

IP OFFERINGS

More tokens per second. More bandwidth. More capacity.

Memory and storage are the bottleneck for modern AI and data center silicon, and they get more expensive every quarter. ZeroPoint hardware compression IP runs inline at line rate, so existing memory and storage systems carry more data than their specs allow.

The result is higher inference throughput, more effective bandwidth, and more usable capacity from the DRAM and NAND you already have, at the lowest latency and best GB/s/mm² of any compression solution, with no model retraining and no changes to the host software stack.

Up to 5x

Lossless compression on streaming data

Up to 2x

More effective bandwidth across links and interfaces

Up to 2x

More usable capacity from existing memory and storage

The ZeroPoint Technologies product portfolio

Lossless capacity and bandwidth improvements across AI inference, chip-to-chip, memory pooling, and storage applications. Every product is delivered as an IP block that integrates directly into your SoC, accelerator, or storage controller, without changing memory technology, widening buses, retraining models, or losing precision.

STREAMING BANDWIDTH

ZeroStream™

Hardware link (de)compression for die-to-die, chip-to-chip, and DRAM interfaces. Expands effective bandwidth with deterministic, data-independent latency, fully pipelined to sustain line rate.

Bandwidth Up to 2x effective

Latency Deterministic, fully pipelined

Designed for Streaming, HBM, LPDDR, DDR, chip-to-chip

CHIP-TO-CHIP / CXL LINKS

ZeroConnect™

Hardware compressed-memory IP for CXL devices, pairing inline compression with full memory management, address remapping, and caching to expand CXL capacity and lower data center TCO.

Capacity Up to 2x more CXL memory

Management Remapping, caching, capacity management

Designed for CXL, die-to-die links, data center TCO

AI MODEL COMPRESSION

ZeroAI™

Software compression with hardware decompression for static model data, plus full hardware compression and decompression for weights, KV cache, and activations. Extremely low latency, no model retraining.

Capacity / Bandwidth Up to 2x within a fixed budget

Retraining None required

Memory HBM, GDDR, LPDDR, DDR

STORAGE APPLICATIONS

ZeroStorage™

Standards-based, LZ4-compatible hardware compressor and decompressor for storage and SSD controllers. Software compatible and transparent to the host file system.

Capacity Up to 2x more usable

Algorithm LZ4-compatible, standards-based

Efficiency Industry-leading per mm²

Why ZeroPoint

- ▶ Lowest latency and best GB/s/mm² of any compression solution.
- ▶ Delivered as IP blocks that drop into your SoC, accelerator, or controller.
- ▶ Experts in compression, compute, and memory architecture.
- ▶ Runs inline at line rate, with no host software changes and no retraining.
- ▶ Proven across AI inference, chip-to-chip, and storage workloads.
- ▶ More than 70 worldwide patents.

Talk to us about doing more with the memory and storage you already have.

Tell us your target node, memory, and workload, and we will scope the right IP and licensing path with you.

ZeroPoint Technologies AB

Falkenbergsgatan 3
412 85 Göteborg, Sweden

ZeroPoint Technologies, Inc.

470 Ramona Street
Palo Alto, CA 94301, USA

Contact info@zeropoint-tech.com

Web zeropoint-tech.com/products