• Memory Mapped I/O

Certain memory addresses correspond to registers in I/O devices and not normal memory.

Control Register: Indicates if it is okay to read/write data register
Data Register: Contains I/O data

Register	Location	Contains
Receiver Control	0xffff0000	Lowest two bits: Interrupt Enable Bit, Ready Bit
Receiver Data	0xffff0004	Received data stored at lowest byte
Transmitter Control	0xffff0008	Lowest two bits: Interrupt Enable Bit, Ready Bit
Transmitter Data	0xffff000c	Transmitted data stored at lowest byte

Write MIPS code to read a byte from the receiver and immediately send it to the transmitter.

• Polling and Interrupts

Operation	Definition	Pro/Con	Good For
Polling			
Interrupts			

Interrupt Enable: Indicates whether or not to cause an interrupt when the ready bit is set

• Network Overhead vs. Bandwidth

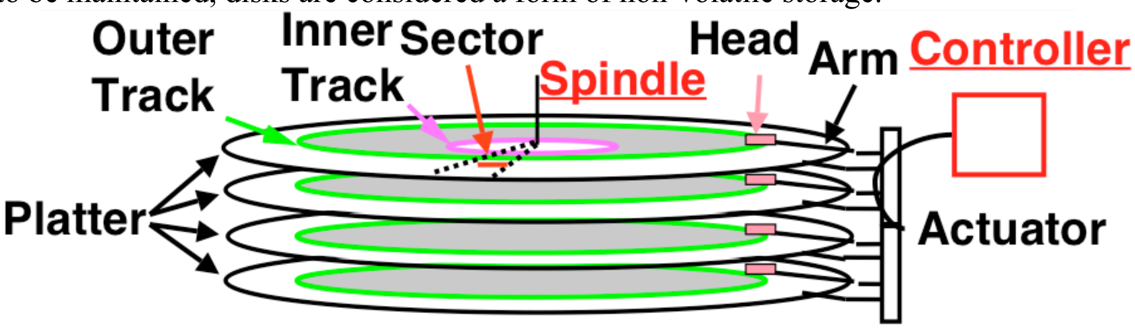
Assume we have two networks A and B.

Network A has a 200us overhead and a peak bandwidth of 10MB/s
Network B has a 500us overhead and a peak bandwidth of 100MB/s
(MB = MebiByte)

How long would it take to send 1000 Bytes over each network?
Which network is better for sending large amounts of data?

Disk Organization

Magnetic disks are one of the most common types of I/O devices. Bits are encoded by the controlling the polarity of magnetic fields on some sort of substrate. Since the magnetic fields do not require power to be maintained, disks are considered a form of non-volatile storage.



An additional term not shown here is that the collection of corresponding tracks across all the platters is called a cylinder.

There are two ways to address disks. Logical addressing treats the disk drive as one big array of blocks. Physical addressing uses a (cylinder, sector, platter) tuple to specify a blocks physical position in the disk drive.

Disk Performance

Disk Latency = Seek Time + Rotation Time + Transfer Time + Controller Overhead

RAID

Big disks are expensive (and dangerous). We can use an array of smaller disks to simulate the behavior of one larger disk with a more reasonable cost.

RAID 0	No redundancy, just multiple disks
RAID 1	Mirroring for redundancy, doubles read bandwidth
RAID 2	Bit-level striping, increases bandwidth further
RAID 3	Parity Disks, allows recovery from a single disk failure
RAID 4	Block-level striping with Parity Disk, increases bandwidth
RAID 5+	Striped Parity, reduces wear and tear

Performance Metrics

In order to get any meaningful definition of performance (used for comparison), we need to develop a quantitative metric that we all can agree on. This is harder than it sounds. We briefly talked about these when discussing pipelining.

Response Time/Execution Time/Latency – the time it takes to complete one task

Throughput/Bandwidth – tasks completed per time unit

We say A has *n* times better performance than B if *n* = performance(A) / performance (B)

Avoid using "percent" comparisons and "increased/decreased" to compare performances. This can be confusing and misleading. Instead we talk about performance improvements.

Based on notes by Aaron Staley, which were in turn based on notes by David Jacobs.

CS 61C Spring 2008	Week 16	TA: Omar Akkawi

Megahertz Myth

A processors performance is determined by more than just the clock speed.
CPU time = Instruction Count * CPI * clock period

Exercises

You are the lead developer of a new video game at AE, Inc. The graphics are quite sexy, but the frame rates (performance) are horrible. Doubly unfortunately, you have to show it off at a shareholder meeting tomorrow. What do you do?

You need to render your latest and greatest über-l33t animation. If your rendering software contains the following mix of instructions, which processor is the best choice?

Operation	Frequency
ALU	30%
Load	30%
Store	20%
Branch	20%

A's CPI	B's CPI	C's CPI
1	1	1
3	5	3
2	3	4
3	2	2

What if the processors had different clock speeds? Assume A is a 1 Ghz processor, B is a 1.5 Ghz processor, and C is a 750 Mhz processor.

But wait, these processors are made by different manufacturers, and use different instruction sets. So the renderer (for the different architectures) takes a different number of instructions on each. Which is best if your main loop on A averages 1000 instructions; on B it averages 800 instructions; and on C it averages 1200 instructions?

Parallel Computing

Parallel computing refers more to multicore and multiprocessor machines. This is sometimes also called "supercomputing." Since the processors are physically closer together, there is a potential for much faster communications between them. However, synchronizing the processors can prove a difficult problem.

Amdahl's Law

The potential speedup from parallelization is limited by the amount a program can be parallelized. Let s be the fraction of the work that must be done sequentially and P be the number of processors. Then,

$$\text{Speedup}(P) \leq 1/s$$

Exercise

What are the contributing factors to Amdahl's law? Why isn't it an equality?

Based on notes by Aaron Staley, which were in turn based on notes by David Jacobs.