groq™

Running Converged HPC & AI Workloads on the Groq AI Inference Accelerator

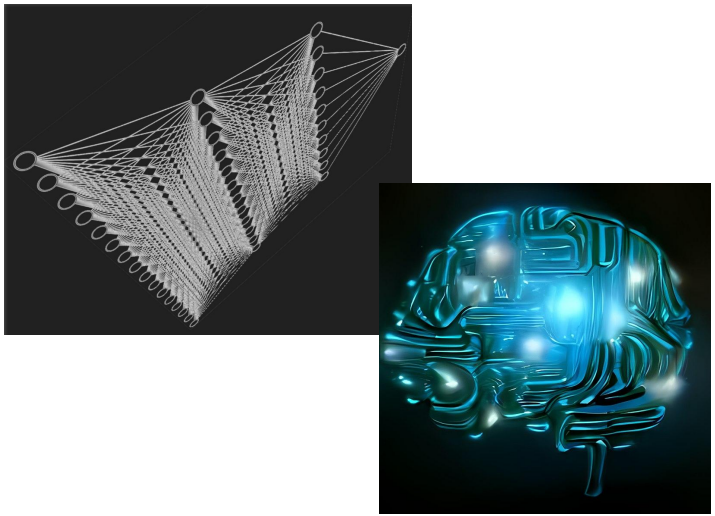
FastML 2023 - Imperial College London

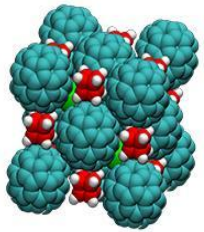
Tobias Becker
tbecker@groq.com

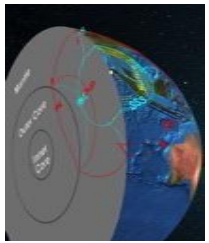
27 Sep 2023

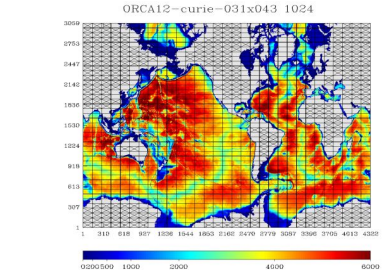
What is converged compute?

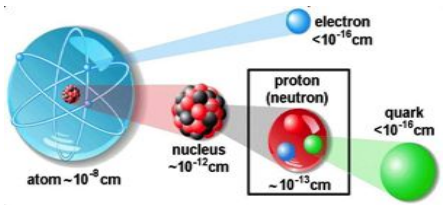
- Infrastructure for combined AI and HPC workloads
- Hybrid applications with combined AI and HPC algorithms







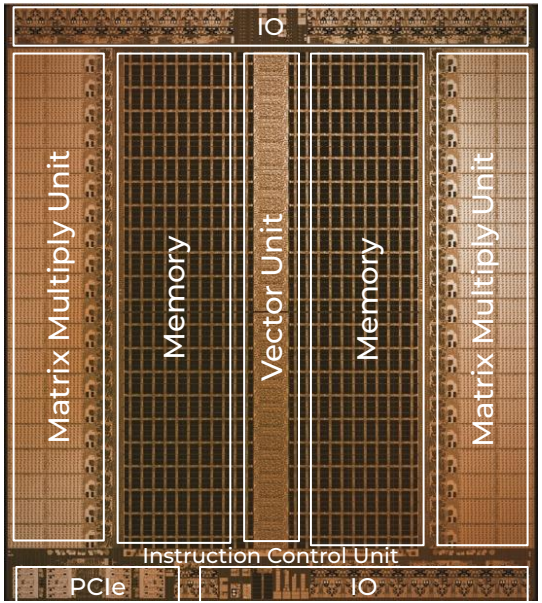




Groq is Radically Simplifying Compute

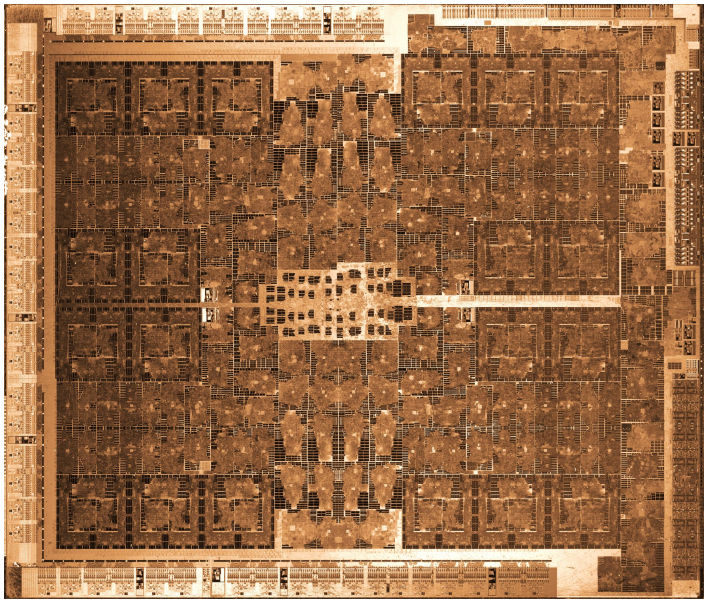
Less hardware control = more compute & memory!

GroqChip™ 1



Simplified design enables **Compute Performance**

Competitive Chip Example



Complex design for processing data results in **Compute Costs**

D. Abts et al. "Think Fast: A Tensor Streaming Processor (TSP) for Accelerating Deep Learning Workloads" ISCA 2020

DEMONSTRATING

Groq Workloads at Scale

A full hierarchy of software & hardware support for scale

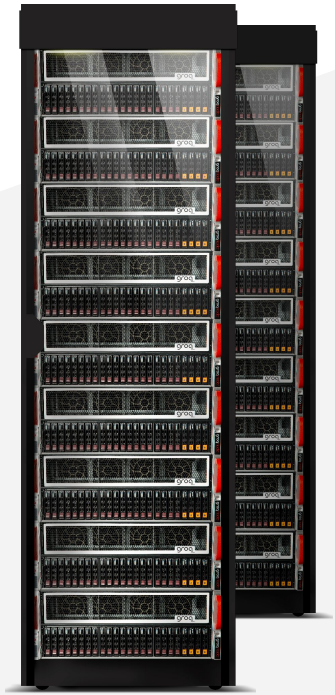
Cards Scale to Nodes

GroqNode contains 8 cards

Nodes Scale to Racks

GroqRack: 8 compute nodes

+ 1 redundant node



GroqChip™ 1 GroqCard™ GroqNode™ GroqRack™ GroqCloud

Converged Use Case: CFD

Computational Fluid Dynamics

CFD benefits a wide range of applications where fluid flow is critical for performance and accuracy for time-market ROI

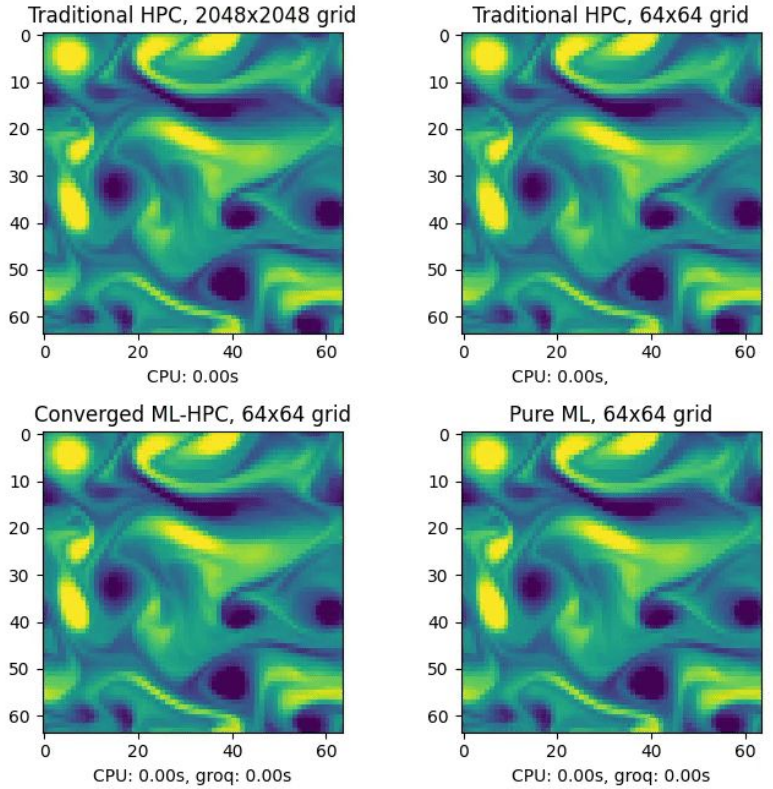
Four approaches using traditional HPC, pure ML, and converged ML-HPC, with varied resolution

Results consider accuracy, throughput, and computation cost

Converged ML-HPC combines high throughput and high accuracy

POTENTIAL APPLICATIONS

Aerospace **Automotive** **Industrial** **Energy** **Medical**

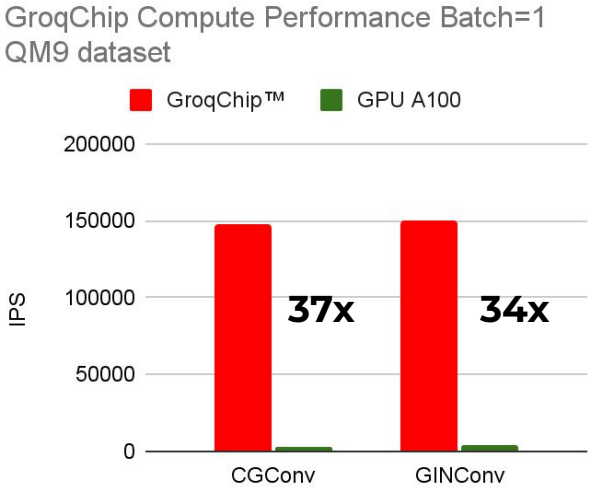


Simulation results and the elapsed time of different solvers.

HydraGNN - Graph Neural Networks

Argonne NATIONAL LABORATORY

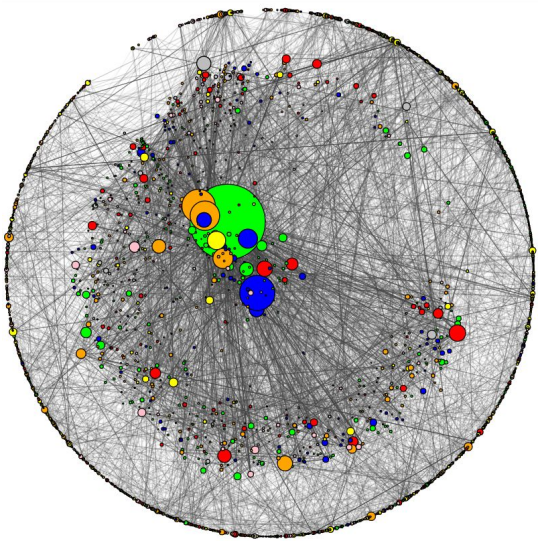
GNNs: Computational chemistry, material science, drug discovery



Performance comparison of CGConv and GINCov graph convolutional layer from PyTorch Geometric (PyG)

Graph Convolution	A100 Speedup (Aug 2022)	A100 Speedup (June 2023)
CGConv_qm9-o1	34x	111x
GIN_qm9-o1	23x	67x
GraphSAGE-_imdb-o1	12x	10x
FiLMConv_qm9-o1	-	20x
PNAConv_MNIST-o1	20x	610x

Evaluation of benchmark results with compiler improvements



Hosseini et al. "Exploring the Use of Dataflow Architectures for Graph Neural Network Workloads" ISC'23

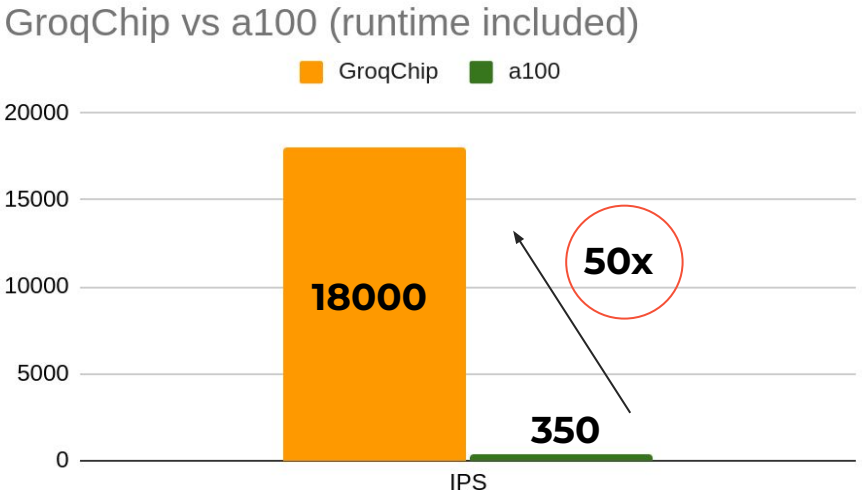
HydraGNN on FePt

Use Case:

- Model predicts total energy, charge density and magnetic moment for each FePt configuration.
- Thus identifying molecules with desired reactivity in a dataset of 10 million molecules.

Need for Scale:

- Needs 10k parallel walks of HydraGNN @ batch 1, that can be parallelized across multiple chips..
- Models currently being trained increase the number of atoms per molecule where Groq can scale to multi-chip execution .

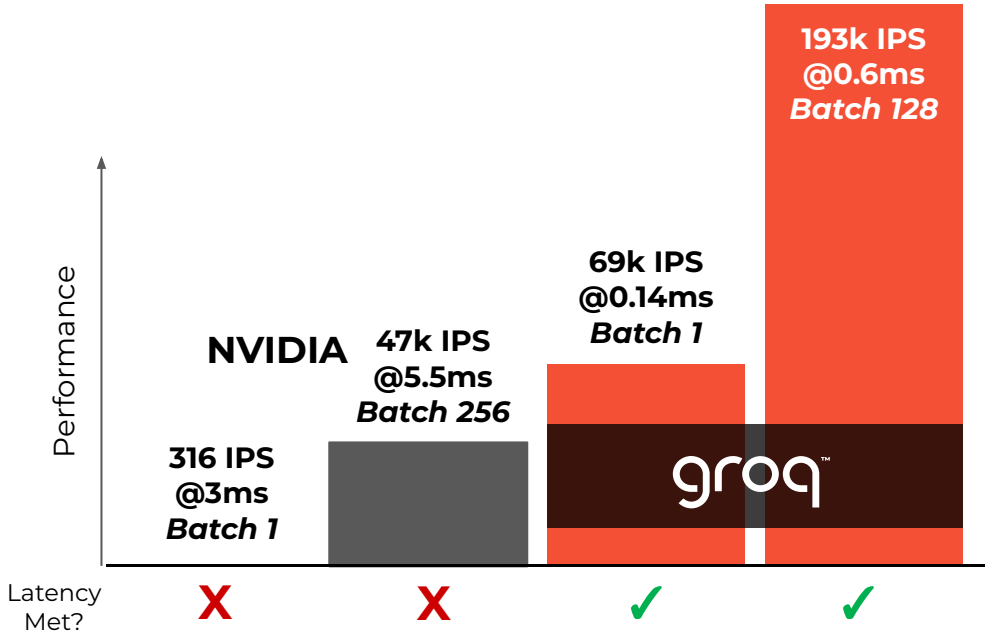


HydraGNN Lsms FePt model

Fusion Reactor Control

Argonne NATIONAL LABORATORY

- Predict, avoid, and mitigate plasma instabilities in Tokamak fusion reactors.
- Goal: highly reliable control loop with <1ms latency.
- Groq offers low batch performance and determinism.



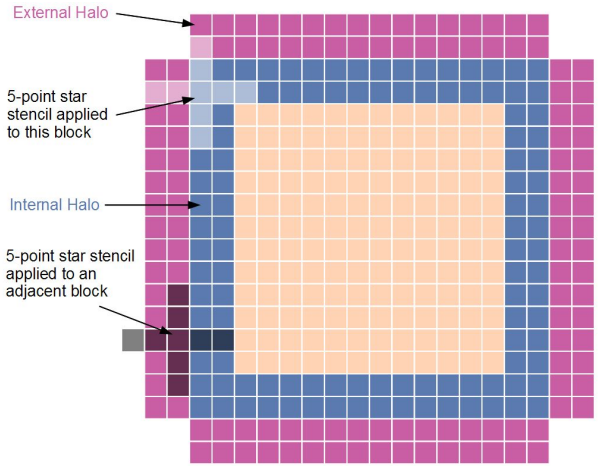
Seismic modelling

- Simulate the propagation of an acoustic wave through the earth by solving the acoustic wave equation

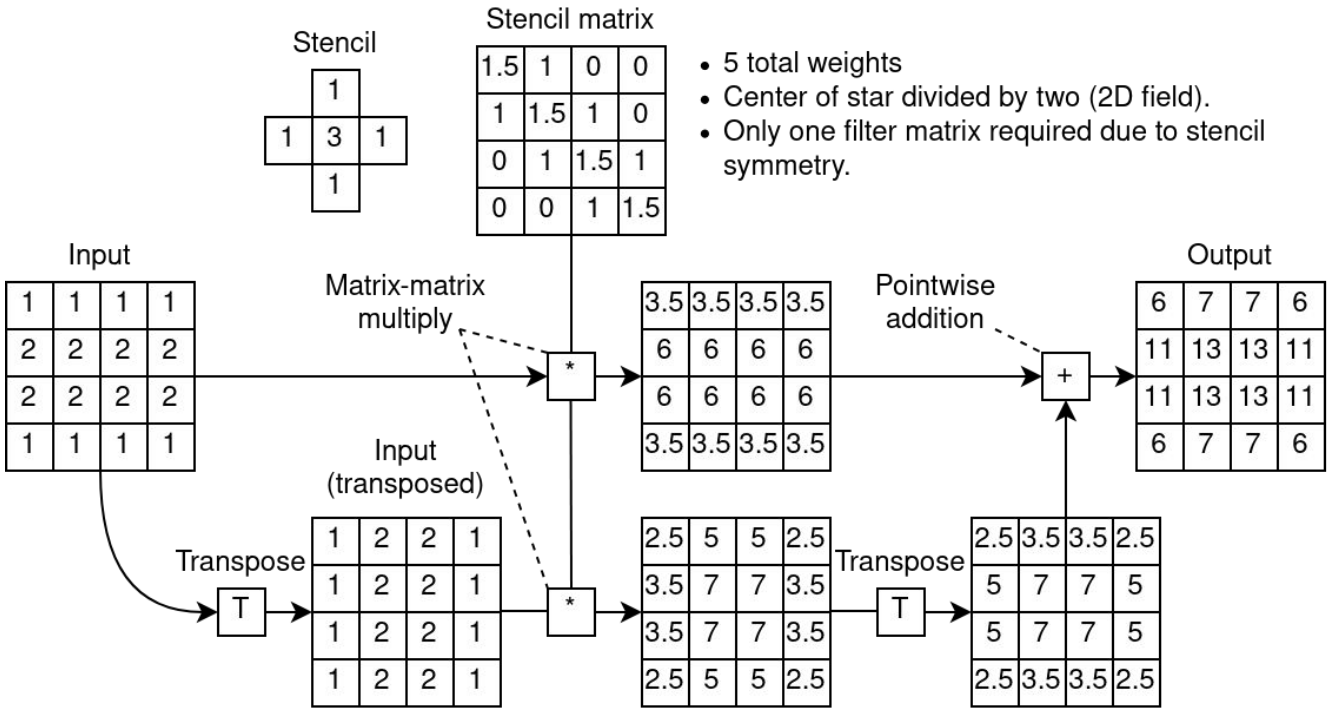
$$\frac{\partial^2 p}{\partial t^2} = v^2 \nabla^2 p + s(t)$$

- Used in Reverse Time Migration (RTM) and Full Waveform Inversion

- Finite difference solver with 3D stencil

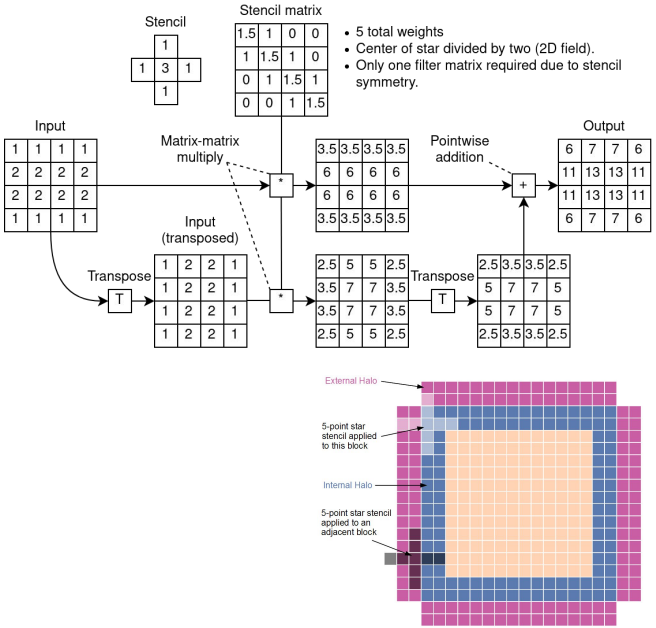


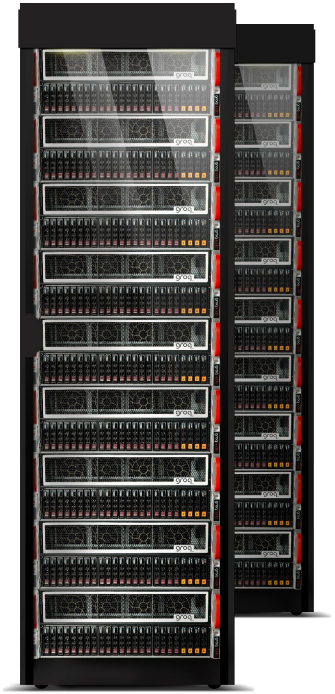
From Stencils to Tensors



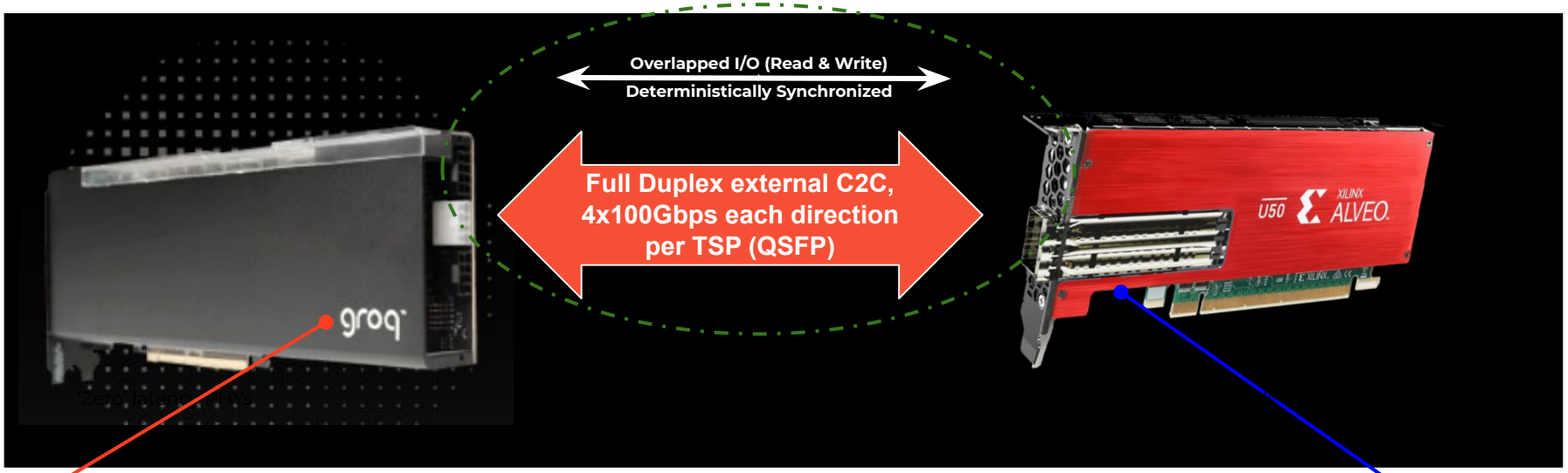
Racklevel scalability

Halo data exchange via Chip2Chip interconnect





Groq IO Accelerator



A very high speed, deterministic processor for:

- Real-time inference
- AI algorithms & compute intensive offload

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

- Low latency data IO via Ethernet
- Application specific interfacing
- Data preprocessing / conversion
- Memory expansion

groq™