The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

AMBIENT'S ANALOG CUTS AI POWER

GPX-10 Boasts Custom SRAM Cells for Low-Power IoT

By Aakash Jani (November 2, 2020)

Startup Ambient Scientific uses analog technology to reduce power for AI inference and training in IoT devices. Its GPX-10 SoC delivers 512 billion operations per second (GOPS) while consuming only 120mW for active inferencing. In standby or always-on mode, the chip requires as little as 10 microwatts. It includes a wide array of sensor interfaces that feed directly to its proprietary sensor-fusion hub.

Ambient was founded in 2017. The startup is led by GP Singh, who helped design the UltraSparc processor at Sun Microsystems and led the engineering team at Wave Computing. It has withheld its funding amount and the identity of investors, but we estimate it has raised enough to cover the design and production cost of its first device. The GPX-10 is currently sampling and is expected to enter production in 2Q21.

The GPX-10 features a single Cortex-M4 CPU with 10 AI cores. It's manufactured in TSMC's 40nm process and has 512KB of on-die flash memory, in addition to support for external flash through QSPI and SPI protocols. The chip handles both analog and digital sensors through its wide range of interfaces.

The GPX-10 is the first chip to bring retraining to IoT devices. Today's neural networks are trained in the cloud using generic data from many users. This approach takes advantage of the massive compute power in cloud data centers, but it creates a one-size-fits-all AI model. To provide truly personal services, an IoT device must adapt the neural network to the user. For example, a voice-recognition model could be trained to recognize a specific voice. This approach doesn't require training the model from scratch, a computationally intensive task; instead, an existing model can be retrained on the device to more accurately respond to a particular user or use case.

Custom SRAM Computes MACs

Multiply-accumulate (MAC) operations are at the heart of many AI workloads. Analog-AI companies are using basic principles such as Ohm's law to perform these operations

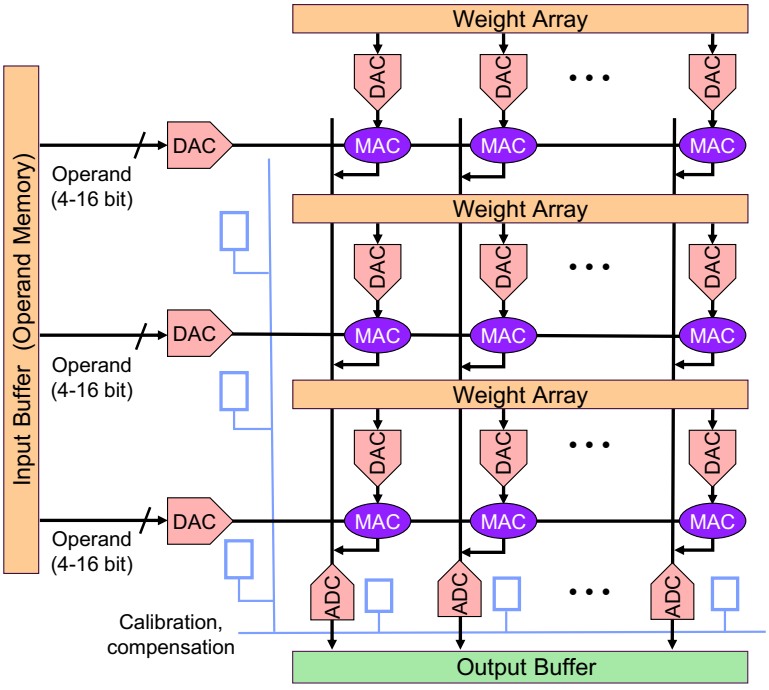

Figure 1. GPX-10 compute engine. Each MAC unit is based on a transistor-only custom SRAM cell. The compute engine employs on-chip calibration and compensation to mitigate noise in analog signals.

in memory, consuming far less power than digital circuitry. IBM, Mythic, and similar vendors have already begun pioneering this new approach, but reliable manufacturing is a high wall to climb.

These vendors believe flash memory cells are best for analog MACs. Embedded flash cells rely on a wide polysilicon layer to be manufactured reliably and prevent shorts. But advanced process nodes require thinner poly layers as to reduce gate actual gate length, making it impossible to support embedded flash at 16nm and below. Ambient realized this dissonance, which led it to use custom SRAM cells for the basis of its MAC units.

The company implements a middle-of-the-road approach by combining analog and digital components to perform MAC operations. Its compute engine executes the equivalent of 512 operations per cycle using 256 analog MAC units. Both the operands and the weights reside in multiple banks of custom SRAM cells. The weights are distributed between 32 memory arrays, as Figure 1 shows. The number of weights depends on the neural-network size, which in turn affects the necessary number of arrays.

The operand memory is multibanked with 64 arrays to support neural-network training as well as inference. Training uses back propagation to calculate the loss-function gradient per layer and iterates the gradient backward to minimize the loss function. Each array stores gradient information on a per-layer basis.

The MAC units employ analog computation, reducing power relative to digital MAC units. Like the SRAMs, these units are based on standard CMOS transistors that omit discrete resistors and capacitors, meaning they can be fabricated in any CMOS process. Each operand must pass through a digital-to-analog converter (DAC) before undergoing analog processing. After the analog multiplier computes a product, the chip sums these values across the wires in accordance with Kirchhoff's Law. It then converts the sums back to digital format through a series of analog-to-digital converters (ADCs).

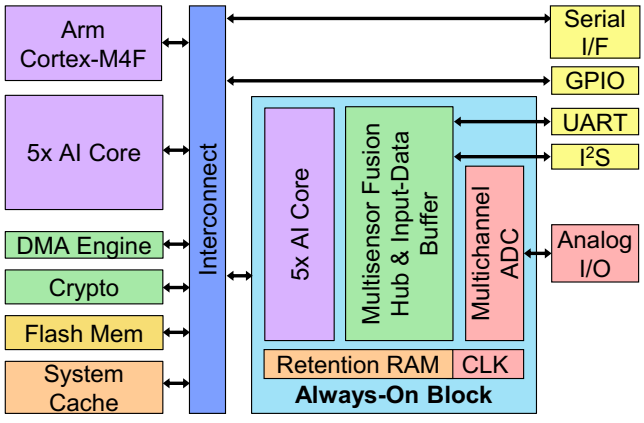

Figure 2. GPX-10 AI core. The design features a RISC engine and an analog compute engine. The activation engine controls the data flow from the analog compute engine.

Analog circuits face different challenges than digital circuits, including matching, biasing, calibration, and compensation. The removal of discrete capacitors and resistors minimizes matching error by reducing design complexity, decreasing process variation and the chance of human error in mask design. The MAC units feature self-compensating and self-biasing circuits to reduce imprecision due to current and temperature variation. The ADCs and DACs are custom designed with matched transistors to reduce voltage and current variation.

In this way, the analog compute engine can process operands with a peak precision of 16 bits. Ambient believes future products will achieve 32 or even 64 bits, but achieving such high precision in analog circuitry is extremely difficult, since even a tiny amount of analog noise can cause errors. Although many digital AI chips offer peak precision of 32 bits, 8 to 16 bits is adequate for most inference and retraining.

Mixed Signals, in a Good Way

Each AI core contains all the components of a self-sufficient deep-learning accelerator (DLA). A custom RISC CPU orchestrates the core's various elements; its simple 32-bit design has three pipeline stages. The AI core relies on two different crossbars to distribute data in and among the cores.

After data leaves the MAC engine, it enters an output buffer. To conserve power, the data sits in this buffer until it receives a signal from downstream circuits. Each AI core features eight ALUs with eight dedicated dual-port register files. Each ALU can perform integer operations at up to 32-bit precision. The core uses the ALUs to execute biasing, normalization, and general arithmetic operations. These units employ custom circuitry to accelerate common AI operations such as ReLU.

Next, data flows to the activation engine, which comprises both digital and analog components. An SRAM array stores a multisegmented lookup table that provides a coarse collection of activation data points for processing by mixed-signal logic. This logic interpolates between the coarse points for a more accurate output. The activation engine asynchronously signals both the compute engine and output buffer. Once this computation is complete, the chip can store the results in memory or recirculate them to the compute engine for another pass.

Always on Guard

The GPX-10 has two logical partitions: the always-on block and the processor subsystem. To save power, the processor subsystem remains powered down most of the time while the always-on (AON) block monitors the sensors. The AON block contains five AI cores, a multisensor-fusion hub, an eight-channel 16-bit ADC, and an input-data buffer. This buffer implements 64KB of custom RAM and stores digital as well as post-ADC analog sensor data. The buffered data then enters the multisensor-fusion hub, which broadcasts

(or selectively communicates) the preprocessed data to the AON AI cores. Only if these cores detect a significant event does the AON block wake the processor subsystem. Essentially, the GPX-10 is an MCU with an added DLA that can either function in concert with the MCU or separately, depending on the workload.

The processor subsystem can run both AI and non-AI workloads using its Cortex-M4F CPU up to 10 AI cores. It employs 512KB of internal flash memory that comprises both NAND and NOR cells. For systems that need more capacity, customers can attach additional flash memory through the QSPI and SPI interfaces. To maximize performance, inferencing code and data resides on the chip. The GPX-10 has 320KB of system-level cache to store mission-critical data. Ambient offloaded encryption operations from the CPU by integrating a dedicated hardware security module (HSM) that implements AES encryption and secures on-chip data through a 128-bit asymmetric key.

The processor features four components that actively execute instructions: the CPU, the security module, the AON partition, and the AI cores. An Advanced High-Performance Bus (AHB) and an Advanced Peripheral Bus (APB) connect the different components in the processor subsystem. As the name suggests, the APB disseminates the peripheral data, which comes from two SPI, one I^2S, and three I^2C interfaces. The two busses are linked by an AHB-APB bridge. Flash memory, the DMA, the HSM, the Cortex M4F, and the AON block are all joined through the performance bus.

Ambient's provides offers a software stack for the GPX-10, including drivers for TensorFlow, Caffe, Pytorch, Theano, and CNTK frameworks. The stack further includes a proprietary compiler, real-time OS (RTOS), and custom libraries. Customers can also upgrade to Ambient DataVault, a subscription stack that provides the software framework to retrain edge devices by storing abstracted data. Retraining models on the edge boosts privacy because it eliminates the risk of information theft between the device and the cloud.

The GPX-10 is uniquely designed to handle retraining. Retraining varies from training because the former uses small batches of new data to generate a modified loss function. Then, the network compares the differential of the new loss function versus the old to incrementally modify the weights. Both retraining and training require FP and INT operations. The GPX-10 parallelizes these operations by offloading the INT operations across the MAC units, while the RISC engines process the FP operations.

Low Power at Every Level

The GPX-10 features many optimizations to reduce power. The CPU and AI cores operate on a wide array of frequencies, ranging from 0.01MHz to 150MHz. A custom low-power oscillator (LPO) provides an energy-efficient 50MHz clock signal in a 20-microwatt power envelope. Traditional high-frequency oscillators, by contrast, employ a combination of digital and analog phase-locked loops (PLLs). When they change frequencies, the PLLs must relock. The GPX-10's custom LPO can change frequencies without halting downstream logic and without glitches.

The SoC can operate in a variety of power states: always-on voice detection, always-on fault detection, and active mode. With all systems running at maximum specifications, it consumes 120mW. For always-on voice commands and fault detection, the power drops to just 80 microwatts, allowing the GPX-10 to operate from a small battery. We estimate the AON block provides 0.3 GOPS at this low power. After activating the processor subsystem, the device typically requires 0.15–1.50mW to process a fault or a voice input.

Memory arrays in the processor subsystem and AON block have custom SRAM cells to reduce power. Each array is equipped with a custom sense amplifier, which prevents the discharging of multiple bit lines during read operations.

For AI workloads, MAC-related memory operations heavily contribute to overall power consumption. Ambient optimized these operations through its memory architecture in the analog compute engine. The multibanked memory arrays continuously feed weights and operands to the MAC units. In fact, the compute engine can load an entire row and multiple columns of a matrix in one cycle. This method not only reduces power from memory read operations, but it also reduces latency.

Price and Availability

The GPX-10 is now sampling, and we expect it to reach production in 2Q21. For more product information, visit *www.ambientscientific.net/products.html*. For pricing information, you may contact the company at *www.ambientscientific.net/contact-us.php*.

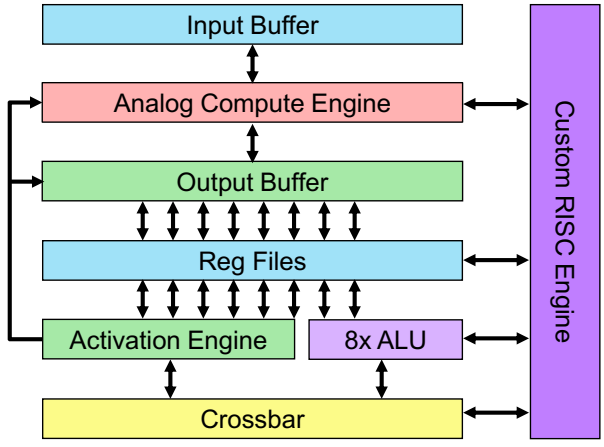

Figure 3. GPX-10 block diagram. The always-on block sends an activation signal to the rest of the SoC when it detects an inference event.

Highly Efficient

Ultra-low-power processors are well suited to the MCU market. Although CPUs can inference small neural networks, chip designers offload large networks to dedicated DLAs. GreenWaves, a French IoT startup, designed its DLA around the open-source RI5CY architecture (see *MPR 1/13/20,* "GreenWaves GAP9 Goes Faster"). Syntiant, an AI-hardware startup, employed a custom approach for the NDP101's DLA (see *MPR 3/18/19,* "Syntiant Knows All the Best Words").

Syntiant designed the NDP101 purely for speech recognition, whereas GAP9 serves in both speech recognition and image processing. Ambient's GPX-10 bests the competition by enabling these tasks as well as sensor fusion and industrial applications. That chip thus offers a wider range of I/O interfaces to provide a kitchen-sink approach to sensor support.

On-chip RAM stores neural networks and time-sensitive data for these three low-power edge chips. The GPX-10 is a middle-of-the-road offering that has 38% less on-die memory than GAP9 but 3x more than the NDP101. For larger networks and files, its embedded flash memory can store weights. GAP9 and NDP101 customers must provide external flash for their devices.

In its lowest-power state, we estimate the GPX-10 performs 0.3 GOPS at 80 microwatts. The NDP101 generates about the same throughput but uses 200 microwatts. Thus, Ambient's performance per watt exceeds even Syntiant's. That efficiency is best explained by Ambient's choice of analog circuitry. GAP9, by contrast, has multiple low-power CPUs with a small MAC array to accelerate deep-learning operations. In summary, at maximum speed the GPX-10 outperforms the GAP9, while at minimum speed it uses less power than the NDP101, providing both high performance and low power in a single chip.

	Ambient GPX-10	GreenWaves GAP9	Syntiant NDP101
Main CPU	Cortex-M4F	RI5CY	Cortex-M0
CPU Speed (peak)	100MHz	250MHz	20MHz
DLA Type	Custom	8x RI5CY	Custom
DLA Speed (max)	150MHz	400MHz	20MHz
Total AI Perf (INT8)	512 GOPS	49 GOPS	0.24 GOPS
On-Chip RAM	320KB	512KB	112KB
Flash Memory	0.5MB internal, external	External	External
Other I/O	GPIO, serial, analog	Serial	2x PDM, PCM over SPI
IC Process	40nm CMOS	22nm FD-SOI	40nm ULP
Active Power (typ)	120mW	50mW	0.2mW
Perf Efficiency	4,250 GOPS/W	980 GOPS/W	1,200 GOPS/W
Production	2Q21 (est)	4Q20 (est)	2Q19

Table 1. Comparison of AI MCUs. The GPX-10 completes over 10x operations per second as GAP9 and does so 4x more efficiently. (Source: vendors)

Getting Production Ready

The GPX-10's analog MAC units provide exceptional power efficiency. Since these units are essentially stripped-down SRAM cells, they mitigate defects from process variation. Mythic was among the first companies to announce a commercial analog-MAC product (see *MPR 8/27/18,* "Mythic Multiplies in a Flash"), but it has struggled ever since to achieve volume production. Ambient hopes to avoid Mythic's arduous journey by relying on its custom MAC units, which implement a different type of analog multiplier.

Ambient designed the GPX-10 as a standalone SoC in 40nm while employing techniques that are necessary for more-advanced process nodes. Each AI core acts as a standalone DLA, which simplifies core-count scaling as long as the data fabric can handle the increased throughput. This forward thinking extends to the startup's SRAM designs and multibank approach. Each dimension can scale to boost throughput per core. Every design employs simplified geometries, so as the company pursues advanced nodes, porting becomes easier. Although 40nm provides an inexpensive starting point, we expect Ambient to move to more advanced nodes once it has enough funding; these nodes should further improve performance per watt.

The GPX-10 features power-saving techniques in every facet of its design. Although these techniques mitigate both leakage and active power for the digital circuitry, the high performance per watt is a direct result of the analog compute unit. In our comparison, the GPX-10 was 3–4x more power efficient than its digital competitors. Although some high-end DLAs achieve up to 8 TOPS per watt—twice the GPX-10's throughput—they require leading-edge 7nm technology.

The combination of its DLA and wide array of sensor interfaces makes the GPX-10 a strong competitor in the low-power edge-AI market. The true test now lies ahead. Many analog-compute companies deliver impressive test chips but are stymied during the production phase. If Ambient clears the production hurdle, the GPX-10 may become a formidable competitor in a burgeoning field. ♦