The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

CEVA SENSPRO FUSES AI AND VECTOR DSP

Applications Include Smart IoT Devices, Drones, and Small Robots

By Mike Demler (April 20, 2020)

Ceva's new SensPro gives designers a versatile and highly configurable set of licensable intellectual-property (IP) cores for scalar and vector signal processing as well as neural-network acceleration. The company designed SensPro, which it introduced at the recent Linley Spring Processor Conference, as a powerful sensor hub that combines technology from its BX2 DSP, XM6 computer-vision engine, and NeuPro deep-learning accelerators (DLAs). Although the new offering's neural-network capacity is less than NeuPro's, its AI/sensor-fusion capabilities are a good fit for a wide range of products, from IoT devices and smartphones to ADASs, augmented/virtual-reality (AR/VR) headsets, drones, and robots.

By integrating the BX2 with an eight-slot VLIW architecture that includes a DLA, SensPro can execute DSP operations in parallel with neural-network inference and thus power devices that require contextual awareness. For example, it can steer a drone around hazards by fusing altitude- and velocity-sensor measurements with image recognition from cameras. Designers can choose from a list of options, but the company is initially offering SensPro in three preconfigured versions: small, medium, and large. All have a BX2 DSP that runs the application software and that handles control and scalar signal processing. Ceva plans to ship production RTL for general licensing in 3Q20.

As Table 1 shows, the SP250 is the small SensPro model, combining the BX2 with a single vector control unit (VCU) that includes 256 single-cycle INT8 multiply-accumulate (MAC) units. The MAC units can also deliver 64 INT16 operations per cycle. The medium-size SP500F integrates two VCUs, each with 256 INT8 MAC units but 32 single-precision floating-point MAC units as well. The SP500F adds support for binary and ternary neural networks (BNNs and TNNs). BNNs reduce storage requirements by using just two weight values: +1 and –1. Doing so reduces processing to simple bit manipulations. TNNs allow quantization of small weight values to zero, effectively pruning the network and thereby increasing power efficiency.

The large SP1000 also has two VCUs, each with 512 INT8 MAC units that can execute 128 INT16 MACs per cycle as well. In addition, each VCU can execute 3,072 BNN operations per cycle. The company expects SensPro to run at up to 1.6GHz in 7nm technology, enabling the SP1000 to deliver more than three trillion operations per second (TOPS) using 8-bit integers and to deliver 20 TOPS for BNNs.

A Higher-Performance Hybrid

SensPro implements multiple functions that Ceva previously developed for its DSPs. As Figure 1 shows, it integrates all the function blocks of the BX2 (see *MPR 2/4/19*, "Ceva's BX Hybrid Boosts DSP Engine"). They include the program-control unit (PCU) and two CPUs the company calls scalar processing cores (SP0 and SP1), although it positions the BX2 as a hybrid DSP owing to its support for five-slot SIMD/VLIW operations. The CPUs work with a configurable four-way instruction cache of undisclosed size. The original BX2, however, can have two- and four-way instruction

	SP250	SP500F	SP1000
INT8 Perf	256 MACs/cycle	512 MACs/cycle	1,024 MACs/cycle
INT16 Perf	64 MACs/cycle	128 MACs/cycle	256 MACs/cycle
BNN Perf	None	1,536 ops/cycle	3,072 ops/cycle
FP32 Perf	None	64 MACs/cycle	None

Table 1. Ceva SensPro IP. The company will initially offer the new sensor-fusion product in three configurations, but it plans to offer a release that allows licensees to customize their designs. (Source: Ceva)

Price and Availability

Ceva plans to begin licensing production SensPro RTL in 3Q20. It doesn't disclose pricing. For more online information, access *www.ceva-dsp.com/product/ceva-senspro*.

caches of up to 128KB. As with all the company's DSPs, customers can add their own instructions using the optional Ceva-Xtend feature.

To support AI and vector-DSP workloads, SensPro has a relatively large load/store bandwidth. The original BX2 can execute a single 128-bit load or store operation per cycle, but the new design can load 2,048 bits in parallel with a 1,024-bit store.

The SensPro architecture makes each of the VCU's four function units optional and configurable. The models with two VCUs require them to have identical configurations, and each unit supports a 1,024-bit vector width. As Figure 2 shows, the architecture can provide 256 or 512 INT8 MAC units per VCU; they can also execute 64/128 INT16 or 16 INT32 MAC operations per cycle. BNNs and TNNs replace MAC operations with simple bit manipulations, which execute in separate hardware units. SensPro handles binary and ternary multiplications of 2-, 4-, and 8-bit activations.

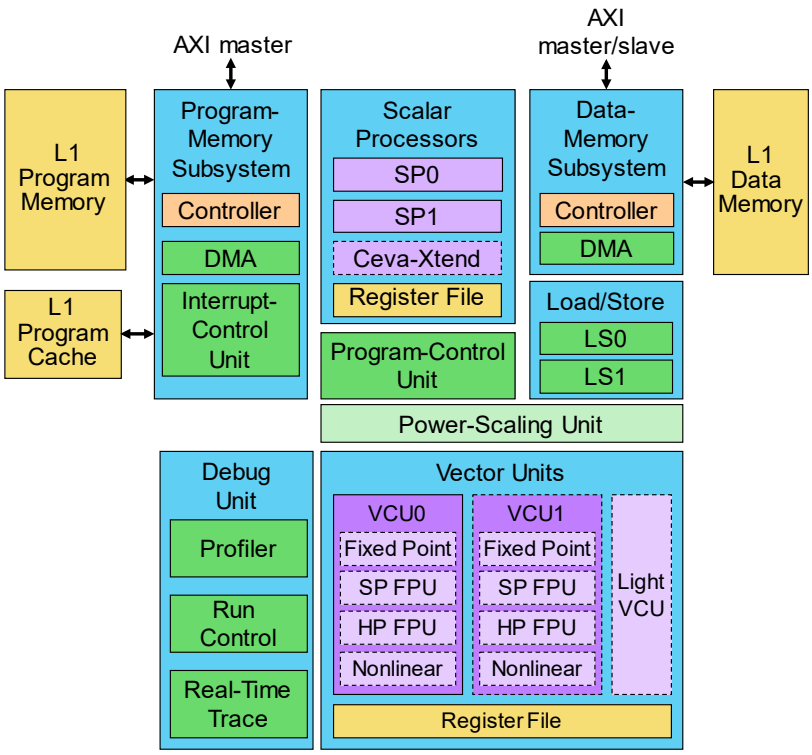

Figure 1. Ceva SensPro DSP. The design integrates the program-control and scalar processing units from the BX2 DSP, along with as many as two configurable vector control units. To feed the vector engine, the data-memory interface can execute parallel 2,048-bit loads and 1,024-bit stores on each clock cycle.

SensPro customers can equip each VCU with a single-precision FPU that includes 32 FP32 MAC units, which can also perform 64 FP16 MACs per cycle. The VCU omits double-precision (FP64) capability, but the CPUs can execute scalar FP64 operations. In its largest configuration, a 7nm SensPro design running at 1.6GHz can deliver more than 400 single-precision gigaflops per second. The VCUs work with an optional two-way data cache that designers can configure, but Ceva withheld the size.

The VCUs accelerate nonlinear operations such as Taylor-series expansions and the complex arithmetic commonly used in radar signal processing. A new feature supports the iterative Newton-Raphson root-finding algorithm, which increases accuracy for nonlinear operations. Intra-vector operations enable functions that output a scalar value, such as the max-pooling that many convolutional neural networks (CNNs) employ. The optional "light" VCU handles only nonarithmetic operations, such as bit manipulation, data movement, and data rearrangement.

Putting Things Into Context

Ceva compares SensPro's AI performance with that of the XM6 computer-vision DSP, which it introduced in 2016 (see *MPR 10/10/16*, "Ceva XM6 Accelerates Neural Nets"). But the new design more closely resembles a scaled-down NeuPro DLA. The latest NeuPro-S combines the XM6 with an engine optimized for CNNs, and it lets customers integrate their own accelerators (see *MPR 10/7/19*, "Ceva and Synopsys Spin More TOPS"). Whereas NeuPro-S comes in 1,024-, 2,048-, and 4,096-MAC configurations, however, the initial SensPro models top out at 1,024 MAC units.

The biggest difference between NeuPro and SensPro is that the former lacks the sensor-processing capabilities of the latter. By integrating an XM6 DSP, NeuPro can target some of the same applications as SensPro, but for AR/VR, drones, and robots, it can only handle their computer-vision functions. The older design is also much larger, owing to its four scalar cores and three vector units. The combination of a BX2 DSP with 1,024-bit VCUs, including FPUs and nonlinear function units, enables SensPro to perform complex signal processing, unlike its predecessor.

Cadence, Synopsys, and other competitors offer a variety of licensable DSPs that target the same workloads as SensPro, but they have yet to integrate all the relevant features in a single product. The closest match is Cadence's Vision Q7 (see *MPR 6/10/19*, "DSP-IP Vendors Target 3D Navigation"), which integrates some, but not all, of the SensPro SP500F's capabilities. Like SensPro, the Q7 can fuse depth- and

motion-sensor data to run simultaneous localization and mapping (Slam), which enables drones and robots to navigate. It also combines scalar cores with a 512-bit vector engine, including 512 INT8 MAC units for computer-vision neural networks. But the SP500F executes twice as many floating-point operations per cycle.

SensPro can perform computer vision and Slam in parallel with audio processing and speech recognition. Ceva supports it with the ClearVox and WhisPro software for the latter two applications, along with the CDNN compiler, a computer-vision library, and a Slam SDK. It integrates those software tools in Senslinq, a sensor-fusion platform comprising APIs and a Linux-based hardware abstraction layer (HAL) that distributes control messages and sensor data between Ceva DSPs, host CPUs, and wireless modems.

Making Sensors Smarter

Ceva has always offered configurable and customizable DSPs, but SensPro marks a departure from its earlier products, which target a single application such as audio processing or computer vision. Although it's initially selling SensPro in just three fixed configurations, they cover a wide range of applications. The SP250 can combine audio, speech, and vision in devices such as doorbell cameras, smart speakers, and smart displays. The SP500F adds the floating-point capabilities necessary to run Slam in drones and smart vacuum cleaners, and the SP1000 goes even further by increasing AI performance for more-sophisticated robotics.

Yet Ceva isn't the first IP vendor to sell multipurpose DSPs. Cadence claims that distinction for its appropriately named Fusion DSPs (see *MPR 8/1/16,* "Fusion G3 Beefs Up Floating Point"). It offers two Fusion families: the Fusion F1 combines signal processing with wireless connectivity for IoT and wearables, and the G3/G6 target higher-performance tasks such as ADAS sensor fusion and radar. But because Fusion DSPs only integrate up to eight MAC units for low-end image processing, they lack the capabilities for AI-enabled contextual awareness. To match SensPro, Fusion customers must add a Cadence Vision DSP.

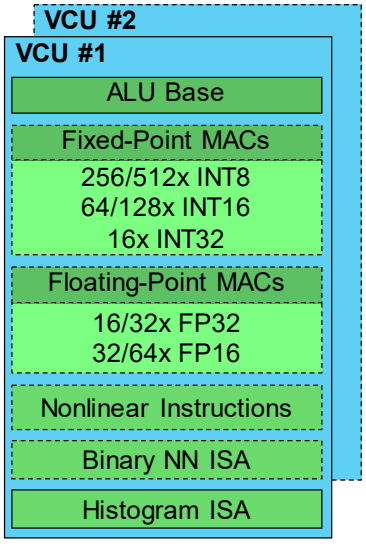

Figure 2. SensPro vector units. The VCUs augment the BX2 DSP with an eight-slot VLIW that handles 1,024-bit vectors. The architecture allows custom configurations that include or exclude the integer, floating-point, and nonlinear execution units.

SensPro addresses two industry trends: on-device AI and smart machines. Just like humans, smart machines must employ multiple senses to properly perceive their environment. Several chip vendors target them by offering powerful camera-based neural-network processors, but they lack the DSP capabilities for sensor fusion.

Ceva has wisely combined multiple hardware and software products into one package. Its three initial preconfigured models are suitable for numerous consumer and industrial systems, but licensees will appreciate the ability to customize their designs when that option arrives in a future release. ♦