

Thousands of RISC-V Cores for AI and Beyond

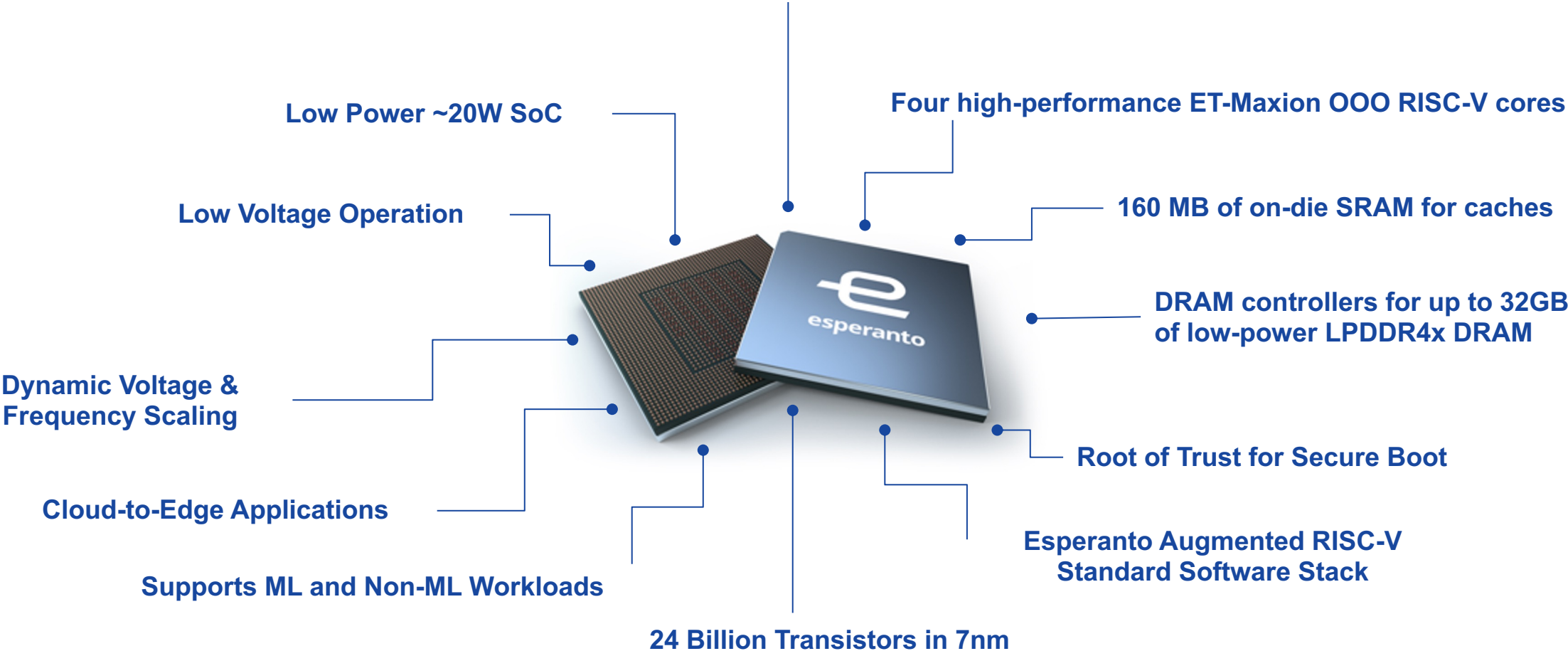


2022 RISC-V Summit Presentation

**Dave Ditzel (dave@esperanto.ai)
Founder and CTO
Esperanto Technologies
December 13, 2022**

How many
high-performance 64-bit RISC-V processors
could be put in a standard computer rack?

1,088 ET-Minion 64-bit RISC-V vector/tensor CPUs



32 RISC-V ET-Minion CPUs and 4 MB Memory Form a “Minion Shire”

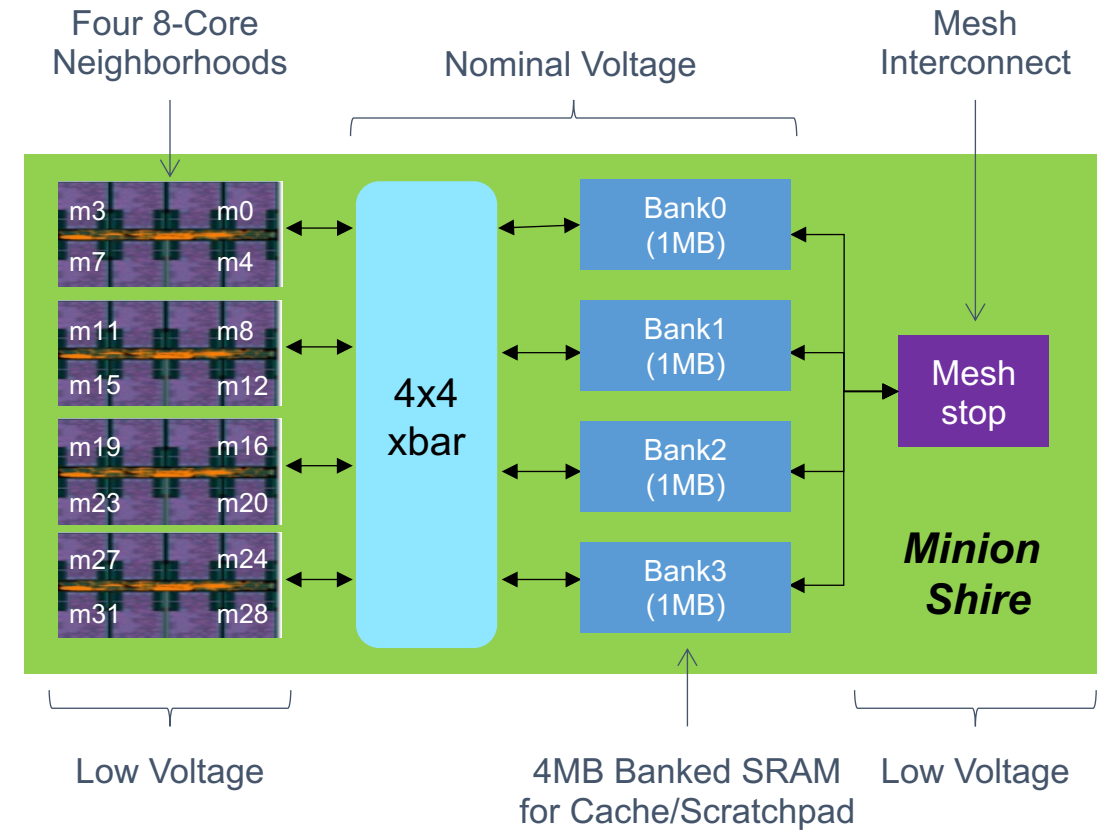
32 ET-Minion RISC-V cores per Minion Shire

- 64-bit RISC-V integer CPU cores each with 512-bit vector/tensor units
- Arranged in four 8-core neighborhoods

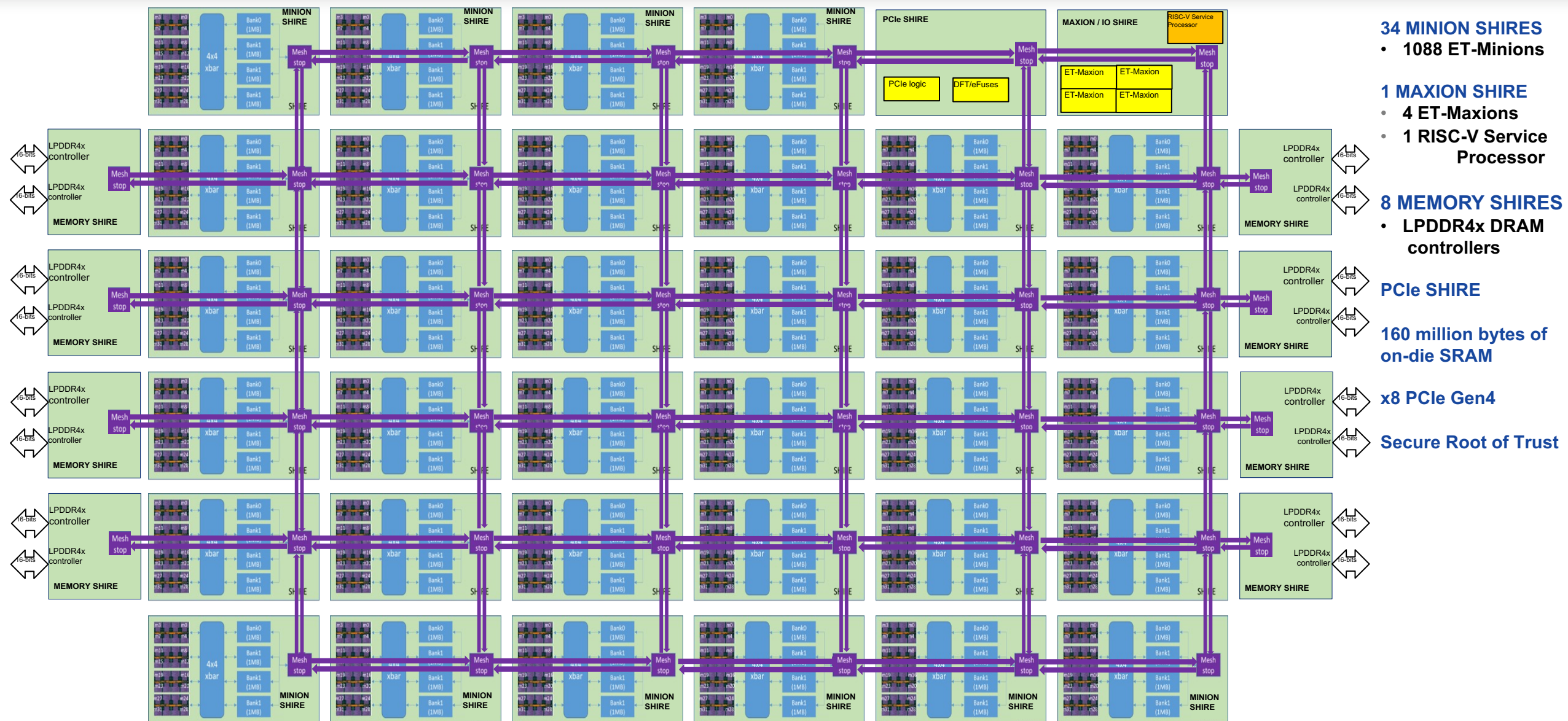
Software configurable memory hierarchy

- 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-on-chip



Block Diagram of ET-SoC-1: Over a thousand (1093) RISC-V processors on a single 7nm chip



Summary Statistics of ET-SoC-1

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

- 24 billion transistors
- Die-area: 570 mm²
- 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 GHz

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

- Typical operation 500 MHz to 1.5 GHz

1 64-bit RISC-V service processor

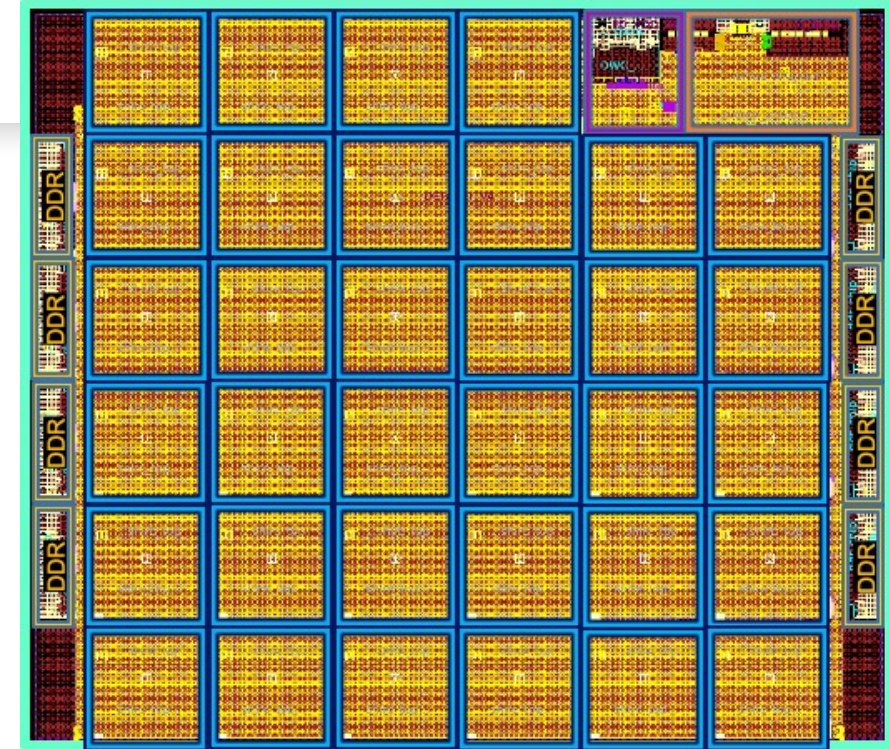
Over 160 million bytes of on-die SRAM used for caches and scratchpad memory

Root of trust for secure boot

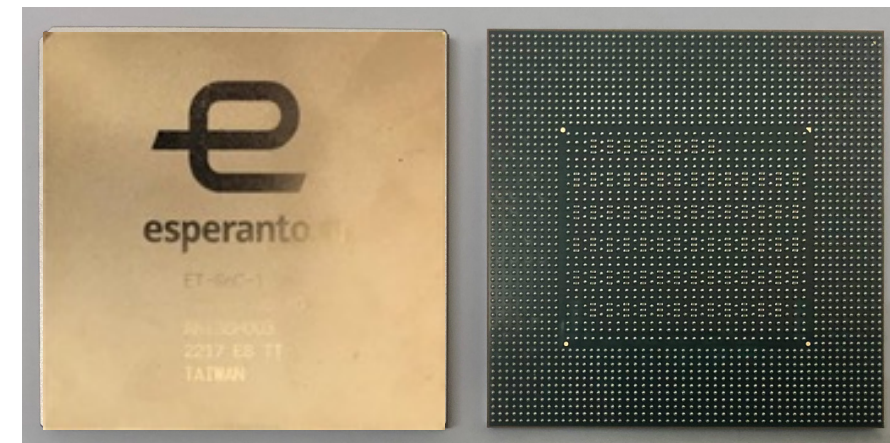
Power typically ~20 watts, can be adjusted for 10 to 60 watts under SW control

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

Status: Shipping to customers



ET-SoC-1 Die Plot



ET-SoC-1 Package

ET-SoC-1 External Chip Interfaces

8-bit PCIe Gen4

- Root/endpoint/both

256-bit wide LPDDR4x

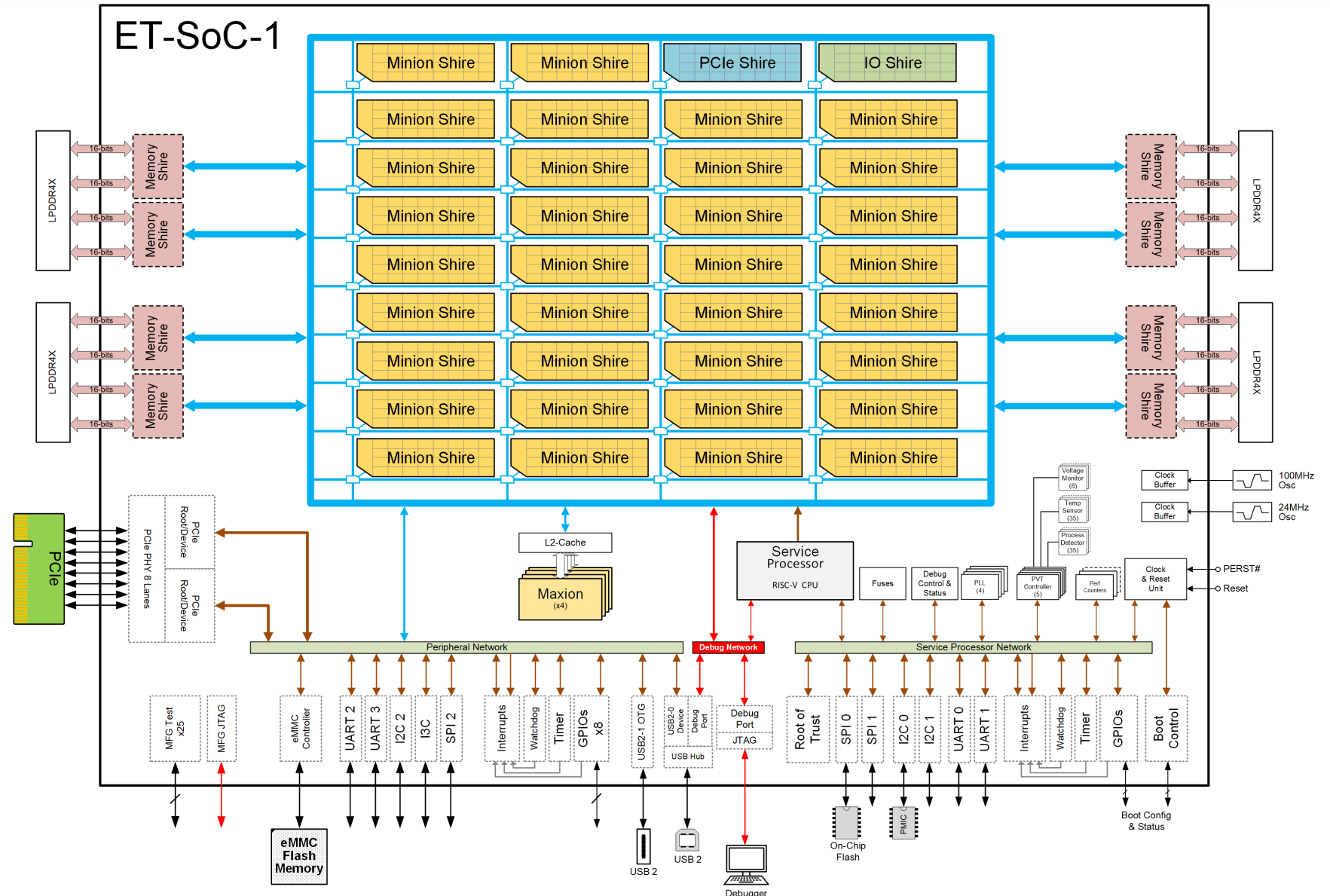
- 32 GB
- 137 GB/s
- ECC support

RISC-V SERVICE PROCESSOR

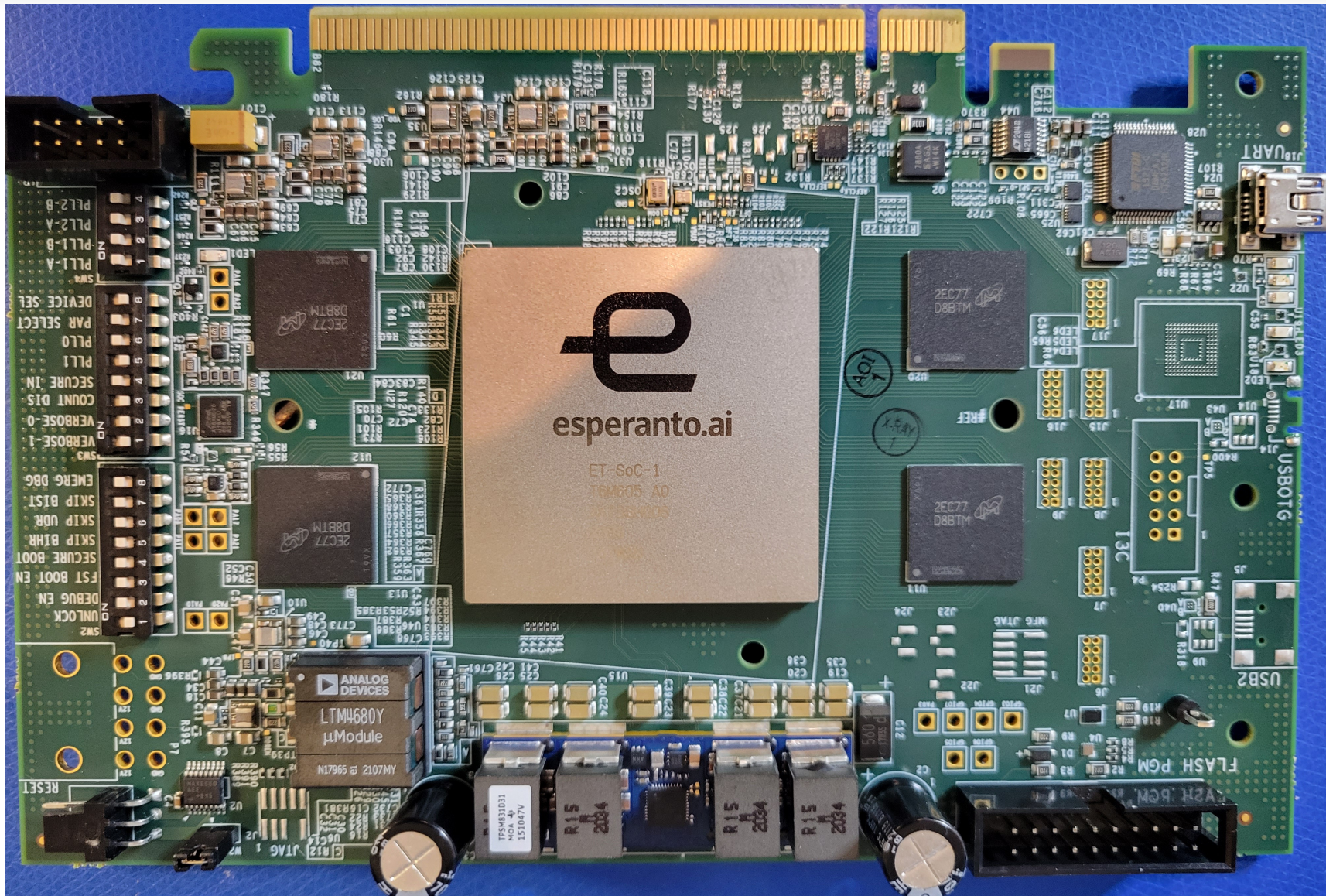
- Secure Boot
- System Management
- Watchdog timers
- eFuse

EXTERNAL IO

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



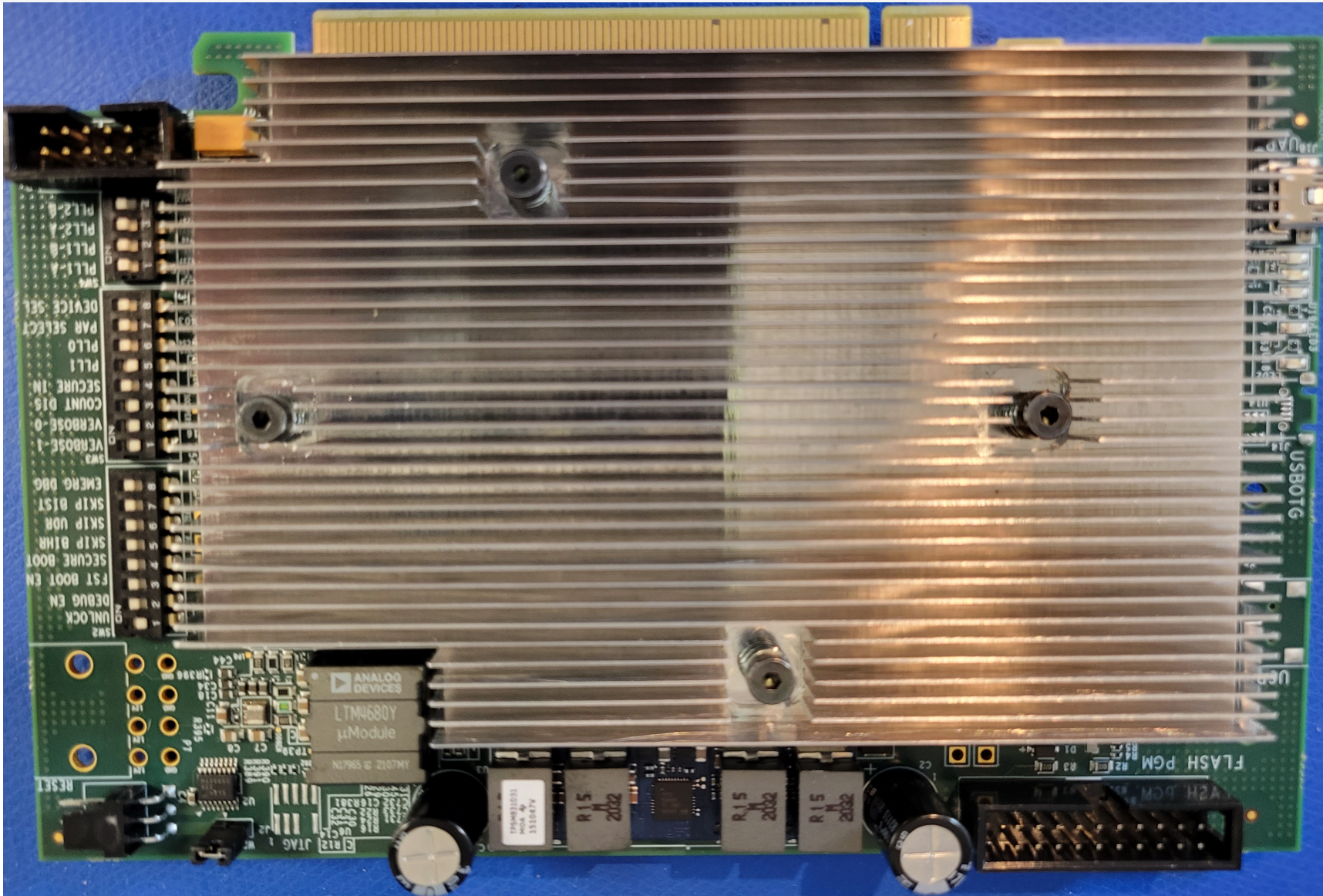
Over 1,000 RISC-V processors and 32 GB DRAM on a PCIe card



Esperanto PCIe evaluation card

- 1 ET-SoC-1
 - 1093 RISC-V processors
- 32 GB LPDDR4x DRAM
- 8 lanes of Gen4 PCIe bus
- Power supplies
- Typically, under 40 watts

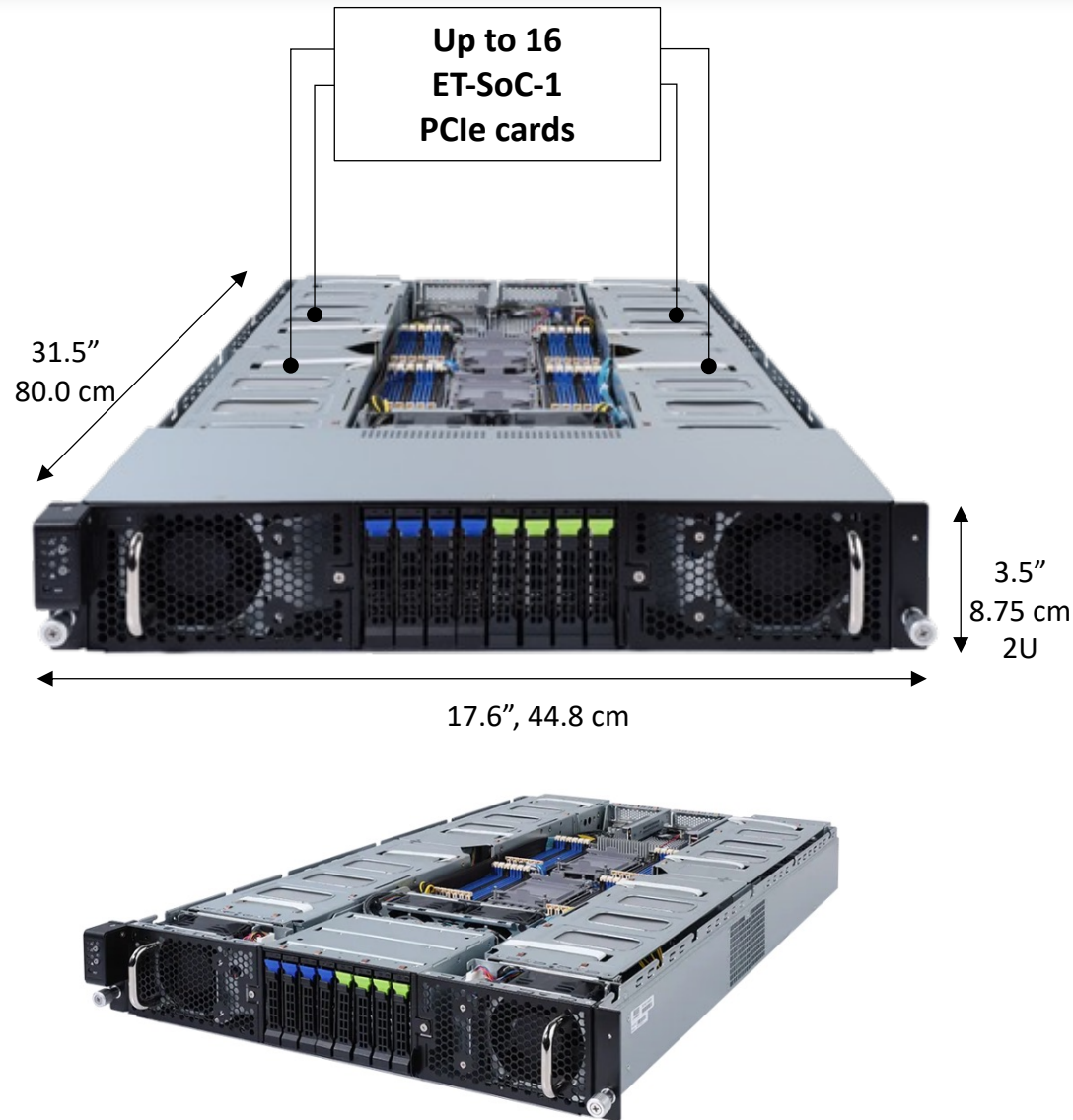
Over 1,000 RISC-V processors with simple air cooling



Esperanto PCIe evaluation card

- Simple heat sink for servers
- Only takes 1 PCIe slot
- Card power typically < 40 watts
- Power depends on application and MHz of ET-Minions

Can put up to 16 Esperanto PCIe cards in a standard 2U server

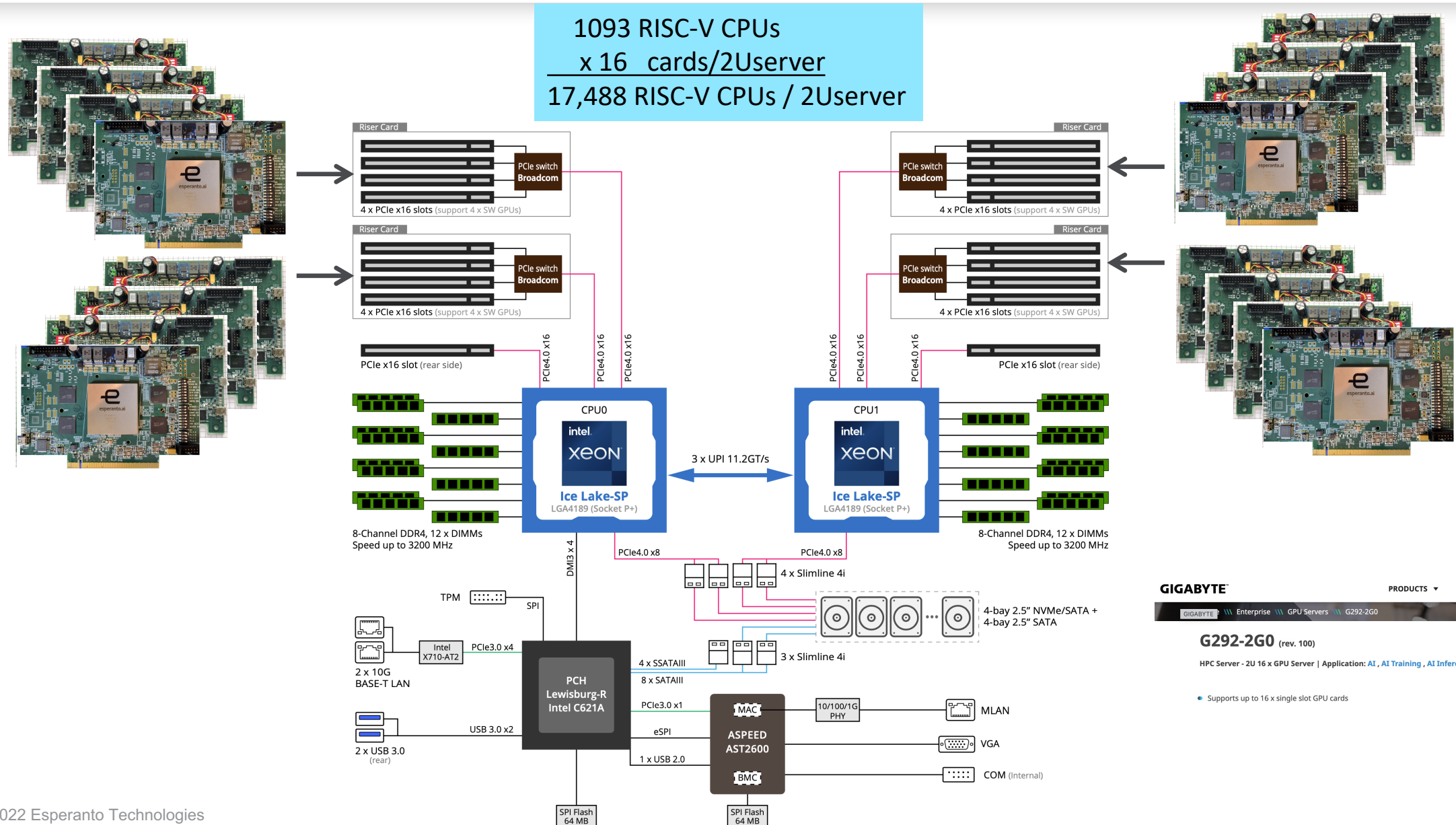


Target environment	Data center	
Server model	GIGABYTE G292-2G0	
Server configuration	Standard 2U 19" rack-mount chassis	
ET-SoC-1 PCIe cards ¹	8 Cards	16 Cards
System host processor	2x Intel Xeon Gold 6326 16-core	2x Intel Xeon Platinum 8358P 32-core
System memory ²	512GB DDR4-3200	1TB DDR4-3200
Storage	2x Samsung PM9A3 3.84TB NVMe U.2 SSDs	
ET-SoC-1 performance	600-800 MHz typical ET-Minion speed	
ET-SoC-1 power consumption	10W to 40W (workload and MHz dependent)	
ET-SoC-1 RISC-V CPUs	17,488 RISC-V processors (with 16 ET PCIe cards)	
ML software	Jupyter Notebook and command-line tools	
Performance, power, and trace analysis tools	Included (ET-Powertop, Perfetto)	
Training and documentation	Included	
Connectivity	2 x 10Gb/s LAN ports built in, 2 free PCIe Gen4 x8 slots	

¹ PCIe Gen 4, x8 lanes, card size FHHL

² Extendable up to 6TB

16 Esperanto PCIe Cards fit into each 2U high server chassis



Three Hundred Fifty Thousand RISC-V Processors in a Rack

Standard rack can be 42 U high. (73.5" / 187 cm)

- Room for 20 2U servers
- 2U leftover for top of rack switch

Twenty 2U servers can hold

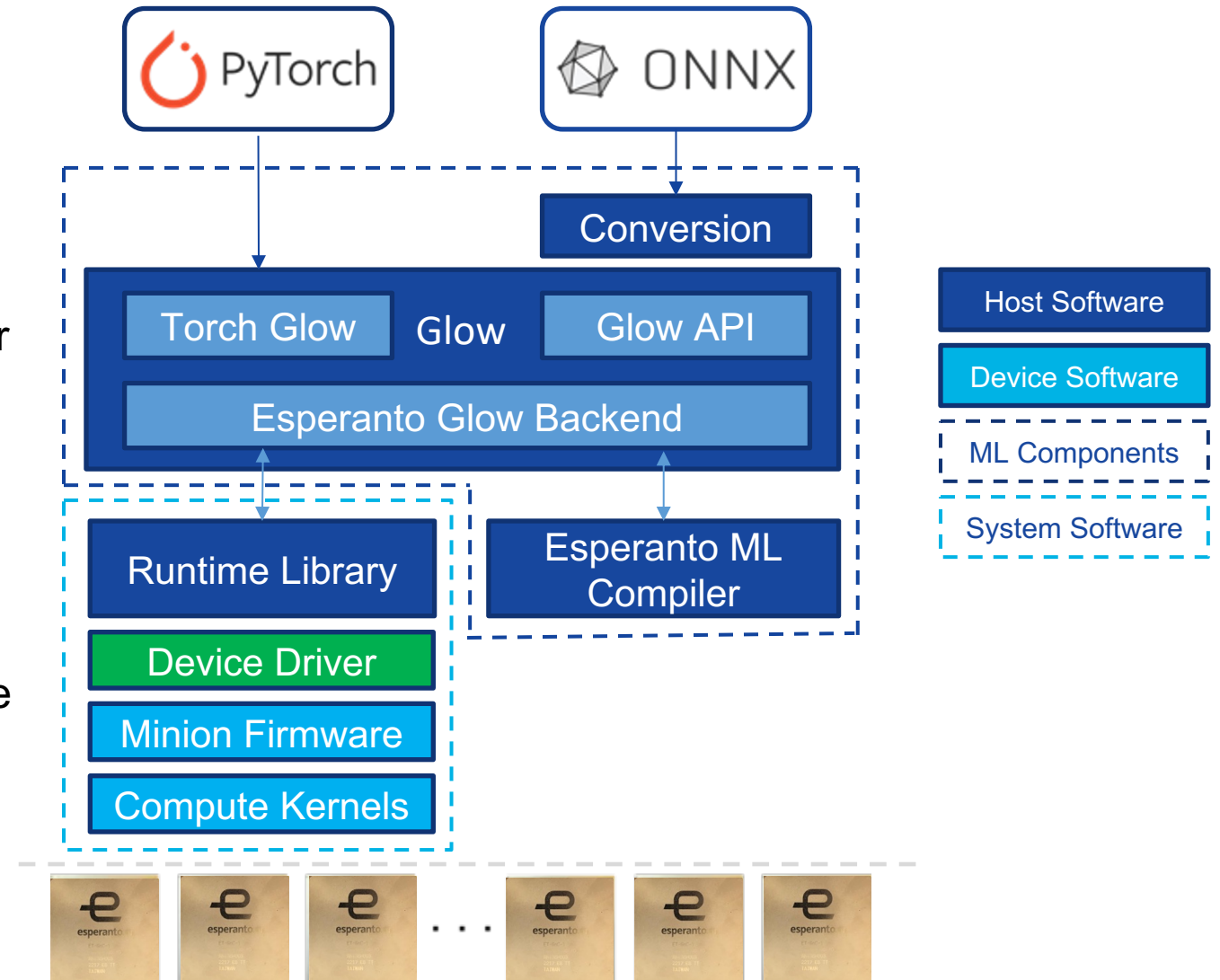
- 320 PCIe cards = 20 x 16 cards / server
- 20 x 1093 RISC-V CPU/card x 16 cards = 349,760 RISC-V Processors per rack
- ~350K RISC-V processors per rack

Peak Performance:

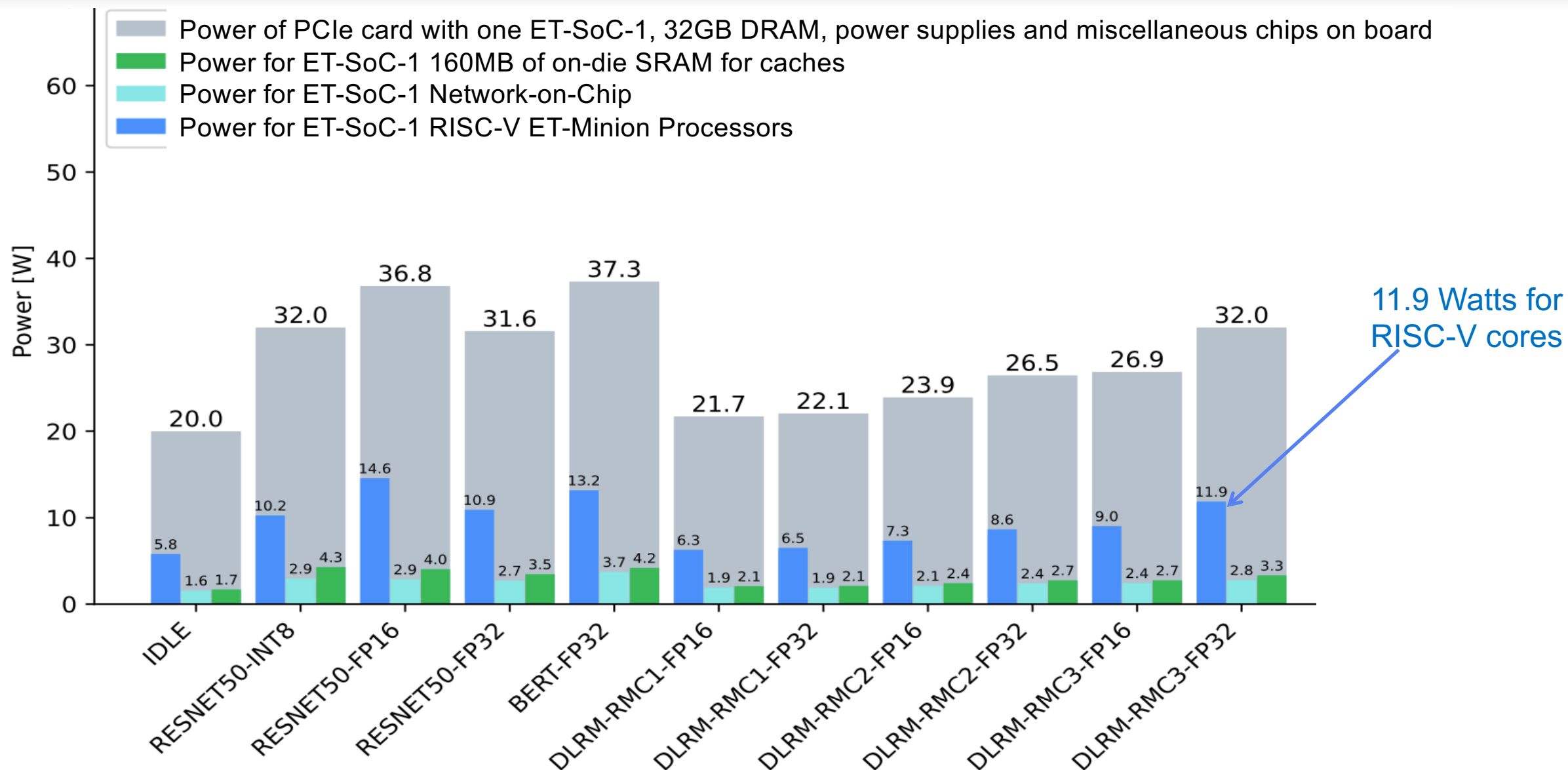
- ET-SoC-1 chips with 600 MHz ET-Minion CPU's
- 84 Int8 TeraOps per card @600MHz
- 1,344 Int8 TeraOps per 2U server
- 26,880 Int8 TeraOps per rack



- The AI/ML framework compiler accepts a model expressed at a high level – PyTorch or ONNX – and generates RISC-V executable code to run on ET-SoC-1
- The ML compiler stack is built on Glow (Graph Lowering), which is an open-source compiler for AI/ML models, as well as an execution engine for hardware accelerators
- Esperanto has also demonstrated mixed mode operation combining ML and non-ML workloads
- General Purpose Software Development Kit enables direct RISC-V programming of compute fabric for non-ML applications

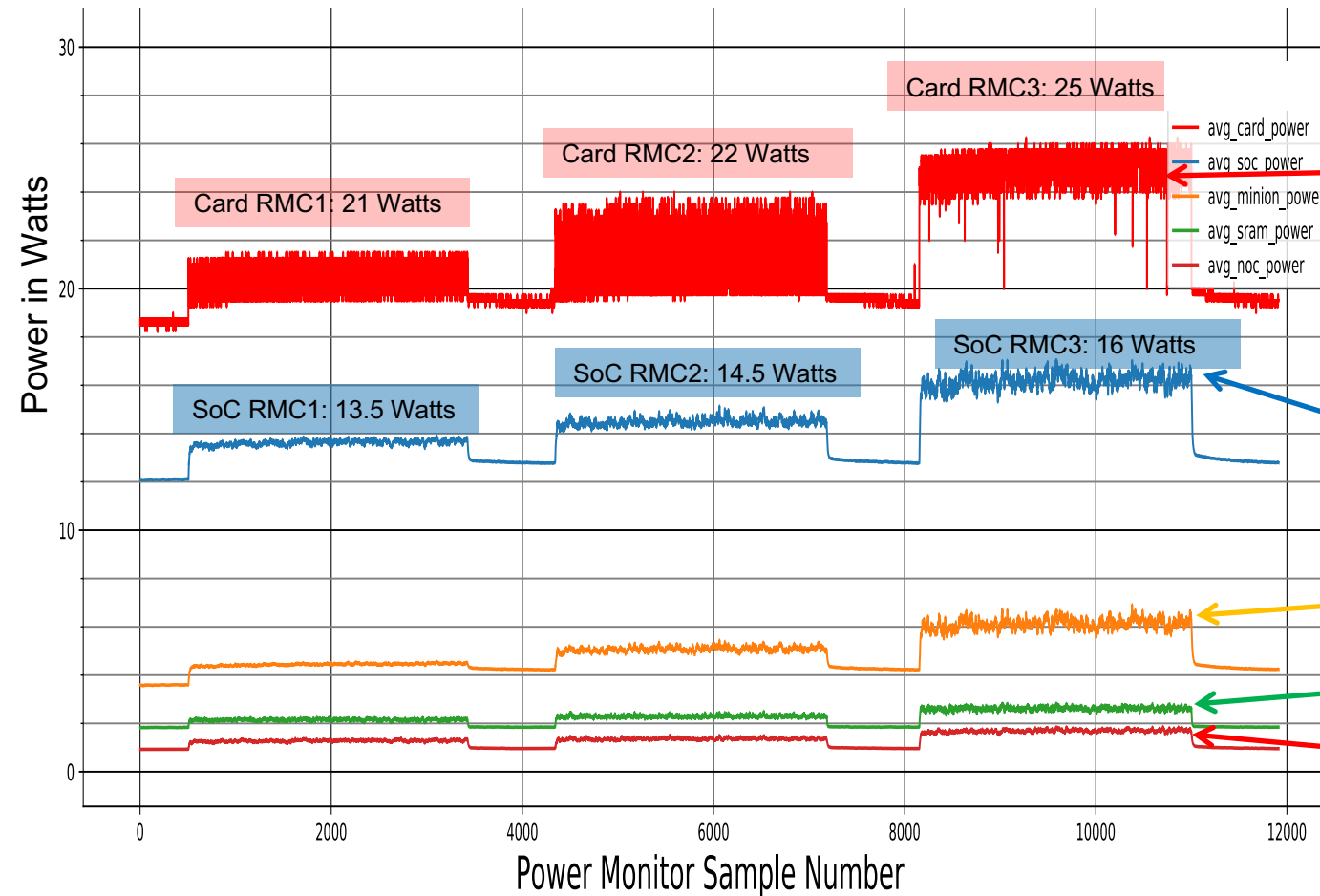


PCIe Card Power typically under 40 Watts



Example Card and ET-SoC-1 Power on ML Recommendation Benchmarks

DLRM FP16 RMC1, RMC2, RMC3



Power will vary depending on benchmark and operating conditions, but here is recently measured data.

Total PCIe card power under 30 watts for DLRM

- ET-SoC-1
- Eight LPDDR4x 32-bit wide DRAM die
- Flash memory and other misc components
- Power supplies

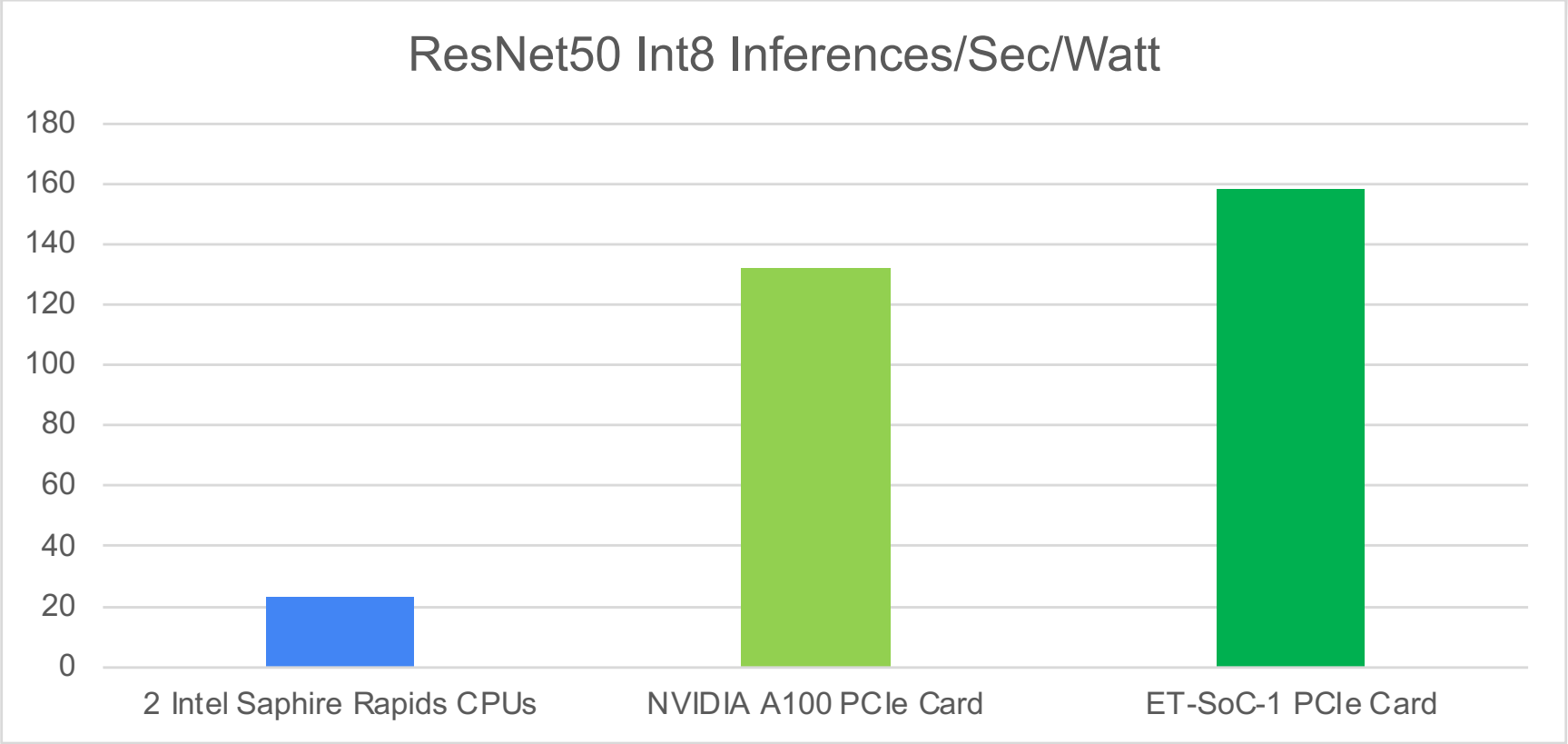
ET-SoC-1 power about 13.5 to 16 watts

A thousand 600 MHz ET-Minion RISC-V processors: 6 watts

128 MB of on-die SRAM for caches: 2.8 watts

44 Network-on-Chip connecting compute shires: 1.8 watts

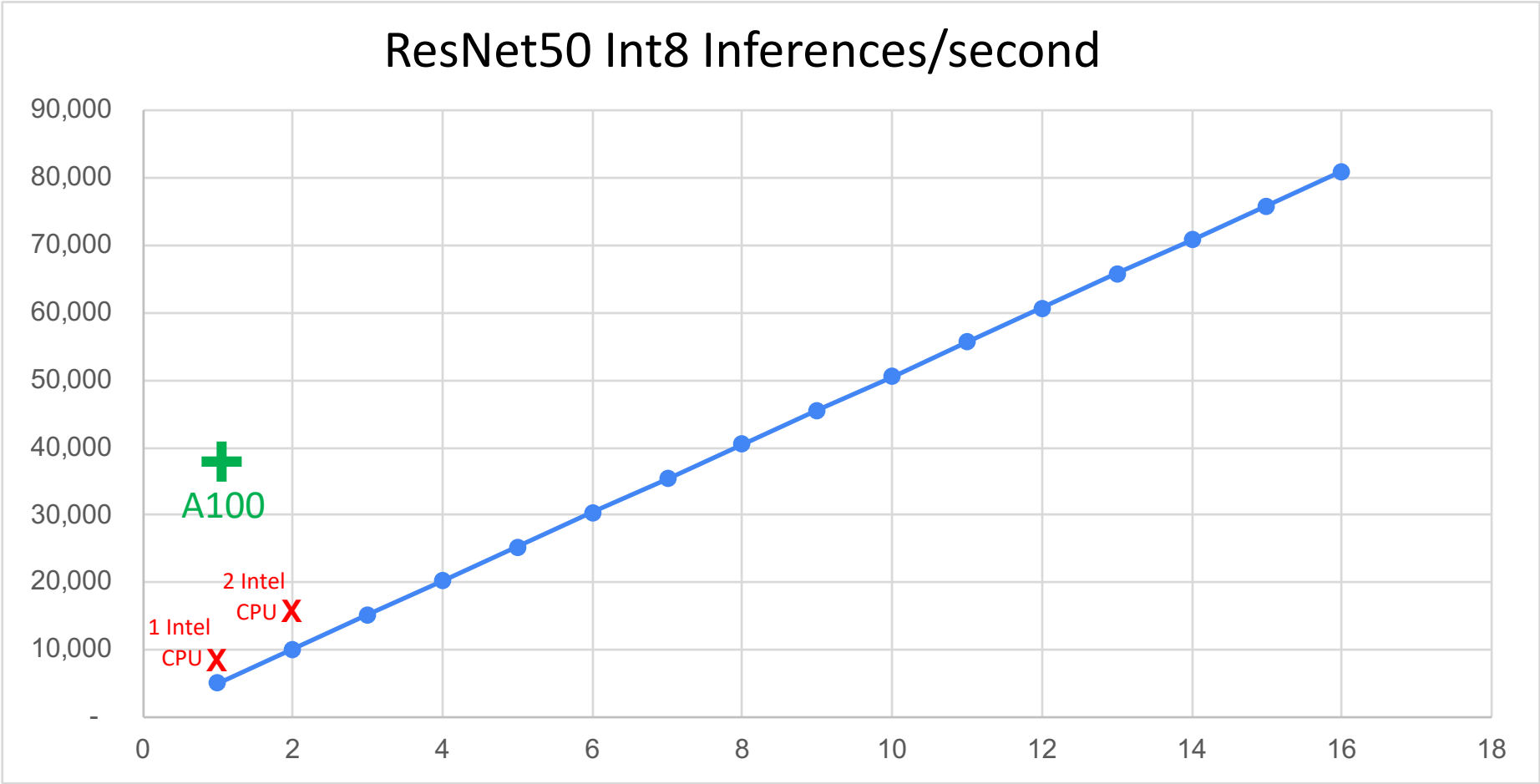
Esperanto RISC-V based ET-SoC-1 has excellent Performance per Watt



	2 Intel Sapphire Rapids CPUs	NVIDIA A100 PCIe Card	ET-SoC-1 PCIe Card
RN50 Inf/Sec	16178	39672	5059
Card or CPU Power	700	300	32
Inf/Sec/Watt	23	132	158

Notes:
Intel performance from MLcommons.org website, CPU TDP power from Intel datasheets
Nvidia A100 power and performance measured by Esperanto running RN50 Int8
Esperanto ET-SoC-1 card power and performance measured by Esperanto running RN50 Int8

32 GB of DRAM in Esperanto Cards enables near linear performance scaling for fully replicated workloads like ResNet50



Number of Esperanto ET-SoC-1 PCIe cards in server

Flexibility to choose the performance and power you need

1,093 RISC-V processors in a single ET-SoC-1 chip

- 1088 ET-Minion 64-bit RISC-V processors with 512-bit Vector/Tensor units
- 4 64-bit RISC-V ET-Maxion high-performance out-of-order processors (e.g. to run Linux)
- 1 RISC-V Service processor

17.5K RISC-V processors on 16 PCIe cards in a 2U high server

350K RISC-V processors could be put into 20 2U servers a single rack

Excellent performance per watt, rivaling the best GPU's but with simple RISC-V programming

Program with our Machine Learning software stack or native RISC-V SDK

Shipping to customers now, contact us if you would like to have thousands of RISC-V cores

Contact info@esperanto.ai for more information



www.esperanto.ai



www.esperanto.ai