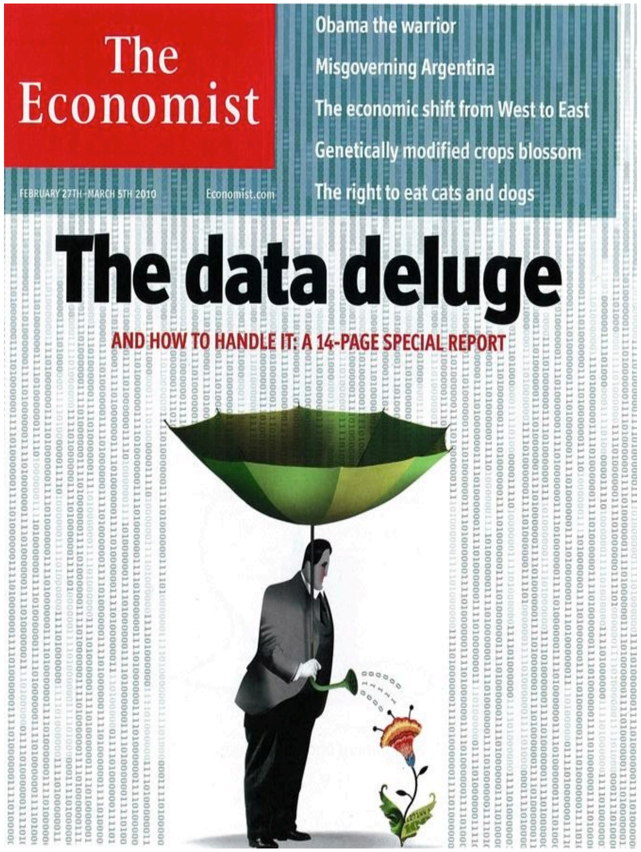
CSE 710 Seminar

Wide Area Distributed File Systems

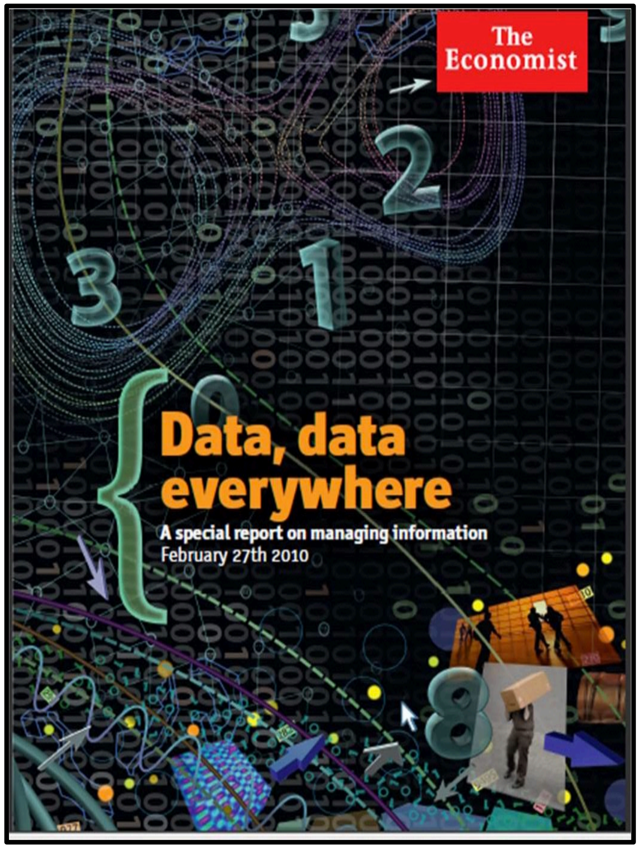
Tevfik Kosar, Ph.D.

Week 1: January 16, 2013

Data Deluge


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

Science
11 February 2011 | $10
scientific
information
climate
research
science
data
researchers
analysis
visualization
access
sharing
challenges
AAAS

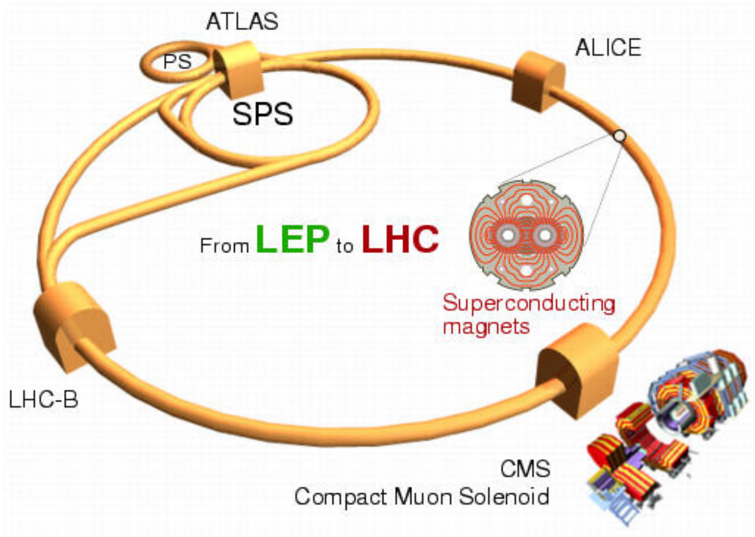

The Economist
Data, data everywhere
A special report on managing information
February 27th 2010

Big Data in Science

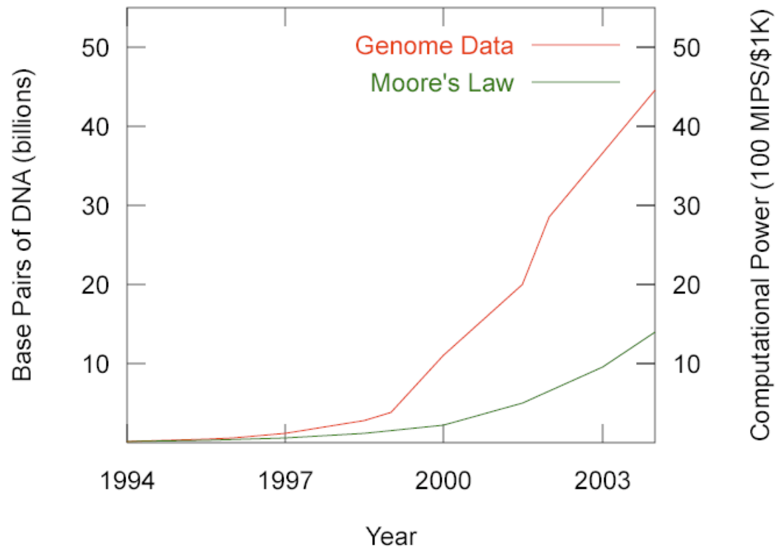
Demand for data in all areas of science!

Application	Area	Data Volume
VISTA	Astronomy	100 TB/year
LIGO	Astrophysics	250 TB/year
WCER EVP	Educational Technology	500 TB/year
LSST	Astronomy	1000 TB/year
BLAST	Bioinformatics	1000 TB/year
ATLAS/CMS	High Energy Physics	5000 TB/year

The Large Hadron Collider (LHC)



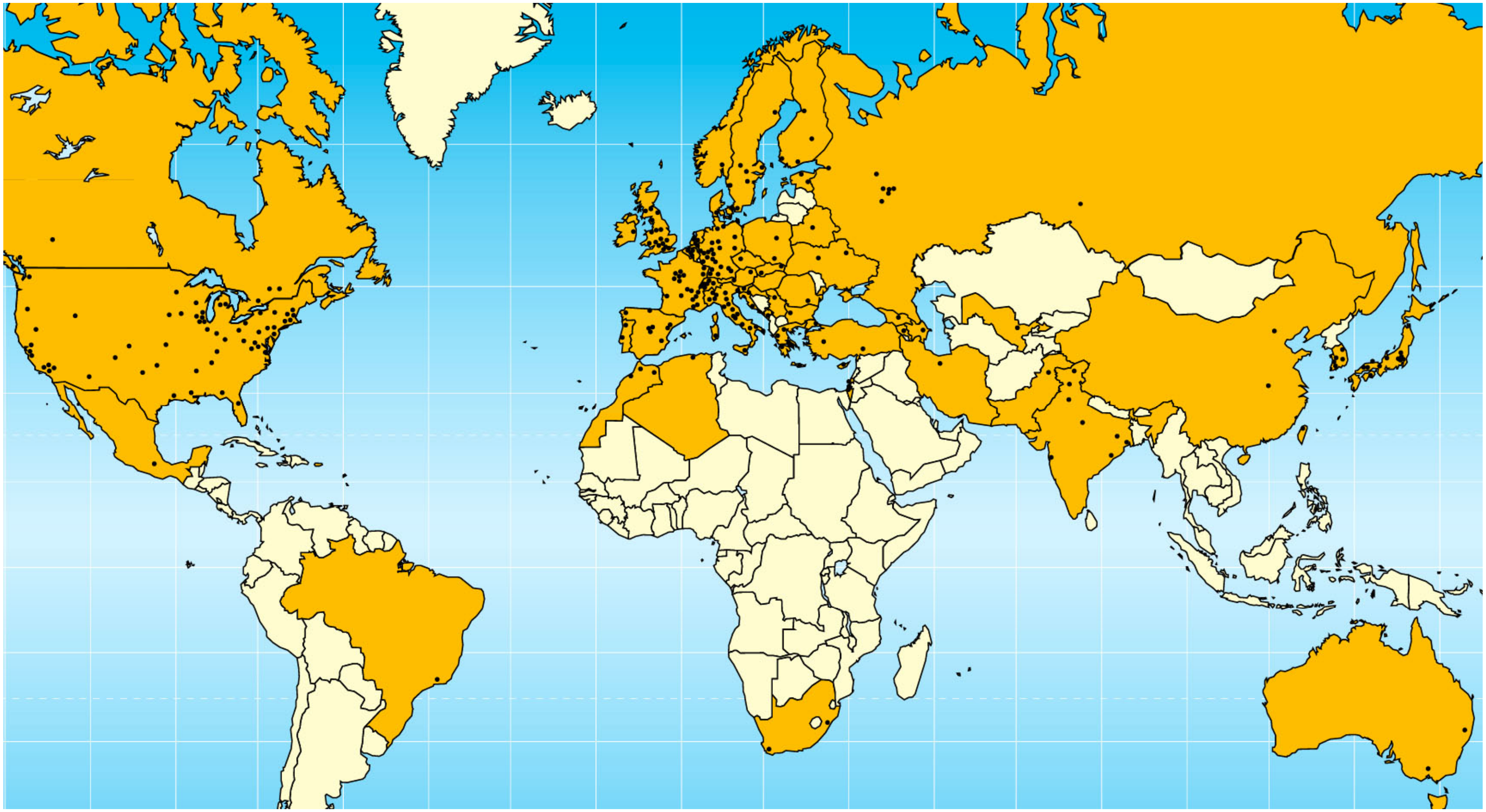
Scientific data outpaced Moore's Law!



Demand for data brings demand for computational power:

ATLAS and CMS applications alone require more than 100,000 CPUs!

ATLAS Participating Sites



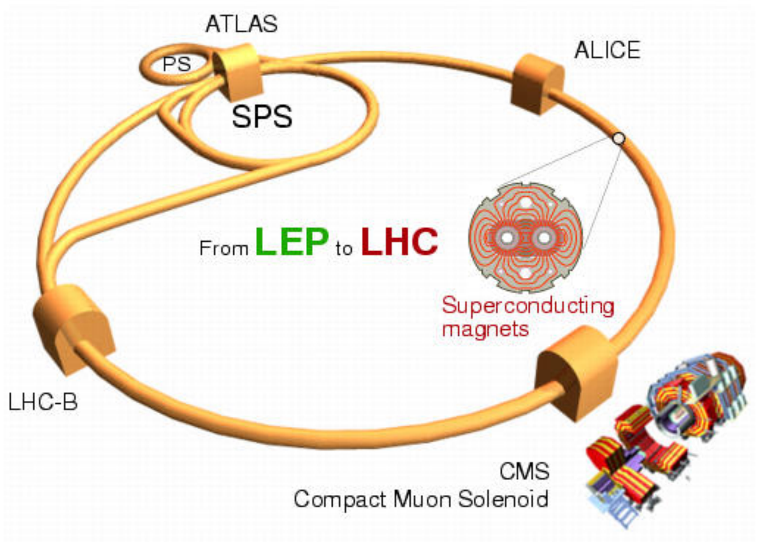
ATLAS: High Energy Physics project

Generates ***10 PB*** *data/year --> distributed to and processed by 1000s of researchers at* ***200 institutions*** *in* ***50 countries****.*

Big Data Everywhere

Science

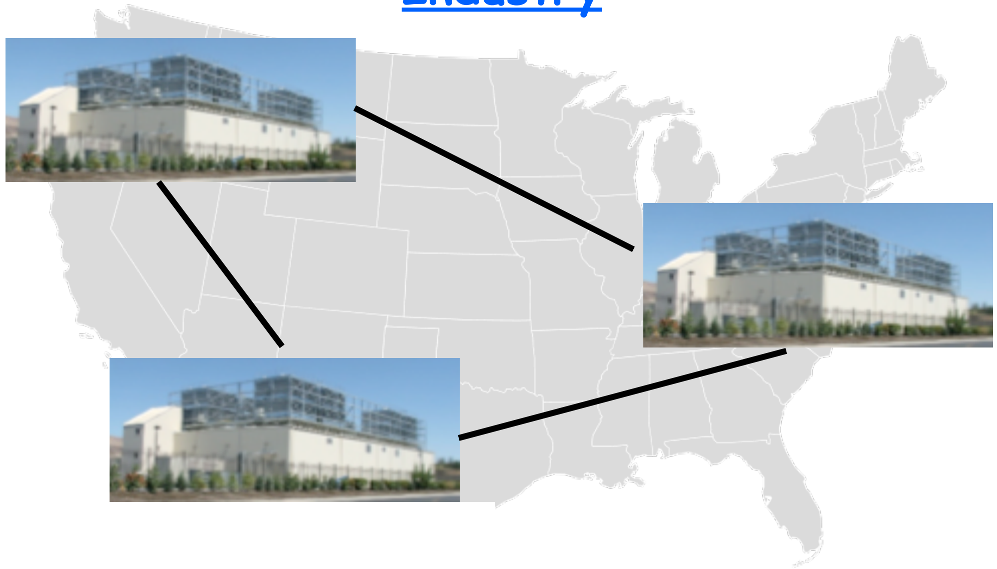
The Large Hadron Collider (LHC)



- 1 PB is now considered "small" for many science applications today

- For most, their data is distributed across several sites

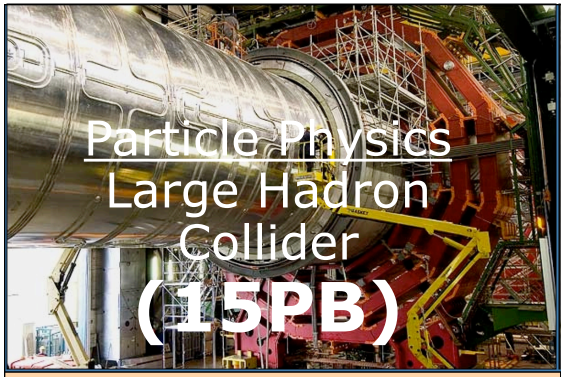
Industry



A survey among 106 organizations operating two or more data centers:

- 50% has more than 1 PB in their primary data center

- 77% run replication among three or more sites



Particle Physics
Large Hadron Collider
(15PB)

Human Genomics
(7000PB)
1GB / person
200PB+ captured

World Wide Web
(10PB)

Wikipedia
400K Articles/ Year

Annual Email Traffic, no spam
(300PB+)

Internet Archive
(1PB+)

Estimated On-line RAM in Google
(8PB)

Personal Digital Photos
(1000PB+)

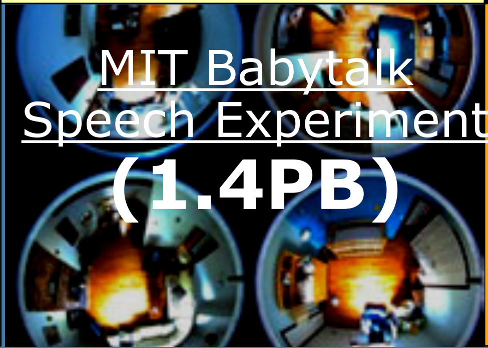
200 of London's Traffic Cams
(8TB/day)

Walmart Transaction DB
(500TB)

Typical Oil Company
(350TB+)

Merck Bio Research DB
(1.5TB/qtr)

UPMC Hospitals Imaging Data
(500TB/yr)



MIT Babytalk Speech Experiment
(1.4PB)

Terashake Earthquake Model of LA Basin
(1PB)

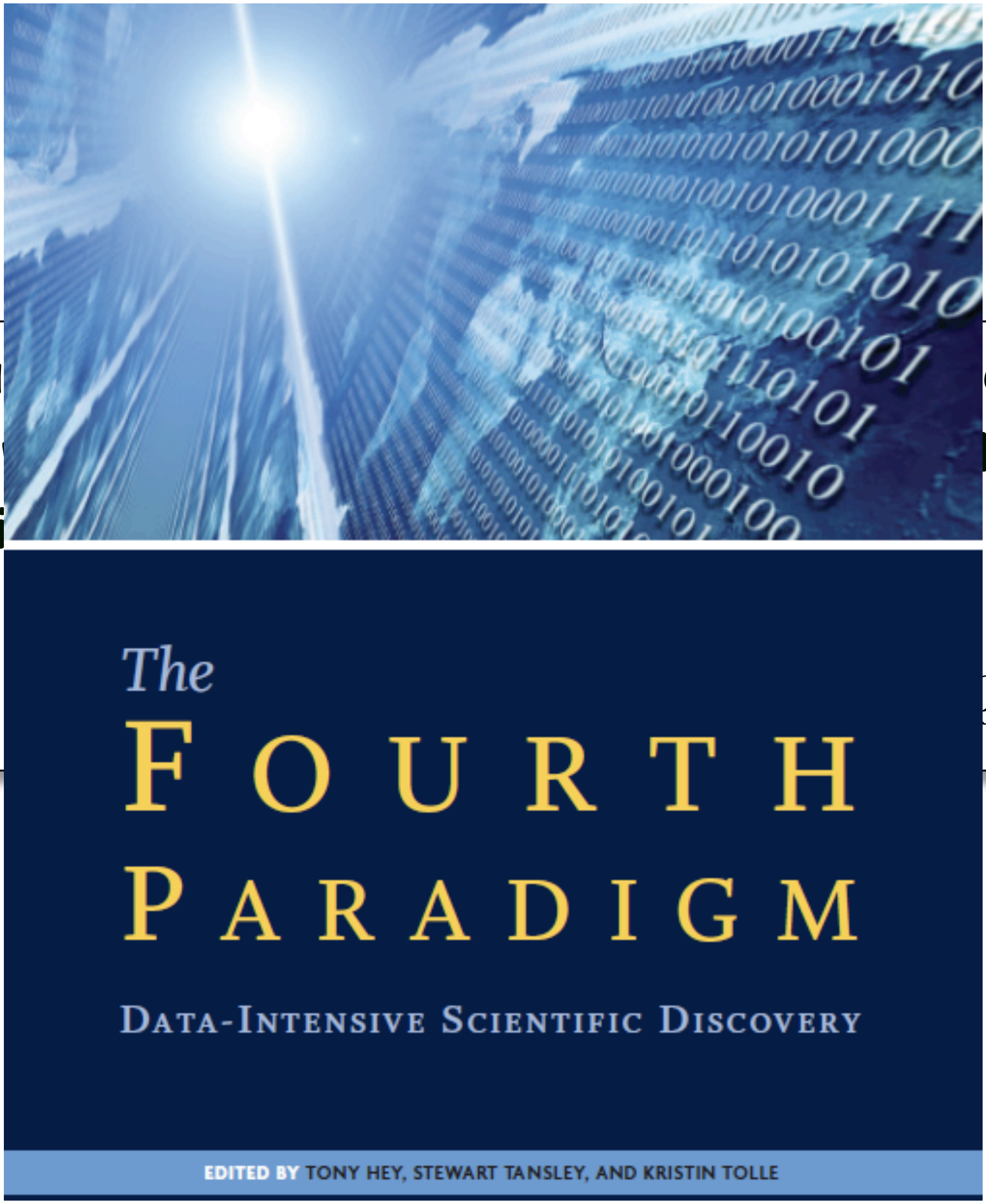
One Day of Instant Messaging
(1TB)

Total digital data to be created this year **270,000PB** **(IDC)**

Future Trends

“In the futu ence and
engineering **ty to**
leverage thi ed in digital
form.”

yberinfrastructure



The

FOURTH PARADIGM

DATA-INTENSIVE SCIENTIFIC DISCOVERY

EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

Emergence of a **Fourth Research Paradigm**

Thousand years ago – **Experimental Science**

- Description of natural phenomena

Last few hundred years – **Theoretical Science**

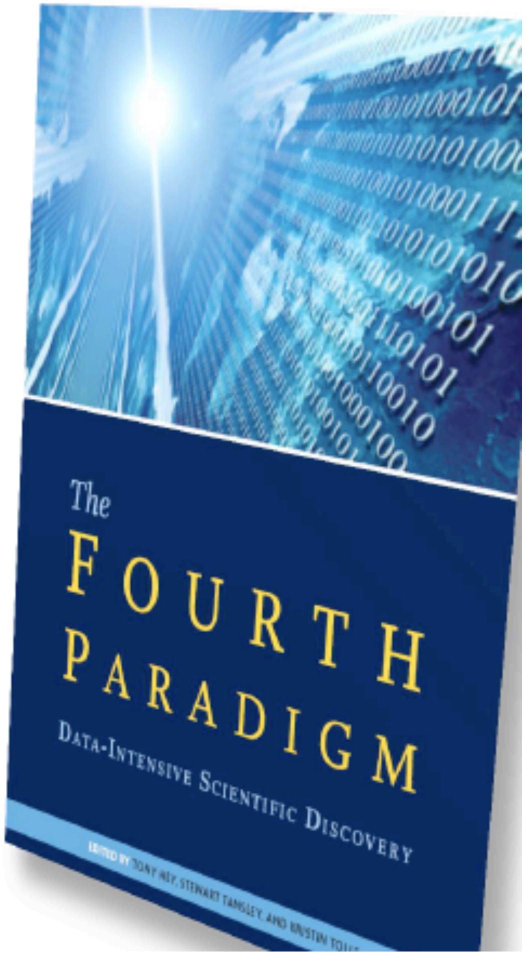
- Newton's Laws, Maxwell's Equations…

Last few decades – **Computational Science**

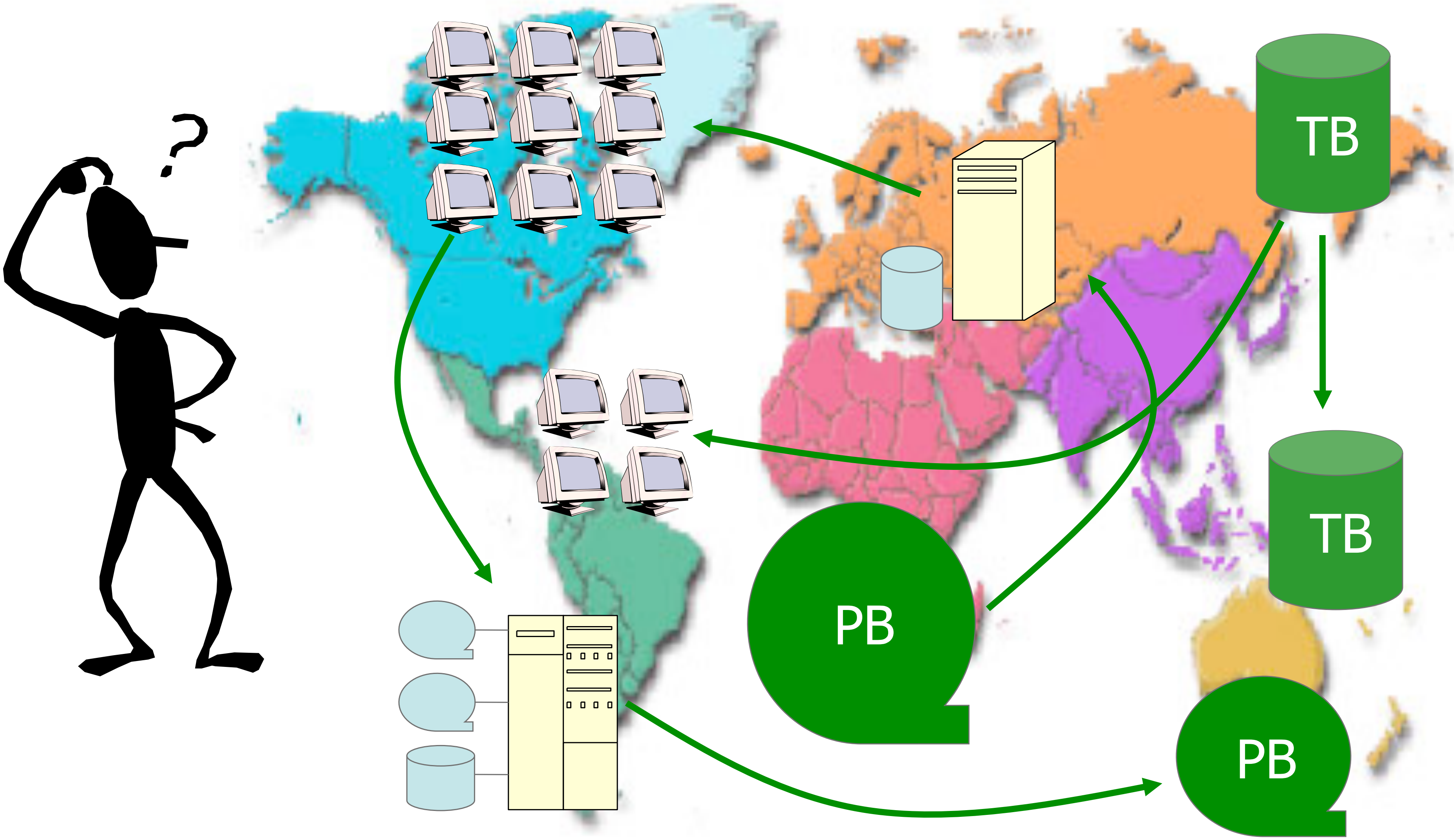
- Simulation of complex phenomena

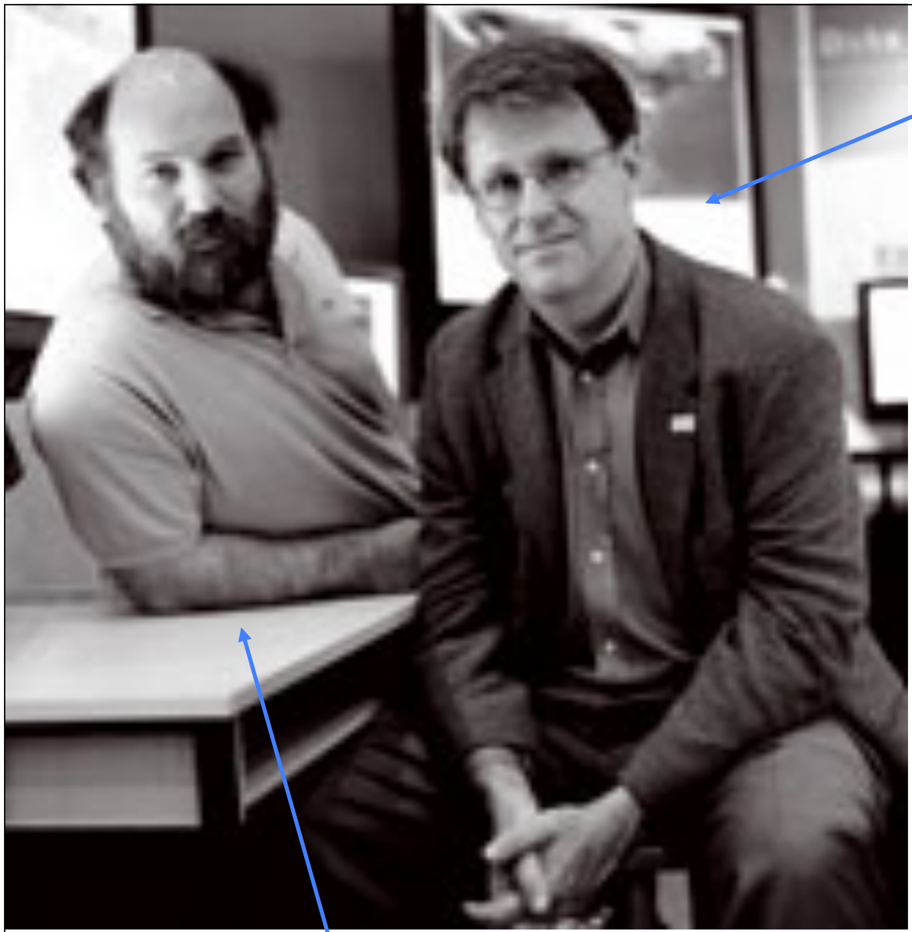
Today – **Data-Intensive Science**

- Large-scale data analysis and data mining; visualization and exploration; scholarly communication and dissemination


The
FOURTH
PARADIGM
DATA-INTENSIVE SCIENTIFIC DISCOVERY
EDITED BY TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE

How to Access and Process Distributed Data?



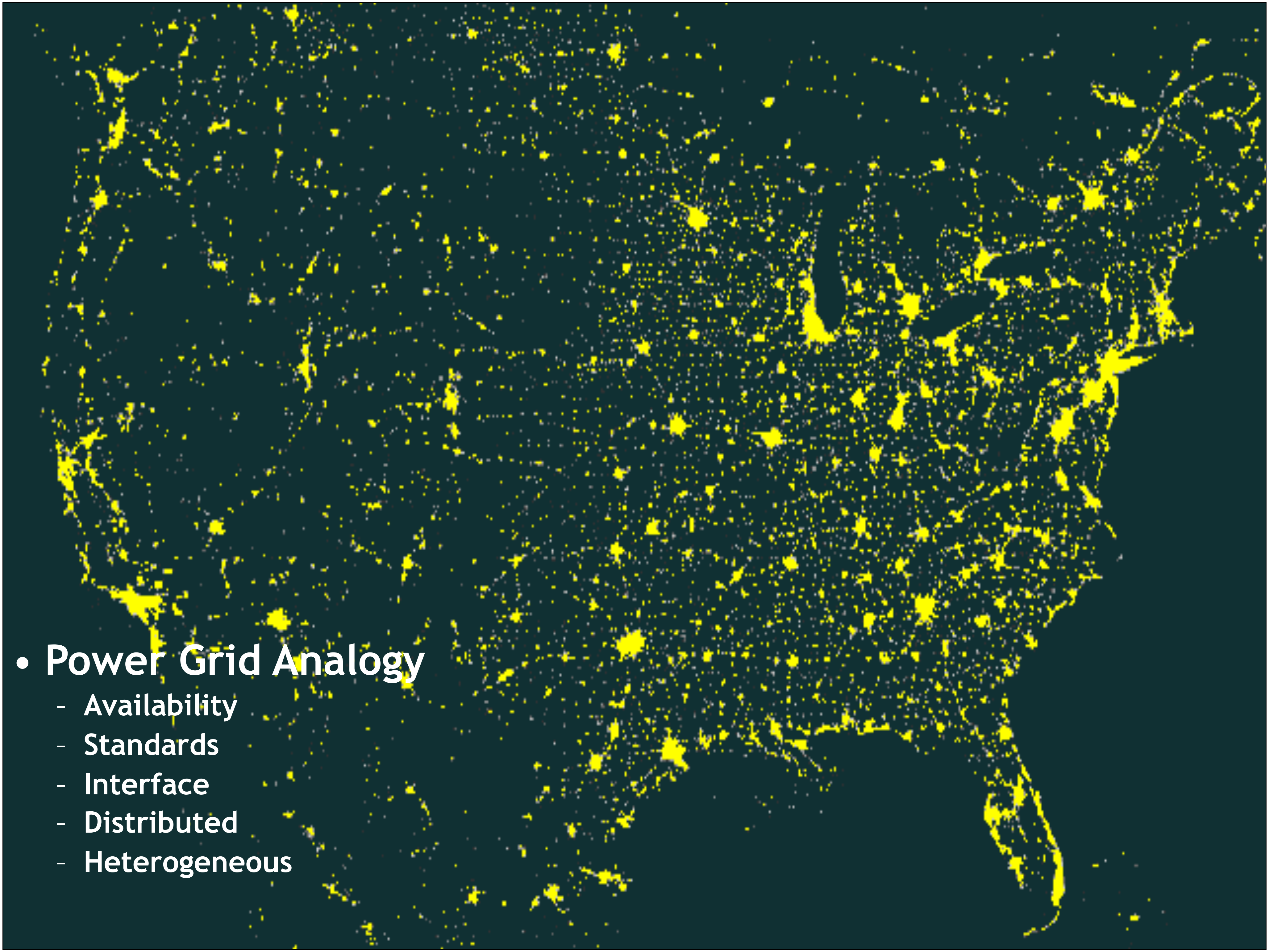


Ian Foster
UChicago/Argonne

Carl Kesselman
ISI/USC

They have coined the term "Grid Computing" in 1996!

In 2002, "Grid Computing" selected one of the Top 10 Emerging Technologies that will change the world!



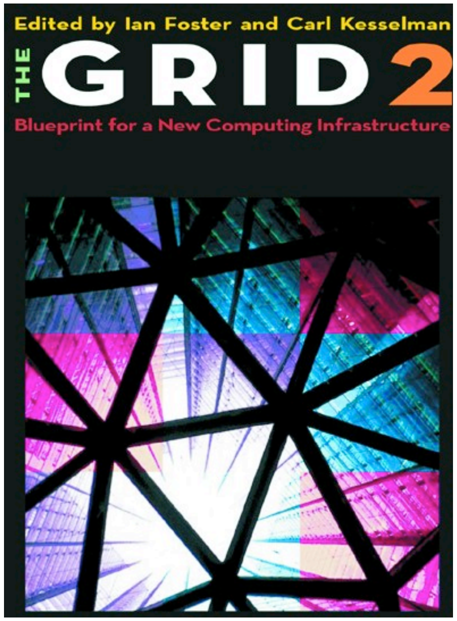
- Power Grid Analogy
  - Availability
  - Standards
  - Interface
  - Distributed
  - Heterogeneous

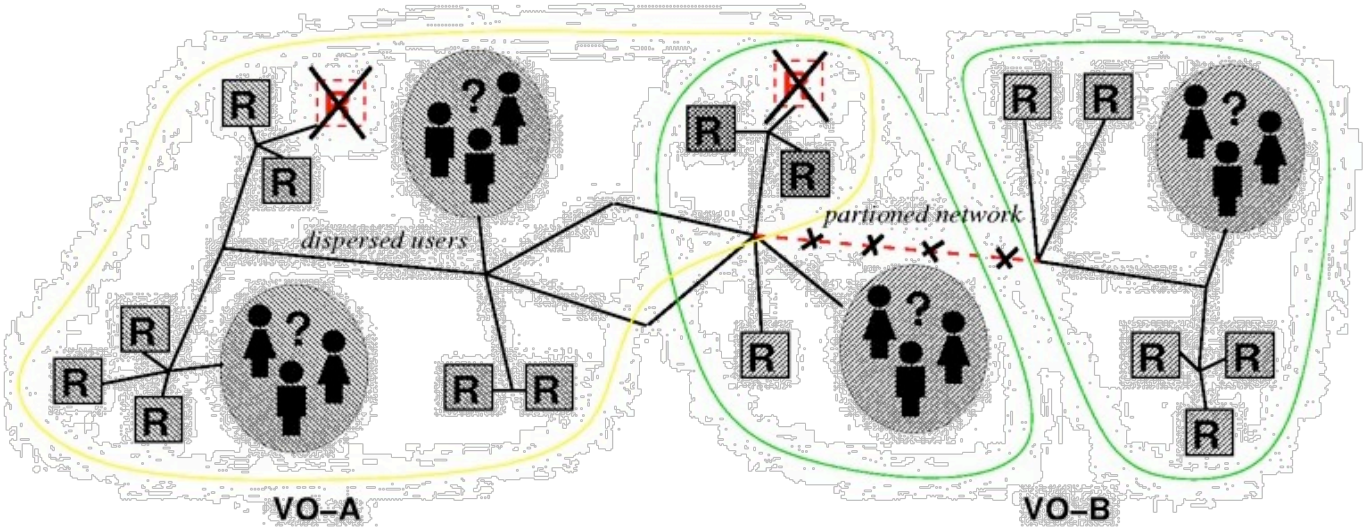
Defining Grid Computing

- There are several competing definitions for “The Grid” and Grid computing
- These definitions tend to focus on:
  - Implementation of Distributed computing
  - A common set of interfaces, tools and APIs
  - inter-institutional, spanning multiple administrative domains
  - “The Virtualization of Resources” abstraction of resources

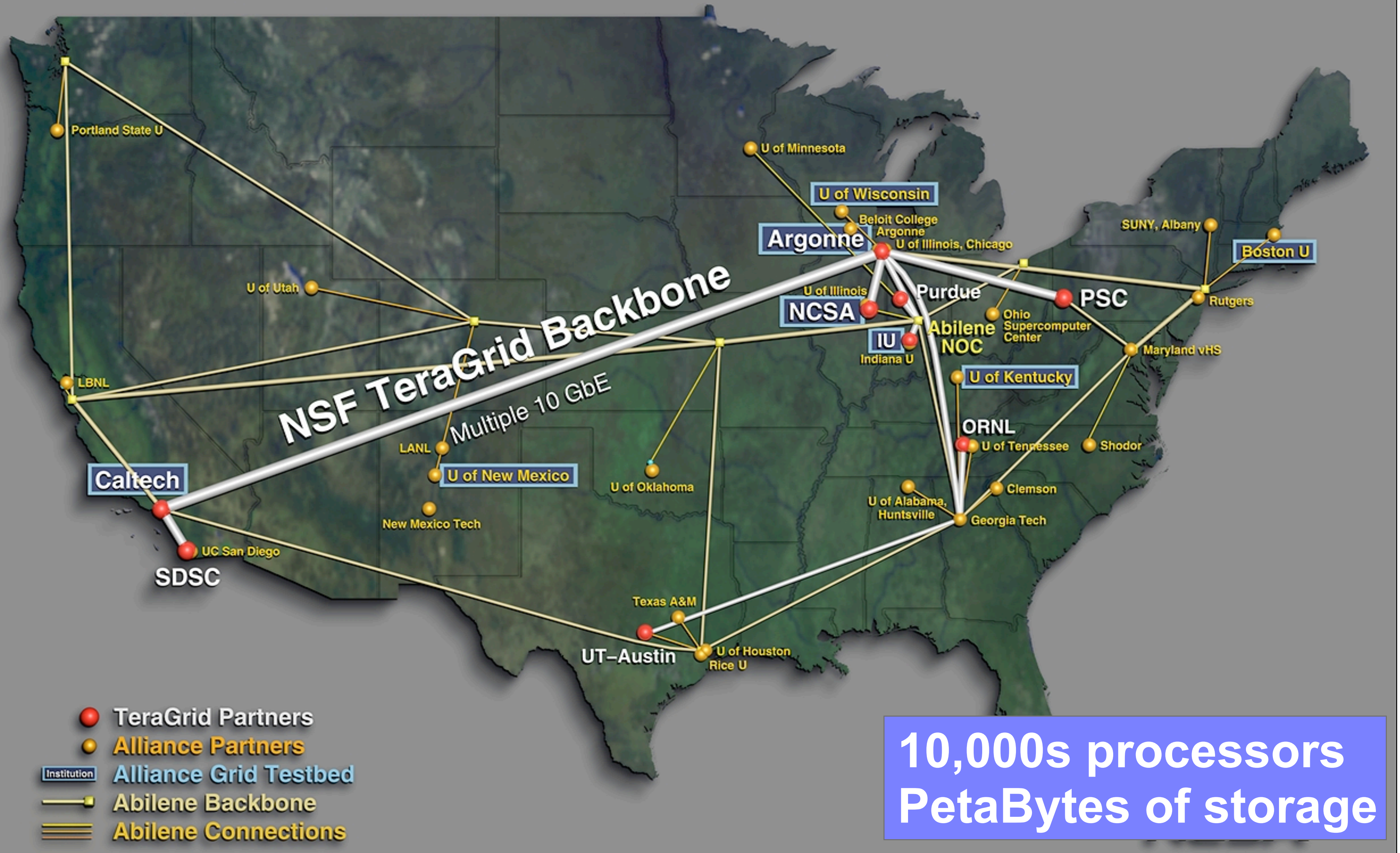
According to Foster & Kesselman:

"coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations" (The Anatomy of the Grid, 2001)


Edited by Ian Foster and Carl Kesselman
THE GRID 2
Blueprint for a New Computing Infrastructure



TeraGrid and the Alliance



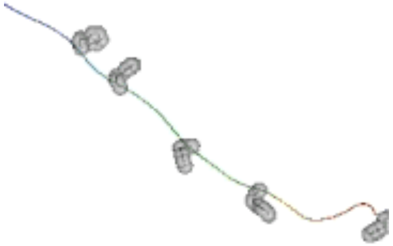
10,000s processors
PetaBytes of storage

Desktop Grids

SETI@home:

- Detect any alien signals received through Arecibo radio telescope
- Uses the idle cycles of computers to analyze the data generated from the telescope

Others: Folding@home, FightAids@home



- Over 2,000,000 active participants, most of whom run screensaver on home PC
- Over a cumulative 20 TeraFlop/sec
  - TeraGrid: 40 TeraFlop/src
- Cost: $700K!!
  - TeraGrid: > $100M

Emergence of Cloud Computing

Grid Computing

- Solving large problems with parallel computing
- Made mainstream by Globus Alliance

Utility Computing

- Offering computing resources as a metered service
- Introduced in late 1990s

Software as a Service

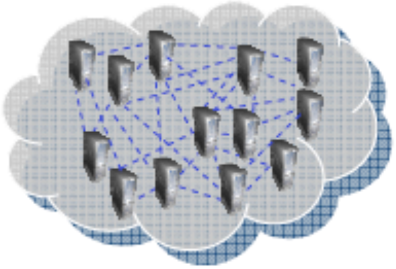
Network-based subscriptions to applications

Gained momentum in 2001

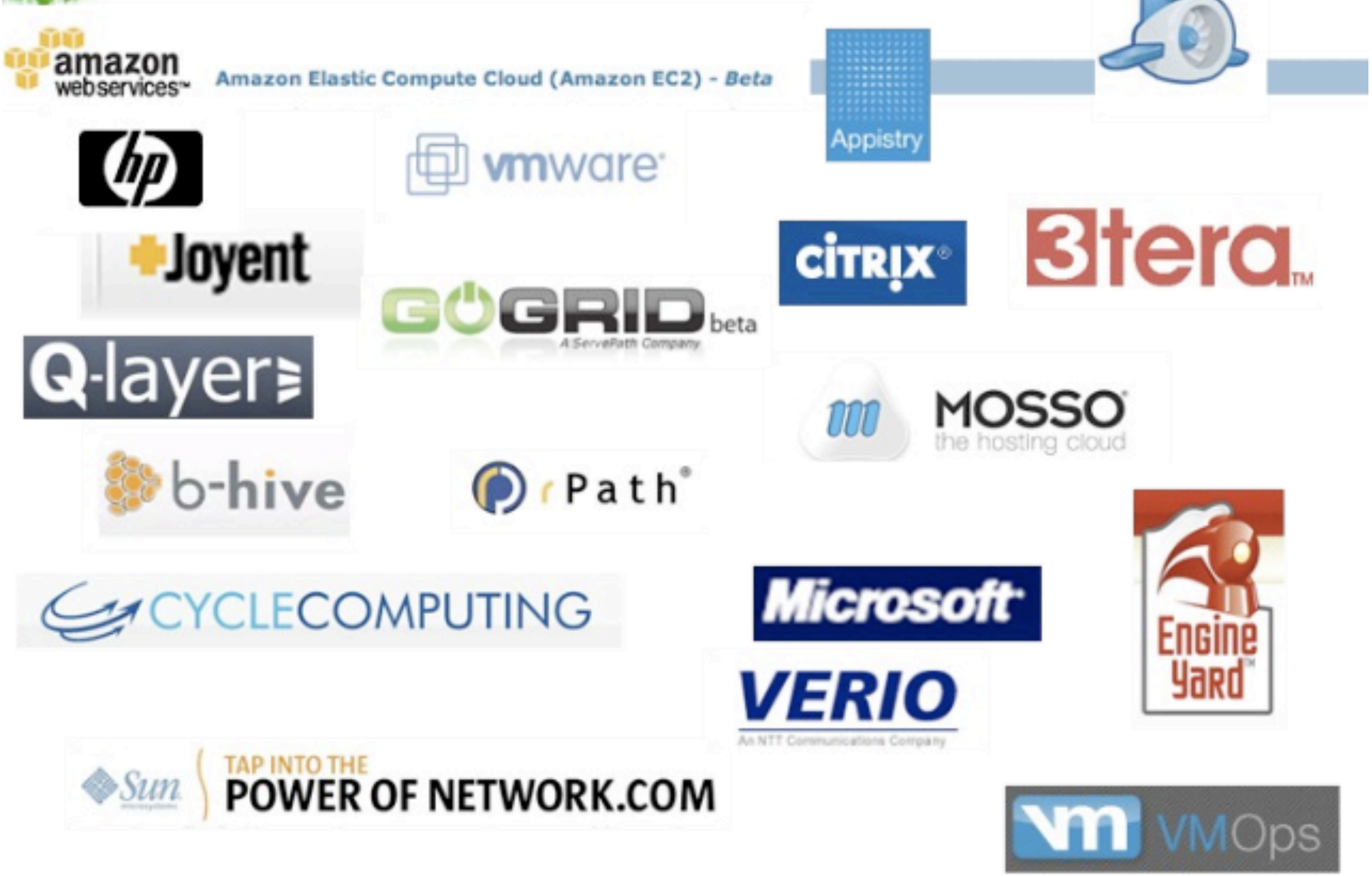
Cloud Computing

Next-Generation Internet computing

Next-Generation Data Centers



Commercial clouds

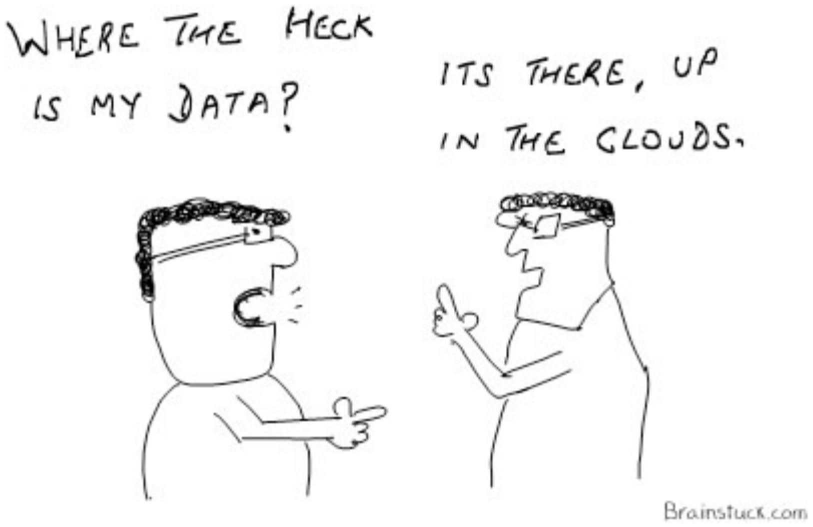


Commercial Clouds Growing...

- Microsoft [NYTimes, 2008]
  - 150,000 machines
  - Growth rate of 10,000 per month
  - Largest datacenter: 48,000 machines
  - 80,000 total running Bing
- Yahoo! [Hadoop Summit, 2009]
  - 25,000 machines
  - Split into clusters of 4000
- AWS EC2 (Oct 2009)
  - 40,000 machines
  - 8 cores/machine
- Google
  - (Rumored) several hundreds of thousands of machines

Distributed File Systems

- Data sharing of multiple users
- User mobility
- Data location transparency
- Data location independence
- Replications and increased availability



- Not all DFS are the same:
  - Local-area vs Wide area DFS
  - Fully Distributed FS vs DFS requiring central coordinator

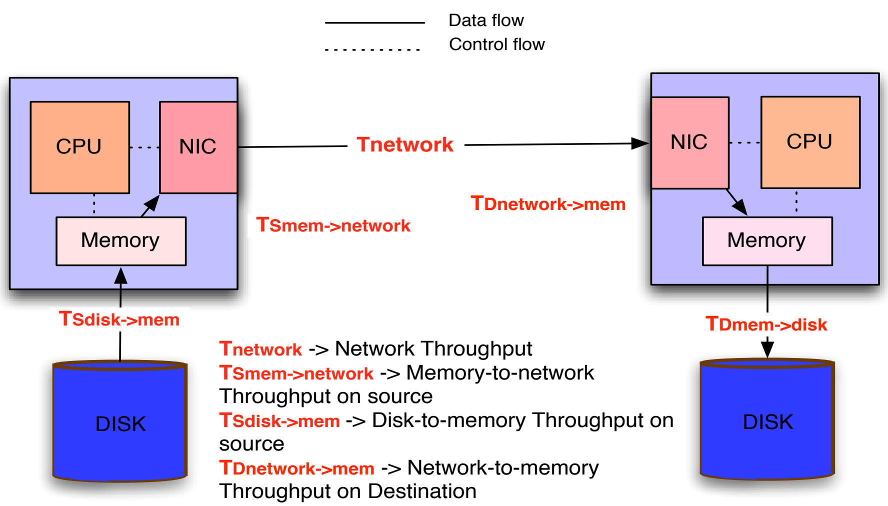
Issues in Distributed File Systems

- **Naming (global name space)**
- **Performance (Caching, data access)**
- **Consistency (when/how to update/synch?)**
- Reliability (replication, recovery)
- Security (user privacy, access controls)
- Virtualization

Moving Big Data across WAFS?

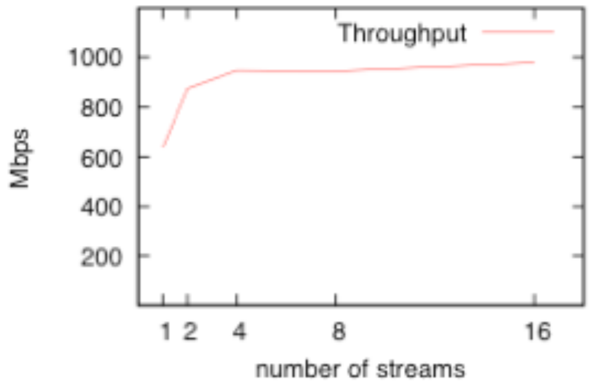
- Sending **1 PB** of data over 10 Gbps link would take **nine days** (assuming 100% efficiency) -- too optimistic!
- Sending **1 TB** Forensics dataset from Boston to Amazon S3 cost $100 and took **several weeks** [Garfinkel 2007]
- Visualization scientists at LANL dumping data to tapes and sending them to Sandia Lab via **Fedex** [Feng 2003]
- Collaborators have the option of moving their data into disks, and sending them as packages through **UPS or FedEx** [Cho et al 2011].
- Will **100 Gbps** networks change anything?

End-to-end Problem

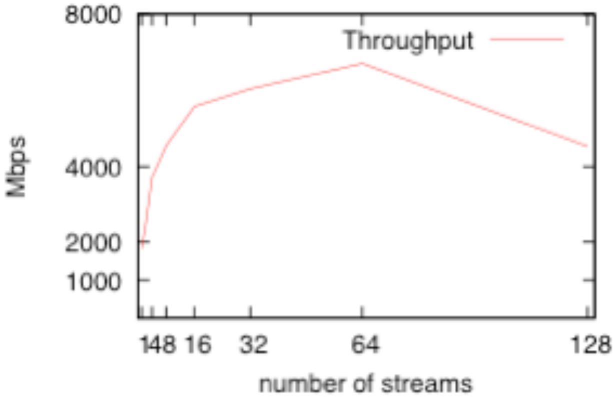


Tnetwork -> Network Throughput
TSmem->network -> Memory-to-network Throughput on source
TSdisk->mem -> Disk-to-memory Throughput on source
TDnetwork->mem -> Network-to-memory Throughput on Destination
TDmem->disk -> Memory-to-disk Throughput on destination

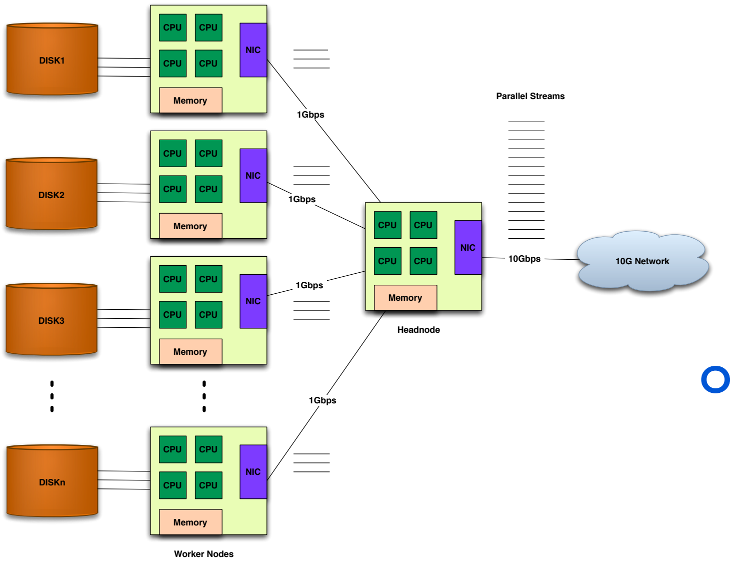
protocol tuning



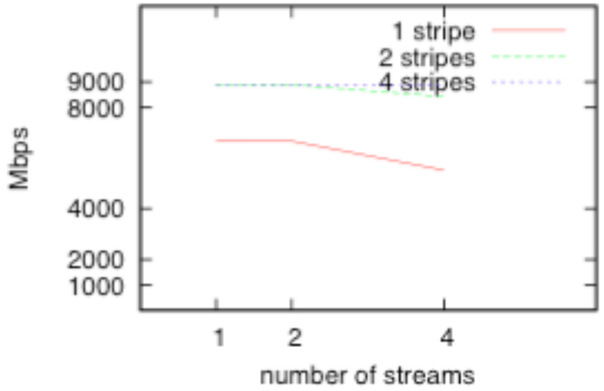
disk I/O optimization



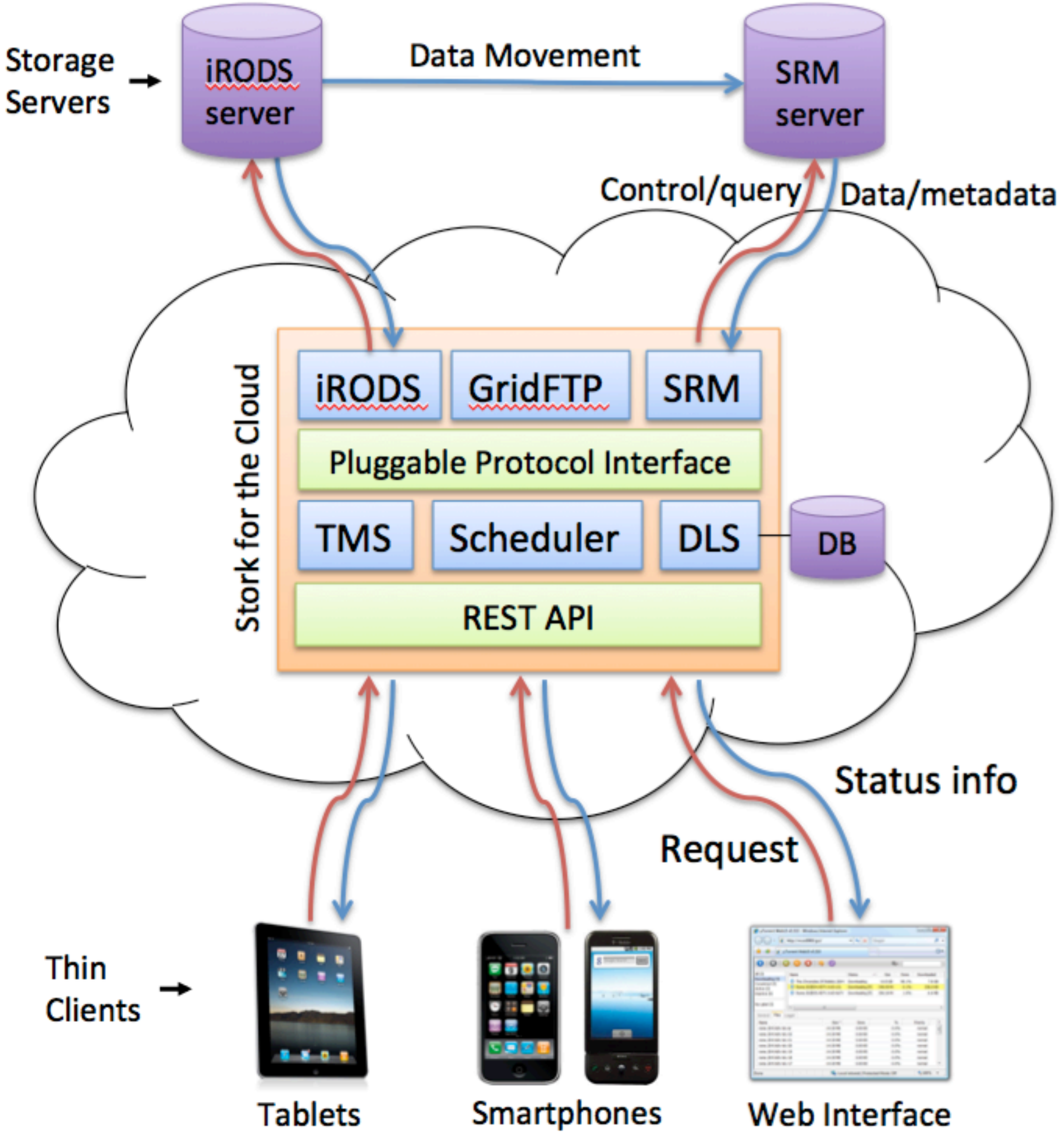
Parameters to be optimized:
- # of streams
- # of disk stripes
- # of CPUs/nodes



CPU optimization



Cloud-hosted Transfer Optimization



CSE 710 Seminar

- State-of-the-art research, development, and deployment efforts in wide-area distributed file systems on clustered, grid, and cloud infrastructures.

- We will review 21 papers on topics such as:
  - File System Design Decisions
  - Performance, Scalability, and Consistency issues in File Systems
  - Traditional Distributed File Systems
  - Parallel Cluster File Systems
  - Wide Area Distributed File Systems
  - Cloud File Systems
  - Commercial vs Open Source File System Solutions

CSE 710 Seminar *(cont.)*

- Early Distributed File Systems
  - NFS (Sun)
  - AFS (CMU)
  - Coda (CMU)
  - xFS (UC Berkeley)

- Parallel Cluster File Systems
  - GPFS (IBM)
  - Panasas (CMU/Panasas)
  - PVFS (Clemson/Argonne)
  - Lustre (Cluster Inc)
  - Nache (IBM)
  - Panache (IBM)

CSE 710 Seminar *(cont.)*

- Wide Area File Systems
  - OceanStore (UC Berkeley)
  - Ivy (MIT)
  - WheelFS (MIT)
  - Shark (NYU)
  - Ceph (UC-Santa Cruz)
  - Giga+ (CMU)
  - BlueSky (UC-San Diego)

  - Google FS (Google)
  - Hadoop DFS (Yahoo!)
  - Farsite (Microsoft)
  - zFS (IBM)

Reading List

- The list of papers to be discussed is available at:

http://www.cse.buffalo.edu/faculty/tkosar/cse710_spring13/reading_list.htm

- Each student will be responsible for:
  - Presenting 1 paper
  - Reading and contributing the discussion of all the other papers (ask questions, make comments etc)
- We will be discussing 2 papers each class

Paper Presentations

- Each student will present 1 paper:

- 25-30 minutes each + 20-25 minutes Q&A/discussion
- No more than 10 slides

- Presenters should meet with me on Tuesday before their presentation to show their slides!

- Office hours: Tue 10:00am - 12:00pm

Participation

- Post at least one question to the seminar blog by Tuesday night before the presentation:

- http://cse710.blogspot.com/

- In class participation is required as well
- (Attendance will be taken each class)

Projects

- ☑ Design and implementation of a Distributed Metadata Server for Global Name Space in a Wide-area File System [3-student teams]

- ☑ Design and implementation of a serverless Distributed File System (p2p) for smartphones [3-student teams]

- ☑ Design and implementation of a Cloud-hosted Directory Listing Service for lightweight clients (i.e. web clients, smartphones) [2-student teams]

- ☑ Design and implementation of a Fuse-based POSIX Wide-area File System interface to remote GridFTP servers [2-student teams]

Project Milestones

- Survey of Related work -- Feb. 6th
- Design document -- Feb 20th
- Midterm Presentations -- March 6th
- Imp. Status Report -- Apr. 3rd
- Final Present. & Demos -- Apr. 17th
- Final Reports -- May 9th

Contact Information

- Prof. Tevfik Kosar
- Office: 338J Davis Hall
- Phone: 645-2323
- Email: tkosar@buffalo.edu
- Web: www.cse.buffalo.edu/~tkosar

- Office hours: Tue 10:00am – 12:00pm
- Course web page: http://www.cse.buffalo.edu/faculty/tkosar/cse710_spring13

Any Questions?