esperanto.ai

Enabling Machine Learning & HPC Workloads with Energy Efficient RISC-V Solutions

SC23 RISC-V Workshop
Contact: lee.flanagin@esperanto.ai
November, 2023

ET-SoC-1: Esperanto's RISC-V Supercomputer on a Chip

Over 1,000 64-bit RISC-V CPUs per Chip

As low as 13W per SoC (workload dependent)

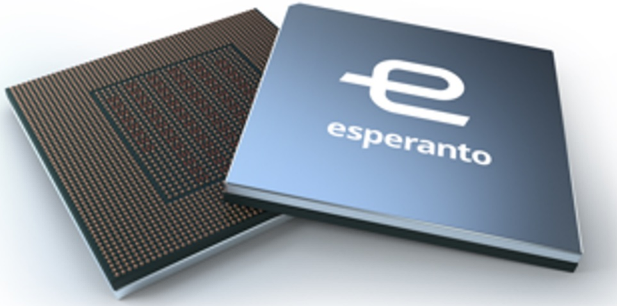
High efficiency operation (performance / watt)

Accelerates wide range of AI/ML workloads

- Language Models (NLP/LLM)
- Visual Models (Detection, Segmentation)
- Recommendation Models (RecSys, DLRM)

High bandwidth interconnect network

Dynamic, tiered architecture of 160 MB on-die SRAM for caches



Allows flexible general-purpose computing eg single precision HPC workloads

Up to 32GB LPDDR4x DRAM

Enables pre- and post-processing

4 ET-Maxion high-performance Out-of-order RISC-V cores

Highly efficient HPC workload through massive parallelism

TSMC 7nm

Summary Statistics of ET-SoC-1

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

- 24 billion transistors
- Die-area: 570 mm^2
- 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 1.5 GHz

1 64-bit RISC-V service processor

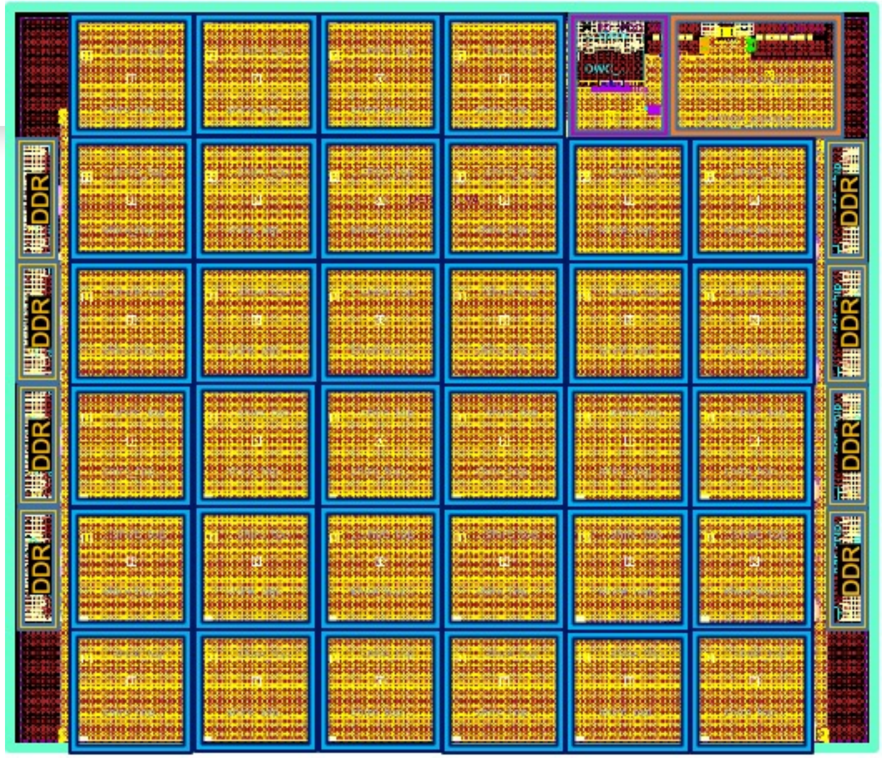
Over 160 million bytes of on-die SRAM used for caches and scratchpad memory

Root of trust for secure boot

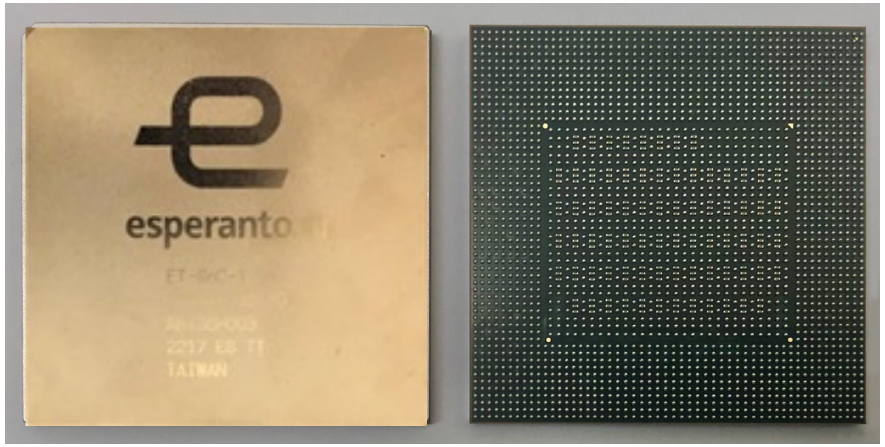
Power typically 20-30 watts, can be adjusted for 15 to 60 watts under SW control

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

Status: Shipping to customers



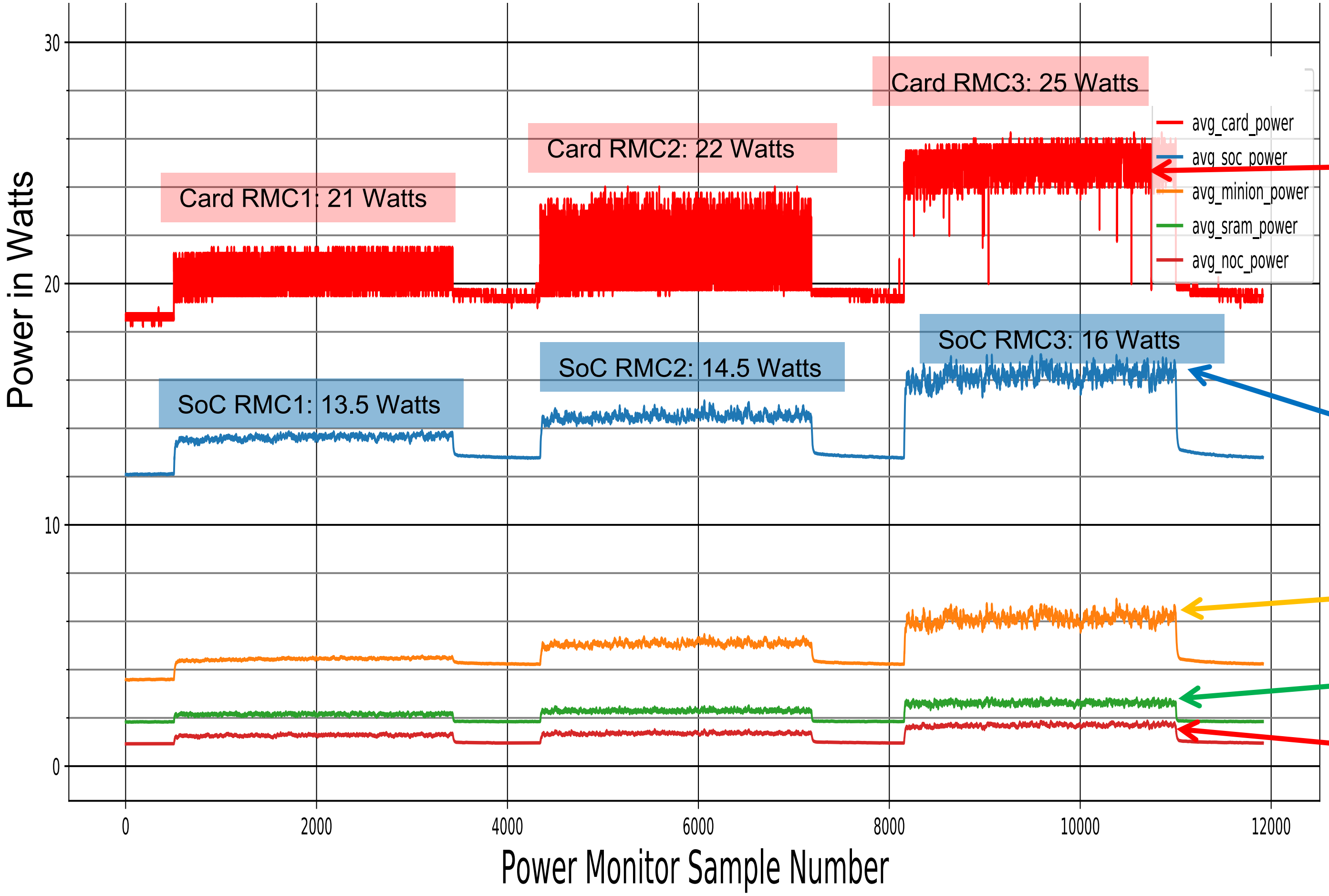
ET-SoC-1 Die Plot



ET-SoC-1 Package

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

DLRM FP16 RMC1, RMC2, RMC3



Total PCIe card power under 25 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

ET-SoC-1 PCIe Production Card

PCIe Gen4 8-lane low-profile form factor

ET-SoC-1 and 32 GB LPDDR4x memory

Safety and emissions certification for multiple geographies

Developed, tested, certified and delivered by Esperanto's partner Penguin Solutions / Smart Global Holdings (SGH)

Implemented across various server form factors

- Single card (ruggedized edge server)
- Up to 6 cards (short-form 2U server for enterprise edge)
- Up to 16 cards (ultra-high density 2U server for data center)

Penguin Computing production ET-SOC-1 PCIe card with and without heatsink

Visit Our Partner's Booth (Penguin Computing) # 843

Esperanto’s RISC-V Evaluation System – Shipping Now

esperanto.ai

- **Up to 16 energy efficient Esperanto PCIe Cards**
- **Being used today by leading researchers for Machine Learning and HPC workloads**

Target environment	Data center	
Server configuration	Standard 2U 19” rack-mount chassis	
ET-SoC-1 PCIe cards (Gen4x8)	8 Cards	16 Cards
System host processor	2x Intel Xeon® Gold 6326 16-core	2x Intel Xeon Platinum 8358P 32-core
System memory	512GB DDR4-3200	1TB DDR4-3200
Storage	2x Samsung® PM9A3 3.84TB NVMe U.2 SSDs	
ET-SoC-1 frequency	600 MHz	
ET-SoC-1 power consumption	15W to 50W (workload dependent)	
ET-SoC-1 RISC-V CPUs	8,448 or 16,896 (high performance, low power)	
Pre-installed AI models	DLRM, ResNet50, BERT, other LLMs, GPT family via Pytorch (updates available regularly)	
ML SDK	Jupyter Notebook and command-line tools	
GP SDK	Included	
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	

Esperanto brings RISC-V to ML and HPC today

Eighty Thousand Esperanto RISC-V processors at work in photo at right

One standard rack can hold twenty 2U servers with

- 320 Esperanto accelerator chips
- 1088 64-bit RISC-V CPU with vector/tensor accelerators per chip
- 348,160 total RISC-V processors
- 24 PetaOps Int8 peak performance
- 6 PetaFlops FP16 peak performance
- 3 PetaFlops FP32 peak performance

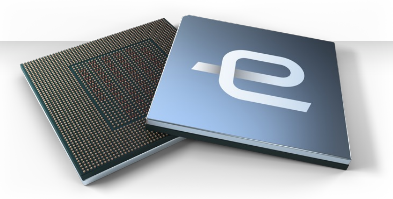
Being used today for Machine Learning workloads

You can program it for HPC as well with Esperanto's General Purpose SDK

esperanto.ai

Sneak Peek: Esperanto’s RISC-V Roadmap

Esperanto Roadmap for merging ML and HPC in a multi-chip system



ET-SoC-1

- Primary application: ML Inference
- 7nm
- Shipping

FP32: 16 TFlops/GHz

FP16: 32 TFlops/GHz

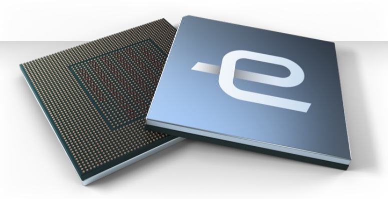
Int8: 128 TFlops/GHz

Up to 32 GB DRAM per SoC

Custom Vector/Tensor unit

Efficiency today:

10 TF SP @40W = 250 GFlops/Watt



Second Generation

- Primary applications: Merged ML and HPC
- 4nm
- In final stages of design

Native support for FP64, FP32, FP16, BF16, FP8 and Int8

Improved peak performance, energy efficiency, memory capacity & memory bandwidth

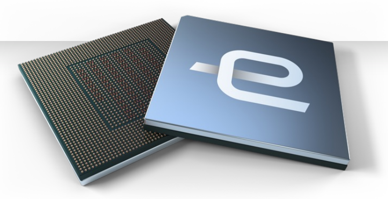
Up to 256 GB DRAM per SoC

All RISC-V RVV 1.0 Vector instructions;

Esperanto Tensor instructions

Efficiency:

@40W SoC >300 DP GFlops/Watt



Third Generation

- Primary applications: Merged ML and HPC
- 3nm
- In architecture definition w/ lead customer

Features for large scale supercomputers

Additional information coming soon

esperanto.ai

www.esperanto.ai