Examination for Advances in Computer Architecture

20th December 2007 09:00-12:00

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) Define what is meant by the term “the memory wall”. Using your understanding of the scaling in CMOS technology, give an explanation of why this occurs. Finally describe what techniques can be used in computer architecture to minimise the impact of this problem. [4%].

(b) Under low-load conditions in a network, virtual cut-through flow control pipelines the delivery of a message over a number of nodes in a network. Assuming a message comprises up to ten 8-bit flits, draw a diagram, with time on one axis and the sequence of nodes visited on the other axis, to show the concurrency in a delivering a message over the network under such load conditions. What is the maximum reduction in latency that can be obtained by using this flow control compared to store and forward routing. Under what conditions does this maximum occur (your answer should be dependent on the number of hops a message propagates – d)? [4%]

(c) A pipeline of L stages and cycle time of t seconds can be characterised by the parameters $r_{infinity}$ and $n_{1/2}$. Define both parameters both informally and mathematically, as a function of L and t. Now derive a formula for the time required to compute n operations in a pipeline and hence a formula for the computational rate of a pipeline as a function of n. Compute and tabulate the computational rate of this pipeline for a number of integer multiples of $n_{1/2}$ and hence sketch a graph of this function using units of $n_{1/2}$ and 1/t. [4%]

(d) Define the following metrics used to characterise the performance of a computer system or its components. What are the units of these metrics? [4%]

(i) latency
(ii) bandwidth
(iii) performance

(e) In a memory system what techniques can be used to increase its bandwidth. Explain why the latency of the memory component cannot be reduced for a given technology. [4%]

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 (compile-time) and dynamic (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 class of 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 scaling instruction issue width. [6% + 6%]

(d) Discuss other alternatives to increase processor throughput and to tolerate memory latency [4%]

Question 3.

(a) Define the following dependencies that occur in the execution of operations from operands taken from a register file. For each, give a fragment of assembly code or pseudo assembly code to illustrate the dependency.

(i) Data dependency [2%]
(ii) Output dependency [2%]
(iii) Anti dependency [2%]

(b) For wide-issue or superscalar architectures, various instructions scheduling strategies can be implemented. Describe all possible strategies, indicating which of the above dependencies must be taken into account in each scheduling strategy [9]

(c) Techniques can be used for eliminating two of the three dependencies from the execution of your assembly code. Describe how this may be implemented in

a superscalar architecture. Is this technique applicable to VLIW architectures? [5]

Question 4.

Compare and contrast the IBM cell and the Niagara II multi-core processors. In your answer you should include discussion and as many implementation parameters as possible on the following topics:

a) The memory model and memory architecture [4%]
b) The floating point computation architecture [4%]
c) The programming model and its implementation [4%]
d) How the processors tolerate latency on long latency operations [4%]
e) The switching network architecture [4%]