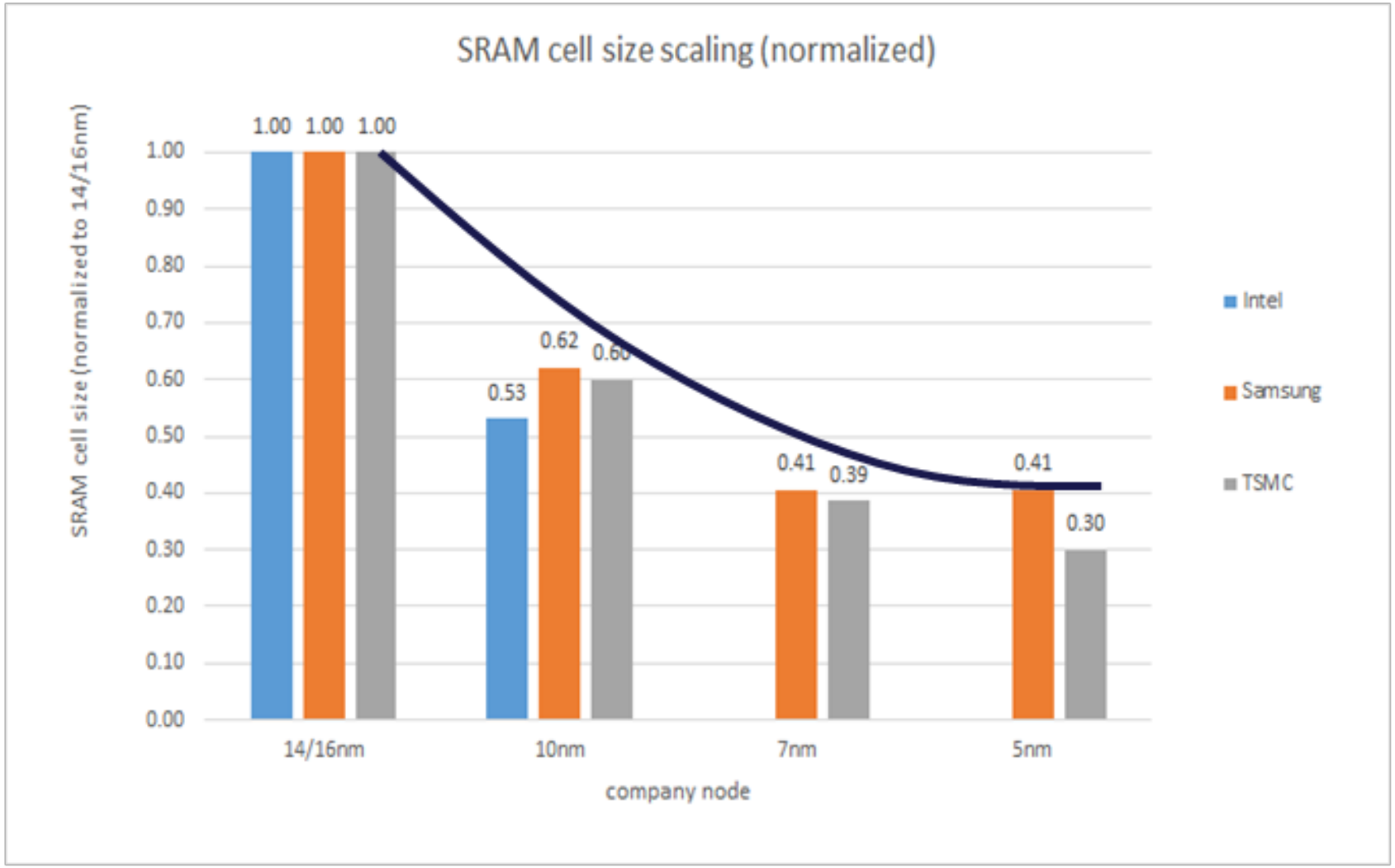
Chiplet-Based Compressed LLC Cache & Memory Expansion

Presenter:
Nilesh Shah, VP Business Development, ZeroPoint Technologies

FMS the Future of Memory and Storage

Memory Challenge: SRAM Leakage , Scaling limitations



(Source: SemiWiki 07/28/20)

Memory Challenge: AI

- AI Architectures
  - Training: Nvidia GPU L2 cache, HBM subsystem [(de)compression]
  - Inference: Custom accelerators (Groq, Tenstorrent)
- Use case :Inference vs training
- Memory capacity + latency Critical for response time
- Compute limited by memory bandwidth wall
- **Cost Optimization**

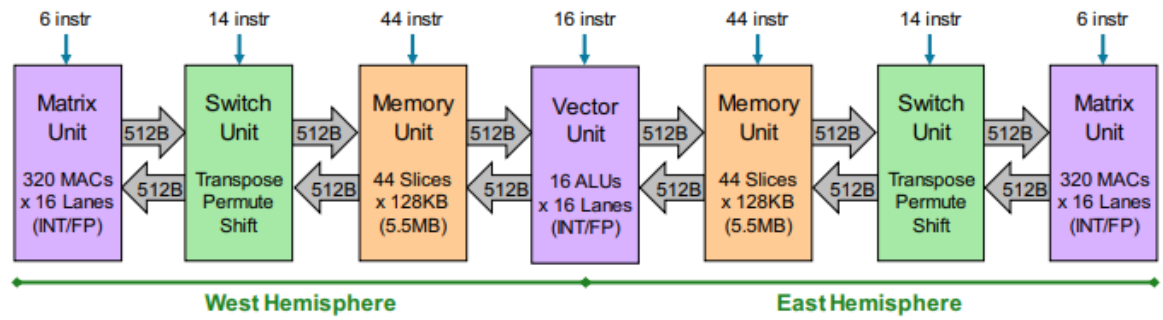
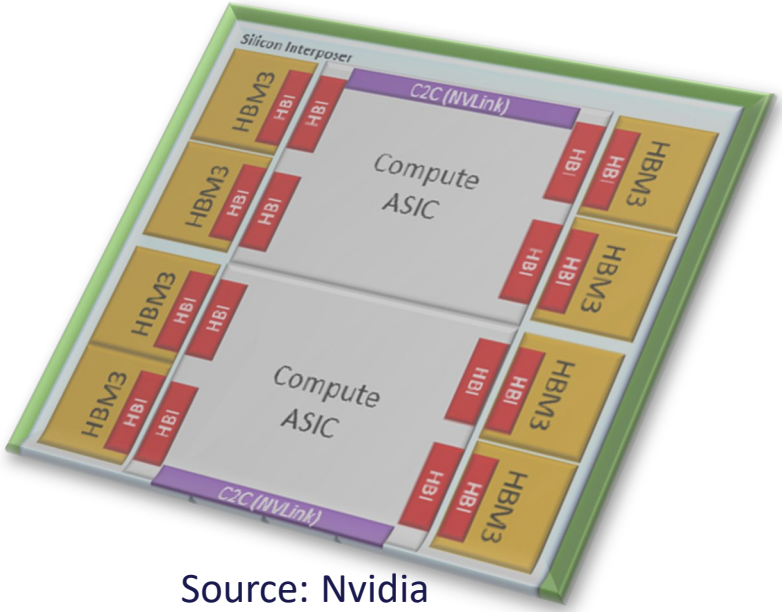
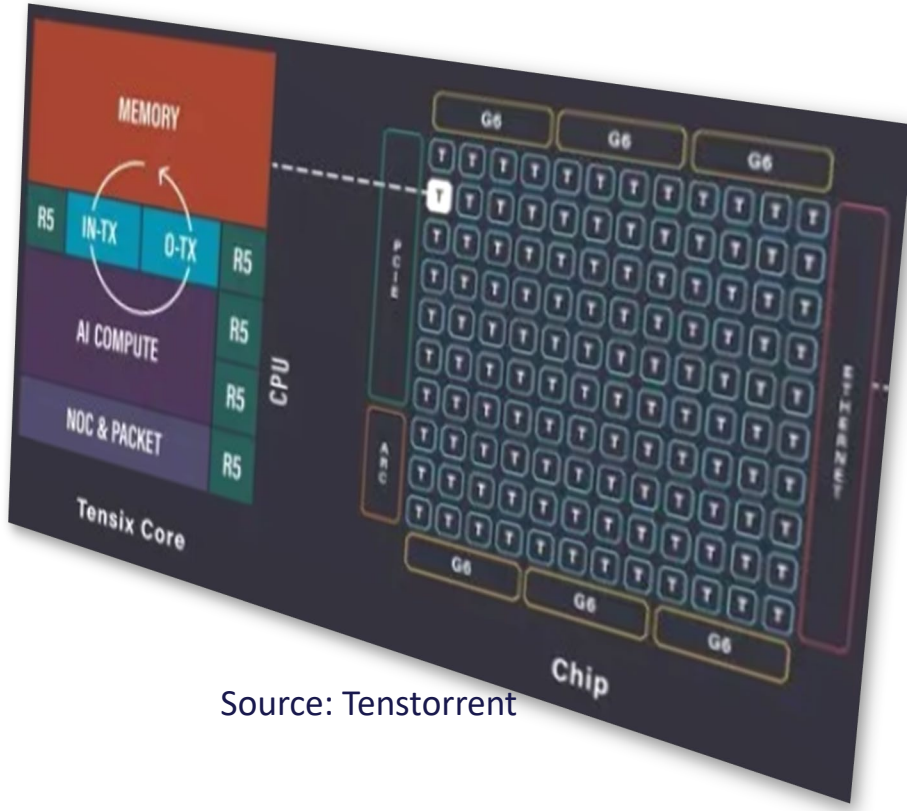


Figure 2. TSP superlane block diagram. Every superlane is bilaterally symmetric with an east side and a west side. It contains 16 lanes, each of which is 8 bits wide. Data flows from east to west and from west to east.

Source: Groq



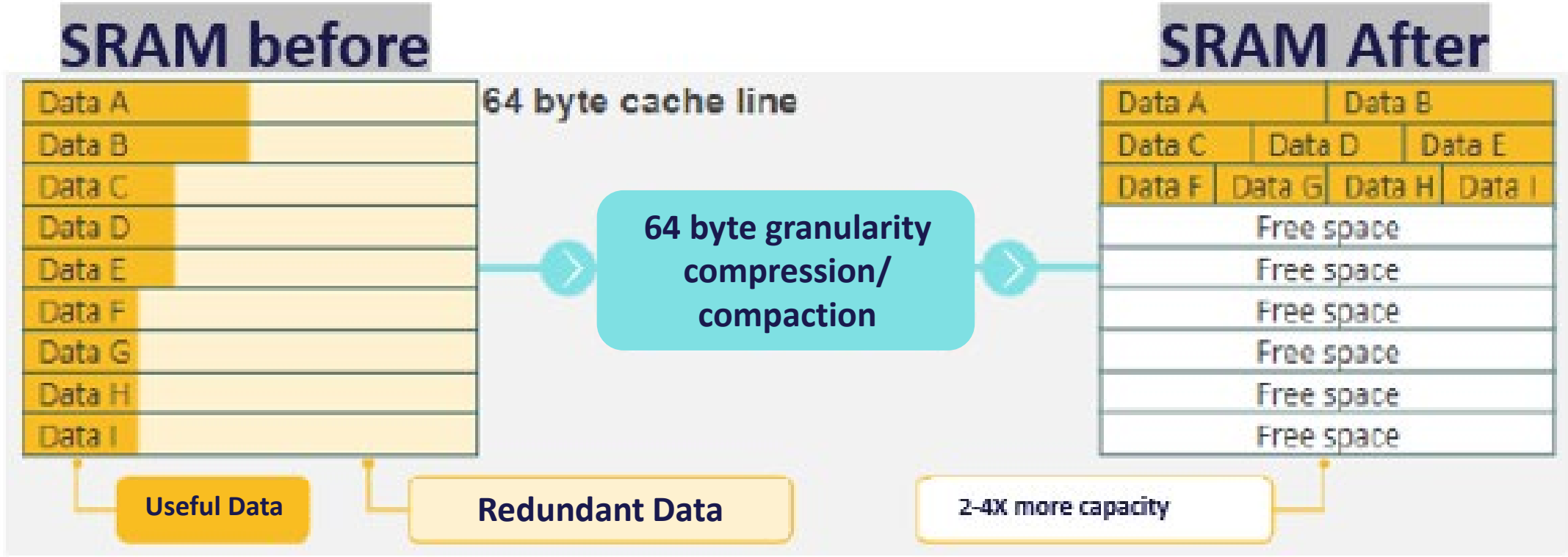
Source: Nvidia



Source: Tenstorrent

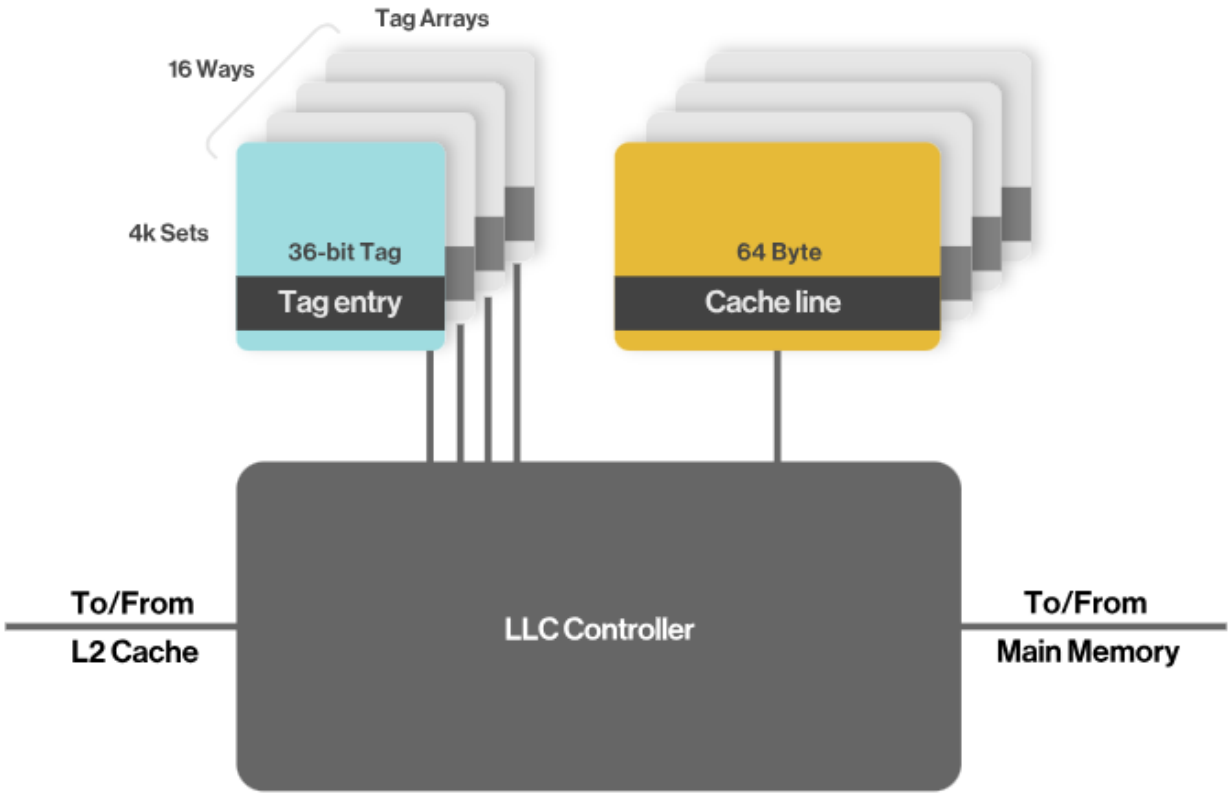
Opportunity #1 | Compressed SRAM

- 2-4X Compression



Cache Compression Opportunity: Requirements

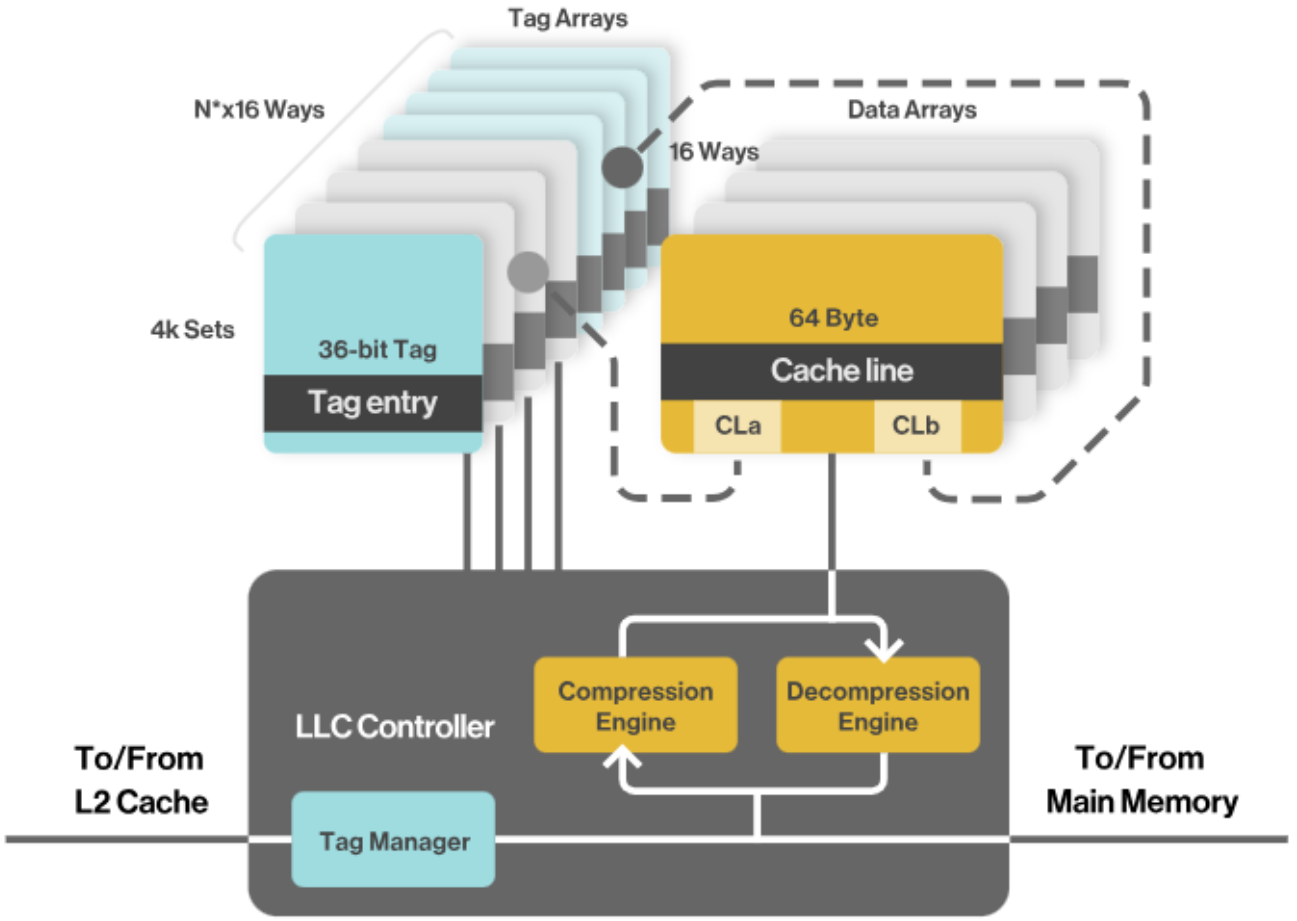
- High Compression
  High Performance
- Low Latency
- Small area overhead/footprint
- Low Power
- Ease of Integration (scratchpad, LLC)
- Transparent to user (self contained)



Proposed Solution | Cache Compression IP

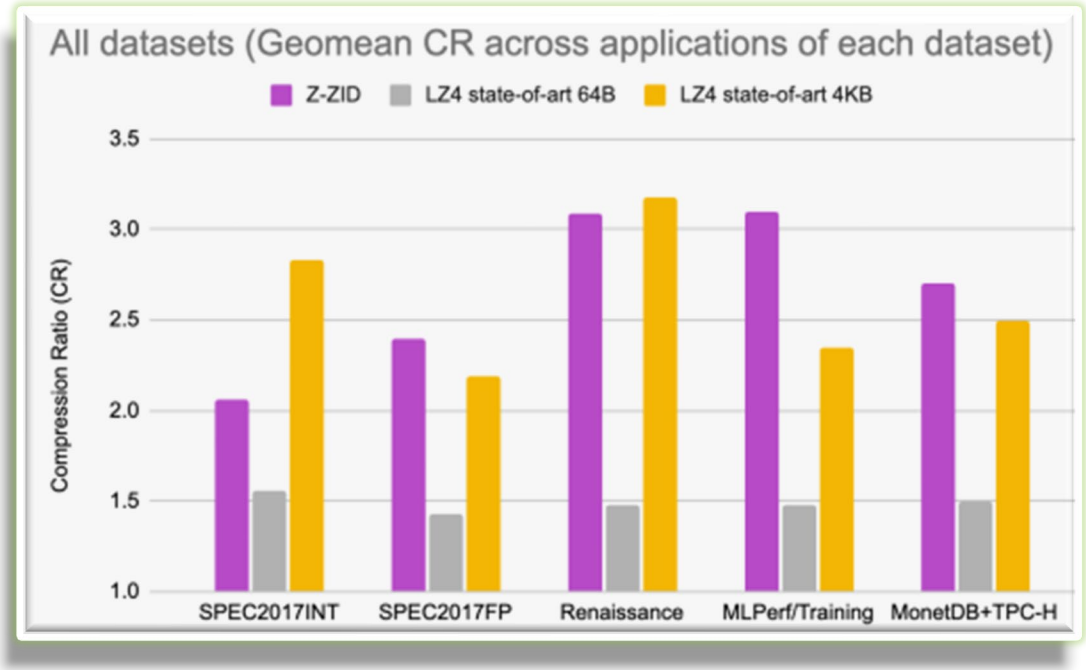
- Process agnostic, PORTABLE IP Solution
  - (De)Compressor
  - Tag Manager
  - Tag Arrays
- Democratize access
  - Integrate into any SoC, chiplet

[IP Spec Sheet]



ZeroPoint Cache Compression IP Results

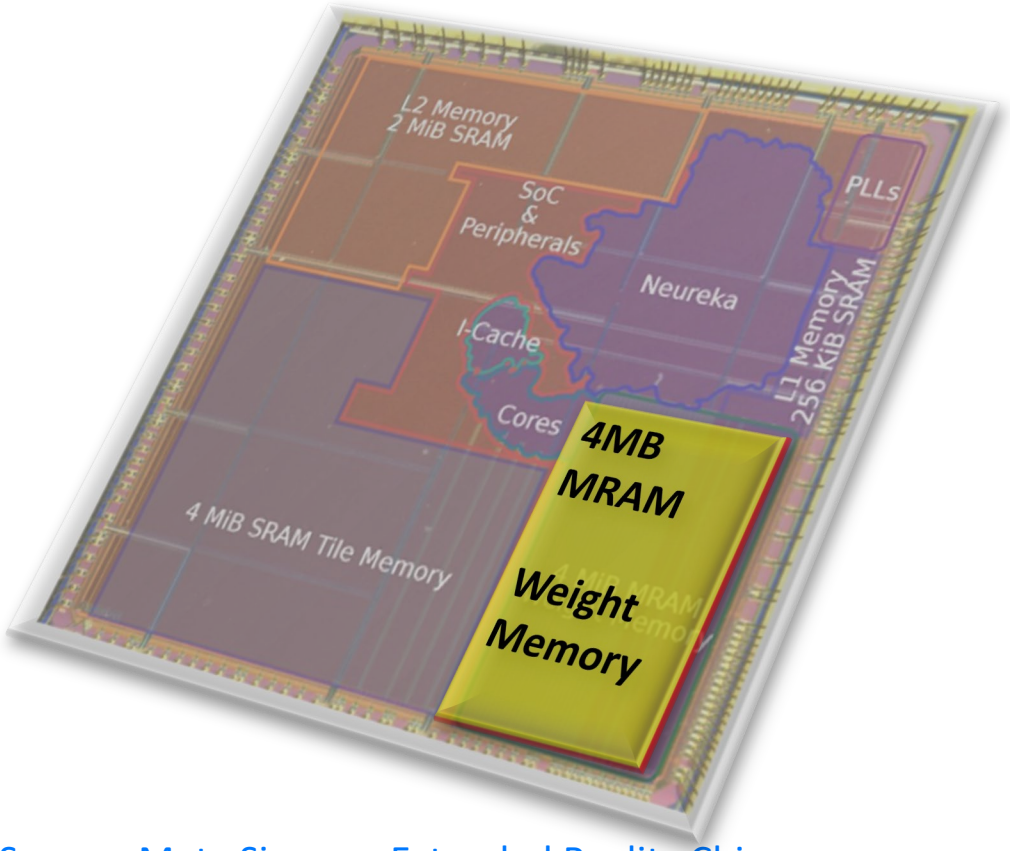
- 2-4X Compression ratio across variety of workloads Cache line granularity compression algorithm
- 15-30% performance acceleration
- Low Latency – 5 cycles (ZSD algorithm)
- Area efficient starting at 0.1mm sq 5nm TSMC
- Operate at line speed for L2$, L3$, SLC



2-4X Compression ratio across variety of workloads

Opportunity #2: SRAM Alternative - MRAM

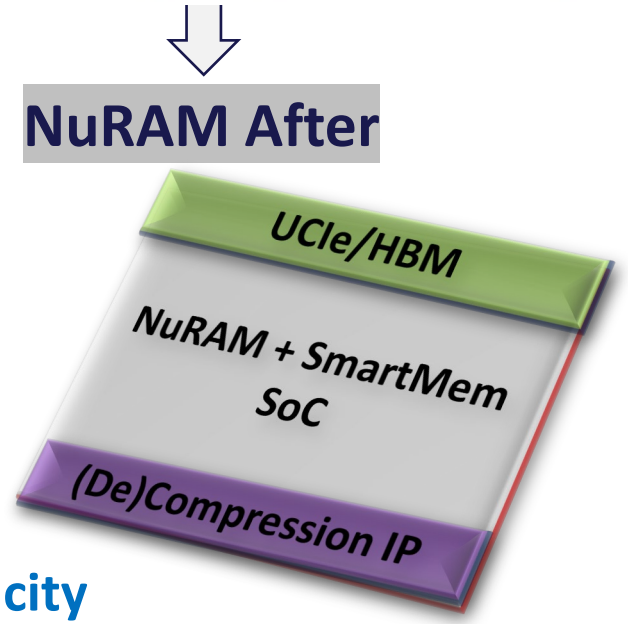
- NuMem NuRAM [MRAM bitcell based memory]
- 2.5X denser than SRAM, Scales down with process geometry
- 85x-2000x lower leakage power than SRAM
- 60-650x improvement in Latency over DRAM
- 2x HBM Bandwidth (at equivalent # of wires)
- Data Retention without power
- Implementation: Meta Siracusa Extended Reality SoC



Source: Meta Siracusa Extended Reality Chip
At-MRAM Neural Engine.
~2.5X denser than SRAM

Opportunity #3: Combine Compression + NuRAM

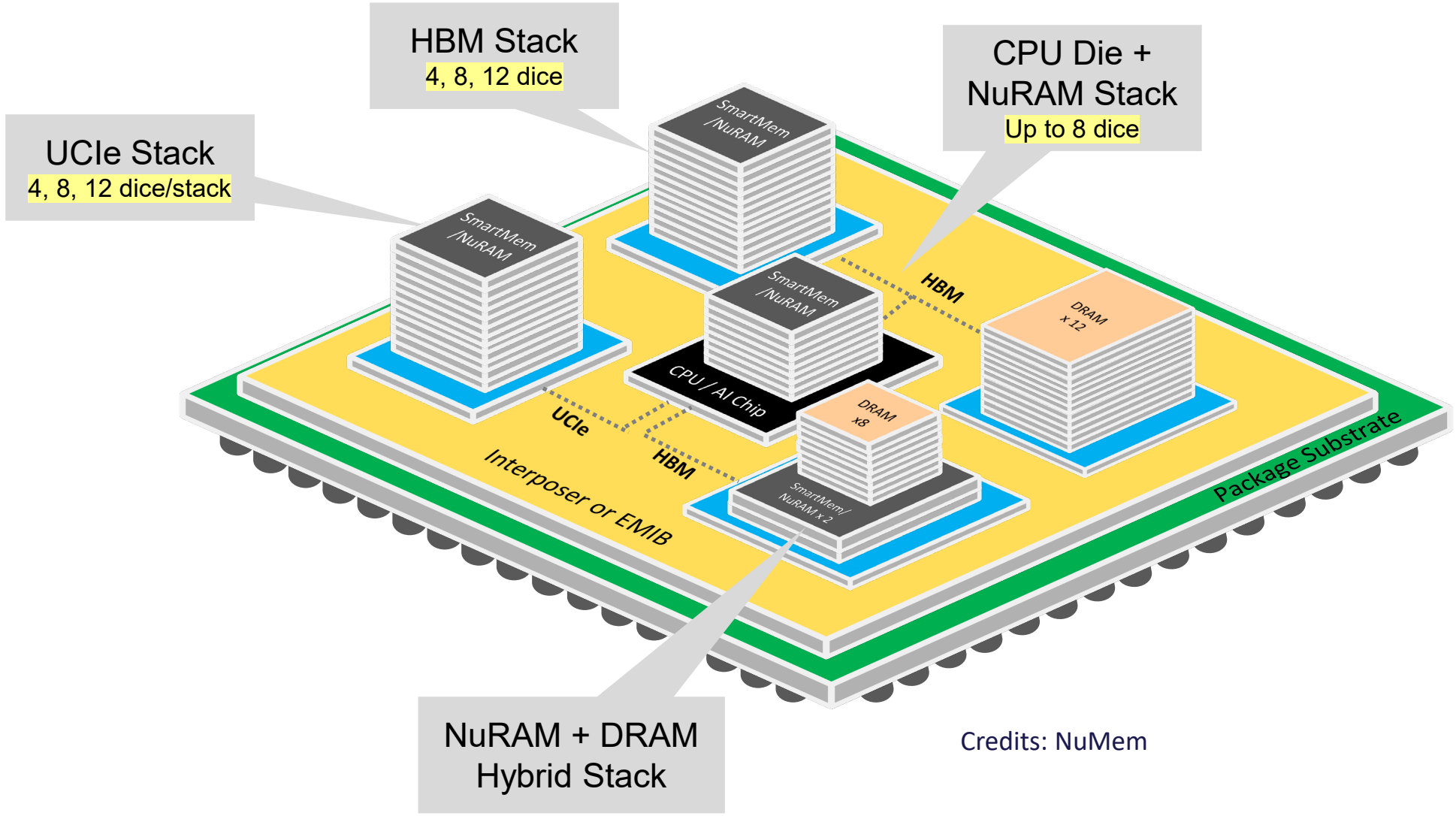
- Chiplet Synergy
  - Compression : 2-4X effective capacity
  - NuRAM: 2.5X capacity ISO area,
    Low density up to 1-3GB per die
  - Compression + NuRAM: up to 5-10X effective capacity compared to SRAM
  - Chiplet[NuMem SmartMem SoC solution]
  - Amortize Chiplet cost over larger effective capacity



5-10X more capacity

Opportunity #3: SoC chiplet options

Chiplet Synergy



Credits: NuMem

Opportunity: HBM Compression

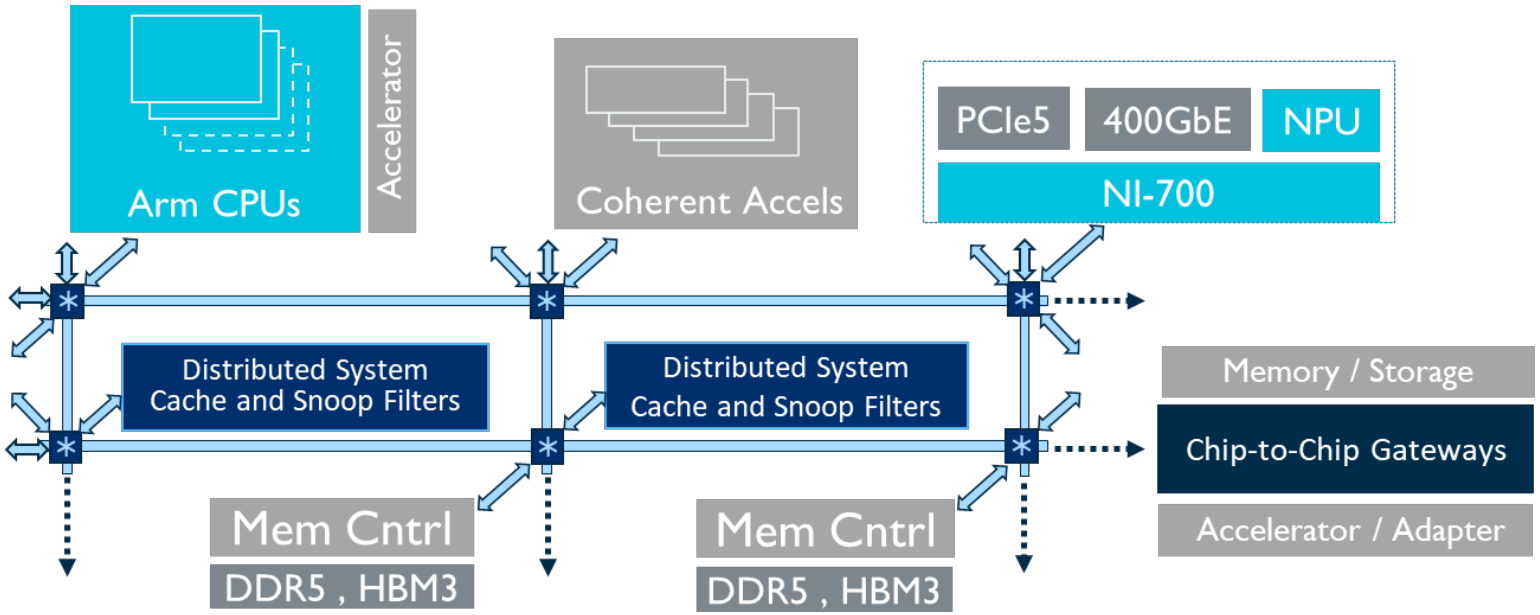
- Nvidia Blackwell : 800GB/s embedded Decompression engine

- Opportunity: (De)Compression engine in HBM chiplet

Source: https://nvidianews.nvidia.com/news/nvidia-blackwell-platform-arrives-to-power-a-new-era-of-computing

Opportunity: Coherent Mesh network integration

- Disaggregated, coherent compression engines
- Extensible across chiplets via gateways
- Ex: ARM CMN-Cyprus scalable coherent mesh architecture



Credit: ARM

Summary/ Call to Action

Summary

- Portable cache compression ZeroPoint IP solution product brief
- NuRAM SmartMem high-bandwidth, low latency memory SOC: www.numem.com
- NuRAM SmartMem with ZeroPoint compression enables 5-10x area reduction or increased density
- Coherent Mesh network available from ARM

Call to Action -Community collaboration:

- Most important use cases?
- Chiplet/test chip implementation collaboration?
- SoC performance and TCO targets for each use case?
- Chiplet: how do we scale beyond point to point connections using coherent mesh?