The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

ESPERANTO MINIONS EXCEL AT AI

Startup Nears Tapeout of Thousand-Core Accelerator

By Linley Gwennap (December 21, 2020)

Startup Esperanto Technologies has designed a chip with a thousand tiny RISC-V cores based on its ET-Minion design. Each core has a custom tensor unit that offloads the matrix multiplication common in neural-network inferencing; working together, the Minions can perform 139 trillion INT8 operations per second (TOPS) when operating at 1.0GHz. Yet the ET-SoC-1 design is estimated to consume only 20W in typical operation. This power efficiency is suited to both data-center accelerators and network-edge applications.

The chip has taken a long road to market. Esperanto completed its RTL design in September 2018 in preparation for a mid-2019 tapeout, but completing the design took far longer than expected. In late 2019, the board replaced founding CEO Dave Ditzel with Art Swift, former CEO of MIPS and Wave Computing. Swift brought in Darren Jones, also from MIPS and Wave, to run the engineering team, although Ditzel remains involved in technology development.

Esperanto went silent for a year, declining to announce even the CEO change, as it worked to complete and verify the complex 7nm design. Having reached this milestone, it expects to tape out to TSMC soon; if so, we expect samples in 2Q21 and full production in 1H22. To support this effort, the company has a total of $82 million in funding and a staff of more than 100 employees and full-time contractors.

The chip contains 34 tiles that each comprise 32 ET-Minion cores, for a total of 1,088 cores and 136MB of SRAM. As Figure 1 shows, the chip also features four supervisory ET-Maxion CPUs, an eight-lane PCI Gen4 interface, and four 64-bit DRAM interfaces. The company withheld the die size for the 24-billion-transistor chip, which we expect measures about 300mm^2. It plans to offer the chip standalone, on a PCIe card, and in the Dual M.2 form factor that Facebook favors. A Glacier Point card can hold six of these M.2 modules in a dense power-efficient design. Using this card, the Esperanto modules can achieve more than 800 TOPS with a typical power of 120W. The company plans to disclose neural-network benchmark results once it completes silicon characterization.

Minions Reduce RISC-V Power

Because the ET-Minion cores comprise most of the die area, Esperanto focused on maximizing their power efficiency.

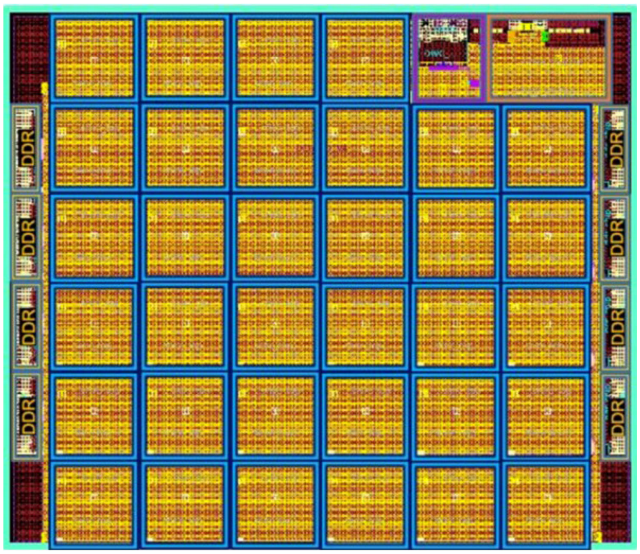

Figure 1. Esperanto ET-SoC-1 die plot. The 34 blue boxes indicate the ET-Minion tiles, each comprising 32 cores. The purple box contains four more-powerful ET-Maxion CPUs, and the orange box surrounds the PCIe controller. The left and right edges show the sixteen 16-bit DDR DRAM interfaces. (Image source: Esperanto)

They operate from a separate voltage plane that typically supplies about 0.4V, which is well below the standard 7nm voltage but above the threshold voltage. Although it's a simple in-order scalar design, the CPU employs a longer pipeline than typical five-stage RISC-V cores, not to enable high-speed operation but to allow moderate speeds at low voltage. As a result, the core can reach 1.5GHz or higher in 7nm but normally operates at 1.0GHz or less to save power.

To enable the entire core to operate at such low voltage, the company designed a custom SRAM cell. This cell is larger than standard TSMC SRAM but can achieve the target clock speeds at low voltage. Each CPU has 4KB of SRAM that's configurable as data cache, scratchpad memory, or a combination of both (with 512-byte granularity) depending on the application. The remainder of the core employs other circuit-design techniques to reduce power, along with extensive clock gating.

The Minion core implements the 64-bit RISC-V base integer instruction set plus the M (multiply), F (FP), and C (compressed) extensions. But it adds custom instructions to optimize performance on neural networks, including tensor multiplication and transcendentals; the latter accelerate certain activation functions.

To increase utilization, Minion is dual threaded. Although this technique doubles the size of the register files, it enables the core to continue processing after a cache miss or other multicycle data fetch by simply switching to the other thread. The second thread can also be used for software-controlled prefetches. The scalar portion of the CPU fetches, decodes, and executes standard RISC-V instructions using an integer ALU and a load/store unit.

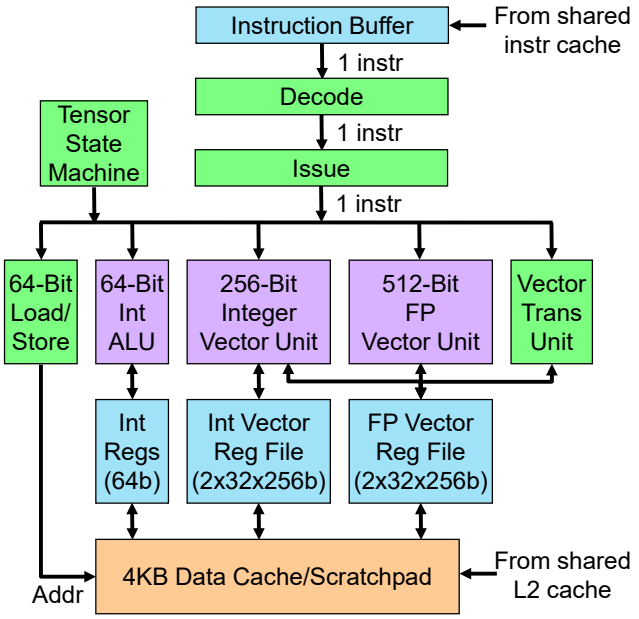

Figure 2. Minion CPU block diagram. Along with a small scalar (64-bit) unit, the core contains two wide vector units that accelerate matrix multiplication, plus a vector unit for transcendental functions (trans).

Autonomous Tensor Operations

The ET-Minion CPU includes two vector units: one for integer data and one for floating-point data. Although the chip targets AI inference, Esperanto and Facebook believe floating point remains an important data type. The 256-bit-wide FP vector unit can perform 8 multiply-accumulate (MAC) operations per cycle for 32-bit floating-point (FP32) data or 16 MAC operations for FP16 data. It draws data from a vector register file that has 32 entries per thread (64 total), as Figure 2 shows. The 512-bit-wide integer vector unit can perform 64 MAC operations per cycle for 8-bit (INT8) data, achieving the chip's peak performance. To satisfy its full width, it simultaneously draws data from its private 256-bit-wide vector register file and the FP vector registers.

Many processors, including some RISC-V designs, provide vector units and vector registers. What's unique about Minion is its custom multicycle tensor instructions that can process arbitrarily large matrices. To execute tensor instructions, the core contain a small state machine that cycles through the data. A single instruction can thus operate for dozens or hundreds of cycles; during this time, the entire scalar portion of the core, including the front end and the scalar units, is clock gated. This approach uses far less power than a standard vector processor that must fetch and execute one instruction per cycle for maximum throughput.

One problem is that the vector registers hold only 32 operands, which must be allocated among two multiplicands and the accumulator on each cycle. The tensor instructions are optimized for common convolution operations in neural networks. When operating on INT8 values, both multiplicands stream from memory while the vector registers hold the accumulators, which are always 32 bits wide to avoid overflow. Because the data cache is so small, the tensor state machine loads these values directly from the level-two (L2) cache in parallel with the computations. The L2 cache can provide 512 bits per cycle, enough to maintain peak compute performance.

The CPU also includes a vector transcendental unit that computes sine and exponential functions. Combining these functions allows quick calculation of sigmoid, hyperbolic tangent (tanh), and other functions that are frequently used to postprocess convolution results. The unit works on the vector registers and can compute four FP32 values per cycle. It employs ROM-based lookup tables, which consume considerable die area but very little power.

A Short Visit to the Shire

The ET-Minion cores are in groups of eight called neighborhoods. Each neighborhood shares a 32KB instruction cache that provides one cache line (16 instructions) per cycle to each of two cores; the cores buffer these instructions so they can continue to operate while they take turns accessing the cache. When executing neural-network code, a core can take dozens of cycles to perform a single matrix instruction, reducing pressure on the instruction cache.

Each tile, which Esperanto calls a shire, contains four neighborhoods for a total of 32 cores. Each neighborhood has a single connection to a four-port crossbar that provides access to four 1MB SRAM banks, as Figure 3 shows. Software can independently configure subbanks as scratchpad memory, L2 cache, or L3 cache as appropriate for the application. Scratchpad memory fits into a unified global address space that all cores can access, allowing software to directly manage the location of all data on the chip. L2 cache is shared among all the Minions in a shire, whereas L3 cache is shared across the entire chip. For example, software could configure three L2 banks and one L3 bank per tile on 32 tiles, providing a 3MB L2 cache for each tile plus a distributed 32MB L3 cache.

As noted, the cores are designed for low-voltage operation to increase power efficiency, but the custom low-voltage SRAM is less dense than standard SRAM. For the shire memory, Esperanto employs high-density SRAM to maximize the chip's storage capacity, but this memory is designed for nominal voltage. Thus, the tile features multiple voltage planes, separating the cores from the shire's SRAM. In fact, the chip contains dozens of power domains, enabling fine-grained power management to optimize efficiency.

Assembling the Tiles

The tiles communicate using a standard mesh architecture. Each has a mesh controller that connects to the north, south, east, and west tiles. In Esperanto's design, the mesh includes one link in each direction, each delivering 128GB/s at 1.0GHz. In addition to the 34 ET-Minion tiles, the mesh connects one ET-Maxion tile and one PCIe tile. The latter controls an eight-lane PCIe interface that operates at up to Gen4 speed for a maximum data rate of 16GB/s. The Dual M.2 module splits this interface across the two connectors. This interface typically links the accelerator to a host processor.

The chip can also operate in standalone mode using its integrated host processor. The Maxion tile contains four Maxion CPUs, which Esperanto designed as general-purpose application cores that can run an RTOS or even Linux (see *MPR 12/10/18,* "Esperanto Maxes Out RISC-V"). Operating at up to 1.5GHz, the Maxion cores perform about 60% better than a Cortex-A55 at the same clock speed. The Maxion tile also contains 4MB of L2 cache and an extra Minion core that acts as a low-power service processor.

Although the Minion tiles provide a total of 136MB of on-chip SRAM, large neural networks need more capacity. The chip supports up to 32GB of external DRAM using sixteen 16-bit channels of LPDDR4X-4266 memory, typically implemented with four 64-bit-wide DRAM chips. The complete memory system delivers up to 132GB/s. Each channel has two mesh stops to support the peak data rate. LPDDR4X, which appears in most smartphones, employs 0.6V I/O to minimize power.

Price and Availability

We estimate Esperanto will sample the ET-SoC-1 in mid-2021. The company has yet to announce product details or pricing. For more online information, access *www.esperanto.ai*.

Like an Nvidia GPU, Esperanto's chip will modulate its clock speed to remain within a specified power limit. Thus, the same chip can hit a variety of performance and power points. The company will announce product configurations after it has characterized the initial silicon. Even within a specific power envelope, the chip can operate at a higher speed on lighter workloads. Esperanto's typical power estimate of 20W is based on running a typical deep-learning model at 1.0GHz; a model with high utilization would use more power at that clock speed, or it would have to throttle down to stay at the same power. We expect the chip will reach 30W TDP when running at peak TOPS for 1.0GHz; although this power would exceed the limits of the Dual M.2 form factor, the PCIe card could easily handle multiple chips.

Esperanto is developing a software stack that builds on Facebook's open-source Glow compiler, which converts models from Pytorch or the ONNX exchange format to an intermediate representation. Esperanto's compiler back end converts this representation to RISC-V code, including the company's custom matrix instructions. It also maps the model onto the chip's numerous cores and handles the memory configuration and allocation.

Aiming for Cloud TOPS

Few AI-accelerator vendors target low-power M.2 modules. Qualcomm recently announced its Cloud AI 100 in that

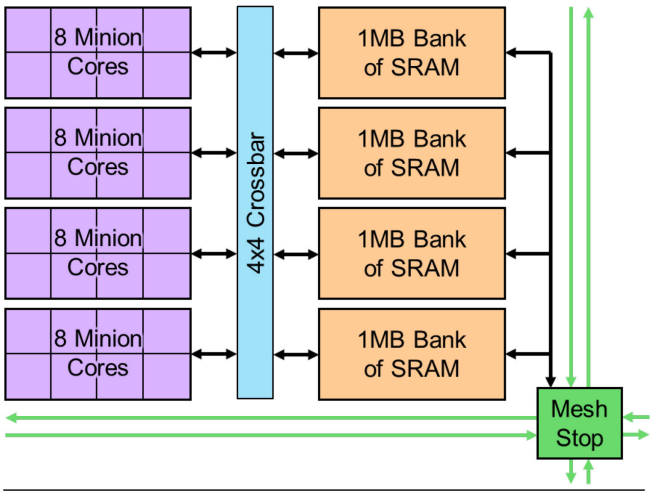

Figure 3. Minion tile block diagram. Each tile, or "shire," includes 32 Minion CPUs grouped into four "neighborhoods." The tile offers 4MB of SRAM divided into four banks to increase bandwidth through the crossbar.

form factor (see *MPR 10/12/20,* "Qualcomm Samples First AI Chip"). Its Dual M.2 module carries a 25W TDP rating; we estimate a 30W TDP for the Esperanto module. At similar power, the Qualcomm chip delivers considerably more INT8 TOPS, although it appears to lack floating-point capability. Thus, Esperanto provides a superior product for FP16 inference.

Both chips have capacious on-chip memory and low-power DRAM subsystems with identical configurations, as Table 1 shows. Both have the same PCIe configuration as well. The chip architectures are different, however, as Qualcomm relies on relatively few cores that each employ large matrix units; Esperanto instead employs more than a thousand cores with smaller matrix units. Both approaches should work well for traditional CNNs such as ResNet, which spend most of their cycles on matrix multiplication. Qualcomm hasn't released benchmarks for other models.

A data-center operator can combine the M.2 modules in carrier cards such as Glacier Point to compete against higher-power designs. Among 75W accelerators, one of the most power efficient is Tenstorrent's Grayskull (see *MPR 4/13/20,* "Tenstorrent Scales AI Performance"). This chip fits onto a bus-powered PCIe card and packs 120 cores, choosing an intermediate point between Qualcomm's big cores and Esperanto's Minions. As Table 1 shows, Grayskull and the Esperanto SoC provide similar performance per watt for both floating-point and integer data. Although both chips are also similar in on-chip and external memory, a Glacier Point card with six modules would far outclass Grayskull in memory capacity.

One drawback for Esperanto is that its chip has yet to reach silicon and won't begin production for at least a year. Both Qualcomm and Tenstorrent have been sampling for months and expect production in the months ahead; these companies may produce second-generation products by the time Esperanto ships its first one. Tenstorrent is already working on a 7nm design that will improve in power efficiency.

	Esperanto ET-SoC-1	Qualcomm Cloud AI 100	Tenstorrent Grayskull
Form Factor	Dual M.2	Dual M.2	PCIe card
AI-Core Count	1,088 cores	16 cores	120 cores
Clock Speed	1.0GHz*	0.8GHz*‡	1.3GHz
Peak FP16 Perf	32Tflop/s	Undisclosed	92Tflop/s
Peak INT8 Perf	128 TOPS	200 TOPS	368 TOPS
Chip Memory	136MB	144MB	120MB
DRAM Channels	4x LPDDR4X	4x LPDDR4X	8x LPDDR4
DRAM Bandwidth	132GB/s	132GB/s	132GB/s
Host Interface	PCIe Gen4 x8	PCIe Gen4 x8	PCIe Gen4 x16
ResNet-50 Inference†	Undisclosed	11,200 IPS	22,431 IPS
Board Power (TDP)	30W‡	25W	75W
TOPS per Watt	4.3 TOPS/W	8.0 TOPS/W	4.9 TOPS/W
IC Process	TSMC 7nm	TSMC 7nm	GF 12nm
Production	1H22‡	1H21 (est)	4Q20 (est)

Table 1. Deep-learning accelerators for inference. Qualcomm targets throughput similar to Tenstorrent's but far better than Nvidia's. *Can operate at 1.5GHz, but at higher TDP; †best batch size, INT8. (Source: vendors, except ‡The Linley Group estimate)

Lowering the Boom on Power

Esperanto brings to the AI market a sharp focus on power efficiency; Ditzel has worked on low-power designs for more than two decades (see *MPR 2/14/00,* "Transmeta Breaks x86 Low-Power Barrier"). The generically named ET-SoC-1 employs a broad range of techniques to boost performance per watt, including low-voltage operation, custom circuit design, and clever microarchitecture. To optimize the critical matrix-multiplication function, the chip uses a simple state machine that minimizes instruction overhead. This approach enables it to pack 128 TOPS into a typical operating power of 20W. By contrast, Nvidia's T4 card requires 70W to provide similar performance. Despite its tensor cores, the GPU is notoriously inefficient in performance per watt.

Other vendors, however, also aim to exploit Nvidia's inefficiency. Qualcomm, in particular, used its power-efficient-smartphone experience to develop the new Cloud AI 100 chip, which offers impressive performance in a power budget and form factor similar to Esperanto's. Neither vendor has disclosed full product details or standard benchmark results, however, so a precise comparison is difficult. Whereas most AI-chip vendors focus mainly on matrix multiplication, we expect Esperanto's many-core design to fare better on RNNs and recommender models that require more general-purpose computation. Having a thousand cores available, even small scalar ones, provides an advantage in this regard. We await benchmark results to confirm this thesis.

Like most AI startups, Esperanto faces the difficult task of building a software stack and optimizing it to efficiently process customer models. Facebook's Glow offers a piece of this puzzle, but it still leaves considerable work for Esperanto. Even if it can demonstrate performance and efficiency advantages, the company must deliver a solid software solution before achieving customer adoption.

Esperanto's chip would've been more competitive had the company stayed on its original schedule. The competition for AI inference is fierce today and will be yet more so in another year. The company's design appears to be highly power efficient yet flexible enough to handle a broad range of models. Once it has achieved silicon, Esperanto must validate the chip's frequency, power, and performance and demonstrate its software stack on real neural networks. Thanks to its hard-working Minions, the company is now ready for tapeout. ♦