



VIETNAM NATIONAL UNIVERSITY HANOI (VNU)
INFORMATION TECHNOLOGY INSTITUTE

Computer Architecture

Lecture 5: Caches

Duy-Hieu Bui, PhD

VNU Information Technology Institute
Laboratory for Smart Integrated System (SISLAB)

Email: hieubd@vnu.edu.vn

<https://duyhieubui.github.io>

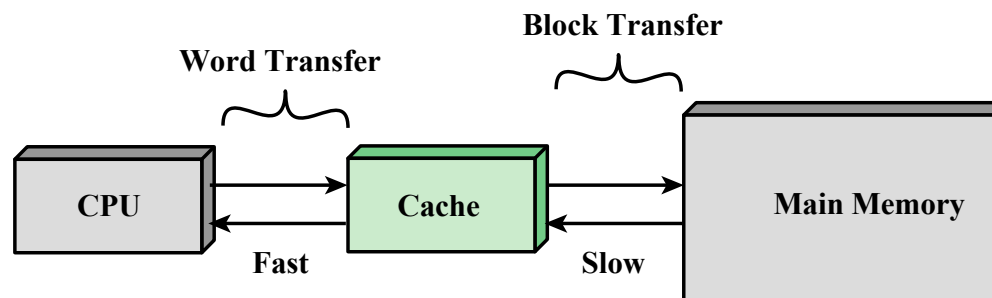


Nội dung

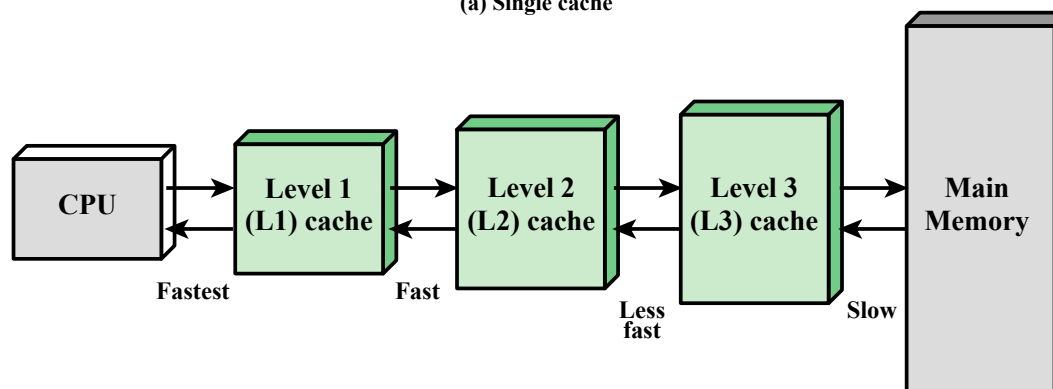
1. Tổ chức máy tính với bộ nhớ cache
2. Các phép ánh xạ cache
 - Ánh xạ trực tiếp
 - Ánh xạ kết hợp
 - Ánh xạ kết hợp theo tập hợp
3. Luật thay thế
4. Thiết kế cache
5. Hiệu năng cache

Tham khảo chương 5, “Computer Organization and Architecture: Designing for Performance”, William Stallings, 10th edition

- Bộ nhớ có tốc độ truy cập nhanh
- Thường nằm giữa main memory và CPU
- Có thể nằm trong CPU chip hoặc một module riêng



(a) Single cache

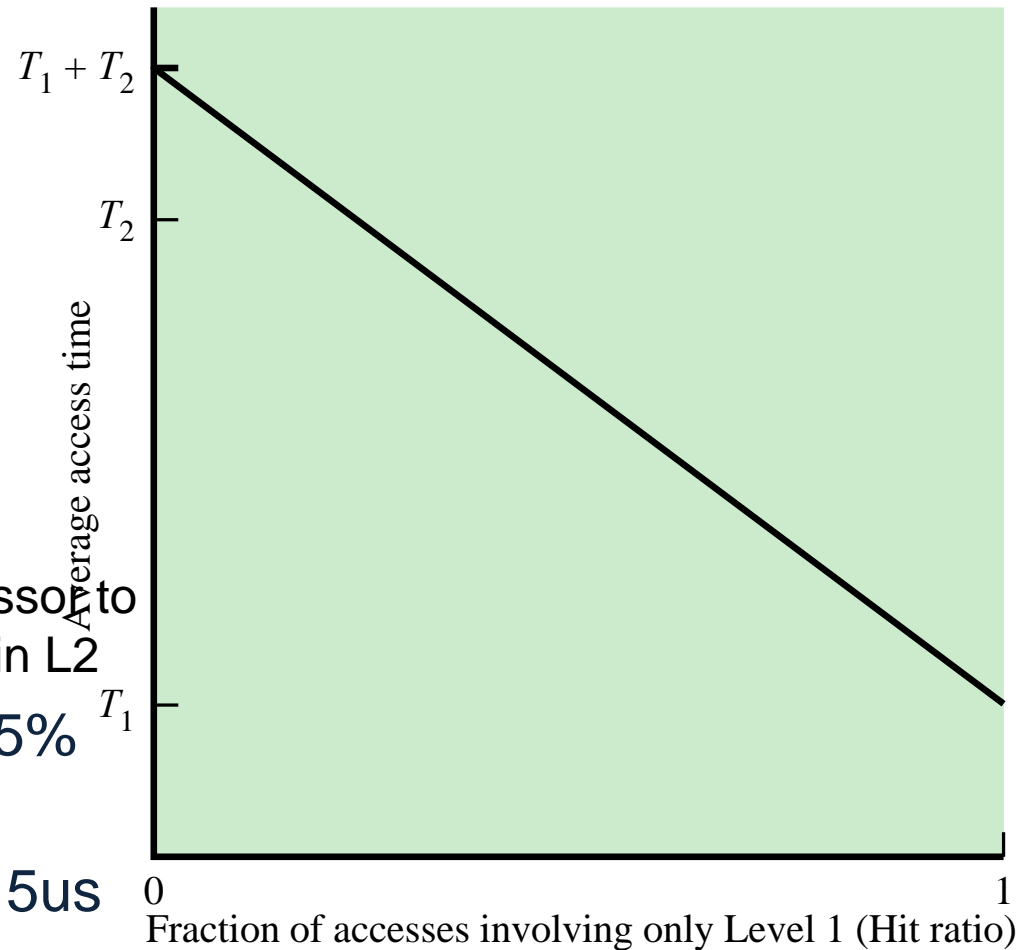


(b) Three-level cache organization

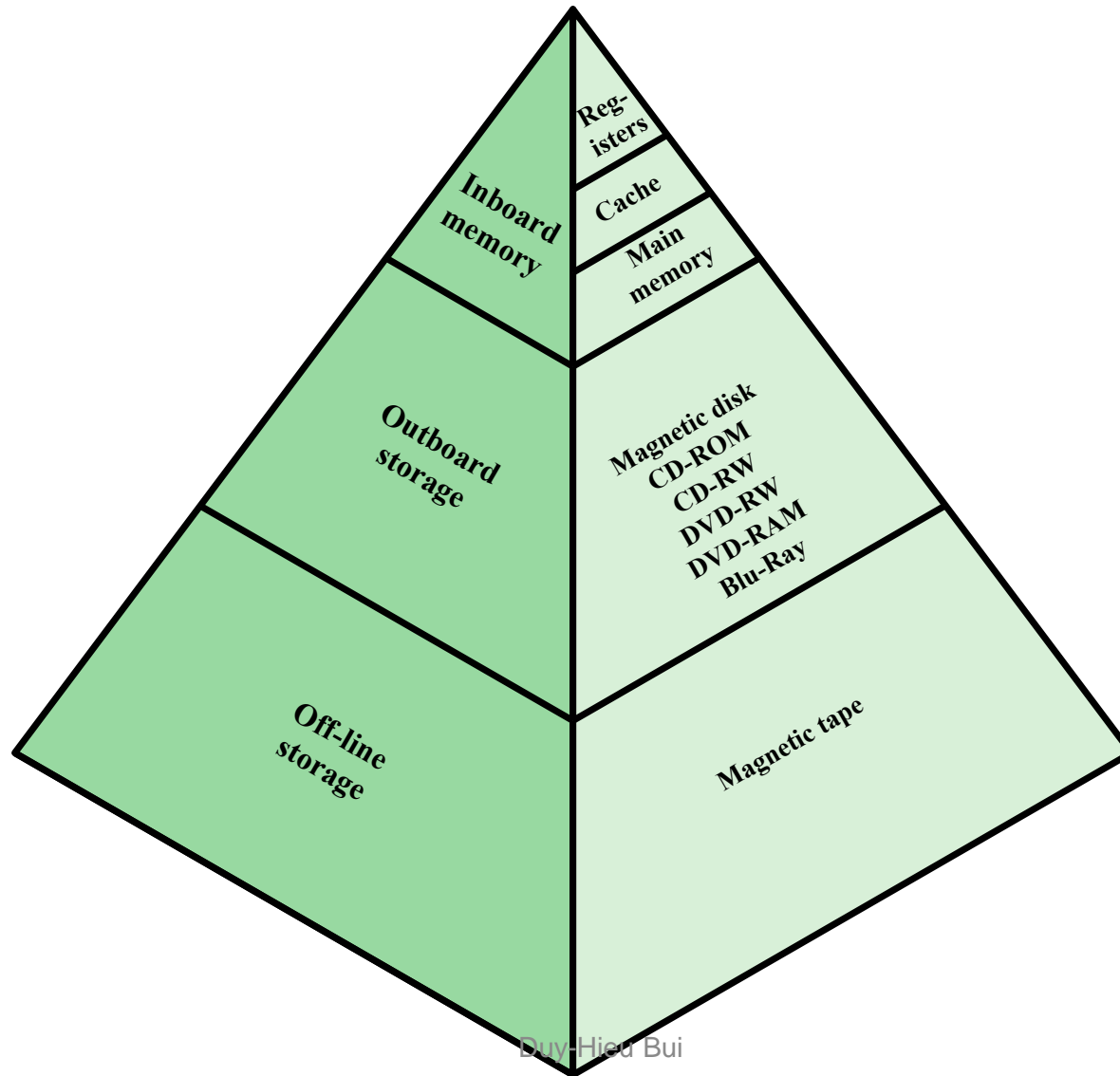
Cache example

- Two levels of memory
- L1: 1000 words
 - Access time: $T_1=0.01\mu s$
- L2: 100,000 words
 - Access time: $T_2=0.1\mu s$
- Access rule:
 - L1: direct processing
 - L2: transfer to L1 first
 - Ignore the time for processor to decide the word in L1 or in L2
- Average access time: 95% hit

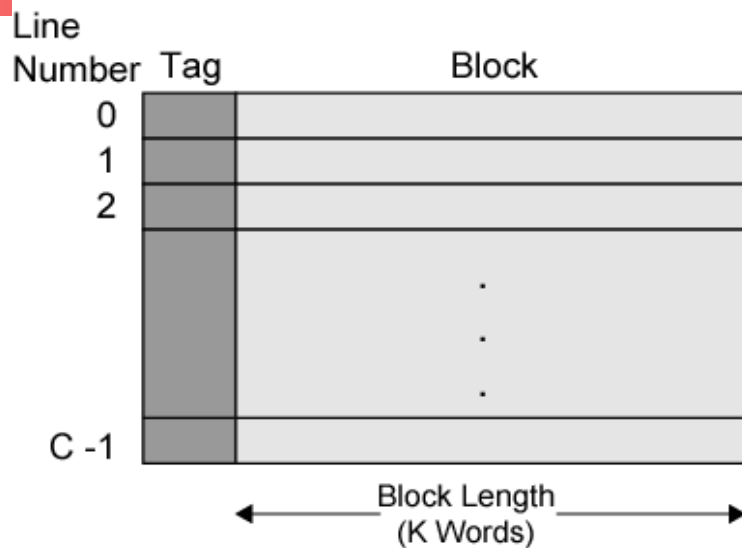
$$0.95T_1 + 0.05(T_1 + T_2) = 0.015\mu s$$



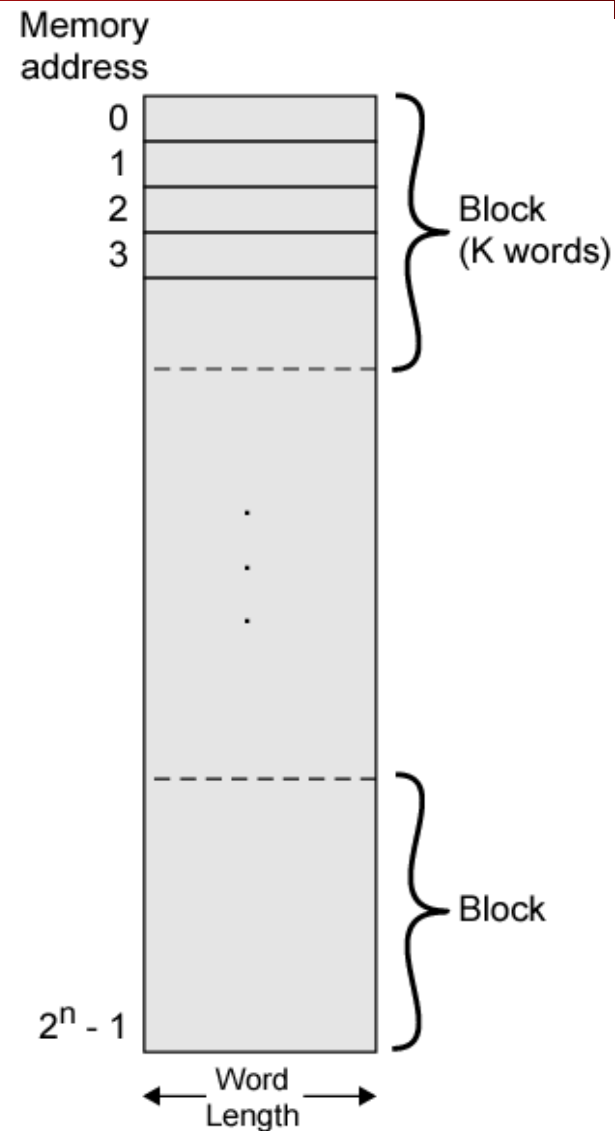
Memory hierarchy



Cấu trúc Cache/Main Memory



(a) Cache

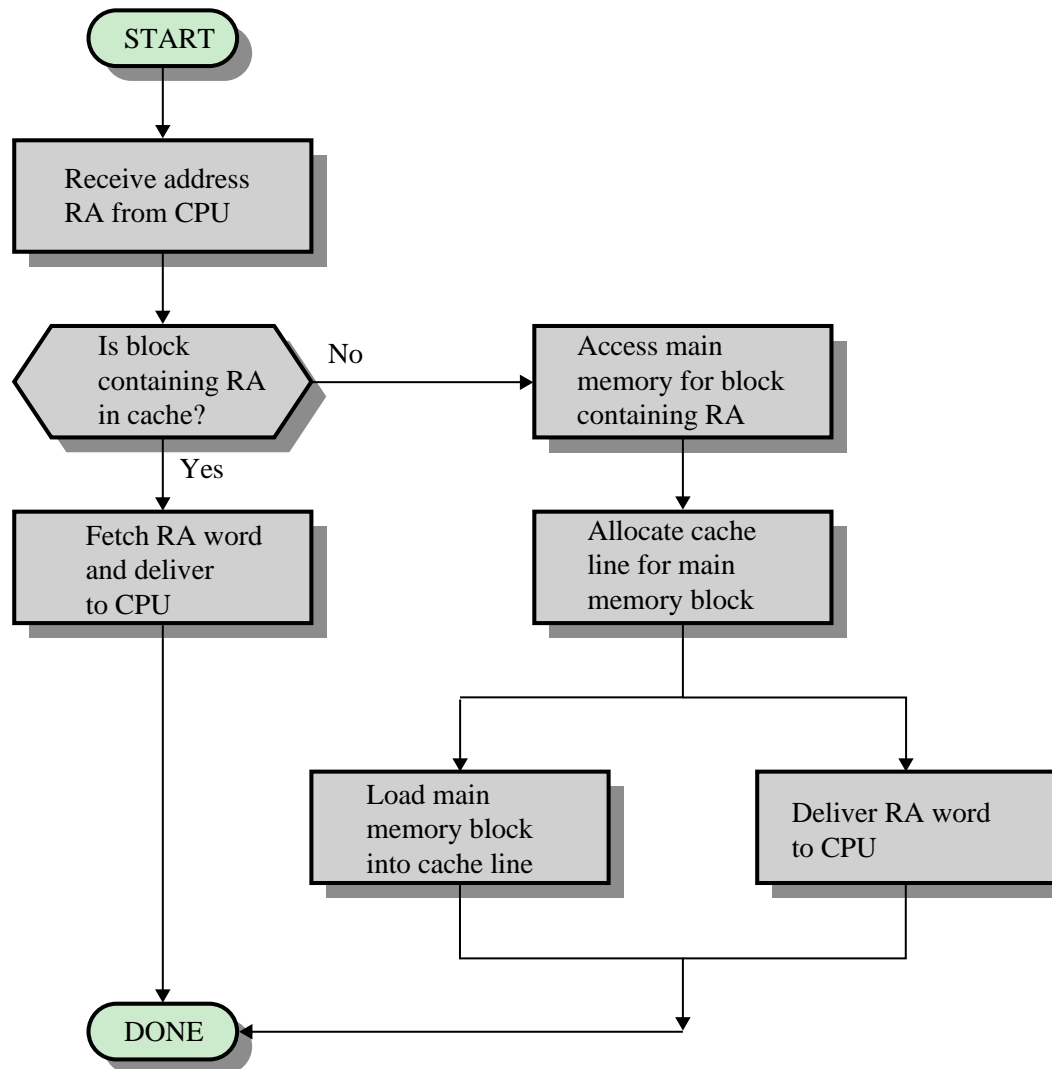


(b) Main memory



Cơ chế hoạt động

- CPU yêu cầu nội dung nhớ tại một vùng nhớ
- Kiểm tra cache đối với dữ liệu đó
- Nếu có, lấy dữ liệu đó từ cache (→ faster)
- Nếu không, đọc block yêu cầu từ main memory vào cache
- Chuyển dữ liệu đó từ cache đến CPU
- Cache sử dụng tags để xác định block nào của main memory trong mỗi khe cache (cache slot)

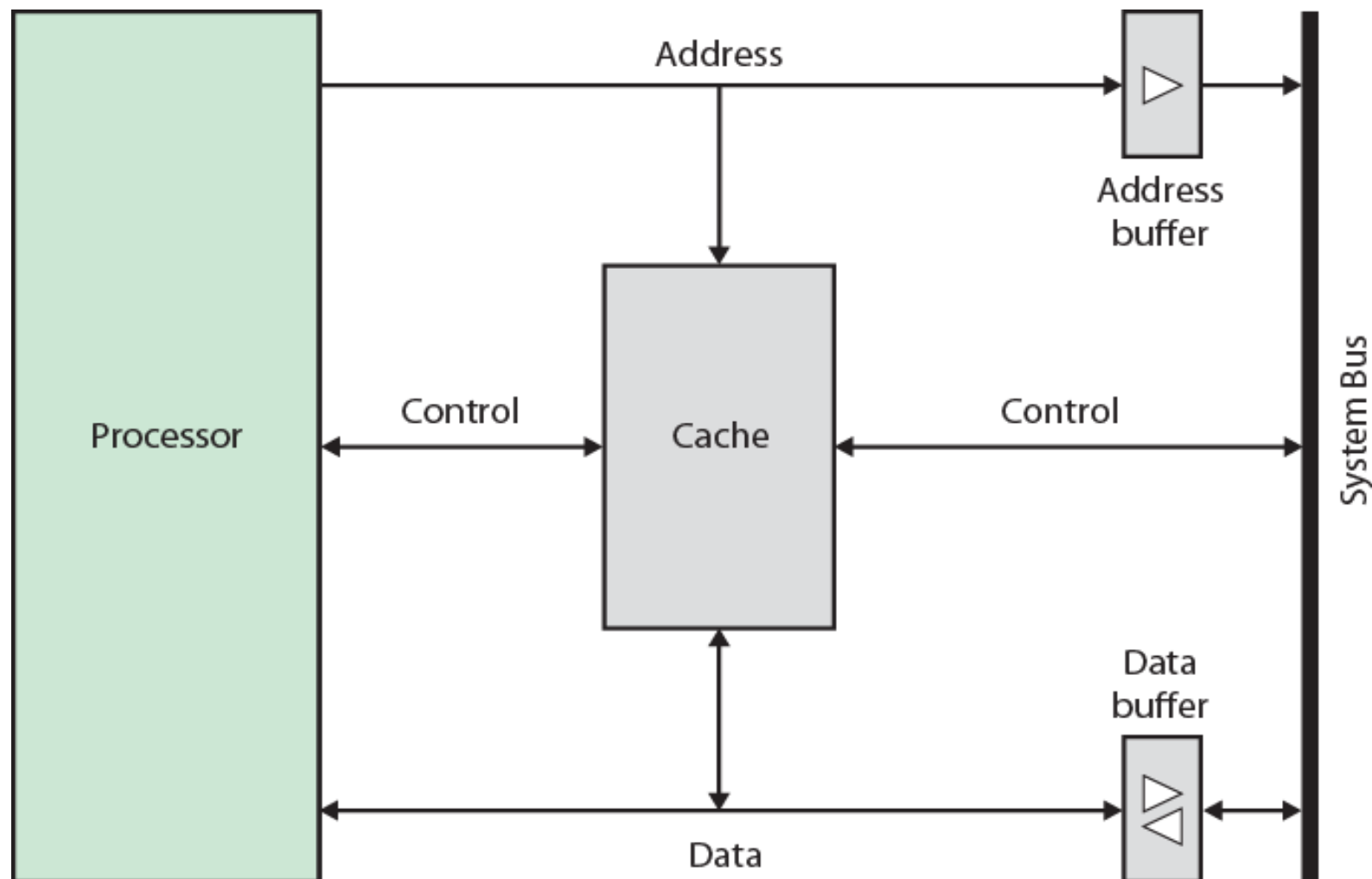




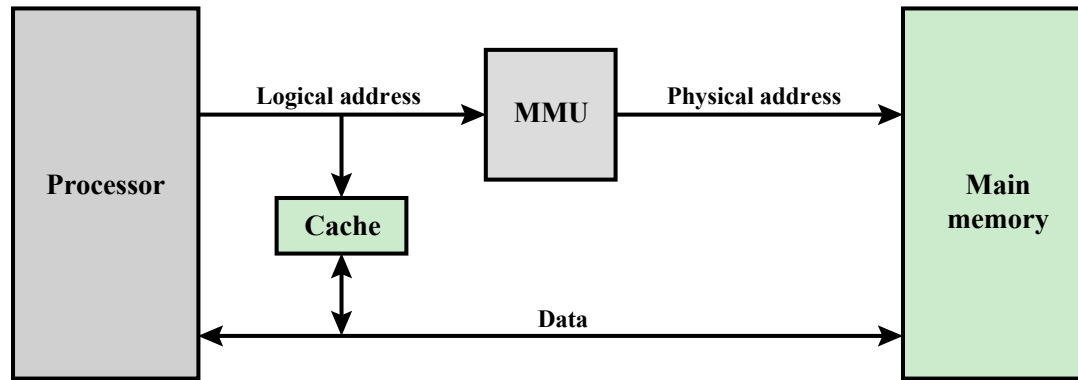
Thiết kế Cache

- Size: tốc độ và chi phí
 - More cache is faster (up to a point)
 - Checking cache for data takes time
 - More cache is expensive
- Mapping Function
 - Ánh xạ trực tiếp
 - Ánh xạ kết hợp
 - Ánh xạ kết hợp theo tập
- Replacement Algorithm
- Write Policy
- Block Size
- Number of Caches
 - 1 hay nhiều caches ở mức 1 (bên trong)
 - 1 cache ở mức 2 (trong hoặc ngoài) (thường $V_{L2} \geq 4V_{L1}$)
 - Đôi khi có thể sử dụng 1 cache ở mức 3 (ngoài)

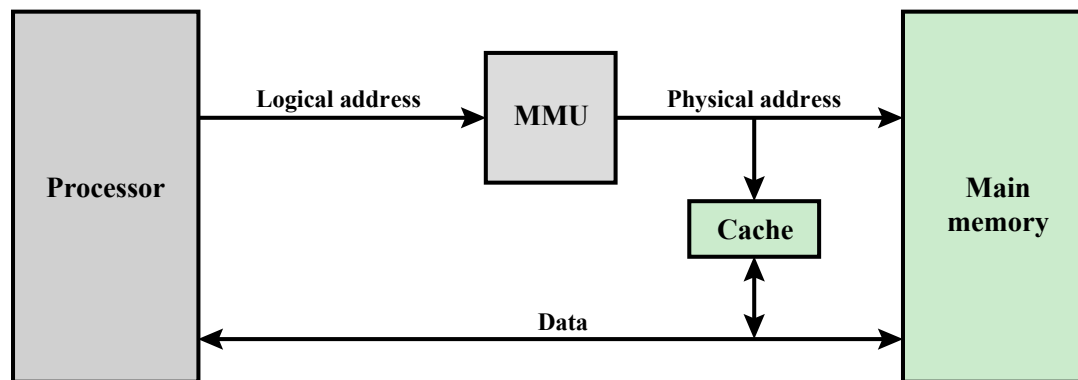
Tổ chức cache điển hình



Cache với bộ nhớ ảo (virtual memory)



(a) Logical Cache



(b) Physical Cache
Duy-Hieu Bui

Bảng so sánh caches một số CPU

Processor	Type	Year of Introduction	L1 Cachea	L2 cache	L3 Cache
IBM 360/85	Mainframe	1968	16 to 32 kB	—	—
PDP-11/70	Minicomputer	1975	1 kB	—	—
VAX 11/780	Minicomputer	1978	16 kB	—	—
IBM 3033	Mainframe	1978	64 kB	—	—
IBM 3090	Mainframe	1985	128 to 256 kB	—	—
Intel 80486	PC	1989	8 kB	—	—
Pentium	PC	1993	8 kB/8 kB	256 to 512 KB	—
PowerPC 601	PC	1993	32 kB	—	—
PowerPC 620	PC	1996	32 kB/32 kB	—	—
PowerPC G4	PC/server	1999	32 kB/32 kB	256 KB to 1 MB	2 MB
IBM S/390 G6	Mainframe	1999	256 kB	8 MB	—
Pentium 4	PC/server	2000	8 kB/8 kB	256 KB	—
IBM SP	High-end server/ supercomputer	2000	64 kB/32 kB	8 MB	—
CRAY MTAb	Supercomputer	2000	8 kB	2 MB	—
Itanium	PC/server	2001	16 kB/16 kB	96 KB	4 MB
Itanium 2	PC/server	2002	32 kB	256 KB	6 MB
IBM POWER5	High-end server	2003	64 kB	1.9 MB	36 MB
CRAY XD-1	Supercomputer	2004	64 kB/64 kB	1MB	—
IBM POWER6	PC/server	2007	64 kB/64 kB	4 MB	32 MB
IBM z10	Mainframe	2008	64 kB/128 kB	3 MB	24-48 MB
Intel Core i7 EE 990	Workstaton/ server	2011	6 ´ 32 kB/32 kB	1.5 MB	12 MB
IBM zEnterprise 196	Mainframe/ Server	2011	24 ´ 64 kB/ 128 kB	24 ´ 1.5 MB	24 MB L3 192 MB L4

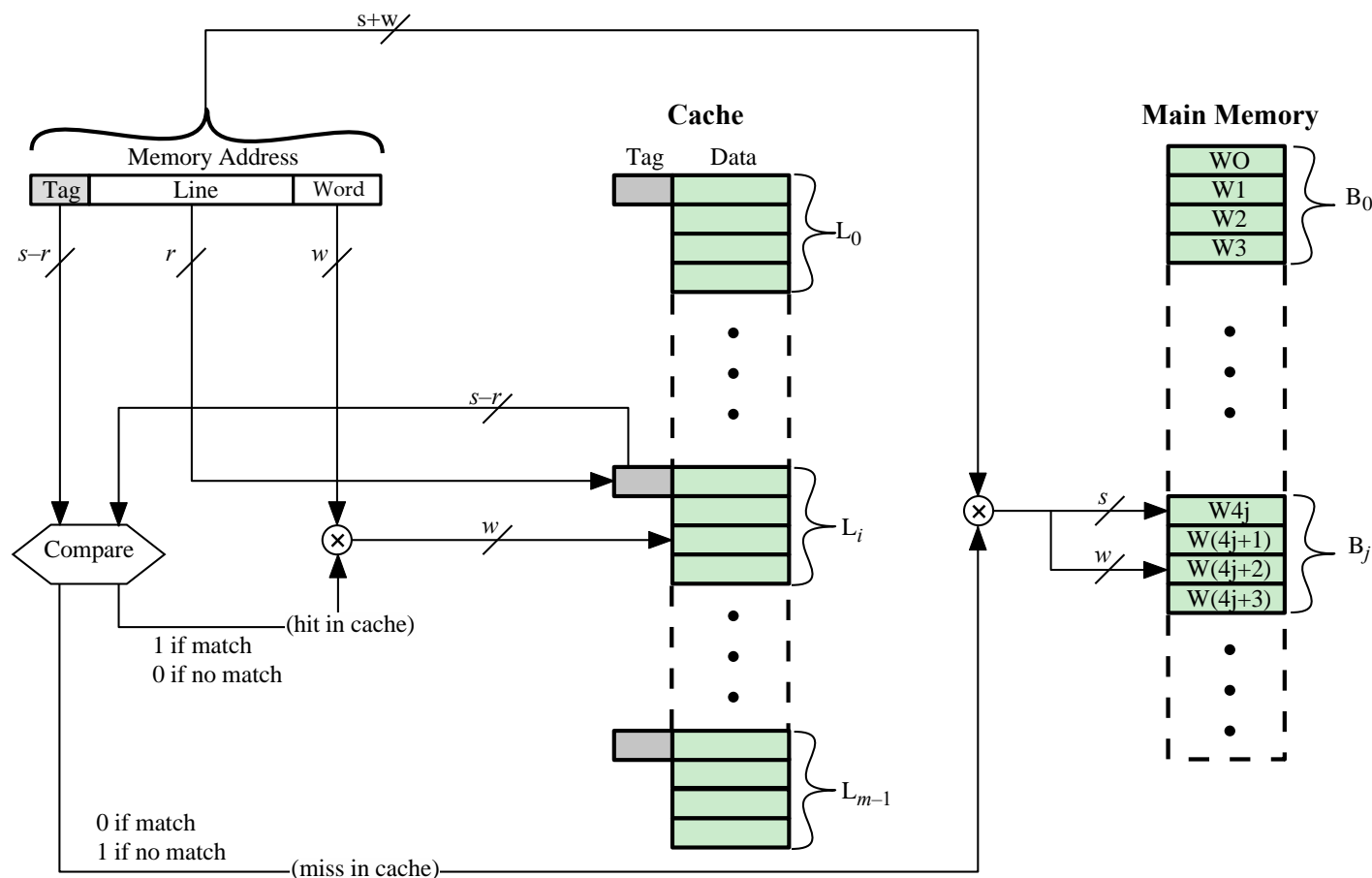


Ánh xạ cache – main memory

- Ánh xạ trực tiếp:
 - Mỗi block của bộ nhớ chính được đặt chỉ vào 1 dòng cache
- Ánh xạ kết hợp:
 - 1 block của bộ nhớ chính có thể được đặt vào bất kỳ dòng nào trong cache
- Ánh xạ kết hợp theo tập:
 - 1 block của bộ nhớ chính có thể được đặt vào bất kỳ dòng nào trong 1 tập n-blocks của cache

Tag s-r	Cache slot r	Word w
---------	--------------	--------

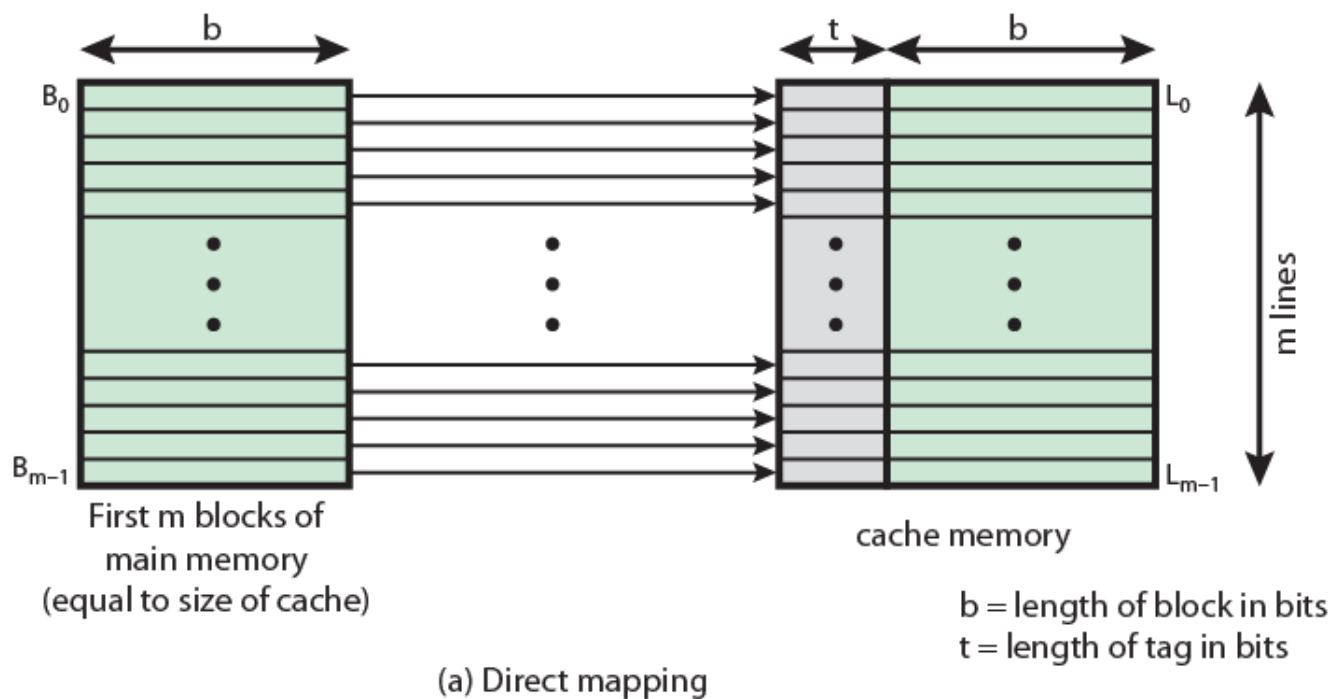
Direct Mapping



- Độ dài: $s+w$ bits
- Số lượng block = 2^s
- Tag = $s-r$ bits

Độ lớn block = 2^w từ
Số slots = 2^r

Ảnh xạ trực tiếp từ Cache đến MM

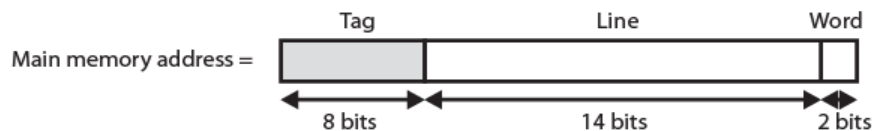
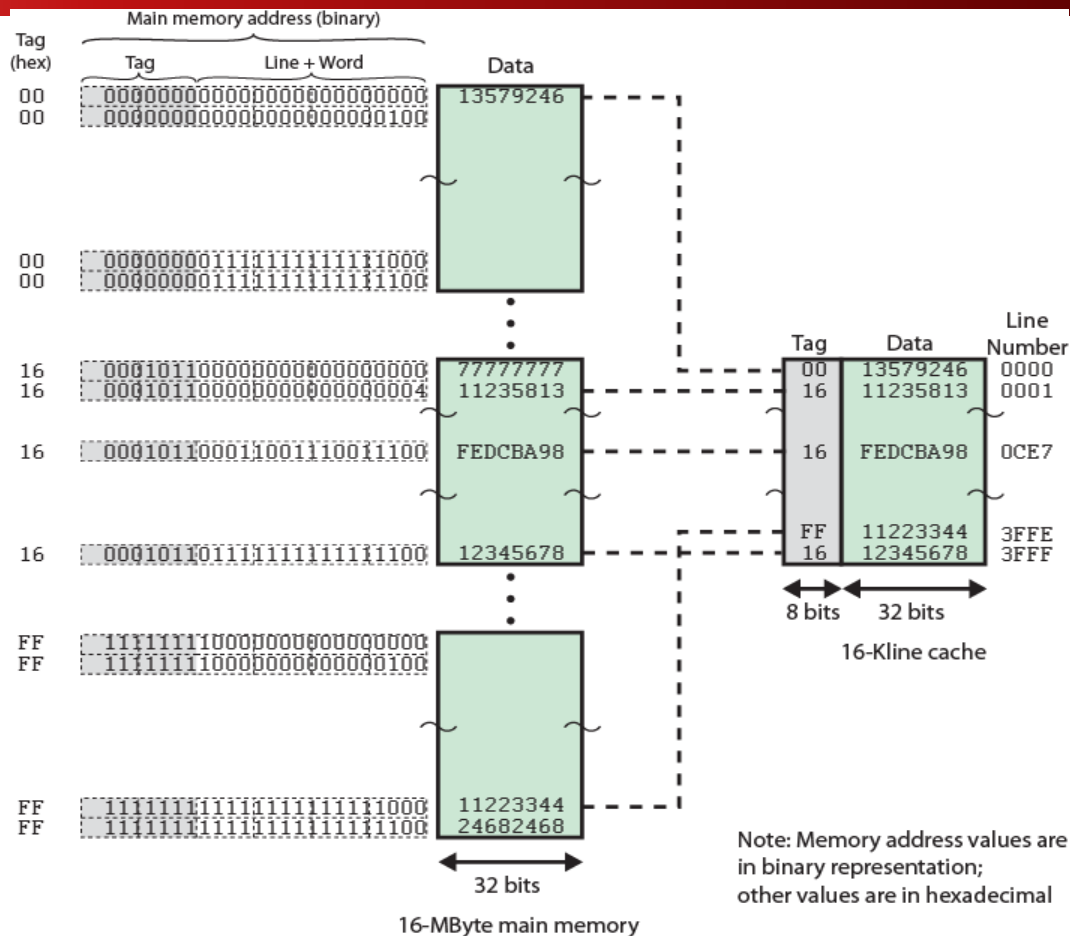




Bảng ánh xạ trực tiếp

Vị trí slot (cache line)	Khối MM có thể chứa	Tag
0	0, m, 2m, 3m...2s-m	
1	1,m+1, 2m+1...2s-m+1	
...		
m-1	m-1, 2m-1,3m-1...2s-1	

- Cache size = 64kByte,
- Cache block = 4 bytes -> có 16k (2^{14}) lines
- Main memory = 16MBytes, 24 bit address ($2^{24}=16\text{M}$)





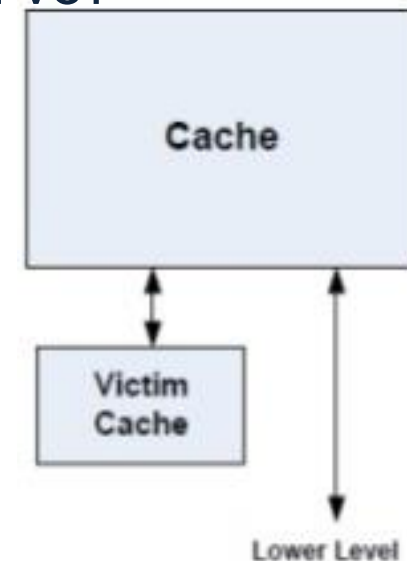
Ưu/Nhược của ánh xạ trực tiếp

- Đơn giản, không cần thời gian tính toán nhiều
- Chi phí cài đặt thấp
- Mỗi block chỉ có một vị trí duy nhất trong cache (slot tương ứng) → có thể dẫn đến tỷ lệ cache miss tăng lên

- Cách tiếp cận cho phép giảm tỷ lệ cache miss đối với phương pháp ánh xạ trực tiếp

- Ý tưởng:

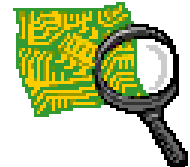
- Bổ sung thêm một số slots để lưu lại những slots bị thay thế, tạo “second chance”
- Từ 4-16 slots phục vụ lưu lại nội dung slot cần phải bị thay thế → nếu CPU cần sử dụng lại slot đó thì chỉ cần lookup trong victim cache
- Victim Cache sử dụng phương pháp ánh xạ kết hợp (fully associative cache)
- Thường được sử dụng ở L1 cache



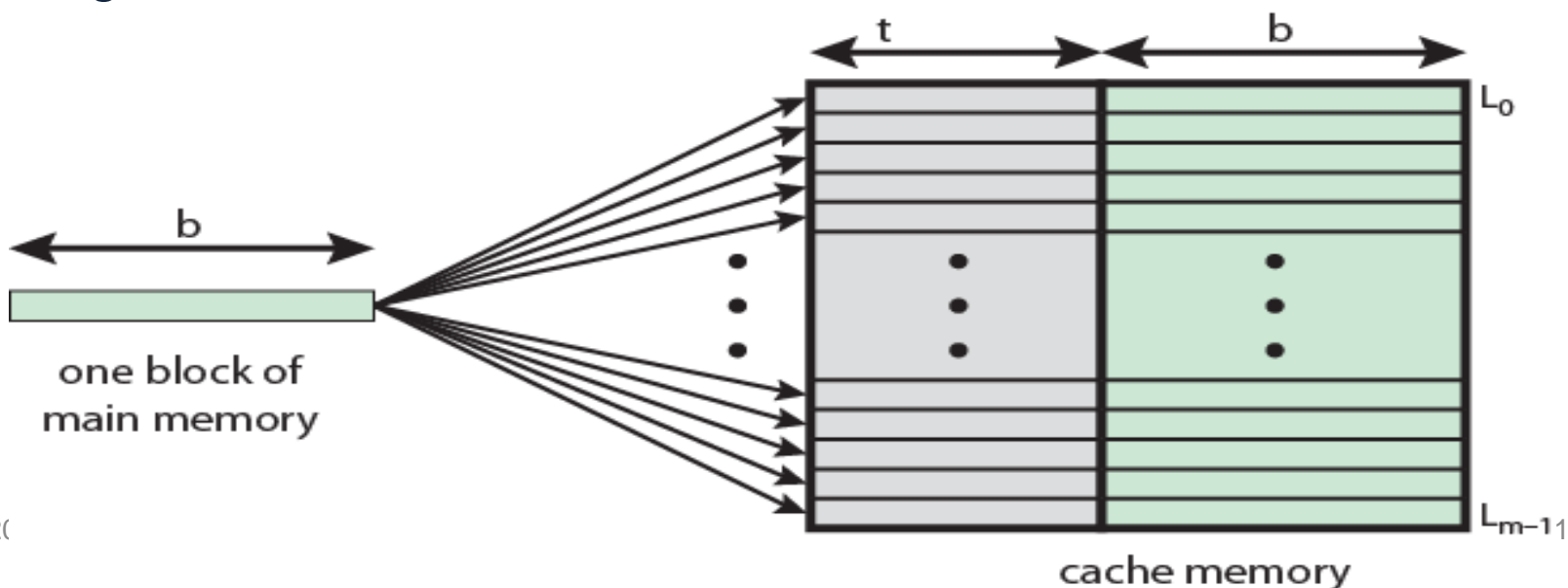


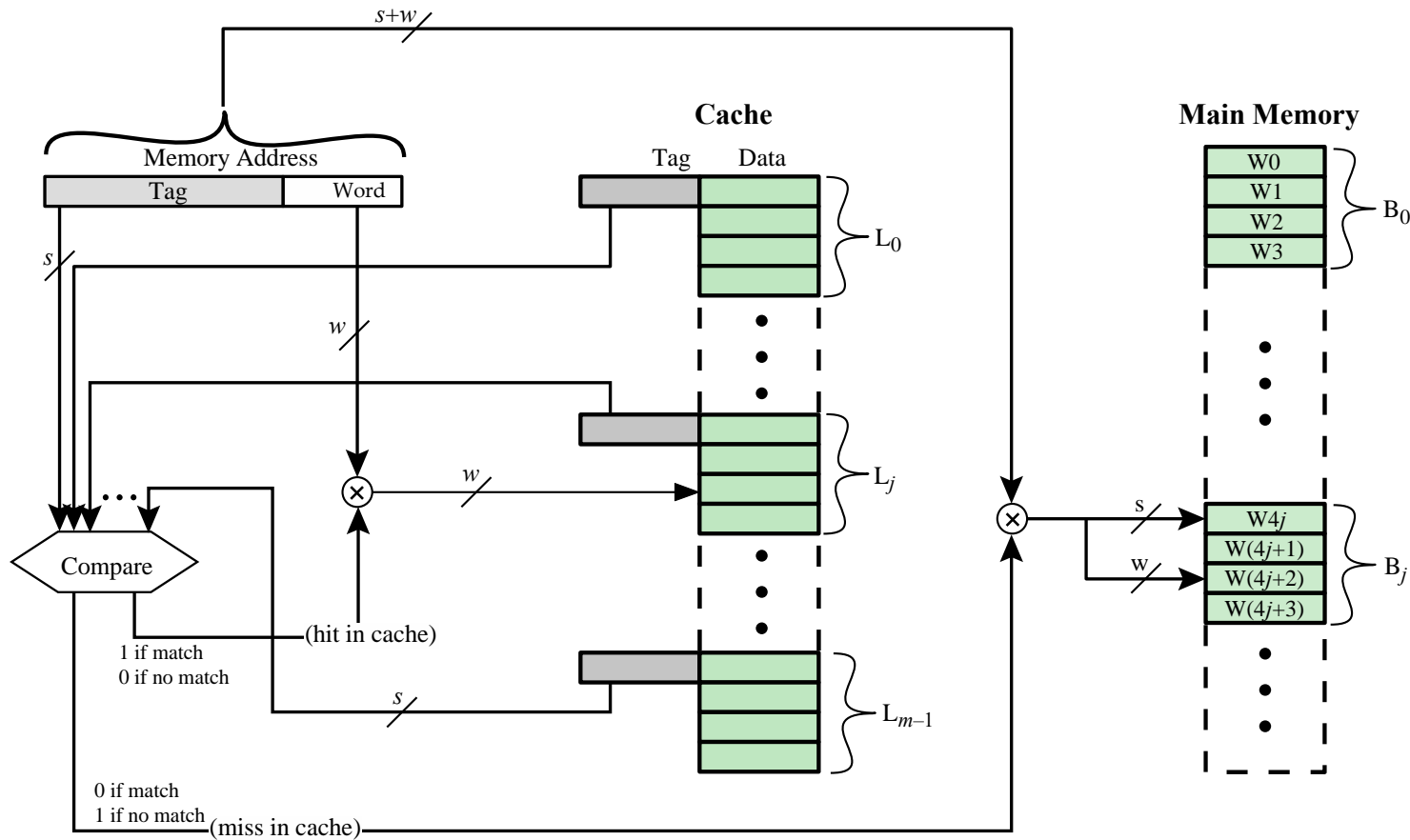
Direct Mapping Summary

- Address length = $(s + w)$ bits
- Number of addressable units = 2^{s+w} words or bytes
- Block size = line size = 2^w words or bytes
- Number of blocks in main memory = $2^{s+w}/2^w = 2^s$
- Number of lines in cache = $m = 2^r$
- Size of tag = $(s - r)$ bits



- MM block có thể được đưa vào bất kỳ dòng cache nào
- Địa chỉ từ nhớ sẽ được chia thành hai phần: tag và word
 - Tag sẽ cho phép xác định địa chỉ block trong MM
 - Word cho phép xác định vị trí từ trong dòng cache tương ứng với địa chỉ Tag
- Việc so khớp cache hit/miss được tiến hành dựa trên trường Tag, có thể dẫn đến trường hợp phải so khớp toàn bộ các dòng cache

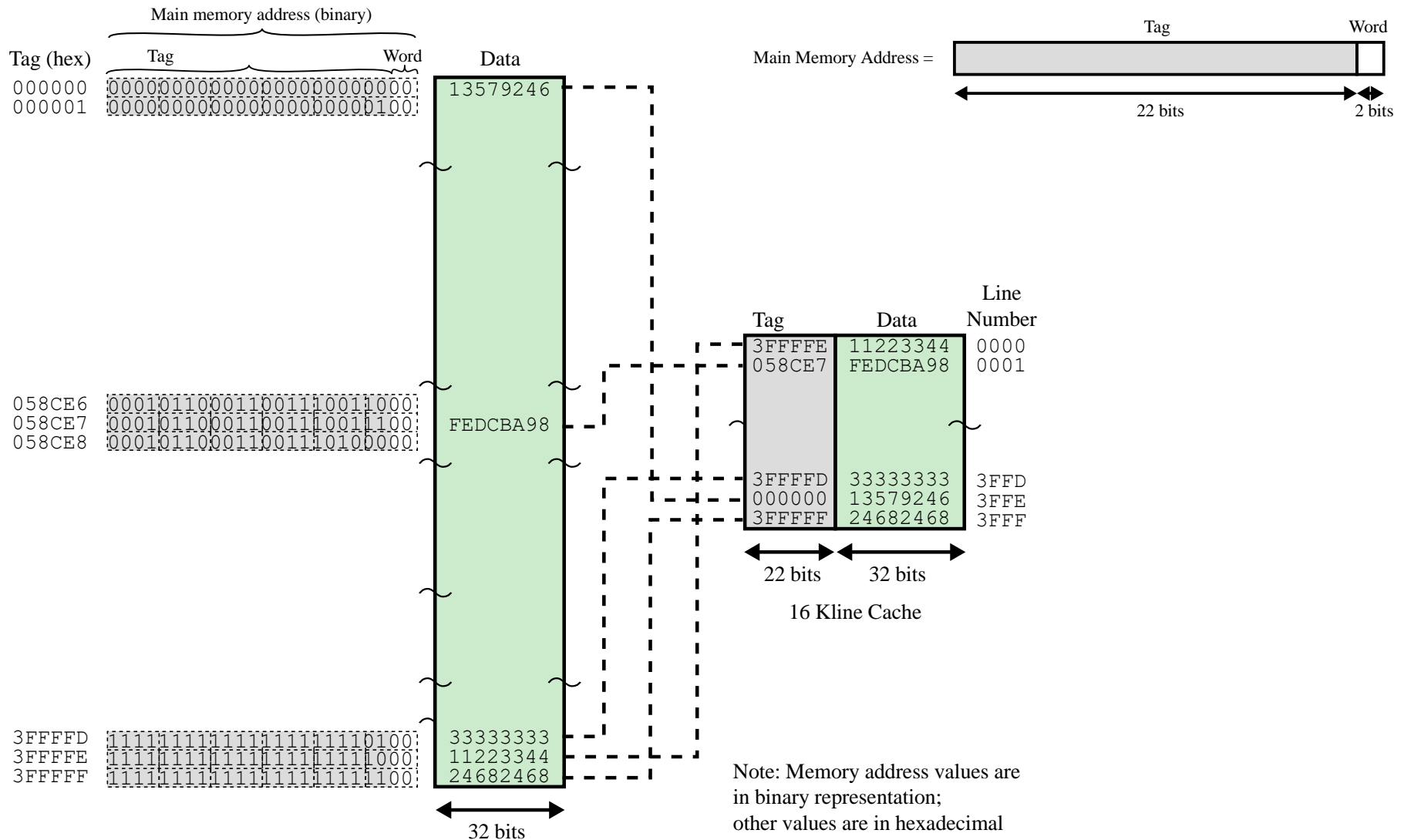




- Độ dài: $s+w$ bits
- Số lượng block = 2^s
- Tag = s bits

Độ lớn block = 2^w từ
Số slots = undefined

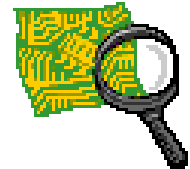
Ví dụ ánh xạ kết hợp





Associative Mapping Summary

- Address length = $(s + w)$ bits
- Number of addressable units = 2^{s+w} words or bytes
- Block size = line size = 2^w words or bytes
- Number of blocks in main memory = $2^{s+w}/2^w = 2^s$
- Number of lines in cache = undetermined
- Size of tag = s bits

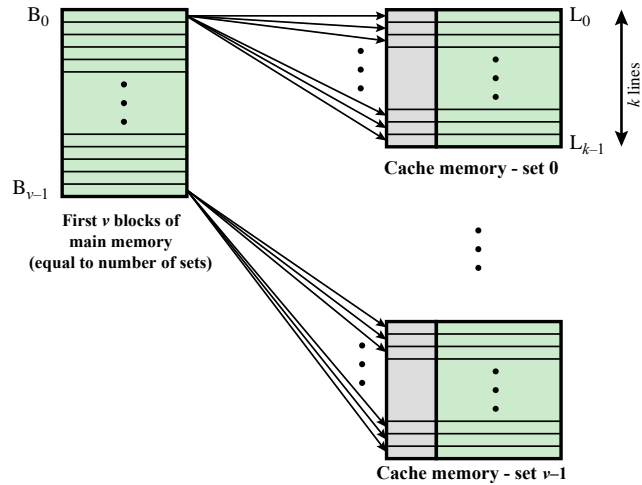




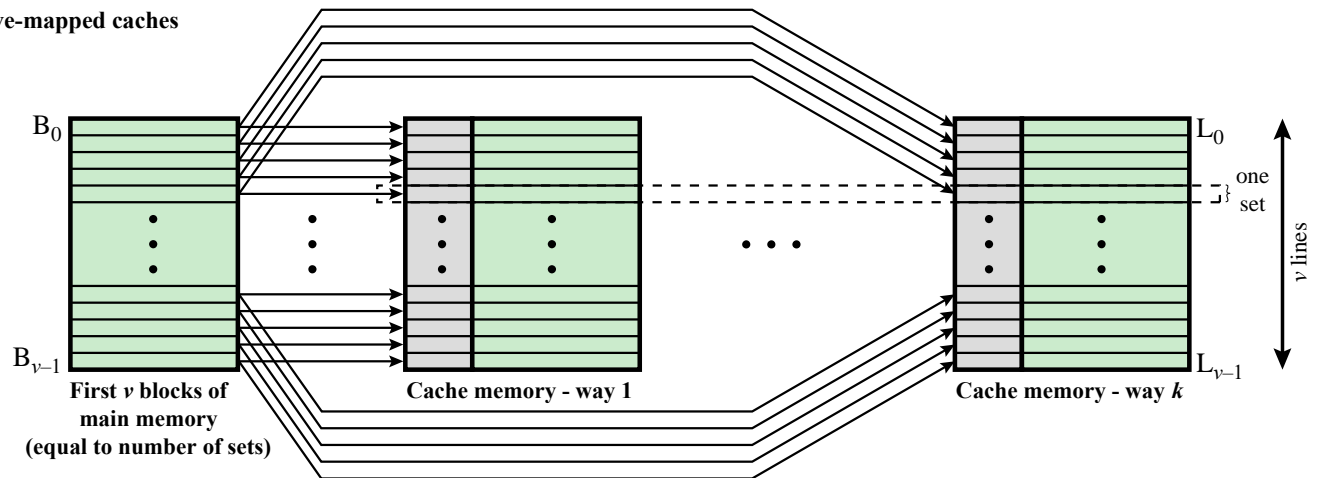
Ánh xạ kết hợp theo tập – Set Associative Mapping

- Cache được chia thành nhiều tập, mỗi tập chứa một số slots
- Mỗi block trong MM được đưa vào một tập duy nhất theo cơ chế ánh xạ trực tiếp; tuy nhiên, nó có thể được đặt vào bất kỳ slot nào trong tập đó theo cơ chế ánh xạ kết hợp
 - ❑ Chẳng hạn với 2 lines mỗi tập, mỗi block có thể ở 1 trong 2 lines tại vị trí một tập duy nhất
- Với k dòng, cơ chế ánh xạ này được gọi là k-Way Set Associative Mapping

Set Associative Mapping...

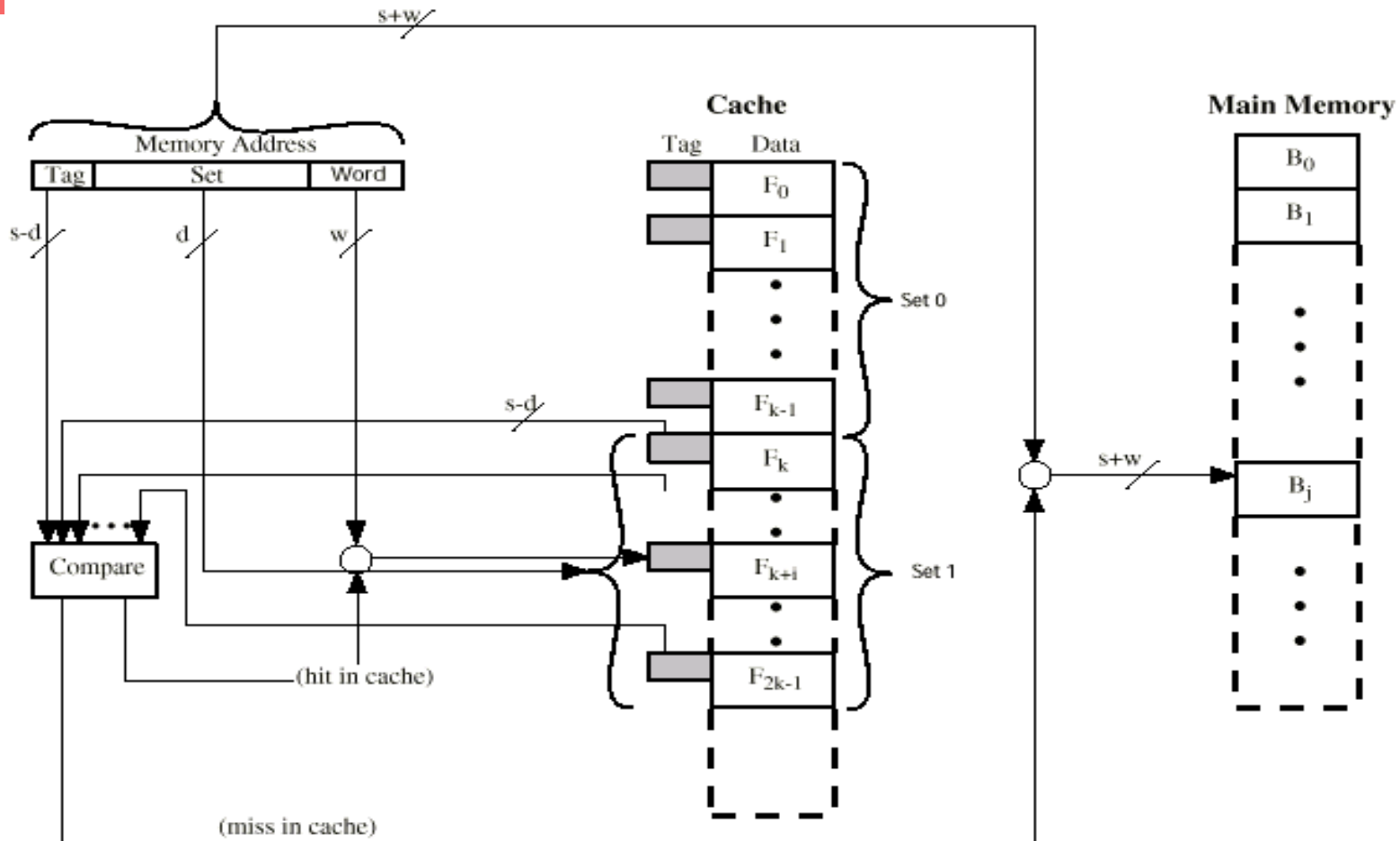


(a) v associative-mapped caches



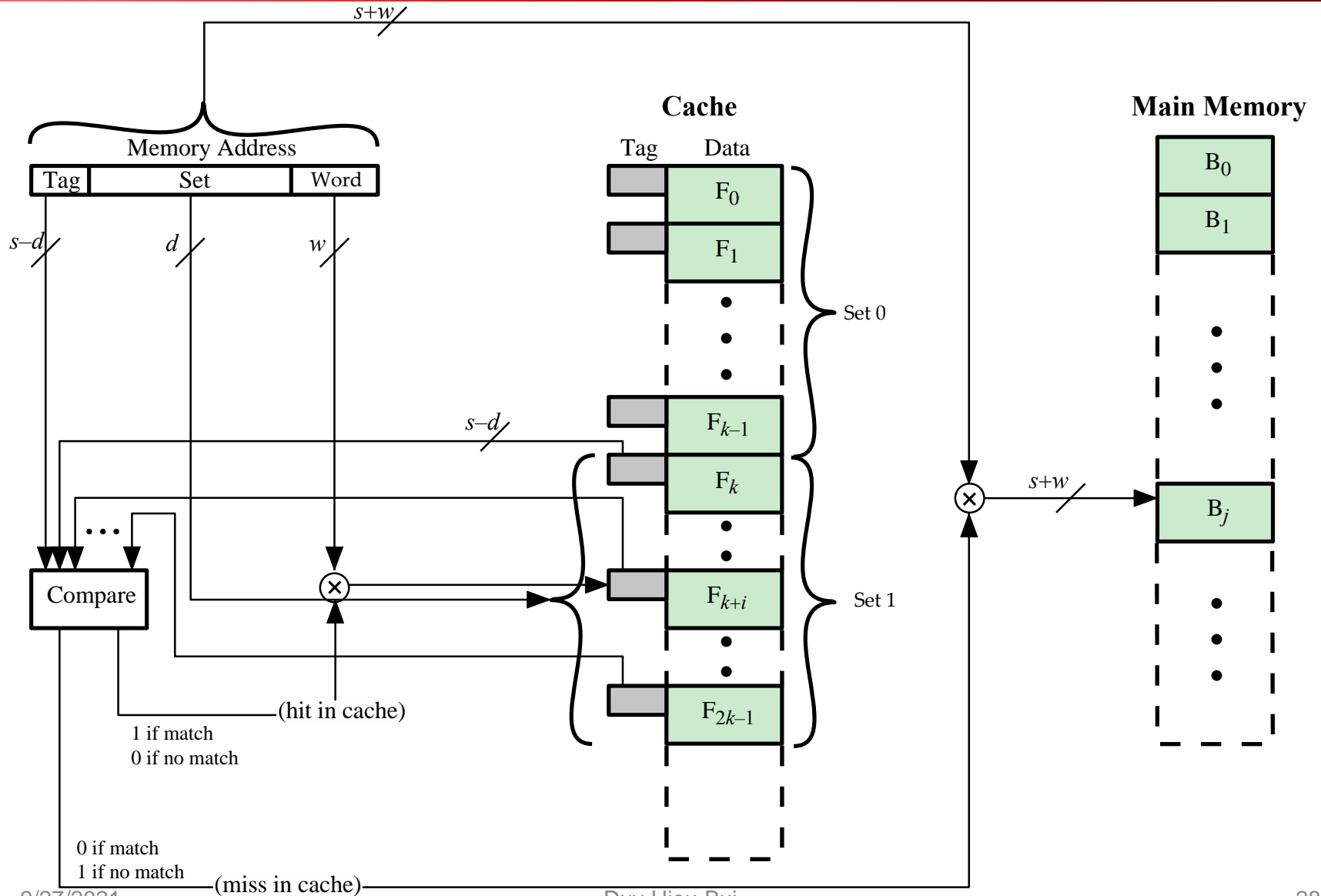
(b) k direct-mapped caches

k-Way Set Associative



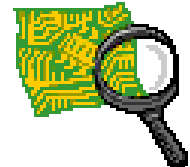
- Độ dài: $s+w$ bits
 - Số lượng block = 2^s
 - Số set $v = 2^d$, Số slots = $k \cdot v$
- Độ lớn block = 2^w từ
Số lines/set = k
Tag = $s-d$ bits

Ví dụ về Ánh xạ kết hợp theo tập hợp



Set Associative Mapping Summary

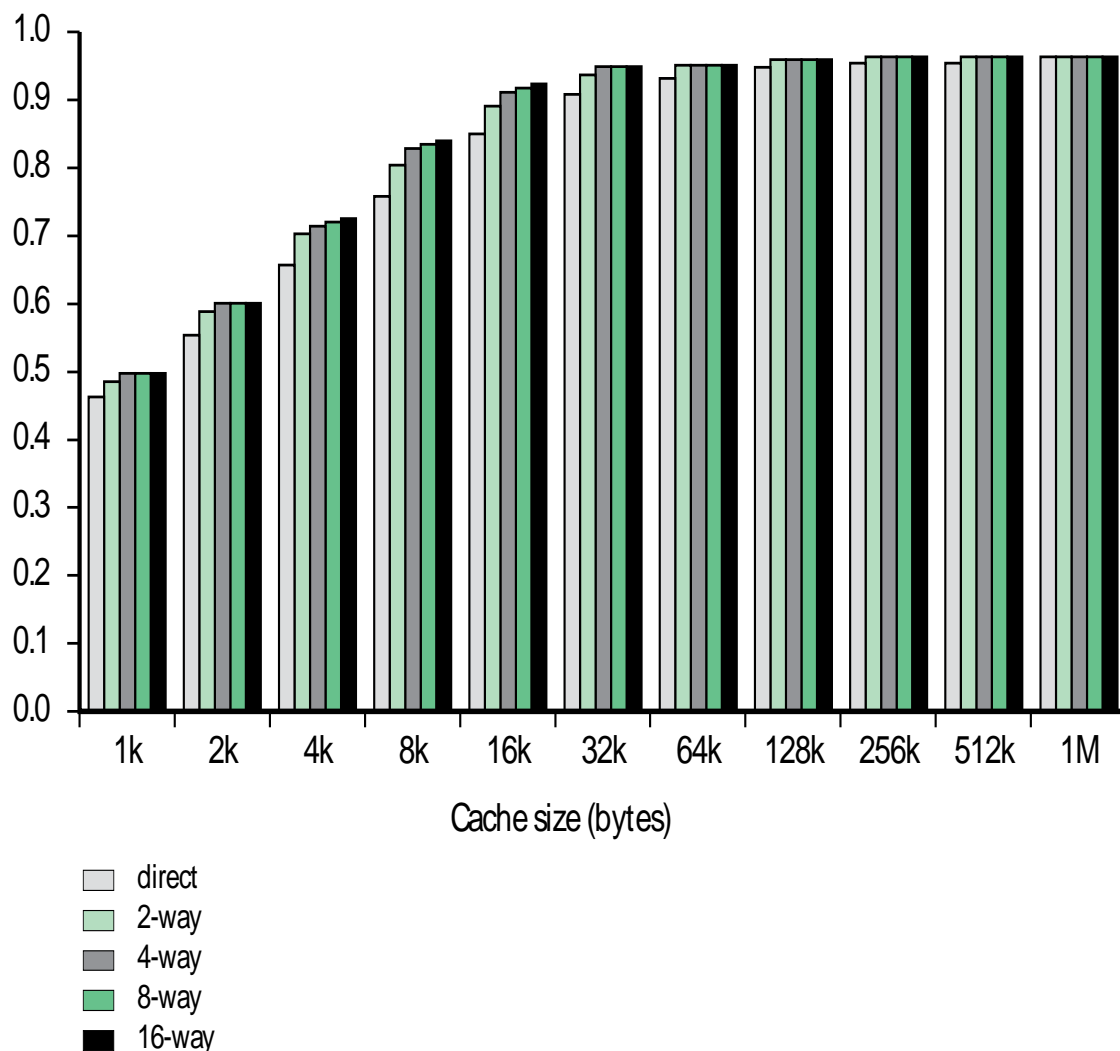
- Address length = $(s + w)$ bits
- Number of addressable units = 2^{s+w} words or bytes
- Block size = line size = 2^w words or bytes
- Number of blocks in main memory = $2^{s+w}/2^w = 2^s$
- Number of lines in set = k
- Number of sets = $v = 2^d$
- Number of lines in cache = $m = kv = k * 2^d$
- Size of cache = $k * 2^{d+w}$ words or bytes
- Size of tag = $(s - d)$ bits





So sánh hiệu năng

- Khác biệt giữa direct và k-way tối đa đến 64kB
- Khác biệt giữa 2-way và 4-way tại cache 4kB nhỏ hơn so với cache 1kB
- Độ phức tạp của cache tăng theo độ kết hợp
- Không có lý do gì để không tăng cache lên 8kB hay 16kB
- Từ 32kB trở lên không cải thiện nhiều về hiệu năng





Truy cập cache





































- Xây dựng địa chỉ nhớ dựa trên hàm ánh xạ
- Địa chỉ nhớ 1 từ cho phép tìm được:
 - block mà nó phụ thuộc
 - Vị trí của nó trong block đó
 - Vị trí block đó trong cache
- Cấu trúc: 2 phần
 - Số block:
 - index: vị trí block trong cache
 - tag: xác định block trong số những blocks có cùng vị trí
 - Độ rời: địa chỉ của từ trong block



Giải thuật thay thế

- Ảnh xạ trực tiếp: không có sự lựa chọn
 - 1 block được đặt chỉ vào 1 vị trí
- Ảnh xạ kết hợp – kết hợp tập:
 - Chọn block ngẫu nhiên
 - FIFO: Chọn block **cũ nhất** trong cache
 - Chọn block được sử dụng sớm nhất (LRU Least Recently Used)
 - Chọn block có tần xuất sử dụng ít nhất (LFU Least Frequently Used)
- Hiệu quả nhất: LFU, LRU, ngẫu nhiên
- Dễ cài đặt: ngẫu nhiên, FIFO

- Xét chuỗi truy cập các khối MM có địa chỉ: ABCABDADBCBCBCB
- Cache có 3 slots

	OPT	FIFO	LRU	LFU
ABC				
A				
B				
D				
A				
D				
B				
C				
B				



Chính sách Write

- Không được ghi đè một cache block trừ khi CPU cập nhật dữ liệu đó
- Mỗi CPU có thể có cache riêng
- I/O có thể truy cập trực tiếp main memory
- Phân biệt:
 - Write through
 - Write back



Write through

- Ghi cả main memory lẫn cache
- Nếu có nhiều CPU, cần quản lý lưu lượng của bộ nhớ chính để giữ cache cục bộ được cập nhật
- Giảm tốc độ ghi, gây nhiều traffic



Write back

- Cập nhật đầu tiên trong cache, lưu trạng thái trong bit cập nhật của cache slot (update bit = set)
- Nếu block có yêu cầu thay đổi dữ liệu, ghi nội dung hiện thời vào main memory chỉ khi update bit đã được thiết lập
- I/O phải truy cập main memory thông qua cache
- Khoảng 15% truy cập bộ nhớ là thao tác ghi



Độ lớn đường cache (slot size)

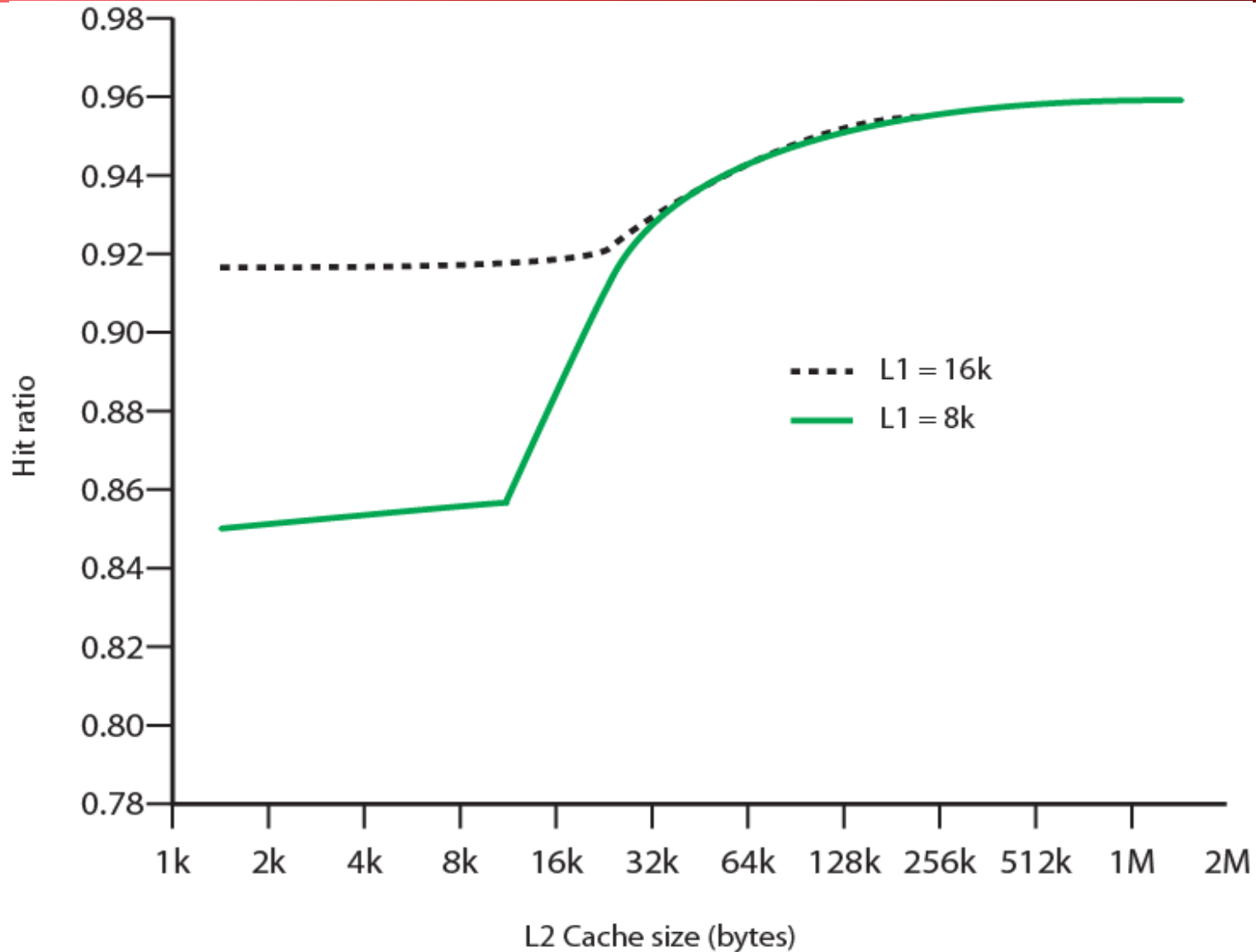
- Tăng kích thước slot sẽ tăng tỷ lệ hit (dựa trên nguyên lý cục bộ của chương trình)
- Tỷ lệ hit sẽ giảm nếu kích thước block quá lớn
 - ❑ Xác suất sử dụng dữ liệu mới tải lên nhỏ hơn xác suất sử dụng lại dữ liệu được thay thế
- Block lớn hơn
 - ❑ Giảm số blocks nằm trong cache
 - ❑ Khả năng dữ liệu bị đè sau khi được tải vào tăng lên
 - ❑ Các từ nhớ có tính cục bộ nhỏ đi → giảm tỷ lệ hit
- Không có giá trị tối ưu được xác định chính thức, từ 8 đến 64 bytes là giá trị hợp lý
- Với những hệ thống tính toán hiệu năng cao, kích thước block thường từ 64 đến 128 bytes

Cache đa mức - Multilevel Caches

- Hầu hết các CPU ngày nay đều có cache on chip (do có mật độ logic gates tăng lên)
 - ➔ Nhanh hơn vì không cần sử dụng bus hệ thống
 - ➔ Bus được sử dụng cho những mục đích khác

- Thường thì cache đa mức được cài đặt trong cả trong và ngoài chip
 - ❑ L1 on chip, L2 off chip in static RAM
 - ❑ L2 access much faster than DRAM or ROM
 - ❑ L2 often uses separate data path
 - ❑ L2 may now be on chip
 - ❑ Resulting in L3 cache
 - Bus access or now on chip...

Tỷ lệ Hit (L1 & L2) với 8 kB và 16 kB L1





Unified v Split Caches

- **Unified Cache:** cache chung cho cả lệnh và dữ liệu
 - ➔ Tỷ lệ hit cao hơn; cân bằng tải lệnh và dữ liệu; chỉ phải thiết kế, cài đặt một bộ nhớ cache
- **Split Cache:** một cache cho lệnh và một cache cho dữ liệu
 - ➔ Loại được sự cạnh tranh cache (cache contention) trong chu trình tải/giải mã lệnh và chu trình thi hành (nâng cao hiệu năng pipeline)



Hiệu năng

- Thời gian truy cập trung bình = Thời gian truy cập thành công + tỷ lệ thất bại x penalty thất bại
- Thời gian truy cập = Thời gian truy cập một dữ liệu trong cache
- Tỷ lệ thất bại = số lần dữ liệu cần đọc không chứa trong cache / số lần truy cập cache

Intel Cache Evolution

Problem	Solution	Processor on which Feature First Appears
External memory slower than the system bus.	Add external cache using faster memory technology.	386
Increased processor speed results in external bus becoming a bottleneck for cache access.	Move external cache on-chip, operating at the same speed as the processor.	486
Internal cache is rather small, due to limited space on chip	Add external L2 cache using faster technology than main memory	486
Contention occurs when both the Instruction Prefetcher and the Execution Unit simultaneously require access to the cache. In that case, the Prefetcher is stalled while the Execution Unit's data access takes place.	Create separate data and instruction caches.	Pentium
Increased processor speed results in external bus becoming a bottleneck for L2 cache access.	Create separate back-side bus that runs at higher speed than the main (front-side) external bus. The BSB is dedicated to the L2 cache.	Pentium Pro
	Move L2 cache on to the processor chip.	Pentium II
Some applications deal with massive databases and must have rapid access to large amounts of data. The on-chip caches are too small.	Add external L3 cache.	Pentium III
	Move L3 cache on-chip.	Pentium 4

Pentium 4 Block Diagram

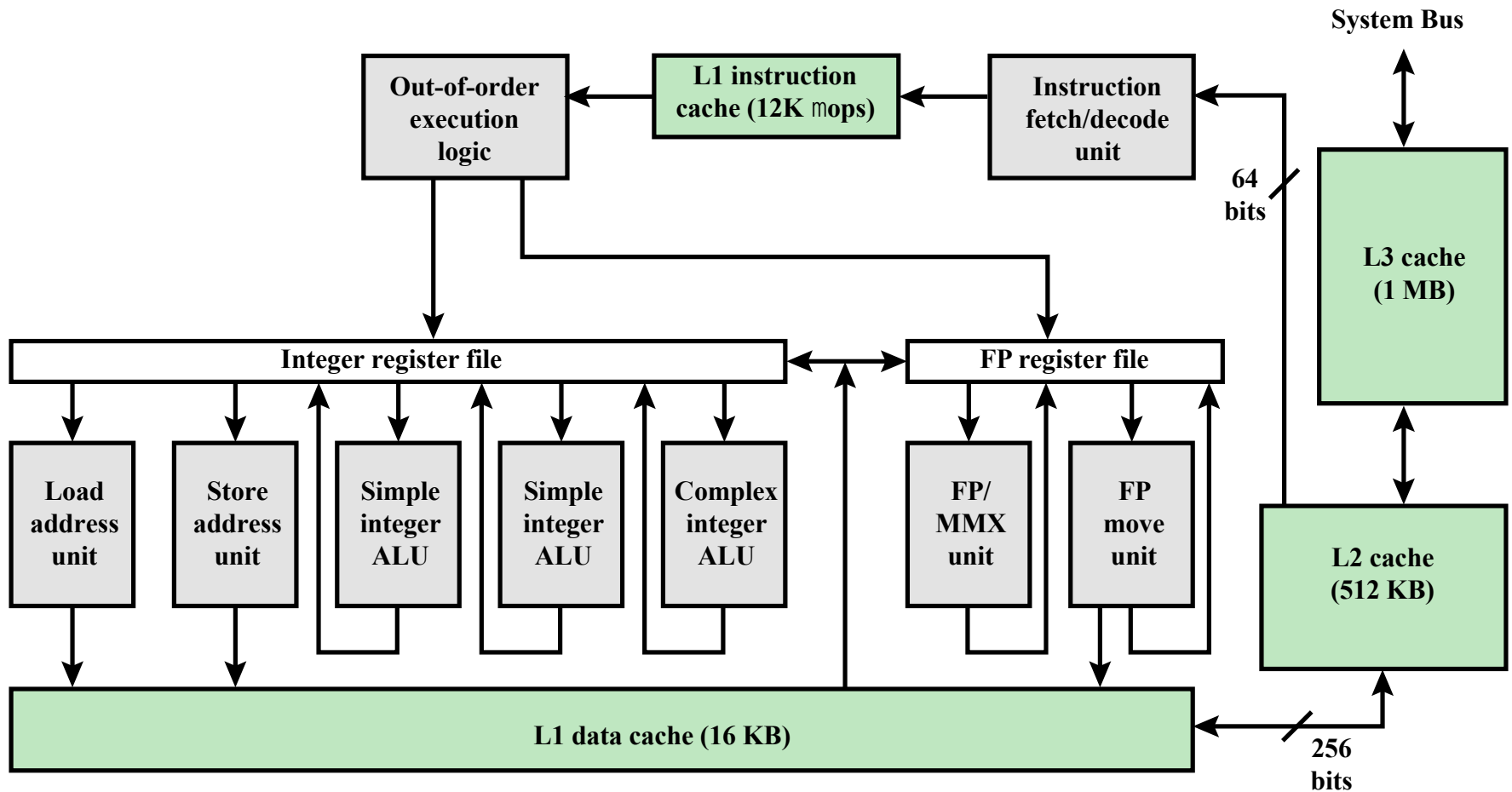


Table 4.5 Pentium 4 Cache Operating Modes

Control Bits		Operating Mode		
CD	NW	Cache Fills	Write Throughs	Invalidates
0	0	Enabled	Enabled	Enabled
1	0	Disabled	Enabled	Enabled
1	1	Disabled	Disabled	Disabled

Note: CD = 0; NW = 1 is an invalid combination.



Intel Coffee Lake

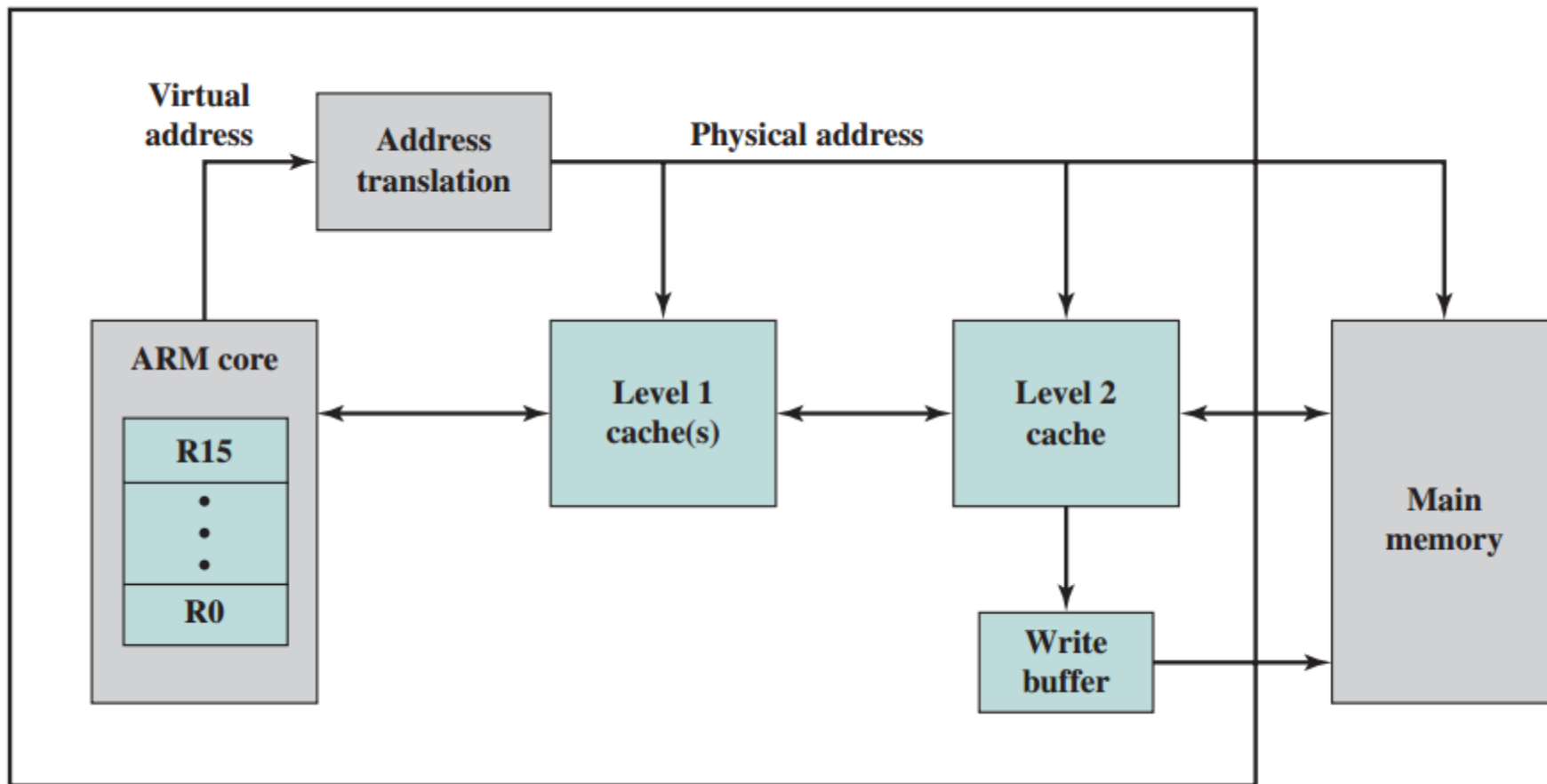
Processor branding ↕	Model ↕	Cores (threads) ↕	CPU clock rate ↕	CPU turbo clock rate				GPU ↕	Maximum GPU clock rate ↕	L3 cache ↕	TDP ↕	Memory support ↕	Socket ↕	Price (USD) ↕
				Single core ↕	Dual core ↕	Quad core ↕	Hexa core ↕							
Core i7	8700K🔗	6 (12)	3.7 GHz	4.7 GHz	4.6 GHz	4.4 GHz	4.3 GHz	UHD Graphics 630 (GT2)	1.20 GHz	12 MB	95 W	DDR4-2666	LGA 1151	\$359
	8700🔗		3.2 GHz	4.6 GHz	4.5 GHz	4.3 GHz					65 W			\$303
Core i5	8600K🔗	6 (6)	3.6 GHz	4.3 GHz	4.2 GHz		4.1 GHz		1.15 GHz	9 MB	95 W			\$257
	8400🔗		2.8 GHz	4.0 GHz	3.9 GHz		3.8 GHz		1.05 GHz		65 W			\$182
Core i3	8350K🔗	4 (4)	4.0 GHz	N/A					1.15 GHz	8 MB	91 W	DDR4-2400		\$168
	8100🔗		3.6 GHz						1.10 GHz	6 MB	65 W			\$117



ARM Cache

Core	Cache Type	Cache Size (kB)	Cache Line Size (words)	Associativity	Location	Write Buffer Size (words)
ARM720T	Unified	8	4	4-way	Logical	8
ARM920T	Split	16/16 D/I	8	64-way	Logical	16
ARM926EJ-S	Split	4-128/4-128 D/I	8	4-way	Logical	16
ARM1022E	Split	16/16 D/I	8	64-way	Logical	16
ARM1026EJ-S	Split	4-128/4-128 D/I	8	4-way	Logical	8
Intel StrongARM	Split	16/16 D/I	4	32-way	Logical	32
Intel Xscale	Split	32/32 D/I	8	32-way	Logical	32
ARM1136-JF-S	Split	4-64/4-64 D/I	8	4-way	Physical	32

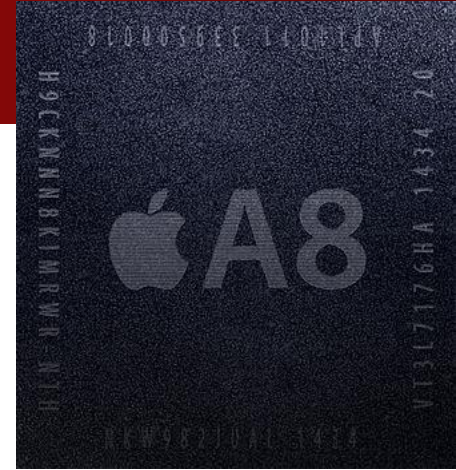
Tổ chức ARM Cache ...





Apple A8 CPU

- 2 cores
- Max. CPU clock: 1.38 GHz
- Min. feature size: 20 nm
- Instruction set: ARMv8-A
- L1 cache: Per core: 64 KB instruction + 64 KB data
- L2 cache: 1 MB shared
- L3 cache: 4 MB
- 1 GB of LPDDR3 RAM included in the package
- GPU: PowerVR Series 6XT GX6450 (quad core)
- 2 billion transistors, physical size reduced by 13% to 89 mm²
- Produced by Taiwan Semiconductor Manufacturing Company Limited (TSMC)



Apple A8X, 10/2014

- Cores: 3
- Max. CPU clock rate: 1.5 GHz
- Min. feature size: 20 nm
- Instruction set: A64, A32, T32
- Microarchitecture: Typhoon ARMv8-A-compatible
- L1 cache Per core: 64 KB instruction + 64 KB data
- L2 cache 2 MB shared
- L3 cache 4 MB
- Predecessor Apple A7
- Successor Apple A9X
- GPU PowerVR Series 6XT GXA6850 (octa-core)





Apple A9

- Cores: 2
- Max. CPU clock rate: 2.16 -2.26 GHz
- Min. feature size: 16 (TSMC)->14(Samsung) nm
- Instruction set: A64, A32, T32
- Microarchitecture: Typhoon ARMv8-A-compatible
- L1 cache/core: 64 KB instruction + 64 KB data
- L2 cache 3 MB shared
- L3 cache 4 MB (not for A9X)
- GPU : PowerVR Series 7XT GT7600 (six-core)
A9X PowerVR Series 7XT (12 cores)
- included 2 GB of LPDDR4 RAM (not for A9X – 4GB)





Apple A10 Fusion

- Cores: 2
- Max. CPU clock rate: 2.34 GHz
- Min. feature size: 16 (TSMC) nm
- Instruction set: A64, A32, T32
- Microarchitecture: Typhoon ARMv8-A-compatible
- L1 cache/core: 64 KB instruction + 64 KB data
- L2 cache 3 MB shared
- L3 cache 4 MB (not for A9X)
- GPU : (six-core)
- included the LPDDR4 RAM: 2 GB – iPhone 7; 3GB for 7+



A11 Bionic

- Min. feature size: 10 nm (4.3 billion transistors on a die 87.66 mm² in size, 41% smaller than the A10)
 - Instruction set: A64, ARMv8-A compatible
 - Cores: Hexa-core (2× Monsoon + 4× Mistral)
 - L1 cache: 64 KB instruction, 64 KB data
 - L2 cache: 8 MB
 - GPU: Apple-designed 3 core
-
- 2 GB of LPDDR4 for iPhone 8
 - 3 GB of LPDDR4 for iPhone 8 Plus and iPhone X



A12 Bionic

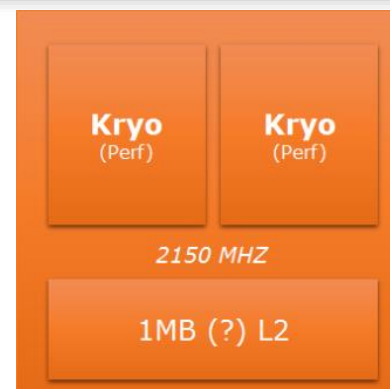
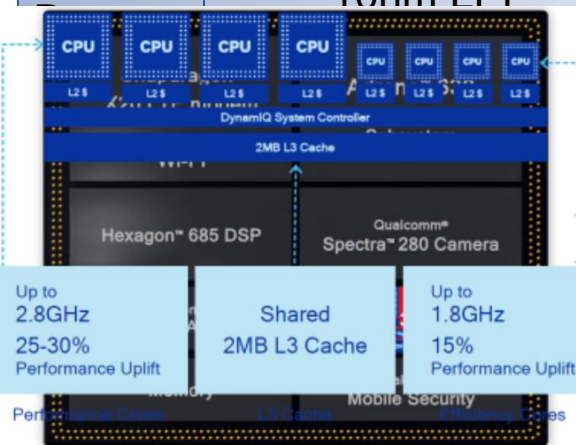
- Min. feature size: 7nm (6.9 billion transistors on a die 83.27mm² in size, 5% smaller than the A11)
 - Instruction set: A64, ARMv8.3 compatible
 - Cores: Hexa-core (2× Vortex + 4×Tempest)
 - L1 cache: 128 KB instruction, 128 KB data
 - L2 cache: 8 MB
 - GPU: Apple-designed 4 core
-
- 3 GB of LPDDR4 for iPhone XR
 - 4 GB of LPDDR4 for iPhone XS, XS Max





Qualcomm Snapdragon

SoC	Snapdragon 845	Snapdragon 835	Snapdragon 820	Snapdragon 810
CPU	4x Kryo 385 Performance 4x256MB L2 4x Kryo 845 Efficiency 4x128KB L2 2MB L3	4x Kryo 280 Performance @ 2.45GHz 2MB L2 4x Kryo 280 Efficiency @ 1.90GHz 1Mb L2	2x Kryo@1.593GHz 512KB(?) L2 cache 2x Kryo@2.150GHz 1MB(?) L2 cache	4x A53@1.555GHz 512KB L2 cache 4x A57@1.958GHz 2MB L2 cache
Memory Controller	2x 32-bit @ 1866MHz LPDDR4x 29.9GB/s	2x 32-bit @ 1866MHz LPDDR4x 29.9GB/s	2x 32-bit LPDDR4 @ 1803MHz 28.8GB/s b/w	2x 32-bit LPDDR4 @ 1555MHz 24.8GB/s b/w
GPU	Adreno 630 @ 710MHz	Adreno 540 @ 710MHz	Adreno 530 @ 624MHz	Adreno 430 @ 600MHz
Mfc.	10nm LPP	10nm LPP	14nm LPP	20nm SoC



Exynos 7 Octa 7420, 2015

- Technology: 14 nm LPE
 - Instruction Set: ARMv8-A
 - Microarchitecture: Cortex-A57+ Cortex-A53 (big.LITTLE with GTS)
 - Cores: 4 (2.1GHz) + 4 (1.5GHz)
 - GPU: Mali-T760 MP8 @ 772 MHz
-
- Samsung Galaxy S6, Samsung Galaxy S6 Edge, Samsung Galaxy S6 Active, Samsung Galaxy S6 Edge+, Samsung Galaxy Note 5





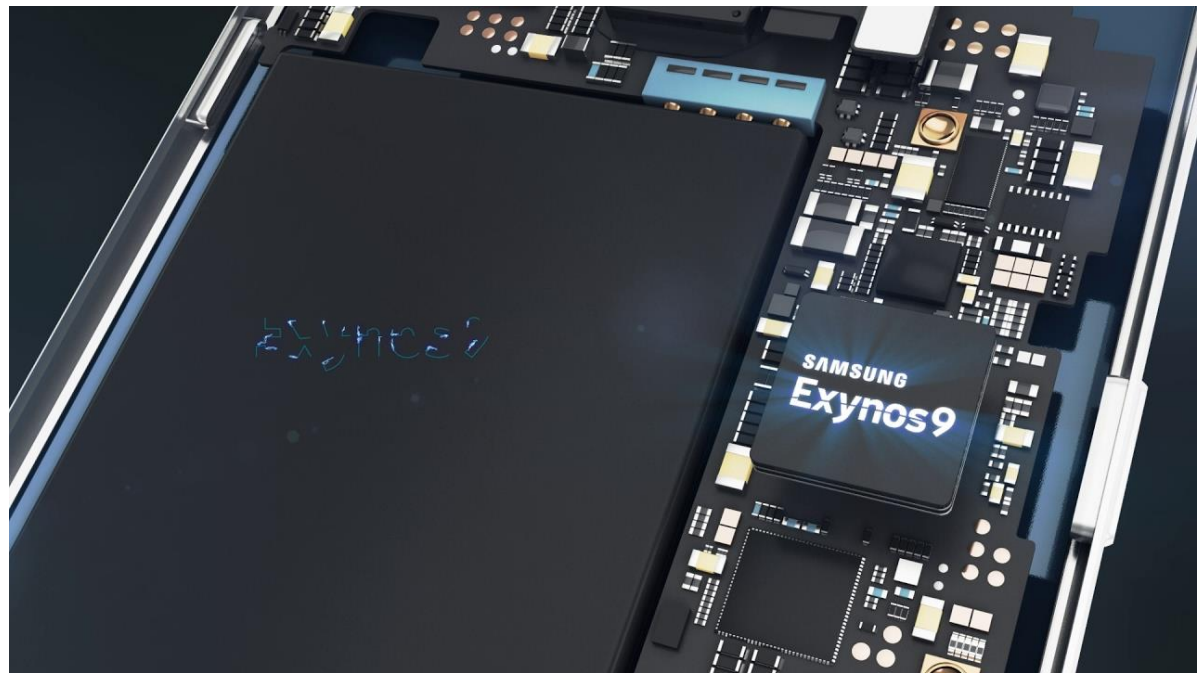
Exynos 8 Octa 8890

- Technology: 14 nm LPE
- Instruction Set: ARMv8-A
- Microarchitecture: Exynos M1 "Mongoose"+ Cortex-A53 (GTS))
- Cores: 4 (2.2-2.6GHz) + 4 (1.6GHz)
- GPU: Mali-T880 MP12 @ 650 MHz
- RAM LPDDR4, 1794 MHz
- Samsung Galaxy S7, S7 Edge, Note 7



Exynos 9 Series (8895)

- CPU: 4 x 2.5GHz Exynos M2 + 4 x 1.7GHz Cortex-A53
- GPU: Mali™-G71 MP20
- 10nm FinFET Process





Tổng kết

- Bộ nhớ cache
 - Nguyên tắc chung
 - Các phương pháp ánh xạ
 - Các giải thuật thay thế, chính sách ghi, ...