CLA CS-550 (A1)
Advanced Computer Architecture

Exam Preview

Total Time: 120 minutes | **Closed Book** | **Total Points: 100/100**

Answer All Questions

State your asumptions clearly

Your grade will depend on the clarity and soundness of your assumptions

1 Give short answers for the following questions:

- Define what is meant by *Die Yield*. How is the die yield related to the number of dies per wafer?
- A floating point unit is used 32 percent of the time. What would be the overall performance gain if the performance of that unit is improved by 100 percent.
- What is meant by ISA *regularity*, *orthogonality*, and *composability*? Why are these important? Exemplify.
- The VAX-11 has a large number of addressing modes, which include *auto-decrement deferred indexed* and *displacement indexed*. Explain how to compute the effective address for both of these modes. For each of these two modes, give an example for a typical use.
- Which machine type is likely to have a higher percentage of *branching*, a *load-store* machine or a *memory-memory* machine ? Justify.
- What type of hazards do *bypassing*, *scoreboarding*, and *Tomasulo* treat? How? Which types of Hazards do these techniques avoid? How?
- How many reservation stations would you need to implement the Tomasulo Algorithm for a processor, knowing that the ALU is fully pipelined and consists of 4 stages.
- *Strip-mining* and *software pipelining* are two optimization techniques used by compilers to produce high-performance compiled code. Explain briefly each of these techniques. Use an example to illustrate your explanation. When does the need for these optimizations arise?

2 Define the topology of an n-dimensional Cube Connected Cycle. What is the total number of nodes in a Cube Connected Cycle of dimension n ? What is the diameter? How many direct neighbor connections does each node in an n-dimensional Cube Connected Cycle have? Suggest an algorithm for routing in a Cube Connected Cycle.

3 The cost/performance of three DLX processors is to be examined.

- Model A: The superscalar DLX issues up to one integer and one floating point instruction each clock cycle.
- Model T: Similar to model A. In addition to issuing one integer and one floating point instruction each clock cycle, it has the necessary hardware for the Tomasulo Algorithm.
- Model V: The vector version of DLX, called DLXV, which supports chaining.

All three models have an instruction cache containing 4096 bytes using 16 byte blocks that uses direct mapping. The models A and T also have a 4096-byte data cache with 128 byte blocks that is also direct mapped. This is a write through cache.

All models have the same hardware building blocks:
- A fully pipelined adder whose latency is 3 clock cycles.
- A fully pipelined multiplier whose latency is 5 clock cycles.
- The memory contains 16 banks, with each bank being 64-bits wide.
- The latency for a memory access is 10 clock cycles. The memory system delivers one 64-bit quantity per bank after the initial memory latency.

The following is a kernel to be used to compare the three models.

```
C***     Kernel 3        Inner Product
C
           Q = 0.0
           DO 3 K = 1,1024
        3  Q = Q + Z(K) * X(K)
```

Assume that Q, X, and Z are double precision floating point numbers, with the dimensions of X and Z being 1024. Q is in Fq (one of the general purpose Floating Point registers), K is in Rk (one of the general purpose Integer registers), and the base address of array Z is in Rz, and the base address of array X is in Rx.

a. Write an efficient DLX version of this program for Model A.

b. What changes (if any) would you make to run this program on model T. Why?

c. Write an efficient DLXV version of this program for Model V.

d. What is the total number of *instruction* misses for each version of the program? Give the number of clock cycles for instruction cache misses. Explain your calculations.

e. What is the total number of *data* misses for each version of the program? Give the number of clock cycles for instruction cache misses. Explain your calculations.

f. What is the total number of clock cycles it takes to execute the programs for each model of the DLX. Explain your calculations.

g. When laid out for fabrication using a CMOS process, the three DLX Models (A, T, and V) had relative silicon areas of 81 : 85 : 114. What would be the relative cost for these three models?

h. Using the result from parts (f and g) above, which one of the three Models achieves the best cost performance. Comment.

4 The Fujitsu SPARClite 86930 processor operates at a 40 MHz clock rate. The performance of this machine on a particular benchmark under a 100% data cache hit ratio was metered at a CPI of 1.08. When the benchmark was run with a smaller cache, the performance was metered at 36 MIPS. The benchmark had an instruction mix where 25% of all instructions were Load/Store (memory access) instructions. Assuming that the cache access time is 20 nanoseconds (latency for a hit) and that the DRAM access time is 82 nanoseconds (latency for a miss). Compute the cache hit ratio.

5 In 1983, James Goodman has proposed a cache coherence protocol for bus-based multiprocessors. This scheme combines the advantages of both write-through and write-back invalidations. In order to reduce bus traffic, the very first write of a cache block uses a write-through policy. After this first write, shared memory is updated using a write-back policy. This scheme can be described by assigning cache lines (blocks) to one of 4 possible states.

- ◇ **Valid**: The cache block, which is consistent with the main memory copy, has been read from shared memory and has not been modified.
- ◇ **Invalid**: The block is not found in the cache or is inconsistent with the memory copy.
- ◇ **Reserved**: Data has been written exactly once since being read from shared memory. The cache copy is consistent with the memory copy, which is the only other copy.
- ◇ **Dirty**: The cache block has been modified (written) more than once, and the cache copy is the only one in the system (thus inconsistent with all other copies).

Draw a state diagram showing all the possible transitions between the states defined above, as a result of *local processor* transactions or *remote processor* (bus) transactions. Give a clear definition of any transaction/action you use in your state diagram/table. Explain why this protocol is likely to reduce bus traffic.

6

a. Suppose the branch frequencies (as a percentage of all instructions) are as follows: Conditional branches (20%); Jumps and calls (5%). Also, assume that 60% of all conditional branches are taken. Consider a 4-stage pipeline, where the branch is resolved at the end of the second cycle for unconditional branches, and at the end of the third cycle for conditional branches. Assuming that only the first pipe stage is always done independent of whether or not the branch is taken and ignoring all other pipeline stalls, how much faster would the machine be without any branch hazards?

b. Suppose we have a deeply pipelined machine, for which we implement a branch-target buffer for the conditional branches only. Assume that the misprediction penalty is always 4 clock cycles and that the buffer miss penalty is always 3 clock cycles. Assume 90% hit rate and 90% prediction accuracy, also assume that the branch statistics are as in Figure 6.18 of the textbook. How much faster is the machine with the branch-target buffer versus a machine that has a fixed 2-cycle branch penalty? Assume that the CPI to be 1 (withouth any branch stalls).

c. How many switches are there in a mesh of trees (MOT), with k dimensions and n nodes along each dimension? What is the diameter of this topology? (note: Check your answer on a 3-D MOT). Write the pseudo code for routing through this topology. Assume that all arcs are bi-directional.

Good Luck!