The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

CENTAUR ADDS AI TO SERVER PROCESSOR

First x86 SoC to Integrate Deep-Learning Accelerator

By Linley Gwennap (December 2, 2019)

Centaur is galloping back into the x86 market with an innovative processor design that combines eight high-performance CPUs with a custom deep-learning accelerator (DLA). The company is the first to announce a server-processor design that integrates a DLA. The new accelerator, called Ncore, delivers better neural-network performance than even the most powerful Xeon, but without incurring the high cost of an external GPU card. The Via Technologies subsidiary began testing the silicon in September; we estimate the first products based on this design could enter production in 2H20, although Via has disclosed no product plans.

Ncore, which operates as a coprocessor, avoids the vogue MAC array, instead relying on a more traditional programmable SIMD engine. But Centaur took the *multiple* in SIMD to the extreme, designing a unit that processes 4,096 bytes in parallel to achieve peak performance of 20 trillion operations per second (TOPS) for 8-bit integers (INT8). To feed the wide compute unit, the accelerator employs 16MB of private SRAM. Glenn Henry, the Ncore architect, likens this approach to "AVX-32,768," indicating the SIMD architecture is 64 times wider than Intel's.

The company also designed a new x86 microarchitecture called CNS, targeting much higher per-clock performance (IPC) than its previous PC-focused CPUs. The new design can decode four x86 instructions per cycle and execute 10 micro-ops in parallel, including three load/store operations. It runs at up to 2.5GHz in TSMC 16nm technology. The eight x86 cores share 16MB of L3 cache. The CHA (pronounced C-H-A) processor handles four channels for DDR4 DIMMs and provides 44 PCI Express 3.0 lanes, as Figure 1 shows. It enables two-socket systems and targets low-cost servers, particularly for edge tasks.

Centaur Technology has designed x86 CPUs and processors for Via for more than 20 years, but we've heard little from the Texan design shop since the debut of the dual-core Nano X2, which was built in a leading-edge 40nm process (see *MPR 1/24/11,* "Via Doubles Down at CES"). Henry, who managed the company since its inception, recently handed the reins to new president Al Loper, another long-time Centaurian. Henry, boasting 50 years of CPU-design experience, continues as the company's AI architect.

One Tractor Pulling 4,096 Trailers

When designing the Ncore accelerator, Centaur was concerned that the rapid pace of neural-network evolution

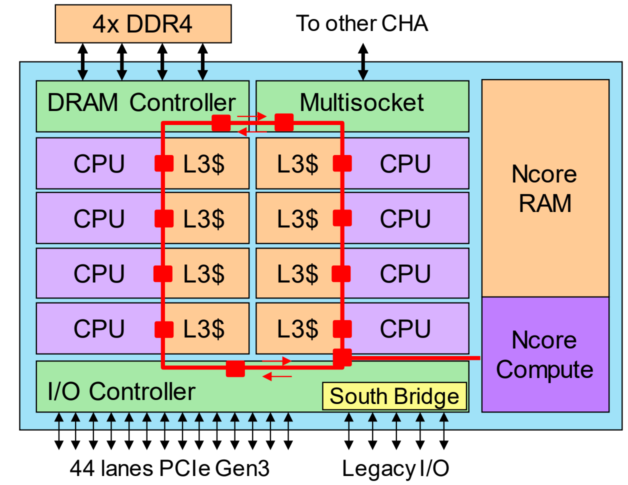

Figure 1. Centaur CHA processor. The design includes a powerful DLA called Ncore and eight x86 CPUs based on the new CNS microarchitecture. A bidirectional ring connects the CPUs, accelerator, memory, and I/O.

Price and Availability

Centaur is currently testing silicon of its CHA design; we expect the first CHA-based chips to reach production in 2H20. Via has announced no CHA-based products and therefore hasn't determined pricing. For more information, access *www.centtech.com*.

could obsolete a MAC-array design. SIMD architectures have more overhead than MAC arrays because they must move data into and out of registers each cycle, but they can handle a wider range of algorithms, including those that frequently perform non-MAC operations. As an on-chip accelerator, Ncore can more easily exchange data with the host CPUs, providing additional flexibility in partitioning the workload. The integrated design also reduces die area, cost, and power relative to a similar accelerator in an external chip.

The DLA implements a VLIW architecture—that is, a single 128-bit instruction controls the entire pipeline. An x86 CPU loads these instructions into Ncore's instruction RAM, which holds 768 instructions (12KB). A 4KB instruction ROM holds self-test code and some common subroutines, reducing the die area needed to store this code. Ncore fetches one instruction per cycle, decodes it, and uses a sequencer to control the compute pipeline and memory. The sequencer contains 16 address registers along with hardware to compute various address modes (e.g., base+offset) with optional auto-increment. It has loop counters and other special registers as well. As Figure 2 shows, the sequencer also controls two DMA engines in the ring interface, allowing instructions to directly transfer data to and from the x86 memory space.

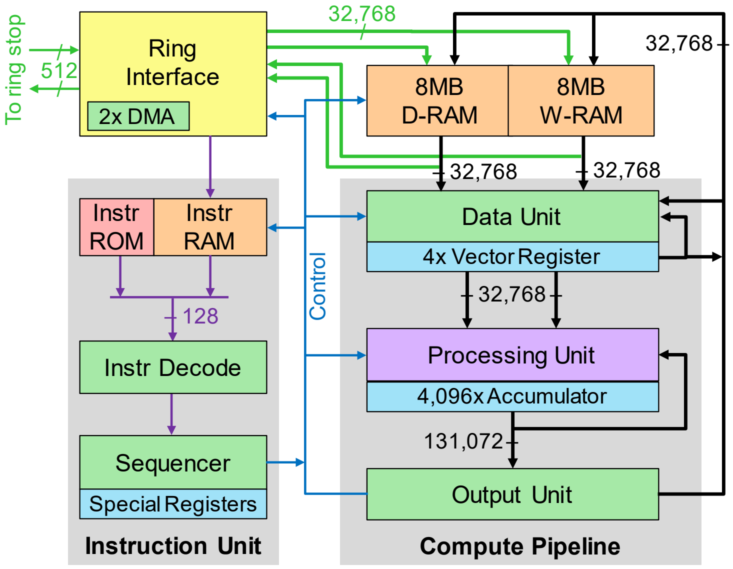

Figure 2. Ncore block diagram. The DLA implements 4,096-byte SIMD processing and has 16MB of private RAM to hold neural-network weights and activation values.

The design divides the accelerator's memory into two 8MB banks: the D-RAM and the W-RAM. Each RAM can supply a 4,096-byte vector on every cycle, producing 20TB/s of total bandwidth at 2.5GHz. Only one RAM can be written on each cycle, matching the output rate of the compute pipeline. Writes from the ring interrupt this sequence, but since it takes 64 bus cycles to load enough data for a single 4,096-byte write, these interruptions are rare. For high-reliability applications, both RAMs implement 64-bit ECC across the entire 4,096-byte output value.

Data from the RAMs first flows into the data unit, which performs various shift and permute functions. Specifically, it can perform up to three functions in a single 2.5GHz clock cycle, such as rotating an entire 4,096-byte vector by up to 64 bytes, broadcasting a single INT8 value (e.g., a weight) to fill a vector, compressing blocks (for pooling), and swapping bytes.

Although such wide vectors require sizable die area for a single register, the data unit contains four such registers. It can read or write any of these registers on each clock cycle. For example, it can merge a RAM value with a register value using one of the other registers as a byte mask. Thus, one or both RAMs can be powered down on many cycles while the unit continues to run at peak throughput.

Powerful SIMD Engine

The processing unit fetches two vector operands per cycle from the data-unit registers. It's optimized for INT8 values, which flow through the pipeline every cycle. The unit can also operate on INT16 and Bfloat16 values, but they require three cycles to compute, reducing throughput. Some users prefer these 16-bit data types for added precision. Because INT8 values are typically quantized, the processing unit converts them to signed INT9 values (by subtracting a variable offset) before further computation. MAC operations employ 4,096x32-bit accumulators (INT32 or FP32) and saturate on overflow. In addition to MAC operations, the processing unit can perform ALU operations including min/max. It also has eight predication registers that allow instructions to conditionally update the accumulators.

When a computation finishes, the output unit performs the final postprocessing. It typically converts the values from the 32-bit accumulators to INT8, INT16, or BF16 format to enable more-efficient storage in memory. If desired, the full 32-bit values can be output as four 4,096-byte vectors. The output unit also implements normalization functions such as ReLU, sigmoid, and hyperbolic tangent (tanh). It can forward results directly to the data unit for the next processing round, or it can store results in either of the two RAMs.

The entire Ncore design requires 34.4mm^2 in a TSMC 16FFC process. As Figure 3 shows, it's roughly half the size of the eight-CPU cluster. About two-thirds of the DLA's die area is the 16MB SRAM. The Ncore die plot reveals that the compute unit is divided into 16 slices to simplify the design process. The green areas show the dense metal routing in the data units; this routing mainly rearranges the data. The accelerator's central belt contains the instruction unit and ring interface.

Centaur has constructed a basic software stack that converts TensorFlow Lite models to an internal graph format that's then compiled into Ncore assembly instructions. It also provides run-time software for the x86 cores that manages the DLA and enables application code to run a precompiled network. The software handles inference functions but not training. Over time, the company plans to add support for other standard frameworks (such as TensorFlow and Pytorch) as well as the standard ONNX format and emerging MLIR format.

More Micro-Ops, Better IPC

Although the CNS CPU is a big step from Centaur's previous microarchitecture (see *MPR 3/10/08,* "Via's Speedy Isaiah"), it embodies many of the same design techniques. Isaiah could decode three x86 instructions per cycle and execute seven micro-ops in its out-of-order execution engine; the CNS extends this design to four decoders and 10 execution units. The intervening decade, however, enables a much larger transistor budget, so the new design implements a larger reorder window, more-accurate branch prediction, and a more sophisticated scheduler. It supports the 256-bit AVX and AVX2 operations as well as the initial AVX-512 extensions. The company also added proprietary instructions to stream data to the DLA. Unlike CPUs from Intel, the CNS executes only one thread per core.

To start the pipeline, the branch predictor determines the next instruction address, and the CPU fetches 32 bytes from the instruction cache. The pre-decoder then determines the instruction boundaries and loads four x86 instructions into the instruction queue. The decoders typically process four instructions per cycle, but certain pairs of x86 instructions can be decoded together, yielding a maximum of five instructions in a single cycle. The decoders convert these instructions into micro-ops.

The register-rename unit maps the requested registers onto the larger physical register file, which contains 192 integer entries and 192 FP/AVX entries. This unit can process six micro-ops per cycle and dispatch them to the unified scheduler. The 64-entry scheduler issues micro-ops when their input values are available; if one micro-op stalls, the scheduler continues to issue subsequent micro-ops while waiting to resolve the stall.

The scheduler issues micro-ops to the 10 execution units, as Figure 4 shows. The CNS CPU has four integer units; two have multipliers and the other two have bit-manipulation units (BMUs).

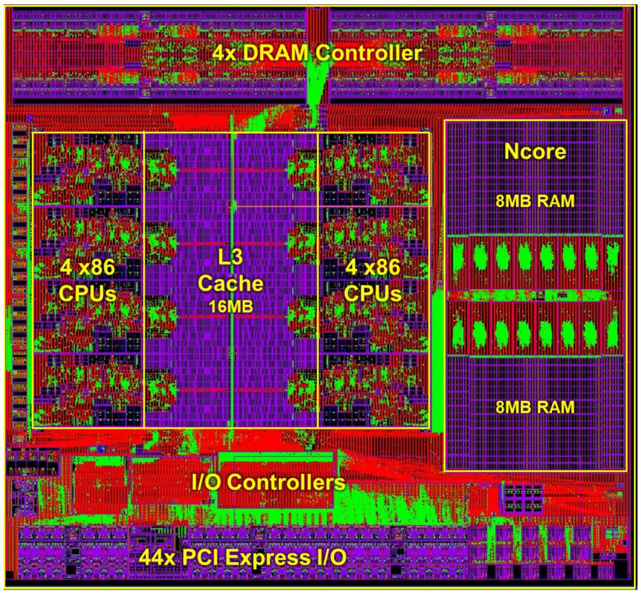

Figure 3. CHA die plot. The die requires 194mm² in a TSMC 16FFC process. The Ncore DLA, which contains 16MB of SRAM, is about half the size of the eight-CPU cluster.

The integer units can process up to two branch instructions per cycle. The CPU features two FP/AVX units that include floating-point multiply-accumulate (MAC) units; a third FP/AVX unit handles FP divide and crypto (AES) acceleration. All three can process AVX integer instructions. Because these units are 256 bits wide, AVX-512 instructions decode into two micro-ops. The MAC units can generate

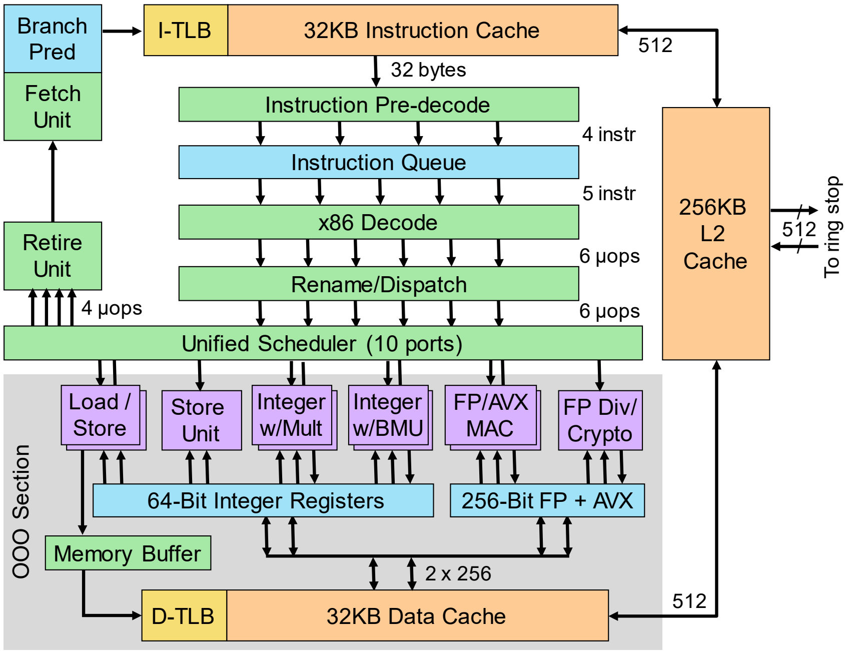

Figure 4. Centaur CNS microarchitecture. The "Haswell-class" CPU can decode four x86 instructions per cycle and execute 10 micro-ops per cycle using three load/store or store units, four integer units, and three FP/SIMD units.

one result per cycle with a three-cycle latency for add or multiply operations and five cycles for the full MAC.

The CPU has two load units and one store-address unit. The load units can also perform store-address operations, as this function is essentially the same as generating the load address. Whereas the integer and FP/AVX units perform their operations immediately, the 116-entry memory order buffer (MOB) holds load and store operations until they can access the 32KB data cache. This cache can service two 256-bit loads or stores (in any combination) per cycle, transferring information to the integer or FP/AVX registers. As with the compute instructions, AVX-512 loads and stores are broken into two micro-ops, but the cache can supply a total of 512 bits per cycle. The two L1 caches are backed by a private 256KB L2 cache. The 192-entry reorder buffer (ROB) eventually retires instructions in order.

The CNS pipeline requires 20 stages to complete a basic integer operation and 22 stages for a load that hits the data cache. It allocates five stages for cache accesses, boosting the clock speed. Even so, the CPU achieves 2.5GHz in 16nm, far slower than Intel and AMD x86 designs. Relative to these competitors, Centaur invests far less time in optimizing its physical designs, limiting its CPUs' top speed.

Each CPU has a 2MB slice of L3 cache, and it can access them directly. The eight CPUs and their L3 slices connect via the ring, which implements two 512-bit-wide buses that run from ring stop to ring stop, as Figure 1 shows. The rings cycle at the same speed as the CPUs, providing a theoretical bandwidth of 320GB/s. Since data can traverse the bus for multiple cycles, the usable throughput is less than half that figure. The processor supports four DDR4-3200 DRAM channels with ECC, providing 102GB/s of peak memory bandwidth. It also sports 44 PCIe Gen3 lanes, which are configurable into ports of varying widths. The processor includes standard server south-bridge functions, creating a fully integrated solution.

	Centaur CNS	Intel Haswell	Intel Skylake	Intel Sunny Cove
Threads/CPU	1 thread	2 threads	2 threads	2 threads
x86 Decoders	4–5 instr	4 instr	5 instr	5 instr
Instr Extensions	AVX-512	AVX2	AVX-512	VNNI
Micro-op Cache	None	1,536 µops	1,536 µops	2,304 µops
Max Ops/Cycle	10 µops	8 µops	8 µops	10 µops
Reorder Buffer	192 ops	192 ops	224 ops	352 ops
Load Buffer	72 loads	72 loads	72 loads	128 loads
Store Buffer	44 stores	42 stores	56 stores	72 stores
Scheduler	64 entries	60 entries	97 entries	Undisclosed
Integer Rename	192 int regs	168 int regs	180 int regs	Undisclosed
FP Rename	192 FP regs	168 FP regs	168 FP regs	Undisclosed
Interconnect	2x512b ring	2x256b ring	4x256b mesh	4x256b mesh
First Production	2H20*	2Q13	3Q15	4Q19

Table 1. CPU-microarchitecture comparison. The CNS is equal to or better than Intel's Haswell CPU on most metrics, but it falls short of the newer Skylake and Sunny Cove. (Source: vendors, except *The Linley Group estimate)

Aiming for the Sky

Centaur refers to the CNS as "Haswell-class"—a fair characterization. As Table 1 shows, the design matches or exceeds Intel's Haswell in most microarchitecture parameters. It offers more rename registers and greater interconnect bandwidth, but it lacks two important features: multithreading (which Intel calls Hyper-Threading) and a micro-op cache. The former provides a 20–30% performance boost on many server workloads, while the latter reduces power by disabling the x86 decoders on most cycles.

Of course, Haswell is an obsolete CPU that first appeared in 2013. Intel's current server processors employ Skylake, which is slightly beefier than the CNS in some ways and continues to offer multithreading and a micro-op cache (see *MPR 9/7/15,* "Skylake SpeedShifts to Next Gear"). It can decode five instructions per cycle for any instruction combination, whereas the CNS is limited to four except in certain cases.

The biggest difference is that Skylake can operate at up to 5.0GHz in Intel's 14nm++ process, doubling the CNS's peak speed. Most server processors, however, operate at much lower speeds to save power, limiting Intel's speed advantage. Furthermore, the x86 leader is preparing new server processors based on the 10nm Sunny Cove CPU that are scheduled to debut in late 2020, about the same time we expect CNS-based processors to reach production. Sunny Cove is a sizable upgrade to Skylake, delivering 18% better IPC according to the vendor.

The CHA processor will compete against the low end of the Intel Xeon lineup. For example, Intel today offers the Xeon Silver 4208 (Cascade Lake) with eight Skylake cores, six DDR4-2400 channels, and 48 PCIe Gen3 lanes (see *MPR 4/8/19,* "2nd-Gen Xeon Scalable Adds Cores"). Centaur's design has similar DRAM and PCI bandwidth. At an 85W TDP, the CPUs in the 4208 operate at a 2.1GHz base frequency—slower than the CNS—but they can turbo to 3.2GHz when lightly loaded. The Xeon 4208 carries a list price of $417. Centaur withheld the CHA's TDP, but we expect it will operate at its peak frequency while using less power than the 4208.

Centaur's big differentiator is its DLA. On the basis of the company's MLPerf Inference results, Ncore is roughly as powerful as a $5,000 Xeon Platinum processor (using VNNI) for imaging tests such as MobileNet and ResNet-50. We estimate it's about 5x faster than a comparable Xeon Silver chip for these tasks. Although Intel can combine Xeon Silver with its new NNP-I accelerator to achieve much greater neural-network performance (see *MPR 9/2/19,* "Spring Hill Sets AI Efficiency Mark"), this combination will be far more expensive. Intel has withheld the NNP-I's price, but it could be $500–$1,000, whereas a GPU-based accelerator such as Nvidia's T4 card adds about $2,000 to the system price.

Good Performance at Low Cost

Centaur's goal is to deliver the best neural-network performance per dollar in its class. Via will ultimately determine the price of CHA-based products, but if they sell for about the same price as a Xeon Silver, customers will essentially get the DLA for free. Even though external DLAs based on the NNP-I or the T4 deliver considerably better performance, they're far from free; in fact, they cost more than the processor. Thus, for essentially no cost, Ncore customers could get a 5x speedup on neural networks relative to a similarly priced system with no external accelerator. Centaur is still optimizing its software (it released MLPerf numbers only a month after receiving working silicon), so its scores could improve further by the time the product reaches the market.

A SIMD design, even one having Ncore's great width, is unlikely to match Spring Hill's more optimized MAC arrays in performance per watt. But integrating the DLA greatly reduces cost, helping Centaur meet its goal. The SIMD design is also more flexible than most MAC-array designs. Even with its early software, Centaur submitted respectable scores on the SSD (single-shot detection) and GNMT (language translation) models; on these tests, Intel didn't even post scores for the NNP-I. In this way, Ncore is nearly as flexible as CPUs and GPUs for a broad range of neural networks.

For CPU-based server workloads, the CNS will likely fall short of Xeon Silver processors in performance, mainly because it lacks multithreading. For a single thread, its IPC should be close to Skylake's, but its CPU speed may fall a bit short on some workloads. The processor is similar to low-end Xeons in core count, memory bandwidth, and PCIe bandwidth, and its integrated south bridge is a plus. Centaur's new design is a good fit for edge servers, meaning systems located either at the network's edge, such as in a 5G base station, or on customer premises, such as in a factory or store. These systems must often be small and low cost, so the processor's high integration works well.

Despite Intel's best efforts, an x86 CPU isn't the best solution for all problems. Many processors today include graphics units (GPUs), video accelerators, DSP cores, and other specialized architectures that either increase performance or reduce power consumption for common workloads. As neural networks become more prevalent, adding a DLA provides similar benefits. Such coprocessors are already common in Arm-compatible smartphone chips, but neither AMD nor Intel has seen fit to add a DLA to its PC or server processors. By announcing the first x86 host processor with an integrated AI accelerator, Centaur is leading the way for these applications. ♦