CSARCH Lecture Series: Cache Memory: Replacement Algorithm

Sensei RL Uy

College of Computer Studies

De La Salle University

Manila, Philippines

'2110

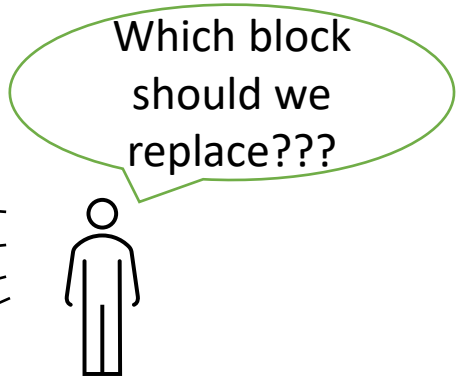
Copyright Notice

Overview

Reflect on the following question:

- When the cache memory is full, and a new memory block is to be transferred. Which cache memory block will be the “victim” (i.e., replaced)?

Block	Data
0	1
1	7
2	5
3	0



Overview

- This sub-module introduces the concept of cache memory replacement algorithm
- The objectives are as follows:
  - ✓Describe the replacement algorithm used by direct mapping
  - ✓Describe the Least Recently Used (LRU) replacement algorithm
  - ✓Describe the Most Recently Used (MRU) replacement algorithm

Cache operation

- In which cache block will the main memory blocks be placed? [Mapping function]
- Which cache block to replace? [replacement algorithm]

Replacement Algorithm

- ***Replacement algorithms*** are used to identify which cache block may be overwritten (released) when all cache blocks a main memory block may occupy are already used.
- The block that will no longer be accessed should be released $\Rightarrow$ very difficult to determine.

Replacement Algorithm

- **Least Recently Used** (**LRU**) – replace the least recently used block first
- **Most Recently Used** (**MRU**) – replace the most recently used block first
- **Random** – replace random block.

Replacement algorithm and Direct Mapping

- Memory is located by comparing the tag of the cache block corresponding to the memory block address.
- There is a large possibility of ***contention*** without full utilization.
- Replacement algorithm is automatically handled by the “modulo” function

Direct Mapping

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; (3) snapshot of the cache memory

Block	Data
0	
1	
2	
3	

Block	Data
0	
1	
2	
3	

Seq	Hit	Miss	Block
1			
7			
5			
0			
2			
1			
5			
6			
5			
2			
2			
0			

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty:

Hit rate: ; miss rate:

Average access time =

Total access time =

Direct Mapping

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; (3) snapshot of the cache memory

Block	Data
0	0
1	1,5,1,5
2	2,6,2
3	7

Block	Data
0	0
1	5
2	2
3	7

Seq	Hit	Miss	Block
1		1	1
7		7	3
5		5	1
0		0	0
2		2	2
1		1	1
5		5	1
6		6	2
5	5		
2		2	2
2	2		
0	0		

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty: 1ns+20ns+1ns = 22ns

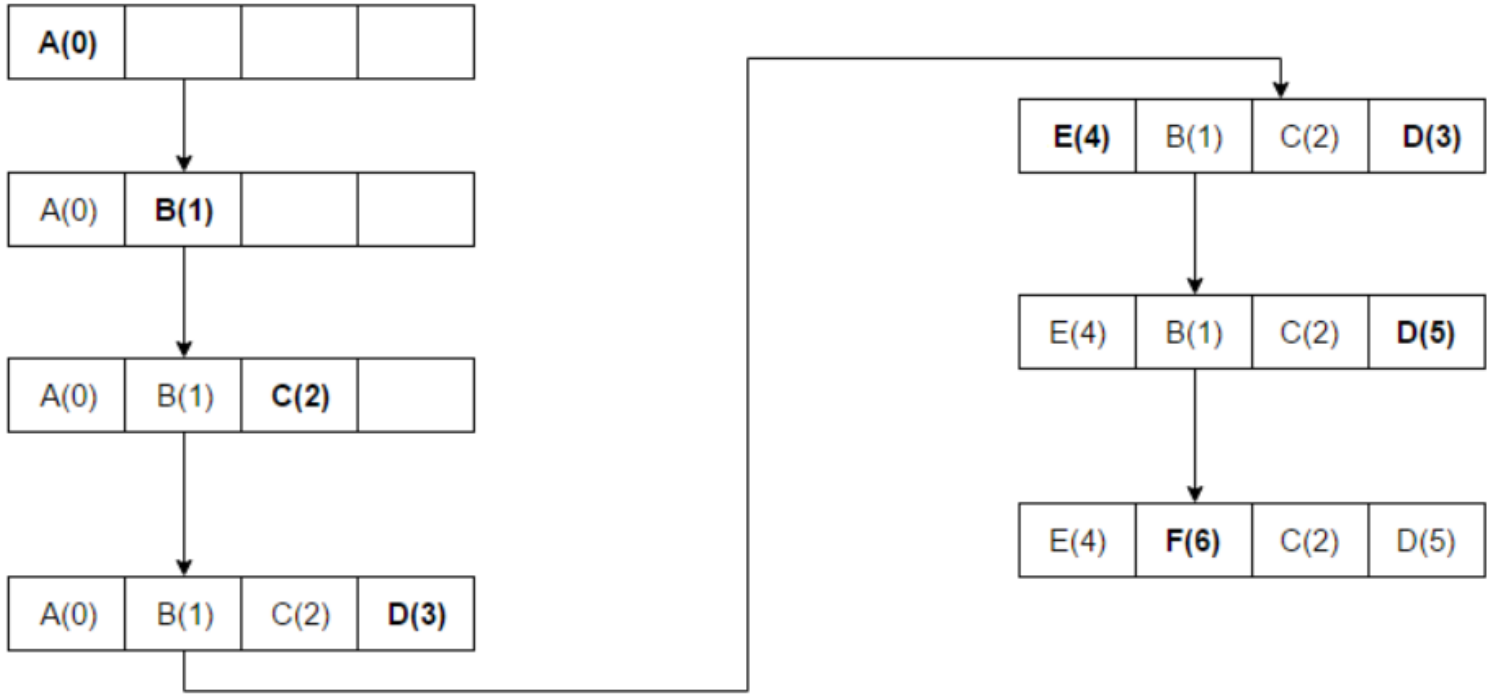
Hit rate: 3/12; miss rate: 9/12

Average access time = 0.25*1ns + 0.75*22ns
= 16.75ns

Total access time = 3*2*1ns + 9*2*11ns+9*1ns
= 204ns+9ns = 213ns

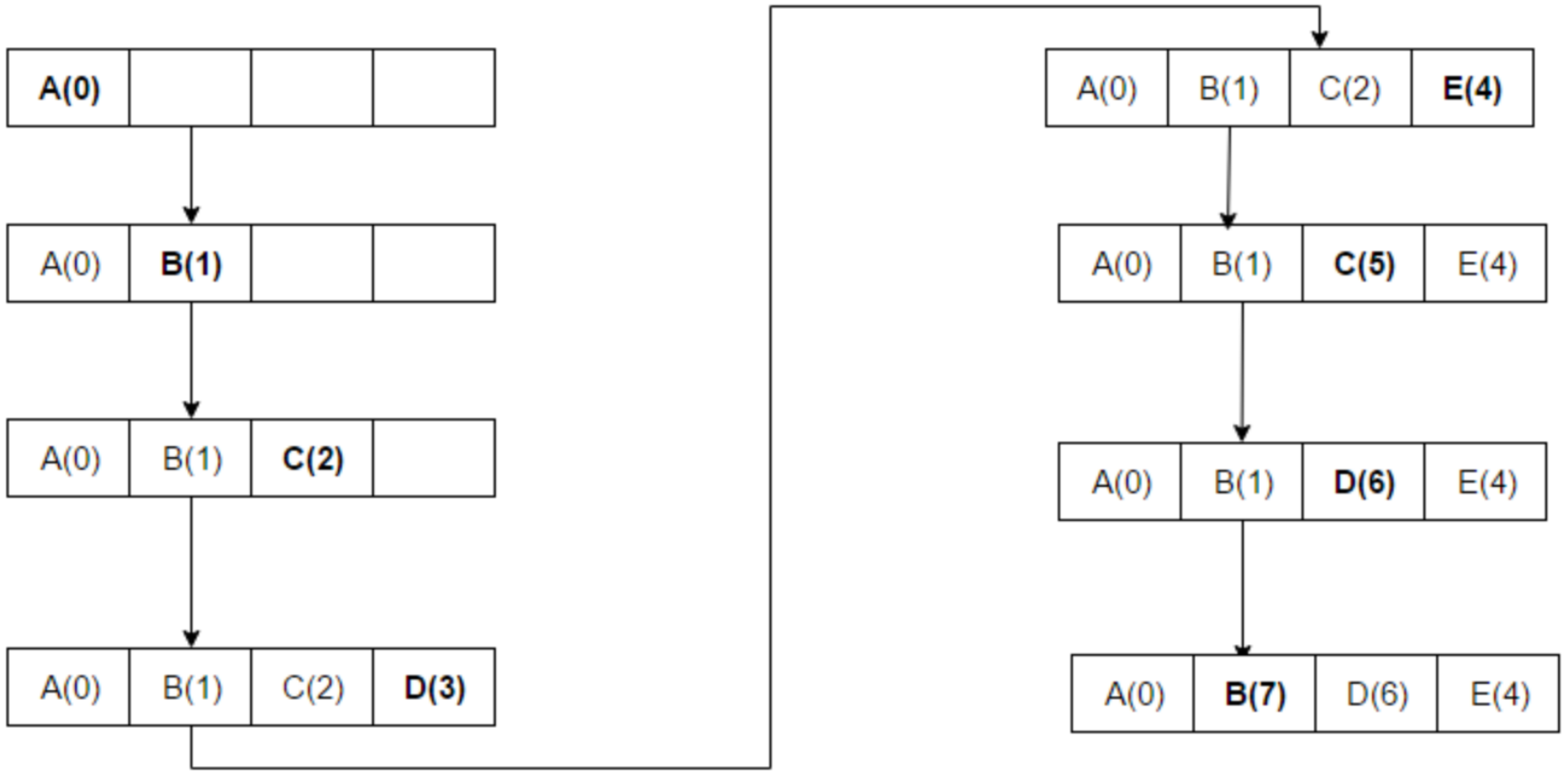
Least Recently Used (LRU)

The access sequence for the below example is A B C D E D F.



Most Recently Used (MRU)

The access sequence for the below example is A B C D E C D B.



Full Associative Mapping (LRU)

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; and (3) snapshot of the cache memory

Block	Age	Data
0		
1		
2		
3		

Block	Data
0	
1	
2	
3	

Seq	Hit	Miss	Block
1			
7			
5			
0			
2			
1			
5			
6			
5			
2			
2			
0			

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty:

Hit rate: ; miss rate:

Average access time =

Total access time =

Full Associative Mapping (LRU)

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; and (3) snapshot of the cache memory

Block	Age	Data
0	10	2
1	11	0
2	8	5
3	7	6

Block	Data
0	2
1	0
2	5
3	6

Seq	Hit	Miss	Block
1		1	0
7		7	1
5		5	2
0		0	3
2		2	0
1		1	1
5	5		
6		6	3
5	5		
2	2		
2	2		
0		0	1

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty: 1ns+20ns+1ns = 22ns

Hit rate: 4/12; miss rate: 8/12

Average access time = 0.33*1ns + 0.67*22ns
= 15.00ns

Total access time = 4*2*1ns + 8*2*11ns + 8*1ns
= 184ns + 8ns = 192ns

Full Associative Mapping (MRU)

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; and (3) snapshot of the cache memory

Block	Age	Data
0		
1		
2		
3		

Block	Data
0	
1	
2	
3	

Seq	Hit	Miss	Block
1			
7			
5			
0			
2			
1			
5			
6			
5			
2			
2			
0			

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty:

Hit rate: miss rate:

Average access time =

Total access time =

Full Associative Mapping (MRU)

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; and (3) snapshot of the cache memory

Block	Age	Data
0	5	1
1	1	7
2	8	5
3	11	0

Block	Data
0	1
1	7
2	5
3	0

Seq	Hit	Miss	Block
1		1	0
7		7	1
5		5	2
0		0	3
2		2	3
1	1		
5	5		
6		6	2
5		5	2
2	2		
2	2		
0		0	3

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty: 1ns+20ns+1ns = 22ns

Hit rate: 4/12; miss rate: 8/12

Average access time = 0.33*1ns + 0.67*22ns
= 15.00ns

Total access time = 4*2*1ns + 8*2*11ns + 8*1ns
= 184ns + 8ns = 192ns

Block Set Associative (LRU)

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; and (3) snapshot of the cache memory

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty:

Hit rate:; miss rate:

Average access time =

Total access time =

Set	Block 0	Block 1
0		
Age		
1		
Age		

Set	Block 0	Block 1
0		
1		

Seq	Hit	Miss	Set
1			
7			
5			
0			
2			
1			
5			
6			
5			
2			
2			
0			

Block Set Associative (LRU)

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; and (3) snapshot of the cache memory

Set	Block 0	Block 1
0	0	2
Age	6	5
1	5	1
Age	5	3

Set	Block 0	Block 1
0	0	2
1	5	1

Seq	Hit	Miss	Set
1		1	1
7		7	1
5		5	1
0		0	0
2		2	0
1		1	0
5	5		
6		6	0
5	5		
2	2		
2	2		
0		0	0

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty: 1ns+20ns+1ns = 22ns

Hit rate: 4/12; miss rate: 8/12

Average access time = 0.33*1ns + 0.67*22ns
= 15.00ns

Total access time = 4*2*1ns + 8*2*11ns + 8*1ns
= 184ns + 8ns = 192ns

Block Set Associative (MRU)

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; and (3) snapshot of the cache memory

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty:

Hit rate: miss rate:

Average access time =

Total access time =

Set	Block 0	Block 1
0		
Age		
1		
Age		

Set	Block 0	Block 1
0		
1		

Seq	Hit	Miss	Set
1			
7			
5			
0			
2			
1			
5			
6			
5			
2			
2			
0			

Block Set Associative (MRU)

- Cache has 4 blocks, set size is 2 blocks, block size is 2 words, cache access time is 1ns, memory access time is 10 ns. Given the main memory block sequence below, (1) compute for the average access time; (2) total access time; and (3) snapshot of the cache memory

Set	Block 0	Block 1
0	0	2
Age	5	4
1	1	5
Age	3	5

Set	Block 0	Block 1
0	0	2
1	1	5

Seq	Hit	Miss	Set
1		1	1
7		7	1
5		5	1
0		0	0
2		2	0
1	1		
5	1		
6		6	0
5	1		
2		2	0
2	2		
0	0		

MM Block sequence: 1,7,5,0,2,1,5,6,5,2,2,0

Miss penalty: 1ns+20ns+1ns = 22ns

Hit rate: 5/12; miss rate: 7/12

Average access time = 0.42*1ns + 0.58*22ns
= 13.25ns

Total access time = 5*2*1ns + 7*2*11ns + 7*1ns
= 164ns + 7ns = 171ns

Try

- Given the main memory block sequence, show the contents of the cache using direct mapping, fully associative mapping(LRU, MRU) and block set associative(LRU, MRU) mapping . Compute for the hit rate, miss rate, total memory access time and average memory access time. Assume there are 4 cache blocks, 4 words/block, 2 blocks/set, cache access time is 1ns, memory access time is 10 ns and non-load through
- MM Block Sequence: 1,2,3,4,5,4,6,3

To recall …

- What have we learned:
  - ✓Describe the replacement algorithm used by direct mapping
  - ✓Describe the Least Recently Used (LRU) replacement algorithm
  - ✓Describe the Most Recently Used (MRU) replacement algorithm