國立中央大學115學年度碩士班考試入學試題

系所： 資工類

科目： 作業系統與計算機組織

＊本科考試禁用計算器

多選題(每一個選項單獨計分)：每題 5 分，共 20 題，答錯每個選項倒扣 1 分
倒扣至多選題零分為止

1. Which of the following statements are TRUE regarding computer architecture?
   A. A typical MIPS arithmetic instruction can directly operate on operands located in memory.
   B. Fixed instruction length simplifies the instruction decoding logic.
   C. A RISC architecture typically has fewer general-purpose registers compared to a CISC architecture, such as x86.
   D. Using fewer addressing modes generally simplifies the datapath control logic.
   E. A RISC architecture's procedure calls usually require a stack in memory to store the return address and saved registers if the register file is insufficient.

2. Which of the following comparisons between single-cycle and multi-cycle implementations are TRUE?
   A. The clock cycle time in a single-cycle implementation is determined by the longest possible path in the logic.
   B. A multi-cycle implementation allows functional units (like the ALU) to be reused within a single instruction execution.
   C. Single-cycle implementations generally have a higher CPI than multi-cycle implementations.
   D. The control unit of a single-cycle implementation is typically purely combinational, whereas a multi-cycle implementation requires a sequential state machine.
   E. Floating-point operations are more efficiently handled in a multi-cycle datapath than a single-cycle datapath due to varying latency.

3. Consider the code sequence:

```
add x1, x2, x3
sub x4, x1, x5
```

   In a 5-stage MIPS pipeline with full forwarding logic, which of the following are TRUE?
   A. A bubble (stall) is required between the add and sub instructions.
   B. The result of the add instruction is forwarded from the EX/MEM pipeline register to the ALU input for the sub instruction.

注意:背面有試題

C. If forwarding were not implemented, the pipeline would need to stall for 2 cycles to avoid the hazard (assuming register write in the first half, read in the second half).
D. Forwarding eliminates all data hazards, including those caused by Load-Use dependencies.
E. The hazard detected here is a RAW (Read After Write) hazard.

4. Which of the following statements about branch prediction are TRUE?
A. Static branch prediction assumes a branch is always taken or always not taken based on the instruction type or direction.
B. A 1-bit dynamic branch predictor will always mispredict a loop exit branch if the loop executes many times.
C. A 2-bit saturating counter predictor changes its prediction immediately after a single misprediction.
D. Branch Target Buffers store the target address of branches to allow fetching from the target in the IF stage.
E. Delayed branching is a hardware technique where the compiler fills the slot after a branch with an independent instruction.

5. Regarding exceptions in a pipelined processor:
A. "Precise exceptions" imply that the processor state can be recovered exactly as it was before the faulting instruction executed.
B. When an exception occurs, the pipeline must flush instructions that occurred after the faulting instruction.
C. Handling exceptions is easier in a pipelined processor than in a single-cycle processor.
D. The EPC (Exception Program Counter) stores the address of the instruction that caused the exception.
E. External interrupts (like I/O requests) are synchronous events tied to the CPU clock.

系所： 資工類

科目： 作業系統與計算機組織

＊本科考試禁用計算器

6. Regarding Virtual Memory and Translation Lookaside Buffers (TLBs) in a modern hierarchical memory system, which statements are correct?
   A. A TLB miss implies a Page Fault must strictly occur.
   B. Huge Pages (e.g., 2MB or 1GB pages) generally reduce TLB miss rates by increasing TLB reach.
   C. In a virtualized cloud environment using Extended Page Tables (EPT) or Nested Paging, a TLB miss requires a two-dimensional page walk, increasing latency compared to native execution.
   D. Virtually Indexed, Physically Tagged (VIPT) L1 caches allow the cache lookup to proceed in parallel with the TLB translation.
   E. None of the above.

7. Consider a Multi-Level Cache (MLC) hierarchy (L1, L2, L3). Which design choices regarding inclusion and writing policies are correct?
   A. Inclusive Caches (e.g., L3 includes all L1/L2 contents) simplify cache coherence because the snoop filter only needs to check the L3 tags to know if a line exists in the core.
   B. Exclusive Caches (e.g., L3 contains only victims evicted from L2) maximize the total effective cache capacity of the hierarchy.
   C. A Virtually Indexed, Physically Tagged (VIPT) L1 cache allows the TLB lookup and the cache set index decoding to proceed in parallel, reducing the critical path latency.
   D. Write-Allocate is typically paired with Write-Through caches to ensure the line is brought into the cache on a write miss.
   E. In a Non-Uniform Cache Architecture (NUCA) L3, the latency to access a specific L3 bank depends on the physical distance between the requesting core and that bank on the die.

8. In Modern GPU Architectures (SIMT - Single Instruction, Multiple Threads), how are memory divergence and control divergence handled?
   A. If threads within a single Warp take different paths of an `if-else` branch, the hardware serializes the execution, executing both paths for all threads but masking off writes for inactive threads.

B. Memory Coalescing occurs when threads in a Warp access contiguous memory addresses, allowing the GPU memory controller to combine them into a single high-bandwidth transaction.

C. GPUs rely on massive L3 caches (hundreds of MBs) rather than thread-level parallelism to hide memory latency.

D. A Shared Memory (or Scratchpad) bank conflict occurs when multiple threads in a Warp try to access different addresses that map to the same physical memory bank in the same cycle.

E. Unlike CPUs, GPUs strictly avoid context switching and run a single kernel to completion before starting another.

9. Which of the following statements correctly describe the principles of Locality of Reference?

A. Temporal Locality implies that if a memory location is referenced, it is likely to be referenced again soon (e.g., variables in a loop).

B. Spatial Locality implies that if a memory location is referenced, nearby addresses are likely to be referenced soon (e.g., traversing an array).

C. Caches rely on Spatial Locality by loading data in "Blocks" or "Lines" rather than single bytes.

D. Temporal Locality is primarily exploited by the Prefetcher, not the Cache.

E. Instruction caches generally exhibit high Spatial Locality because programs usually execute instructions sequentially.

10. Modern GPUs often use GDDR (Graphics Double Data Rate) or HBM. How does the memory subsystem of a GPU differ from a CPU's DDR subsystem regarding latency and parallelism?

A. CPU memory controllers are optimized for Low Latency to ensure quick response times for serial tasks and OS interrupts.

B. GPU memory controllers are optimized for Throughput and are designed to tolerate high latency by switching between thousands of active threads (latency hiding).

C. Memory Coalescing: GPU hardware attempts to combine memory accesses from adjacent threads into a single transaction to maximize bus utilization, a feature less critical in standard CPU scalar execution.

D. CPUs typically use massive L1/L2/L3 caches to reduce the effective memory latency, whereas GPUs devote more die area to ALUs and use smaller caches relative to their compute throughput.

E. GDDR memory is pin-compatible with standard DDR memory, meaning you can plug GDDR6 chips directly into a standard PC motherboard DDR4 slot.

11. Which of the following statements accurately describe the system startup and boot process?

A. The BIOS/UEFI resides in volatile RAM and is the first program executed when the computer is powered on.

B. The Bootloader (e.g., GRUB) is typically responsible for loading the Operating System kernel from the disk into the main memory.

C. UEFI (Unified Extensible Firmware Interface) supports "Secure Boot," which checks the cryptographic signature of the bootloader and kernel to prevent malware injection.

D. The kernel initialization phase includes probing hardware, mounting the root file system, and starting the first user-space process (e.g., init or systemd).

E. The Master Boot Record (MBR) partition scheme is required for drives larger than 4 TB to function as a boot drive.

12. Which of the following are valid Process States and transitions in a standard 5-state process model?

A. Running ➔ Ready: This transition happens when a process is preempted by the scheduler (e.g., time quantum expires).

B. Blocked (Waiting) ➔ Running: A process moves directly to Running when its I/O operation completes.

C. Ready ➔ Running: This transition is managed by the Dispatcher.

D. Running ➔ Blocked (Waiting): This transition occurs when a process requests an I/O operation or waits for an event.

E. Terminated ➔ Ready: A zombie process can be restarted by the parent process.

13. What specific OS features are critical for high-performance AI/Machine Learning workloads?
    A. Huge Pages / Large Pages support reduces TLB miss rates when training models with massive datasets.
    B. NUMA (Non-Uniform Memory Access) awareness allows the OS to schedule threads on CPU cores that are physically closer to the memory bank holding the training data.
    C. GPUDirect (or similar RDMA technologies) enables network adapters to move data directly to/from GPU memory, bypassing the CPU and system RAM.
    D. AI Workloads typically prefer Fine-Grained Context Switching (frequent switching) to maximize CPU responsiveness during training.
    E. Containerization (e.g., Docker/Kubernetes) is preferred over full Virtual Machines for AI deployment due to lower overhead and direct access to hardware drivers.

14. How does OS design adapt for Solid State Drives (SSDs) compared to HDDs?
    A. SSDs perform "in-place updates" just like HDDs, allowing data to be overwritten physically in the same location without erasure.
    B. The Flash Translation Layer (FTL) maps logical block addresses (LBA) to physical pages, hiding the complexity of erase blocks from the OS.
    C. The TRIM command allows the OS to inform the SSD which data blocks are no longer considered in use (e.g., after file deletion), aiding internal Garbage Collection.
    D. Write Amplification is a phenomenon where the actual amount of data written to the flash chips is greater than the amount of data written by the host OS.
    E. Wear Leveling algorithms are implemented by the OS kernel to ensure all SSD cells effectively have the same lifespan.

15. Which statements accurately describe I/O Subsystem mechanisms?
    A. DMA (Direct Memory Access) allows an I/O device to transfer data directly to/from memory without continuous CPU intervention.
    B. Interrupt-driven I/O requires the CPU to constantly poll the device status register to check if data is ready.
    C. Blocking I/O suspends the execution of the calling process until the I/O operation is complete.
    D. Double Buffering allows the producer (device) and consumer (CPU) to operate in parallel, smoothing out speed mismatches.
    E. The Kernel I/O Subsystem manages issues like caching, spooling, and device reservation.

16. Choose the correct statements from the multiple choices
    A. In a multithreaded process using kernel-level threads, a blocking I/O operation in one thread causes the entire process to block.
    B. During a context switch between two threads belonging to the same process, the operating system must update the CPU's page table base register.
    C. When two processes share the same physical memory page, they use the same virtual address to refer to that page.
    D. If two processes share a physical memory page, any modification to that page is immediately visible to both processes.
    E. None of the above.

17. Choose the correct statements from the multiple choices
    A. An attack that compromises data confidentiality does not necessarily affect data integrity.
    B. If Alice encrypts a document using her private key in an asymmetric cryptographic system, the operation provides authentication rather than confidentiality.
    C. Transport-layer encryption protects data from eavesdropping and also prevents traffic analysis.
    D. Side-channel attacks are fundamentally designed to exploit mathematical weaknesses or logical flaws within a cryptographic algorithm's design.
    E. None of the above.

系所： 資工類

科目： 作業系統與計算機組織

＊本科考試禁用計算器

18. Choose the correct statements from the multiple choices
    A. When the OS switches to a process with a different address space, the existing TLB entries are typically no longer valid and therefore need to be flushed or invalidated.
    B. A TLB miss is a sufficient condition to conclude that the referenced page is not resident in physical memory.
    C. Reducing the frequency of TLB misses can significantly improve system performance.
    D. The TLB indexes translations by virtual address, so different virtual addresses mapping to the same physical page require separate TLB entries.
    E. None of the above.

19. Choose the correct statements from the multiple choices
    A. Spinlocks are generally more efficient than binary semaphores because they avoid putting threads to sleep.
    B. In an interrupt handler, a spinlock is usually a better synchronization choice than a semaphore because interrupt handlers are not allowed to sleep.
    C. Disabling interrupts while holding a spinlock can prevent deadlock between a thread and an interrupt handler.
    D. The optimal maximum spin time for a spinlock is independent of the time required to perform a context switch.
    E. None of the above.

20. Choose the correct statements from the multiple choices
    A. GPU internal thread scheduling is physically managed and allocated directly by the OS kernel scheduler.
    B. When a GPU reads system memory via DMA, the CPU is not directly involved in the data movement during the transfer.
    C. The OS interrupt handling mechanism can be used to notify the CPU upon completion of a GPU computation.
    D. The OS, through its GPU drivers, enforces access control and resource-sharing policies that determine which processes are permitted to submit work to the GPU.
    E. None of the above.