Quick Take

Intel Selects Habana Technology to Drive AI Accelerator Roadmap

Michael Feldman and Alex Norton
March 2020

HYPERION RESEARCH OPINION

In February 2020, Intel announced that it will center its AI acceleration roadmap for the data center on Habana technology, abandoning the Nervana development that has driven Intel's AI accelerator strategy for the last four years. As a result, Habana's Goya and Gaudi AI platforms will replace the Nervana-based Neural Network Processor (NNP) platforms, the NNP-I and NNP-T products, which were launched last year. The decision was precipitated by Intel's $2 billion acquisition of Habana Labs, an Israeli startup specializing in AI processors for the data center, in December 2019. Intel had purchased Nervana in 2016 for close to $400 million to bring that startup's AI processor hardware and associated software technology in-house. The Habana purchase led to speculation about the disposition of the two competing product lines and which would be favored in Intel's AI strategy going forward. That decision has been made.

SITUATION OVERVIEW

Intel's switch from Nervana's to Habana's AI processor technology reflects the company's willingness to make some big bets and, when necessary, cut its losses to capture the critical AI data center market. The company has previously stated that it expects the entire AI silicon market to be worth more than $25 billion by 2024.

Although Intel has yet to offer an explicit technical or market rationale for swapping the Nervana technology with that of Habana, Intel suggested that the change was informed by feedback from early customers who had access to both technologies.

- The two sets of hardware have superficially similar architectures and capabilities, but customer evaluations in the field can often illuminate strengths and weaknesses not apparent from an engineer's block diagram.

Detailed evaluations of Habana's technology have not been made public, but Hyperion Research can offer some assessment.

- Internal tests of the Goya processor conducted by Habana using ResNet, BERT, and a handful of other models point to a highly performant platform, providing both high throughput and low latency compared with results delivered by NVIDIA GPUs. The same tests indicate Goya is power-efficient, a critical criterion for large-scale deployments envisioned for data center inference.

- There are fewer benchmark results for Gaudi, which was launched in June 2019. However, using internal tests at Habana, Gaudi has performed exceptionally well on ResNet-50, especially in large-scale configurations. Performance on other types of models have not been made public, if evaluated at all.

Scalability with Gaudi is unique among AI training hardware, utilizing integrated RDMA over Converged Ethernet (RoCE) with a total of ten 100 GbE ports per processor. The RoCE network capability provides a cost-effective way for customers to construct large scale-out training systems. Given that the size and complexity of neural network models is increasing exponentially at a time when Moore's Law is slowing, the most practical way to meet these growing computational and memory demands is via scale-out designs.

- In contrast, the Nervana technology incorporated a network capability for scale-out training as well but had a more custom implementation, potentially driving up cost.

The weakest aspect of the Habana offerings is its relatively limited software support, and it is here where Intel can leverage its extensive expertise with software tools and runtime libraries. Although Habana had a toolset based on its own SynapseAI compiler and runtime, to make the platform appealing to the widest array of developers, it needs a complete stack that supports the popular machine learning frameworks.

- Making that a reality requires a level of resources and long-term commitment that is typically feasible only by larger IT providers.
- By offering a wide array of AI-capable hardware that includes Xeon CPUs, X^e GPUs, Movidius VPUs, FPGAs, and now Habana SoCs, Intel is uniquely positioned to amortize its AI software investment across a large product portfolio.

The significant challenge for Intel is to develop a software stack as broad and deep as what NVIDIA provides with its CUDA-X AI environment. Intel's advantage is that it can apply that software over a much wider and more diverse portfolio than that of NVIDIA, or for that matter any other AI hardware provider. Although Habana-based systems may initially offer superior performance across some set of deep learning applications, NVIDIA and others can counter with similarly designed features. As a consequence, Intel's long-term success in the AI data center market could hinge on its ability to build a more attractive software ecosystem.

FUTURE OUTLOOK

The AI accelerator market for the data center is in its earliest stages. In particular, significant use of purpose-built AI accelerators is currently limited to Google's deployment of its Tensor Processing Units (TPUs). NVIDIA provides purpose-built AI acceleration in the form of specialized logic in its GPUs, but that logic is embedded in graphic processors that can be used for other types of applications. Intel, AMD, and ARM chip merchants are also building CPUs and GPUs infused with AI capabilities for mixed application environments.

- That said, the expanding market demand for more powerful and efficient AI hardware for the data center could drive greater specialization. Habana is one of the first major acquisitions in this space and marks the beginning of a period of innovation and competition centered around AI workloads.

Superimposed on this hardware trend is the continuing evolution and diversification of machine learning and deep learning models, which tend to shorten the useful life and scope of specialized hardware. That kind of volatility creates both opportunities for new architectures to emerge and the kind of risk represented by Intel's earlier acquisition of Nervana. Nonetheless, Hyperion Research expects the AI startup-acquisition cycle to continue and perhaps even intensify until the market reaches a much greater level of maturity.

About Hyperion Research, LLC

Hyperion Research provides data-driven research, analysis and recommendations for technologies, applications, and markets in high performance computing and emerging technology areas to help organizations worldwide make effective decisions and seize growth opportunities. Research includes market sizing and forecasting, share tracking, segmentation, technology and related trend analysis, and both user & vendor analysis for multi-user technical server technology used for HPC and HPDA (high performance data analysis). We provide thought leadership and practical guidance for users, vendors and other members of the HPC community by focusing on key market and technology trends across government, industry, commerce, and academia.

Headquarters

365 Summit Avenue
St. Paul, MN 55102
USA
612.812.5798
www.HyperionResearch.com and www.hpcuserforum.com
