MAXELER™
a groq company

Accelerating Real-time AI

FIRE: FPGA Innovation Research Exchange

Max Engelen
mengelen@groq.com

Maxeler: a Groq Company

DataFlow driven solutions

- Maxeler has 20 years of experience with data flow compute on FPGAs
- We build and maintain our own compiler for Dataflow model's mapping onto FPGAs
- Joined Groq last year as daughter company

GroqChip™ 1 Architecture Overview

Scalable compute architecture ~ https://groq.com/isca-2020-conference/

SRAM Memory

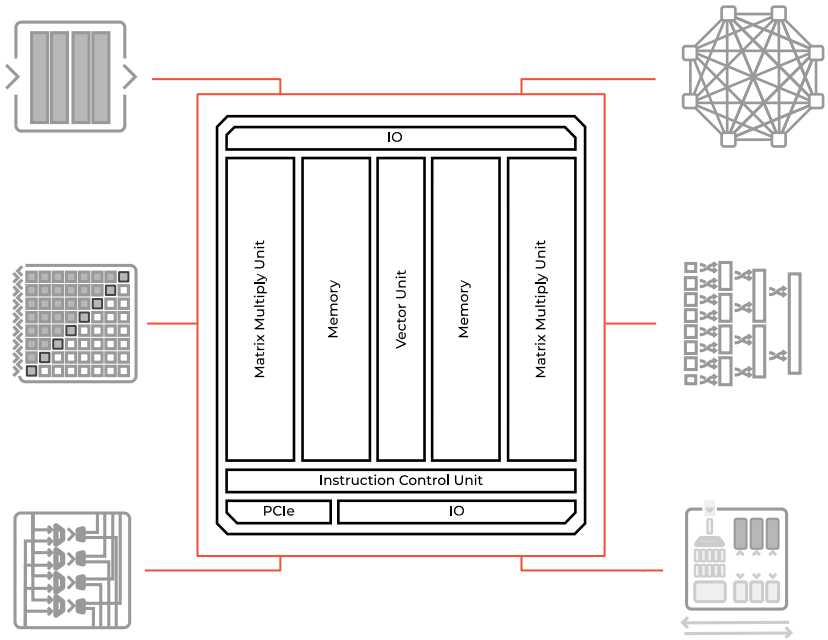
Massive concurrency
80 TB/s of BW
Stride insensitive
220 MiBytes

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

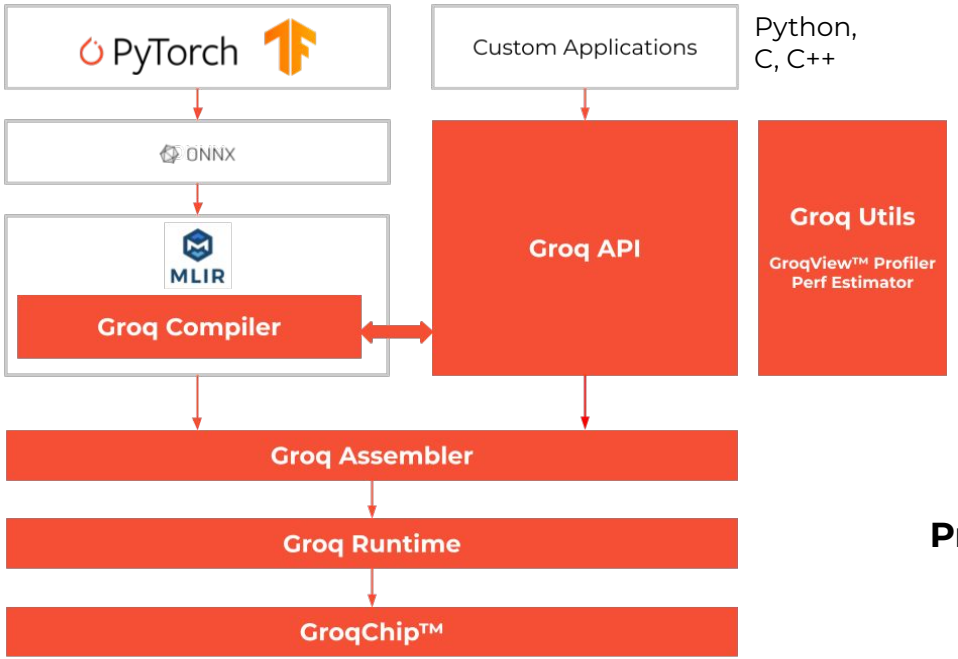
Shift, Transpose, Permuter for improved data movement and data reshapes

Instruction Control

Multiple instruction queues for instruction parallelism

Current GroqWare™ Suite At-a-glance

Accelerating ML & HPC developer velocity



Out-of-Box

Fine Grained Control

Productivity Tools

A Diverse Suite of Development Tools

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

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

+

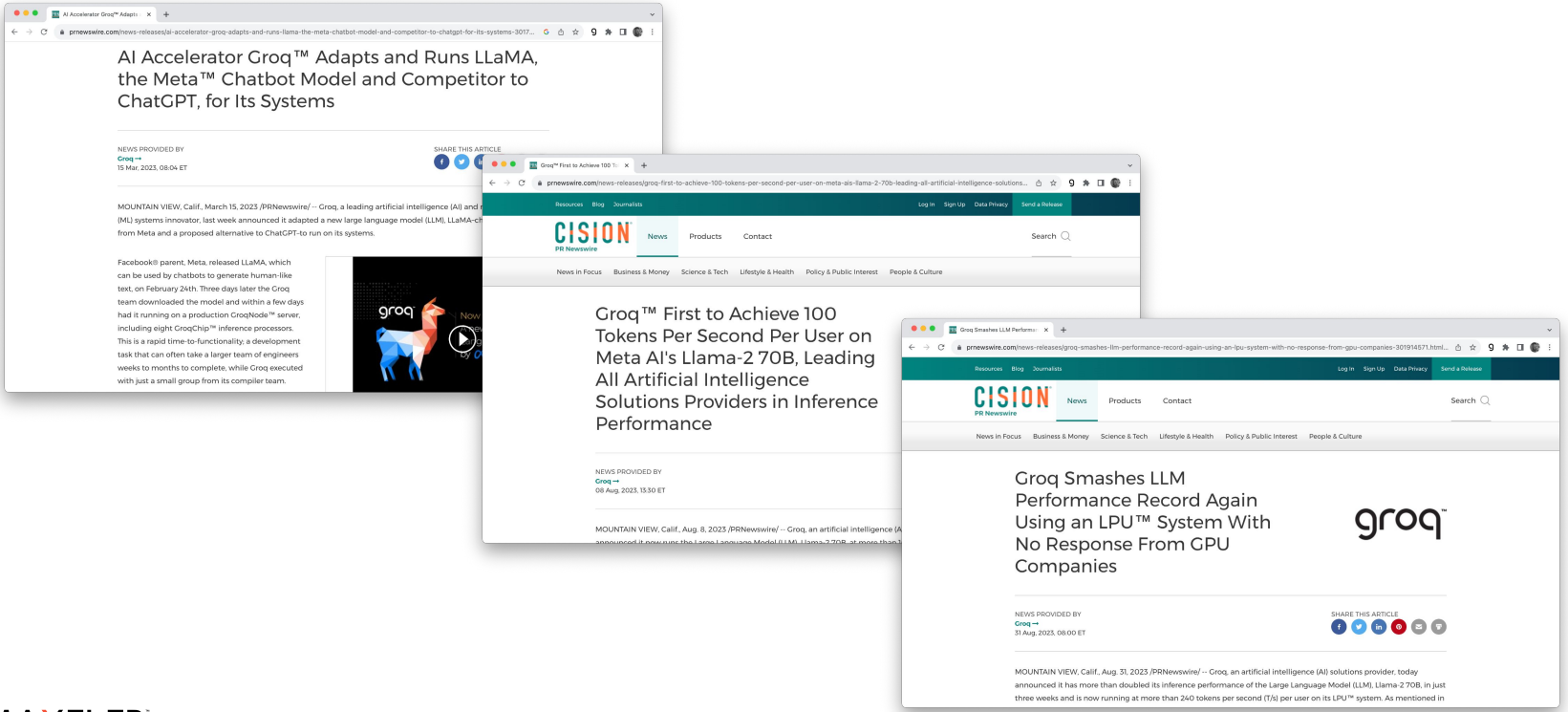
GroqFlow™ Toolchain automatically runs PyTorch models with just one line of code

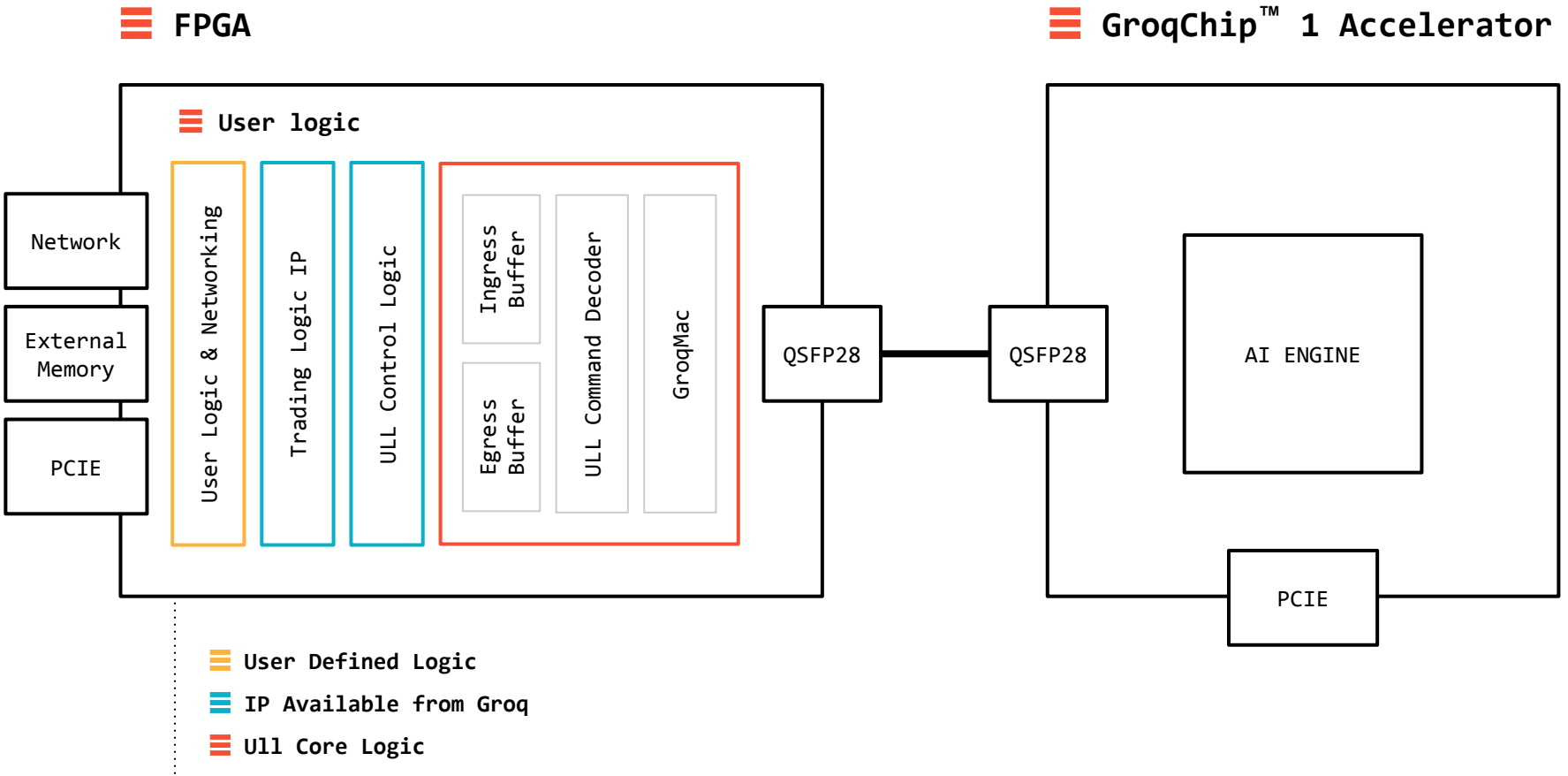
GroqView™ Profiler provides visualization of the chip's compute and memory usage at compile time

Performance Estimator provides accurate predictions even without access to hardware

DEMO: LLMs Running On an LPU™ System

Running Llama-2 70B and achieving ultra-low latency inference





MAXELER™
a groq company

Questions

mengelen@groq.com

MAXELER™

a groq™ company