Note: You can give the answers in English or Swedish.

1. a) Why is memory access the bottleneck of a computer system?

   b) Describe two different methods to increase the bandwidth of the main memory, without considering the cache. What are the advantages and disadvantages of the described methods, respectively?

   (3 p)

2. a) What is the basic idea of associative mapping for cache organization? What are the advantages and disadvantages of the associative mapping organization?

   b) Why is the fully associative cache organization seldom used in practical computers? Which cache organization is commonly used? Why?

   (3 p)

3. The following sequence of virtual page numbers is encountered in the course of execution on a computer with virtual memory:

   5 7 5 4 1 2 3 6 4 7 1 2 1 7

   Assume that the least-recently used (LRU) page replacement policy is used. Assume also that the main memory has four page frames, and is initially empty. How many page misses will be encountered during this execution? Which are the virtual pages in the main memory when this execution finishes?

   (3 p)

4. What are the most important issues to be considered when designing an instruction set for a computer architecture? Explain in which way each of these issues has an impact on the performance of the computer.

   (2 p)

5. a) Why is instruction pipeline *widely* used to enhance performance of modern computers?

   b) What are the three main types of hazards that can reduce the performance of an instruction pipeline? What are the impacts of each of these hazards?

   c) In your opinion, which hazard causes the biggest problem for instruction pipeline? Why?

   (4 p)

Note: You can give the answers in English or Swedish.

6. Consider the following instruction sequence:

```
L1: MUL R3,R4          Note: R3 <- R3 * R4
    ADD R1,#1          Note: R1 <- R1 + 1
    BLT R1,#16, L2     Note: Branch to L2, if R1 > 16
    SUB R2, R3         Note: R2 <- R2 - R3
    ...
L2: ...
```

Transform this sequence in order to make use of the delayed branch technique. Show how the original sequence and the transformed one are executed in a pipelined CPU with two pipeline stages, and illustrate the reduction of the delay (draw two figures to illustrate the corresponding pipelined executions of the original sequence and the transformed sequence, respectively).

(3 p)

7. a) Provide a taken/not-taken execution pattern consisting of a sequence of 4 branches where a bimodal predictor will perform better than a one-bit predictor. Assume that the one-bit predictor is initially set at 0 (not taken) while the bimodal predictor is initially set at 01 (weakly not taken). Your answer should be of the form {T, T, T, NT}, for example.

b) Explain why the bimodal predictor gives better prediction results than those of the one-bit predictor.

(3 p)

8. The design of RISC architectures is based on certain characteristics of program execution.

a) What are the characteristics concerning procedure calls and returns?

b) What is the mechanism used in a RISC architecture to make procedure calls and returns efficient? How does this mechanism work?

(3 p)

9. a) What are the most essential characteristics of superscalar architecture and VLIW architecture, respectively?

b) Compare superscalar architecture with VLIW architecture from a performance point of view. Which architecture gives usually better performance? Why?

c) Compare superscalar architecture with VLIW architecture from a power consumption point of view. Which architecture consumes usually less power? Why?

(3 p)

Note: You can give the answers in English or Swedish.

10. a) Define the concept of loop unrolling. Why is loop unrolling very useful for a VLIW processor?

    b) Can the loop unrolling technique be used in a superscalar processor? Why?

    c) If a loop is unrolled completely (i.e., there is no need for loop control any longer), what will happen? Discuss the negative side effects of unrolling a loop completely?

    (3 p)

11. a) What is a NUMA computer system? What are the motivations for using such a system?

    b) Draw a picture of a typical NUMA system. Use the picture to illustrate and discuss the important concepts and components of such a system.

    c) What is the purpose of having directories in a NUMA system? What information is stored in them?

    (3 p)

12. a) There are two basic approaches to implement a snoopy protocol: write-invalidate and write-update. How do they work, respectively?

    b) Describe the situation when the write-invalidate approach works better, and the situation when the write-update works better, respectively.

    c) Both these approaches suffer from false sharing overheads. What does it mean by false sharing here?

    (4 p)

13. a) Why has Graphics Processing Unit (GPU) become very popular recently?

    b) Describe all features of a GPU that have contributed to its high performance.

    c) Describe all features of a GPU that have contributed to its power efficiency.

    (3 p)