

Demystifying CXL Memory Bandwidth Expansion for Analytical Workloads

Georgiy Lebedev, Hamish Nicholson, Musa Ünal,
Sanidhya Kashyap, Anastasia Ailamaki

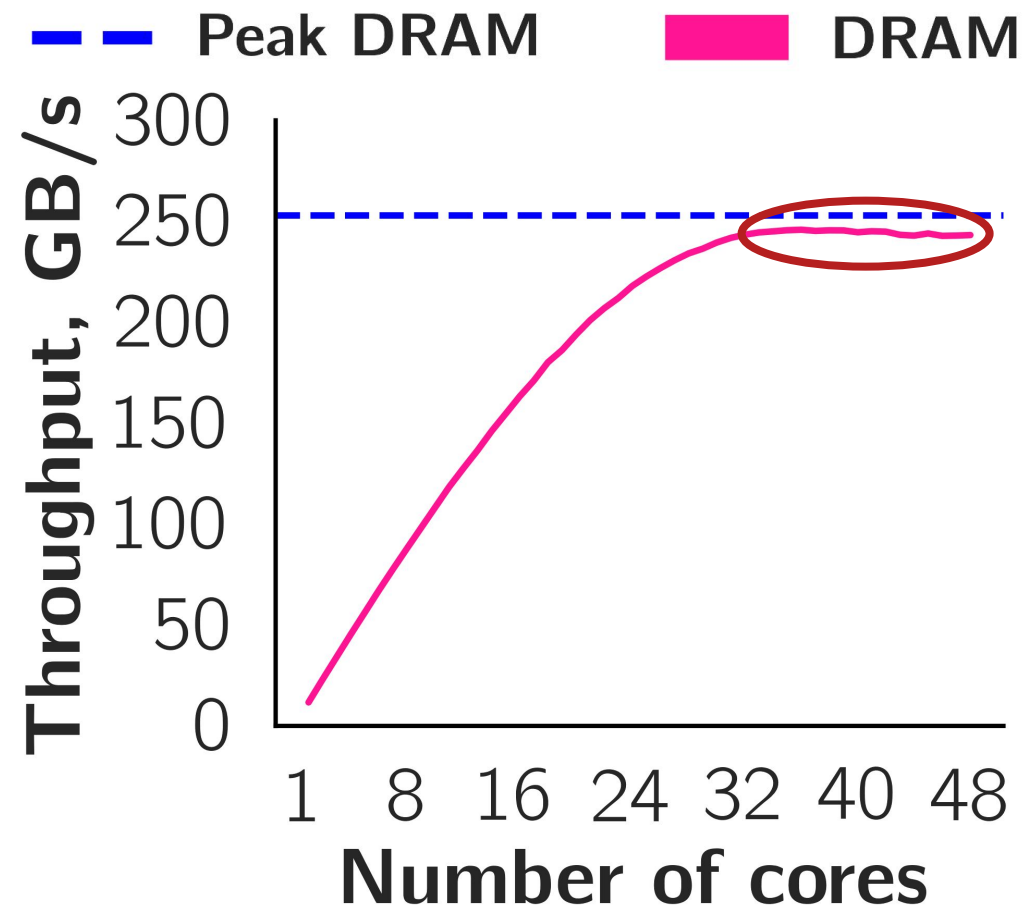
Data-intensive systems rely on DRAM for performance

CPU: 48 cores, 1 socket

DRAM: 253 GB/s

Sequential access: 32 GB array, 8-byte integers

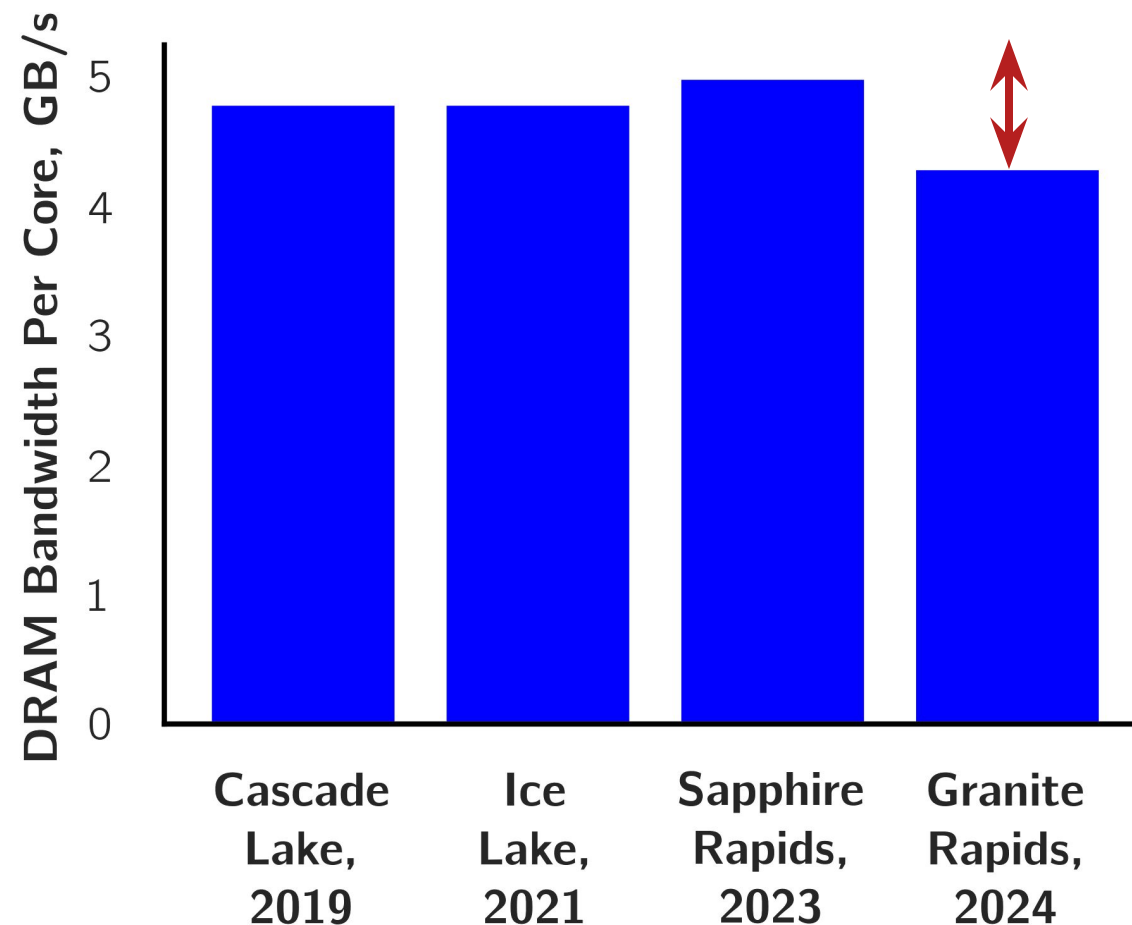
- Bandwidth-bound stalls limit performance
- CPU cores wasted



Bottleneck: memory bandwidth

Memory bandwidth per core improvement is stagnating

- # of chips per channel — signal integrity
- # of channels per socket — pin budget
- Bandwidth/pin — DDR interconnect



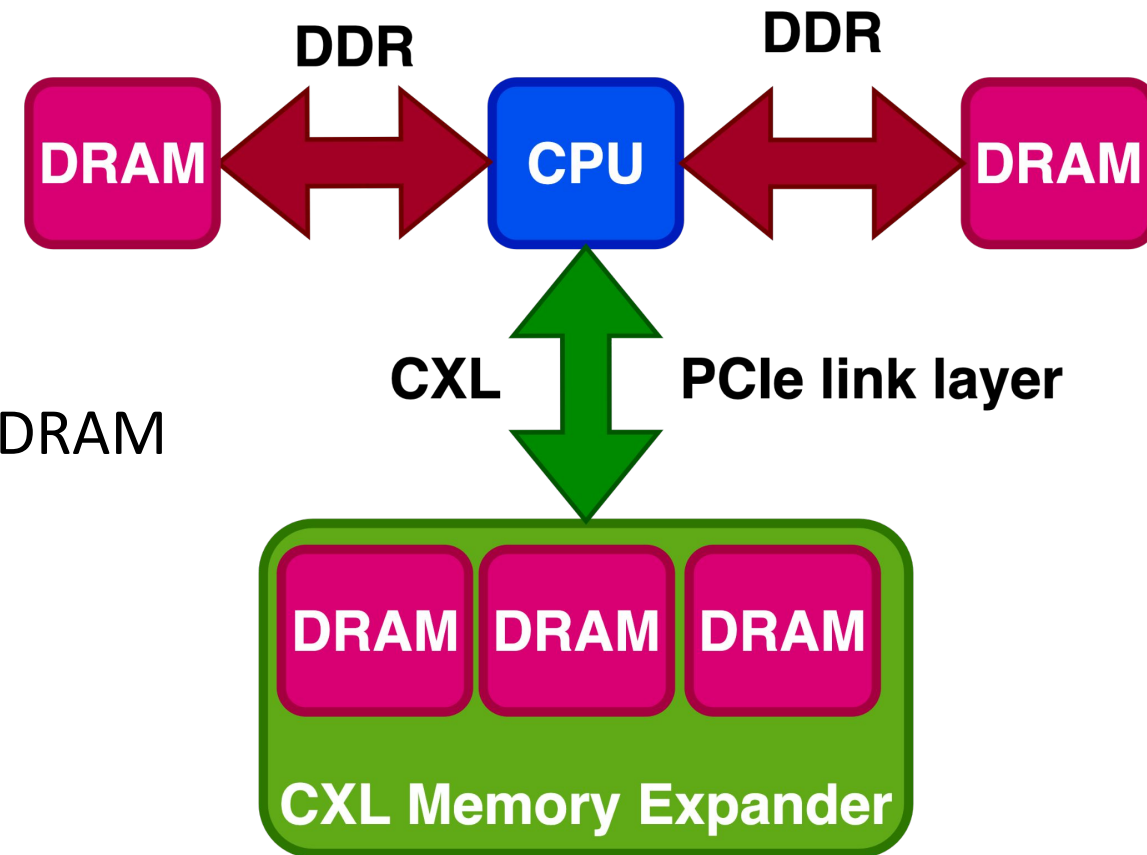
Additional source of bandwidth is needed

CXL adds memory bandwidth from PCIe interconnect

DRAM: 253 GB/s

CXL: 41 GB/s

- Bandwidth scales with PCIe generations
- Performance characteristics differ from local DRAM



Memory heterogeneity demands a new allocation model

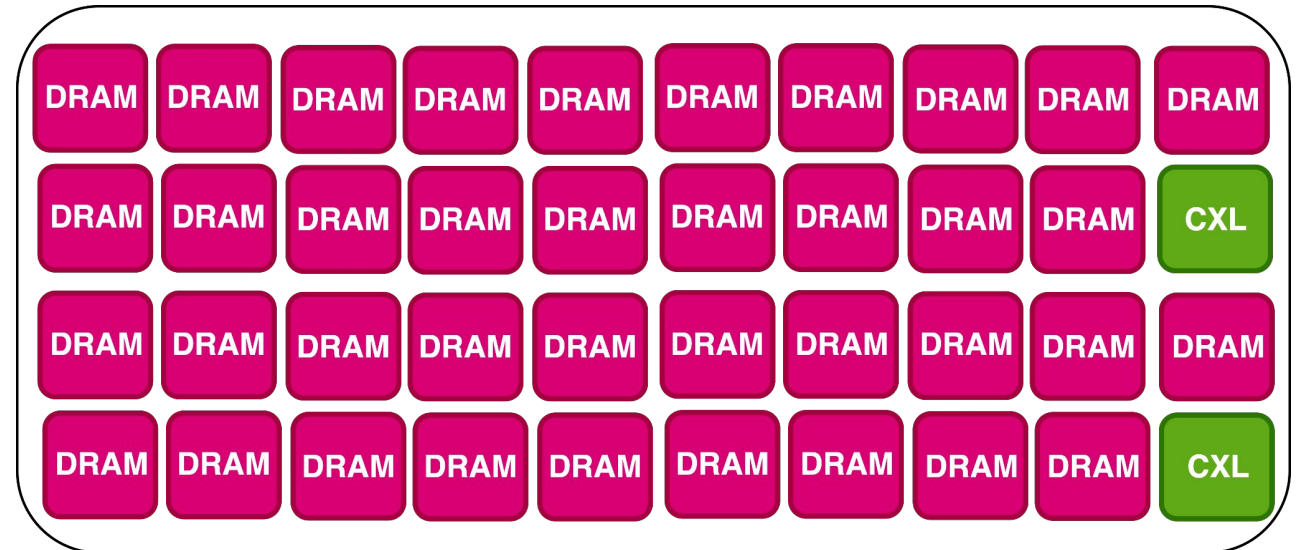
Weighted heterogeneous memory interleaving

DRAM: 253 GB/s

CXL: 41 GB/s

- avoid software modification
- adjust to difference in bandwidths
- maximize system bandwidth

OS allocation policy



DRAM:CXL at 19:1

CXL adds to DRAM as if it was another memory channel

CXL improves performance under high bandwidth utilization

CPU: 48 cores, 1 socket

DRAM: 253 GB/s

CXL: 41 GB/s

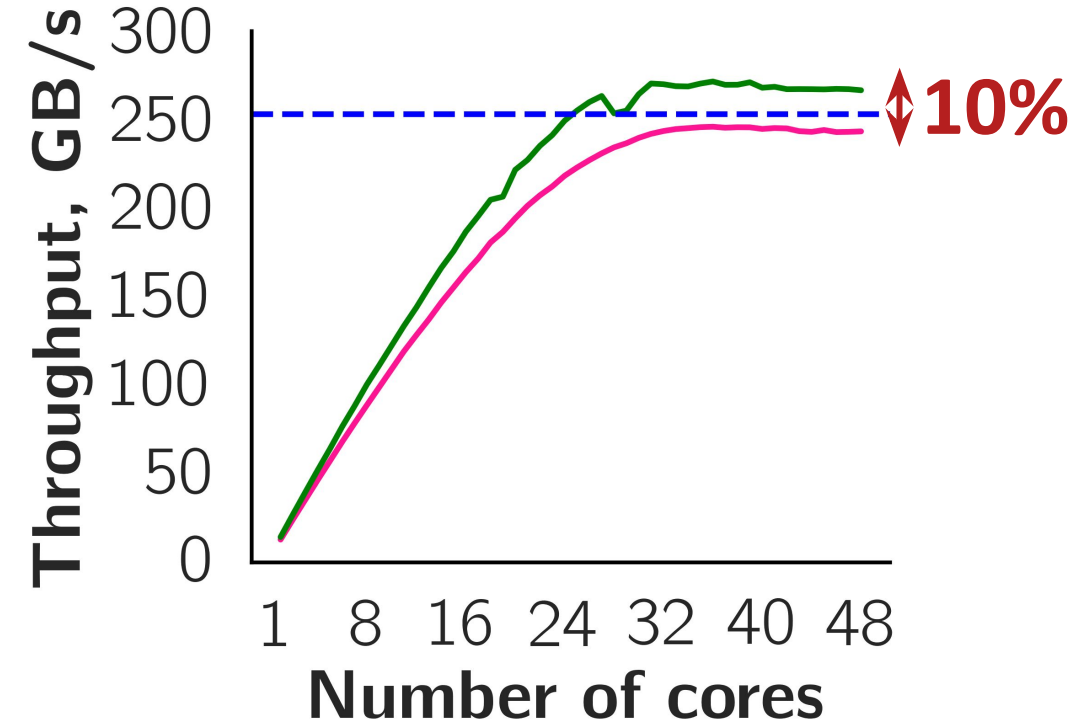
DRAM:CXL interleaving at 19:1

Sequential access: 32 GB array, 8-byte integers

— Peak DRAM ■ DRAM ■ DRAM:CXL

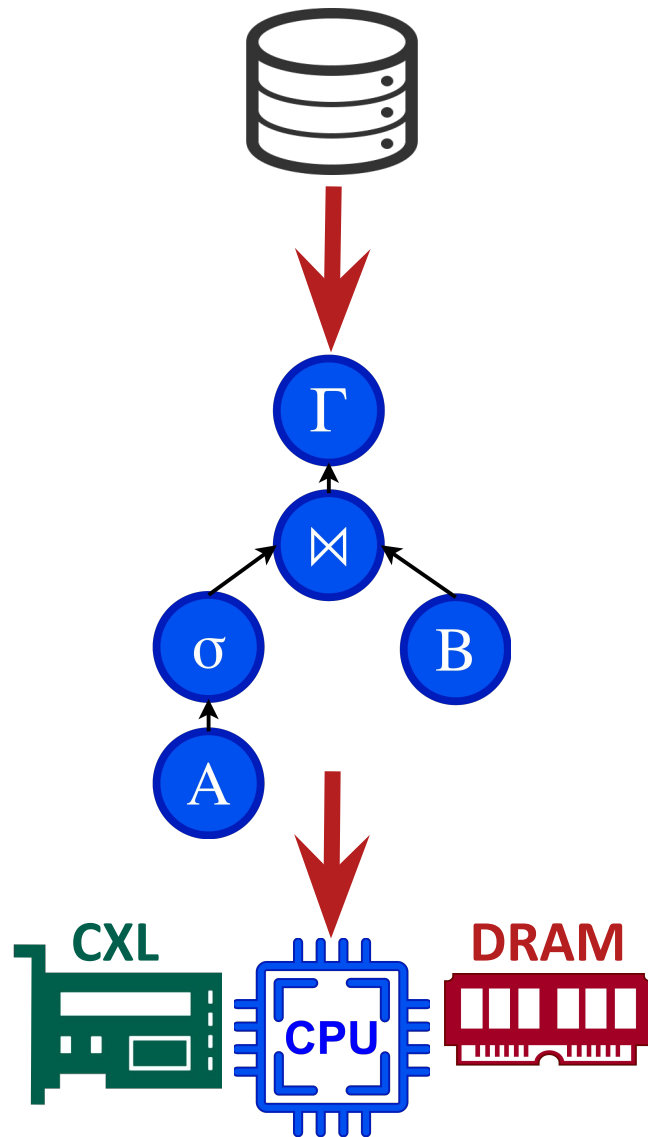
DRAM latency > CXL latency:

111 ns + bandwidth-bound stall > 230 ns



Interleaving CXL with DRAM reduces bandwidth-bound stalls

CXL memory bandwidth expansion challenges



- Is it a drop-in solution?
- Is workload adaptivity needed?
- What determines performance gains?

Transparent DRAM:CXL interleaving degrades performance

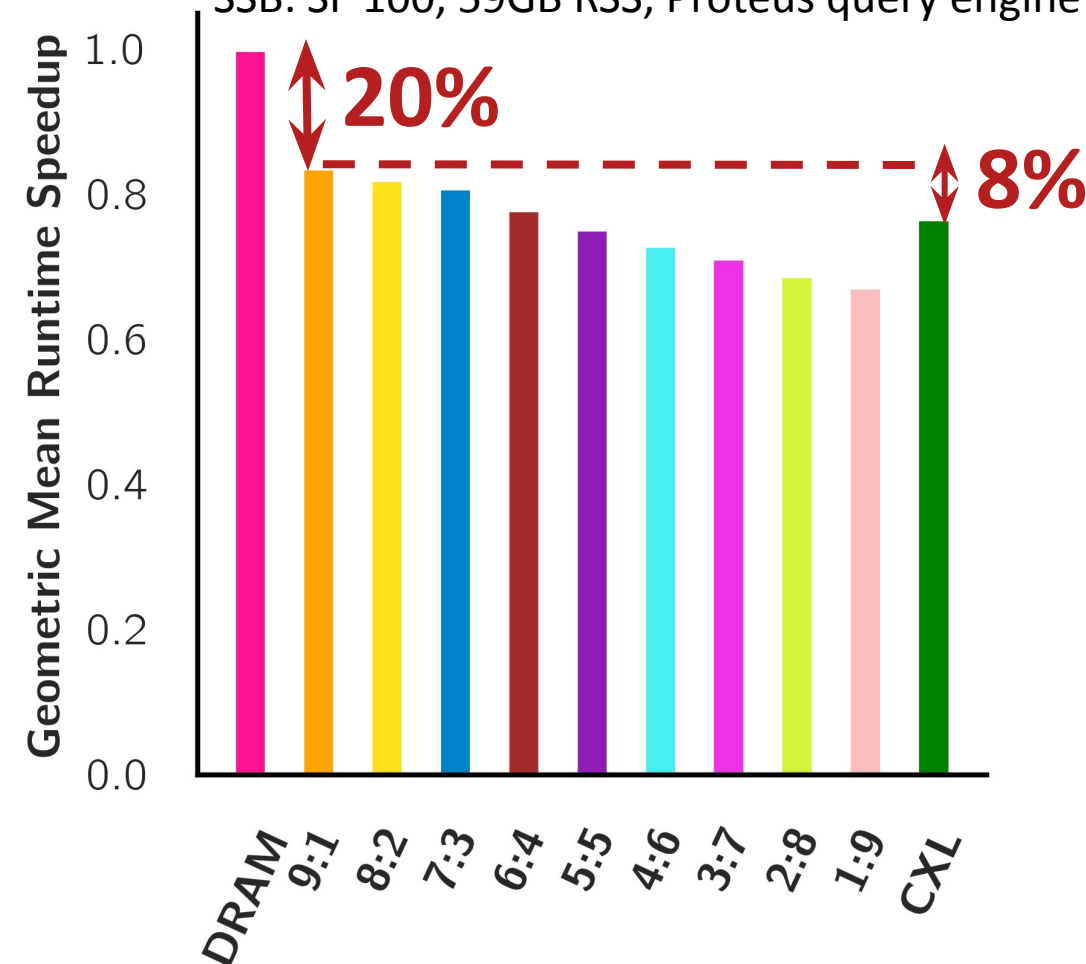
CPU: 48 cores, 1 socket

DRAM: 253 GB/s

CXL: 41 GB/s

SSB: SF 100, 59GB RSS, Proteus query engine

- Latency-critical memory is accessed from CXL
- Latency-bound memory stalls dominate
- % of memory-bound CPU cycles increases



Memory management must be latency- and bandwidth- conscious 8

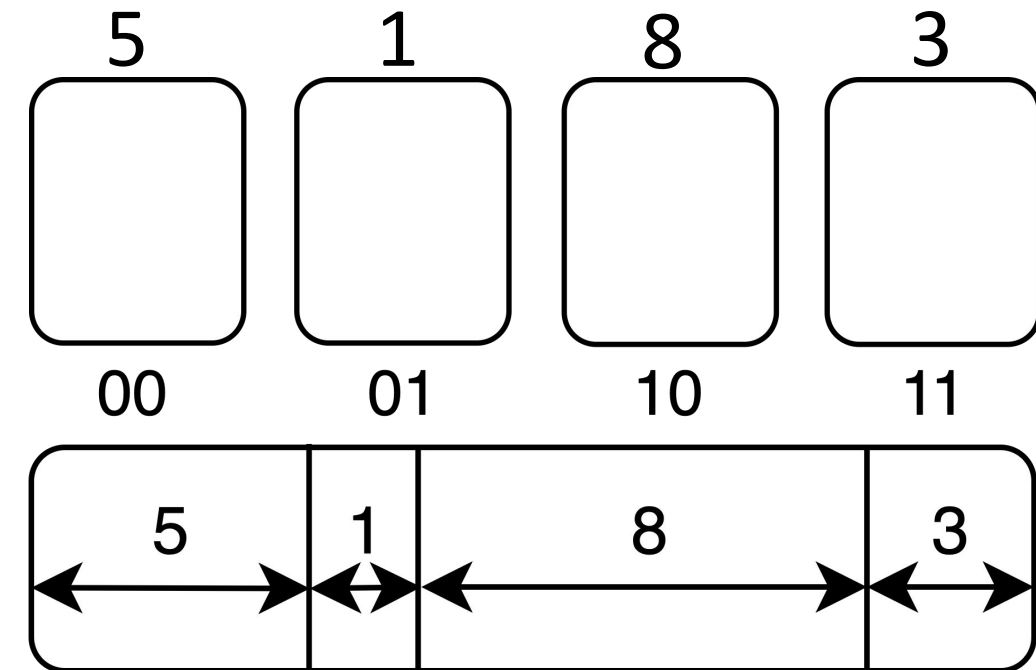
Workload adaptivity is essential for optimal performance

CPU: 48 cores, 1 socket

DRAM: 126 GB/s (SNC-2)

CXL: 41 GB/s

Radix partitioning: 16 GB array, 8-byte integers



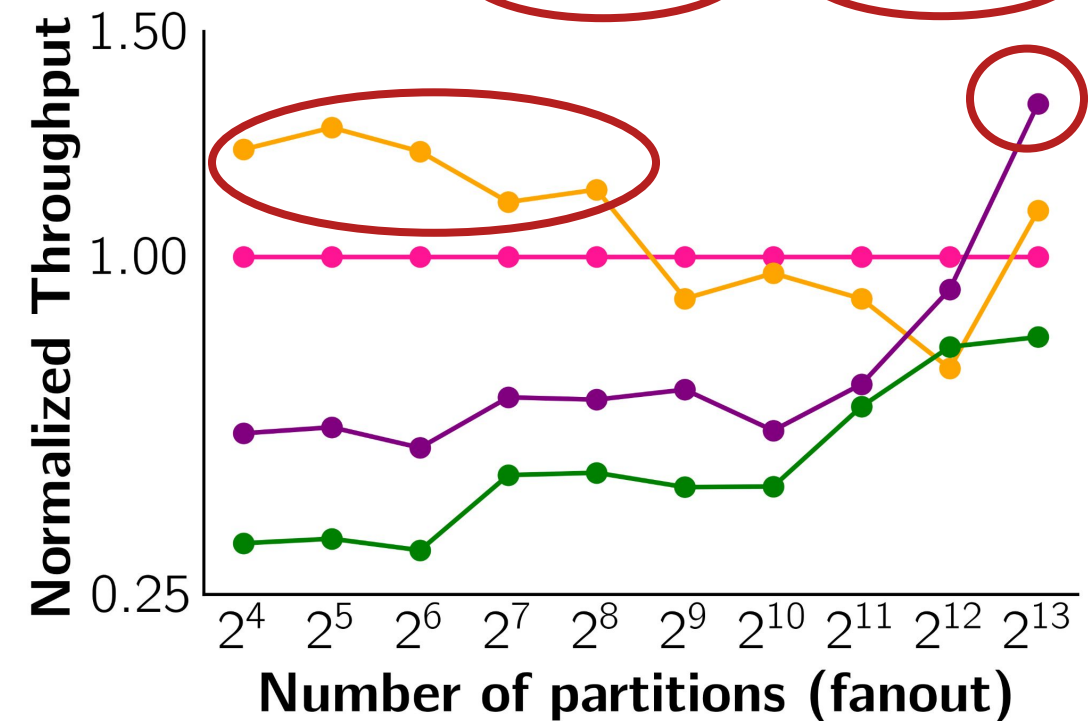
- Read-write ratio moves towards more writes
- Store-bound memory stalls reduction

DRAM

CXL

8:2

4:6



Optimal interleaving weights vary with read-write ratio by up to 88%

High degree of memory-level parallelism is required

CPU: 48 cores, 1 socket

DRAM: 253 GB/s

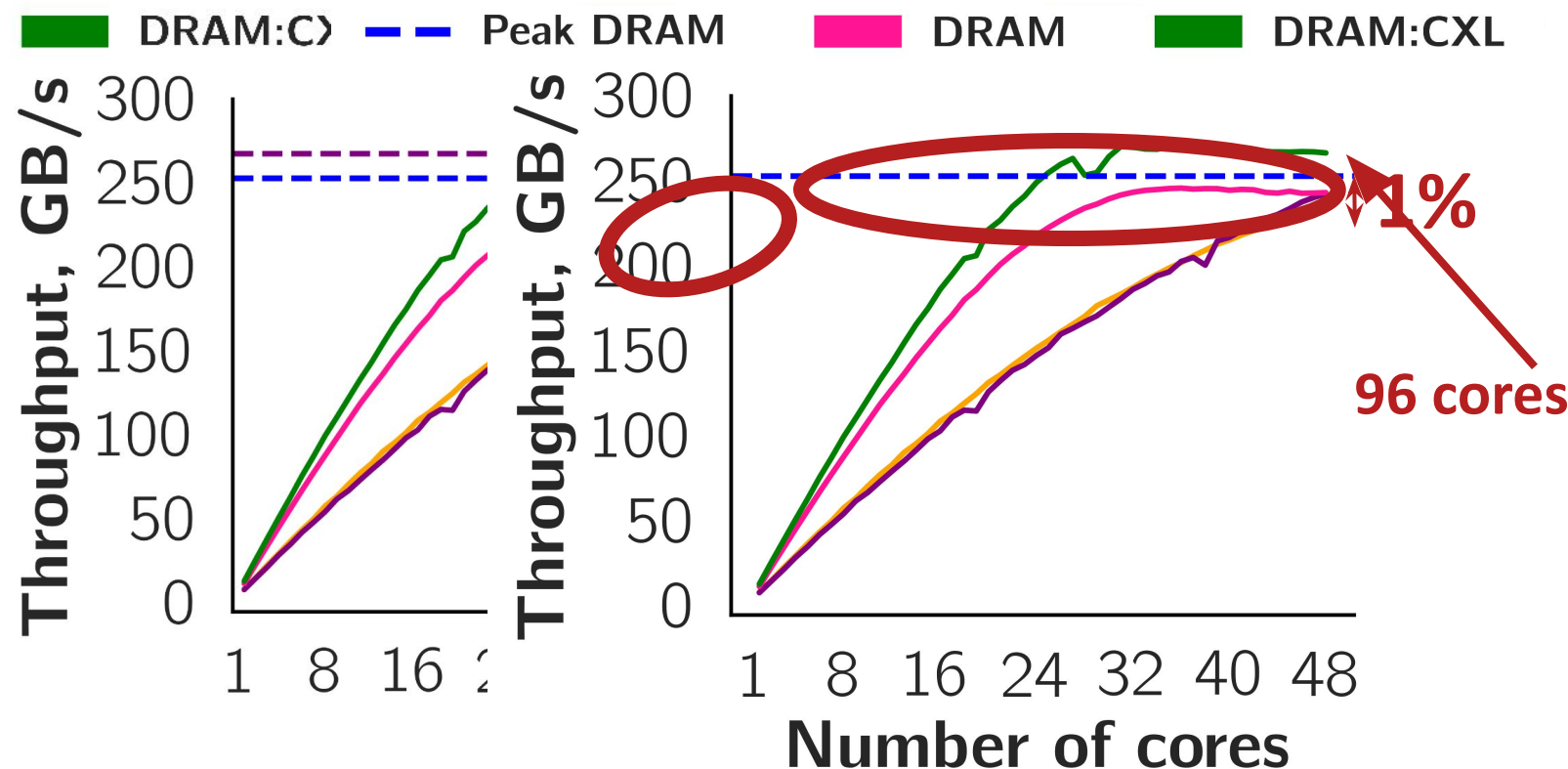
CXL: 41 GB/s

DRAM:CXL interleaving at 19:1

Sequential access: 32 GB array, 8-byte integers

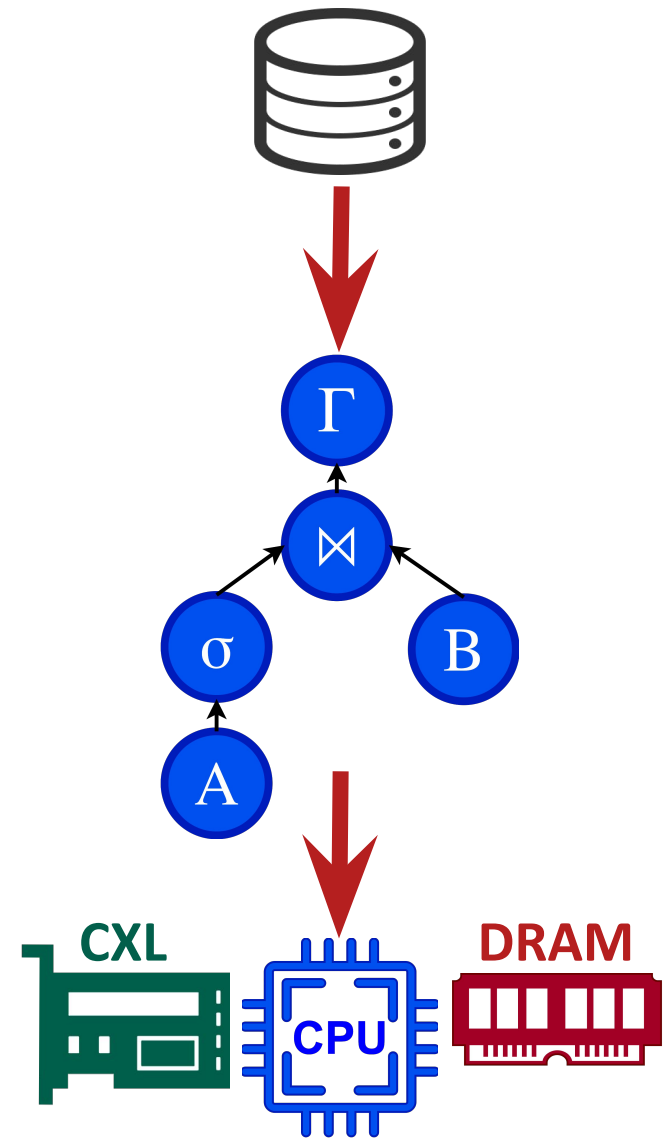
Peak DRAM DRAM SIMD

- Aggressive prefetching
- SIMD instructions
- Hyper-threading



Poor memory-level parallelism diminishes performance gains

CXL memory bandwidth expansion takeaways



- **Transparent DRAM:CXL interleaving degrades performance**
 - SSB workload performance degrades by 20%
- **Workload adaptivity is essential for optimal performance**
 - Optimal interleaving weights vary by 88%
- **High degree of memory-level parallelism is required**
 - Otherwise performance gains diminish from 10% to 1%

Realizing CXL potential requires workload and hardware consciousness ¹¹

Backup

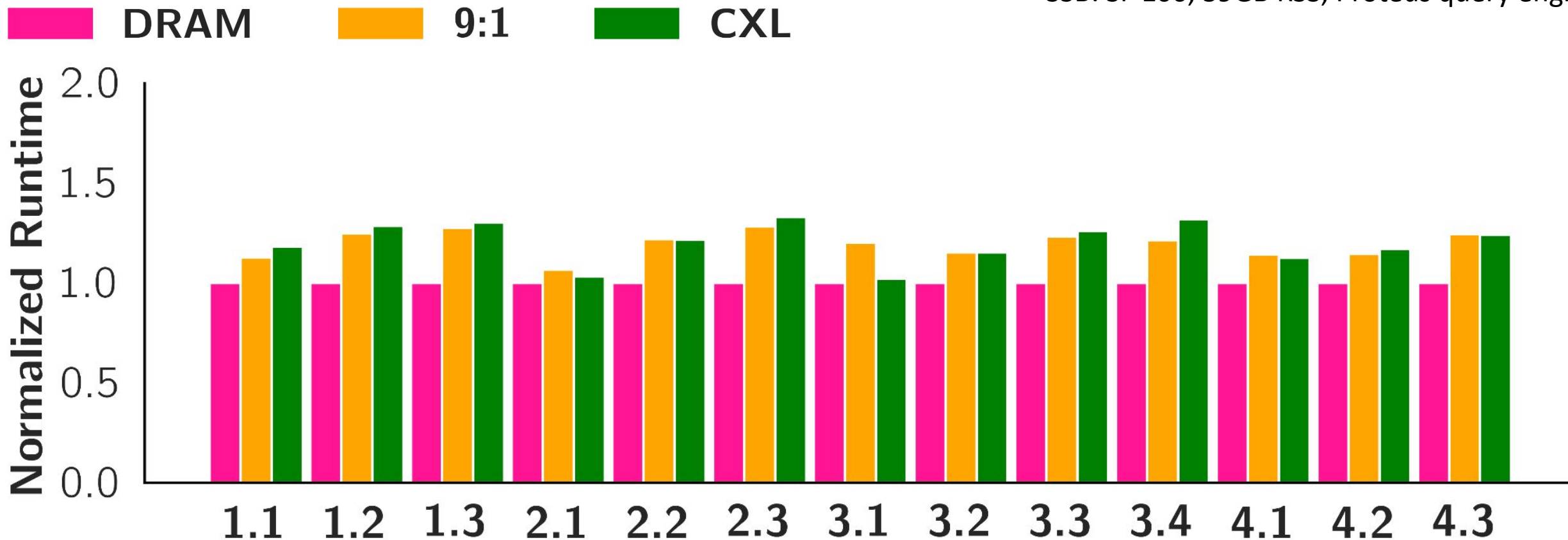
SSB query runtimes

CPU: 48 cores, 1 socket

DRAM: 253 GB/s

CXL: 41 GB/s

SSB: SF 100, 59GB RSS, Proteus query engine



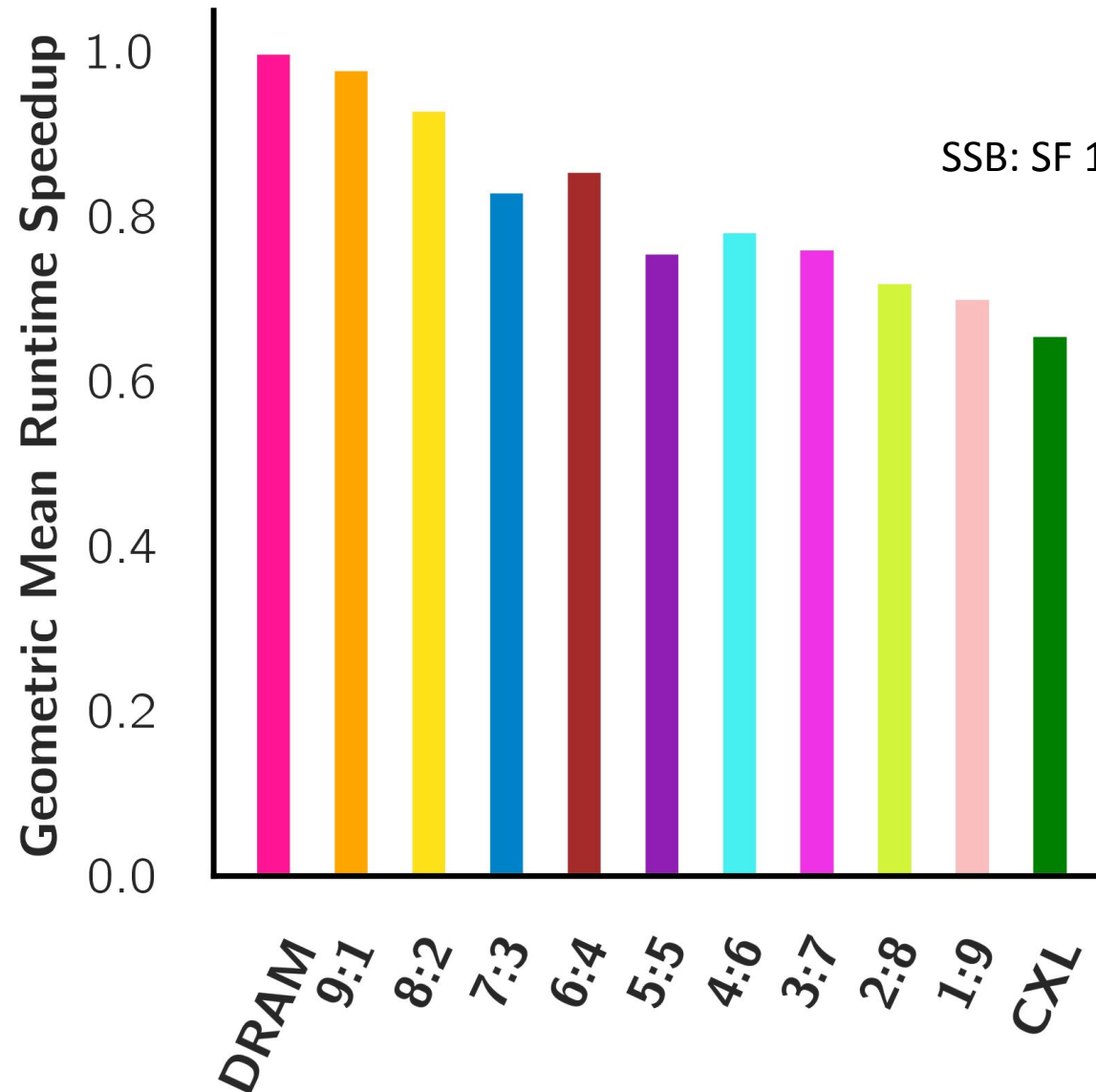
SSB with disabled hardware prefetching

CPU: 48 cores, 1 socket

DRAM: 253 GB/s

CXL: 41 GB/s

SSB: SF 100, 59GB RSS, Proteus query engine



Radix partitioning

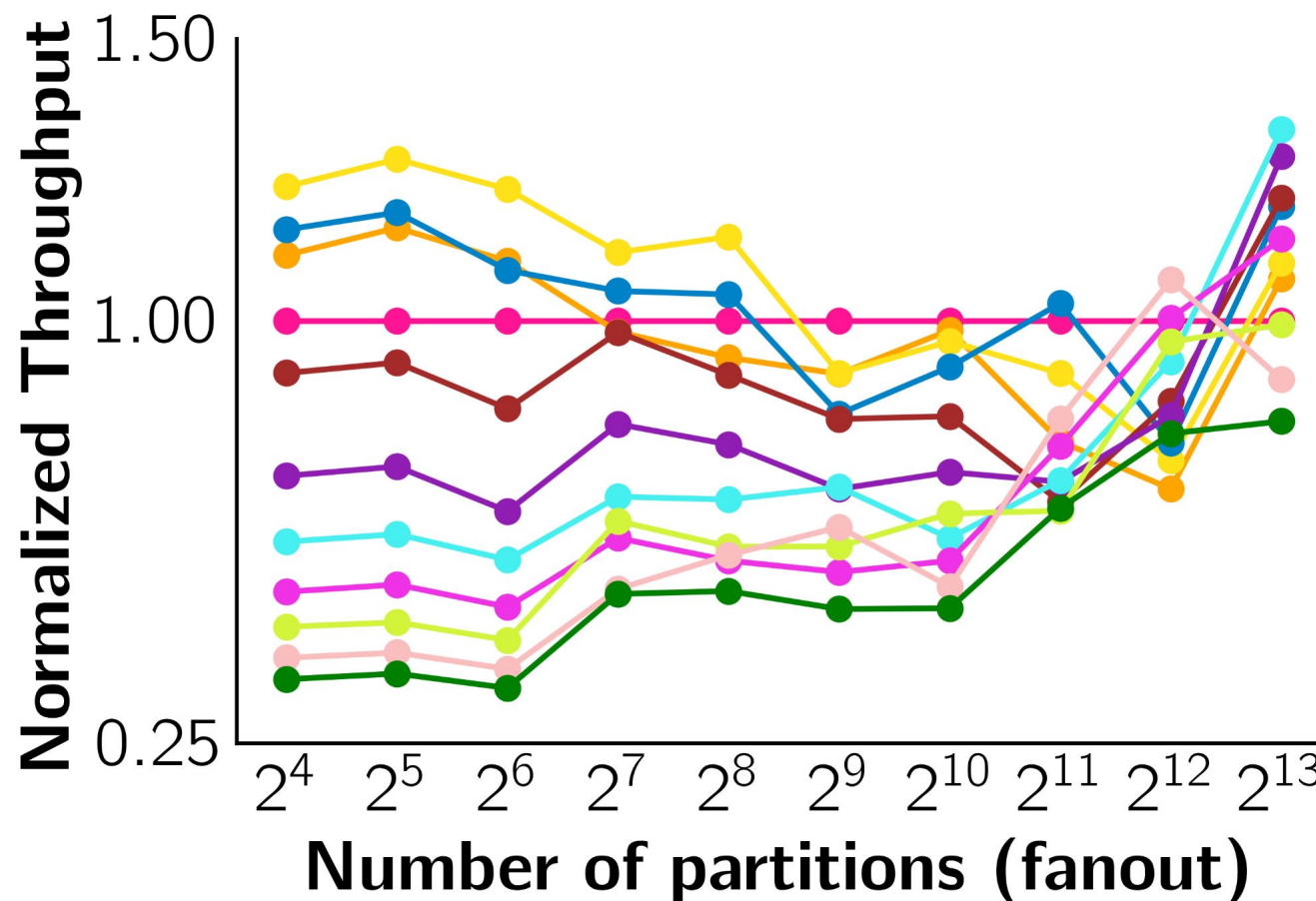
CPU: 48 cores, 1 socket

DRAM: 253 GB/s

CXL: 41 GB/s

SSB: SF 100, 59GB RSS, Proteus query engine

■ DRAM
 ■ 9:1
 ■ 8:2
 ■ 7:3
 ■ 6:4
 ■ 5:5
 ■ 4:6
 ■ 3:7
 ■ 2:8
 ■ 1:9
 ■ CXL



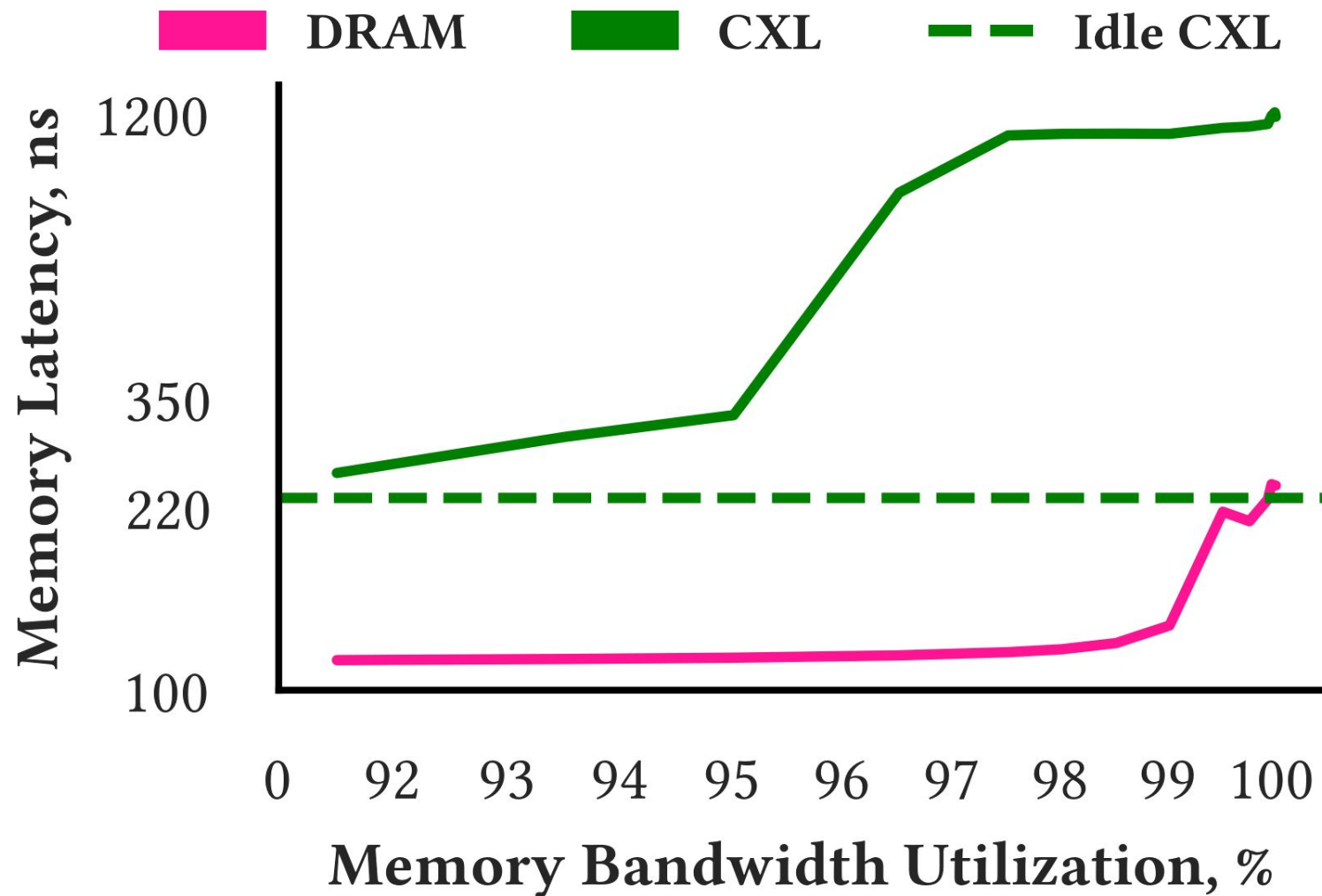
Latency under high bandwidth utilization

CPU: 48 cores, 1 socket

DRAM: 253 GB/s

CXL: 41 GB/s

Intel MLC Loaded Latency

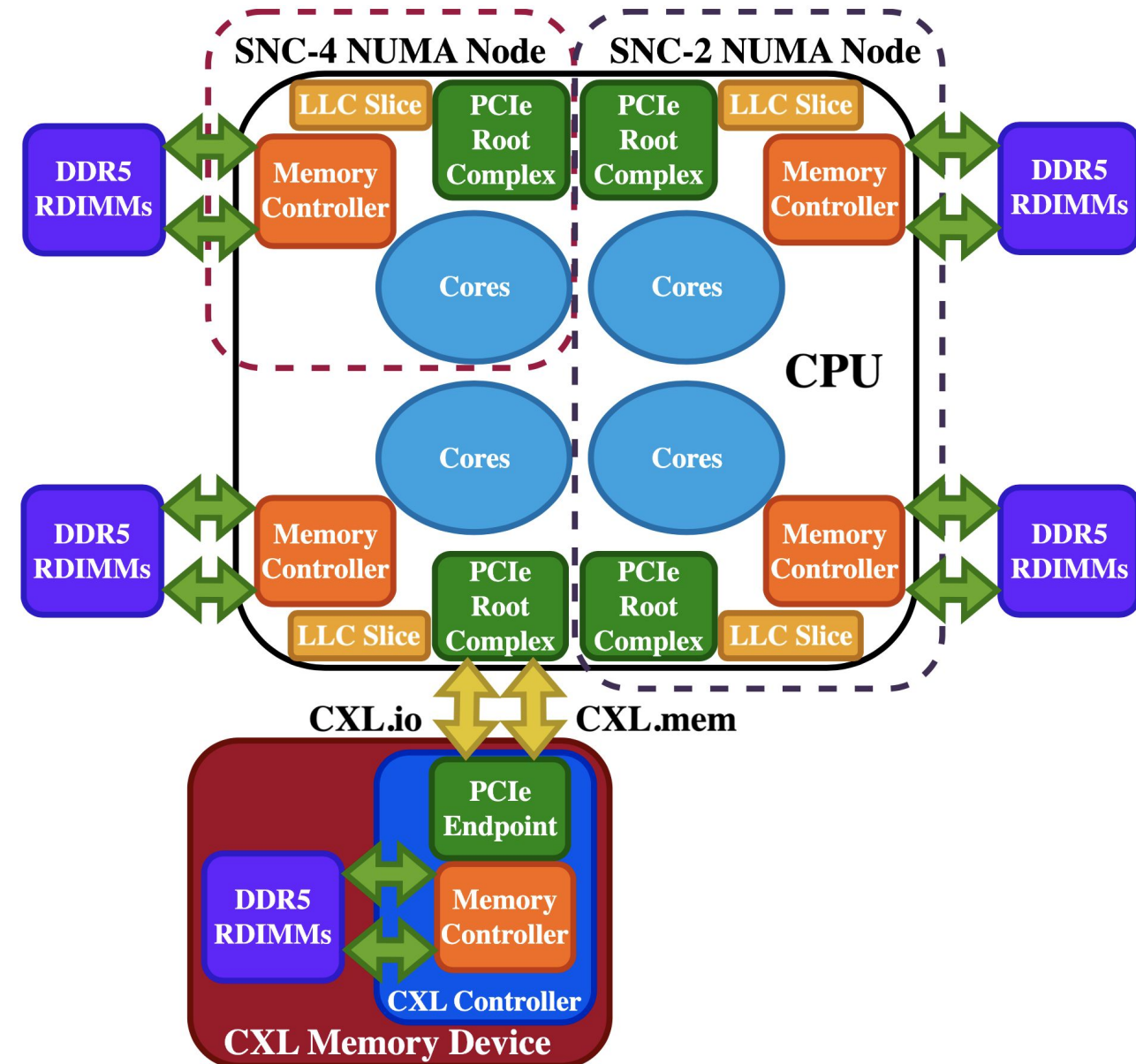


Testbed specs

- CPU: Intel Sapphire Rapids Xeon® Platinum 8468 with 48 cores, 210 MB LLC
- DRAM: 8x32GB DDR5, 253 GB/s, 111 ns
- CXL: SMART® CXA-4F1W with 4x64GB DDR5 over x16 PCIe 5.0, 41 GB/s, 230 ns

Sub-NUMA Clustering

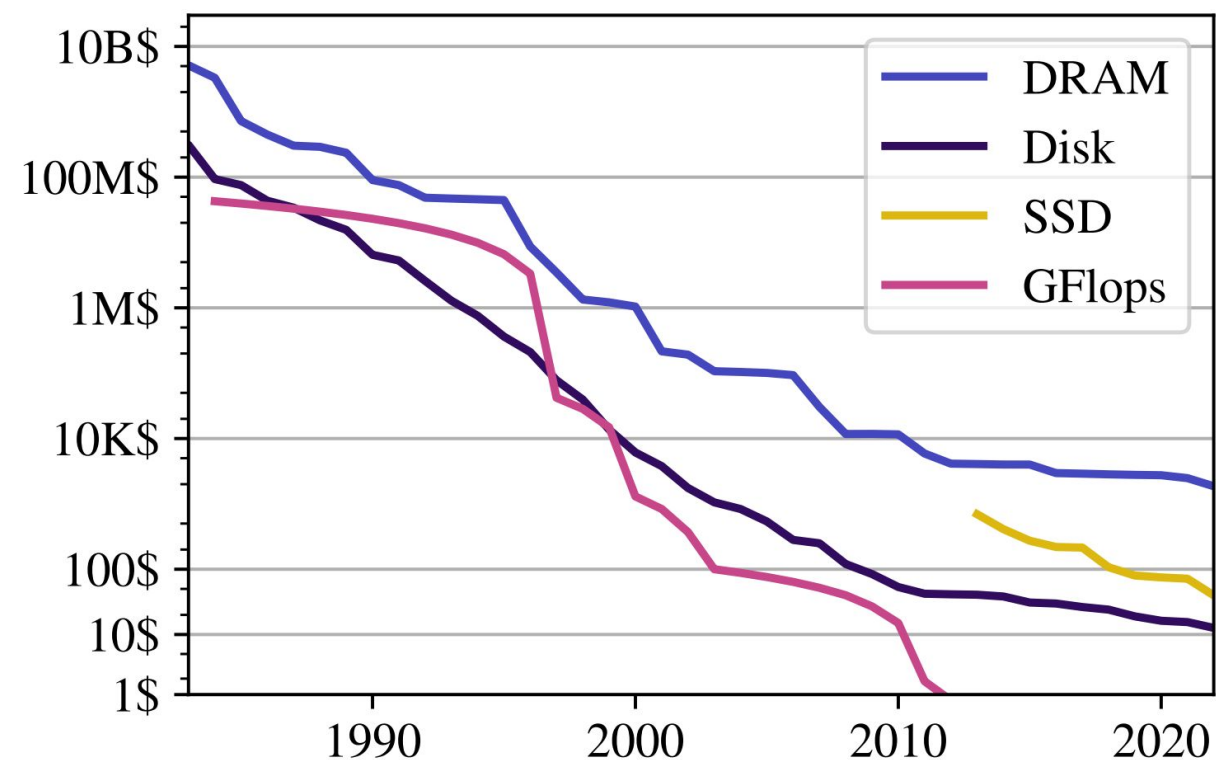
- Full Socket
 - 8 DRAM channels, 253 GB/s
- SNC-2
 - 4 DRAM channels, 126 GB/s
- SNC-4
 - 2 DRAM channels, 59 GB/s
- CXL
 - x16 PCIe 5.0, 41 GB/s



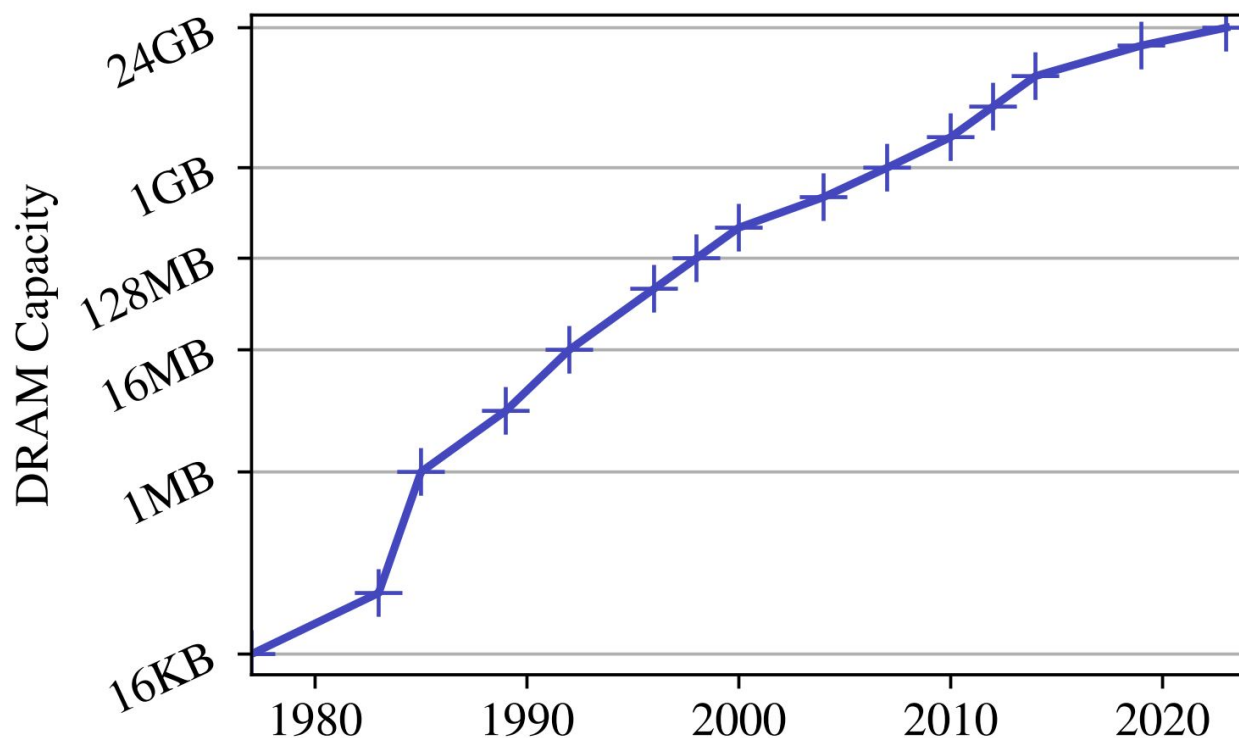
Memory bandwidth per core calculation

Intel Server CPU Generation, Year	Max Core Count	Max Bandwidth, GB/s
Cascade Lake, 2019	56	$2933 \times 10^6 \times 8 \times 12$
Ice Lake, 2020	40	$3200 \times 10^6 \times 8 \times 8$
Sapphire Rapids, 2022	60	$4800 \times 10^6 \times 8 \times 8$
Granite Rapids, 2024	128	$6400 \times 10^6 \times 8 \times 12$

DRAM cost and capacity stagnation

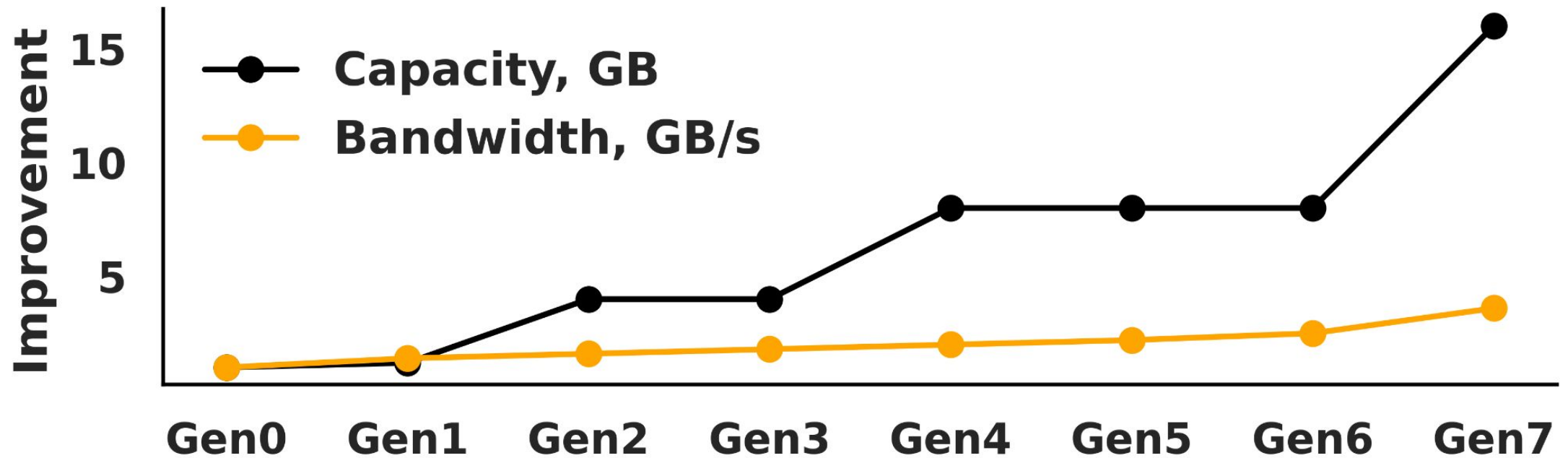


Databases in the Era of Memory-Centric Computing



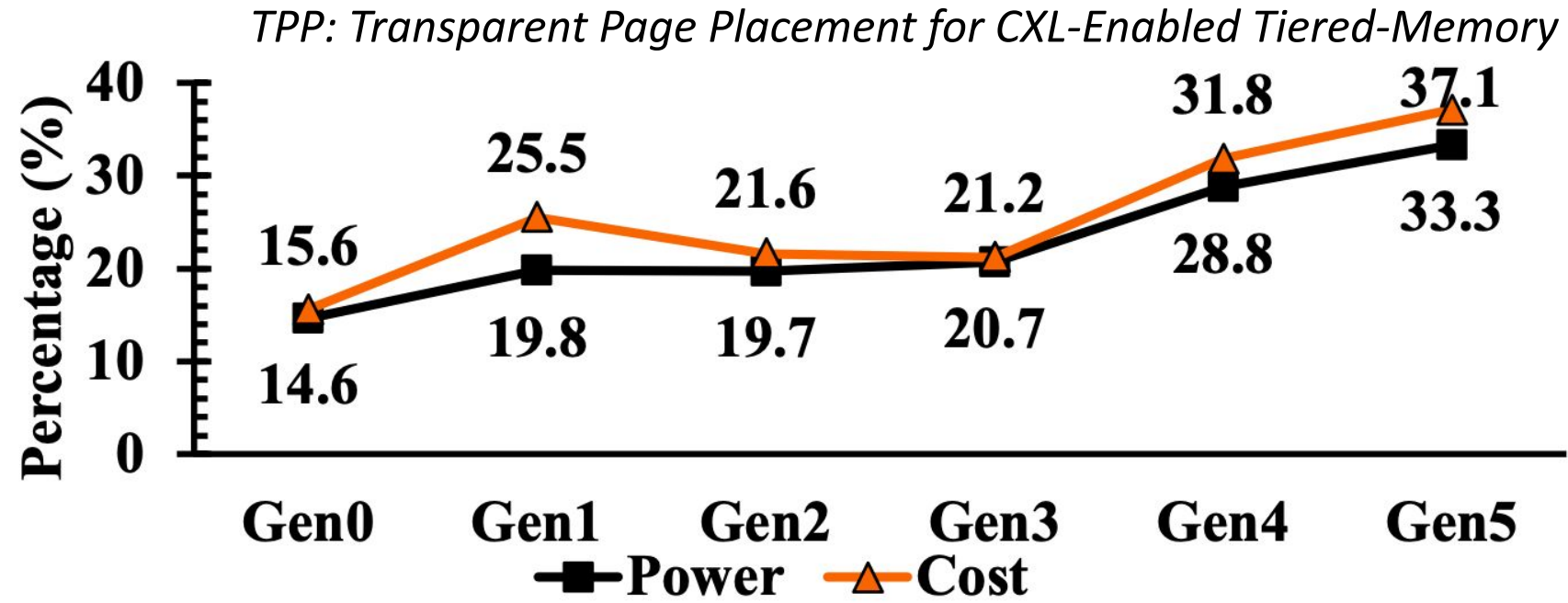
Memory trends for several generations of servers at Meta

Hasan Al Maruf et al. TPP: Transparent Page Placement for CXL-Enabled Tiered-Memory



Memory dominates TCO

- 40% for Meta
- 50% for Azure
- Similar for Google



Memory optimization is largest factor in reducing TCO