MYTHIC: HOW AN ANALOG PROCESSOR COULD REVOLUTIONIZE EDGE AI

Alberto Romero, LLM Analyst
Karl Freund, Founder and Principal Analyst
Cambrian-AI Research

Mythic is an analog AI processor company conceived to overcome the growing limitations of digital processors. Founded by Mike Henry and Dave Fick, and based in Texas, Austin, and Redwood City, California, Mythic aims to solve the technical and physical bottlenecks that limit current processors through the use of analog compute in a world dominated by digital technology. Mythic wants to prove that, contrary to common belief, analog isn't a relic of the past, but a promise for the future.
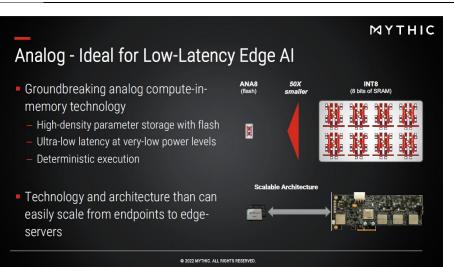
Two main problems inhibit the pace of development of digital hardware: The end of Moore's Law and the Von Neumann architecture. For 60 years we've enjoyed ever-increasing powerful hardware as predicted by Gordon Moore in 1965, but as we approach the minimum theoretical size of transistors, his well-harnessed law seems to be coming to an end. Another well-known issue is the need in the Von Neumann architecture to move data from memory to the processor and back to make the computations. This approach is increasingly being replaced by compute-in-memory (CIM) or compute-near-memory approaches that significantly reduce memory bandwidth and latency while increasing performance.

THE COMEBACK OF ANALOG COMPUTE?

Mythic claims it has built a unique, paradigm-shifting solution that promises to tackle digital's limitations while providing improved specifications compared to the best-in-class digital solutions: an analog compute engine (ACE). Historically, analog computers were replaced by digital due to the latter's reduced cost and size and their general-purpose nature. However, the current landscape of AI is dominated by deep neural networks (DNNs) which don't require extreme precision and, more importantly, the majority of the computing bulk goes into a single operation: matrix multiplication. The perfect opportunity for analog compute.
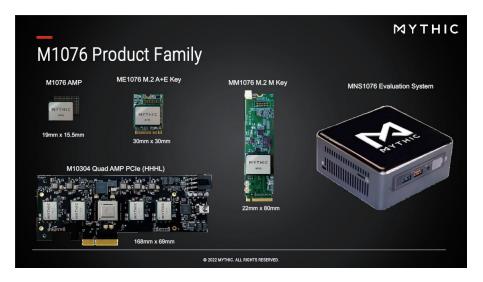
On top of it, Mythic is exploiting the advantages of CIM and dataflow architecture to obtain impressive early results. They've taken CIM to the extreme by computing directly in the flash memory cells. Their analog matrix processors take the inputs as voltage, the weights are stored as resistance, and the output is the resulting current. In addition, the dataflow design keeps these processes running in parallel, which allows for extremely fast and efficient calculations while maintaining high performance. A clever combination of analog computing, CIM, and dataflow architecture defines the Mythic ACE, the company's main differentiating technology.

Analog - Ideal for Low-Latency Edge AI

MYTHIC'S ACE MEETS THE REQUISITES OF EDGE AI INFERENCE

Mythic's tech promises high performance at very low power, ultra-low latency, low cost, and small form factor. The basic element is their Analog Matrix Processor (AMP) which features an array of tiles, each containing the ACE complemented by digital elements: SRAM, a vector SIMD unit, a NoC router, and a 32-bit RISC-V nano processor. The innovative design of the ACE eliminates the need for DDR DRAM, reducing latency, cost, and power consumption. AMP chips can be scaled, providing support for large or multiple models. Their first product, the single-chip M1076 AMP (76 AMP tiles) can handle many endpoint applications and can be scaled up to 4-AMPs or even 16-AMPs on a single PCI express card, adequate for edge server-level high-performance usage.

M1076 Product Family

The hardware is complemented with a software stack that provides a seamless pipeline going from the graph (ONNX and PyTorch) to an AMP-ready package through a process of optimization (including a quantization to analog INT8) and compilation. Mythic's platform also supports a library of ready-to-go DNNs, including object detection/classification (YOLO, ResNet, etc.) and pose estimation models (OpenPose).

The company's full-stack solution leverages the potential of analog processors while maintaining relevant features of the digital world. It makes the M1076 AMP a great option to handle AI workloads for inference at the edge faster and more efficiently—the company claims it provides the "best-in-class TOPS/W"—than its fully-digital counterparts. That, and the company's broad offering of products and AI models, make it well-posited to target fast-growing edge AI-focused markets like video surveillance, smart home devices, AR/VR, drones, and robotics.

So far, it seems Mythic has transformed an innovative idea into promising tech to compete for edge inference AI. Now, let's see the numbers. The company claims the M1076 AMP performs at up to 25 TOPS running at around 3W. Compared to similar digital hardware, that's a reduction in power consumption of up to 10x. And it can store up to 80M weights on-chip. The MP10304 Quad-AMP PCIe card can deliver up to 100 TOPS at 25W and store 320M weights. When we compare these claims to those of many others, we can't help but be impressed.

CONCLUSIONS

The success of analog AI will depend on achieving high density, high throughput, low latency, and high energy efficiency, while simultaneously delivering accurate predictions. Compared to pure digital implementations analog circuits are inherently noisy, but despite this challenge, the benefits of analog compute become apparent as processors like the M1076 are able to run larger DNN models that feature higher accuracy, higher resolution or lower latency.

As Mythic continues to refine its hardware and software, we will look forward to seeing benchmarks that can demonstrate the platform's capabilities and power efficiency. But we have seen enough already to be excited by the potential of this unique approach.

IMPORTANT INFORMATION ABOUT THIS PAPER

Authors: Alberto Romero and Karl Freund, Cambrian-AI Research, LLC

INQUIRIES:

Contact us if you would like to discuss this report, and Cambrian-AI Research will respond promptly.

CITATIONS

This paper can be cited by accredited press and analysts but must be cited in context, displaying the author's name, author's title, and "Cambrian-AI Research." Non-press and non-analysts must receive prior written permission from Cambrian-AI Research for any citations.

LICENSING

This document, including any supporting materials, is owned by Cambrian-AI Research. This publication may not be reproduced, distributed, or shared in any form without Cambrian-AI Research's prior written permission.

DISCLOSURES

This document was developed with QTI funding and support. Although the document may utilize publicly available material from various vendors, including QTI, it does not necessarily reflect the positions of such vendors on the issues addressed in this document.

DISCLAIMER