TechInsights
Microprocessor Report

Dylan McGrath September 6, 2024

Ceva NPU Core Targets TinyML Workloads

Ceva's NeuPro-Nano licensable neural processing unit (NPU) targets processors that run TinyML workloads, offering up to 200 billion operations per second (GOPS) for power-constrained edge IoT devices. In contrast to competitive NPU IP offerings aimed at the IoT edge, NeuPro-Nano can act as a stand-alone, self-contained solution for AI and machine-learning applications; it includes control and management functions and can be implemented without a host processor in some instances, saving die area.

AI has been moving to the edge of networks for reduced latency and bandwidth consumption and increased data security (MPR January 2020, "AI is Livin' On The Edge"). TinyML has emerged as a specialized field of machine learning that focuses on deploying models to the smallest, most power-efficient edge devices, typically running on tiny microcontrollers (MCUs) that cost a few dollars or less and can run for years on a small battery (MPR March 2020, "Deep Learning Gets Small"). TinyML typically runs on a processor's main CPU, but moving AI to an NPU is usually much more power efficient, extending the life of battery-powered devices.

The NeuPro-Nano comes to market at a time of rapid evolution for TinyML chips. Amid increasing machine-learning model complexity and heightened expectations for performance at the edge, more TinyML chips are incorporating NPUs to increase the power efficiency of AI acceleration.

NeuPro-Nano is available for licensing now in two different configurations. It features a fully programmable, scalable architecture and supports several AI acceleration features such as sparsity acceleration, nonlinear activation types, and fast quantization. It expands the company's line of NPU IP, which now covers the gamut from power-constrained edge IoT applications all the way to performance-demanding applications in automotive, mobile, and generative AI that require several hundred times more AI compute horsepower. To support NeuPro-Nano, Ceva also expanded its NeuPro Studio AI software development kit (SDK), which supports open AI frameworks include TensorFlow Lite Micro and microTVM to help with development of TinyML applications.

AI on a Nanoscale

A growing number of applications demand on-device intelligence, necessitating extremely compact TinyML models. In consumer electronics, TinyML is used for wearables, smart home, and smart audio devices. In the industrial IoT space, TinyML devices are being used to collect and process data from various sensors to monitor the health of equipment for predictive maintenance, aiding in diagnosing the root cause of equipment failures by analyzing sensor data for fault diagnosis, simple visual inspection to inspect products for defects or inconsistencies, process monitoring, anomaly detection, defect classification, and other tasks. Battery-powered TinyML systems are also used in healthcare (for remote patient monitoring applications), agriculture (such as livestock monitoring devices), automotive, and smart cities.

Microcontrollers, sensors, and audio SoCs targeting edge AI are increasingly leveraging TinyML to execute models directly. Most TinyML models currently run on MCUs or DSPs, often in conjunction with hardware accelerators, to overcome the computational limitations of these general-purpose devices. However, we expect to see an increase in chips specifically designed for TinyML applications over the next few years.

Battery-powered TinyML chips must be optimized for low power consumption (often less than 10 milliwatts) and deployment on resource-constrained hardware. They require as much as 10 MB of flash, ROM, or RAM and less than 500 KB of code memory as well as computational resources for inference tasks. They also typically feature the ability to interface with various sensors for data collection and wireless connectivity for data transfers and updates. Some TinyML devices use hardware accelerators to accelerate computations. But today most TinyML applications run on MCUs, most of which lack AI acceleration.

High-Performance Quantization Unit

At a high level, the NeuPro-Nano core consists of five blocks, as Figure 1 shows: a control unit, a signal processing unit, separate data and program memory subsystems, and the NPU block.

The NPU block contains a tensor multiplication engine to perform tensor operations such as matrix multiplication for AI and convolution for digital signal processing. A special quantizer unit converts high-precision floating-point data into integers to reduce the model sizes.

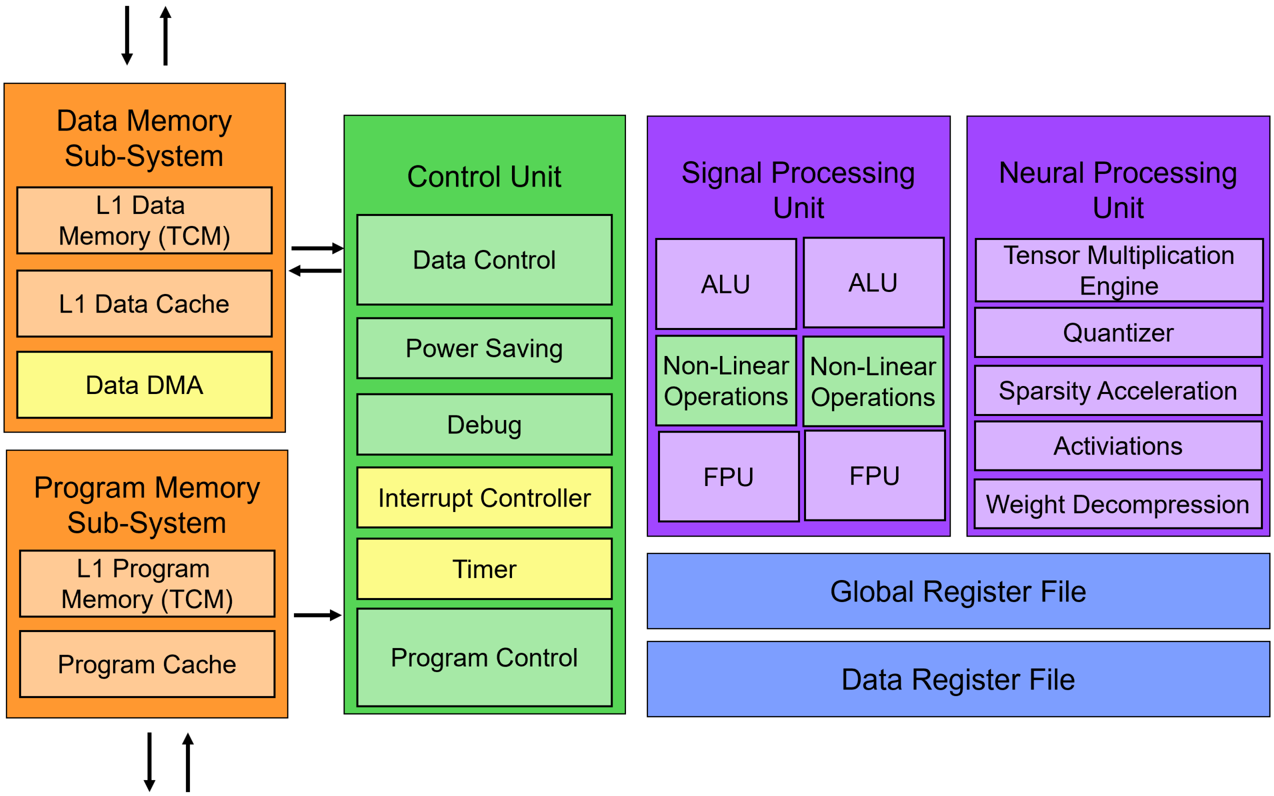

Figure 1 – NeuPro-Nano block diagram. The NeuPro-Nano's architecture differs substantially from Ceva's other licensable NPU cores, including the NeuPro-M. It is a more general-purpose NPU design that strives for a balance between data path cycle and compute cycle to enable optimized MAC utilization. (Data source: Ceva)

To handle sparsity acceleration, the NPU block features a sparsity unit similar to that in Ceva's NeuPro-M NPU (MPR January 2022, "Ceva Tackles Unstructured Sparsity"). However, the sparsity acceleration function in the NeuPro-Nano is more rigid than the sparsity engine in NeuPro-M (MPR September 2023, "NeuPro-M Enhances BF16, FP8 Support"). Sparsity support is comprised of two aspects: lossless memory compression, branded NetSqueeze by Ceva, which does not reduce the precision of the model, and performance acceleration. For performance acceleration, NeuPro-Nano requires the sparsity to be semi-structured and supports up to 2× acceleration. The NeuPro-M sparsity engine can work with additional levels of sparsity and can accelerate sparsity up to 4×.

Using proprietary mechanisms and algorithms, the NetSqueeze on-the-fly weight decompression technology removes zero weights to compress the memory footprint. NetSqueeze mechanisms work directly with the compressed weights, eliminating the need for any intermediate decompression stage, without any intervention from the software programmer. An advanced hardware prefetch mechanism reduces traffic overhead and maintains balance between the data path cycle and compute cycle, enabling higher MAC utilization.

The NPU block also houses weight decompression to optimize the storage and retrieval of neural network weights and the activation functions, including ReLU, sigmoid, and tanh.

Integrated Control Capabilities

Ceva claims that NeuPro-Nano derives a significant footprint advantage of up to 45% over competing solutions because its integrated CPU functionality can in some cases eliminate the need for a host CPU. Existing TinyML NPUs typically require a companion CPU or MCU to handle general-purpose tasks such as feature extraction, sensor interfacing, peripheral control, and system management. NeuPro-Nano, by contrast, can be implemented unaccompanied, capable of executing the neural network, feature extraction, and control code.

Many existing TinyML devices feature a Cortex-M or other CPU for control functions. Device vendors may prefer to continue using the same controller in future generations to avoid adding a software port. NeuPro-Nano can also be used in multicore implementations to handle neural network computations alongside a CPU such as the Cortex-M, which can continue to handle the general-purpose and system-management functions.

NeuPro-Nano’s scalability is an important attribute for a TinyML NPU operating at the edge. It supports up to 64 int8 MACs per cycle.

The architecture is also fully programmable to execute various neural network algorithms and operations. It supports the most common advanced machine-learning data types ranging from 4-bit to 32-bit precision. It natively supports transformer models.

Machine-Learning Model Optimizations

The NeuPro-Nano offers several enhancements to improve the efficiency and performance of machine-learning models. It implements sparsity acceleration techniques to reduce computations for weights in a neural network with zero values. It accelerates the performance of nonlinear activation functions like ReLU, sigmoid, and tanh.

Quantization techniques are increasingly popular in machine learning to reduce model sizes, decreasing memory and power requirements while increasing the speed of computation by converting high-precision floating-point numbers like 32-bit floats to lower-precision integer formats like 8-bit integers. The trade-off for quantization is that it reduces the accuracy of the model, though the amount of accuracy loss varies depending on several factors, including the specific quantization techniques used and the architecture of the model.

Ceva has developed a fast quantization technology that uses a dedicated proprietary instruction set architecture and natively supports fast approximate TFLM quantization, a technique for speeding quantization of TensorFlow Lite Models (TFLM) without sacrificing significant accuracy. Ceva claims that its implementation of these mechanisms accelerates the quantization of TFLM models by up to five times. In addition, the NeuPro-Nano implements an on-the-fly weight decompression tool that reduces the memory footprint required for models.

Standing Alone with Stand-alone Capability

Other leading IP vendors offer licensable NPU cores that target edge applications. The Arm Ethos-U55 (MPR March 2020, "Cortex-M55 Supports Tiny-AI Ethos"), Synopsys ARC NPX NPU (MPR April 2022, "Synopsys NPX6 Expands AI Options"), and the Cadence Neo NPU (MPR October 2023, "Cadence Boosts DLA Speed by 20×") are scalable IP cores with configurations that support a wide range of computational capability. Ceva offers separate licensable NPU cores that compete with these products at the higher end.

Of the three competitors, Ethos-U55 most closely resembles the NeuPro-Nano's performance and acceleration profile; like NeuPro-Nano the Ethos-U55 is a highly specialized NPU for edge AI and is specifically designed to accelerate machine-learning inference in area-constrained embedded and IoT devices (Arm offers the Ethos-U65 and Ethos-U85 for embedded AI applications that demand greater performance). Neo and NPX are scalable and cover a much greater range of AI performance and MAC acceleration capabilities to target higher-performance applications at the high end.

Among the products considered, only Ceva's NeuPro-Nano can be implemented as a self-sufficient solution with integrated CPU functions, as Table 1 illustrates. The Ethos-U85, ARC NPX, and Neo require a host processor to act as the central control unit, handling control and management functions, assigning tasks, transferring data, and controlling communication. As noted above, few customers are likely to implement NeuPro-Nano as a self-sufficient core, at least initially.

Like Ceva, Arm, Synopsys, and Cadence offer comprehensive software stacks and SDKs to simplify the development and deployment of TinyML applications.

	Ceva NeuPro-Nano	Arm Ethos-U55	Arm Ethos-U85	Cadence Neo	Synopsys ARC NPX	Ceva NeuPro-M
AI Performance	10 to 200 GOPS	64 to 512 GOPS	256 GOPs to 4 TOPs	600 GOPS to 80 TOPS	Up to 440 TOPS	32 to 256 TOPS
Peak Performance Clock Speed	1.5 GHz	1 GHz	1 GHz	1.25 GHz	1.3 GHz	2 GHz
AI Acceleration Throughput Per Cycle (int8)	32 to 64 MACs	32 to 256 MACs	128 to 2,048 MACs	256 to 32,000 MACs	1,000 to 96,000 MACs	8,000 to 64,000 MACs
MAC-Array Data Types	int4, int8, int16, int32, float32, FP16	int8, int16	int8, int16	int4, int8, int16, FP16	int4, int8, int16, FP16, bfloat16	int4, int8, int16, int32, FP8, FP16
Transformer Support	Yes	Yes	Yes	Yes	Yes	Yes
Host Processor Required?	No	Yes	Yes	Yes	Yes	Yes

Table 1 – Licensable NPU IP comparison. NeuPro-Nano optimized for TinyML workloads, slots in at the low end of Ceva's IP portfolio, while the NeuPro-M packs substantially more AI performance for applications such as computer vision and ADAS. Arm's Ethos product line also scales to higher levels of performance, as do Cadence's Neo and Synopsys' NPX. (Data source: vendors)

Small but Growing Opportunity

Ceva's NeuPro-Nano is specialized for resource-constrained edge devices running TinyML workloads but incorporates variations on many of the features available in Ceva's more powerful NPUs like NeuPro-M as well as energy-efficiency optimization techniques like on-the-fly energy tuning, sparsity acceleration, and dynamic voltage acceleration.

Most current IoT chips that run AI run it in the cloud. Among those chips that do some AI on the device, most run the AI in software on the CPU rather than dedicate additional die area to an NPU. Battery-powered IoT systems, however, derive critical power savings from the implementation of an NPU, which provides a more efficient way to execute neural networks. As a result, there is an increasing need for NPUs in edge IoT, especially for battery-powered systems. But while the market may be growing, it is also rife with competition. Due to the low cost of developing and offering a simple AI core, NeuPro-Nano faces competition from the rival products from established IP vendors discussed above as well as AI IP startups with different capabilities, and area, performance, and power consumption profiles.

NeuPro-Nano is distinct among licensable NPU cores for TinyML devices in that it can be implemented as a stand-alone core that not only executes neural networks but also the control code, DSP code, and feature extraction. While this sets NeuPro-Nano apart, we expect this capability to appeal only to a relatively small subset of the market. Many architectures will continue to require a host CPU due to the NeuPro-Nano's relatively limited performance. Chip companies are generally reluctant to port their Arm software to a different architecture, particularly a brand-new one with limited and unproven software tools and OS support (note the limited uptake of RISC-V in MCUs, despite years of trying). Those customers that are willing to make this leap, however, will benefit from the smaller footprint of implementing a stand-alone NPU with integrated CPU functionality, which Ceva estimates can save up to 45% die area compared to the combination of an NPU and host processor.

In summary, NeuPro-Nano enters a relatively small but growing market for NPUs in chips for battery-powered IoT edge systems with a distinct value proposition. While the market opportunity for a self-sufficient NPU core with integrated CPU functionality appears limited in the near term, it could accelerate with demand for edge AI. For applications where AI is the primary function, MCU vendors may create products that lack a traditional Arm-based host CPU and rely solely on an NPU such as NeuPro-Nano. Keeping a close watch on the leading MCU vendors should offer a clear indication of how quickly the self-sufficient NPU approach is taking off.

Learn more

- An early example of an edge AI device without a significant CPU core is the Syntiant NDP250: Syntiant's NDP250 Boosts Memory.
- Interested in exploring the market for datacenter AI chips? Have a look at TechInsights' AI Chips and Accelerator Forecast.
- Looking for a report on Arm's Ethos-U85 NPU? Refer to Ethos-U85 Adds Transformer Support.
- Want more details on the NPU IP market? See Synopsys NPX6 Expands AI Options.
- Need more background on TinyML? Check out Deep Learning Gets Small.