Next Gen MTIA -
Recommendation Inference Accelerator

Mahesh Maddury, Pankaj Kansal, Olivia Wu

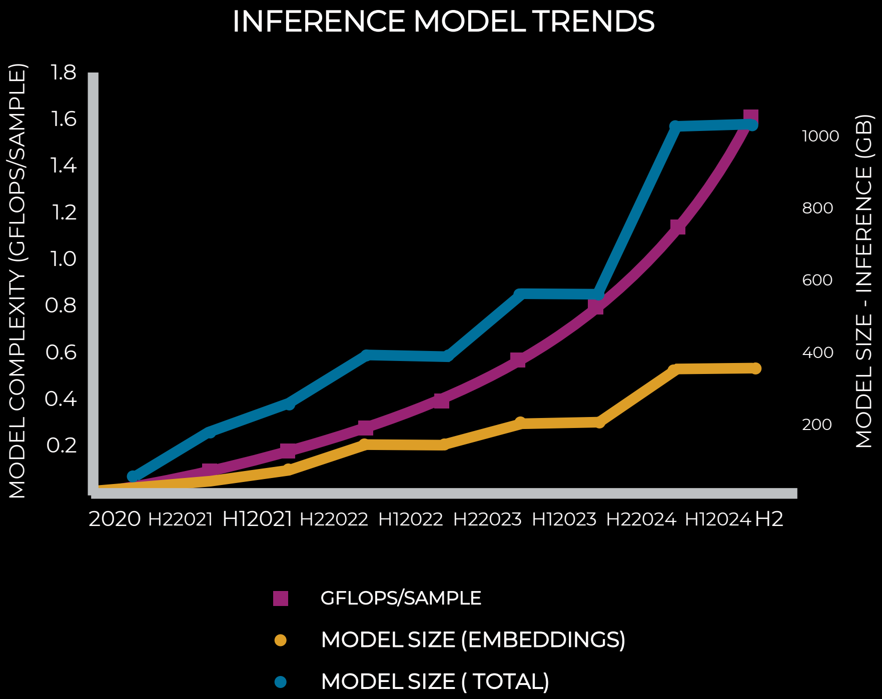
Meta

Acknowledgement

Prahlad Venkatapuram, Linda Cheng, Willie Wei, Harikrishna Reddy, Nagesh Sreedhara, Dheepak Jayaraman, Jae Won Seo, Soheil Gharahi, Anurag Jain, Vinod Bhat, Krishna Datla, Srinu Sunkara, Surendra Rajupalem, Inder Bhasin, Krishna Noru, Adam Hutchin, Rakesh Nattoji, Brian Ko, Kevin Quinn, Himanshu Sanghavi, Yogesh Agrawal, Ushma Bharucha, Harshitha Pilla, Adrian Lewis, Eric Qin, Duckhwan Kim, Giri Prasanna M, Bob Dreyer, Amin Firoozshahian, Gurdeepak Grewal, Sameer Shripad, Krishna Thottempudi, Ramakoteswara Rao Chennamsetti, K Rajesh Jagannath, Xiao He, Tuo Wang, Keerthi Manthati, Eric Wang, Kalyan Chakravarthy Kurapati, Joyce Wu, Vaibhav Desai, Sai Ma, Tongji Wang, Aadesh Karki, Teja Kala, Bhargav Alluri, Hao Jin, Eda Sahin, Arun Karthik Sengottuvel, Satyanarayana Arvapalli, Kalyan Chakravarthy Kurapati, Kaushal Gandhi, Samantak Gangopadhyay, Andrew Xi, Siddharth Bhargav, Michael Wrighton, Hengsheng Chen, Bala Balasubramanian, Harish Aepala, Ram Gupta, Aarti Ramesh, Ejaz Mohammed, Sivanandh Ramadass, Anshul Bansal, Gautham Suryanarayana, Changru Liu, Michael Xiong, Pritam Chopda, Joseph Saiju Thirunilathu, Gaurav Sudheer Bhangle, Shayan Ebrahimi, Ferze Patawaran, Lei Gao, Sumanth Batchu, Brian Dodds, Pritesh Modi, Bhasker Jakka, Anmol Shridhar Poojary, Ashwin Narasimha, Mick Baikie, Bradley Hall, Greeshma Sasikumar, Lara Khamisy, Lawrence Jacob, Mrinalini Padmanabhan, Patrick Hanley, Shayan Raisi, Erum Kazi, Athul Asokan Thulasi, Krish Skandakumaran, Pavan Shetty, Sanjay Desai, Indu Kalayraman, Karthik Narayanan, Roman Levenstein, Kaustubh Gondkar, Jack Montgomery, Sameer Abu Asal, Joe Shajrawi, Kunming Ho, Saritha Dwarakapuram, Mike Tsai, Hangchen Yu, Richard Li, Xiaolong Xie, Siyan Lin, Jordan Fix, Junhan Hu, Ilia Cherniavskii, Aleksandr Malyshev, Mandeep Singh Grang, Garima Gupta, Giancarlo Colmenares, Wang Xu, Cyrus Daruwala, Site Cao, Jun Luo, Mitch Gamburg, Ryan Xia, Qiming Lu, Jay Banerjee, Raviteja Chinta, Joel Coburn, Cao Gao, Kimble Houck, Aravind Sukumaran-Rajam, Andrey Kareev, Satish Nadathur, Jiyuan Zhang, Nasibeh Nasiri, Shintaro Iwasaki, Amy Yang, Jiawei Zhang, Jongsoo Park, Summer Deng, Fangran Xu, Eric Ouyang, Gautam Patil

Motivation

Meta Inference Workload Trends



Deep Learning Recommendation Models (DLRM) are increasing in model size (GB) and complexity (GFLOPS)

Models evolved beyond SparseNN for better accuracy and user experience

Emergence of GenAI with LLMs and wide array of models across different use cases

Meta

GPU Deployment Challenges

Peak performance **not always** equal to effective performance

Large deployments can be **resource intensive**

Capacity constraints due to GenAI demand

Next Gen MTIA Development Goals

Improve perf/TCO and perf/W compared to previous generation

Handle models across multiple Meta services efficiently

Developer efficiency to quickly reach high volume deployments

Meta

Features

GEN-O-GEN PERFORMANCE

Increased GEMM TOPs by 3.5x to 177 TFLOPS @BF16

Sparse matrix support with 2x TFLOPS

ANS weight decompression with 50% compression ratio and 20% better memory to compute tensor transfer performance

Balance compute, memory and data transfer to achieve over 80% utilization

INTEGER DYNAMIC QUANTIZATION

Hardware based tensor quantization

Deliver accuracy comparable to FP32

PYTORCH EAGER MODE SUPPORT

New hardware job launch time <1us

Completed job replacement time <0.5us

TBE (TABLE BATCH EMBEDDING) OPTIMIZATION

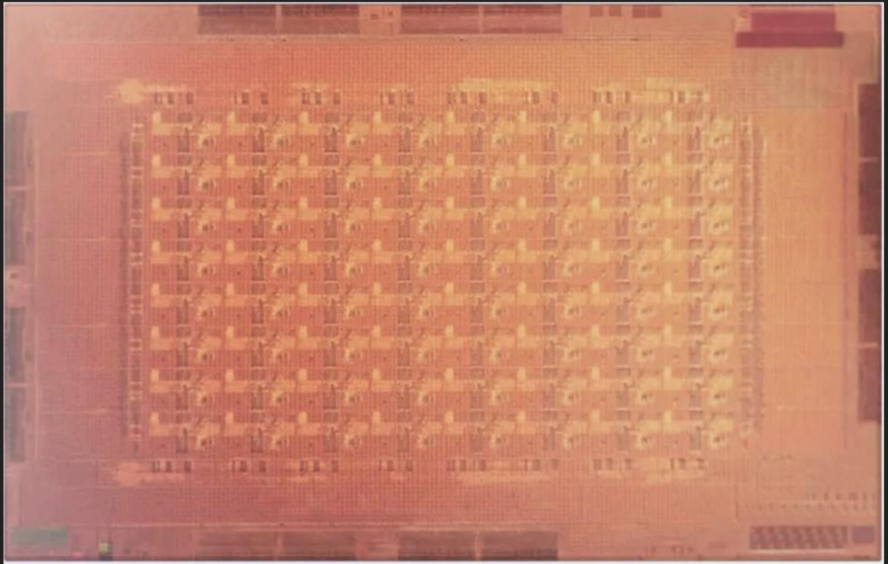
HW optimization for download and prefetch of embedding indices

2-3x faster run time compare to prev gen

Meta

Accelerator Architecture

Specification



TECHNOLOGY	TSMC 5nm
FREQUENCY	1.35 GHz
GATE COUNT	2.35B gates, 103M flops
DIMENSIONS	25.6 x 16.4 mm (421 mm2)
PACKAGE	50mm x 40mm
TDP	90 Watts
GEMM TOPS	354 (INT8), 177 (FP16) 2x with sparsity
MEMORY	128GB LPDDR5 6400 BW 204.8GB/s

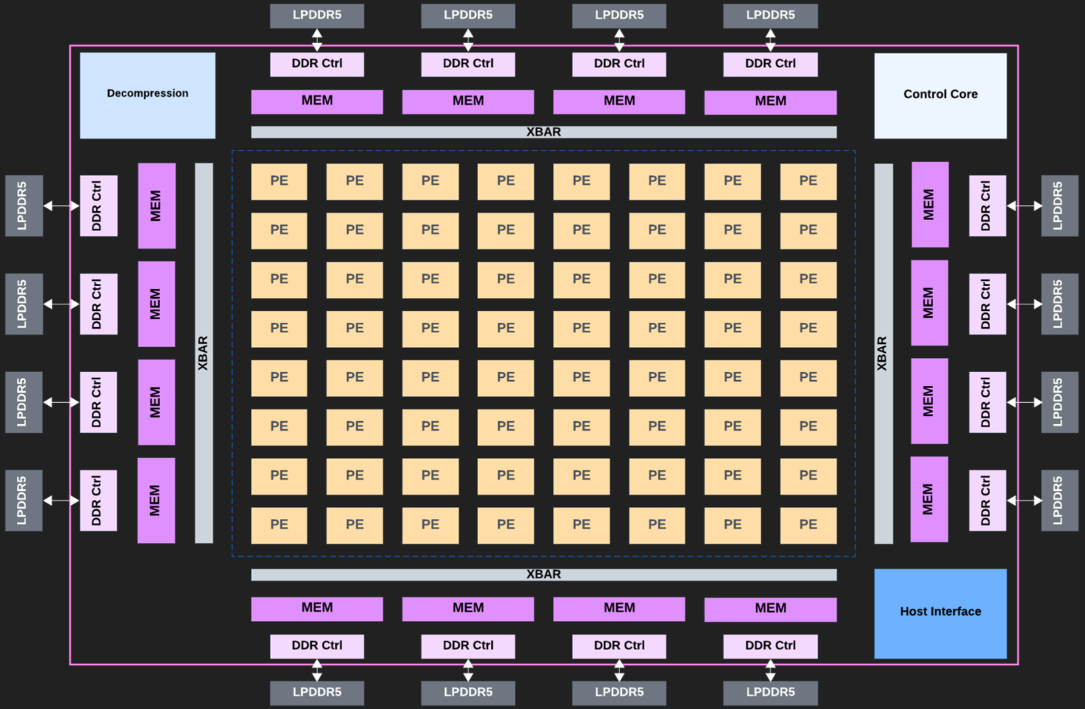
Architecture Overview

8x8 grid of processing elements connected via custom mesh network

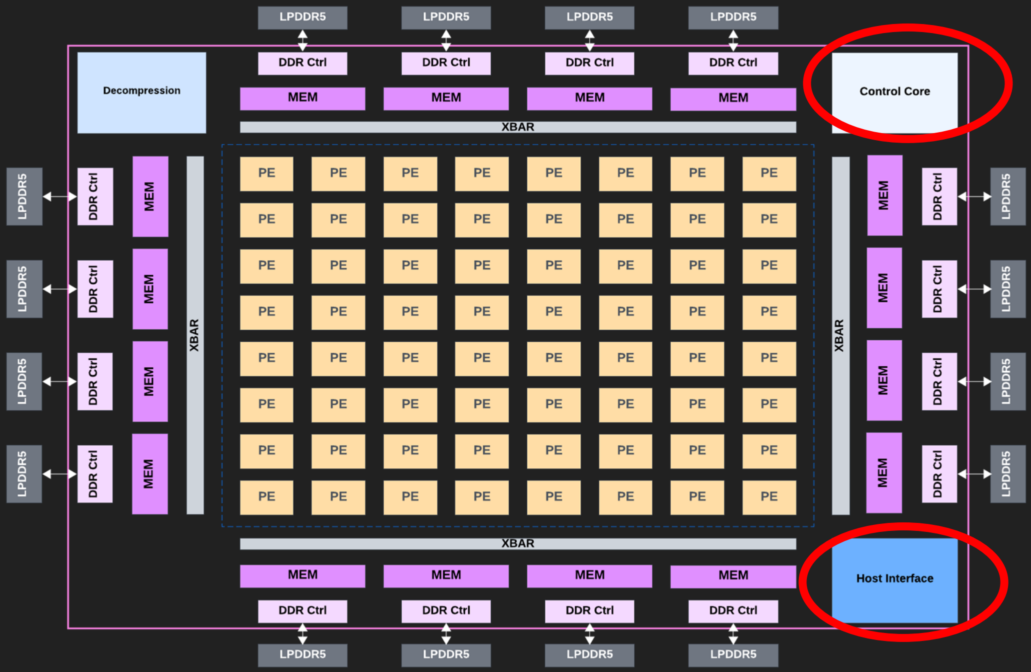
256MB of on-chip SRAM, distributed across 4 sides with 2.7 TB/s BW

16 channels of LPDDR5 memory on 4 sides, up to 128GB capacity with 204.8GB/s BW

Control subsystem & host interface



Host Interface & Control Core



Host Interface

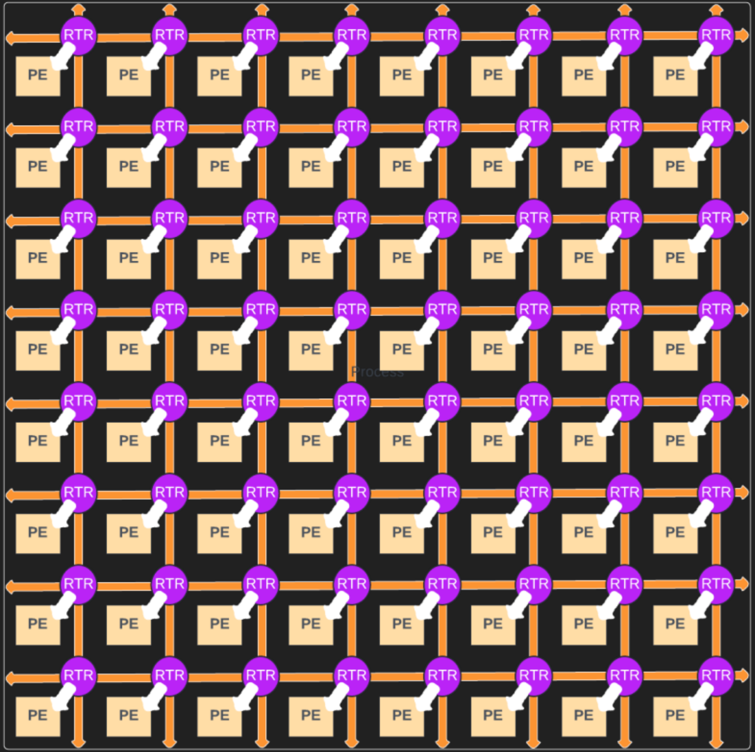
- Gen5 x8 - 32GB/s
- 4MB PCIe Descriptor SRAM for fast descriptor fetch

Control Core Subsystem

- Quad Core Scalar RISC-V
- 8MB L2 Cache.
- 4MB Context SRAM for fast workload distribution

Meta

Network on Chip (Noc)



Data NoC

- Increased PE to Memory subsystem BW by ~3.4x to 2.76 TB/s
- Non-blocking and QoS support
- Multicast reads from PE

Config NoC

- Non-blocking and QoS support
- Subsystem level broadcast
- Selective multicast for eager mode support

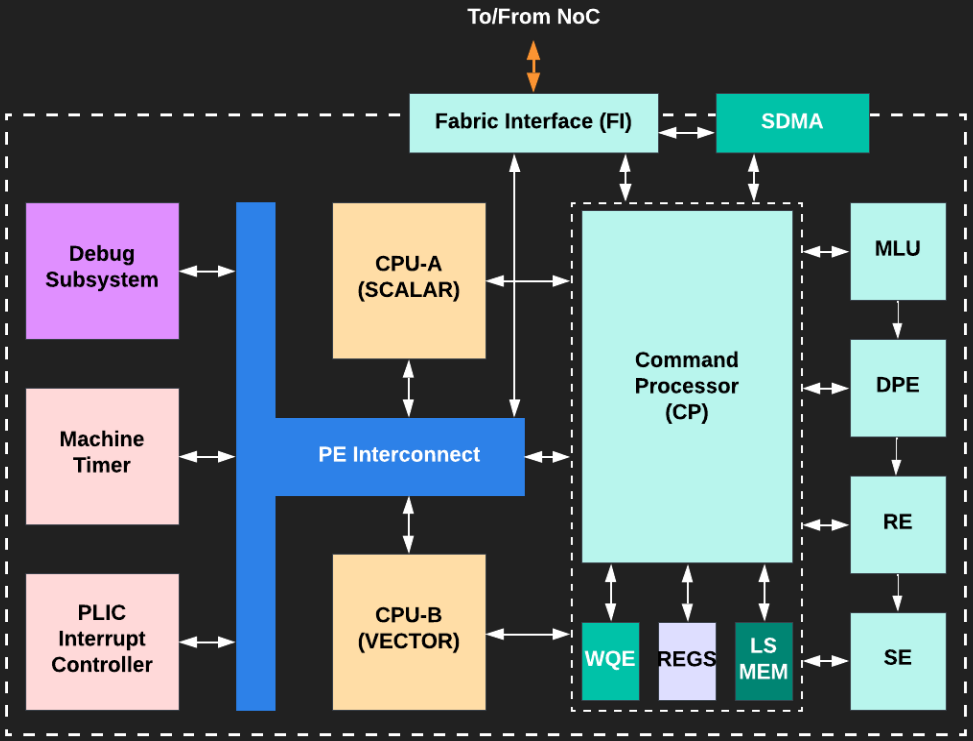
Processing Elements (PE) Overview

Dual RISC-V cores, one scalar, one with vector extension

Command Processor (CP) to coordinate execution of functional blocks in PE

Fixed-function units to accelerate:

- Matrix multiplications with sparsity support (DPE)
- Non-Linear functions (SE)
- Data movement (MLU)
- Dynamic quantization (RE)
- Weight decompression (SDMA)
- Eager mode (WQE)

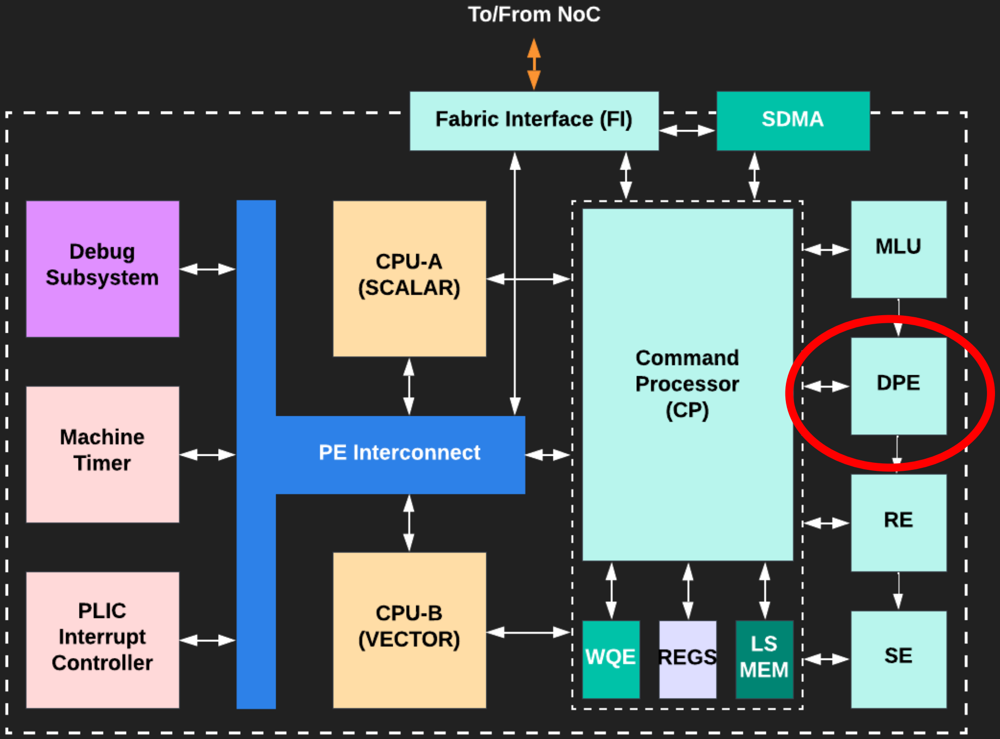


PE Compute

Dot Product Engine (DPE) delivers 2.77 TF/s (FP16) per PE

Added sparsity matrix support, providing 5.54 TF/s (FP16) in sparse mode

Widened data paths in MLU, RE & SE to match DPE performance



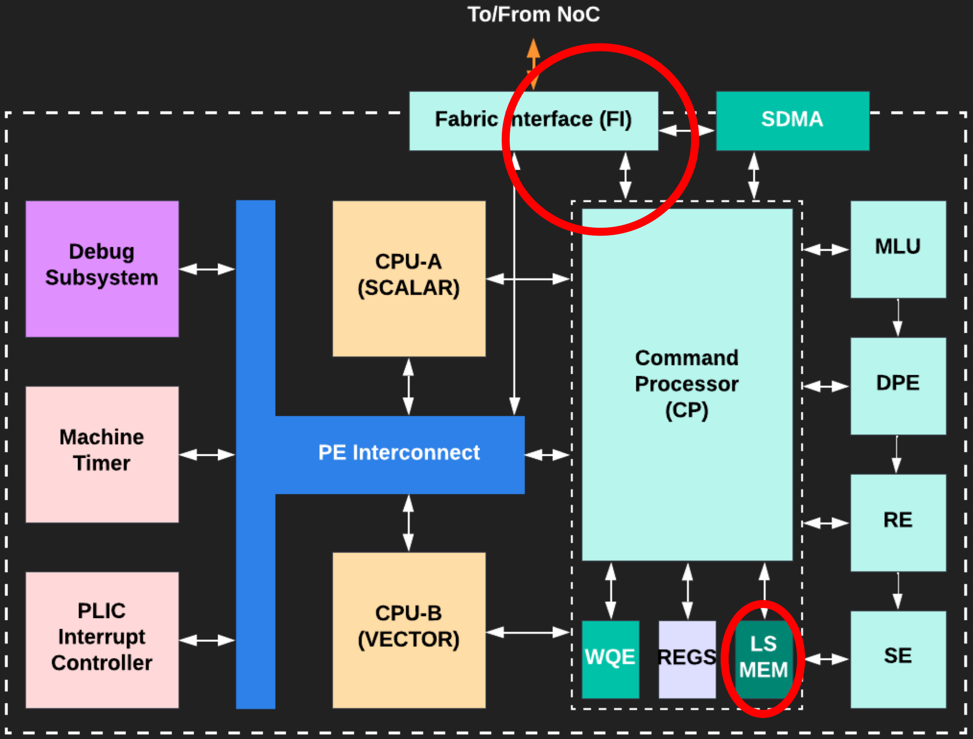
Meta

PE Memory Subsystem

384KB PE local memory to support larger and more complex workloads

Increased PE local SRAM and Fabric Interface BW to balance compute and memory access and provide over 80% PE utilization

Index aligned DMA support to speed up index prefetch



Meta

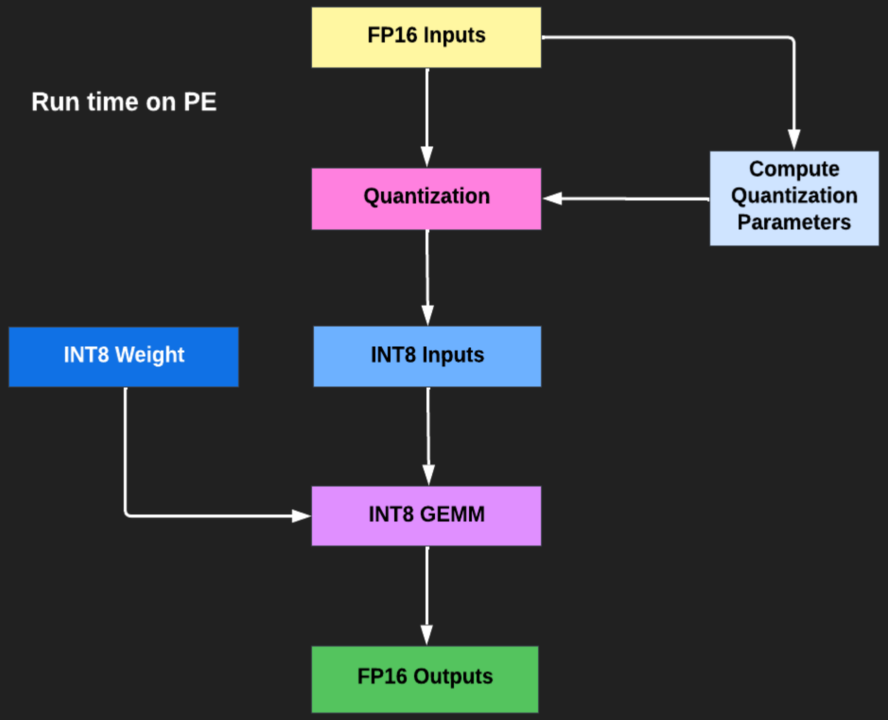
Integer Dynamic Quantization

Built-in hardware that provide capability to offload the task of adjusting quantization parameters in real time

- Collect min/max per batch during run time
- Support rowwise quantization

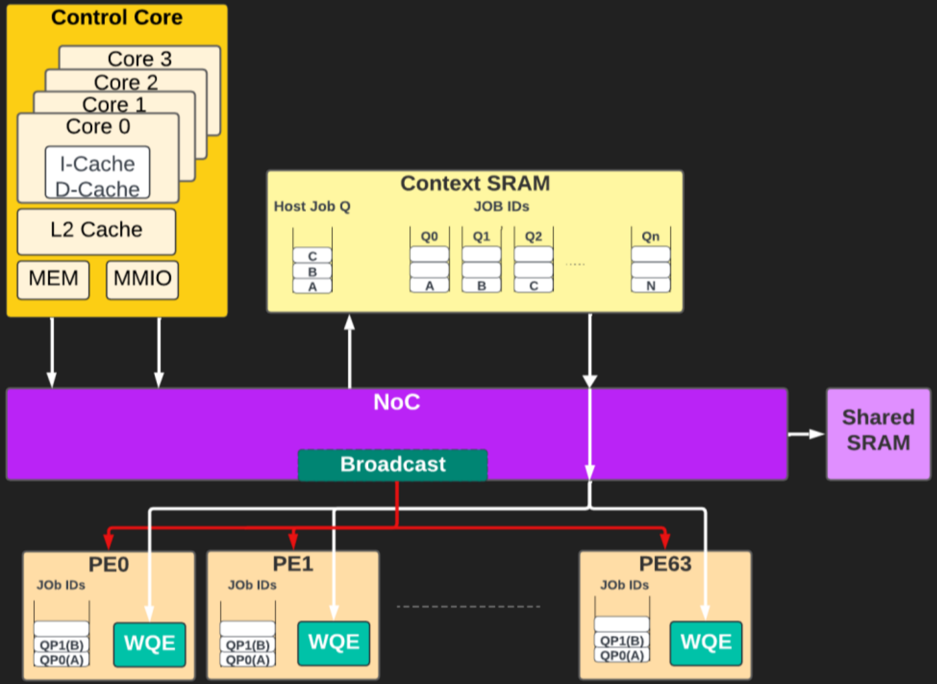
Enable channel-wise symmetric dynamic quantization for FC operators

Achieved over 99.95% accuracy comparing to baseline FP32 result



Meta

Eager Mode Enhancements

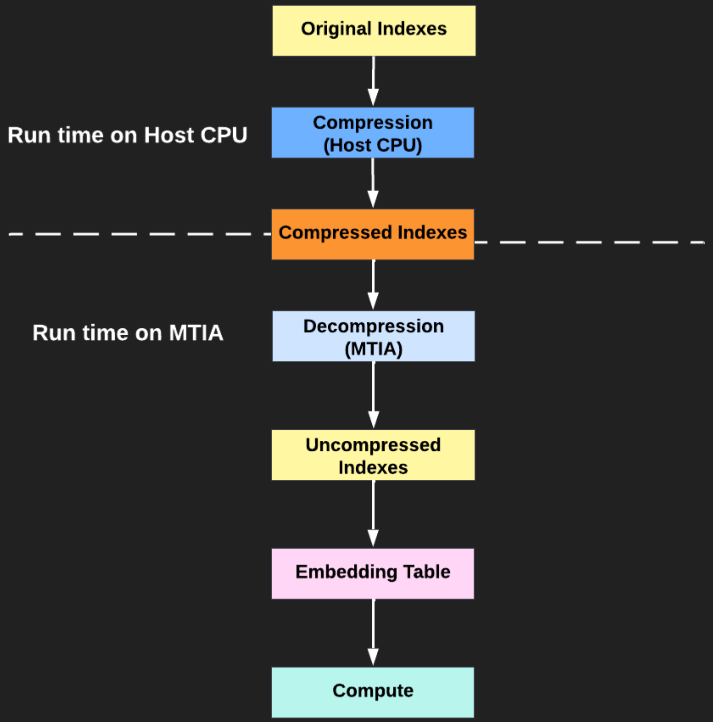


Added multicast write groups to allow Control Core to broadcast Eager Mode Work Queue (WQ) descriptors to select PEs

Added Work Queue Engine (WQE) in PE to DMA WQ requests descriptors from Control Core

More than 80% reduction in PE job launch time

Hardware Decompression



Transfer time of large embedding indices over PCIe impact SLS performance

Added Decompression Engine to alleviate PCIe and network congestion

Support for RFC1952 (GUNZIP/GZIP) standard encapsulating RFC1951 (Deflate Compression Format)

- Support for static and dynamic Huffman coded blocks

4 Decompression Cores

- Decompression rate up to 25 GB/s

PE Weight Decompression

Lossless Asymmetric Numerical System (ANS) algorithm

Data processed on a 32x32B granularity

32B / cycle decompression rate

Achieved close to 50% compression ratio

- improve the on-chip memory footprint
- reduce the PE to NoC read BW

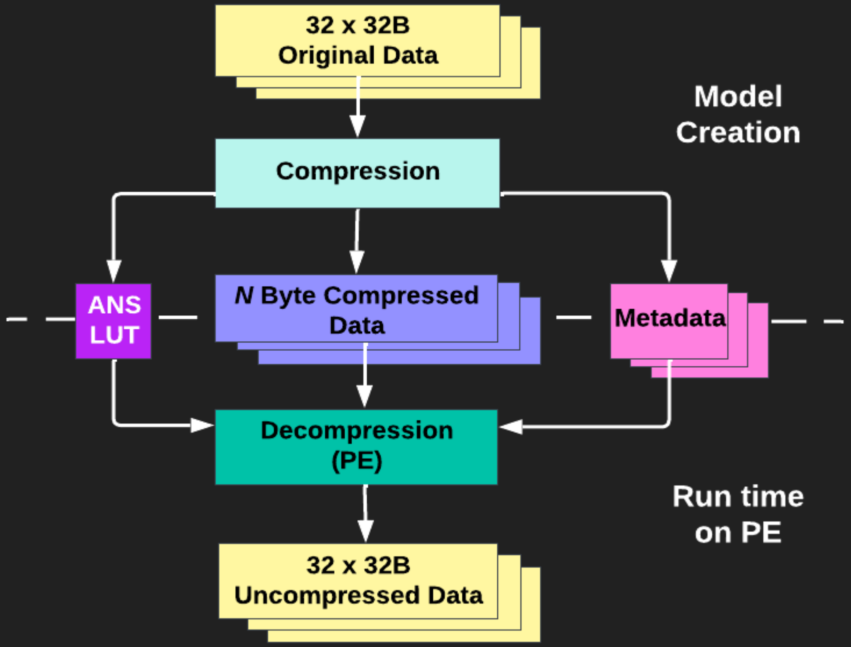


Table Batch Embedding (TBE)

TBE

Combines tables from separate embedding batch ops into one single table

HARDWARE FEATURES

Index aligned DMA for faster index prefetch

Multiple context unrolling with SW prefetch for indices

HW decompression engine to speed up large embedding indices from host via PCIe

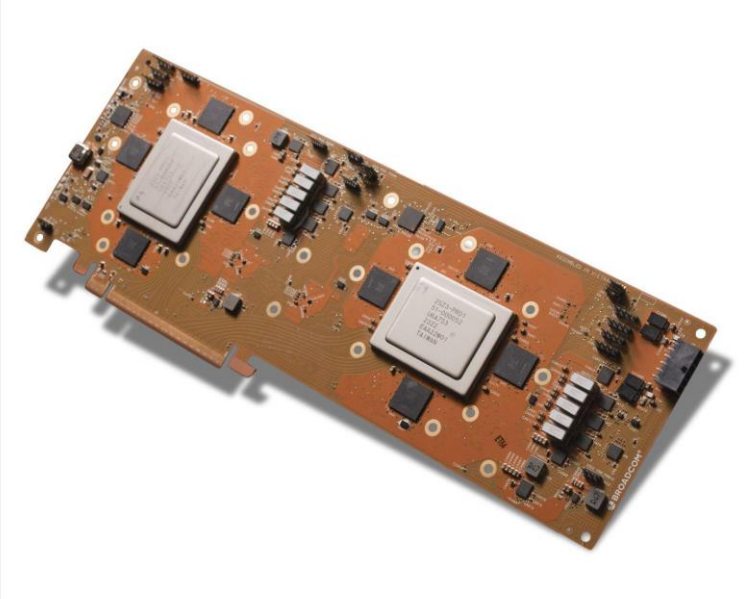
PERFORMANCE

Improve runtime by 2-3x over previous generation

Meta

System Design

Accelerator Module



PCIe CEM FHFL Form Factor

- 2 MTIAs per Module

Board TDP of 220W

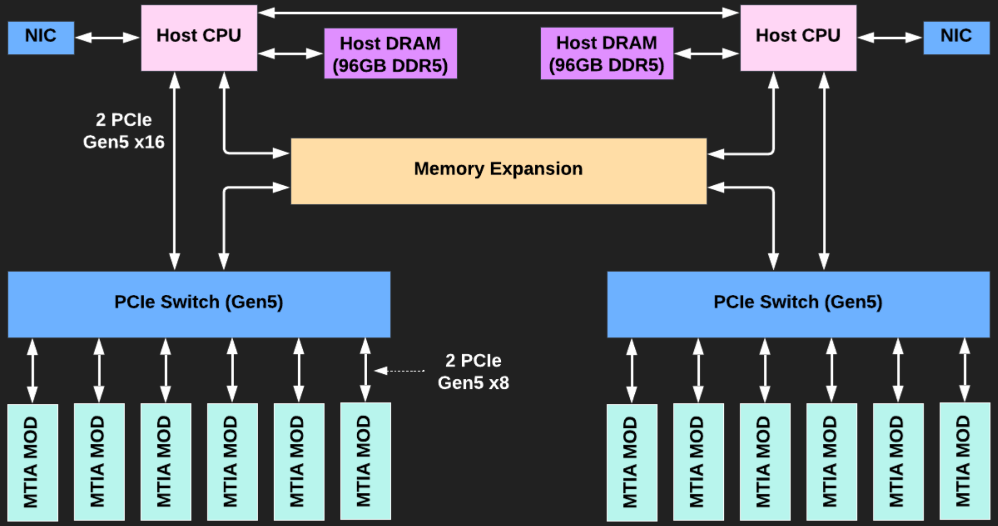
64GB/s Gen5 PCIe Interface

- 2 Gen5 x8

Up to 256GB LPDDR5 Memory

- 409.6 GB/s total memory BW

System Topology



12 modules per chassis

3 chassis per rack

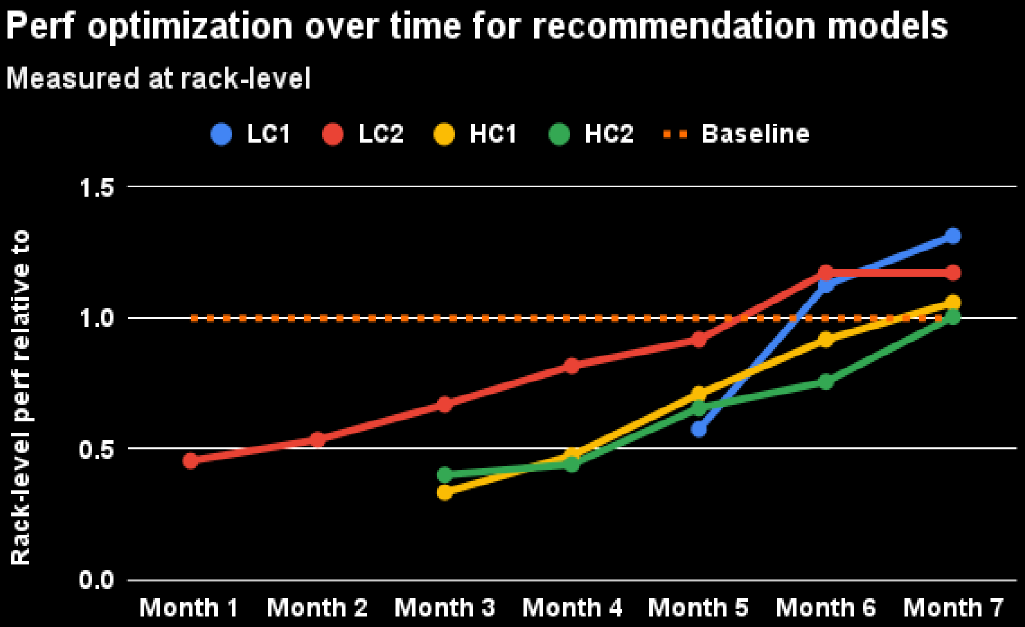
72 MTIA ASICs per rack

Deployed in DC since H1’ 24

Meta

Performance

Model Optimization



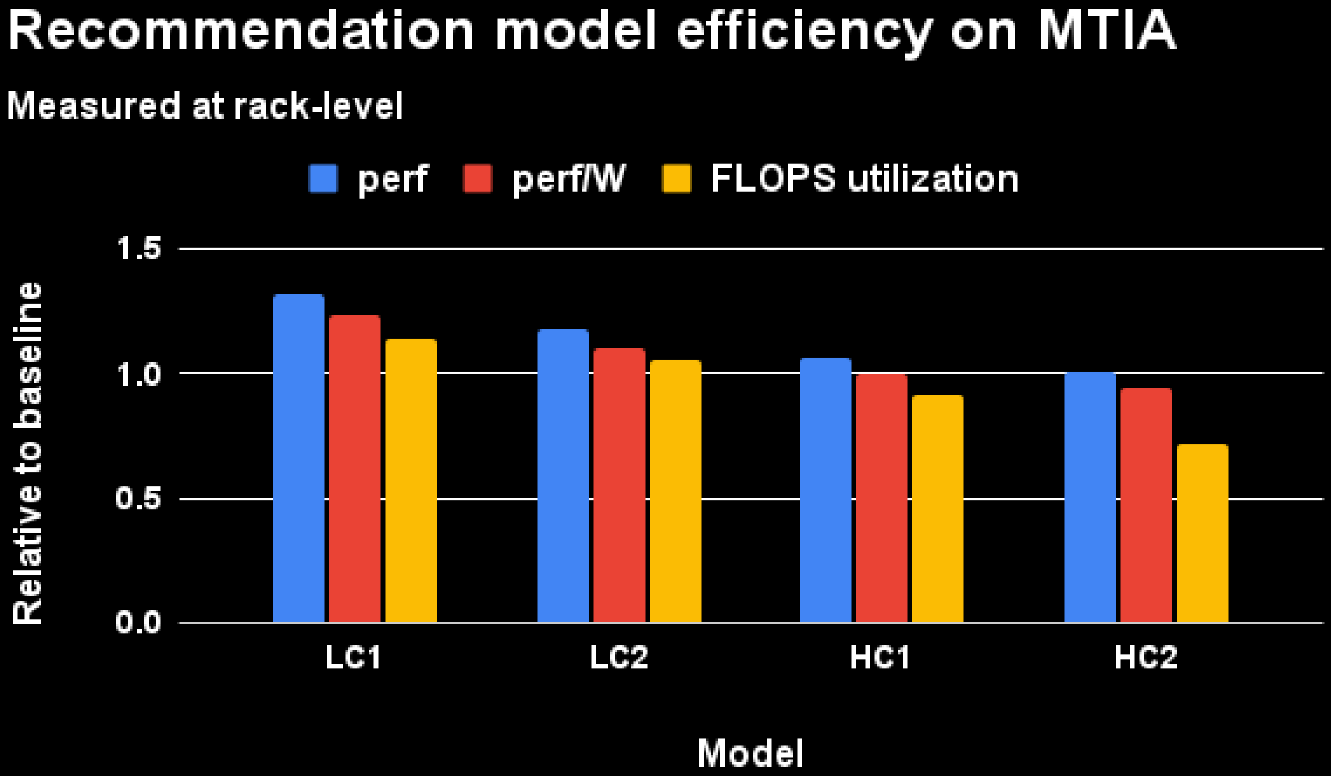
Continuous improvement in model performance

Low complexity (LC) models have an out-of-the-box advantage with the large on-chip SRAM

High complexity (HC) models require more optimization to effectively block data in SRAM and realize higher effective FLOPS

- More than 2x performance improvement over 4-6 months

Model Performance



Recommendation model efficiency on MTIA

Measured at rack-level

Q & A

Meta