The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

# CEVA SHARPENS COMPUTER VISION

## *New Ceva-XM4 DSP Core Adds FPUs and 32/64-Bit Vectors*

*By Tom R. Halfhill (April 27, 2015)*

As more machines gain the gift of sight, engineers are rediscovering a principle long known to biologists: vision is equally a sensory perception and a cerebral function. The eyes see, but the brain interprets and reacts. Thus, processing power is as vital to computer vision as image capture.

To augment those back-end functions, Ceva introduced at the recent Linley Mobile Conference a new licensable DSP core optimized for vision processing. The Ceva-XM4 is a fourth-generation design that has numerous improvements over the previous Ceva-MM3101. As Table 1 shows, it quadruples the number of multiply-accumulate (MAC) units, quadruples the width of VLIW operations, adds 32-bit floating-point units and vector operations, and doubles the number of scalar units. It also boosts the I/O bandwidth by 100% and memory bandwidth by 33%.

These and other enhancements to the memory system, instruction-set architecture, and software tools increase the XM4's performance per milliwatt and per square millimeter of silicon by 2.5x under the same conditions, the company says. To further boost performance, the XM4 can reach a 20% faster clock frequency in the same process technology—1.2GHz in TSMC's 28nm high-performance mobile (HPM).

Ceva licenses the XM4 for a broad range of computer-vision applications, including low-power mobile devices (smartphones, tablets, and wearables), higher-power mobile systems (vehicles, robots, and drones), and fixed systems (security and surveillance cameras, industrial machinery, and IoT). The XM4 can process video streams of up to 4K resolution, and some features are customizable for specific applications. Ceva says the production RTL is available now.

### Vision Processing Goes to School

State-of-the-art vision processors must enable 3D imaging with depth perception and machine learning, which demand more processing power than rudimentary object recognition. To that end, Ceva has significantly enhanced the XM4's VLIW architecture. Its very long instruction words can now execute eight parallel operations versus six in the MM3101.

In addition, the XM4 can process up to 4,096 bits of data per clock cycle compared with 1,024 bits for its predecessor. To reach this peak performance, the XM4's dual vector processing units can execute up to 128 MACs per cycle. Although the memory system can deliver only

| | Ceva-XM4 | Ceva-MM3101 |
|---|---|---|
| **Clock Frequency (max)** | 1.2GHz | 1.0GHz |
| **VLIW Architecture** | 8 parallel operations | 6 parallel operations |
| **VLIW Data Width** | 4,096 bits | 1,024 bits |
| **Scalar Processing** | 4 scalar units | 2 scalar units |
| **Vector Processing** | 2 vector units | 2 vector units |
| **32-Bit FPUs** | 1 scalar FPU, 1 vector FPU | None |
| **L1 Cache** | Configurable, 4-way | Configurable, 2-way |
| **Memory Interfaces** | 2x 256-bit | 2x 256-bit + 1x 128-bit |
| **Data Memories** | Unified data memory | Scalar and vector memories |
| **AXI Interfaces** | Configurable 128- or 256-bit AXI4 | 128-bit AXI3 |
| **Multiply-Accumulate Perf** | 128 MACs per cycle | 32 MACs per cycle |
| **32-Bit Operations** | Fixed & FP, scalar & vector | Fixed-point scalar |
| **64-Bit Operations** | Fixed-point scalar & vector | None |
| **RTL Availability** | 1Q15 | 2012 |

**Table 1. Ceva-XM4 DSP core versus its predecessor.** Maximum clock frequency assumes fabrication in TSMC 28nm HPM. (Source: Ceva, except *The Linley Group estimate)

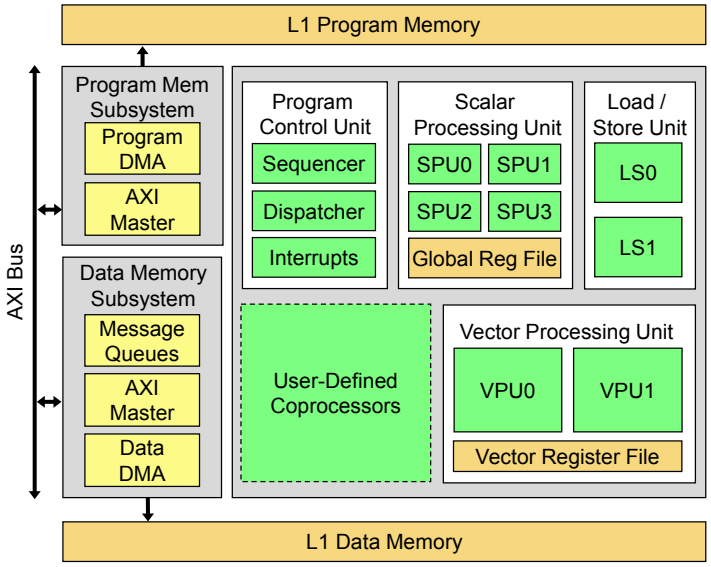


**Figure 1. Ceva-XM4 block diagram.** Dotted lines indicate optional features. To optimize performance, the DSP contains both scalar and vector units. The design also contains a power-management unit that can adjust the DSP's voltage and frequency to suit workloads or to rest between tasks.

512 bits per cycle, the company says the memory bandwidth rarely limits computer-vision algorithms, which often reuse the same coefficients and vectors.

Unlike the 32-bit fixed-point MM3101, the XM4 supports both fixed- and floating-point arithmetic. It can perform 8-, 16-, 32-, and 64-bit fixed-point operations as well as 32-bit floating-point math. The single-precision FPU complies with the IEEE 754 standard. Ceva decided against adding 64-bit FPUs, however, because few customers need them and they would have significantly enlarged the core's footprint.

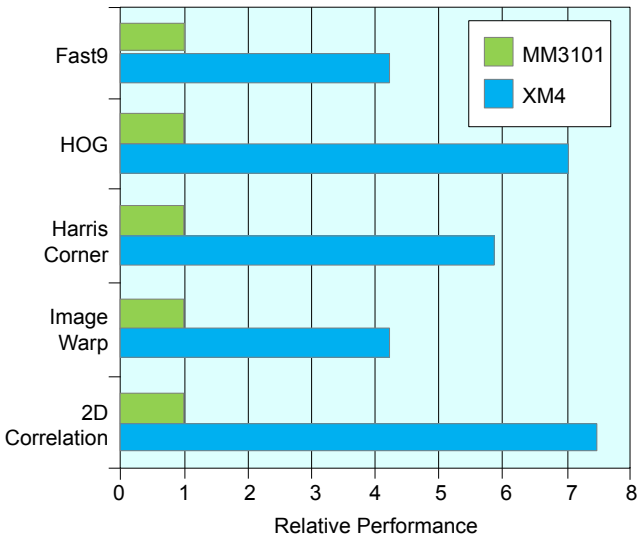


**Figure 2. Ceva-XM4 performance versus Ceva-MM3101.** The XM4 shows a big advantage over its predecessor on common computer-vision algorithms, even when running at the same clock speed. At its maximum frequency, the XM4 widens its lead by another 20%. Fast9 and HOG (histogram of gradients) are feature-detection algorithms. (Source: Ceva)

As the block diagram in Figure 1 shows, the XM4 has four scalar processing units, two vector processing units, and one (optional) vector FPU. Each scalar unit can perform one 32-bit FP operation per cycle. The vector FPU can perform eight FP operations per cycle. By contrast, the MM3101 has only two 32-bit scalar units, no 32-bit vector units, and no FPU.

Ceva also expanded the XM4's instruction set to accelerate computer-vision algorithms and other applications that use nonlinear math. For example, some vision algorithms require numerous divides, square roots, and inverse square-root operations. Whereas the MM3101 used lookup tables, the XM4 has native instructions that can perform multiple operations per clock cycle on either fixed- or floating-point data.

## XM4 Easily Outruns MM3101

By combining the 1.33x improvement in VLIW operations per cycle (eight versus six), the 4x improvement in SIMD width (4,096 versus 1,024), and the 1.2x faster clock speed (1.2GHz versus 1.0GHz), we estimate the XM4's peak throughput is about 6.4x greater than the MM3101's. The new vision-optimized instructions add still more performance.

Sure enough, as Figure 2 shows, the XM4 easily outperforms the MM3101 on common vision algorithms that Ceva optimized for the new core. On these tests, the XM4's average performance gain is about 5.7x at the same clock speed. When running at its 20% higher clock speed, the new DSP should gain about 6.8x, which exceeds our 6.4x estimate. The benchmarks suggest that these vision algorithms are compute intensive and can take advantage of the XM4's improvements without saturating the memory subsystem.

To boost the maximum clock frequency to 1.2GHz in 28nm HPM, Ceva lengthened the instruction pipeline to 14 stages versus 10 in the MM3101. It says the XM4 uses 60% less energy than the MM3101 when performing the same tasks because it finishes sooner, thus enabling it to spend more time relaxing at a lower clock frequency and voltage. Still, we estimate that the wider SIMD units and numerous other architectural improvements double the operating power of the new design relative to the MM3101. The die-area increase is likely similar.

Among other things, higher performance enables better machine learning. Ceva says the XM4 can implement a convolutional neural network (CNN) that enables computer-vision systems to learn new things from their observations and adapt to their environments.

One example might be a security camera that learns to distinguish between typical and atypical movements

within its visual range. The camera might start recording video and alert a guard only when it detects something unusual. A neural network also enables a system to learn new tasks without requiring programmers to rewrite code.

## Accessing Memory in Parallel

Keeping all the function units fed with instructions and data is a challenge, so the XM4 improves on the MM3101's memory subsystem. It supports up to 4GB of program memory and 4GB of data memory, each with its own 256-bit interface. The DSP core can access both memories at once to sustain parallel operations. A dedicated DMA controller for the data memory uses up to eight traffic managers and queue managers to automatically handle data traffic to the optional and configurable L1 cache, which is now four-way set associative.

To ease programming, the DMA controller can automatically fetch data in various formats, such as byte-aligned transfers and two-dimensional arrays. Byte-aligned transfers automatically maintain memory alignment when moving data from one place to another, and 2D data transfers are useful for fetching video frames as arrays instead of as single rows of pixels. The latter technique can reduce memory transactions by a factor of eight, conserving memory bandwidth and power.

As Figure 3 shows, the XM4 can also load data from multiple memory addresses in a single parallel operation. This capability enables scatter-gather operations that fetch several operands from disparate memory locations and automatically load them into a vector register. Computer-vision applications frequently need this capability. For example, building a histogram requires access to random pixels throughout the image. Parallel memory accesses also enable Ceva's high-level compilers to vectorize serial code.

The XM4's traffic managers and queue managers can transfer data to and from external memory, external hardware engines, the host CPU, or another XM4 core. In addition, they buffer the data without involving the host CPU. Some of these features depend on the user's system design, so they are configurable when synthesizing the DSP core. For instance, users can specify the number of AXI4 master and slave ports and configure their widths to 128 or 256 bits. By contrast, the MM3101 is limited to 128-bit ports using the older AXI3 standard.

## Vectorizing Compilers Ease Development

Ceva provides software-development tools and code libraries for computer vision and general-purpose signal processing. The Eclipse-based integrated development environment (IDE) includes a vectorizing C/C++ compiler that can automatically convert serial code into vector code using SIMD instructions. Some serial code can't be automatically vectorized, so the compiler also supports the Vec-C extensions, which use the same vector data types as OpenCL. These data types replace the ANSI C types but use standard operators. For even more control, programmers can use intrinsic functions, such as VMPYADD (vector multiply-add).

The tool chain includes a linker, a profiler, an instruction-set simulator, a cycle-accurate simulator, support for an on-chip emulator, and a multicore debugger that works with the software simulators or emulator. In addition, the company offers a prototype development board that implements an XM4 DSP in an FPGA.

Ceva's application developer kit (ADK) includes the Ceva-CV library, which is based on OpenCV, an open-source library of computer-vision and machine-learning functions that works with Android, iOS, Linux, Mac OS, and Windows. These C/C++ libraries implement more than 2,500 algorithms commonly used for 3D vision, facial recognition, object tracking, and other functions. The Ceva-CV library further optimizes these algorithms for the company's DSPs by improving their performance and power efficiency and by abstracting the CPU-DSP interface and memory system.

For instance, Ceva's SmartFrame module can transfer entire video frames or tiles to or from memory without requiring programmers to manipulate individual bytes or pixels. Another example is depth perception, which requires processing the inputs from at least two cameras or from one camera and a depth sensor. The Ceva-CV library can also work alongside a customer's proprietary code libraries.

Note that these libraries can run on a real-time operating system (RTOS) hosted on the XM4 DSP core instead of on a CPU core. The company's scalar-processing units are capable of hosting an RTOS for code scheduling and other low-level tasks. This control code can run faster on the XM4 than on the MM3101 because the new core has twice as many scalar units.

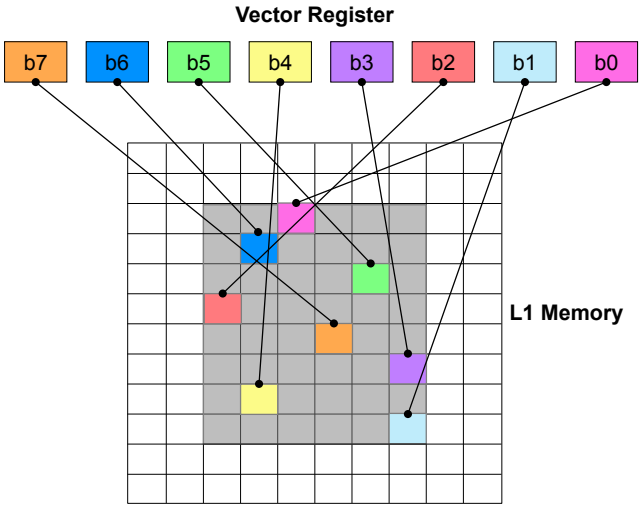


**Figure 3. Scatter-gather memory operations.** The Ceva-XM4 can gather operands from scattered memory locations and load them into a vector register using a single operation.

### Price and Availability

Production RTL for the Ceva-XM4 vision-processor core is available now. Like most IP vendors, Ceva does not disclose licensing fees. For more information, point your browser to *launch.ceva-dsp.com/xm4*.

Ceva also provides an OpenVX-compatible API for offloading computer-vision functions from a host CPU. The Khronos OpenVX 1.0 specification provides a higher-level abstraction for executing these algorithms, enabling developers to build applications using a directed-graph approach that combines software modules with hardware acceleration.

## Ceva Challenges Competitors

Licensable-IP cores marketed as vision processors are available from several vendors, including Cadence (Tensilica), Cognivue, Synopsys (ARC), and Videantis. Their architectures vary widely, defying apples-to-apples comparisons. Additionally, as with all soft-IP cores, their size, performance, and power consumption greatly depend on the core configuration, synthesis flags, physical-IP libraries, target IC process, and other variables. Nevertheless, all of them plausibly claim to deliver better vision performance for less silicon and power than a conventional CPU, DSP, or GPU core.

The Ceva and Cadence VLIW architectures have the most in common. Ceva's vision processors are based on its general-purpose DSPs, and Cadence's IVP processor is based on the configurable Xtensa CPU/DSP. But the Ceva-XM4's VLIW implementation is much wider—up to eight parallel operations on 4,096 bits of data versus the IVP's three parallel operations on 512 bits. In partial compensation, the IVP has twice the memory bandwidth (see *MPR 3/4/13,* "Tensilica Sets Its Sights on Vision").

Synopsys is entering this market with its new DesignWare EV52 and EV54 licensable IP. These products integrate two or four 32-bit ARC HS cores with up to eight programmable accelerators optimized for computer vision and convolutional neural networks. They are designed for face and gesture recognition, home automation, traffic-sign identification, video surveillance, and videogames. Synopsys plans to deliver the production RTL in May (see *MPR 4/13/15,* "Synopsys Embeds Vision Processing").

Instead of starting with a general-purpose DSP or CPU architecture, Cognivue and Videantis designed their vision-processor cores specifically for digital video and computer vision. Their primary applications are advanced driver-assistance systems (ADASs) and in-vehicle infotainment (see *MPR 11/17/14,* "Putting the 'Auto' in Automobile"). Videantis was a pioneer in this field, having introduced its first video core in 2004 (see *MPR 11/7/05,* "Videantis Chases Digital Video").

Cognivue's second-generation G2-Apex combines 32-bit RISC controllers with arrays of SIMD/VLIW computational units (CUs). Each CU has its own local memory. Hard-wired accelerators perform generic vision functions, such as image-pyramid construction. The architecture is configurable: the Apex-321 has one controller and 32 CUs, the Apex-641 has one controller and 64 CUs, and the Apex-1282 has one or two controllers and a pair of 64x CU arrays. By contrast, Videantis uses independent VLIW processing elements or cores that can perform multiple fixed-point operations per clock cycle, and the core count is configurable.

Ceva's existing MM3101—introduced in 2012—was already competitive with the Cognivue and Videantis cores in vision performance, die area, and power. Owing to its numerous improvements, the new XM4 could gain an advantage. As with all soft-IP cores, however, much depends on the configuration options, logic synthesis, and overall chip design. One comfort for prospective customers is that the OpenCV platform helps programmers write software that's more portable, so changing vendors is less prohibitive than before.

## It's Bigger but Better

Ceva has not disclosed the XM4's typical die size when fabricated in 28nm HPM. Although the XM4 is considerably larger than the MM3101, its benchmark results lend credence to the company's claims: 2.5x better performance per milliwatt and per square millimeter of silicon. These improvements are necessary to keep Ceva competitive with the rush of new and improved vision-processing cores coming from competitors.

The market is still wide open. Although computer vision has been around for years, it's finally getting small enough to go mobile and cheap enough for affordable consumer products. It's also becoming more sophisticated, demanding more processing power. An arms race similar to that with smartphone processors is underway. One result will be more-frequent upgrades from the leading vision-processor vendors.

Demand for computer vision is increasing in many applications, mainly in automotive systems, virtual-reality gaming, security, surveillance, computational photography, and natural user interfaces for smartphones and tablets. Processor vendors are offering various solutions, such as Nvidia's Cuda-based GPUs, Qualcomm's Hexagon DSPs, and dedicated coprocessors like the Movidius Myriad (see *MCR 9/15/14,* "Movidius Eyes Computational Vision"). Purpose-built soft-IP cores like Ceva's XM4 offer the most area- and power-efficient solution for mobile processors, so we expect them to become commonplace alongside GPUs and ISPs.

Because Ceva is the leading vendor of general-purpose DSP IP, many chip designers already are using a

Ceva DSP core whose architecture and development tools are similar to those of the new XM4. Familiarity and tool maturity are persuasive factors in the company's favor.

The XM4 has so many improvements over the MM3101, however, that it's almost like adopting a new architecture. That's why the abstraction of computer-vision libraries like OpenCV and Ceva-CV is so important. As with other processor types, vendors must strive to offer the best combination of hardware and software—and the Ceva-XM4 leaps forward on both fronts. ♦