The Linley Group

MICROPROCESSOR report

Insightful Analysis of Processor Technology

MYTHIC MULTIPLIES IN A FLASH

Analog In-Memory Computing Eliminates DRAM Read/Write Cycles

By Mike Demler (August 27, 2018)

Machine learning is a hot field that attracts high-tech investors, causing the number of startups to explode. As these young companies strive to compete against much larger established players, some are revisiting technologies the veterans may have dismissed, such as analog computing. Mythic is riding this wave, using embedded flash-memory technology to store neural-network weights as analog parameters—an approach that eliminates the power consumed in moving data between the processor and DRAM.

The company traces its origin to 2012 at the University of Michigan, where CTO Dave Fick completed his doctoral degree and CEO Mike Henry spent two and a half years as a visiting scholar. The founders worked on a DoD-funded project to develop machine-learning-powered surveillance drones, eventually leading to the creation of their startup. Over the past two years, Mythic has attracted more than $55 million from a group that includes Lockheed Martin and SoftBank along with venture-capital (VC) firms and individual investors. Its goal is to apply analog in-memory computing to neural networks in low-power clients, beginning with smart cameras for surveillance and drones. We estimate the company has about 50 employees, divided between its Austin, Texas and Silicon Valley offices.

By performing multiplication operations directly in flash memory, thus eliminating DRAM read/write cycles, Mythic estimates it can match the accuracy of inference engines running on GPUs but at a fraction of the power. Many neural networks store 10 million to 100 million weights or more in DRAM, since there are far too many to store in an inexpensive client processor. For example, the popular ResNet-50 uses more than 25 million weights. But Mythic's prototype runs that network using just 5MB of on-chip SRAM, consuming less than 5W.

The startup plans by the end of this year to tape out its first product, which can store up to 50 million weights. At that time, it also aims to release the alpha version of its software tools and a performance profiler. The company expects to begin volume shipments in 4Q19.

Solving a Weighty Problem

Mythic uses Fujitsu's 40nm embedded-flash cell, which it plans to integrate in an array along with digital-to-analog converters (DACs) and analog-to-digital converters (ADCs), as Figure 1 shows. By storing a range of analog voltages in the bit cells, the technique uses the cell's voltage-variable conductance to represent weights with 8-bit resolution. This

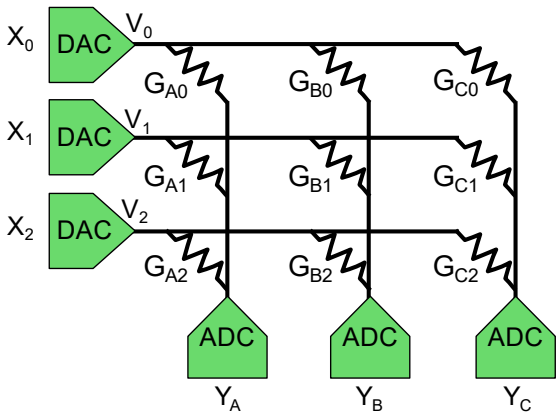

Figure 1. Mythic flash-based neural-network accelerator. The 8-bit DACs program voltage-variable conductances in the memory cells to represent the neural-network weights. When inferencing, multiplication follows Ohm's law, producing currents that are the product of the input signal voltage and the stored weight.

technique is similar to what IBM uses in its analog neural networks based on phase-change memory (PCM), which also employs variable-conductance cells to store synaptic weights (see *MPR 8/13/18,* "IBM Trains in Analog to Save AI Power"). That company withheld details, but its architecture has the advantage of a proprietary sensing technique that works without DACs or ADCs.

In digital applications, each flash cell typically stores only one of two voltage levels such that its storage transistor is either completely on or off. When a bit cell is conducting, a sense amp detects the current as a logic one. In Mythic's architecture, the 8-bit DACs enable storage of 256 different voltages, yielding 256 different conductance (G) values between off and fully on. Connecting a group of cells to a voltage signal is equivalent to performing matrix multiplication, but using Ohm's Law ($I = V \times G$) instead of digital logic. Rather than using sense amps to detect on or off current with 1-bit resolution, the ADCs measure it with 8-bit resolution. Although this technique lacks the dynamic range needed for training, it's suitable for inference engines.

For its first product, Mythic omitted the DAC circuits; users must therefore program the device with an external voltage source. The flash cells operate in pairs, and subtracting the currents in a pair can represent weights ranging from –127 to +127. The company withheld details of the ADC, but it employs a fully differential architecture that performs the necessary subtraction.

The programming uses closed-loop calibration in which the ADCs measure each cell's equivalent weight. By comparing the measurement to the desired value, the calibration program can adjust the weight to within 0.4% ($1/2^8$) accuracy. Because this technique requires multiple iterations, the programming time is longer than for standard flash. Mythic's design target is a one-minute total programming time for the complete neural-network array.

Although the DAC/ADC processing and flash storage cells have nonlinearities and other distortions, the company's software compensates for them by adjusting the weights during offline conversion of pretrained networks. The adjustment is one time only, however, so it doesn't account for drift that results from temperature and voltage variations. Because the flash memory isn't for long-term storage, it has much shorter retention requirements than it would in digital applications. Nevertheless, to maintain weight accuracy, the system can execute periodic refresh cycles. The company expects a daily refresh to be adequate.

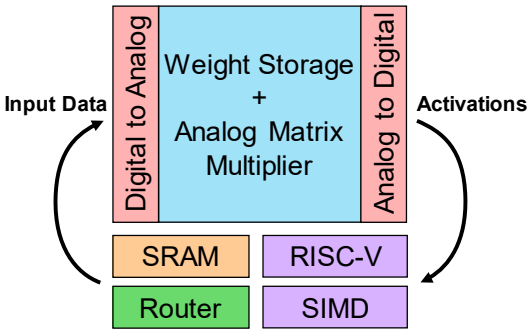

Figure 1. Mythic flash-based neural-network accelerator. The 8-bit DACs program voltage-variable conductances in the memory cells to represent the neural-network weights. When inferencing, multiplication follows Ohm's law, producing currents that are the product of the input signal voltage and the stored weight.

Digital Logic Completes the Picture

Although MAC operations constitute the vast majority of inference-engine calculations in convolutional neural networks (CNNs), other functions such as pooling layers and activations still require digital logic. Neural networks also need a CPU controller to oversee operations. To handle the non-MAC layers, Mythic complements the analog flash array with a RISC-V microcontroller core and a SIMD engine, as Figure 2 shows. It selected RISC-V because it can extend the ISA with special-purpose neural-network instructions and because licensing is royalty free (see *MPR 3/28/16,* "RISC-V Offers Simple, Modular ISA"). We expect the ISA extensions include the SIMD operations, which support matrix multiplication, pooling, and other neural-network instructions.

Mythic integrates the analog compute engine, CPU, and SIMD engine with a network-on-a-chip (NoC) and SRAM. The blocks connect in a tile structure that can be instantiated multiple times on a chip. The NoC routes signals between tiles. The SRAM stores input data as well as the results of each operation layer.

The company calls its SoC design an intelligent processing unit (IPU); it integrates multiple tiles that can handle different tasks in parallel. Users will employ the IPU as a coprocessor, connecting it to a host processor through the chip's four-lane PCIe 2.1 interface. The host programs the flash memory during the initial boot. During inference operations, it uses the PCIe interface to send image data to the IPU. The IPU runs the complete neural network without requiring any subsequent host processing or DRAM access, finally signaling the host when it has completed classification.

The initial product can store 50 million weights, which is sufficient to run the Inception v4 or ResNet-50 object classifiers entirely on chip without accessing external DRAM. Larger networks such as VGG-16, however, exceed that capacity and would require multiple IPUs. The PCIe interface enables connection of multiple IPUs for the larger networks. Alternatively, a single chip can hold multiple small neural networks, each in a separate group of tiles.

Keeping Its Focus on Cameras

Following the legacy of its earlier drone work, Mythic is initially focusing on smart cameras. For example, one group of tiles could handle image enhancement while a second group handles scene segmentation and a third runs an

object-tracking algorithm. The company plans to offer first samples on single- and quad-IPU PCIe development boards. For volume production, it intends to add a 16-IPU board along with BGA packages. Mythic has working silicon for the flash-based analog compute engine, but for its initial test results, it emulated the digital portions of the design.

On the basis of those preliminary performance tests, the company estimates one 40nm IPU can run ResNet-50 on 224x224 images at 900 fps. By comparison, it delivers more than 3x the performance of Nvidia's 16nm NVDLA IP, which integrates 2,048 MACs (see *MPR 3/26/18,* "Nvidia Shares Its Deep Learning"). The IPU's estimated ResNet power consumption is 2W, which yields power efficiency similar to the NVDLA's, but Nvidia's specifications exclude the on-chip SRAM as well as the off-chip DRAM. Its PCIe port and SRAM each consume 20% of the total power, with the control logic consuming the rest.

The Mythic design is a good fit for its target markets, including drones and surveillance cameras. Because such devices require neither frequent reprogramming nor the ability to seamlessly switch applications, the one-minute programming time isn't an issue. But the IPU must compete for those applications against other low-power coprocessors, such as the Gyrfalcon Lightspeeur 2801 (see *MPR 2/19/18,* "Gyrfalcon Shrinks AI Accelerator").

The IPU can deliver higher frame rates than Lightspeeur, but at half the power efficiency—although the latter product requires that non-MAC layers run on a host CPU. The IPU also requires connection to a host, but it offloads all neural-network operations. Mythic estimates its design can deliver two trillion MAC operations per watt (2TMAC/W), whereas Gyrfalcon designed its 28nm chip to deliver 4.7TMAC/W. The 900fps ResNet performance is sufficient to handle a 1080p stream at 30fps, but lower frame rates will likely reduce the IPU's power consumption. To provide a net system power savings, however, the host must be able to enter a lower-power state while the IPU is running.

Analog Variability Is No Myth

Other companies are also working on analog in-memory computing, including flash-based designs, so Mythic is likely to have even more competition by the time it begins production. Syntiant is another startup using a flash-based technique, but it has yet to disclose details. Although IBM's analog neural network is still under research, the company estimates 28TOPS/W for its PCM-based approach, and the design offers the advantage of on-device training.

Price and Availability

Mythic expects to sample its IPU in 2H19 and to begin volume production in 4Q19. It withheld prices. For more information, access *www.mythic-ai.com*.

Mythic has fabricated prototypes of its analog-compute engine that omit the DACs, but to show the design is a viable competitor to the plethora of digital alternatives, it must first demonstrate a complete working solution. It must overcome technical hurdles as well. The first is analog variability. Unlike digital logic, the precision of ADCs, DACs, and flash cells varies with process, voltage, and temperature, and it's also subject to intradie variations.

Converting network weights to 256 analog values is the theoretical maximum for an 8-bit design, but the real resolution will be less. Classification accuracy will degrade owing to analog variability, hindering the IPU from serving in automotive and other safety-critical systems. Mythic says its analog design and software compensates for these problems, but recalibration consumes power and time and is also subject to drift. Digital inference engines avoid that dilemma.

The company is relying on Fujitsu's 40nm embedded-flash technology, which UMC has acquired. Although TSMC and other foundries are developing 28nm versions, embedded processes lag several nodes behind leading-edge digital processes. Designers of digital inference engines can take advantage of new manufacturing technologies to increase performance and reduce power, but analog designs don't scale as well.

Nevertheless, if Mythic can back up its simulations with working silicon, we expect it will succeed in applications that can tolerate the long boot time. Few consumers will wait a full minute for their smart cameras to become usable, but industrial inspection cameras, military drones, and other such devices need not respond the second they're turned on. The company is well funded, so it has sufficient room to refine its solution. The IPU is unique in its ability to run a complete inference engine for high-performance image recognition using only on-chip flash memory, and it promises to deliver real cost and power savings. ♦