Final exam preparation questions.

DRAM refresh overhead usually decreases with an increase in DRAM device density.
: False, density increase, refresh increase

Parallel tag and data access is commonly used in L2 and L3 caches.
: False,

- Parallel use in L1.
- Serial use in L2, L3. Check Tag first then only fetch data when hit.

The relationship between a TLB and page table is the same as the relationship between a cache and main memory.
: True,
Page table = Main memory
TLB(Translation Lookaside Buffer) = Cache of Page table

Which of the following techniques can be used to reduce the cache hit time?
i. Way prediction
ii. Increase the associativity of the cache
iii. Increase the size of cache
iv. Use multiple cache banks

: Way prediction and Use multiple cache banks.

- Way prediction: predicts which "way" (among the multiple cache lines in a set)
- Use multiple cache banks: parallel operations
- Increasing cache size or associativity can improve the hit rate but typically increases hit time due to the added complexity.

Why Google Drive suck?
: Google Drive is designed as a scalable distributed storage system rather than a traditional hierarchical file system.
1. ไม่สามารถดูได้ว่าใน folder มีทั้งหมดกี่ files
Metadata is stored separately from file data.
2. Can't move file by drag and drop
Google Drive operates over HTTP(S) protocols, which are not natively designed for low-latency, real-time file manipulation.
3. Can't move file by path
virtualized directory structure, and their actual locations are abstracted.
4. การ download file ใหญ่ ที่มีจำนวนมาก ต้องลุ้นว่าไฟล์จะมาครบหรือไม่
stores files across multiple servers

Distributed Storage, Virtualization and Abstraction, Scalability Over Performance, Network Dependency

Concept of Overlapped Cache & TLB Access.
: Parallel (overlap) the cache look up (index) and translation from virtual address to physical address at the same time.
Possible when index is not changed in address translation

Design storage(nested RAID e.g. RAID05) for spacecraft, that need a lot of capacity, no need to maintenance frequently.
: RAID50 because RAID 0's large capacity and RAID 5's fault tolerance.

Calculate tag, set, line, hit, and miss of cache.

Concept of Write buffer.
: With write through, a write access is bad for performance. To make write faster, use write buffer.
Because writing direct to main memory is slow. Prevents the CPU from stalling
Write Buffer Saturation(overflowed/fills up), a naive solution is to introduce a second level cache (L2). CPU -> Buffer -> L2 -> Mem.

GPU vs CPU.
:
CPU:
- general-purpose computing
- multitasking
- 4–16 cores
- Optimized for low latency.
- Traditional & Simultaneous Multithreading
- 5-stage Pipeline: FDEMW
  Many things are added to improve single thread performance (Fetch&Decode).
- Memory Heirarchy: CPU->L1->L2->L3->Mem
- OS, App, Background

GPU:
- SIMD/SIMT
- parallel processing
- High Throughput
- specific workloads e.g. repetitive mathematical computations
- Many type cores
  - GPU Cores (SM-streaming multiprocessor)

        - CUDA Core = 1 execution unit
- Fine-grain Multithreading
- 5-stage Pipeline
    - Execute: Many ALUs
    - Writeback: Many register file for each lane.
- Memory Heirarchy: SM->Shared Memory->Scheduler->DRAM
- Memory management
    - Normal: MEM<->CPU<->GPU
    - Unified Virtual Memory (same address space): CPU<->MEM<->GPU
- High-Bandwidth
- graphics, ML

GPU Memory utilization calculation.

Concept of TLB of virtual memory.
: A cache for address translations(virtual addr. to physical addr.).
Special cache of recently used page table entries.
Fully associative because TLB is small and miss penalty is very high. (Mostly, but if you poor will use small n-way set assoc.)

Relation between Block size and Access Time, and Miss&Hit Rate, Time.
: If Block Size (+), then
Access Time (+): complex logic and data transfer.
Hit Rate (+): better spatial locality.
Miss Rate (-): reduces cold misses but may increase conflict misses.
Miss Penalty (+): increases due to longer memory transfer times.

Tradeoff between Access Time and Miss Penalty
: Small, fast caches improve access time but suffer from higher miss penalties, while larger caches reduce miss penalties at the cost of slower access.
Compromise solution is Multi-level cache.

Software Pipeline, Loop Unroll.
Performance of Instruction-Level Parallelism.