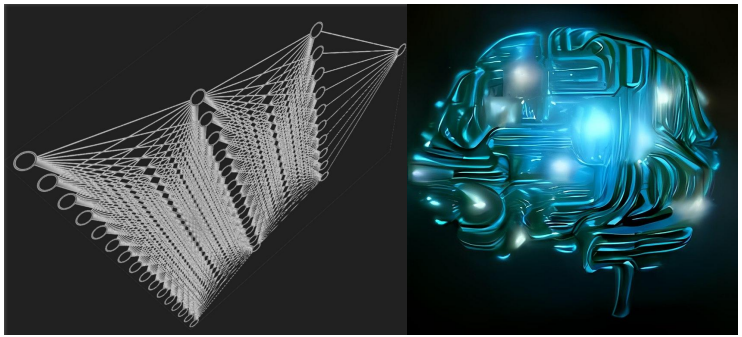


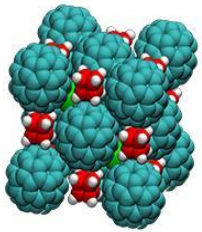
EXPLORING CONVERGED HPC & AI WORKLOADS ON
The Groq AI Inference Accelerator

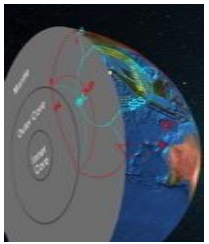
Tobias Becker
November 2023

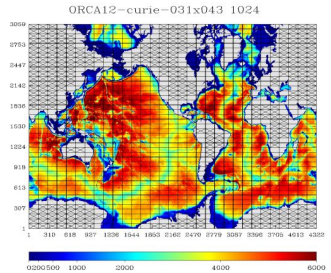
What Is Converged Compute?

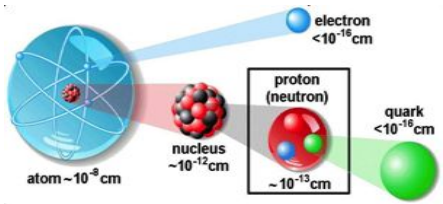
- Infrastructure for combined AI and HPC workloads
- Hybrid applications with combined AI and HPC algorithms





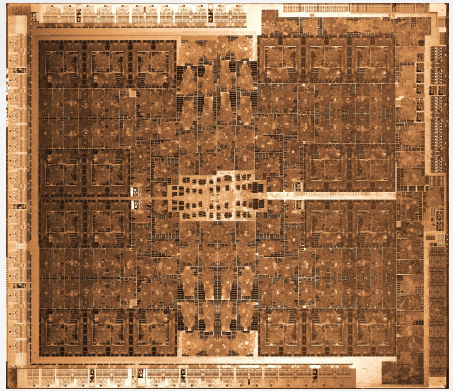


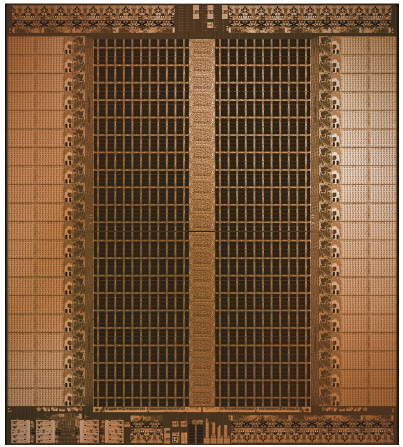




© Groq, Inc.

Groq Simplifies Compute





Typical GPU
Graphic Processor

COMPLEX
Difficult programming
Less responsiveness
Non-deterministic execution
Higher costs

GroqChip™ 1
First LPU™ Accelerator

SIMPLIFIED

© Groq, Inc.

GroqChip™ 1 Overview

Scalable compute architecture

SRAM Memory

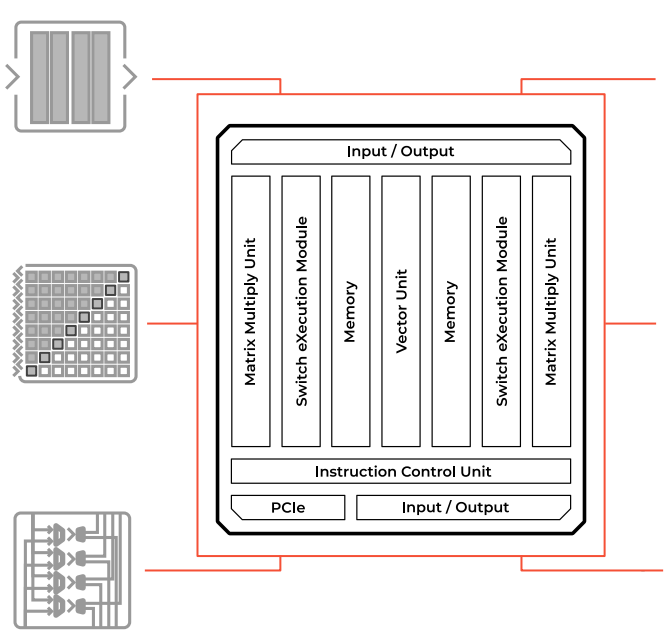
Massive concurrency
80 TB/s of BW
230MB capacity
Stride insensitive

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

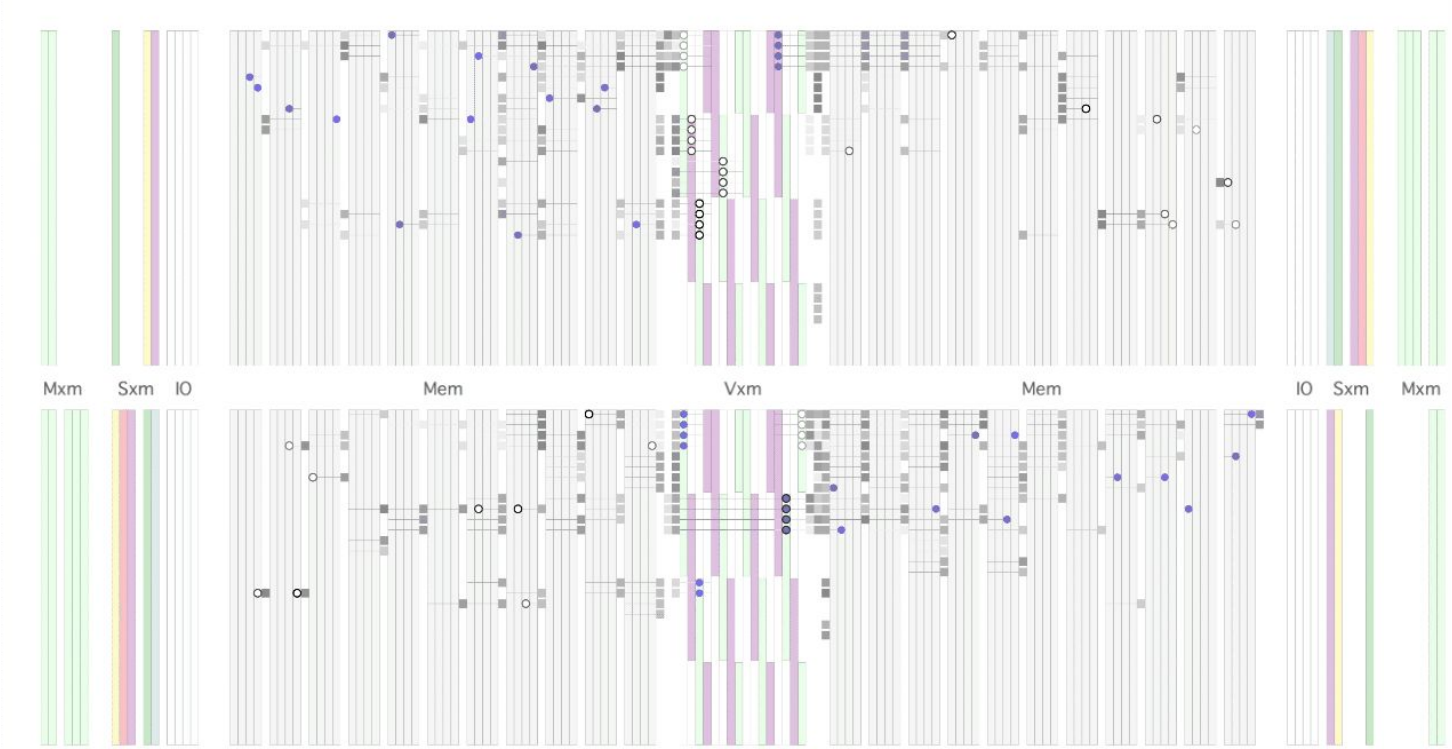
Instruction Control

Multiple instruction queues for instruction parallelism

© Groq, Inc.

Visualizing Data Orchestration

Given to Groq™ Compiler



© Groq, Inc.

Groq Workloads at Scale

TSP Architecture
Provides Near-linear Scaling Performance

Cards Scale to Nodes
GroqNode contains 8 cards

Cluster Ready
8 x RealScale external ports

2U Server with 4x GroqCard

Nodes Scale to Racks
GroqRack: 8 compute nodes

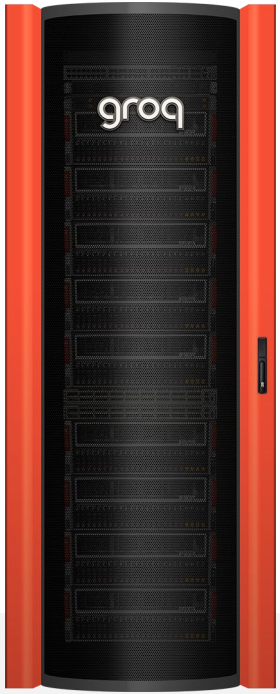
+ 1 redundant node











GroqChip™ 1

Synchronous Tensor Streaming Processor architecture

RealScale™ enabled

GroqCard™

Up to 750 TOPs, 188 TFLOPs (INT8, FP16 @ 900MHz), 240W*

Dell R750XA

Up to 3 POPs, 752 TFLOPs

(INT8, FP16)

GroqNode™

Up to 6 POPs, 1.5 PFLOPs (INT8, FP16)

GroqRack™

Up to 48 POPs, 12 PFLOPs (INT8, FP16)

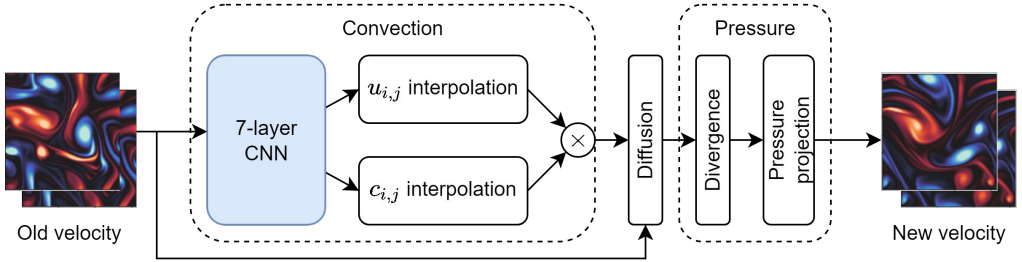
© Groq, Inc.

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

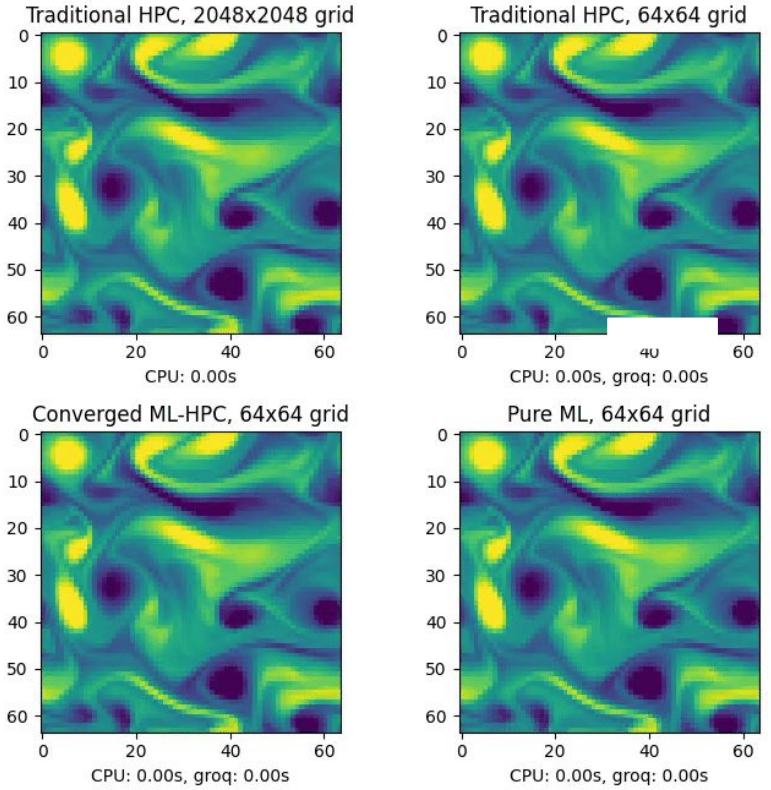
groq™ © Groq, Inc.

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.

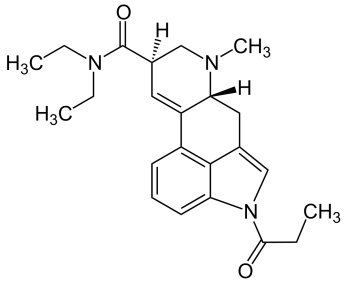
POTENTIAL APPLICATIONS

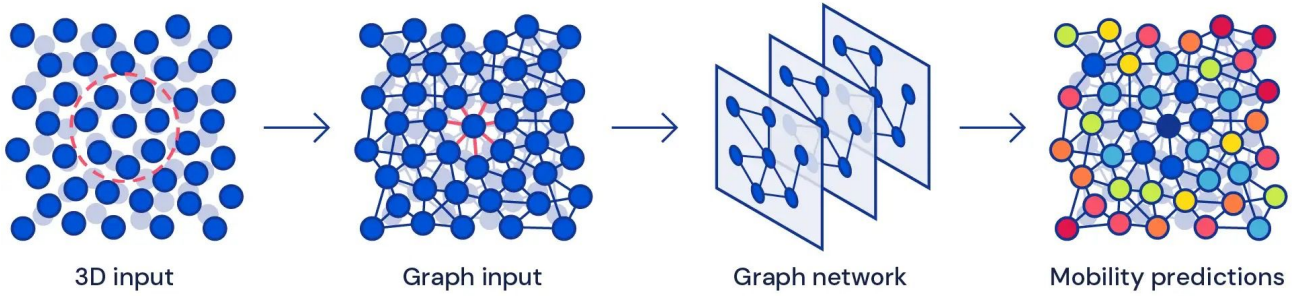
Aerospace **Automotive** **Industrial** **Energy** **Medical**

© Groq, Inc.

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





© Groq, Inc.

HydraGNN on Iron-Platinum (FePt)

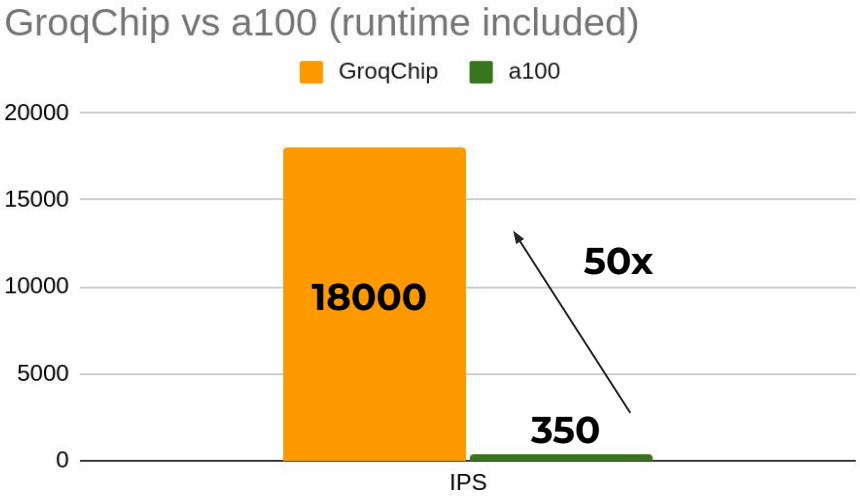
End-to-end GNN with end-to-end benchmark including runtime

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

Iron-Platinum FePt Dataset

Static dataset with quantum chemical properties of 32, 000 configurations of FePt

- The FePt dataset consists of 32, 000 configurations of the FePt binary allow with body-centered cubic (BCC) structure.
- The configurations span all compositions from 0% Fe - 100% Pt through to 100% Fe - 0% Pt
- Total energy, charge density and magnetic moment are for each configuration is predicted by running ab-initio DFT with LSMS-3 on the OLCF Titan.
- This is a static problem with fixed 32 fixed nodes and 206 edge connections per molecule.

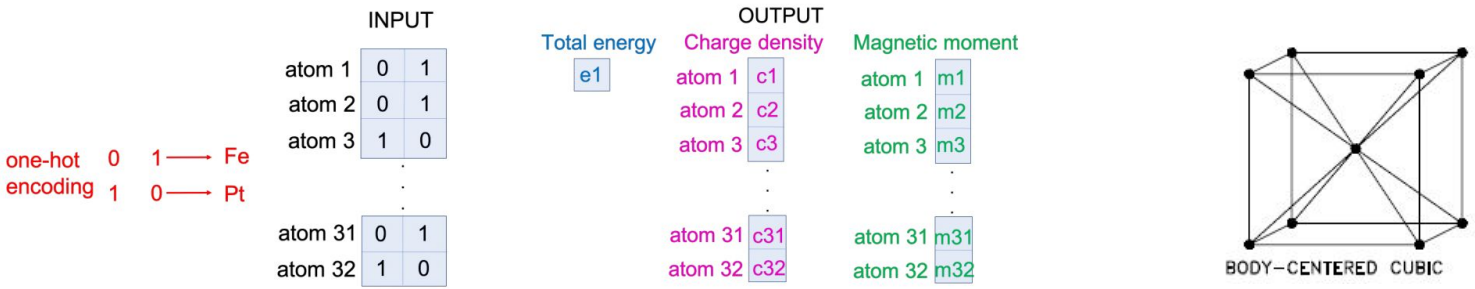


Figure 2. Input data (left) and output data (right) formats for each FePt configuration.

© Groq, Inc.

Chemprop: Messaging Passing GNN

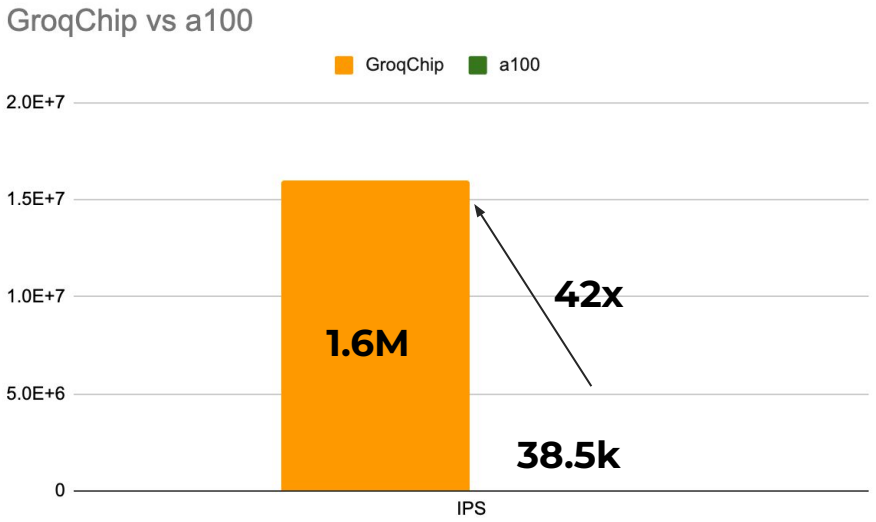
Machine Learning Package for Chemical Property Prediction

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

© Groq, Inc.

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.

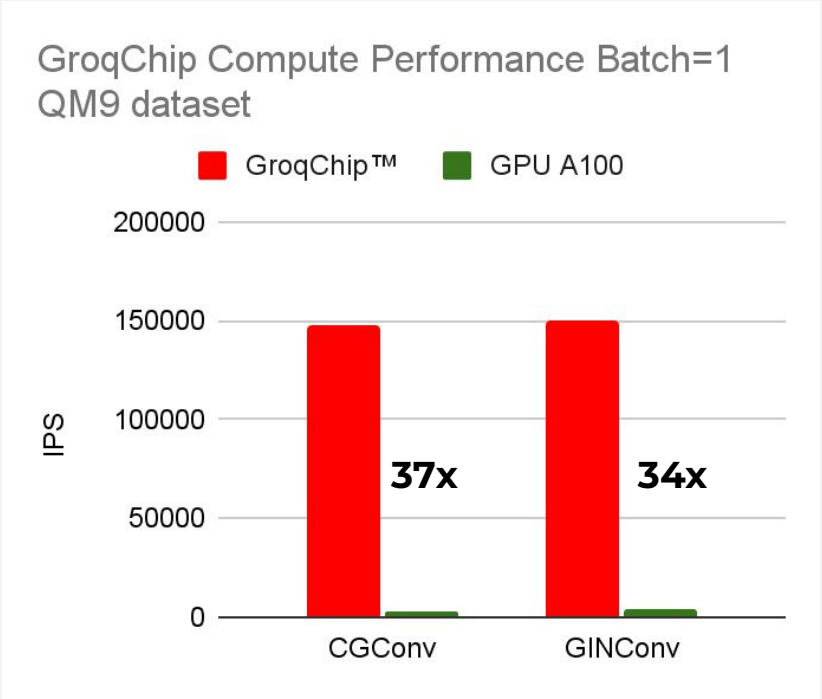


Figure 1 (August 2022): Performance comparison of GroqChip vs GPU A100 on the CGConv and GINCov graph convolutional layer from PyTorch Geometric (PyG)

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

© Groq, Inc.

Fusion Reactor Control

Smart Power Grids

“Mission Impossible”

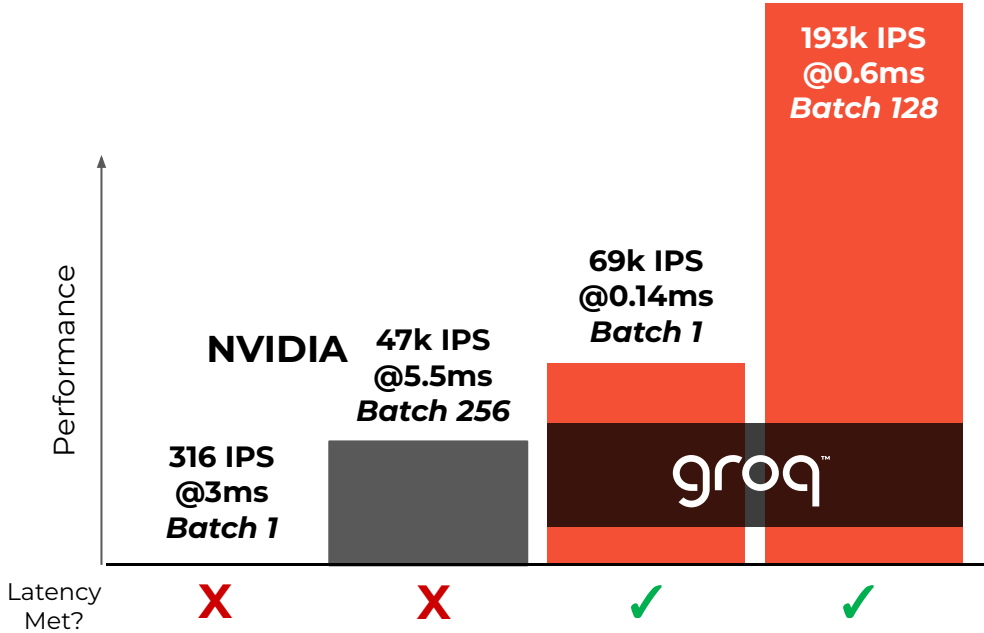
Forecasting plasma instabilities in Tokamak fusion reactor simulation

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].

© Groq, Inc.

Summary

Scale

Purpose-built dataflow architecture for inference at scale

Performance

at batch 1 and highly scalable

Simplicity

through Groq Compiler

Converged Application Example

CFD

GNN

Tokamak

© Groq, Inc.

groq™

Tobias Becker
tbecker@groq.com

LEARN MORE AT GROQ.COM







groq™ © Groq, Inc.

Groq Proprietary

groq™

