Esperanto's RISC-V Pre-Production Evaluation Systems For Datacenter AI Inference Applications

Esperanto Technologies is offering its RISC-V solutions to qualified customers for evaluation in pre-production system form factors. Esperanto's pre-production systems enable developers to have direct access to thousands of high-performance, low-power RISC-V cores, which deliver superior compute efficiency compared to alternative CPU- and GPU-based offerings. Target applications include AI inference workloads, such as Recommendation, Transformer and Computer Vision, as well as non-AI workloads.

Esperanto pre-production servers come in standard 2U form factors, are rack mountable, and offer a configurable number of ET-SoC-1 inference accelerators.

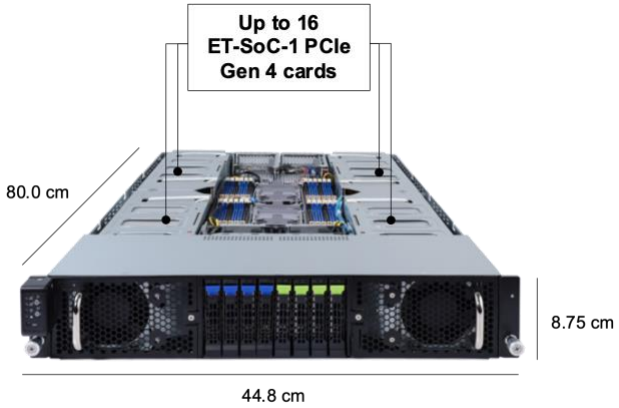
ET-SoC-1 is a high-performance, energy-efficient AI inference accelerator, delivering high performance per watt. Flexible and scalable, ET-SoC-1 supports a range of AI inference applications and delivers highly parallelized performance with low Total Cost of Ownership (TCO).

ET-SoC-1 Specifications

- 1,093 RISC-V processors on a single 7nm chip
- Over 1,000 energy-efficient ET-Minion 64-bit RISC-V in-order cores, each with an ML-optimized vector/tensor unit
- Tensor instructions optimized for machine learning matrix computation
- 4 high-performance ET-Maxion 64-bit RISC-V out-of-order cores
- Over 160MB of on-chip SRAM
- Interfaces for external low-power LPDDR4x DRAM and eMMC FLASH
- Compatible with PCI Express Gen4
- Low power architecture and circuit design techniques
- Patented voltage scaling technology for adjustable compute and power
- Operation at 10 to 60 watts per chip for ML recommendation workloads

Esperanto's Datacenter AI Inference Evaluation Server

Esperanto's evaluation server is for pre-production use and is a delivered as a 19" rack-mounted chassis for the data center environment. It comes with either 8 or 16 ET-SoC-1 inference accelerator cards, dual Intel Xeon Gold 16-core or Platinum 32-core host CPUs and 512GB to 1 TB of DDR4-3200 system memory. Also included are several pre-installed AI models including DLRM RMC, RMC2 and RMC3, ResNet-50 and BERT-Base. Additional AI models will be added, and customers are able to leverage their own pre-trained models and datasets via Esperanto's software development toolkit (SDK).

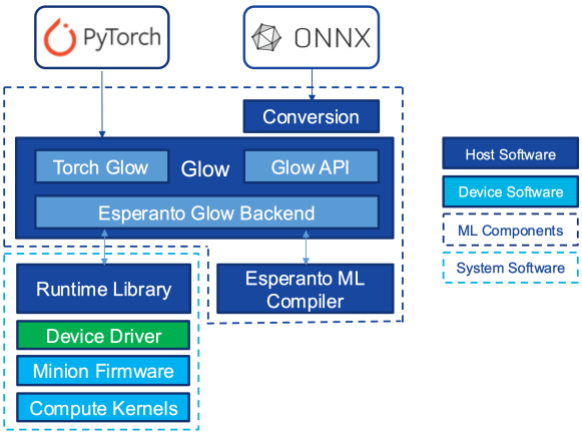


Target environment	Data center	
Server configuration	Standard 2U 19" rack-mount chassis	
ET-SoC-1 PCIe cards [1]	8 Cards	16 Cards
System host processor	2x Intel Xeon Gold 6326 16-core	2x Intel Xeon Platinum 8358P 32-core
System memory [2]	512GB DDR4-3200	1TB DDR4-3200
Storage	2x Samsung PM9A3 3.84TB NVMe U.2 SSDs	
Operating System	Ubuntu 20.04 LTS	
ET-SoC-1 performance	600-800 MHz	
ET-SoC-1 power consumption	10W to 60W (workload dependent)	
ET-SoC-1 RISC-V CPUs	8,704 (8 cards)	17,408 (16 cards)
Pre-installed AI models	DLRM RMC 1, DLRM RMC 2, DLRM RMC 3, ResNet50, BERT-base; additional models to be supported throughout 2H 2022	
ML software	Jupyter Notebook and command-line tools	
Performance, power, and trace analysis tools	Included (et-powertop, Perfetto)	
Training and documentation	Included	
System Power	Dual redundant 3,200W power supplies, 100V–240VAC	
Connectivity	2 x 10Gb/s LAN ports built in, 2 free PCIe Gen4 x16 slots	

Esperanto Software Development Environment

ET-SoC-1 is supported by a ML software environment, which enables customers to run pre-loaded reference models (ResNet50, DLRM and BERT) and datasets as well as compile and use their own models and datasets.

Data scientists can explore ET-SoC-1 via Jupyter Notebook/Lab running in a browser or leverage the software stack via a command line interface. The monitoring of power and real time statistics, and analyzing the resulting traces using the Perfetto visualization framework also is supported.



Esperanto's ML compiler stack is built on Glow (Graph Lowering), which is a compiler for AI/ML models, as well as an execution engine for hardware accelerators. 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(s).

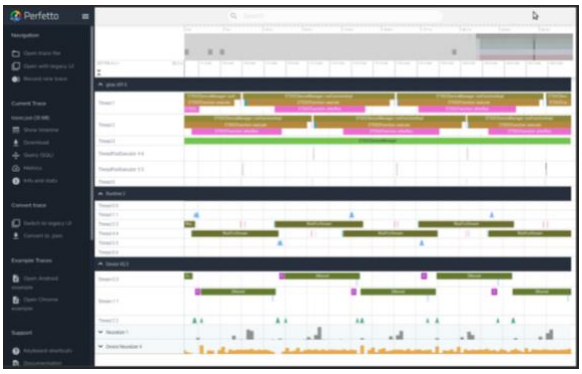
Preloaded ML Models and Datasets

Esperanto systems ship with preloaded models for ResNet-50, DLRM RMC1, DLRM RMC2, DLRM RMC3 and BERT-Basic.

Dataset	Count	Path
ImageNet Validation Test	50,000	imagenet2012/data/
SQuAD v1.1	128	bertMlPerfInfFiles/data/
Criteo Kaggle (26 tables)	131,072	dlrmKaggle26InfFiles/data/
Criteo Kaggle (21 tables)	131,072	dlrmCatInfFiles/data/

Data Visualization

Perfetto is a visualization framework for software profiling and trace analysis. Esperanto's engineers use Perfetto to optimize every aspect of the company's device driver, device management software, and compute kernels. Esperanto's customers can use Perfetto to examine the execution of application-layer code and ML models, allowing them to improve hardware utilization, efficiency, and throughput.



Support

Qualified customers will receive dedicated support by Esperanto throughout their evaluation process.