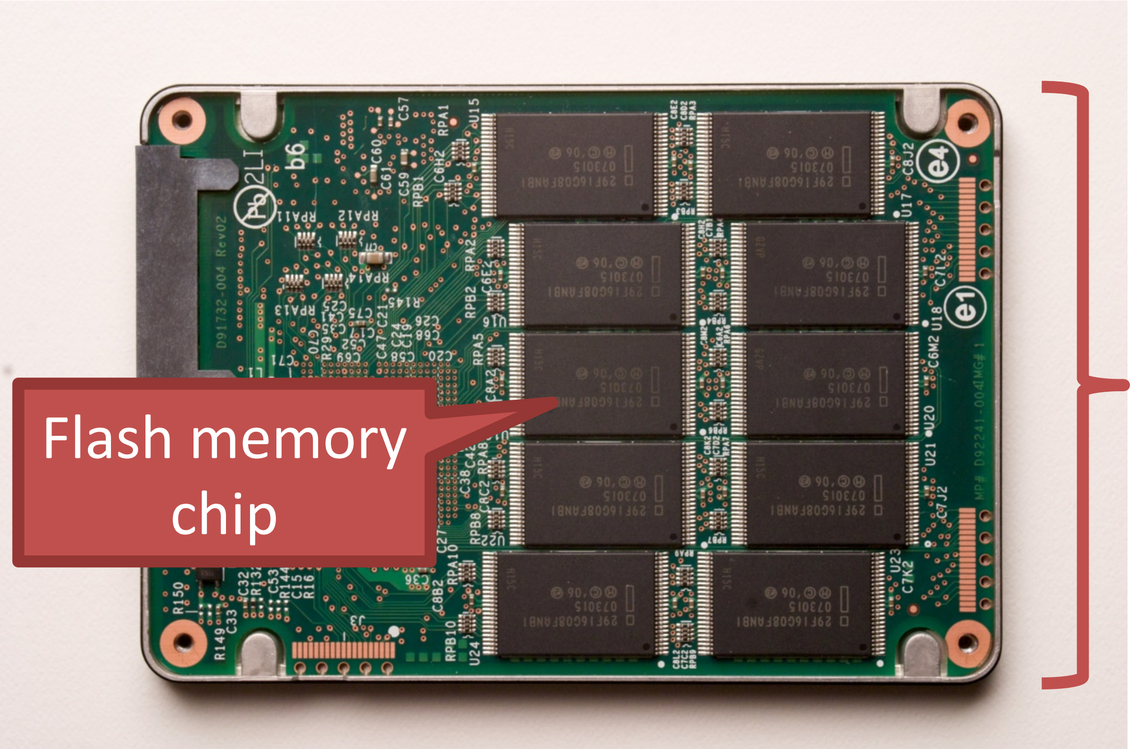
Beyond Spinning Disks

- Hard drives have been around since 1956
  - The cheapest way to store large amounts of data
  - Sizes are still increasing rapidly
- However, hard drives are typically the slowest component in most computers
  - CPU and RAM operate at GHz
  - PCI-X and Ethernet are GB/s
- Hard drives are not suitable for mobile devices
  - Fragile mechanical components can break
  - The disk motor is extremely power hungry

Solid State Drives

- NAND flash memory-based drives
  - High voltage is able to change the configuration of a floating-gate transistor
  - State of the transistor interpreted as binary data



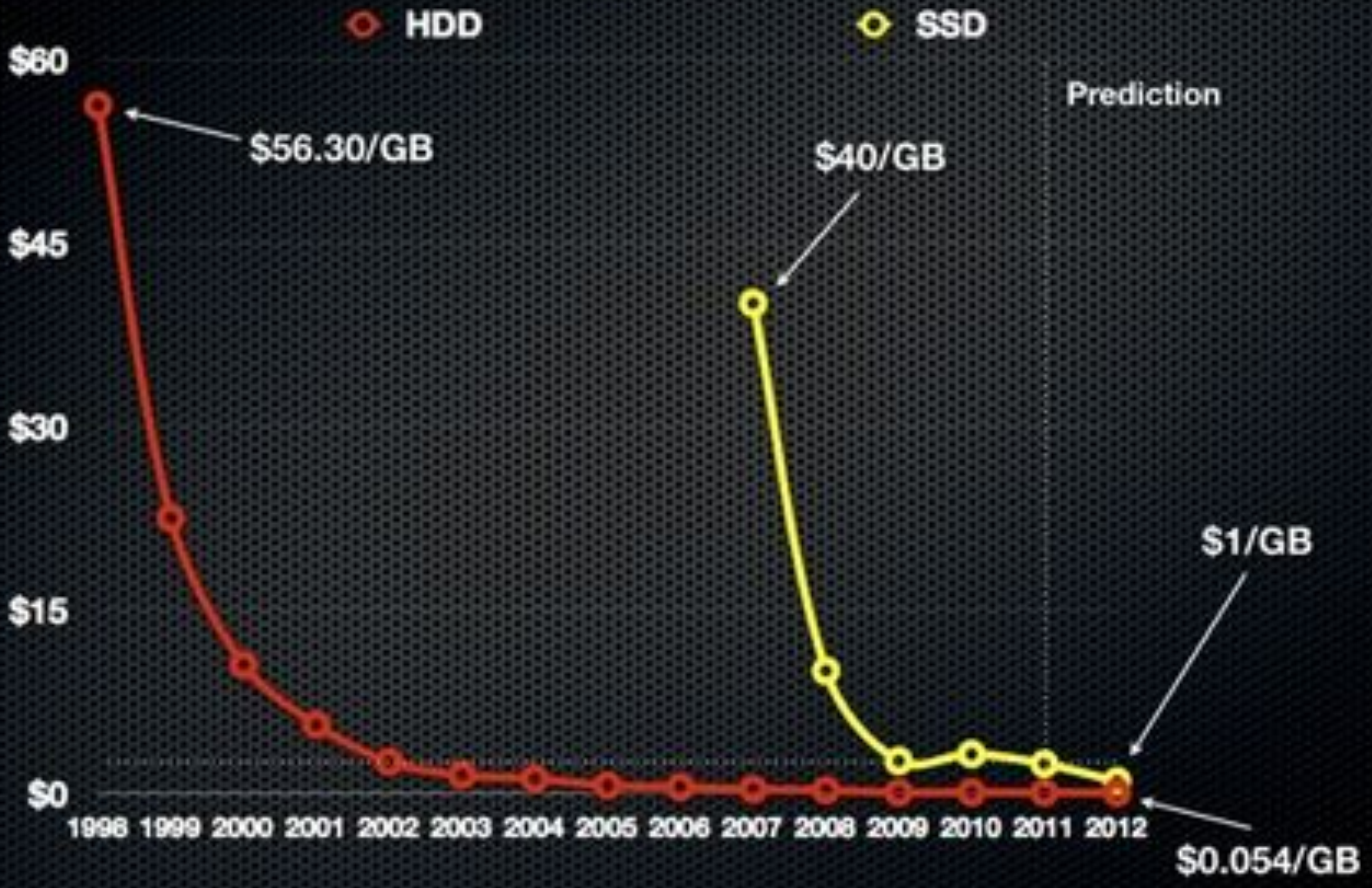
Flash memory chip

Data is striped across all chips

Advantages of SSDs

- More resilient against physical damage
  - No sensitive read head or moving parts
  - Immune to changes in temperature
- Greatly reduced power consumption
  - No mechanical, moving parts
- Much faster than hard drives
  - >500 MB/s vs ~200 MB/s for hard drives
  - No penalty for random access
    - Each flash cell can be addressed directly
    - No need to rotate or seek
  - Extremely high throughput
    - Although each flash chip is slow, they are RAIDed

Average HDD and SSD prices in USD per gigabyte



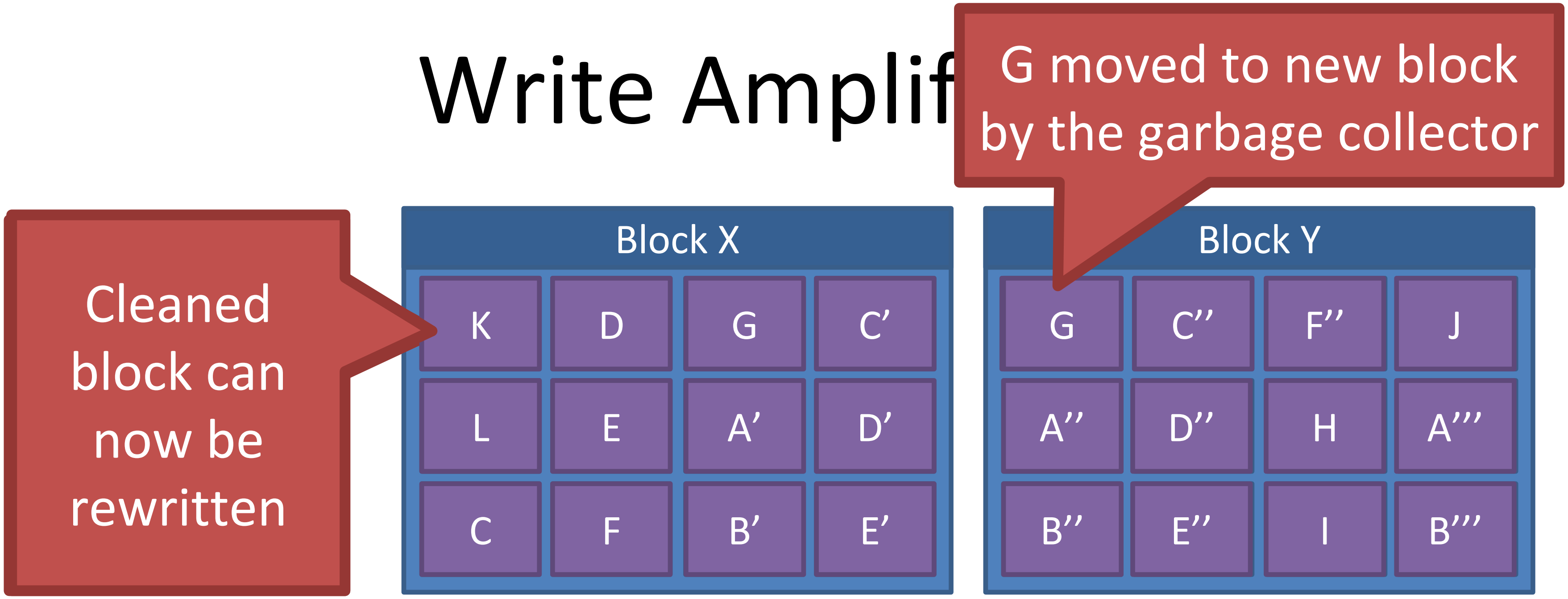
Data sources: Mkomo.com, Gartner, and Pingdom (December 2011)

www.pingdom.com

Challenges with Flash

- Flash memory is written in pages, but erased in blocks
  - Pages: 4 – 16 KB, Blocks: 128 – 256 KB
  - Thus, flash memory can become fragmented
  - Leads to the write amplification problem
- Flash memory can only be written a fixed number of times
  - Typically 3000 – 5000 cycles for MLC
  - SSDs use wear leveling to evenly distribute writes across all flash cells

Write Amplif



G moved to new block by the garbage collector

Cleaned block can now be rewritten

Block X			
K	D	G	C’
L	E	A’	D’
C	F	B’	E’

Block Y			
G	C’’	F’’	J
A’’	D’’	H	A’’’
B’’	E’’	I	B’’’

- Once all pages have been written, valid pages must be consolidated to free up space
- Write amplification: a write triggers garbage collection/compaction
  - One or more blocks must be read, erased, and rewritten before the write can proceed

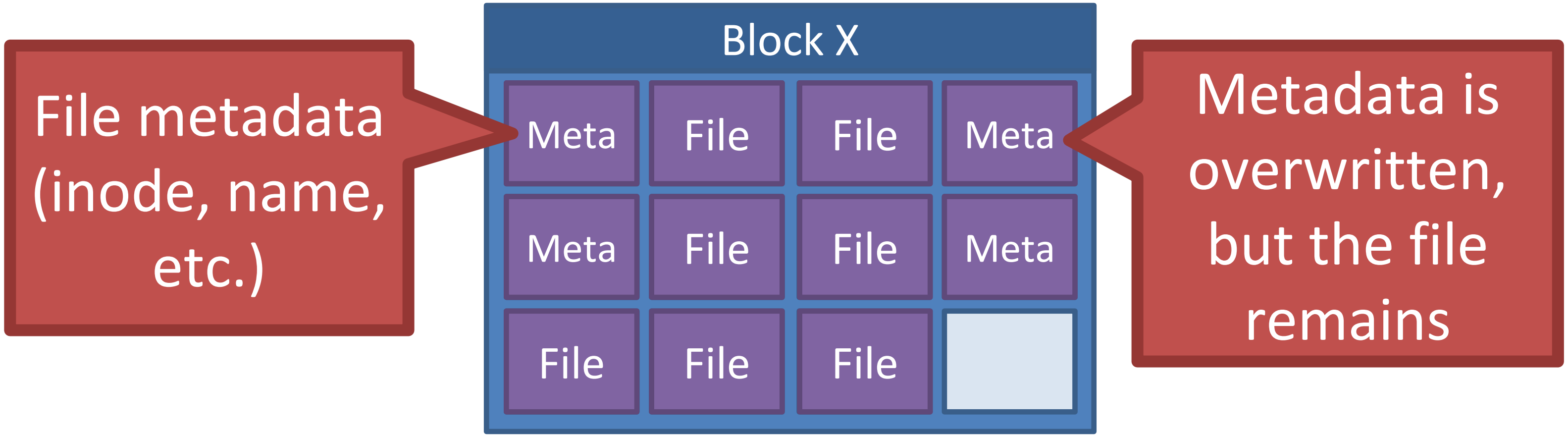
Garbage Collection

- Garbage collection (GC) is vital for the performance of SSDs
- Older SSDs had fast writes up until all pages were written once
  - Even if the drive has lots of “free space,” each write is amplified, thus reducing performance
- Many SSDs over-provision to help the GC
  - 240 GB SSDs actually have 256 GB of memory
- Modern SSDs implement background GC
  - However, this doesn’t always work correctly

The Ambiguity of Delete

- Goal: the SSD wants to perform background GC
  - But this assumes the SSD knows which pages are invalid
- Problem: most file systems don’t actually delete data
  - On Linux, the “delete” function is unlink()
  - Removes the file meta-data, but not the file itself

Delete Example



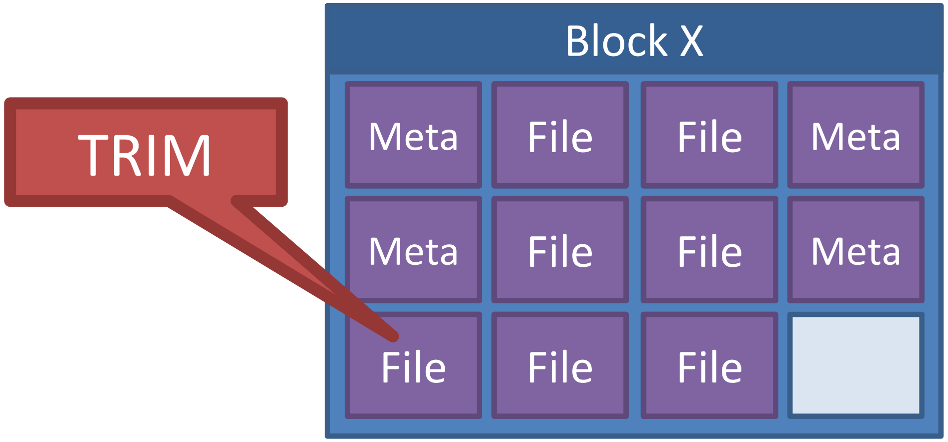
File metadata (inode, name, etc.)
Block X
Meta
File
File
Meta
Meta
File
File
Meta
File
File
File
Metadata is overwritten, but the file remains

1. File is written to SSD
2. File is deleted
3. The GC executes
   - 9 pages look valid to the SSD
   - The OS knows only 2 pages are valid

- Lack of explicit delete means the GC wastes effort copying useless pages
- Hard drives are not GCed, so this was never a problem

TRIM

- New SATA command TRIM (SCSI – UNMAP)
  - Allows the OS to tell the SSD that specific LBAs are invalid, may be GCed



- OS support for TRIM
  - Win 7, OSX Snow Leopard, Linux 2.6.33, Android 4.3
- Must be supported by the SSD firmware

Wear Leveling

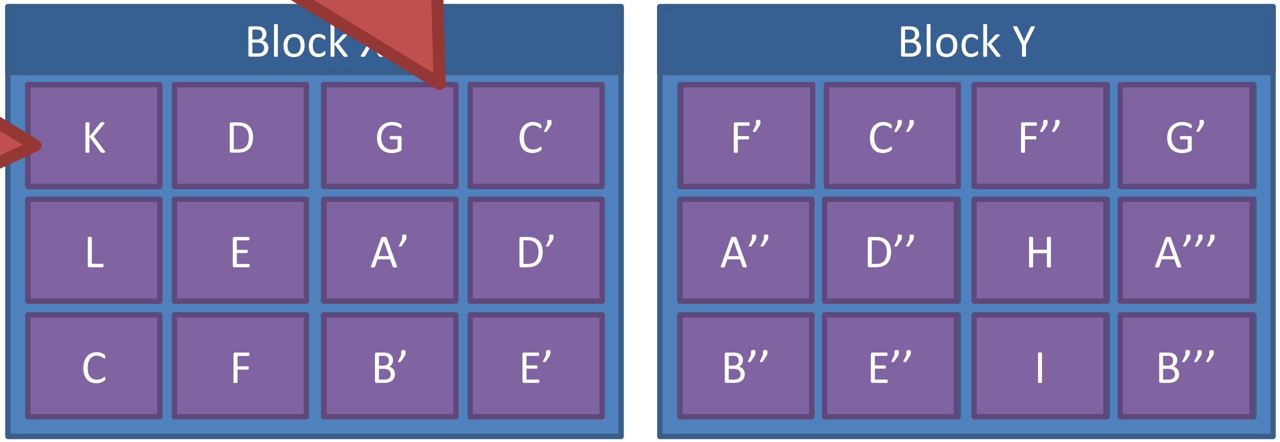
- Recall: each flash cell wears out after several thousand writes
- SSDs use wear leveling to spread writes across all cells
  - Typical consumer SSDs should last ~5 years

W xamples

If the GC runs now, page G must be copied

Dynamic Wear Leveling

Wait as long as possible before garbage collecting



Block X

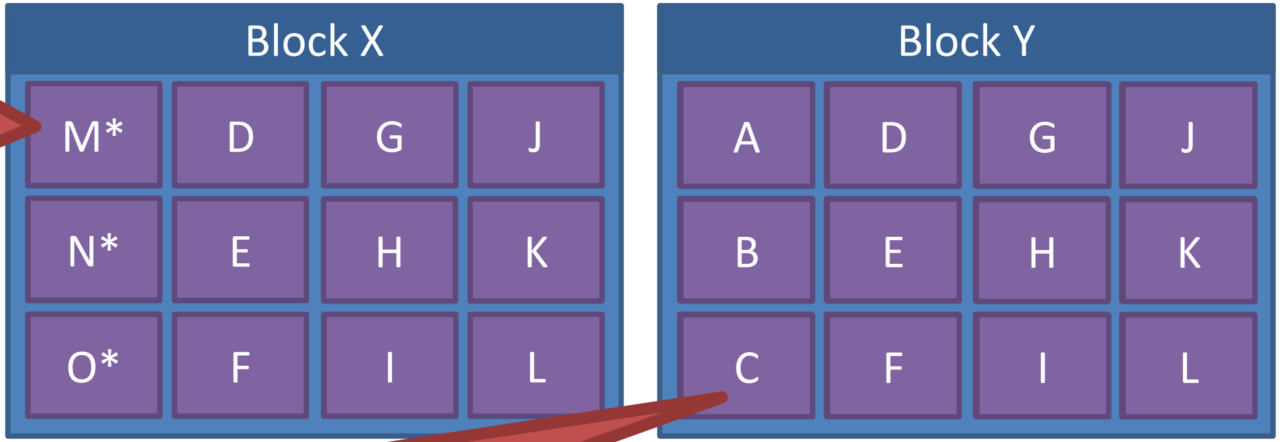
K	D	G	C’
L	E	A’	D’
C	F	B’	E’

Block Y

F’	C’’	F’’	G’
A’’	D’’	H	A’’’
B’’	E’’	I	B’’’

Static Wear Leveling

Blocks with long lived data receive less wear



Block X

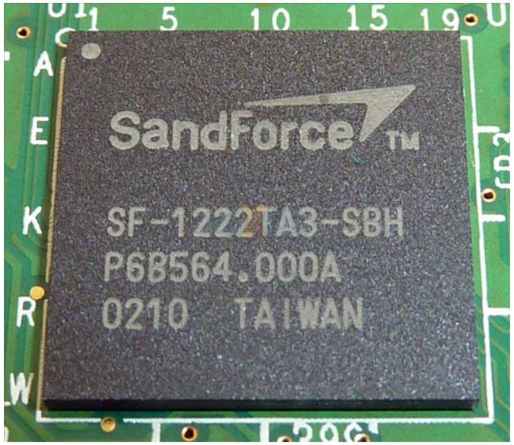
M*	D	G	J
N*	E	H	K
O*	F	I	L

Block Y

A	D	G	J
B	E	H	K
C	F	I	L

SSD controller periodically swap long lived data to different blocks

SSD Controllers



- SSDs are extremely complicated internally
- All operations handled by the SSD controller
  - Maps LBAs to physical pages
  - Keeps track of free pages, controls the GC
  - May implement background GC
  - Performs wear leveling via data rotation
- Controller performance is crucial for overall SSD performance

Flavors of NAND Flash Memory

Multi-Level Cell (MLC)

- Multiple bits per flash cell
  - For two-level: 00, 01, 10, 11
  - 2, 3, and 4-bit MLC is available
- Higher capacity and cheaper than SLC flash
- Lower throughput due to the need for error correction
- 3000 – 5000 write cycels
- Consumes more power

Consumer-grade drives

Single-Level Cell (SLC)

- One bit per flash cell
  - 0 or 1
- Lower capacity and more expensive than MLC flash
- Higher throughput than MLC
- 10000 – 100000 write cycles

Expensive, enterprise drives