Computing

Richard P. Mount

Director, SLAC Computing Services

Assistant Director, Research Division

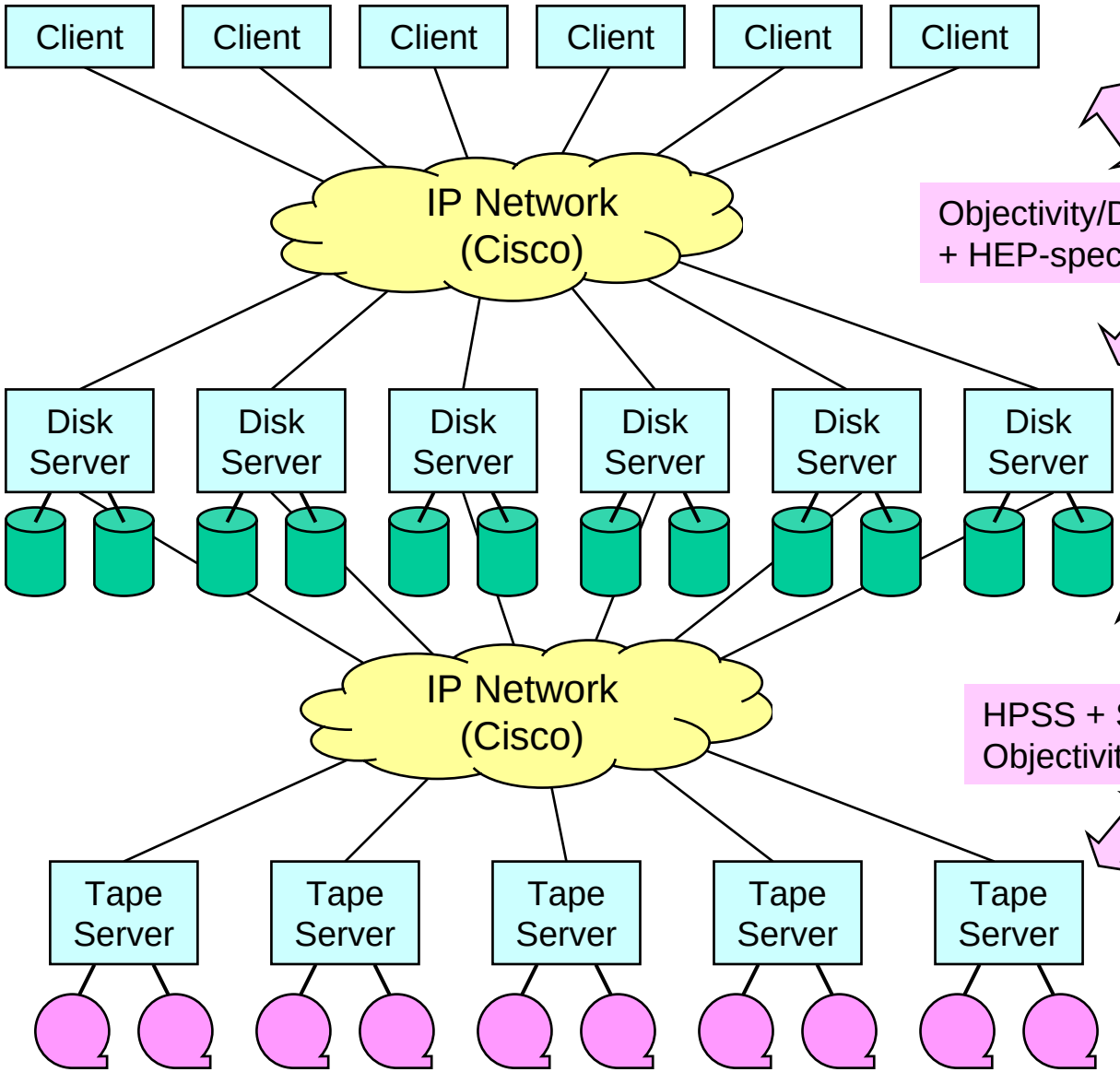
DOE Review

June 3, 2004

SLAC Computing

- BaBar
  - The world's most data-driven experiment
- KIPAC
  - Immediate and future challenges
- Research and development: The science of applying computing to science
  - Scalable, Data-Intensive Systems
  - Particle Physics Data Grid (SciDAC)
  - Network research and monitoring (MICS/SBIR/DARPA etc.)
  - GEANT4, OO simulation code

SLAC-BaBar Computing Fabric



1400 dual CPU Linux
900 single CPU
Sun/Solaris

Objectivity/DB object database
+ HEP-specific ROOT software (Xrootd)

120 dual/quad CPU
Sun/Solaris
400 TB Sun
FibreChannel RAID
arrays

HPSS + SLAC enhancements to
Objectivity and ROOT server code

25 dual CPU
Sun/Solaris
40 STK 9940B
6 STK 9840A
6 STK Powderhorn
over 1 PB of data

BaBar Computing at SLAC

- Farm Processors (4 generations)
- Servers (the majority of the complexity)
- Disk storage (3+ generations)
- Tape storage
- Network “backplane”
- External network

- Planning and cost management
- Tier-A Centers: the distributed approach to BaBar’s data-intensive computing.

Sun Netra-T1 Farm

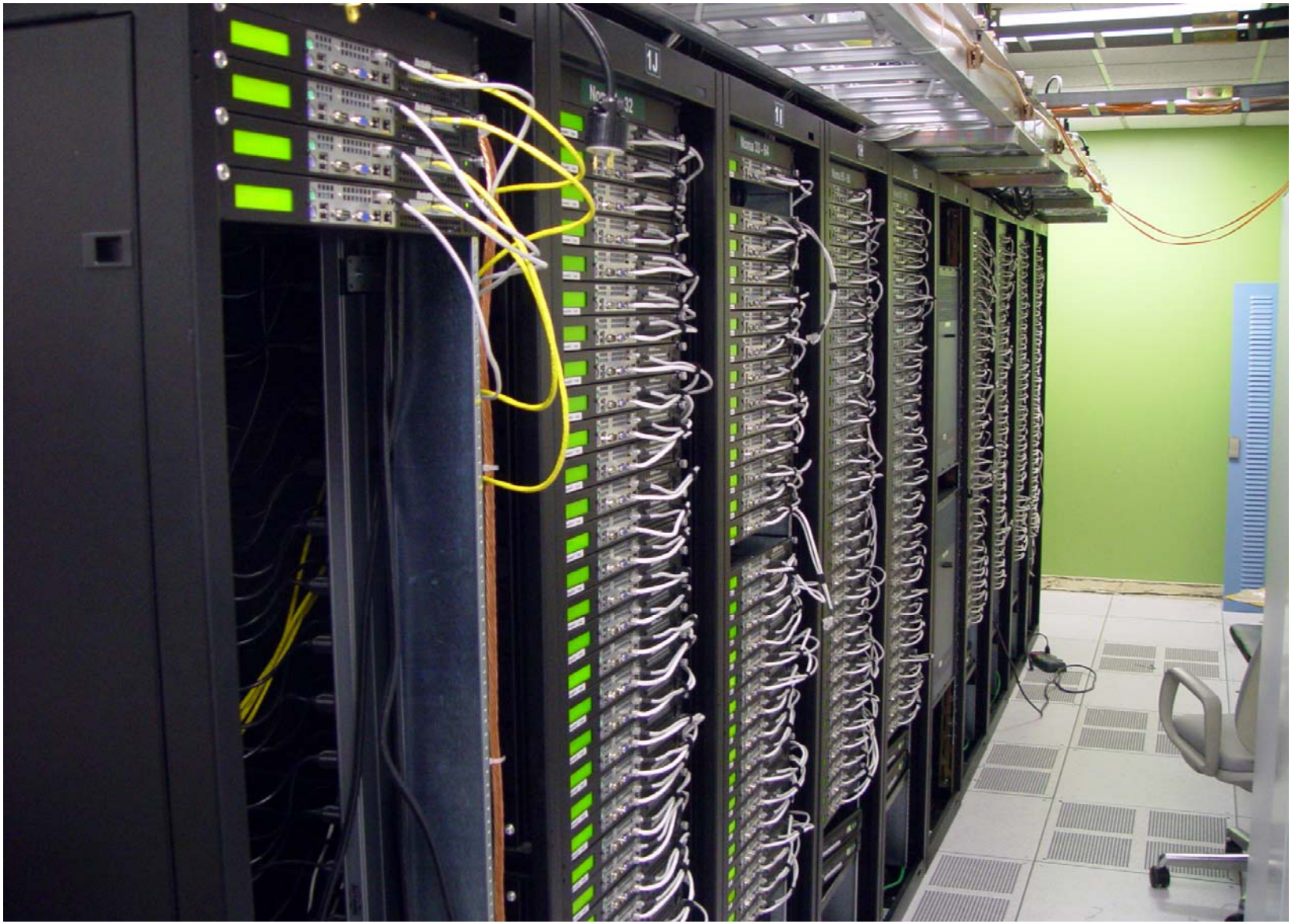
900 CPUs Bought in 2000 (to be retired real soon now)

VA Linux Farm (bought in 2001)

512 machines, each 1 rack unit, dual 866 MHz CPU

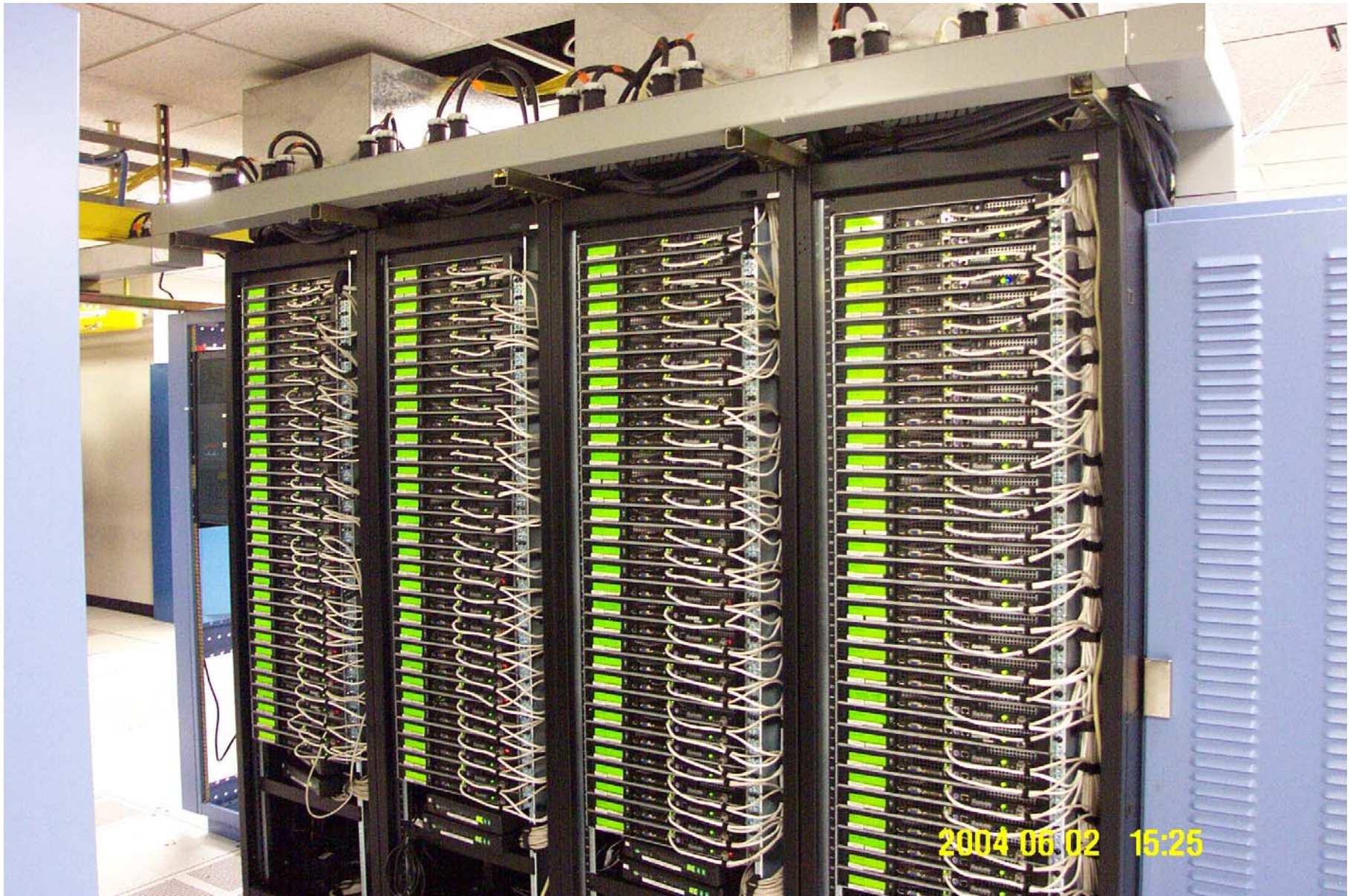
Rackable Intel PIII Farm (bought in 2002)

512 machines, 2 per rack unit, dual 1.4 GHz CPU



Rackable Intel P4 Farm (bought in 2003/4)

384 machines, 2 per rack unit, dual 2.6 GHz CPU



Sun Raid Disk Arrays (Bought 1999, 2000)

about 60 TB in 300 trays (Retired 2003)

Sun T3 FibreChannel Raid Disk Arrays

0.5 TB usable per tray, (144 trays bought 2001)
1.2 TB usable per tray, (68 trays bought 2002)

Electronix IDE-SCSI Raid Arrays

0.5 TB usable per tray, 22 trays bought 2001

(Retired 2003)

Sun 6120 “T4”

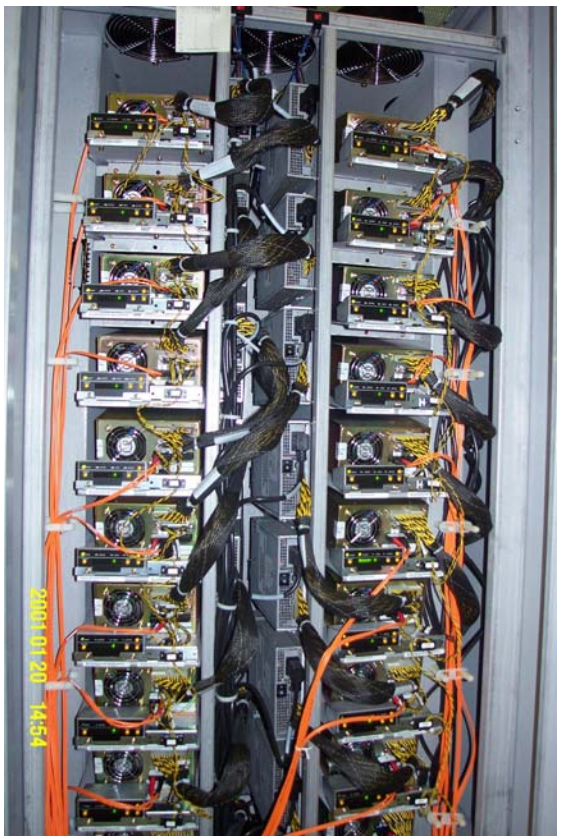
1.6 TB usable per tray, ~160 trays bought 2003/4

Tape Drives

40 STK 9940B (200 GB) Drives

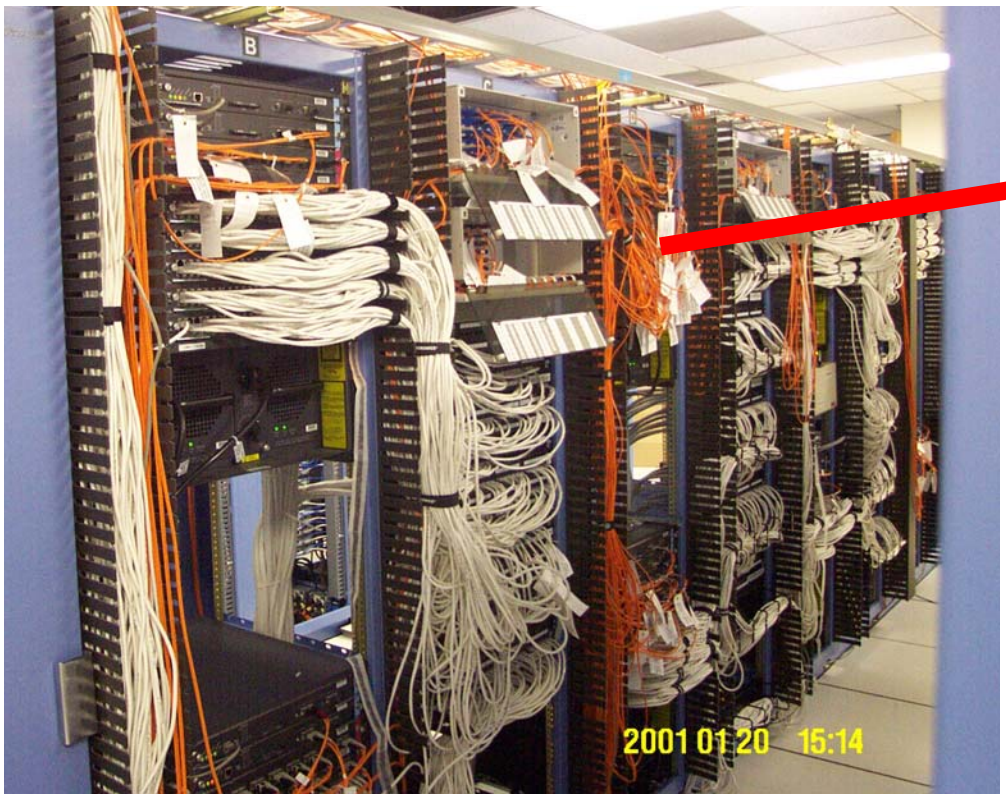
6 STK 9840 (20 GB) Drives

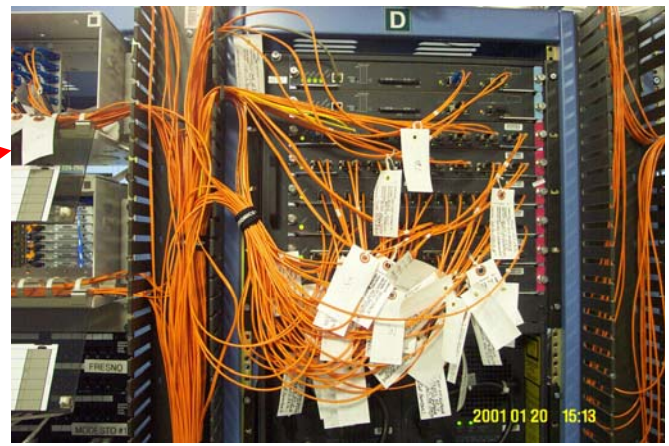
6 STK Silos (capacity 30,000 tapes)



BaBar Farm-Server Network

~22 Cisco 65xx Switches





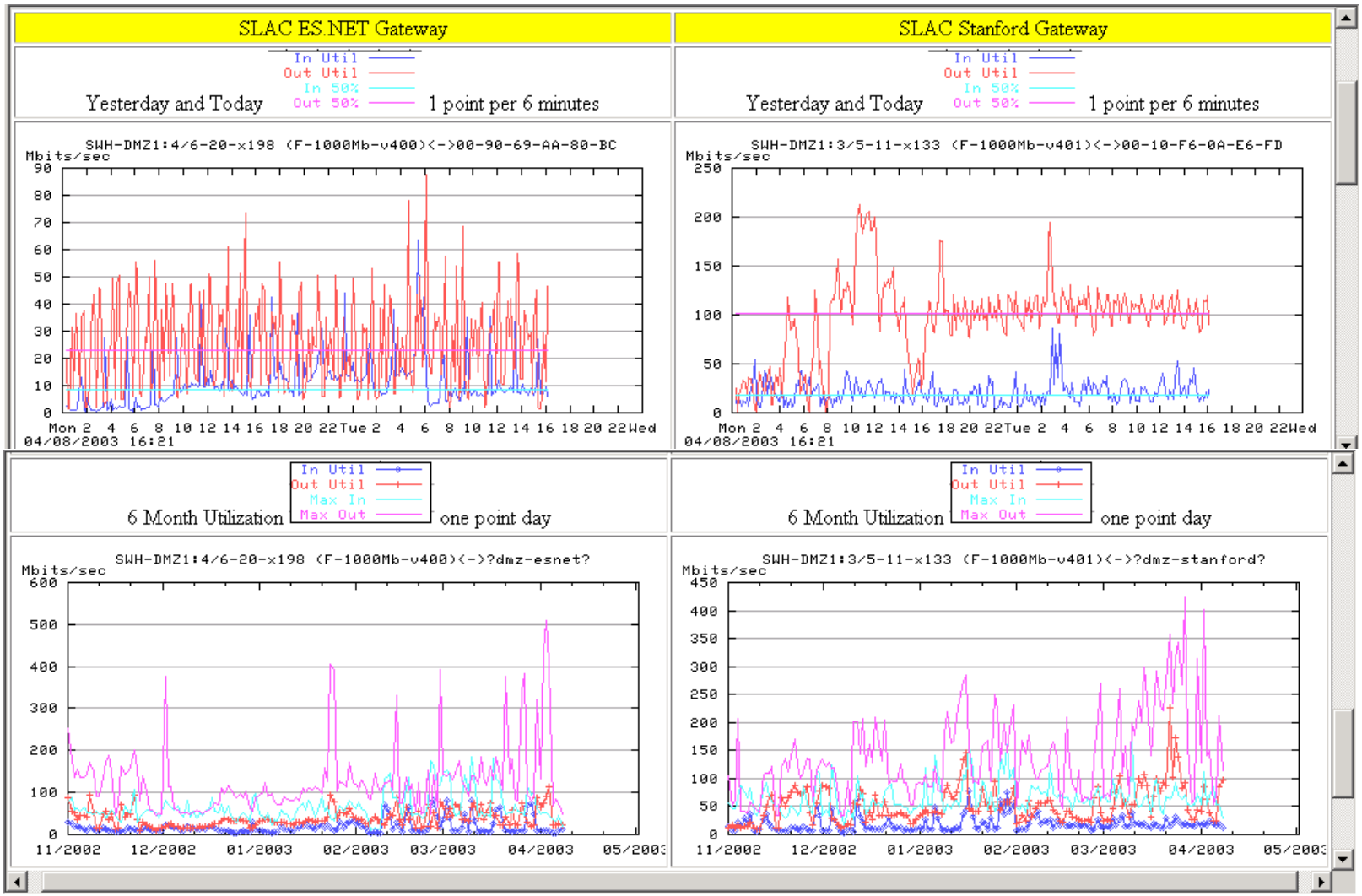
Farm/Server Network

SLAC External Network (April 8, 2003)

622 Mbits/ to ESNet

622 Mbits/s to Internet 2

~ 120 Mbits/s average traffic

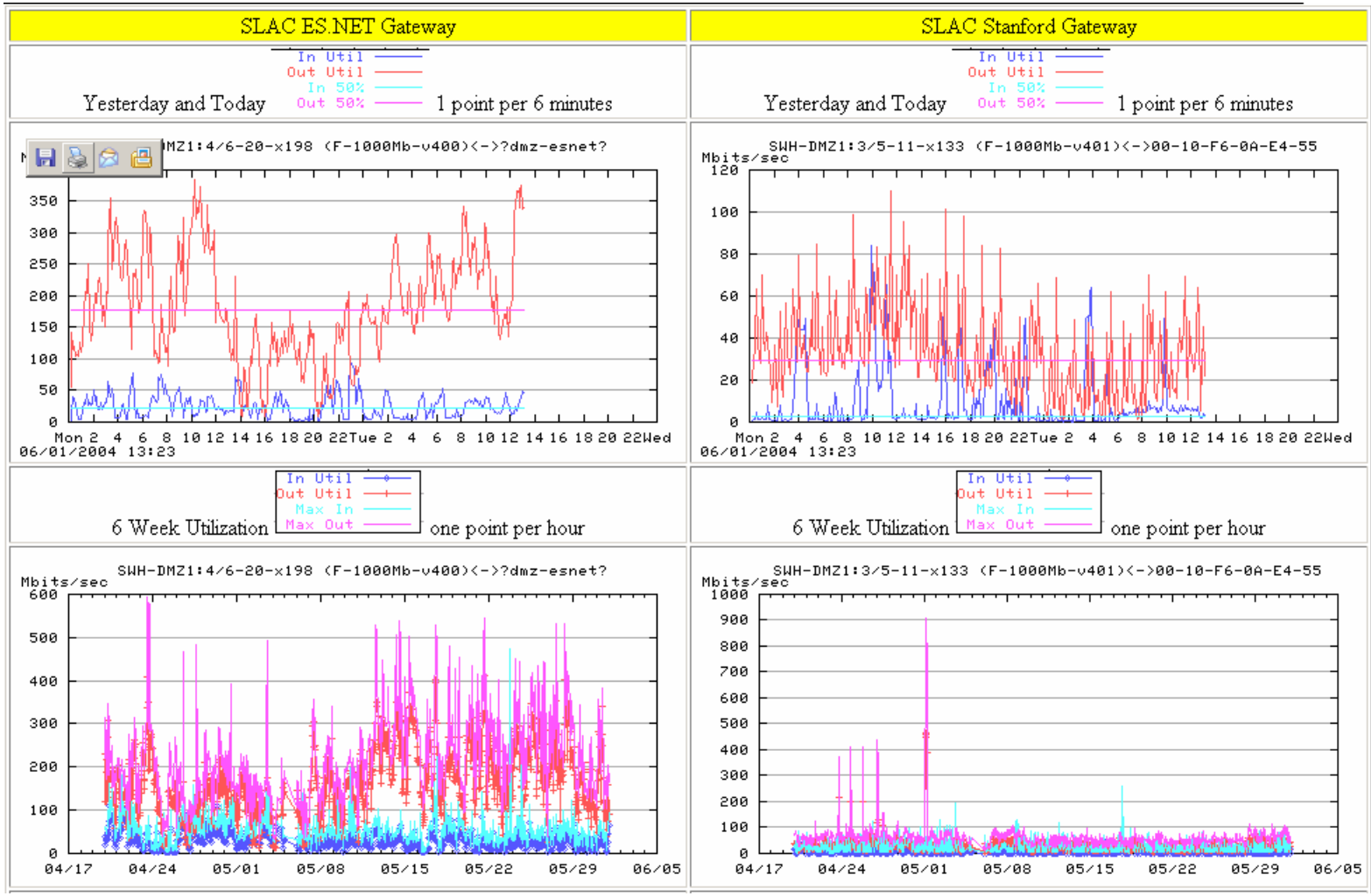


SLAC External Network (June 1, 2004)

622 Mbits/ to ESNet

1000 Mbits/s to Internet 2

~210 Mbits/s average traffic



Infrastructure Issues

No fundable research here!

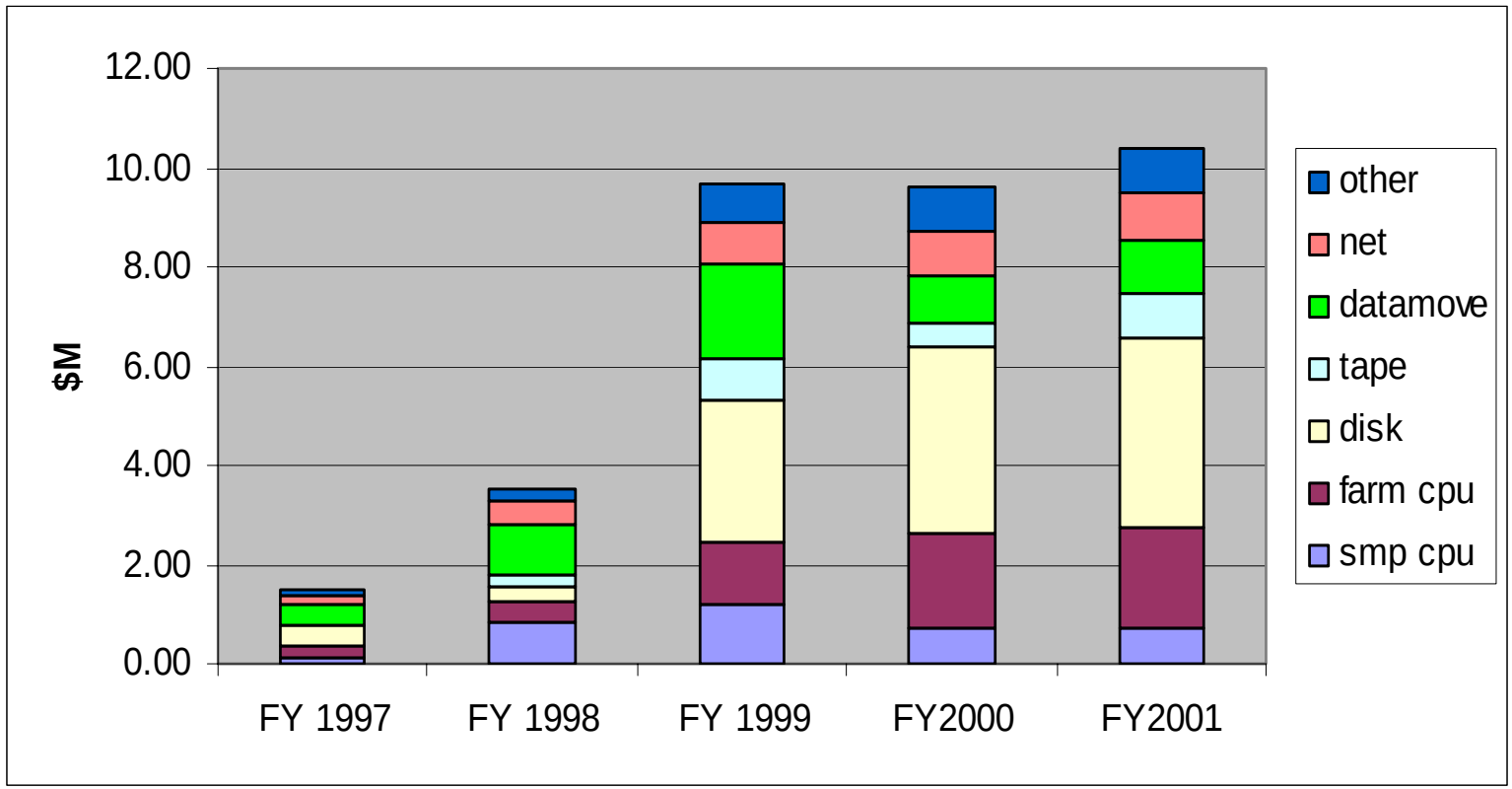
- Power and Cooling
  - UPS system (3x225 KVA) installed, additional capacity planned;
  - Diesel generator postponed *sine die;*
  - Most of available 1500 KVA in use;
  - Power monitoring system almost complete;
  - New 4.0MVA substation almost complete;
  - Cooling capacity close to limit (installing additional raised-floor “air handlers”);
  - Planning further power and cooling upgrades for 2004 on;
  - Logistics of power/cooling installations/modifications are horrendous (24x365 operation).
- Seismic
  - Computer center built like a fort;
  - Raised floor is (by far) the weakest component;
  - Phased (2-year) replacement now underway;
- Space
  - Extension to computer building in 2007-11 plan;
  - Exploring use of cheap commercial space to ease near-term pressures and logistics.

From April 2000 DOE Review

BaBar Offline Computing at SLAC:

Costs other than Personnel

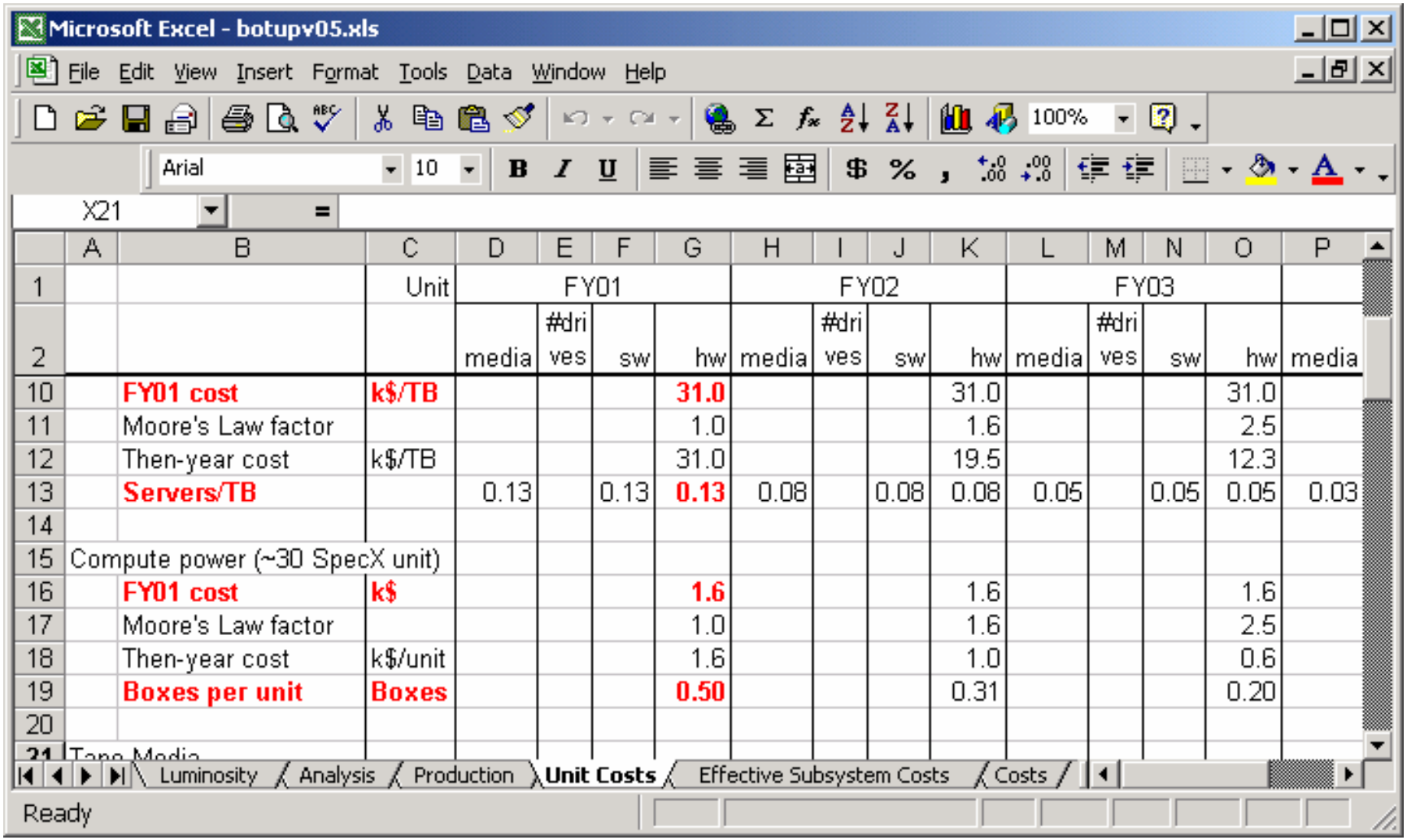
(does not include "per physicist" costs such as desktop support, help desk, telephone, general site network)



Does not include tapes

Bottom-up Cost Estimate

December 2000, January 2002, January 2003, January 2004



Microsoft Excel - botupv05.xls

File Edit View Insert Format Tools Data Window Help

Arial 10 100%

X21 =

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1			Unit	FY01				FY02				FY03				
2				media	#drives	sw	hw	media	#drives	sw	hw	media	#drives	sw	hw	media
10		**FY01 cost**	**k$/TB**				**31.0**				31.0				31.0	
11		Moore's Law factor					1.0				1.6				2.5	
12		Then-year cost	k$/TB				31.0				19.5				12.3	
13		**Servers/TB**		0.13		0.13	**0.13**	0.08		0.08	0.08	0.05		0.05	0.05	0.03
14																
15	Compute power (~30 SpecX unit)															
16		**FY01 cost**	**k$**				**1.6**				1.6				1.6	
17		Moore's Law factor					1.0				1.6				2.5	
18		Then-year cost	k$/unit				1.6				1.0				0.6	
19		**Boxes per unit**	**Boxes**				**0.50**				0.31				0.20	
20																
21	Tape Media															

Luminosity | Analysis | Production | **Unit Costs** | Effective Subsystem Costs | Costs

Ready

http://www-user.slac.stanford.edu/rmount/BaBar/botupv05.xls
http://www-user.slac.stanford.edu/rmount/BaBar/botup_jan02_final.xls
http://www-user.slac.stanford.edu/rmount/babar/botup_ifc_jan15_03.xls
http://www-user.slac.stanford.edu/rmount/babar/botup_dec03_v04.xls

Computing Model Approach

- Production:
  - OPR Must keep up with *Peak* Luminosity
  - Reprocessing must keep up with *Integrated* Luminosity
  - Skimming must keep up with *Integrated* Luminosity
- Analysis:
  - Must keep up with *Integrated* Luminosity
    (Must be able to re-analyze all previous year's data plus analyze this year's data during this year.)
- Simulation
  - Capacity to simulate 3 x hadronic data sample
  - Simulation capacity not costed (mainly done at universities)
  - Analysis capacity for simulated data is costed in the model

Costing the BaBar Computing Model

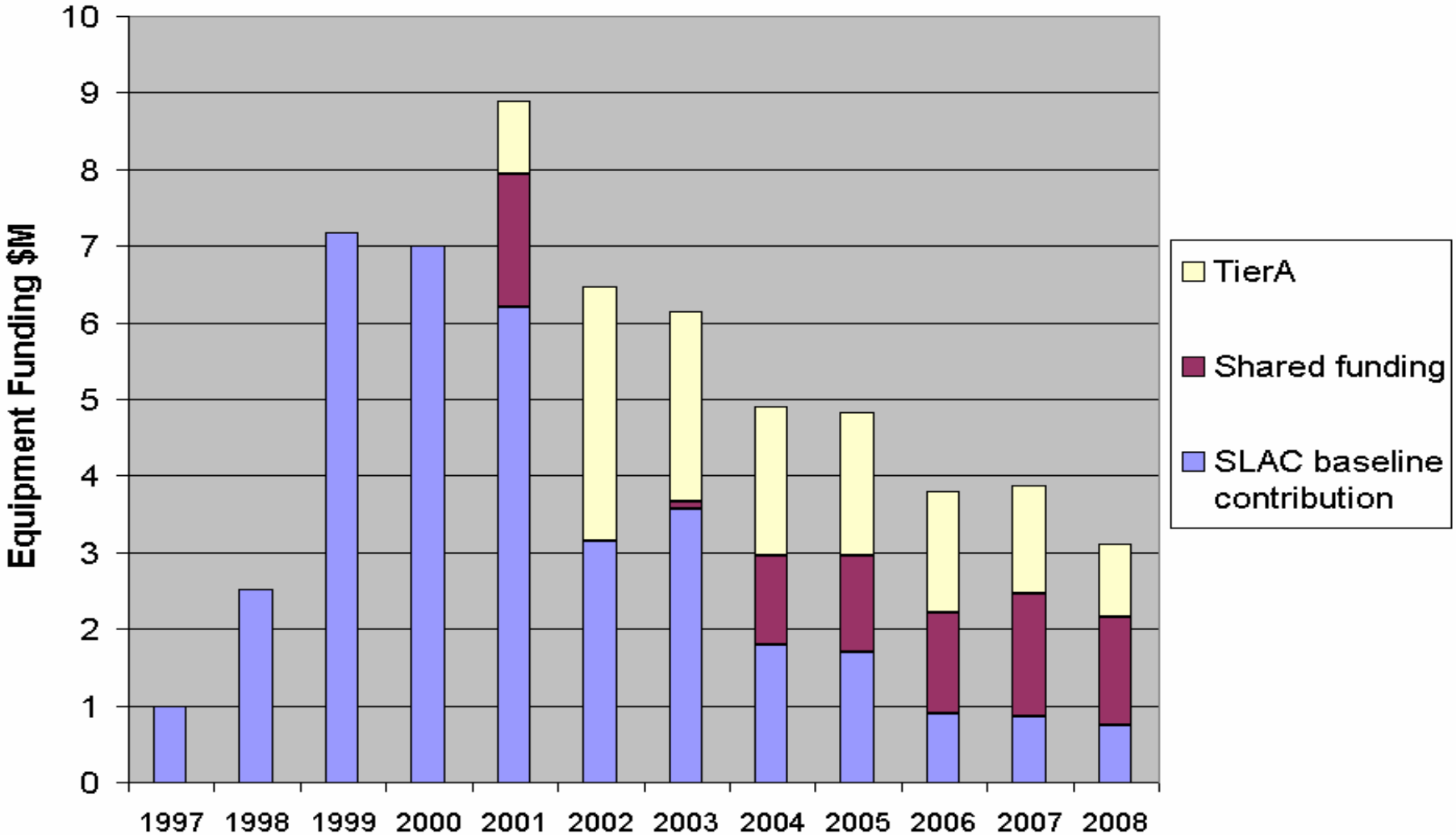
Major drivers of analysis cost:

1. Disk arrays (plus servers, fiberchannel, network, racks …)
2. CPU power (plus network, racks, services …)

Major subsidiary cost:

– Tape drives (plus servers, fiberchannel, network, racks …).
Driven by disk-cache misses due to analysis CPU I/O

BaBar Offline Computing Equipment
Bottom-up Cost Estimate (December 2003)
(To be revised annually)



The Science of Scientific Computing

- Between
  - The commercial IT offering (hardware and software) and
  - The “application science”
- The current SLAC “application” is principally experimental high-energy physics
  - Geographically distributed
  - Huge volumes of data
  - Huge real-time data rates
- Future SLAC growth areas include
  - Astrophysics
    - Data-intensive sky surveys – LSST …
    - Simulation – computational cosmology and astrophysics
  - SSRL Program
    - The explosion of compute and data-intensive biology
  - Accelerator Physics: A simulation and instrumentation-intensive future

Research Areas (1)

(Funded by DOE-HEP and DOE SciDAC and DOE-MICS)

- Scalable Data-Intensive Systems
  - "The world's largest database (OK not really a database any more)"
  - How to maintain performance with data volumes growing like "Moore's Law"?
  - How to improve performance by factors of 10, 100, 1000, … ? (intelligence plus brute force)
  - Robustness, load balancing, troubleshootability in 1000 – 10000-box systems.
- Grids and Security:
  - **PPDG**: Building the US HEP Grid – OSG;
  - Security in an open scientific environment;
  - Monitoring, troubleshooting and robustness.

Research Areas (2)

(Funded by DOE-HEP and DOE SciDAC and DOE MICS)

- Network Research (and stunts) – Les Cottrell
  - Land-speed record and other trophies
- Internet Monitoring and Prediction:
  - **IEPM:** Internet End-to-End Performance Monitoring (~5 years)
    SLAC is the/a top user of ESNet and the/a top user of Internet2.
    (Fermilab doesn't do so badly either)
  - **INCITE**: Edge-based Traffic
    Processing and Service Inference for High-Performance Networks
- GEANT4: Simulation of particle interactions in million to billion-element geometries
  - BaBar, GLAST, LCD …
  - LHC program
  - Space
  - Medical

Grids

The Particle Physics Data Grid Collaboratory Pilot

Proposal in Response to SciDAC Announcements LAB 01-06 and LAB 01-11 and Grant Notices 01-06 and 01-11

DOE Laboratory Contact:
Richard P. Mount, SLAC MS 97, 2575 Sand Hill Road, Menlo Park, CA 94025, Tel: 650 926 2467, Fax: 650 926 3329
University Contact:
Miron Livny, Computer Sciences Dept., Room 5372, University of Wisconsin-Madison, 1210 West Dayton St., Madison, WI 53706, Tel. 608-262-4694, Fax: 608-262-9777

Submitted March 15, 2001

Approved at $3.18M per year for 3 years

Renewal Proposal Submitted February 2004

Approved at $3.25M per year for 2 years

Particle Physics Data Grid

www.ppdg.net

PPDG

PARTICLE PHYSICS DATA GRID

ANL · BNL · Caltech · FNAL · JLAB · ISI · LBNL · SDSC · SLAC · Wisconsin · UCSD

- Search
- News & Events
- Teams
- Email lists
- PPDG RA
- PPDG at Work
- Papers
- Presentations
- Software
- Meetings
- Calendars
- SciDAC Contacts
- Other Grid Projects
- Management
- Acknowledgements

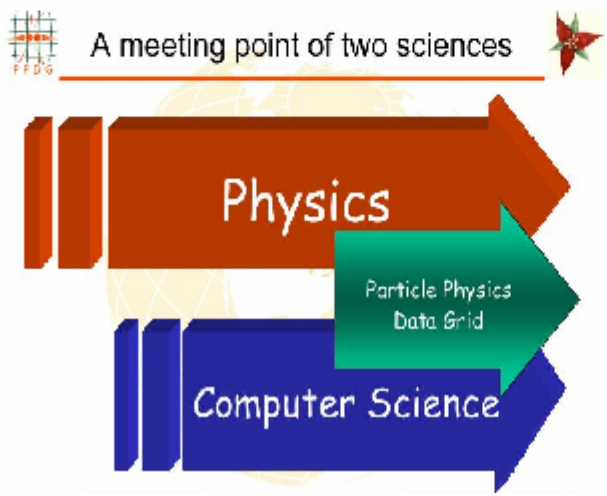
SciDAC Scientific Discovery

Notice to users

Webmaster

The Particle Physics Data Grid Collaboratory Pilot (PPDG) is developing and deploying production Grid systems vertically integrating experiment-specific applications, Grid technologies,Grid and facility computation and storage resources to form effective end-to-end capabilities. PPDG is a collaboration of computer scientists with a strong record in Grid technology, and physicists with leading roles in the software and network infrastructures for major high-energy and nuclear experiments. Our goals and plans are guided by the immediate and medium-term needs of the physics experiments and by the research and development agenda of the computer science groups. Ongoing status and achievements are given in the project's Quarterly Reports.

What's New:
D0 achieves physics results through reprocessing >500M events using SAMGrid, JIM and Grid tools.
Pre-Register Now for June 28,29 Collaboration Meeting.



PPDG has released the following News Items:

STAR Physics - Utilizing the Grid
Grid2003 - A Multi-VO Application Grid
Using the Virtual Data Toolkit
STAR/Hierachical Resource Manager
US/CMS Testbed Production
DZero uses DOESG certificates across the Atlantic

Sustained Production Data Movement over the Grid has resulted in: a factor of 2-10 more data transfer throughput, operational effort reduced by a factor of 2, a paradigm shift for distributed data processing from manual to

PPDG Project

- Just renewed for an additional two years
- Program of work has a significant new focus on creating and exploiting the “Open Science Grid” (OSG)
- OSG is, initially an ad-hoc effort by SLAC, Fermilab, Brookhaven to create a Grid based on existing computation, storage and network resources
- OSG builds on and learns from Grid 2003

SLAC-BaBar-OSG

- BaBar-US has been:
  - Very successful in deploying Grid data distribution (SRB US-Europe)
  - Far behind BaBar-Europe in deploying Grid job execution (in production for simulation)
- SLAC-BaBar-OSG plan
  - Focus on achieving massive simulation production in US within 12 months
  - make 1000 SLAC processors part of OSG
  - Run BaBar simulation on SLAC and non-SLAC OSG resources

GEANT4 at SLAC

Geant4 Members at SLAC

Last modified : *03/11/2004 10:17:24*

Member	Working Groups	Other Duties
Makoto Asai	Global Architecture, Run&Event, Detector Response, Intercoms	Deputy Spokesman, Tech. Steering Board
Mark Donszelmann	Interfaces, Visualization	
Tony Johnson	Interfaces, Analysis	
Tatsumi Koi	Hadronics	
Willy Langeveld	Processes, Materials	
Bill Lockman (UC Santa Cruz)	Geometry	
Richard Mount		Collaboration Board Chairman
Joseph Perl	Visualization	Tech. Steering Board
Terry Schalk (UC Santa Cruz)		Collaboration Board
Doug Smith	Documentation	
Max Turri	Interfaces, Visualization	
David Williams (UC Santa Cruz)	Geometry and Transport	
Dennis Wright	Documentation, Hadronic	Tech. Steering Board

SLAC Computing Philosophy

- Achieve and maintain collaborative leadership in computing for high-energy physics
- Exploit our strength in the science of applying IT to science, wherever it is synergistic with SLAC's mission
- Make SLAC an attractor of talent – a career enhancing experience, and fun.

A Leadership-Class Facility for Data-Intensive Science

Richard P. Mount

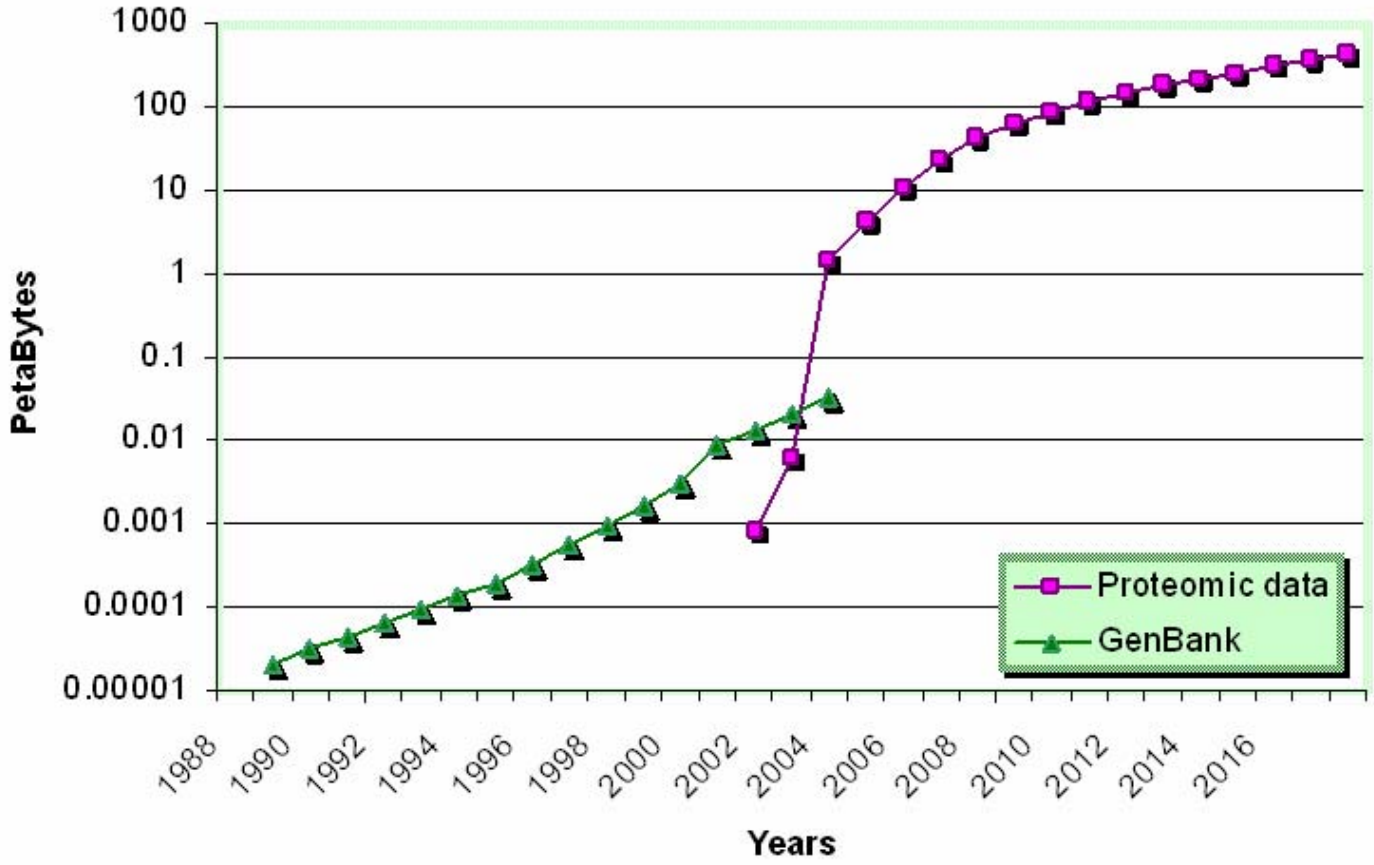
Director, SLAC Computing Services
Assistant Director, SLAC Research Division

Washington DC, April 13, 2004

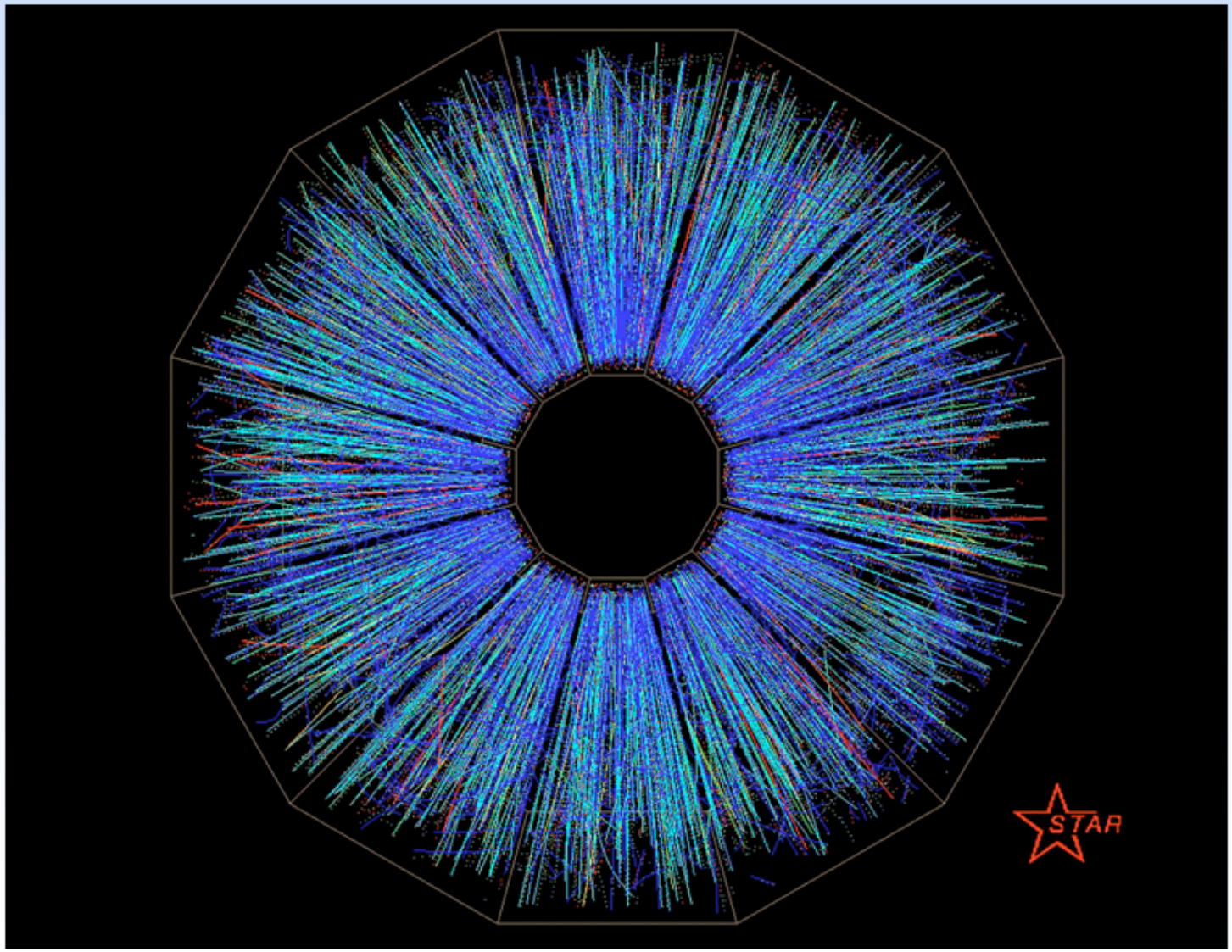
Outline

- The Science Case for a Leadership-Class Initiative
  - DOE Office of Science Data Management Workshop
    Richard Mount
  - Astronomy and Astrophysics
    Roger Blandford/Director KIPAC
  - High-energy physics – Babar
    David Leith/SLAC
- Proposal Details (Richard Mount)
  - Characterizing scientific data
  - Technology issues in data access
  - The “solution” and the strategy
  - Development Machine
  - Leadership Class Machine

Growth of Proteomic Data vs. Sequence Data



Digitized Event In STAR at RHIC



NP Analysis Limitations (2)

- It seems that it is less frequently possible to do Topological Analyses in NP than in HEP so Statistical Analyses are more often required
  - Evidence for this is rather anecdotal – not all would agree
  - To the extent that it is true, final analysis data sets tend to be large
  - These are the data sets accessed very frequently by large numbers of users … thus exacerbating the data management problem
- In any case the extraction and the delivery of distilled data subsets to physicists for analysis currently most limits NP analyses

Data management challenges for combustion science

- 2D complex chemistry simulations today: 200 restart files ($x, y, Z_1, \ldots Z_{50}$) skeletal n-heptane 41 species, 2000x2000 grid, 1.6 Gbytes/time x200 files = 0.32 Tbyte, 5 runs in parametric study 1.6 Tbytes raw data
- Processed data: 2 Tbyte data

- 3D complex chemistry simulations in 5 years: 200 restart files ($x, y, Z_1, \ldots Z_{50}$) skeletal n-heptane 41 species, 2000x2000x2000 grid, 3.2 Tbytes/time x 200 files = 640 Tbytes per run, 5 runs = 3.2 Petabytes raw data
- Processed data: 3 Petabytes
- Combustion regions of interest are spatially sparse
- Feature-borne analysis and redundant subsetting of data for storage
- Provenance of subsetted data
- Temporal analysis must be done on-the-fly
- Remote access to transport subsets of data for local analysis and viz.

Today's HENP Data Management Challenges

- Sparse access to objects in petabyte databases:
  - Natural object size 100 bytes – 10 kbytes
  - Disk (and tape) non-streaming performance dominated by latency
  - Approach - current:
    - Instantiate richer database subsets for each analysis application
  - Approaches – possible
    - Abandon tapes (use tapes only for backup, not for data-access)
    - Hash data over physical disks
    - Queue and reorder all disk access requests
    - Keep the hottest objects in (tens of terabytes of) memory
    - etc.

Computing at BaBar

- Computing for BaBar is done at five Tier A sites around the world – specifically in the US/SLAC; France/Lyon; Italy/Padua; UK/Rutherford Lab and Germany/Karlsruhe.

- 400 cpu's on production [300 It; 100 US]
- 1500 cpu's on simulation [Tier-A's and ~20 university sites, ½ in the U.S.]
- 2500 cpu's on skimming [2100 US; 250 Ge; 150 Fr]
- 1250 cpu's on analysis [750 US;250 Fr; 250 UK]

Managing Data for the World Wide Telescope aka: The Virtual Observatory

Jim Gray

Alex Szalay

SLAC Data Management Workshop

- Download (FTP and GREP) are not adequate
  - You can GREP 1 MB in a second
  - You can GREP 1 GB in a minute
  - You can GREP 1 TB in 2 days
  - You can GREP 1 PB in 3 years.
- Oh!, and 1PB ~3,000 disks
- At some point we need
  indices to limit search
  parallel data search and analysis
- This is where databases can help
- Next generation technique: **Data Exploration**
  - Bring the analysis to the data!

The Proposal

A Leadership Class Facility for Data-Intensive Science

Characterizing Scientific Data

My petabyte is harder to analyze than your petabyte

- Images (or meshes) are bulky but simply structured and usually have simple access patterns
- Features are perhaps 1000 times less bulky, but often have complex structures and hard-to-predict access patterns

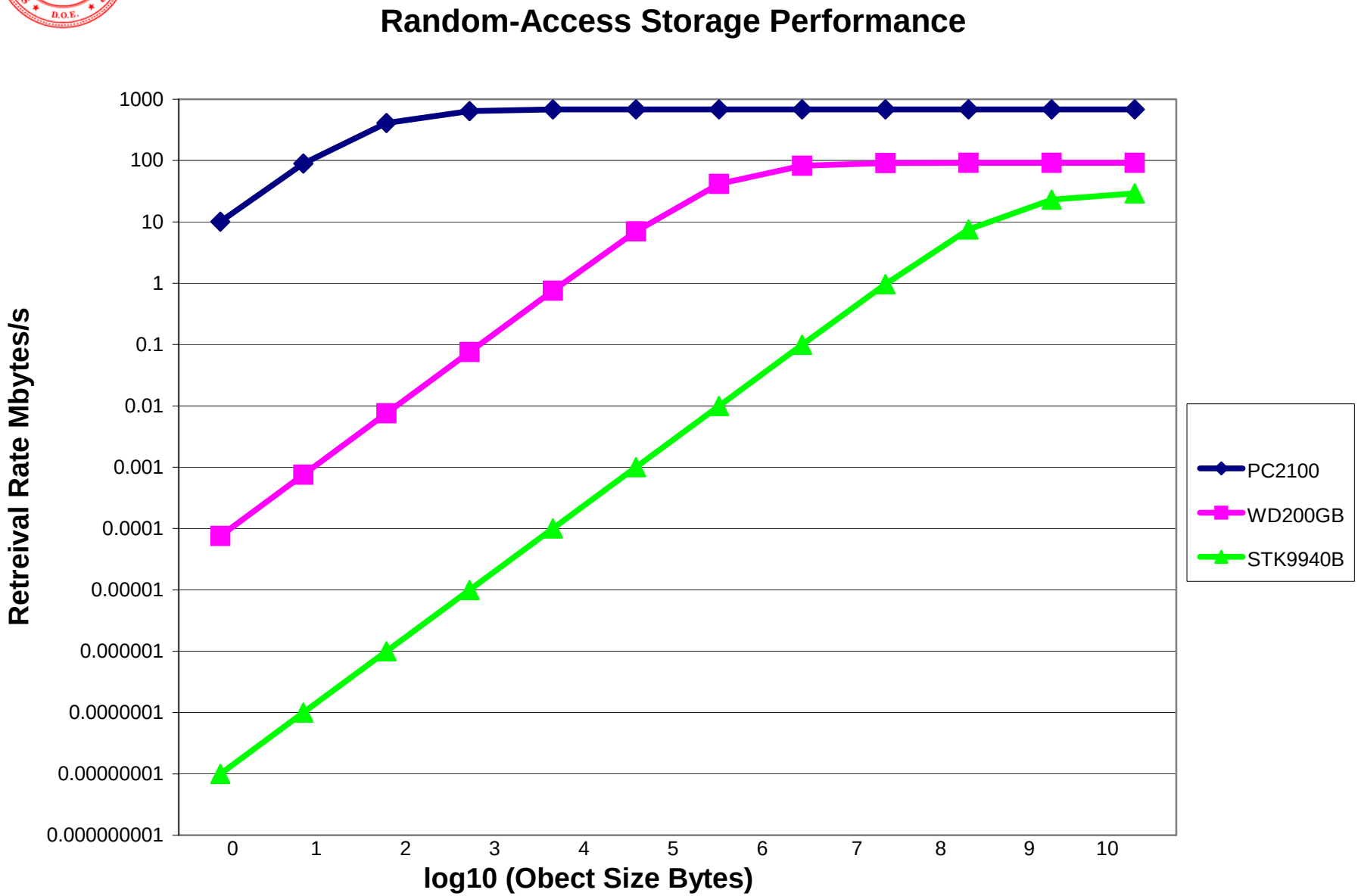
Characterizing Scientific Data

- This proposal aims at revolutionizing the query and analysis of scientific databases with complex structure.
- Generally this applies to feature databases (terabytes–petabytes) rather than bulk data (petabytes–exabytes)

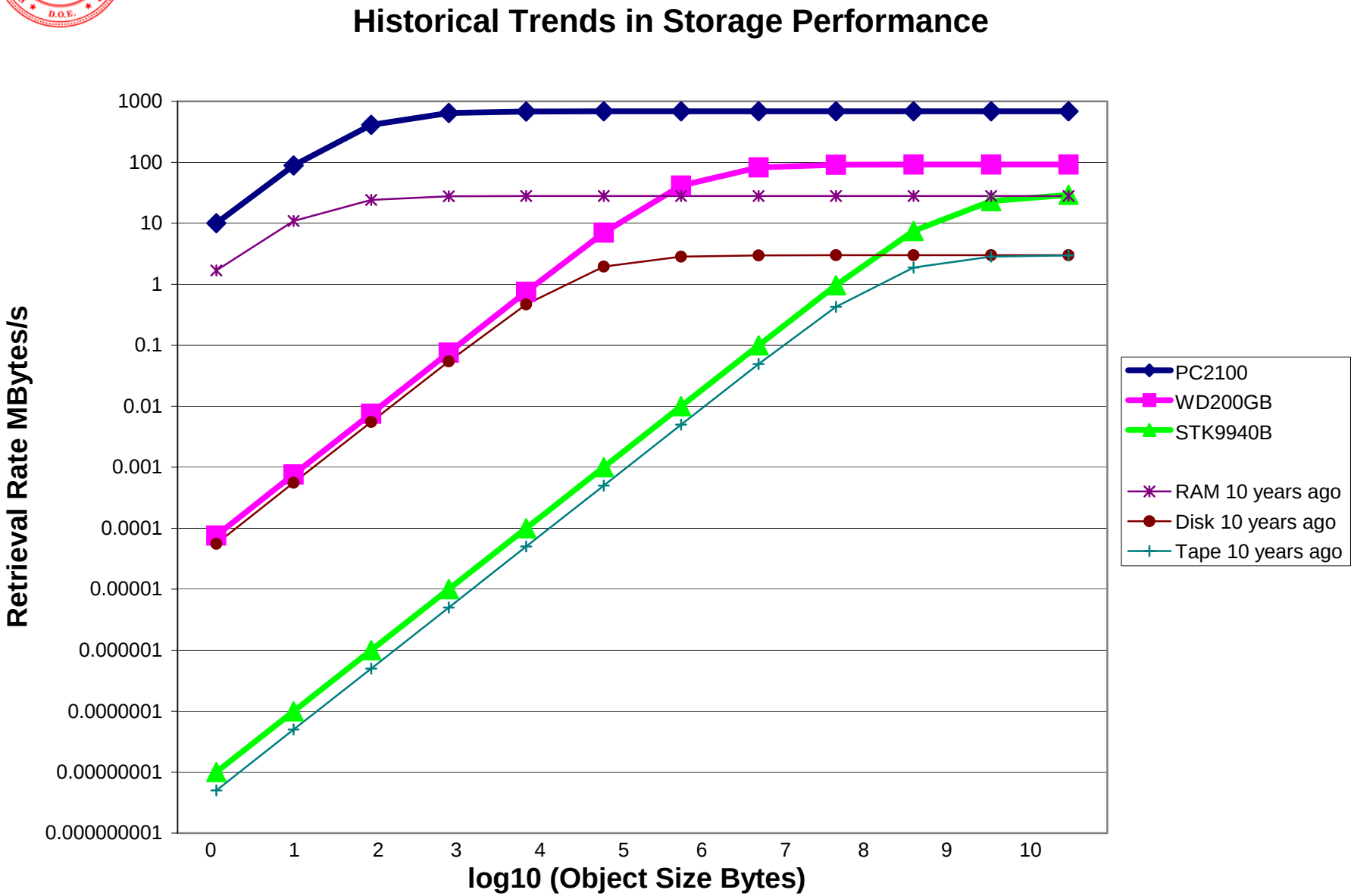
Technology Issues in Data Access

- Latency
- Speed/Bandwidth
- (Cost)
- (Reliabilty)

Latency and Speed – Random Access



Latency and Speed – Random Access



Storage Issues

- Disks:
  - Random access performance is lousy, unless objects are megabytes or more
    - independent of cost
    - deteriorating with time at the rate at which disk capacity increases

(Define random-access performance as time taken to randomly access entire contents of a disk)

The “Solution”

- Disk storage is lousy and getting worse
- Use memory instead of disk
  (“Let them eat cake”)
- Obvious problem:
  - Factor ≥ 100 in cost
- Optimization:
  - Brace ourselves to spend (some) more money
  - Architecturally decouple data-cache memory from high-performance, close-to-the-processor memory
  - Lessen performance-driven replication of disk-resident data

The Strategy

- There is significant commercial interest in an architecture including data-cache memory
- **But**: from interest to delivery will take 3-4 years
- **And**: applications will take time to adapt not just codes, but their whole approach to computing, to exploit the new architecture
- **Hence**: two phases
  1. Development phase (years 1,2,3)
     - Commodity hardware taken to its limits
     - BaBar as principal user, adapting existing data-access software to exploit the configuration
     - BaBar/SLAC contribution to hardware and manpower
     - Publicize results
     - Encourage other users
     - Begin collaboration with industry to design the leadership-class machine
  2. Leadership-Class Facility (years 3,4,5)
     - New architecture
     - Strong industrial collaboration
     - Facility open to all

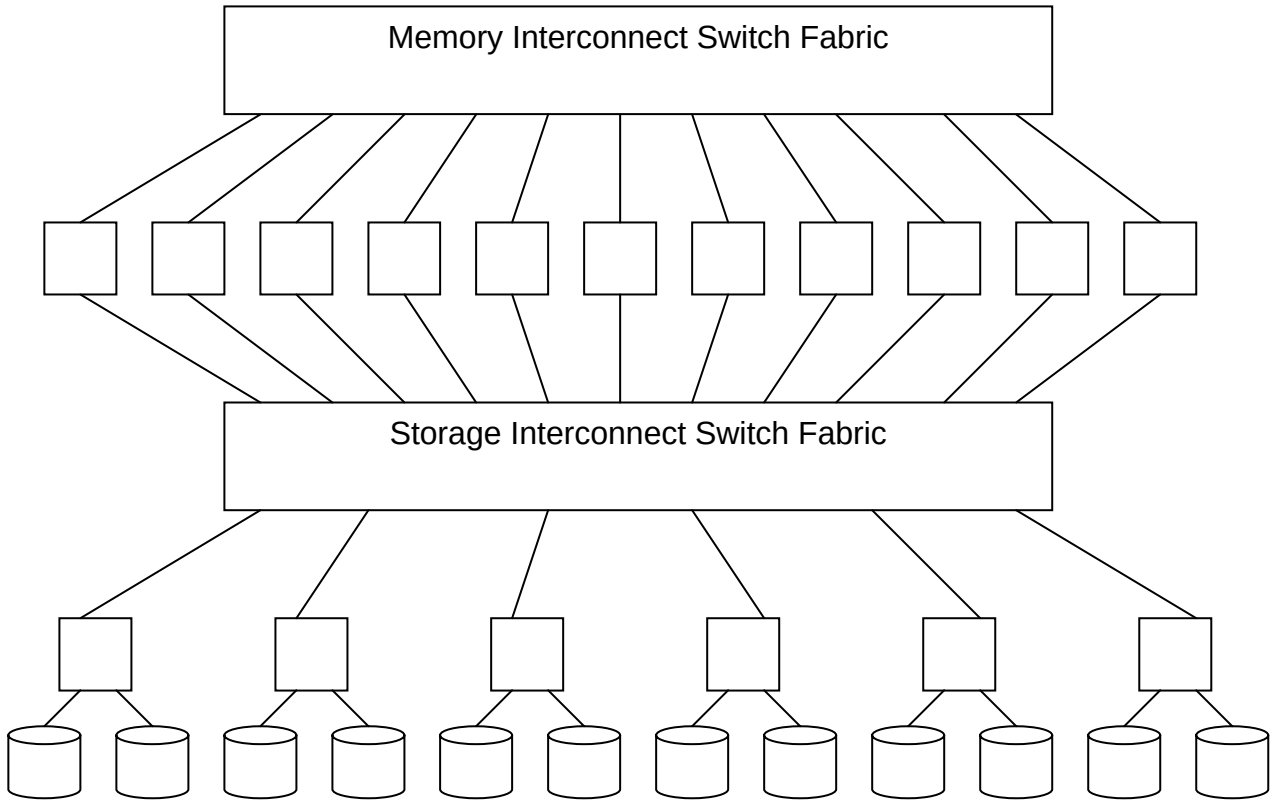
Development Machine Design Principles

- Attractive to scientists
  - Big enough data-cache capacity to promise revolutionary benefits
  - 1000 or more processors
- Processor to (any) data-cache memory latency < 100 μs
- Aggregate bandwidth to data-cache memory > 10 times that to a similar sized disk cache
- Data-cache memory should be 3% to 10% of the working set (approximately 10 to 30 terabytes for BaBar)
- Cost effective, but acceptably reliable
  - Constructed from carefully selected commodity components

Development Machine Design Choices

- Intel/AMD server mainboards with 4 or more ECC dimm slots per processor
- 2 Gbyte dimms (4 Gbyte too expensive this year)
- 64-bit operating system and processor
  - Favors Solaris and AMD Opteron
- Large (500+ port) switch fabric
  - Large IP switches are most cost-effective
- Use of ($10M+) BaBar disk/tape infrastructure, augmented for any non-BaBar use

Development Machine Deployment – Year 1



Cisco/Extreme/Foundry

650 Nodes, each
2 CPU, 16 GB memory

Cisco/Extreme/Foundry

> 100 Disk Servers
Provided by BaBar

BaBar/HEP Object-Serving Software

- AMS and XrootD (Andy Hanushevsky/SLAC)
  - Optimized for read-only access
  - Make 100s of servers transparent to user code
  - Load balancing
  - Automatic staging from tape
  - Failure recovery
- Can allow BaBar to start getting benefit from a new data-access architecture within months without changes to user code
- Minimizes impact of hundreds of separate address spaces in the data-cache memory

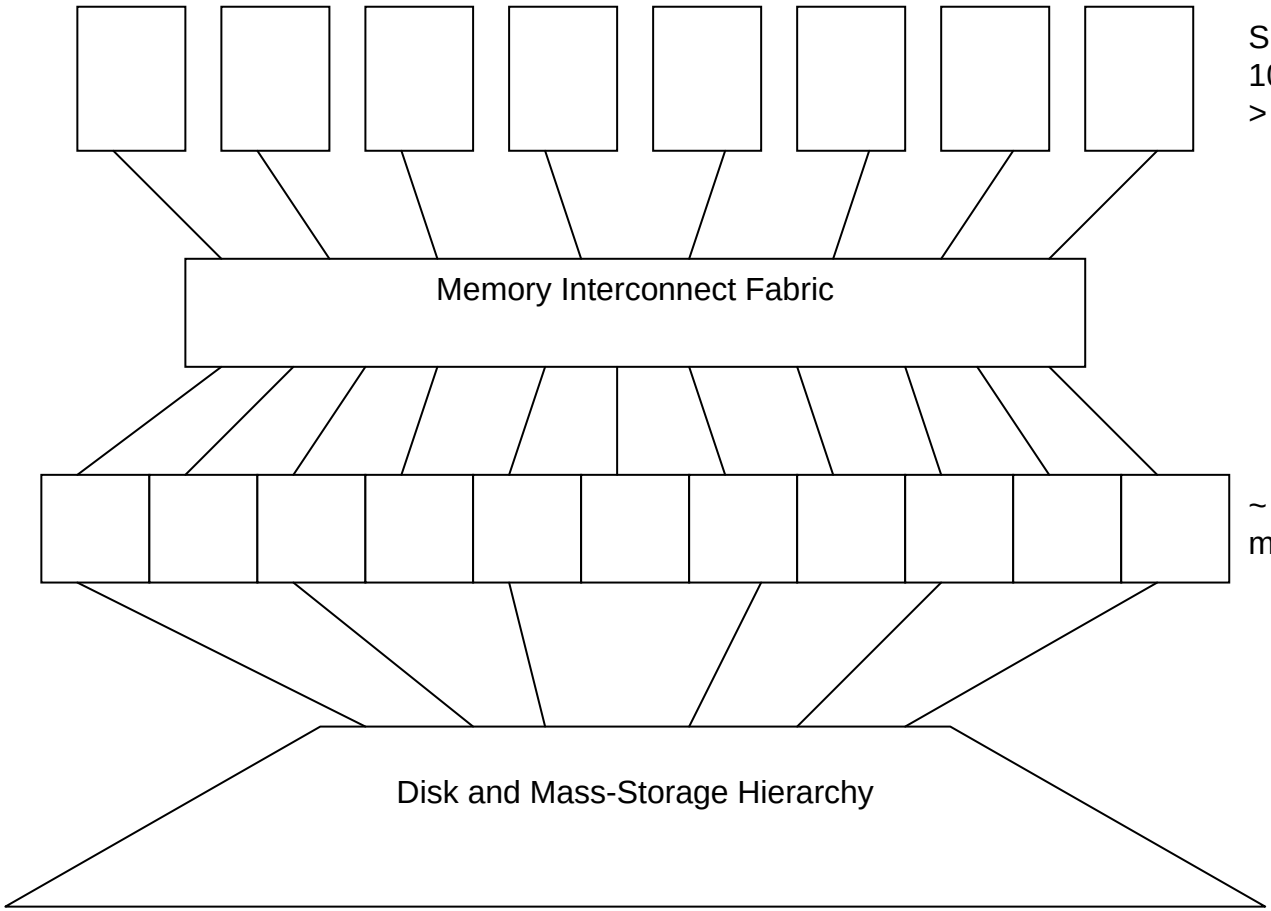
Leadership-Class Facility Design Principles

- All data-cache memory should be directly addressable by all processors
- Optimize for read-only access to data-cache memory
- Choose commercial processor nodes optimized for throughput
- Use the (then) standard high-performance memory within nodes
- Data-cache memory design optimized for reliable bulk storage
  - 5μs latency is low enough
  - No reason to be on the processor motherboard
- Operating system should allow transparent access to data-cache memory, but should also distinguish between high-performance memory and data-cache memory

Leadership-Class Facility Design Directions

- ~256 terabytes of data-cache memory and ~100 teraops/s by 2008
- Expandable by factor 2 in each of 2009,10,11
- Well-aligned with mainstream technologies but:
  - Operating system enhancements
  - Memory controller enhancements (read-only and coarse-grained locking where appropriate)
- Industry partnership essential
- Excellent network access essential
  - (SLAC is frequently the largest single user of *both* ESNet and Internet 2)
- Detailed design proposal to DOE in 2006

Leadership-Class Facility



SMP systems
100 teraops/s
> 10,000 threads

~ 256 terabytes data-cache memory

Summary

- The Office of Science is a leader in data-intensive science
- Data-intensive science will demand:
  - New architectures for its computing
  - Radically new approaches to exploiting these architectures
- We have presented an approach to
  - Creating a leadership facility for data-intensive science
  - Driving the revolutions in approaches to data analysis that will drive revolutions in science