Cache Coherence

- Cache coherence ensures that multiple processors have a consistent view of memory despite having private caches. The problem arises because each processor's view of memory is through its own cache.

Time	Event	Cache contents for processor A on X	Cache contents for processor B on X	Memory contents on X
0	Processor A write 1 to X	1		1
1	Processor A reads 1 from X	1		1
2	Processor B reads 1 from X	1	1	1
3	Processor B writes 0 to X	1	0	0
4	Processor A writes 1 to X	1	0	1

Cache Coherence Requirements

- Read-after-Write Consistency
  - A read by a processor to a location it previously wrote must return the value it wrote, assuming no other processor wrote to that location in between.
- Write Propagation
  - A read to a location that follows a write by another processor must return the written value if sufficient time has passed and no other writes occurred.
- Write Serialization
  - Writes to the same location must be seen in the same order by all processors. This prevents processors from seeing different sequences of values.

Coherence vs. Consistency

Coherence

Defines what values can be returned by a read operation to a specific memory location.

- Focuses on a single memory location
- Ensures all processors see the same sequence of values
- Maintains the illusion of a single memory location despite multiple copies

Consistency

Determines when a written value will be visible to other processors.

- Concerns relationships between different memory locations
- Defines ordering rules for memory operations
- Affects how programmers reason about parallel code

Both properties are essential for writing correct shared-memory programs. Coherence is necessary but not sufficient for parallel correctness.

Cache Coherence Protocol

- Directory based
  - The sharing status of a particular block of physical memory is kept in one location, called the directory
  - Two types
    - One centralized directory in SMP (Symmetric Multiprocessors)
    - Distributed directories in DSM (Distributed Shared Memory)
- Snoop based
  - Every cache that has a copy of the data from a block of physical memory could track the sharing status of the block