1. A gshare predictor uses saturating counters state machines. The branch predictor state is 00 and the branch is predicted not taken. The next state will be:

a. 00

b. 01

c. 10

d. 11

e. Need more information to determine next state.

1. A local correlating branch predictor uses 10 history bits and 12 address bits. The size of the state machines array is:

a. 512 entries

b. 1K entries

c. 2K entries

d. 4K entries

e. 2M entries

1. The gshare and bimodal arrays are usually non-tagged arrays.

a. True

b. False

1. A direct mapped cache is the same as a 1-way associative cache

a. True

b. False

1. Capacity misses are inversely proportional to the number of bytes in the cache.

a. True

b. False

1. Choose the one TRUE statement about write-through cache.

a. On a store miss, the store is written to memory, but only if the cache is not write-allocate.

b. On a store hit, the store is written to memory, but only if the cache is not write-allocate.

c. On a store hit, the store is written to memory, but only if the cache is write-allocate.

d. A store is written to memory regardless if the store hit or miss the cache.

1. Conflict misses decrease as the set associativity in the cache decreases.

a. True

b. False

1. Hardware prefetching reduces compulsory misses.

a. True

b. False

1. A gshare branch predictor uses 12 history bits and 10 address bits. The size of the state machines array is:

a. 512 entries

b. 1K entries

c. 2K entries

d. 4K entries

e. None of the above

1. A cache has a total capacity of 32K bytes. It is implemented as 4-way set associative, with block size of 32 bytes. The physical address on the machine consists of 32 bits. Which of the following statements is TRUE?

a. Number of set index bits = 7 and number of tag bits = 20

b. Number of set index bits = 8 and number of tag bits = 19

c. Number of set index bits = 9 and number of tag bits = 18

d. Number of set index bits = 8 and number of tag bits = 20

e. None of the above

1. A cache has a total capacity of 32K bytes. It is implemented as a fully set associative cache, with block size of 32 bytes.

The physical address on the machine consists of 32 bits. Which of the following statements is TRUE?

a. Number of tag bits = 20

b. Number of tag bits = 13

c. Number of tag bits = 27

d. Number of tag bits = 32

e. None of the above

1. Consider the following sequence of address references issued by a CPU to a cache.

0xFFFF0108, 0xFFFFC100, 0X1FFF7100, 0XFFFF4100, 0xFFFFC104, 0xFFFF0100, 0xFFFF4100

Assuming a 2-way set associative cache with block size of 16 bytes, total capacity of 8Kbytes and LRU replacement policy. Which of these statements is TRUE?

a. There will be 6 compulsory misses and 1 hit

b. There will be 3 compulsory misses, 1 conflict miss and 3 hits

c. There will be 4 compulsory misses, 2 conflict misses and 1 hit

d. There will be 4 compulsory misses and hits

e. There will be 4 compulsory misses and 3 conflict misses

1. Which of the following statements is FALSE?

a. Virtual memory allows large applications that do not fit in DRAM to execute with good performance

b. Virtual memory makes DRAM effectively a cache for the very slow hard disk

c. Virtual memory is implemented completely in software

d. Virtual memory is implemented using combination of hardware and software

e. Virtual memory also allows multiple running applications to share the DRAM

1. Which of the following statements is false

a. Access bit is set for a page in DRAM when the page is read or written.

b. Access bits are periodically cleared by the operating system.

c. Access bit indicates if a page has been accessed since it has been loaded from the disk.

d. Access bits are stored with the physical address in the page tables.

e. Access bits are used in page replacement algorithm.

1. A TLB miss indicates that a page is not in DRAM and needs to be loaded from the disk.

a. TRUE

b. FALSE

1. A page fault exception always occur after a TLB miss

a. TRUE

b. FALSE

1. A processor uses the fact that the page offset bits are the same in both virtual and physical addresses to reduce the latency it takes to perform address translation and cache lookup. The processor uses the page offset bits and starts the set access to read data and tags in the same cycle that the TLB performs the virtual to physical translation. Assuming a page size of 4K bytes, cache block size of 32 bytes and 4-way set associative cache. What is the maximum cache size that this processor can implement?

a. 4K bytes

b. 8K bytes

c. 16K bytes

d. 32Kbytes

e. Any size that is a power of 2 is possible

1. A CPU write hits a block in shared state and the block is in the cache of another processor. Which of the following statements is TRUE?

a. The other processor writes back its block and invalidates the block in its cache.

b. The other processor invalidates the block in its cache without writing the block back.

c. The other processor does nothing.

d. None of the above.

1. A CPU write misses and the block is in exclusive-modified state in another processor. Which of the following statements is TRUE?

a. The other processor writes back its block and invalidates the block in its cache.

b. The other processor writes back the block and changes its state to Shared.

c. The other processor does nothing.

d. None of the above.

1. A CPU read hits a block in shared state and the block is in the cache of another processor. Which of the following statements is TRUE?

a. The other processor writes back its block and invalidates the block in its cache.

b. The other processor writes back its block and invalidates the block in its cache.

c. The other processor does nothing.

d. None of the above.

1. Multi-cycle execution units cause R after W hazards.

a. True.

b. False.

1. Multi-cycle execution units cause W after W hazards.

a. True.

 b. False.

1. Multi-cycle execution units cause W after R hazards.

a. True.

b. False.

1. Which of the following statements is TRUE:

a. The reorder buffer was first used in Tomasulo’s algorithm.

b. Most microprocessors that feature out-of-order execution use a future file.

c. Instructions read operands from the reorder buffer in-order.

 d. A history buffer provides better performance than a future file.

e. Instructions write results to the reorder buffer in-order.

1. A branch predictor uses saturating counters state machines. The branch predictor state corresponding to a branch being predicted is 01 and the branch is mispredicted. The next state will be:

a. 00

b. 01

c. 10

 d. 11

e. Need more information to determine the next state.

1. Choose the pair of terms that are most related:

a. History buffer, branch predictor

b. History buffer, branch target buffer

c. History buffer, branch mispredictions

 d. History buffer, register renaming

e. History buffer, Pentium Pro processor

1. A correlating global branch predictor uses 7 history bits and 8 address bits. The size of the state machines array is:

a. 128 entries

b. 256 entries

c. 384 entries

d. 32K entries

 e. None of the above.

1. Describe what happens in a snoopy cache coherence multiprocessor system when
2. A CPU write misses and the block is in exclusive-modified state in another processor.
3. A CPU write hits a block in shared state and the block is in the cache of another processor.
4. A CPU read misses a block which is in exclusive-modified state in another processor.

In your discussion, give the state transition of the involved block in both processor caches, the transaction that is sent on the external system bus, and which hardware unit supplies the data of the accessed block.

1. Name the hardware mechanism used to speed up virtual memory. Very briefly describe what it is.
2. Name six techniques for improving cache and memory access performance.
3. In virtual memory, what is a TLB page miss, and what is a page fault?
4. Explain how virtual memory provides performance and improves the utilization of computers by allowing them to perform multiple tasks concurrently.
5. A processor uses the fact that the page offset bits are the same in both virtual and physical addresses to reduce the latency it takes to perform address translation and cache lookup. The processor uses the page offset bits and starts the set access to read data and tags in the same cycle that the TLB performs the virtual to physical translation. Assuming a page size of 4K bytes, cache block size of 32 bytes and 4-way set associative cache. What is the maximum cache size that this processor can implement? What would you change in the cache organization to increase the size of the cache?
6. Calculate the performance improvement of a 4-wide superscalar processor over a 2-wide superscalar processor, both processors having equal cycle time and 20 cycle pipeline length from fetch to branch execution. Assume 10% branch misprediction rate and that 1 out of 7 instructions are conditional branches. Also assume unlimited instruction level parallelism, i.e. no stalls due to data hazards or cache misses.
7. A history buffer and a reorder buffer are two methods for implementing precise faults and for recovering from mispredicted branches. Which method gives better performance and why?
8. What type of hazards is created in a pipeline with variable-latency functional units that would not occur with equal-latency functional units? Give an example.
9. Name three mechanisms to provide precise exceptions in a processor that performs out-of-order execution. Which method has the highest branch misprediction penalty and why?
10. Both local and global branch predictors use branch history registers for indexing an array of state machines. What is the difference between the two history registers?
11. Explain the principles of temporal and spatial locality in programs and describe three hardwarevperformance optimizations that exploits temporal and spatial locality.
12. Use a block diagram to describe how a hybrid bimodal-gshare predictor works. Assume that the predictors’ predictor has 4K entries, the bimodal predictor has 4K entries and the gshare predictor has 16K entries.
13. Assume that we make an enhancement to a computer that improves some mode of execution by a factor of 5. Enhanced mode is used 80% of the time, measured as a percentage of the execution time *when the enhanced mode is in use.* Recall that Amdahl’s Law depends on the fraction of the original, *unenhanced* execution time that could make use of the enhanced mode. Thus, we cannot directly use this 80% measurement to compute speedup with Amdahl’s Law.
	1. What is the speedup we have obtained from fast mode?

b. What percentage of the original execution time has been converted to fast mode?