CEVA XM6 ACCELERATES NEURAL NETS

By Mike Demler (October 10, 2016)

Ceva's new XM6 DSP core enables deep learning in embedded computer-vision (CV) processors. The synthesizable intellectual property (IP) targets self-driving cars, augmented and virtual reality, surveillance cameras, drones, and robotics. At the recent Linley Processor Conference, the company described the new features of the XM6 as well as the updated capabilities of its development kit for deep neural networks (CDNN2). It plans to license the IP to lead customers by the end of the year and to make it available for general licensing in 1Q17.

Despite the nonsequential numbering, the XM6 is the company's fifth-generation architecture for imaging and CV. It adds processing capabilities that build on those of the previous-generation XM4 (see *MPR 4/27/15*, "Ceva Sharpens Computer Vision"). Foremost among the enhancements is a new neural-network hardware accelerator that offers 512 additional single-cycle multiplier-accumulators (MACs) that are 16x16 bits. The accelerator connects to the DSP core's vector-processing cluster through an AXI4 interface, providing a boost specifically for the convolution layers that consume most neural-network processing cycles (see *MPR 3/7/16*, "Accelerating Machine Learning").

The normalization, pooling, and other layers that constitute a convolutional-neural-network (CNN) model run on the XM6's 512-bit vector processing units (VPUs). The new design increases the number of VPUs from two to three, all of which share 128 single-cycle 16x16-bit MACs, bringing the XM6's total MAC count to 640. The core also includes four 32-bit scalar processing units (SPUs), which are the same as in the XM4.

According to Ceva's tests, the addition of a third VPU and hardware accelerator increases CNN performance by 8x compared with the XM4. MAC utilization increases to 95%, yielding an average 2x boost for a variety of complete vision kernels.

Designers can include the optional vector floating-point unit, which can handle half-resolution FP16 calculations—a capability the XM4 lacks. The half-resolution format is better suited to CNN training than embedded inference engines, which typically use reduced-resolution fixed-point calculations. Nevertheless, the company's customers requested the feature, and providing it eases the transition from GPU-based CNNs. The vector FPU can perform eight FP32 operations or 16 FP16 operations per cycle.

The XM6 can use multiple CNN accelerators connected to the configurable AXI4 interface. The core integrates a dedicated de-warp accelerator to process images from a wide-angle fisheye lens or the 360-degree surround views that some vehicles employ. The accelerator supports the ARM Frame Buffer Compression (AFBC) protocol that the IP provider implements in its display controllers, GPUs, and VPUs. The combination of fixed-function and programmable units enables users to easily implement new network topologies. Autonomous vehicles, for example, are likely to require frequent CNN-model updates as manufacturers collect data from previously untrained driving scenarios.

Additional improvements in the XM6's performance come from refinements to the scatter-gather memory interface and Ceva's sliding-window mechanism. The company withheld details, but the DRAM bandwidth remains unchanged relative to the 512-bit XM4 interface. The sliding window reduces external-memory accesses by enabling reuse of pixel data in overlapping image subframes.

To support development and installation of neural-network models on the XM6 and XM4, the company in June released its second-edition deep-neural-network development kit (see *MPR 10/26/15*, "Ceva Enables Deep Learning"). CDNN2 automates conversion of pretrained 32-bit floating-point models developed using the Caffe and TensorFlow frameworks to the 16-bit fixed-point precision that the XM6 CNN accelerator employs. It supports a variety of popular CNN topologies, including AlexNet and GoogleNet. Ceva also offers software libraries for OpenCV and OpenVX, and the XM6 release will add OpenCL support.

The race to develop self-driving cars is accelerating, pushing CV-IP vendors to rapidly produce new designs that will enable these vehicles to safely navigate without human intervention. The XM6 aims to take the lead in programmable CV cores from Cadence and Synopsys. Cadence's P6 offers 2x the memory bandwidth and superior floating-point performance, but its recent upgrade to 256x8-bit MACs per cycle is unable to match the XM6 DSP and CNN accelerator (see *MPR 10/12/15*, "Cadence P5 Boosts Embedded Vision"). The quad-core EV64 from Synopsys is a close competitor in scalar and vector processing performance, but it lacks FP16 support. Also, its CNN engine uses 12-bit fixed-point calculations, which may be inadequate for some tasks (see *MPR 7/4/16*, "Synopsys Improves Vision With DSP").

Compared with these competitors, Ceva has demonstrated the most complete package for self-driving cars. It recently showed the predecessor XM4 running AdasWorks Drive 2.0, a self-driving-car software stack. That startup will likely take advantage of the XM6's performance boost in the prototype self-driving car it plans to roll out at the 2017 Consumer Electronics Show. The CV cores come with an ISO 26262–compliant safety design package, which together with the CDNN2 software kit enables ADAS-processor developers to quickly enter the fast-changing autonomous-vehicle market. ♦