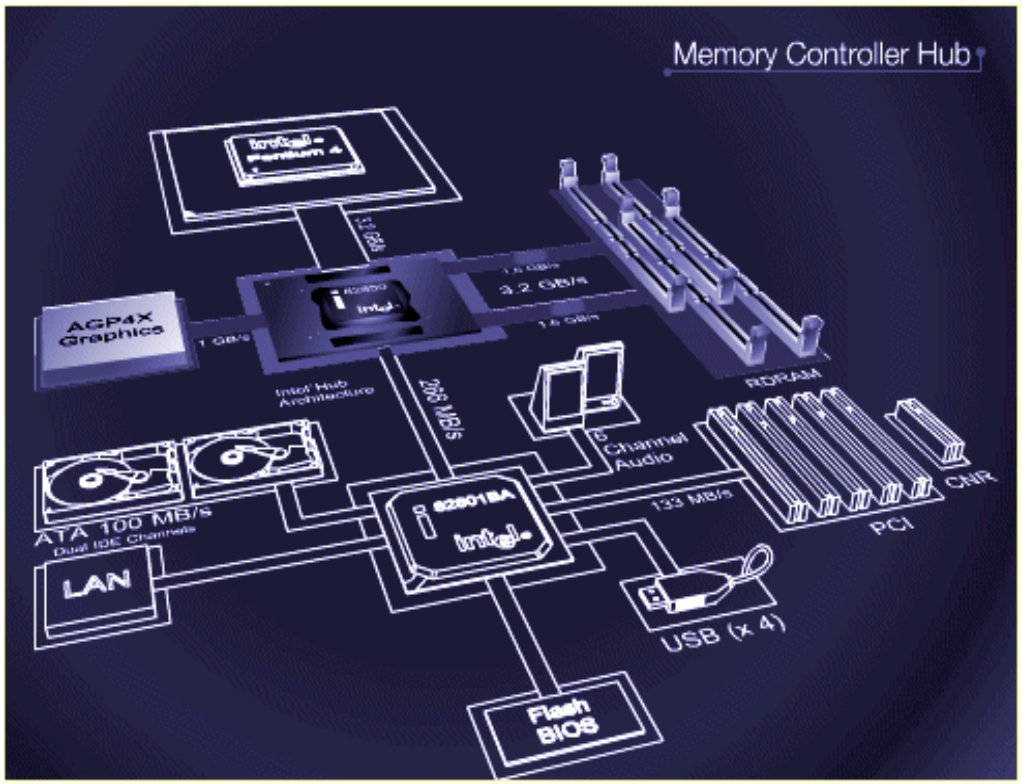
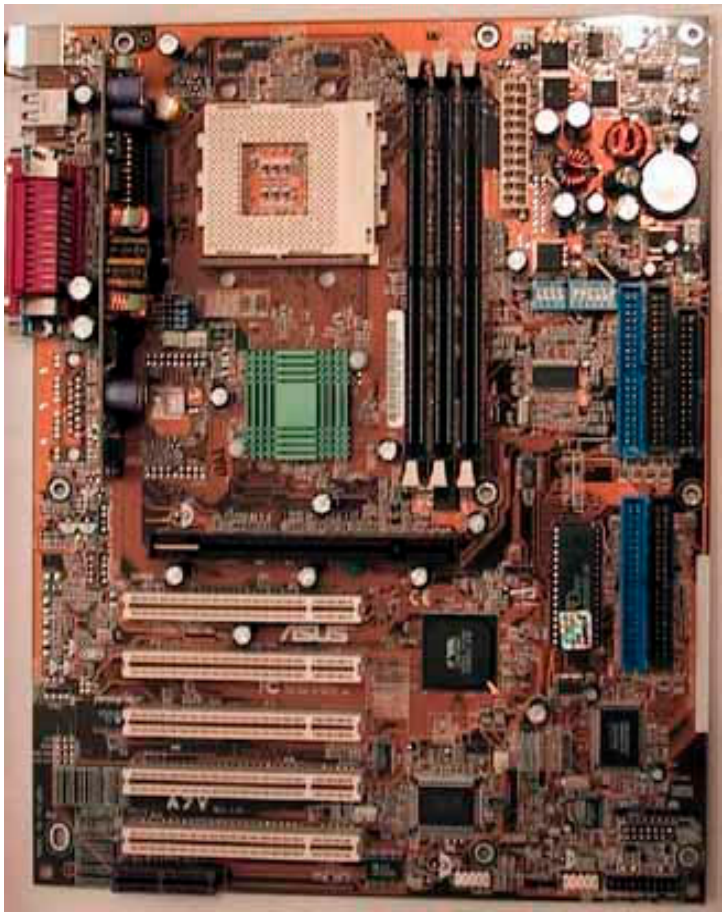
PC I/O

- Today wraps up the I/O material with a little bit about PC I/O systems.
  - Internal buses like PCI and ISA are critical.
  - External buses like USB and Firewire are becoming more important.
- Today also happens to be the last day of class.

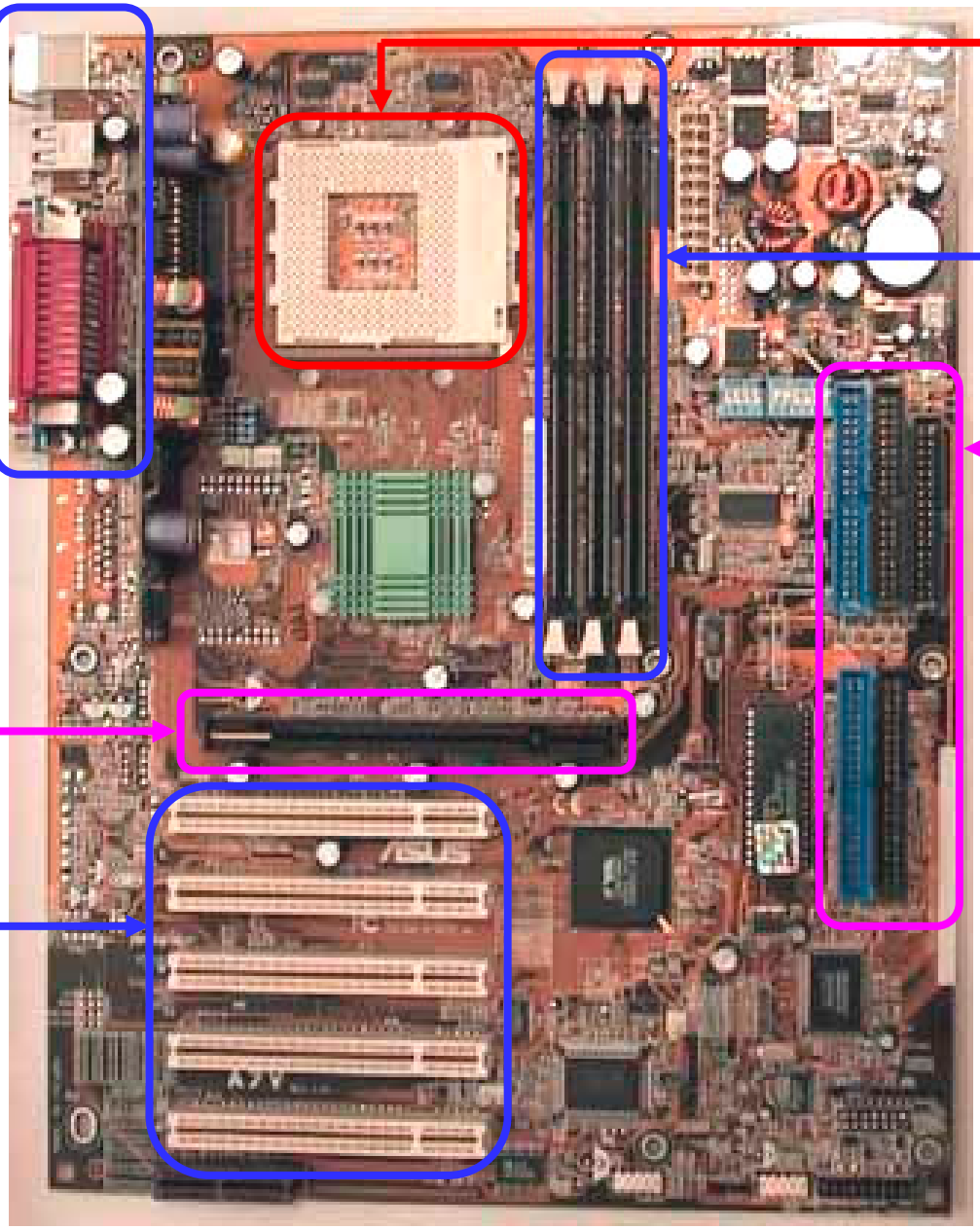


Open up and say ahhh

- If you open up your computer, you'll find the motherboard, which connects everything together.



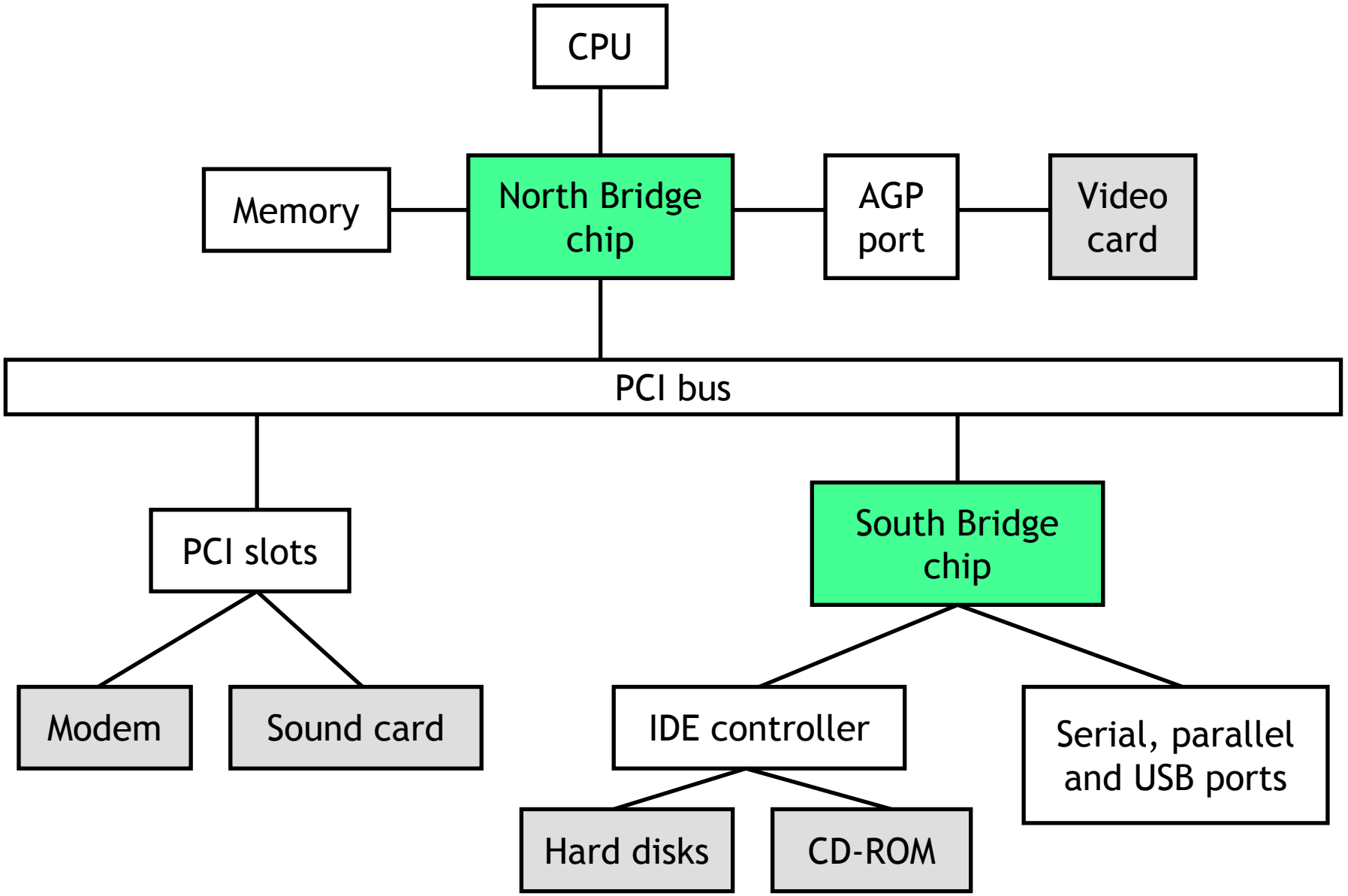
The mothership



What is all that stuff?

- Different motherboards support different CPUs, types of memories, and expansion options.
- The picture is an Asus A7V.
  - The CPU socket supports AMD Duron and Athlon processors.
  - There are three DIMM slots for standard PC100 memory. Using 512MB DIMMs, you can get up to 1.5GB of main memory.
  - The AGP slot is for video cards, which generate and send images from the PC to a monitor.
  - IDE ports connect internal storage devices like hard drives, CD-ROMs, and Zip drives.
  - PCI slots hold other internal devices such as network and sound cards and modems.
  - Serial, parallel and USB ports are used to attach external devices such as scanners and printers.

How is it all connected?

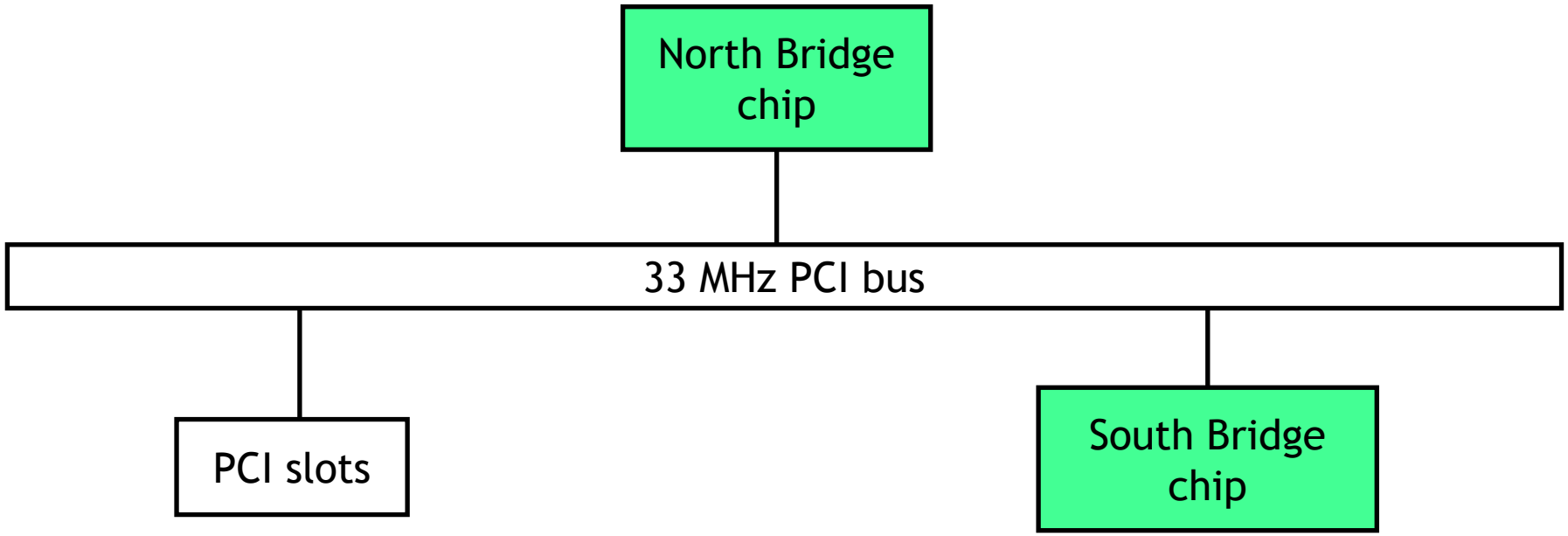


PCI

- Peripheral Component Interconnect is a synchronous 32-bit bus running at 33MHz, although it can be extended to 64 bits and 66MHz.
- The maximum bandwidth is about 132 MB/s.

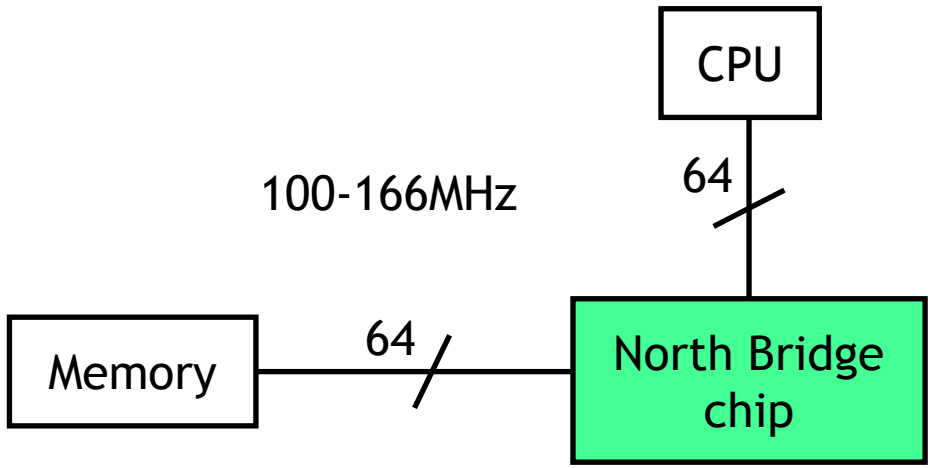
33 million transfers/second × 4 bytes/transfer = 132MB/s

- Cards in the motherboard PCI slots plug directly into the PCI bus.
- Devices made for the older and slower ISA bus standard are connected via a "south bridge" controller chip, in a hierarchical manner.



Fist of the North Bridge

- The CPU and main memory exchange large amounts of data frequently.
- The 33MHz PCI bus is relatively slow for this, so a frontside bus is often dedicated to the CPU and main memory.
  - Some newer systems use 64-bit frontside buses running 100-166MHz. By transferring data on both the positive and negative clock edges, the effective frequency increases to 200-333MHz.
  - Other systems feature 16-bit buses running at 800MHz, and making four data transfers per clock cycle.
- All of this goes through the "north bridge" controller, which can get very hot. The north bridge in the A7V is cooled by a green heatsink.



Frequencies

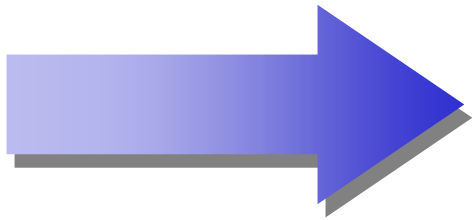
- CPUs actually operate at two frequencies.
  - The internal frequency is the clock rate inside the CPU, which is what we've been talking about all along.
  - The external frequency is the speed of the processor bus, which limits how fast the CPU can transfer data.
- The internal frequency is usually a multiple of the external bus speed.
  - A 2.66 GHz Pentium 4 might use a 533 MHz bus (533 × 5).
  - An 2.167 GHz Athlon XP sits on a 166 MHz bus (166 × 13).
- Processors can often be overclocked to run *faster* than advertised.
  - Some motherboards allow you to change the external bus frequency, the "clock multiplier" (e.g., 5 or 13), or both.
  - This can cause problems with some devices or the CPU itself if you're not careful or if you overclock too much.

Video cards are memory hogs

- Graphics cards need a lot of fast memory for high-resolution images.
  - A 1024 × 768 resolution image in 32-bit color requires 3 MB of RAM.

    1024 pixels/row × 768 rows × 4 bytes/pixel = 3 MB

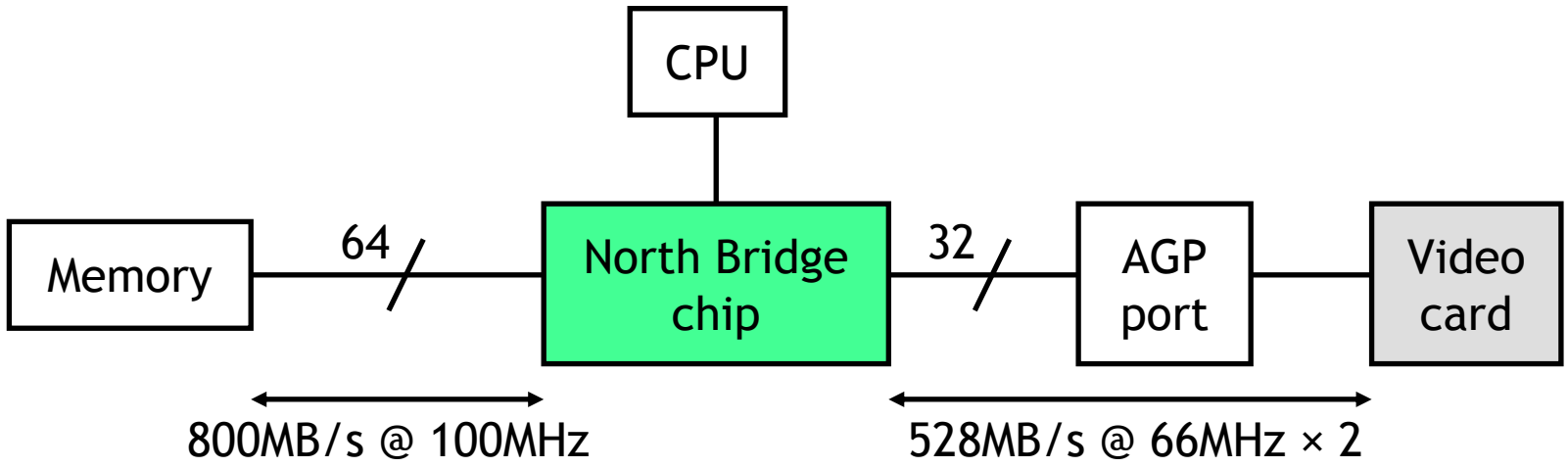
  - If the screen is refreshed at 75 Hz, this 3 MB image must be read 75 times per second.
- Games also use textures, which are images that get pasted onto shapes. Textures can each require 256 KB of memory or more.



Chris Lattner,
CS319 HW1

AGP

- In the old days video cards used PCI, but that involved a slow 33MHz bus with a lot of contention and a maximum transfer rate of 132MB/s.
- New machines include an Accelerated Graphics Port.
  - The graphics card gets its own 32-bit, 66MHz path to the north bridge.
  - With clock doubling (AGP 2x), transfer rates can go up to 528MB/s from the video card to main memory.
- The CPU and video card *both* need to access main memory quickly, so we still have possible contention problems.



External buses

- External buses can be much more convenient for end users.
  - PCI, ISA and AGP devices must be attached inside the computer. This is difficult for some users, and laptops cannot even be opened.
  - External buses also allow devices to be connected from further away.
- Some common external buses include serial and parallel connections, USB and Firewire.

Serial and parallel connections

- A serial connection transmits data one bit at a time.
- A parallel port transmits multiple bits (typically 8) at a time. Although this can be faster, there are a couple of problems too.
  - Thicker cables are needed to transmit more bits at once. This can be a hindrance for smaller devices like MP3 players.
  - With many wires, interference becomes an issue. Special cables may be necessary for high-speed parallel connections.
  - Delay or skew is also a problem, especially with longer external connections. All of the bits in a parallel transfer could arrive at slightly different times.

The serial USB connector is the small one, and the parallel connector is the big one.

USB

- The Universal Serial Bus is a newer asynchronous serial bus standard, with many advantages over older serial and parallel buses.
- The USB 1.1 standard has a bandwidth of 1.5 MB/s, which is enough for keyboards, mice and small home networks.
- The newewr USB 2.0 standard boosts speeds up to 60 MB/s, which works better for faster devices like CD-ROMs or hard drives, which can transfer at rates of 7-40 MB/s.
- The bus can be shared by up to 127 devices. As usual, more devices leads to more bus contention, but most personal computers have only a couple of USB peripherals.

USB is friendly

- USB has several other advantages.
  - It supports plug-and-play standards, so devices can be configured with software instead of flipping switches or setting jumpers.
  - It also supports hot plugging, so you don't have to turn off a machine to add or remove a peripheral. This is important for portable devices that aren't always connected to a PC, like an MP3 player or a PDA.
  - The cable and connectors are small, with just four wires.
  - The cable transmits power! No more power cables, extension cords or electrical fires needed!

Firewire

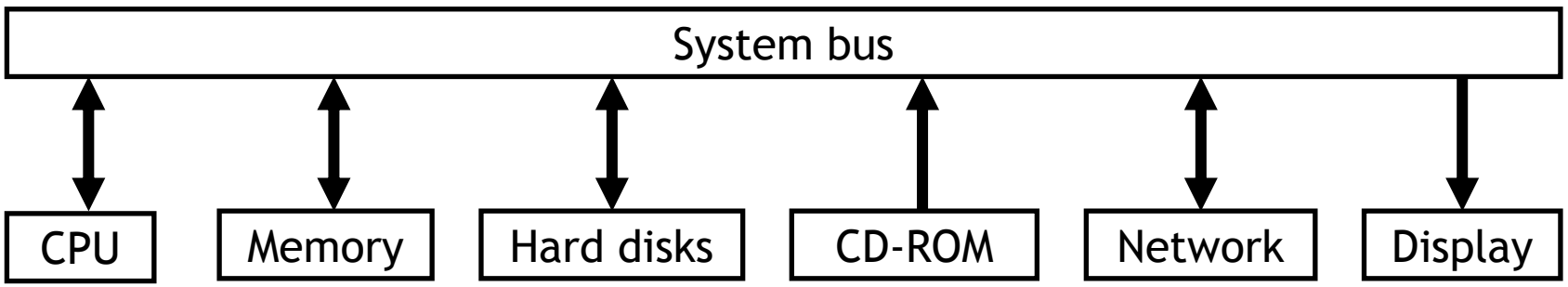
- Firewire, developed and trademarked by Apple, is very similar to USB.
- It has three main advantages.
  - Most current implementations transfer up to 50MB/s, and the newest versions of Firewire can go twice as fast.
  - No PC is needed! It's possible to connect two devices directly, which can make it easy to do things like copy data between two digital video camcorders.
  - It provides more power to peripheral devices.
- These advantages make it attractive for video processing applications.

Common bus themes

- Internal buses are taxed by fast devices that need large bandwidths, as well as many devices contending for the bus.
  - Clock doubling increases transfer rates with the same bus width.
  - Adding shorter and faster dedicated buses for high-bandwidth devices like the CPU, main memory and video cards reduces contention.
- External buses have some different requirements.
  - Distances between devices are greater, so asynchronous serial data transfers are often faster and more reliable.
  - Plug-and-play and hot plugging are important for end users.
- Numbers are always just general guidelines! Most buses never reach their theoretical maximum bandwidth, for instance.

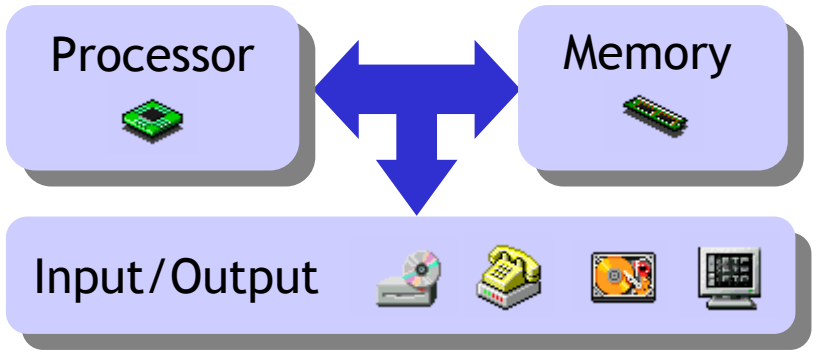
New bus trends

- Emerging bus standards like 3GIO, HyperTransport and RapidIO combine many ideas from previous bus designs.
  - Serial connections avoid delay skews and interference, can run at high speeds, and are less expensive to manufacture.
  - Using multiple buses or moving to "hub" or "switch" organizations can reduce contention.
  - Data from several devices can be multiplexed onto a bus, much like packets on a network. Thus, many devices can transmit data at once, instead of having to wait for previous transfers to complete.
- New features like error detection, quality-of-service, and power saving are also important for applications like servers, real-time broadcasting, and embedded systems.



Instant replay

- The semester was split into roughly three parts.
  - The first third covered instruction set architectures—the connection between software and hardware.
  - In the middle of the course we discussed processor design. We focused on pipelining, which is one of the most important ways of improving processor performance.
  - Finally we talked about fast and large memory systems, I/O, and how to connect everything together.
- We also introduced many performance metrics to estimate the actual benefits of all of these fancy designs.



Some recurring themes

- There were several recurring themes throughout the semester.
  - Instruction set and processor designs are intimately related.
  - Parallel processing can often make systems faster.
  - Amdahl's Law quantifies performance limitations.
  - Hierarchical designs combine different parts of a system.
  - Hardware and software depend on each other.

Instruction sets and processor designs

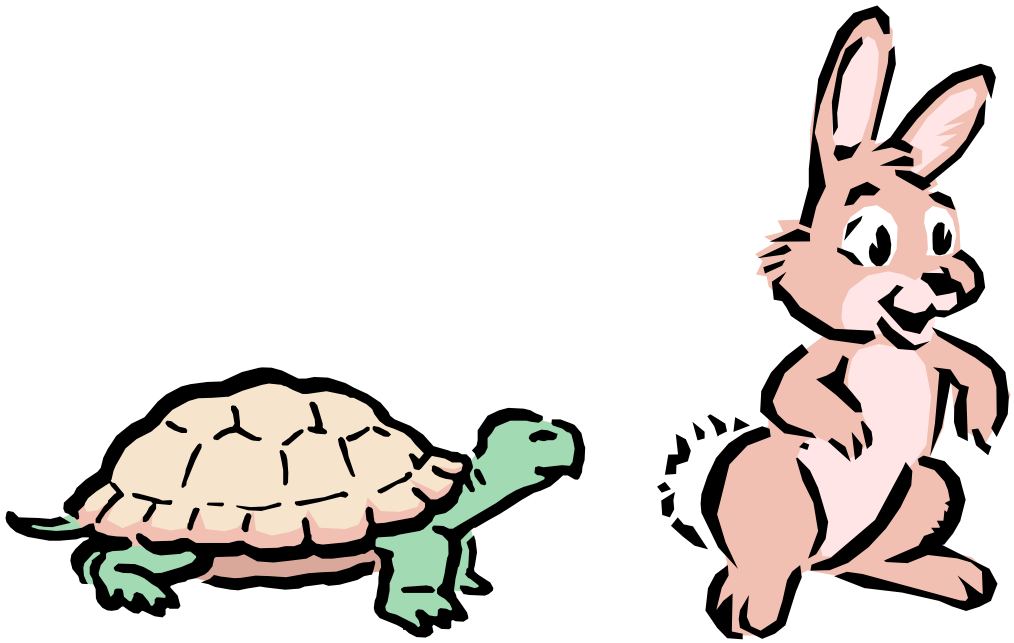
- The MIPS instruction set was designed for pipelining.
  - All instructions are the same length, to make instruction fetch and jump and branch address calculations simpler.
  - Opcode and operand fields appear in the same place in each of the three instruction formats, making instruction decoding easier.
  - Only relatively simple arithmetic and data transfer instructions are supported.
- These decisions have multiple advantages.
  - They lead to shorter pipeline stages and higher clock rates.
  - They result in simpler hardware, leaving room for other performance enhancements like forwarding, branch prediction and on-die caches.

Parallel processing

- One way to improve performance is to do more processing at once.
- There were several examples of this in our CPU designs.
  - Multiple functional units can be included in a datapath to let single instructions execute faster. For example, we can calculate a branch target while reading the register file.
  - Pipelining allows us to overlap the executions of several instructions.
  - Superscalar CPUs permit parallel execution of entire instructions.
- Memory and I/O systems also provide many good examples.
  - A wider bus can transfer more data per clock cycle.
  - Memory can be split into banks that are accessed simultaneously. Similar ideas may be applied to hard disks, as with RAID systems.
  - A direct memory access controller performs I/O operations while the CPU does compute-intensive tasks instead.

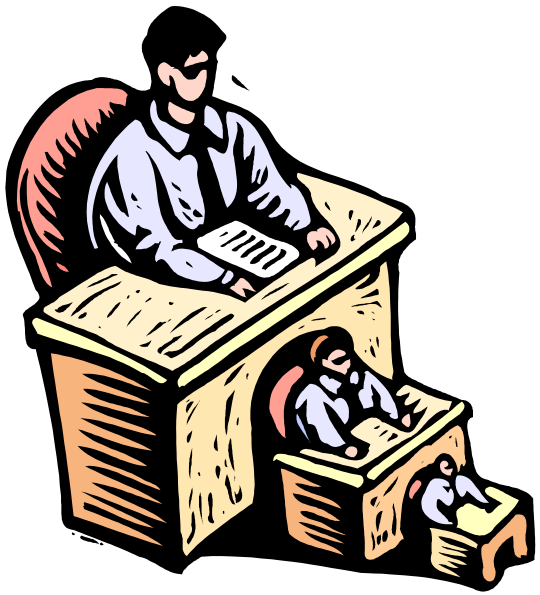
Amdahl's Law

- Performance is limited by the slowest component of the system.
- We've seen this in regard to cycle times in our CPU implementations.
  - Single-cycle clock times are limited by the slowest instruction.
  - Pipelined cycle times depend on the slowest individual stage.
- Amdahl's Law also holds true outside the processor itself.
  - Slow memory or bad cache designs can hamper overall performance.
  - I/O performance is also becoming increasingly important.



Hierarchical designs

- Hierarchies separate fast and slow parts of a system, and minimize the interference between them.
  - Caches are fast memories which speed up access to frequently-used data and reduce traffic to slower main memory.
  - Buses can also be split into several levels, allowing higher-bandwidth devices like the CPU, memory and video card to communicate without affecting or being affected by slower peripherals.



Architecture and software

- Computer architecture plays a vital role in many areas of software.
- Compilers are critical to achieving good performance.
  - They must take full advantage of a CPU's instruction set.
  - Optimizations can reduce stalls and flushes, or arrange code and data accesses for optimal use of system caches.
- Operating systems interact closely with hardware.
  - They should take advantage of CPU features like support for virtual memory and I/O capabilities for device drivers.
  - The OS handles exceptions and interrupts together with the CPU.
- Demanding applications rely on advanced system architectures.
  - The latest multimedia, database and supercomputing applications all benefit from modern architectures.
  - Consumer devices such as cell phones and Gameboys frequently have different requirements and need custom hardware and software.

Wait! Don't stop now!

- Those wacky electrical engineering people also have architecture classes.
  - ECE291 teaches Intel 8086 assembly language programming, and also presents some architectural concepts along the way.
  - ECE312 is similar to CS232, but a little more advanced and with actual processor simulations.
- The CS department has several higher-level architecture classes.

CS331	Embedded Systems Architectures and Software
CS333	Computer System Organization

- There are other related classes in our department also.

CS321	Programming Languages and Compilers
CS323	Operating Systems Design
CS326	Compiler Construction

Good luck on your exams and have a great summer!