CS 110
Computer Architecture

Advanced Caches

Instructor:
Sören Schwertfeger and Chundong Wang

https://robotics.shanghaitech.edu.cn/courses/ca/21s

School of Information Science and Technology SIST

ShanghaiTech University

Slides based on UC Berkeley's CS61C (2015)

Review

- Networking
  - Connecting computers, and networks
- Use abstraction to cope with complexity of communication
- Hierarchy of layers:
  - Application (chat client, game, etc.)
  - Transport (TCP, UDP)
  - Network (IP)
  - Data Link Layer (Ethernet)
  - Physical Link (copper, wireless, etc.)

Protocol Family Concept

- *Protocol*: packet structure and control commands to manage communication
- *Protocol families (suites)*: a set of cooperating protocols that implement the network stack

- Key to protocol families is that communication occurs logically at the same level of the protocol, called peer-to-peer…

  …but is implemented via services at the next lower level

- Encapsulation: carry higher level information within lower level "envelope"

Inspiration…

- CEO A writes letter to CEO B
  - Folds letter and hands it to assistant
- Assistant
  - Puts letter in envelope with CEO B's full name
  - Takes to China Post
- China Post Office
  - Puts letter in larger envelope
  - Puts name and street address on China Post envelope
  - Puts package on China Post delivery truck
- China Post delivers to other company

Dear Jack,

6.18 is coming.

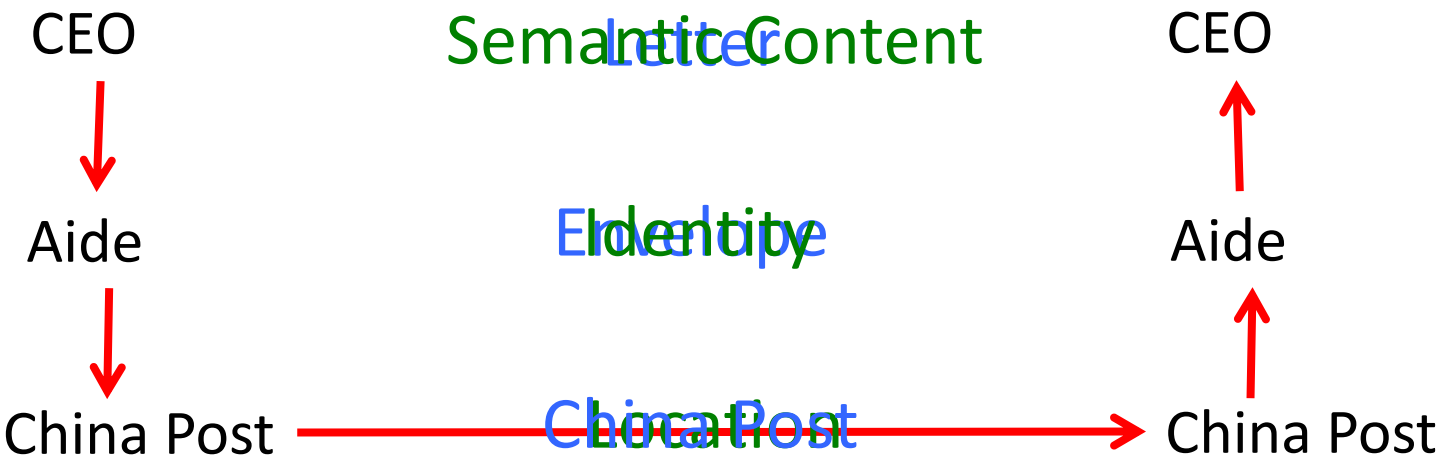
-Pony

The Path of the Letter

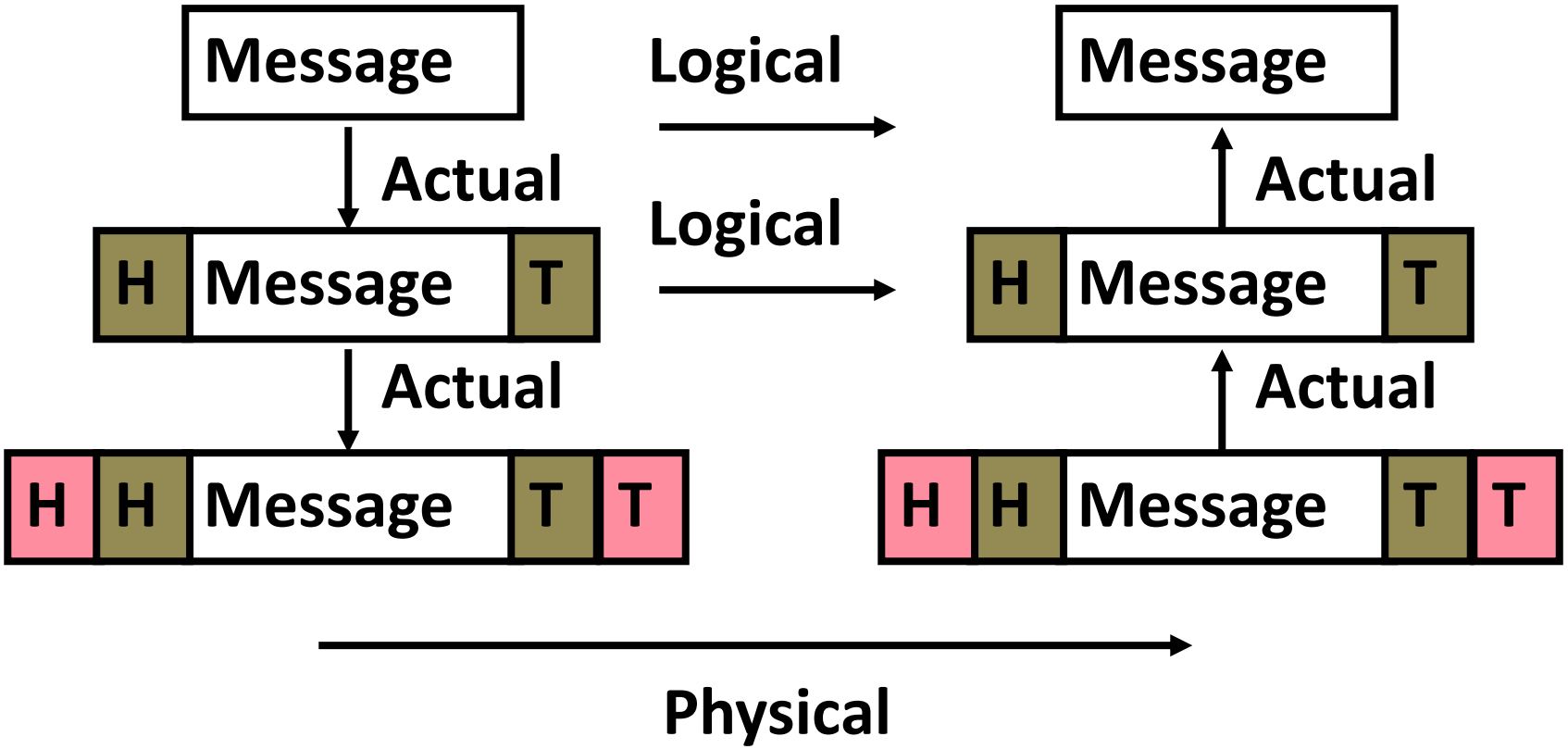
"Peers" on each side understand the same things

No one else needs to

Lowest level has most packaging



Protocol Family Concept



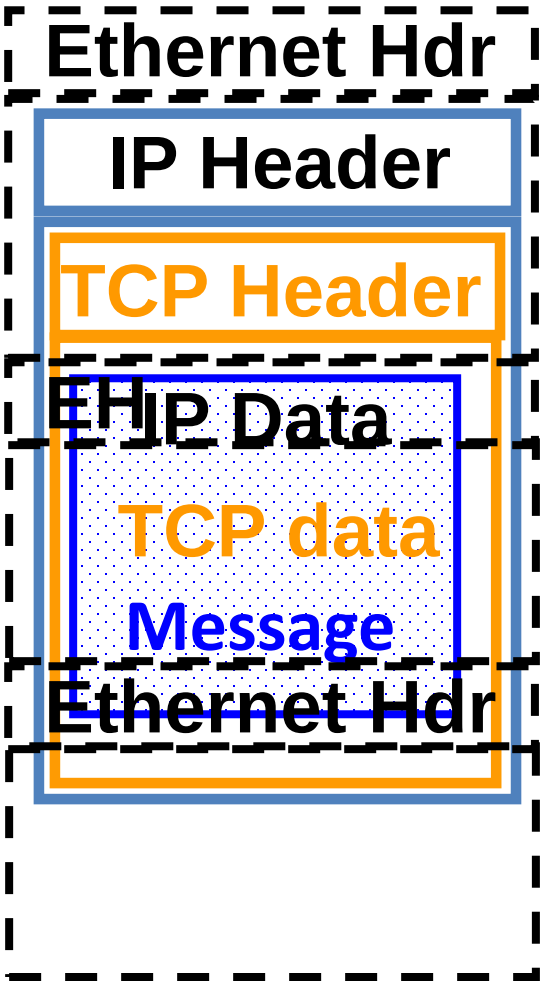
Each lower level of stack "encapsulates" information from layer above by adding header and trailer.

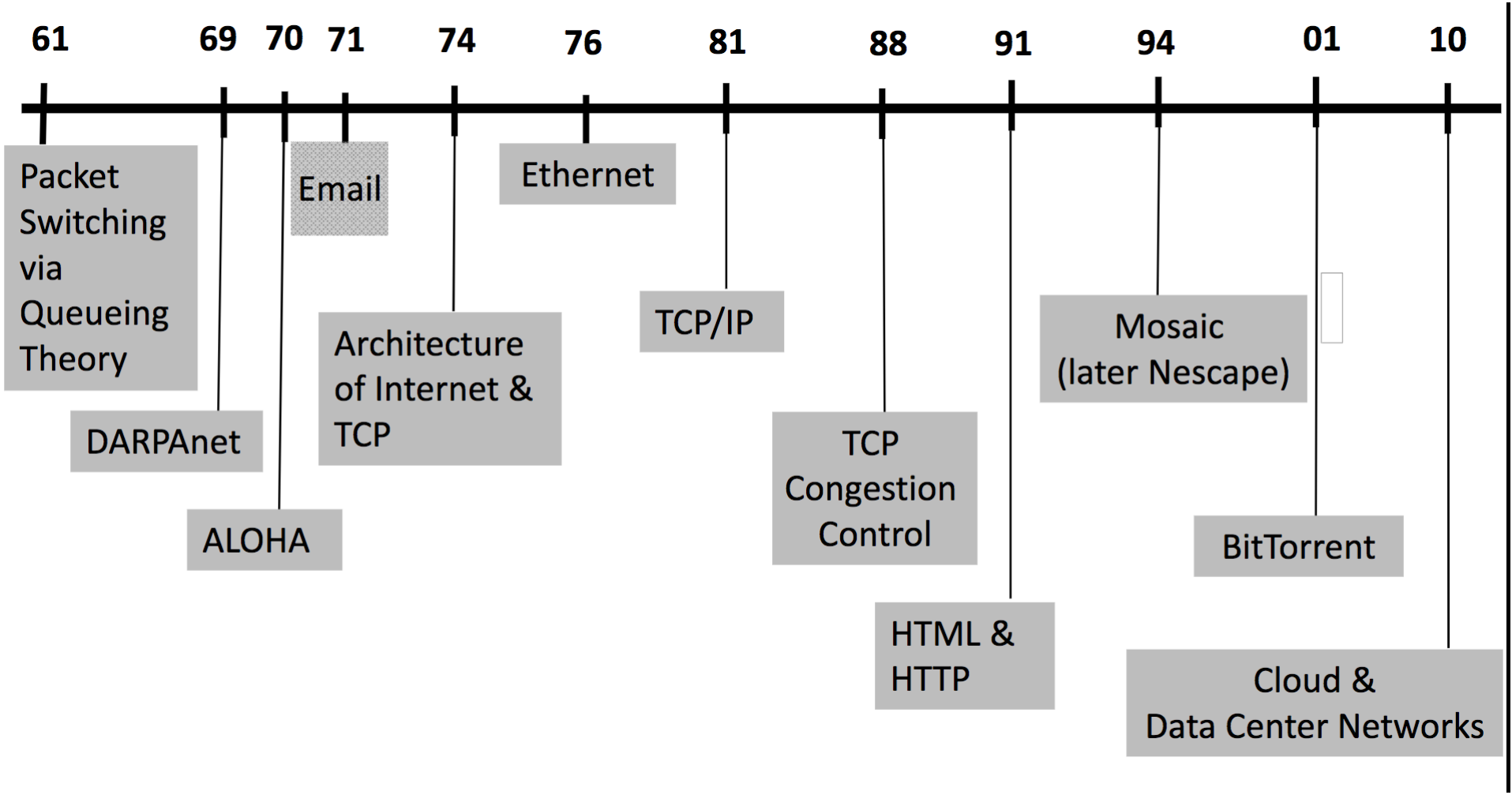
Most Popular Protocol for Network of Networks

- Transmission Control Protocol/Internet Protocol (TCP/IP)
- This protocol family is the basis of the Internet, a WAN (wide area network) protocol
  - IP makes best effort to deliver
    - Packets can be lost, corrupted
  - TCP guarantees delivery
  - TCP/IP so popular it is used even when communicating locally: even across homogeneous LAN (local area network)
  - UDP/IP: video or sound streaming; video call….

TCP/IP packet, Ethernet packet, protocols

- Application sends message
- **TCP breaks into 64KiB segments, adds 20B header**
- **IP adds 20B header, sends to network**
- **If Ethernet, broken into 1500B packets with headers, trailers**

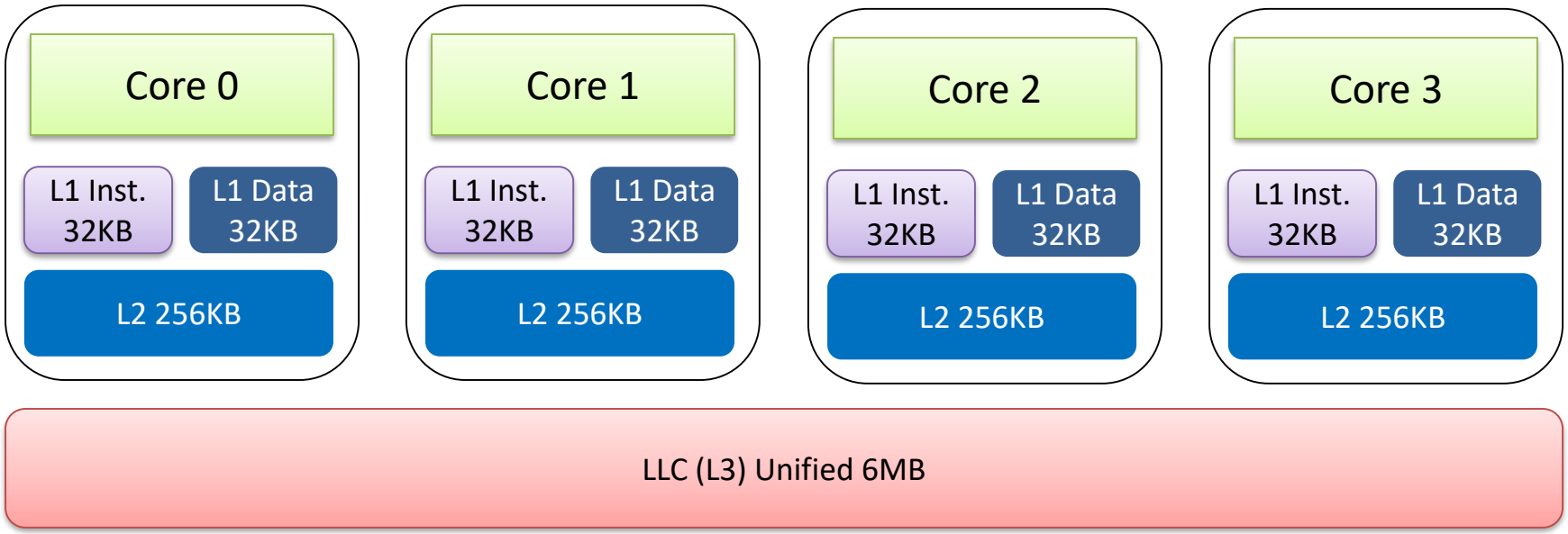



61
69
70
71
74
76
81
88
91
94
01
10
Packet Switching via Queueing Theory
DARPAnet
ALOHA
Email
Architecture of Internet & TCP
Ethernet
TCP/IP
TCP Congestion Control
HTML & HTTP
Mosaic (later Nescape)
BitTorrent
Cloud & Data Center Networks

Advanced Caches: MRU is LRU

Cache Inclusion

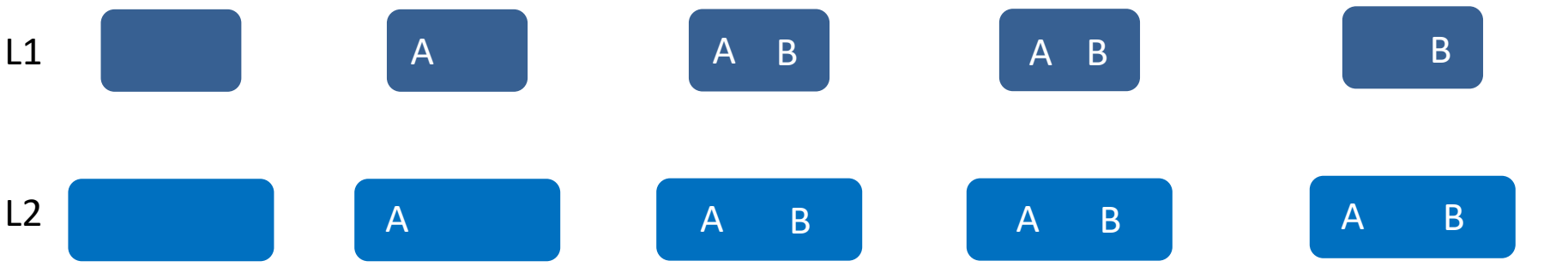
- Multilevel caches



Intel Ivy Bridge Cache Architecture (Core i5-3470)

If all blocks in the higher level cache are also present in the lower level cache, then the lower level cache is said to be **inclusive** of the higher level cache.

Inclusive



Initial state

Read A miss; load A into L1 and L2

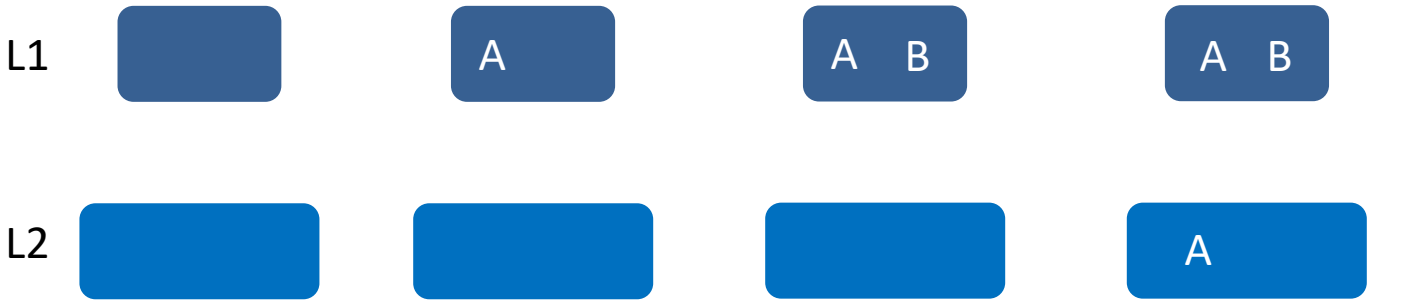
Read B miss; load B into L1 and L2

Evict A from L1 due to cache replacement

Evict B from L2 due to cache replacement

Back invalidation

Exclusive



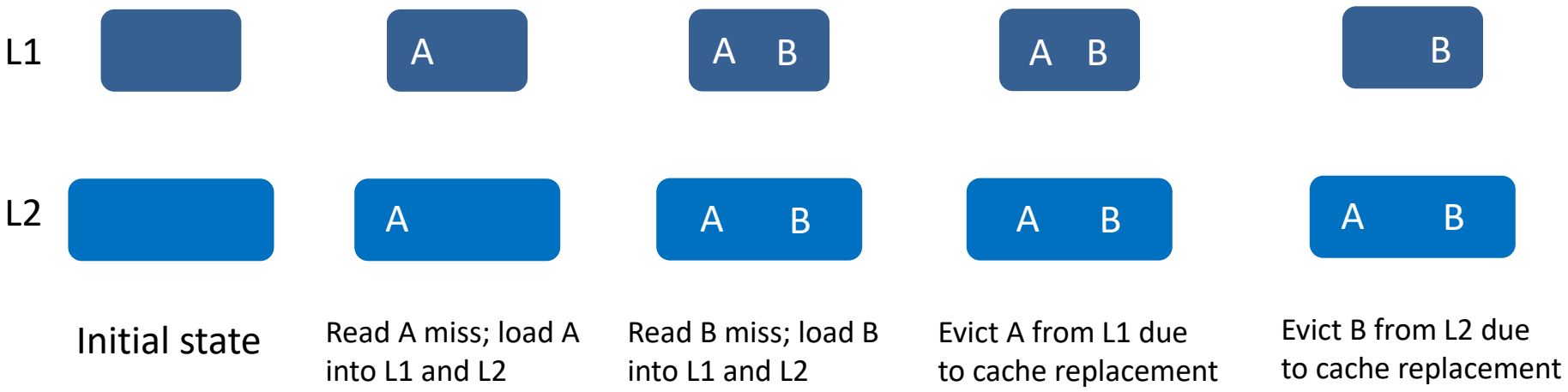
Initial state

Read A miss; load A into L1

Read B miss; load B into L1

Evict A from L1 due to cache replacement and place in L2

Non-inclusive



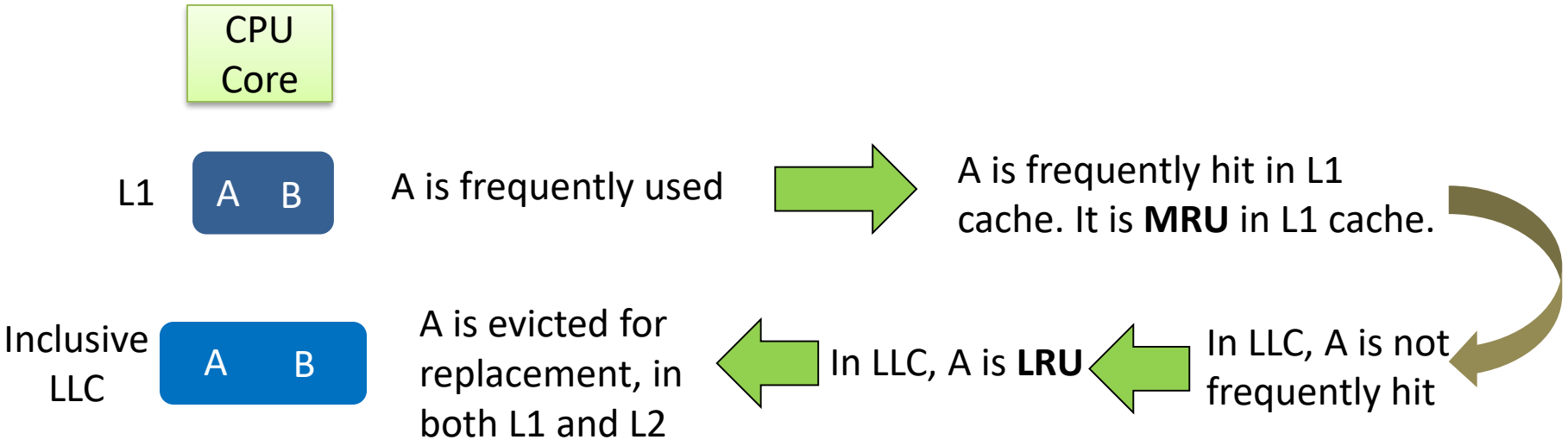
Real-world CPUs

- Intel Processors
  - Sandy bridge, inclusive
  - Haswell, inclusive
  - Skylake-S, inclusive
  - Skylake-X, non-inclusive
- ARM Processors
  - ARMv7, non-inclusive
  - ARMv8, non-inclusive
- AMD
  - K6, exclusive
  - Zen, inclusive
  - Shanghai, LLC non-inclusive

Inclusive, or not?

- Inclusive cache eases coherence
  - A cache block in a higher-level surely existing in lower-level(s)
  - A non-inclusive LLC, say L2 cache, which needs to evict a block, **must** ask L1 cache if it has the block, because such information is not present in LLC.
- Non-inclusive cache yields higher performance though, why?
  - No back invalidation
  - More data can be cached ← larger capacity

'*Sneaky*' LRU for Inclusive Cache



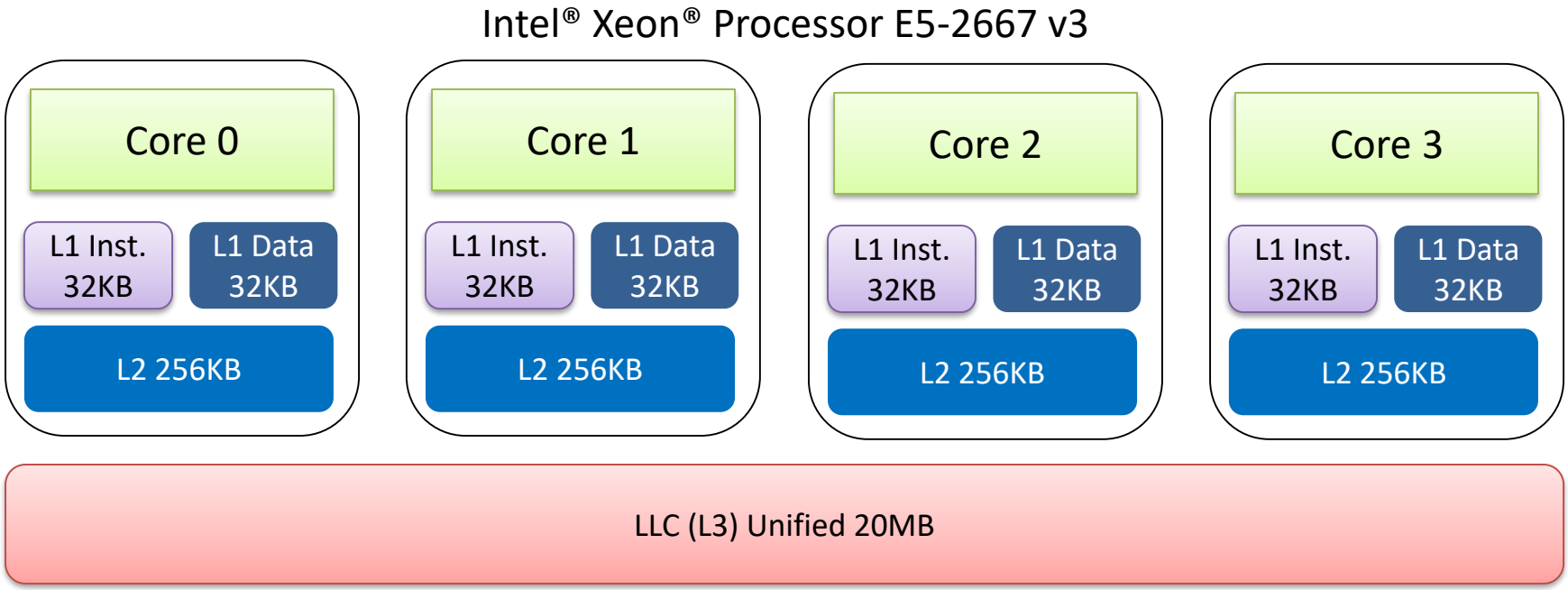
As a result, MRU block that should be retained might be evicted, which causes performance penalty.

What if LLC is non-inclusive?

Should you be interested, you can click *https://doi.org/10.1109/MICRO.2010.52* to read the related research paper for details.

Advanced Caches:
LLC is not monolithic

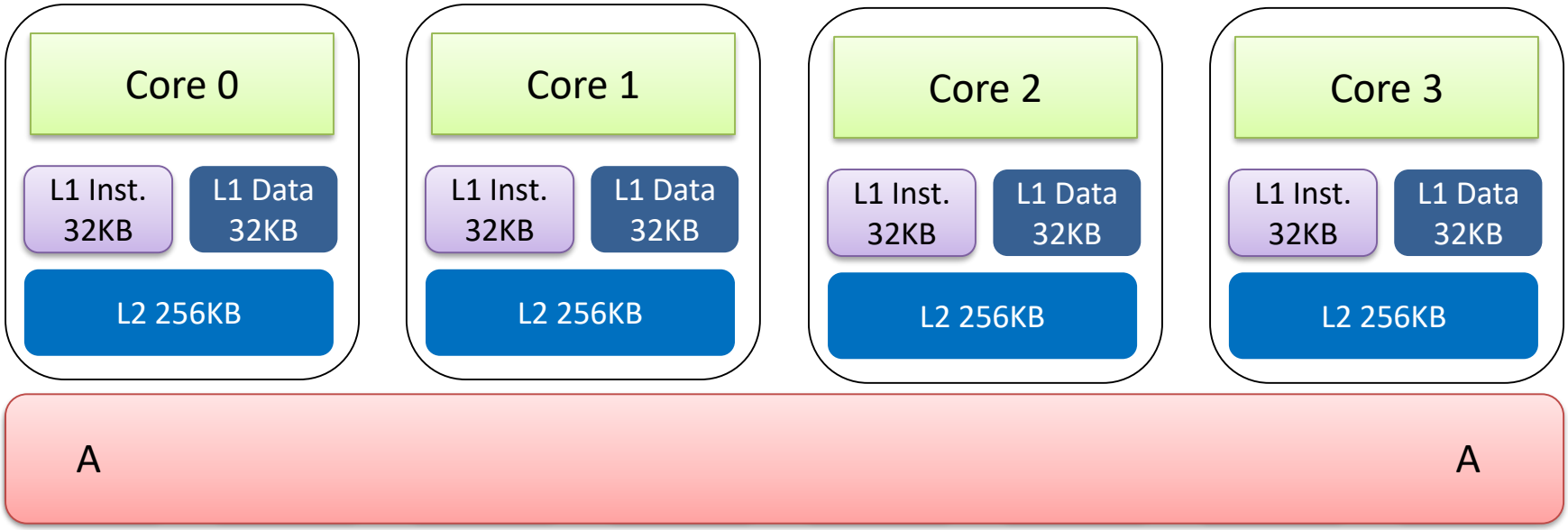
LLC is not monolithic

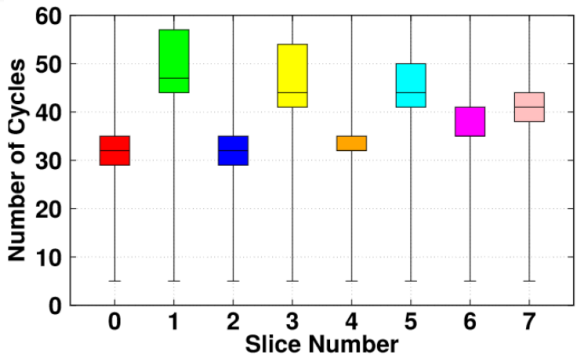


Previously, it's considered that, to CPU cores, LLC is monolithic. No matter where a cache block in the LLC, a core would load it into private L2 and L1 cache with **the same** time cost.

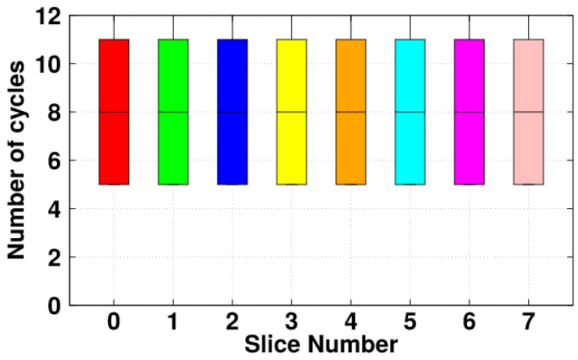
LLC is fine-grained

Intel® Xeon® Processor E5-2667 v3





(a) Read.



(b) Write.

From the paper https://doi.org/10.1145/3302424.3303977

Slice-aware memory management

- The idea seems simple
  - Put your data closer to your program (core)
- But it not *EASY* to do so
  - Cache management is undocumented, not to mention fine-grained slices
  - Researchers did a lot of efforts
    - Click https://doi.org/10.1145/3302424.3303977 for details
    - They managed to improve the average performance by 12.2% for GET operations of a key-value store.
    - 12.2% is a lot, if you consider the huge transactions every day for Google, Taobao, Tencent, JD, etc.

Advanced Caches:
Yes, you can control the cache

Scratchpad Memory

Scratchpad Memory

- Strictly speaking, scratchpad memory (SPM) is **not** cache
  - Widely used in embedded systems
  - On-chip SRAM, like cache, close to ALU
  - *Software controlled*: software decides what data sections to be placed in SPM
    - By the programmer or the compiler before running
  - Memory-mapped to a predefined address range
    - Remember <Base, Bound> registers?

Why SPM?

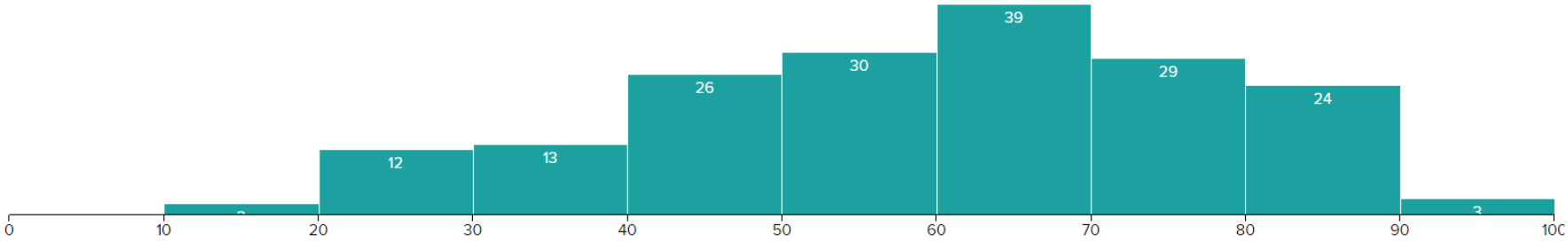
- To control the execution time
  - More predictable than hardware-controlled cache
  - Especially for WCET (worst-case execution time)
- With reduced area and energy consumptions
  - More space- and energy-efficient

Conclusion

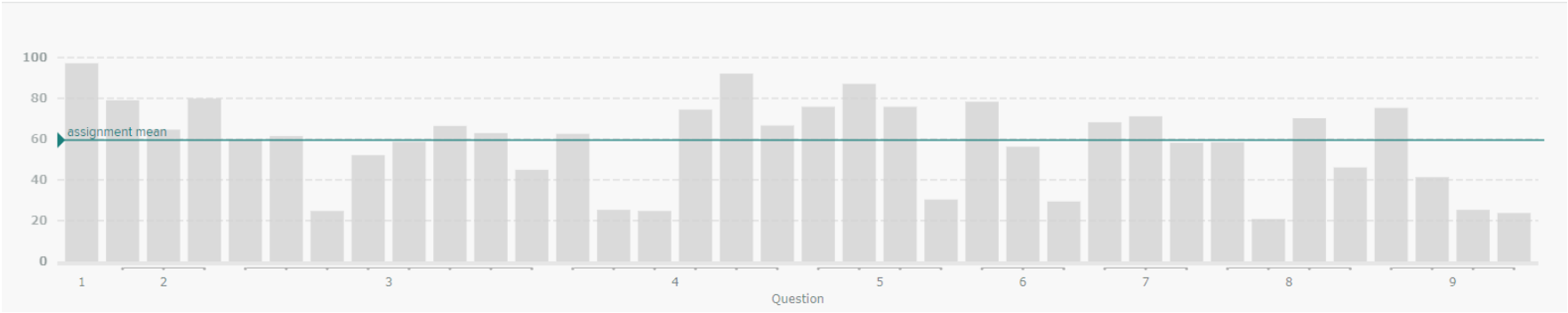
- There are many interesting facts of CPU cache
- To make the best of cache can boost your program's performance!

Midterm II Review

Statistics



MINIMUM	MEDIAN	MAXIMUM	MEAN	STD DEV
16.0	61.5	94.5	59.65	18.0



Q2

- Most can be found in L12
  - P7, P11, P26, etc.

- mv
  - addi rd, rs1, 0 ← P8@L05

Q3

- L13

Q4

- (a)
  - P9@L14, Chapter 3.3, P17@L14, P24@L14
  - P8@L14, P4@L14, P9@L14, P33@L14
- (d)
  - P10@L14
- (f)
  - P21@L14

Q5

- Time!

Q6, Q7

- L16, L17, L18

Q8

- L19

Q9

- L21