Determinism and the Tensor Streaming Processor.

How Deterministic Processing Delivers Predictability and ROI

Legal Statements

Determinism and the Tensor Streaming Processor.

Groq Tech Doc

Overview

Humans have an innate desire for life to be predictable and repeatable. However, real-world variability tends to get in the way. But, what if we could remove the variability?

Think of your morning commute. Each day variables like traffic, stop lights, accidents, and departure time all prevent you from arriving exactly on time each day. Instead, imagine waking up in the morning, being told the exact time you need to leave, and traveling at max speed (without stop lights) to arrive on time.

One might call this a miracle, but Groq calls it determinism, an underlying tenant and key differentiator of Groq's Tensor Steaming Processor (TSP) architecture.

A deterministic architecture delivers predictable and repeatable performance. The total execution time is known at compile time, so you know exactly how much time (how many clock cycles) it will take to run a workload and exactly the order of operations—with zero variance. From a single GroqChip™ accelerator to a network of chips, workloads always run the exact same way the first time and the millionth time—with no performance variation.

Take financial services. Quantitative research, back-testing, and trading demand performance and low latency, but equally important are predictable and repeatable results. If you are performing a risk assessment calculation to inform a trading event, you want to bound the duration of that process. If the first process takes a millisecond and the next iteration is 10 milliseconds, that's not good enough. Every single one — from the first to the millionth — has to be exactly the same. Furthermore, what if the exact same risk calculation produced a slightly different numerical result based on the order of execution in the underlying hardware? Groq's deterministic architecture maintains the order of operations every single time and executes in a known, exact amount of time.

This concept of perfectly scheduled determinism is exactly what Groq's hardware seeks to execute and the key is in the software, the GroqWare™ suite.

Unlike traditional CPUs (Central Processing Unit), GPUs (Graphics Processing Unit), or FPGAs (Field Programmable Gate Array), a GroqChip has no control flows, no hardware interlocks, no reactive components like arbiters or replay mechanisms that would perturb or permute the order of events and thereby vary processing performance on the same workload. Rather, Groq takes a software-defined approach where flow control is orchestrated by the compiler. When a tensor is read out of memory and destined to a functional unit to be operated on, the compiler knows

exactly how long it is going to take. This approach allows for a simpler and more efficient hardware architecture that executes a predetermined script.

The diagram in Figure 1 shows how this simplified software-defined hardware approach—where all execution planning happens in software—frees up precious silicon space for additional processing capabilities. By removing reactive components that introduce an element of non-determinism, Groq effectively addresses variance in tail latency (a GroqChip doesn't have tail latency) thereby eliminating the need for developers to spend precious time profiling workloads. Furthermore, since you know exactly what is happening across the chip each clock cycle, you can predict power in a precise manner. This also means you can optimize workload kernels to execute at a desired power threshold, enhancing TCO.

Software-defined Hardware

All execution planning happens in software, freeing up valuable silicon real estate and providing additional cache for memory bandwidth

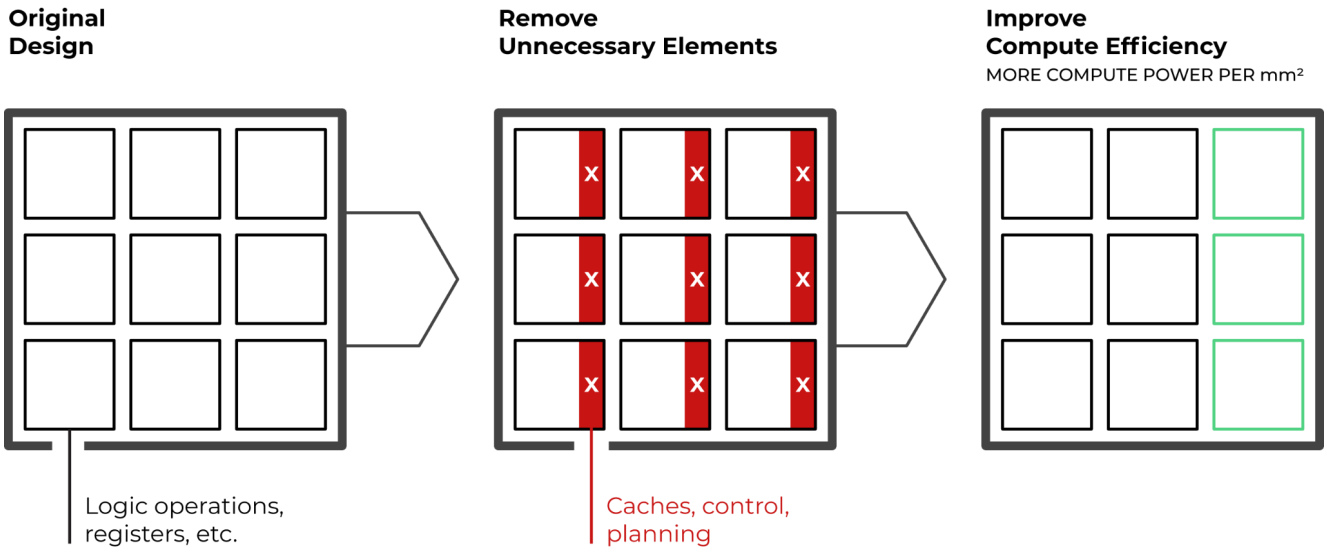

Figure 1. Groq's simplified software-defined hardware approach frees up silicon space for additional processing capabilities.

Dependable execution time also enables you to run an ensemble of models within a fixed time budget. In autonomous driving, an ensemble of models are required to be executed within a demanding, and fixed perception time window. For example, you must employ object detection (looking for a stop sign), then behavioral prediction (looking for objects like pedestrians around the stop sign and what paths they might take), and a suite of other sensor processing in order to provide the accuracy required in this critical application. Since each model's execution is deterministic, you know with absolute precision how to schedule the suite of sensor processing needed to maintain acceptable quality of service.

Three Points That Matter

1

Compiler decision making gives you determinism

2

Determinism plus low latency equals predictable performance

3

No reactive components means no tail latency

Benefits

Ease of Design and Debuggability

Since workloads always run the same way, once you get a workload working, you never have to debug the same workload twice. Gone are the days of worrying about variance from 'noisy neighbors' or dynamic profiling of workloads which complicate and bog down the development process. Furthermore, Groq's static profiling immediately provides you with an in-depth performance report and visualization of the chip's compute and memory usage at compile time — all without the need to run the program on hardware — enabling true developer velocity.

Predictability

Because control has moved into the software stack, the hardware is consistent and predictable. The compiler assembles deep learning models into instruction streams, all orchestrated in advance. When you start a program you know with 100% certainty exactly when it's going to be finished.

QoS (Quality of Service)

Determinism plus low latency equals cutting-edge predictability. This means guarantees on throughput, latency, and numerical results. This unmatched QoS becomes even more valuable when scheduling ensemble models or multi-input systems. There's zero variance, translating to rock solid QoS.

TCO (Total Cost of Ownership)

Determinism also provides flexibility, since you have control of how the chip uses power. You can slow down a program if you're looking for efficiency. And you'll no longer incur costs associated with over-provisioning power supplies for huge GPU spikes. Finally, Groq's single core, deterministic architecture eliminates any guess work from the development process, accelerating time to market.

Use Cases

Government

Quality of Service calculates the response times to complete a task. For instance, a 99% QoS score represents the time it takes to complete 99% of tasks. While government agencies admire performance in their mission critical applications, consistency is much more critical. And while consistent performance is in huge demand, meeting these standards has been historically difficult. Determinism takes predictability to a new level: by allowing ensemble model execution and providing deterministic execution time allows end-to-end service agreements to be negotiated and delivered reliably.

Industrial

Industrial automation uses computer vision (CV), so they can "see" and extract, process, and analyze information from visual inputs. CV, paired with machine learning, increases productivity and grows revenue with vision-guided robots, anomaly detection, and efficient inspection, scanning, labeling, tracking, and tracing. Computer vision requires low latency, high performance, and extreme accuracy. Groq's simple architecture delivers all three, with industry leading batch-1 speed, Groq TruePoint™ engine's numerical accuracy, and chip-to-chip low latency interconnect. These industrial processes also require 100% predictable behavior to optimize efficiency. Since Groq is deterministic all the way to the network, you know exactly how many clocks it will take to run a workload—or multiple chain workloads— every single time.

By the Numbers

- The metric is simple. The compute behavior is the same the first time and the millionth time. Because decision making is moved to the software stack, there are no longer hardware schedulers figuring out how to route things onto the chip (a GroqChip doesn't even have hardware schedulers.) The compiler schedules all events happening in different times on different functional units (vector unit, memory unit, MXM, etc.) and the hardware simply executes a predetermined script. It knows what it's going to do on every single cycle on every functional unit across the chip, all in advance.
- Groq's architecture can perform over 400,000 Deep Learning operations in the time it takes a conventional CPU to fetch a cache line.
- Meanwhile, the architecture provides incredible performance per watt — driving down total cost of ownership, while reducing your carbon footprint.

At Groq, we are radically simplifying compute with the end-user in mind. Do you have a workload that could benefit from determinism? Reach out via the contact information below.

Interested in Learning More?

For more information on Groq technology and products, contact us, follow us on YouTube and Twitter, and connect with us on LinkedIn.

Additional Related Documents

- Accuracy Tech Doc
- Latency Tech Doc
- Scalability Tech Doc
- Velocity Tech Doc