

Introduction to Cache Memory – Part#2

↳ “설계적인 저마다의 캐시 메모리에 대해서 이야기해볼 것인.”

Yoonjin Kim

Full Professor

Dept. of Computer Science
Sookmyung Women's University

Outline

- Cache memory examples of commercial CPU products
 - Selected commercial CPU products
 - Questions about commercial CPU products
(파생된 문제)
 - Deriving issues of cache memory
 - Physical characteristics of SRAM for handling issues of cache memory
(SRAM의 물리적인 특성)
※
 - Issues of cache memory
 - Multi-level cache
 - L1 cache size
 - Separate cache vs unified cache
 - Private cache vs shared cache for multi-core
- ※ **Cache 문제** (Cache Issues)
- ↳ 초기화로 갖을 수 있는 궁금증을 헤아려면서
이론을 살펴보는 것임. (교재의 단원 질문)
- ≠ 상용화된 CPU 제품 예시를
볼 것임.
-

제품의 "high performance"을 발휘하는 경우.
(우리는 95% 정도 맞음)
→ 여기서는 주제에 맞지 않음.

일반 PC "Laptop"
vs
일비디드 "터틀 PC"

구글Glass 것 ⇒ cache의 구조
→ 딜레이면 딜레이 피드백 + feedback.
FPGA는 동작하지 않을 때.
→ 일비디드와 PC 정렬을 찾음.
→ 매우 광범위한 네트워크를 넘어서.

* Group #1 과 Group #2를 놓아두

Cache Memory Examples of Commercial CPU Products

• Selected commercial CPU products – Group#1

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq	ISA	No. of Cores	L1 Cache		
① Embedded Computer Systems	Smart Phone	iPhone 4	2010. 06	45 nm	?	Apple A4	ARM Cortex-A8	32	800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None
		iPhone 4s	2011. 03	32 nm	?	Apple A5	ARM Cortex-A9	32	1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None
		iPhone 5s	2013. 09	28 nm	2~3 W	Apple A7	Apple Cyclone	64	1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC
		iPhone 6s	2015. 09	16 nm	?	Apple A9	Apple Twister	64	1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
		iPhone 7	2016. 09	16 nm	?	Apple A10 Fusion	2 x Hurricanes for high performance 2 x Zephyr for energy efficiency	64	2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	3MB shared 1MB shared	4MB shared by the entire SoC
		Galaxy S8 (High Performance)	2017. 04	10 nm	?	Exynos 8895	4x Mongoose2 (M2) for high performance 4x Cortex-A53 for energy efficiency	64	2.3 GHz 1.7 GHz	ARM v8-A	8	Per-core: I-64KB, D-32KB Per-core: I-32KB, D-32KB	2MB shared 256KB shared	None
		iPhone X	2017. 11	10 nm	?	Apple A11 Bionic	2 x Monsoon for high performance (Low performance) 4 x Mistral for energy efficiency	64	2.39 GHz 1.42 GHz	ARM v8.2-A	6	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	8MB shared 1MB shared	None
		iPhone XS	2018. 09	7 nm	?	Apple A12	2 x Vortex for high performance 4 x Tempest for energy efficiency	64	2.49 GHz 1.52 GHz	ARM v8.3-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-32KB	8MB shared 2MB shared	None
		iPhone 11 Pro	2019. 09	7 nm	?	Apple A13	2 x Lightning for high performance 4 x Thunder for energy efficiency	64	2.66 GHz 1.82 GHz	ARM v8.4-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-48KB	8MB shared 4MB shared	None
Tablet PC	HP x2 210 G2 3ML39PA	2018. 02	14 nm	2W	Intel Atom® Processor x5-Z8350	Airmont	64	1.44 ~ 1.92 GHz	x86-64	4	Per-core: I-32KB, D-24KB	2MB (1MB shared by 2 cores)	None	

→ Galaxy S9, Galaxy Note10이 출시한 날짜가 4년만 예상일 (아마도 아직까지 있는)

Intel의 ARM
ARM (Advanced RISC Machines)

→ 현재 풍경이 더 정교화되었지. 일반적인 general PC 경쟁력 부족으로.
(2011 ~ 2018년 4월)

현재는 - 노트폰은 예상
ARM 기반 경쟁력 (ARM의 부족 때문)

각종 SoC 블록의 "SoC" 생활화.

특정 "CPU" 칩을 부기하고
"Application processor" 2개를.

GPU HW Accelerator

Cache Memory Examples of Commercial CPU Products

Selected commercial CPU products – Group#2

Type	Sub Type	Product Name	Release date (Year. Month)	SoC		SoC (System-on-Chip) = CPU + GPU + HW Accelerators								
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache
① Embedded Computer Systems	Smart Phone	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance 4 x Cortex-A55 for energy efficiency	64	2.9 GHz	ARM v8-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
		Galaxy Note10	2019. 08	7 nm	?	Exynos 9825	2 x Cheetah (M4) for high performance 2 x Cortex-A75 for moderate performance 4 x Cortex-A55 for energy efficiency		1. 9 GHz	ARM v8.2-A		Per-core: I-32KB, D-32KB	256KB Shared	
	Laptop	SAMSUNG NT900X3N-K59SS	2015. 03	14 nm	15W	Intel® Core™ i5-7200U	7th Gen. Kaby Lake	64	2.73 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
		LG GRAM 17ZD90N-VX70K	2019. 12	10 nm	15W	Intel® Core™ i7-1065G7	10th Gen. Ice Lake		2.4 GHz	ARM v8.2-A		Per-core: I-64KB, D-64KB	Per-core: 256KB	
② General PC	Desktop	SAMSUNG DB400T6B	2015. 12	14 nm	65W	Intel® Core™ i7-6700	6th Gen. Skylake	64	3.4 ~ 4 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores and GPU
		Danawa 190721	2019. 07	14 nm	95W	Intel® Core™ i7-9700K	9th Gen. Coffee Lake		3.6 ~ 4.9 GHz	x86-64		Per-core: I-32KB, D-32KB	Per-core: 256KB	
	High performance Workstation	TYAN KFT46	2011. 09	32 nm	80W	Intel® Xeon ® E5606	Westmere WikiChip Block diagram	64	2.13 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores
		Lenovo ThinkStation P500TW 850W	2015. 07	22 nm	140W	Intel® Xeon ® E5-1630 v3	Haswell		3.7 ~ 3.8 GHz	x86-64		Per-core: I-32KB, D-32KB	Per-core: 256KB	
③ Server/Workstation	TYAN TAKO-KQT44	2020. 04	14 nm	150W	Intel® Xeon Gold 6226R	Cascade Lake	64	2.9 ~ 3.9 GHz	x86-64	16	Per-core: I-32KB, D-32KB	Per-core: 1MB	22MB shared by 16 cores	

General PC/Workstation 대비 14nm 제작. (디자인이나 차이 때문... why?)

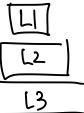
[02 : 30]

Cache Memory Examples of Commercial CPU Products

- Questions about commercial CPU products

- Question#1: Why are caches organized in the multi-level manner?

→ L1 cache memory (L1, L2, L3)
multi-level cache organization?



Group#1

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache		
Embedded Computer Systems	Smart Phone	iPhone 4	2010. 06	45 nm	?	Apple A4	ARM Cortex-A8	32	800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None
		iPhone 4s	2011. 03	32 nm	?	Apple A5	ARM Cortex-A9	32	1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None
		iPhone 5s	2013. 09	28 nm	2~3 W	Apple A7	Apple Cyclone	64	1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC
		iPhone 6s	2015. 09	16 nm	?	Apple A9	Apple Twister	64	1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
		iPhone 7	2016. 09	16 nm	?	Apple A10 Fusion	2 x Hurricanes for high performance 2 x Zephyr for energy efficiency	64	2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
		Galaxy S8	2017. 04	10 nm	?	Exynos 8895	4 x Mongoose2 (M2) for high performance 4 x Cortex-A53 for energy efficiency	64	2.3 GHz	ARM v8-A	8	Per-core: I-32KB, D-32KB	1MB shared	
		iPhone X	2017. 11	10 nm	?	Apple A11 Bionic	2 x Monsoon for high performance 4 x Mistral for energy efficiency	64	1.42 GHz	ARM v8.2-A	6	Per-core: I-64KB, D-32KB	2MB shared	None
		iPhone XS	2018. 09	7 nm	?	Apple A12	2 x Vortex for high performance 4 x Tempest for energy efficiency	64	2.39 GHz	ARM v8.3-A	6	Per-core: I-32KB, D-32KB	256KB shared	
		iPhone 11 Pro	2019. 09	7 nm	?	Apple A13	2 x Lightning for high performance 4 x Thunder for energy efficiency	64	1.52 GHz	ARM v8.4-A	6	Per-core: I-128KB, D-128KB	8MB shared	None
		Tablet PC	HP x2 210 G2 3ML39PA	2018. 02	14 nm	2W	Intel Atom® Processor x5-Z8350	Airmont	64	1.44 ~ 1.92 GHz	x86-64	4	Per-core: I-32KB, D-24KB 2MB (1MB shared by 2 cores)	4MB shared

Cache Memory Examples of Commercial CPU Products

- Questions about commercial CPU products

- Question#1:** Why are caches organized in the multi-level manner?

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache
Group#2	Embedded Computer Systems	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance	64	2.9 GHz	ARM v8-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
							4 x Cortex-A55 for energy efficiency		1. 9 GHz	ARM v8.2-A		Per-core: I-32KB, D-32KB	256KB Shared	
		Galaxy Note10	2019. 08	7 nm	?	Exynos 9825	2 x Cheetah (M4) for high performance	64	2.73 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
							2 x Cortex-A75 for moderate performance		2.4 GHz			Per-core: I-64KB, D-64KB	Per-core: 256KB	
	General PC	SAMSUNG NT900X3N-K59SS	2015. 03	14 nm	15W	Intel® Core™ i5-7200U	7th Gen. Kaby Lake		2.5 ~ 3.1 GHz	x86 -64	2	Per-core: I-32KB, D-32KB	Per-core: 256KB	3MB shared by 2 cores and GPU
							LG GRAM 17ZD90N-YX70K		1.3 ~ 3.9 GHz	x86 -64		Per-core: I-32KB, D-48KB	Per-core: 512KB	8MB shared by 4 cores and GPU
		SAMSUNG DB400T6B	2015. 12	14 nm	65W	Intel® Core™ i7-6700	6th Gen. Skylake		3.4 ~ 4 GHz	x86 -64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores and GPU
		Danawa 190721	2019. 07	14 nm	95W	Intel® Core™ i7-9700K	9th Gen. Coffee Lake		3.6 ~ 4.9 GHz	x86 -64		Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores and GPU
Server/ Workstation	Floating License/ Data Sever	TYAN KFT46	2011. 09	32 nm	80W	Intel® Xeon ® E5606	Westmere WikiChip Block diagram	64	2.13 GHz	x86 -64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores
	High performance Workstation	Lenovo ThinkStation P500TW 850W	2015. 07	22 nm	140W	Intel® Xeon ® E5-1630 v3	Haswell		3.7 ~ 3.8 GHz	x86 -64		Per-core: I-32KB, D-32KB	Per-core: 256KB	10MB shared by 4 cores
		TYAN TAKO-KQT44	2020. 04	14 nm	150W	Intel® Xeon Gold 6226R	Cascade Lake	64	2.9 ~ 3.9 GHz	x86 -64	16	Per-core: I-32KB, D-32KB	Per-core: 1MB	22MB shared by 16 cores

→ L1 캐시는 딱히 필요.
(in Group 2)

Corest
I-cache, D-cache
→ 디자인 특성.

I : Instruction Fetch (Fetch)
D : Data Fetch (Memory)

Cache Memory Examples of Commercial CPU Products

- Questions about commercial CPU products → 왜 L1 캐시 I, D 캐시는 32KB 를 선택하는가!?

- Question#2-1: Why do most of non-embedded CPUs have 32 KB L1 caches?
(= General PC & Work station)

Type	Sub Type	Product Name	Release date (Year, Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators									
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	
Group#2	Embedded Computer Systems	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance	64	2.9 GHz	ARM v8-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB
							4 x Cortex-A55 for energy efficiency		1. 9 GHz	ARM v8.2-A		Per-core: I-32KB, D-32KB	256KB Shared
		Galaxy Note10	2019. 08	7 nm	?	Exynos 9825	2 x Cheetah (M4) for high performance	64	2.73 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB
							2 x Cortex-A75 for moderate performance		2.4 GHz			Per-core: I-64KB, D-64KB	256KB
	General PC	SAMSUNG NT900X3N-K59SS	2015. 03	14 nm	15W	Intel® Core™ i5-7200U	7th Gen. Kaby Lake		2.5 ~ 3.1 GHz	x86-64	2	Per-core: I-32KB, D-32KB	3MB shared by 2 cores and GPU
							LG GRAM 17ZD90N-YX70K		1.3 ~ 3.9 GHz	x86-64		Per-core: I-32KB, D-48KB	Per-core: 512KB
		SAMSUNG DB400T6B	2015. 12	14 nm	65W	Intel® Core™ i7-6700	6th Gen. Skylake	64	3.4 ~ 4 GHz	x86-64	4	Per-core: I-32KB, D-32KB	8MB shared by 4 cores and GPU
		Danawa 190721	2019. 07	14 nm	95W	Intel® Core™ i7-9700K	9th Gen. Coffee Lake		3.6 ~ 4.9 GHz	x86-64		Per-core: I-32KB, D-32KB	12MB shared by 4 cores and GPU
Server/Workstation	Floating License/ Data Sever	TYAN KFT46	2011. 09	32 nm	80W	Intel® Xeon ® E5606	Westmere WikiChip Block diagram	64	2.13 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB
	High performance Workstation	Lenovo ThinkStation P500TW 850W	2015. 07	22 nm	140W	Intel® Xeon ® E5-1630 v3	Haswell		3.7 ~ 3.8 GHz	x86-64		Per-core: I-32KB, D-32KB	Per-core: 256KB
		TYAN TAKO-KQT44	2020. 04	14 nm	150W	Intel® Xeon Gold 6226R	Cascade Lake	64	2.9 ~ 3.9 GHz	x86-64	16	Per-core: I-32KB, D-32KB	Per-core: 1MB

Cache Memory Examples of Commercial CPU Products

- Questions about commercial CPU products

- Question#2-2:** Why do most of embedded CPUs have larger L1 caches (64KB, 128KB) than L1 caches (32 KB, 48KB) of non-embedded CPUs?

→ 100% non-embedded CPU often have L1 cache 2x larger than embedded CPU
 L1 cache size is larger?

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache		
Group#1	Embedded Computer Systems	iPhone 4	2010. 06	45 nm	?	Apple A4	ARM Cortex-A8	32	800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None
		iPhone 4s	2011. 03	32 nm	?	Apple A5	ARM Cortex-A9	32	1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None
		iPhone 5s	2013. 09	28 nm	2~3 W	Apple A7	Apple Cyclone	64	1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC
		iPhone 6s	2015. 09	16 nm	?	Apple A9	Apple Twister	64	1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
		iPhone 7	2016. 09	16 nm	?	Apple A10 Fusion	2 x Hurricanes for high performance 2 x Zephyr for energy efficiency	64	2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
		Galaxy S8	2017. 04	10 nm	?	Exynos 8895	4 x Mongoose2 (M2) for high performance 4 x Cortex-A53 for energy efficiency	64	2.3 GHz 1.7 GHz	ARM v8-A	8	Per-core: I-32KB, D-32KB	1MB shared	
		iPhone X	2017. 11	10 nm	?	Apple A11 Bionic	2 x Monsoon for high performance 4 x Mistral for energy efficiency	64	2.39 GHz 1.42 GHz	ARM v8.2-A	6	Per-core: I-64KB, D-64KB	8MB shared	None
		iPhone XS	2018. 09	7 nm	?	Apple A12	2 x Vortex for high performance 4 x Tempest for energy efficiency	64	2.49 GHz 1.52 GHz	ARM v8.3-A	6	Per-core: I-32KB, D-32KB	1MB shared	
		iPhone 11 Pro	2019. 09	7 nm	?	Apple A13	2 x Lightning for high performance 4 x Thunder for energy efficiency	64	2.66 GHz 1.82 GHz	ARM v8.4-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-32KB	8MB shared 2MB shared	None
		Tablet PC	HP x2 210 G2 3ML39PA	2018. 02	14 nm	2W	Intel Atom® Processor x5-Z8350	Airmont	64	1.44 ~ 1.92 GHz	x86-64	4	Per-core: I-32KB, D-24KB Per-core: I-32KB, D-48KB	2MB (1MB shared by 2 cores)

Cache Memory Examples of Commercial CPU Products

- Questions about commercial CPU products

- Question#2-2:** Why do most of embedded CPUs have large L1 caches (64KB, 128KB) compared with L1 caches (32 KB, 48KB) of non-embedded CPUs?

Core of [29M12]
[TC] [TJL]
[N2]
"TKI" [TJL]
[D2]

Type	Sub Type	Product Name	Release date (Year, Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	CPU			ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache
							Micro-architecture	Bit-Width	Clock Freq.					
Group#2														
Embedded Computer Systems	Smart Phone	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance 4 x Cortex-A55 for energy efficiency	64	2.9 GHz 1. 9 GHz	ARM v8-A ARM v8.2-A	8	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-core: 512KB 256KB Shared	4MB shared
		Galaxy Note10	2019. 08	7 nm	?	Exynos 9825	2 x Cheetah (M4) for high performance 2 x Cortex-A75 for moderate performance 4 x Cortex-A55 for energy efficiency	64	2.73 GHz 2.4 GHz 1.95 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-core: 512KB 256KB Per-core: 128KB	4MB shared
General PC	Laptop	SAMSUNG NT900X3N-K59SS	2015. 03	14 nm	15W	Intel® Core™ i5-7200U	7th Gen. Kaby Lake	64	2.5 ~ 3.1 GHz	x86-64	2	Per-core: I-32KB, D-32KB	Per-core: 256KB 3MB shared by 2 cores and GPU	
		LG GRAM 17ZD90N-VX70K	2019. 12	10 nm	15W	Intel® Core™ i7-1065G7	10th Gen. Ice Lake	64	1.3 ~ 3.9 GHz	x86-64	4	Per-core: I-32KB, D-48KB	Per-core: 512KB 8MB shared by 4 cores and GPU	
	Desktop	SAMSUNG DB400T6B	2015. 12	14 nm	65W	Intel® Core™ i7-6700	6th Gen. Skylake	64	3.4 ~ 4 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB 8MB shared by 4 cores and GPU	
		Danawa 190721	2019. 07	14 nm	95W	Intel® Core™ i7-9700K	9th Gen. Coffee Lake	64	3.6 ~ 4.9 GHz	x86-64	8	Per-core: I-32KB, D-32KB	Per-core: 256KB 12MB shared by 4 cores and GPU	
Server/Workstation	Floating License/ Data Sever	TYAN KFT46	2011. 09	32 nm	80W	Intel® Xeon ® E5606	Westmere WikiChip Block diagram	64	2.13 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB 8MB shared by 4 cores	
	High performance Workstation	Lenovo ThinkStation P500TW 850W	2015. 07	22 nm	140W	Intel® Xeon ® E5-1630 v3	Haswell	64	3.7 ~ 3.8 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB 10MB shared by 4 cores	
		TYAN TAKO-KQT44	2020. 04	14 nm	150W	Intel® Xeon Gold 6226R	Cascade Lake	64	2.9 ~ 3.9 GHz	x86-64	16	Per-core: I-32KB, D-32KB	Per-core: 1MB 22MB shared by 16 cores	

L1 cache, D-cache \rightarrow L1 cache
L2, L3 cache \rightarrow L2, L3 cache
 \rightarrow L1+L2+L3 cache \rightarrow L1+L2+L3 cache

Cache Memory Examples of Commercial CPU Products

- Questions about commercial CPU products

- Question#3: Why are L1 caches separate and L2/L3 caches unified?
(separate L1, L2, L3) (unified L1+L2+L3)

Group#1

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache		
Embedded Computer Systems	Smart Phone	iPhone 4	2010. 06	45 nm	?	Apple A4	ARM Cortex-A8	32	800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None
		iPhone 4s	2011. 03	32 nm	?	Apple A5	ARM Cortex-A9	32	1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None
		iPhone 5s	2013. 09	28 nm	2~3 W	Apple A7	Apple Cyclone	64	1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC
		iPhone 6s	2015. 09	16 nm	?	Apple A9	Apple Twister	64	1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
		iPhone 7	2016. 09	16 nm	?	Apple A10 Fusion	2 x Hurricanes for high performance	64	2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
							2 x Zephyr for energy efficiency					Per-core: I-32KB, D-32KB	1MB shared	
		Galaxy S8	2017. 04	10 nm	?	Exynos 8895	4 x Mongoose2 (M2) for high performance	64	2.3 GHz	ARM v8-A	8	Per-core: I-64KB, D-32KB	2MB shared	None
							4 x Cortex-A53 for energy efficiency		1.7 GHz			Per-core: I-32KB, D-32KB	256KB shared	
		iPhone X	2017. 11	10 nm	?	Apple A11 Bionic	2 x Monsoon for high performance	64	2.39 GHz	ARM v8.2-A	6	Per-core: I-64KB, D-64KB	8MB shared	None
							4 x Mistral for energy efficiency		1.42 GHz			Per-core: I-32KB, D-32KB	1MB shared	
		iPhone XS	2018. 09	7 nm	?	Apple A12	2 x Vortex for high performance	64	2.49 GHz	ARM v8.3-A	6	Per-core: I-128KB, D-128KB	8MB shared	None
							4 x Tempest for energy efficiency		1.52 GHz			Per-core: I-32KB, D-32KB	2MB shared	
		iPhone 11 Pro	2019. 09	7 nm	?	Apple A13	2 x Lightning for high performance	64	2.66 GHz	ARM v8.4-A	6	Per-core: I-128KB, D-128KB	8MB shared	None
							4 x Thunder for energy efficiency		1.82 GHz			Per-core: I-32KB, D-48KB	4MB shared	
Tablet PC	HP x2 210 G2 3ML39PA	2018. 02	14 nm	2W	Intel Atom® Processor x5-Z8350	Airmont		64	1.44 ~ 1.92 GHz	x86-64	4	Per-core: I-32KB, D-24KB	2MB (1MB shared by 2 cores)	None

Cache Memory Examples of Commercial CPU Products

- Questions about commercial CPU products

- Question#3: Why are L1 caches separate and L2/L3 caches unified?

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators											
				Fab. Tech	Power	Name	CPU					No. of Cores	L1 Cache	L2 Cache	L3 Cache
							Micro-architecture	Bit-Width	Clock Freq.	ISA					
Group#2	Embedded Computer Systems	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance	64	2.9 GHz	ARM v8-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared	
							4 x Cortex-A55 for energy efficiency		1. 9 GHz	ARM v8.2-A		Per-core: I-32KB, D-32KB	256KB Shared		
		Galaxy Note10	2019. 08	7 nm	?	Exynos 9825	2 x Cheetah (M4) for high performance	64	2.73 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB		
							2 x Cortex-A75 for moderate performance		2.4 GHz			Per-core: I-64KB, D-64KB	256KB		
							4 x Cortex-A55 for energy efficiency		1.95 GHz			Per-core: I-32KB, D-32KB	128KB		
		SAMSUNG NT900X3N-K59SS	2015. 03	14 nm	15W	Intel® Core™ i5-7200U	7th Gen. Kaby Lake	64	2.5 ~ 3.1 GHz	x86-64	2	Per-core: I-32KB, D-32KB	Per-core: 256KB	3MB shared by 2 cores and GPU	
							LG GRAM 17ZD90N-VX70K		1.3 ~ 3.9 GHz	x86-64		Per-core: I-32KB, D-48KB	Per-core: 512KB	8MB shared by 4 cores and GPU	
General PC	Laptop	SAMSUNG DB400T6B	2015. 12	14 nm	65W	Intel® Core™ i7-6700	10th Gen. Ice Lake	64	3.4 ~ 4 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores and GPU	
		Danawa 190721	2019. 07	14 nm	95W	Intel® Core™ i7-9700K	6th Gen. Skylake	64	3.6 ~ 4.9 GHz	x86-64	8	Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores and GPU	
	Desktop	TYAN KFT46	2011. 09	32 nm	80W	Intel® Xeon ® E5606	Westmere	64	2.13 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores	
		Lenovo ThinkStation P500TW 850W	2015. 07	22 nm	140W	Intel® Xeon ® E5-1630 v3	WikiChip		3.7 ~ 3.8 GHz	x86-64		Per-core: I-32KB, D-32KB	Per-core: 256KB	10MB shared by 4 cores	
Server/Workstation	High performance Workstation	TYAN TAKO-KQT44	2020. 04	14 nm	150W	Intel® Xeon Gold 6226R	Haswell	64	2.9 ~ 3.9 GHz	x86-64	16	Per-core: I-32KB, D-32KB	Per-core: 1MB	22MB shared by 16 cores	

Cache Memory Examples of Commercial CPU Products

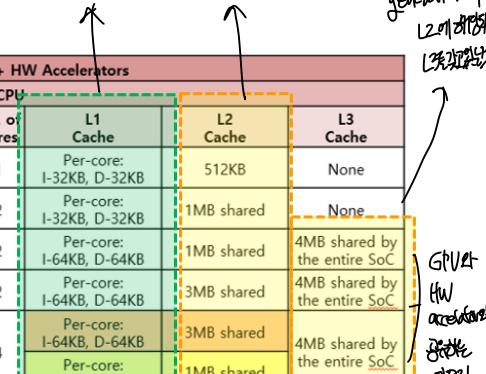
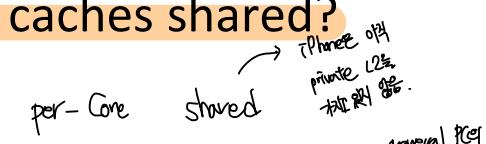
- Questions about commercial CPU products

- Question#4: Why are some caches private and other caches shared?

⇒ 어떤 캐시가 private이고 다른 캐시는 shared인가?

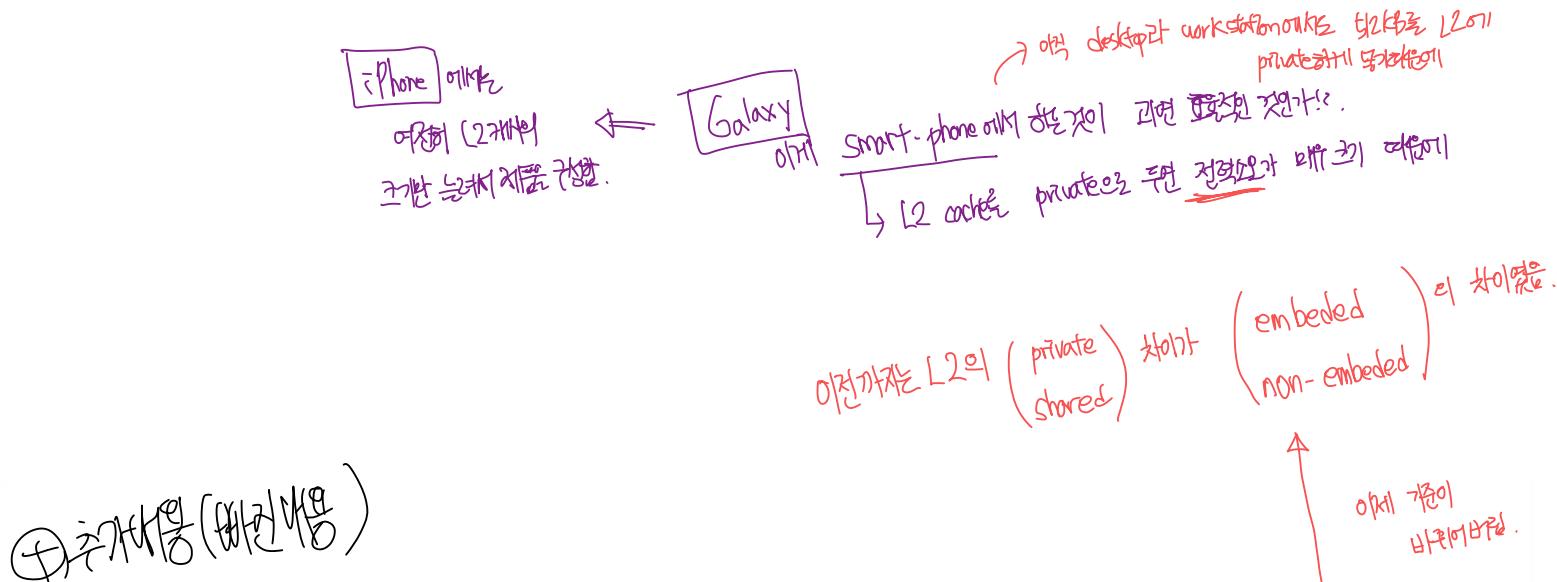
Type	Sub Type	Product Name	Release date (Year, Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	CPU			No. of Cores	L1 Cache	L2 Cache	L3 Cache	
							Micro-architecture	Bit-Width	Clock Freq.	ISA				
Group#1	Smart Phone	iPhone 4	2010. 06	45 nm	?	Apple A4	ARM Cortex-A8	32	800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None
		iPhone 4s	2011. 03	32 nm	?	Apple A5	ARM Cortex-A9	32	1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None
		iPhone 5s	2013. 09	2~3 nm	W	Apple A7	Apple Cyclone	64	1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC
		iPhone 6s	2015. 09	16 nm	?	Apple A9	Apple Twister	64	1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
		iPhone 7	2016. 09	16 nm	?	Apple A10 Fusion	2 x Hurricane for high performance 2 x Zephyr for energy efficiency	64	2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	3MB shared 1MB shared	4MB shared by the entire SoC
		Galaxy S8	2017. 04	10 nm	?	Exynos 8895	4 x Mongoose2 (M2) for high performance 4 x Cortex-A53 for energy efficiency	64	2.3 GHz 1.7 GHz	ARM v8-A	8	Per-core: I-64KB, D-32KB Per-core: I-32KB, D-32KB	2MB shared 256KB shared	None
		iPhone X	2017. 11	10 nm	?	Apple A11 Bionic	2 x Monsoon for high performance 4 x Mistral for energy efficiency	64	2.39 GHz 1.42 GHz	ARM v8.2-A	6	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	8MB shared 1MB shared	None
		iPhone XS	2018. 09	7 nm	?	Apple A12	2 x Vortex for high performance 4 x Tempest for energy efficiency	64	2.49 GHz 1.52 GHz	ARM v8.3-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-32KB	8MB shared 2MB shared	None
		iPhone 11 Pro	2019. 09	7 nm	?	Apple A13	2 x Lightning for high performance 4 x Thunder for energy efficiency	64	2.66 GHz 1.82 GHz	ARM v8.4-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-48KB	8MB shared 4MB shared	None
	Tablet PC	HP x2 210 G2 3ML39PA	2018. 02	14 nm	2W	Intel Atom® Processor x5-Z8350	Airmont	64	1.44 ~ 1.92 GHz	x86 -64	4	Per-core: I-32KB, D-24KB	2MB (1MB shared by 2 cores)	None

private ⇒ core당 개별적인
shared ⇒ core끼리 공유되는



GPU HW accelerators have their own cache.

L2 cache shared.
Tablet PC는
Laptop은 T2이라는 차이가 있다.
(Windows 10은 대체로 T2이다.)



Cache Memory Examples of Commercial CPU Products

- Questions about commercial CPU products

- Question#4: Why are some caches private and other caches shared?

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache
Group#2	Embedded Computer Systems	Smart Phone	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance 4 x Cortex-A55 for energy efficiency	64 2.9 GHz 1. 9 GHz	ARM v8-A ARM v8.2-A	8	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-core: 512KB Shared	4MB
			Galaxy Note10	2019. 08	7 nm	?	Exynos 9825	2 x Cheetah (M4) for high performance 2 x Cortex-A75 for moderate performance 4 x Cortex-A55 for energy efficiency	64 2.73 GHz 2.4 GHz 1.95 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-cores 512KB Per-cores 256KB Per-cores 128KB	4MB
		Laptop	SAMSUNG NT900X3N-K59SS	2015. 03	14 nm	15W	Intel® Core™ i5-7200U	7th Gen. Kaby Lake	64 2.5 ~ 3.1 GHz	x86-64	2	Per-core: I-32KB, D-32KB	Per-cores 256KB	MB shared by 2 cores and GPU
			LG GRAM 17ZD90N-VX70K	2019. 12	10 nm	15W	Intel® Core™ i7-1065G7	10th Gen. Ice Lake	64 1.3 ~ 3.9 GHz	x86-64	4	Per-core: I-32KB, D-48KB	Per-cores 512KB	8MB shared by 4 cores and GPU
	General PC	Desktop	SAMSUNG DB400T6B	2015. 03	14 nm	65W	Intel® Core™ i7-6700	6th Gen. Skylake	64 3.4 ~ 4 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-cores 256KB	MB shared by 4 cores and GPU
			Danawa 190721	2017. 10	14 nm	95W	Intel® Core™ i7-9700K	9th Gen. Coffee Lake	64 3.6 ~ 4.9 GHz	x86-64	8	Per-core: I-32KB, D-32KB	Per-cores 256KB	12MB shared by 4 cores and GPU
		Floating License/ Data Server	TIAN KFT46	2011. 09	32 nm	80W	Intel® Xeon ® E5606	Westmere WikiChip Block diagram	64 2.13 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-cores 256KB	8MB shared by 4 cores
Server/ Workstation	High performance Workstation	Lenovo ThinkStation P500TW 850W	2015. 07	22 nm	140W	Intel® Xeon ® E5-1630 v3	Haswell	64 3.7 ~ 3.8 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-cores 256KB	10MB shared by 4 cores and GPU	
		TIAN TAKO- KQT44	2020. 04	14 nm	150W	Intel® Xeon Gold 6226R	Cascade Lake	64 2.9 ~ 3.9 GHz	x86-64	16	Per-core: I-32KB, D-32KB	Per-cores 1MB	22MB shared by 16 cores	

(32bit shared) ← core of L2 cache ← general PC & workstation ← L2 cache shared by 16 cores.

Deriving Issues of Cache Memory

- Cache memory examples of commercial CPU products (교제용 메모리 단위 품질)
 - Question#1: Why are caches organized in the multi-level manner?
 - Question#2-1: Why do most of non-embedded CPUs have 32 KB L1 caches?
 - Question#2-2: Why do most of embedded CPUs have larger L1 caches (64KB, 128KB) than L1 caches (32 KB, 48KB) of non-embedded CPUs?
 - Question#3: Why are L1 caches separate and L2/L3 caches unified?
 - Question#4: Why are some caches private and other caches shared?
-
- Issues of cache memory
 - Multi-level cache : Answer#1 for Question#1
 - L1 cache size : Answer#2 for Question#2
 - Separate cache vs unified cache: Answer#3 for Question#3
 - Private cache vs shared cache for multi-core: Answer#4 for Question#4

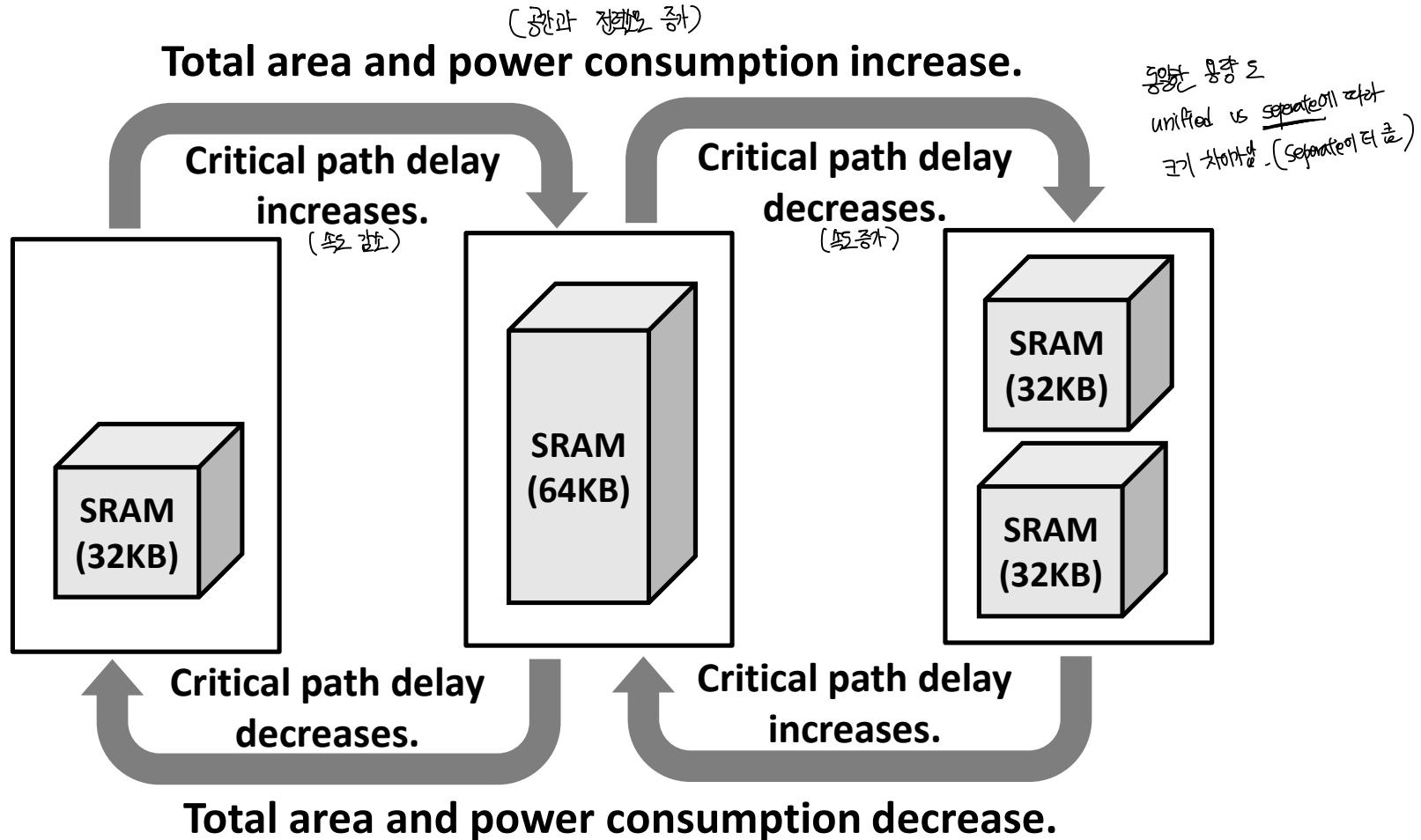
↳ 복수개의 코어가 있는 시스템에서는 각각의 코어가 자신의 독립적인 캐시를 갖고 있다.

Separate 캐시
Unified 캐시를 보관하는
방법은 종종 접근성을 보장하기 어렵다.

(SRAM의 물리적 특징 → 캐시에 있는 handling issue)

Physical Characteristics of SRAM for Handling Issues of Cache Memory

- For example,



Issues of Cache Memory

→ **Multi-level Cache** (Multi-level 캐시는 어떤가요?)

- Multi-Level Cache

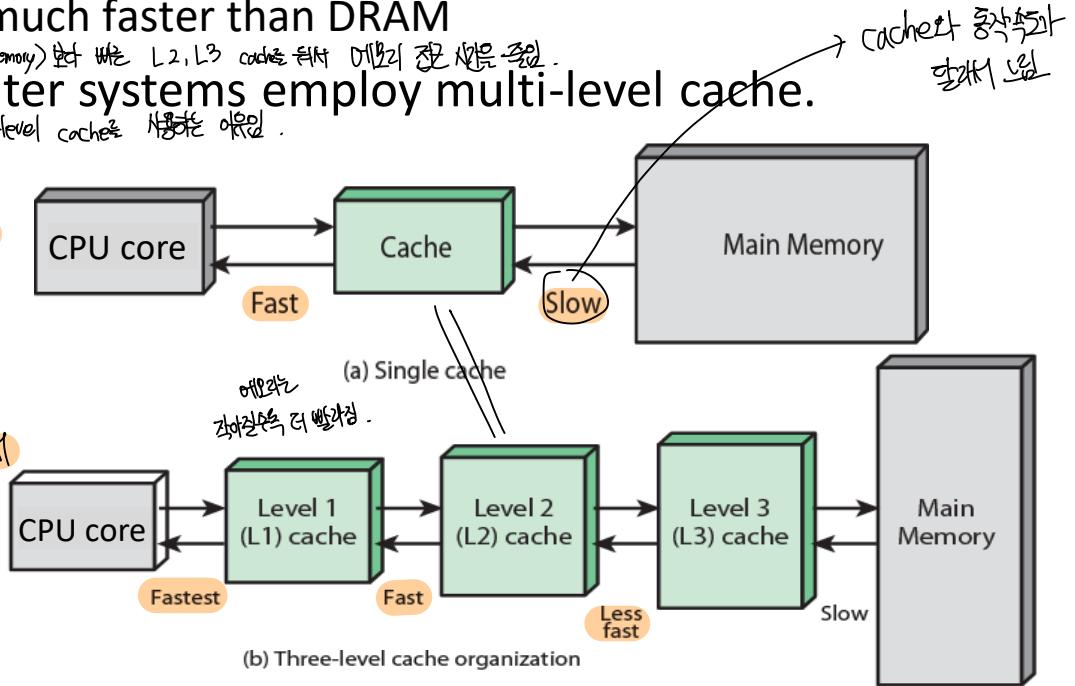
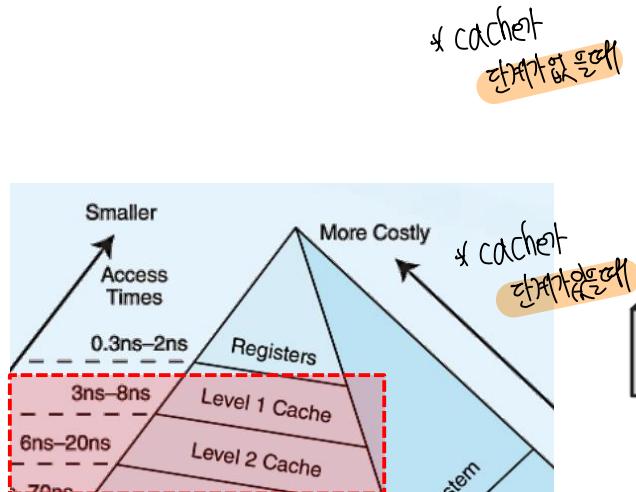
- Questions about commercial CPU products
 - Question#1: Why are caches organized in the multi-level manner?
 - Answer#1: For improving performance by reducing memory access time

				Type	Sub Type	Product Name	Release date (Year. Month)	Fab. Tech	Power	Name	SoC (System-on-Chip) = CPU + GPU + HW Accelerators											
											Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	CPU						
																L1 Cache	L2 Cache	L3 Cache				
Group#1				Group#2							Galaxy S9	2018. Q3	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance	2.9 GHz	ARM v8-A	8	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-core: 512KB 256KB Shared 256KB 128KB	4MB shared
											SoC (System-on-Chip) = CPU + GPU + HW Accelerators											
											CPU											
											Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache				
Embedded Computer Systems	Smart Phone	iPhone 4	2010. 06	45 nm	?	Apple A4	ARM Cortex-A8	32	800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None	8	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-core: 512KB 256KB Shared 256KB 128KB	4MB shared				
		iPhone 4s	2011. 03	32 nm	?	Apple A5	ARM Cortex-A9	32	1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None	8	Per-core: I-64KB, D-64KB Per-core: I-64KB, D-64KB	Per-core: 512KB 256KB	4MB shared				
		iPhone 5s	2013. 09	28 nm	2~3 W	Apple A7	Apple Cyclone	64	1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC	8	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-core: 512KB 256KB	4MB shared				
		iPhone 6s	2015. 09	16 nm	?	Apple A9	Apple Twister	64	1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC	8	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-core: 512KB 256KB	4MB shared				
		iPhone 7	2016. 09	16 nm	?	Apple A10 Fusion	2 x Hurricanes for high performance 2 x Zephyr for energy efficiency	64	2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	3MB shared	4MB shared by the entire SoC	2	Per-core: I-32KB, D-32KB	Per-core: 256KB	3MB shared by 2 cores and GPU				
		Galaxy S8	2017. 04	10 nm	?	Exynos 8895	4 x Mongoose2 (M2) for high performance 4 x Cortex-A53 for energy efficiency	64	2.3 GHz 1.7 GHz	ARM v8-A	8	Per-core: I-64KB, D-32KB Per-core: I-32KB, D-32KB	2MB shared 256KB shared	None	4	Per-core: I-32KB, D-48KB	Per-core: 512KB	8MB shared by 4 cores and GPU				
		iPhone X	2017. 11	10 nm	?	Apple A11 Bionic	2 x Monsoon for high performance 4 x Mistral for energy efficiency	64	2.39 GHz 1.42 GHz	ARM v8.2-A	6	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	8MB shared 1MB shared	None	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores and GPU				
		iPhone XS	2018. 09	7 nm	?	Apple A12	2 x Vortex for high performance 4 x Tempest for energy efficiency	64	2.49 GHz 1.52 GHz	ARM v8.3-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-32KB	8MB shared 2MB shared	None	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores				
		iPhone 11 Pro	2019. 09	7 nm	?	Apple A13	2 x Lightning for high performance 4 x Thunder for energy efficiency	64	2.66 GHz 1.82 GHz	ARM v8.4-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-48KB	8MB shared 4MB shared	None	4	Per-core: I-32KB, D-32KB	Per-core: 1MB	10MB shared by 4 cores				
	Tablet PC	HP x2 210 G2 3ML39PA	2018. 02	14 nm	2W	Intel Atom® Processor x5-Z8350	Airmont	64	1.44 ~ 1.92 GHz	x86-64	4	Per-core: I-32KB, D-24KB	2MB (1MB shared by 2 cores)	None	16	Per-core: I-32KB, D-32KB	Per-core: 1MB	22MB shared by 16 cores				

Issues of Cache Memory

- Multi-Level Cache

- Cache performance can be improved by two cases.
– Case#1: Enhancing hit-rate ⇒ 빠른 데 달라졌다
– Case#2: Reducing memory access time ⇒ 메모리 접근 시간을 줄인다.
- Purpose of multi-level cache organization ⇒ Multi-level로 cache를 구성을 하면↓
– In order to improve cache performance by case#2
 ▪ L2/L3 cache access much faster than DRAM
 ⇒ DRAM (Main Memory) 보다 빠른 L2, L3 cache에서 메모리 접근 시간은 짧다.
– Most of today's computer systems employ multi-level cache.
 ⇒ 이것은 multi-level cache는 네트워크 때문.

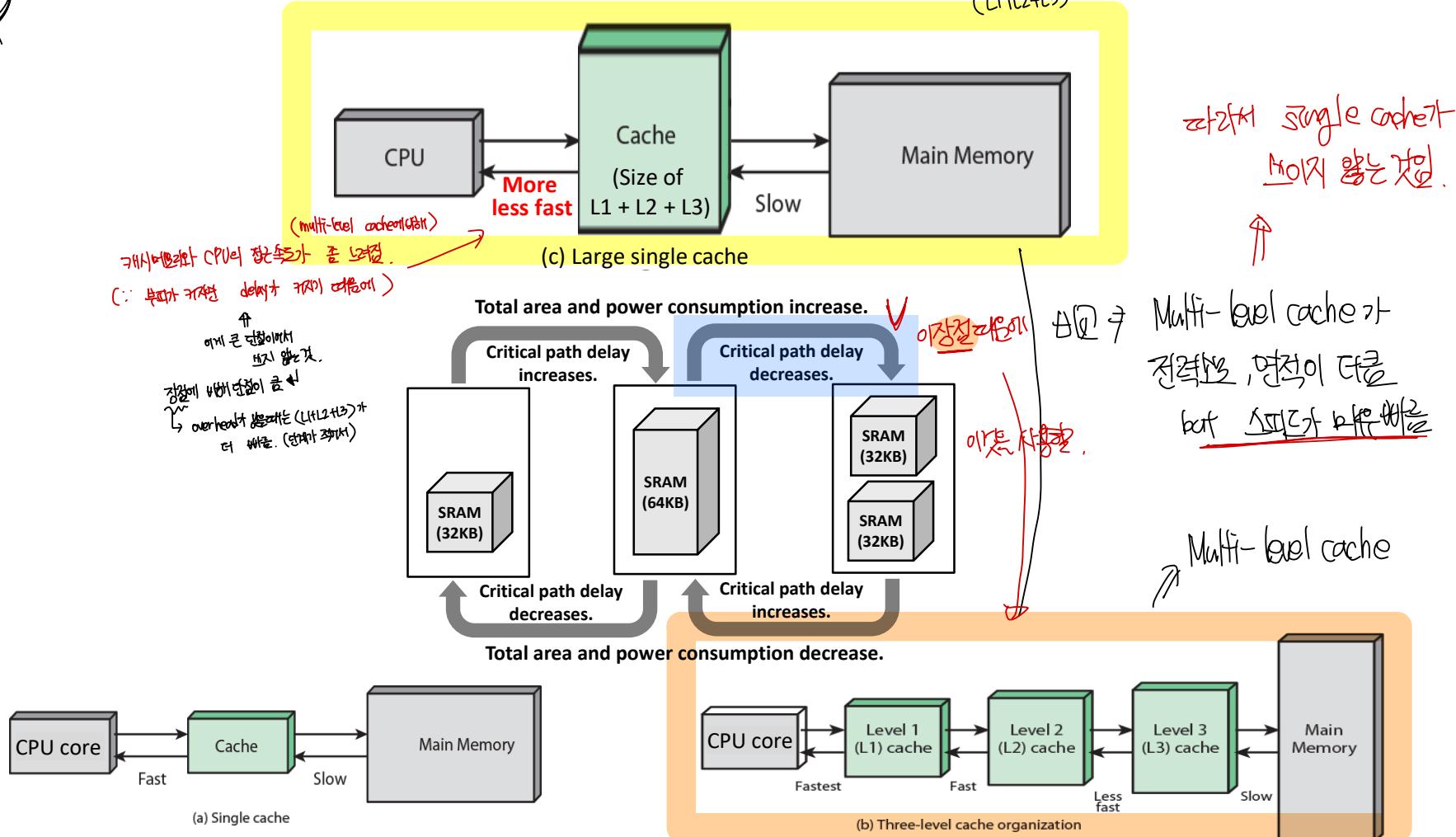


Issues of Cache Memory

- Multi-Level Cache



How about only a large single cache? ⇒ 허나의 큰 single cache를 주면 안되는 것인가?
(L1+L2+L3)

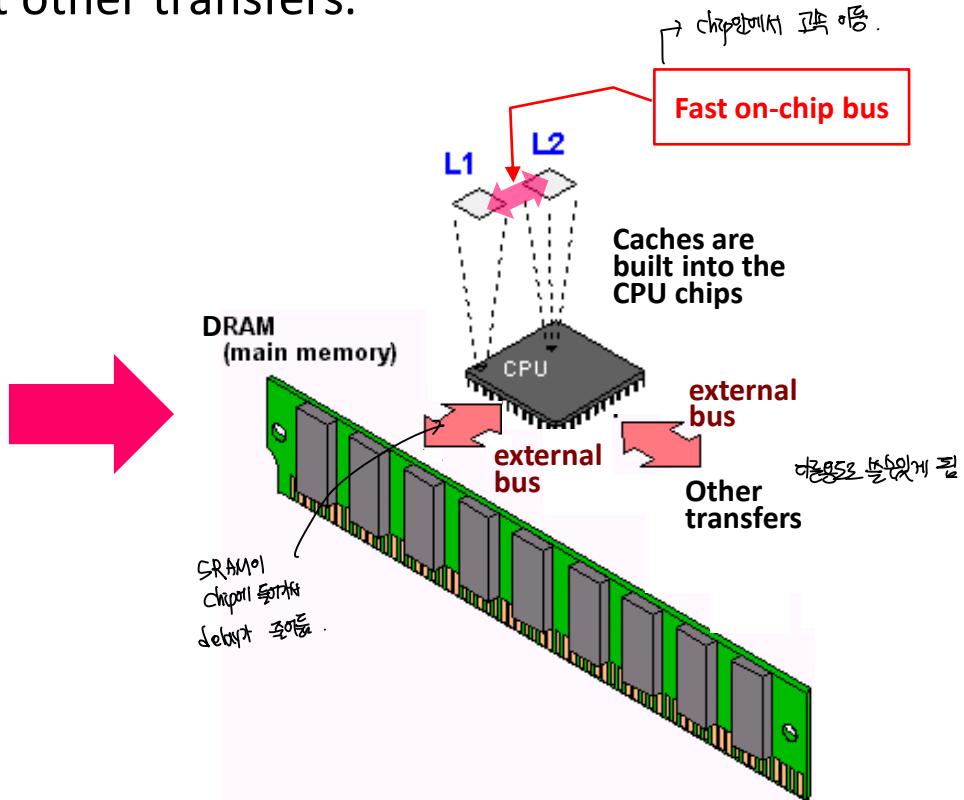
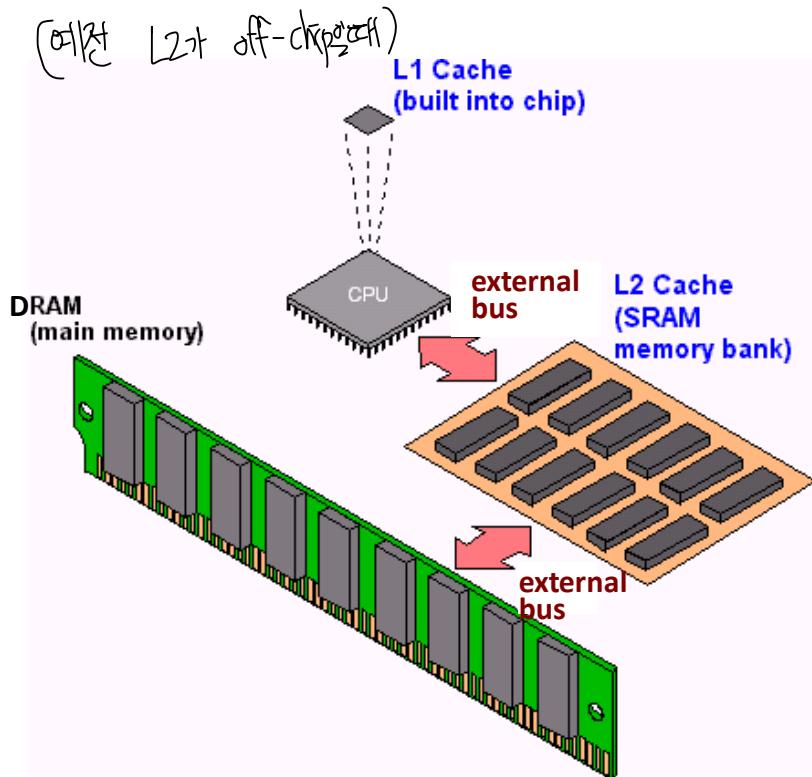


Issues of Cache Memory

→ L1, L2, L3 가 모두
1개의 chip(칩)에 있는 상황
off-chip(외부)
(로마자 표기)

- Multi-Level Cache

- High logic density enables multi-level caches on a chip.
 - On-chip cache reduces the CPU's external bus delay, speeds up execution times and increases system performance.
 - External bus is free to support other transfers.



Issues of Cache Memory

- Multi-Level Cache

여기

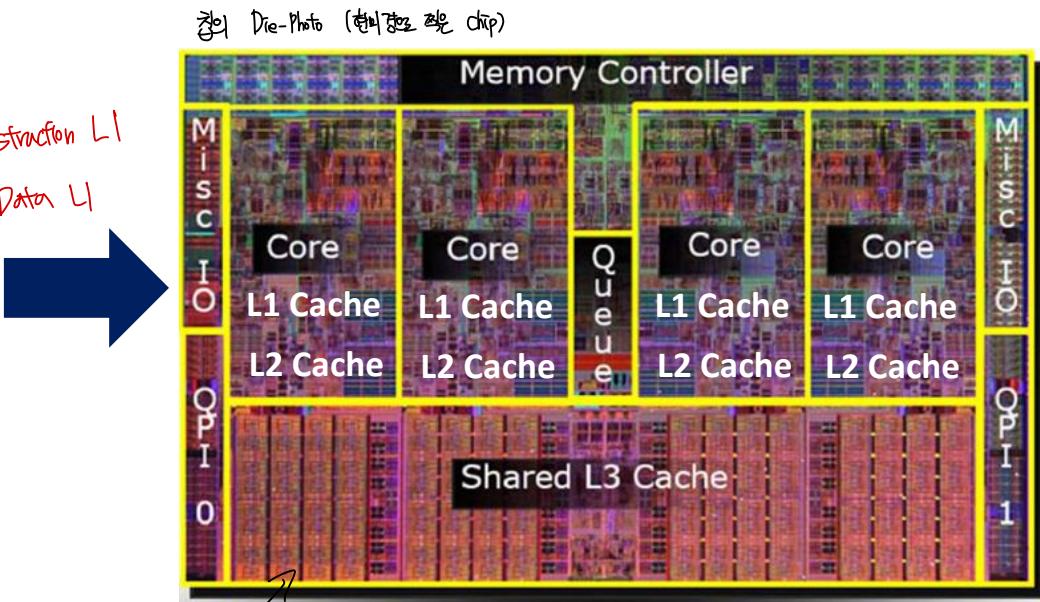
- On-chip cache examples

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators									
				Fab. Tech	Power	Name	CPU						
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache
General PC	Desktop	SAMSUNG DB400T6B	2015. 03	14 nm	65W	Intel® Core™ i7-6700	6th Gen. Skylake	64	3.4 ~ 4 GHz	x86 -64	4	Per-core: I-32KB, D-32KB 256KB	Per-core: 4 cores and G PU 8MB shared by 4 cores and G PU

→ CPU



Intel® Core™ i7-6700
Multi-Level Cache Architecture



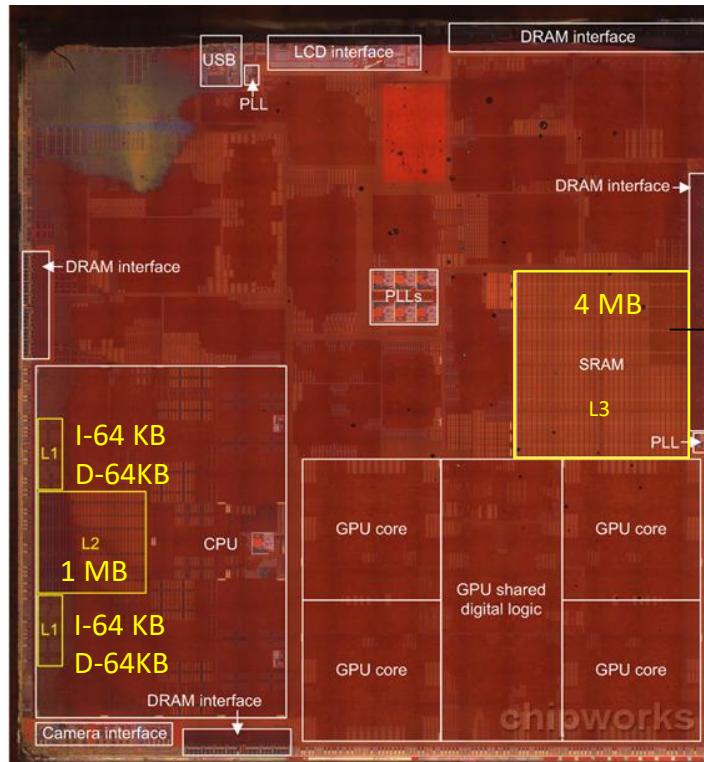
Intel® Core™ i7-6700 Chip
Die-Photo
한국어로 된 사진, cache memory

Issues of Cache Memory

- Multi-Level Cache

- On-chip cache examples

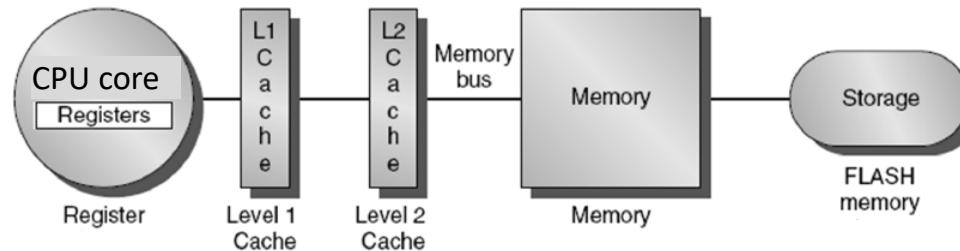
Type	Sub Type	Product Name	Year. Month	System-on-Chip											
				Tech	Power	Name	CPU					No. of Cores	L1 Cache	L2 Cache	L3 Cache
							Micro-architecture	Bit-Width	Clock Freq.	ISA					
		iPhone 5s	2013. 09	28 nm	2~3 W	Apple A7	Apple Cyclone	64	1.3 GHz	ARM v8-A	2	Per-core: I-64KB, D-64KB	1MB shared by 2 cores	4MB shared by the entire System-on-Chip	



Issues of Cache Memory

- Multi-Level Cache

- Multi-level cache organization



Size: 500 bytes
Speed: 500 ps

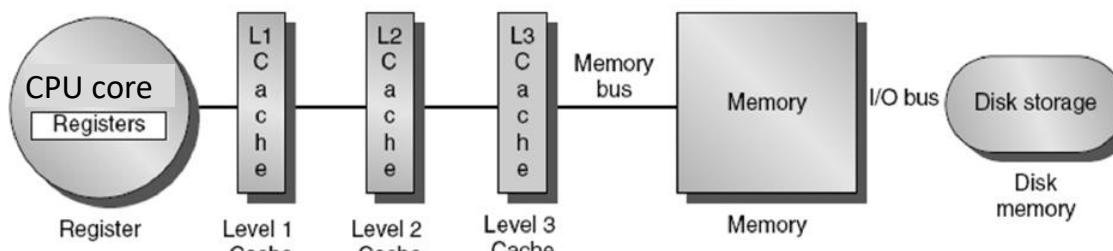
32 - 128 KB 256 - 512 KB
2 ns **(Single core)**

1 – 8 MB
(shared by
multi-core),
10–20 ns

256 MB - 3 GB
50–100 ns

4 – 256 GB
25–50 us

(a) Memory hierarchy for embedded CPU



Size: 1000 bytes
Speed: 300 ps

32 - 48 KB **1024KB**

2 - 22 MB
10-20 ns

4–16 GB
50–100 ns

4–16 TB
5–10 ms

(b) Memory hierarchy for general PC/server/workstation CPU

Issues of Cache Memory

- Multi-Level Cache

- Multi-level cache organization with commercial CPU products

diff[erent] multi-cache embedded

Type	Sub Type	Product Name	Release date (Year, Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators								
				Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache
Group#1	Embedded Computer Systems	Smart Phone	iPhone 4	2010. 06	45 nm							
			iPhone 4s	2011. 03	32 nm							
			iPhone 5s	2013. 09	28 nm							
			iPhone 6s	2015. 09	16 nm							
			iPhone 7	2016. 09	16 nm							
			Galaxy S8	2017. 04	10 nm							
			iPhone X	2017. 11	10 nm							
			iPhone XS	2018. 09	7 nm							
			iPhone 11 Pro	2019. 09	7 nm							
	Tablet PC	HP x2 210 G2 3ML39PA	2018. 02	14 nm								

The diagram illustrates the memory hierarchy for an embedded CPU. It shows a CPU core containing Registers connected to a Level 1 Cache (L1 Cache). This is followed by a Memory bus connecting to a larger Level 2 Cache (L2 Cache), then to a central Memory block, and finally to a Storage unit (FLASH memory).

(a) Memory hierarchy for embedded CPU

Level	Size	Speed
Register	500 bytes	
L1 Cache	32 - 128 KB 2 ns	
L2 Cache	256 - 512 KB (Single core) 1 - 8 MB (shared by multi-core) 10-20 ns	
Memory	256 MB - 3 GB 50-100 ns	
Storage	4 - 256 GB 25-50 us	

Issues of Cache Memory

- Multi-Level Cache

- Multi-level cache organization with commercial CPU products

설명의 시도. (이걸 어떤지 풍자하고 단정지는게 좋을 듯) ↗ 예를 : L2와 같은 캐시를 공유하는 경우 (Shared)
설명 : per-core L2는 실행.

Type	Sub Type	Product Name	Release date (Year, Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators											
				Fab-Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq	ISA	No. of Cores	L1 Cache			
Group#2 ?	Embedded Computer Systems	Smart Phone	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	64	2.9 GHz	ARM v8-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB		
			Galaxy Note10	2019. 08	7 nm	?	Exynos 9825		1. 9 GHz	ARM v8.2-A		Per-core: I-32KB, D-32KB Shared	256KB		
General PC	Laptop	SAMSUNG NT900X3N-K59SS	2015. 03	14 nm	CPU										
		LG GRAM 17ZD90N-VX70K	2019. 12	10 nm	2.73 GHz										
	Desktop	SAMSUNG DB400T6B	2015. 12	14 nm	2.4 GHz										
		Danawa 190721	2019. 07	14 nm	1.95 GHz										
	Floating License/ Data Sever	TYAN KFT46	2011. 09	32 nm	4MB shared										
Server/ Workstation	High performance Workstation	Lenovo ThinkStation P500TW 850W	2015. 07	22 nm	non-embedded 7nm										
		TYAN TAKO-KQT44	2020. 04	14 nm	Disk storage										

The diagram illustrates the memory hierarchy for general-purpose CPUs. It shows a central CPU core with registers connected to a Level 1 Cache (L1 C a c h e). This is followed by a Level 2 Cache (L2 C a c h e) and a Level 3 Cache (L3 C a c h e), both connected via a Memory bus to a main Memory block. The Memory block is also connected to an I/O bus, which leads to Disk storage and Disk memory. The diagram includes handwritten notes: 'Register' under the CPU core, 'Level 1 Cache' under the first cache, 'Level 2 Cache 256 -' under the second cache, and 'Level 3 Cache' under the third cache. Below the diagram, specific values are listed for each component:

Size:	1000 bytes	32 - 48 KB	1024KB	2 - 22 MB	4-16 GB	4-16 TB
Speed:	300 ps	1 ns	3-10 ns	10-20 ns	50-100 ns	5-10 ms

(b) Memory hierarchy for general PC/server/workstation CPU

Issues of Cache Memory – L1 Cache Size

- Questions about commercial CPU products

- Question#2-1:** Why do most of non-embedded CPUs only have 32 KB L1 caches?
- Answer#2-1:** if L1 cache size exceeds 32KB under private L2 cache organization, its performance is hardly improved any more whereas its area, power and energy consumption increase a lot.

• 메인 메모리 Maximization

private = per-core.

32KB를 초과하면 성능향상이 없음.
면적, 전력소모가 높아지며 성능향상이 어렵다.

(private L2 cache 확장)
→ 가중치에 따라

이전 대형화된 쪽은
Micro-architecture
선택한 편

SM I-32kBuf!

마지막
선택한 편

Type	Sub Type	Product Name	Release date (Year, Month)	Fab. Tech	Power	Name	SoC (System-on-Chip) = CPU + GPU + HW Accelerators								
							CPU				ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache
Group#2	General PC	Laptop	SAMSUNG NT900X3N-K59S	2015. 03	14 nm	15W	Intel® Core™ i5-7200U	7th Gen. Kaby Lake	64	2.5 ~ 3.1 GHz	x86-64	2	Per-core: I-32KB, D-32KB	Per-core: 256KB	3MB shared by 2 cores and GPU
			LG GRAM 17ZD90N-VX70K	2019. 12	10 nm	15W	Intel® Core™ i7-1065G7	10th Gen. Ice Lake	64	1.3 ~ 3.9 GHz	x86-64	4	Per-core: I-32KB, D-48KB	Per-core: 512KB	8MB shared by 4 cores and GPU
	Desktop	Desktop	SAMSUNG DB400T6B	2015. 12	14 nm	65W	Intel® Core™ i7-6700	6th Gen. Skylake	64	3.4 ~ 4 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores and GPU
			Danawa 190721	2019. 07	14 nm	95W	Intel® Core™ i7-9700K	9th Gen. Coffee Lake	64	3.6 ~ 4.9 GHz	x86-64	8	Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores and GPU
Server/ Workstation	Floating License/ Data Sever	TYAN KFT46	Lenovo ThinkStation P500TW 850W	2011. 09	32 nm	80W	Intel® Xeon ® E5606	Westmere	64	2.13 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores
			TYAN TAKO-KQT44	2020. 04	14 nm	150W	Intel® Xeon Gold 6226R	WikiChip							
			Haswell	2015. 07	22 nm	140W	Intel® Xeon ® E5-1630 v3	Block diagram							
	High performance Workstation	Cascade Lake	2020. 04	14 nm	150W	Intel® Xeon Gold 6226R			64	3.7 ~ 3.8 GHz	x86-64	4	Per-core: I-32KB, D-32KB	Per-core: 256KB	10MB shared by 4 cores
									64	2.9 ~ 3.9 GHz	x86-64	16	Per-core: I-32KB, D-32KB	Per-core: 1MB	22MB shared by 16 cores

Issues of Cache Memory – L1 Cache Size

- Generally, increasing L1 cache size ↗ 일반적으로 L1 캐시가 커지면 히트률도 커짐.

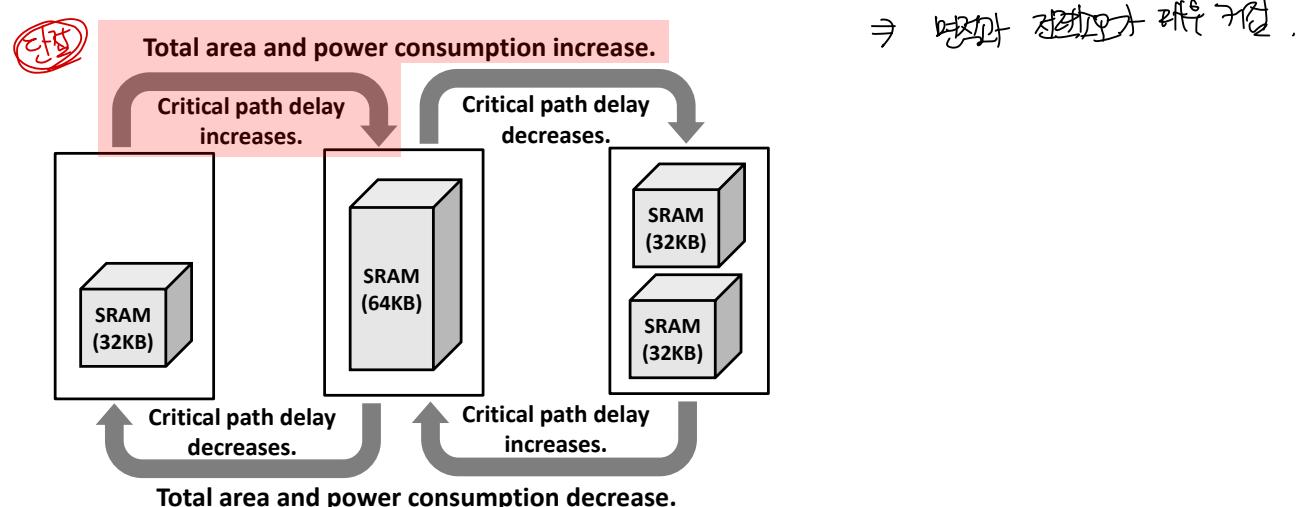
- improves cache performance by enhancing hit-rate. (case#1).

↗ 예전, L2 메모리가 per-core로 갖는 거정도면, L1이 32KB를 넘어야 한다.

- However, if L1 cache size exceeds 32 KB under private L2 cache organization

- its performance is hardly improved any more. ↗ 성능상의 개선이 없어졌다.

- whereas its area and power consumption increase a lot.



Single-core 쪽에의 편집.

Issues of Cache Memory – L1 Cache Size

- A survey paper by 'Mutaz Al-Tarawneh' selected for showing comparison between L1 instruction cache memory size and its performance/power/energy under multi-level cache organization ⇒ L1 cache가 32KB를 넘어서면 범위를 벗어나기 때문에 속도에 영향을 끼친다.
(L1 cache의 크기)

→ Mutaz Al-Tarawneh, "An Investigation of the Impact of Instruction Cache (I-Cache) Organization on Power-Performance Trade-offs in the Design of Scalar Processors," *European Journal of Scientific Research*, vol. 115, no. 1, pp. 7-26, November 2013
(제작 안내한 저널)
↳ But, 실제 퍼포먼스가 종래 기대한 것과는 달리.

(9장 내용간략화하기) ↳ 순서대로 매 cycle 멀티프로세서는 수행하는 processor를 선택한다.

- Scalar processors mean simple execution cores that can execute a single instruction per clock cycle in the order.
- Multi-level cache organization: 256 KB L2 cache memory
- An execution-driven power simulation tool has been used for obtaining performance/power/energy cost. ↳ 이 결과를 사용하여

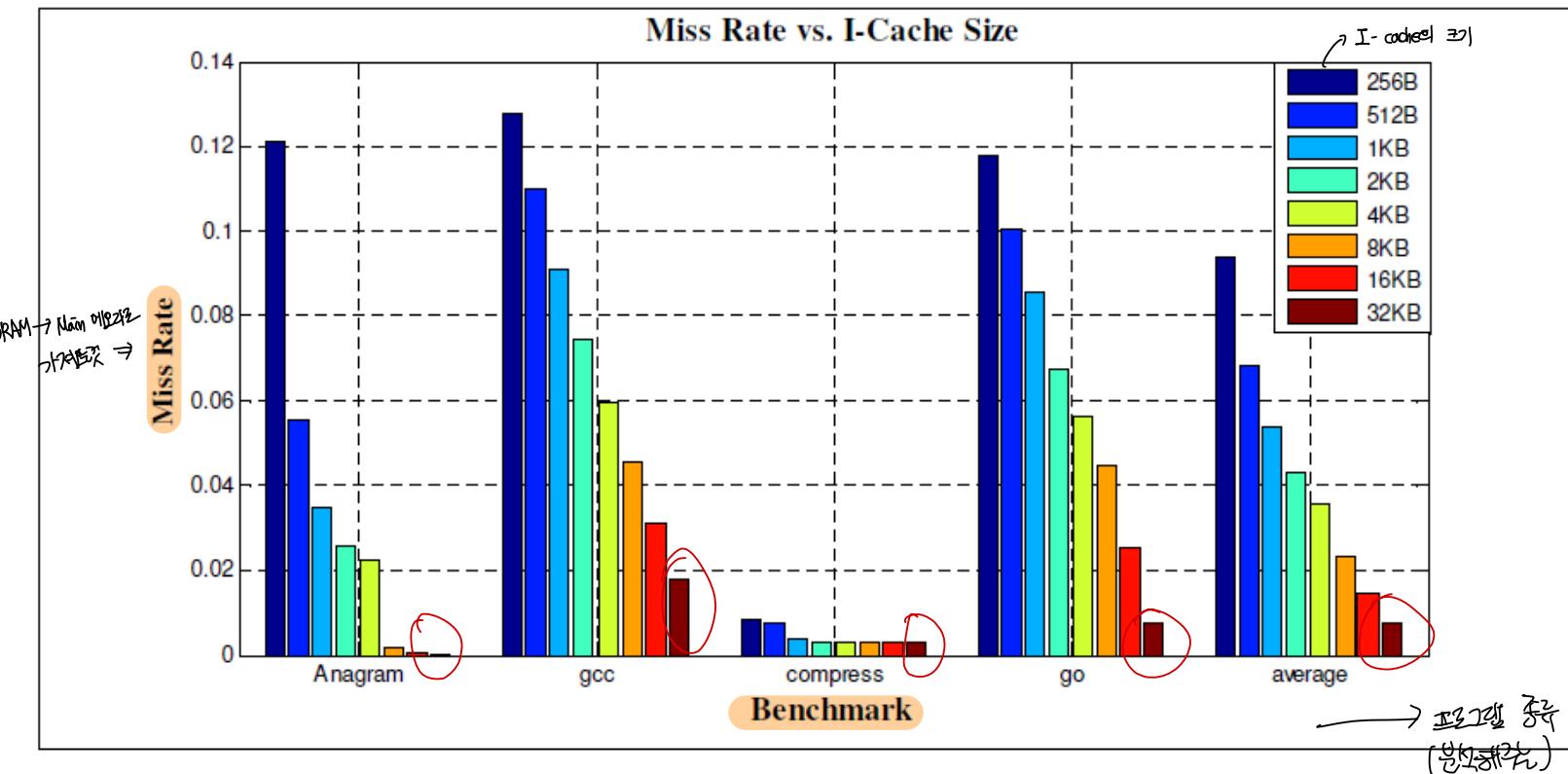
기능적 성능 / 전력 / 에너지 효율을 조사함.

Issues of Cache Memory – L1 Cache Size

- Survey paper by 'Mutaz Al-Tarawneh'
 - Size versus performance

고려해야 하는
I-cache의 크기
32KB까지 I-cache를 사용하면
Miss Rate가 증가함.
(장치별로 (L1 + 256KB)가 있는 전자장비)
→ I-cache 크기

Figure 1: The Impact of I-Cache Size on its Miss Rate.

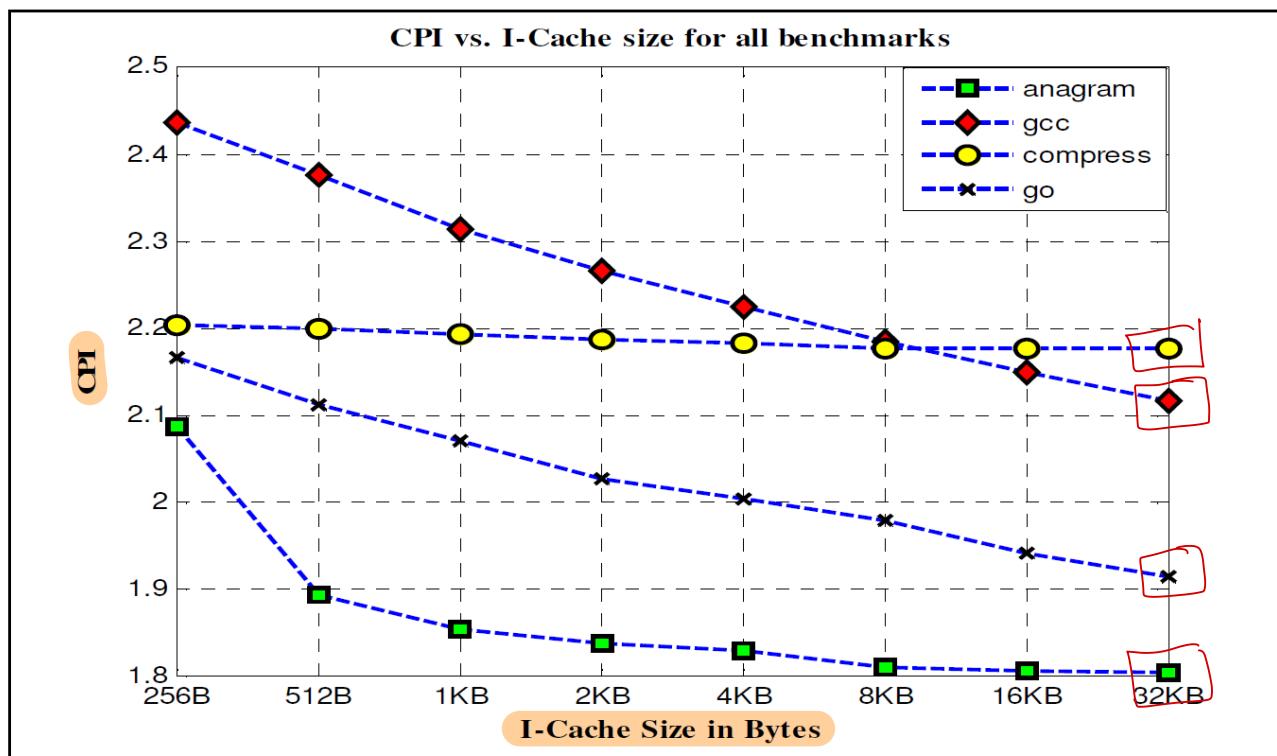


Issues of Cache Memory – L1 Cache Size

- Survey paper by ‘Mutaz Al-Tarawneh’

- Size versus performance
 - 명령어 (캐시 사용 여부에 따라) 주연 주연 성능이 좋다.
 - 미쓰리지 높아면 CPI도 높아짐.

Figure 2: CPI of different benchmarks as I-Cache size increases.
(Cycle per Instruction)

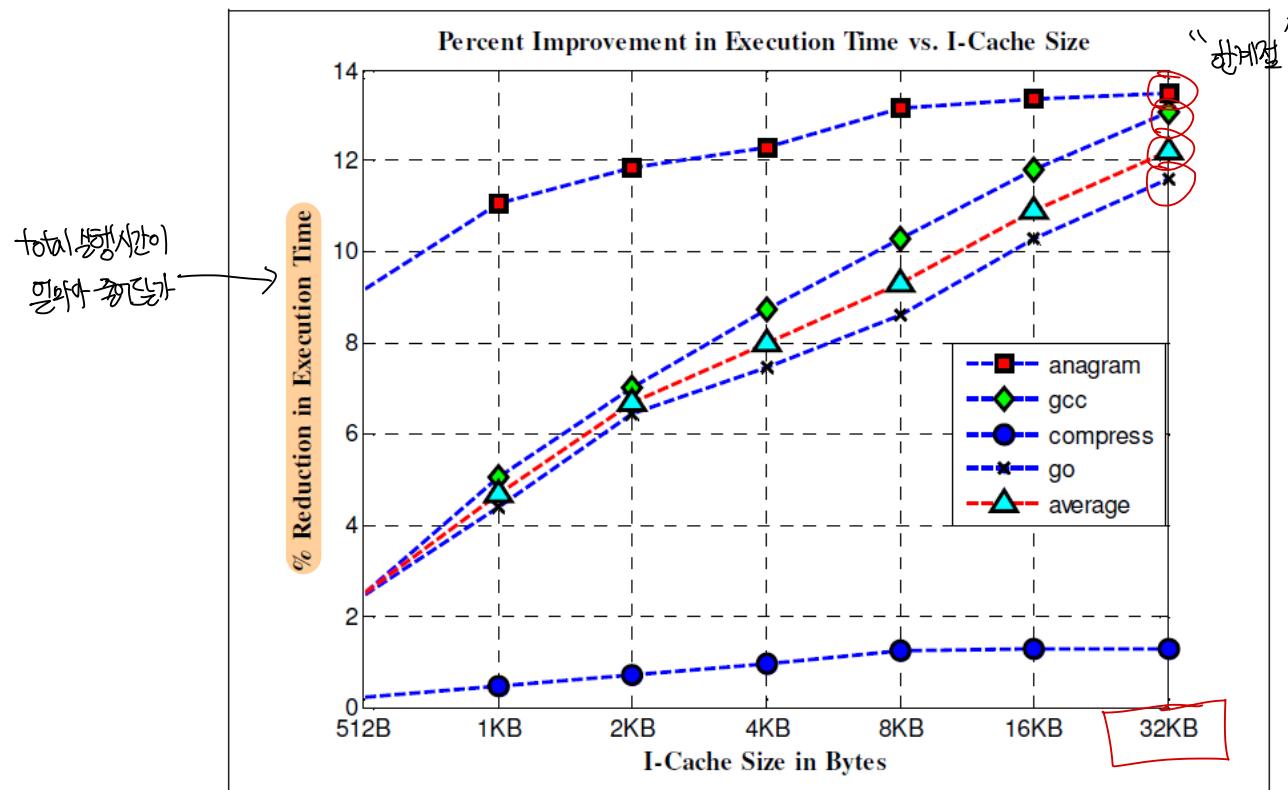


평가
⇒ CPI가 계속 줄어듬.
But 32KB 와다
기억해 더 큰쪽으로
증가하지 않음.
(증가가 멈음)

Issues of Cache Memory – L1 Cache Size

- Survey paper by ‘Mutaz Al-Tarawneh’
 - Size versus performance

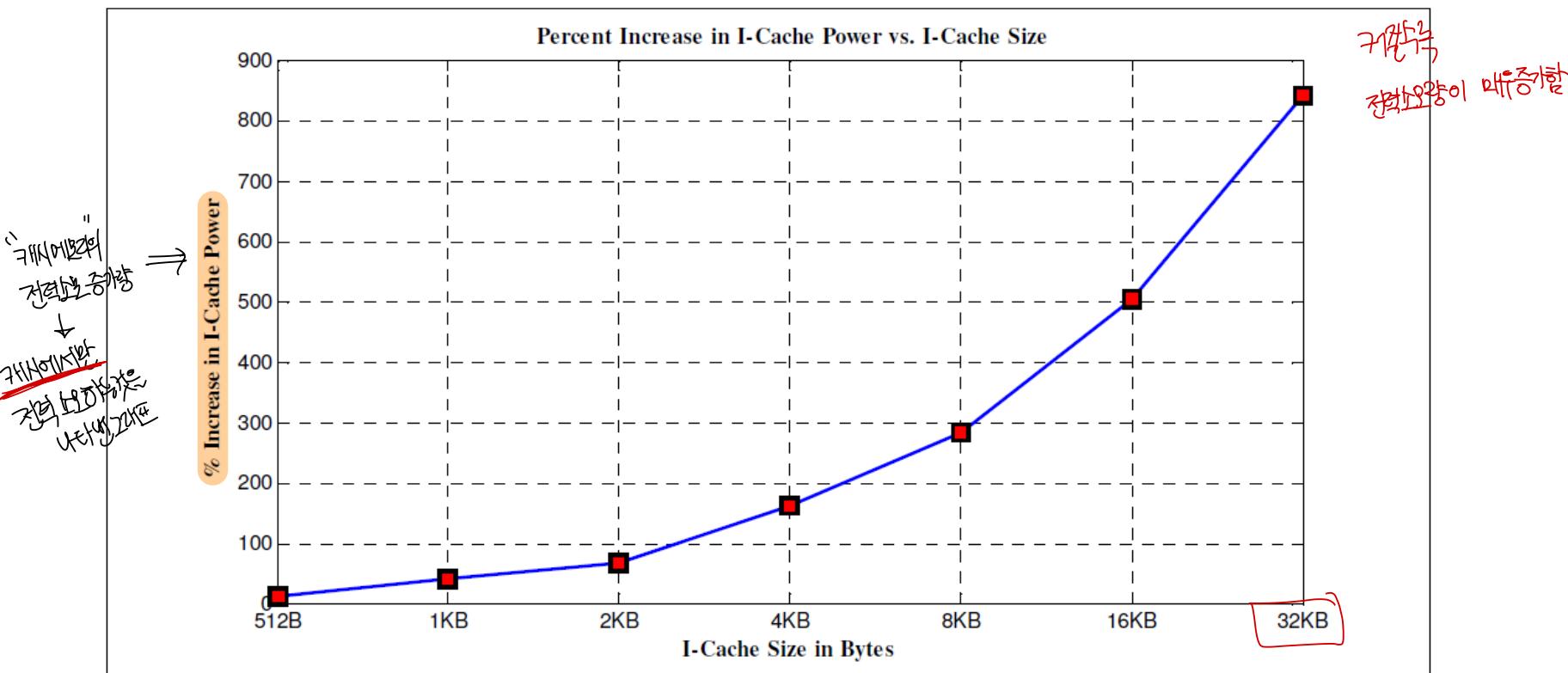
Figure 3: Percent improvement in execution time as cache size increases.



Issues of Cache Memory – L1 Cache Size

- Survey paper by ‘*Mutaz Al-Tarawneh*’
 - Size versus power

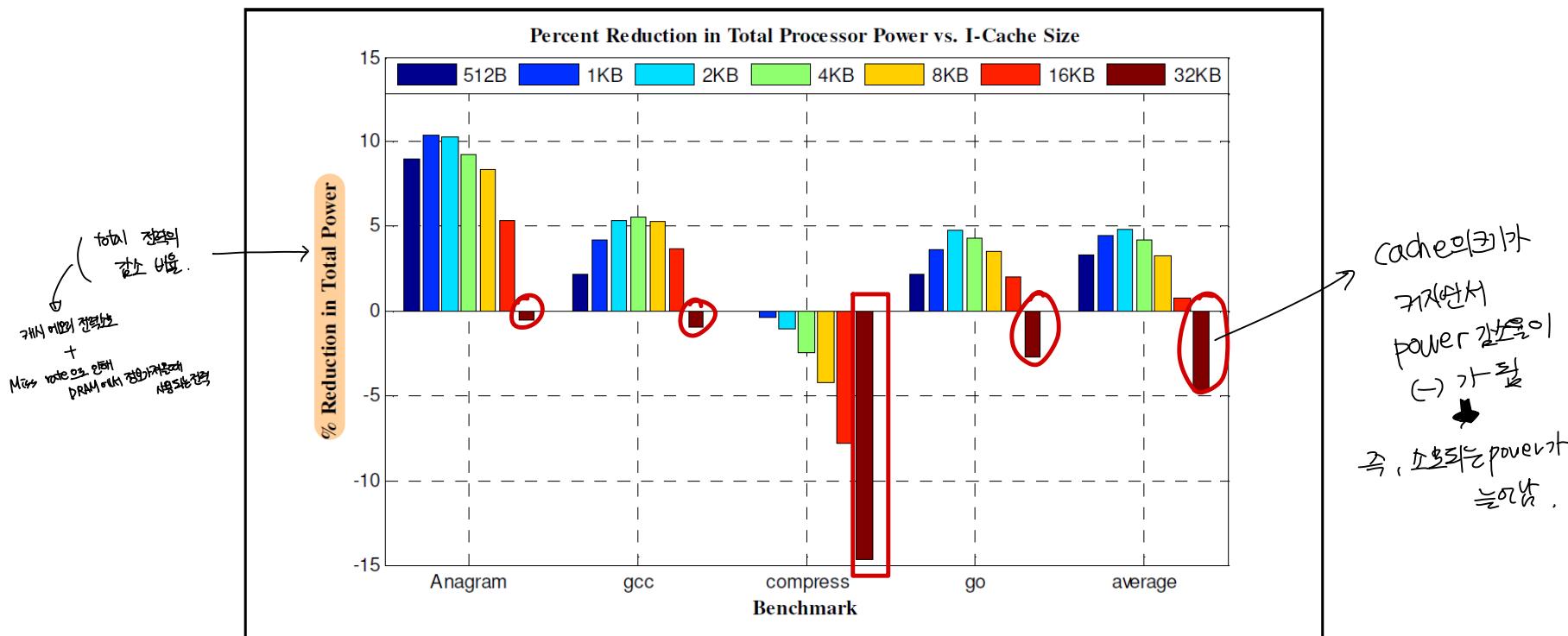
Figure 4: Percent increase in I-Cache power as cache size increases.



Issues of Cache Memory – L1 Cache Size

- Survey paper by ‘Mutaz Al-Tarawneh’
 - Size versus power

Figure 5: Percent reduction in total processor power as I-Cache size increases.



Issues of Cache Memory – L1 Cache Size

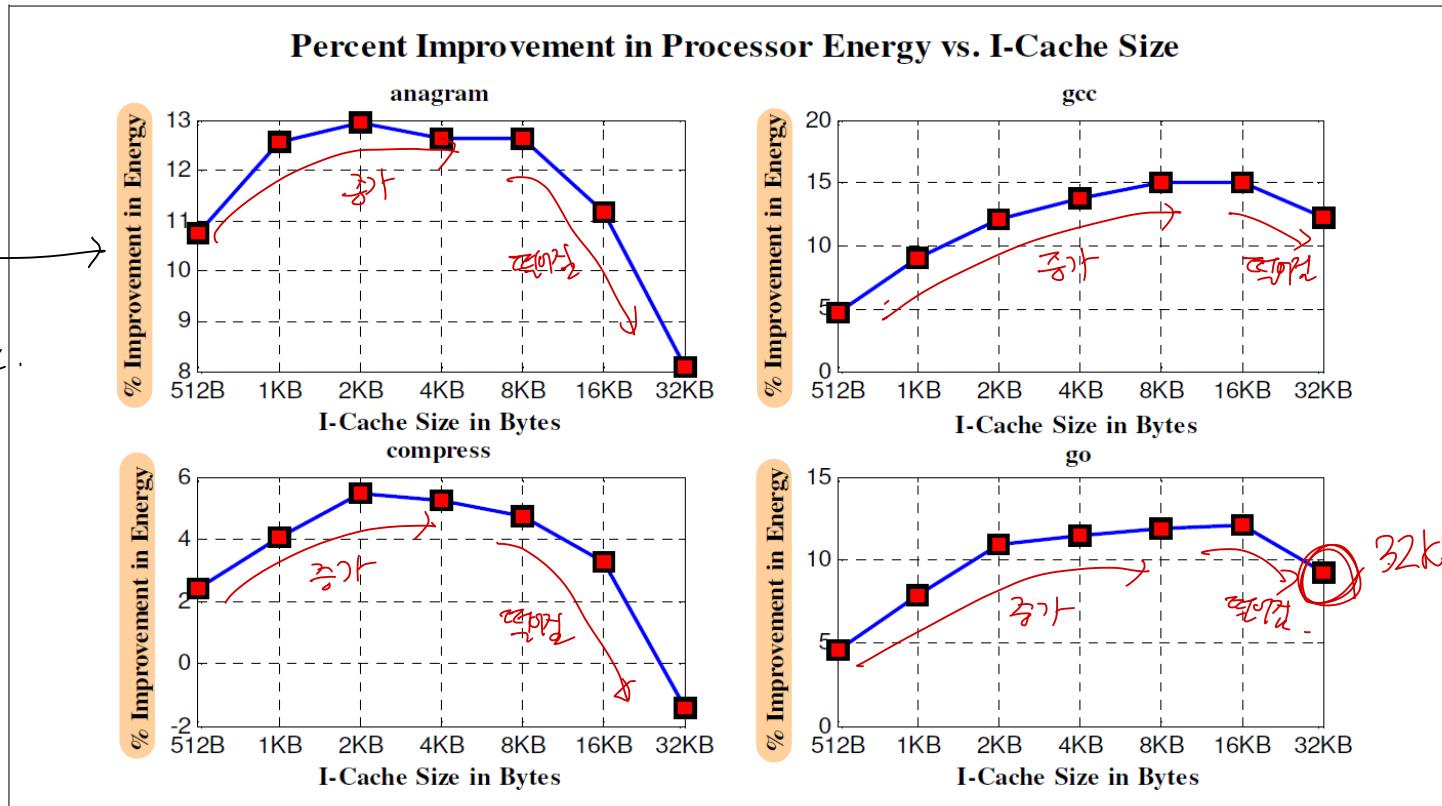
- Survey paper by 'Mutaz Al-Tarawneh'
 - Size versus energy

panov는 시간과 소모하는 에너지

cache에 예외가 가지면서
Miss rate이 감소하여
소모되는 에너지 감소

↳ 결과: 에너지 감소의 폭넓음

Figure 6: Percent reduction in processor energy as cache size increases.

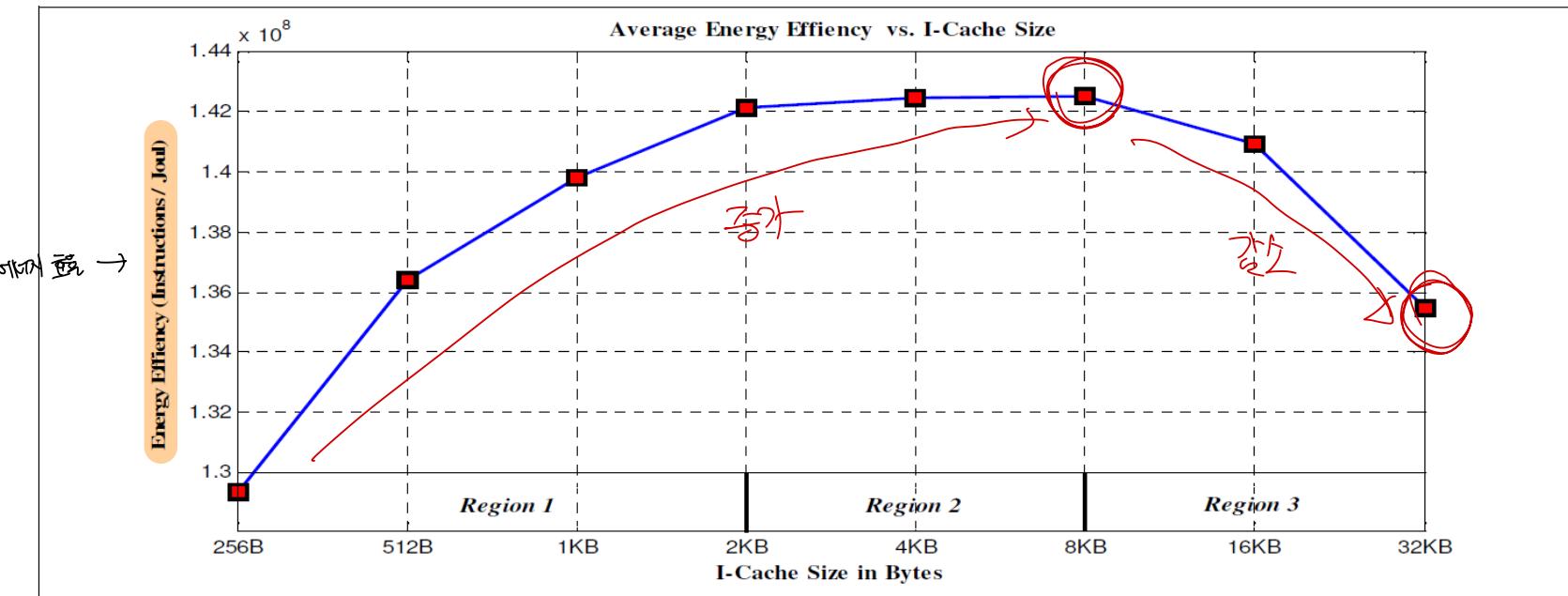


Issues of Cache Memory – L1 Cache Size

- Survey paper by 'Mutaz Al-Tarawneh'
 - Size versus energy

이전 특성 대비
L1 cache memory는 32KB 까지면
특성이 좋았지만 7KB.
특성이 좋았지만 7KB.

Figure 7: Average energy efficiency as I-Cache size increases.



Issues of Cache Memory – L1 Cache Size

- Questions about commercial CPU products

- Question#2-2: Why do most of embedded CPUs have larger L1 caches (64KB, 128KB) than L1 caches (32 KB, 48KB) of non-embedded CPUs?

SoC (System-on-Chip) = CPU + GPU + HW Accelerators														
Type	Sub Type	Product Name	Release date (Year, Month)	Fab. Tech	Power	Name	CPU							
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache
Group#1	Embedded Computer Systems	Smart Phone	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance 4 x Cortex-A55 for energy efficiency	64 1. 9 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB Shared	Per-core: 512KB	4MB shared
			Galaxy	2019.	7	?		2 x Cheetah (M4) for high performance 2 x Cortex-A75 for	2.73 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	Per-core: 512KB Per-core: 1256KB Per-core: 128KB	4MB shared
		Smart Phone	iPhone 4	2010. 06	45 nm	?	Apple A4	ARM Cortex-A8	32 800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None
			iPhone 4s	2011. 03	32 nm	?	Apple A5	ARM Cortex-A9	32 1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None
			iPhone 5s	2013. 09	28 nm	2~3 W	Apple A7	Apple Cyclone	64 1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC
			iPhone 6s	2015. 09	16 nm	?	Apple A9	Apple Twister	64 1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC
			iPhone 7	2016. 09	16 nm	?	Apple A10 Fusion	2 x Hurricanes for high performance 2 x Zephyr for energy efficiency	64 2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	3MB shared 1MB shared	4MB shared by the entire SoC
Group#2	Embedded Computer Systems	Smart Phone	Galaxy S8	2017. 04	10 nm	?	Exynos 8895	4 x Mongoose2 (M2) for high performance 4 x Cortex-A53 for energy efficiency	64 2.3 GHz 1. 7 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-32KB Per-core: I-32KB, D-32KB	2MB shared 256KB shared	None
			iPhone X	2017. 11	10 nm	?	Apple A11 Bionic	2 x Monsoon for high performance 4 x Mistral for energy efficiency	64 2.39 GHz 1.42 GHz	ARM v8.2-A	6	Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	8MB shared 1MB shared	None
			iPhone XS	2018. 09	7 nm	?	Apple A12	2 x Vortex for high performance 4 x Tempest for energy efficiency	64 2.49 GHz 1.52 GHz	ARM v8.3-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-32KB	8MB shared 2MB shared	None
			iPhone 11 Pro	2019. 09	7 nm	?	Apple A13	2 x Lightning for high performance 4 x Thunder for energy efficiency	64 2.66 GHz 1.82 GHz	ARM v8.4-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-48KB	8MB shared 4MB shared	None
			Tablet PC	HP x2 210 G2	2018. 02	14 nm	2W	Intel Atom® Processor x5-Z8350	Airmont	64 1.44 ~ 1.92 GHz	x86-64	4	Per-core: I-32KB, D-24KB 2MB (1MB shared by 2 cores)	22MB shared by 16 cores

Issues of Cache Memory – L1 Cache Size

- Questions about commercial CPU products

- Question#2-2: Why do most of embedded CPUs have larger L1 caches (64KB, 128KB) than L1 caches (32 KB, 48KB) of non-embedded CPUs?

- Answer#2-2:

- Most of embedded CPUs do not have private L2 caches.
→ embedded CPU는 private (=per-core) L2 캐시를 가지고 있지 않아서
 - Therefore, the below principle is not true of the embedded CPUs.
 - ✓ Answer#2-1: if L1 cache size exceeds 32KB under private L2 cache organization, its performance is hardly improved any more.
→ private L2 캐시보다 성능을 올리기 어렵다. L1 캐시를 쓰는 경우 (즉) ⇒ 이를 보통 report하는 경우 .
 - In order to maximize their performance without the private L2 caches
 - ✓ Most of embedded CPUs have large L1 caches (64 KB, 128 KB) compared with L1 caches (32KB, 48KB) of non-embedded CPUs .
 - However, only large L1 cache without private L2 cache is a second-best solution.
그러나, private L2는 있어 L1 캐시를 크게 하는 것은 best-solution은 아님 .

Issues of Cache Memory – L1 Cache Size

- Questions about commercial CPU products

- Question#2-2:** Why do most of embedded CPUs have larger L1 caches (64KB, 128KB) than L1 caches (32 KB, 48KB) of non-embedded CPUs?
- Answer#2-2:** In order to maximize their performance without the private L2 caches
 ↗ 주거지 embedded CPU가 이성호. ⇒ per-core L2를 가지고 있을까도 끝까지 L1 cache는 I-64KB다.
- However, the following two embedded CPUs are strange.

Group#2

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	CPU							
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache
Embedded Computer Systems	Smart Phone	Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x Meerkat (M3) for high performance	64	2.9 GHz	ARM v8-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
							4 x Cortex-A55 for energy efficiency		1. 9 GHz	ARM v8.2-A		Per-core: I-32KB, D-32KB	256KB Shared	
		Galaxy Note10	2019. 08	7 nm	?	Exynos 9825	2 x Cheetah (M4) for high performance	64	2.73 GHz	ARM v8.2-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
							2 x Cortex-A75 for moderate performance		2.4 GHz			Per-core: I-64KB, D-64KB	Per-core: 256KB	
							4 x Cortex-A55 for energy efficiency		1.95 GHz			Per-core: I-32KB, D-32KB	Per-core: 128KB	

정도 있고 I-64KB를 한 것임

설명 정리 설명 정리.
(여기 설명의 정리)

Issues of Cache Memory

- Separate Cache versus Unified Cache

- Questions about commercial CPU products

- Question#3: Why are L1 caches separate and L2/L3 caches unified?

- Answer#3

- L1 cache : merit of separate cache > merit of unified cache
- L2/L3 cache : merit of separate cache < merit of unified cache

Group#2										SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
Type	Sub Type	Product Name	Release date (Year, Month)	Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	CPU		L1 Cache		L2 Cache		L3 Cache		
Group#1	Embedded Computer Systems	iPhone 4	2010. 06	45 nm	?	Apple A4	ARM Cortex-A8	32	800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None	Per-core: I-64KB, D-64KB		Per-core: 512KB		4MB shared	
		iPhone 4s	2011. 03	32 nm	?	Apple A5	ARM Cortex-A9	32	1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None	Per-core: I-32KB, D-32KB		256KB		Shared	
		iPhone 5s	2013. 09	28 nm	2~3 W	Apple A7	Apple Cyclone	64	1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC	Per-core: I-64KB, D-64KB		Per-core: 512KB		4MB shared	
		iPhone 6s	2015. 09	16 nm	?	Apple A9	Apple Twister	64	1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC	Per-core: I-64KB, D-64KB		Per-core: 256KB		4MB shared	
		iPhone 7	2016. 09	16 nm	?	Apple A10 Fusion		64	2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC	Per-core: I-32KB, D-32KB		Per-core: 128KB		4MB shared	
		Galaxy S8	2017. 04	10 nm	?	Exynos 8895						Per-core: I-64KB, D-32KB	1MB shared		Per-core: I-32KB, D-32KB		Per-core: 256KB		3MB shared by 2 cores and GPU	
		iPhone X	2017. 11	10 nm	?	Apple A11 Bionic		64	2.3 GHz 1.7 GHz	ARM v8-A	8	Per-core: I-64KB, D-32KB	2MB shared	None	Per-core: I-32KB, D-32KB		Per-core: 512KB		8MB shared by 4 cores and GPU	
		iPhone XS	2018. 09	7 nm	?	Apple A12						Per-core: I-64KB, D-64KB Per-core: I-32KB, D-32KB	8MB shared		Per-core: I-64KB, D-64KB		Per-core: 256KB		8MB shared by 4 cores	
		iPhone 11 Pro	2019. 09	7 nm	?	Apple A13		64	2.49 GHz 1.52 GHz	ARM v8.3-A	6	Per-core: I-128KB, D-128KB Per-core: I-32KB, D-32KB	1MB shared	None	Per-core: I-32KB, D-32KB		Per-core: 256KB		12MB shared by 4 cores and GPU	
		Tablet PC	HP x2 210 G2	14 nm	2W	Intel Atom® Processor x5-Z8350						Airmont	Per-core: I-32KB, D-24KB	8MB shared	Per-core: I-32KB, D-32KB		Per-core: 256KB		10MB shared by 4 cores	
												Per-core: I-32KB, D-24KB	2MB (1MB shared by 2 cores)	None	Per-core: I-32KB, D-32KB		Per-core: 1MB		22MB shared by 16 cores	

L1 cache 뒤에 separate 이고 I-32KB D-32KB로 작게 되어 있다.
L2, L3은 합쳐져 있다.

(separate cache)의 장점으로 인해서 선택됨.
↓
내장형 메모리로 편리.

제작자 차별화, 설계 단순화 등이 있다.

Issues of Cache Memory

- Separate Cache versus Unified Cache

- Initial cache memory designs used unified caches.

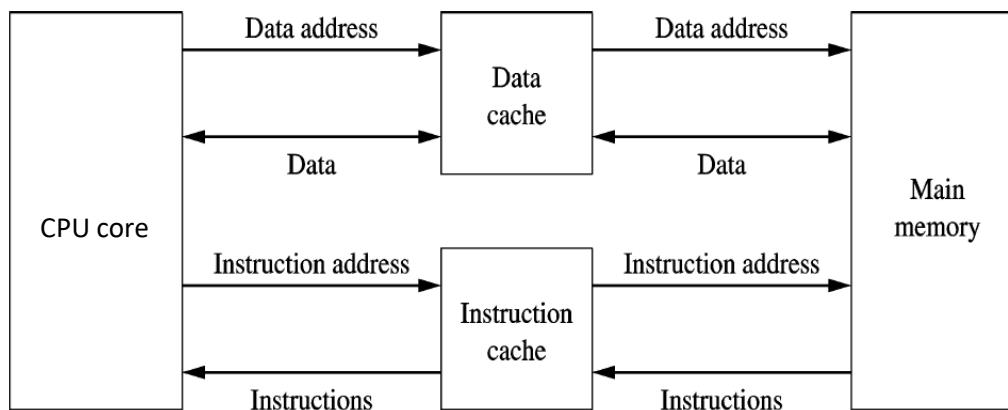
Processor	Type	Year of Introduction	L1 Cache
IBM 360/85	Mainframe	1968	16 to 32 KB
PDP-11/70	Minicomputer	1975	1 KB
VAX 11/780	Minicomputer	1978	16 KB
IBM 3033	Mainframe	1978	64 KB
IBM 3090	Mainframe	1985	128 to 256 KB
Intel 80486	PC	1989	8 KB
Pentium	PC	1993	I-8 KB, D-8 KB
PowerPC 601	PC	1993	32 KB
PowerPC 620	PC	1996	I-32 KB, D-32 KB
PowerPC G4	PC/server	1999	I-32 KB, D-32 KB

IBM

Unified cache memory

L1은 unified 입니다.

- Current trend is to use separate caches the nearest to CPU core.



→ CPU의 디자인
I, D가 분리된 형태

Issues of Cache Memory

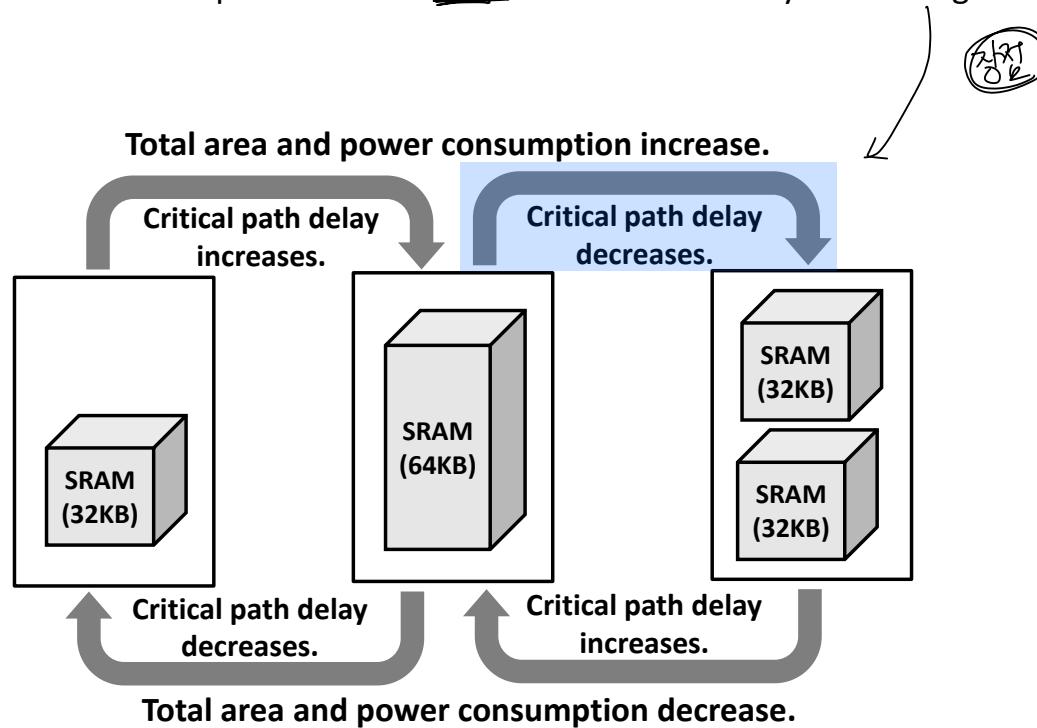
- Separate Cache versus Unified Cache

Separate



Advantage of separate cache memory.

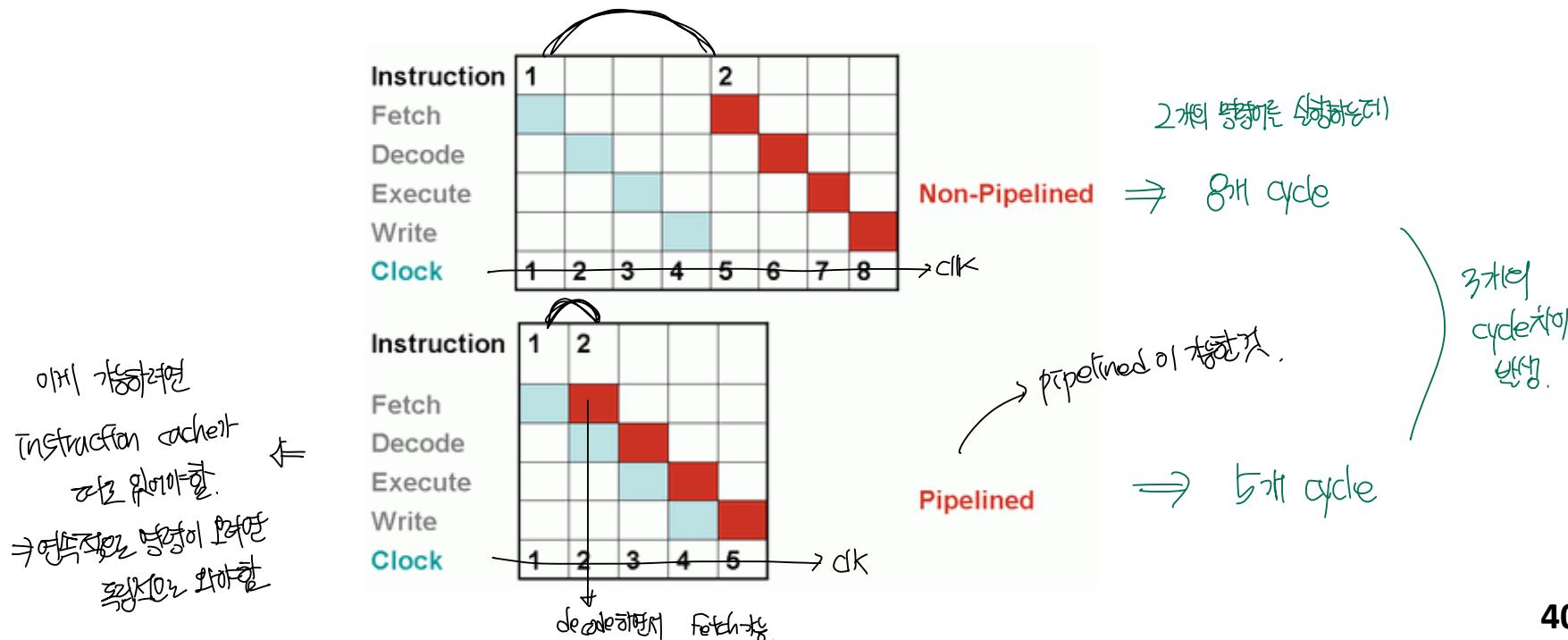
- It may offers faster access than unified cache. ⁽¹⁾ ⇒ unified cache ~~更快~~ 더 빠르다.
 - In the aspect of critical path delay
 - ✓ Two small separate caches may show shorter delay than a large unified cache.



Issues of Cache Memory

- Separate Cache versus Unified Cache

- **Advantage of separate cache memory.** ② 연속적인 instruction (명령어)를 fetch 하면 수 있다. ⇒ 그걸로 통해 speed up이 가능하다.
 - Consecutive instruction-fetch (pipelining) for speedup is possible.
 - pipelining is only available under independent instruction-cache memory without confliction with data-load/store operation. ↳ I-cache는 pipelining을 허용하지 않는다.
 - Of course, the previous four-instruction processor example does not support pipelining.
 - However, all commercial CPUs operate in the manner of pipelining in order to achieve performance improvement. ↳ 대부분의 CPU는 pipelining을 지원한다.

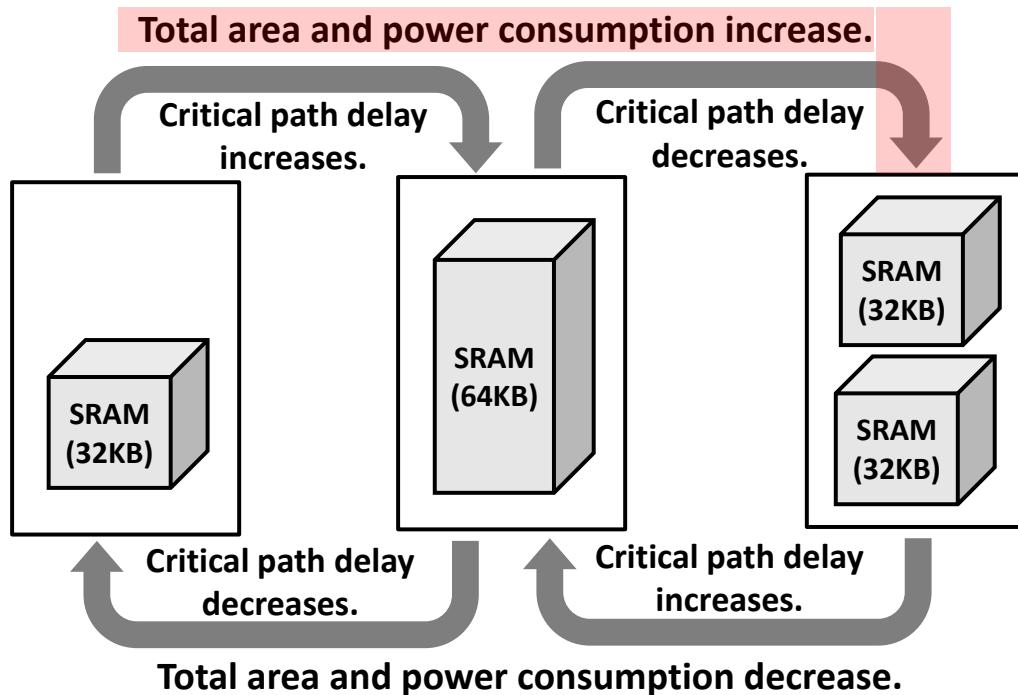


Issues of Cache Memory

- Separate Cache versus Unified Cache

• Disadvantage of separate cache memory.

- Total area and power consumption of the separate caches/is larger than the area and the power consumption of an unified cache when they have the same size.
① *separate cache*는 면적과 전력소모가 큽니다.



Issues of Cache Memory

- Separate Cache versus Unified Cache

- Disadvantage of separate cache memory.
 - Rigid boundaries between data and instruction caches
 - Dynamic load balancing between two caches is not possible.
 - Therefore, entire miss-rate is higher than unified cache memory that combines two caches. \Rightarrow 전체 miss-rate는 separate의 unified보다 높다.
(용량이 제한적이라서)
- (2) unified cache의 경우, I 와 D의 내용을 조정하기 어렵다.
하지만, separate의 경우, 정해진 I와 D 사이에 조정이 가능하다.
ex) (I + 48KB) 조정할 때, separate는 쉽게 가능하지만 miss-rate가 높다.

Miss rates for instruction, data, and unified caches of different sizes.

Size	Instruction cache	Data cache	Unified cache
1 KB	3.06%	+ 24.61%	$\Rightarrow 27 > 13.34\%$
2 KB	2.26%	+ 20.57%	$\Rightarrow 23 > 9.78\%$
4 KB	1.78%	+ 15.94%	$\Rightarrow 17 > 7.24\%$
8 KB	1.10%	+ 10.19%	$\Rightarrow 11 > 4.57\%$
16 KB	0.64%	+ 6.47%	$\Rightarrow 7 > 2.87\%$
32 KB	0.39%	+ 4.82%	$\Rightarrow 5 > 1.99\%$
64 KB	0.15%	+ 3.77%	$\Rightarrow 4 > 1.35\%$
128 KB	0.02%	+ 2.88%	$\Rightarrow 3 > 0.95\%$

어려워
다른
방법

20% Si. 암호화

Separate case

Miss-rate가 Separate의 경우는 .

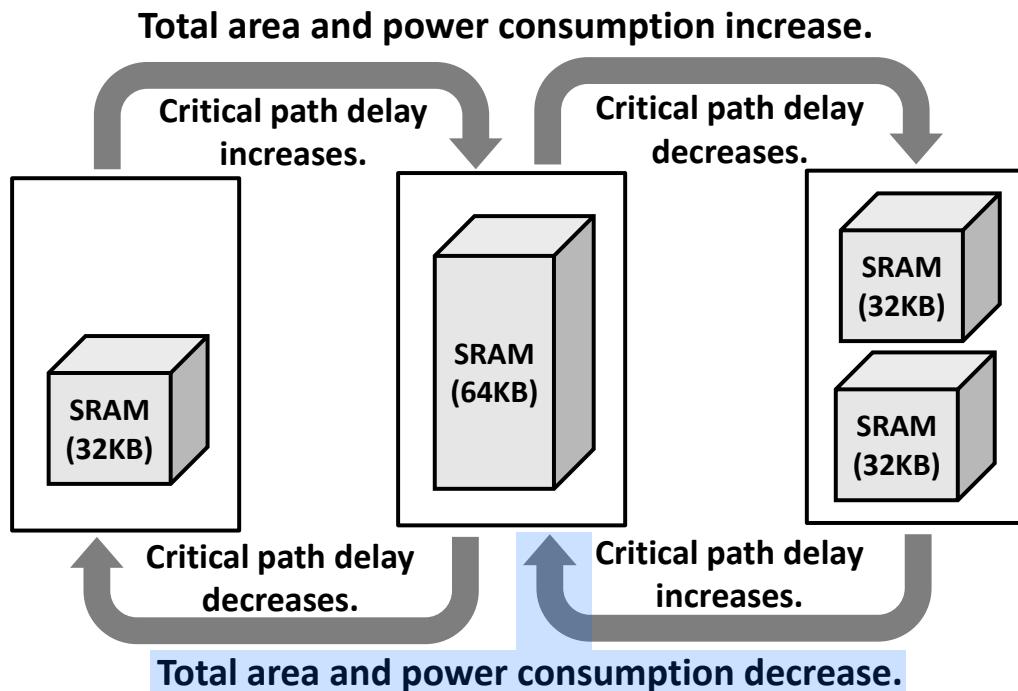
Issues of Cache Memory

- Separate Cache versus Unified Cache



Advantage of unified cache memory.

- The area and the power consumption of an unified cache is smaller than total area and power consumption of the separate caches when they have the same size.
① total 면적과 전력소비 감소한다.



Issues of Cache Memory

- Separate Cache versus Unified Cache

- **Advantage of unified cache memory.** ② Flexible boundaries between data and instruction caches
 - Dynamic load balancing between two caches is possible.
 - Therefore, if an execution pattern involves more instruction fetches than data-load/store operation, the unified cache memory will tend to fill up with instructions – vice versa.
↳ 이때는 주로 명령어에 separate cache의 miss rate가 높다.
 - Therefore, entire miss-rate is lower than separate cache memory.

Miss rates for instruction, data, and unified caches of different sizes.

Size	Instruction cache	Data cache	Unified cache
1 KB	3.06%	24.61%	13.34%
2 KB	2.26%	20.57%	9.78%
4 KB	1.78%	15.94%	7.24%
8 KB	1.10%	10.19%	4.57%
16 KB	0.64%	6.47%	2.87%
32 KB	0.39%	4.82%	1.99%
64 KB	0.15%	3.77%	1.35%
128 KB	0.02%	2.88%	0.95%

Issues of Cache Memory

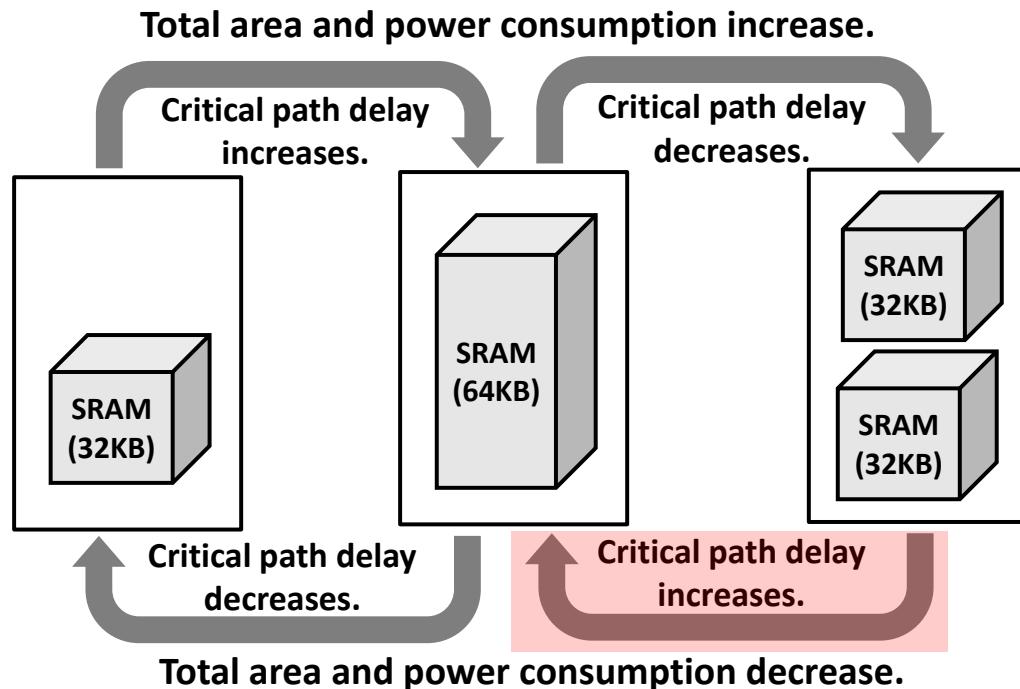
- Separate Cache versus Unified Cache



Disadvantage of unified cache memory.

① 장이 뒷다.

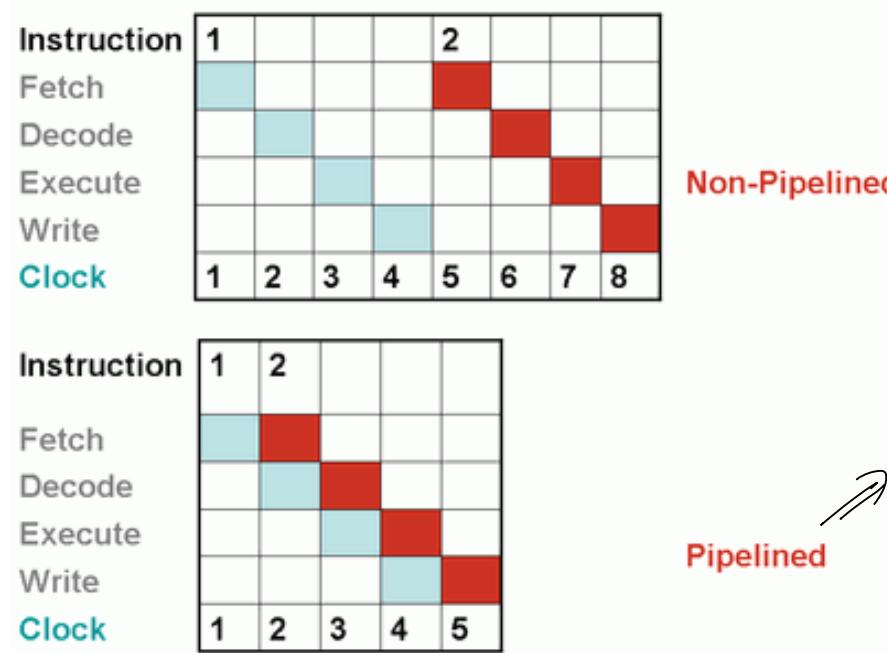
- It may offers slower access than separate cache.
 - In the aspect of critical path delay
 - ✓ Two small separate caches may show shorter delay than a large unified cache.



Issues of Cache Memory

- Separate Cache versus Unified Cache

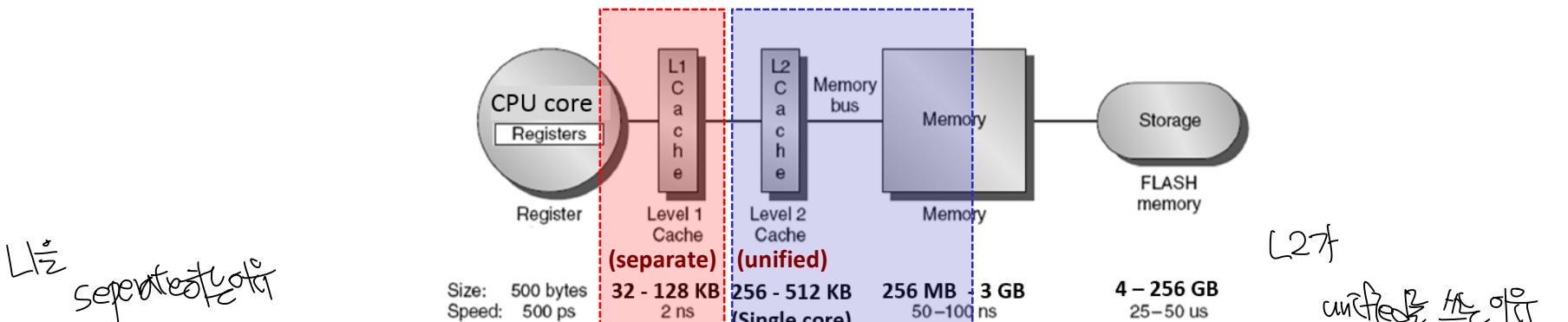
- Disadvantage of unified cache memory.
 - Consecutive instruction-fetch (pipelining) is not possible.
 - Therefore, it may causes performance degradation even though high-performance CPU supports multiple stage-pipelining.



Issues of Cache Memory

- Separate Cache versus Unified Cache

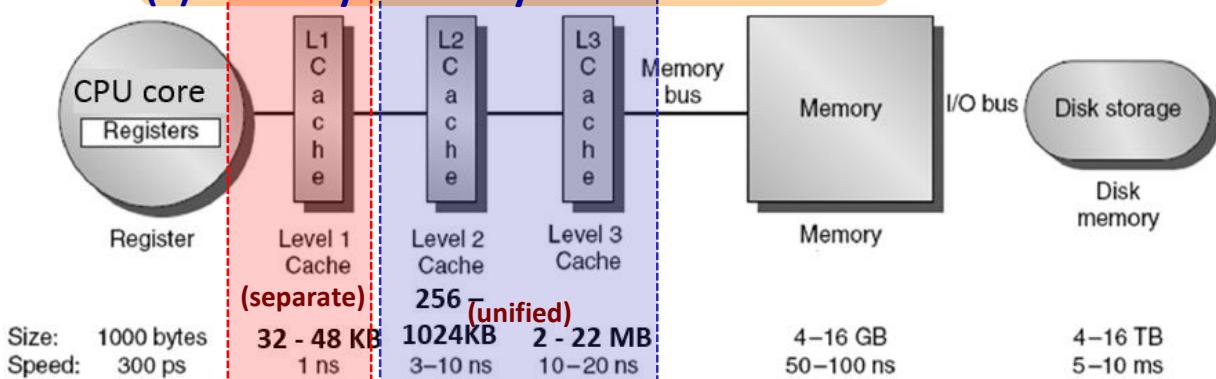
- Multi-level cache organization



Advantage of separate cache > Advantage of unified cache

Advantage of separate cache < Advantage of unified cache

(a) Memory hierarchy for embedded CPU



(b) Memory hierarchy for desktop/server/workstation CPU
(+ Latest SAMSUNG embedded CPUs)

L1은 개별자 32 ~ 128 KB 사이 RAM과 함께 캐시로 사용되는 경우.
캐시가 있거나 (⇒ separate 캐시로 사용하는 경우)
L2는 256KB ~ 1MB 사이에 이어지는 경우 캐시로 사용되는 경우.
(⇒ unified 캐시로 사용하는 경우)

Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- Questions about commercial CPU products

- Question#4:** Why are some caches private and other caches shared?

↳ private 과 shared의 차이로 나온다.

- Answer#4**

- Some caches : *merit of private cache > merit of shared cache*
- Other caches : *merit of private cache < merit of shared cache*

→ (private
shared) 와 정답과의 차이를 보여함.

				Type	Sub Type	Product Name	Release date (Year, Month)	Fab. Tech	Power	Name	SoC (System-on-Chip) = CPU + GPU + HW Accelerators				CPU				L1 Cache			L2 Cache		L3 Cache
											Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores									
Group#2											2018 1n 4 x Meerkat (M3) for high performance					2.9 GHz	ARM v8-A							
Group#1											SoC (System-on-Chip) = CPU + GPU + HW Accelerators				CPU				L1 Cache			L2 Cache		L3 Cache
Embedded Computer Systems	Smart Phone	Type	Sub Type	Product Name	Release date (Year, Month)	Fab. Tech	Power	Name	Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache								
		iPhone 4		Apple A4	2010. 06	45 nm	?	ARM Cortex-A8	32	800 MHz	ARM v7-A	1	Per-core: I-32KB, D-32KB	512KB	None					Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared		
		iPhone 4s		Apple A5	2011. 03	32 nm	?	ARM Cortex-A9	32	1 GHz	ARM v7-A	2	Per-core: I-32KB, D-32KB	1MB shared	None					Per-core: I-64KB, D-64KB	Per-core: 256KB	4MB shared		
		iPhone 5s		Apple A7	2013. 09	2~3 W	?	Apple Cyclone	64	1.3 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	1MB shared	4MB shared by the entire SoC					Per-core: I-64KB, D-64KB	Per-core: 256KB	4MB shared by 2 cores and GPU		
		iPhone 6s		Apple A9	2015. 09	16 nm	?	Apple Twister	64	1.85 GHz	ARM v8.0-A	2	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC					Per-core: I-32KB, D-32KB	Per-core: 512KB	8MB shared by 4 cores and GPU		
		iPhone 7		Apple A10 Fusion	2016. 09	16 nm	?	2 x Hurricanes for high performance 2 x Zephyr for energy efficiency	64	2.34 GHz	ARM v8.1-A	4	Per-core: I-64KB, D-64KB	3MB shared	4MB shared by the entire SoC					Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores and GPU		
		Galaxy S8		Exynos 8895	2017. 04	10 nm	?	4 x Mongoose2 (M2) for high performance 4 x Cortex-A53 for energy efficiency	64	2.3 GHz	ARM v8-A	8	Per-core: I-64KB, D-32KB	2MB shared	None					Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores and GPU		
		iPhone X		Apple A11 Bionic	2017. 11	10 nm	?	2 x Monsoon for high performance 4 x Mistral for energy efficiency	64	2.39 GHz	ARM v8.2-A	6	Per-core: I-64KB, D-64KB	8MB shared	None					Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores and GPU		
		iPhone XS		Apple A12	2018. 09	7 nm	?	2 x Vortex for high performance 4 x Tempest for energy efficiency	64	2.49 GHz	ARM v8.3-A	6	Per-core: I-128KB, D-128KB	8MB shared	None					Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores		
		iPhone 11 Pro		Apple A13	2019. 09	7 nm	?	2 x Lightning for high performance 4 x Thunder for energy efficiency	64	2.66 GHz	ARM v8.4-A	6	Per-core: I-128KB, D-128KB	2MB shared	None					Per-core: I-32KB, D-32KB	Per-core: 256KB	10MB shared by 4 cores		
Tablet PC		HP x2 210 G2	2018. 02	Airmont	14 nm	2W	Intel Atom® Processor x5-Z8350		1.44 ~ 1.92 GHz	x86-64	4	Per-core: I-32KB, D-24KB	2MB (1MB shared by 2 cores)	None					Per-core: I-32KB, D-32KB	Per-core: 1MB	22MB shared by 16 cores			

Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

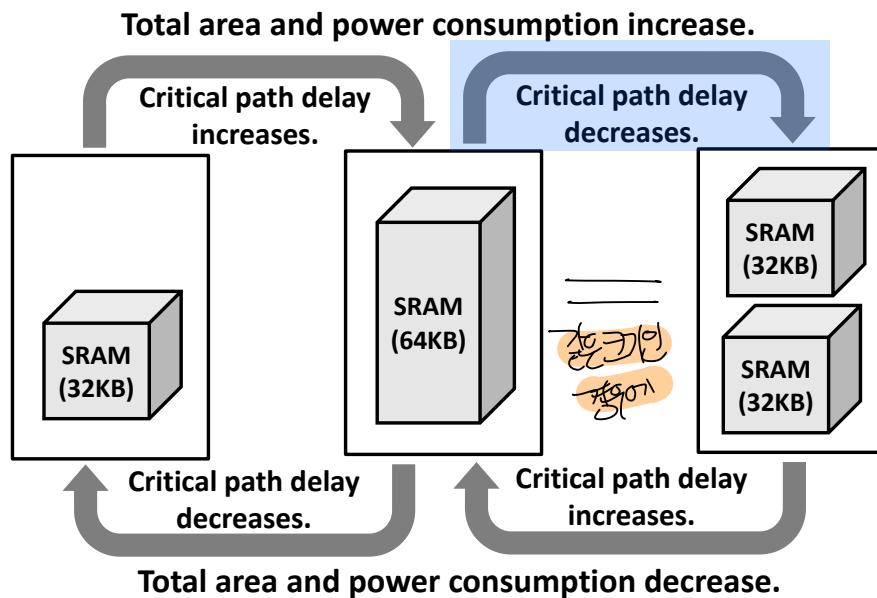


Advantage of private cache memory

① private 캐시가 다른 캐시와 충돌하지 않는다.

- It offers much faster access than shared cache.
 - In the aspect of critical path delay
 - ✓ Small private caches show much shorter delay than a large shared cache.
 - In the aspect of cycle count
 - Private cache is free of delay cycles caused by contention among cores sharing a cache .
↳ shared의 경우 어떤 코어를 핸들링하는 경우는 delay cycle이 있는데
private의 경우 이러한 delay가 없음. (여러 코어가 캐시를 공유)

Separate 캐시와
다른 절.
(다른 장애에 대해서 없음)



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

Disadvantage of private cache memory

- Total area and power consumption of the private caches is larger than the area and power consumption of a shared cache when they have the same size.

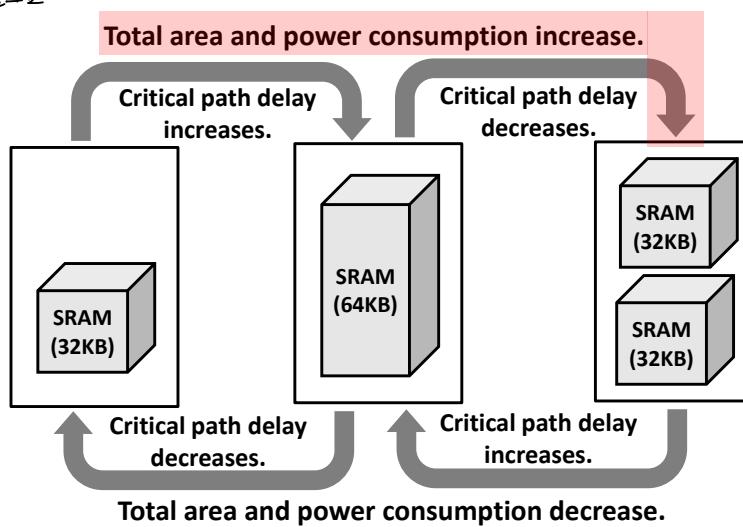
⊕ private 캐시의 흐름, cache coherence problem이 생김. ⇒ 이를 해결하기 위한 부수가 추가로 필요.

- Additional hardware is required to solve **cache coherence problem**.

multi-core↑
private cache
→ cache coherence problem
⇒ H.W를 위한 부수의 필요성

✓ **Cache coherence problem** is a general one with multi-core with private caches.

✓ **additional hardware** means increase of power and area raising the price of product.



한국어로
data를 주입하는 순간에도 퍼포먼스가 감소함.
⇒ 이를 위한 H.W가 추가로 필요함.

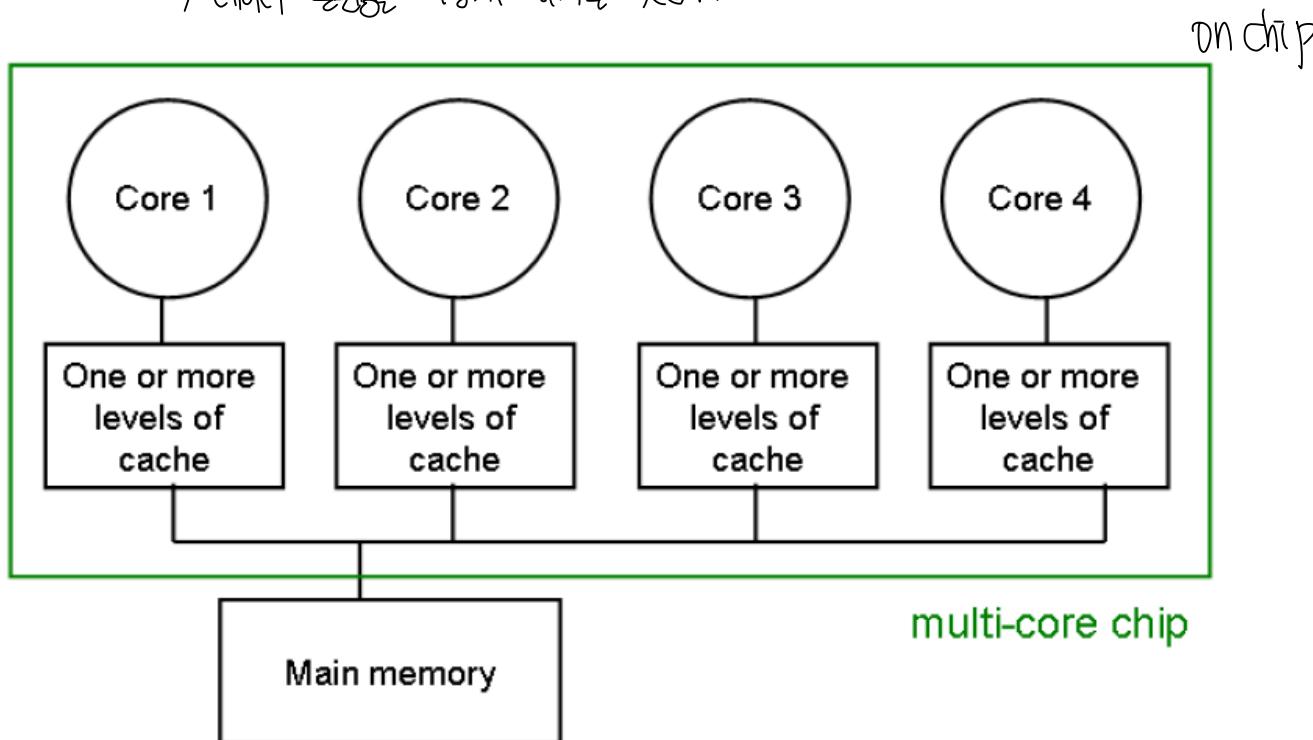
Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- **Cache coherence problem**

- Let's assume a multi-core chip as below.
 - Four cores have private caches.
- How to keep the data consistency among the private caches?

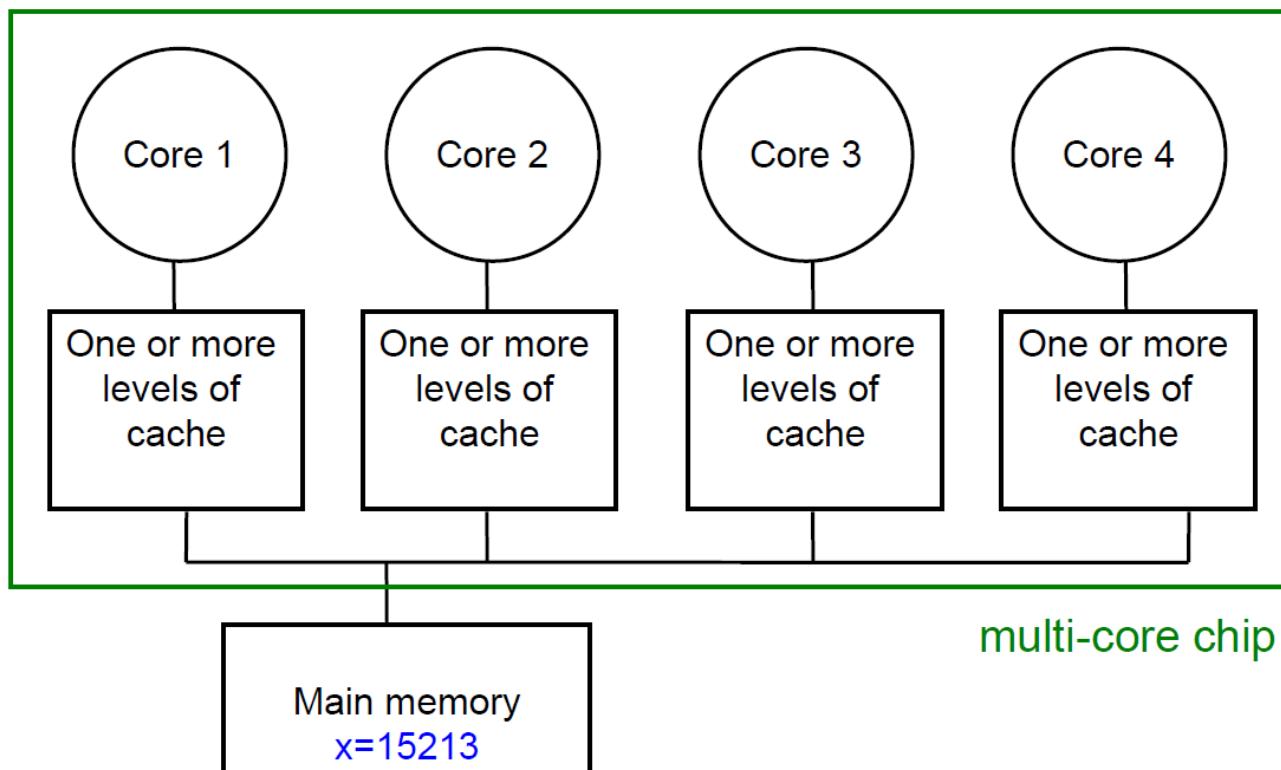
⇒ 케이스의 문제는 주제가 되어야 한다.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- **Cache coherence problem**
 - Suppose variable x initially contains 15213.

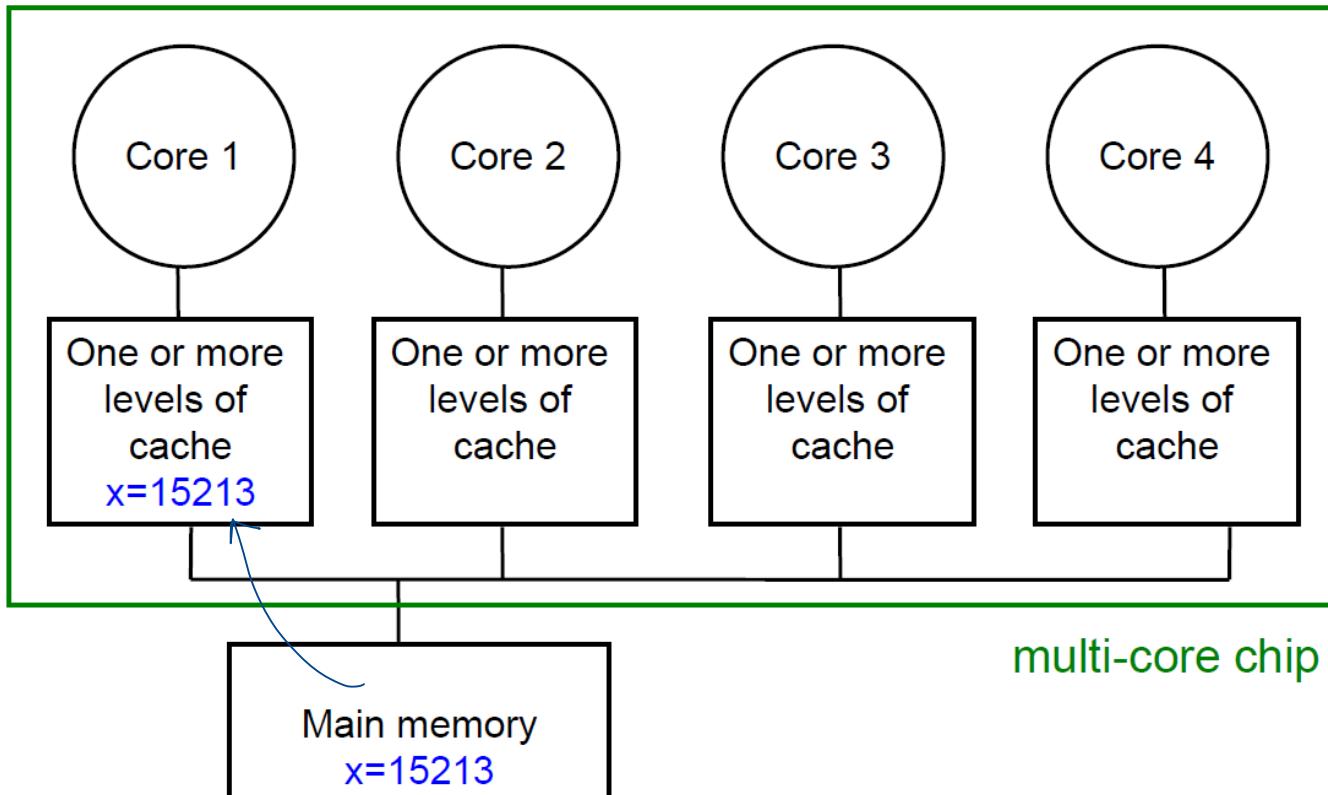


Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- **Cache coherence problem**

- Core 1 reads x.

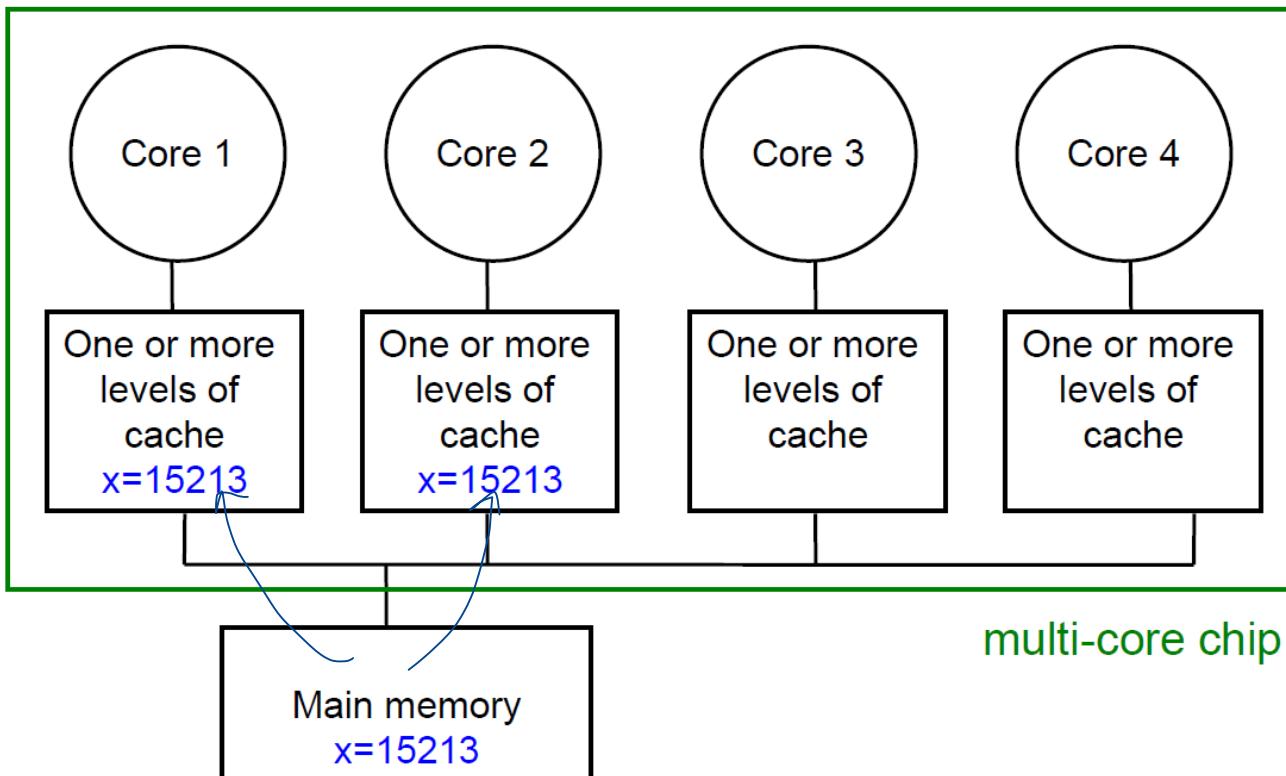


Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- **Cache coherence problem**

- Core 2 reads x.

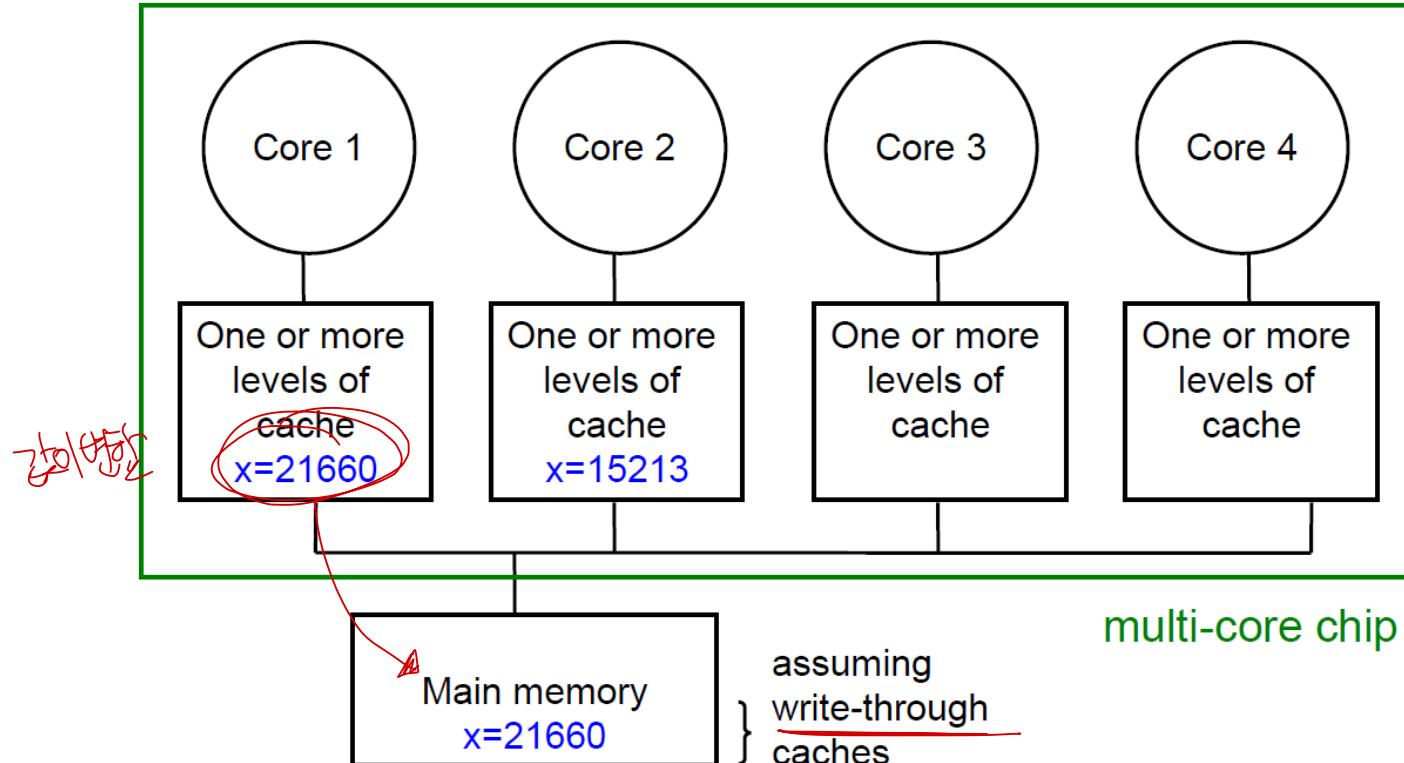


Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- **Cache coherence problem**

- Core 1 writes to x , setting it to 21660.

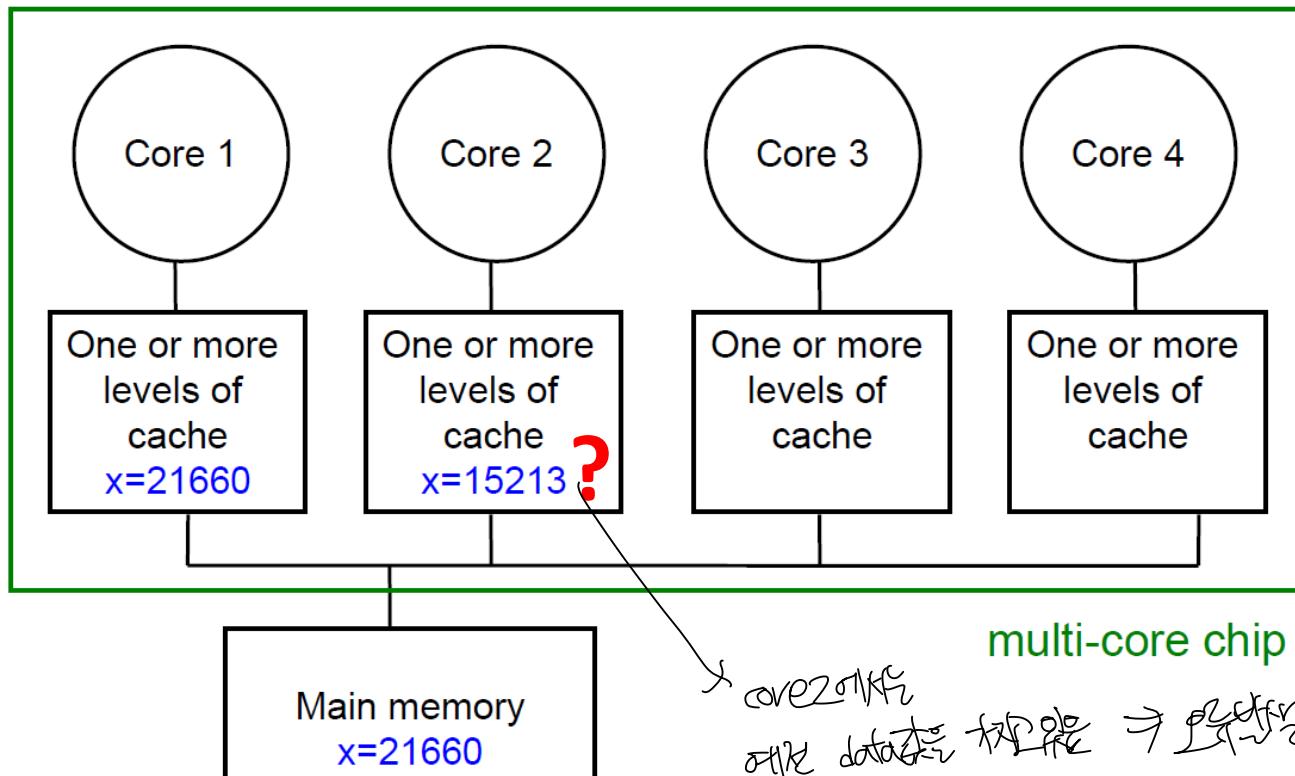


Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- **Cache coherence problem**

- Core 2 attempts to read x but gets a old copy.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- Solutions for **cache coherence problem**

- Additional hardware is needed to support invalidation protocol with snooping. ⇒ 추가적인 HW를 끌고. 부정화 프로토콜.

- Invalidation

- If a core write to a data, all other copies of this data in other caches are invalidated.
↓
이전 데이터를 무효화하고 새로운 데이터는 셋는다.
(Core가 데이터를
쓰면 무효화)

- Snooping

- All cores continuously “snoop” (monitor) the bus connecting the cores.

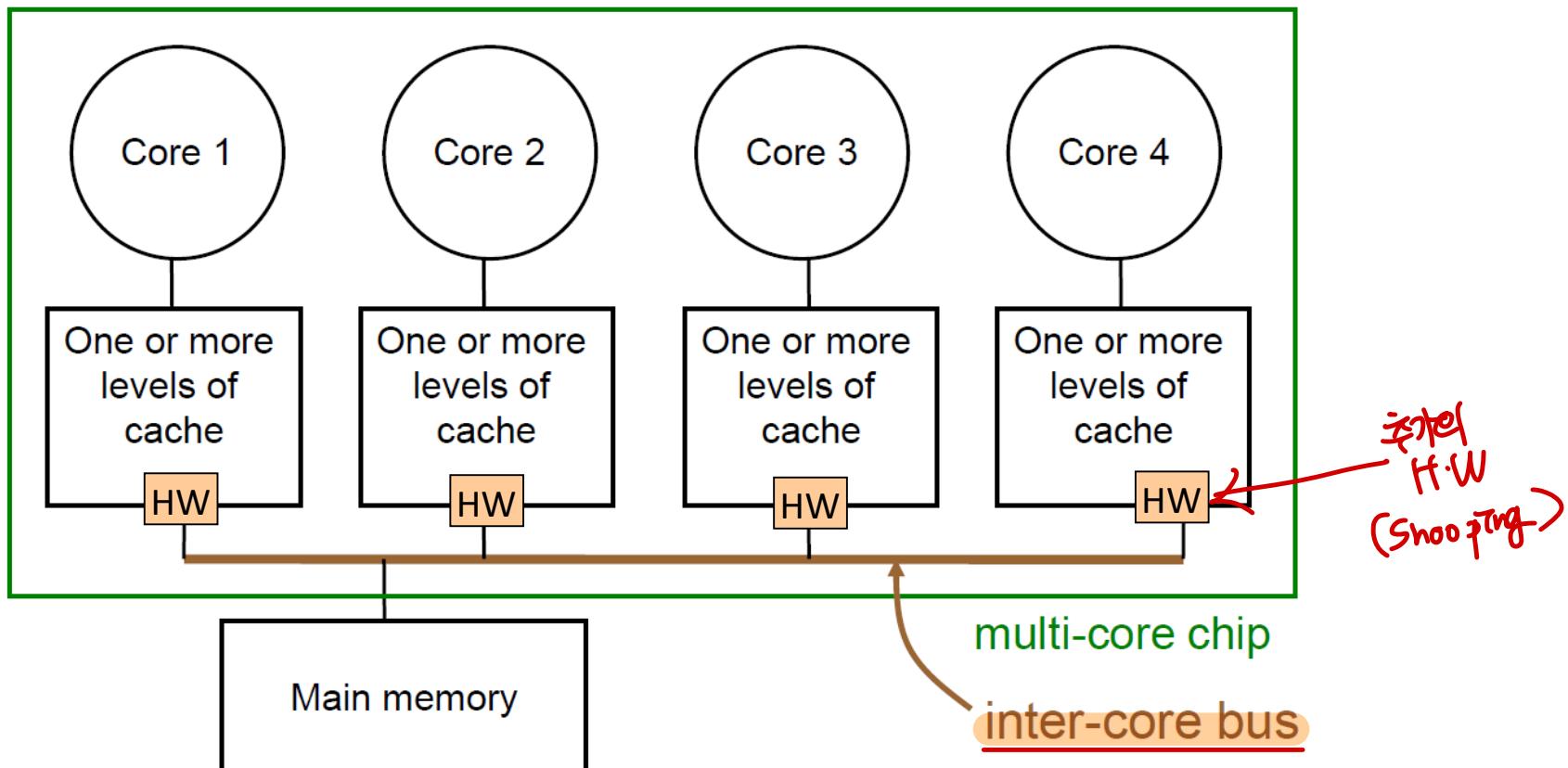
→ 무언가를 찾는다.

Write가 발생했을 때, 그의 번역에 발생한
일의 core에게 차교 읽는다. 이 old 데이터를 얻자고 부정화 신호
invalidation.

Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

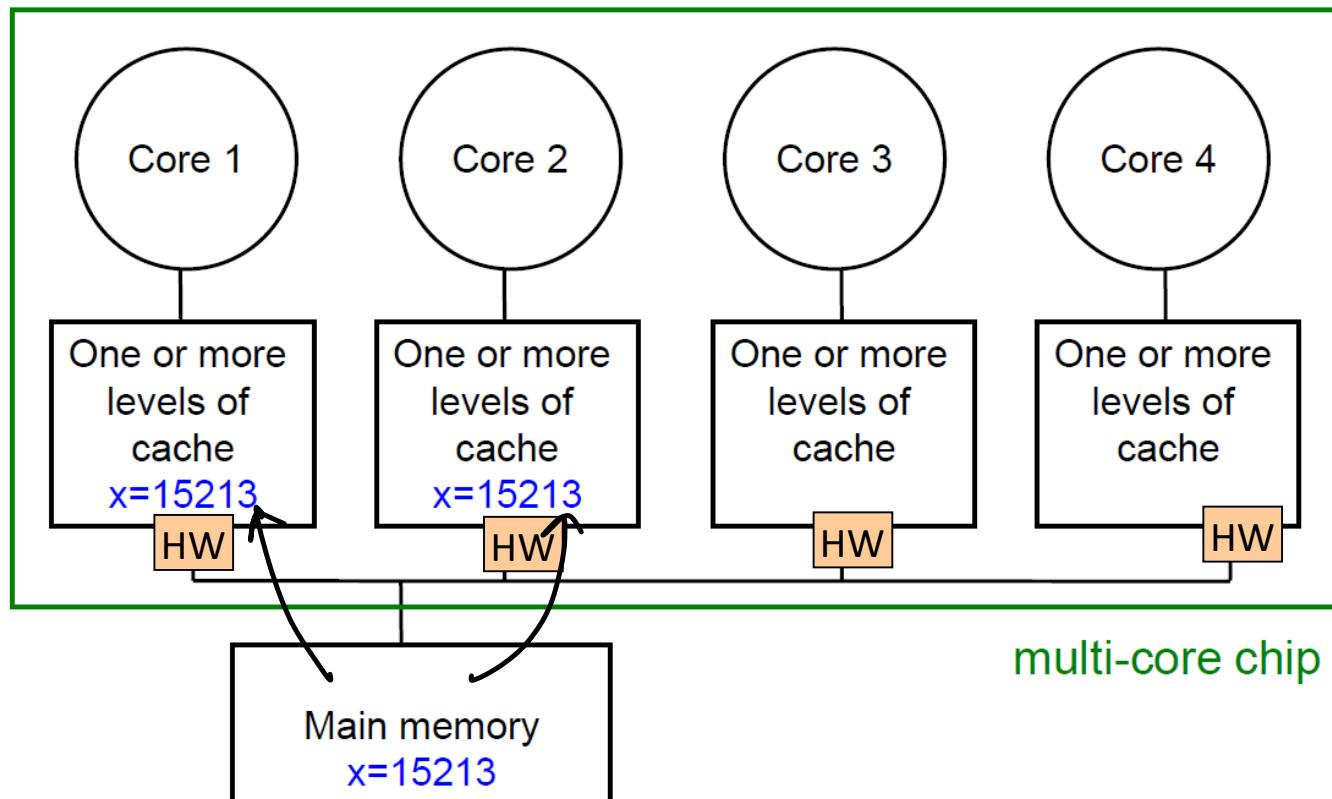
- Bus based “Snooping” multi-core processor



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- Invalidation based cache coherence protocol
 - Cores 1 and 2 have both read x.

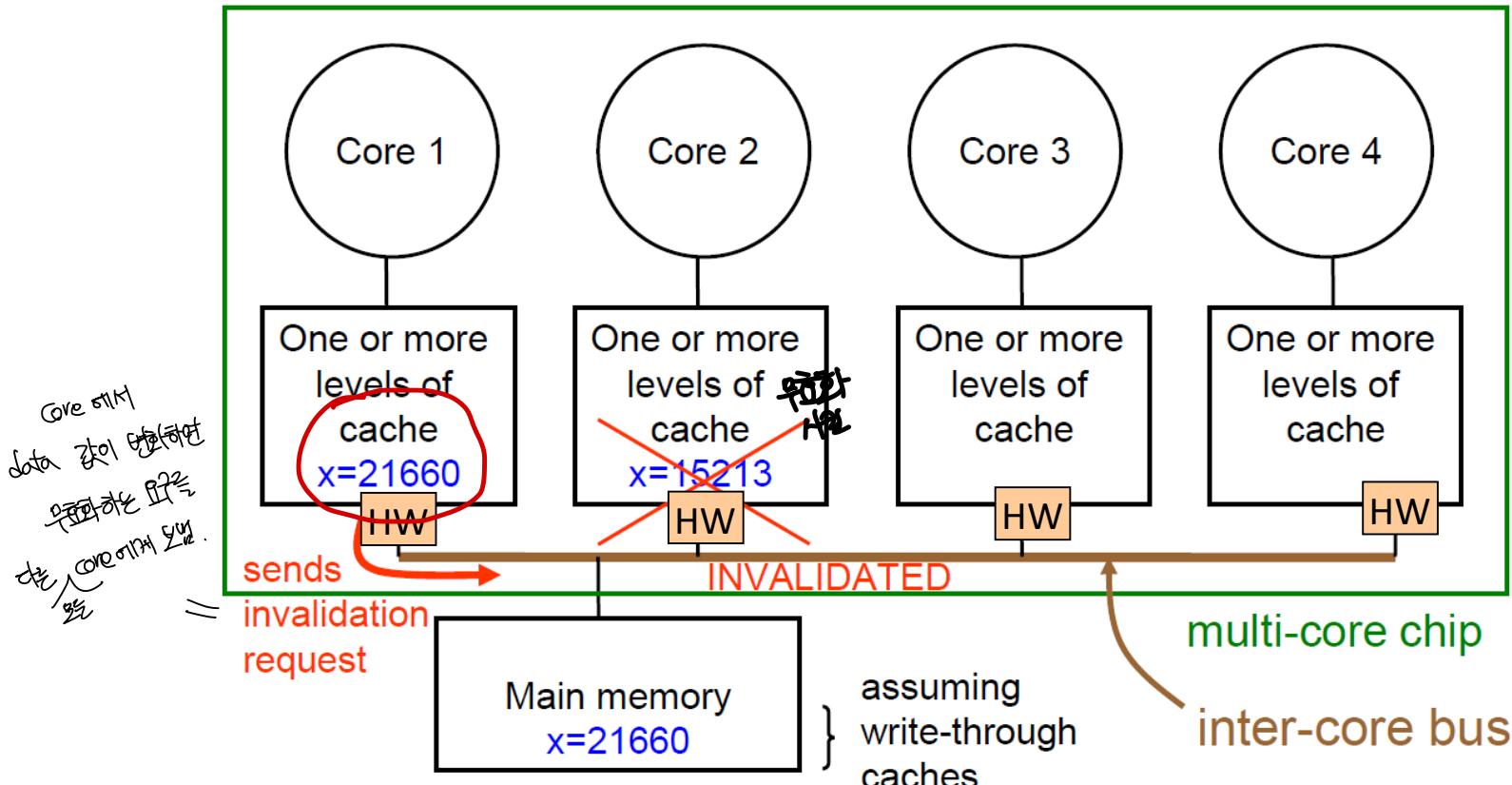


Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- Invalidation based cache coherence protocol

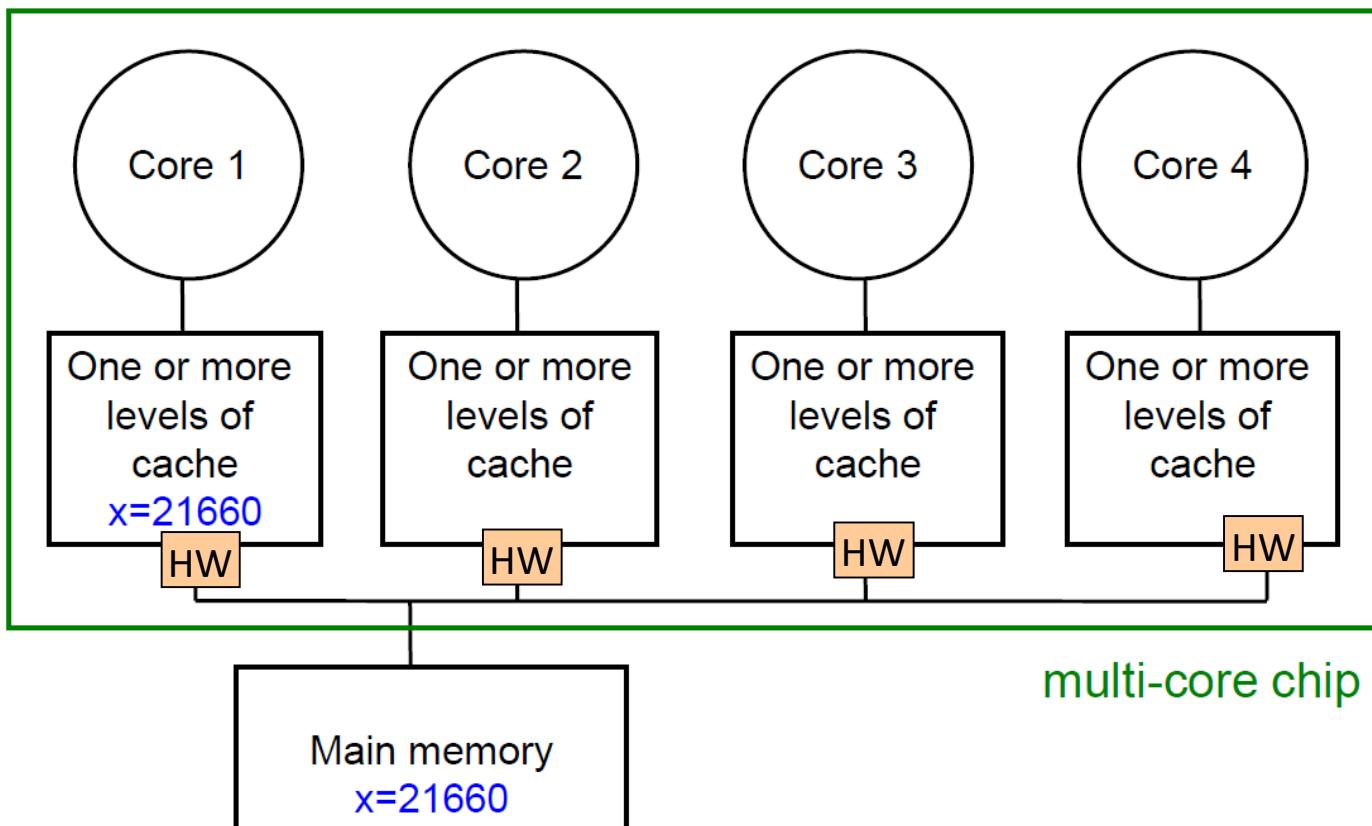
- Cores 1 writes to x , setting it to 21660.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

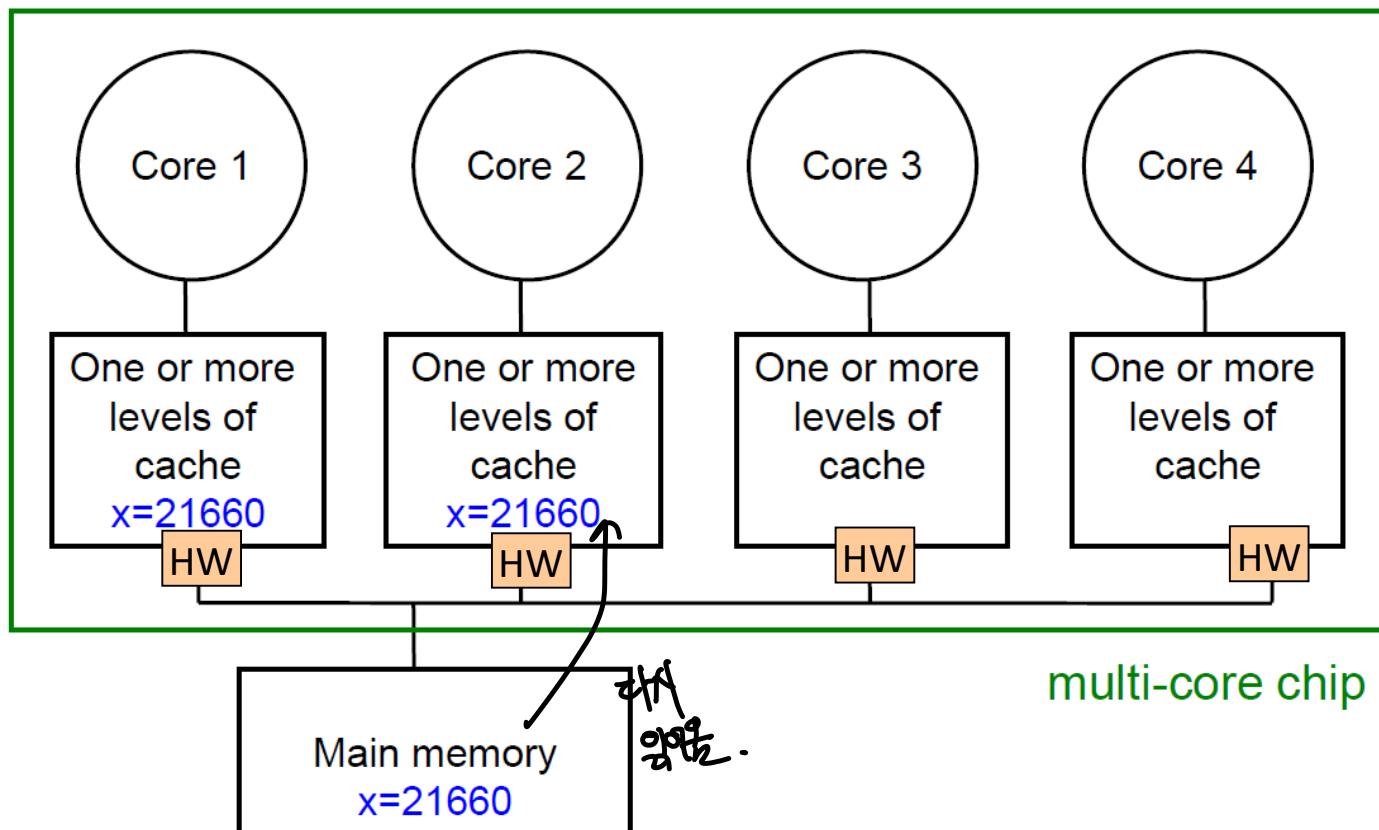
- Invalidation based cache coherence protocol
 - After invalidation



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- Invalidation based cache coherence protocol
 - Core 2 reads x. Cache misses, and loads the new copy.



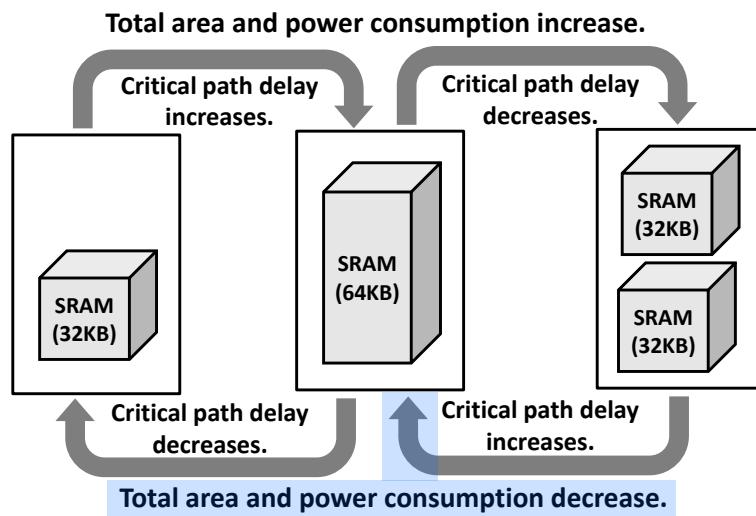
Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

Advantage of shared cache memory

- The area and the power consumption of a shared cache is smaller than total area and power consumption of the private caches when they have the same size. \Rightarrow ① 총 면적과 전력소비가 적다.
- Threads on different cores can share the same cache data without **cache coherence problem**. \Rightarrow ② 캐시 콘센스 문제를 고려할 필요가 없다.
- More cache space are available if a single (or a few) high-performance thread runs on the system.

단일 코어에서 thread 사용량이
증가할 때, 다른 코어에 대한
접근성이 좋다.



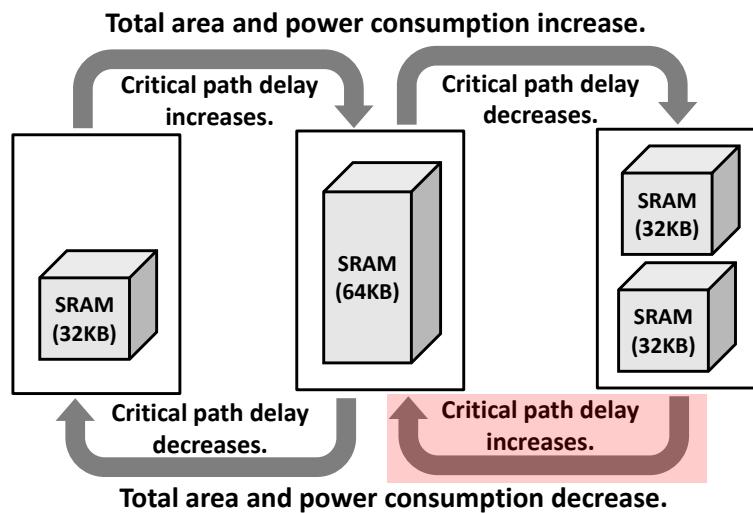
Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core



• Disadvantage of shared cache memory

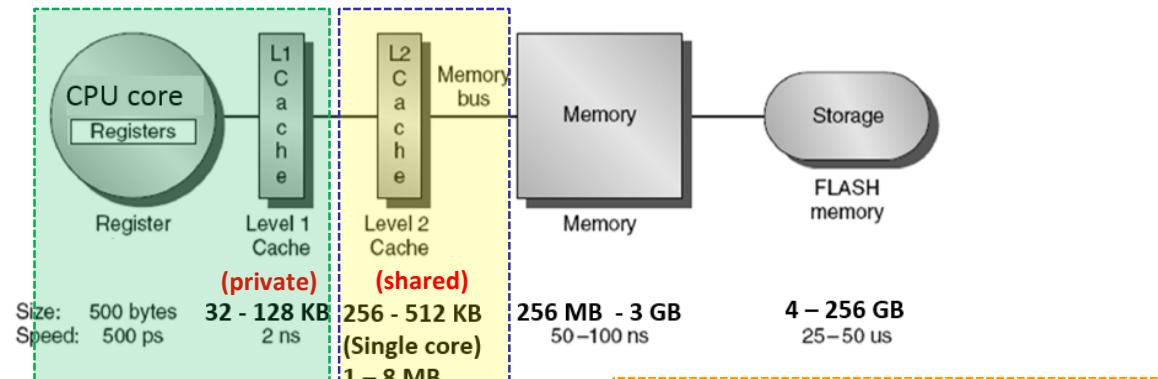
- It offers much slower access than private cache.
 - In the aspect of critical path delay \Rightarrow ① critical path delay \rightarrow 증가
✓ Small private caches show much shorter delay than a large shared cache.
 - In the aspect of cycle count \Rightarrow ② cycle count \rightarrow 증가.
 - Private cache is free of delay cycles caused by contention among cores sharing a cache .
 \hookrightarrow 어떤 core를 접속하지 결정하는데
delay cycle이 발생함.



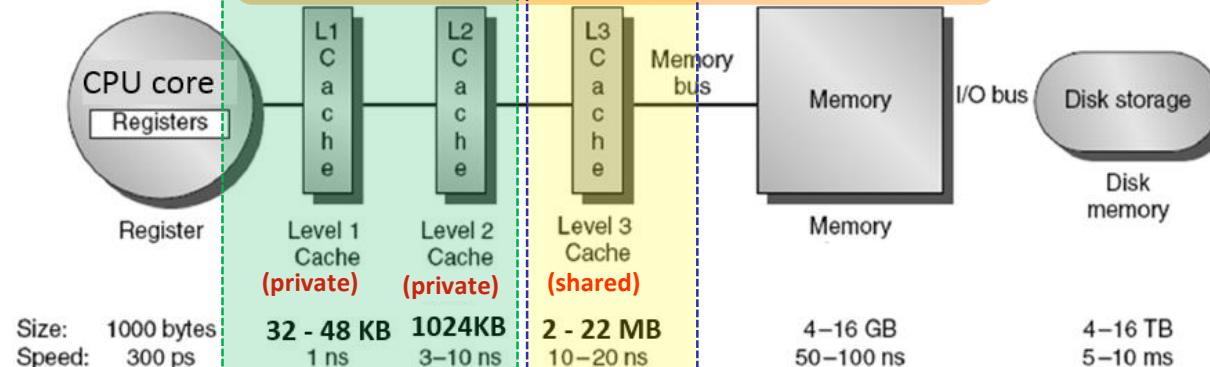
Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- Multi-level cache organization



(a) Memory hierarchy for embedded CPU



(b) Memory hierarchy for desktop/server/workstation CPU
(+ Latest SAMSUNG embedded CPUs)