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

1. a) What does it mean by a memory hierarchy? Why it is useful to build a memory hierarchy?

   b) What is the fundamental assumption which makes a memory hierarchy work efficiently in a computer system?

   (3p)

2. a) There are several write policies that are used to keep the cache contents and the contents of the main memory consistent. Describe briefly each of these policies and discuss the advantages and disadvantages of each of them, respectively.

   b) Describe the additional problems we have when applying these policies in a multiprocessor system.

   (3p)

3. A computer has a main memory and a cache. If a referenced word is in the cache, 5 *ns* are required to access it, while the main memory's access time is 50 *ns*. Assume that the cache block size (line size) is 4, the cache hit ratio is 0.98, and the time needed to check for cache hit/miss is 1 *ns*. What is the average time in *ns* required to access a word in this memory system? (Note that you should give the steps of your calculation, not just the final result.)

   (3p)

4. a) What does it mean by virtual memory? Describe how a virtual memory works.

   b) Why is it not needed to have all of the pages of a program to be in the main memory while the program is being executed?

   c) How is a logical (virtual) address converted into a physical address of the main memory?

   (3p)

5. a) What are the main issues to be considered when designing an instruction set for a computer architecture?

   b) Explain in which way each of these issues has an impact on the overall performance and implementation cost of a computer system.

   (3p)

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

6. a) What is the basic principle of an instruction pipeline? Why is it useful to have such a pipeline?

   b) Instruction pipeline has been widely used in almost all modern computers. Why?

   c) In general, a larger number of pipeline stages gives a better performance. However, this has not led to the situation that we have a huge number of pipeline stages. Why?

   (3p)

7. a) Explain the difference between static and dynamic branch prediction.

   b) Describe briefly an example of static branch prediction and dynamic branch prediction, respectively.

   c) Which branch prediction method gives better prediction results? Why?

   (3p)

8. a) What are the main characteristics of a superscalar architecture? You should provide at least two of them.

   b) Why is the window of execution an important mechanism for a superscalar architecture?

   c) Why do we have a “commit” mechanism in a superscalar architecture? How does this mechanism work?

   (3p)

9. a) Identify all the different types of data dependencies in the following code. Indicate the type of dependency you have identified for each one, and give the reasons for your answers.

```
L1: move r3,r9        Note: r3 <- r9
    load r8,(r3)      Note: r8 <- memory location pointed by r3
    add r4,r3,4       Note: r4 <- r3 + 4
    load r9,(r4)      Note: r9 <- memory location pointed by r4
    ble r8,r9,L1      Note: branch to L1 if r8 <= r9
```

   b) Which of the identified dependencies can be eliminated? How?

   (3p)

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

10. a) A VLIW architecture is said to support explicit parallel instruction execution. Define the concept of explicit parallelism. What are the advantages of explicit parallelism?

    b) What is the main problem of a VLIW computer? How is this problem addressed by the IA-64 architecture?

    (3p)

11. 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?

    (4p)

12. a) Which features of a graphics processing unit (GPU) have contributed to its high performance? Why?

    b) Discuss the concept of divergent execution in a GPU processor. What is the main impact of such divergent execution?

    c) Discuss one technique that can be used to address the divergent execution problem.

    (3p)

13. Describe all low-power techniques and principles that can be used for architecture design (Note: we are not interested in low-power techniques that are used at circuit, logic, micro-architecture, and software levels). Explain why the techniques and principles you have described can help to reduce the power consumption of a computer.

    (3p)