The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

CEVA NEUPRO ACCELERATES NEURAL NETS

New IP Comprises a General-Purpose Machine-Learning Processor

By Mike Demler (January 29, 2018)

Ceva has extended its line of neural-network-processor intellectual property (IP) by launching NeuPro, which comprises a family of cores optimized for general-purpose machine learning. It previously developed the XM-series of deep-learning accelerators (DLAs), but whereas they use an architecture optimized for computer vision (CV), the NeuPro ISA supports any machine-learning application and doesn't require a separate host-CPU controller.

As Figure 1 shows, NeuPro employs two separate cores: the DSP-based NeuPro Vector Processor Unit (VPU), and the NeuPro Engine. The VPU plays more of a supervisory role while offloading most of the computation to the NeuPro Engine. Although the company withheld details, the VPU is similar to its previous DSP-based CV cores. It includes a scalar unit and vector unit that run the deep-neural-network (DNN) control code, and it's programmable for handling layer functions that the NeuPro Engine omits.

The NeuPro Engine is the new element in Ceva's architecture. Whereas the XM6 uses the vector DSP to run the normalization, pooling, and other layers in a convolutional neural network (CNN), NeuPro shifts those functions to specialized hardware in the NeuPro Engine. The XM6 employs a separate accelerator that only runs CNN multiplier-accumulate (MAC) functions, but the NeuPro Engine includes hardware that accelerates all layers. Compared with its predecessor, the new design offers a more streamlined data flow that reduces external-memory transactions, which often bottleneck neural-network performance. Keeping the DNN data local to the execution units also saves power.

NeuPro's target applications include advanced driver-assistance systems (ADASs), augmented-reality (AR) headsets, drones, smartphones, and surveillance cameras. The IP will become available to lead customers in 2Q18; the company plans a general release in 3Q18.

A Fine-Tuned Engine

As Figure 1 shows, the two NeuPro cores communicate through an AXI bus interface. The NeuPro Engine comprises a set of specialized execution units designed to handle number-crunching tasks for all popular neural-network layers. It has a convolution controller linked to an array of matrix MAC units, which run the fully connected and convolution layers that make up most CNN calculations. The results of one MAC flow immediately to the next unit in the matrix, avoiding the need to store the result in a register (or even DRAM) and then fetch it again.

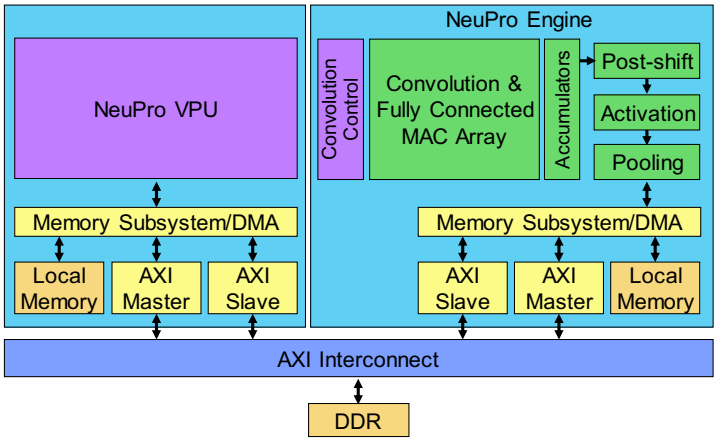

Figure 1. Ceva NeuPro deep-learning accelerator. The new design has two cores: the NeuPro VPU and the NeuPro Engine. The former issues a sequence of DNN tasks to the latter, which executes a streamlined data flow that increases performance by reducing external-memory transactions.

Price and Availability

Ceva plans to begin sampling NeuPro RTL to its lead customers in 2Q18; general release is scheduled for 3Q18. The company doesn't disclose pricing. For more information, access *www.ceva-dsp.com/product/ceva-neupro.*

This data flow is identical to the approach of Google's tensor processing unit (TPU), which integrates 64,000 MACs that operate on 8-bit integers. Whereas the search giant uses TPUs to accelerate DNNs in its data centers, NeuPro is a much smaller AI processor designed for client devices (see *MPR 5/8/17,* "Google TPU Boosts Machine Learning"). Ceva offers the NeuPro Engine in 512/1,024/2,048-MAC (NP500/1000/200) and 4,096-MAC (NP4000) configurations. Each unit can perform an 8x8-bit MAC in a single cycle, but the array is programmable to perform 16x8-bit MACs at half that rate or 16x16-bit MACs at a quarter of that rate. Software can mix the MAC widths as needed.

The NeuPro Engine's other execution units include 32-bit accumulators along with specialized hardware to handle post-shift functions, neural-network activations, and pooling layers. The activation hardware supports rectified linear units (ReLUs), parametric ReLUs, sigmoid, tangent hyperbolic, and other activation functions. The pooling-layer hardware enables 2x2 and 3x3 average and maximum calculations. The entire engine performs 8- and 16-bit operations, and it allows customers to set the precision layer by layer. The company's CDNN software automatically determines the resolution for each layer when it maps a pretrained network graph to the engine.

Like the Google TPU, the NeuPro Engine implements a direct pipeline from the activation layer to the pooling layer. This pipeline passes results directly between stages without accessing local memory, further reducing bandwidth requirements. Designers can configure separate local memories for storing frequently used convolution data and weights; typical sizes range from 0.5MB to 2MB.

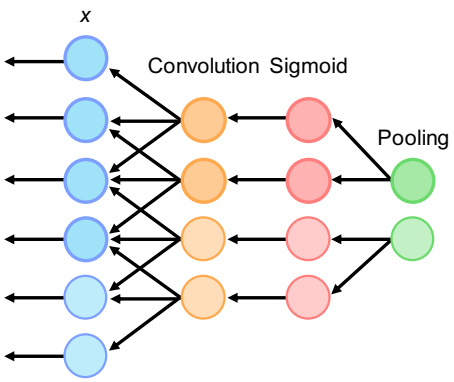

Figure 2. Neural-network training using back propagation. Retraining requires an optimization routine that reverses the sequence of operations that a feed-forward DNN uses. The training algorithm iteratively adjusts the weights of preceding layers to minimize error at the output.

Ceva specifies the NeuPro Engine's performance at 2,000 GOPS for the smallest 512-MAC configuration, assuming a 2.0GHz clock frequency. This performance is similar to that of the Kirin 970's neural processing unit, the fastest accelerator available in today's smartphones (see *MPR 1/22/18,* "Neural Engines Rev Up Mobile AI"). In a 16nm process, the largest 4,096-MAC NeuPro configuration can deliver about 12,500 GOPS (6,267GMAC/s) when running at 1.53GHz. This performance is high enough for Level 2 ADAS and some Level 3 autonomous-driving systems.

A DSP Controls the Show

The second NeuPro core is the programmable VPU, which also includes a scalar processing unit (SPU). It's an eight-stage VLIW core with a 14-stage pipeline. Although the ISA is optimized for neural-network processing, it uses the same pipeline depth and VLIW width as the company's XM-series (see *MPR 10/10/16,* "Ceva XM6 Accelerates Neural Nets"). As with other Ceva DSPs, designers can build custom instructions into the VPU. The NeuPro VPU can run the FreeRTOS operating system, although some users will run bare metal.

The VPU includes a new DMA controller optimized for 3D tensors (see *MPR 12/12/16,* "Many Options for Machine Learning"). To ensure that reading and writing 3D-tensor data conform to the network's dimensions, the DMA engine automatically adds zeros to pad input data and crops output data with no additional overhead.

Code in the VPU employs a driver that transmits processing jobs to the NeuPro Engine. The engine's hardware has a task queue that receives these jobs and executes them in sequence. This approach decouples the engine from the VPU and reduces control overhead relative to the more common coprocessor model. Although the NeuPro Engine is designed to handle all common neural-network layers, Ceva anticipates that new approaches may emerge over time; in this situation, the VPU's programmability allows it to handle these new neural-network layers while continuing to offload the remaining layers.

Although the two new cores work together to form a complete machine-learning processor, Ceva licenses the NeuPro Engine separately for users wanting to combine a general-purpose DLA with one of its vision-optimized XM products. The XM cores also have built-in SPUs that can perform control tasks.

On-Device Retraining

Neural-network-processor IP typically serves in inference engines, which run networks previously trained to classify objects from a particular data set—for example, images representing facial features or road signs, as well as sounds representing parts of spoken words. To most accurately

determine the weights for a particular classifier, application developers minimize error by training their networks on servers and PCs using FP16 or FP32 calculations. Floating-point math is too area and power hungry for most embedded applications, however, so IP vendors offer software that fits the trained network weights to the fixed-point precision of their DLA cores.

Like other DLA-IP vendors, Ceva supplies its deep-neural-network (CDNN) software framework for converting previously trained models. Unlike some, however, it retained a 16-bit integer option in its IP for cases where 8 bits has proven insufficient. The company counts safety-critical automotive tasks among applications that need the higher precision. The INT16 format still requires more area and power than INT8, but it's a reasonable compromise compared with floating point.

For NeuPro, Ceva is adding to its CDNN software the capability to retrain networks on the client device. This feature initially targets app developers rather than end users, but it allows them to employ the Ceva network generator to update networks without having to upload a database to a server, run Caffe or some other training framework, and then convert the computational graph from floating point to NeuPro's fixed-point architecture.

As Figure 2 shows, NeuPro retraining involves running the network-layer operations in reverse. This process adjusts the weights in the hidden layers in accordance with a known training set and repeats to achieve minimal classification error at the output. An app developer can use the CDNN tools on a simulator or test device and then roll out an updated model to devices in the field.

Staying Ahead of the Class

As Table 1 shows, the NeuPro architecture extends Ceva's performance lead. The company's previous XM6 computer-vision cores also optionally integrate up to 4,096 MACs, but in that architecture, the normalization, pooling, and other layers run on a separate set of VPUs. The new design can handle all the layers on the NeuPro Engine, which uses specialized hardware to increase energy efficiency by eliminating data movement between the two cores. By avoiding that bottleneck, NeuPro is likely to deliver higher performance on real DNN code relative to DSP-based architectures using the same number of MAC units.

But Ceva isn't the first company to build a DLA capable of running all DNN layers in a single core. Cadence also takes that approach in its Vision C5 (see *MPR 5/29/17,* "Cadence C5 Flexes for Neural Networks"). That core can perform 8- or 16-bit MAC operations, but it integrates just 1,024 MACs. To run more-complex networks, designers must instantiate multiple cores connected through the SoC bus. The C5 lacks mixed-precision-network support, as well as the special-purpose hardware in the NeuPro Engine.

Imagination's PowerVR 2NX provides a larger mix of variable-precision operations than NeuPro, giving customers a choice of 4- to 8-bit precision in 1-bit increments, along with a choice of 10-, 12-, or 16-bit precision. It offers half of NeuPro's maximum MAC complement, however, and delivers just a third of the performance. The Cadence and Imagination cores are well suited to integration in power-constrained mobile processors, as are the NP500/1000, but the larger NeuPro NP4000 is better for ADASs and autonomous vehicles.

The Synopsys EV64 is NeuPro's closest competitor in performance, but it strikes a compromise by integrating fixed 12-bit MACs. Hence, it lacks the capability to mix precision layer by layer. The EV64 couples a CPU and DSP cluster with the MAC accelerators, though, providing additional options for handling DNN-layer operations.

Expanding Machine-Learning Opportunities

Ceva is the first DLA-IP vendor to support on-device neural-network retraining. That feature will be attractive to app developers who want to add new classification functions to their existing networks without a complete offline retraining and remapping cycle. Using the same hardware for retraining and inferencing makes it easier to update machine-learning applications as new training sets become available, and it enables tuning with data from individual users.

The NeuPro Engine gives designers a unique DLA that's more efficient than competing architectures yet retains 16-bit precision for systems requiring high accuracy. It can scale from less than 2,000 GOPS to 12,500 GOPS, making it well suited to a wide range of tasks. The small configurations are useful for mobile devices, and the large ones can handle more-complex object recognition and video surveillance. The NeuPro VPU matches the capability of the Cadence and Synopsys IP, both of which also have other DSP-derived architectures, but it's important for supporting new network models as they emerge.

Ceva has more experience than other DLA-IP providers, and it has adapted more than 120 neural-network models to run on its cores. The CDNN software supports the popular TensorFlow and Caffe frameworks, offering designers a complete package for developing embedded neural networks. ♦

	Ceva NeuPro NP4000	Cadence Vision C5	Imagination PowerVR 2NX	Synopsys EV64
Clock Speed*	1.53GHz	1.10GHz	1.00GHz	1.28GHz
Pipeline Depth	14 stages (VPU)	10 stages	Not applicable	10 stages
Integer MACs per Cycle*	4,096x 8-bit MACs	1,024x 8-bit MACs	2,048x 8-bit MACs	3,520x 12-bit MACs†
Peak DNN Performance*	6,267 GMAC/s 8x8-bit	1,126 GMAC/s 8x8-bit	2,048 GMAC/s 8x8-bit	4,500 GMAC/s 12x12-bit
Production RTL	3Q18 (est)	3Q17	3Q17	3Q17

Table 1. Comparison of DLA IP cores. NeuPro has the highest-performance DLA available in a licensable core. *Maximum configuration, 16nm FinFET, typical conditions; †including optional CNN engine. (Source: vendors)