Bringing Speed to Mission with the Groq™ LPU Inference Engine

Groq® is a generative AI (GenAI) solutions company. Our LPU™ Inference Engine runs GenAI applications at **10X better speed and precision, with 10X better energy efficiency**. This performance is foundational for real-time AI solutions that help people do what they do, much better. Real-time, interactive, conversational solutions in fields such as customer service, data analytics, cybersecurity, and software development can leverage “wow” performance by Groq to put the power of AI knowledge and insights at the fingertips of human experts in the moment.

Without **10X speed**, these AI solutions aren’t feasible. With it, they have the power to transform missions and tackle big challenges.

Who is Groq?

Groq is a GenAI solutions company. Our LPU Inference Engine runs GenAI applications at 10X better speed and precision, opening the door to an entirely new class of real-time AI solutions that will transform organizations and solve big challenges.

Why does Groq exist?

Current solutions for deploying GenAI are slow and expensive. For starters, getting trained LLMs ready for inference (compiling) on current hardware is slow and resource-intensive, tremendously hindering the pace of innovation. The market of available solutions simply cannot support real-time inference, i.e. running AI at conversational speeds and better. Finally, most current solutions are sourced outside the U.S.

What does Groq do?

The Groq LPU Inference Engine is designed and assembled in North America and is a software-first solution for compiling and running GenAI, especially LLMs, at scale. Our primary benefits are superior performance, pace, and energy efficiency.

- **Performance:** LLMs running on the Groq LPU Inference run at speeds up to 10X faster than other inference solutions, at scale. This enables a whole new class of real-time AI solutions. See our latest public LLM inference benchmarks from Anyscale and ArtificialAnalysis.ai[1].
- **Pace:** Compiling LLMs to run on Groq takes hours or days, not months, and requires minimal engineering resources. You can get a model from training to inference at a fraction of the cost and time, accelerating the pace of innovation.
- **Energy:** The Groq LPU consumes about 1/10th of the energy on a per token basis compared to GPUs.

Really? How can you be 10X better?

The Groq architecture is fundamentally different from that of current GPU-based solutions. It was designed from the outset to be software centric and support AI models.

Our chip's clean design enhances speed and efficiency. It has ample SRAM on board, minimizing the number of calls required to retrieve data from memory on another chip. It is software-centric: the software schedules and controls all activity on the hardware, so the program knows exactly how long it takes to execute. The compiler is kernel-less, so compiling a model to run on Groq does not require manually writing custom kernels..

Add this up, and you have a new type of inference solution - the LPU inference engine - that delivers 10X better performance, efficiency, and pace.

> **The Groq architecture is fundamentally different from that of current GPU-based solutions. It was designed from the outset to be software-centric and support AI models.**

Wow, 10X faster is great. Now what? Why does that matter?

Groq 10X performance enables an entirely new class of AI solutions. Real-time AI assistance, running on Groq, can empower people to make optimal decisions in the moment, helping them do what they do, much better. A defining characteristic of these solutions is that they are real-time. They work in concert with humans almost seamlessly, assisting exactly when needed. There is no delay; people using AI assistance stay completely "in the flow."

These real-time AI assistance solutions are not feasible with incumbent AI solutions, which are too slow.

1. Read more about our benchmark results at groq.link/anyscaleblog and groq.link/aabenchmark.

What could I do with Groq real-time inference that I can't do with GPUs?

Assume you have a call center and a customer reaches out with a question about a product or service.A real-time AI assistant, running on Groq, listens in on the conversation between a customer service agent and the customer. As the agent is talking to the customer and going through their standard scripting options, the assistant listens to the conversation, understands the customer's situation, reviews the customer's data, and provides the agent timely suggestions to help solve the customer's problem.

Relying on the assistant's help, the agent is able to ask better questions to get to the heart of the problem. The AI assistant employs sentiment analysis to coach the agent in matching and addressing a customer's mood. The assistant's suggested responses could be more understanding and sympathetic for a deeply frustrated customer, or more playful for a friendlier one. The assistant provides consistent and clear suggestions to the agent, reducing the problem customers may experience of getting different answers from different agents.

As the agent, with the AI assistant's help, solves the customer's problem, the assistant may suggest additional steps the customer can take to improve their experience, or additional products or services that may be helpful. These suggestions are tailored to meet the customer's background, history, and current sentiment.

When the call concludes, the assistant gives the agent a quick suggestion or two about what to remember about and learn from the call, along with a well-earned pat on the back. The real-time AI assistant has helped the agent solve the caller's problem faster and better, provided targeted suggestions for additional sales, and offered coaching advice so the agent can learn and get better. All in real-time.

Imagine how this sort of real-time AI assistant could transform call centers, making them more responsive, more successful in addressing customer issues and introducing them to new services, and much more efficient. All without adding personnel.

Here's another example. Think about the job of an analyst in a fast moving environment, spending their day sifting through large and disparate data sets, looking for patterns and insights that inform decisions and actions. In some cases (e.g. financial markets, cybersecurity), they need to draw their conclusions very quickly, which is very challenging given the breadth and complexity of the data at their fingertips.

Now suppose a real-time AI analytical assistant works alongside the analyst, continuously integrating numerous datasets of different types and modalities. It proactively notices patterns and situations that demand further attention, and flags them for the analyst. The analyst converses with the assistant to delve into areas demanding further attention. To help the model find signal in noise (or a needle in a haystack), the analyst employs different conversational frameworks. These are algorithmic approaches to prompt engineering that help the human get optimal insights from their real-time assistant as quickly as possible.

As the analyst discovers a situation requiring immediate attention or action, the AI assistant helps the analyst concisely describe the situation, including recommended actions, relevant information, and access to additional information. The analyst has the opportunity to challenge the assistant's findings and reporting, and vice versa. Each gets better as a result of their conversation.

After the analyst provides recommendations, insights, and supporting information to their colleagues, they engage their AI assistant in a conversation about how they could improve the process. Working in concert with a real-time AI solution, the analyst can move faster, develop deeper, more nuanced insights, and learn and get better.