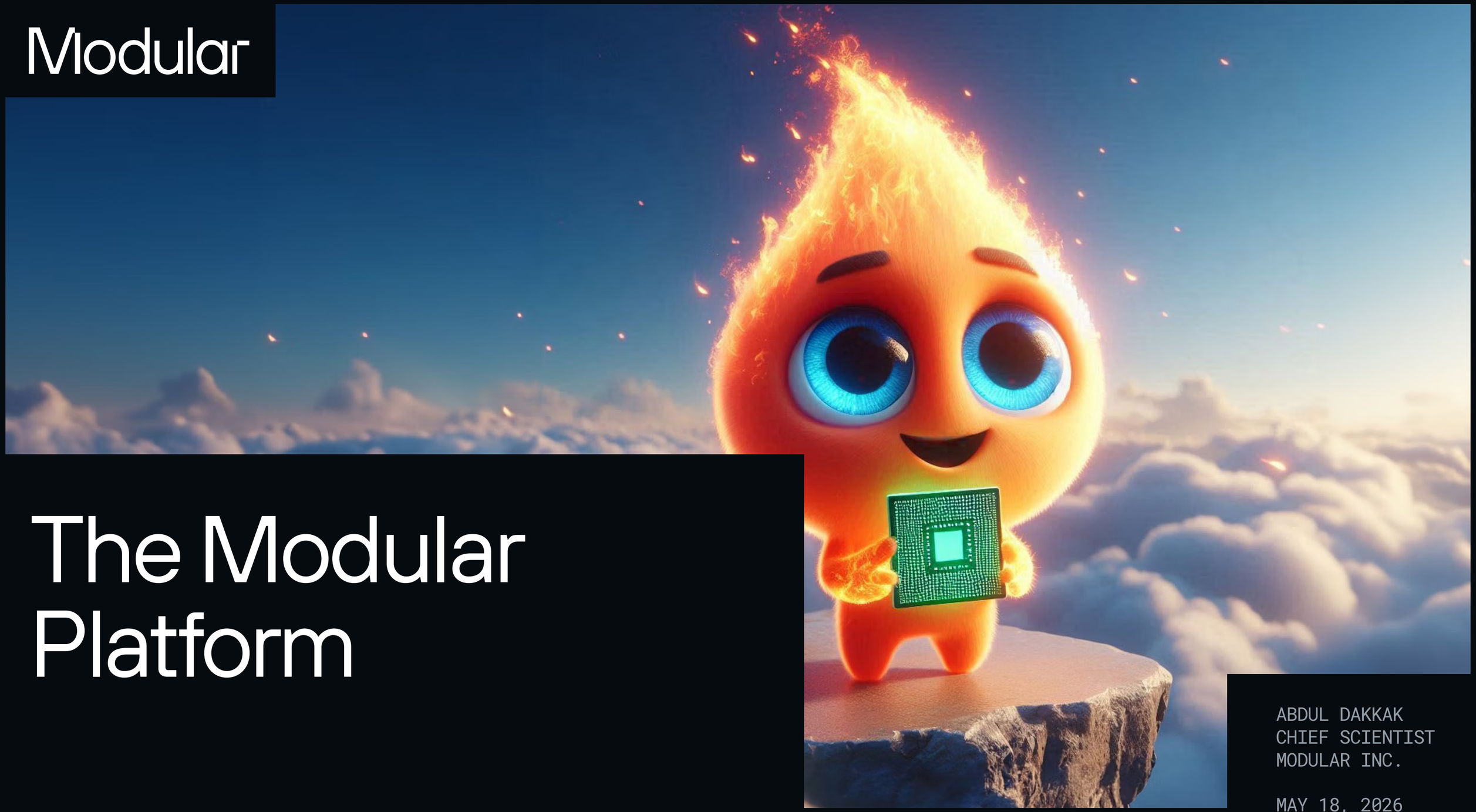
Modular



The Modular Platform

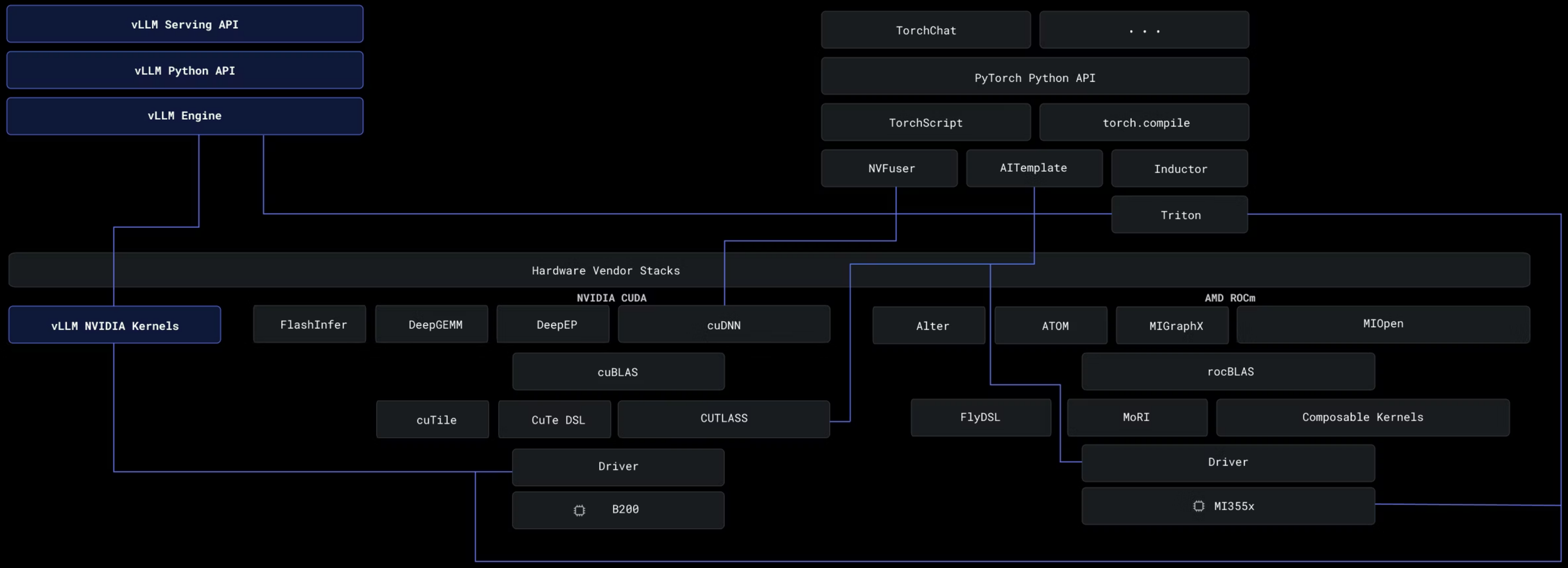
ABDUL DAKKAK
CHIEF SCIENTIST
MODULAR INC.

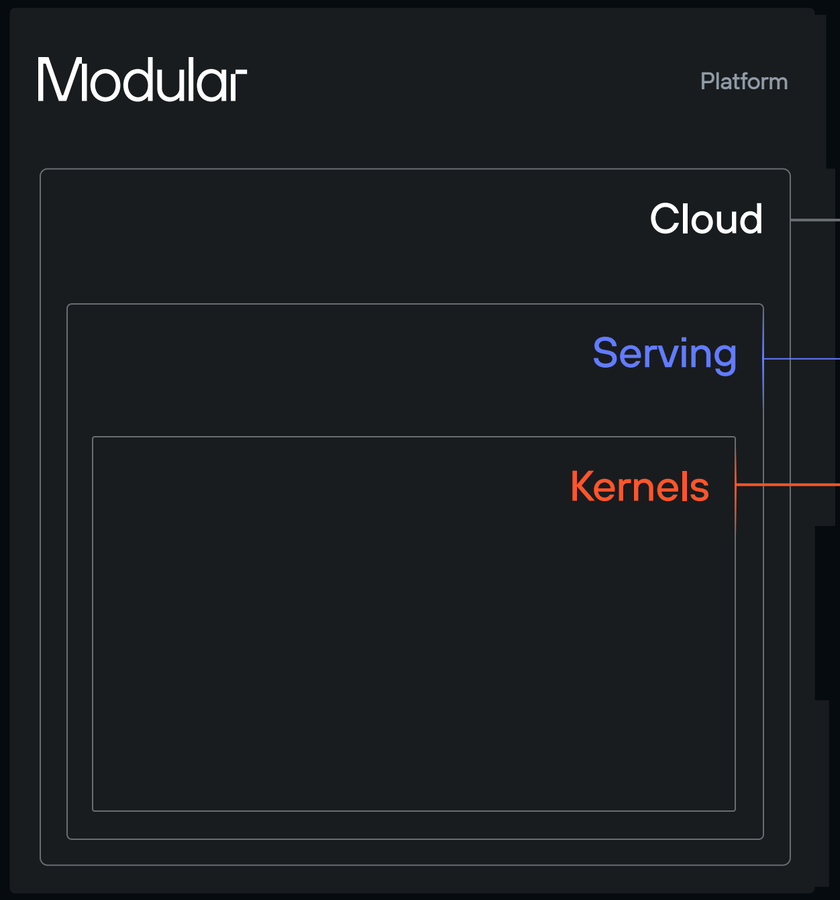
MAY 18, 2026

AI Hardware is exploding and AI software is not scaling

The problem is at the bottom.
Old technologies tied to individual vendors' hardware.

This forces complexity into AI systems upward to support different hardware options.





MCloud ☁️

A performance-native, “glass-box” cloud that turns MAX and Mojo into production GenAI endpoints.

ENDPOINT | CLUSTER | LAB

MAX 🧑‍🚀

Serve GenAI models with SOTA Perf on NVIDIA+AMD GPU & CPUs with one container & OpenAI Python API

SERVING | MODELS | 🔥 KERNELS

Mojo 🔥

Enables novel AI algorithms with a Pythonic systems language that runs on AI compute hardware with cutting-edge tooling

COMPILER | LIBRARIES | TOOLS

01 Switched to MAX because couldn't meet their targets as they scaled with existing solutions.

02 MAX delivered sub-500ms TTFT keeping every conversation instant for clinicians & patients.

03 30% faster P99 end-to-end latency eliminating latency spikes that break trust.

04 22% faster mean end-to-end latency letting Hippocratic scale to more patients per node

Hippocratic AI

— Do No Harm —

[CUSTOMER]

Modular's SOTA Perf on MAX

Kimi-K2.5 (vs. vLLM)

B200

5.5x P50 TTFT

1.5x Throughput

Gemma-4-31B-it (vs. vLLM)

B200

2.5x P99 TTFT

1.5x Throughput

Flux.2-dev (vs. PyTorch Diffusers with torch.compile)

B200

6.9x faster generation

MI355x

3.8x faster generation

Get involved

01

Join our OSS community

Explore open-source tools for portable, high-performance compute across hardware.

github.com/modular/modular

02

Explore AI Skills

Learn modern GPU patterns, optimizations, and workflows designed to go beyond CUDA.

github.com/modular/skills

03

Join the Team

Come and be part of a world-class team that is rebuilding AI for everyone - we have roles open across product and engineering for our entire stack.

modular.com/company/careers