Solid State Drive Architecture and Garbage Collection Algorithms

Colby Whitaker, Gurjeet Singh, Hassan Lashari & Garrett Clay
Outline

- SSD Architecture
- Access Operations
- Wear Leveling
- Garbage Collection
- Simulators
- Future Work
SSD Architecture

- Host interface
- Buffer
- Flash translation layer (FTL)
  - Addressing
  - Wear-leveling
  - Bad block management
  - Garbage collection
- NAND flash memory
Flash Architecture

Die

Plane

Plane

2-4 planes

Block

Page

1024-2048 blocks

String

64-256 pages

2-16 kB
Outline

● SSD Architecture
● Access Operations
● Wear Leveling
● Garbage Collection
● Simulators
● Future Work
Write / Read Operation

- Page level operations
- Write is second slowest operation (300 us)
- Read is fastest operation (often less than 50 us)

Write Steps:
1. Apply voltage to chosen bit line
2. Apply large voltage to wordline

Read Steps:
1. Chosen bit line is read by connected ADC
2. Chosen wordline is grounded (while others are kept at a high voltage)

Write / Read Buffer

- Optimized performance
- Buffer management
  - Block padding least recently used (BPLRU)
  - Flash aware buffer (FAB)
  - Popularity aware buffer (PAB)

Erase

- Block level operation
- Slowest operation (up to a few ms)
- Reduces SSD lifetime

Erase Steps:
1. Apply high negative voltage to repel electrons out of the floating gate region

Error Correcting Codes

- Deterioration from Read/Write operations can cause random bit errors
- Decrease in cell size requires advanced ECC
  - BCH codes
  - LDPC (Low Density Parity Check)
- HDDs use ECC techniques as well
Outline

● SSD Architecture
● Access Operations
● Wear Leveling
● Garbage Collection
● Simulators
● Future Work
Wear-leveling

- Fixed erase cycles for floating gate transistors due to trapping of charge in floating gate during erase operations
- About 10K cycles for MLC and 100K cycles for SLC
- Wear-leveling is the technique to ensure even life span for all the transistors in the SSD

Beug, M.F., NAND Flash Technology, Inside Solid State Drives, p-96
Software Architecture of SSDs

- Flash Translation Layer helps operating systems to interface SSDs in a same way as HDDs
Evenness-Aware Algorithm

- Maintains a cyclic queue based on the usage of individual memory location and the count of erase cycles
- New memory location is chosen according to this queue
- Two Important elements of algorithm:
  - Branch Erasing Table
  - SW leveler

Evenness-Aware Algorithm (Contd.)

Dual-pool Algorithm

- Dual pool algorithm keep two kinds of pools
- One pool with address which are extensively used known as hot-pool and one with rarely used addresses known as cold pool
- Occasionally, this algorithm change rarely used addresses of cold pool with frequently used addresses of hot pool

Dual-pool Algorithm (contd.)

Bad Block Management

- Bad Block Management module creates and maintains a map of bad blocks and replace those blocks with good physical blocks.
- Contains the list of the bad blocks already present during the factory testing of the NAND Flash memory modules[1].
- It is updated during device lifetime whenever a block becomes bad.

Reference:[1]Eshghi, K; SSD Architecture and PCI Express Interface; Inside Solid State Drives; p-30
Increasing Erase Cycles for SSD

- Heat plates of Chalcogenide glass can be added to the transistors to provide heating to the device and free up the electrons[1].

[1] Chiu, Y; Flash Memory Survives 100 Million Cycles; spectrum.ieee.org, 30 Oct,
Outline

- SSD Architecture
- Access Operations
- Wear Leveling
- Garbage Collection
- Simulators
- Future Work
Garbage Collection

Erase operation takes place at the block level, the process of copying valid pages, then erasing the block is called *Garbage Collection (GC)*.

- Erase, read & write operations (ms)
- Erase & write degrade cells - wear leveling
- GC targets invalid blocks, makes them free

Garbage Collection

- GC efficiency metric - average number of invalid pages in a block to be erased
  - space
  - write amplification
- Block and system thresholds

“The ratio of the total number of writes and the number of externally requested writes is termed write amplification and it is a measure for the effectiveness of a GC algorithm.” [1]

Garbage Collection - Published Results

320GB Fusion IO ioDrive

120GB Intel 320 SSD

Garbage Collection Algorithms

FIFO - selects blocks to erase in a cyclical manner

Ex. threshold = 60% (invalid pages)

Garbage Collection Algorithms

Greedy - selects blocks with fewest number of invalid pages

Ex. threshold = 60% (invalid pages)

Garbage Collection Algorithms

Windowed - FIFO with greedy window

Ex. threshold = 60% (invalid pages), window size = 4

Garbage Collection Algorithms

d-choices - selects blocks with the fewest number of valid pages out of a set of d randomly chosen blocks

Ex. threshold = 60% (invalid pages), window size = 6

Outline

● SSD Architecture
● Access Operations
● Wear Leveling
● Garbage Collection
● Simulators
● Future Work
Simulators

- SSD simulators not as common as HDD simulators
- Two types of SSD simulators
  - Hardware-based simulators
  - Software simulators
- We will focus on FlashSim and Microsoft SSD extension for DiskSim
FlashSim

- Single-threaded object-oriented simulator
- Each class represents hardware or software component
- Ability to simulate your own FTL (Flash Translation Layer)
- Allows the simulation of energy characteristics
Microsoft SSD Extension

- Add-on to the popular HDD simulator DiskSim
- Ideal NAND flash storage
- Single FTL
- Can be integrated into a bigger system
- Trace driven and internal synthetic workload
- Various parameters:
  - Background cleaning
  - Interleaving

Simulation Results

Average Response Time vs Garbage Collection System Threshold

Response Time vs GC System Threshold with Greedy Algorithm
Outline

● SSD Architecture
● Access Operations
● Wear Leveling
● Garbage Collection
● Simulators
● Future Work
Future Work

- Power consumption simulator for SSDs
- Thermal characteristics simulator for SSDs
- Adding GC algorithms to simulators
- Move processors closer to memory - Small processors with dedicated tasks can predict the memory locations required for processing and thus can move those locations up in cache, thus reduce miss rate and increase processing speed.
References


Business advantage of SSDs over HDD (PCs)

Data Centers

[Image of an Intel Solid-State Drive Data Center TCO Calculator]

SSD Solution – Drive Lifetime TCO Analysis

- TCO Impact: $6,168
- I/O Performance: 947.6%
- IOPS Gain: 179,100
- Reduction in power: 92%
- Total HDDs: 50
- Total SSDs: 2
- Total HDD Capacity Consumed: 360 GB
- Total Usable SSD Capacity: 3,200 GB

TCO Results

- Power Cost
  - HDD: $921
  - SSD: $74
- Cooling Cost
  - HDD: $1,106
  - SSD: $98
- Total Energy Cost
  - HDD: $2,027
  - SSD: $162
  - Savings: $1,865

Disk Array

- Cost of Enclosures
  - HDD: $6,500
  - SSD: $0
- Enclosure Energy Draw
  - HDD: $1,365
  - SSD: $0
- Total Enclosure Cost
  - HDD: $7,865
  - SSD: $0
  - Savings: $7,865

Drive and Maintenance

- Cost
  - HDD
  - SSD

http://estimator.intel.com/ssdpro/
SSD as cache

Windows Boot Time
BootRacer 2nd Reboot

The Elder Scrolls V: Skyrim
Level Load Time

Reference: http://www.hardcoreware.net/ssd-cache-performance/