Practice Questions

1) Define Instruction-Level Parallelism (ILP). List the two main approaches to achieve ILP, and briefly describe compiler based static ILP.
2) Explain how does hardware-based dynamic ILP approach differ from compiler-based static ILP?
3) What is a basic block and why parallelism is limited within a basic block.
4) Describe the formula for calculating Pipeline CPI. How does pipeline CPI influence the overall speedup of a system?
5) Identify and describe the types of data dependence. Explain how does data dependences affect pipeline performance?
6) What is loop unrolling in the context of compiler techniques for ILP? Describe the steps involved in loop unrolling.
7) Discuss one limitation of loop unrolling and explain how strip mining can be used as an alternative.
8) Define dynamic scheduling and explain the process of dynamic scheduling through a diagram.
9) What is register renaming, and how does it help resolve Write After Read (WAR) and Write After Write (WAW) hazards?
10) Explain Tomasulo's Algorithm step-by step with diagrams. Illustrate this algorithm with example.
11) Outline the steps involved in speculative execution, including instruction commit and the role of the Reorder Buffer (ROB).
12) Describe the operations in a ROB.
13) Compare Tomasulo's algorithm with and without the use of ROB. Using a diagram, highlight key differences.
14) Based on your comparison, evaluate the performance benefits of using a ROB in Tomasulo's algorithm.
15) Define superpipelining. What is multiple issue in pipelining? Demonstrate through example how it is handled with speculation.
16) Define multithreading and describe how it enhances processor efficiency.
17) Describe the VLIW (Very Large Instruction Word) processor architecture. What are the primary disadvantages of using VLIW architectures?
18) Define the concept of hyperthreading along with a diagram.

19) Define a GPU and Compare the roles of a CPU and a GPU in a computing system.
20) Explain CPU-GPU interaction process and describe how this interaction affects processing speed.
21) Describe Flynn's Classification**.** Show the comparison of MIMD and SIMD architectures.
22) What is vectorization, and how does it differ from general parallel processing?
23) Describe the execution models for vectorization: SISD, SIMD, and multithreaded. Provide an example of each model.
24) Explain how GPU functions as a SIMT (Single Instruction, Multiple Threads) machine.
25) Describe the concept of warps and warp-level execution in a SIMT model.
26) Define and illustrate SIMD execution units within a SIMT framework.
27) Explain warp instruction-level parallelism and discuss how memory access works in SIMT systems to optimize throughput.
28) Explain the memory hierarchy in a GPU. Describe thread scheduling in GPUs and explain how it enhances parallel execution.
29) What is CUDA. Illustrate the processing flow in CUDA, with its flow architecture for 1D and 2D grids.
30) Describe latency hiding and memory coalescing in GPU memory management. Differentiate between coalesced and uncoalesced memory access.
31) What is data reuse in the context of GPU processing? Describe the concept of tiling as a data reuse technique.
32) Define SIMD warp utilization and explain techniques for warp divergence.
33) Describe asynchronous data transfers between the CPU and GPU, and explain how overlapping communication with computation can improve performance.
34) Explain the concept of memory hierarchy. Describe the role of cache memory in improving processor performance.
35) Define the principle of locality. What are spatial and temporal locality?
36) Discuss the general organization of a cache memory with a diagram. How does a processor interact with cache memory in a typical CPU-cache interaction?
37) Describe how index and offset are calculated in cache addressing.
38) Summarize the four main cache memory design choices.
39) Display a comparative study of block placement techniques.
40) Describe block identification in direct-mapped, set-associative, and fully associative caches.

41) Explain the concept of cache indexing and how it affects cache efficiency.

42) What are the main methods of block replacement in cache memory?

43) List and briefly describe each method of block replacement.

44) Compare look-aside and look-through caches. Discuss the trade-offs associated with look-aside and look-through cache architectures in terms of latency and cache access speed.

45) Define write hits and write miss strategies and explain which strategy pairs of write hits and miss performs best.

46) Explain in detail the Least Recently Used (LRU) and Least Frequently Used (LFU) block replacement policies.

47) Define write-through and write-back strategies for handling write hits in cache memory.

48) Define the three main types of cache misses and provide an example of each type.

49) Describe how each type of cache miss impacts cache performance and discuss one technique for reducing each type of miss.

50) Consider a 2-way set associative mapped cache of size 16 KB with block size 256 bytes. The size of main memory is 128 KB. Find-

Number of bits in tag 2.Tag directory size

51) Consider an 8-way set associative mapped cache of size 512 KB with block size 1 KB. There are 7 bits in the tag. Find-

Size of main memory 2.Tag directory size

52) Consider a 4-way set associative mapped cache with block size 4 KB. The size of main memory is 16 GB and there are 10 bits in the tag. Find-

Size of cache memory 2.Tag directory size

53) Consider an 8-way set associative mapped cache. The size of cache memory is 512 KB and there are 10 bits in the tag. Find the size of main memory.

54) Consider a 4-way set associative mapped cache. The size of main memory is 64 MB and there are 10 bits in the tag. Find the size of cache memory.

55) If memory size is 128 KB and memory is Byte addressable then how many bits required to Represent Physical address of main Memory

56) If memory size is 128 KB and memory is word addressable and word size is 8 byte then how many bits required to Represent Physical address of main Memory

57) If we have 8 words in block and each word is of 32 bit then how many Bits are required for Block offset? Main memory is Byte addressable

58) If memory size is 128 KB and memory is Byte addressable then how many bits required to Represent Physical address of main Memory

59) If memory size is 128 KB and memory is word addressable and word size is 8 byte then how many bits required to Represent Physical address of main Memory

60) If we have 8 words in block and each word is of 32 bit then how many Bits are required for Block offset? Main memory is Byte addressable

61) Consider a direct mapped cache of size 16 KB with block size 256 bytes. The size of main memory is 128 KB. Find-

- Number of bits in tag
- Tag directory size

62) Consider a direct mapped cache of size 512 KB with block size 1 KB. There are 7 bits in the tag. Find-

- Size of main memory
- Tag directory size

63) Consider a direct mapped cache with block size 4 KB. The size of main memory is 16 GB and there are 10 bits in the tag. Find-

- Size of cache memory
- Tag directory size
- No. of physical address bits
- No. of bits in block offset

64) In two level hierarchy, the cache has an access time of 12ns and the main memory access time of 120ns, the hit rate of cache is 90%. If the block size of cache is 16 bytes, then what is the average memory access time including Miss Penalty? (Miss Penalty: Time to bring main memory block to cache memory when cache miss occurs)