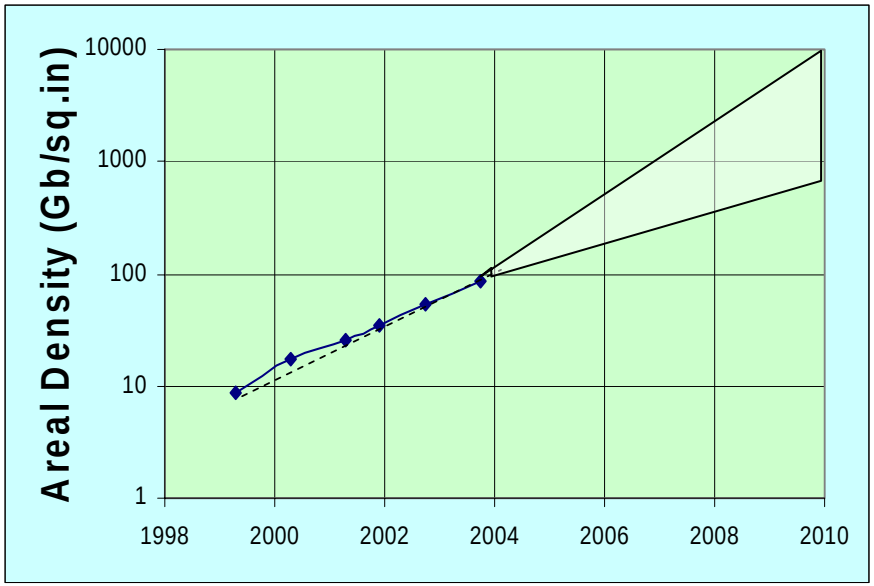
IBM TotalStorage

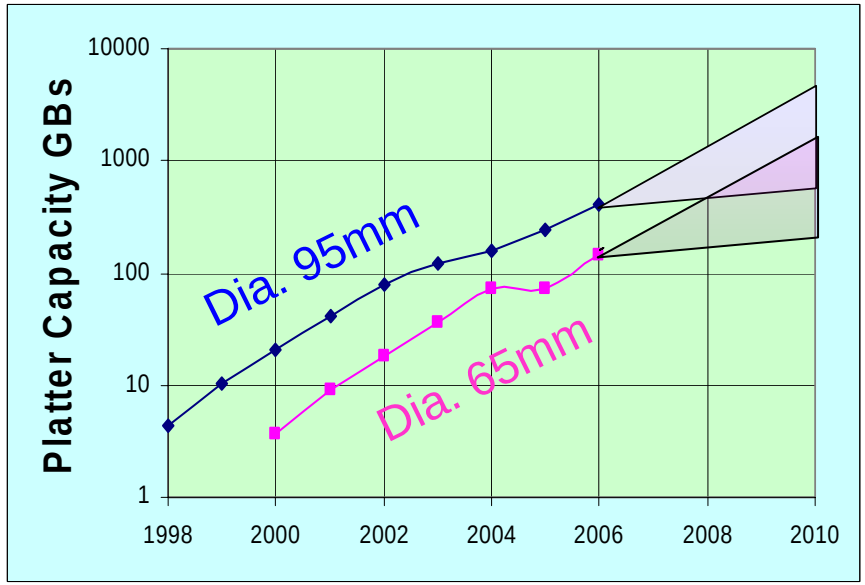
Storage Futures and Research

Jai Menon
IBM Fellow

September 2004

Disk Drive Technology Trends

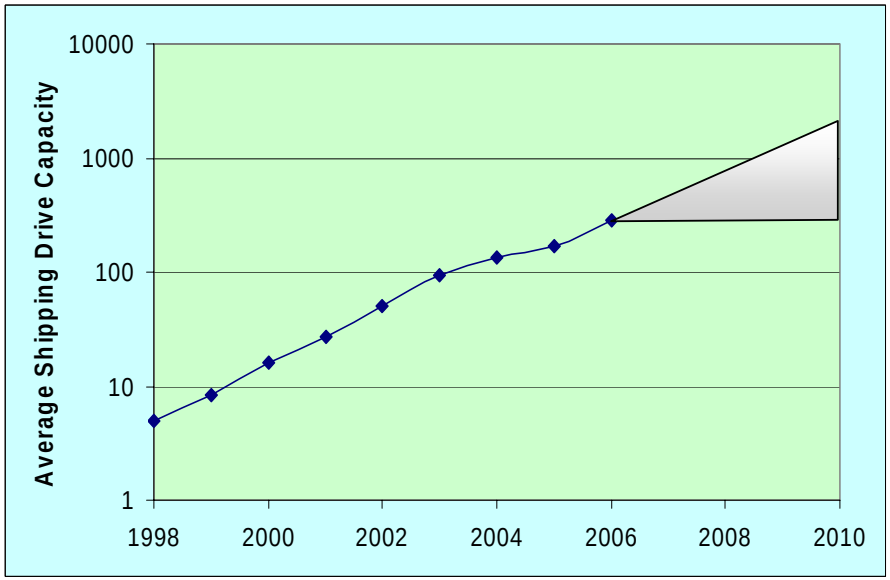


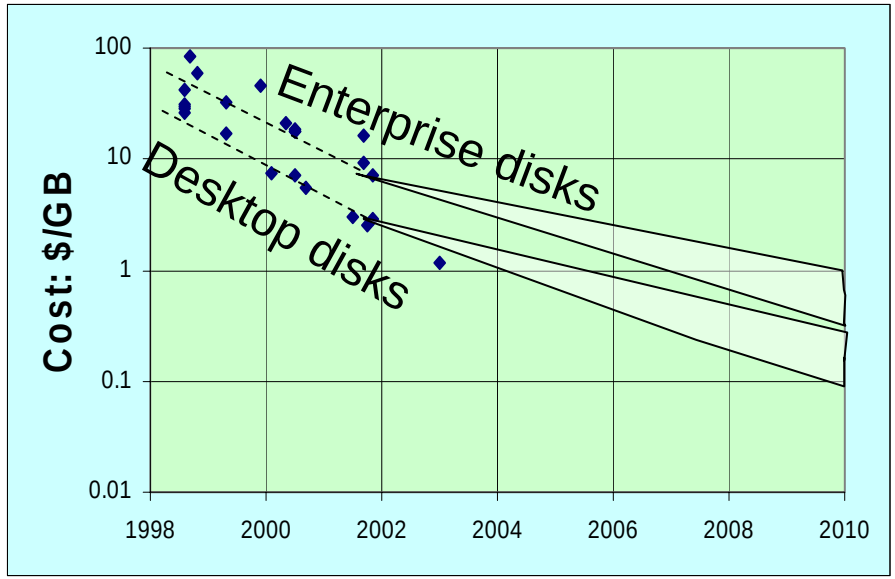


- Recent past ~100% CAGR
- Industry view of future mixed
- Forecasts are now 40% CAGR

- Most vendors moving to 65mm diameter drives
- Drive FF likely to also be reduced to 2 ½”
- 3 ½” FF drives may be gone by 2010

Disk Drive Cost and Capacity Trends

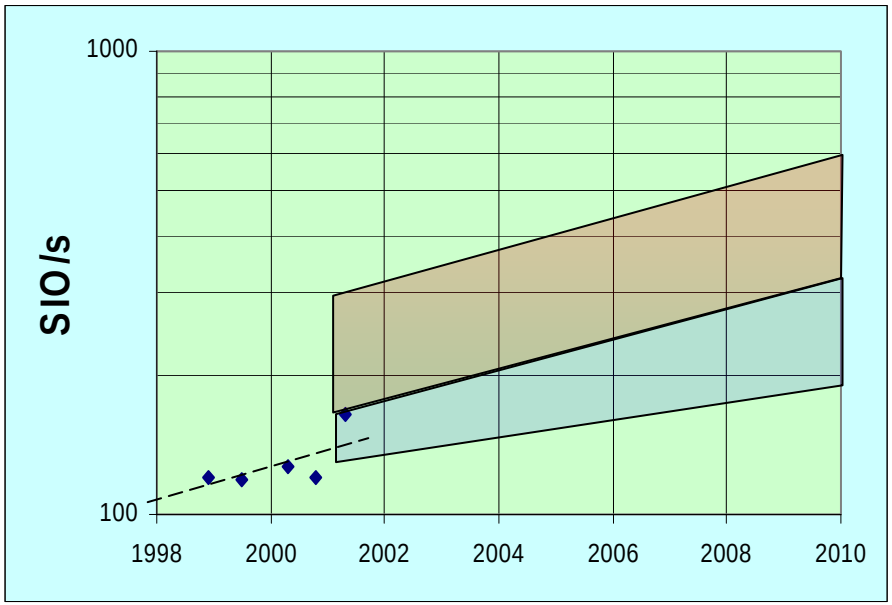


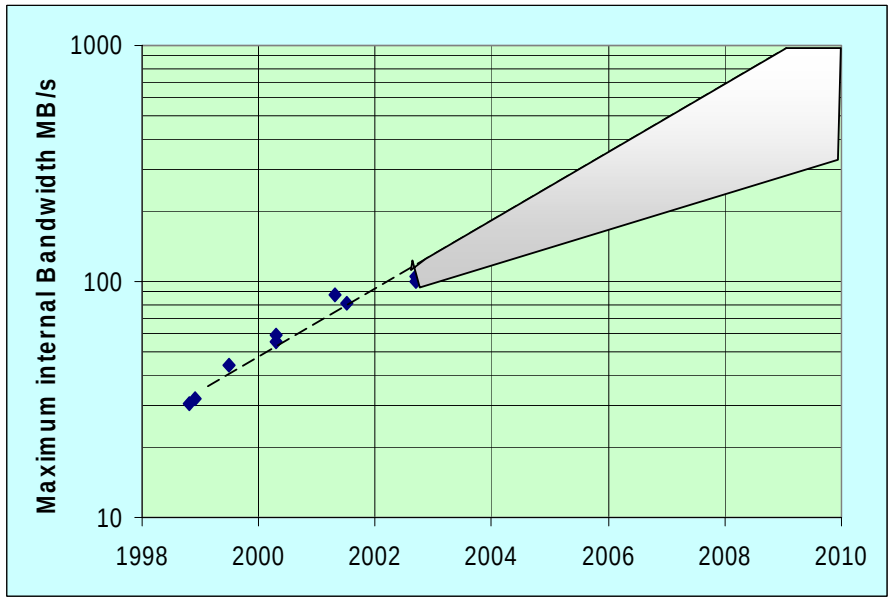


- Technology has been outpacing needs
- Single platter drive most likely
- 2010 sweet spot: ~1TB 2 ½" disk

- 2010 2.5" desktop disk will likely cost ~ $100
- Enterprise ~3X Desktop drives in $/GB

Disk Drive Performance Trends

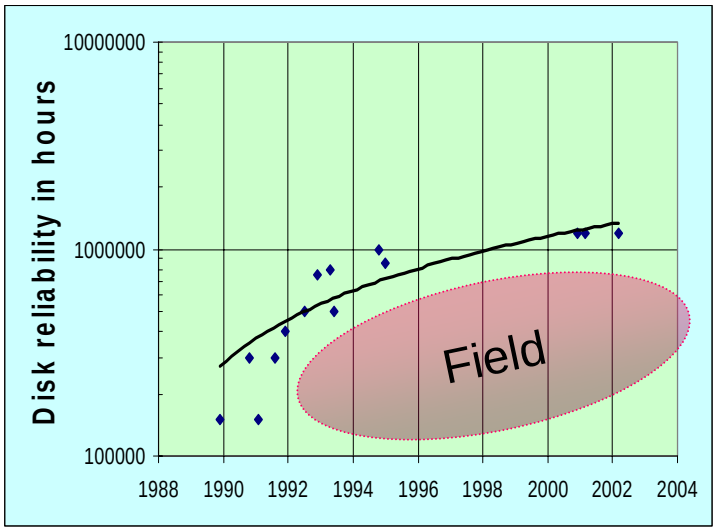


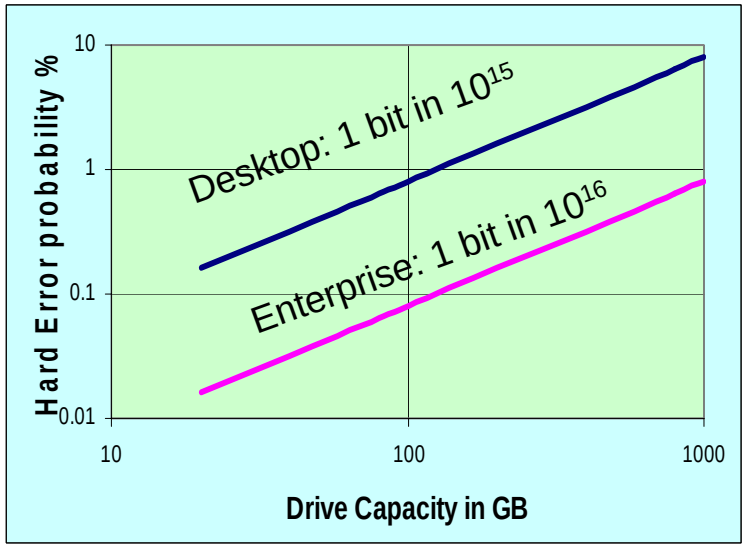


- Historical trend: ~8% /yr, will continue
- Command queuing can help significantly

- Historical trend: ~40%/yr
- Will track linear density
- Assume channels can keep up

Disk Drive Reliability Trends





- Actual information from field not as good as vendor specs

- Drive hard error rate 1 in 10^{16} for enterprise
- Issue as drive capacity increases
- .8% probability of hard error (HE) reading 1 TB

Large customers will need more than R-5/Mirroring

Failure rates in a system with 1 PB

	RAID5	Mirror	RAID-6	New
Drive Loss/Y	**46**	**80**	**46**	**80**
Strip Loss/Y	**6**	**2**	**2e-3**	**2e-9**
Array Loss/Y	**2e-3**	**3e-4**	**1e-6**	**7e-12**

When a disk fails, there is a 1% to 8% chance of being unable to read all sectors from other disks in the array -- this causes a strip loss

	# of errors	Ios/write	Eff. (16)
RAID 5	1	4	94%
Mirror	1	2	50%
RAID 6	2	6	88%
New	2	4	50%
New	2	5	82%
3xMirror	2	3	33%
RAID 51	3	6	44%
New RAID	3	6	50%

Disks in 2010

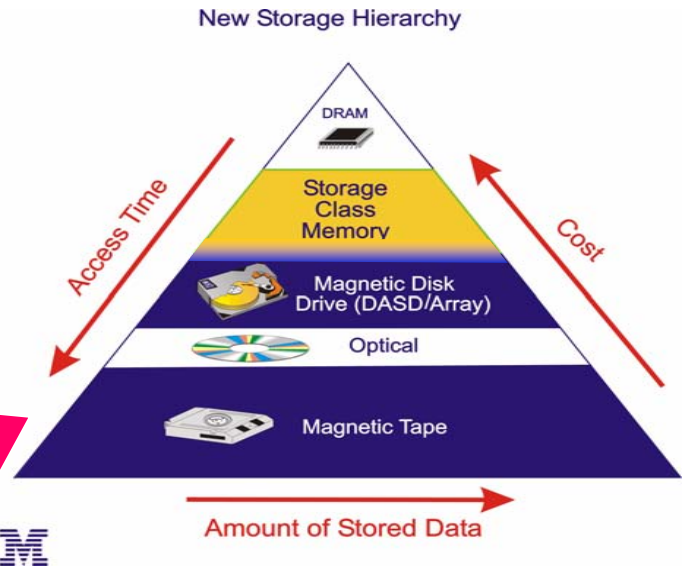
- **Drive Characteristics**
  - 1-10TB capacity
  - 3 ms Access
  - 700MB/s streaming data rate
  - 70MB/s random data rate for 256 KB records
  - Random I/O rate 500 IOs/sec (with queueing)
  - MTBF 1,200,000 hours

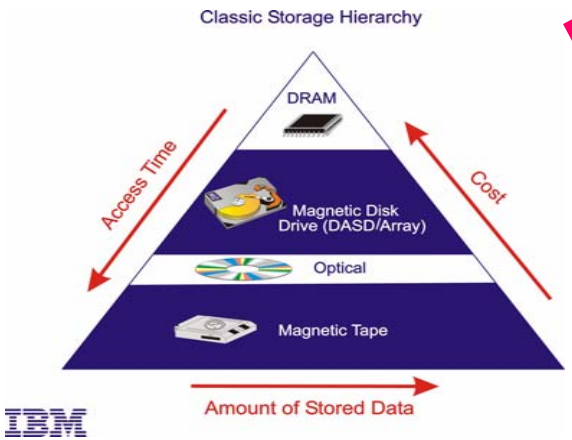
System Consequences

Number of disks		10,000		35,000	
Raw Capacity	Cabinets (2 ½")	10PB	10	35PB	35
Streaming Bandwidth		7 TB/s		24 TB/s	
Random Bandwidth		.7 TB/s		2.4TB/s	
Failures		~1.5 /wk		~5/wk	

SCM – Storage Class Memory

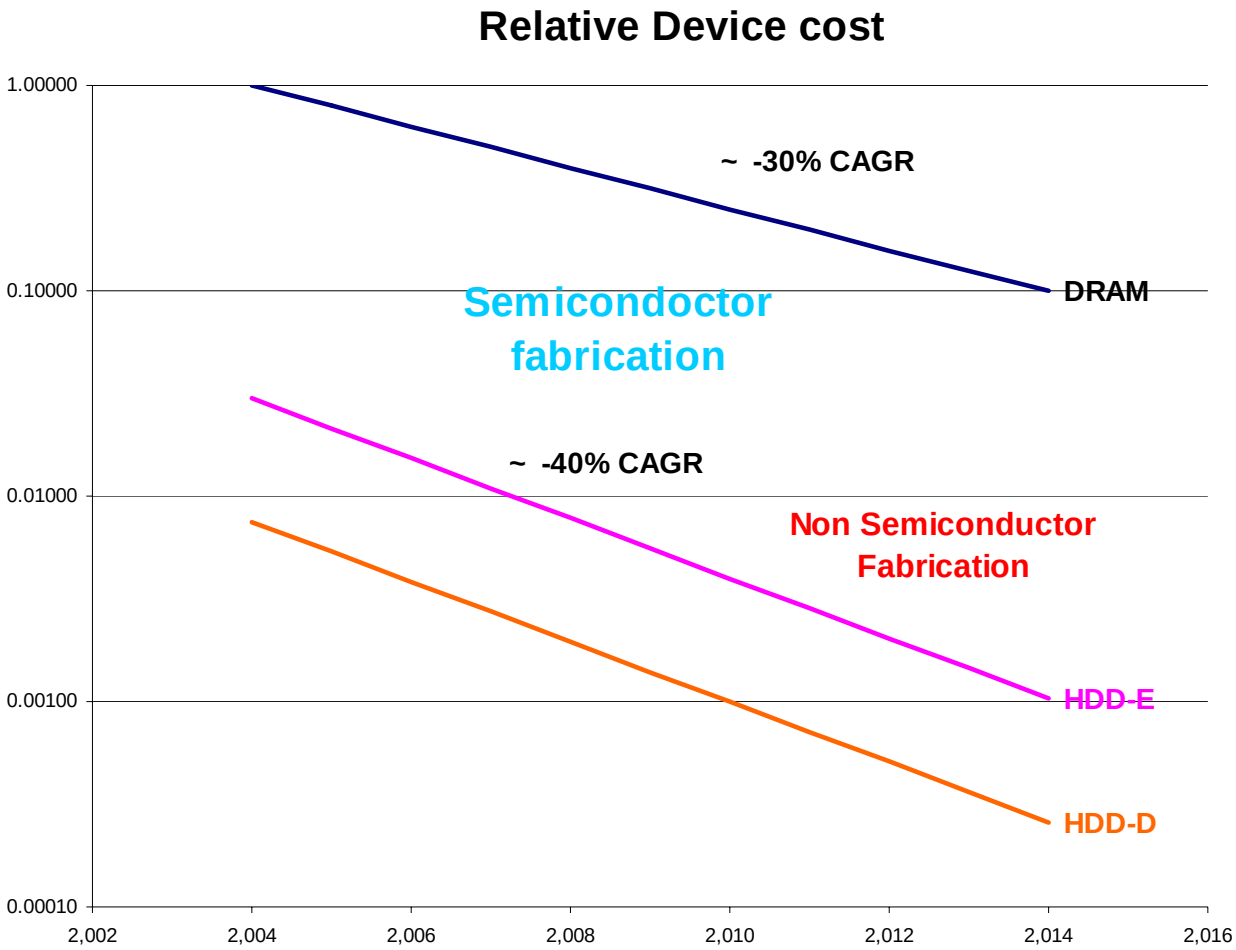
- New nonvolatile memory technologies
  - Starts as Flash replacement
  - No erase cycle needed
  - Performance ~DRAM
  - Write Endurance ~10^{12} ops
  - Semiconductor Reliability



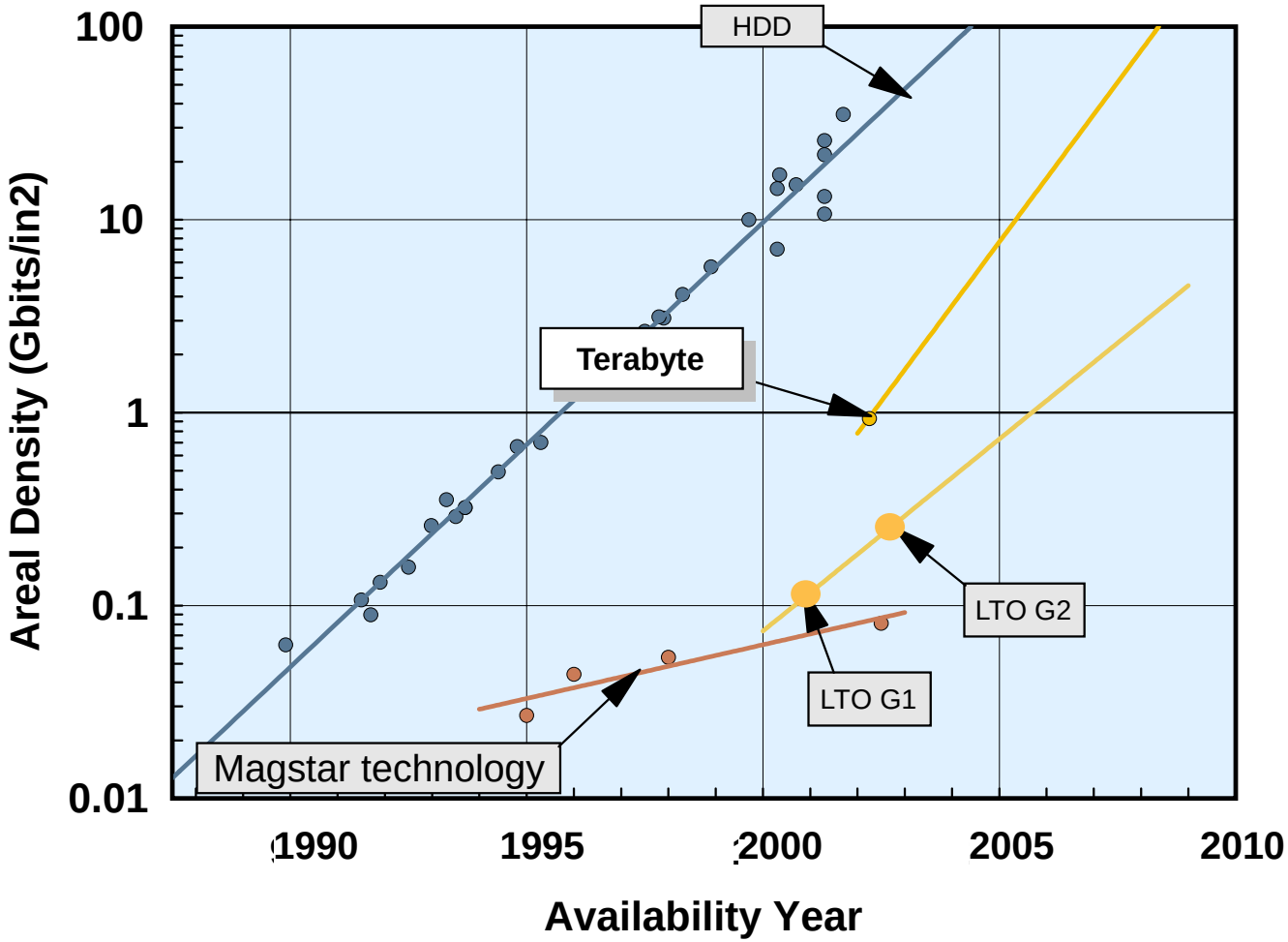


- New layer in memory device hierarchy
- Eventually competitive with Disk in Cost
- NV cache in storage controllers
- HSM
- Eventual replacement for Enterprise DASD ?

SCM = 1/10 of DRAM in 2010

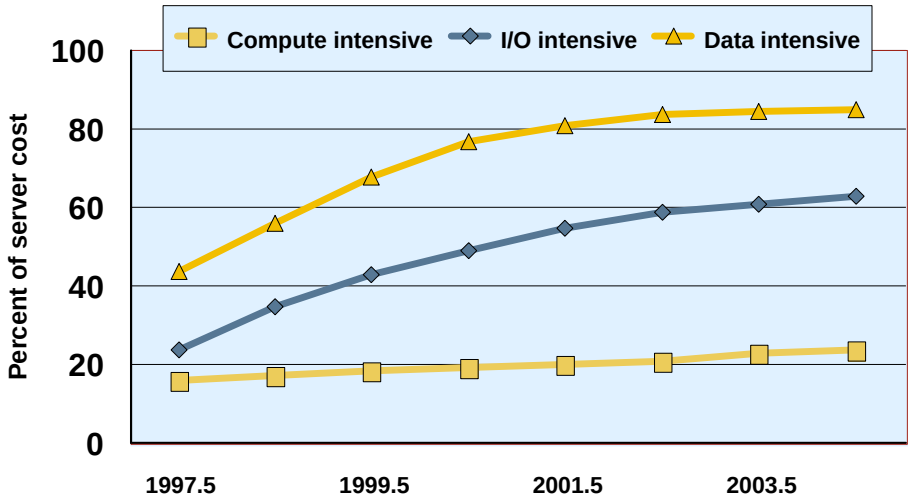


Tape Roadmap



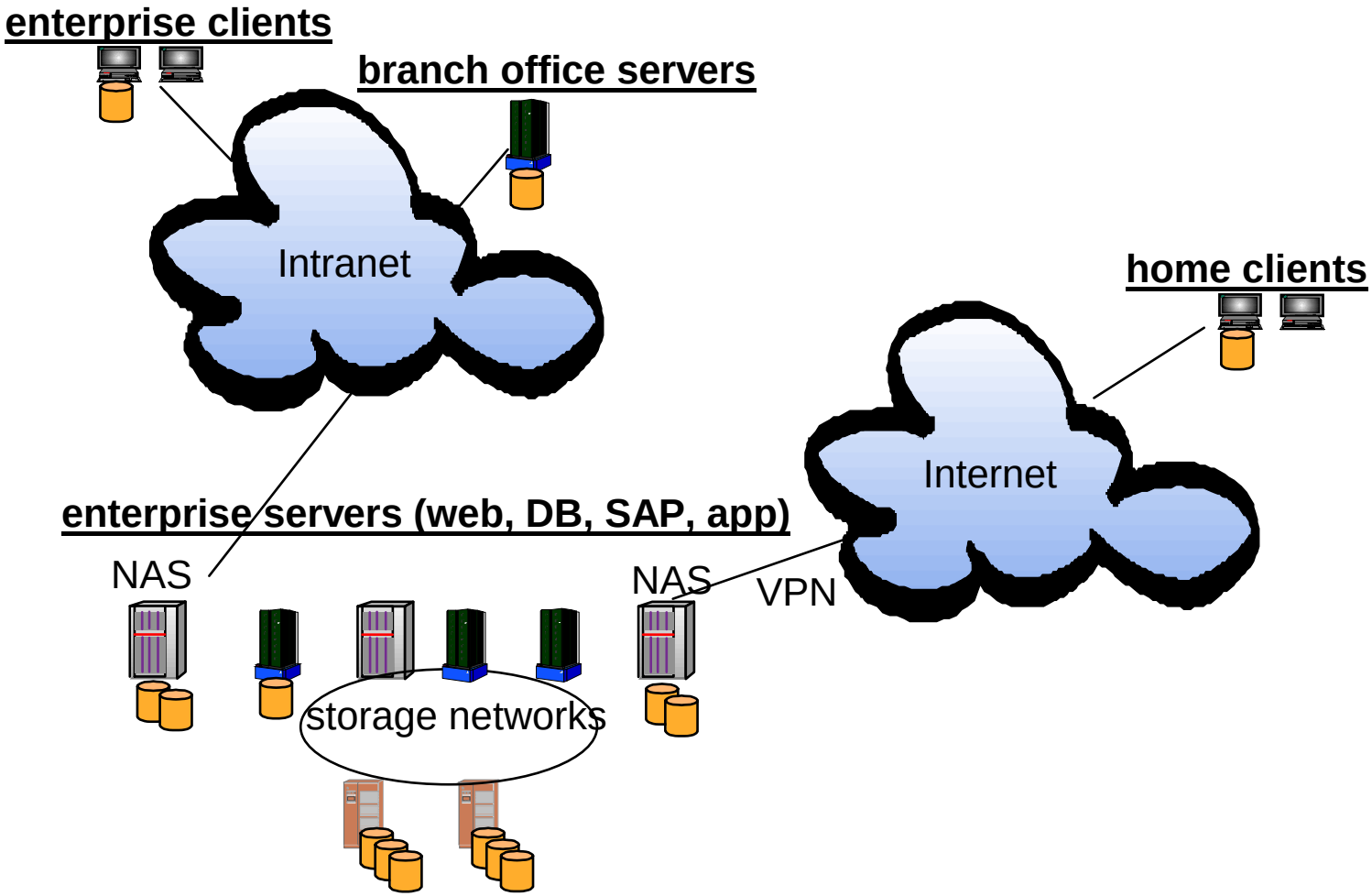
Storage Management is a big challenge

- **Storage cost increasing as fraction of total server HW (>50%)**



- **Storage management is a major concern of customers**
  - Management personnel cost conservatively 2-3x purchase
  - Not getting simpler: More entities, more options, more tools, more acronyms
    - configuring, identifying failed components, understanding scope of impact
  - Requirements are increasing (e.g., 24x7 operation)

Today – diversity of storage infrastructures and management



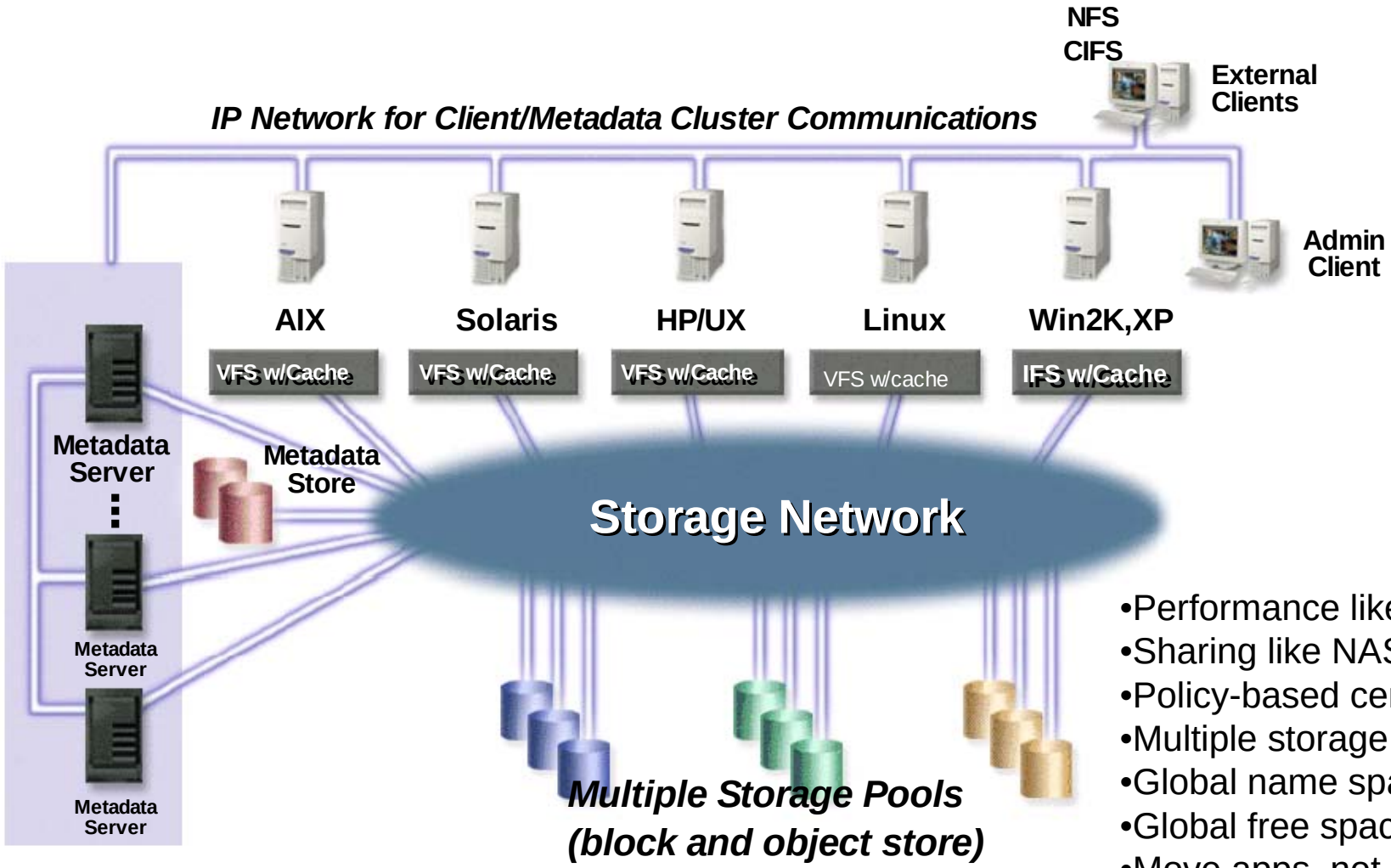
Approach to deal with storage management

- **Single storage infrastructure and common management for the enterprise**

- **Minimize complexity of storage hardware**
  - minimize number of components
  - eliminate cables
  - reduce environmentals
  - fail-in-place

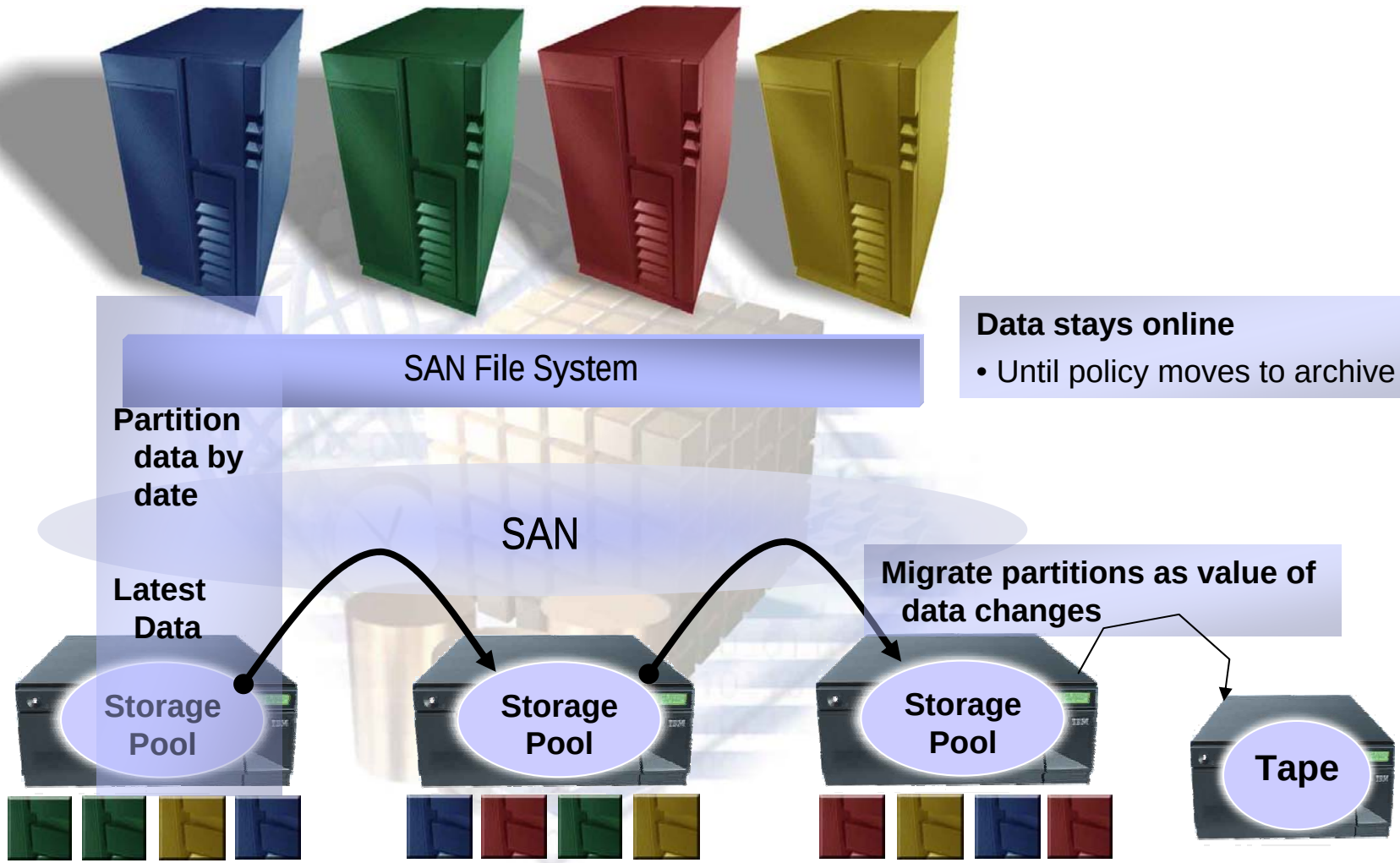
- **Move from 1 TB/storage admin to 1 PB/storage admin**

Common File Systems and Management



- Performance like SAN
- Sharing like NAS
- Policy-based central mgmt
- Multiple storage pools
- Global name space
- Global free space
- Move apps, not data

Integrated Life Cycle management of Data

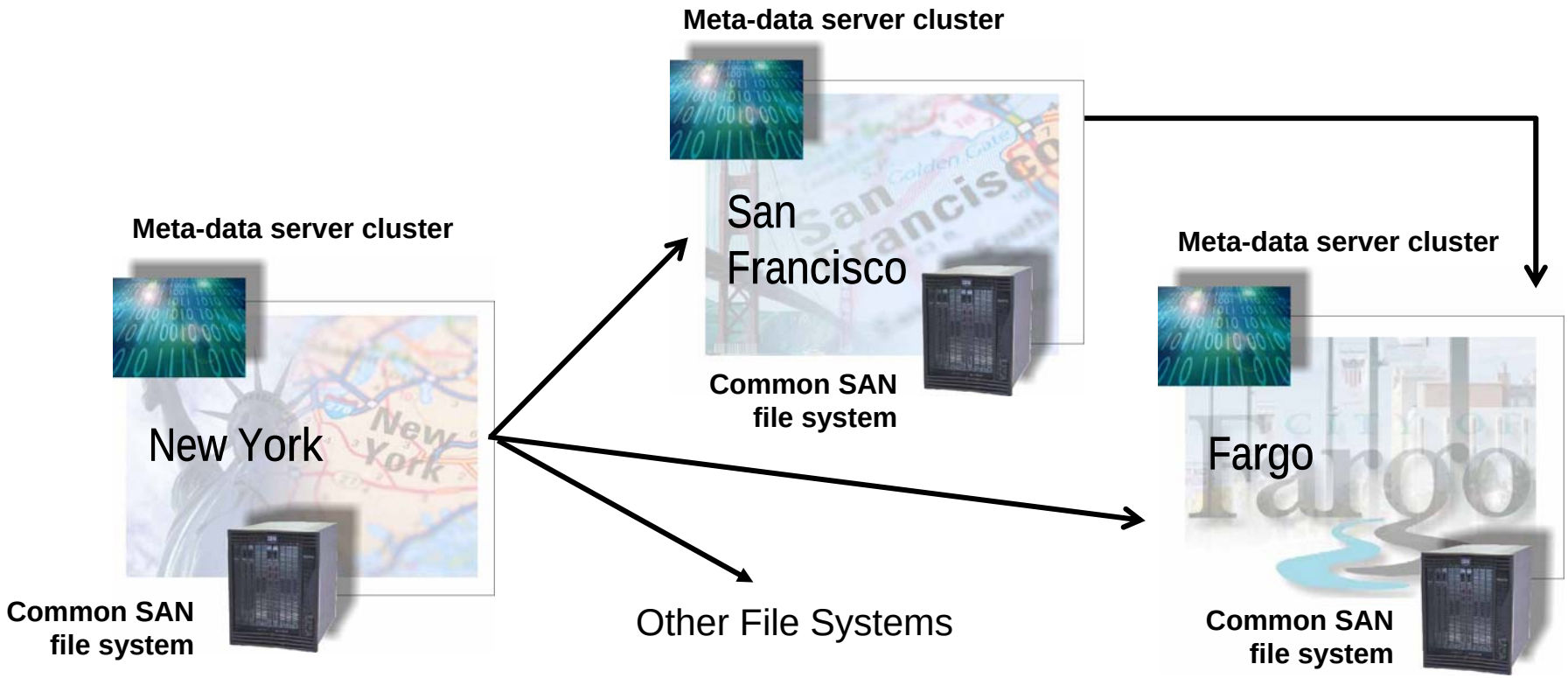


Data stays online

- Until policy moves to archive

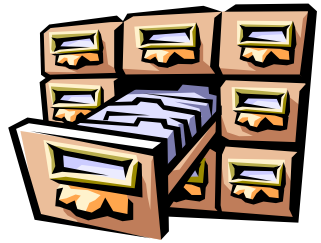
Unified File Federation Architecture (UFFA)

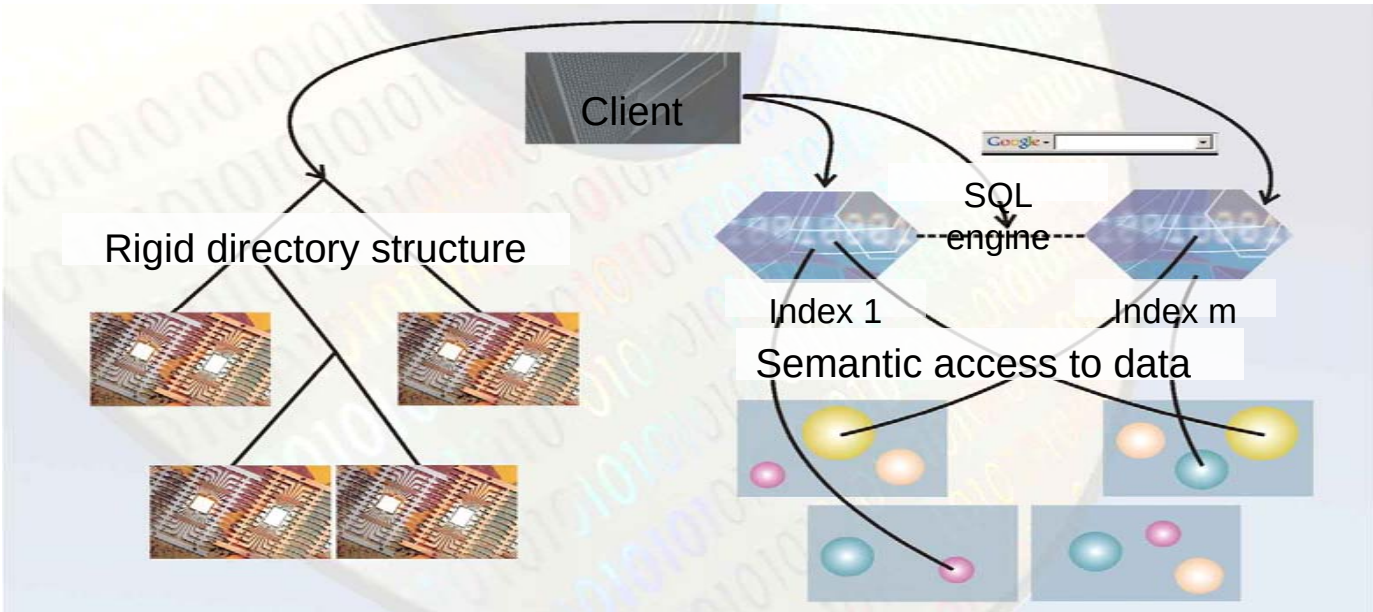
- Geographically distributed server clusters forming single namespace
  - Replication of files/containers for good local performance
  - Extended protocols for consistency across replicas
  - Migration of primary copy of data to point of use



Google for the Enterprise

- File system directory structure based upon file cabinet metaphor
  - Each file exists in one place in a fixed hierarchy
  - To find a file must remember where it was placed
- Metaphor has not scaled with growth in number of files
  - Modern scalable file systems aim for storing a billion files
  - Need paradigm shift to a more flexible mechanism





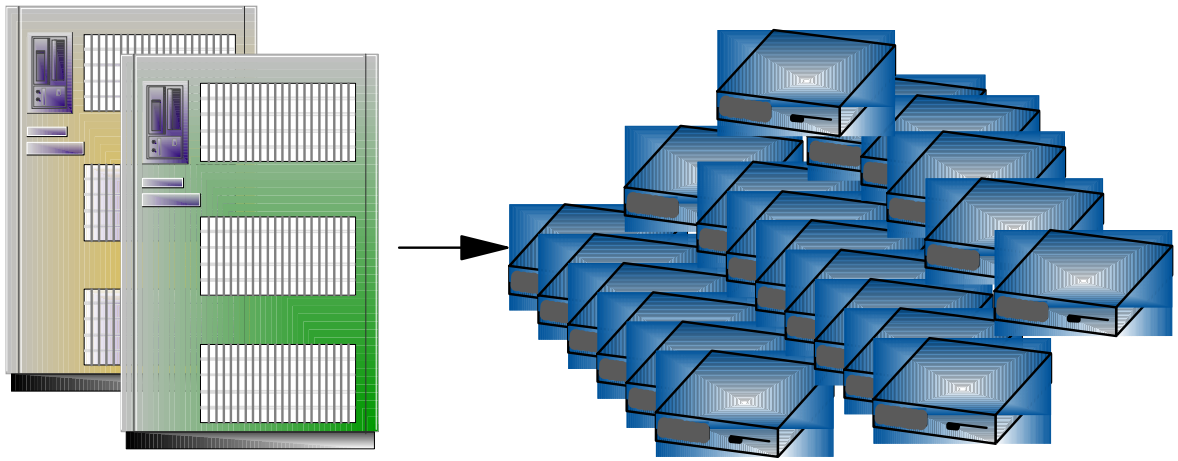
Collective Intelligent Bricks

Key Ideas for low maintenance

- One basic building block
- Deferred maintenance
- Eliminate high maintenance parts like cables, fans
- Continuous, autonomic data migration, cache mgmt, throttling
- New bricks work with old bricks

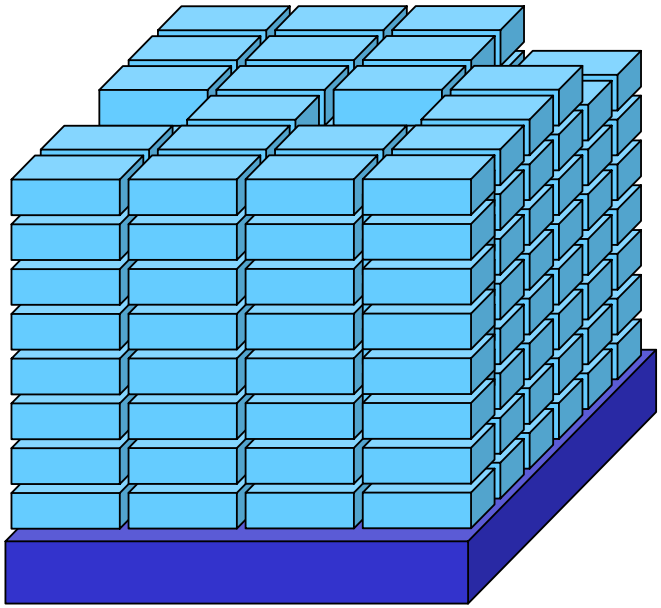
Key Ideas for low cost

- ATA drives
- low-cost distributed switching
- cheap, low-power processors
- dense packaging



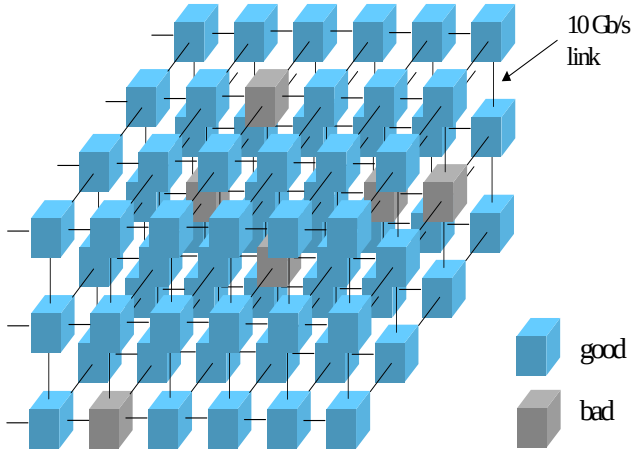
Collective Intelligent Bricks (CIB)

- **Bricks are stacked on vertical columns (not shown) for power insertion and heat removal**
- **Bricks communicate with neighbors in a 3D mesh**



Cube

=

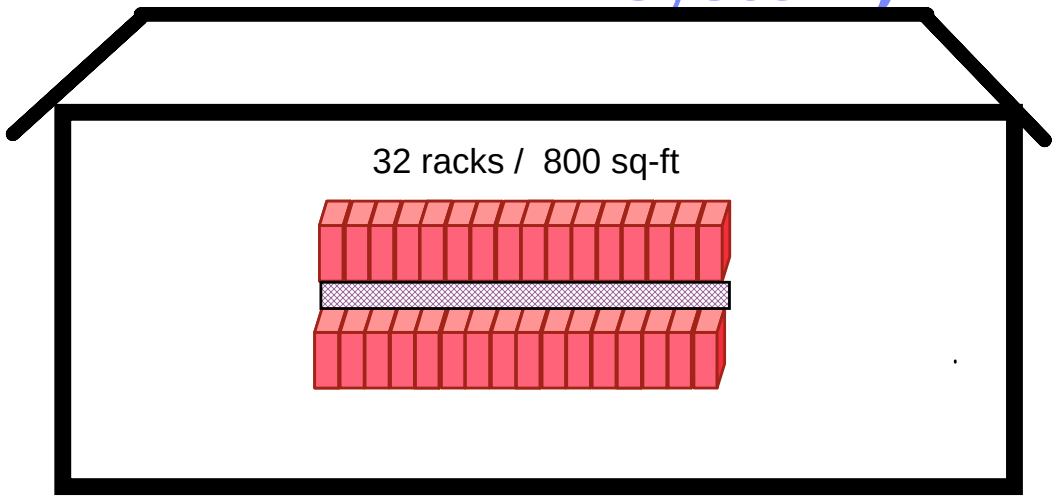


Logic View = 3D Mesh

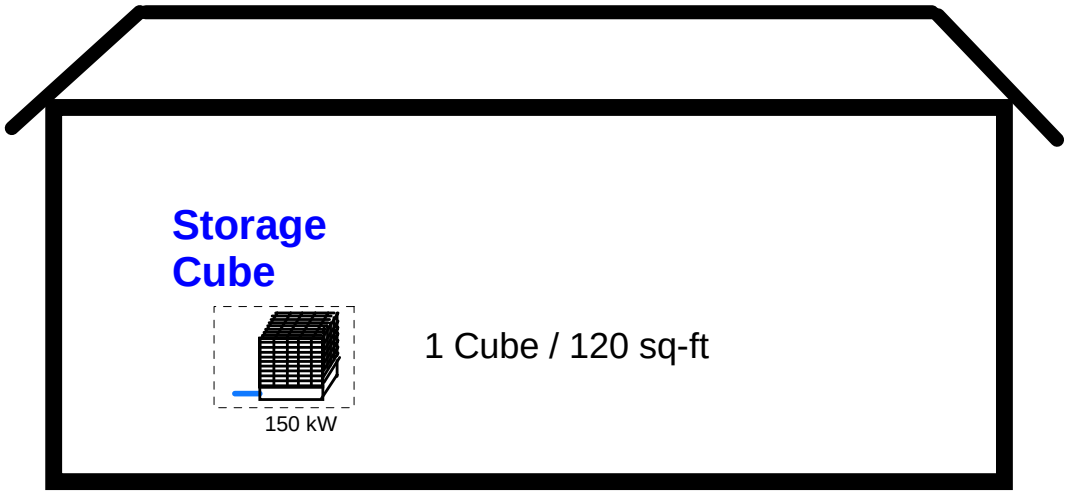
Almaden prototype - 3x3x3 bricks, 324 disks, 32 TB, 2 ft on a side => could store all the books of the library of congress, June 2004

CIB Prototype

1PB: Conventional vs Storage Cube (w/ 250-GB, 18-W, 3.5" disks in either system)



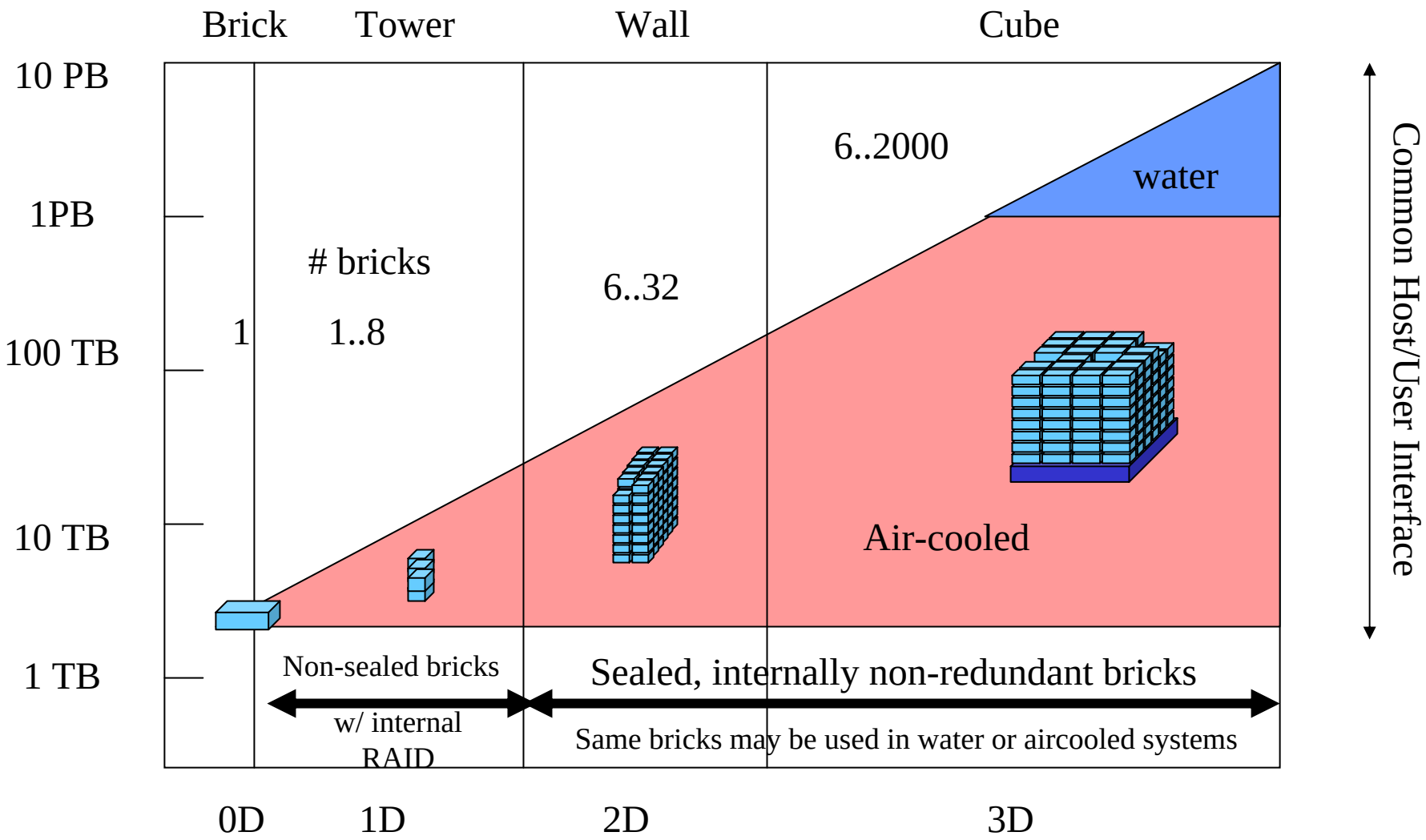
Today...



Future...

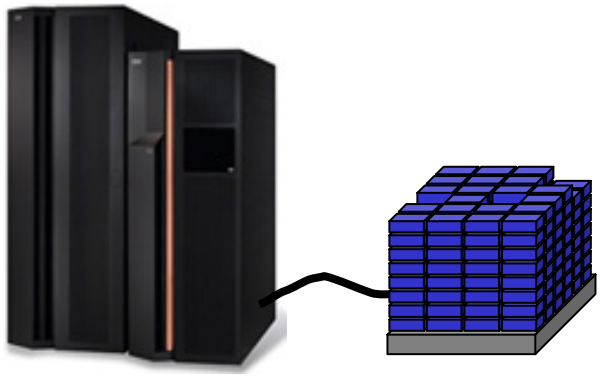
14% Floorspace

CIB Family

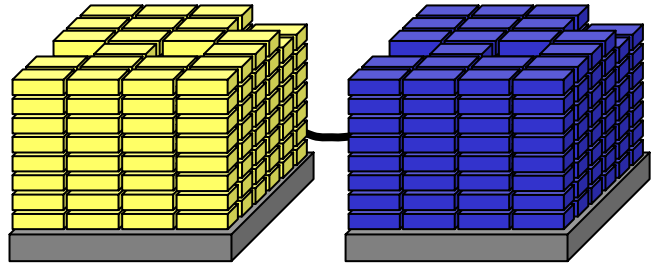


IceCube: Storage and/or Compute Server

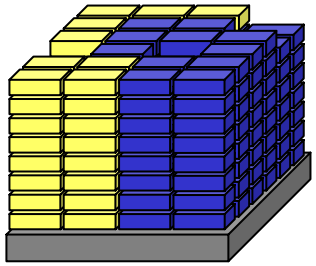
Color code: Blue=Storage Bricks Yellow=Compute Bricks Green=Mixed Bricks



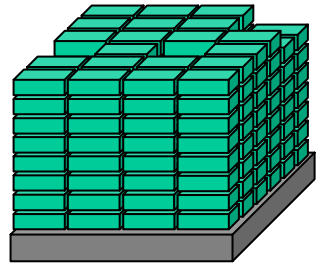
A: Host Storage Server



B: Compute Server Storage Server



C: Functions mixed in one Cube



D: Functions mixed within bricks