1 Power Equation

Which equation is used for first order power estimation?

Select an alternative:

○ P = (1/2)*C*V*(f^2)

○ P = (1/2)*(C^2)*V*f

○ P = (1/4)*C*(V^2)*f

○ P = (1/2)*C*(V^2)*f ✓

Maximum marks: 3

2 Moore's Law

Moore's law predicts that

Select an alternative:

○ the number of transistors doubles roughly every 18 months. ✓

○ the power reduces by half roughly every 18 months.

○ the number of cores doubles roughly every 18 months.

○ the performance doubles roughly every 18 months.

Maximum marks: 3

3 Flynn's taxonomy

Which of the following is not part of Flynn's taxonomy?

Select an alternative:

○ SIMT ✓

○ SISD

○ MIMD

○ SIMD

Maximum marks: 3

4 Dennard scaling

Which is the most accurate statement regarding Dennard Scaling?
Select an alternative:

○ Dennard Scaling neglects the impact of voltage scaling.

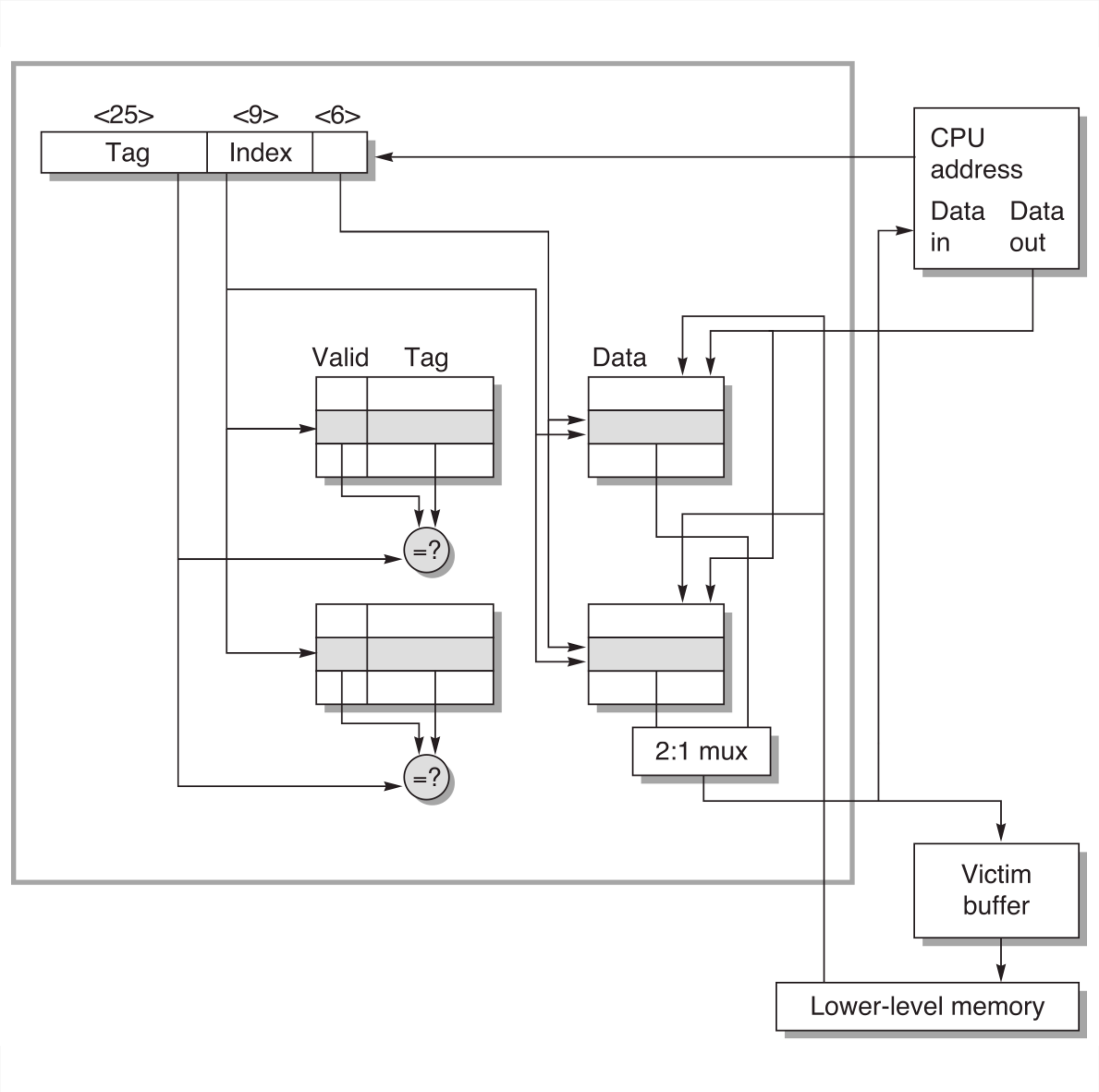
○ The power density remains constant across technology nodes.

○ Dennard Scaling neglects the impact of leakage currents. ✓

○ The power of a chip remains constant across technology nodes.

Maximum marks: 3

5 Cache property



Which statement is correct regarding the cache shown in the figure?

Select an alternative:

○ The cache size is 64 KiB. ✓

○ The cache size is 32 KiB.

○ The cache block size is 512 bytes.

○ The cache has 64 blocks.

Maximum marks: 3

6 Cache misses

Which of the following is not one of the four Cs?

Select an alternative:

○ Capacity

○ Conflict

○ Compulsory

○ Concurrent ✓

Maximum marks: 3

7 VIPT cache

Which of the following statements are incorrect given a VIPT cache?

Select an alternative:

○ Requires the associativity to be increased to increase the cache size.

○ Enables parallel access to the TLB.

○ Limits the cache way size to the page size.

○ Limits the cache block size. ✓

Maximum marks: 3

8 Cache access time

Given a system with a memory hierarchy consisting of three cache levels and DRAM as main memory. The access time for the caches are

L1=1 cycle,

L2=10 cycles,

L3=30 cycles, and

DRAM=70 cycles.

What is the average access time for an application with the following characteristics:

L1: 95% hit 5% miss

L2: 90% hit 10% miss

L3: 99% hit 1% miss

Select an alternative:

○ 1.85

○ 1.96

○ 1.50

○ 1.65 ✓

Maximum marks: 3

9 Data dependency

Which of the following is a true data dependency?

Select an alternative:

○ Write after read

○ Read after write ✓

○ Write after write

○ Read after read

Maximum marks: 3

10 Hazards

Which of the following is not a hazard?

Select an alternative:

- ○ Control hazards
- ○ Structural hazards
- ○ Data hazards
- ○ Execution hazards ✓

Maximum marks: 3

11 Loop unrolling

Loop unrolling is a compilation technique for improving instruction level parallelism.

Which of the following statements are incorrect?

Select an alternative:

- ○ It can be complicated to apply on loops with dynamic bounds.
- ○ It can reduce the number of instruction cache misses. ✓
- ○ It can reduce the total number of executed instructions.
- ○ It can increase register pressure.

Maximum marks: 3

12 VLIW

Which of the following statements are incorrect regarding a very long instruction word architecture?

Select an alternative:

○ Relies on compiler optimizations for extracting ILP.

○ Reduces the amount of hazard detection logic needed.

○ Increases the binary code compatibility. ✓

○ Instructions are executed in lock step.

Maximum marks: 3

13 SIMD

Which of the following statements are false?
Select an alternative:

○ SIMD improves energy efficiency by reducing the instruction overhead per performed operation.

○ SIMD can reduce the code size.

○ SIMD is usually very efficient on streaming (e.g., video, audio) applications.

○ SIMD as a concept can be used for all types of applications. ✓

Maximum marks: 3

14 Vector architecture

Which of the following is not a conventional vector architecture technique?
Select an alternative:

○ Scatter gather

○ Mask register

○ Chaining

○ Threading ✓

Maximum marks: 3

15
SIMD vs Vector

Which of the following options is true for SIMD extensions when compared against a vector architecture?

Select an alternative:

○ Has flexible number of operands.

○ Easier to implement in hardware. ✓

○ More sophisticated addressing modes.

○ Has a mask register.

Maximum marks: 3

16
Roofline model

What is a roofline model used for?

Select an alternative:

○ A roofline model is a way to visualise the maximum required power given an applications arithmetic and bandwidth intensity.

○ A roofline model is a way to visualise the arithmetic intensity of an application given its performance and bandwidth requirements.

○ A roofline model is a way to visualise the maximum performance given ✓ an applications arithmetic intensity.

○ A roofline model is a way to visualise the maximum required bandwidth given an applications arithmetic intensity.

Maximum marks: 3

17
GPUs

GPUs hide memory latency by

Select an alternative:

○ GPUs have short memory latency, which doesn't have to be hidden.

○ Implementing large caches.

○ Many levels of caches.

○ Massive amount of threads. ✓

Maximum marks: 3

18 Simultaneous multi-threading

Which of the following statements are false regarding SMT?
Select an alternative:

○ Requires a large register file.

○ Can increase the cycle time.

○ Reduces cache contention. ✓

○ Much of the hardware support exists already in an out-of-order scheduled processor.

Maximum marks: 3

19 Branch divergence

What is branch divergence?
Select an alternative:

○ Branch divergence is when a branch misprediction occurs and the execution diverges down the wrong execution path.

○ Branch divergence is when a branch causes the execution of a vector operation to diverge.

○ Branch divergence is when an interrupt occurs and the threads need to branch to the interrupt routine.

○ Branch divergence is when multiple threads in lock step execute different code paths. ✓

Maximum marks: 3

20 Speedup

What is the maximum speedup for any computer system with 300 cores executing a parallel application with a sequential section of 5%?

Select an alternative:

○ 188x

○ 18.8x ✓

○ 20x

○ 300x

Maximum marks: 3

21 Multicore benefits

Which of the following statements are incorrect?

Select an alternative:

○ It is often easier to achieve high performance using multiple small co ✓ than a few large cores.

○ Many small cores can provide better energy efficiency than a few large cores.

○ Wire lengths dominate the delay so keeping the core small can improve performance.

○ According to Pollack's rule it's more efficient use of chip area (in terms of performance) to build many small cores than one large core.

Maximum marks: 3

22 Cache coherency

Which of the following statements are correct?

Select an alternative:

○ Cache coherency protocols specify the order of memory operations.

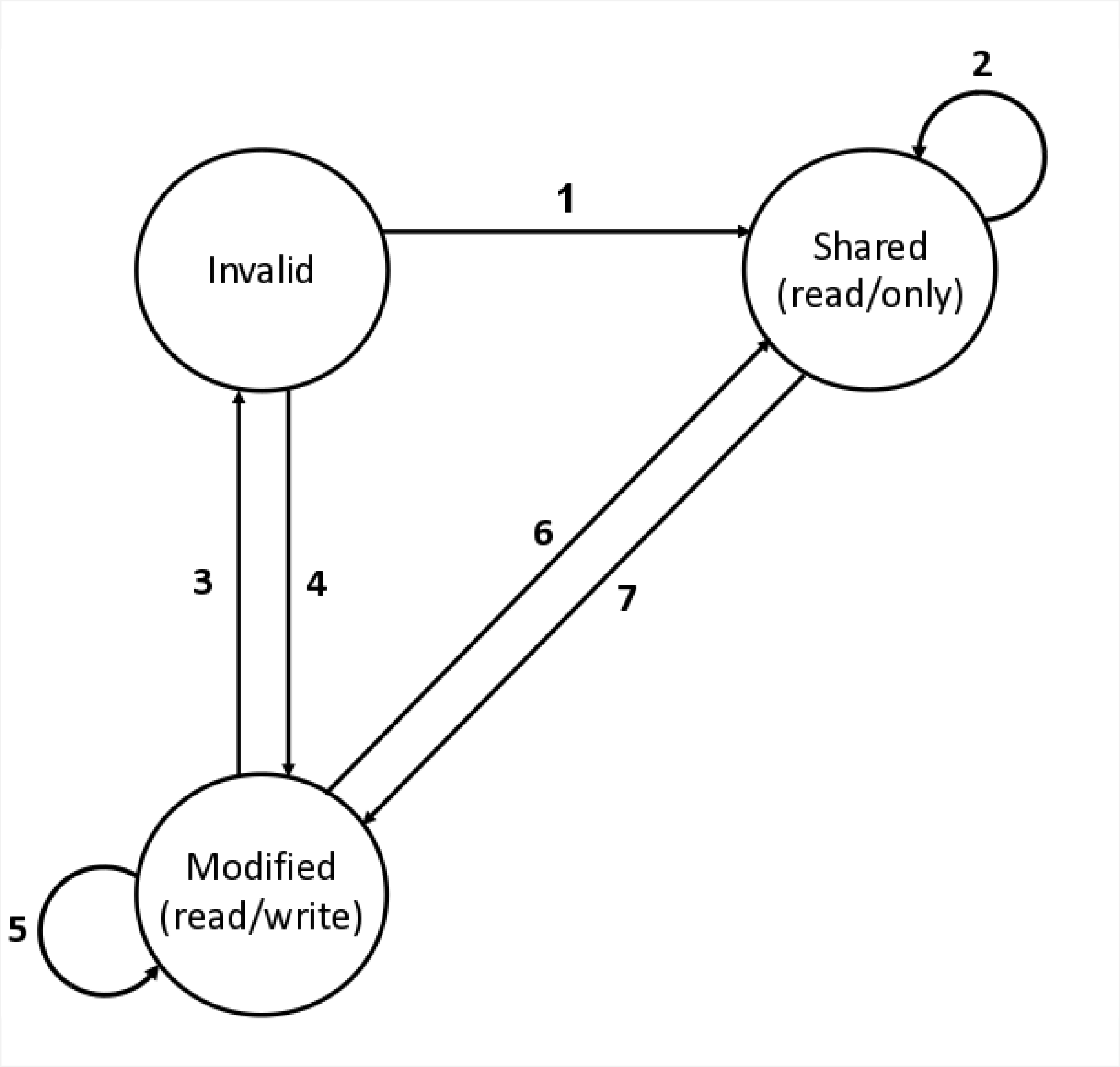
○ Snooping based scales better than directory based cache coherency.

○ Write invalidate is commonly more energy efficient than write update ✓

○ Write update is commonly used together with write-back caches.

Maximum marks: 3

23 Cache coherency FSM



The figure shows a so called MSI L1 cache coherency finite state machine (FSM).

Match each event to the correct number.

(1p for each correct match, NO minus points)

Please match the values:

	CPU write hit/miss	CPU read hit or CPU write hit/miss	CPU read miss or Bus read miss	Bus Write miss	CPU read miss	CPU read hit/miss or Bus read miss	CPU write miss
6	○	○	○ ✓	○	○	○	○
7	○ ✓	○	○	○	○	○	○
5	○	○ ✓	○	○	○	○	○
4	○	○	○	○	○	○	○ ✓
2	○	○	○	○	○	○ ✓	○
3	○	○	○	○ ✓	○	○	○
1	○	○	○	○	○ ✓	○	○

Maximum marks: 7

24 Sequential consistency

What does a sequential consistency model specify?

Select an alternative:

- ○ How write buffering is performed in a system.
- ○ That the execution of memory operations appear in the same order f ✓ all cores in a system.
- ○ The order of memory operations performed in a system.
- ○ The order of which threads are executed in a system.

Maximum marks: 3

25 Race Condition

What of the following are not solutions to avoid memory race conditions ?

Select one or more alternatives:

- □ To use memory barriers. ✓
- □ To use the sequential consistency model. ✓
- □ To use atomic memory instructions.
- □ To use locks around shared data.

Maximum marks: 3

26 Interconnect Topologies

Which of the following topologies has the smallest bisection bandwidth

Select an alternative:

- ○ Ring
- ○ Tree ✓
- ○ Crossbar
- ○ Omega

Maximum marks: 3

27 Hypercube

Which is not a correct statement regarding hypercube interconnects?
Select an alternative:

○ Maximum of O(log N) hops between two nodes

○ The router out-degree increases with the network size.

○ Good bisection bandwidth

○ Scales easily ✓

Maximum marks: 3

28 Dimension order routing

Dimension order routing is a way to
Select an alternative:

○ do flow control.

○ do cut-through routing.

○ avoid deadloocks. ✓

○ do arbitration.

Maximum marks: 3

29 Circuit switching

For which scenario is circuit switching most beneficial?

Select an alternative:

○ When a node communicates with many other nodes and a small amount of data.

○ When a node communicates with few other nodes and a small amount of data.

○ When a node communicates with many other nodes and a large amount of data.

○ When a node communicates with few other nodes and a large amou✓of data.

Maximum marks: 3

30 **Pollack\'s Rule**

Which of the following statements best describe Pollack's rule?

Select an alternative:

○ The number of transistors doubles roughly every 18 months.

○ The performance increases with the square root of the number of transistors. ✓

○ The energy increases with the square root of the number of transistors.

○ The power density is constant between technology generations.

Maximum marks: 3

31 **Mesh network**

Which is not a reason for why the Tilera designers chose a mesh network?

Select an alternative:

○ A mesh network provides ample amount of bandwidth.

○ A mesh network provides a low hop count between nodes. ✓

○ A mesh network maps easily to a 2D silicon substrate.

○ A mesh network reduces the wire length and wire congestion compared to 2D toroids.

Maximum marks: 3

32 Alpha 21164

What were the reasons for the designers of the Alpha 21164 to opt for an L2 cache?

Select one or more alternatives:

□ To reduce the average memory access latency. ✓

□ To increase the energy efficiency by sequentially accessing the L2 ca✓e.

□ To increase the total cache size to fit applications with larger working ✓ sets.

□ To enable the L1 caches to be small enough to be accessed in 1 cyc ✓

Maximum marks: 3

33 Utilization wall

The utilization wall is another expression for

Select an alternative:

○ the difficulty utilizing all the different types of cores and accelerators in a heterogeneous system.

○ the difficulty utilizing all the cores in a multicore system.

○ the difficulty utilizing the memory hierarchy efficiently.

○ dark silicon. ✓

Maximum marks: 3

34 Tomasulo

Describe the Tomasulo algorithm (with a reorder buffer) and its purpose.

The terms Issue, Execution, Write Results, and Commit might be of help when describing the algorithm.

Fill in your answer here

Maximum marks: 10

35 Amdahl's law

Describe Amdahl's law and the implications the law has on the performance of the execution of different applications on single- and multi-core architectures.

Fill in your answer here

Maximum marks: 10

36 Energy proportionality

Describe what energy proportionality is and its implication on warehouse scale computing.

Fill in your answer here

Maximum marks: 10