Introduction to Distributed Systems

Ioan Raicu
Computer Science Department
Illinois Institute of Technology

CS 553: Cloud Computing
August 27th, 2014

Logistics

- Piazza online discussion forum
  - https://piazza.com/iit/fall2014/cs553/home
  - You will receive an invite later today via email
- BlackBoard
- TA and Professor alias
  - cs553-f14@datasys.cs.iit.edu
- Office hours today:
  - CANCELED (will reschedule and announce over Piazza)
- Form your groups:
  - https://piazza.com/class/hza36qliup33pg?cid=5
- CS553 next offered in Spring 2016

Famous Quotes

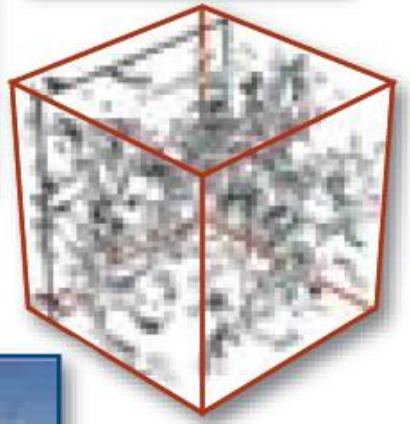
The advent of computation can be compared, in terms of the breadth and depth of its impact on research and scholarship, to the invention of writing and the development of modern mathematics.

Ian Foster, 2006

Science Paradigms

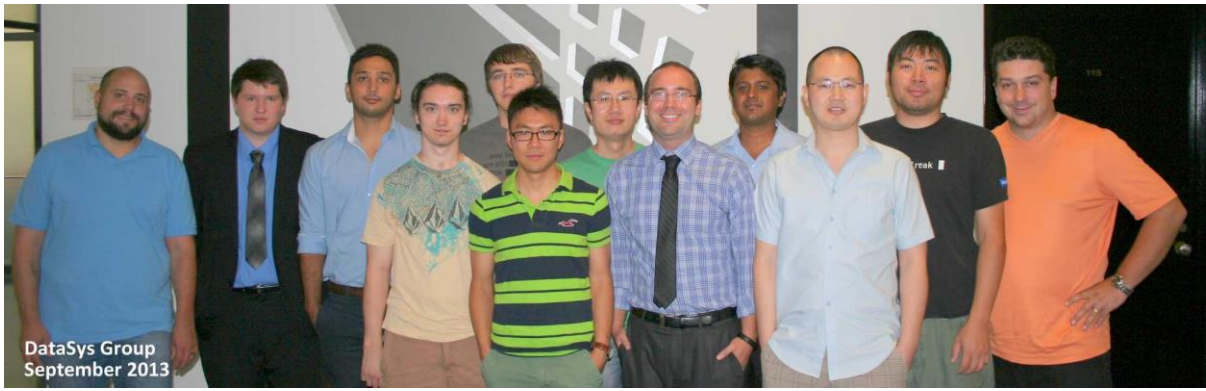
- Thousand years ago:
  science was **empirical**
  describing natural phenomena
- Last few hundred years:
  theoretical branch
  using models, generalizations
- Last few decades:
  a **computational** branch
  simulating complex phenomena
- Today: **data exploration** (eScience)
  unify theory, experiment, 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

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



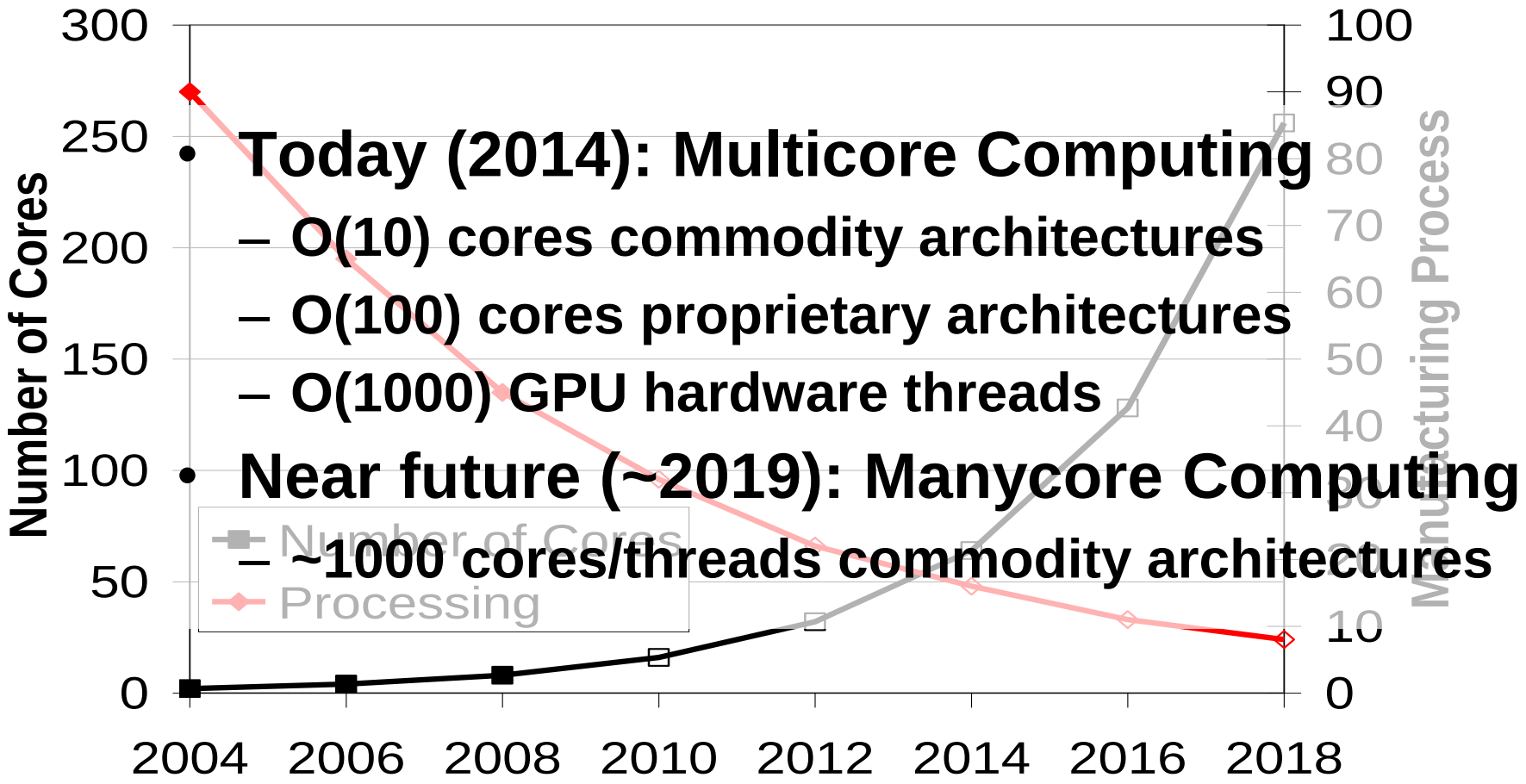
DataSys: Data-Intensive Distributed Systems Laboratory

- **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**
  - 1 Faculty member
  - 6 PhD students
  - 3 MS students
  - 6 UG student



- **More information**
  - http://datasys.cs.iit.edu/

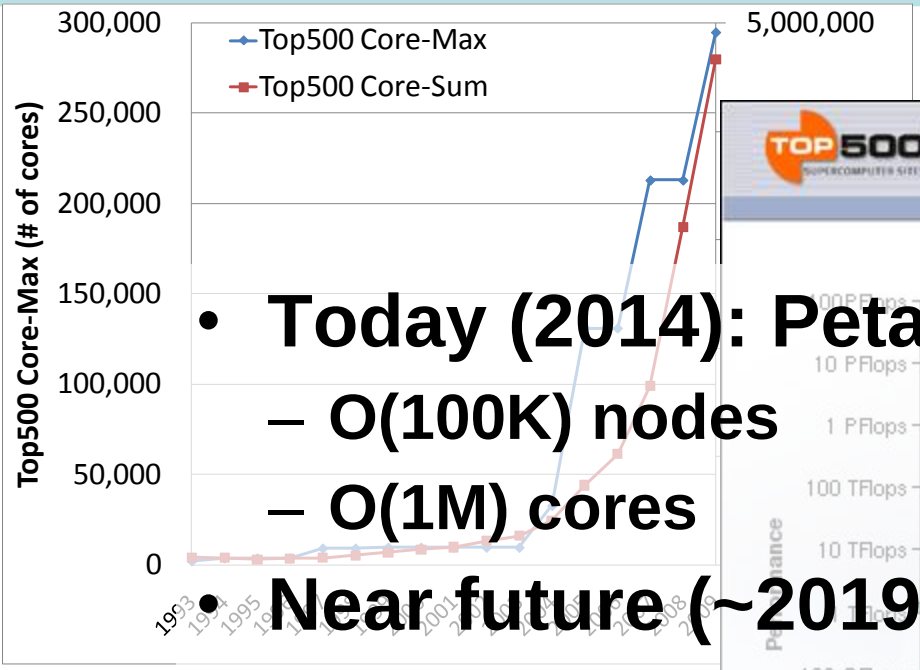
Manycore Computing

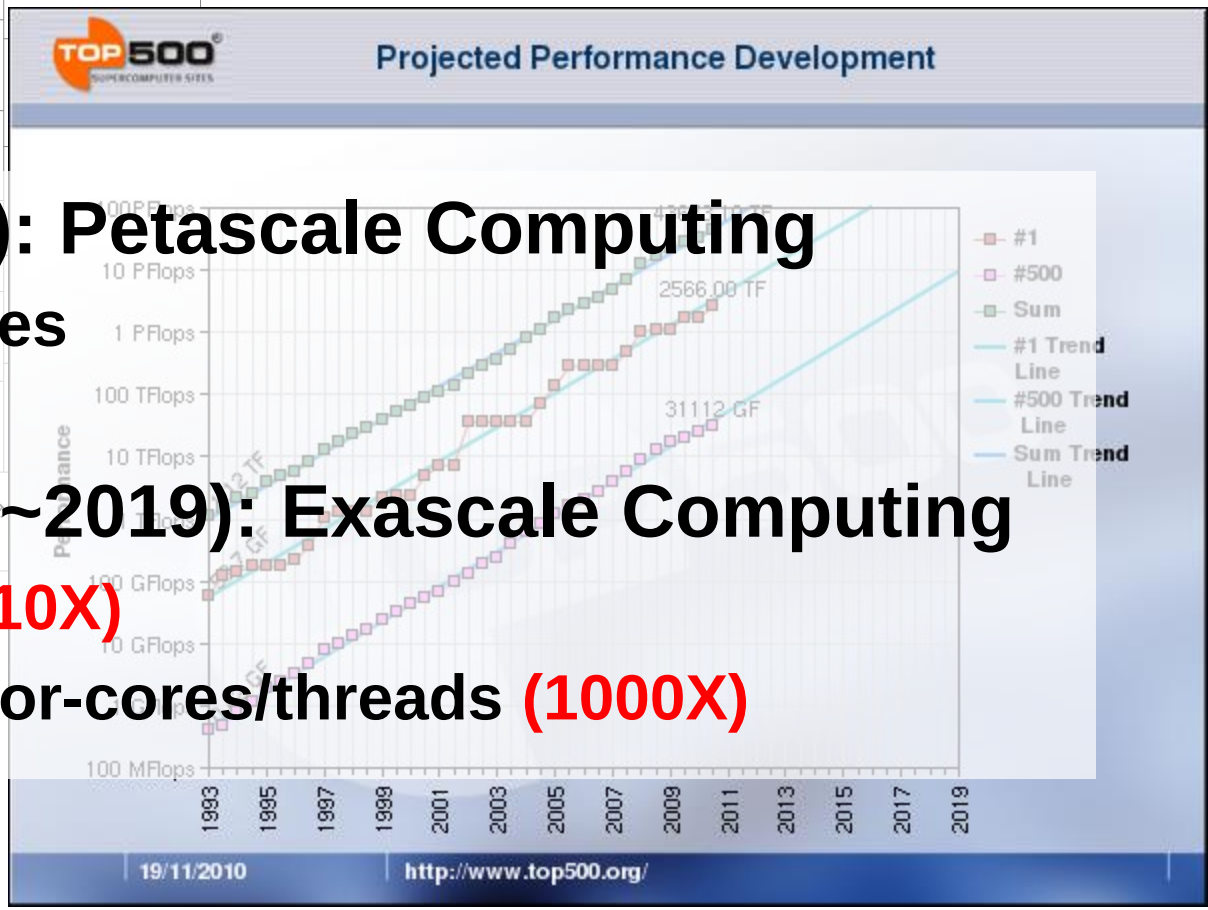


- **Today (2014): Multicore Computing**
  - **O(10) cores commodity architectures**
  - **O(100) cores proprietary architectures**
  - **O(1000) GPU hardware threads**
- **Near future (~2019): Manycore Computing**
  - **~1000 cores/threads commodity architectures**

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

Exascale Computing





- **Today (2014): Petascale Computing**
  - **O(100K) nodes**
  - **O(1M) cores**
- **Near future (~2019): Exascale Computing**
  - **~1M nodes (10X)**
  - **~1B processor-cores/threads (1000X)**

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

Cloud Computing

- Relatively new paradigm… ~6 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
    - http://www.siliconvalleywatcher.com/mt/archives/2009/10/meausuring_amaz.php
- Amazon in 2019
  - 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

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. Middleware

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)

My Research: Current Projects

- Swift: a Parallel Programming System
  - http://www.ci.uchicago.edu/swift/
- FusionFS: Fusion Distributed File System
  - http://datasys.cs.iit.edu/projects/FusionFS/index.html
- ZHT: Zero-Hop Distributed Hashtable
  - http://datasys.cs.iit.edu/projects/ZHT/index.html
- MATRIX: MAny-Task computing execution fabRIc at eXascales
  - http://datasys.cs.iit.edu/projects/MATRIX/index.html
- GeMTC: Virtualizing GPUs to Support MTC Applications
  - http://datasys.cs.iit.edu/projects/GeMTC/index.html
- CloudKon: a Cloud enabled Distributed Task Execution Framework
  - http://datasys.cs.iit.edu/projects/CloudKon/index.html
- HRDBMS: A Scalable Distributed Relational Database for Commodity Hardware
  - http://datasys.cs.iit.edu/projects/HRDBMS/index.html

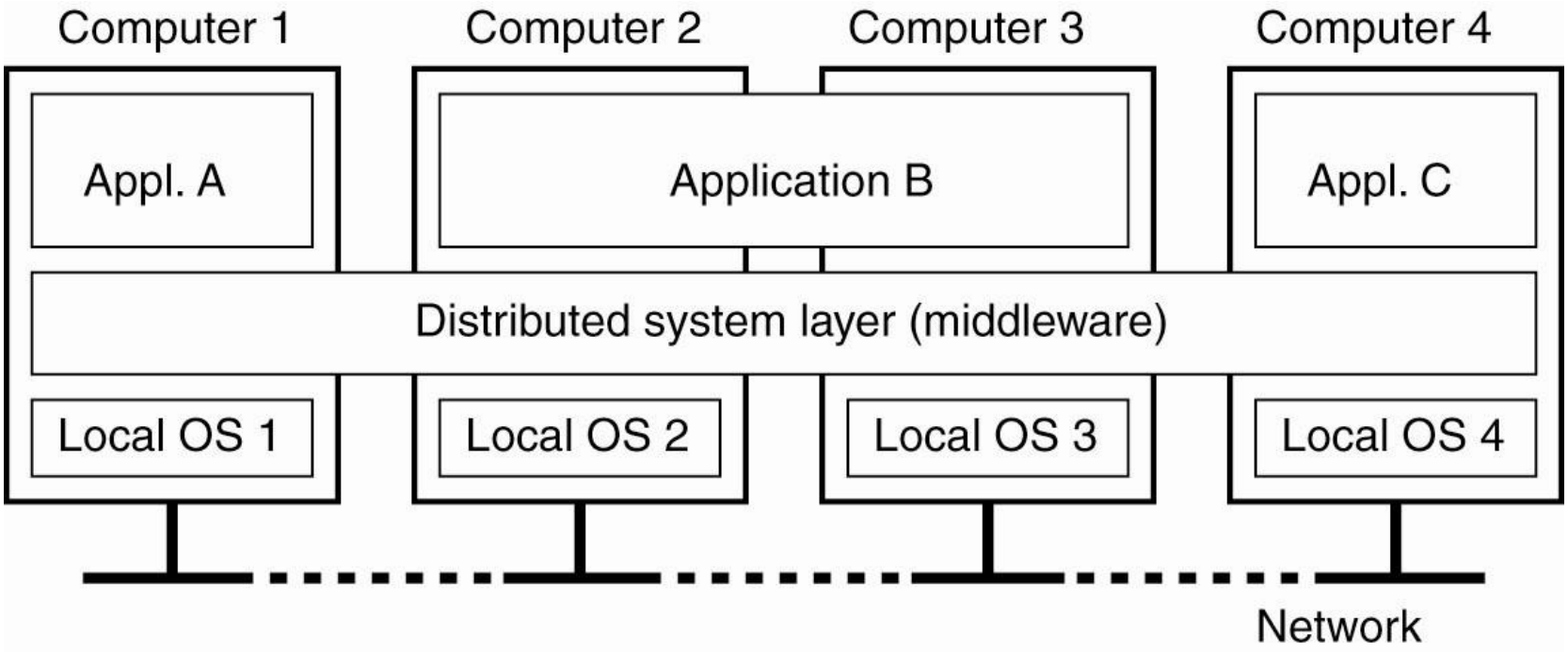
Distributed Systems

- What is a distributed system?

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

-A. Tanenbaum

Distributed Systems



A distributed system organized as middleware. The middleware layer extends over multiple machines, and offers each application the same interface.

Distributed vs. Centralized Systems

- Economics
  - Microprocessors have better price/performance than mainframes
- Speed
  - Collective power of large number of systems
- Geographic and responsibility distribution
- Reliability
  - One machine's failure need not bring down the system
- Extensibility
  - Computers and software can be added incrementally

Disadvantages of Distributed Systems

- Software
  - Little software exists compared to PCs
- Networking
  - Still slow and can cause other problems (e.g. when disconnected)
- Security
  - Data may be accessed by unauthorized users

Concurrency

- In a single system several processes are interleaved
- In distributed systems: there are many systems with one or more processors
  - Many users simultaneously invoke commands or applications
  - Many servers processes run concurrently, each responding to different client request

Scalability

- Scale of system
  - Few PCs servers ->dept level systems->local area networks->internetworked systems->wide are network…
  - Ideally, system and application software should not change as systems scales
- Scalability depends on all aspects
  - Hardware
  - Software
  - networks

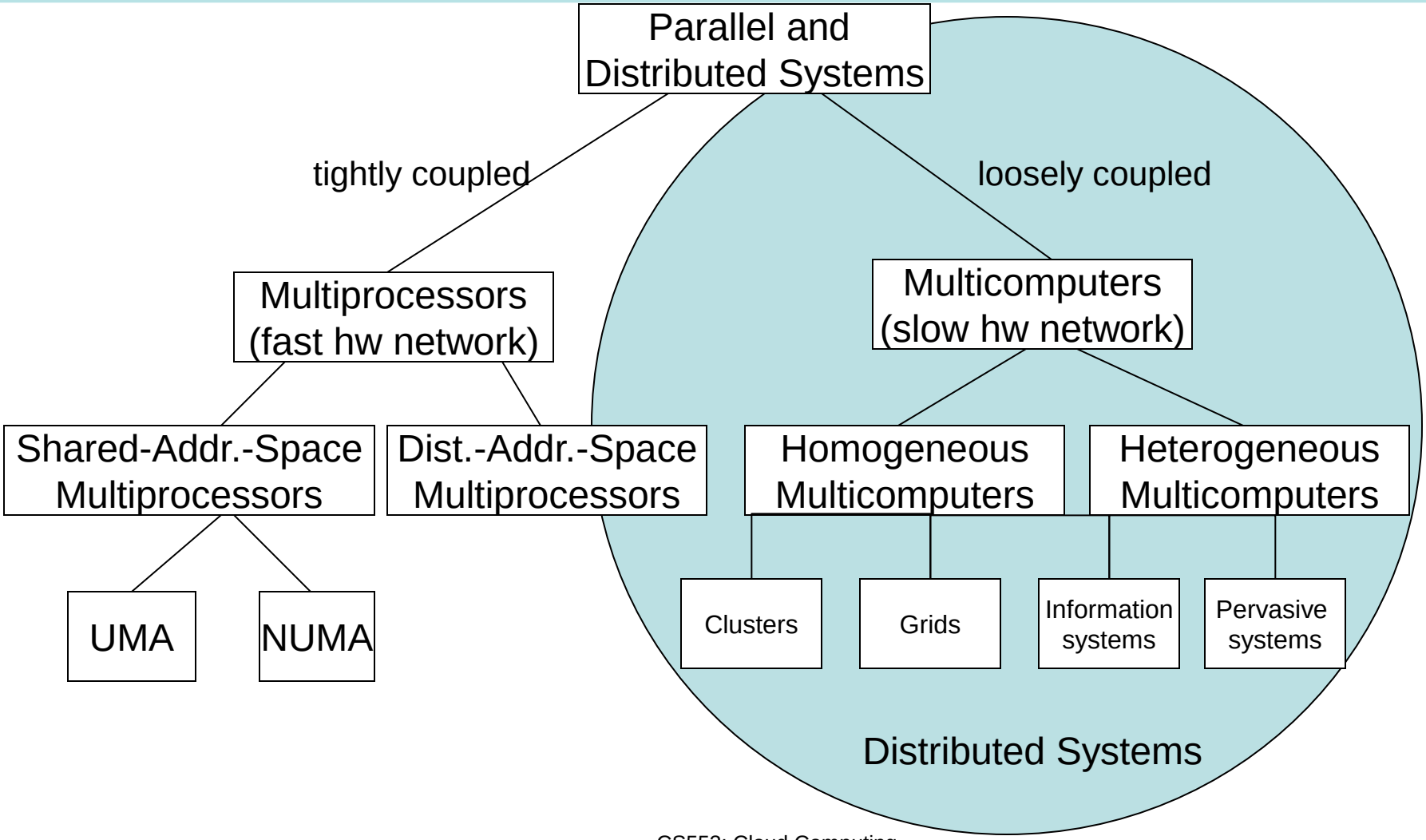
Fault Tolerance

- Definition?
- Two approaches:
  - Hardware redundancy
  - Software recovery
- In distributed systems:
  - Servers can be replicated
  - Databases may be replicated
  - Software recovery involves the design so that state of permanent data can be recovered

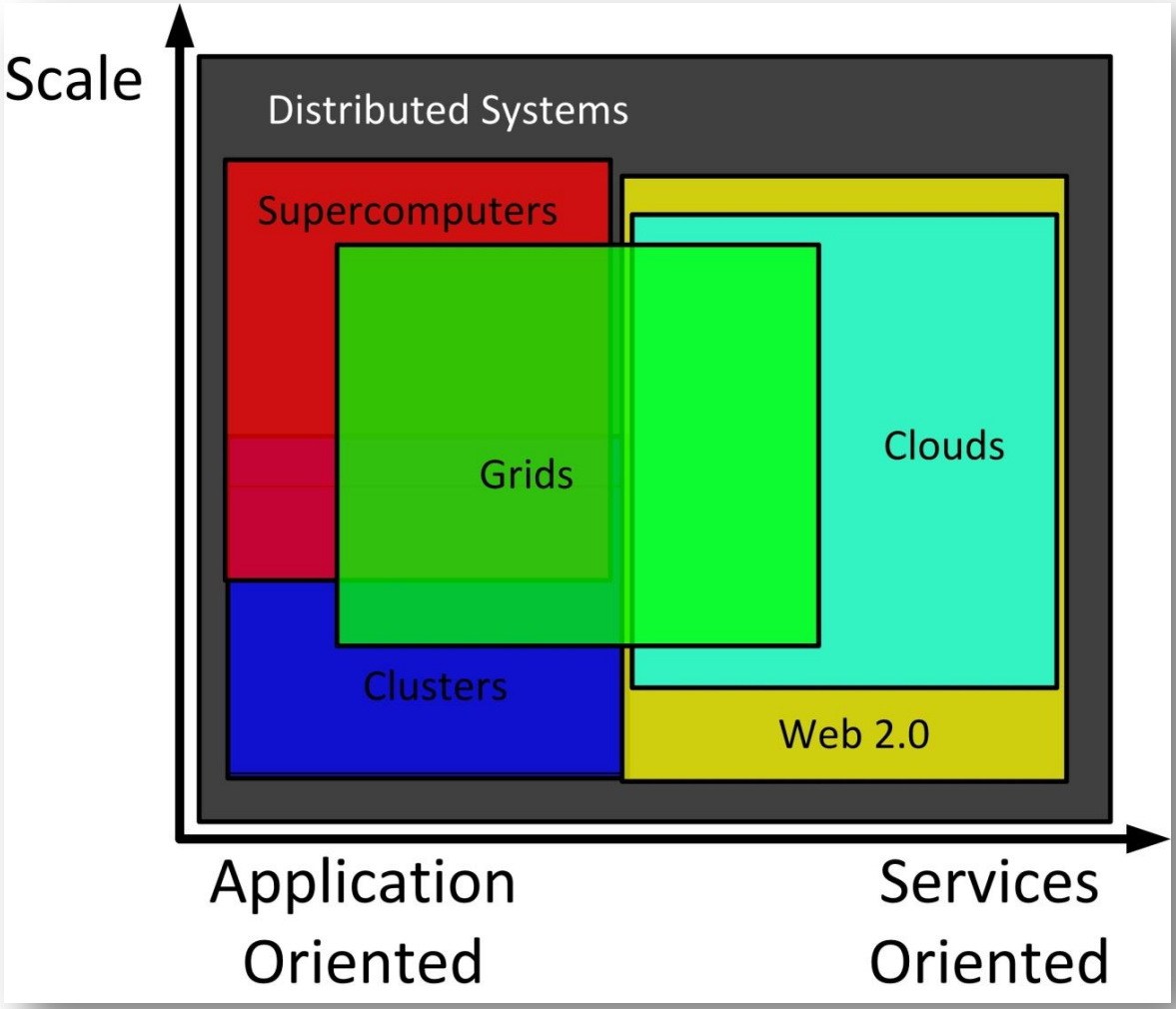
Pitfalls When Developing Distributed Systems

- False assumptions made by first time developer:
  - The network is reliable.
  - The network is secure.
  - The network is homogeneous.
  - The topology does not change.
  - Latency is zero.
  - Bandwidth is infinite.
  - Transport cost is zero.
  - There is one administrator.

Distributed Systems

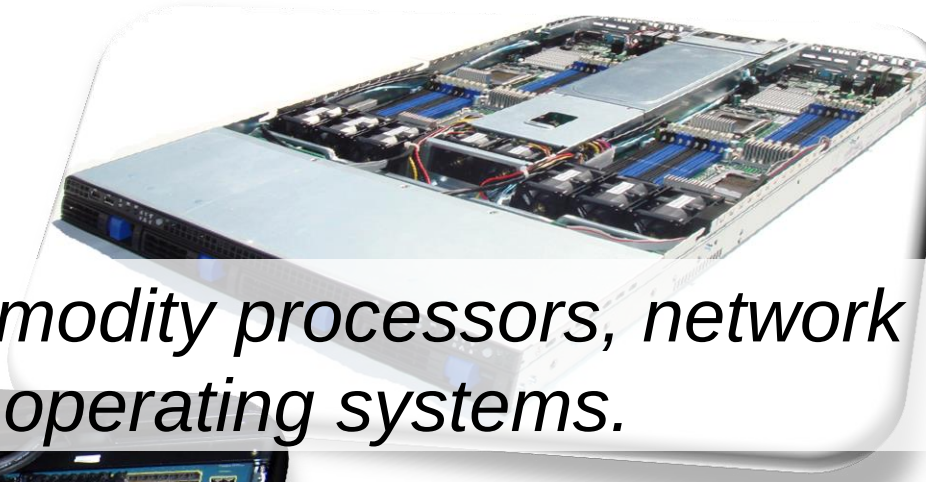


Distributed Systems:
Clusters, Grids, Clouds, and Supercomputers



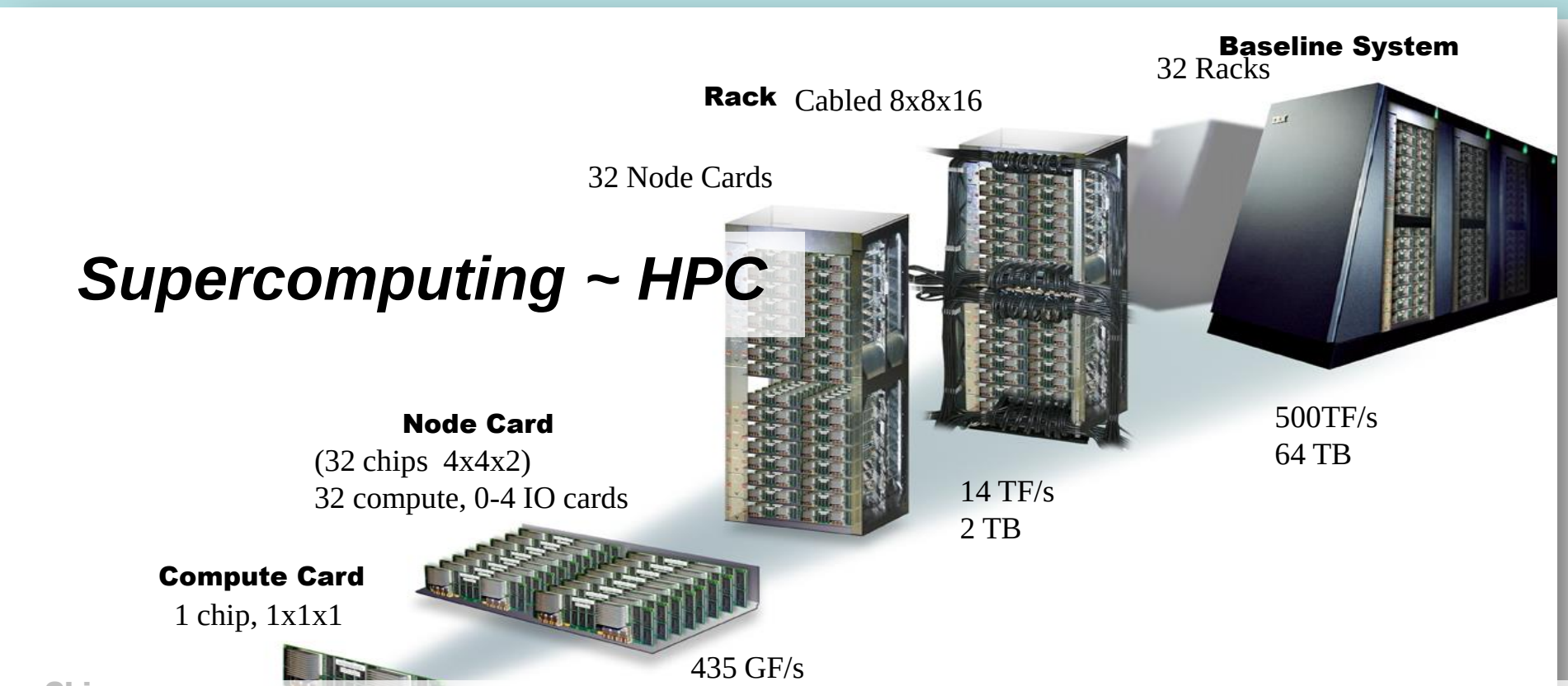
[GCE08] "Cloud Computing and Grid Computing 360-Degree Compared"

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

Top Supercomputers from Top500

- Cray XT4 & XT5
  - Cray #1
  - Cielo #18
  - Hopper #19
- IBM BlueGene/L/P/Q
  - Sequia #2
  - Mira #4
  - Juqueen #5
  - Fermi #9
- GPU based
  - Titan #1
  - Tianhe-1A #8
  - Nebulae #12
- SGI Altix ICE
  - Plaiedas #14
- SPARC64 VIIIfx
  - K #3

Top500 List - November 2012

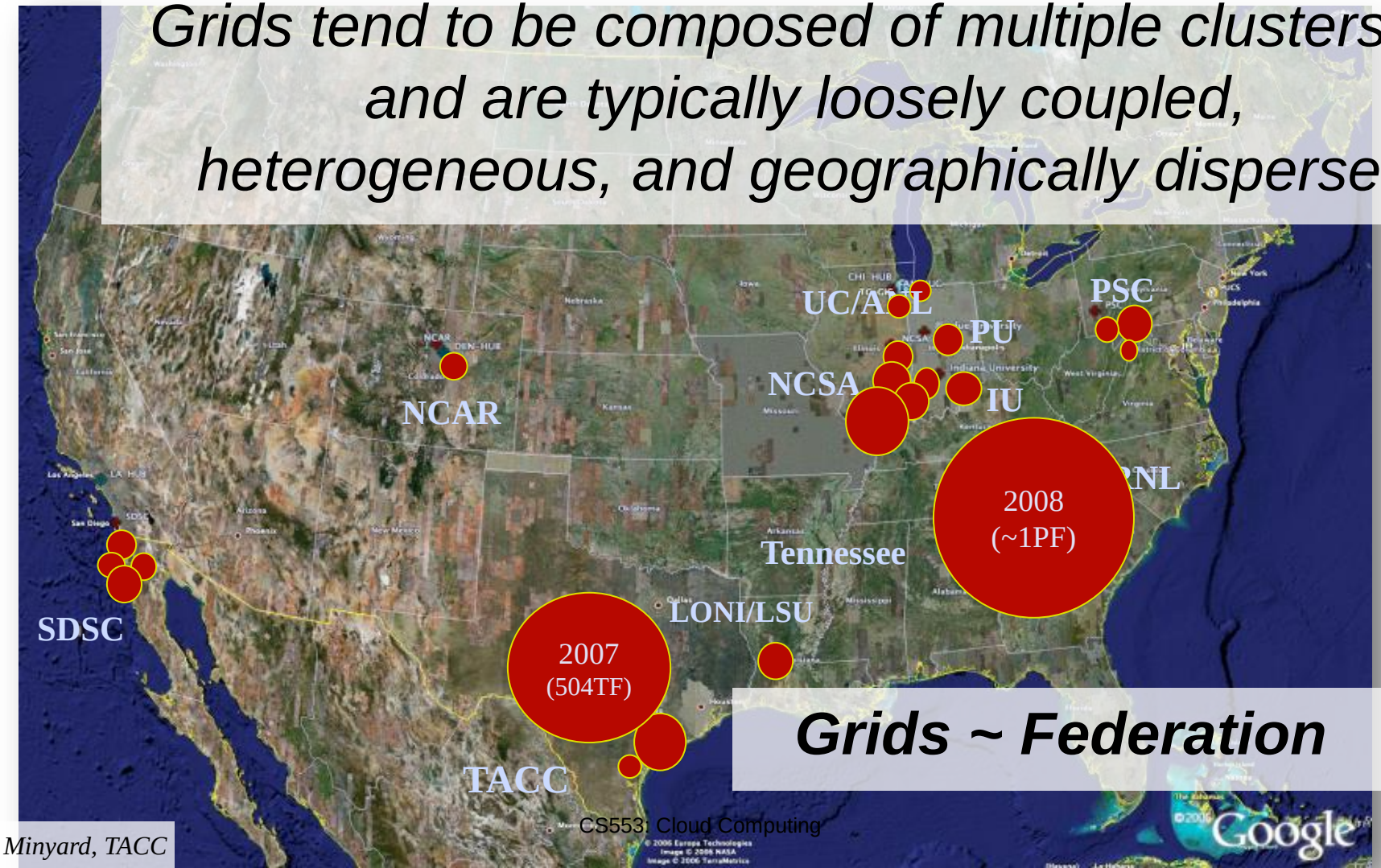
R_{max} and R_{peak} values are in TFlops. For more details about other fields, check the TOP500 description.

previous | 1 | 2 | 3 | 4 | 5 | next

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560640	17590.0	27112.5	8209
2	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1572864	16324.8	20132.7	7890
3	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12660
4	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786432	8162.4	10066.3	3945
5	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393216	4141.2	5033.2	1970

Grid Computing

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



Grids ~ Federation

Tommy Minyard, TACC

Major Grids

- XSEDE (Formerly TeraGrid)
  - 200K-cores across 11 institutions and 22 systems over the US
- Open Science Grid (OSG)
  - 43K-cores across 80 institutions over the US
- Enabling Grids for E-sciencE (EGEE)
- LHC Computing Grid from CERN
- Middleware
  - Globus Toolkit
  - Unicore

Cloud Computing: A Mature Paradigm

Explore trends

Hot searches

Search terms

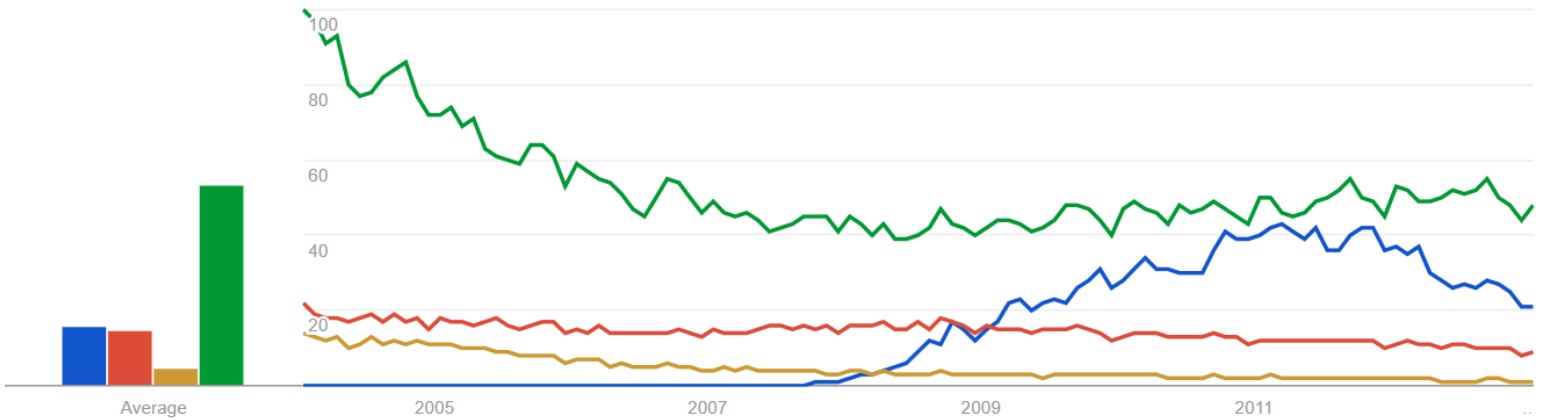
- cloud computing
- hpc
- grid computing
- computer science
- + Add term

Other comparisons

Interest over time

The number 100 represents the peak search volume

News headlines Forecast

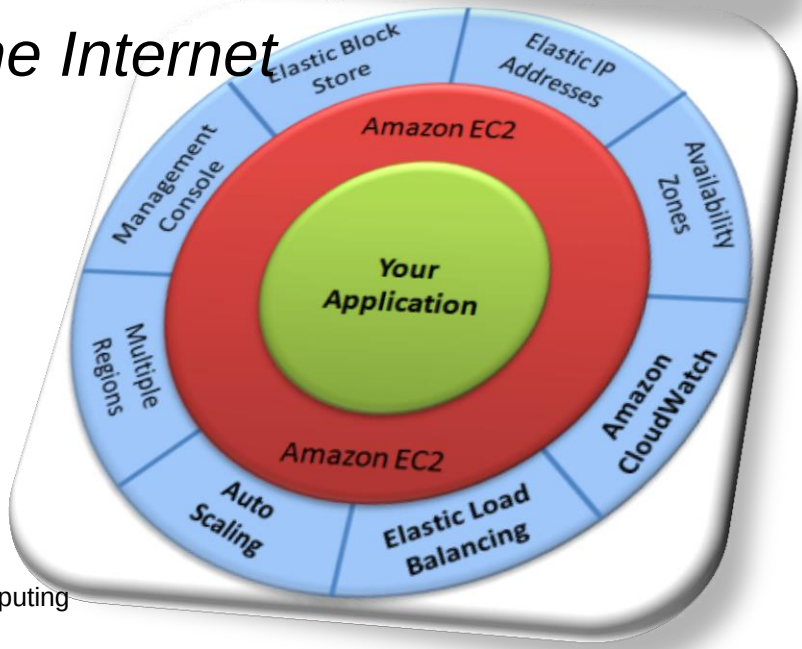


Embed

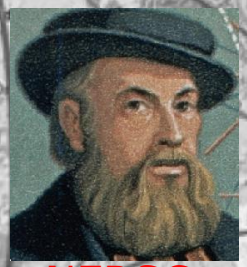
Cloud Computing

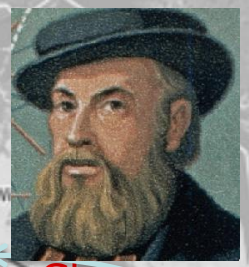
- *A large-scale distributed computing paradigm driven by:*
  1. *economies of scale*
  2. *virtualization*
  3. *dynamically-scalable resources*
  4. *delivered on demand over the Internet*

Clouds ~ hosting



Magellan + DOE's Advanced Network Initiative





NERSC

Chicago

ALCF

NYC

Sunnyvale

Nashville

OLCF

— 100 gigabit/sec

Major Clouds

- Industry
  - Google App Engine
  - Amazon
  - Windows Azure
  - Salesforce
- Academia/Government
  - Magellan
  - FutureGrid
- Opensource middleware
  - Nimbus
  - Eucalyptus
  - OpenNebula
  - OpenStack
  - CloudStack

Questions