The Linley Group
Tech Insights
MICROPROCESSOR report
Insightful Analysis of Processor Technology

SAKURA DEBUTS FOR EDGE AI

Silicon Implementation Uses DNA IP for Low-Latency Inference

By Bryon Moyer (April 25, 2022)

EdgeCortix has made a strategic shift from selling AI intellectual property (IP) to selling its own edge-AI inference chips for line-powered systems. The new die, dubbed Sakura, started as a test chip, but the company says customer interest convinced it to offer the chip as a product. The chips come mounted on one of two cards: a dual-M.2 (M-key) card and a low-profile PCIe card.

Based in Japan, EdgeCortix has supplemented its $13.5 million in total funding with revenue from FPGAs and ASICs that employ its soft IP. It targets perception—vision, lidar, and related technologies—for transportation, augmented/virtual reality, industry, smart cities, and drones.

Sakura, revealed first at the recent Linley Spring Processor Conference, implements the company's dynamic neural accelerator (DNA) engine, adding on-chip SRAM, two LPDDR4X ports, and I/O. The chip has no host CPU, so it operates under the control of an external host. Sakura has a maximum performance of 40 TOPS; on ResNet-50, it achieves 0.4ms latency at 4.7W, yielding 533 inferences per second per watt (IPS/W). The company plans to ship samples on a development board in July, with production anticipated in 1Q23.

EdgeCortix announced its DNA architecture as IP last year. Although its primary focus will be on selling chips, it intends to entertain IP business opportunistically. The company may also license Sakura hard IP for chiplets. It's open sourcing the front end of its Mera compiler so future formats will be available to the tool.

DNA Blossoms in Sakura

Sakura combines six engines for standard 3D convolutions, six engines for 2D depth-wise convolutions, and three vector engines for activations and other miscellaneous operations, as Figure 1 shows (see *MPR 4/26/21,* "EdgeCortix IP Lowers AI Latency"). Standard convolutions combine all three color channels, whereas depth-wise convolutions separate the channels, combining results later to reduce computing operations. To boost utilization, the Mera tools can convert between standard and depth-wise convolutions to keep all engines occupied. Together, the engines are capable of 25,000 MACs per clock cycle.

Within a vector engine, dedicated blocks implement activation functions. Each block is configurable for a range of variants in a family of activation functions. Although this approach yields better performance, it's less flexible than a fully programmable engine. A 20MB on-chip SRAM can hold weights and activations or act as a scratchpad. The run-time software reconfigures the network-on-a-chip (NoC) on the fly, reassigning engines as part of EdgeCortix's focus on achieving high hardware utilization.

Sakura provides a PCIe Gen3 interface to connect to the host processor. It lacks other I/O, including a direct

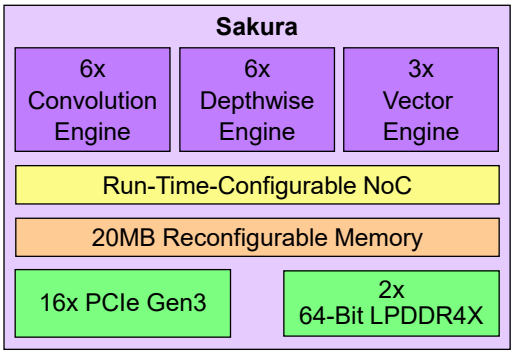

Figure 1. Sakura block diagram. The bulk of the inference work happens in the convolution and depth-wise engines. Vector engines handle other math, including activation functions.

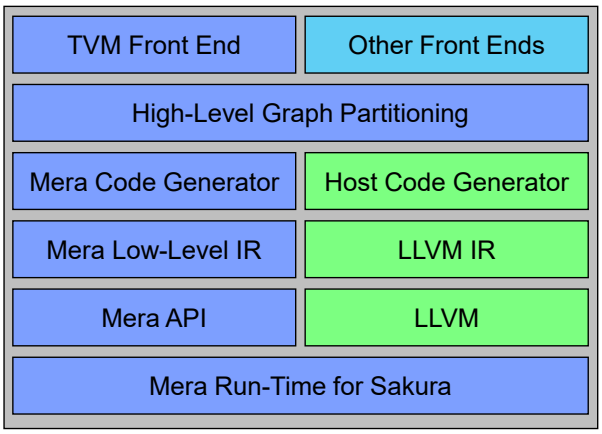

Figure 2. EdgeCortix Mera software stack. Blue boxes are Mera-specific functions; green boxes reflect host code. TVM is an Apache AI framework. The company intends to open-source the front end for future formats. It can also customize the compiler and run time for non-DNA architectures.

camera interface such as MIPI. Typical systems connect one or more cameras to the host processor, which feeds the uncompressed video streams across the PCIe bus to the inference chip. Eight 1080p video streams consume 1.25GB/s—roughly two PCIe lanes.

Designers can also link multiple Sakura chips across the PCIe bus. The 16-lane interface, however, provides only 16GB/s for chip-to-chip communication, host-processor traffic, and uncompressed video streams, so this approach is unlikely to scale beyond a handful of chips. Because PCIe isn't memory coherent, software divides the neural network and manages all data transfers among the chips.

Chips, IP, Tools, or All Three?

EdgeCortix has three separate business lines—a tall order for a small company. The original business offered DNA IP and FPGA implementations of that IP. The company has deprioritized but not eliminated this line in favor of another: chips. It's also considering whether to add a chiplet interface and license the Sakura hard IP for chiplets (see *MPR 3/28/22,* "UCIe Addresses Chiplet Interconnect"). Sakura may compete with chips made by some IP customers, softening interest in the IP for certain applications.

The third EdgeCortix business line is software: the company sells its Mera tools independently of the DNA IP and Sakura chip. It can customize the compiler and run time to architectures other than DNA. Renesas is one tool customer, with Mera compiling to an unspecified AI architecture (which "may or may not" contain EdgeCortix IP) in the RZ family of AI-accelerated microprocessors.

For Sakura, the tool flow requires no hardware-specific model retraining; as Figure 2 shows, the tools compile any model as is. The front end accepts a variety of input formats, and the open-sourcing move lets others adapt it for new ones. Mera has its own low-level intermediate representation (IR) for Sakura code and uses LLVM for host code.

Sakura Numbers Bear Fruit

Sakura targets applications with 5 – 20W power, where the most obvious competitor is from Hailo (see *MPR 6/24/19,* "Hailo Illuminates Low-Power AI Chip"). Nvidia's Orin aims at similar workloads, although it's a complete SoC with higher power (see *MPR 4/18/22,* "Nvidia Orin Appears in MLPerf"). Qualcomm's Cloud AI 100 chip also plays in this realm, at least in its smaller form factors (see *MPR 10/12/20,* "Qualcomm Samples First AI Chip"). Even though the company markets its chip for the cloud, edge servers are in range as well.

In this group, Orin has the highest peak AI performance (TOPS), as Table 1 shows, although we halved its advertised value because Nvidia's public claim assumes sparsity. The full Cloud AI 100 has a top capacity of 400 TOPS, but we used the 15W dual-M.2 configuration, which competes directly with one of the Sakura configurations; Qualcomm rates that unit at 70 TOPS. In addition, Qualcomm has by far the most on-chip memory, enabling it to better handle larger models. Critically, Hailo-8 omits a DRAM interface, necessitating either multiple-chip clusters for models too big to fit into SRAM or use of the PCIe interface to retrieve additional parameters. Neither the EdgeCortix nor Qualcomm part has a dedicated camera input, requiring the host processor to deliver video streams.

	EdgeCortix Sakura	Hailo Hailo-8	Nvidia AGX Orin 32MB	Qualcomm Cloud AI 100 Dual M.2e
Main CPU	None	1x Cortex-M4	8x Cortex-A78	None
DLA Architecture	DNA	Custom	Ampere	Custom
Peak AI Perf (INT8)	40 TOPS	26 TOPS	100 TOPS	70 TOPS
On-Chip Memory	20MB	32MB	4MB	72MB*
DRAM Interface	2x 64-bit LPDDR4X	None	4x 64-bit LPDDR5	2x 64-bit LPDDR4X*
Camera I/O	None	2x MIPI	6x MIPI	None*
Power (TDP)	10W	9W*	40W	15W
IC Process	TSMC 12nm	16nm	TSMC 7nm	TSMC 7nm
Production	1Q23 (est)	4Q21	2H22 (est)	3Q21

Table 1. Sakura versus competitors. Orin and the Cloud AI 100 target larger and higher-power applications than Sakura. Hailo-8 is a coprocessor, like Sakura, but it has a small CPU. (Source: vendors, except *TechInsights estimate)

The only benchmark data available for all four companies is ResNet-50 with batch=1. We've normalized latency, power, and power efficiency (IPS/W) to illustrate relative performance, as Figure 3 shows. Sakura has the lowest latency; Hailo-8 has the lowest power. But Sakura's latency is better by a sufficient margin to lead in power efficiency. Although a single batch at a time is typical for some edge applications, others pull in multiple camera streams, allowing largest batch

sizes. Orin and Cloud AI 100 have plenty of MACs to handle multiple streams; a single-stream benchmark therefore underutilizes them.

On this model, Sakura utilizes roughly 50% of the available TOPS. EdgeCortix touts high utilization as a strength, saying it can process more streams thanks to the reconfigurable interconnect that makes hardware available as needed when the workload balance changes. But the company withheld data that would demonstrates utilization at the high end.

Sakura Excels at Inference in Isolation

Numerous chip companies are trying to make their mark on the AI edge. Most have yet to ship a product, giving Nvidia a solid lead—especially considering the availability and maturity of their software. EdgeCortix has already sold its DNA IP to customers, but IP customers are chipmakers; they're system builders, and EdgeCortix is a latecomer to that market.

For ResNet-50, Sakura boasts less than half the latency and 25% better power efficiency than Hailo-8. The latter advantage is a result of its low latency rather than low power; Hailo-8 uses roughly two-thirds of Sakura's power. Orin power is much higher than Sakura, but Nvidia equipped it to process video streams from six or more cameras at once; it fares less well on single-stream workloads. Although EdgeCortix targets perception, Sakura lacks a camera interface as well as the computer-vision, media, and display capabilities that Orin has. Absent a direct connection, cameras must feed their data to Sakura through the PCIe interface, which is less efficient than a direct MIPI port. An external host processor with these capabilities would add considerable power and cost, diluting Sakura's advantage.

Price and Availability

EdgeCortix has yet to set pricing. Samples are scheduled to appear on development boards in July 2022, with production following in 1Q23. For more information, access www.edgecortix.com.

EdgeCortix's business model will be a challenge to fulfill, since it divides resources among hard silicon, soft IP, and Mera tools—all of which may have different customers. The company must demonstrate that it can keep all those plates spinning while posing no competitive threat to its IP customers. If it can do so, Sakura can be a serious contender in the edge-AI market. ♦

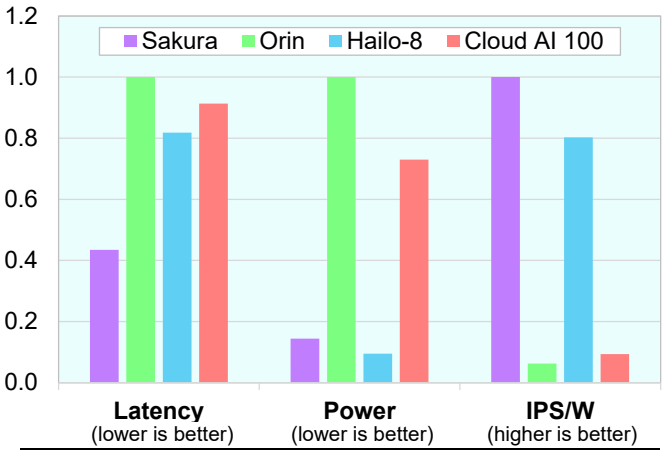

Figure 3. ResNet-50 Comparison for batch=1. Sakura leads in latency and power efficiency (IPS/W). Hailo-8 uses less energy per inference, but the longer latency hurts efficiency. The Orin and the Cloud AI 100 target larger designs. (Data source: vendors)