The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

EdgeCortix IP Lowers AI Latency

Startup's High-End DNA IP Targets Level 3 ADAS

By Aakash Jani (April 26, 2021)

EdgeCortix targets the high-end edge market with its custom intellectual property (IP), delivering class-leading efficiency through a dual-accelerator approach. In simulation, the IP achieves top scores in among Level 3 ADAS accelerators for batch=1 performance on ResNet-50 and Yolo v3, with equally impressive power efficiency. The startup has yet to disclose any customers, but it is profitable thanks to its IP and FPGA businesses.

To demonstrate this performance, the company is developing a test chip that it estimates will execute 54 trillion operations per second (TOPS) in an 8W power envelope, as disclosed at the recent Linley Spring Processor Conference. The design reaches this peak throughput using an 800MHz clock and eight convolution engines. The chip will be manufactured in TSMC's 12nm process; EdgeCortix expects first silicon early next year.

The startup is headquartered in Tokyo and has a design center in San Francisco. It has raised a mere $6 million through two funding rounds, one for US investors and the other for Asian investors. Although profitable, it's now raising Series A funds to allow it to grow more quickly. EdgeCortix employs approximately 20 people led by CEO and founder Sakyasingha Dasgupta, who previously led an AI-optimization team at IBM Research.

EdgeCortix began as an IP licenser, combining its DNA-branded deep-learning-accelerator (DLA) IP with Xilinx FPGAs for initial customers in the automobile, drone, and robot markets. Although the IP is highly configurable, the company offers predesigned solutions, as Table 1 shows. At the base of the product stack is the DNA-A050, which delivers 1.8 TOPS from a single convolution engine. Above it are the A100, A200, A400, and A800.

In parallel with its IP business, the company runs a profitable FPGA vertical, winning designs in the embedded (sub-25W) and edge-server (over 50W) markets with its F-series, which operates at 300MHz. That family mirrors the company's IP line, except for the F200 and F400 lines, which come in two variants: v1 employs 64x64 systolic arrays with fewer convolution engines, and v2 has 32x32 systolic arrays with more convolution engines. The latter variant thus provides more parallelism for small networks. Unlike most other IP vendors and AI chip startups, EdgeCortix recently posted MLPerf 1.0 results for the DNA F200 using a Xilinx Alveo U50 FPGA card, showcasing the robustness of its software stack and the low latency achieved.

Convolution Engine Lowers Power

If the EdgeCortix IP is the DNA, its complementary base pairs are two different engines that form a complete accelerator. What the company calls a standard convolution engine employs a systolic array of processing elements (PEs) that can scale from 3x3 to 64x64 arrays. Licensees that use bigger models can opt for larger arrays to avoid the waste of splitting convolutions. But beyond a 64x64 configuration, the systolic array can become underutilized

	DNA-A800	DNA-A400	DNA-A200	DNA-A100	DNA-A050
Systolic Arrays	8 arrays	4 arrays	2 arrays	1 array	1 array
Internal Memory	30MB	16MB	10MB	9MB	3.5MB
MACs per Cycle	67,000	34,300	17,100	8,600	2,200
Peak INT8 Perf*	54.0 TOPS	26.0 TOPS	13.0 TOPS	6.5 TOPS	1.8 TOPS
ResNet-50*†	2,500 fps	2,000 fps	1,250 fps	571 fps	159 fps
Yolo v3*†	555 fps	333 fps	167 fps	83 fps	27 fps

Table 1. EdgeCortix DNA IP. The DNA IP portfolio offers a range of products with widely differing total AI performance. The broad scope allows the company to target embedded and edge-server AI applications. *At 800MHz; †inference throughput at batch size of 1. (Source: EdgeCortix)

and provide diminishing returns for many models. EdgeCortix pairs each PE with two weight registers to boost utilization. Elements are summed by column and by row across the array and accumulate at the end.

A separate engine that the company call depth-wise convolution follows a vector approach to processing convolution operations. These engines directly implement depth-wise convolutions that use two-dimensional kernels (vectors) rather than the 3D kernel of standard convolutions, as Figure 1 shows. Neural-network models such as MobileNet and ShuffleNet employ the technique to reduce model parameters with little precision loss. For other models, a 3D kernel can be converted to 2D slices and fed to the convolution engine.

External weight values enter the convolution unit and are stored in the weight buffer, while external feature values are stored in the activation buffer. The feature and weight values then enter a different set of activation and weight buffers that feed the adder tree to produce a partial sum. The DNA IP is highly configurable. Licensees can increase the pipeline width and the number of parallel pipelines, but the former incurs considerable performance penalties. For every additional multiplication unit, the adder tree must also grow, decreasing the cycle efficiency of the pipeline. Additionally, wider convolution engines reduce utilization for smaller filter and kernel sizes, in turn decreasing power and area efficiency. For this reason, convolution engines are typically optimized for 3x3 kernels, the most common type. Both DNA engines support signed and unsigned INT8 multiply-accumulate operations with INT32 accumulators.

Separate vector units include multiple pooling, activation, and scaling engines, each of which is configurable in number. The pooling engine solely handles max pooling, while the activation engine accelerates a wider berth of activation functions such as ReLU, leaky ReLU, H-swish, and ReLU6. Lastly, the scaling engine processes bilinear up- and downsampling, enlarging or shrinking images, respectively, through linear interpolation. Parameter values for the activation and scaling engines reside locally in each subunit and arrive from either the systolic array or the convolution engine.

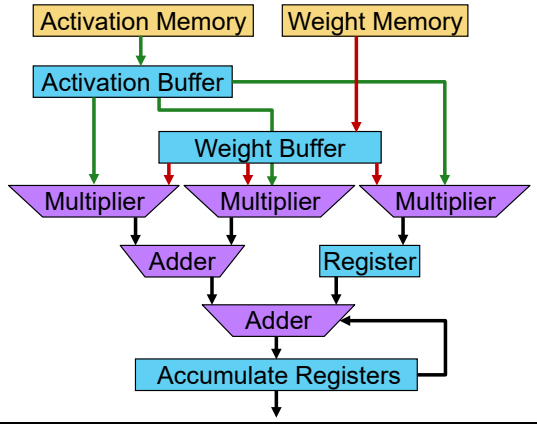

Figure 1. Convolution engine. The figure depicts a three-wide convolution engine with one pipeline. Licensees can increase the number of pipelines or the pipeline width, but the latter requires deepening the adder tree, decreasing efficiency.

Flexible NoC Halves Latency

As a DLA processes a neural network, it encounters layers with different channel and matrix sizes. Fixed-hardware structures vary in compute utilization from layer to layer. EdgeCortix pounced on the opportunity to differentiate its portfolio through reconfigurable interconnect and memory structures.

Once per layer, the DNA cores can change the data flow using circuit switching to optimize MAC utilization. Circuit switching, like its macroscopic counterpart, routes data to varying end points on the basis of "select" signals, which are configured for each layer during compilation. This approach lets the reconfigurable network on chip (NoC) reduce data-dependency stalls and increases tile-, kernel-, and model-level parallelism. It also does so by parallelizing low-rank kernels across multiple convolution engines or, for 3D kernels, by splitting the operation across the systolic array. For example, DarkNet-53's middle layers are one-sixteenth the size of the input image. In this case, the DNA core can reconfigure its NoC to shift from tile-level parallelism toward channel-level instead, keeping the engines busy.

Using the Cadence Xcellium Logic Simulation tool, EdgeCortix quantized the effect of its reconfigurable interconnect on power, performance, and area for two common neural networks: ResNet-50 and Yolo v3. It simulated five preconfigured IP cores: the A050, A100, A200, A400, and A800. For batch=1 inferencing, the A800 saw a 50% latency decrease relative to the version without a reconfigurable interconnect; the improvement declined for smaller configurations. The A800 has the most convolution engines, giving the compiler more granularity when reconfiguring resources.

The reconfigurability doubles the A800's throughput when running ResNet-50. That network's layers shrink exponentially after the initial image. Switching from the systolic array to the convolution engine dramatically reduces data-related stalls, boosting throughput. The power efficiency of EdgeCortix's IP peaks with the A400 and declines by 10% for the A800. We suspect the A800 is overqualified to handle ResNet-50, leading to poorer MAC utilization. Most customer models are more complex, however.

Lacking Camera Interfaces

Shifting from a purely IP/FPGA design house, EdgeCortix selected the DNA-A800 to power its test chip. As Figure 2 shows, that design features 30MB of on-die memory, enough to hold the entire ResNet-50 model in INT8 format. Most customer models, however, won't fit in this

memory, requiring it to swap layers in and out. Thus, the chip integrates two LPDDR4X DRAM controllers that drive 51GB/s of peak bandwidth.

EdgeCortix partitions the on-die memory into three virtual blocks, allocating contiguous arrays to weight, intermediate accumulation, and activation values. The size of these partitions is flexible and can be configured at runtime. The weight and data blocks connect directly to the LPDDR4X controllers through a load/store unit, which reduces external read and write latencies. The startup's test chip employs eight 64x64 systolic arrays and four convolution engines to accelerate MAC operations. To process auxiliary functions such as activations and pooling, it integrates four vector engines.

As a coprocessor, the chip must connect to a host CPU or SoC. The host connection uses the PCIe Gen3 interface, which has up to 16 lanes, although a small M.2 module would implement only 4 lanes. The chip lacks discrete interfaces for cameras and microphones, offering only GPIO. Instead, it relies on the host SoC to drive these peripherals. EdgeCortix is developing an SoC combining the DNA IP with a cluster of automotive-grade Arm Cortex-A CPUs along with MIPI-CSI camera interfaces and other peripherals.

The company developed its Mera compiler by extending the framework of the Apache TVM deep-learning compiler, which is open source. It supports post-training INT8 quantized models in Pytorch and Tensorflow Lite. It also accepts ONNX models. The open-source TVM compiler handles target-independent graph optimizations, with operators supported by DNA IP translated to a quantization-aware intermediate representation called quantized neural networks (QNN) that can execute on a built-in interpreter for functional simulation. Next, Mera performs DNA-specific optimizations such as constant folding, operator merging, and layer fusion to reduce external memory access. The operations are then scheduled and allocated to the hardware. Operations unrecognized by Mera are compiled using LLVM for the CPU, which is less efficient than the specialized hardware.

A800 Latency Leads by 5x

The test chip's closest competitors for machine-vision systems are Nvidia's Xavier and Hailo's Hailo-8 (see *MPR 6/24/19,* "Hailo Illuminates Low-Power AI Chip"). These three chips can power Level 3 ADASs, using their DLA to watch the road ahead. Hailo-8 employs a heterogeneous-resource map to closely track a neural network's compute needs, offering 26 TOPS of total AI performance. The A800 doubles that number, and we estimate it also leads in power efficiency (TOPS per watt), as Table 2 shows. We expect this lead to diminish for ResNet-50 efficiency, since both chips carefully manage convolution-unit utilization for real-world models.

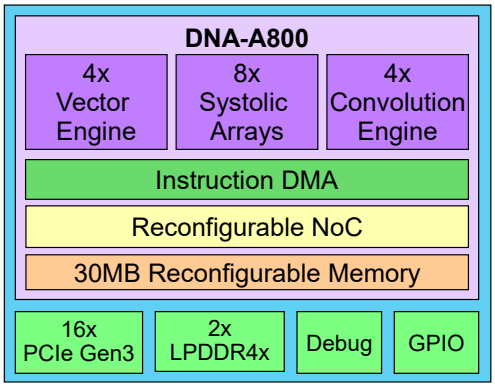

Figure 2. DNA-A800 block diagram. This coprocessor uses both systolic arrays and convolution engines to process neural networks. For pooling and scaling operations, it utilizes its collection of vector engines. The A800 test chip connects to the host through a PCIe Gen3 interface.

The A800's estimates top both Hailo-8 and Xavier in total performance and ResNet-50 latency. At a batch size of one, which is common for automotive applications that analyze real-time video, it inferences ResNet-50 five times as fast as the Nvidia part and consumes 73% less power based on simulated values. Xavier, however, comes with a high-performance host-CPU complex (see *MPR 2/19/18,* "Nvidia Xavier Drives to Carmel"). The A800 instead needs a discrete host processor, increasing system power. Even with a companion SoC, however, it would still lead in power efficiency given the current gap.

The A800 approximately matches Hailo-8 for on-die memory but differentiates itself with LPDDR4X support. Lacking a DRAM inference, Hailo-8 handles only smaller models. Since it's a test chip, the A800 lacks camera interfaces that are all but standard in machine-vision accelerators.

	EdgeCortix DNA-A800	Nvidia Xavier	Hailo Hailo-8
Main-CPU Type	None	8x Carmel VLIW	1x Cortex-M4
CPU Speed	0.8GHz	2.5GHz	Undisclosed
DLA Type	Convolution and systolic engines	NVDLA + GPU	Custom
Peak AI Perf (INT8)	54 TOPS	30 TOPS	26 TOPS
ResNet-50 Latency	0.4ms	1.5ms	1.5ms
On-Die Memory	30MB	Undisclosed	32MB
DRAM Interface	2x 64-bit LPDDR4X-3200	1x 256-bit LPDDR4-2166	None
Camera Interface	None	16x MIPI CSI	2x MIPI
Power (TDP)	8W	30W	7W*
Efficiency (INT8)	6.8 TOPS/W	1.0 TOPS/W	3.7 TOPS/W
IC Process	12nm	12nm	16nm
Production	2H22 (est)	3Q18	1H20

Table 2. Machine-vision-DLA comparison. The A800 blows away the competition by 5x in ResNet-50 latency. But it lacks a host CPU and camera interfaces, increasing system cost and power. (Source: vendors, except *The Linley Group estimate)

Price and Availability

EdgeCortix withheld pricing and royalty fees for its DNA IP portfolio. The company is now providing access to the DNA IP and Mera software for deployment on FPGAs and for benchmarking models using its simulator. It plans to sample A800 chips toward the middle of 2022. For more details, access *www.edgecortix.com*.

The company's Linley Spring Processor Conference presentation will be available at *www.linleygroup.com/SPC21* shortly after the conclusion of the event (registration required).

Both Xavier and Hailo-8 have MIPI interfaces, whereas the A800 must rely on its companion chip, which adds system-level image latency and cost.

Awaiting Hardened Silicon

For a newcomer, the EdgeCortix IP makes a splash in the machine-vision segment. The company leveraged its experience from initial FPGA deployments to license IP with a reconfigurable interconnect architecture and two convolution-engine styles to effectively address a variety of AI operations. Its IP-licensing business has helped it harden the software stack. EdgeCortix transposed an open-source compiler framework to its own architecture and tested it on customer silicon; few startups get the opportunity to actively test their software without the foreboding pressure of an upcoming tapeout.

To prove the capabilities of its IP, the company is developing a test chip that it expects will surpass Xavier and Hailo-8 in TOPS performance and streaming-inference latency while maintaining impressively low power. As a co-processor, it requires an external host SoC, but it leaves enough power headroom for system-level add-ons. EdgeCortix hadn't planned to bring the test chip to production, but the impressive specs are attracting interest from chip customers. The startup will need additional funding to support a full tapeout. And since the test chip is due to tape out late this year, production won't occur before late 2022. By that time, the competitive landscape will completely transform.

In the meantime, EdgeCortix will continue to license its IP for other chip designs. Although best suited to ADAS applications, the IP can be used for a variety of camera-based systems, including drone and robots. As an IP company, however, the company faces a wider field of competition, including Arm, Cadence, Ceva, and Synopsys. Each of these IP vendors has a longstanding history of providing trusted solutions, creating a high barrier for the startup to overcome.

EdgeCortix's DNA IP is a promising portfolio. Based on simulated results, we expect the unique architecture to deliver best-in-class latency. The IP is configurable and scalable to support a range of automotive and consumer applications. Potential customers can experiment with the DNA design using the company's FPGA systems and validate using next year's test chip. This combination should help the startup find additional licensees. ♦

This article is updated to reflect that the EdgeCortix test chip originally planned for 1Q21 has been delayed until 1Q22.