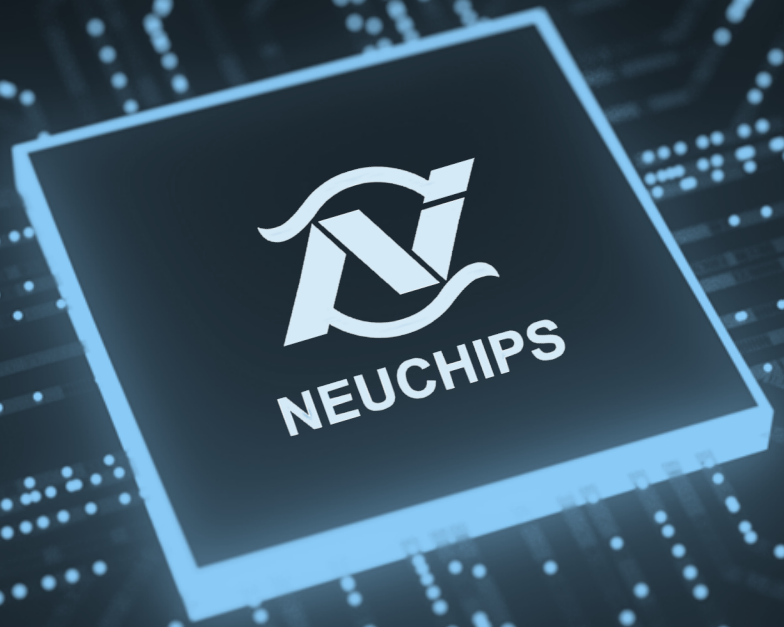
Design of a High Efficiency Accelerator for Full Scale Deep Learning Recommendation Models (DLRM) in the Datacenter



Alan Pita, Software Architect, NEUCHIPS Inc.

Kinny Chen, BD Manager ,NEUCHIPS Inc.

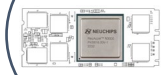
Agenda

- DLRM Challenges in Datacenter
- Accelerator Design Challenges & Solutions
- Results
- Summary

Introducing NEUCHIPS

2019 — Formed in 2019 in Hsinchu, Taiwan by an experienced team from Mediatek, Novatek, Realtek, GUC & TSMC

First product is 7nm inference accelerator tuned for DLRM, on track to perform 20M inferences per second per 20 watt with very high accuracy



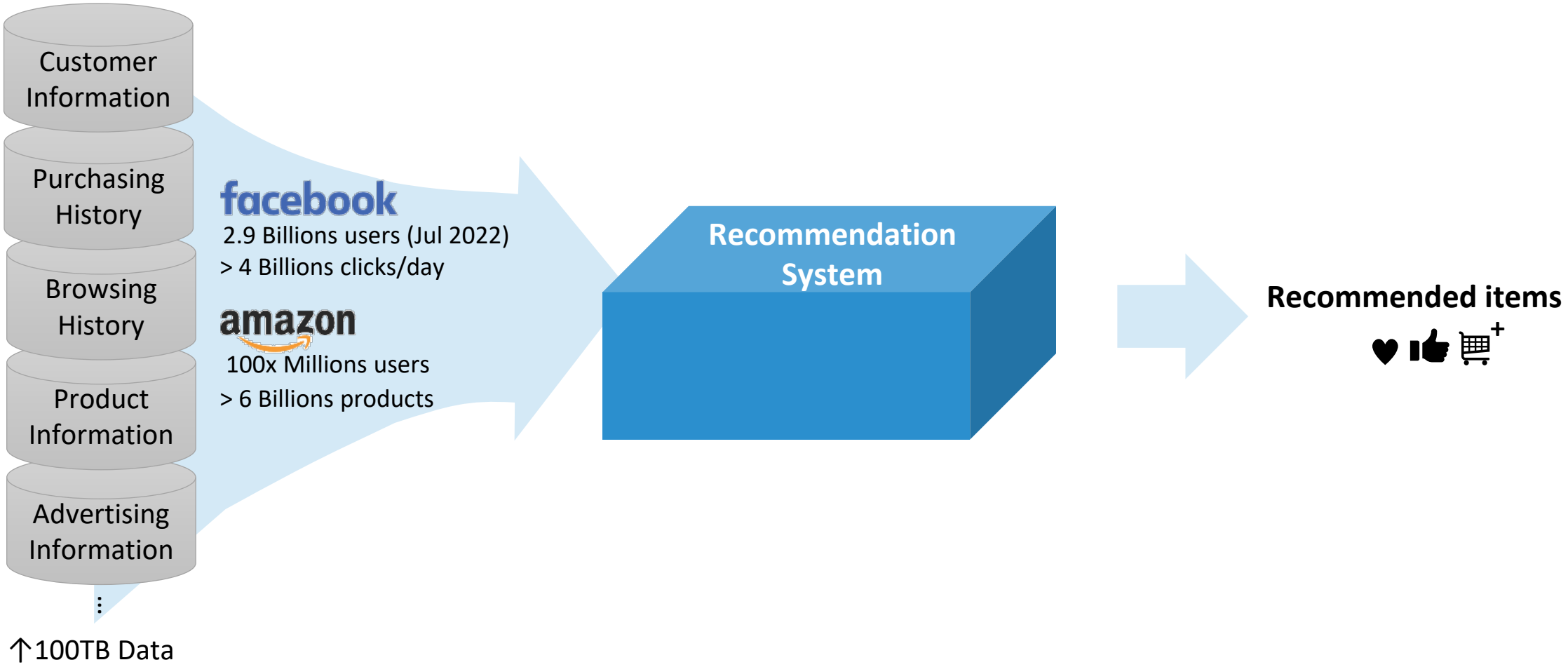
Partnered with Taiwan leading server OEMs to deliver samples of DM.2 card and PCIe card in Q4'22

Strategic partners include GUC, Wistron, Gigabyte

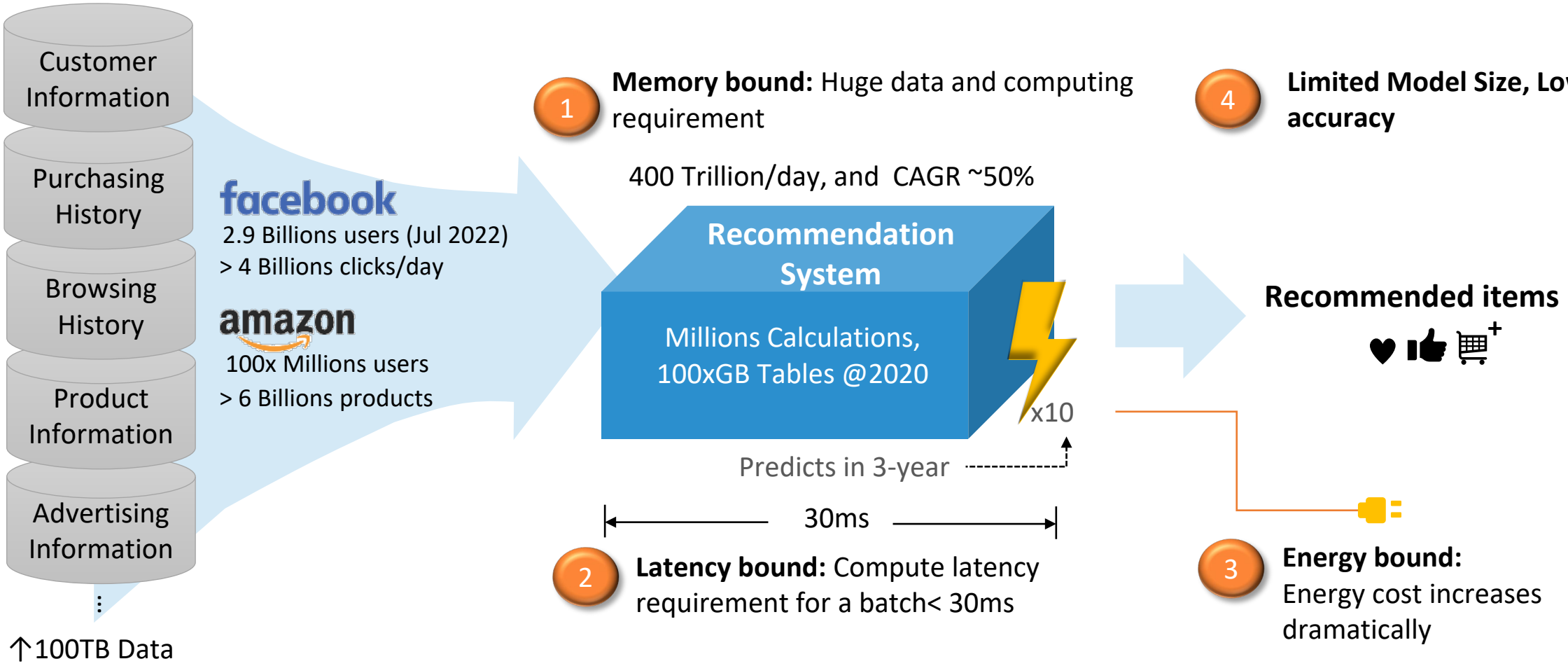
Investors include wistron, Powerchip, SUNPLUS, JAFCO ASIA

Recommendation Inference Challenges in Datacenter

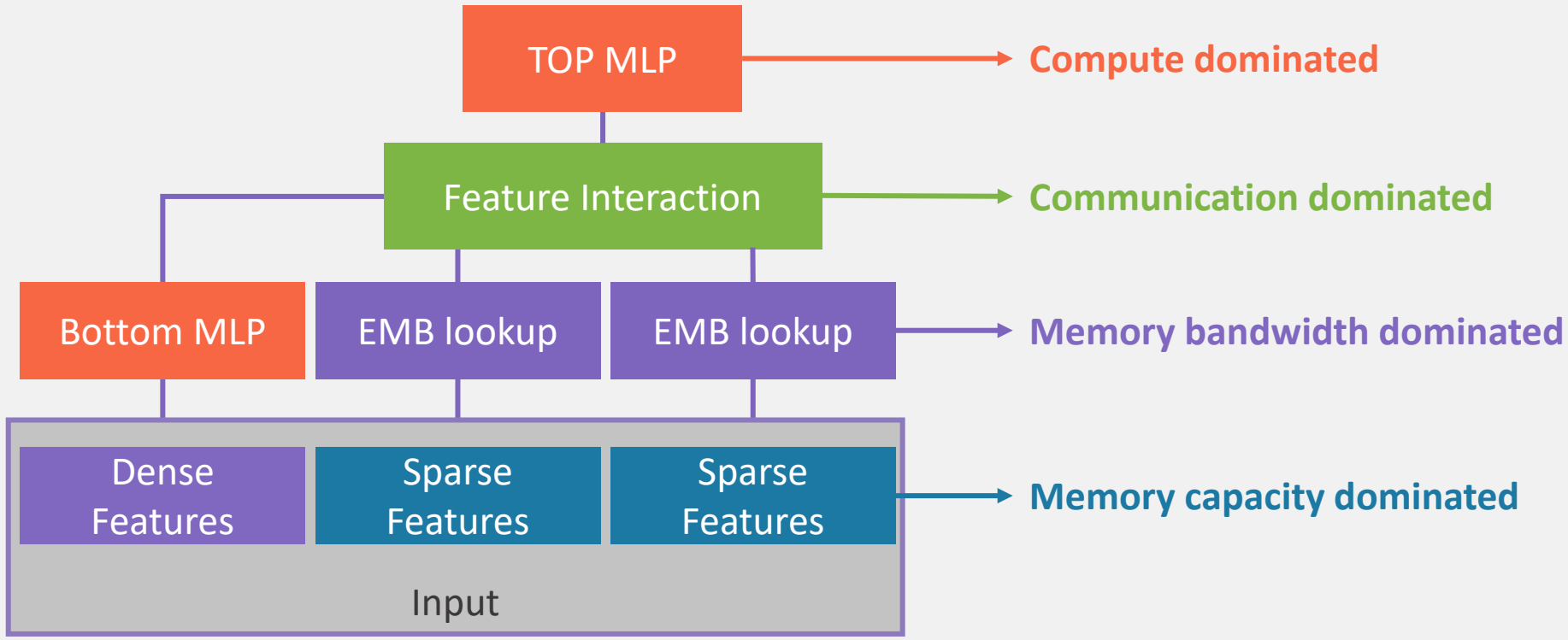
Cloud Recommendation Inferences



Cloud Recommendation Inferences Challenges

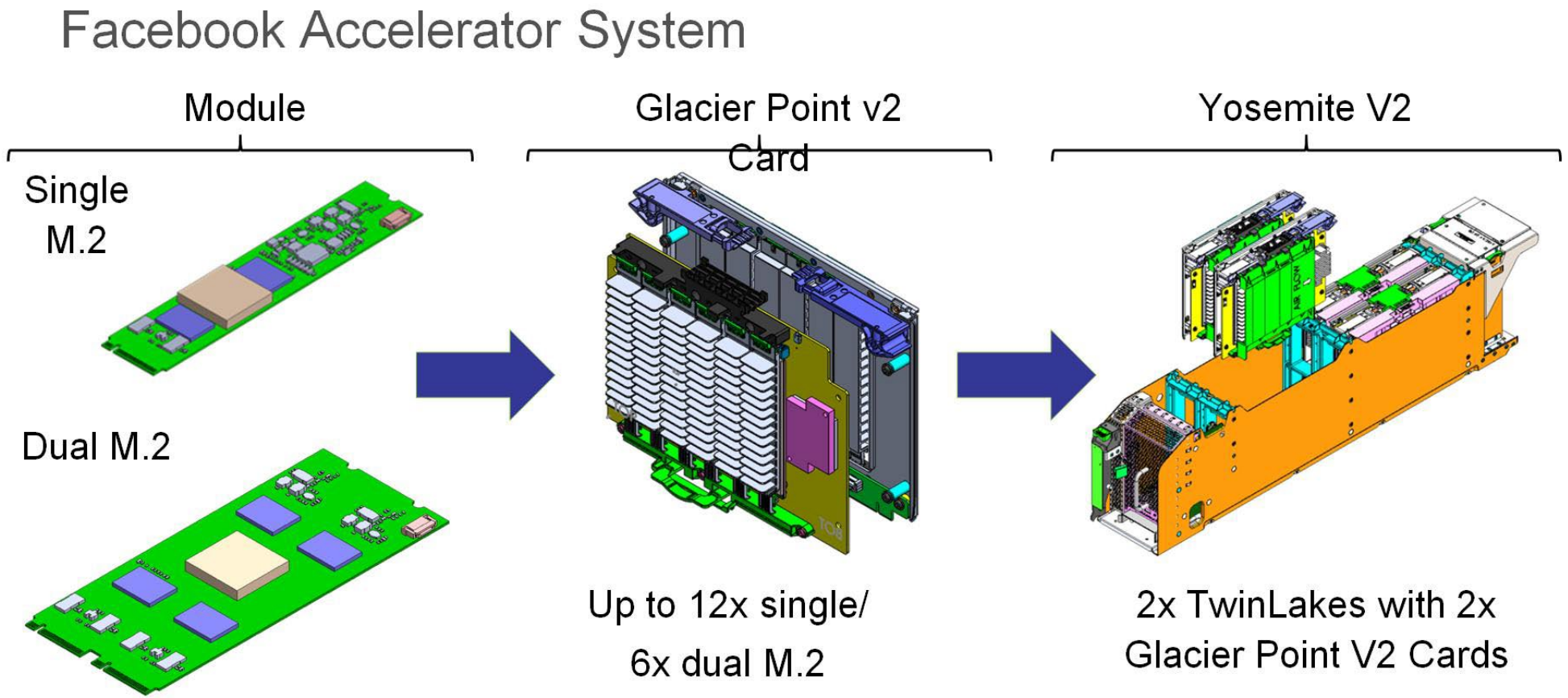


Meta's Grand Challenge to the Industry



https://ai.facebook.com/blog/dlrm-an-advanced-open-source-deep-learning-recommendation-model/

Meta Accelerator System for Inferencing

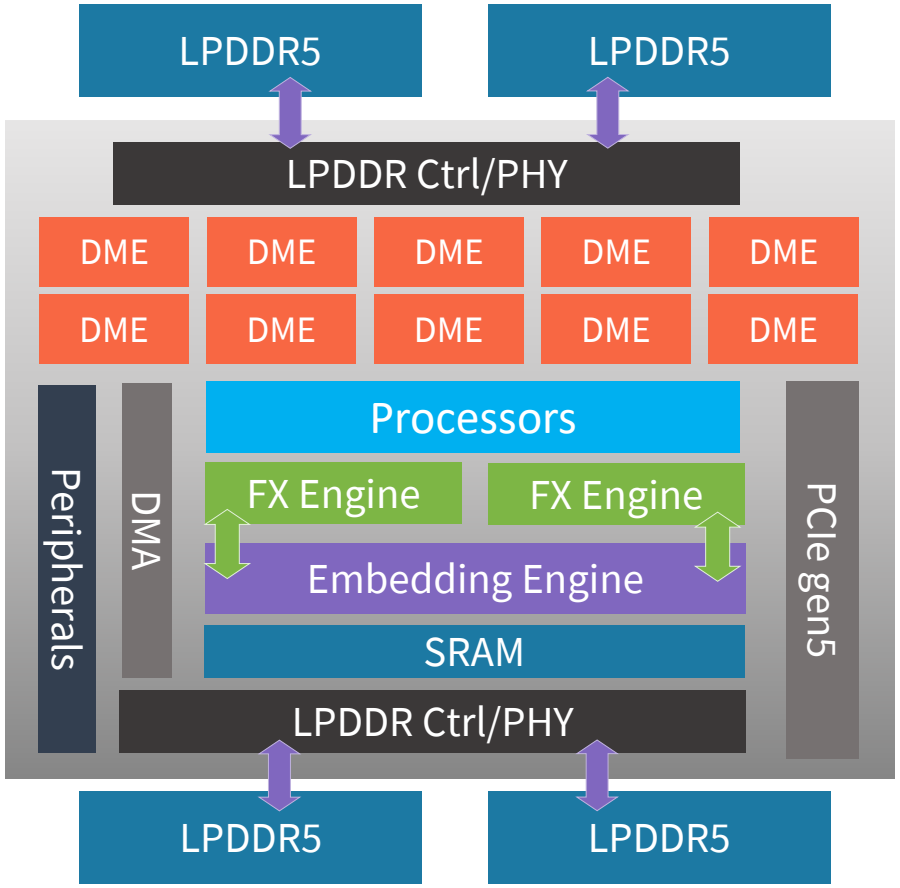


Source: OCP

DLRM Accelerator Design Challenges & Solutions

RecAccel™ - more than GPUs and systolic arrays

➢ Purpose-built IPs driven by deep understanding of DLRM dataflow



Compute solution:

✓10 Compute Engines, 192 TOPS, Low power circuits

Communication solution:

✓1.6 Tb/s inter-engine communication

Memory Bandwidth solution:

✓3.2 Tb/s table lookup

Memory Capacity solution:

✓32GB LPDDR5 per card, up to 128GB LPDDR5 per module
✓160MB SRAM per chip, FFP8 and calibrated Int8

RecAccel™ End to End Innovations - Hardware

Others

- PCIe Gen5
- LPDDR5
- 160M SRAM

Secure

- Protection
- Detection
- Recovery

Embedding Engine

Patented traffic-balanced memory allocation

- Patented smart EMB tiering
- Direct access multi-DDR for enhancing random lookup
- Up to 2TB/s to FX Engine (Feature-cross engine)

2Tb/s

FX Engine

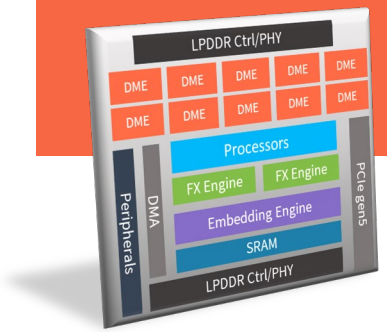
Recommender systems

- DLRM
- Wide n Deep
- Deep Cross Net
- NCF

DME (Dynamic MLP Engine)

Silicon Engineering

- Near-threshold voltage(NTV) computing
- PPA-optimized mega-cell



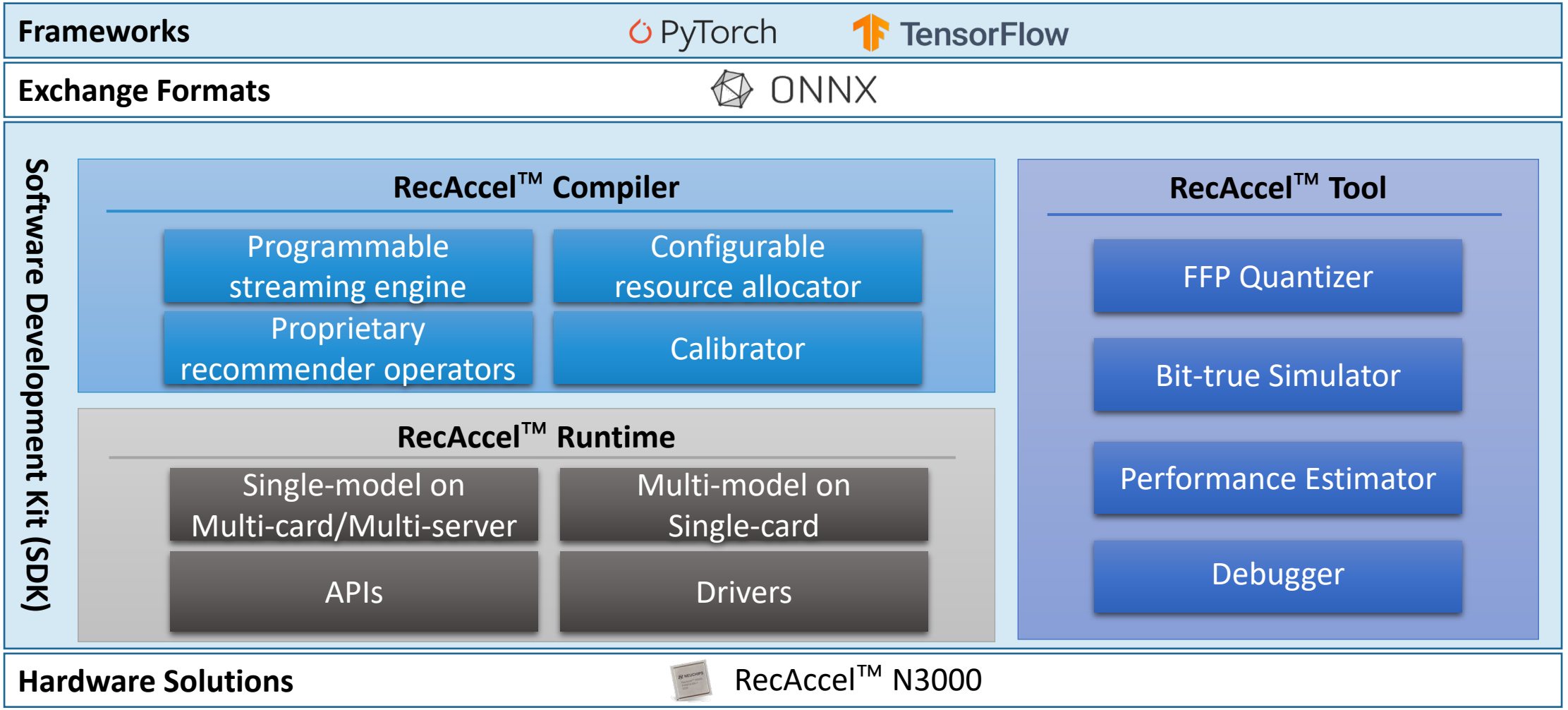
Patented Technology

1. Fast MAC
2. Sparse matrix optimization
3. Power-saving compute sharing
4. Power/performance efficient approximate computing

Features

1. Power-saving data broadcasting
2. Data movement reduction
3. Dynamic compute deployment

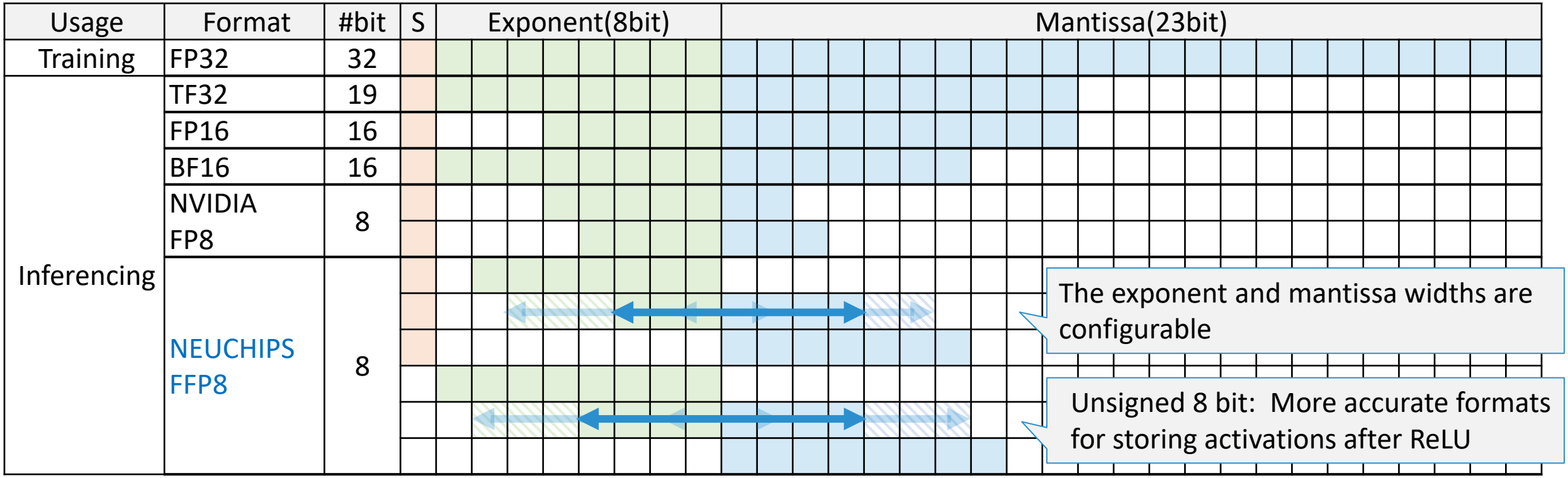
RecAccel™ End to End Innovations - Software



Patented FFP8 (US Patent: 17/238,226)

Flexible 8-bit Floating-Point Format (FFP8) delivers highest inference accuracy

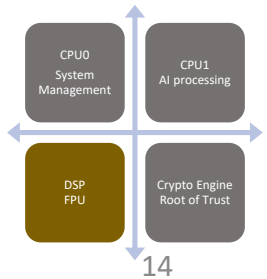
Usage	Format	#bit	S	Exponent(8bit)	Mantissa(23bit)
Training	FP32	32			
Inferencing	TF32	19			
	FP16	16			
	BF16	16			
	NVIDIA FP8	8			
	NEUCHIPS FFP8	8			



S: Sign bit

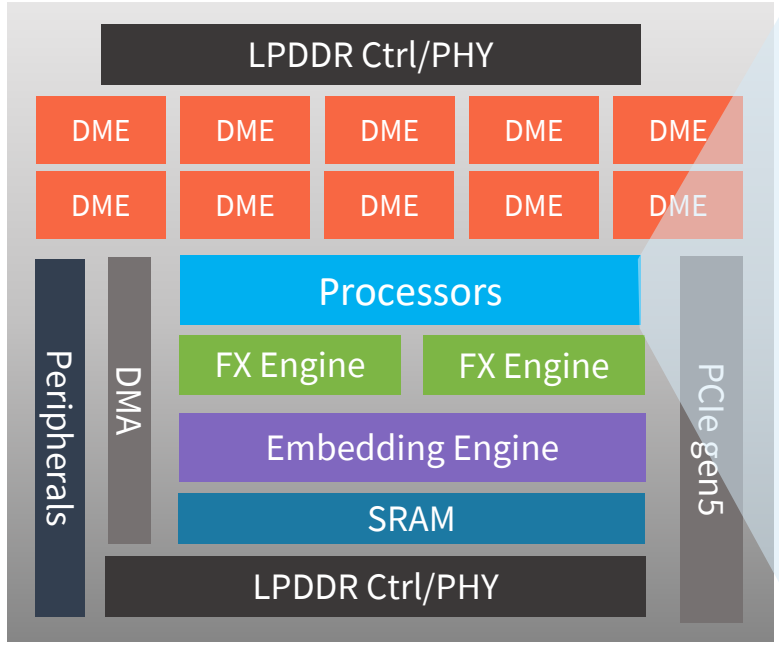
DSP: ARC EV72 Processor

- The EV Processors are designed to integrate seamlessly into the system and can be used with processors and operate in parallel with the hosts.
  - Two vector processing units (~ VPX5 x2)
  - Integrates 32-bit scalar core and 512-bit vector processor
  - IEEE 754-compliant vector floating point unit (FPU)
    - Single or half precision
    - Advanced math functions such as
      - div,
      - sqrt(x)
      - 1/sqrt(x)
      - 2^x
      - sin(x)
      - cos(x)
      - e^x



OCP Design
Challenges & Solutions

Processor Requirements for RecAccel™-N3000



Efficient

- Processors should be able to provide **efficient configuration** for PPA saving

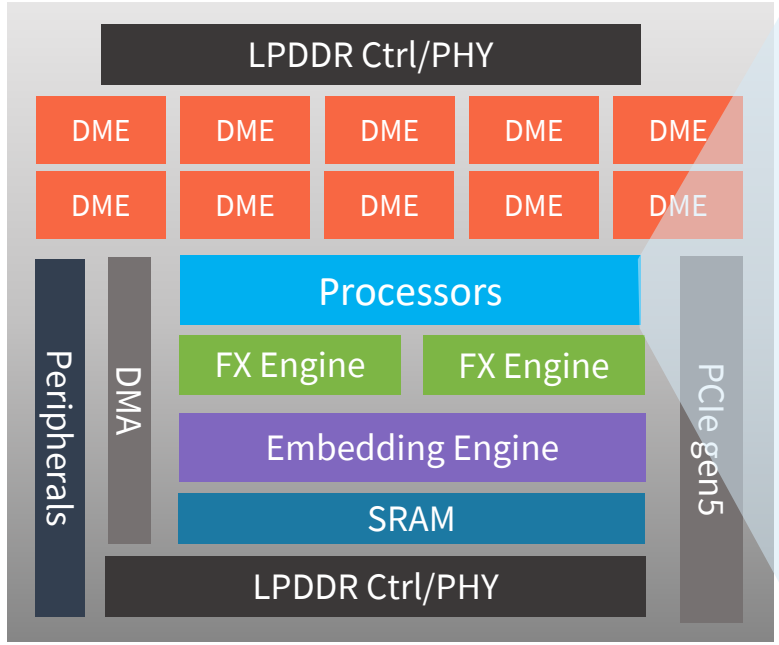
Flexible

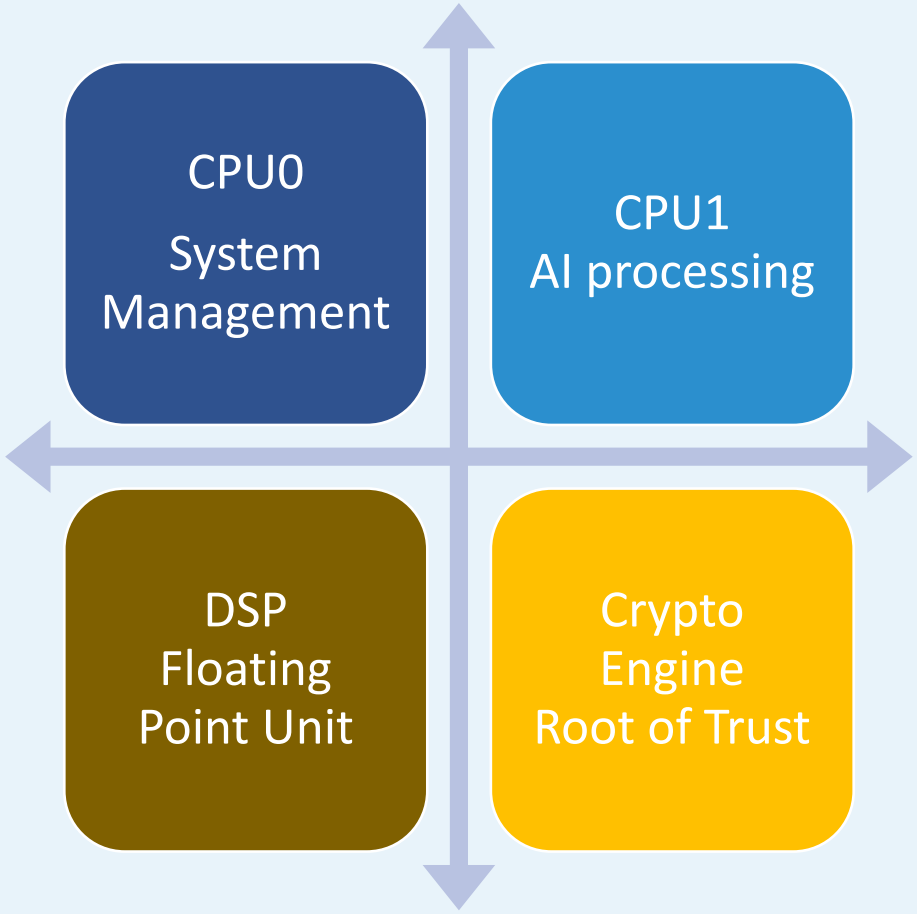
- The processor/DSP should be able to support general mathematic operations for various AI recommendation models

Secure

- Support OCP security requirements for datacenter

RecAccel™ N3000 - Processors





CPU0: System Management

Mission

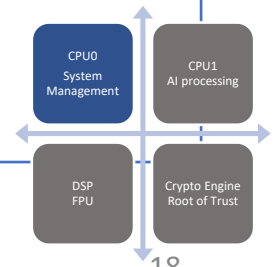
- System boot-up control
- Error & Interrupt handling

Requirement

- Access all blocks in the system for event handling
- Cooperate with Cryptal engine for secure-boot & authentications

Configuration

- Quad-core processor
- Data cache & Instruction cache is suitable for Linux OS
- Memory management unit (MMU) with **40-bit physical address**



CPU1: AI processing

Mission

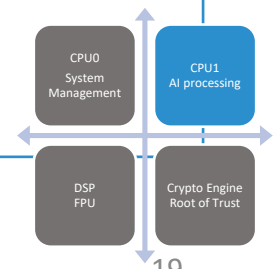
- AI engine handling
- Data I/O handling

Requirement

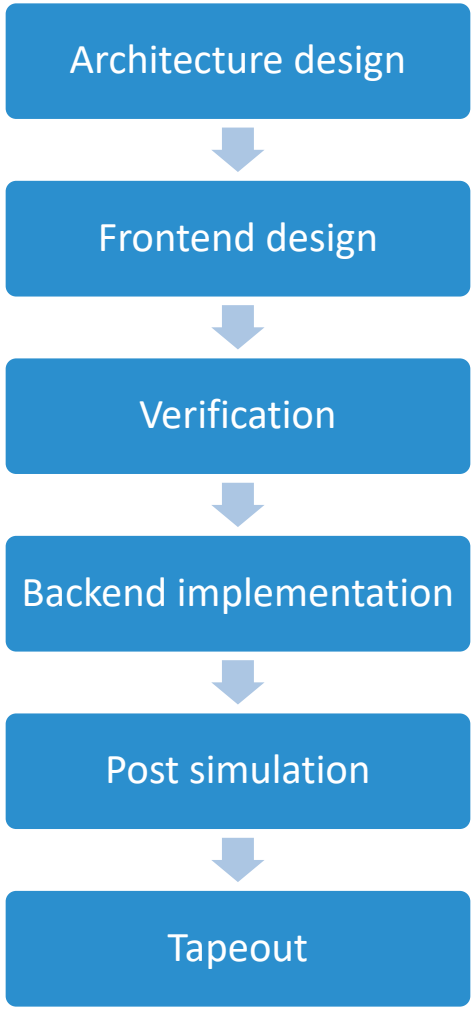
- Performance driven, able to handle inference data movement to maximum system performance

Configuration

- Dual-core processor
- Small Data cache and instruction cache per core
- High data closely coupled memories (DCCM)
- High instruction closely coupled memories (ICCM)
- High Cluster shared memory (CSM)



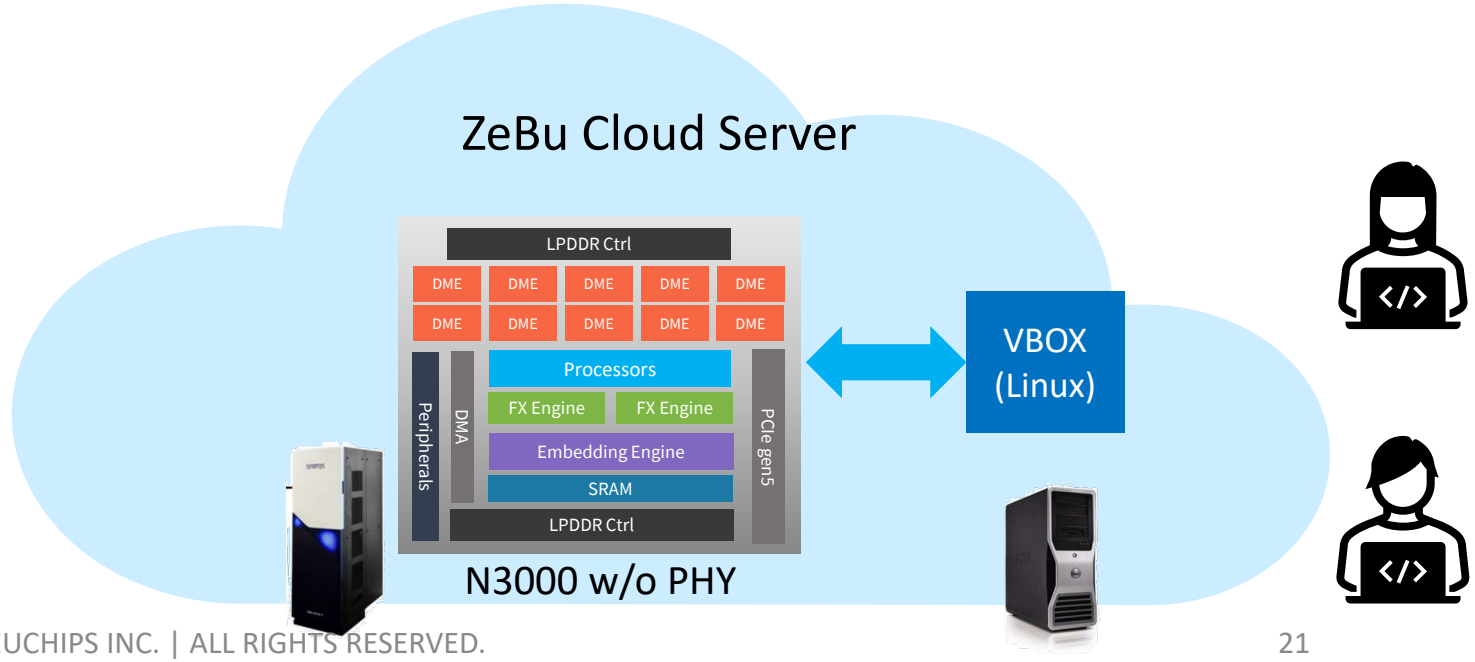
Implementation Challenges



- Design size is huge
  - NO suitable prototyping solution
- Simulation time is long
  - For Linux booting, RTL simulation takes a week
- Performance tuning & optimization
  - Dynamic workload balancing
  - Correctness vs. accuracy balancing
  - Data movement (activation/weight re-use) optimization

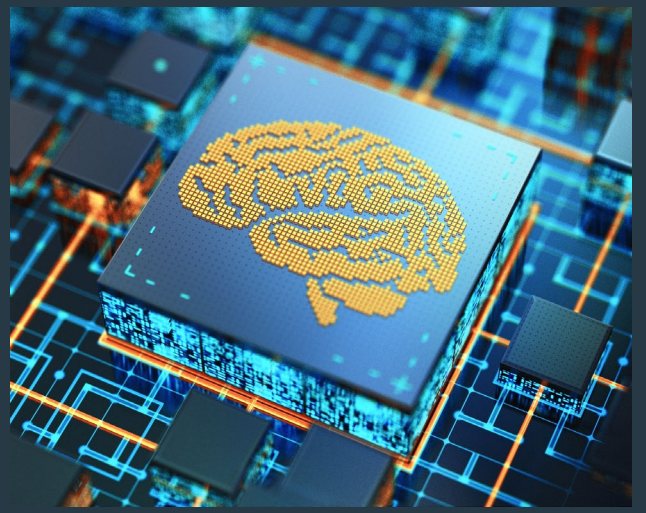
System Verification at Zebu Server

- Use Zebu emulator for whole system & application verification. By removing the PCIe PHY, and connect it with PCIe transactor, the Virtual Box on host server will be able to link with the PCIe for access the DUT (N3000).

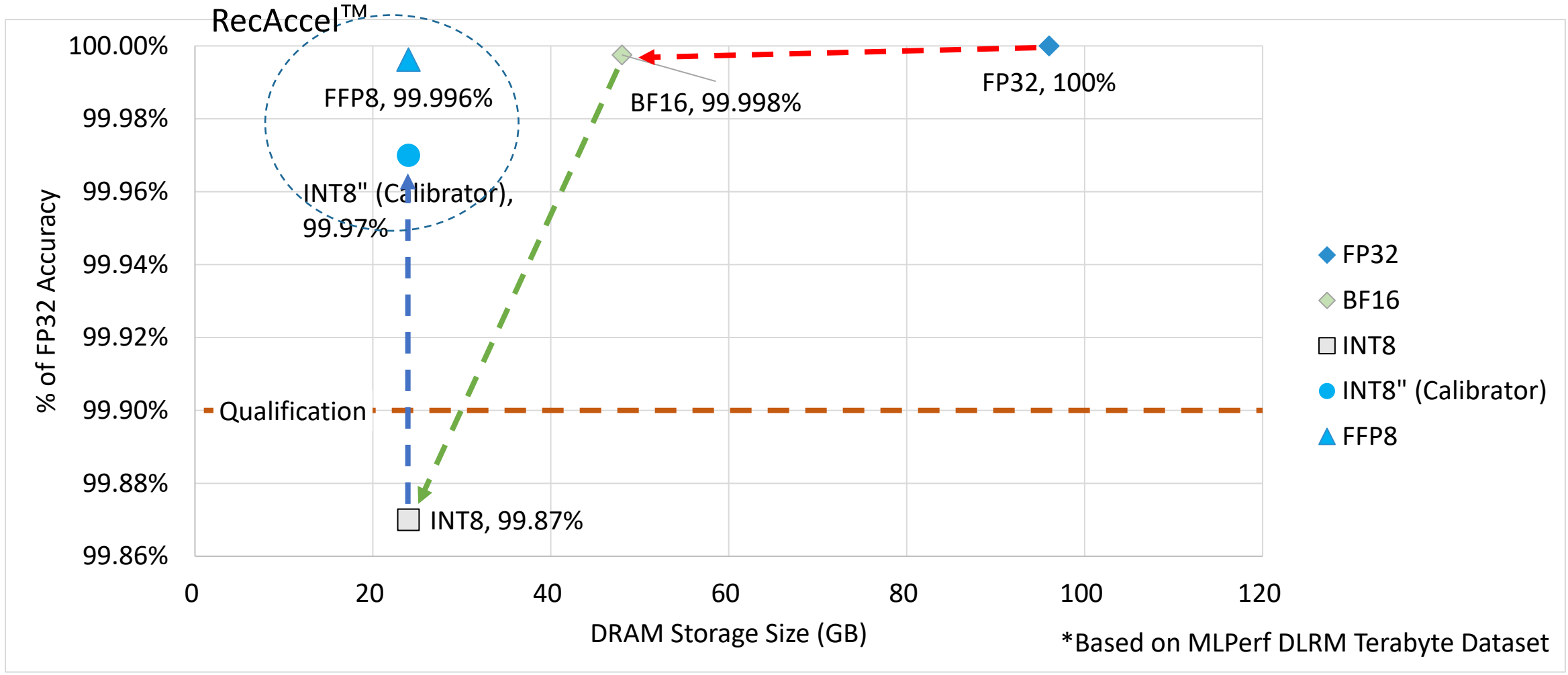


NEUCHIPS

RecAccel™ Results

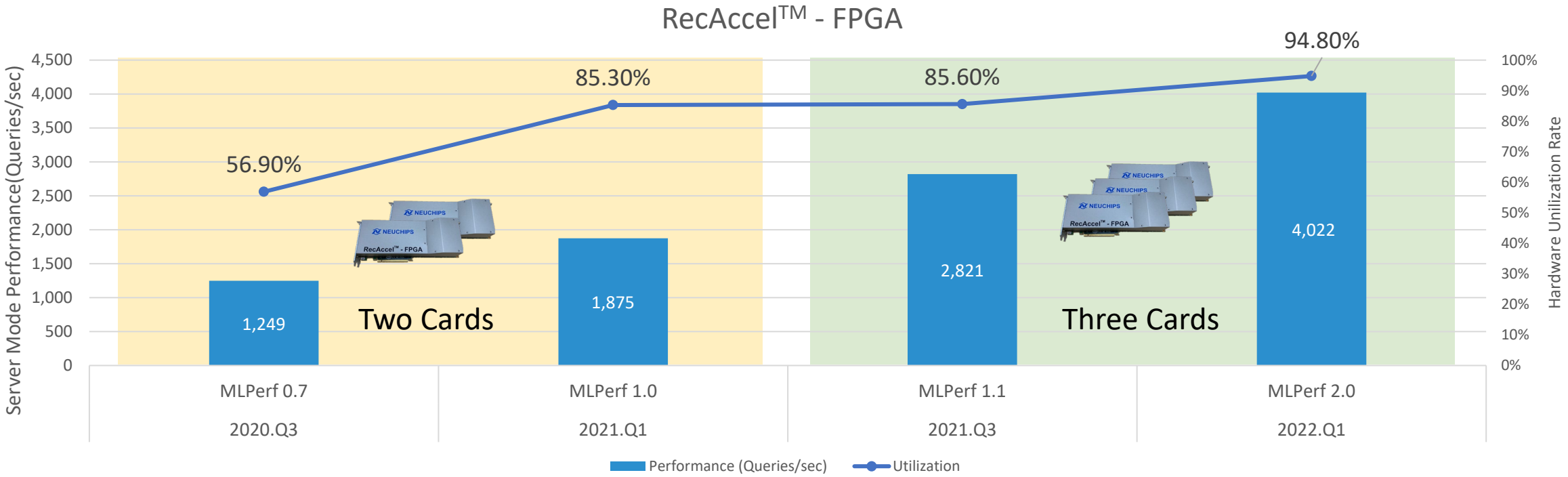


High Accuracy Coefficient Quantization and Calibration

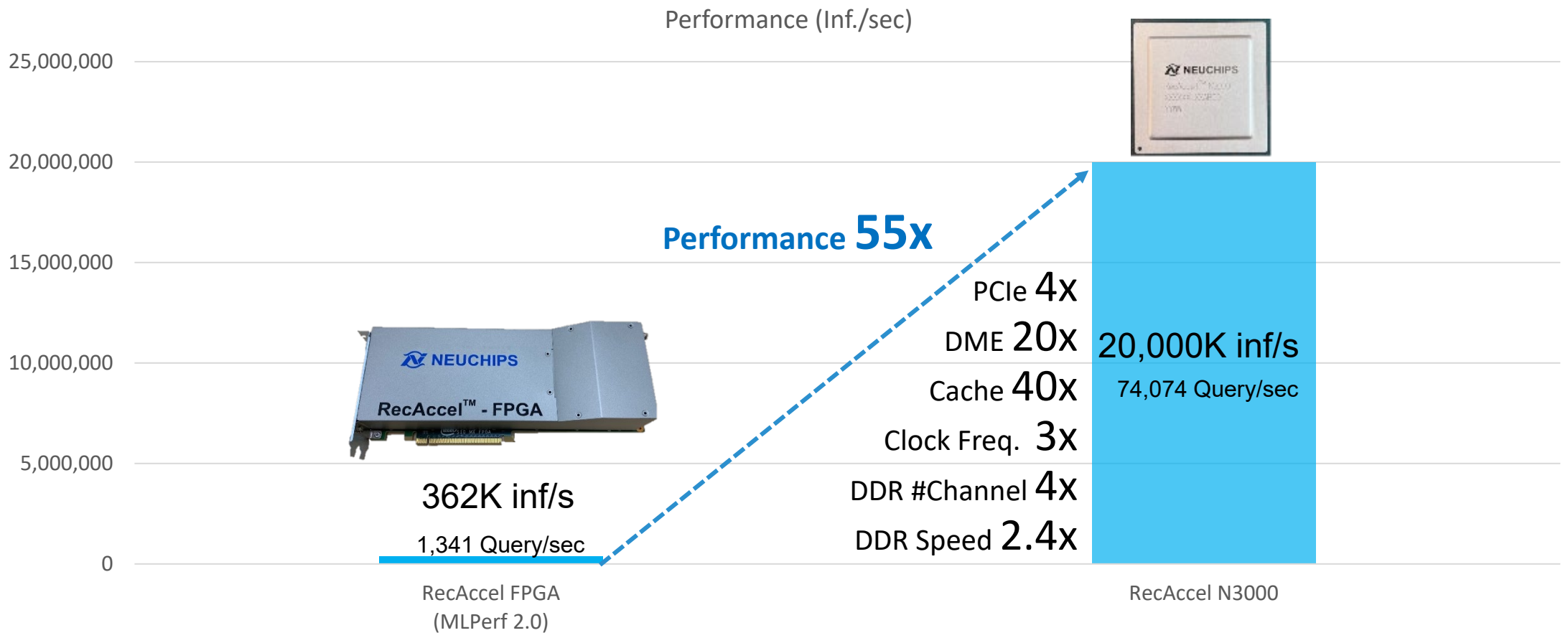


FPGA Proof-of Concept: RecAccel™- MLPerf

- NEUCHIPS is a founding member of MLCommons
- RecAccel™-FPGA is the world's 1st DLRM domain-specific accelerator at ***MLPerf*** benchmarking (datacenter inference) since 2020/Q3 (MLPerf 0.7)



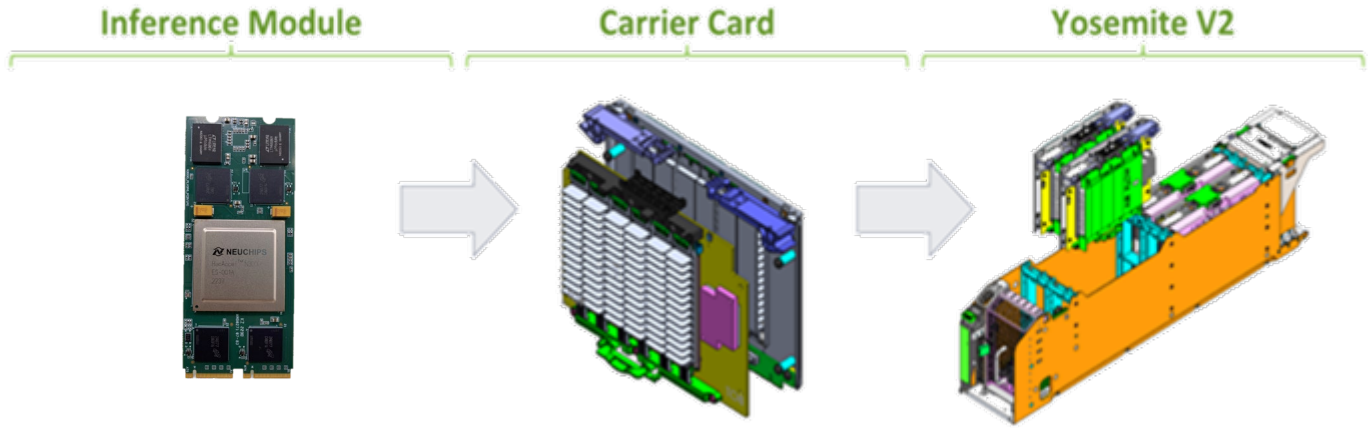
RecAccel™-FPGA vs. RecAccel™-ASIC



DME: Dynamic MLP Engine

NEUCHIPS Product Plan – RecAccel™ ASIC, Module and System

RecAccel™ N3000



Dual M.2 module (25W)

Up to 6x dual M.2

2x Twin Lakes with 2x carrier cards

OPEN Compute Project SOLUTION PROVIDER

OCP READY

Source: OCP

Appendix

IEEE Spectrum FOR THE TECHNOLOGY INSIDER

https://spectrum.ieee.org/ai-benchmarks-mlperf-performance

NEWS | COMPUTING

Benchmark Shows AIs Are Getting Speedier

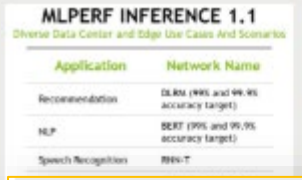
MLPerf stats show some systems have doubled performance this year, competing benchmark coming

BY SAMUEL K. MOORE | 24 SEP 2021 | 4 MIN READ

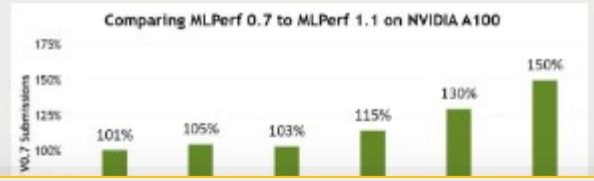
This week, AI industry group MLCommons released a new set of results for AI performance. The new list, MLPerf Version 1.1, follows the first official set of benchmarks by five months and includes more than 1800 results from 20 organizations, with 350 measurements of energy efficiency. The majority of systems improved by between 5-30 percent from earlier this year, with some more than doubling their previous performance stats, according to MLCommons. The new results come on the heels of the announcement, last week, of a new machine-learning benchmark, called TCP-AIx.

In MLPerf's inferencing benchmarks, systems made up of combinations of CPUs and GPUs or other accelerator chips are tested on up to six neural networks performing a variety of common functions—image classification, object detection, speech recognition, 3D medical imaging, natural language processing, and recommendation. For commercially available datacenter-based systems they were tested under two conditions—a simulation of real datacenter activity where queries arrive in bursts and "offline" activity where all the data is available at once. Computers meant to work onsite instead of in the data center—what MLPerf calls the edge—were measured in the offline state and as if they were receiving a single stream of data, such as from a security camera.

Although there were datacenter-class submissions from Dell, HPE, Inspur, Intel, LTech Korea, Lenovo, Nvidia, Neuchips, Qualcomm, and others, all but those from Qualcomm and Neuchips used Nvidia AI accelerator chips. Intel used no accelerator chip at all, instead demonstrating the performance of its CPUs alone. Neuchips only participated in the recommendation benchmark, as their accelerator, the RecAccel, is designed specifically to speed up recommender systems—which are used for recommending e-commerce items and for ranking search results.



For the results Nvidia submitted itself, the company used software improvements alone to eke out as much as a 50 percent performance improvement over the past year. The systems tested were usually made up of one or two CPUs along with as



A100 accelerators were paired with server-class Arm CPUs instead of x86 CPUs. The results were nearly identical between Arm and x86 systems across all six benchmarks. "That's an important milestone for Arm," says Salvator. "It's also a statement about the readiness of our software stack to be able to run the Arm architecture in a datacenter environment."

Samuel K. Moore

Samuel K. Moore is the senior editor at *IEEE Spectrum* in charge of semiconductors coverage. An IEEE member, he has a bachelor's degree in biomedical engineering from Brown University and a master's degree in journalism from New York University.

Although there were datacenter-class submissions from Dell, HPE, Inspur, Intel, LTech Korea, Lenovo, Nvidia, Neuchips, Qualcomm, and others, all but those from Qualcomm and Neuchips used Nvidia AI accelerator chips. Intel used no accelerator chip at all, instead demonstrating the performance of its CPUs alone. Neuchips only participated in the recommendation benchmark, as their accelerator, the RecAccel, is designed specifically to speed up recommender systems—which are used for recommending e-commerce items and for ranking search results.

https://www.eetimes.com/neuchips-tapes-out-recommendation-accelerator-for-world-beating-accuracy/

DESIGNLINES | AI & BIG DATA DESIGNLINE

Neuchips Tapes Out Recommendation Accelerator for World-Beating Accuracy

By Sally Ward-Foxton 06.23.2022 0

recommendation models. Emulation of the chip suggests it will be the only solution on the market to achieve one million DLRM inferences per Joule of energy (or 20 million inferences per second per 20-Watt chip). The company has already demonstrated that its software can achieve world-beating INT8 DLRM accuracy at 99.97% of FP32 accuracy.

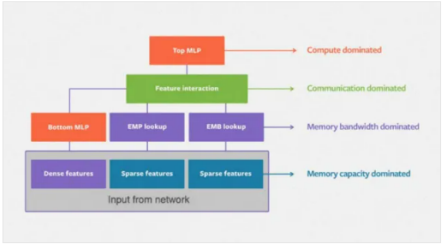
Neuchips was founded in response to a call by Facebook (now Meta) in 2019 for the industry to work (hardware acceleration for recommendation inference. The Taiwanese startup set out to do exactly thi and the company is one of only two startup entrants specifically targeting recommendation (the othe Esperanto with its 1000-core RISC-V design).

"According to many reports, most of the AI inference cycles in the data center are actually for recommendation models, not vision or language… so we think recommendation is an important market," Neuchips CEO Youn-Long Lin told EE Times, adding that the number of recommendation inferences required is growing steadily. "The power consumption is fixed, so the essential issue is that we have to do as much as possible within an energy budget in order to increase prediction accuracy."

Youn-Long Lin (Source: Neuch

Prediction accuracy is very important for recommendation applications, such as online shopping, wh any loss in accuracy means a corresponding loss in revenue for online shopping platforms.

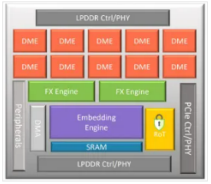
DLRM (deep learning recommendation model), Meta's open-source recommendation model, has quit different characteristics compared to the CNNs widely used for computer vision. Dense features, thos with continuous values such as customer age or income, are extracted by multilayer perceptron (MLP type of neural network) while sparse features (yes or no questions) use embedding tables. There may many hundreds of features or more, and embedding tables can be gigabytes in size. Interactions betw these features would indicate the relationship between products and users for online shopping platfo These interactions are computed explicitly — DLRM uses a dot product. And then these interactions through another neural network.



Structure of the DLRM recommendation network. Neural networks are marked in orange, embedding tables in purple and dot product in green (Source: Meta)

While neural network computation may be compute-bound, the other operations required for DLRM may be bound by memory capacity, memory bandwidth, or communication. This makes DLRM a very hard model to accelerate with general-purpose AI accelerators, including those developed for applications such as image processing.

Neuchips' ASIC solution, RecAccel, includes specially designed engines to accelerate embeddings (marked purple in diagram below), matrix multiplication (orange) and feature interaction (green).



Neuchips' recommendation inference accelerator chip includes hardware engines designed for the key parts of the recommendation workload (Source: Neuchips)

"In the embedding engine, mostly the issue is to look up multiple tables simultaneously and very fast," Lin said. "Recommendation model sizes vary a lot — some are very small, some are very large. The important issue is how to allocate tables to both off-chip and on-chip memory appropriately."

Neuchips' embedding engine reduces access to off-chip memory by 50% and increases bandwidth utilization by 30%, the company said, via a novel cache design and DRAM traffic optimization techniques.

Different recommendation models use different operations for feature interaction — DLRM uses dot product, but there are others. Lin said Neuchips' feature interaction engine supports this kind of flexibility.

The chip has 10 compute engines with 16K MAC per engine.

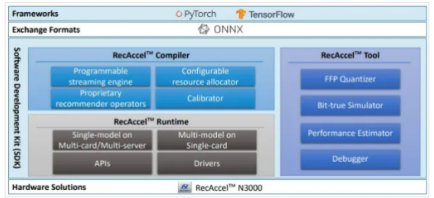
"The important issue here is how to implement this compute engine with low power consumption and so it can handle sparse matrices efficiently," Lin said. The compute engines consume 1 microjoule per inference at the SoC level.

Lin added that hardware features can also terminate computation when a certain level of accuracy is reached, to save power.

SOFTWARE STACK

Neuchips already has a complete software stack up and running, including compiler, runtime, and toolchain, as evidenced by two successful MLPerf submissions.

The SDK supports both splitting big models across multiple chips/cards and running multiple smaller inferences per chip (Lin said that Meta has several hundred DLRM models in production with vastly different sizes and characteristics).



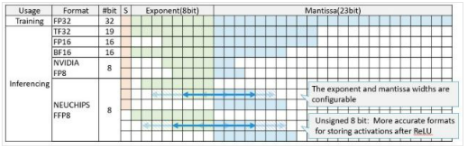
Neuchips' software development kit (SDK) includes compiler, runtime and toolchain and has already been demonstrated successfully in previous MLPerf rounds (Source: Neuchips)

Neuchips' secret weapon is the new 8-bit number format it invented, and patented, called flexible floating point or FFP8.

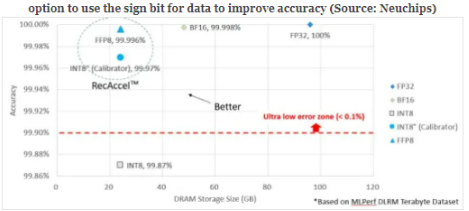
"[FFP8] means our circuit can be more adaptive to the model, and that's how we achieve high accuracy," Lin said. "The training part is always in 32-bit, and you can use 32-bit to inference, if you don't care about the energy consumption, but with 8-bit, the energy consumption is one-sixteenth… The problem is the trade off between how much accuracy loss you are willing to suffer to gain the computing efficiency."

Companies such as Nvidia and Tesla are moving towards 8-bit floating point formats where possible, pointing towards a consensus on 8-bit computation for inference, Lin said. Neuchips' FFP8 is a superset of these formats, with configurable exponent and mantissa widths. There is also an unsigned version which uses the extra bit to increase accuracy of stored activations after ReLU operations.

Neuchips' calibrator block (part of the compiler) "defines the quantization and representation format according to model and data characteristics," said Lin. This calibrator was able to achieve what Neuchips says is the world's best DLRM accuracy at INT8 — 99.97% of the accuracy of an FP32 version of the model. Use calibration in combination with FFP8 (to determine the exact format used for different parts of the model), and accuracy improves to 99.996%, close to what can be achieved with bigger formats like BF16.



Neuchips' FFP8 format has configurable exponent and mantissa widths, and the option to use the sign bit for data to improve accuracy (Source: Neuchips)



Neuchips' accuracy results for its calibration process, and for calibration plus FFP8 format, normalized to FP32 accuracy (Source: Neuchips)

PATENTS FILED

Neuchips was founded in 2019 by Lin, a computer science professor at the National Tsing Hua University in Taiwan, previously co-founder and CTO of design services company Global Unichip Corp (now part of TSMC), along with an experienced team from Mediatek, Novatek, Realtek, GUC, and TSMC.

The company employs 38 people in Taiwan, of which 30 are engineers, including many former students of Lin's. The company has filed 30 patents so far, and received 8 U.S. and 12 Taiwan patents.

Neuchips' RecAccel chip has taped out and will be manufactured in TSMC 7nm, occupying 400mm^2. The chip will be available on dual M.2 modules ready to go onto Glacier Point carrier cards (6 modules per Glacier Point) and PCIe Gen 5 cards. Both cards will begin sampling in Q4 '22.

Sally Ward-Foxton

Sally Ward-Foxton covers AI technology and related issues for EETimes.com and all aspects of the European industry for EETimes Europe magazine. Sally has spent more than 15 years writing about the electronics industry from London, UK. She has written for Electronic Design, ECN, Electronic Specifier: Design, Components in Electronics, and many more. She holds a Masters' degree in Electrical and Electronic Engineering from the University of Cambridge.