Persistence: Solid-State Storage Devices

CS 537: Introduction to Operating Systems

Louis Oliphant

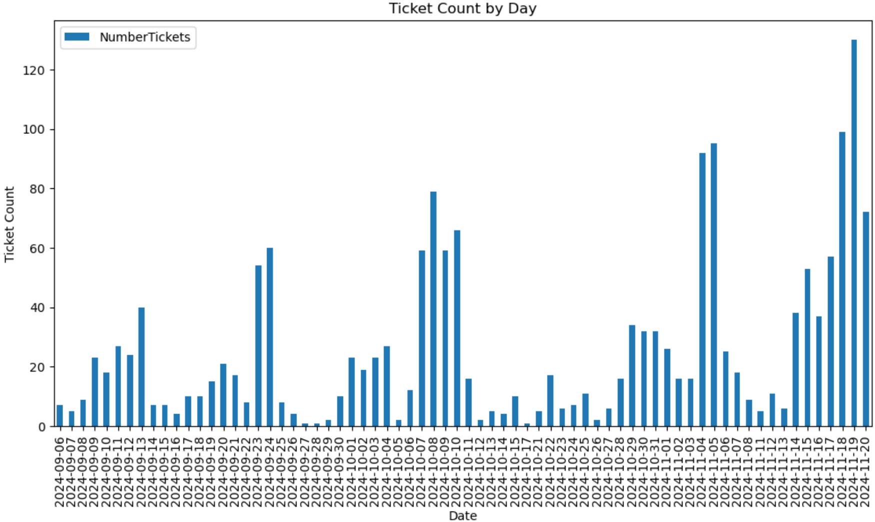
University of Wisconsin - Madison

Fall 2024

Administrivia

- Project 6 due Nov 27 and Dec 6
- Due to CSL issues, one extra slip day added to all students
- OH and discussions cancelled next Wed, Nov 27th
- Exam 2 Regrades done

Office Hour Tickets



Review FSCK, Journaling & Log-Structured File Systems

- FSCK
  - `fsck` attempts to scan and correct inconsistencies found in the file system.
  - build **used data blocks** from inode table, checks inodes and directory entries for consistency
- 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

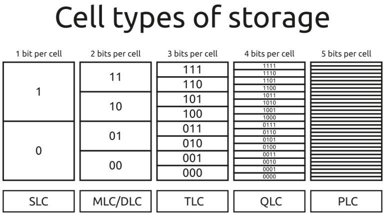
Quiz 19 LFS

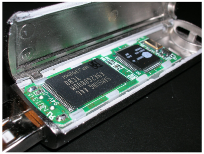
https://tinyurl.com/cs537-fa24-q19

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 Storage

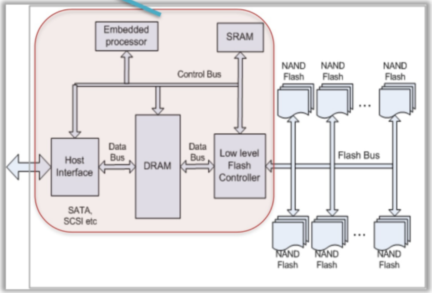




- Single Level Cell (SLC) = 1 bit per cell (faster, more reliable)
- Multi Level Cell (MLC) = 2 bits per cell (slower, less reliable)
- Triple level Cell (TLC) = 4 bits per cell (even more so)

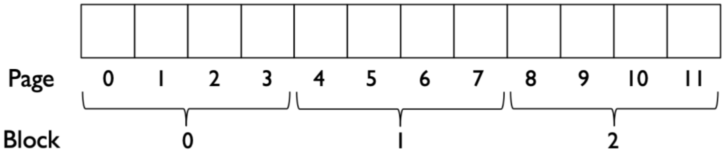
SSD Structure

Flash Translation Layer
(Proprietary firmware)



Simplified block diagram of an SSD

SSD Layout and Operations



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

- **SSD Operations**:
  - **Read** a page
    - ~ 25-75 microseconds
    - independent of page number, prior requests
  - **Erase** a block (reset all pages in block to all 1s)
    - ~ 1.5 to 4.5 milliseconds
    - Pages must be erased before they can be programmed
  - **Program** (i.e. write) a page
    - 200 to 1400 microseconds

- **Page States**
  - INVALID – Can only be erased
  - ERASED – Ready to be programmed
  - VALID – Programmed, ready to be read

FLASH TRANSLATION LAYER

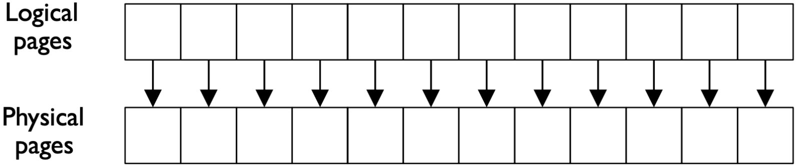
1. Translate reads/writes to logical blocks into reads/erases/programs

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

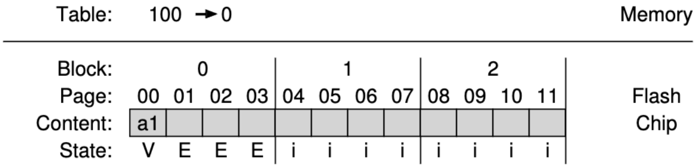
FTL: DIRECT MAPPING



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

Table: 100 → 0   101 → 1   2000 → 2   2001 → 3   Memory

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

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

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

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)

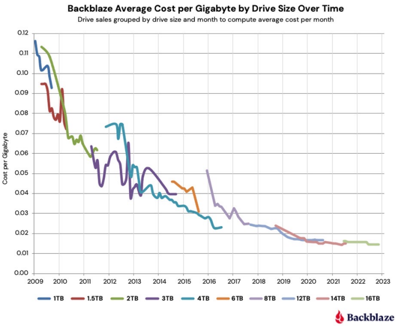
Occasionally shuffle live (i.e., non-garbage) blocks that never get overwritten

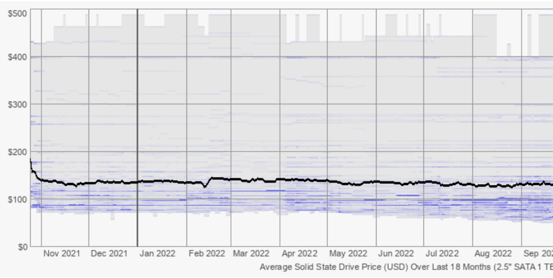
- Enforces wear levelling

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

More Modern Drive - Samsung 970 EVO Plus 2 TB SSD /w Cache

Solid-State-Drive	
Capacity:	2 TB (2000 GB)
Variants:	250 GB · 500 GB · 1 TB · 2 TB
Hardware Versions:	• Elpis + V6 250 GB · 500 GB · 1 TB · 2 TB • Phoenix + V5 250 GB · 500 GB · 1 TB · 2 TB
Overprovisioning:	185.4 GB / 10.0 %
Production:	Active
Released:	Feb 2019
Price at Launch:	220 USD
Part Number:	MZ-V7S2T0B/AM

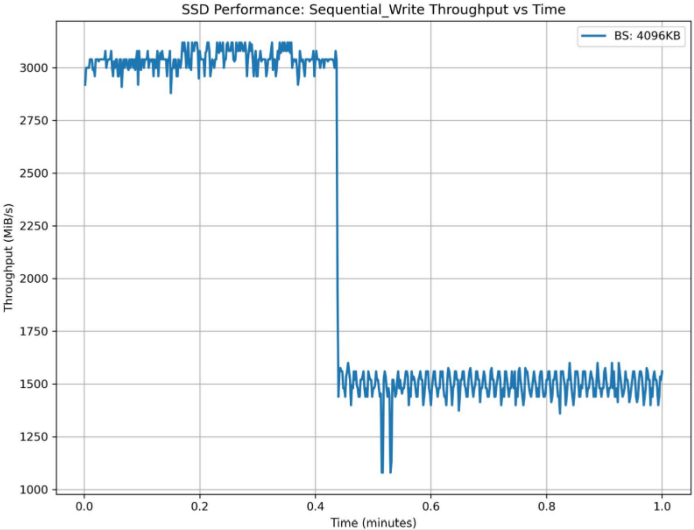
Performance	
Sequential Read:	3,500 MB/s
Sequential Write:	3,300 MB/s
Random Read:	620,000 IOPS
Random Write:	560,000 IOPS
Endurance:	1200 TBW
Warranty:	5 Years
MTBF:	1.5 Million Hours
Drive Writes Per Day (DWPD):	0.3
SLC Write Cache:	approx. 78 GB (72 GB Dynamic + 6 GB Static)
Speed when	

SSD More Specs

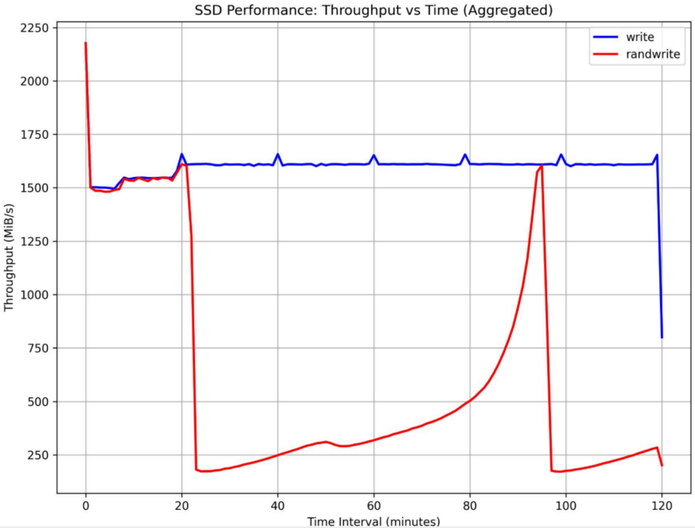
NAND Flash	
Manufacturer:	Samsung
Name:	V-NAND V5
Type:	TLC
Technology:	92-layer
Speed:	533 MT/s .. 1400 MT/s
Capacity:	2 chips @ 8 Tbit
Toggle:	4.0
Topology:	Charge Trap

Read Time (tR):	73 µs
Program Time (tProg):	500 µs
Block Erase Time (tBERS):	3.5 ms
Die Read Speed:	438 MB/s
Die Write Speed:	64 MB/s
Page Size:	16 KB

SSD Performance Drop 1



SSD Performance Drop 2



Summary

Persistence Summary

- IO Devices / Disks
- File System API
- File System Implementation / Fast File System
- Journaling
- Log Structured FS
- SSDs

Advanced Topics

- Virtual Machines
- Multiprocessor Scheduling
- Distributed Systems