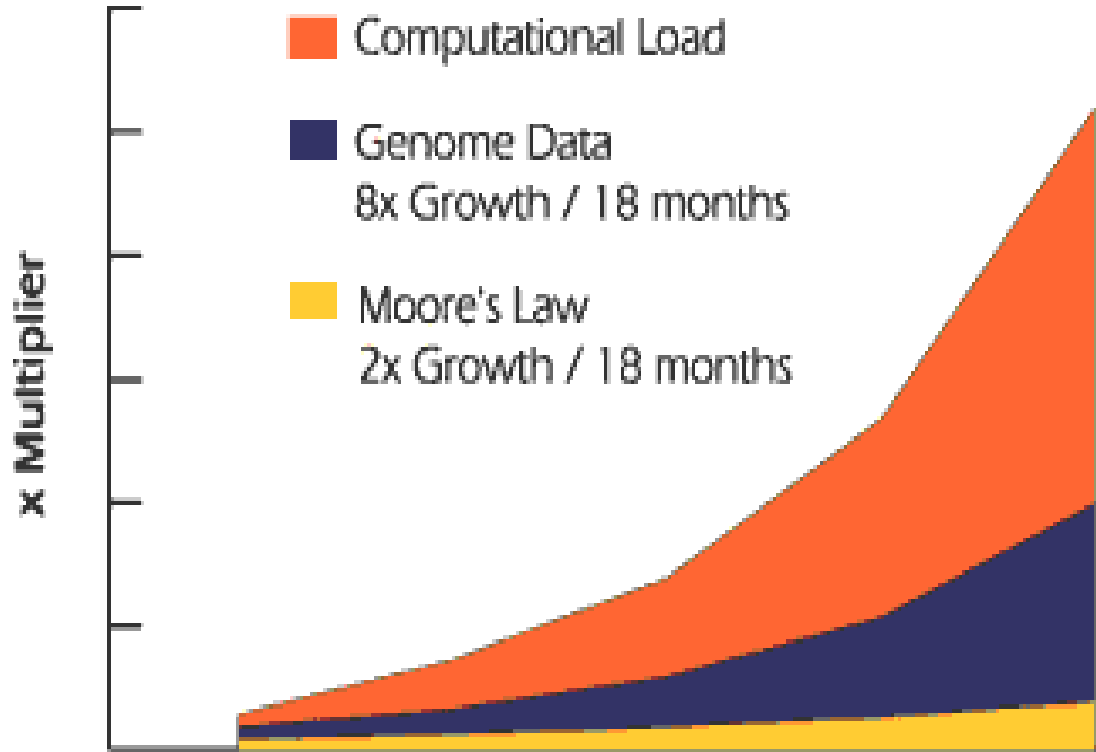


Figure 3. Growth of protein sequences in UniProtKB/TrEMBL database. Release 38.2 (08-Apr-2008): it contains 5577054 sequence entries comprising 1803615193 amino acids. 216176 sequences have been added since release 38, the sequence data of 422 existing entries has been updated and the annotations of 1612992 entries have been revised. [34]

The Gap → Computational Challenge



Growth of GenBank

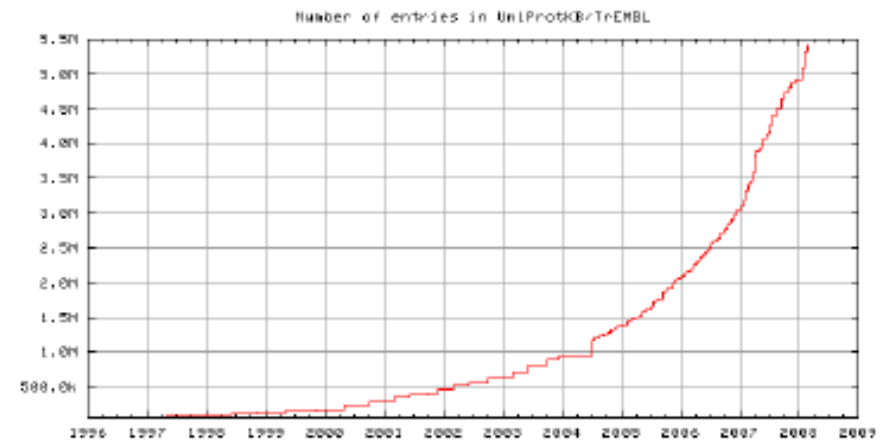
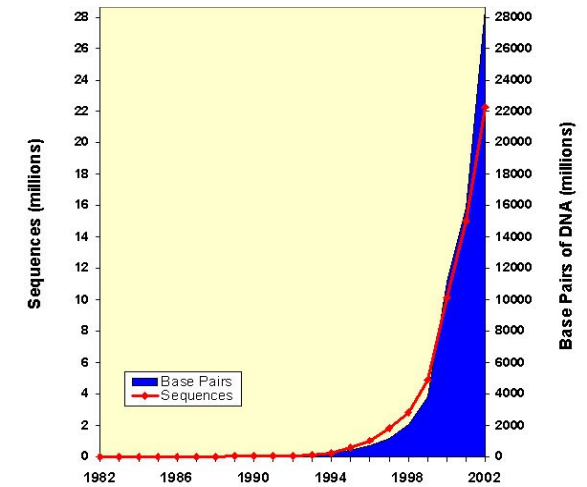
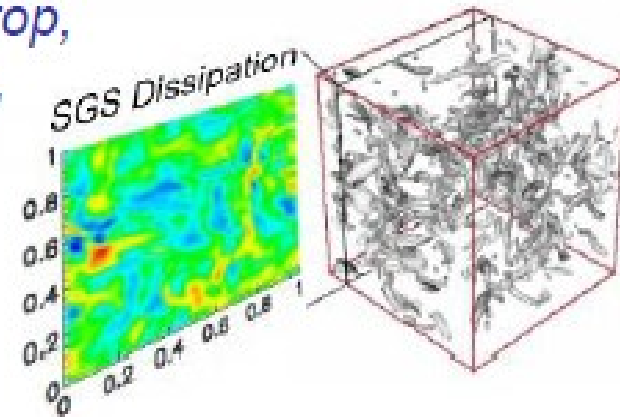
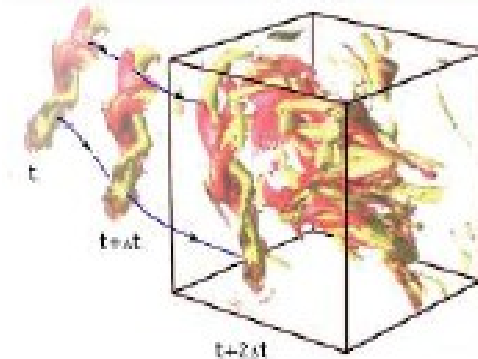
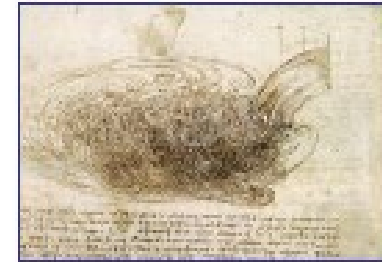


Figure 3. Growth of protein sequences in UniProtKB/TrEMBL database. Release 38.2 (08-Apr-2008): it contains 5577054 sequence entries comprising 1803615193 amino acids. 216176 sequences have been added since release 38, the sequence data of 422 existing entries has been updated and the annotations of 1612992 entries have been revised. [34]

Immersive Analysis - Turbulence

- **Unique turbulence database**
 - Consecutive snapshots of a $1,024^3$ simulation of turbulence: now 30 Terabytes
 - Hilbert-curve spatial index
 - Soon $6K^3$ and 300 Terabytes
 - Treat it as an experiment, observe the database!
 - Throw test particles in from your laptop, immerse yourself into the simulation, like in the movie *Twister*
- New paradigm for analyzing HPC simulations!

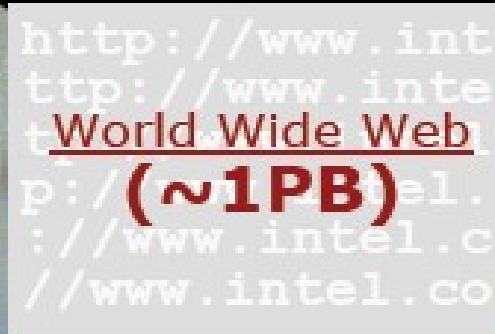




Particle Physics Large Hadron Collider
(15PB)



Human Genomics
(7000PB)
1GB / person
200PB+ captured
200% CAGR



World Wide Web
(~1PB)



Wikipedia
(10GB)
100% CAGR

Annual Email Traffic, no spam
(300PB+)



Internet Archive
(1PB+)

Estimated On-line RAM in Google
(8PB)



Personal Digital Photos
(1000PB+)
100% CAGR

200 of London's Traffic Cams
(8TB/day)

2004 Walmart Transaction DB
(500TB)

Typical Oil Company
(350TB+)

Merck Bio Research DB
(1.5TB/qtr)

UPMC Hospitals Imaging Data
(500TB/yr)



MIT Babytalk Speech Experiment
(1.4PB)

Terashake Earthquake Model of LA Basin
(1PB)

One Day of Instant Messaging in 2002
(750GB)

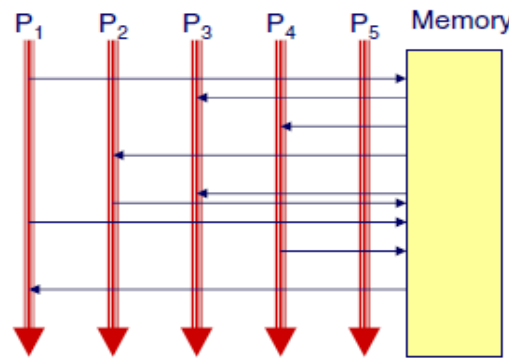
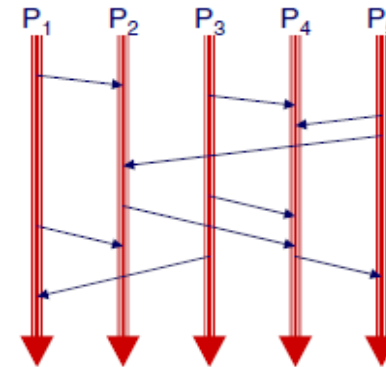
Total digital data to be created this year **270,000PB** (IDC)

Progamming Environment

Programming Model

Traditional HPC Prog Model

- Distributed Memory
 - > Message Passing Interface (MPI)
 - > PGAS
 - > UPC
- Shared Memory
 - > Pthreads
 - > OpenMP

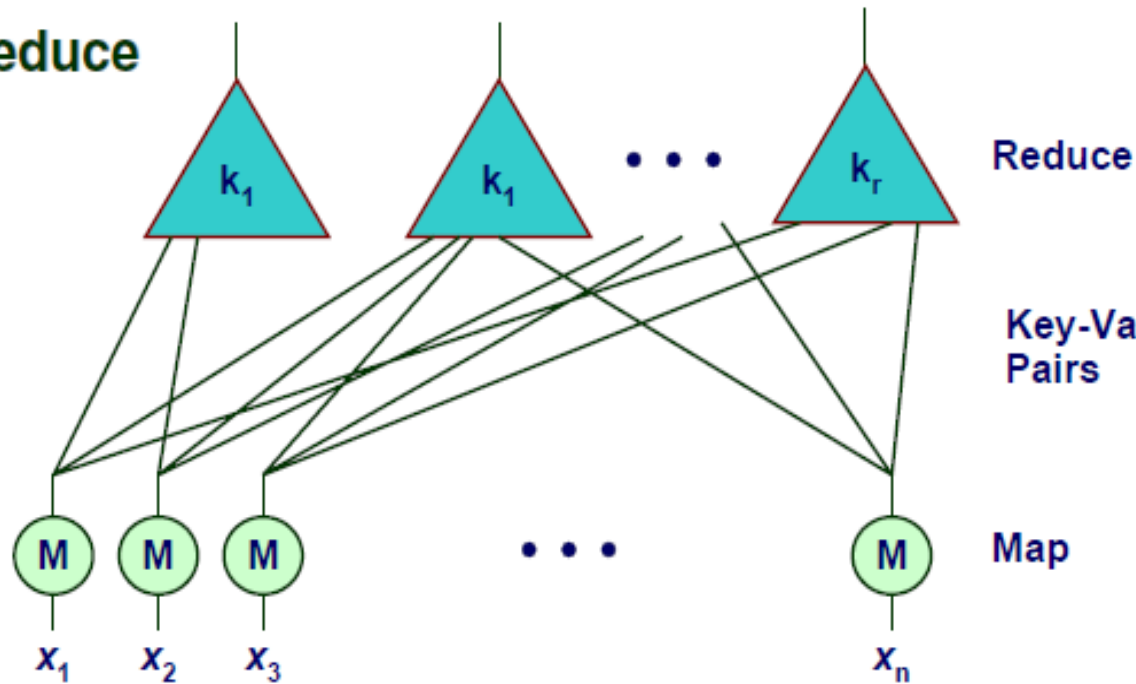


Message Passing / Shared Memory

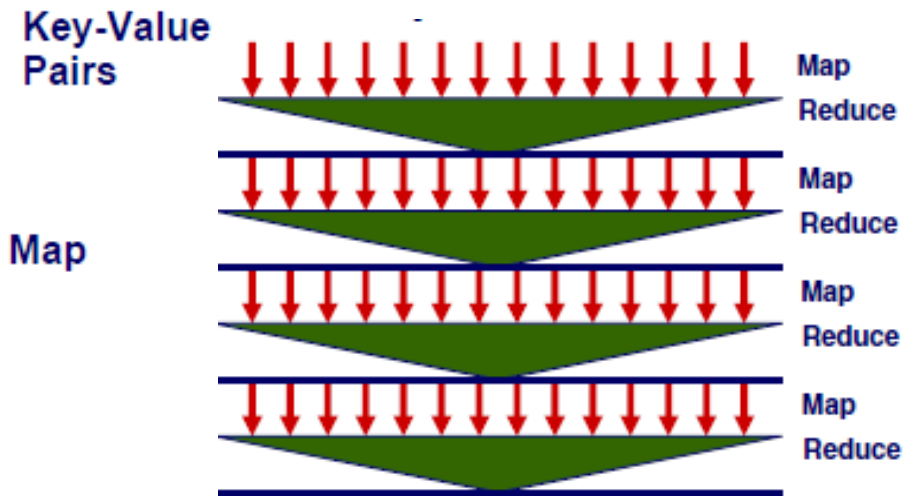
- ▣ Achieves very high performance when everything works well
- ▣ Requires careful tuning of programs
- ▣ Vulnerable to single points of failure

MapReduce Programming Model

MapReduce



- Map computation across many objects
 - E.g., 10^{10} Internet web pages
- Aggregate results in many different ways
- System deals with issues of resource allocation & reliability



Source : Dean GhemawatGhemawat: : ““MapReduceMapReduce: Simplified Data : Processing on Large ClustersProcessing Clusters””, OSDI 2004, 2004

Applications

Aggregate, store, and analyze data related to in-stream viewing behavior of Internet video audiences.

Analytics

Analyze and index textual information

Analyzing similarities of user's behavior.

Build scalable machine learning algorithms like canopy clustering, k-means and many more to come (naive bayes classifiers, others)

Charts calculation and web log analysis

Crawl Blog posts and later process them.

Crawling, processing, serving and log analysis

Data mining and blog crawling

Facial similarity and recognition across large datasets.

Filter and index our listings, removing exact duplicates and grouping similar ones.

Filtering and indexing listing, processing log analysis, and for recommendation data.

Flexible web search engine software

Gathering world wide DNS data in order to discover content distribution networks and configuration issues

Generating web graphs

Image based video copyright protection.

Image content based advertising and auto-tagging for social media.

Image processing environment for image-based product recommendation system

Image retrieval engine

Large scale image conversions

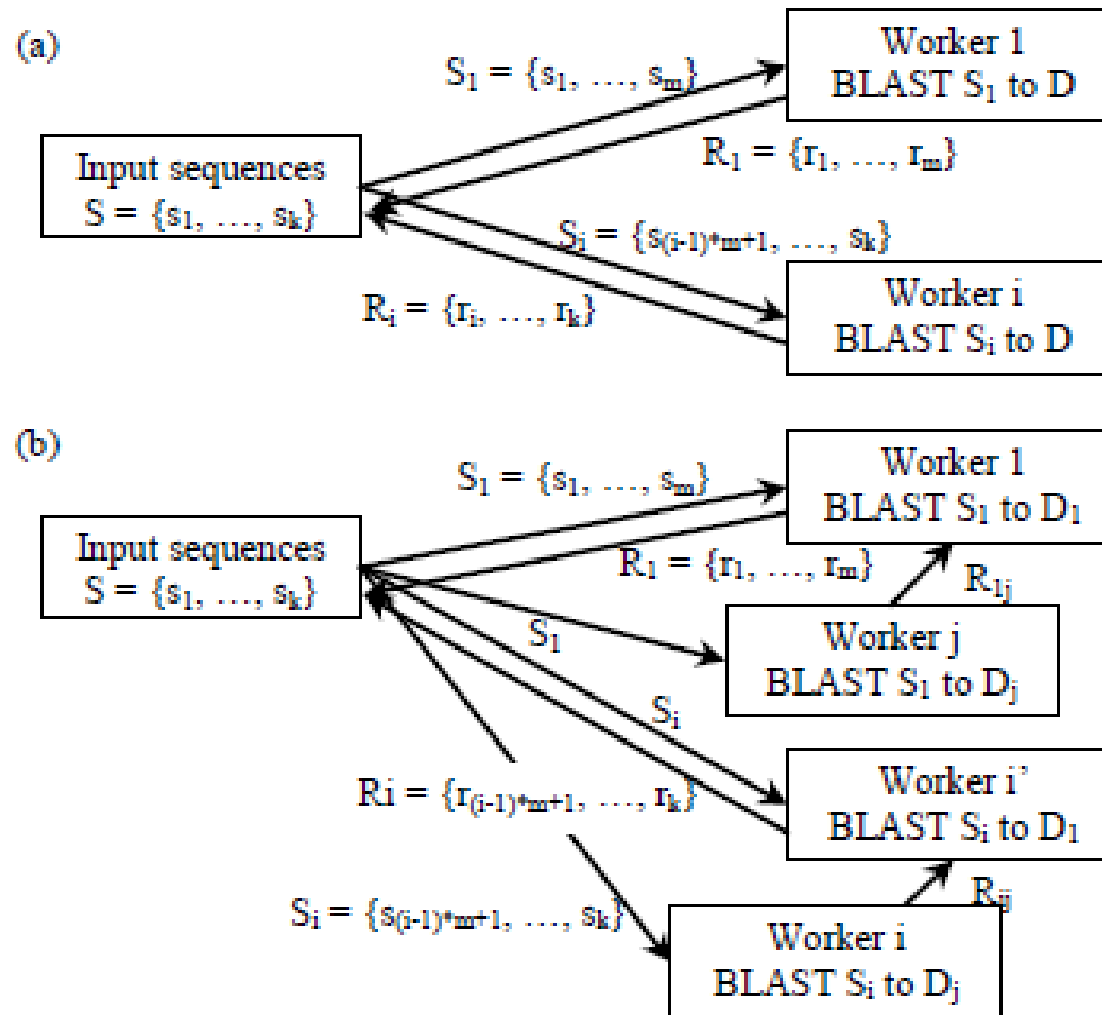
Generating web graphs

Image based video copyright protection.

Image content based advertising and auto-tagging for social media.

Image processing environment for image-based product recommendation system

Image retrieval engine



Source: Andréa Matsunaga, CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications

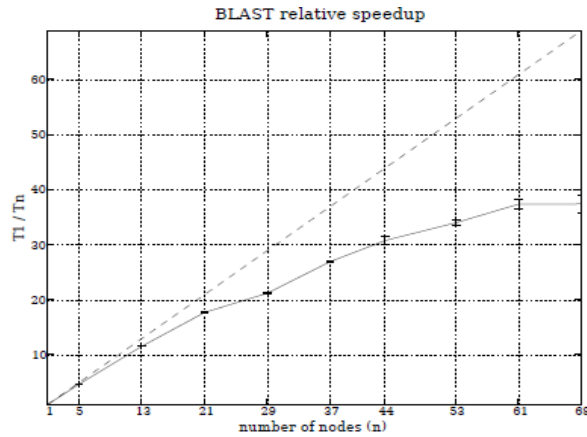


Fig. 1. BLAST relative speedup, average performance on 3 runs. Every node has four CPU cores, for a total of 276 cores with 69 nodes. Here, we ran a 10 sequences, 814 bases average length query against the nt database, using the tblastx program with an e-value threshold of 10^{-50} . We set the HDFS block size to 100 MB.

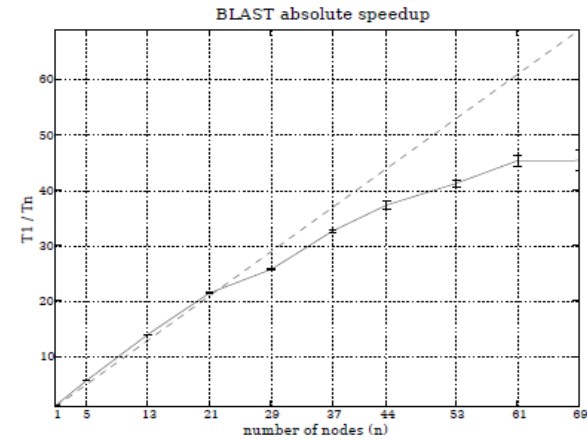


Fig. 2. BLAST absolute speedup, average performance on 3 runs. Every node has four CPU cores, for a total of 276 cores with 69 nodes. Parallel performance data is the same as in relative speedup. To get the reference implementation timing, we scheduled four blastall threads (one per CPU core) on a single machine. Two of the four threads were assigned a 3 query file, while the remaining two got a 2 query file.

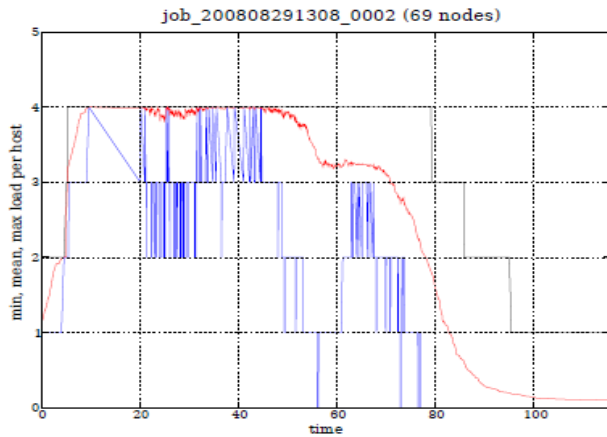


Fig. 3. Machine load pattern for a 69 nodes run. The blue, red and black lines indicate respectively the minimum, mean and maximum load per node.

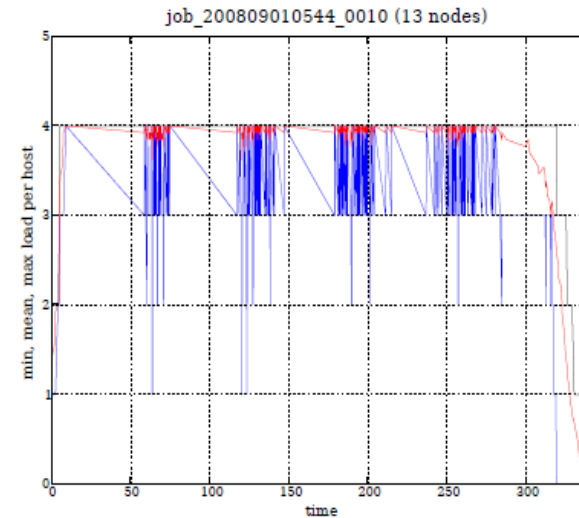


Fig. 4. Machine load pattern for a 13 nodes run. The blue, red and black lines indicate respectively the minimum, mean and maximum load per node.

Source : Massimo Gaggero, et. Al, Parallelizing bioinformatics applications with MapReduce

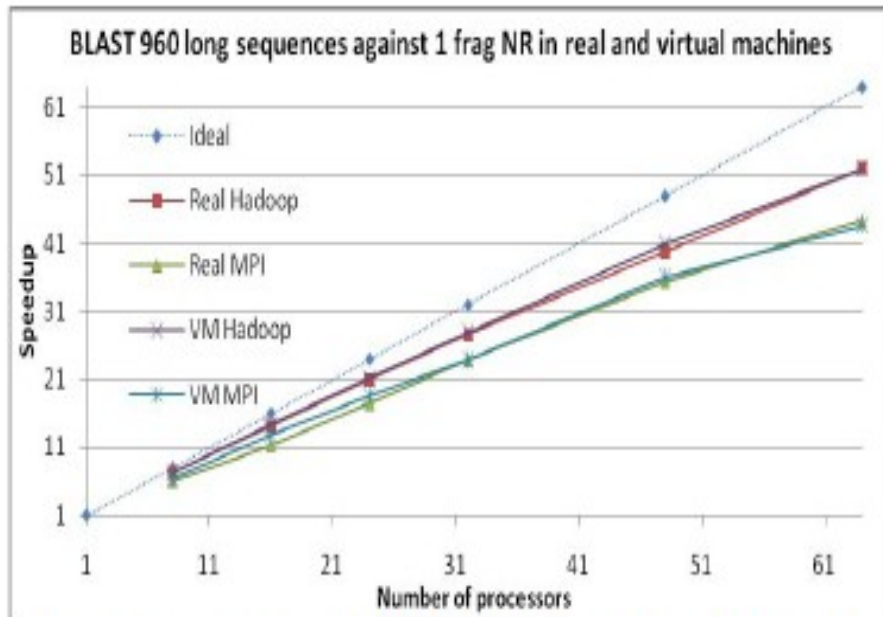


Figure 6. Comparison of BLAST speed-ups on physical and virtual machines. Speedup is calculated relative to sequential BLAST on virtual resource. VMs deliver performance comparable with physical machines when executing BLAST.

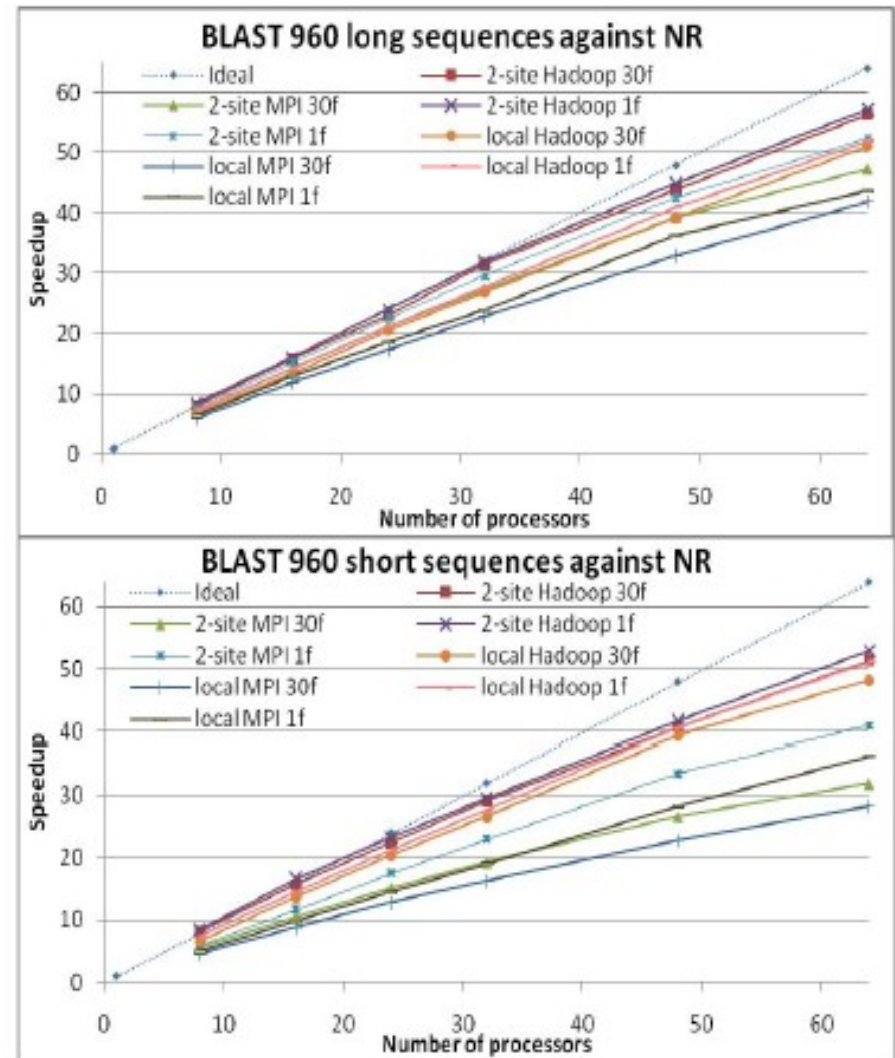
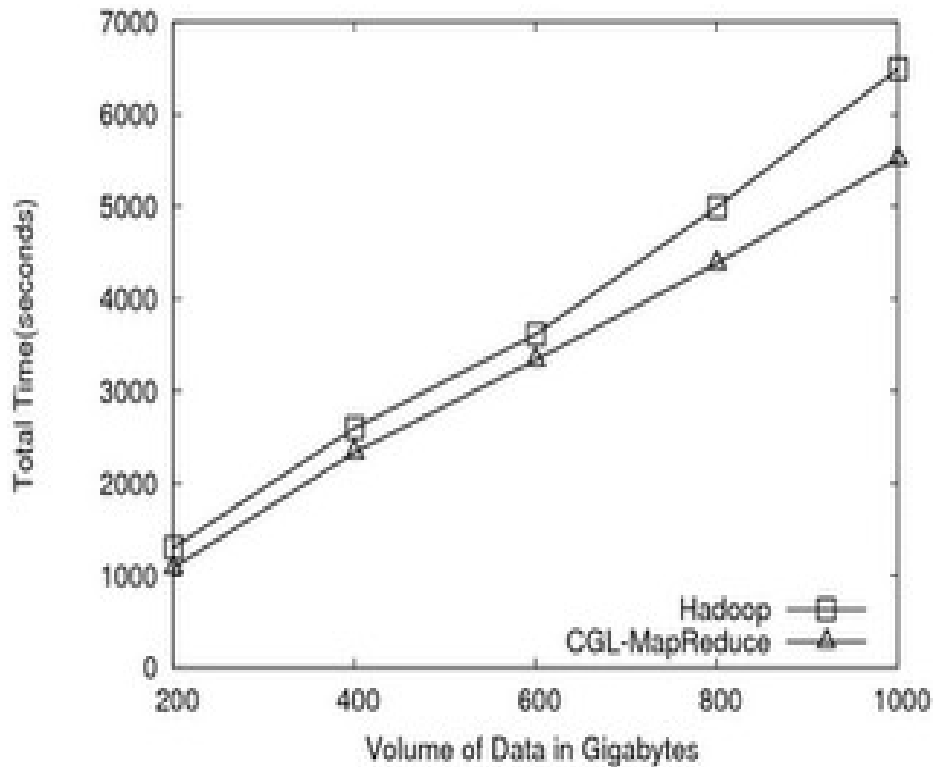


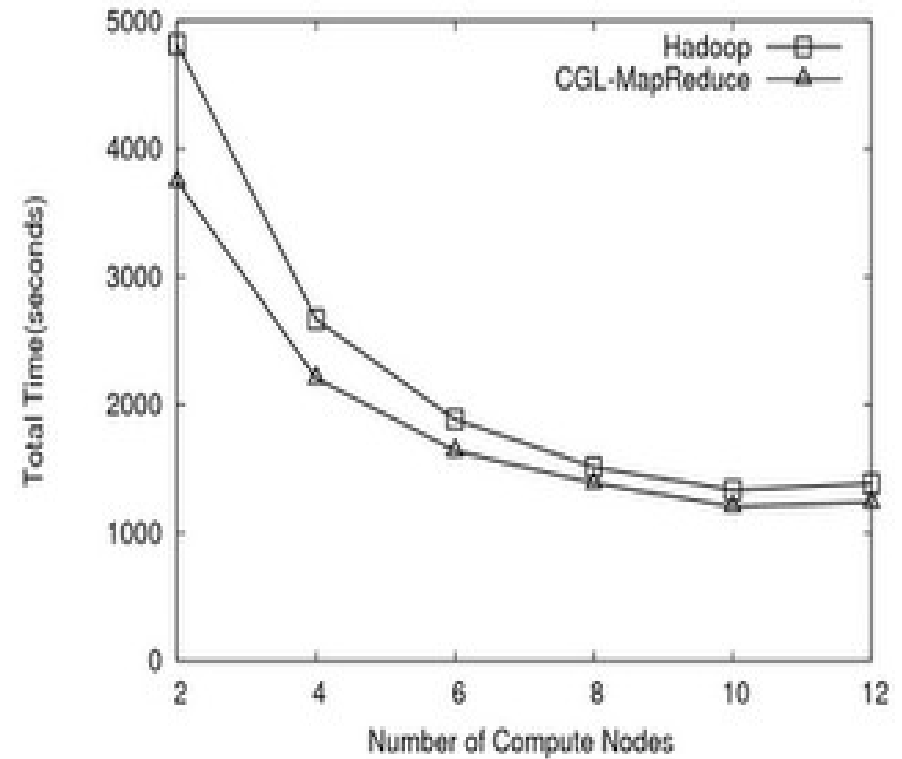
Figure 7. Speedup curves for CloudBLAST (Hadoop) and mpiBLAST: 960 short and long sequences against NR database segmented into 1 and 30 fragments on 1-site and 2-site resources. CloudBLAST presents a slightly better performance in particular when sequences are shorter.

Source: Andréa Matsunaga, CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications

High Energy Physics



HEP data analysis, execution time vs. the volume of data (fixed compute resources)



Total time vs. the number of compute nodes (fixed data)

Source: Jaliya Ekanayake

MapReduce Variation

DAG for the computation - Isard et al. [13] proposed to allow programmers to specify a DAG for the computation. The DAG specification is sophisticated and a manual DAG representing an SQL query on an astronomy database was illustrated

PIG Latin - Olston et al. [15]. SQL-style declarative language

Hive - SQL-Like Constructs + Hadoop Streaming

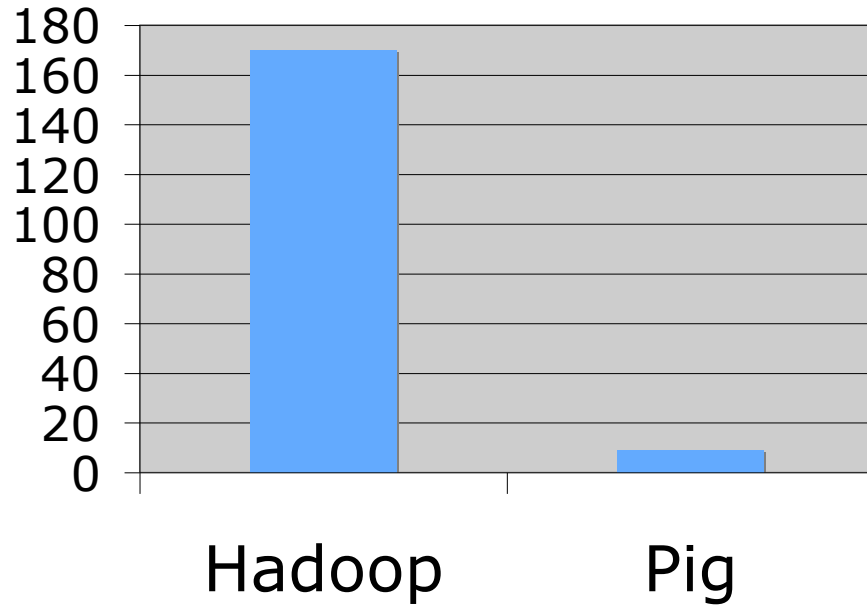
RHIPE- a java package that integrates the R environment with Hadoop

HiMach - a framework inspired from MapReduce for the molecular dynamics solutions. HiMach allows users to write trajectory analysis programs sequentially, and carries out the parallel execution of the programs automatically.

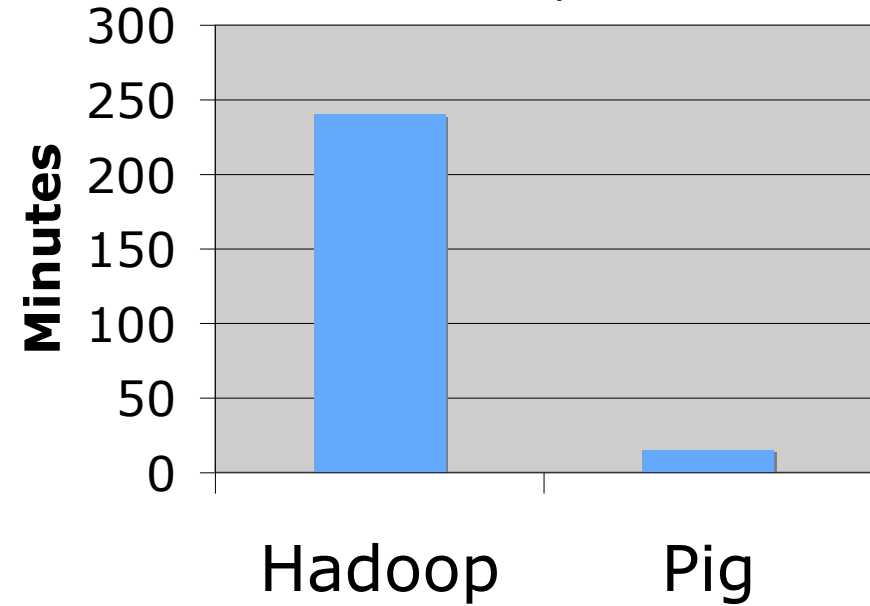
DryadLinQ

Comparison

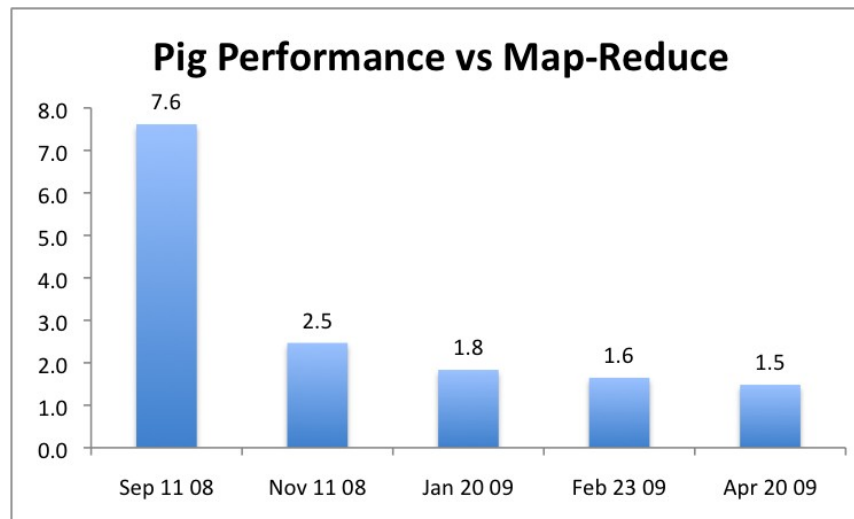
1/20 the lines of code



1/16 the development time

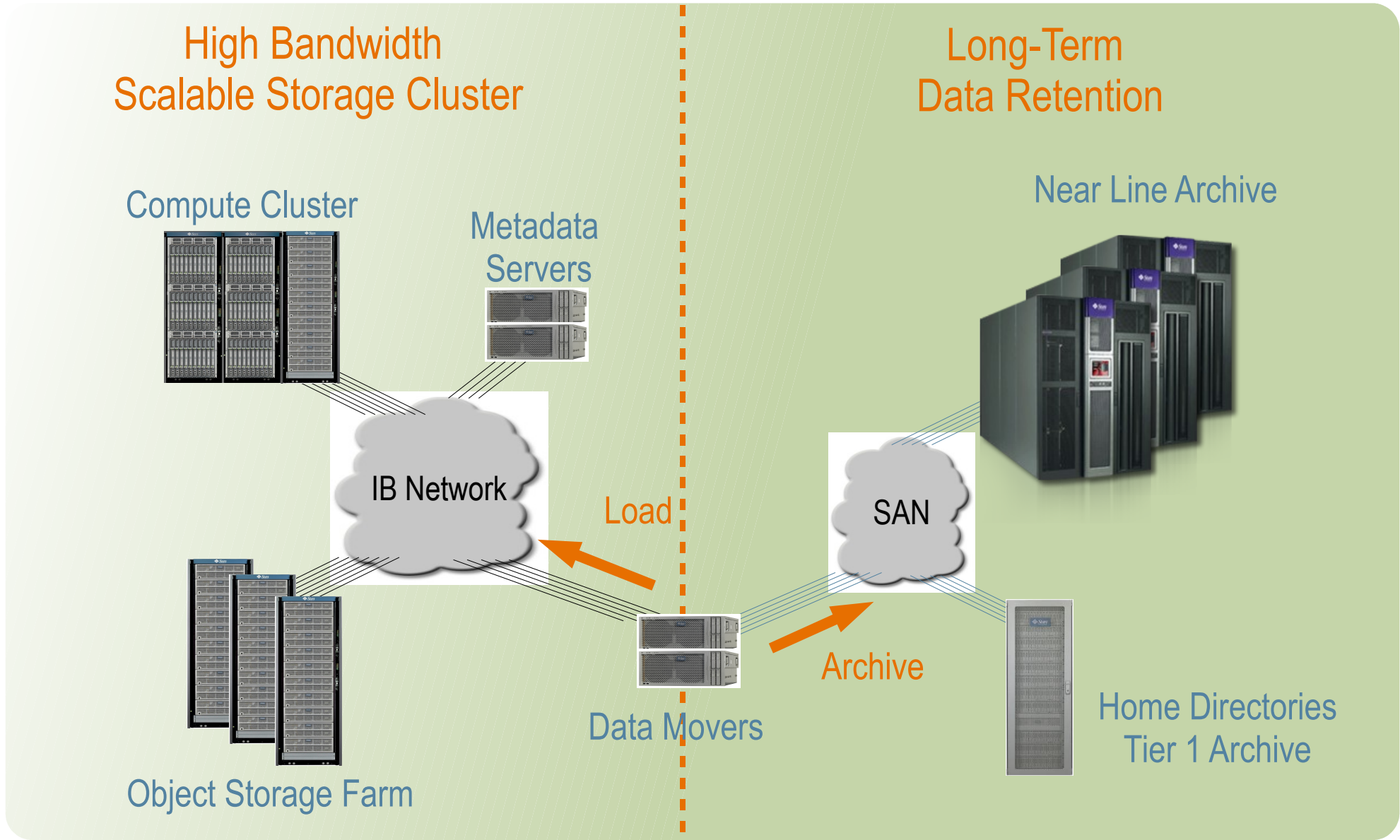


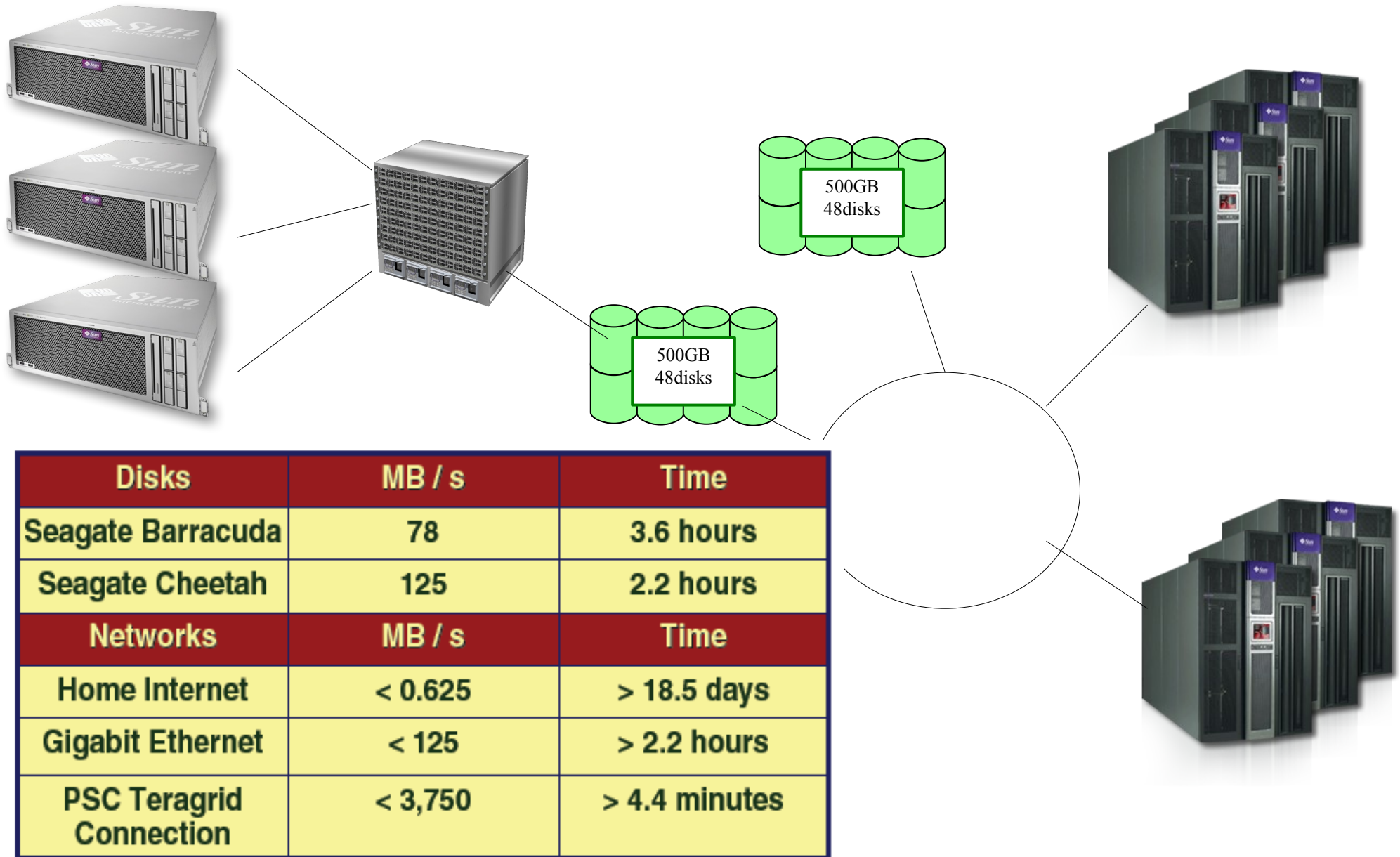
Performance:
1.5x Hadoop



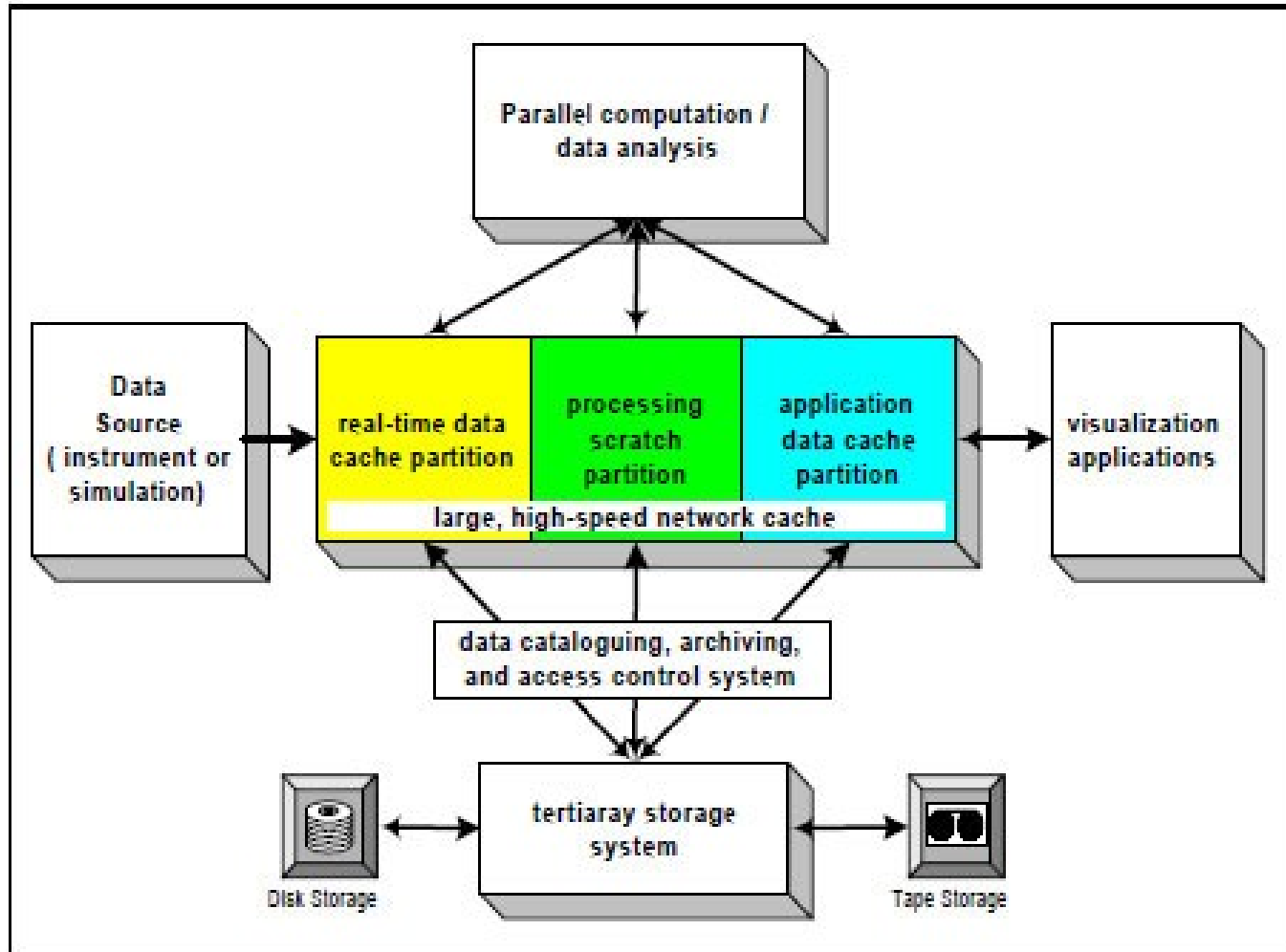
Architecture

Traditional Compute and Storage Solution



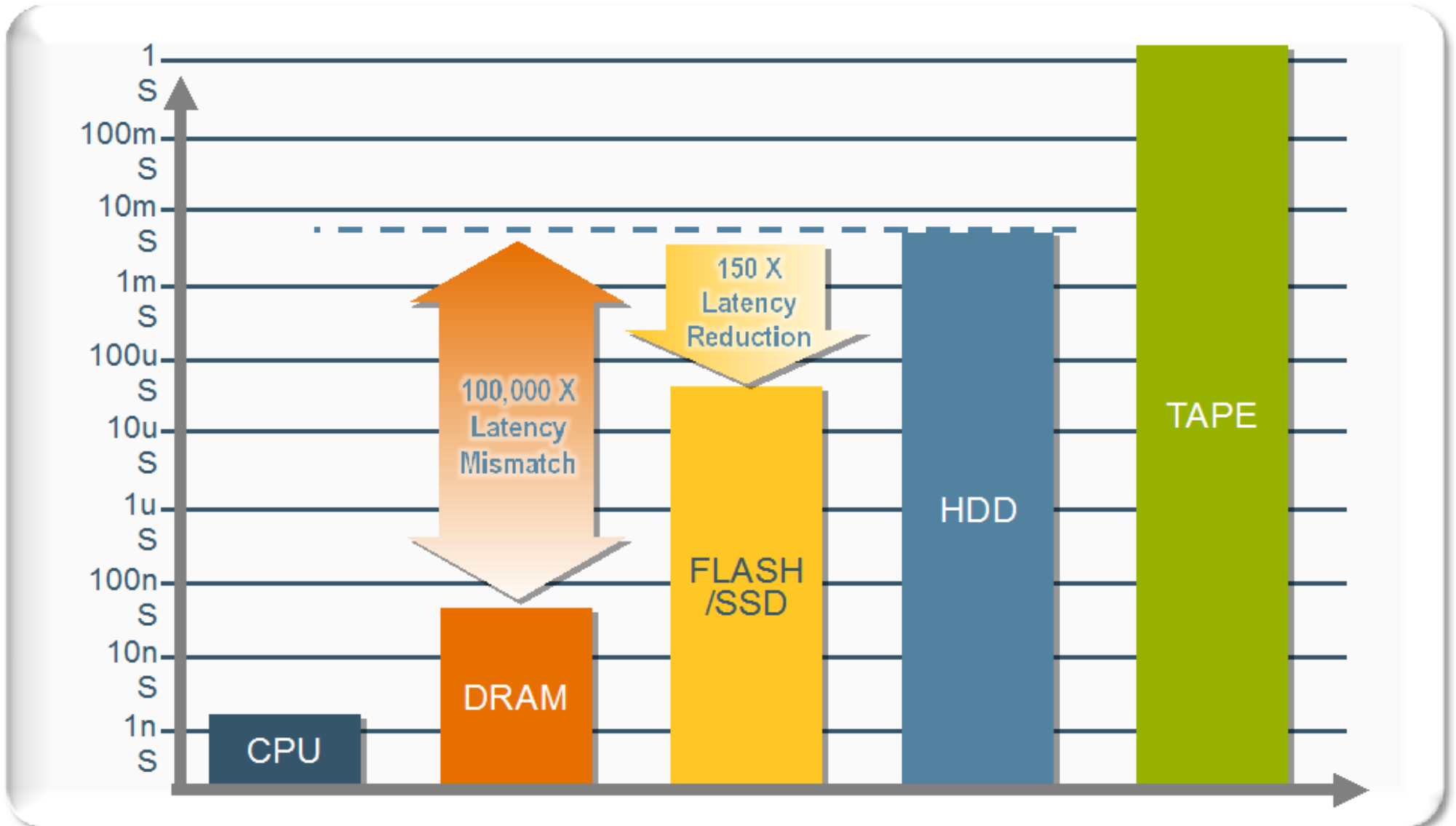


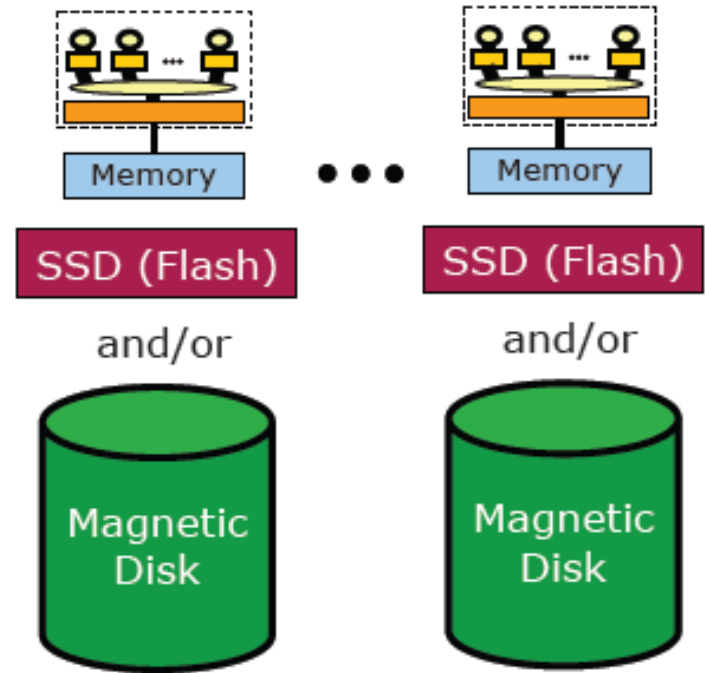
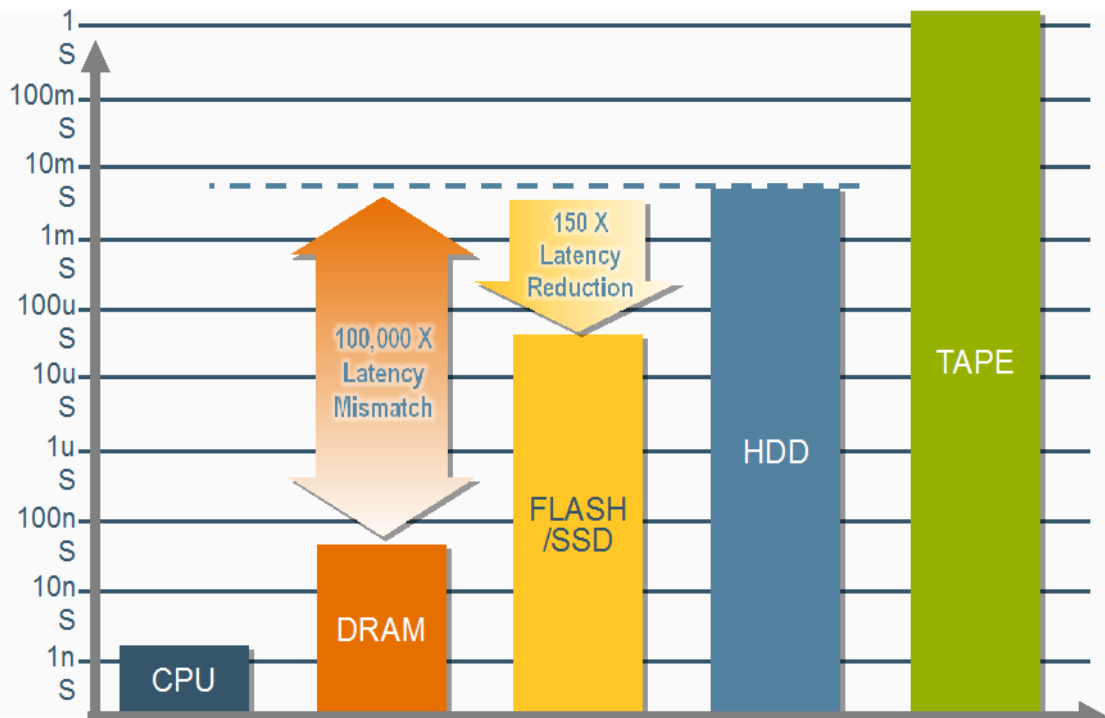
Disks	MB / s	Time
Seagate Barracuda	78	3.6 hours
Seagate Cheetah	125	2.2 hours
Networks	MB / s	Time
Home Internet	< 0.625	> 18.5 days
Gigabit Ethernet	< 125	> 2.2 hours
PSC Teragrid Connection	< 3,750	> 4.4 minutes



Latency Comparison

Bridging the DRAM to HDD Gap





Hierarchy-Savvy Parallel Algorithm Design (HI-SPADE) project

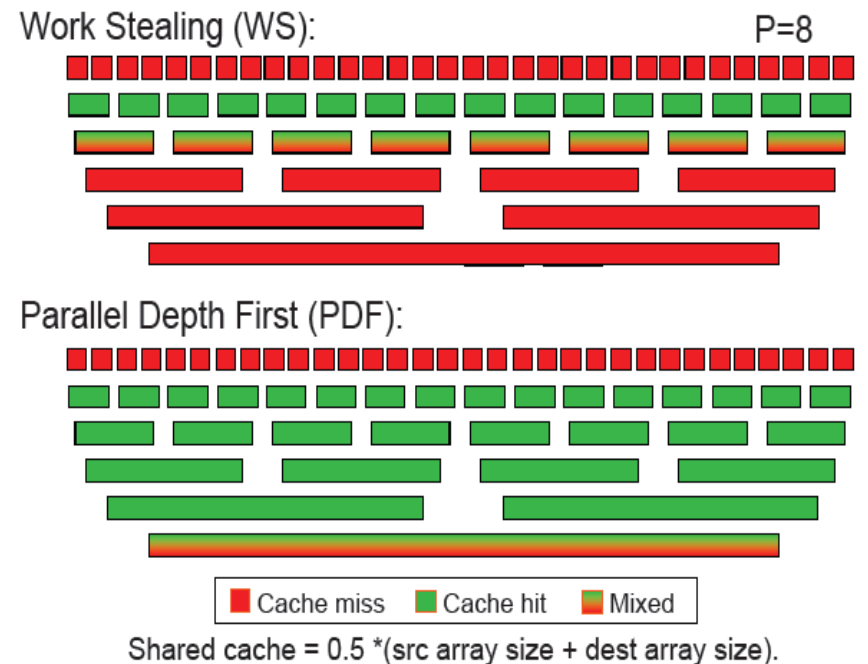
Goal: Support a hierarchy-savvy model of computation for parallel algorithm design

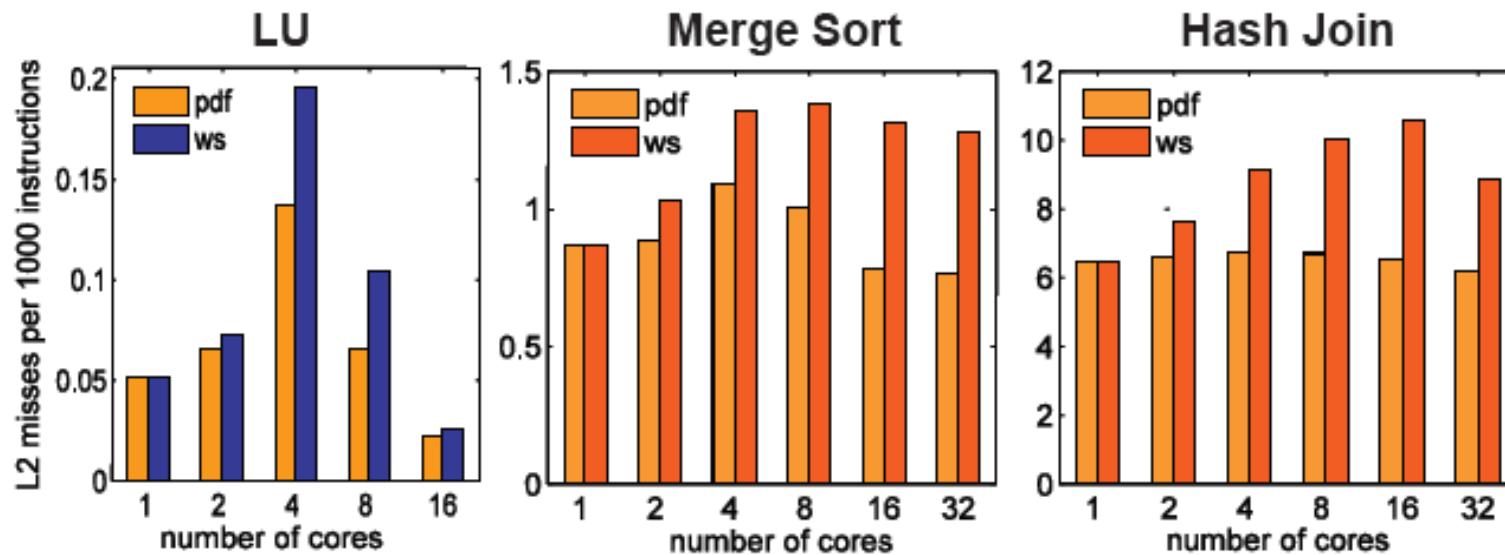
Effectively Sharing a Cache among Threads [Blelloch & Gibbons, SPAA'04]

- First thread scheduling policy (PDF) with provably-good shared cache performance for any parallel computation
- W.r.t. sequential cache performance

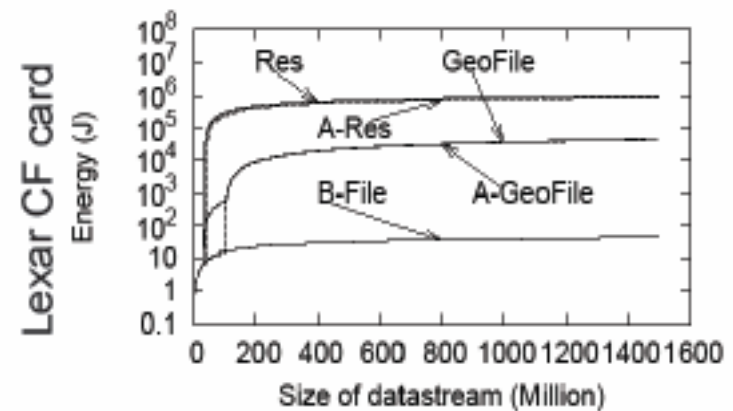
Scheduling Threads for Constructive Cache Sharing on CMPs [Chen et al, SPAA'07]

Example: Parallel Merging in Merge Sort





Work Stealing (ws) vs. Parallel Depth First (pdf); simulated CMPs



Summary

- Data Explosion at Exponential Rate or worse !!!
 - > Coming from new sensors, instruments and large simulation
 - > It has to be Archived !
- Some Experiment will add Petabytes or 10s of Petabytes per year
 - > Need scalable solutions (compute, network and storage)
 - > Locate analysis to the data
 - > Spatial and temporal features essential
- Need a New Paradigm
 - > Computational methods, algorithms
- Architecture Design or “Plumbing” has to be carefully taken into account.

A photograph showing an underwater view of a blue wave tunnel, with water swirling and creating a circular opening.

Thank you

Simon.See@Sun.com

Agenda

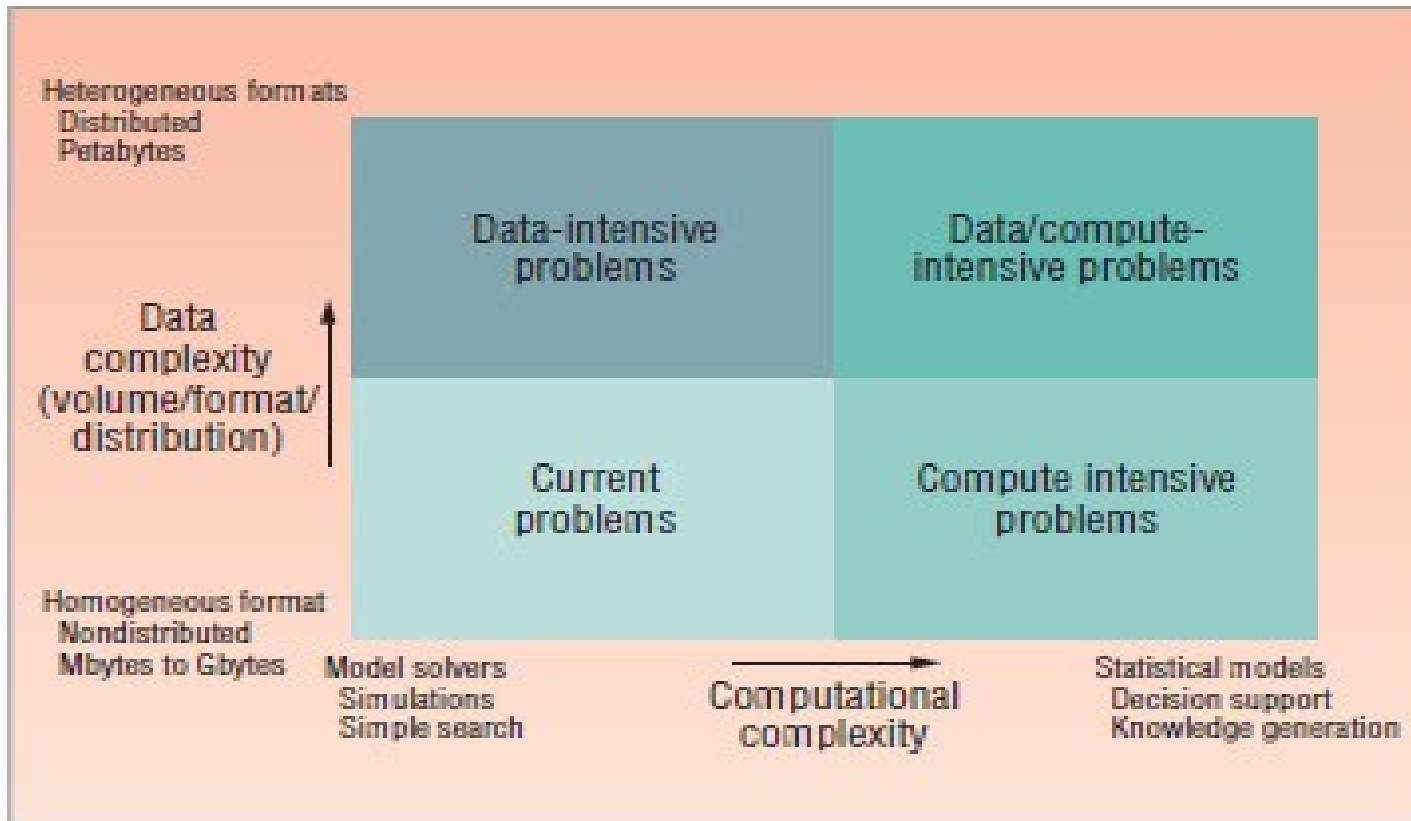
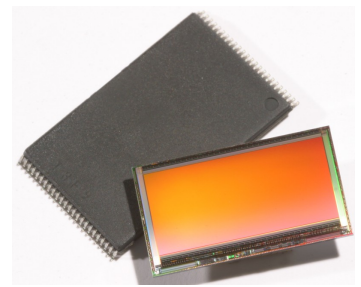


Figure 1. Research issues. Data-intensive computing research encompasses the problems in the upper two quadrants.

Storage Comparison - 147GB Server-Class



DRAM



Server SSD



HDD

Budgetary Cost	\$45/GB 37 DIMMS - \$6,615	\$20/GB 192GB SSD - \$3,840	\$3.30/GB 147GB HDD - \$485
Power Consumption	463 Watts	2.5 Watts	12 Watts
Random I/O	1,000,000 IOP/s	10,000-50,000 IOP/s	350 IOP/s

Data Intensive Computing

