Example examination for Computer Architecture

Date: Nov 12th, 2012

Instructions

You must answer question 1, plus any 1 of the remaining 2 questions. You may not bring books or lecture notes into the examination and you have three hours to complete the examination.

The examination constitutes 30% of the final evaluation (the lab assignments constitute 50%, and the homework 20%). Each question of the examination carries an equal weight (15%) in the final evaluation. Question parts are labeled with individual percentage marks in square brackets (eg. [x%]), expressed over the total.

Question 1 (15%)

a. A pipeline implements concurrency in an operation by dividing the operation into a number of dependent stages and performing each operation concurrently over a sequence of operations. Explain in your own words what are the advantages of a pipeline over a non-pipelined unit that executes each operation in turn. [3%]

b. Describe how using multiple cores on chip can increase the ratio of performance over energy budget. [3%]

c. Explain in your own words, if possible using examples, the difference between each of the following pairs of words [3%]:

   i. bandwidth vs. throughput

   ii. replacement policy vs. write policy (for caches)

   iii. static vs. dynamic RAM (SRAM vs. DRAM)

   iv. hardware vs. software multithreading

d. A processor makes 234 loads and 115 stores to a write-through cache. The read hit rate is 10%, the write hit rate is 60%. Determine the number of requests sent by the cache to the next memory level. [3%]

e. Describe what is meant by the term “the memory wall”. [3%]

Question 2 (15%)

a. Define in your own words the following dependencies that may occur in the concurrent execution of operations. For each, give a fragment of assembly code or pseudo assembly code to illustrate the dependency. Describe how you would detect such dependencies when issuing instructions with operands taken from a register file.

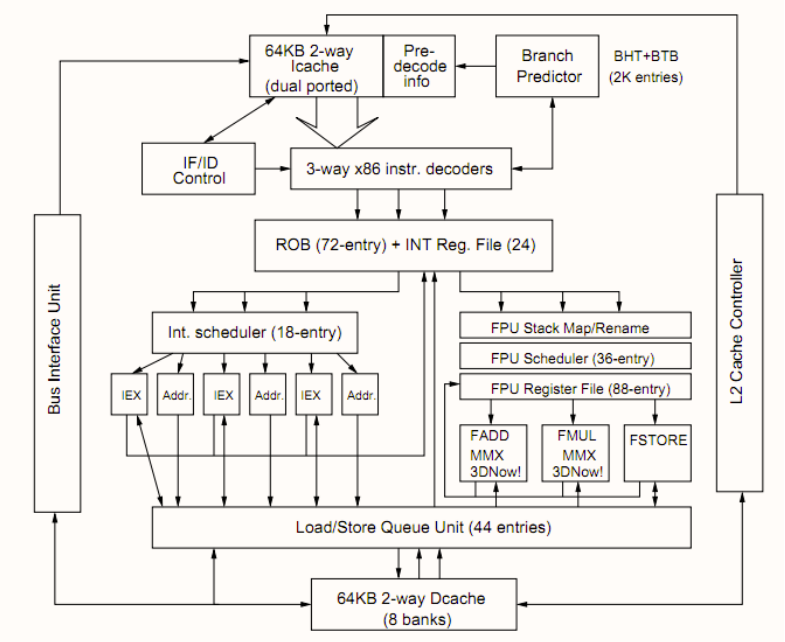
   i. Data dependency [2%]

   ii. Output dependency [2%]

iii. Anti dependency [2%]

(Assembly format: op rega <- regb, regc; rega is the target register.)

b. The AMD Athlon pipeline is described in the image below. Explain the purpose and function of the branch predictor. [3%]



c. The AMD Athlon processor described above uses uncached memory-mapped I/O. Explain what this means, then indicate where in the processor diagram the data and control lines to the I/O subsystem are connected. [2%].

d. The Intel processors use a CISC instruction set for historical reasons, yet all modern Intel-compatible processors use a RISC pipeline. Explain the impact of keeping CISC compatibility on performance, and what architectural features can be introduced to reduce this impact. Mention two different processor architectures as examples. [4%]

Question 3 (15%)

a. Explain the terms “CAS”, “RAS” and “row buffer” in the context of D-RAM memory. [3%]

b. A given D-RAM controller can answer a single load request in x second if the row for the address is already loaded in the row buffer; it needs to wait y seconds between subsequent requests, and it also needs z seconds to change rows. Express the minimum read latency, maximum read latency, and maximum read bandwidth as functions of x, y and z. [3%]

c. An embedded application uses a dedicated processor to run a single image processing algorithm. This algorithm reads image data linearly from

one area in memory, makes a simple transformation then writes data linearly into a different area in memory. Explain which policies you would recommend for the L1 cache of this processor and why. [3%]

d. Explain the difference between cache coherency and memory consistency in your own words, with examples. [3%]

e. A 2-way set associative cache with LRU policy can be much simpler and cheaper to implement than a cache with higher associativity. Why? [3%]