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

<table>
<thead>
<tr>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>Shipping now to select customers. Broadly available 2H’22.</td>
</tr>
<tr>
<td>Chassis</td>
<td>GroqRack 42U Server Chassis</td>
</tr>
<tr>
<td>GroqNode Servers</td>
<td>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)</td>
</tr>
<tr>
<td>Realscale Network</td>
<td>288 x QSFP28 GroqNode connectors creating a switchless routing fabric with 3.2TBps total global bisection bandwidth across a single global hop</td>
</tr>
<tr>
<td>Server Management</td>
<td>Primary Rack Controller Server with AMD EYPC 7413 Processor</td>
</tr>
<tr>
<td>NVMe Server</td>
<td>NVMe Server with dual AMD EYPC 7413 Processors with 8 x 7.68TB NVME installed for data (61.44TB total storage)</td>
</tr>
<tr>
<td>Ethernet</td>
<td>2 x 100GE Ethernet Switches (data network + redundancy) with min 32 x QSFP28 ports</td>
</tr>
<tr>
<td></td>
<td>1GE Ethernet Switch (management network) with min 48 x RJ45 ports</td>
</tr>
<tr>
<td>Power</td>
<td>4x 17.3kW PDUs (2N Redundancy)</td>
</tr>
<tr>
<td>OS and Software</td>
<td>Ubuntu Linux with GroqWare™ Suite (SDK and Utilities) pre-installed on dual SATA SSD drives (RAID1 config). Optional support for RHEL &amp; Rocky Linux.</td>
</tr>
</tbody>
</table>

For more information visit [groq.com](http://groq.com) or contact us at [info@groq.com](mailto:info@groq.com).