August 2024

Intel Gaudi 3 AI Accelerator: Architected for Gen AI Training and Inference

Roman Kaplan, Ph.D.
Principal AI Performance Architect

Intel Corporation
August 2024

Gaudi Product Generations

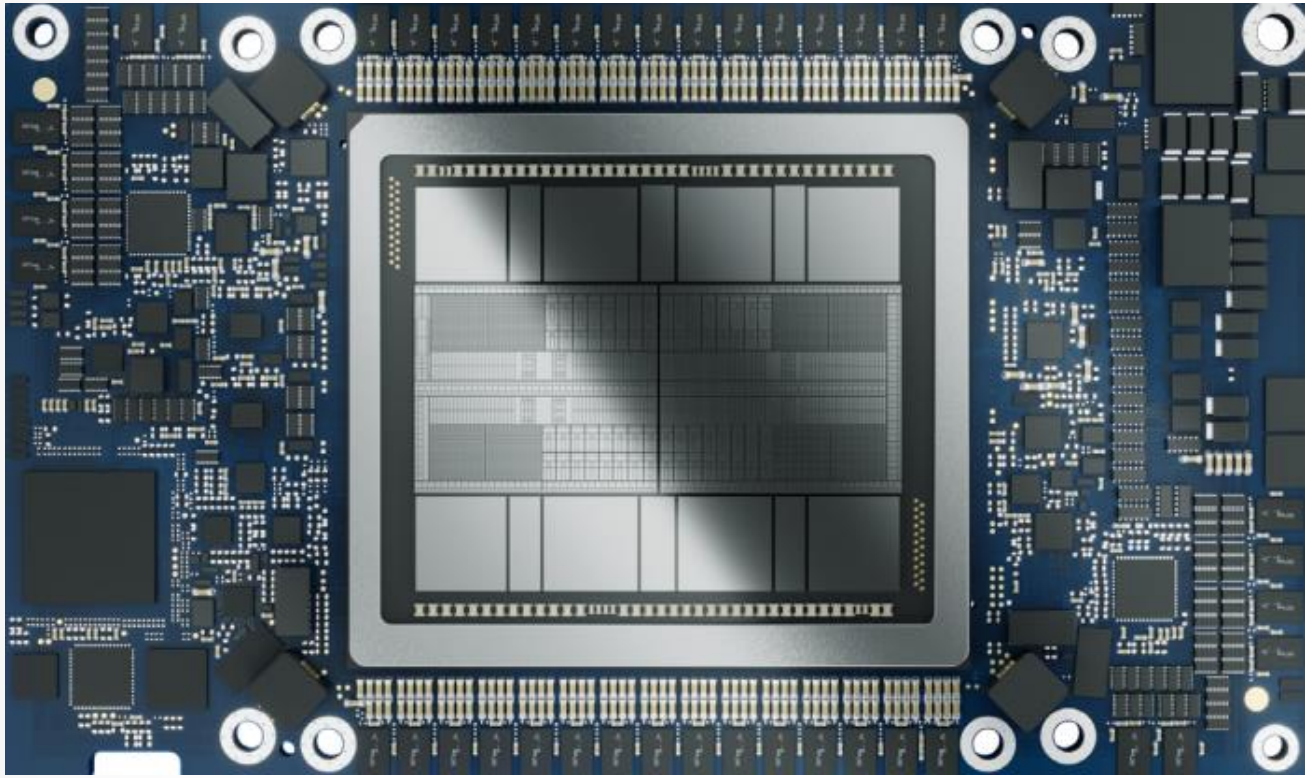
Intel® Gaudi® accelerator	Intel® Gaudi® 2 accelerator	Intel® Gaudi® 3 accelerator
		
(16nm)	(7nm)	(5nm)

Product Parameter	Gaudi	Gaudi 2	Gaudi 3
TDP (OAM)	400 W	600W	900W (Air) / 1200W (Liquid)
Peak Compute (BF16)	60 TFLOPs	432 TFLOPs	1835 TFLOPs
HBM Capacity	32 GB	96 GB	128 GB
Peak HBM BW	900 GB/s	2.46 TB/s	3.67 TB/s
Peak PCIe BW (bi-directional)	64 GB/s	64 GB/s	128 GB/s
Embedded NIC BW (bi-directional)	2 Tb/s	4.8 Tb/s	9.6 Tb/s

Intel® Gaudi® 3 AI Accelerator OAM

- OAM: Open Compute Platform Acceleration Module
- 2 compute dies connected over an interposer bridge
- 8 HBM2e stacks
- Up to 900W with air cooling
- Up to 1200W with liquid cooling
- PCIe Gen5 x16
- 24x 200GbE RoCE via 48 112G PAM4 Serdes

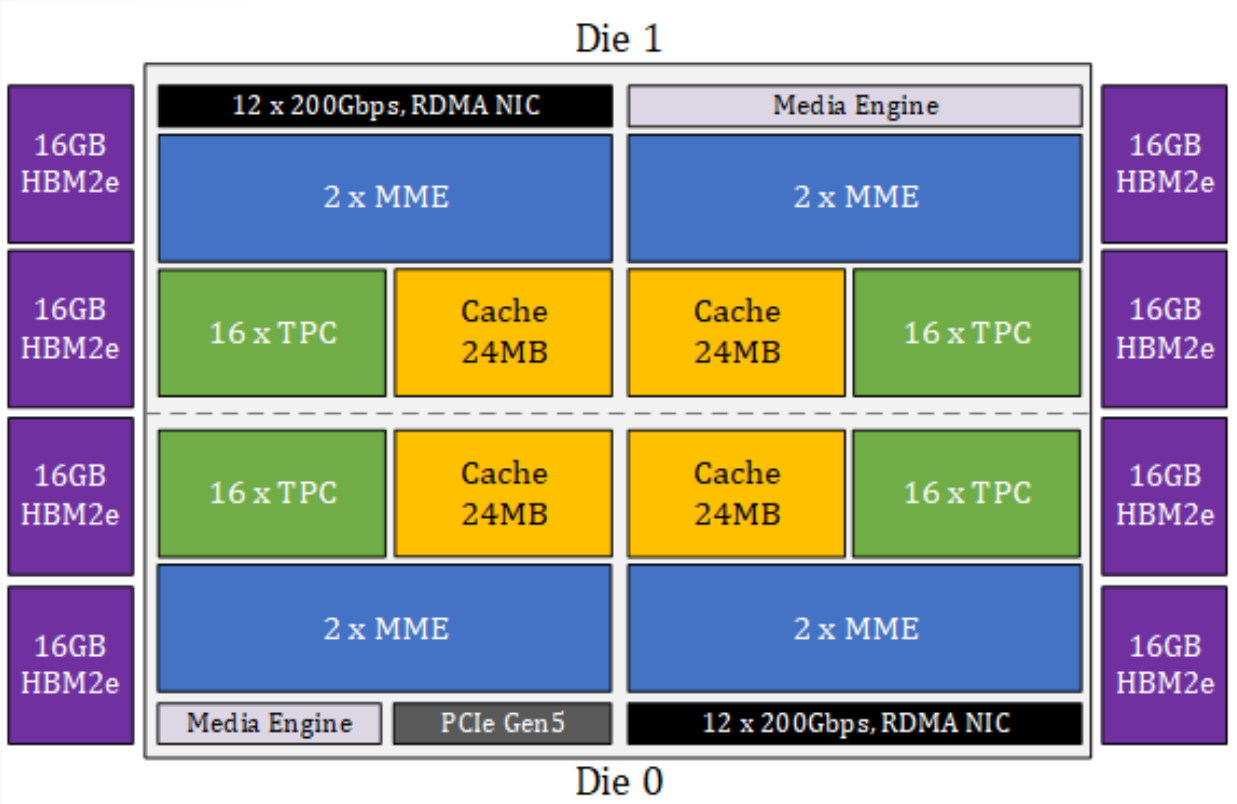
Intel Gaudi 3 AI Accelerator OAM



Intel® Gaudi® 3 Spec and Block Diagram

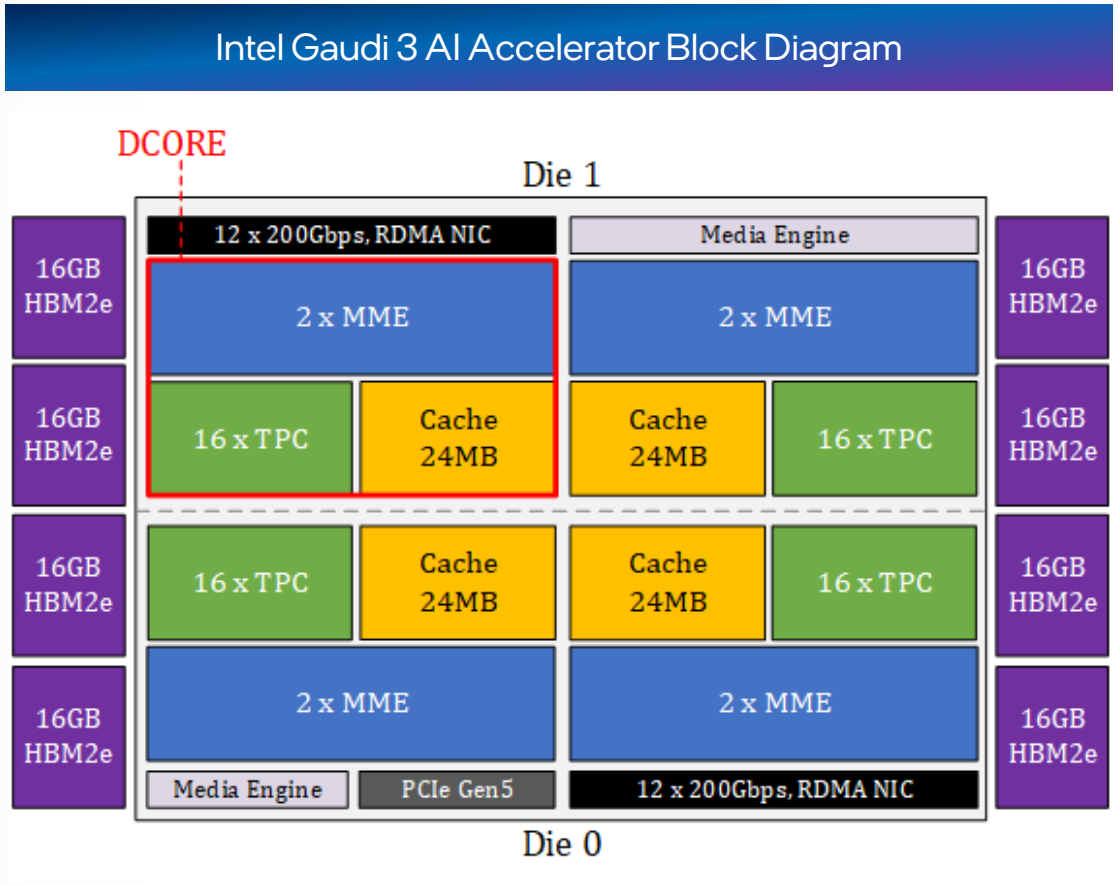
Feature/Product	Intel® Gaudi® 3 Accelerator
BF16 Matrix TFLOPs	1835
FP8 Matrix TFLOPs	1835
BF16 Vector TFLOPs	28.7
MME Units	8
TPC Units	64
HBM Capacity	128 GB
HBM Bandwidth	3.67 TB/s
On-die SRAM Capacity	96 MB
On-die SRAM Bandwidth (L2 Cache)	12.8 TB/s
Networking	1200 GB/s bidirectional
Host Interface	PCIe Gen5 x16
Host Interface Peak BW	128 GB/s bidirectional
Media Engine	Rotator + 14 Decoders (HEVC, H.264, JPEG, VP9)

Intel Gaudi 3 AI Accelerator Block Diagram



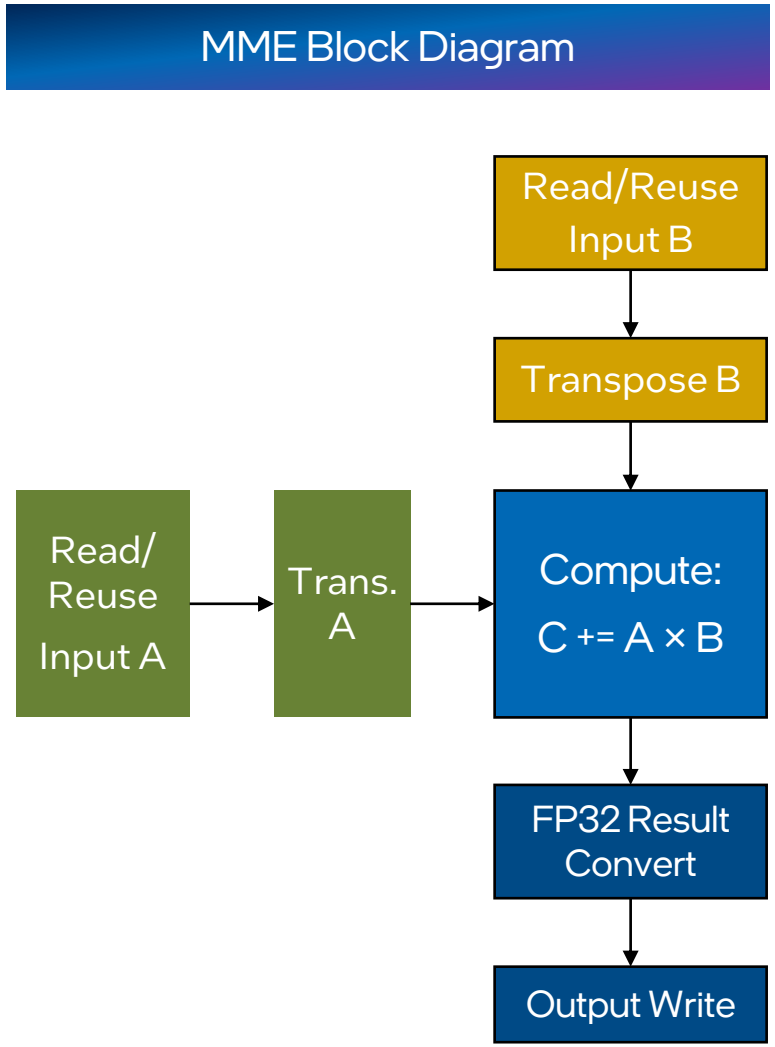
Architecture in Depth

- Uniform memory mapping of HBM by MMU
- Compute is clustered:
  - 4 Deep Learning Cores (DCORE)
  - Each DCORE: 2xMME, 16xTPC, 24MB cache
- L2 and L3 data caches:
  - L2: Allocated only in DCORE cache
  - L3: Uniformly distributed across all DCORE caches
- Media accelerators:
  - Decoder and Rotator
- NW Sub-system containing:
  - 24 RDMA NIC 200GbE ports
  - (details in a separate slide)
- Control has a separate block and NOC fabric



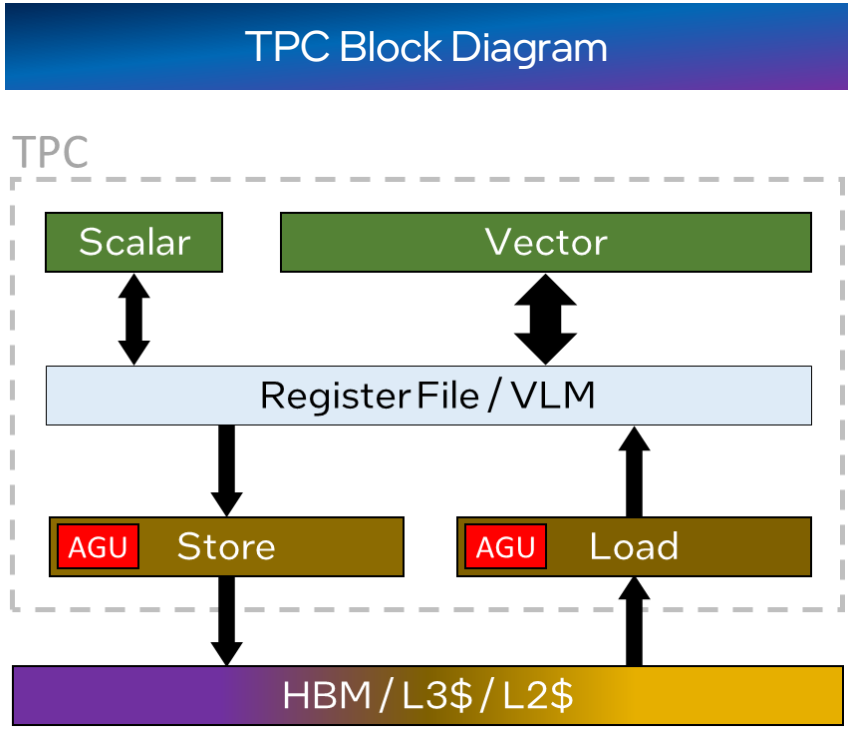
Matrix Multiplication Engine (MME)

- Executes all matrix multiplication operations
- Configurable, not programmable
- Each MME is a large output stationary systolic array:
  - 256x256 MAC structure w/ FP32 accumulators
  - 64k MACs/cycle for BF16 and FP8
- Internal pipeline to maximize compute throughput:
  - Input read, compute & output write all execute in parallel
  - Integrated transpose engines for zero-overhead input transpose
  - Accumulated result is converted to any precision before write out
- Internal buffers for input reuse – replacing L1$
- Integrated Address Generation Unit (AGU):
  - Address calculation within 5-D data tensor
  - OOB read padding and write prevention



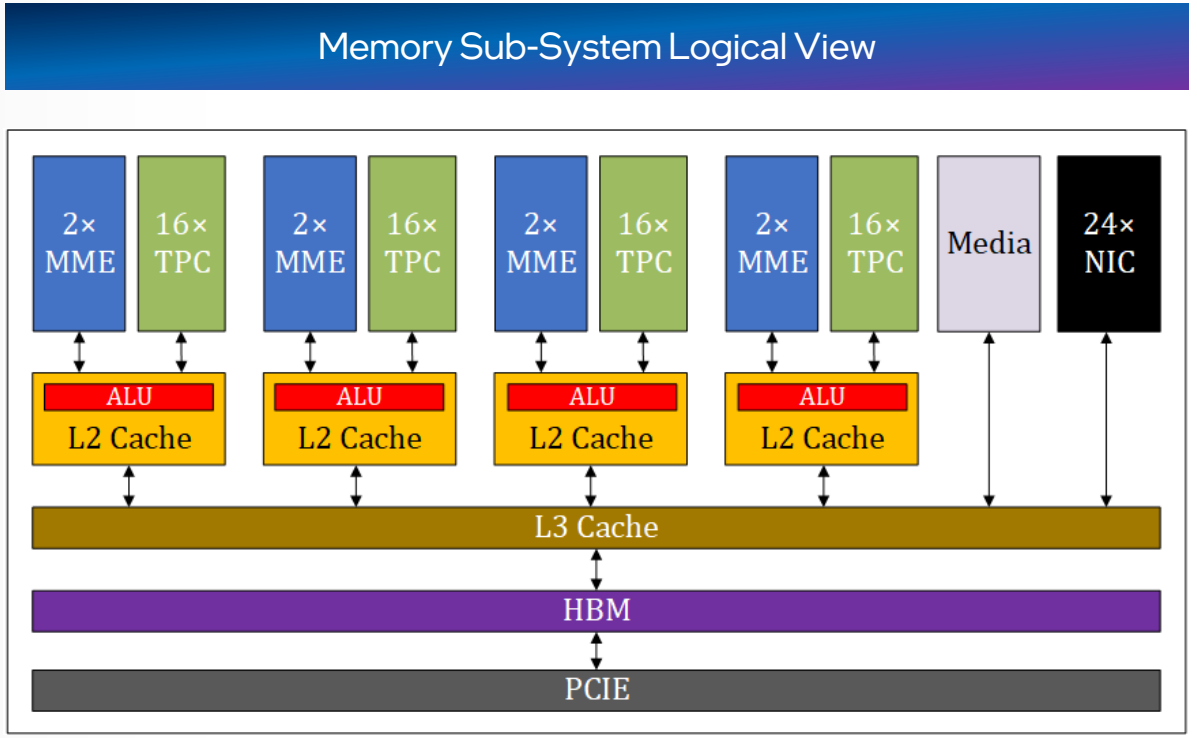
Tensor Processor Core (TPC)

- Total of 64 TPC units across the chip
- Executes non-Matmul operations
- Programmable: C enhanced with TPC intrinsics
- VLIW with 4 separate pipeline slots:
  - Vector: 256B-wide SIMD
  - Scalar
  - Load: Vector or scalar
  - Store: Vector or scalar
- Load and Store slots integrate Address Generation Unit
  - Calculates the memory address within a 5-D data tensor
- Supports main 1/2/4-Byte datatypes: FP and integer
- 12KB vector register file
- 80KB of vector local memory (VLM)
- Latency hiding mechanism



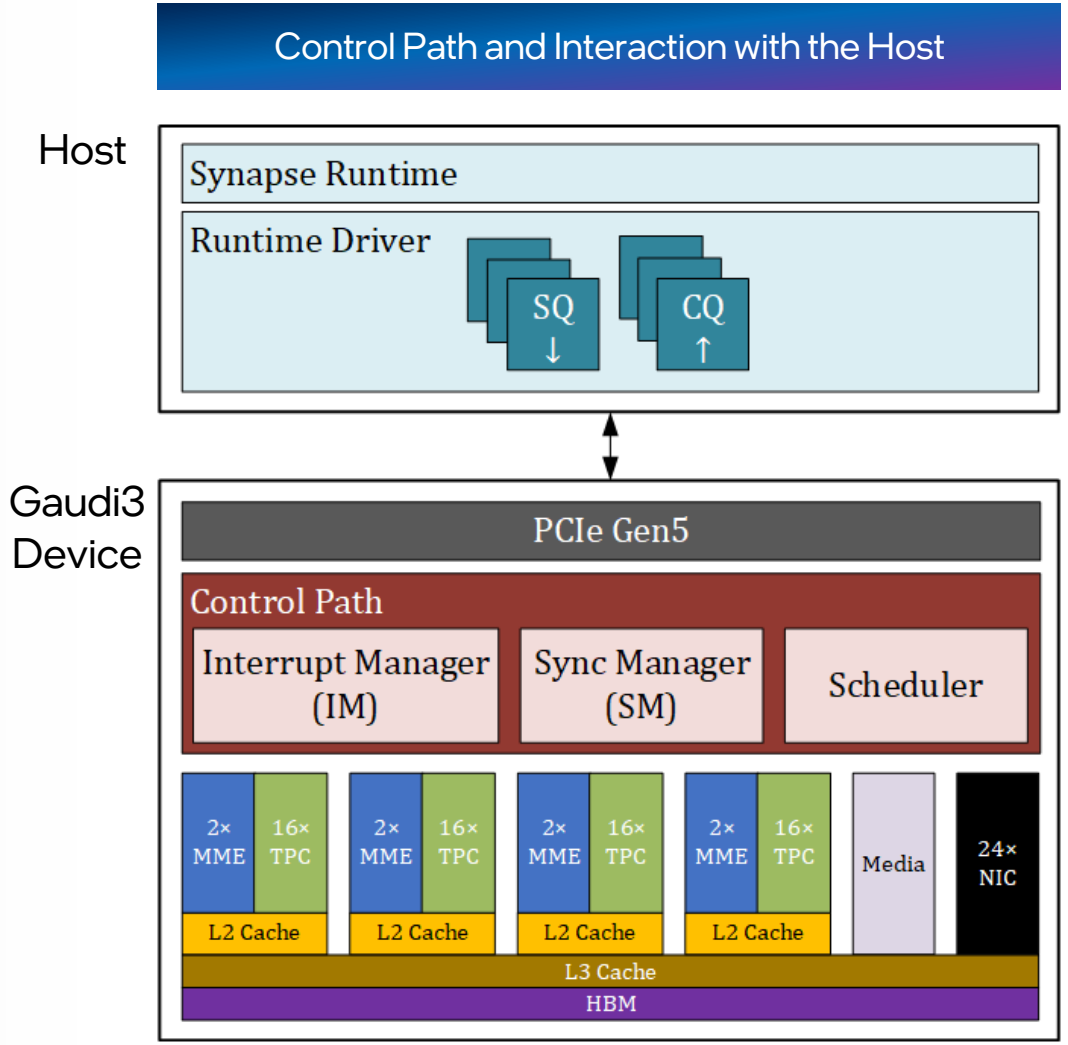
Memory Sub-System

- Unified Memory space of L2 / L3/ HBM
- Near Memory Compute:
  - Add / Sub
  - Max / Min
- Usage of Memory Context ID (MCID) to tag cache lines with shared algorithmic usage
- Cache Directives:
  - No-$, L2$, L3$, L2$+L3$
  - Discard: Invalidate all same-MCID CLs
  - Degrade: Reset same-MCID CLs hit count



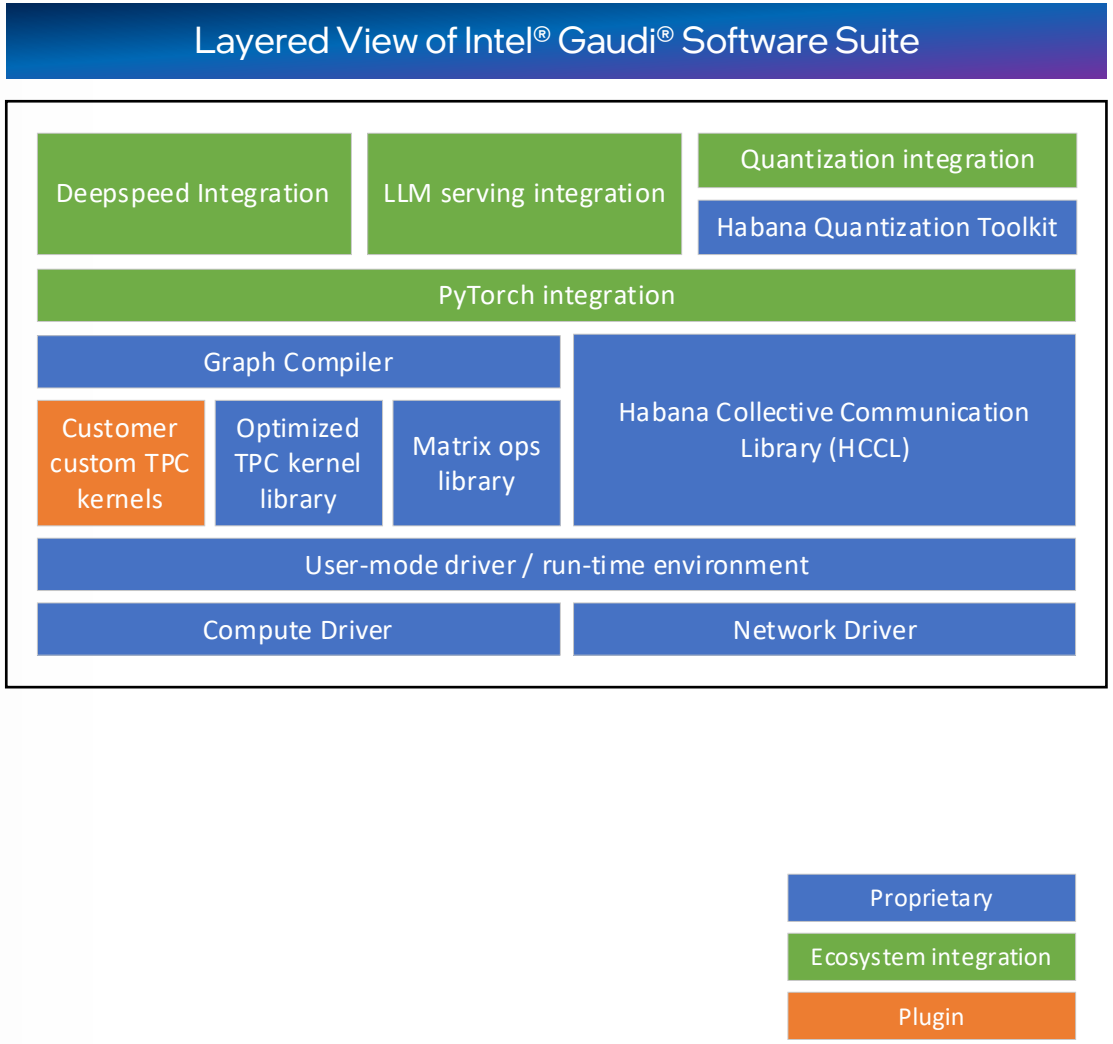
Control Path and Runtime Driver

- Separate NOC fabric for control path messages
- Sync Manager handles on-die control logic:
  - Dispatches work to the designated units
  - Waits on counted events, and triggers start-of-job to the units
- Runtime Driver :
  - Sets engine work dependencies by configuring the Sync Manager
  - Submits jobs through Submission Queues
  - Accepts completion events through Completion Queues
- Interrupt Manager enables passing events asynchronously to the driver



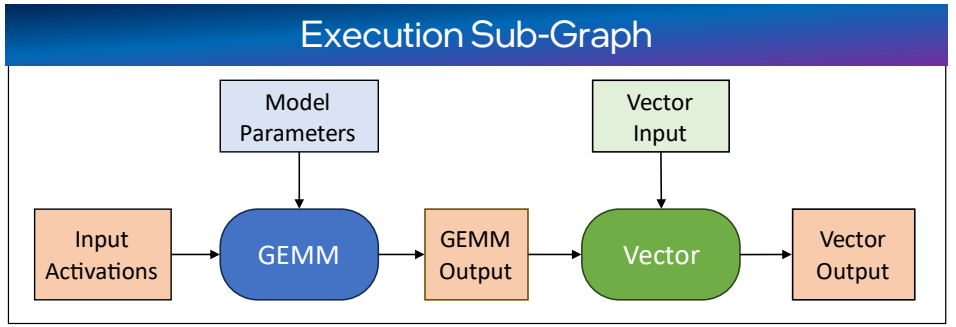
Intel® Gaudi® Software Suite

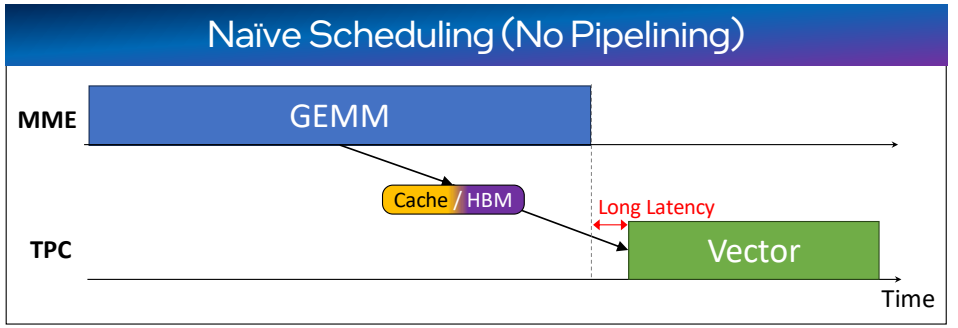
- Integrates the main frameworks used today
- Supports FP16/BF16 → FP8 quantization
- Main proprietary SW layers:
  - Graph Compiler: Determines all engine dependency and scheduling logic
  - Matrix operations: Configuring the MME
  - TPC kernels: All non-Matrix operations
  - Collective Communication Library (HCCL)
- Several sources for TPC Kernels:
  - Habana's optimized TPC kernel library
  - Custom user kernels
  - MLIR-based fused kernels: generated during graph compilation

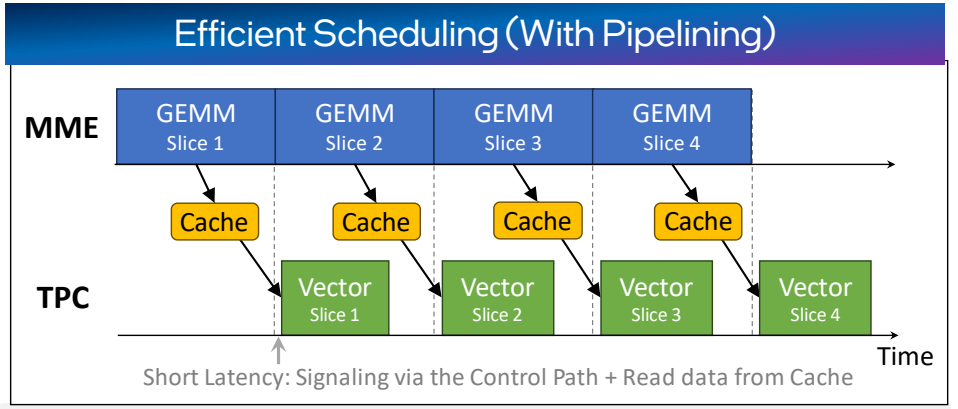


MME-TPC Parallelism via Pipelining

- Graph Compiler orchestrates MME & TPC parallelism
  - Long chunks of work are split to smaller independent slices
  - Pipeline using producer→consumer relation
- Slice size is determined to balance between the following:
  - High compute utilization
  - Fit within the cache capacity
  - Maximize engine parallelism
- NOC fabric was designed to support the parallel work of MME and TPC



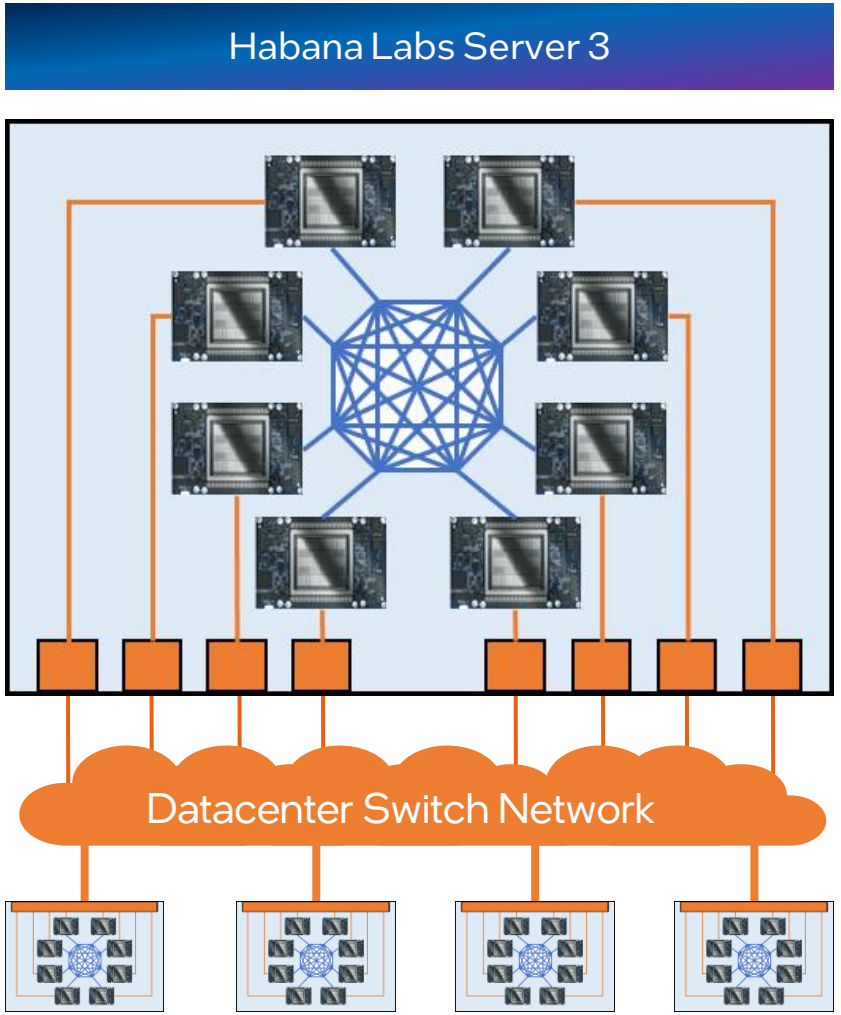




Networking and Habana Labs Server 3 (HLS3)

- Network ports exposed as NICs to driver
- NICs are activated via RDMA verbs over Device Virtual Space
- Collective operations execute w/ low control overhead

- **Habana Labs Server (HLS) Node**
  - 2x Intel Xeon Host CPUs
  - 8x Gaudi 3 OAM Cards
  - Peer-to-peer (P2P) connection between each pair of Gaudi3 cards
  - Gaudi NICs are used both for scale-up and for scale-out
  - No need for a network switch inside the node to support scale-out
  - Scale-up BW: Total of 67.2Tb/s bi-directional
  - Scale-out BW: Total of 9.6Tb/s bi-directional



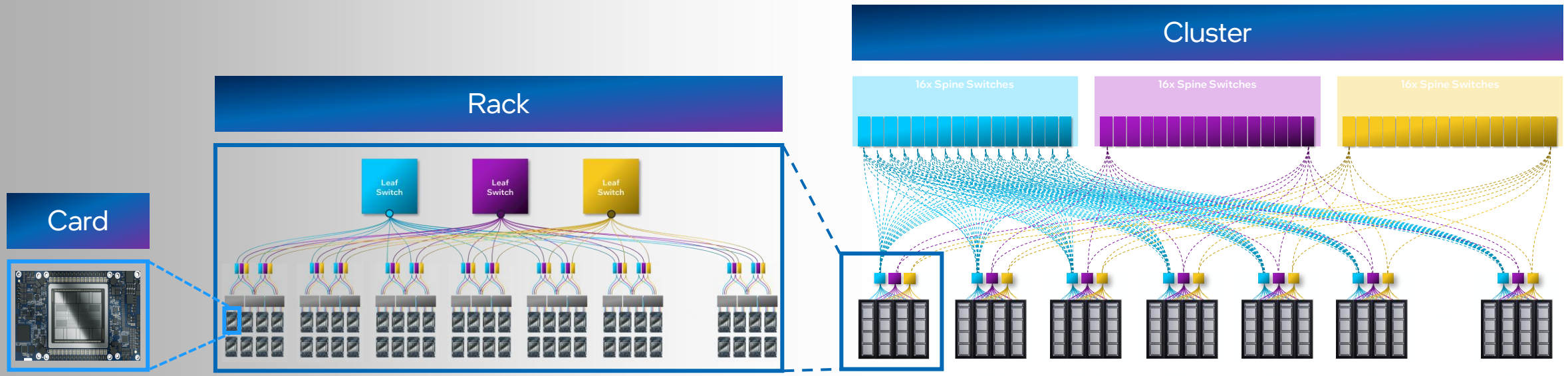
Performance: GenAI Inference Benchmarks

- Initial measured benchmarks:
  - FP8 precision
  - Intermediate results using development version before 1.18 release
  - 1.18 performance results will be published during September
- Gen-over-gen improvement of up to 2.8x
- All main LLM models and GenAI models are functional on Gaudi3
- **Software optimization continues on LLMs for future software releases – working on wide range of use-cases**

Model & Execution Parameters				Gaudi2 Measured		**Gaudi3 Measured**	
Model	TP (# Devices)	Input Length	Output Length	Batch Size	Throughput (tokens/sec)	Batch Size	Throughput (tokens/sec)
LLAMA2-7B	1	128	2048	163	4,789	217	6,574
	1	2048	2048	81	1,969	81	2,427
LLAMA3-8B	1	128	2048	289	11,098	768	18,769
	1	2048	2048	155	5,380	364	6,922
LLAMA2-70B	1	128	2048	88	1,126	164	3,218
	1	2048	2048	45	499	60	1,160
LLAMA2-70B	2	128	2048	327	3,212	512	6,225
	2	2048	2048	78	1,394	240	2,550

Gaudi 3: AI Acceleration at any Scale

- Close collaboration between SW and HW teams to reach high performance
- SW stack is continuously improving performance and expanding functionality
- Gaudi3 provides AI acceleration at any scale: Card → Node → Rack → Cluster
- Multiple HLS3 clusters are being designed and built
- More details in the Whitepaper ("Gaudi3 Whitepaper" on Google):
  https://www.intel.com/content/www/us/en/content-details/817486/intel-gaudi-3-ai-accelerator-white-paper.html



Thank You


intel