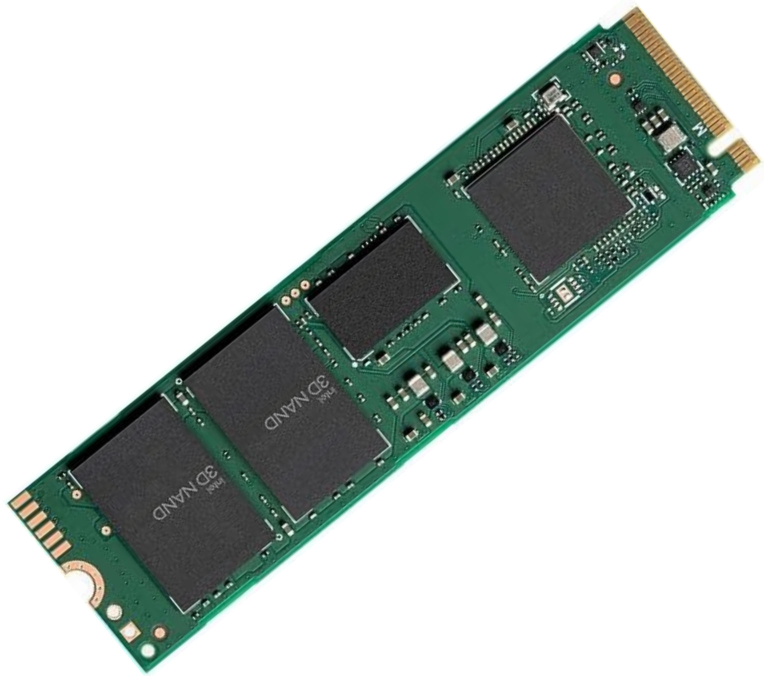
Solid State Drives (SSDs)

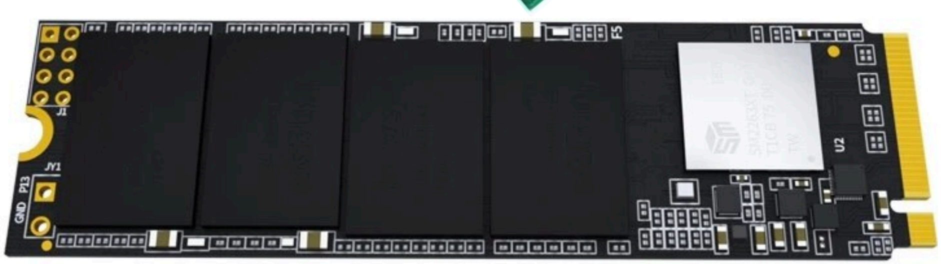
ICS332
Operating Systems

Henri Casanova (henric@hawaii.edu)

Solid State Drives (SSDs)

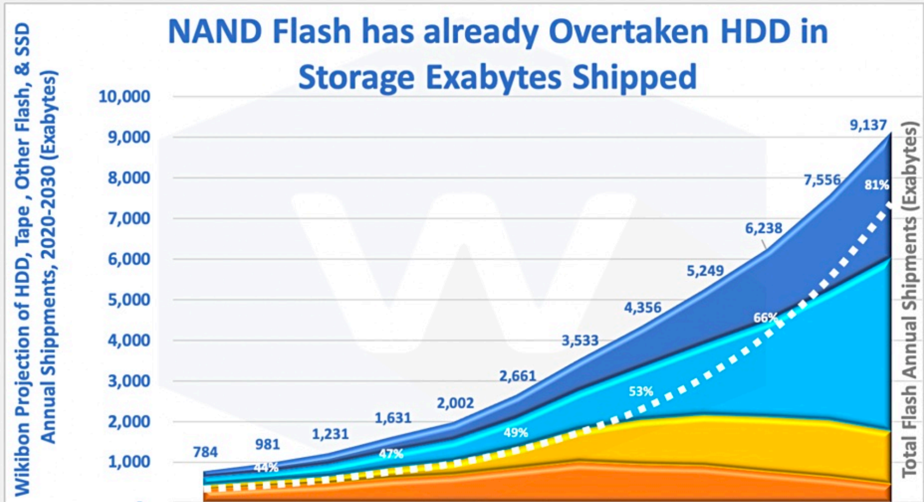
- Flash-based storage, no moving parts





Advantages over HDDs

- Faster, silent, lower power, more reliable, lighter
- Their market share has been increasing
- The only clear advantage of HDDs **for now**: $ / byte
  - "Greenness"? HDDs better for en environment to manufacture, but worse to operate

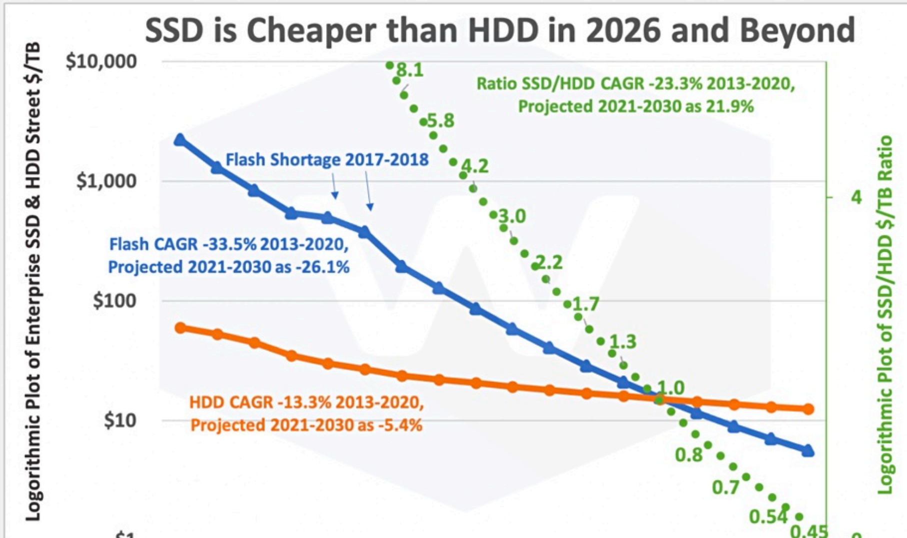


	2019	2020	2021	2022	2023	2024	2025	2026	2027	2028	2029	2030
SSD Flash Exabytes	152	217	272	355	444	586	774	1021	1348	1779	2349	3100
Other Flash Exabytes	191	218	304	412	524	707	955	1289	1740	2350	3172	4282
Tape Exabytes	115	137	177	259	352	533	812	1,105	1,253	1,350	1,404	1,298
HDD Exabyes	326	410	478	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.



	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

SSD $/TB HDD $/TB Ratio SSD/HDD $/TB

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.

SDD Storage Structure

- SSDs store bits into cells
  - Each cell can store 1, 2, or 3 bits depending on the technology
- Cells are organized into **pages** (e.g., 4KB)
- Pages are grouped into **blocks** (e.g., 128KB, 256KB)
- Blocks are grouped into banks (or planes)

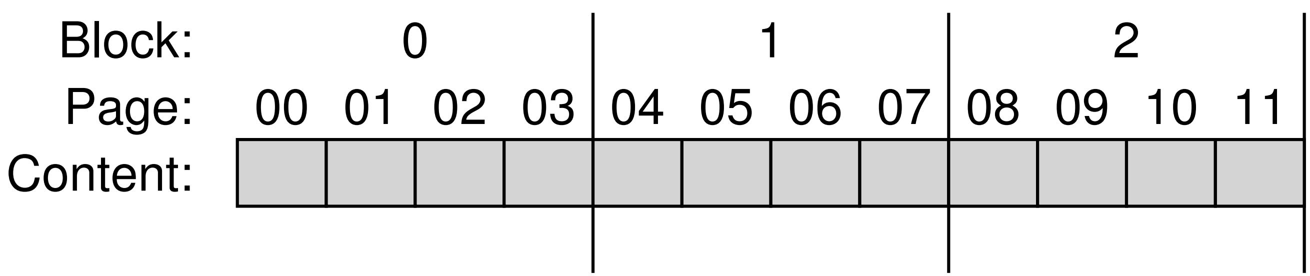


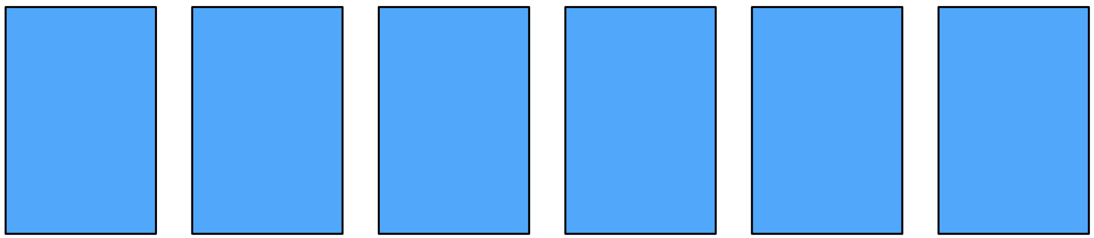
Figure 44.1: **A Simple Flash Chip: Pages Within Blocks**

SSD Operations

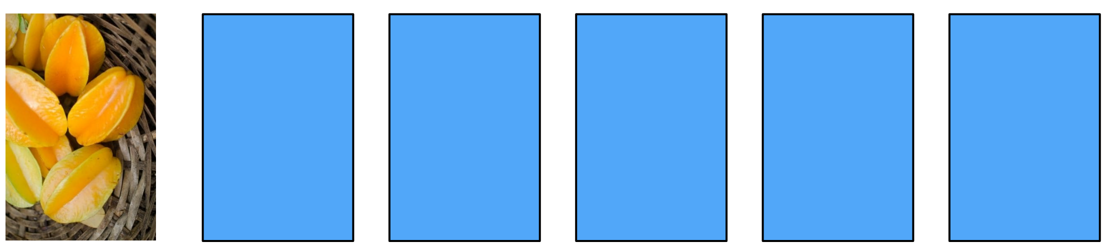
- **Read a page**
  - Very fast (μs), random access makes no difference (the major advantage over HDDs: locality doesn't matter)
- **Erase a block**
  - Much more expensive (ms)
- **Write (a.k.a. *program*) a page**
  - Requires that the page's block has been erased first!!!!
    - This is the "SSD weirdness": To update data in a page, you need to erase the whole block of pages
  - To make things worse: this causes wear out of the flash cells
  - Other problem: if you want to update, say, only one of the pages in a block, you need to first copy all other pages somewhere (e.g., the SSD's controller, RAM), then erase the whole block, and then write to all pages
  - This is called write amplification (we wrote more data than needed)
- Let's see this on an example….

Write Amplification Example (1)

- Say we have a 6-page block (each page is 4KB)

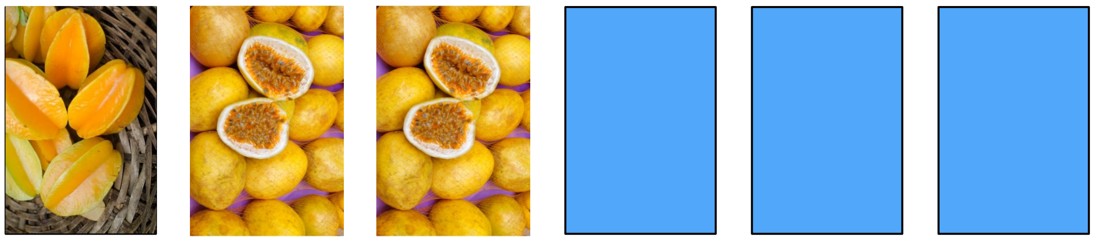


- Let's write a 4KB file



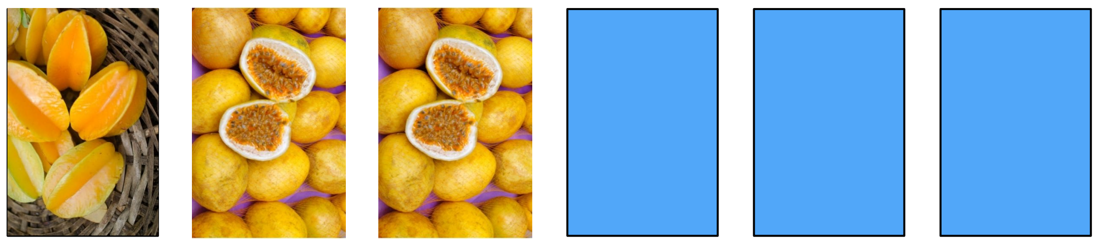
Write Amplification Example (2)

- Let’s write a 8KB file

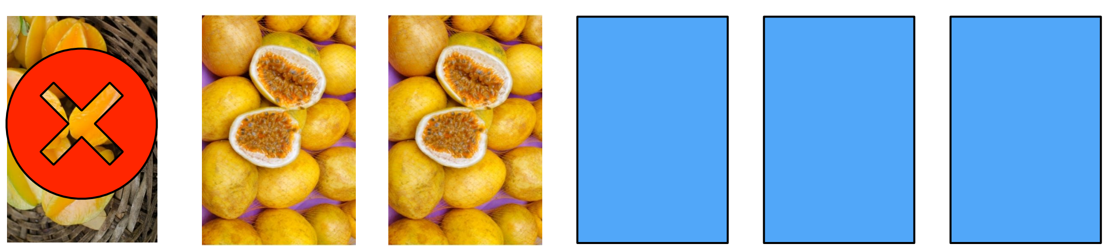


Write Amplification Example (2)

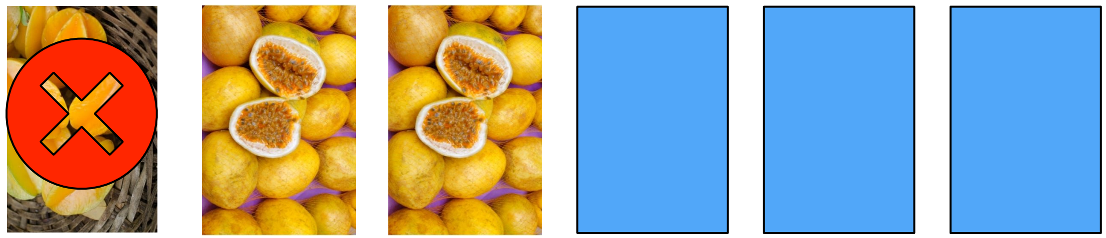
- Let's write a 8KB file



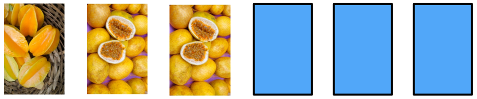
- Let's erase the first file
  - Instead of erasing the whole block, the SSD controller just marks the first page as invalid



Write Amplification Example (3)



- Now we want to write a 16KB file 
- The whole block's data is loaded into the SSD's controller's cache



- The data is updated in the cache

- The data is written to the block on the device

Write Amplification Example (4)

- We wanted to write 4KiB + 8KiB + 16KiB = 28KiB of application data
- We had to write 4KiB + 8KiB + 24KiB = 36KiB of data to the SSD
- We could come up with an example where we write about 5x more than what we need to write
- For this reason, the controller keeps writing on the SSD until full, before it attempts any rewrite
- Once full, writes are then more and more amplified
  - i.e., there is fragmentation everywhere, and no “free” blocks where data can be written
- And rewriting a block over and over leads to a wear out
- In the end, performance is still very good relative to that of an HDD, in part because SSDs employ several techniques….

SSD Techniques

- **Increase Performance:** the controller can clean up blocks with invalid pages at any time so that they're easily writable later
  - This is called garbage collection!
  - Same idea as the OS writing back dirty pages to disk once in a while when the disk is idle so that at the next page fault there is no need to write the page back
- **Decrease wear out:** the controller tries to spread writes over all pages, so that they would all wear out together as late as possible
  - In practice, for most "normal" workloads, wear out isn an issue
  - i.e., your SSD will be ok until you get your next laptop unless you do something very unusual

Conclusion

- SSDs are here to stay
  - By now we all have one!
- They are still more expensive than HDDs
  - HDDs will continue being used to store large dataset for the upcoming years
  - Just like tape storage is still in use!

- OSTEP has many more details in Chapter 44