SynoGate ThinkTank

The ideal hardware for LLM inference: the most cost-efficient solution available, compatible to all popular models and OpenAI’s API, and scalable from SME to cloud and enterprise setups.

As more and more businesses adopt LLM inference, two factors can stand in the way to success: availability of affordable hardware, and data privacy*.

We offer dedicated hardware for on-premise LLM inference, optimized for efficiency and ease of use. It can be customized to your needs, with 200% of the performance per dollar compared to NVIDIA’s DGX H100.

https://medium.com/@davidfagb/challenges-facing-llm-tools-and-solutions-9c5802939054

The **ThinkTank**

Dedicated plug-and-play hardware for LLM inference, on-prem, fully within your control.

No maintenance or adaptation of models or applications is required. Connect to your own LLMs or any model of your choice via REST API, just like OpenAI’s API.

It is assembled in Germany by Thomas-Krenn.AG with boards manufactured by Prodesign. Synogate’s circuit design runs on programmable semiconductors from Intel.

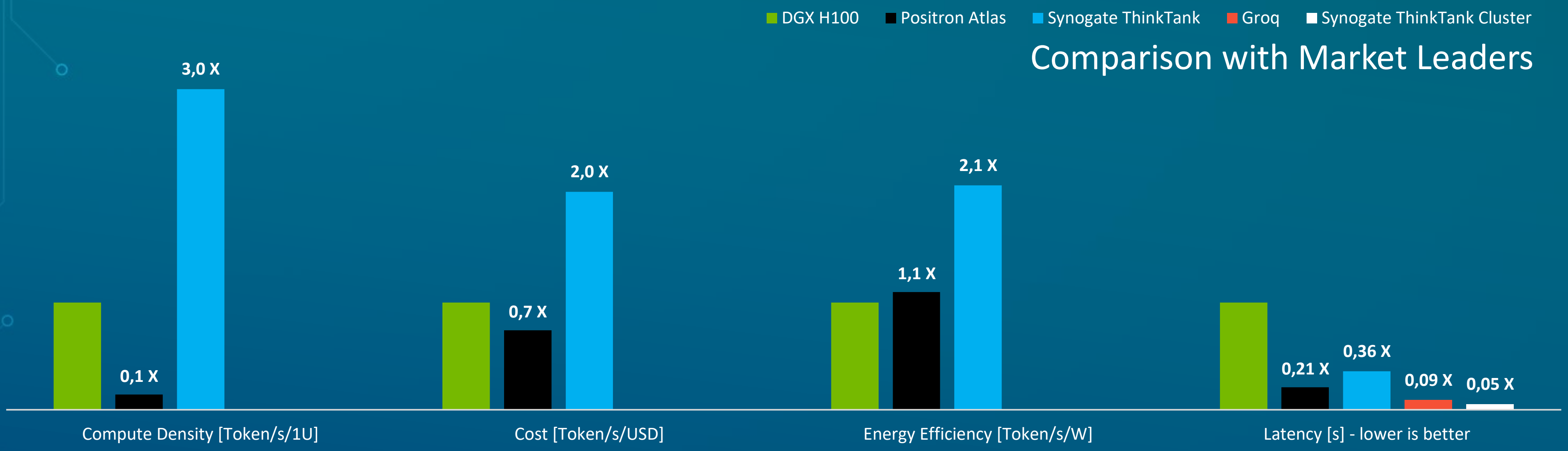
The **ThinkTank** – Made In Germany

- Form factor: 1U rack-mountable server
- 144 GiB of HBM
- Cost: ~ € 80k
- Power intake: ~ 1.3 kWh*
- Availability: from Q2/2025* *through Thomas-Krenn.AG*

The ThinkTank is in development; values marked with * are our current projections.

Sales and distribution will go through our partner, Thomas-Krenn.AG, a German hardware manufacturer with high production capacity.

Benchmarks


Comparison with Market Leaders

2x performance per USD compared to NVIDIA’ DGX H100

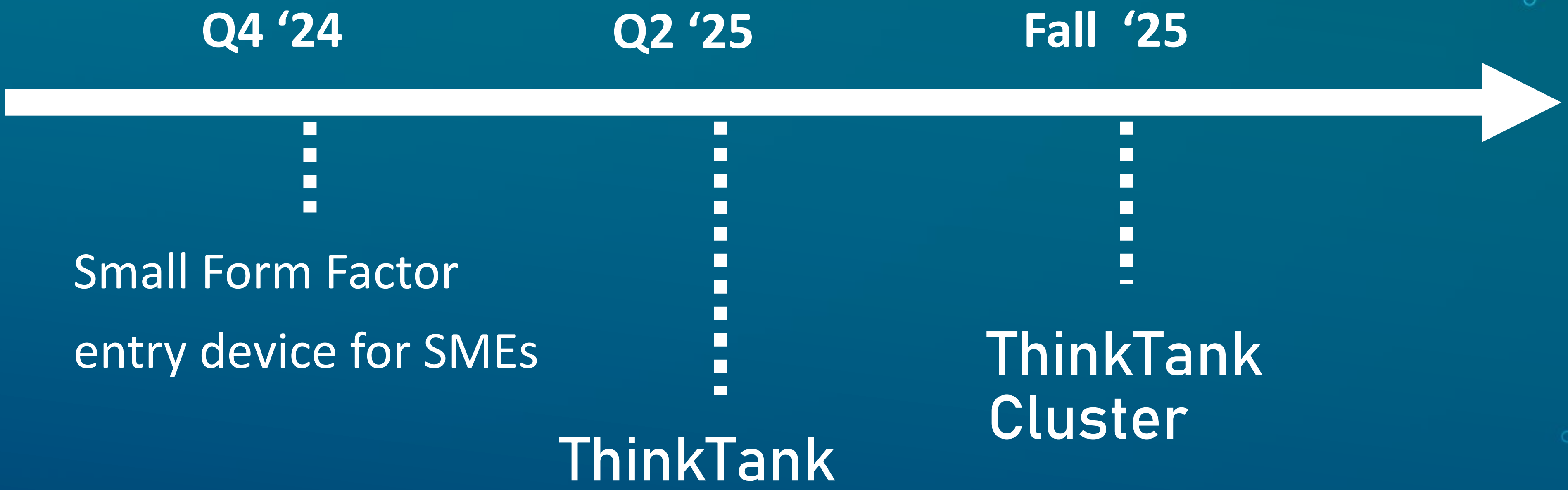
- *All numbers refer to Llama 2 70B*
- *Latency is measured in seconds until all tokens are computed*
- *Benchmark uses queries of 1024 input tokens, and 512 output tokens*
- *Not all metrics available for Groq*

Confidential

The **ThinkTank Cluster**

Ultimate, future-proof customizability: scale up performance by adding ThinkTanks to meet increasing workloads, or combine ThinkTanks into a ThinkTank Cluster. Optimized for ultra-low latency, the ThinkTank Cluster is tailored to the needs of long-running RAG/agent-framework-based applications.

Timeline



Q4 ‘24

Small Form Factor

entry device for SMEs

Q2 ‘25

ThinkTank

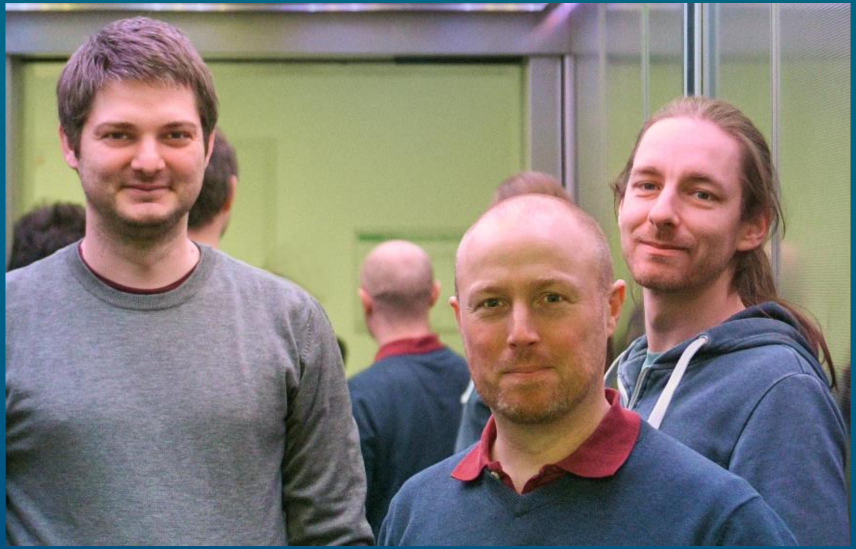
Fall ‘25

ThinkTank Cluster

Why we can do this

- We develop hardware-accelerated cybersecurity and highspeed networking solutions
- We combine expertise in semiconductor design, machine learning, software development, and cybersecurity
- Our Hardware Construction Library “Gatery” allows for fast, flexible circuit design

Andreas Ley
Software Development



Michael Offel
RTL Design

Philipp Keydel
Business Development

SYNOGATE

- Visit our Website:
  - https://www.synogate.com/products/llm-accelerator.html
  - more products & solutions
- CONTACT US FOR MORE INFORMATION
  - mail@synogate.com

Thank you.