CS310 Operating Systems

Lecture 41: Solid State Drive (SSD) – Flash Drive

Ravi Mittal

IIT Goa

References

- CS162, Operating Systems and Systems Programming, University of California, Berkeley
- Various sources on the Internet

Reading

- CS162, Operating Systems and Systems Programming, University of California, Berkeley
- Book: Operating System Concepts, 10th Edition, by Silberschatz, Galvin, and Gagne

Lecture Contents

- Flash Storage – Introduction
- Flash Operation
- Flash Architecture

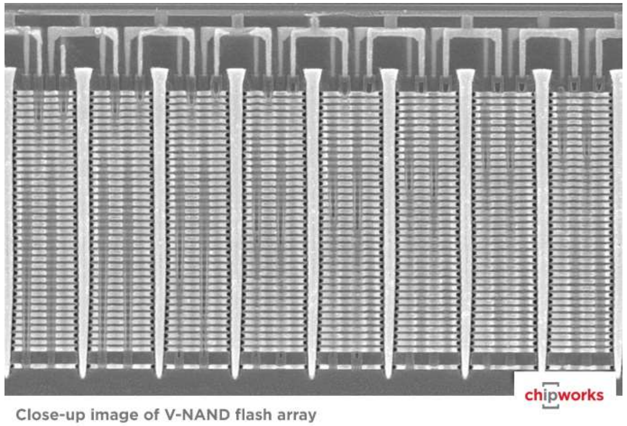
Last Class

Storage Technologies

Magnetic Disks

- Store on magnetic medium
- Electromechanical access

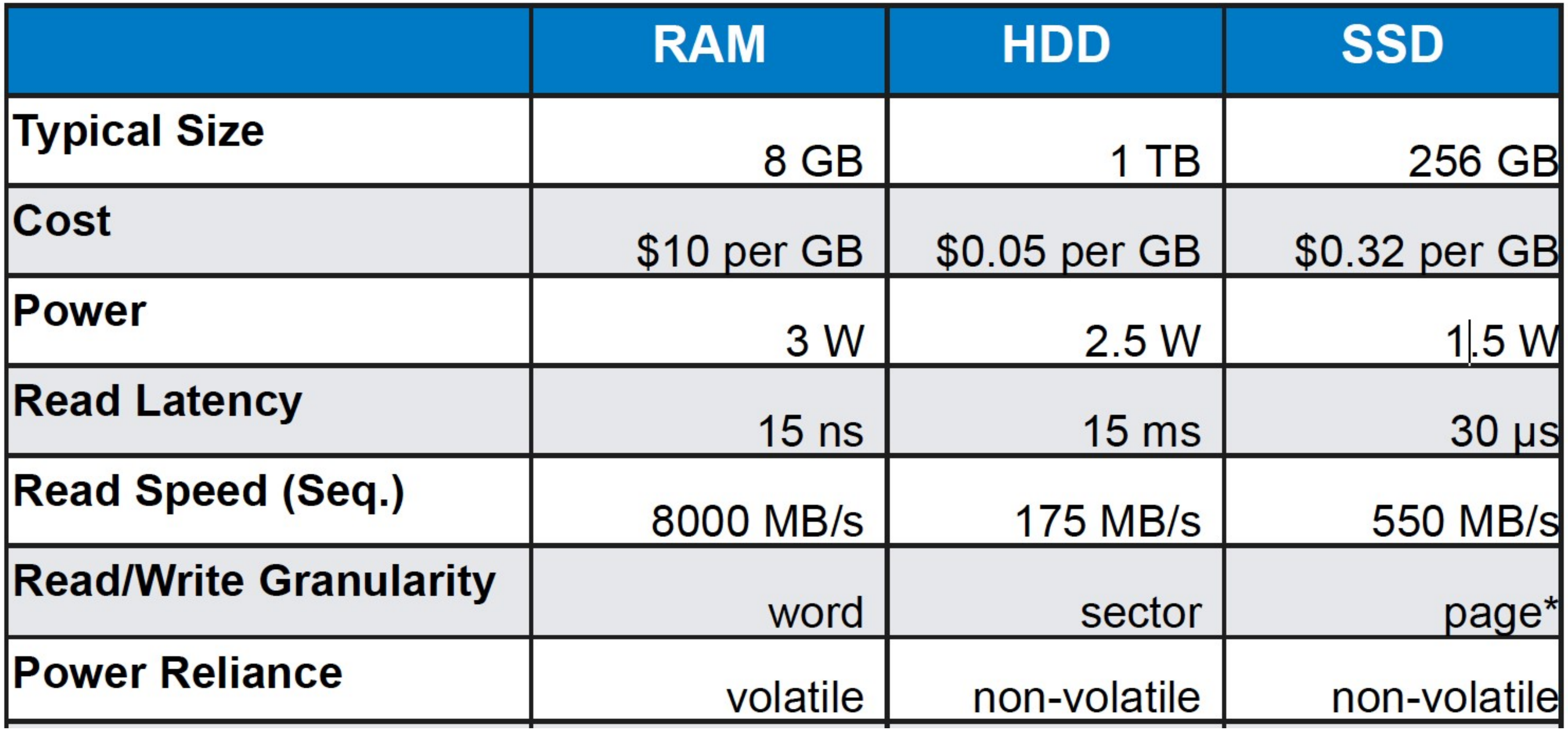
Nonvolatile (Flash) Memory



Close-up image of V-NAND flash array

- Store as persistent charge
- Implemented with 3-D structure
  - 100+ levels of cells
  - 3 bits data per cell

RAM vs Hard Disk vs SSD - 2018



	RAM	HDD	SSD
Typical Size	8 GB	1 TB	256 GB
Cost	$10 per GB	$0.05 per GB	$0.32 per GB
Power	3 W	2.5 W	1.5 W
Read Latency	15 ns	15 ms	30 µs
Read Speed (Seq.)	8000 MB/s	175 MB/s	550 MB/s
Read/Write Granularity	word	sector	page*
Power Reliance	volatile	non-volatile	non-volatile

In SSD Each cell has limited program/erase lifetime (thousands, for modern devices) – Cells become slowly less reliable

Popular Storage Devices

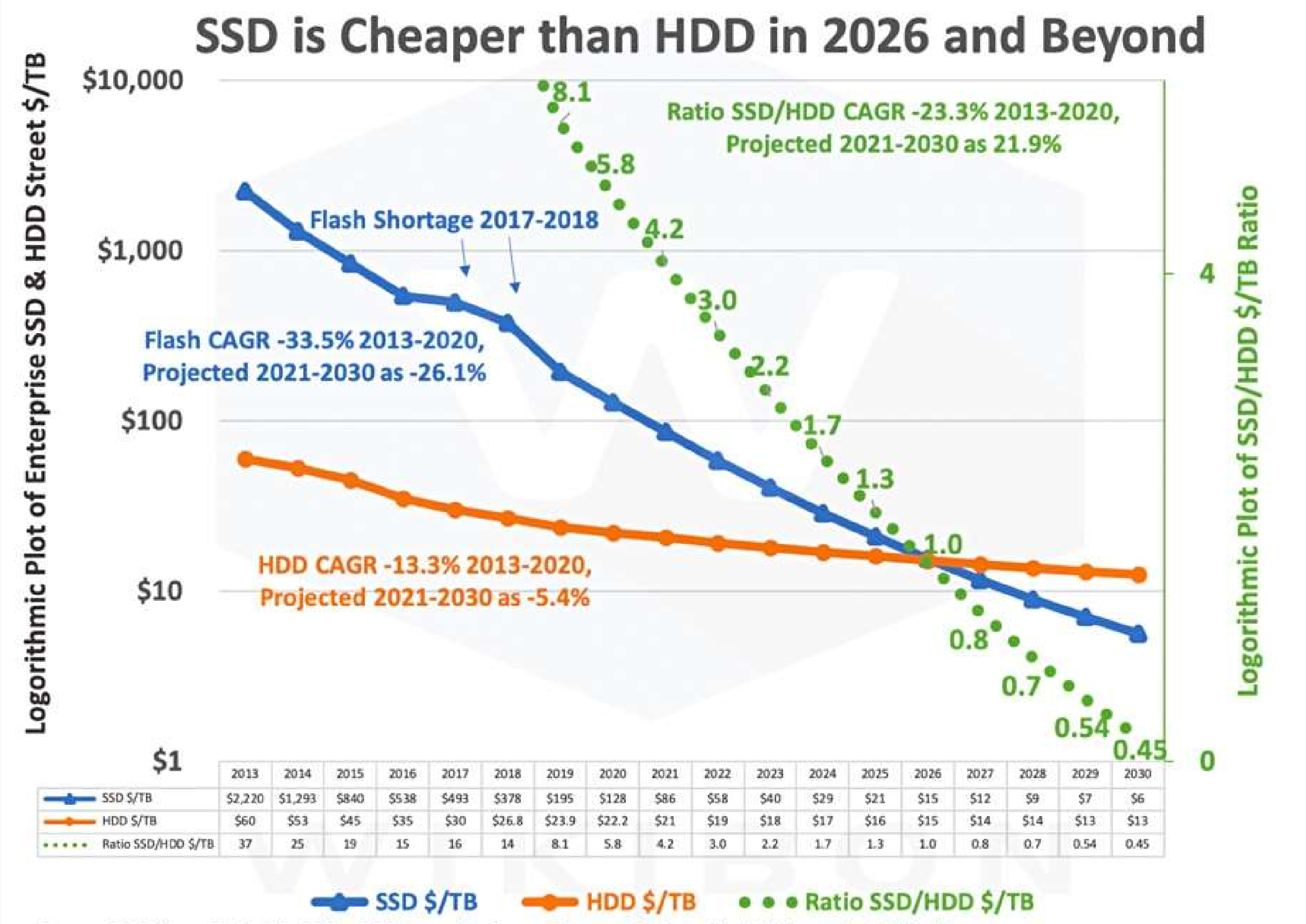
Magnetic Disks

- Rarely becomes corrupted
- Traditionally: large capacity at low cost
- Block level random access
- Slow performance for random access
- Better performance for sequential access

Flash Memory

- Rarely becomes corrupted
- Increasingly larger and cheaper
- Block level random access
- Good performance for reads, worse for random writes
- Have to erase data in large blocks
- Challenge: Wear Levelling

Emergence of SSDs



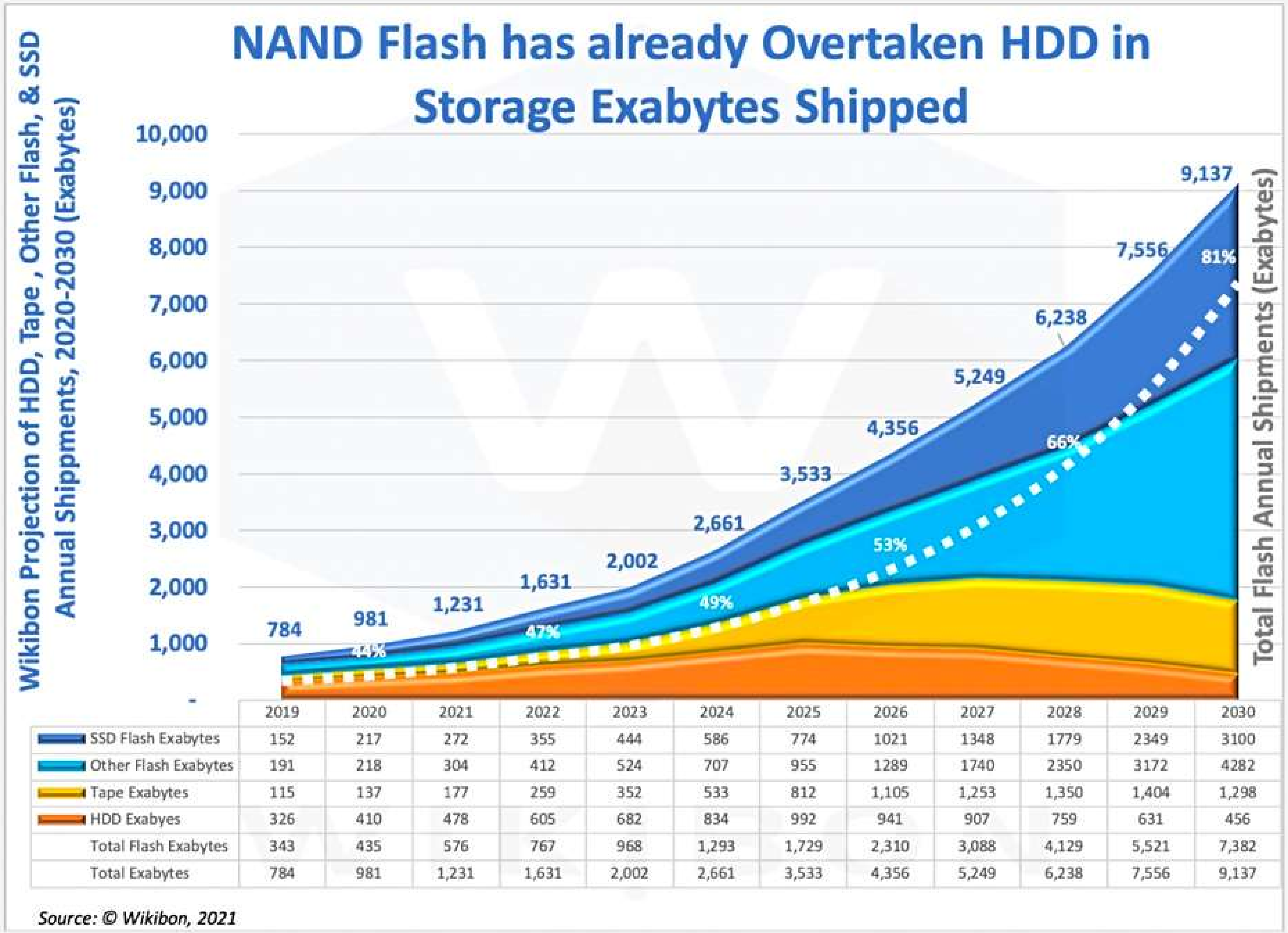
SSD is Cheaper than HDD in 2026 and Beyond

	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
SSD $/TB	$2,220	$1,293	$840	$538	$493	$378	$195	$128	$86	$58	$40	$29	$21	$15	$12	$9	$7	$6
HDD $/TB	$60	$53	$45	$35	$30	$26.8	$23.9	$22.2	$21	$19	$18	$17	$16	$15	$14	$14	$13	$13
Ratio SSD/HDD $/TB	37	25	19	15	16	14	8.1	5.8	4.2	3.0	2.2	1.7	1.3	1.0	0.8	0.7	0.54	0.45

Source: © Wikibon, 2021. The 2021-2030 Enterprise Street Prices are Projected by Wikibon using Wright's Law

Figure 4 - SSD/HDD Pricing Ratio 2013 - 2030

Source: © Wikibon, 2021.



	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
SSD Flash Exabytes	152	217	272	355	444	586	774	1021	1348	1779	2349	3100
Other Flash Exabytes	191	218	304	412	524	707	955	1289	1740	2350	3172	4282
Tape Exabytes	115	137	177	259	352	533	812	1,105	1,253	1,350	1,404	1,298
HDD Exabyes	326	410	478	605	682	834	992	941	907	759	631	456
Total Flash Exabytes	343	435	576	767	968	1,293	1,729	2,310	3,088	4,129	5,521	7,382
Total Exabytes	784	981	1,231	1,631	2,002	2,661	3,533	4,356	5,249	6,238	7,556	9,137

Source: © Wikibon, 2021

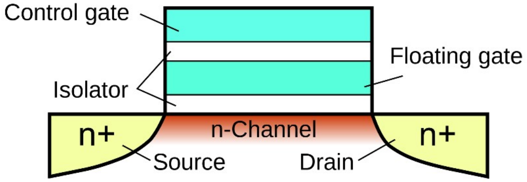
Figure 9 – Exabyte Storage Shipments Split by SSD, Other Flash, HDDs, and Tape

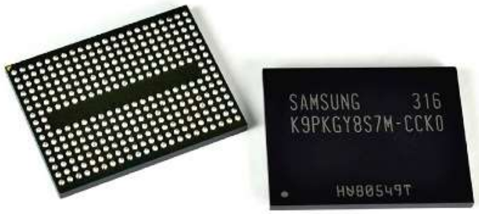
Source: © Wikibon, 2021.

Flash Storage - Introduction

Flash Storage

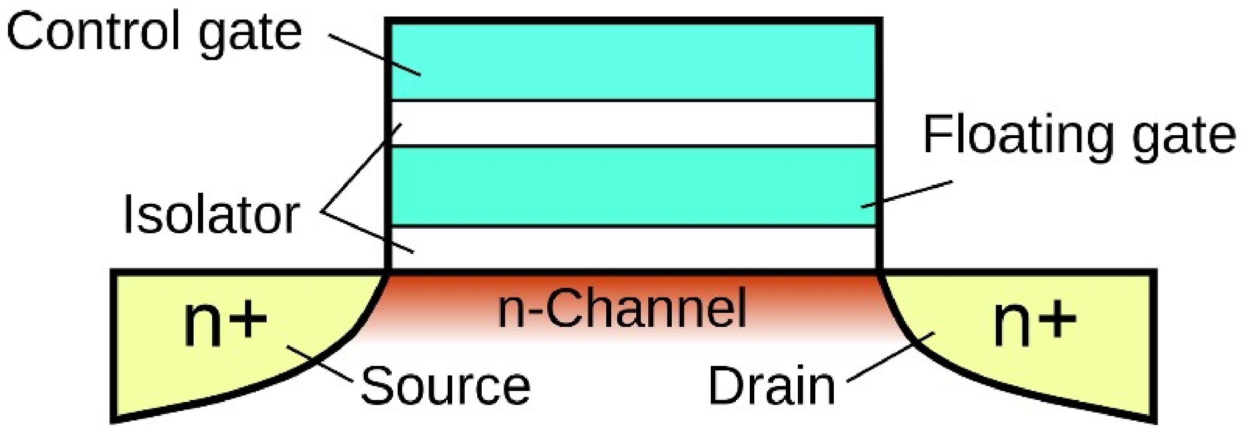
- Most prominent solid state storage technology
- NAND- and NOR- flash types available
    - NOR-flash can be byte-addressed, expensive
    - NAND-flash is page addressed, cheap
    - Except in very special circumstances, all flash-storage we see are NAND-flash
    - SD Cards, USB Drives, SSDs are based on NAND memory
    - NAND: Each cell can not be written and deleted independently
    - NOR: each cell can be handled independently



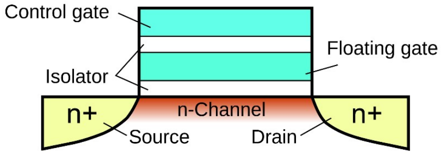


The Flash Cell

- Flash cells store data in floating gate by charging it at high voltage
- Encode bit by trapping electrons into a cell



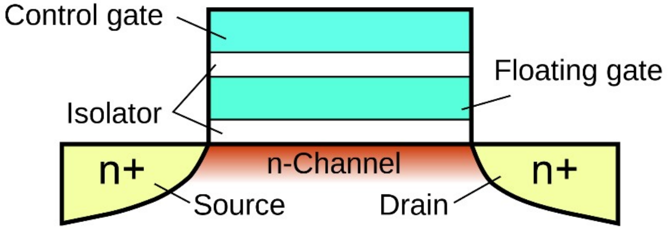
The Flash Cell

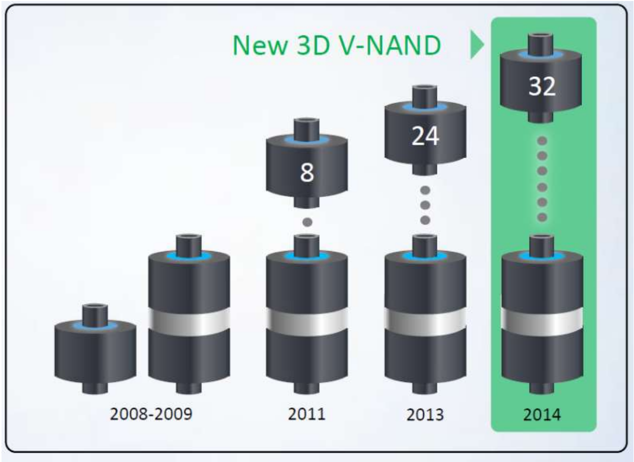


- Single-level cells (SLC)
  - Single bit stored within a transistor
  - faster, more lasting (100K writes before wear out)
- Multi-level cells (MLC)
  - Many bits can be stored in a cell by differentiating between the amount of charge in the cell
  - It can store 2, 3, even 4 bits
  - It cheaper to manufacture
  - It wears out faster (1k to 10K writes)
  - It is more fragile (stored value can be disturbed by accesses to nearby cells)

3D NAND-Flash

- 3D NAND is a type of non-volatile flash memory in which the memory cells are stacked vertically in multiple layers
  - Creates larger storage capacity
  - Smaller footprint
  - Shorter overall connections for each memory cell
  - Lower cost per byte compared to 2D NAND





SSD Storage Hierarchy

Flash Chip
Several banks that can be accessed in parallel

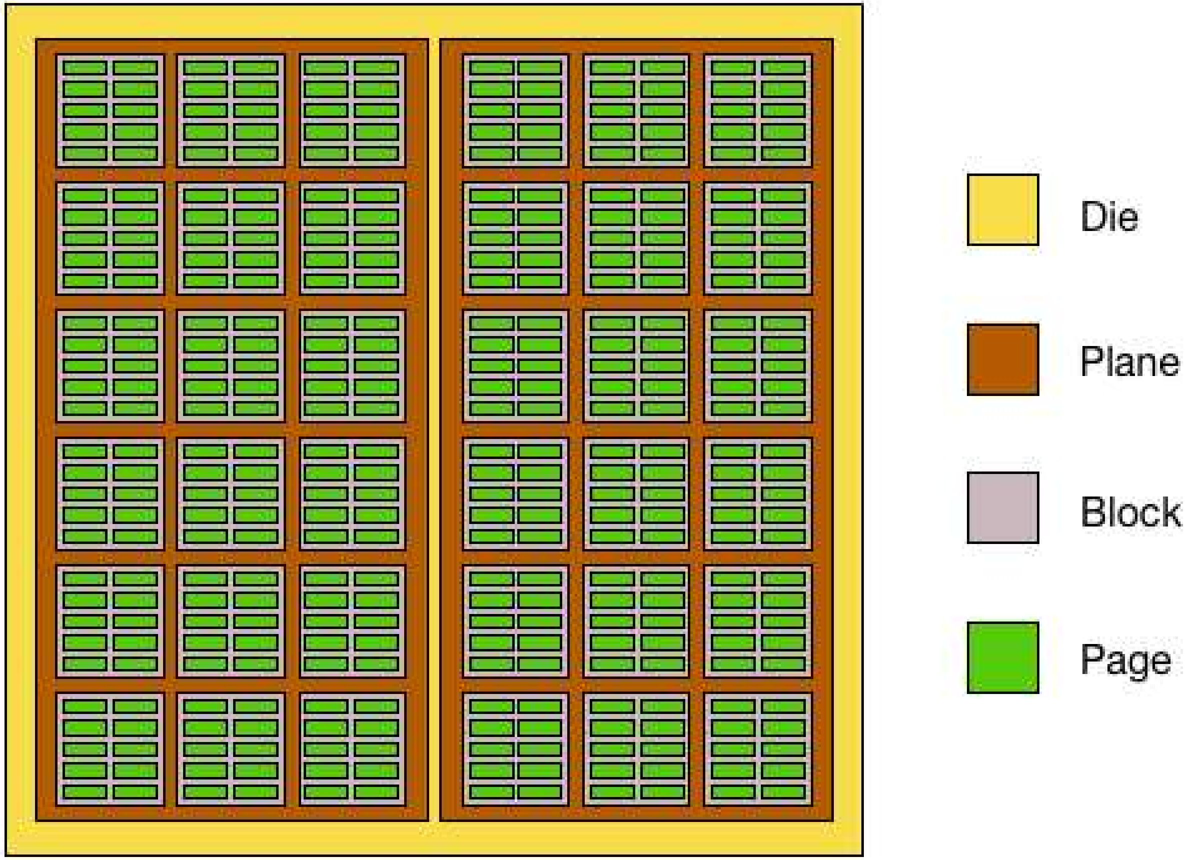
Plane/Bank
Many blocks (Several Ks)

Block
64 to 256 pages

Page
2 to 8 KB

Cell
1 to 4 bits

NAND Flash Die Layout



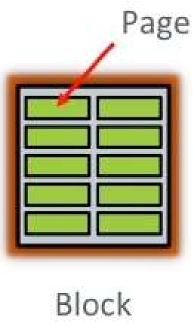
(image courtesy of AnandTech)

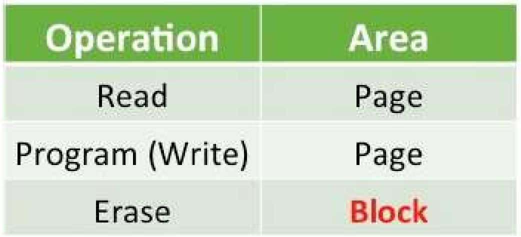
Die, Planes, Blocks, Pages

- Each package (chip) contains one or more **dies** (for example one, two, or four)
  - The die is the smallest unit that can independently execute commands or report status
- Each die contains one or more **planes** (usually one or two)
  - Identical, concurrent operations can take place on each plane, although with some restrictions
- Each plane contains a number of **blocks**
  - Block is the smallest unit that can be erased
- Each block contains a number of **pages**
  - Page is the smallest unit that can be programmed (i.e. written to)

Die, Planes, Blocks, Pages

- Write take place on a page
  - typically 8-16KB in size
- Erase operations take place to a block
  - 4-8MB in size
- A block needs to be erased before it can be programmed again





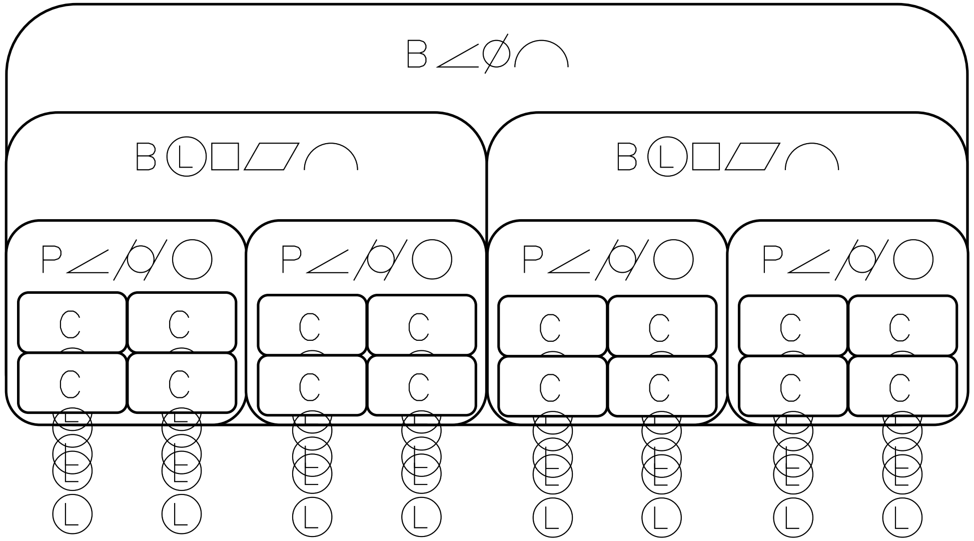
Operation	Area
Read	Page
Program (Write)	Page
Erase	Block

Interesting way of working!

- There is no update operation for flash
- No undo or rewind mechanism for changing what is currently in place
- Just the **erase** operation

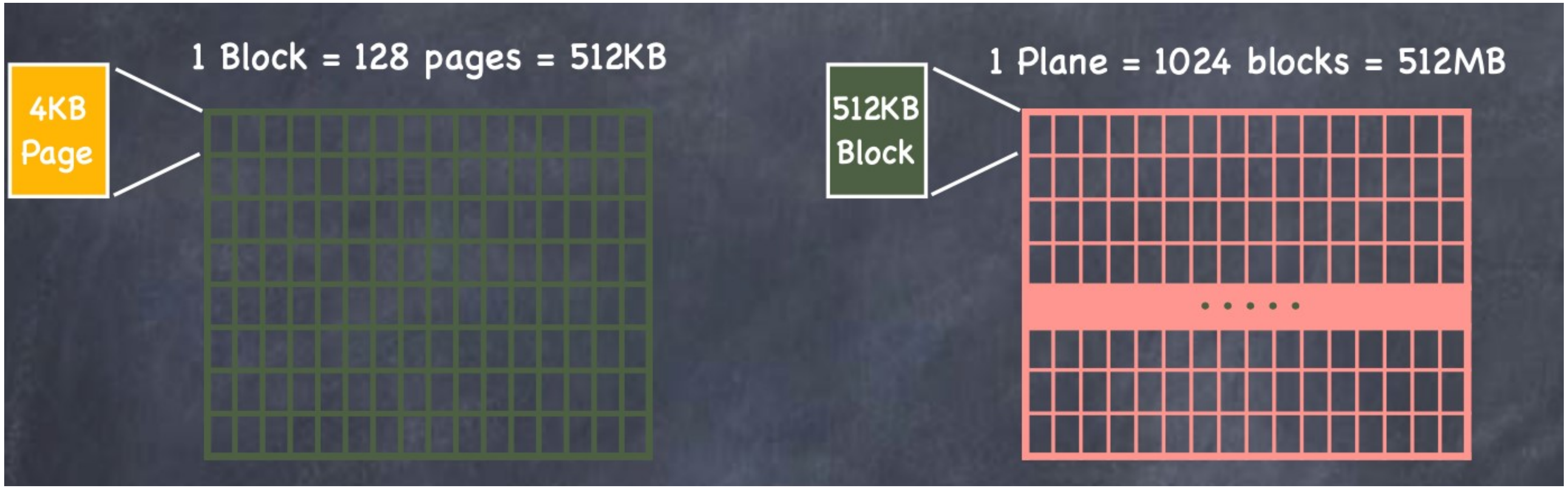
- An erase operation on a flash chip clears the data from all pages in the block
- If some of the other pages contain active data (stuff you want to keep)
  - Either have to copy it elsewhere first
  - Or don’t do erase

Of banks, blocks, cells



- Flash chips organized in **banks**
  - Banks can be accessed in parallel
- **Blocks**: 128 KB/256KB
  - (64 to 258 pages)
- **Pages**: Few KB
- **Cells:** 1 to 4 bits
- Distinction between blocks and pages important in operations!

Example: NAND Flash Units



SSD Operations

- Erase
    - Before a block can be written it is erased
        - Set to logical 1
            - Under the hood: erase sets all bits to 1, write can only change some to 0
        - Operation takes several milliseconds – high latency
- Write a page
    - Tens of microseconds to hundreds of microseconds
- Reading a page
    - Read takes 10s of micro seconds

Flash Drive - Data

- Flash drive specs 4 KB page
  - 3ms to erase erasure block
  - 512KB erasure block
  - 50µs read page/write page

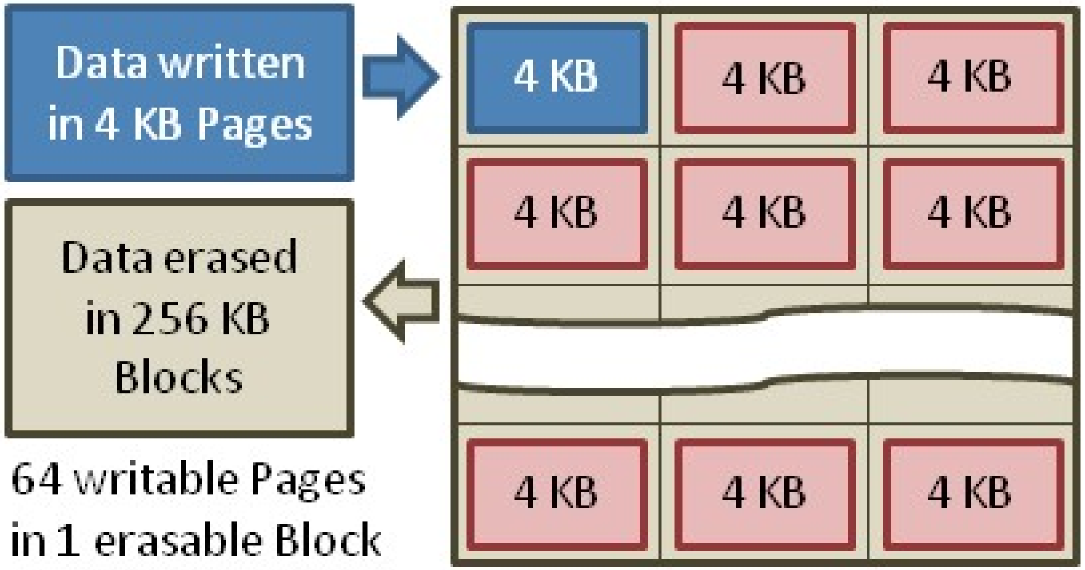
Read Block, Erase Block, Write Block

- How long to naively read/erase/and write each page?
  - 128 x (50 x 10-3) + 3 + 128 x (50 x 10-3) = 15.8ms per write

Flash Operations: Erase, Write, Read

SSD Architecture – Writes (I)

- Writing data is complex! (~200μs – 1.7ms)
- Write be done only on empty pages in a block
- Erasing a block takes ~1.5ms
- Controller maintains pool of empty blocks by coalescing used pages (read, erase, write), also reserves some % of capacity

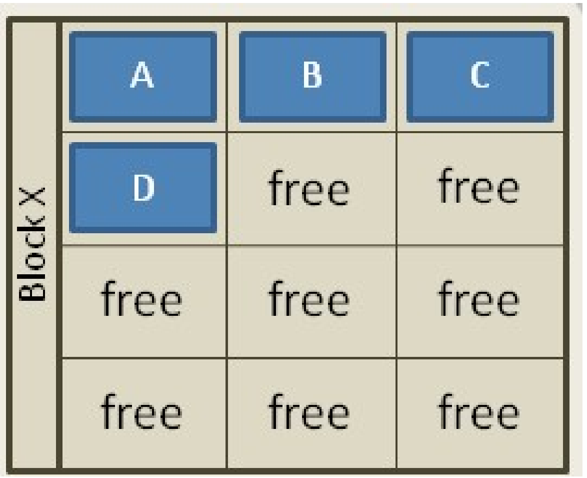

Data written in 4 KB Pages
Data erased in 256 KB Blocks
64 writable Pages in 1 erasable Block
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB
4 KB

Typical NAND Flash Pages and Blocks

https://en.wikipedia.org/wiki/Solid-state_drive

SSD Architecture – Writes (II)

- Write A, B, C, D



Block X

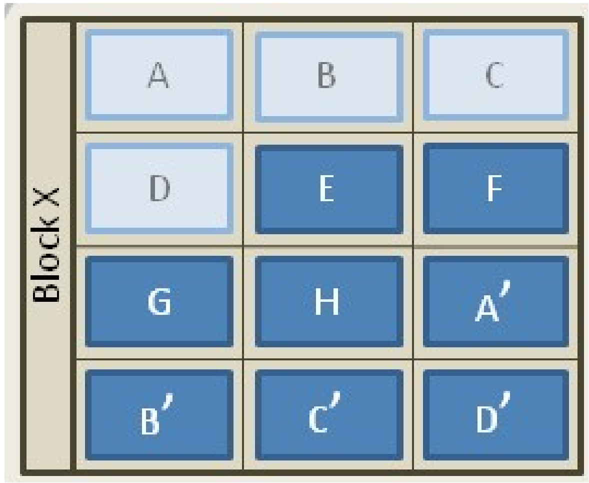
A	B	C
D	free	free
free	free	free
free	free	free

https://en.wikipedia.org/wiki/Solid-state_drive

SSD Architecture – Writes (II)

- Write A, B, C, D

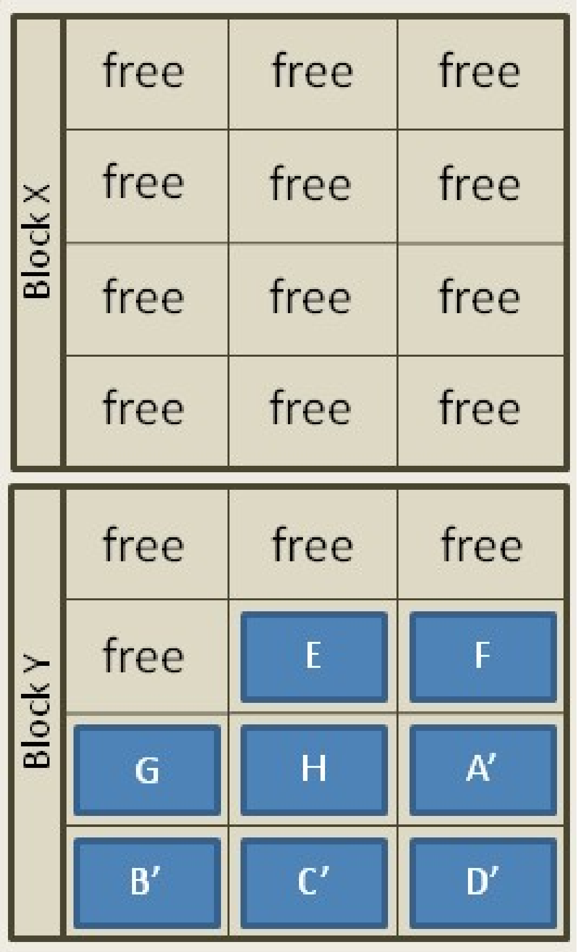
- Write E, F, G, H and A’, B’, C’, D’
  - Record A, B, C, D as obsolete



https://en.wikipedia.org/wiki/Solid-state_drive

SSD Architecture – Writes (II)

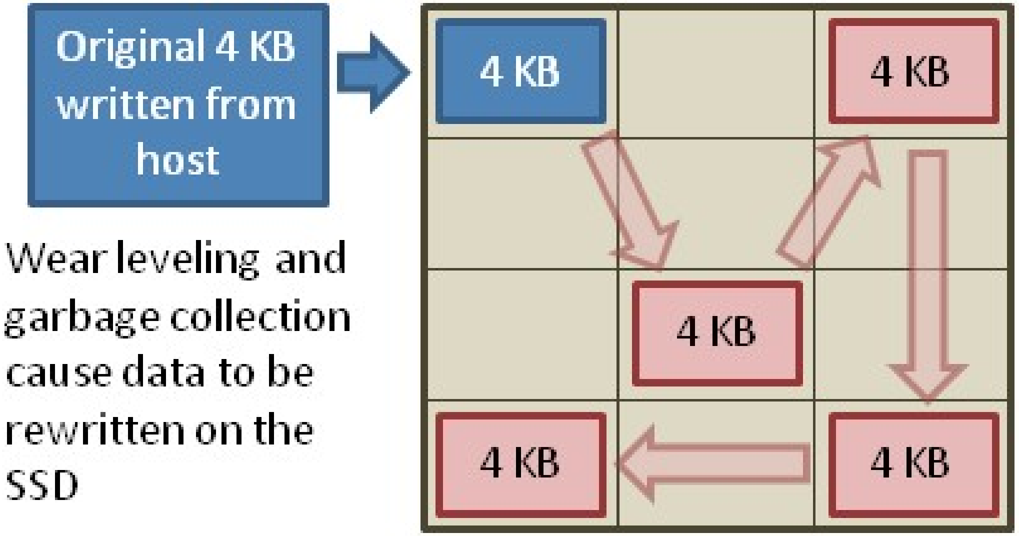
- Write A, B, C, D
- Write E, F, G, H and A’, B’, C’, D’
  - Record A, B, C, D as obsolete
- Controller *garbage collects* obsolete pages by copying valid pages to new (erased) block



Block X	free	free	free
	free	free	free
	free	free	free
	free	free	free
Block Y	free	free	free
	free	E	F
	G	H	A’
	B’	C’	D’

SSD Architecture – Writes (III)

- Write and erase cycles require "high" voltage
  - Damages memory cells, limits SSD lifespan
  - Controller uses ECC, performs wear leveling



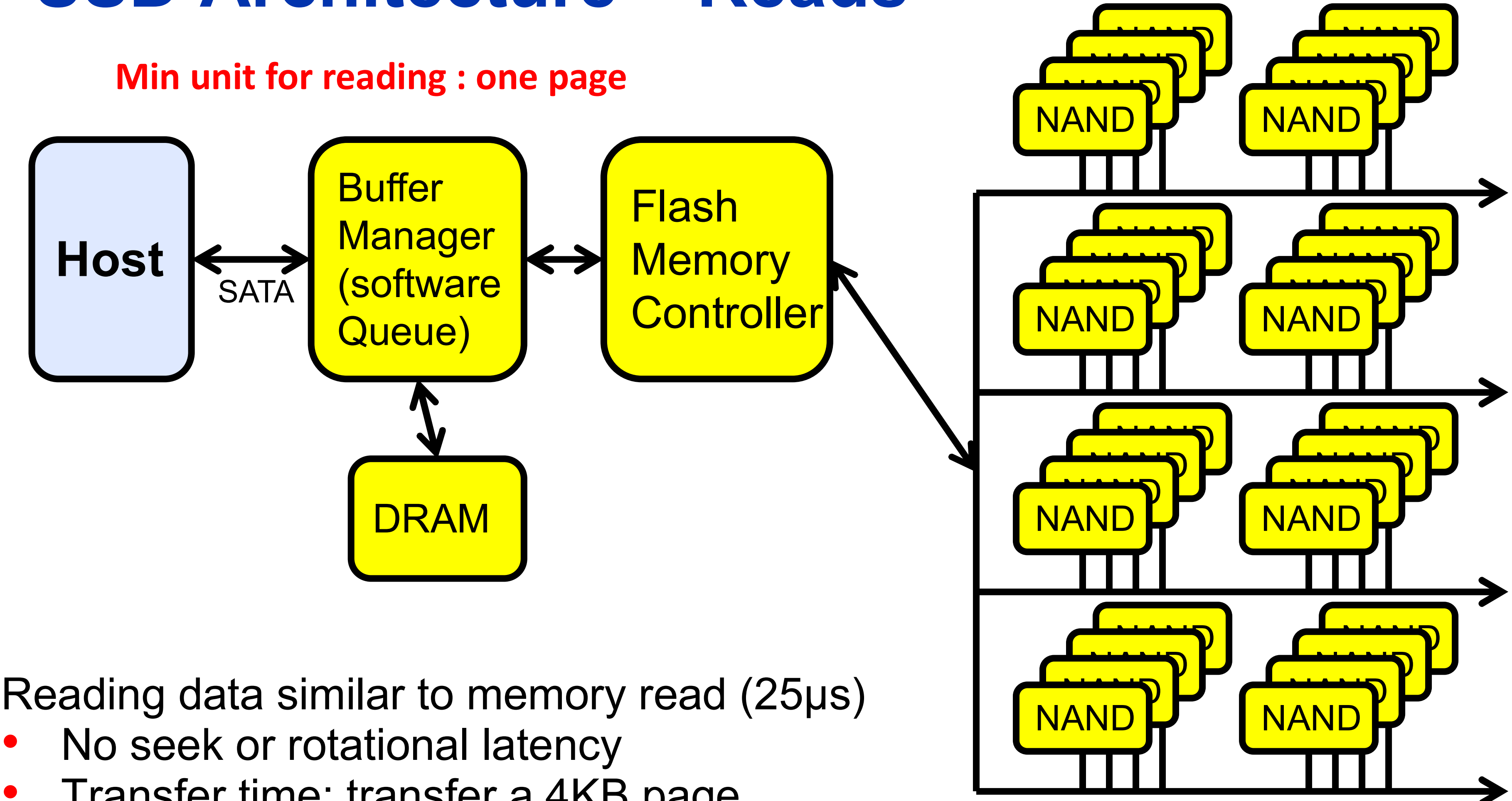
Wear leveling and garbage collection cause data to be rewritten on the SSD

- Result is very workload dependent performance
  - Latency = Queuing Time + Controller time (Find Free Block) + Xfer Time
  - Highest BW: Seq. OR Random writes (limited by empty pages)

Rule of thumb: writes 10x more expensive than reads, and erases 10x more expensive than writes

SSD Architecture – Reads

Min unit for reading : one page



Reading data similar to memory read (25μs)

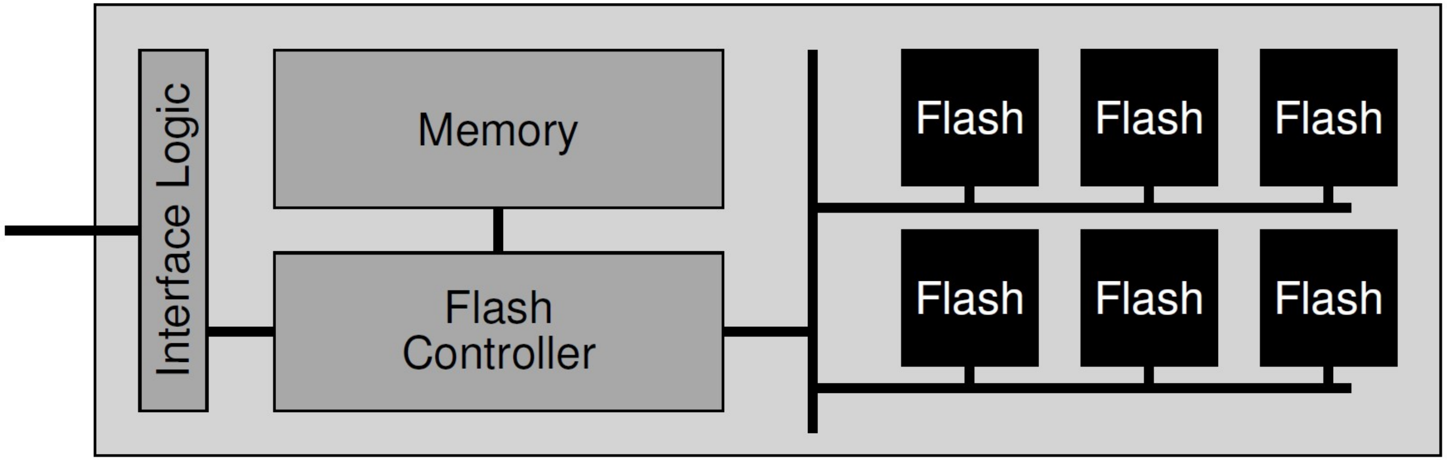
- No seek or rotational latency
- Transfer time: transfer a 4KB page
  - Limited by controller and disk interface (SATA: 300-600MB/s)
- Latency = Queuing Time + Controller time + Xfer Time
- Highest Bandwidth: Sequential OR Random reads

Flash Durability

- Flash memory stops reliably storing a bit
    - After many erasures (in the order of 10^3 to 10^6)
    - After a few years without power
    - After nearby cell is read many times (read disturb)
- To improve durability
    - Error correcting codes
        - extra bytes in every page
    - Management of defective pages and blocks
    - Spreads updates to hot logical pages uniformly over all blocks

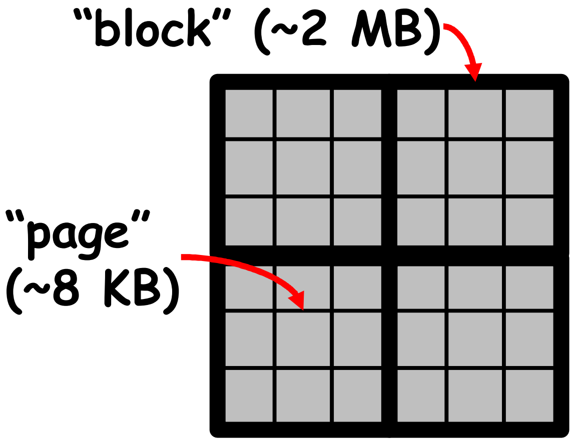
Flash Architecture

SSD Architecture (Simplified)



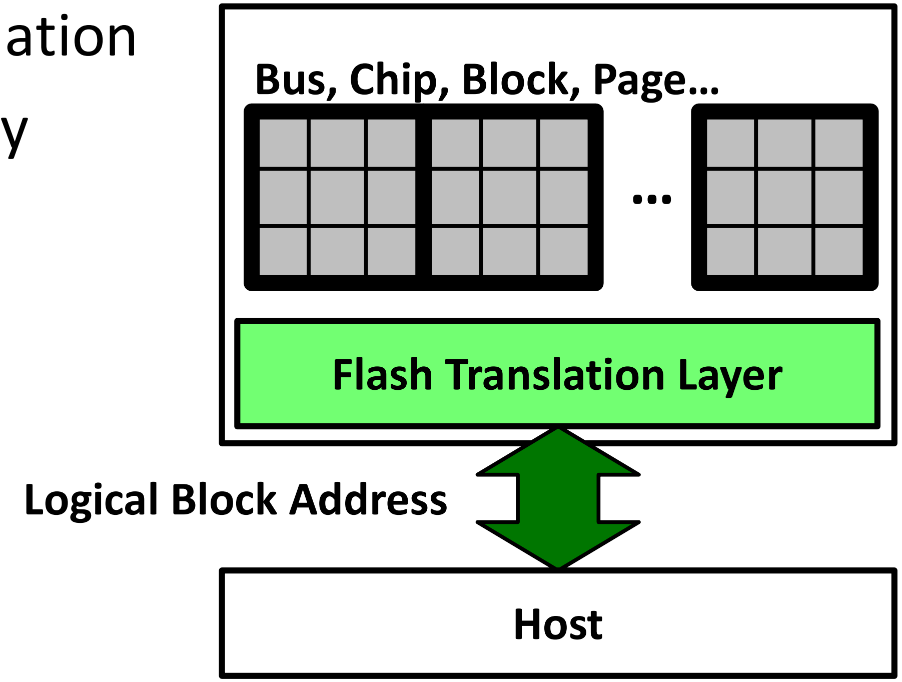
NAND-Flash Fabric Characteristics

- Performance impact of high-latency erase mitigated using large erase units (“blocks”)
  - Hundreds of pages erased at once
- What these mean: in-place updates are no longer feasible
  - In-place write requires whole block to be re-written
  - Hot pages will wear out very quickly
- People would not use flash if it required too much special handling

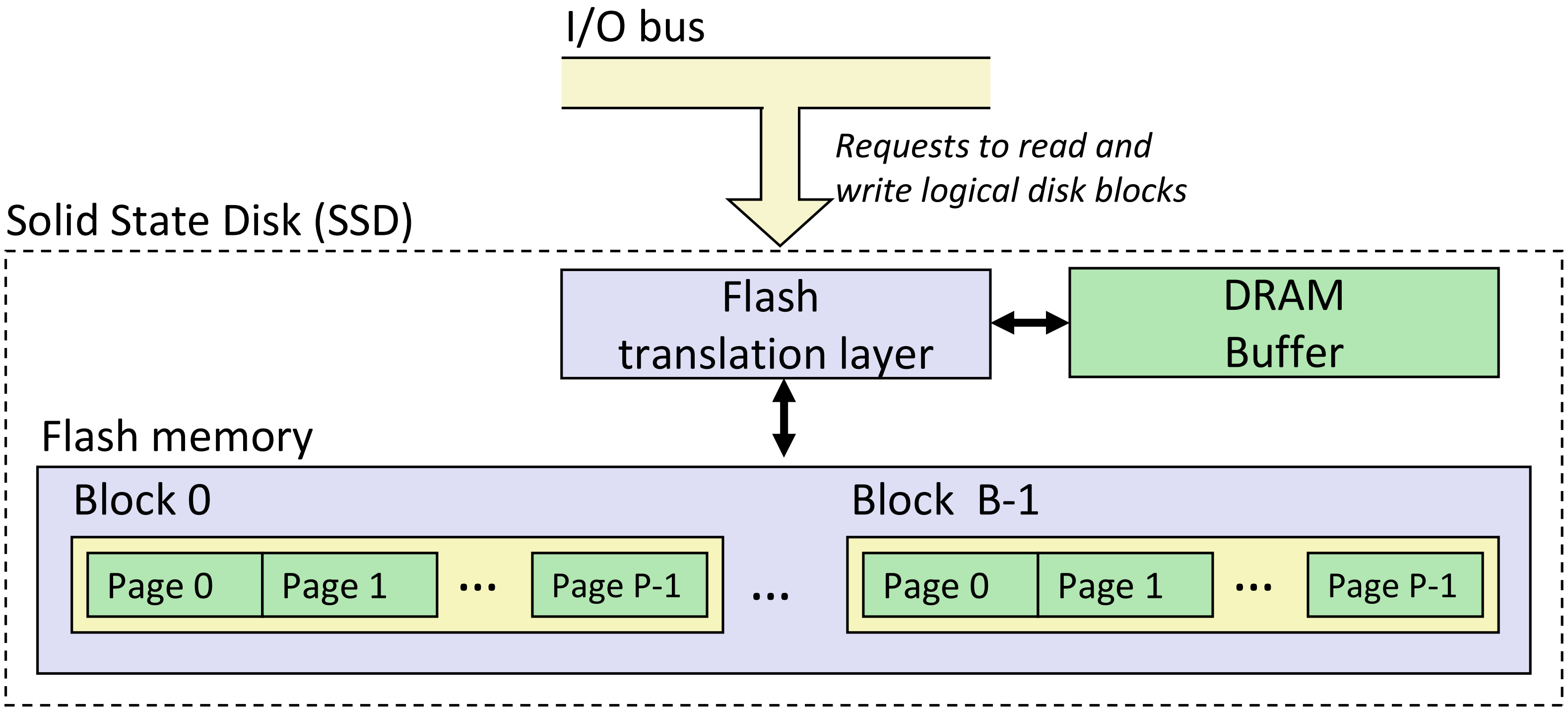


The Solution: Flash Translation Layer (FTL)

- Exposes a logical, linear address of pages to the host
- A “Flash Translation Layer” keeps track of actual physical locations of pages and performs translation
- Transparently performs many functions for performance/durability



Solid State Disks (SSDs)



- A block wears out after about 100,000 repeated writes

Some Jobs of the Flash Translation Layer

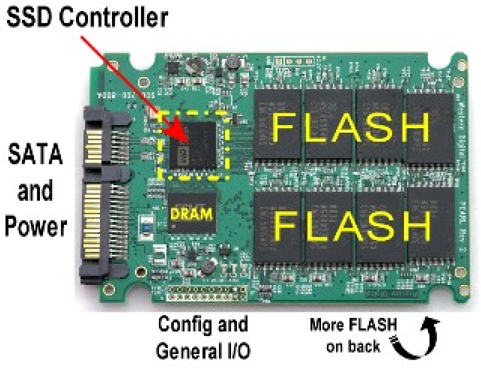
- Flash translation table maps logical page to several physical pages; Logical page is written to already erased physical page

- Logical-to-physical mapping
- Bad block management
- Wear leveling: Assign writes to pages that have less wear
- Error correction: Each page physically has a few more bits for error codes
  - Reed-Solomon, BCH, LDPC, …
- Deduplication: Logically map pages with same data to same physical page
- Garbage collection: Clear stale data and compact pages to fewer blocks
- Write-ahead logging: Improve burst write performance
- Caching, prefetching,…

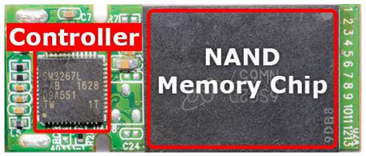
That’s a Lot of Work for an Embedded System!

- Needs to maintain multi-GB/s bandwidth
- Typical desktop SSDs have multicore ARM processors and gigabytes of memory to run the FTL
  - FTLs on smaller devices have sacrifice various functionality

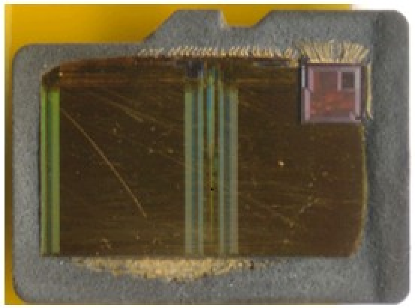
SATA SSD



USB Thumbdrive



MicroSD



Thomas Rent, “SSD Controller,” storagereview.com
Jeremy, “How Flash Drives Fail,” recovermyflashdrive.com
Andrew Huang, “On Hacking MicroSD Cards,” bunniestudios.com

Lecture Summary

- Pros (vs. hard disk drives):
  - Low latency, high throughput (eliminate seek/rotational delay)
  - No moving parts:
    - Very light weight, low power, silent, very shock insensitive
  - Read at memory speeds (limited by controller and I/O bus)
- Cons
  - Small storage capacity – compared to HDD
    - Though the ratio is changing
  - Expensive compared to SSD
    - SSD: 10 cents per GB
    - HDD: 4-6 cents per GB
  - Asymmetric block write performance: read pg/erase/write pg
    - Controller garbage collection (GC) algorithms have major effect on performance
  - Limited Drive lifetime
    - 1-10K writes/page for Multi level cell NAND
    - Avg failure rate is 6 years, life expectancy is 9–11 years

Solid State Disks (SSDs)

- 1995 – Replace rotating magnetic media with non-volatile memory (battery backed DRAM)
- 2009 – Use NAND Multi-Level Cell (2 or 3-bit/cell) flash memory
  - Sector (4 KB page) addressable, but stores 4-64 “pages” per memory block
  - Trapped electrons distinguish between 1 and 0
- No moving parts (no rotate/seek motors)
  - Eliminates seek and rotational delay (0.1-0.2ms access time)
  - Very low power and lightweight
  - Limited “write cycles”
- Rapid advances in capacity and cost ever since!
- Very popular now
  - Phones, Cameras, thumb drive, laptops, slates etc