The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

XMOS XCORE.AI ADDS VECTOR UNIT

New Iteration Targets Neural Networks on IoT Devices

By Linley Gwennap (May 4, 2020)

XMOS has extended its microcontroller architecture to incorporate a vector unit that generates up to 51 billion operations per seconds (GOPS). The new Xcore.ai chip retains the company's flexible I/O and DSP capability, allowing it to serve as the main processor in a variety of embedded systems. The best part is the price: it sells for as little as $1 in high volume. XMOS recently received first silicon and expects production late this year.

The chip features two CPUs that implement the latest version of the company's Xcore architecture. Each runs at 800MHz and, as in previous versions, handles up to eight threads. The new 256-bit vector unit operates on a variety of popular AI data types, including 8-bit integers and binary values. Each core has 512KB of SRAM and 64 bits of flexible I/O, as Figure 1 shows. The chip optionally connects to LPDDR DRAM and provides USB2.0 and MIPI interfaces.

Xcore.ai's main target is intelligent IoT devices for smart-home and industrial deployment. At about 500mW (typical), it's best suited to line-powered and frequently tethered devices. For example, a smart speaker could use one CPU to handle far-field-voice processing and the other to handle speech recognition, employing the core's DSP extensions in the former case and the neural-network acceleration in the latter. The chip's flexible I/O makes it an ideal SoC, letting it connect to almost any low-speed serial or parallel device, including digital microphones and cameras.

A Teenage Startup

XMOS spun off from the University of Bristol in 2005 and three years later delivered its first product, the XS1, which introduced a custom multithreaded CPU and software-driven I/O. The XS2 debuted in 2015, adding DSP capability to the CPUs to allow processing of audio and other signals from the flexible I/O connections. In 2016, the company split in two, with then CEO Nigel Toon taking about a third of the employees to found Graphcore, which has since released a high-performance deep-learning accelerator for data centers (see *MPR 9/17/18,* "Graphcore Makes Big AI Splash"). Mark Lippett became XMOS CEO after the split and now leads a team of about 60 people, most in the UK.

The XS1 and XS2 appeal mainly to low-volume designs that implement custom or unusual interfaces. Having many small design wins that require lots of customer support, XMOS generated revenue of about $10 million in 2018 and 2019 while incurring sizable losses. To fund its development, the private company has over the years raised nearly $100 million from major venture-capital firms such as Amadeus and Draper. It expects recent higher-volume design wins for the XS2 to help it achieve breakeven revenue by the end of this year.

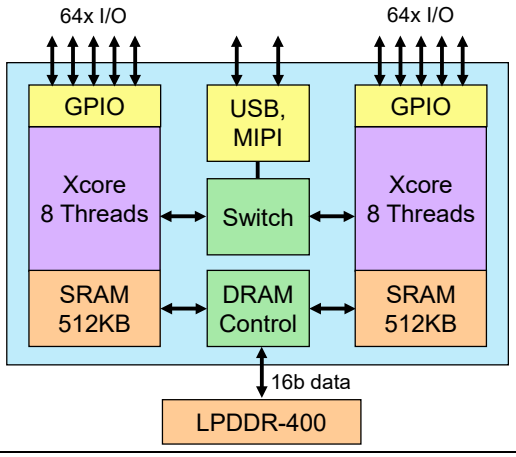

Figure 1. Xcore.ai block diagram. The SoC contains two Xcores, each acting as eight logical CPUs. Each Xcore has a vector unit and controls 64 GPIO connections that can emulate various protocols.

Price and Availability

XMOS expects to sample Xcore.ai later this quarter, with production scheduled for the fourth quarter. The chip will sell for about $2 in 10,000-unit quantities and about $1 in high volumes. For more information, access www.xmos.com.

The company's flexible I/O relies on Xcore's multithreaded design and real-time capabilities. Each thread is guaranteed a fixed cycle time, so software can precisely control the timing of any general-purpose-I/O (GPIO) signal to emulate standard and custom I/O protocols (see *MPR 8/6/07*, "XMOS Redefines Silicon"). This capability allows Xcore.ai to implement any serial port (e.g., UART, I^2C, or I^2S), drive LEDs and other peripherals at 1.8V or 3.3V, and connect to many different sensors. The chip also contains a high-speed USB PHY and MIPI PHY. The two-lane MIPI port connects to cameras and other sensors but not to displays. Although a single thread can handle several low-speed serial ports, the USB and MIPI interfaces typically require 2–3 threads each.

For applications that don't fit into the chip's 1MB of SRAM, designers can add a single LPDDR1 chip through a 16-bit 200MHz interface. Xcore.ai also includes 8KB of one-time-programmable (OTP) ROM. Designers can attach external flash through the flexible I/O. An internal switch connects the two cores at 25Gbps; the design can connect to other chips via four 500Mbps links as well.

Extending the Microarchitecture

The Xcore.ai CPU can decode one instruction per cycle, dispatching it to either the scalar unit or the vector unit, as Figure 2 shows. To reduce code space, most instructions require only 16 bits, although a few require 32-bit encodings. The scalar unit handles math and control operations as well as memory accesses (load/store) using 32-bit registers. Each instruction can access 12 scalar registers, and the CPU includes 96 total registers for the eight threads. The scalar unit generates the effective memory address for both scalar and vector accesses. XMOS rates the 800MHz scalar unit at 1,600 mips on the basis of its many instructions that execute two operations. The scalar unit can also perform 32-bit DSP or floating-point operations at 800Mflop/s.

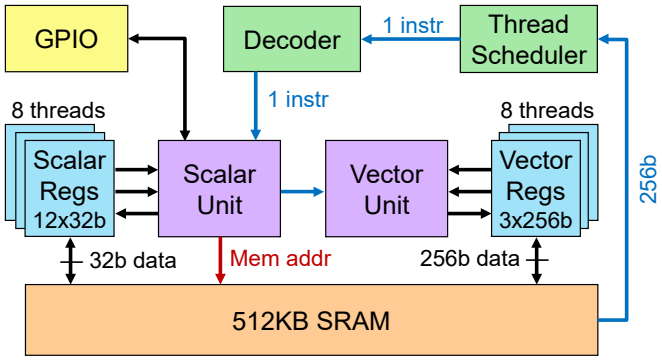

Figure 2. Xcore microarchitecture. The multithreaded CPU dispatches instructions to a scalar unit and a vector unit, each with its own register file.

The vector register file has only three registers per thread, but each holds 256 bits. The vector unit can compute 256 bits at a time, partitioning them into multiple values. It supports 32-, 16-, and 8-bit integers as well as complex integers (32-bit real and 32-bit imaginary components), plus single-bit values that represent either –1 or +1 in binary neural networks. The instruction set includes operations that multiply 8- or 16-bit values and accumulate the products into 32-bit values to avoid overflow. Vector instructions also accelerate common neural-network operations such as batch normalization.

Once an instruction is fetched, the scalar pipeline has five basic stages. The vector pipeline, using a single instruction, can load 256 bits from memory, multiply that data by a value stored in a vector register, and accumulate that data into a third register. (This compound instruction has at least four operands and requires a longer 32-bit encoding.) Thus, the vector unit can perform 256-bit multiply-accumulate (MAC) operations at 800MHz, generating a peak rate of 51.2 GOPS for 8-bit data (per Xcore). Because of the tiny register file, however, a typical neural network would need to load weight values between MAC operations, halving the sustainable throughput.

The CPU switches threads every cycle in round-robin fashion. Ideally, each thread advances every eight cycles, achieving an effective speed of 100MHz. If a thread is inactive, the CPU skips to the next one, but because the five-stage pipeline has no forwarding or interlocks, the minimum number of active threads is five. The thread scheduler buffers 256 bits (at least eight instructions) per thread and issues them to the decoder as needed. When the buffer is empty, it fetches another 256 bits from the SRAM; since the memory is single ported, this access prevents an instruction from executing on that cycle.

An MCU Plus an Accelerator

The Xcore CPU lacks an MMU, but it can run a real-time operating system such as FreeRTOS on a single thread or multiple threads. For the flexible I/O, the company supplies library code for many protocols. It also offers libraries of DSP kernels and neural-network functions. Users can create applications in high-level code and link them to these library routines. XMOS provides an LLVM compiler, Gnu debugger, linker, software-timing-closure tool, and other tools for directly programming its architecture. Note that instead of using interrupts, programmers can allocate threads to time-critical tasks; these threads can restart immediately when an input is received, eliminating interrupt latency.

Customers can employ Xcore.ai in two ways. As Figure 3a shows, it can serve as an accelerator for an external

host processor, allowing the host to run the system software and user code. In this case, almost all the threads on both cores are available to run DSP or neural-network code; the USB connection to the host processor will consume a few threads, reducing the cycles available for vector processing. To minimize system cost, the XMOS design can instead serve as the main processor, as Figure 3b shows. In this case, customers typically assign one Xcore to run the RTOS, user code, and any I/O controllers that the system requires, while the other Xcore handles DSP or neural-network acceleration. This approach further reduces the available vector processing.

Xcore.ai sells in two packages: a 14mm BGA-265 that enables all the I/Os and a 7mm QFN-60 that provides only a few GPIOs. The company declined to reveal the process node, but older XS2 products employ 65nm, so the new product likely uses 40nm or 28nm. XMOS is still characterizing the silicon but expects the chip will consume about 7mW in standby mode and 4mW in suspend mode. Active power depends on how the chip is used. In accelerator mode (Figure 3a), the typical power is about 800mW at peak performance. In SoC mode (Figure 3b), it's 500mW at peak performance or perhaps 250mW for a lighter workload such as audio processing.

For neural-network development, XMOS has ported TensorFlow Lite for MCUs (see *MPR 3/23/20,* "Deep Learning Gets Small"). Developers working in TensorFlow must transfer their models to TensorFlow Lite, then use the company's converter to create a run-time model that automatically calls its AI-library code. Developers who prefer a different high-level framework can export their models in ONNX format and go from there to TensorFlow Lite. Users must quantize their models to an integer format before running them on Xcore.ai. For developing binary neural networks, Plumerai will offer its Larq tools for the new chip. XMOS also provides a C compiler and other tools for developing application software, including optimized libraries of DSP and voice functions, but these tools are less complete than those for Arm and even RISC-V.

Faster Than a Speeding MCU

Using its vector engine, Xcore.ai delivers much better neural-network performance than any standard microcontroller, even more-expensive models. A common Cortex-M7 MCU such as NXP's RT1020 can sustain less than five GOPS at 600MHz. Xcore.ai's advantage is even bigger on real neural networks. XMOS simulated its chip on the same small image-recognition model that Arm reported for the RT1020 and achieved 39x better performance using a single Xcore. Arm expects its newest microcontroller CPU, the Cortex-M55 with 64-bit Helium extensions, will run neural networks 6x faster than Cortex-M7, but that still leaves it far behind Xcore.ai.

Other vendors offer low-cost MCUs with AI acceleration. For example, GreenWaves recently announced its

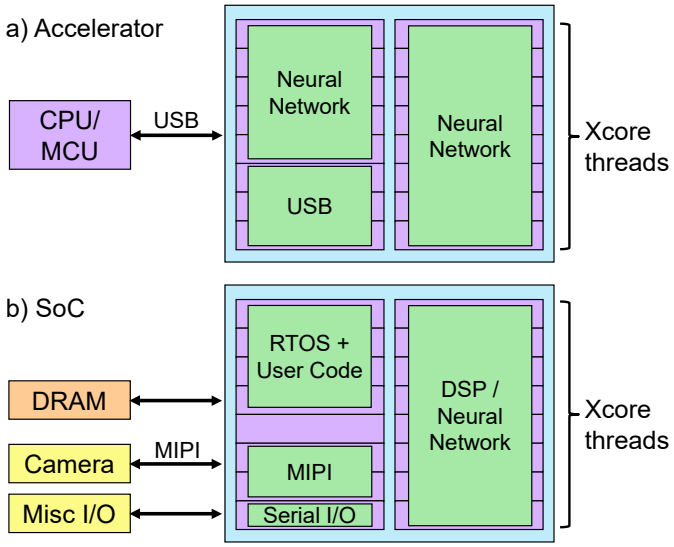

Figure 3. Xcore.ai system configurations. The chip can serve as (a) an accelerator for an external host processor or (b) the main SoC, running system code as well as vector acceleration.

GAP9 processor, which features a RISC-V application CPU plus nine additional RISC-V cores for AI inference (see *MPR 1/13/20,* "GreenWaves GAP9 Goes Faster"). At a peak rate of 49 GOPS, GAP9 delivers about the same INT8 performance as Xcore.ai, as Table 1 shows. It consumes only 50mW at this performance, however, making it better suited to battery-powered devices. For line-powered systems, customers can save money by adopting Xcore.ai.

Kneron developed a custom convolution engine for its KL520 MCU, which features a 200MHz Cortex-M4 CPU (see *MPR 2/24/20,* "Kneron Delivers Efficient AI"). Using this engine, the chip can achieve seven times Xcore.ai's INT8 TOPS, as Table 1 shows, while consuming similar power. The KL520 also offers analog I/O that Xcore.ai lacks and is compatible with Arm user code and development tools.

	XMOS Xcore.ai	GreenWaves GAP9	Kneron KL520
Real-Time CPU	Xcore	32-bit RISC-V	Cortex-M4
RT-CPU Clock	800MHz	250MHz	200MHz
AI Perf (INT8)	51 GOPS*	49 GOPS	346 GOPS
Binary Networks	408 GOPS*	49 GOPS	346 GOPS
DSP Perf (FP32)	0.8Gflop/s*	Undisclosed	None
Internal SRAM	1,024KB	1,638KB	96KB
DRAM Interface	16-bit LPDDR1	None	16-bit LPDDR2
Camera Inputs	1x MIPI	Parallel	2x MIPI, DVP
Video Output	Emulated†	VGA	VGA
Motor Control	PWM†	None	PWM
Power (typical)	500mW	50mW	500mW
IC Process	Undisclosed	22nm FD-SOI	40nm
Volume Price	$1	$3	$4†
Production	4Q20	4Q20	4Q19

Table 1. Comparison of three low-cost AI SoCs. XMOS provides similar performance as GreenWaves but at a lower price. *Using one Xcore; †using flexible I/O. (Source: vendors, except †The Linley Group estimate)

But again, Xcore.ai is much less expensive. It supports binary networks as well, unlike the other two chips, allowing it to outperform even the faster Kneron design on those calculations.

In the longer term, stiffer competition may arise from MCUs based on Arm's new Ethos-U55 accelerator (see *MPR 3/9/20,* "Cortex-M55 Supports Tiny-AI Ethos"). The U55 can deliver up to 256 MACs per cycle, depending on the configuration size. Arm benchmarked Cortex-M55 and a 128-MAC Ethos-U55 at 50x faster than a standard Cortex-M7 when running neural-network inference. It estimates this combination consumes 40% more die area than the M7 CPU. Leading MCU suppliers NXP and STMicroelectronics have already licensed the M55 and U55, and we expect the first products with these cores to reach production in 2H21.

Making IoT Devices Smarter

Lots of companies are jumping into the edge-AI market. Many target camera-based applications that require at least 1,000 GOPS or more. Some instead target battery-powered systems that consume just a few GOPS but are limited to milliwatts of power. Xcore.ai fits somewhere in the middle. Its performance is best suited to audio tasks that are more sophisticated than mere wake-word detection. The chip can also handle simple vision workloads such as detecting human presence. Keep in mind that most end applications will reserve 4–12 threads for the RTOS, user code, and possibly DSP-based audio processing, leaving as little as 25% of the chip's capacity for neural networks.

Among the AI-chip vendors offering similar performance, XMOS stands out mainly on price. At $1 in high volume, Xcore.ai is much less expensive than competing chips. Bolting on a vector unit is less efficient than a more-specialized accelerator design, but it's relatively simple. XMOS has years of experience selling low-cost chips, and its unique flexible I/O can replace external glue chips in certain system designs. The company already has a revenue stream and is in better financial shape than some of its new competitors.

Most potential customers, however, are also considering standard MCUs from leading suppliers. Xcore.ai ranks among the least expensive 32-bit microcontrollers, giving customers a big AI-performance boost essentially for free. Because many neural networks that run on the XMOS chip simply won't work on a Cortex-M4 or even Cortex-M7 MCU, designers looking to add AI to their devices are willing to consider new alternatives. One concern is that XMOS customers must port their code to a new instruction set, so this path is best for new applications and ones that have a simple code base. Customers willing to take the plunge will find a fast, flexible chip that handles a range of IoT functions while reducing system cost. ♦