SEAGATE | GOVERNMENT SOLUTIONS

: Designing Storage Architectures for Digital Collections

March 14, 2022

LIBRARY LIBRARY OF CONGRESS

Some really smart people think things will happen faster than they do

AND SOME NOT SO SMART PEOPLE ALSO ☺

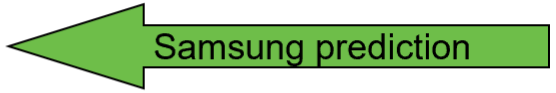
- Disruptive changes in storage were predicted in December 2006
- To Quote Jim Gray
  - Famous computer scientist who received the Turing Award in 1998 "for seminal contributions to database and transaction processing research
  - Tape is Dead
  - Disk is Tape
  - Flash is Disk
  - RAM Locality is King
    - People often forget this 1
  - I have the whole presentation as backup slides for feel free to search Jim Gray Tape is Dead”

What was Jim Predicting

VENDORS MAKE BOLD CLAIMS

- Jim said:

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



- We got to 1 Tb in 2017 and the cost is now below the less than ~$200 for 2 TB consumer SSD, but what happened and why are not where Jim said they would be?

1st a Back Story

- I invited Jim to dinner in March of 2007
- Restaurant Alma on University Ave
  - I was **honored** that he accepted
- I was a very serious tape proponent in those days
- I wanted to convince him that he was wrong
  - Guess what I was wrong, but Jim was wrong on the timeline
  - Why?

Why are predictions usually right by smart people but

Timeline is often very wrong

- Carl Watts and I talked about the end of tape over 8 years ago at DSA
  - It is still here
- Jim talked about disk being relegated to archive in 2006
- Change does not happen fast as fast as many predict
  - But sometimes it does
    - Linux, x86, PCI bus, are examples of fast change
- What makes some change fast and others much slower?

Looking back, it is Requirements vs. Nice To Have

What is really a requirement?

If you are building a system, you need to interface with peripheral devices

- In the past everyone had their own interface
- Not workable for innovation
- Not workable for time to market
- Not workable for cost

PCI (1992) took over very quickly and eliminate Intergraph's market control for graphics on the low end and then eventually SGI's high-end dominance

- Anyone could build a graphics card and the engineering required was the graphics not the interface to the system
- This kick started many companies we have today like NVIDIA (1993)

What about Jim's predication on storage

It will be become true, but it will still take more time

Let's look at HDDs 1st- 246 EB shipped

HDD Shipments in Q2 2021						
Data by Trendfocus						
		HDDs in million	**Q/Q growth**	**Avg HDD (TB)**	**Exabytes**	**Market share**
Vendor	**Seagte**	28.17	2.30%	5.67	152.3	41.80%
	Toshibia	13.98	3.20%	3.75	49.97	20.80%
	WDC	25.4	9.20%	6.13	148.43	37.40%
Client PC	**Total Desktop**	14.61	-1.20%	2.59	36.12	10.30%
	Total Mobile	18.78	-6.90%	1.74	31.18	8.89%
	Total Client	33.39	-4.51%	2.11	67.3	19.19%
Enterprise	**3.5" enterprise**	19.31	19.90%	13.2	243	69.29%
	2.5" enterprise	3.44	16.40%	1.41	4.61	1.31%
	Total enterprise	22.75	19.30%	11.41	**247.61**	70.60%
CE						
	3.5" CE	8.89	19.80%	3.88	32.86	9.37%
	2.5" CE	2.34	-7.50%	0.79	1.76	0.50%
	Total CE	11.23	12.90%	3.23	34.62	9.87%
Total		67.6	5.00%	5.43	350.7	100%

Now SSDs

Lots more time

35 EB shipped

SSD Shipments in Q2 2021					
Data by Trendfocus					
		SSDs in million	**Avg SSD (TB)**	**Exabytes**	**Q/Q Unit Growth**
Client SSDs	**2.5 Inc**	15.38	0.6	8.74	?
	M.2 Modules	71.48	0.48	32.81	1.50%
	Total	86.86	1.08	41.55	1.70%
Enterprise SSDs	**SATA**	5.79	0.95	5.25	~17%
	SAS	1.1	3.51	3.68	1%
	PCIe	5.84	3.26	18.16	14.90%
	Total	12.74	2.95	**35.79**	*?*
All SSDs		99.596	0.72	68.63	0.16%

This is a factor 6.92 times. That is a lot of NAND fabs to build 4Q21 only 69.38 EB. No real changes

Storage is the forgotten child until you need something

HDDs are often not workable for innovation

- You can still get much of your job done with disk
- The parts of the problems that are not workable such as ML/AI, HPC problems, etc. have moved to flash

Not workable for time to market

- Putting disk in cars could be done but ..
- Not workable for cost and reliability

What I realized that change is all about requirements along with innovation

- Innovation alone is not going to change markets quickly.

THANKS to LOC Especially Jane

BACK IN JUNE OF 2001 JANE CALLED ME, 1 DAY

I have enjoy working with and previously for LOC for over ~22 year now.

This is my last DSC and I appreciated and truly enjoy the interactions I have had here in this building and other buildings. I have really enjoyed my interactions with LOC!

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

Jim Gray
Microsoft
December 2006

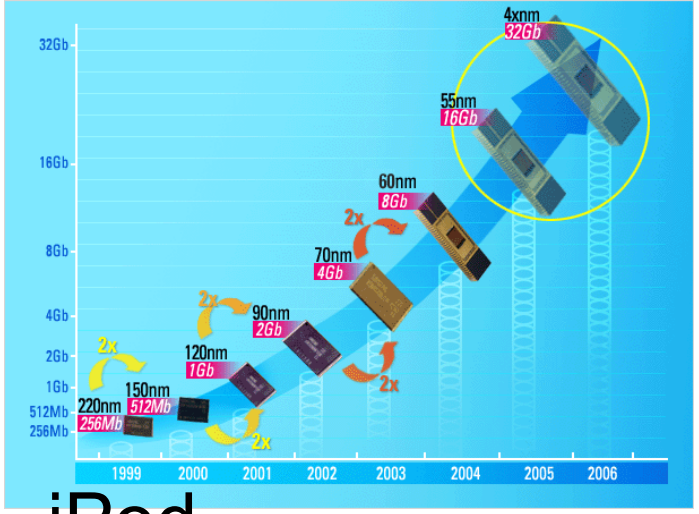
Tape Is Dead
Disk is Tape

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

Samsung prediction

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 µs to start read,
~ 100 µs to read a “2K page”
~ 2,000 µs to erase
~ 200 µs 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…

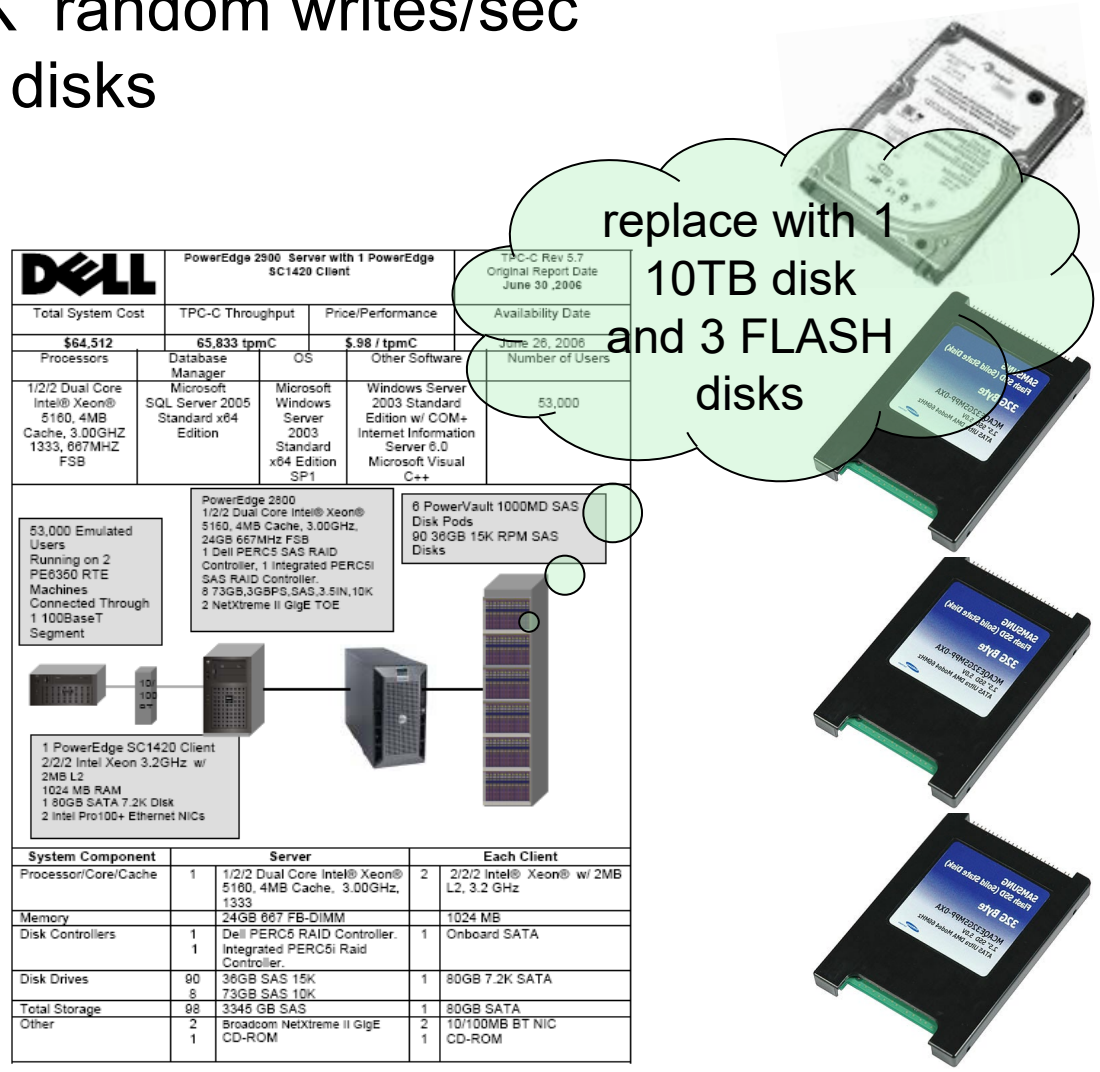
Less space, …

See

"A Design for High-Performance Flash Disks"

Birrell, Isard, Thacker, Wobber

MSR-TR-2005-176

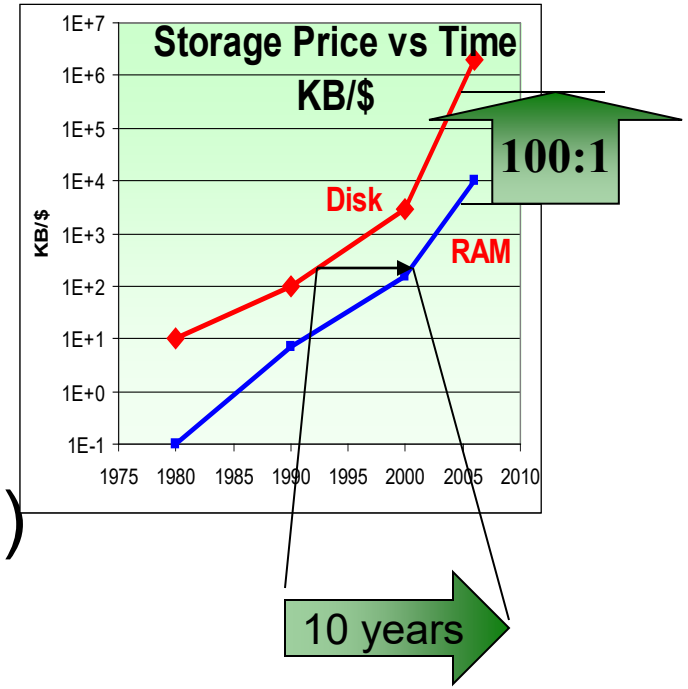


We Are Not There Yet

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