Flash-based SSDs

[ Material based on slides from Tyler Caraza-Harter ]
www: https://tyler.caraza-harter.com
Cost: HDD vs. SSD

Cost: HDD vs. SSD

Note: These are trends, not the most up-to-date data.

There are different classes of HDDs and SSDs which complicate this graph, but the thing to note is that there is a gap, but it is narrowing and all costs are trending downward.

Disk Overview

I/O requires: seek, rotate, transfer

Inherently:
- not parallel (only one head)
- slow (mechanical)
- poor random I/O (locality around disk head)

Random requests take 10ms+
Flash

Hold charge in cells. No moving parts!

Inherently parallel.

No seeks!
SLC: Single-Level Cell

NAND Cell
SLC: Single-Level Cell

NAND Cell

charge

1
SLC: Single-Level Cell

NAND Cell

charge

0
MLC: Multi-Level Cell

NAND Cell

charge

00
MLC: Multi-Level Cell

NAND Cell

charge
MLC: Multi-Level Cell

NAND Cell

charge

10
MLC: Multi-Level Cell

NAND Cell
Single- vs. Multi- Level Cell

SLC

 MLCD
Single- vs. Multi-Level Cell

SLC
expensive
robust

MLC
cheap
sensitive
Wearout

Problem: flash cells wear out after being overwritten too many times.

MLC: ~10K times
SLC: ~100K times

Usage strategy:
Wearout

Problem: flash cells wear out after being overwritten too many times.

MLC: ~10K times
SLC: ~100K times

Usage strategy: wear leveling.
- prevents some cells from wearing out while others still fresh.
Banks

Flash devices are divided into banks (aka, planes).

Banks can be accessed in parallel.

Bank 0  Bank 1  Bank 2  Bank 3
Banks

Flash devices are divided into banks (aka, planes).

Banks can be accessed in parallel.

read

Bank 0

read

Bank 1

read

Bank 2

read

Bank 3
Banks

Flash devices are divided into banks (aka, planes).

Banks can be accessed in parallel.

Bank 0  Bank 1  Bank 2  Bank 3
Banks

Flash devices are divided into banks (aka, planes).

Banks can be accessed in parallel.
Banks

Flash devices are divided into banks (aka, planes).

Banks can be accessed in parallel.

Bank 0  Bank 1  Bank 2  Bank 3
Flash Writes

Writing 0's:
- fast, fine-grained

Writing 1's:
- slow, course-grained
Flash Writes

Writing 0’s:
- fast, fine-grained
- called “program”

Writing 1’s:
- slow, course-grained
- called “erase”
Flash Writes

Writing 0’s:
- fast, fine-grained [pages]
- called “program”

Writing 1’s:
- slow, course-grained [blocks]
- called “erase”
A Bank Consists of Blocks

Bank 0  Bank 1  Bank 2  Bank 3
A Bank Consists of Blocks

each bank contains many “blocks”
A Block Consists of Pages

1111 1111 1111 1111
1111 1111 1111 1111
1111 1111 1111 1111
1111 1111 1111 1111
A Block Consists of Pages

one block
The Heirarchy of SSD components:

- One NAND flash Chip
- Is made of up several Banks
- Is made up of several blocks
- Is made up of several pages
<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1111</td>
<td>1111</td>
<td>1111</td>
<td>1001</td>
</tr>
<tr>
<td>1111</td>
<td>1111</td>
<td>1111</td>
<td>1111</td>
</tr>
<tr>
<td>1111</td>
<td>1111</td>
<td>1111</td>
<td>1111</td>
</tr>
<tr>
<td>1111</td>
<td>1111</td>
<td>1111</td>
<td>1111</td>
</tr>
</tbody>
</table>

Block

program
Block

```
1111 1111 1111 1001
1111 1111 1111 1111
1111 1111 1111 1111
1111 1111 1111 1111
```
program
Block

erase
Block
<table>
<thead>
<tr>
<th>APIs</th>
<th>disk</th>
<th>flash</th>
</tr>
</thead>
<tbody>
<tr>
<td>read</td>
<td>read sector</td>
<td>read page</td>
</tr>
<tr>
<td>write</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### APIs

<table>
<thead>
<tr>
<th></th>
<th>Disk</th>
<th>Flash</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Read</strong></td>
<td>read sector</td>
<td>read page</td>
</tr>
<tr>
<td><strong>Write</strong></td>
<td>write sector</td>
<td>program page (0’s) erase block (1’s)</td>
</tr>
</tbody>
</table>
Flash Hierarchy

**Plane**: 1024 to 4096 blocks
- planes accessed in parallel

**Block**: 64 to 256 pages
- unit of erase

**Page**: 2 to 8 KB
- unit of read and program
Disk vs. Flash Performance

Throughput:
- disk: ~130 MB/s (sequential)
- flash: ~200 MB/s - 550 MB/s
Disk vs. Flash Performance

**Throughput:**
- disk: ~130 MB/s (sequential)
- flash: ~200 MB/s - 550 MB/s

**Latency**
- disk: ~10 ms (one op)
- flash
  - read: 10-50 us
  - program: 200-500 us
  - erase: 2 ms
Traditional File Systems

File System

Storage Device

Traditional API:
- read sector
- write sector
Traditional File Systems

File System

Storage Device

Traditional API:
- read sector
- write sector

not same as flash
Options

1. Build/use new file systems for flash
   - JFFS, YAFFS
   - lot of work!

2. Build traditional API over flash API.
   - use FFS, LFS, whatever we want
Traditional API with Flash

read(addr):
    return flash_read(addr)

write(addr, data):
    block_copy = flash_read(block of addr)
    modify block_copy with data
    flash_erase(block of addr)
    flash_program(block of addr, block_copy)
FS wants to write 0001.

Memory:

Flash:

<table>
<thead>
<tr>
<th>block 0</th>
<th>block 1</th>
<th>block 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>00 00</td>
<td>11 11</td>
<td>00 01</td>
</tr>
<tr>
<td>00 11</td>
<td>11 11</td>
<td>11 11</td>
</tr>
<tr>
<td>11 11</td>
<td>11 11</td>
<td>11 11</td>
</tr>
<tr>
<td>00 11</td>
<td>11 11</td>
<td>11 11</td>
</tr>
</tbody>
</table>
Memory:

Block 0

Block 1

Block 2

Read all other pages in block
modify target page in memory

Memory:

Flash:

modify target page in memory
program all pages in block

Memory:

Flash:

block 0

block 1

block 2
Write Amplification

Random writes are extremely expensive!

Writing one 2KB page may cause:
- read, erase, and program of 256KB block.
Write Amplification

Random writes are extremely expensive!

Writing one 2KB page may cause:
- read, erase, and program of 256KB block.

Would FFS or LFS be better with flash?
File Systems over Flash

Copy-On-Write FS may prevent some expensive random writes.

What about wear leveling? LFS won’t do this.

What if we want to use some other FS?

Perhaps some other FS has features or APIs our applications rely on, so we must use it.
Better Solution

Add copy-on-write layer between FS and flash. Avoids RMW (read-modify-write) cycle.

Translate logical device addrs to physical addrs.

FTL: Flash Translation Layer.

Should translation use math or data structure?
Flash Translation Layer

write 1101

logical:

0 1 2 3 4 5 6 7

physical:

block 0
00 00 00 00
01 10 11 00

block 1
10 11 11 11
01 11 11 11
Flash Translation Layer

write 1101

logical:

```
0 1 2 3 4 5 6 7
```

physical:

```
block 0
00 00 00 00
01 10 11 00
```

```
block 1
10 11 11 11
01 01 11 11
```


Flash Translation Layer

write 1101

logical:

0 1 2 3 4 5 6 7

physical:

block 0

00 00 00 00
01 10 11 00

block 1

10 11 11 11
01 01 11 11

11 11
Flash Translation Layer

logical: 0 1 2 3 4 5 6 7

downward arrows:

physical: block 0

00 00 00 00
01 10 11 00

block 1

10 11 11 11
01 01 11 11
11 11
Flash Translation Layer

logical:

must eventually be garbage collected

physical:

block 0

block 1
FTL

Could be implemented as device driver or in firmware (usually the latter).

Where to store mapping? SRAM.

Physical pages can be in three states:
- valid, invalid, free
States

- free
- valid
- invalid

Transition:
- free to valid: program
- valid to invalid: erase
- invalid to free: erase
States

free → program → valid

erase

invalid

relocate or TRIM
SSD Architecture

SSD: looks like disk

(Traditional block API)

FTL
SRAM: mapping tbl
Problem: Big Mapping Table

Assume 200GB device, 2KB pages, 4-byte entries.

SRAM needed: \( \frac{200\text{GB}}{2\text{KB}} \times 4 \text{ bytes} = 400 \text{ MB} \).

Too big, SRAM is expensive!
Page Translations

logical:

0 1 2 3 4 5 6 7

physical:

block 0
00 00 00 00
01 10 11 00

block 1
10 01 11 11
01 01 11 11
11 11 11 11
2-Page Translations

logical:

physical:

block 0

block 1
Larger Mappings

Advantage: larger mappings decrease table size.

Disadvantage?
2-Page Translations

logical:

physical:

block 0

block 1
2-Page Translations

write 1011

logical:

0 1 2 3 4 5 6 7

physical:

block 0

00 00 00 00
01 10 11 00

block 1

10 01 11 11
01 01 11 11
2-Page Translations

write 1011

logical:

physical:

block 0

block 1

copy
2-Page Translations

**logical:**

0 1 2 3 4 5 6 7

**physical:**

block 0: 00 00 00 00

block 1: 10 01 10 01

write 1011
2-Page Translations

write 1011

logical:

0 1 2 3 4 5 6 7

physical:

block 0:

00 00 00 00
01 10 11 00

block 1:

10 01 10 01
01 01 11 01
Larger Mappings

Advantage: larger mappings decrease table size.

Disadvantages?
- more read-modify-write updates
- more garbage
- less flexibility for placement
Hybrid FTL

Use course-grained mapping for most (e.g., 95%) of data. Map at block level.

Use fine-grained mapping for recent data. Map at page level.
Log Blocks

Write changed pages to designated log blocks.

After blocks become full, merge changes with old data.

Eventually garbage collect old pages.
Merging

Merging technique depends on I/O pattern.

Three merge types:
- full merge
- partial merge
- switch merge
Merging

Merging technique depends on I/O pattern.

Three merge types:
- full merge
- partial merge
- switch merge
write D2

logical:

0 1 2 3 ...

physical:

A B C D

block 0

11 11 11 11 11

block 1 (log)

11 11 11 11 11

block 2

11 11 11 11 11
write D2

logical: 0 1 2 3 ...

physical:

block 0

A B C D

block 1 (log)

D2 11 11 11 11

block 2

11 11 11 11 11
logical: 0 1 2 3 ...

physical: A B C D D2 11 11 11 11 ...

block 0  block 1 (log)  block 2
eventually, we need to get rid of red arrows, as these represent expensive mappings
logical: 0 1 2 3 ...

physical:

block 0: A B C D
block 1 (log): D2 11 11 11 11
logical: 0 1 2 3 ...

physical:

block 0: A B C D
block 1 (log): D2 11 11 11 11
block 2: A B C D2
logical: 0 1 2 3 ...

physical: A B C D D2
               11 11 11 11
               A B C D2

block 0         block 1 (log)       block 2

A block 0 contains A, B, C, D, and D2. Block 1 contains 11, 11, 11, 11. Block 2 is labeled as garbage.
Merging

Merging technique depends on I/O pattern.

Three merge types:
- full merge
- partial merge
- switch merge
logical: 0 1 2 3 ...

physical: A B C D block 0 block 1 (log) block 2
write D2

logical: 0 1 2 3 ...

physical: A B C D

block 0

block 1 (log)

block 2
write D2

logical: 0 1 2 3 ...


block 0
logical: 0 1 2 3 ...

physical:
block 0: A B C D
block 1 (log): A B C D2
block 2: 11 11 11 11 11
logical: 0 1 2 3

physical: A B C D block 0
          A B C D2 block 1
          11 11 11 11 11 block 2
Merging

Merging technique depends on I/O pattern.

Three merge types:
- full merge
- partial merge
- switch merge
logical: 0 1 2 3 ...

physical: A B C D

block 0

block 1 (log)

block 2
write A2

logical: 0 1 2 3 ...

physical: A B C D

block 0

block 1 (log)

block 2
write A2

logical:

0 1 2 3 ...

physical:

write B2

logical:

physical:
write C2

logical: 0 1 2 3 ...

physical:

block 0 A B C D
block 1 (log) A2 B2 C2 11 11
block 2 11 11 11 11 11
write D2

logical: 0 1 2 3 ...

physical: A B C D A2 B2 C2 D2

block 0 block 1 (log) block 2
Merging

Merging technique depends on I/O pattern.

Three merge types:
- full merge
- partial merge
- switch merge
Summary

Flash is much faster than disk, but...

It is more expensive.

It’s not a drop-in replacement beneath an FS without a complex layer for emulating hard disk API.