LHC Computing Grid Project

PASTA III

26 September 2002

David Foster, CERN
David.foster@cern.ch

http://lcg.web.cern.ch/LCG/PEB/PASTAIII/pasta2002Report.htm

Approach to Pasta III

- Technology Review of what was expected from Pasta II and what might be expected in 2005 and beyond.
- Understand technology drivers which might be market and business driven. In particular the suppliers of basic technologies have undergone in many cases major business changes with divestment, mergers and acquisitions.
- Try to translate where possible into costs that will enable us to predict how things are evolving.
- Try to extract emerging best practices and use case studies wherever possible.
- Involve a wider number of people than CERN in major institutions in at least Europe and the US.

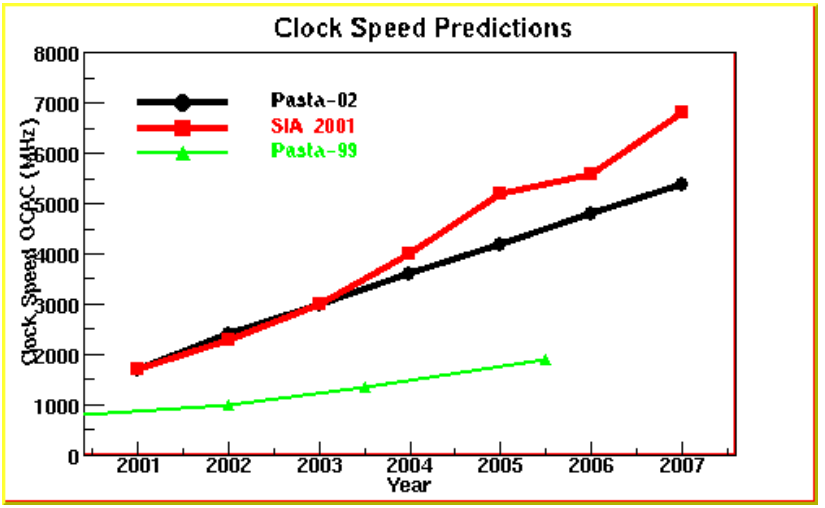
Participants

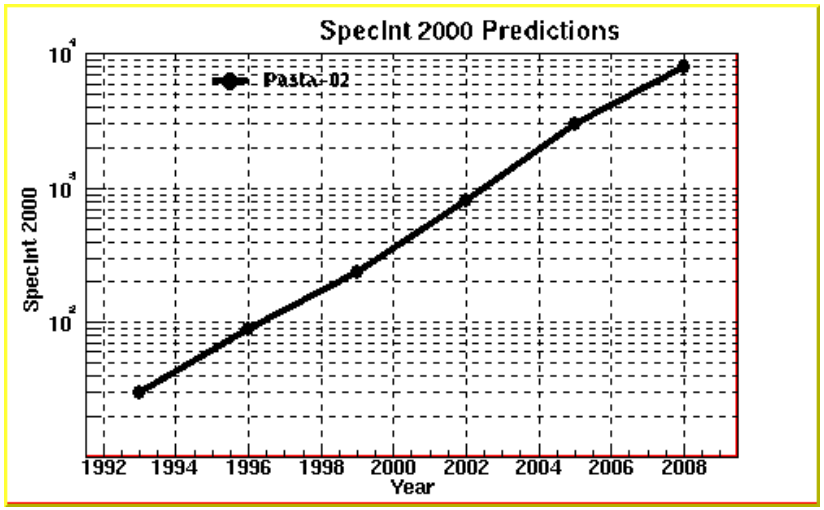
- A: Semiconductor Technology
  - Ian Fisk (UCSD) Alessandro Machioro (CERN) Don Petravik (Fermilab)
- B:Secondary Storage
  - Gordon Lee (CERN) Fabien Collin (CERN) Alberto Pace *(CERN)*
- C:Mass Storage
  - Charles Curran (CERN) Jean-Philippe Baud (CERN)
- D:Networking Technologies
  - Harvey Newman (Caltech) Olivier Martin (CERN) Simon Leinen *(Switch)*
- E:Data Management Technologies
  - Andrei Maslennikov (Caspur) Julian Bunn *(Caltech)*
- F:Storage Management Solutions
  - Michael Ernst (Fermilab) Nick Sinanis (CERN/CMS) Martin Gasthuber (*DESY*)
- G:High Performance Computing Solutions
  - Bernd Panzer (CERN) Ben Segal (CERN) Arie Van Praag *(CERN)*

- Editorial Work Gordon Lee

Basic System Components - Processors

- *1999 Pasta report was conservative in terms of clock speed BUT, clock speed is not a good measure, with higher clock speed CPU's giving lower performance in some cases*
- *Predictions beyond 2007 hard to make, CMOS device structures will hit limits within next 10 years, change from optical litho to electron projection litho required => new infrastructure*





Specint 2000 numbers for high-end CPU.
Not a direct correlation with CERN Units.
P4 Xenon = 824 SI2000 but only 600 CERN units

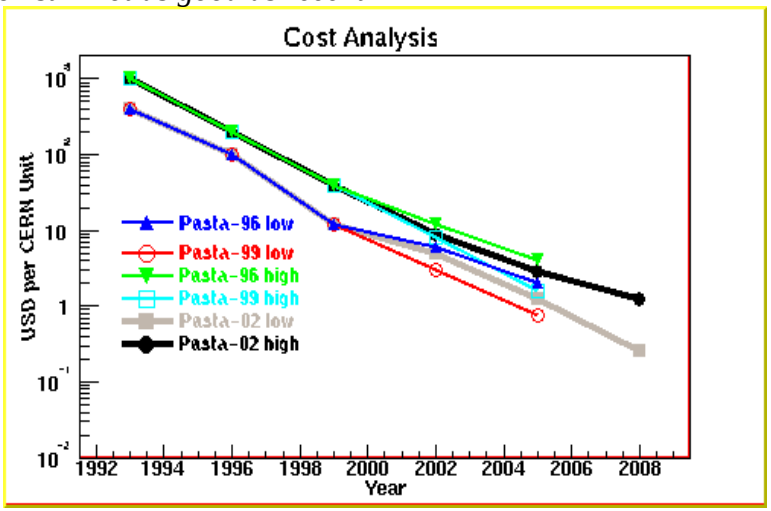
Compilers have not made great advances but Instruction Level Parallelism gives you now 70% usage (CERN Units) of quoted performance.

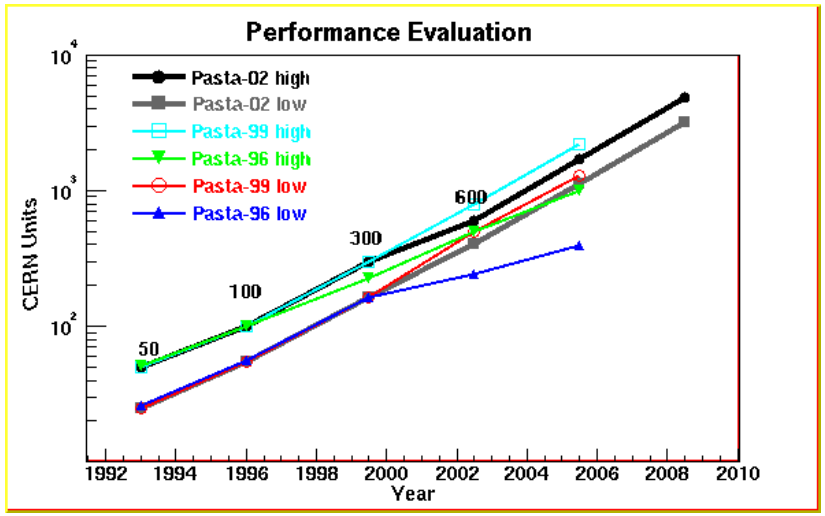
Basic System Components - Processors

Performance evolution and associated cost evolution for both High-end machines (15K$ for quad processor) and Low-end Machines (2K$ for dual CPU)

Note 2002 predictions revised down slightly from the 1999 Predictions of actual system performance

- *'99 report: expect 50% of what Intel quotes, trend holds*
- *with hyperthreading (P4 XEON) agrees with '96 predictions reducing the gap from 50% to 30%*
- *ILP has not increased significantly*
- *IA-64 still not as good as recent P4*





Fairly steep curve leading to LHC startup suggesting delayed purchases will save money (less CPU's for the same CU performance) as usual

Basic System Components - Interconnects

- PCI Developments
  - PCI 66/64 mostly on servers
  - PCI-X introduction slow
  - PCI-X 2 standard with 266 MHz (2.13 GB/s) and 533 MHz (4.26 GB/s)
    - Supports DDR and QDR technology
  - PCI Express (alias 3GIO, project Arapahoe)
    - Internal Serial Bus, NOT an Interconnect
    - Primarily for high-end systems
- New Interconnects
  - 3GIO, Intels industrial proposal
  - HyperTransport, AMD (12.8 GB/s asym, bi-directional, 64 bit Bus)
  - Chipset includes routing crossbar switch
  - Connection to outside to connect peripherals
  - Superior to Intel, but will the market accept it ?

Basic System Components Summary

- Memory capacity increased faster than predicted, costs around 0.15 $/Mbit in 2003 and 0.05 $/Mbit in 2006
- Many improvements in memory systems 300 MB/sec in 1999 now in excess of 1.2 GB/sec in 2002.
- PCI bus improvements improved from 130MB/sec in 1999 to 500 MB/second with 1GB/sec foreseen.
- Intel and AMD continue as competitors. Next generation AMD (Hammer) permits 32bit and 64bit code. And is expected to be 30% cheaper than equivalent Intel 64bit chips.
- The "2.5kCHF" dual processor cheap box will probably have an SI2k rating of about 3000/processor in 2007. We will need of the order of 4000 such boxes for LHC startup.

http://www.aceshardware.com/SPECmine/cpus.jsp#cpu1

Top 20 SPEC systems

	Top 20 SPECint2000					Top 20 SPECfp2000				
#	MHz	Processor	int peak	int base	Full results	MHz	Processor	fp peak	fp base	Full results
1	3000	Pentium 4	1200	1164	HTML	1700	POWER4+	1699	1598	HTML
2	1800	Opteron	1170	1095	HTML	1150	Alpha 21364	1482	1124	HTML
3	3066	Pentium 4 Xeon	1138	1089	HTML	1000	Itanium 2	1431	1431	HTML
4	1700	POWER4+	1113	1077	HTML	1250	Alpha 21264C	1365	1019	HTML
5	2200	Athlon XP	1080	1044	HTML	1350	SPARC64 V	1322	1047	HTML
6	1250	Alpha 21264C	928	845	HTML	1300	POWER4	1281	1200	HTML
7	1350	SPARC64 V	892	767	HTML	3000	Pentium 4	1229	1213	HTML
8	1150	Alpha 21364	877	795	HTML	1800	Opteron	1219	1122	HTML
9	1300	POWER4	848	822	HTML	1200	UltraSPARC III Cu	1118	953	HTML
10	2000	Athlon MP	766	737	HTML	3066	Pentium 4 Xeon	1063	1053	HTML
11	1200	UltraSPARC III Cu	722	642	HTML	2200	Athlon XP	982	873	HTML
12	1000	Itanium 2	719	711	HTML	1002	UltraSPARC IIIi	841	722	HTML
13	875	PA-RISC 8700+	678	642	HTML	833	Alpha 21264B	784	643	HTML
14	1400	Pentium III	664	648	HTML	800	Itanium	701	701	HTML
15	810	SPARC64 GP	624	512	HTML	875	PA-RISC 8700+	674	600	HTML
16	750	PA-RISC 8700	604	568	HTML	2000	Athlon MP	656	605	HTML
17	833	Alpha 21264B	571	497	HTML	833	Alpha 21264A	644	571	HTML
18	1002	UltraSPARC IIIi	555	485	HTML	810	SPARC64 GP	644	483	HTML
19	1400	Athlon	554	495	HTML	750	PA-RISC 8700	581	526	HTML
20	833	Alpha 21264A	533	511	HTML	600	MIPS R14000	529	499	HTML

Basic System Components Conclusions

- No major surprises so far, but
  - New semiconductor fab's very expensive squeezing the semiconductor marketplace.
  - MOS technology is pushing again against physical limits – gate oxide thickness, junction volumes, lithography, power consumption.
  - Architectural designs are not able to efficiently use the increasing transistor density (20% performance improvement)
  - A significant change in the desktop market machine architecture and form factor could change the economics of the server market.
- Do we need a new HEP reference application ?
  - Using industry benchmarks still do not tell the whole story and we are interested in throughput.
  - Seems appropriate with new reconstruction/analysis models and code

LTO Ultrium Roadmap

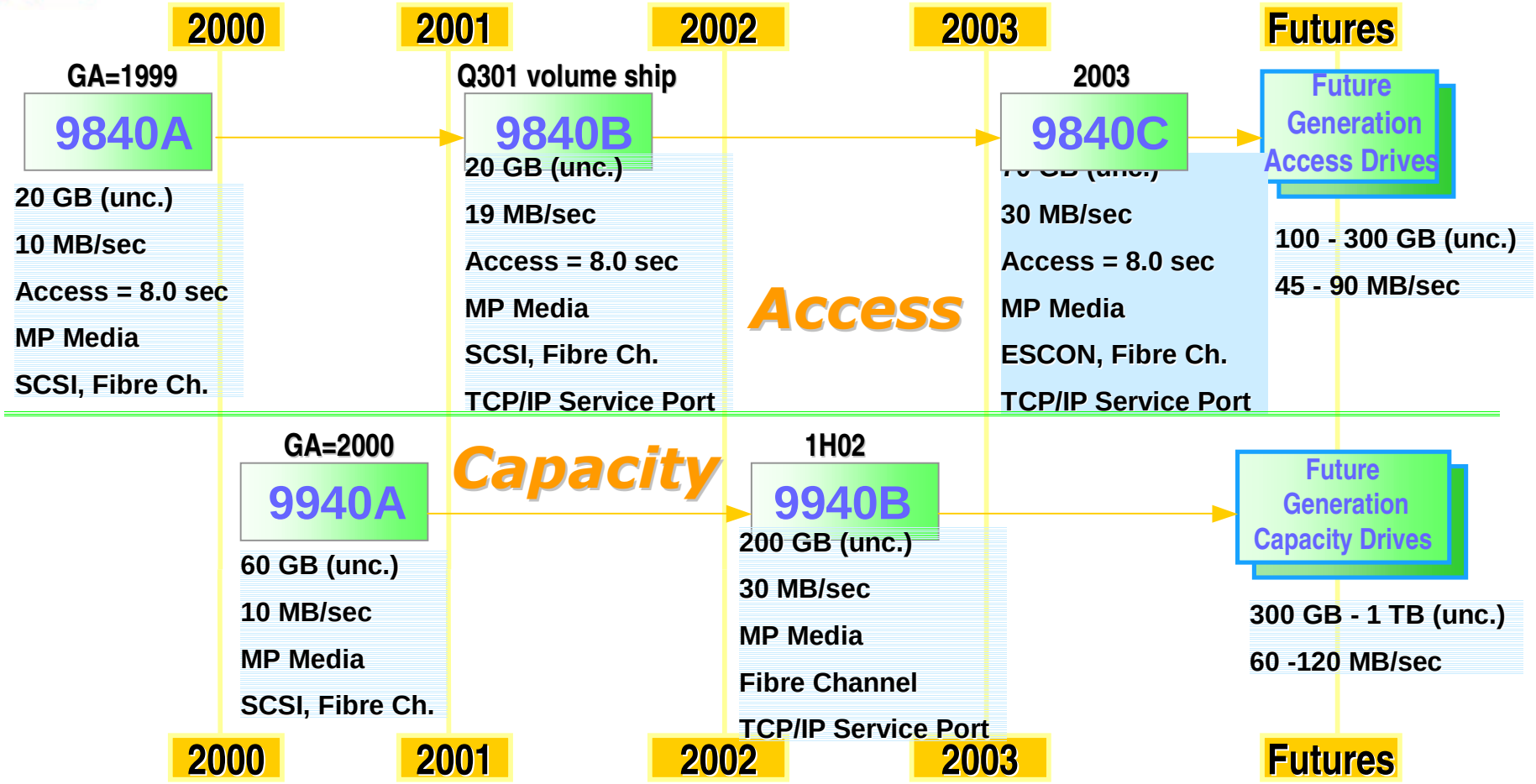
	Generation 1	Generation 2	Generation 3	Generation 4
	GA Start of Q3-00	18-24 Months after Gen1	18-24 Months after Gen2	
Capacity	100 GB	200 GB	400 GB	800 GB
Transfer Rate	10-20 MB/s	20-40 MB/s	40-80 MB/s	80-160 MB/s
Enabling Technology	?	?	?	?
Number of Channels	8	8	16	16
Recording Method	RLL 1,7	PRML	PRML	PRML
Media Type	MP2	MP	MP	Thin Film
Tape Length	580 m	580 m	800 m	800 m

Media Swap

Media Swap

Media Swap

9840 *and* 9940 Tape Drive Roadmap



Tapes - 1

- New format tape drives (9840, 9940, LTO) are being tested.
- Current Installation are 10 STK silo's capable of taking 800 new format tape drives. Today tape performance is 15MB/sec so theoretical aggregate is 12GB/sec
- Cartridge capacities expected to increase to 1TB before LHC startup but its market demand not technical limitations that is the driver.
- Using tapes as a random access device is a problem and will continue to be.
  - Need to consider a much larger, persistent disk cache for LHC reducing tape activity for analysis.

Tapes - 2

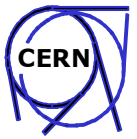
- Currents costs are about 50 CHF/slot for a tape in the Powderhorn robot.
- Current tape cartridge (9940A) costs 130 CHF with a slow decrease over time.
- Media dominates the overall cost and a move higher capacity cartridges and tape units sometimes require a complete media change.
  - Current storage costs 0.6-1.0 CHF/GB in 2000 could drop to 0.3 CHF/GB in 2005 but probably would require a complete media change.
- Conclusions: No major challenges for tapes for LHC startup but the architecture should be such that they are used better than today (write/read)

Network Progress and Issues for Major Experiments

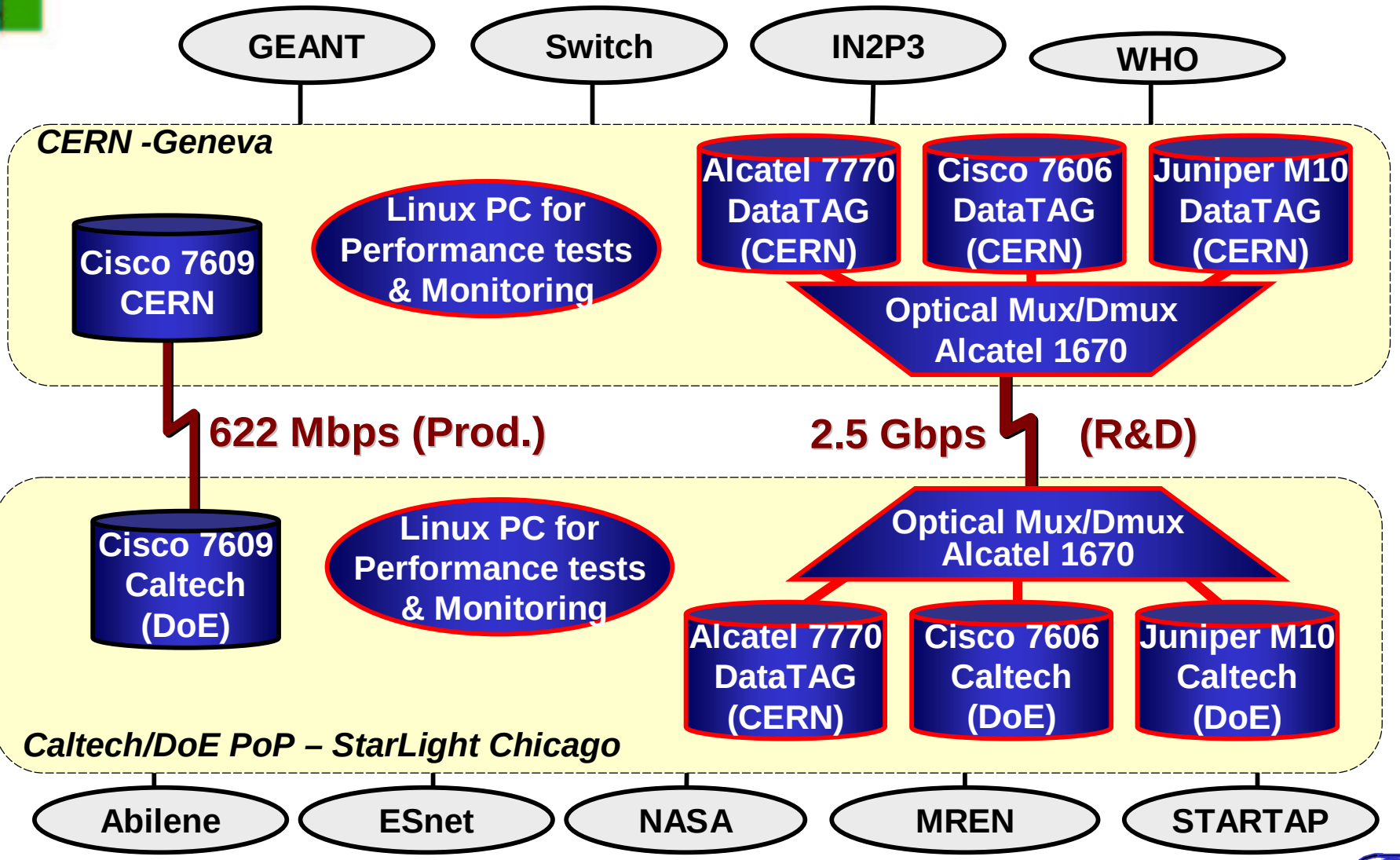
- Network backbones are advancing rapidly to the 10 Gbps range
  - "Gbps" end-to-end throughput data flows will be in production soon (in 1-2 years)
- Network advances are changing the view of the net's roles
  - This is likely to have a profound impact on the experiments' Computing Models, and bandwidth requirements
- Advanced integrated applications, such as Data Grids, rely on seamless "transparent" operation of our LANs and WANs
  - With reliable, quantifiable (monitored), high performance
  - Networks need to be integral parts of the Grid(s) design
- Need new paradigms of real network and system monitoring, and of new "managed global systems" for HENP analysis
  - These are starting to be developed for LHC

Network Progress and Issues for Major Experiments

- Key Providers in Bankruptcy
  - KPNQwest, Teleglobe, Global Crossing, FLAG; Worldcom
- Switching to Others, Where Needed and Possible
  - E.g. T-Systems (Deutsche Telecom) for US-CERN
- Strong Telecom Market Outlook
  - Good pricing from T-Systems
  - MCI/Worldcom network will continue to operate (?): 20 M customers in US; UK academic & research network
  - Aggressive plans by major and startup network equipment providers
- Strong Outlook in R&E Nets for Rapid Progress
  - Abilene (US) Upgrade On Schedule; GEANT (Europe), and SuperSINET(Japan) Plans Continuing
  - ESNet Backbone Upgrade: 2.5 Gbps "Now"; 10 Gbps in 2 Yrs.
  - Regional Progress, and Visions;
    - **E.g. CALifornia Research and Education Network: *"1 Gbps to Every Californian by 2010"***



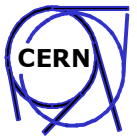
US Link: Late 2002



Development and tests

HENP Major Links: Bandwidth Roadmap (Scenario) in Gbps

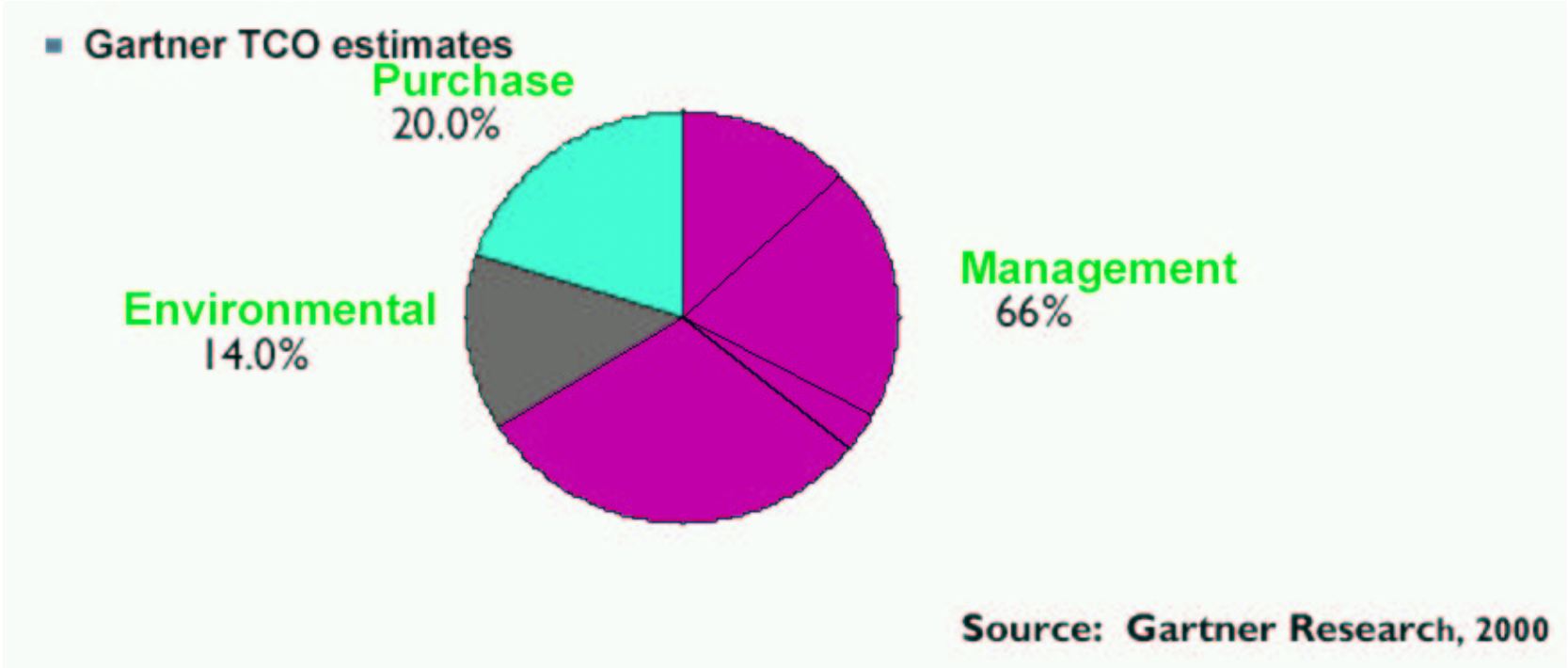
Year	Production	Experimental	Remarks
2001	0.155	0.622-2.5	SONET/SDH
2002	0.622	2.5	SONET/SDH DWDM; GigE Integ.
2003	2.5	10	DWDM; 1 + 10 GigE Integration
2005	10	2-4 X 10	λ Switch; λ Provisioning
2007	2-4 X 10	~10 X 10; 40 Gbps	1st Gen. λ Grids
2009	~10 X 10 or 1-2 X 40	~5 X 40 or ~20-50 X 10	40 Gbps λ Switching
2011	~5 X 40 or ~20 X 10	~25 X 40 or ~100 X 10	2nd Gen λ Grids Terabit Networks
2013	~Terabit	~MultiTerabit	~Fill One Fiber



Networking Summary

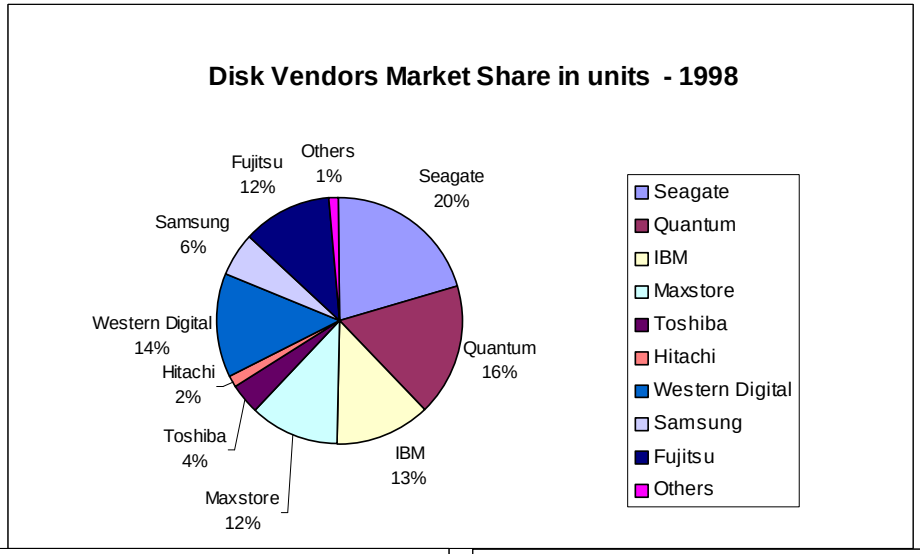
- Major cost reductions have taken place in wide-area bandwidth costs.
  - 2.5 Gbit common for providers but not academic in 1999. Now, 10Gbit common for providers and 2.5Gbit common for academic.
  - Expect 10GBit by end 2003. Vastly exceeds the target of 622 Mbit by 2005.
- Wide area data migration/replication now feasible and affordable.
  - Tests of multiple streams to the US running over 24hrs at the full capacity of 2Gbit/sec were successful.
- Local area networking moving to 10 Gbit/sec and this is expected to increase. 10Gbit/sec NIC's under development for end systems.

Storage Cost

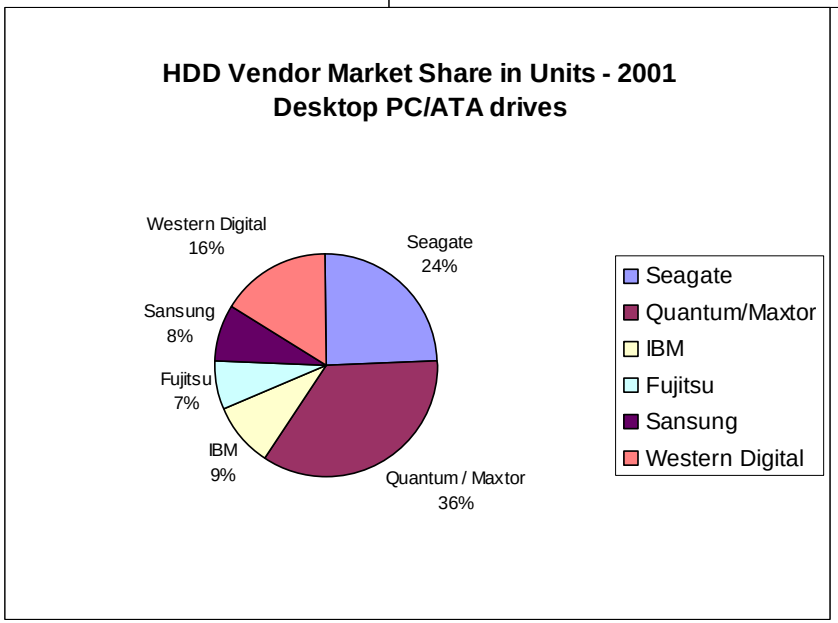


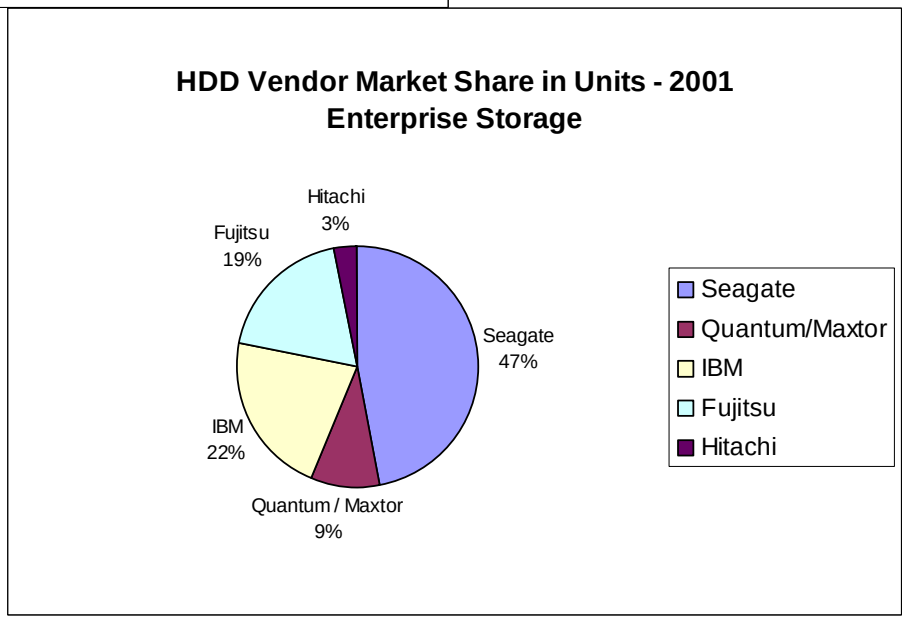
Cost of managing storage and data are the predominate costs

Disk Technology

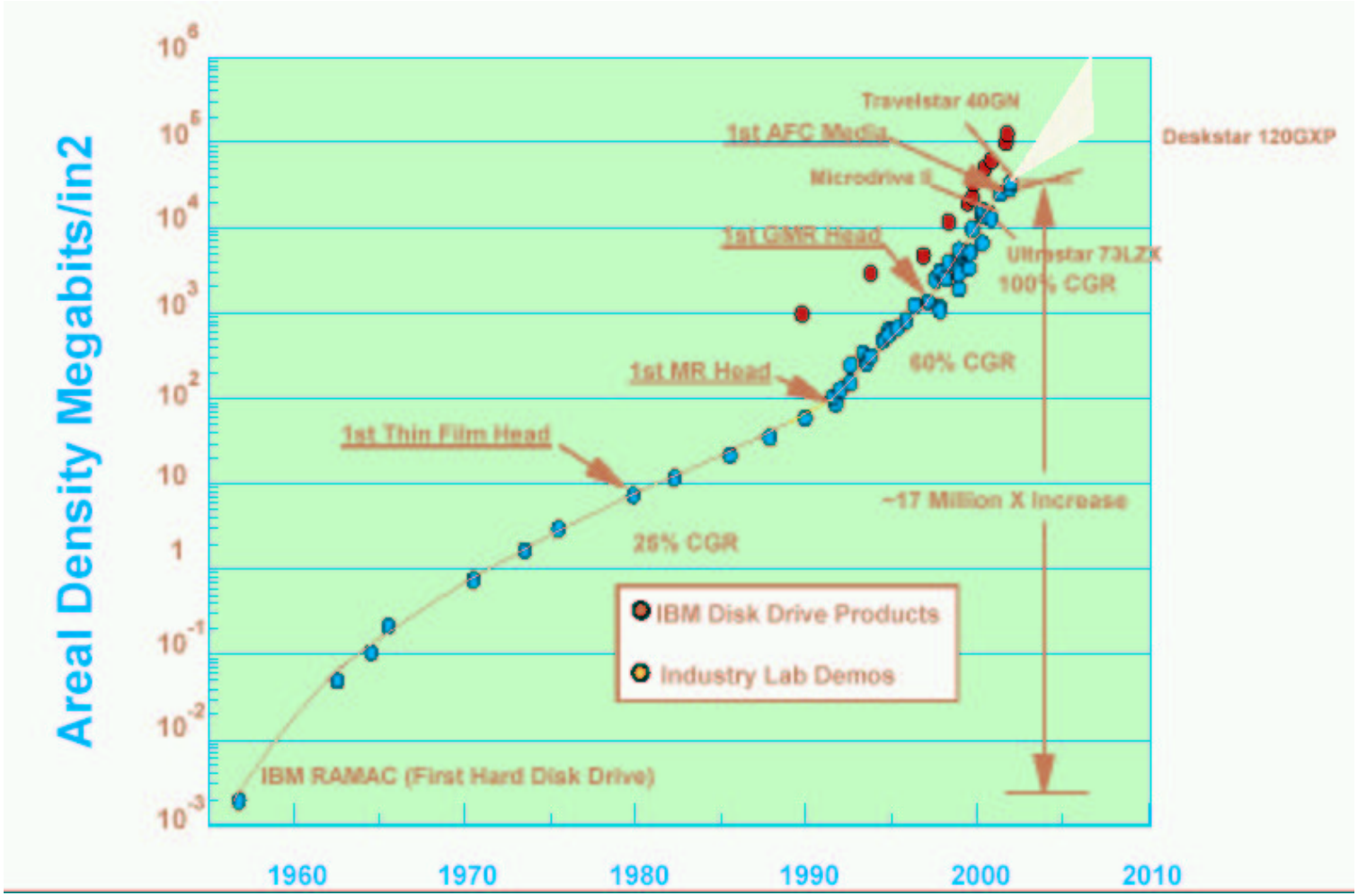


Specialisation and consolidation of disk manufacturers





Historical Progress



Disk Drive Projections

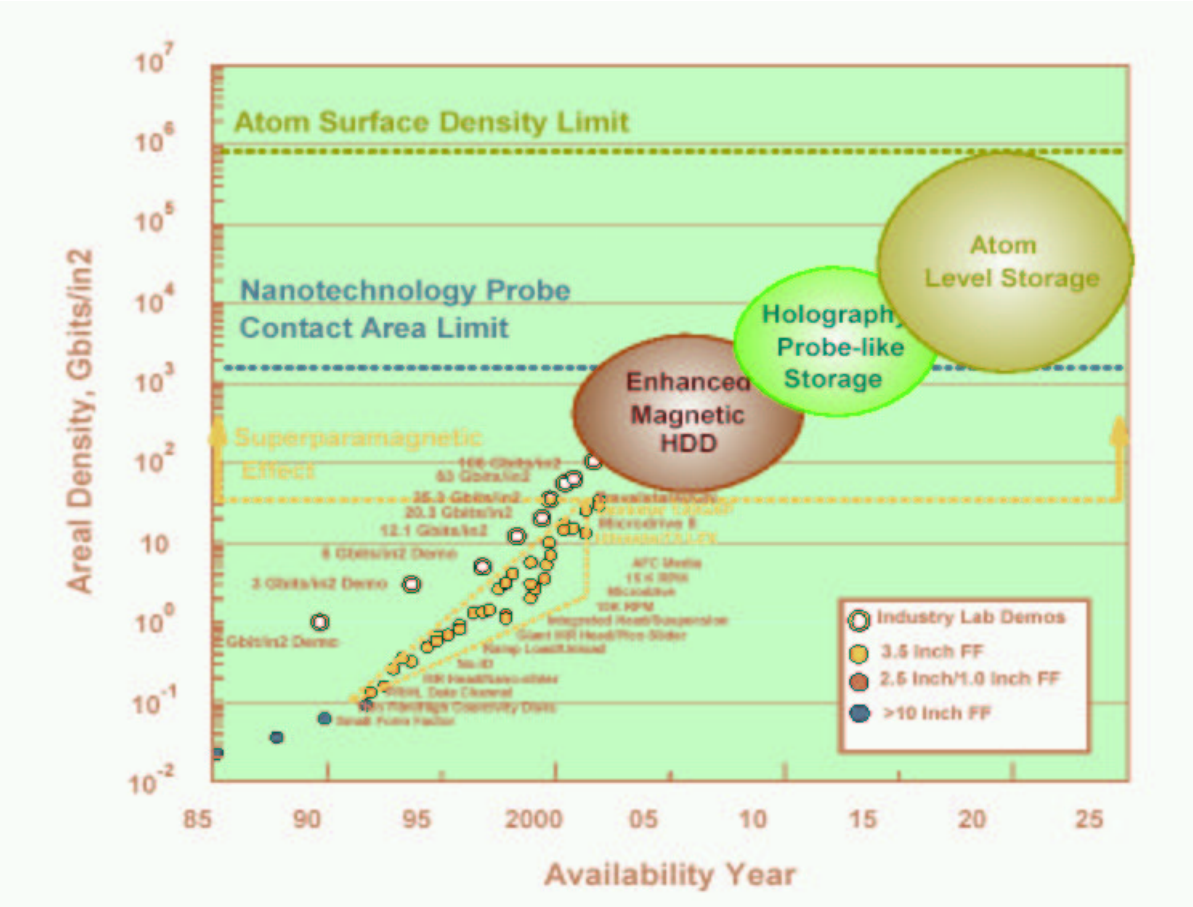
Performance Desktop

	RPM	Avg. Seek (ms)	1KB Random SIO/s (RPO)
2000 (75GB)	7200	8.5	137
2001	7200	7.8	146
2002	10K	7.0	171
2003	10K	6.3	187
2004	10K	5.6	205
2005 (1050GB)	15K	4.8	252

Mainstream Server

	RPM	Avg. Seek (ms)	1KB Random SIO/s (RPO)
2000 (36GB)	10K	4.9	226
2001	10K	4.5	244
2002	15K	4.1	283
2003	15K	3.6	317
2004	15K	3.2	345
2005 (500GB)	20K	2.8	408

Advanced Storage Roadmap



Disk Trends - 1

- Capacity is doubling every 18 months but not access times
- Super Paramagnetic Limit (estimated at 40GB/in^2) has not been reached. Seems that a platter capacity of 2-3 times todays capacity can be foreseen.
- "Perpendicular recording" aims to extend the density to 500-1000GB/in^2. Disks of 10-100 times todays capacity seem to be possible. The timing will be driven my market demand.
- Rotational speed and seek times are only improving slowly so to match disk size and transfer speed disks become smaller and faster. 2.5" with 23500 RPM are foreseen for storage systems.

Disk Trends - 2

- SCSI still being developed, now at 320MB/sec transfer speed.
- Serial ATA is expected to dominate the commodity disk connectivity market by end 2003. 150MB/sec moving to 300 MB/sec
- Fiber channel products still expensive.
- DVD solutions still 2-3x as expensive as disks. No industry experience managing large DVD libraries.
- Some new technologies starting to be talked about in the early LHC timeframe (nano fabrication) as replacement for rotating disk storage so will probably have an effect during during the LHC lifetime

Storage Management

- Very little movement in the HSM space since the last PASTA report.
  - HPSS still for large scale systems
  - A number of mid-range products (make tape look like a big disk) but limited scaling possible
- HEP still a leader in tape and data management
  - CASTOR, Enstore, JASMine
  - Will remain crucial technologies for LHC.
- Cluster file systems appearing (StorageTank - IBM)
  - Provide "unlimited" (PB) file system through SAN fabric
  - Scale to many 000's of clients (CPU servers).
  - Need to be interfaced to tape management systems (e.g. Castor)

Storage - Architecture

- Possibly the biggest challenge for LHC
  - Storage architecture design (seamless integration from CPU caches to deep archive required)
  - Data management. Currently very poor tools and facilities for managing data and storage systems.
- SAN vs. NAS debate still alive
  - SAN, scalable and high availability, but costly
  - NAS, cheaper and easier to implement
- Object storage technologies appearing
  - Intelligent storage system able to manage the objects it is storing
  - Allowing "light-weight" Filesystems

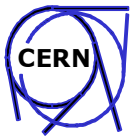
CERN

Storage - Connectivity

- FiberChannel market growing at 36%/year from now to 2006 (Gartner). This is the current technology for SAN implementation.
- iSCSI or equivalent over Gigabit Ethernet is an alternative (and cheaper) but less performant implementation of SAN gaining in popularity.
  - It is expected that GigE will become a popular transport for storage networks.
- InfiniBand (up to 30 Gbps) is a full-fledged network technology that could change the landscape of cluster architectures and has much, but varying, industry support.
  - Broad adoption could drive costs down significantly
  - FIO (Compaq, IBM, HP) and NGIO (Intel, MS, Sun) merged to IB
  - Expect bridges between IB and legacy Ethernet and FC nets
  - Uses IPv6
  - Supports RDMA and multicast
- May expect NAS/SAN models to converge

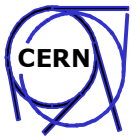
Disk Pricing

- In 2002, the cost of storage in the ATA disk servers used with CASTOR at CERN was 24 CHF/GB of effective disk capacity. Servers are configured in mirrored mode giving 500 GB capacity and the cost includes the server, power supply and network infrastructure. **The raw disk capacity represents about 20% of the cost with a 120 GB drive costing 240 CHF.**
- It is reasonable to predict that in 2006/2007, 480 GB drives will be available at a similar price to the 120 GB drives of today.
- We believe that for the foreseeable future, IDE/ATA will remain the technology that minimises cost per GB at reasonable speed.
- SAN architectures using either SATA or Fiber Channel drives should be considered for applications that demand higher levels of performance and availability. For configurations of 100 TB and higher, the cost overhead compared to ATA servers is estimated to be about 40%
- Using the figures above, 1 PB of raw disks could be purchased for about 0.5M CHF in 2007. Assuming this represents 20% of the total server cost, the price moves to 2.5M CHF With mirroring, the effective capacity is 500 TB; with RAID5 it is about 800 TB



Some Overall Conclusions

- Tape and Network trends match or exceed our initial needs.
  - Need to continue to leverage economies of scale to drive down long term costs.
- CPU trends need to be carefully interpreted
  - The need for new performance measures are indicated.
  - Change in the desktop market might effect the server strategy.
  - Cost of manageability is an issue.
- Disk trends continue to make a large (multi PB) disk cache technically feasible, but ….
  - The true cost of such an object remains unclear, given the issues of reliability, manageability and the disk fabric chosen (NAS/SAN, iSCSI/FC etc efc)
  - File system access for a large disk cache (RFIO, StorageTank) is also unclear.
- More architectural work is needed in the next 2 years for the processing and handling of LHC data.
  - NAS/SAN models are converging, access patterns are unclear, many options for system interconnects.



… Sounds like we are in pretty good shape …..

… but let’s be careful …

PASTA has addressed issues exclusively on the Fabric level

- It is likely that we will get the required technology (Processors, Memory, Secondary and Tertiary Storage Devices, Networking, Basic Storage Management)
- Complete user level solutions allowing truly distributed Computing on a Global Scale are complex to design and build.
  - Will the Grid Projects meet our expectations (in time) ?