

Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Jim Gray

Microsoft

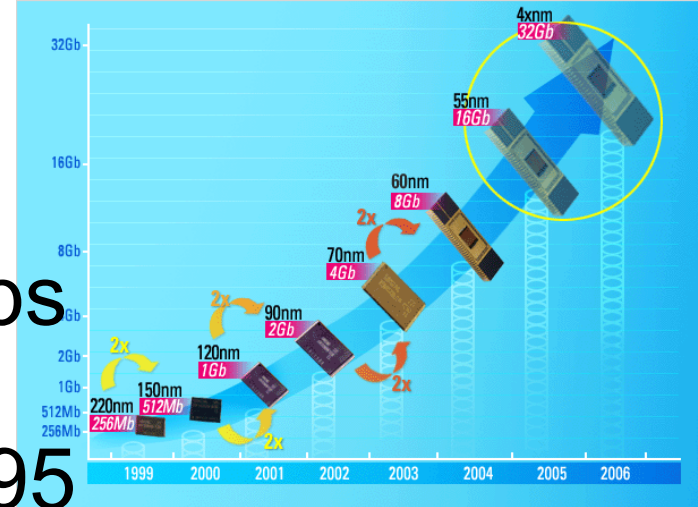
December 2006

Tape Is Dead Disk is Tape

- 1TB disks are available
- 10+ TB disks are predicted in 5 years
- Unit disk cost: ~\$400 → ~\$80
- But: ~ 5..15 **hours to read (sequential)**
- ~15..150 **days to read (random)**
- Need to treat **most of disk as Cold-storage archive**

FLASH Storage?

- 1995 16 Mb NAND flash chips
- 2005 16 Gb NAND flash
- Doubled each year since 1995



- Market driven by Phones, Cameras, iPod,...

Low entry-cost,
~\$30/chip → ~\$3/chip

- 2012 1 Tb NAND flash

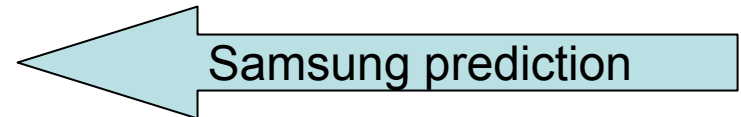
== 128 GB chip

== 1TB or 2TB “disk”

for ~\$400

or 128GB disk for \$40

or 32GB disk for \$5



FLASH Some Parameters

5,000 IO/s per chip!

- Chip read ~ 20 MB/s
write ~ 10 MB/s
N chips have N x bandwidth
- Latency ~ 25 μ s to start read,
~ 100 μ s to read a “2K page”
~ 2,000 μ s to erase
~ 200 μ s to write a “2K page”
- Power ~ 1W for 8 chips and controller

What's Wrong With FLASH?

- Expensive: \$/GB
 - 50x more than disk today
 - Ratio may drop to 10x in 2012
- Limited lifetime
 - ~100k to 1M writes / page
 - requires “*wear leveling*”
but, if you have 1B pages,
then 15,000 years to “use” ½ the pages.
- Slow to write
you can only write 0's,
so erase (set all 1) then write.

Obvious Uses For Flash

- PDAs, cameras, iPod,
- Laptop disks
 - power, rugged, quiet, big enough, ...
- Not so obvious use:
 - ARCHIVE for photo/music/..
because it's simple to understand.
 - Enterprise drives (lots of IO/s per \$
per watt
per liter)

One Could Make a Flash Disk (or a Flash File System)

- 6K random reads/sec, 3K random writes/sec
- The IO capacity of 30..45 disks
- Uses 1 W vs 500W...

Less space, ...

- See

[“A Design for High-Performance Flash Disks”](#)

Birrell, Isard,
Thacker, Wobber

MSR-TR-2005-176

Dell		PowerEdge 2900 Server with 1 PowerEdge SC1420 Client		TPC-C Rev 5.7 Original Report Date June 30, 2006	
Total System Cost		TPC-C Throughput		Price/Performance	
\$64,512		65,833 tpmC		\$.98 / tpmC	
Processors		Database Manager		OS	
1/2/2 Dual Core Intel® Xeon® 5160, 4MB Cache, 3.00GHz 1333, 667MHz FSB		Microsoft SQL Server 2005 Standard x64 Edition		Microsoft Windows Server 2003 Standard x64 Edition SP1	
Other Software		Availability Date		Number of Users	
Windows Server 2003 Standard Edition w/ COM+ Internet Information Server 6.0 Microsoft Visual C++		June 26, 2006		53,000	
53,000 Emulated Users Running on 2 PE6350 RTE Machines Connected Through 1 100BaseT Segment		PowerEdge 2900 1/2/2 Dual Core Intel® Xeon® 5160, 4MB Cache, 3.00GHz, 24GB 667MHz FSB 1 Dell PERC5 SAS RAID Controller, 1 Integrated PERC5i SAS RAID Controller 8 73GB 3.0GSP/S, SAS 3.5IN, 10K 2 NetXtreme II GigE TOE		6 PowerVault 1000MD SAS Disk Pods 90 36GB 15K RPM SAS Disks	
1 PowerEdge SC1420 Client 2/2/2 Intel Xeon 3.2GHz w/ 2MB L2 1024 MB RAM 1 80GB SATA 7.2K Disk 2 Intel Pro100+ Ethernet NICs					
System Component		Server		Each Client	
Processor/Core/Cache	1	1/2/2 Dual Core Intel® Xeon® 5160, 4MB Cache, 3.00GHz, 1333	2	2/2/2 Intel® Xeon® w/ 2MB L2, 3.2 GHz	
Memory		24GB 667 FB-DIMM		1024 MB	
Disk Controllers	1	Dell PERC5 RAID Controller, Integrated PERC5i Raid Controller.	1	Onboard SATA	
Disk Drives	90	36GB SAS 15K	1	80GB 7.2K SATA	
		8 73GB SAS 10K			
Total Storage	98	3345 GB SAS	1	80GB SATA	
Other	2	Broadcom NetXtreme II GigE CD-ROM	2	10/100MB BT NIC CD-ROM	
	1		1		

replace with 1
10TB disk
and 3 FLASH
disks

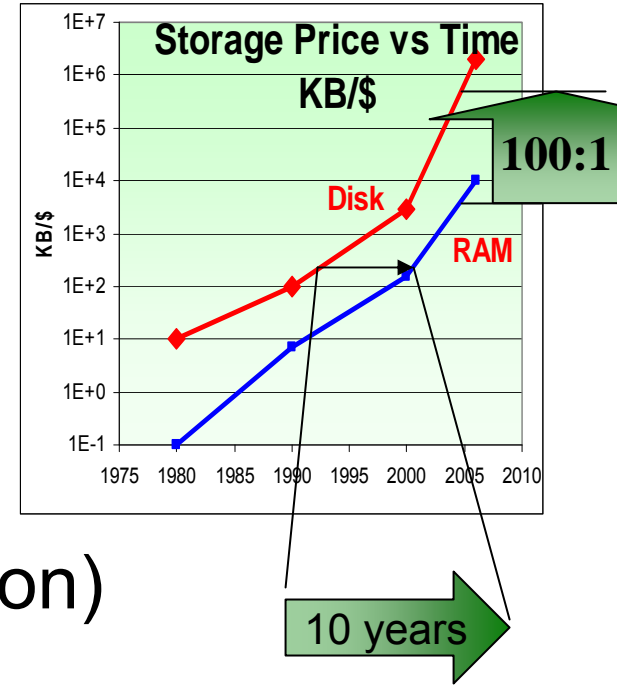


We Are Not There Yet

- Current FLASH disks could do much better on writes (100x better (!))
Algorithms are known but...
- This changes many ratios
Access time is 20x less (~200us)
IOps is 100x more
- Re-evaluate page sizes MSR-TR-2006-168
[FlashDB: Dynamic Self-tuning Database for NAND Flash](#), Suman Nath, Aman Kansal

RAM Locality is King

- The cpu mostly waits for RAM
- Flash / Disk are 100,000 ... 1,000,000 clocks away from cpu
- RAM is ~100 clocks away unless you have locality (cache).
- If you want 1CPI (clock per instruction) you have to have the data in cache (program cache is “easy”)
- This requires cache conscious data-structures and algorithms sequential (or predictable) access patterns
- Main Memory DB is going to be common.



Tape is Dead
Disk is Tape
Flash is Disk
RAM Locality is King

Jim Gray

Microsoft

December 2006