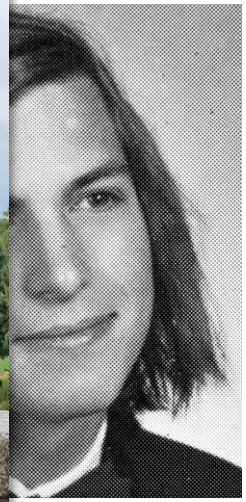

groq
World
Summit AI
09-10 October 2024 | Amsterdam



THE SPEED OF ITERATION

=

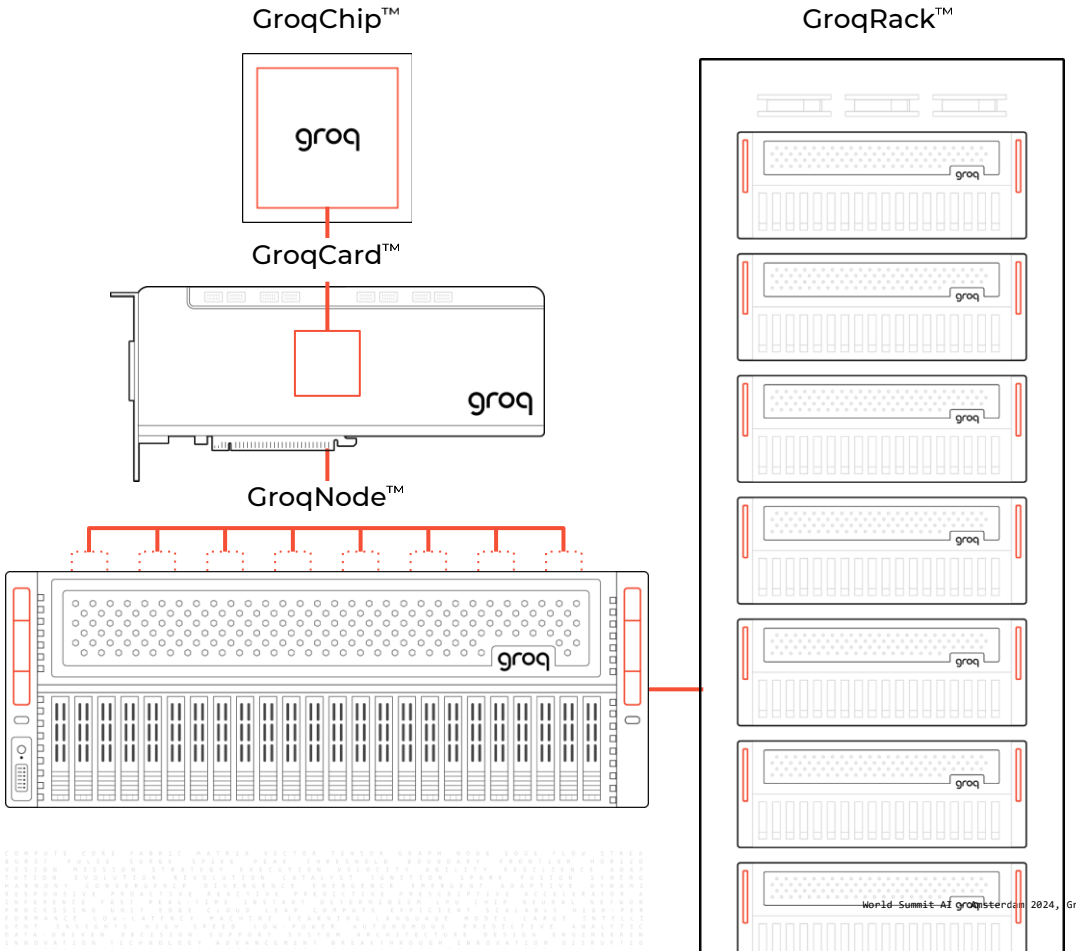
THE SPEED OF INNOVATION

groq

Full Stack Innovation

Hardware + Software

Designed, Engineered, and Fabbed in North America



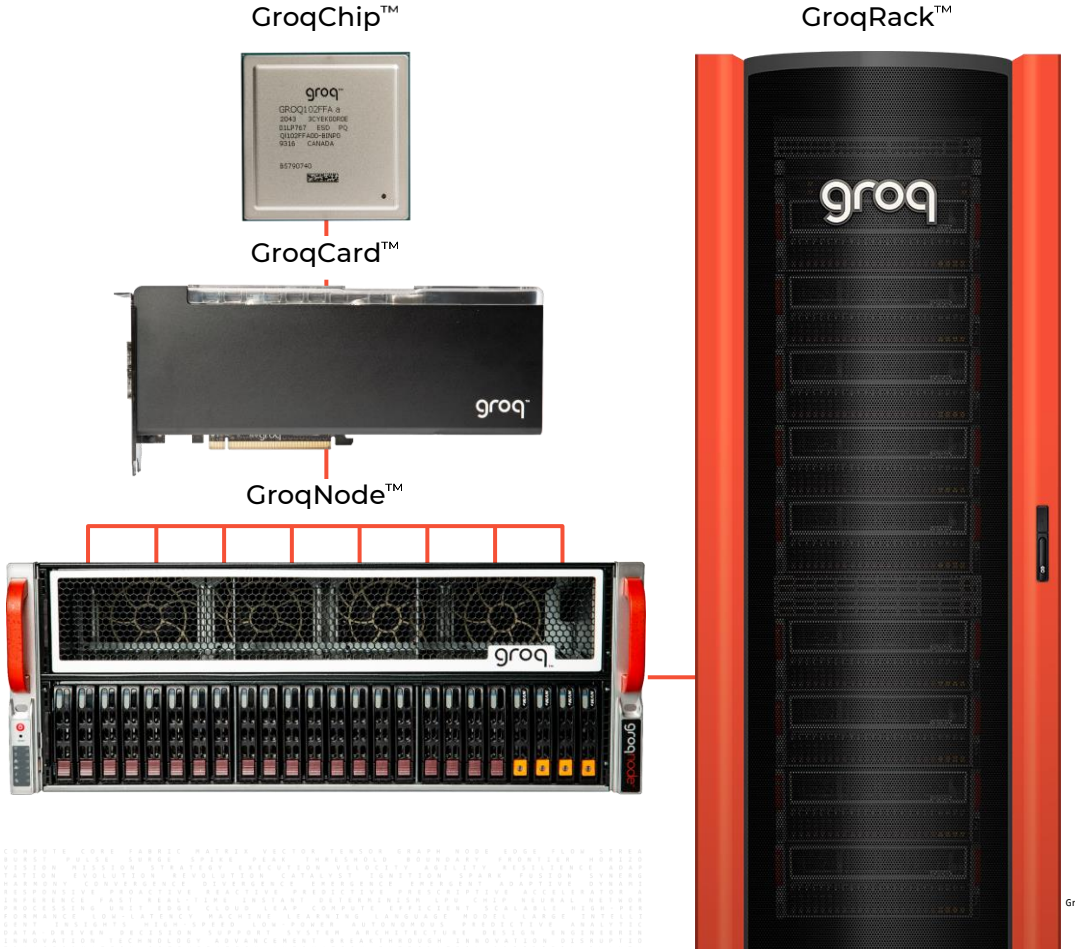
Full Stack Innovation

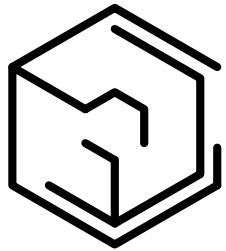
Hardware + Software

Designed, Engineered, and Fabbed in North America

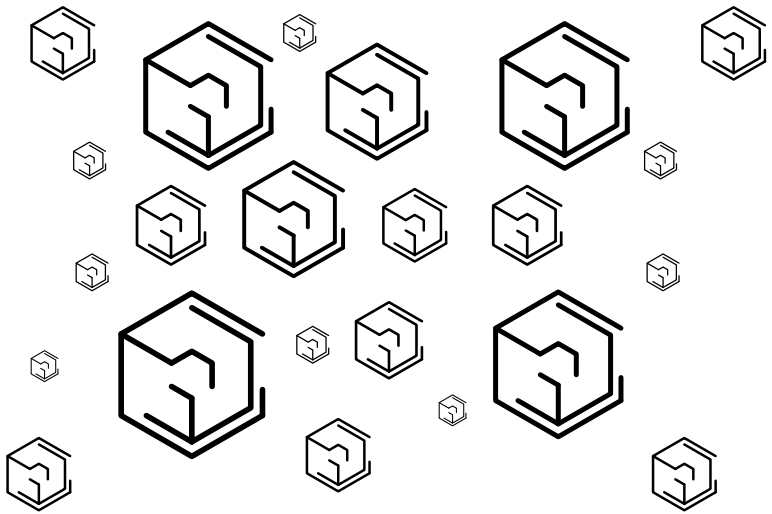
LPU™ - Language Processing Unit.

The AI inference technology is designed to scale and is more power-efficient, with greater performance, than a GPU for AI applications like LLMs when used at scale.

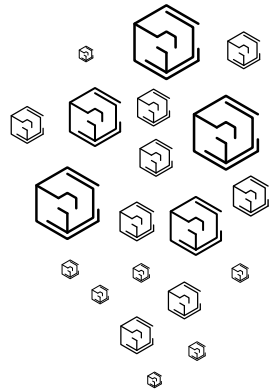




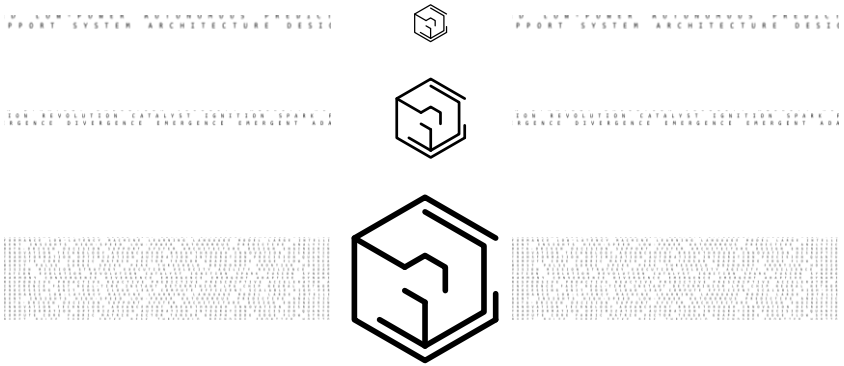
Model



Training



Training

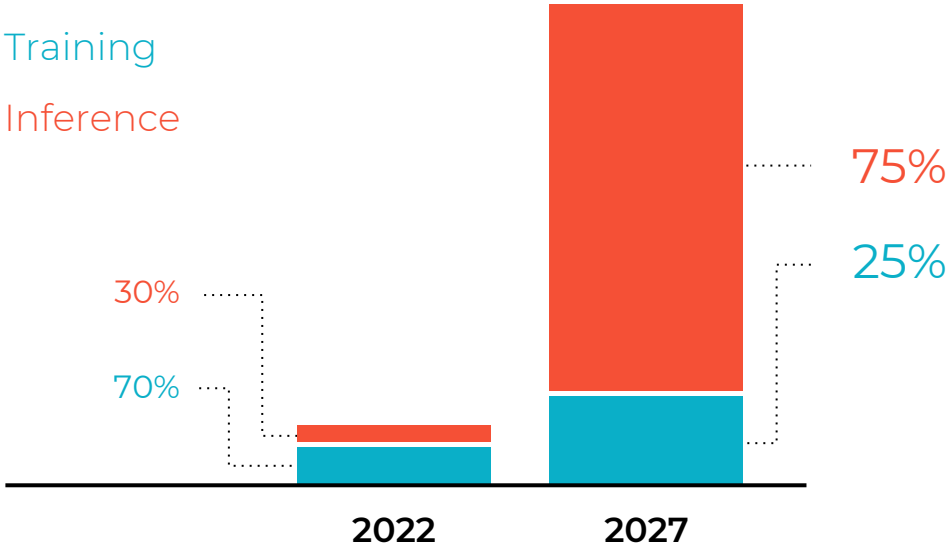


Production

We're at a tipping point

Training ↦ Inference

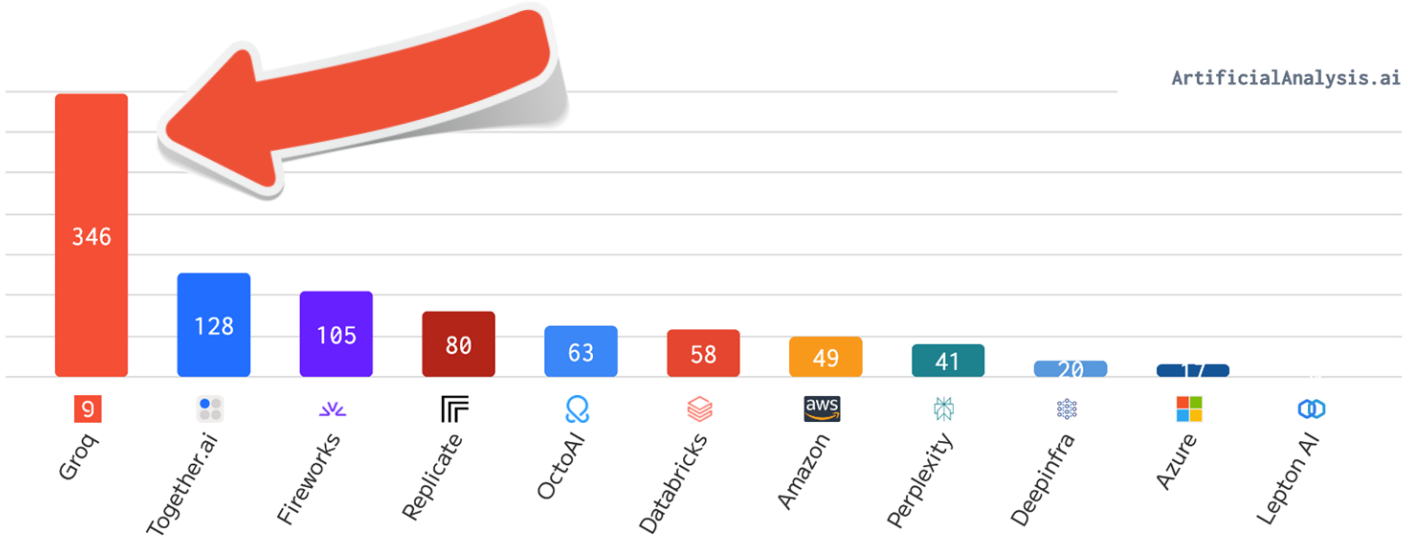
AI Semiconductors Deployed in Data Centers



Gartner, Forecast Analysis AI Semiconductors Worldwide, April 2023

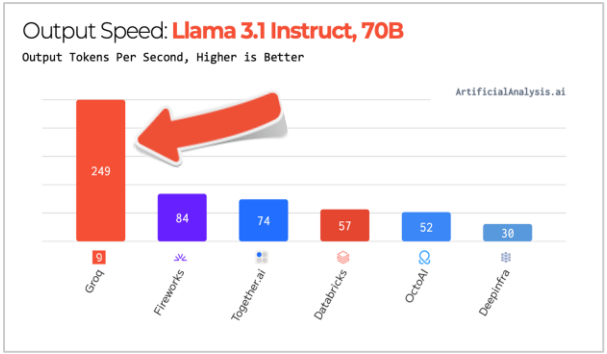
Output Speed: Llama 3, 70B

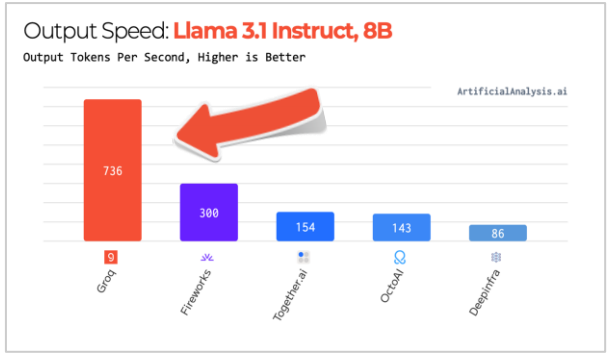
Output Tokens Per Second, Higher is Better

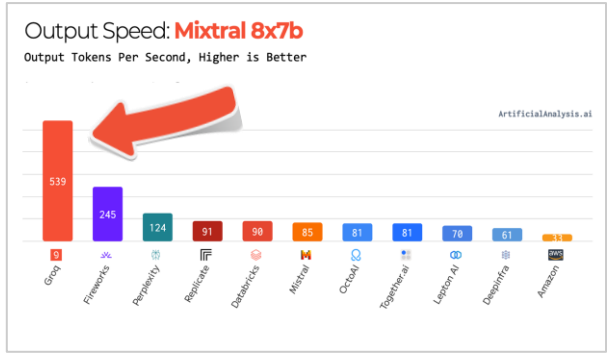


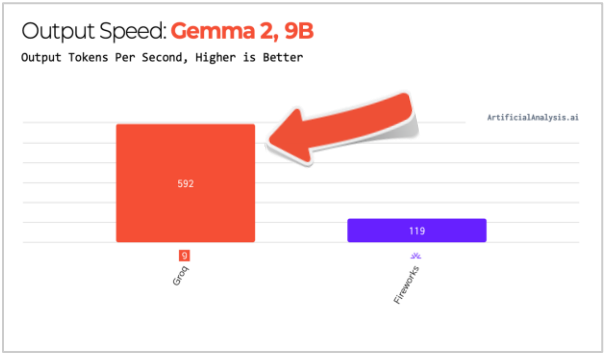
Independent Benchmarks: Inference Speed

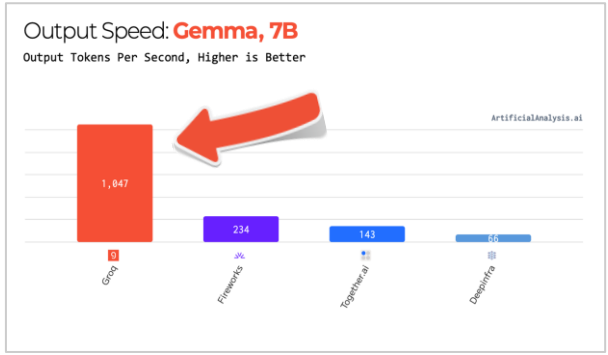
ArtificialAnalysis.ai

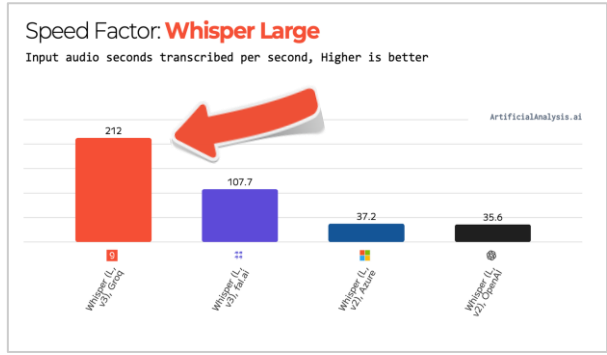


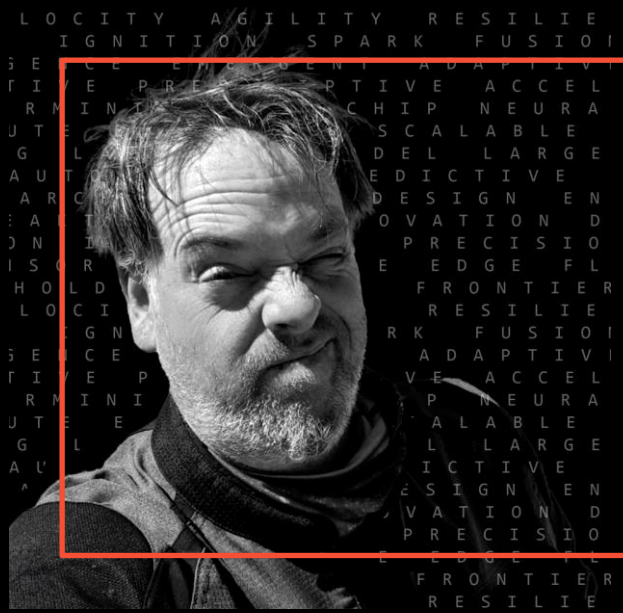






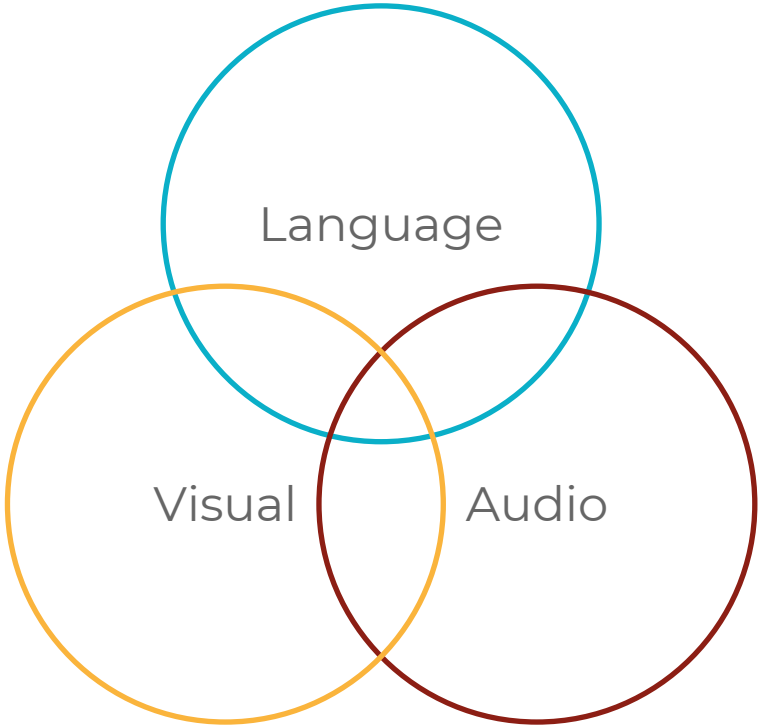






Mark Heaps

Chief Tech Evangelist
VP of Brand



AI Diverse Ecosystem

More complex applications will continue to be created

<Demo>

Demo

POWERED BY groq

Mark Heaps 5:11 PM
I really wish Vision had a batch capability where I could put in 5 images and have it write descriptions for us to all 5 and then give a summary of the collective

That would be different enough to reduce imagination gaps

for potential use cases

like CCTV, batch metadata generation, e'

Benjamin Klieger 5:12 PM
Oh, I can build that

Mark Heaps 5:12 PM
Video frame accessibility extraction

Really?

Benjamin Klieger 5:12 PM
It's just looped API calls (edited)

That would be simple actually

Mark Heaps 5:12 PM
Don't tease me Ben

I'm 48 hours away from presenting

lol

Benjamin Klieger 5:13 PM
This is the one thing that's actually so simple I can do it right now and be confident for the demo

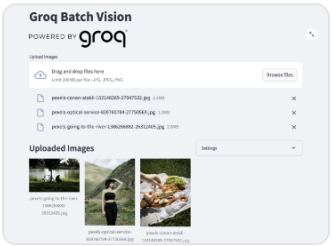
Mark Heaps 5:23 PM
Just mocking to align our thoughts.
Foundationally you have a drag and drop or paste from clipboard image load area. Like in Console now.
then you see all the images, and then you have a prompt area where you can ask it questions about the images or for a description of each image, etc...

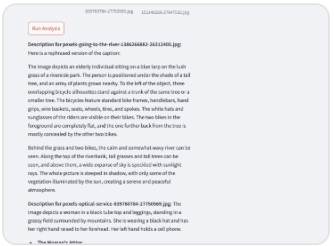
2 files

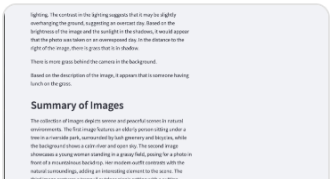
Then ideally, a copy button from the results area

Benjamin Klieger 5:41 PM
Bringing online for you to play with

3 files







<Demo>

Demo

POWERED BY groq

GroqChat™

GroqCloud™

Experience the speed and join the community yourself at Groq.com for free.


groq
World
Summit AI
09-10 October 2024 | Amsterdam