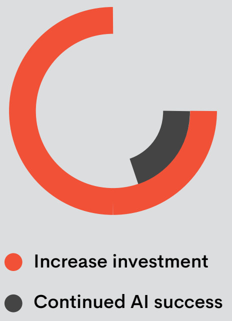
Tensor Streaming Architecture Delivers Unmatched Performance for Compute-Intensive Workloads

AUTHOR
Dale Southard, Ecosystem Solutions Distinguished Architect, Groq

© 2019 Groq. All rights reserved.

Businesses and governmental entities are increasingly turning to compute-intensive applications, such as machine learning and artificial intelligence (AI), to enhance customer experience, increase competitive advantage and improve security and safety in communities. However, achieving and maintaining the high-performance processing that these workloads require is extremely difficult due to the growing complexity of hardware processor models.

As reported in Forbes, "There are still significant challenges to companies wishing to adopt smart, cognitive computing processes into their operations." According to a McKinsey report cited by Forbes, nearly three-quarters of over 2,000 organizations surveyed expected to increase investments in AI in the future, but just one in five respondents claimed they had already successfully rolled out AI in more than one process. A separate CIO/IDG survey indicated that 53 percent of companies found that quick and reliable deployment of AI models to production is either very or extremely challenging, and reported that only one in three AI projects is successful.



AI Adoption

Nearly three-quarters of over 2,000 organizations surveyed expected to increase investments in AI in the future, but just one in five respondents claimed they had already successfully rolled out AI in more than one process.

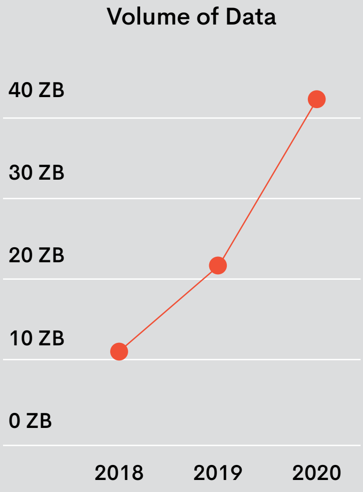
- McKinsey

To gain the benefits of AI, smart infrastructure and predictive intelligence will require a much simpler and more scalable processing architecture that can sustainably accelerate the performance of compute-intensive workloads. A less complex chip design is the answer.

The Challenge of Achieving High-Performance Machine Learning Processing

Why is it so difficult to achieve high-performance compute processing? In part, the challenge involves managing rapidly increasing volumes of data. Data scientists estimate that the volume of data is doubling every two years, and will reach 44 zettabytes by 2020 — in other words, there will be more than 40 times more bytes of data than there are stars in the observable universe.

In addition, meeting human-like inference performance with neural networks will require exponential increases in model complexity and computing throughput. However, achieving faster, more efficient neural net processing won't come from scaling up the number of physical processors, as investments in traditional server clusters are reaching a computational cost wall. Meanwhile, standard computing architectures are crowded with hardware features and elements that offer no advantage to inference performance. Inference has reached a bottleneck.


Volume of Data

Data scientists estimate that the volume of data is doubling every two years, and will reach 44 ZB by 2020.

- IDC

Machine learning computations also make unprecedented demands on processors – and developers. To perform more and more operations per second, chips have become larger and much more complex, with multiple cores, multiple threads, on-chip networks, and complicated control circuitry. To accelerate software performance and output, developers struggle with complicated programming models, security problems, and loss of visibility into compiler control due to layers of processing abstraction. To yield higher machine learning performance within these constraints relies on laborious hand-tuning optimization that is based on intimate knowledge of the hardware architecture.

Another reason that achieving high-performance compute processing is so challenging is that processing systems and architectures have grown so complex. As chip designers have strived to wring more and more performance out of silicon, they have integrated more and more components and building blocks on the chips, driving up chip complexity with feature and processor bloat.

For machine learning developers seeking inference performance, these feature-rich architectures are restrictive, as they present a patchwork of "dark silicon" – areas that are dedicated to hardware elements and functionality that offer no advantage for AI or machine learning processing.

For instance, CPU architectures are the dominant inference platform but they are limited for machine learning performance by latencies introduced by execution models like out-of-order execution, speculative execution, and multiple branch prediction that stall high-performance processing. In addition, the transistors dedicated to managing out-of-order sequencing and other instruction cycle queuing techniques are effectively dark silicon, not available for inference performance. Also, increasingly large inference models stretch the platform's cache capabilities, reducing the efficacy of the CPU multi-level cache architecture.

GPU architectures are optimized for DRAM bandwidth and were historically built on multi-data, or multi-task, fixed-structure processing engines. Though GPUs are designed for massively parallel processing, these platforms were originally created for real-time graphics rendering, and the chip is crowded with various rendering engines dedicated to creation of images in a frame buffer. These graphics rendering features don't benefit AI performance, meaning that the portions of the chip dedicated to this hardware are essentially dark silicon in terms of inference performance. In addition, the overhead for fixed-feature GPU engines is the memory access latency, and deep learning is already pushing the limits of external memory bandwidth.

As if these conventional computing platforms weren't complicated enough, a chipmaker has recently released a new type of processor for AI acceleration that it claims is the "most complex processor chip that's ever been built."

At Groq, we believe that moving toward greater complexity to gain more performance is a move in the wrong direction. To illustrate this point, look at internal combustion engines (ICEs), which have been in use for over a century. To gain more power and more performance, engineers have added more and more complexity to the engine. Turbochargers, variable valve timing, fuel injection, computerized ignition systems, and dual-clutch transmissions have made the ICE run much faster and more efficiently. These technologies have also made the ICE much, much more complex. A typical four-cylinder ICE has hundreds of moving parts.

At Groq, we believe that increased processing performance for compute-intensive workloads must come from simpler, more innovative and efficient technologies.

But if the ultimate job of an engine is to provide power and propel a passenger vehicle, the far less complex electric motor is a much more efficient and effective alternative. Brushless DC motors (BDCMs), which are powered by electromagnets, change the automotive power system paradigm completely. With no sparking and producing much less heat and noise, the BDCM reduces overall cost and maintenance and doesn't even require a clutch or a transmission. It's a much simpler and more reliable power source for a passenger vehicle, and that's before we begin tallying up the obvious environmental and sustainability benefits.

Simplicity is the Answer

At Groq, we believe that increased processing performance for compute-intensive workloads must also come from simpler, more innovative and efficient technologies. The current complexity of processor architectures is the primary inhibitor that slows developer productivity and hinders the adoption of AI applications and other compute-heavy workloads. Current processor complexity decreases developer productivity. Moore's law is slowing, making it harder to deliver ever-greater compute performance.

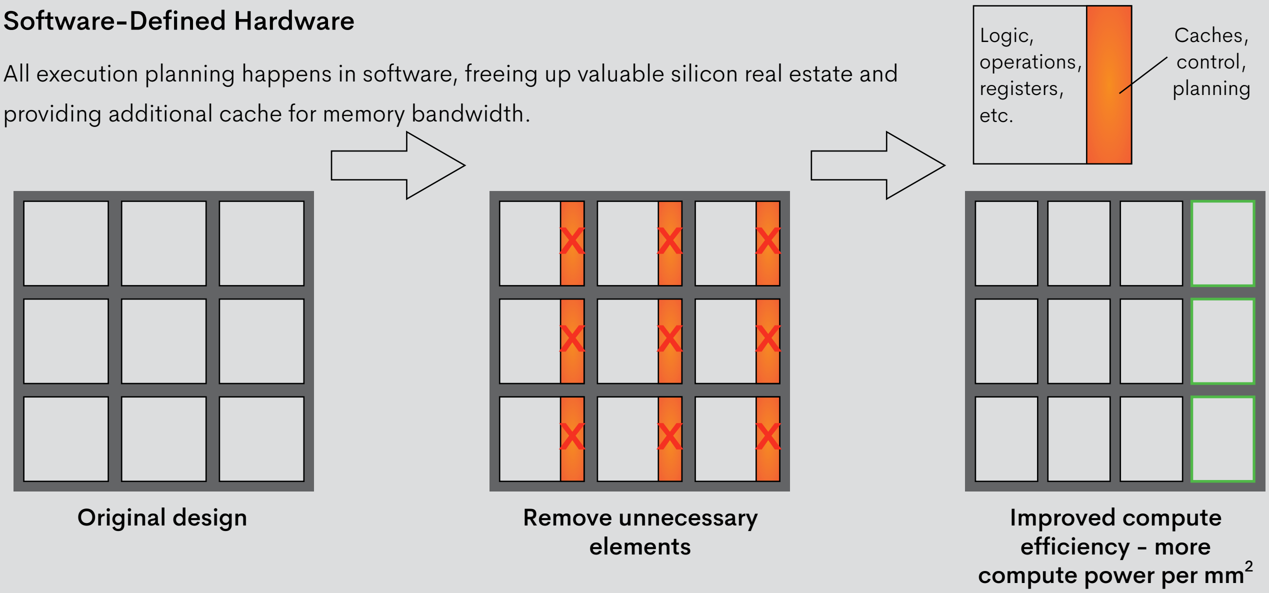
Because current processor complexity decreases developer productivity, Moore's law is slowing, making it harder to deliver ever-greater compute performance.

Groq is introducing a new, simpler processing architecture designed specifically for the performance requirements of machine learning applications and other compute-intensive workloads. Groq's chip design reduces the complexity of the traditional hardware-focused development, so developers can concentrate on algorithms (or solving other problems) instead of adapting their solutions to the hardware. The simpler hardware also saves developer resources by eliminating the need for profiling, and also makes it easier to deploy AI solutions at scale.

Inspired by a software-first mindset, Groq's overall product architecture provides an innovative and unique approach to accelerated computation. In Groq's architecture, the compiler choreographs the operation of the hardware. All execution planning happens in software, freeing up valuable silicon space for additional processing capabilities. The tight control provided by this architecture leads to the deployment of better and faster models using industry standard frameworks and results in fast and predictable performance on current and future workloads.

Software-Defined Hardware

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


Logic, operations, registers, etc.
Caches, control, planning
Original design
Remove unnecessary elements
Improved compute efficiency - more compute power per mm²

This architecture, implemented in our tensor streaming processor (TSP), provides a new paradigm for achieving both flexibility and massive parallelism without the limitations and communication overheads of traditional GPU and CPU architectures. The Groq compiler orchestrates everything: Data flows into the chip and is plugged in at the right time and the right place to make sure calculations occur immediately, with no stalls.

This allows Groq's chip performance to be deterministic. The compiler reconfigures the hardware dynamically to perform each calculation so there is no abstraction between the compiler and the chip. Because it understands the hardware and the speed of each instruction, the compiler can tell the hardware exactly what to do, and when. In a traditional architecture, not only does it take a lot of power and time to move data from DRAM into the processor, but the processing performance on the same workload is variable. In a typical workflow, a developer profiles and tests a workload or program by running it over and over to validate and measure its average processing performance. Due to variables in how the processor receives and sends data, this processing can reveal slightly different results, and it's the job of the developer to hand tune the program to attain a predetermined level of reliability.

But with Groq hardware and software, the compiler knows exactly how the chip works and precisely how long it takes to perform each computation. The compiler moves the data and the instructions into the right place at the right time so that there are no delays. The flow of instructions to the hardware is completely choreographed, making processing fast and predictable.

Developers can run the same model 100 times on the Groq chip and receive precisely the same result each time. Our software-defined hardware delivers deterministic results – the same performance over and over again, exactly.

Developers can run the same model 100 times on the Groq chip and receive precisely the same result each time. Our software-defined hardware delivers deterministic results – the same performance over and over again, exactly.

This precision of performance is extremely valuable, especially for applications where safety is paramount. In an autonomous vehicle, the navigation system may need to react within milliseconds to avoid hitting an obstacle, so having a processing architecture where minimum response times are ensured each and every time is essential for safety.

Systems designed using Groq hardware do not suffer from long tail latencies, and AI systems can be tailored to operate within a specific power or latency budget. The run-to-run performance, which is reported at compile time, is predictable for performance-critical and Quality of Service (QoS) guarantees. Since the software knows what data and instructions will be needed for each computation, the compiler can move the data from DRAM to the processor – at the right place at the right time – to make processing as efficient as possible.

Our flexible architecture simplifies the development process and eliminates the need for hand optimization and profiling, spurring development velocity.

Groq's software-first mindset, in which the compiler determines the hardware architecture, leads to a simpler, higher performance design to accelerate inferencing workflows. Our flexible architecture simplifies the development process and eliminates the need for hand optimization and profiling, spurring development velocity.

Groq is ideal for deep learning inference processing for a wide range of AI applications, but it is critical to understand that the Groq chip is a general-purpose, Turing complete, compute architecture. It is an ideal platform for any high-performance, compute-intensive workload.

More Performance Per Transistor

Groq's tensor streaming architecture delivers where it counts — performance.

Compared to best-in-class GPU and CPU designs, Groq's processor delivers an astounding 3x-6x performance per transistor. This advantage in raw performance means much higher delivered performance, decreasing latency and reducing cost.

The result is a design architecture that is simpler to use and capable of much higher performance than traditional compute platforms. Because of its simplicity of design, Groq is positioned to be the only hardware and software solution with a sustainable performance advantage beyond the limitations of process scaling.

Learn more about Groq and get in contact by visiting www.groq.com.