



SHAPING THE NEXT GENERATION OF ELECTRONICS

JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA



JUNE 23-27, 2024

MOSCONE WEST CENTER
SAN FRANCISCO, CA, USA

OCP ready Hardware Accelerated CXL Memory Compression IP for Data Center Applications

Nilesh Shah

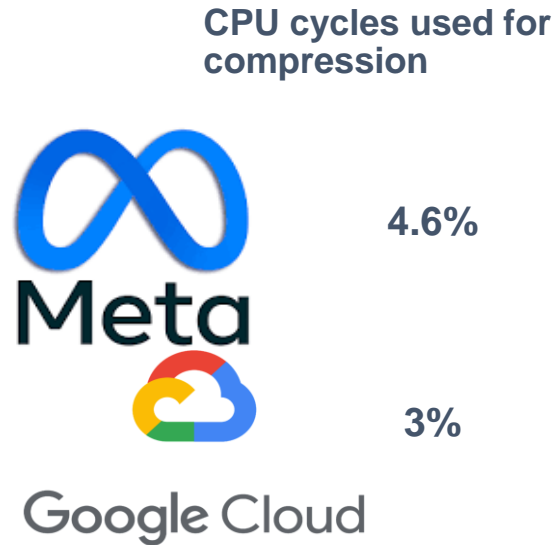
VP Business Development, ZeroPoint Technologies

Nilesh.shah@zptcorp.com



Challenge

Hyperscalers spending significant
\$\$ on software based compression



Hyperscaler requirement:
hardware compression is a MUST-have



Hyperscale CXL Tiered Memory Expander Specification

Revision 1

Version 1.0

Base Specification Template v1.2

Effective October 27, 2023

Opportunity | Add Compressed CXL Memory Tier

Deploy New Tiers: Ordinary + Compressed DRAM memory on CXL

| | Benefit | Spec Component |
|---|--------------------------------------|--|
| 1 | Reduction in total cost of ownership | Standardization, Hardware accelerated, Lossless Compressed Memory Tier |
| 2 | Energy Efficiency, Sustainability | Transparent Hardware accelerated Compression |
| 3 | Preserve Software Investments | Support for legacy Compression Algorithms |

Removes barriers, enables diversity of Hyperscalers + Enterprise Customers



Hyperscale CXL Tiered Memory Expander Specification
Revision 1
Version 1.0
Base Specification Template v1.2
Effective October 27, 2023



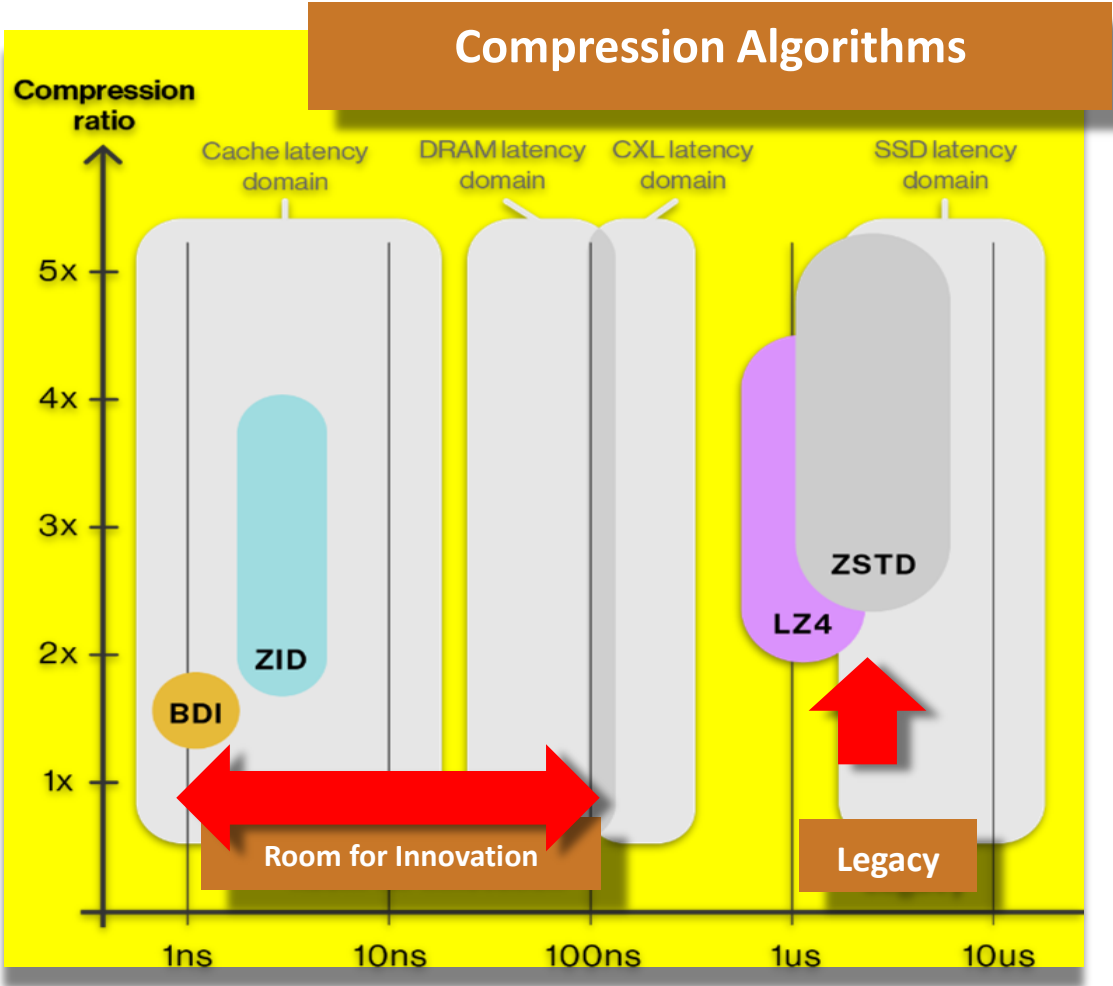
Requirement: Hyperscale CXL Tiered Memory Expander Spec

Preserving Software Investment, without compromising performance

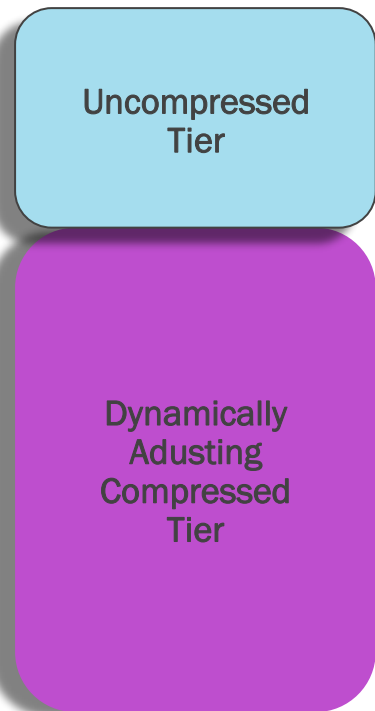
- Support for Legacy Compression Algorithms
- Future Proof , supports Algorithm Innovation
- Stringent Latency & Bandwidth Specifications

| Parameter | Specification |
|-----------------------------|---------------|
| Latency Uncompressed Access | 90 to 150ns |
| Latency Compressed Access | 250ns to <1us |
| Bandwidth Efficiency | 80% / 75% |
| Read only/ Write only | |

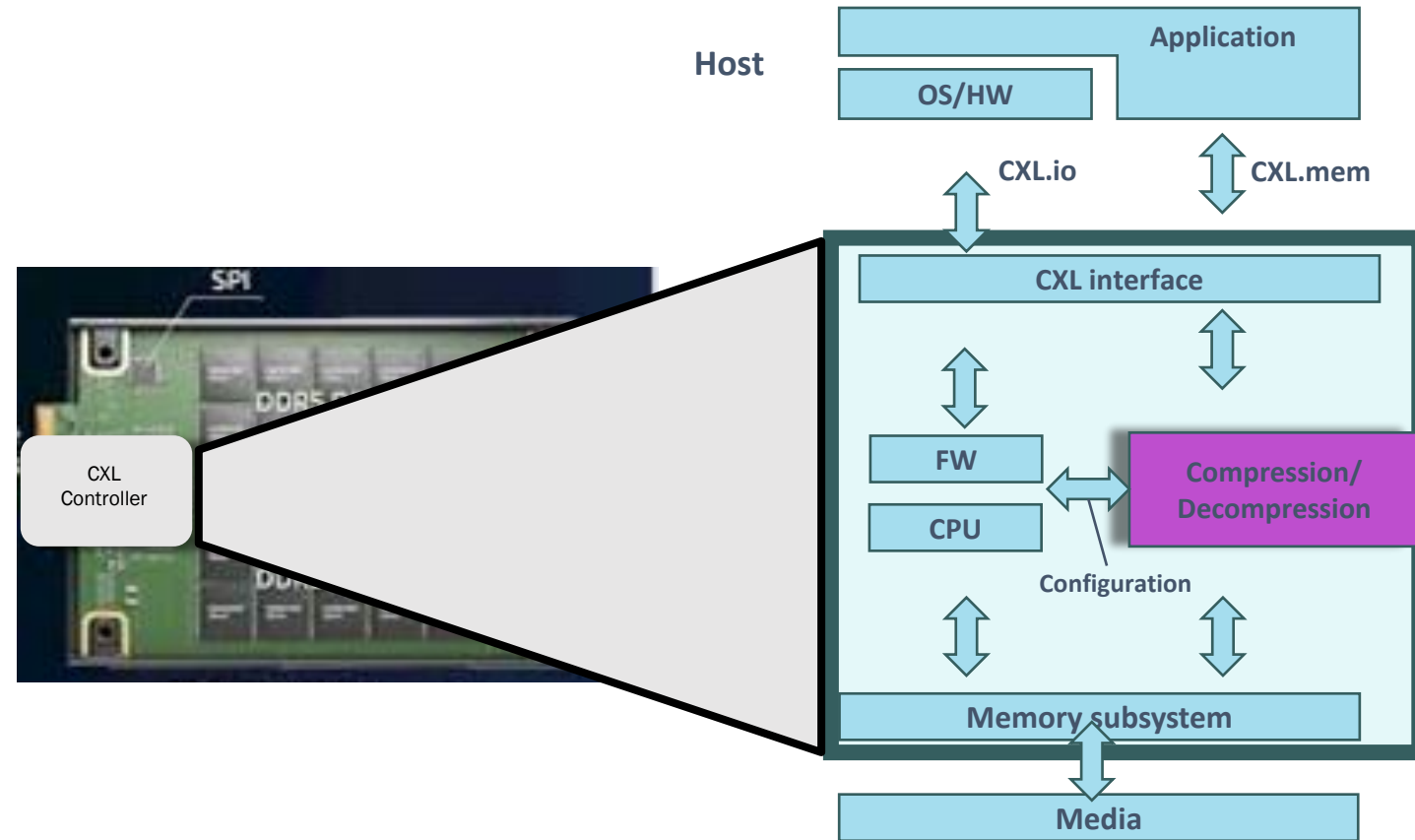
Stringent Requirements



Proposed Solution: CXL Controller with integrated Hardware Acceleration

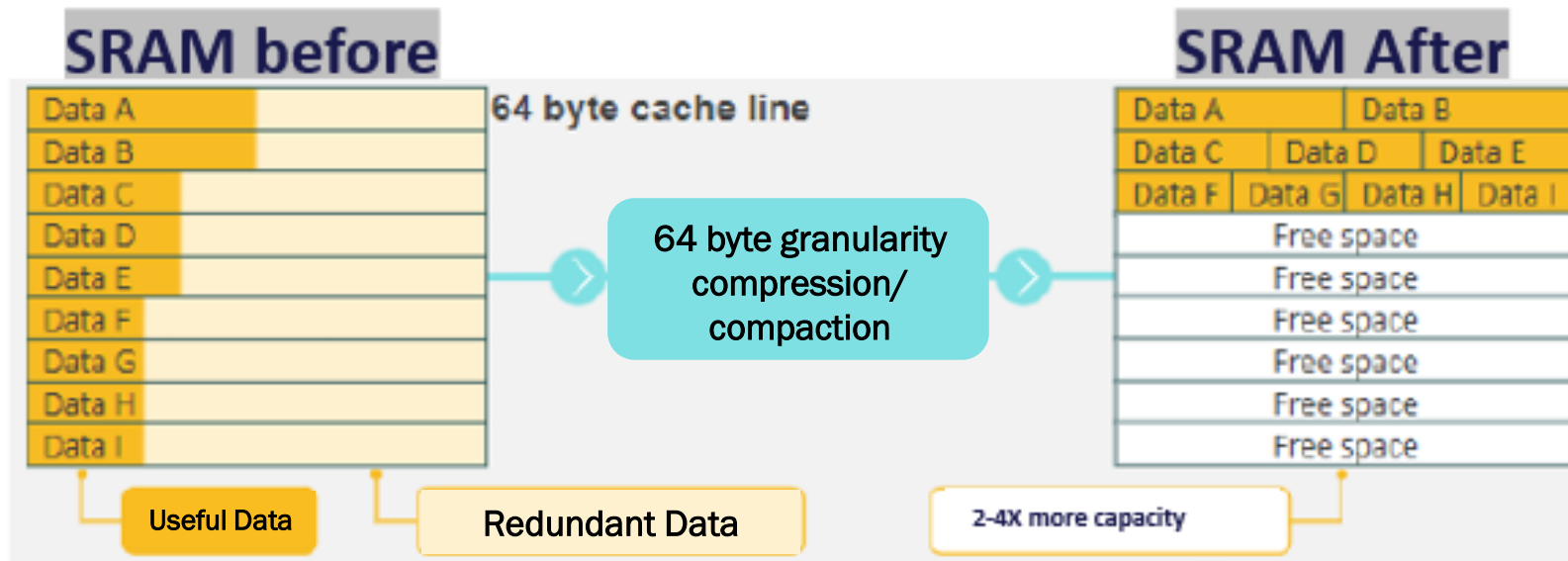


CXL Type 3 Device Address Space



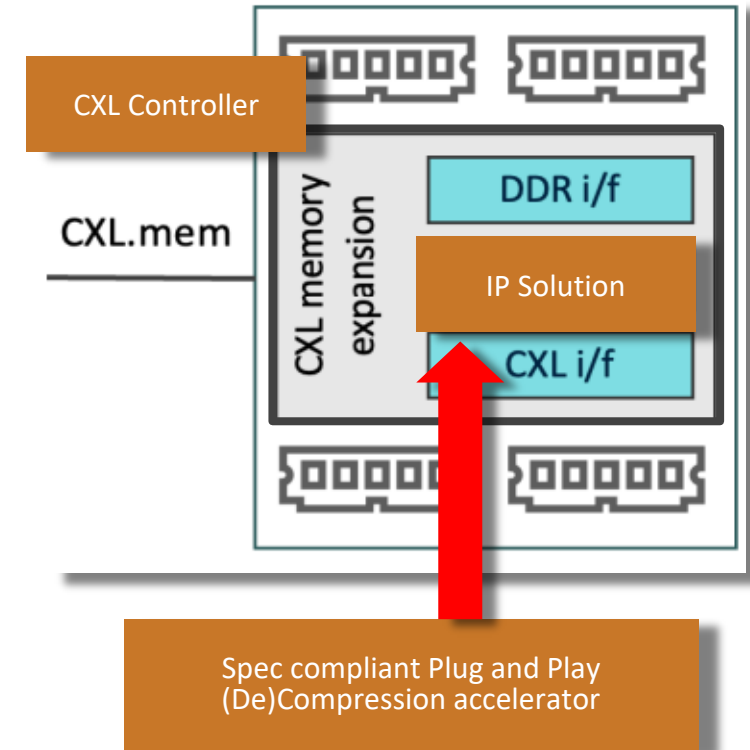
Proposed Solution: Cacheline Compression

- 2-4X Compression Ratio



Solution | Portable, Integrated IP

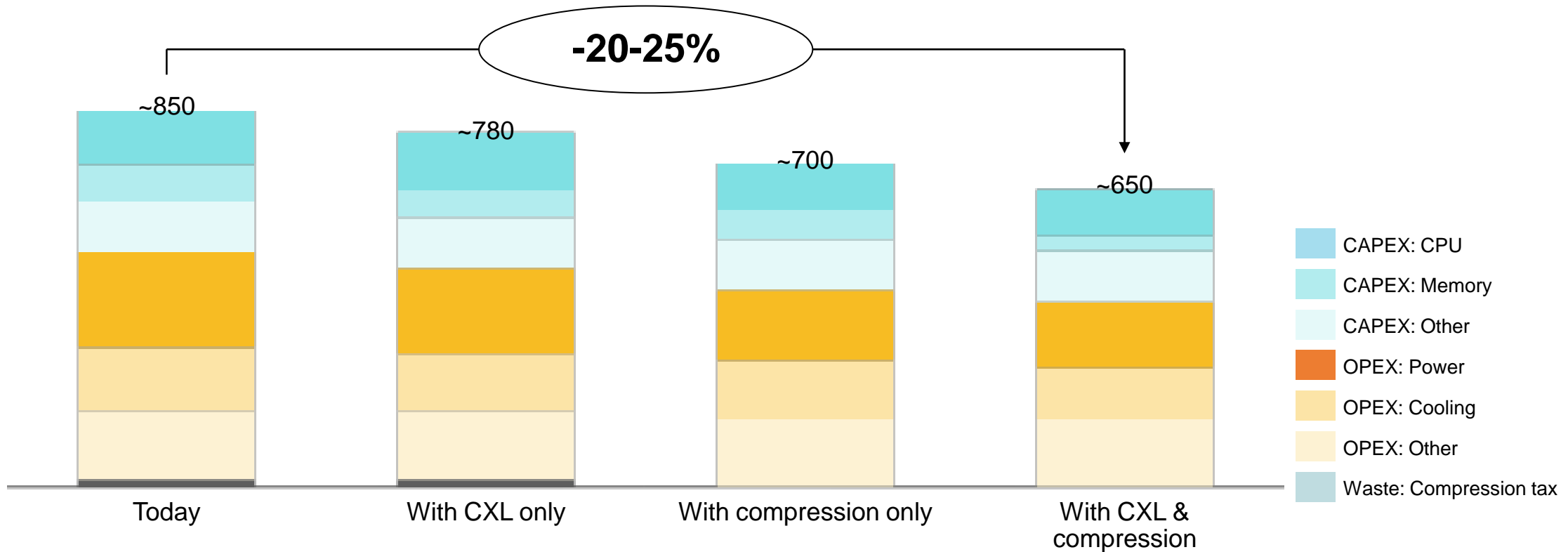
- OCP Spec compliant Hardware Accelerated CXL memory **(De)Compression + Compaction + Transparent** memory management IP block
- **2-4x** transparent (de)compression major Datacenter workloads
- Compression Algorithms: LZ4, ZID (proprietary)
- Portable: AXI4, CHI, Leading process node support
- Verified @ 1.2Ghz , 0.9mm² (at 4nm Samsung)
 - SRAM 75% of the IP solution SRAM.



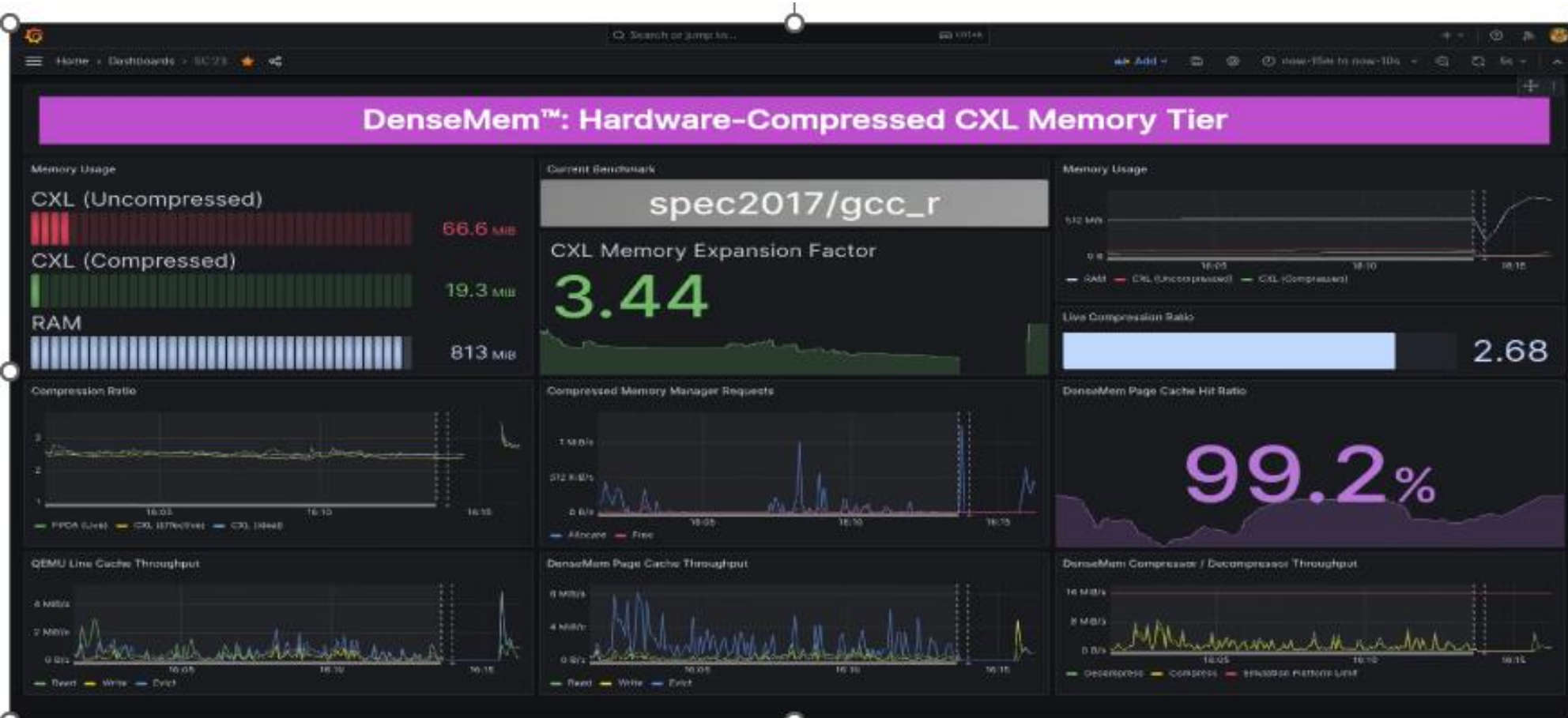
| IP Solution Performance Characteristics | Value |
|---|--------|
| Compression Ratio | 2-3X |
| Block cache (SRAM) hit latency | <30ns |
| Cache line in uncompressed region latency | <90ns |
| Cache line in an uncompressed block latency | <150ns |
| Cache line in a compressed block latency | <250ns |
| Tail latency [cache line in a compressed block] | <1us |
| Decompress bandwidth[4x 1867MT/s] | >46G/s |

Solution | Reduce Data Center TCO 20-25%

Total Cost of Ownership (TCO) for 40 server rack over 3y lifetime [kUSD]



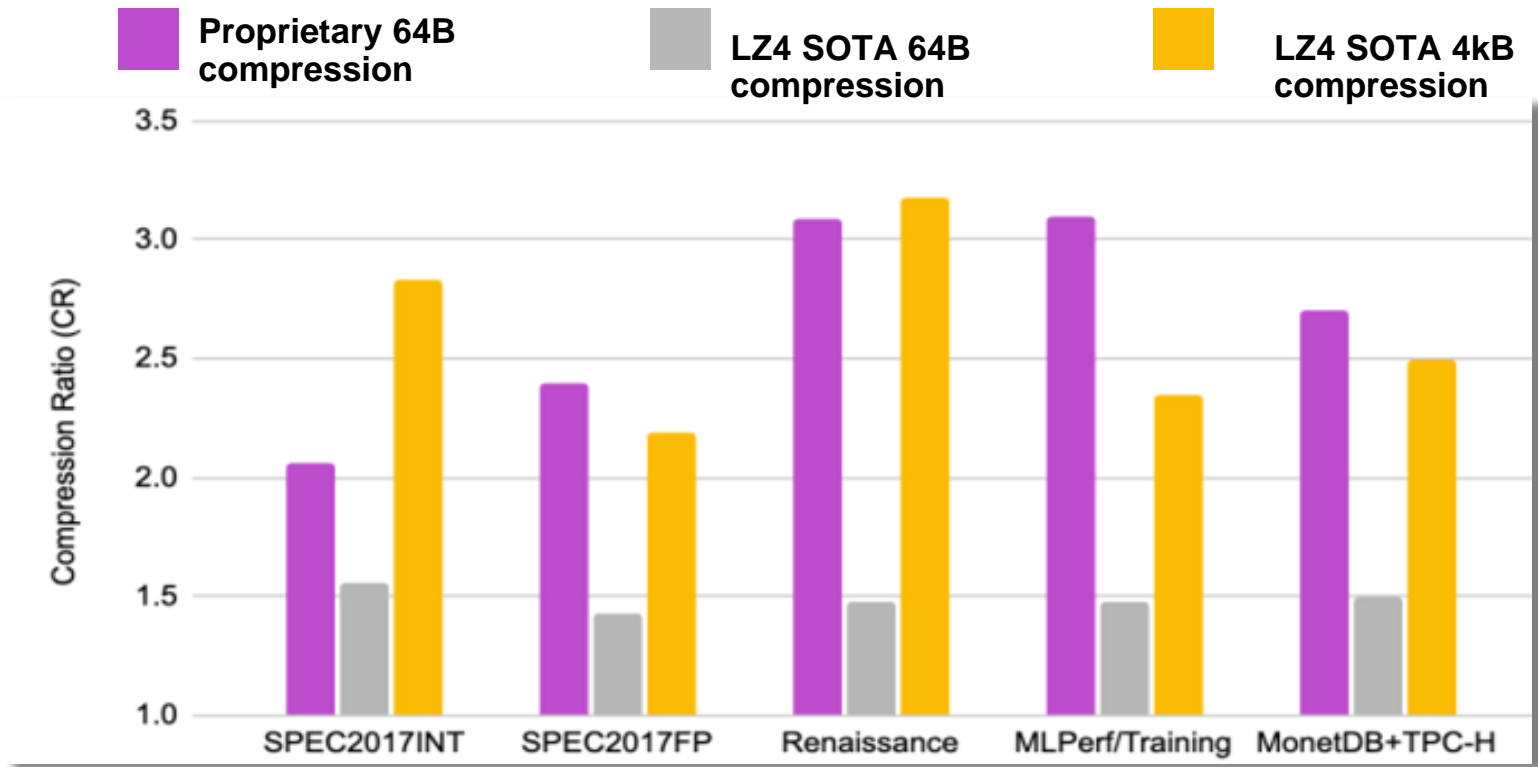
Integrated IP Demo | QEMU + FPGA



[Link to Video](#)

IP Solution | Performance across Datacenter workloads

Geomean Compression Ratio (across applications of each dataset)



Summary | Call To Action

- **Summary**

- Integrated IP Solution: OCP Spec Compliant
- Portable across process nodes, AXI/CHI interface
- Production ready mid '24, Performance verified

- **Call To Action**

- **Controller manufacturers:** Collaboratively address Hyperscale OCP requirements
- **ISVs:** Host software integration, target workloads
- Link Compression and zswap IP
- Additional information ([URL link](#))

Open Discussion/ Q&A