Persistence:
Flash-based Solid State Disks

OSTEP Chapter 44:
http://pages.cs.wisc.edu/~remzi/OSTEP/file-ssd.pdf

Slides based on Youjip Won's (https://oslab.kaist.ac.kr/people/) material.

Jan Reineke
Universität des Saarlandes

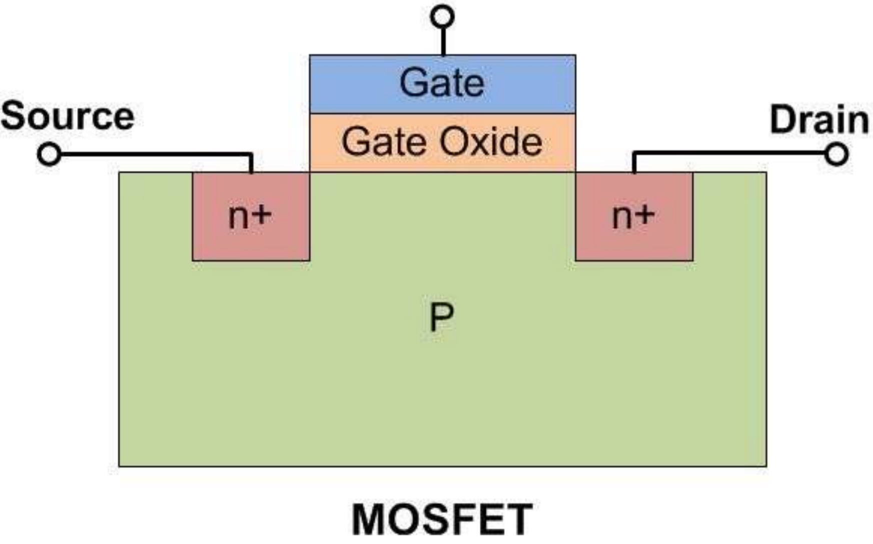
Solid-state storage devices

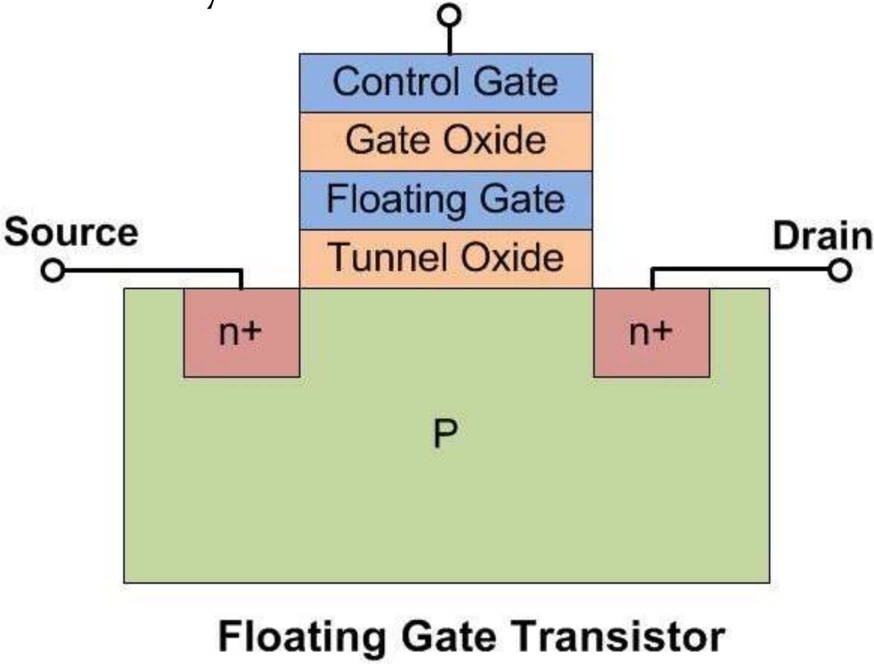
- No mechanical or moving parts like HDD
- Built out of transistors (like memory and processors)
- Retain information despite power loss unlike typical RAM

Memory cells: Floating gate transistors

p-type transistor:
gate controls the conductivity between source and sink

floating-gate transistor:
trapped electrons in floating gate controls the conductivity between source and sink



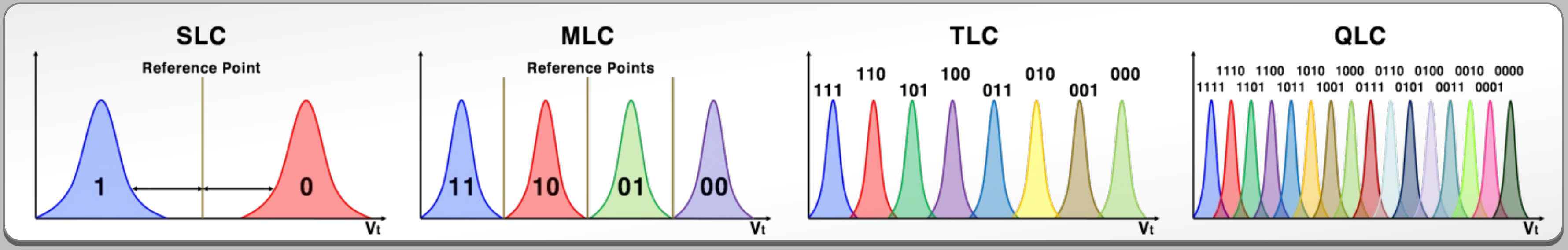


- electrons can be **trapped in** and **removed from** the floating gate
- electrons do not escape otherwise → **persistent memory**

[Source: https://www.embedded.com/flash-101-types-of-nand-flash/]

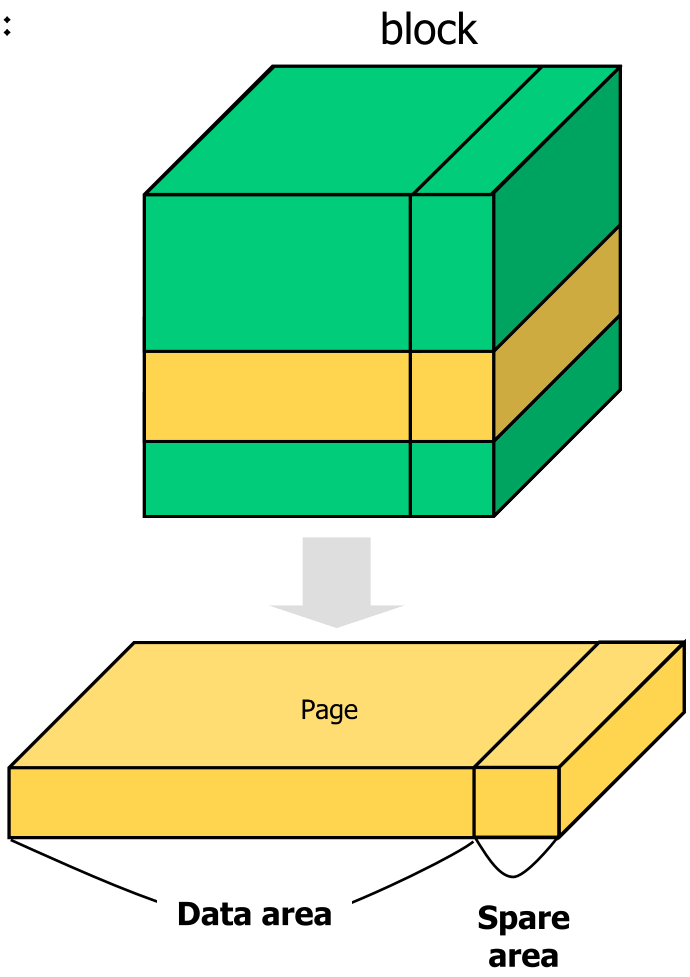
Types of cells

- Single-level cell (SLC): a single bit per cell
- Multi-level cell (MLC): two bits per cell
- Triple-level cell (TLC): three-bits per cell
- … Penta-level cells (PLC) currently under development



Structure of Flash

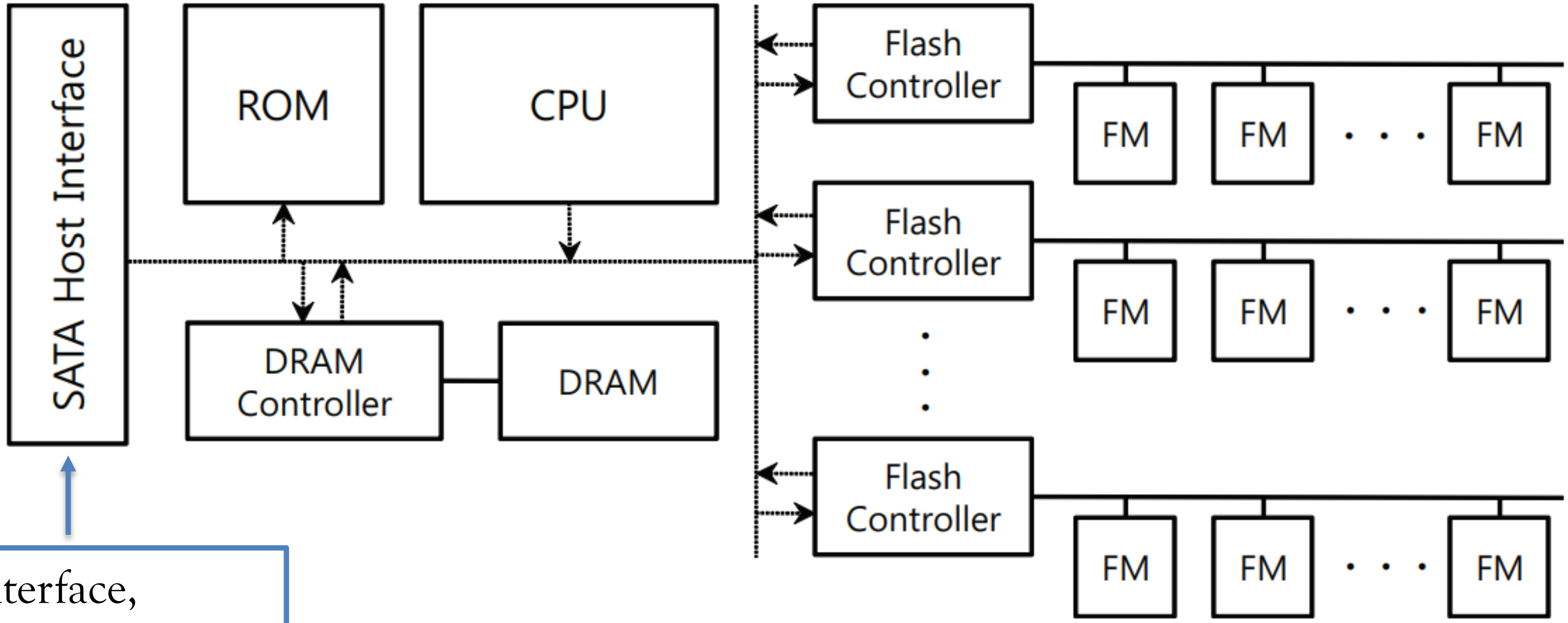
Hierarchical organization:



4 – 64 pages per block

Array of memory cells:

Structure of Flash SSDs



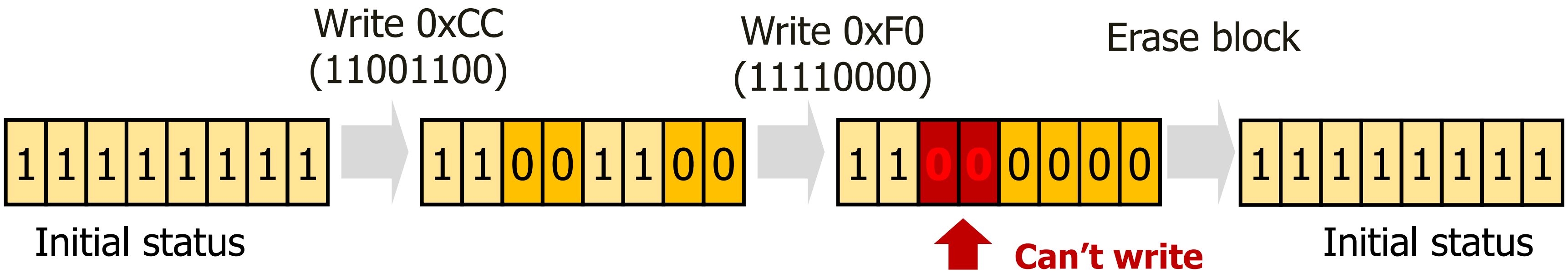
Bus interface, connection to CPU

Complex translation layer

Flash Memory

Basic operations

- Read: at page granularity
- Write (“program”): 1 → 0: at page granularity
- Erase: 0 → 1: **only** at block granularity



Write 0xCC (11001100)

Write 0xF0 (11110000)

Erase block

1 1 1 1 1 1 1 1

1 1 0 0 1 1 0 0

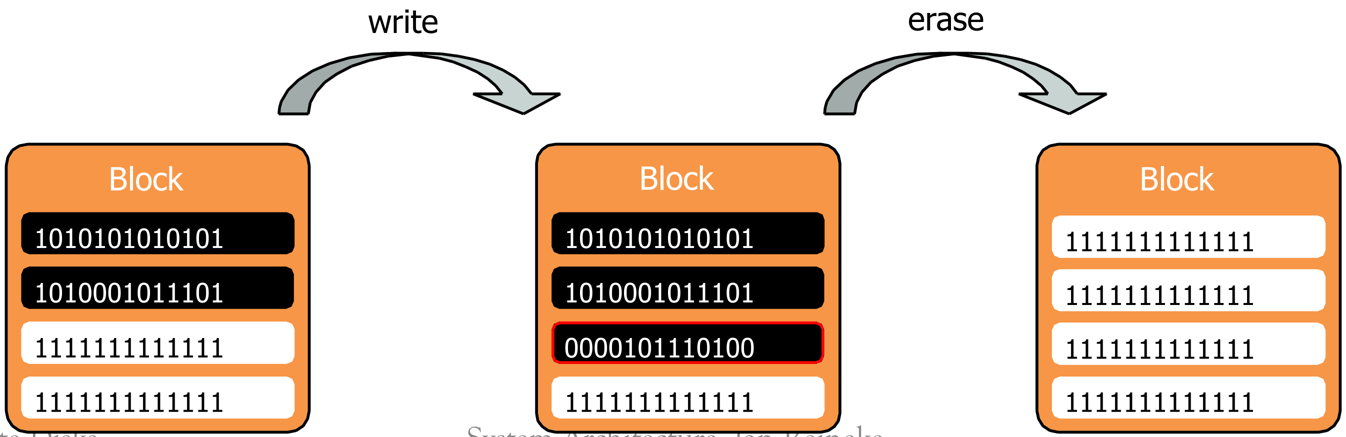
1 1 0 0 0 0 0 0

1 1 1 1 1 1 1 1

Initial status

Can’t write

Initial status



write

erase

Block

10101010101010

10100010111101

11111111111111

11111111111111

Block

10101010101010

10100010111101

0000101110100

11111111111111

Block

11111111111111

11111111111111

11111111111111

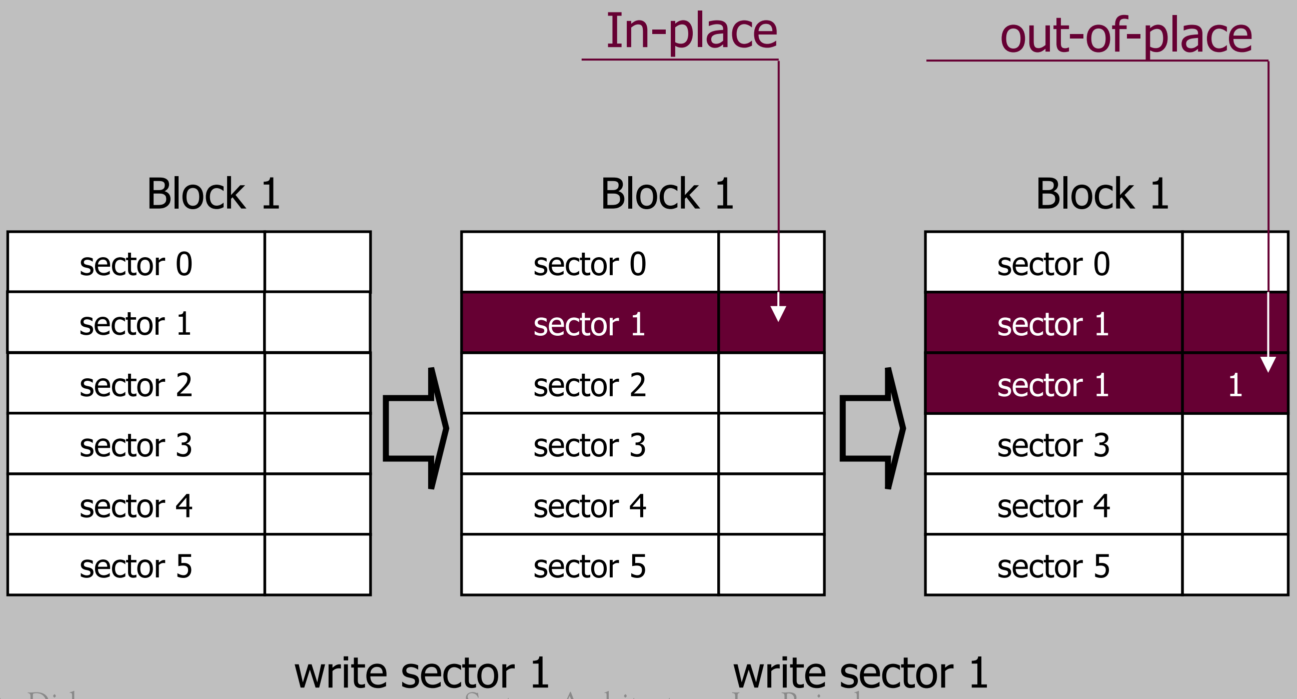
11111111111111

Reliability of Flash

- Wear out
  - Flash cell "wears out" as we program/erase it
  - Eventually, cells may become unusable
  - Typical erase/wear out cycle
    - MLC-based block: 10,000 P/E (Program/Erase)
    - SLC-based block: 100,000 P/E

Out-of-place update in Flash memory

- Need to erase block before writing to page
- *Implication:*
  Flash SSD uses “out-of-place” update for writes



In-place

out-of-place

Block 1

sector 0
sector 1
sector 2
sector 3
sector 4
sector 5

write sector 1

Block 1

sector 0
sector 1
sector 2
sector 3
sector 4
sector 5

write sector 1

Block 1

sector 0
sector 1
sector 1 1
sector 3
sector 4
sector 5

Flash Translation Layer (FTL)

A software layer that makes SSDs look like HDDs

- Address translation (yet another level!)
  - program pages within an erased block in order
- Wear leveling
  - tries to spread writes evenly across all blocks (locality is “bad”)
- Garbage collection

Comparison with Hard disks

	Hard disk	Flash-based SSD
Sequential access performance (throughput)	250 MB/s	several GB/s 15 GB/s (demonstrated) 7 GB/s (available commercially)
Random access latency	3-12 ms	< 0.1 ms
Cost	˜12 Euro/TB	˜35 Euro/TB
Density	1.2 TB/sq. inch	2.8 TB/
Lowest operating temperature	Most modern HDDs can operate at 0 °C	SSDs can operate at −55 °C
Highest altitude	HDDs will fail to operate at altitudes above 12,000 meters	no constraint

Summary

- Flash-based SSD is much faster than disk, in particular for random access patterns, but …
- It is more expensive
- It is not a drop-in replacement for a disk beneath a file system without a complex emulation layer
  - Challenging due to erasure granularity