Master Degree Program in Computer Science and Networking

High Performance Computing

2014-15

Homework 8

All the answers must be properly and clearly explained.

Exercise (*Complete analysis of a parallel program on a multiprocessor*):

A stream equivalent computation is composed of three processes *P1*, *P2* and *P3* interconnected in a pipeline fashion. *P1* produces an input stream of arrays *B[M]* (M=1K) that are rows of a matrix *V[R][M]*. *P2* encapsulates an interger array *A[M]*. For each received *B* the output stream element is a new value of *B* calculated as follows:

```
for i = 0 .. M - 1 : B[i] = F (A[i], B[i])
```

The sequential function *F* has calculation time $T_F = 100\tau$. *P3* receives arrays *B* and fills an output matrix *Z[R][M]*. The program is executed in a single-chip CMP with the following features:

- 16 PEs and 4 MINFs. The internal interconnect is a crossbar. Each MINF is *logically* connected to a group of 4 PEs, and it is physically connected to an external memory subsystem composed of 4 mutually interleaved macro-modules. A group of 4 PEs, with the associated MINF and the *local* memory subsystem, has a SMP organization. Globally the four groups have a NUMA organization;
- CPUs have D-RISC pipelined scalar architecture, with private primary cache (32K + 32K) and secondary cache (1 M). The service time per instruction is τ.
- cache coherence is automatic, directory-based, invalidation-based, with home flushing *(up to L2 cache)*. Home nodes are chosen statically;
- processes are mapped onto PEs in an exclusive fashion (*exclusive mapping*). Run-time support is according to the Rdy-Ack solution based on I/O communications.

For this program give the analysis of the original system in term of *base latency* assuming the homing of each symmetric channel in the PE of the receiver process. Evaluate the ideal service times and the effective ones of *P1*, *P2* and *P3* and the optimal parallelism degree of the bottleneck. Evaluate the efficiency of the system and of the three processes and the completion time. For the *under-load analysis* indicate the values of $\boldsymbol{T_p, p, T_S, R_{q-0}}$ and suppose that $\boldsymbol{R_q / R_{q-0} \sim 1}$.

Provide a data parallel parallelization of the bottleneck and discuss the performance of the system with the *base latency*. For the *under-load analysis* indicate the values of $\mathbf{T_p, p, T_S, R_{q-0}}$ and suppose that $\mathbf{R_q / R_{q-0} = 1}$.5. Determine the effective service time of the parallel implementation and its completion time. Discuss, from the qualitative point of view, the behavior of the parallel implementation according to the basic invalidation semantics.