Very short answer questions. "True" and "False" are considered very short answers.

[1] Does peak performance track observed performance?

[1] Correctly predicting that a branch is taken is not enough. What else is necessary?

[1] Write down the average memory access time equation.

[1] How do two processes communicate when running on a shared memory machine?

[1] Which type of cache miss can be changed by altering the mapping scheme?

[1] Which type of cache miss can be reduced by using longer lines?

[1] Which type of cache miss can be increased by using longer lines?

[1] If we make transistors and wires smaller and smaller but make no other changes, what happens to the power density?

[1] Superscalar and VLIW are 2 ways to increase ILP. What is the primary difference between them?

[2] There are two main ways to define performance - what are they?

[2] When we talk about the number of operands in an instruction (a 1-operand or a 2-operand instruction, for example), what do we mean?

[2] As speculation _____ power consumption _____

[1] The NVIDIA GPU exploits multiple _____ in order to achieve maximum performance.

[2] Coherence refers to _____ is returned while consistency has to do with _____.

[2] What are the two biggest challenges to obtaining a substantial decrease in response time when using a MIMD parallel processor?

[2] What is "leakage" current? If Vdd is lowered, what happens to the amount of leakage current?

[2] What is the main difference between a commodity cluster and a custom cluster?

Short Answers (25 or fewer words)

[3] The power consumed by a chip has increased over time, but the clock rate has increased at a far greater relative rate. How was this possible? Explain how designers keep the chip from melting.

[2] Why is it so difficult for the processing elements on a CMOS-based chip to communicate with things that located off the chip?

[2] Area on the die used to be the most critical design constraint. That is no longer true - what is the most critical factor now? Why?

[3] Does the design of the instruction set have a significant impact on the ability to pipeline a processor? If so, explain your answer.

[2] What is a benchmark program?

[2] Do benchmark programs remain valid indefinitely? Why or why not?

[2] Why is it difficult to come up with good benchmarks that will work across all types of parallel processors?

[2] Which is easier to write a program for, a shared memory machine or a message passing machine? Why?

[2] Which is more expensive to build - a shared memory machine, or a message passing machine? Why?

[9] What do the following acronyms stand for?

MSI	**SIMD**	**UMA**
WAR	**GPU**	**TLB**
VMM	**VLIW**	**ILP**

[4] In your first design of a 4-stage pipeline (F,D,E,M/W) F takes 27 time units, D takes 24, E takes 52, and M/W takes 25.

a) What will the clock cycle time be for this pipeline?

b) Is it a balanced pipeline? If not, explain what you could do to fix it. What would the cycle time be now?

[3] Briefly outline how a Vector machine works, and what type of parallelism it is exploiting.

[4] Supporting precise interrupts in a machine that allows out of order completion is a challenge. Why do we care? What is a precise interrupt, and why is it important to support precise interrupts?

[3] What is the definition of a basic block? Why is there a desire to create a bigger one? Give one example of a way to create a bigger basic block.

[3] Slow and wide architectures can be more power efficient than fast and narrow architectures. Explain why. Also, explain the underlying assumption that is being made, and why it is that we are still making narrow fast machines.

[2] What makes one instruction set harder to write a virtual machine monitor for than another one?

[3] Compilers have a hard time guaranteeing program correctness when doing static scheduling. Why is that? Give an example of a case where something might go wrong. (Hint - think Memory System)

[3] What is a Tournament predictor? Describe what it is, how it works, and why it is used. Your description can (and probably should) include sketches and drawings.

[8] For each of the following techniques, indicate how (on average) the 3 terms of the AMAT equation are affected. Only mark the ones that change - if a term is unaffected, just leave it blank. The first one is done for you as an example.

Technique	Hit Time	Miss Rate	Miss Penalty
Increasing Block Size		**Down**	**Up**

Technique	Hit Time	Miss Rate	Miss Penalty
Increasing Associativity			
Compiler optimizations			
Decreasing Cache Size			
Non-blocking Cache			
Hardware/Software Prefetching			
Critical Word First			
Multilevel Cache			
Victim Cache			

[7] The standard MIPS has a 5-stage pipeline, and uses a load and a branch delay slot. Assume the machine is redesigned to be a 9-stage pipeline, with the following stages:

F D RR E1 E2 E3 M1 M2 WB (where **RR** stands for Register Read)

In this pipeline, logical operations are finished during E1, arithmetic during E2, and multiplication/division during E3. There are no load or branch delay slots.

a) Assuming this machine has a branch predictor and the branch condition is calculated by the end of the E2 stage, how big is the branch penalty (measured in cycles) when the prediction is incorrect? What if the branch condition is not calculated until the end of E3?

b) How many load delay slots would this machine need (assuming it has forwarding logic and you are forwarding to E1) assuming the memory returns the value by the end of M1? M2?

c) What type of data hazard does the above pipeline need to worry about?

d) If the above pipeline were modified to support out of order completion, what new data hazard would be introduced?

e) If in addition to completing out of order, instructions were allowed to issue out of order, what new data hazard would be introduced?

[3] The designer has the choice of using a physically addressed cache or a virtually addressed cache. Explain the difference, and give 1 advantage for each.

[3] Assuming a 23-bit address and a 1k-byte Direct Mapped cache with a linesize=2, show how an address is partitioned/interpreted by the cache.

[3] Assuming a 23-bit address and a 160-byte 10-way SA cache with a linesize=8, show how an address is partitioned/interpreted by the cache.

[2] Assuming a 23-bit address and a 212-byte FA cache with a linesize=4, show how an address is partitioned/interpreted by the cache.

[2] Given a 32 Megabyte physical memory, a 31 bit Virtual address, and a page size of 1K bytes, write down the number of entries in the Page Table, and the width of each entry. Is there a problem with this configuration? If so, how can you fix it?

[4] Given a 1 Gigabyte physical memory, a 48 bit Virtual address, and a page size of 2K bytes, write down the number of entries in the Page Table, and the width of each entry. Is there a problem with this configuration? If so, how can you fix it?

[2] The CPI of the WhizBang 8000 is 2, assuming a perfect memory system (all references to memory take a single cycle.) Assume we are able to keep a perfect instruction memory system but must use a more realistic data memory system - a data cache with a miss rate of 10% and a miss penalty of 40 cycles. If 35% of all instructions are loads or stores, what does the CPI become in this case?

[9] You have been writing C programs for a simple, non-pipelined machine. You have recently received a promotion, and now your job is to write C programs for a heavily pipelined, high performance processor. This processor has an advanced tournament-style branch predictor, and your code will be compiled using a highly optimizing compiler. Your programs must execute as fast as possible (the emphasis is on response time, not throughput). Give at least 3 examples of things you should do differently in your program now, and be sure to explain in detail what the problem is you are addressing, and how you are going to overcome it in the software.

[3] You are responsible for designing a new embedded processor, and for a variety of reasons you must use a fixed 18 bit instruction size. You would like to support 64 different opcodes, use a 3-operand instruction format, and have 32 registers. If it is possible to do this, draw what an instruction would look like. If it is not possible, explain why, and give at least 2 different ways to solve the problem.

[2] An important program spends 70% of its time doing Integer operations, and 30% of its time doing floating point arithmetic. By redesigning the hardware you can make the Integer unit 40% faster (take 60% as long), or you can make the Floating Point unit 99% faster (take 1% as long). Which should you do and why? (You must show your work to get full credit.)

[5] Here is a code sequence.

```
load    R1, 0(R10)

sub     R2, R1, R2

store   R3, 20(R2)

load    R6, 10(R8)
```

Assuming a standard 5-stage pipeline that supports hazard detection and does forwarding,

a) Indicate all dependencies (draw lines/arrows between them, and write beside each line/arrow which hazard is involved).

b) Insert as many No Operations (NOPS) as required in order to ensure this code runs correctly. (Remember, writes to the register file occur on the first half of the cycle, and reads occur during the second half).

c) Schedule the code to remove as many stalls as possible. How many NOPS are left?

d) There are things the static scheduler does not know which makes guaranteeing correctness difficult. Give an example of something that might be true in the above code which will make scheduling and removing all NOPs a problem.

[5] In an MOS device, there is a gate, drain, and source. Briefly explain how this device works. Pictures are welcome.

[8] Here is a code sequence.

```
        lw      R8, 8(R9)

        add     R7, R1, R7

Label:  lw      R3, 0(R10)

        add     R4, R2, R3

        sw      R4, 20(R2)

        sub     R4, R6, R7

        add     R5, R9, R6

        breq    R3,R6,Label         ; branch to Label if R3=R6
```

Assuming a 6-stage pipeline (F,D,E1,E2,M1,WB) that does not support hazard detection, does no forwarding, and does not use a branch delay slot,

a) Indicate all dependencies (draw lines/arrows between them, and write beside each line/arrow which hazard is involved).

b) Insert as many No Operations (NOPS) as required in order to ensure this code runs correctly. (Remember, writes to the register file occur on the first half of the cycle, and reads occur during the second half).

c) Circle the NOPs that can be removed if forwarding and hazard detection logic is implemented. Assume data is needed at the beginning of E1.

d) Assuming the pipeline has been modified so that it does support hazard detection and forwarding, schedule the code to remove as many stalls as possible. Assume there are no hazards through memory. How many NOPS are left?

[6] In class, we talked about the cycle by cycle steps that occur on different interrupts. For example, here is what happens if there is an illegal operand interrupt generated by instruction i (note this machine has a 7 stage pipeline and supports imprecise interrupts. RR stands for Register Read):

	1	2	3	4	5	6	7	8	9	10	11	12
i	IF	ID	RR	EX	M1	M2	WB	<- Interrupt detected				
i+1		IF	ID	RR	EX	M1	M2	WB	<- Instruction Squashed			
i+2			IF	ID	RR	EX	M1	M2	WB	<- Trap Handler fetched		
i+3				IF	ID	RR	EX	M1	M2	WB		
i+4					IF	ID	RR	EX	M1	M2	WB	

Fill out the following table if for the same machine, instruction i experiences a fault in the EX stage (Overflow, for example):

	1	2	3	4	5	6	7	8	9	10			
i	IF	ID	RR	EX	M1	M2	WB						
i+1		IF	ID	RR	EX	M1	M2	WB					
i+2			IF	ID	RR	EX	M1	M2	WB				
i+3				IF	ID	RR	EX	M1	M2	WB			
i+4					IF	ID	RR	EX	M1	M2	WB		
i+5						IF	ID	RR	EX	M1	M2	WB	
i+6							IF	ID	RR	EX	M1	M2	WB

Assuming precise interrupts are being supported, what happens in this case?

	1	2	3	4	5	6	7	8	9	10			
i	IF	ID	RR	EX	M1	M2	WB	<- M1 stage has page fault					
i+1		IF	ID	RR	EX	M1	M2	WB	<- Inst Decode has Illegal Instruction				
i+2			IF	ID	RR	EX	M1	M2	WB				
i+3				IF	ID	RR	EX	M1	M2	WB			
i+4					IF	ID	RR	EX	M1	M2	WB		
i+5						IF	ID	RR	EX	M1	M2	WB	
i+6							IF	ID	RR	EX	M1	M2	WB
i+7								IF	ID	RR	EX	M1	M2	WB

What is the maximum number of exceptions that could happen at a single time in the above machine? Explain how you got your answer.