



# Building the Teraflops/Petabytes Production Supercomputing Center at NERSC

*Horst D. Simon*

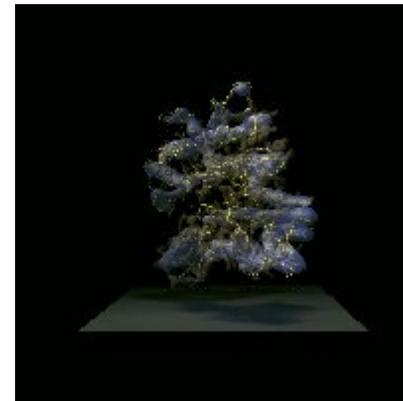
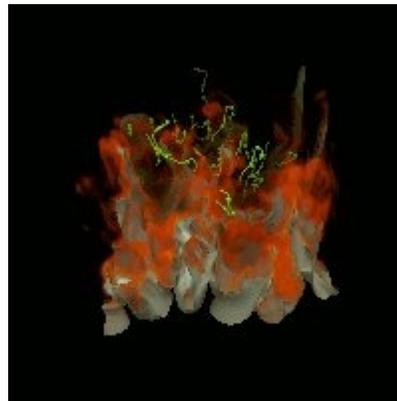
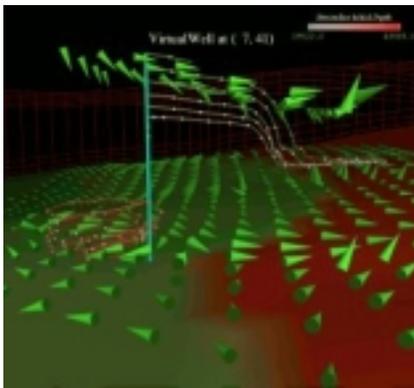
Director, National Energy Research Scientific  
Computing Center (NERSC)

Berkeley, CA 94720

[hdsimon@lbl.gov](mailto:hdsimon@lbl.gov)

November 11, 1999

- **the** Department of Energy, Office of Science, supercomputer facility
- unclassified, open facility; serving >2000 users in all DOE mission relevant basic science disciplines
- 25th anniversary in 1999

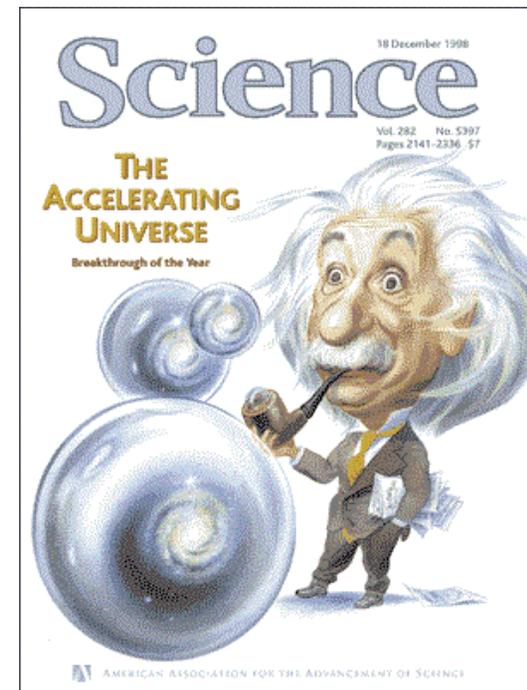


# NERSC Overview

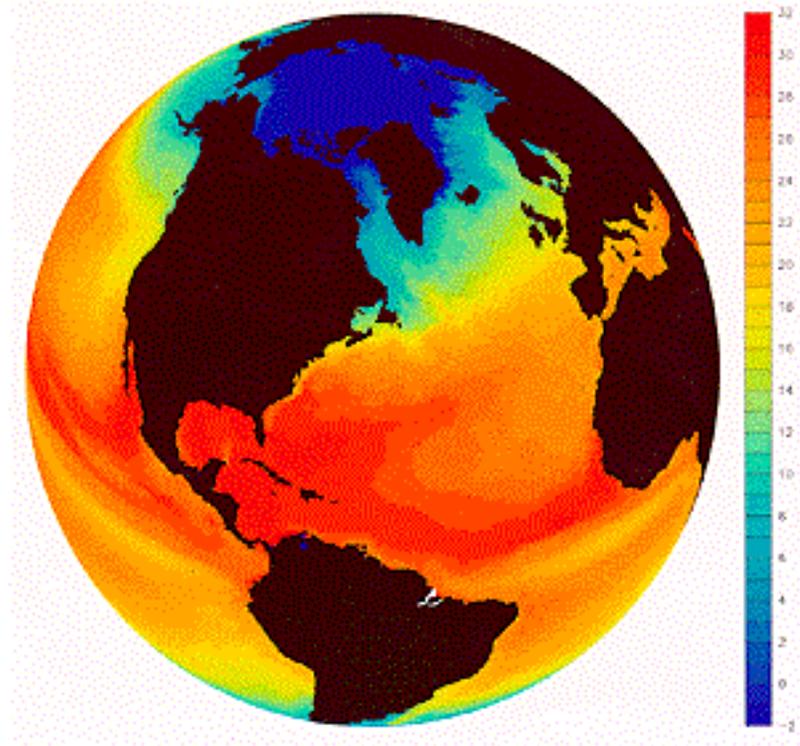
- Located in the hills next to University of California, Berkeley campus
- close collaborations between university and NERSC in computer science and computational science



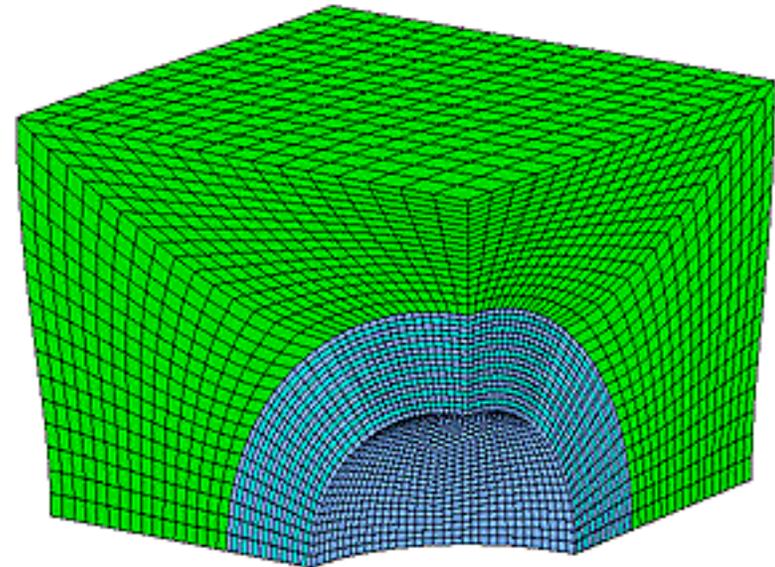
- Saul Perlmutter was co-recipient of Science magazine's "Breakthrough of the year award" (1998): universe expands at an accelerating rate
- This project is a "graduate" of the computational science program at LBNL (since 1996); LDRD funded; started with the arrival of NERSC



- Warren Washington, NCAR
- PCM - parallel climate model
- Sustained 17 Gflop/s on NERSC T3E
- Highest sustained performance on climate model in the U.S.



- Mark Adams and Jim Demmel, UC Berkeley developed Prometheus, a parallel multigrid solver package for unstructured finite element problems in solid mechanics
- Recipients of the 1999 “Carl Benz Award” , awarded by Daimler-Chrysler for the best industrial applications development on parallel machines



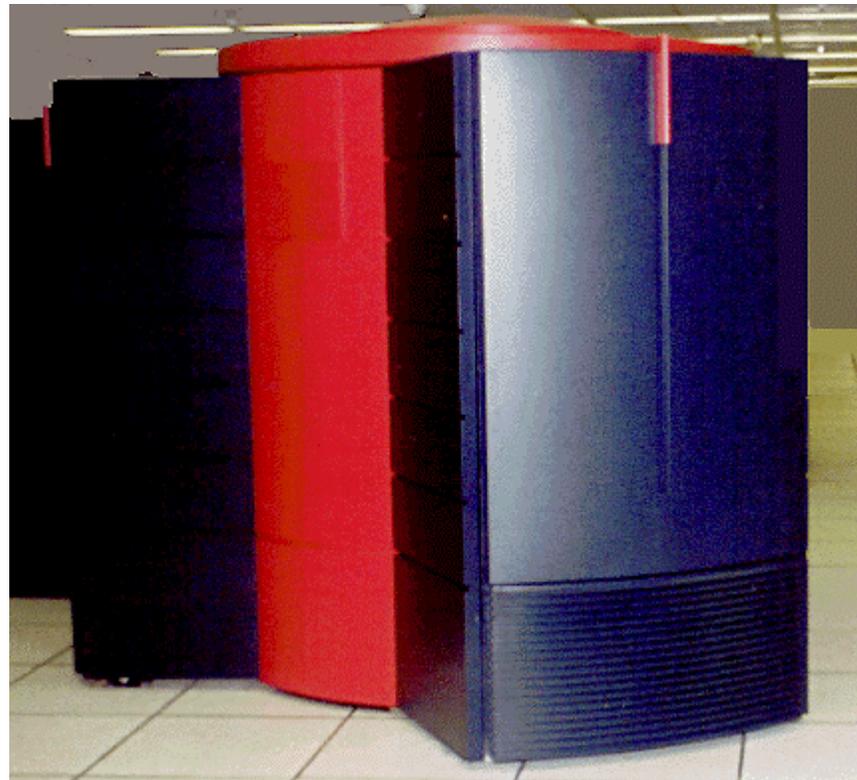
Model of a sphere within a soft material with parameterized mesh

- Overview
- Two technology transitions in the 1990s
- Clusters of SMPs as production platforms
- Petabyte data challenges
- Integration into the data grid
- Building a high performance organization
- Conclusions

# NERSC-1

## Cray C90 installed in Dec. 1991

- Cray C90 installed in December 1991
- ended contract with CCC for a Cray-3
- stable high end production platform for seven years until 12/31/98



# NERSC- 2

## Cray T3E-900 installed in 1996

The 644 processor T3E-900 is one of the most powerful unclassified supercomputers in the U.S.

- eight out of twelve DOE Grand Challenge Projects compute at NERSC
- 50% of the resource dedicated to GC projects
- about 100 other projects allocated on the NERSC T3E-900
- 1997 GAO report judged NERSC to have the best MPP utilization (75%) -- 1999 utilization >90%





# NERSC-3 IBM SP3 installed in 6/99



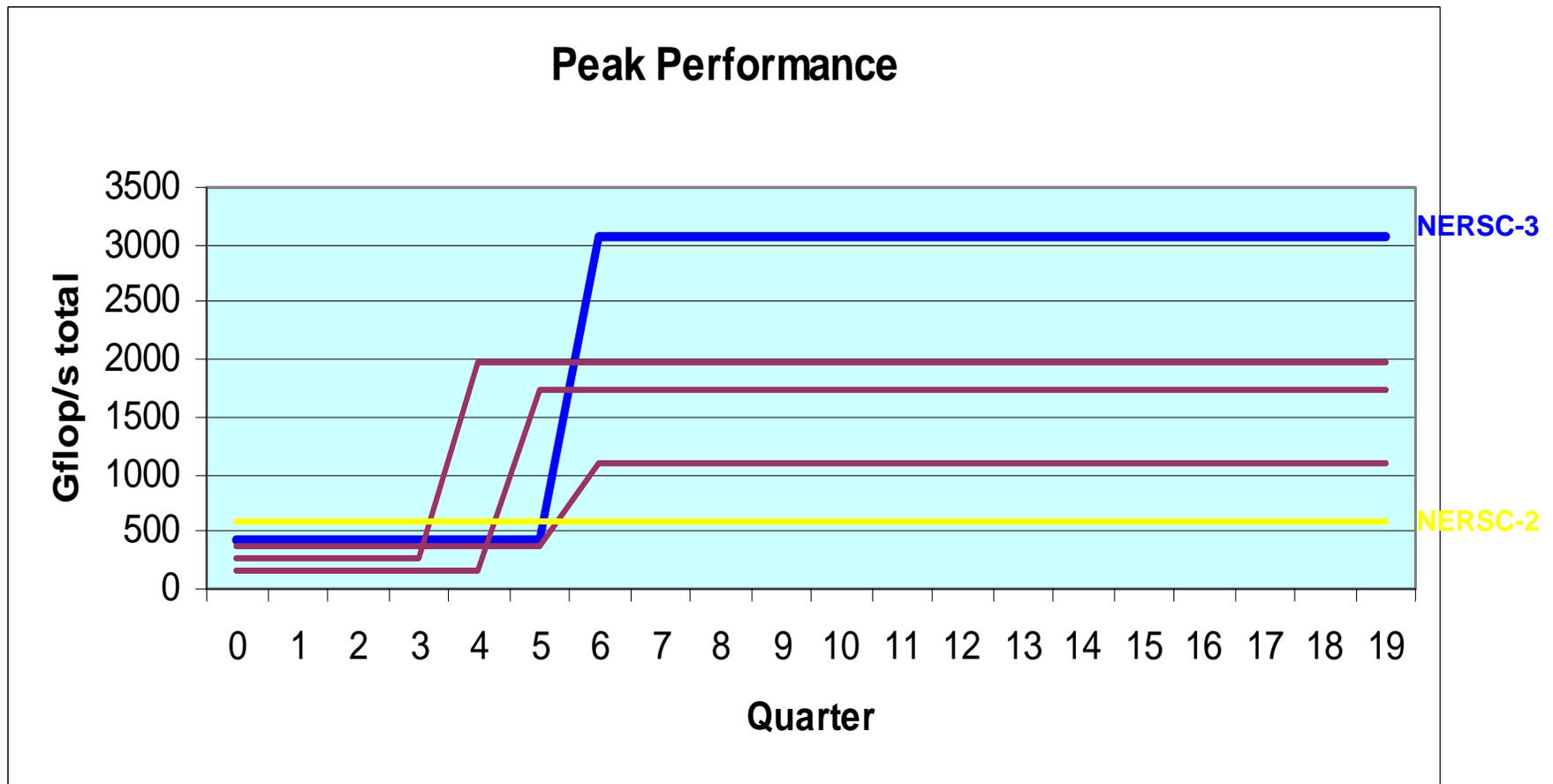
- New contract with IBM announced in April 1999
- IBM was clearly the best value for the primary award
  - provides the best absolute performance
  - has lowest absolute cost
  - provides the best price performance
  - provides acceptable functionality
  - guarantees performance - low risk

# NERSC-3 Supercomputer

- IBM selected to provide NERSC-3 (IBM SP3/RS 6000)
- Phase I: June 1999 installation
  - 608 processors
  - 410 gigaflop peak performance
  - Provides one teraflop NERSC capability
- Phase II: December 2000 completion
  - 2,432 processors
  - 3.2 teraflop peak performance
  - 4 teraflop total NERSC capability



# NERSC-3 Peak Performance Comparison

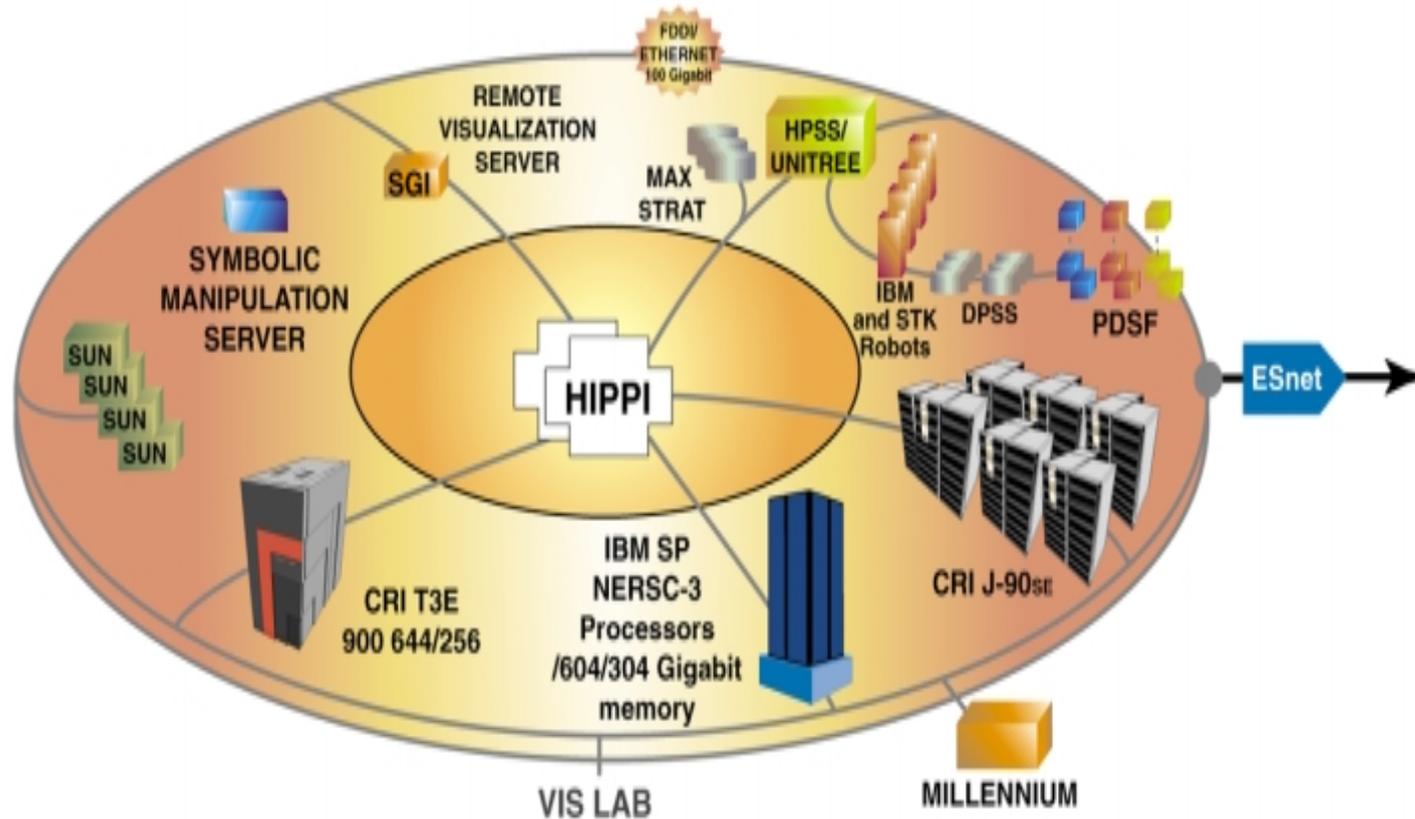


# NERSC-3 Node Configuration

	Phase 1	Phase 2a	Phase 2b (If different)
CPUs	Power 3	Power 3+	
CPUs per node	2	16	
Switch	TB3MX2	Next Generation SP Switch (aka "Colony") (Single-Single)	(Double-Double)
Memory per CPU	500 MB	500 MB	
Local System Disk per Node	18 GB - mirrored	18 GB-mirrored	
Clock Speed	200 MHz	>333 MHz	
Peak Speed per CPU	800 Mflop/s	> 1,330 Mflop/s	
Bandwidth per CPU	60 MB/s	40 MB/s	80 MB/s
Parallel Disk	10 TB - GPFS	20 TB - GPFS	
Number of Nodes	304 Total	152 Total	
Number of Compute Nodes	264 (256+8)	130 (128+2)	
Network Interfaces	6 HiPPI 2 Gigabit Ethernet 1 ATM OC-3	16 HiPPI 6 Gigabit Ethernet 2 ATM OC-12, 1 OC-3	

## NERSC-3 and ASCI System Comparison

System Attributes	NERSC Phase 1	NERSC Phase 2	ASCI Blue	ASCI White
Total nodes	304 2-way SMP nodes	152 16-way SMP nodes	1464 4-way SMP nodes in 3 488-way systems	512 16-way SMP nodes
Total processors	608	2432	5856	8192
Processor technology	200 MHz Power 3	333+MHz Power 3	332 MHz Power PC 604e	310+MHz Power 3
Interconnect	SP Switch MX Adapter-2	next generation adapter 4 planes	SP Switch MX Adapter-2 within system SP Switch router between systems	next generation adapter 2 planes
Computer nodes	256 (512)	128 (2048)	1296 (5184)	472 (7552) or 488 7808
GPFS nodes	16 (32)	16 (256)	168 (672)	32 (512) or 16 (256)
Network nodes	8 (16)	2 (32)	x Power 2 processors	8 (128)
Interactive nodes	8 (16)	2 (32)		unknown
Service nodes	16 (32)	4 (64)	none	none
Node Memory	1-1.5 GB	8 GB	1.5-2.5 GB	8 GB
System Memory	.3 TB	1.2 TB	2.6 TB	4 TB
System Disk	10 TB	20 TB	62.5 TB	150 TB
System Peak	.5 Tflop/s	3.2 Tflop/s	3.9 TeraOPs	10.2 TeraOPs
Delivery	6/99	12/00	1/99	2/00





# HPC Systems at NERSC in the 90s



	<b>NERSC-1</b> Cray C90	<b>NERSC-2</b> Cray T3E	<b>NERSC-3</b> IBM SP-3
Year of Installation	1991	1996	1999
Number of Processors	16	640	2048
Processor Technology	Custom ECL	Commodity CMOS	Commodity CMOS
Peak System Perform.	16 Gflop/s	580 Gflop/s	3000 Gflop/s
Architecture	Shared memory, parallel vector	Distributed memory	128 nodes with 16 processor SMP
System	Fully integrated custom system	Fully integrated custom system with commodity CPU and memory	Loosely integrated system with commodity system components
System Software	Vendor supplied, ready on delivery	Vendor supplied, completed after nearly 3 years development	Vendor supplied, contractual complete in about 3 years
Floor space	588 ft	360 ft	4000 ft
Power consumption	500 kW	288 kW	1400 kW

# Impact of Technology Transitions

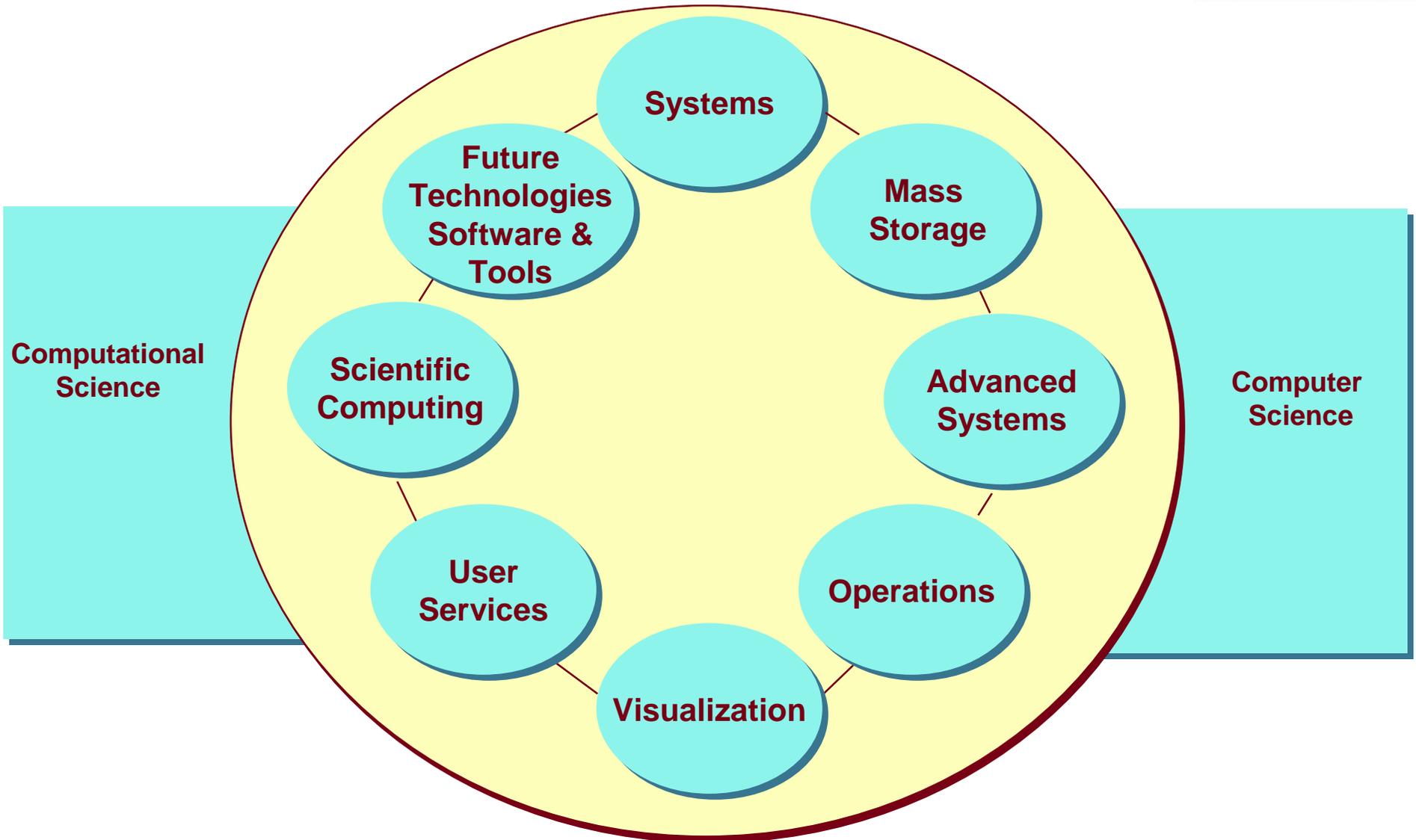
	<b>1994 – 1996 transition</b>	<b>1998 – 2000 transition</b>
Economic Driver	Price performance of commodity processors and memory	16 – 64 CPU “sweet spot” for SMP technology in the commercial market place
Advantages of transition	Higher performance and better price performance	Higher performance
Challenges of transition	<ol style="list-style-type: none"> <li>1) Applications transition to distributed memory, message passing model (MPI)</li> <li>2) More complex system software (scheduling, checkpoint restarting)</li> </ol>	<ol style="list-style-type: none"> <li>1) Applications transition to hierarchical, distributed memory model (threads + MPI)</li> <li>2) New development efforts for even more complex systems software</li> <li>3) Increased cost of facilities</li> </ol>

**Table 2.** Impact of the two technology transitions of the 1990s.

- In 1995-96 DOE and NSF competitively re-examined the role of centers:
  - Rapidly changing technology
  - Better local facilities everywhere
  - Growth of computational approaches in all disciplines
- New Model: Intellectual Services + a Major Facility
  - New algorithms and strategies developed in medium and long-term collaborations with scientific user community
  - The Center is the working interface between computer science and physical science



Necessary but  
not sufficient!



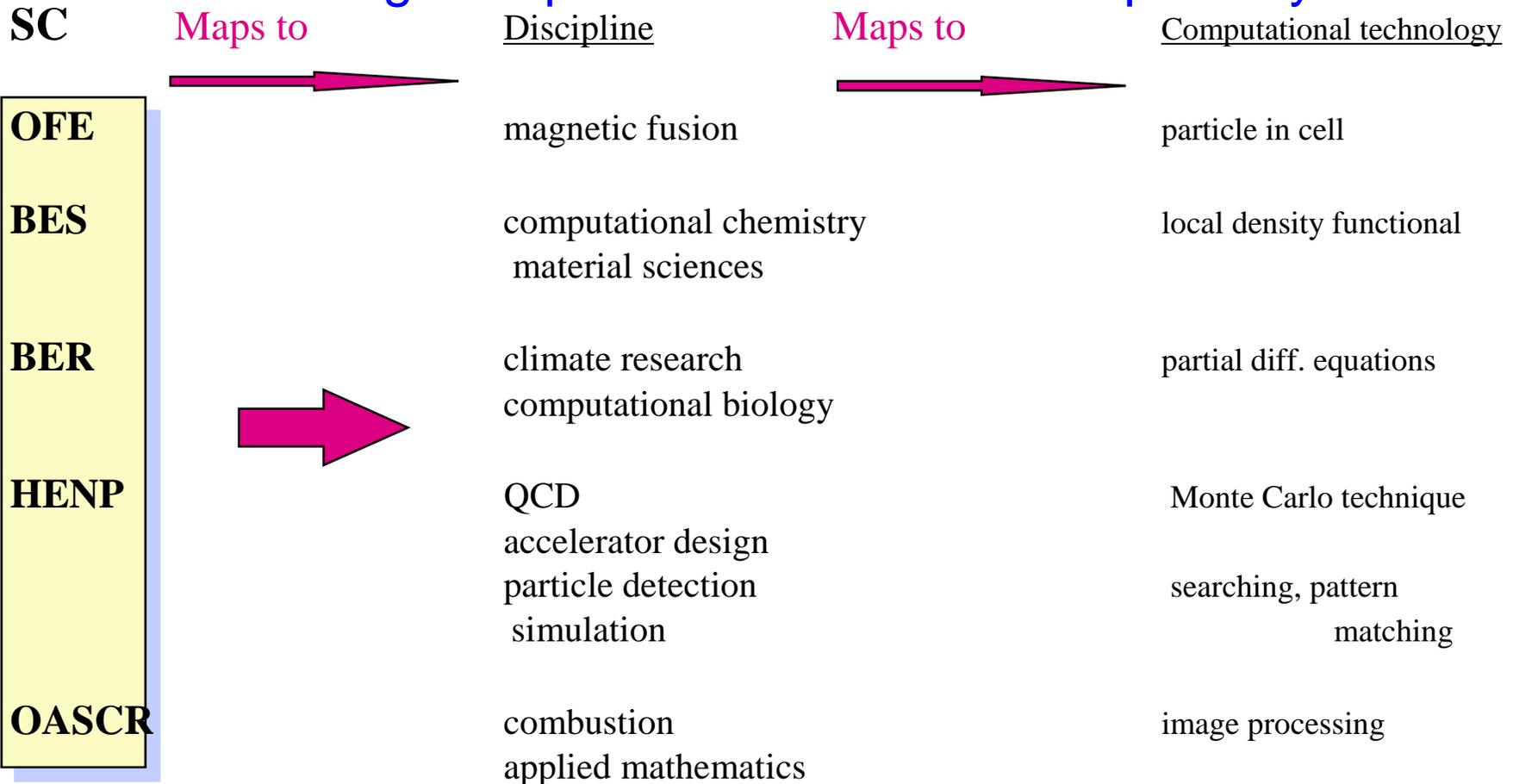
- Overview
- Two technology transitions in the 1990s
- Clusters of SMPs as production platforms
- Petabyte data challenges
- Integration into the data grid
- Building a high performance organization
- Conclusions

# Three Challenges

- applications that can tolerate an increase in communication latency and parallelism as well as a distributed, hierarchical memory model need to be written
- system software for increasingly complex, more difficult to manage, one-of-a-kind systems will have to be developed anew
- center management will be forced to take creative new approaches to solve the space and power requirements for the new systems.

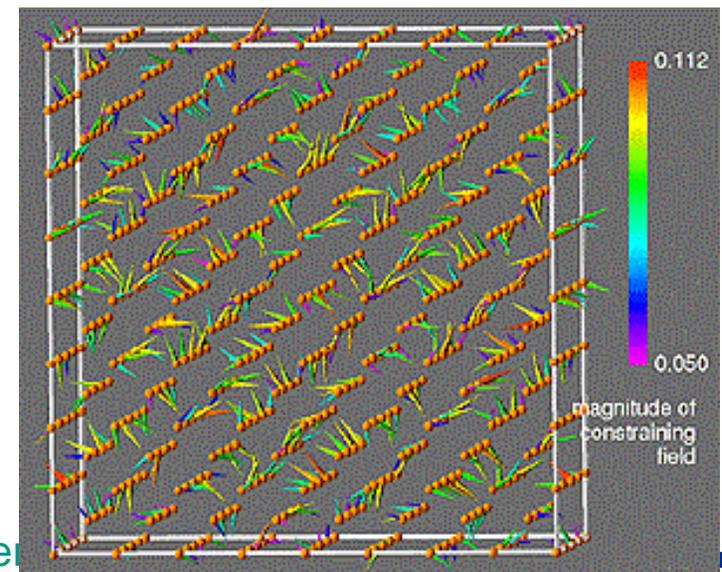
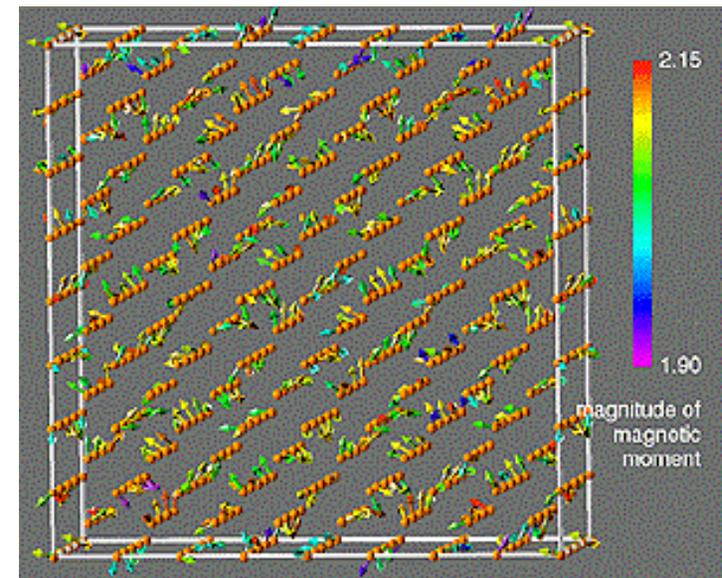
# Meeting the Applications Transition Challenge

## Building Computational Science Competency



NERSC has or will build competency in all technological areas of relevance to SC research

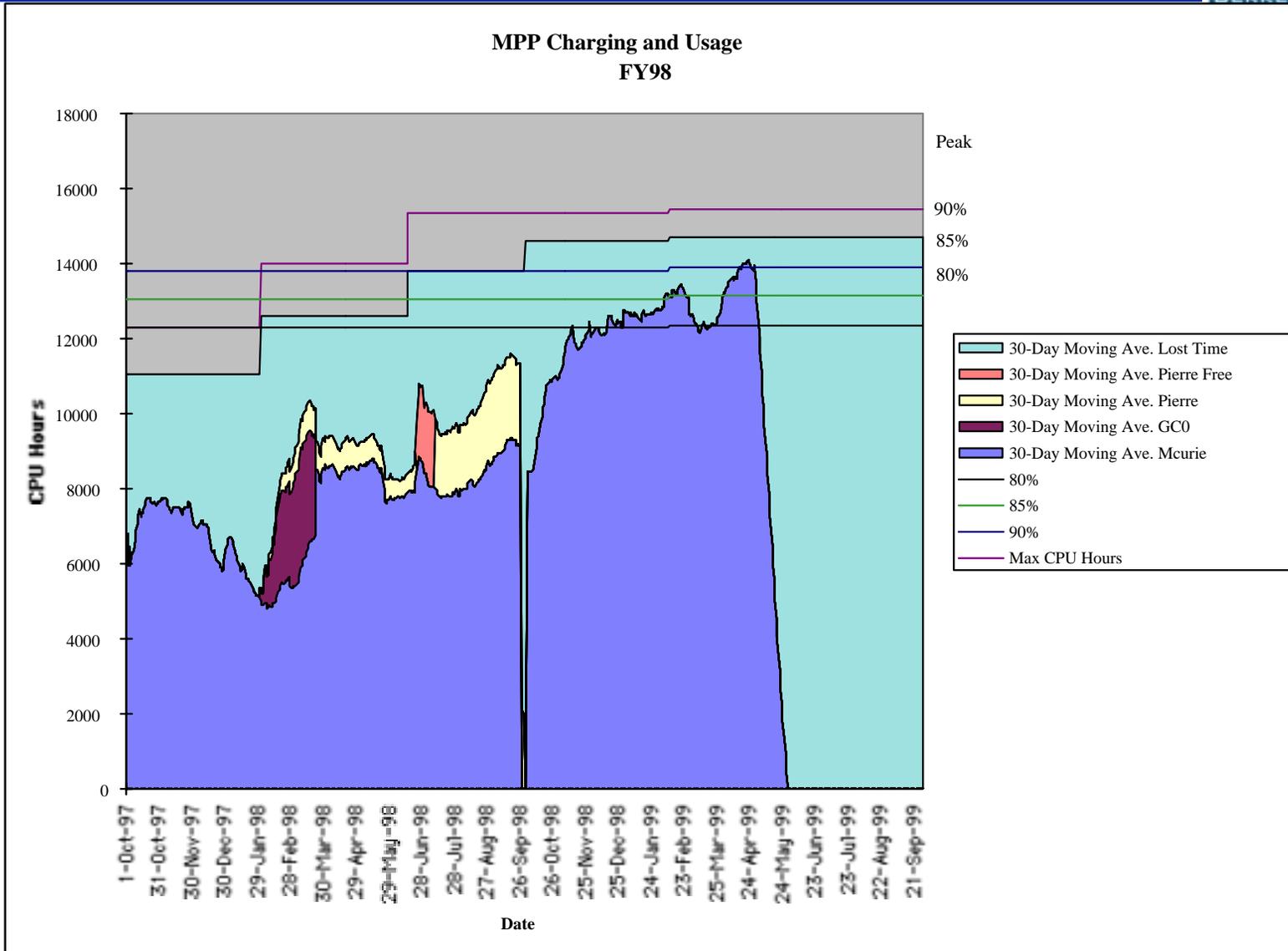
- 1998 Gordon Bell Prize for best performance of a parallel supercomputer application for a team of collaborators from DOE's Grand Challenge on Materials, Methods, Microstructure, and Magnetism.
- **Andrew Canning** (NERSC) made significant algorithmic contributions to this project, and was the key force behind the large scale simulations
- 1024-atom first-principles simulation of metallic magnetism in iron
- **first complete application to break the 1Tflops barrier.**



NERSC collaborated with vendors in the past and had major breakthroughs

- first site to demonstrate checkpoint/restarting on highly parallel platform (Sept. 97)
- >90% utilization on highly parallel platform due to Psched software and intensive tuning

# T3E Utilization > 90%



# Historical Perspective

- In the late '70s, DOE had intellectual leadership in High-End computing
- This ended in the mid-'80s with the increasing stability of PVP systems
- Today's MPP systems are getting harder to use, so DOE again needs to take a position of intellectual leadership.



# ASCI Approach: Above the line / Below the line

## QUOTED FROM PAUL MESSINA: 30 TERAOPS ASCI PROCUREMENT

Parallel HW/SW Boutiques, Universities,  
Labs supply items above the line:

**Economic Incentive**

- |                         |                               |
|-------------------------|-------------------------------|
| Parallel file system    | Profilers                     |
| Parallel debugger       | Resource management           |
| Visualization software  | Parallel scientific libraries |
| Message Passing (tuned) | Global file system            |



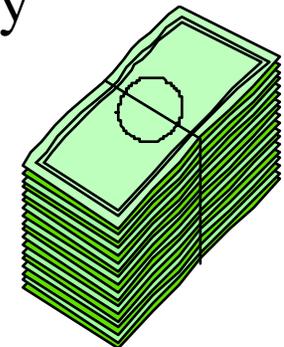
“The line”

### Computer Industry

- SMP servers (4-128 way)
- Compilers
- Operating System
- Message Passing

### Communications Industry

- Switch fabric
- Network interface cards



- NERSC will participate in any joint development effort
- but “line” is drawn too low
- lowering expectations too much, will let vendors increasingly deliver hardware without the necessary production system software
- NERSC and IBM have jointly developed an approach for developing high-quality, production system software

## However

- NERSC also invests significantly in PC cluster development (and production use) as potential NERSC-4 or NERSC-5 platforms



# NERSC-3: System Availability and Reliability

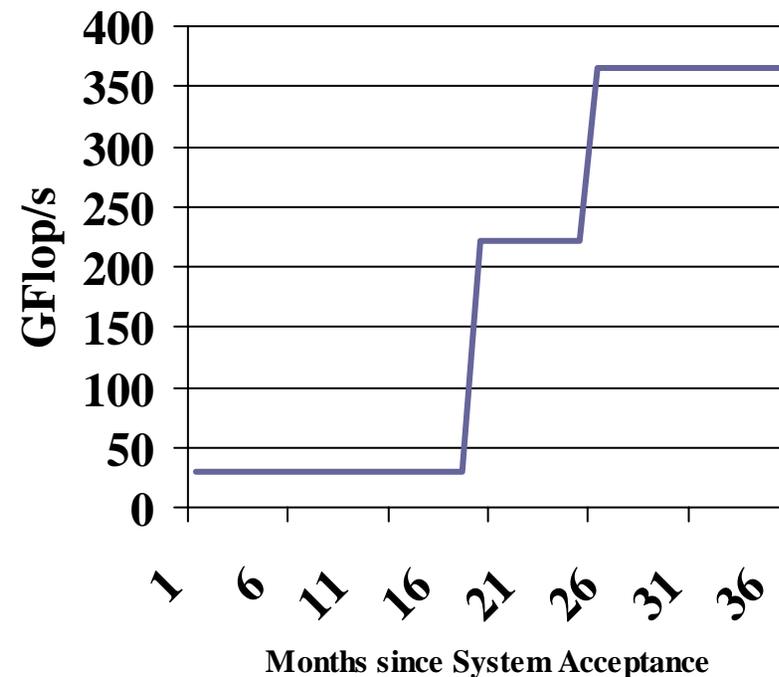


- First time IBM has agreed to Software and Hardware Availability and Reliability Measures
- First time NERSC has Availability and Reliability for the life of a contract
- System Availability of 94% annually on a node basis
  - Use Service Nodes as “hot spares” so there is always at least 128 compute nodes and all other functions operate
  - Service Nodes can also be used for rolling upgrades or put into service
- System Wide MTBF of 14 days

# NERSC-3 : Sustained System Performance Measure

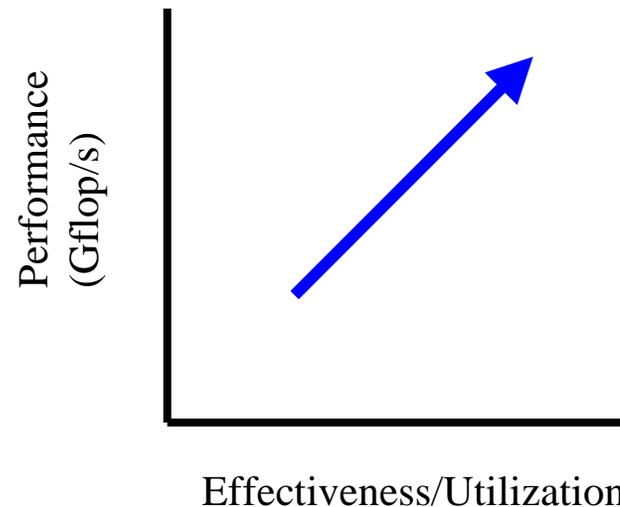
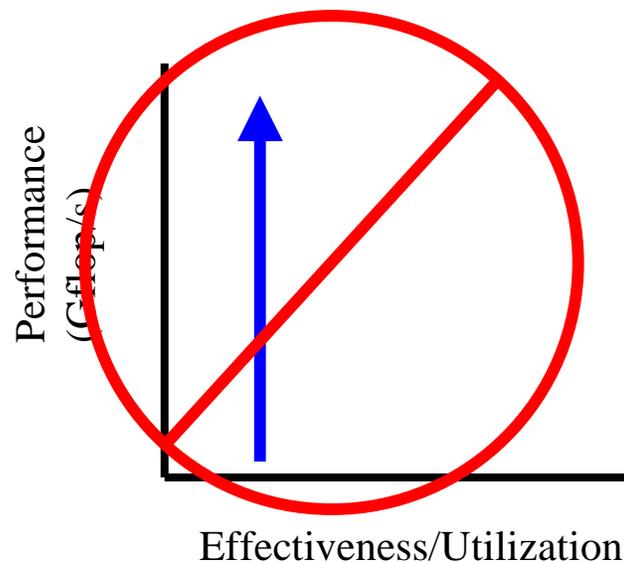
- NPBs are a tough but honest measure for vendors
  - NPBs indicate T3E is a 30 GFlop/s system yet Gordon Bell prize code runs at ~260 GFlop/s
  - NPBs typically indicate the lower level of what a good code should get
- Vendor projections are <130 Gflop/s but they committed to meet this measure
  - by faster CPUs,
  - earlier delivery of Phase 2a/b
  - more CPUs

NPB Average > 155 Gflop/s  
5.6 Tflop/s-months



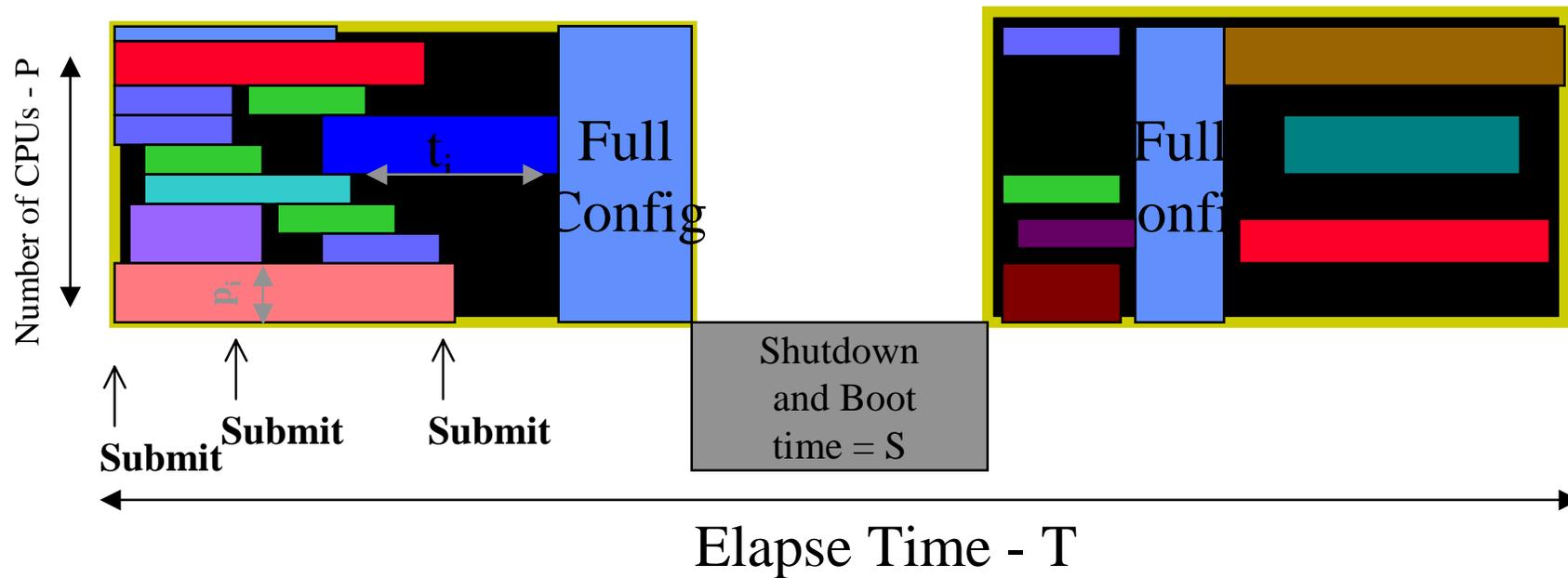
# Effective System Performance (ESP) Test

- This innovative test was added in negotiations to set goals for improving system utilization
  - Performance - how much scientific work can be done for a given quantum of CPU time
  - Effectiveness - How many quanta of CPU time can be made available to scientific programs



# ESP Goals

- Determine how well an existing system supports a particular scientific workload
- Assess systems for that workload before purchase
- Provide quantitative information regarding system enhancements
- Compare different systems on a single workload or discipline
- Compare system-level performance on workloads derived from different disciplines
- Compare different systems for different workloads



$$\text{Effectiveness} = (\sum_{i=1,N} p_i * t_i) / [P * (S + T)]$$



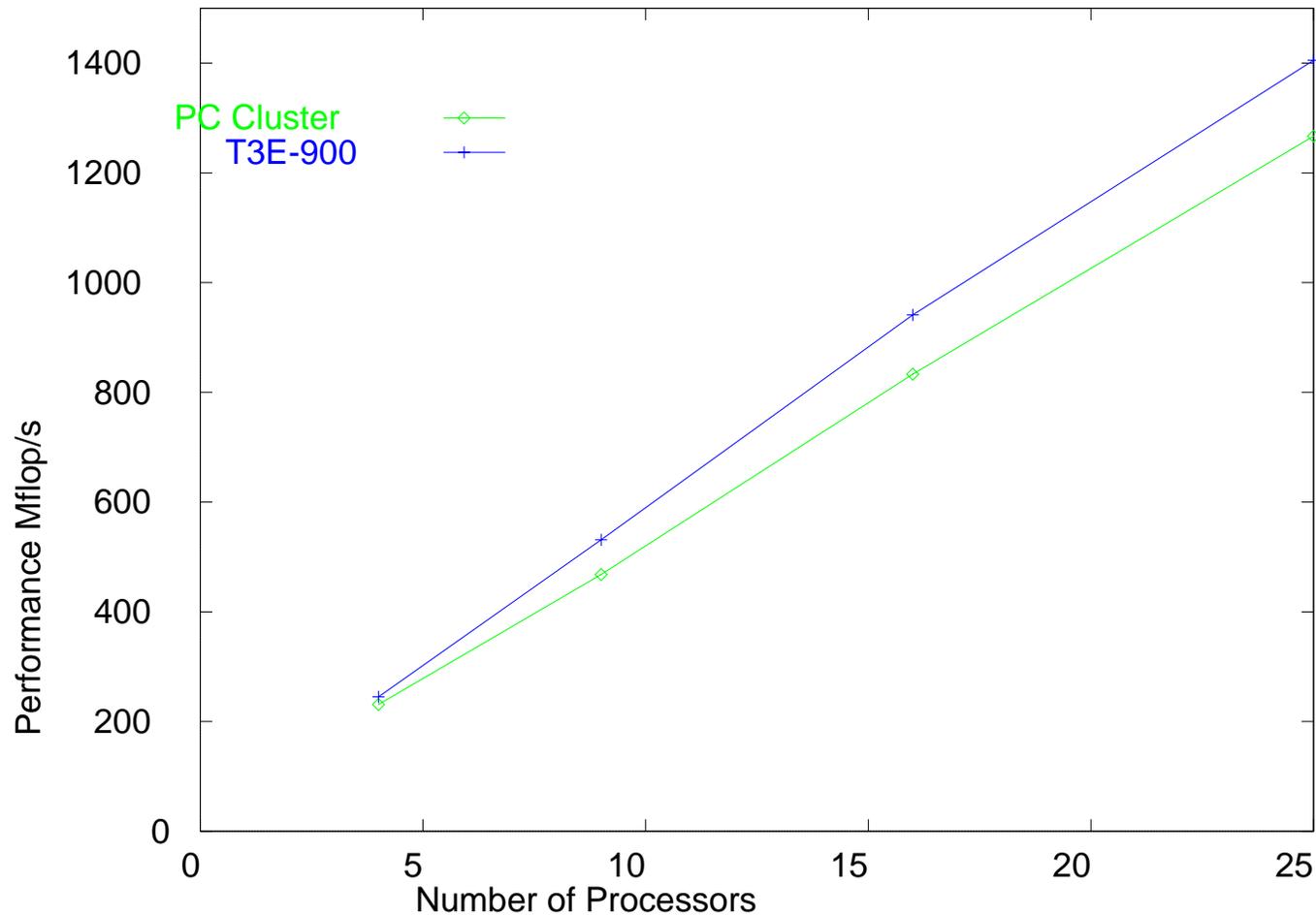
# NERSC has two PC Cluster Projects



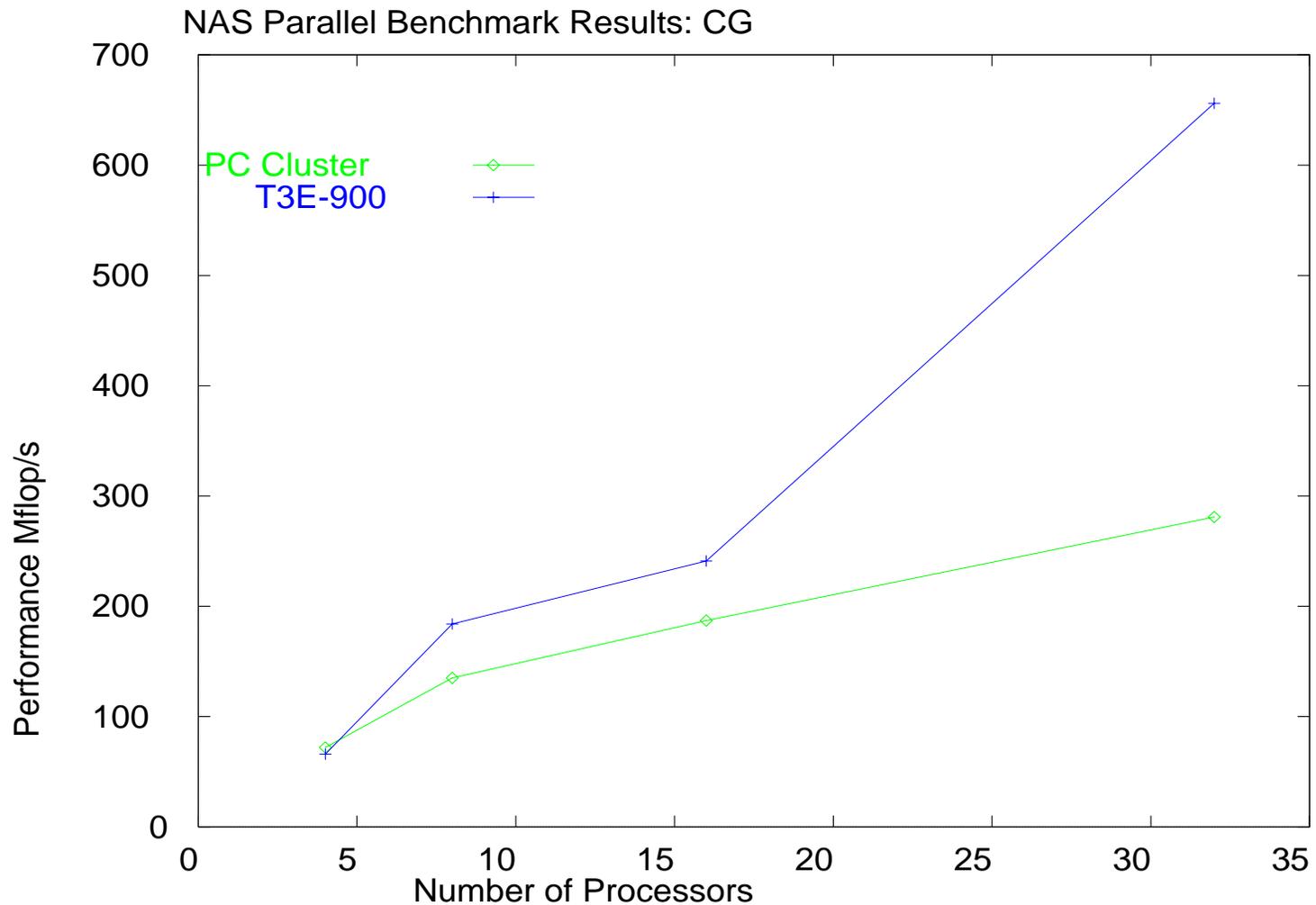
- Future Technologies Group
  - Developing/collecting software for “plug&play” clusters.
  - Working with NERSC clients interested in building small clusters
  - Developing VIA for more tightly coupled apps.
  - 32 processor software development cluster in house. Available for “test drives”
  - UCB Millennium collaboration
- PDSF - medium sized production cluster ( ~48 nodes) for high energy physics applications

# Good news

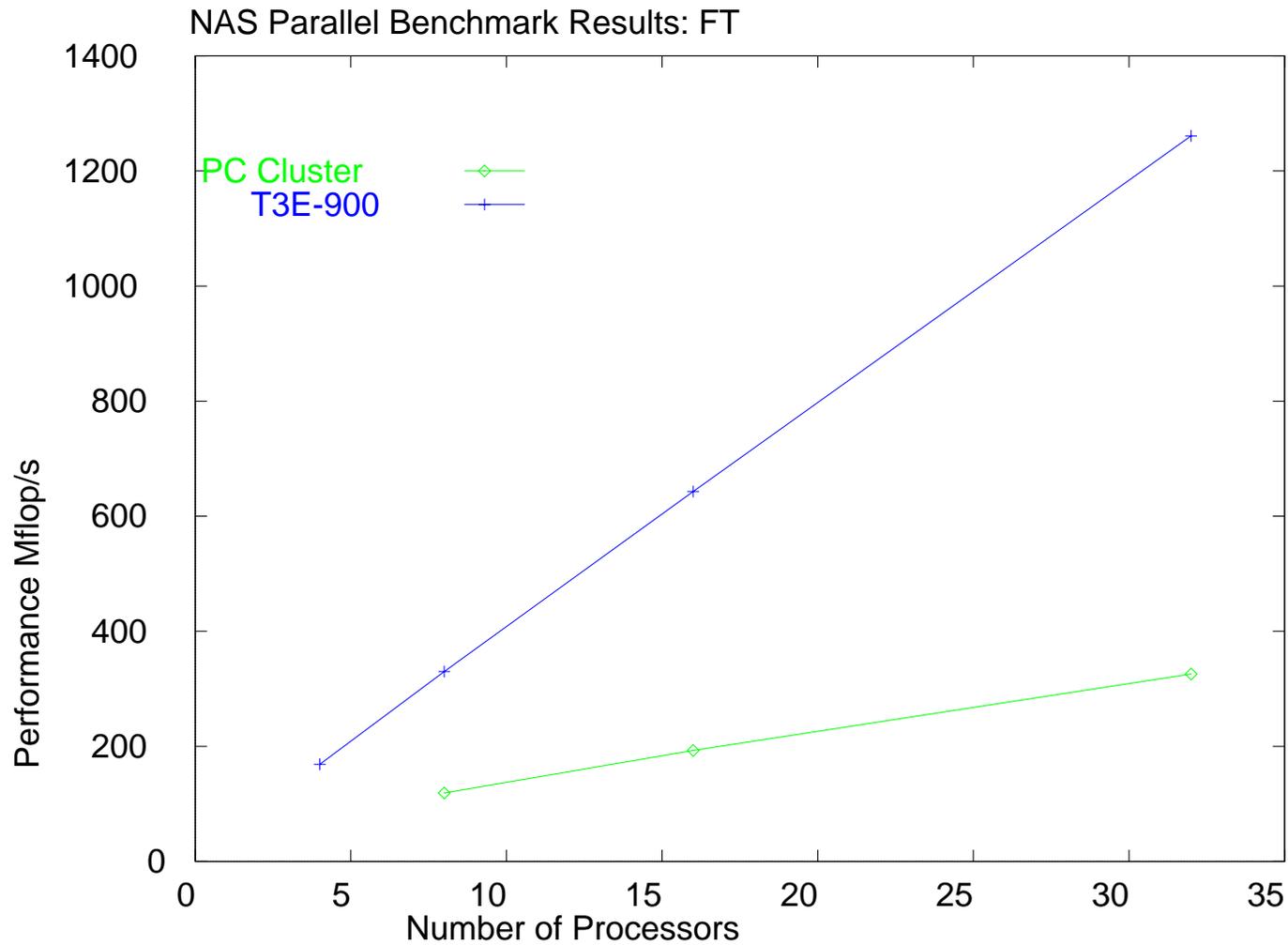
NAS Parallel Benchmark Results: BT



# Bad News



# Downright Ugly





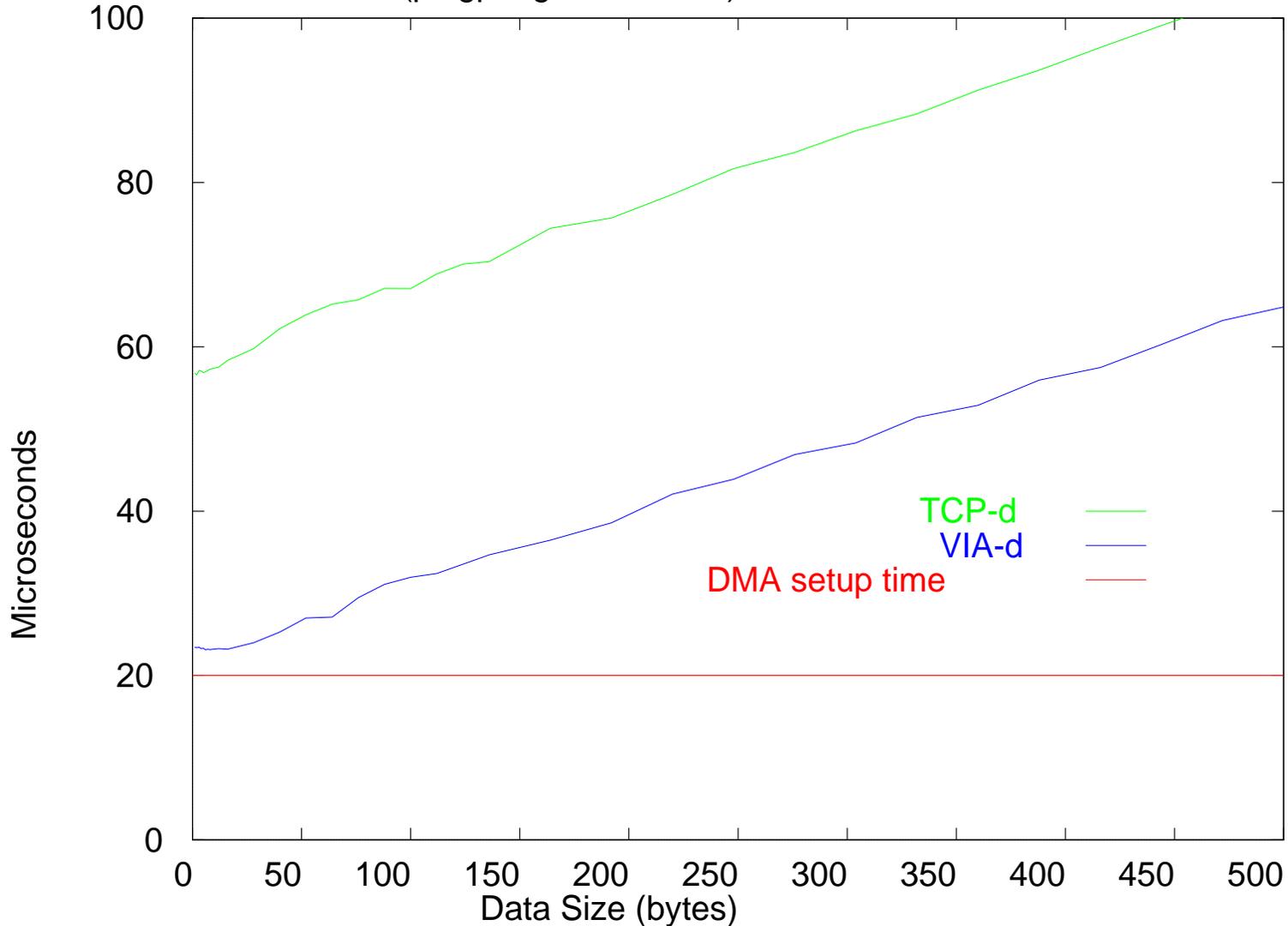
# Virtual Interface Architecture (VIA) - CRADA with Intel and ANL



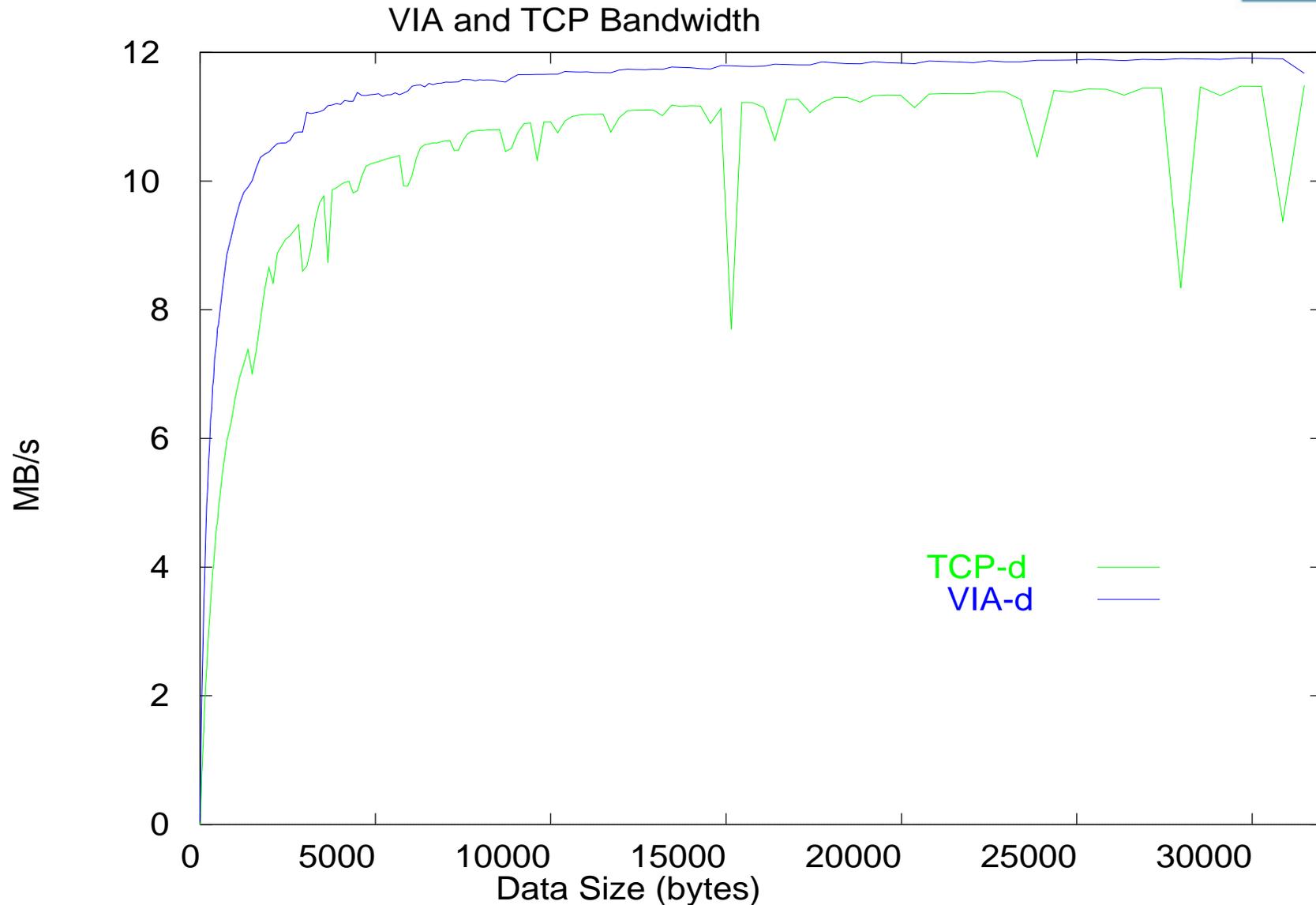
- VIA is the only commercially important standard for low-overhead communication on system area commodity networks.
  - Software only or software/hardware
  - Ideal for high performance on PC clusters
  - Basis for NGIO and Future I/O
- LBL implementation M-VIA (Modular VIA)
  - First and only Portable, high performance, network-independent, full implementation of VIA for Linux
  - Released 12/98 (mail to [via@nersc.gov](mailto:via@nersc.gov))

# VIA performance - latency

VIA and TCP Transfer Time (pingpong benchmark)



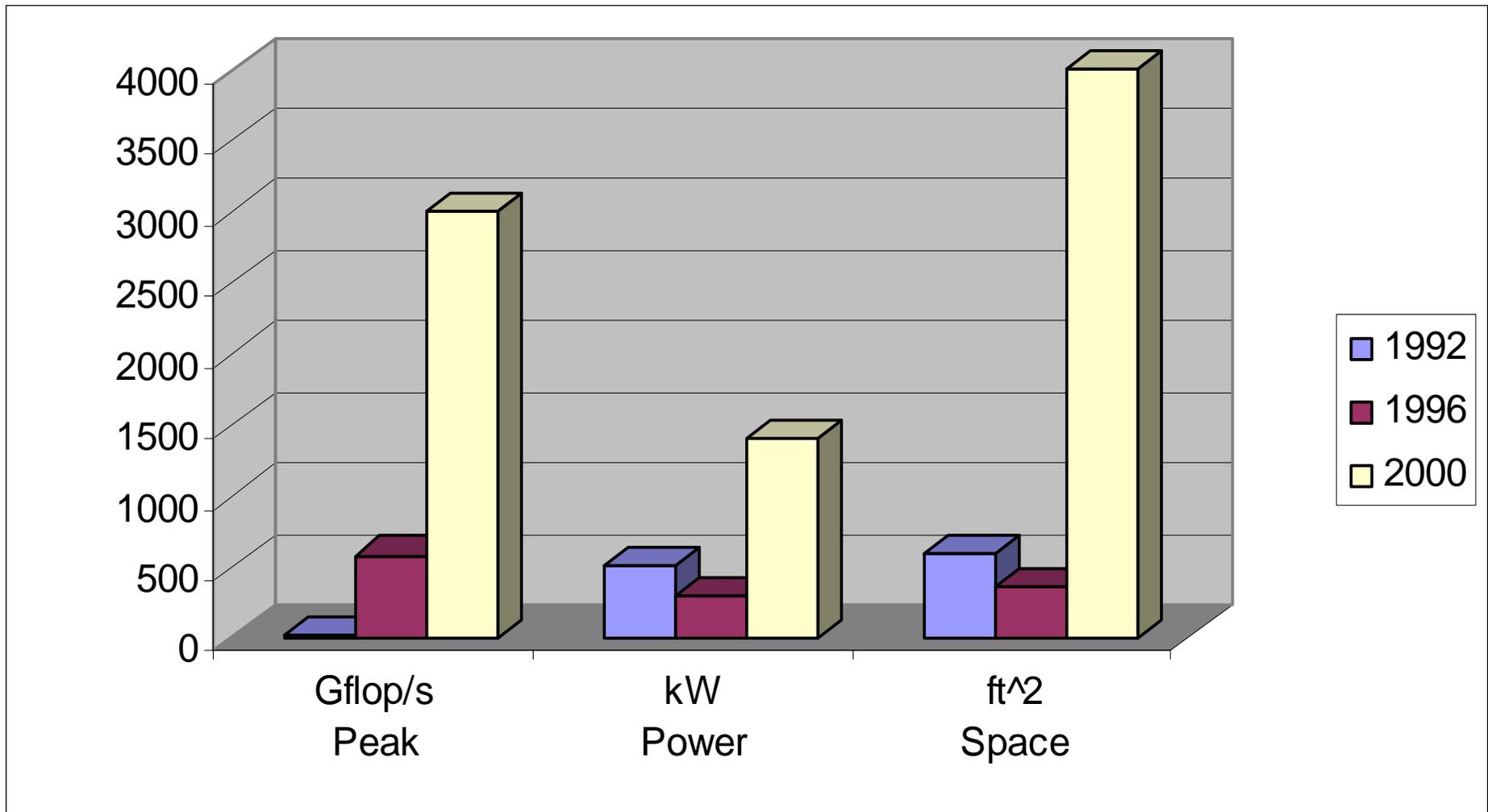
# VIA performance - bandwidth



- Berkeley Lab Distribution (BLD)
  - OS leadership in the BSD tradition
  - Distribution/development in the Linux tradition
  - Reference implementation, suitable as base for commercial development
- NERSC-4 ??, NERSC-5?
- Collaborative Research Agreement with IBM about Linux on SP



# Increase in Power and Space Requirements





# Increase in Power and Space Requirements



- 1996: factor of 6 increase in sustained compute power, and **decrease** in power and space requirements
- 2000: factor of 6 increase in sustained compute power; factor of **5 increase in power**, factor of **11 increase in space**
- requires new investment in building, infrastructure
- NERSC is using well developed commercial real estate in urban environment; lease for a new building in Oakland signed
- I would prefer to pay a higher price to vendors as opposed to develop real estate





# Berkeley Lab Computing Facility



- Will house NERSC together with LBNL Administrative Computing and other activities
- 28K-square-foot leased area
  - 11K-20K-sf computer room
  - 4K-8K-sf operations room, videoconference room, project offices
  - Options for additional 25K-sf computer room and 35K-sf office space
- Occupancy July 2000
- Ten-year lease with 3 five-year options
- 4MW-12MW power

- Overview
- Two technology transitions in the 1990s
- Clusters of SMPs as production platforms
- **Petabyte data challenges**
- Integration into the data grid
- Building a high performance organization
- Conclusions



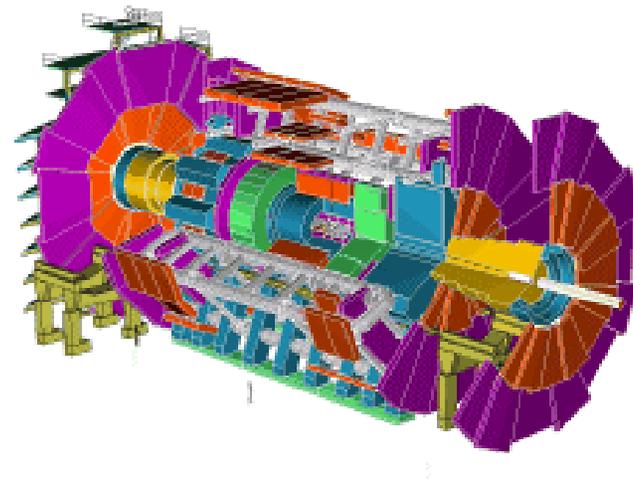
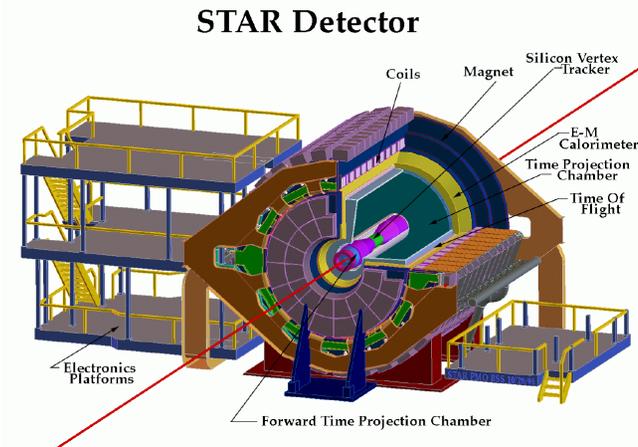
# Increased Requirements in Data Storage



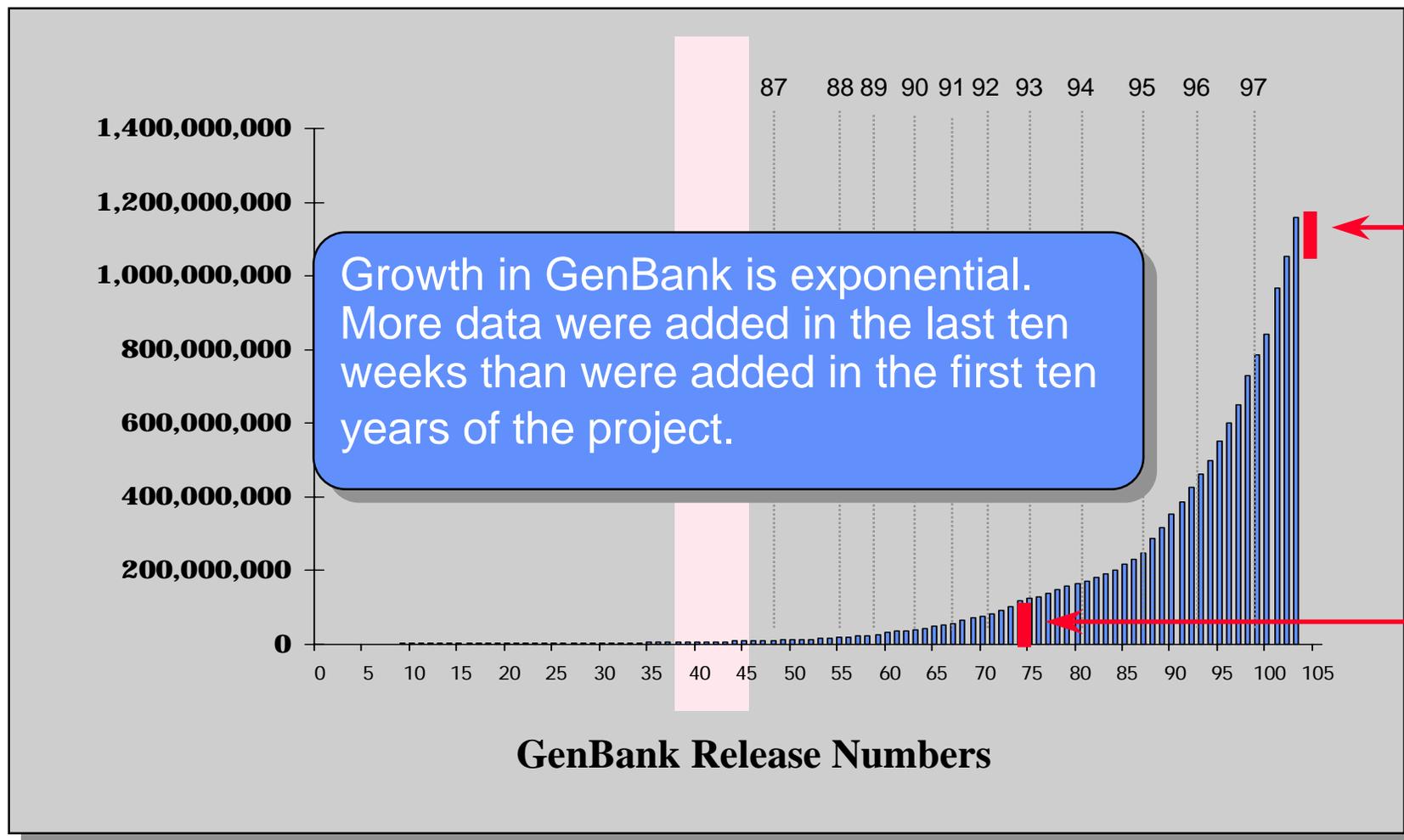
Many applications acquire simulation or experimental data at an accelerated rate

- high energy physics experiments
- genome data
- climate simulation and diagnostics
- combustion simulation and experiment
- crystallography
- protein engineering
- seismology
- cosmology
- medical imaging

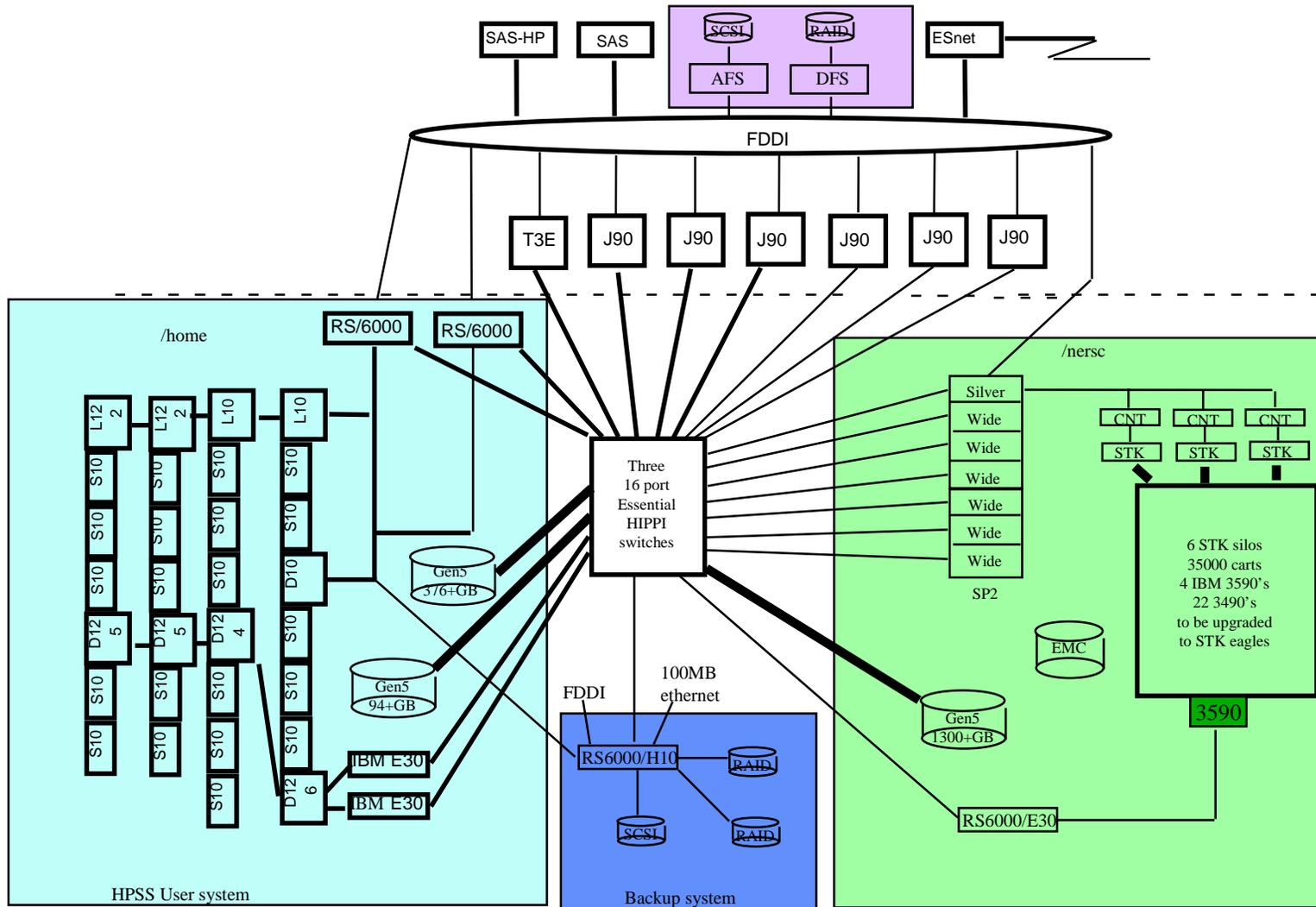
- STAR- RHIC at Brookhaven
  - Fall 1999
  - 266 TB/yr (year 2001)
  - 10,100 SPECint95
  - 73 MB/sec
- ATLAS-LHC at CERN
  - 2004
  - 1.1 PB/yr
  - 270e6 MIPS
  - 100 MB/s DAQ
  - 1 MB/s to desktop/ARC
  - 1Gb/s to MDS/ARC



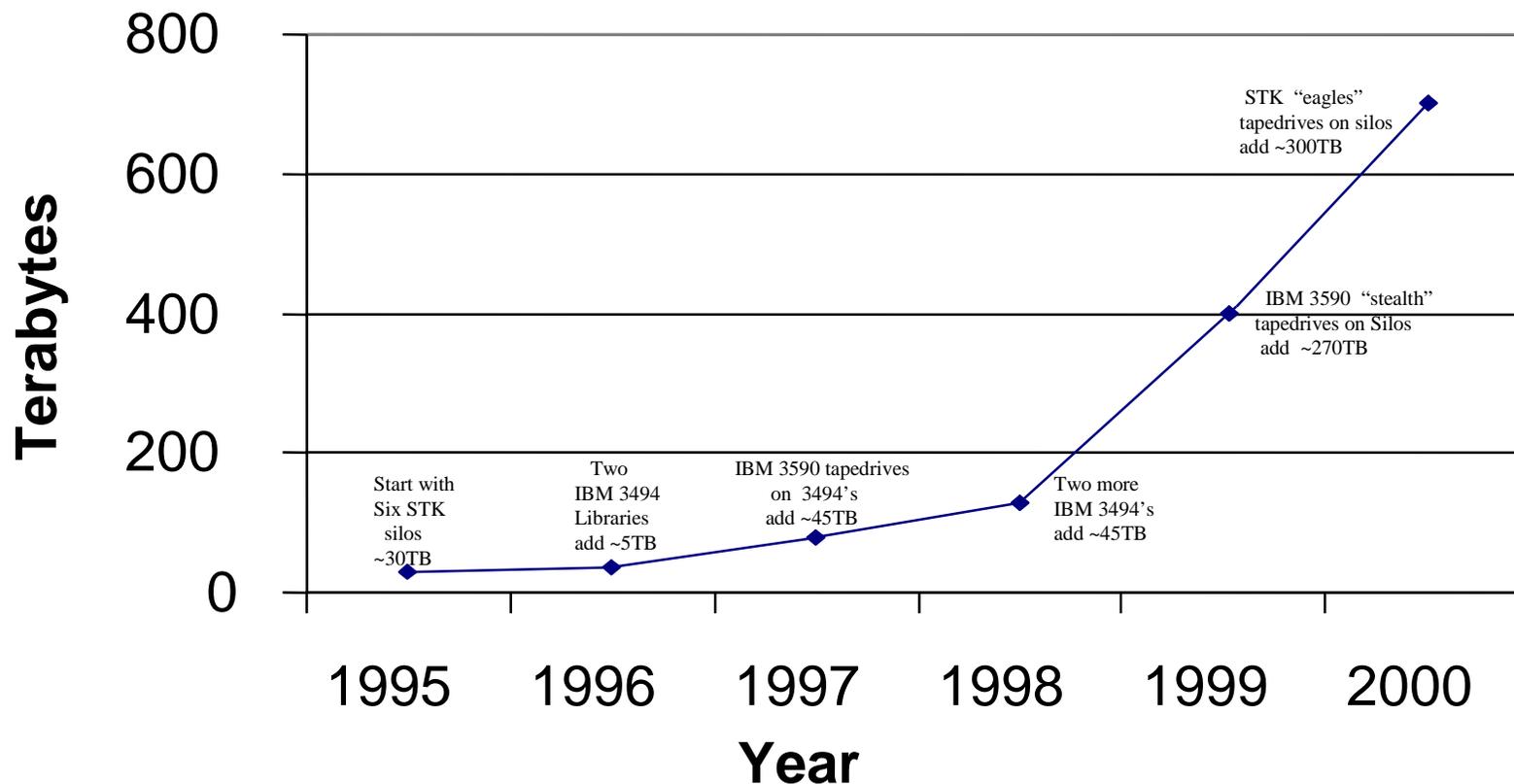
# Human Genome



# Current Storage Configuration



# NERSC Storage Capacity (raw - theoretical)



# Increased Requirements in Data Handling

- The amount of raw data is manageable ...
- ... but processing scientific data is increasingly a part of the mission
  - Users import data from external sources
    - Accelerators and other scientific instruments
    - Simulations run elsewhere (I.e., PCMDI)
  - Users analyze data
  - Users visualize data
- In the long run, effectively managing this data will be as critical to NERSC's success as managing our computing platforms.

- NERSC must maintain active involvement in the CS research community to accomplish its mission w.r.to new technology introduction
- Distinguish between **compute** intensive and **data** intensive technology
  - for **compute** intensive CS research, rely on leverage and technology transfer from universities
  - for data intensive CS research build key competency in house - **SCIENTIFIC DATA MANAGEMENT GROUP**
- Why? HPC technology is well researched and supported in many places, but few sites have 100 Terabyte - 1 Petabyte storage, and focus on data related problems

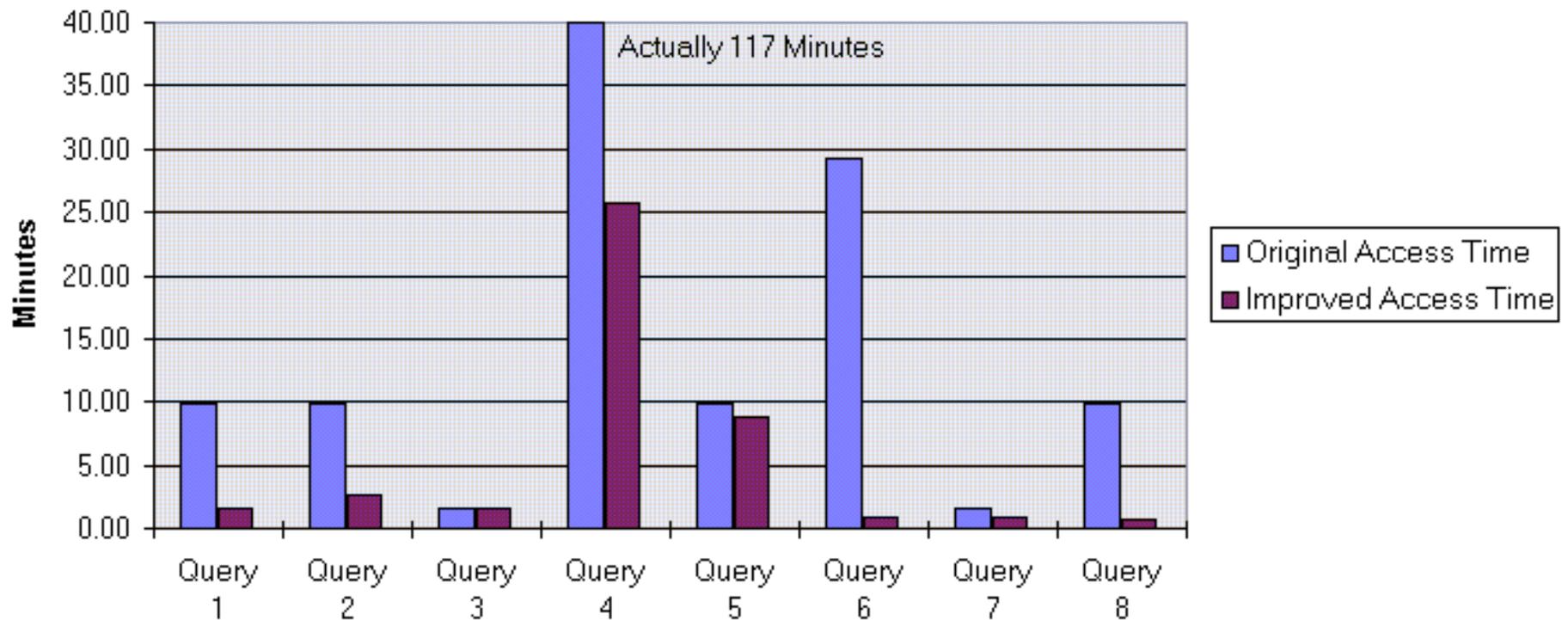


# Scientific Data Management Group

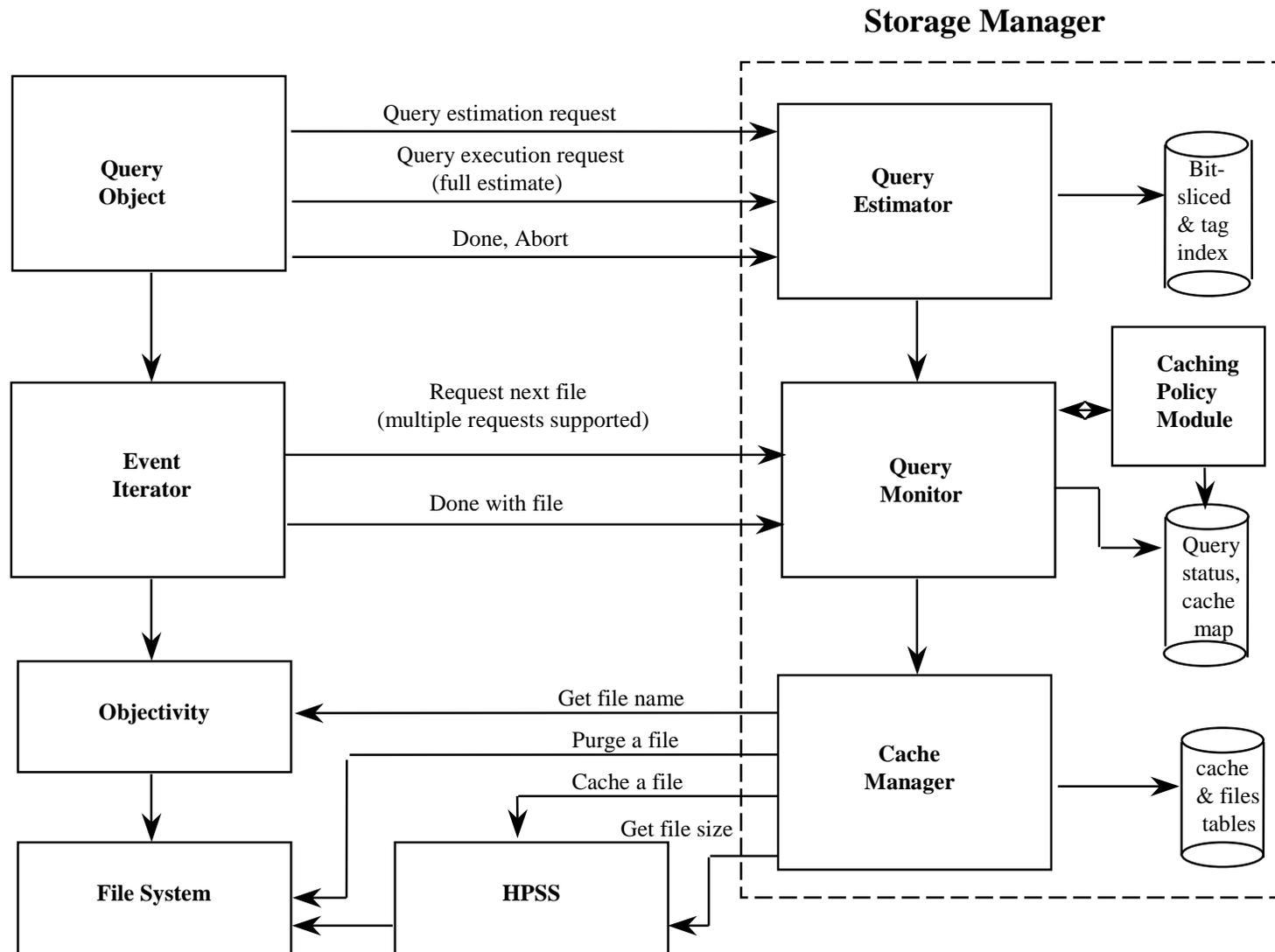


- Research Areas
  - Database tools on top of commercial database systems
    - Adaptation of object-based tools to various scientific applications
  - Data mining from large high-dimensional datasets
  - Data organization and indexing methods
    - for parallel disks
    - for tertiary storage
  - Specialized data structures and operations
    - emphasis on: temporal, spatial, sequence, multi-dimensional, and classification hierarchies
- Methodology
  - Research and Software products applied to specific scientific projects
  - Work performed by staff, faculty, students, developers, and visitors

Improved Access Time for 8 Typical Queries as a Result of Dataset Reorganization



# STACS- Storage Tape Access Coordination System



## Vision:

A national center for understanding information and information systems in modern biology

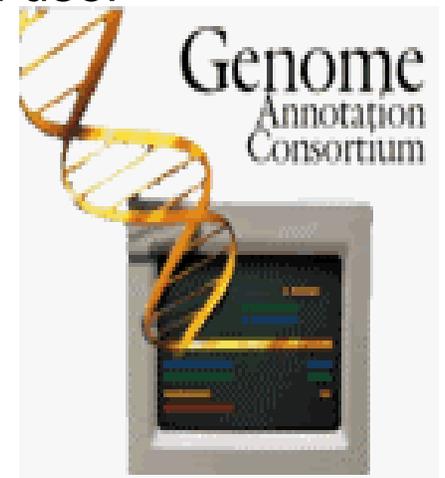
Provide initial support for creating a Center for Bioinformatics and Computational Genomics at LBNL

- **Bioinformatics**

The application of computer science, computational science, mathematics, and statistics to the analysis of large, complex biological data. This includes database and tool development, algorithm design, application and user interfaces, and data visualization.

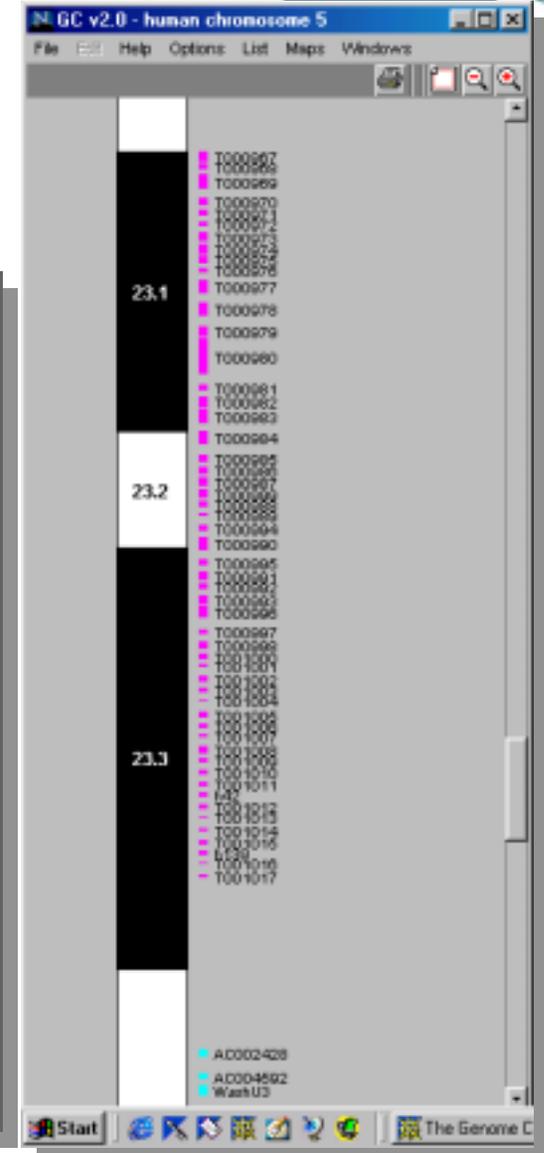
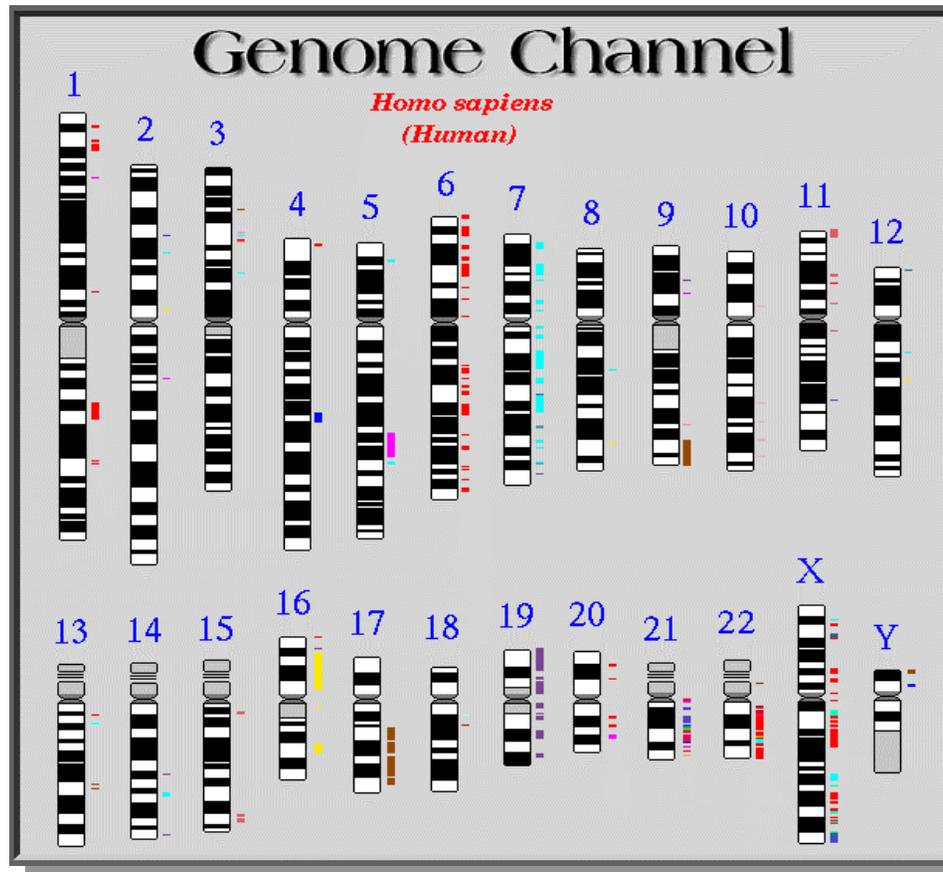
- **Computational Genomics**

Development and application of computational and statistical methods for analysis and comparison of genomic data, including genome annotation, whole genome analysis, and systematics.



# Genome Channel

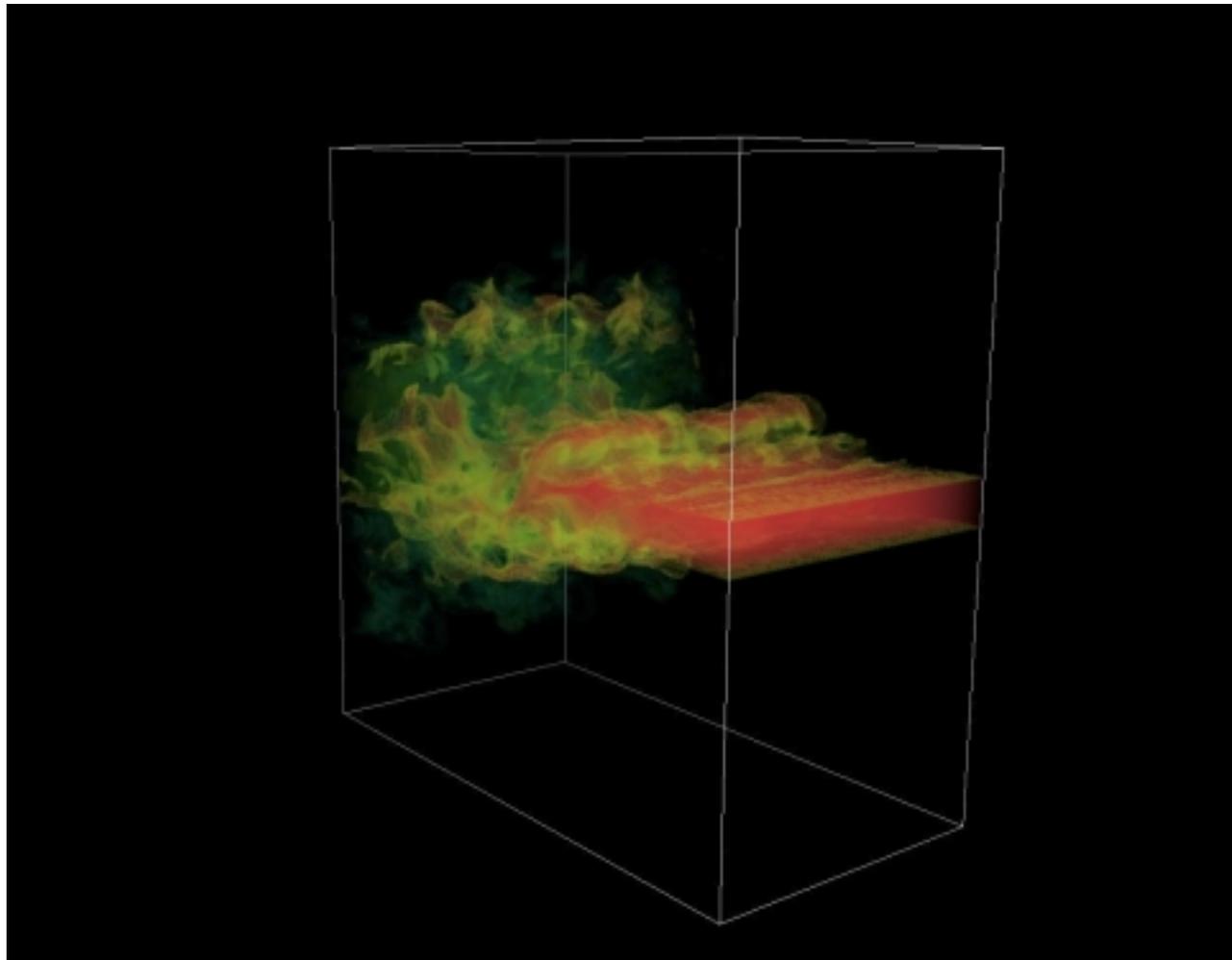
The **only place** to access all human genome sequences determined so far; data from all sequencing efforts combined



- Software Volume Renderer
  - Prototyped software volume renderer on the T3E.
  - Display images onto a workstation
  - Large data sets resident in T3E main memory
    - Upper bound on data set size (on the T3E) is  $3700^3$ .
  - MPI for interprocess communication

Example: CCSE has been testing this renderer and will be including it with their visualization tool.

# Visualization of Large Data Software Volume Renderer



- Overview
- Two technology transitions in the 1990s
- Clusters of SMPs as production platforms
- Petabyte data challenges
- **Integration into the data grid**
- Building a high performance organization
- Conclusions

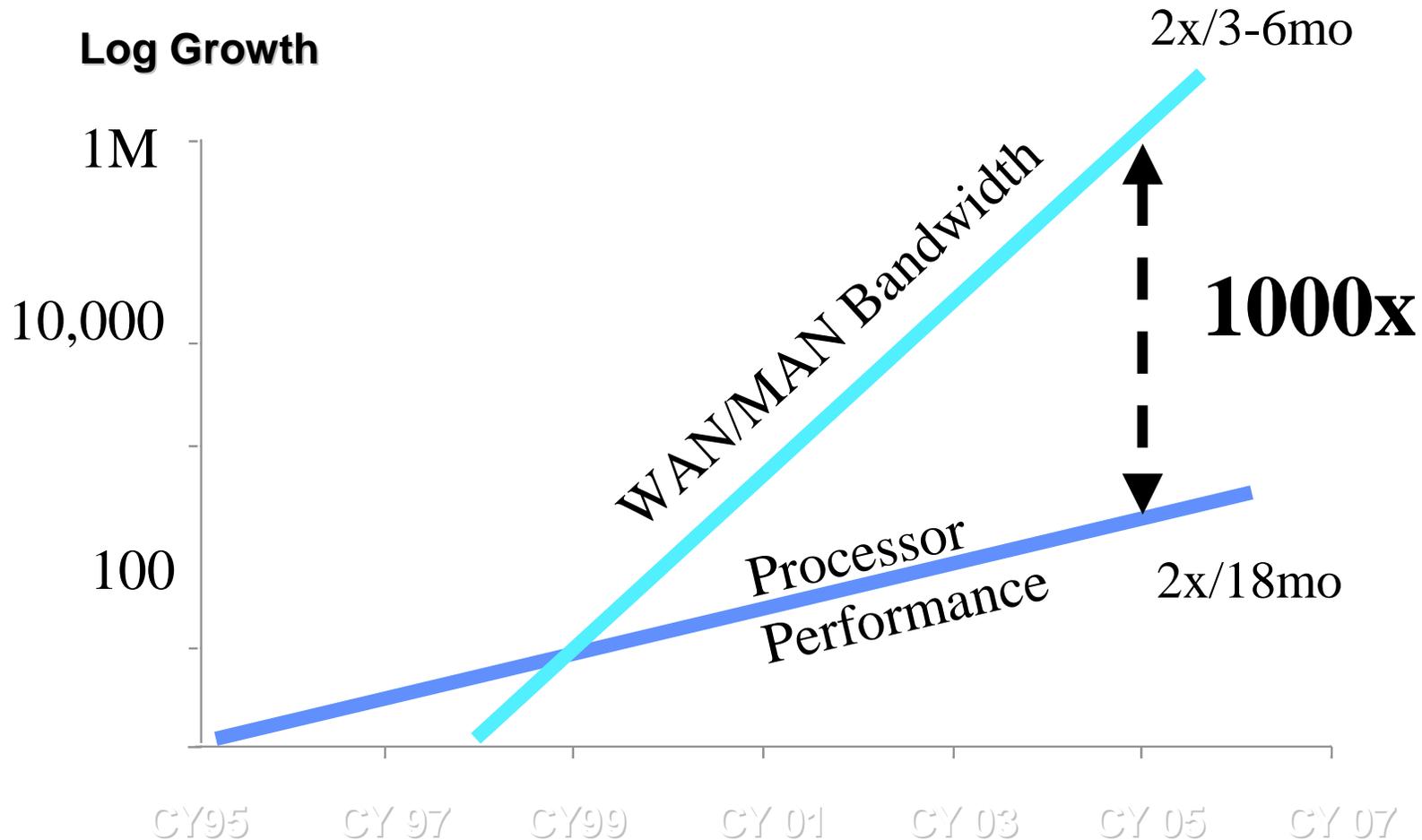


# Networking Bandwidth Increases



- Example: Nortel has announced 1.6 Terabit networking bandwidth (160 way WDM) -- technical trials in 4th quarter 1999
- At 1.6 Terabits, the equivalent of the Library of Congress (~25 Terabytes) takes 125 seconds to transmit
- Data are everywhere!

Adapted from G. Papadopoulos, Sun



# Computational Grids

New paradigm for large-scale computational science

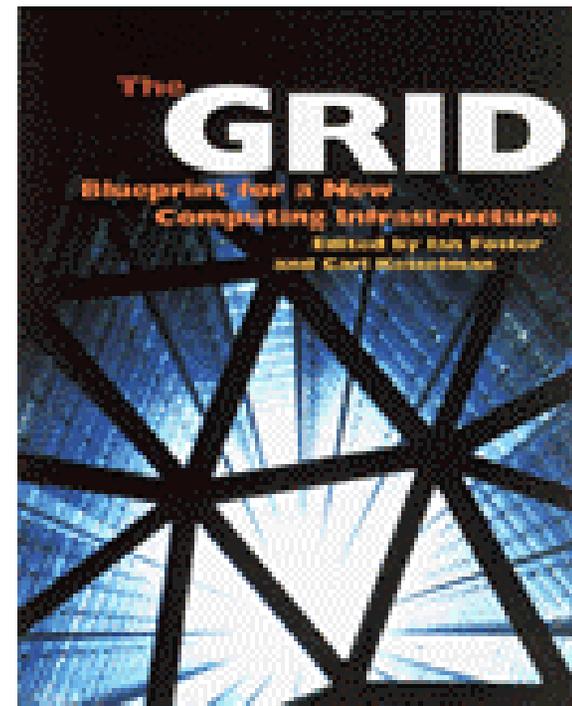
Transparent application access to remote data and systems

Supports moving, visualizing, and analyzing large data sets

Analogous to electric power grid

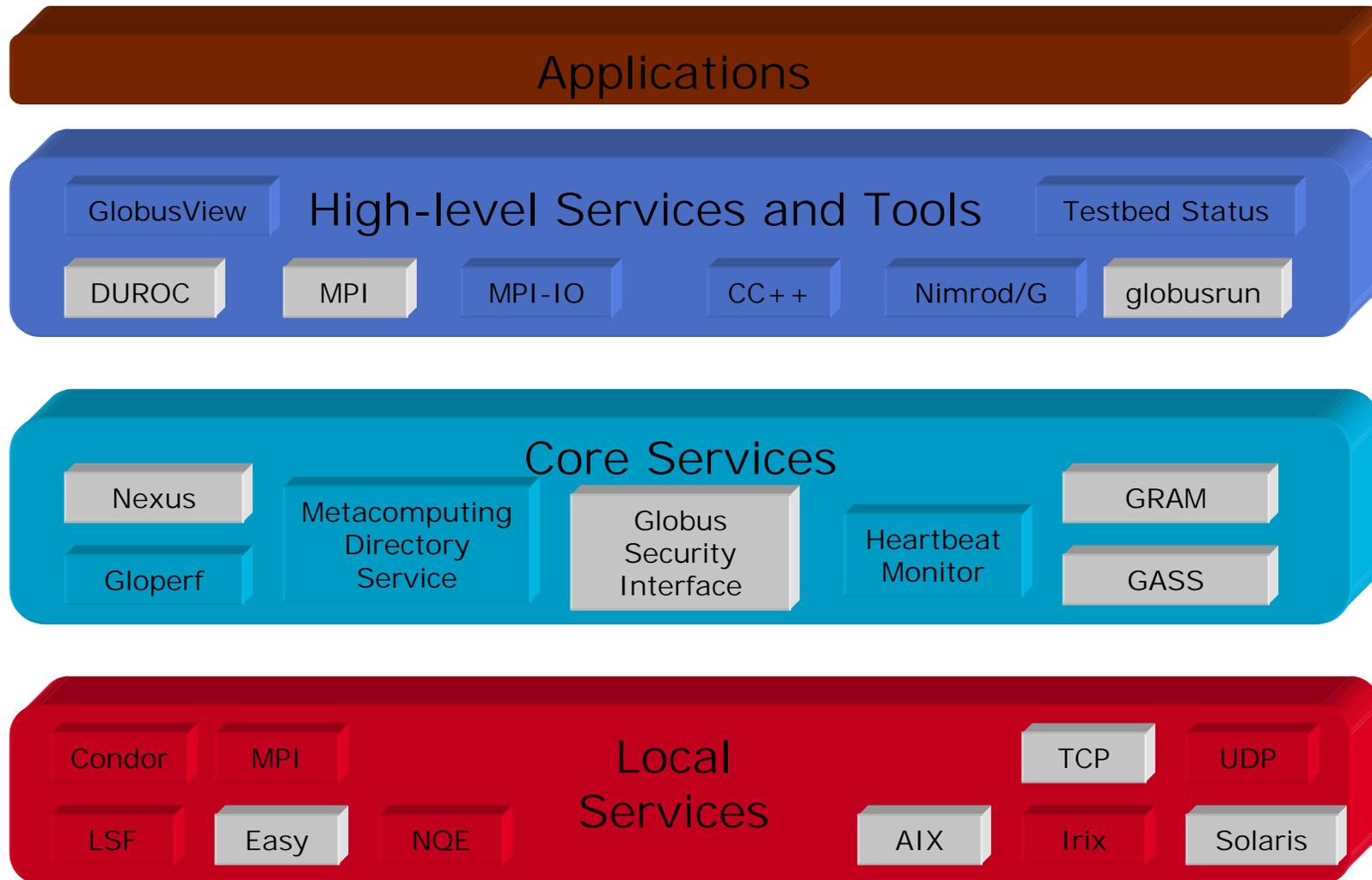
Aggregates existing computational infrastructure

- Computers
- Networks
- Mass storage systems
- Visualization devices

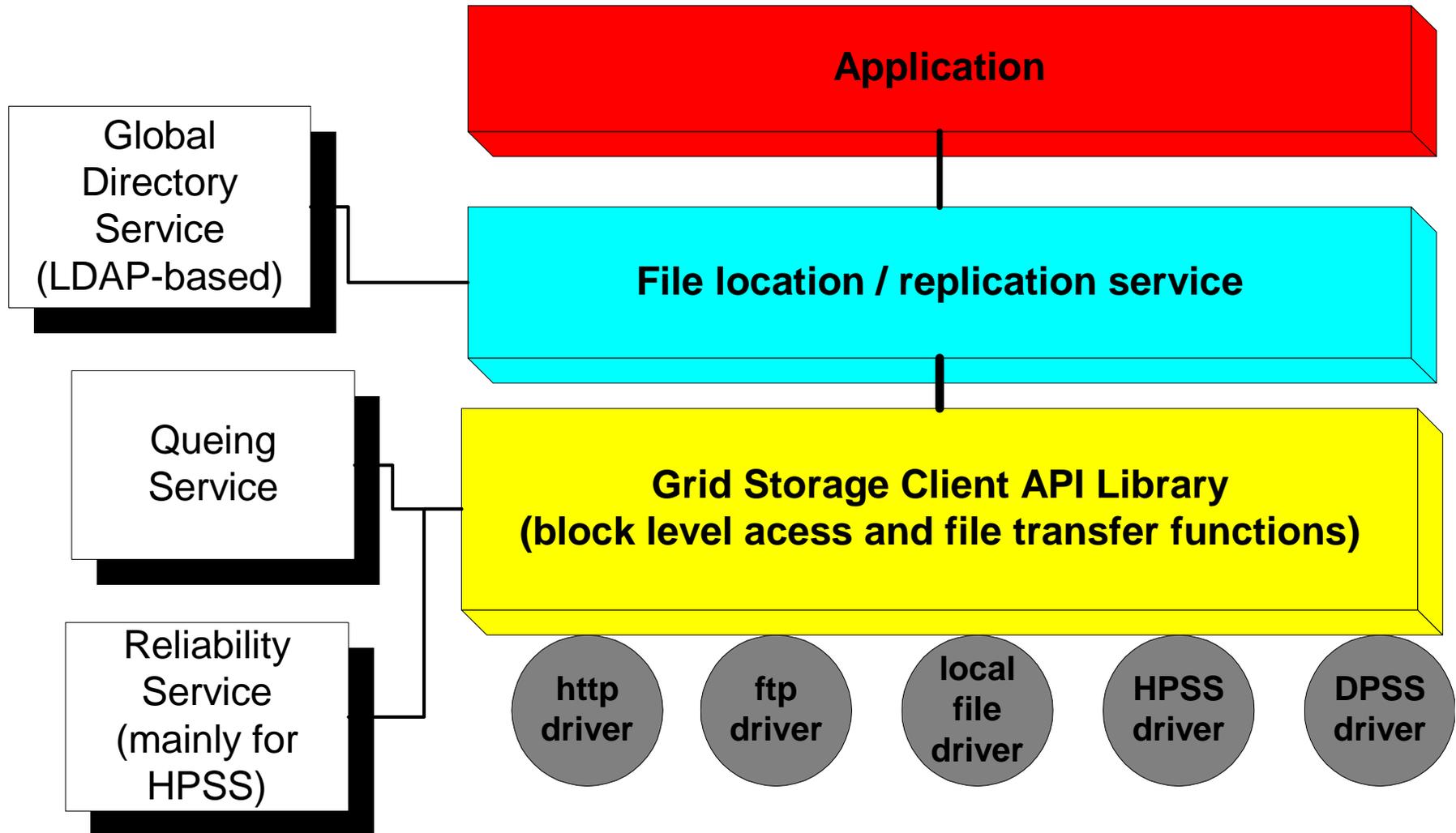


- Grid / Computational Grid:
  - The integration of various approaches used for integrating dispersed resources
  - analogy with the grid that supplies ubiquitous access to electric power.
  - Basic grid services are those that locate, allocate, coordinate, utilize these resources
- Data Grid:
  - services for handling remote access to large data sets in a grid environment
- Working with Globus group at ANL to build “Data Grid” services

# Layered Architecture (Globus)



- We use the term “Data Grid” to describe additional services that are unique to data intensive grid applications. These services include:
  - data migration tools that are optimized for transferring large data sets over WANs
  - data set discovery and replication tools
  - data caches / cache management services
  - metadata service:
    - global name space for data archived at multiple sites
    - file access control
    - file collections (data set = many files)
    - replica management



# Building a Data Grid: Building Blocks

Ingest/  
catalog  
service

pftp  
GASS

MCAT,  
SRB

STACS,  
Condor  
others

Globus toolkit: security, information,  
fault detection, resource management,  
communication, etc.

MPI-IO  
Akenti

Netlogger  
Autopilot

Condor  
...

HPSS  
Archival, multi-PB  
Access > 100 MB/s?  
No QoS

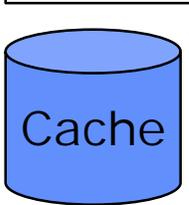
DPSS  
Fast disk cache  
No QoS

Computers  
Preliminary QoS  
work (e.g., DSRT)  
XFS: QoS for disk

ESnet, MREN,  
NTON  
QoS: e.g., diffserv



1-10 PB  
Archival  
GB/s net  
QoS



10-100 TB  
Nonarchival  
GB/s net  
QoS

Analysis  
computer

1-10 TF/s  
GB/s net  
On-demand  
QoS

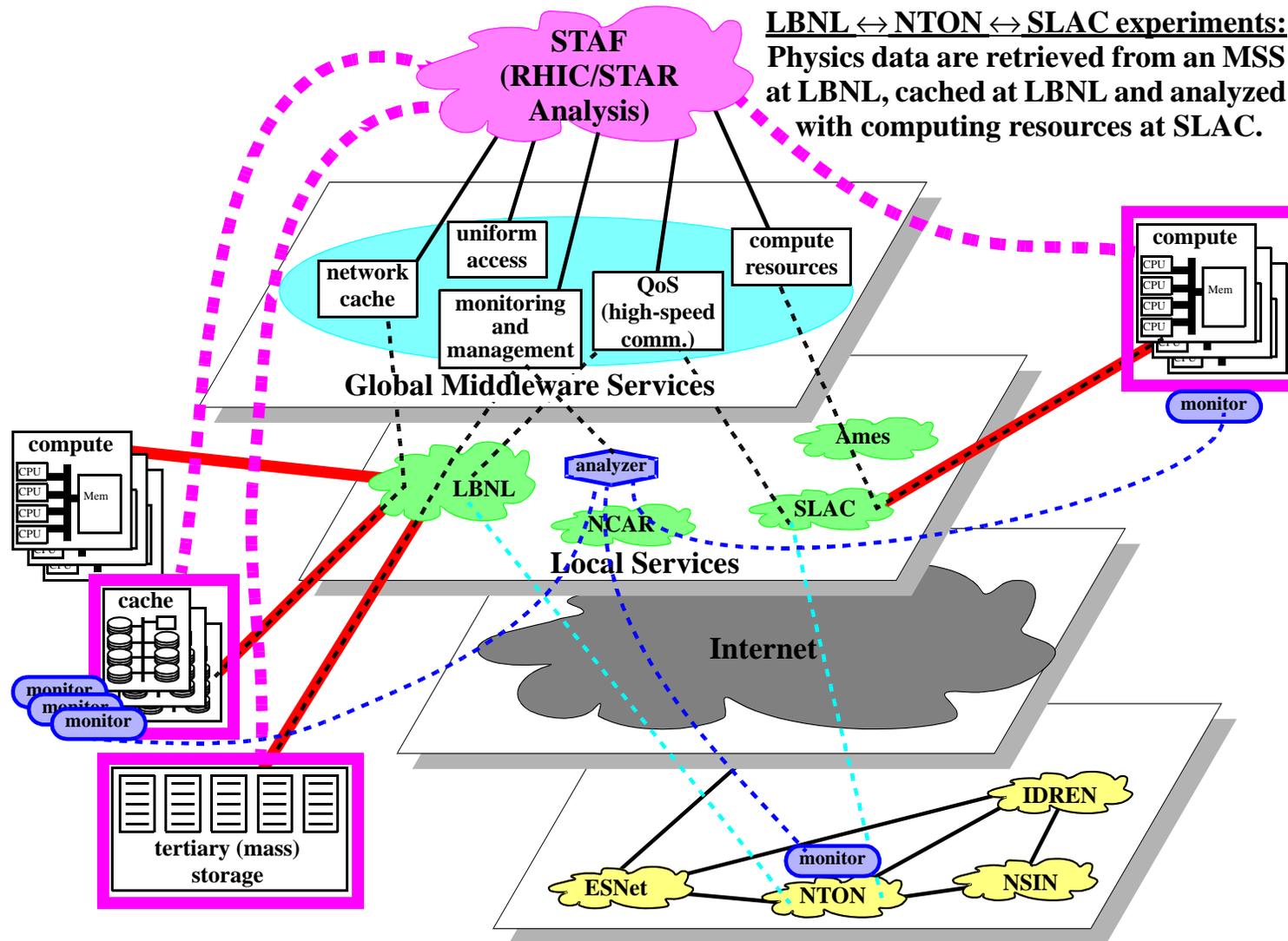
Network  
Striped  
Secure  
QoS

# China Clipper Project

- Goals
  - Develop technologies required for distributed data-intensive applications
  - Apply to high energy physics (HEP) data analysis
- Participants
  - Argonne National Laboratory
  - Lawrence Berkeley National Laboratory
  - Stanford Linear Accelerator Center (SLAC)

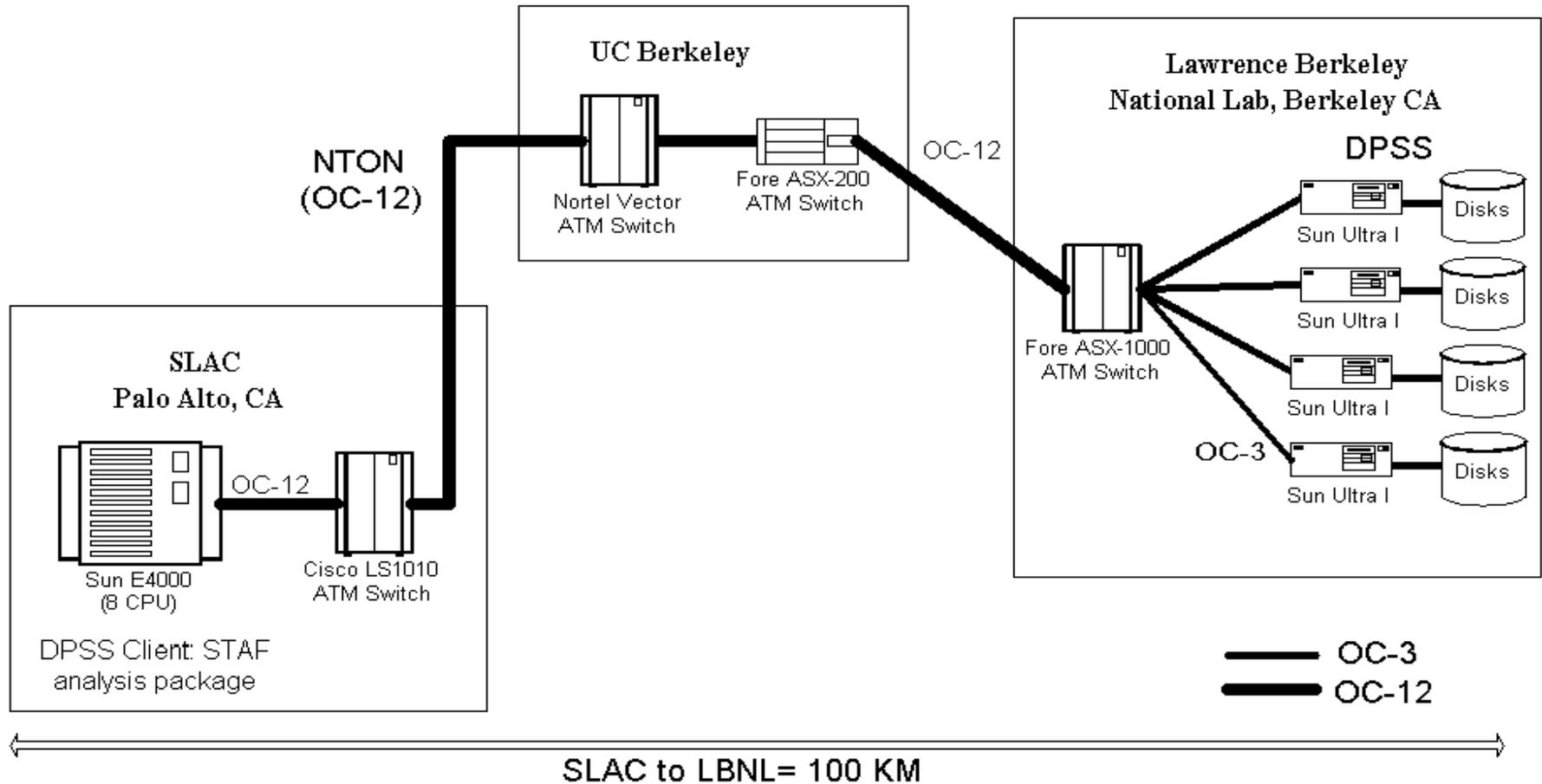


# Clipper Architecture



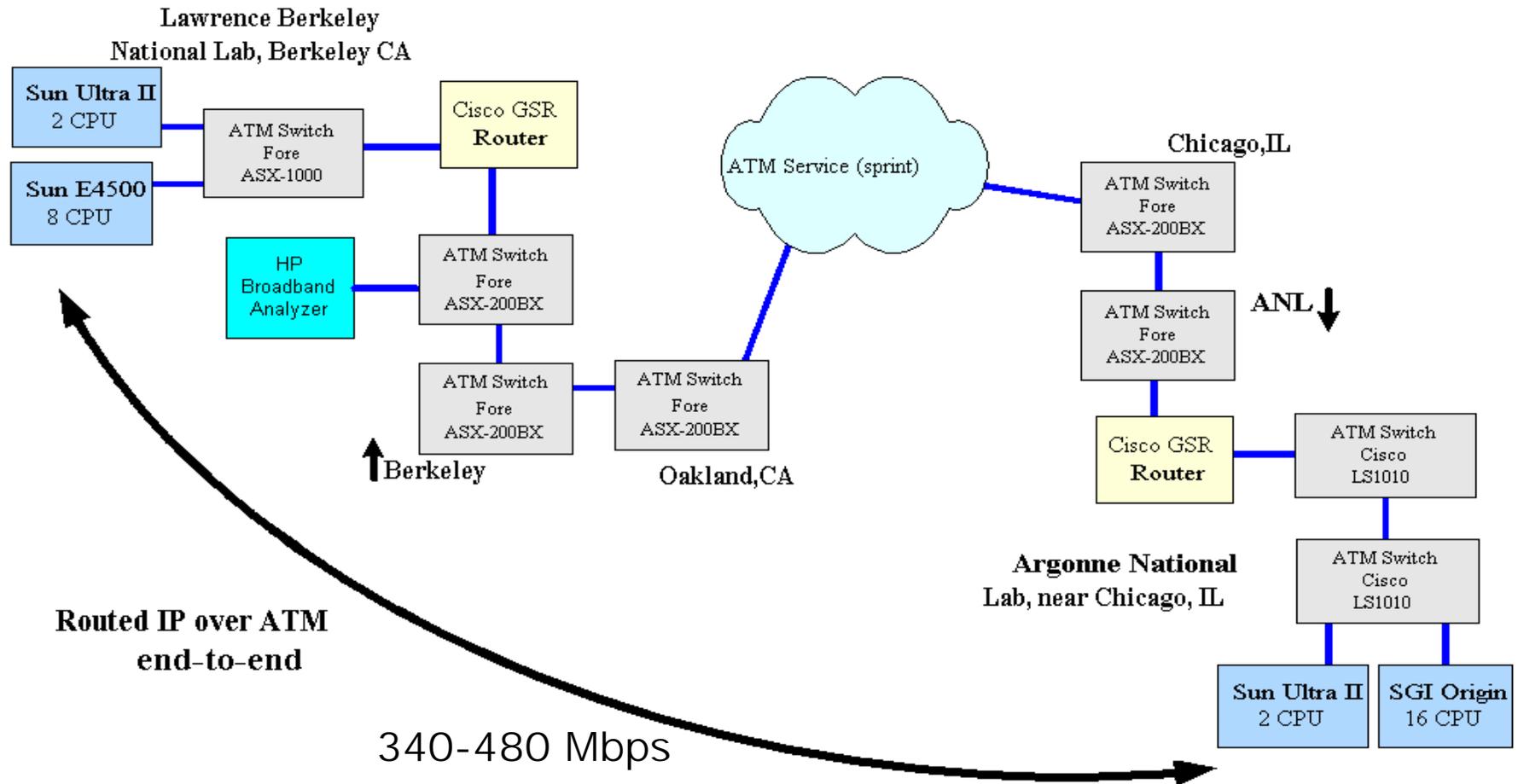
- Distributed Parallel File System (DPSS)
  - High-speed, low-cost data cache
- Globus
  - End-to-end resource management
- ESnet and NTON
  - OC12 networks
- HPSS and Objectivity
  - Data archives

# LBLN / SLAC Application Experiment



Achieved 57 MBytes/sec (450 Mbits/sec) of user data delivered to the application (equivalent to 4.5 TBytes/day)

# LBL/ANL TCP Experiments



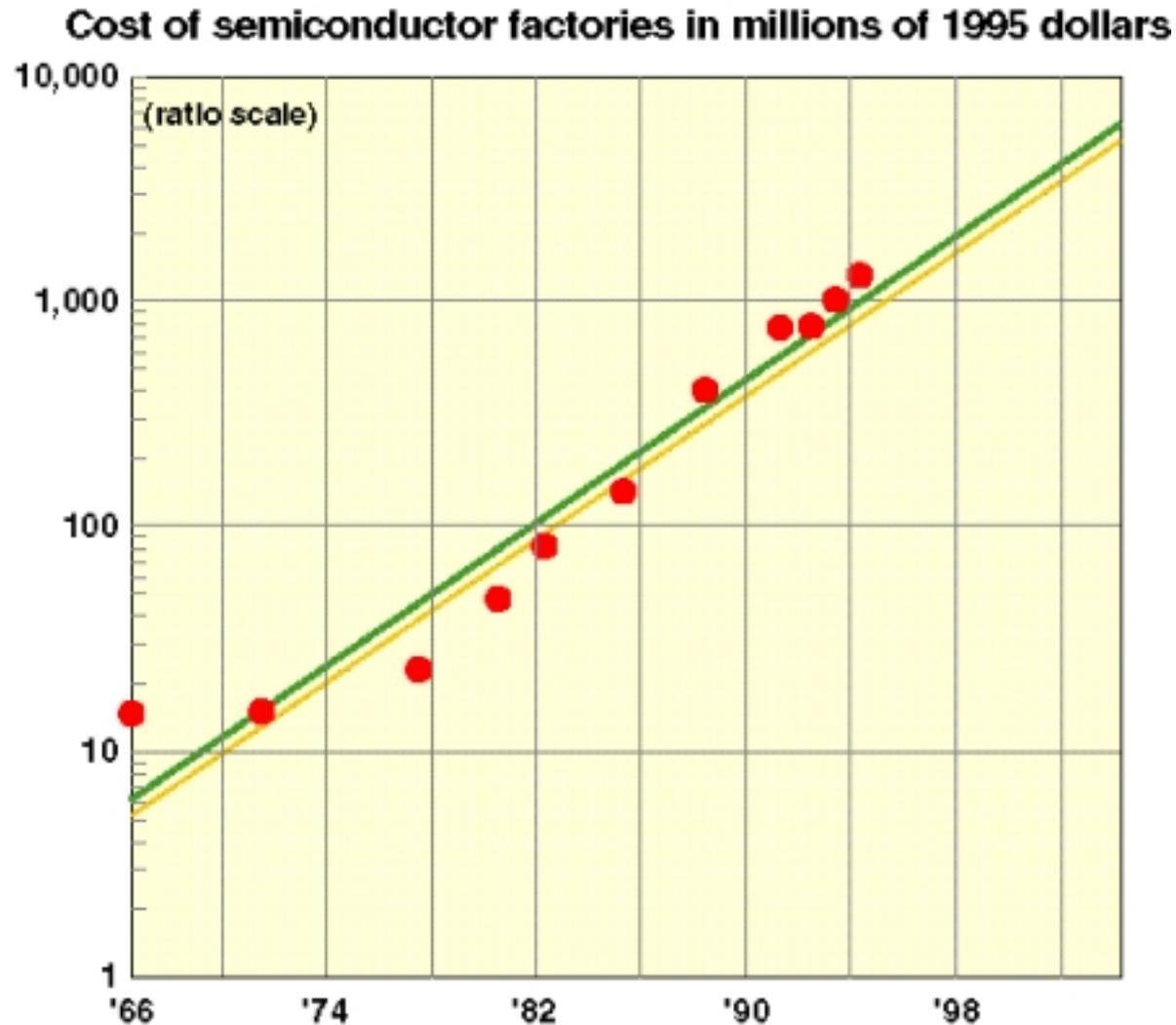
2/9/99 - RLN

# NGI projects at NERSC

- Four major applications projects funded
  - Earth Systems Grid (NCAR, LLNL, LANL, LBNL, ANL)
  - Combustion Corridor (LBNL, Sandia et al.)
  - Corridor One (ANL, ASCI, LANL, LBNL)
  - High Energy Physics (CalTech, LBNL, et al.)
- Three research projects funded as well
- 3 Year projects starting in July 99, about \$3M at NERSC

- Overview
- Two technology transitions in the 1990s
- Clusters of SMPs as production platforms
- Petabyte data challenges
- Integration into the data grid
- Building a high performance organization
- Conclusions

# Moore's Law - the traditional (linear) view



# Moore's Wall - the real (exponential) view





# High Performance Organization



- must keep an environment where staff will continue to explore the new and unknown
- in the past used technologies and outside consulting to facilitate continuous organizational change -- change will be the normal state
- current advantages: SF Bay Area dynamics and young staff (only 15% of NERSC staff has been with the lab from before 1996)

# Summary

- NERSC has established itself as the new model of a supercomputer center in less than two years from 1996 to 1997, and reached “maturity” in 1998
- The next three years will be again a major challenge because of fundamental technology transitions in several areas.
- NERSC provides excellent staff, facilities, and intellectual infrastructure and is ready to take on the Teraflops/Petabyte challenge



# NERSC Vision



**NERSC aspires to be a world leader in accelerating scientific discovery through computation. Our vision is to provide high-performance computing tools and expertise to tackle science's biggest and most challenging problems, and to play a major role in advancing large-scale computational science and computer science.**



# NERSC Division

