Software Is Critical to Edge-AI Deployment

By Mike Demler, Senior Analyst

September 2021

The Linley Group
Powered by TechInsights

About TechInsights

TechInsights is the most trusted source of technology analysis and market information for the semiconductor and microelectronics industry.

TechInsights leads the world in microelectronics reverse engineering. Our technology and market analysis informs leaders of the world's most innovative companies and enables the creation and monetization of intellectual property.

TechInsights monitors major industry events across consumer electronics markets and semiconductor devices. Our technology analysis starts with teardown – IC design wins and components costing – and spans every facet of design, from the system to the atomic level. Our market analysis adds context to today's developments and looks toward the future of equipment manufacturers, key industry players, and emerging technologies.

TECHNOLOGY INTELLIGENCE

TechInsights helps decision makers in semiconductor, system, financial, and communication service provider companies:

- Discover what products are winning in the highest-growth markets and why
- Spot or anticipate disruptive events, including the emergence of new players
- Understand state-of-the-art technology through independent, objective analysis
- Make better, faster product decisions with greater confidence
- Understand product costs and bill of materials

INTELLECTUAL PROPERTY SERVICES

We help IP Professionals in global technology companies, licensing entities and legal firms to:

- Build higher quality, more effective patents
- Identify patents of value and gather evidence of use to demonstrate this value
- Obtain accurate data for planning a potential defensive strategy or assertion case
- Make better portfolio management decisions to invest, abandon, acquire or divest
- Understand their competition, identify strategic partners, acquisition targets and business threats

Software Is Critical to Edge-AI Deployment

This white paper describes the importance of a flexible and robust software stack to edge-AI deployment. Processor vendors and intellectual-property licensers often tout the theoretical performance of their designs, but the neural-network compiler, run-time engine, and scheduler are just as critical to realizing that potential in production systems. The Linley Group prepared this paper, which EdgeCortix sponsored, but the opinions and analysis are those of the author.

Introduction

AI is rapidly moving out of data centers and into edge computing. Developers typically employ general-purpose CPU and GPU cores to develop and train their neural-network models, but those cores are much less efficient than purpose-built accelerators for infer-ence tasks. Although rapid growth in this market has led numerous entrepreneurs to launch edge-AI startups, few have produced complete solutions that include both the hardware and software necessary for production systems. Algorithm developers have published numerous neural-network models that are freely available online, and tech giants such as Amazon, Google, and Microsoft offer platforms that run those models in the cloud, but edge-AI deployment demands finer-grain optimization for each segment.

The edge-AI market comprises many diverse applications, from embedded devices to on-premises servers for enterprise, industrial, retail, and smart-city-management sys-tems. These applications cover a wide performance range, starting at less than a trillion operations per second (TOPS) in embedded and reaching tens or hundreds of TOPS in edge servers. To avoid employing a different platform for each segment, designers should choose an accelerator and software stack that scales to meet the varied require-ments of the segments they target.

The trend in edge servers is to offload AI from the CPU by adding a special-purpose accelerator on a PCIe plug-in card. FPGA-based accelerators have the advantage of configurability to support diverse applications, along with much shorter time to market than custom silicon. They also let users modify the architecture to install new models in the field as well as optimize the accelerator to handle different workloads. On the other hand, ASICs are better suited to embedded systems with fixed requirements and to high-volume consumer electronics.

Several intellectual-property (IP) vendors offer licensable AI accelerators for either ASICs or FPGAs, but few offer both. AI startup EdgeCortix is an exception. The com-pany developed a highly configurable hardware platform it calls the Dynamic Neural Accelerator (DNA), which delivers from 1.2 to 15 TOPS in an FPGA and up to 54 TOPS in an ASIC. For systems that need more throughput, designers can install multiple PCIe cards or connect multiple ASIC cores to an SoC's AXI bus.

But in addition to such a highly scalable architecture, the key to serving both ASIC and FPGA designs is the company's Multi-module Efficient Reconfigurable Accelerator (Mera) software. By combining Mera and the DNA cores, designers can target pro-grammable logic and custom silicon using the same tools. They can also benefit from Mera's cloud-based-platform support; an example is the Xilinx app store, which further accelerates deployment and enables over-the-air (OTA) updates to Internet-connected devices. The startup has proven its technology on Alveo and Zynq UltraScale+ FPGAs, reporting results on the MLPerf test suite, and it's developing a test chip to demonstrate DNA's performance in an ASIC.

Edge AI Requires Scalable Accelerators

The EdgeCortix DNA-F-series is well suited to edge servers, whereas the DNA-A-series addresses a broader range of ASIC power budgets and performance requirements. The underlying architectural features are the same in both products, however, giving custo-mers a highly configurable and scalable inference engine.

As Figure 1 shows, the DNA architecture scales by connecting multiple cores to a con-figurable interconnect bus. Within each core, the convolution engines execute most operations using INT8 data. The vector engines handle activation functions, along with pooling, sampling, and other nonconvolution layers. The composition and number of engines in a DNA design is configurable.

The size and number of the internal SRAMs is also configurable. DNA stores data and network parameters in three different memory blocks, but these blocks share the same physical resource. The AccMem is an accumulator that holds intermediate re-sults be tween activation and convolution operations. The DataMem stores activations, and as its name implies, the WeightMem stores the neural-network weights and bias values.

EdgeCortix offers five DNA-F models optimized for Xilinx Alveo and Zynq UltraScale+ FPGAs. The DNA-A cores are more customizable, and they support run-time configur-ability per layer or per inference, allowing reallocation of the physical storage to adjust the capacities of the three memory blocks. The inter-connect between the compute engines and memory is also run-time configurable, so the DNA cores can dynamically allocate resources for channel, kernel, model, and tile parallelism.

The Mera compiler automatically determines for each layer or model the optimum arrangement of compute blocks and memories, which the DNA cores dynamically con-figure through a circuit-switching technique. This run-time configurability is a unique feature of the Mera software and DNA architecture, ensuring all neural-network models benefit from maximum hardware utilization and minimum inference latency.

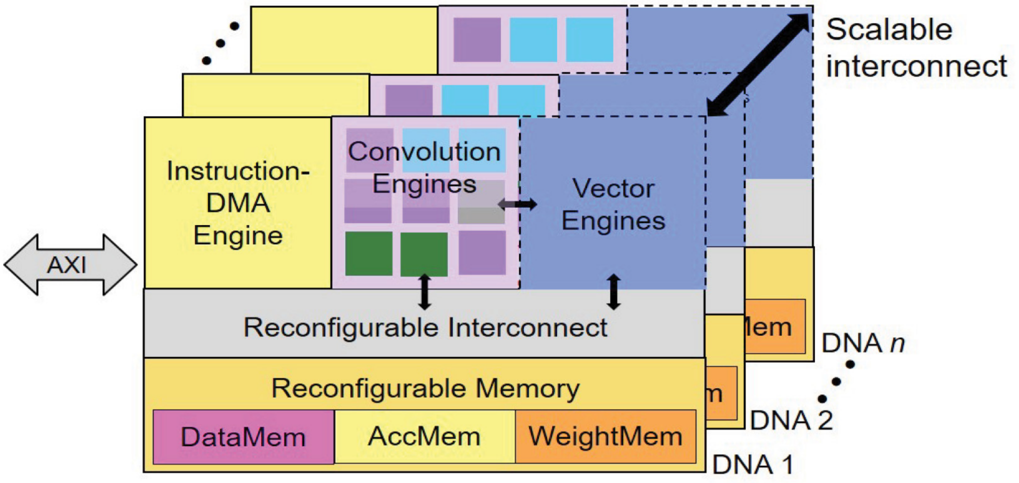


Figure 1. EdgeCortix DNA architecture. Each DNA core includes two convolution engines: one for point-wise operations and the other for depth-wise operations. The vector units handle most common activation functions, as well as pooling, sampling, and other nonconvolution layers. The architecture allows configuration of the various engines' number and type, as well as connection of multiple cores to scale performance.

Seamless Flows Link Training to Inference

Neural-network developers typically train their models in machine-learning frameworks such as Pytorch and TensorFlow, using full-precision floating-point (FP32) calculations running on CPUs and GPUs. All DLA-IP vendors offer software-development kits (SDKs) for compiling pretrained models to run on their inference engines, but to maintain accuracy, many require additional optimization or retraining that delays deployment.

In comparison, the Mera software stack directly compiles and runs models built in Pytorch or TensorFlow Lite without requiring additional post-training optimization. EdgeCortix created Mera by extending the Apache Software Foundation's TVM deep-learning compiler, a popular open-source software stack. Developers using TVM benefit from the contributions of major AI-technology companies, including Amazon, Facebook, Google, and Microsoft.

Figure 2 shows a high-level view of the Mera software flow and its integration with the machine-learning frameworks. Developers run Mera using C++ or Python scripts. Functions built into Pytorch and TensorFlow Lite can quantize network parameters to INT8 format for inference. The open-source ONNX exchange format allows conversion of neural-network models trained in other frameworks to TensorFlow Lite.

After quantization, the TVM front end handles the initial high-level graph partitioning, which is independent of the target inference engine. Using the Mera extension that EdgeCortix developed, the compiler translates supported operators from the open-source Relay intermediate representation (IR) to DNA instructions. It also detects unsupported operators, translating them to the LLVM IR for execution on the target hardware's Arm- or x86-based host CPU.

Once graph partitioning is complete, Mera's code generator lets developers perform a quick functional simulation using the Mera interpreter. The functional simulator provides a check to ensure the compiled model matches the original pretrained version without losing accuracy. After functional verification, Mera performs low-level optimi-zations and partitioning for the specific target. Optimizations include fusing layers and operators to maximize throughput as well as efficiently tiling operations to match feature-map dimensions. Developers can select as a target the built-in performance simulator, which accurately estimates the network latency. They can also use the RTL code that Mera generates in Verilator or in another open-source cycle-accurate simulator.

Mera's scheduler is a key to DNA's efficiency. It maximizes utilization by taking advan-tage of network parallelism, distributing the workload on the basis of the hardware configuration. The software minimizes latency for batch=1, which is typical in real-time object recognition. Depending on their objectives, developers can choose either a fast or slow scheduler. The former lacks the low-level optimizations of the latter, but it provides quicker proof that the network compiles properly; for some customers, that's enough for deployment.

Because FPGAs such as the Xilinx Alveo family ship with various amounts of DRAM, and because some include High Bandwidth Memory (HBM), the slow scheduler can take advantage of those details by performing finer-grain optimizations than the fast scheduler. This feature enables greater flexibility than compilers that only optimize for the resources built into the accelerator cores.

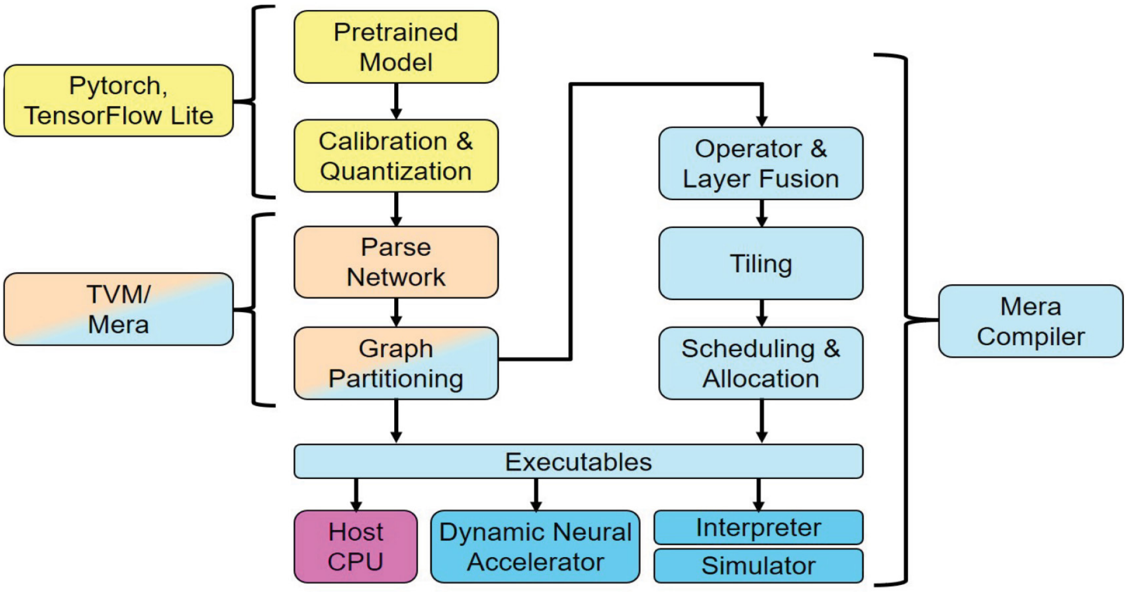
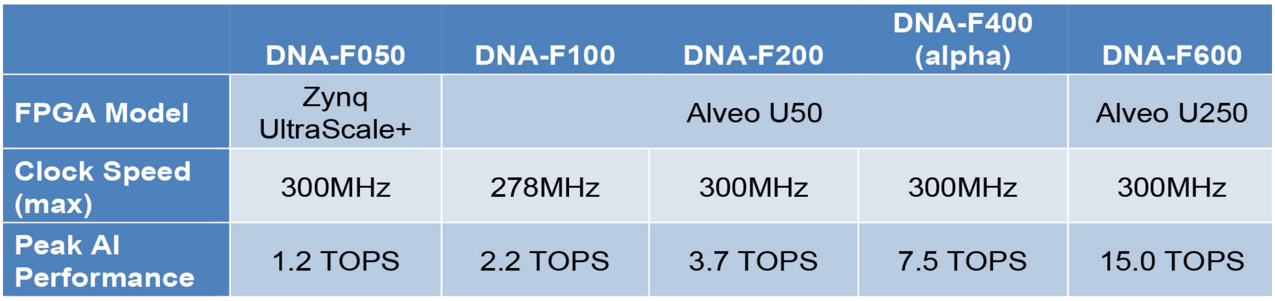


Figure 2. Mera compiler flow. Mera supports models trained in Pytorch or TensorFlow Lite. The quantization tools in those platforms convert weights to the INT8 format used for inference. Mera optimizes the model for the target hardware. It includes a functional simulator that verifies accuracy and a dynamic simulator that estimates latency, and it works with Ventilator as well as other open-source cycle-accurate simulators.

Cloud Platforms Reduce Time to Market

For its ASIC products, EdgeCortix uses a traditional IP-licensing model that includes an upfront licensing fee along with royalties based on the number of chips sold. For FPGA-based customers, however, it offers a cloud-based software-as-a-service (SaaS) model, which accelerates deployment. Regardless of the target hardware, though, the Mera software stack is the same.

Customers can select from five DNA-F products optimized for Xilinx PCIe cards, as Table 1 shows. The F050 works with Zynq UltraScale+ SoCs that integrate Arm CPUs. Designers can either employ Zynq's programmable logic to stream video directly to the DNA cores or use the chip's Cortex-A53 CPUs as a host processor. The F100, F200, and F400 are optimized for Xilinx U50 PCIe cards, delivering AI throughput ranging from 2.2 to 7.5 INT8 TOPS. The F600 is optimized for the Alveo U250; running at 300MHz, it achieves 15 TOPS.



	DNA-F050	DNA-F100	DNA-F200	DNA-F400 (alpha)	DNA-F600
FPGA Model	Zynq UltraScale+	Alveo U50			Alveo U250
Clock Speed (max)	300MHz	278MHz	300MHz	300MHz	300MHz
Peak AI Performance	1.2 TOPS	2.2 TOPS	3.7 TOPS	7.5 TOPS	15.0 TOPS

Table 1. EdgeCortix DNA-F-series. The DNA-F050 targets Zynq UltraScale+ FPGAs, which include Arm Cortex-A CPUs that can serve as the host processor. Designers can compile the F100, F200, and F400 to run on Alveo U50 FPGAs; the F600 targets the Alveo 250. The Alveo cards have PCIe interfaces for connection to an x86 or Arm host.

Designing a system using FPGAs historically required RTL-programming skills, but EdgeCortix makes that task much easier by packaging the DNA-F bit streams and the corresponding Mera software in ready-to-use Docker containers, as Figure 3 shows. The SaaS model gives customers a renewable subscription locked to each device. The F100 and F200 are available for cloud or on-premises deployment via the Xilinx app store, as well as for cloud deployment on the Nimbix platform. By the end of 2021, EdgeCortix plans to offer the DNA-F products on other cloud platforms, such as AWS and Microsoft Azure.

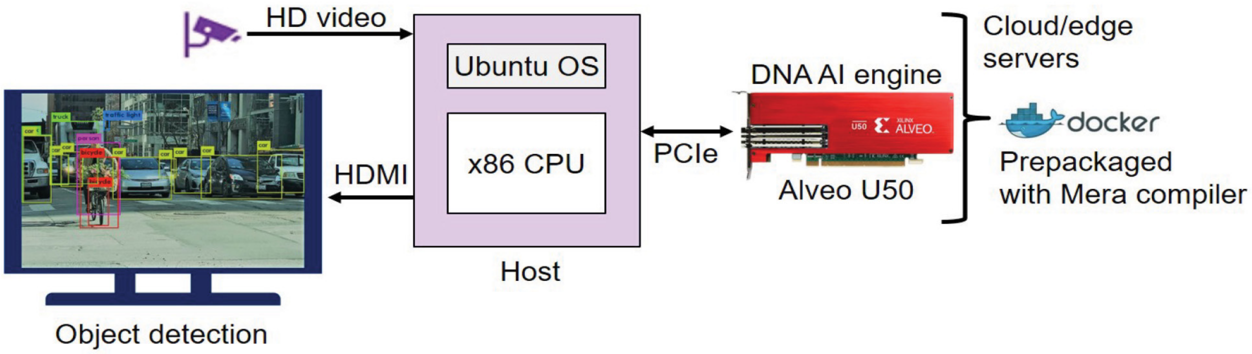


Figure 3. DNA computer-vision system. In this example configuration, the host processor streams images at HD resolution to the Alveo PCIe card. Designers can easily install the DNA accelerator in the FPGA by downloading Docker containers that include the hardware-specific bit streams as well as the Mera compiler running on the host processor.

Standard Benchmarks Validate Performance

Edge-AI-accelerator vendors often specify inference throughput only as the total num-ber of multiply-accumulate (MAC) operations or TOPS, because those operations can represent 90% or more of a computer-vision network's computations. But hardware utilization in most accelerators is typically less than 50%, and it varies by model, yield-ing much less throughput than the data sheets imply. We therefore advise customers evaluating such devices to test their own models, or at least request that the vendor provide results using standard publicly available benchmarks.

EdgeCortix demonstrated the capabilities of its edge-AI platform by benchmarking the DNA-F200 design on a variety of popular neural networks, including samples from the industry-standard MLPerf test suite. Because the DNA cores target computer-vision systems that perform real-time object recognition, it employed the batch=1 processing typical of such applications. The company also optimized the models to minimize latency, a critical factor in automotive and industrial systems.

As Figure 4 shows, the published EdgeCortix benchmark results include a variety of MobileNet and ResNet configurations, along with more-complex Yolo v3 models. MobileNet v2 and ResNet-50 are two components of MLPerf Inference v1.0, which requires that vendors demonstrate 99% accuracy on 50,000 images in the ImageNet validation database. Running ResNet-50 at 300MHz and batch=1, the DNA-F200 meets that criterion, delivering 152fps with just 6.6ms latency.

MobileNet requires only about 10% as many parameters as ResNet-50, enabling the accelerator to boost performance to 390fps with just 2.6ms latency. Customers should keep in mind that these MLPerf results came from an early v0.2 version of Mera; the production release is likely to perform better. After its MLPerf submission, the company ran ResNet-50 again using a later version of its compiler. As Figure 4 shows, latency improved to 4.4ms, and throughput increased by nearly 50%, from 152fps to 225fps.

Yolo v3 isn't part of MLPerf, but it's a more challenging network than ResNet-50, comprising 33 billion MAC operations in 106 layers, along with 62 million parameters. Its native input-image resolution is 416x416 pixels, but EdgeCortix demonstrated the DNA-F200's ability to accurately classify images by downscaling full-HD (1,920x1,080) video streams, delivering 25fps with 40ms latency. Reducing the input resolution to Yolo v3's native frame size increases throughput to 43fps, reducing latency to 23ms.

These benchmark results demonstrate Mera's versatility in compiling neural networks that span a 10:1 range of model sizes, as well as the DNA IP's ability to efficiently increase performance. For example, although Yolo v3 comprises more than 8x as many MAC operations as ResNet-50, the DNA-F200 executes the larger network with just 7x the latency, demonstrating Mera's ability to maximize hardware utilization.

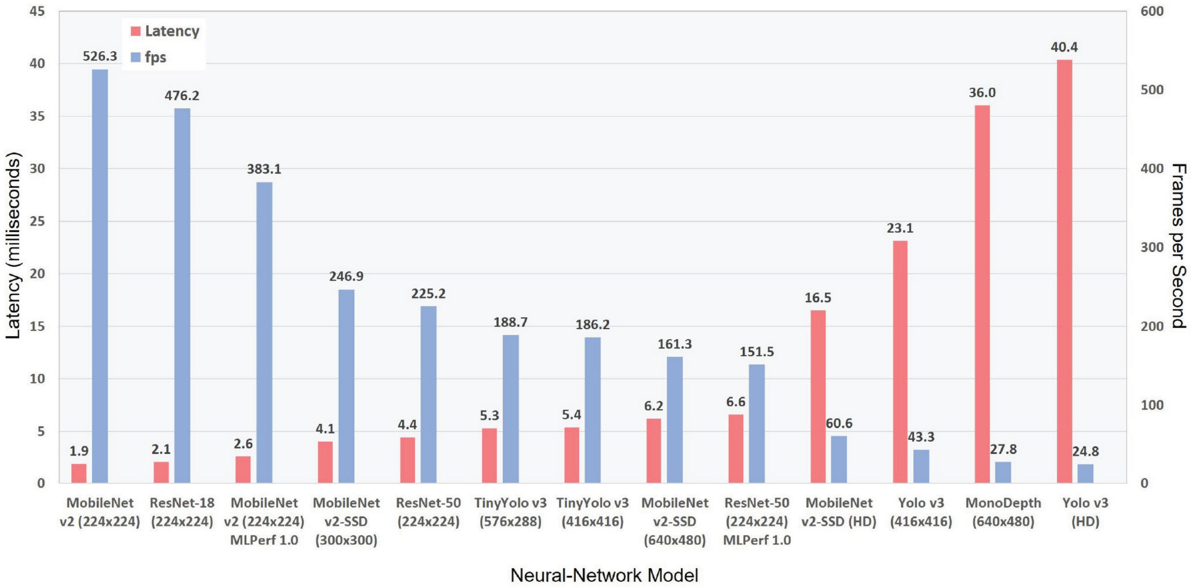


Figure 4. DNA-F200 batch=1 benchmarks. EdgeCortix has tested its DNA accelerators on computer-vision networks ranging from MobileNet v2, which comprises just 300 million MAC operations, to others that execute 100x as many operations. It published results for MobileNet v2 and ResNet-50 in compliance with MLPerf requirements, demonstrating at least 99% object-classification accuracy on samples drawn from the ImageNet database. (Source: EdgeCortix)

Conclusion

Because edge-AI devices have a vast range of performance and power requirements, choosing an accelerator that can support the different workloads and deliver the required throughput on popular computer-vision models can be a challenge. But the edge-AI software stack can be even trickier, because it must include a compiler capable of optimizing diverse pretrained models to achieve peak hardware utilization in different target devices.

The need for a seamless interface to standard training frameworks is a given. The soft-ware stack must also include a scheduler and run-time engine that meet the require-ments of latency-sensitive applications. Because algorithm developers are constantly releasing new models, the software platform should have a cloud component for delivering model updates to installed systems.

The combination of EdgeCortix's DNA IP and Mera software stack meets all of these requirements. The DNA-A-series is a good fit for ASIC designs with fixed requirements, but custom silicon takes much longer to bring to market than an FPGA. By employing Mera with Xilinx Alveo and Zynq PCIe cards, customers can immediately deploy neural-network models to their edge servers.

EdgeCortix is rare among startups for publishing benchmarks on a wide range of stan-dard models, from MobileNet to ResNet to Yolo v3. Many vendors are reluctant to dis-close their results, instead releasing a theoretical TOPS number that users never realize. By contrast, the DNA-F200 test results demonstrate excellent latency and throughput processing video at up to HD-resolution, making the design a strong candidate for real-time computer vision. EdgeCortix's hardware+software platform is a complete edge-AI solution well suited to industrial, retail, and smart-city-infrastructure systems.

www.techinsights.com
info@techinsights.com

The Linley Group
Powered by TechInsights