OPERATING SYSTEMS

IO SYSTEMS

Jerry Breecher

IO SYSTEMS

This material covers Silberschatz Chapters 12 and 13.

Mass Storage - hardware

This is about Disk Behavior and Management.

- Disk Characteristics
- Space Management
- RAID
- Disk Attachment

IO Interface – how the OS interfaces to the hardware

The busses in the computer and how the O.S. interfaces to it.

- Talking to the IO – Polling, Interrupts and DMA
- Application IO Interface
- Kernel IO Subsystem

Mass-Storage Structure

Disk Characteristics

- A disk can be viewed as an array of blocks. In fact, a file system will want to view it at that logical level.
- However, there's a mapping scheme from logical block address B, to physical address (represented by a track / sector pair.)
- The smallest storage allocation is a block - nothing smaller can be placed on the disk. This results in unused space (internal fragmentation) on the disk, since quite often the data being placed on the disk doesn't need a whole block.

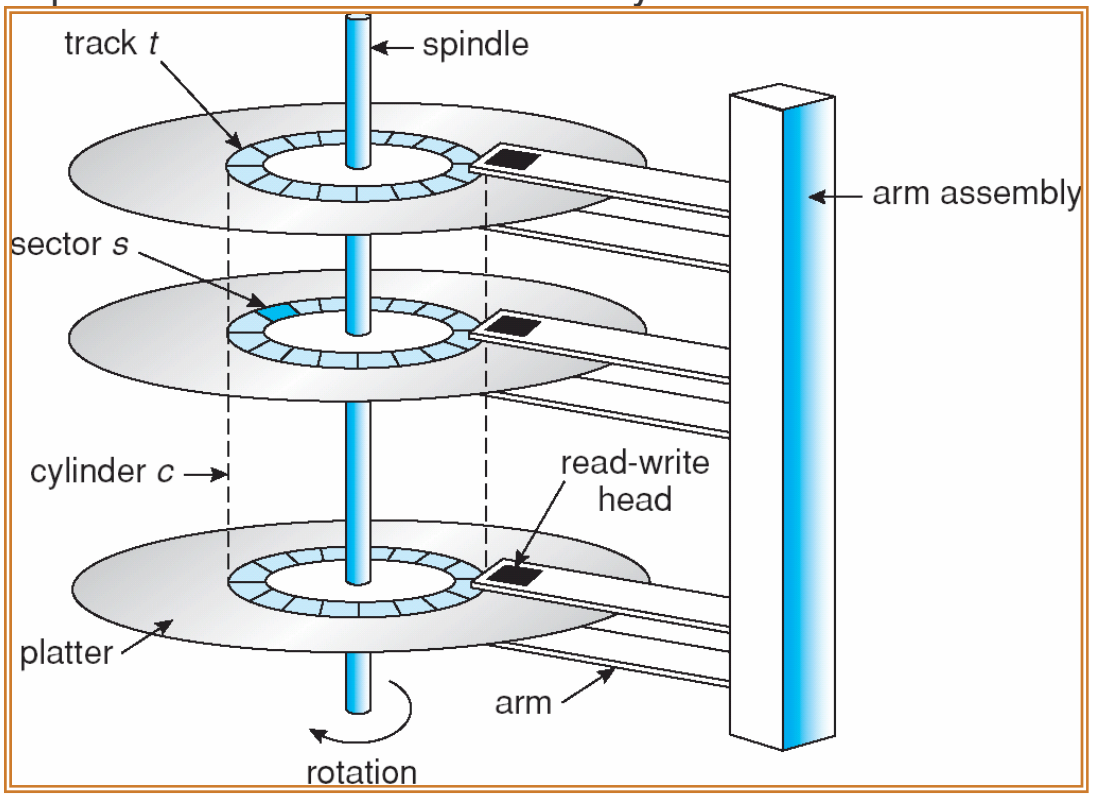
Mass-Storage Structure

Disk Scheduling

The components making up disk service time include:

time = setup + seek + rotation time + transfer + wrap-up

The methods discussed below try to optimize seek time but make no attempt to account for the total time. The ideal method would optimize the total time and many controllers are now able to accomplish this.



Mass-Storage Structure

Disk Management

Disk formatting — Creates a logical disk from the raw disk. Includes setting aside chunks of the disk for booting, bad blocks, etc. Also provides information needed by the driver to understand its positioning.

Boot block — That location on the disk that is accessed when trying to boot the operating system. It's a well-known location that contains the code that understands how to get at the operating system - generally this code has a rudimentary knowledge of the file system.

Bad blocks — The driver knows how to compensate for a bad block on the disk. It does this by putting a pointer, at the location of the bad block, indicating where a good copy of the data can be found.

Swap Space Management — The Operating System requires a contiguous space where it knows that disk blocks have been reserved for paging. This space is needed because a program can't be given unshared memory unless there's a backing store location for that memory.

Mass-Storage Structure

Disk Attachment

Host-attached storage

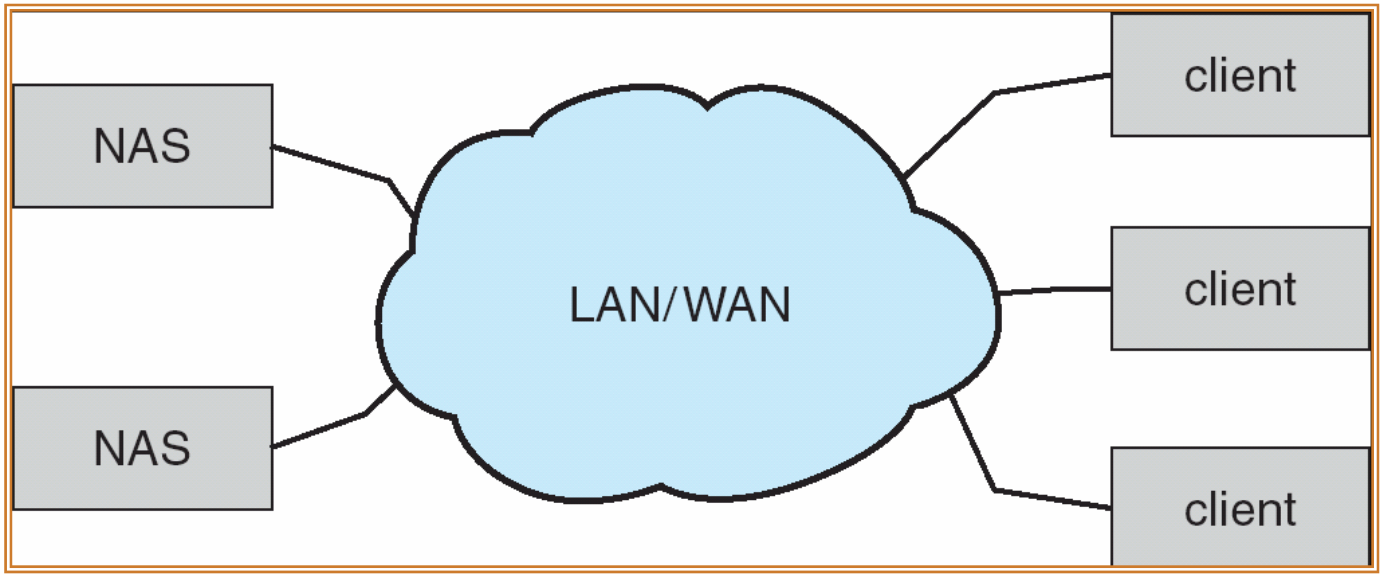
- accessed through I/O ports talking to I/O busses
- SCSI itself is a bus, up to 16 devices on one cable, **SCSI initiator** requests operation and **SCSI targets** perform tasks
  - Each target can have up to 8 **logical units** (disks attached to device controller
- Fibre Channel (FC) is high-speed serial architecture
  - Can be switched fabric with 24-bit address space – the basis of **storage area networks** (**SAN**s) in which many hosts attach to many storage units

Mass-Storage Structure

Disk Attachment

Network-attached storage

- Network-attached storage (**NAS**) is storage made available over a network rather than over a local connection (such as a bus)
- NFS and CIFS are common protocols
- Implemented via remote procedure calls (RPCs) between host and storage
- New iSCSI protocol uses IP network to carry the SCSI protocol

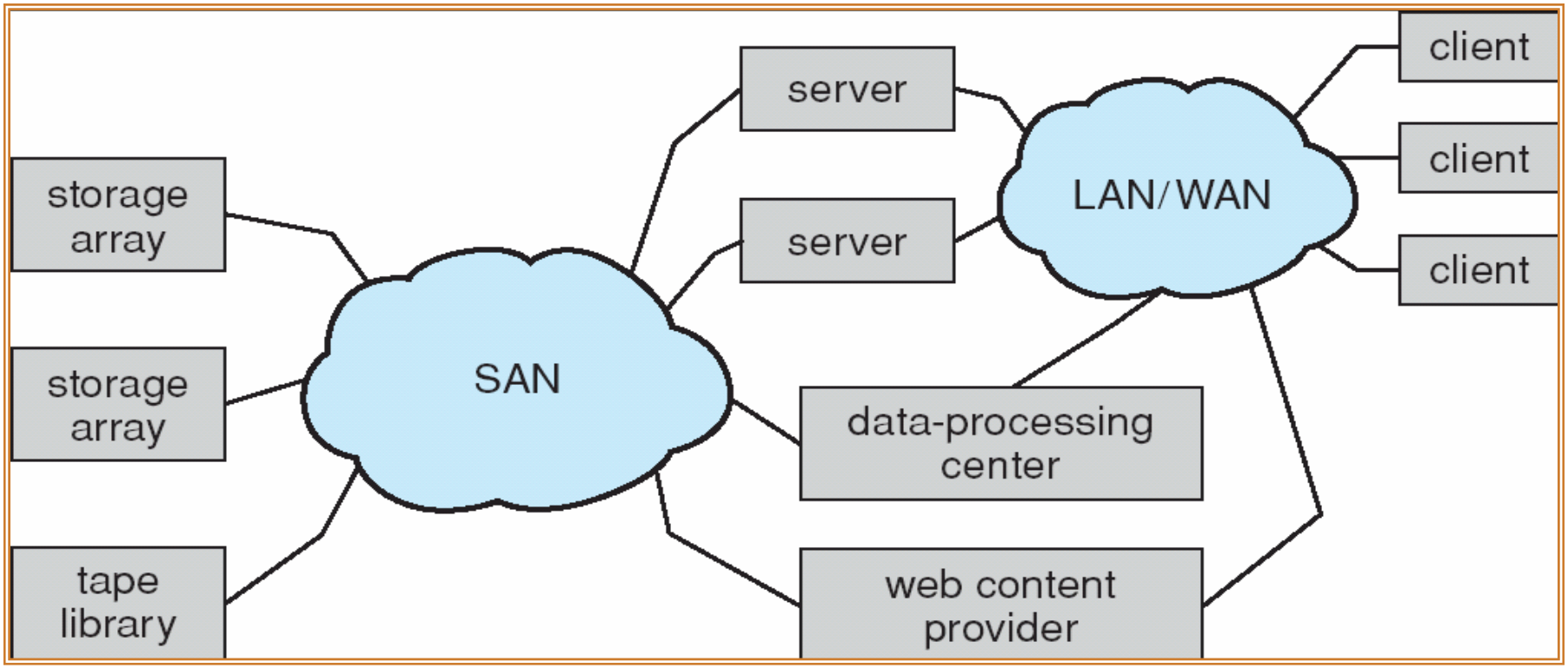


Mass-Storage Structure

Disk Attachment

Storage-Area Network

- Common in large storage environments (and becoming more common)
- Multiple hosts attached to multiple storage arrays - flexible



Mass-Storage Structure

Reliability

MIRRORING One way to increase reliability is to "mirror" data on a disk. Every piece of data is maintained on two disks - disk drivers must be capable of getting data from either disk. Performance issues: a read is faster since data can be obtained from either disk - writes are slower since the data must be put on both disks.

RAID Redundant Array of Inexpensive Disks: Rather than maintain two copies of the data, maintain one copy plus parity. For example, four disks contain data, and a fifth disk holds the parity of the XOR of the four data disks. Reads slower than mirroring, writes much slower. But RAID is considerably CHEAPER than mirroring.

DISK STRIPING Disks tend to be accessed unevenly - programs ask for a number of blocks from the same file, for instance. Accesses can be distributed more evenly by spreading a file out over several disks. This works well with RAID. Thus block 0 is on disk 0, block 1 is on disk 1, block 4 is on disk 0.

Consider how to recover from a failure on these architectures.

Mass-Storage Structure

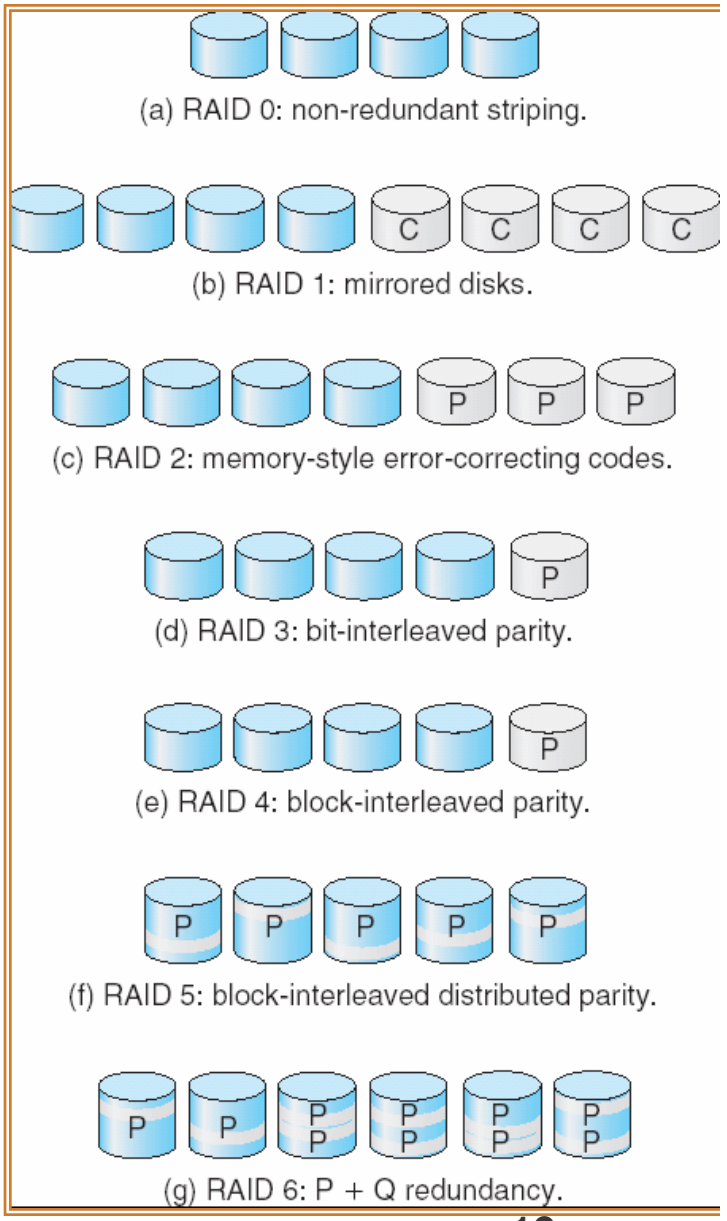
These are the various levels of RAID.

The reliability increases with higher levels.

In practice, only levels 0, 1, 5 and 10 are typically used.

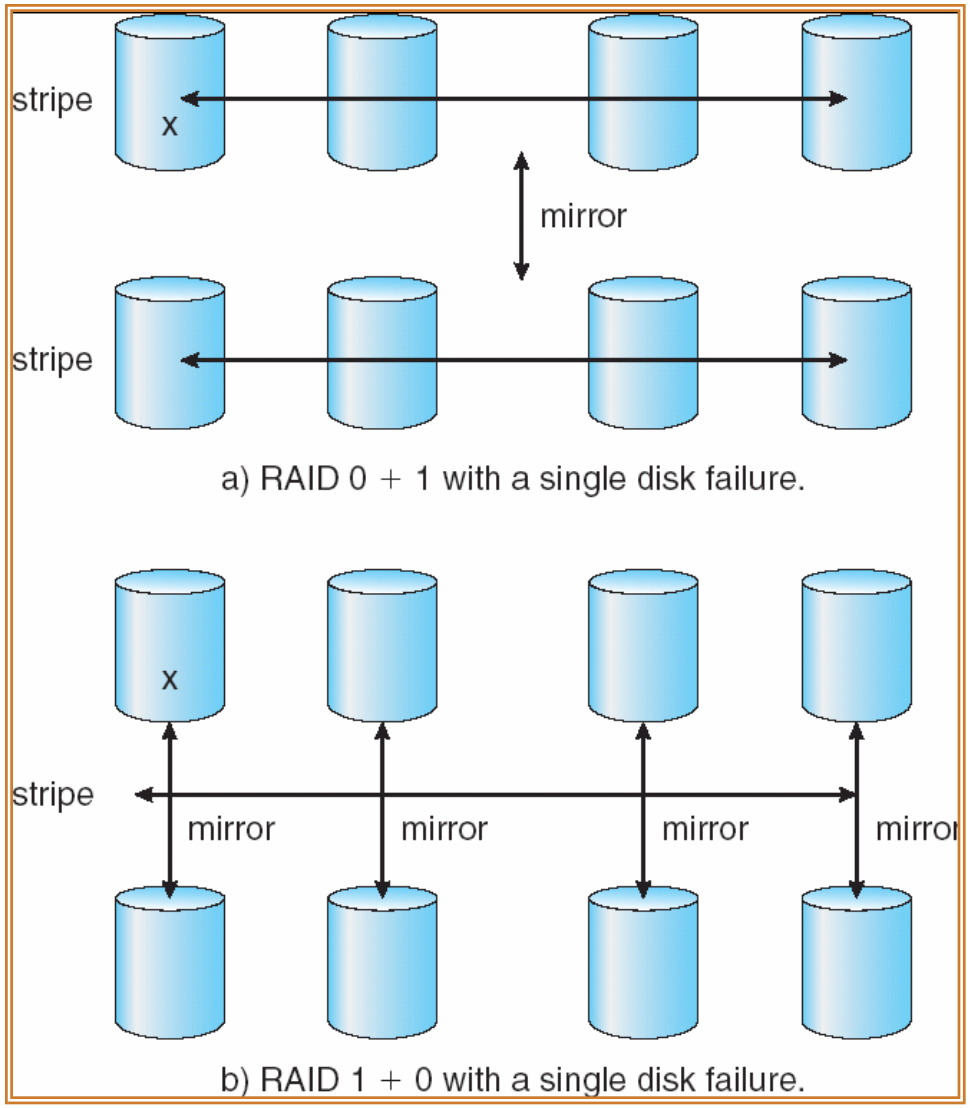
- Several improvements in disk-use techniques involve the use of multiple disks working cooperatively.
- **Disk striping** uses a group of disks as one storage unit.
- RAID schemes improve performance and improve the reliability of the storage system by storing redundant data.
  - *Mirroring* or *shadowing* keeps duplicate of each disk.
  - *Block interleaved parity* uses much less redundancy.

RAID



Mass-Storage Structure

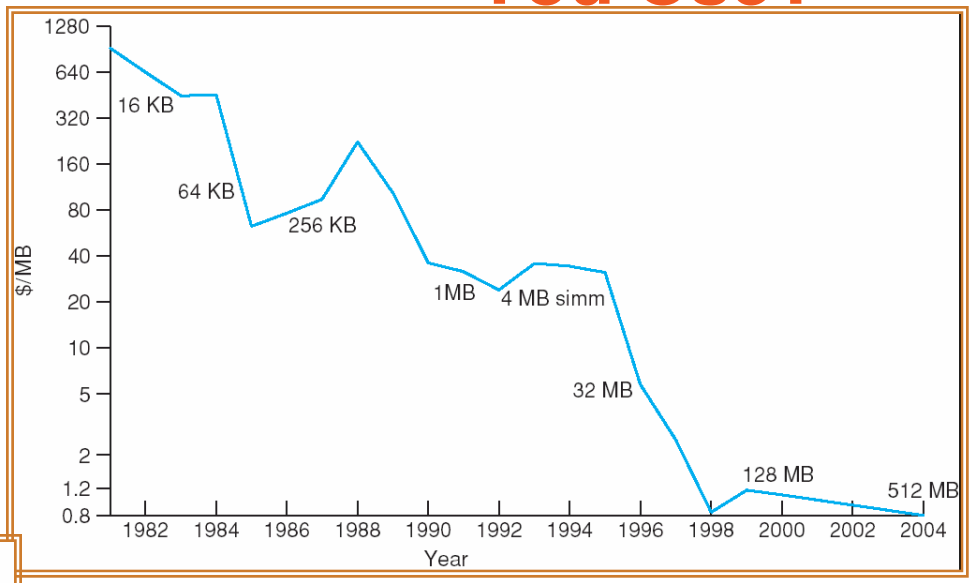
RAID 10 becoming more and more popular.

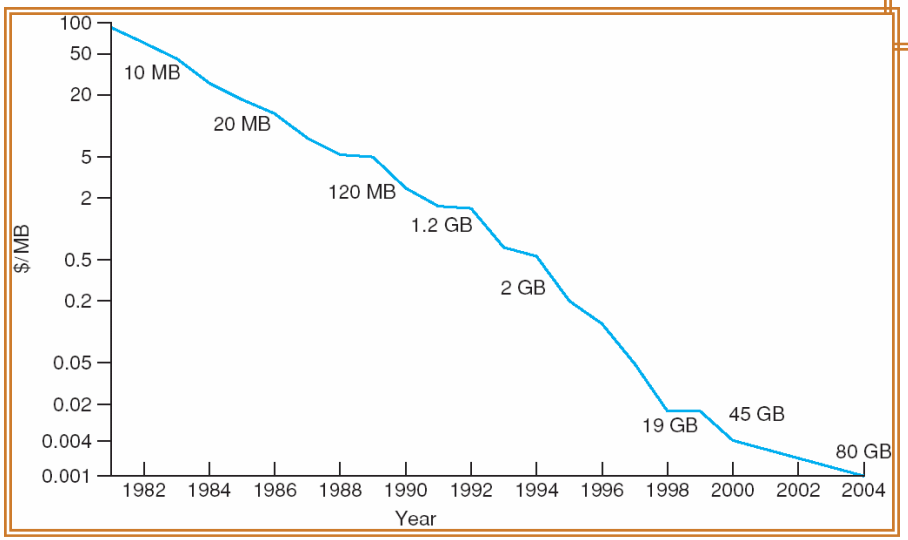


Mass-Storage Structure

What Kind Of Storage Should You Use?

Price per Megabyte of DRAM, From 1981 to 2004





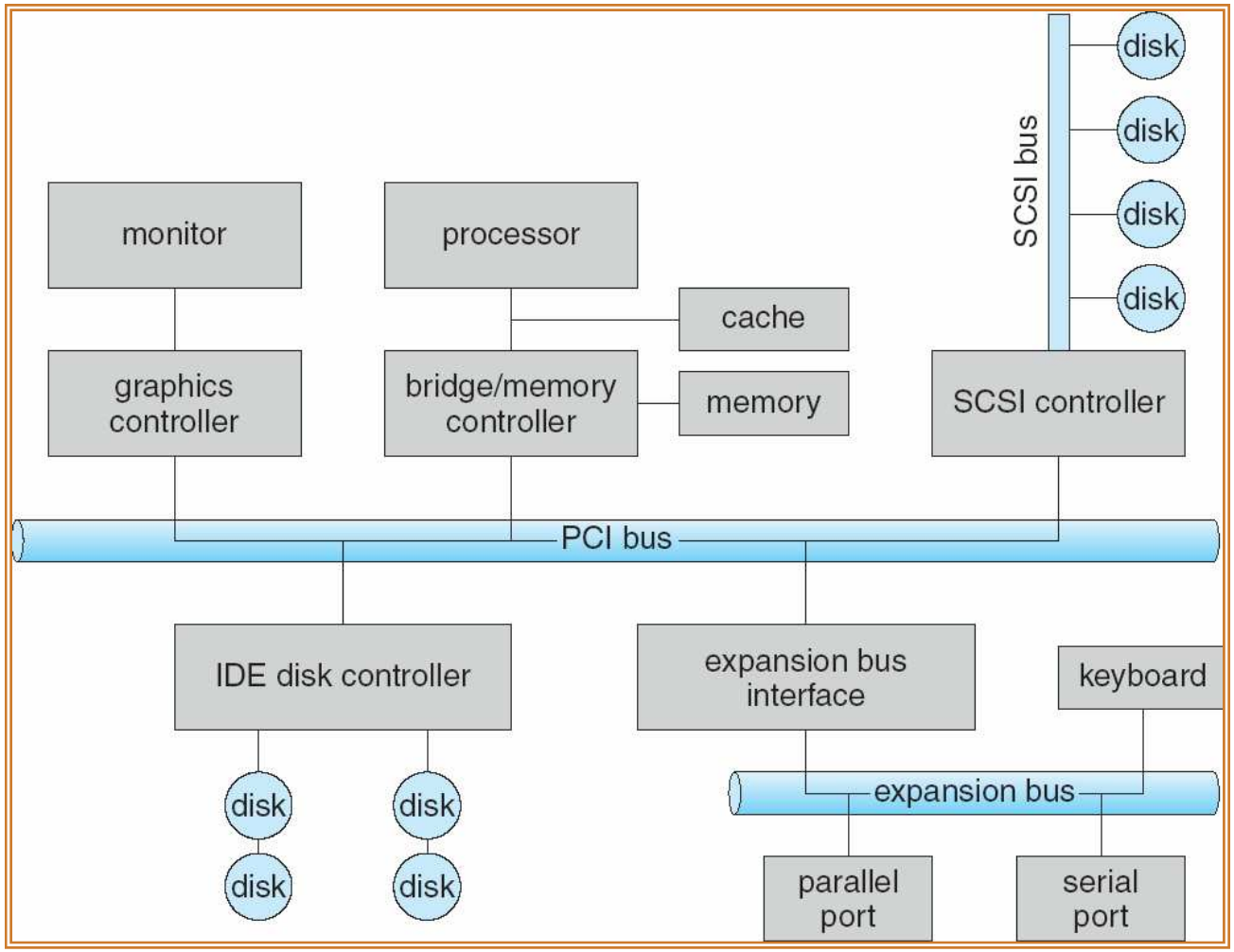
Price per Megabyte of Hard Disk, From 1981 to 2004

Is it only about price?? Where does speed fit in?

IO SYSTEMS

IO Hardware

- Incredible variety of I/O devices
- Common concepts
  - **Port**
  - **Bus** (**daisy chain** or shared direct access)
  - **Controller** (**host adapter**)
- I/O instructions control devices
- Devices have addresses, used by
  - Direct I/O instructions
  - **Memory-mapped** I/O



IO SYSTEMS

IO Hardware

Memory Mapped IO:

Works by associating a memory address with a device and a function on that device.

I/O address range (hexadecimal)	device
000–00F	DMA controller
020–021	interrupt controller
040–043	timer
200–20F	game controller
2F8–2FF	serial port (secondary)
320–32F	hard-disk controller
378–37F	parallel port
3D0–3DF	graphics controller
3F0–3F7	diskette-drive controller
3F8–3FF	serial port (primary)

IO SYSTEMS

Polling and Interrupts

CPU Interrupt request line triggered by I/O device

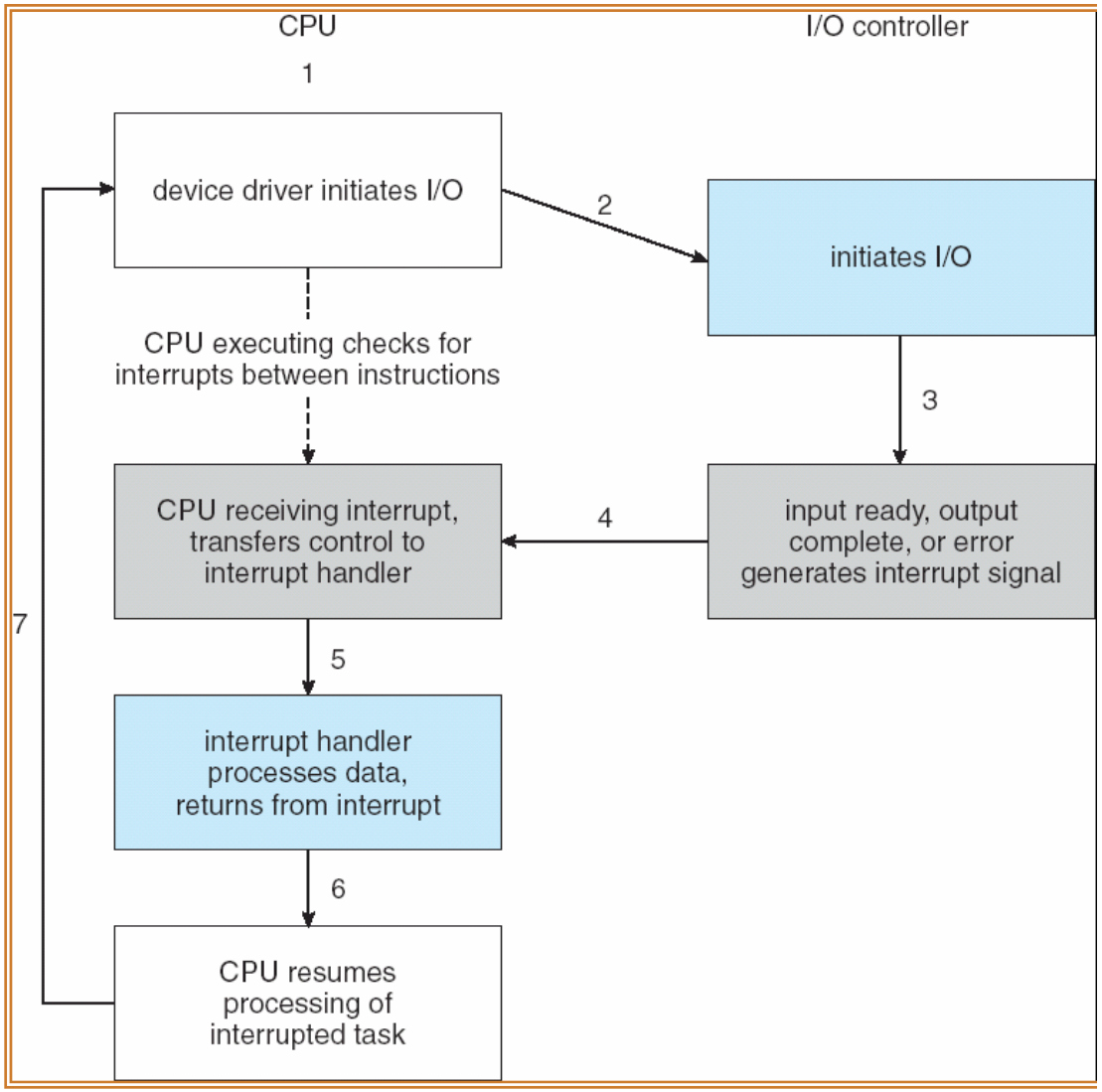
Interrupt handler receives interrupts

Maskable to ignore or delay some interrupts

Interrupt vector to dispatch interrupt to correct handler

- Based on priority
- Some unmaskable

Interrupt mechanism also used for exceptions.



IO SYSTEMS

Polling and Interrupts

When you get an interrupt, you need to be able to figure out the device that gave you the interrupt.

These are the interrupt vectors for an Intel Processor.

Notice that most of these are actually exceptions.

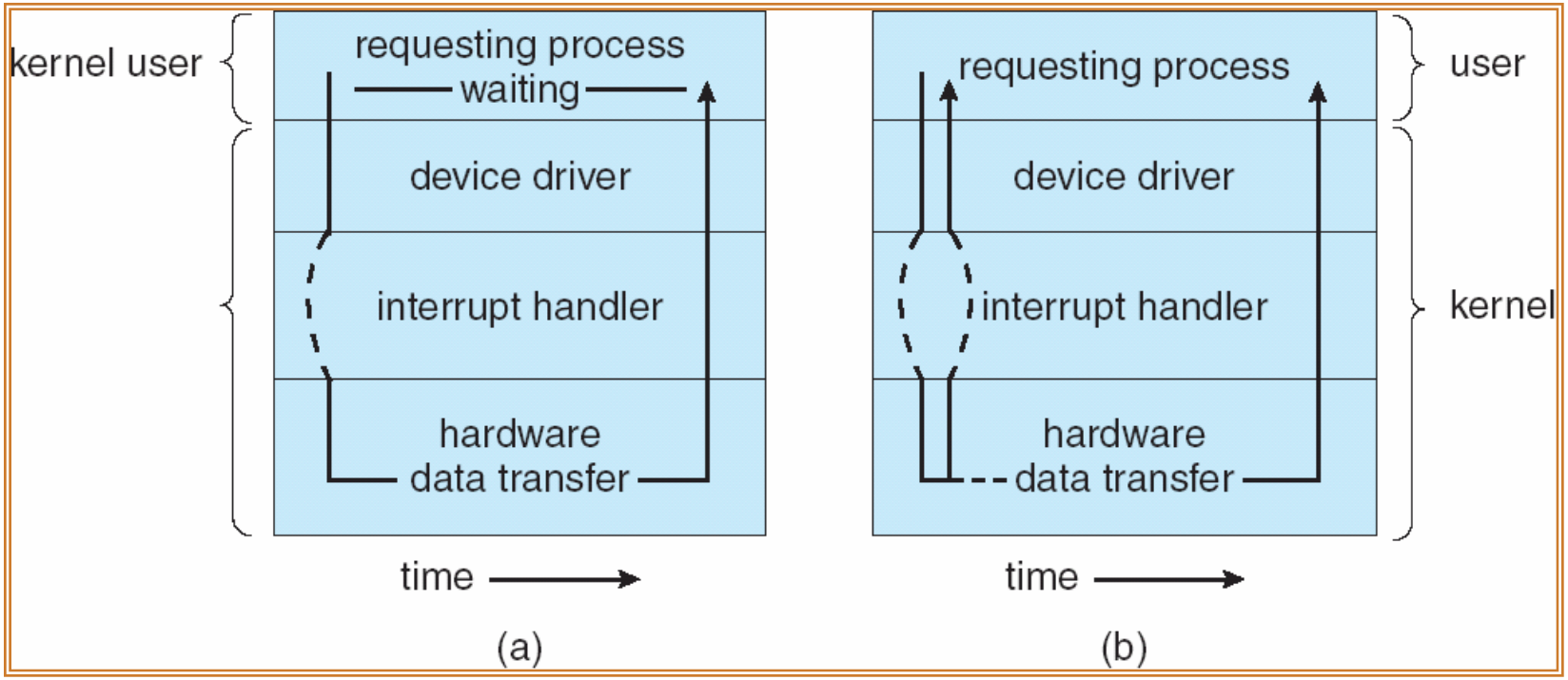
vector number	description
0	divide error
1	debug exception
2	null interrupt
3	breakpoint
4	INTO-detected overflow
5	bound range exception
6	invalid opcode
7	device not available
8	double fault
9	coprocessor segment overrun (reserved)
10	invalid task state segment
11	segment not present
12	stack fault
13	general protection
14	page fault
15	(Intel reserved, do not use)
16	floating-point error
17	alignment check
18	machine check
19Ð31	(Intel reserved, do not use)
32Ð255	maskable interrupts

IO SYSTEMS

Synchronous or Asynchronous

Synchronous does the whole job – all at one time – data is obtained from the device by the processor.

Asynchronous has the device and the processor acting in time independent of each other.



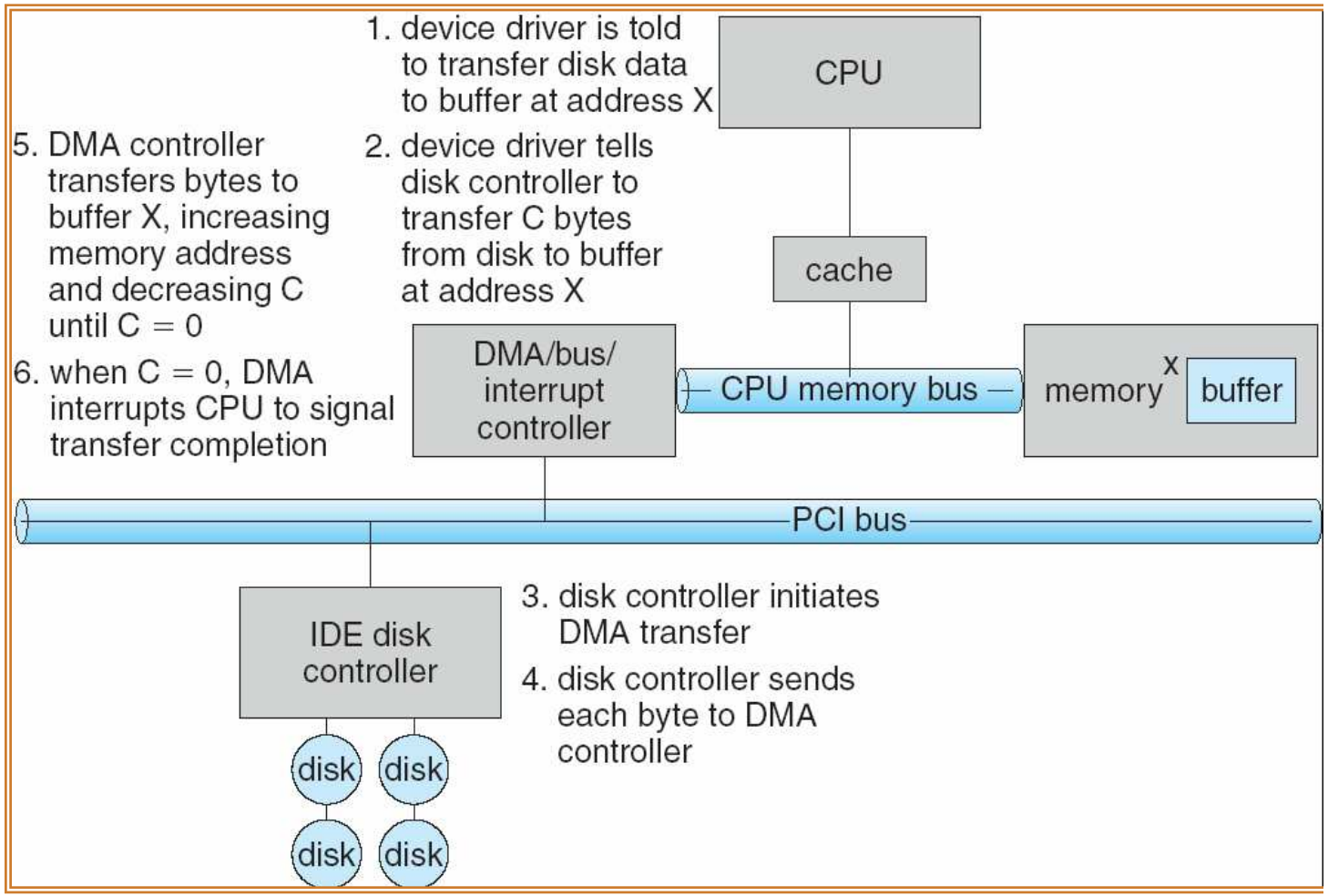
IO SYSTEMS

DMA

Used to avoid programmed I/O for large data movement

Requires DMA controller

Bypasses CPU to transfer data directly between I/O device and memory



IO SYSTEMS

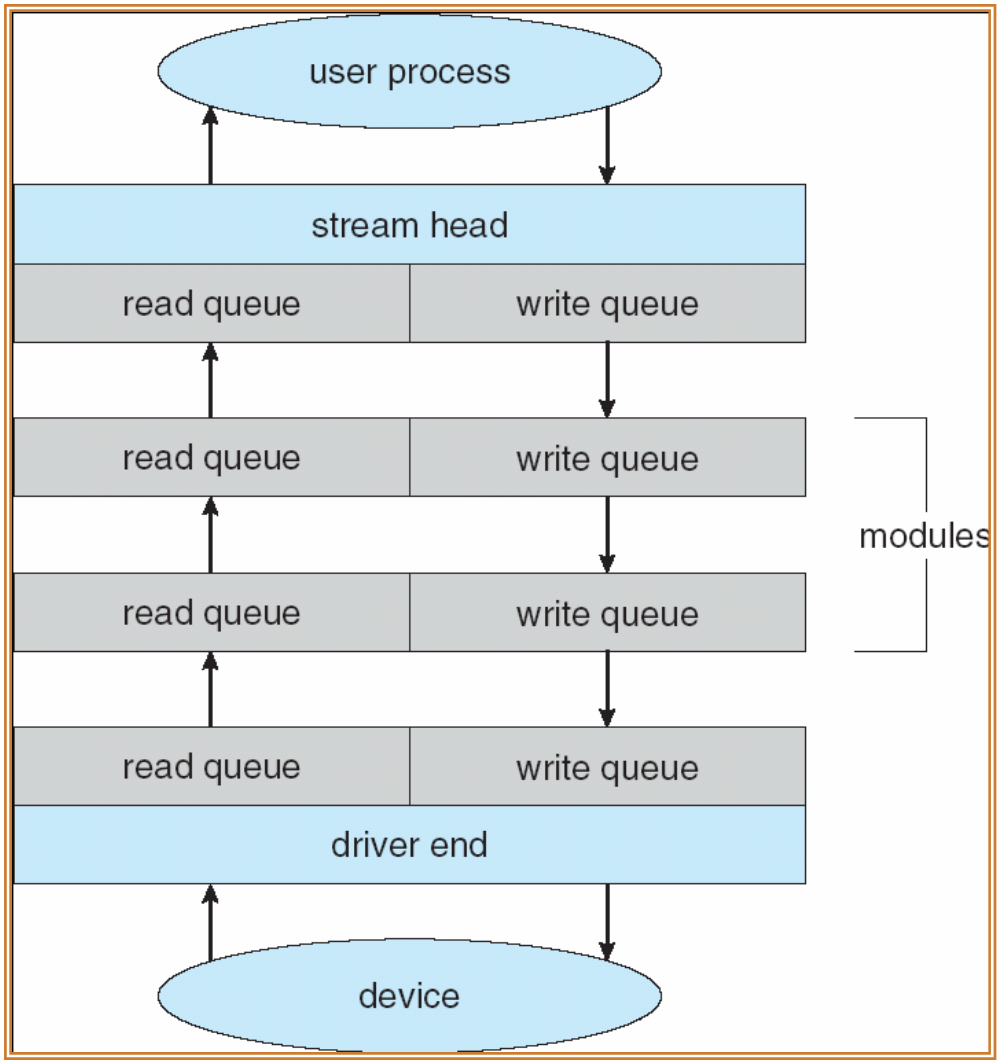
- **STREAM** – a full-duplex communication channel between a user-level process and a device in Unix System V and beyond

- A STREAM consists of:
  - STREAM head interfaces
  - driver end interfaces with the device
  - zero or more STREAM modules between them.

- Each module contains a **read queue** and a **write queue**

- Message passing is used to communicate between queues

Streams



IO SYSTEMS

Interfaces

Block and Character Devices

- Typical for disks – use read(), write(), seek() sequence.

Network Devices

Clocks and Timers

- The OS uses an incredible number of clock calls – many events are timestamped within the OS

Blocking and Non-Blocking IO

- Non- Blocking: Some devices are started by the OS, and then proceed on without further OS intervention. The delay timer in our project works this way.
- Blocking: Any Read-Device will be blocking since the program can't proceed until it gets the information it wanted from the device.

IO SYSTEMS

Kernel IO Subsystem

Buffering

- Used to interface between devices of different speeds (modem and disk for instance.)
- Interface between operations having different data sizes. (small network packets as part of a bigger transfer.)
- Users often read or write small number of bytes – but the disk wants 4096 bytes. The filesystem maintains this buffer.

Spooling

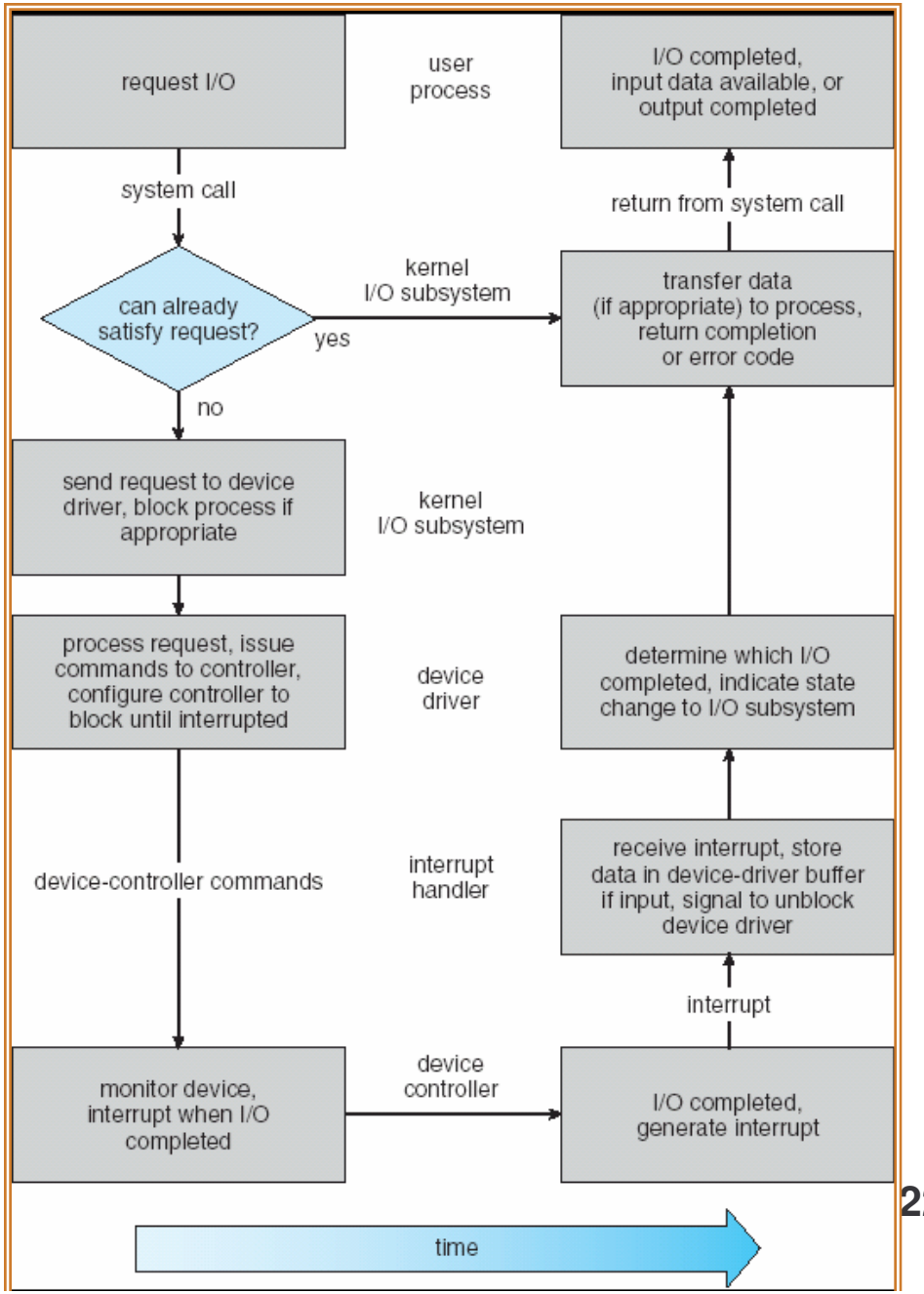
Kernel data structures – what needs to be maintained to

- Support the device
- Support an instance of opening the device.

IO SYSTEMS

The steps in an IO request.

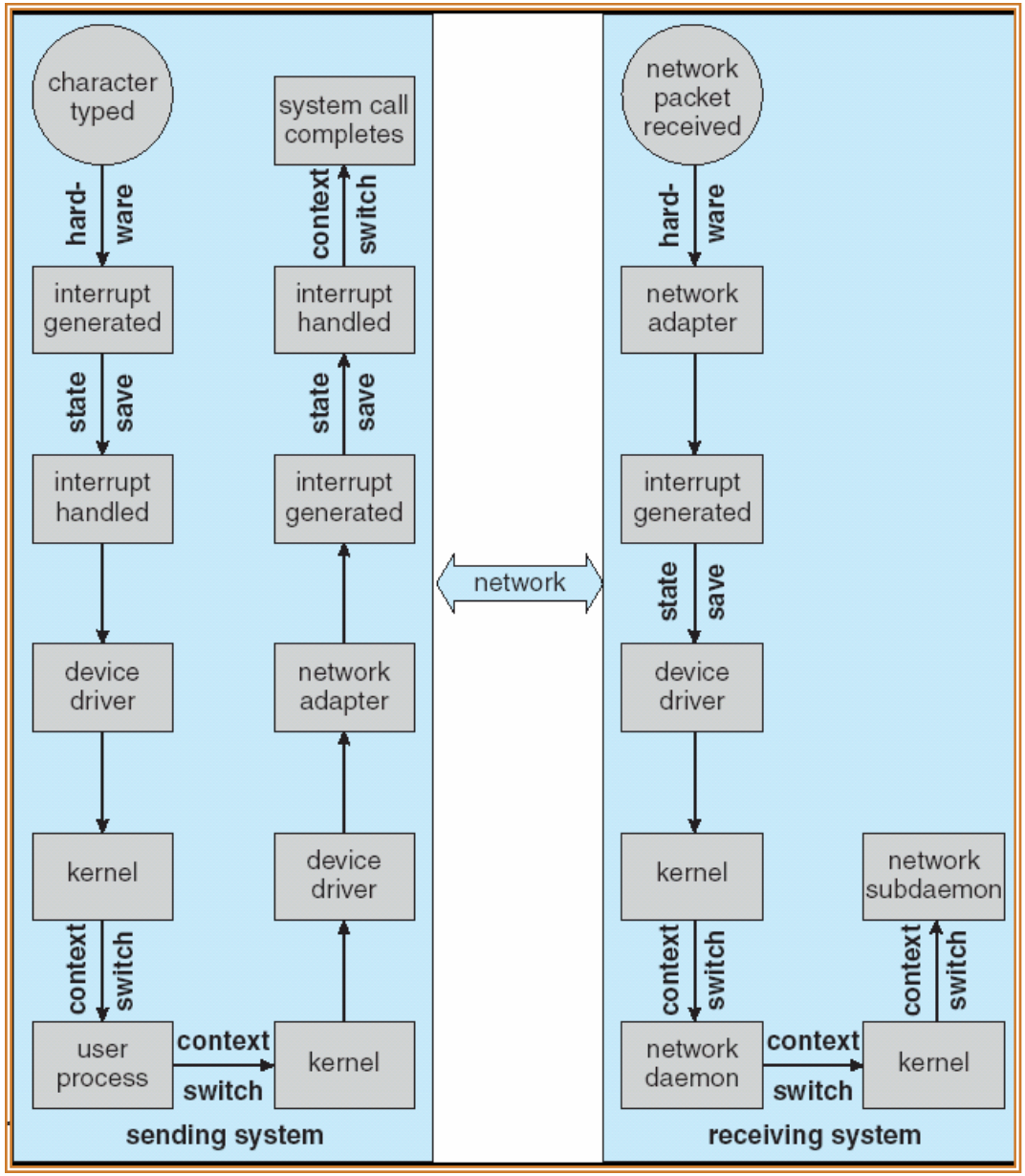
Kernel IO Subsystem



IO SYSTEMS

Performance

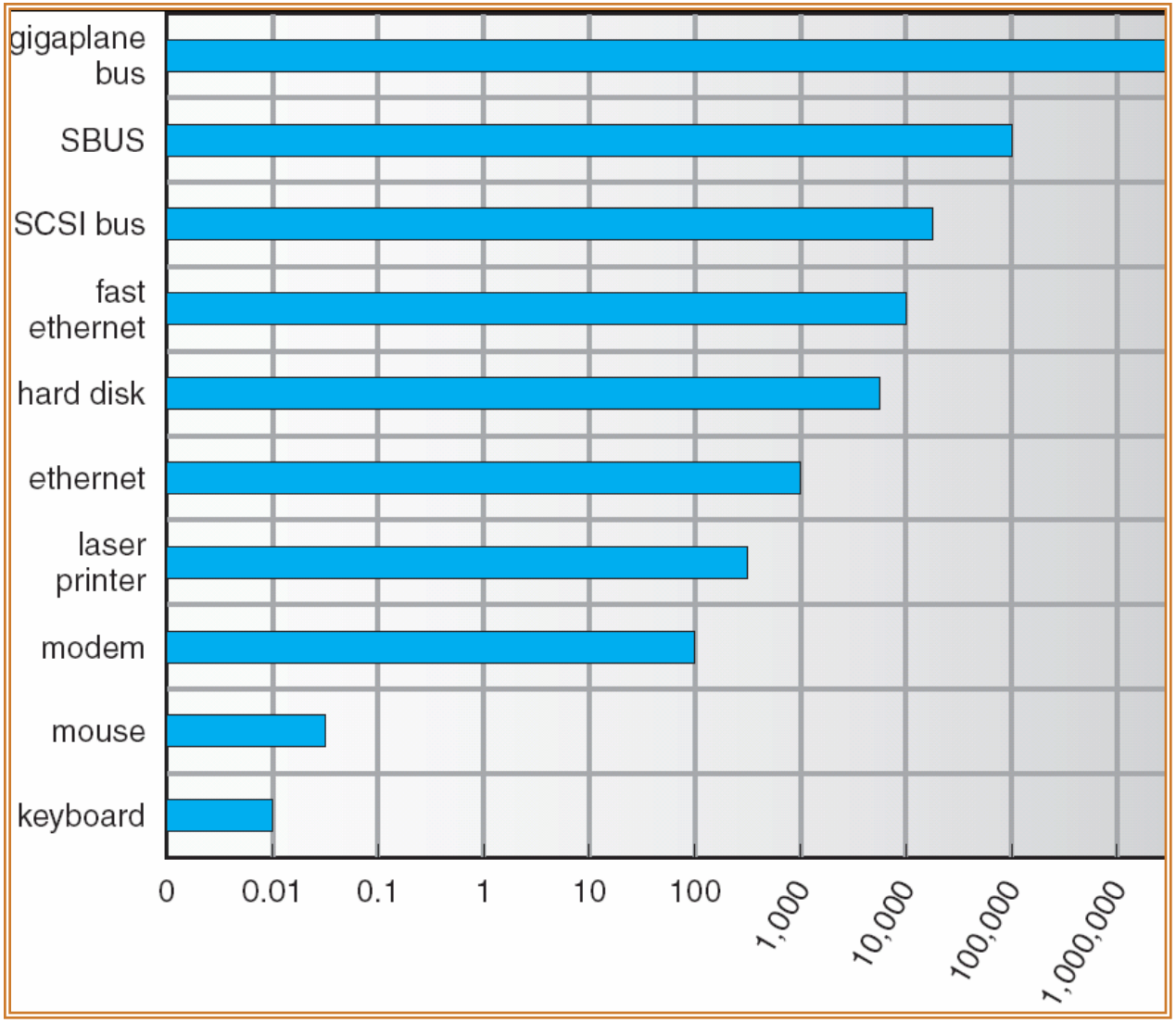
The steps required to handle a single keystroke across the network.



IO SYSTEMS

Performance

Throughput for various devices.



FILE SYSTEMS

Wrap Up

Mass Storage

This is about Disk Behavior and Management.
- Disk Characteristics
- Space Management
- RAID
- Disk Attachment

IO Interface

The busses in the computer and how the O.S. interfaces to it.
- Talking to the IO – Polling, Interrupts and DMA
- Application IO Interface
- Kernel IO Subsystem