39 Lecture - CS501

Important Subjective

1. **What is a cache? How does it improve computer performance?**
   Answer: A cache is a small, high-speed memory that stores frequently accessed data to reduce the number of times the CPU has to access the slower main memory. It improves computer performance by providing faster access to data, reducing the average memory access time.

   What is the difference between a direct-mapped cache and an associative cache?
   Answer: In a direct-mapped cache, each memory location can only be stored in one specific location in the cache. In an associative cache, each memory location can be stored in any location in the cache.

   What is cache coherence? How is it maintained?
   Answer: Cache coherence is the property that ensures that all copies of a memory location in different caches have the same value. It is maintained through a protocol such as MESI (Modified-Exclusive-Shared-Invalid) that controls how cache copies are updated and invalidated.

   What is a cache hit? What is a cache miss?
   Answer: A cache hit occurs when the CPU requests data that is already stored in the cache. A cache miss occurs when the CPU requests data that is not stored in the cache and must be retrieved from main memory.

   What is the principle of locality? How does it relate to the cache?
   Answer: The principle of locality states that memory accesses tend to cluster around a small set of memory locations. This principle is important for the cache because it allows the cache to store the most frequently accessed data, reducing the number of cache misses.

   What is a write-back cache? How does it differ from a write-through cache?
   Answer: A write-back cache only writes data to main memory when it is evicted from the cache. In contrast, a write-through cache immediately writes data to main memory. Write-back caches can be more efficient because they reduce the number of main memory writes.

   What is a cache line? How is it related to cache performance?
   Answer: A cache line is the smallest unit of data that can be stored in the cache. The size of the cache line can affect the cache performance because larger cache lines can reduce the number of cache misses, but smaller cache lines can reduce the cache access time.

   What is the difference between a level 1 (L1) cache and a level 2 (L2) cache?
   Answer: An L1 cache is a small, fast cache that is built into the CPU. An L2 cache is a larger, slower cache that is located outside the CPU, typically on the motherboard or in a separate chip.

   What is cache bypassing? When is it useful?
   Answer: Cache bypassing is the process of skipping the cache and accessing main memory

directly. It can be useful in certain situations where the cache may be slowing down memory accesses, such as when accessing large, contiguous blocks of memory.

What is cache thrashing? How can it be prevented?
Answer: Cache thrashing occurs when the cache is repeatedly filled with data that is immediately evicted, causing a high number of cache misses. It can be prevented by increasing the size of the cache, increasing the cache line size, or optimizing the program to reduce unnecessary memory accesses.