Extra

INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR
End-Spring Semester 2017-18

Date of Examination : _____ Session (FN/AN) _Duration: 3 hrs Marks: 90

Subject No:CS60003 Subject:High Performance Computer Architecture

Department of Computer Science and Engineering

Specific charts, log book etc.: Use of simple non-programmable calculators is permitted.

Special Instructions:

- **Answer all questions from both Section A and Section B.**
- **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.**

Section A

1. Consider the following **for** loop that iterates 100,000 times. Out of the total number of iterations, in how many iterations the two branches: branch1 and branch2 are correctly predicted using 2-bit predictors, each initialized to strongly not taken? Assume that the size of the BPB (BHT) is extremely large. **[2]**

```
for (i=0; i<100,000; i++) {
        if (i%2==1) {     // branch 1: if i is odd
            ... ... ...   // do something
         }
}                         // branch 2: loop
```

2. An obvious disadvantage of multilevel cache inclusion (that is, blocks in a cache are also present in a lower level cache) is wastage of cache space. Yet, processors such as Intel's Core i7 use multilevel cache inclusion policy. Using only one short sentence identify a possible reason for this. **[2]**

3. In one short sentence write the purpose served by the **return address stack (RAS)** in a processor. **[2]**

4. Present day processor manufacturers often allow the processor clock frequency to be temporarily boosted above the so-called Thermal Design Power (TDP)—the sustainable thermal budget of the CPU. If the frequency is temporarily boosted by 20%, would you expect performance to increase by less than 20%, more than 20%, or exactly 20%? Justify your answer using one short sentence. **[2]**

5. Using one short sentence explain how the L2 cache controller of a multiprocessor determines the response corresponding to which one of the several outstanding memory requests is received over a packet switched bus? **[2]**

6. Is there any loop-carried dependence in the following code? If so, identify the dependence or dependences and mention which loop (that is, i-loop or j-loop) carries it/them. In case there is one (or more), for each identify whether it is cyclic and show your workout. **[2]**

```
for (i=1; i < n-1; i++)
    for (j=1; j < n-1; j++)
        A[i+2][2*j+1] = A[i+2][4*j-1] + A[i+2][4*j+1];
```

7. Suppose you want to run a program containing four separate threads of execution. Two of these threads experience an L2 cache miss (L2 miss penalty = 70 cycles) every 20 instructions, while the other two threads suffer only one L1 cache misses (L1 miss penalty = 4cycles) every 20 instructions. Would this program run faster: 1) On a processor using fine-grained or 2) On a similar processor using coarse-grained multithreading? Justify your answer using one short sentence. **[2]**

8. Suggest at least one scheme using which false sharing misses in a multiprocessor system can be reduced to a

large extent? Justify your answer using one short sentence. **[2]**

9. GPUs do not rely on large caches to hide the latency of accessing off-chip memory. Name the technique that they do use instead. Explain your answer using a short sentence. **[2]**
10. Almost no commercially available multicore processor supports shared L1 cache. Identify two important shortcomings of incorporating shared L1 caches as compared to private L1 caches. **[2]**

Section B

11. The clock rate of a simple MIPS processor is 200 MHz. Simulation results of the processor with a perfect cache system indicate a CPI of 1 while running a certain program. The processor is next fitted with a cache system having a hit time of 1 clock cycle and a hit rate of 95%. It takes 50 ns to access the first word from main memory and 80 ns to load the complete cache line. The cache uses critical word first technique. What is the average utilization of the main memory while running the program? **[5]**

12. Simulation experiments on the Dolphin processor for an important workload indicated that two branch instructions at the following two 32-bit addresses (in binary) execute frequently:
    - **Address A:1000 1101 1100 0011 0110 1011 0100 1001**
    - **Address B: 1000 1101 0110 1001 1110 1011 0100 1001**

    a) Based on an analysis of the low branch prediction accuracy that was observed, the processor designers suspect that the above two branch instructions must be conflicting (hashing to the same entry) in the BPB (BHT. For a simple "bimodal" predictor of two-bit saturating counters, at most how many entries must the predictor have to prevent these two branches from interfering (conflicting) with each other? In this case, at most how much storage space(in bytes) is required for implementing the BPB? **[1+1]**

    b) Simulation results indicated that a predictor with just 2048 entries (512 bytes) would be sufficient if it wasn't for the above two troublesome branches. The processor designers had a brilliant idea: "Why not create a set-associative BPB?" They considered a two-way set-associative predictor that uses a straightforward tagging strategy (each two-bit counter has a tag, instructions have 32-bit addresses). How large (in KBs) is a two-way set-associative 2048-entry tagged predictor? Assume that an LRU replacement would be used. **[3]**

13. Consider a MIPS-based multiprocessor that supports LL and SC instructions for synchronization.
    a) Write an efficient lock routine using LL and SC pair for mutual exclusion. **[1]**
    b) Assume that the processor also includes an atomic **TEST&RESET** instruction in its ISA. The atomic instruction **TEST&RESET R2, 0(R1)** fetches the content of the memory location 0(R1) into R2 and resets the content of 0(R1) to zero. Write an efficient lock routine using the TEST&RESET instruction. **[2]**
    c) What is the advantage of using the LL and SC pair as compared to using the TEST&RESET instruction? **[2]**

14. In a certain multiprocessor, the caches at various nodes are connected using a snooping bus. Assume that each node is a multiple-issue, out-of-order processor and can achieve an average IPC of 1.1. The average memory access per instruction is 1.2. Assume that the miss rate (global miss rate) of the L2 (last-level cache) is 1.5%. The bus speed (cycle time) is four times slower than the processor speed (cycle time). The bus can serve one request every 2 bus cycles. Furthermore, on the average 30% of the replaced blocks are dirty and need to be written back to the memory. Calculate at most how many processors the snooping bus can support. Note that you may assume that data bus connecting the memory as well as the memory are not bottlenecks. **[5]**

15. A designer is investigating two different cache designs for the memory subsystem of a new processor.

Option 1: A direct-mapped cache with a 90% hit rate and 2 ns hit time.

Option 2: A set-associative cache with a 93% hit rate and 10 ns hit time. To improve the performance of the cache, the designer has incorporated a way predictor. If the way predictor is correct, cache hits take only 3 ns, while way prediction misses that still produce cache hits take 12 ns.

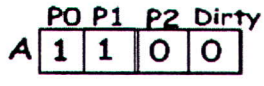
Assume that the cache miss penalty is 200 ns, regardless of the type of cache being used. How accurate must the way predictor be for the option 2 to result in better AMAT? **[5]**

16. Suppose the threads running on a 10 node multiprocessor each try to lock a shared variable simultaneously. The processor supports the load-lock (LL) and a store-conditional (SC) instructions. You can assume that each bus transaction (read miss or write miss) is 100 clock cycles long. You can ignore the time of the actual read or write of a lock held in the cache, as well as the time duration for which a lock is held. Further assume that all the threads are all spinning when the lock is released at time 0. Determine the number of bus transactions required for all 10 processors to acquire the lock. **[5]**

17. Consider a simple bus-based symmetric multiprocessor. Each processor has a single private **write-through** cache. Cache coherence maintained with a snooping, invalidate protocol. How many states would the state machine implementing the protocol have? Name the states. Draw the state machine diagram showing the events causing the transitions among the states and the associated actions if any. **[1+4]**

18. The designers of the R-Star Processor are trying to improve the performance of their branch resolution logic. Their base processor design achieves an IPC of 0.8 for the SPEC benchmark suite. Their proposed change shortens the branch execution pipeline, shaving 2 cycles off the average branch misprediction penalty. However, the design change increases the clock period by 10%. Over the entire benchmark suite, 20% of instructions are branches. However, we unfortunately do not know the prediction accuracy of the branch predictor. For what range of prediction accuracies will this change result in a performance improvement? Clearly show the details of your work out. **[5]**

19. Consider a multi-processor system with three nodes, P0, P1, and P2. The threads running on these three processors share a block having address A. The system uses a write-invalidate, **directory-based coherence protocol**. Initially, the directory entry for the block A is as in the adjoining diagram:



If P2 now attempts to write to block A, what messages are sent between the nodes and the directory. What would be the new directory entry for the block A? You may want to draw a diagram to support your answer. **[4+1]**

20. Consider a multi-processor having the following characteristics:
    - 8-processor bus-based multiprocessor with following private caches : L1 instruction, L1 data, and L2 cache. The L3 cache is shared.
    - Snoopy invalidation protocol on write back caches is used for cache coherence.
    - Each processor is hyperthreaded and on the average two instructions are issued per cycle, out of which one is a memory operation (load or store).
    - At each processor, the observed L1 data cache miss rate is 20%, L1 instruction miss rate is 10%, and L2 cache local miss rate is 25%.

    In this system, for any of the L2 caches, much more accesses to the tag occur than data. Why? How many tag accesses occur per cycle? **[1+4]**

21. In the dynamic instruction scheduling (Tomasulo's algorithm) we studied, the steps carried out for an ALU instruction are the following:
    1. Issue when reservation station and ROB entry is available
       - Read already available operands from registers and instruction

- Send instruction to reservation station
- Tag unavailable operands with ROB entry
- Tag destination register with ROB entry
- Write destination register to ROB entry
- Mark ROB entry as busy

2. Execute after issue
   - Wait for operand values on CDB (if not already available)
   - Compute result
3. Write result when CDB and ROB available
   - Send result on CDB to reservation stations
   - Update ROB entry with result, and mark as ready
   - Free reservation station
4. Commit when at head of ROB and ready
   - Update destination register with result from ROB entry
   - Untag destination register
   - Free ROB entry

What are the corresponding steps for handling a store instruction? **[5]**

22. Consider a pipelined processor that has an average CPI of 1.8 when there are no memory stalls. The instruction cache has a hit rate of 95% and the data cache has a hit rate of 98%. Assume that memory reference instructions account for 30% of all the instructions executed. Out of these 80% are loads and 20% are stores. On the average, the read-miss penalty is 20 cycles and the write-miss penalty is 5 cycles. Compute the effective CPI of the processor by accounting for memory stalls. **[5]**

23. Assume that the base CPI for a pipelined datapath in a single core processor is 1, when there are no *cache misses.* Profiles of a benchmark suite that was run on this *single core chip* with an L1 cache (code named C1) suggest that for every 10,000,000 accesses to the cache, there are 300,000 L1 cache misses.
    - If data is found in the L1 cache, it can be accessed in 1 clock cycle, and there are no pipe stalls
    - If data is not found in the L1 cache, it can be accessed in 10 clock cycles

    Now, consider a dual-core version of the processor in which the private L1 cache at each processor is code named C2. All cores access a shared L2 cache. Coherency of L1 data is maintained by deploying an MSI coherency protocol.

    Benchmark profiling obtained by running the same benchmark suite on the dual-core system suggests that, on an average, there are now 450,000 misses (considering both the L1 caches) per total of 10,000,000 accesses.
    - If data is found in a cache, it can still be accessed in 1 clock cycle
    - However, on the average, 14 cycles are now required to satisfy an L1 cache miss.

    What must the CPI of the multi-core system with a perfect cache, so that when fitted with C2 cache it will outperform the single core processor fitted with C1 cache? **[5]**

24. A 5-stage MIPS processor C1 running at 2.2 GHz is used to execute a program P1. The instruction statistics for P1 are as follows: Branches: 20%, Loads: 20%, Stores: 10%, and Arithmetic Instructions: 50%. Assume that the program P1 has no data dependencies. C1 uses a dynamic branch predictor and a branch target buffer to predict the outcome of the branch instructions with a prediction accuracy of 80%. The computation of actual branch outcomes is carried out in the "decode" stage. Also assume that C1 uses a cache system in which 100% of the instruction fetches and x % of the data accesses hit in the cache. The penalty to access the main memory for a cache miss is 10 cycles. A customer requires that the processor C1 must achieve a throughput of 2 billion instructions per second. Calculate the minimum value of x (data cache hit rate) that would satisfy this requirement. **[5]**

-----The End ----