PALMA DE MALLORCA
7-10 SEPTEMBER
ESSERC 2026

CHIMERA: A Flexible and Scalable 3.1 TOPS/W AI-MCU with Transformer Accelerator and 563 Gb/s Shared-L2 Memory Subsystem with QoS Guarantees

Lorenzo Leone, Philip Wiese, Gamze İslamoğlu, Michael Rogenmoser,
Davide Rossi, Francesco Conti and Luca Benini

IIS, ETH Zurich, Switzerland
University of Bologna, Italy
Chips-IT, Pavia, Italy
lleone@iis.ee.ethz.ch

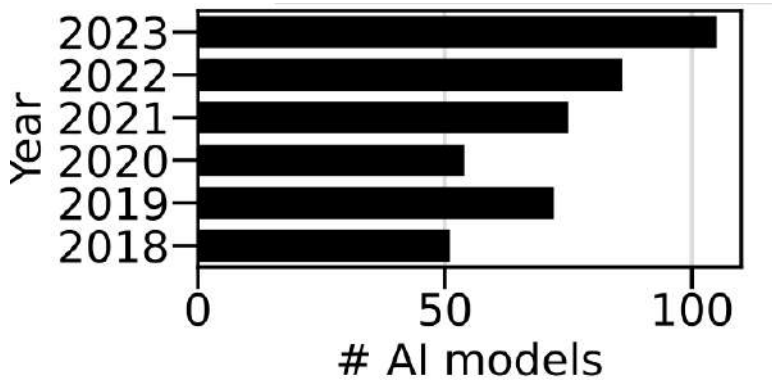
Outline

- Background and Motivation
- Architecture Overview
- Results
- Conclusion

Outline

- Background and Motivation
- Architecture Overview
- Results
- Conclusion

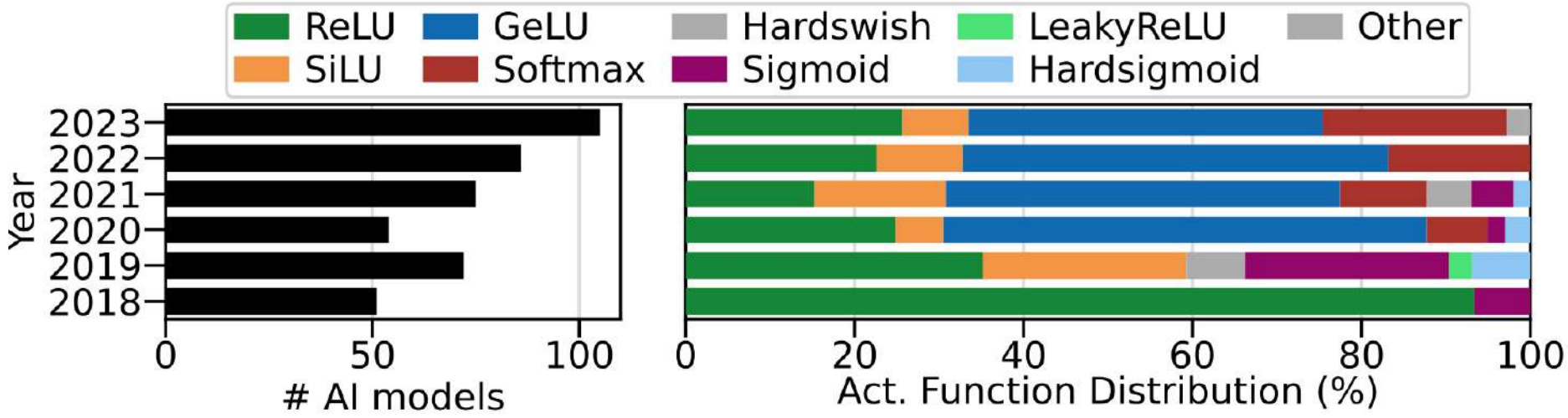
The AI Model Race: From Datacenter to Edge



- Attention-based models are rapidly expanding across modalities:
  - Vision (ViTs), Speech (ASR), Language (NLP)
- Rapidly evolving model and operator diversity:
  - Increasing activation-function diversity across workloads

N. Maslej et al., "Artificial intelligence index report 2025," 2025
R. Andri et al., "Flex-SFU: Activation Function Acceleration With Nonuniform Piecewise Approximation,"2025

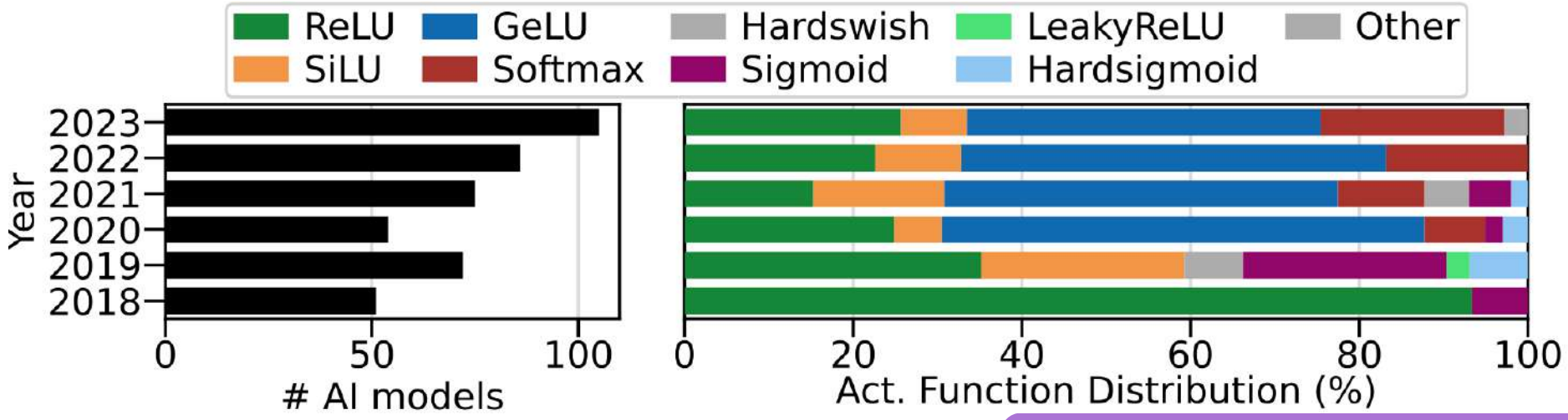
The AI Model Race: From Datacenter to Edge



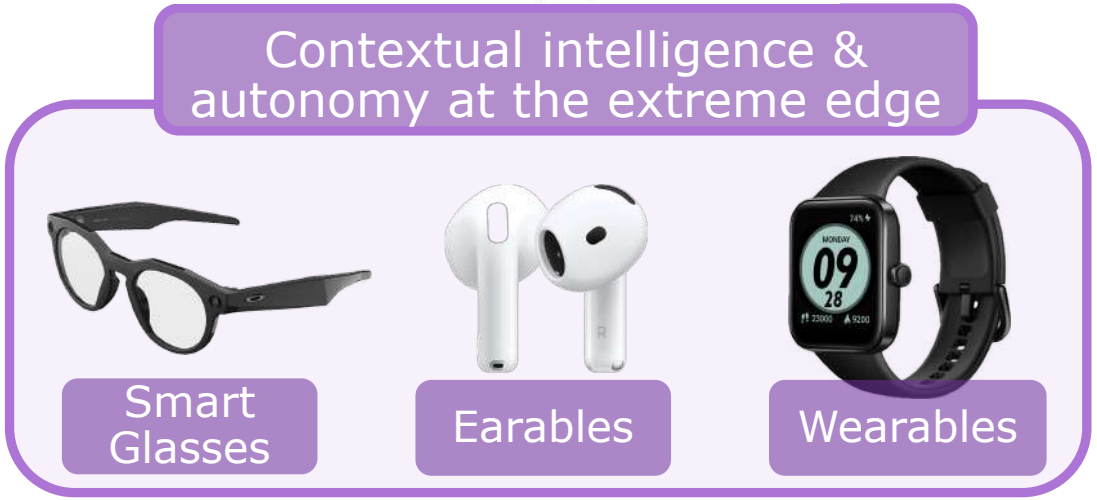
- Attention-based models are rapidly expanding across modalities:
  - Vision (ViTs), Speech (ASR), Language (NLP)
- Rapidly evolving model and operator diversity:
  - Increasing activation-function diversity across workloads

N. Maslej et al., "Artificial intelligence index report 2025," 2025
R. Andri et al., "Flex-SFU: Activation Function Acceleration With Nonuniform Piecewise Approximation,"2025

The AI Model Race: From Datacenter to Edge



- Attention-based models are rapidly expanding across modalities:
  - Vision (ViTs), Speech (ASR), Language (NLP)
- Rapidly evolving model and operator diversity:
  - Increasing activation-function diversity across workloads



N. Maslej et al., "Artificial intelligence index report 2025," 2025
R. Andri et al., "Flex-SFU: Activation Function Acceleration With Nonuniform Piecewise Approximation,"2025

The Power-Flexibility Trade-Off

Power

7-25 W
Robotics

1-5 W
Embedded AI

10-100 mW
Wearables

< 1 mW
AoN sensing

Jetson Orin Nano
- High-performance edge AI
- Highly programmable

Coral Edge TPU/HAILO-8
- Low-power edge AI
- Dedicated ML acceleration

106 mW

Our Work

Syntiant NDP200
- Ultra-low-power
- Very Specialized

Flexibility

The Power-Flexibility Trade-Off

Power

7-25 W
Robotics

1-5 W
Embedded AI

10-100 mW
Wearables

< 1 mW
AoN sensing

Need for system-level co-design to preserve **flexibility** while sustaining accelerator-level **energy efficiency**

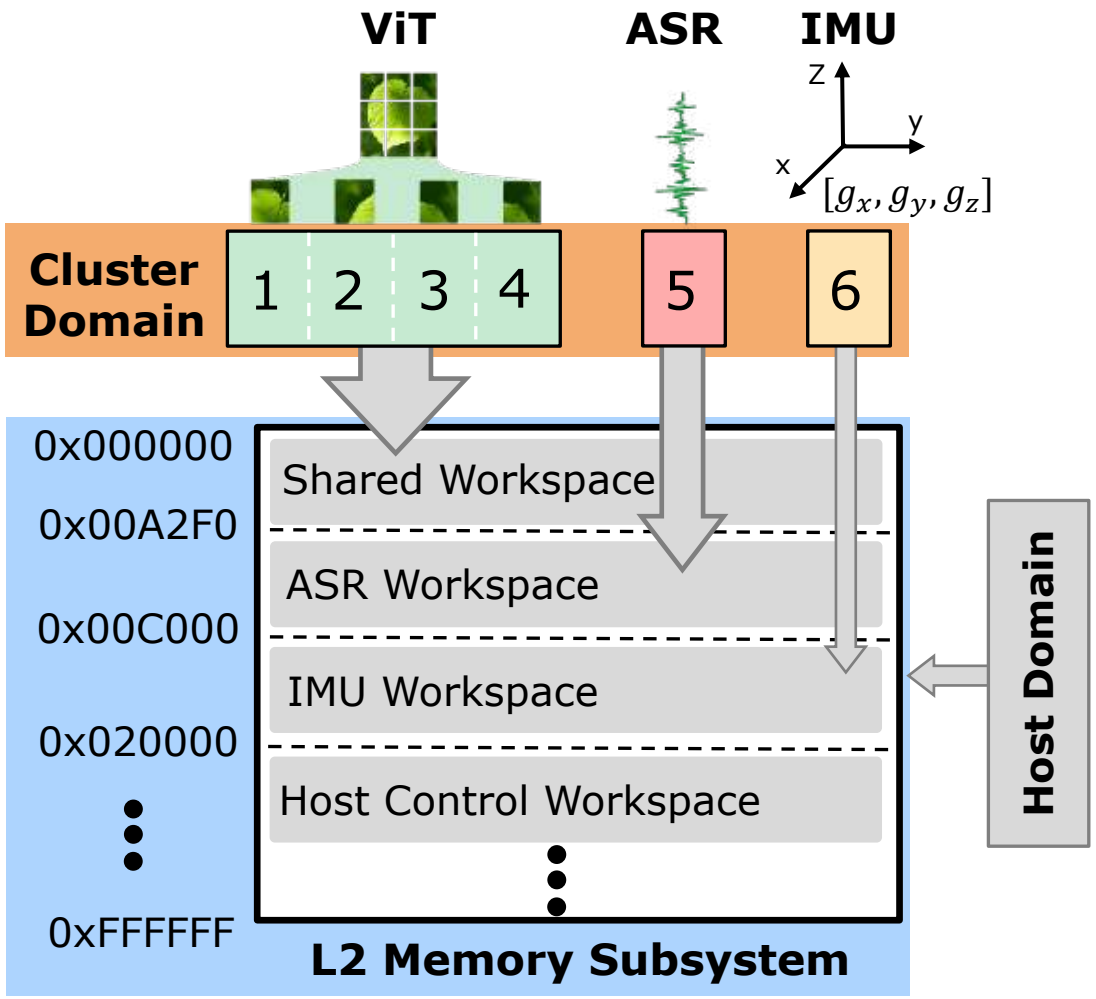
Syntiant NDP200
- Ultra-low-power
- Very Specialized

106 mW

Our Work

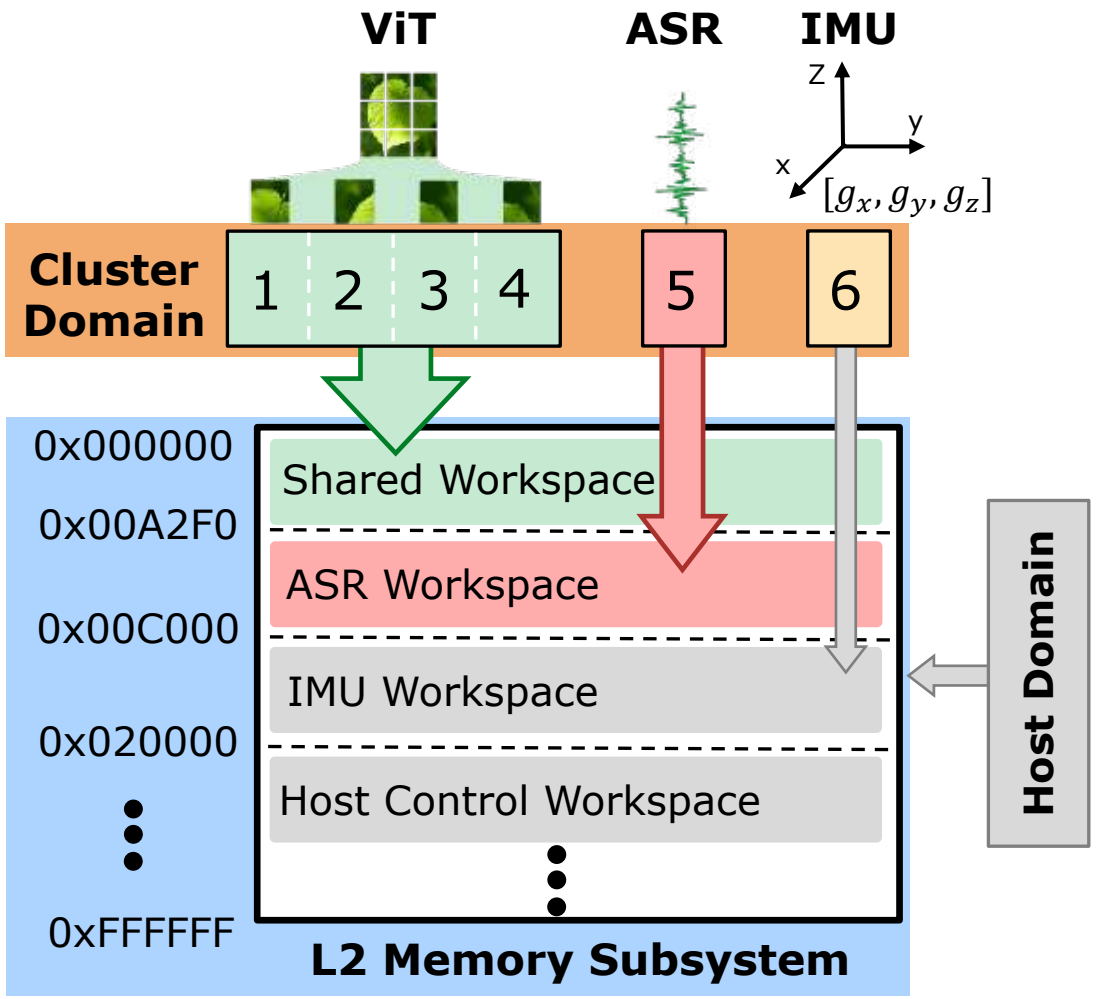
Flexibility

The Ingredients for Modern AI-MCU



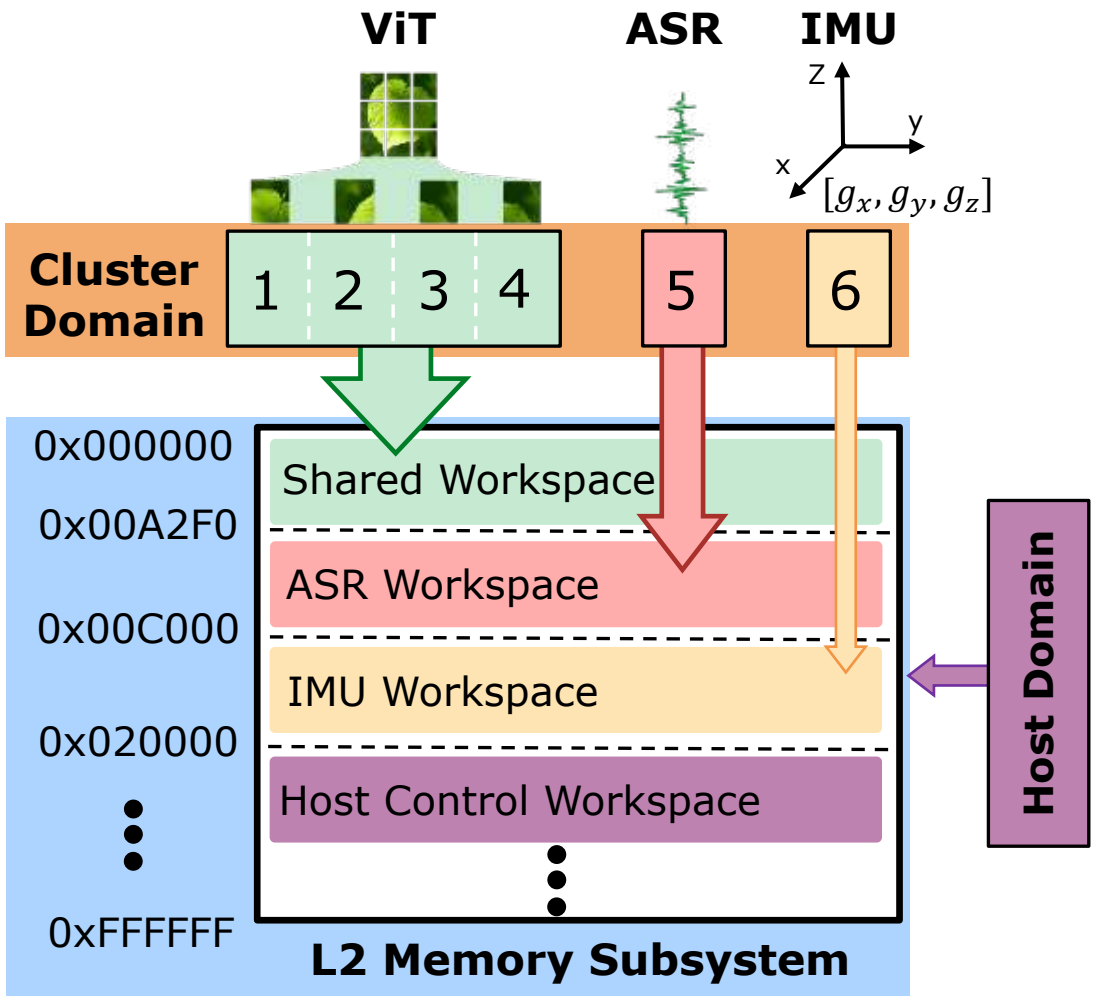
- **Flexible** Compute:
  - Heterogeneous workloads
  - Programmable processors + specialized acceleration
- **Heterogeneous** L2 traffic:
  - Concurrent multi-cluster accesses
  - High-throughput read/write
  - Latency-critical inter-cluster synchronization
- **QoS** guarantee:
  - Host control core message passing
  - Fast and predictable memory access

The Ingredients for Modern AI-MCU



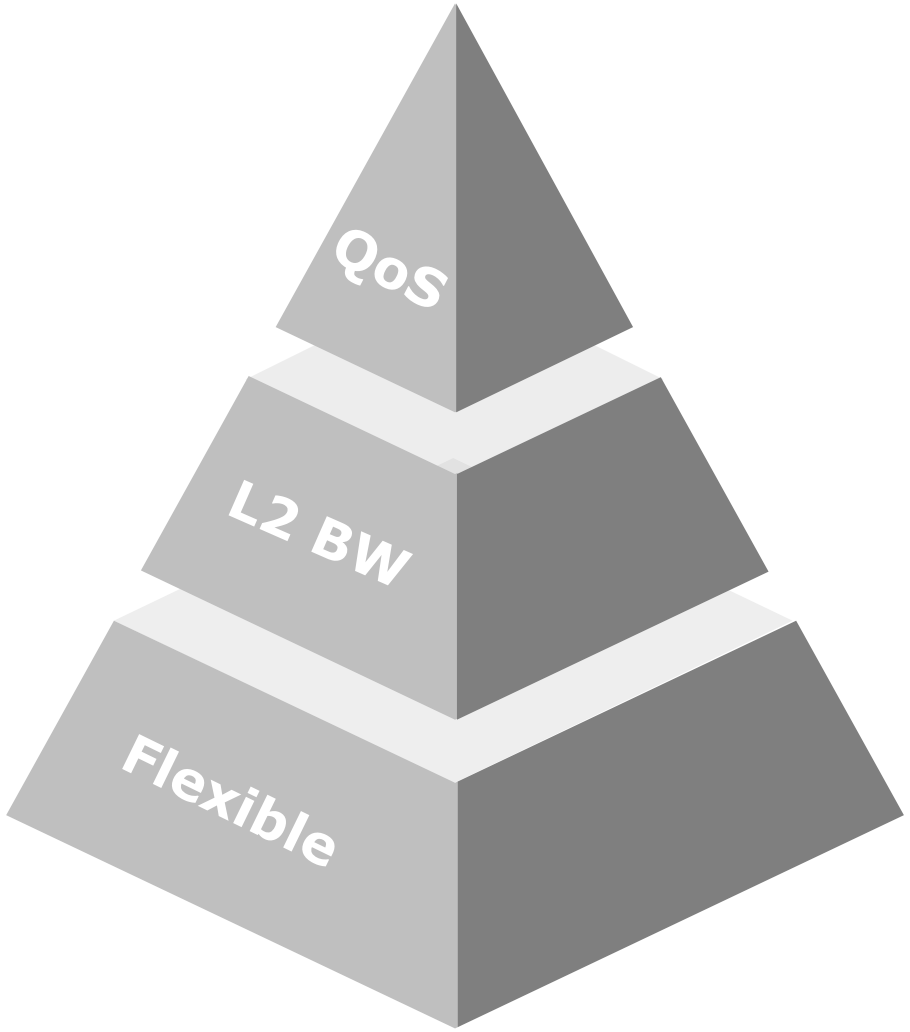
- **Flexible** Compute:
  - Heterogeneous workloads
  - Programmable processors + specialized acceleration
- **Heterogeneous** L2 traffic:
  - Concurrent multi-cluster accesses
  - High-throughput read/write
  - Latency-critical inter-cluster synchronization
- **QoS** guarantee:
  - Host control core message passing
  - Fast and predictable memory access

The Ingredients for Modern AI-MCU

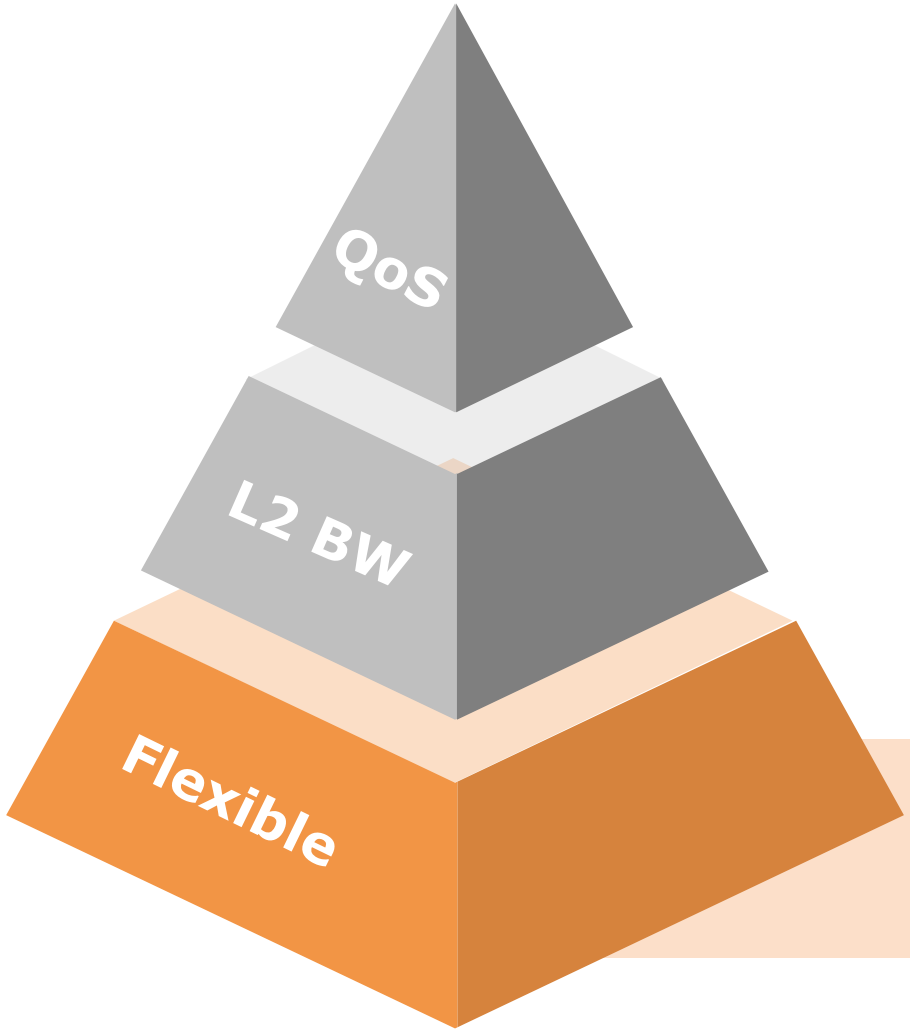


- **Flexible** Compute:
  - Heterogeneous workloads
  - Programmable processors + specialized acceleration
- **Heterogeneous** L2 traffic:
  - Concurrent multi-cluster accesses
  - High-throughput read/write
  - Latency-critical inter-cluster synchronization
- **QoS** guarantee:
  - Host control core message passing
  - Fast and predictable memory access

CHIMERA: Putting the Ingredients Together



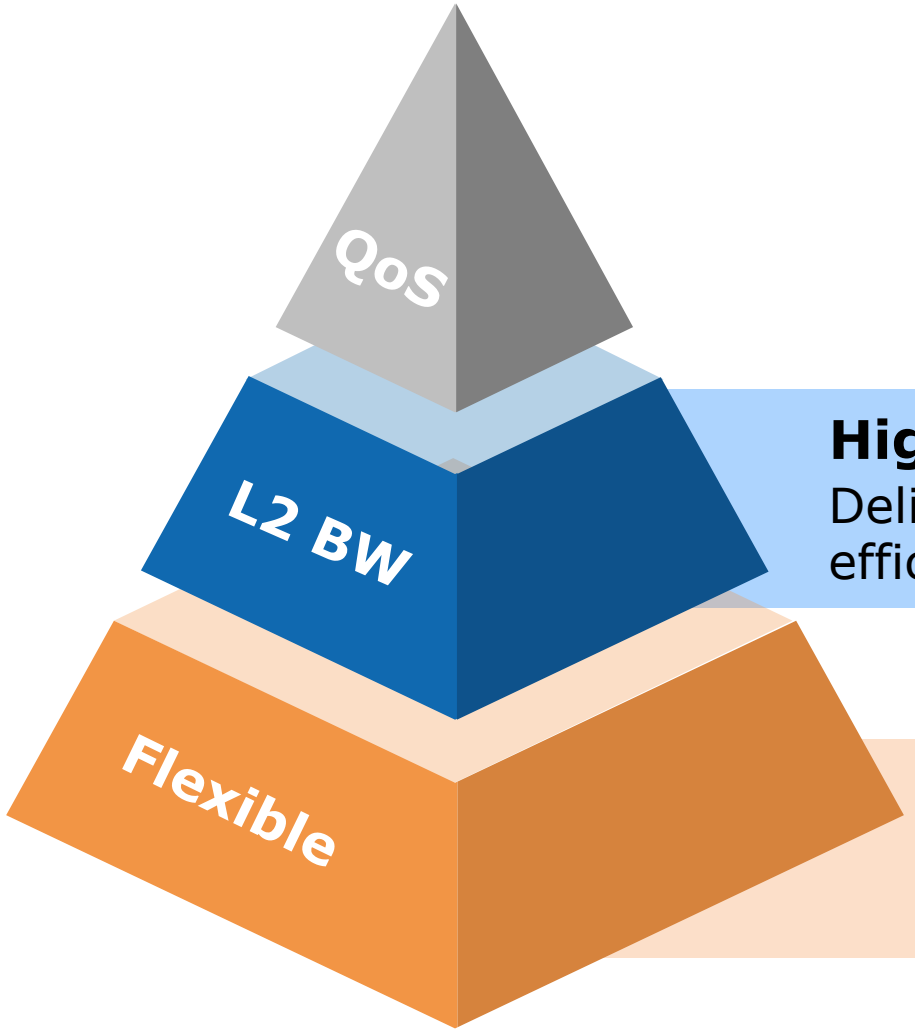
CHIMERA: Putting the Ingredients Together



Transformer Acceleration Cluster (TAC)

Tightly coupled Transformer accelerator with fully programmable RV32IMA cores for efficient and flexible AI inference.

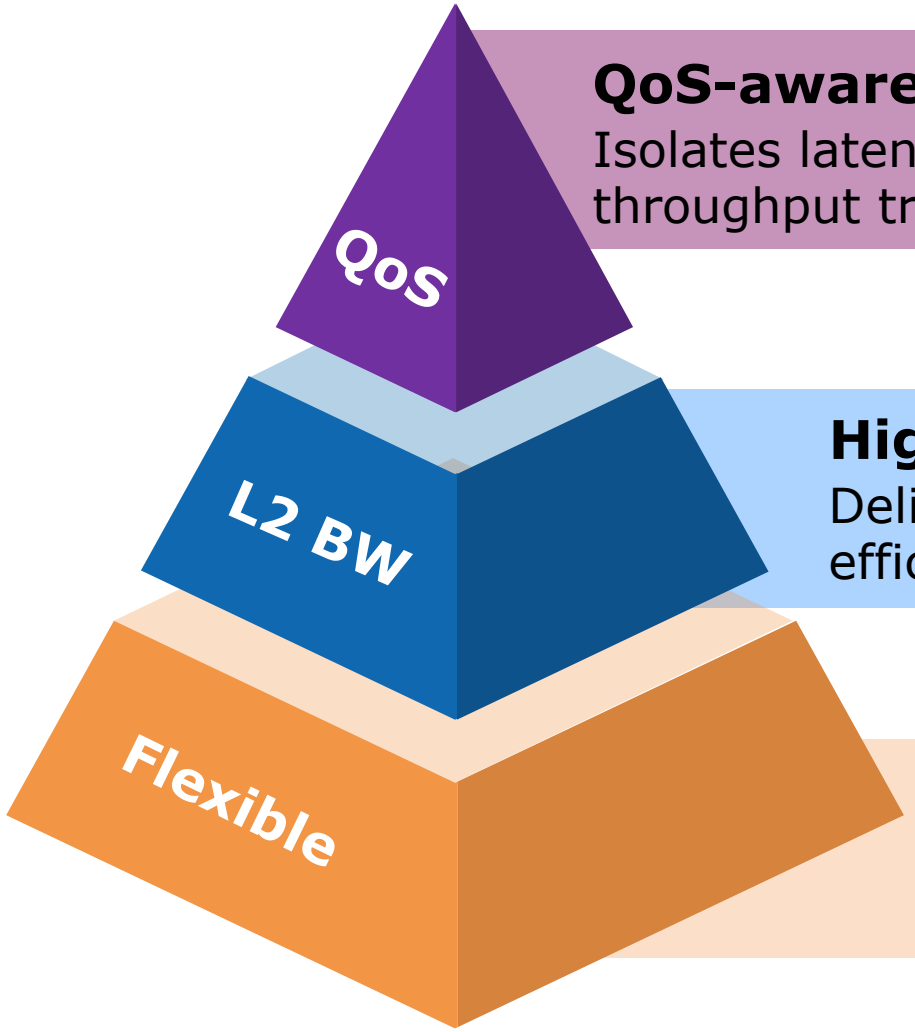
CHIMERA: Putting the Ingredients Together



High-Bandwidth Memory Island
Delivers up to 563 Gb/s aggregate bandwidth for efficient multi-cluster workload parallelization.

Transformer Acceleration Cluster (TAC)
Tightly coupled Transformer accelerator with fully programmable RV32IMA cores for efficient and flexible AI inference.

CHIMERA: Putting the Ingredients Together



QoS-aware L2 architecture
Isolates latency-critical accesses from concurrent high-throughput traffic for fast and predictable service.

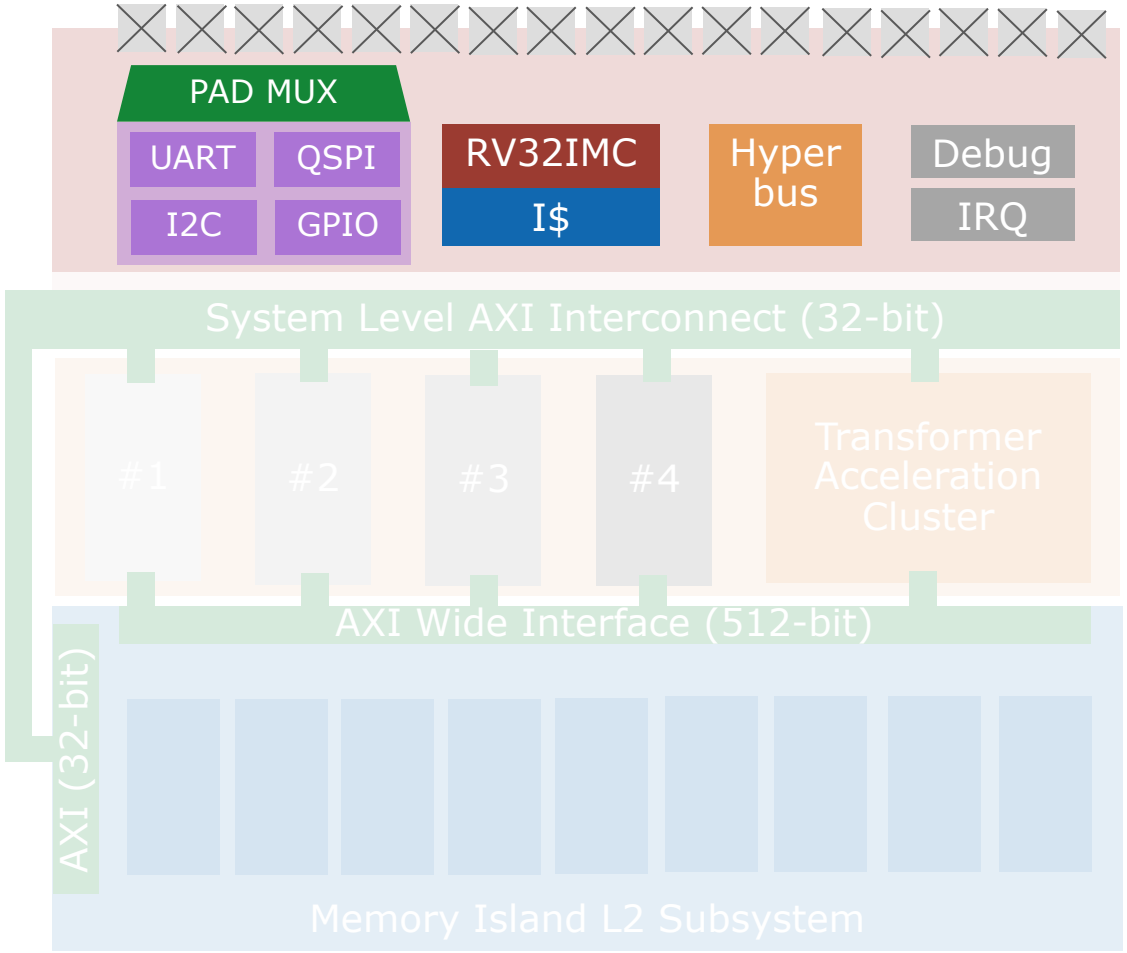
High-Bandwidth Memory Island
Delivers up to 563 Gb/s aggregate bandwidth for efficient multi-cluster workload parallelization.

Transformer Acceleration Cluster (TAC)
Tightly coupled Transformer accelerator with fully programmable RV32IMA cores for efficient and flexible AI inference.

Outline

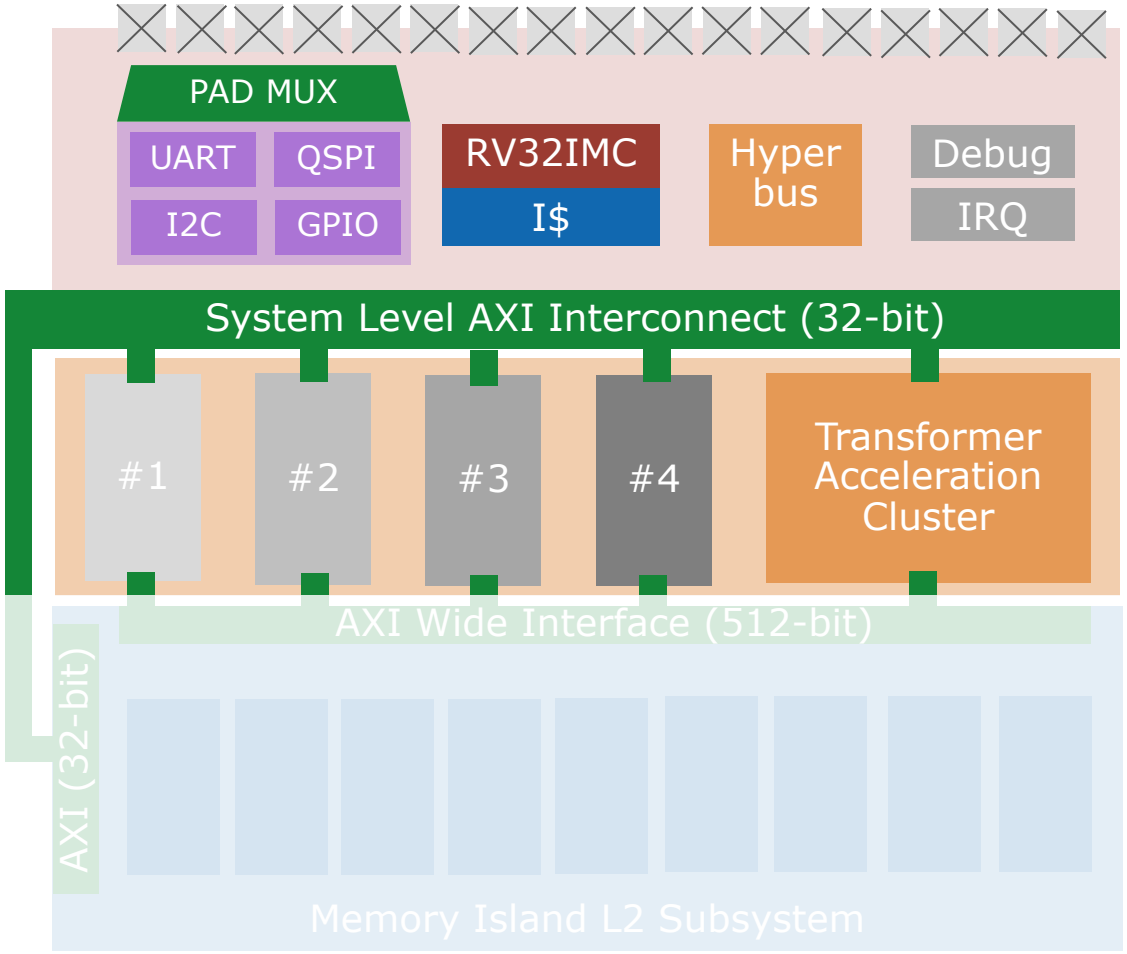
- Background and Motivation
- Architecture Overview
- Results
- Conclusion

CHIMERA: System-Level Overview



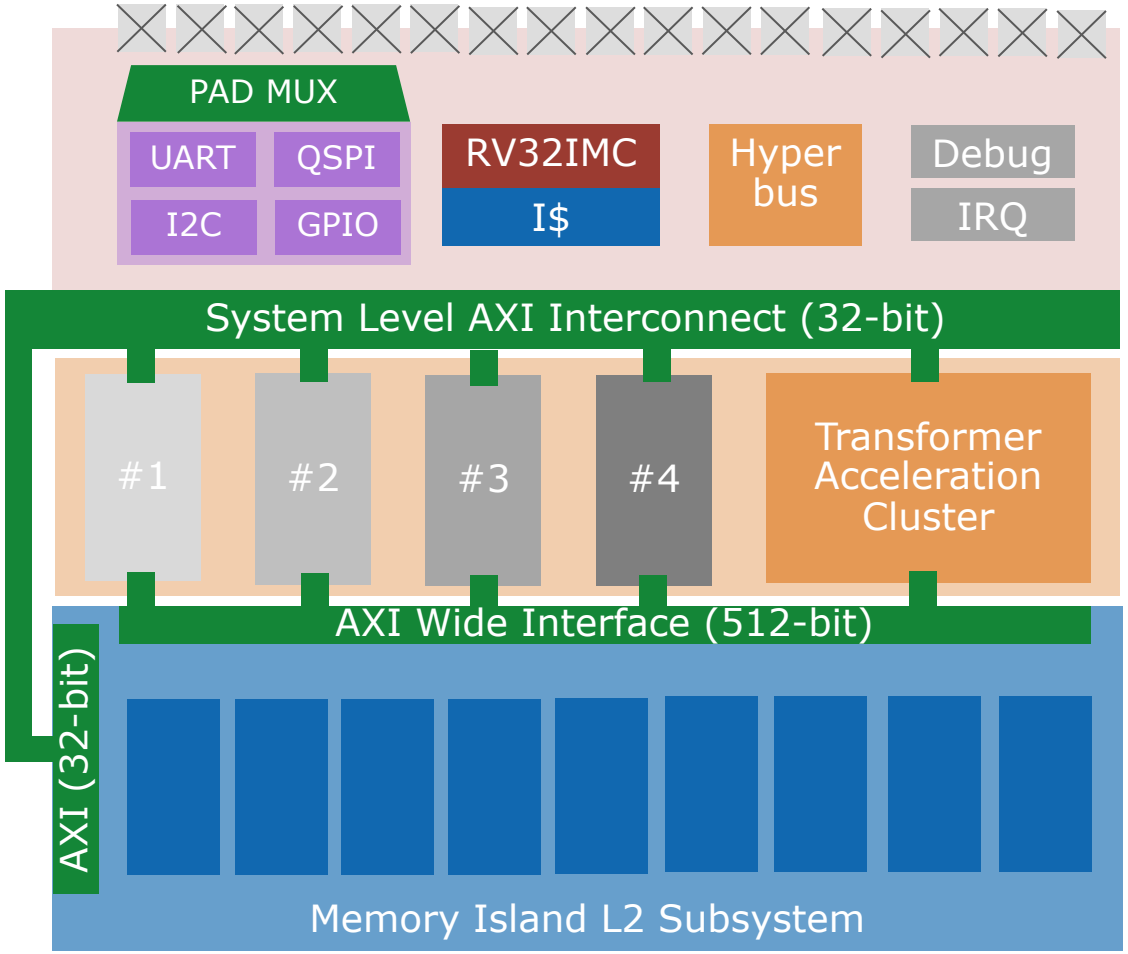
- Host domain:
  - RV32IMC system control core
  - Peripheral subsystem
- Heterogeneous clusters
  - 5 Cluster domains
  - Dedicated Transformer Acceleration Cluster
- L2 Heterogeneous Memory Subsystem:
  - Shared 256 KiB L2
  - 512-bit AXI4 wide interface
  - 32-bit AXI4 narrow interface
  - Internal QoS-aware arbiter

CHIMERA: System-Level Overview



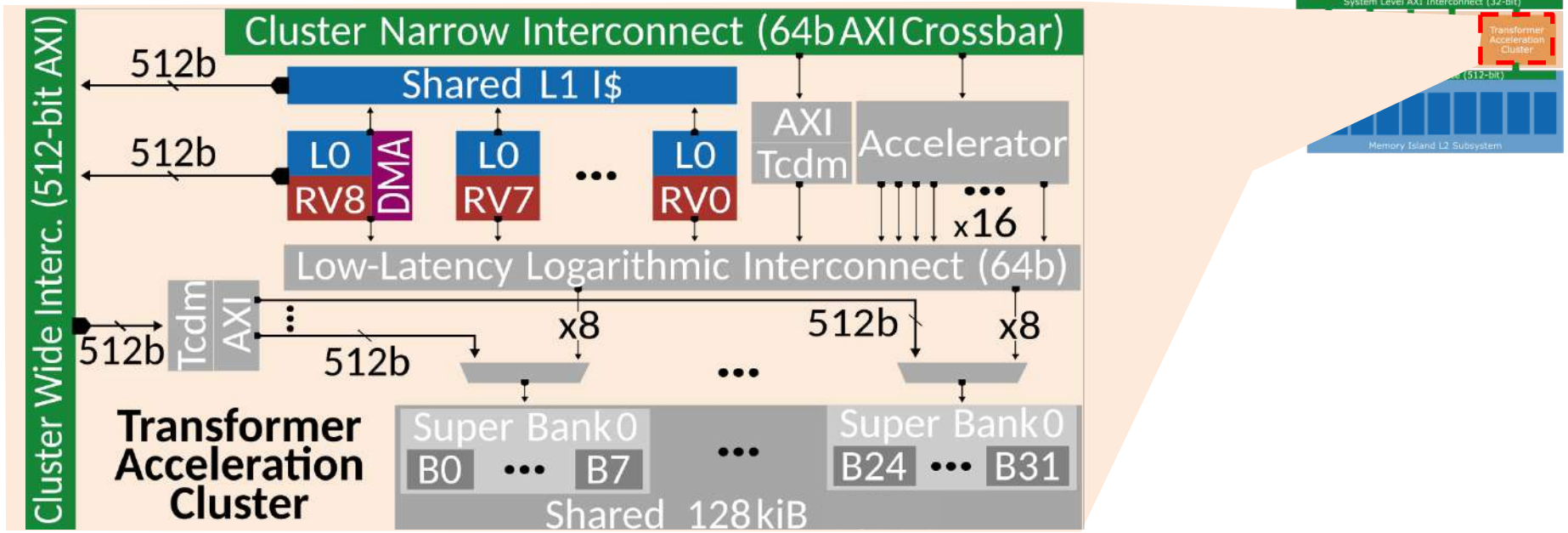
- Host domain:
  - RV32IMC system control core
  - Peripheral subsystem
- Heterogeneous clusters
  - 5 Cluster domains
  - Dedicated Transformer Acceleration Cluster
- L2 Heterogeneous Memory Subsystem:
  - Shared 256 KiB L2
  - 512-bit AXI4 wide interface
  - 32-bit AXI4 narrow interface
  - Internal QoS-aware arbiter

CHIMERA: System-Level Overview



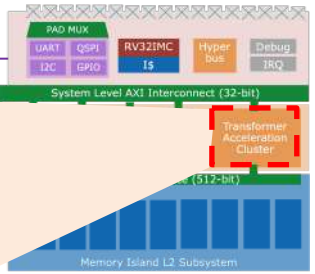
- Host domain:
  - RV32IMC system control core
  - Peripheral subsystem
- Heterogeneous clusters
  - 5 Cluster domains
  - Dedicated Transformer Acceleration Cluster
- L2 Heterogeneous Memory Subsystem:
  - Shared 256 KiB L2
  - 512-bit AXI4 wide interface
  - 32-bit AXI4 narrow interface
  - Internal QoS-aware arbiter

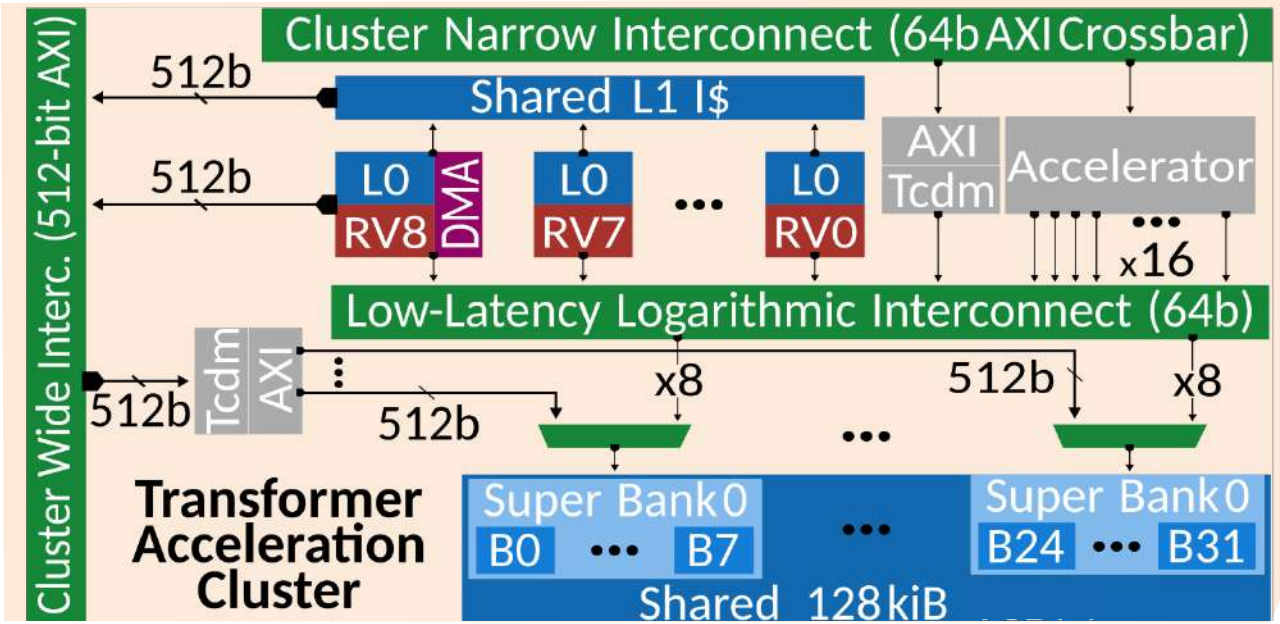
Compute: Transformer Acceleration Cluster



- 8 RV32IMA compute cores + 9th core for DMA control:
  - **512-bit wide ports** to L2 for data-intensive transfers
  - **Narrow control path** for synchronization and control traffic
- Low-Latency Logarithmic Interconnect to SPM
- Tightly coupled Transformer accelerator for GEMM and attention mechanism

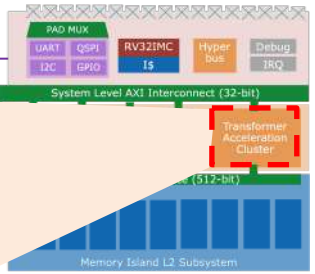
Compute: Transformer Acceleration Cluster

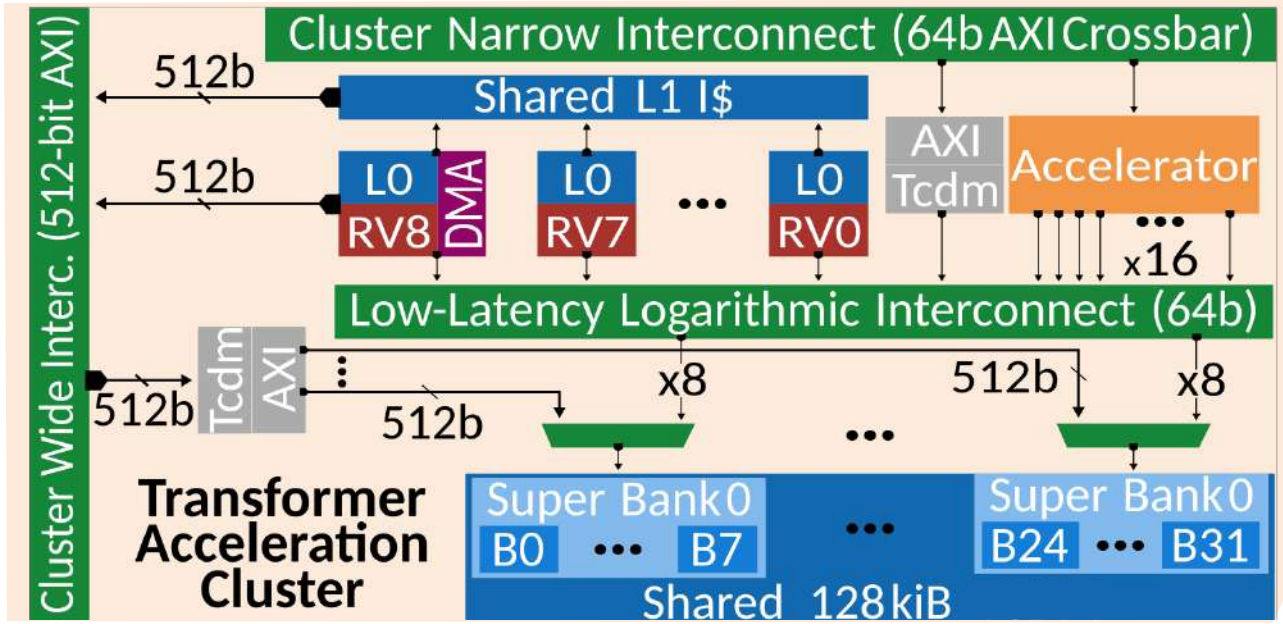




- 8 RV32IMA compute cores + 9th core for DMA control:
  - **512-bit wide ports** to L2 for data-intensive transfers
  - **Narrow control path** for synchronization and control traffic
- Low-Latency Logarithmic Interconnect to SPM
- Tightly coupled Transformer accelerator for GEMM and attention mechanism

Compute: Transformer Acceleration Cluster

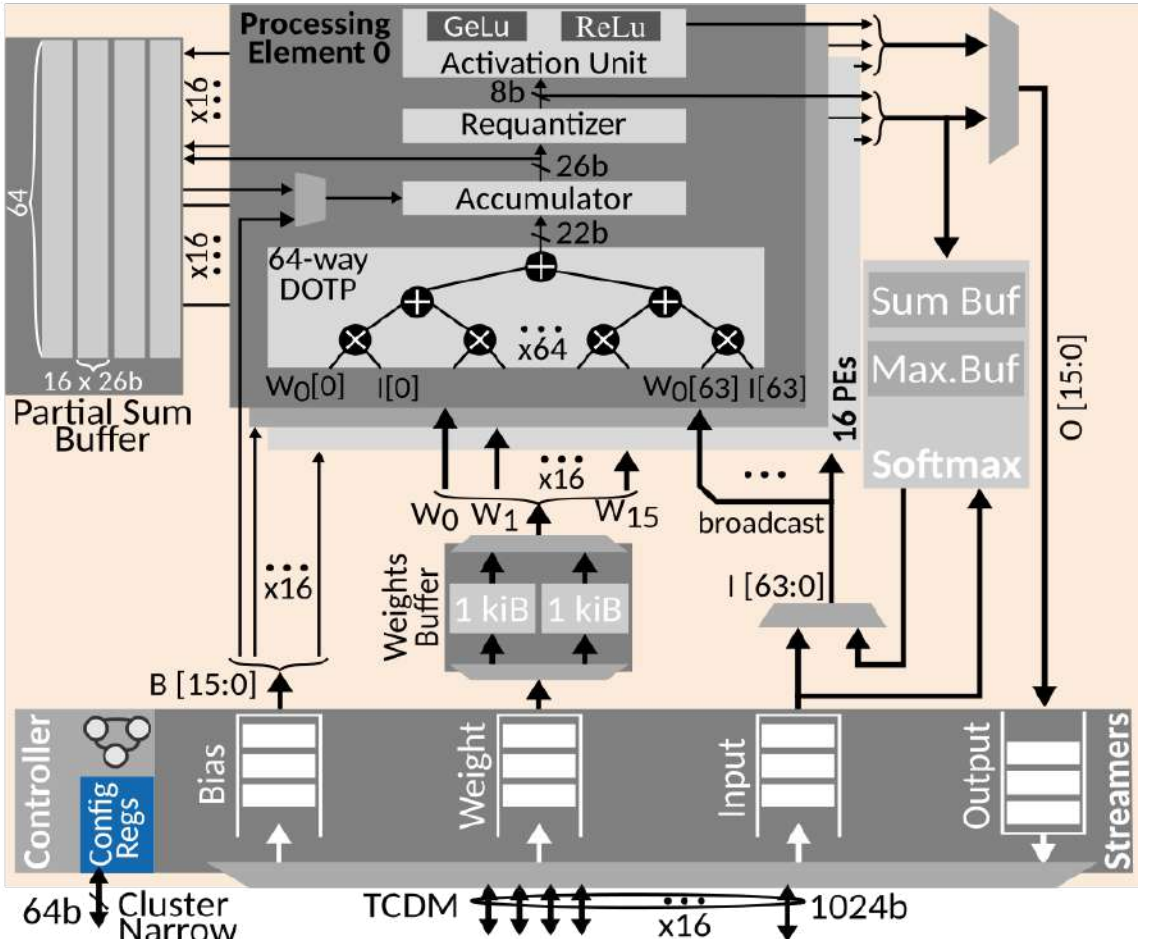




- 8 RV32IMA compute cores + 9th core for DMA control:
  - **512-bit wide ports** to L2 for data-intensive transfers
  - **Narrow control path** for synchronization and control traffic
- Low-Latency Logarithmic Interconnect to SPM
- Tightly coupled Transformer accelerator for GEMM and attention mechanism

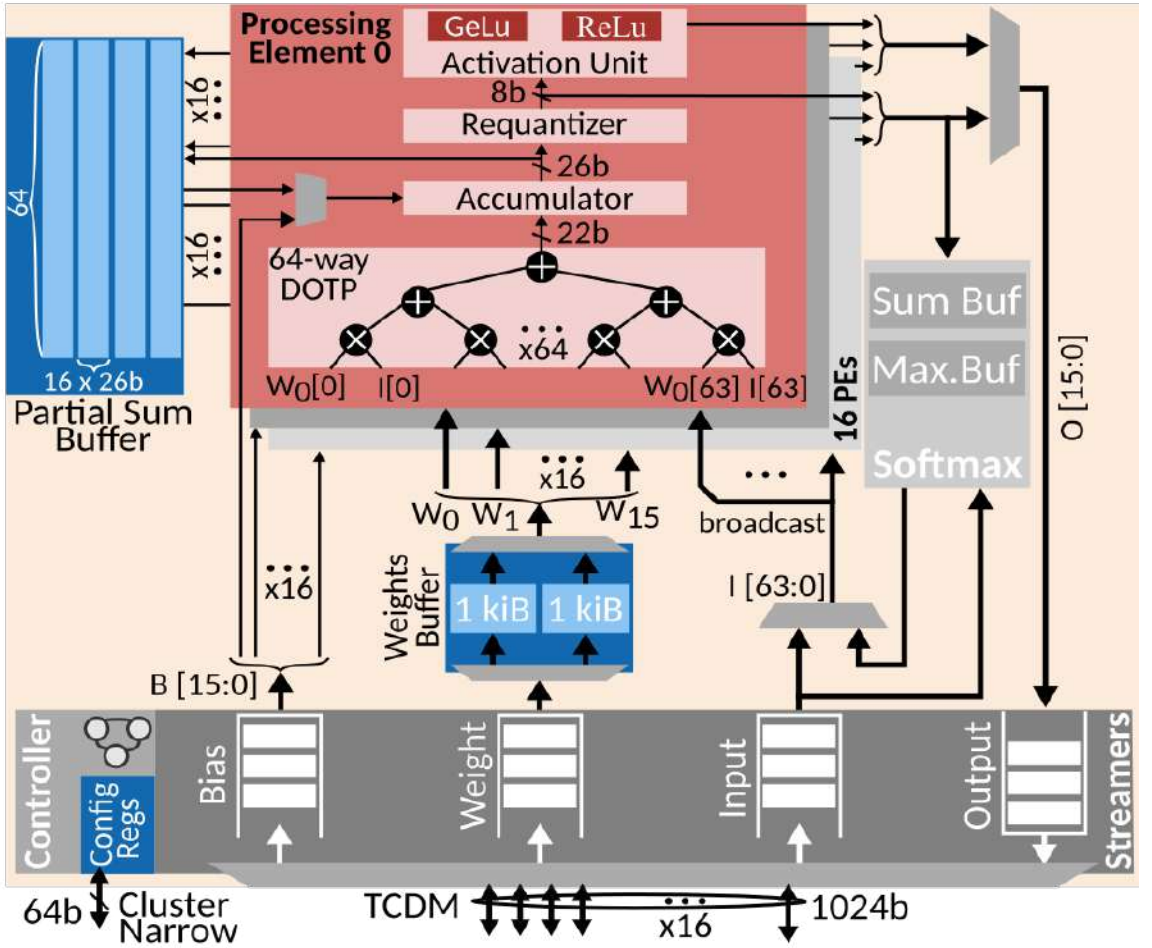
Compute: Transformer Accelerator

- GEMM and Attention support
- 16 Processing Elements:
  - 8-bit weights (*W*) and activations (*I*)
  - 64-way DOTP per PE
  - 2048 ops total
- 4 Data Streamers:
  - 3 for input (*I*), weights (W), and Bias (*B*)
  - 1 for output write-back (*O*)
  - 128 B/cycle bandwidth
- Softmax Engine for Attention:
  - Parallel to PEs DOTP computation
  - 64 softmax/cycle, 44 kGE



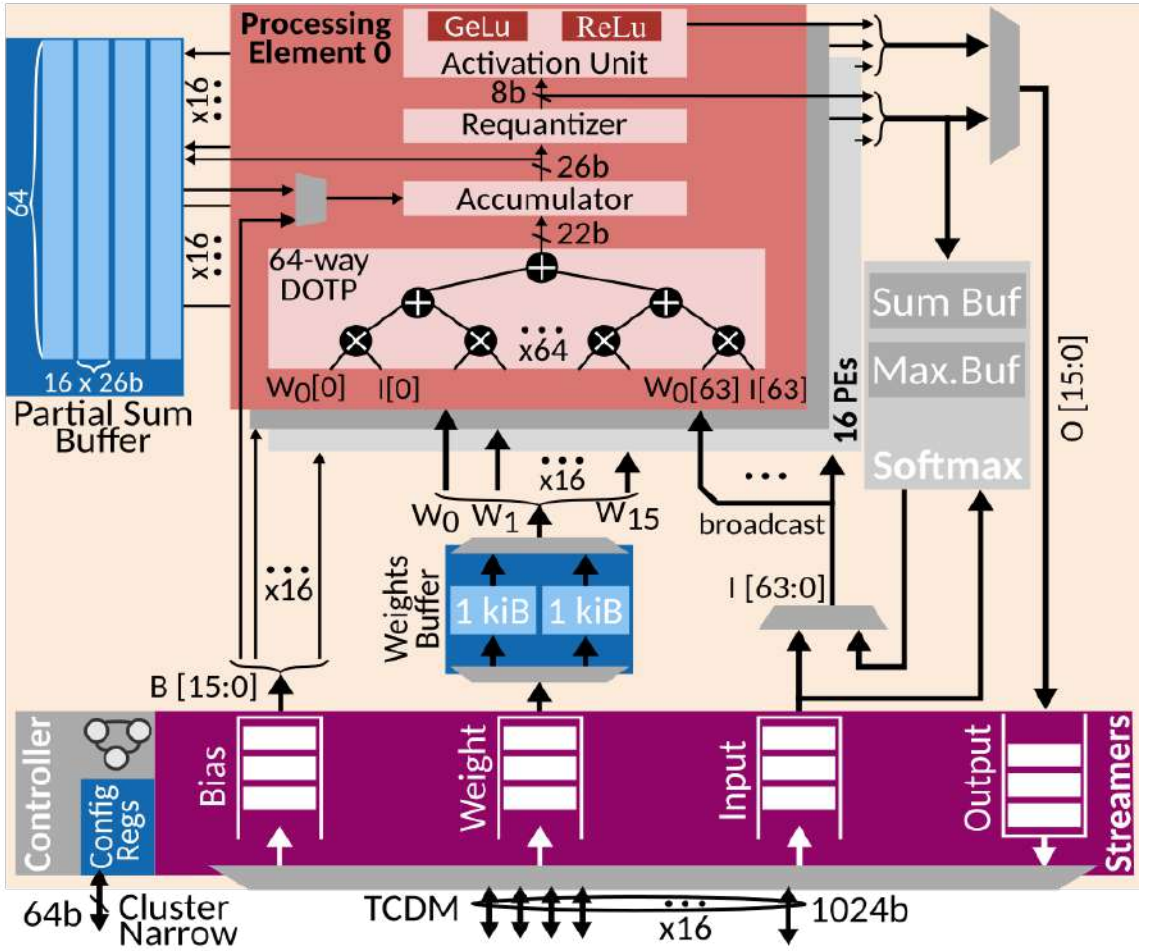
Compute: Transformer Accelerator

- GEMM and Attention support
- 16 Processing Elements:
  - 8-bit weights (*W*) and activations (*I*)
  - 64-way DOTP per PE
  - 2048 ops total
- 4 Data Streamers:
  - 3 for input (*I*), weights (W), and Bias (*B*)
  - 1 for output write-back (*O*)
  - 128 B/cycle bandwidth
- Softmax Engine for Attention:
  - Parallel to PEs DOTP computation
  - 64 softmax/cycle, 44 kGE



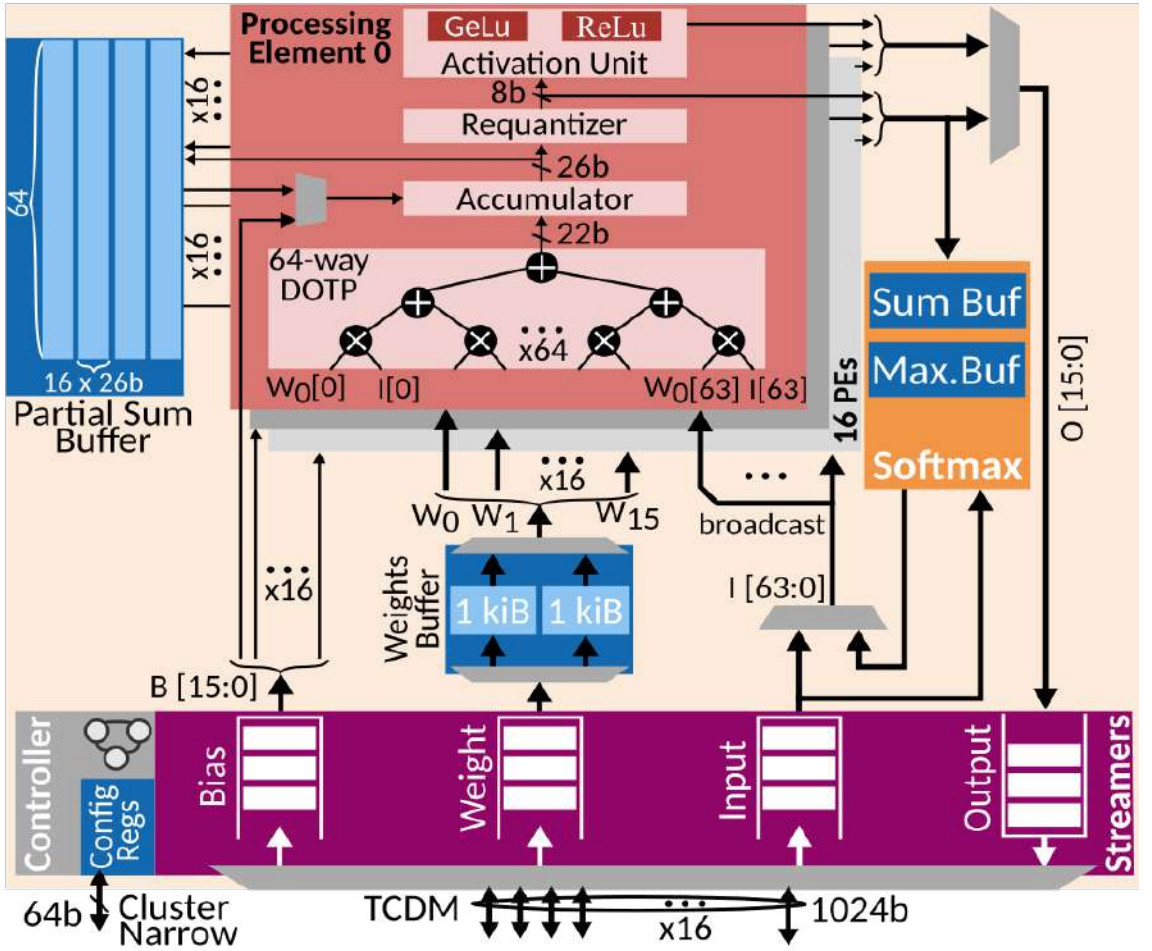
Compute: Transformer Accelerator

- GEMM and Attention support
- 16 Processing Elements:
  - 8-bit weights (*W*) and activations (*I*)
  - 64-way DOTP per PE
  - 2048 ops total
- 4 Data Streamers:
  - 3 for input (*I*), weights (W), and Bias (*B*)
  - 1 for output write-back (*O*)
  - 128 B/cycle bandwidth
- Softmax Engine for Attention:
  - Parallel to PEs DOTP computation
  - 64 softmax/cycle, 44 kGE

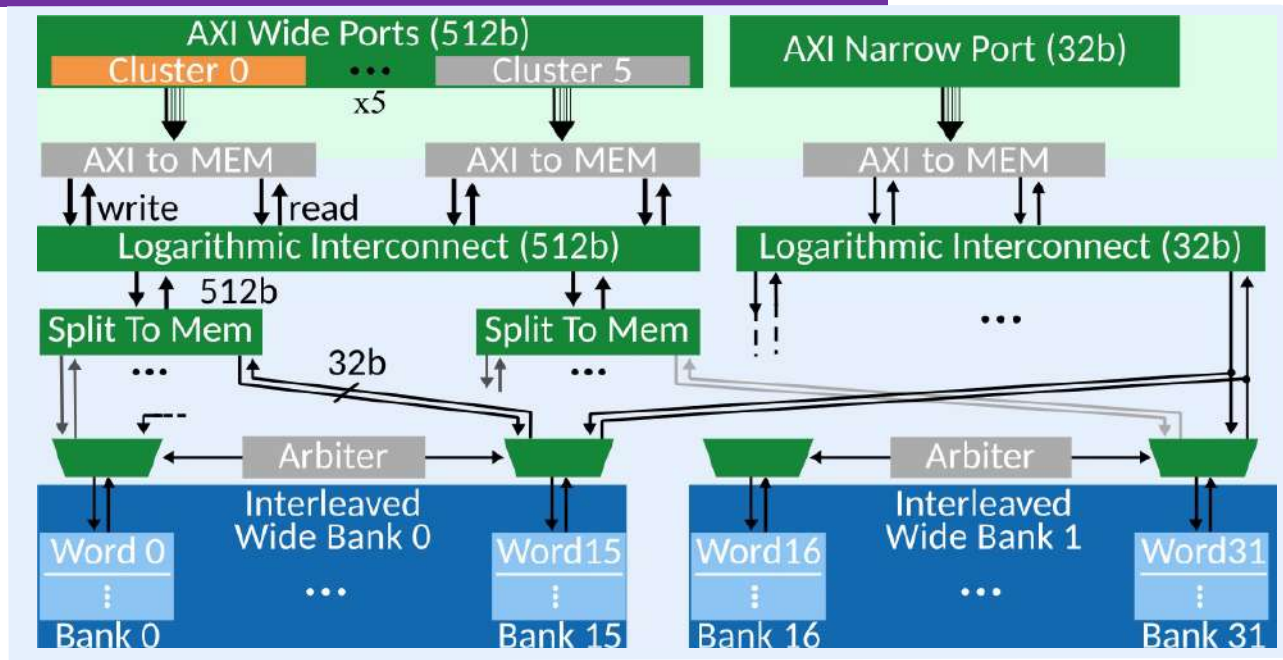


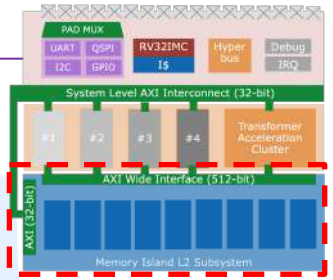
Compute: Transformer Accelerator

- GEMM and Attention support
- 16 Processing Elements:
  - 8-bit weights (*W*) and activations (*I*)
  - 64-way DOTP per PE
  - 2048 ops total
- 4 Data Streamers:
  - 3 for input (*I*), weights (W), and Bias (*B*)
  - 1 for output write-back (*O*)
  - 128 B/cycle bandwidth
- Softmax Engine for Attention:
  - Parallel to PEs DOTP computation
  - 64 softmax/cycle, 44 kGE



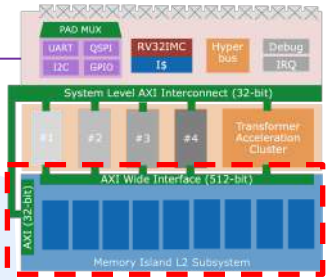
L2: High-Bandwidth Memory Island



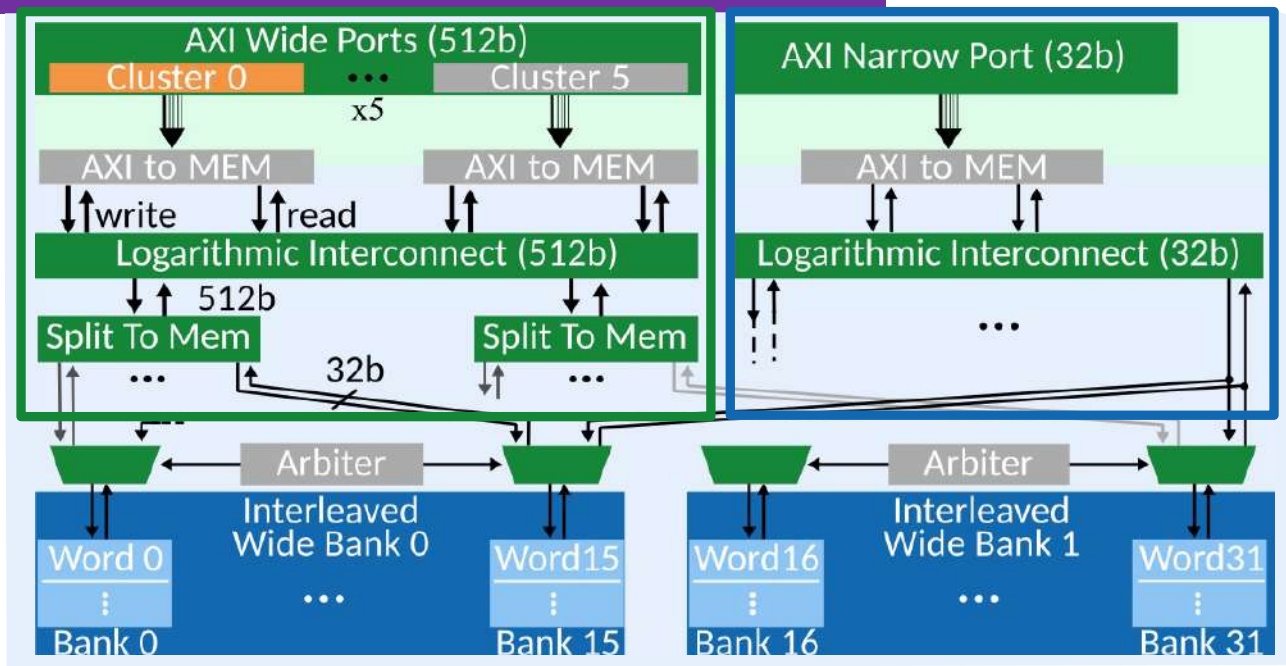


- Heterogeneous interface:
  - **Wide Banks** (512-bit word) for high-throughput accesses
  - **Narrow Banks** (32-bit word) for latency-critical messages
- Two interleaved 128 KiB-wide banks
  - Mitigate access conflicts under parallel traffic
  - Sustain up to **128 B/cycle** read/write bandwidth

L2: High-Bandwidth Memory Island



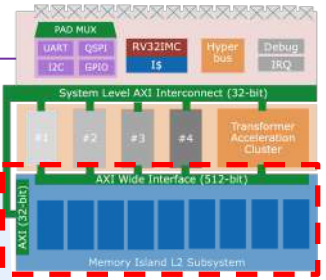
WIDE DATAPATH

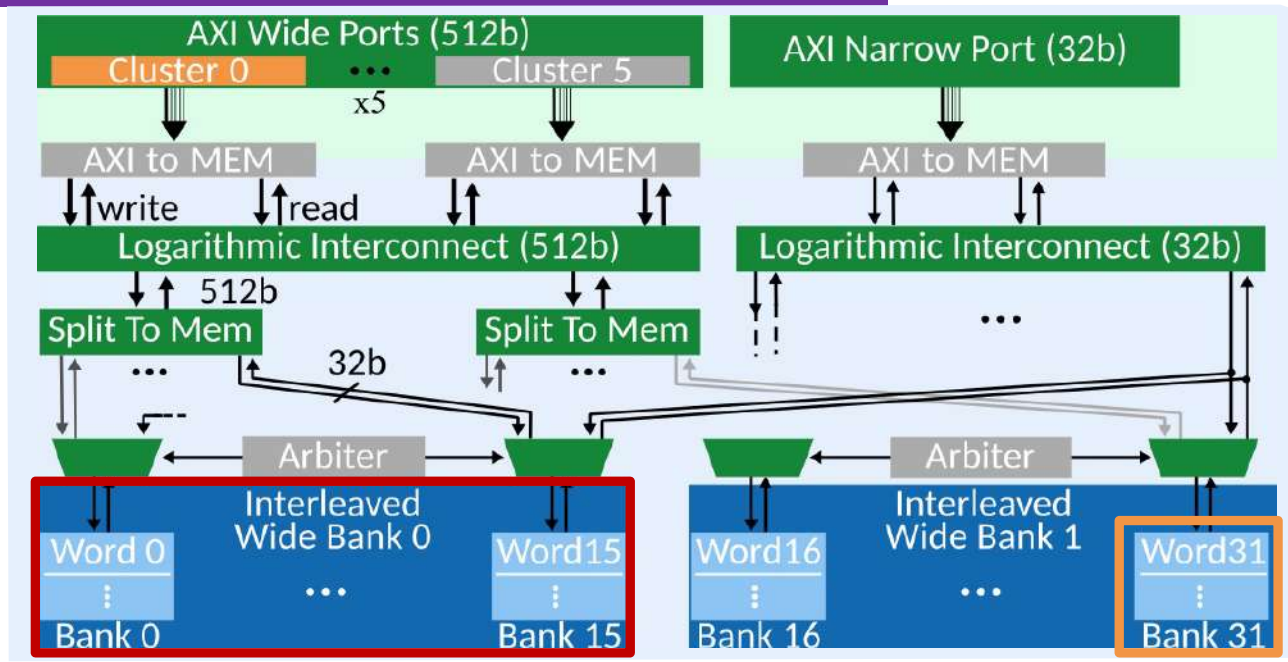


NARROW DATAPATH

- Heterogeneous interface:
  - **Wide Banks** (512-bit word) for high-throughput accesses
  - **Narrow Banks** (32-bit word) for latency-critical messages
- Two interleaved 128 KiB-wide banks
  - Mitigate access conflicts under parallel traffic
  - Sustain up to **128 B/cycle** read/write bandwidth

L2: High-Bandwidth Memory Island



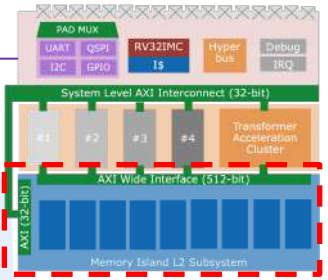


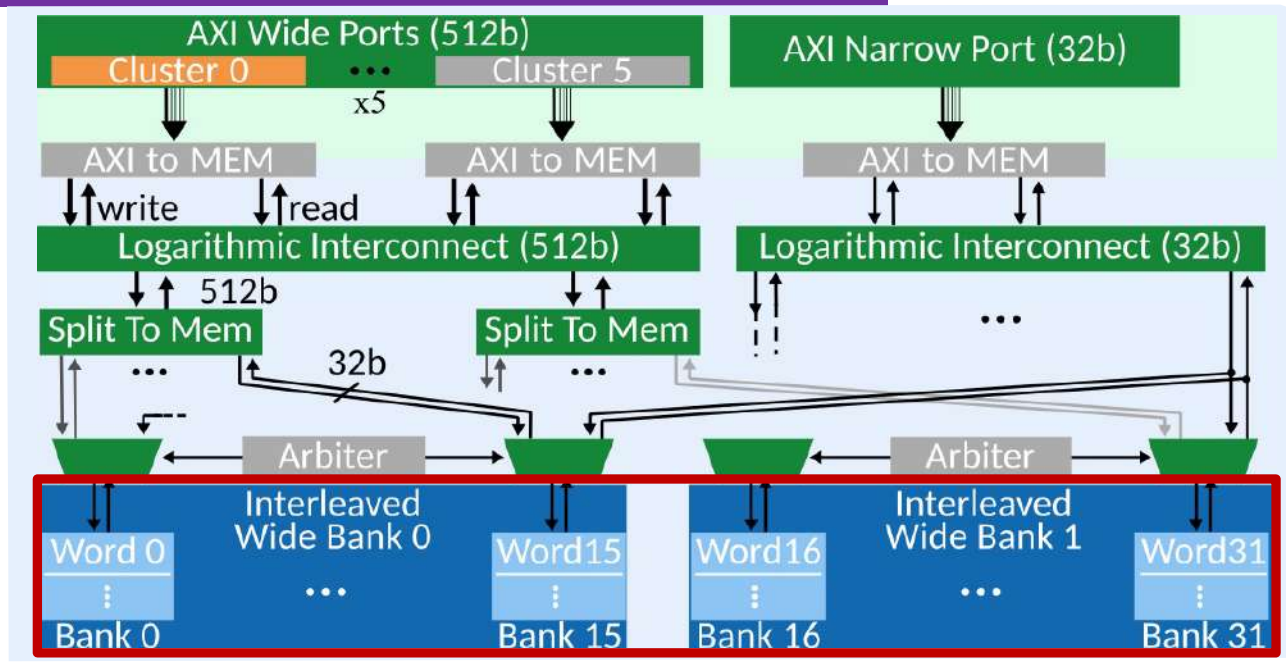
WIDE DATAPATH

NARROW DATAPATH

- Heterogeneous interface:
  - **Wide Banks** (512-bit word) for high-throughput accesses
  - **Narrow Banks** (32-bit word) for latency-critical messages
- Two interleaved 128 KiB-wide banks
  - Mitigate access conflicts under parallel traffic
  - Sustain up to **128 B/cycle** read/write bandwidth

L2: High-Bandwidth Memory Island



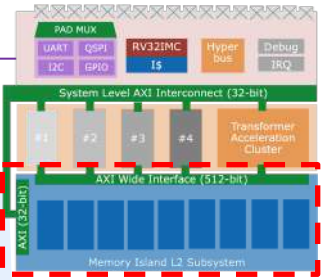


WIDE DATAPATH

NARROW DATAPATH

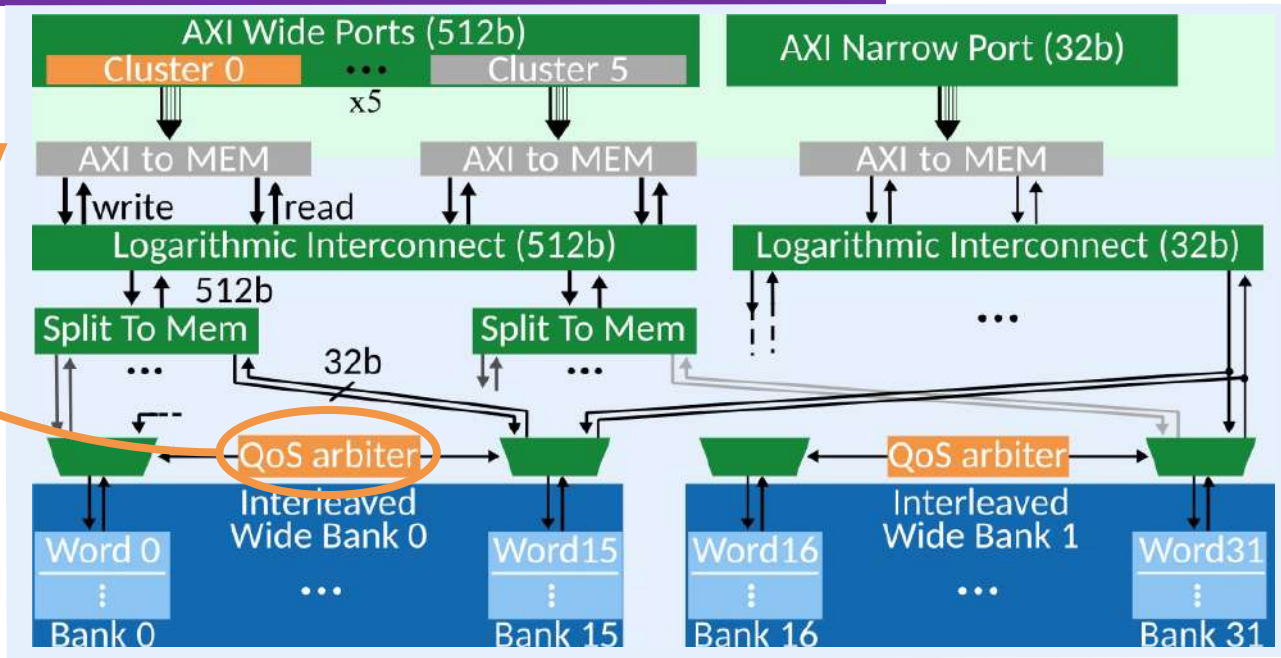
- Heterogeneous interface:
  - **Wide Banks** (512-bit word) for high-throughput accesses
  - **Narrow Banks** (32-bit word) for latency-critical messages
- Two interleaved 128 KiB-wide banks
  - Mitigate access conflicts under parallel traffic
  - Sustain up to **128 B/cycle** read/write bandwidth

L2: QoS-Aware Memory Island



Fixed or Bounded Priority

WIDE DATAPATH



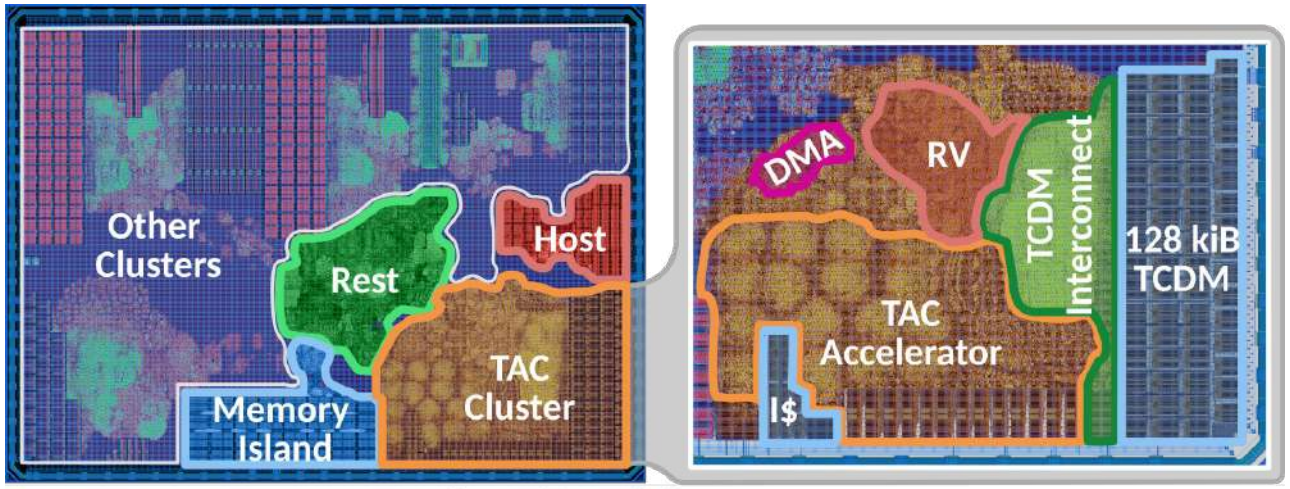
NARROW DATAPATH

- QoS-aware arbitration between wide and narrow accesses:
  - Low-latency service for latency-critical messages
- Two arbitration schemes:
  - **Fixed priority** to narrow accesses when narrow traffic is system-regulated
  - **Bounded priority** to prevent starvation of wide traffic under continuous contention

Outline

- Background and Motivation
- Architecture Overview
- Results
- Conclusion

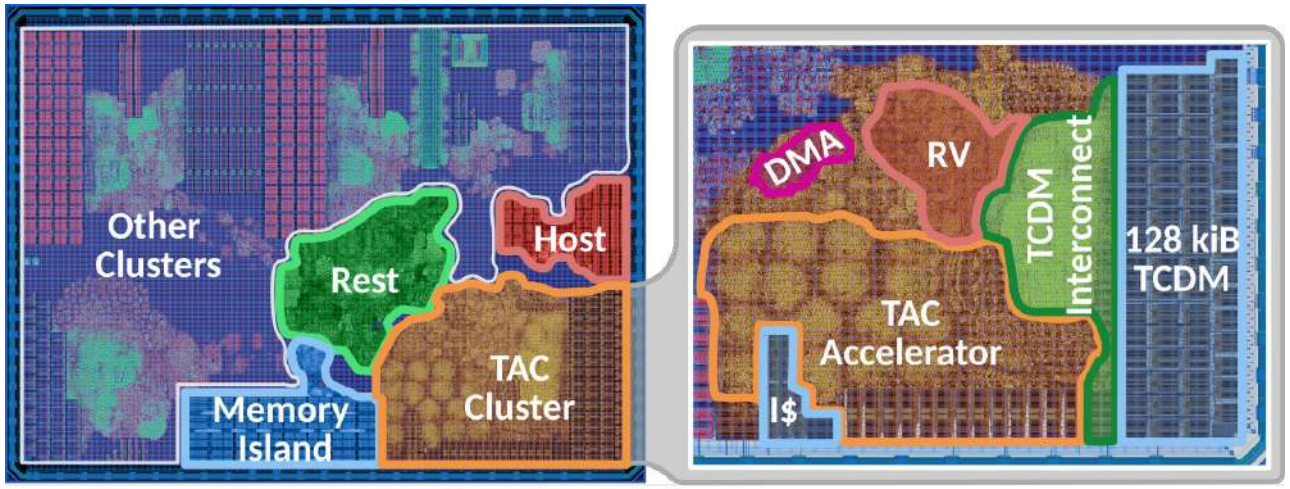
Chimera Die Overview



Technology	22nm FDX
Die Area	3.19mm^2
Core Supply	0.8V
Frequency	500MHz
L2 BW	563 Gb/s
Power*	106 mW

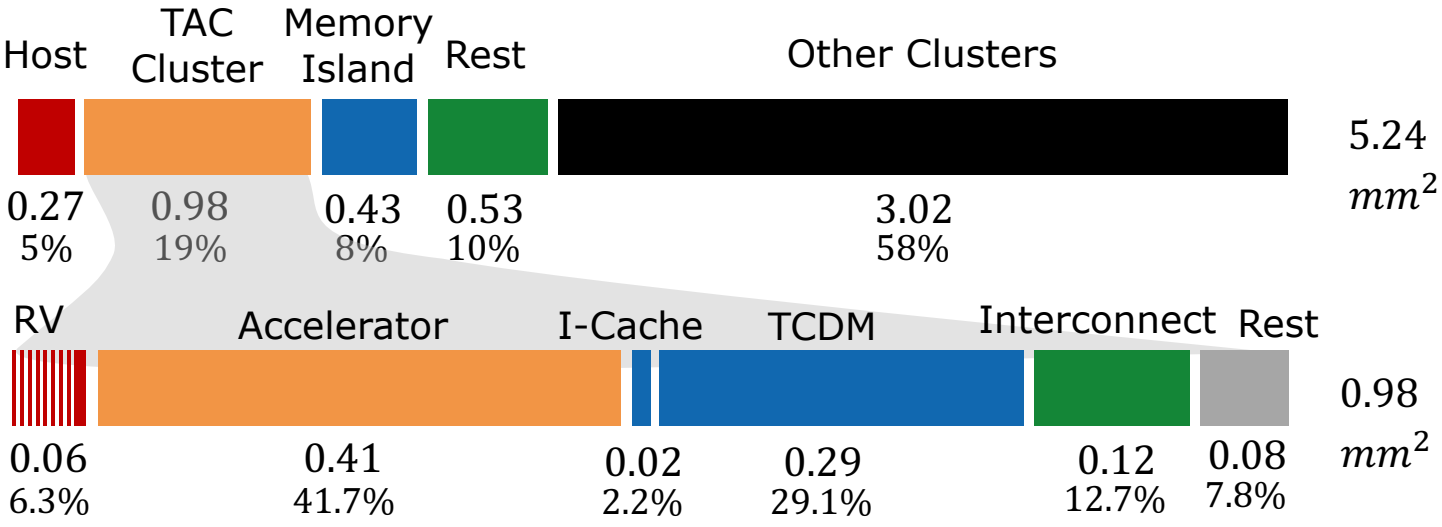
* @0.6V

Chimera Die Overview

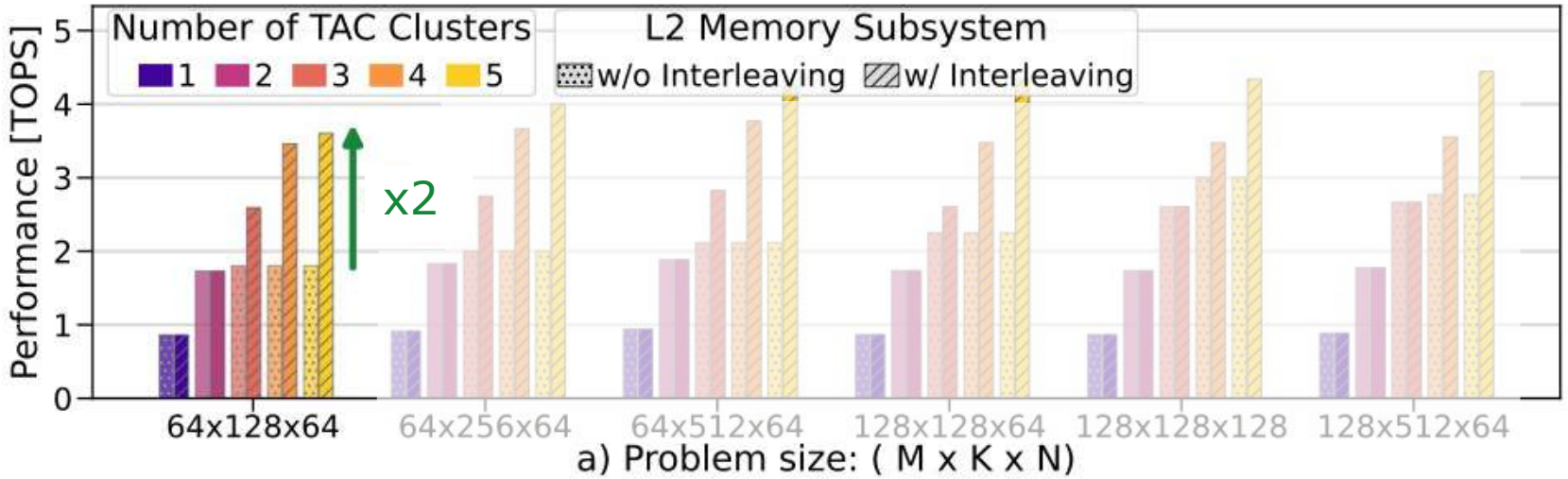


Technology	22nm FDX
Die Area	3.19mm^2
Core Supply	0.8V
Frequency	500MHz
L2 BW	563 Gb/s
Power*	106 mW

* @0.6V

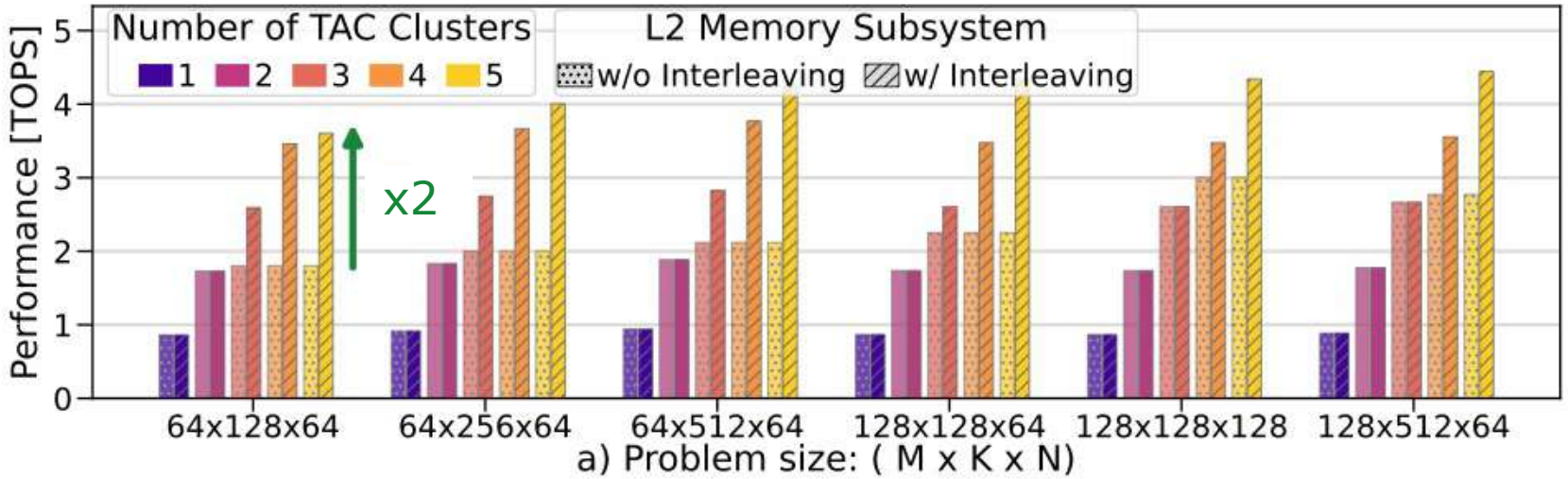


Memory Island Evaluation



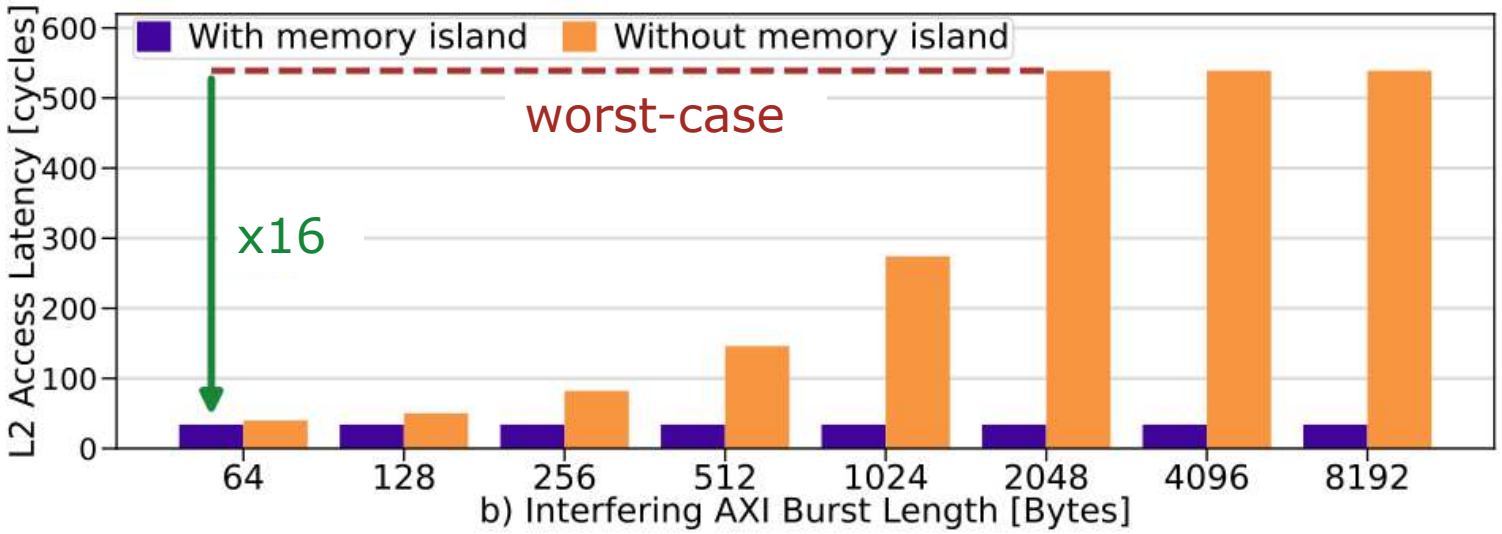
- Aggregated bandwidth under L2 multi-cluster access:
  - Memory-bound beyond 2 active clusters
  - Interleaved wide banks mitigate shared-L2 conflicts → Up to **2x** MATMUL performance
- QoS evaluation:
  - 20,000 32-bit host accesses through the narrow L2 interface
  - Concurrent DMA bursts target the same memory region
  - Bounded, burst-independent latency with up to a **16x** reduction

Memory Island Evaluation



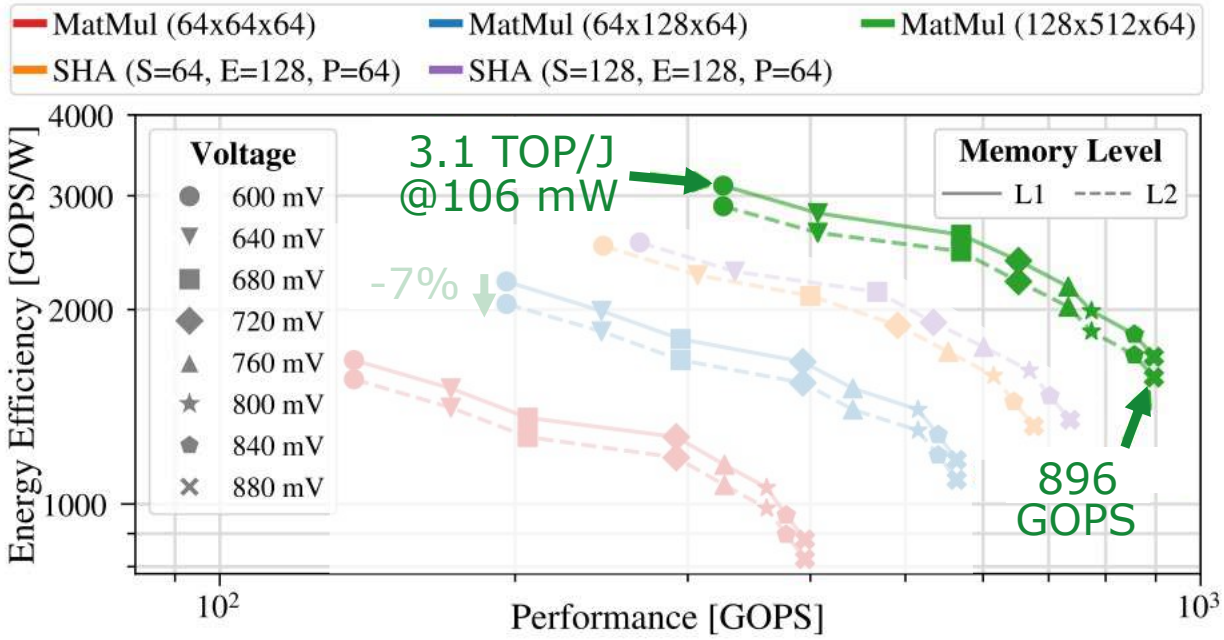
- Aggregated bandwidth under L2 multi-cluster access:
  - Memory-bound beyond 2 active clusters
  - Interleaved wide banks mitigate shared-L2 conflicts → Up to **2x** MATMUL performance
- QoS evaluation:
  - 20,000 32-bit host accesses through the narrow L2 interface
  - Concurrent DMA bursts target the same memory region
  - Bounded, burst-independent latency with up to a **16x** reduction

Memory Island Evaluation



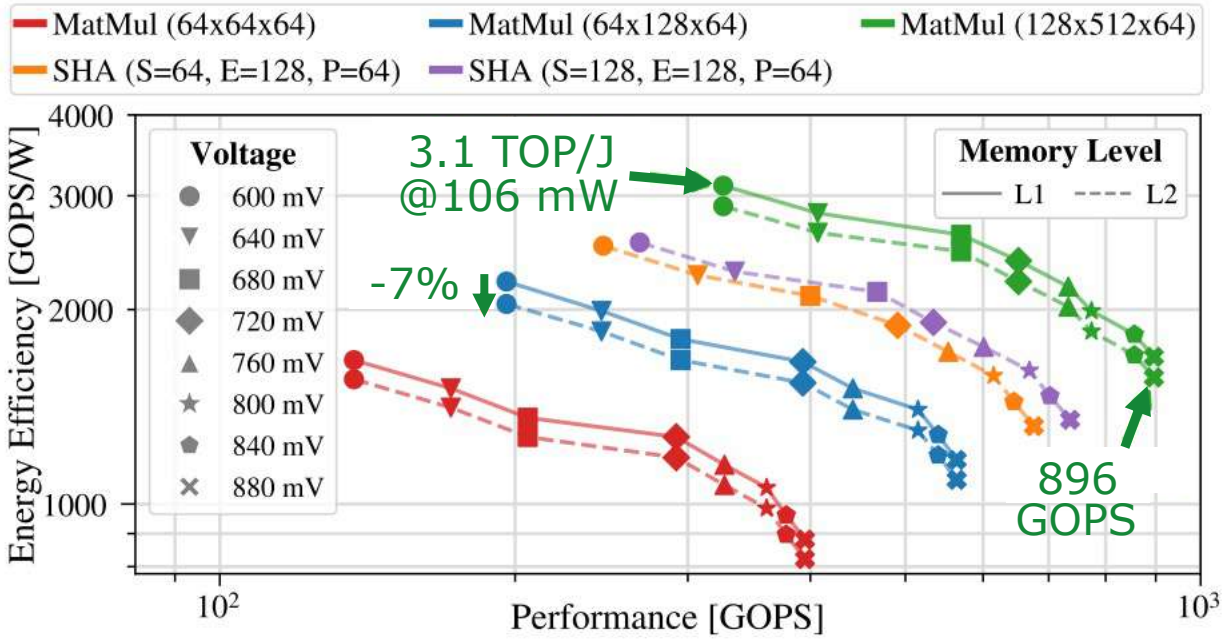
- Aggregated bandwidth under L2 multi-cluster access:
  - Memory-bound beyond 2 active clusters
  - Interleaved wide banks mitigate shared-L2 conflicts → Up to **2x** MATMUL performance
- QoS evaluation:
  - 20,000 32-bit host accesses through the narrow L2 interface
  - Concurrent DMA bursts target the same memory region
  - Bounded, burst-independent latency with up to a **16x** reduction

Energy and Perfomance Evaluation



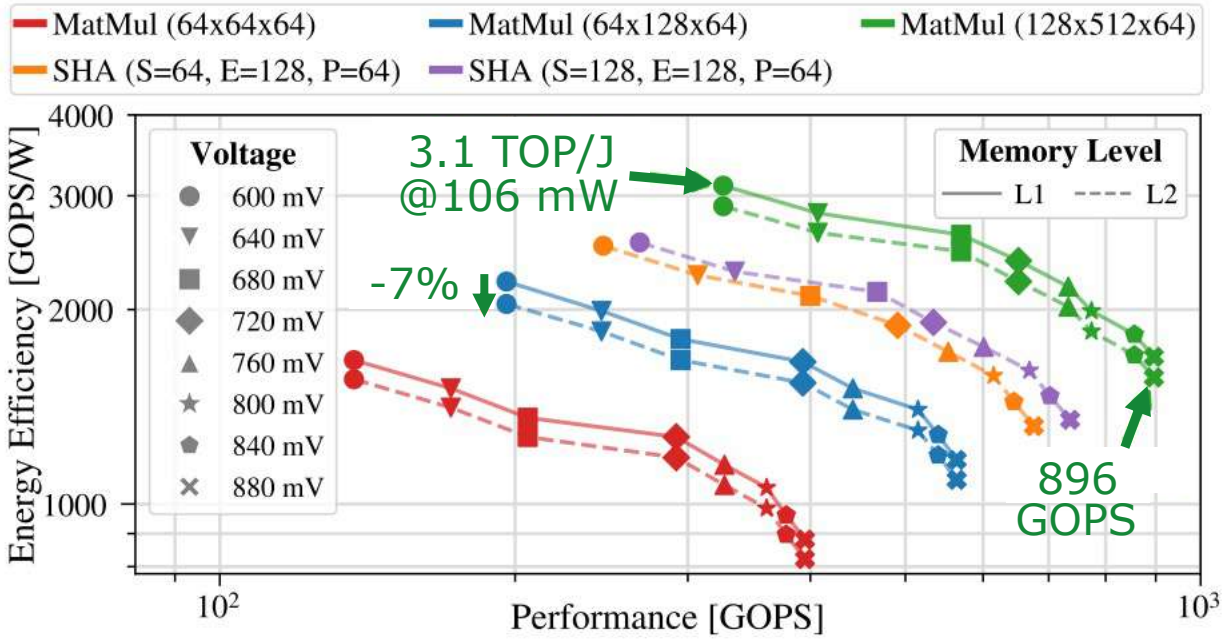
- Low-Voltage (@0.6 V) → up to **3.1 TOPS/W**
- High-Voltage (@0.88V) → up to **896 GOPS**
- L1-to-L2 transfers → only **7%** energy decrease when running at the system level
- Transformer workloads → up to **737 GOPS** and **2.54 TOPS/W**

Energy and Perfomance Evaluation



- Low-Voltage (@0.6 V) → up to **3.1 TOPS/W**
- High-Voltage (@0.88V) → up to **896 GOPS**
- L1-to-L2 transfers → only **7%** energy decrease when running at the system level
- Transformer workloads → up to **737 GOPS** and **2.54 TOPS/W**

Energy and Perfomance Evaluation



	MobileBERT	Whisper-Tiny Encoder	DINOv2-S
Model Complexity [GOP]	7.4	9.7	11.7
Throughput [1/s]	7.7-21	2.0-5.4	1.3-3.3
Energy [mJ]	9.2-16	36-72	60-118

- Low-Voltage (@0.6 V) → up to **3.1 TOPS/W**
- High-Voltage (@0.88V) → up to **896 GOPS**
- L1-to-L2 transfers → only **7%** energy decrease when running at the system level
- Transformer workloads → up to **737 GOPS** and **2.54 TOPS/W**

Comparison with State-of-the-Art

	Ayaka JSSCC'24	EVA VLSI'25	VLSI'25	TinyVers VLSI'22	ESSERC'24	This Work
Technology	28nm 430MHz@1V	18nm 1500MHz@0.9V	22nm 400MHz@0.9V	22nm 150MHz@0.9V	18nm 500MHz@0.87V	22nm, 550MHz@0.88V
Processing Element	Transformer Accelerator	1 × RISC-V + PE Array	Transformer Accelerator	1 × RI5CY + ML Accelerator	1 × RISC-V + 128 PEs TPU	**1 × RV32IMC + 9 × RV32IMA + 1 × TAC Accel.**
GP-Host/Prog. Accelerator	✗ / ✗	✗ / ✗	✗ / ✗	✓ / ✗	✓ / ✗	✓ / ✓
L2 Aggr. Bandwidth	-	-	-	-	-	**563 Gb/s**
Data Precision	INT16/18	FP16	INT4/8	INT2/4/8	INT8	INT8
Peak Efficiency	**49.7*** TOPS/W (@0.68 V)	0.71 TOPS/W (@0.49 V)	12.52** TOPS/W (@0.55 V)	2.47 TOPS/W (@0.4 V)	3.25*** TOPS/W (@0.5V)	**3.1 TOPS/W** (@0.6 V)
Area Efficiency	**607** GOPS/mm^2	96.5 GOPS/mm^2	158.62 GOPS/mm^2	2.82 GOPS/mm^2	-	**281 GOPS/mm^2**

* The highest values are measured assuming 90% output sparsity.
** Value obtained with one dense and one 87.5% sparse input.
*** Scaling from 0.5 to 0.6 V, the efficiency is 27% lower than Chimera.

Outline

- Background and Motivation
- Architecture Overview
- Results
- Conclusion

Conclusion

CHIMERA

A **flexible** and **efficient** AI-MCU for **evolving** extreme-edge AI, with a **high-bandwidth QoS-aware** memory subsystem

3.1 TOPS/W and **1.37x** higher energy efficiency over SoA

563 Gb/s aggregate L2 **bandwidth** for multi-cluster execution

QoS-aware L2 for predictable service, with **34-cycle** worst-case latency and up to **16x** reduction

https://github.com/pulp-platform/chimera