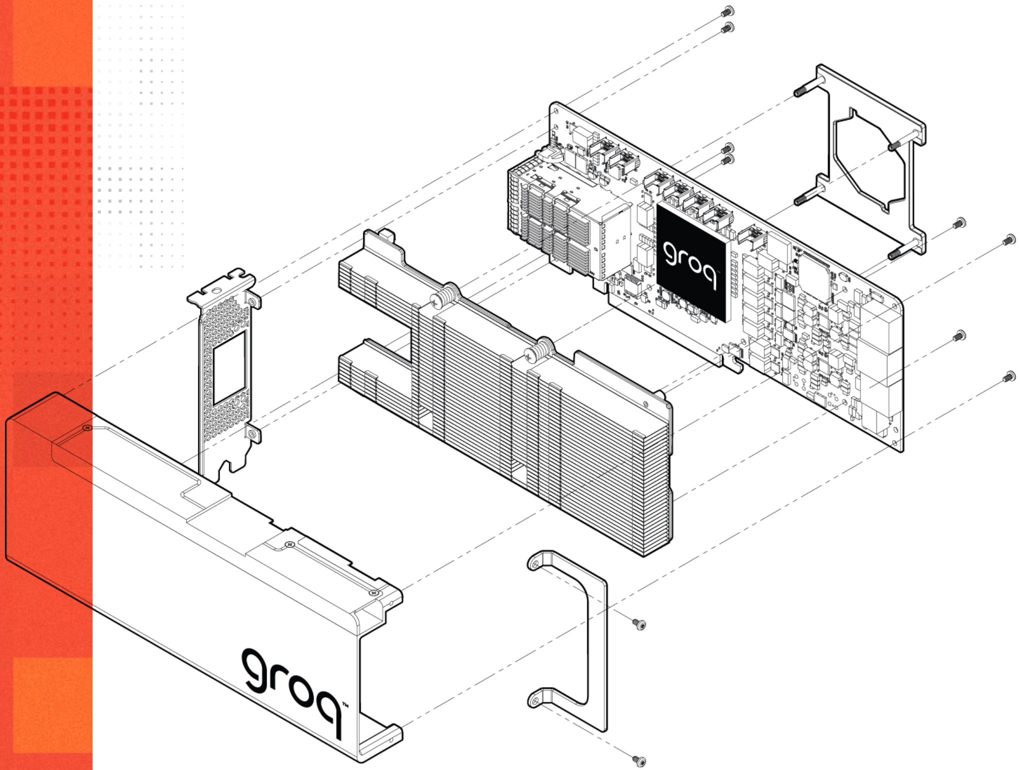
MAXELER
a groq company

Running Scalable Applications on the Groq AI & HPC Platform

Gary Robinson



ABOUT

Groq + Maxeler

DATAFLOW

Compute on data **while it moves**

Dataflow

The data processing factory

Much like the advent of Ford Motor Company's moving assembly line—Maxeler achieves massive scale through computation on deep pipelines

Highly efficient

High throughput

Predictable

No dynamic control or synchronisation issues

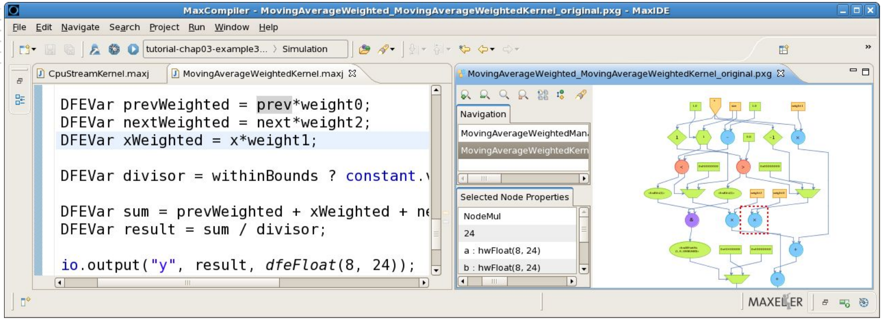
MAXELER

DATAFLOW COMPUTING ON

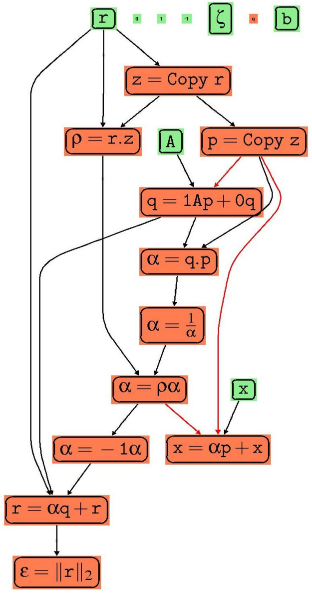
FPGAs with MaxCompiler

Maxeler tools for FPGA acceleration projects

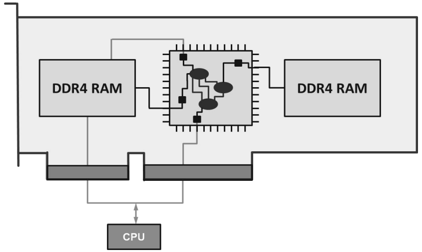
Maxeler MaxCompiler

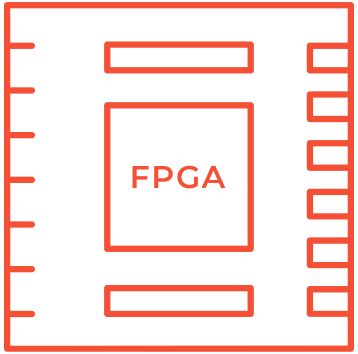


Dataflow Model



FPGA Card





Fine-grained,
programmable logic

Massive scale matrix and vector
operations in a dataflow architecture

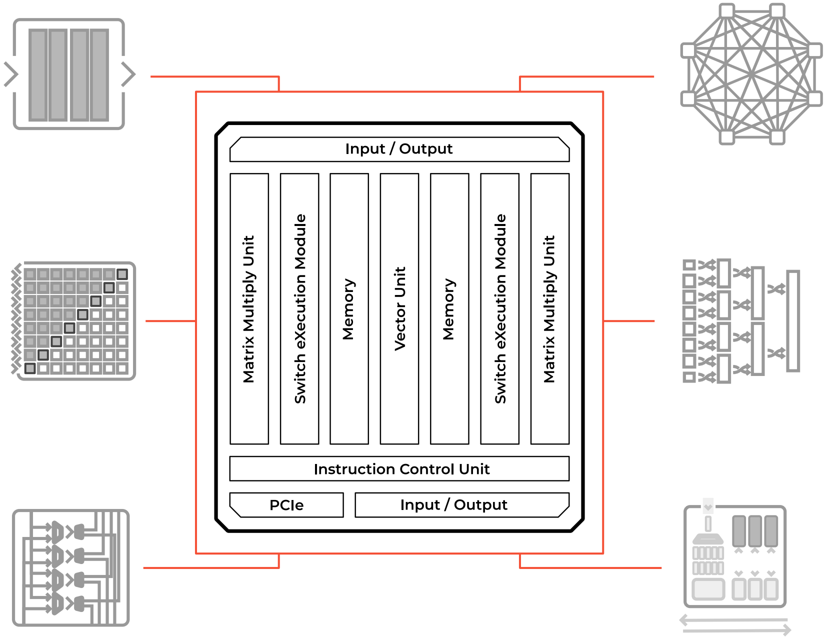
GroqChip™ 1 Overview

Scalable compute architecture

SRAM Memory
Massive concurrency
80 TB/s of BW
Stride insensitive

Groq TruePoint™ Matrix
4x Engines
320x320 fused dot product
Integer and floating point

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

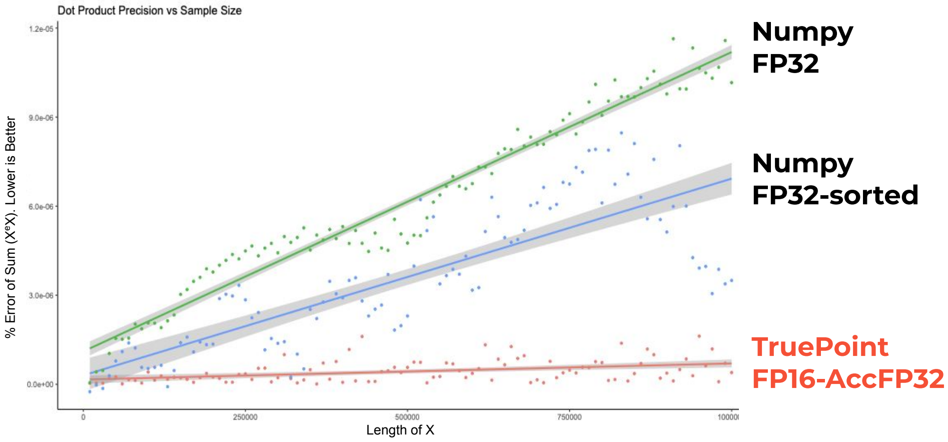
Instruction Control
Multiple instruction queues for instruction parallelism

Groq TruePoint™

High accuracy with fast compute times and low power usage

Residual MSE vs Dot Product Length

ML workloads can take advantage of lower-precision numerics like FP16 or INT8 for quantized models



Linear fits with 95% confidence intervals shown (robust improvement in precision).
Compares against inputs in FP32 but within the range of FP16 values (remove quantization error effects).
Sorted line shows best-case FP32 MSE assuming deterministic compute, like the GroqChip.
Compared against FP64 oracle.

After quantization, losses can continue to accumulate through series of discrete computations

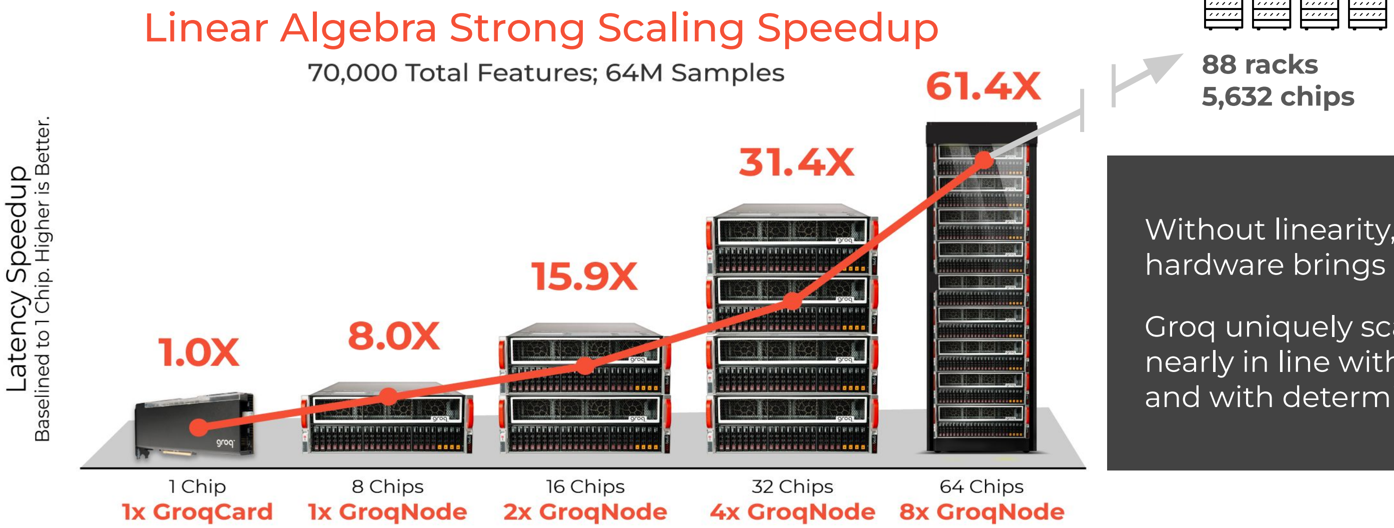
TruePoint takes advantage of mixed-precision in a **320-element fused dot product** with a single rounding step, each dot product then accumulated in FP32

Lower energy to compute FP16 data than wider formats like FP32 or FP64

TruePoint **outperforms** standard IEEE FP32 over long compute lengths

Interactive Compute at Massive Scale

Linear algebra workloads scale near-linearly on Groq architecture



Without linearity, adding more hardware brings diminishing returns

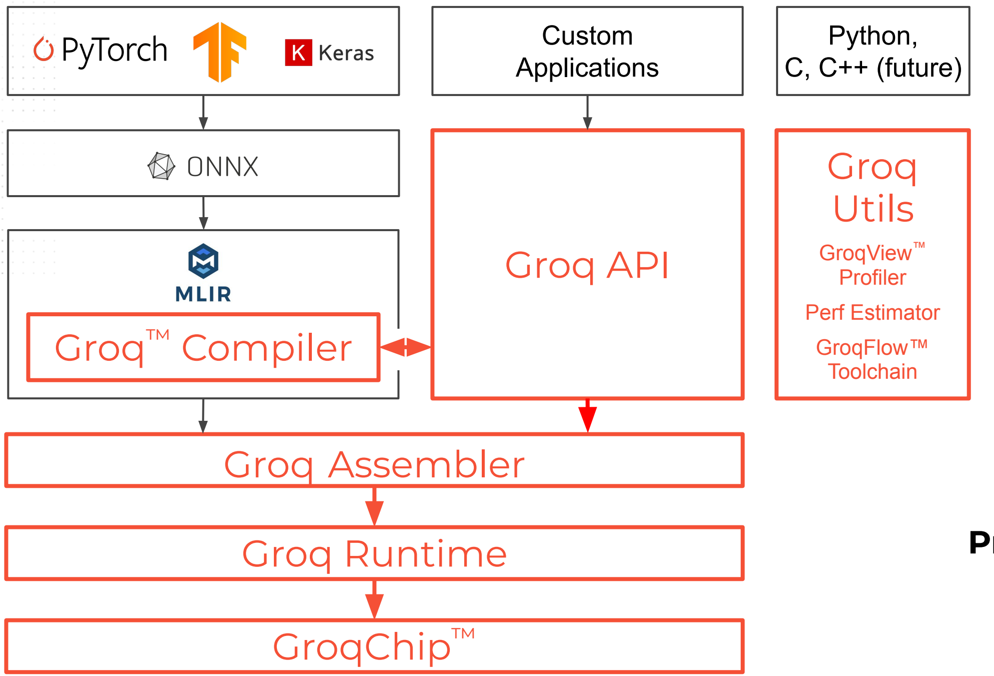
Groq uniquely scales performance nearly in line with volume deployed, and with deterministic certainty

Strong scaling from 1 to 64 chips with 96% linear scaling[1]

[1]As measured by the relative latency of 1 chip vs. 64 chips when running a custom linear algebra workload with 70,000 total features and 64,000,000 samples.

GroqWare™ Suite

Components



Out-of-the-box

Fine-grained Control

Productivity Tools

A Diverse Suite of Development Tools

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

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

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

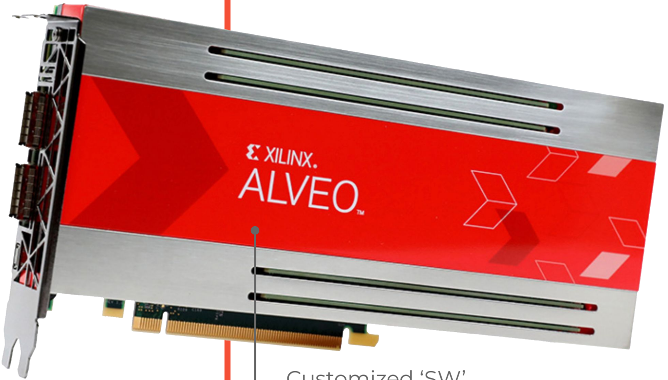
Performance Estimator provides accurate predictions for model compiler performance even before compilation

GroqFlow™ Toolchain enables a single line of Pytorch or TensorFlow code to import and transform models through a fully automated toolchain to run on Groq hardware

Overlapped I/O (Read & Write)
Deterministically Synchronized

Full Duplex
x2/x4 each up to 92 Gbps

Interface Code



'Zero' Latency PHYs

Customized 'SW'
- Converters
- Networking
- Transformers
- Data Cache

DATA CENTER

We have developed a high speed interface to an FPGA which in turn provides:

An interface to the data center

Data preprocessing/conversion

Real-time Image Classification

I/O Accelerator Design

Image processing and classification on GroqChip and I/O Accelerator in real time

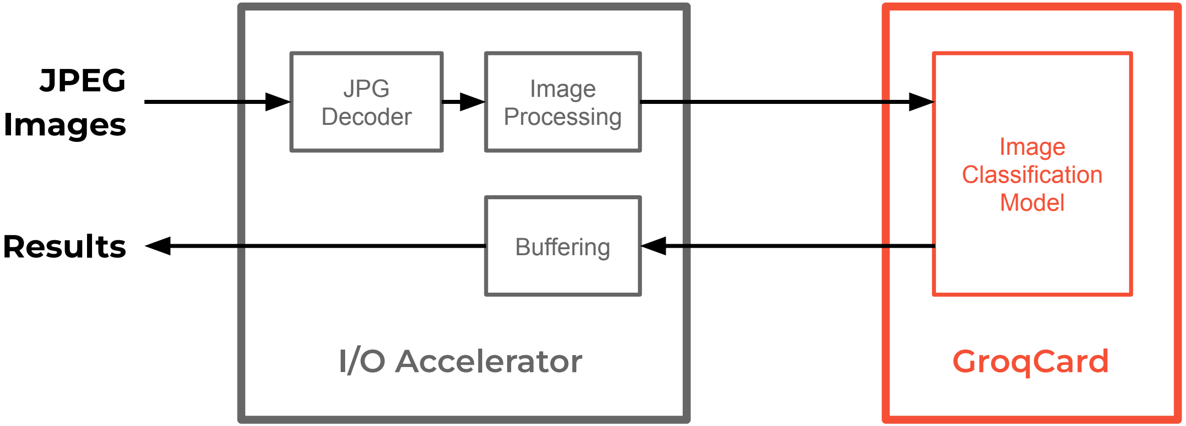
Classification of handwritten numbers

JPEG Decoding and Image Preprocessing on Groq I/O Accelerator

Image classification model on GroqChip

Image data and results transferred using RealScale™ chip-to-chip interconnect

- Ensures that communication between GroqCard and I/O Accelerator does not become the bottleneck



Real-time Image Classification

JPEG decoding

Decompression
(Huffman decoding)

Dequantization

IDCT

Colour space conversion
(YCbCr to RGB)

Image preprocessing

Grayscale conversion

Thresholding

Image centring

Image Classification Model

Simple 2 Layer Neural Network

2 dense linear layers

1st layer uses ReLu activation

2nd layer has Softmax activation

Implemented using Groq API

Trained using MNIST dataset

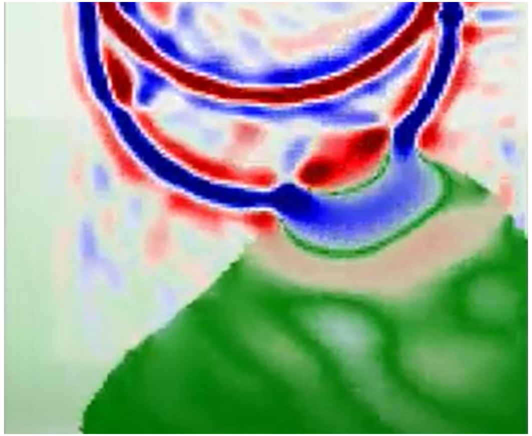
HPC Applications on GroqChip™

HPC DEMOS

Seismic & CFD on GroqChip™

HPC applications running on the AI-inspired Groq architecture

Seismic

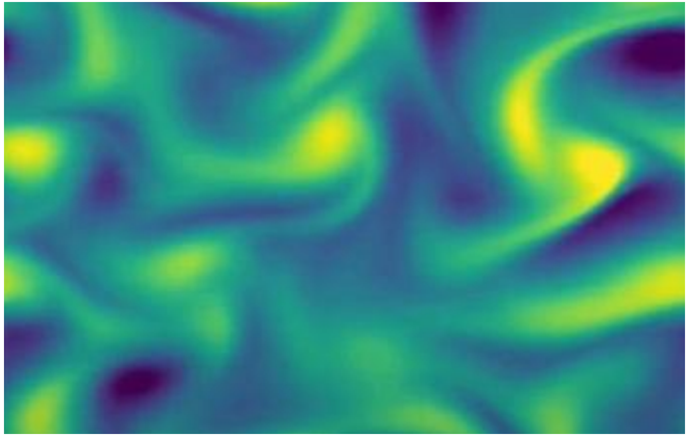


3D finite difference solver for seismic

Scales to multiple nodes

60x speedup

CFD



Finite volume solver

Structured grid method

80x speedup

Acoustic Wave Propagation

HPC applications running on the AI-inspired Groq architecture

Simulate propagation

by solving the acoustic wave equation using explicit finite difference

p	Pressure at a given point
v	Local speed of sound, and is variable in space over the model
$s(t)$	Wave stimulus at a given point at time *t*

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

Implementation on GroqChip™

Main calculation involves applying a seven point star to every point in the wavefield

Split star stencil into 3 dimensions

Calculate each dimension as a matrix-vector multiplication for each ‘row’ of the model in that dimension

Stencil elements are arranged on the diagonal of each row of the matrix

Utilises GroqChip processor’s fast matrix multiplication hardware

Larger domain sizes can be decomposed into blocks

Block size is limited by the size of on-chip SRAM

Fast internal SRAM on GroqChip has capacity for a 128x128x128 cube

Transfer of blocks over PCIe is a performance bottleneck

Use Groq I/O Accelerator to expand the memory capacity of the GroqChip

64 GB of DRAM attached to the FPGA gives enough space for a large domain

GroqChip loads a block from the I/O accelerator, calculates a timestep on it, then writes results back

Calculation Precision

Achieve maximum performance through analysis and dedicated optimisations

Maxeler's investigation suggests that 10/11 bit fixed point arithmetic is sufficient for this application

Matrix Multiply units on GroqChip are optimised for FP16

FP16 has 10 mantissa bits and one sign bit, totalling 11 bits of precision

FP16 has greater range due to the exponent

Groq TruePoint™ arithmetic improves accuracy

Studies involving use of FP16 for seismic modelling achieve speed up proportional to space savings*

*https://gfabieno.github.io/bib/eage_2018.pdf

Questions

MAXELER™

a groq™ company

MAXELER™

a groq company