Cambrian AI Research
explore the explosion

GrAI Matter Labs: Brain-Inspired AI For The Edge

Karl Freund

This article was written by Alberto Romero, Cambrian-AI Analyst, and Karl Freund, Cambrian-AI Founder.

We recently tweeted about the startup GrAI Matter Labs (GML) and received a lot of questions about the company's products and strategy. As one of the first startups to launch a neuromorphic AI platform for edge AI, the company is deserving of a little more attention, so let's take a closer look.

Background

GML is an AI hardware start-up targeting Edge AI with near-real-time computation. Founded in 2016 by a group of experts in silicon design and neuromorphic computing, GML believes they are revolutionizing inference at the endpoint device, focussing initially on audio and video processing with very low latencies. By processing data closest to its source, AI algorithms can provide almost instant insight and transformation without incurring the higher latencies and costs typical of cloud servers. GML's "Life-ready" AI provides solutions that here-to-for were simply impossible at such low cost and power. After a demo, we were amazed by the quality and instant latencies they were able to produce.

GML is currently focused on industrial robotics and drones for near-sensor understanding, which is a ~50M device market in 2025. Now the company's plans will expanding its reach to include high-fidelity data transformation in

mobile and consumer devices, a market which the company estimates is 20 times larger with over 1 billion devices in 2025.

GML is turning its attention to data transformation in consumer devices.

High-fidelity 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 in-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 concept, GML VIP (not yet available for production) is an SoC (System on Chip) that integrates a neuron engine, GrAICore, with the required characteristics for low-power, ultra-low latency, and high-precision inference processing at the endpoint.

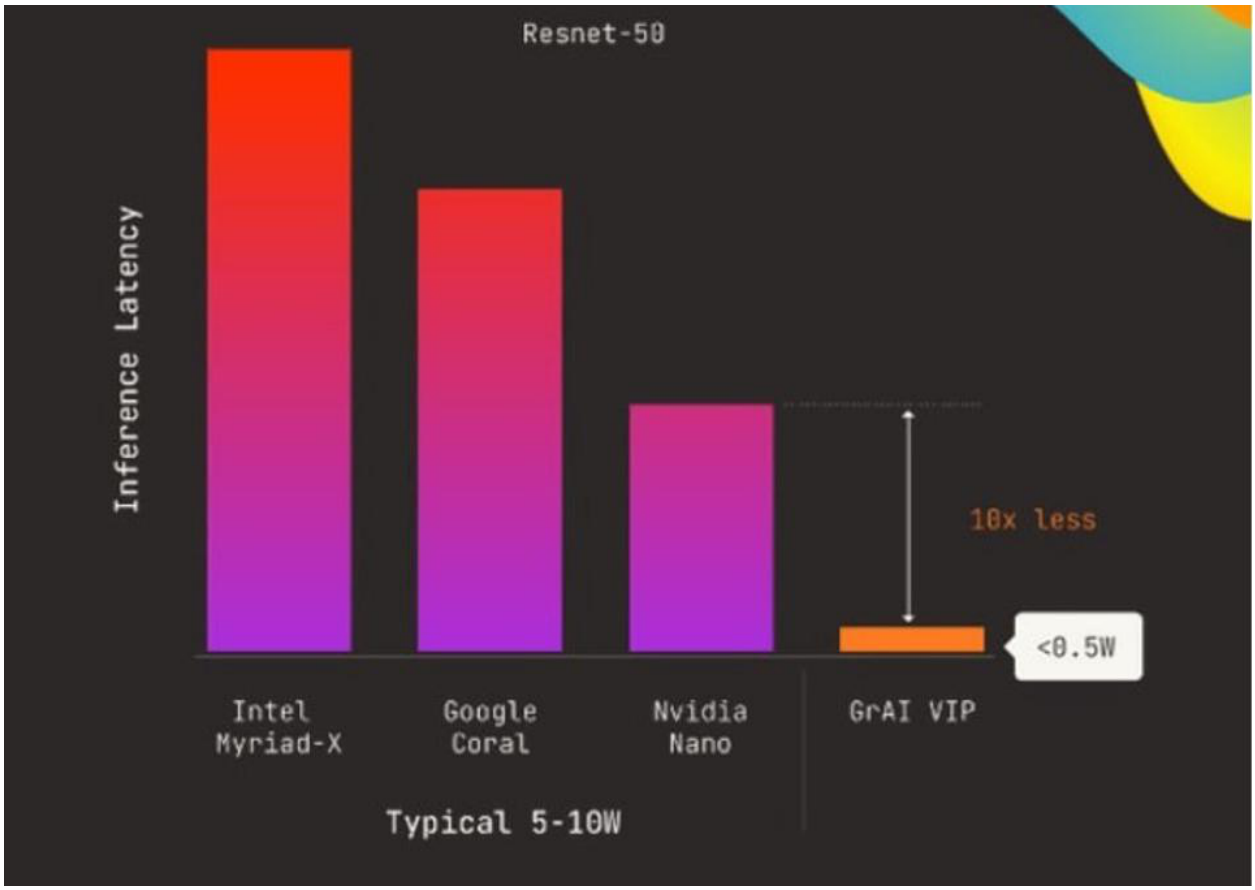
GrAICore employs brain-inspired NeuronFlow technology. Apart from sparse processing, NeuronFlow is based on the dataflow architecture paradigm, which allows for efficient fine-grained parallelization. Together with in-memory compute, which reduces performance bottlenecks caused by moving data between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

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 best endpoint competitors by focusing on high-fidelity 16-bit floating-point real-time "content transformation" instead of just "understanding" (categorizing) which typically uses 8-bit computation.



GML has better performance on Resnet50 compared to other edge devices from Intel, Google, and ... [+]

GRAI MATTER LABS

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, we note that the Jetson Nano is a comprehensive edge platform, while the GML platform is focused on doing a few tasks very well.

GML potentially 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 unique differentiation could help the company grow rapidly in a segment where they can enjoy a first-mover advantage.

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.