

Multicore Computer Architecture - Storage and Interconnects

Lecture 6

Basic Cache Optimization Techniques



NPTEL Online Course : Multi-Core Computer Architecture and Interconnects Dr. John Jose, IIT Guwahati

Dr. John Jose

Assistant Professor

**Department of Computer Science & Engineering
Indian Institute of Technology Guwahati, Assam.**

Accessing Cache Memory



$$\text{Average memory access time} = \text{Hit time} + (\text{Miss rate} \times \text{Miss penalty})$$

- ❖ **Hit Time:** Time to find the block in the cache and return it to processor [*indexing, tag comparison, transfer*].
- ❖ **Miss Rate:** Fraction of cache access that result in a miss.
- ❖ **Miss Penalty:** Number of additional cycles required upon encountering a miss to fetch a block from the next level of memory hierarchy.

How to optimize cache ?

- ❖ Reduce Average Memory Access Time
- ❖ $AMAT = \text{Hit Time} + \text{Miss Rate} \times \text{Miss Penalty}$
- ❖ Motives
 - ❖ Reducing the miss rate
 - ❖ Reducing the miss penalty
 - ❖ Reducing the hit time

Larger Block Size

- ❖ **Larger block size to reduce miss rate**

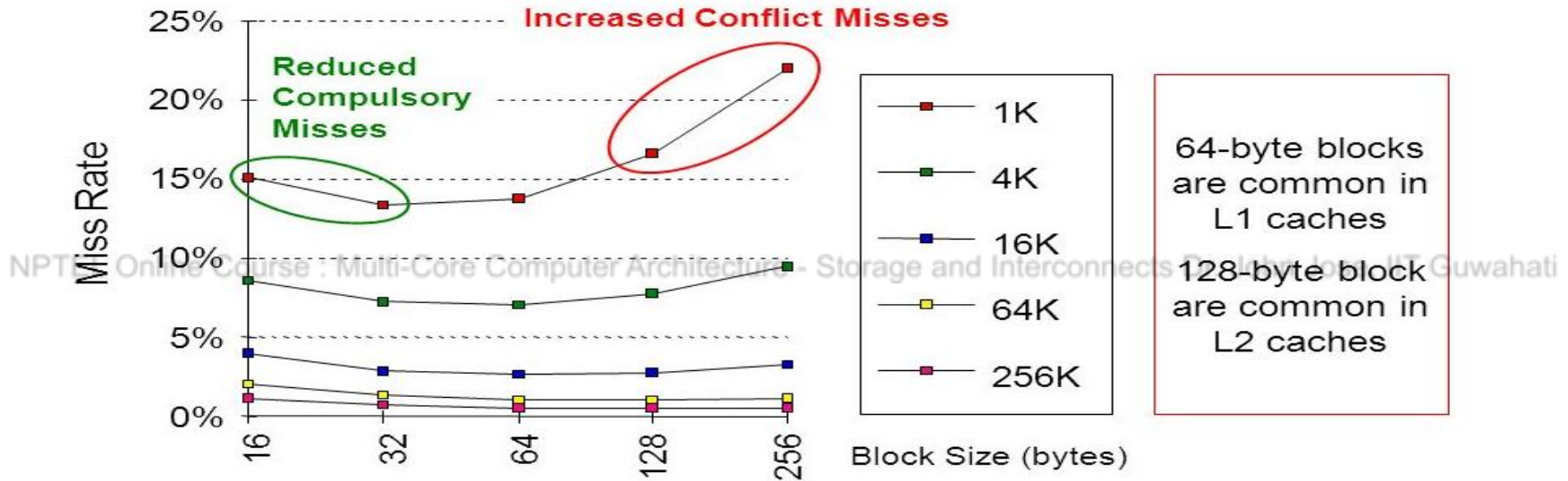
- ❖ **Advantages**

- ❖ Utilize spatial locality
- ❖ Reduces compulsory misses

- ❖ **Disadvantages**

- ❖ Increases miss penalty
- ❖ More time to fetch a block to the cache [bus width issue]
- ❖ Increases conflict misses
- ❖ More number of blocks will be mapped to the same location
- ❖ May bring useless data and evict useful data [pollution]

Larger Block Size



Larger Caches

- ❖ **Larger cache to reduce miss rate**

- ❖ **Advantages**

 - ❖ Reduces capacity misses

 - ❖ Can accommodate larger memory footprint

NPTEL Online Course : Multi-Core Computer Architecture - Storage and Interconnects Dr. John Jose, IIT Guwahati

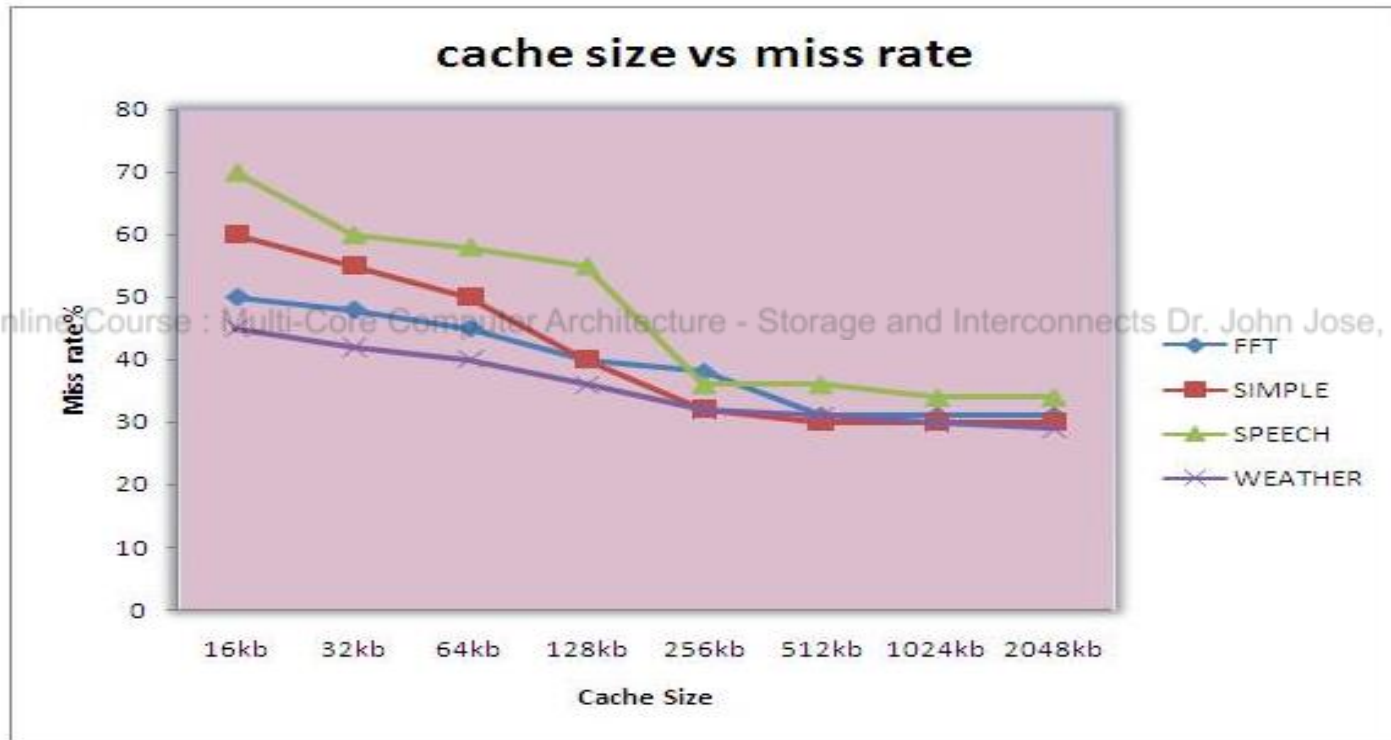
Block size	Cache size			
	4K	16K	64K	256K
16	8.57%	3.94%	2.04%	1.09%
32	7.24%	2.87%	1.35%	0.70%
64	7.00%	2.64%	1.06%	0.51%
128	7.78%	2.77%	1.02%	0.49%
256	9.51%	3.29%	1.15%	0.49%

- ❖ **Drawbacks**

 - ❖ Longer hit time

 - ❖ Higher cost, area and power

Larger Caches



Higher Associativity

❖ Higher associativity to reduce miss rate

- ❖ Fully associative caches are the best, but high hit time.
- ❖ So increase the associativity to an optimal possible level

❖ Advantages

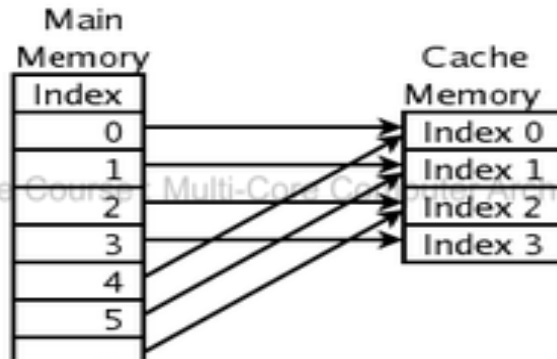
- ❖ Reduce conflict miss
- ❖ Reduce miss rate and eviction rate

❖ Drawbacks

- ❖ Increase in the hit time
- ❖ Complex design than direct mapped
- ❖ More time to search in the set (tag comparison time)

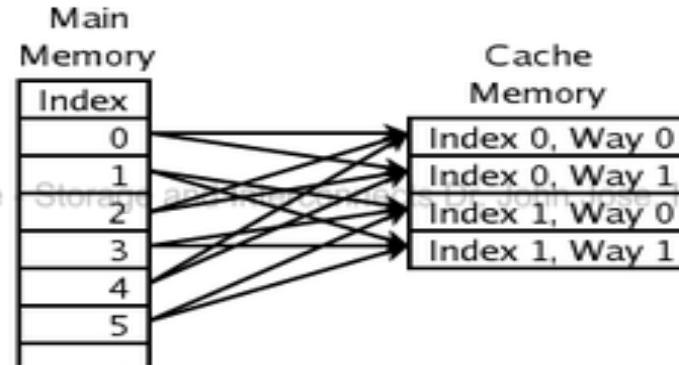
AMAT vs cache associativity

Direct Mapped
Cache Fill



...
Each location in main memory can be
cached by just one cache location.

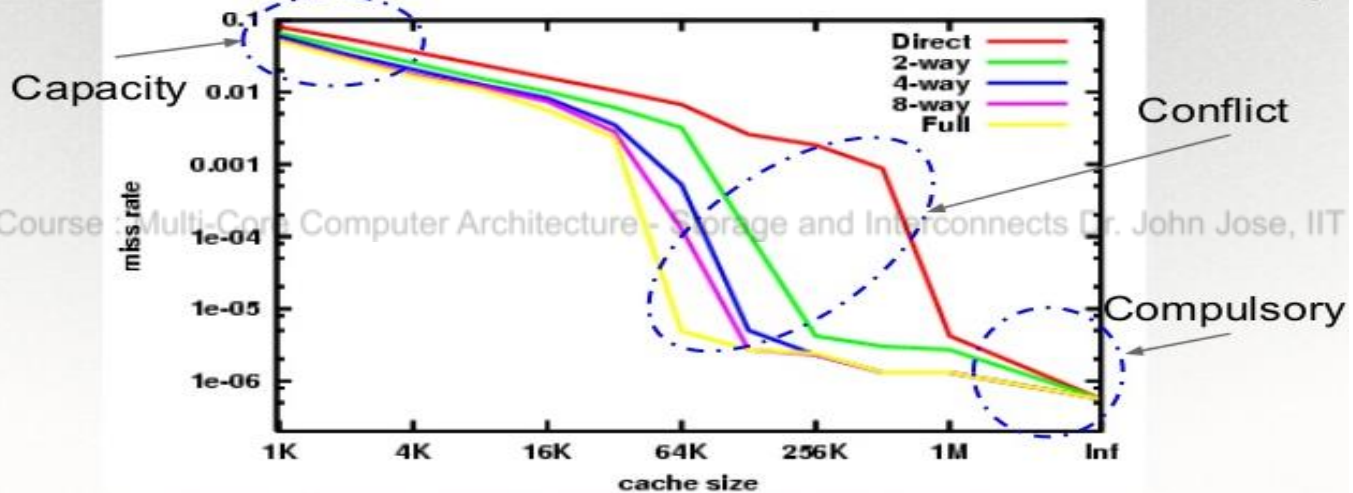
2-Way Associative
Cache Fill



...
Each location in main memory can be
cached by one of two cache locations.

AMAT vs cache associativity

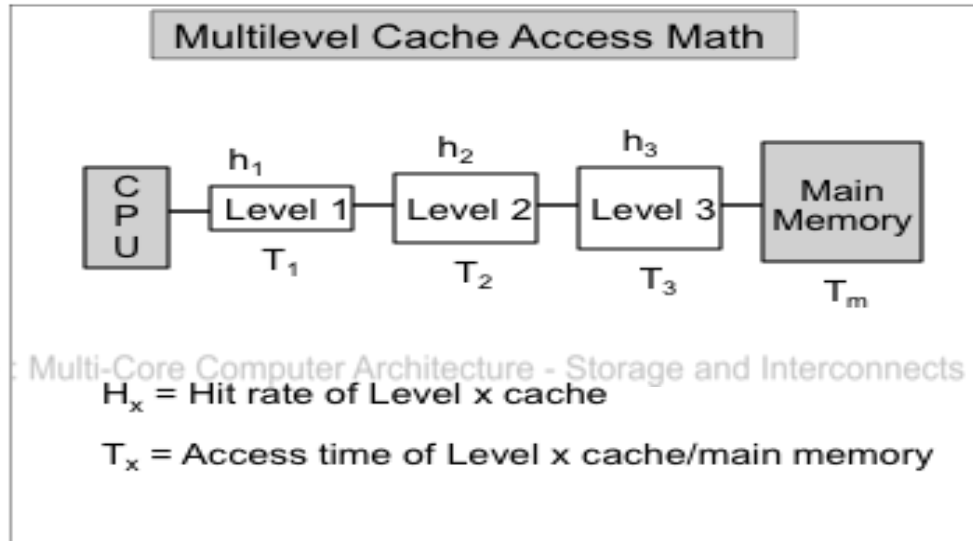
Miss rate vs cache size (SPEC CPU2000)



Multilevel caches

- ❖ **Multilevel caches to reduce miss penalty**
- ❖ Performance gap between processors and memory
- ❖ **Caches should be faster** to keep pace with the speed of processors, **AND cache should be larger** to overcome the widening gap between the processor and main memory
- ❖ Add another level of cache between the cache and memory.
- ❖ The first-level cache (L1) can be small enough to match the clock cycle time of the fast processor. [Low hit time]
- ❖ The second-level cache (L2) can be large enough to capture many accesses that would go to main memory, thereby lessening the effective miss penalty. [Low miss rate]

Multilevel caches



NPTEL Online Course : Multi-Core Computer Architecture - Storage and Interconnects Dr. John Jose, IIT Guwahati

Average memory access time = Hit time_{L1} + Miss rate_{L1} × Miss penalty_{L1}

Miss penalty_{L1} = Hit time_{L2} + Miss rate_{L2} × Miss penalty_{L2}

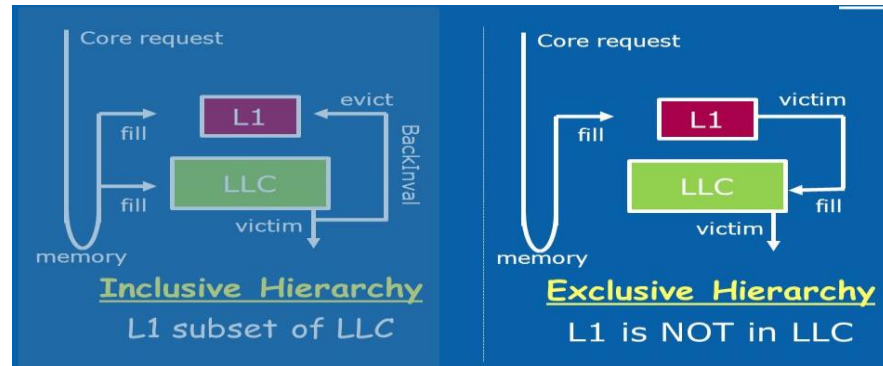
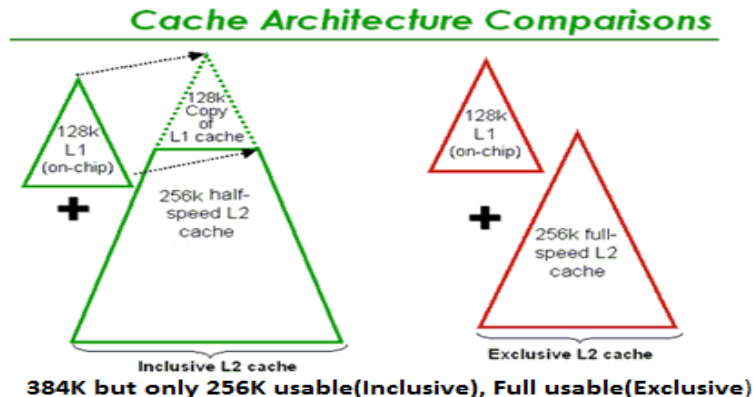
Average memory access time = Hit time_{L1} + Miss rate_{L1}
× (Hit time_{L2} + Miss rate_{L2} × Miss penalty_{L2})

Multilevel caches

- ❖ **Multilevel caches to reduce miss penalty**
- ❖ **Local miss rate:** Number of misses in a cache level divided by number of memory access to this level.
- ❖ **Global miss rate:** Number of misses in a cache level divided by number of memory access generated by the CPU.

NPTEL Online Course : Multi-Core Computer Architecture - Storage and Interconnects Dr. John Jose, IIT Guwahati

- ❖ **Inclusive and Exclusive caches**



Prioritize read miss over writes

- ❖ **Prioritize read misses to reduce miss penalty**

- ❖ If a read miss has to evict a dirty memory block, the normal sequence is write the dirty block to memory and read the missed block

```
SW R3, 512(R0)      ;M[512] ← R3      (cache index 0)
LW R1, 1024(R0)     ;R1 ← M[1024]    (cache index 0)
LW R2, 512(R0)      ;R2 ← M[512]    (cache index 0)
```

NPTEL Online Course : Multi-Core Computer Architecture - Storage and Interconnects Dr. Sunil Joshi, IIT Guwahati

- ❖ Optimization: copy the dirty block to a buffer, read from memory and then write the block - reduces CPU's waiting time on read miss
- ❖ In write through caches, write buffers may hold the updated values of a block that encountered a read miss
- ❖ Either wait till write buffer is empty or search in write buffer before going to memory.

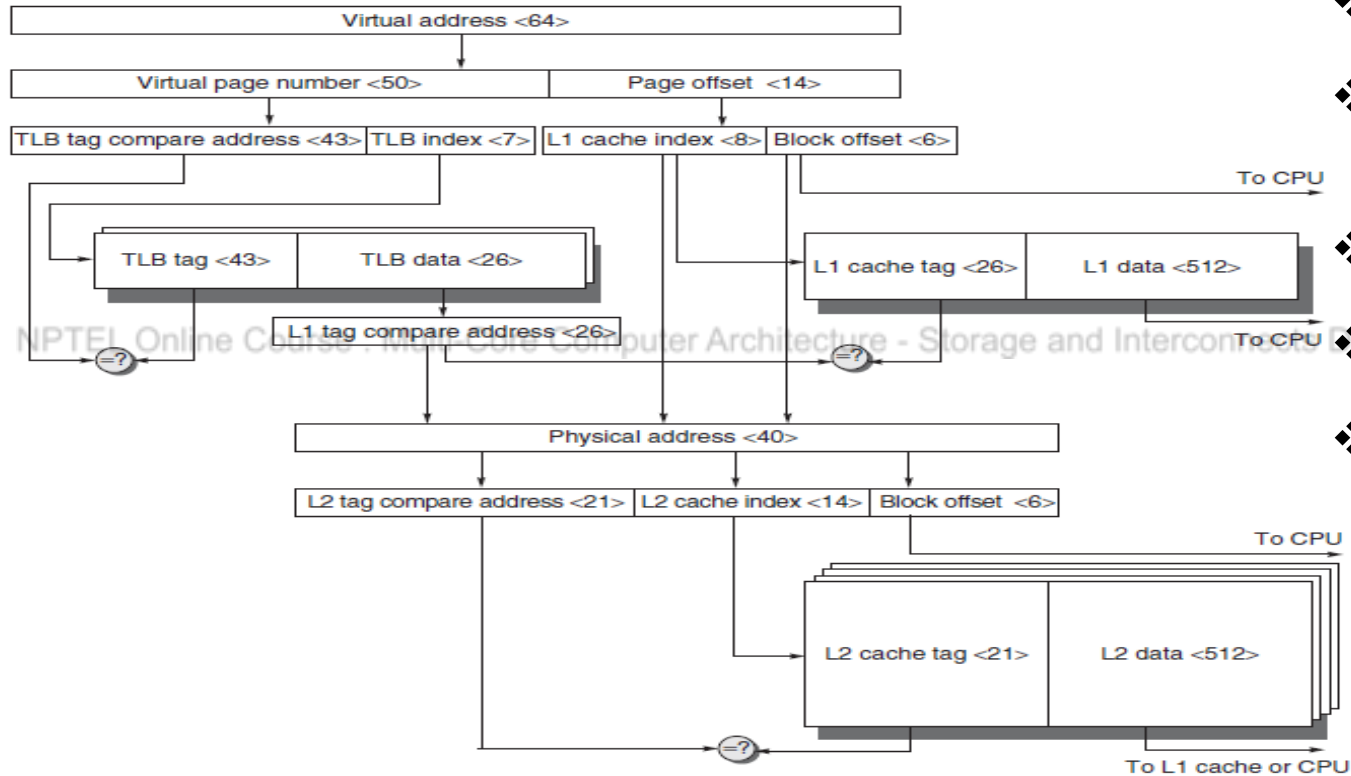
Avoid address translation for indexing

- ❖ **Avoid address translation in cache indexing to reduce hit time**
- ❖ Software uses virtual address (VA), but memory accessed using physical address (PA)
- ❖ VA to PA has to be done before cache look up **[MMU, TLB]**
- ❖ If indexing & tag comparison in virtual address, better hit time
- ❖ **Solution:** Start indexing with VA
- ❖ In the meantime do address translation and get PA and tag.
- ❖ Role of TLB, TLB access in critical path.
- ❖ Indexing by virtual address and tagging by physical address (VIPT cache)

Avoid address translation for indexing

- ❖ **Avoid address translation in cache indexing to reduce hit time**
- ❖ Software uses virtual address (VA), but memory accessed using physical address (PA) - Illustration

Avoid address translation for indexing



- ❖ 16 KB page size
- ❖ 256 entry 2 way TLB
- ❖ 16KB DM L1
- ❖ 4MB 4W L2
- ❖ 64B block

**THANK
YOU!**

FOR

YOUR

PRESENCE

NPTEL Online Course : Multi-Core Computer Architecture - Storage and Interconnects Dr. John Jose, IIT Guwahati

johnjose@iitg.ernet.in

<http://www.iitg.ernet.in/johnjose/>