جلسه‌ی بیست و سوم

معماری کامپیوتر

۱۳۹۵-۱۱-۱۳

حافظه ۵

دانشگاه شهید بهشتی

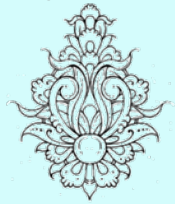
دانشکده‌ی مهندسی برق و کامپیوتر

بهار ۱۳۹۵

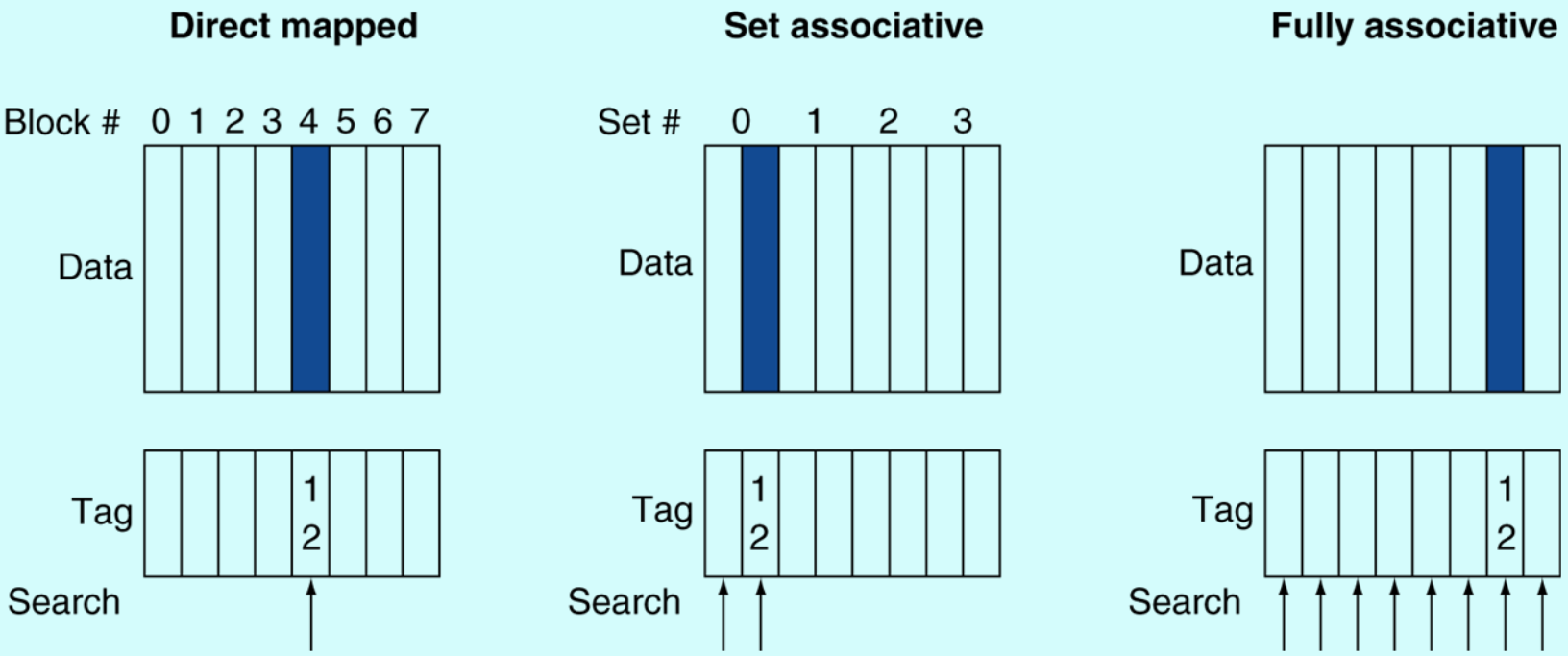
احمد محمودی ازناوه

فهرست مطالب

- مروری بر جلسه‌ی پیش
- مثال
- حافظه‌ی نهان چند سطحی
- حافظه‌ی مجازی



حافظه‌های نهان اشتراکی



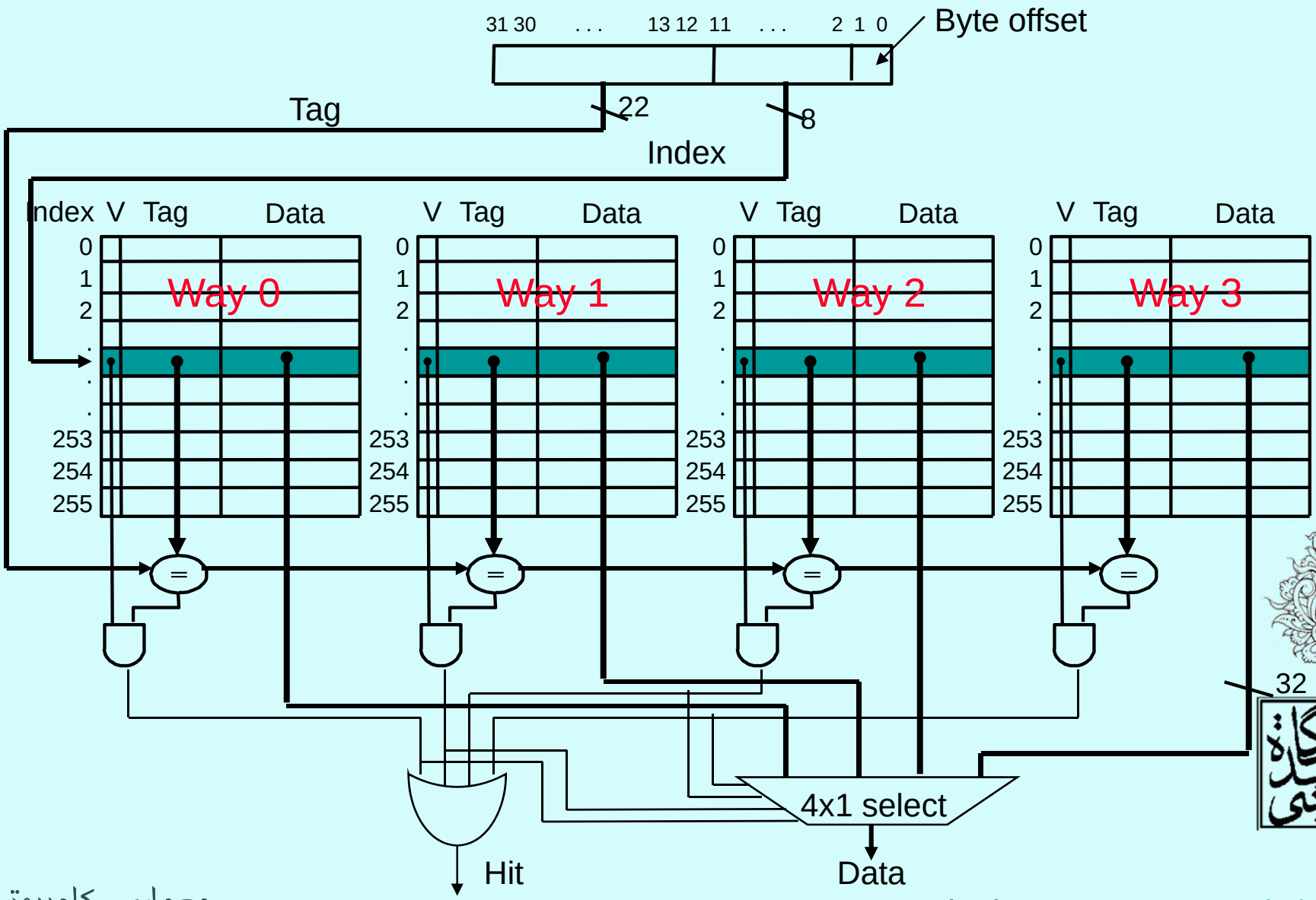
(Block address) modulo (#Blocks in cache)

Direct Map

(Block address) modulo (#Sets in cache)

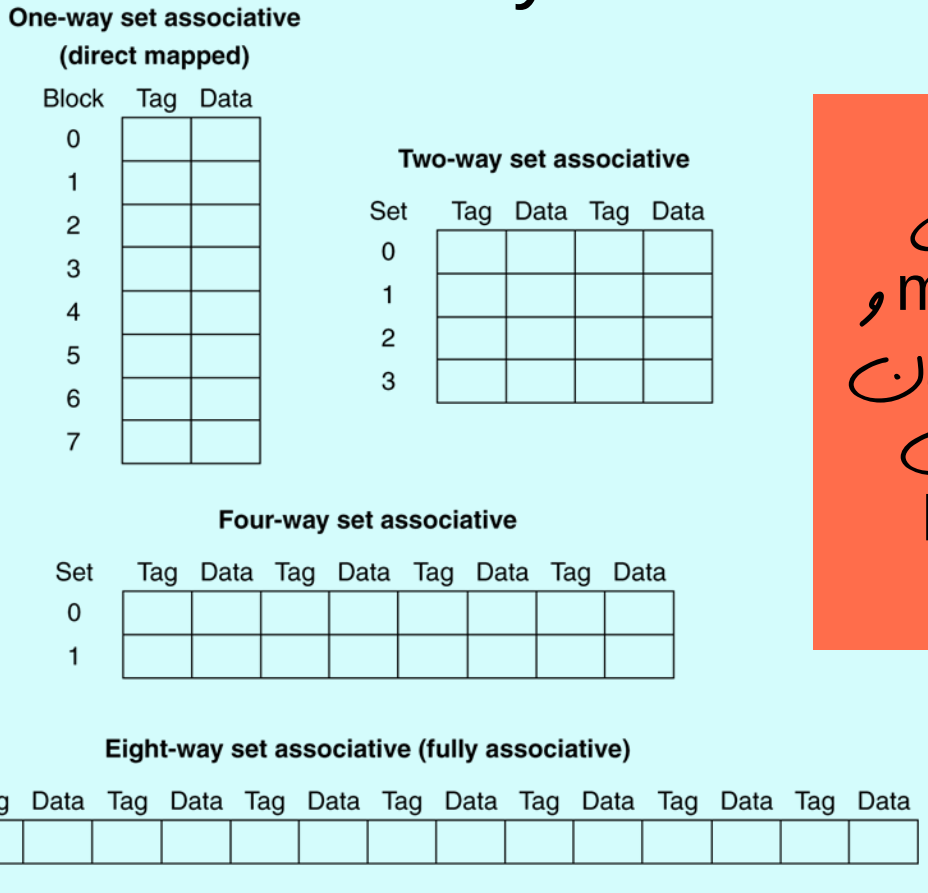
Set Associative

Four-Way Set Associative Cache



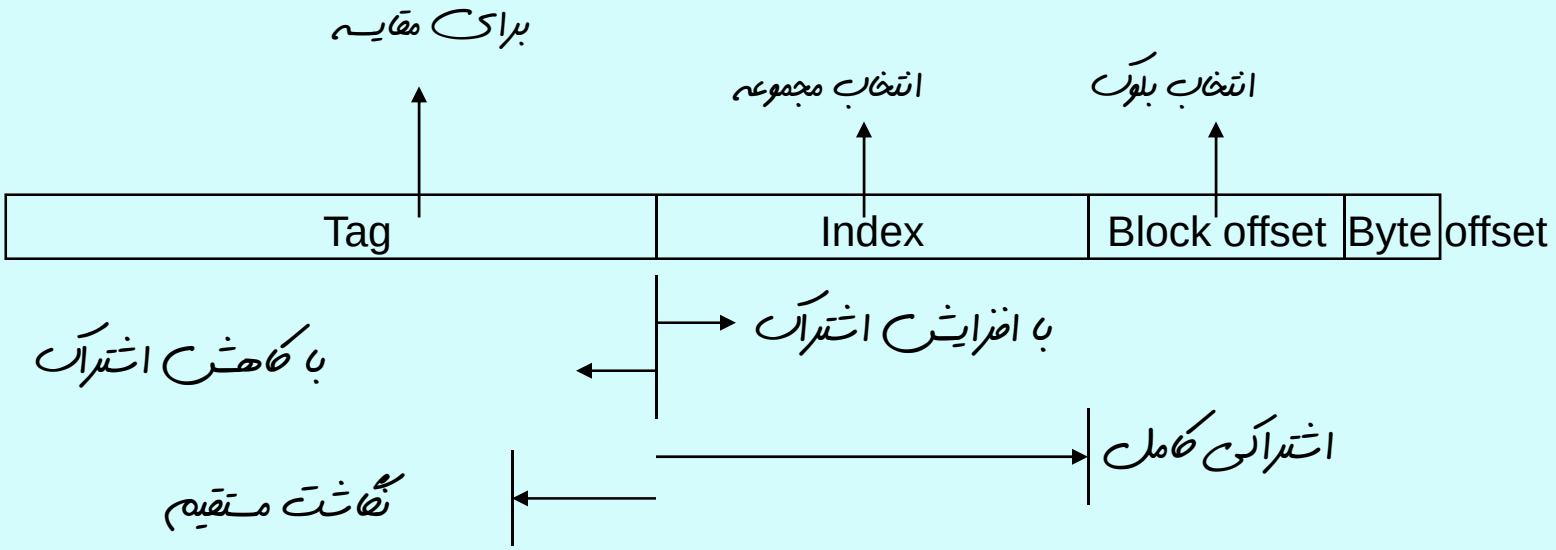
طیف اشتراک

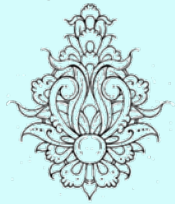
- تمام شیوه‌های جای‌دهی را به نوعی می‌توان n-way set associative دانست.



مهمترین برتری شیوه‌های اشتراکی کاهش miss-rate و بزرگ‌ترین مشکل آن افزایش مشکل آن افزایش hit-time است

طیف اشتراک (ادامه...)





مثال- نگاشت مستقیم

- 0 1 2 3 4 3 4 15

0 miss

00	Mem(0)

1 miss

00	Mem(0)
00	Mem(1)

2 miss

00	Mem(0)
00	Mem(1)
00	Mem(2)

3 miss

00	Mem(0)
00	Mem(1)
00	Mem(2)
00	Mem(3)

4 miss

01 4

00	Mem(0)
00	Mem(1)
00	Mem(2)
00	Mem(3)

3 hit

01	Mem(4)
00	Mem(1)
00	Mem(2)
00	Mem(3)

4 hit

01	Mem(4)
00	Mem(1)
00	Mem(2)
00	Mem(3)

15 miss

11 15

01	Mem(4)
00	Mem(1)
00	Mem(2)
00	Mem(3)

مثال

- یک حافظه‌ی نهان با شرایط زیر مفروض است:
  - 4K blocks, 4-word block size, 32 bit address
- تعداد مجموعه‌ها و طول برچسب را برای حالات زیر حساب کنید؟
  - نگاشت مستقیم، حافظه‌ی نهان اشتراکی دوبلوکی، حافظه‌ی نهان اشتراکی چهار بلوکی و حافظه‌ی اشتراکی کامل به دست آورید.

نگاشت مستقیم

16(=2^4) byte per block

برای آدرس شاخص و برچسب 32-4=28

log2(4K)=12

تعداد بلوک‌ها در نگاشت مستقیم طول شاخص را مشخص می‌کنند

28-12=16

تعداد بیت‌های برچسب

مثال (ادامه...)

با افزایش درجه‌ی اشتراک، تعداد بیت‌های شاخص کاهش یافته و بیت‌های برچسب افزایش خواهد یافت.
بنابراین برای اشتراک با دو بلوک 2K مجموعه خواهیم داشت.

28-log2(2K)=17

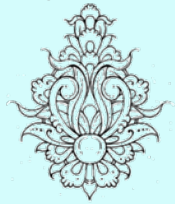
2-way associative

28-log2(1K)=18

4-way associative

28

fully associative



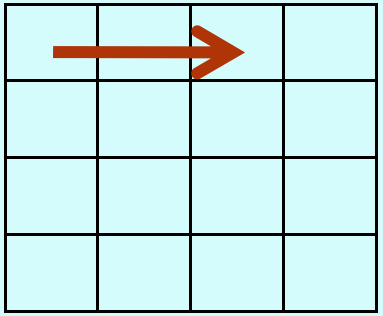
مثال

```
for( i=0; i<N; i++ )
    for( j=0; j<N; j++ )
        for( k=0; k<N; k++ )
            c[i][j] += a[i][k] * b[k][j];
```

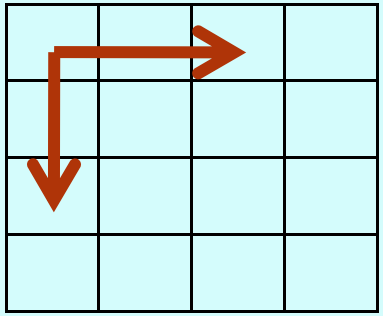
real	**0m10.688s**
user	0m10.581s
sys	0m0.068s

real	**0m5.730s**
user	0m5.668s
sys	0m0.052s

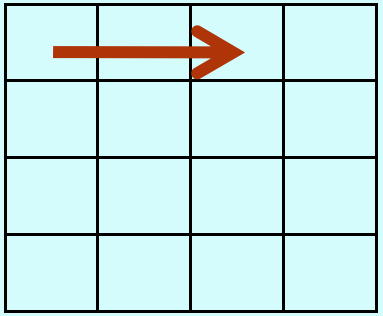
A

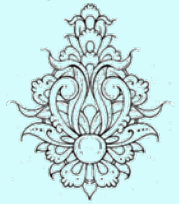


B



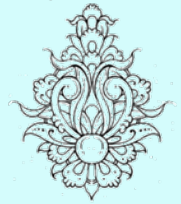
C





حافظه‌ی نهان چند سطحی

- حافظه‌های نهان متصل به پردازنده‌ها
  - کوچک، اما بسیار سریع هستند.
- حافظه‌ی نهان سطح ۲ (levlel 2 cache)
  - در صورتی که در حافظه‌ی نهان سطح ۱ داده موجود نباشد، این سطح پاسخگو خواهد بود.
  - بزرگ‌تر، اما کندتر هستند، ولی در هر حال از حافظه‌ی اصلی سریع‌تر هستند.
- حافظه‌ی اصلی پاسخگوی نبود داده در حافظه‌ی نهان سطح ۲ می‌باشد.



مثال

- سیستمی با مشخصات زیر مفروض است:
  - CPU base CPI = 1, clock rate = 4GHz
  - Miss rate/instruction = 2%
  - Main memory access time = 100ns
- در صورتی که از یک سطح حافظه‌ی نهان استفاده کنیم:

Miss penalty = 100ns/0.25ns = 400 cycles

Effective CPI = 1 + 0.02 × 400 = 9

مثال (ادامه...)

- با افزودن یک سطح دیگر حافظه‌ی نهان با مشخصات زیر:
  - Access time = 5ns
  - Global miss rate to main memory = 0.5%

Penalty = 5ns/0.25ns = 20 cycles

CPI = 1 + 0.02 × 20 + 0.005 × 400 = 3.4

Performance ratio = 9/3.4 = 2.6

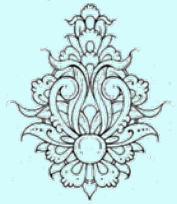
Total CPI= Base CPI+ Primary stalls per instruction

+

Secondary stalls per instruction

حافظه‌ی نهان چند سطحی

- حافظه‌ی نهان سطح ۱
  - تمرکز بر روی hit time
- حافظه‌ی سطح ۲
  - تمرکز بر روی کاهش miss rate است



حافظه‌ی نهان در دو پردازنده‌ی واقعی

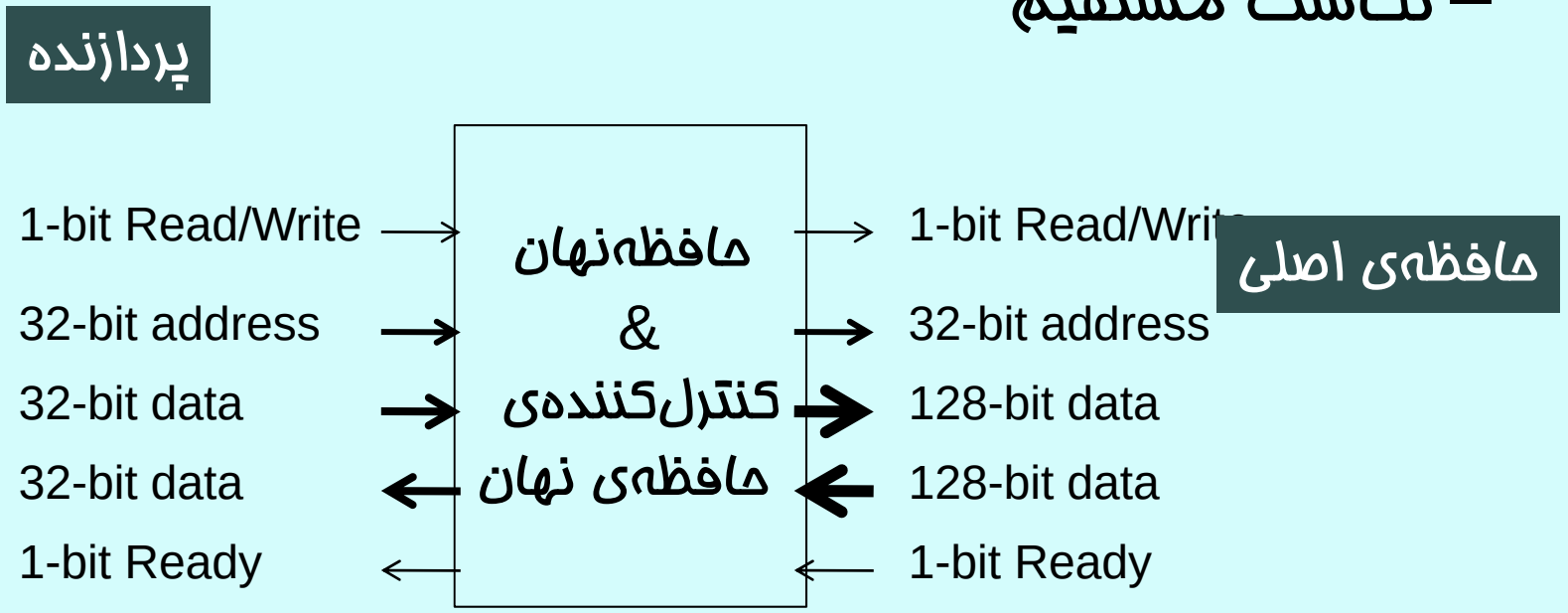
	Intel P4	AMD Opteron
L1 organization	Split I$ and D$	Split I$ and D$
L1 cache size	8KB for D$, 96KB for trace cache (~I$)	64KB for each of I$ and D$
L1 block size	64 bytes	64 bytes
L1 associativity	4-way set assoc.	2-way set assoc.
L1 replacement	~ LRU	LRU
L1 write policy	write-through	write-back
L2 organization	Unified	Unified
L2 cache size	512KB	1024KB (1MB)
L2 block size	128 bytes	64 bytes
L2 associativity	8-way set assoc.	16-way set assoc.
L2 replacement	~LRU	~LRU
L2 write policy	write-back	write-back

حافظه‌ی نهان در دو پردازنده‌ی واقعی ۲

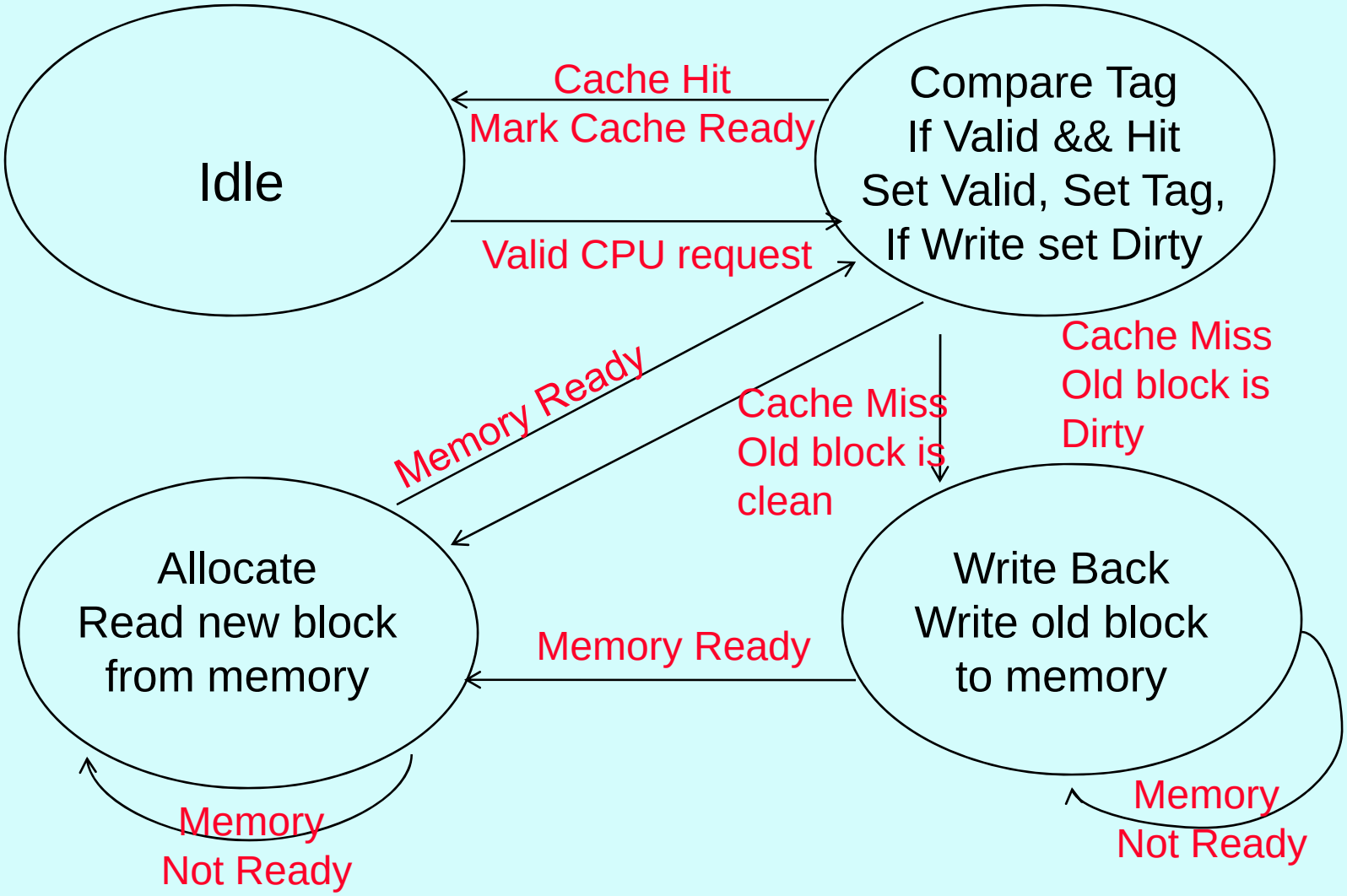
	Intel Nehalem	AMD Barcelona
L1 cache organization & size	Split I$ and D$; 32KB for each per core; 64B blocks	Split I$ and D$; 64KB for each per core; 64B blocks
L1 associativity	4-way (I), 8-way (D) set assoc.; ~LRU replacement	2-way set assoc.; LRU replacement
L1 write policy	write-back, write-allocate	write-back, write-allocate
L2 cache organization & size	Unified; 256MB (0.25MB) per core; 64B blocks	Unified; 512KB (0.5MB) per core; 64B blocks
L2 associativity	8-way set assoc.; ~LRU	16-way set assoc.; ~LRU
L2 write policy	write-back	write-back
L2 write policy	write-back, write-allocate	write-back, write-allocate
L3 cache organization & size	Unified; 8192KB (8MB) shared by cores; 64B blocks	Unified; 2048KB (2MB) shared by cores; 64B blocks
L3 associativity	16-way set assoc.	32-way set assoc.; evict block shared by fewest cores
L3 write policy	write-back, write-allocate	write-back; write-allocate

کنترل حافظه‌ی نهان

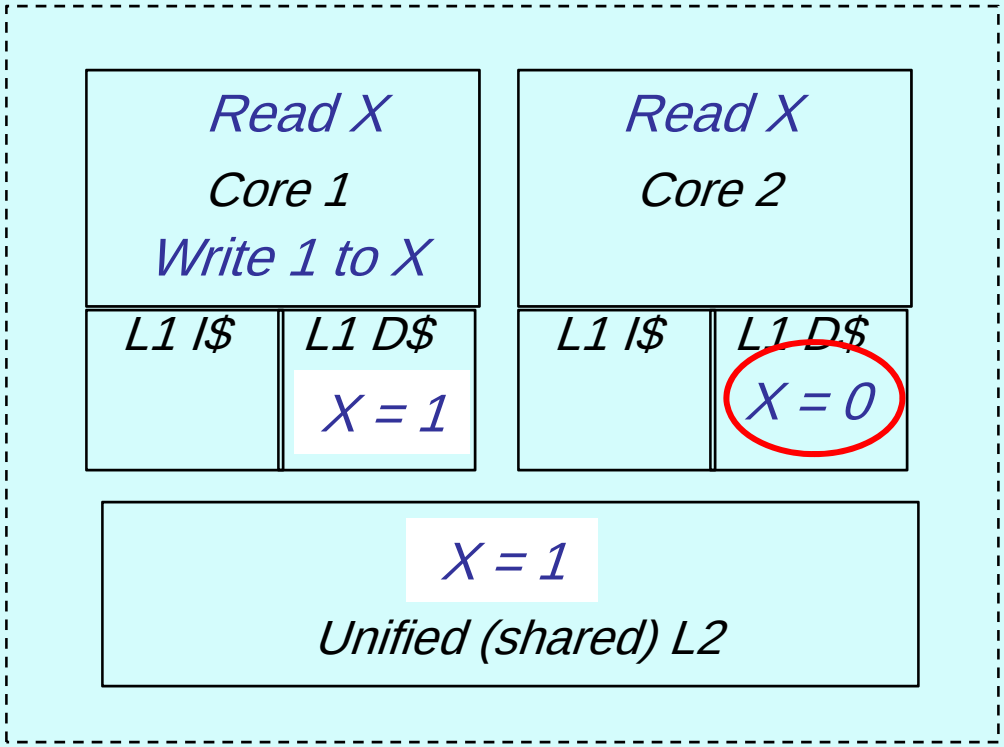
- **یک حافظه‌ی نهان با مشخصات زیر مفروض است:**
  - Write back –
  - اندازه‌ی بلوک‌ها چهار کلمه –
  - اندازه‌ی حافظه‌ی نهان 16KB –
  - نگاشت مستقیم –



نمودار حالت کنترل حافظه‌ی نهان

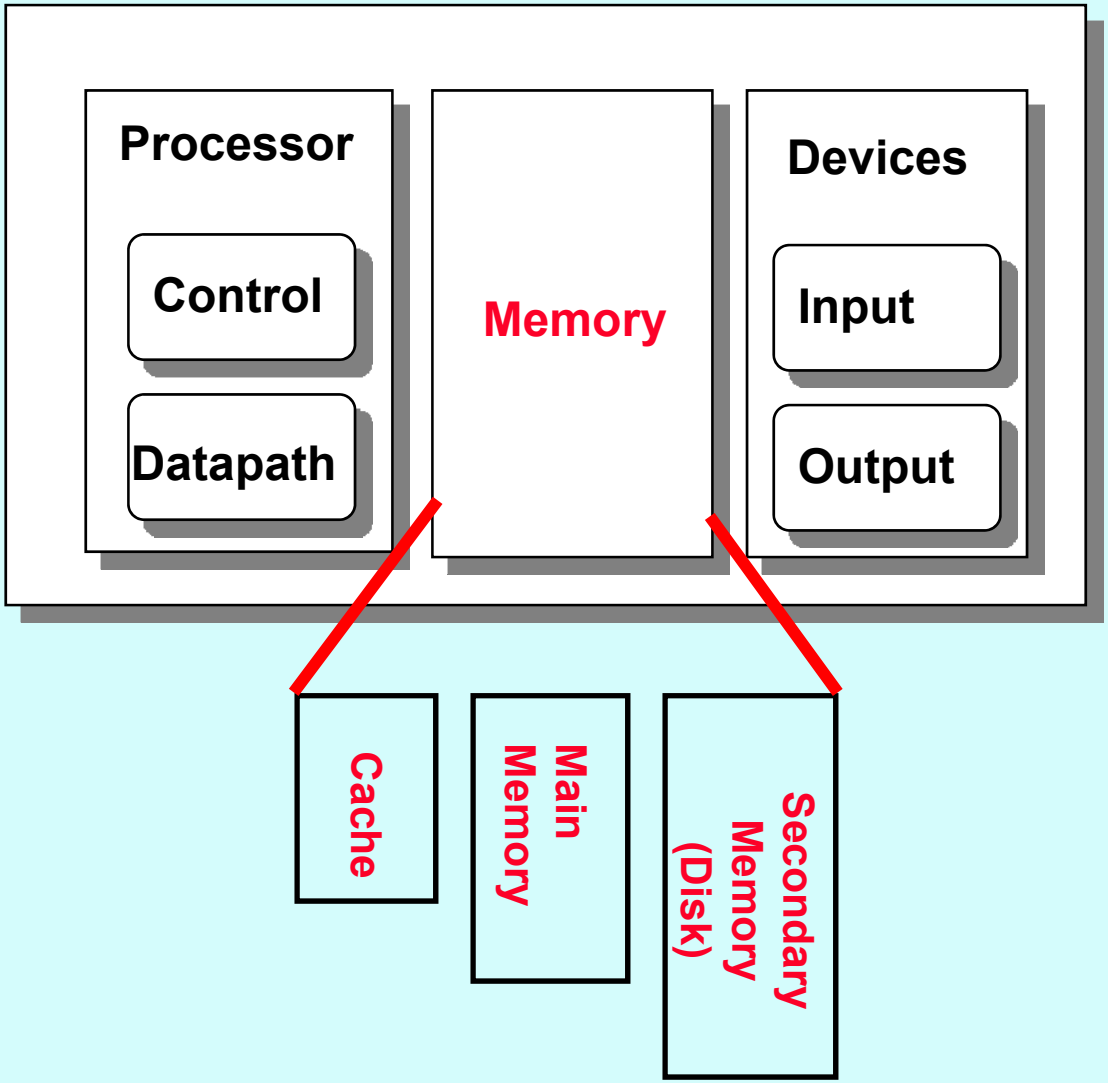


حافظه‌ی نهان در پردازنده‌های چند هسته‌ای



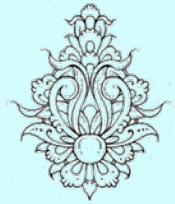
cache coherence problem

ساختار کلی یک کامپیوتر



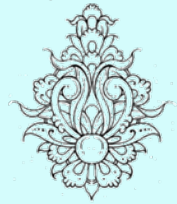
سلسله مراتب در حافظه‌ی اصلی

- ثبات‌ها و حافظه‌ی نهان
  - کامپایلر یا برنامه‌نویس
- حافظه‌ی نهان و حافظه‌ی اصلی
  - کنترل‌کننده‌ی حافظه‌ی نهان
- حافظه‌ی اصلی و حافظه‌ی ثانویه



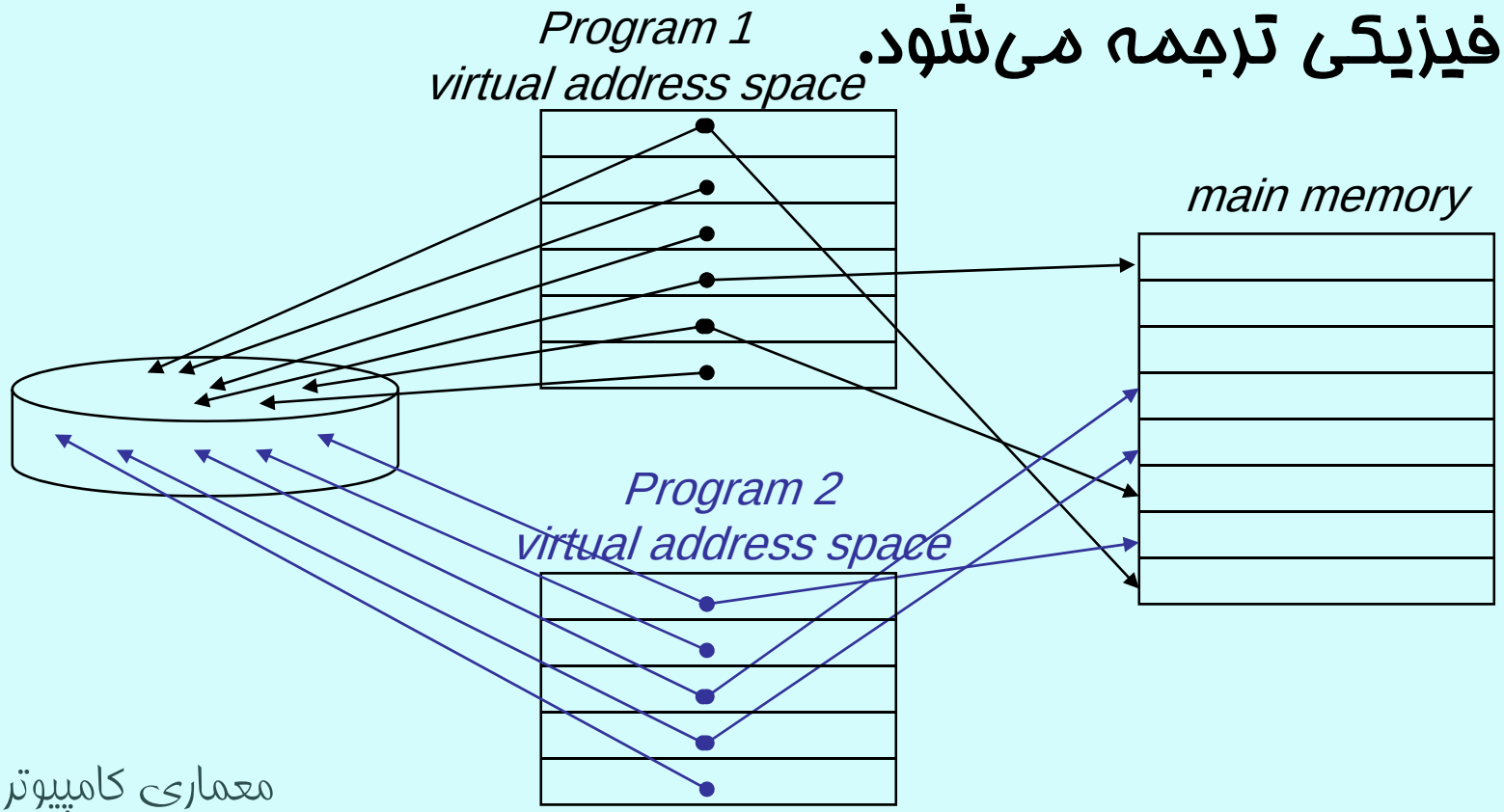
حافظه‌ی مجازی

- حافظه‌ی اصلی نقشی مانند حافظه‌ی نهان را برای حافظه‌ی اصلی ایفا می‌کند.
  - مدیریت آن به صورت مشترک توسط پردازنده و سیستم‌عامل صورت می‌پذیرد.
- با کمک آن می‌توان به گونه‌ای کارا و امن حافظه را بین چندین برنامه به اشتراک گذاشت.
- می‌توان به کمک آن برنامه‌هایی را اجرا کرد، که دارای حجمی بیش از حجم حافظه‌ی فیزیکی هستند.
- بارگذاری برنامه در حافظه با سهولت بیش‌تری صورت می‌گیرد.



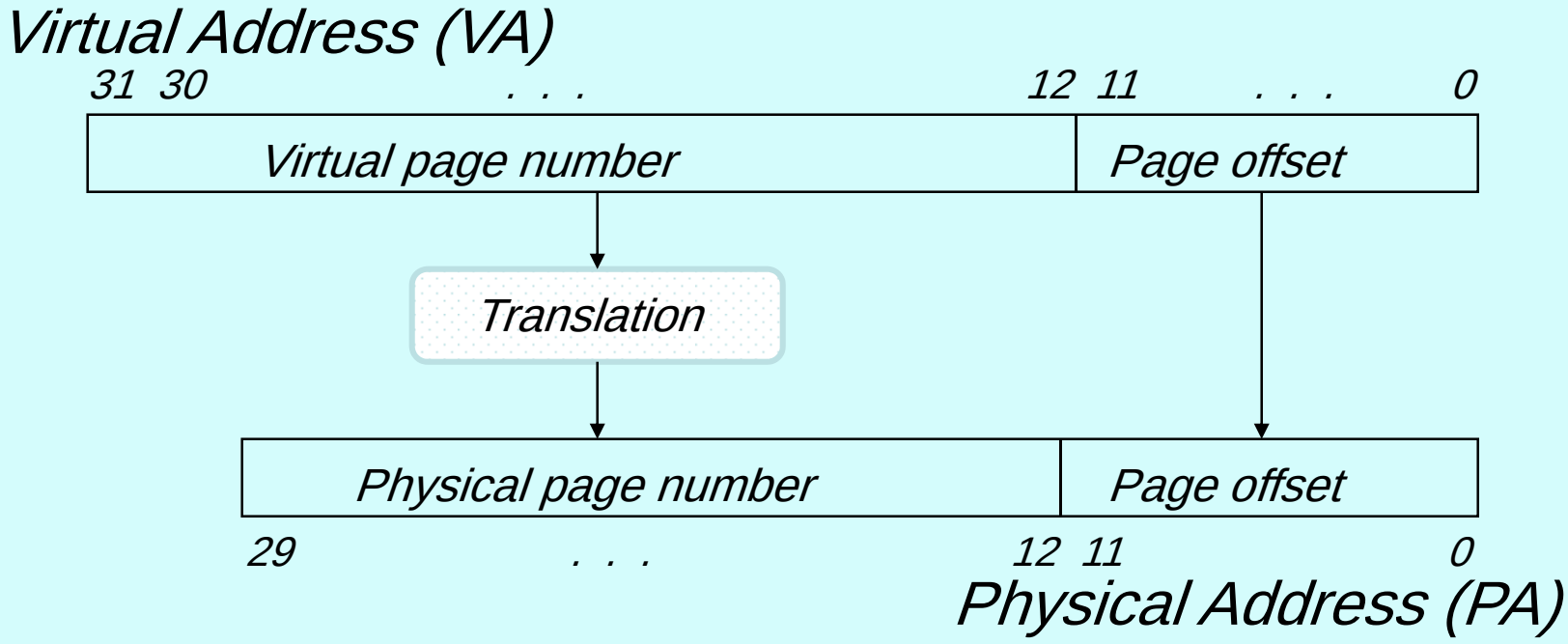
حافظه‌ی مجازی

- در واقع به هر برنامه در زمان کامپایل فضایی اختصاص داده می‌شود.
- در هنگام اجرای برنامه آدرس مجازی به آدری فیزیکی ترجمه می‌شود.

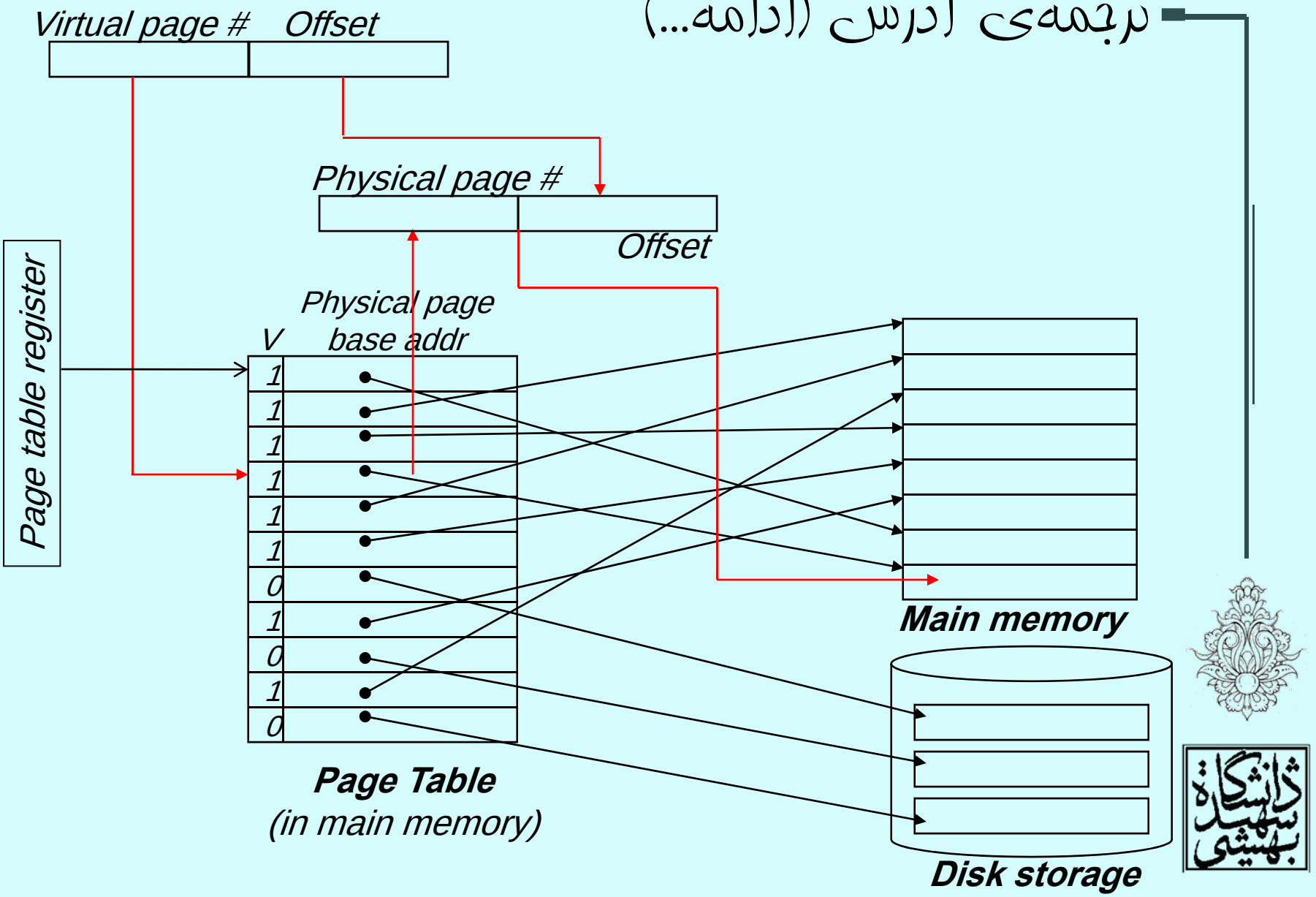


ترجمه‌ی آدرس

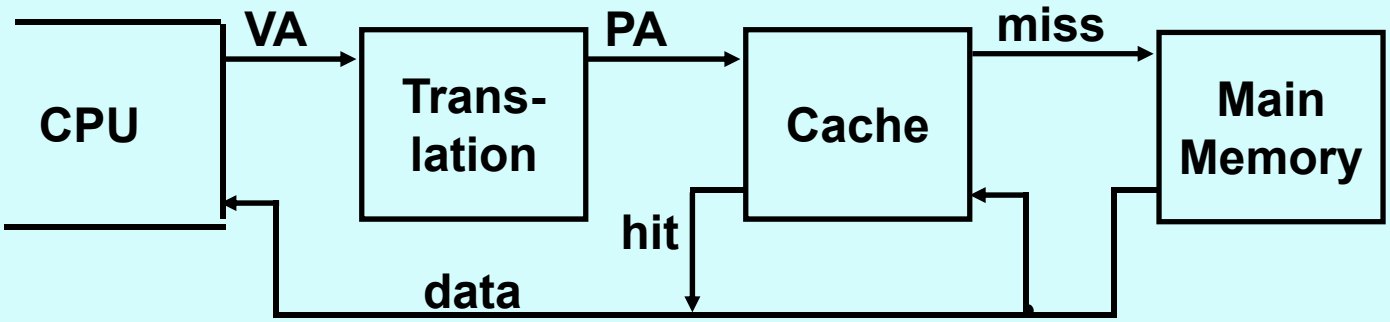
- ترجمه‌ی آدرس با همکاری پردازنده و سیستم‌عامل صورت می‌پذیرد.
- در صورتی که داده در حافظه‌ی اصلی نباشد، «page fault» رخ می‌دهد.



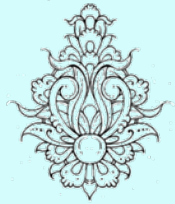
ترجمه‌ی آدرس (ادامه...)



ترجمه‌ی آدرس (ادامه...)



- با این حساب عمل دستیابی به حافظه نهان خیلی زمان‌بر خواهد شد!
- با کمک سخت‌افزار و در نظر گرفتن یک میان‌گیر این مشکل برطرف می‌شود.



Translation Lookaside Buffer

ترجمه‌ی آدرس (ادامه...)