Introduction to Cache Memory – Part#2

↳ "실질적인 제품의 캐시메모리에" 대해 이야기해볼 것임.

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 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	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
		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	
							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

- **Selected commercial CPU products – Group#2**

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
① 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	
② 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
③ 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

Cache Memory Examples of Commercial CPU Products

- **Questions about commercial CPU products**
  - Question#1: Why are caches organized in the multi-level manner?

Group#1

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	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	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

Cache Memory Examples of Commercial CPU Products

- **Questions about commercial CPU products**
  - Question#1: Why are caches organized in the multi-level manner?

Group#2

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	CPU					L1 Cache	L2 Cache	L3 Cache
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores			
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	
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

Cache Memory Examples of Commercial CPU Products

- **Questions about commercial CPU products**
  - Question#2-1: Why do most of non-embedded CPUs have 32 KB L1 caches?

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	
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

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?

Group#1

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	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#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?

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	
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

Cache Memory Examples of Commercial CPU Products

- **Questions about commercial CPU products**
  - Question#3: Why are L1 caches separate and L2/L3 caches unified?

Group#1

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators										
				Fab. Tech	Power	Name	CPU					L1 Cache	L2 Cache	L3 Cache
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores			
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?

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	
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

CPU만 쓸 수 있는 메모리를 cache라고 함.

Cache Memory Examples of Commercial CPU Products

- **Questions about commercial CPU products**
  - Question#4: Why are some caches private and other caches shared?

private → core당 가지고 있는 것
shared → core끼리 공유하는 것

→ 왜 똑같이 cache는 private하고 다른 cache는 shared 인가?

per-core　　shared　　iPhone은 아직 private L2를 채택하지 않음.

Group#1

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	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

GPU나 HW accelerator와 공유하는 메모리

L2 cache가 shared임. (window 10을 원활하게 돌리려고) ← Tablet PC를 Laptop으로 구분하는 이유

iPhone 에서는 여전히 L2캐시의 크기만 늘려서 처리를 진행. ⇐ Galaxy 이거

Smart-phone 에서 하는 것이 과연 효율적인 것인가!?
↳ L2 cache를 private으로 두면 전력소모가 매우 크기 때문에
→ 아직 desktop과 workstation에서는 빅리틀을 L2에 private하게 두기 때문에

이전까지는 L2의 (private / shared) 차이가 (embeded / non-embeded)의 차이였음.

이제 기준이 바뀌어버림.

⊕ 추가내용 (빠진내용)

Cache Memory Examples of Commercial CPU Products

- **Questions about commercial CPU products**
  - Question#4: Why are some caches private and other caches shared?

private 하게 / shared

Group#2

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators			CPU							
				Fab. Tech	Power	Name	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
							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
							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	
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	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-core: 512KB	8MB shared by 4 cores and GPU
	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-core: 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-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 and GPU
		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

→ 실험에서 사전적으로 해보는 것

L3를 shared함. ⇐ core당 L2를 가지고 있음. ⇐ general PC & workstation은 L2가 shared 가 아님.

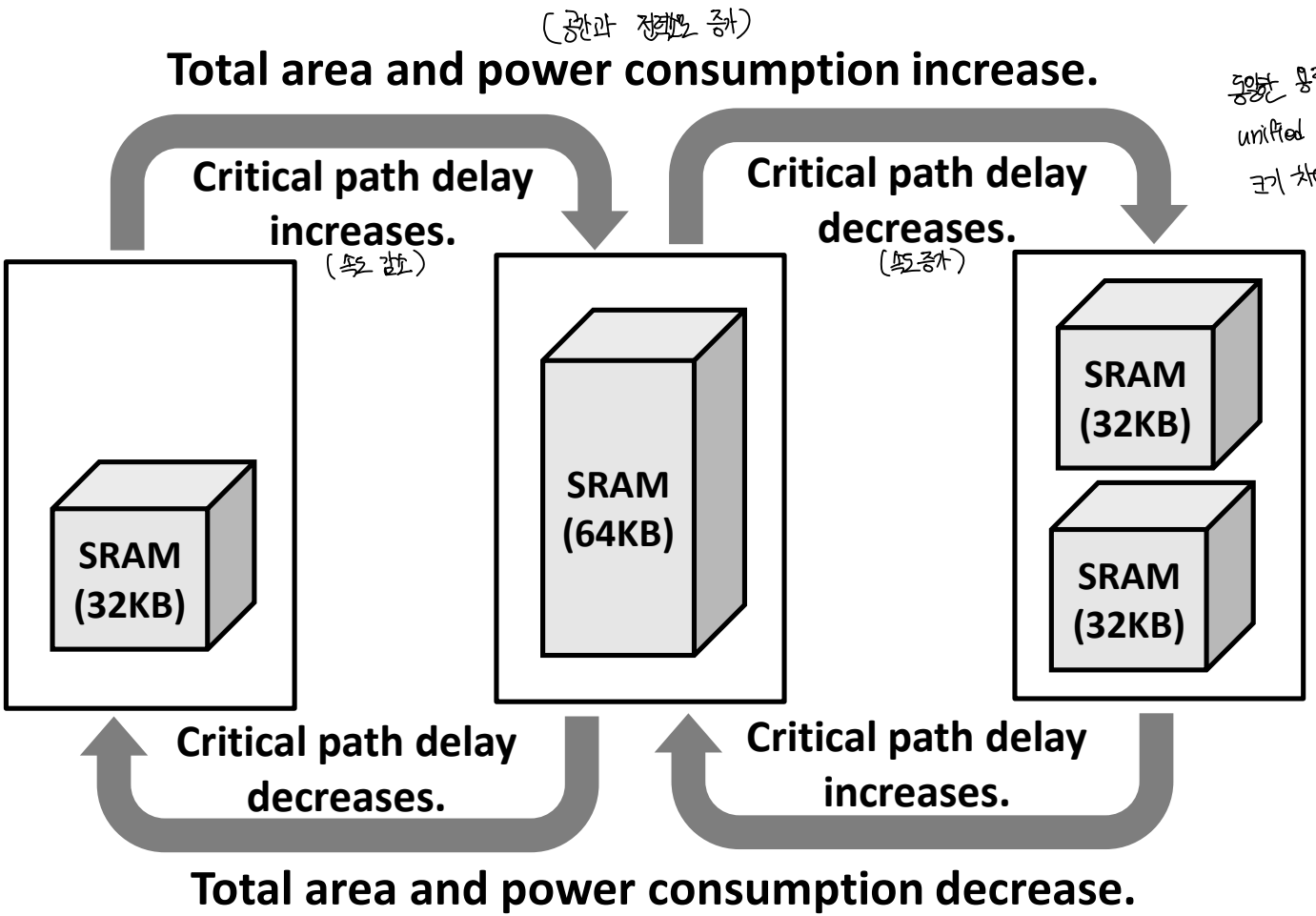
Deriving Issues of Cache Memory

- **Cache memory examples of commercial CPU products**
  - 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

Physical Characteristics of SRAM for Handling Issues of Cache Memory

- For example,



Issues of Cache Memory - 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

Group#2

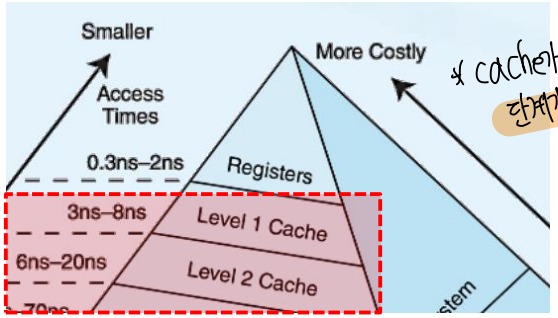
Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators											
				Fab. Tech	Power	Name	CPU					L1 Cache	L2 Cache	L3 Cache	
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores				
		Galaxy S9	2018. 03	10 nm	?	Exynos 9810	4 x **Meerkat (M3)** for high performance	64	2.9 GHz	ARM v8-A	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared	
												Per-core: I-32KB, D-32KB	256KB Shared		

Group#1

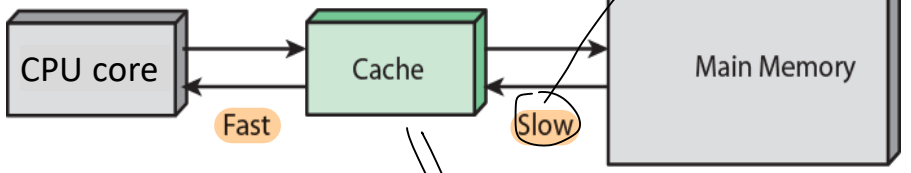
Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators											
				Fab. Tech	Power	Name	CPU					L1 Cache	L2 Cache	L3 Cache	
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores				
Embedded Computer Systems	Smart Phone	iPhone 4	2010. 06	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	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	

Issues of Cache Memory
- Multi-Level Cache

- **Cache performance can be improved by two cases.** (cache 성능 향상방법 – 2가지)
  - Case#1: Enhancing hit-rate ⇒ hit를 더 많이 한다
  - Case#2: Reducing memory access time ⇒ Memory 접근 시간을 줄인다.
- **Purpose of multi-level cache organization** ⇒ Multi-level로 cache를 구성하는 이유
  - In order to improve cache performance by case#2 ⇓ 메모리 접근 시간을 줄이기 위해서
    - L2/L3 cache access much faster than DRAM ⇒ DRAM (Main Memory)보다 빠른 L2, L3 cache를 두어 메모리 접근 시간을 줄임.
  - Most of today's computer systems employ multi-level cache. ⇒ 이것은 multi-level cache를 사용하는 이유임.



※ cache가 단계적일 때

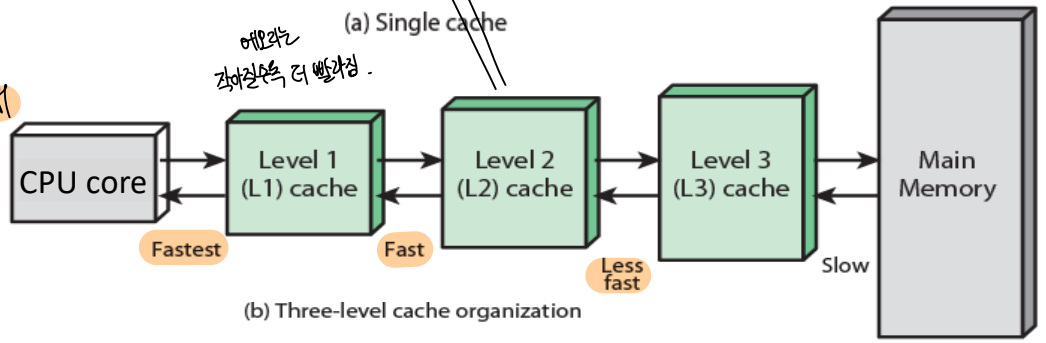


(a) Single cache

cache와 중간속도 칩셋 있음

메모리 접근속도 더 빨라짐.

※ cache가 단계적일 때

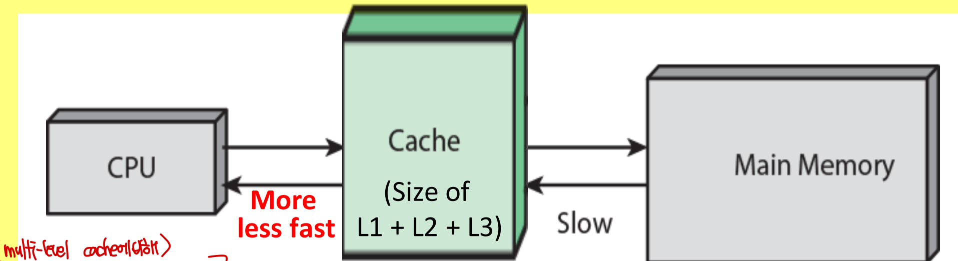


(b) Three-level cache organization

Issues of Cache Memory
- Multi-Level Cache

How about only a large single cache?

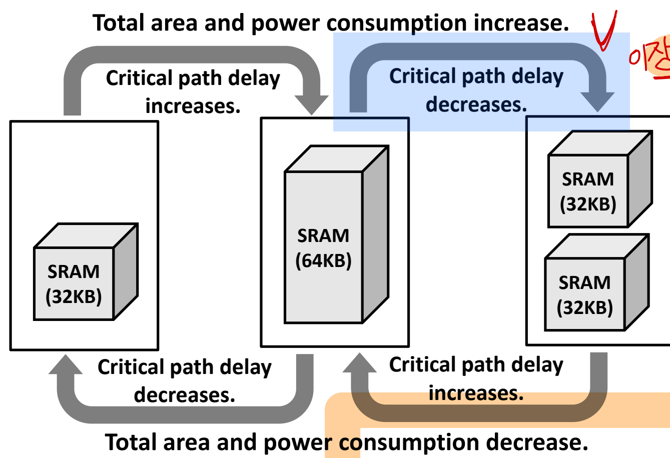
⇒ 하나의 큰 single cache를 두면 안되는 것인가? (L1+L2+L3)



(c) Large single cache

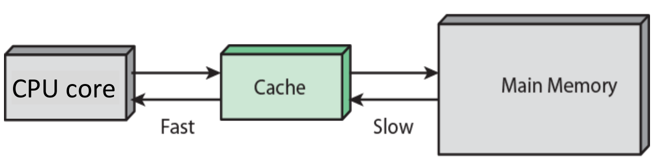
(multi-level cache에 비해) 캐시메모리와 CPU의 접근속도가 좀 느려짐. (∵ 부피가 커지면 delay가 커지기 때문에)

따라서 single cache가 쓰이지 않는 것임.

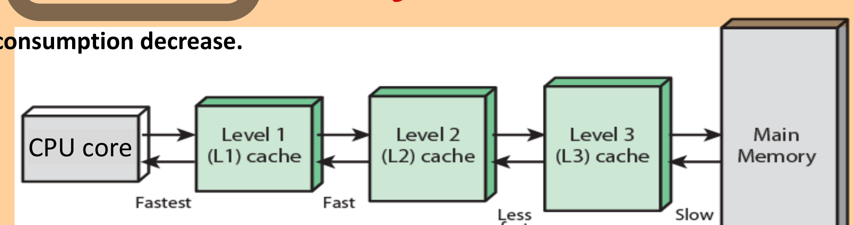


비교 ⇒ Multi-level cache가 전력소모, 면적이 더 큼 but 스피드가 매우 빠름

Multi-level cache



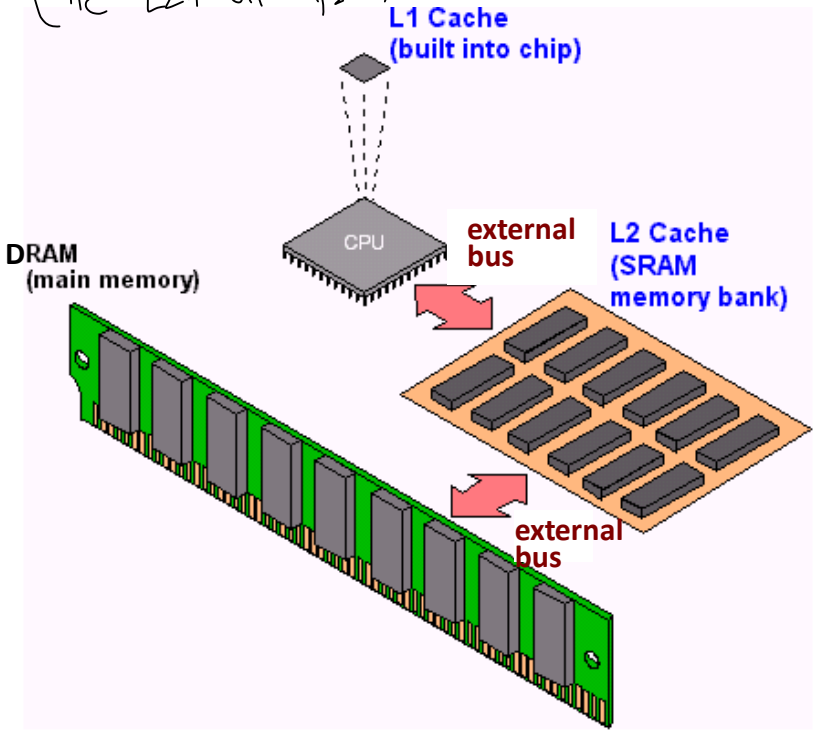
(a) Single cache

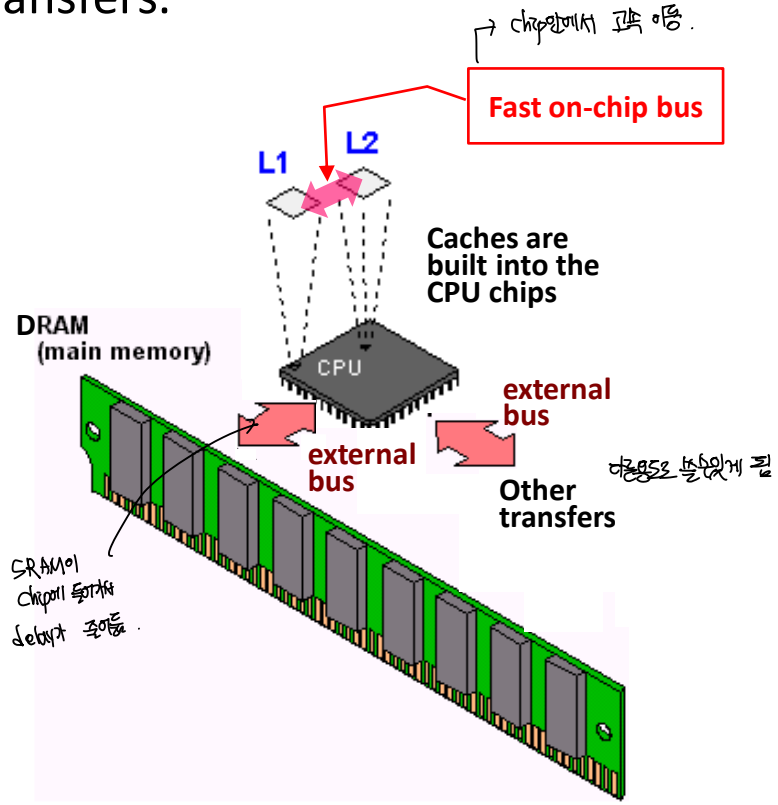


(b) Three-level cache organization

Issues of Cache Memory - 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.



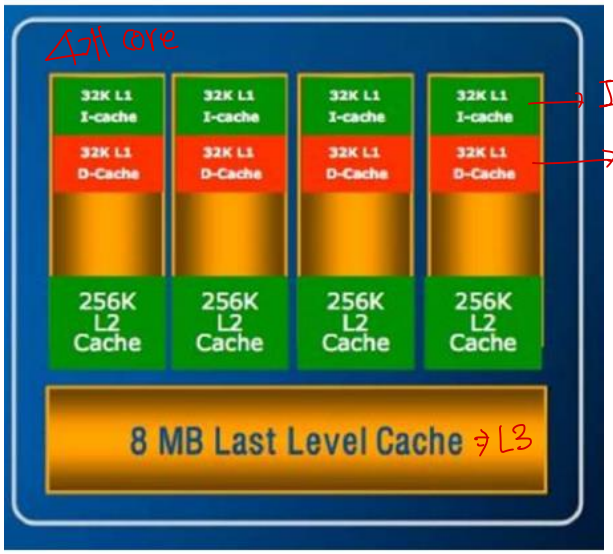


Issues of Cache Memory

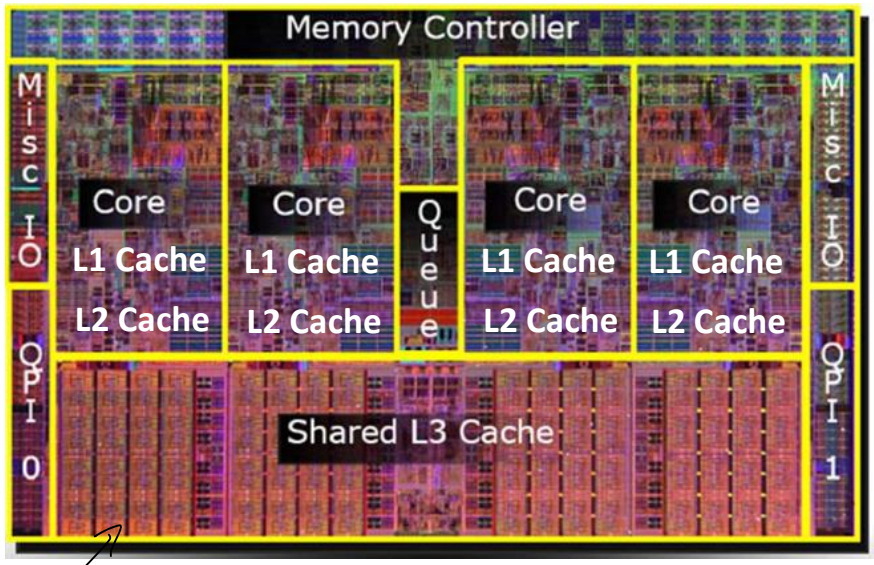
- Multi-Level Cache

- **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	L3 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	Per-core: 256KB	8MB shared by 4 cores and GPU



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



Intel® Core™ i7-6700 Chip
Die-Photo

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							
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores	L1 Cache	L2 Cache	L3 Cache
		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

per-core로 못 넣음 ⇒ power를 너무잡아먹어서

(shared는 per-core와 반대됨).

Apple A7 Chip Die-Photo

↳ 이건 대부분 안보여줌.

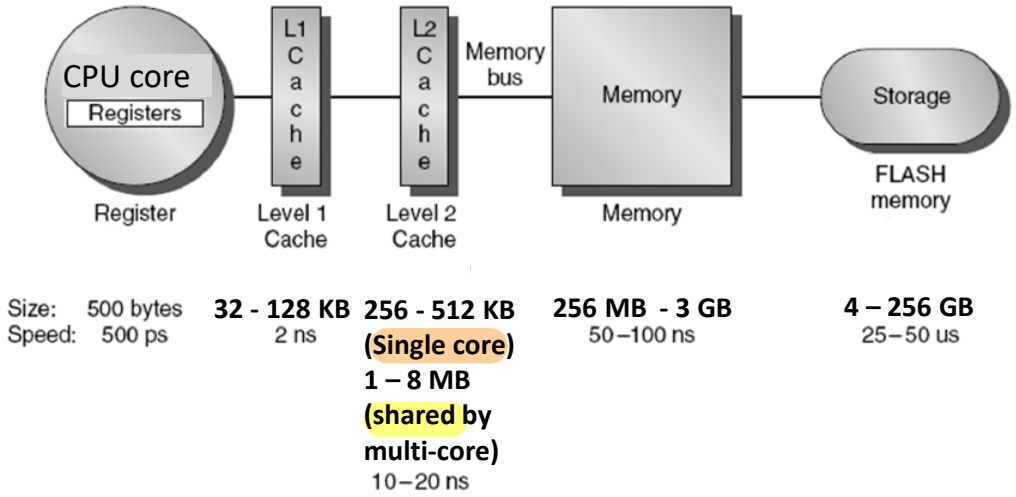
→ H-W의 local buffer로 쓸 것임.

(이름은 L3지만)

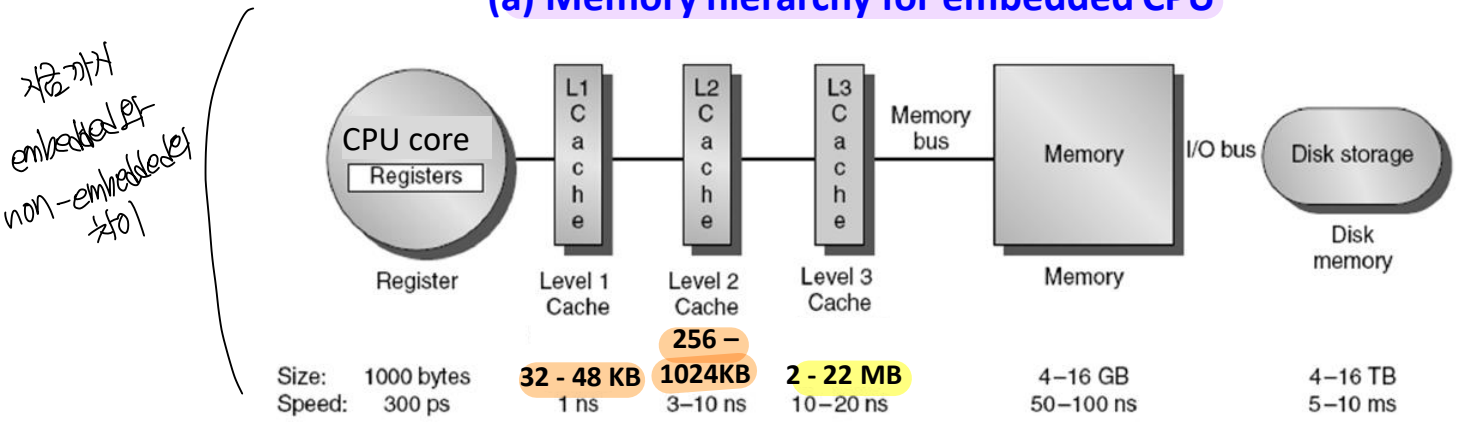
↳ L3라고 이야기하기 눈탐...

Issues of Cache Memory
- Multi-Level Cache

- **Multi-level cache organization**



(a) Memory hierarchy for embedded CPU



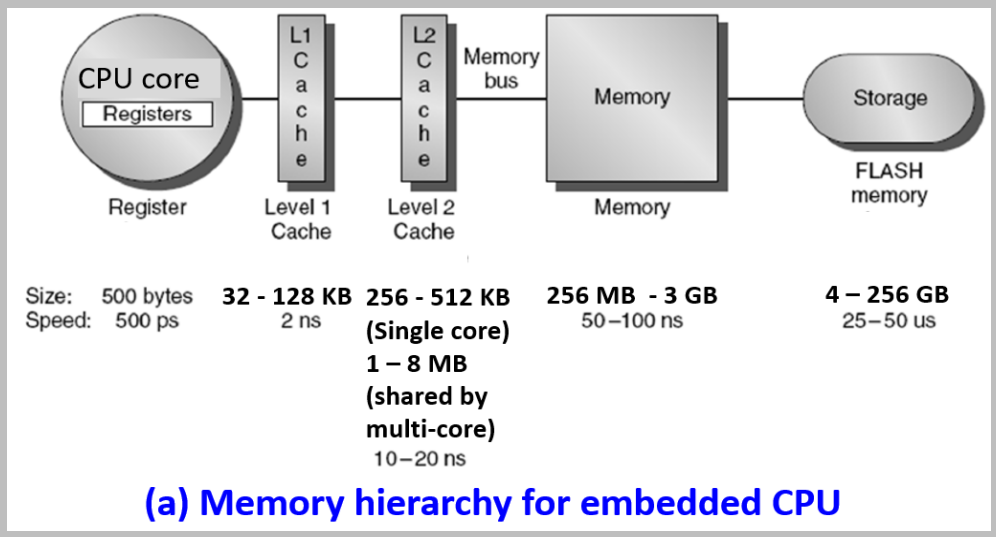
(b) Memory hierarchy for general PC/server/workstation CPU

Issues of Cache Memory
- Multi-Level Cache

- **Multi-level cache organization with 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	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										
		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										



(a) Memory hierarchy for embedded CPU

Issues of Cache Memory

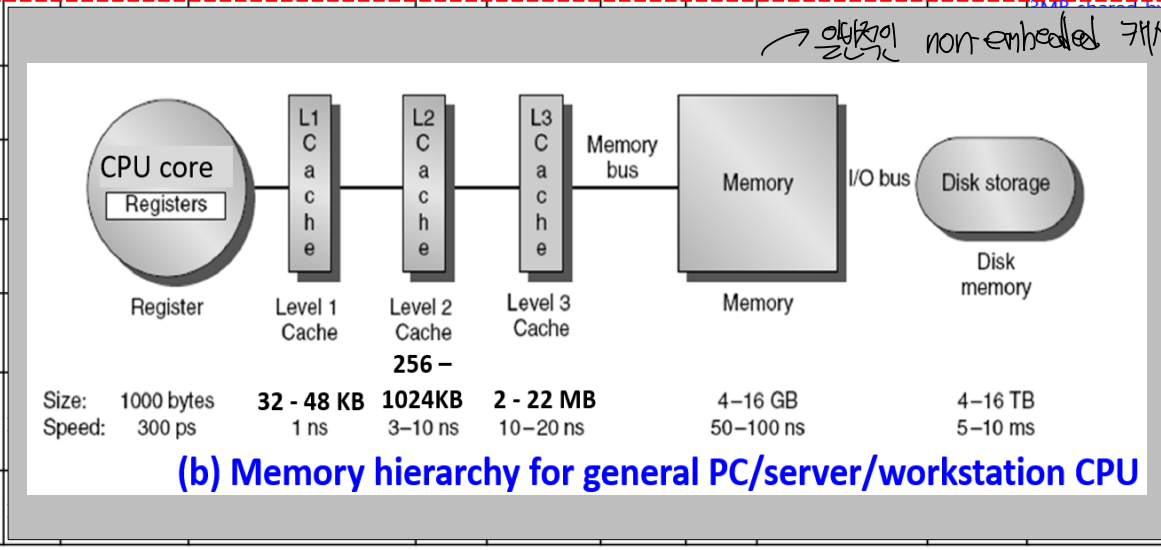
- Multi-Level Cache

- **Multi-level cache organization with commercial CPU products**

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	
General PC	Laptop	SAMSUNG NT900X3N-K59SS	2015. 03	14 nm										
		LG GRAM 17ZD90N-VX70K	2019. 12	10 nm										
	Desktop	SAMSUNG DB400T6B	2015. 12	14 nm										
		Danawa 190721	2019. 07	14 nm										
Server/ Workstation	Floating License/ Data Sever	TYAN KFT46	2011. 09	32 nm										
	High performance Workstation	Lenovo ThinkStation P500TW 850W	2015. 07	22 nm										
		TYAN TAKO-KQT44	2020. 04	14 nm										



(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.

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
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

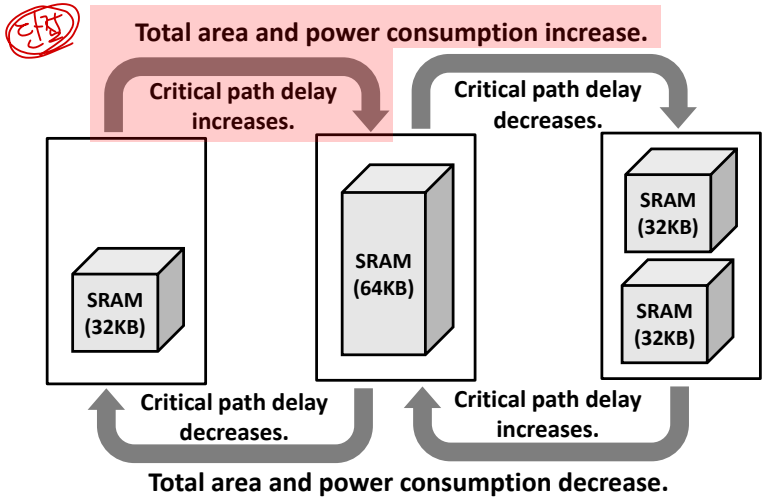
Issues of Cache Memory – L1 Cache Size

- **Generally, increasing L1 cache size** ⇒ 일반적으로 L1 cache가 커지면 hit-rate도 커짐.
  - improves cache performance by enhancing hit-rate. (case#1).

↗ 그러나, L2 캐시가 per-core에 있다는 가정하에, L1이 32 KB를 넘기면

- **However, if L1 cache size exceeds 32 KB under private L2 cache organization**
  - its performance is hardly improved any more. ⇒ 성능향상이 거의 없음.
  - whereas its area and power consumption increase a lot. ⇒ 면적과 전력소모가 매우 커짐.

L1이 32KB를 넘기면 → L2가 성능을 저하시키게 됨.



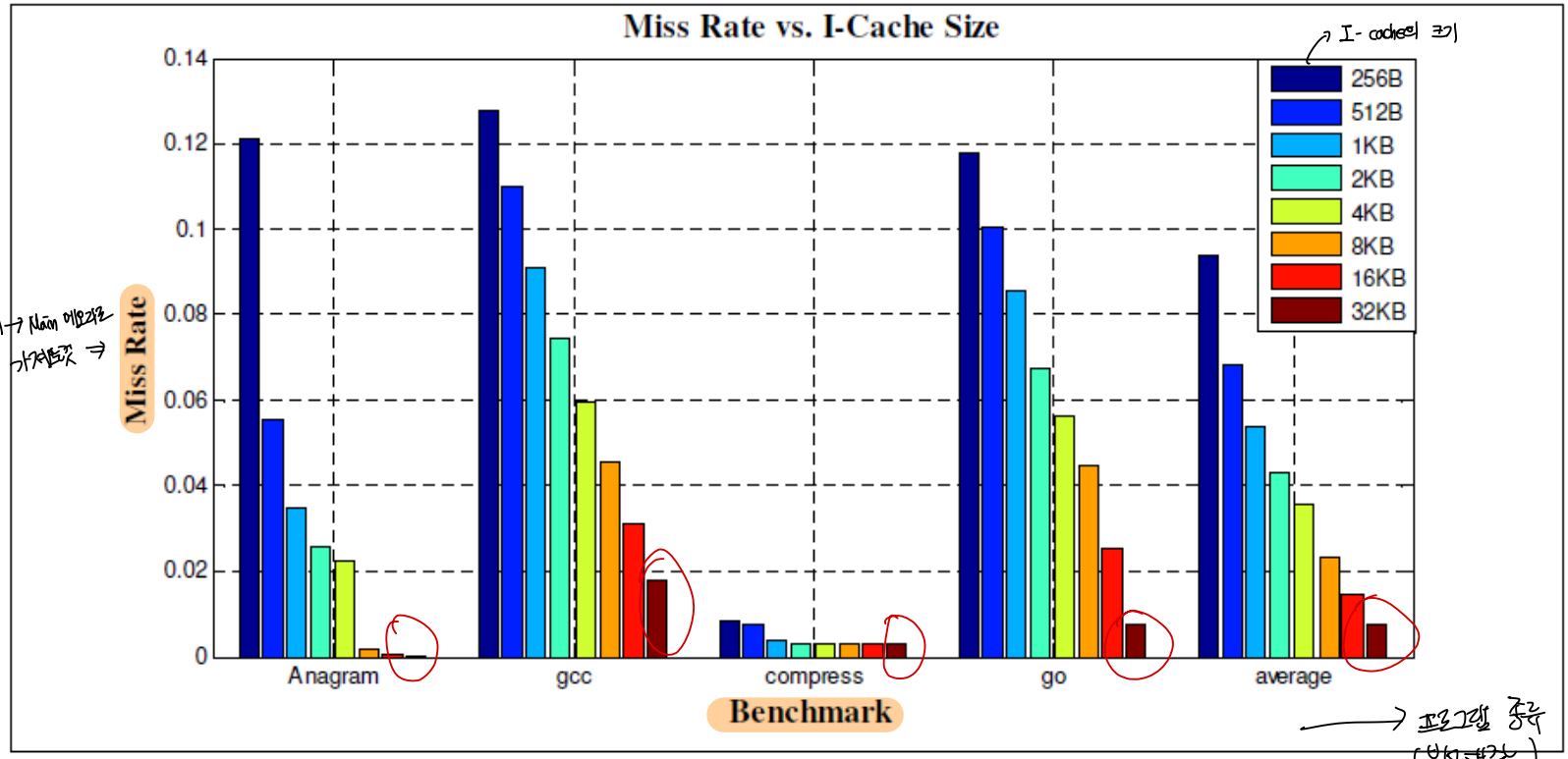
Issues of Cache Memory – L1 Cache Size

- **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**
  - 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
    - 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.

Issues of Cache Memory – L1 Cache Size

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

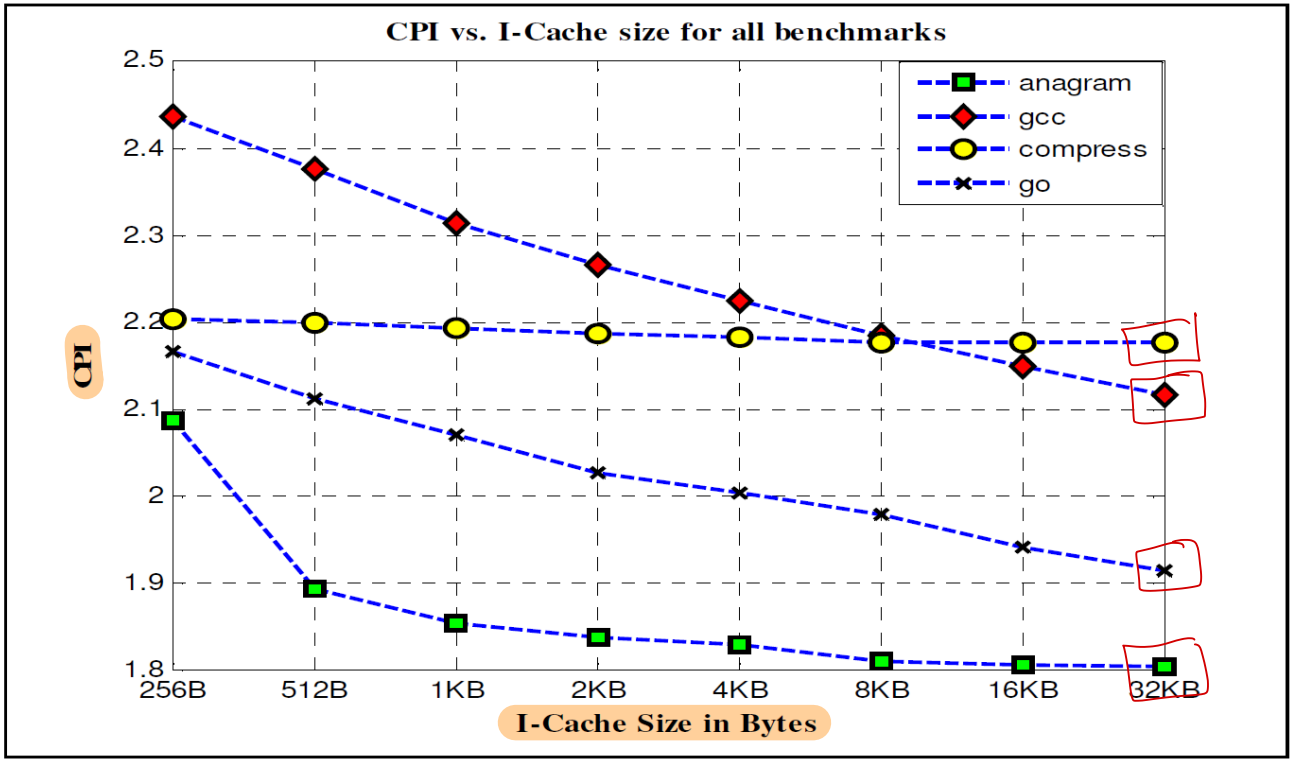
Figure 1: The Impact of I-Cache Size on its Miss Rate.



Issues of Cache Memory – L1 Cache Size

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

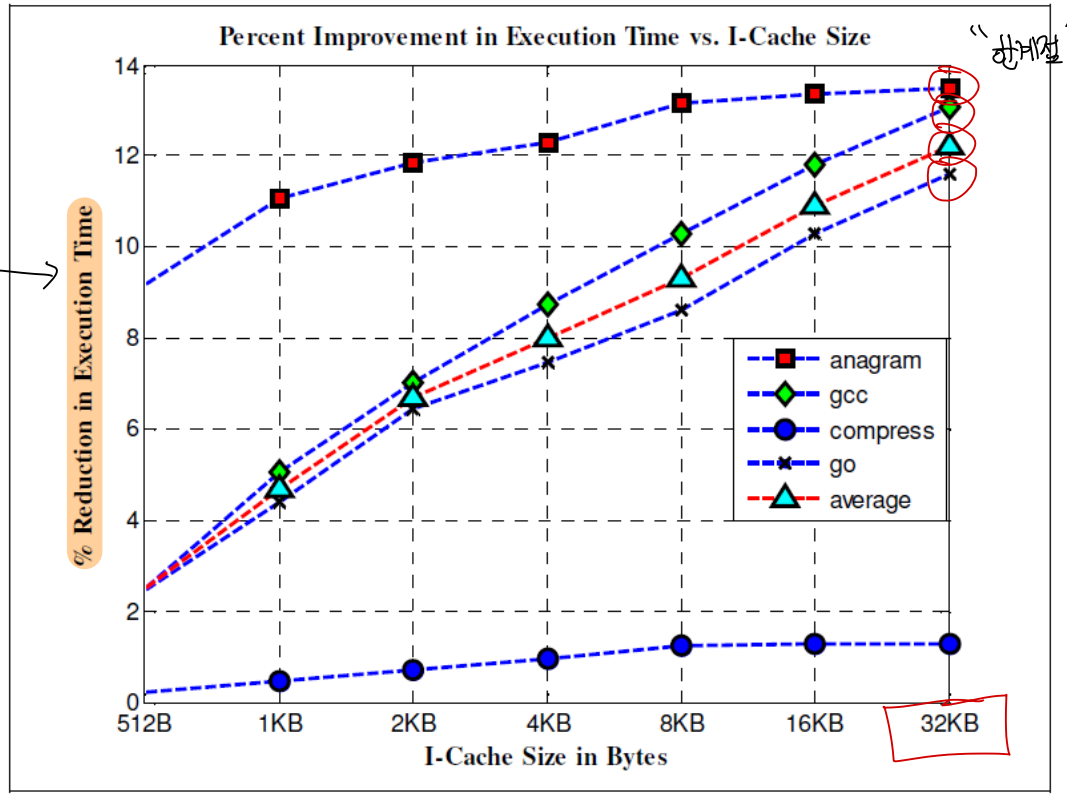
Figure 2: CPI of different benchmarks as I-Cache size increases.



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.

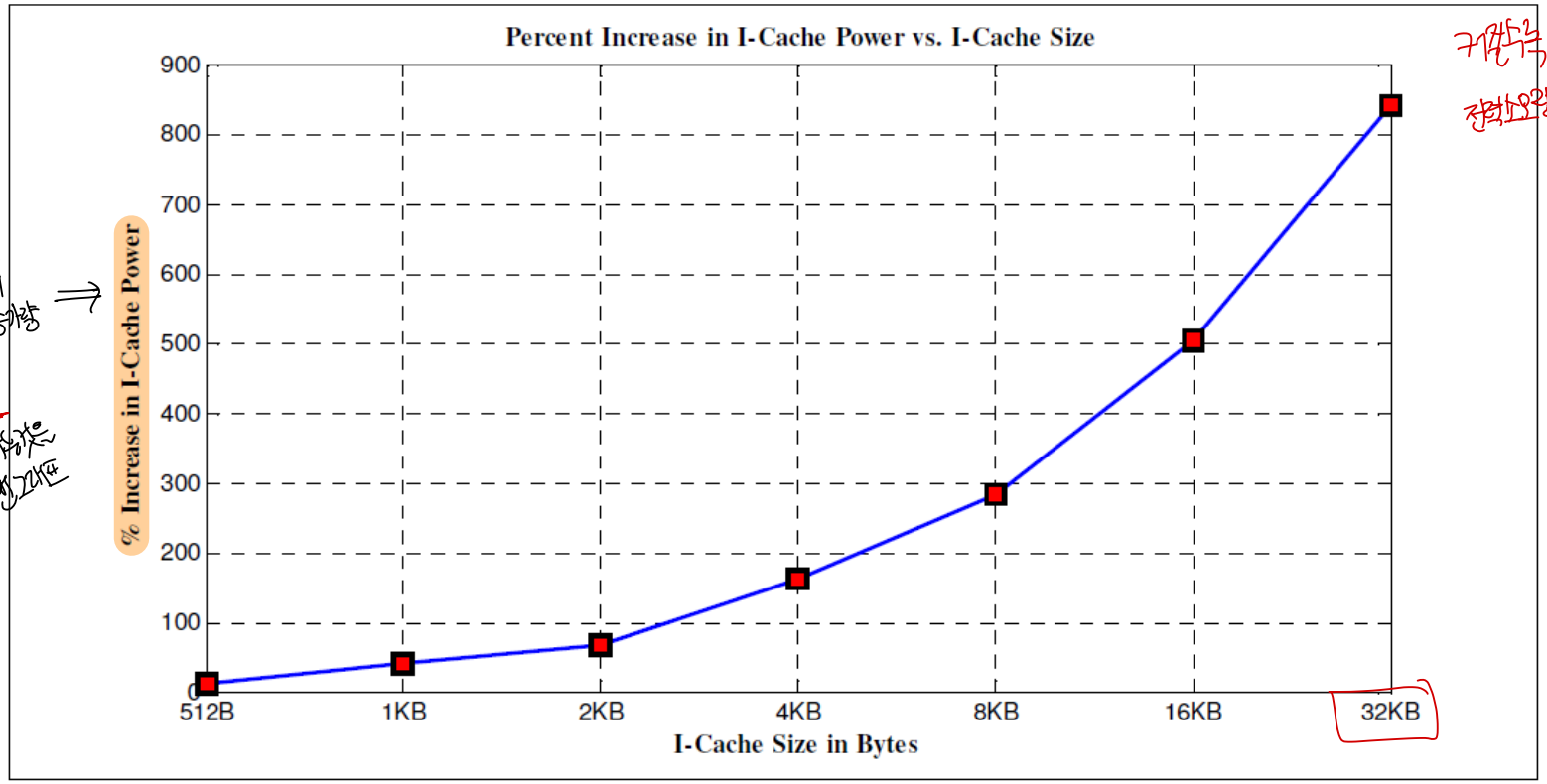


29

Issues of Cache Memory – L1 Cache Size

- **Survey paper by '*Mutaz Al-Tarawneh*'**
  - **Size versus power**

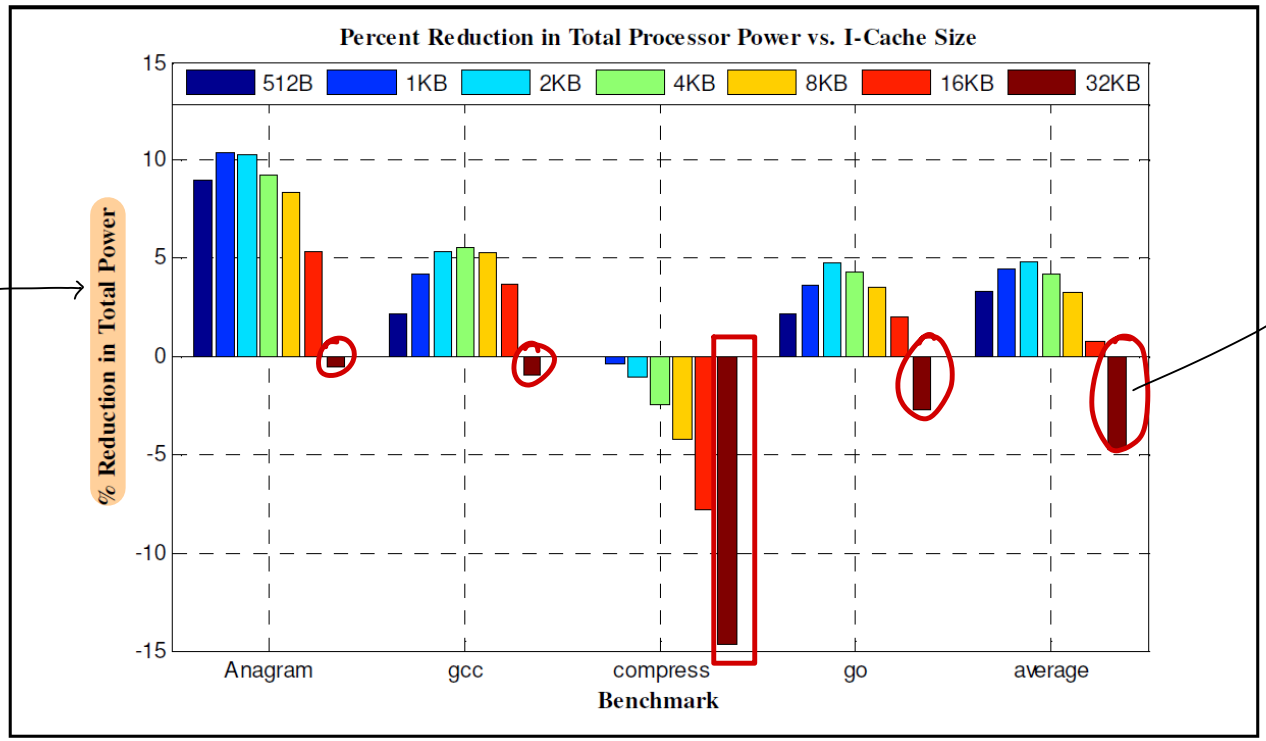
Figure 4: Percent increase in I-Cache power as cache size increases.



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

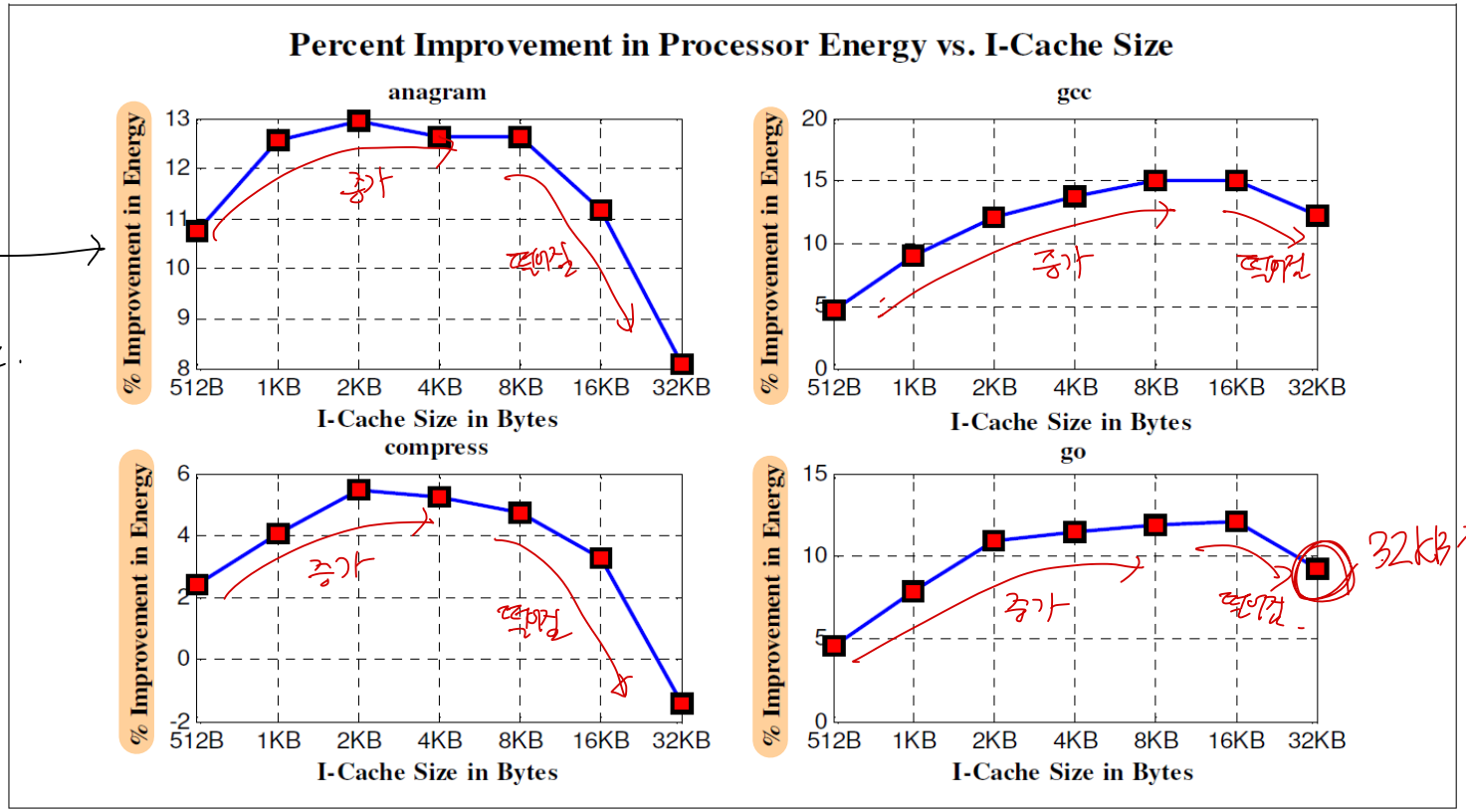
- **Survey paper by '*Mutaz Al-Tarawneh*'**
  - **Size versus energy**

power는 시간당 소모되는 에너지

캐시 메모리가 커지면서 Miss rate이 감소하여 소비되는 에너지감소 → 캐시 메모리가 커지면서 증가하는 에너지 소모

↳ 결론: 에너지 감소율이 적어짐.

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

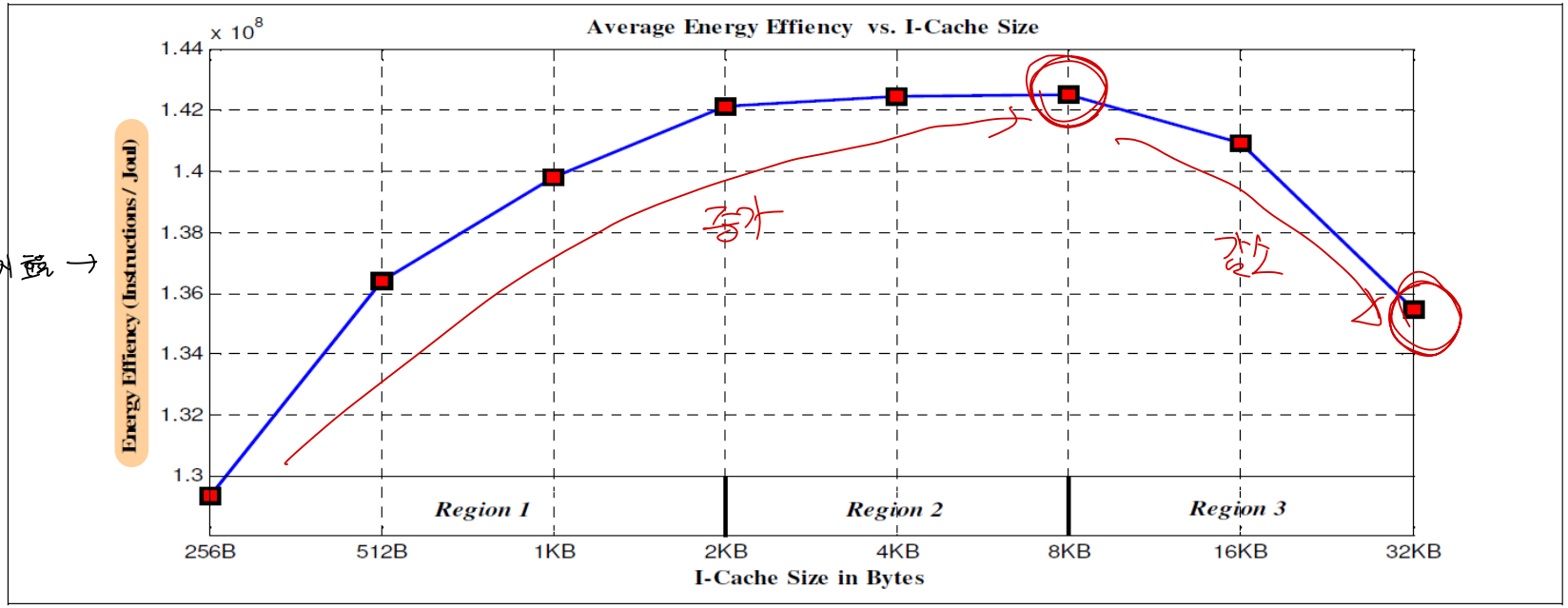


Issues of Cache Memory – L1 Cache Size

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

이러한 특징 때문에 L1 cache memory는 32KB까지만 묶여서 사용하는 것임.

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



Issues of Cache Memory – L1 Cache Size

- **Questions about commercial CPU products** 왜 embedded는 L1 캐시가 32KB보다 클 것인가?
  - 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?

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	2019.	7			2 x **Cheetah (M4)** for high performance		2.73 GHz	ARM	8	Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
							2 x **Cortex-A75** for					Per-core: I-64KB, D-64KB	Per-core: 256KB	
												Per-core: I-32KB, D-32KB	Per-core: 128KB	

Group#1

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	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

No. of Cores	L1 Cache	L2 Cache	L3 Cache
2	Per-core: I-32KB, D-32KB	Per-core: 256KB	3MB shared by 2 cores and GPU
4	Per-core: I-32KB, D-48KB	Per-core: 512KB	8MB shared by 4 cores and GPU
4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores and GPU
8	Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores and GPU
4	Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores
4	Per-core: I-32KB, D-32KB	Per-core: 256KB	10MB shared by 4 cores
16	Per-core: I-32KB, D-32KB	Per-core: 1MB	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.
    - 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.
    - 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.**

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

→ 두가지 embedded CPU가 이상함. ⇒ per-core L2를 가지고 있음에도 불구하고 L1 cache가 I-64KB임.

  - However, the following two embedded CPUs are strange.

→ workstation 등 메모리 비교 있음.

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	

공정을 믿고 I-64KB를 한 것임
성능 향상은 별로 없을 것임.
(오버 스펙인 것이다)

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#1

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	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

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
							4 x **Meerkat (M3)** for		2.9 GHz	ARM		Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
												Per-core: I-32KB, D-32KB	256KB Shared	
												Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
												Per-core: I-64KB, D-64KB	Per-core: 256KB	
												Per-core: I-32KB, D-32KB	Per-core: 128KB	
												Per-core: I-32KB, D-32KB	Per-core: 256KB	3MB shared by 2 cores and GPU
												Per-core: I-32KB, D-48KB	Per-core: 512KB	8MB shared by 4 cores and GPU
												Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores and GPU
												Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores and GPU
												Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores
												Per-core: I-32KB, D-32KB	Per-core: 256KB	10MB shared by 4 cores
												Per-core: I-32KB, D-32KB	Per-core: 1MB	22MB shared by 16 cores

Issues of Cache Memory
- Separate Cache versus Unified Cache

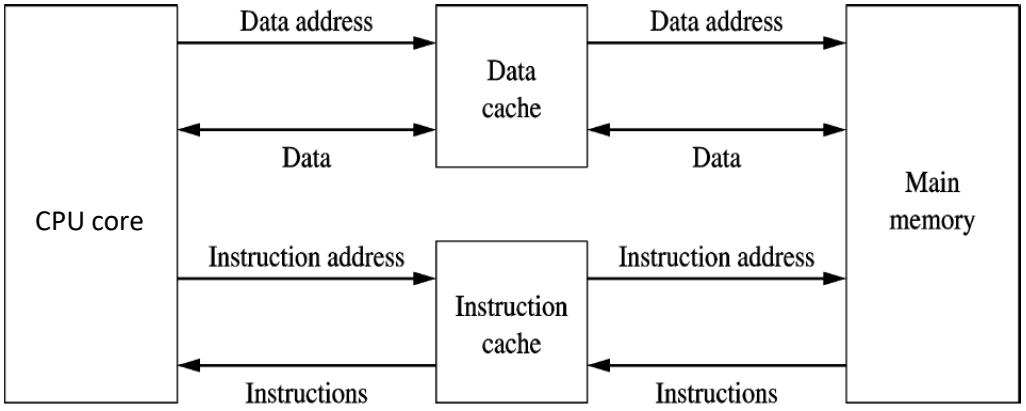
- **Initial cache memory designs used unified caches.**

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

Unified cache memory

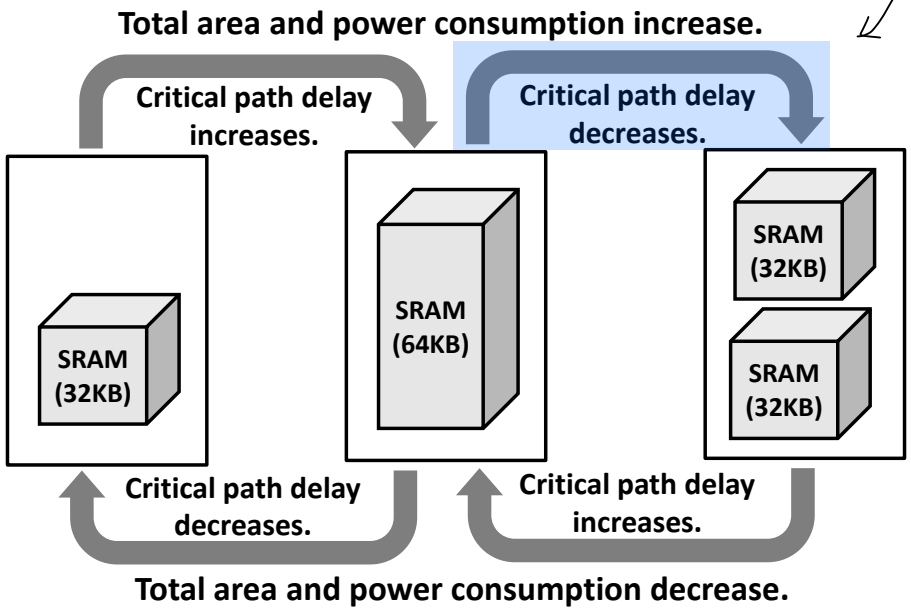
IBM

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



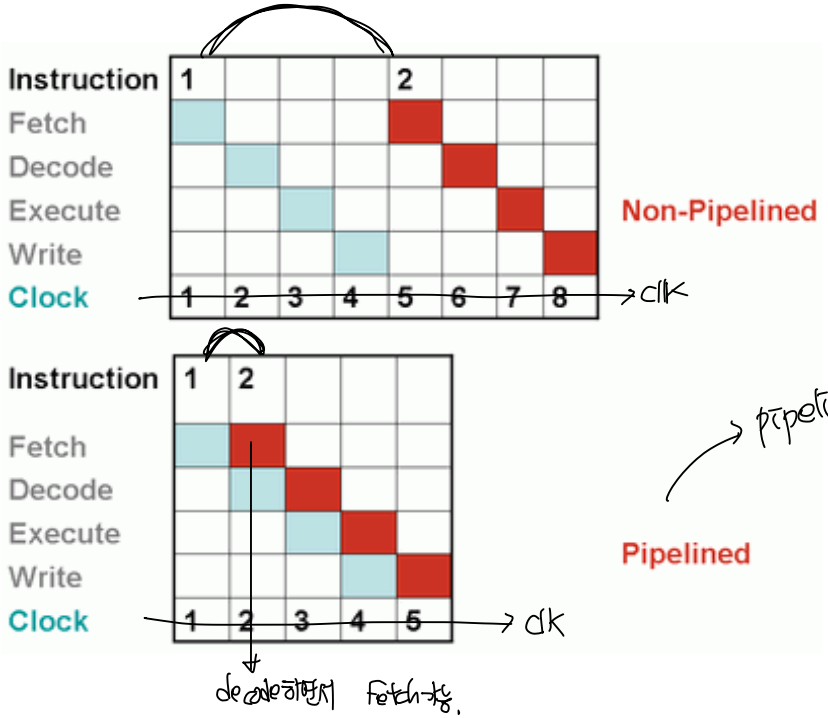
Issues of Cache Memory
- Separate Cache versus Unified Cache

- **Advantage of separate cache memory.**
  - It may offers faster access than unified 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

- **Advantage of separate cache memory.**
  - Consecutive instruction-fetch (pipelining) for speedup is possible.
    - pipelining is only available under independent instruction-cache memory without confliction with data-load/store operation.
    - 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.



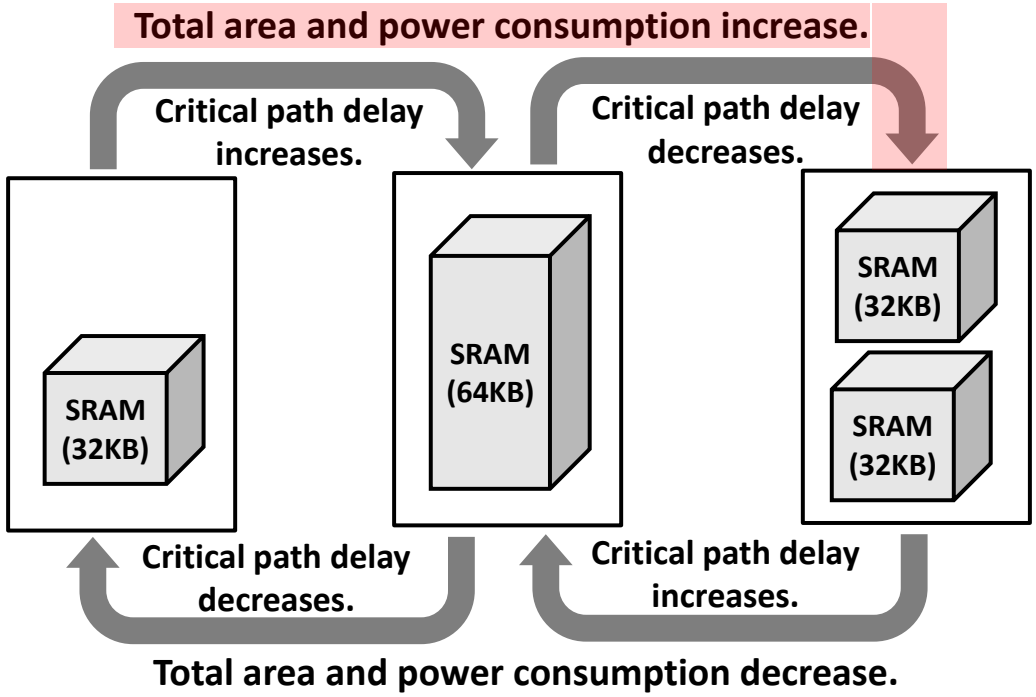
② 명령어를 instruction (명령어)를 fetch 해올 수 있다.
⇒ 그것으로 인해 speed up이 가능하다.
↳ I-cache를 독립적으로 두어야 pipelining이 가능.
↳ 상용화된 CPU는 pipelining으로 작동.
2개의 명령어를 실행하는데
⇒ 8개 cycle
3개의 cycle차이 발생.
↳ pipelined이 가능한 것.
⇒ 5개 cycle
이게 가능하려면
Instruction cache가 따로 있어야함.
⇒ 연속적으로 명령이 오려면 독립적으로 와야함
decode하면서 fetch가능.

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 seperate cache는 면적과 전력소모가 증가한다.



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.

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

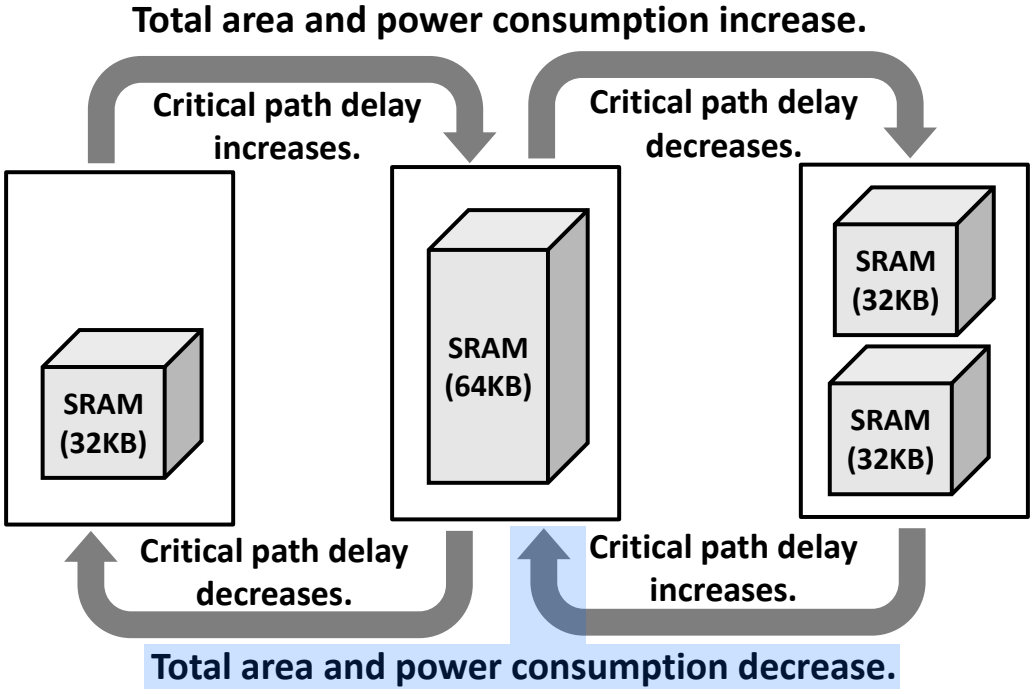
Size	Instruction cache	Data cache	Unified cache
1 KB	3.06%	24.61%	13.34%
2 KB	2.26%	20.57%	9.78%
4 KB	1.78%	15.94%	7.24%
8 KB	1.10%	10.19%	4.57%
16 KB	0.64%	6.47%	2.87%
32 KB	0.39%	4.82%	1.99%
64 KB	0.15%	3.77%	1.35%
128 KB	0.02%	2.88%	0.95%

Issues of Cache Memory
- Separate Cache versus Unified Cache

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 전력소모가 감소한다.



Issues of Cache Memory
- Separate Cache versus Unified Cache

- **Advantage of unified cache memory.**
  - 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.
    - Therefore, entire miss-rate is lower than separate cache memory.

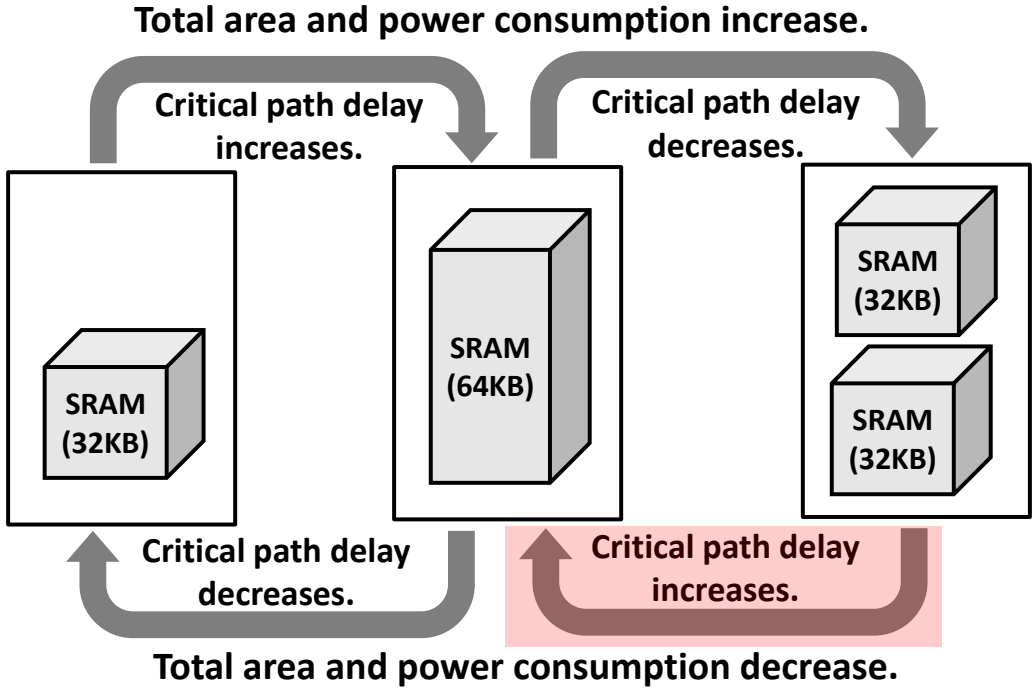
Miss rates for instruction, data, and unified caches of different sizes.

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

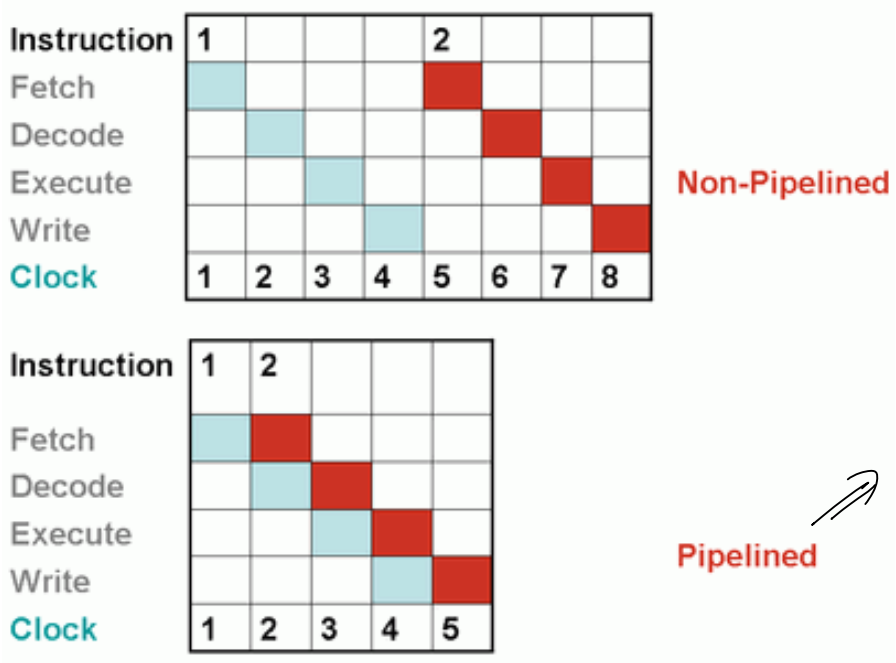
Disadvantage of unified cache memory. ① 장점이 느려짐.

- 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.



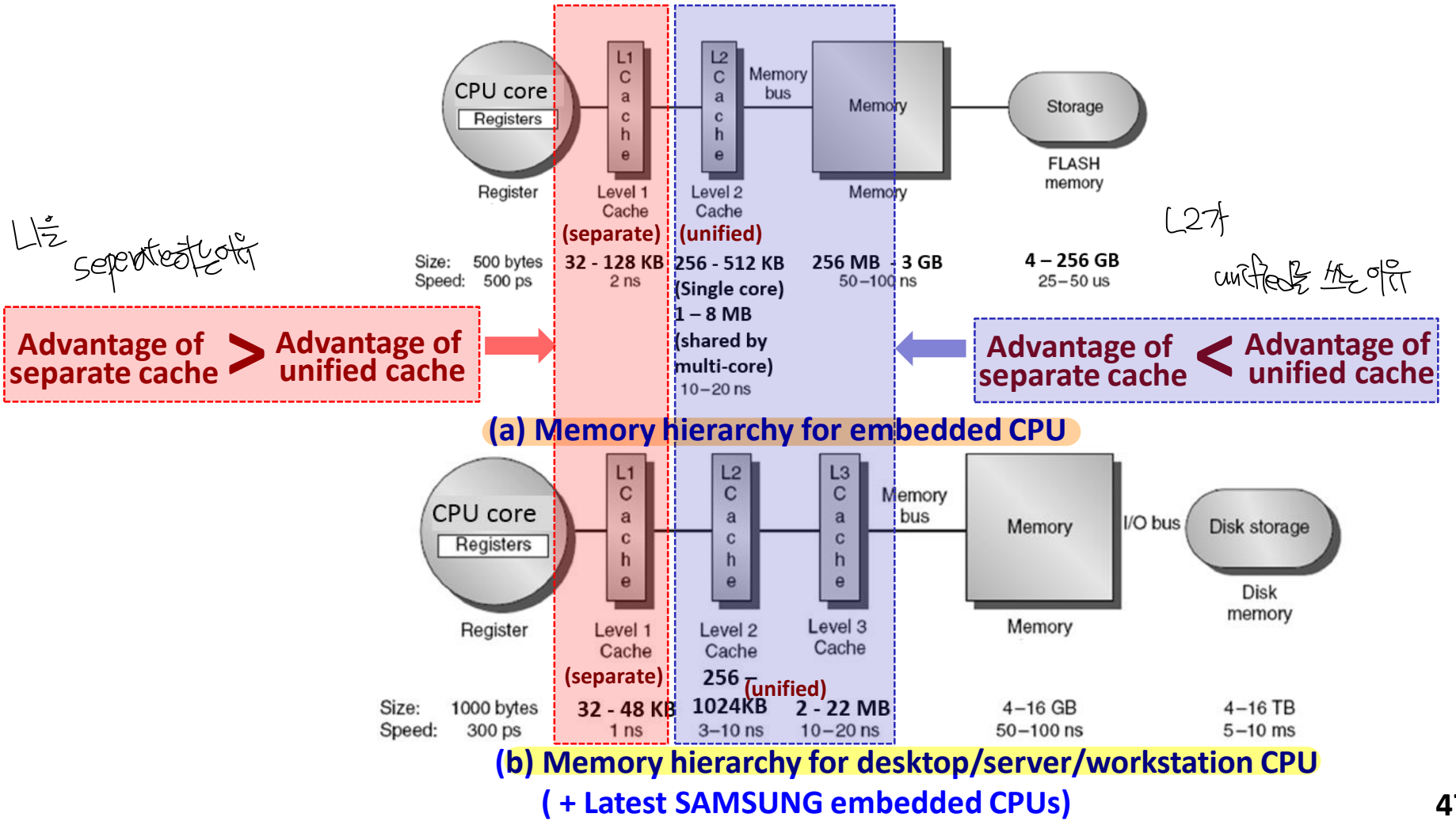
Issues of Cache Memory
- Separate Cache versus Unified Cache

- **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**



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?
  - Answer#4
    - Some caches : *merit of private cache > merit of shared cache*
    - Other caches : *merit of private cache < merit of shared cache*

↳ private 과 shared의 차이에 대해 답하시오.

↳ (private / shared) 의 장단점이 무엇인지 파악할 것.

Group#2

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators												
				Fab. Tech	Power	Name	CPU					L1 Cache	L2 Cache	L3 Cache		
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores					
			2018.	10			4 x Meerkat (M3) for high performance		2.9 GHz	ARM v8-A		Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared		
												Per-core: I-32KB, D-32KB	256KB Shared			

Group#1

Type	Sub Type	Product Name	Release date (Year. Month)	SoC (System-on-Chip) = CPU + GPU + HW Accelerators												
				Fab. Tech	Power	Name	CPU					L1 Cache	L2 Cache	L3 Cache		
							Micro-architecture	Bit-Width	Clock Freq.	ISA	No. of Cores					
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	Per-core: I-64KB, D-64KB	Per-core: 512KB
		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
		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
		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	Per-core: I-32KB, D-32KB	Per-core: 256KB
							2 x Zephyr for energy efficiency					Per-core: I-32KB, D-32KB	1MB shared		Per-core: I-32KB, D-48KB	Per-core: 512KB
		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	Per-core: I-32KB, D-32KB	Per-core: 256KB
							4 x Cortex-A53 for energy efficiency		1. 7 GHz			Per-core: I-32KB, D-32KB	256KB shared		Per-core: I-32KB, D-32KB	Per-core: 256KB
		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	Per-core: I-32KB, D-32KB	Per-core: 256KB
							4 x Mistral for energy efficiency		1.42 GHz			Per-core: I-32KB, D-32KB	1MB shared		Per-core: I-32KB, D-32KB	Per-core: 256KB
		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		

L1 Cache	L2 Cache	L3 Cache
Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
Per-core: I-32KB, D-32KB	256KB Shared	
Per-core: I-64KB, D-64KB	Per-core: 512KB	4MB shared
Per-core: I-64KB, D-64KB	Per-core: 256KB	
Per-core: I-32KB, D-32KB	Per-core: 128KB	
Per-core: I-32KB, D-32KB	Per-core: 256KB	3MB shared by 2 cores and GPU
Per-core: I-32KB, D-48KB	Per-core: 512KB	8MB shared by 4 cores and GPU
Per-core: I-32KB, D-32KB	Per-core: 256KB	3MB shared by 4 cores and GPU
Per-core: I-32KB, D-32KB	Per-core: 256KB	12MB shared by 4 cores and GPU
Per-core: I-32KB, D-32KB	Per-core: 256KB	8MB shared by 4 cores
Per-core: I-32KB, D-32KB	Per-core: 256KB	10MB shared by 4 cores
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

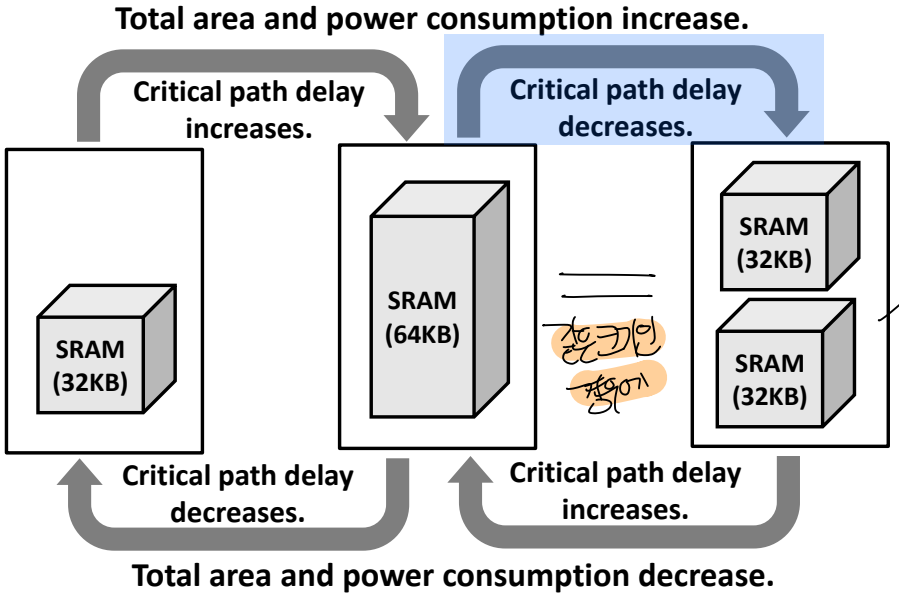
Advantage of private cache memory

① private 캐시가 더 빠른 접근이 가능하다.

- 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 .

Seperate 캐시나 ⊕ 라는 점. (라는 장점이 shared 더 많은)

↳ shared의 경우 여러 코어를 접근할지 결정하는 delay cycle이 있는데 private의 경우 이러한 delay가 없음. (많은 core를 처리 할 때)



Issues of Cache Memory

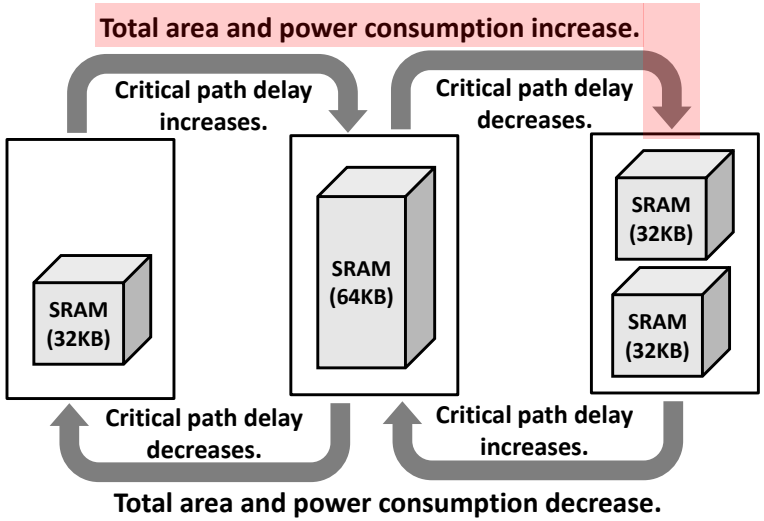
- Private Cache versus Shared Cache for Multi-Core

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

✓ **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.

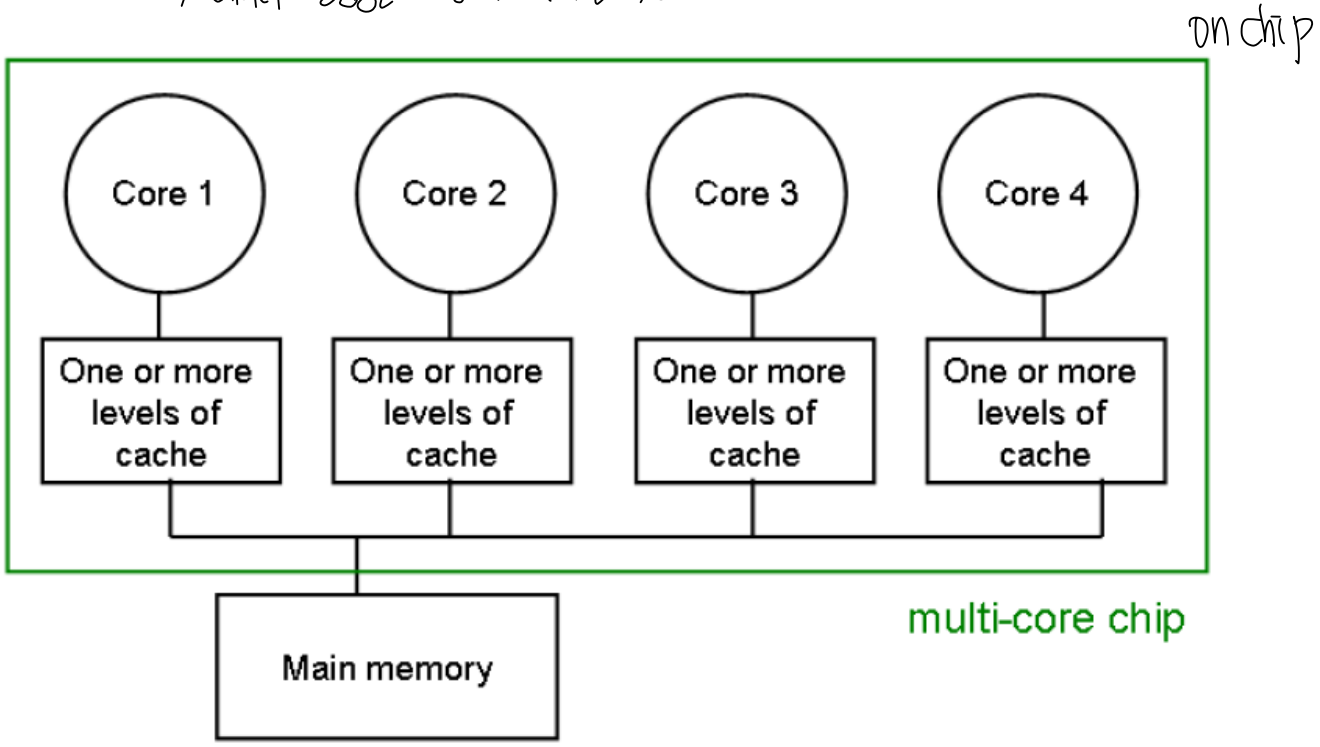


Issues of Cache Memory

- 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?

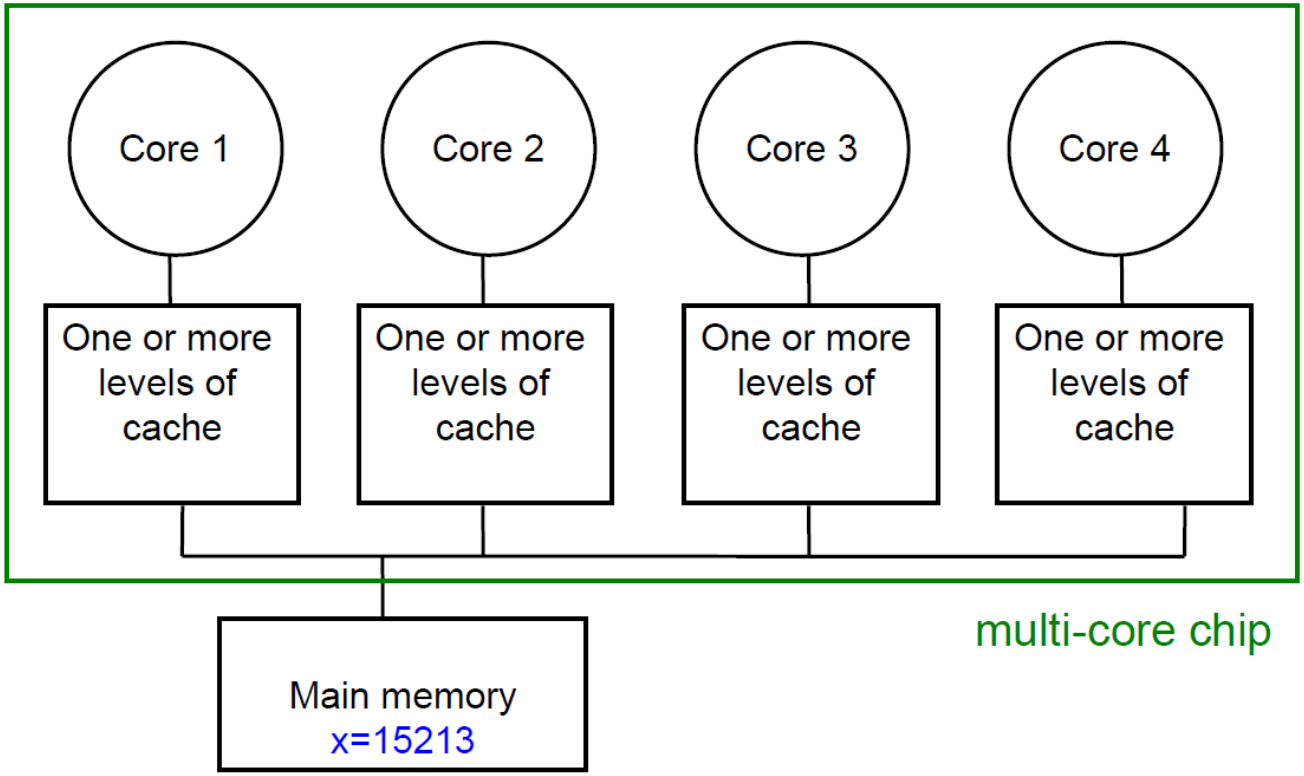
⇒ 데이터 일관성을 어떻게 유지할 것인가.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

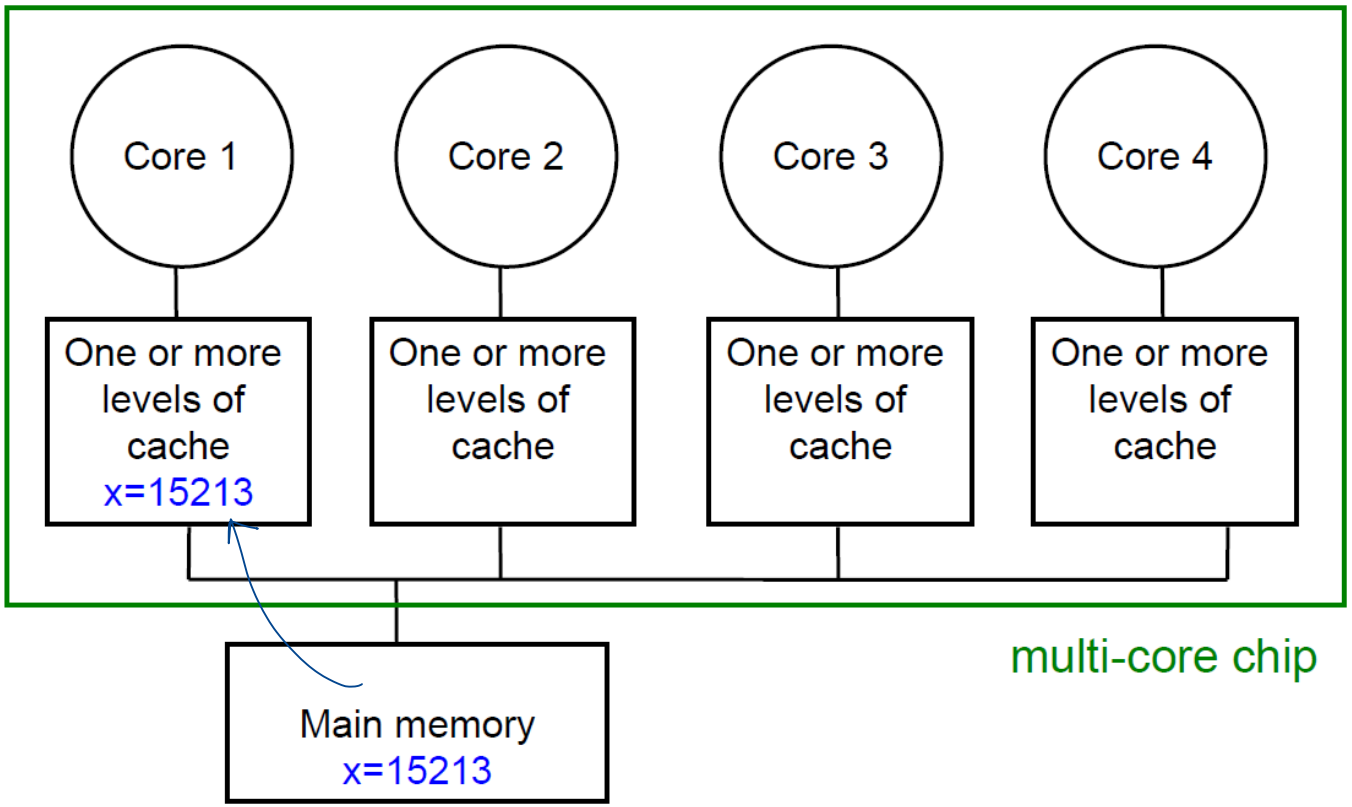
- **Cache coherence problem**
  - Suppose variable x initially contains 15213.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

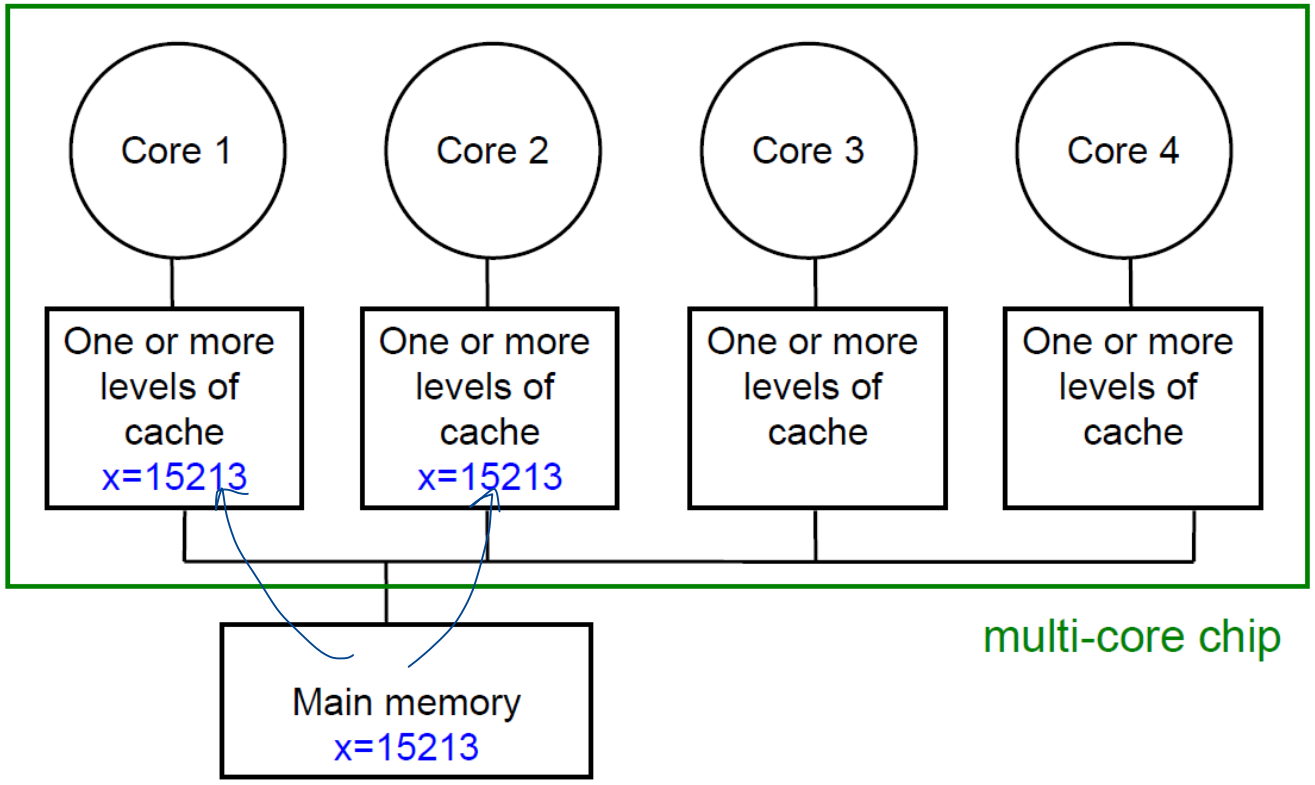
- **Cache coherence problem**
  - Core 1 reads x.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

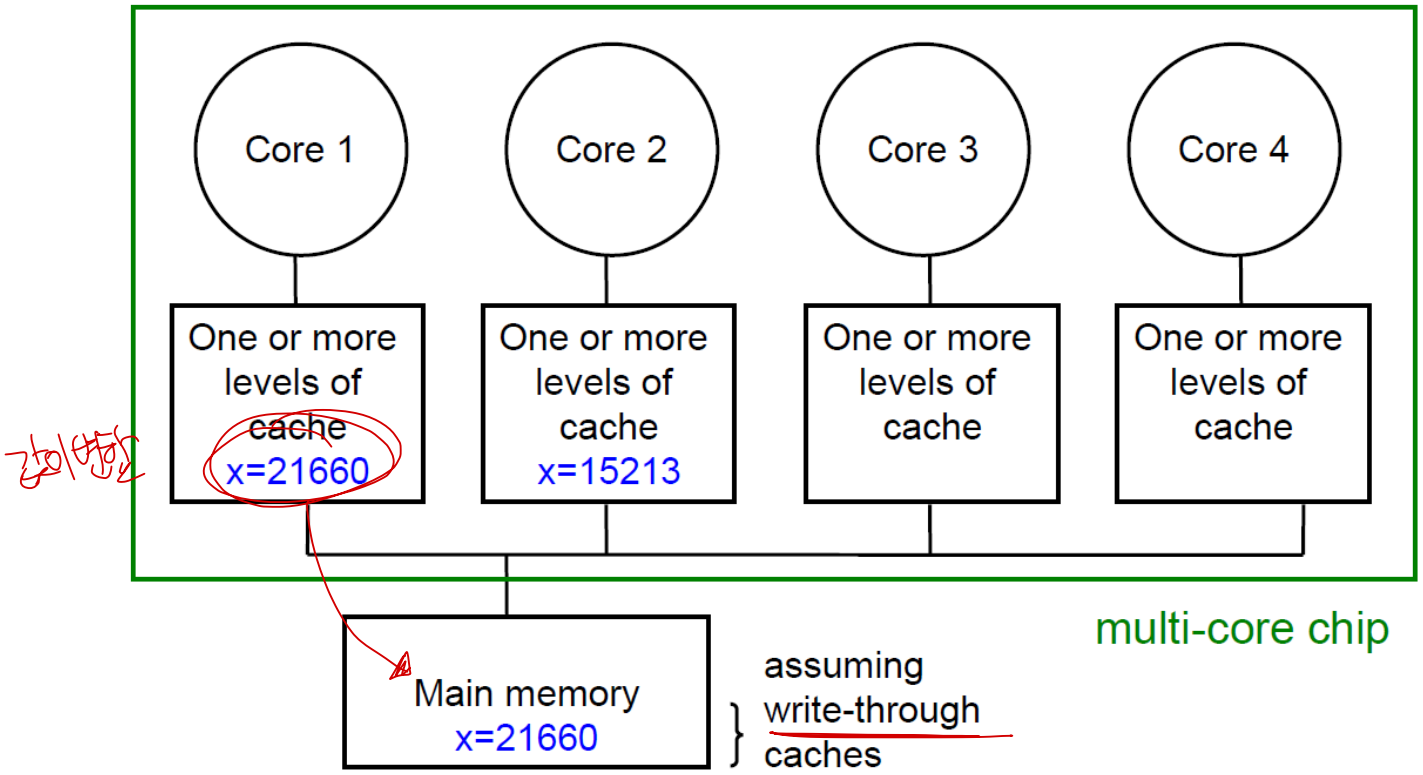
- **Cache coherence problem**
  - Core 2 reads x.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

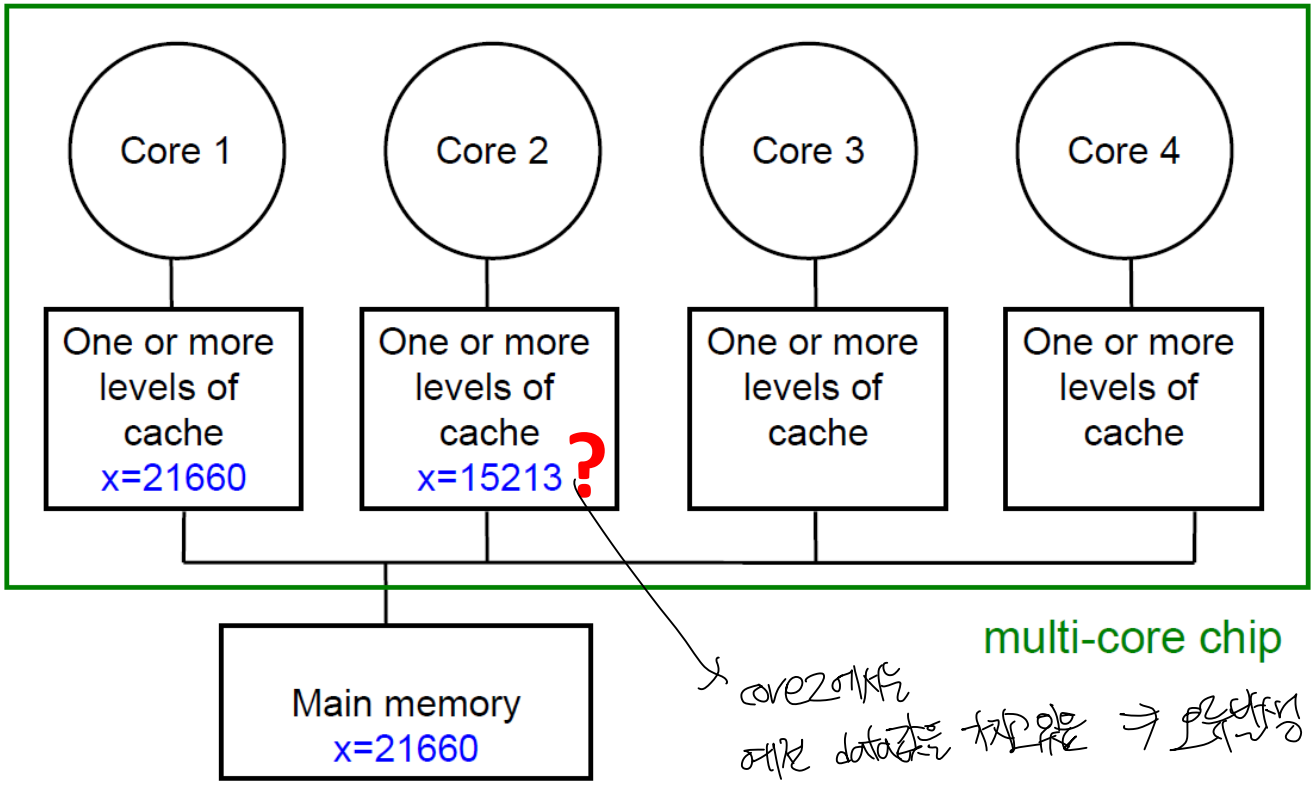
- **Cache coherence problem**
  - Core 1 writes to x, setting it to 21660.



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

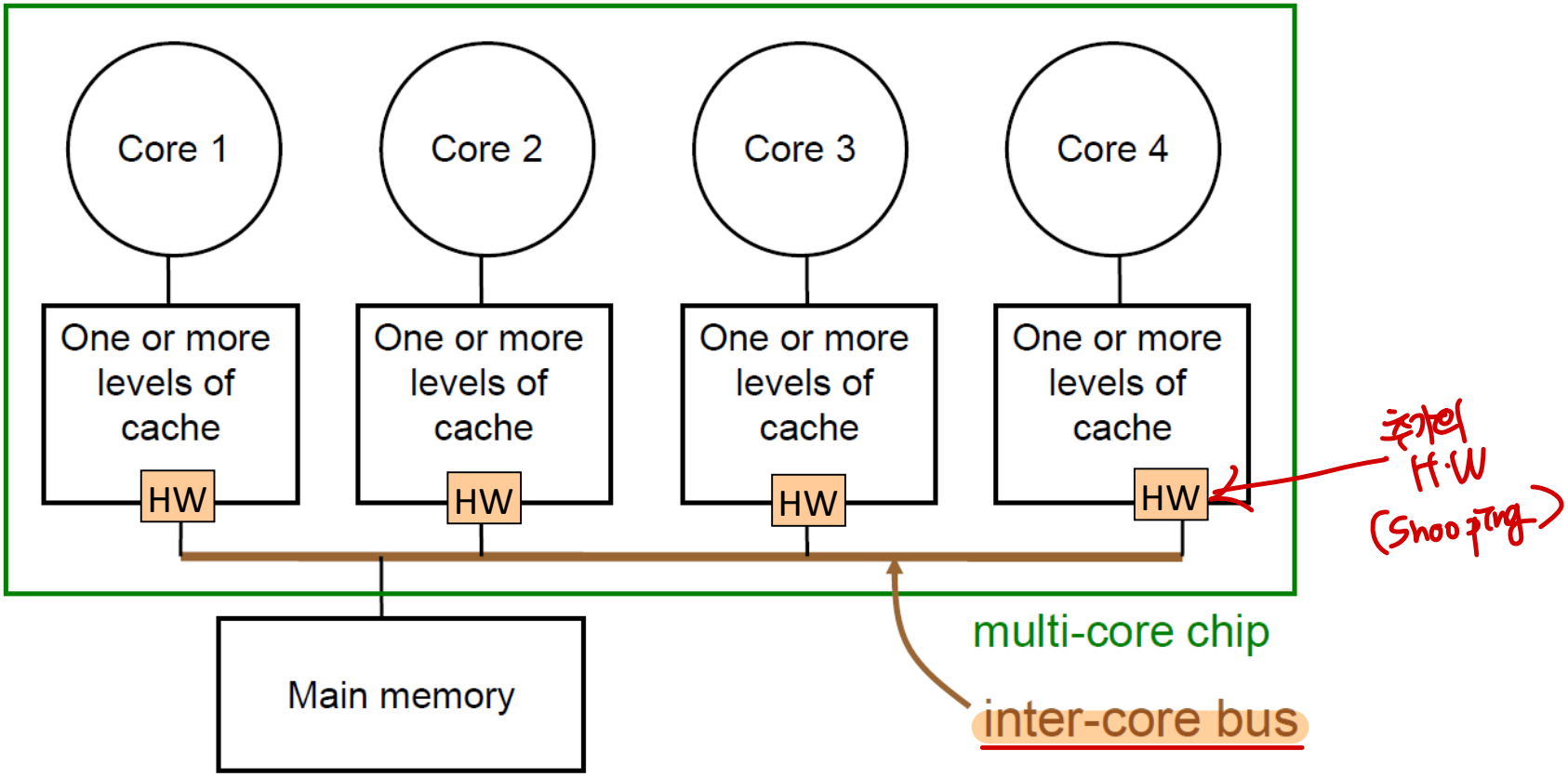
- Private Cache versus Shared Cache for Multi-Core

- **Solutions for cache coherence problem**
  - **Additional hardware** is needed to support **invalidation protocol with snooping**.
  - **Invalidation**
    - If a core write to a data, all other copies of this data in other caches are invalidated.
  - **Snooping**
    - All cores continuously "snoop" (monitor) the bus connecting the cores.

Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

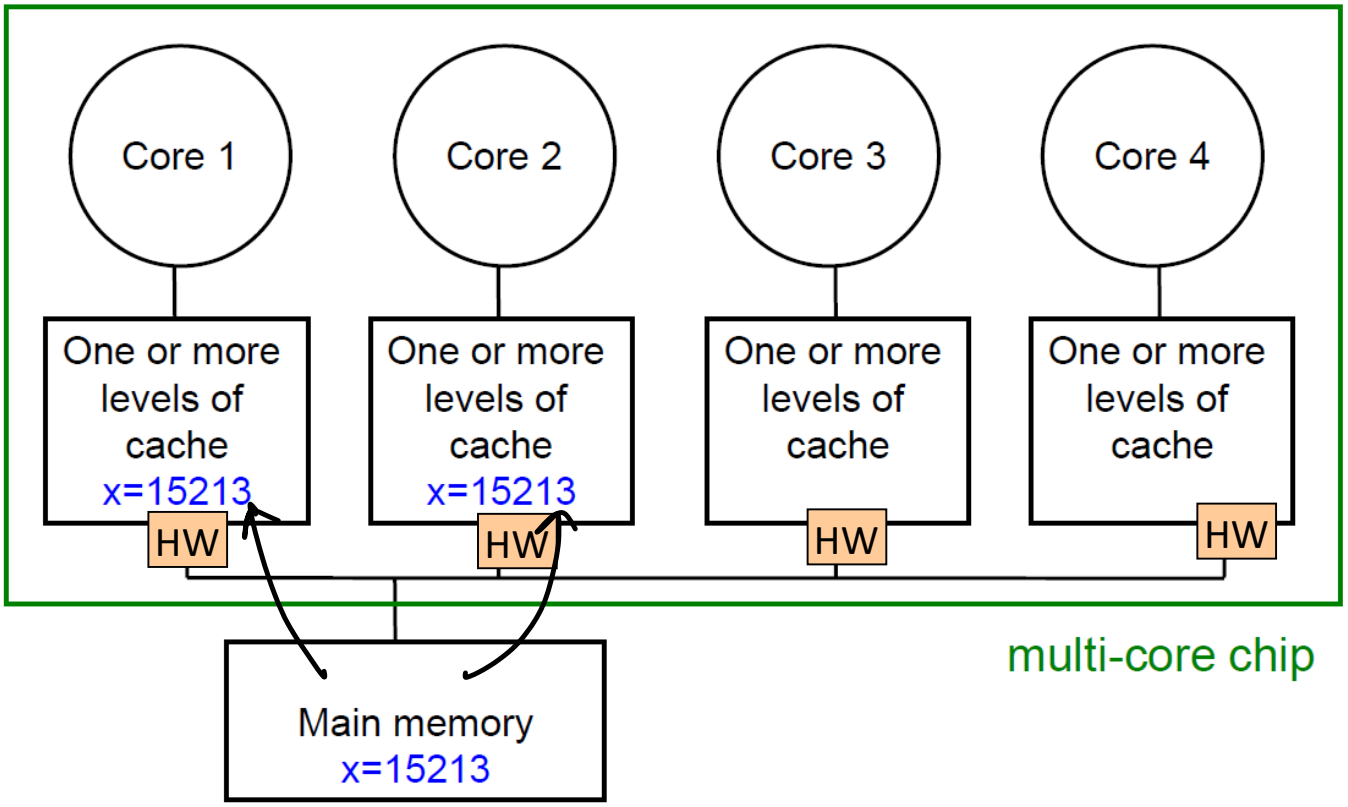
- **Bus based "Snooping" multi-core processor**



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

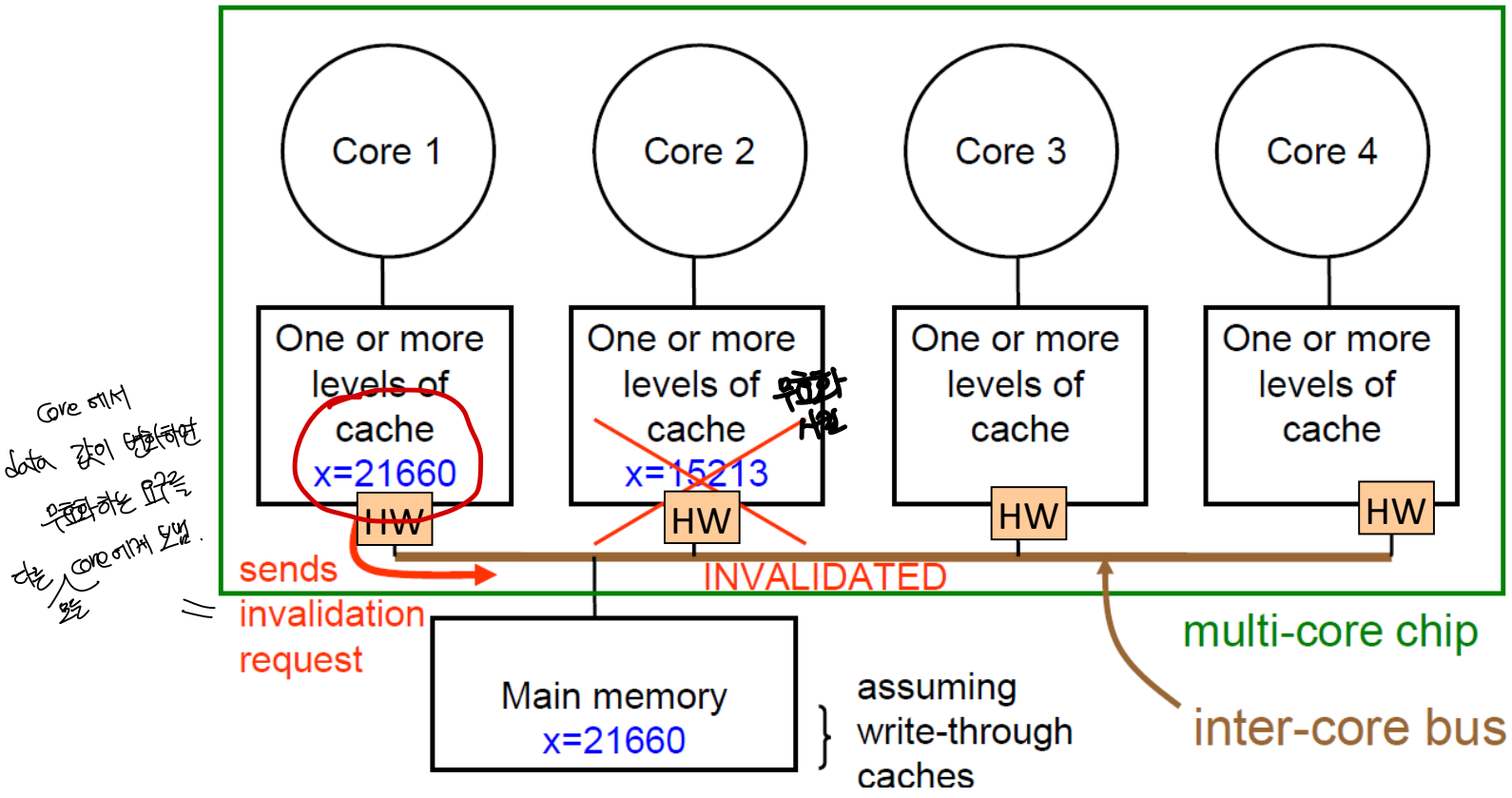
- **Invalidation based cache coherence protocol**
  - Cores 1 and 2 have both read x.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

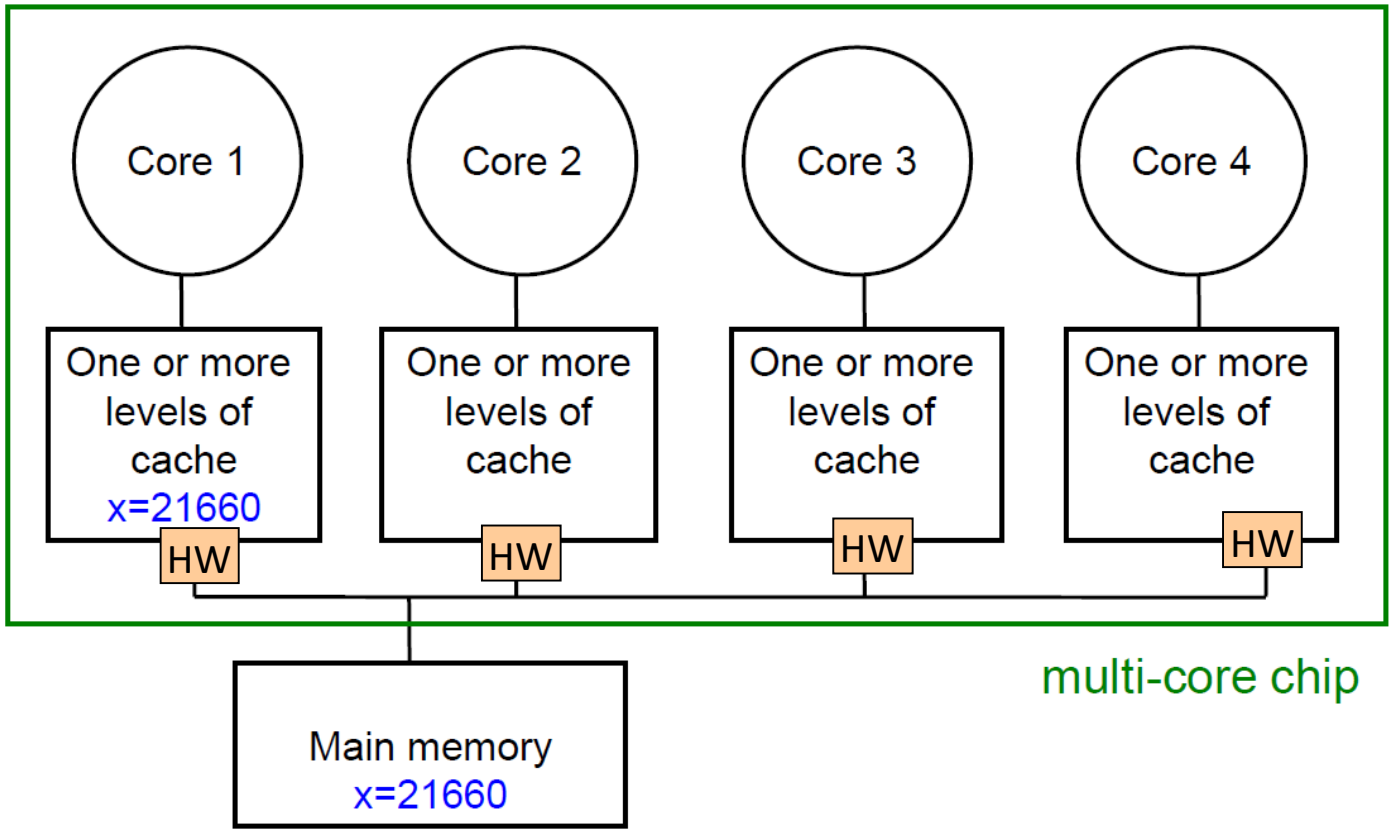
- **Invalidation based cache coherence protocol**
  - Cores 1 writes to x, setting it to 21660.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

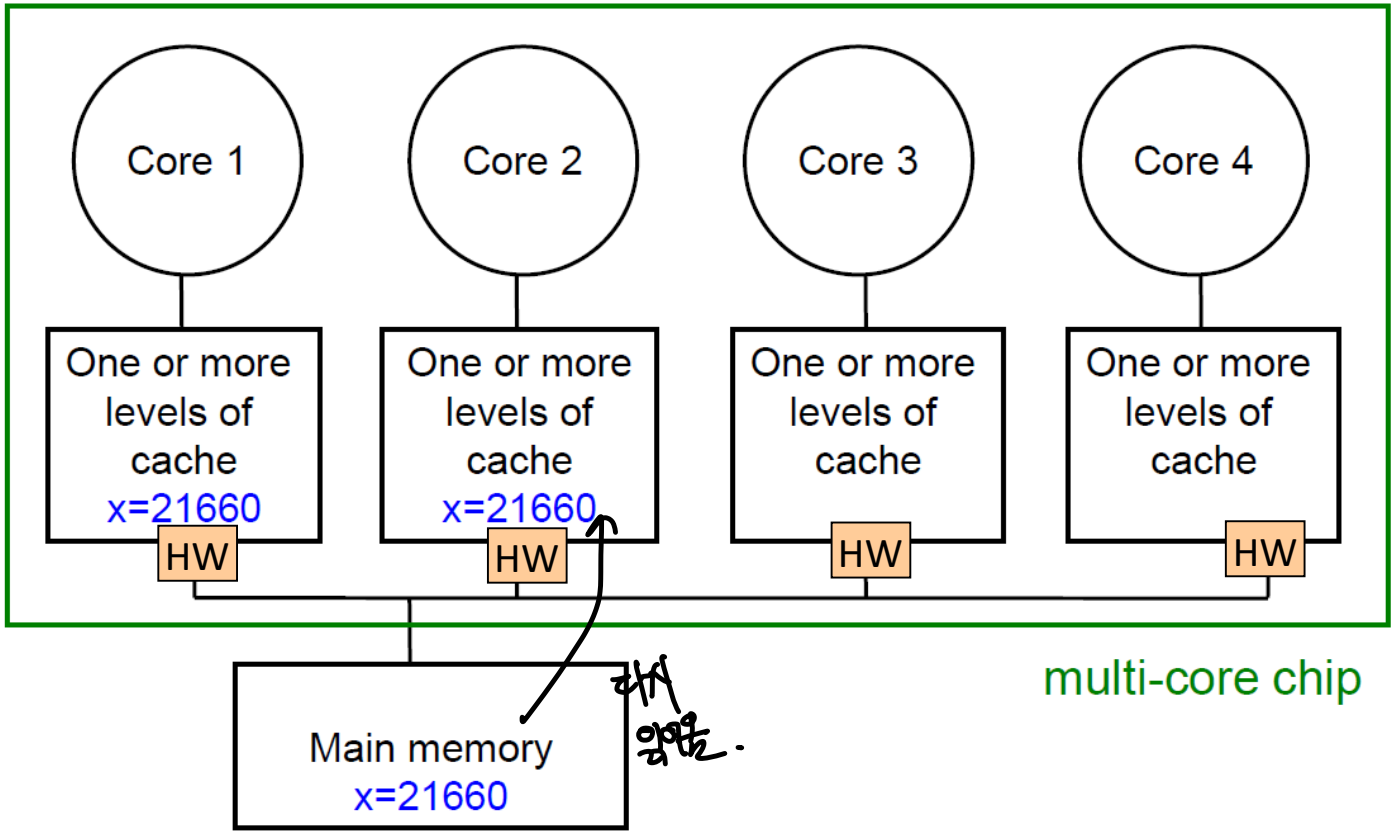
- **Invalidation based cache coherence protocol**
  - After invalidation



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.



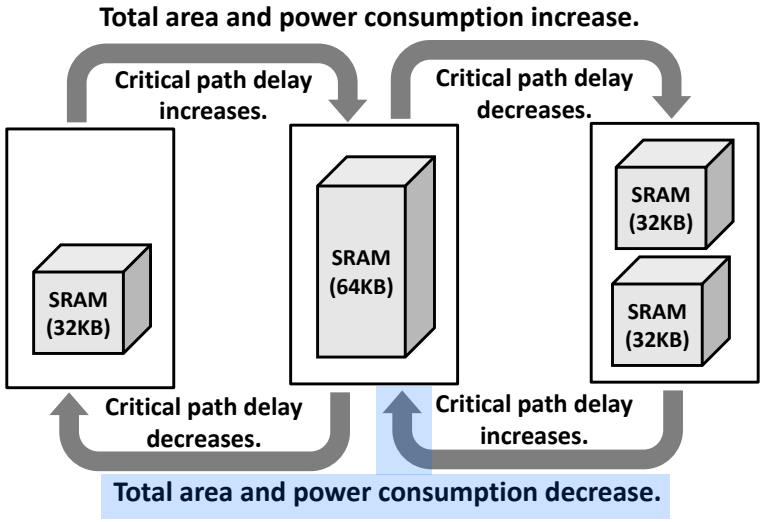
Issues of Cache Memory

- 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. ⇒ ① 총 면적과 전력소모가 적음.
- Threads on different cores can share the same cache data without **cache coherence problem**. ⇒ ② cache coherence problem을 고려할 필요가 없음.
- More cache space are available if a single (or a few) high-performance thread runs on the system.

↓
단일로 고성능 thread가 시스템에서 실행될 때, 더 많은 캐시 공간을 사용할 수 있다.

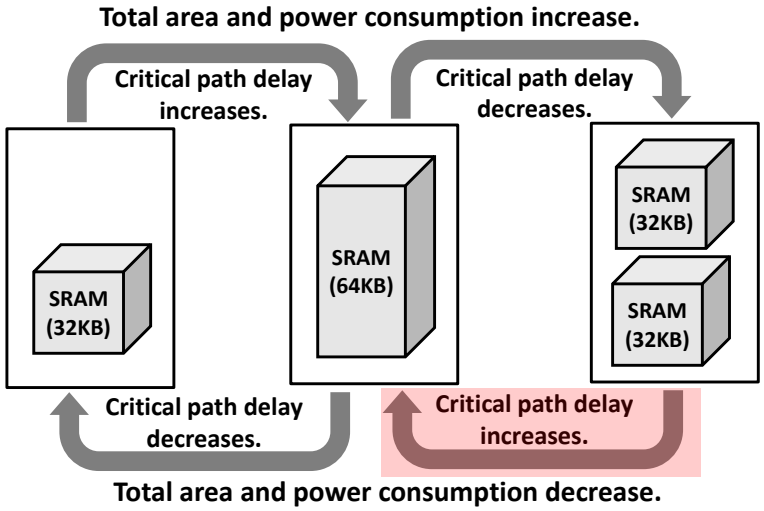


Issues of Cache Memory

- 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** ⇒ ① critical path delay가 증가
    - ✓ Small private caches show much shorter delay than a large shared cache.
  - **In the aspect of cycle count** ⇒ ② cycle count가 증가.
    - Private cache is free of delay cycles caused by contention among cores sharing a cache .
      ↳ 여러 core를 접근하지 못하도록 delay cycle이 발생함.



Issues of Cache Memory

- Private Cache versus Shared Cache for Multi-Core

- **Multi-level cache organization**