The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

ANDES PLOTS RISC-V VECTOR HEADING

NX27V CPU Supports up to 512-bit Operations

By Mike Demler (May 25, 2020)

RISC-V International, formerly the RISC-V Foundation, expects to release its v1.0 vector-extension (RVV) specifications in August. But Andes has already delivered the first CPU core to support the new features. The NX27V is a 64-bit design that implements the RISC-V RV64 ISA, RVV, and proprietary Andes extensions. The scalar processing unit employs the same architecture as the company's NX25: a single-issue five-stage microcontroller CPU. The new vector processing unit (VPU) adds configuration options for 128-, 256-, and 512-bit SIMD operations. Andes recently delivered production RTL to its lead customer, which is using it in a 7nm design. The NX27V should be available for general licensing in October.

Andes chairs the RISC-V P-extension task group developing SIMD-DSP specifications. It says the SIMD and vector extensions target different workloads, but with some overlap. The company contributed the AndeStar V3 ISA, which it designed for power-efficient embedded DSPs, as the P-extension starting point. Whereas that extension enables one 64-bit SIMD operation per cycle on 8-, 16-, or 32-bit elements, RVV reduces issue bandwidth by letting the CPU issue one instruction that executes a series of SIMD operations on larger vectors. Vector CPUs consume more area and power than SIMD CPUs owing to their additional control logic, execution pipelines, and register files, but they're better suited to performing operations on large data sets.

Andes customers are evaluating the NX27V for digital basebands, image/vision processing, and neural-network acceleration. In those cases, the company expects designers will employ multiple NX27Vs along with a small RISC-V or Arm Cortex-A CPU that handles control operations. The new core also supports FFTs and other data-intensive DSP functions.

Table 1 shows the performance increase that the 512-bit vector extensions bring to an NX27V FPGA prototype, compared with the same functions running on scalar C code. Basic single-precision floating-point math runs 19x faster, as do MobileNet convolutions. As another example, a 32x32x32 matrix-multiplication operation using scalar C-code requires 207,331 cycles to execute the load/store, multiply-accumulate (MAC), address increment, and loop-control functions. Using RVV extensions, the NX27V completes that operation in just 3,620 cycles, a 57x boost.

Going to Great Lengths

The NX27V combines the proprietary AndeStar V5 architecture with the RV64 integer ISA, including support for RISC-V memory atomics, 16-bit compressed instructions, a single-precision FPU, a hardware multiplier/divider, and user-level interrupts (A, C, F, M, and N extensions). The scalar core employs the same five-stage in-order pipeline as the Andes AX25, AX25MP, and NX25 CPUs (see *MPR 4/15/19,* "Andes Strengthens Its RISC-V Arsenal"). The A-series cores include an MMU that can run Linux, unlike the NX27V and other N-series models. The company plans to offer vector extensions in a later AX27V release.

Function	Speed Increase
Basic FP32 math functions	19x
MobileNet CNN_RGB	18x
MobileNet CNN_Depthwise	18x
MobileNet CNN_1x1	21x
FP32 matrix multiplication (32x32x32)	57x

Table 1. NX27V speed improvement. Andes tested a variety of compute kernels on its NX27V FPGA prototype. Basic floating-point math and convolutional-neural-network operations garnered an average 19x boost, but 32x32x32 matrix multiplication garnered 57x. (Source: Andes)

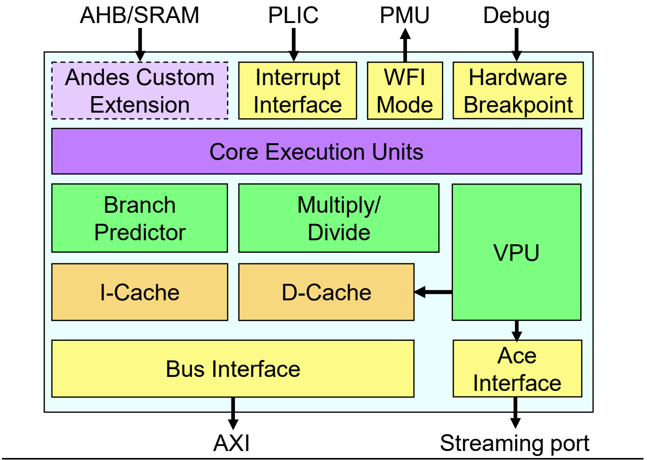


Figure 1. Andes NX27V CPU. PLIC=platform-level interrupt controller; WFI=wait for interrupt. The scalar core implements the same microarchitecture as the predecessor NX25, but it adds RISC-V's vector-processor feature and Andes custom extensions.

By implementing RVV (see *MPR 4/27/20,* "RISC-V Vectors Know No Limits"), the new CPU can serve as an accelerator or coprocessor in an SoC. In the proposed RVV ISA, the vector-length (VLEN) parameter theoretically allows up to 2^{32} bits, but designers must fix the value in hardware during implementation. Andes views 512 bits as a practical limit for the area- and power-efficient designs the NX27V targets; larger vectors are more difficult to implement, requiring more area for the wider data paths.

Customers can equip the NX27V with 8–64KB instruction and data caches, along with optional ECC or parity protection. The CPU's load/store unit supports three independent memory-access paths from system SRAM: one to the general-purpose registers, one to the floating-point registers, and one to the VPU registers. To prevent the scalar and vector pipelines from stalling, the NX27V allows up to 16 outstanding data accesses. The VPU can access data in shared system memory through the CPU's 256-bit AXI bus interface unit (BIU), and it can access data in the data cache, as Figure 1 shows. Because the VPU requires much greater memory bandwidth than the scalar unit, Andes used its custom extensions (Ace) to implement a dedicated streaming port that handles 512-bit operations.

The NX27V dispatches vector instructions to the VPU after the integer-execution stage, as Figure 2 shows. Because some applications combine scalar and vector floating-point operations, the company chose to integrate the scalar FPU downstream in the vector pipeline. The VPU handles the standard RVV data formats—8- to 64-bit integer as well as half-, full-, and double-precision floating point—but the custom extensions add Bfloat16 and INT4, which are popular for neural-network acceleration.

In the VPU, multiple vector execution units operate independently, allowing out-of-order (OOO) execution. A scoreboard keeps track of dependencies in the vector-instruction queue. Vector operations begin execution after all previous integer instructions retire, ensuring sequential program execution when exceptions or interrupts delay operations.

Masks and Chains

By applying the RISC-V length multiplier (LMUL) at run time, NX27V users can increase the vector length from 512 bits to as many as 4,096. The microarchitecture accelerates these large operations by chaining. In the Andes implementation, this feature works by allowing each SIMD operation to feed its output directly to the next vector instruction in the chain.

Figure 3 shows an example operation chaining when LMUL is eight, increasing effective vector length to 8xVLEN, or 4,096 bits. This operation divides the 32 vector registers into four groups: v0–v7, v8–v15, v16–v23, and v24–v31. In this example, the vfcvt instruction converts INT32 data in the v16 group to FP32 format, storing the results in the 4,096-bit v8 group. The next vfadd instruction adds the floating-point data in v8–v15 to the data in v0–v7, storing the results in v24–v31. The 512-bit VPU must execute these instructions as eight consecutive micro-ops, but rather than wait for all the floating-point conversions to finish, chaining allows the first one to feed its result directly to the vfadd instruction in cycle 3, so it can start eight cycles earlier than without chaining.

The NX27V also supports the RVV mask feature, which enables element-wise predication. RVV implements this feature by allocating one vector register to hold the mask, adjusting its element size and length so that each element in the mask register corresponds to one execution lane. The value of each mask element determines whether or not that lane executes its designated operation.

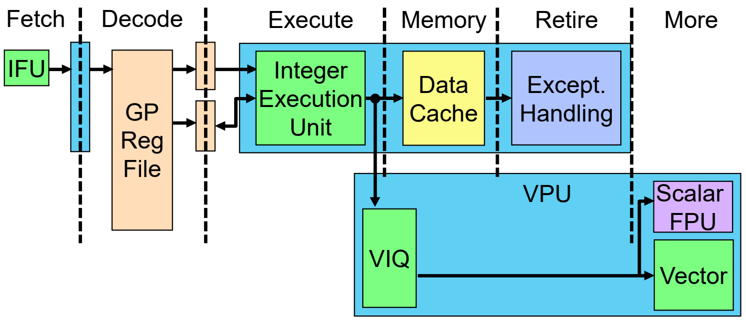


Figure 2. NX27V scalar and vector pipelines. The five-stage in-order scalar pipeline is unchanged from the NX25. It dispatches instructions to the vector instruction queue (VIQ) after the integer instructions execute. Vector instructions begin execution once the integer instructions retire.

Competition Is Coming

Although the NX27V is currently available to lead customers only, it holds an early lead in RISC-V vector processors. Andes competes with SiFive, which offers just a vector evaluation platform (VEP) for its forthcoming VI-series. That VEP is useful for software development and simulations with the RVV features, but Andes provides synthesizable RTL to lead customers, and it has completed a 7nm layout.

The company's performance estimates come from a physical model, whereas the VEP only lets users evaluate vector computing software. SiFive is developing three CPUs with RVV, but it has revealed few details about the low-end VI2 and none about the midrange VI7 and high-end VI8. Like the NX27V, the VI2 sets VLEN to 512. To target modest compute requirements in IoT devices, however, its default configuration has a 128-bit data path that requires four cycles to compute each 512-bit result.

Although the VI2 also has an in-order scalar core, we expect the NX27V will deliver much higher vector performance owing to its 512-bit VPU. SiFive estimates its vector extensions will boost F32 matrix multiplications by 31x. But whereas the VI2 executes four 32-bit MAC operations per cycles, it needs 8,192 cycles just for the 32x32x32 MAC operations, more than twice the cycles the NX27V uses to complete the entire instruction. Both vendors plan to begin licensing their CPUs in 4Q20; we expect SiFive's will be a preproduction beta release.

Andes aims to compete with Arm CPUs as well: both the new Cortex-M55 with its Helium vector extension (see *MPR 3/9/20,* "Cortex-M55 Supports Tiny-AI Ethos") and Cortex-A models with optional Neon SIMD units. Like the NX27V, the M55 has a five-stage integer pipeline (including the retire stage Arm omits from its specifications), but Helium supports a maximum 128-bit vector length. Because the Cortex-M line targets low-power embedded systems, the M55 needs a two-cycle chained operation to complete one 128-bit vector computation. A future RVV-equipped AX27V will compete with small 64-bit Arm cores, such as Cortex-A35, but the A35 handles vector operations in a single 128-bit Neon unit. To match Andes's 512-bit vector-compute capabilities, designers need a quad-core configuration.

In designs requiring both DSP operations and neural-network acceleration, the NX27V will compete with other licensable cores that combine a small CPU with vector-processing units, such as the Ceva SensPro (see *MPR 4/20/20,* "Ceva SensPro Fuses AI and Vector DSP"). Like the NX27V, Ceva's midrange SP500F can execute 512 INT8 operations per cycle, and it supports INT16 as well as FP32 operations. Both products can accelerate FFTs, but SensPro implements a DSP ISA that handles sensor and other signal-processing workloads, a feature the NX27V lacks. The AX25F implements a DSP ISA, however, so we expect the forthcoming AX27V will do so, too.

Price and Availability

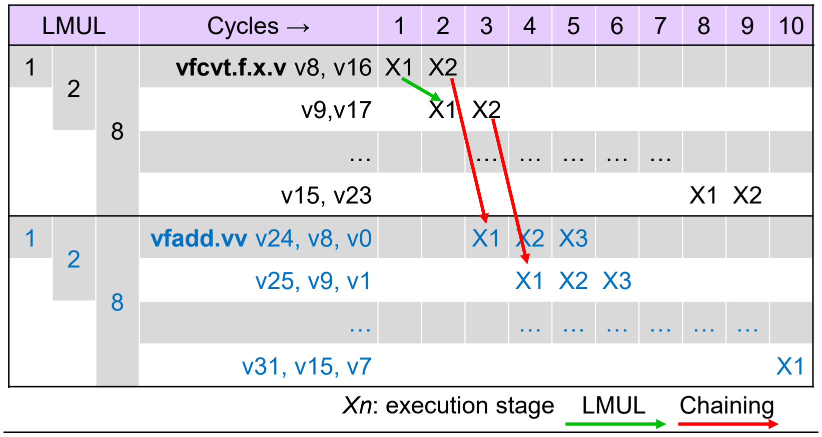
Andes doesn't disclose licensing or royalty fees for its IP cores. Production RTL for the NX27V is now available to lead customers, and the company plans to begin general licensing in 4Q20. For more information about Andes RISC-V CPUs, access *www.andestech.com/en/products-solutions/andescore-processors*.

The Freedom to Differentiate

Like all open-source RISC-V ISAs, the RVV extension gives designers much more freedom than commercial architectures from Arm and other intellectual-property (IP) suppliers. In addition to providing the flexibility to pick and choose features, RISC-V reserves a portion of the opcode space for custom features, allowing Andes to differentiate its products. The CPUs comply with standard RISC-V ISAs, but their custom extensions mean code that employs the noncompliant features won't run on CPUs from other vendors.

RVV's freedom to set the vector length with no hard limit is unique, but it requires scaling of all other microarchitecture features accordingly. A big bolted-on vector engine would be more than a tiny CPU can handle, so Andes limits designers to a 128-, 256-, or 512-bit VPU—a reasonable range for low-power embedded systems. Using the smallest 128-bit option, the NX27V competes directly with Arm's Cortex-M55, but the vector-chaining feature provides more power than Arm v8.1-M, which must break 128-bit instructions into consecutive 64-bit "beats."

On the basis of its MobileNet performance, we expect the first Andes customer has modest DLA requirements, because the 7nm design only delivers 128 billion operations



LMUL			Cycles →	1	2	3	4	5	6	7	8	9	10
1	2	8	**vfcvt.f.x.v** v8, v16	X1	X2								
			v9,v17		X1	X2							
			…			…	…	…	…	…			
			v15, v23								X1	X2	
1	2	8	**vfadd.vv** v24, v8, v0			X1	X2	X3					
			v25, v9, v1				X1	X2	X3				
			…				…	…	…	…	…	…	
			v31, v15, v7										X1

Figure 3. NX27V vector chaining. The RVV extensions allow the effective vector length to grow by setting LMUL to 2, 4, or 8, but doing so equivalently increases the number of cycles necessary to execute an instruction. By chaining vector operations, the NX27V enables one SIMD step to feed its result directly to the next, boosting throughput.

per second (GOPS) per core. Whereas dedicated deep-learning accelerators (DLAs) typically deliver at least 1,000 GOPS, Andes customers need eight cores to match that performance, complicating software development. The tool kit includes an AI compiler that works with LLVM, but the company is also developing an Apache TVM compiler that supports MXNet, Pytorch, TensorFlow, and other popular frameworks. It backs the NX27V with a standard set of development tools, including the AndeSim cycle-accurate simulator, an assembler, a compiler, a debugger, and a computation library.

Andes is a founding member of the RISC-V Foundation, and CEO Frankwell Lin recently joined the board of directors. The company contributes to various RISC-V International task groups, helping to drive adoption, but it has added custom extensions to all its RISC-V-based CPUs. Although this tactic enables the company to differentiate, it also causes ecosystem fragmentation. Nevertheless, Andes customers are more interested in performance and in a lower-cost alternative to Arm and other licensed CPUs.

The NX27V will be attractive to customers that need more vector-processing performance than Arm's Cortex-M55 provides, but designers should evaluate the power, performance, and area (PPA) tradeoffs. A dedicated DLA is better for neural-network inference, typically offering a more streamlined data-flow architecture with larger arrays of MAC units. Yet the Andes core is a capable and flexible acceleration coprocessor that can boost audio/video compression, handle networking memcpy operations, and perform other array-processing tasks. By delivering a product ahead of the final RVV ratification, the company has established more than a six-month lead over its rivals. ♦