

Hello |  **Communities**





Low Latency and Strong Scaling Inference With Groq
Software-scheduled Deterministic Architecture.

Max Engelen

Dataflow Engineer

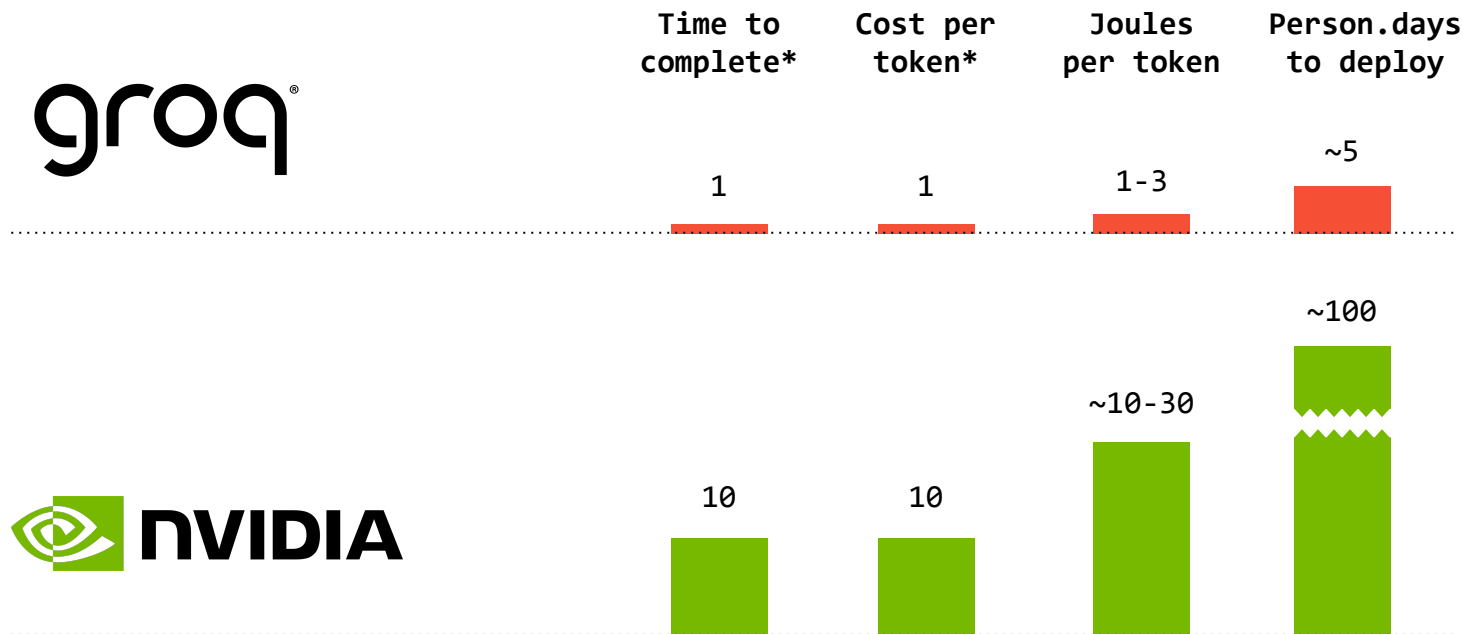
LLM Demo

LPU™ Inference Engine

Ten GroqRack™ Compute Clusters



Groq's LPU™ Inference Engine



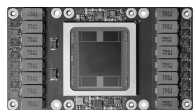
Public Latency Benchmark: <https://github.com/ray-project/llmperf-leaderboard?tab=readme-ov-file>



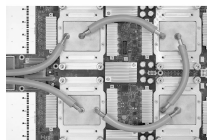
LPU™ Inference Engine

Golden Age of Computer
Architecture

Explosion of Domain Specific
Architectures (DSA)



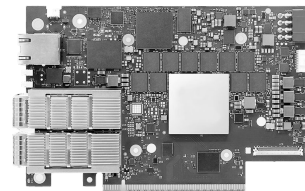
GPU



TPU



VPU



DPU

The Challenge with AI Accelerators : **Compilers**

Ref. A.Bitars, Presented at Crossroads 3D-FPGA Academic Research Center - December 2022

Algorithms — Compilers —→ Hardware

Dataflow
dominated

Statically predictable set
of executed operations

Highly-parallel
vector operations

✓ **PREDICTABLE**

Remain a challenge

Reliant on hand-tuned
libraries

Fragmented front-end
ecosystem

Require iterative
hardware profiling

High-density compute
using SIMD

Less silicon area spent
on re-ordering and
speculation

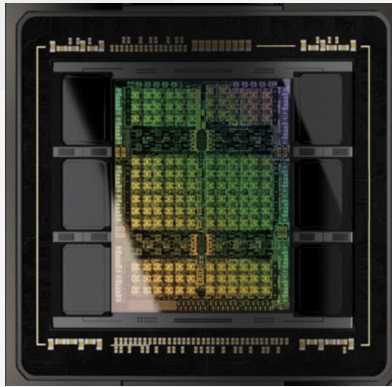
More memory
bandwidth

✗ **UNPREDICTABLE**

✖ UNPREDICTABLE

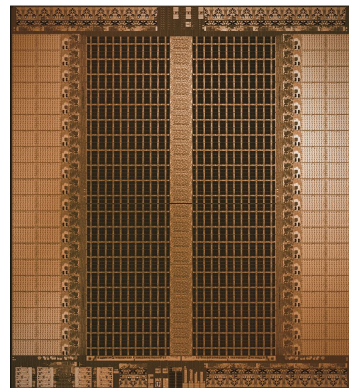
✔ PREDICTABLE

4nm



No CoWoS
No HBM
1/11x silicon area
1/31x devices

14nm



Groq
Simplifies
Compute

Graphic Processor

COMPLEX

Difficult programming
Less responsiveness
Non-Deterministic execution
Higher costs

LPU™ Inference Engine

SIMPLIFIED

Easier compilation
Lower latency
Deterministic / Predictable execution
Massive scale

GroqChip™ Overview

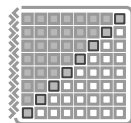
SRAM Memory

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



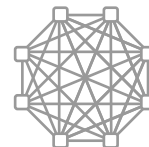
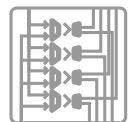
Groq TruePoint™ Matrix

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



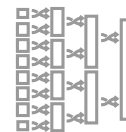
Programmable Vector Units

5,120 Vector ALUs for high performance



Networking

480 GB/s bandwidth
Extensible network scalability
Multiple topologies



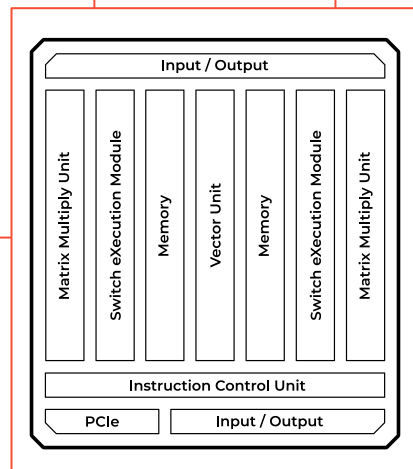
Data Switch

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



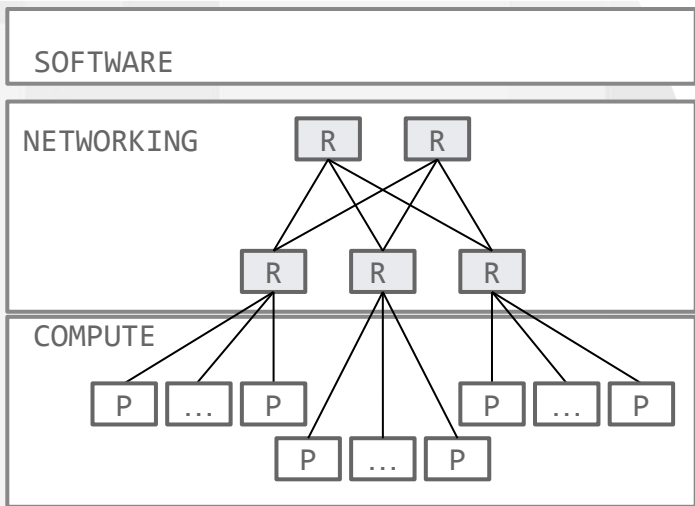
Instruction Control

Multiple instruction queues for instruction parallelism



Groq Simplifies Interconnect

❌ UNPREDICTABLE



Conventional Network

Disjoint Compute/Networking

COMPLEX

Per-hop Router arbitration

(High Latency)

Hardware-based global adaptive routing

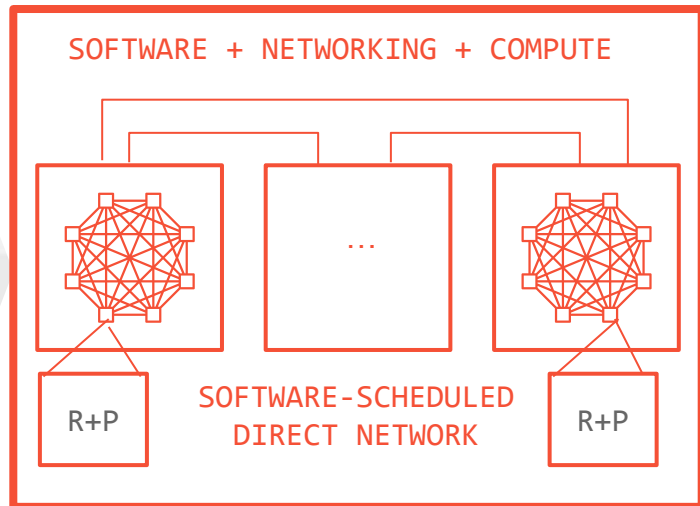
(Weak Scaling)

Congestion sensing in the network through backpressure

(Non-deterministic delay)

High network load sensitivity

✅ PREDICTABLE



SW Controlled Network

SW Orchestrated Compute & Network

SIMPLIFIED

No hardware-arbitration for dynamic contention

(Lower latency)

No hardware routing

(Massive Scale)

No congestion sensing

(Deterministic Delay)

Low network load sensitivity

Software-scheduled Network

Synchronous chip-to-chip communication

RealScale™ chip-to-chip (C2C) interconnect enables synchronous communication

- Clock drift across chips is accounted for and mitigated deterministically

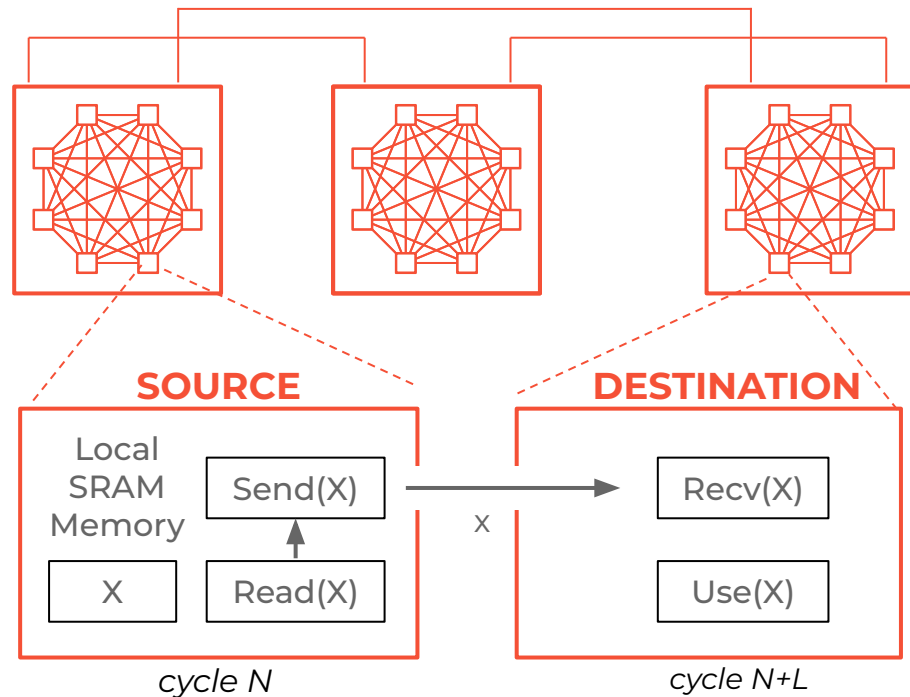
Each chip acts as both a processor and router

- Compiler schedules messages as part of programs loaded onto each chip

No adaptive routing / congestion sensing needed

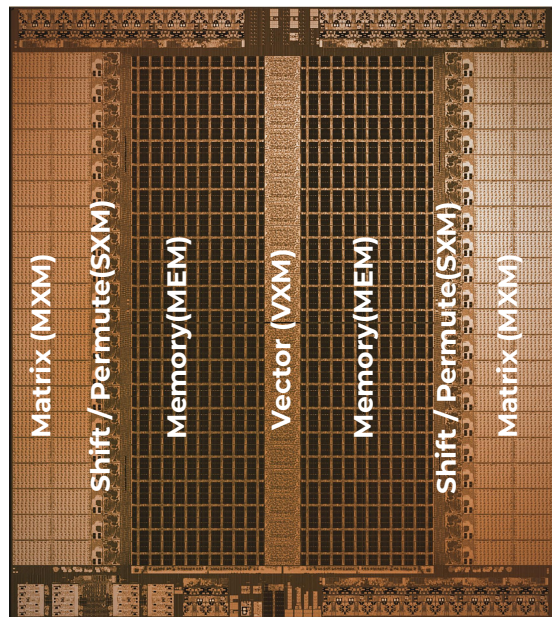
- Compiler knows exact cycle data should be sent from one chip and received at another

SOFTWARE-SCHEDULED DIRECT NETWORK

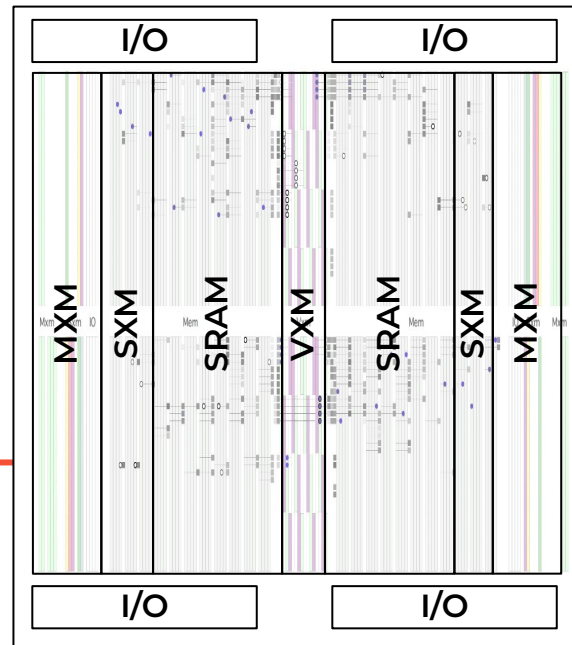


GROQ Data Orchestration Enables

Software & Hardware Co-optimization



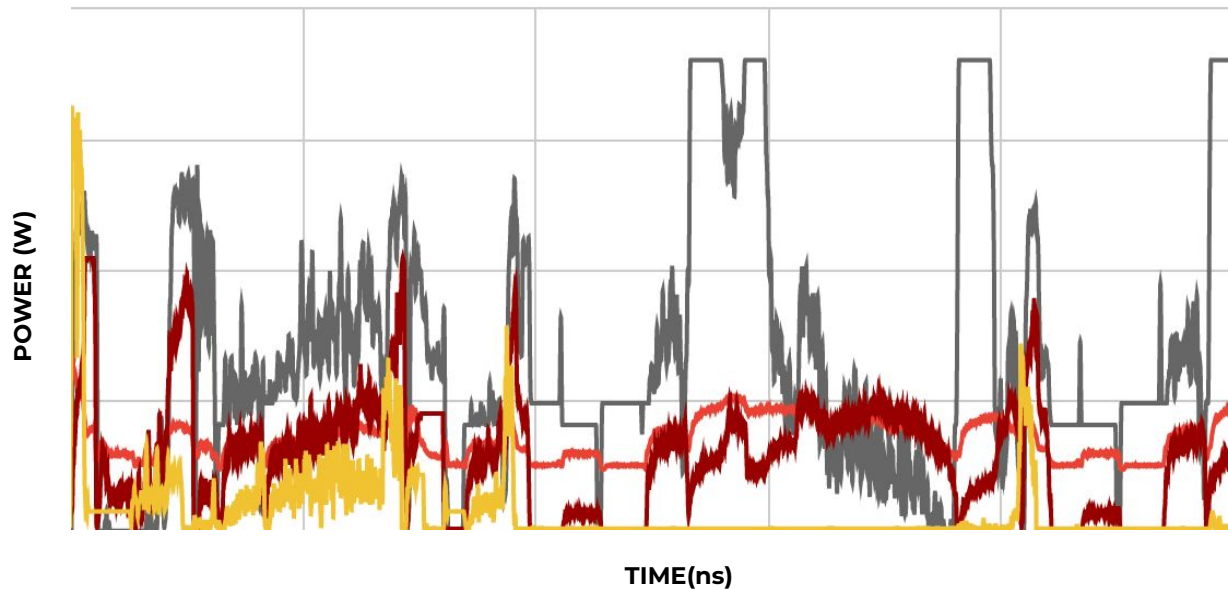
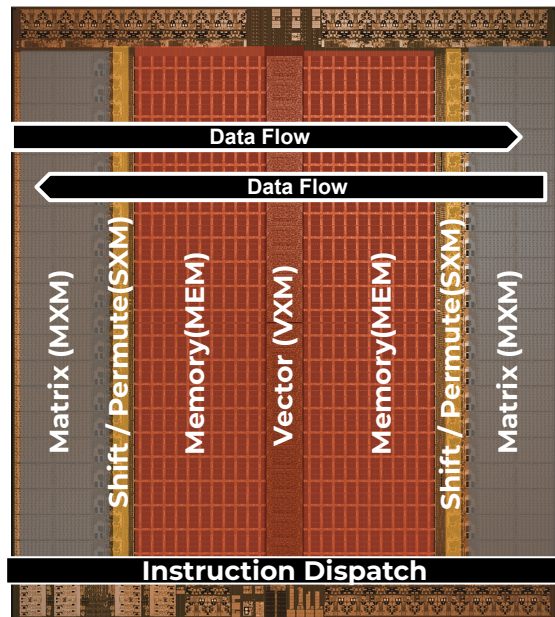
GROQ® COMPILER ENABLES
Hardware & Software
Co-optimization



Enables Performance, Power, Ldi/dt, & Thermal Profiling

GroqChip™ Functional Units Power Over Time

— MXM — MEM — VXM — SXM



Groq Compiler can profile 100% deterministic power, temp, di/dt down to a “ns”

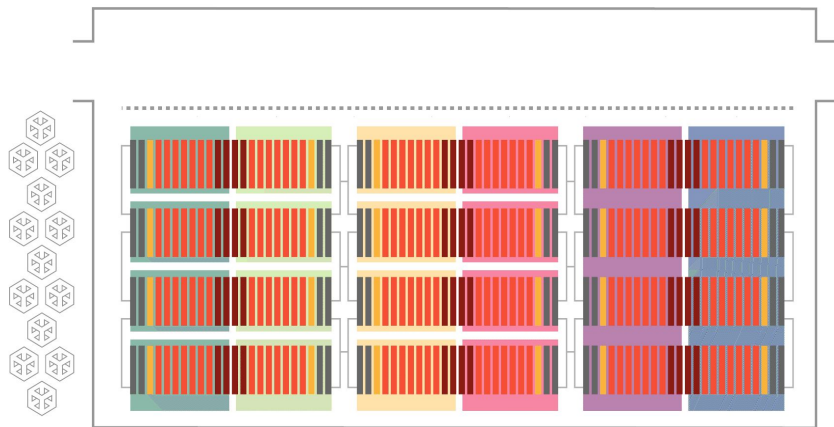
GPUs scale largely in time, some in space (clustering)

For large compute volume, GPUs iterate over multiple partitions of model code, weights etc in time

- Most of GPU time/energy spent paging weights & KV cache in/out of **HBM**
- **Highly Inefficient → High Cost / Token**
 - ✗ Low HBM bandwidth 1/100X of on-chip SRAM
 - ✗ High HBM access latency (300ns-1300ns)
 - ✗ High HBM access power (4-6 pJ/bit for R/W)
 - ✗ Need high-batch size to saturate compute (100s-1000s)
 - ✗ Poor GPU-to-GPU collectives with asynchronous communication (through switches) and high batch sizes results in high latency
- **Expensive BOM & Supply Concerns** with HBM, exotic packaging, network switches, etc.

Groq TSP scales largely in space as an assembly line

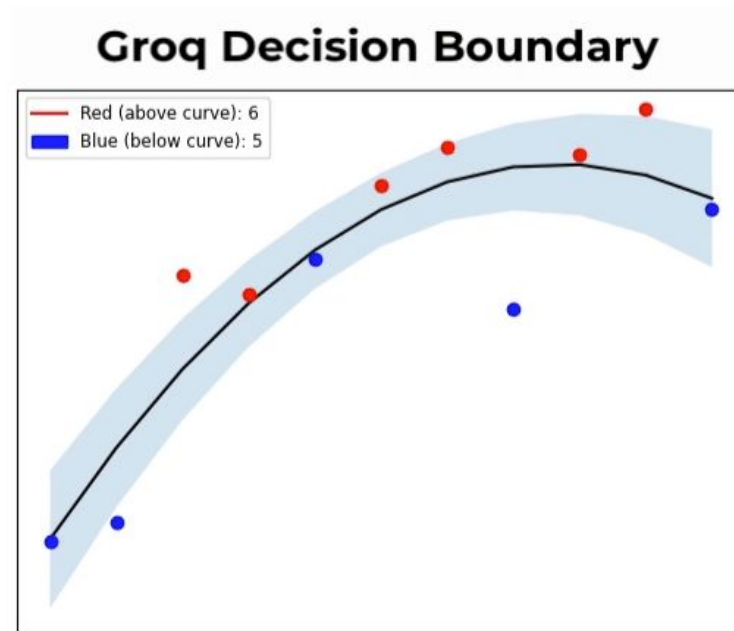
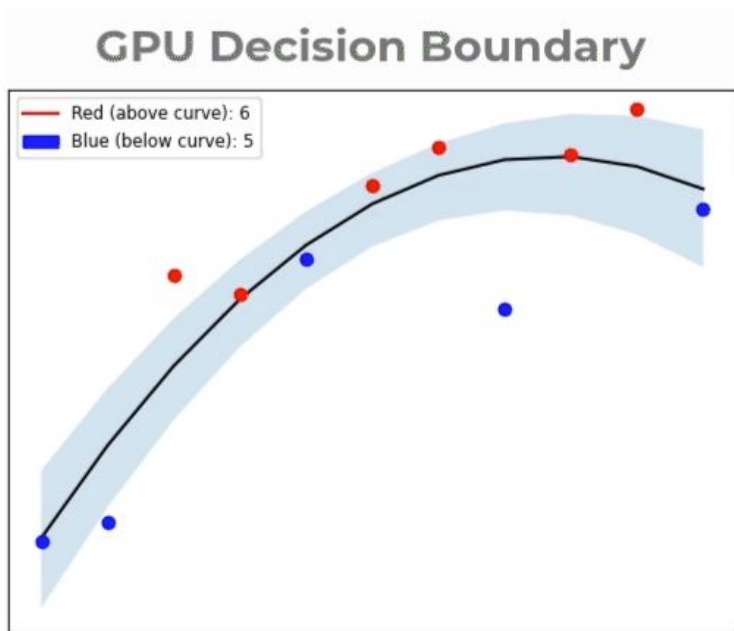
For large compute volume, TSPs partition model code across multiple chips to form an assembly line



- Break through the memory wall by computing on weights & KV cache from **SRAM**
- **Highly efficient** → **Lowest Cost / Token**
 - ✓ 100X higher bandwidth than HBM
 - ✓ Lowest SRAM access latency (<5 ns)
 - ✓ Lowest SRAM power (0.3 pJ/bit for R/W)
 - ✓ Saturate compute at low batch sizes
 - ✓ Efficient LPU-to-LPU collectives due to deterministic communication and low batch size
- Single chip module, no HBMs, no expensive external switches → abundant supply

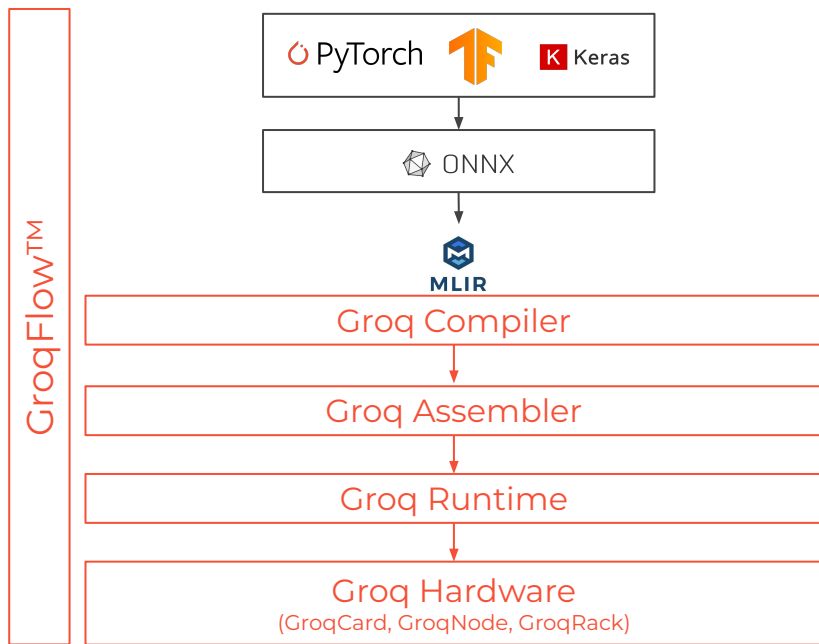
Deterministic Computing

Guarantees reproducibility and ensures every inference produces the same prediction



Programmability

GroqWare™ Suite



DIVERSE SUITE OF DEVELOPMENT TOOLS

Out-of-Box

Groq Compiler provides out-of-box support for standard Deep Learning models



Productivity Tools

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

GroqFlow Tool Chain enables a single line of Pytorch or TensorFlow code to import and transform models through a fully automated tool chain to run on Groq hardware

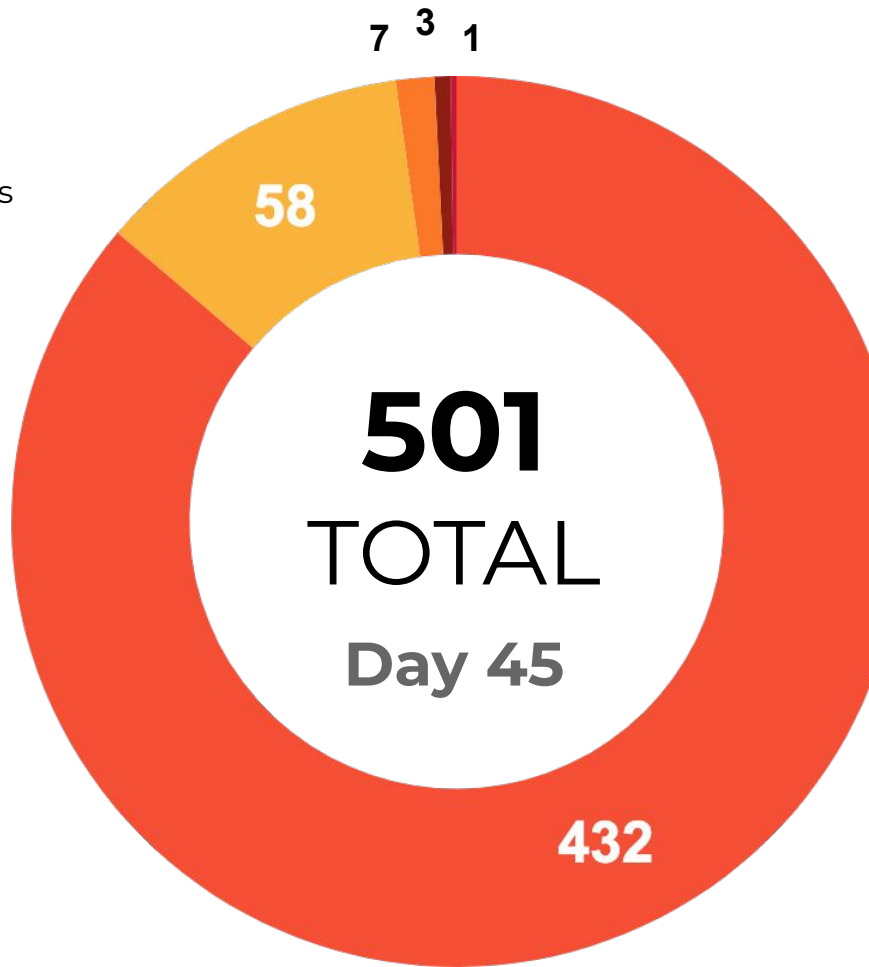
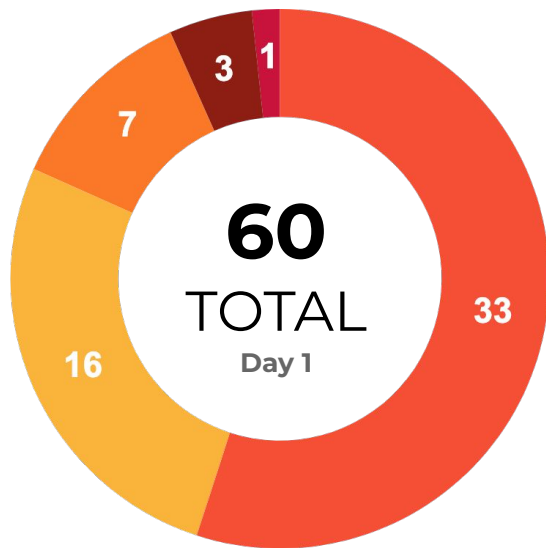
Better Than We Imagined

Push-button deployment from a diversity of public repositories

We estimate **~20,000,000** person-hours to develop all of this with CUDA/Kernels—but **Groq is kernel-less**

COMPILED MODELS

- Hugging Face
- Tourchhub: CL
- ONNX CNN
- Graph Convolutions
- Tourchhub: De



We maximize human capital
giving advantages to
organizations that must
compete for talent.



~35

SW Engineers

groq™



>50,000

Kernel Developers

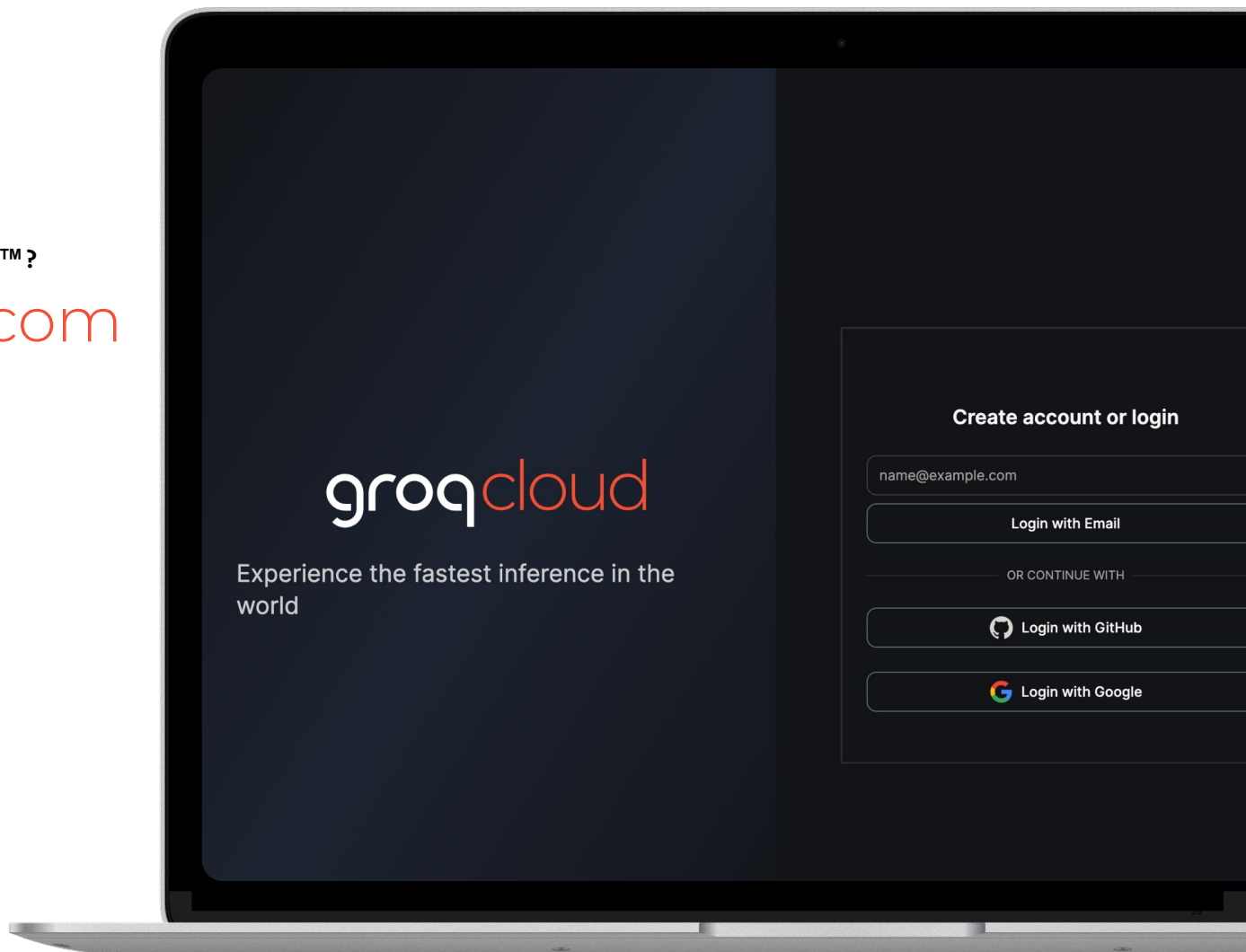
 **nvidia**

GroqCloud

How to get access to GroqCloud™?

console.groq.com

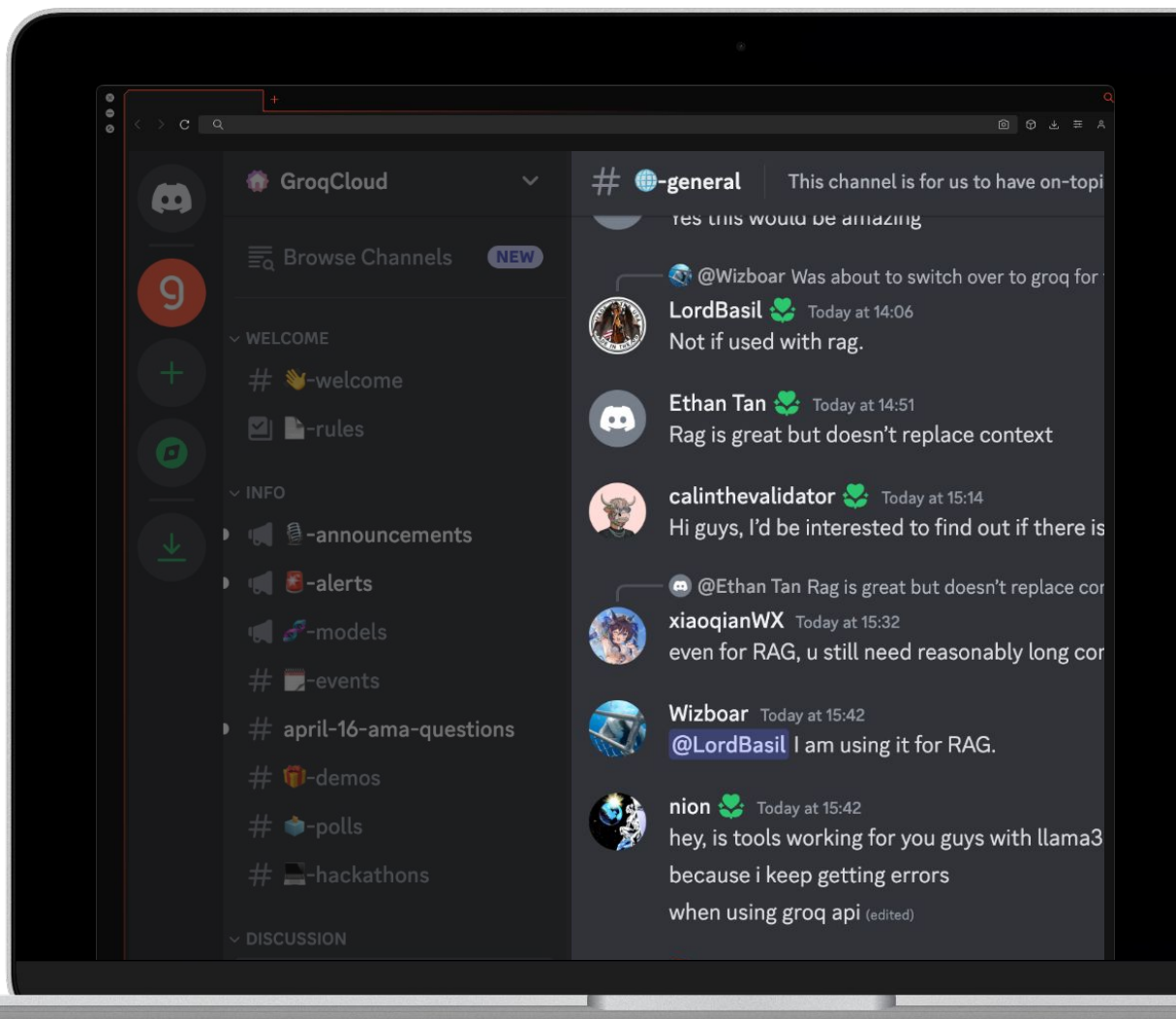
groq



Join our Developer Community



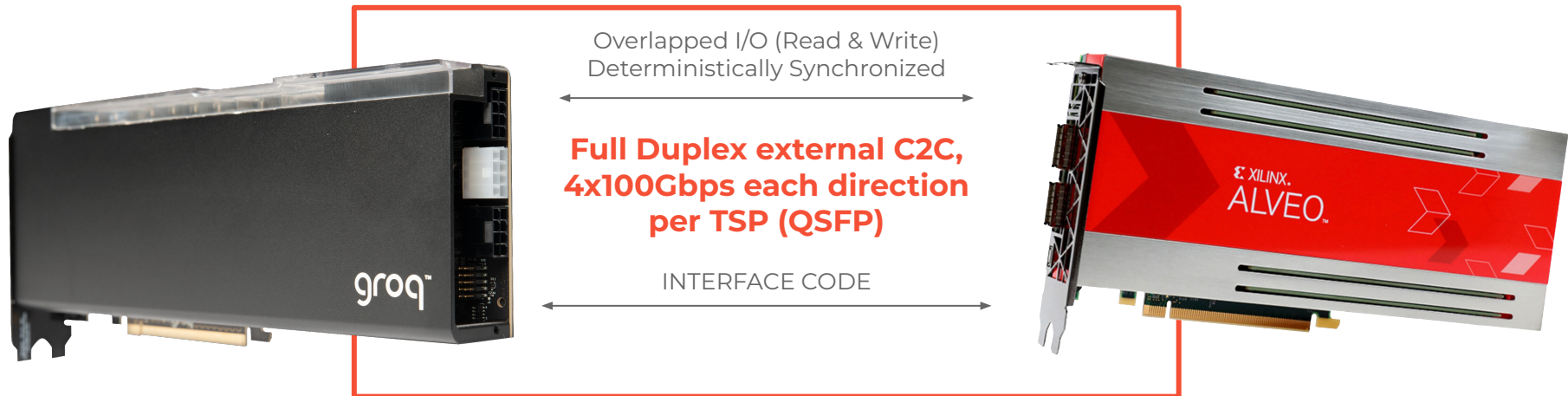
groq



Groq IO Accelerator

Groq IO Accelerator

GroqChip + FPGA vastly extends use cases and connectivity



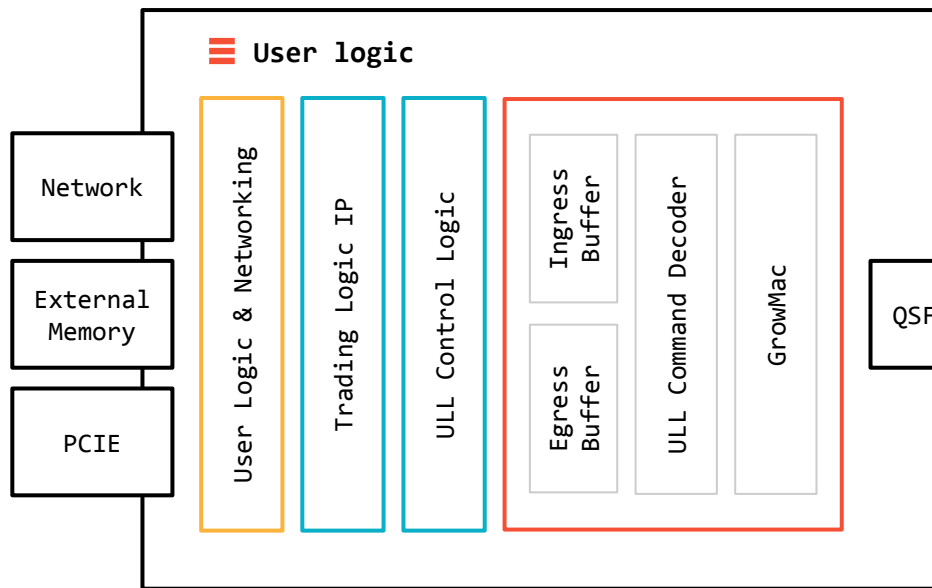
A very high speed, deterministic processor for:

- Real-time AI inference
- Compute intensive offload

A very high speed, synchronized, interface which in turn can provide:

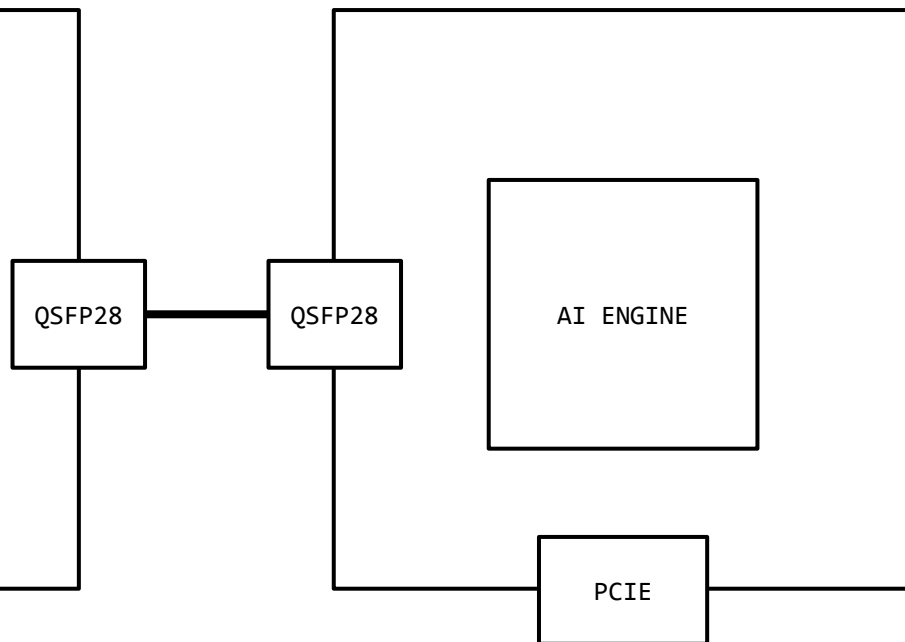
- Real-time data IO
- Application specific interfacing
- Data preprocessing/conversion
- Memory expansion

FPGA

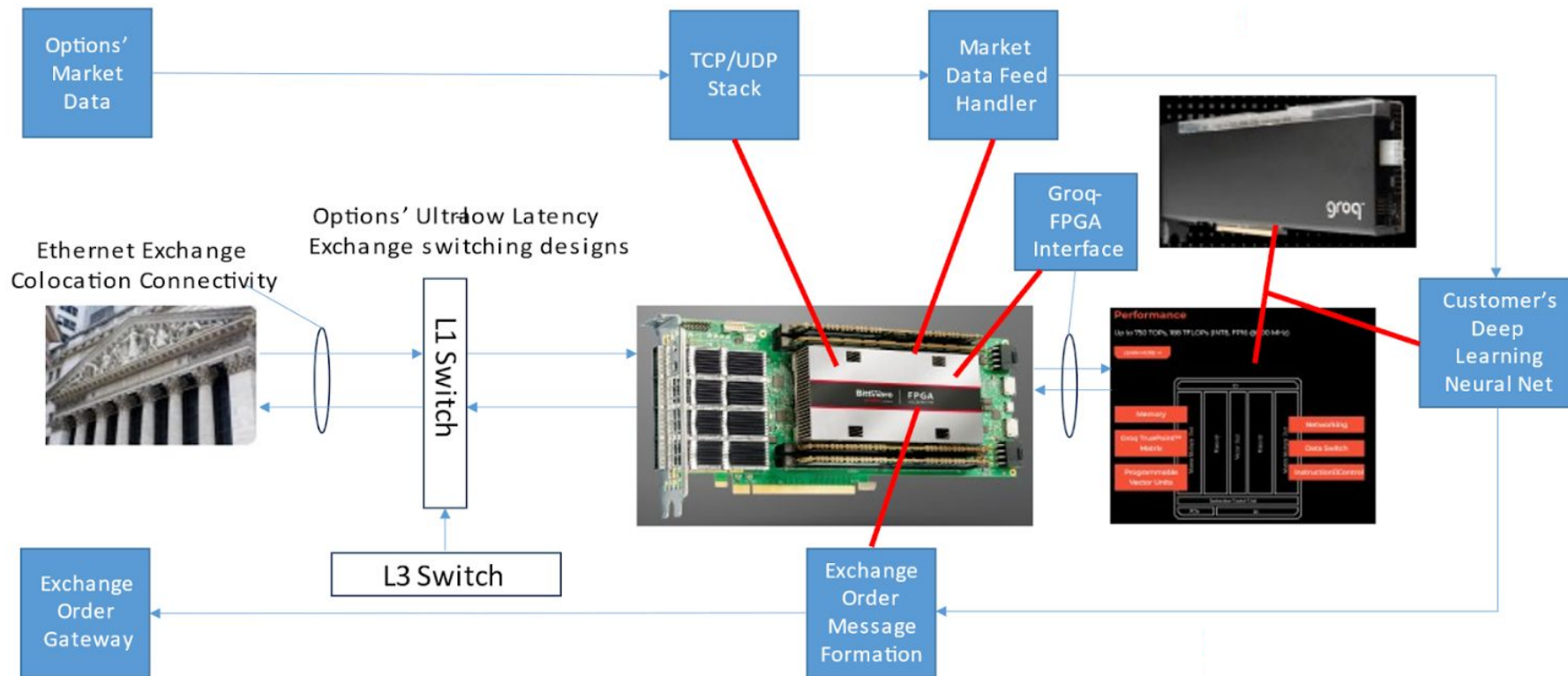


- User Defined Logic
- IP Available from Groq
- Ull Core Logic

GroqChip™ 1 Accelerator



options | Collaboration With Options IT



Thank You

Max Engelen
mengelen@groq.com

Groq's LPUs are set to power AI

**Join us to make AI accessible,
preserve human agency and to
leave your mark.**

Interested ? Lets Talk!

LEARN MORE AT [GROQ.COM](https://groq.com)

