Computer Architecture
计 算 机 体 系 结 构

Lecture 6. Data Storage and I/O
第六讲、数据存储和输入输出

Chao Li, PhD.
李超 博士

SJTU-SE346, Spring 2019

Review

- Memory hierarchy
- Cache and virtual memory
- Locality principle
- Miss cache, victim cache, prefetch buffer
- DRAM concept
- 1T1C cell, data access
- Memory design challenges

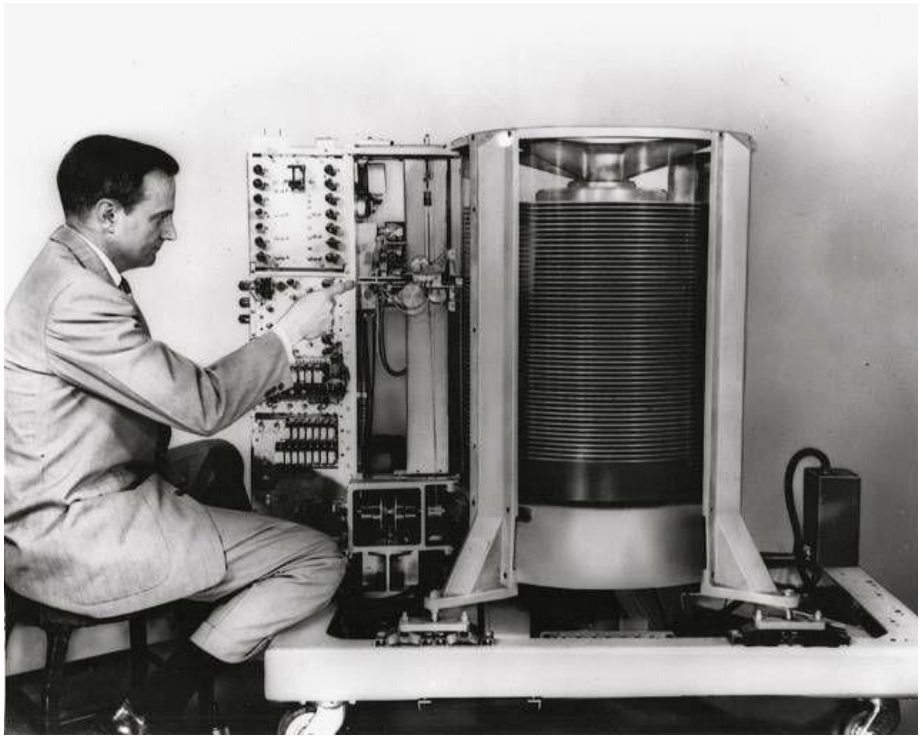
The I/O Problem

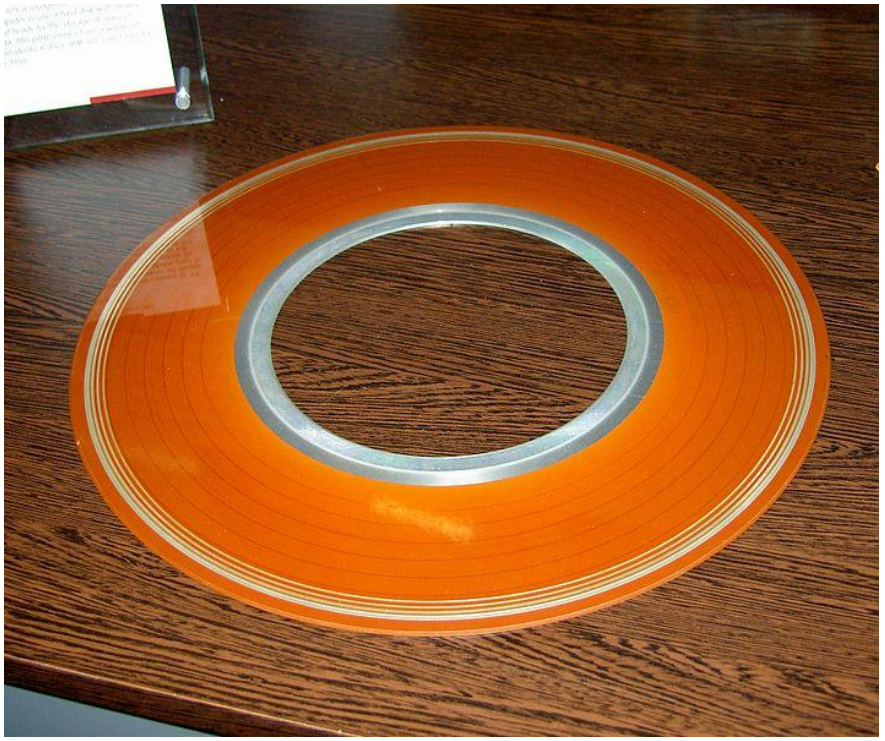
- Current processor performance
  - Intel Core i5-4690 CPU@ 4.9GHz, 5.25 GFLOPS/core
- Memory Bandwidth
  - 166 MHz * 2 lines/clock * 64 bits/line * 2 channel / 8 = 5.3 G/s
- Disk drive performance
  - Seagate Barracuda: 200 MB/s with SATA
  - I/O performance has improved less than 10% per year

Outlines

- Basic Concept
- Disk Interface
- Disk Array and RAID
- NAS and SAN
- Flash Storage Device

IBM 305 RAMAC



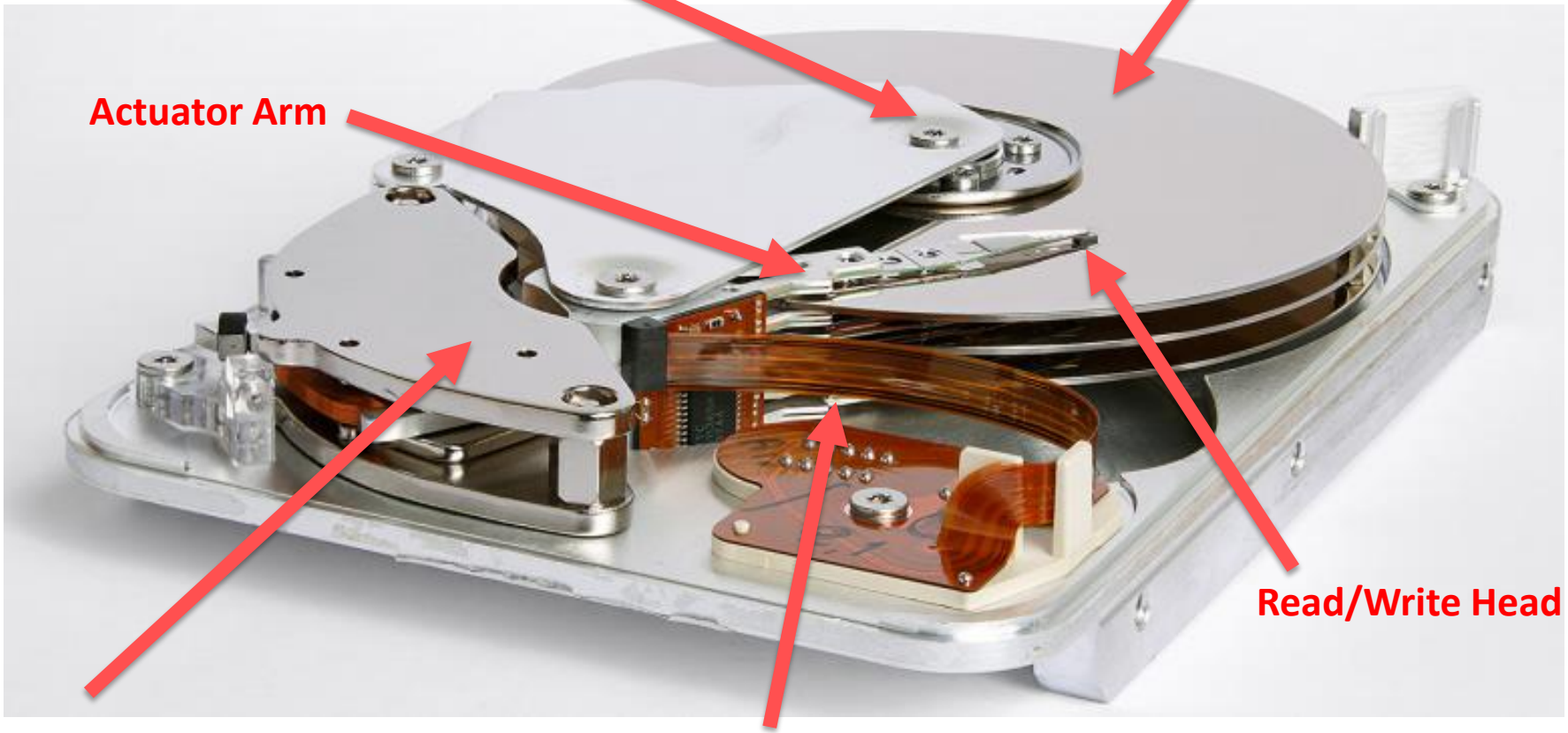


- The first commercial computer that used a moving-head hard disk drive (magnetic disk storage) [1955]
  - Fifty 24 inch aluminum disks; 100 tracks per surface; 1200 rpm; 5 million characters (7 bits each)

Evolution Timeline

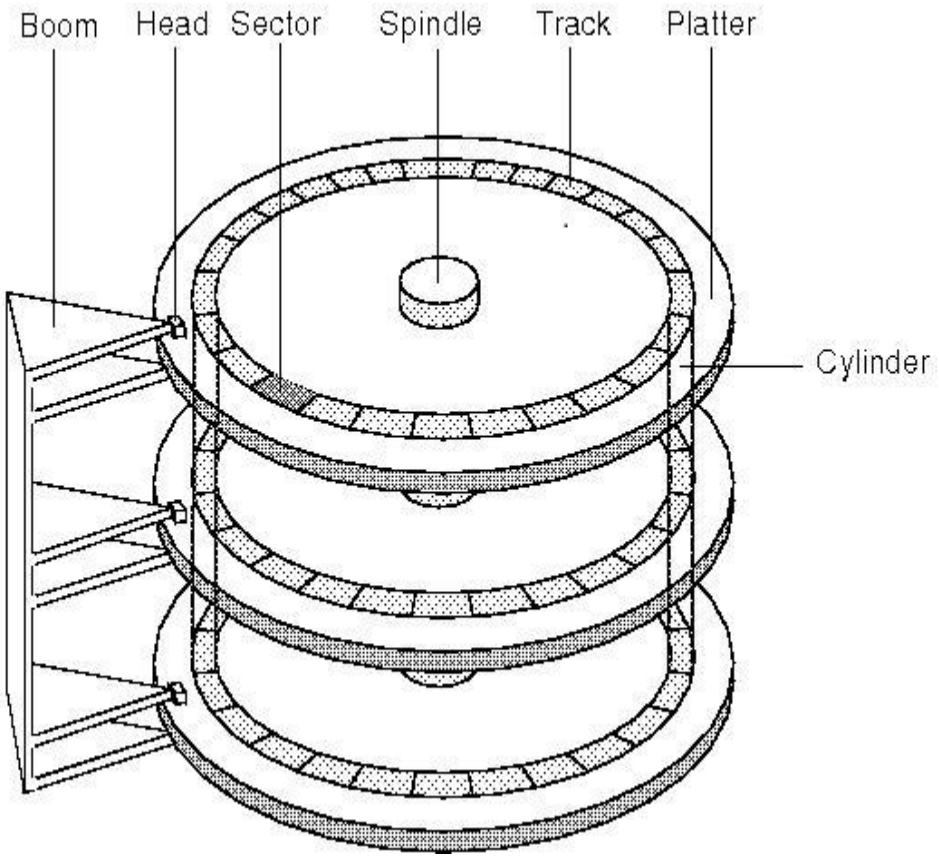
- 1956: first commercial disk drive (IBM)
- 1980: first gigabyte-capacity disk (IBM)
- 1986: standardization of SCSI
- 1988: first 2.5 inch HDD (PrairieTek)
- 2002: 137 GB addressing barrier broken
- 2003: serial ATA introduced
- 2007: first 1TB hard drive (Hitachi)
- 2008: first 1.5TB hard drive (Seagate)
- 2009: first 2.0TB hard drive (Western Digital)

Major Components of a Typical Disk Drive



Disk Parameters

Platter: A non-magnetic storage surface with data on both sides



Track: A circular "slice" of area on a platter's surface

Sector: A uniform subsection of a track

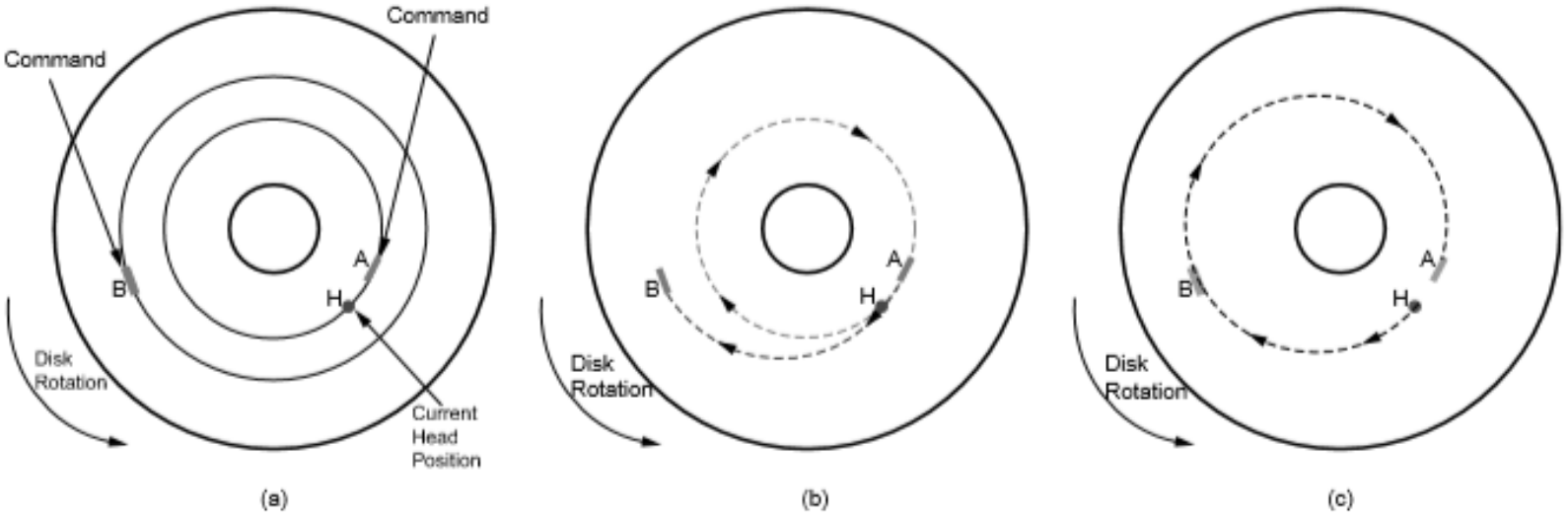
Cylinder: A set of vertically overlapping tracks

Example: Seagate Barracuda

- 3 platters (6 heads)
- 75 nanometers track width
- 63 sectors per track
- 4096 bytes per sector
- 72000 RPM

Access Cost

- **Seek Time**: move the head to the proper track
- **Rotational Latency**: bring the target sector to the head
- **Data Transfer Time**: the time needed to read the data
- Others: controller delay and queueing delay

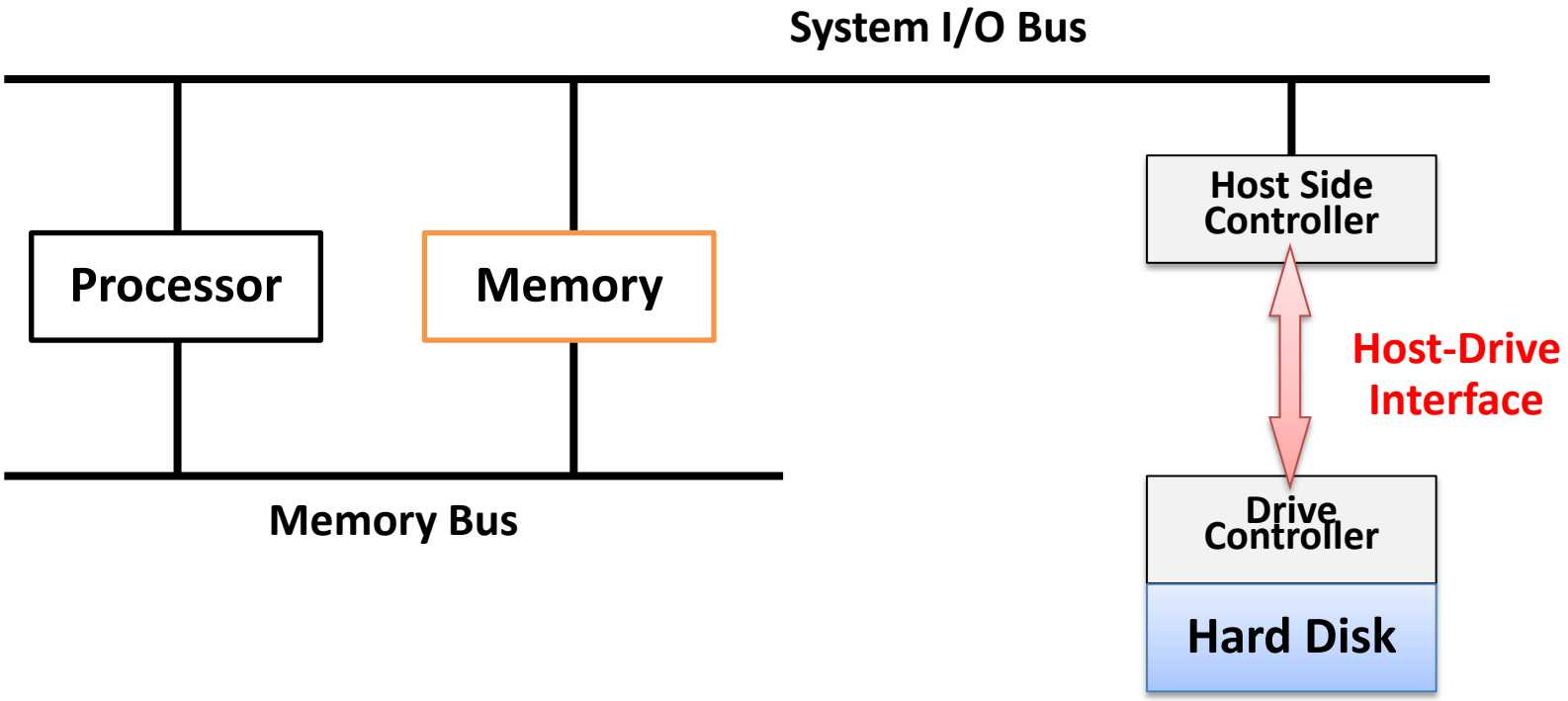


Total-Access-Time-Based Scheduling

Outlines

- Basic Concept
- Disk Interface
- Disk Array and RAID
- NAS and SAN
- Flash Storage Device

Drive Interface

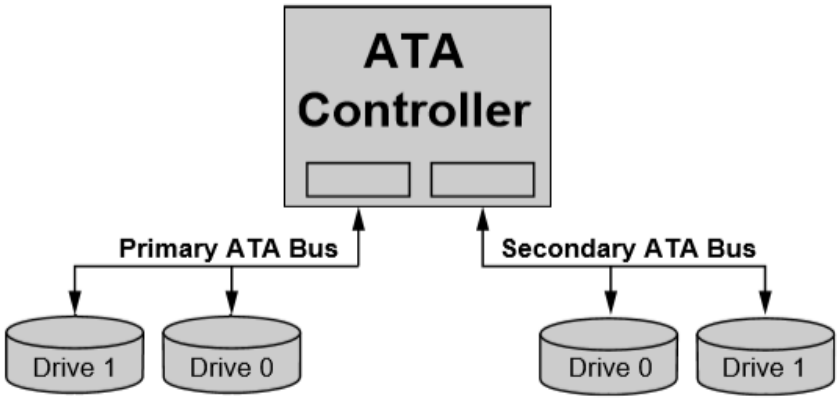


- The drive interface is a bridge between host and the disk
  - The communication channel for I/O requests
  - Allows data transfers for reading and writing

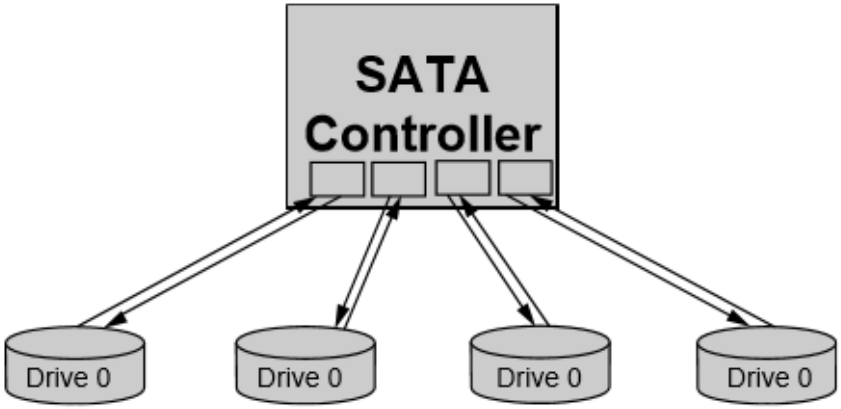
Desirable Characteristics of Drive Interfaces

- **Simple protocol**
  - Fewer handshakes => lower communication overhead
- **High autonomy**
  - Less host processor involvement => lower computation overhead
- **High data rate, up to a point** **Q: what if not?**
  - Higher than media data rate of the drive is desirable
- **Overlapping commands**
  - Support high utilization with multiple connected disk drives
- **Command queueing**
  - Greatly improve disk throughput

Advanced Technology Attachment (ATA)



Configuration of a dual channel PATA controller

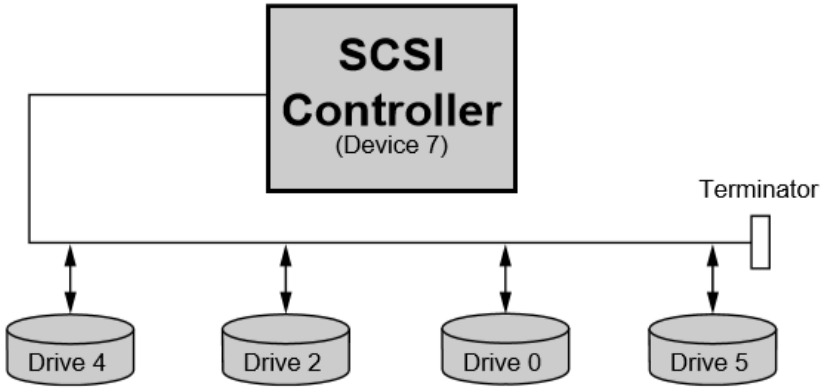


Configuration of a four-port SATA controller

- **Parallel ATA**
  - A parallel interface
  - Typical interfaces for old PCs
  - Less than 18 inch cable length
  - Supports different data transfer modes
- **Serial ATA**
  - Point-to-point interface
  - Dominant in newer systems
  - Up to 1 meter cable length
  - Backward compatible

Small Computer System Interface (SCSI)

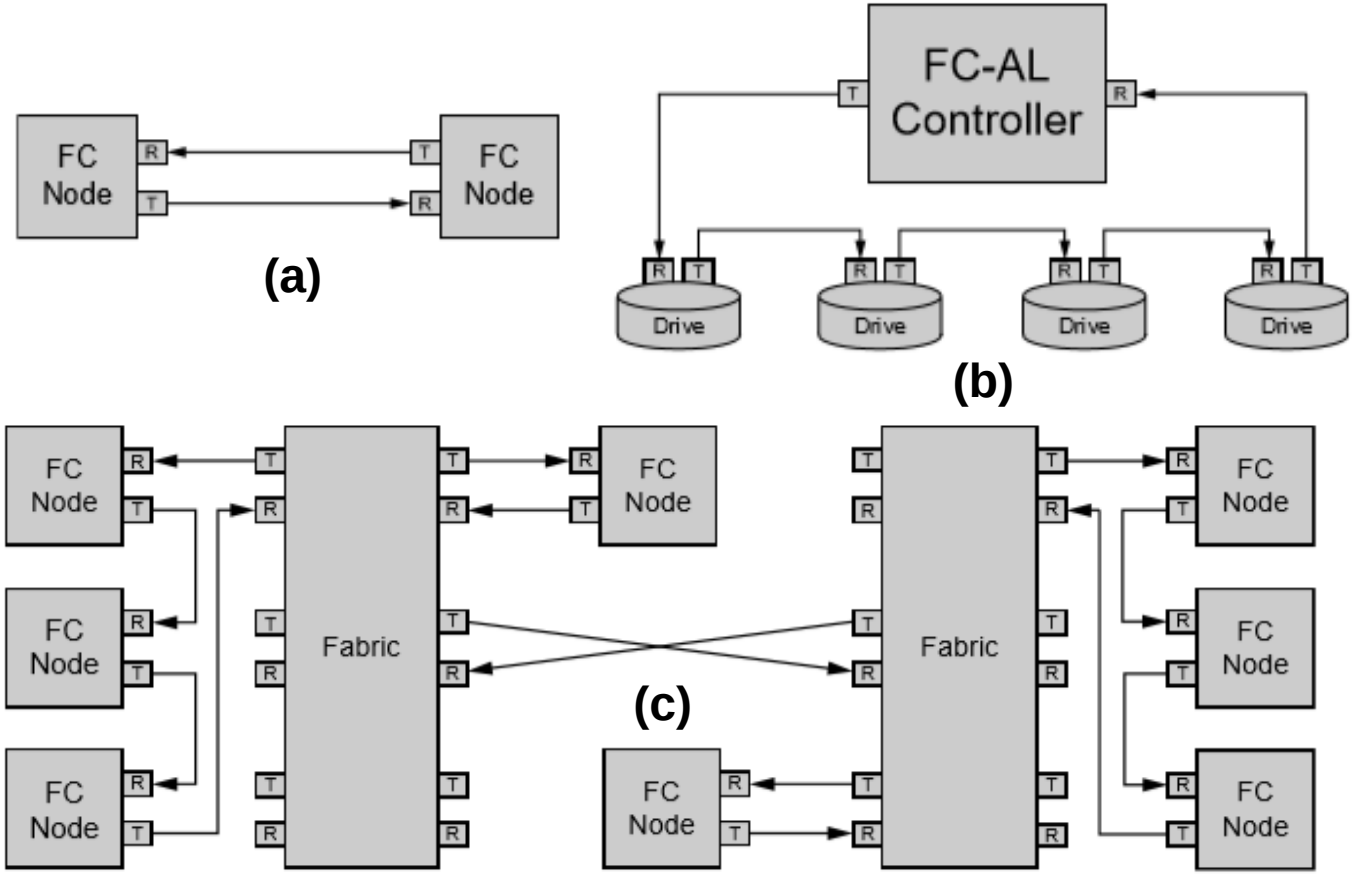
- A more advanced interface with some functionalities and features not available in ATA
- Available in a variety of interfaces: Parallel/Serial



	SATA	Serial Attached SCSI (SAS)
Advantages	Inexpensive, large storage, less power consumption	Fast data rate, higher reliability, longer cables
Application	PC, normal storage	Enterprise, server system

Fibre Channel (FC)

- Fibre Channel is a high-end, feature-rich, serial interface
  - Three topologies: (a) Point-to-point; (b) Arbitrated Loop; (c) Switched Fabric

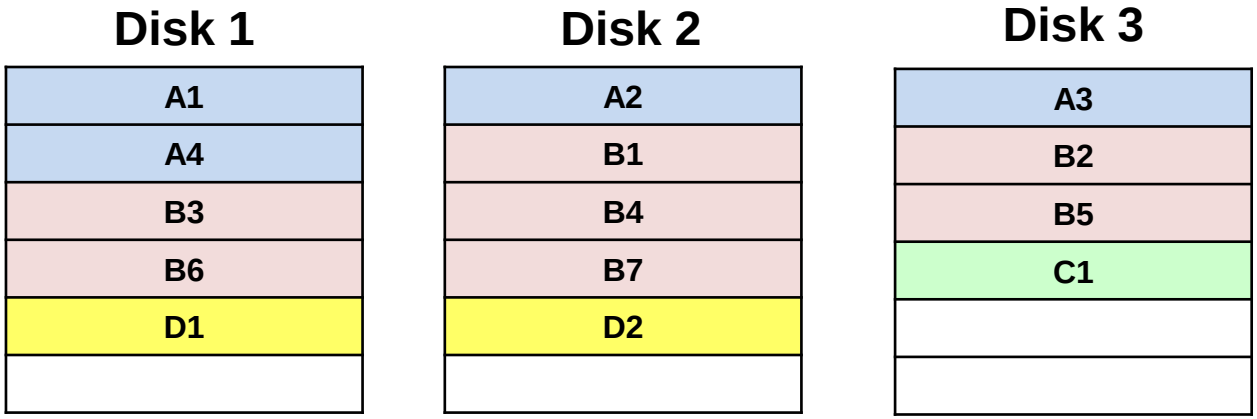


Outlines

- Basic Concept
- Disk Interface
- Disk Array and RAID
- NAS and SAN
- Flash Storage Device

Data Striping

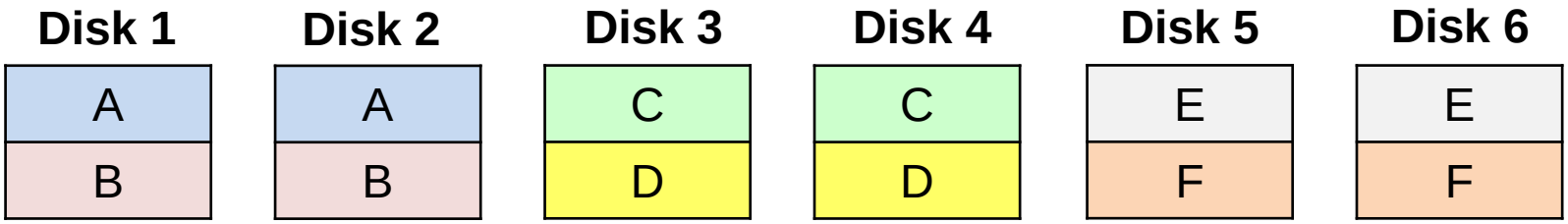
- **JBOD**: Just a bunch of disks
  - A set of disk drives that have no logical relationship in-between
  - Co-located solely for sharing physical resources such as power
- **Data striping** is the technique of segmenting data
  - Stripe factor/width: The number of disks
  - Strip unit: The fixed-sized data block specified
  - Stripe size/depth: The size of a stripe unit



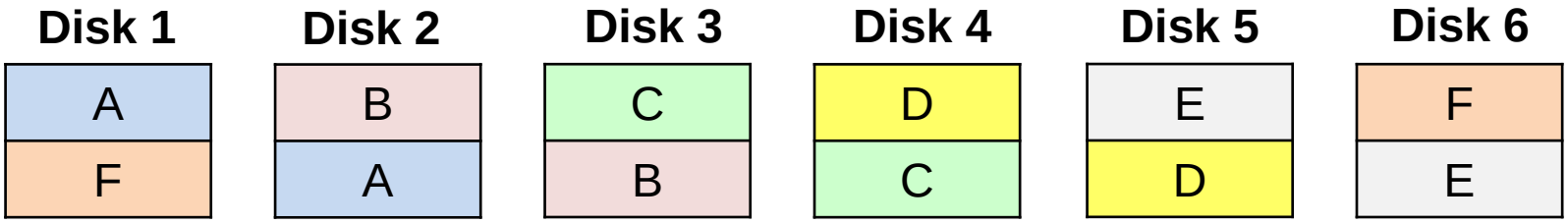
Stripe width = 3. Four user files A, B, C, and D of different sizes are shown

Data Mirroring

Basic mirroring with M=6 drives



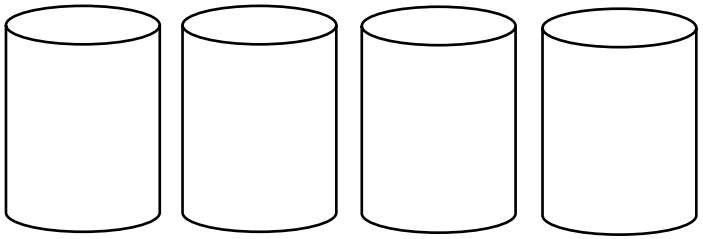
Chained cluster mirroring with N=6 drives



For chained cluster mirroring, M does not have to be an even number, which makes this approach more flexible than the basic mirroring method

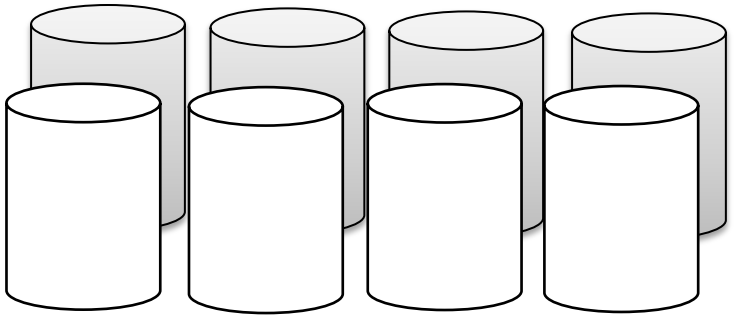
RAID Organization

- Redundant Array of Inexpensive Drives (RAID)
  - Providing fault tolerance in a collection of disk drives



RAID-0

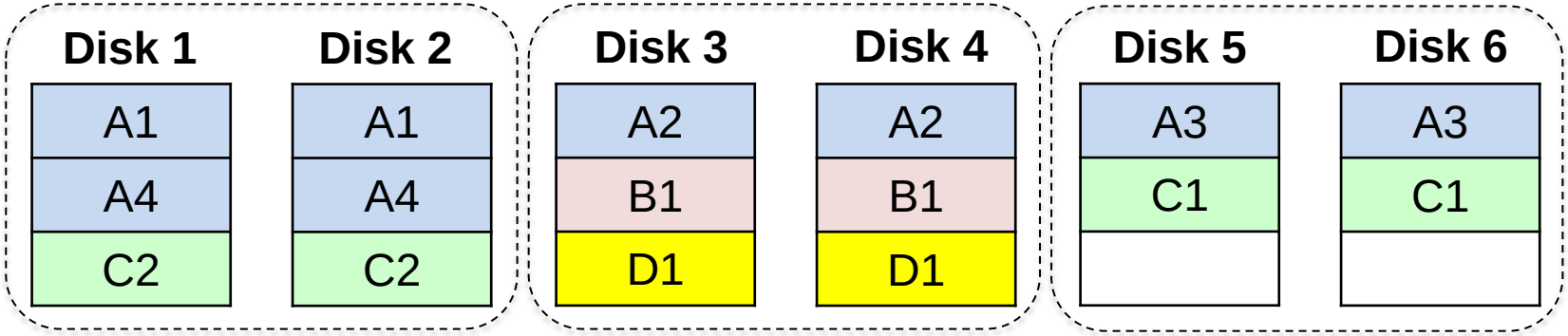
- Simply data striping
- No redundancy
- “marketing hype”



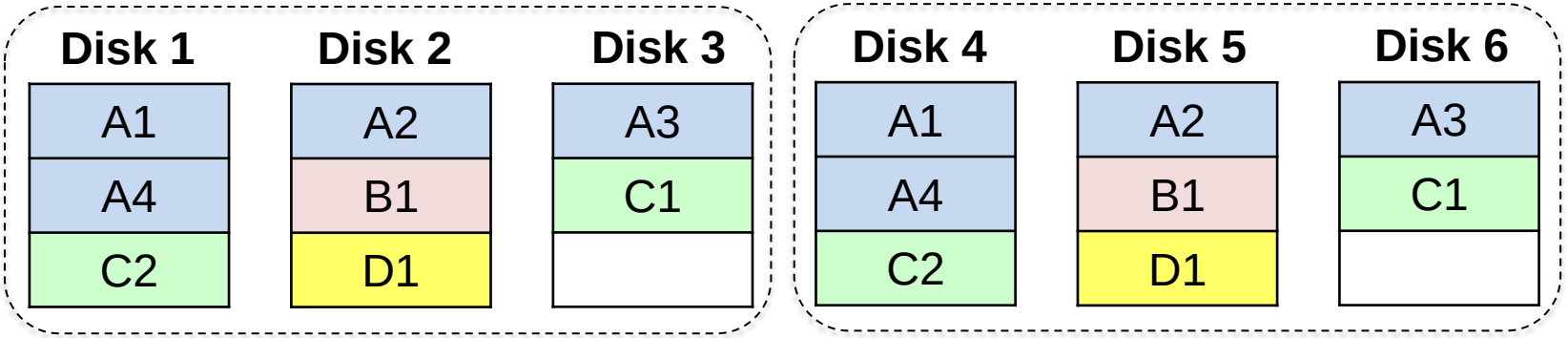
RAID-1

- Basic mirroring
- Most costly solution
- Very simple to implement

RAID 10 vs. RAID 01



RAID 10, a.k.a. RAID 1+0, strip of mirrors

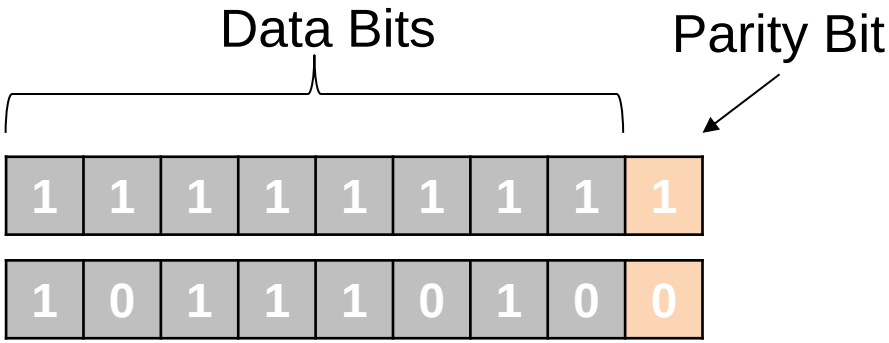


RAID 01, a.k.a. RAID 0+1, mirror of strips

Q: Minimum required number of disks? Comparison of fault tolerance?

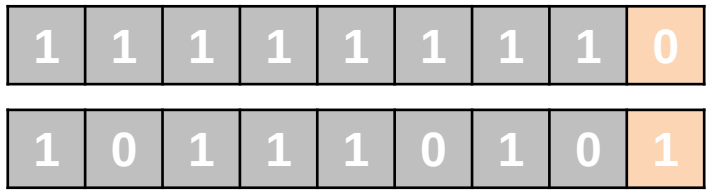
Fault Tolerance

- Data storage requires high data availability
  - **Data replication**: effective and simple way, but expensive
  - **Error correcting coding** (ECC): effective and cost-efficient



Odd Parity

- The parity bit ensures that the total number of 1s to be odd

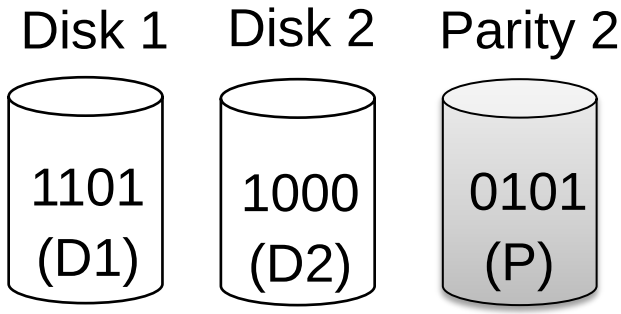


Even Parity

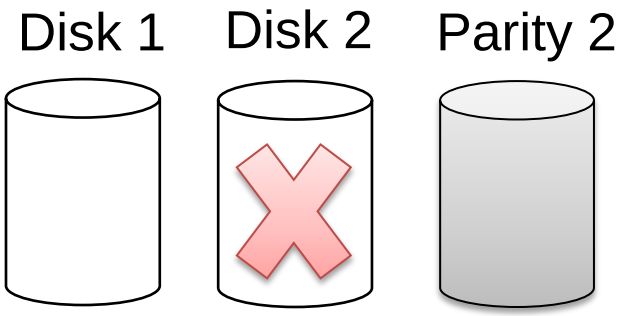
- The parity bit ensures that the total number of 1s to be even

XOR-based Redundancy Scheme

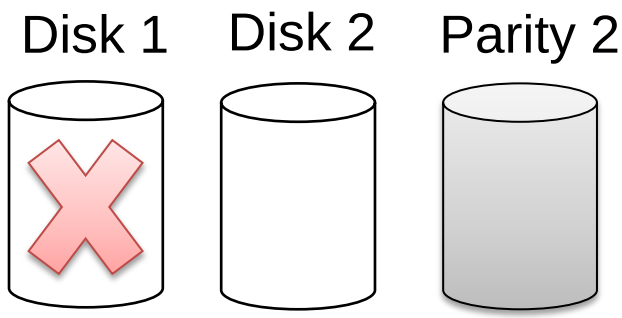
XOR	0	1
0	0	1
1	1	0



P = D1 xor D2



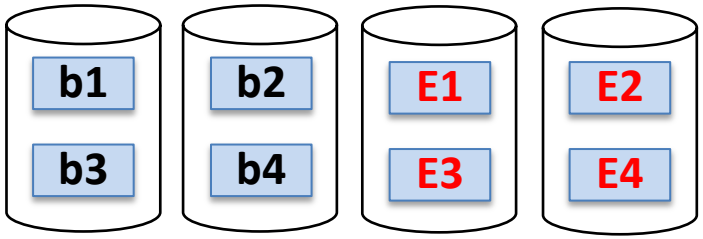
D2 = D1 xor P = 1000



D1 = D2 xor P = 1101

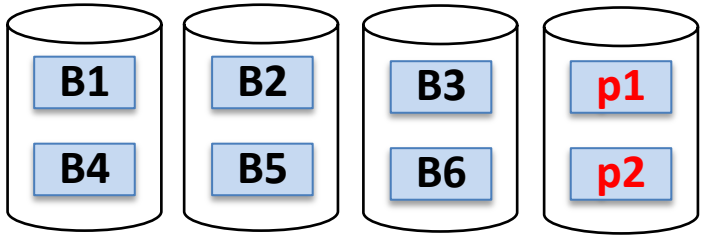
RAID 2, 3, 4, 5, 6

RAID-2



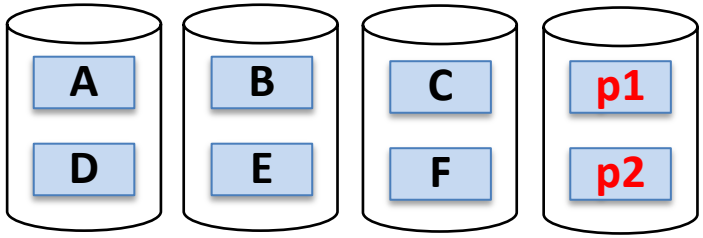
Bit-level striping with dedicated Hamming-code parity (not used by the storage industry)

RAID-3



Byte-level striping with dedicated parity (rarely used by the storage industry)

RAID-4



Block-level striping with dedicated parity (not commonly used)

RAID-5

Block-level striping with distributed parity. Offer a single drive failure protection

RAID-6

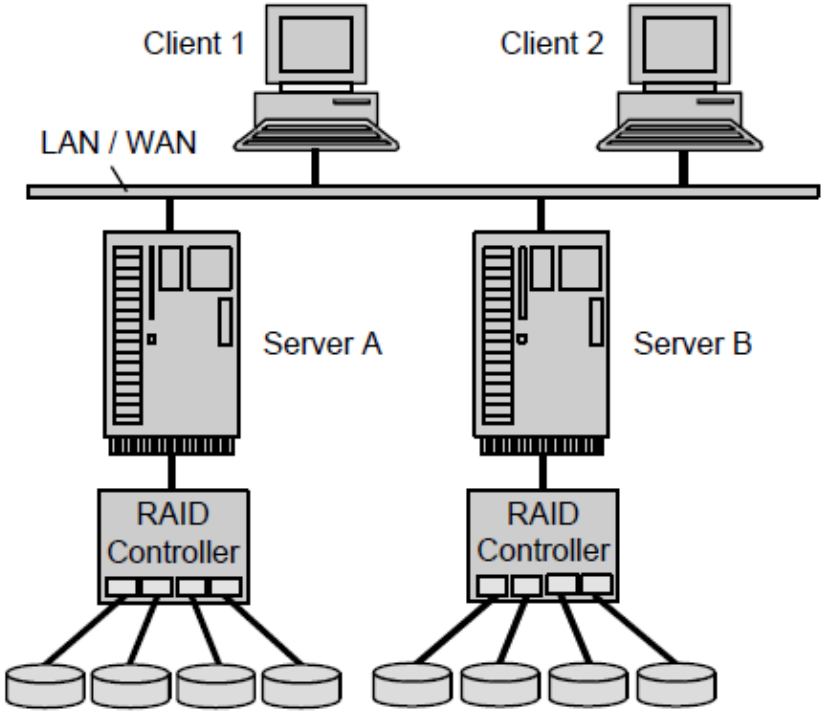
Block-level striping with double distributed parity. Offer a double failure protection

Outlines

- Basic Concept
- Disk Interface
- Disk Array and RAID
- NAS and SAN
- Flash Storage Device

Direct Access Storage (DAS)

- Management of data storage is distributed
- Servers send data over LAN/WAN
- Additional server access over the network



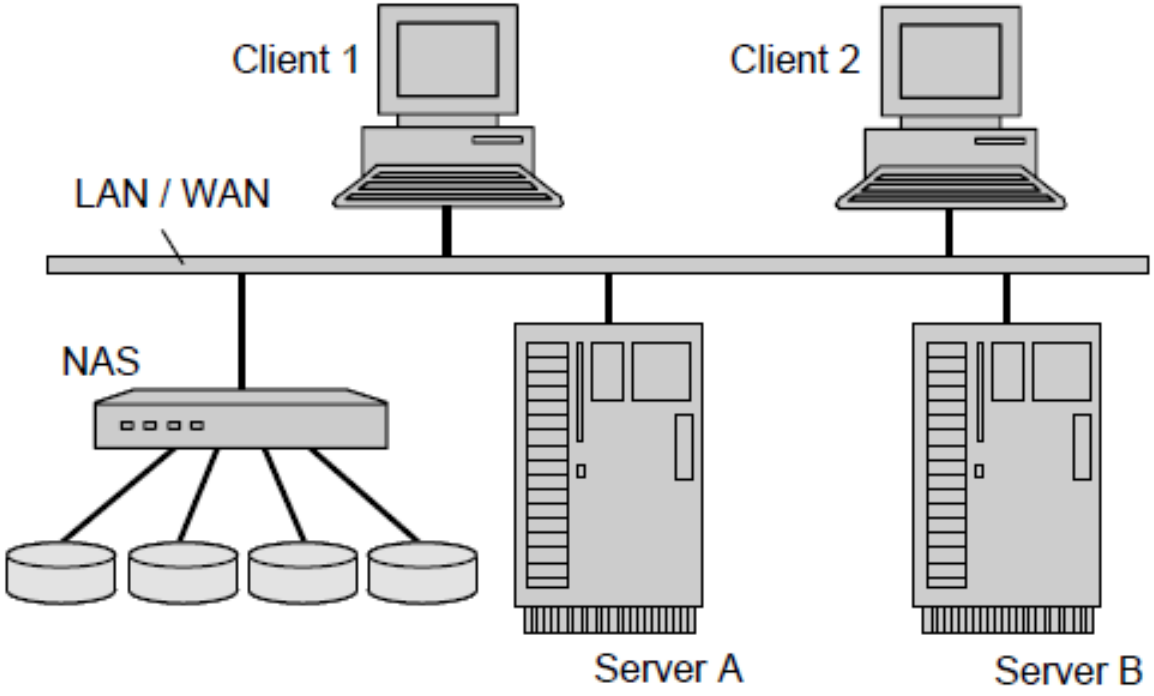
Limitations of DAS

- Accessing data in a different machine on the LAN suffer from poor performance
- Sending bulk data over the LAN/WAN can affect other communications
- If a server was down, its DAS became unavailable to the rest of the system

What are the advantages of DAS?

Network Attached Storage

- NAS is a specialized device
  - Composed of storage, a processor, and an operating system, dedicated to function solely as a file server



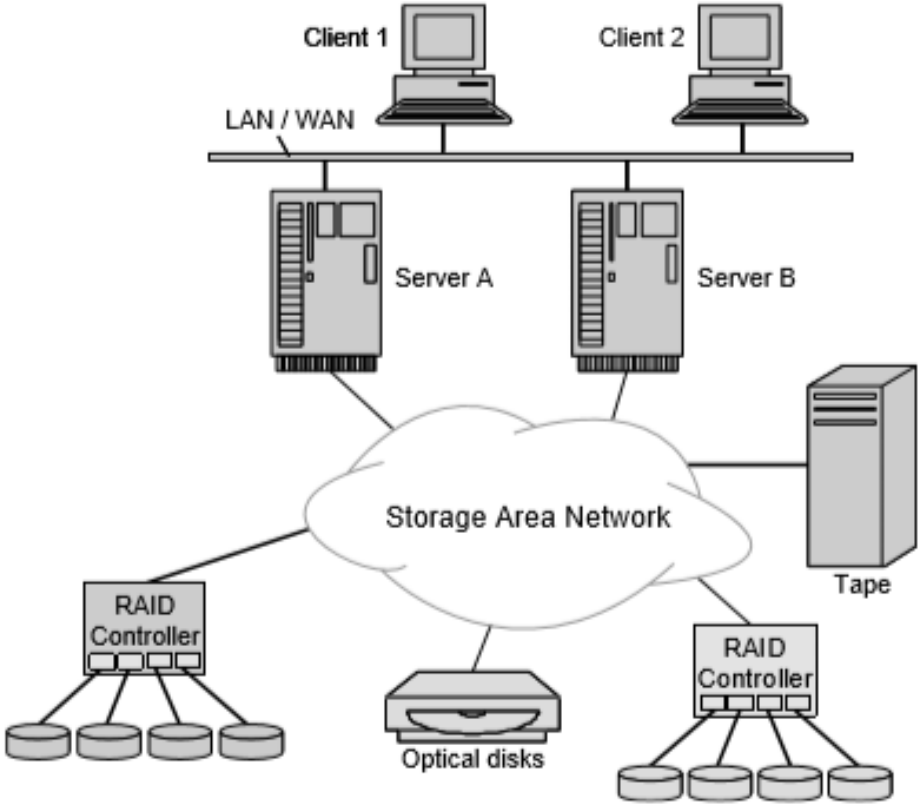
Advantages of NAS

- Economical way of storage sharing
- Easier to setup and configure
- Readily support RAID
- Higher utilization of storage resources

What are the disadvantages of NAS?

Storage Area Network (SAN)

- A vast array of standard storage devices
- Dedicated, high-speed, and scalable backend network
- Decoupling of storage from direct attachment to server



Advantages of SAN

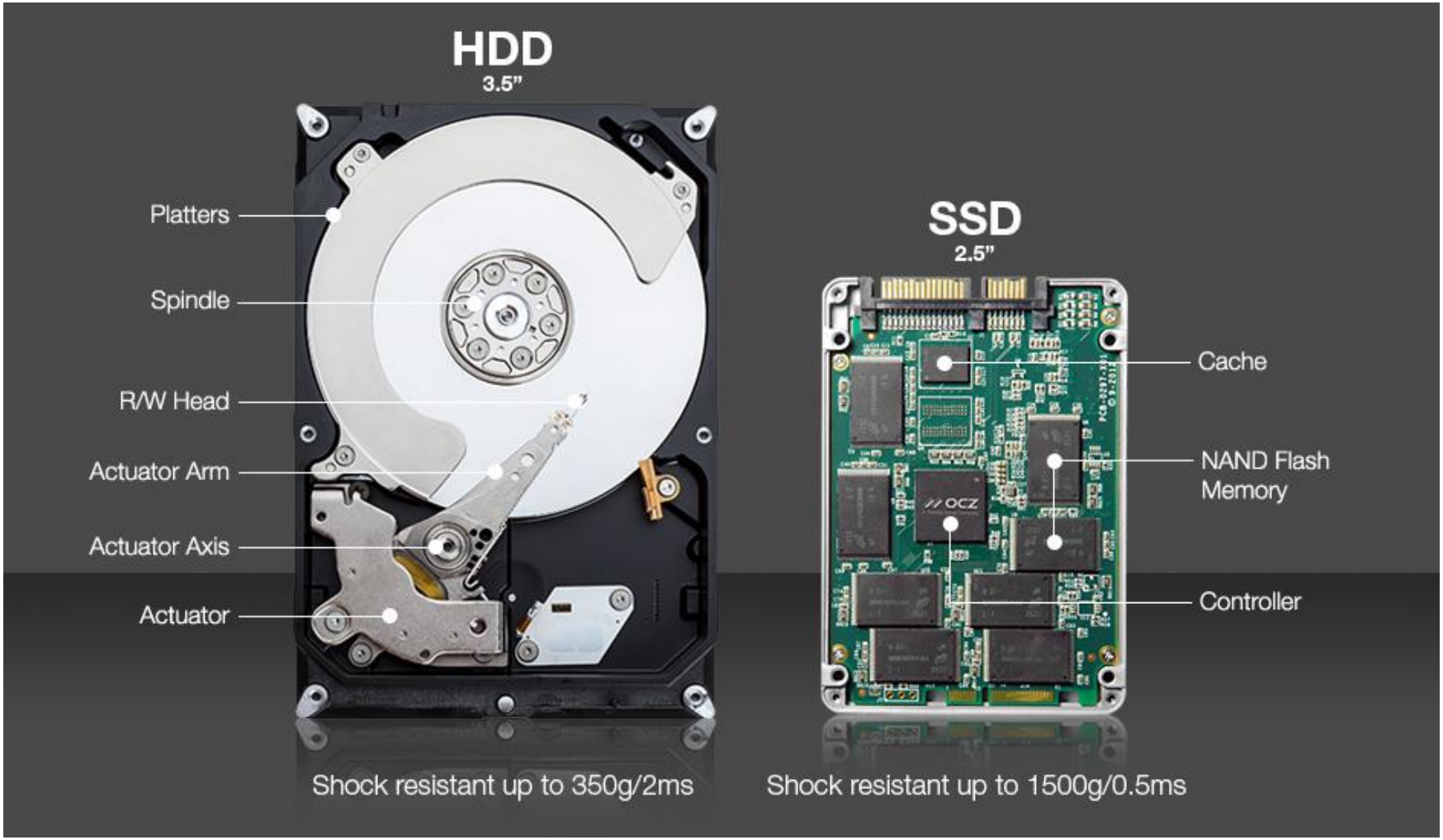
- Saves LAN/WAN bandwidth
- Better data availability
- Maintenance becomes easier
- Support heterogeneous devices
- Readily accept centralized management
- Higher hardware utilization and high performance

	SAN	NAS
Usage Model	Mission-critical data	Serve files
Network	Fibre channel	Ethernet
Data Access	Blocks of data	File level
Cost	Very high	Cost-efficient

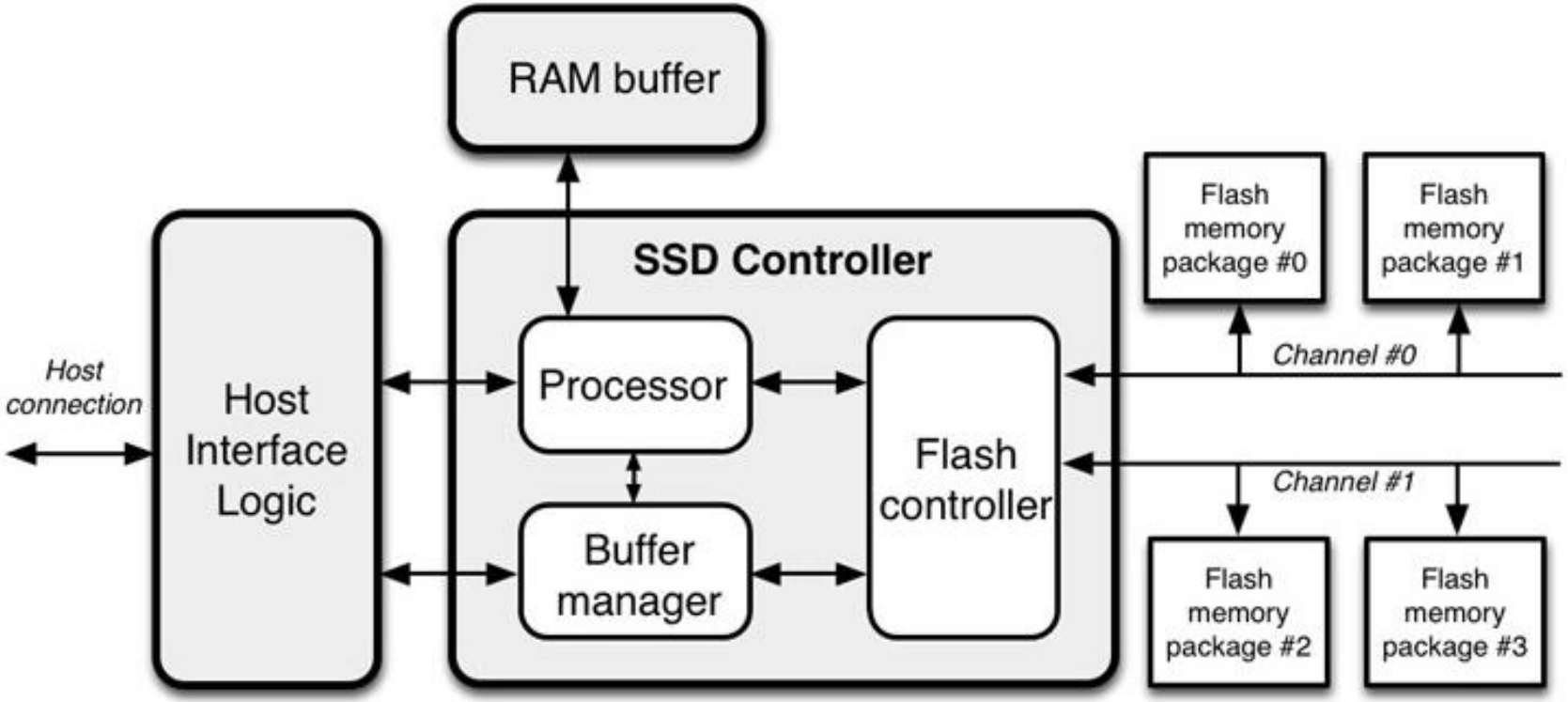
Outlines

- Basic Concept
- Disk Interface
- Disk Array and RAID
- NAS and SAN
- Flash Storage Device

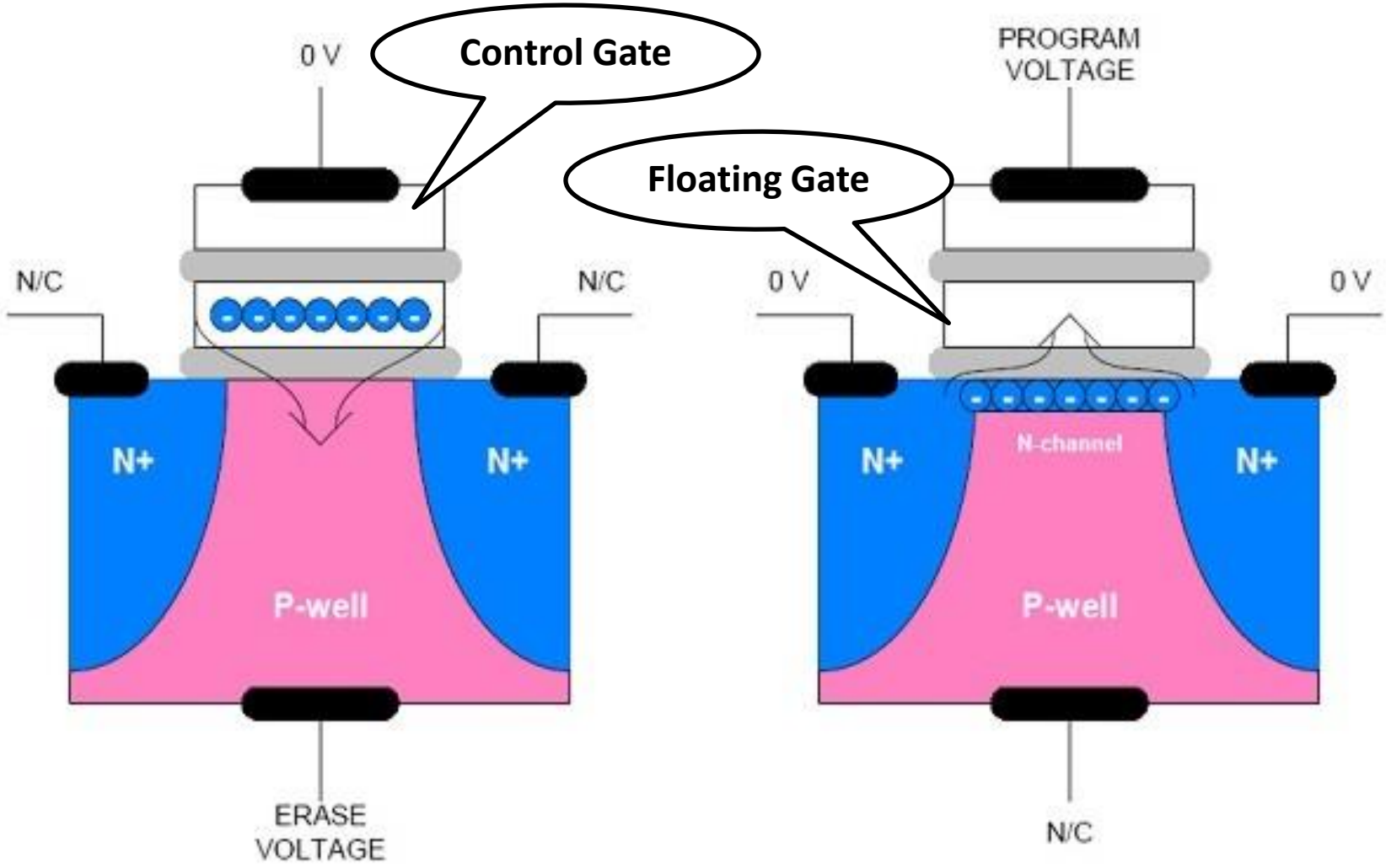
Solid-State Drive (SSD)



Architecture of a SSD

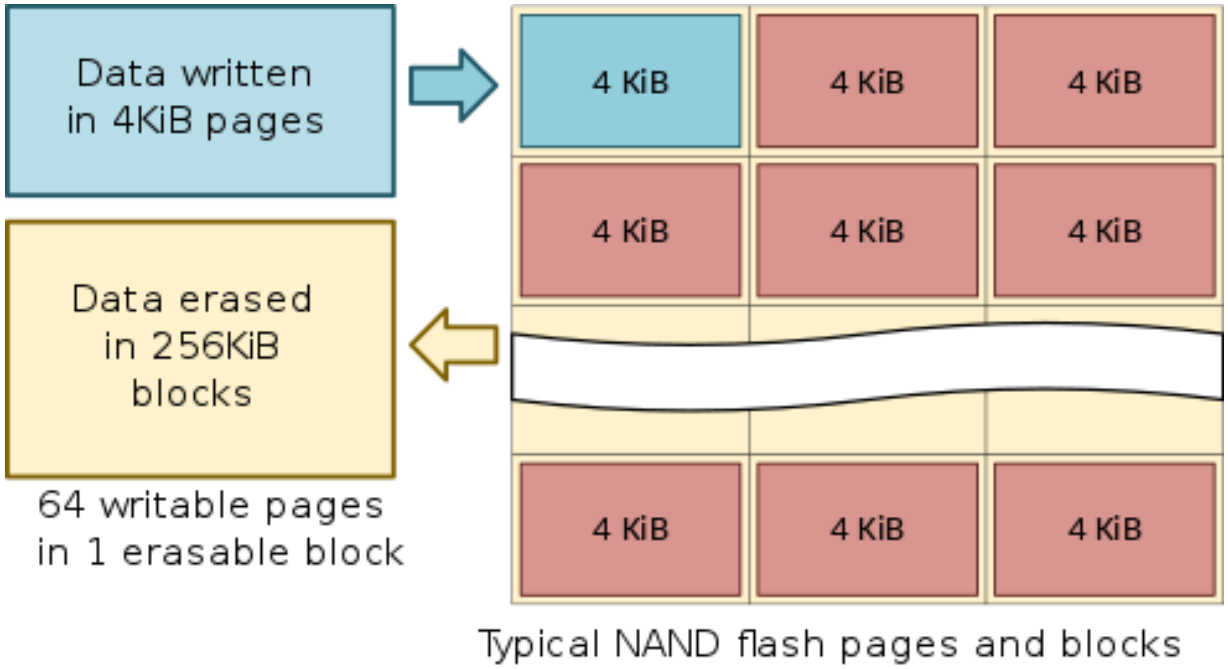


NAND Flash Cell



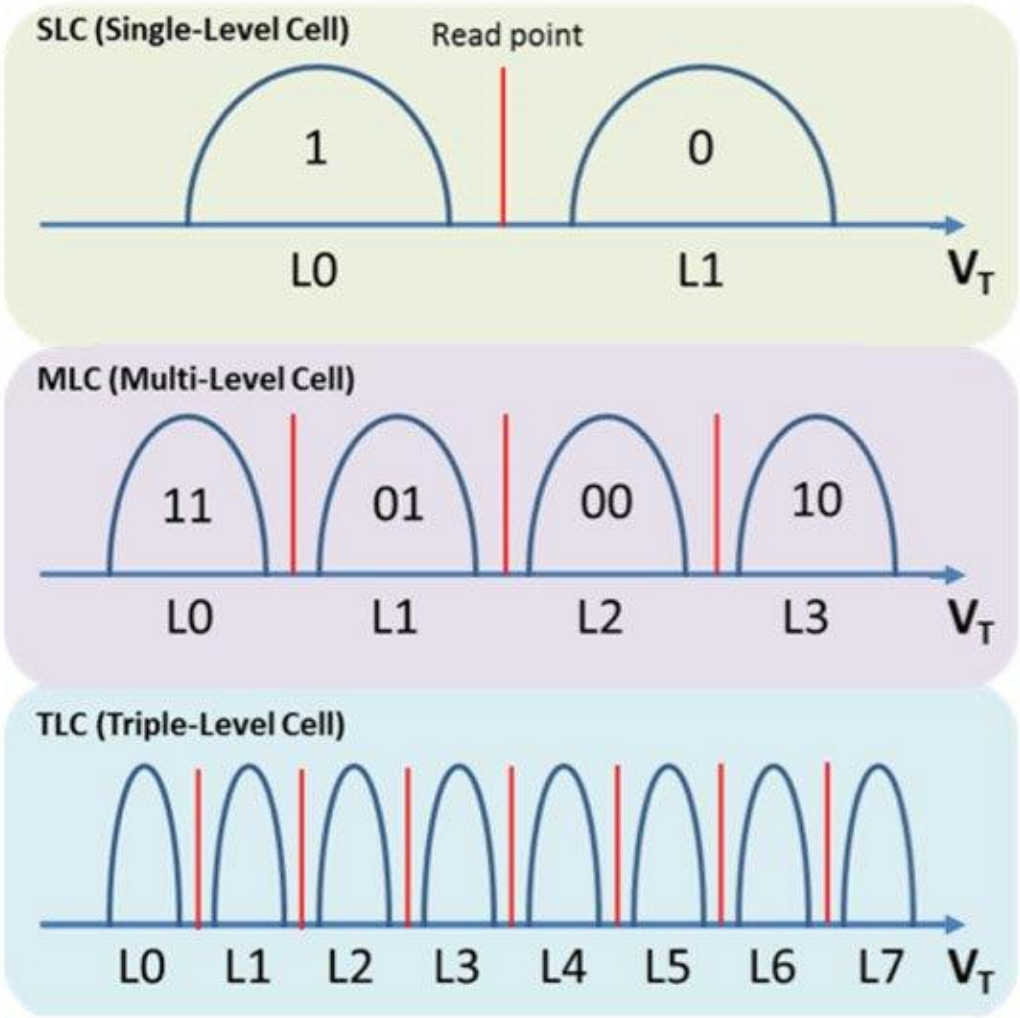
Write Amplification

- Write amplification (WA): the actual amount of information physically written to the storage media is a multiple of the logical amount intended to be written.



Can increase writes on the drive and reduce its life

SLC and MLC



- SLC
  - 1 bit data per cell
  - Higher cost per bit
  - Lower density
  - Lower power cons.
  - Shorter program time
- MLC
  - 2 bits data per cell
  - Lower cost per bit
  - Higher density
  - Higher power cons.
  - Longer program time

Advantages of SSD

- Super low latency:
  - Orders of magnitude less than HDD; zero seek time
- Very fast read and write speed
  - 2700 MB/s (Intel SSD P3700 series)
  - Excel at small/short reads and writes
- Physically more robust
  - Shock resistance
  - Zero moving parts
- Immune to data fragmentation

SSD in Existing Storage System

- Hybrid design: magnetic media + non-volatile cache
  - Improved power management
  - Improved drive reliability
  - Faster boot and loading times
  - Cost efficiency

- Flash-only:
  - Guaranteed high performance
  - Much simpler to manage
  - No mechanical moving parts
  - Capacity and cost issue

Summary

- Disk concept; platter/track/sector
- Design good drive Interfaces
- Parallel/Serial ATA; Parallel/Serial SCSI
- RAID Organization
- DAS, NAS, SAN
- Flash memory cell, SLC/MLC
- SSD advantages, hybrid storage