Advanced Computing Technology Overview

Richard P. Mount

Director: Scientific Computing and Computing Services
Stanford Linear Accelerator Center

May 25, 2005

Advanced Computing Technology
My Viewpoint

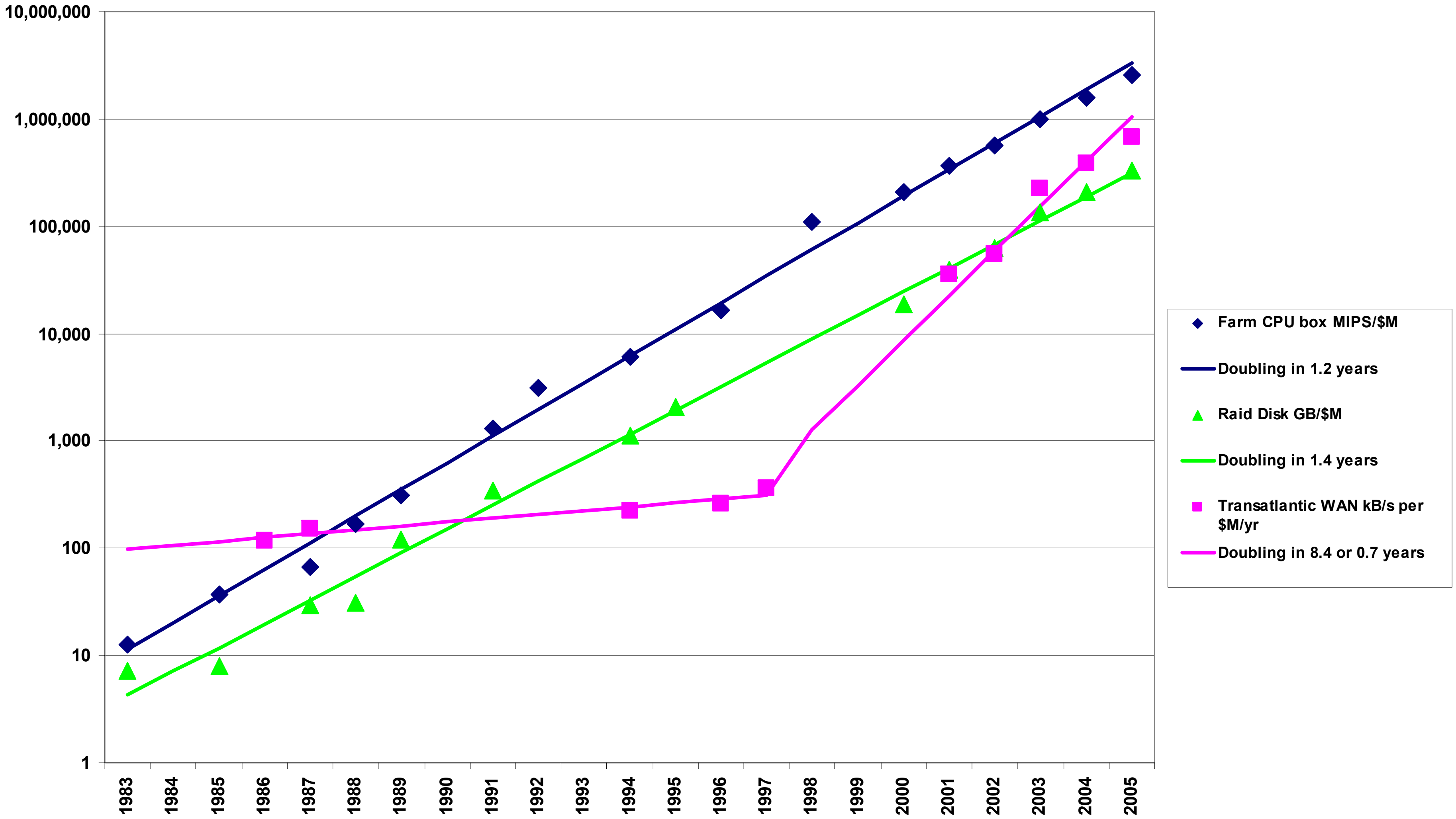
- Decades of computing for experimental HEP;
- Decades of data-intensive computing;
- Belief that the future will be even more data-intensive for HEP;
- Belief that the many other sciences are also facing a data-intensive future.

My History

- Circa 1980
  - The EMC experiment
  - World's largest collaboration (99 physicists)
  - 10,000 tapes/year
  - No advance planning for computing resources
  - I had to invent a data-handling system to be able to do physics
- 1982 – 1997
  - The L3 experiment at LEP
  - Responsible for L3 computing at CERN
- 1997 – now
  - BaBar at SLAC
  - Future HEP, Particle-astro and X-ray science at SLAC/Stanford

CPU, Disk and Network History

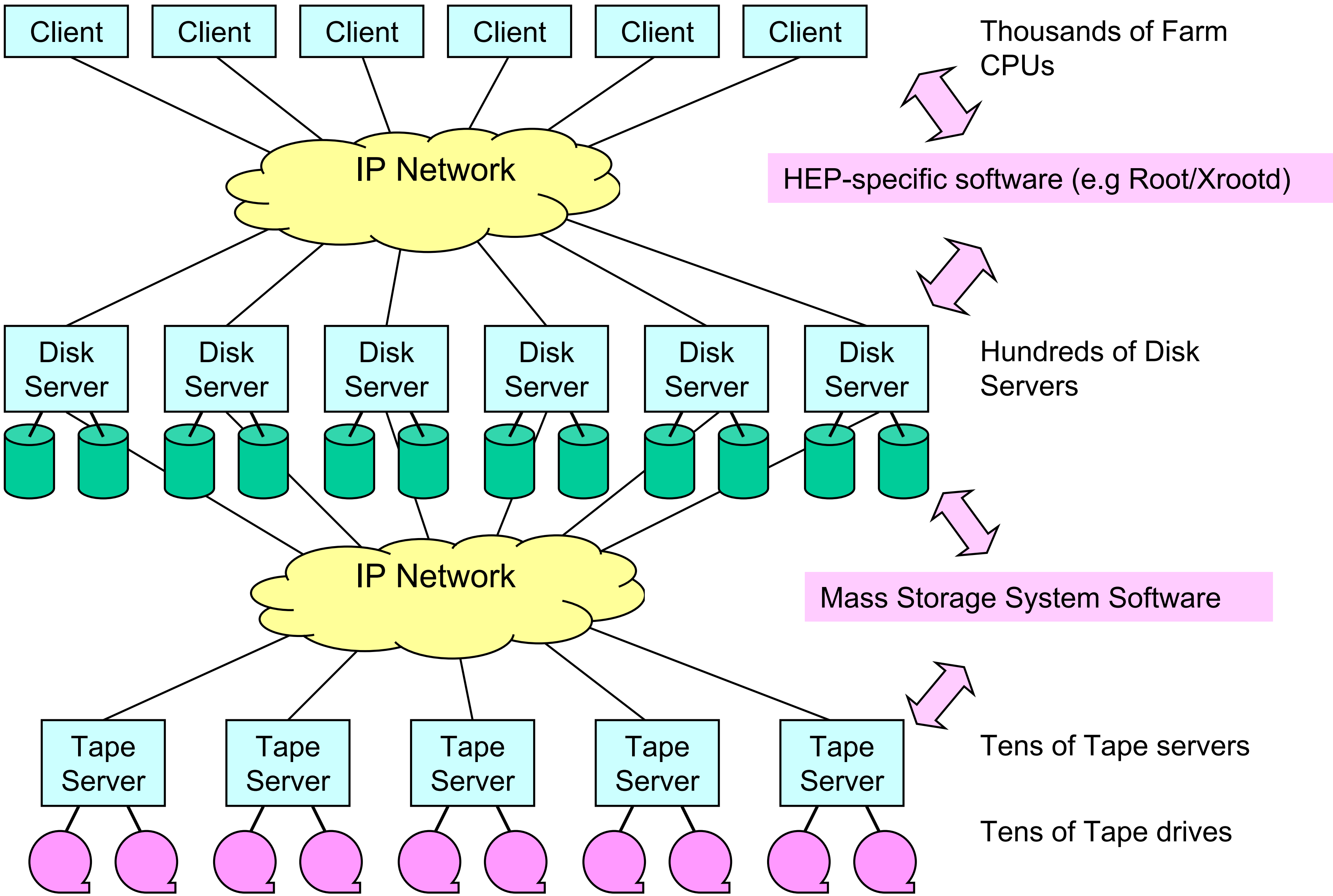
i.e. what I (and Harvey) have bought



What about the future?

- CPU
  - The clock-speed ramp up has run out of steam
  - Intel/AMD response: multicore chips
    - Fairly easy to use for HEP data processing
  - Intel predicts 25 Tflops/chip in 2015 (100 cores)
    - Close to the “doubling every 1.2 year” extrapolation (costs are very dependent on memory)
- Disk
  - “increasing requirements for disk drive improvements provides a unending challenge to extend GMR technology to its limits, and then to look beyond” (Hitachi GST)
  - It seems to be working – no end to ***capacity*** growth is in sight yet.

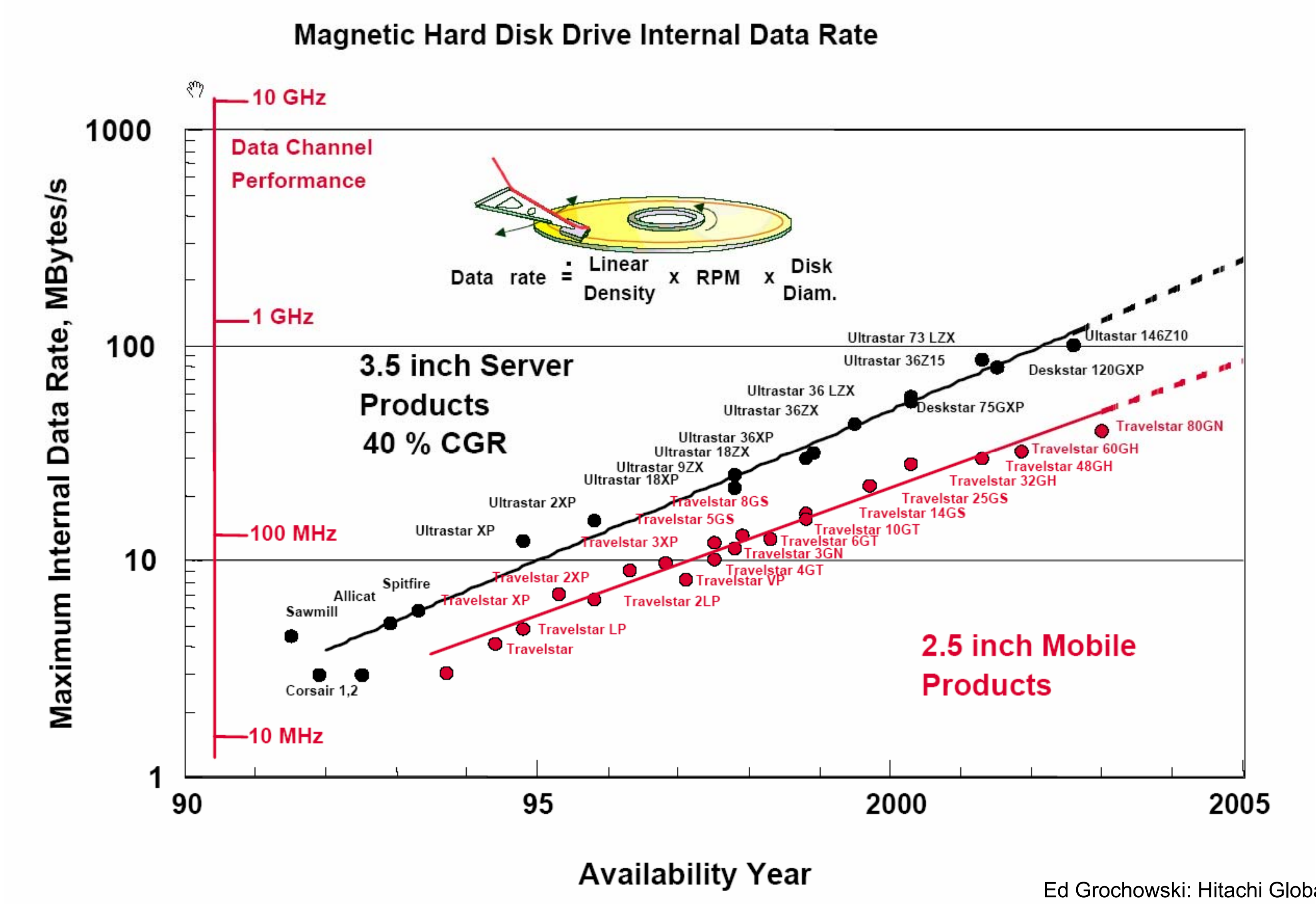
Generic HEP Computing Fabric



Some Issues

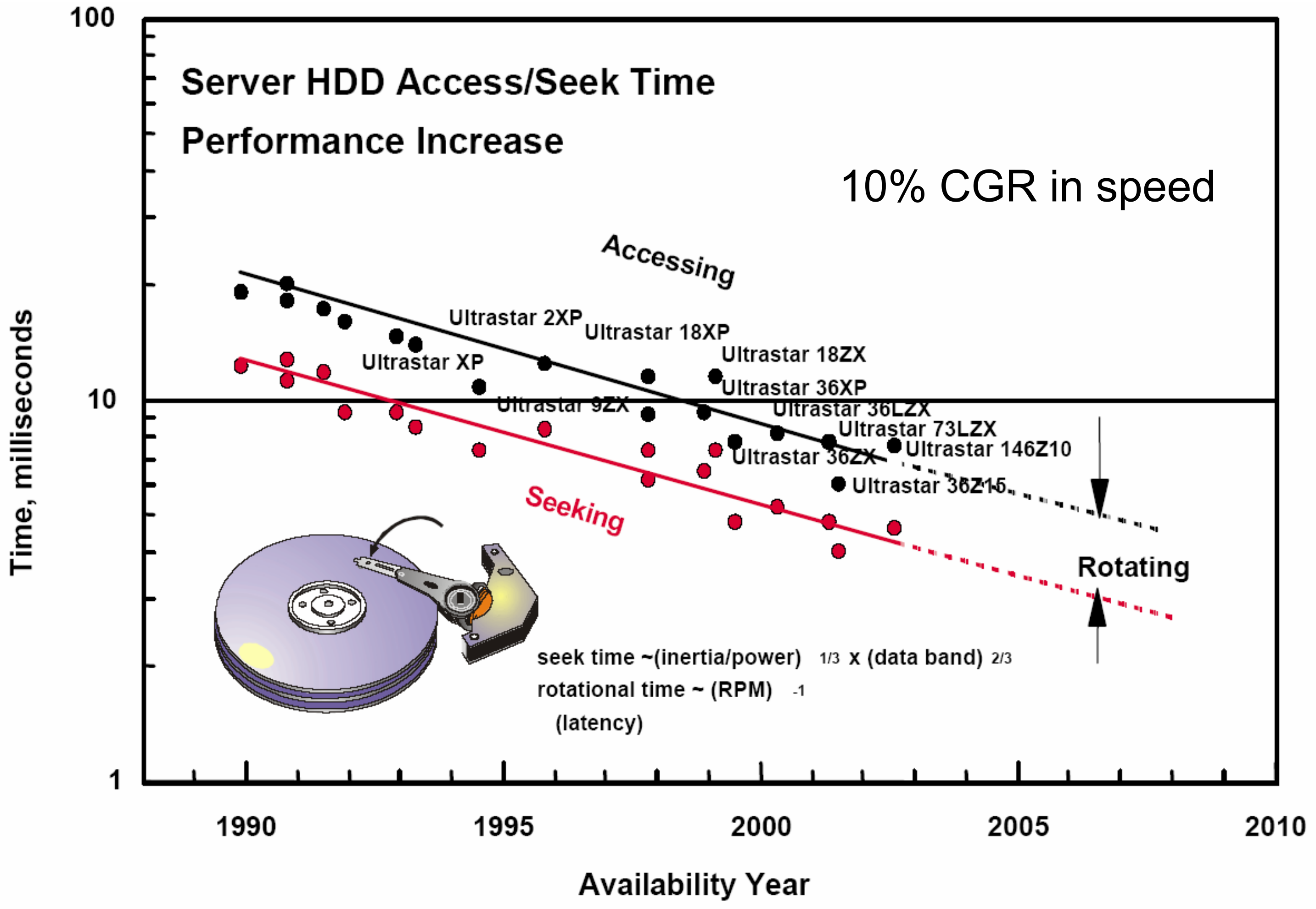
- While CPU power per $ has been doubling every 1.2 years, Watts per $ have been increasing too.
  - Infrastructure (power, cooling, space) now costs as much per year as the computers
- Boxes per $ are also increasing
  - 10,000 – 100,000 box systems are in sight
  - Scalability is vital
  - Fault tolerance is a requirement
- The raised-floor switched network (= system backplane) is a potential bottleneck
  - But if you have a few times $10M then a Cisco CSR-1 can provide about 10,000 non blocking 10Gbit Ethernet ports on a single switch fabric.

Disks: It's not just about capacity (1)



Ed Grochowski: Hitachi Global Storage Technologies

Disks: It's not just about capacity (2)



Ed Grochowski: Hitachi Global Storage Technologies

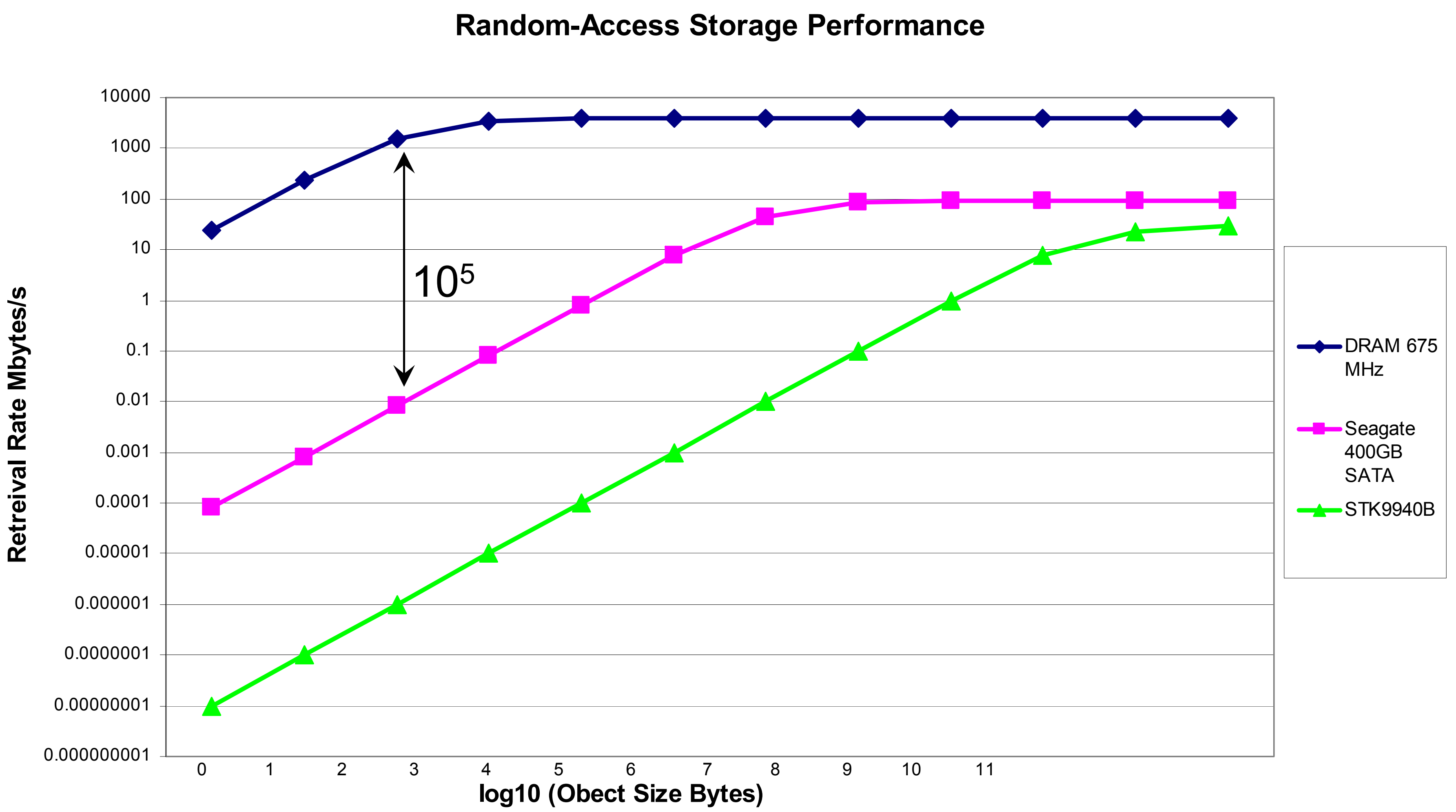
Disks: Gloom and Doom

- Giora Tarnopolski (TarnoTek)
  - "I do not believe that there will be much shorter access times in the future"
- Ed Grochowski (Hitachi GST)
  - "While rotation rates beyond 15K are possible in the future, these will likely occur at longer product time intervals"
- BaBar reality:
  - Micro DST events are replicated about tenfold on disk
  - Millions of $$$ per year, and many months of delay, are spent on data reorganization to allow efficient access by thousands of concurrent jobs

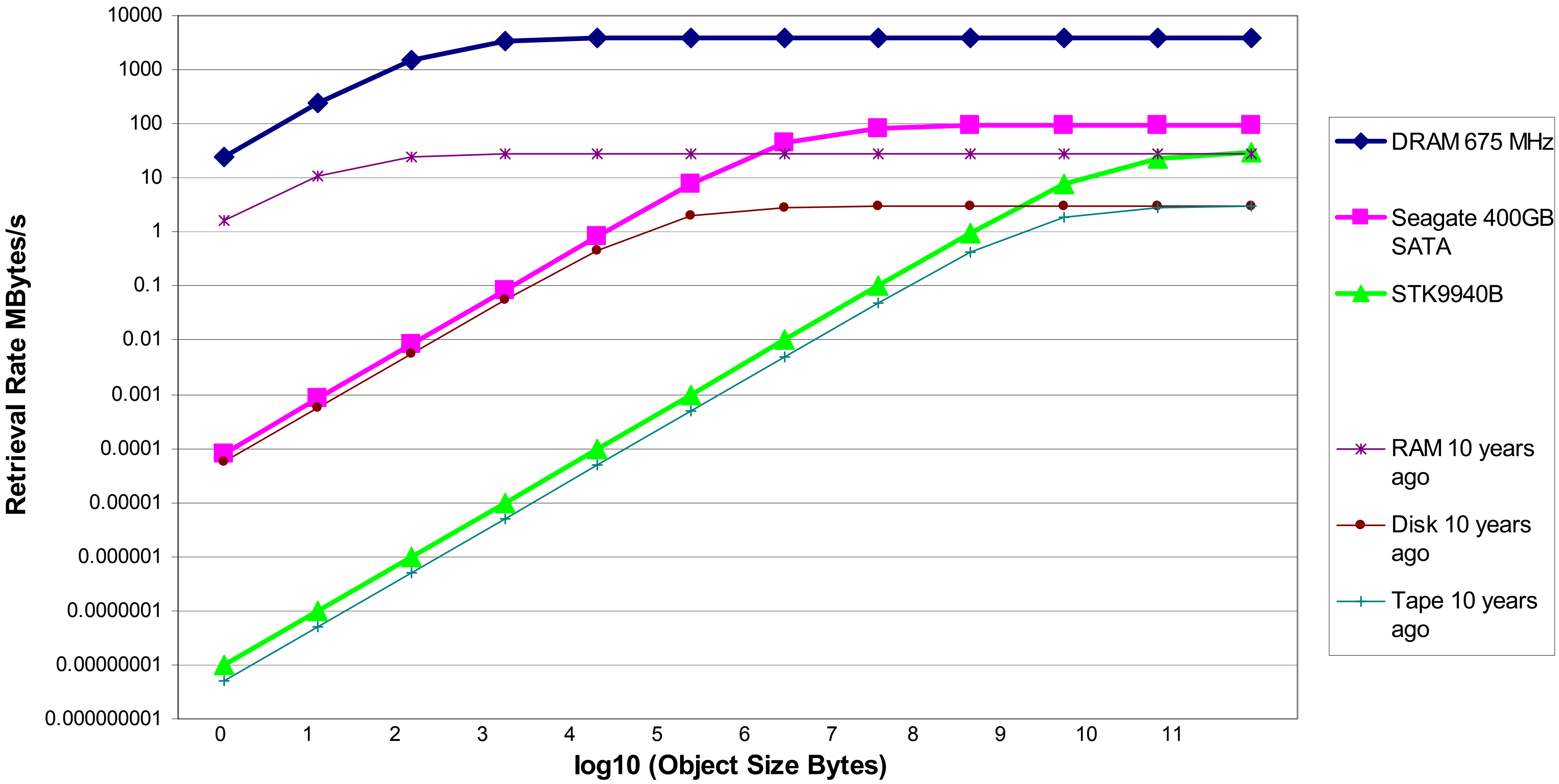
Technology Issues in Data Access

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

Latency and Speed – Random Access



Latency and Speed – Random Access



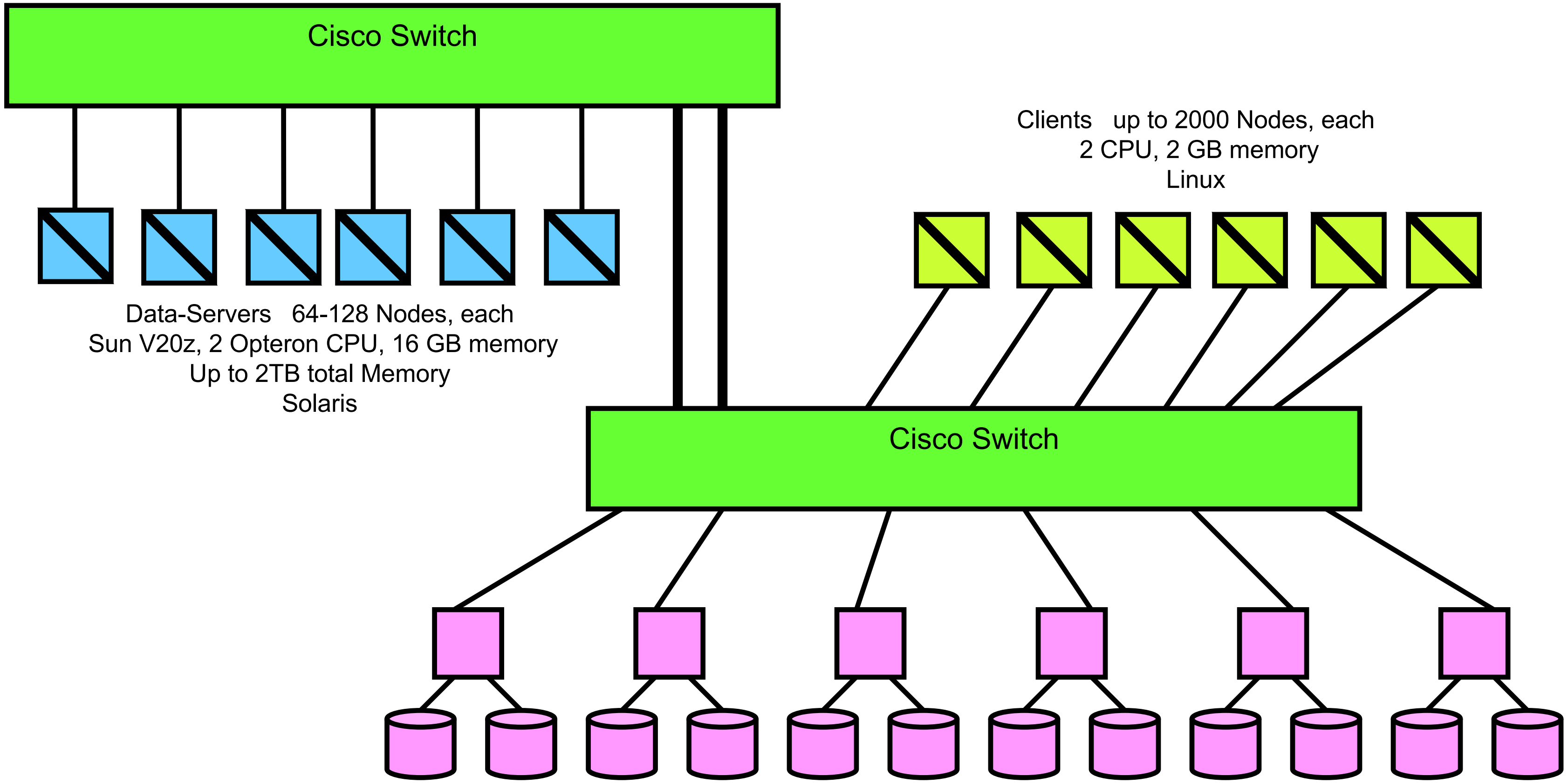
Death to Disks

- 10^5 latency gap with respect to memory will be intolerable, eventually even in consumer applications;
- Storage-class memory is in development (see Jai Menon's talk at CHEP 2004);
- In the meantime we can use DRAM or even Flash memory in latency-critical or throughput-critical applications;
- For example, May 23, 2005 news item:
  - "Samsung develops flash-based 'disk' for PCs"
    - "It uses memory chips instead of a mechanical recording system"
    - http://www.computerworld.com/hardwaretopics/storage/story/0,10801,101946,00.html?source=NLT_AM&nid=101946
- Market forces are not yet aligned with scientific needs for massive, random access storage-class memory.

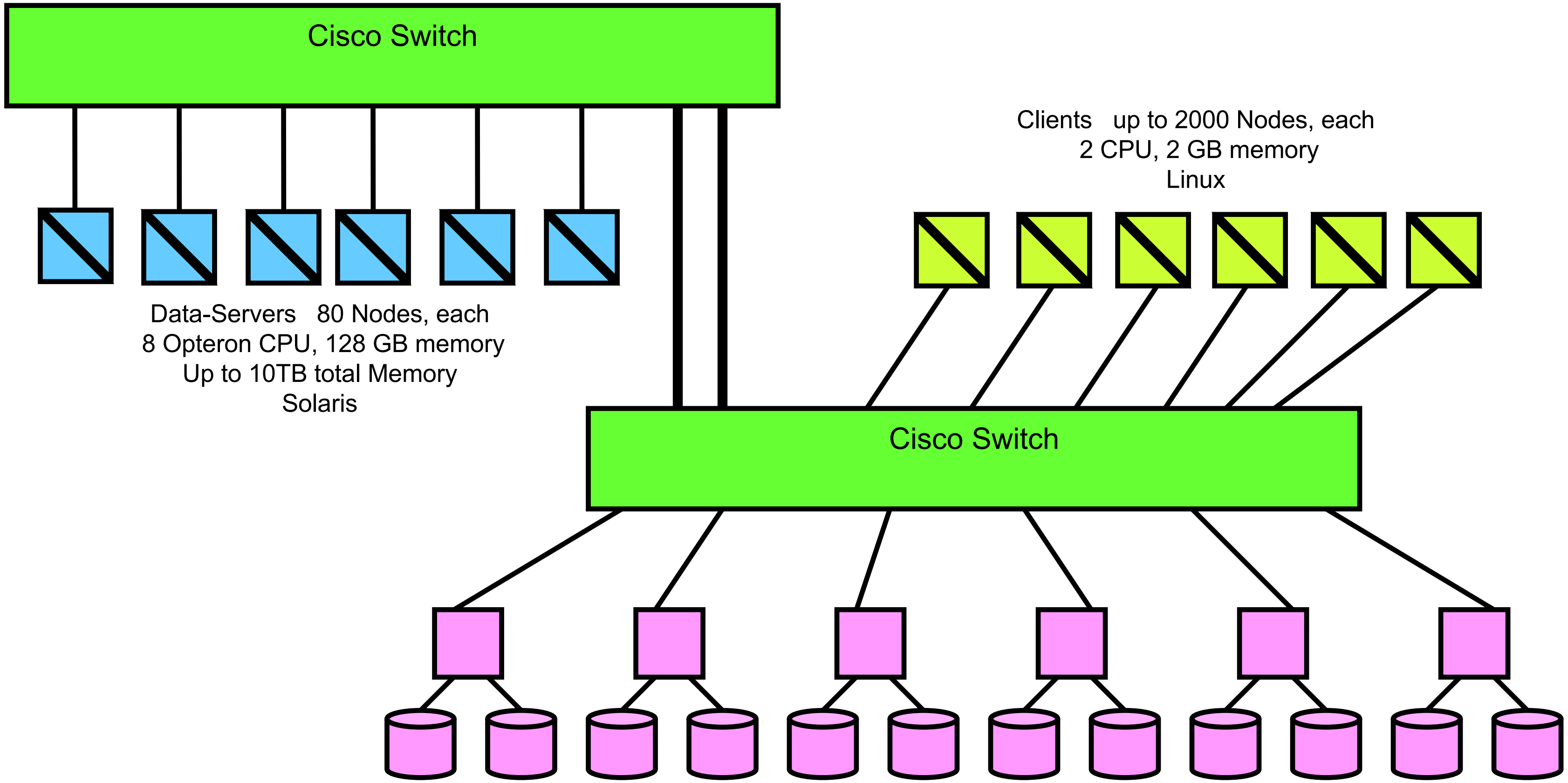
Storage-Class Memory Architecture: SLAC Strategy (PetaCache)

- There is significant commercial interest in architectures including massive 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 random-access 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. Operational Facility (years 3,4,5)
     - New architecture
     - Strong industrial collaboration
     - Wide applicability

Development Machine Deployment – Currently Funded



Development Machine Deployment – Possible Next Step



Scalable Object-Serving Software Example

- Xrootd (Andy Hanushevsky/SLAC)
  - Optimized for read-only access
  - Contributes ~29 microseconds to server latency
  - Make 1000s 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

Summary

- Moore's Law (at least the generalized version) is alive and well for CPU throughput and disk capacity;
- Moore's Law seems dead for single-threaded CPU power;
- Moore's Law never applied to random-access to data;
- At constant cost, computing is getting hotter!
- Disks are now playing the same role in HEP that tapes were in 1990 (i.e. they are not random-access devices);
- Prepare for a random-access future;
- Prepare for a 10,000 to 100,000 box future;
- Scalability and fault tolerance are the challenges we must address.