Introduction to Computer Science

The What, How, and Why of CS

Ioan Raicu
Computer Science Department, Illinois Institute of Technology
Math and Computer Science Division, Argonne National Laboratory

digitalblasphemy

February 14th, 2013

What is Computer Science?

- The scientific and mathematical approach in information technology and computing
- Started in the 1960s from Mathematics or Electrical Engineering
- Today:
  - Arguably one of the most fundamental discipline that touches all other disciplines and people

Famous Quotes

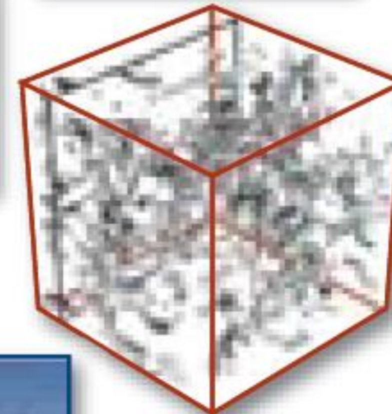
The advent of computation can be compared, in terms of the breadth and depth of its impact on research and scholarship, to the invention of writing and the development of modern mathematics.

Ian Foster, 2006

Science Paradigms

- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations

$$\left(\frac{\dot{a}}{a}\right)^2 = \frac{4\pi Gp}{3} - K\frac{c^2}{a^2}$$



- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, and simulation
  – Data captured by instruments
  or generated by simulator
  – Processed by software
  – Information/knowledge stored in computer
  – Scientist analyzes database/files
  using data management and statistics

Computer Science Theory

- Theory
  - Theory of computation
  - Information and coding theory
  - Algorithms and data structures
  - Programming language theory
  - Formal methods
- Systems

Computer Science Systems

- Theory
- Systems
  - Artificial intelligence
  - Computer architecture
  - Computer graphics and visualization
  - Computer security and cryptography
  - Computational science
  - Databases and information retrieval
  - Distributed systems
  - Health Informatics
  - Information science
  - Programming Languages
  - Software engineering

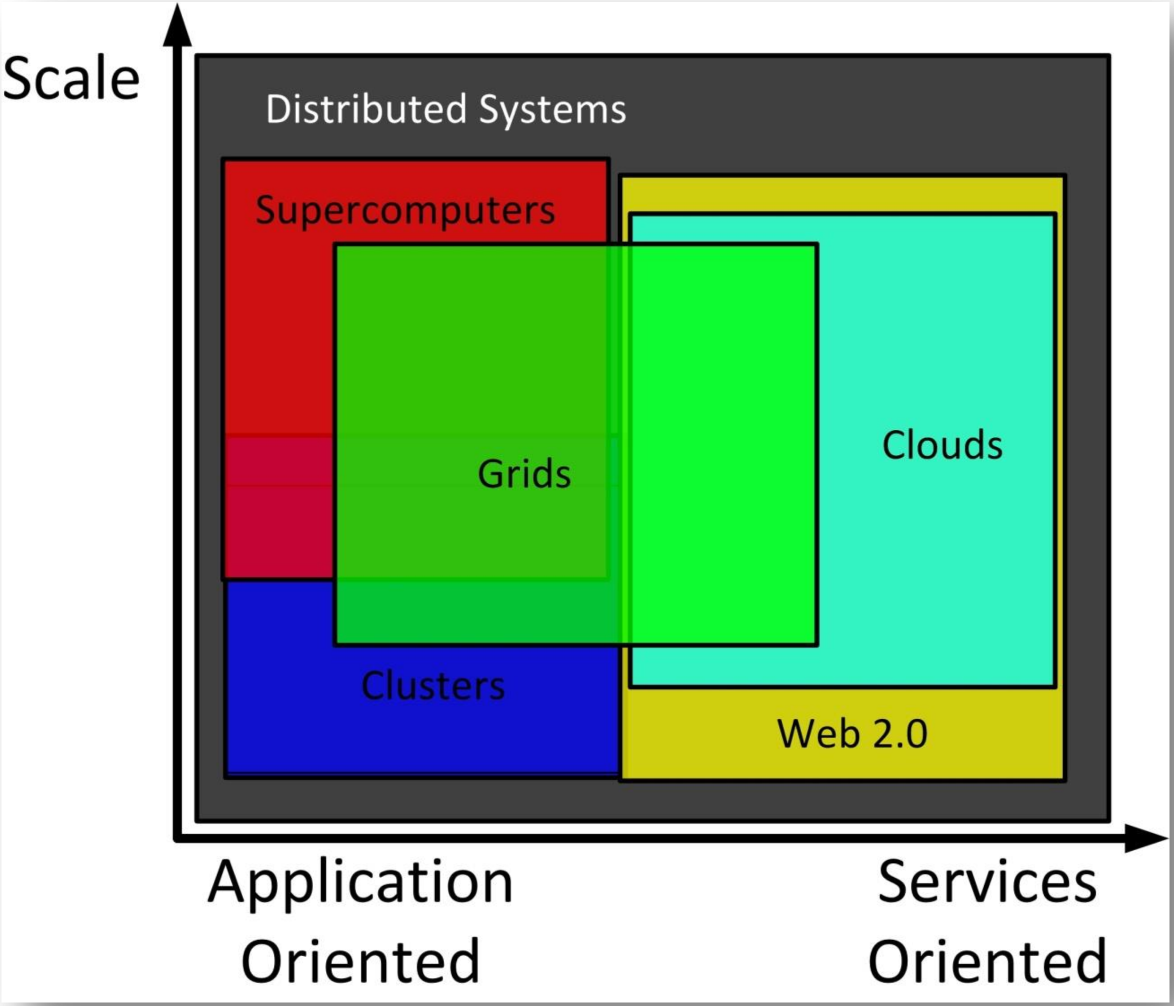
How?

- What is a distributed system?

“A collection of independent computers that appears to its users as a single coherent system”

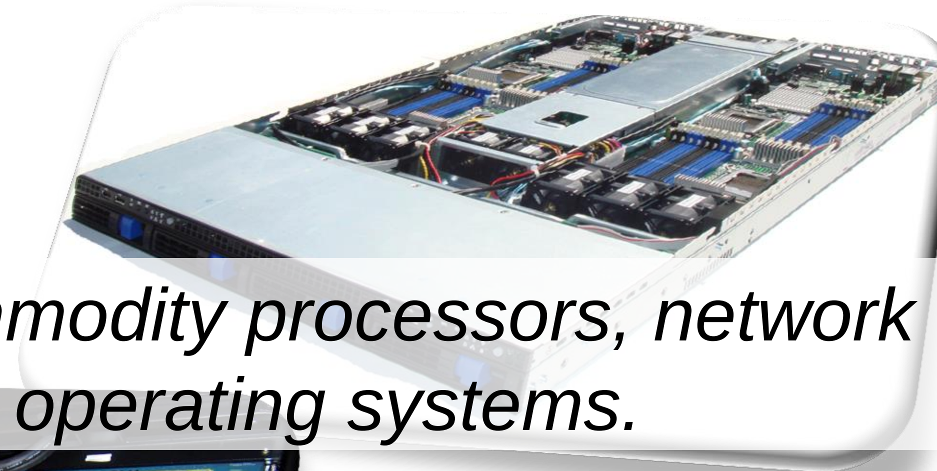
-A. Tanenbaum

Distributed Systems:
Clusters, Grids, Clouds, and Supercomputers



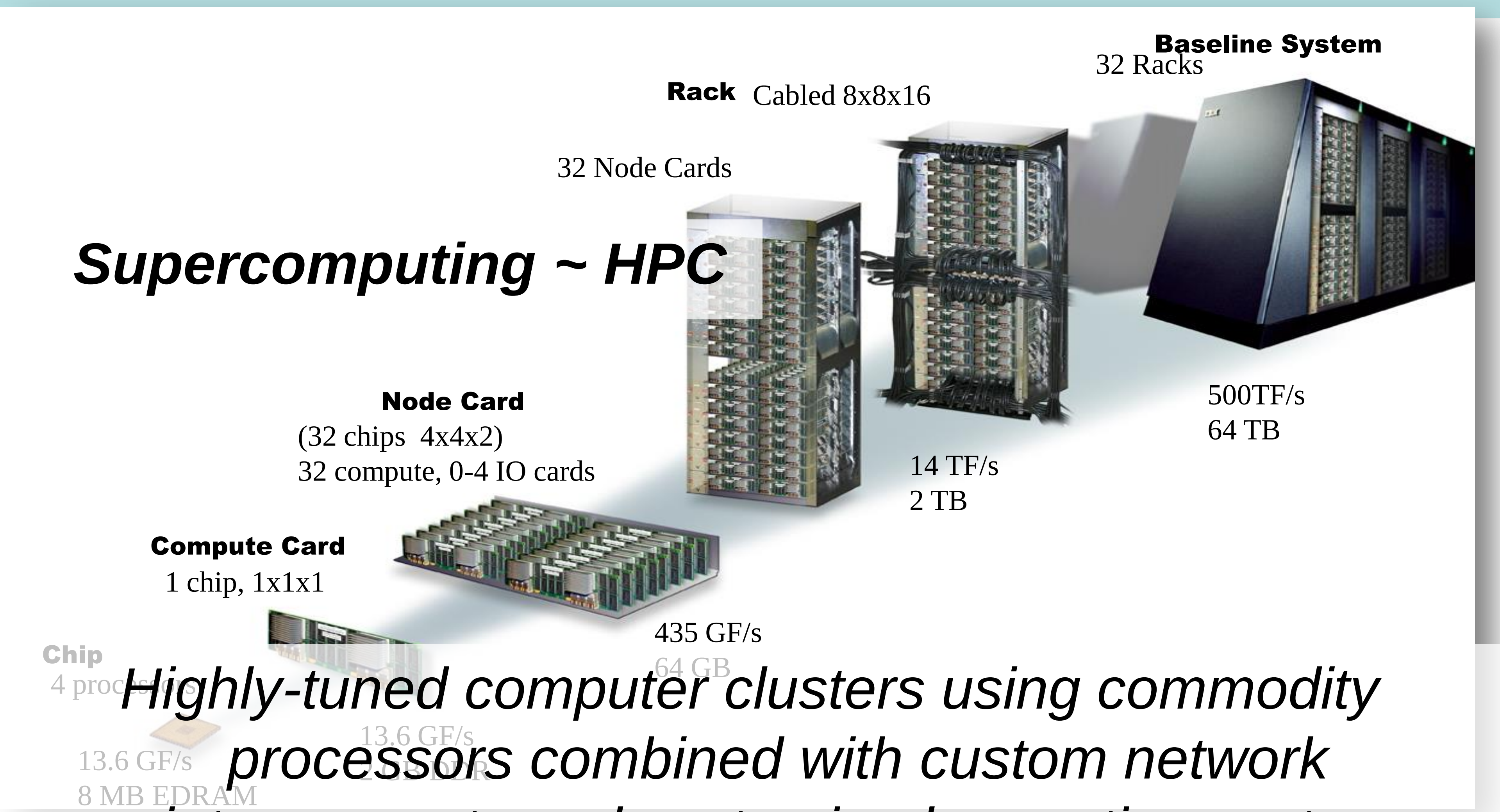
[GCE08] "Cloud Computing and Grid Computing 360-Degree Compared"

Cluster Computing



Computer clusters using commodity processors, network interconnects, and operating systems.

Supercomputing

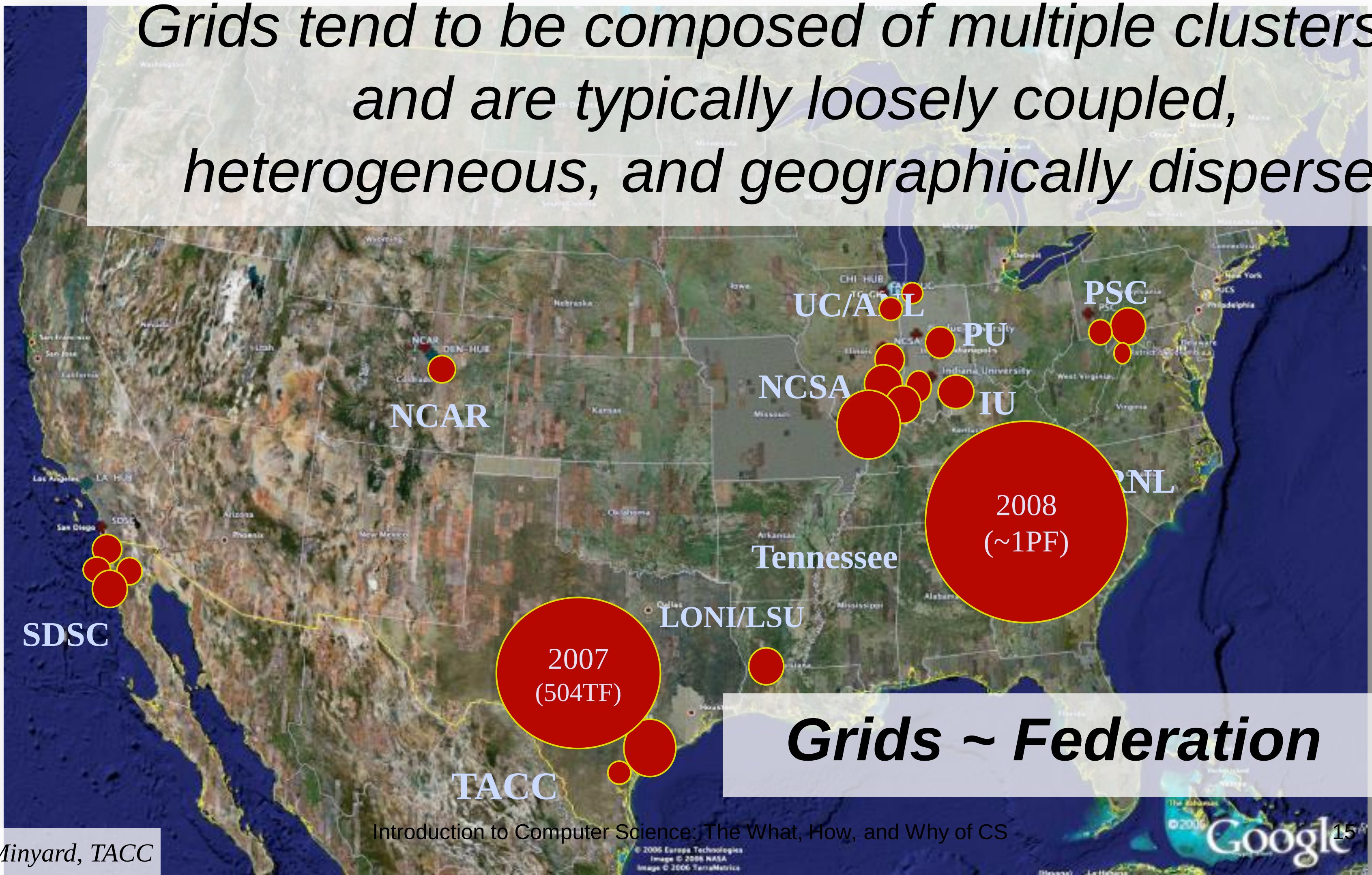


Supercomputing ~ HPC

Highly-tuned computer clusters using commodity processors combined with custom network interconnects and customized operating system

Grid Computing

Grids tend to be composed of multiple clusters, and are typically loosely coupled, heterogeneous, and geographically dispersed

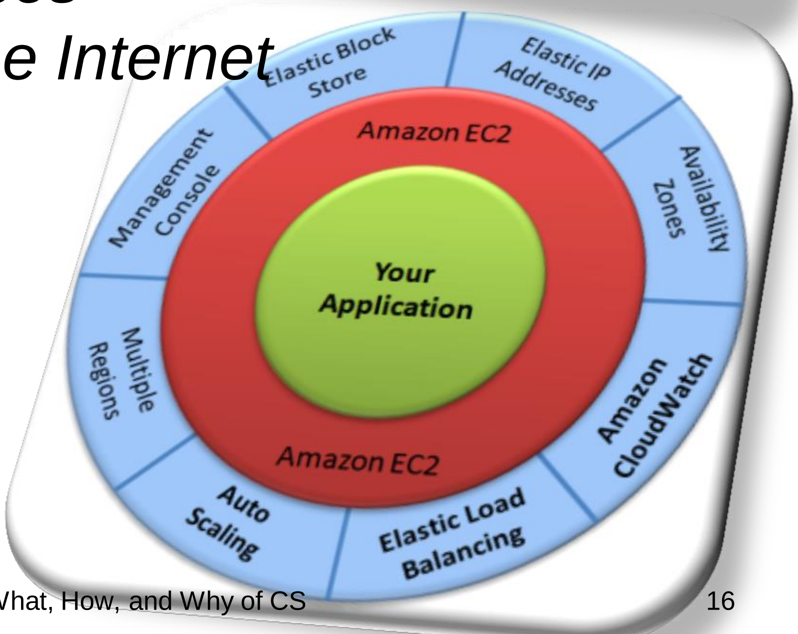


Grids ~ Federation

Tommy Minyard, TACC

Cloud Computing

- *A large-scale distributed computing paradigm driven by:*
  1. *economies of scale*
  2. *virtualization*
  3. *dynamically-scalable resources*
  4. *delivered on demand over the Internet*



Clouds ~ hosting

Coursework

- CS 451 Introduction to Distributed Computing
- CS 546 Parallel and Distributed Processing
- CS 550 Advanced Operating Systems
- CS 552 Distributed Real-Time Systems
- CS 553 Cloud Computing
- CS 554 Data-Intensive Distributed Computing
- CS 570 Advanced Computer Architecture

Faculty

- Xian-He Sun

- Zhiling Lan

- Shangping Ren

- Ioan Raicu

SCS
Scalable Computing Software Laboratory

DataSys
Data-Intensive Distributed Systems Laboratory

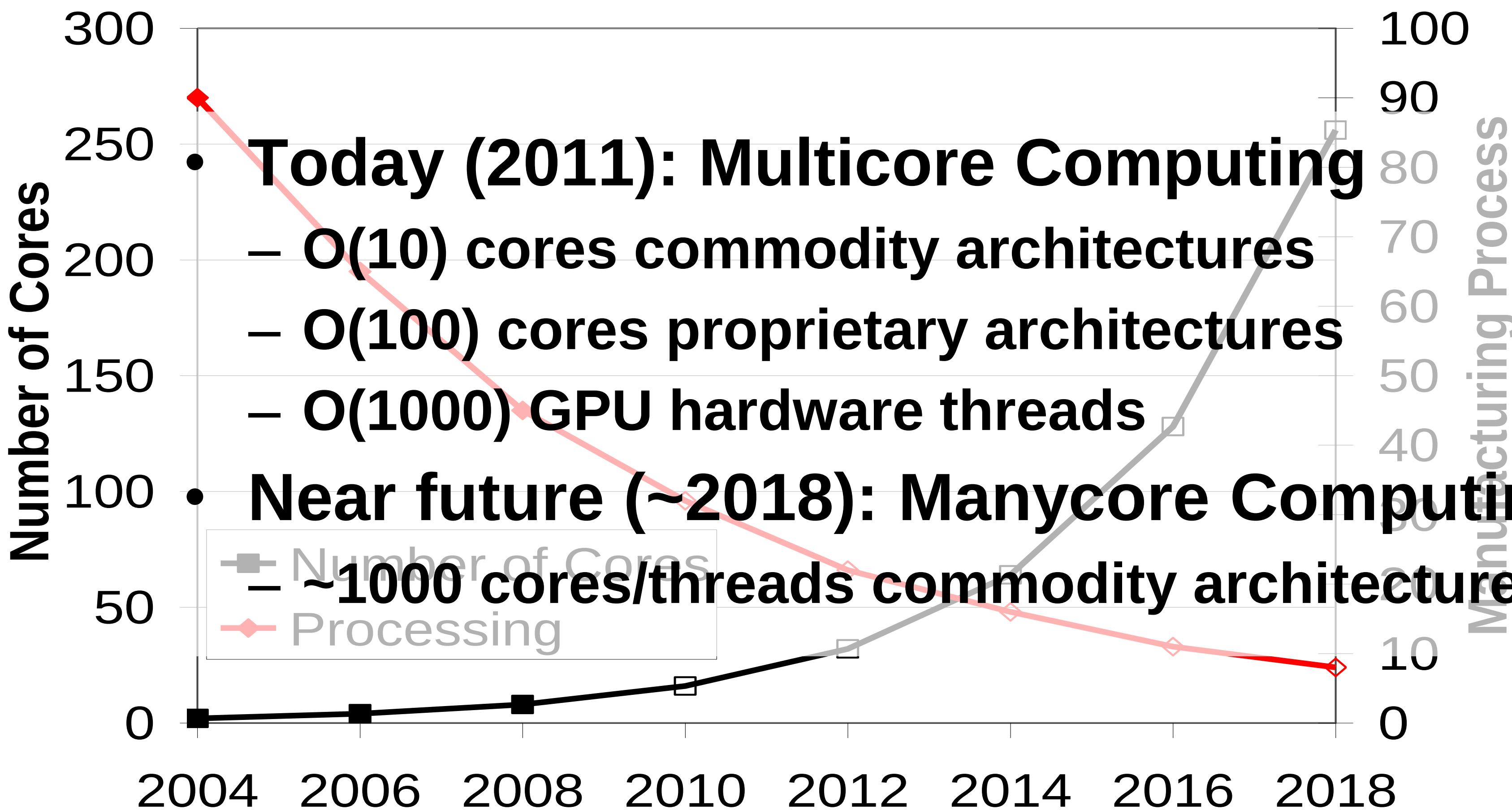
DataSys: Data-Intensive Distributed Systems Laboratory

- **Research Focus**
  - Emphasize designing, implementing, and evaluating systems, protocols, and middleware with the goal of supporting **data-intensive applications on extreme scale distributed systems**, from many-core systems, clusters, grids, clouds, and supercomputers
- **People**
  - Dr. Ioan Raicu (Director)
  - 5 PhD Students
  - 4 MS Students
  - 5 UG Students

- **Contact**
  - http://datasys.cs.iit.edu/
  - iraicu@cs.iit.edu

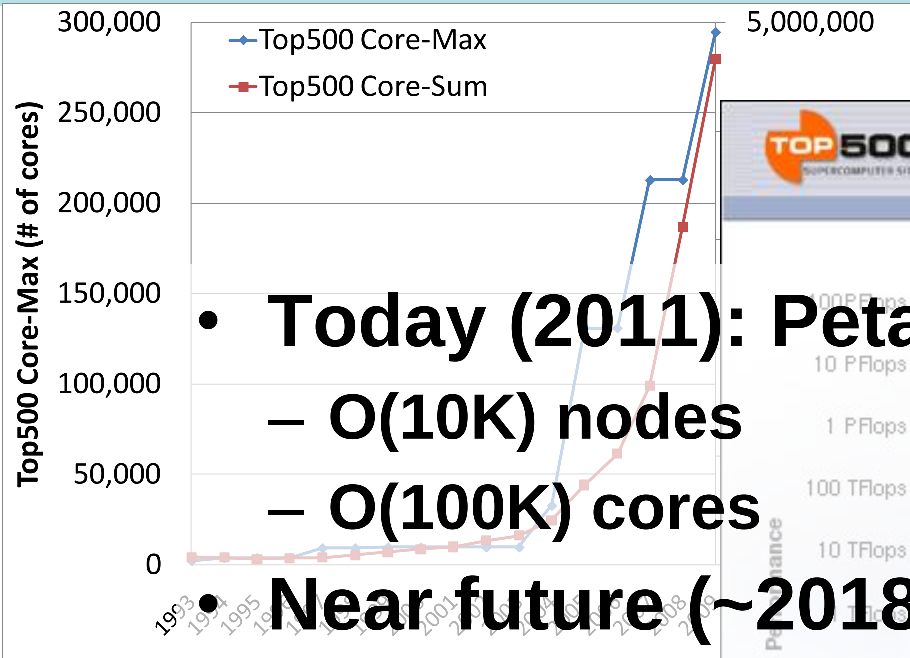
Manycore Computing

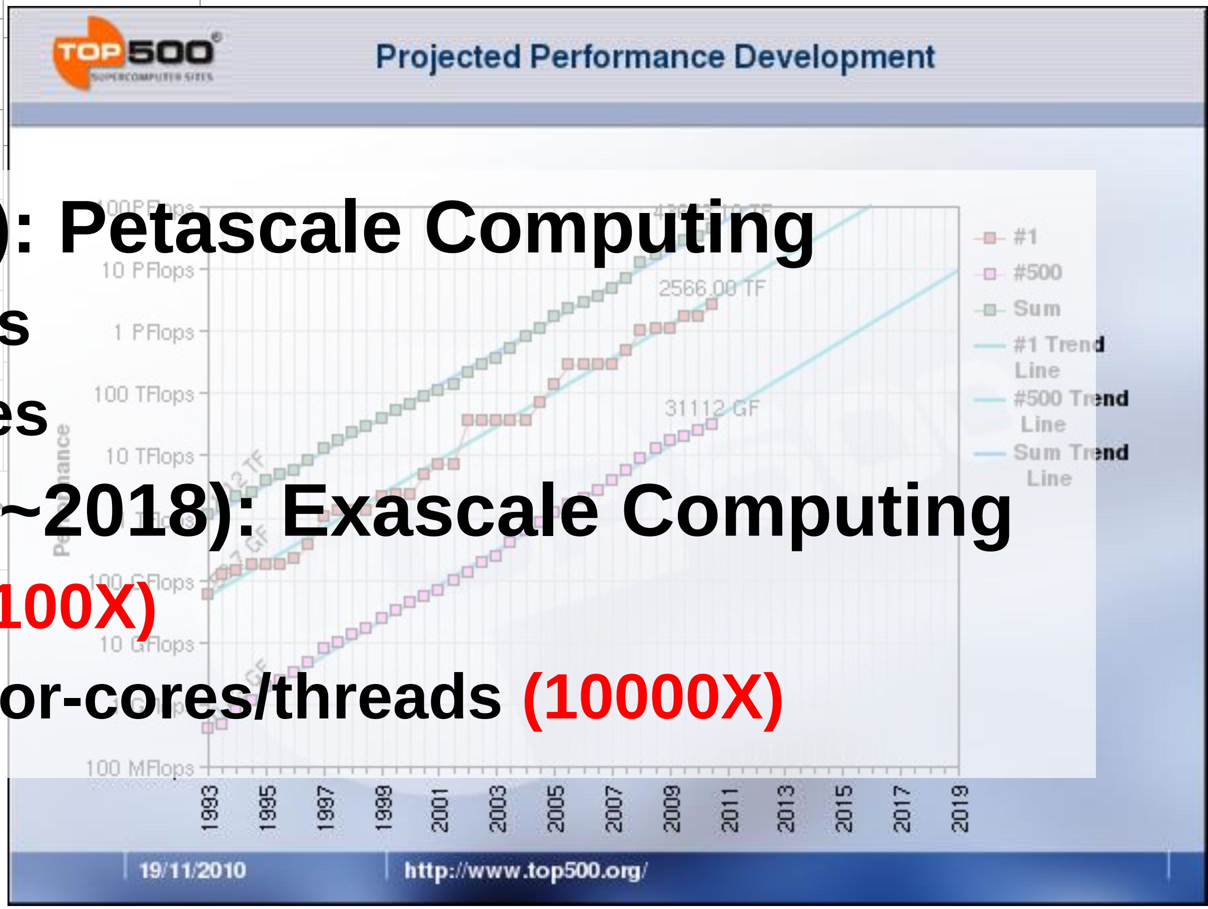
- **Today (2011): Multicore Computing**
  - **O(10) cores commodity architectures**
  - **O(100) cores proprietary architectures**
  - **O(1000) GPU hardware threads**
- **Near future (~2018): Manycore Computing**
  - **~1000 cores/threads commodity architectures**



Pat Helland, Microsoft, The Irresistible Forces Meet the Movable Objects, November 9th, 2007

Exascale Computing





- **Today (2011): Petascale Computing**
  - **O(10K) nodes**
  - **O(100K) cores**
- **Near future (~2018): Exascale Computing**
  - **~1M nodes (100X)**
  - **~1B processor-cores/threads (10000X)**

Top500 Projected Development,
http://www.top500.org/lists/2010/11/performance_development

Cloud Computing

- Relatively new paradigm… 3 years old
- Amazon in 2009
  - 40K servers split over 6 zones
    - 320K-cores, 320K disks
    - $100M costs + $12M/year in energy costs
    - Revenues about $250M/year
- Amazon in 2018
  - Will likely look similar to exascale computing
    - 100K~1M nodes, ~1B-cores, ~1M disks
    - $100M~$200M costs + $10M~$20M/year in energy
    - Revenues 100X~1000X of what they are today

Common Challenges

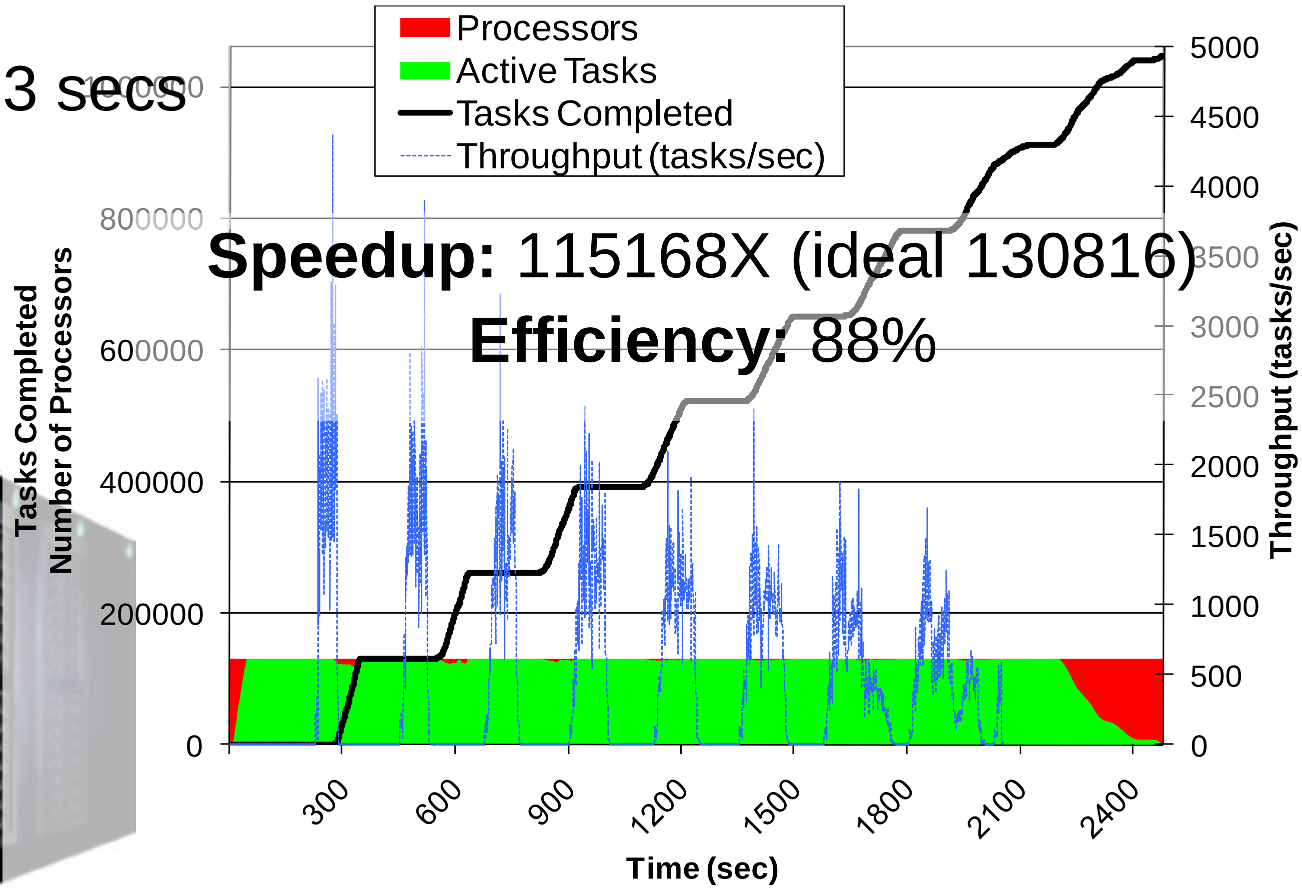
- Power efficiency
  - Will limit the number of cores on a chip (Manycore)
  - Will limit the number of nodes in cluster (Exascale and Cloud)
  - Will dictate a significant part of the cost of ownership
- Programming models/languages
  - Automatic parallelization
  - Threads, MPI, workflow systems, etc
  - Functional, imperative
  - Languages vs. Middlewares

Common Challenges

- Bottlenecks in scarce resources
  - Storage (Exascale and Clouds)
  - Memory (Manycore)
- Reliability
  - How to keep systems operational in face of failures
  - Checkpointing (Exascale)
  - Node-level replication enabled by virtualization (Exascale and Clouds)
  - Hardware redundancy and hardware error correction (Manycore)

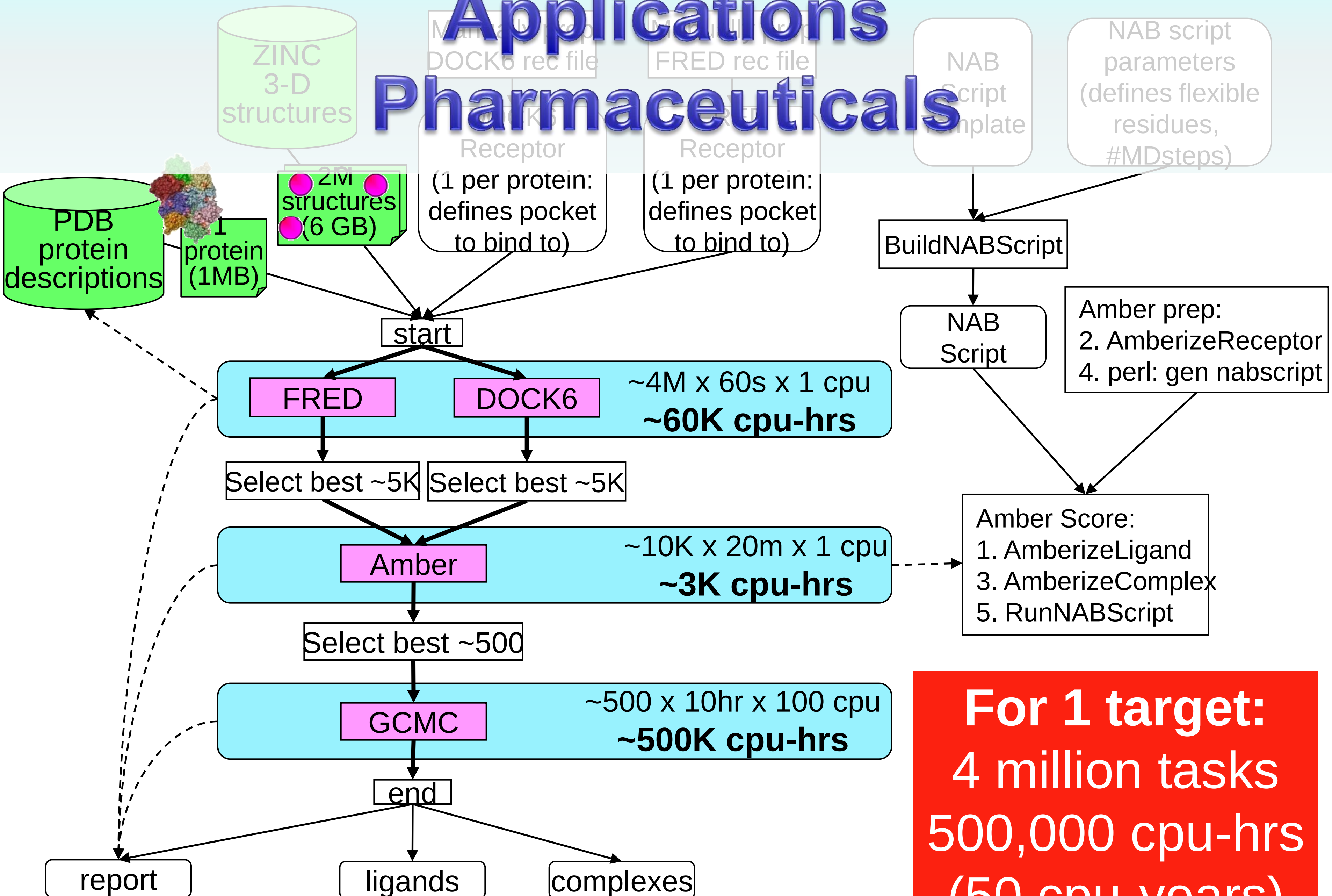
Applications
Economic Modeling: MARS

- CPU Cores: 130816
- Tasks: 1048576
- Elapsed time: 2483 secs
- CPU Years: 9.3



[SC08] "Towards Loosely-Coupled Programming on Petascale Systems"

Applications Pharmaceuticals



For 1 target:
4 million tasks
500,000 cpu-hrs
(50 cpu-years)

[SC08] “Towards Loosely-Coupled Programming on Petascale Systems”

Applications
Pharmaceuticals: DOCK

CPU cores: 118784

Tasks: 934803

Elapsed time: 2.01 hours

Compute time: 21.43 CPU years

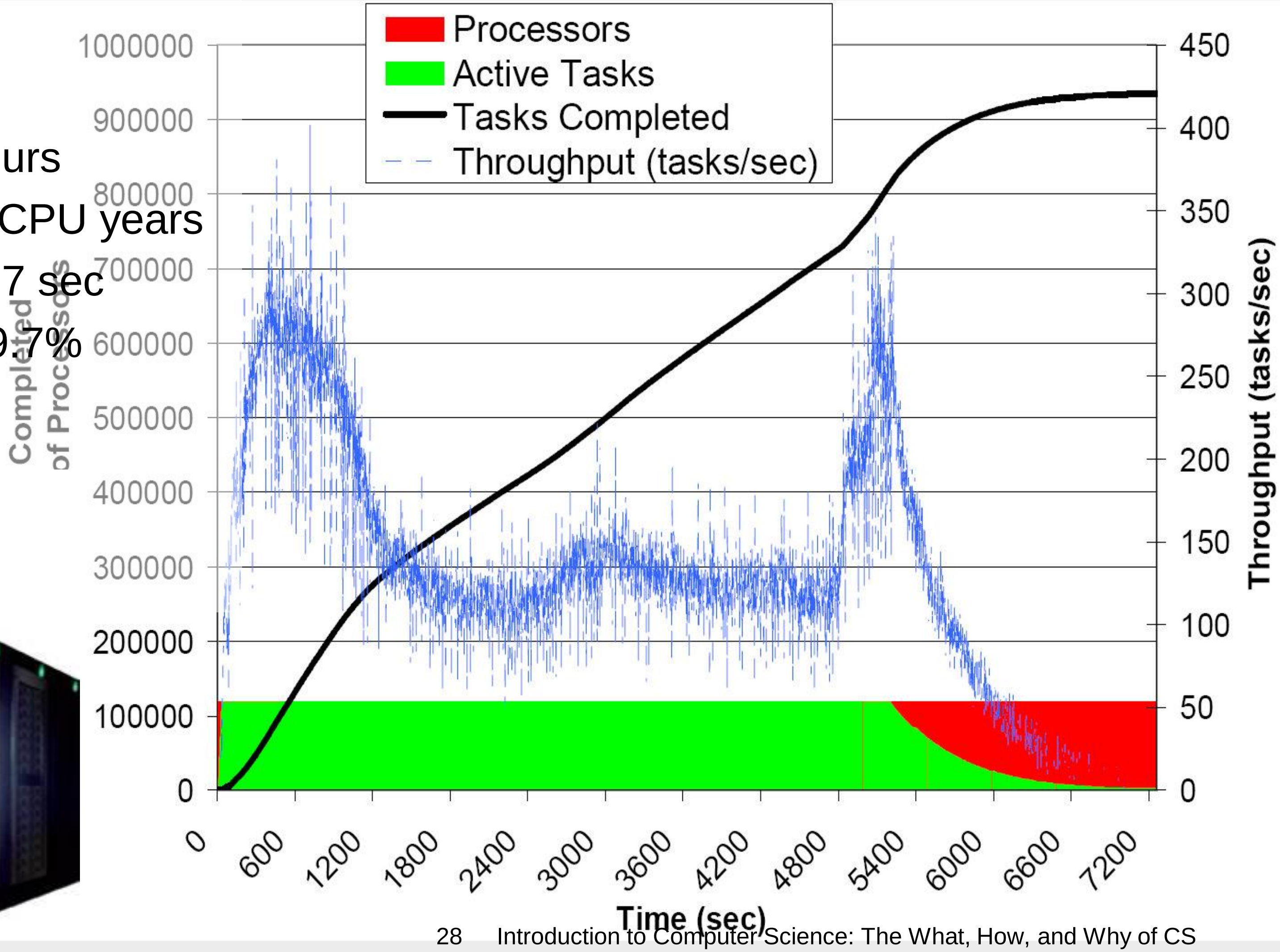
Average task time: 667 sec

Relative Efficiency: 99.7%

(from 16 to 32 racks)

Utilization:

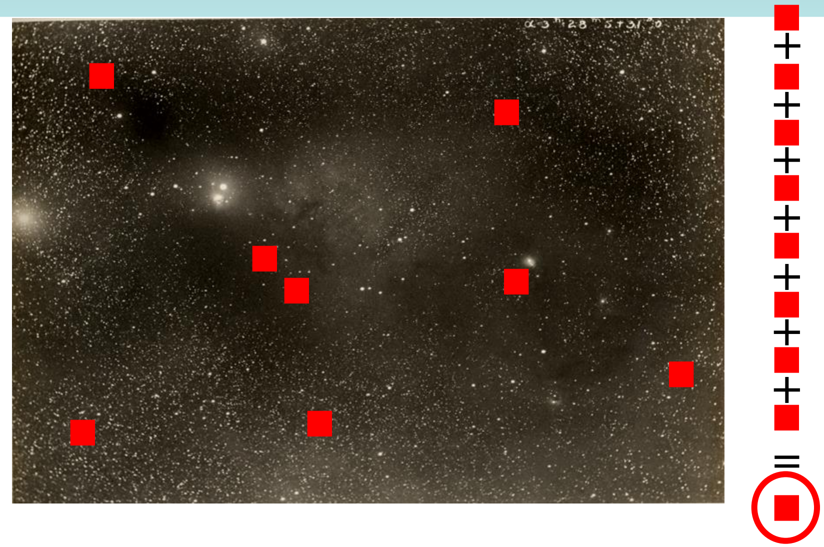
- Sustained: 99.6%
- Overall: 78.3%

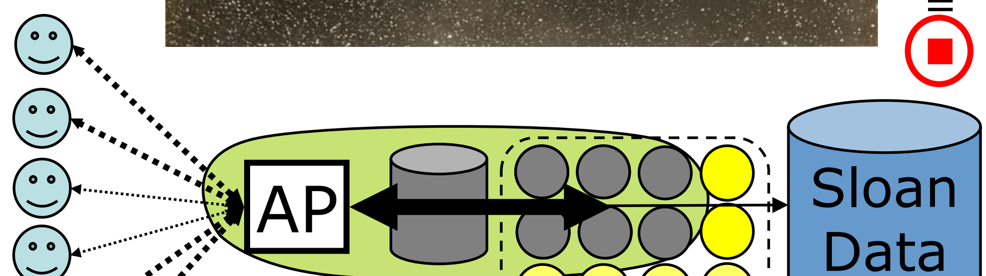


[SC08] "Towards Loosely-Coupled Programming on Petascale Systems"

Applications
Astronomy: AstroPortal

- Purpose
  - On-demand “stacks” of random locations within ~10TB dataset
- Challenge
  - Processing Costs:
    - O(100ms) per object
  - Data Intensive:
    - 40MB:1sec
  - Rapid access to 10-10K “random” files
  - Time-varying load





Locality	Number of Objects	Number of Files
1	111700	111700
1.38	154345	111699
2	97999	49000
3	88857	29620
4	76575	19145
5	60590	12120
10	46480	4650
20	40460	2025
30	23695	790

[DADC08] “Accelerating Large-scale Data Exploration through Data Diffusion”
[TG06] “AstroPortal: A Science Gateway for Large-scale Astronomy Data Analysis”

Why?

- Be the one creating and shaping the future of technology, not just the user
- Employment at the best technology companies in the world (see next slide)
- Be the next **Steve Jobs** (Apple), **Bill Gates** (Microsoft), **Sergei Brin** (Google), or **Zach Zuckerberg** (Facebook)
- Be part of the most amazing revolution to date: **The Computing Revolution!**

Employment Opportunities Distributed Systems

- Google
- Yahoo
- Microsoft
- Amazon
- IBM
- Apple
- VMWare
- Netflix
- Cray
- Intel
- NVIDIA
- Facebook
- LinkedIn
- Salesforce.com
- Rackspace
- Red Hat
- Cleversafe
- UnivaUD
- Greenplum
- AsterData
- Proprietary Trading Companies
- Department of Energy Laboratories
- NASA
- Academic supercomputer centers
- Many more…

More Information

- More information:
  - http://www.cs.iit.edu/~iraicu/
  - http://datasys.cs.iit.edu/
- Contact:
  - iraicu@cs.iit.edu
- Questions?