UNIVERSITY OF EDINBURGH

FACULTY OF SCIENCE AND ENGINEERING

MASTER OF SCIENCE IN COMPUTER SCIENCE

CONCURRENT COMPUTER ARCHITECTURE

Monday 24 April 2000

9:30 to 22:00

Examiners:

I. Watson (External)

R.N. Ibbett (Chairman)

INSTRUCTIONS TO CANDIDATES

Answer any TWO questions.

All questions carry equal weight.

1. (a) With the aid of diagrams, briefly explain the organisation and operation of each of the following types of cache:

    i. fully associative
    ii. set associative
    iii. direct mapped

    [*9 marks*]

   (b) A number of processors, each with its own direct-mapped, write-through cache, are connected together by means of a bus. Explain how the caches can be kept consistent when several of the processors share common variables. What is the difference between an *update* and an *invalidate* protocol? [*7 marks*]

   (c) In an attempt to improve performance, the system is modified so that the caches operate in copy-back, write allocate mode (*i.e.* main memory is updated only when a modified line has to be re-allocated and a line in cache is allocated whenever there is a read or a write miss). With the aid of state transition diagrams, explain how the caches are now kept consistent using a write-invalidate protocol. [*9 marks*]

2. (a) Explain why the maximum potential execution rate of the CDC 7600 could not be achieved in practice and describe how the implementation of vector instructions in the Cray-1 overcame the limitations of the 7600 design. Explain how *chaining* works and how it helps to maximise the utilisation of the floating-point capacity of the Cray-1. [*11 marks*]

   (b) A certain vector processor has a clock period τ and the time taken to start up a vector operation is $S\tau$. Derive an expression for the average time taken to produce an element of the result vector formed by the addition of two n-element vectors ($n \leq$ vector register length) in an arithmetic pipeline containing L stages. [*4 marks*]

   (c) i. Explain what is meant by *half performance length* and determine its value if S=4 and L=6.
   If two successive vector operations are chained and half of the start-up time of the second instruction is masked by the chaining, what is the half performance length of the combined operation if the second instruction is executed in a 5-stage pipeline?

   ii. In the same processor a (scalar) program loop used to execute the same composite function as in (c) without the use of vector instructions takes 5 clock periods to execute one iteration of the loop. What is the *breakeven length* i.e. the vector length at which the time to produce a result element is the same using scalar or vector instruction sequences?

[*10 marks*]

3. (a) Write short notes explaining what is meant by each of the following terms in the context of a computer interconnection network.

   i. Network topology
   ii. Network diameter
   iii. Latency
   iv. Bisection Width

   [*8 marks*]

   (b) Explain what a *routing function* is and give a mathematical definition of each of the following functions.

   i. perfect shuffle
   ii. butterfly
   iii. exchange

   [*6 marks*]

   (c) i. Eight processors are connected together by a unidirectional ring network. In any one time step a single data item can be moved from each of the processors to its immediate neighbour. A set of 8 data items, one per processor, is to be permuted using an *exchange* routing function. For each of the three possible exchanges, draw a diagram which shows the pattern of data movements between the sources and the destinations.

   ii. Determine the minimum number of time steps required in each case to move the data items round the ring to their final destination.

   iii. If the network is upgraded to become a bidirectional ring in which each processor can send a data item to each of its neighbours in any one time step, what are the new times required to move the data items? [*11 marks*]