

DDCA

by dcamenisch

Binary Numbers

An N-bit binary number ranges from 0 to $2^N - 1$.
The rightmost bit is the least significant bit (LSB).
The MSB is defined as the opposite.

Big Endian: größtes Byte zuerst

$$2^0 = 1 \quad 2^3 = 8 \quad 2^6 = 64$$

$$2^1 = 2 \quad 2^4 = 16 \quad 2^7 = 128$$

$$2^2 = 4 \quad 2^5 = 32 \quad 2^8 = 256$$

Little Endian: kleinstes Byte zuerst

$$2^3 = 8 \quad 2^6 = 512$$

$$2^4 = 16 \quad 2^7 = 1024$$

$$2^5 = 32 \quad 2^8 = 1'048'576$$

= Kilo

Hex. Dez. Binary	Hex. Dez. Binary	Hex. Dez. Binary	Hex. Dez. Binary
0 0 0000	4 4 0100	8 8 1000	C 12 1100
1 1 0001	5 5 0101	9 9 1001	D 13 1101
2 2 0010	6 6 0110	A 10 1010	E 14 1110
3 3 0011	7 7 0111	B 11 1011	F 15 1111

2's complement: negative numbers go by inverting every bit then adding 1, MSB used as sign flag.

Boolean Algebra

Rules: $X + X = X$ $X + \bar{X} = 1$ $X \cdot (Y + Z) = (X \cdot Y) + (X \cdot Z)$
 $X \cdot X = X$ $X \cdot \bar{X} = 0$ $X \cdot (Y \cdot Z) = (X \cdot Y) \cdot (X \cdot Z)$

De Morgan: $(X + Y + Z + \dots) = \bar{X} \cdot \bar{Y} \cdot \bar{Z} \dots$, $(X \cdot Y \cdot Z \dots) = \bar{X} + \bar{Y} + \bar{Z} + \dots$

Product of Sum:

Sum of Product:		
A	B	X
0	0	1
1	0	0 : $\bar{A} + B$
0	1	1
1	1	0 : $\bar{A} + \bar{B}$
\maxterm		
\minterm		
0	0	0
1	0	1 : $\bar{A} \cdot B$
0	1	0
1	1	1 : $A \cdot \bar{B}$

NAND Operations

NOT: \overline{AA} OR: $\overline{AA \cdot BB}$ AND: $\overline{AB} \cdot \overline{AB}$

Karnaugh Maps: used to minimize boolean equations, they work well with up to 4 variables. Some rules:

- use fewest circles possible
- only size $2^n \times 2^m$
- all must only contain 1's

Grey Code		
A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

$\Rightarrow X = \bar{B} + A \cdot \bar{C}$

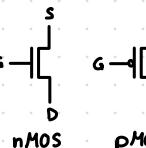
Logic Gates

AND	NAND	OR	NOR
A B X	A B X	A B X	A B X
0 0 0	0 0 1	0 0 0	0 0 1
0 1 0	0 1 1	0 1 1	0 1 0
1 0 0	1 0 1	1 0 1	1 0 0
1 1 1	1 1 0	1 1 1	1 1 0

XNOR	XOR	NOT
A B X	A B X	A X
0 0 1	0 0 0	0 1
0 1 0	0 1 1	1 0
1 0 0	1 0 1	1 0
1 1 1	1 1 0	0 1

Transistors

MOS transistors work like a switch there are two types of MOS transistors. pMOS conducts when G is LOW, nMOS is the opposite.



Verilog

D Flip-Flop

```
always @ (posedge clk)
begin
    q <= d;
end
```

↳ synchronous, if it should be asynchronous add rst to sensitivity list!

→ if d and q are multiple bits wide, the code implements multiple flip-flops.

Correct Code:

- wires: can't be on the left of $= / \leftarrow$ in an always block, are used to connect input and output.
- egs: can't be connected to output port of a module, can't be used in input port declaration, only on the left in always blocks.
- I/O: check if names match and if all ports are assigned.
- not multiple assignments to the same signal
- names can't start with numbers: ~~2good~~
- no module recursion

Combinational vs. Sequential:

Comb: all left hand signals get assigned, all inputs are in sensitivity list, all outputs are assigned

Building Blocks

MUX:

select one input, $\log_2(n)$ -bit control signal

PLA: array of AND gates, followed by an array of OR gates

Tri-State-Buffer: enables gating of signals onto wire

Decoder:

n inputs, 2^n outputs, one output is 1 depending on the input pattern

Combinational / Sequential Logic

- Comb: - no memory
Seq: - has memory
- no cyclic paths
- combines inputs to get output
- signals are assigned in every possible condition
ex. MUX
- all addends in sensitivity list

Finite State Machine

Goes through different state, where each state depends on the prev. state and the input.

Moore: Output depends only on current state

Mealy: Output depends on the current state and the input

State Encodings: - Binary Encoding (minimizes flip-flops), 00, 01, 10, 11
- One-Hot (max. flip-flops, min. next state logic), ~~0000~~
- Output (min. output logic)

Designing a FSM

- identify inputs/outputs
- state transition diagram
- write state transition & output table
- write boolean equations for next state

Area of FSM

FF = # bits for state $\times 2$

logic gates = count next state and output logic

Correctness of state diagram

- reset-line
- not multiple transitions for the same input
- no missing transitions
- no unmarked transitions
- initial state (if no reset)
- no mix of Moore/Mealy labeling

MIPS

J-Type: Jump / Branch Instructions

R-Type: Register for all operands ($OP = 0$)

I-Type: Instructions with an immediate/constant value

The caller calls a function, while the callee gets called.

The caller needs to take care of the temporary registers \$t0 - \$t9, while the callee needs to save and restore the preserved registers \$s0 - \$s7.

ISA and Microarchitecture

The ISA is the interface between software and hardware ("what the programmer sees").

The microarchitecture specifies the underlying implementation that actually executes the instructions.

Blocking and Non-Blocking Assignment Guidelines

- Use δ (posedge clk) and non-blocking ($=$) to model synchronic sequential logic
- Use continuous assignments to model simple combinational logic
- Use always $\delta(^*)$ and blocking ($=$) assignments to model more complicated combinational logic where the always statement is helpful
- Do not make assignments to the same signal in more than one always statement or continuous assig...

- mechanism to cache in a system, call in the QD
- number of groups - proposed register
- index of each bit in programmatic read/write configuration register

ISA

vs. Microarchitecture

- number of bits required for destination register of a load instruction
- Instructions: opcodes, addressing modes, data types, instruction type and format, registers, condition codes
- Memory: address space, alignment, addressability, virtual memory management, multiprocessor memory location
- Call, interrupt and exception handling
- I/O: memory mapped vs. instructions
- Power & Thermal management
- Multiprocessing / Multithreading support
- Access control, priority and privilege
- Memory-mapped location of exception vectors
- Function of each bit in a programmable branch prediction register
- Order of execution of loads and stores in multi-core CPU
- Program counter width
- Hardware FP-exception support
- Vector instruction support
- CPU endianness
- Virtual page size

Performance Evaluation

- CPI: cycles per instruction
 - IPC: instruction per cycle
 - MHz: frequency, 10^9 cycles/sec.
 - higher MHz \Rightarrow higher MIPS, IPS could be lower
 - higher MIPS \Rightarrow less time, could need more instructions
- \cdot MIPS: million instructions / sec. = MHz/CP
- \cdot Time = # instr. \cdot CPI \cdot Hz
- \cdot Speedup = oldTime/newTime

Single-Cycle Machines

Each instruction takes a single clock cycle and all state updates are made at the end of the cycle.

- slowest instruction determines cycle time
- + easy to build

Multi-Cycle Machines

Instruction processing is broken into multiple stages/cycles, state updates happen during execution and architectural updates at the end. Instruction processing consists of two components:

- Data path - relay and transform data
- Control logic - FSM that determines control signals
- + slowest stage determines cycle time

Dataflow

In a dataflow machine, a program consists of dataflow nodes. A node fires (executes) when all its inputs are ready.

Write-through

data written to cache block is simultaneously written to main memory

Write-back

dirty bit (1) is associated with each cache block. write back to main memory when evicted from cache

Pipelining

The idea is to process multiple instructions at once by keeping each stage occupied. In reality there are a few problems:

- Resource contention, can be fixed by duplication, increased throughput or detection and stalling
- Long latency operations
- Data dependencies, there are flow (read after write), output (write after write) and anti (write after read) dependencies. The last two exist due to a limited amount of registers.

Handling flow dependencies

- stall
- eliminate at software level
- predict values
- data forwarding
 - $\hookrightarrow W \rightarrow D$: internal/register file forwarding
 - $\hookrightarrow M \rightarrow E$: operand forwarding

Pipeline Stages

- Fetch: CPU reads instructions from instruction memory
- Decode: CPU reads source operands from register file and decodes instruction to control signals
- Execute: CPU performs a computation with the ALU
- Memory: CPU reads/writes data memory
- Writeback: CPU writes result to register file

Interlocking & Scoreboarding

Detection of data dependencies to ensure correct execution.

Out-of-Order Execution

Idea to move dependent instructions out of the way of independent ones. Reservation stage as rest are for dependent instructions.

Reorder Buffer

Complete instructions OoO but reorder them before making results visible to architectural state.

Tomasulo's Algorithm

Implementation of OoO-Execution. Uses register renaming to eliminate output and anti-dependencies. It further uses reservation stations for individual operations.

1. If reservation station is available:
 - instr. + renamed operands inserted into reservation station
 - rename destination registerElse stall
2. While in reservation station:
 - watch common data bus for tag of sources
 - if tag seen grab value
 - if both operands are valid inst. ready for dispatch
3. Dispatch instr. to functional unit
4. After instr. finishes:
 - put tagged value onto common data bus
 - if register file contains tag, update its value and set valid bit
 - reclaim rename tag \rightarrow no valid copy of tag in the system

VLIW

The idea is that the compiler finds independent instructions and statically schedules them into single VLIW instructions.

Lock step execution: if one instruction stalls, the whole VLIW stalls

- + simple hardware
- compiler needs to find N independent instructions per cycle
- + no dependency checking
- + no instruction distribution
- lock step causes stalls

Superscalar Execution

Idea is to fetch/decode/... multiple instructions per cycle.

- + higher IPC
- higher complexity for dependency checking \Rightarrow more hardware

Systolic Arrays

Instead of a single processing element (PE) we have a array of PE and carefully orchestrate the data-flow between them \Rightarrow Maximize computation done on a single element.

Difference from pipelining: Array structure is non-linear and multi-dimensional. PE connections can be multi-directional with different speeds. PEs can have local memory and execute kernels.

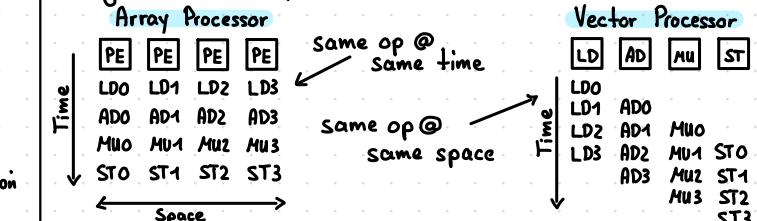
Fine Grained Multithreading

Hardware has multiple thread contexts (PC+reg) and each cycle the fetch engine fetches from a different thread.

- + no dependencies
- + no branch prediction
- + improved throughput, latency, tolerance, utilization
 - \hookrightarrow dependency checking between threads
- extra hardware
- reduced single-thread performance
- resource contention

SIMD

Single instruction operates on multiple data.



Formulas

$$\text{Execution time} = (\# \text{ instructions}) / (\text{instructions}/\text{cycle}) \cdot \frac{\text{cycles}}{\text{seconds}}$$

MIPS: million instructions per second

$$\text{Miss Rate} = \frac{\text{Number Of Misses}}{\text{Number Of Total Memory Accesses}} = 1 - \text{Hit Rate}$$

$$\text{Average Memory Access Time} = t_{cache} + MR cache (t_{main} + MR main t_{main}) / AMAT$$

L = main memory, VM = virtual memory, MR = miss rate

Number of set bits: $\log_2 S$

Remaining tag bits indicate memory address of the data stored in a given cache set. Two least significant bits \rightarrow byte offset

Vector Processing

Performs operation on a whole array. This is only possible if the operations on each element are independent from each other.

The data gets stored in vector registers. Vector chaining describes the vector version of data forwarding. It allows a operation to start as soon as an individual element is ready. The stride is the distance of the vector elements in memory. If a vector is too long it can be split into multiple vectors (strip mining).

- + a lot of work per instruction - works only if parallelism is regular, else it is very inefficient
- + regular memory access pattern
- + no need for loops

GPU

GPUs are SIMD engines but programmed using threads (SPMD). A set of threads executing the same program are grouped into a warp.

Dynamic Warp merging: Merge threads executing the same instr. after branch divergence. This forms new warps from the warps waiting.

Delayed Branching

Means that some instructions after a branch are executed regardless of which way the branch goes. A compiler can instructions in such a delay slot if they don't influence the branch itself, else they are filled with NOPs.

Branch Prediction

A technique used to predict the next address after a branch. If the prediction is wrong, we have to flush the pipeline (misprediction penalty).

Prediction Direction Schemes

- always (not)-taken (30-40%) 60-70% accuracy
- BTBN: backwards taken forwards not taken (good with loops)
- Last time predictor: single bit stored in BTB (branch target buffer) indicates last direction. Loop accuracy = $\frac{N-2}{N}$.
- 2-bit counter based prediction:
 - Local: no interference between different branches
 - Global: single counter for all branches

Cycles

Total number of cycles can be expressed as $C = P(I - 1) + B + D$

- C = total number of cycles taken
- P = total number of pipeline stages
- I = total number of branches
- B = total number of conditional branch instructions executed (including executed and predicted)
- D = number of cycles stalled for each condition of branch

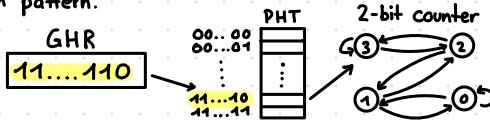
Delays

Propagation delay: typ: maximum time from when an input changes until its output changes reach their final value.
Computation delay: typ: minimum time from when an input changes until any output starts to change its value.

Global Branch Correlation

The idea is that recently executed branch outcomes are correlated with the outcome of the next branch.

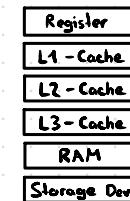
- First level: Global branch history register, keeps track of the last branch outcomes.
- Second level: Pattern history table, keeps a 2-bit counter for each pattern.



Memory Hierarchy

Memory Array: stores data, address selection logic selects row, readout circuitry reads data.

Memory banking: Multiple memory units with a common data and address bus, helps to resolve long latency. Units can be accessed individually.



Locality: temporal = access to same address in short time
spatial = access to nearby address

Blocks & Addressing cache:

- memory is divided into fixed-size blocks
- each block maps to a location in the cache (index bits)
- for a cache hit the tags need to match



Offset = Byte im Block
Index = Zeile in Tag Store
Tag = Welches Block im Set

Associativity:

- multiple blocks have the same index → conflict misses
- n-way associative allows n-blocks with same index
- 1-way → direct mapped no index → fully associative

Cache Performance:

- cache size, total data c
- block size b
- associativity n

$$\# \text{blocks} : B = c / b$$

$$\# \text{sets} : S = B / n$$

Replacement Policies:

- FIFO, first-in-first-out
- LRU, least-recently-used
- Random

Handling Writes

Writeback: write to lower levels when the block is evicted, needs a dirty bit.

Writethrough: write to all levels immediately, simpler but bandwidth intensive.

Classification of Misses

Compulsory Miss: first reference is always a miss (prefetching)

Capacity Miss: cache is too small

Conflict Miss: all other misses (more associativity)

Improvement Ideas

- reduce miss rate
- reduce miss latency or cost
- reduce hit latency or cost

$$\text{bits of storage} = \text{associativity} \cdot (\text{tag} + \text{dirty} + \text{valid}) + \text{LRU}$$

Prefetching

The idea is to improve cache performance by preloading data to avoid misses. There are different techniques to prefetching, some are software based while others hardware dependent.

Stride prefetcher: prefetches cache block in a pattern with a certain stride (if stride = 0, next block prefetching)

Runahead execution: allows the processor to pre-process instr. during cache misses instead of stalling. Therefore it can detect potential cache misses earlier.

- accuracy = #pref. used / #pref. total
- coverage = #acc. predicted / #total acc.

Virtual Memory

Much larger than physical memory. Virtual address space is divided in pages, while physical address space is divided into frames. Page Table Stores mapping: Virtual → Physical together with a valid bit (and more meta data).

$$\# \text{Virtual Pages} = \frac{\text{Virtual Address}}{\text{Page Size}}$$

$$\# \text{Physical Pages} = \frac{\text{Physical Address}}{\text{Page Size}}$$

$$\text{VA} : \frac{\text{Virtual Page Number}}{\text{Page Offset}}$$

$$\text{PA} : \frac{\text{Page Table}}{\text{P. Page Number}}$$

$$\downarrow \text{same a physical tag}$$

$$\text{Physical Address Space} = \frac{\text{PPN}}{\text{Page Offset}}$$

Multi-level page tables: keeps PT size small

Memory protection: different PT for each program

Translation Lookaside Buffer TLB: cache PT entries to speed up address translation

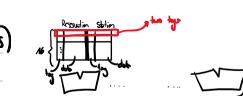
A memory address a is valid to one digit when a is a multiple of n (where n is a power of 2). In this case, a is the smallest unit of memory access, i.e. each memory address specifies a different byte. An n-digit address would have n bytes (b) and significant parts like most significant byte.

Cache capacity = (Block Size in Bytes) * (Blocks per Set) * (Number of Sets)

Index Bits = log₂(Blocks per Set)

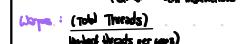
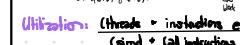
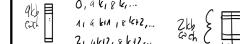
Block Offset Bits = log₂(Block Size in Bytes)

Tag Bits = (Address Bits) - (Index Bits) - (Block offset Bits)



Software Interlocking: NOPs

Hardware Interlocking: Detects dependencies and stalls pipeline accordingly



Min Utilization: following different paths when in general as many parallel tasks like the shared path, but for each other parallel tasks we use 1 thread. If there is a copy thread, the inner threads there are just not send to further threads + 0 (but needs to be in the memory)

Ex. Find the simplest sum-of-products form for this equation: $F = B + (A + \bar{C}) \cdot (\bar{A} + \bar{B} + \bar{C})$

$$\begin{aligned} F &= B + A\bar{A} + A\bar{B} + A\bar{C} + \bar{C}\bar{A} + \bar{C}\bar{B} + \bar{C} \\ &= B + A + \bar{C} \end{aligned}$$

Ex. Simplify the following min-terms: $\sum(3, 5, 7, 11, 13, 15)$. $\{3, 5, 7, 11, 13, 15\} = \{0011, 0101, 0111, 1011, 1101, 1111\}$

$$\begin{aligned} F &= \bar{A}\bar{B}\bar{C}\bar{D} + \bar{A}\bar{B}\bar{C}\bar{D} + \bar{A}\bar{B}\bar{C}D + \bar{A}\bar{B}CD + \bar{A}B\bar{C}\bar{D} + ABCD \\ &= CD \cdot (\bar{A}\bar{B} + \bar{A}\bar{B} + \bar{A}\bar{B} + AB) + BD \cdot (\bar{A}\bar{C} + A\bar{C}) \\ &= CD + BD\bar{C} \\ &= D(B + C) \end{aligned}$$

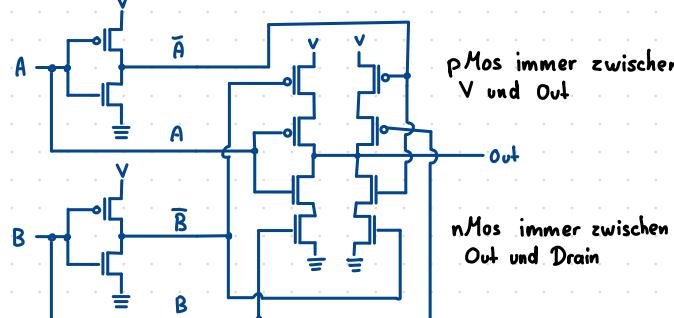
Ex. Convert the following equation to only contain NANDs.

$$\begin{aligned} F &= (\bar{A}\bar{B} + C) + AC = (\bar{A}\bar{B} + C) \cdot \bar{AC} = (\bar{A}\bar{B} \cdot \bar{C}) \cdot \bar{AC} \\ &= \bar{AB} \cdot \bar{\bar{AC}} = \bar{AB} \cdot \bar{AB} \cdot \bar{AC} \end{aligned}$$

Ex. Convert the following equation to only contain NANDs.

$$\begin{aligned} F &= (\bar{A} + BC) + \bar{C} = (\bar{A} + BC) \cdot \bar{C} = (\bar{A} + BC) \cdot C \\ &= (\bar{A} \cdot (BC)) \cdot C = (\bar{A} \cdot A) \cdot (B \cdot C) \cdot C \end{aligned}$$

Ex. Draw a XOR-Gate with transistors.



Ex. Sequential or Combinational circuit?

```
module one (input clk, input o, input b, output reg [1:0] q);
  always @(*)
    if (b)
      q <= 2'b01;
    else if (a)
      q <= 2'b10;
endmodule
```

This code results in a sequential circuit because a latch is required to store old values of q if both conditions are not satisfied.

Ex. Is this code a correct multiplexer?

```
module four (input sel, input [1:0] data, output reg z);
  always @(*)
    if (sel)
      z = data[1];
    else
      z = data[0];
endmodule
```

No, the input data is missing in the sensitivity list. A update would not be reflected to the output z.

Ex. Does this result in a D-FlipFlop with a synchronous active-low reset?

```
module mem (input clk, input reset, input [1:0] d, output reg [1:0] q);
  always @(*)
    if (!reset) q <= 0;
    else q <= d;
endmodule
```

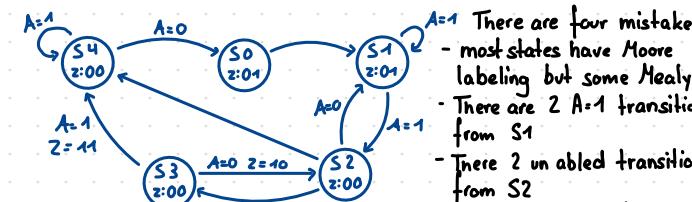
The code implements 2 D-FlipFlops, each works with a asynchronous active low reset.

Ex. Is this code syntactically correct?

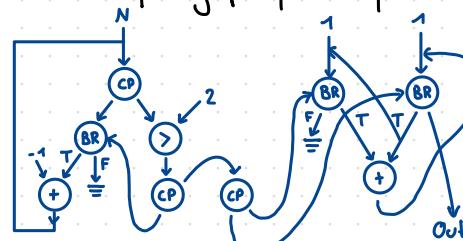
```
module fulladd (input a, b, c, output reg s, c.out);
  assign s = a ^ b;
  assign c.out = (a & b) | (a & c) | (b & c); we use assign therefore these have to be wires
endmodule
```

```
module top (input wire [5:0] instr, input wire op, output z);
  reg [1:0] r1, r2; should be wires
  wire [3:0] w1, w2;
  fulladd FA1 (.a(instr[0]), .b(instr[1]), .c(instr[2]),
                .c.out(r1[1]), .z(r1[0]));
  fulladd FA2 (.a(instr[3]), .b(instr[4]), .c(instr[5]),
                .c.out(r2[1]), .z(r2[0]));
  assign z = r1 | op;
  assign w1 = r1 + 1;
  assign w2 = r2 << 1;
  assign op = r1 ^ r2; multiple drivers
endmodule
```

Ex. List all the mistakes in this diagram.



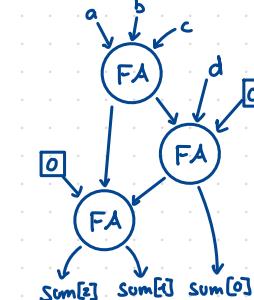
Ex. Draw a data flow graph for the fibonacci function.



Ex. Which designs are compatible with each other?

superscalar - in-order precise exceptions - out-of-order retirement
superscalar - out-of-order branch prediction - fine-grained multithreading
single cycle - branch prediction fine-grained multithreading - pipelining
reservation station - microprogramming Tomasulo's algorithm - in-order
fine-grained multithreading - single core direct mapped cache - LRU replacement

Ex. Draw the dataflow graph for a four 1-bit addition, you can use Full Adder nodes.



Ex. Any $n \geq 3$ 1-bit addition can be implemented only using Full Adders. Fill out the table.

n	# required FAs	n	# required FAs
3	1	6	4
4	3	7	4
5	3	8	7

Ex. Two programs A, B run on the same machine, both have the same # memory requests, but A needs to stall way more. Why could this be?

A could have a lot of row buffer conflicts, while B has a lot of row buffer hits.

Ex. If a processor executes more IPS, does a program always finish faster?

No, the number of instructions for a program could be different for different processors.

Ex. If a program runs on a processor with a higher frequency, does this imply that it executes more IPS?

No, a processor with a lower frequency can have a much higher number of IPC.

Ex. Write a MIPS 64-bit subtraction (2s-complement) where $\$4 \$5 - \$6 \7 .

```
subu $3, $5, $7
sltu $2, $5, $7
add $2, $6, $2
sub $2, $4, $2
```

Ex. A machine with 5 pipeline stages uses delay slots to handle control dependences. Jump and branch are resolved during execution stage. How many delay slots are needed?

2, since we can fill them during fetch and decode of the jump / branch instruction.

Can we modify the pipeline to reduce the number of delay slots?

Yes, if we move the resolution of the jump/branch target to the decode stage, we only need one delay slot.

Ex. How many delay slots are needed for the following implementations?
 In-order with branch resolving during 4th stage: 3
 000 with 64 reservation stages, branch resolving during 2nd cycle of branch execution and 16 stages before the execution stage: Don't know

Ex. Given the following microbenchmark for a pipelined machine.
 Calculate #dynamic instructions executed, # pipeline stages and #cycles of stall caused by branch instruction.

LOOP1: Initial R1 #Cycles

SUB R1, R1, #1	4	51
BGT R1, LOOP1	8	63
	16	87

LOOP2: B LOOP2 all runs execute the same #dynamic instr.

Let: $C = \# \text{cycles}$

$P = \# \text{stages}$

$I = \# \text{dynamic instr.}$

$B = \# \text{branch instr.}$

$D = \# \text{cycles stall / branch} \Rightarrow P + I = 40, D = 3$

$$C = P + T - 1 + B \cdot D$$

$$51 = P + I - 1 + 4D$$

$$63 = P + I - 1 + 8D$$

$$87 = P + I - 1 + 16D$$

Ex. Given a scalar processor with in-order fetch, out-of-order dispatch and in-order retirement. It has 4 pipeline stages, and 2 reservation stations (one for each type). If the following program gets executed, answer the questions?

```

MOV R0<-1000 FDE1E2E3E4W
LD R1<[R0] FD - - E1E2E3E4E5E6E7E8W
BL R1,100,LB1 FD - - - - E1E2E3E4W
MUL R1<R1,5 FD E1E2E3 // killed
ST [R0]< R1 FDE1E2E3E4W
ADD R1<R1,R0 FD - - - E1W
ST [R0]< R1

```

Cache hit latency? 1 cycle, the last ST instr. is a hit.

Cache miss latency? 8 cycles, the first LD instr. misses.

Cache line size? Unknown

#entries in each reservation station? ALU at least 2, MU unknown

#ALUs? if pipelined at least 1, else at least 2.

Is the ALU pipelined? If there is only 1 ALU yes, else unknown

Does the processor have branch prediction? Yes, because there are instr. that get killed.

At which stage do branches get resolved? At the end of E4, because in the next cycle the previously fetched instr. get killed.

Ex. Given Tomasulo's Algorithm with: 8 functional units with their own tag/data bus, 32x64 bit registers, 16 reservation stations per functional unit and 2 source register per reservation station, calculate:

$$\# \text{tag comparator / reservation station entry} = 2 \times 8 = 16$$

$$\# \text{tag comparators} = 16 \times 16 \times 8 + 32 \times 8 = 2304$$

$$\min. \text{tag size} = \log(16 \times 8) = 7$$

$$\min. \text{size of register alias table} = 32 \times (7 + 64 + 1) = 2304$$

$$\min. \text{total size of tag store} = 32 \times 7 + 8 \times 16 \times 2 \times 7 = 2048$$

Ex. Comparing a VLIW and an in-order superscalar processor with the same machine width and frequency. For a program A, the VLIW machine is much faster, why could this be?

The superscalar proc. is in-order, requiring bubbles in the pipeline, while the VLIW proc. can re-order instr.

For some other program B, the VLIW is slower, why could this be?

VLIW needs NOPs, while the superscalar proc. doesn't. These NOPs can lead to lower L1-cache hit rate and higher fetch bandwidth.

Ex. Which of the following are goals of VLIW?

- i. Simplify code compilation
- ii. Simplify application development
- iii. Reduce overall hardware complexity
- iv. Simplify hardware dependence checking
- v. Reduce processor fetch width

Ex. Given a vector proc. with these fully interleaved / pipelined instr. VLD/VST 50 cycles, VADD 4 cycles, VMUL 16 cycles, VDIV 32 cycles and VRSHFA 1 cycle. Assume: in-order pipeline, chaining between functional units, first element bank 0, 8KB row buffer / bank, 64 bit vector elements, each memory bank has 2 ports and there are 2 load/store units. What is the minimum (power of 2) # banks so memory accesses never stall? 64 banks, because access latency is 50ms and 64 is the next power of 2.

Executing this program takes 111 cycles, what is the vector length?

VLD	V1, A	50	L-1	111 = 51 + 4 + 16 + 1 + L-1
VLD	V2, B	111	50	=> L = 40
VADD	V3, V1, V2	4	L-1	
VMUL	V4, V1, V3	16	L-1	
VRSHFA	V5, V4, 2	111	L-1	

Reducing the banks by a factor of 2, how long does the program take?

VLD	[0]	50		
	[31]	50		
	[33]	7	50	1 + 50 + 7 + 50 + 4 + 16 + 1 = 129
	[32]	7	50	=> 129 cycles
VLD	[0]	11	50	
	[31]	50		
	[33]	7	50	
VADD		4		
VMUL		16		
VRSHFA		11		

tracking the last element

Now the #banks get reduced further and it takes 279 cycles. How many banks are there?

$$279 = 1 + 16 + 4 + 1 + 7 + \lceil \frac{1}{4} \rceil \cdot 50$$

$$\Rightarrow 5 = \lceil \frac{1}{4} \rceil \cdot 7 \Rightarrow x = 8 \text{ memory banks}$$

In a new version 4 vector proc. share the same memory with 4 times the banks. However the execution is slower than if each program ran on a single proc. with 1/4 banks, why could this be?

Row buffer conflicts as all cores interleave their vectors across all banks.

How can this be fixed?

Partition the memory mappings, or use better memory scheduling.

Ex. Consider the following warps, how can dynamic warp formation be used?

$$X = \{1\ 0\ 0\ 1\ 0\ 1\ 1\ 1\}$$

$$Y = \{1\ 0\ 0\ 0\ 1\ 0\ 0\ 1\}$$

$$Z = \{0\ 1\ 0\ 0\ 0\ 0\ 0\}$$

X' = {1 0 0 1 0 1 1 1}
 Y' = {1 1 0 0 1 0 0 1}
 Z' = {0 0 0 0 0 0 0 0}

There are several answers, but notice that X, Y can't be merged.

Ex. How effective is a 16KB, 4-way associative cache with 8B instructions?

Not effective, since the block size is 4B, each instruction needs two accesses. Further it can't exploit spatial locality.

Ex. Given the following access pattern and hit rate for a cache determine its characteristics.

Addresses Accessed	must miss	Hit rate
1. 0 4 8 16 64 128		1/2
2. 31 8192 63 16384 4096 8192 64 16384		5/8
3. 32368 0 123 1024 3072 8192 1	hit	1/3

Cache block size: 8, 16, 32, 64 or 128B

From ① we can see that only 32 or 64 are possible. From ② we can see that 63 must be a hit and therefore it can only be 64B.

Cache Associativity: 1, 2, 4 or 8 way

Combining this with the possible cache sizes of 4 or 8KB we can see that 1 and 2 way would cause too much misses in ② and 8 way would cause another miss in ③, therefore it must be 4 way.

Cache size: 4 or 8 KB

In ③ the access to 0 is a miss and therefore 8192 should be a hit, but with 4KB, 1024 and 3072 would map to the same set and therefore it couldn't be a hit. So cache size must be 8KB.

Replacement Policy: LRU or FIFO

For 8192 to hit in ③ it must be LRU.

