Homework 4

All the answers must be properly and clearly explained.

Question 1

The general problem is: for a given multiprocessor architecture Arch and for a given computation Comp to be parallelized, study the parallelization of Comp for Arch. In particular, determine the parallel paradigm(s) and the cost model of the parallel program (parallelism degree, mapping, ideal and effective service time, relative efficiency) by properly exploiting the cost model of the architecture and of the sequential computation.

Assume the following characteristics:

- every PE contains a D-RISC pipelined scalar CPU with on-demand 32K primary instruction cache and data cache with 8-word blocks, and 512K on chip secondary cache with 128-word blocks; two variants to be studied: secondary cache works on-demand or with prefetching of the next block;
- each macro-module is composed of 8 interleaved modules with clock cycle \( \tau_M = 20 \tau \);
- network flow control is wormhole with 1-word links and flits, and 1\( \tau \) link transmission latency;
- single buffering applied to any inter-chip and intra-chip firmware communication;
- overlappable communications; for interprocess communication latency parameters we consider two variants : \( 10^3 \tau \) or \( 10^2 \tau \) for \( T_{\text{setup}} \), and \( 10 \tau \) or \( 10^2 \tau \) for \( T_{\text{trans}} \) (to be chosen according to the specific case study);
- exclusive mapping.

This general problem can to instantiated in many case studies, which can also be defined by the student (see the note at the end of Question 2, Homework 2). Some proposed case studies can be defined by combining the following architectural and computational characteristics:

1) architecture (number of PEs \( N \) to be chosen):
   a. SMP with \( n_M \) macromodules (\( n_M \) to be chosen) and generalized binary fat tree network;
   b. NUMA with binary fat tree network;
   c. NUMA with \( k \)-ary 2-cube network.

2) module Q operating on streams (interarrival time and \( M \) to be chosen):
   a. Q encapsulates an integer variable \( C \) and receives \((A[M], d)\) with \( A \) integer array and \( d \) boolean. If \( d \) is true then \( C \) is incremented, else the value equal to the number of occurrences of \( C \) in \( A \) is sent onto the output stream;
   b. Q encapsulates an integer array \( A[M][M] \), receives an integer array \( B[M] \), and sends an integer array \( C[M] \), where \( C \) is the result of the matrix-vector product of \( A \) and \( B \).

When feasible, solve the problem in a general, parametric way and, after that, assign values to the unspecified parameters (e.g. communication latency, interarrival time, \( M \)) in order to recognize different cases (e.g. primary or secondary cache reuse exploitation).
Question 2
Consider the zero-copy run-time support of LC interprocess communication.

a) Recognize all the possible instances of shared pointers, and describe their implementation.

b) Define the type of object returned by the receive library.

c) Assuming that the receive library as defined in b) is available to you, compile the following process fragment:

```
... channel in ch (3); int x; ...

...;
receive (ch, x);
x = x + 1;
...
```

d) Compile the code for copying the message value into the target variable and evaluate its latency.

Question 3
With reference to a client-server computation with strict request-reply behavior, explain whether the following statements are true, or false, or true under certain conditions:

a) the effective bandwidth of the whole computation is equal to the interarrival rate of the server;

b) the effective bandwidth of the whole computation is equal to the interdeparture rate of the server;

c) with just one client and requests of integer type, a farm parallelization of the server is not meaningful;

d) with just one client and requests of array type, a farm parallelization of the server might be meaningful.

Question 4
With reference to the pipelined CPU architecture:

a) Prove that the calculation time and the ideal service time of every unit of the physical architecture are equal to one clock cycle and to $L_{\text{com}}$ respectively.

b) Is it possible that the Execution Unit (EU) becomes a bottleneck?

c) Is it possible that the EU functional units are load unbalanced? If so, is this a reason for the CPU performance degradation?

d) How the Bernstein conditions are exploited by the compiler?

Note: exercises of program compilation and evaluation for a pipelined CPU are part of exercises on parallel programs cost model, e.g. as in Question 1.

Question 5

a) Evaluate the ideal service time and the latency of the Processing Element Interface Unit (W) for a SMP and for a NUMA multiprocessor architecture.

b) Evaluate the ideal service time and the latency of the Switch Unit SW of the various classes of interconnection networks, assuming wormhole flow control.

In both a) and b) the student must be able to give the evaluation in a convincing way, not necessarily by writing the detailed/complete microprogram.