GroqChip™ Processor.

Unlock the power of ultra-low latency.

The revolutionary, fully deterministic GroqChip processor, built on the LPU™ architecture, is the core of scalable performance. Built from the ground up to accelerate AI, ML, and HPC workloads, GroqChip reduces data movement for predictable low-latency performance, bottleneck-free.

Key Features

Fully deterministic processor provides predictable and repeatable performance with no run-to-run variation.

16 chip-to-chip interconnects allow chips to talk directly to each other for scalable low-latency performance without the need for extra switches, cards, or CPUs.

Up to 80 TBs on-die memory bandwidth facilitates massive concurrency and data parallelism needed for bandwidth sensitive applications.

Simplified SIMD architecture leads to an easy-to-use software suite, accelerating developer velocity.

230 MB of on-die memory delivers large globally sharable SRAM for high-bandwidth, low-latency access to model parameters without the need for external memory.

End-to-end on-chip protection improves uptime and reliability with error-correction code (ECC) protection throughout the entire GroqChip data path.

Specifications

Item	Description
Availability	Available as apart of a GroqRack™ compute cluster
Process Node	14nm
Performance	Up to 750 TOPs, 188 TFLOPs (INT8, FP16 @900 MHz)
Memory	230 MB SRAM per chip Up to 80 TB/s on-die memory bandwidth
Chip Scaling	16 integrated RealScale™ chip-to-chip interconnects
I/O	Integrated PCIe Gen4 x16 controller
Numerics	INT8, INT16, INT32 & TruePoint™ technology MXM: FP32 VXM: FP16, FP32
Power	Max: 300W; TDP: 215W; Average: 185W

Groq Inc. HQ
301 Castro St. Suite 200
Mountain View, CA 94041

Mailing Address
PO Box 1778
Mountain View, CA 94041

www.groq.com

For more information visit groq.com or contact us at info@groq.com.