Cache Memory Questions

1. **Cache Replacement Policies:** When a cache miss occurs in a set-associative or fully associative cache and the target set (or the entire cache, for fully associative) is full, an existing block must be evicted to make space for the new block.
    - Describe the operational difference between the **Least Recently Used (LRU)** and **First-In, First-Out (FIFO)** replacement policies.
    - What are the potential advantages and disadvantages of using LRU compared to FIFO?
2. **Cache Block Size:** The size of the data block (or cache line) transferred between main memory and the cache is a critical design parameter.
    - Discuss the trade-offs involved in selecting a cache block size.
    - What are the potential performance benefits and drawbacks associated with using a larger block size versus a smaller block size?

Cache Memory Exercises

1. **Set-Associative Cache Address Decomposition (Scenario 1):**
   Suppose a computer system uses a set-associative cache. The main memory contains 2^{16} words. The cache has a total capacity of 32 blocks, and each cache block contains 8 words.

    a. If this cache is **2-way set-associative**, determine the format of a memory address as seen by the cache. Specify the number of bits required for the **tag**, **set index**, and **word offset** fields.
    b. If this cache is **4-way set-associative**, determine the format of a memory address as seen by the cache, again specifying the sizes of the **tag**, **set index**, and **word offset** fields.
2. **Set-Associative Cache Address Decomposition (Scenario 2):**
   Consider a system with a 32-bit memory address space. It features an 8-way set-associative cache with a total size of 128 KB and a cache block size of 64 Bytes.

a. Calculate the total number of blocks in the cache.
b. Calculate the total number of sets in the cache.
c. Show the format of the 32-bit memory address by specifying the number of bits for the **tag**, **set index**, and **byte offset** fields.

3. **Effective Access Time (EAT):**

A processor accesses memory through a Level 1 (L1) cache. The L1 cache has an access time (hit time) of 2 ns and achieves a hit rate of 95%. Accessing main memory directly takes 100 ns (this is the miss penalty if only L1 cache exists). Calculate the Effective Access Time (EAT) for this single-level cache system.

4. A direct-mapped cache consists of eight blocks. Main memory contains 4K blocks of eight words each. Access time for the cache is 22 ns and the time required to fill a cache slot from main memory is 300ns (this time will allow us to determine the block is missing and bring it into cache). Assume a request is always started in parallel to both cache and to main memory (so if it is not found in cache, we do not have to add this cache search time to
the memory access). If a block is missing from cache, the entire block is brought into the cache and the access is restarted. Initially, the cache is empty.
a. Show the main memory address format that allows us to map addresses from main memory to cache. Be sure to include the fields as well as their sizes.
b. Compute the hit ratio for a program that loops 4 times from locations 0 to (67)10 in memory.
c. Compute the effective access time for this program.
5. Consider a byte-addressable computer with 24-bit addresses, a cache capable of storing a total of 64K bytes of data and blocks of 32 bytes. Show the format of a 24-bit memory address for:
a. direct mapped
b. associative
c. 4-way set associative