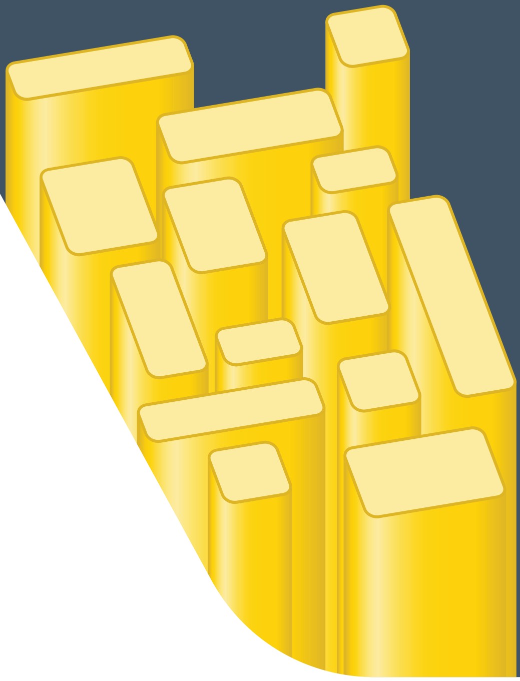
BittWare
a molex company

Partner Solution



groq™

GroqCard™ Accelerator

Real-Time AI Acceleration

The GroqCard is a double-width PCIe form factor ML accelerator that's hassle-free to integrate. The GroqWare™ suite follows a software-defined hardware approach, giving easy deployment paths for your PyTorch, TensorFlow, and ONNX-trained deep learning models.

Scalability is a core feature of the GroqCard, with 9 RealScale chip-to-chip connections that ensure deployment of multiple cards is as efficient as one. An internal software defined network provides predictable, repeatable performance with no run-to-run variations.

Fully deterministic processor

Predictable and repeatable performance with no run-to-run variation

9 RealScale™ chip-to-chip connectors

Near-linear multi-server and multi-rack scalability without the need for external switches

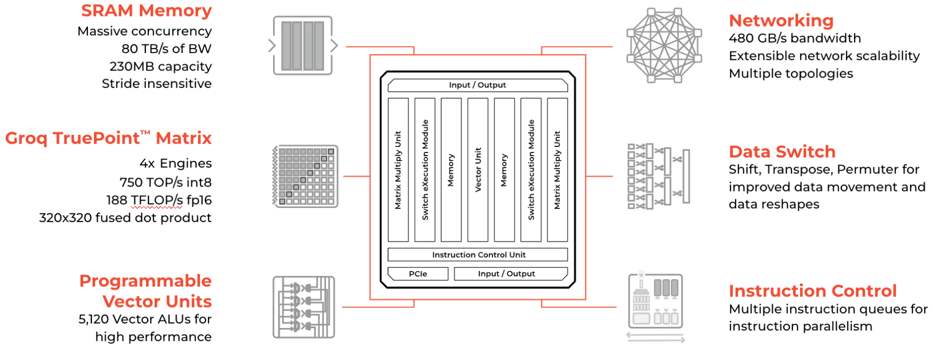
End-to-end on-chip protection

Improves uptime and reliability with error-correction code (ECC) protection

Ready, set, done.
Guaranteed low latency.

GroqChip™ 1 Overview

Scalable compute architecture



SRAM Memory
Massive concurrency
80 TB/s of BW
230MB capacity
Stride insensitive

Groq TruePoint™ Matrix
4x Engines
750 TOP/s int8
188 TFLOP/s fp16
320x320 fused dot product

Programmable Vector Units
5,120 Vector ALUs for high performance

Networking
480 GB/s bandwidth
Extensible network scalability
Multiple topologies

Data Switch
Shift, Transpose, Permuter for improved data movement and data reshapes

Instruction Control
Multiple instruction queues for instruction parallelism

GroqCard Accelerator

Real-Time AI Acceleration

Simplify programming with

GroqWare™ Suite

GroqWare Suite is a comprehensive and versatile software stack designed to accelerate a variety of HPC and ML workloads. Composed of Groq™ Compiler, Groq API, and Utilities, the suite eases deployment implementations with an open source driver/runtime and support for industry standard AI/ML frameworks.

GroqFlow™ Tool Chain (included in the GroqWare Suite) enables a single line of Pytorch or TensorFlow code to import and transform existing models through a fully automated tool chain to run on Groq hardware.

Card Specifications

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 9 RealScale™ chip-to-chip connectors

Numerics

INT8, INT16, INT32 & TruePoint™ technology
MXM: FP32
VXM: FP16, FP32

Power

Max: 375W; TDP: 275 ; Typical: 240W

Sales Part Number

RS-GQ-GC1-0109	GroqCard PCIe ML accelerator card

Looking for a different configuration? Ask us about other configuration options.

What is GroqChip™ Processor?

A scalable processor built from the ground up to accelerate AI, ML, and HPC workloads.

The revolutionary, fully deterministic GroqChip processor 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. This standalone chip provides flexible integration into compute intensive applications.

The architecture is much simpler than a GPU and is designed with a software-first focus, making it easier to program and providing predictable performance with lower latency.

Get GroqCard pre-integrated in a high-density server with a comprehensive warranty.

To learn more, visit **www.BittWare.com**

Rev 2023.10.23 | October 2023

BittWare
a molex company