groq™

Exploring Converged HPC & AI on Dataflow Architectures

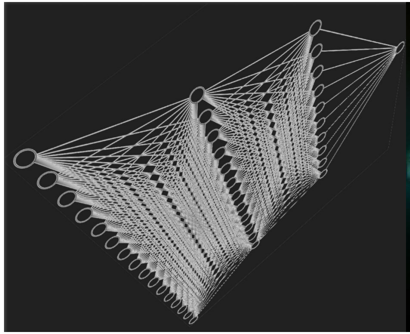
WRC - HiPEAC 2024

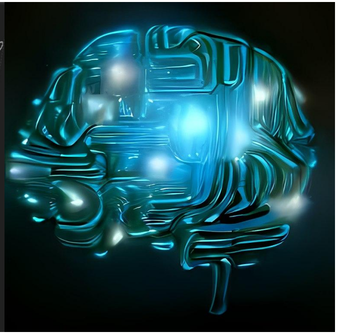
Tobias Becker
tbecker@groq.com

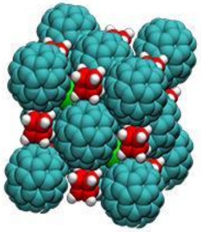
17 Jan 2024

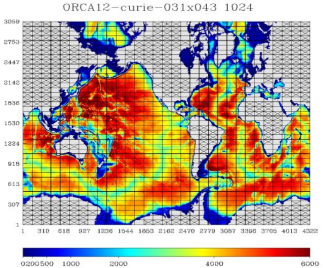
What Is Converged Compute?

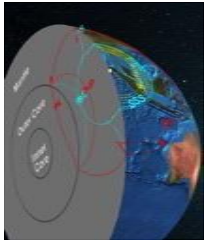
- Hybrid infrastructure and applications for combined AI and HPC
- Specialisation vs Generalisation

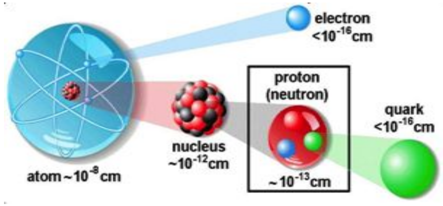




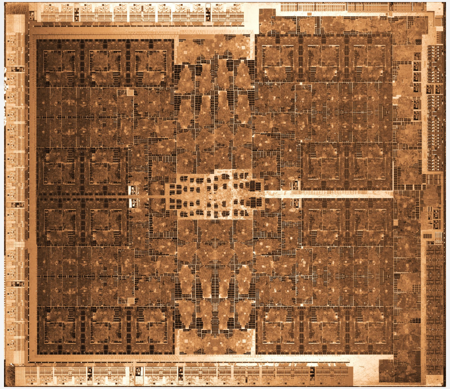


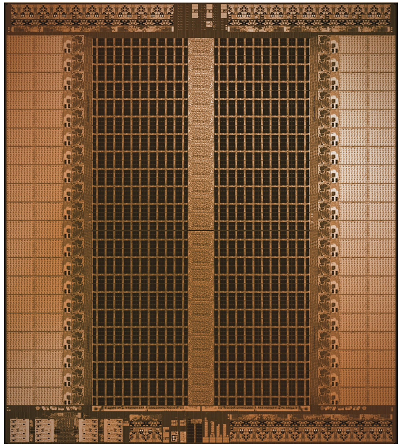






Groq Simplifies Compute





Typical GPU

Graphic Processor

COMPLEX

Difficult programming

Less responsiveness

Non-deterministic execution

Higher costs

GroqChip™ 1

First LPU™ Accelerator

SIMPLIFIED

GroqChip™ 1 Overview

Scalable compute architecture

SRAM Memory

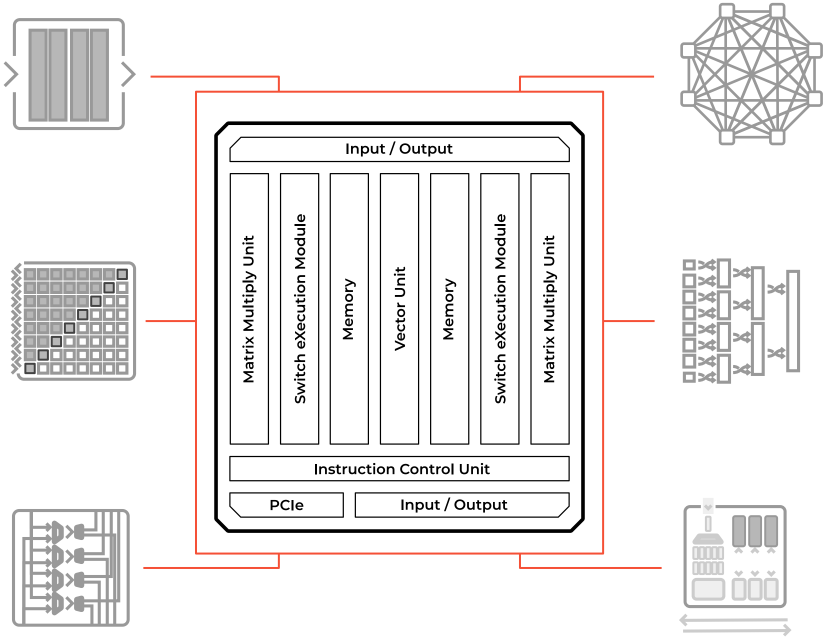
Massive concurrency
80 TB/s of BW
230MB capacity
Stride insensitive

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

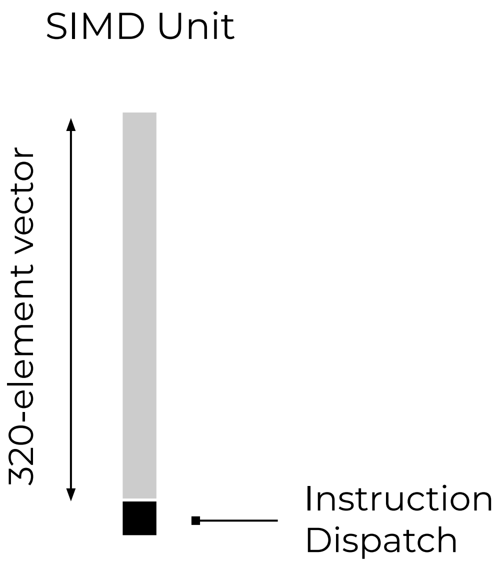
Data Switch

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

Instruction Control

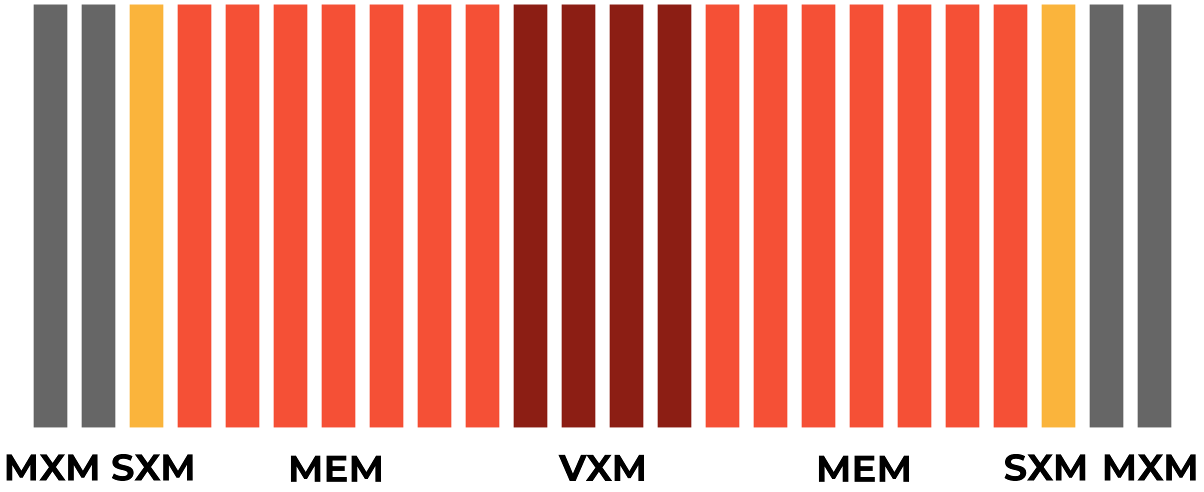
Multiple instruction queues for instruction parallelism

GroqChip™ Building Blocks



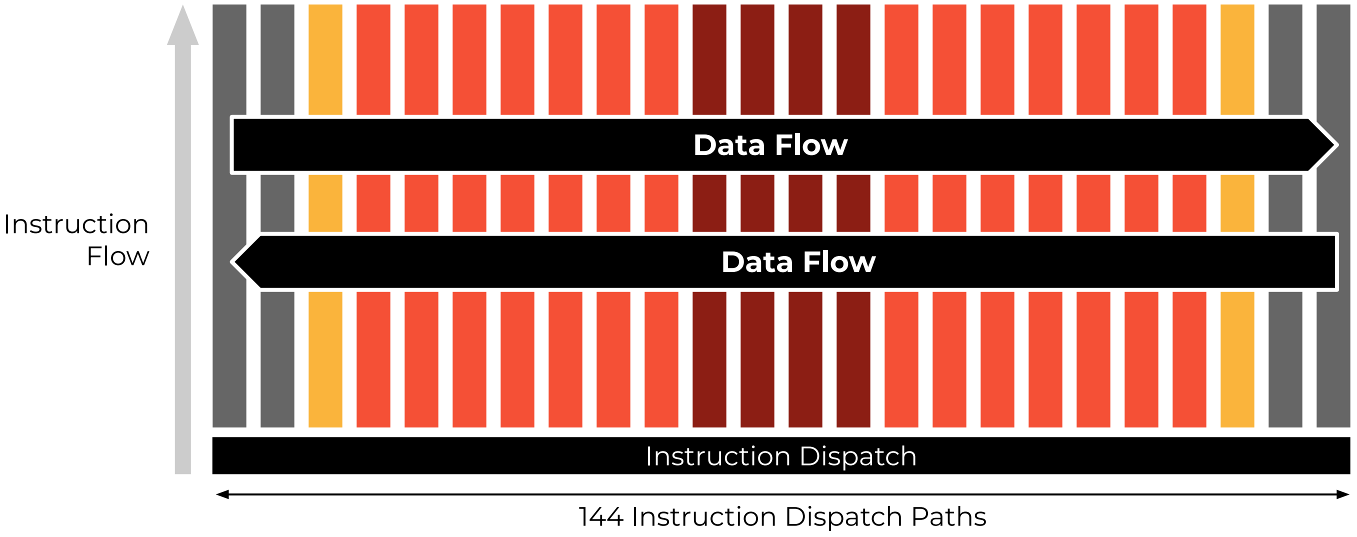
GroqChip™ Building Blocks

Lay out SIMD units across chip area



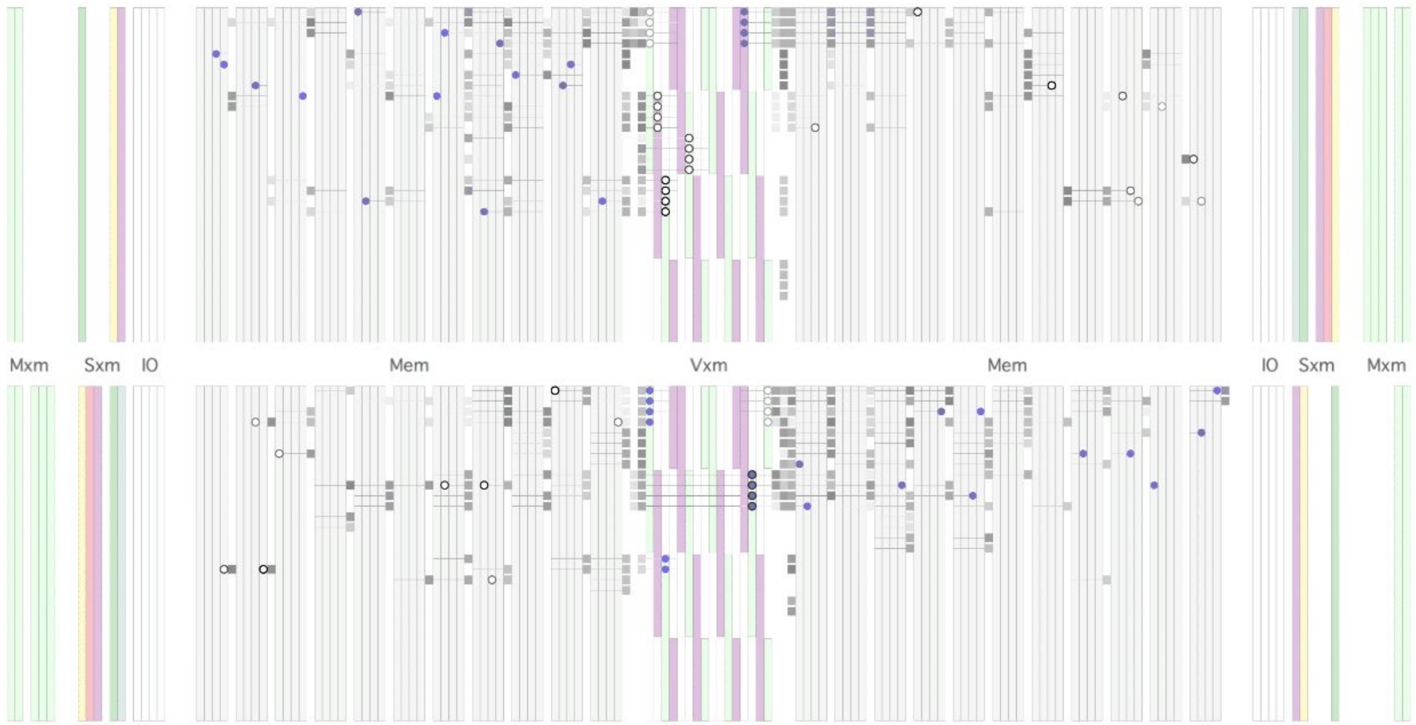
GroqChip™ Building Blocks

High-bandwidth "Stream Registers" for passing data between units



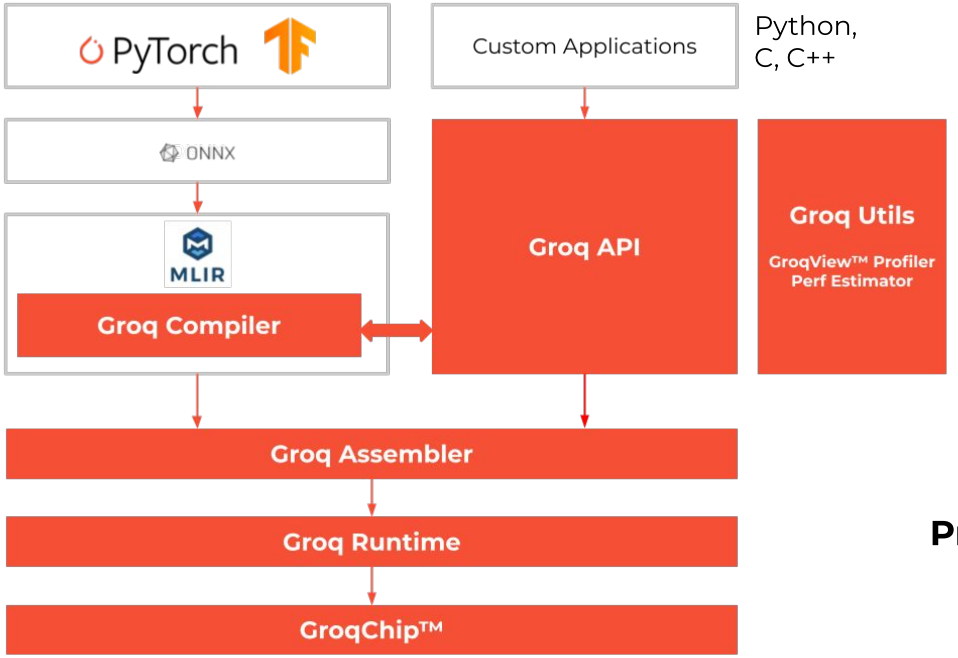
Visualizing Data Orchestration

Given to Groq™ Compiler



GroqWare™ Suite At-a-glance

Accelerating ML & HPC developer velocity



A Diverse Suite of Development Tools

Out-of-Box

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

Fine Grained Control

Groq API provides finer grained control of GroqChip in order to support custom applications

Productivity Tools

GroqFlow™ Toolchain automatically runs PyTorch models with just one line of code

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

Performance Estimator provides accurate predictions even without access to hardware

Groq Workloads at Scale

TSP Architecture
Provides Near-linear Scaling Performance

GroqChip™ 1

Cards Scale to Nodes
GroqNode contains 8 cards

GroqCard™

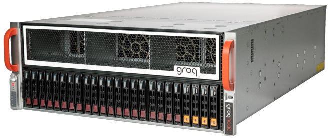
Cluster Ready
8 x RealScale external ports

2U Server with 4x GroqCard

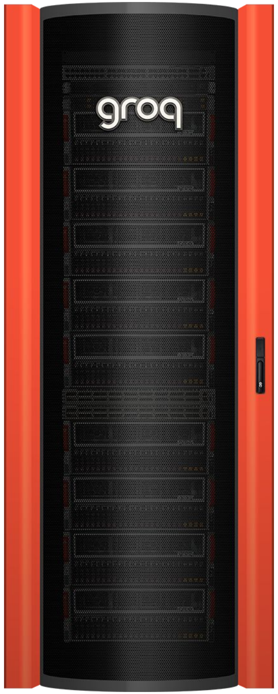
Dell R750XA

Nodes Scale to Racks
GroqRack: 8 compute nodes

+ 1 redundant node



GroqNode™



GroqRack™

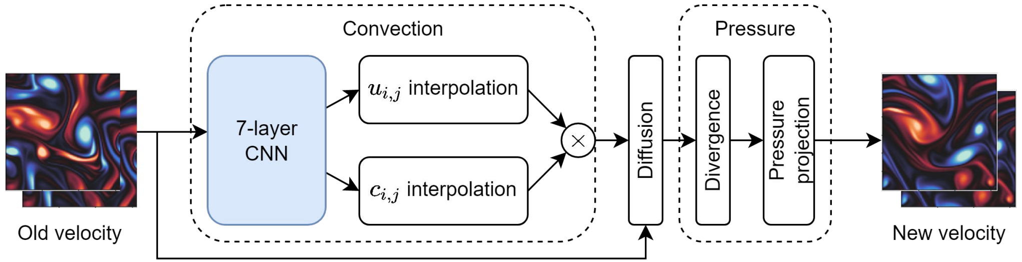
groq™

Converged Compute: CFD

Conventional and AI based solvers for structured grid methods

Solver Summary

- 2D structured grid
- Incompressible airflow
- Explicit time integration
- Framework in JAX-CFD



Pure DNS based on incompressible Navier-Stokes equations

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\,\mathbf{u} = -\frac{1}{\rho}\nabla p + \nu \nabla^2 \mathbf{u}$$

$$\nabla \cdot \mathbf{u} = 0$$

Convection Diffusion Pressure

D. Kochkov et al. "Machine learning accelerated computational fluid dynamics" PNAS 2021

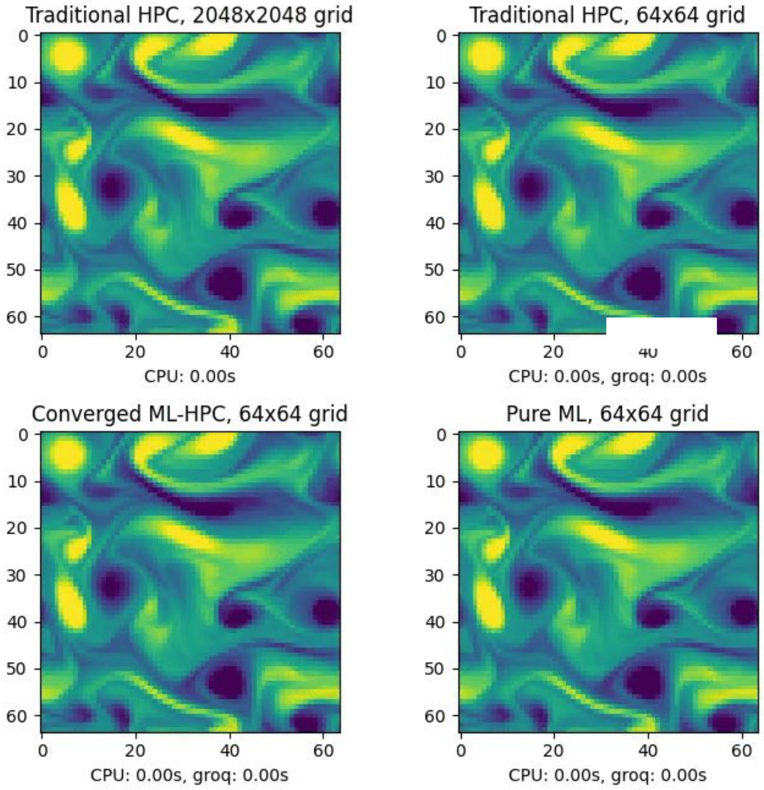
Direct numerical simulation (DNS) can be replaced or augmented with AI

Hybrid CFD with AI Augmentation

Compare traditional and AI based approaches

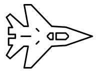
Four Approaches:

- Traditional DNS: standard solver based on pressure projection (high and low res)
- Learned correction: Small grid DNS with CNN-based correction
- Pure ML: LSTM-based encoder-process-decoder
- Converged ML-HPC combines high throughput and high accuracy



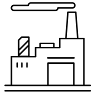
Simulation results and the elapsed time of different solvers.

POTENTIAL APPLICATIONS

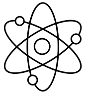


Aerospace

Automotive



Industrial

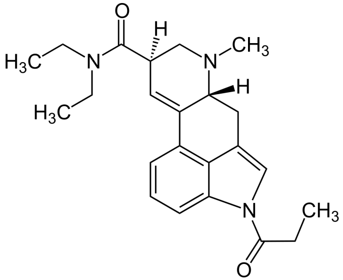


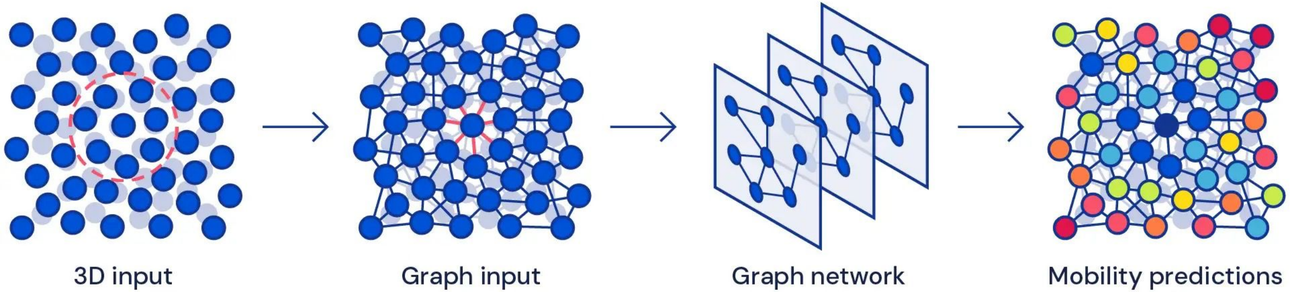
Energy

Medical

Graph Neural Networks (GNNs)

- Generalization of common deep neural network (DNN) architectures to non-euclidean data
- Consider graph representation of a problem:
    - Molecules in computational chemistry
    - Recommendation systems for social media
- Computational chemistry use case: Replace conventional DFT based algorithms





HydraGNN on Iron-Platinum (FePt)

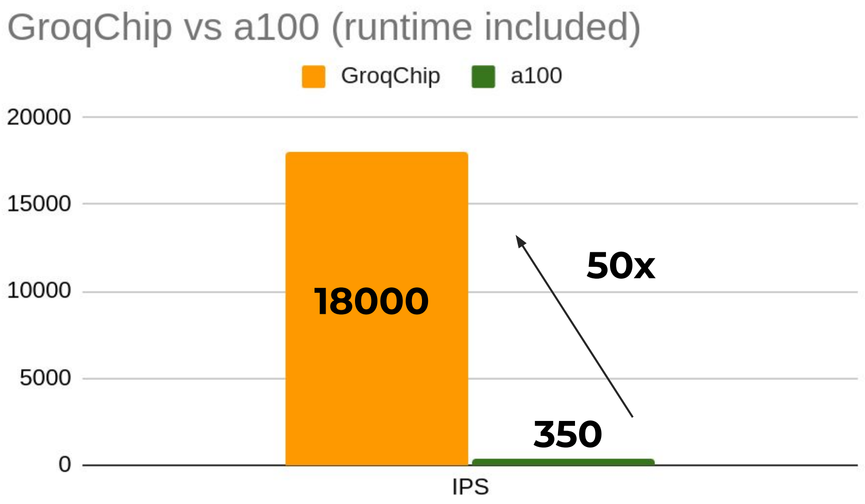
End-to-end GNN with end-to-end benchmark including runtime

Use Case:

- Model predicts total energy, charge density and magnetic moment (multiple predictions, i.e hydra model)for each FePt configuration
- This allows us to identify molecules with desired reactivity in a dataset of 10 million molecules

Need for Scale:

- Production needs 10k parallel walks of HydraGNN @ batch 1
- Can be parallelized across an entire GroqRack
- Models currently being trained at ORNL increases the number of atoms per molecule where Groq can scale to multi-chip execution



HydraGNN Lsms FePt model (M Lupo Pasini et al 2022.)

Multi-task graph neural networks for simultaneous prediction of global and atomic properties in ferromagnetic systems

Chemprop: Messaging Passing GNN

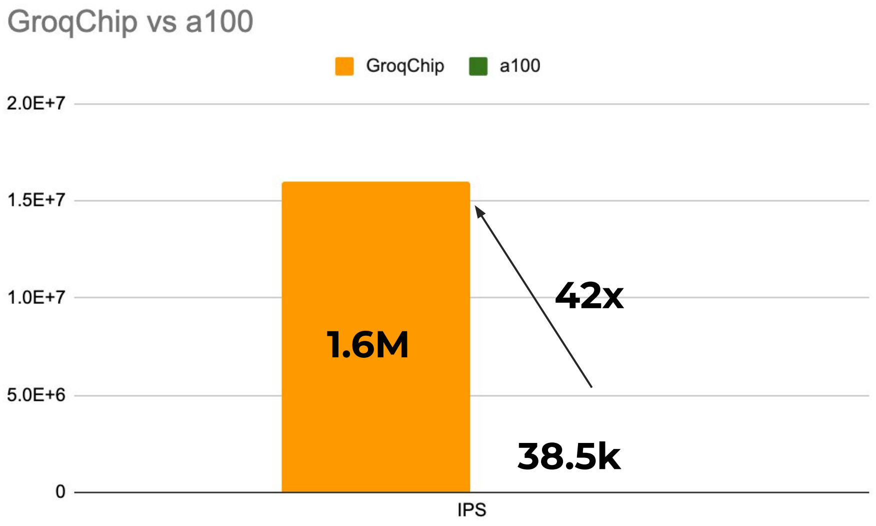
Machine Learning Package for Chemical Property Prediction

Use case:

- ChemProp is a message passing neural networks for molecular property prediction capabilities across a range of properties.
- Specifically tested for drug discovery with smile string inputs.

Scalability:

- Production configuration involves processing **4 Billion Compounds**.
- A **42x** speed up drastically improves the speed of iteration with less hardware.



ChemProp model repo

ISC 2023 Workshop Paper (May '23)

Exploring the Use of Dataflow Architectures for Graph Neural Network Workloads
(Hosseini et al.) In collaboration with Argonne National Laboratory and Sambanova.

Results

- In August 2022 during the paper write up, GroqChip™ achieved up to 37x speed-up for GNN convolution layers (CGConv, GINConv etc.)
- In the previous year, Groq™ Compiler optimizations delivered an additional speed-up of up to 50x for these GNN convolution layers*
- This speed-up is achieved as a result of the dataflow paradigm and speed-up of up to 10x on operator microbenchmarks, which frequently appear in GNN architectures. This provides an additional speed up to converged HPC workloads on non Euclidean data.

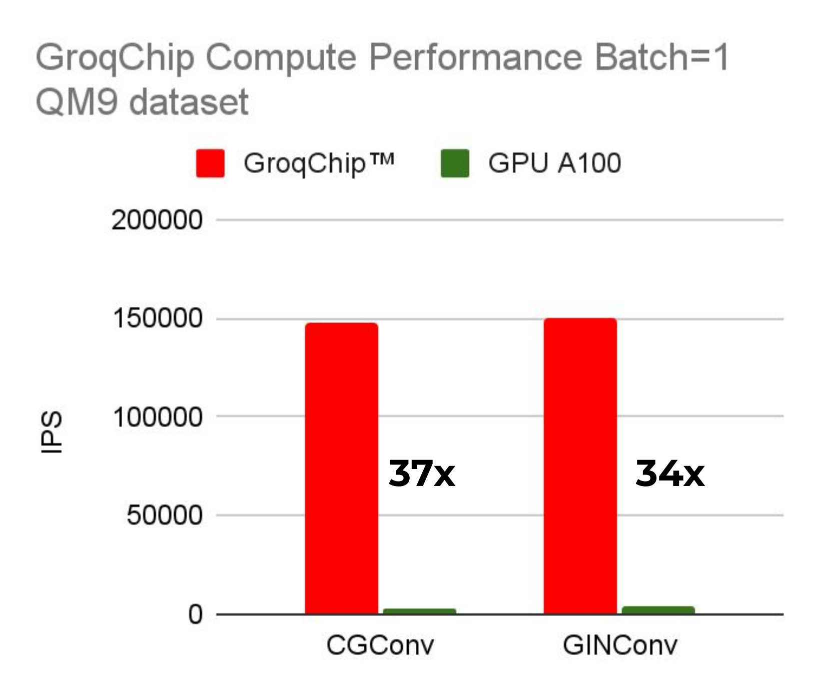


Figure 1 (August 2022): Performance comparison of GroqChip vs GPU A100 on the CGConv and GINCov graph convolutional layer from PyTorch Geometric (PyG)

*Groq Compiler is continually improving and has delivered 35% speed-up (April 2023) since the paper write-up (August 2022)

Fusion Reactor Control

Smart Power Grids

“Mission Impossible”

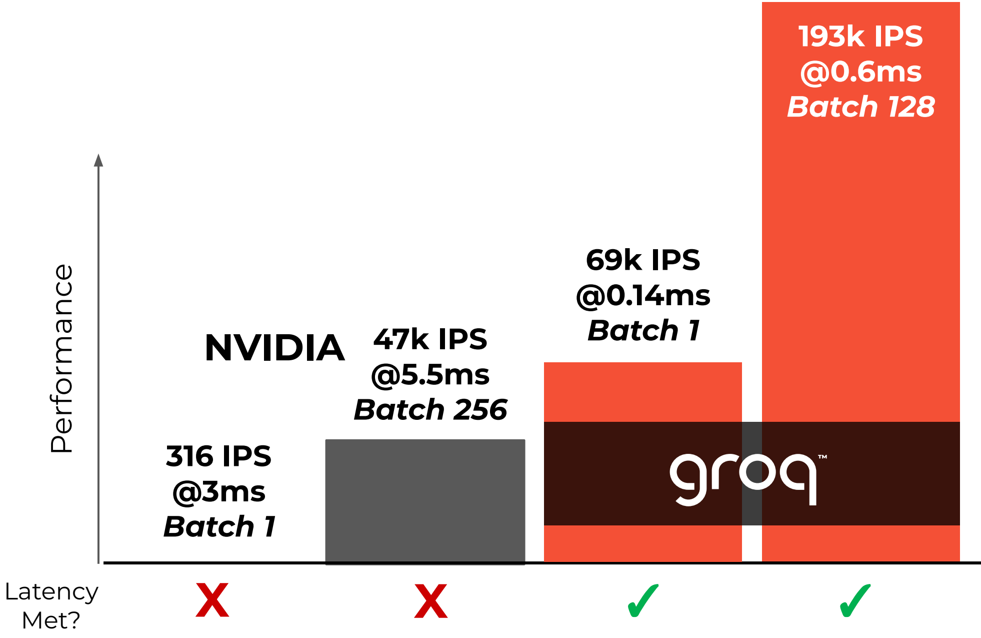
Forecasting plasma instabilities in Tokamak fusion reactor simulation

Maximize performance of LSTM model within 1ms hard requirement

Groq Advantage

Deterministic AI processor delivers ultra-low latency

Enables highly reliable real-time control



Groq architecture delivers deterministic ML at ultra-low latency, with **5–20x performance that meets 1ms response window**[1].

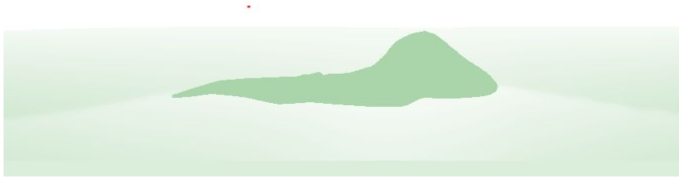
Seismic Modelling

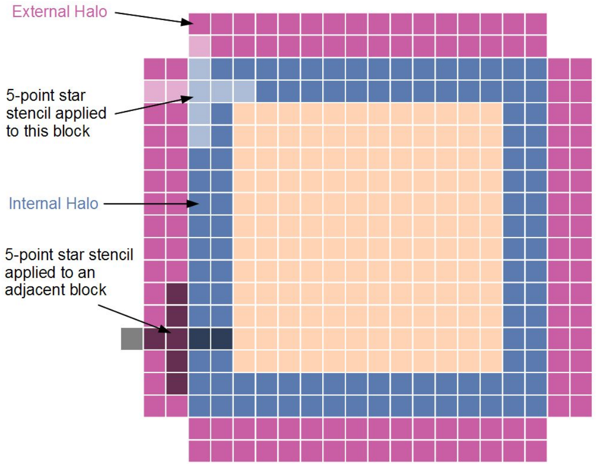
- Simulate the propagation of an acoustic wave through earth / water by solving the acoustic wave equation:

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

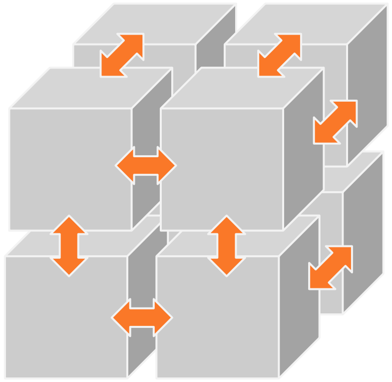
- Used in Reverse Time Migration (RTM) and Full Waveform Inversion

- Finite difference solver with 3D stencil





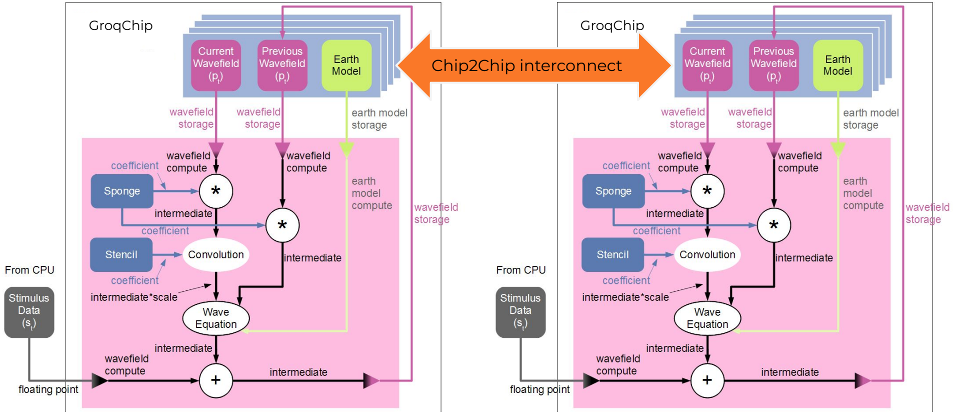
Scaling seismic: Multi Chip



One chip supports up to 128^3 domain size

Larger Domains:

- Split into subcubes
- Requires halo data exchange at the edge
- Use Groq RealScale Chip2Chip interconnect to avoid PCIe bottlenecks
- Single-chip (1) performance for 128^3: 10 Gpt/s
- Multi-chip (64) performance for 512^3: 400 Gpt/s



Groq IO Accelerator

Overlapped I/O (Read & Write)
Deterministically Synchronized

Full Duplex external C2C, 4x100Gbps each direction per TSP (QSFP)

INTERFACE CODE

A very high speed, deterministic processor for:

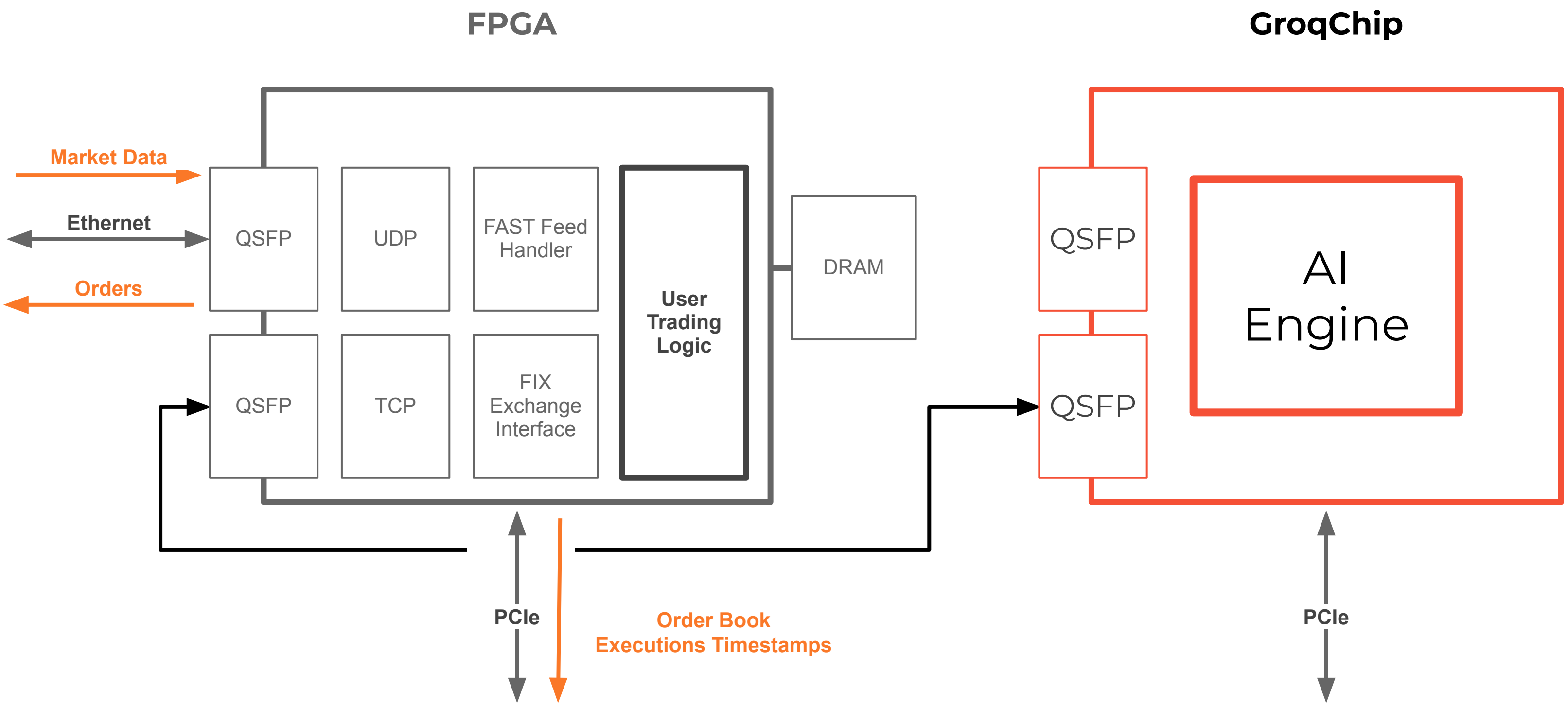
- Real-time series processing
- AI algorithms & compute intensive offload

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

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

Ultra-Low Latency Trading with FPGA + GroqChip™

High-powered inference engine



Summary

- The current landscape of computing is dominated by CPUs and GPUs
- In order to compete you need to be sufficiently different
  - GroqChip: SRAM & C2C
  - FPGAs: fine-grain reconfigurable
- Applications evolve to leverage new hardware
  - New: AI, LLM
  - Old: HPC
- Specialised hardware evolves to enable more applications
- Hybrid architectures, interconnect & dataflow
- Programming is key

Tobias Becker
tbecker@groq.com

LEARN MORE AT GROQ.COM

groq™