GrAI Matter Labs: At-Memory Brain-inspired Compute for AI at the Edge

GrAI Matter Labs (GML) is an AI hardware start-up targeting Edge AI with near-real-time computation. Seed-funded by DARPA in 2016 and a group of experts in silicon design and neuromorphic computing, GML believes they're revolutionizing deep learning at the endpoint device, focussing on audio and video processing with very low latencies. By processing data closest to its source, AI algorithms can provide almost instant insights and transformations without incurring higher latencies and costs typical of cloud servers. GML's "Life-ready" AI provides solutions that heretofore were simply impossible at such low foot-print and power. After experiencing a private demo, we were amazed by the quality and instant latencies they were able to produce.

Transforming content at the endpoint device with high fidelity

IoT devices are proliferating in smart security cameras in the streets, robotic arms in factories, voice assistants in our homes, and smartphones in our pockets. All these devices have sensors that capture data. Most companies applying AI at the edge of the network are focusing on understanding or categorizing that data to enable predictions. GML is literally transforming the audio-visual user experience on the fly. To achieve this, they combine four pillars of technology: high-precision (16-bit Floating-Point) processing to deliver high-quality content, dynamic data flow to exploit data-dependent sparsity, neuromorphic design to improve efficiency, and at-memory computing to reduce power consumption and latency. The bottom line: 1/10th the response time at 1/10th the power.

GML's value proposition is therefore building on these pillars that, combined, create a uniquely differentiated solution: Endpoint computing with AI at low latency and high-power efficiency to transform raw data into high-fidelity consumable content in real-time, allowing for instant applicability in many daily situations.

Sparsity is the key to transforming content at low latency and low power

Power restrictions at the edge of the network force endpoint AI devices to keep consumption low. GML's innovative solution produces high fidelity content by exploiting sparsity—the fact that audio and video content doesn't change everywhere, nor all at once—at high precision.

A prototypical example to illustrate the upside of this approach is a smart security camera. The recorded background remains largely constant across the day, so it gives no new information. By processing and analyzing only people, vehicles, and other moving objects, the savings in power consumption and reductions in latency can range up to 95%.

A silicon implementation of GML's solution: GrAI VIP

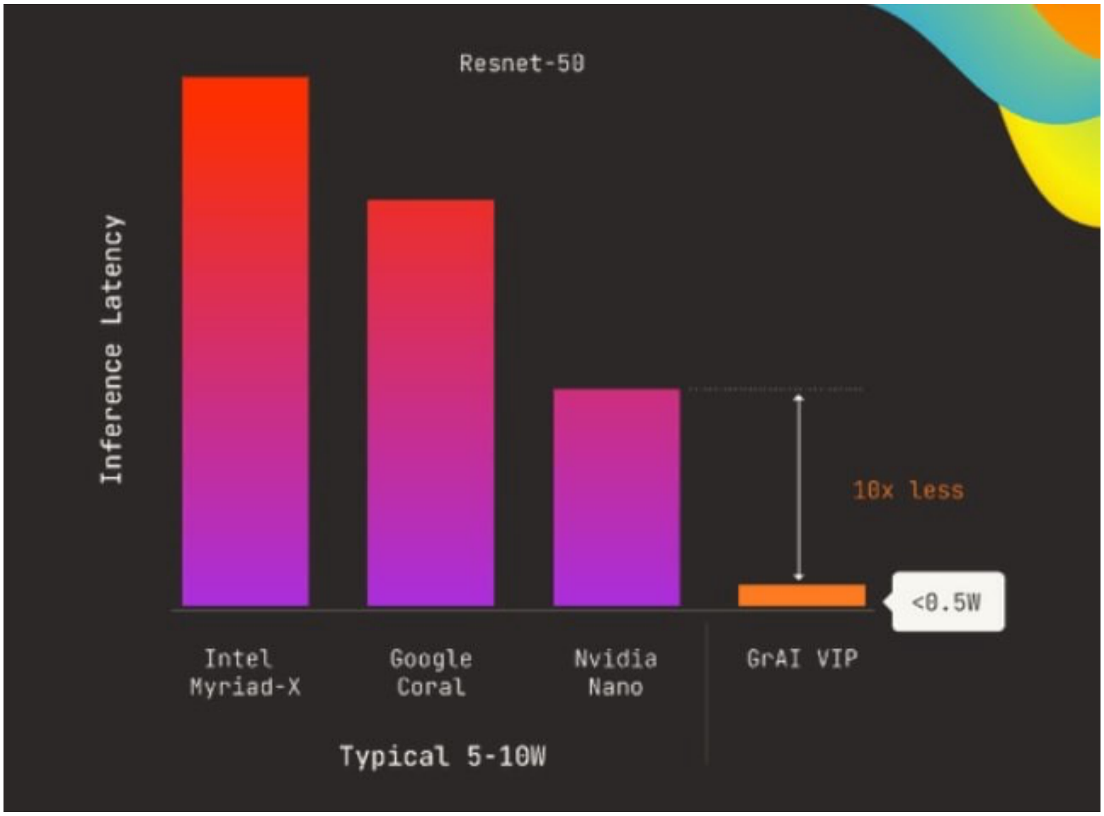
GML's forthcoming hardware product, GrAI VIP (available in engineering samples) is an SoC (System on Chip) that integrates a patented neuron engine, GrAICore, two ARM processors, all the memory for the use-case, audio and video sensor interfaces with the required characteristics for low-power, ultra-low latency, and high-precision inference processing at the endpoint.

GrAICore employs brain-inspired NeuronFlow technology. Apart from sparse processing, NeuronFlow is based on the dataflow architecture paradigm, which allows for efficient fine-grained parallelization. Together with at-memory compute, which reduces performance bottlenecks caused by moving data between memory and processor, these features accelerate the computations by several orders of magnitude.

GrAI VIP's full-stack is completed with the GrAIFlow SDK, compatible with the usual ML frameworks, TensorFlow and PyTorch, to implement custom models. It also provides a library of ready-to-deploy models. Both custom and pre-trained models can be optimized and compiled with the ML toolkit to be deployed for inference at the edge device with the last component, the GrAIFlow Run-Time Ready.

Conclusions

GML is targeting the $1 billion+ fast-growing market (20%+ per year) of endpoint AI with a unique approach backed by innovative technology. They further beat endpoint competitors by focusing on high-fidelity 16-bit floating-point real-time "content transformation" instead of just "understanding" (categorizing) using 8-bit or lower computations.



According to the company, the four pillars combine to outperform NVIDIA's leading-edge platform, the Jetson Nano, by 10X, at > 10X lower power for Resnet50. However, the Nano is a much more comprehensive edge platform, while the GML platform is focussed on doing a few tasks very well. As a result, GML stands to revolutionize consumer and enterprise audio-visual experiences with everyday devices at high fidelity while meeting the strict power and cost requirements of endpoint content manipulation.

We believe GML's pervasive applicability and unique differentiation could help the company grow rapidly in a segment where they have very little competition if any.

IMPORTANT INFORMATION ABOUT THIS PAPER

AUTHORS: Alberto Romero and Karl Freund, Cambrian-AI Research

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 mentioned in the 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

DISCLOSURES

This document was developed with Qualcomm Technologies, Inc. (QTI) funding and support. Although the paper 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

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions, and typographical errors. Cambrian-AI Research disclaims all warranties as to the accuracy, completeness, or adequacy of such information and shall have no liability for errors, omissions, or inadequacies in such information. This document consists of the opinions of Cambrian-AI Research and should not be construed as statements of fact. The views expressed herein are subject to change without notice.

Cambrian-AI Research provides forecasts and forward-looking statements as directional indicators and not as precise predictions of future events. While our forecasts and forward-looking statements represent our current judgment on what the future holds, they are subject to risks and uncertainties that could cause actual results to differ materially. You are cautioned not to place undue reliance on these forecasts and forward-looking statements, which reflect our opinions only as of the date of publication for this document. Please keep in mind that we are not obligating ourselves to revise or publicly release the results of any revision to these forecasts and forward-looking statements in light of new information or future events.