GroqRack™ Compute Cluster.

The backbone of low latency, large-scale deployments.

For data center deployments, GroqRack provides an extensible accelerator network. Combining the power of an eight GroqNode™ set, GroqRack features up to 64 interconnected chips. The result is a deterministic network with an end-to-end latency of only 1.6µs for a single rack, ideal for massive workloads and designed to scale out to an entire data center.

Key Features

Eight GroqNode™ servers with 64 interconnected cards plus 1 additional redundant node reduces unexpected downtime impact.

14 GB shared global SRAM delivers large globally sharable SRAM for high-bandwidth, low-latency access to model parameters.

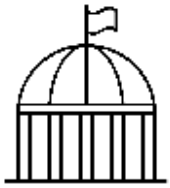
Low latency and high performance delivers large globally sharable SRAM for high-bandwidth, low-latency access to model parameters.

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

Up to 3.2 TBps global bisectional bandwidth facilitates massive concurrency and data parallelism needed for bandwidth-sensitive applications.

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

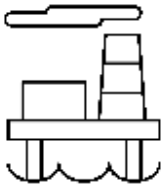
Targeted Applications



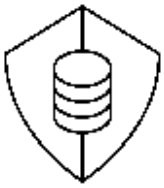
Government

Financial Services

Enterprise Comms



Oil & Gas



Cyber & InfoSec

Research & Sciences

GR1-B9A Specifications

Feature	Description
Availability	Shipping now to select customers. Broadly available 2H'22.
Chassis	GroqRack 42U Server Chassis
GroqNode Servers	Up to 9 x GroqNode 1 (GN1-B8C) servers with a fully connected internal RealScale network delivering accelerated compute performance up to 48 POPs (INT8), 12 PFLOPs (FP16)
Realscale Network	288 x QSFP28 GroqNode connectors creating a switchless routing fabric with 3.2TBps total global bisection bandwidth across a single global hop
Server Management	Primary Rack Controller Server with AMD EYPC 7413 Processor
NVMe Server	NVMe Server with dual AMD EYPC 7413 Processors with 8 x 7.68TB NVME installed for data (61.44TB total storage)
Ethernet	2 x 100GE Ethernet Switches (data network + redundancy) with min 32 x QSFP28 ports 1GE Ethernet Switch (management network) with min 48 x RJ45 ports
Power	4x 17.3kW PDUs (2N Redundancy)
OS and Software	Ubuntu Linux with GroqWare™ Suite (SDK and Utilities) pre-installed on dual SATA SSD drives (RAID1 config). Optional support for RHEL & Rocky Linux.

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.