Final Exam Study Guide

Very short answer questions. You must use 10 or fewer words. "True" and "False" are considered very short answers.

(1) **[1]** Predicting the direction of a branch is not enough. What else is necessary?

Branch Target Address
(Exam 2 2021 Question 5)

(2) **[1]** Which is on average more effective, dynamic or static branch prediction?

Dynamic

(3) **[1]** Does an average program's locality behavior remain the same during the entire run of the program?

No
(Exam 2 2020 Question 2)

(4) **[1]** Is peak performance usually the same as sustained performance?

{Charlie } No

(5) **[1]** Which type of cache miss can be reduced by using shorter lines?

Coherence miss

(6) **[1]** Which type of cache miss can be reduced by using longer lines?

Compulsory miss
(Exam 2 2020 Question 4)

(7) **[1]** Using a different mapping scheme will reduce which type of cache miss?

Conflict miss
(Exam 2 2020 Question 4)

(8) **[1]** What pipeline hazard can be avoided by "throwing money at the problem"?

Structural hazard

(9) **[1]** What pipeline hazard can be avoided using a technique known as value prediction?

Data hazard (RAW)
(Textbook, page 217 -- Annie confirmed)

(10) **[1]** Give 1 advantage to using a VLIW.

Simple, does not need that much hardware and therefore fast & cool
(Exam 2 2020 Question 5)

(11) **[1]** Give 1 disadvantage to using a VLIW.

Compiler has no knowledge of runtime info (can lead to memory hazards), terrible at handling unexpected events (interrupts), bad at branch prediction, change in hardware means you'd need to recompile, has to insert NOPs if can't find enough instructions.
(Exam 2 2020 Question 5 -- help me make this a "short answer"?)

(12) **[1]** Give 1 advantage to using a superscalar. Your answer must be different from your VLIW answer.

(13) **[1]** Give 1 disadvantage to using a superscalar. Your answer must be different from your VLIW answer.

has a limited window over which it can schedule code.
(Exam 2 2020, Question 13)

(14) **[1]** As transistors and wires shrink, what happens to the power density?

Increases

(15) **[2]** There are two main ways to define performance - what are they?

Response time & throughput

(16) **[2]** There are two major challenges to obtaining a substantial decrease in response time when using the MIMD approach. What are they?

Communication and finding parallelism

(17) **[2]** If I add processors but keep the job size the same, am I measuring strong or weak scaling? Does this correspond most closely to response time or throughput?

Strong scaling. It corresponds most closely to response time.
(Exam3.w20.pdf, Question 10)

(18) **[2]** Give a one-word definition of coherence, and a one-word definition of consistency.

Coherence: What. Consistency: When

Short Answers (20 or fewer words)

(19) **[2]** What is a benchmark program?

It's a representative program that represents what we do with the machine.
(Exam 1 2020, Question 12)

(20) **[2]** Do benchmark programs remain valid indefinitely? Why or why not?

Benchmarks don't remain valid indefinitely. If the way we use the machine changes, benchmarks must change to adapt.
(Exam 1 2021, Question 13 - similar question)

(21) **[3]** Why is it difficult/impossible to create a benchmark that will work across all classes of parallel processors?

It's difficult/impossible to create such a benchmark because the shared memory machines and message passing machines require different underlying program models.

(22) **[3]** Over the years, clock rates grew by a factor of 1000 while power consumed only increased by a factor of 30. How was this accomplished without melting the chip?

P = CfV^2. We were able to lower the power by lowering the voltage squared term (lowering voltage by a little lead to significantly lowered power.)

(23) **[3]** Which is harder to write a program for, a shared memory machine or a message passing machine? Why?

Message passing machine because the programmer has to handle the explicit communication.

It is easier to write a program for a shared memory machine.

(24) **[3]** Which is more expensive to build - a shared memory machine, or a message passing machine? Why?

Shared memory machine because it requires a lot of hardware (BUS, additional caches) that ensures coherence and consistency. The tradeoff for it being expensive is that it's easy to code.

(25) **[3]** What is "leakage" current? If Vdd is lowered, what happens to the amount of leakage current, and why?

Leakage current is the current that flows through a machine in its OFF state. If Vdd is lowered, it forces the threshold voltage to go down which causes the amount of leakage current to increase.

(26) **[3]** Why is it so difficult for the processing elements on a CMOS-based chip to communicate with things that are located off the chip?

CMOS is charge-based (moves a small number of electrons short distance.) CMOS must be able to move a large amount of electrons (but it can't) in order to communicate with things off chip because they must be able to detect the charge.

(27) **[4]** Speculation is a very useful technique for improving performance. However, it is not being used as extensively as it once was - why not?

If we speculated it wrong, we are wasting energy in power and we are concerned about power consumption these days. We should not increase power consumption by wasting unnecessary energy

(28) **[3]** What is the definition of a basic block? Why is there a desire to create a bigger one?

Basic block has one entry point and one exit point. Bigger basic blocks allow for code motion (you can replace NOP instructions with useful instructions.)

(29) **[2]** What is the definition of a precise interrupt?

All of the instructions before the instruction with a problem have completed and none of the ones after have.

(30) **[2]** Why is it important to support precise interrupts in modern pipelined processors?

Virtual Memory Support (Sai confirmed with the professor)

(31) **[6]** 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.

(32) **[4]** Processors have been built that were able to issue 8 instructions at a time using a fast clock. However, these processors are no longer being built - why not? Why would you choose a 3-issue machine over an 8-issue machine, if the clock rates were the same?

There is a lot of extra hardware to support 8 issue, and some of that hardware does not scale linearly. Since the average number of instructions that can be started at the same time is approximately 3, there is a lot of wasted energy in an 8-issue machine.
(Exam 2 2021, Question 13. Professor's Answer: > 20 words)

(33) **[4]** Vector machines are an example of a SIMD style of parallel processing. They feature instructions that look like VR0 = VR1 + VR2. Explain briefly why these machines are able to fetch and decode many fewer instructions than a traditional processor does. Use pictures if that will help get your point across.

Vector machines operate on multiple data elements with a single instruction (though not at the same time) through vector data. For example, the above instruction would add the first element of VR1 to the first element of VR2 and store that as the first element of VR0, and so on. (Annie's Answer)

(34) **[6]** The designer has the choice of using a physically addressed cache or a virtually addressed cache. Explain the difference (drawing a picture is fine!), and give 1 advantage for each.

Virtually addressed cache: **advantage** - no translation, faster
disadvantage - aliasing problem
Physically addressed cache: **advantage** - no aliasing problem
disadvantage - translation, slower

(35) **[4]** An important program spends 30% of its time doing memory operations (loads and stores). By redesigning the memory hierarchy you can make the memory operations 80% faster (take 20% as long), or you can redesign the hardware to make the rest of the machine 30% faster (take 70% as long). Which should you do and why? (You must show your work to get full credit.)

(36) **[6]** You are responsible for designing a new embedded processor, and for a variety of reasons you must use a fixed 23 bit instruction size and you must support at least 32 different opcodes. You would like to use a 3-operand instruction format, and have 128 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.

(37) **[4]** Find the Average Memory Access Time (AMAT) for a processor with a 1 ns clock cycle time, a miss penalty of 20 clock cycles, a miss rate of 0.10 misses per instruction, and a cache access time (including hit detection) of 1 clock cycle. Assume that the read and write miss penalties are the same and ignore other write stalls.

AMAT = 1 + 20 * .10
AMAT = 1 + 2
AMAT = 3