D-Matrix: Digital In-memory Compute For Data Center Inference On Transformer-based AI Models

D-Matrix was founded in 2019 by two veterans in the field of AI hardware, Sid Sheth and Sudeep Bhoja, who previously worked together at Inphi (Marvell) and Broadcom. The company was born at a singular moment for the field of AI, just two years after the popular transformer architecture was invented by Google Brain scientists. By 2019, the world was starting to realize the massive significance of transformer-based models and D-Matrix saw an opportunity to define its AI hardware specifically to excel in using these Large Language Models.

Transformers are eating the world

GPT-3, MT-NLG, Gopher, DALL·E, PaLM, and virtually every other large language model is based on the now ubiquitous transformer architecture. Tech companies keep announcing potentially amazing models that remain inaccessible to the world due to one insurmountable obstacle: deploying these models into production for inference at the data center is virtually unfeasible with current AI hardware. That's what D-Matrix is aiming to solve and, as a company developing in parallel to the already world-changing wave of transformers and LLMs, they're well-posited to bring a clean-slate approach to this problem.

Focusing on large multimodal models (those that use different types of data) is what differentiates the company from its competitors. Transformer-based models are usually trained on high-performance GPUs (where Nvidia enjoys a multi-year edge), but performing inferences is a power efficiency story, not just performance at any cost. D-Matrix has found an innovative solution with which they claim to achieve 10–30x the efficiency of the current hardware. Once tech companies begin to embed transformer-based NLP models in all kinds of applications and spread them across industries, this type of ultra-efficient hardware will be appealing to handle the inference workloads.

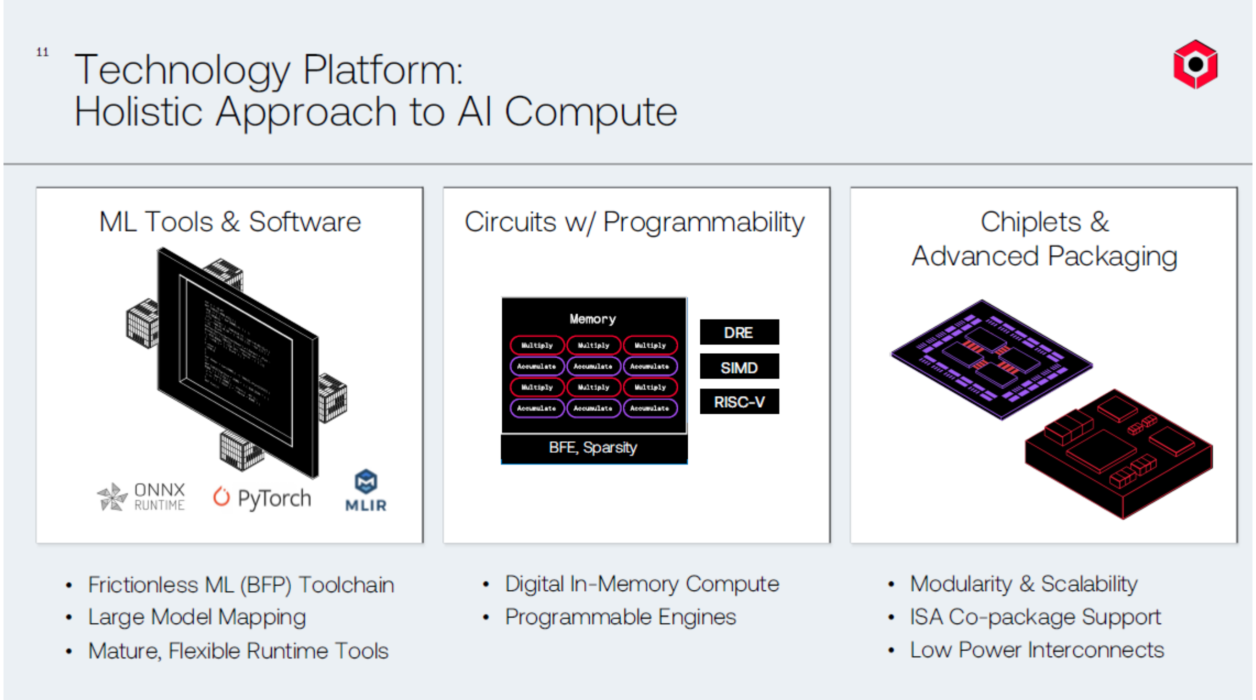
The key to the next generation of AI hardware: In-memory compute

D-Matrix's solution is currently a proof-of-concept chiplet-based architecture called Nighthawk. Together with Jayhawk, its soon-to-be second chiplet that will also implement die-to-die interfaces, they form the basis for Corsair, D-Matrix's hardware product planned to be released in the second half of 2023. Nighthawk comprises an AI engine with four neural cores and a RISC-V CPU. Each neural core is composed of two octal compute cores (OC), each of which has eight digital in-memory compute cores where weights are stored, and matrix multiplication is performed.

Nighthawk emerges from the novel combination of three technological pillars. First is digital in-memory compute (Digital IMC). The efficiency barrier that existing hardware suffers is due to the costs and performance limits caused by moving data around to do the computations. D-Matrix has mixed the accuracy and predictability of digital hardware with super-efficient IMC to create what D-Matrix believes is the first DIMC architecture for inference at the data center. Nighthawk's projected performance seems to back D-Matrix's idea of bringing both data and compute into the SRAM, which is the current best memory type that serves the IMC solution. D-Matrix claims its hardware is 10x more efficient than an NVIDIA A100 for inference workloads.

The second pillar is the use of lego-like modular chiplet architecture. Chiplets can be interconnected with Jayhawk—the complementary IP piece for Nighthawk—for scaling up and scaling out the hardware. Up to 8 chiplets can be arranged in a single card while keeping the efficiency capabilities intact. These chiplets can be "plugged in" with existing hardware and be used specifically to handle transformer-related workloads. In the future, D-Matrix believes its hardware could store models as large as the 175-billion-parameter GPT-3 in a single card.

Finally, D-Matrix applies transformer-specific numerics, sparsity, and other ML tools that further enhance their efficiency-focused solution. They also provide a model zoo and ML libraries ready-to-use, which also reinforces their AI-first approach to their hardware.



11

Technology Platform:
Holistic Approach to AI Compute

ML Tools & Software

Circuits w/ Programmability

Chiplets &
Advanced Packaging

- Frictionless ML (BFP) Toolchain
- Large Model Mapping
- Mature, Flexible Runtime Tools

- Digital In-Memory Compute
- Programmable Engines

- Modularity & Scalability
- ISA Co-package Support
- Low Power Interconnects

Conclusions

It won't be an easy ride for D-Matrix and other start-ups in this space. Its competitors, some considerably more mature, also realized the potential of a transformer architecture. Nvidia recently unveiled the Hopper H-100, its next-generation GPU architecture, capable of up to 10x the performance of the previous hardware on large-model AI, albeit at significantly higher power consumption and cost. Another company with similar ambitions is Cerebras Systems. Its latest Wafer-scale system, the Cerebras CS-2, is the largest AI server in the market and the company claims a cluster of those could soon support a 120-trillion-parameter model for training and inference.

However, although D-Matrix is a new company entering a highly competitive space, it has an edge; it appeared just at the right time when transformers were clearly promising but still young enough that most companies hadn't had time to react. There are plenty of opportunities and strategies for companies that, like D-Matrix, are trying to get a share of the transformer market. The D-Matrix hardware could fill a space that could grow significantly in the coming years. And the vast expertise and knowledge of its founders will help them transform this advantage into a reality.

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

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 D-Matrix Inc. funding and support. Although the paper may utilize publicly available material from various vendors, including D-Matrix, 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.