Discussion 9 - Cache

CS 110 COMPUTER ARCHITECTURE

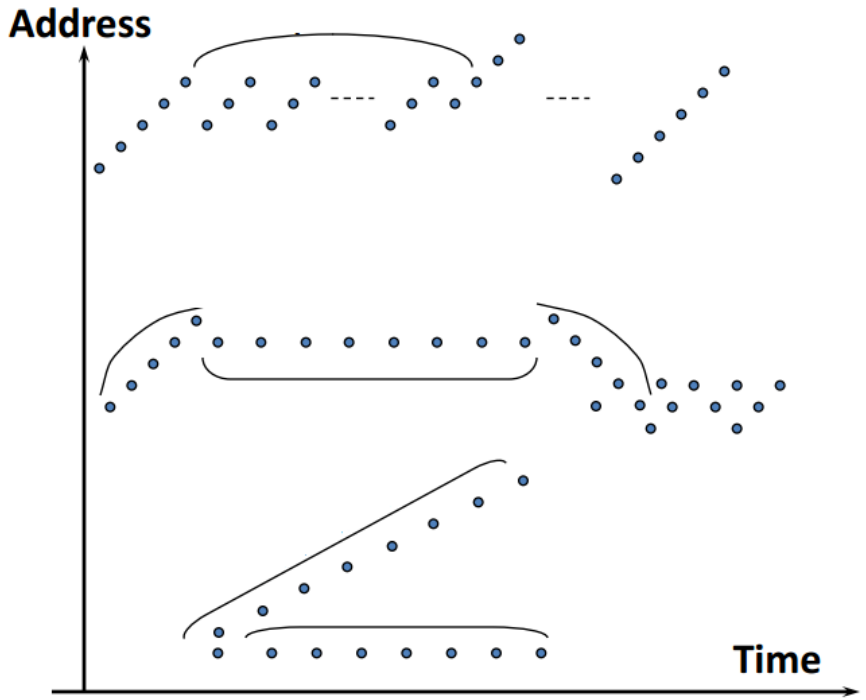
by 沈喆奇(Shen, Zheqi)

Review: Big Idea - Locality

Principle of Locality
Programs access small portion of address space at any instant of time (spatial locality) and repeatedly access that portion (temporal locality)

Memory Reference Patterns

instruction accesses & data accesses
separated L1 instruction/data cache design



Why the memory cannot be faster?
Why the caches cannot be larger?

Review: Types of caches

Direct Mapping cache

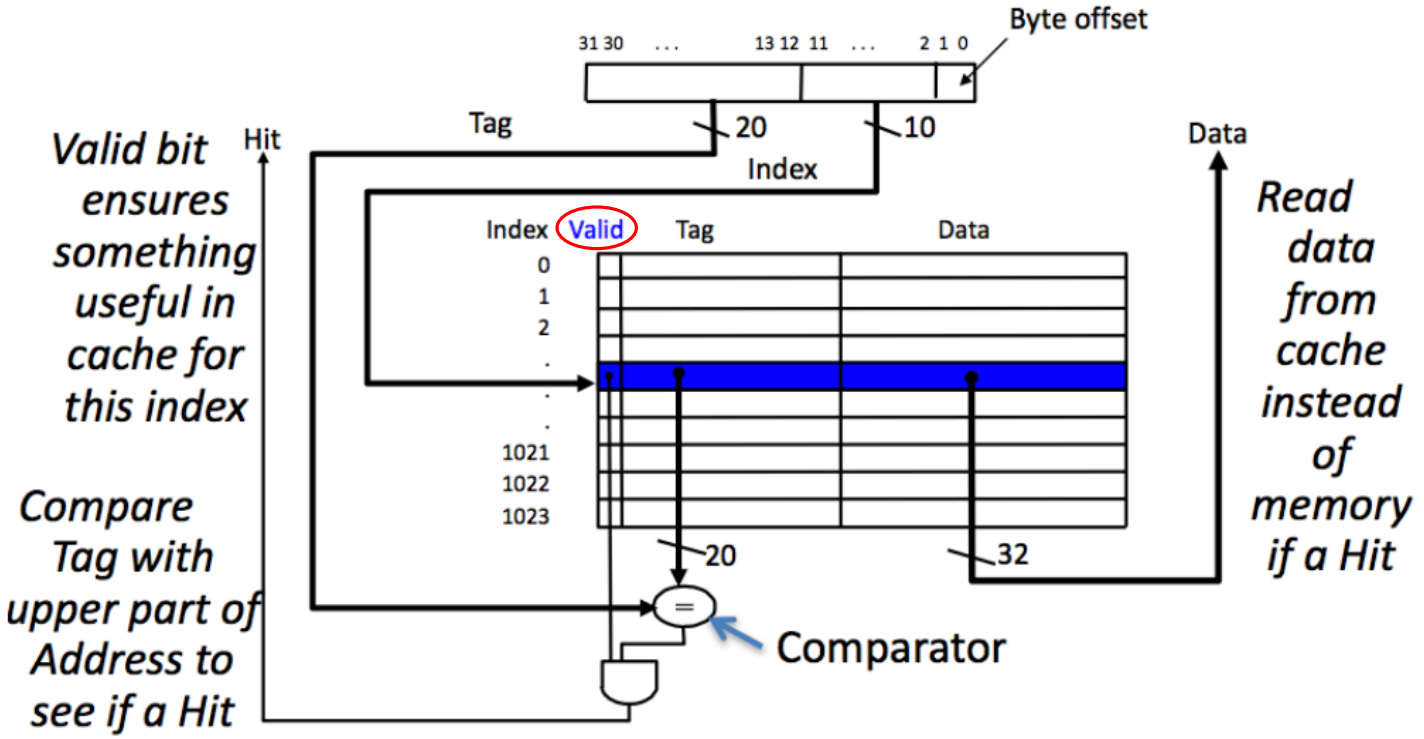
size of set = 1

N-Way Set Associative

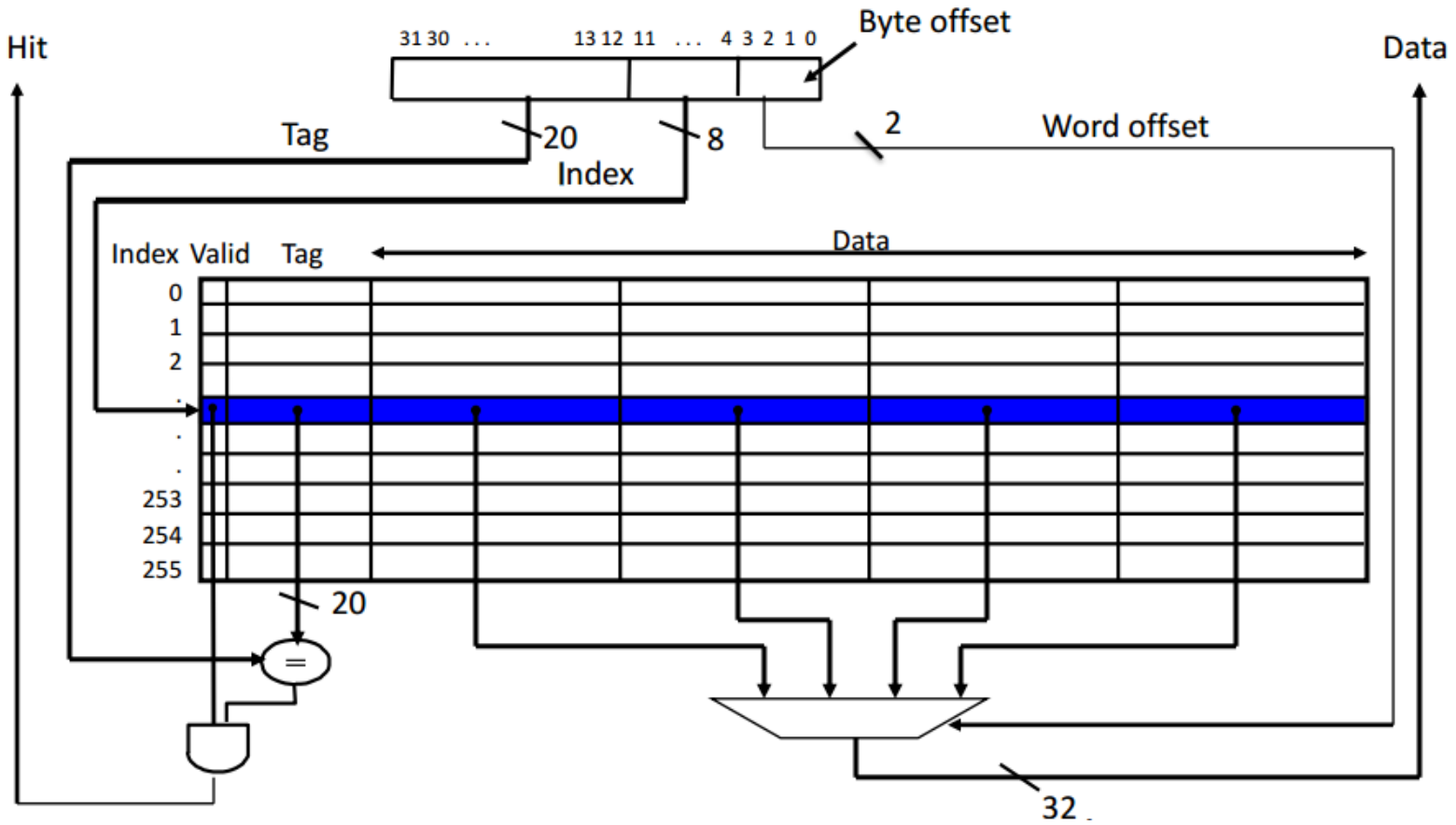
# of set = 1

Fully Associative

Extra bits in a cache entry
What's Valid flag used for?



Multiword-Block Cache



x86 cache line size is 64 bytes

Write Policy

Write-Through Policy

Immediately write data to the memory

Invovle Write Buffer to reduce the overhead

Predictable timing – what is it?

Reliable

Write-Back Policy

Write data until it is evited from the cache

Invovle the Dirty flag

Variable timing – what are the possibilities? 0, 1, or 2

Less reliable – why? inconsistency

What are the cache replacement policies?
What about the application scenarios?

Average Memory Access Time (AMAT)

Average time to access memory considering both hits and misses in the cache

What is the formula?

Why the time for a hit is always counted?

What about the multi-level caches?

从CPU到	大约需要的 CPU 周期	大约需要的时间
主存		约60-80纳秒
QPI 总线传输 (between sockets, not drawn)		约20ns
L3 cache	约40-45 cycles,	约15ns
L2 cache	约10 cycles,	约3ns
L1 cache	约3-4 cycles,	约1ns
寄存器	1 cycle	

3CS Misses

Compulsory

Capacity

Conflict

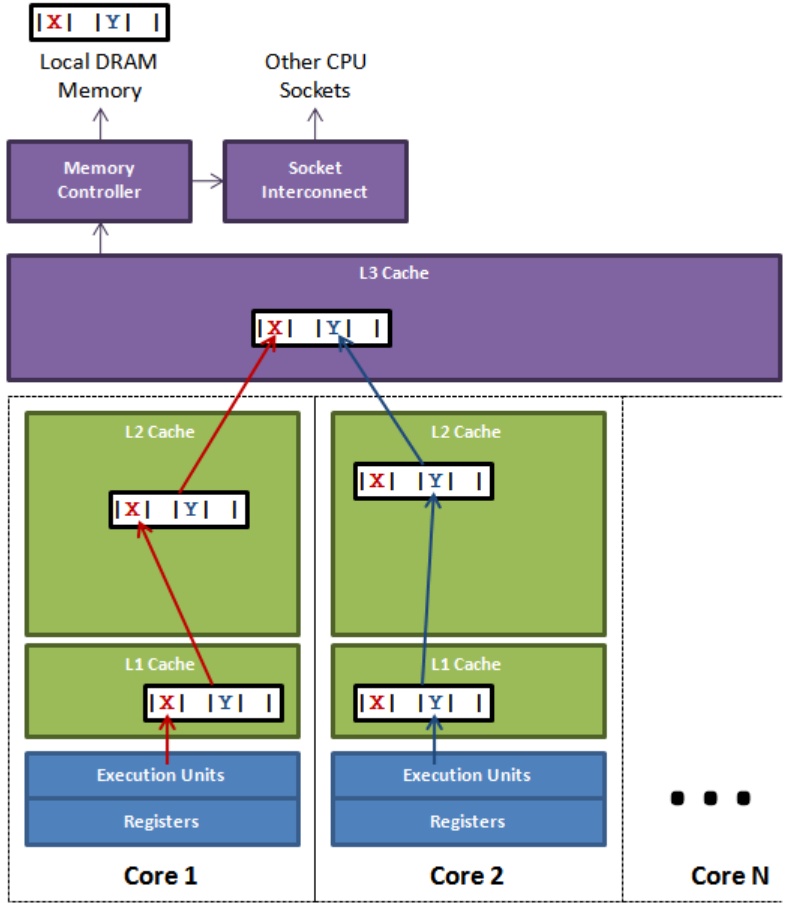
Belady's Anomaly

May occurs in FIFO, but not in LRU

	Cache with 3 slots											
time	0	1	2	3	4	5	6	7	8	9	10	11
w	d	c	b	a	d	c	e	d	c	b	a	e
slot 1	d	d	d	a	a	a	e	e	e	e	e	e
slot 2		c	c	c	d	d	d	d	d	b	a	a
slot 3			b	b	b	c	c	c	c	c	c	c
#miss	1	2	3	4	5	6	7	7	7	8	9	9

	Cache with 4 slots											
time	0	1	2	3	4	5	6	7	8	9	10	11
w	d	c	b	a	d	c	e	d	c	b	a	e
slot 1	d	d	d	d	d	d	e	e	e	e	a	a
slot 2		c	c	c	c	c	c	d	d	d	d	e
slot 3			b	b	b	b	b	b	c	c	c	c
slot 4				a	a	a	a	a	a	b	b	b
#miss	1	2	3	4	4	4	5	6	7	8	9	10

False Sharing



Does cache line always take benefits?
- Actually not.

Search MESI protocol and RFO request on internet if you want to learn more about this.

Exercises

Assume we have a direct-mapped byte-addressed cache with capacity 32 Bytes and block size of 8 Bytes.

Of the 32 bits in each address, which bits do we use to find the index of the cache to use?

Which bits are out offset?

Exercises

Given the follow chunk of code, analyze the hit rate given that we have a byteaddressed computer with a total memory of 1 MiB. It also features a 16 KiB Direct-Mapped cache with 1 KiB blocks.

```
#define NUM_INTS 8192    // 2^13
int A[NUM_INTS];         // A lives at 0x10000
int i, total = 0;
for (i = 0; i < NUM_INTS; i += 128) {
    A[i] = i;            // Line 1
}
for (i = 0; i < NUM_INTS; i += 128) {
    total += A[i];       // Line 2
}
```

How many bits make up a memory address on this computer? - log(1M) = 20

How many bits is used for Offset and Index?
- 10 for offset, 4 for index

What is the hit rate for Line 1? - 50%

What is the hit rate for Line 2? - 50%

Q&A

THANKS FOR YOUR ATTENDANCE