معماری کامپیوتر (۱۳۹۱-۱۱-۱۳)

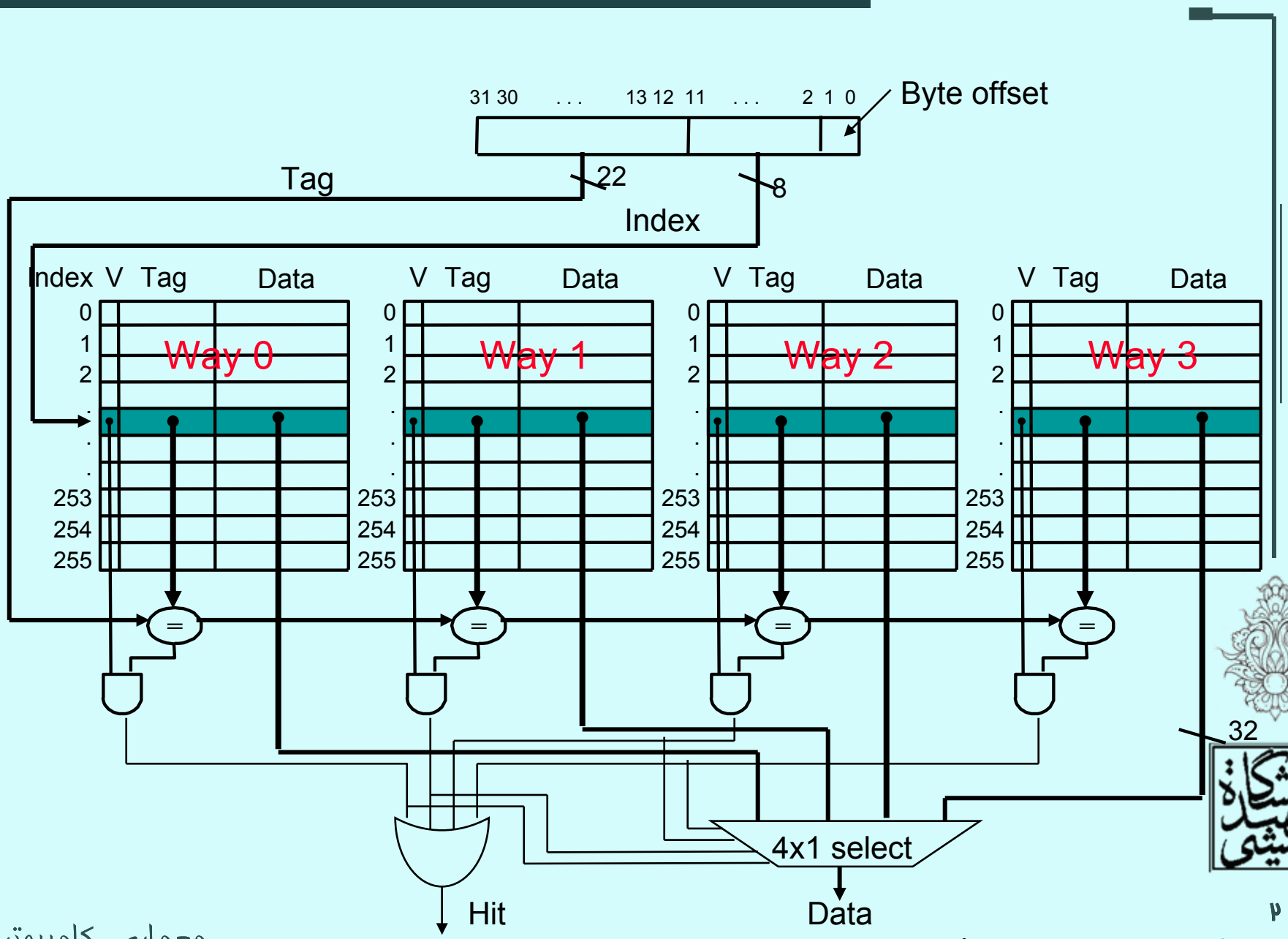
جلسه‌ی بیست و سوم

دانشگاه شهید بهشتی
دانشکده‌ی مهندسی برق و کامپیوتر
بهار ۱۳۹۱
احمد محمودی ازناوه

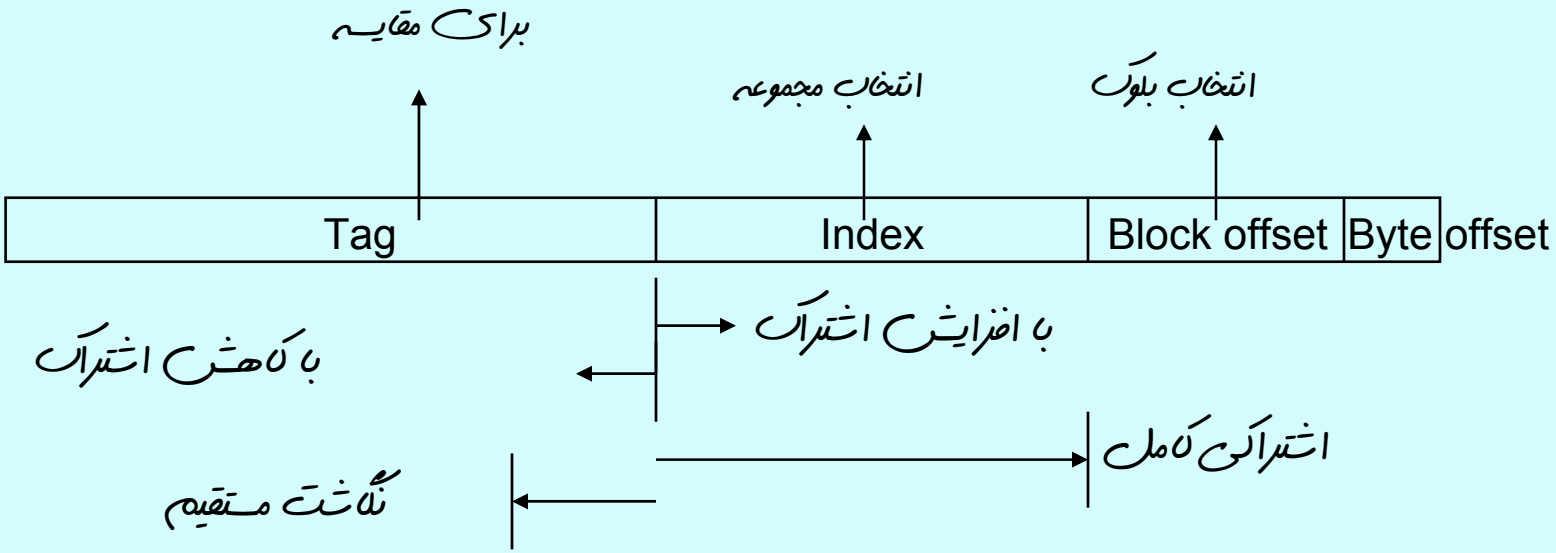
فهرست مطالب

- حافظه‌ی تداعی‌گر
- سیاست‌های جایگزینی
- حافظه‌ی نهان چند سطحی
- حافظه‌ی مجازی

Four-Way Set Associative Cache



طیف اشتراک (ادامه...)



Content Addressable Memory (CAM)

Associative memory

حافظه‌ی انجمنی!!

حافظه‌ی تداعی‌گر

- در بسیاری کاربردها، لازم است یک آیتم در یک جدول (حافظه) جستجو شود، که فرآیندی زمان‌بر است.
- در صورتی که بتوان حافظه‌ای ساخت که با ارائه‌ی داده، آدرس را بیابد، کارایی فرآیند جستجو به صورت مؤثری بهبود خواهد یافت. چنین حافظه‌ای، حافظه‌ی تداعی‌گر خوانده می‌شود.
- این نوع حافظه‌ها، هزینه‌ی بالاتر نسبت به حافظه‌های معمولی دارند، و بدین‌سبب در کاربردهایی که زمان جستجو نقشی حیاتی دارد، به کار می‌روند.

در حافظه‌های نهان انجمنی، علاوه بر داده، (بخشی) از آدرس را نیز ذخیره می‌کند. چنین حافظه‌ای از یک حافظه‌ی معمولی و یک حافظه‌ی تداعی‌گر تشکیل شده است.

Replacement Policy

سیاست جایگزینی

- در نگاشت مستقیم، جایی که بلوک باید در آن قرار گیرد مشخص است.
- در روش‌های اشتراکی بلوک در چند محل متفاوت می‌تواند قرار گیرد.
  - در درجه‌ی اول مکانی انتخاب می‌شود که بیت‌اعتبار آن غیر فعال است.
  - در غیر این صورت از بلوکی که کم‌تر مورد استفاده قرار گرفته است، از حافظه‌ی نهان خارج می‌شود.

هر چه تعداد مجموعه‌های مشترک افزایش یابد، هزینه‌ی سخت‌افزاری LRU افزایش خواهد یافت.

Least-recently used (LRU)

راه دیگر، انتخاب تصادفی است

برای حالت اشتراک با مجموعه‌های بزرگ کارایی یکسانی با LRU دارد

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 شخص 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

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

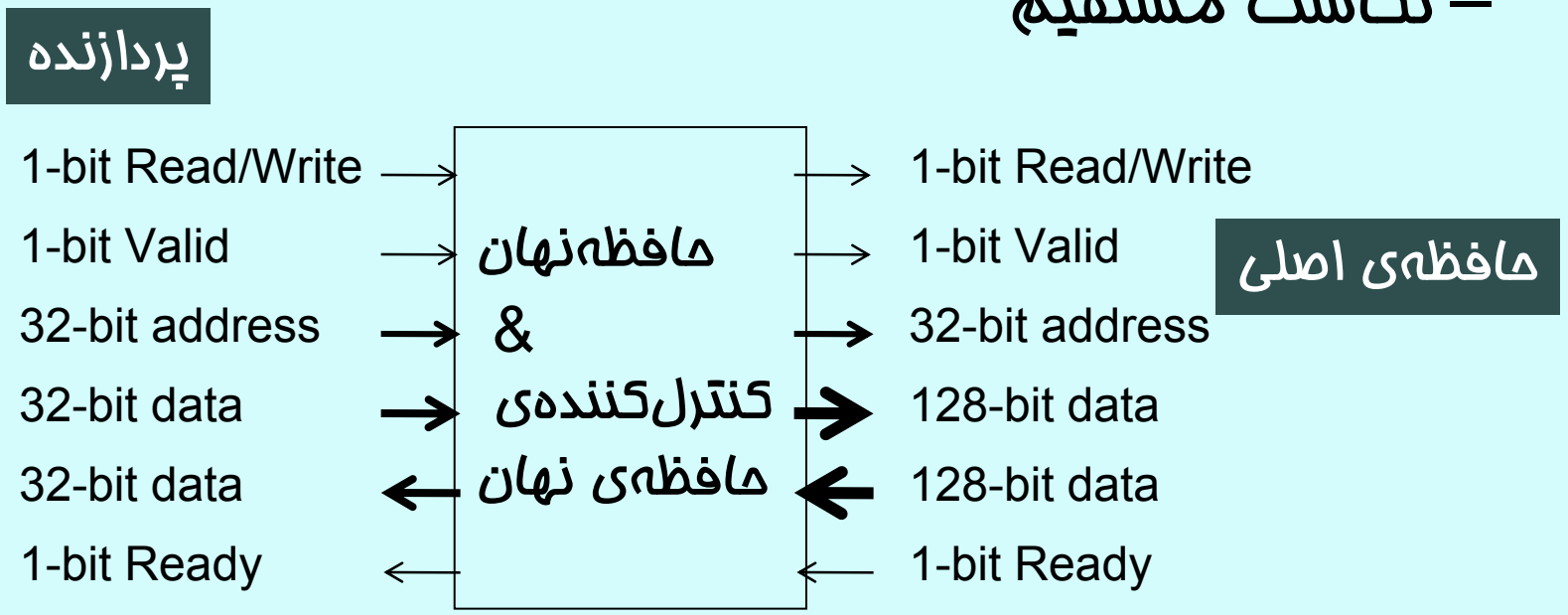
- **حافظه‌ی نهان سطح ۱**
  - تمرکز بر روی hit time
- **حافظه‌ی سطح ۲**
  - تمرکز بر روی کاهش miss rate است

حافظه‌ی نهان در دو پردازنده‌ی واقعی

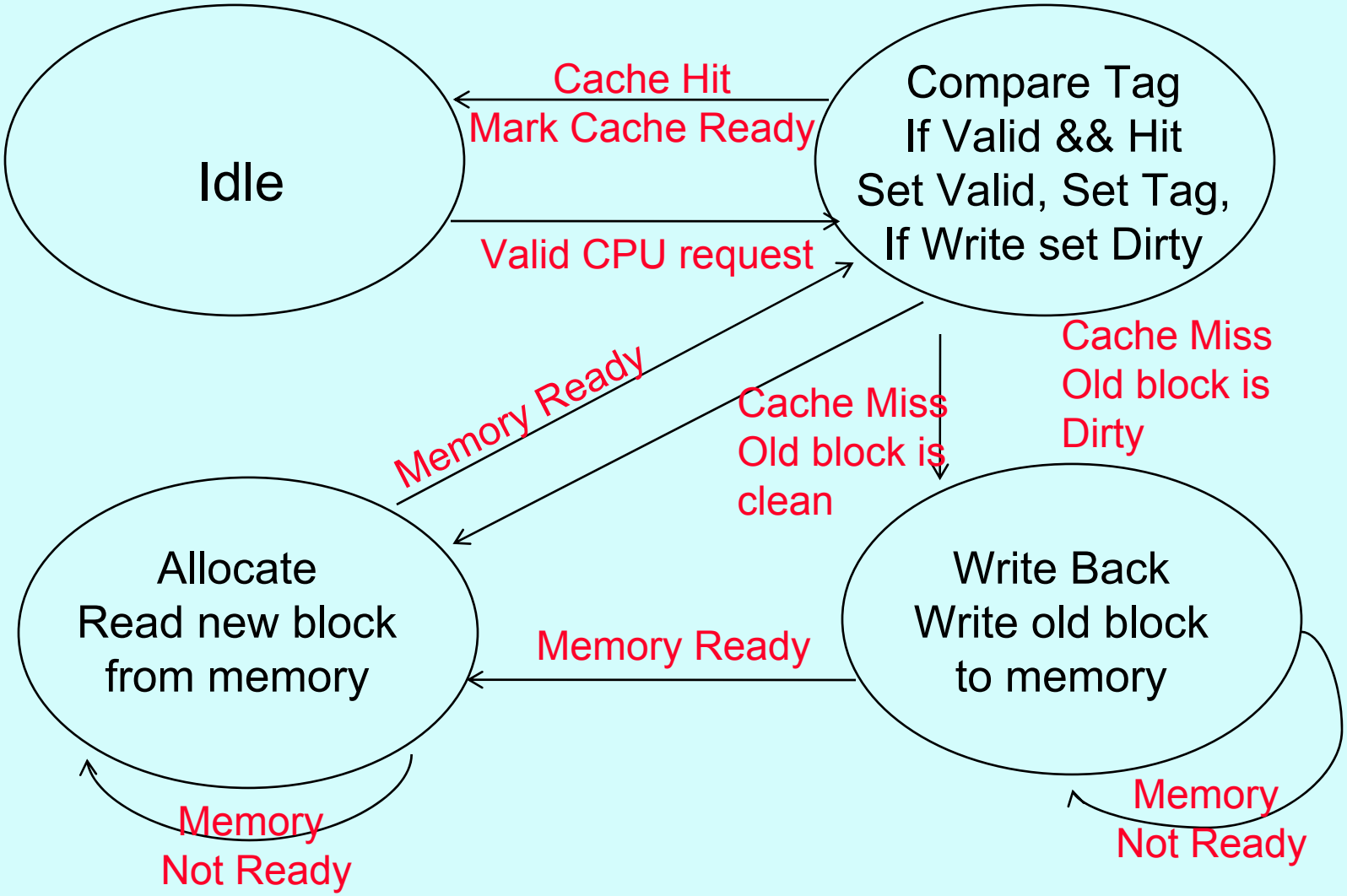
	Intel Nehalem	AMD Barcelona
L1 cache organization & size	Split I$ and D$; 32KB for each per core; 64B blocks	Split I$ and D$; 64KB for each per core; 64B blocks
L1 associativity	4-way (I), 8-way (D) set assoc.; ~LRU replacement	2-way set assoc.; LRU replacement
L1 write policy	write-back, write-allocate	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

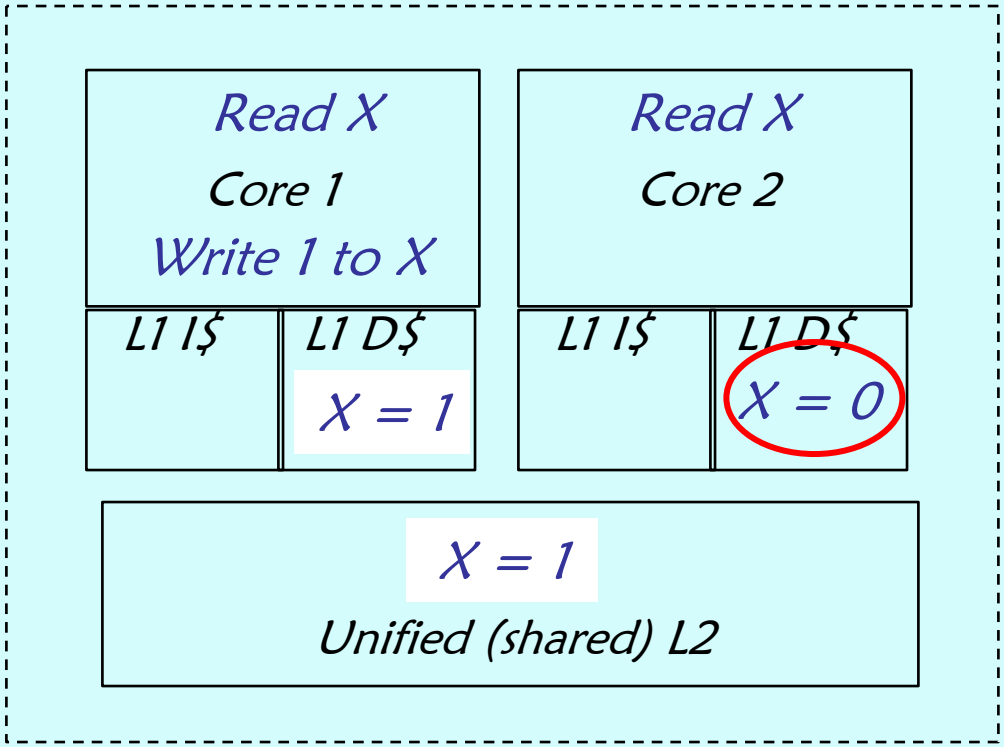
- یک حافظه‌ی نهان با مشخصات زیر مفروض است:
  - Write back
  - اندازه‌ی بلوک‌ها چهار کلمه
  - اندازه‌ی حافظه‌ی نهان 16KB
  - نگاشت مستقیم



نمودار حالت کنترل حافظه‌ی نهان

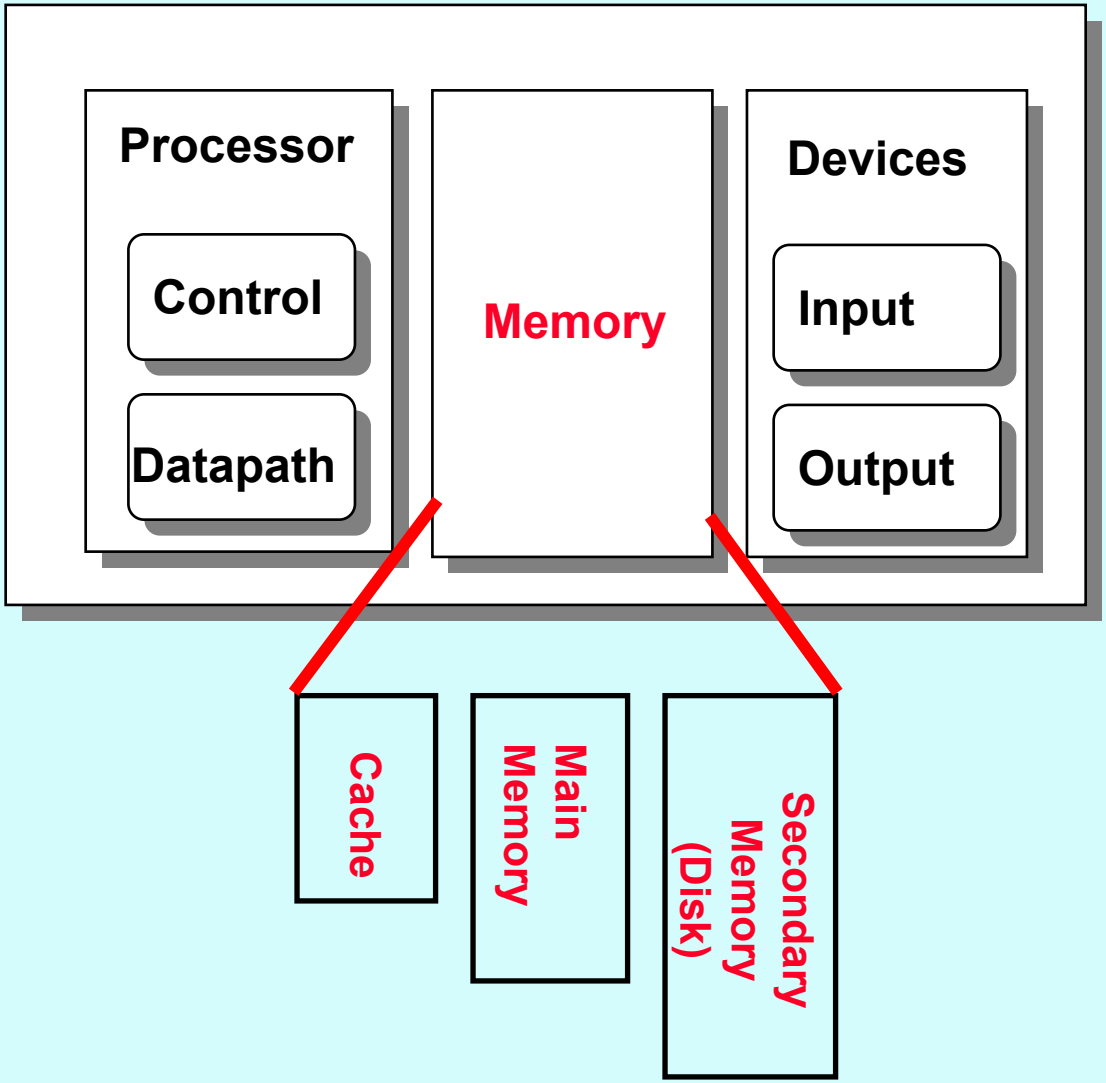


حافظه‌ی نهان در پردازنده‌های چند هسته‌ای



cache coherence problem

ساختار کلی یک کامپیوتر



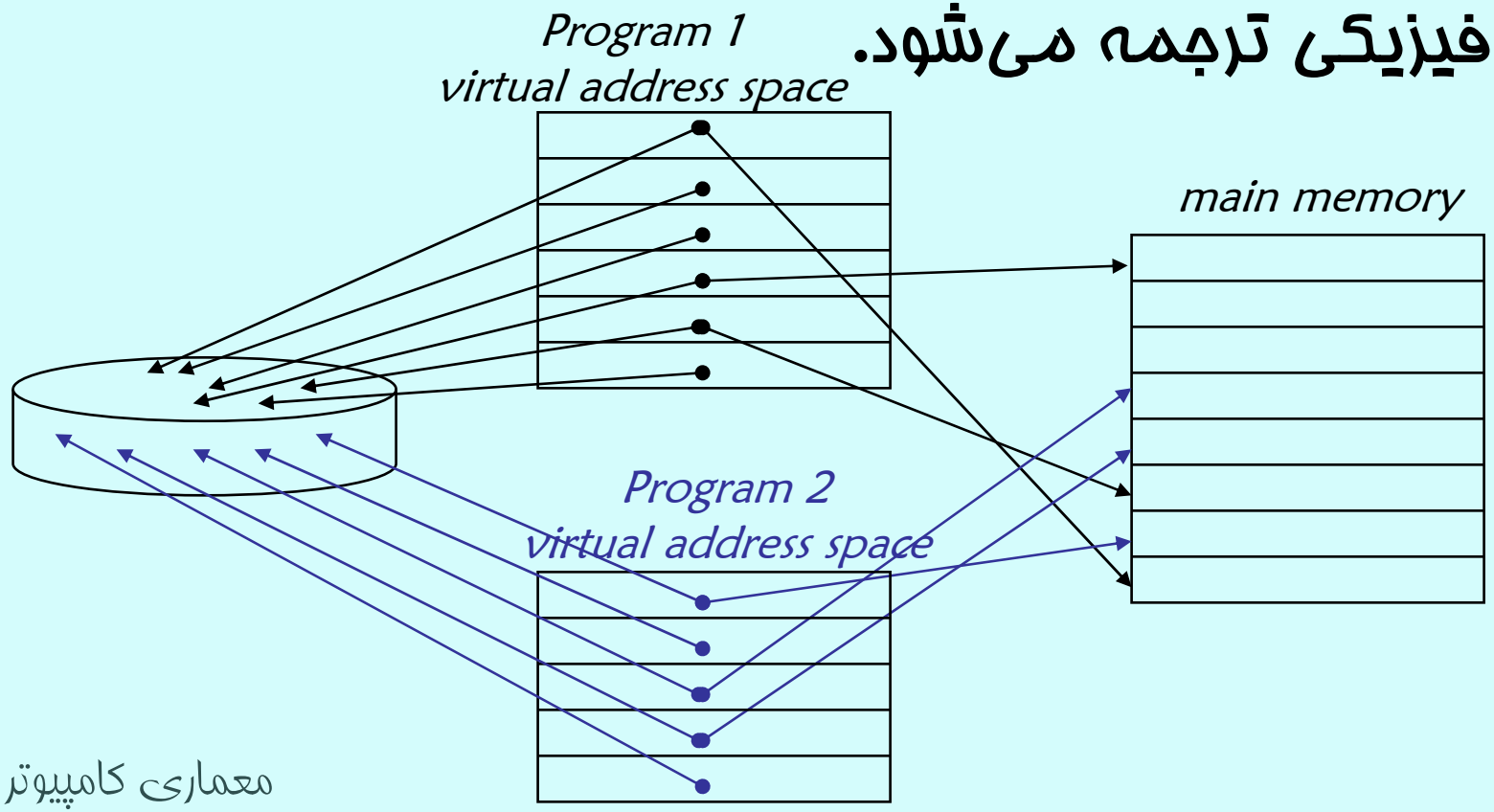
سلسله مراتب در حافظه‌ی اصلی

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

حافظه‌ی مجازی

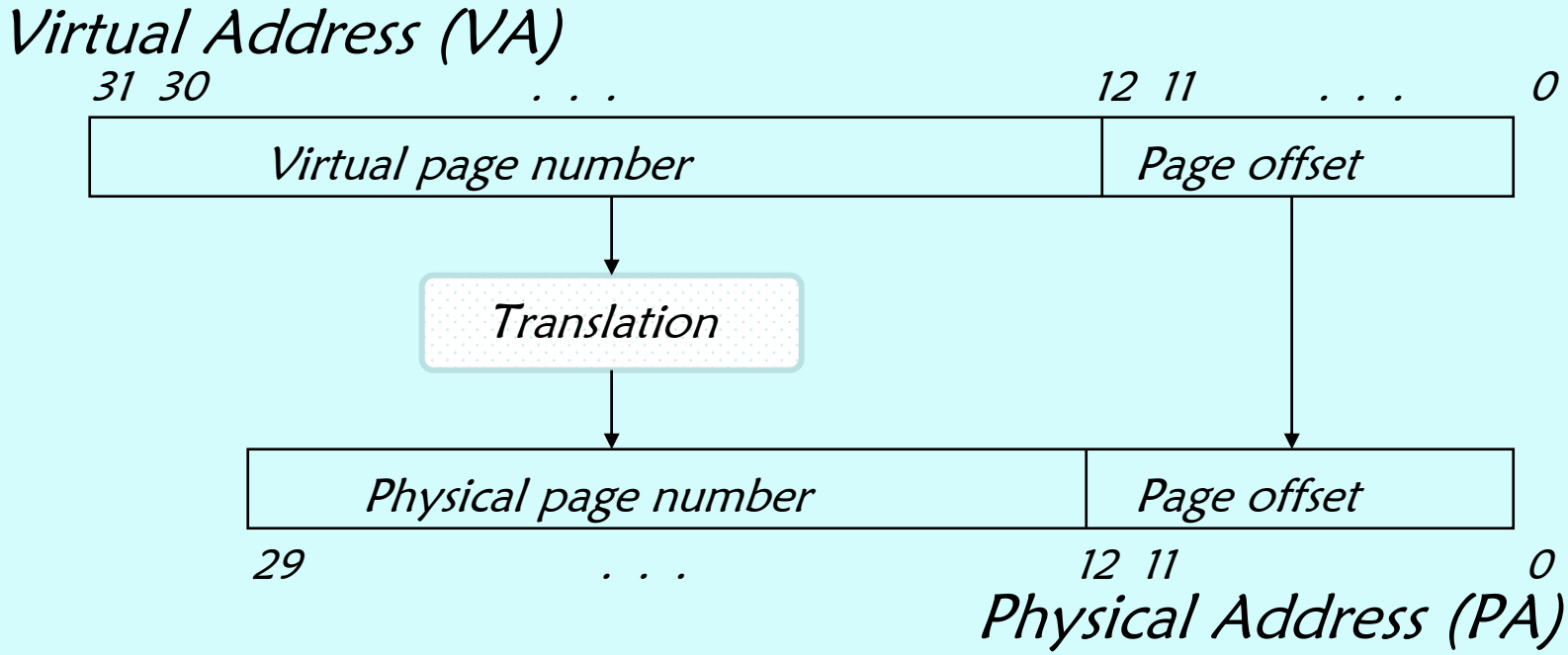
- حافظه‌ی اصلی نقشی مانند حافظه‌ی نهان را برای حافظه‌ی اصلی ایفا می‌کند.
  - مدیریت آن به صورت مشترک توسط پردازنده و سیستم‌عامل صورت می‌پذیرد.
- با کمک آن می‌توان به گونه‌ای کارا و امن حافظه را بین چندین برنامه به اشتراک گذاشت.
- می‌توان به کمک آن برنامه‌هایی را اجرا کرد، که دارای حجمی بیش از حجم حافظه‌ی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

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

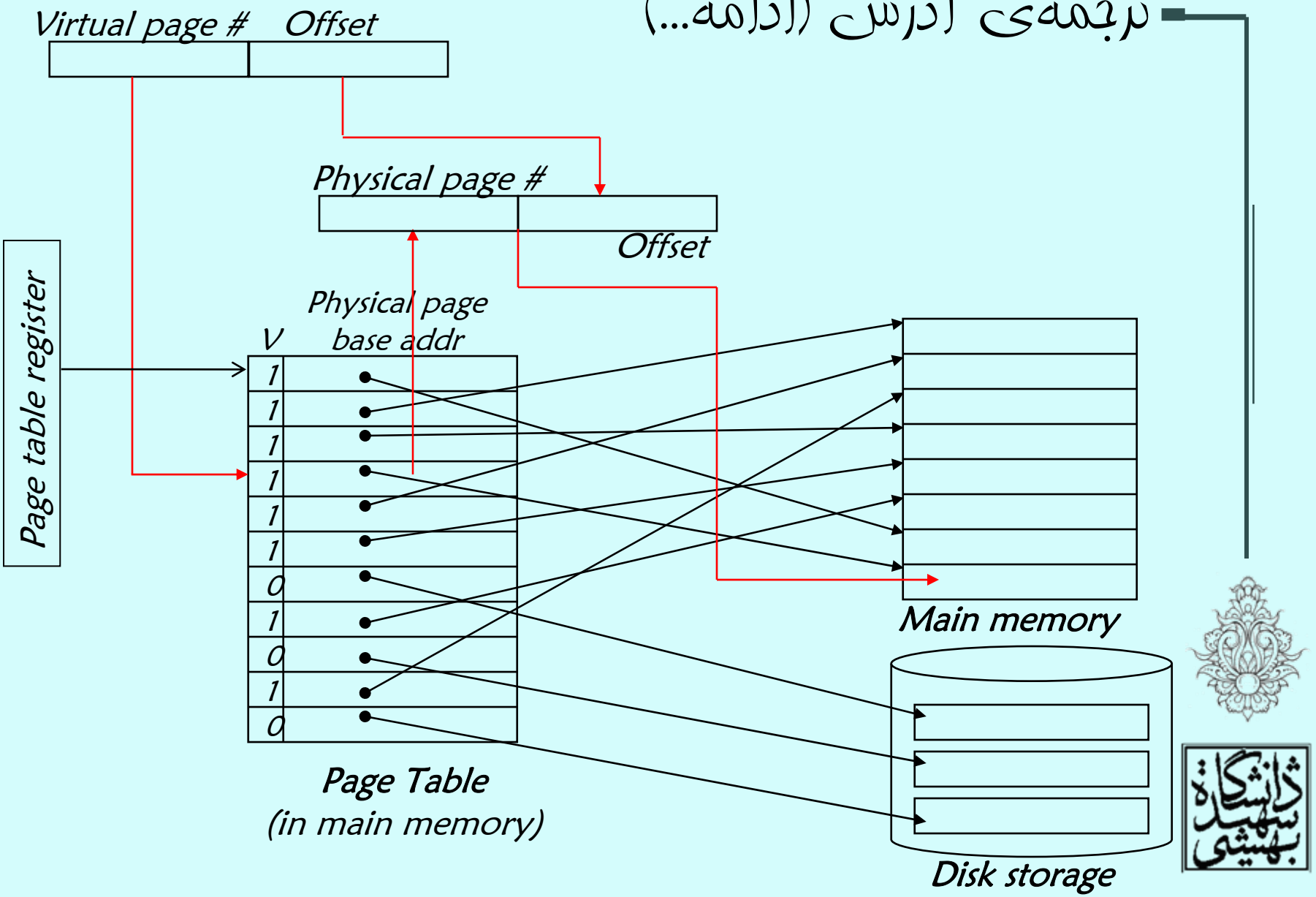


ترجمه‌ی آدرس

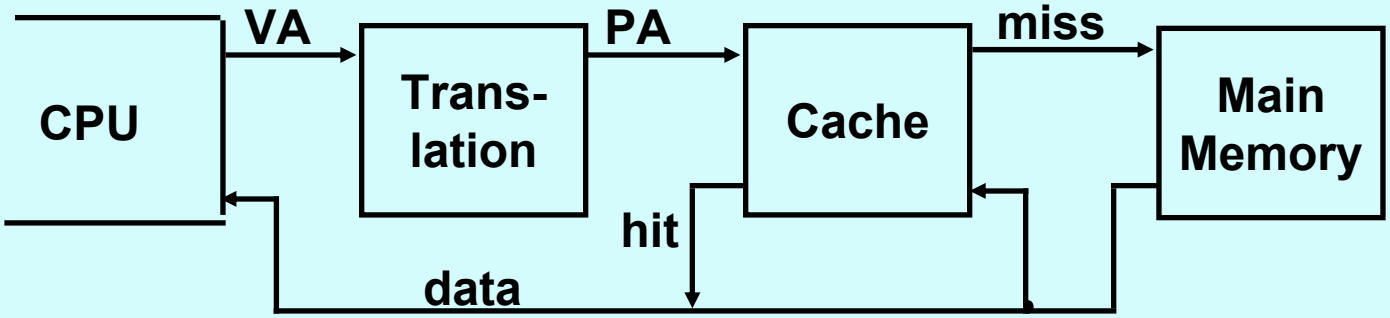
- ترجمه‌ی آدرس با همکاری پردازنده و سیستم‌عامل صورت می‌پذیرد.
- در صورتی که داده در حافظه‌ی اصلی نباشد، «page fault» رخ می‌دهد.



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...)