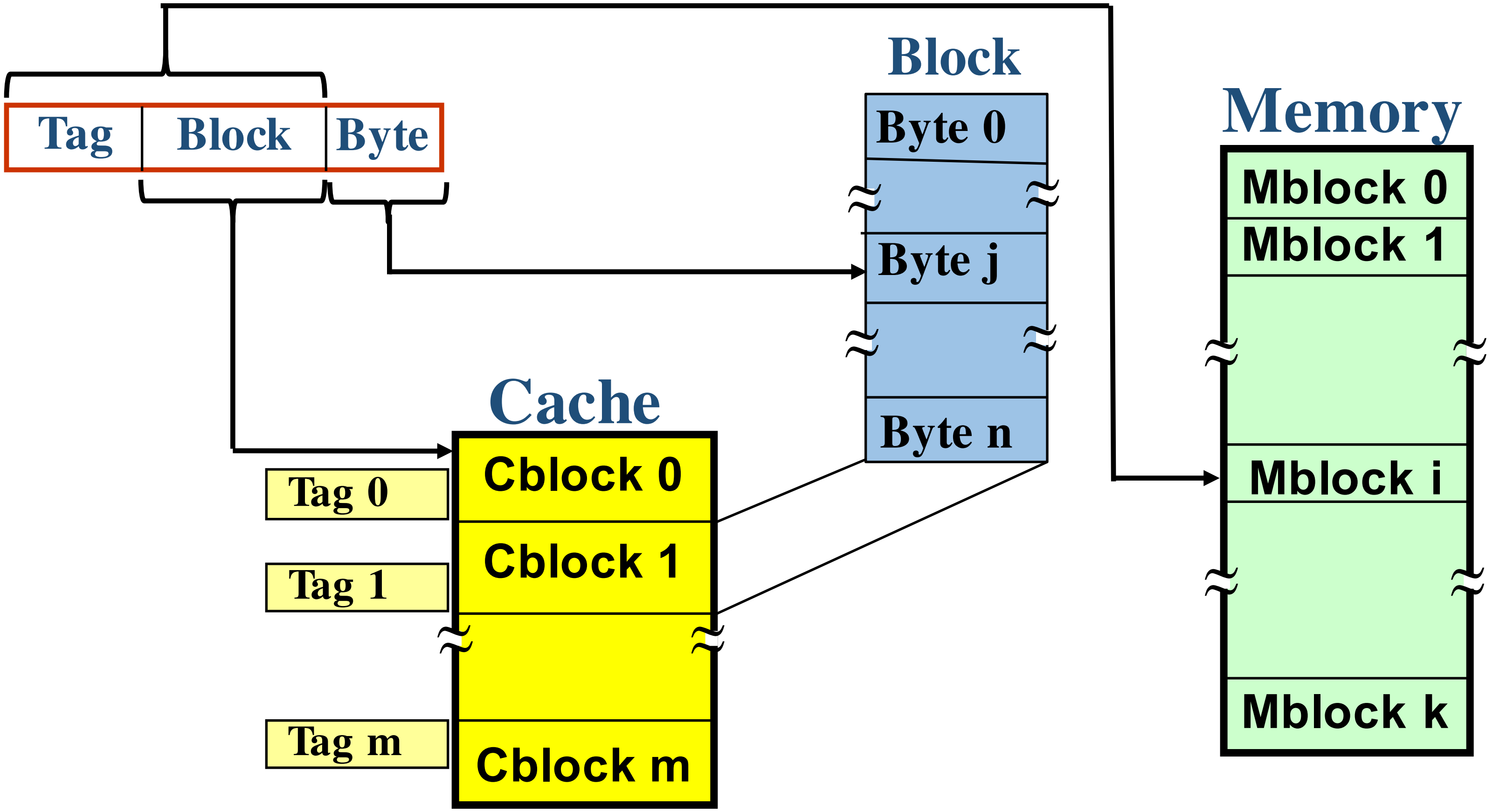
Caches: Memory Coherence and Consistency



On this Lesson

- Difference between memory coherence and consistency

- Mechanisms for dealing with memory coherence and consistency
  - *Snooping*
  - *Directory-based*

Coherence vs. Consistency

- In a mutiltiprocessing system the main memory is shared by many processors with local caches.
- At a given time a memory block could be allocated in the local cache of several processors.
- When any change in a shared memory block is updated in each of the copies on the local cache of the sharing processors, the memory system is said to be **coherent**.
- When every change in a shared memory block is updated in each of the copies on the local cache of the sharing processors, **in the same order**, the memory system is said to be **consistent**.

Source of Memory Inconsistency

- In addition to the accesses generate by the CPU, the main memory of a microprocessor is also accessed by external devices through a mechanism known as Direct Memory Access (DMA).
- Due to this DMA mechanisms the data in main memory is not always consistent with corresponding data in the cache.
- This problem arises under two circumstances:
  - *An external device writes a memory block that has a copy in the cache, but the cache copy is not updated*
  - *An external device reads a memory block that has a copy in the cache that was written by the CPU, but not updated in main memory.*
- The primary memory or the cache needs to be refreshed when a data inconsistency arises

Old Mechanisms for Memory Consistency

- Write-back
  - *Memory is refreshed with a victim block that has been written in the cache*

- Write-through
  - *Memory is refreshed every time a cache block is written*

Snooping for Memory Coherence Mechanism

- Used for uniprocessing systems as well as multiprocessors systems

- The cache is notified when any external device attempts to access a block of primary memory that has a copy in the cache.

- The action taken depends on the type of access (read/write) and the type of snooping mechanism.

- There are two types of snooping mechanisms
  - *Write- invalidate*
  - *Write-update*

Snooping Write-Invalidate

- Write Access:
  - *The external device writes on a primary memory block*
  - *Its corresponding block in the cache is invalidated*

- Read Access:
  - *An external device attempts to read a block in primary memory*
  - *Its corresponding block in the cache is written back into primary memory if it was written in the cache*
  - *Access to the external device is then granted*

Snooping Write-Update

- Write Access:
  - *The external device writes a block of primary memory*
  - *The corresponding block of the cache is replaced with the new block placed by the external device in primary memory*
- Read Access:
  - *An external device attempts to read a block in primary memory*
  - *Its corresponding block in the cache is written back into primary memory if it was written in the cache*
  - *Access to the external device is then granted*

Directory-Based Memory Coherence Mechanism

- Is the mechanism mostly used in large shared-memory multiprocessing systems.
- Coherence is maintain through a directory of shared memory blocks.
- Local caches must ask permission to the directory to access blocks from main memory.
- When a shared memory block is changed in main memory, the directory either invalidates or updates the copies in local caches.

Snooping vs. Directory-Based Coherence

- Snooping
  - *Faster*
  - *Does not scale well because memory accesses are broadcast to all local caches*
- Directory-Based
  - *Slower*
  - *Scales better because memory accesses are only sent to the local caches sharing the memory block*

Lesson Outcomes

- Understand the difference between memory coherence and consistency
- Understand the snooping and directory-based memory coherence mechanisms