GroqCard™ Accelerator.

Ready, set, done. Guaranteed low latency.

For plug and play low latency, scalable performance, GroqCard accelerator packages a single GroqChip™ processor into a standard PCIe Gen4 x16 form factor providing hassle-free server integration. Featuring up to 11 RealScale™ chip-to-chip connections alongside an internal software-defined network, GroqCard enables near-linear multi-server and multi-rack scalability without the need for external switches.

Key Features

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

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.

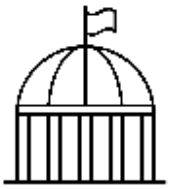
Up to 80 TBs on-die memory bandwidth facilitates massive concurrency and data parallelism needed for bandwidth sensitive applications.

11 RealScale™ chip-to-chip connectors enable near-linear multi-server and multi-rack scalability without the need for external switches.

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

PCIe Gen4 x16 interface delivers up to 31.5GBs of bi-directional bandwidth in an industry standard interface for fast device and network connections - all with a lightweight open source driver and no CPU burden.

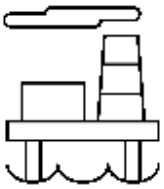
Targeted Applications



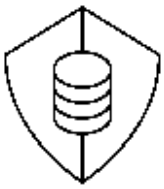
Government

Financial Services

Enterprise Comms



Oil & Gas



Cyber & InfoSec

Research & Sciences

GroqCard SKUs

Available within GroqNode™ Servers	**GC1-010B**	11 RealScale chip-to-chip connectors Modified PCIe CEM specification form factor Designed for users that need the ultimate scalability
Available as an add-in card supported with select 3rd Party Qualified Servers. Contact Groq for supported servers.	**GC1-0109**	9 RealScale chip-to-chip connectors PCIe CEM specification form factor Designed for users that require OEM server support
	GC1-0100	0 RealScale chip-to-chip connectors PCIe CEM specification form factor Designed for users on a budget without scalability needs

Specifications

Item	Description
Availability	Shipping now
Form Factor	Dual width, full height, ¾ length PCI Express Gen4 x16 adapter
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	Up to 11 RealScale™ chip-to-chip connectors
Numerics	INT8, INT16, INT32 & TruePoint™ technology MXM: FP32 VXM: FP16, FP32
Power	Max: 375W; TDP: 275 ; Typical: 240W

Groq Inc. HQ
400 Castro St. Suite 600
Mountain View, CA 94041

Mailing Address
PO Box 1778
Mountain View, CA 94041

www.groq.com
support.groq.com

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