Use CXL, not DDR, for Scalable Server Processors

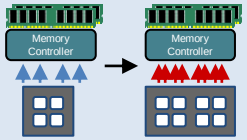
Albert Cho, Anish Saxena, Srikar Vanavasam, Moinuddin Qureshi, Alexandros Daglis
Georgia Institute of Technology

Georgia Tech College of Computing
Center for Research into Novel Computing Hierarchies

Memory Bandwidth Bottleneck

Higher core counts improve workload consolidation and cost efficiency

Increasing only core count:
- Higher contention at memory
  → Increased queuing delay for memory accesses
  → Longer processor stalls

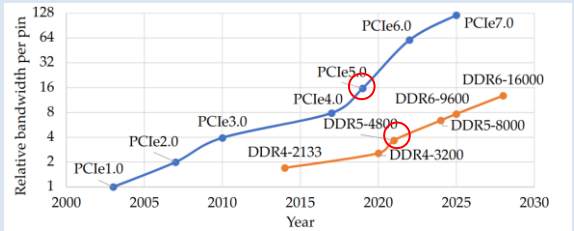


Want to match available memory BW with core count
- Signal integrity limits controller frequency
- Processor pin count limits number of channels

Challenging to scale available memory BW with DDR

CXL as an Alternative to DDR

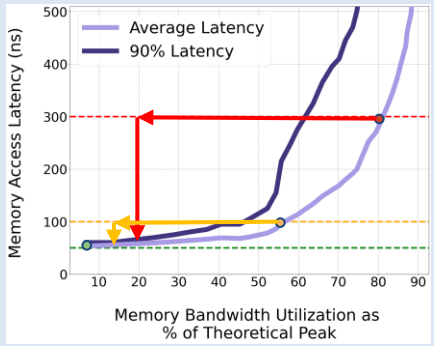
Compute Express Link (CXL)
- Rising interconnect Standard
- Unified solution for wide range of devices
- PCIe as underlying physical layer
- Expecting industry-wide adoption
- Type-3 (memory) CXL devices:
  - Attach DDR-based memory over PCIe
  - Provides load/store semantics
  - Currently 4x BW per pin vs DDR, expected to grow
  - 30~70ns latency penalty vs DDR



** Projected BW support for DDR and PCIe*

Memory Bandwidth, Queuing Delay, and Latency

Latency vs Bandwidth Utilization (*DDR5-4800 modeled in DRAMSIM*)

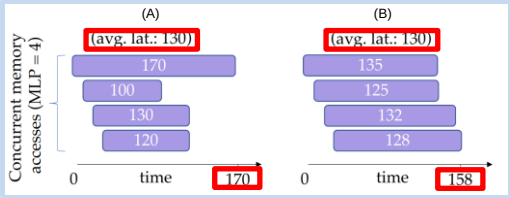


Zero-load Latency: 50ns
Queuing adds 50ns @ 55% util.
Queuing adds 250ns @ 80% util.
- 80% of total latency

Replacing DDR w/ CXL that adds 4x BW and 50ns latency penalty:
- 80% util. → 20% util.
  - ~zero queuing delay
  - -250ns + 50ns = 200ns net gain
- 55% util. → 14% util.
  - zero queuing delay
  - CXL breaks even on latency

Memory Access Latency Variance Matters

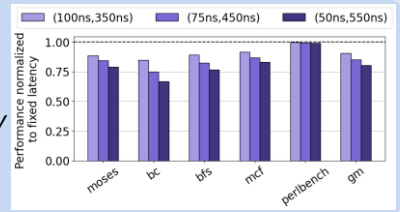
We tend to focus on *average* memory access latency, but *variance* also matters
- i.e., how far tail latencies are from avg
- Slowest access dictates processor stall
- Occur due to temporal access bursts → increased queuing
  - Increasing available BW decreases variance



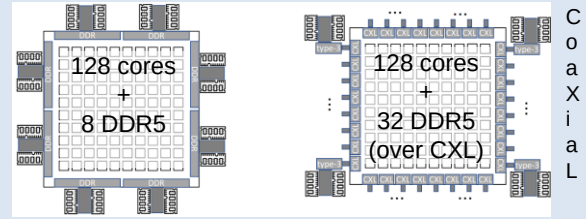
Example:
Series of memory accesses with higher latency variance resulting in longer runtime, despite same avg latency

Controlled simulation with synthetic memory access time shows variance effect on perf.:

- *(x,y): 80%/20% of memory accesses take x/y*
- *Avg latency fixed at 150ns*

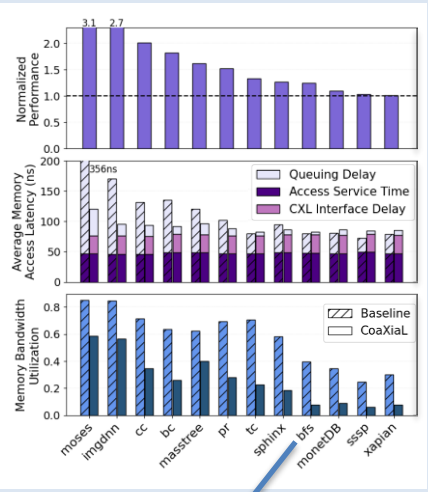


Proposed Design (CoaXiaL)



- Each DDR5 channel replaced by 4 CXL channels (4x BW boost)
- Each CXL channel connected to a type-3 device that supports one DDR5 memory channel

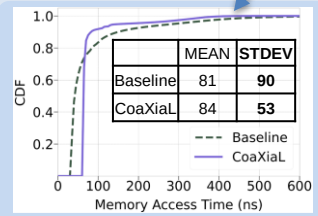
Evaluation Results



Simulation results with 4x BW and 30ns latency penalty for CoaXiaL

1.5x Performance Gain (avg)
- Higher BW mitigates queuing
- Similar gains with 50ns penalty

System utilization	Speedup vs baseline
1%	0.95x
25%	1.00x
50%	1.13x
100%	1.50x



Example workload (BFS):
Memory access latency distribution for CoaXiaL and baseline.

CoaXiaL outperforms baseline even with higher average memory latency, due to lower variance