

CS 636: A Brief Introduction to Cache Coherence

Swarnendu Biswas

Department of Computer Science and Engineering,
Indian Institute of Technology Kanpur

Sem 2025-26-II



Types of Parallelism

Instruction-level Parallelism

Overlap instructions within a single thread of execution (e.g., pipelining, superscalar issue, and out-of-order execution)

Data-level Parallelism

Execute an instruction in parallel on multiple data values (e.g., vector instructions)

```
for (int i = 0; i < N; i++) {  
    c[i] = a[i] + b[i];  
}
```

Thread-level Parallelism

Concurrently execute multiple threads

Shared Memory Multiprocessor Architecture

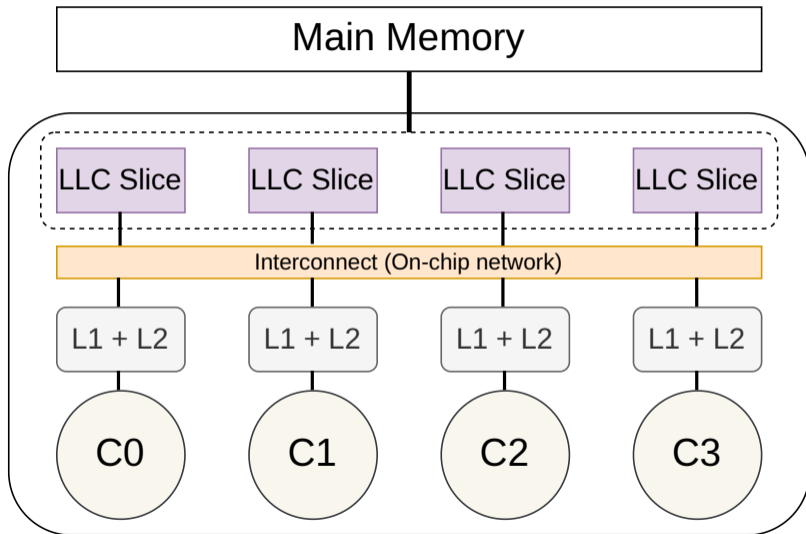
Single address space shared by multiple cores

- + Exploits TLP by having a number of cores
- + Can share data efficiently, communication is implicit through memory instructions (i.e., loads and stores)
- Cost for accessing shared memory can be uniform or non-uniform across cores

Processors privately cache data to improve performance

Reduces average data access time and saves interconnect bandwidth

Block Diagram of a SMP

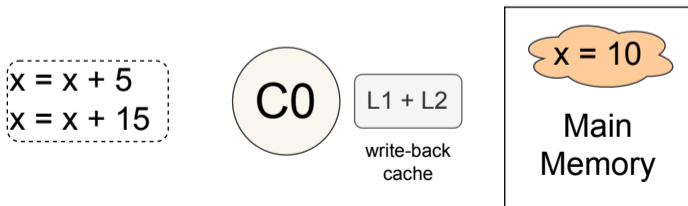


Data Coherence

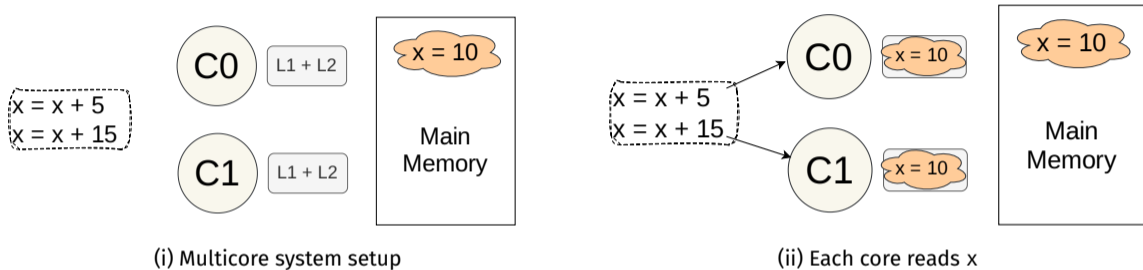
⚠ Private caches create data coherence problem

- Copies of a variable can be present in multiple caches
- Private copies of shared data must be **coherent**, i.e., all copies must have the same value (okay if the requirement holds eventually)

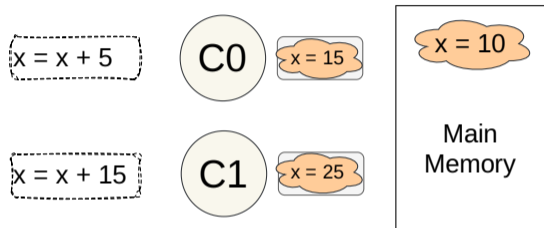
Consider the following sequence of operations on a single core system with write-back caches but without coherence



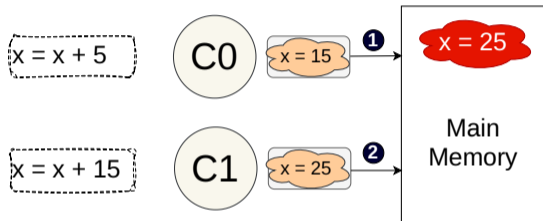
Coherence Challenge with Multicores



Coherence Challenge with Multicores



(iii) Each core updates x in its private cache



(iv) Cores write back x , a store is lost depending on the order of write backs

Can Write-through Caches Avoid the Coherence Problem?

Assume 3 cores with write-through caches

- (i) Core C₀ reads x from memory, caches it, and gets the value 10
- (ii) Core C₁ reads x from memory, caches it, and gets the value 10
- (iii) C₁ writes x=20, and updates its cached and memory values
- (iv) C₀ reads x from its cache and gets the value 10
- (v) C₂ reads x from memory, caches it, and gets the value 20
- (vi) C₂ writes x=30, and updates its cached and memory value

Sources of Errors in the Previous Setups

✘ Write-back cache

- Stores are not visible to memory immediately
- Order of write backs are important
- **Lesson learned:** do not allow more than one copy of a cache line in dirty state

⚠ Write-through cache

- The value in memory may be correct if the writes are correctly ordered
- A store proceeded when there is already a cached copy
- **Lesson learned:** must invalidate all cached copies before allowing a store to proceed

A memory system is coherent if the following hold:

- (i) A read from a location X by a core C after a write by C to X always returns the value written by C , provided there are no writes of X by another processor between the two accesses by C .
- (ii) A read from a location X by a core C after a write to X by another core returns the written value if the read and write are sufficiently separated in time and no other writes to X occur between the two accesses.
- (iii) Writes to the same location are serialized. That is, two writes to the same location by any two cores are seen in the same order by all processors.

Correctness Requirement

✓ For sequential programs, there is only one correct output

A read from a memory location must return the “latest” value written to it

⚠ For parallel programs, there can be multiple correct outputs

- Defining “latest” precisely is crucial
- Assume that the latest value of a location is the latest value “committed” by any thread/process

Cache Coherence Protocol

Multicore processors implement a cache coherence protocol to keep private caches in sync

💡 A “cache coherence protocol” is a set of actions that ensure that a load to address A returns the “last committed” value to A

- Essentially, makes one core’s write to A visible to other cores by propagating the write to other caches
- Aims to make the presence of private caches functionally invisible
- Coherence protocols usually operate at the granularity of whole cache blocks (e.g., 64 bytes)

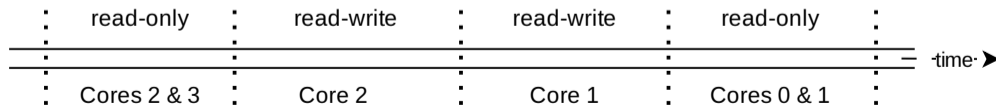
Cache Coherence Protocol Invariants

1. Enforces the Single-Writer-Multiple-Reader (SWMR) invariant

For any given memory location, at any given moment in time, there is either a single core that may write it (including read) or some number of cores that may read it

2. Data values must be propagated correctly (data invariant)

The value of a memory location at the start of a read-only time period is the same as the value of the location at the end of its last read-write time period



Alternate Definitions of Coherence

Definition 2

A coherent system must appear to execute all threads' loads and stores to a **single** memory location in a total order that respects the program order of each thread

Definition 3

A coherent system satisfies two invariants:

write propagation every store is eventually made visible to all cores

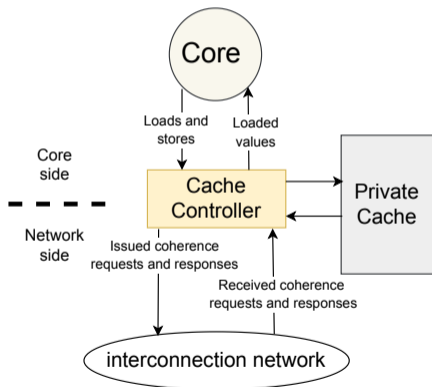
write serialization writes to the same memory location are serialized (i.e., observed in the same order by all cores)

Cache Coherence Protocols

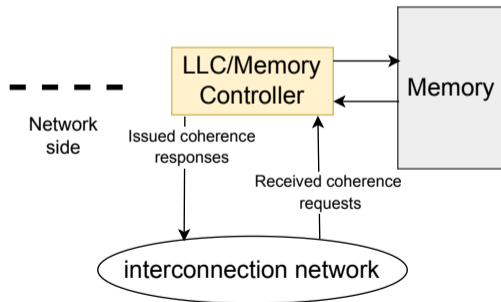
Implementing Coherence Protocols

Protocols are implemented as finite state machines called coherence controllers

A protocol formalizes the interactions between the different coherence controllers



Cache controller



Memory controller

Important Characteristics of a Cache Block

Coherence protocols are implemented by associating states with each cache block

Validity A valid block has the most up-to-date values for the block. The block may be read. It can be written if it is also exclusive.

Dirtyness A cache block is dirty if its value is the most up-to-date, and the value differs from the value in the LLC/memory.

Exclusivity A cache block is exclusive if it is the only privately cached copy of that block except perhaps in the shared LLC.

Ownership A cache or memory controller is the owner of a block if it is responsible for responding to coherence requests for that block. In most protocols, there is exactly one owner of a given block at all times.

Stable States

M, S, and I are commonly-used states

Modified (M) The block is valid, exclusive, owned, and potentially dirty. The cache has the only valid copy of the block, the cache must respond to requests for the block, and the copy of the block at the LLC/memory is potentially stale.

Shared (S) The block is valid but not exclusive, not dirty, and not owned. The cache has a read-only copy of the block. There may be multiple processors caching a line in S state.

Invalid (I) The cache either does not contain the block (not present) or it contains a potentially stale copy that it may not read or write.

Different protocol extensions add additional states (e.g., E, O, and F) to optimize for certain sharing patterns

Common Coherence Transactions

Transaction	Goal of Requestor
GetS	Obtain block in Shared (read-only) state
GetM	Obtain block in Modified (read-write) state
Upg	Upgrade block state from read-only (Shared or Owned) to read-write (Modified); Upg (unlike GetM) does not require data to be sent to requestor
PutS	Evict block in Shared state
PutE	Evict block in Exclusive state
PutO	Evict block in Owned state
PutM	Evict block in Modified state

Communication between Core and Cache Controller

Event	Response from Cache Controller
Load	If cache hit, respond with data from cache; else initiate GetS transaction
Store	If cache hit in state E or M, write data into cache; else initiate GetM or Upg transaction
Atomic RMW	If cache hit in state E or M, atomically execute RMW semantics; else initiate GetM or Upg transaction
Instruction fetch	If I-cache hit, respond with instruction from cache; else initiate GetS transaction
Read-only prefetch	If cache hit, ignore; else (optionally) initiate GetS transaction
Read-write prefetch	If cache hit in state M, ignore; else (optionally) initiate GetM or Upg transaction
Replacement	Depending on state of block, initiate PutS, PutE, PutO, or PutM transaction

Types of Coherence Protocols

💡 Protocols differ in when and how writes are propagated

- The writes can be propagated synchronously or asynchronously
- Synchronous propagation means a write is made visible to other cores *before* returning

i Two main axes to classify synchronous protocols

- (i) Invalidation-based protocol and Update-based protocol
- (ii) Snoopy protocol and Directory protocol

Invalidation-Based Protocols

Invalidate all cached copies before allowing a store to proceed

Need to know the location of cached copies

Solution 1 : Broadcast that a core is going to do a store and sharers invalidate themselves

Solution 2 : Keep track of the sharers and invalidate them when needed

- + Only store misses go on bus and subsequent stores to the same line are cache hits
- Sharers will miss next time they try to access the line

Update all cached copies with the new value of the store

- + Sharers continue to hit in the cache, do not need to initiate and wait for a GetS transaction to complete
- Prevalence of spatial and temporal locality can lead to unnecessary updates, leading to increased bandwidth requirements
- Complicates implementing many memory consistency models

Snoopy Protocol

Cache controller initiates a request for a block by broadcasting a request message to all other coherence controllers

- Each cache controller snoops (i.e., continuously monitors) the shared medium (e.g., bus or switch) for write activity concerned with its cached data addresses
- Assumes a global bus structure where communication can be seen by all
- Relies on the interconnection network to deliver the broadcast messages in a consistent order to all cores

How do you prevent simultaneous writes from different controllers?

Snoopy Protocol

Invalidate on a write

- Core that wants to write to an address grabs a bus cycle and broadcasts a “write invalidate” message
- All snooping caches invalidate their private copy of the appropriate cache line
- Core writes to its cached copy
- Any future read in other cores will now miss in cache and refetch new data

Update on a write

- Core that wants to write to an address grabs a bus cycle and broadcasts new data as it updates its own copy
- All snooping caches update their copy

Cache controller initiates a request for a block by unicasting it to the block's home memory controller

- Memory controller maintains a directory that holds state about each block in the LLC/memory (e.g., coherence state, the current owner ID, and a bitvector for the list of current sharers)
- If the LLC/memory is the owner, the memory controller completes the transaction by sending a data response to the requestor
- If a cache controller is the owner, the memory controller forwards the request to the owner cache
- When the owner cache receives the forwarded request, it completes the transaction by sending a data response to the requestor

Snoopy vs Directory Protocol

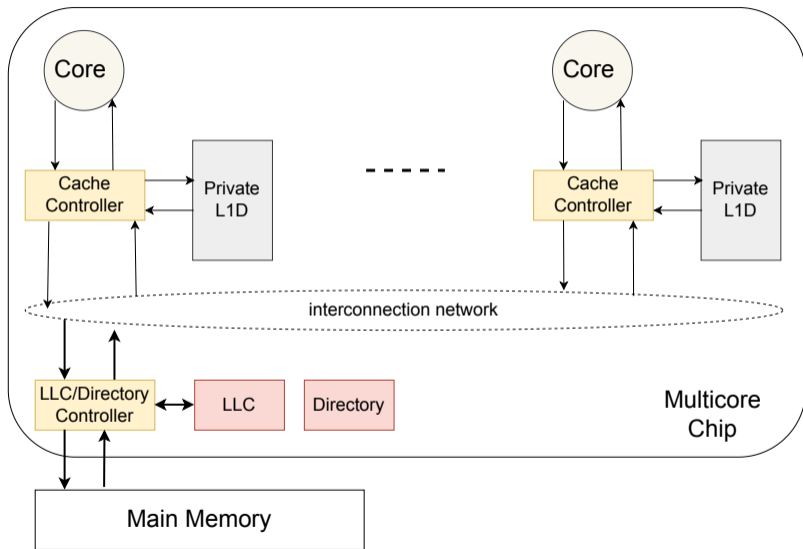
Snoopy Protocol

- Does not scale to large core counts because of broadcast messages
- Requires some ordering guarantees on messages which limits network optimizations

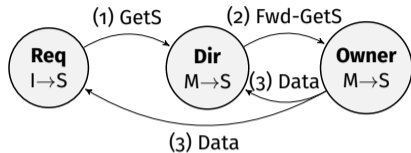
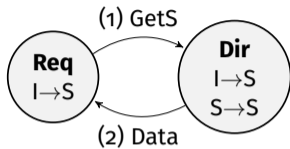
Directory Protocol

- + Scalable because messages are unicast
- + The directory can be distributed to improve scalability
- Few transactions take more cycles when the home is not the owner
- Memory requirement increases with core count

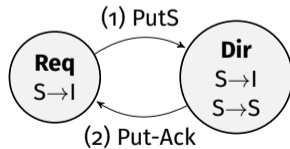
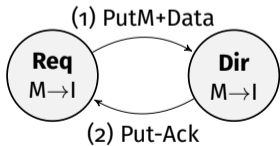
Directory System Model



MSI Protocol

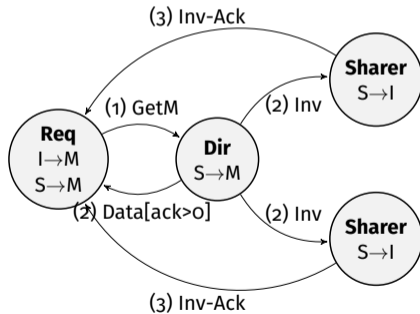
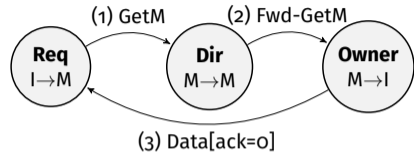
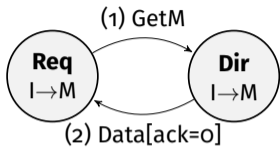


Transitions from I to S



Transitions from M or S to I

MSI Protocol



Transitions from I or S to M

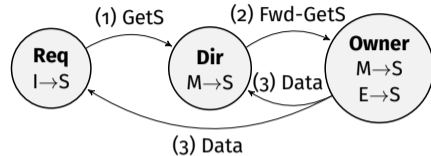
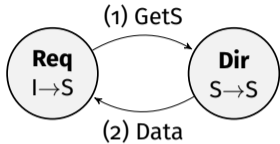
💡 Cores often read data before updating it

- Oftentimes, there is only one sharer in the system (also applicable for single-threaded programs)
- With MSI, the core will issue two coherence transactions: GetS followed by an Upg

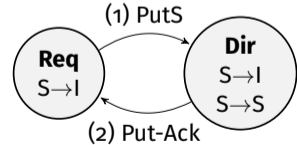
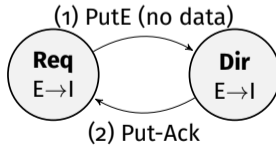
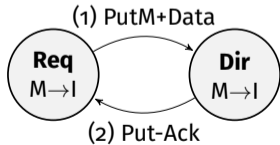
👍 Optimization with E state

- The core that issues GetS for a block gets it in E state if there are no existing sharers
- E state indicates the cache line is clean and is the only cached copy
- The core may then silently upgrade the block from E to M without issuing another coherence request
- We will assume E is an ownership state, which implies evictions cannot be silent

MESI Protocol

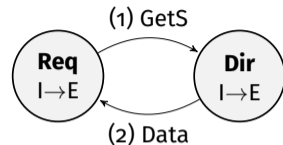
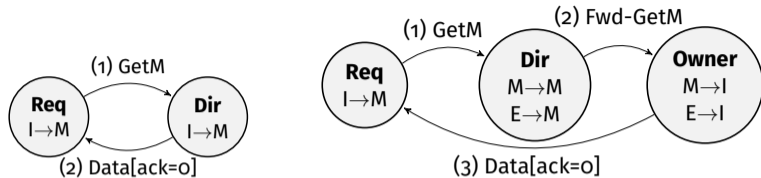


Transitions from I to S

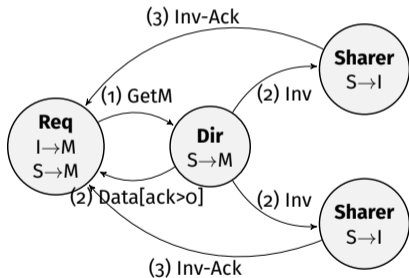


Transitions from M or E or S to I

MESI Protocol



Transitions from I to E



If the only sharer is the requestor, then no Inv messages are sent and the Data message from the Dir to Req has an Ack count of zero.

Transitions from I or S to M

Adding an Owned (O) State

💡 Suppose a cache has a block in state M or E and receives a GetS

The cache changes the block state from M or E to S and sends the data to **both** the requestor and the memory controller

- Why is it necessary to send the data to the memory controller?

Adding an Owned (O) State





Owned state indicates that the block is valid, dirty, and shared, and the cache is the owner

- The owner cache does **not** have permission to modify the block, and is responsible for eventually updating memory

👍 Advantages of MOESI protocol (used in AMD Opteron)

- + Eliminates the extra data message to update the LLC/memory when a cache receives a GetS request in the M or E state
- + Eliminates potentially unnecessary writes to the LLC if the block is written again before being written back to the LLC
- + Allows subsequent requests to be satisfied by the private owner cache instead of the slower LLC/memory

References

-  V. Nagarajan et al. A Primer on Memory Consistency and Cache Coherence. Chapters 1,2,6–8, 2nd edition, Morgan and Claypool.
-  J. Hennessy and D. Patterson. Computer Architecture: A Quantitative Approach. Sections 5.2, 5.4, 6th edition, Morgan Kaufmann.
-  A. Gupta et al. Parallel Computer Architecture: A Hardware/Software Approach. Morgan Kaufmann.
-  Mainak Chaudhuri. Cache Coherence. Computer Architecture Summer School, IIT Kanpur, 2018.