R.M.

Indian Institute of Technology, Kharagpur
Department of Computer Science and Engineering
End-Semester Examination

High Performance Computer Architecture (CS60003)

Time=3 Hours **Max Marks=87**

Important Instructions:

- **Answer all questions from both the Sections. Do not mix up answers to questions from the two sections.**
- **No clarification to any of the questions shall be provided.** In case you have any queries, you can make suitable assumptions, but please write down your assumptions clearly.
- All answers should be brief and concise. Lengthy and irrelevant answers will be penalized.
- **Use of simple calculators is permitted.**

---

1. Name any two techniques that can be used to eliminate the need for hardware-based interlocking in a pipelined processor. **[2]**

2. Assume that you have a 16-core processor and that the number of cores to be used for executing a program can be set as a compiler parameter. Assume that 1% code of the program is purely sequential and the rest of the code is ideally parallel. You need to achieve a speed up of 10 over a single core performance while using a minimum number of cores. How many cores would you configure the compiler to use? **[3]**

3. A computer uses a single cache of line size 64 bytes. The synchronous DRAM is clocked at 133 MHz, and requires 8 clocks for transmitting the first word on the memory bus. Subsequent words are transmitted at the rate of one word per cycle on the memory bus which is 32 bits wide. What is the miss penalty of the cache in nano seconds? **[3]**

4. Suppose for a certain program, the branch frequencies (as percentage of all instructions) are as follows: Conditional branches=15% and Jumps and calls =1%. Of the Conditional branches, 60% are taken on the average. This program is run on a four-stage pipeline, in which the unconditional branches are resolved at the end of the second cycle and the conditional branches get resolved at the end of the third cycle. If an instruction is decoded to be branch (at the end of the second stage), no further instructions are fetched till the branch is resolved. There are no other hazards. How much faster would the program run, if a dynamic branch predictor with 100% accuracy is used? **[3]**

5. Suppose you are designing a processor and your simulation results for an important workload indicate that branch instructions at the following two 32-bit addresses (in binary) are executed most frequently:
   Address A: 1000 1101 1100 0011 0110 1011 0100 1001
   Address B: 1000 1101 0110 1001 1110 1011 0100 1001

   The two branch instructions at A and B are likely to conflict (hash to the same entry) in a branch predictor. For a simple "bimodal" predictor of two-bit saturating counters, how many entries must the predictor have, in order to prevent these two branches from interfering with each other? What would be the size of the predictor in KBytes? **[2+1]**

6. How many stalls would the following code cause when executed on a simple 5-stage MIPS processor? Now assume that the Exe stage of the simple MIPS pipeline is split into 2 stages, without altering any of the other stages. How many stalls would result in executing the code on this modified pipeline. **[2+2]**

```
ADD R1,R2,R1
LW R2, 0(R2)
ADD R1,R2,R1
```

7. Assume the miss rate of an instruction cache is 2% and the miss rate of the data cache is 4%. If a processor has a CPI of 2 without any memory stalls and the miss penalty is 100 cycles for all misses, determine how much faster a processor would run with a perfect cache that never missed. Assume the frequency of all loads and stores is 36%. **[5]**

8. A 200 MHz simple MIPS pipeline-based processor deploys a single-level unified cache and on the average makes one memory reference per clock. The unified cache has an access time of 1 clock and has a hit rate of 95%. The cache deploys critical word first policy. It takes 50 ns to access the first word from main memory and 80 ns to load the complete cache line. What is the average utilization of the main memory? **[5]**

9. A 2.5 GHz processor achieves a CPI of 3 when simulated with a perfect cache. Assume that the cache being deployed has a miss penalty is 40ns and cache miss rate is 6%. What is the CPI of the processor with this cache? If a second level (L2) cache with a hit time 10 ns is added, the miss rate to the main memory is reduced to 2%. What speedup would be observed after adding L2 cache? **[2+3]**

10. Consider a cache coherent 32-processor computer that uses a directory-based cache coherence protocol. If each local memory stores 4 GBytes and the size of the cache block 128 bytes, how much space is necessary to store directory information in each processor? **[3]**

    How does the space required to store directory information in this computer change if the number of processors is increased to 64 processors? **[2]**

11. Consider that the following code for a 5-stage MIPS processor. Show how static scheduling could be used to eliminate Load-Use hazards in the code. Assume that various suitable forwarding schemes have been implemented. **[5]**

```
LW $s0,0($t0)
LW $s1,0($t1)
ADD $s2,$s0,$s1
SW $t2,0($s2)
LW $s3,0($t3)
LW $s4,0($t4)
```

12. Consider that a for a certain processor, the better L1 cache configuration is to be chosen for running a program having 33% load/store instructions. When a split L1 instruction and data cache of sizes 16KB each were used, Inst miss rate=0.64%, Data miss rate=6.47% was observed. When a unified

32KB L1 cache was used, a miss rate of 1.99 was observed. For both cases, hit time of 1 cycle and miss penalty is 50 cycles. Which cache configuration would give better performance and by how much? **[5]**

13. Consider the following CUDA program along with the input specifications for the array A.

```
__global__ is_divergent(int *A){
   int tid= blockIdx.x*blockDim.x + threadIdx.x;
   if(A[tid] %32==0) A[tid]+=1;
   if(A[tid] %32==1) A[tid]+=2;
   if(A[tid] %32==2) A[tid]+=3;
   if(A[tid] %32==3) A[tid]+=4;
         -----
         -----
   if(A[tid] %32==31) A[tid]+=32;
}
```

Content specifications for array A

i. A is filled such that every block of 4 elements contains the same number. For example, A[] = 111122223333…

ii. Every block of 8 elements in the array contains the same number

iii. Every block of 16 elements in the array contains the same number.

Suppose you have two alternate GPU architectures available to you: GPU1 (warp size =8) and GPU2 (warp size = 16). Threads inside a warp are scheduled concurrently. For each of three array content specifications (i, ii and iii) which one of the following conditions would hold true. **[2+2+2]**

a) The program exhibits divergent behavior for both architectures.

b) The program exhibits divergent behavior on GPU1 and not on GPU2.

c) The program exhibits divergent behavior on GPU2 and not on GPU1.

d) The program exhibits no divergent behavior on both the architectures.

14. The average CPI a processor is 2 with a perfect cache. Each instruction requires, on an average, 1 memory access for the instruction and 2 accesses for data. The internal cache is implemented as a Harvard cache. The hit ratio to the L1 instruction cache is 92% and the hit ratio to the L1 data cache is 87%. A unified L2 cache requires 6 cycles to access and has a global hit ratio of 98%. The AMAT for main memory is 25 cycles. What is the average MIPS rate of the processor if it is clocked at 400MHz? **[7]**

15. Designers of the Yoyodyne 370 processor are trying to decide the best organization for the data cache. Their primary customer, a video game console manufacturer, is very concerned about the performance of the Y-370 on a game-physics simulation engine. The Yoyodyne designers have been provided a benchmark of the simulation engine that contains 15% load instructions and 85% other instructions. A main memory access requires 50 ns. Assume there are no stalls in the pipeline except for those caused by data cache misses. The team is considering three possible data cache designs as indicated below. What would be the performance of each design in millions of instructions per second (MIPS)? **[3+3+3]**

a) **Option 1:** 16KB direct-mapped cache that has a hit rate of 85% and hit time of 1ns. **[3]**

b) **Option 2:** 16KB 2-way associative cache, which achieves a hit rate to 90%, but requires a memory pipeline stage of 1.1ns. **[3]**

c) **Option 3:** 32KB direct-mapped cache. The larger cache improves the hit rate to 92%, but slows the memory pipeline stage to 1.25ns. **[3]**

16. Consider an Intel P4 microprocessor with a 16 Kbyte unified L1 cache. The miss rate for this cache is 3% and the hit time is 2 Cycles. The processor also has an 8 Mbyte, on-chip L2 cache with hit time of 15 cycles. If data is not found in the L2 cache, a request is made to a 4 Gbyte main memory. The average latency for serving a memory request is 100,000 Cycles. It was observed that the processor on the average stalls 3.45 Cycles per memory request. How often is the main memory accessed (express in terms of number of main memory accesses per memory access request generated by the processor)? **[7]**

17. Consider a simple bus-based dual processor cache coherent SMP. Each processor has a single private cache, and coherence is maintained with the help of a snooping, write-back protocol. Each cache is direct mapped, with 2 blocks each holding 1 word. The even addresses are mapped to Block B0 and the odd addresses are mapped to Block B1. For example, addresses 0 and 2 are mapped to block B0, and addresses 1 and 3 are mapped to block B1.

    For each of the following questions, assume the initial cache state is as shown in the diagram below. Show the resulting state of the caches and memory after each action. Show only the blocks in processor caches and memory that change. For instance, after **[P0: read 3],** the changes are **[P0 B1: (Shared, 3, 7)],** indicating Processor 0's block B1 now has state = Shared, addr = 3, and data = 7. Also indicate the value returned by each read operation. **[2*5=10]**

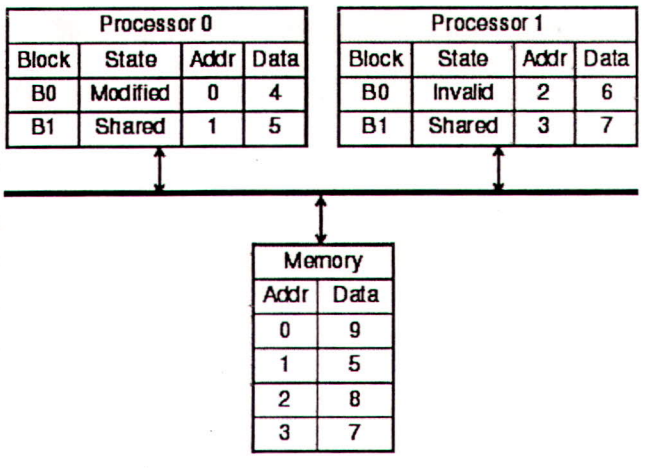
    a) Processor 1 reads 2

    b) Processor 1 writes value 10 to addr 2

    c) Processor 0 writes value 11 to addr 1

    d) Processor 1 writes value 12 to addr 0

    e) Processor 0 reads addr 2



Processor 0			
Block	State	Addr	Data
B0	Modified	0	4
B1	Shared	1	5

Processor 1			
Block	State	Addr	Data
B0	Invalid	2	6
B1	Shared	3	7

Memory	
Addr	Data
0	9
1	5
2	8
3	7

----The End----