groq™

Accelerating Systems with Real-time AI Solutions

EXPLORE THE PRODUCT SUITE ▶

Groq provides new solutions to address the evolving challenges of real-time AI and HPC.

Comprehensive end-to-end acceleration solutions from scalable, ultra-low latency systems to generalized software that deliver orders of magnitude performance improvements.

GroqChip™

GroqCard™

GroqNode™

GroqRack™

GroqChip Processor

- The Tensor Streaming Processor architecture allows for model ideation and experimentation while large SRAM capacity allows for low latency processing of high accuracy models
- Built for long-term viability, with the flexibility to accommodate AI model evolution

GroqNode Server

- Provides maximum bandwidth with eight fully connected GroqCard accelerators, utilizing a dragonfly topology interconnect to support larger models with enhanced accuracy
- Preconfigured with an operating system, pre-installed drivers, system monitoring, and the full GroqWare™ Suite

GroqCard Accelerator

- A passively cooled, dual slot, three quarter length card with a PCIe Gen4x16 interface
- Plug-and-play compatibility with existing server deployments saves setup time and cost

GroqRack Compute Cluster

- Accelerates deployment of state-of-the-art models with predictable performance at scale, enabling enterprises to meet the most demanding SLAs and keep pace with the evolving computing landscape
- Contact us to prepare for your system's future demands with GroqRack at info@groq.com

Scan and Download the Full Product Spec Sheets

Learn more at groq.com

Developers Love groq™

TWO PATHS FOR DEVELOPMENT

1

Groq™ Compiler

OUT-OF-THE-BOX

Provides out-of-the-box support for standard Deep Learning models

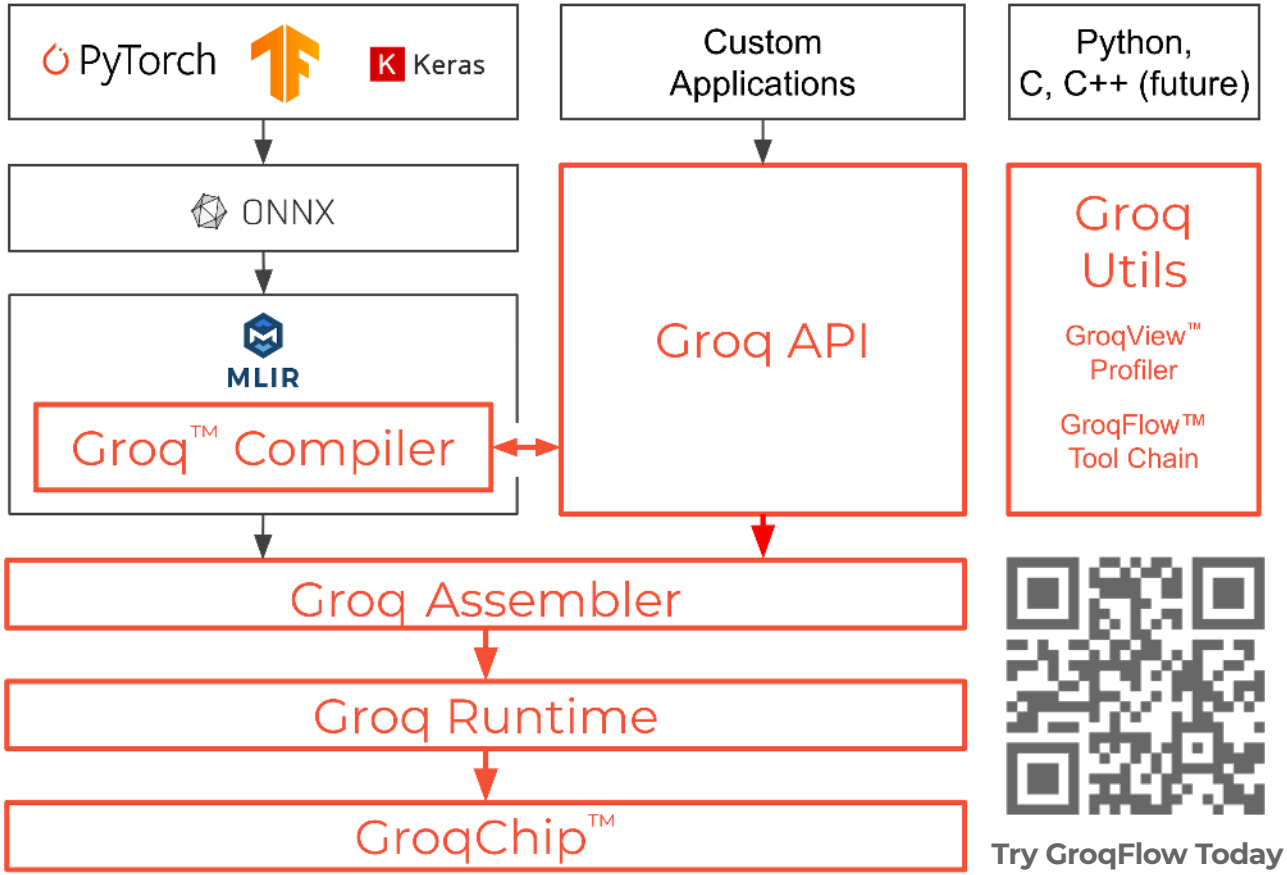
2

Groq API

FINE-GRAINED CONTROL

Provides finer grained control of GroqChip™ in order to support custom applications

GroqWare™ Suite Components



Try GroqFlow Today

Kills the slow, painful dynamic profiling process for developers.

GroqView™ Profiler provides visualization of the chip's compute and memory usage at compile time.

GroqFlow™ enables a single line code to import and transform models through a fully automated tool chain to run on Groq hardware.