Press Release
ZeroPoint Technologies AB

GOTHENBURG, SWEDEN
9:00 AM ET, FEB 20, 2025

ZeroPoint Technologies Unveils Groundbreaking Compression Solution to Increase Foundational Model Addressable Memory by 50%

Gothenburg, Sweden – ZeroPoint Technologies AB today announced a breakthrough hardware-accelerated memory optimization product that enables the nearly instantaneous compression and decompression of deployed foundational models, including the leading large language models (LLMs).

The new product, AI-MX, will be delivered to initial customers and partners in the second half of 2025 and will enable enterprise and hyperscale datacenters to realize a 1.5 times increase in addressable memory, memory bandwidth, and tokens served per second for applications that rely on large foundational models. The full technical specifications of AI-MX are available here.

"Foundational models are stretching the limits of even the most sophisticated datacenter infrastructures. Demand for memory capacity, power, and bandwidth continues to expand quarter-upon-quarter," said Klas Moreau, CEO of ZeroPoint Technologies. "With today's announcement, we introduce a first-of-its-kind memory optimization solution that has the potential to save companies billions of dollars per year related to building and operating large-scale datacenters for AI applications."

"Futurum Intelligence currently predicts the total AI software and tools market to reach a value of $440B by 2029 and Signal65 believes that ZeroPoint is positioned to address a key challenge within this fast-growing market with AI-MX," said Mitch Lewis, Performance Analyst at Signal65. "Signal65 believes that AI-MX is currently a unique offering and that with ongoing development and alignment with leading technology partners, there is strong growth opportunity for both ZeroPoint and AI-MX."

ZeroPoint's proprietary hardware-accelerated compression, compaction, and memory management technologies operate at low nanosecond latencies, enabling them to work more than 1000 times faster than more traditional compression algorithms.

For foundational model workloads, AI-MX enables enterprise and hyperscale datacenters to increase the addressable capacity and bandwidth of their existing memory by 1.5 times, while simultaneously gaining a significant increase in performance per watt. Critically, the new AI-MX product works across a broad variety of memory types, including HBM, LPDDR, GDDR and DDR – ensuring that the memory optimization benefits apply to nearly every possible AI acceleration use case.

A summary of the benefits provided by the initial version of AI-MX include:

ZeroPoint Technologies AB
Falkenbergsgatan 3
412 85 Göteborg
Sweden

web: www.zeropoint-tech.com
email: klas.moreau@zeropoint-tech.com
phone: +46-725-268101

Expands effective memory capacity by up to 50%

- This allows end-users to store AI model data more efficiently. For example, enabling 150GB of model data to fit within 100GB of HBM capacity.

Enhances AI accelerator capacity

- An AI accelerator with 4 HBM stacks and AI-MX can operate as if it has the capacity of 6 HBM stacks.

Improves effective memory bandwidth

- Achieve a similar 1.5 times improvement in bandwidth efficiency by transferring more model data per transaction.

The above benefits are specifically associated with the initial implementation of the AI-MX product. ZeroPoint Technologies aims to further exceed the 1.5 times increases to capacity and performance in subsequent generations of the AI-MX product.

Given the exponentially increasing memory demands of today's applications, partially driven by the explosive growth of generative AI, ZeroPoint addresses the critical need of today's hyperscale and enterprise data center operators to get the most performance and capacity possible from increasingly expensive and power-hungry memory.

For more general use cases (those not related to foundational models) ZeroPoint's solutions are proven to increase general memory capacity by 2-4x while also delivering up to 50% more performance per watt. In combination, these two effects can reduce the total cost of ownership of hyperscale data center servers by up to 25%.

ZeroPoint offers memory optimization solutions across the entire memory hierarchy - all the way from cache to storage. ZeroPoint's technology is agnostic to data load, processor type, architectures, memory technologies and processing node, and the company's IP has already been proven on a TSMC 5nm node.

About ZeroPoint Technologies AB

ZeroPoint Technologies is the leading provider of hardware-accelerated memory optimization solutions for a variety of use cases, ranging from enterprise and hyperscale datacenter implementations to consumer devices. Based in Gothenburg, Sweden, ZeroPoint has developed an extensive portfolio of intellectual property. The company was founded by Professor Per Stenström and Dr. Angelos Arelakis, with the vision to deliver the most efficient memory compression available, across the memory hierarchy, in real-time, based on state-of-the-art research. For more information, visit **https://www.zeropoint-tech.com/**.

For further information contact **Klas Moreau, CEO** at +46-725-268101