RISC-V's revolutionary role for simultaneously supporting Machine Learning and HPC

David R. Ditzel

dave@esperanto.ai
Esperanto Technologies Europe, S.L.U.

Abstract

High-end computing is undergoing revolutionary change. After many years of most high-end computers being used to serve web pages, the last decade saw a remarkable growth in the use of machine learning. Separately, HPC computers continued to grow peak flop rates, but system design was largely constrained to x86 server CPUs coupled with GPU accelerators. The thesis of this presentation is that the rise of generative-AI system is the final tipping point that will push for merged ML/HPC system design, and that RISC-V is well positioned to be able to take a leading role in this revolution.

Introduction

Esperanto Technologies is one of a few providers of high-end RISC-V computing systems. Esperanto's ET-SoC-1 processor has over a thousand 7nm RISC-V RV64GC processors each with a custom vector/tensor unit optimized for and is currently shipping to customers. One rack of Esperanto systems could hold over 348 thousand RISC-V processors. That experience has given Esperanto a unique view into how RISC-V can compete with CPUs and GPUs in machine learning and has been instrumental as we look to extend our next generation processors to be capable of handling both machine learning and HPC applications. This talk will cover industry trends, particularly generative AI, how the Esperanto current product is applied to machine learning and give some of the first previews of what Esperanto is planning for our next generation products suitable for both ML and HPC, and how Esperanto's technology would fit with RISC-V processors from other vendors to create full HPC/ML supercomputing systems.

RISC-V can compete with GPU's on ML

General-purpose CPU's have a bad reputation for not being energy efficient or performant at running machine learning software. Empirically GPU's have better performance per watt and overall higher performance than most x86 CPUs. But the blame for CPU's being worse is not solely due to being based on a general-purpose instruction set, but rather on the extreme instruction set bloat and other baggage of the x86 instruction-set in particular. Esperanto set out to show that the simplicity of the RISC-V ISA could enable multi-core implementations that could be as energy efficient as GPUs, and we described this challenge and our intentions as Esperanto began implementations [1].

The key challenge Esperanto sought to prove, was that general-purpose instruction sets could be just as efficient as GPUs, if based on the free and open RISC-V instruction set.

Esperanto's ET-SoC-1

This talk will start with a brief overview of the Esperanto ET-SoC-1 chip and its use in high performance systems.

ET-SoC-1: The Esperanto Supercomputer-on-a-Chip (ET-SoC-1) is fabricated in TSMC's 7nm technology node. Each ET-SoC-1 chip contains 1088 ET-Minion RISC-V processors, four ET-Maxion RISC-V processors, over 148 MB of SRAM used for caches and scratchpads, a mesh "Network-on-Chip" for inter-CPU communication and I/O, sixteen LPDDR4x DRAM controllers, a root of trust for secure boot, PCI Express 4.0 interfaces, an eMMC flash memory interface, USB interfaces and various serial peripheral and JTAG interfaces. The chip may be operated either as an accelerator attached via PCI Express to a host CPU or in a stand-alone configuration.

ET-Minion Processors: The ET-Minion processor is an Esperanto-designed RISC-V RV64GC in-order dual-threaded core. Each ET-Minion processor also has an attached custom vector/tensor unit that is optimized to process common ML data types, including 8-bit integers, 16-bit (half-precision) floating point and 32-bit (single-precision) floating point. The vector/tensor unit can process 128 8-bit integer, 32 half precision or 16 single precision operations per cycle. The integer pipeline, vector/tensor unit and an L1 data cache are all designed for low-voltage operation to both reduce power and improve energy efficiency. At relatively low voltages the ET-Minion processors can operate at frequencies between 500 MHz and 1.0 GHz.

ET-Maxion Processors: The ET-SoC-1 chip also includes four ET-Maxion processor cores. ET-Maxion uses a high-performance 64-bit out-of-order pipeline to implement the RISC-V RV64GC instruction set, derived from the BOOM RISC-V implementation and roughly doubling BOOM's performance per clock. While less energy efficient than the in-order ET-Minion cores, ET-Maxion cores provide higher single-thread throughput and are particularly useful in stand-alone applications, such as serving as an on-chip host processor running the Linux operating system to drive the array of ET-Minion processors. The ET-Maxion can fetch up to 8 RISC-V instructions per cycle, issue up to 5 instructions per cycle and retire up to 4 instructions per cycle. Each ET-Maxion processor has a private 64KB data cache and a 32KB instruction cache. The four ET-Maxion processors share a coherent unified 4MB L2 cache and operates at 1.5 GHz.

Internal Memory System: In order to keep all these processors fed with both data and instructions, three levels of caches are distributed across the die, providing several terabytes per second of on-chip bandwidth. 148 MB of on-die SRAM can be flexibly allocated to either caches or scratchpad memory.

External Memory System: Even our large amount of on-die memory is not sufficient for many important inferencing applications, such as recommendation systems, where important models already require 100 GB+ of accelerator memory and are expected to continue to grow. To accommodate these models, each ET-SoC-1 chip provides sixteen 16-bit LPDDR4x controllers supporting up to 32GB of DRAM with a total throughput of 137 GB/s. Models larger than 32GB can be accommodated by operating multiple chips in parallel, as described later. We chose LPDDR4x as it is both cost effective and energy efficient. By using a 256-bit interface we can achieve the bandwidth required for the most demanding ML inferencing applications, and which can service HPC type applications as well.

PCIe Interface: Each ET-SoC-1 chip implements eight PCI Express 4.0 lanes for a peak bandwidth of 16 GB/s in each direction. The PCIe controllers allow the chip to act as a host or endpoint or be split to support simultaneous host and endpoint configurations.

Low-Power Energy-Efficient Circuits: Running over a thousand cores along with vector units at the same time might be expected to consume hundreds of watts, as is common with other accelerator chips. Esperanto used a number of design techniques to enable very low voltage operation for the ET-Minion processors and other performance-critical circuits. The entire ET-Minion processor is designed to operate robustly at voltages as low as 400 millivolts. Operating at these low voltages also improves energy efficiency per operation. As a result, the entire ET-SoC-1 chip operates with around 20-40 watts of power on its targeted applications.

Esperanto's experience with RISC-V for ML

The proof is in building real live systems and shipping them to customers. Esperanto has been able to show that RISC-V can enable system to compete against GPUs [2]. This talk will review some of those real-world results of ET-SoC-1 systems.

The ET-SoC-1 chip typically consumes less than 40 watts for ML or HPC compute intense workloads such as matrix multiplies, which makes it appropriate to implement on a simple PCIe card, keeping under the standard 75 watt per card limit without additional external power supplies, and allowing for simple air cooling. This allows up to 16 PCIe cards to be put into a simple 2U high conventional server such as the Gigabyte G292-2G0. This particular system also has two high-end Xeon processors to server as the host CPU.

With 1088 RISC-V processor per ET-SoC-1 chip, and one ET-SoC-1 per PCIe card in a 16-card system, that 2U system can hold up to 17,408 ET-Minion RISC-V processors.

Up to 20 of these 2U high servers can be put into a single rack with 348,160 RISC-V ET-Minion processors. This is a great example of the density that can be achieved with RISC-V. Figure 1 shows a somewhat smaller system, with ten 2U servers each holding 8 PCIe accelerator cards, for a total of 87,040 RISC-V ET-Minion Processors. If each ET-Minion processor were to run at 1 GHz, peak performance of one rack with 20 2U servers would be about 40 PetaOps Int8, 10 PetaFlops of FP16 or 5 PetaFlops of FP32.

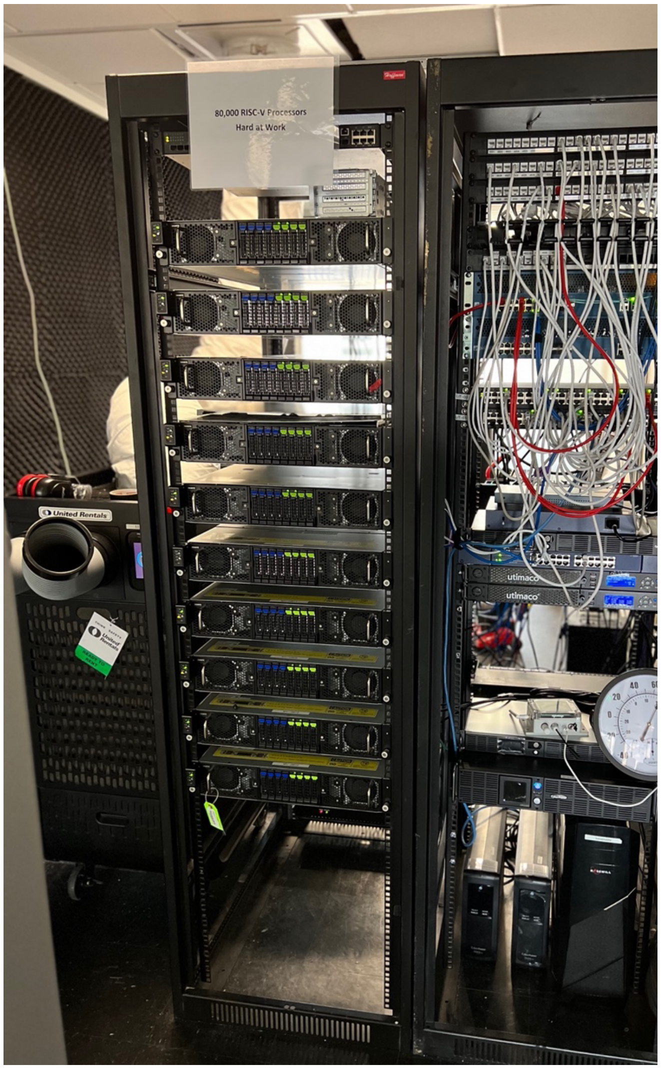

Figure 1. Example of Ten 2U Esperanto Servers with 87K RISC-V processors in this rack.

Customers want merged ML/HPC systems

In discussions with potential customers of ET-SoC-1 systems for machine learning applications, two comments were frequently heard. First, customers liked that the processors were based on RISC-V and were interested in being able to program at the RISC-V level to add additional functionality beyond pure ML applications. Second, many wished for a broader range of data types, particularly high-performance double-precision floating-point, so that the same systems could also run HPC applications. Conversations with large potential HPC customers repeatedly said that they wished for and expect a convergence of ML and HPC systems. For example, this was a view shared by the U.S. Department of Energy as part of their future supercomputer acquisitions [3].

Generative AI is the tipping point for merged ML/HPC

The recent success of GPT-4 [4] has caused a dramatic interest in systems that can run a variety of Generative AI workloads. We have analyzed generative AI workloads and will comment on what it would take for RISC-V based systems to run these well.

Esperanto Roadmap for ML/HPC

With the experience Esperanto has gained from the ET-SoC-1, our focus is now on next-generation chips and systems. Particularly with the recent ratification of the RISC-V Vector Specification (RVV) Version 1.0, there is the opportunity to have a single vector implementation that can be good for both ML and HPC. Esperanto will comment on its findings for penalties caused by adding HPC support, and its impact on ML workloads.

Esperanto will comment on directions for its next generation processors, and how they are expected to excel at both ML and HPC workloads.

Large scale system will likely have many components, and there is opportunity to utilize RISC-V technologies from many vendors. In particular, several companies have announced high-performance out-of-order cores expected to rival the performance of high end x86 CPUs. Recent announcements of these high-performance cores are Ventana's Veyron V1[5] and Tenstorrent's Ascalon processor [6]. This talk will comment on how these might be combined to build great ML/HPC systems completely based on RISC-V.

Summary: RISC-V's opportunity for ML/HPC

Yes, RISC-V can be a great base for future ML/HPC large scale computer systems and has a chance to be some of the best performing systems for Generative AI. This talk will expand on these comments, and in particularly why RISC-V is likely to be the most successful general-purpose processor for future high-end ML/HPC systems.

References

[1] D. Ditzel, Industrial-Strength High-Performance RISC-V Processors for Energy-Efficient Computing, 7th RISC-V Workshop, November 28, 2017

[2] D. Ditzel, Thousands of RISC-V Cores for AI and Beyond, 2022 RISC-V Summit, December 13, 2022.

[3] Private communication with Rick Stevens, Argonne National Labs, 2022.

[4] GPT-4 Technical Report, OpenAI (2023).

[5] B. Baktha, Announcing Veyron V1: World's First Data Center Class RISC-V Processor, 2022 RISC-V Summit.

[6] W. Lien, High Performance RISC-V Processor for Computation Acceleration and Server, 2022 RISC-V Summit.

Author Bio: Dave Ditzel is the founder and CTO of Esperanto Technologies. Esperanto builds energy-efficient processors for AI and HPC based on RISC-V, with over a thousand RISC-V vector/tensor processors on a chip. Dave spent six years at Intel Corporation as VP of Hybrid Computing, leading a team building a high-performance processor using binary translation to run x86 applications. In 2007 he founded ThruChip Communications, to reduce IO energy between 3D stacked die using inductive communication. In 1995 Dave founded Transmeta Corporation, which developed low-power x86-compatible processors using Code Morphing Software. Dave spent 10 years at Sun Microsystems as CTO for the SPARC Technology Business. Prior to Sun, Dave spent 10 years at AT&T Bell Laboratories, where he worked on RISC processors optimized for the C language. Dave was a graduate student under U.C. Berkeley Professor David Patterson and in 1980 they co-authored "The Case for the Reduced Instruction Set Computer".