GroqNode™ Server.

Unprecedented low latency meets uncompromised scalability.

For large scale deployments, GroqNode server provides a rack-ready scalable compute system. GroqNode, an eight GroqCard™ accelerator set, features integrated chip-to-chip connections alongside dual server-class CPUs and up to 1 TB of DRAM in a 4U server chassis. GroqNode is built to enable high performance and low latency deployment of large deep learning models.

Key Features

Eight GroqCard accelerators combine the power of interconnected chips to tackle large model deployments.

1.76 GB of on-die memory delivers large globally shared SRAM for high-bandwidth, low-latency access to model parameters.

4U server chassis simplifies rack integration with a standard form factor and 10x PCIe Gen 4 x16 slots.

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

Up to 640 TB/s on-die memory 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 (ECC) protection throughout the entire GroqChip™ data path.

GN1-B8C / GN1-B8C-C1 Specifications

Feature	Description
Availability	Available as apart of a GroqRack™ compute cluster
Chassis	GroqNode 4U server chassis, 7.0” (H) x 17.2” (W) x 29” (D)
Accelerators	Up to 8 x GroqCard 1 (GC1-010B) accelerators with a fully connected internal RealScale network delivering accelerated compute performance up to 6 POPs, 1.5 PFLOPs (INT8, FP16)
Model Memory	1.76 GB on-die SRAM (230 MB per GroqChip™) Up to 640 TB/s on-die memory bandwidth (80 TB/s per GroqChip)
Cluster-ready	Up to 32 x RealScale external ports 200Gb/s HDR Infiniband or Ethernet NIC
Host Processor	2x AMD EPYC™ 7313 processors (3GHz, 16C/32T, 155W TDP each)
Host Memory	1TB DDR4-3200 DRAM with ECC (16 x 64GB RDIMMs)
Optional High-performance Storage	Optional integrated 7.68TB PCIe NVMe SSD device provides on-host data storage with low latency and high bandwidth
Server Management	In-band: 2 x 1GbE ports for host OS access Out-of-band: Dedicated 1GbE port for baseband management
Power	4 x 2000W (220-240VAC) with 4 x C13-C14 cables (4ft length)
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
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.