Persistence: Solid-State Storage Devices

CS 537: Introduction to Operating Systems

Louis Oliphant

University of Wisconsin - Madison

Fall 2023

Administrivia

- Project 7 due Dec 8th @ 11:59pm
- Discussion Section tomorrow
  - Project 7 Support
  - Midterm 3 Preparation
- Midterm 3 scheduled for Dec 12th in-class
  - Alternate time Dec 18th @ 12:25pm (email me)
  - Alternate time also for McBurney accommodations

Review Journaling & Log-Structured File Systems

- Data Journaling and Metadata (or ordered) Journaling
  - Understand protocol of what gets written where and what waits occur to insure consistency
- Log-structured File System
  - Layout on disk – checkpoint region, segments (data, inodes, imap, segment summary),
  - Memory caching – imap and buffered writes
  - Garbage Collection – block liveness, which blocks to clean
  - Crash Recovery – multiple CRs, roll forward

Quiz 20 Journaling

https://tinyurl.com/cs537-fa23-q20

Solid-State Storage Devices

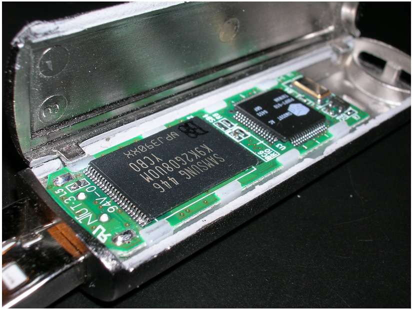
- Physical Storage System
  - SLC, MLC, TLC
  - Banks, Blocks, and Pages
- Flash-based Operations
  - Read (a page), Erase (a block), Program (a page)
- Flash Translation Layer (FTL)
- Log-Structured FTL
- Garbage Collection
- Mapping Tables
- SSD Performance and Cost

NAND FLASH

Single Level Cell (SLC) = 1 bit per cell
(faster, more reliable)

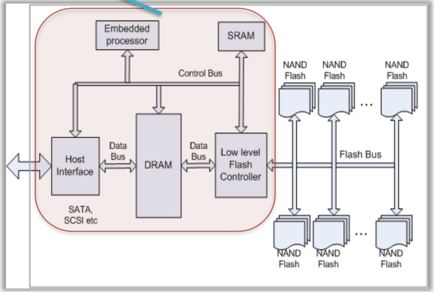
Multi Level Cell (MLC) = 2 bits per cell
(slower, less reliable)

Triple Level Cell (TLC) = 4 bits per cell
(even more so)



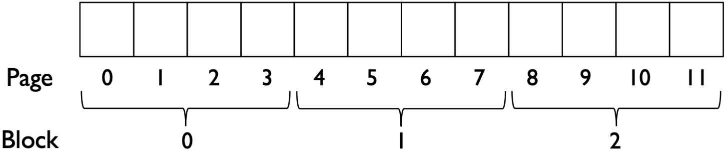
SSD STRUCTURE

Flash Translation Layer
(Proprietary firmware)



Simplified block diagram of an SSD

SSD PROPERTIES



Page ~ 4KB,
Block ~ 128 KB
or 256 KB

Read

Write

Failures: Block likely to fail after a certain number of erases
(~10000 for MLC flash, ~100,000 for SLC flash)

SSD OPERATIONS

Read a page: Retrieve contents of entire page (e.g., 4 KB)

- Cost: 25—75 microseconds
- Independent of page number, prior request offsets

Erase a block: Resets each page in the block to all 1s

- Cost: 1.5 to 4.5 milliseconds
- Much more expensive than reading!
- Allows each page to be written

Program (i.e., write) a page: Change selected 1s to 0s

- Cost is 200 to1400 microseconds
- Faster than erasing a block, but slower than reading a page

FLASH TRANSLATION LAYER

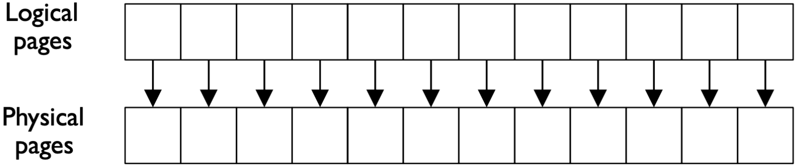
1. Translate reads/writes to logical blocks into reads/erases/programs

2. Reduce write amplification (extra copying needed to deal with block-level erases)

3. Implement wear leveling (distribute writes equally to all blocks)

Typically implemented in hardware in the SSD, but in software for some SSDs

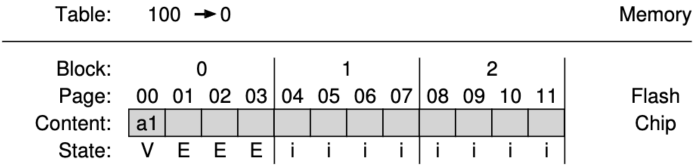
FTL: DIRECT MAPPING



Cons?

FTL: LOG-BASED MAPPING

Idea: Treat the physical blocks like a log



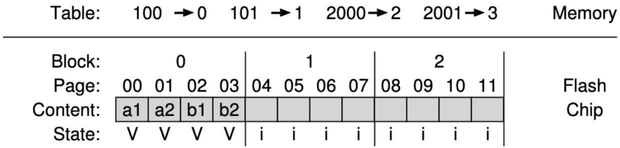
FTL: LOG-STRUCTURED ADVANTAGES

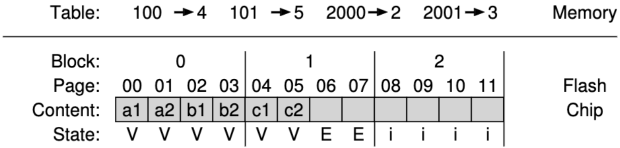
Avoids expensive read-modify-write behavior

Better wear levelling: writes get spread across pages,
even if there is spatial locality in writes at logical level

Challenges? Garbage!

GARBAGE COLLECTION





GARBAGE COLLECTION

Steps:

Read all pages in physical block

Write out the alive entries to the end of the log

Erase block (freeing it for later use)

Table: 100 → 4   101 → 5   2000 → 2   2001 → 3   Memory

Block:	0				1				2				Flash Chip
Page:	00	01	02	03	04	05	06	07	08	09	10	11	
Content:	a1	a2	b1	b2	c1	c2							
State:	V	V	V	V	V	V	E	E	i	i	i	i	

Table: 100 → 4   101 → 5   2000 → 6   2001 → 7   Memory

Block:	0				1				2				Flash Chip
Page:	00	01	02	03	04	05	06	07	08	09	10	11	
Content:					c1	c2	b1	b2					
State:	E	E	E	E	V	V	V	V	i	i	i	i	

OVERHEADS

Garbage collection requires extra read+write traffic

Overprovisioning makes GC less painful

- SSD exposes logical space that is smaller than the physical space
- By keeping extra, "hidden" pages around, the SSD tries to defer GC to a background task (thus removing GC from critical path of a write)

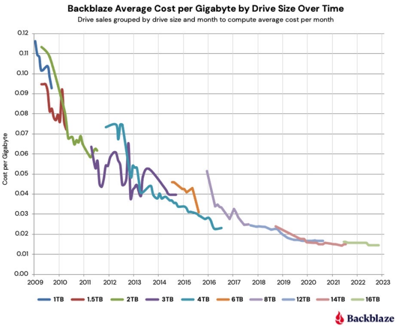
Occasionally shuffle live (i.e., non-garbage) blocks that never get overwritten

- Enforces wear levelling

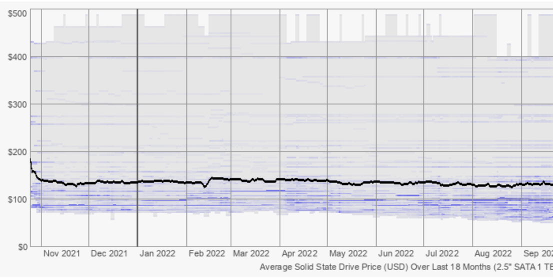
OVERALL PERFORMANCE

Device	Random Reads (MB/s)	Random Writes (MB/s)	Sequential Reads (MB/s)	Sequential Writes (MB/s)
Samsung 840 Pro SSD	103	287	421	384
Seagate 600 SSD	84	252	424	374
Intel SSD 335 SSD	39	222	344	354
Seagate Savvio 15K.3 HDD	2	2	223	223

COST?



~1.5 cents / GB



1TB ~ $150 on average
~15 cents / GB

Next Time – Data Integrity, Student Projects, Review