The Linley Group

Tech Insights

MICROPROCESSOR report

Insightful Analysis of Processor Technology

VSORA SHIFTS GEARS, OFFERS AI CHIPS

DLA-IP Vendor Plans to Sample First Tyr Chips This Year

By Linley Gwennap (February 14, 2022)

A year ago, Vsora gained attention for its powerful deep-learning-accelerator (DLA) design, which it initially offered to license. After gaining only one major customer, the tiny French company has decided to increase its addressable market by selling chips based on its DLA architecture. It plans to sample the first chip by the end of this year, with two related ones to follow in 2023. The family scales to 1,000 trillion operations per second (TOPS), which Vsora expects to deliver at a typical power of just 50W. The company initially targets automotive customers, but its technology applies to a broad range of applications.

The new chips are branded Tyr (pronounced "tier"), named after the Norse god of war and brother of the better-known Thor. The initial Tyr 1 chip starts with a small version of Vsora's architecture rated at 256 TOPS for AI operations. Tyr 2 will double those specifications, and Tyr 3 will deliver the full 1,024 AI TOPS. All three chips employ TSMC 7nm technology. The company hopes to tape out Tyr 2 and Tyr 3 within a few months of Tyr 1.

Vsora has about 20 employees but plans to double that figure by the end of this year. To produce three chips with such a small team, we believe it's outsourcing Tyr's physical design to an ASIC-design house. Vsora has reported less than $2 million in external funding but needs tens of millions to design and tape out three 7nm chips. Using multi-project wafers would reduce the tapeout fees but prevent the chips from entering volume production without a full tapeout (about $10 million per chip). To cover these costs, the startup has unspecified internal funding—possibly from its primary customer, a major European carmaker.

Building Around the DLA

The Tyr chips employ Vsora's AD1028 architecture, which combines DSP and DLA units (see *MPR 12/7/20,* "Vsora Drives to Deliver Petaflops"). Each DLA unit contains an array of 16K MAC units that operate on 8-bit floating-point (FP8) data, which has about the same range as 8-bit integers (INT8) but can represent numbers as small as 0.004. FP8 also requires less power than INT8, since each MAC operation involves two small multiplies (for the 4-bit exponent and 3-bit mantissa) instead of a single 8x8-bit multiply. Each DSP unit implements a 24-bit ALU that operates on FP data (FP24). This combination allows the design to handle both AI inference and signal processing, making it well suited to automotive systems.

The Tyr 1 design instantiates a total of 64K MAC units and 1,024 DSP units, as Figure 1 shows. At 2.0GHz, its MAC units can reach 256 TOPS for FP8 matrix operations. The DSP units can perform another 4 TOPS for FP24 data. The compute units share 26MB of tightly coupled memory

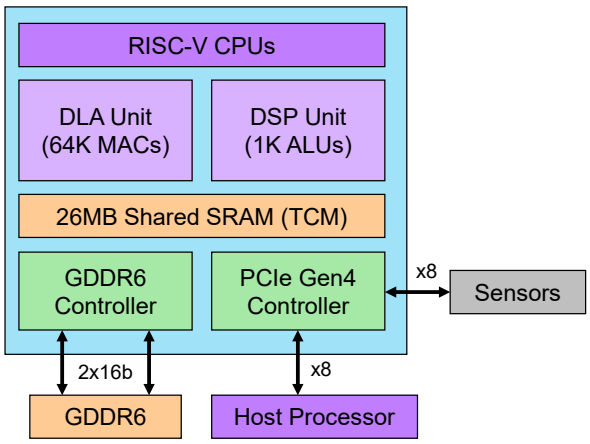

Figure 1. Vsora's Tyr 1 chip. The DLA units can generate 256 TOPS for AI models while the DSP units generate 4 TOPS for signal processing.

Price and Availability

Vsora plans to sample Tyr 1 in 4Q22, followed by Tyr 2 and Tyr 3 in 1Q23. The chips are scheduled to reach production for automotive customers in 2024. The company withheld pricing. For more information, access www.vsora.com.

(TCM) that can store neural-network parameters and other data. For models that don't fit in the TCM, Tyr 1 features two 16-bit GDDR6 channels, enabling 64GB/s to external memory. It provides a x16 PCIe Gen4 interface that can connect to the host processor and sensors. The chip also has several small RISC-V cores that supervise data movement and other low-level tasks.

To develop three chips at nearly the same time, the startup is using a modular approach: the larger ones simply replicate the Tyr 1 physical design to implement more compute units, more memory, and more DRAM connections. Tyr 2 doubles the number of DLA and DSP units in addition to doubling the on-chip memory size and the number of GDDR6 channels. Tyr 3 enables the maximum performance using its 256K MAC units and 2K DSP units. It supports 105MB of on-chip memory as well as 256GB/s of GDDR6 bandwidth.

Because the physical design is in progress, Vsora lacks a final power estimate for its chips, but it expects all three to achieve more than 20 TOPS per watt. Although this figure is lower than the 30 TOPS/W the startup specifies for its AD1028 intellectual property (IP), a complete chip includes clock and power distribution, network-on-a-chip and other data buses, on-chip memory, I/O- and memory-control logic, and high-speed interfaces, all of which consume considerable power. If the products meet this specification, Tyr 3 would require less than 50W and Tyr 1 less than 13W.

	Vsora Tyr 1	Qualcomm AI 100	Nvidia Orin
Application CPUs	None	None	12x Cortex-A78
AI Performance	256 TOPS (FP8)	200 TOPS (INT8)	254 TOPS (INT8)
DSP Performance	4 TOPS (FP24)	2 TOPS (FP32)‡	None
On-Chip Memory	26MB	72MB‡	Undisclosed
DRAM Interface	2x 16-bit GDDR6	2x 64-bit LPDDR4‡	4x 32-bit LPDDR5‡
DRAM Bandwidth	64GB/s	66GB/s‡	102GB/s‡
Host Interface	PCIe Gen4 x16	PCIe Gen4 x8	4x 10GbE
Power Rating	13W TDP	25W TDP	100W TDP
ResNet-50 Perf*	15,835 IPS	11,333 IPS	12,750 IPS‡
Perf per Watt†	1,760 IPS/W	650 IPS/W	180 IPS/W
ResNet-50 Latency	0.13ms	1.02ms	Undisclosed
IC Process	TSMC 7nm	TSMC 7nm	TSMC 7nm
Production	1H24‡	1H21	2H22‡

Table 1. Automotive neural-network chips. Like the AI 100, Tyr 1 requires an external host processor, but it delivers more performance at half the rated power. Nvidia's Orin burns far more power but integrates a beefy host processor. *ResNet-50 v1.5, maximum batch size; †assumes ResNet uses 70% of TDP. (Source: vendors, except ‡The Linley Group estimate)

A Higher Tyr

For automotive systems, the Vsora chips will compete against Nvidia's Orin, a popular choice for high-end ADAS. Nvidia rates the processor at 254 TOPS, almost exactly the same as Tyr 1, but Orin requires eight times more power, as Table 1 shows. This comparison isn't exactly fair, as Orin integrates several components that Tyr lacks, including a powerful host processor with 12 Cortex-A78 and 8 Cortex-R52 CPUs as well as a GPU, image processor (ISP), video engine, and camera interfaces (see *MPR 5/17/21*, "Nvidia Orin Turbocharges ADAS"). Even so, a carmaker could easily add an external processor that would similar functions while consuming far less than 87W. Note that Vsora extends its performance lead on real neural networks (e.g., ResNet-50) owing to the greater utilization of its architecture.

Qualcomm's Cloud AI 100 chip has gained some ADAS traction by winning a design at BMW. The company offers versions in different form factors; at 200 TOPS, the dual-M.2 module matches up best against Tyr 1. At 25W, however, the AI 100 achieves less than half of Vsora's efficiency, assuming the startup's power rating holds up, and its multicore architecture has much worse latency. Like the Tyr chips, the AI 100 requires an external host processor. Qualcomm's AI cores feature Hexagon DSPs, so like Tyr, the AI 100 can handle workloads that mix AI and signal processing (see *MPR 9/27/21*, "Qualcomm Spills Cloud AI 100 Guts"). It also integrates more memory, enabling it to execute larger neural networks on the chip.

Although Vsora provides greater performance per watt, customers may choose Nvidia or Qualcomm for other reasons. These large vendors offer comprehensive and proven software stacks that support a variety of AI frameworks and libraries. Vsora, by contrast, supplies a compiler and a basic TensorFlow interface. The AI 100 is already in production, with validated performance and power data submitted to MLCommons; Orin is sampling and due to enter production at least a year before Tyr 1. Nvidia has extensive experience with automotive-safety certifications such as ASIL D, whereas Vsora includes ASIL D features in Tyr but hasn't validated them.

The startup plans to deliver Tyr 2 and Tyr 3 quickly, pushing AI performance to 1,024 TOPS at just 50W. By the time these chips reach production, however, Nvidia plans to deploy Atlan, an Orin follow-on that targets 1,000 TOPS at 200W. The AI 100 can scale to 400 TOPS at 75W, but Qualcomm will likely have a second-generation product available by 2024. Thus, these competitors may be able to match even Tyr 3's performance, although Vsora should retain its power-efficiency advantage.

An Expensive Path

Many small startups become IP vendors because of the modest engineering cost of this model: they need create only a single function block while avoiding physical design and tapeout fees. Chip startups, however, typically spend $50 million to $100 million to bring their first product to market. This business model can be particularly brutal for automotive suppliers, as years may elapse before the customer begins to buy chips in volume. Vsora appears to lack the engineering resources and the funding to bring three 7nm chips to volume production; instead, we expect it'll use shuttle runs to fabricate test chips, then attempt to raise additional funds after demonstrating the chips' performance. Alternatively, its primary customer could simply acquire the startup.

Vsora's advantage is the impressive efficiency of its architecture. At 30 TOPS/W, the DLA core outperforms all competing IP, including Ceva's NeuPro-M (see *MPR 1/31/22,* "Ceva Tackles Unstructured Sparsity"). Thus, we'd expect it to excel compared with other DLA processors, and Tyr's initial ResNet-50 numbers are impressive. Nvidia and Qualcomm, however, are experienced chip companies that know how to optimize an SoC design. Vsora may lose some of its advantage if the components around the DLA are inefficient. Until it receives and tests the initial silicon, we won't know whether the company can meet its aggressive power target.

Software is a big challenge for all AI startups, and Vsora is no exception. Nvidia provides not only a broad set of development tools but also higher-level software for building and testing autonomous vehicles (AVs). The startup's approach is suited to customers that have their own software and are willing to port it to a new architecture.

Vsora's architecture features both AI and DSP engines connected to a large SRAM that facilitates rapid data sharing. The company believes this approach is critical for AVs, which must perform both functions—sometimes in the same algorithm. Most AV designers, however, seem willing to split AI and DSP between two separate chips. For those that want a single-chip solution, Qualcomm's AI 100 accelerator also checks all the boxes, although it lags Tyr in performance.

Companies typically struggle to sign new IP licensees once they announce plans to sell chips, owing to potential competition with those licensees. Vsora's primary licensee presumably approves of, and may have even initiated, the change to a chip strategy. But one customer isn't enough to support this expensive change. Vsora has developed a strong DLA design, but extending it into a family of chips will require a tremendous effort. The revenue from selling chips is much greater than from licensing IP, so if the startup can build a viable chip business, the payoff will be equally big. ♦