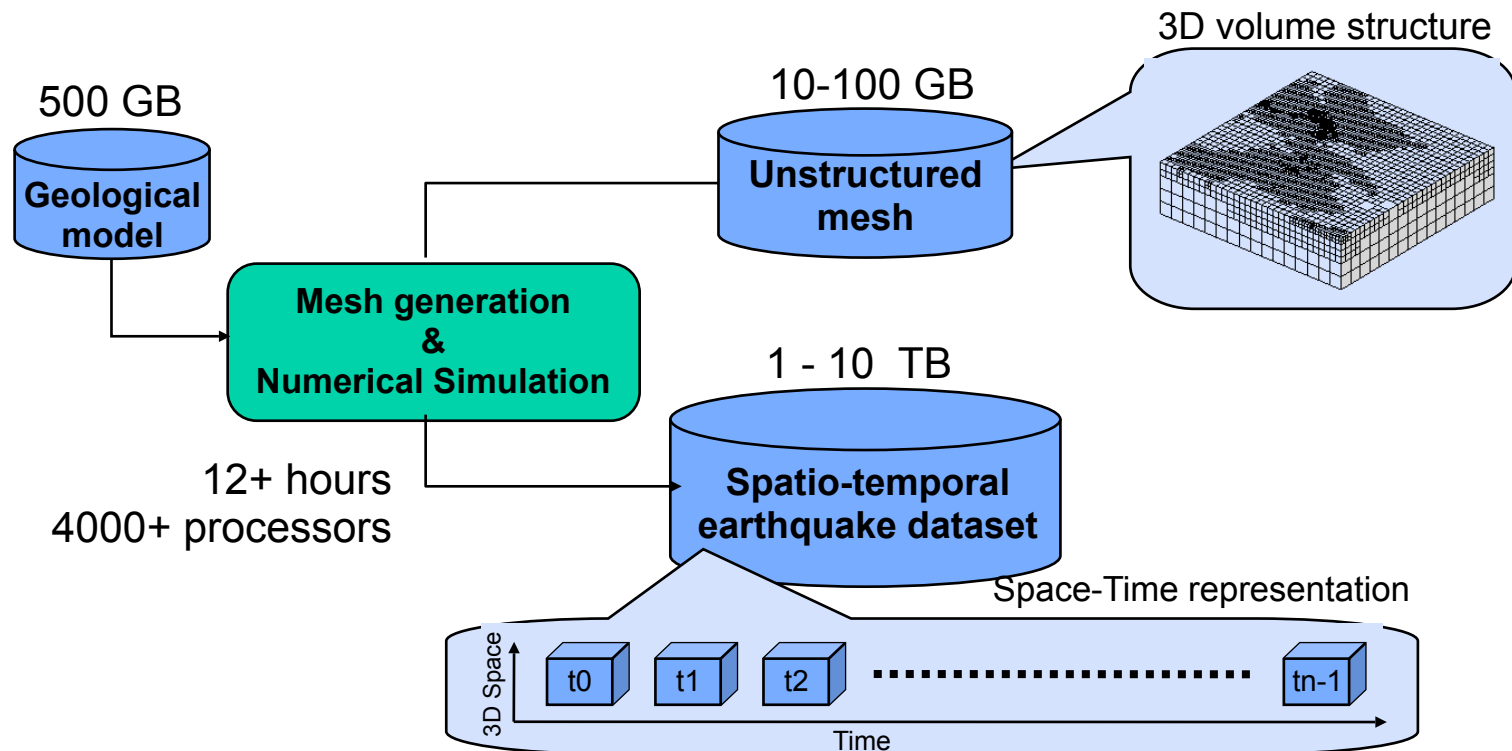

Data
Intensive
Scalable
Computing

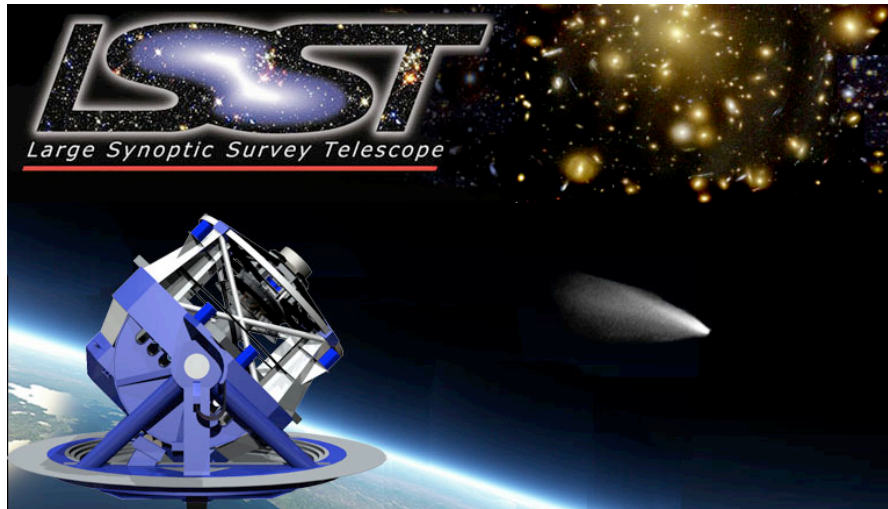
Thanks to: Randal E. Bryant
Carnegie Mellon University
<http://www.cs.cmu.edu/~bryant>

Big Data Sources: Seismic Simulations

- Wave propagation during an earthquake
- Large-scale parallel numerical simulations
- Quake 4D wavefields: $O(\text{TB})$



Big Data Sources: LSST



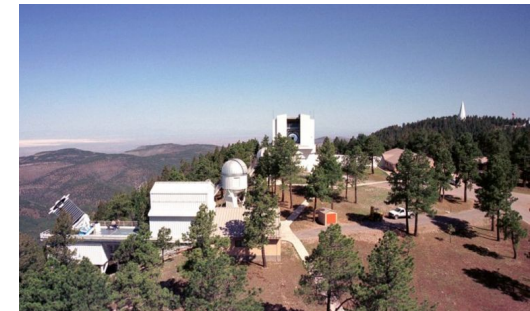
- LSST: Large Synoptic Survey Telescope (2016)
 - Pixel count: 3.2 Gpixels, Dynamic range: 16 bits
 - Readout time: 2 sec.
 - Nightly data generation rate: 15 TBytes (16 bits)
 - Avg. yearly data generation rate: 6.8 PBytes
 - Total disk storage: 200 Pbytes

Examples of Big Data Sources

- Wal-Mart



- 267 million items/day, sold at 6,000 stores
- HP building them 4PB data warehouse
- Mine data to manage supply chain, understand market trends, formulate pricing strategies



- 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

Our Data-Driven World


- **Science**
 - Data bases from astronomy, genomics, natural languages, seismic modeling, ...
- **Humanities**
 - Scanned books, historic documents, ...
- **Commerce**
 - Corporate sales, stock market transactions, census, airline traffic, ...
- **Entertainment**
 - Internet images, Hollywood movies, MP3 files, ...
- **Medicine**
 - MRI & CT scans, patient records, ...

Why So Much Data?

- We Can Get It
 - Automation + Internet
- We Can Keep It
 - 1 TB Disk @ \$199 (20¢ / GB)
- We Can Use It
 - Scientific breakthroughs
 - Business process efficiencies
 - Realistic special effects
 - Better health care
- Could We Do More?
 - Apply more computing power to this data



Oceans of Data, Skinny Pipes



No more blaming connection speeds for your losses.

Plans as low **\$39.99/month** (up to 5 Mbps).
Plus, order online & **get your first month FREE!**



- 1 Terabyte
 - Easy to store
 - Hard to move

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

Data-Intensive System Challenge

- For computation that accesses 1 TB in 5 mins
 - Data distributed over 100+ disks
 - Assuming uniform data partitioning
 - Compute using 100+ processors
 - Connected by at least gigabit Ethernet

- System Requirements
 - Lots of disks
 - Lots of processors
 - Located in close proximity
 - Within reach of fast, local-area network

Google's Computing Infrastructure

- System

- ~ 3 million processors in clusters of ~2000 processors each
- Commodity parts
 - x86 processors, IDE disks, Ethernet communications
 - Gain reliability through redundancy & software management
- Partitioned workload
 - Data: Web pages, indices distributed across processors
 - Function: crawling, index generation, index search, document retrieval, Ad placement

Barroso, Dean, Hölzle, "Web Search for a Planet: The Google Cluster Architecture" IEEE Micro 2003

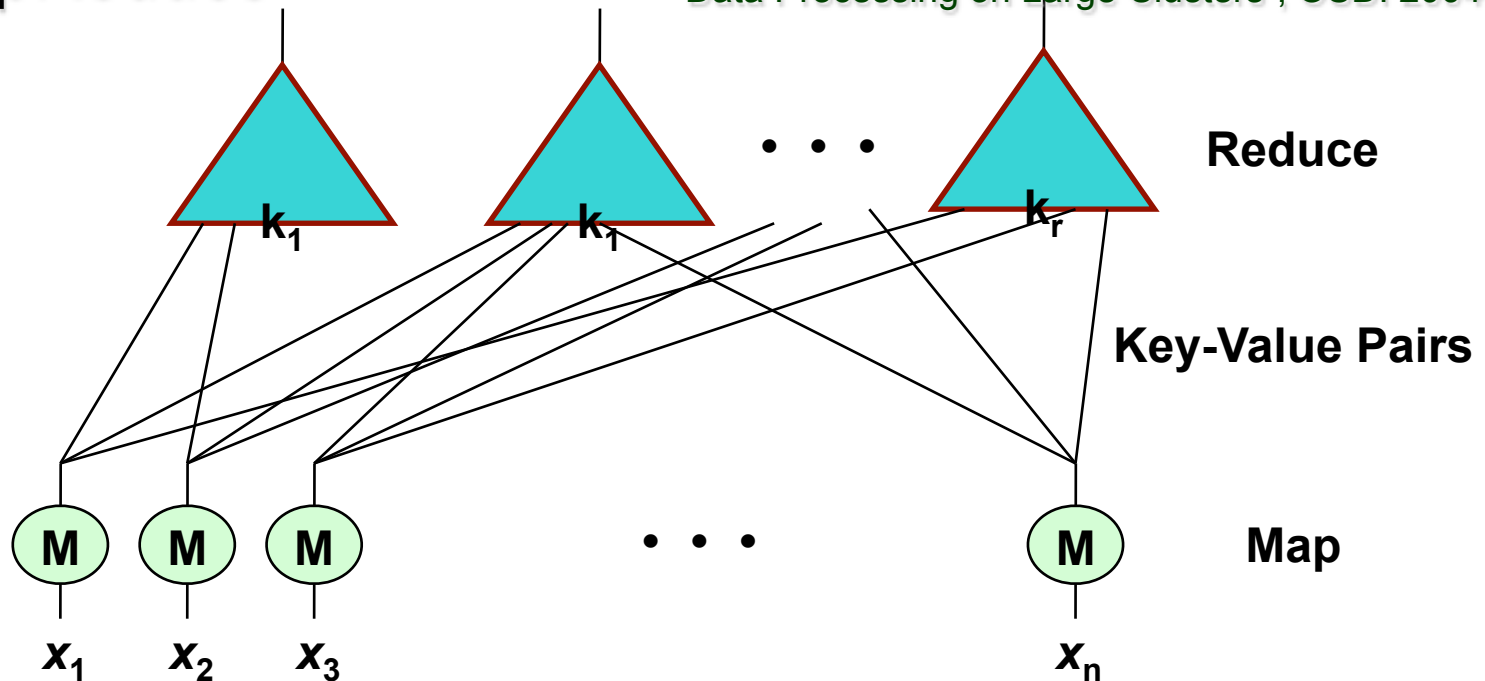
- A Data-Intensive Scalable Computer (DISC)

- Large-scale computer centered around data
 - Collecting, maintaining, indexing, computing
- Similar systems at Microsoft & Yahoo

MapReduce Programming Model

- MapReduce

Dean & Ghemawat: "MapReduce: Simplified Data Processing on Large Clusters", OSDI 2004



- 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

Comparing Parallel Computation Models

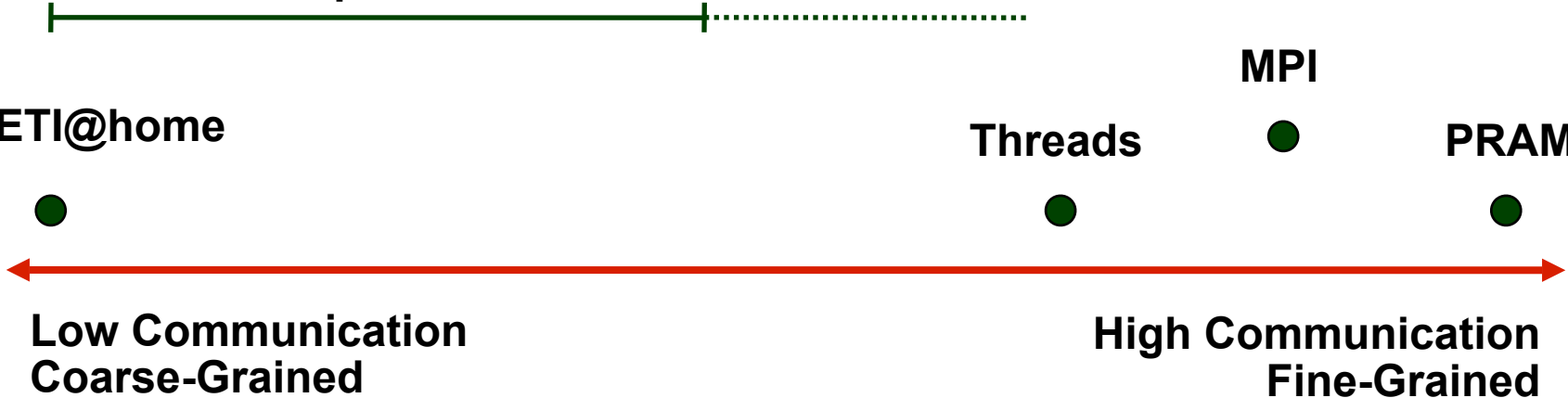
MapReduce

MPI

SETI@home

Threads

PRAM



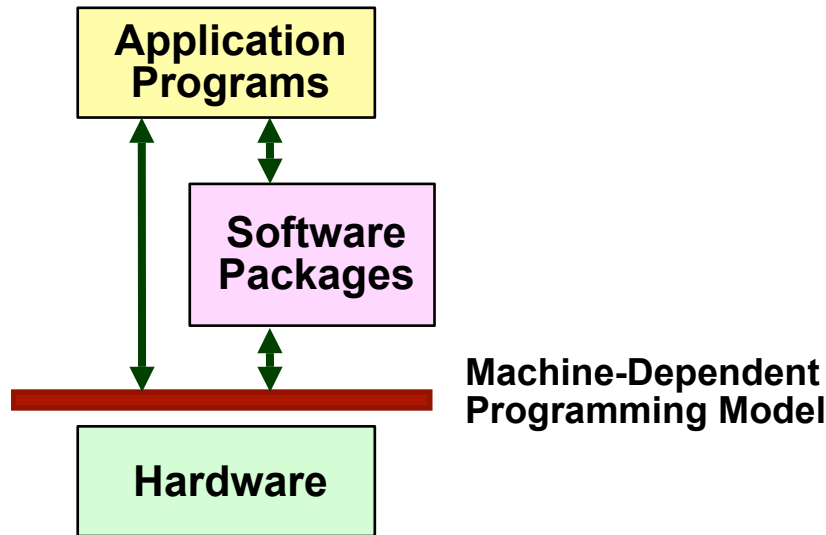
- DISC + MapReduce Provides Coarse-Grained Parallelism
 - Computation done by independent processes
 - File-based communication
- Observations
 - Relatively “natural” programming model
 - Research issue to explore full potential and limits
 - Dryad project at MSR
 - Pig project at Yahoo!

Desiderata for DISC Systems

- Focus on Data
 - Terabytes, not tera-FLOPS
- Problem-Centric Programming
 - Platform-independent expression of data parallelism
- Interactive Access
 - From simple queries to massive computations
- Robust Fault Tolerance
 - Component failures are handled as routine events

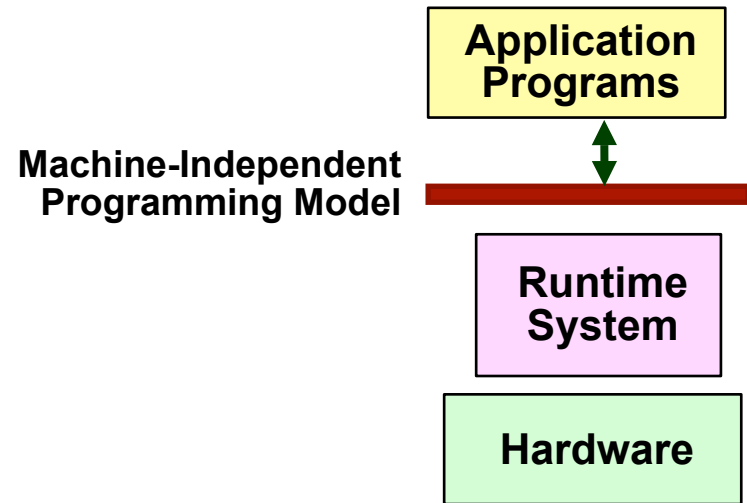
System Comparison: Programming Models

Conventional Supercomputers



- Programs described at very low level
 - Specify detailed control of processing & communications
- Rely on small number of software packages
 - Written by specialists
 - Limits classes of problems & solution methods

DISC



- Application programs written in terms of high-level operations on data
- Runtime system controls scheduling, load balancing, ...

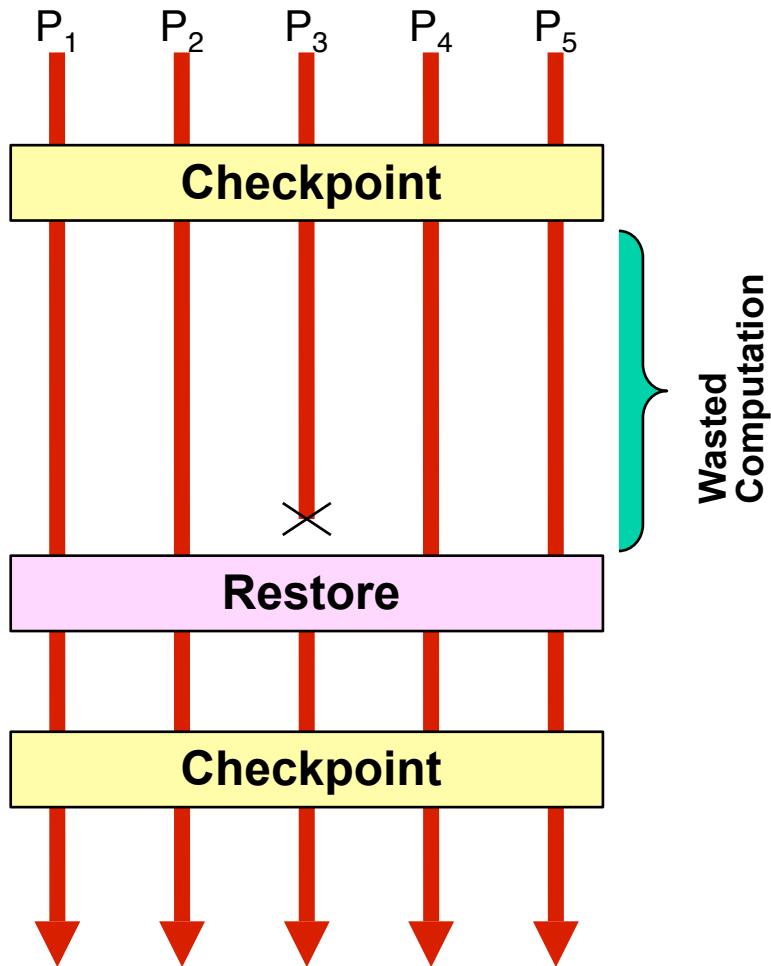
Compare to Transaction Processing

- **Main Commercial Use of Large-Scale Computing**
 - Banking, finance, retail transactions, airline reservations ...
- **Stringent Functional Requirements**
 - Only one person gets last \$1 from shared bank account
 - Must not lose money when transferring between accounts
 - Small number of high-performance, high-reliability servers
- **Our Needs are Different**
 - More relaxed consistency requirements
 - Fewer sources of updates, write-once, read-many data.
 - Individual computations access more data

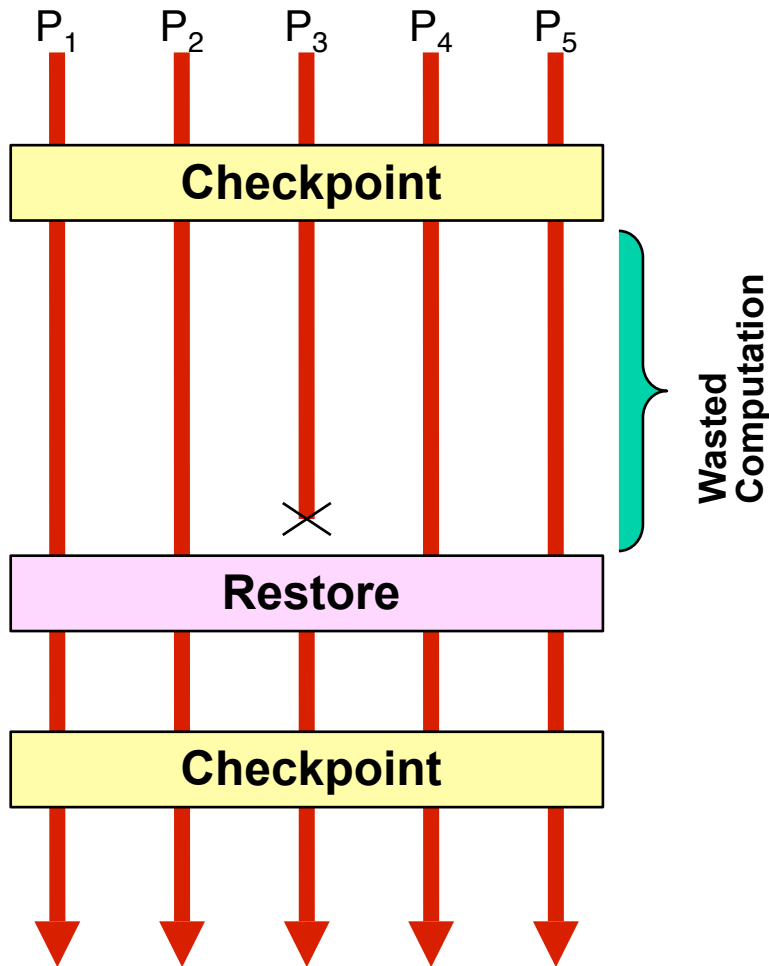
CS Research Issues

- Applications
 - Language translation, image processing, ...
- Application Support
 - Machine learning over very large data sets
 - Web crawling
- Programming
 - Abstract programming models to support large-scale computation
 - Distributed databases
- System Design
 - Error detection & recovery mechanisms
 - Resource scheduling and load balancing
 - Distribution and sharing of data across system

HPC Fault Tolerance



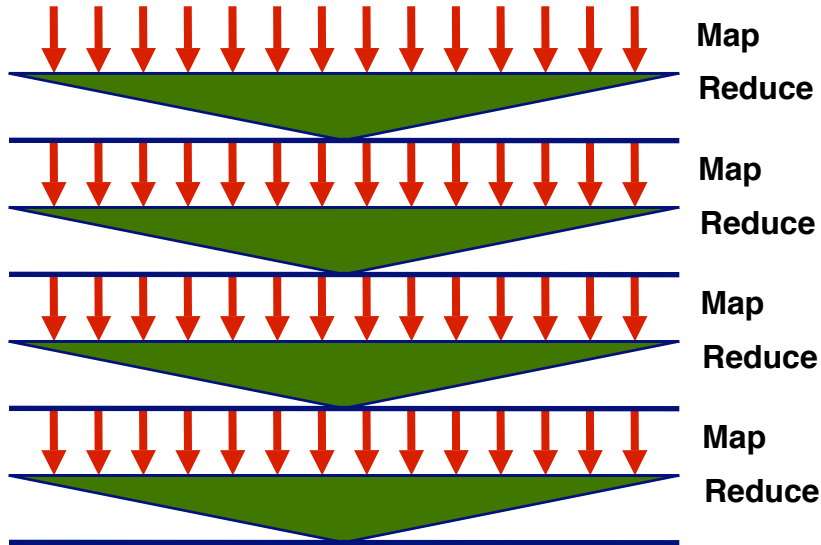
HPC Fault Tolerance



- **Checkpoint**
 - Periodically store state of all processes
 - Significant I/O traffic
- **Restore**
 - When failure occurs
 - Reset state to that of last checkpoint
 - All intervening computation wasted
- **Performance Scaling**
 - Very sensitive to number of failing components

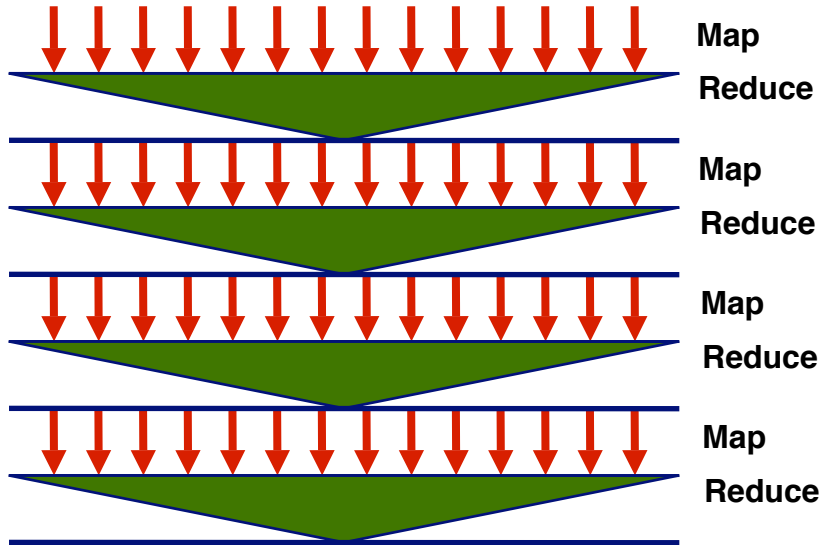
Map/Reduce Operation

Map/Reduce



Map/Reduce Operation

Map/Reduce



• Characteristics

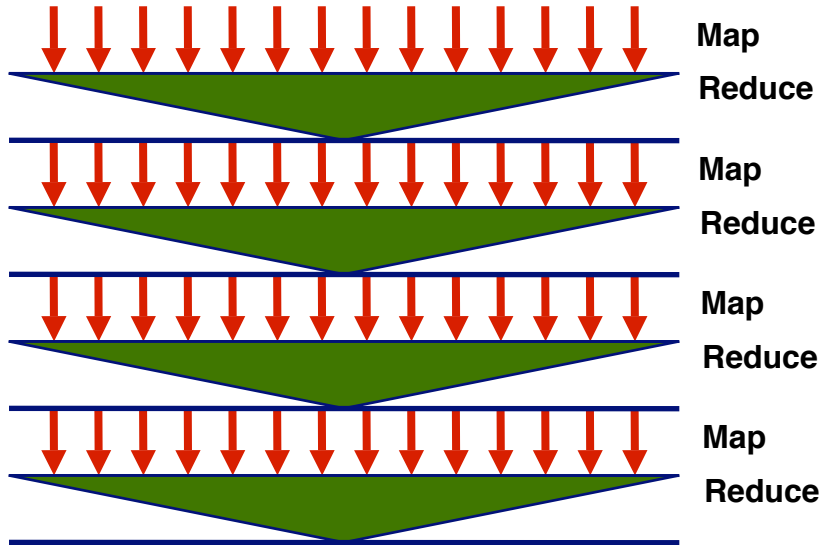
- Computation broken into many, short-lived tasks
Mapping, reducing
- Use disk storage to hold intermediate results

• Strengths

- Great flexibility in placement, scheduling, and load balancing
- Handle failures by recomputation
- Can access large data sets

Map/Reduce Operation

Map/Reduce



• Characteristics

- Computation broken into many, short-lived tasks
Mapping, reducing
- Use disk storage to hold intermediate results

• Strengths

- Great flexibility in placement, scheduling, and load balancing
- Handle failures by recomputation
- Can access large data sets

• Weaknesses

- Higher overhead
- Lower raw performance

Choosing Execution Models

- **Message Passing / Shared Memory**
 - Achieves high performance when everything works well
 - Requires careful tuning of programs
 - Vulnerable to single points of failure
- **Map/Reduce**
 - Allows for abstract programming model
 - More flexible, adaptable, and robust
 - Performance limited by disk I/O
- **Alternatives?**
 - Is there some way to combine to get strengths of both?

Getting Started

- Goal: Get faculty & students active in DISC
- Hardware: Rent from Amazon
 - Elastic Compute Cloud (EC2)
 - Generic Linux cycles for \$0.10 / hour (\$877 / yr)
 - Simple Storage Service (S3)
 - Network-accessible storage for \$0.15 / GB / month (\$1800/TB/yr)
 - Example: maintain crawled copy of web
50 TB, 100 processors, 0.5 TB/day refresh ~ \$250K / year



- Software

- Hadoop Project
 - Open source project providing file system and MapReduce
 - Supported and used by Yahoo
 - Prototype on single machine, map onto cluster



Rely on Kindness of Others

Press Release 08-031

NSF Partners With Google and IBM to Enhance Academic Research Opportunities

Computer science researchers at universities and colleges will be able to utilize large-scale computing cluster

February 25, 2008

Today the National Science Foundation's Computer and Information Science and Engineering (CISE) Directorate announced the creation of a strategic relationship with Google Inc. and IBM. The Cluster Exploratory (CluE) relationship will enable the

- Google setting up dedicated cluster for university use
- Loaded with open-source software
 - Including Hadoop
- IBM providing additional software support
- NSF will determine how facility should be used.

More Sources of Kindness

Yahoo, Carnegie Mellon Switch On Supercomputer



Submitted by [David A. Utter](#) on Mon, 11/12/2007 - 11:08.

[Comment](#) | [Email](#) | [Print](#)

The M45 supercomputer provided by Yahoo opened its ports to its partners at Carnegie Mellon University, where the initiative should help boost research that benefits the broader Internet community.



For those of you firing up the old faithful laptop for a morning of surfing, blogging, maybe a little development work, get a load of what some of the lucky geeks at [Carnegie Mellon University](#) got to play with this morning:

The M45, Yahoo's supercomputing cluster, has approximately 4,000 processors, three terabytes of memory, 1.5 petabytes of disks, and a peak performance of more than 27 trillion calculations per second (27 petaflops), placing it among the top 50 fastest

- Yahoo! is a major supporter of Hadoop
- Yahoo! plans to work with other universities


Beyond the U.S.

March 24 2008

Yahoo, Tata Subsidiary In Research Pact

Duncan Riley

[9 comments >>](#)

Yahoo **has announced**  an agreement with Computational Research Laboratories (CRL, a wholly owned subsidiary of Indian conglomerate Tata) to jointly undertake cloud computing research.



Under the deal, CRL will give access to one of world's top five supercomputers "that has substantially more processors than any supercomputer currently available for cloud computing research."

Testbed for system research in DISC systems

HP, Yahoo and Intel Create Compute Cloud

[Stacey Higginbotham](#), Tuesday, July 29, 2008 at 10:37 AM PT

[Comments \(8\)](#)

Related Stories

[HP Weds Cloud and High-performance Computing](#)

[Intel Friends Facebook to Make x86 Chips Sexy](#)

[Elastra Gets \\$12M — Is It Amazon's Enterprise Play?](#)

Powered by [Sphere](#)

Updated at the bottom: At long last, Hewlett-Packard is stepping up with an answer to cloud computing by inking a partnership with two other big technology vendors and three universities to create a cloud computing testbed. Through its R&D unit, HP Labs, the computing giant had teamed up with Intel, Yahoo, the Infocomm Development Authority of Singapore (IDA), the University of Illinois at Urbana-Champaign, the National Science Foundation (NSF) and the Karlsruhe Institute of Technology in Germany.

More Information

“Data-Intensive Supercomputing: The case for DISC”

Tech Report: CMU-CS-07-128

Available from <http://www.cs.cmu.edu/~bryant>