

Data **Intensive** **Super** **Computing**

Randal E. Bryant
Carnegie Mellon University

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

Data

Intensive

**~~Super~~ Scalable
Computing**

**Randal E. Bryant
Carnegie Mellon University**

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

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

- Seagate 1 TB Barracuda @ \$270 (27¢ / 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.

Verizon FiOS – the fastest
Internet available.

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

Enter your home phone number below to check availability.

Don't have a Verizon phone number? [Qualify your address.](#)



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 minutes

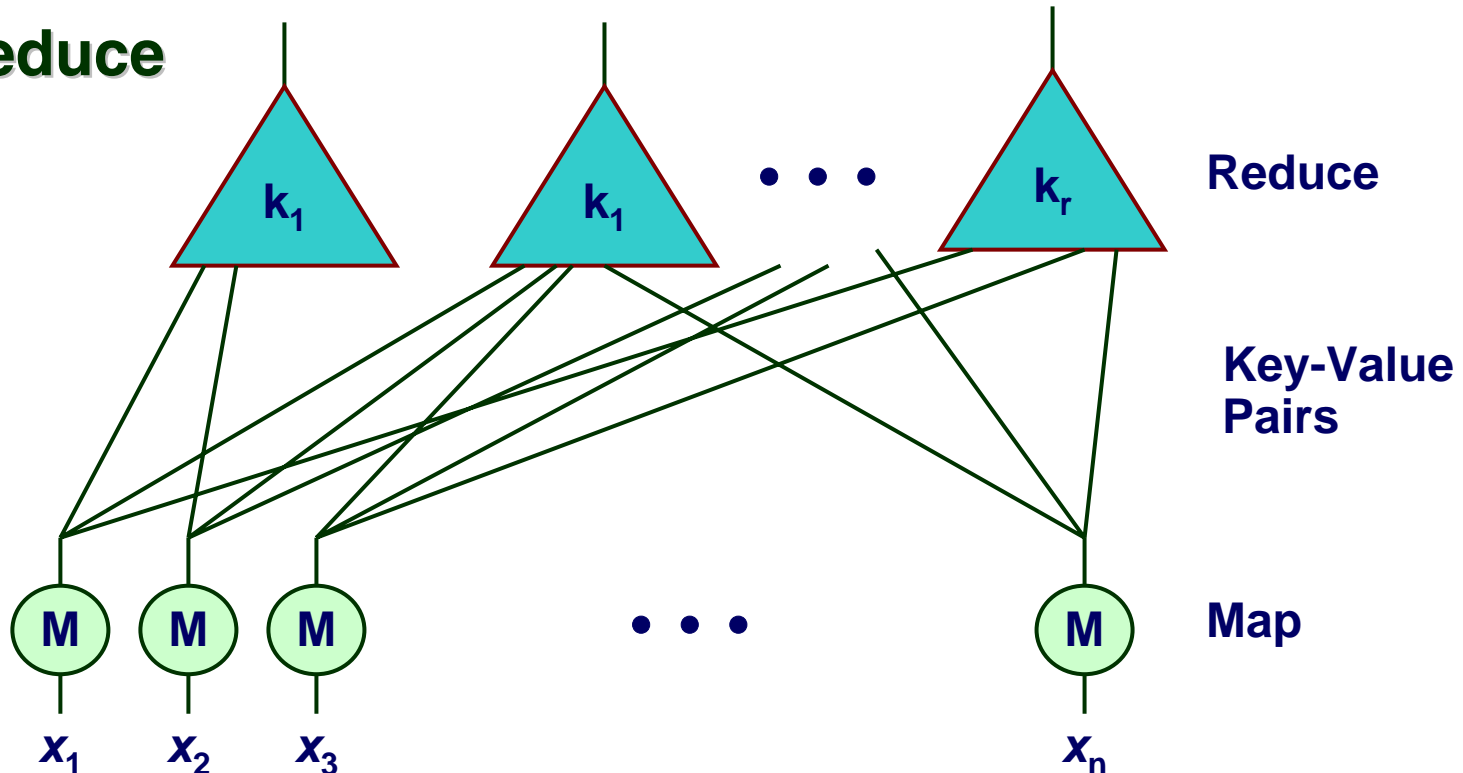
- **Data distributed over 100+ disks**
 - Assuming uniform data partitioning
- **Compute using 100+ processors**
- **Connected by gigabit Ethernet (or equivalent)**

System Requirements

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

MapReduce Programming Model

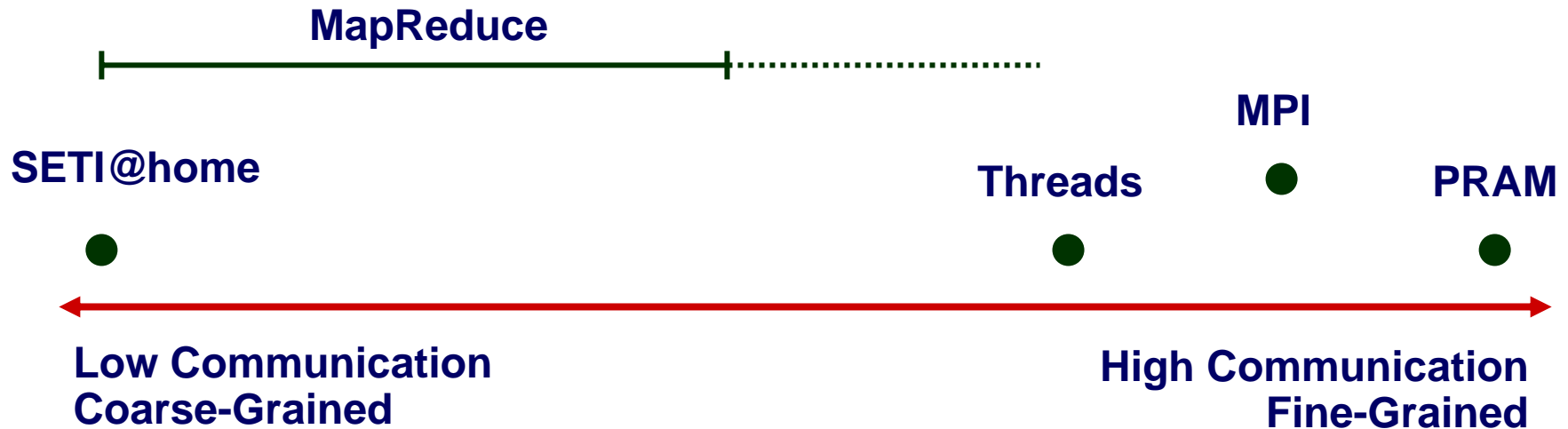
MapReduce



- 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

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

Comparing Parallel Computation Models



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

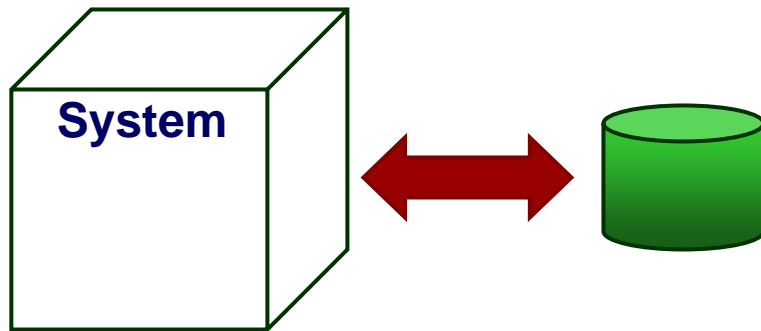
Robust Fault Tolerance

- Component failures are handled as routine events

Contrast to existing supercomputer / HPC systems

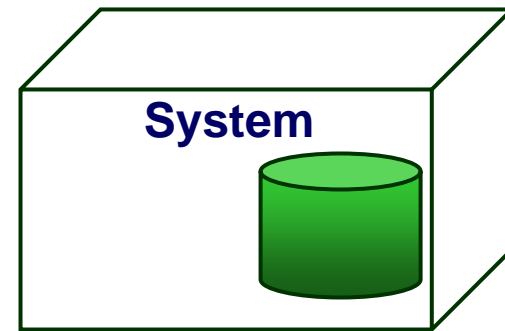
System Comparison: Data

Conventional Supercomputers



- **Data stored in separate repository**
 - No support for collection or management
- **Brought into system for computation**
 - Time consuming
 - Limits interactivity

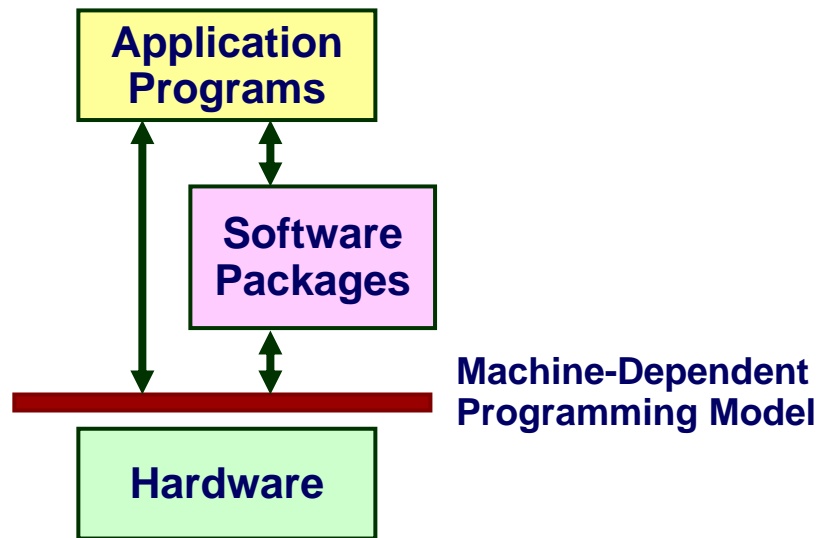
DISC



- **System collects and maintains data**
 - Shared, active data set
- **Computation colocated with storage**
 - Faster access

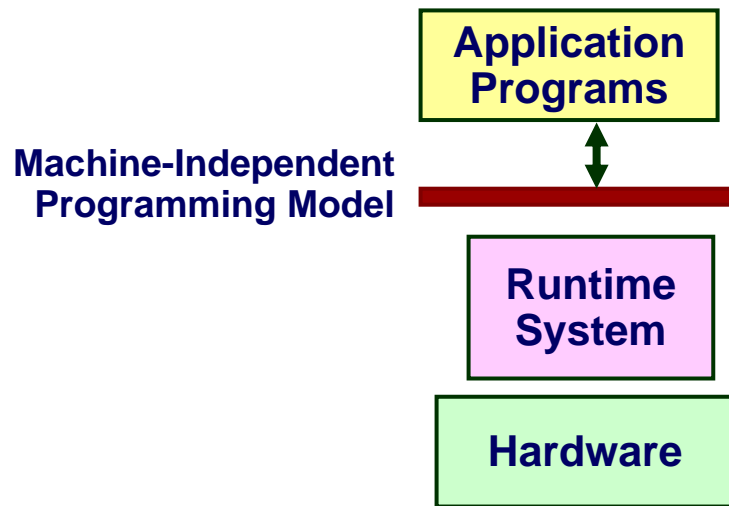
System Comparison: Programming Models

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, ...**

System Comparison: Interaction

Conventional Supercomputers

Main Machine: Batch Access

- Priority is to conserve machine resources
- User submits job with specific resource requirements
- Run in batch mode when resources available

Offline Visualization

- Move results to separate facility for interactive use

DISC

Interactive Access

- Priority is to conserve human resources
- User action can range from simple query to complex computation
- System supports many simultaneous users
 - Requires flexible programming and runtime environment

System Comparison: Reliability

Runtime errors commonplace in large-scale systems

- Hardware failures
- Transient errors
- Software bugs

Conventional Supercomputers

“Brittle” Systems

- Main recovery mechanism is to recompute from most recent checkpoint
- Must bring down system for diagnosis, repair, or upgrades

DISC

Flexible Error Detection and Recovery

- Runtime system detects and diagnoses errors
- Selective use of redundancy and dynamic recomputation
- Replace or upgrade components while system running
- Requires flexible programming model & runtime environment

What About Grid Computing?

- “Grid” means different things to different people

Computing Grid

- **Distribute problem across many machines**
 - Geographically & organizationally distributed
- **Hard to provide sufficient bandwidth for data exchange**

Data Grid

- **Shared data repositories**
- **Should colocate DISC systems with repositories**
 - It's easier to move programs than data

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
 - Beware of replicated data
- Must not lose money when transferring between accounts
 - Beware of distributed data
- Favors systems with small number of high-performance, high-reliability servers

Our Needs are Different

- More relaxed consistency requirements
 - Web search is extreme example
- Fewer sources of updates
- 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

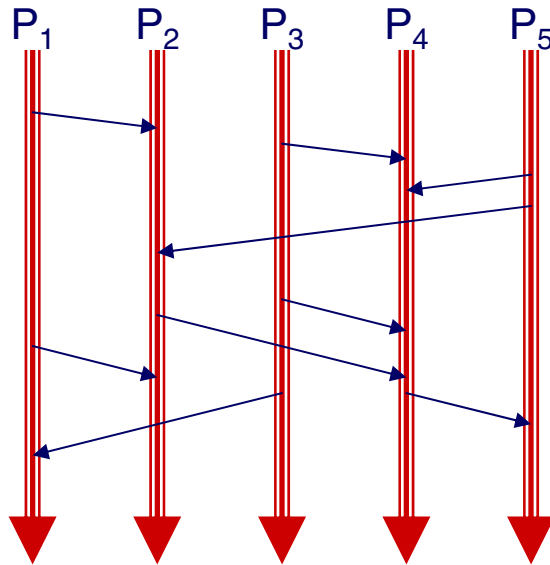
- Abstract programming models to support large-scale computation
- Distributed databases

System Design

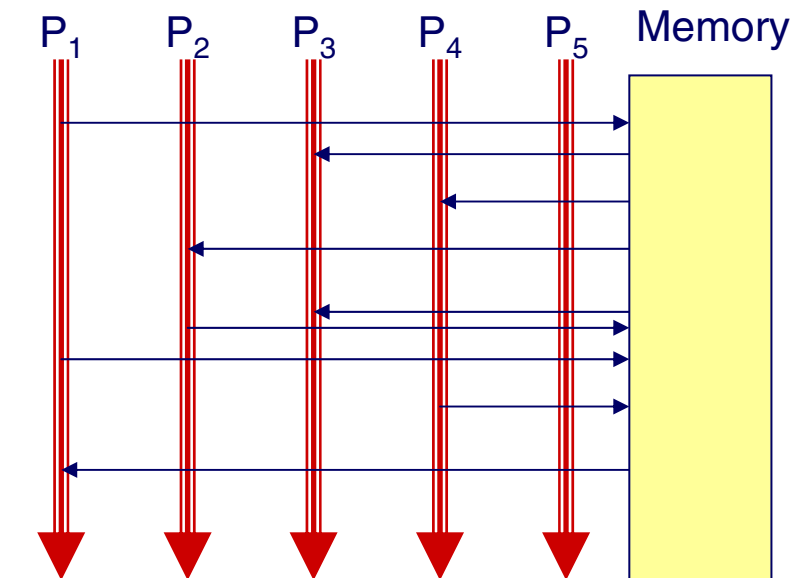
- Error detection & recovery mechanisms
- Resource scheduling and load balancing
- Distribution and sharing of data across system

Existing HPC Machines

Message Passing



Shared Memory



Characteristics

- Long-lived processes
- Make use of spatial locality
- Hold all program data in memory
- High bandwidth communication

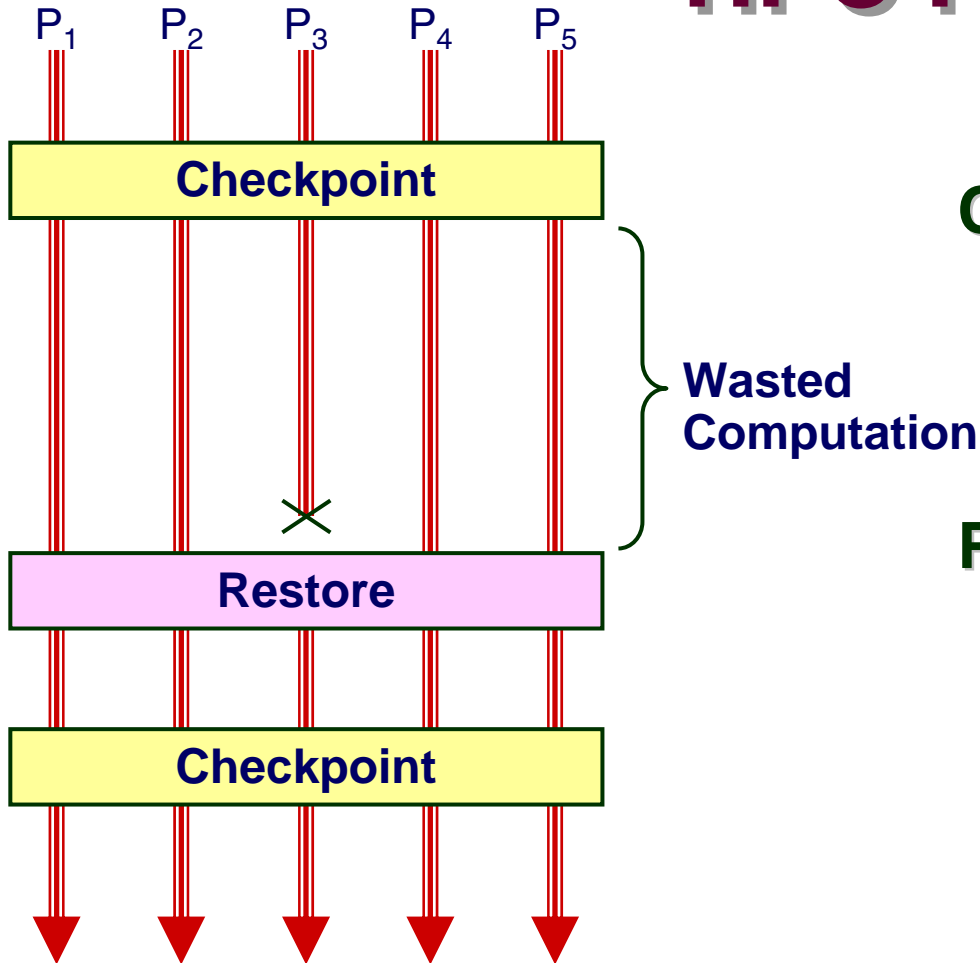
Strengths

- High utilization of resources
- Effective for many scientific applications

Weaknesses

- Very brittle: relies on everything working correctly and in close synchrony

HPC Fault Tolerance



Checkpoint

- Periodically store state of all processes
- Significant I/O traffic

Restore

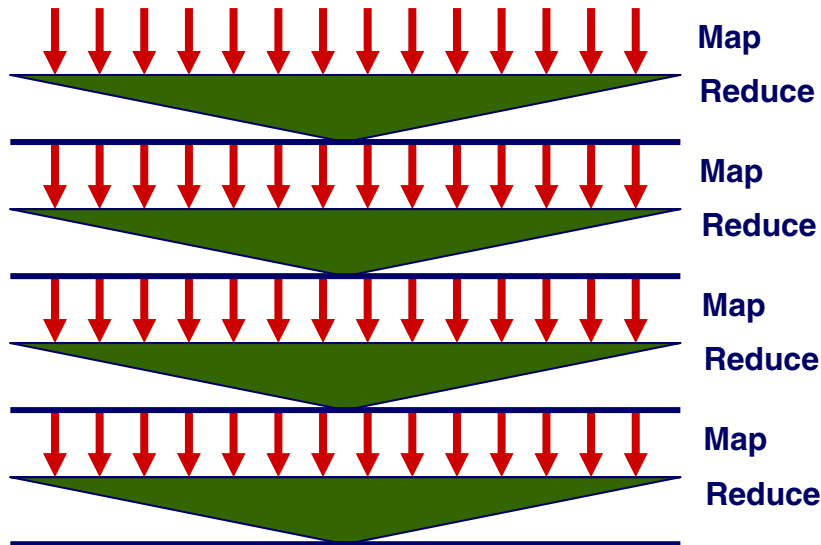
- When failure occurs
- Reset state to that of last checkpoint
- All intervening computation wasted

Performance Scaling

- Very sensitive to number of failing components

Map/Reduce Operation

Map/Reduce



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 very 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 academic research community to conduct experiments and test new theories and ideas using a large-scale, massively distributed computing cluster.

- **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 teraflops), placing it among the top 50 fastest supercomputers in the world.

- **Yahoo: 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."

Concluding Thoughts

The World is Ready for a New Approach to Large-Scale Computing

- Optimized for data-driven applications
- Technology favoring centralized facilities
 - Storage capacity & computer power growing faster than network bandwidth

Industry is Catching on Quickly

- Large crowd for Hadoop Summit

University Researchers / Educators Eager to Get Involved

- Spans wide range of CS disciplines
- Across multiple institutions

More Information

“Data-Intensive Supercomputing: The case for DISC”

- **Tech Report: CMU-CS-07-128**
- **Available from**
<http://www.cs.cmu.edu/~bryant>