rPeAnut

xvar : block #1 ; apply for a global variable xvar initialized to 1

xvar: block 10 ; a global array xvar of length 10, all values 0

store R3 xvar ; assign value to global variable xvar

Memory

Dynamic random access memory (DRAM) only requires a single transistor and capacitor per bit, this enables much higher density memory (compared with SRAM). However, it is considerably slower and more complex to access.

Static random access memory (SRAM) is fast volatile memory that is simple to interface with, however, it uses a lot more area than DRAM.SRAM will be found in: CPU caches, buffers for various devices within a computer system.

ROM is non-volatile so state is maintained without power.

**Flash memory** is an electronic [non-volatile](http://en.wikipedia.org/wiki/Non-volatile_memory) [computer storage](http://en.wikipedia.org/wiki/Computer_storage) device that can be electrically erased and reprogrammed.

Building a CPU

two basic approaches for the design of the control unit

Hardwire - these control units basically are finite state machines that are specially designed to sequence the CPU based on the instruction it is executing. (not particularly flexible and can be difficult to design, however, good performance)

Microcode - this approach uses ROM type memory within the CPU to act like machine code for sequencing the different action within the CPU. simpler and more flexible approach.

Performance

PUs have moved caches onto the CPU die which enables the CPU to be physically closer to the cache. This reduces latency.

Pipeling: involves in a number of independent stages, IF, ID, E, WB

Superscale architectures involve duplicating functional units within the cpu and then starting more than one instruction on the same clock cycle in the pipeline. This enables a larger throughput of instructions.

Sometimes instructions will require data from memory before they can execute, this will stall the pipeline. Slow the CPU.

The "Outer of order execution" approach loads the next few instructions and starts executing the instruction that has the required data

Multithread: CPUs can maintain the programming context of multiple threads (so duplication of state and register information), without the duplication of processing units, caches, TLBs, etc, This enables multiple threads to be executed within the one core. This can hide latency.

Multicores: The CPU can be duplicated, share the same memory and L2 Cache.

.

Linking

1. This binding of instructions and data to memory addresses may occur at :

compile time, load time, or execution time.Li

2. Libraries may be either:

statically linked - the library is compiled into the final binary executable.

dynamically linked - just a 'stub' is includes in the binary executable, the library code is obtained as needed during execution.

Memory Management and Paging

Addresses generated by the CPU are referred to as logical addresses. These are the addresses 'seen' by the user's programs.

Addresses seen by the main memory are referred to as physical addresses.

Physical memory is partitioned into fixed-sized blocks called frames.

Logical memory is partitioned into blocks of the same size called pages.

Logical addresses (produced by the CPU)are divided into two parts: the page number and the offset.

Each process has a page table.

A user or process would like to view memory as a set of variable-sized segments. Each segment has a name and there is no necessary ordering among segments.

Virtual Memory

Swapping is an approach that involves bringing an entire process from disk into main memory so it may be executed.

Demand-paging is a lazy "swapper" which brings into memory the pages of the process that are needed.

One approach to paging is to only bring pages into memory when they are needed. This is called Pure demand paging.

CPU Scheduling

In [*multiprogramming*](http://en.wikipedia.org/wiki/Multiprogramming) systems, the running task keeps running until it performs an operation that requires waiting for an external event (e.g. reading from a tape) or until the computer's scheduler forcibly swaps the running task out of the CPU. (multiprogramming is a CPU scheduling strategy.)

There are two types of schedulers:

non-preemptive - The CPU is not 'forcefully' taken from the process, and a

preemptive - The CPU may be forcefully taken from the process and switched to another process.

A variety of criteria are used to compare different scheduling algorithms.

CPU utilization - The percentage of time the CPU is in use.

Throughput - The rate at which processes are being completed.

Turnaround time - The interval of time from the starting a process to completing the process.

Wait time - The amount of time a process spends in the ready queue.

Response time - The amount of time it takes a process to start responding. This does not include the time to output the

response.

Three algorithms for CPU Scheduling:

First Come First Serve (FCFS)

Shortest Job First (SJF) provably shortest waiting time

Round Robin (RR)

Networks

Most Ethernet controllers have unique 48bit MAC address.

Ethernet's topology started as bus topology. But now LAN would generally use star topology with a switch at the center.

Weakness of bus topology: Only one device may transmit information to the bus at any time. If two units transmit at the same time a collision will occur, both transmission with usually be compromised and will need to be re-transmitted.

UDP/TCP

0.0.0.0 - This host (only source).

127.X.X.X - loopback.

10.X.X.X - private network

192.168.X.X - private network

255.255.255.255 - Broadcast on the local network.

IPv4 (Internet Protocol) provides 32bit addresses (decimal) for addressing computing devices that span the globe.

TCP(Transmission Control Protocol) provides a way of sending larger amounts of data to particular 'ports' on a host.

TCP will resend missing packets and check that packets arrive in the correct order.