Making the Case for CXL Native Memory

Wolley Inc.

Presenter: San Chang

Fast Evolving CXL Specs and Products

Wolley

Flash Memory Summit

Samsung 512GB
CXL Memory Module
(2022-05)

Montage CXL-to-DDR
Controller (2022-05)

SK Hynix
CXL Memory Module
(2022-08)

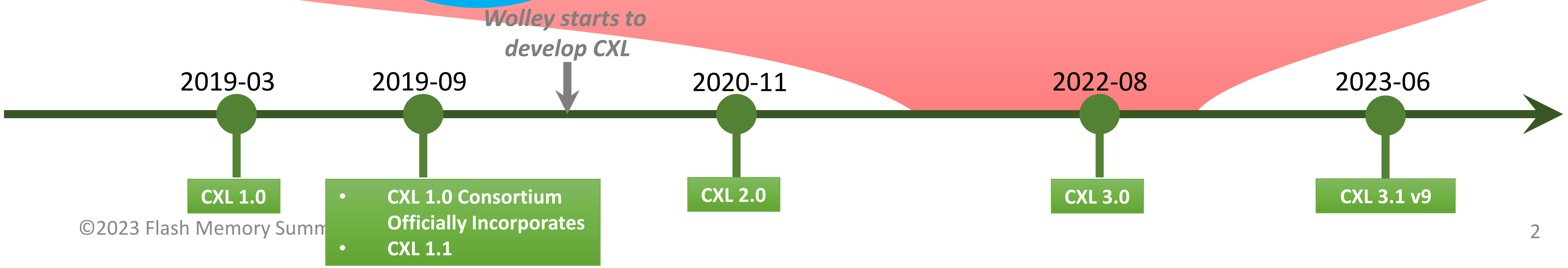
Leo Memory
Connectivity Platform

CXL-attached memory expansion and pooling for cloud servers

Astera Labs CXL-to-DDR
Controller (2022-09)

Samsung
Memory-Semantic SSD (2022-08)

SK Hynix
CMS (2022-10)



Wolley starts to develop CXL

2019-03

CXL 1.0

2019-09

- CXL 1.0 Consortium Officially Incorporates
- CXL 1.1

2020-11

CXL 2.0

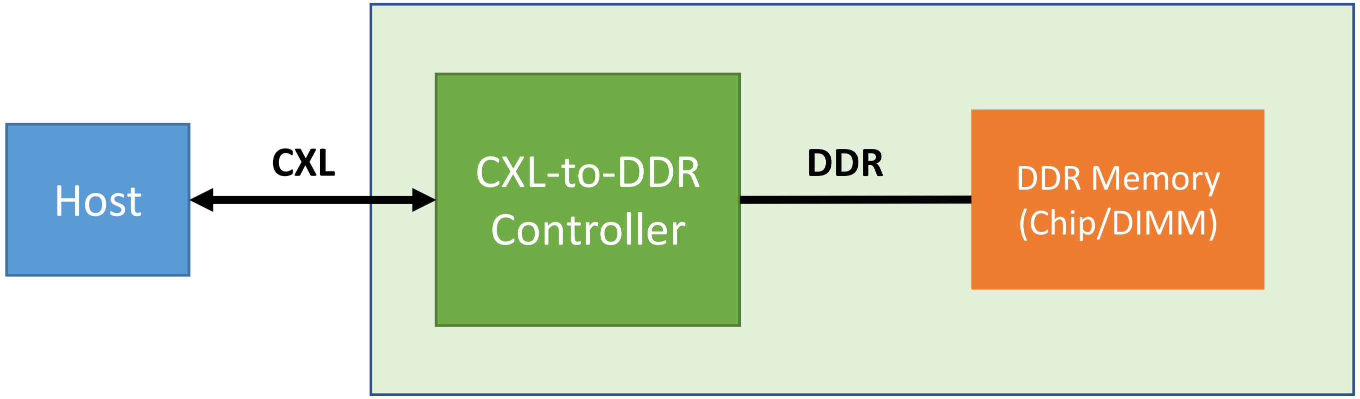
2022-08

CXL 3.0

2023-06

CXL 3.1 v9

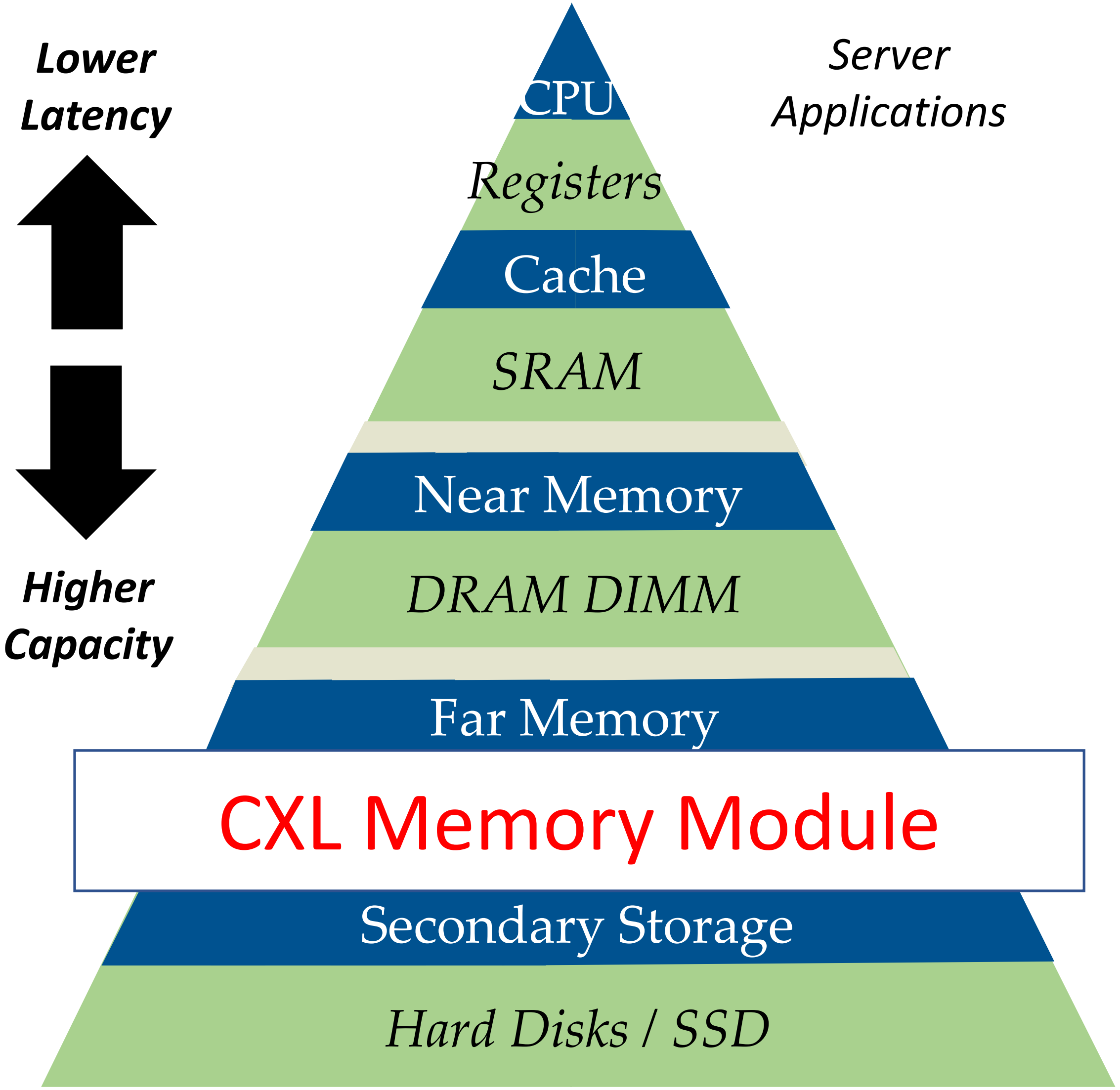
CXL Memory Module for Enterprise/Server Applications

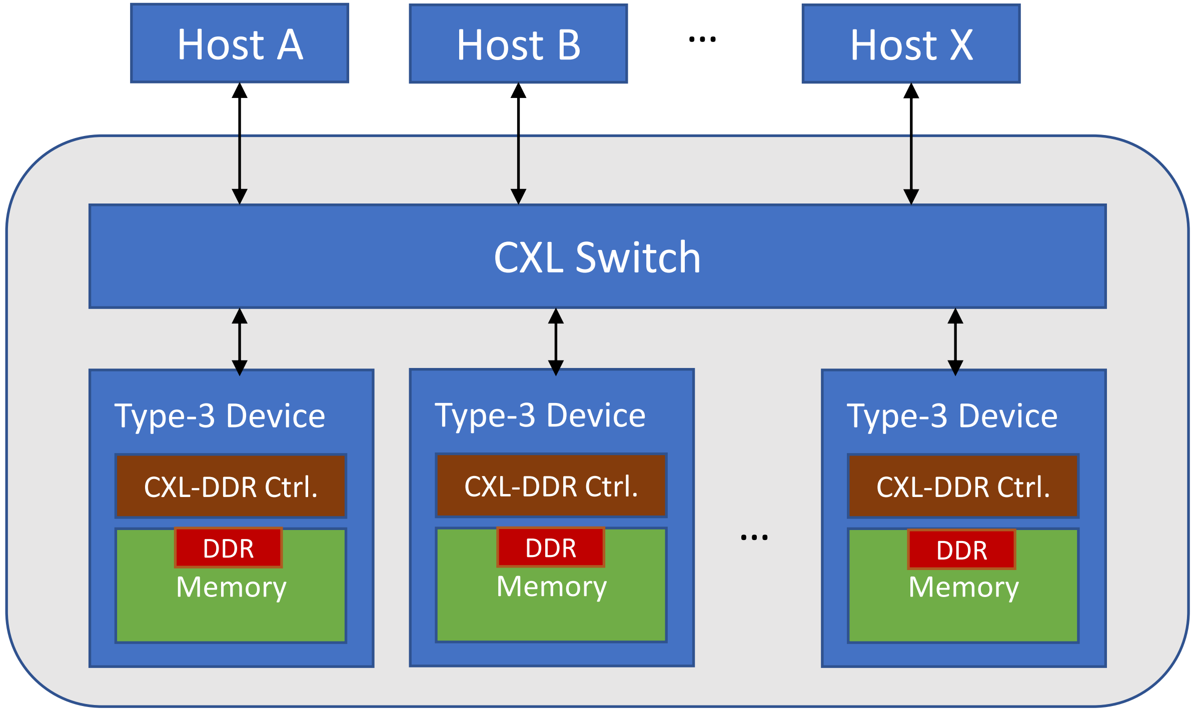


CXL Memory Module

- CXL Memory – the memory "**module**" with CXL interface
  - Usually built with a CXL-to-DDR controller and a number of DDR chips
  - The module can carry multiple DDR chips to offer high memory capacity

CXL Memory Module in Memory Hierarchy



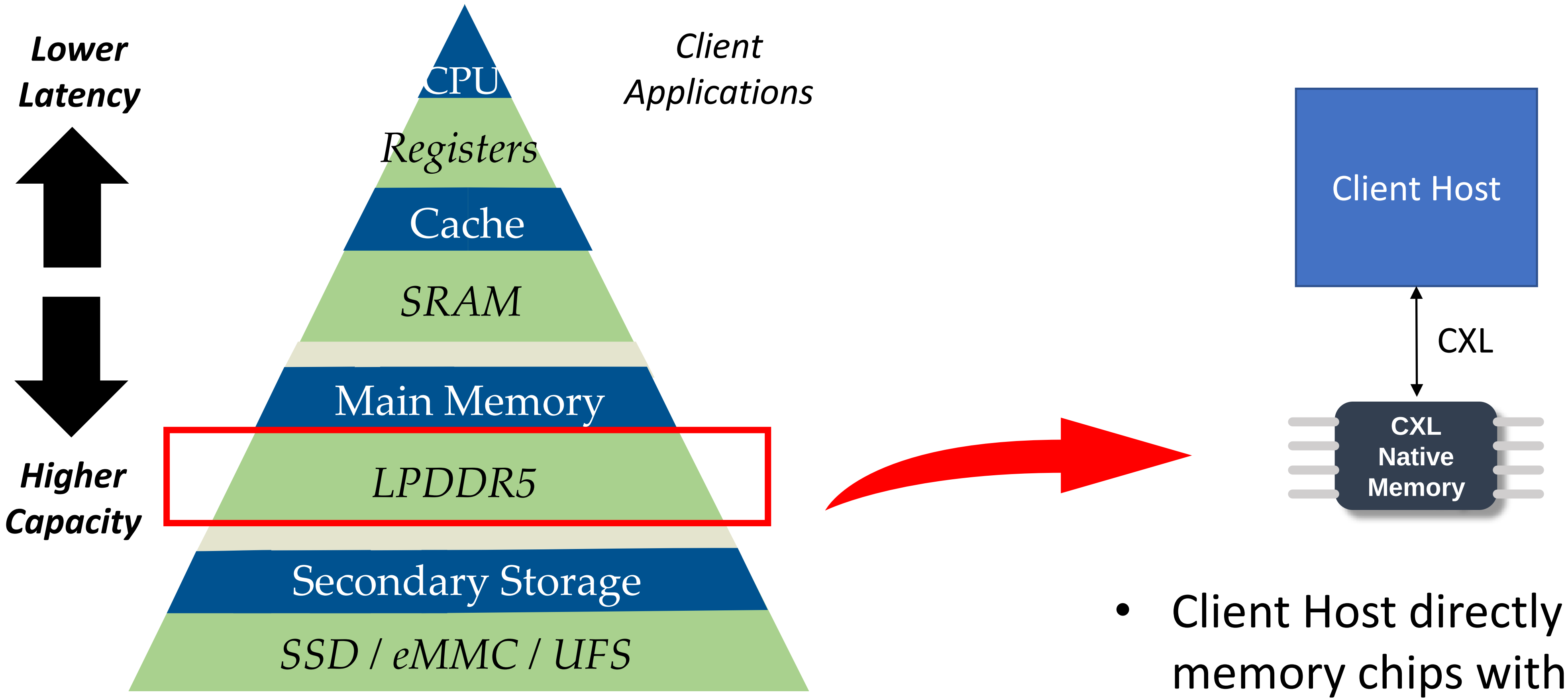


Memory expansion

Memory pooling

Memory sharing

Can CXL Memory Benefit Client Applications?!



- Typical client applications do not require far memory
- Both CXL switch and CXL-to-DDR controllers are "overkill" for client applications

- Client Host directly attaches to memory chips with CXL interface
- LPDDR5 replacement?

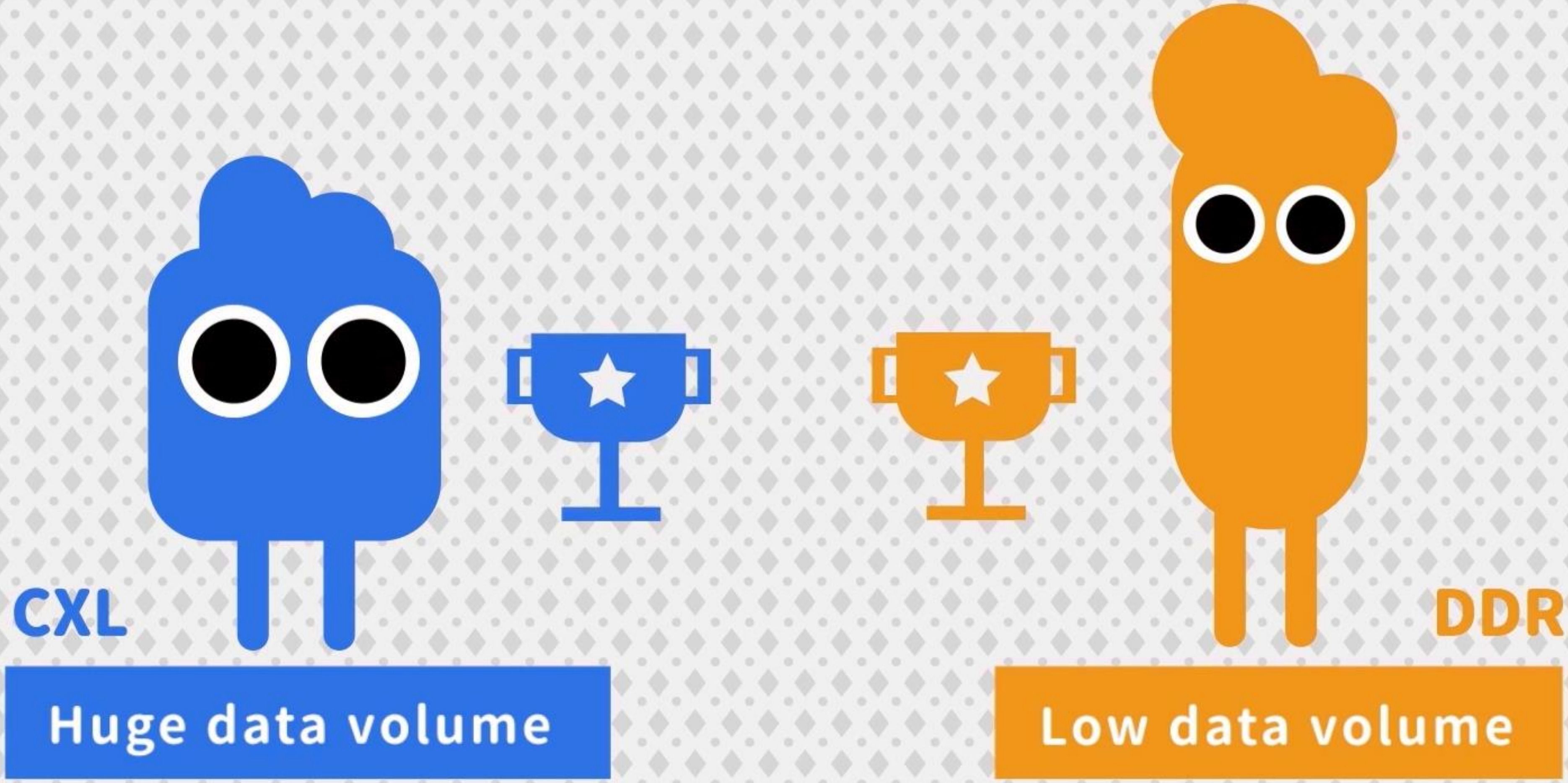
Comparison of CXL and LPDDR5 IP (N7)

	Serial Interfaces	Parallel Interfaces
	CXL (8-lane PCIe Gen5)	**LPDDR5 (2x LPDDR5 6400 16-bit)**
Bandwidth	256Gbps, Full duplex	204.8Gbps, Half duplex
Power Consumption (PHY only)	1.5 ~ 2 pJ/bit	2 ~ 2.5 pJ/bit
IO (Pin Count)	49	68 x 2
Single-command Latency	80-100 ns	40-45 ns

- Longer Latency (***really?***)
- Smaller Area
- Device Controller

- Shorter Latency
- Bigger Area
- Host Controller

Areas of expertise



CXL

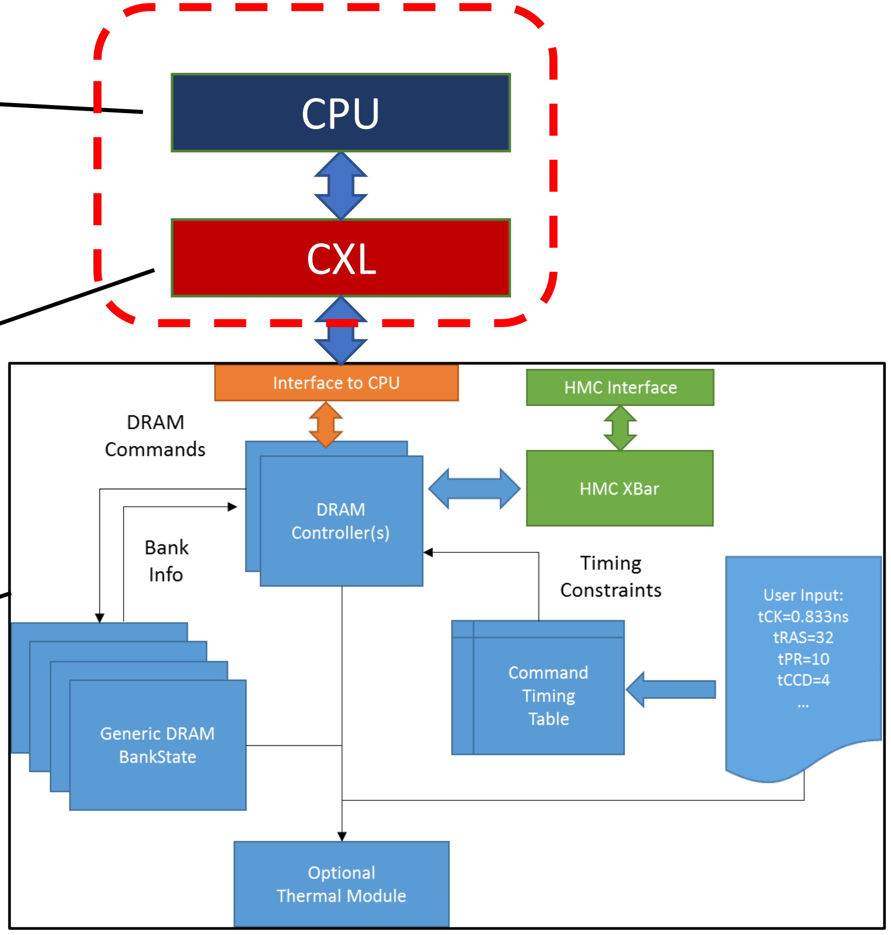
Huge data volume

DDR

Low data volume

System Simulation
- [CPU + CXL] ⟷ CXL Native Memory

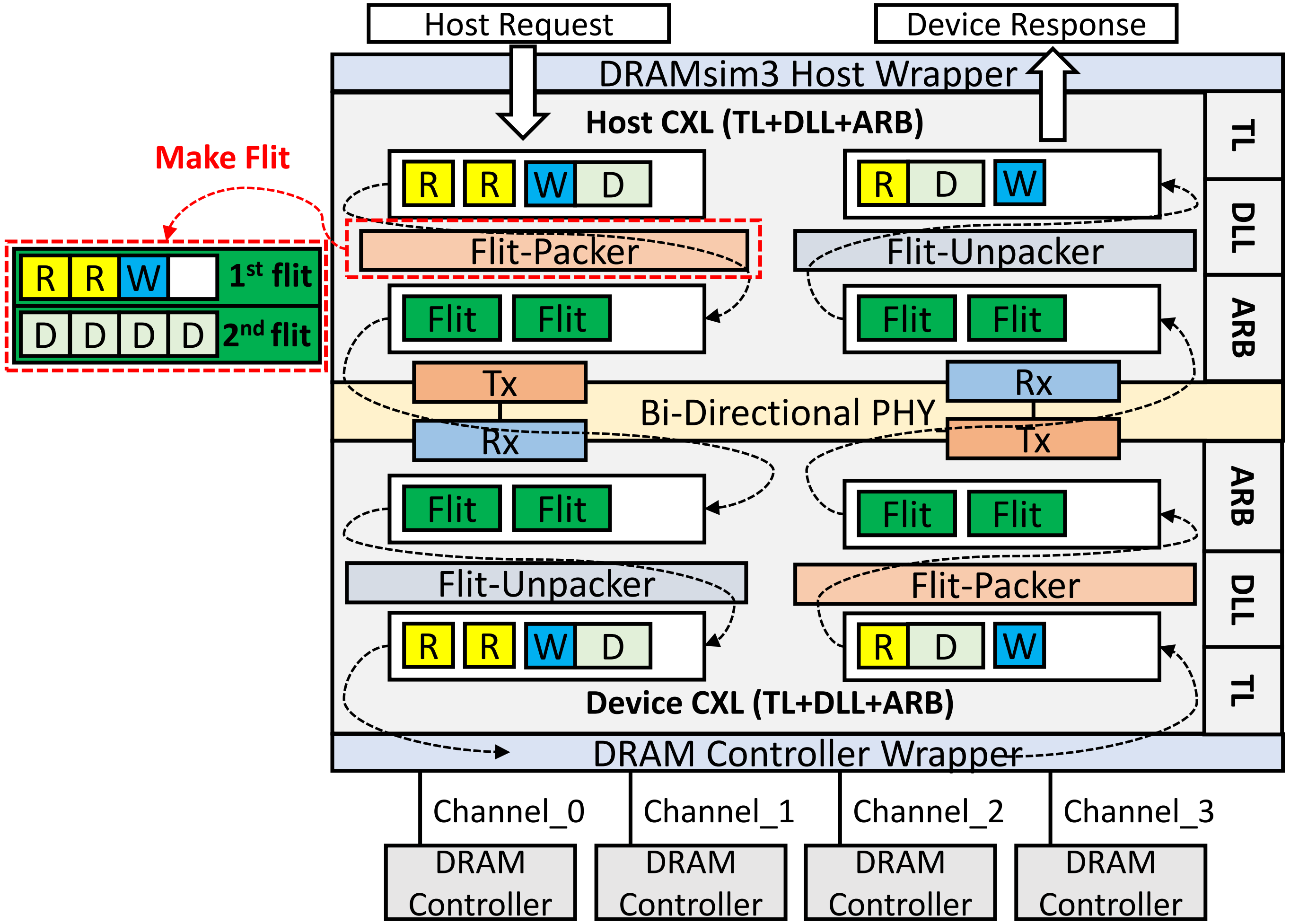
- Gem5
  - A system-level processor simulator widely used by both academic research and industry companies
    - ARM Research, AMD Research, Google, Micron, Metempsy, HP, and Samsung
  - Multiple ISAs (x86, Alpha, ARM, SPARC, MIPS, POWER, RISC-V)
- CXL Modeling
  - Wolley-developed model to simulate full-duplex interface
  - Gen5x8, 32GB/s: 1GHz, 1ns
    (PIPE: 32bit per lane => 32B 8-lane @1GHz)
- DRAMSim3
  - An extension of the well-known DRAM model DRAMSim2
  - Cycle-accurate and multiple protocols support (DDR3, DDR4, LPDDR3, LPDDR4, GDDR5, GDDR6, HBM, HMC, STT-MRAM)



CXL Native Memory

System Simulation
- CXL Interface + CXL Native Memory

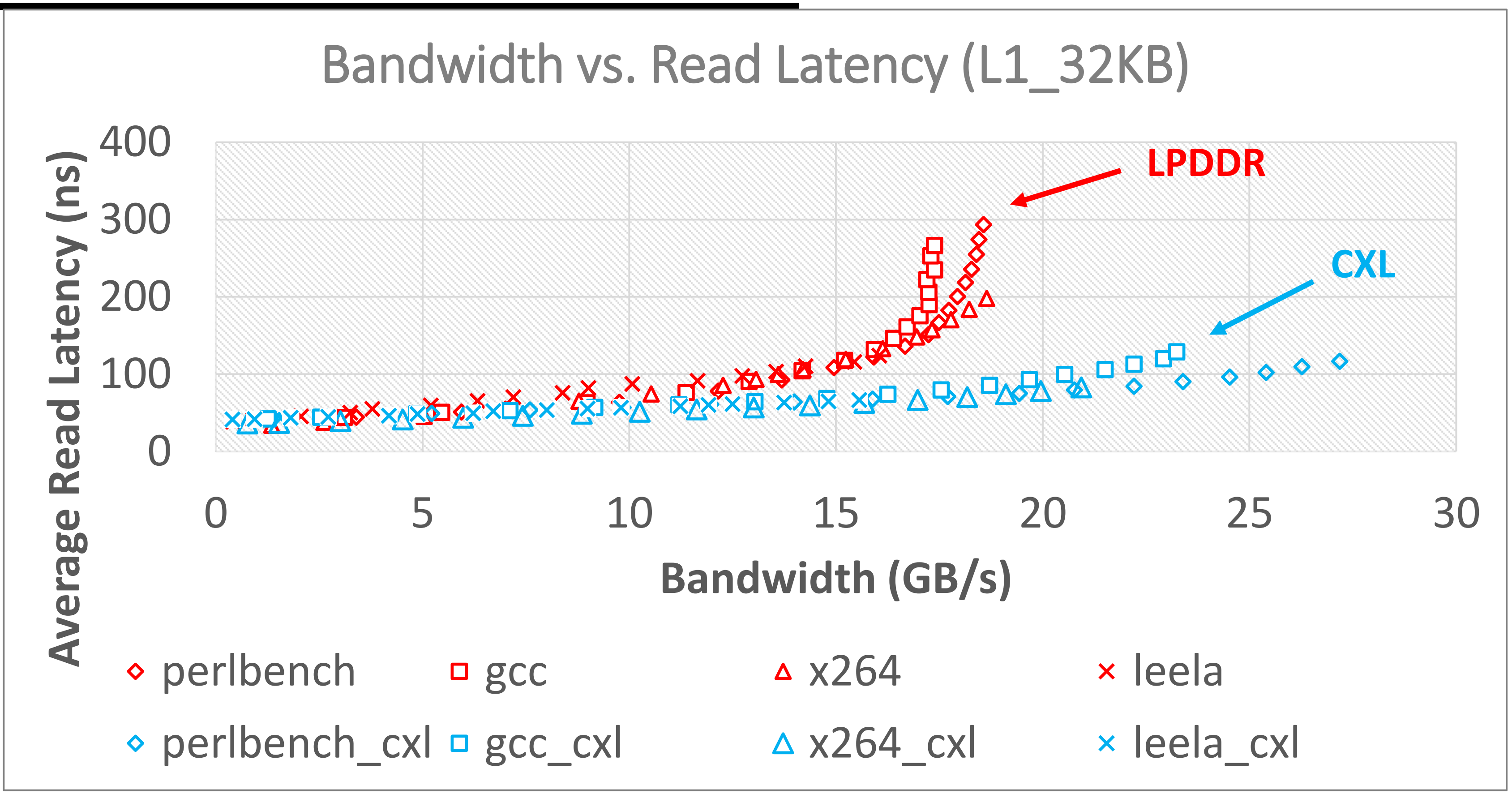
- Simulation includes
    - CXL translation layer
    - CXL link layer
    - CXL ARB/MUX
- Detail setup of timing parameters refers to our IP design
- We use existing DRAM controller wrapper to emulate the memory management inside the CXL Native Memory
    - Without DDR, real implementation could have a shorter latency



CXL Latency Outperforms in Congested Bandwidth – Case 1 (L1_32KB)

- Application latency is much shorter with CXL Native Memory in real applications, particularly when memory bandwidth is under high utilization

- R/W Ratio
  - perlbench (r53/w47)
  - gcc (r54/w46)
  - x264 (r64/w36)
  - leela (r79/w21)

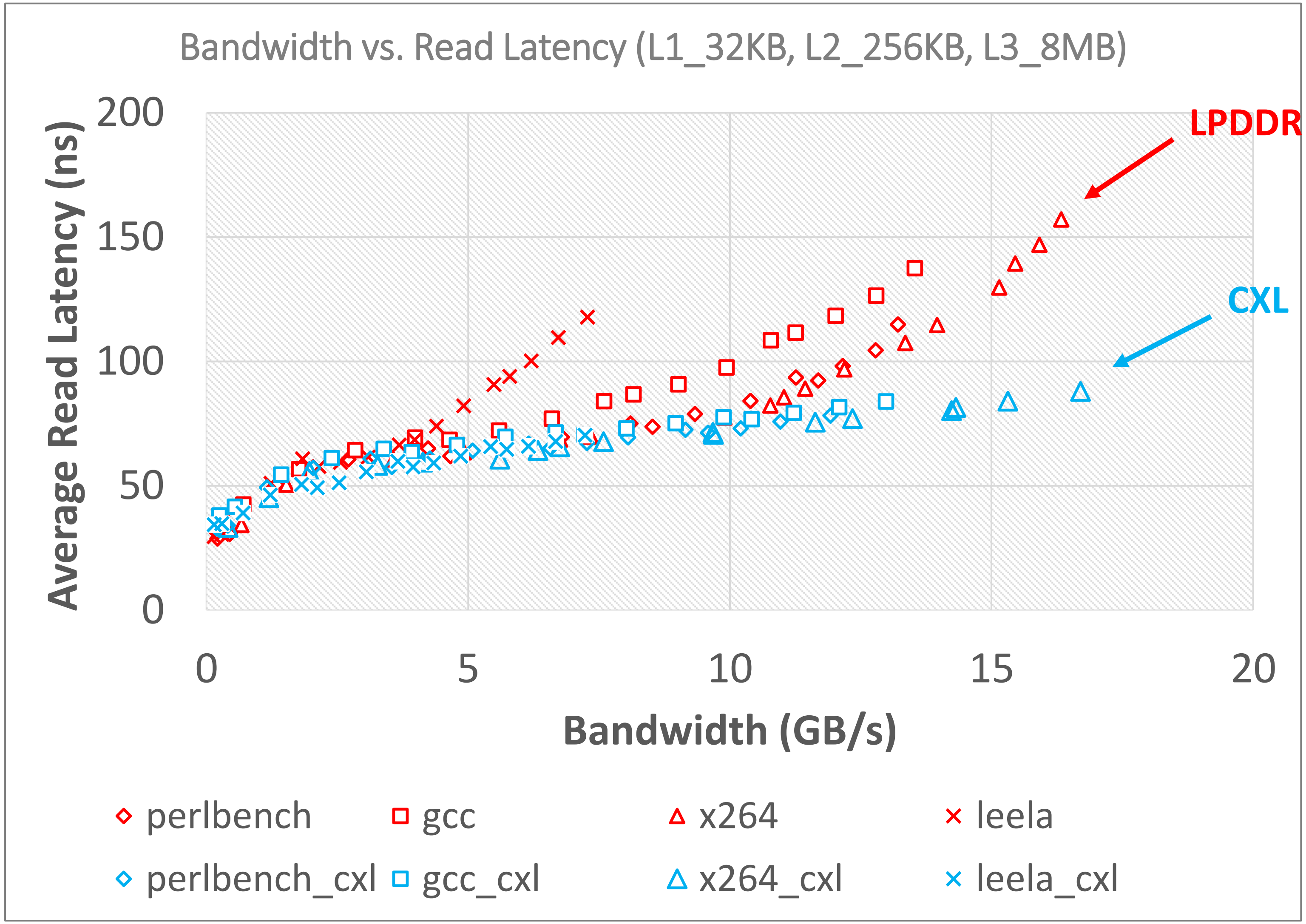


- **LPDDR**
  - **6.4Gb/s, 16-bit, 2-channel**
  - **Max bandwidth: 25.6GB/s**
- **CXL**
  - **Gen5x8**
  - **Max bandwidth: 32GB/s**

CXL Native Memory

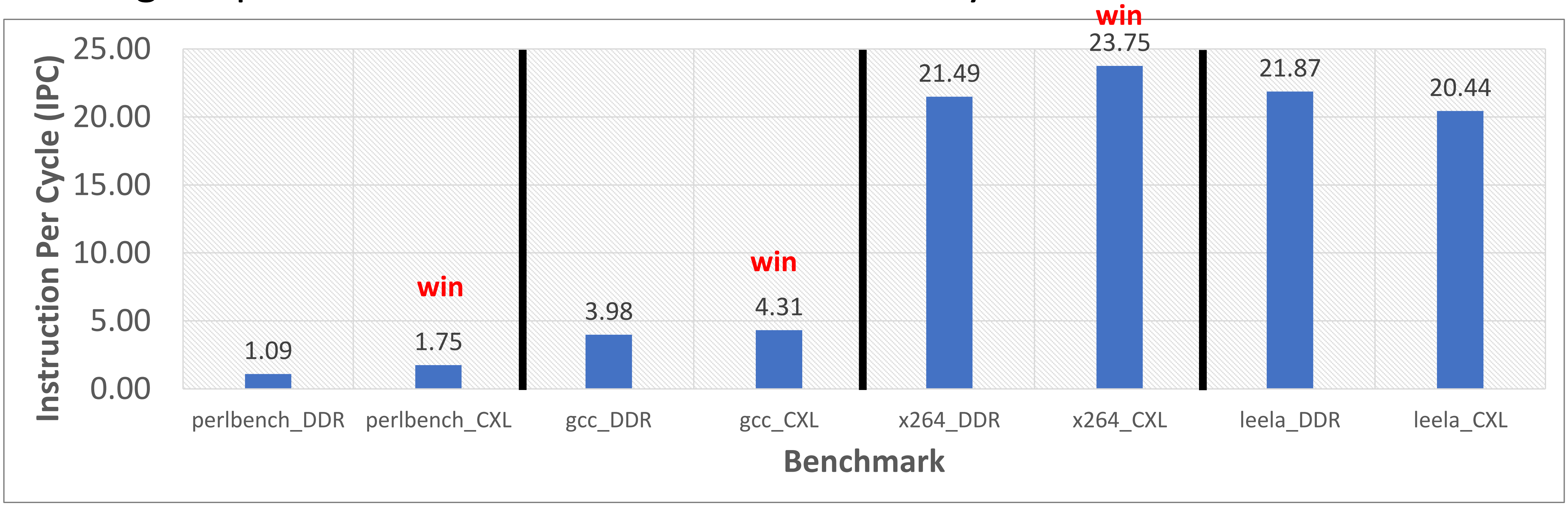
CXL Latency Outperforms in Congested Bandwidth – Case 2 (L1_32KB, L2_256KB, L3_8MB)

- Adopting three-level cache architecture reduces latency apparently
  - Below 150ns (w/o CXL)
  - Below 100ns (w/ CXL)
- Observation
  - Without CXL, latency generally gets longer when memory bandwidth utilization is higher



CXL Memory Performance Better than LPDDR

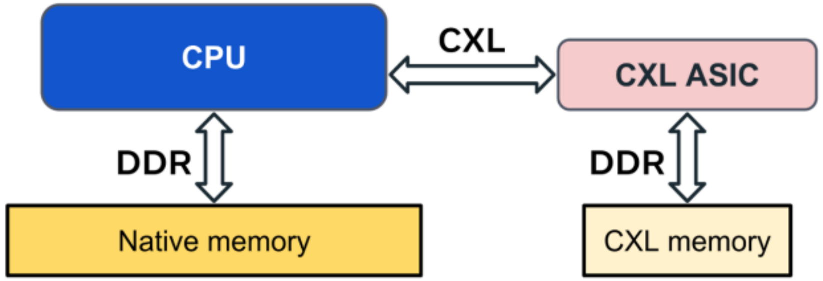
- Every configuration executes the same number of CPU cycles
  - More instructions executed means better performance
- Higher performance consumes more memory bandwidth

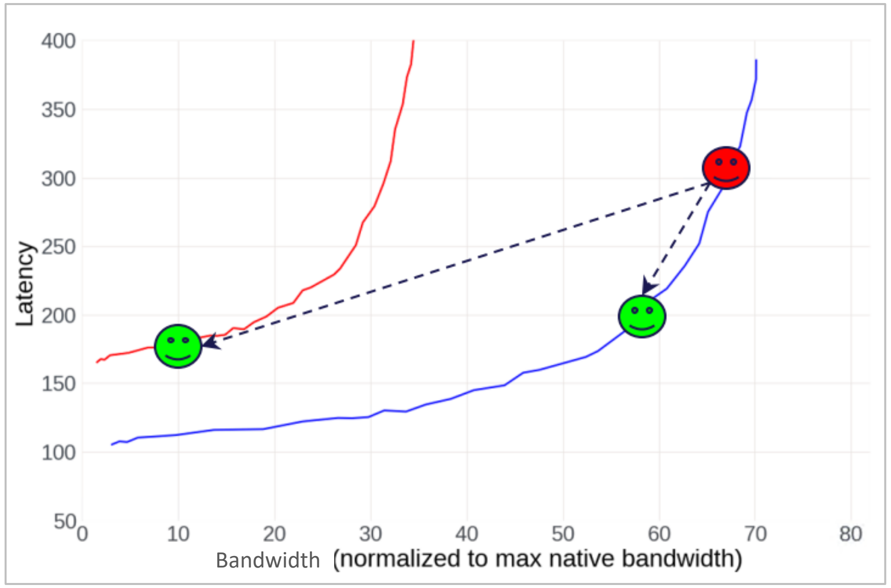


The Viewpoint of Meta on System Latency

- Memory Source
  - DDR Memory (blue line)
  - CXL Memory (red line)
- Summary
  - When DDR bandwidth hits the bottleneck, the latency suffers apparently (red face)
  - CXL memory provides extra bandwidth
    - Move the “right” amount of pages to CXL memory
  - Overall application latency decreases

Bandwidth utilization have a significant impact on application latency

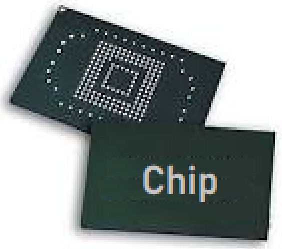




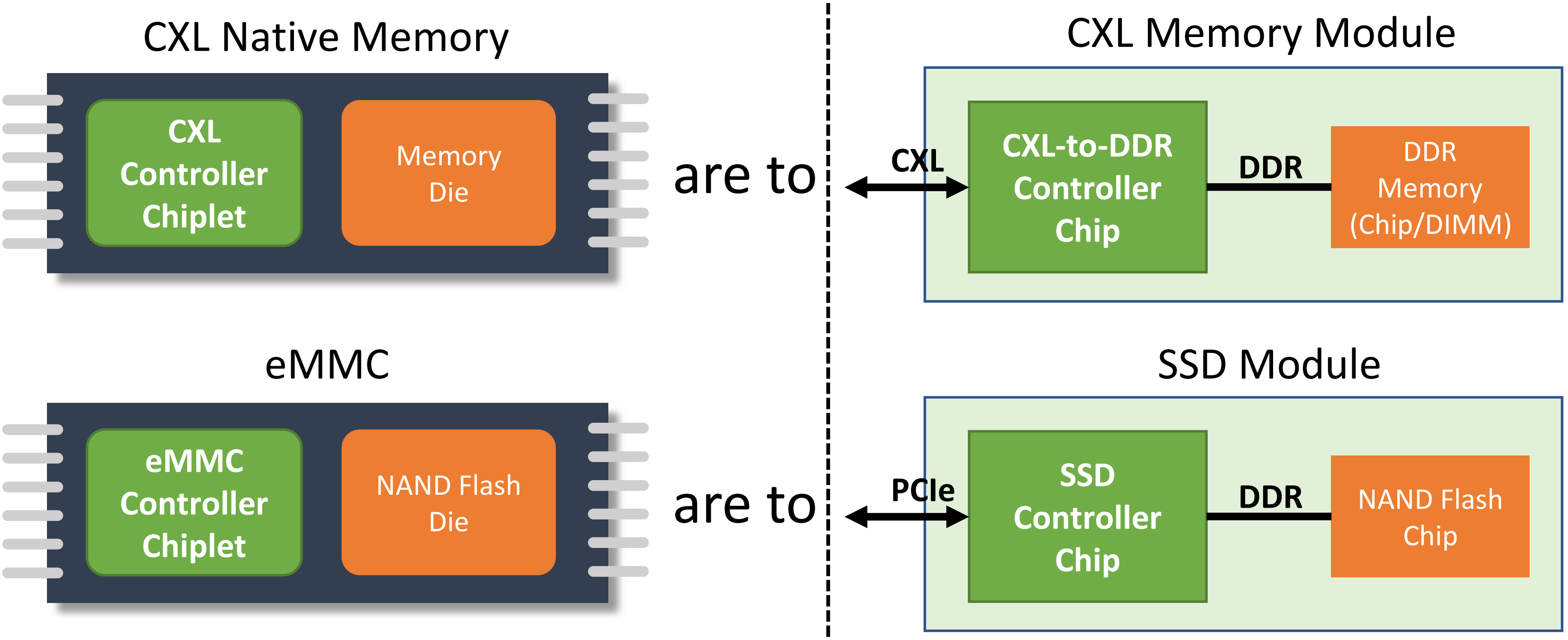
Understanding memory usage in datacenters and Enabling software for CXL-Memory, OCP,2022 (https://www.youtube.com/watch?v=lS2CE-1sgsE)

CXL Native Memory as "eMMC for Memory"

	CXL Native Memory	CXL Memory Module
Applications	Client	Server
Definition	the memory "chip" with CXL interface	the memory "module" with CXL interface
Implementation	CXL controller and memory will be tightly-coupled inside the chip	Usually built with a CXL-to-DDR controller and a number of DDR chips



as



CXL Native Memory

CXL Controller Chiplet

Memory Die

are to

CXL Memory Module

CXL

CXL-to-DDR Controller Chip

DDR

DDR Memory (Chip/DIMM)

eMMC

eMMC Controller Chiplet

NAND Flash Die

are to

SSD Module

PCIe

SSD Controller Chip

DDR

NAND Flash Chip

Client Application Space

Server Application Space

Call-For-Action from Client/Mobile Users

- In order to make progress on CXL Native Memory, the baton lies in the hands of client/mobile users
  - Memory companies won’t create new products without customer interest
- Technical/Business benefit of CXL Native Memory to client/mobile users
  - Better performance (higher bandwidth, lower application latency)
  - More future-proof (LPDDR5 reaching per-pin performance limit)
  - Small/cheaper client host processor (die size no longer IO bound due to LPDDR5 IP)
  - Host processor no longer needs to be responsible for media management – this allow easier switching to different CXL Native Memory chips – which has significant business negotiation advantage
- Dear client/mobile people: let us work together to promote and deliver CXL Native Memory to the industry

CXL
Native
Memory

Summary

- *CXL memory module* has caught most of the attention so far for server applications requiring memory disaggregation
- In this presentation, we highlight *CXL Native Memory* as another interesting memory device for client applications
- Most people have the impression that due to the inherent serial/parallel operation, CXL has a higher latency than LPDDR5
  - But this is only true for a single-command comparison – we showed through simulations that for any practical workload, CXL actually has a lower application latency than LPDDR5
- Wolley will work with memory companies and client/mobile users on CXL Native Memory products

CXL
Native
Memory