What is a Language Processing Unit?

Groq & the Language Processing Unit

Groq builds fast AI inference. The Groq LPU™, AI Inference Technology, delivers exceptional compute speed, affordability, and energy efficiency at scale.

Groq solutions are based on the Language Processing Unit (LPU), a new category of processor. Groq is the creator of the LPU and built it from the ground up to meet the unique characteristics and needs of AI. LPUs run Large Language Models (LLMs) at substantially faster speeds and, on an architectural level, up to 10x better energy efficiency compared to GPUs.

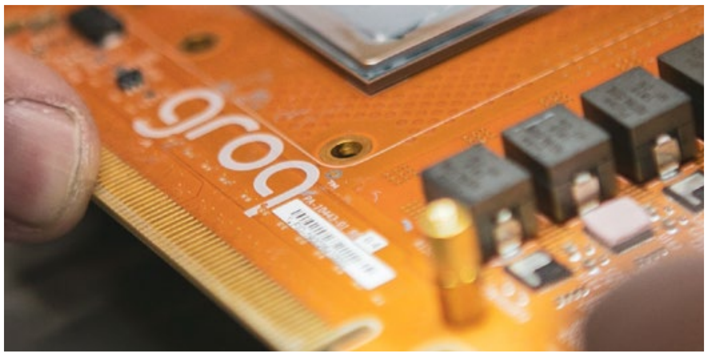
This paper explains the design principles of the Groq LPU and why its architecture delivers such exceptional performance.

Background

From Moore's Law to AI Inference

For decades computer software was the beneficiary of Moore's law, Gordon Moore's self-fulfilling 1965 prophecy that the processing power of a chip would double roughly every two years while keeping costs steady. The law held for several decades, aided by the growing use of multi-core processors (CPUs and GPUs).

Each step of this hardware progression introduced more complexity into systems. Multi-core CPUs and GPUs, for example, are quite powerful and can handle a range of applications, but require ancillary components on the silicon – caches, buffers, prefetchers – to optimize execution. This complexity creates inconsistencies in the runtime execution of a program. This can be managed by software kernels, which create better execution consistency but are quite complex themselves.



With the shift towards inference plus the emergence of LLMs and similar AI workloads, Groq took the opportunity to rethink software and hardware architecture. LLMs are highly powerful, but when run in inference, they rely on a limited set of linear algebra operations, primarily matrix multiplication tasks. AI inference compute boils down to running a massive amount of linear algebra operations on large-scale data. While GPU hardware can host these operations, it isn't designed for it. GPUs will always be limited in how much they can increase inference speed and efficiency given that their legacy architecture was built for independent parallel operations like graphics processing.

So Groq built the LPU. Its four core design principles deliver performance advantages today and tomorrow.

- Software-first
- Programmable assembly line architecture
- Deterministic compute and networking
- On-chip memory

LPU Design Principle 1

Software-first

The Groq LPU architecture started with the principle of software-first. The objective was to make the software developer's job of maximizing hardware utilization easier and put as much control as possible in the developer's hands.

GPUs are versatile and powerful; they can handle many different compute tasks. But they are also complex, putting extra burden on the software. It must account for variability in how a workload executes, within and across multiple chips, making scheduling runtime execution and maximizing hardware utilization much more challenging. To maximize hardware utilization on GPUs, every new AI model requires coding of model-specific kernels. This is where our software-first principle is so important – with GPUs, the software is always secondary to the hardware.

The Groq LPU was designed from the outset for linear algebra calculations – the primary requirement for AI inference. By limiting the focus to linear algebra compute and simplifying the multi-chip computation paradigm, Groq took a different approach to AI inference and chip design. The LPU employs a programmable assembly line architecture, which enables the AI inference technology to use a generic, model-independent compiler and stay true to its software-first principle. The software is always primary, in complete control of every step of inference.

Software-first isn't just a design principle though – it is actually how Groq built its first generation chip, GroqChip 1 processor. We didn't touch chip design until the compiler's architecture was designed. The compiler accepts workloads from several different frameworks, running those workloads through multiple stages. As the compiler maps and schedules a program to run across one or multiple LPUs, it optimizes performance and utilization. The result is a program encompassing all data movement information throughout execution.

The objective was to make the software developer's job of maximizing hardware utilization easier and put as much control as possible in the developer's hands.

LPU Design Principle 2

Programmable Assembly Line Architecture

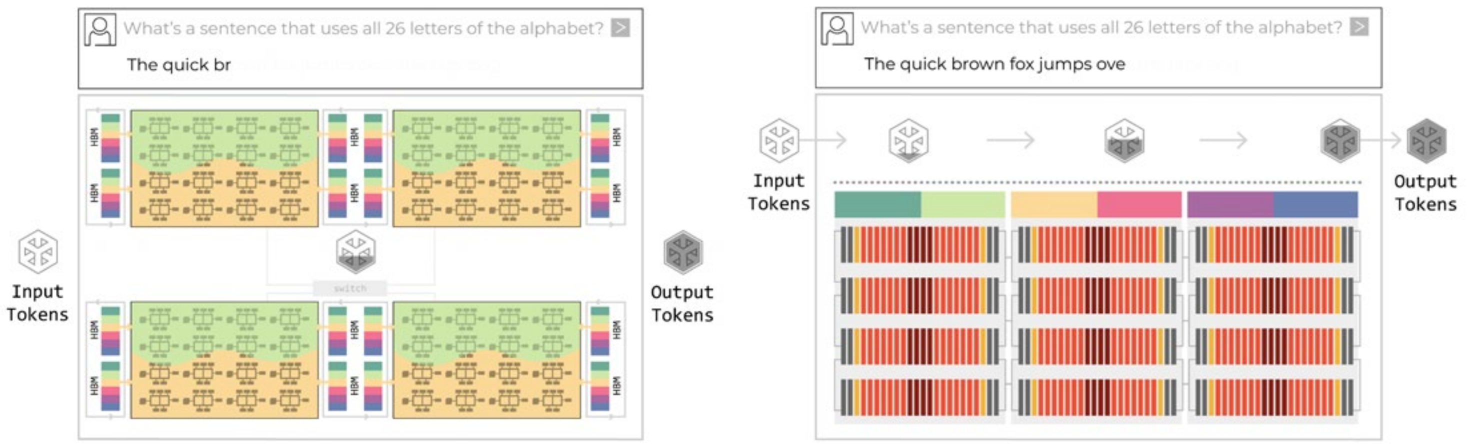
The primary defining characteristic of the Groq LPU is its programmable assembly line architecture.

The LPU features data "conveyor belts" which move instructions and data between the chip's SIMD (single instruction/multiple data) function units. At each step of the assembly process, the function unit receives instructions via the conveyor belt. The instructions inform the function unit where it should go to get the input data (which conveyor belt), which function it should perform with that data, and where it should place the output data. This process is all software-controlled; no synchronization is required within the hardware.

The LPU programmable streaming architecture supports an assembly line process within a chip as well as between chips. There is ample chip-to-chip bandwidth, which enables the data conveyor belts to flow between chips as easily as within a chip. There is no need for routers or controllers for inter-chip connectivity, even at maximum capacity.

The assembly line process within and across chips eliminates bottlenecks. There is no waiting for compute or memory resources to complete a task. There is no need for additional controllers on the chip given there are no bottlenecks to manage. The assembly line moves smoothly and efficiently, perfectly in sync.

This is a big improvement compared to how GPUs work. GPUs operate in a multi-core "hub and spoke" model, where an inefficient data paging approach requires significant overhead to shuttle data back and forth between the compute and memory units within and across chips. GPUs also utilize multiple hierarchies of external switches and networking chips, both within and across racks, to communicate among themselves, further exacerbating the software's scheduling complexity. The result is a hard-to-program, multi-core approach.


What's a sentence that uses all 26 letters of the alphabet?
The quick br
HBM
switch
Input Tokens
Output Tokens
What's a sentence that uses all 26 letters of the alphabet?
The quick brown fox jumps ove
Input Tokens
Output Tokens

The Groq LPU programmable assembly line architecture (right) is much faster and more efficient than the GPU's "hub and spoke" approach (left).

LPU Design Principle 3

Deterministic Compute & Networking

For an assembly line to operate efficiently, there needs to be a high degree of certainty about exactly how long each step will take. If there is excessive variability in how long a particular task takes to execute, that variability manifests across the entire assembly line. An efficient assembly line requires highly precise determinism.

The LPU architecture is deterministic, meaning every execution step is completely predictable to the smallest execution period (also known as clock cycle). The software-controlled hardware knows with a high degree of precision exactly when and where an operation will occur and how long it will take.

The Groq LPU achieves its high degree of determinism by eliminating contention for critical resources, namely data bandwidth and compute. There is ample capacity for routing data around the chip (the conveyor belts) and plenty of compute in the chip's functional units. There is no issue with different tasks using the same resource, so there are no execution delays due to resource bottlenecks.

The same is true for routing data between chips. The LPU data conveyor belts also operate between chips, so connecting chips results in a larger programmable assembly line. Data flow is statically scheduled by the software during compilation, and executes the same way every time the program runs.

LPU Design Principle 4

On-chip Memory

LPUs include both memory and compute on-chip, vastly improving the speed of storing and retrieving data while eliminating timing variation. While determinism ensures the assembly line runs efficiently and eliminates the variability of each compute stage, on-chip memory enables it to run much faster.

GPUs utilize separate high-bandwidth memory chips, introducing complexity – multiple layers of memory cache, switches, and routers to move the data back and forth – while also consuming significant energy. Having the memory on the same chip improves the efficiency and speed of each I/O action and removes complexity and uncertainty.

Groq on-chip SRAM has memory bandidth upwards of 80 terabytes/second, while GPU off-chip HBM clocks in at about eight terabytes/second. That difference alone gives LPUs up to a 10X speed advantage, on top of the boost LPUs get from not having to go back and forth to a separate memory chip to retrieve data.

> **The assembly line process within and across chips eliminates bottlenecks. There is no waiting for compute or memory resources to complete a task.**

Conclusion

The LPU is Fast AI Inference

The Groq LPU delivers exceptional speed, affordability, and energy efficiency at scale. Because of its inherent design principles, LPU performance superiority is durable. GPUs will continue to improve their speed and cost, but so will Groq, and at a much faster clip. Our current chip set is built on a 14 nanometer process. As we move towards a 4 nanometer process, the performance advantages of LPU architecture will only increase.

These are the "first principles" at Groq that guide LPU product development. They ensure we will sustain our substantial performance advantage even as GPU manufacturers attempt to close the gap.

About Groq

Groq builds fast AI inference. The Groq LPU™ AI Inference Technology delivers exceptional compute speed, quality, and energy efficiency. Groq, headquartered in Silicon Valley, provides cloud and on-prem inference at scale for AI applications. The LPU and related systems are designed, fabricated, and assembled in North America.