Exploring Innovations in Storage & Memory Compression at Hyperscale

Presenter:
John Kim, Sr. Director, AI Memory Solutions, SK Hynix
Nilesh Shah, VP Business Development, ZeroPoint Technologies

FMS
the Future of Memory and Storage

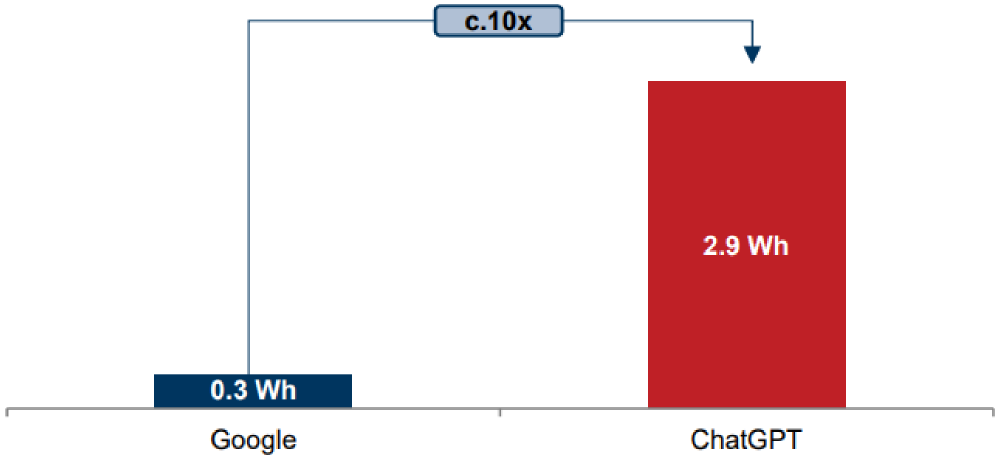
Problem: The AI Energy Efficiency Challenge

Schneider Electric estimate	2023	2028
Total data center power consumption	57 GW	93 GW
AI power consumption	4.5 GW	14.0-18.7 GW
AI power consumption (% of total)	8%	15-20%
AI workload (Training vs Inference)	20% Training, 80% Inference	15% Training, 85% Inference

Source: Schneider Electric White Paper 110, The AI Disruption: Challenges and Guidance for Data Center Design

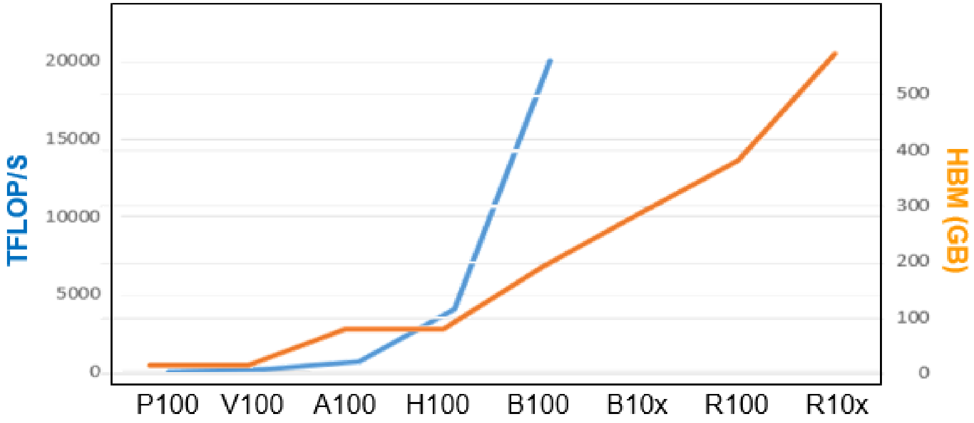
ChatGPT queries are 10x as power intensive as Google searches

Power consumption per query/search, Watt-hour (Wh)

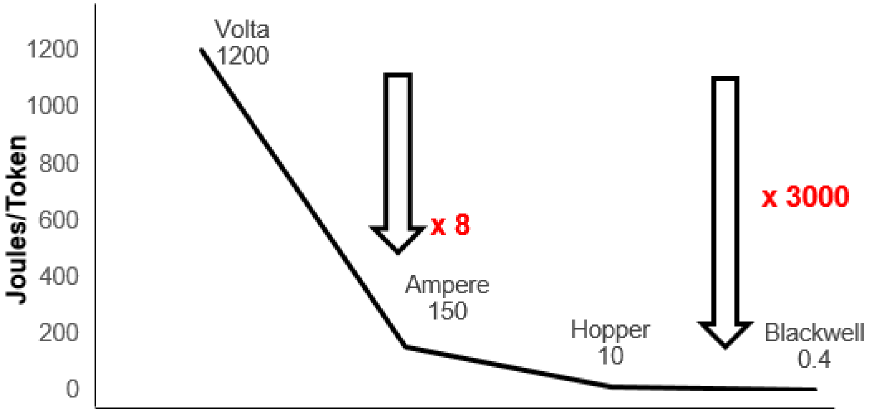


Source: Google, SemiAnalysis, compiled by Goldman Sachs GIR.

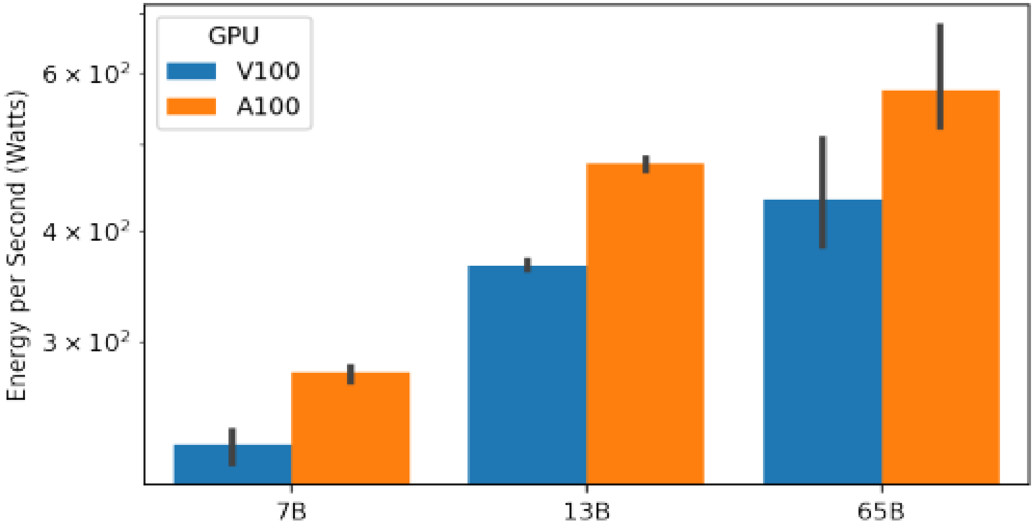
Can GPU solve the energy program?



Nvidia GPU Generation



Nvidia GPU Inferencing Energy (GPT-1.8T)



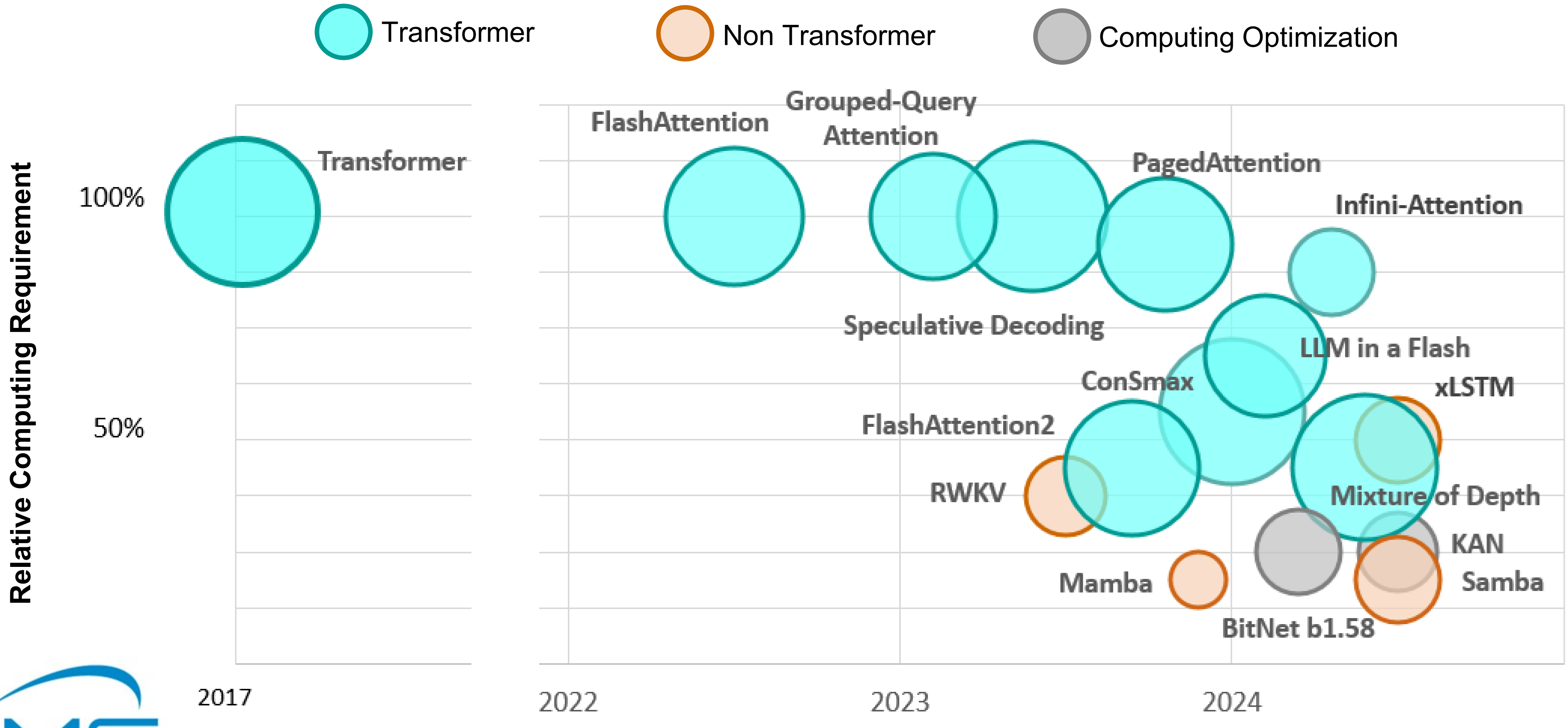
Llama 7B/13B/65B inferencing Power for fixed # of Tokens

TDP
V100: 300W
A100: 400W

Transformer Inference
Arithmetic Intensity is very low
→ Memory bound

Source: Samsi et al (MIT). IEEE HPEC`23

AI Algorithm Evolution



AI Algorithm Computing Requirement

*The circle size is relative memory requirement

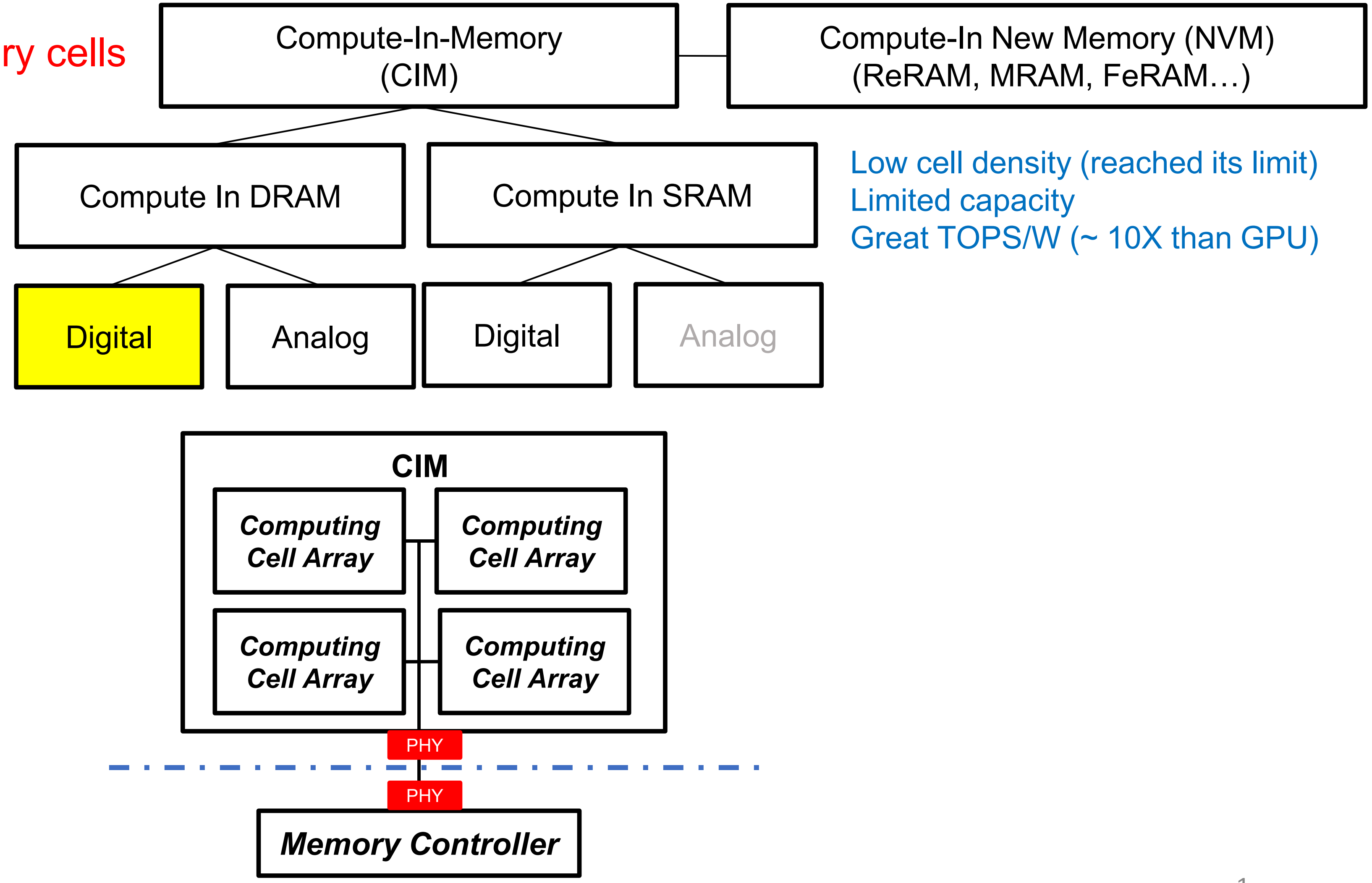
Energy Efficient Memory Solutions

- Compute-In-Memory
- Process-In-Memory
- Processing-Near-Memory (Computational Memory System)

Compute-In-Memory

Computation happens in the memory cells

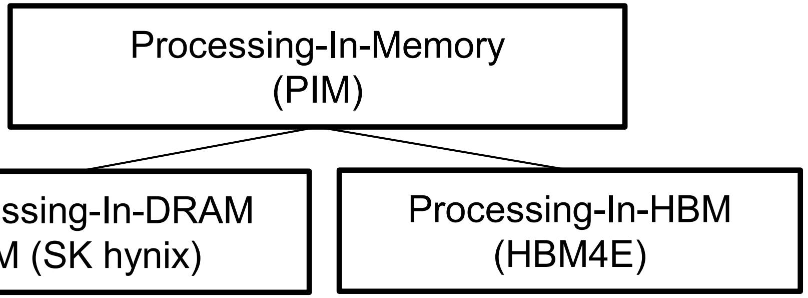
High cell density & capacity
Bit-serial computation (Limited OP/s)
Massive parallel operation



Low cell density (reached its limit)
Limited capacity
Great TOPS/W (~ 10X than GPU)

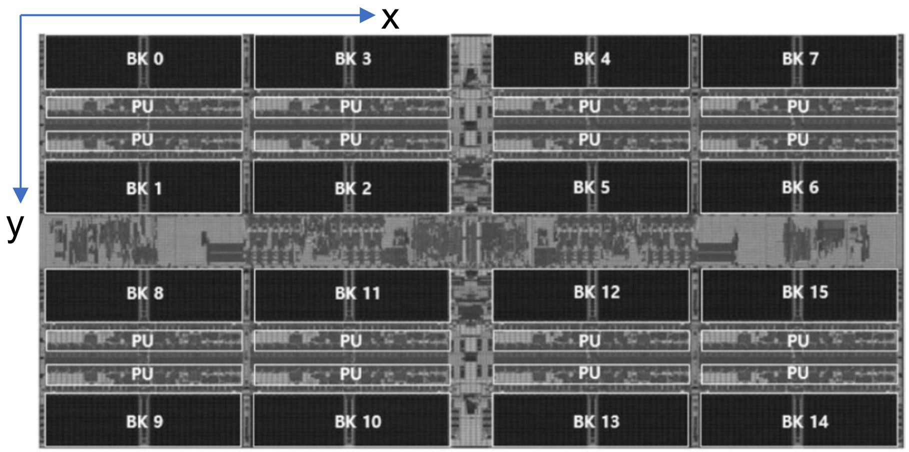
Process-In-Memory (PIM)

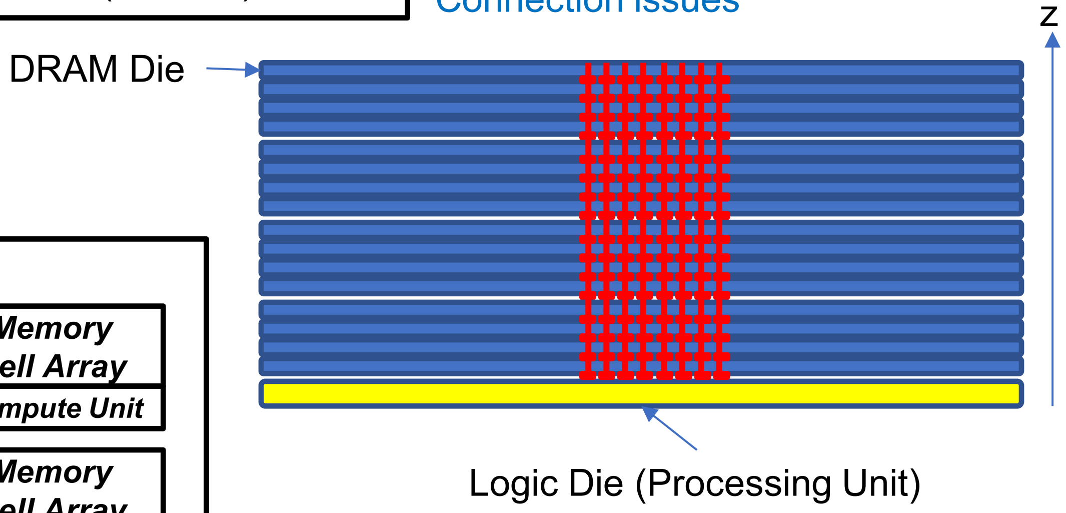
Computation happens in the memory package

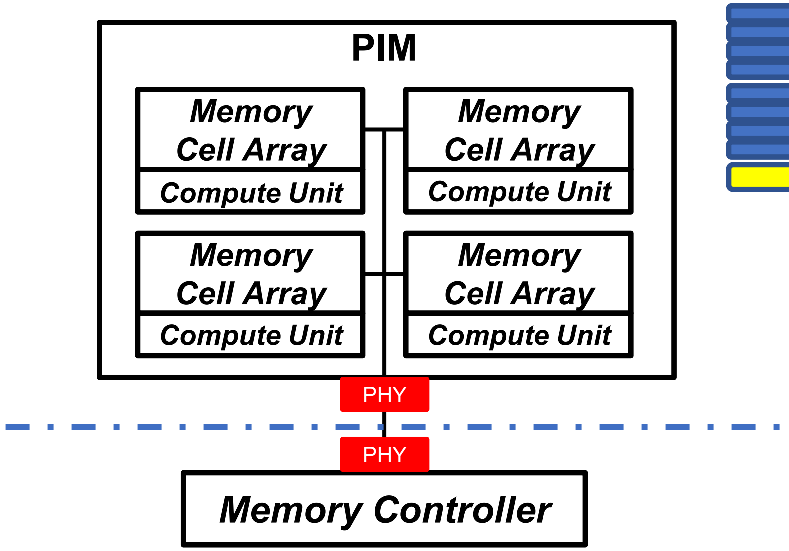


Monolithic Die
Limited PU area & lower cell density

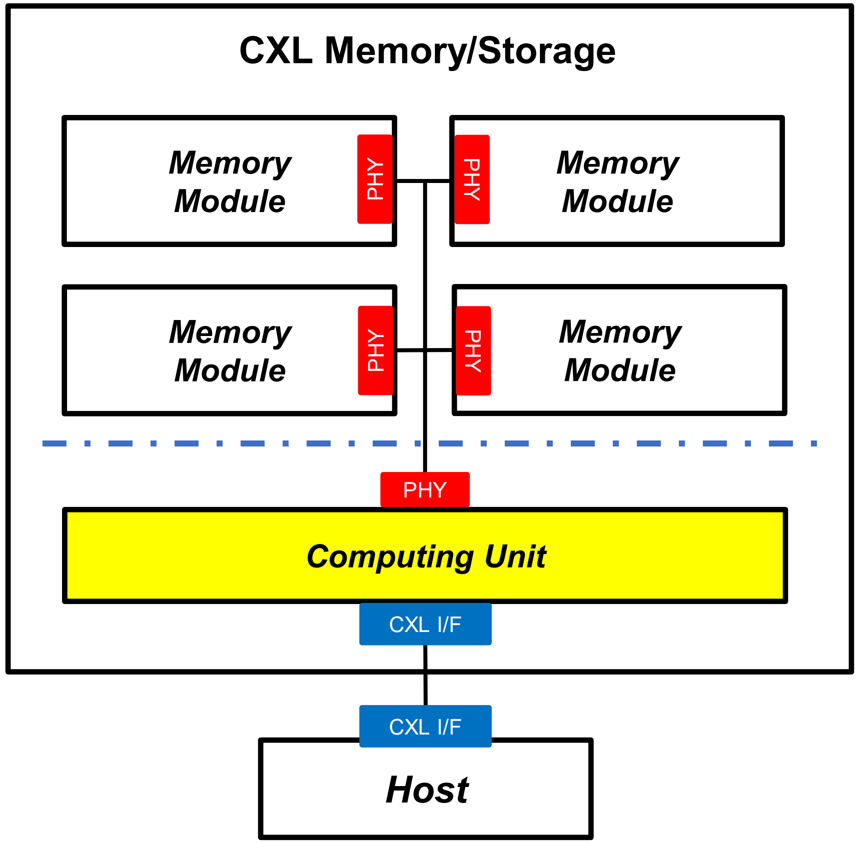
3D Packaging (Hybrid Bonding)
Thermal issues
Connection issues







Computational Memory System



Bandwidth Improvement

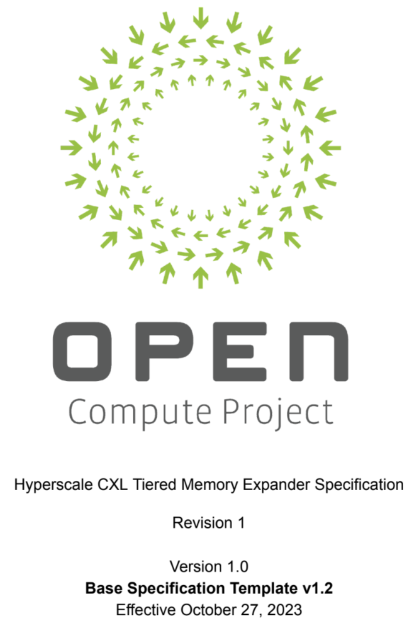
- (De)compression
- Model training computation

Hyperscale Data Challenge/ compression

Hyperscalers spending significant $$ on software based compression

	CPU cycles used for compression
Meta	4.6%
Google Cloud	3%

Hyperscaler requirement:
hardware accelerated compression is a **MUST-have**



OPEN Compute Project

Hyperscale CXL Tiered Memory Expander Specification

Revision 1

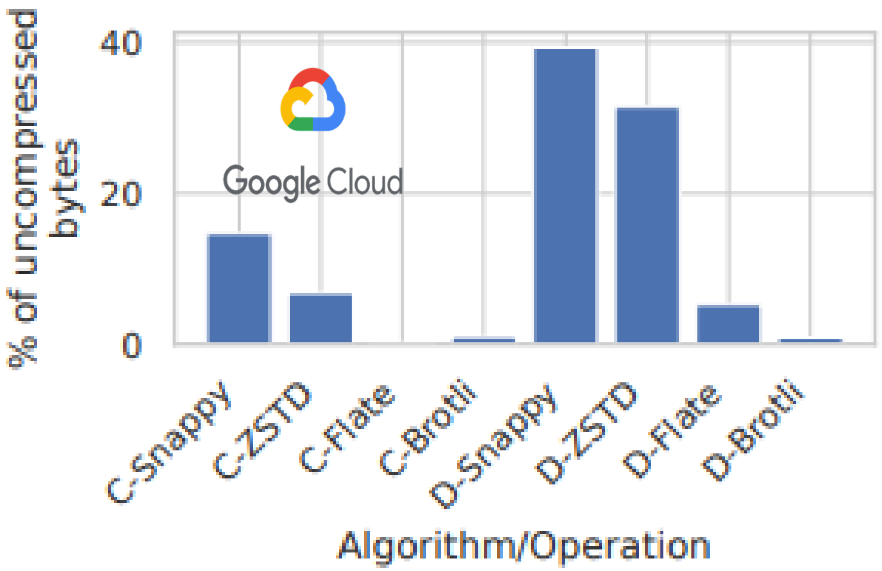
Version 1.0
Base Specification Template v1.2
Effective October 27, 2023

OCP Hyperscale CXL Tiered Memory Expander Spec

FMS

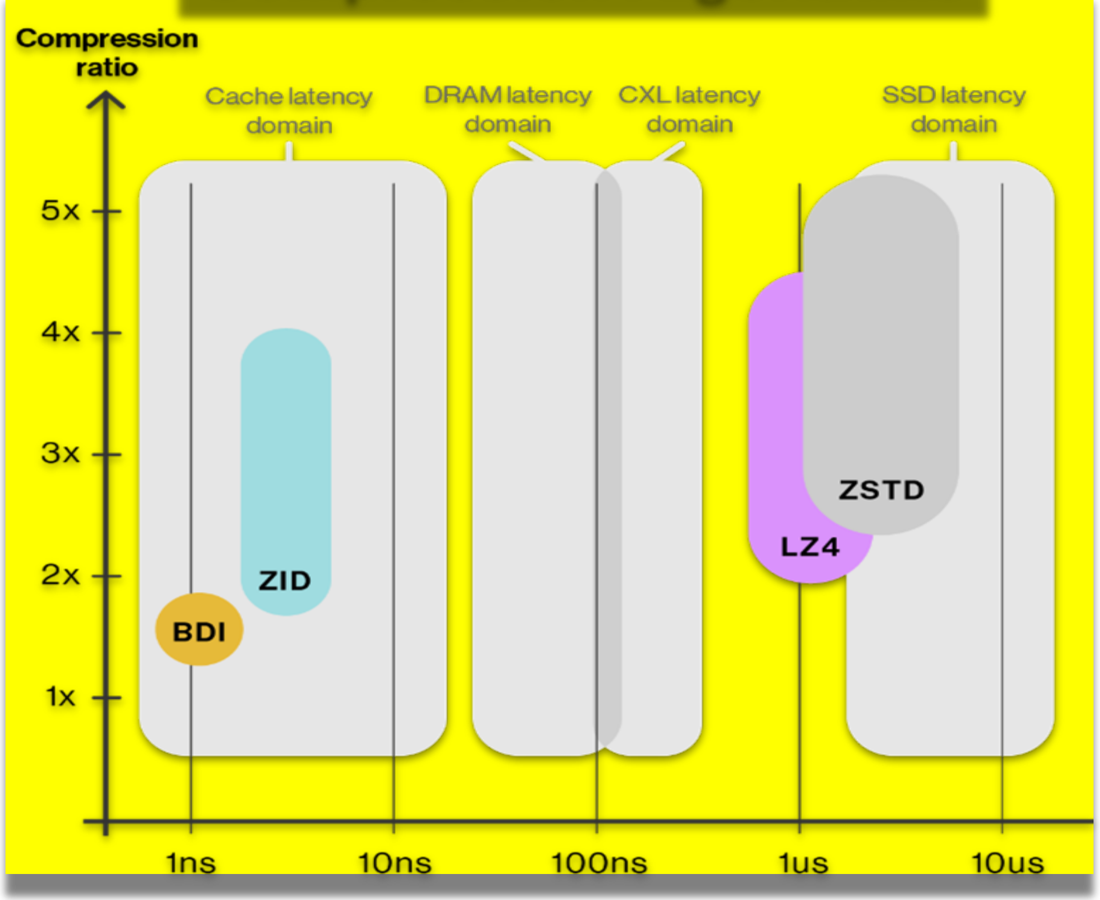
Compression Algorithms: Memory, storage

Google Fleetwide De(compression)



(a) Fleet-wide uncompressed bytes handled by (de)compression, broken down by algo.

Compression Algorithms



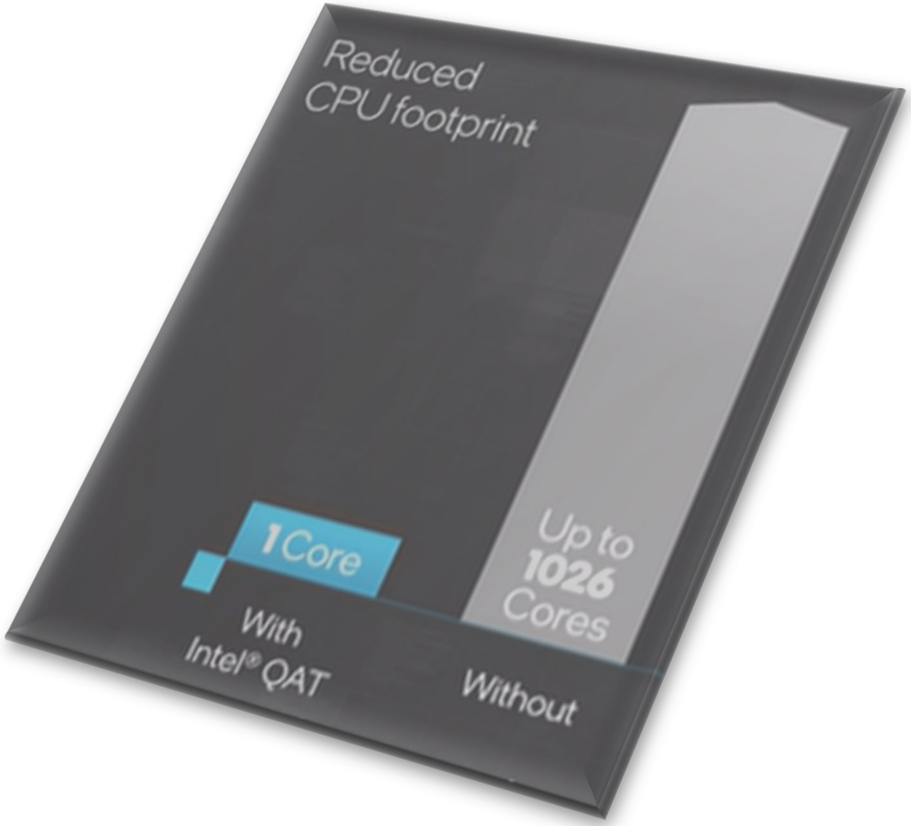
Room for Innovation

Legacy

LZ4, ZSTD IP

https://www.zeropoint-tech.com/products/nvme-expansion-flash-mx

Case Study: Implementations



Intel QAT: CPU integrated (De)compression engine

AI SUPERCHIP
208B Transistors
2nd GEN TRANSFORMER ENGINE
FP4/FP6 Tensor Core
5th GENERATION NVLINK
Scales to 576 GPUs
RAS ENGINE
100% In-System Self-Test
SECURE AI
Full Performance
Encryption & TEE
DECOMPRESSION ENGINE
800 GB/sec

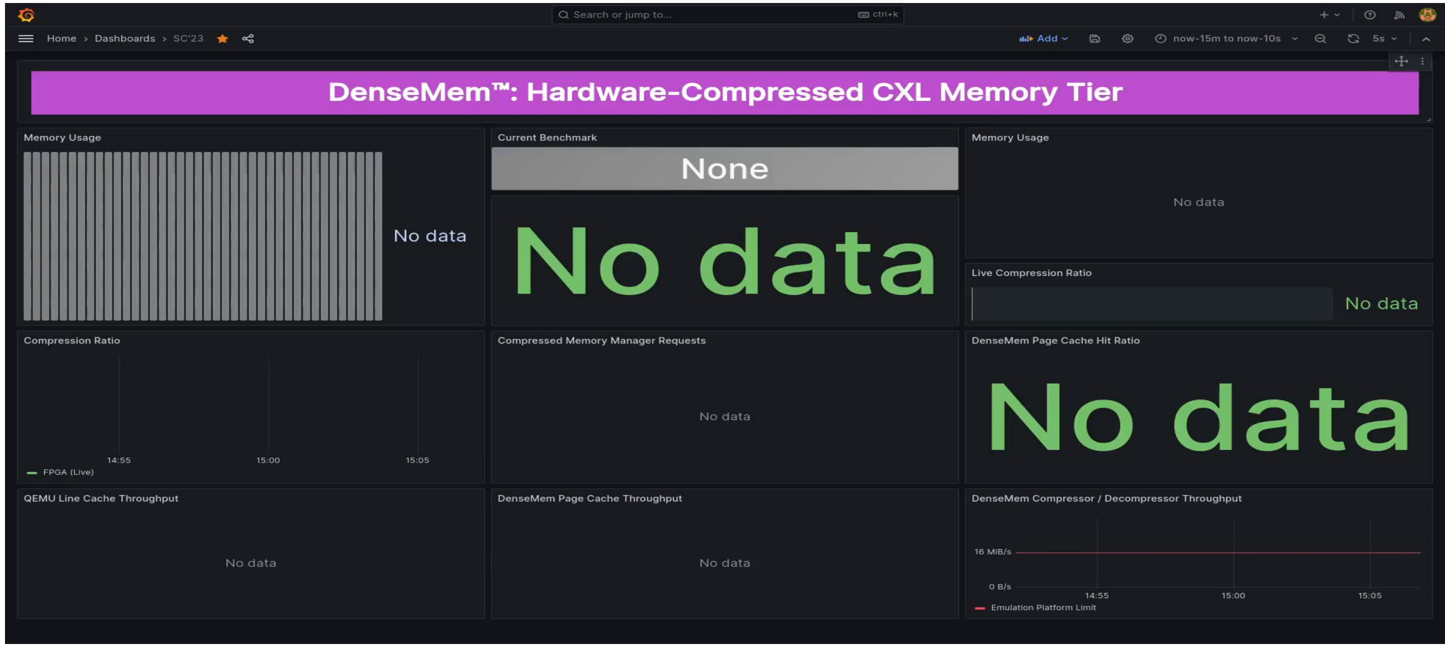
Nvidia Blackwell : 800GB/s embedded Decompression engine

Solutions, Challenges and considerations

Compaction, Software Transparent Address Translation – Dynamically Adjusting Capacity

Computational Programming

Standardized Data Movers



DenseMem™: Hardware-Compressed CXL Memory Tier

Memory Usage

No data

Current Benchmark

None

No data

Memory Usage

No data

Live Compression Ratio

No data

Compression Ratio

Compressed Memory Manager Requests

No data

DenseMem Page Cache Hit Ratio

No data

QEMU Line Cache Throughput

No data

DenseMem Page Cache Throughput

No data

DenseMem Compressor / Decompressor Throughput

SNIA®

Advancing storage & information technology

Smart Data Accelerator Interface ("SDXI") Specification

Version 1.0

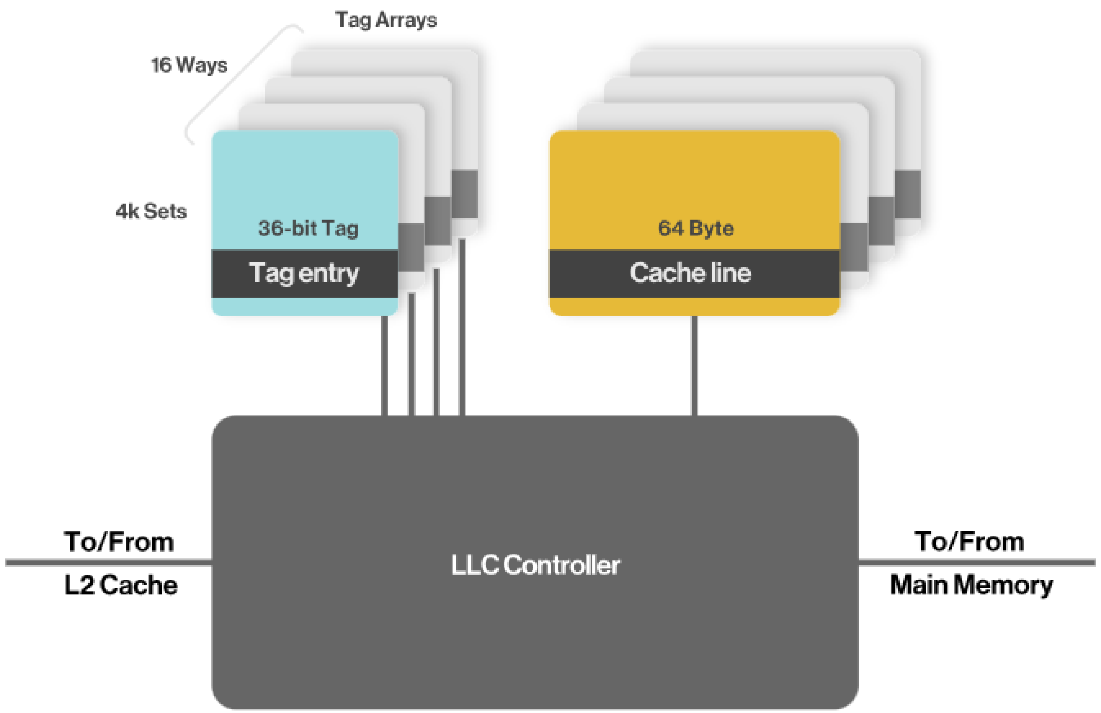
SNIA: Data Mover Spec

COMPUTATIONAL MEMORY & STORAGE: PRINCIPLES & BEST PRACTICES

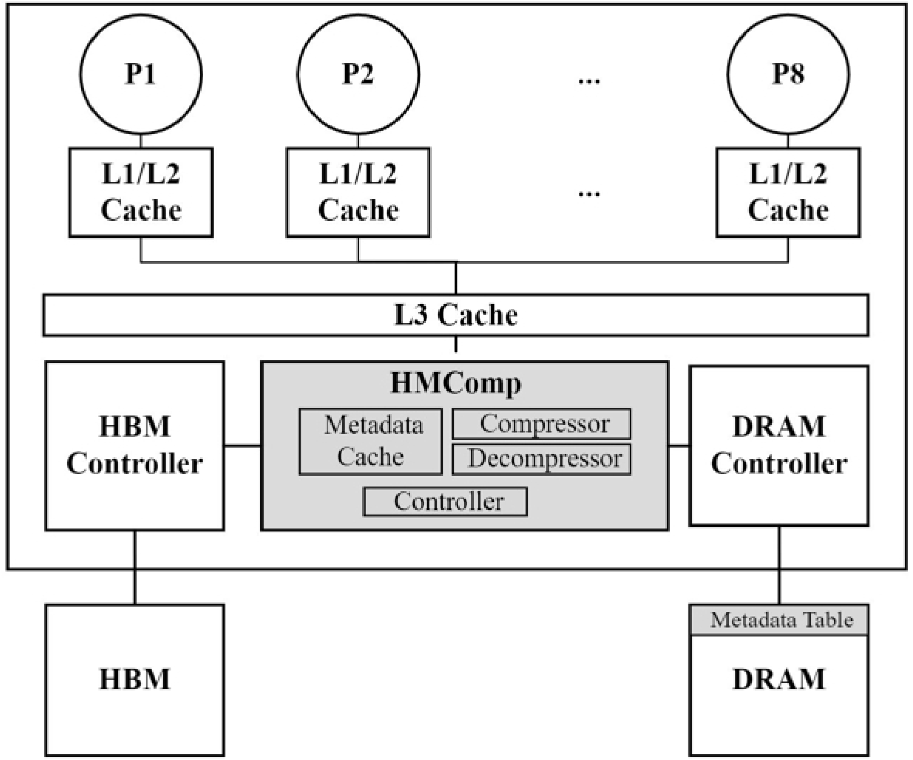
V0.1r4

OCP CMS Computational Programming Work Group

Future Trends / opportunities



CacheMX Cache (De)compressor IP



HMComp: Extending Near-Memory Capacity using Compression in Hybrid Memory

On chip SRAM cache compression

HBM Compression

Summary Call to Action

- Joint Collaboration to accelerate Energy Efficient Solution deployment
- Start with Low hanging fruit : (De)Compression
- Prove out other use cases, influence/ shape standards