Examination for Grid Hardware Infrastructure

21st December 2005, 12am to 3pm

Instructions

You are required to answer question 1 plus any 2 of the remaining 3 questions. You may not bring books or lecture notes into the examination and you have three hours to complete the examination. Each question carries an equal weight in the assessment. Questions parts are labeled with percentage marks for that part in square brackets e.g. [x%]. The percentage is of the final assessment, i.e. a total of 60% for the examination or 20% for each question answered.

Question 1.

(a) Memory is a very important component of computer systems. It has two characteristic measures, bandwidth and latency. How are these defined and in what units are they expressed. Describe two ways in which concurrency in code can be exploited in memory design to increase bandwidth. [4%].

(b) 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. Draw a timing diagram to illustrate this using an SDRAM as an example, which should show how the sequence of memory accesses performed over time. Assume the memory access is divided 4 stages, which are synchronously clocked, namely {address latch; decode; read/write; output}. [4%]

(c) Define the parameters $r_{infinity}$ and $n_{1/2}$ that define the performance of a generic pipeline in the absence of any bubbles. Derive a formula for the performance of a pipeline defined as the number of operations performed per second for a pipeline with L stages and a cycle time of t seconds. Using this formula define $r_{infinity}$ and $n_{1/2}$ in terms of these the parameters L and t. [4%]

(d) A shared memory is constructed with the SDRAM described in part (b) for a bus-based multi-computer. The memory and bus can operate at 400Mhz and provide 256 bits of data (one cache line) in each request. The bus and memory system can support one new memory request in each bus cycle and the bus requires two cycles to present the address to the memory subsystem and two cycles to present the data read back to the processor requesting it. Calculate the peak bandwidth of this memory system and the average bandwidth seen by one processor accessing arrays of 5, 50 and 500 elements. Roughly plot the performance of the memory system against number of consecutive accesses. You may assume that no other processors are accessing memory and that the processor is much faster than the bus. [8%]

Question 2.

(a) Concurrent instruction issue can be used to increase processor throughput and to tolerate latency in memory accesses. Identify two generic types of processor architecture that use static (or compile-time) and dynamic (or run-time) scheduling of instructions to realize one or both of these goals, briefly describe how each processes instructions in order to achieve these goals. [4%]

(b) and (c) Discuss each architecture in detail paying particular attention to all of the mechanisms involved in instruction issue and paying particular attention to any problems they each face in achieving wide concurrent instruction issue. [6% + 6%]

(d) Assuming that a statically-scheduled processor has 2 Integer ALUs with latency of 1 cycle, a floating point ALU with latency of 2 cycles, a branch unit of latency 1 cycle and a load/store unit of latency 2 cycles, attempt to create a schedule of assembler instructions to implement the loop:

```
For I=1,n
        A[i] = B[i] + c[i+k]
```

Where A, B and C are floats and i and k are integers. The latencies are defined from the time data is available from the register file to producing a result. You may assume operations with latencies of 2 cycles can accept new operations every cycle. You need not perform a register allocation and so can use symbolic variables instead of register specifiers in your assembly instructions. [4%]

Question 3.

Describe the following mapping strategies in relation to cache memory. Make use of diagrams in your answers where appropriate.

(i) direct mapped [2%]
(ii) 4-way set associative [2%]
(iii) fully associative [2%]

A Level-1 Instruction cache of 64Kbytes with a cache line of 256 bits is addressed by a 30-bit address to access a 4-byte instruction word. Given these parameters, then:

(a) For each mapping strategy, identify how many different cache lines an arbitrary address in memory can be mapped to and how many tag matches are involved in the access of that address? [2%]

(b) For each of the mapping strategies, divide the 30-bit address into fields and describe how each field is used in accessing an instruction in the cache. Use diagrams of the components that make up the cache in your answer. [4% for each strategy]

Question 4.

Explain what a data hazard is in an in-order, pipelined datapath. Give an example of assembler code that contains a data hazard. Assuming that a datapath has the following stages: {Instruction fetch, register read, execute and writeback}, draw a diagram showing the various stages of the execution of your code in a pipeline, showing where data is required and how bubbles are formed in the pipeline's execution because of the data hazard in your code. [6%]

Data forwarding is a means of reducing the impact of data hazards. Draw a diagram of a pipelined data path showing the data forwarding busses and how these provide alternative input to the ALU stage. Discuss what information is required to control the selection of the appropriate data into the ALU. [8%]

A pipelined datapath has 3 pipelined execution units with different operation latencies and uses in-order issue and out-of-order completion of instructions. Two instructions can be issued to the 3 execution units in each cycle. Draw a diagram of the datapath showing all possible inputs to each execution unit, assuming that bypassing is implemented on the result of each functional unit. [4%]

What dependencies are introduced in implementing out-of-order completion of instructions and how may these be avoided. [2%]