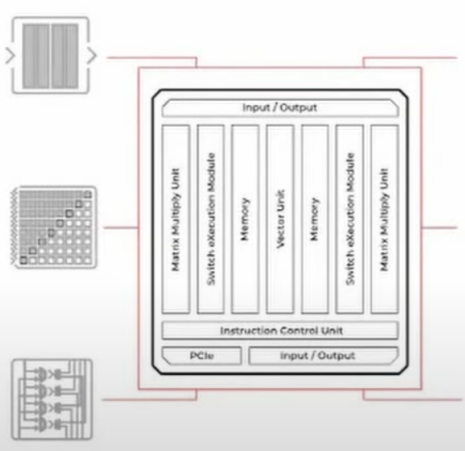
GroqChip™ Overview

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

Architecture Empowering Software

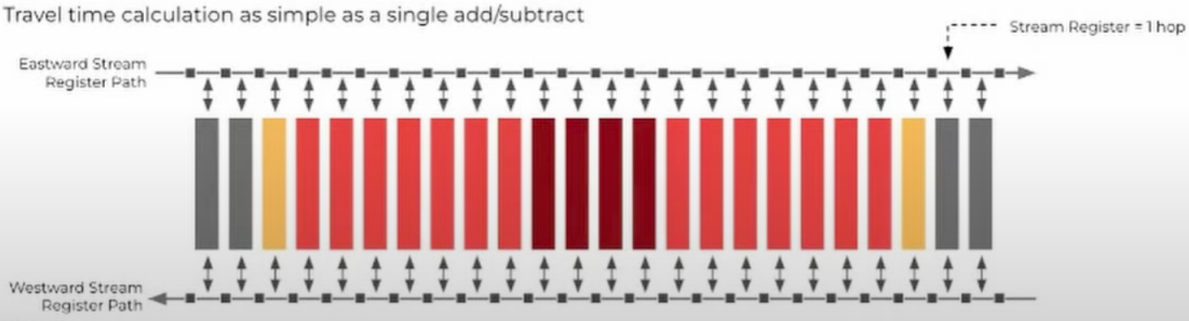
Simple, one-dimensional interconnect for inter-FU communication

Compiler can quickly reason about all data movement between FUs

- Eastward and westward paths made up of arrays of "stream registers"
- Stream register = one-cycle hop

No arbiters / queues = software can easily reason about exact data movement without simulation

Travel time calculation as simple as a single add/subtract



groq

≡ Traffic Nightmare

Stop-and-go bumper-to-bumper traffic gives **long travel times** and **poor gas mileage.**

Unpredictable nature leads to **poor utilization** of roads

A Smart City for AI ≡

Non-stop point to point transport gives **shortest travel time** and **best gas mileage.**

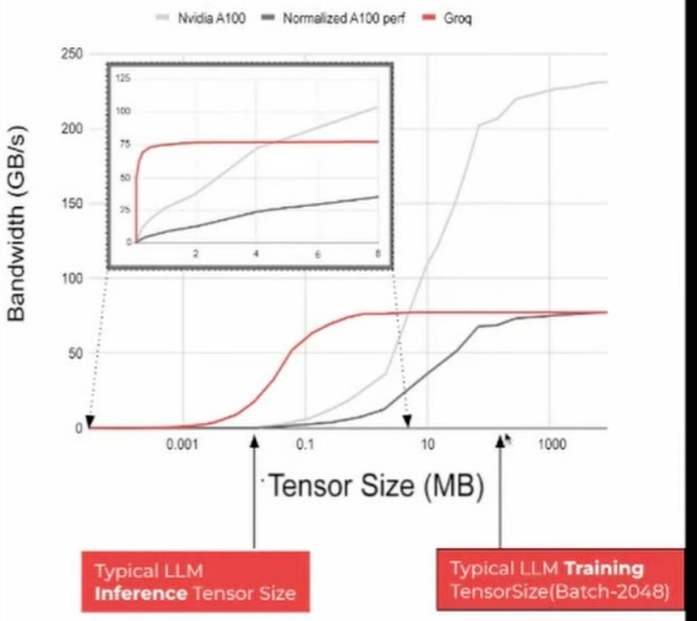
Predictability means **high utilization** and no accidents **EVER!** Nothing is left to chance.

groq

AllReduce Comparison

Comparison made with an eight GPU A100 system with NCCL

- Only a handful of cycles to Read(vector) ↦ Send(vector) enables fine-grained communication across the 16 directly connected links on each TSP
- A100 system has approximately 3x higher network channel bandwidth
- When normalized, the GroqChip device matches the bandwidth at large tensor size while significantly improving bandwidth at intermediate tensor size



Results for A100 were measured on an 8 A100 GPU system with 300 GB/s of NVlink bandwidth per GPU connected through NVSwitch.
Nvidia results are from publicly available data on https://github.com/nvidia/nccl-tests/

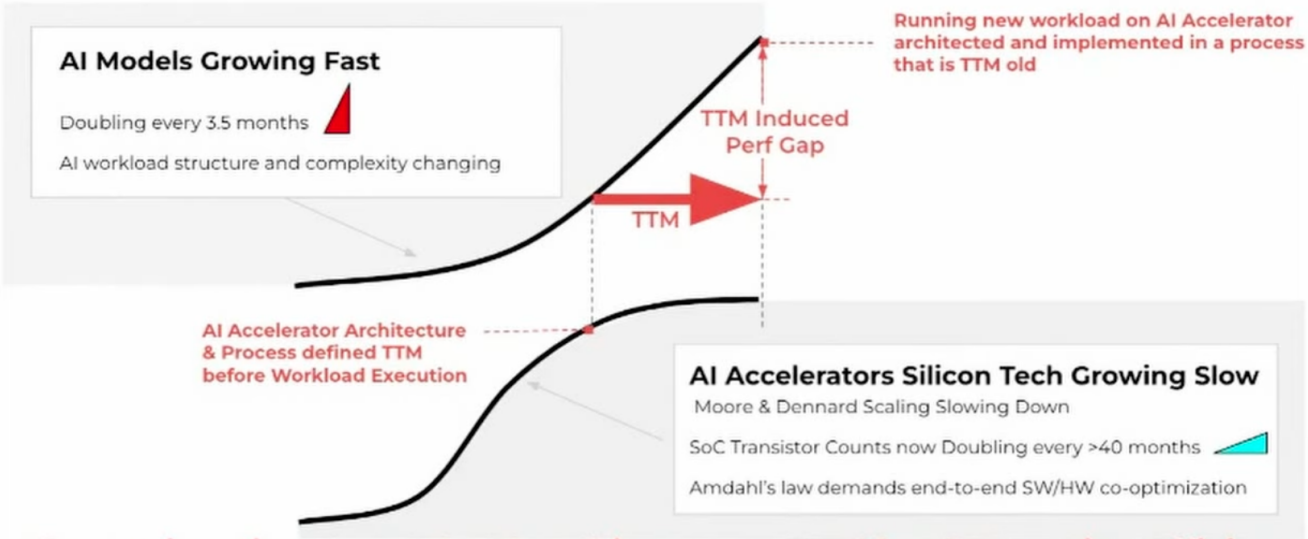
Solution Diversity

Customer Problem Statement	Value Delivered by Groq
Drug discovery: Accelerate time to discovery from days to minutes	**>200x speed-up** when evaluating candidate COVID drugs
Cyber security: Improve accuracy and reduce false positives	**>600x speed up** for real-time cyber-threat anomaly detection; with superior accuracy
Fusion reactor: Enable fully predictable real-time controls systems (<1sec)	**>600x speed up** to make real-time plasma stabilization possible
Capital markets: Enable rapid hypothesis testing at Scale	**>100x speed-up** enabling rapid trading hypothesis testing
General ML: Support a diverse set of popular models	**>500 common models** natively compilable with performance ahead of GPUs

groq

THE BATTLE OF EXPONENTIALS

Reducing TTM improves performance and energy efficiency



Shortening Time to Market (TTM) improves Workload Execution Efficiency

groq

SCALABLE

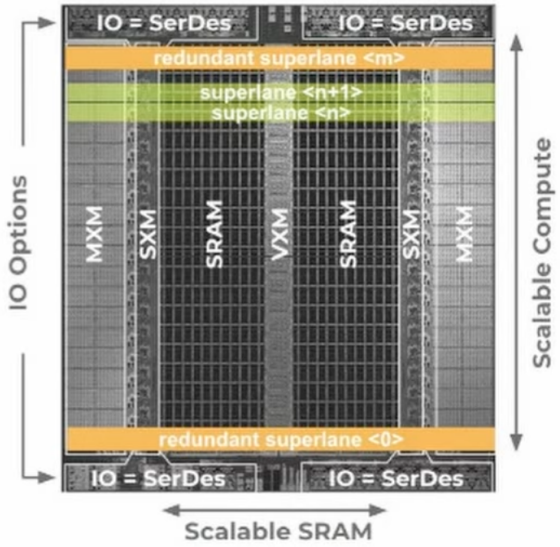
Silicon Tiler For Fast Time-to-market

Multiple Interconnect Options

- C2C for high-radix interconnect
- UCIe for MCM connected sidecar accelerator
- Scalable SXM for BW to/from IO and Compute

Scalable compute architecture

- SRAM scalable capacity
- VXM with scalable number of PEs
- MXM with scalable matrix sizes



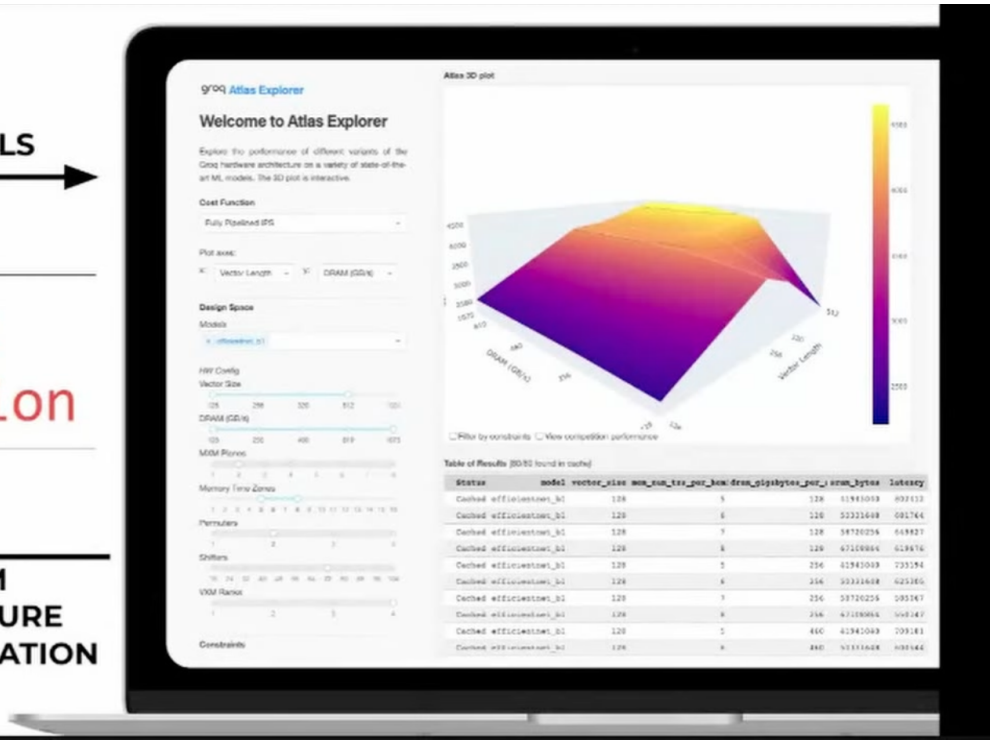
groq

AI/HPC MODELS

DEMOS

≡ Design Exploration

CUSTOM ARCHITECTURE RECOMMENDATION



groq

SILICON TILER

Groq Silicon Tiler Ecosystem

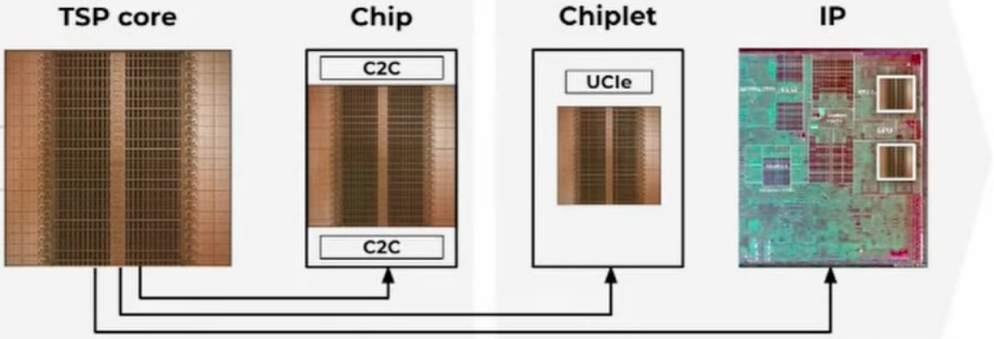
Same
AI SOFTWARE
for the full
HW Ecosystem

Scalable SRAM

(220–440MiB)
with 3D SRAM
extension

Scalable Compute

16 SL: 256x256
20 SL:, 320x320
24 SL: 384x384



groq

Summary

State-of-Art LLM
Performance using
Software-scheduled LPU
Compute and Networking
across 576 chips

Groq Superpowers:

Chip Determinism unlocks performance and efficiency while paving the way for more compute density with 3D chip integration

Low-diameter dragonfly network with abundant path diversity achieves better latency and peak AllReduce performance

Synchronous global communication extends multiple GroqChips into a massive multi-chip single-core cluster

Software Scheduled Deterministic Compute and Network uses **Global time** to provide efficient C2C utilization enabling massive scale

groq