The Groq® LPU™ Inference Engine

10x More Energy Efficient Than GPUs. 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 generation.

Authors:
Santosh Raghavan, Igor Arsovski &
Dinesh Maheshwari

The Groq LPU™ Inference Engine runs Generative AI (GenAI) models, including Large Language Models (LLMs), 10X more efficiently, from an energy consumption standpoint, than existing GPU solutions. This means that all other factors being equal, running an LLM or other type of generative AI model on the Groq Inference Engine consumes less than 1/10th the energy of running it on a GPU. Since energy consumption is one of the biggest cost factors in running AI models, 10X better energy efficiency generally means about 10X 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 Groq LPU Inference Engine run 10X more efficiently? (And, let's not forget, with 10X lower latency!)

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 10X more energy efficient than GPUs because Groq 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.**

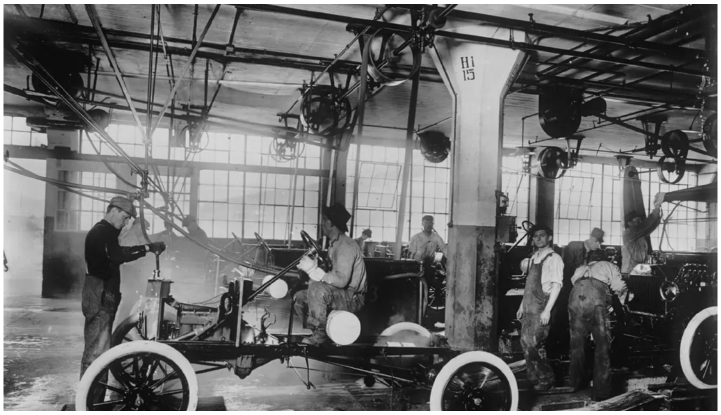
This gets expensive, both in time and energy. Not only does every compute stage require a fresh set of parameters and intermediate results, most of that data needs to come from another piece of silicon. Moreover, all data movement within the small cluster of GPUs needs to be moderated by external routers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

Now, let's look at how the Groq LPU Inference Engine does things (spoiler alert: it's better). When a model is compiled to run on Groq 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 no off-chip HBM? LPUs have enough physical space to embed all memory they need in SRAM directly on the 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!

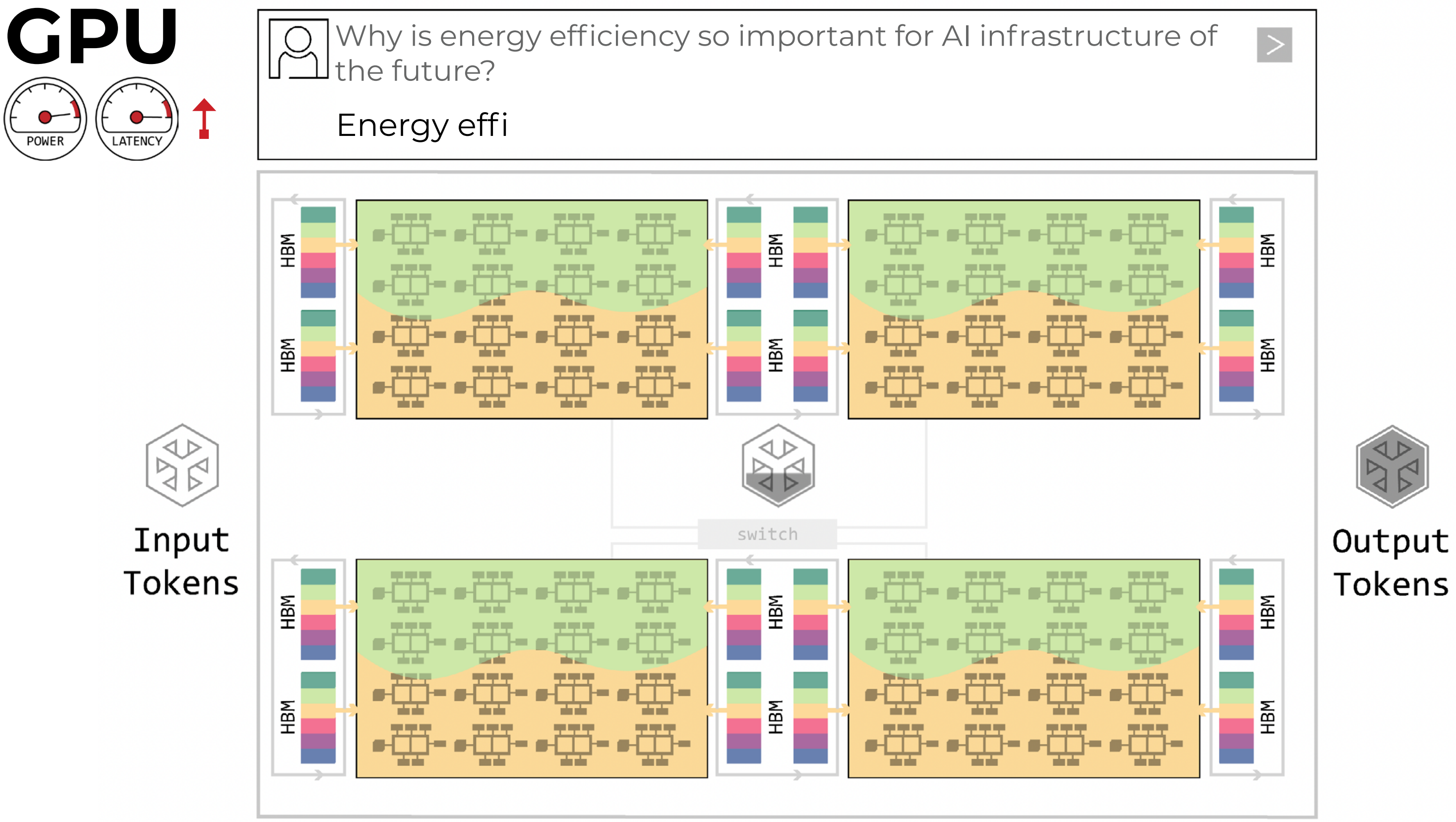
The Groq assembly line approach to inference is only feasible because the Groq LPU Inference Engine is entirely deterministic. That means that from the moment a new workload is compiled to run on Groq, the system knows exactly what is happening at each stage on each chip at each moment. Each step in the computation is perfectly synchronized, making each stage of the process as efficient as possible. This is why, for example, no router is required to move data between LPU chips. The chip sending the data already knows exactly where it needs to go. It doesn't need a router to direct it. This perfect determinism is what enables the assembly line to work at peak efficiency.

Manufacturing a Token



The difference between how the Groq LPU Inference Engine works versus GPUs is analogous to how Henry Ford manufactured the Model T versus how cars were previously built.

Prior to 1913, the manufacturing process was GPU-like. Cars were built 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.



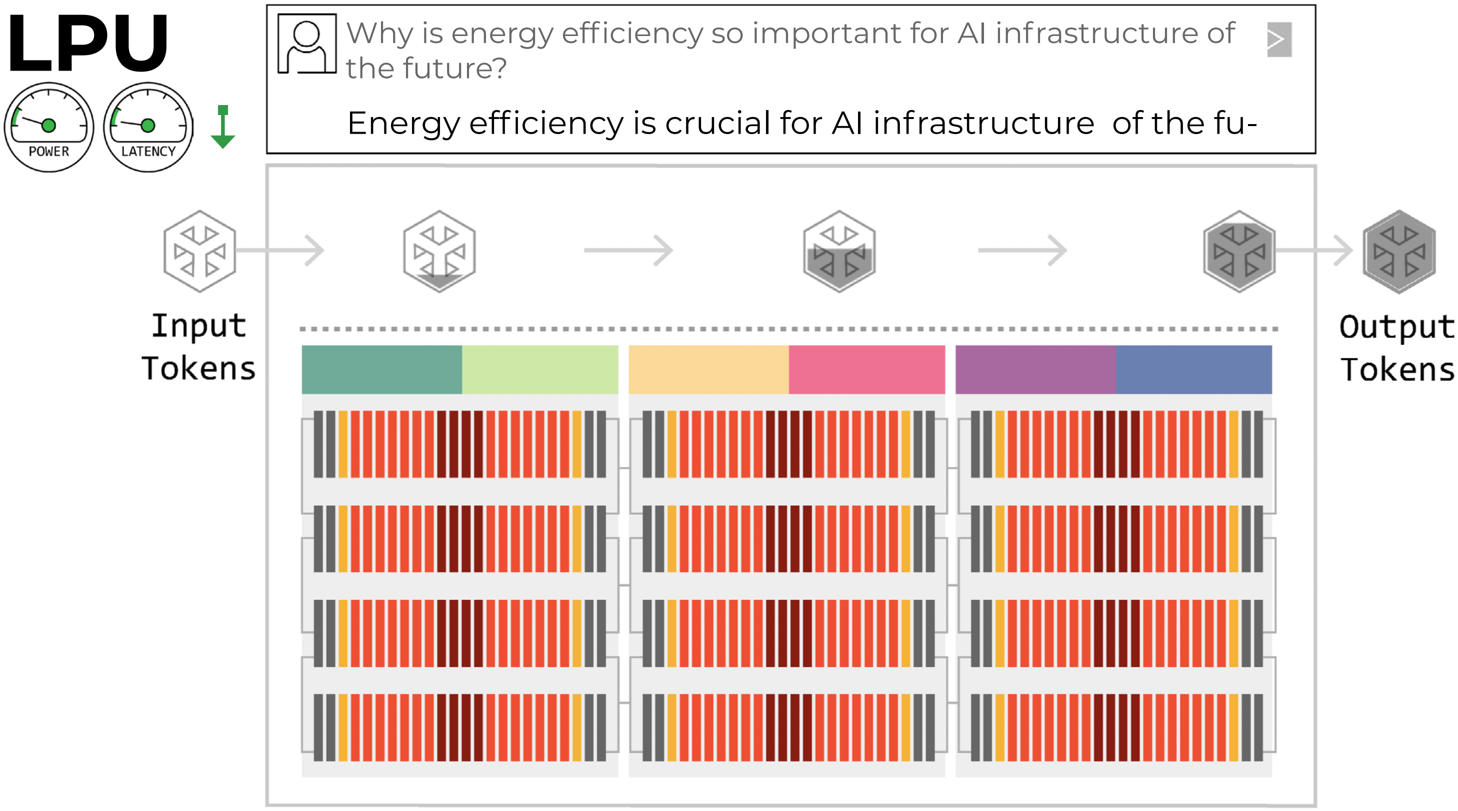
Why is energy efficiency so important for AI infrastructure of the future?

Energy effi

Then, in 1913, Ford adopted the moving assembly line method of building cars. The car moved on a conveyor belt while the workers stayed in place. The Model T comprised 3,000 parts, and the line assembled those parts 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.

The Groq LPU Inference Engine runs like the Model T assembly line.

The model is mapped across the system so that each LPU has a specific task and the right "parts" (data) needed to complete that task. On the Ford assembly line, workers didn't need to go back and forth to a central location to retrieve the parts they needed; they had everything right there and the car came to them. On a Groq LPU, each stage doesn't need to retrieve a bunch of data from HBM; it already has what it needs and it knows exactly where it fits in the compute sequence. This is a far more efficient design.



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

Energy efficiency is crucial for AI infrastructure of the fu-

About Groq

Groq® builds the world's fastest AI inference technology. The LPU™ Inference Engine by Groq is a hardware and software platform that delivers exceptional compute speed, quality, and energy efficiency. Groq, headquartered in Silicon Valley, is committed to serving the U.S. Public Sector and U.S. allies with cloud and on-prem solutions at scale for AI applications. The LPU and related systems are designed, fabricated, and assembled in North America.