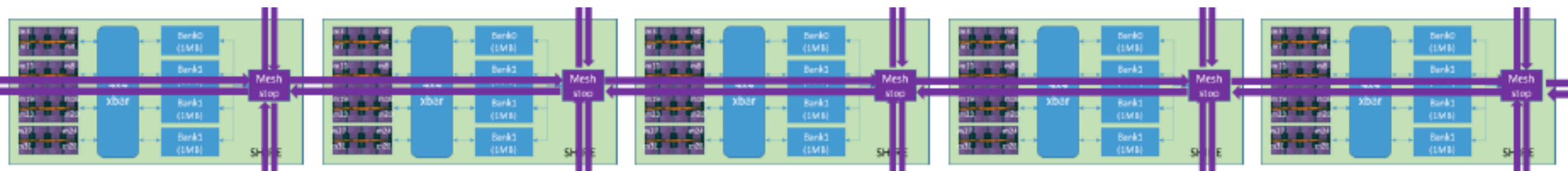


# Accelerating ML Recommendation with over a Thousand RISC-V/Tensor Processors on Esperanto's ET-SoC-1 Chip

Dave Ditzel, email: [dave@esperanto.ai](mailto:dave@esperanto.ai), Founder and Executive Chairman, Esperanto Technologies,

Roger Espasa, Nivard Aymerich, Allen Baum, Tom Berg, Jim Burr, Eric Hao, Jayesh Iyer, Miquel Izquierdo, Shankar Jayaratnam, Darren Jones, Chris Klingner, Jin Kim, Stephen Lee, Marc Lupon, Grigorios Magklis, Bojan Maric, Rajib Nath, Mike Neilly, Duane Northcutt, Bill Orner, Jose Renau, Gerard Reves, Xavier Reves, Tom Riordan, Pedro Sanchez, Sri Samudrala, Guillem Sole, Raymond Tang, Tommy Thorn, Francisco Torres, Sebastia Tortella, Daniel Yau



# The ET-SoC-1 is the highest performance commercial RISC-V chip

---

The ET-SoC-1 has over a thousand RISC-V processors on a single TSMC 7nm chip, including:

- 1088 energy-efficient ET-Minion 64-bit RISC-V in-order cores each with a vector/tensor unit
- 4 high-performance ET-Maxion 64-bit RISC-V out-of-order cores
- >160 million bytes of on-chip SRAM
- Interfaces for large external memory with low-power LPDDR4x DRAM and eMMC FLASH
- PCIe x8 Gen4 and other common I/O interfaces
- Innovative low-power architecture and circuit techniques allows entire chip to
  - Compute at peak rates of 100 to 200 TOPS
  - Operate using under 20 watts for ML recommendation workloads

This general-purpose parallel-processing system on a chip can be used for many parallelizable workloads

But today, we want to show why it is a compelling solution for Machine Learning Recommendation (inference)

- ML Recommendation is one of the hardest and most important problems for many hyperscale data centers

# Requirements and challenges for ML Recommendation in large datacenters

---

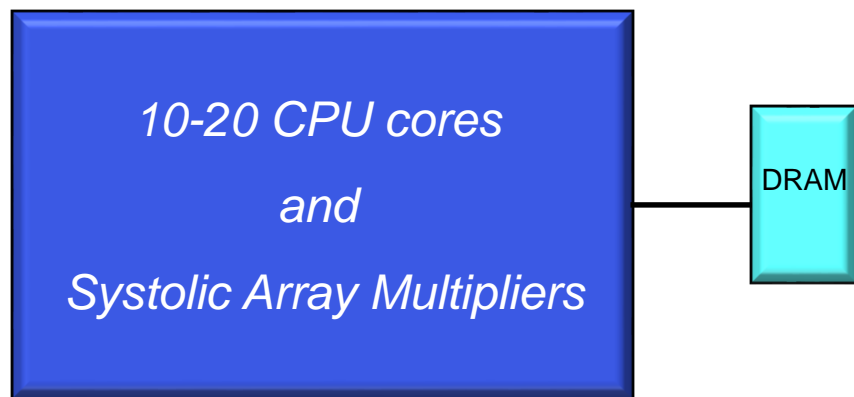
Most inferencing workloads for recommendation systems in large data centers are run on x86 servers

Often these servers have an available slot for an accelerator card, but it needs to meet key requirements:

- **100 TOPS to 1000 TOPS** peak rates to provide better performance than the x86 host CPU alone
- Limited power budget per card, perhaps **75 to 120 watts**, and must be air-cooled<sup>[1]</sup>
- Strong support for **Int8**, but must also support **FP16 and FP32** data types<sup>[1,2]</sup>
- **~100 GB** of memory capacity on the accelerator card to hold most embeddings, weights and activations<sup>[3]</sup>
- **~100 MB** of on-die memory<sup>[5]</sup>
- Handle both **dense and sparse** compute workloads. Embedding look-up is sparse matrix by dense matrix multiplication<sup>[5]</sup>
- Be **programmable** to deal with rapidly evolving workloads<sup>[1]</sup>, rather than depending on overly-specialized hardware<sup>[4,5]</sup>

# Esperanto's approach is different... and we think better for ML Recommendation

## Other ML Chip approaches

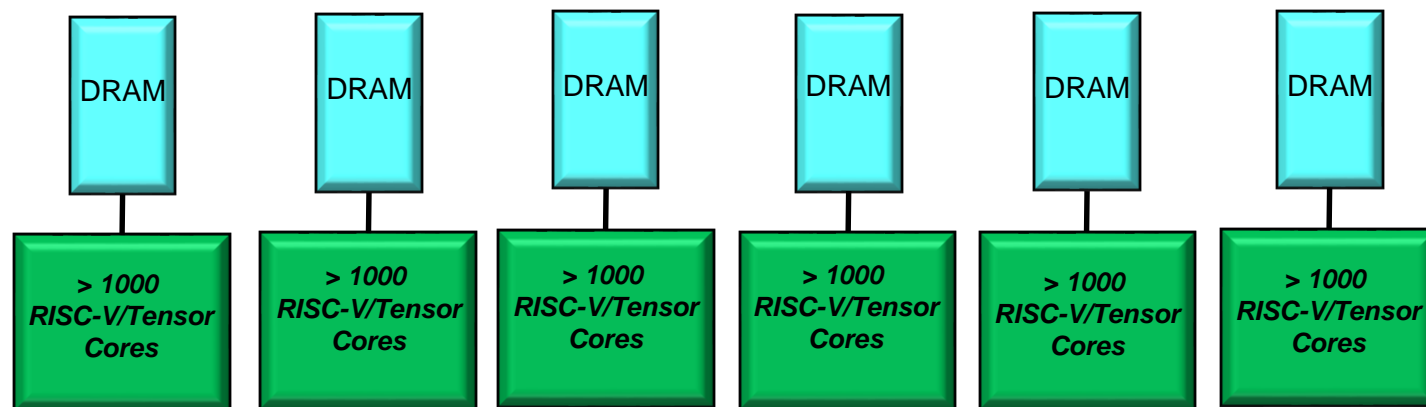


**One Giant Hot Chip** uses up power budget  
**Limited I/O** pin budget limits memory BW  
**Dependence on systolic array multipliers**

- Great for high ResNet50 score
  - Not so good with large sparse memory
- Only a **handful (10-20) of CPU cores**
- **Limited parallelism** with CPU cores when problem doesn't fit onto array multiplier

**Standard voltage:** Not energy efficient

## Esperanto's better approach



Use **multiple low-power** chips that still fit within power budget  
Performance, pins, memory, bandwidth **scale up with more chips**  
**Thousands** of general-purpose RISC-V/tensor cores

- **Far more programmable** than overly-specialized (eg systolic) hw
  - **Thousands of threads** help with large sparse memory latency
- Full parallelism** of thousands of cores always available  
**Low-voltage** operation of transistors is **more energy-efficient**
- Lower voltage operation also reduces power
  - Requires both **circuit and architecture innovations**

*Challenge: How to put the highest ML Recommendation performance onto a single accelerator card with a 120-watt limit?*

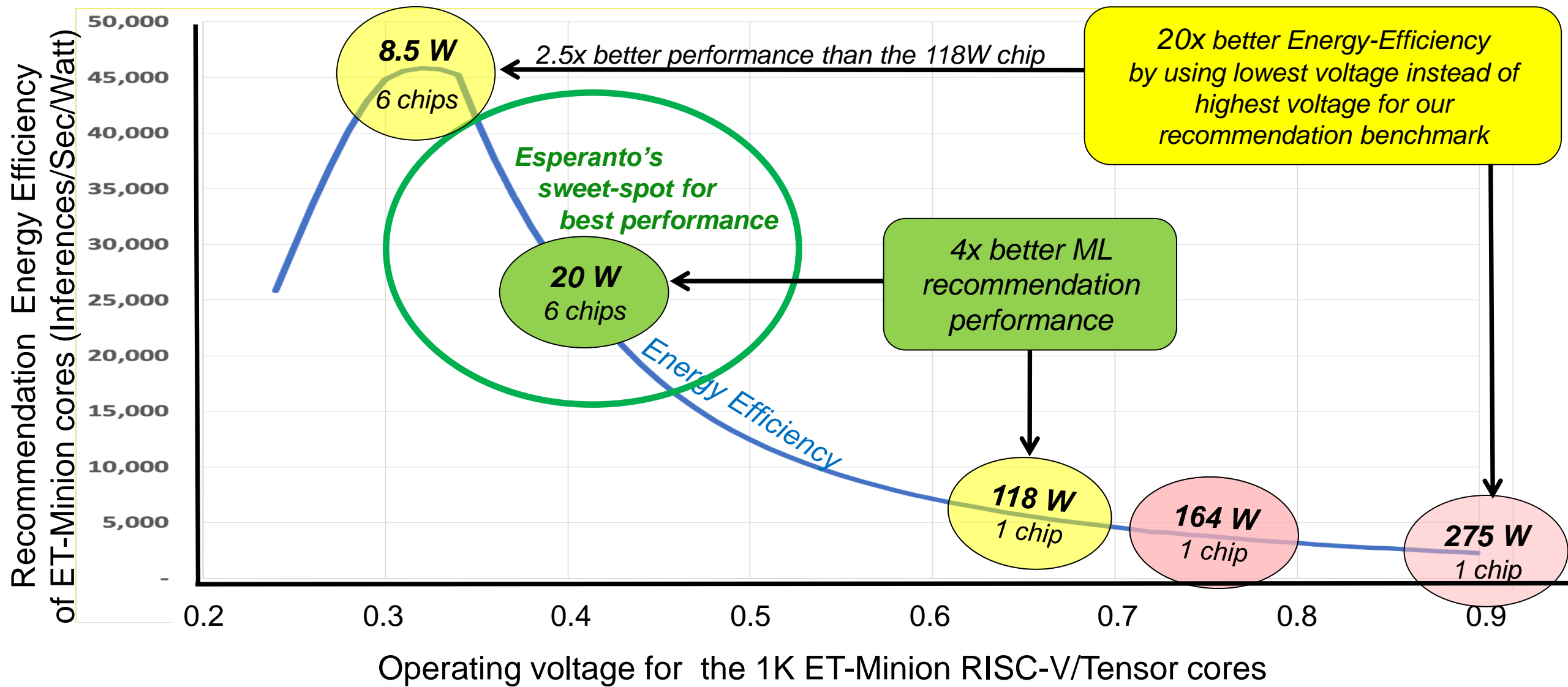
# Could fit six chips on 120W card, if each took less than 20 watts

Assumed half of 20W power for 1K RISC-V cores, so only 10 mW per core!

*Power (Watts) = C<sub>dynamic</sub> x Voltage<sup>2</sup> x Frequency + Leakage*

|   | <u>Power/core</u> | <u>Frequency</u> | <u>Voltage</u> | <u>Cdynamic</u> |
|---|-------------------|------------------|----------------|-----------------|
| Generic x86 Server core (165W for 24 cores) | 7 W               | 3 GHz            | 0.850v         | 2.2nF           |
| 10mW ET-Minion core (~10W for 1K cores)     | 0.01 W            | 1 GHz            | 0.425v         | 0.04nF          |
| Reductions needed to hit goals              | ~700x             | 3x               | 4x             | 58x             |
|   |                   | Easy             | Hard           | Very Hard       |
|   |                   |                  | Circuit/SRAM   | Architecture    |

# Study of energy-efficiency and number of chips to get best ML Performance in 120 watts<sup>[6]</sup>



# ET-Minion is an Energy-Efficient RISC-V CPU with a Vector/Tensor Unit

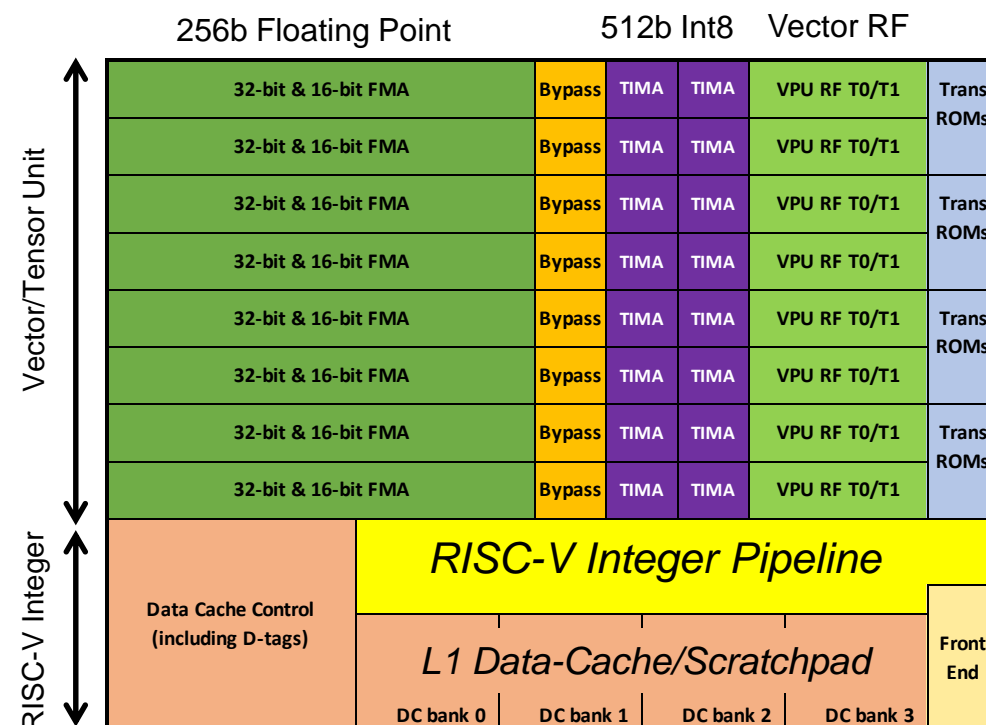
## ET-MINION IS A CUSTOM BUILT 64-BIT RISC-V PROCESSOR

- In-order pipeline with low gates/stage to improve MHz at low voltages
- Architecture and circuits optimized to enable low-voltage operation
- Two hardware threads of execution
- Software configurable L1 data-cache and/or scratchpad

## ML OPTIMIZED VECTOR/TENSOR UNIT

- 512-bit wide integer per cycle
  - 128 8-bit integer operations per cycle, accumulates to 32-bit Int
- 256-bit wide floating point per cycle
  - 16 32-bit single precision operations per cycle
  - 32 16-bit half precision operations per cycle
- New multi-cycle Tensor Instructions
  - Can run for up to 512 cycles (up to 64K operations) with one tensor instruction
  - Reduces instruction fetch bandwidth and reduces power
  - RISC-V integer pipeline put to sleep during tensor instructions
- Vector transcendental instructions

## OPERATING RANGE: 300 MHz TO 2 GHz



*ET-Minion RISC-V Core and Tensor/Vector unit  
optimized for low-voltage operation  
to improve energy-efficiency*

Optimized for energy-efficient ML operations. Each ET-Minion can deliver peak of 128 Int8 GOPS per GHz



## 8 ET-Minions form a “Neighborhood”

### NEIGHBORHOOD CORES WORK CLOSELY TOGETHER

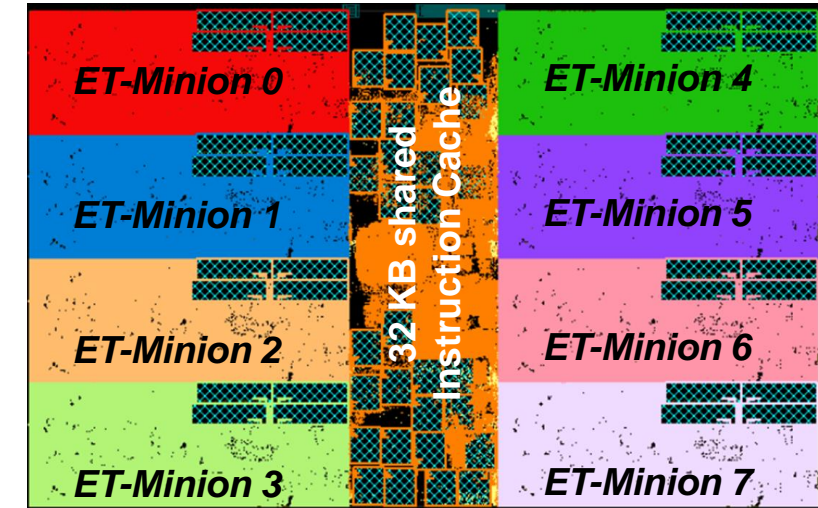
- Architecture improvements capitalize on physical proximity of 8 cores
- Take advantage that almost always running highly parallel code

### OPTIMIZATIONS FROM CORES RUNNING THE SAME CODE

- 8 ET-Minions share single large instruction cache, this is more energy efficient than many separate instruction caches.
- “Cooperative loads” substantially reduce memory traffic to L2 cache

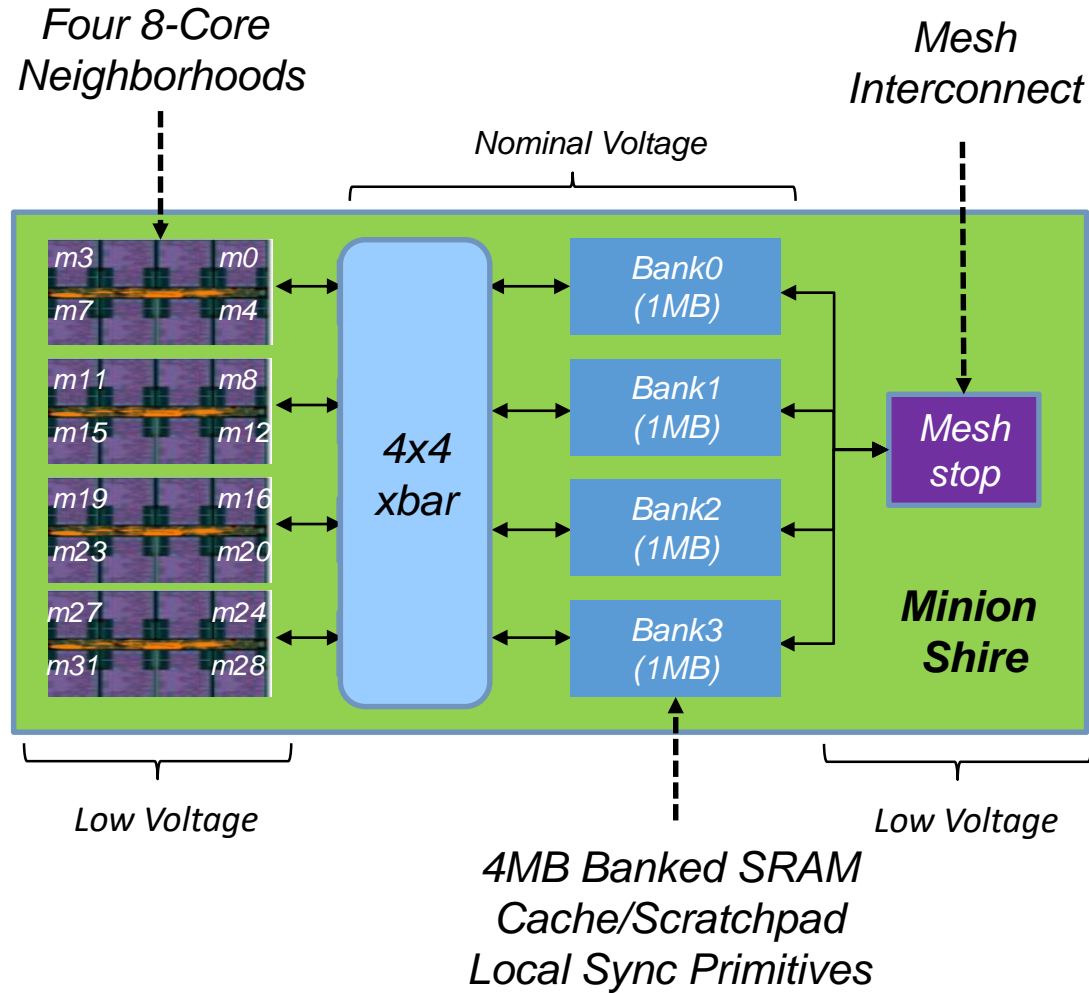
### NEW INSTRUCTIONS MAKE COOPERATION MORE EFFICIENT

- New Tensor instructions dramatically cut back on instruction fetch bandwidth
- New instructions for fast local synchronization within group
- New Send-to-Neighbor instructions
- New Receive-from-Neighbor instructions





# 32 ET-Minion CPUs and 4 MB Memory form a “Minion Shire”



## 32 ET-MINION RISC-V CORES PER MINION SHIRE

- Arranged in four 8-core neighborhoods

## SOFTWARE CONFIGURABLE MEMORY HIERARCHY

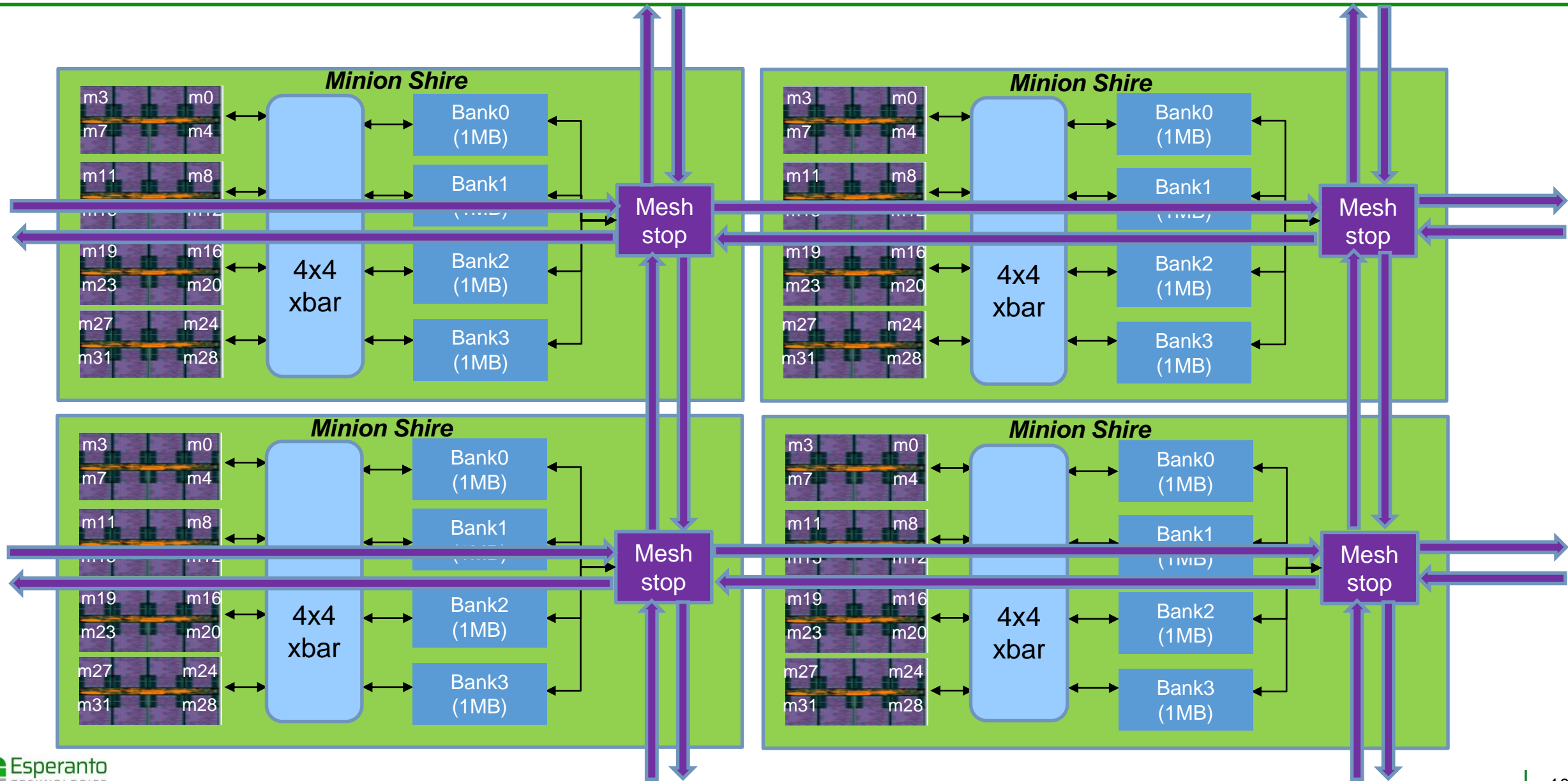
- L1 data cache can also be configured as scratchpad
- Four 1MB SRAM banks can be partitioned as private L2, shared L3 and scratchpad

## SHIRES CONNECTED WITH MESH NETWORK

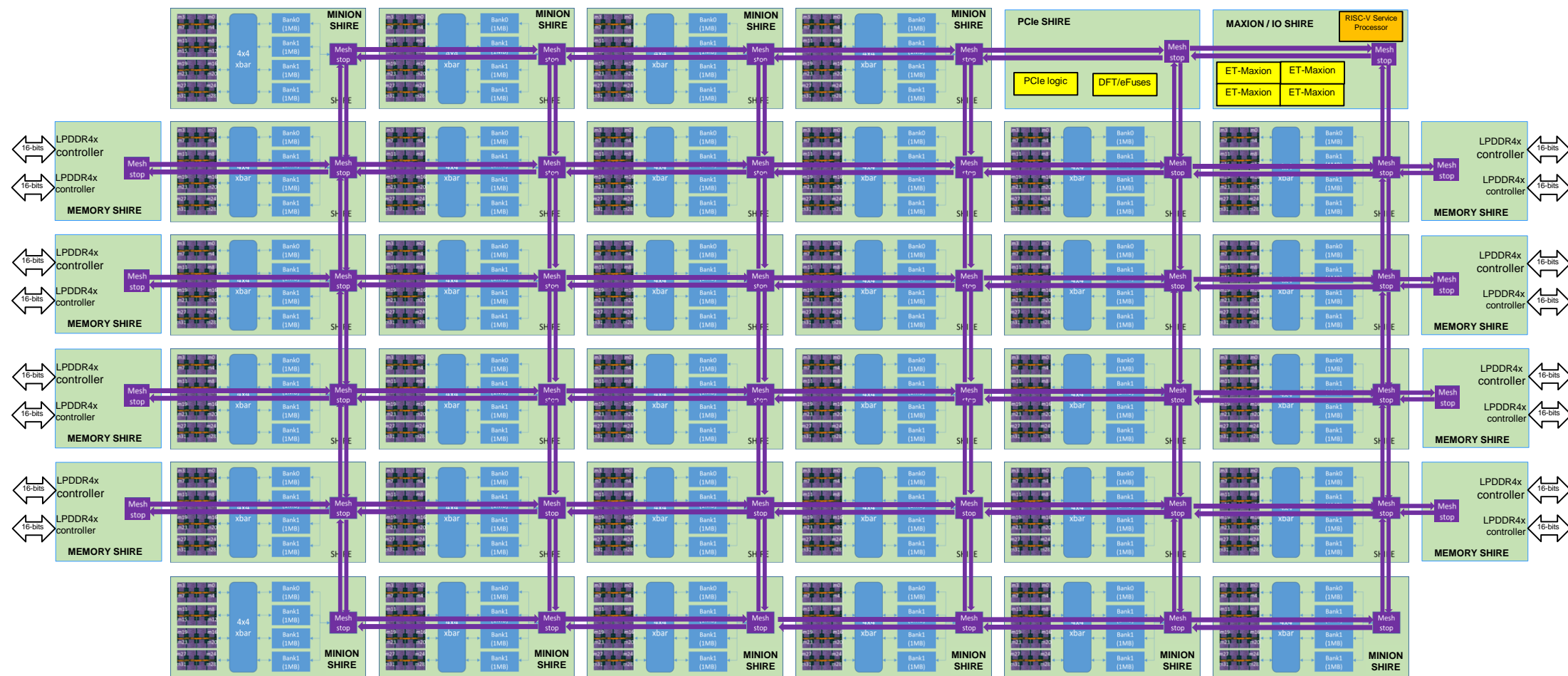
## NEW SYNCHRONIZATION PRIMITIVES

- Fast local atomics
- Fast local barriers
- Fast local credit counter
- IPI support

# Shires are connected to each other and to external memory through Mesh Network



# ET-SoC-1: Full chip internal block diagram



## 34 MINION SHIREs

- 1088 ET-Minions

## 8 MEMORY SHIREs

- LPDDR4x DRAM controllers

## 1 MAXION / IO SHIRE

- 4 ET-Maxions
- 1 RISC-V Service Processor

## PCIe SHIRE

160 million bytes of on-die SRAM

x8 PCIe Gen4

Secure Root of Trust

# ET-SoC-1 External Chip Interfaces

## 8-bit PCIe Gen4

- Root/endpoint/both

## 256-bit wide LPDDR4x

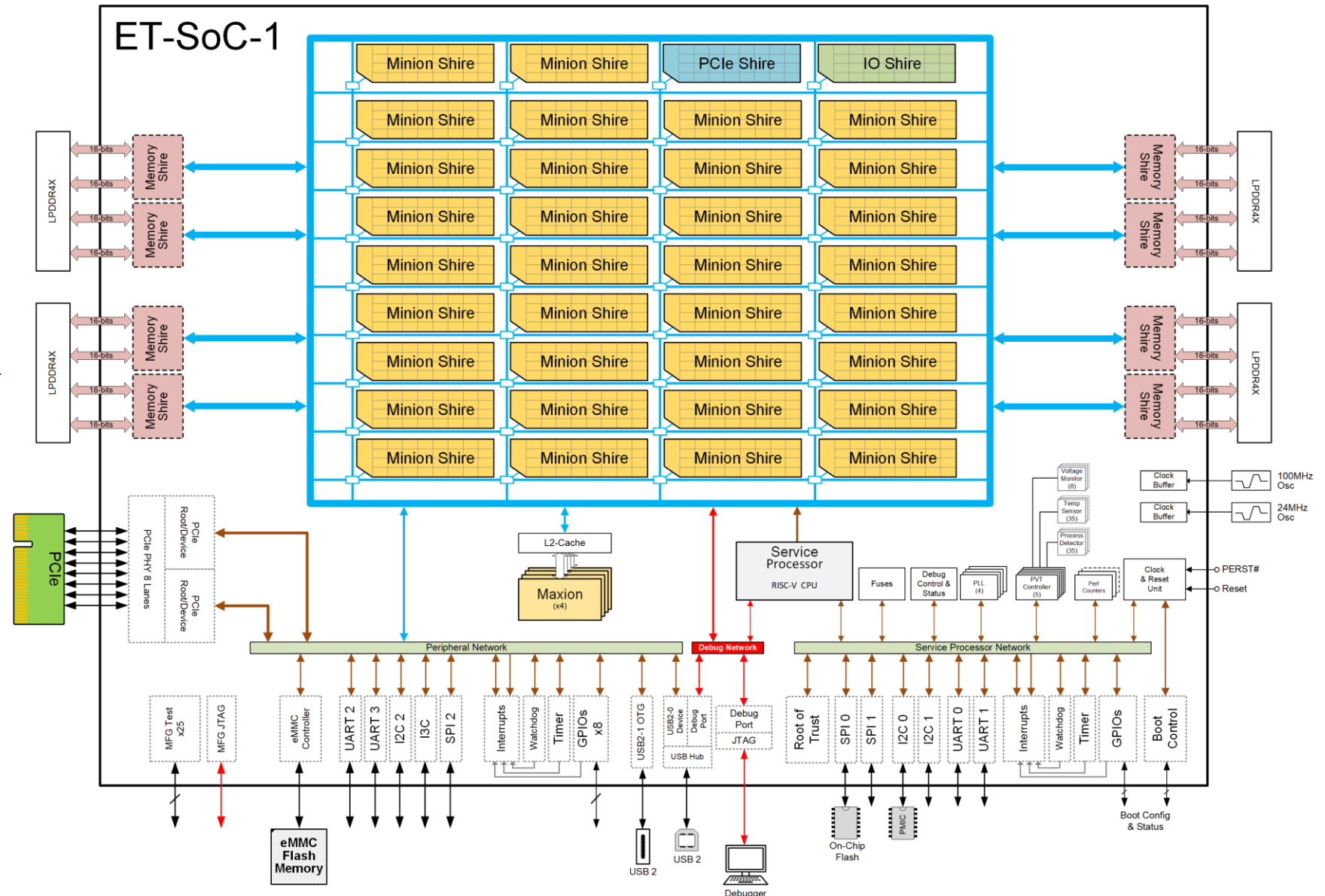
- 4267 MT/s
- 137 GB/s
- ECC support

## RISC-V SERVICE PROCESSOR

- Secure Boot
- System Management
- Watchdog timers
- eFuse

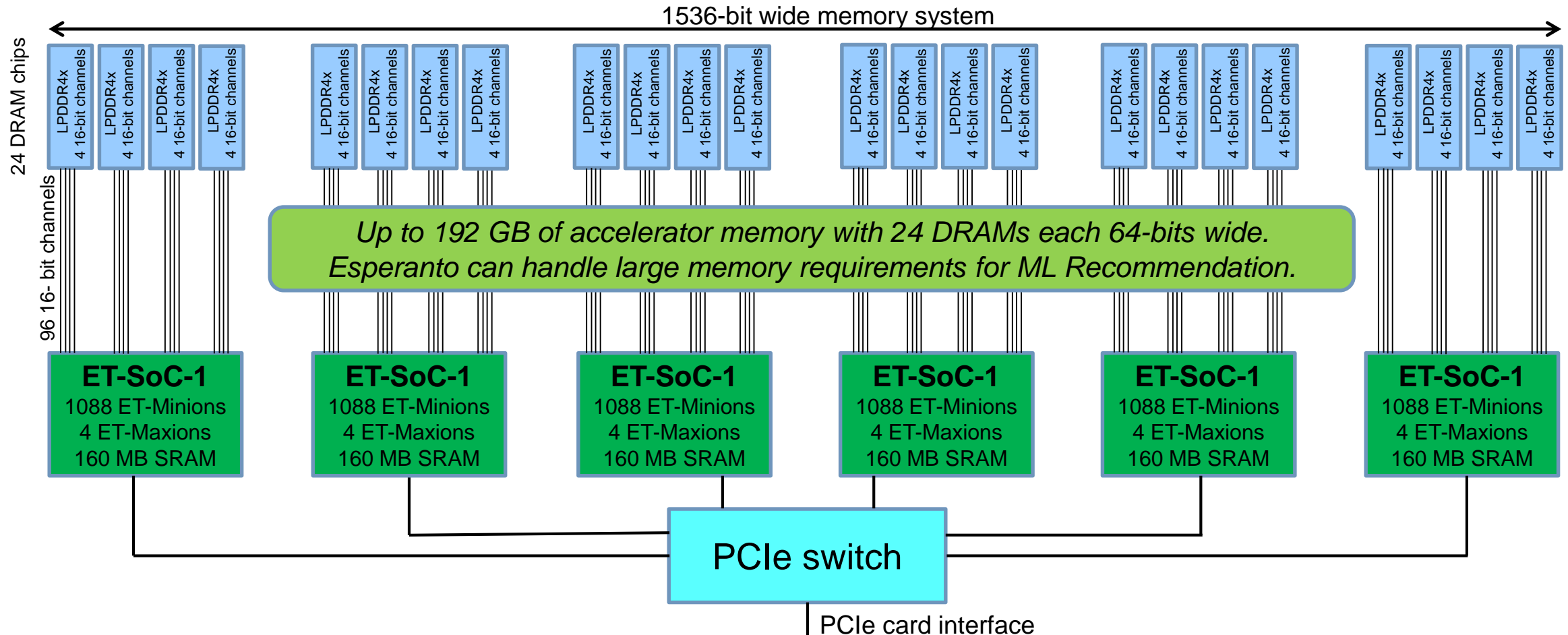
## EXTERNAL IO

- SMBus
- Serial – I2C/SPI/UART
- GPIO
- FLASH



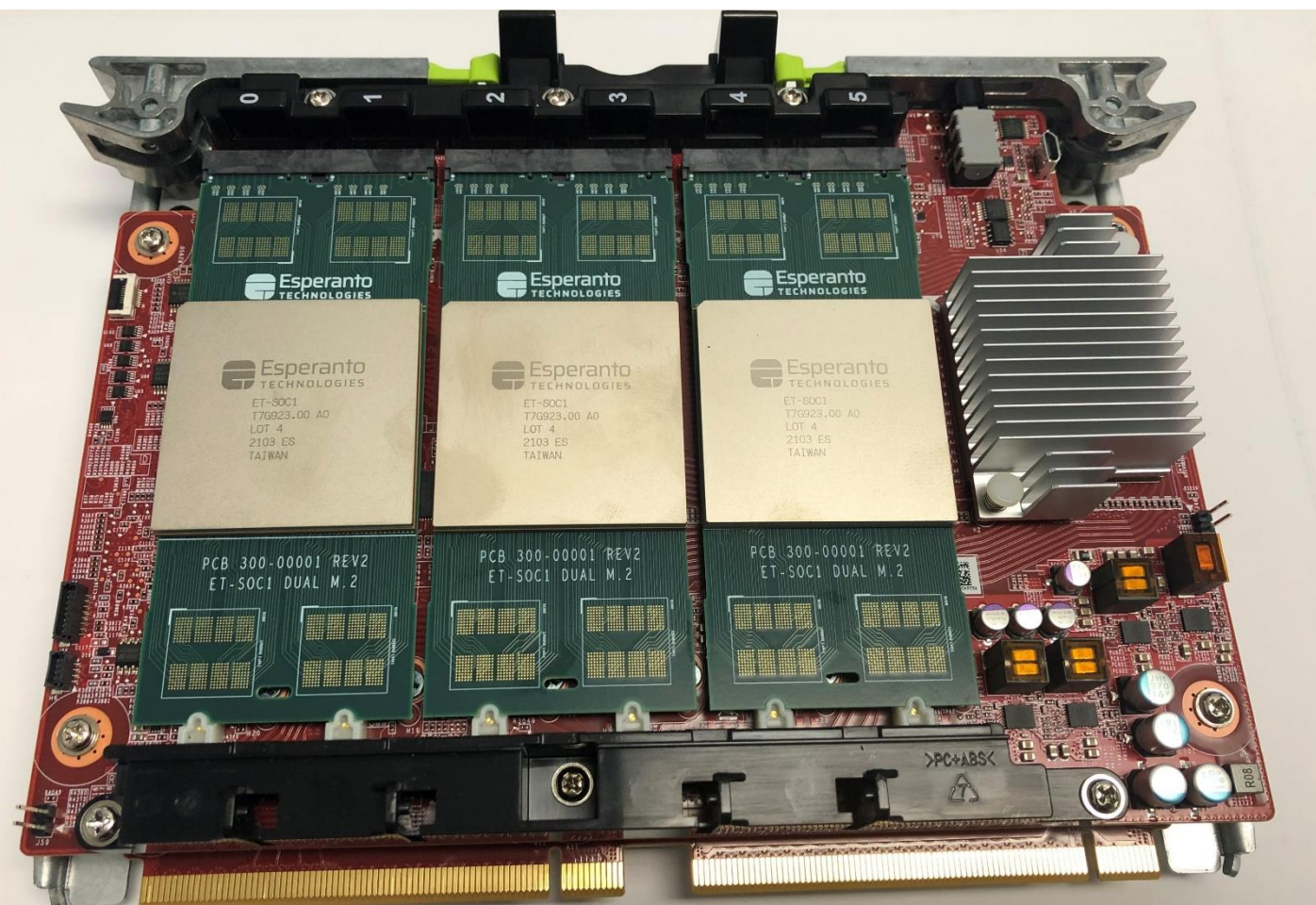
# Card with six ET-SOC-1 chips for large sparse ML Recommendation models

- Esperanto's low-power technology allows six Esperanto chips and 24 DRAM chips to fit into 120-Watt power budget of customer's PCIe card
- A single ML model on one accelerator card can use up to 192 GB of low-cost LPDDR4x DRAM with up to 822 GB/s of memory bandwidth
- Over 6K cores with 12K threads handles memory latency on 96 memory channels and performs well for ML Recommendation (and other) tasks





# Six ET-SoC-1 chips fit on an OCP Glacier Point v2 Card



## GLACIER POINT V2 CARD SHOWN:

- 6,558 RISC-V cores
- Up to 192 GB of DRAM
- Up to 822 GB/s DRAM bandwidth
- ~120 W maximum power consumption

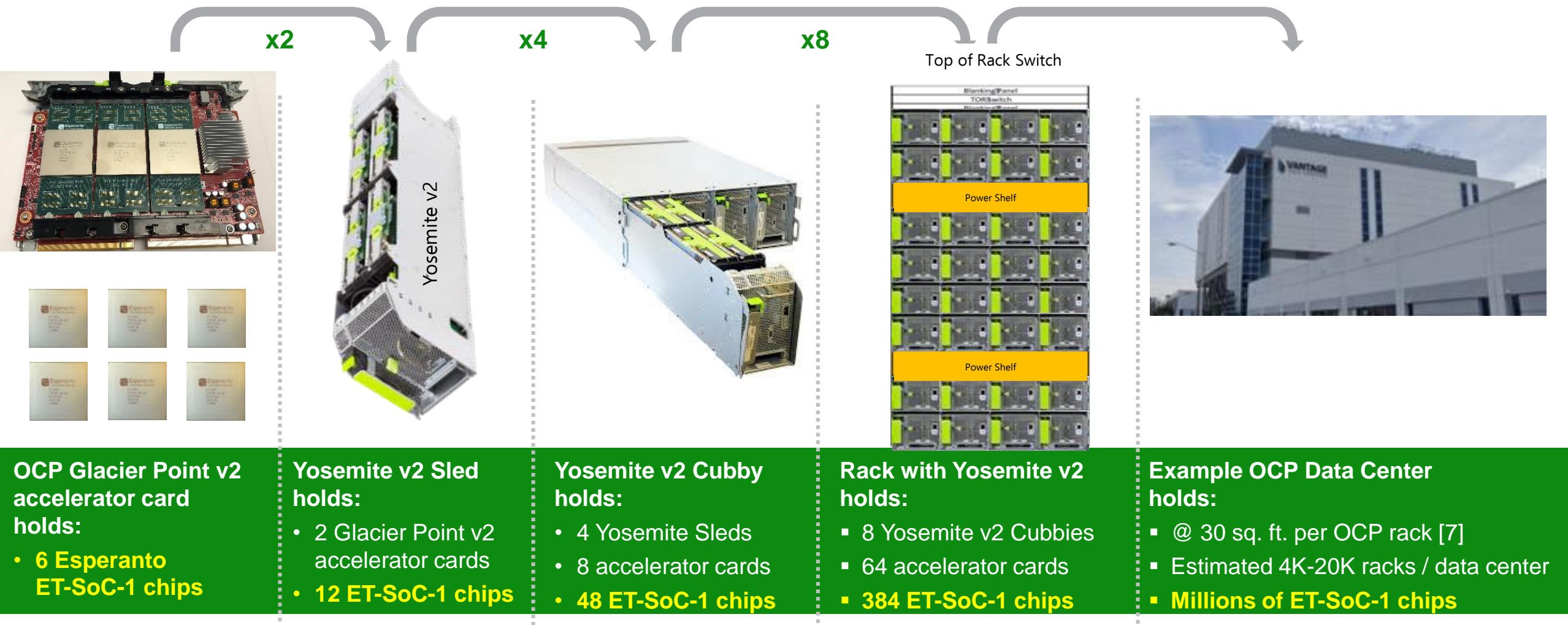
*Based on the open-source Glacier Point v2 board design from the Open Compute Project. Three Esperanto Dual M.2 modules can mount on the top side and three on the bottom side.*

## ON OTHER PCI EXPRESS CARDS:

- One ET-SoC-1 fits easily on a Low-Profile (half-height, half-length) PCIe add-in card
- ET-SoC-1 power budget increases to ~60W
- Many PCIe cards can be operated in parallel

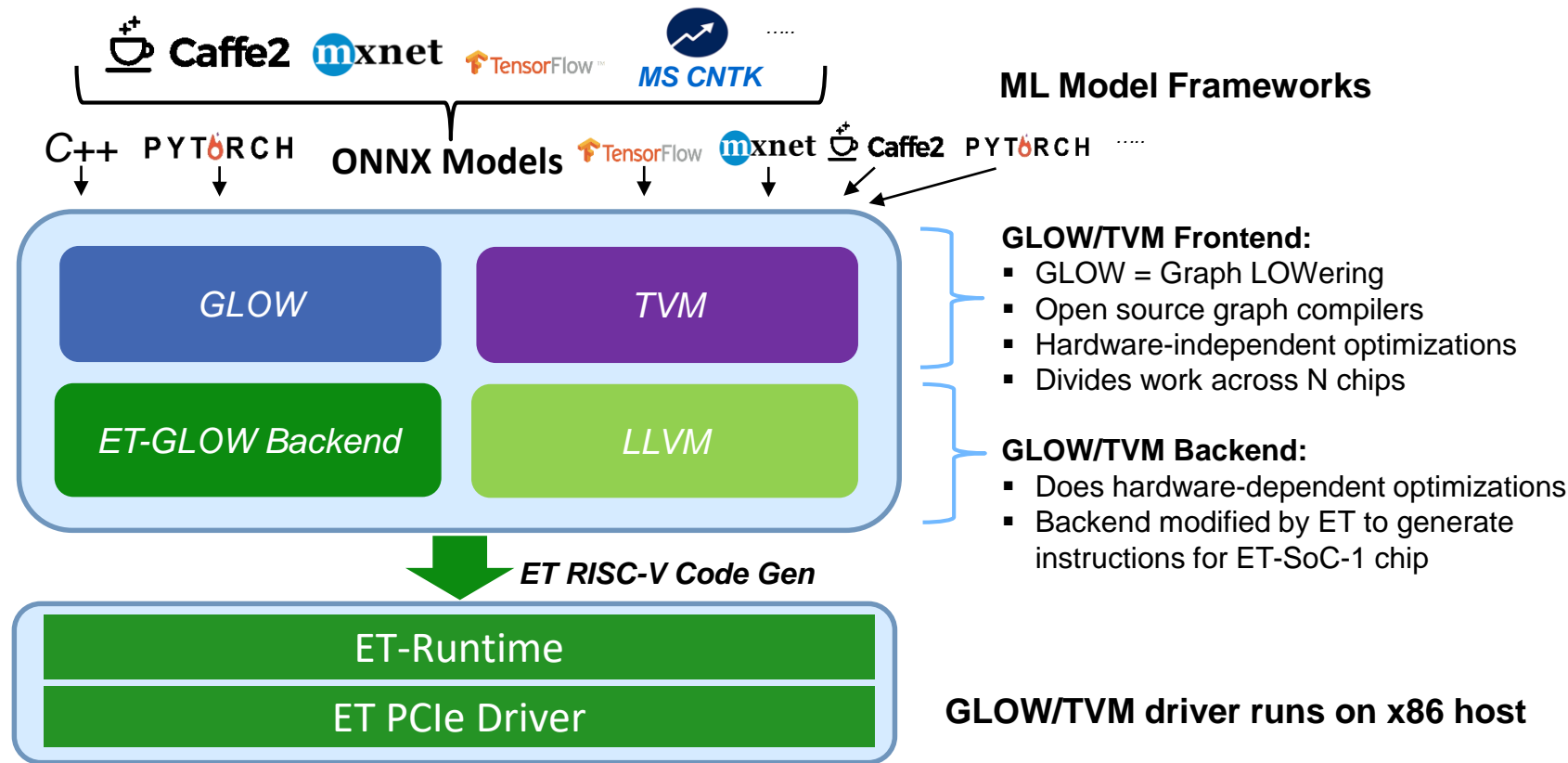
Peak performance > 800 TOPS<sub>8</sub> when all ET-Minions on six chips are operating at 1 GHz

# ET-SoC-1 can be deployed at scale in existing OCP Data Centers





# Software: Esperanto Supports C++ / PyTorch and Common ML Frameworks



## Development Tools

Console /  
Debugger

Performance  
Monitor

## Management Utilities

Diagnostics

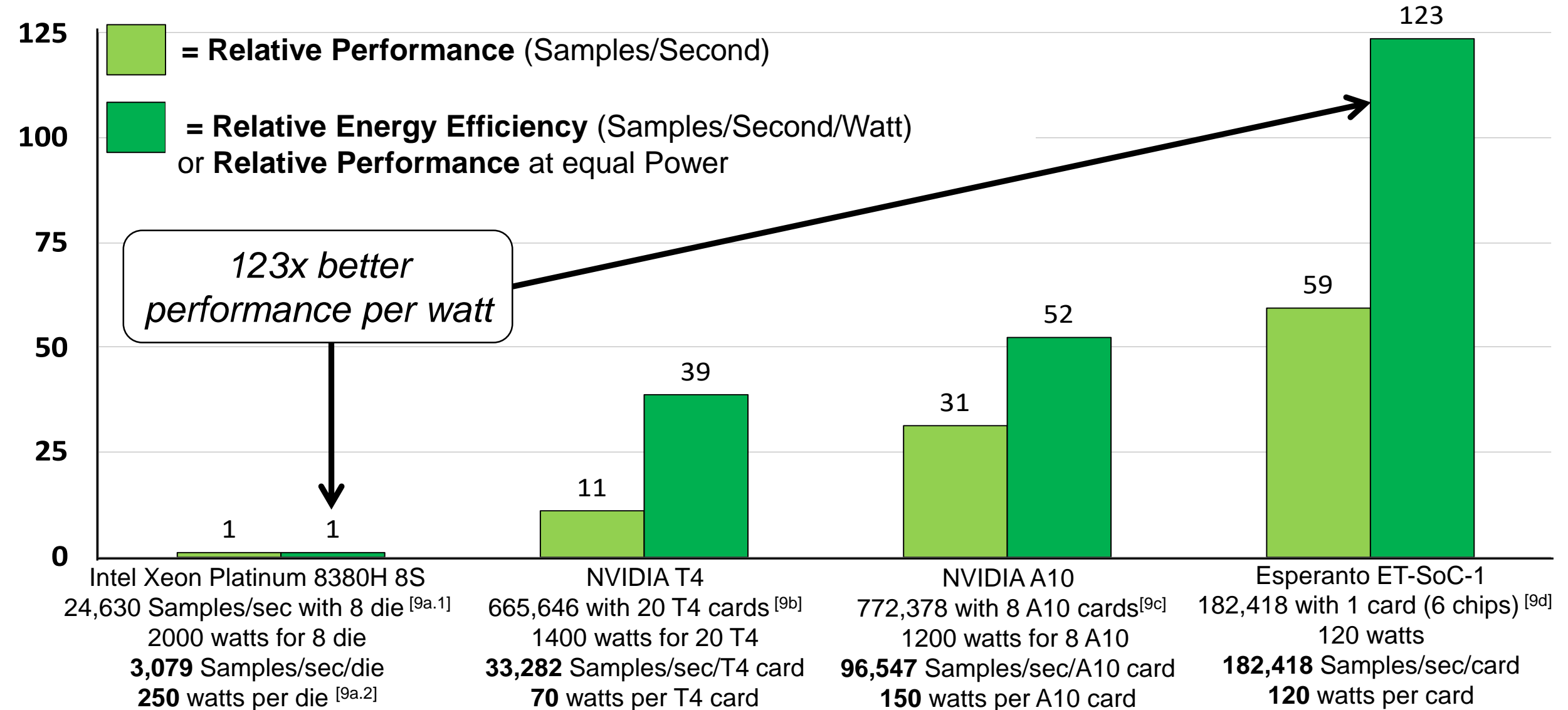
Firmware  
Updater



ML Models run across multiple  
ET-SoC-1 chips

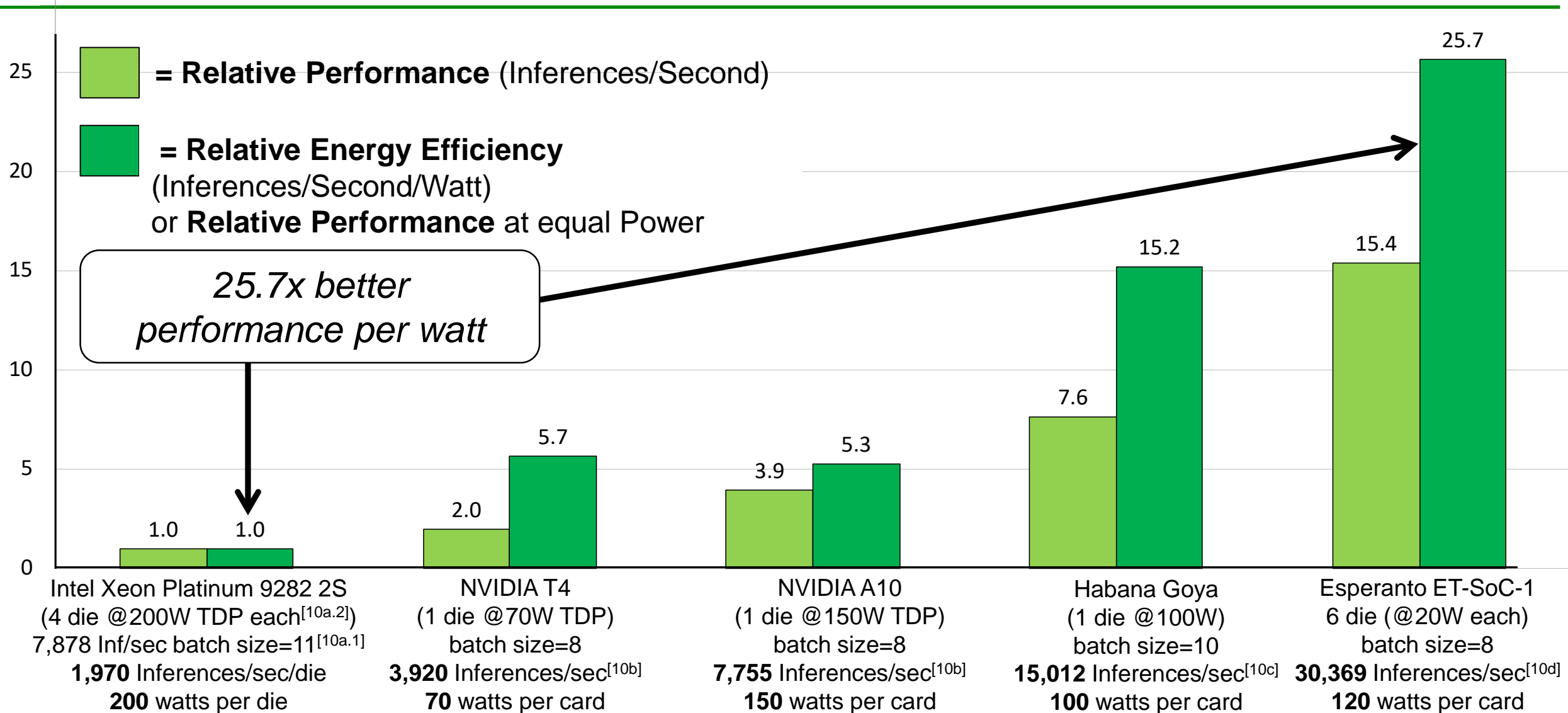
# ML Recommendation performance per card comparisons

Based on MLPerf Deep Learning Recommendation Model benchmark [8]



# Image Classification performance per card comparisons

Based on ResNet-50 benchmark <sup>[10]</sup>



ResNet-50 numbers taken from respective company websites, power numbers from data sheets.

Esperanto estimates

# Four ET-Maxions: High-Performance Out-of-Order RISC-V Processors

## FULL RISC-V RV64GC ISA SUPPORT

- Support for compressed ISA
- Privileged ISA
- Fully respects relaxed consistency model

## SUPPORTS 4 CORE SMP LINUX OS

- Allows chip to run in standalone configuration

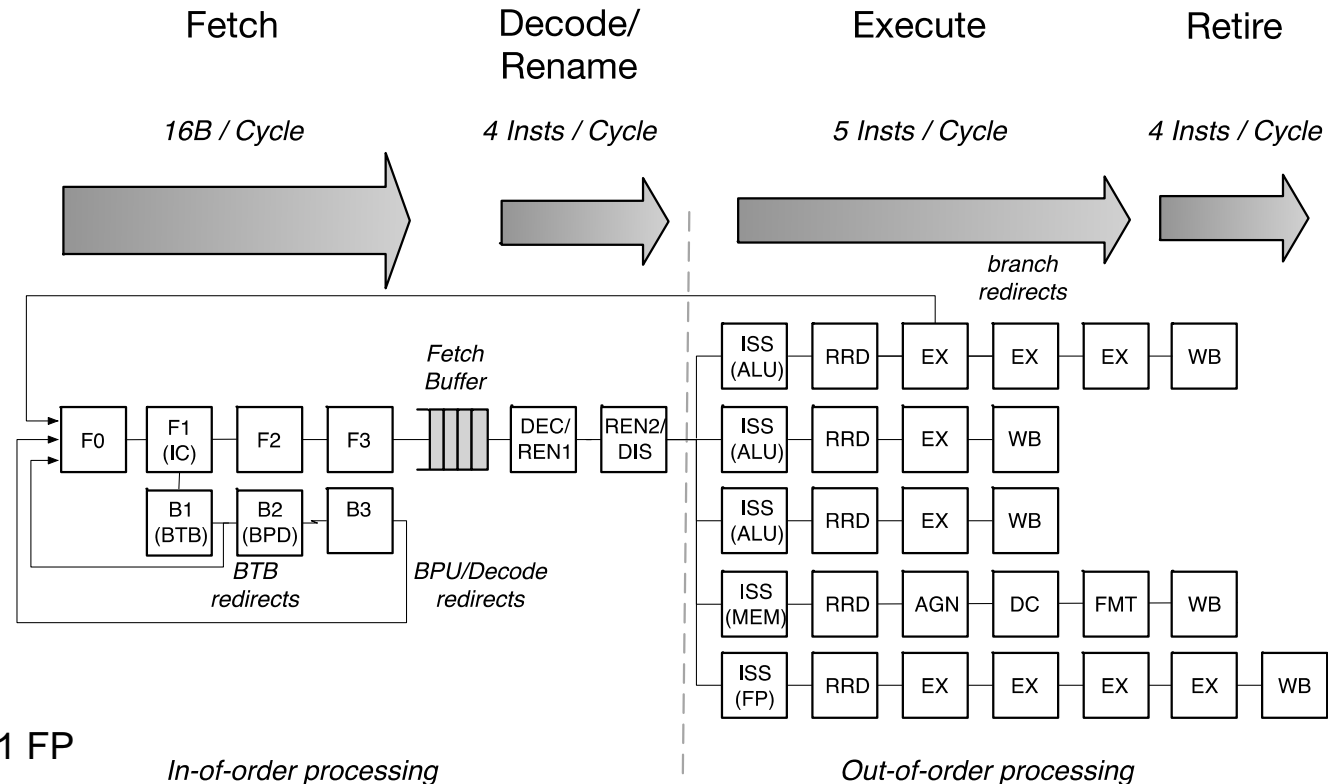
## OUT-OF-ORDER EXECUTION

- 64-entry distributed scheduler
- 128-entry ROB
- 32-entry load queue, 32-entry store queue
- 8R/4W 128-entry integer physical register file
- 3R/2W 64 floating-point physical register file
- Execution units:
  - 1 load/store, 2 simple ALU, 1 complex ALU/Branch, 1 FP

## MEMORY SUB-SYSTEM

- 32 entry data and instruction TLBs and 512 entry unified L2 TLB
- Fully coherent 64KB data cache backed by a unified 4MB L2 cache
- ECC for both data and L2 caches
- Aggressive stride prefetchers for L1 and L2 data caches

## OPERATING RANGE: 500 MHz to 2 GHz



*For details on ET-Maxion,  
see references [11,12]*



# Summary Statistics of ET-SoC-1

The ET-SoC-1 is fabricated in TSMC 7nm

- 24 billion transistors
- Die-area: 570 mm<sup>2</sup>
- 89 Mask Layers

1088 ET-Minion energy-efficient 64-bit RISC-V processors

- Each with an attached vector/tensor unit
- Typical operation 500 MHz to 1.5 GHz expected

4 ET-Maxion 64-bit high-performance RISC-V out-of-order processors

- Typical operation 500 MHz to 2 GHz expected

1 RISC-V service processor

Over 160 million bytes of on-die SRAM used for caches and scratchpad memory

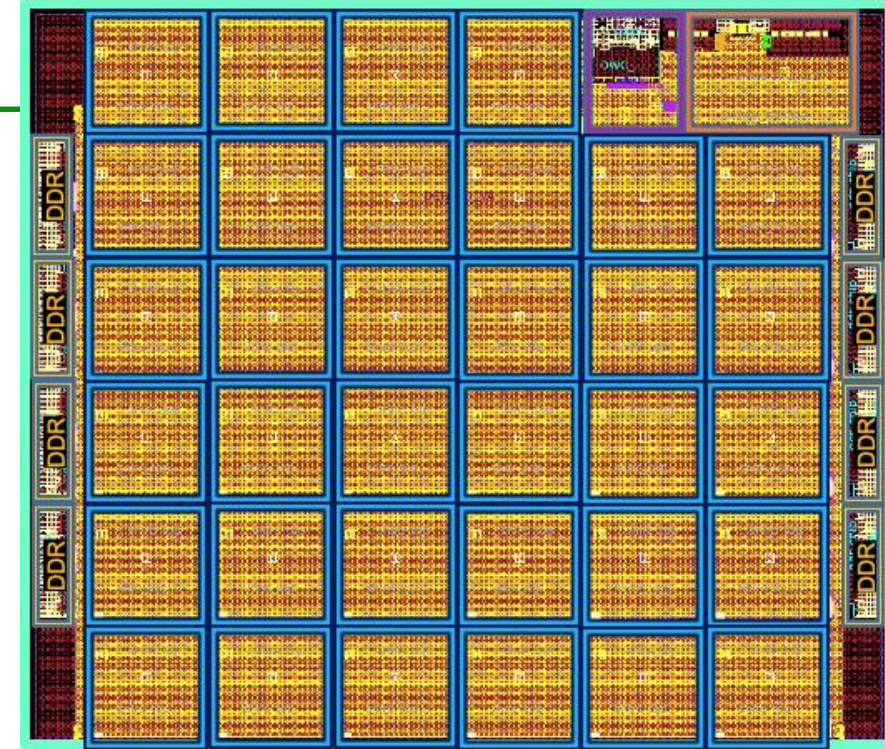
Root of trust for secure boot

Power typically < 20 watts, can be adjusted for 10 to 60+ watts under SW control

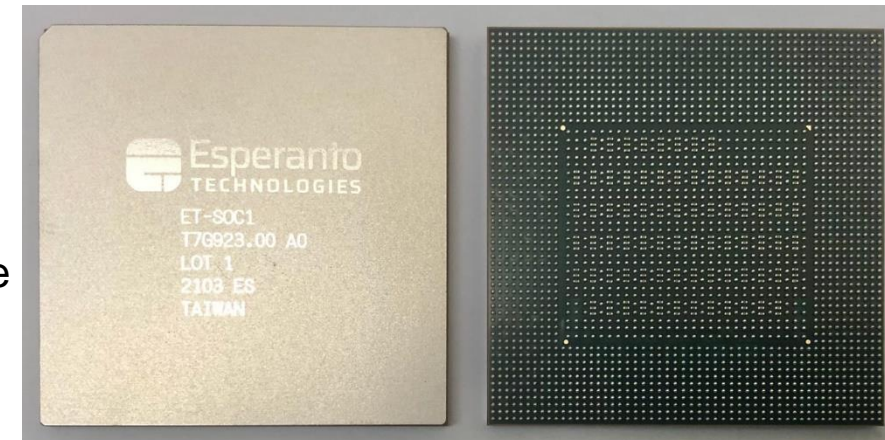
Package: 45x45mm with 2494 balls to PCB, over 30,000 bumps to die

- Each Minion Shire has independent low voltage power supply inputs that can be finely adjusted to mitigate  $V_t$  variation effects and enable DVFS

Status: Silicon currently undergoing bring-up and characterization



ET-SoC-1 Die Plot



ET-SoC-1 Package

# Summary

---

The Esperanto ET-SoC-1 is the highest performance commercial RISC-V chip to date

- More RISC-V cores on a single chip
- More RISC-V aggregate instructions per second on a single chip
- Highest TOPS driven by RISC-V cores

Esperanto's low-voltage technology provides differentiated RISC-V processors with the best performance per watt

- Energy efficiency matters!
- Best performance per watt delivers the best performance in a fixed number of watts
- Solution delivers energy efficient acceleration for datacenter inference workloads, especially recommendation

The hard part was making all the tradeoffs combining

- Processor and memory system architecture
- Circuits and techniques for low voltage operation

Esperanto now has a highly scalable design

- Efficient for ML recommendation
- Thousands of general-purpose RISC-V cores can be applied to many other highly parallel computing tasks
- Modular approach allows design to scale up and down, and to other semiconductor processes

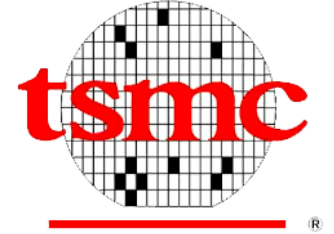
Early Access Program for qualified customers beginning later in 2021 (for info, contact: [chips@esperanto.ai](mailto:chips@esperanto.ai))

# Thanks to our Key Development Partners

---

SYNOPSYS®

R RISC-V®



Mentor®  
A Siemens Business



MOVELLUS semidynamic<sup>s</sup>  
silicon design and verification services



Thanks to all our partners for their help in bringing our vision into reality! Sorry we can't name everyone!



# Footnotes and References

- [1] N. Jouppi, et al., Ten Lessons from Three Generations Shaped Google's TPUv4i, 2021 ACM/IEEE 48<sup>th</sup> Annual International Symposium on Computer Architecture. Page 4 lesson 5 concludes "Inference DSAs need air cooling for global scale".
- [2] J. Park, et al., Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications, arXiv:1811.09886v2, 29 November 2018.
- [3] M. Anderson, et al., First Generation Inference Accelerator Deployment at Facebook, arXiv: 2107.04140v1, 8 Jul 2021
- [4] M. Smelyanskiy, Facebook, Linley Fall Processor Conference 2019 "Challenges and Opportunities of Architecting AI Systems at Datacenter Scale"
- [5] M. Smelyanskiy, AI System Co-Design: How to Balance Performance & Flexibility, AI Hardware Summit, September 17, 2019. Slide 12 indicated desired Inference characteristics with 100 TOPs of INT8, 100 MB of SRAM. Slide 19 talks about the need for programmability over fixed function hardware.
- [6] Note that a core optimized for high voltage and high frequency (2-3GHz) operation will require higher power gate drive strengths to overcome wire delays and hence will have higher C<sub>dyn</sub> than a processor optimized for low-voltage operation. Each of the power/frequency points shown on this energy efficiency curve therefore represents a different physical design, i.e. not the same silicon, to take this changing C<sub>dyn</sub> into account. Designs were synthesized at high and low voltages to estimate potentially achievable frequencies. Performance at each frequency was estimated using our internal ML Recommendation benchmark, based on running this benchmark on a full chip hardware emulation engine (Synopsys Zebu system) providing clock level accuracy at a few points and interpolating the other points. The goal was to understand the shape of the energy efficiency curve to find voltages for best energy efficiency (Inferences/second/watt). Different benchmarks would likely have different curves, though we would expect the overall shape to be similar. Repeating, this was a design study and does not represent any specific silicon results or design, each point on the curve is a differently synthesized design, though with the same architecture, i.e., we used the full ET-Minion as the input to be synthesized.
- [7] Estimate of 30 square feet per rack comes from "The Case for the Infinite Data Center" – Gartner, Source: Gartner, Data Center Frontier
- [8] MLPerf DLRM Inference Data Center v0.7 & v1.0: <https://mlcommons.org/en/>
- [9] Measured by MLPerf DLRM Samples / Second; FP32, Offline scores
- Additional source information:
- a.1. Submitter: Intel; MLPerf DLRM score 24,630: Inference Data Center v0.7 ID 0.7-126; Hardware used (1-node-8S-CPX-PyTorch-BF16); BF16; <https://mlcommons.org/en/inference-datacenter-07/>
  - a.2 Intel 8380H Processor TDP Power of 250W from: <https://ark.intel.com/content/www/us/en/ark/products/204087/intel-xeon-platinum-8380h-processor-38-5m-cache-2-90-ghz.html>
  - b. Submitter: NVIDIA; T4 MLPerf DLRM score 665,646: Inference Data Center v0.7 ID 0.7-115; Hardware used (Supermicro 6049GP-TRT-OTO-29 (20x T4, TensorRT)); INT8; <https://mlcommons.org/en/inference-datacenter-07/>
  - c. Submitter: NVIDIA; A10 MLPerf DLRM score 772,378: Inference Data Center v1.0 ID 1.0-54; Hardware used (Supermicro 4029GP-TRT-OTO-28 (8x A10, TensorRT)); INT8; <https://mlcommons.org/en/inference-datacenter-10/>
  - d. Internal estimates by Esperanto for MLPerf DLRM: Inference Data Center v0.7; ET-SOC-1; Unverified result is from Emulated/Simulated pre-silicon projections; INT8; Result not verified by MLCommons™ Association.
- [10] Measured by ResNet-50 Images per second (Esperanto INT8 Batch 8, NVIDIA INT8 Batch 8, Habana INT8 Batch 10, Intel INT8 Batch 11)
- Additional measurement source information:
- a.1. Intel ResNet-50: <https://software.intel.com/content/www/us/en/develop/articles/intel-cpu-outperforms-nvidia-gpu-on-resnet-50-deep-learning-inference.html>
  - a.2. Intel 9282 has 2 die in the package, CPU TDP power for both die from: <https://ark.intel.com/content/www/us/en/ark/products/194146/intel-xeon-platinum-9282-processor-77m-cache-2-60-ghz.html>
  - b. NVIDIA (T4, A10) ResNet-50: <https://developer.nvidia.com/deep-learning-performance-training-inference>
  - c. Habana ResNet-50: <https://habana.ai/wp-content/uploads/2018/09/Goya-Datasheet-HL-10x-Nov14-2018.pdf>
  - d. Esperanto ResNet-50: Emulated/Simulated projections; INT8
- [11] P. Xekalakis and C. Celio, The Esperanto ET-Maxion High Performance Out-of-Order RISC-V Processor, 2018 RISC-V Summit, presentation at [https://www.youtube.com/watch?v=NjEsIX\\_-t0Q](https://www.youtube.com/watch?v=NjEsIX_-t0Q)
- [12] Maxion is described in "Esperanto Maxes out RISC-V - High-End Maxion CPU Raises RISC-V Performance Bar", Microprocessor Report, December 10, 2018.