TeraGrid and High-end Computing: Lessons and Futures

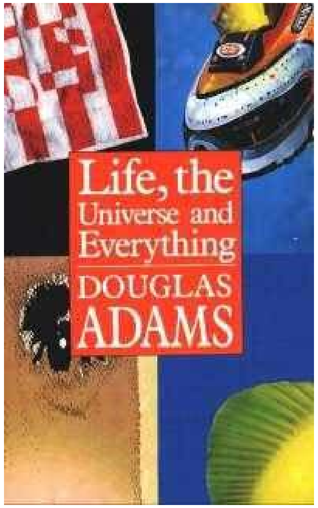
Director, NCSA and the Alliance
Chief Architect, NSF ETF TeraGrid
William and Jane Marr Gutgsell Professor

University of Illinois

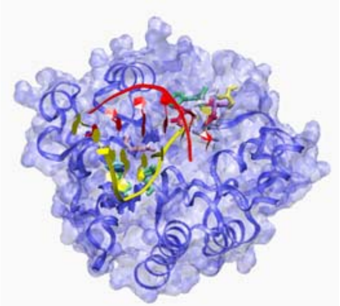
reed@ncsa.uiuc.edu

Outline

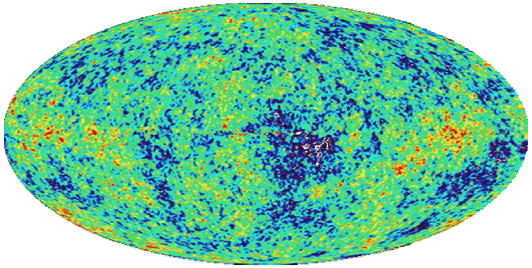
- **The really big questions**
  - life, the universe, and everything
    - why they matter and how we react
- **TeraGrid and NCSA**
  - status and directions
    - lessons and capabilities
    - applications and needs
- **High-end futures**
  - petascale system design
    - challenges and opportunities
  - international networks



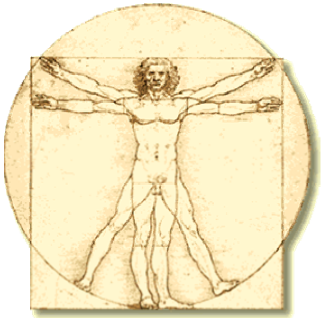
The Big Questions



- **Life and nature**
  - structures, processes and interactions
- **Matter and universe**
  - origins, structure, manipulation and futures
  - interactions, systems, and context

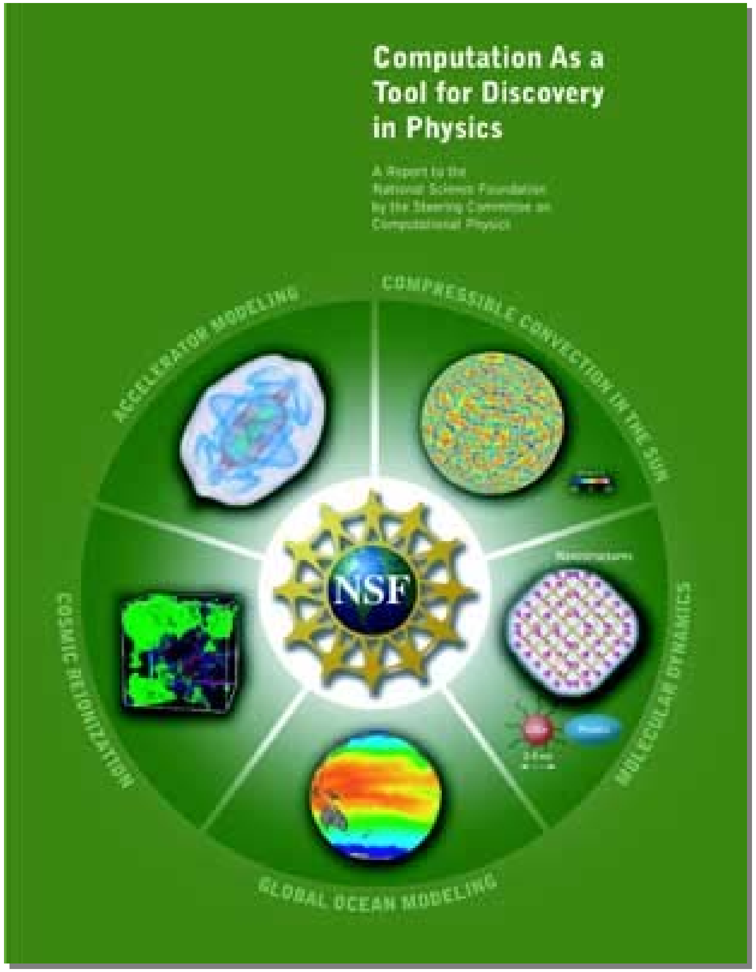


- **Humanity**
  - creativity, socialization and community
- ***Answering big questions requires***
  - *boldness to engage opportunities*
  - *expandable approaches*
  - *world-leading infrastructure*
  - *broad collaborations and interdisciplinary partnerships*



HPC Application Studies

- **Contributors**
  - Berger (NYU)
  - Ceperley (UIUC)
  - Koonin (Caltech)
  - McCurdy (LBL)
  - Mount (SLAC)
  - Ostriker (Princeton)
  - Reed (UIUC)
  - Sugar (UCSB)
  - Wilkins (Ohio State)
  - Woodward (Minnesota)



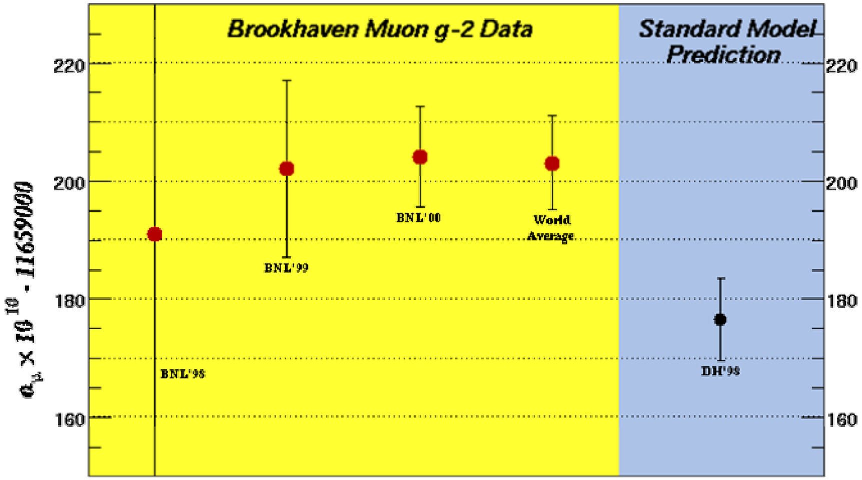
Computation As a
Tool for Discovery
in Physics

http://www.nsf.gov/pubs/2002/nsf02176/start.htm

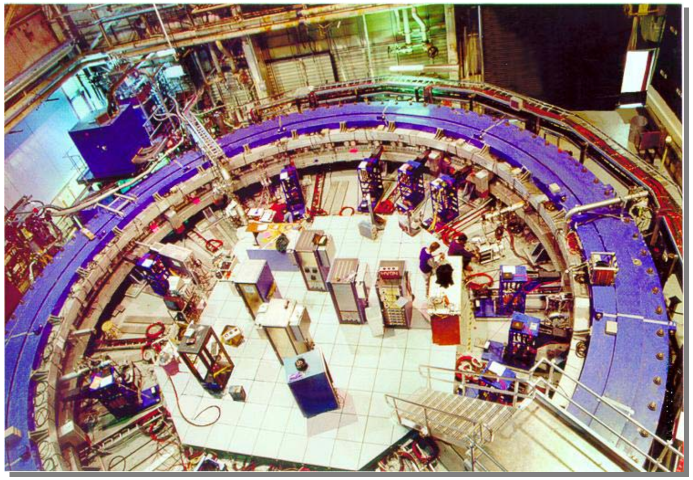
G-2 Data Analysis

- **Brookhaven g-2 experiment**
  - 60 scientists from 11 institutions
  - standard model testing
    - muon anomalous magnetic moments
  - large-scale cluster data analysis
    - Herzog *et al* (Illinois), 10X time decrease

- **The story is in the data …**



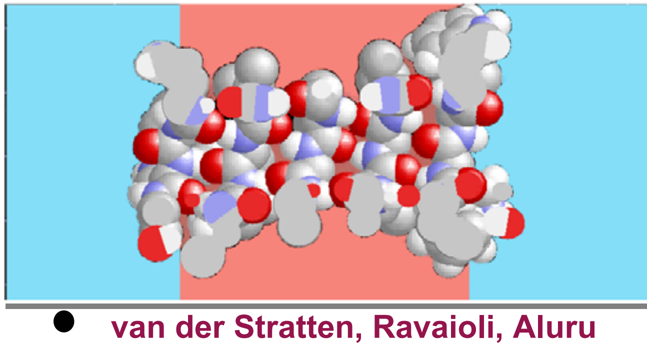
References: BNL'98 PRL 86 2227
BNL'99 PR 62D 091101
BNL'00 accepted for publication in PRL
DH'98 a ,(had;1) from PL 435B 427



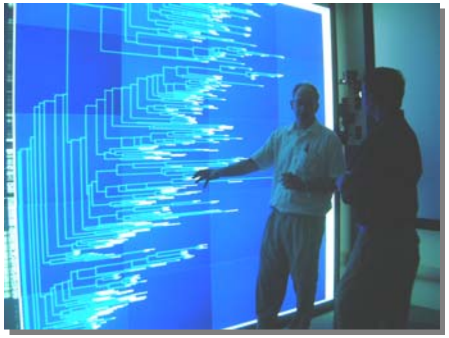
Evolutionary Biology

- **Phylogenetic tree display**
  - all known prokaryotic potassium channels
  - all potassium channels from two animals
    - *homo Sapiens* and *c. Elegans*
  - three classes of animal channels
    - voltage gated, ligand modulated and “other”
- **Observations**
  - Ligand modulated/voltage-gated prokaryotic
    - mingled with the animal channels
    - indicating origin in prokaryotics
  - “other” class
    - no prokaryotes, indicating an origin in eukaryotes
  - roots of two other channels (sodium and calcium)
    - arose in a region in the ligand-gated group

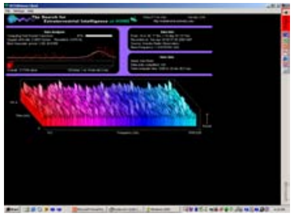
Gramicidin ionic channel



van der Stratten, Ravaioli, Aluru



The Computing Continuum



SETI@home
The Search for Extraterrestrial Intelligence

TERAGRID

Loosely Coupled

"SETI" "Grids" Clusters SMPs

Tightly Coupled

- **Each strikes a different balance**
  - computation/communication coupling
- **Implications for execution efficiency**
- ***Applications for diverse needs***
  - *computing is only one part of the story!*

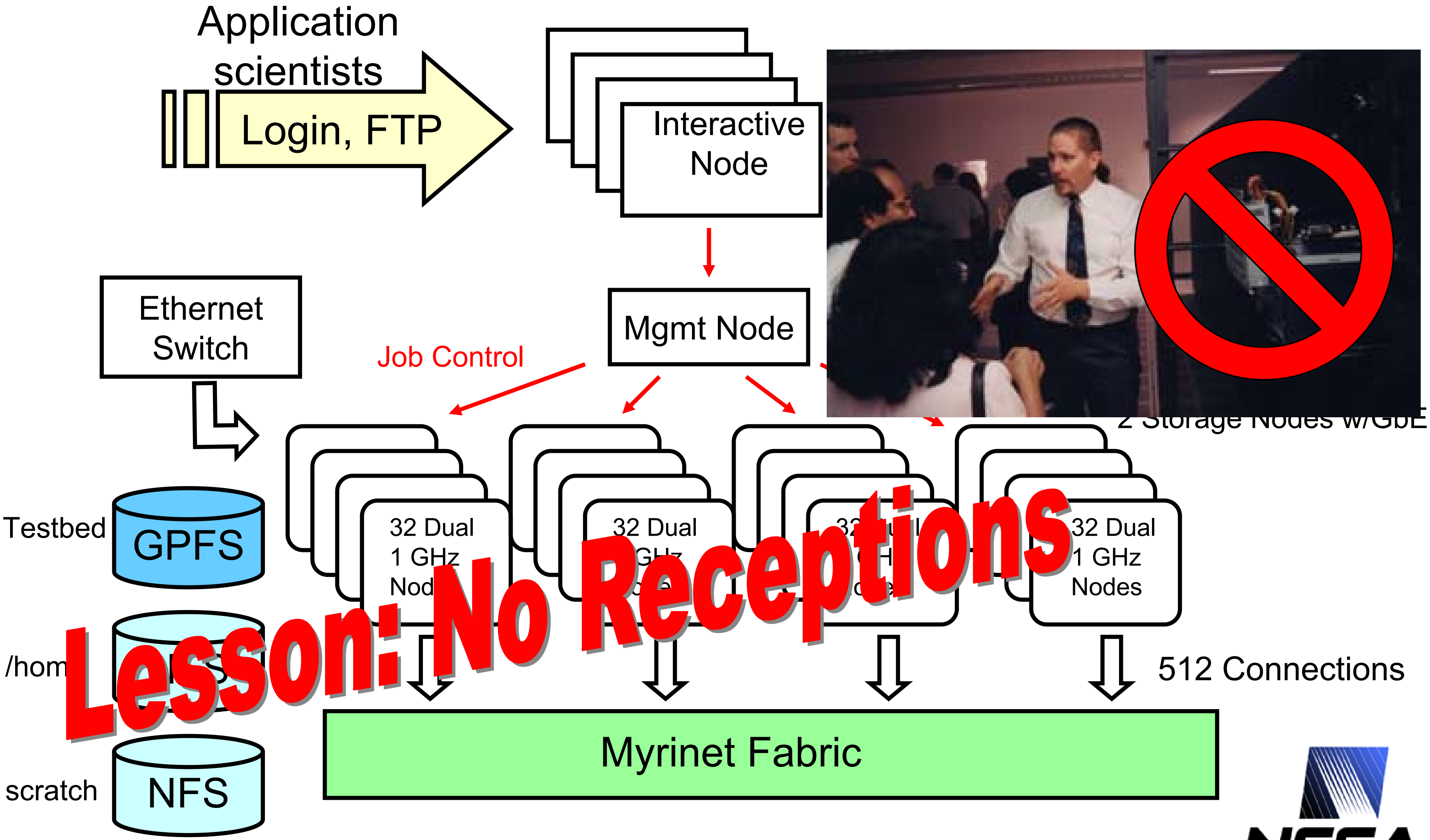
2003 NCSA System Status: 30+TF

- **Three new major computational systems**
  - *17.7 TF Dell Xeon replacement for 1 TF Pentium III cluster*
    - *1474 Dell servers, dual Intel Xeon 3.06 GHz nodes*
    - *installation in progress now and scheduled for operation in late 2003*
  - TeraGrid cluster with Itanium2/Madison nodes
    - 2 TF Itanium2 systems delivered, upgraded to 1.3 GHz Madison
      - production December 2003
    - additional 8 TF of Madison in fall 2003 (production April 2004)
  - IBM p690 32p SMP systems
    - operational in spring 2003
    - 2 TF, 12 systems, 384 1.3 GHz Power4 processors
    - 4 large memory systems with 256 GB of memory
- **Two other production clusters**
  - 1 TF Pentium III and 1 TF Itanium
- **Condor resource pools**
    - parameter studies and load sharing
- **~500 TB of spinning storage**
  - Brocade SAN fabric with DataDirect, IBM and LSI storage arrays

Dell PowerEdge 1750

pSeries 690

1 TF IBM Pentium III Cluster



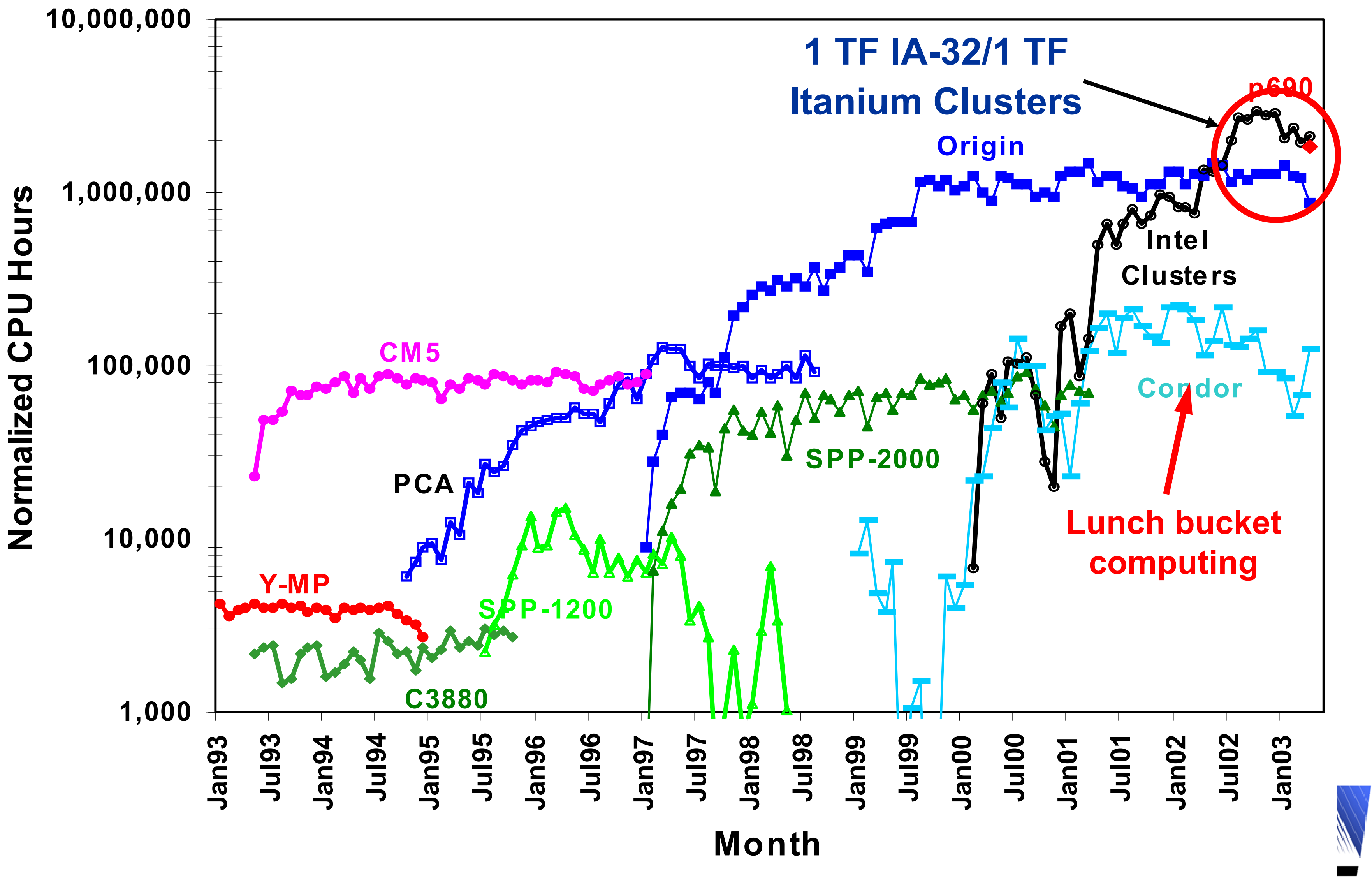
Lesson: No Receptions

Cluster in a Box/OSCAR

- **Community code base with strong support**
  - –Bald Guy Software, Dell, IBM, Intel, Indiana
  - –MSC.Software, NCSA, ORNL, Sherbrooke University, …
- **Six releases within the past year**
  - –29,000 downloads during this period
- **Recent additions**
  - –HDF4/HDF5 I/O libraries
  - –OSCAR database for cluster configuration
  - –Itanium2 and Gelato consortium integration
  - –NCSA cluster monitor package (Clumon)
  - –NCSA VMI 2 messaging layer
    - Myrinet, gigabit Ethernet and Infiniband
  - –PVFS added for parallel filesystem
- ***First Annual OSCAR Symposium***
  - –*May 11-14, 2003, Québec, CANADA*

MSI Cluster Workshop
May 6-7, 2002

NCSA Resource Usage



1 TF IA-32/1 TF Itanium Clusters

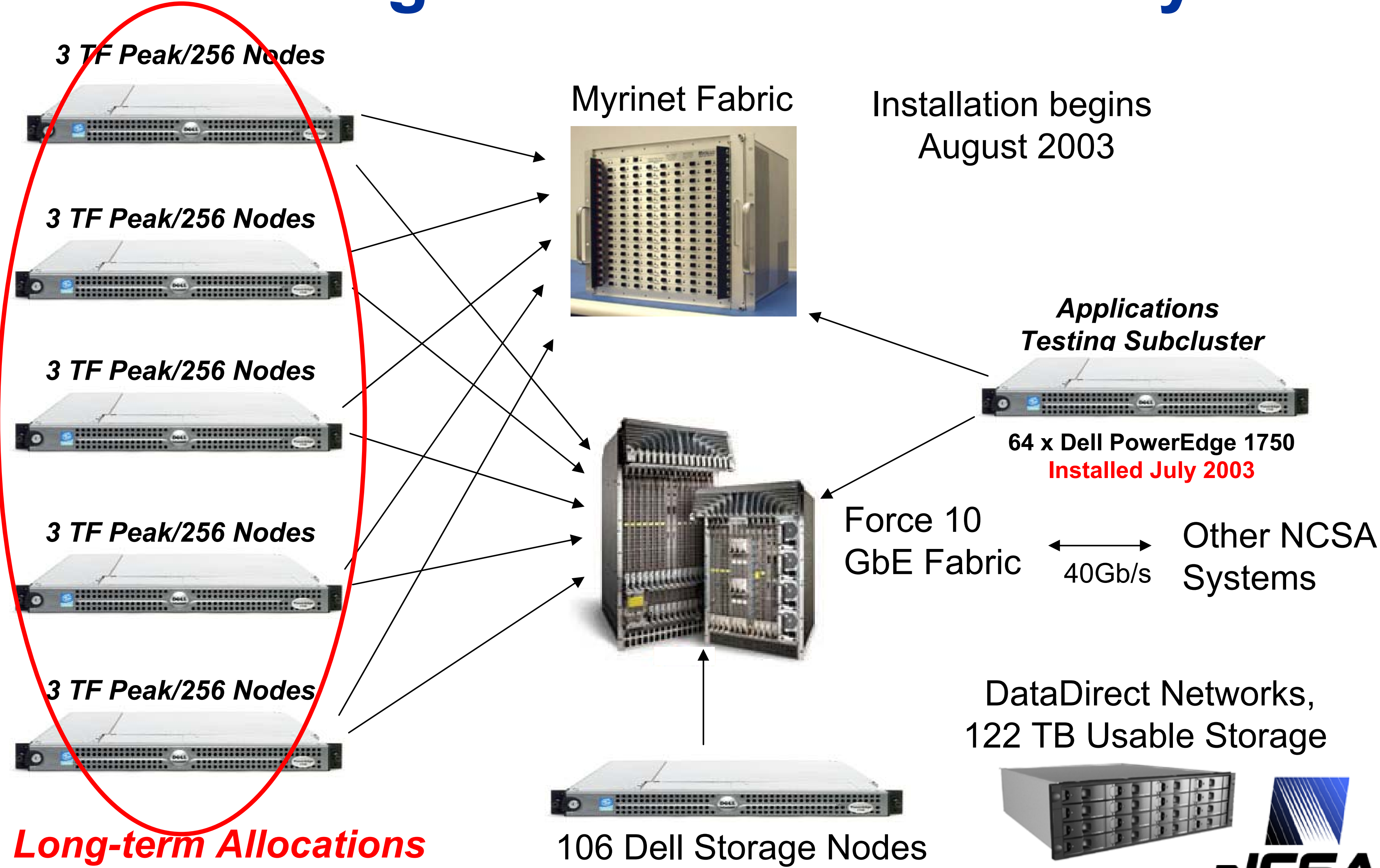
Lunch bucket computing

National Center for Supercomputing Applications

Tungsten: 17.7 TF Peak from Dell

- **Dell Xeon cluster for highly scalable applications**
  - successor to production 512 node/1024p Pentium III cluster
  - scalable configuration for administration and management
    - LSF, Red Hat 9, LSF, Lustre, Dell OpenManage …
- **Maximize increase in capability**
  - 3.06 GHz, 512 KB (L2), dual Xeon nodes
    - application interest in hyperthreading
- **Increase in storage capacity and capability**
  - 100-200 TB and 2GB/s per TF peak
- **Scalable administration**
  - 256 nodes/administration server
- ***Support for very large, long running applications***
  - *256 node/3 TF administrative "subclusters"*
  - *run times of at least 1 week on 100s of nodes*
    - *significant numbers of nodes dedicated to projects*

NCSA Tungsten: Xeon Scalable System



3 TF Peak/256 Nodes

3 TF Peak/256 Nodes

3 TF Peak/256 Nodes

3 TF Peak/256 Nodes

3 TF Peak/256 Nodes

Myrinet Fabric

Installation begins
August 2003

Applications Testing Subcluster

64 x Dell PowerEdge 1750
Installed July 2003

Force 10
GbE Fabric

40Gb/s

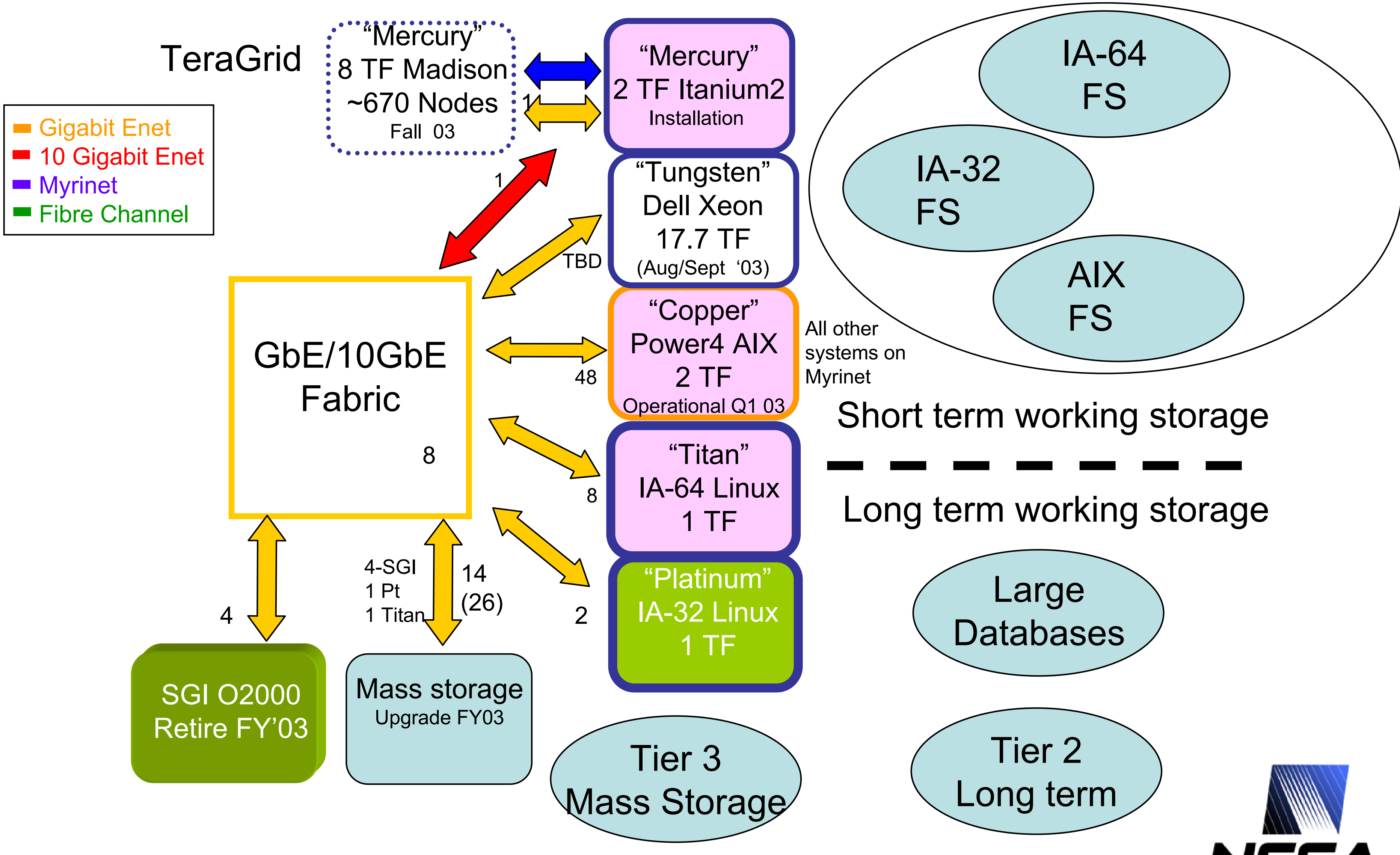
Other NCSA
Systems

DataDirect Networks,
122 TB Usable Storage

106 Dell Storage Nodes

Long-term Allocations

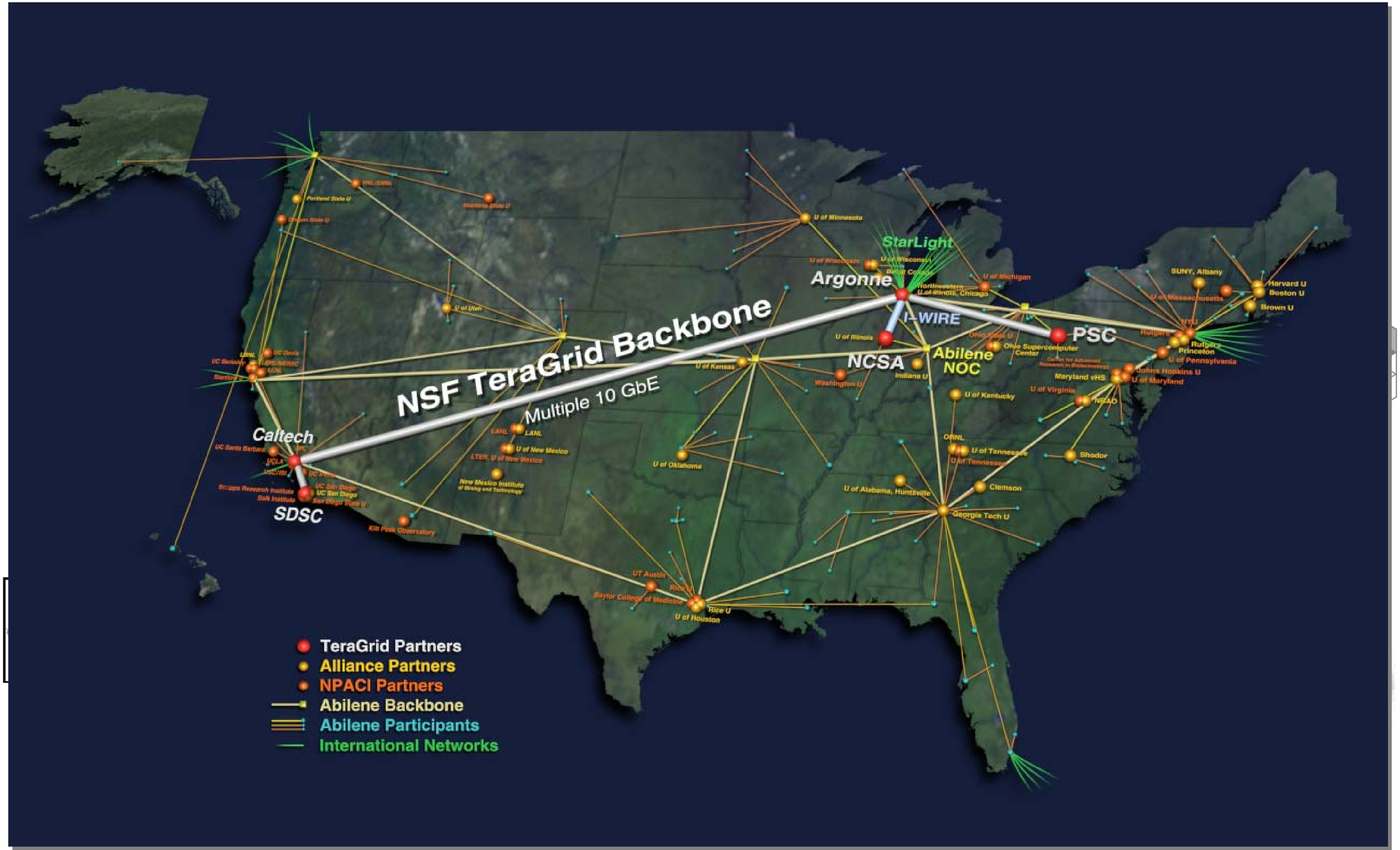
NCSA System Integration



National Center for Supercomputing Applications

NCSA

Science and Engineering Grids



TeraGrid Objectives

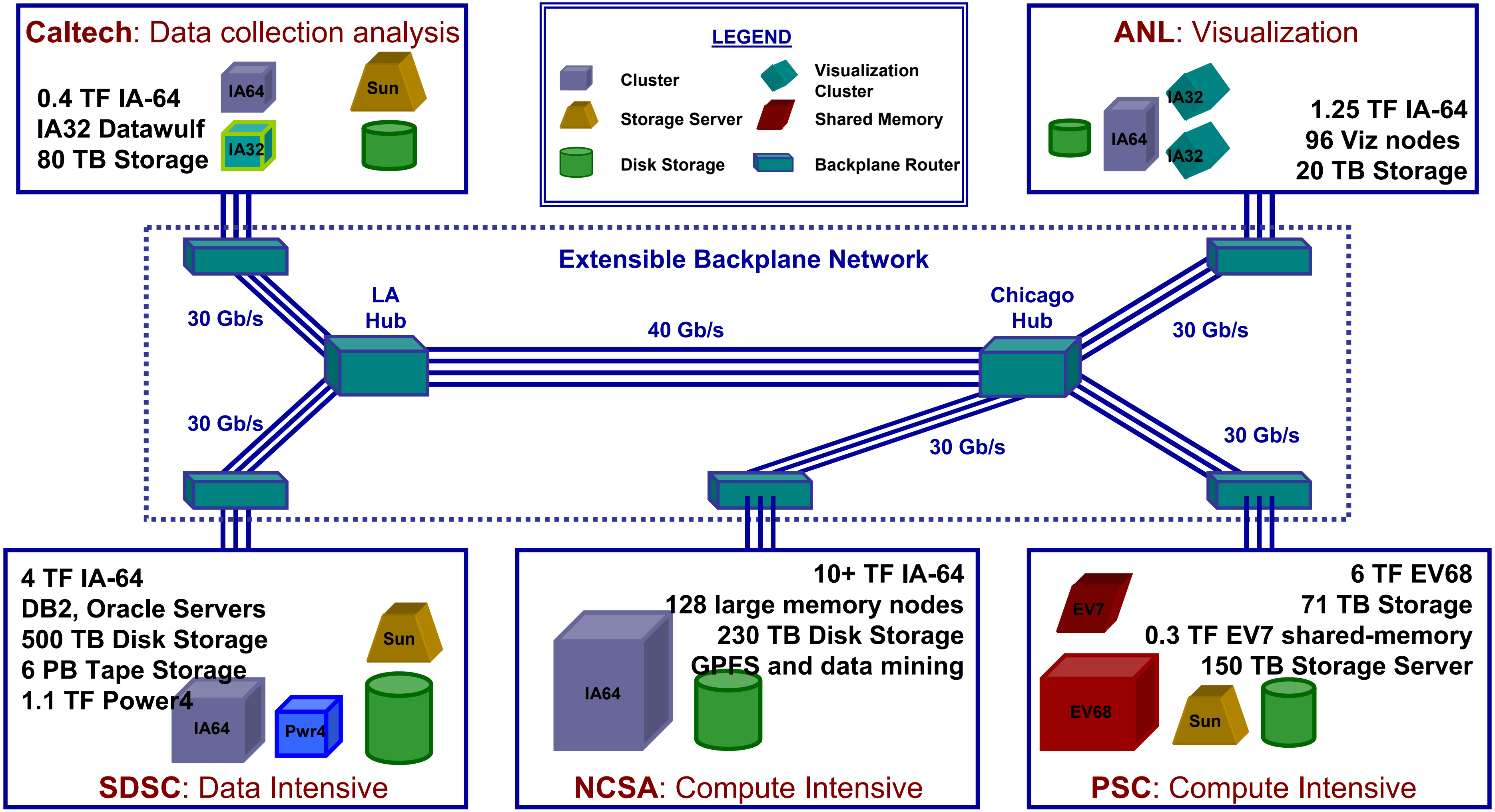
- **Enable and empower new science**
  - “traditional” supercomputing made simpler
    - remote access to data and computers
  - distributed data archive access/correlation
  - remote rendering and visualization
  - remote sensor and instrument coupling
- **Deploy a balanced, distributed system**
  - not a “distributed computer” but rather
  - a distributed “system” using Grid technologies
    - computing and data management
    - visualization and scientific application analysis

- **Define an open and extensible infrastructure**
  - an “enabling cyberinfrastructure” for scientific research
  - extensible beyond original sites with additional funding
    - NCSA, SDSC, ANL, Caltech, PSC, …

TeraGrid Components and Partners

- **Intel/HP Itanium Processor Family™ nodes**
  - Itanium2/3 IA-64 processors for commodity leverage
- **IBM Linux clusters**
  - open source software and community
- **Very high-speed network backbone**
  - high bandwidth for rich interaction and tight coupling
- **Large-scale storage systems**
  - hundreds of terabytes of secondary storage
- **Grid middleware**
  - Globus, data management, …
- **Next-generation applications**
  - breakthrough versions of today’s applications
  - but also, reaching beyond “traditional” supercomputing

Extensible TeraGrid Facility (ETF)



Caltech: Data collection analysis

0.4 TF IA-64
IA32 Datawulf
80 TB Storage

LEGEND

▨	Cluster	◆	Visualization Cluster
▨	Storage Server	▨	Shared Memory
▨	Disk Storage	▨	Backplane Router

ANL: Visualization

1.25 TF IA-64
96 Viz nodes
20 TB Storage

Extensible Backplane Network

LA Hub

Chicago Hub

40 Gb/s

30 Gb/s

30 Gb/s

30 Gb/s

30 Gb/s

30 Gb/s

4 TF IA-64
DB2, Oracle Servers
500 TB Disk Storage
6 PB Tape Storage
1.1 TF Power4

SDSC: Data Intensive

10+ TF IA-64
128 large memory nodes
230 TB Disk Storage
GPFS and data mining

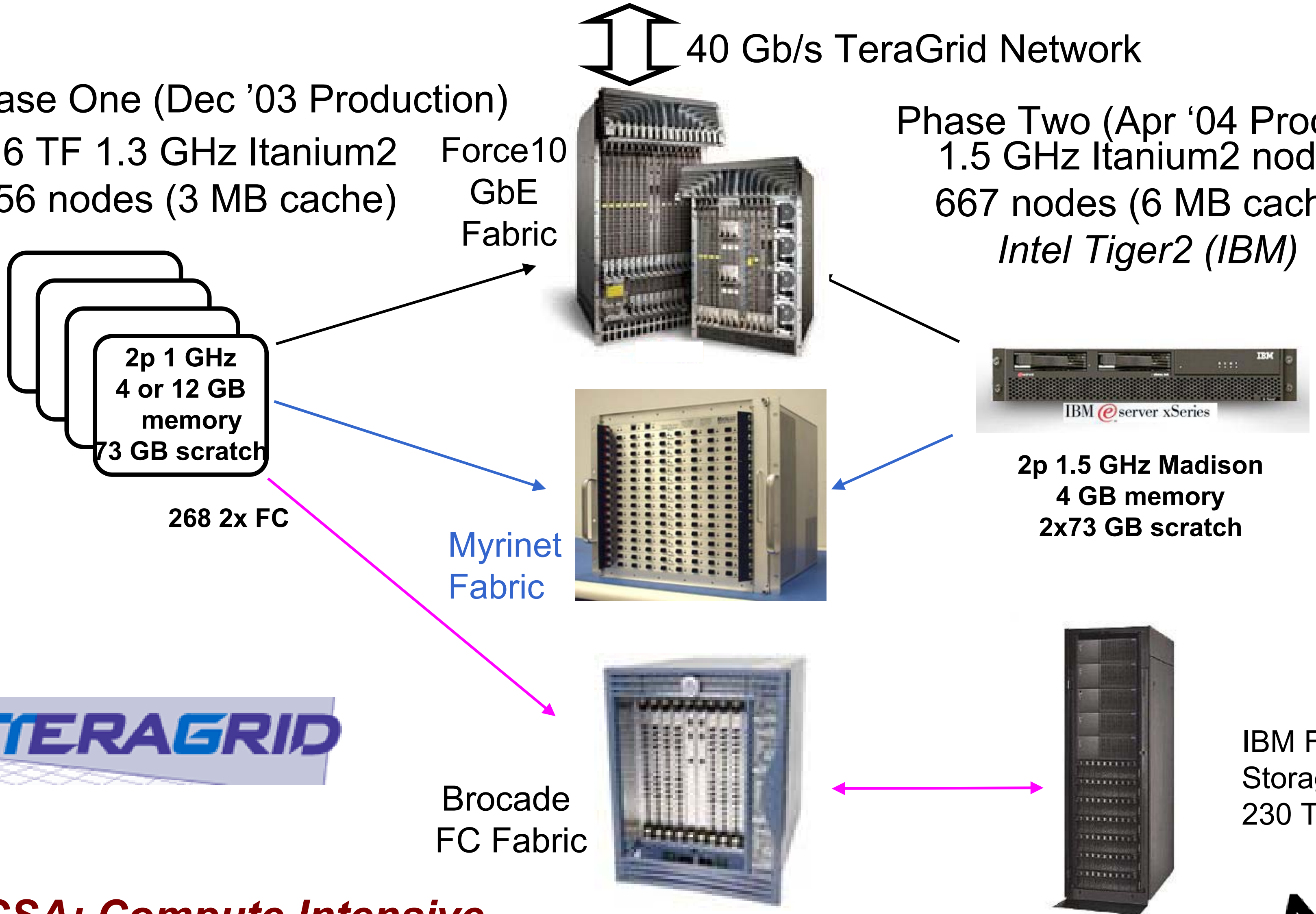
NCSA: Compute Intensive

6 TF EV68
71 TB Storage
0.3 TF EV7 shared-memory
150 TB Storage Server

PSC: Compute Intensive

Proposed 2002, Operational in 2003/2004

NCSA TeraGrid: 10.6 TF IPF and 230 TB



40 Gb/s TeraGrid Network

Phase One (Dec ’03 Production)
2.6 TF 1.3 GHz Itanium2
256 nodes (3 MB cache)

2p 1 GHz
4 or 12 GB memory
73 GB scratch

268 2x FC

Force10
GbE
Fabric

Myrinet
Fabric

Brocade
FC Fabric

Phase Two (Apr ‘04 Production)
1.5 GHz Itanium2 nodes
667 nodes (6 MB cache)
Intel Tiger2 (IBM)

2p 1.5 GHz Madison
4 GB memory
2x73 GB scratch

IBM FastT
Storage
230 TB

NCSA: Compute Intensive

TeraGrid Network Backbone

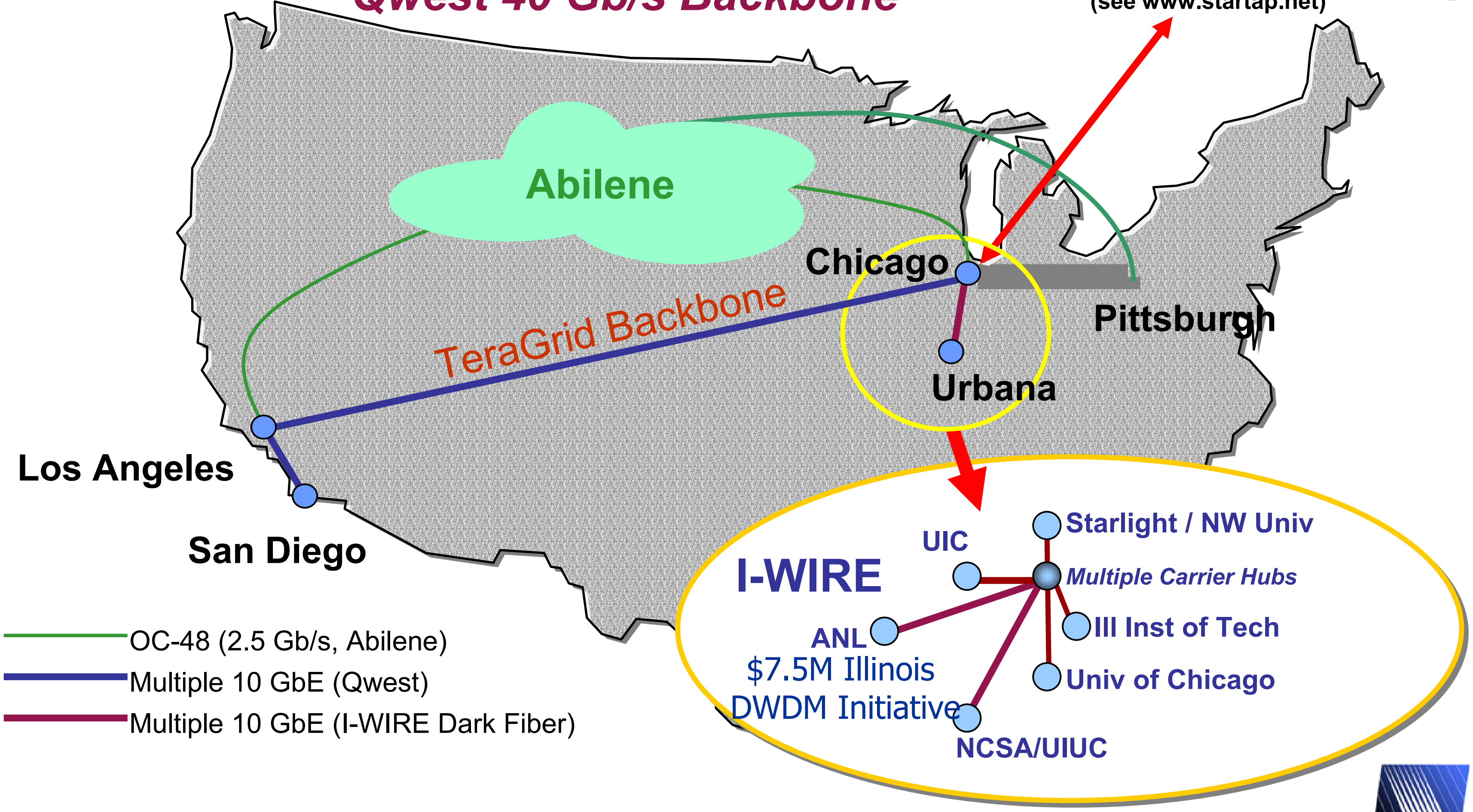
ride the light

Qwest

Qwest 40 Gb/s Backbone

StarLight

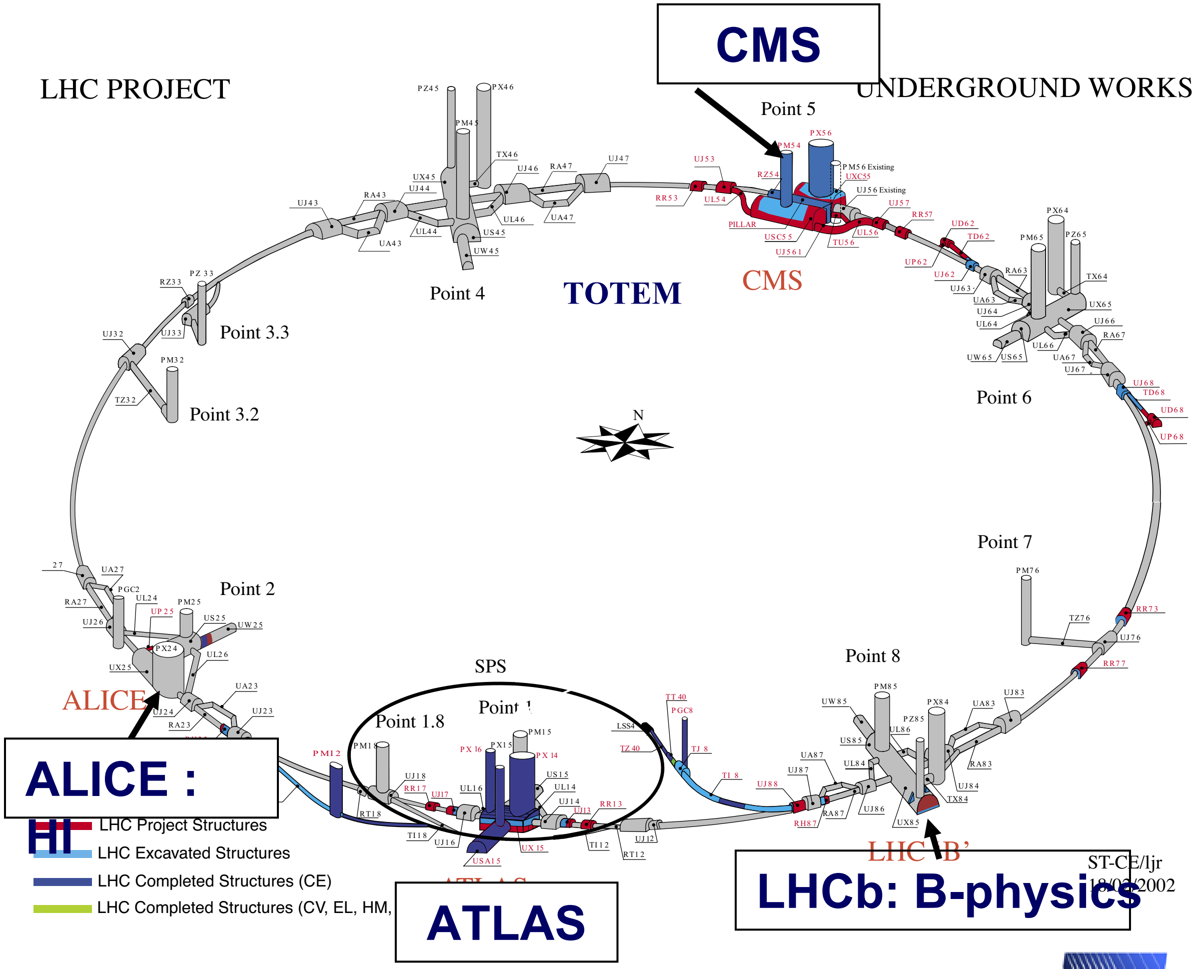
International Optical Peering
(see www.startap.net)



$7.5M Illinois DWDM Initiative

- OC-48 (2.5 Gb/s, Abilene)
- Multiple 10 GbE (Qwest)
- Multiple 10 GbE (I-WIRE Dark Fiber)

Large Hadron Collider: 2007



CMS

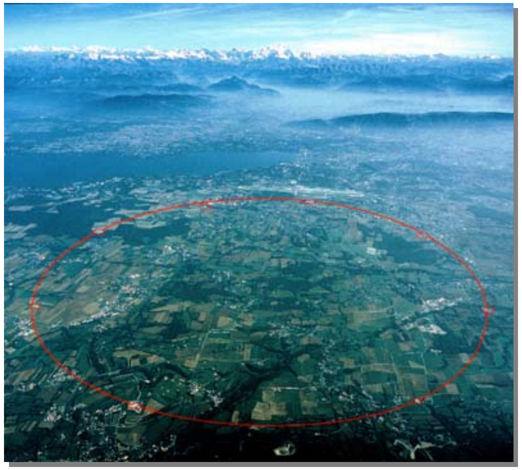
TOTEM

ALICE : HI

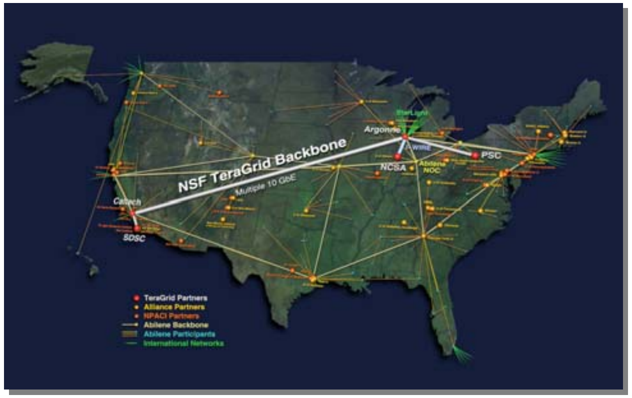
ATLAS

LHCb: B-physics

CMS Data Preparations

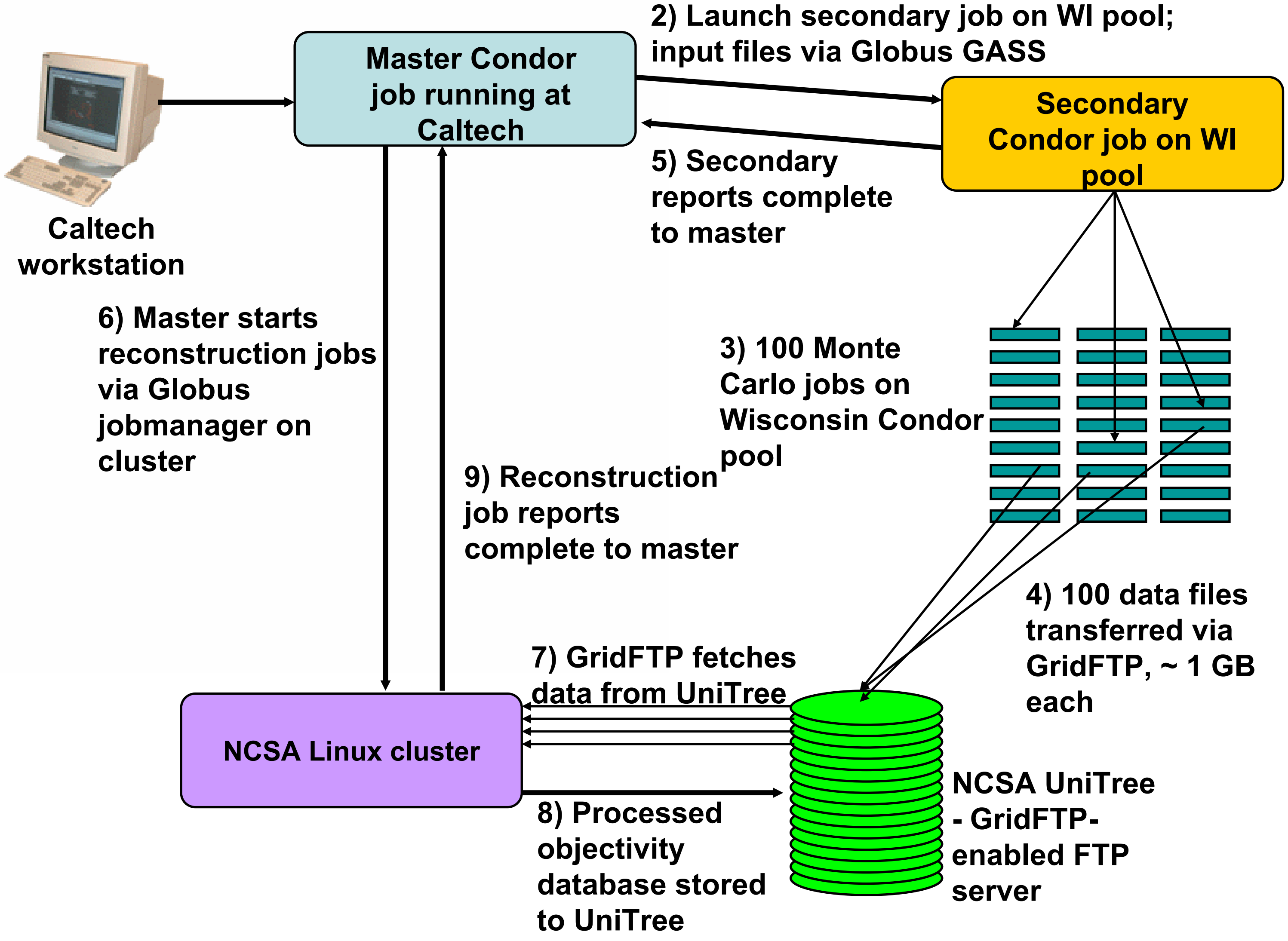


- **Data and software testing challenge**
  - test and validate analysis software
    - 100,000,000 events
- **Testing approach**
  - particle-detector interaction simulator (CMSIM)
    - energy deposition in the detector
  - ORCA (Object Reconstruction for CMS Analysis)
    - reconstruct QCD background sample
  - tracks and reconstructed particles, ready for analysis



- **Computing, storage and networking**
  - 2,600,000 SUs on the TeraGrid
    - 400 processors through April 2005
  - 1,000,000 SUs on IA-32 cluster
  - 1 TB for production TeraGrid simulations
    - 400 GB for data collection on IA-32 cluster
  - 2-5 MB/s throughput between NCSA and Caltech

GriPhyN: CMS Data Reconstruction



Source: Scott Koranda, Miron Livny and many others

Building Something New

One Organization
(merge institutions)

- One sysadmin team
- One management tea
- D ted mac ine r m, c tralize c trol
  - e.g. Googl ters

Not a Grid

The TeraGrid
(A Grid hosting environment)

- Single development environment
- Single stack to learn
- Develop here, run there
- Run here, store there

TERAGRID

Applications are developed for the Grid because the barriers are low and the return large

Very Loose Collaboration
(current situation)

- Different MPIs
- Hit-a miss id software:
  - G us vers n?
  - Cond G?
  - G2?
- Unique opment environment

Not a Grid, but with significant user investment, Grid applications can be developed

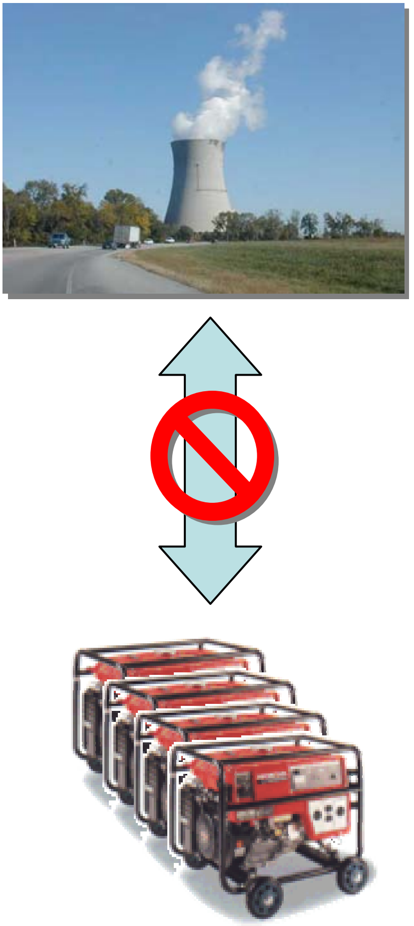
Source: TG Team

National Center for Supercomputing Applications

NCSA

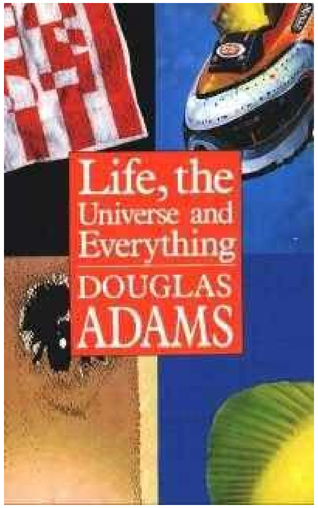
Grids and Capability Computing

- ***Not an "either/or" question***
  - *each addresses different needs*
  - *both are part of an integrated solution*
- **Grid strengths**
  - coupling *necessarily* distributed resources
    - instruments, archives, and people
  - eliminating time and space barriers
    - remote resource access and capacity computing
  - *Grids are not a cheap substitute for capability HPC*
    - *the latency/bandwidth continuum rules*
- **Capability computing strengths**
  - supporting foundational computations
    - terascale and petascale "nation scale" problems
  - engaging tightly coupled teams and computations



Outline

- **The really big questions**
  - – life, the universe, and everything
    - why they matter and how we react
- **TeraGrid and NCSA**
  - – status and directions
    - lessons and capabilities
    - applications and needs
- **High-end futures**
  - – petascale system design
    - challenges and opportunities
  - – international networks



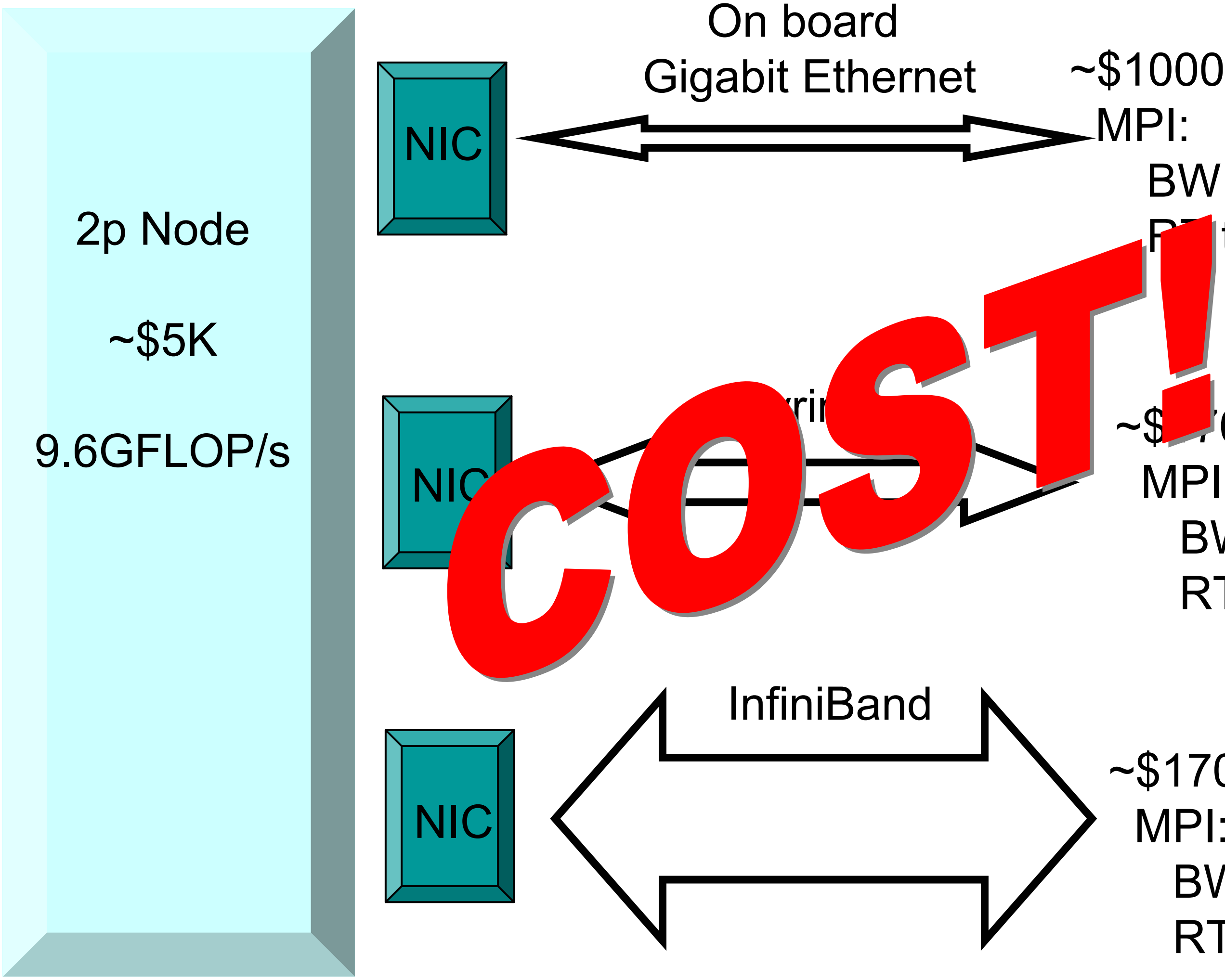
High-end Computing Challenges

- **Funding and long-term R&D**
  - ☺ and ☹
- **Time to solution**
  - too difficult to program and to optimize
  - better programming models/environments needed
- **Often, efficiency declines with more processors**
  - adversely affects time to solution and cost to solution
- **Support overhead for system parallelism**
  - management of large-scale concurrency
- **Processor-memory latency and bandwidth**
  - can be constraining for HEC applications
    - scatter-gather and global accesses
- **I/O and data management**
  - volume and transfer rates
- **Power consumption, physical size and reliability**

HPC Clusters

- ✓ **Processors**
  - ✓ x86, Itanium, Opteron, …
- ✓ **Memory systems**
  - ✓ the jellybean market
  - memory bandwidth ☹
- ✓ **Storage devices**
  - ✓ vibrant storage market
- **Interconnects**
  - ✓ Ethernet (10/100, GbE, 10GbE)
  - ✓ Infiniband (maybe)
  - Myrinet, Quadrics, SCI, …☹

Node Interconnects (2003)



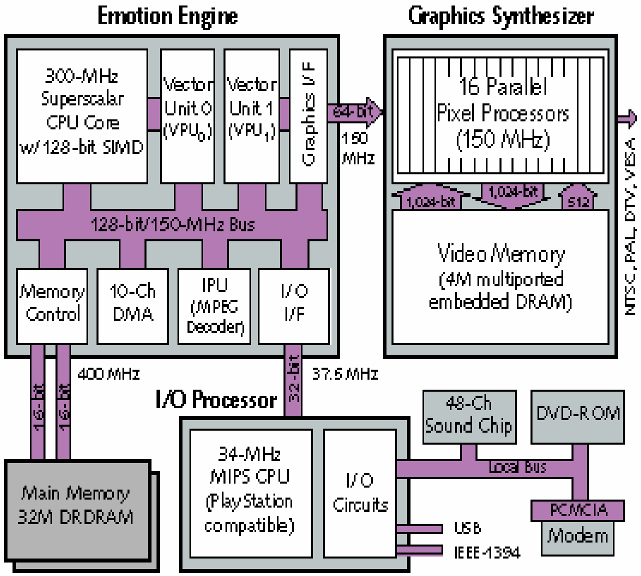
~$1000/port
MPI:
BW ~115 MB/s
RT time ~52 usec

~$[illegible]700/node
MPI:
BW ~220 MB/s
RT time ~12 usec

~$1700-2000/node
MPI:
BW ~760 MB/s
RT time ~12 usec

Computing On Toys

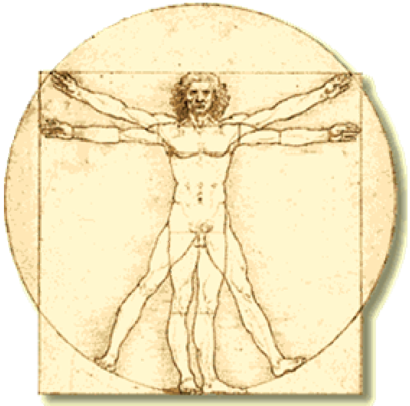
- **Sony PlayStation2 features**
  - 6.2 GF peak
  - 70M polygons/second
  - 10.5M transistors
  - superscalar RISC core
  - plus vector units, each:
    - 19 mul-adds & 1 divide
    - each 7 cycles
  - *$199 retail*
    - *loss leader for game sales*



- **70 unit cluster at NCSA**
  - Linux software and vector unit use
    - over 0.5 TF peak but difficult to program
  - vector assembly code
    - linear algebra libraries (BLAS1, 2, 3)
    - adaptive version selection
  - application porting atop vector code
    - MILC QCD (conjugate gradient dominated)
      - primary PACI cycle consumer

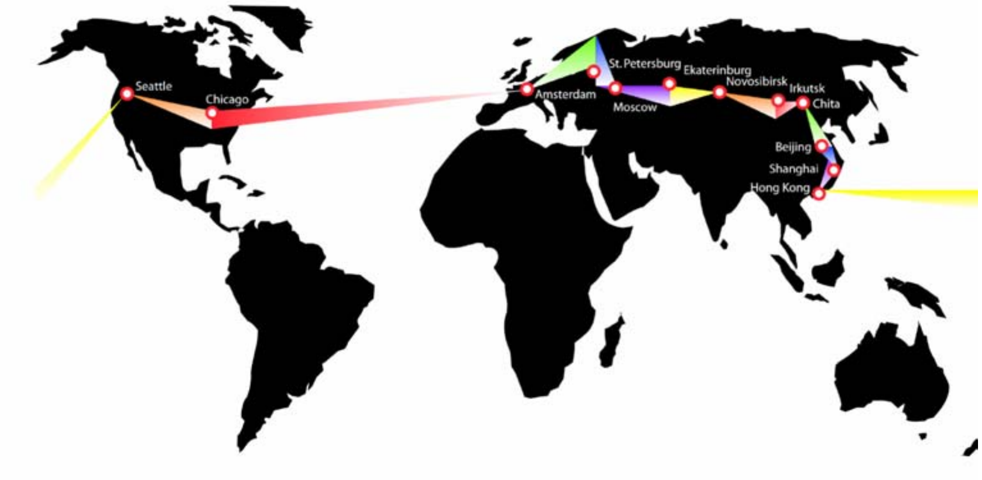
Trans-Petascale Vision

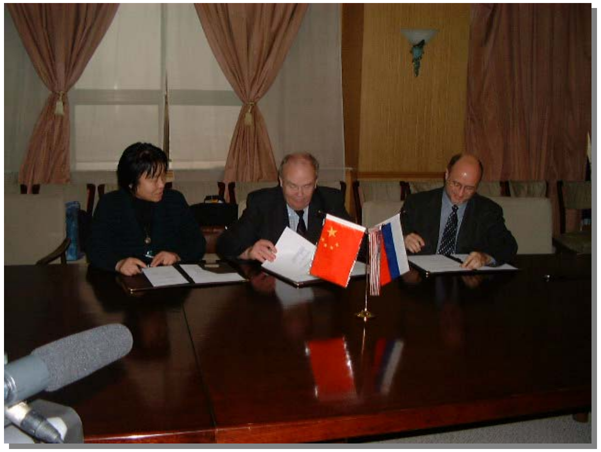
- **Multiple petabyte data archives**
  - 1-10 petabytes of secondary storage
  - tens to hundreds of petabytes of tertiary storage
- **DWDM terabit wide area networks (WANs)**
  - hundreds to thousands of lambdas
    - each operating at >10 Gb/s
- **Petascale computing systems**
  - dense, low-power packaging
  - memory access optimized
- **Responsive environments**
  - ubiquitous, mobile information sharing
- **Coupled by distributed Grid infrastructure**



GLORIAD

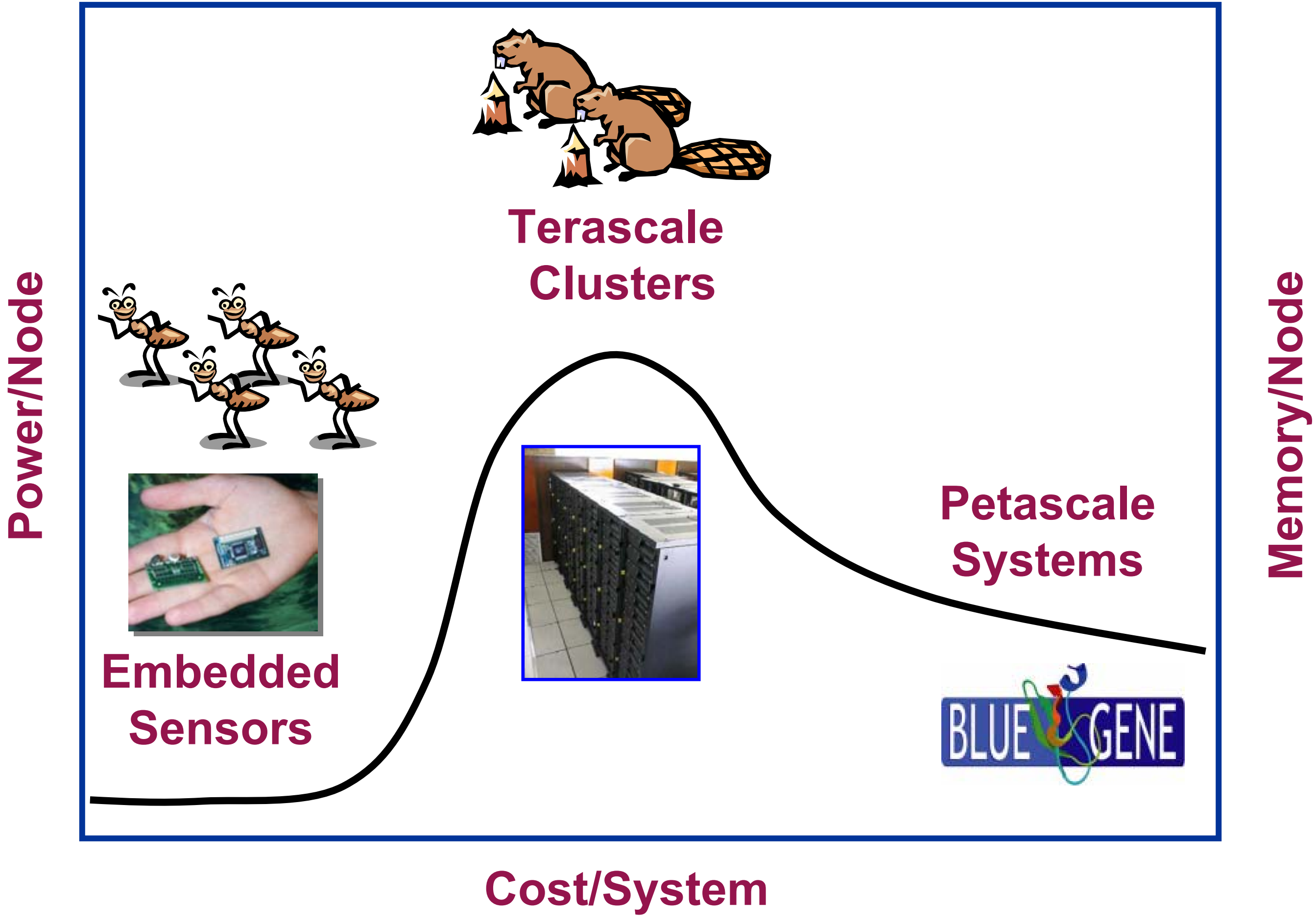
- **GLObal RIng Network for Advanced Applications Development**
  - Russia-China-USA science and education Network
    - 10 Gb/s transglobal network led by NCSA
    - builds on NCSA-led USA-Russia NaukaNet project
- **Funding (beginning in 2004, we hope)**
  - cooperatively funded by USA, China and Russia.
    - USA commitment anticipated at $2.5M annually
  - example applications
    - HEP, ITER, IVO, climate change, nanomaterials





The Computing Continuum

It’s weird (and cool) all along the curve!



Terascale Clusters

Embedded Sensors

Petascale Systems

Power/Node

Memory/Node

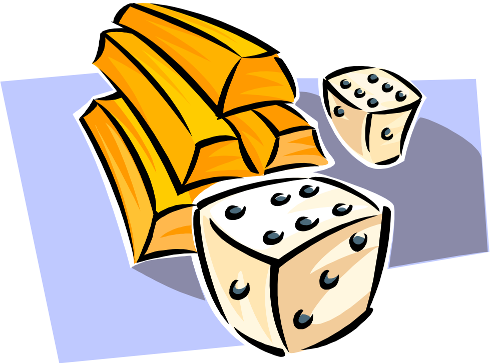
Cost/System

Building A Petaflop System

- **Technology trends**
  - dual-core processors
    - IBM Power4 and SUN UltraSPARC IV
    - quad-core is coming …
  - reduced power consumption (e.g, Intel Banias)
    - laptop and mobile market drivers
  - increased I/O and memory interconnect integration
    - PCI Express, Infiniband, …
- **Let’s look forward five years to 2008**
  - 4-way or 8-way cores (4 or 8 processors/chip)
  - ~10 GF cores (processors)
  - 4-way nodes (4, 4-way cores/node)
  - Infiniband-like interconnect

Building A Petaflop System

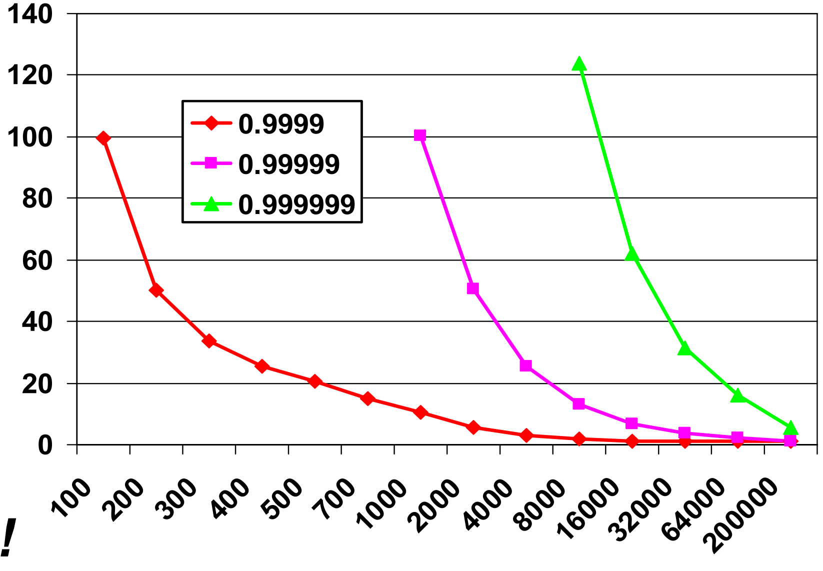
- **With 10 GF/processors**
  - 100K processors are required
  - 6200 nodes (4-way with 4 cores each)
- **Power consumption**
  - more than a portable generator
  - but, quite a bit less than a nuclear plant
- **Software challenges**
  - reliability and recovery



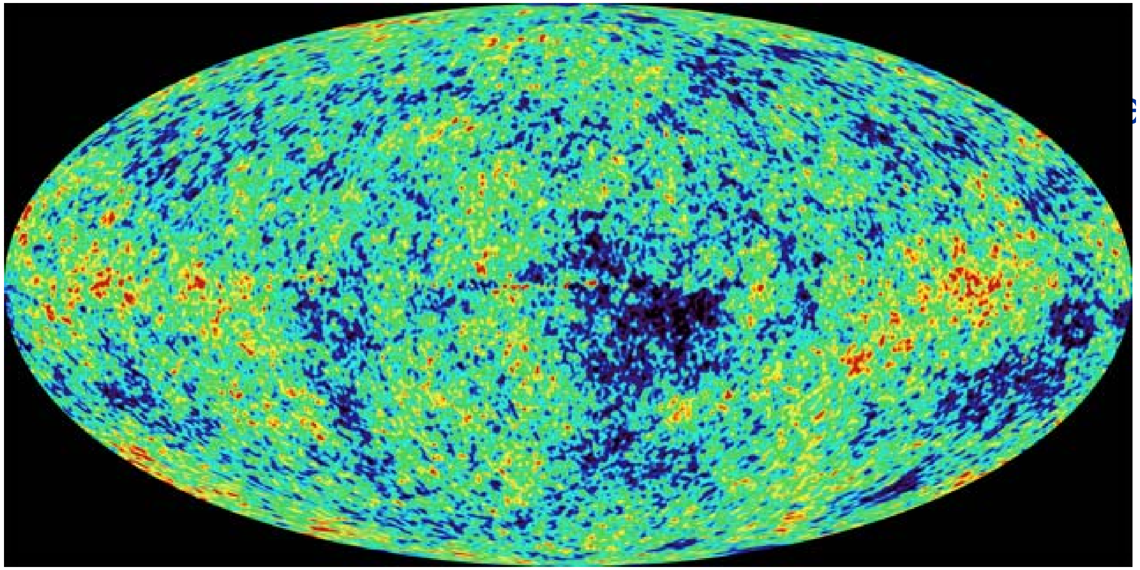
- **Cost of a petaflop system, O($100M)**
  - value of scientific breakthroughs … priceless

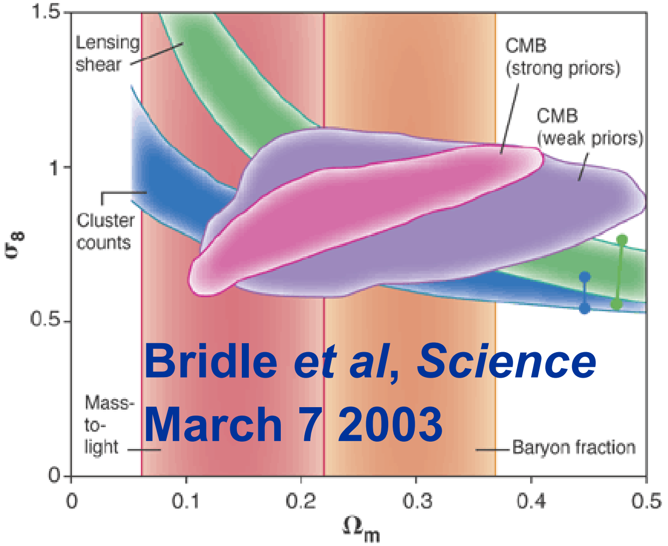
Very Large Scale Implications

- **Single node failure during application execution**
  - causes blockage of the overall simulation
  - data is lost and must be recovered/regenerated
  - key physics require neighbor exchanges
  - each spatial cell exists in one processor memory
- **~10^6 hours for component MTTF**
  - sounds like a lot until you divide by 10^5!
- ***It's time to take RAS seriously***
  - *systems do provide warnings*
    - *soft bit errors – ECC memory recovery*
    - *disk read/write retries, packet loss*
  - *status and health provide guidance*
    - *node temperature/fan duty cycles*
- ***We have to expect components to fail!***
- **Software and algorithmic responses**
  - diagnostic-mediated checkpointing and algorithm-based fault tolerance
  - domain-specific fault tolerance and loosely synchronous algorithms

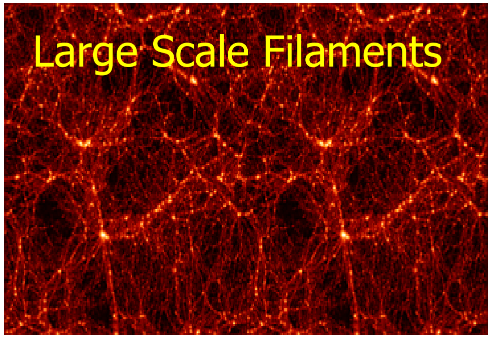


Physics Challenges





Bridle *et al*, *Science* March 7 2003



    - dynamic field (quintessence)
- **Possible missing mass candidates**
  - baryonic
    - Massive Compact Halo Objects (MACHOS)
  - non-baryonic
    - Weakly Interacting Massive Particles (WIMPS)
- **Recent experiment/theory interactions**
  - Wilkinson Microwave Anisotropy Probe (WMAP)
    - universe is 13,400 ± 300 million years old and flat

Discipline Coupling

Ask The Big Questions

Our immediate neighborhood we know intimately. But with increasing distance our knowledge fades. ...The search will continue. The urge is older than history. It is not satisfied, and it will not be denied.

Edwin Hubble