The Future of Storage and Filesystem: Some Ruminations

Theodore Ts'o
September, 2009
It was the best of times;
It was the worst of times
It was a dark and stormy night...
It's all about the money
“economics”, not “economy”
The Cloud and Other Buzzwords

- “Game Changer”
- “Only buy what you need”
- All good things...
- But it means people will buy what they need!
Economies of Scale

- Drove adoption of Linux...
- ... and commoditization of hardware
- Changes in the economies of scale can make or break companies
  - Example: SGI and Sun
People aren't buying as many PC's

- Because of the economy?
- Because of Windows Vista?
- Because of they don't need a new computer?
  - Many of the people who would buy a new computer already have purchased one
  - People are asking the question “do we really need a new computer every 2-3 years?”
  - No killer application that requires bigger, badder machines
  - In fact, netbooks and the Atom processor is proof we don't need bigger machines → cloud computing!
Consequences

- Especially deadly for HDD manufacturers
  - A commodity business to begin with
  - R&D costs have to be amortized across razor-thin profit margins.
- Will cloud computing take up the slack?
  - Maybe...
  - One of the arguments of cloud computing is that it's more efficient. Ergo, cloud providers will buy less storage than the systems they are replacing
- What about SSD's?
Very insightful talk by Dr. Steve Hetzler. IBM Fellow

Manager of the Storage Architecture group at IBM Research at Alamaden

“The Storage Chasm: Implications for the Future of HDD and Solid State Storage”

Find the presentation; it's on Google. Well worth reading.

Just a few slides from his talk....
Why Economics Matters

Imagine a Terabyte HDD for $100...
Why Economics Matters

BitTorrent...

But that’s $250,000.00 worth of MP3s! How will I ever afford them?

Take Away

We need to consider supply costs.

Feb, 2009

Steven R. Hetzler, © Copyright IBM Corp. 2008, 2009
# NAND Flash and HDD Fabrication YE2008

<table>
<thead>
<tr>
<th></th>
<th>MLC NAND Flash</th>
<th>HDD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wafer diam.</td>
<td>300mm (12&quot;)</td>
<td>Head Wafer: 6&quot;</td>
</tr>
<tr>
<td>Node</td>
<td>45nm</td>
<td>Head Node: 60nm</td>
</tr>
<tr>
<td>Die capacity*</td>
<td>2GB</td>
<td>Disk diam.: 3.5&quot;</td>
</tr>
<tr>
<td>Dies/wafer</td>
<td>425</td>
<td>Heads/wafer: 30,000</td>
</tr>
<tr>
<td>Wafer Capacity</td>
<td>850 GB</td>
<td>Disk Capacity**: 375 GB</td>
</tr>
<tr>
<td>Daily output</td>
<td>1,250 wafers</td>
<td>Disks/day: 100,000</td>
</tr>
<tr>
<td>PB/line/day</td>
<td>1.1 PB/day</td>
<td>PB/line/day: 38PB/day</td>
</tr>
<tr>
<td>PB/line/year</td>
<td><strong>390 PB/year</strong></td>
<td>PB/year: 14,000PB/year</td>
</tr>
<tr>
<td>Fab cost</td>
<td><strong>$3.4B</strong>*</td>
<td>Fab cost (est): $1B</td>
</tr>
</tbody>
</table>

* Assumes highest density 2-bit MLC flash. Wafer GB for SLC flash will be lower.
** Assumes highest density SATA disk.
*** Source: IBS

---

Feb, 2009

Steven R. Hetzler, © Copyright IBM Corp. 2008, 2009
300mm Fab Cost

- Capital costs increase as node size shrinks
  - Today’s 45nm fab $3.4B*
  - 22nm fab increases to $7.5B
    - Expected to start around 2012**

* IBS  
** Gartner, Jan 2008
If all of the world's Si fabs converted to manufacture MLC flash, could manufacture 16kPB/year.

Total WW HDD production 125kPB

Total WW HDD revenue $35B

Value of total SI fab output revenue $280B

Take home: It will be challenging for SDD prices to drop below HDD on a $/GB basis
Implications

- Storage advances from HDD may not be coming as quickly as in the past decade
  - Seagate has already cut back its R&D
- SDD's, while exciting may not have costs drop as quickly as we might like, remaining niche/luxury item.
- As high-end storage starts becoming more of a commodity, who will fund academic (and commercial) research into filesystems? The FAST conference seems healthy so far, but....
How to Overcome Depression

By Dr. F. Flach

One of the country's leading authorities in the field of Depression. He is associate clinical professor of Psychiatry at Cornell University Medical College and attending psychiatrist at the Payne Whitney Clinic of the New York Hospital.

Depression may be caused by:
- Insomnia
- Marital problems
- Sexual frustration
- Loss of job
- Financial crises
- Severe sickness
- Constant pain
- Death of kin
- No zest for life
Part II: “Always look on the bright side of life!”
New interest and work in filesystems

- In the Linux world, *3* new filesystems were merged during the 2.6.30 merge window: Nilfs2, POHMELEFS, and EXOFS
  - Other recent additions to Linux: btrfs, ext4, ubifs, read-only bind mounts
- Advent of SSD's is spurring changes and improvement in the storage stack
  - Tuning storage algorithms for different storage devices
  - Multi-tier caching
  - Very large numbers of I/O transactions/second
SDD's can make sense

- But not if you think in terms of raw $/GB
- SSD's advantages
  - $/GB scales to sizes < 80GB
  - Shock resistance / power savings
  - $/Random I/O Performance in ($/MB/s)
- Takeaway: “horses for courses”
- “SSD's are the last storage technology you need” makes as much sense as “ZFS is the last word in filesystems”.
Opportunities

- People are at least talking about changing some storage interfaces!
  - Will the shift to cloud computing make this easier?
  - Commodity hardware made storage cheaper
  - ... but it has also held back evolution in storage interfaces
Storage Interface Changes

- 512 → 4k sector change
- Object based storage
- Disk ↔ Raid ↔ LVM ↔ Filesystem
  - Getting information about the most efficient way to do reads/writes to the storage device
- Filesystem API changes
  - fallocate()
  - refilexk()
Opportunities

- People are at least talking about changing some storage interfaces!
  - Will the shift to cloud computing make this easier?
- Open Source → increasing experimentation
  - Linux as a filesystem developer's laboratory
  - Potential challenges: copyright compatibility and fears of patent contamination
Interesting problems

- Proper benchmarking tools
  - With filesystem aging
  - For different workloads... most importantly your workload.
  - Lies, d*mn lies, and benchmarks
Large File Creates, Raid, 1 Threads

![Graph showing FFSB Ops per Sec with various configurations and tests.](image)
Large File Creates, Raid, 16 Threads

FFSB Ops per Sec

Ops/sec

0
10000
20000
30000
40000
50000
60000
70000
80000
90000
100000
110000
120000
130000
140000
150000

0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150

2.6.27-rc7-ext3-transactions-ext3
2.6.27-rc7-ext4-transactions-ext4
2.6.27-rc7-xfs-transactions-xfs
2.6.27-rc7-jfs-transactions-jfs
2.6.27-rc7-btrfs-transactions-btrfs
2.6.27-rc7-btrfs-transactions-btrfs-nodatsum
2.6.27-rc7-btrfs-transactions-btrfs-nocow-nosum
Large File Creates, RAID, 128 threads
Large File Random Writes, RAID, 1 thread

FFSB Ops per Sec

- 2.6.27-rc7-ext3-transactions-ext3
- 2.6.27-rc7-ext4dev-transactions-ext4
- 2.6.27-rc7-xfs-transactions-xfs
- 2.6.27-rc7-ifs-transactions-ifs
- 2.6.27-rc7-btrfs-transactions-btrfs
- 2.6.27-rc7-btrfs-transactions-btrfs-nodatasum
- 2.6.27-rc7-btrfs-transactions-btrfs-nocow-nosum
Large File Random Writes, RAID, 16 threads
Large File Random Writes, Single Disk, 8 threads

FFSB Ops per Sec

Ops/sec

2.6.27-rc7-ext3-transactions-ext3
2.6.27-rc7-ext4dev-transactions-ext4
2.6.27-rc7-xfs-transactions-xfs
2.6.27-rc7-jfs-transactions-jfs
2.6.27-rc7-btrfs-transactions-btrfs
2.6.27-rc7-btrfs-transactions-btrfs-nodatacow
2.6.27-rc7-btrfs-transactions-btrfs-nodatacum
2.6.27-rc7-btrfs-transactions-btrfs-nocw-nocsum
Interesting problems

- Proper benchmarking tools
- Storage Management
  - Backup management
  - Unstructured data retrieval
- Global Filesystems
  - Beyond cluster filesystems
  - Time to revisit distributed filesystems again?
  - What does disconnected operations mean in the era of mobile devices and mobile networking?
The Road Ahead
This work represents the view of the author(s) and does not necessarily represent the view of IBM or of the Linux Foundation.

IBM is a registered trademark of International Business Machines Corporation in the United States and/or other countries.

Linux is a registered trademark of Linus Torvalds.

Microsoft and Windows are trademarks of Microsoft Corporation in the United States, other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.
Photo Credits

- Slide 2, The Writings of Charles Dickens v2 p 164, Wikimedia commons
- Slide 3, "Stormy Night over Mula”, Diego Sevilla, (flikr, cc-by-nc-2.0)
- Slide 4, "Money money money”, Gavin Anderson (flikr, cc-by-sa-2.0)
- Slide 5, "RIP Good Times”, Sequoia Capital, fair use
- Slide 18, "How to overcome depression”, Kevin Doeley (flikr, cc-by-2.0)
- Slide 19, "Smiley Face”, Jovike (flikr, cc-by-nc 2.0)
- Slide 26, "The Road Ahead”, Florian Boyd (flikr, cc-by-sa-2.0)