CISC 3320

C28a Disk Structure and Storage Device Management

Hui Chen

Department of Computer & Information Science

CUNY Brooklyn College

Acknowledgement

- These slides are a revision of the slides provided by the authors of the textbook via the publisher of the textbook

Memory Hierarchy

~1ns	Registers	<KB
~2ns	Cache	~MB
~10ns	Main Memory	~GB
~10ms	Secondary Memory (e.g., Magnetic Disk)	~TB
	Tertiary Memory (e.g., Magnetic Tapes)	

Outline

- Error Detection and Correction
- Disk Structures
- Storage Device Management
- Swap-Space Management
- Storage Attachment

- RAID Structure

Error Detection and Correction

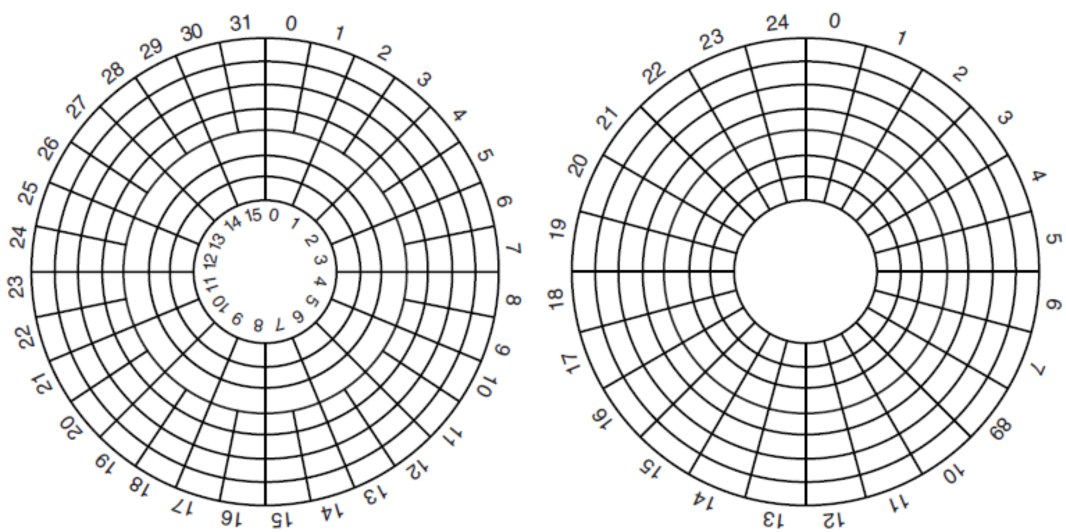
- Fundamental aspect of many parts of computing (memory, networking, storage)
- Error detection determines if there a problem has occurred (for example a bit flipping)
  - If detected, can halt the operation
  - Detection frequently done via parity bit
- Parity one form of checksum – uses modular arithmetic to compute, store, compare values of fixed-length words
  - Another error-detection method common in networking is cyclic redundancy check (CRC) which uses hash function to detect multiple-bit errors
- Error-correction code (ECC) not only detects, but can correct some errors
  - Soft errors correctable, hard errors detected but not corrected

Logical Block Addressing

- Logical blocks
  - To the host, the storage devices are a one-dimensional array of logical blocks
  - Smallest unit of data transfer between the disk and the host
  - Called logical block addressing (LBA)
- Inherited from addressing HDDs

Logical and Physical Disk Geometry

- Logical (virtual) geometry and physical geometry are differ
  - Traditionally, (x, y, z): (cylinders, heads, sectors), i.e., CHS
    - PC: (65535, 16, 63), a sector is typically 512 bytes



(a) (b)

- [Figure 5-19 in Tanenbaum & Bos, 2014] Figure 5-19. (a) Physical geometry of a disk with two zones. (b) A possible virtual geometry for this disk

Disk Structure

- Logical (virtual) geometry and physical geometry are differ
- Traditionally, (x, y, z): (cylinders, heads, sectors), i.e., CHS
  - PC: (65535, 16, 63), a sector is typically 512 bytes
- Modern approach: logical block addressing (LBA), disk sectors numbered consecutively starting at 0
  - A sector is typically 2^9 = 512 bytes
  - A logical block mapped to one or more disk sectors

Block and Sector Mapping

- The 1-dimensional array of logical blocks is mapped into the sectors of the disk sequentially
  - Sector 0 is the first sector of the first track on the outermost cylinder
  - Mapping proceeds in order through that track, then the rest of the tracks in that cylinder, and then through the rest of the cylinders from outermost to innermost
  - Logical to physical address should be easy
    - Except for bad sectors
    - Non-constant # of sectors per track via constant angular velocity

Questions?

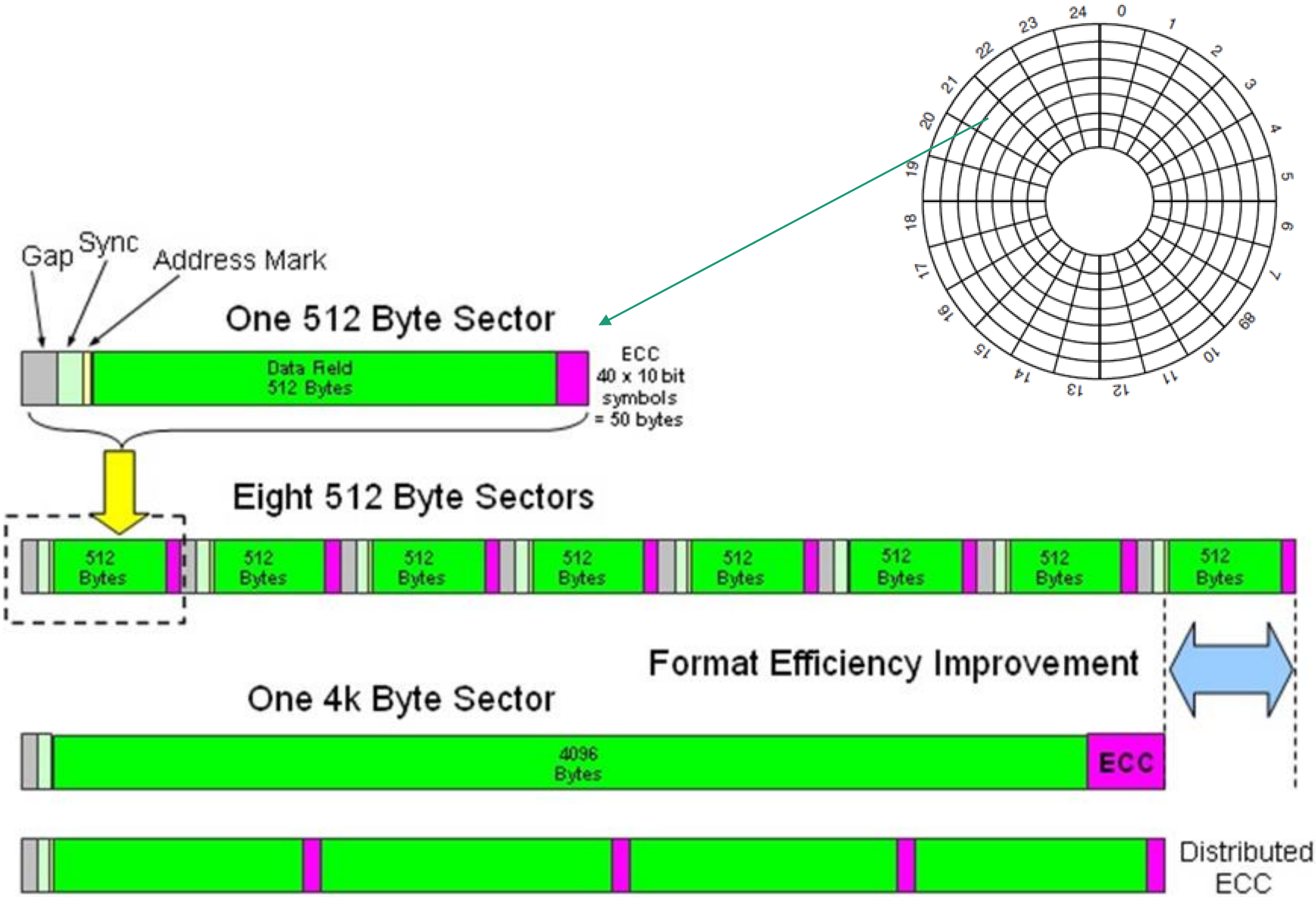
- LBA
- Disk structure
- How about NVM?

Storage Device Management

- Drive Formatting, Partitions, and Volumes
    - Low-level formatting
    - Volume creation and management
    - Partition and logical formatting
    - Boot block

Drive Formatting

- **Low-level formatting**, or **physical formatting**
  - Dividing a disk into sectors that the disk controller can read and write
  - Low-level formatting creates **logical blocks** on physical media
  - Each sector can hold header information, plus data, plus error correction code (**ECC**)
  - Usually 512 bytes of data but can be selectable

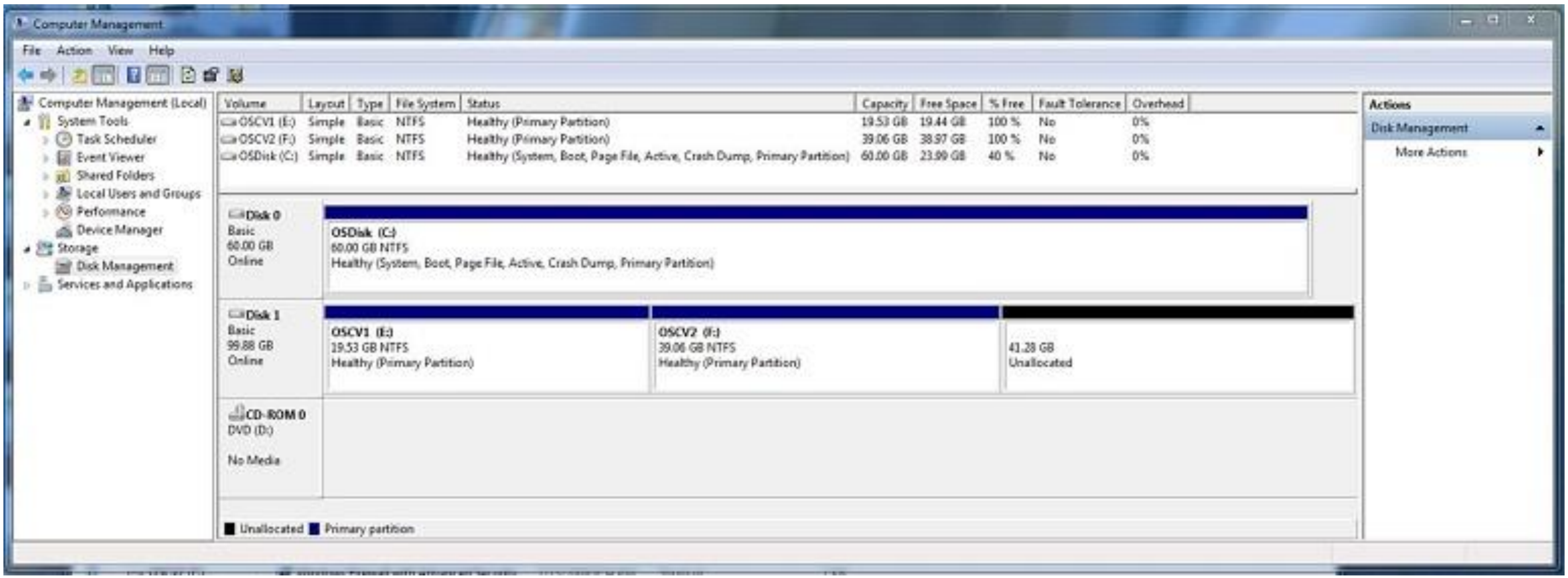


Gap
Sync
Address Mark
One 512 Byte Sector
Data Field
512 Bytes
ECC
40 x 10 bit
symbols
= 50 bytes
Eight 512 Byte Sectors
512 Bytes
512 Bytes
512 Bytes
512 Bytes
512 Bytes
512 Bytes
512 Bytes
512 Bytes
Format Efficiency Improvement
One 4k Byte Sector
4096
Bytes
ECC
Distributed
ECC

Partition and Logical Formatting

- To use a disk to hold files, the operating system still needs to record its own data structures on the disk
- Partition the disk into one or more groups of sectors/cylinders, each treated as a logical disk
- Logical formatting or "making a file system"
- To increase efficiency most file systems group blocks into clusters
  - Disk I/O done in blocks
  - File I/O done in clusters

Windows



Windows Disk Management tool showing devices, partitions, volumes, and file systems

Linux

cat /proc/diskstats

fdisk –l /dev/sda

sfdisk -d /dev/sda

Root and Non-Root Partitions

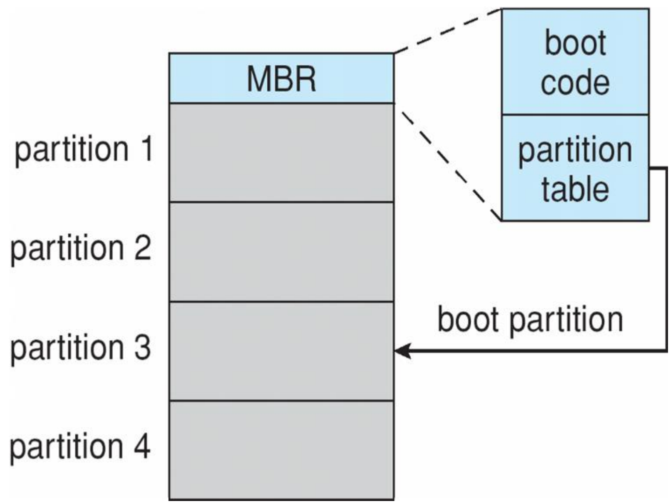
- Root partition contains the OS
  - Mounted at boot time
    - At mount time, file system consistency checked
    - Is all metadata correct?
      - If not, fix it, try again
      - If yes, add to mount table, allow access
- Other partitions can hold other OSes, other file systems, or be raw
  - Other partitions can mount automatically or manually

Boot Block

- Boot block can point to boot volume or boot loader set of blocks
  - that contain enough code to know how to load the kernel from the file system
- Or point to a boot management program for multi-OS booting

Boot Block Initialization

- Boot block initializes system
  - The bootstrap is stored in ROM, firmware
  - **Bootstrap loader** program stored in boot blocks of boot partition



Booting from secondary storage in Windows

Linux Grub MBR

- On a Debian Linux system, assume 32bit architecture,

```
# apt-get install nasm
# cat /proc/diskstats
# dd if=/dev/sda of=mbr.bin bs=512 count=1
# hextedit mbr.bin
# ndisasm -b16 -o7c00h mbr.bin > mbr.asm
# vi mbr.asm
```

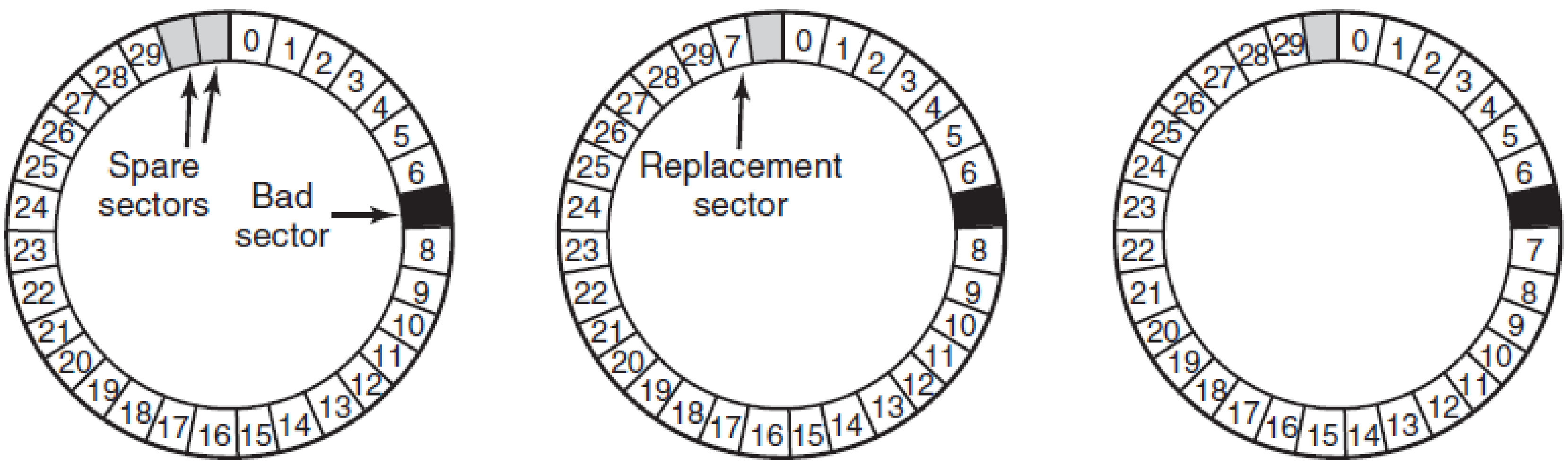
- Partition table starts at offset 446
- For more see https://thestarman.pcministry.com/asm/mbr/GRUB.htm

Bad Blocks

- Methods such as **sector sparing** used to handle bad blocks
  - A typical bad-sector transaction might be as follows:
    1. The operating system tries to read logical block 87.
    2. The controller calculates the ECC and finds that the sector is bad. It reports this finding to the operating system as an I/O error.
    3. The device controller replaces the bad sector with a spare.
    4. After that, whenever the system requests logical block 87, the request is translated into the replacement sector's address by the controller.
- Alternatively, some controllers can be instructed to replace a bad block by **sector slipping**

Spare Substitution

- Two approaches shown in (b) and (c)



(a) (b) (c)

- [Figure 5-26 in Tanenbaum & Bos, 2014]

Raw Disk

- Raw disk access for apps that want to do their own block management
  - Keep OS out of the way
  - Example: databases

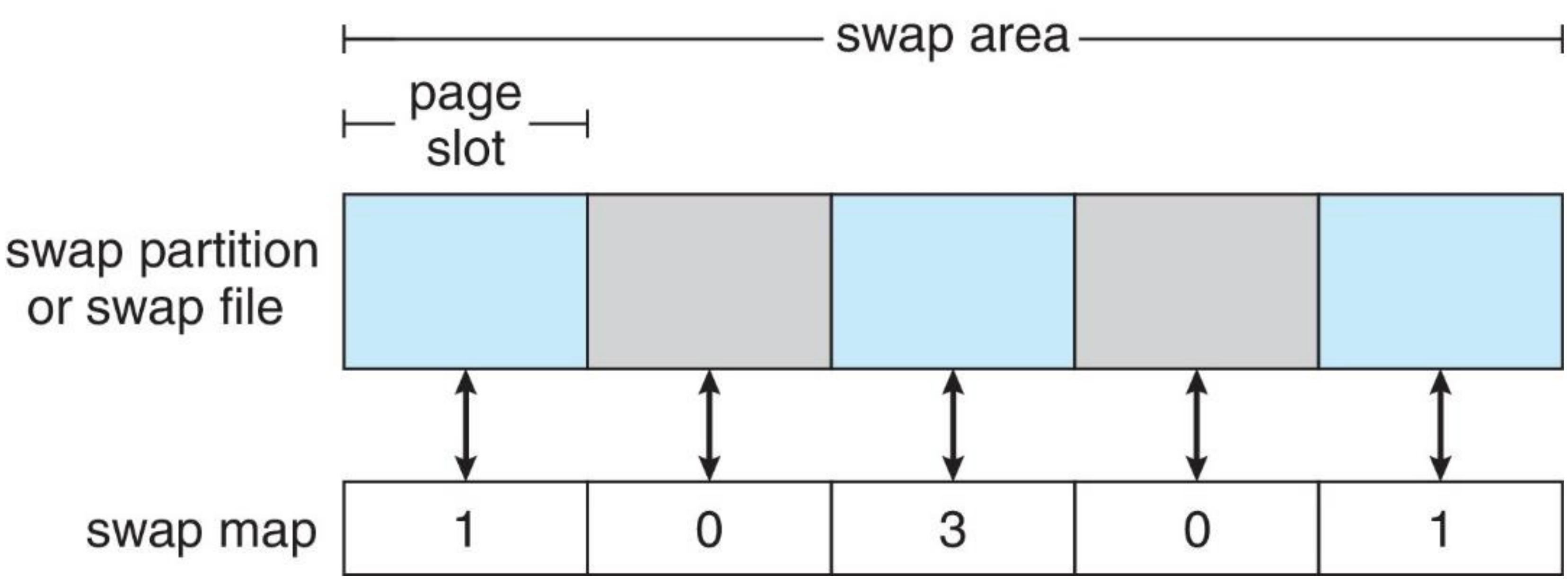
Questions?

- Drive Formatting, Partitions, and Volumes
  - Low-level formatting
  - Volume creation and management
  - Partition and logical formatting
  - Boot block
  - Raw disk

Swap-Space Management

- Used for moving entire processes (swapping), or pages (paging), from DRAM to secondary storage when DRAM not large enough for all processes
- Operating system provides **swap space management**
  - Secondary storage slower than DRAM, so important to optimize performance
  - Usually multiple swap spaces possible – decreasing I/O load on any given device
  - Best to have dedicated devices
  - Can be in raw partition or a file within a file system (for convenience of adding)

Data Structures for Swapping on Linux Systems



Questions?

- Swap-space management?

Storage Attachment

- Computers access storage in three ways
  - host-attached
  - network-attached
  - Cloud
- Storage arrays
- Storage array network

Host-Attached Storage

- Host-attached storage accessed through I/O ports talking to I/O buses
- Several busses available, including
  - advanced technology attachment (ATA),
  - serial ATA (SATA, most common)
  - eSATA,
  - serial attached SCSI (SAS),
  - universal serial bus (USB),
  - thunderbolt, and
  - fibre channel (FC), for high-end systems
    - serial architecture using fibre or copper cables
    - Multiple hosts and storage devices can connect to the FC fabric

NVMe

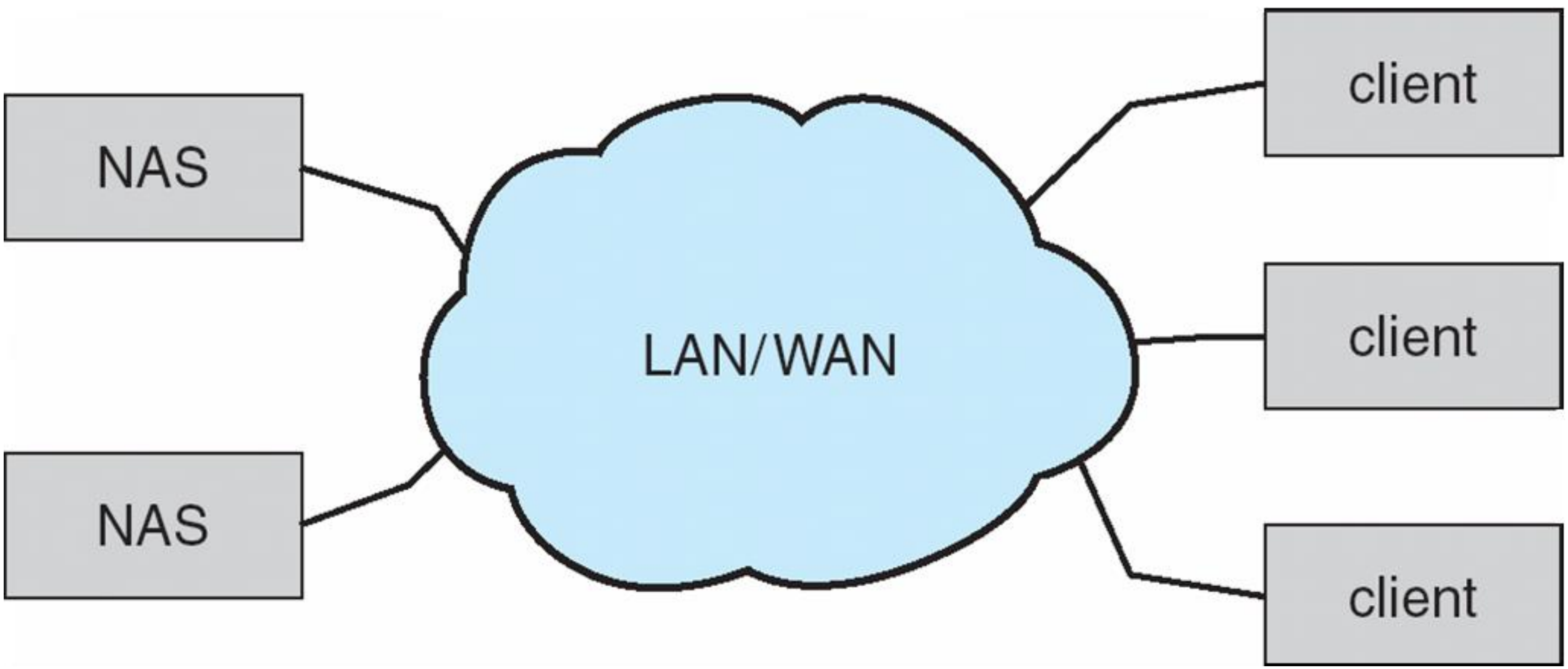
- NVM much faster than HDD
  - New fast interface for NVM called **NVM express** (**NVMe**), connecting directly to PCI bus

Host-Bus Adapters (HBAs)

- Data transfers on a bus carried out by special electronic processors called **controllers** (or **host-bus adapters**, **HBAs**)
  - Host controller on the computer end of the bus, device controller on device end
  - Computer places command on host controller, using memory-mapped I/O ports
    - Host controller sends messages to device controller
    - Data transferred via DMA between device and computer DRAM

Network-Attached Storage

- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
  - Remotely attaching to file systems
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage over typically TCP or UDP on IP network
- **iSCSI** protocol uses IP network to carry the SCSI protocol
  - Remotely attaching to devices (blocks)


NAS
NAS
LAN/WAN
client
client
client

Cloud Storage

- Similar to NAS, provides access to storage across a network
  - Unlike NAS, accessed over the Internet or a WAN to remote data center
- NAS presented as just another file system, while cloud storage is API based, with programs using the APIs to provide access
  - Examples include Dropbox, Amazon S3, Microsoft OneDrive, Apple iCloud
  - Use APIs because of latency and failure scenarios (NAS protocols wouldn't work well)

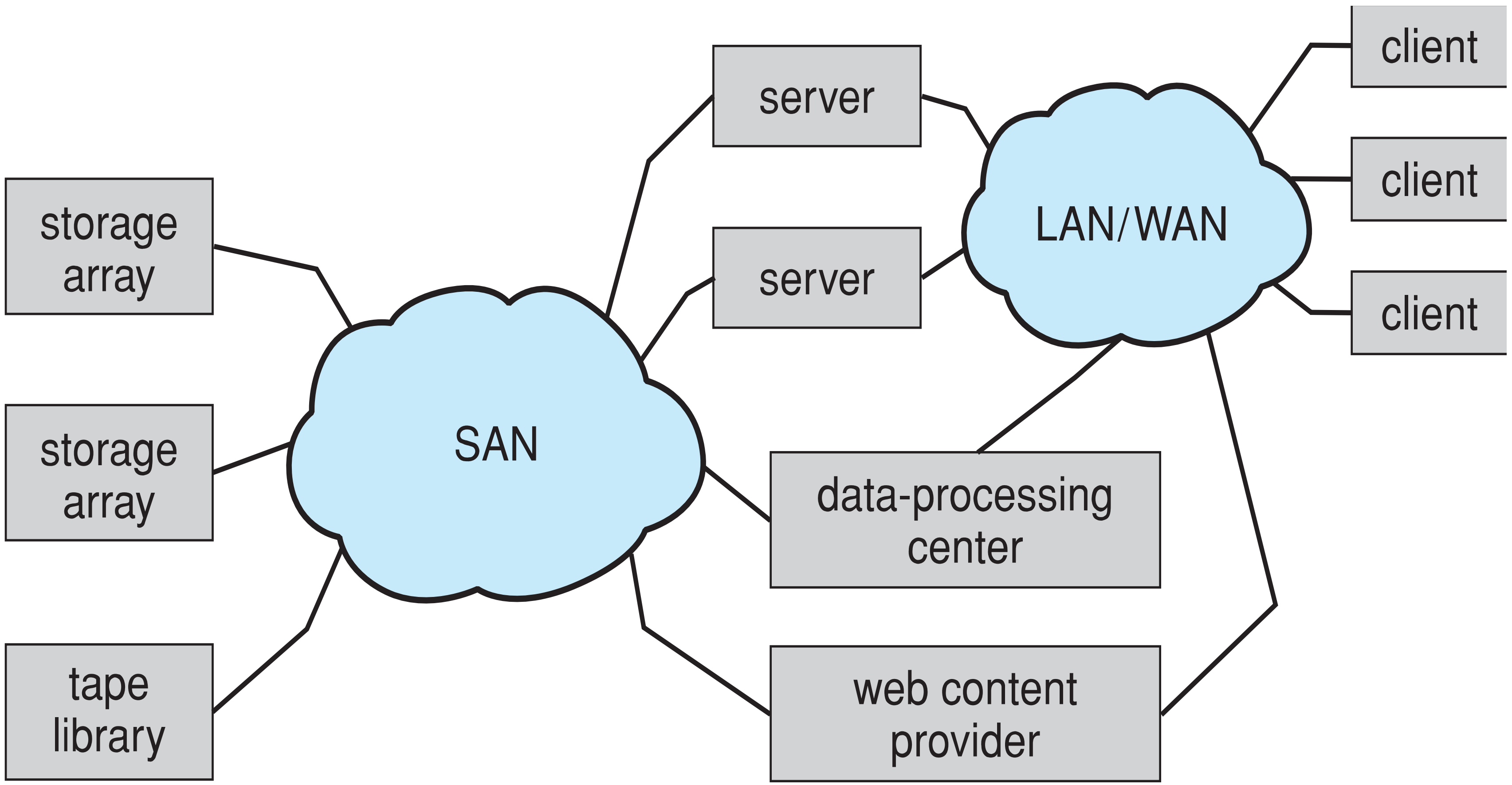
Storage Array

- Can just attach disks, or arrays of disks
- Avoids the NAS drawback of using network bandwidth
- Storage Array has controller(s), provides features to attached host(s)
  - Ports to connect hosts to array
  - Memory, controlling software (sometimes NVRAM, etc)
  - A few to thousands of disks
  - RAID, hot spares, hot swap (discussed later)
  - Shared storage -> more efficiency
  - Features found in some file systems
    - Snaphots, clones, thin provisioning, replication, deduplication, etc

A Storage Array

Storage Area Network

- Common in large storage environments
- Multiple hosts attached to multiple storage arrays
  - Connected to one or more Fibre Channel switches or **InfiniBand** (**IB**) network
- Hosts also attach to the switches
- Storage made available via **LUN Masking** from specific arrays to specific servers
- Easy to add or remove storage, add new host and allocate it storage
- Why have separate storage networks and communications networks?
  - Consider iSCSI, Fibre Channel (FC), Fibre Channel over Ethernet (FCOE), InifiniBand (IB)


client
client
client
server
server
LAN/WAN
storage array
storage array
SAN
tape library
data-processing center
web content provider

Questions?

- Computers access storage in three ways
  - host-attached
  - network-attached
  - Cloud
- Storage arrays
- Storage array network