Data Intensive 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 built them 4 PB data warehouse
- Mine data to manage supply chain, understand market trends, formulate pricing strategies

LSST

- Chilean telescope will scan entire sky every 3 days
- A 3.2 gigapixel digital camera
- Generate 30 TB/day of image data

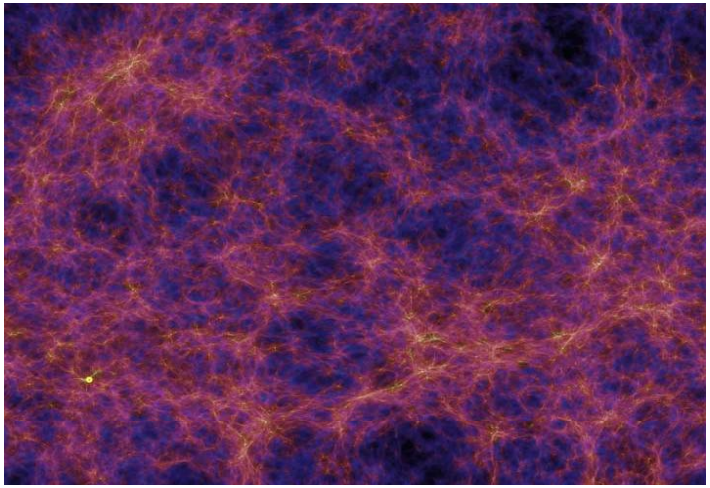
Big Data Sources: Astronomy

Sky Surveys

- **Sloan Survey: 60 TB**
- **Catalog: 350 Million objects**
- **Pan-STARRS: 10x (2009)**
- **LSST: 100x (2016)**

Simulation Results

- **McWilliams Center**
  - 15 B particles
  - 14 TB/run



Why So Much Data?

We Can Get It

- Automation + Internet

We Can Keep It

- Seagate Barracuda
- 1.5 TB @ $120 (8¢ / GB)

Seagate Barracuda 7200.11 1.5 TB SATA 32 MB Cache Bulk/OEM Hard Drive ST31500341AS

Other products by Seagate

(235 customer reviews) | More about this product

Size Name:

1.5 TB: $119.99

List Price: ~~$199.99~~

Price: $119.99 Free Standard Shipping (3-5 days)

Details

You Save: $80.00 (40%)

Special Offers Available

In Stock.

Ships from and sold by **Amazon.com.** Gift-wrap available.

Want it delivered Wednesday, October 14? Order it in the next 20 hours and 15 minutes, and choose **One-Day Shipping** at checkout. Details

11 new from $109.90 **2 used** from $103.00

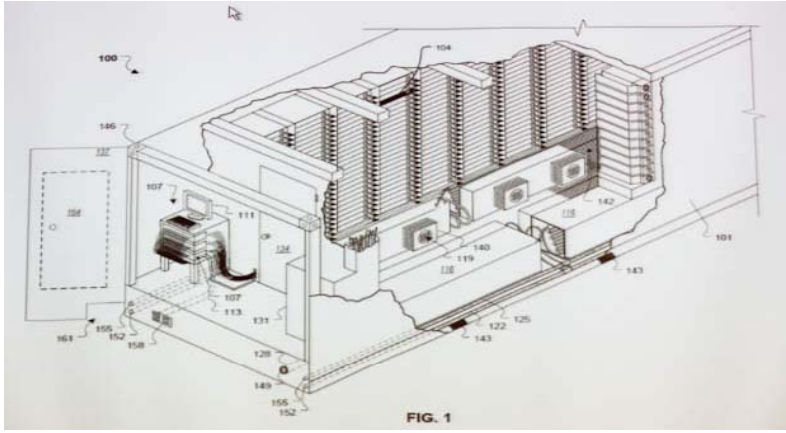
1 refurbished from $114.99

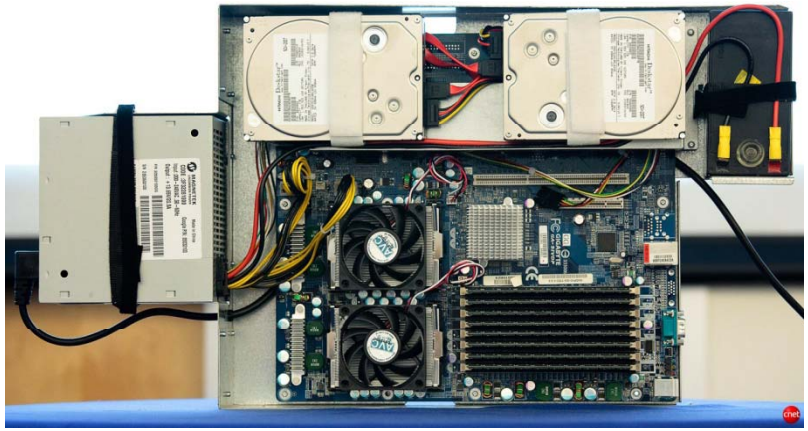
Ships Separately in Original Packaging: To conceal its contents mark it as a gift.

Could We Do More?

- Apply more computing power to this data

Google Data Centers





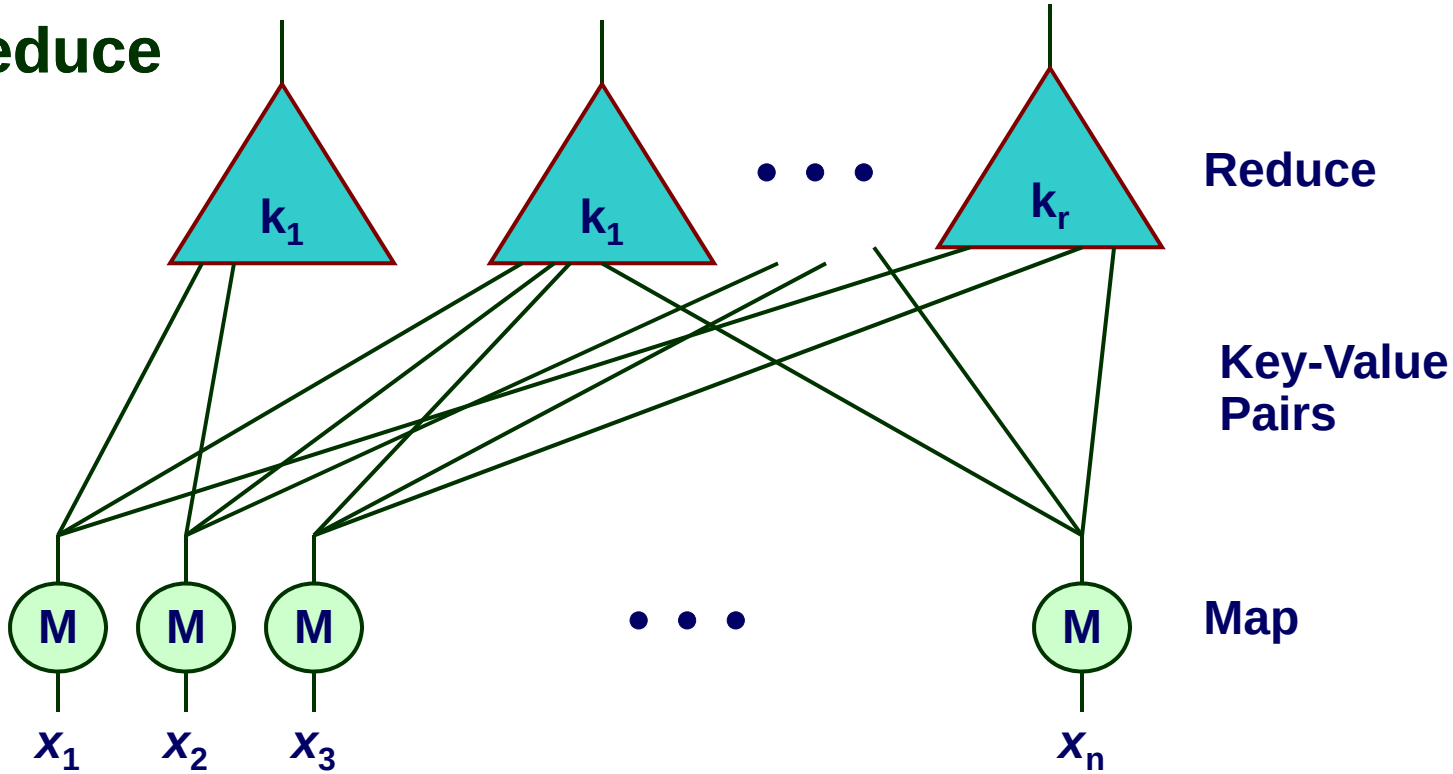
Dalles, Oregon

- **Hydroelectric power @ 2¢ / KW Hr**
- **50 Megawatts**
  - Enough to power 6,000 homes

- **Engineered for maximum modularity & power efficiency**
- **Container: 1160 servers, 250KW**
- **Server: 2 disks, 2 processors**

Google's Programming Model

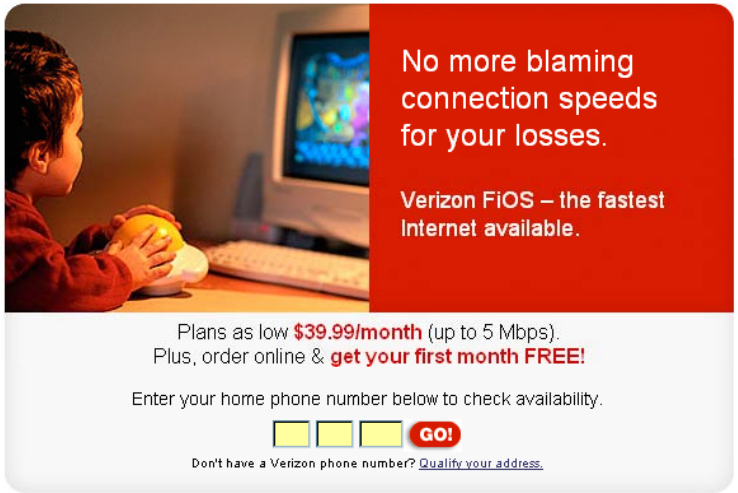
MapReduce



- **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

Oceans of Data, Skinny Pipes



Seagate

1 Terabyte

- **Easy to store**
- **Hard to move**

Disks	MB / s	Time
Seagate Barracuda	115	2.3 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

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

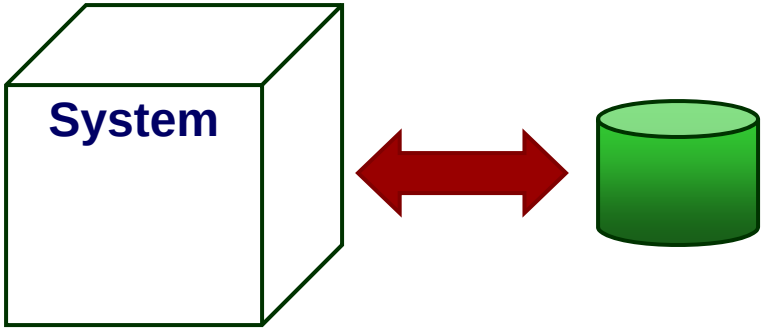
Robust Fault Tolerance

- Component failures are handled as routine events

Contrast to existing supercomputer / HPC systems

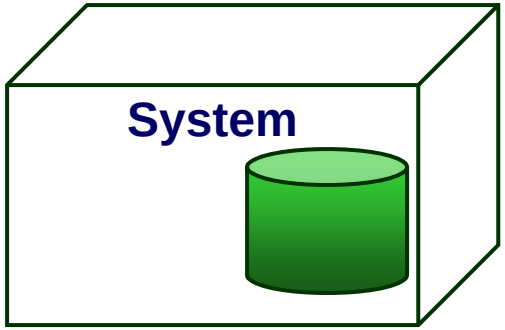
System Comparison: Data

Conventional Supercomputers



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

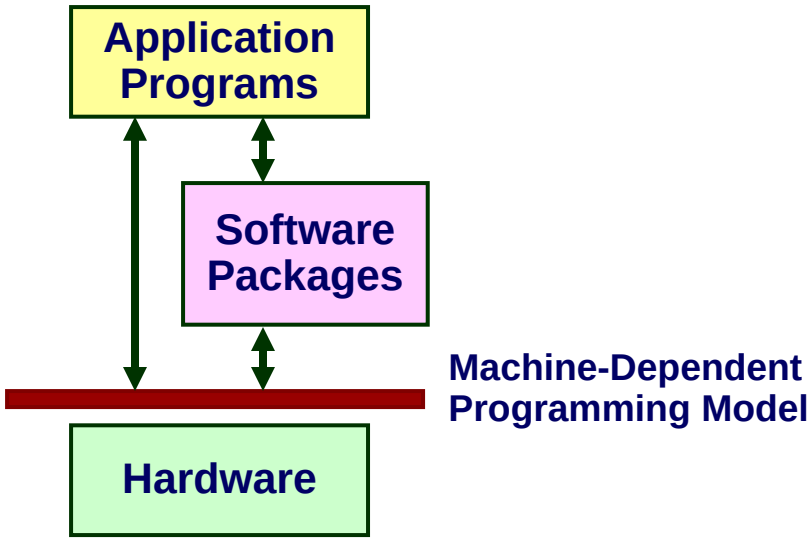
DISC



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

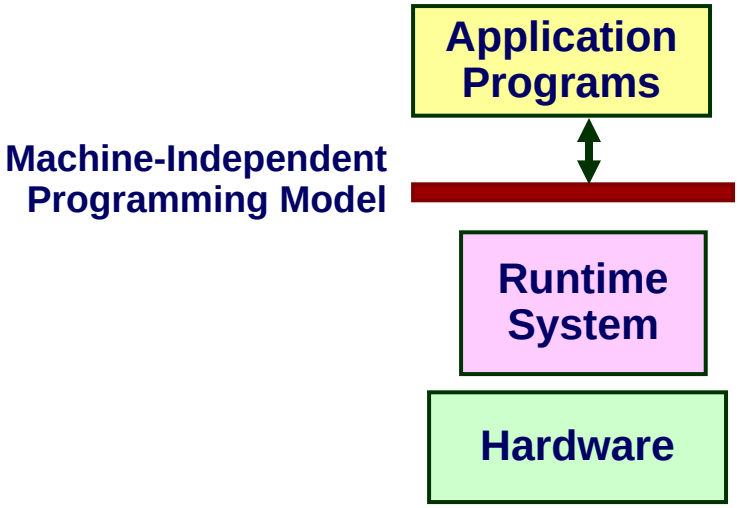
System Comparison: Programming Models

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: 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

CS Research Issues

Applications

- Language translation, image processing, …

Application Support

- Machine learning over very large data sets
- Web crawling

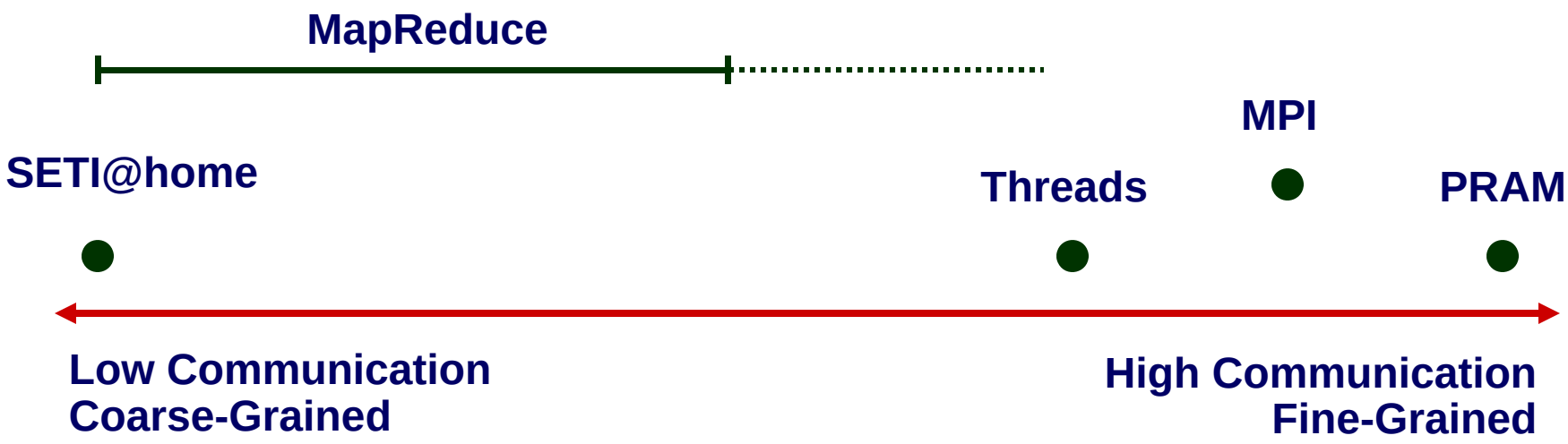
Programming

- 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

Exploring Parallel Computation Models



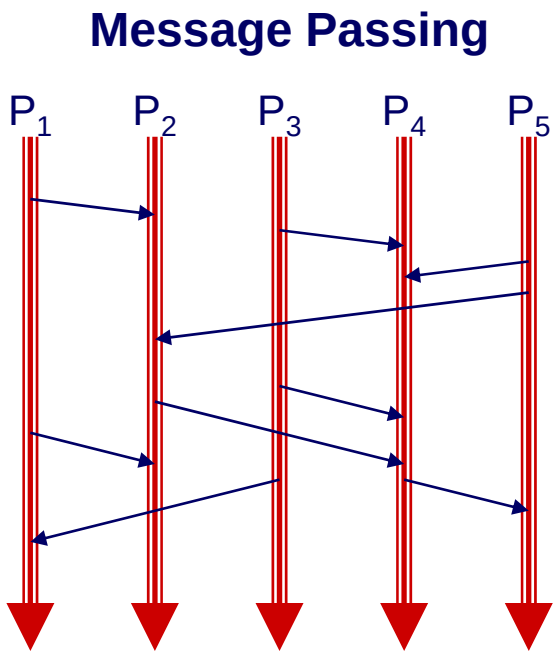
DISC + MapReduce Provides Coarse-Grained Parallelism

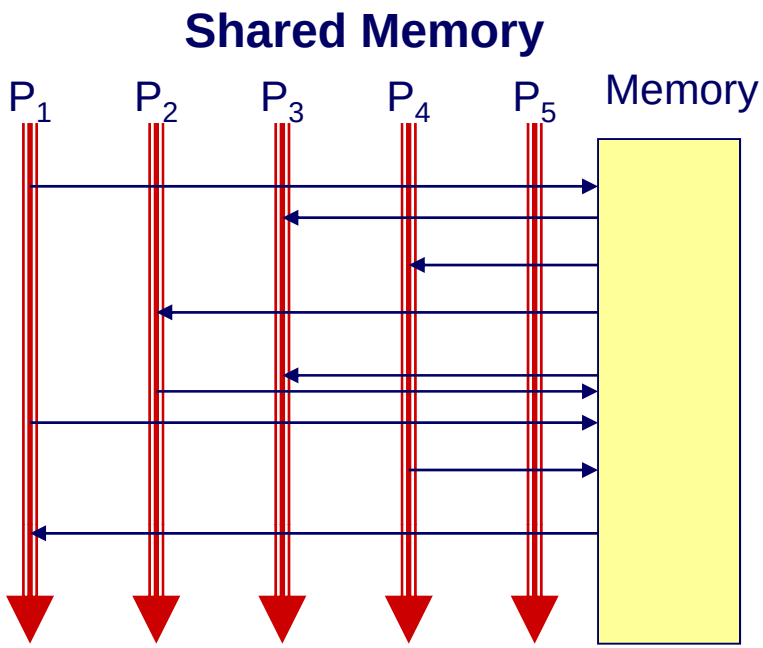
- Computation done by independent processes
- File-based communication

Observations

- Relatively "natural" programming model
- Research issue to explore full potential and limits
  - Dryad project at MSR
  - Pig project at Yahoo!

Existing HPC Machines





Characteristics

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

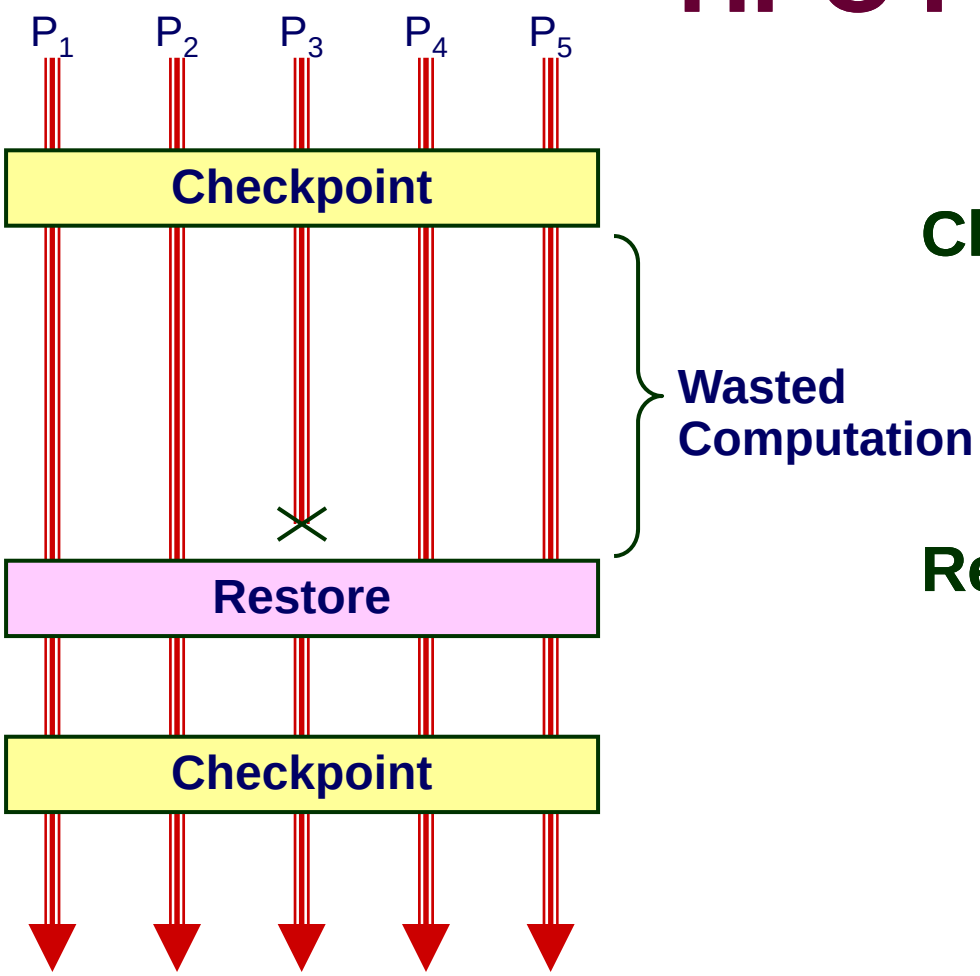
Strengths

- 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

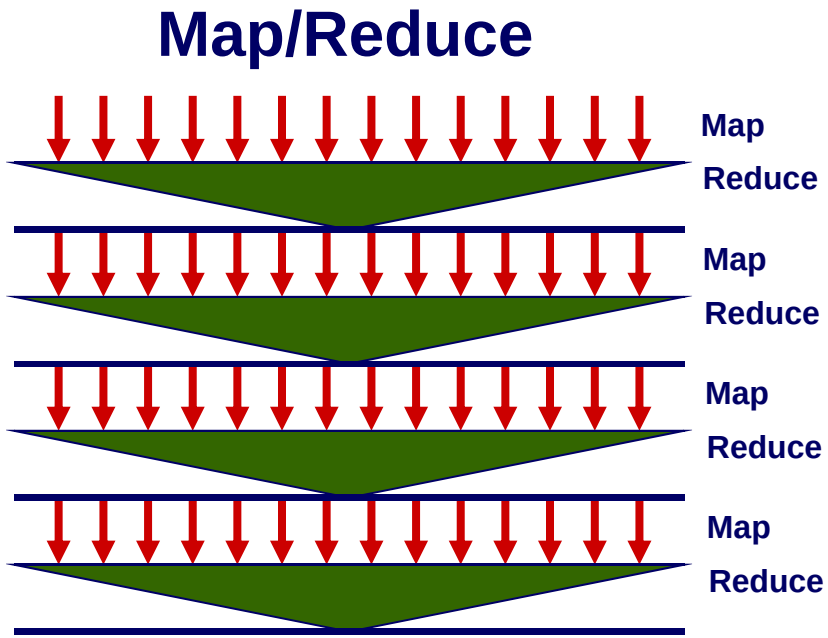
Restore

- 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



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

Generalizing Map/Reduce

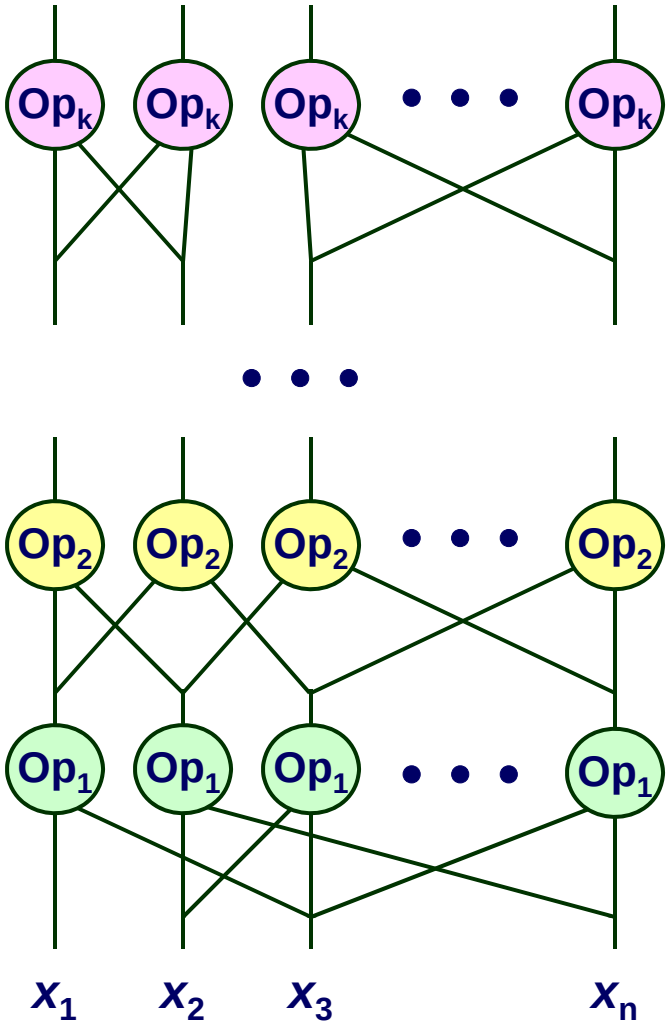
- **E.g., Microsoft Dryad Project**

Computational Model

- **Acyclic graph of operators**
  - But expressed as textual program
- **Each takes collection of objects and produces objects**
  - Purely functional model

Implementation Concepts

- **Objects stored in files or memory**
- **Any object may be lost; any operator may fail**
- **Replicate & recompute for fault tolerance**
- **Dynamic scheduling**
  - # Operators >> # Processors



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

Many Opportunities for Academics to Get Involved

- **Challenging research problems across CS**
- **Applications in multiple disciplines**
- **Across multiple institutions**