Terascale Clusters and Data Grids: A Look at the Future

Dan Reed

Director, NCSA and the Alliance

Chief Architect, NSF TeraGrid

reed@ncsa.uiuc.edu

The HPC Conundrum

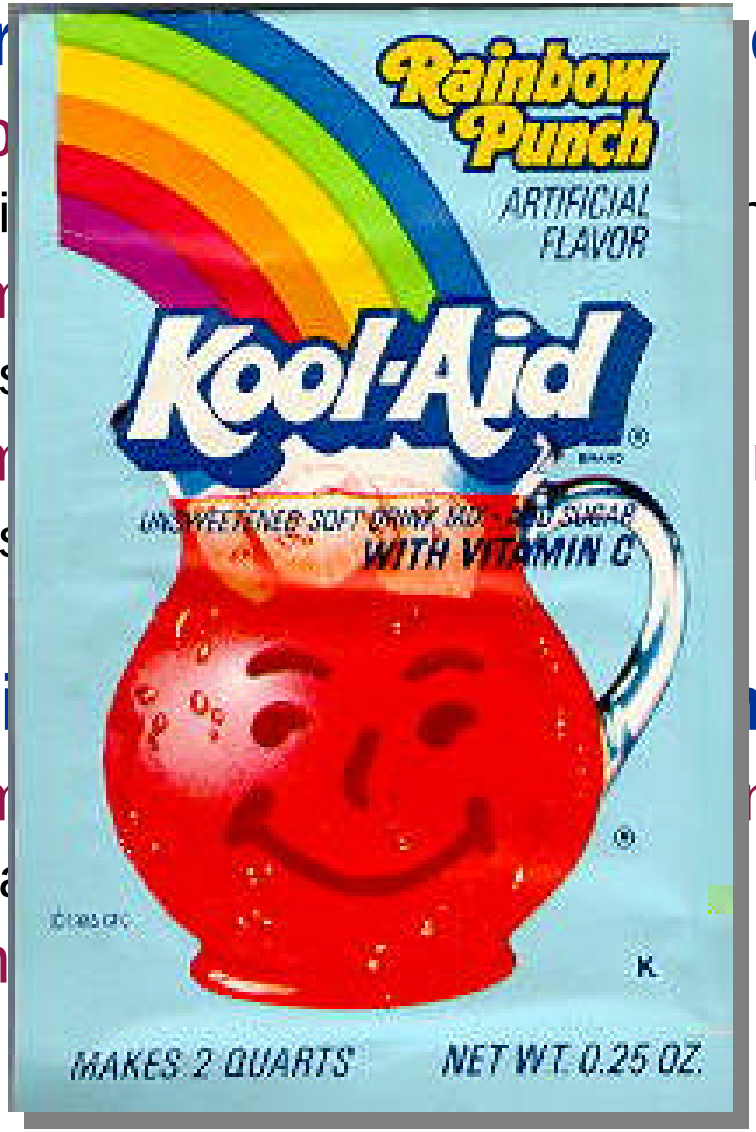
- **The conundrum**
  - large NRE costs
    - particularly software
  - modest size markets
    - TMC, KSR, Cray, SGI, …
  - NCSA's previous vendors
    - almost all RIP
- **Implications**
  - 1980s-1990s market shakeout
  - limited base to amortize software costs
- **Rescue (Ben Franklin)**
  - open source
  - commodity clusters

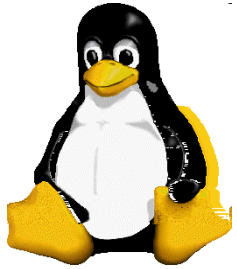
It's The Software, Duh

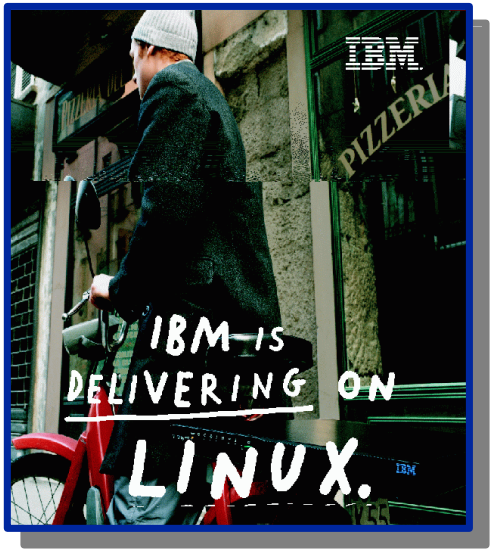
- **We have this fascination with hardware**
  - it's tangible and obvious
  - it's necessary but not sufficient
    - like a car without a road
  - look at our experiences
    - CDC 6600, Cray 1, CM-5, KSR, …
- **Software is always overlooked**
  - both application and infrastructure
  - application embodies the science
  - tools enable/hinder productivity
  - it's not cheap!
    - think about the cost of a single Internet startup **L**
  - you must keep supporting it

Why Linux Clusters?

- "Ever em
  - app
    - i nes
  - com
    - s vehicles
  - com roups
    - s
- Thrivi nmunity
  - com nt
    - a and tools
  - ven







Similar ad campaigns for other vendors

On The Other Hand …

… we could just be crazy! J

But We Don't

Terascale Clusters: First Steps

- **1 TF IA-32 Pentium III cluster (Platinum)**
  - 512 1 GHz dual processor nodes
  - Myrinet 2000 interconnect
  - 5 TB of RAID storage
  - #30 on Top500 list

- **1 TF IA-64 Itanium cluster (Titan)**
  - 164 800 MHz dual processor nodes
  - Myrinet 2000 interconnect
  - shared disk storage

NCSA Platinum Cluster Timeline

Jan 2001

Order placed with IBM 512 compute node cluster

Feb 2001

2/23 First four racks of IBM hardware arrive

Mar 2001

3/1 Head nodes operational
3/10 First 126 processor Myrinet test jobs
3/13 Final IBM hardware shipment
3/22 First application for compute nodes (CMS/Koranda/Litvin)
3/26 Initial Globus installation
3/26 Final Myrinet hardware arrives
3/26 First 512 processor MILC and NAMD runs

Apr 2001

4/4 NFS problem resolved for CMS
4/5 Myrinet static mapping in place
4/7 CMS runs successfully
4/11 400 processor HPL runs completing
4/12 Myricom engineering assistance

May 2001

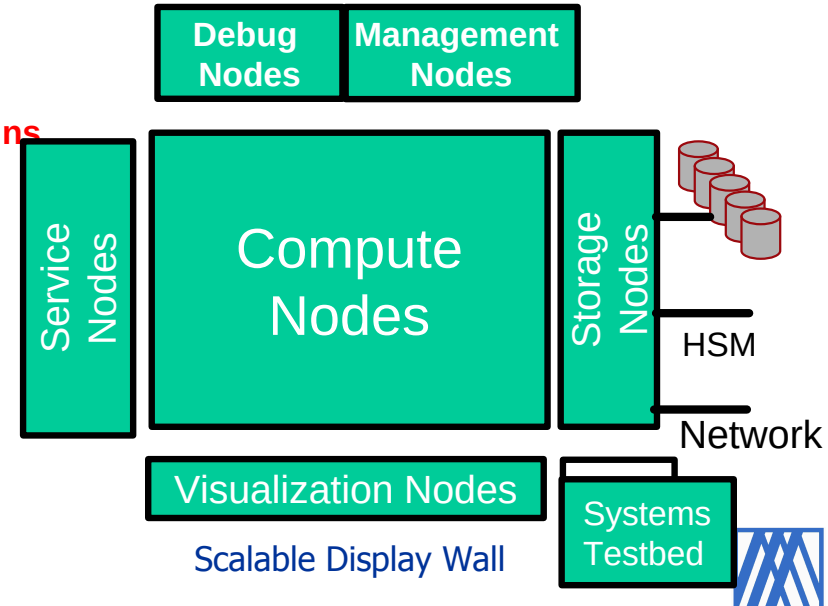
5/8 1000p MP Linpack runs
5/11 1008 processor Top500 run @ 594GF
5/14 2.4 Kernel testing
5/28 RedHat 7.1 testing

June 2001

6/1 Friendly user period begins

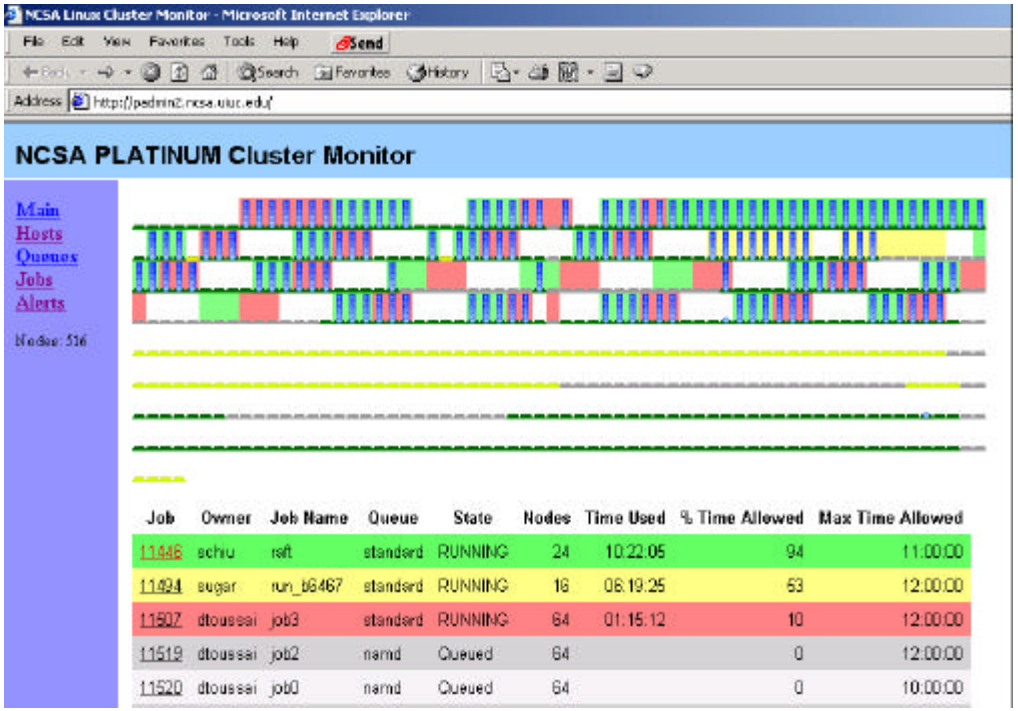
July 2001

Production service since July



Platinum Software Configuration

- **Linux**
  - RedHat 6.2 and Linux 2.2.19 SMP Kernel
- **Open PBS**
  - resource management and job control
- **Maui Scheduler**
  - advanced scheduling
- **Argonne MPICH**
  - parallel programming API
- **NCSA VMI**
  - communication middleware
    - MPICH and Myrinet
- **Myricom GM**
  - Myrinet communication layer
- **NCSA cluster monitor**



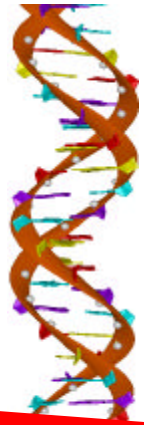
The Biology Revolution

DNA sequence

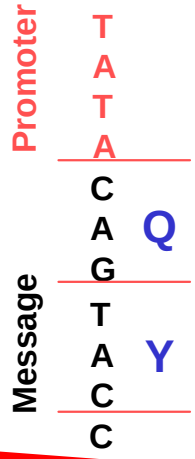
Protein sequence and regulation

Protein structure

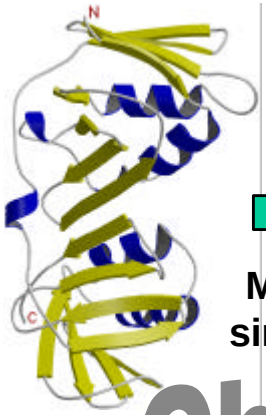
Protein/enzyme function



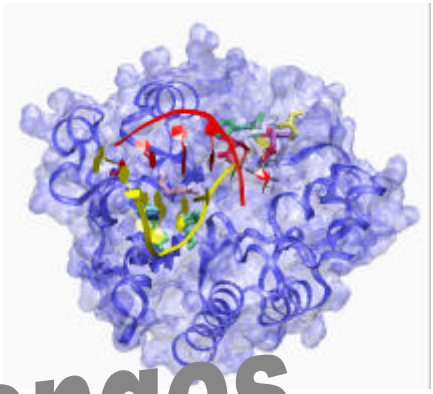
Sequence Annotation



Homology based protein structure prediction



Molecular simulations



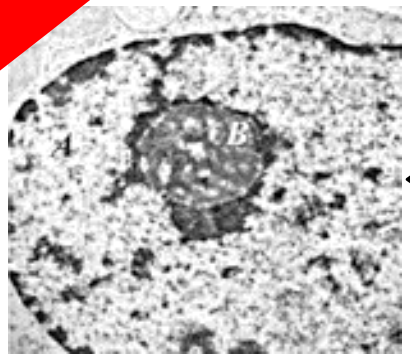
New Challenges

Data integration

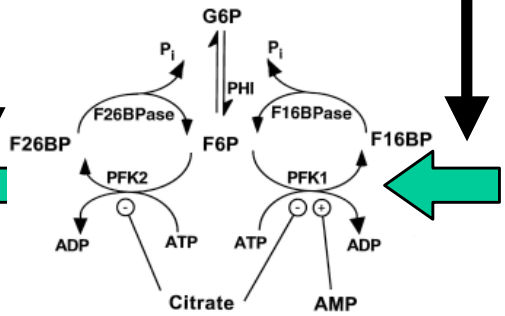
Organ

Pathway simulations

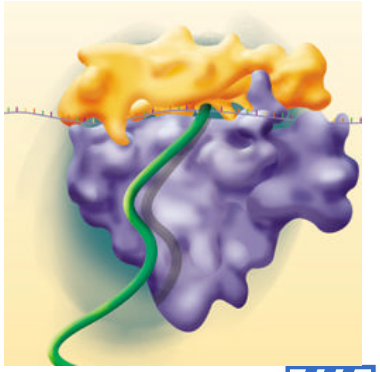
Network analysis



Bacteria and cells



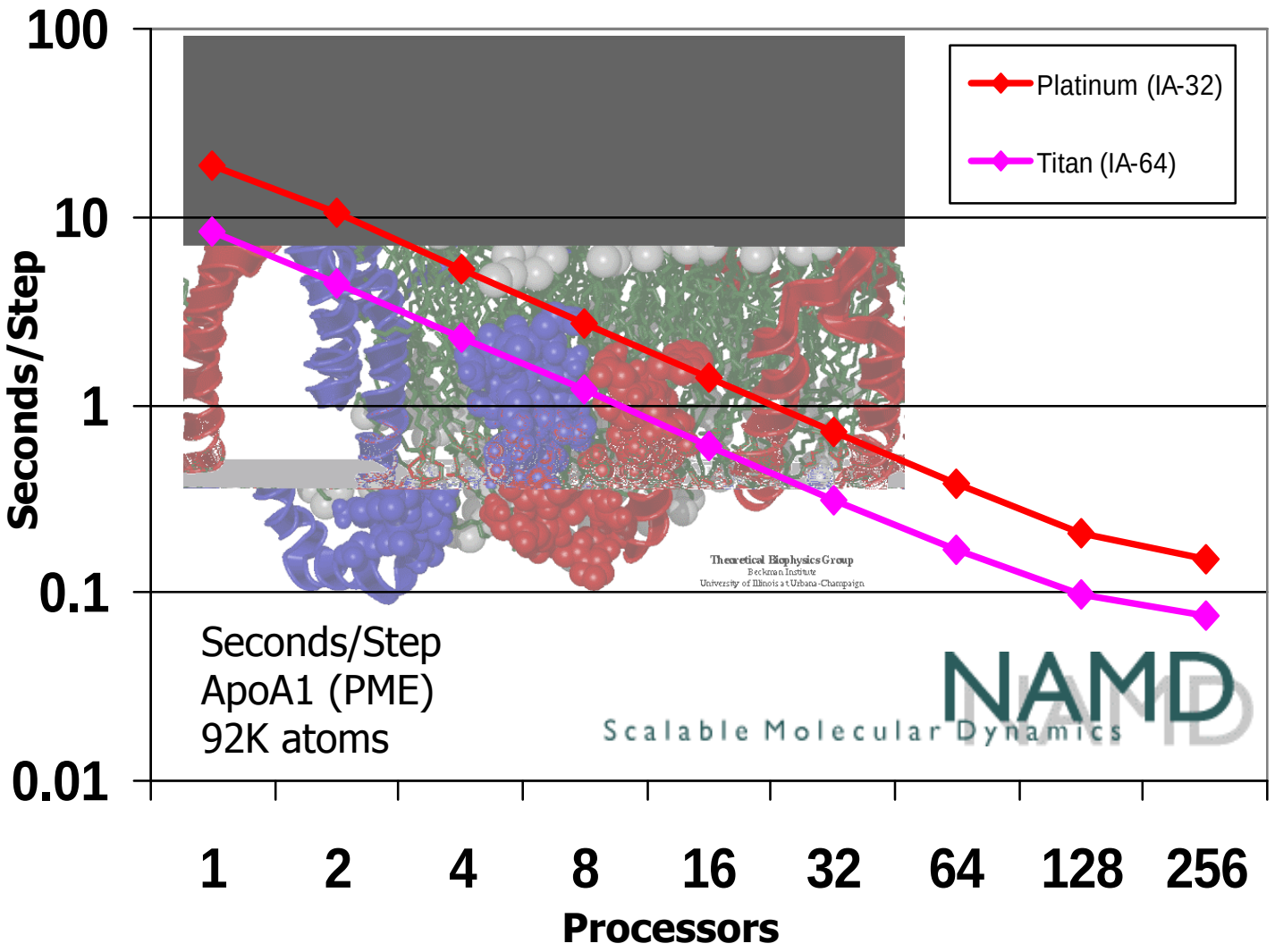
Metabolic pathways and regulatory networks



Multi-protein machines

NCSA

NAMD: Scalable Molecular Dynamics



Simulation of large biomolecular systems on parallel computers

File compatible with CHARMM & AMBER

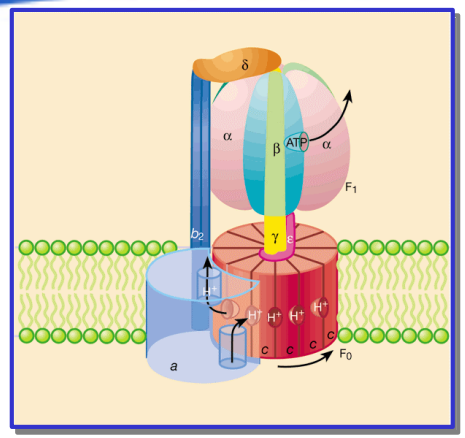
Message-driven and object-oriented design implemented with Charm++/Converse (from PPL at UIUC)

Ported to PACI systems, clusters, and desktop PCs

Available for **FREE**, includes source code

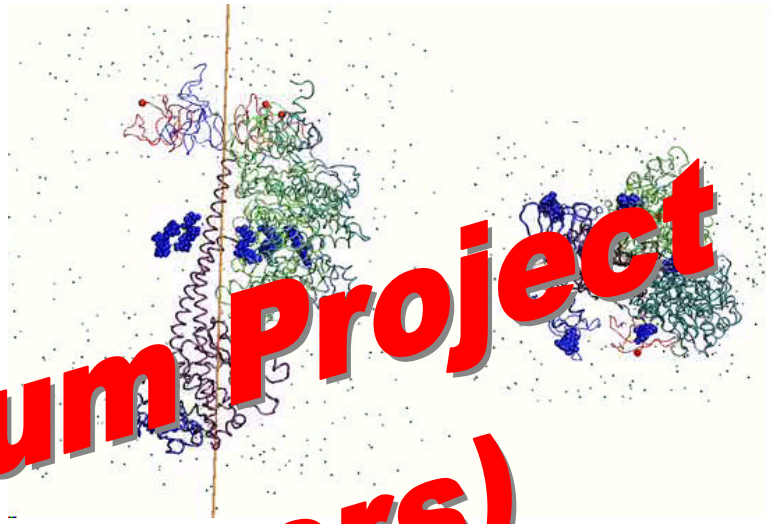
Source: Klaus Schulten

ATP Synthase F1 Stalk Rotation



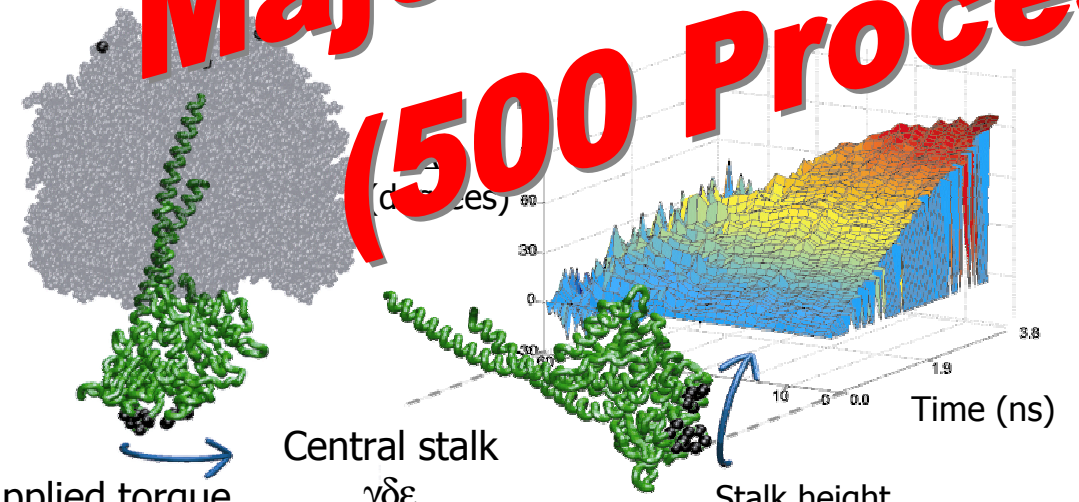
nH^+ (F_0 side)+ ADP + Pi ⟶ ATP +H_20 + nH^+ (F_1 side)

ATP synthase F_1 is a cellular m[illegible] turns rotation into A[illegible]



Trial simu[illegible] [illegible]que applied to 361 [illegible] stalk; w=10,800°/ns

Major Platinum Project (500 Processors)



Applied torque

Central stalk
γδε

Stalk height
(Å)

Production simulation. Torque applied to 16 atoms at bottom of stalk; w=24°/ns. Rotation smoothly propagates along central stalk

The system contains 327,000 atoms. The 3.8 ns MD simulation (particle mesh Ewald electrostatics, periodic boundary conditions) required 135,000 SU on NCSA Platinum

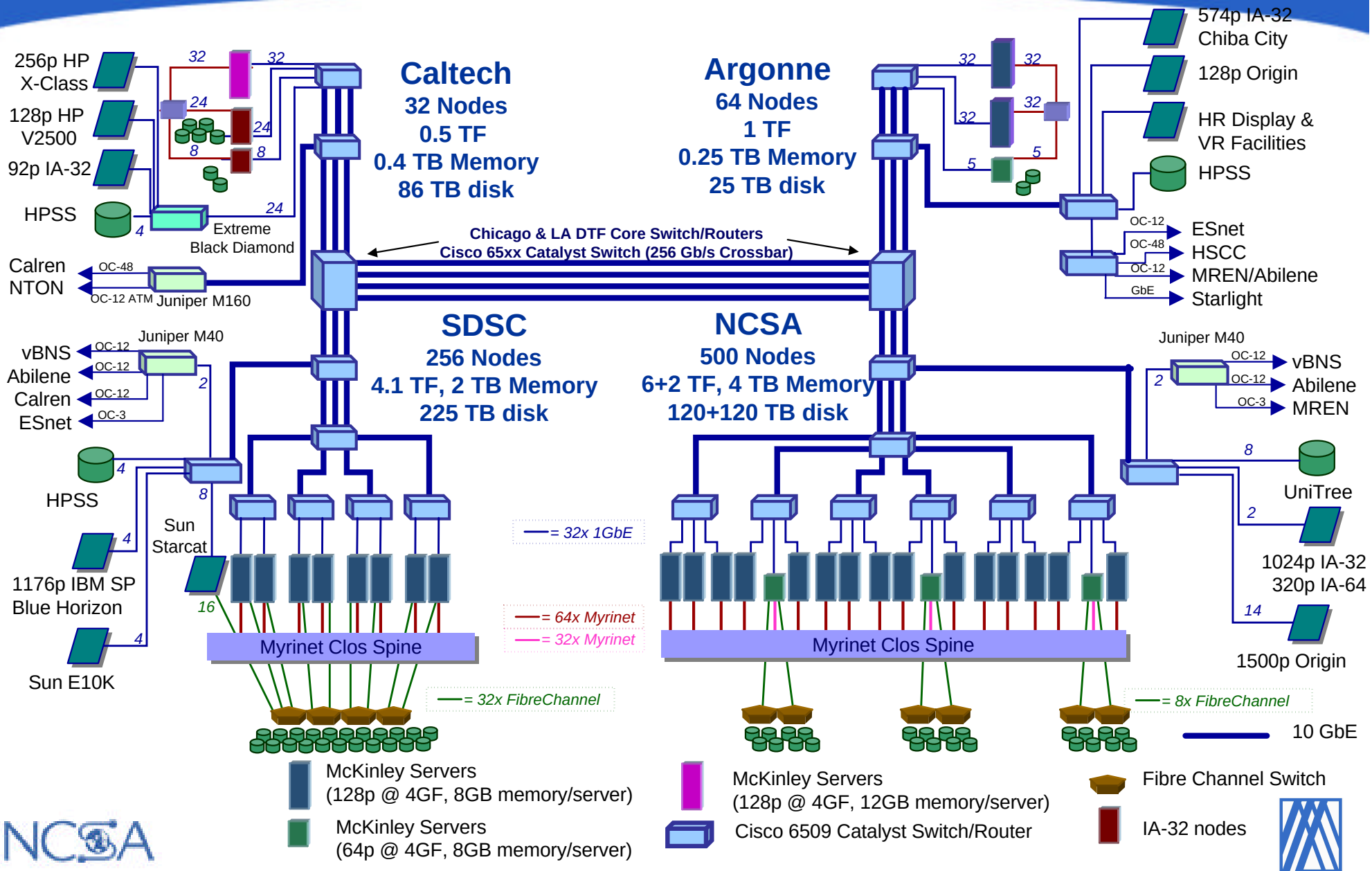
Source: Klaus Schulten

TeraGrid: Linux Writ Large

- **TeraGrid partners**
  - NCSA
    - Dan Reed
  - SDSC
    - Fran Berman
  - Argonne
    - Rick Stevens
    - Ian Foster
  - Caltech
    - Paul Messina

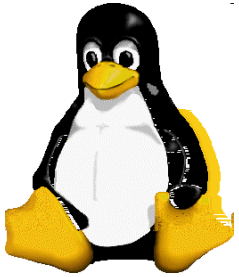
NCSA

PACI Linux McKinley TeraGrid



TeraGrid Strategic Partners

- **IBM**
  - cluster integration and RAS
  - GPFS parallel file system
- **Intel**
  - McKinley IA-64 software and compilers
- **Oracle/IBM**
  - data archive management and mining
- **Qwest**
  - 40 Gb/s DTF WAN backbone
- **Myricom**
  - cluster interconnect
- **SUN**
  - metadata service



IBM

intel

ORACLE

ride the light
Qwest

Myricom

NCSA

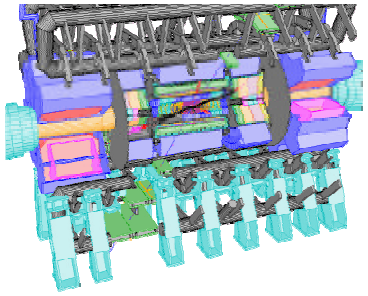
TeraGrid Application Targets

- **Multiple classes of user support**
  - each with differing implementation complexity
    - minimal change from current practice
    - new models, software, and applications

- **Usage exemplars**
  - “traditional” supercomputing made simpler
    - remote access to data archives and computers
  - distributed data archive access and correlation
  - remote rendering and visualization
  - remote sensor and instrument coupling

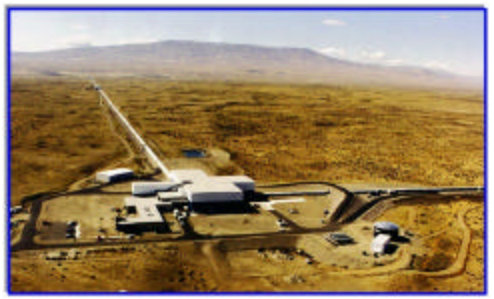
Challenges and Opportunities

ALMA



ATLAS and CMS

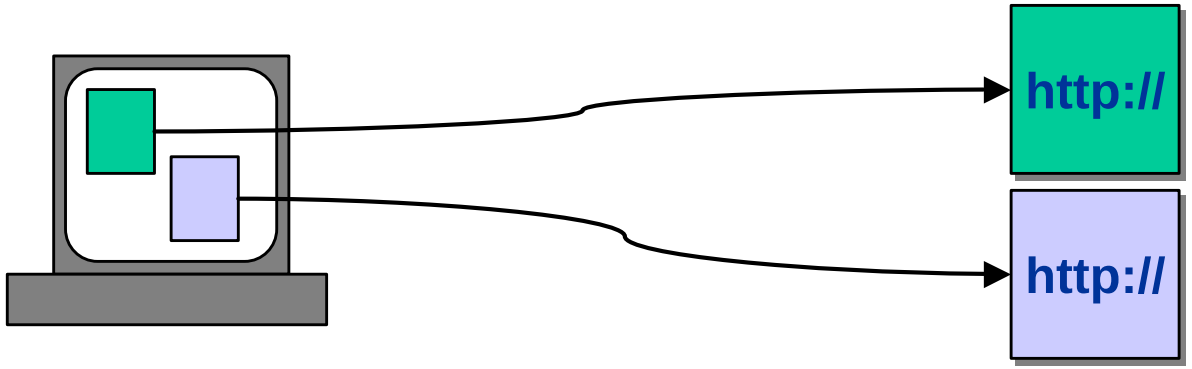
Sloan DSS



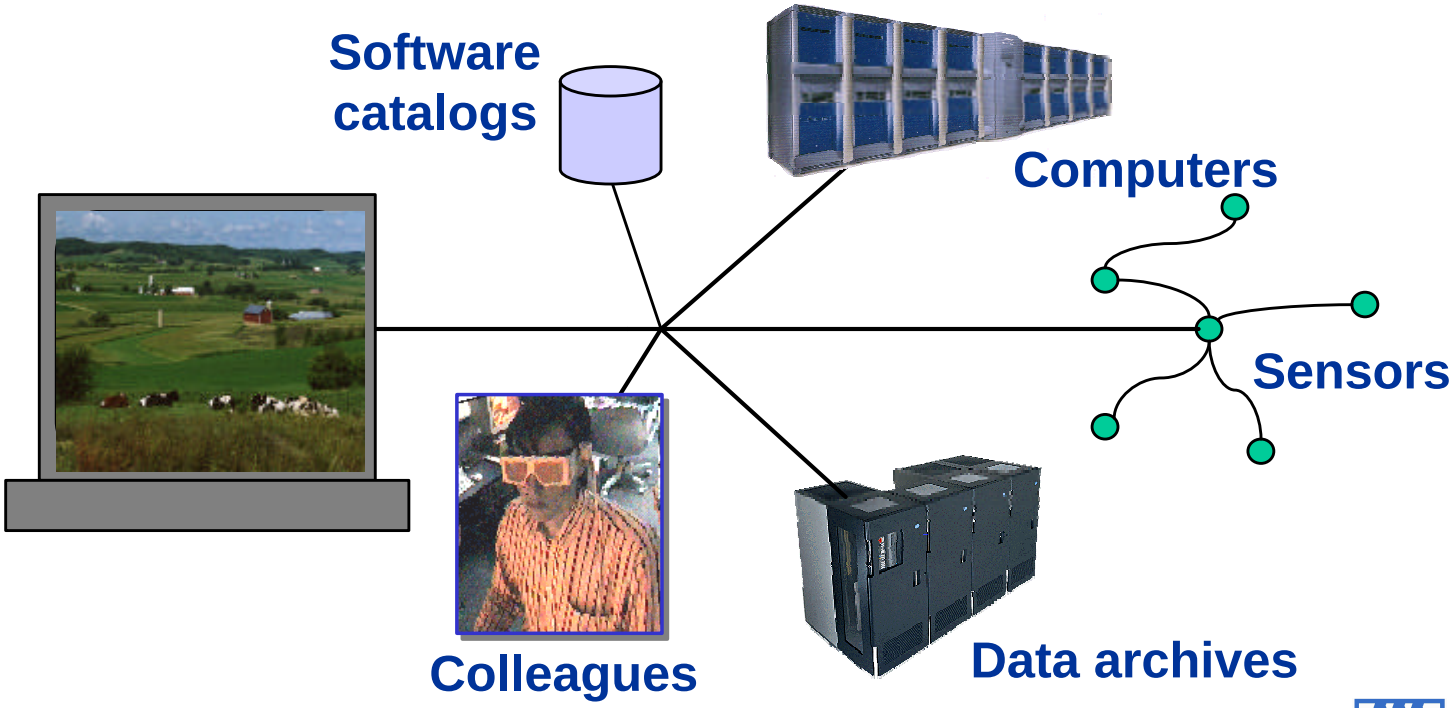
LIGO

Grids: The New Internet

***Web*: Uniform access to HTML documents**

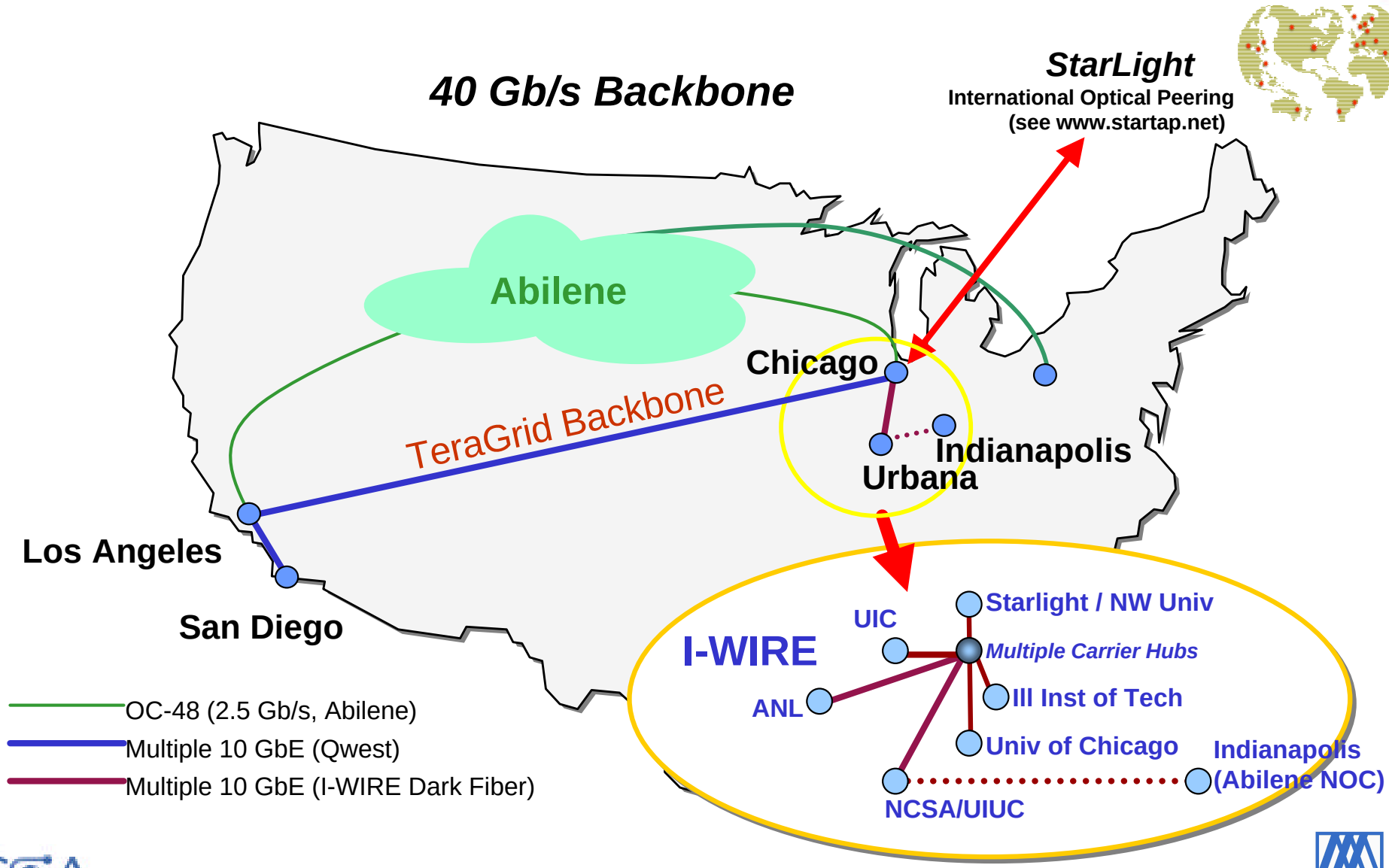


***Grid*: Flexible, high-performance access to resources**



Source: Harvey Newman, Caltech

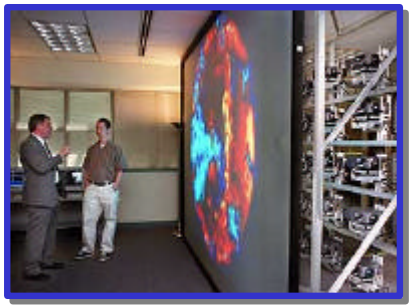
TeraGrid Network Backbone



$7.5M Illinois DWDM Initiative

Alliance X-in-a-Box Initiatives

- **Access Grid**
  - Stevens/Childers (ANL)
- **Display Wall**
  - Baker (NCSA)
- **Clusters**
  - OSCAR/Pennington (NCSA)
  - http://oscar.sourceforge.net
  - http://www.ncsa.uiuc.edu/TechFocus/Deployment/CiB/index.html
- **Grid**
  - Butler (NCSA) and Foster (ANL)
- **Data**
  - Welge/Folk (NCSA)
- **Applications**
  - Crutcher (NCSA)

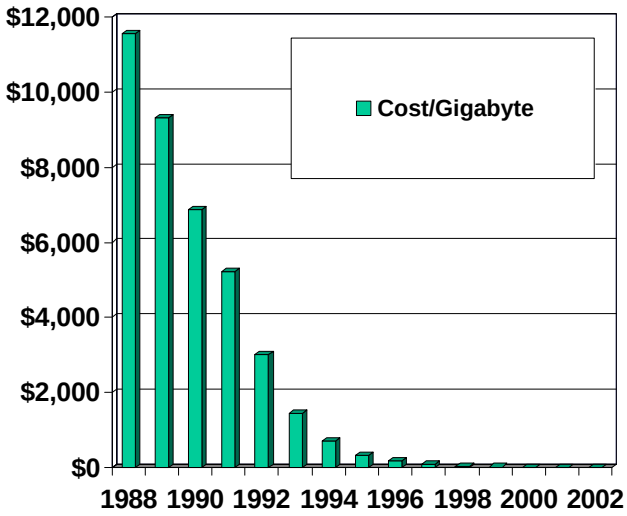


Scalable Display Walls

- **Large format with human scale**
  - commodity components
    - projectors, Linux cluster, cheap frames
  - high resolution (8192 x 3840 pixels)
    - now doubling in size

Petabytes and Beyond

- **"Massive" data is being redefined**
  - O($5000) terabyte disk systems
    - commodity disks
    - software RAID
  - petabyte disk farms very near
    - personal petabytes to follow

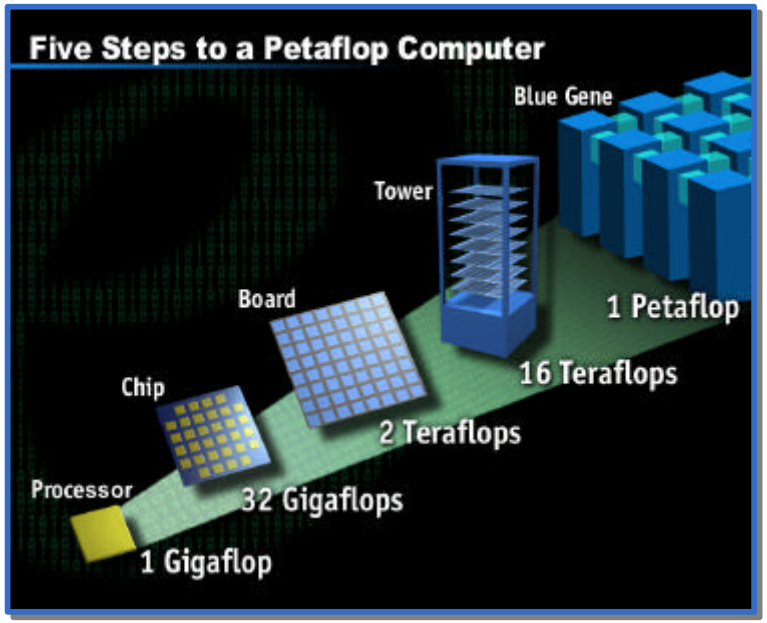


- **Look at the science**
  - experimental (LHC, ALMA, LIGO, …)
  - computational (lots of great examples)
  - petabytes/day coming RSN!

- **Implications**
  - data explosion of unprecedented proportions
  - data is *not* knowledge
  - data mining is becoming critical

Petaops: Thinking Out of the Box

- **Riding the commodity technology curve**
  - petaflop PC clusters (now)
    - $1B-$4B cost
    - 100 MW power
  - new approaches
    - deep integration and packaging
    - new algorithms and software
- **Deskside teraflops**
  - O($50K) teraflop systems
  - empowering new research
- **The new commodities**
  - IBM, Sony, Toshiba PlayStation3
    - 1000X PlayStation2 performance
    - $400M investment
    - teraop processor design
  - IBM research projects
    - Blue Gene and protein folding
  - rethinking technology leverage

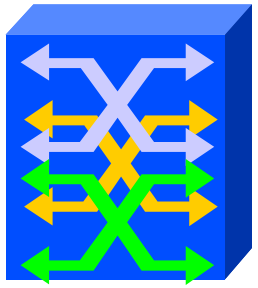


IBM Research

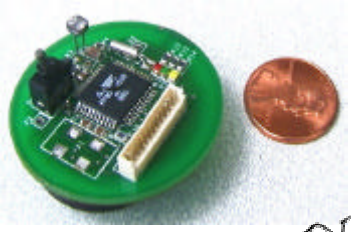
$50,000 teraflops

The Ubiquitous Infosphere

- **Terabit networking, the dominance of lambdas**
  - Dense Wavelength Division Multiplexing (DWDM)
    - thousands of lambdas/fiber
  - bandwidth growth to terabits/second
    - fiber being laid at 10,000 km/hour
  - the "revenge" of circuit switching
    - wavelengths as instrument of exchange
  - *the commodity cluster Grid writ large*



- **Mobile sensors, pixie dust gets real**
  - micro-electro-mechanical systems (MEMS)
  - wireless, sub-millimeter sensor/actuators
  - *the commodity cluster Grid writ small*



NCSA

From *In Vitro* to *In Silica*

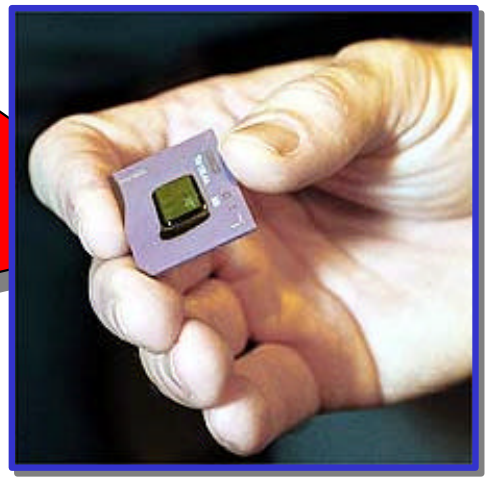
Associated Press

In Vitro

New York Times

In Vivo

In Silica



Agence France-Presse

Teraflop in a Minute

NCSA IA-64 Cluster Deployment