The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

VSORA DRIVES TO DELIVER PETAFLOPS

AI Engines and DSPs Target Level 4/5 Autonomous Vehicles

By Mike Demler (December 7, 2020)

Vsora has jumped into the race to provide DSPs and deep-learning accelerators (DLAs) for autonomous vehicles (AVs). The French startup aims to pass its more established rivals by offering the new ADxxxx intellectual property (IP), which it designed to handle perception and sensor fusion in Level 4 and 5 self-driving cars. Unlike most DSP+DLA combos, all of the ADxxxx's compute units execute floating-point operations, but they implement a configurable architecture that allows designers to select the exponent and mantissa precision to fit their area, power, and precision requirements.

The company's first product is the AD1028, which integrates 24-bit floating-point ALUs in the DSP and 8-bit floating-point multiply-accumulate (MAC) units in the DLA. Running at 2.0GHz in 7nm technology, it delivers one quadrillion FP8 AI operations per second (1.0Pflop/s), along with four trillion FP24 DSP operations per second (4.0Tflop/s). Typical power consumption is 35W, yielding an unparalleled 30 TOPS per watt.

Most Vsora executives previously worked at DIBcom, a designer of programmable DSP receivers for the European Digital Video Broadcast-Handheld (DVB-H) standard and its terrestrial counterpart (DVB-T). DIBcom's work on mobile DVB-H receivers led to product development for automotive entertainment systems. In 2011, Parrot acquired the company, which it subsequently sold to Faurecia Automotive (now Faurecia Clarion). But in 2015, Vsora's founders launched their new venture, initially using their DSP expertise to develop 5G-baseband IP, which constitutes its digital-communications product line.

Khaled Maalej is CEO; he was previously CTO at DIBcom. In 2018, the startup closed a $1.7 million Series A funding round, led by French VC firms Omnes Capital and Partech Ventures. We estimate it has about 20 employees.

A Quadrillion Flops Need Big SRAMs

The AD1028 is the middle child of the ADxxxx family, comprising a 1,024-ALU DSP and a DLA that includes 16 MAC arrays, each integrating 16K configurable MAC units, as Figure 1 shows. Vsora is licensing that model now, and it plans to release two more models by year-end. As its names implies, the AD0514 integrates half as many ALUs and MAC units as the AD1028, and the AD2056 combines a massive 2,048-ALU DSP with 512K MAC units.

The ADxxxx architecture allows designers to select the exponent and mantissa precision. In the AD1028, however, the DSP ALUs are 24-bit floating-point units (7 exponent bits, 16 mantissa bits, and 1 sign bit), and the MAC units execute 8-bit floating-point operations (4 exponent bits, 3 mantissa bits, and 1 sign bit). The company's lead customer

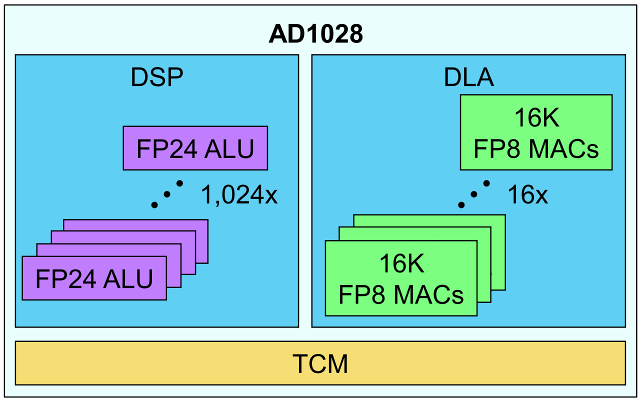

Figure 1. Vsora AD1028. The design includes a DSP that executes 24-bit floating-point operations using 1,024 ALUs and a DLA that comprises a total of 256K MAC units executing FP8 operations. Running at 2.0GHz, a 7nm implementation delivers a total of 1.0Pflop/s.

Price and Availability

Vsora doesn't disclose licensing or royalty fees for its ADxxxx IP. Production RTL is available now. For more information, access *www.vsora.com/products/ad-adas/autonomous-drive-platforms*.

has validated the device using 7nm place-and-route design rules, measuring 35mm^2 of die area for the logic, excluding memory.

Vsora revealed few details, but the DSP processes arbitrary-size matrices by directly reading data from and writing it to the tightly coupled memory (TCM). The ALUs can run in parallel, executing the same instruction, or each can execute a different instruction. To ease programing, customers can define macroinstructions comprising a sequence of DSP instructions (up to 1,024 operations). The ALUs can execute several complex instructions per cycle, such as multiplying a cosine by a floating-point variable.

The company recommends sizing the TCM to minimize the number of cycles to fetch data from DRAM—a critical metric for latency-sensitive computer-vision tasks in AVs. To maximize efficiency, the compiler treats the TCM like a large register file. Storing data on chip ensures high utilization of the massive MAC arrays. For its performance simulations, the company used 35 million 24-bit words of storage, equivalent to a 105MB SRAM. Although that amount of memory consumes a large die area, it's common in AV processors. For example, Tesla's 12nm FSD ASIC integrates a 64MB SRAM that occupies one-third of the 260mm^2 die (see *MPR 5/13/19,* "Tesla Rolls Its Own Self-Driving Chip").

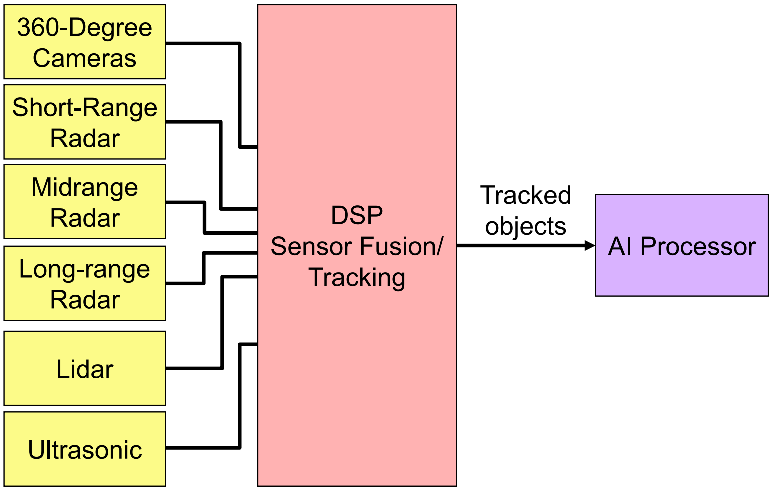

Figure 2. Vsora software-development platform. Designers begin by developing their application on a PC or similar workstation using C++, Matlab, and a TensorFlow framework. The LLVM-based compiler enables simulations to run on Amazon's cloud FPGA as well as in the profiler and RTL simulators.

To enable high-level ADAS and AV capabilities, manufacturers are moving from one- or two-megapixel cameras to eight megapixels (MP). The ADxxxx's MAC arrays can handle high-resolution images in single-batch mode, but because the design is algorithm and application agnostic, it can run other types of neural-network algorithms as well. For example, Vsora has also used it to run Facebook's deep-learning recommendation model (DLRM).

To keep the MAC units occupied, the compiler slices and distributes computational graphs to the available compute resources through a tiling technique similar to that of the Imagination Series4 DLAs (see *MPR 11/16/20,* "Imagination Series4 Tiles Tensors"). Using its code-profiling tools on a variety of neural-network models, Vsora typically measures about 75% MAC utilization. Running Yolo v3 on 8MP images, the DLA delivers 167 images per second with 6ms latency, much less than the 10ms that AVs require.

Fast Fusion Ensures Safety

To achieve Level 4 autonomy, a self-driving car must be able to navigate on its own without any human backup, at least under limited conditions such as in a premapped geofenced area. GNSS receivers and digitized maps enable coarse localization, but precise navigation requires fusing data from multiple sensors to create a real-time 360-degree model that identifies pedestrians, road hazards, traffic signs/signals, and other vehicles in the environment. The AV must also employ a control loop that uses accelerometers and other sensors to monitor its orientation, speed, and trajectory. Much disagreement still plagues the AV industry regarding the necessary type and number of sensors, but most Level 4 vehicles will combine multiple cameras, lidars, radars, and ultrasonic sensors, as Figure 2 shows (see *MPR 10/30/17,* "Lidar Points the Way for Self-Driving").

By combining a high-performance DSP with an AI engine, the AD1028 can handle sensor processing and fusion directly from raw sensor data. An alternative approach is to perform preprocessing and object detection at the sensor, transmitting only the tracked-object list to the central processor. That technique reduces communications bandwidth, but it risks losing critical information owing to the lower dynamic range and precision of the integer compute units that such devices typically employ. Distributed processing also increases system latency.

AVs fuse sensor data to form 3D occupancy grids (see *MPR 6/19/17,* "Xavier Simplifies Self-Driving Cars"), combining objects identified by the cameras, the point cloud created by lidars, and the range and velocity data provided by radars. These algorithms are critical to enabling the AV to calculate a safe navigable path, avoid collisions by predicting other vehicles' movements, and track at-risk road users. Processing all the sensor data requires that DSPs process the radar/lidar signal and a DLA to execute AI algorithms.

To test the sensor-fusion capabilities of the AD1028, Vsora evaluated its FPGA prototype running a 16-million-element particle filter comprising an eight-million-cell occupancy grid. This technique more accurately predicts the paths of multiple objects than Kalman filters do (see *MPR 6/22/20,* "EnSilica Drives Automotive Radar"). Other tests include an eight-million-cell clustering algorithm that groups target detections associated with the same object, along with a chamfer algorithm that estimates the distance of images in that grid. The AD1028 runs the particle filter with just 6ms latency, but the clustering algorithm finishes in 1ms and the chamfer-distance algorithm in just 0.26ms.

A Heterogeneous Programming Model

The ADxxxx IP works as a coprocessor to a host CPU, but it operates asynchronously, communicating with the host through mailboxes. The Vsora software tool kit eases combination of the AI and DSP algorithms that run on the AD1028's various compute units with the code that runs on the host CPU.

Programmers develop applications on a PC or similar workstation, combining C++ with Matlab-like code for the DSP and a TensorFlow model for the DLA. The LLVM-based compiler transparently separates the code, distributing operations to the appropriate compute engines. As Figure 3 shows, the software stack works with multiple simulation platforms. Designers can run a prototype on Amazon's cloud-based F1 FPGA, and they can run a high-level model on the profiler, which enables optimization of the core configurations, memory, and pretrained-model quantization. A TLM 2.0 System-C model lets software engineers develop their code before they receive working silicon.

Insatiable Demand for Horsepower

Eleven years have passed since Google began its self-driving-car project, and over that time, AV developers have learned that they need far more AI-compute horsepower than they anticipated. In 2018, for example, Nvidia announced its first Pegasus system targeting Level 5 AVs. It combined two Xavier SoCs with two discrete GPUs, delivering a total of 320 TOPS at 400W (see *MPR 2/19/18,* "Nvidia Xavier Drives to Carmel"). Although Pegasus has been popular for AV development, the GPU developer announced earlier this year its next-generation Drive AGX robotaxi platform, which uses two Orin processors and two Ampere GPUs to deliver 2.0 quadrillion operations per second, but at the expense of 800W.

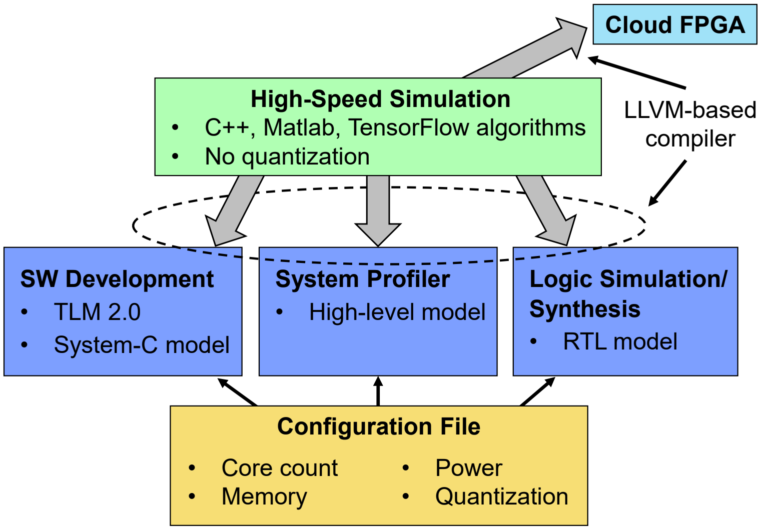

Figure 3. Vsora software-development platform. Designers begin by developing their application on a PC or similar workstation using C++, Matlab, and a TensorFlow framework. The LLVM-based compiler enables simulations to run on Amazon's cloud FPGA as well as in the profiler and RTL simulators.

	Vsora AD1028	Ceva NeuPro-S4000	Imagination Series4
Clock Speed	2.0GHz	1.5GHz	1.5GHz
DSP Pipeline Depth	Undisclosed	14 stages	Not applicable
Int Data Types	Optional	INT8, INT16	INT4, INT8, INT16
FP Data Types	FP8, FP24	FP16, FP32	None
VLIW Architecture	1,024 parallel ops	8 parallel ops	Not applicable
Scalar Units	Undisclosed	4 scalar units	Not applicable
Vector Units	Not applicable	3x 512-bit	Not applicable
MAC Units (single-/multicore)	16,384 / 262,144 FP8 MACs	4,096 / 32,768 INT8 MACs	4,096 / 32,768 INT8 MACs
Max AI Perf (single-/multicore)	65.5 TOPS / 1,048 TOPS	12.5 TOPS / 100 TOPS (INT8)	12.5 TOPS / 100 TOPS (INT8)
Production RTL	4Q20	3Q18	4Q20

Table 1. Automotive neural-network accelerators. The NeuPro-S4000 includes a DSP and DLA, but it delivers just 10% of the AD1028's performance, and its integer MAC units offer much less dynamic range. The Series4 product targets automotive systems, but its throughput is similar to NeuPro's. (Source: vendors)

Several IP vendors have introduced DLAs targeting ADASs and AVs, but none scales close to the AD1028's performance. Before Vsora introduced its IP, 100 TOPS was the highest throughput in a licensable DLA. Designers can scale throughput using multiple instances of a competitor's IP, but these cores aren't designed to distribute and synchronize operations across 10 copies, greatly complicating the programming task.

Like the AD1028, Ceva's NeuPro-S4000 combines a DSP with a multicore MAC engine (see *MPR 10/7/19,* "Ceva and Synopsys Spin More TOPS"). As Table 1 shows, the MAC units only handle INT8 and INT16, however, providing just a tiny fraction of the AD1028's dynamic range. Although NeuPro's XM6 DSP supports an optional vector FPU, it can only perform eight FP32 operations or 16 FP16 operations per cycle; the AD1028, on the other hand, has 1,024 FP24 ALUs. Ceva specifies the NeuPro-S4000 clock frequency in 16nm technology, but we expect a 7nm version can match the AD1028's 2.0GHz.

The Imagination Series4 DLAs lack a DSP, but their MAC units maximize efficiency by handling mixed precision using INT4, INT8, or INT16 operations. The UK-based vendor designed Series4 specifically for the automotive market, and the IP has function-safety features that support ISO 26262 compliance. The AD1028 targets ASIL D as well, including ECC-protection for the instruction and data memories and the DMA paths. Customers can optionally add a small safety-monitor DSP with dedicated memory, providing redundancy by sampling results from the big DSPs.

Getting Self-Driving Cars on Track

Many IP vendors are targeting the ADAS and AV market, but none offers capabilities and performance that matches Vsora's AD1028. Most have focused their products solely on computer-vision perception, progressively introducing DLAs with larger MAC arrays, but that approach only addresses part of the problem. A complete solution begins with the ability to build an environmental model that fuses data from cameras, lidars, radars, and ultrasonic sensors. The AV's central brain must employ information from that model to track other road users, create a navigation plan that calculates inputs to various electronic control units (ECUs), and then check the vehicle's position and response through GNSS, inertial-measurement units (IMUs), and maps.

The French startup leveraged the relationships its founders built working with automotive customers on DVB-H receivers, designing the ADxxxx to address all perception, sensor-fusion, and object-tracking tasks that Level 4 self-driving cars require. Level 5 is still an unattainable dream that no claims of "full self-driving" can realize. As the evolution of Nvidia's Drive systems demonstrates, automotive customers demand ever increasing AI performance, but they also need precision DSPs that ensure the accuracy of complex sensor fusion. The AD1028 aims to address both.

Despite the hardware challenges, however, the software challenge is even greater. Although its AIware3 DLA is no match for the AD1028, Hungarian automotive specialist AImotive offers its IP along with a more comprehensive software stack (see *MPR 11/5/18,* "AIware3 Adds Rings to Neural Engine"). The company has its own self-driving test vehicles, which run its AIdrive perception and sensor-fusion software. It also supplies the AIsim platform for design and verification of software running on its platform.

In comparison, the Vsora software platform is immature and untested on the road. But the company is only three years old, having introduced its first product with support from less than $2 million in Series A funding. AImotive, in contrast, began five years ago and has received a total of $68 million from three funding rounds. Vsora is off to a good start, rightly focusing its initial product on the needs of automotive customers. The French company's technology looks promising, but it will need additional funding to keep pace with more-established vendors in the highly competitive AV market. ♦

To subscribe to *Microprocessor Report,* access www.linleygroup.com/mpr or phone us at 408-270-3772.