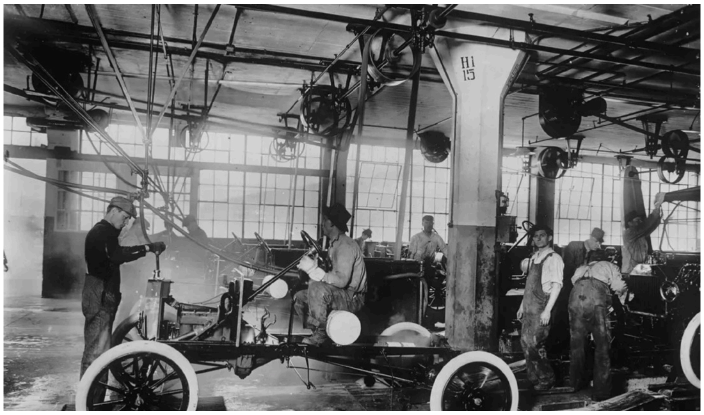
The Groq® LPU™, AI Inference Technology, Delivers More Energy Efficiency Than GPUs for AI Inference. And LPUs Are Faster Too. Here's Why.

Similar to how Henry Ford's assembly line revolutionized manufacturing over a century ago, LPUs are transforming the AI computing landscape by adopting an entirely new, radically more efficient approach to AI inference.

By Santosh Raghavan, Igor Arsovski and Dinesh Maheshwari

In 1913, Henry Ford revolutionized manufacturing by adopting an assembly line approach. Up to that point, cars were worked on in place by a small team of workers, each of whom could handle multiple tasks. The workers and parts went back and forth between workshops and warehouses, wasting a lot of time and energy, while the emerging automobile didn't move.

Then, in 1913, Ford created the modern assembly line. The car moved on a conveyor belt while the workers stayed in place. Over 3,000 parts were assembled into a car in 84 stages. Each stage had exactly what it needed - parts and a one-function worker group - to complete its task and send the emerging car to the next stage. Production time for a Model T dropped from 12 hours to 90 minutes. Ford dropped the price of the car by 65% and sales skyrocketed.



Today, the Groq LPU™ AI Inference Technology works much like the Model T assembly line did. It builds Generative AI (GenAI) tokens in stages, and each stage has exactly the instructions and data it needs to complete the task. A major benefit of this approach is that the Groq LPU runs GenAI models, including LLMs, much faster and more efficiently, from an energy perspective, at an architectural level, than existing solutions such as GPUs.

At an architectural level, Groq LPUs are up to 10X more energy efficient than GPUs. Since energy consumption is one of the biggest cost factors in running AI models, better energy efficiency generally translates to lower costs of operation.

Wow, that's amazing right? Maybe a bit too amazing. For the skeptics out there who think these stats may be too good to be true, let's dive in. How does the LPU run up to 10X more efficiently? (And, let's not forget, with 10X lower latency!)

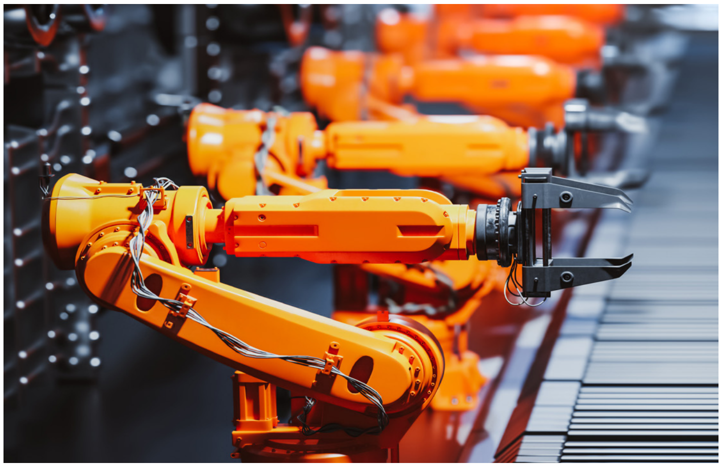
Here's How it Works

The bulk of energy consumption in GenAI computer hardware comes from moving data around, within a chip and especially between chips. The less data a system has to move around, the more energy efficient it is. The LPU is up to 10X more energy efficient than GPUs because Groq's architecture employs a fundamentally different, much more efficient approach to inference computing.

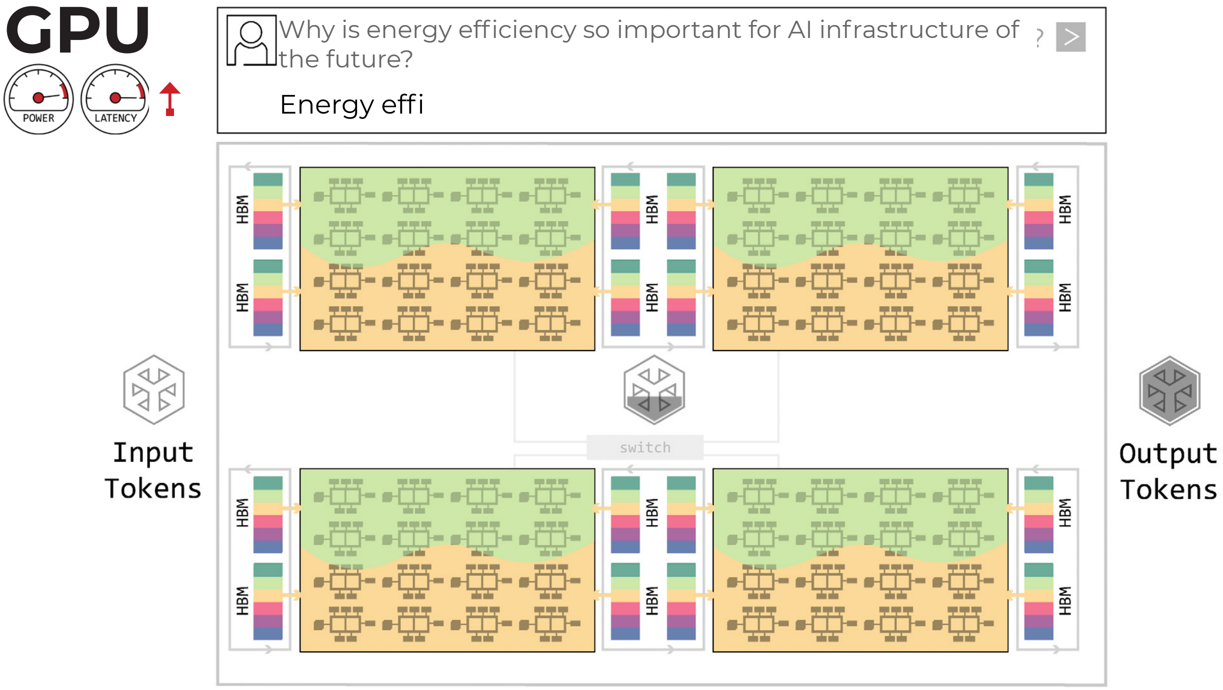
GenAI models, such as LLMs, are composed of a large number of parameters (data). For any given input, such as a query, the model generates an output (a token or series of tokens) via a sequence of compute stages. At each compute stage, the computer chip executes an operation (e.g. an arithmetic calculation) on the output from the previous stage, using the model's parameters of the current stage.

GPUs usually operate in a small team of chips. This team executes all the compute stages required to generate the output. Each stage starts by the GPUs retrieving the complete set of model parameters, along with intermediate outputs (the cumulative outputs of all the prior compute stages), from high bandwidth memory (HBM) storage on a separate chip. Each stage concludes with the chips consolidating the output of that stage and sending it back to the HBM. So every compute stage starts and ends with a lot of data going back and forth to HBM off-chip memory.

> **10X better energy efficiency generally means about 10X lower costs of operation.**



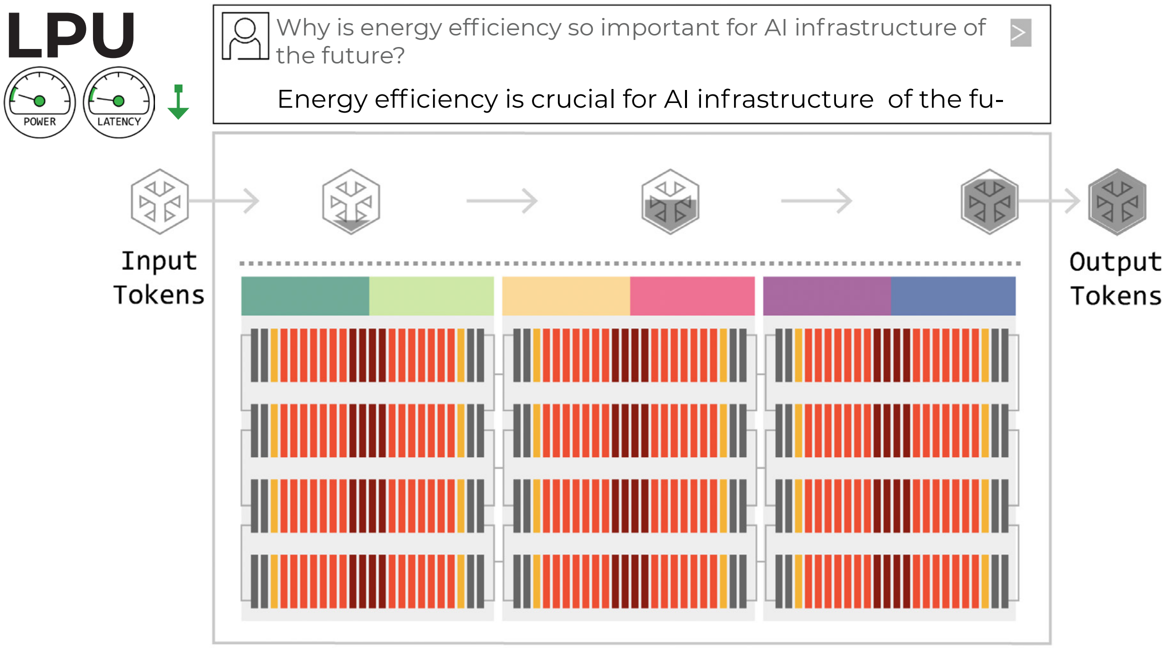
This gets expensive, both in time and energy. Not only does every compute stage require a fresh set of parameters and intermediate results, most of that data needs to come from another piece of silicon. Moreover, all data movement within the small cluster of GPUs needs to be moderated by external routers which sit on yet another piece of silicon. All of these off-chip transactions add up to a highly inefficient design.



Why is energy efficiency so important for AI infrastructure of the future?

Energy effi

Now, let's look at how the Groq LPU does things (spoiler alert: it's better). When a model is compiled to run on LPUs, the compiler partitions the model into smaller chunks which are spatially mapped onto multiple LPU chips. The result is like a compute assembly line. Each cluster of LPUs on this line are set up to run a particular compute stage, and they store all of the data needed to perform that task in their local on-chip memory (called SRAM). The only data they need to retrieve from other chips is the intermediate output that has been generated by either the previous compute stage or the current compute stage. This data transfer is entirely LPU to LPU, requiring no external HBM chips and no external router.



Why is energy efficiency so important for AI infrastructure of the future?

Energy efficiency is crucial for AI infrastructure of the fu-

Why no off-chip HBM? LPUs have enough physical space to embed all memory they need in SRAM directly on-chip. This obviates the need for expensive storage and retrieval trips to off-chip HBM, which generates huge energy cost savings. Getting data from HBM on another chip costs about six picojoules per bit of data, while retrieving it from local on-chip SRAM costs only 0.3 picojoules per bit. That's a 20X energy cost saving!

The Groq assembly line approach to inference is only feasible because the LPU is entirely deterministic. That means that from the moment a new workload is compiled to run on the LPU, the system knows exactly what is happening at each stage on each chip at each moment. Each step in the computation is perfectly synchronized, making each stage of the process as efficient as possible. The assembly line works within a chip and seamlessly across chips as well. There is ample interconnect bandwidth and no need for a network router between chips, as the software knows exactly where the data should go. This perfect determinism is what enables the assembly line to work at peak efficiency.

These various aspects of the Groq LPU is what enables our AI inference technology, at an architectural level, to be up to 10X more energy efficient than GPUs. LPUs are transforming the AI computing landscape by adopting an entirely new, radically more efficient approach to AI inference.

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 Groq LPU and related systems are designed, fabricated, and assembled in North America.