

Accelerating AI Workloads with Composable Memory & Hardware Acceleration

Klas Moreau, CEO
ZeroPoint Technologies

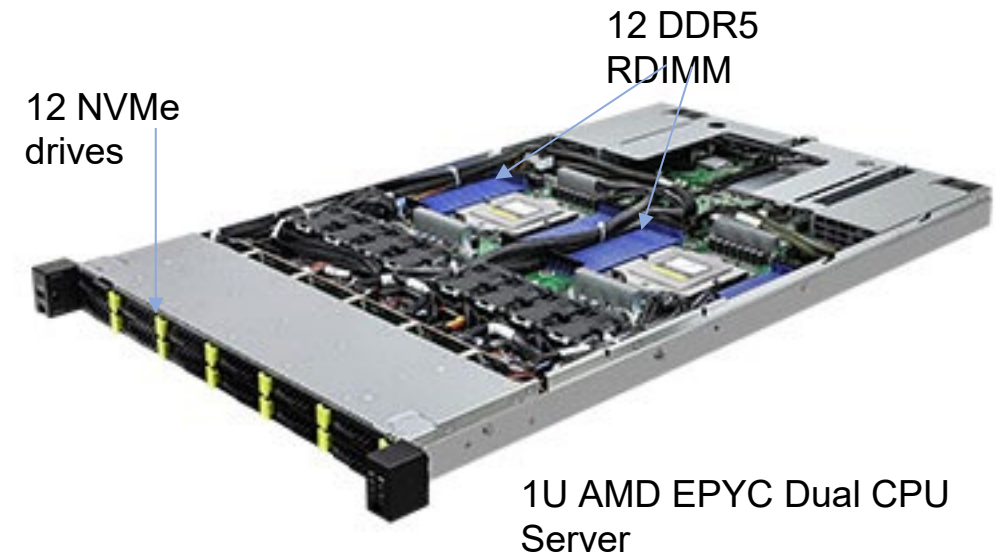
Agenda

- Problem statement: High Cost of Memory in Servers/AI
- Compute Memory Architecture Options
- CXL & Compression Use Case
- DenseMem™: Memory Compression IP
- Compressibility of AI Workloads

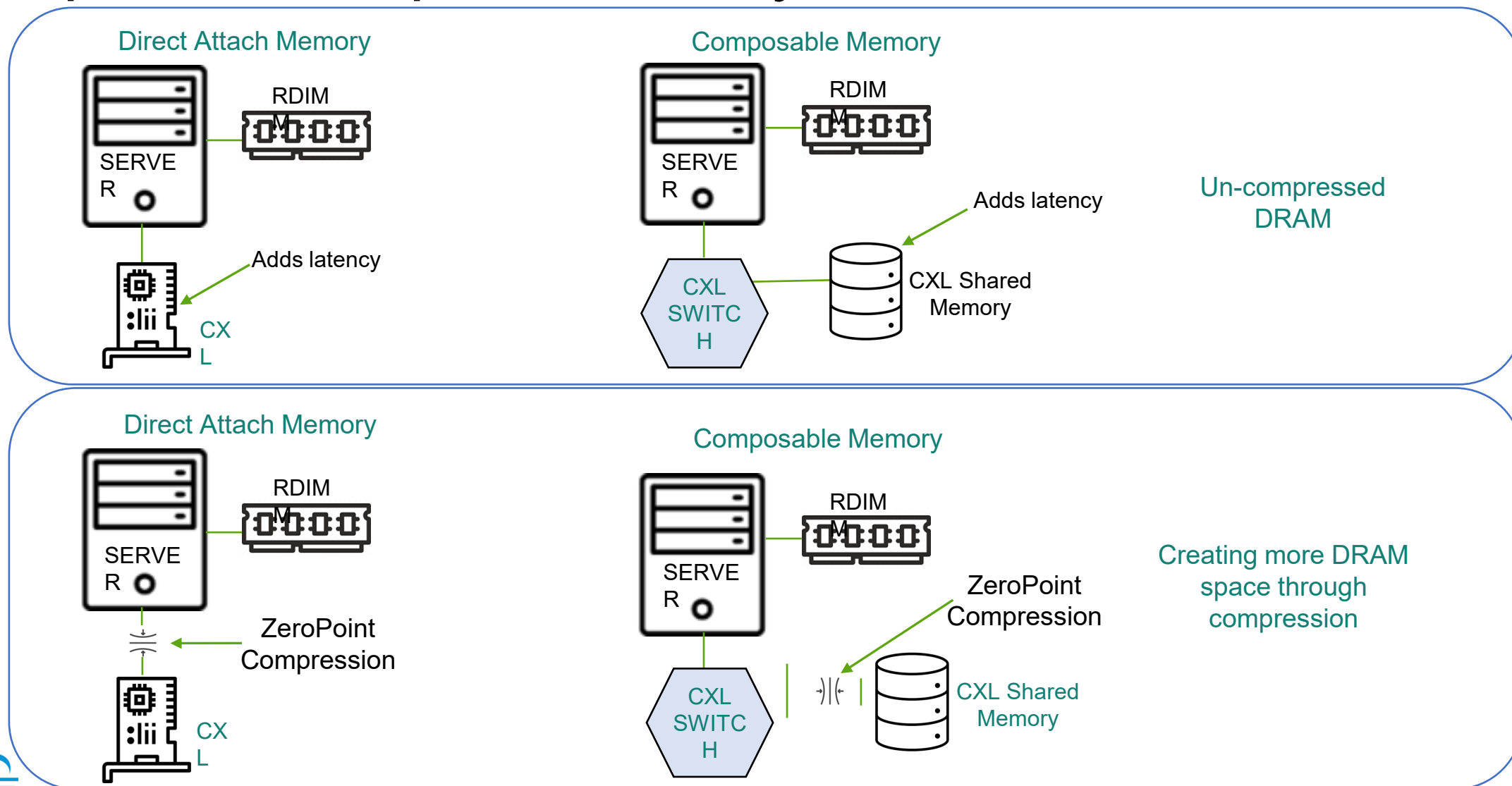
Problem Statement: High Cost of Memory in Servers/AI

- High server cost – largest contributor is memory.
- Inefficient use of memory in data centers:
 - Workloads that require different memory capacity and/or bandwidth.
 - Option to increase capacity/BW via Non-Uniform Memory Access (NUMA) hop.
 - When CPU 0 accesses memory attached to adjacent CPU 1.
 - A need for multiple copies of data.
- Varying compute demands.
- Localized power & thermal density.

Stranded Memory Savings Example
Memory cost: ~50% of server cost
Stranded memory: ~40% of total memory
Eliminating stranded memory saves: ~20% Server Cost reducing TCO (Total Cost of Ownership)



Proposed Compute Memory Architecture (to reduce TCO)

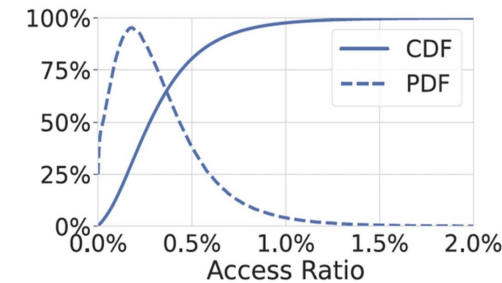
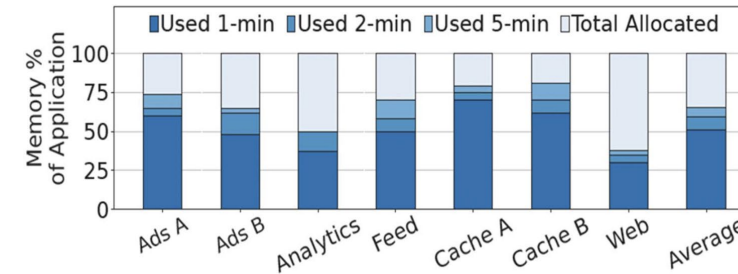


Balancing Memory Latency & Data Center TCO

- Using composable memory is a path to reducing total cost of memory in a data center.
 - Memory sharing helps minimize stranded memory and reduces and balances power consumption.
- A primary challenge with composable memory is the increase in latency.
- Industry architecture improvements to address latency challenges include:
 - Developing additional tiers in memory to balance the CPU demand for data.
 - Lowering latency data transfer with new coherent protocol CXL.
 - Addition of new protocol features and re-architecting IP to reduce latency.
 - Leveraging novel compression techniques to reduce TCO.

The Need for CXL Compression

- Half of memory has not been used in the past minute.
- Cold data has good compressibility.
- Cold data only requires ~1% of system memory bandwidth.

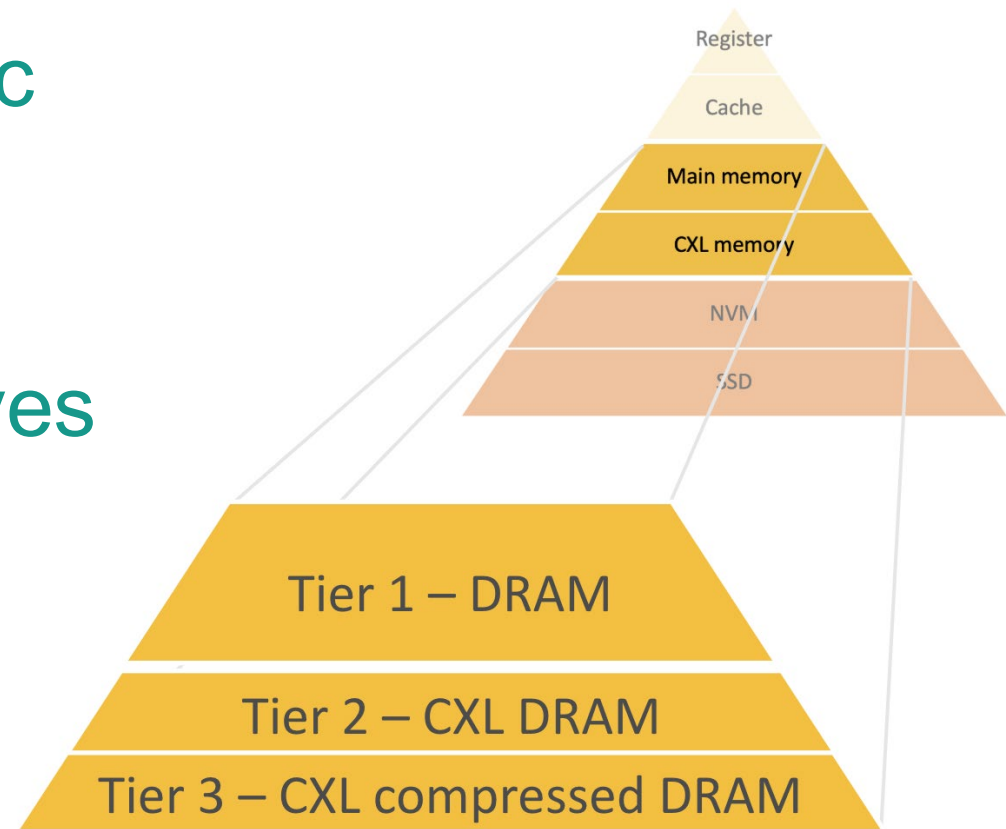


(a) CDF and PDF of tier2 access ratio

Compressed Memory: Cheaper Tier

2:1 compression ratio is realistic in a variety of workloads.

Data with 2:1 compression halves the media cost.



From SW to HW based compression

Data centers are spending capacity on software-based compression.

Meta & Google have stated that a hardware compressed memory tier is a must-have.



CPU cycles used for compression:

4.6% *



Google Cloud

3% **



OPEN
Compute Project

Hyperscale CXL Tiered Memory Expander Specification

Revision 1

Version 1.0

Base Specification Template v1.2
Effective October 27, 2023

*<https://ieeexplore.ieee.org/document/10158161>

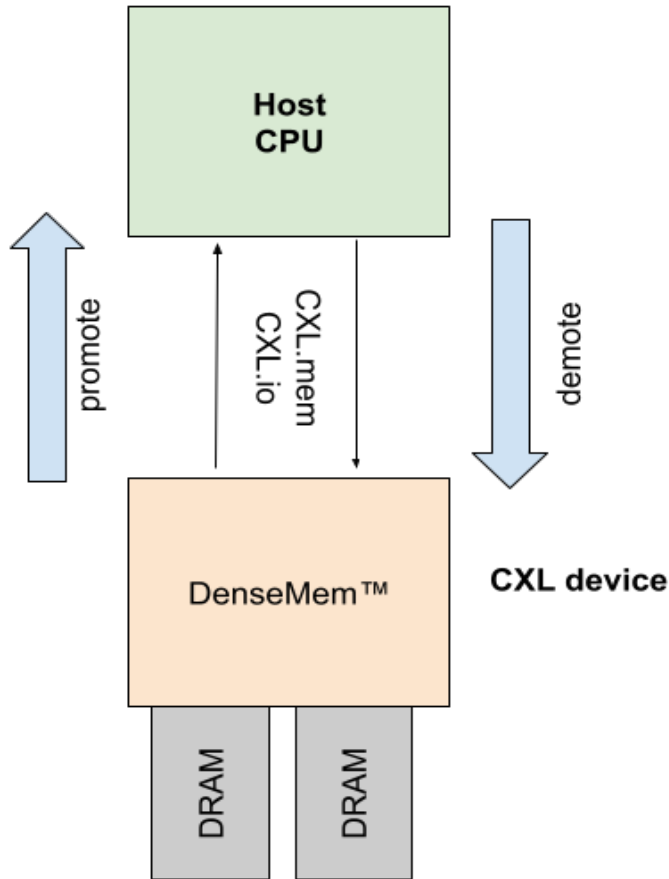
**<https://dl.acm.org/doi/abs/10.1145/3579371.3589074>

Compression IP for CXL Controllers

DenseMem is an IP solution for CXL devices to expand their memory capacity, with:

- Ultra-fast hardware-accel., inline compression and decompression, and
- Real-time compressed memory management.

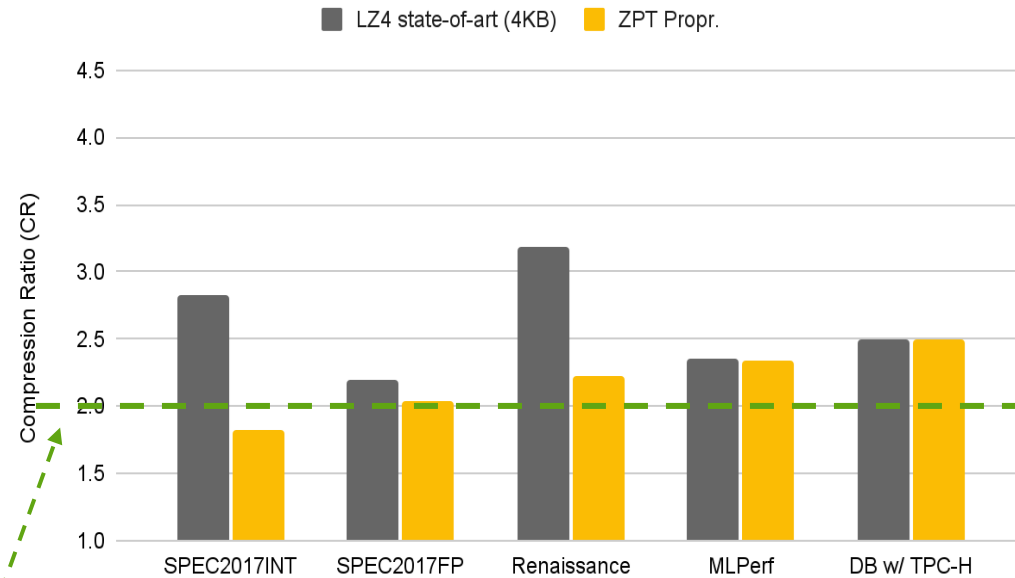
Value prop: Reduce TCO of the data center by 15%



CXL device target	CXL type 3 devices: SLD, (MH-)MLD
Example use scenarios	<p>Dynamic expansion of CXL's memory capacity.</p> <ul style="list-style-type: none">• Compressed memory can be used for storing cold or semi-warm data.• Compressed memory can be used for storing data needing more memory given CXL/compression latency/b/w characteristics.
Functionality exposure to host	<p>Transparent – DenseMem operates on data req/resp issued by host to device over CXL.mem.</p> <ul style="list-style-type: none">• Commands in CXL.io for explicit ctrl (non-critical path).
IP delivery	<p>Soft IP for CXL device ASICs & FPGAs</p> <ul style="list-style-type: none">• RTL, tests framework• Integration support• Firmware running on CXL device• [Host-based software driver]

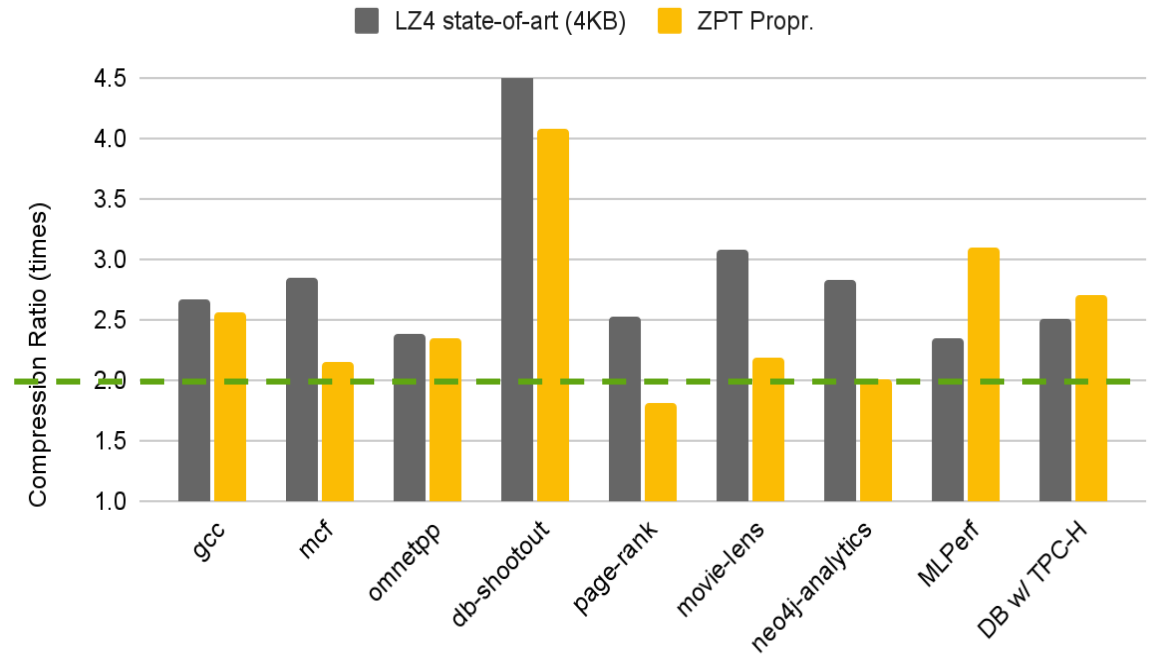
Cache Line Compression

All datasets (Geomean CR across applications of each dataset)



CR = 2x

Our customers are usually more interested in these applications

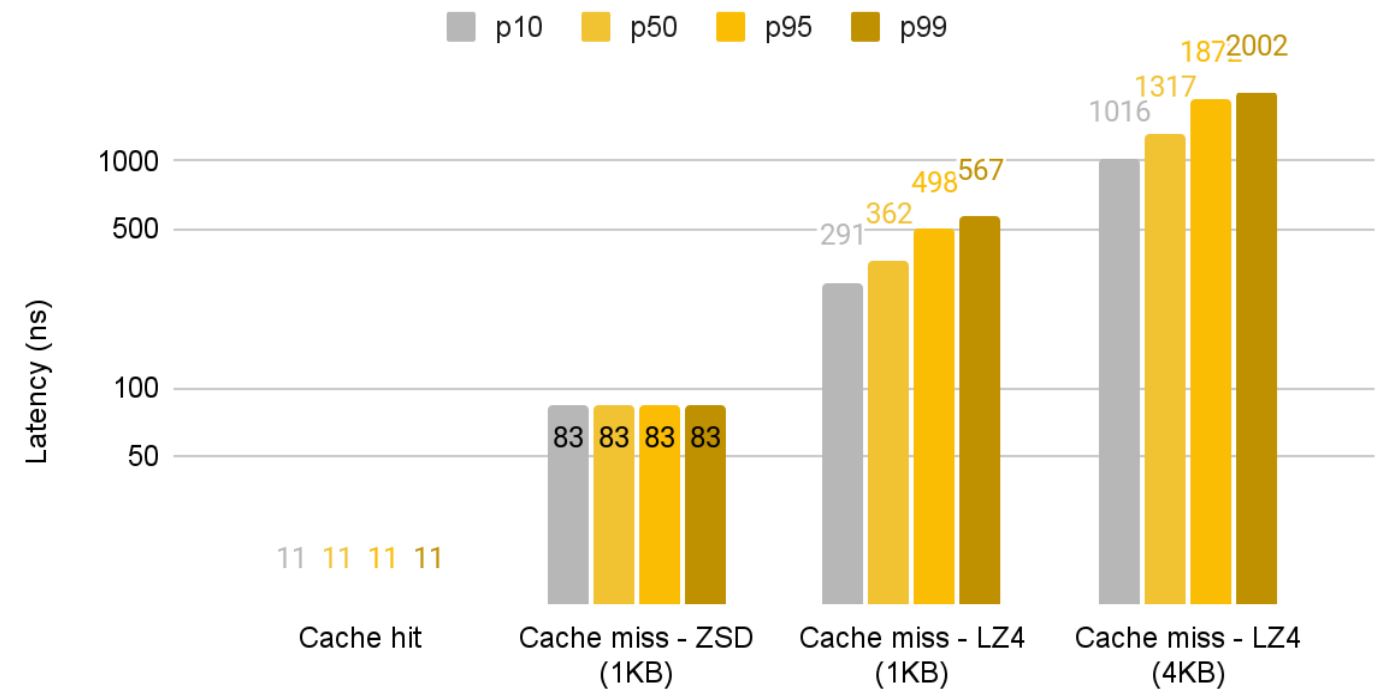


IP supports LZ4 & cacheline compression algorithms selectable at boot
Competitive Compression Ratios, low latency SLA option with cacheline compression.

Latency Breakdown

All IP latencies are measured on RTL.

DenseMem latency block-based -- Unloaded DM req-to-resp latency -- min. block size



* cache: IP cache

1-2 orders of magnitude lower latency
when cache-line compression algorithm used.

Results with Actual Hyperscaler Trace

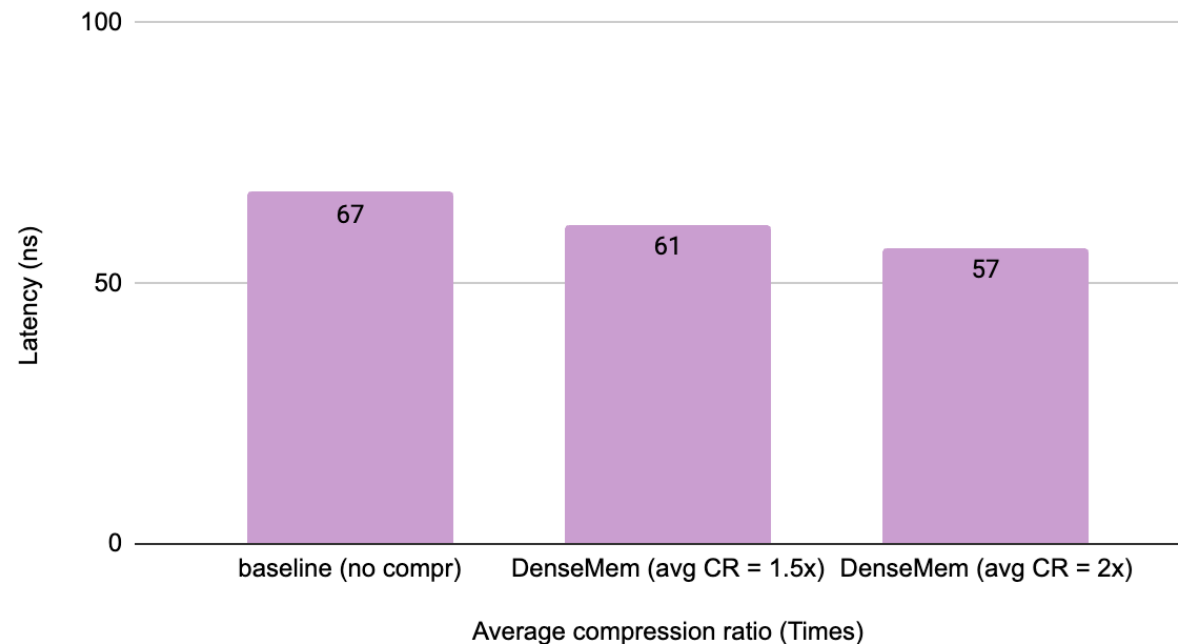
Memory device with DenseMem:

- Memory device latency w/ DenseMem is on par or lower with the memory device latency w/o DenseMem,

and

- At the effect of increased memory capacity.

System latency (DenseMem + DDR4)



* no CXL controller latency

Cacheline Compression: Increased memory capacity with low latency.

Are AI Workloads Compressible?

- Foundational models are optimized during training:
 - Quantization – less accurate weights
 - Pruning – fewer connections between weights
- Lossy compression



Yes!



Block-based Compression Algorithms

Industry Standard Algorithm	Compression Ratio	Block size
LZ4	1.0X (no compression)	64Kb
ZSTD	1.25X (us Latency)	
Deflate	1.25X (us Latency)	
Snappy	0.99X (no compression)	

Cacheline Algorithm

	Compression Ratio	Block size
ZeroPoint	1.5X +	64 byte

Cacheline compression on Foundational Models: 1.5X real time (de)compression with nanosecond latencies.

Compression Alone is Not Enough



Data Compression

ZeroPoint Proprietary Algorithms

- Ultra-fast, Deterministic low latency, suitable for Inline compression
- 2-4x General purpose and Lossless compression



Data Compaction

ZeroPoint Proprietary Algorithms

- Real-time, high performance and low latency
- Cache line granularity



Memory Management

ZeroPoint Developed Driver

- Transparent to operating system and application
- Hardware accelerated

CXL Compression IP: Complete Solution.

ZeroPoint - Altera - Rambus: Better Together

Turnkey Altera FPGA CXL Expander card:

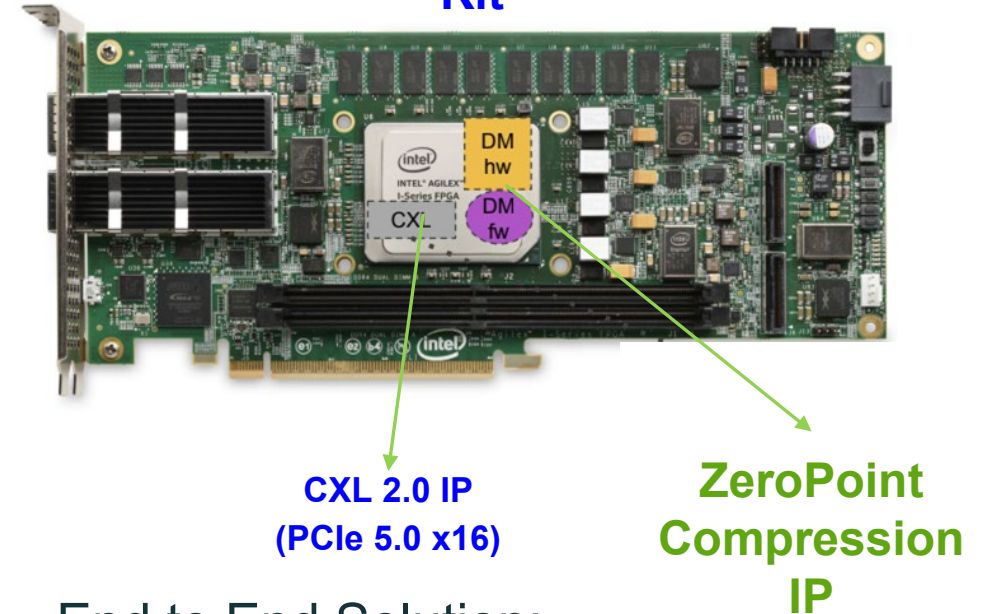
- ZeroPoint (de)compression IP integrated and verified
- DDR4 Memory up to 128GB
- Best FPGA CXL IP solution for highest performance FPGA

Rambus ASIC implementation CXL IP

- ZeroPoint (de)compression IP integrated and verified
- Best solution for highest performance ASIC

Hyperscale-ready integrated platform to test real world workloads.

AlteraTM 7 FPGA I-Series Dev Kit



End to End Solution:

- Compatible with TPP (Transparent Page Placement)
- Host Kernel, software
- Telemetry

Summary & Call To Action

Summary

- High performance CXL IP and compression IP are better together for composable systems
- Integrated FPGA platform with Altera & ZeroPoint IP
- Integrated ASIC platform with Rambus & ZeroPoint IP
- AI & Foundational models practical with CXL



Call to Action

- Sign up for early access to our FPGA prototype!
- Collaborate on host SW integrations:
 - Working with the Linux Kernel community
 - Give us input on your tiering stack
- Bring your AI models to CXL – collaborate!