INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI

Department of Computer Science and Engineering

CS523 (**Advanced Computer Architecture**): End Semester Examination

Date: 20th Nov 2012 **Time: 1.00PM-4.00PM** **Answer all questions**

1. [20] **Interconnection Network**

   a. What is graph embedding and how it helps in designing application in multi-core environments? What are the factors to be considered at the time of embedding? **[2.5+2.5]**

   b. Calculate bisection bandwidth and diameter of array, mesh, tree and hypercube connected multiprocessors. Which network is better in cost and performance tradeoff? **[5]**

   c. Suppose 100 packets of data need to send from node A to C via B with each packet header size is 64bits and payload is 2048bits. Calculate the overall latency of sending all 100 packets when link bandwidth is 1kbits/s for both wormhole routing scheme and store/forward routing scheme. **[5]**

   d. Explain how ***WEST-FIRST*** routing (WF-Partial Adoptive Routing) ensures no deadlock situation happened in MESH network. **[5]**

2. [28] **Vector Architecture and GPU**

   a. Given the following code, which multiplies two vector that contain single precision complex values

```
for(i=0;i<300;i++){
     c_re[i]= a_re[i]*b_re[i]  -  a_im[i]*b_im[i];
     c_im[i]= a_re[i]*b_im[i] +  a_im[i]*b_re[i];
}
```

   Calculate the arithmetic intensity for above code. Suppose vector processor with vector length is 64, vector register load/store time is 15 cycles, number of pipeline multiplier is 4 and with pipe-depth 8 cycles, number of pipeline Adder/Substractor is one with pipe-depth 5 cycles. Calculate total number of cycle required to complete the operation on the assumed vector processor. **[3+8]**

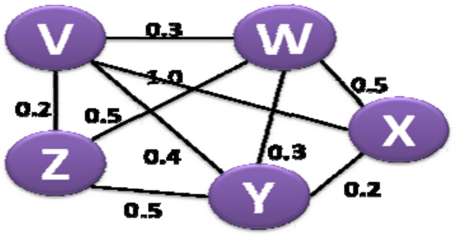
   b. What is an affine access? Does the following loop have loop-carried dependency (identify using GCD test)? **[3+4]**

```
for(i=0;i<100;i++) A[3*i]=B[2*i+4];
```

   c. Describe scatter and gather operation in vector processor or GPU. **[4]**

   d. Given a **GPU** having 8 streaming multiprocessor with each having 16 stream processor, find out the value of ThreadBlock size (nt) and number of Threadblock (nb) for the code given kernel code to execute on the **GPU** with less synchronization overhead and better balanced parallelization. (N=2048, assume any other required assumptions/values). [6]

```
DAXPY<<<nb,nt>>>(2048, 2.0,x,y);
void DAXPY(int n, double a, double *x, double*y){
     int  i;
     i=blockIdx.x*blockDim.x +threadIdx.x;
     If (i<n)   y[i]=a*x[i]+y[i];
}
```



3. [18] **Data Placement Model and Access Classifications** *(L:logical, P:Physical/Partitioned, S:Shared, D:Distributed)*

   a. Explain benefits of LSPD memory model in multicore architecture? **[3]**

   b. Explain "Rotational Interleaving" in placing shared data in "LSPD Cache with copying memory model" in multicore architecture? **[4]**

   c. Compare LSPS and LPPS model in terms of benefit and demerit. **[4]**

   d. Assuming a LSPS(2) memory model architecture and suppose an application "App" is running on this architecture with 5 Array namely V[100], W[100], X[200], Y[200] and Z[100]. Given that size of memories are 400 (M1= M2 = 400) and conflict graph (***Fig 1, above***), place the data to memories for efficient execution of application. **[7]**

4. [22] **Memory Coherence, Consistency and Lock/Barrier**

- **a.** Explain “***Load Locked/Linked***” and “***Store Conditional***” instructions and how these two instructions can be used as universal locking instructions, i.e. can be used to implement (a) test & set, (b) atomic exchange and (c) fetch and increment. **[3+2+2+2]**
- b. Explain spin lock and how performance of spin lock can be improved by exponential back up? [3]
- **c.** Explain difference between Snoopy and Directory based coherence protocol. How cache coherence help to improve locking performance in multiprocessor environment? **[3+3]**
- d. Assume two processes P1 and P2 given bellow to be run concurrently on two OOO processors. Which memory model (among SC, TSO, PSO, PC) ***ensure that if B is printed as “1” then A is also printed as “1”.*** **[4]**

P1	P2
A = 1;	print B;
B = 1;	print A;

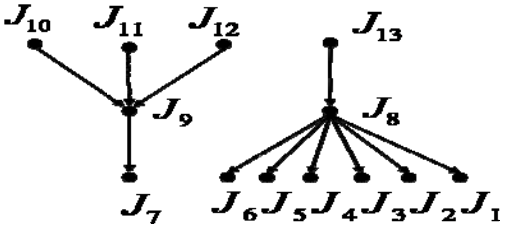
SC: Sequential Consitency
TSO : Total Store Order, Writes are atomic
PC : Processor Consistency, Writes need not be atomic - Invalidations may gradually propagate
PSO: Partial Store Order, Write can be ordered

5. [22] Multicore Cache and Bandwidth partitioning

- **a.** Explain difference between way based and set based LPPS model cache partitioning. **[4]**
- **b.** What is aggressive block acquire policy and graceful block acquire policy in dynamic cache partitioning? **[3]**
- **c.** Explain how off-chip memory Bandwidth can be partition using Token based algorithms. **[3]**
- d. What is harmful prefetch? What are the measure can be taken to reduce the effect of prefetch in at cache level? **[2+3]**
- e. Explain Network Fair Queuing model of Bandwidth sharing. Explain an efficient way to cater (memory scheduling service) pre-fetch requests and demand requests in multicore processors. **[3+4]**

6. [28] Multiprocessor Scheduling

- **a.** Schedule the following graph (node with unit processing cost and dependency) on 3 processor to generate Gantt chart with (a) *list scheduling*, (b) *critical path scheduling* and (c) *most-successor first scheduling* policies. Describe the problem in α| β| γ format. **[3+3+3+2]**



- b. What is difference between Centralized and Distributed scheduling? Work stealing corresponds to what kind of distributed scheduling. If we choose the victim processor randomly in work stealing, it is proved to be optimal: justify reason why! **[3+2+4]**
- c. List schedule the string of following independent jobs on 3 machines and generate the Gantt Chart.

Job	J1	J2	J3	J4	J5	J6	J7	J7	J8	J9	J10
Size	2	6	4	6	2	1	8	3	2	5	2

Also generate Gantt chat using ***List with Largest Processing Time First*** scheduling. **[4+4]**