Master Degree Program in Computer Science and Networking

High Performance Computing

2014-15

Homework 5

All the answers must be properly and clearly explained.

1) Consider the precedence relation computation of page 321, executed on a *cache coherent* architecture. If *STORE b* is asynchronous, explain what is the impact of cache coherence support on the computation semantics and, in particular, which value is returned by $LOAD_2$ *b*.

2) The run-time version of a parallel computation for an exclusive-mapping, cache-coherent architecture (directory-based, invalidation-based, inclusive 2-level caching) contains a pair of processes with a critical section implemented by locking with explicit notify.

Study the behavior of the process pair in terms of cache coherence events and operations, distinguishing between *home* PE and *non-home* PE.

Evaluate the synchronization latency for a single-CMP system with 4 MINFs. The internal interconnect is a 4-ary 2-cube network; 4 switching units are used for connecting the MINFs. Processes are mapped randomly onto PEs.

3) Consider the stream-based module Q defined as in Homework 2:

```
Q:: int A[M], B[M], C[M]; channel in input_stream (4); channel out output_stream;
        while (true) do
                { receive (input_stream, (A, B));
                  ∀ i = 0 .. M – 1:
                        { C[i] = 0;
                          ∀ j = 0 .. M – 1:
                                C[i] = F (C[i], A[i], B[j])
                        };
                  send (output_stream, C)
                }
```

(M = 4K, function F available as a library with service time 100τ), executed on the architecture of Question 2) above with exclusive mapping.

The input stream is generated by a process P, and the output stream is used up by a process R. Q is the bottleneck.

Assume that the *send* implementation copies the messages into the local cache of the sender PE. Don't consider the synchronization issues of the *send* and *receive* run-time support (assume they are implemented correctly, with low T_{setup}).

Evaluate the ideal service time of Q, and the parameters p, T_p, R_{Q0} and T_s to be used for the evaluation of under-load latency R_Q.