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

<table>
<thead>
<tr>
<th>Vendor</th>
<th>HDDs in million</th>
<th>Q/Q growth</th>
<th>Avg HDD (TB)</th>
<th>Exabytes</th>
<th>Market share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seagate</td>
<td>28.17</td>
<td>2.30%</td>
<td>5.67</td>
<td>152.3</td>
<td>41.80%</td>
</tr>
<tr>
<td>Toshiba</td>
<td>13.98</td>
<td>3.20%</td>
<td>3.75</td>
<td>49.97</td>
<td>20.80%</td>
</tr>
<tr>
<td>WDC</td>
<td>25.4</td>
<td>9.20%</td>
<td>6.13</td>
<td>148.43</td>
<td>37.40%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Client PC</th>
<th>HDDs in million</th>
<th>Q/Q growth</th>
<th>Avg HDD (TB)</th>
<th>Exabytes</th>
<th>Market share</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Desktop</td>
<td>14.61</td>
<td>-1.20%</td>
<td>2.59</td>
<td>36.12</td>
<td>10.30%</td>
</tr>
<tr>
<td>Total Mobile</td>
<td>18.78</td>
<td>-6.90%</td>
<td>1.74</td>
<td>31.18</td>
<td>8.89%</td>
</tr>
<tr>
<td>Total Client</td>
<td>33.39</td>
<td>-4.51%</td>
<td>2.11</td>
<td>67.3</td>
<td>19.19%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Enterprise</th>
<th>HDDs in million</th>
<th>Q/Q growth</th>
<th>Avg HDD (TB)</th>
<th>Exabytes</th>
<th>Market share</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5” enterprise</td>
<td>19.31</td>
<td>19.90%</td>
<td>13.2</td>
<td>243</td>
<td>69.29%</td>
</tr>
<tr>
<td>2.5” enterprise</td>
<td>3.44</td>
<td>16.40%</td>
<td>1.41</td>
<td>4.61</td>
<td>1.31%</td>
</tr>
<tr>
<td>Total enterprise</td>
<td>22.75</td>
<td>19.30%</td>
<td>11.41</td>
<td>247.61</td>
<td>70.60%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CE</th>
<th>HDDs in million</th>
<th>Q/Q growth</th>
<th>Avg HDD (TB)</th>
<th>Exabytes</th>
<th>Market share</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5” CE</td>
<td>8.89</td>
<td>19.80%</td>
<td>3.88</td>
<td>32.86</td>
<td>9.37%</td>
</tr>
<tr>
<td>2.5” CE</td>
<td>2.34</td>
<td>7.50%</td>
<td>0.79</td>
<td>1.76</td>
<td>0.50%</td>
</tr>
<tr>
<td>Total CE</td>
<td>11.23</td>
<td>12.90%</td>
<td>3.23</td>
<td>34.62</td>
<td>9.87%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th>HDDs in million</th>
<th>Q/Q growth</th>
<th>Avg HDD (TB)</th>
<th>Exabytes</th>
<th>Market share</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>67.6</td>
<td>5.00%</td>
<td>5.43</td>
<td>350.7</td>
<td>100%</td>
</tr>
</tbody>
</table>

Data by Trendfocus
Now SSDs

Lots more time

35 EB shipped

<table>
<thead>
<tr>
<th>SSD Shipments in Q2 2021</th>
<th>Data by Trendfocus</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SSDs in million</strong></td>
<td><strong>Avg SSD (TB)</strong></td>
</tr>
<tr>
<td>Client SSDs</td>
<td></td>
</tr>
<tr>
<td>2.5 Inc</td>
<td>15.38</td>
</tr>
<tr>
<td>M.2 Modules</td>
<td>71.48</td>
</tr>
<tr>
<td>Total</td>
<td>86.86</td>
</tr>
<tr>
<td>Enterprise SSDs</td>
<td></td>
</tr>
<tr>
<td>SATA</td>
<td>5.79</td>
</tr>
<tr>
<td>SAS</td>
<td>1.1</td>
</tr>
<tr>
<td>PCIe</td>
<td>5.84</td>
</tr>
<tr>
<td>Total</td>
<td>12.74</td>
</tr>
<tr>
<td>All SSDs</td>
<td>99.596</td>
</tr>
</tbody>
</table>

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 \( \mu \text{s} \) to start read,
         ~  100 \( \mu \text{s} \) to read a “2K page”
         ~  2,000 \( \mu \text{s} \) to erase
         ~  200 \( \mu \text{s} \) 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

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