

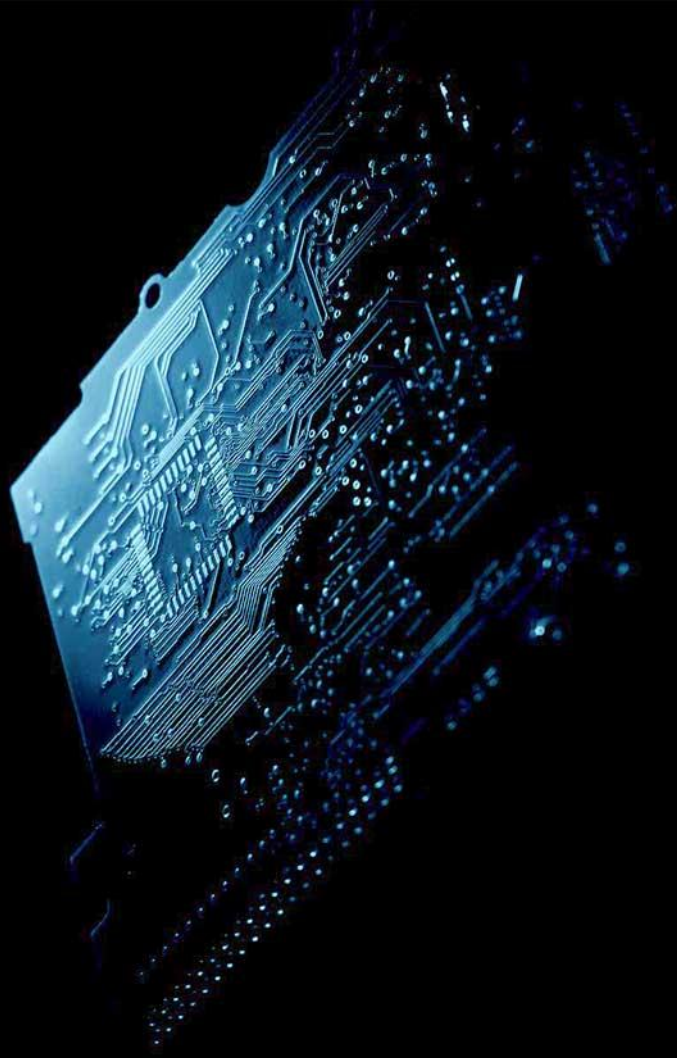


Infrastructure for the AI Era:

Enabling Enterprises to Upgrade from
Hardware and Software to Architecture and AI
Applications

Speaker

Fred Jhang, Lead of CE Front End Engineering, GMI Cloud



We all know AI is important,
but the real question is —

how do we actually start?

Who we are: Introducing GMI Cloud



One out of 6
Nvidia Reference
Platform Cloud
Partner



APAC
Focused



AI native
GPU
Platform

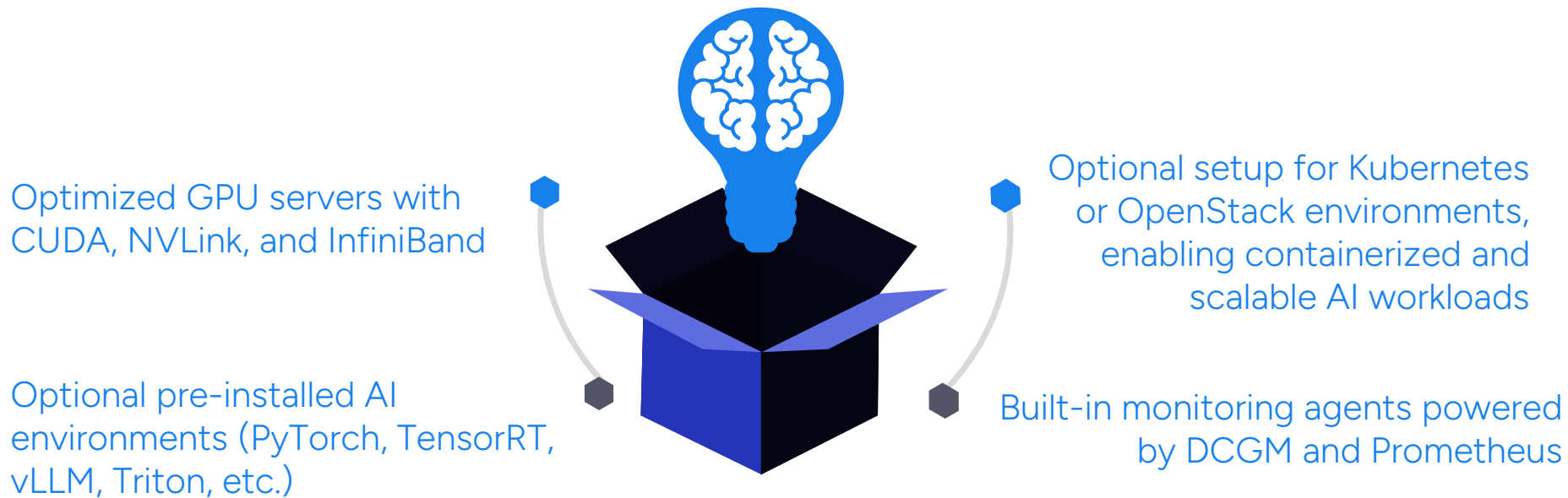


Full-stack
Cluster Engine
Inference Engine

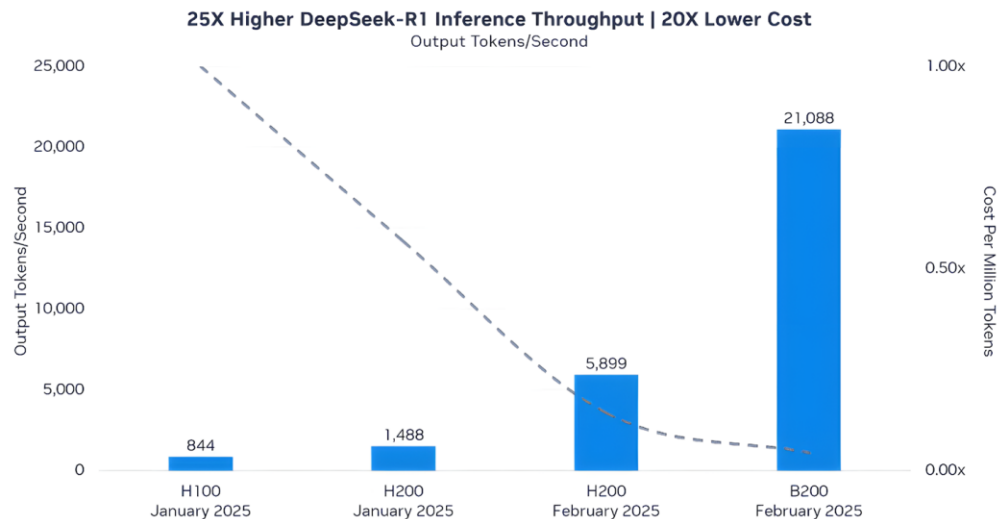


Preferred
Partner

Level 1: Hardware Layer



Always on the Lookout for Best Hardware



Throughput tested on single
8 x H100 、 H200 、 B200
server node

*Data provided by NVIDIA DeepSeek-FP4

B200 is more than 25x
performant than H200. Latest
hardware hosts latest model
better.



Hardware upgrades is one of the easiest way to improve performance.

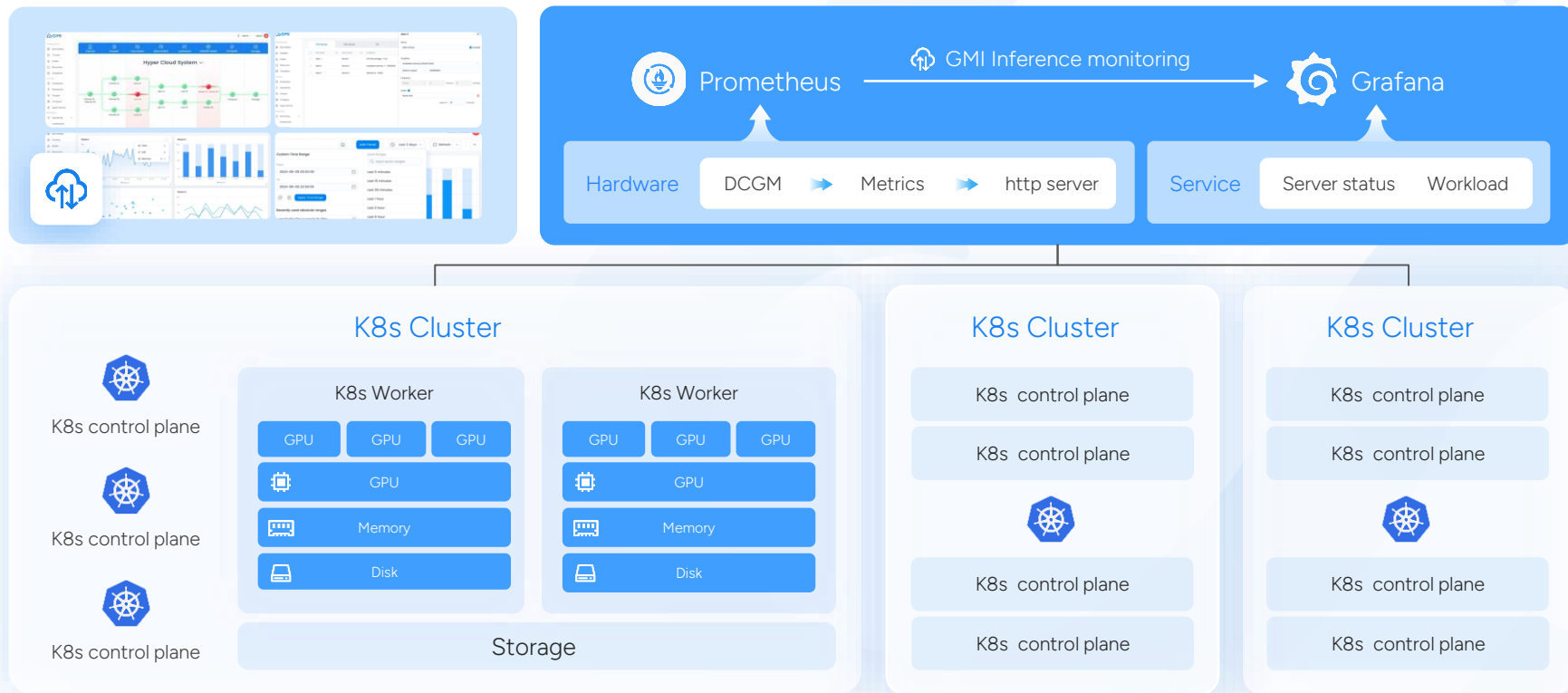
Level 2: Management Layer

- **Cluster Engine** provides both Web and API control interfaces
- One-click creation of Bare Metal, Container, or Kubernetes Clusters
- Integrated monitoring and health checks

The screenshot displays the GMI (Gigamon Inc.) Cloud Management Interface. On the left is a sidebar with navigation options: Bare Metal, Bare Metals, Firewalls, VPC & Subnets, Elastic IP, Container (highlighted), Containers (selected), and Templates. Below these are links for Reservation, Docs, and Company. The main panel is titled 'Containers' and features a search bar labeled 'Search container name'. Below the search bar is a table listing containers. The table has four columns: Name, Template, Data Center, and Spec. Two containers are listed, both using the 'container.l40S.x1' template in the 'tw-lab-1' data center. The first container has ID 'container-20251016-154424' and the second has ID 'container-20251020-133236'. Both specify 'L40S GPU N' and '16 CPU | 128' in their specifications. At the bottom right, there is a pagination control showing '< 1'.

Name	Template	Data Center	Spec
container-20251016-154424 1fab707b-63ce-4494-bb77-fd...	container.l40S.x1	tw-lab-1	L40S GPU N 16 CPU 128
container-20251020-133236 e828850d-635c-4498-add1-f2...	container.l40S.x1	tw-lab-1	L40S GPU N 16 CPU 128

Global End to End Monitoring



Level 3:

Model Serving Layer

- **Inference Engine** provides a unified platform for model deployment
- One-click deployment of vLLM / SGLang inference engines
- Supports Auto Scaling, OpenAI-compatible APIs, and load balancing

The screenshot displays the GMI Cloud interface. On the left is a sidebar with navigation options: Inference, My Models, Deployments (highlighted), Fine-Tuning, and Storage. At the bottom of the sidebar are links for Reservation, Docs, and Company, along with a copyright notice: © Copyright GMI Cloud 2025. The main panel is titled 'Deployments' and features a search bar labeled 'Search by task name'. Below the search bar is a table with two columns: 'Name' and 'Status'. The table contains two entries: 'Llama-3.3-70B-Instruct vLLM-Public-20250220' and 'QwQ-32B-Preview vLLM-Public-20250214'. Each entry includes a subtext line indicating it was 'Created by Demo DDD' and the time since last update ('a month ago' and '8 months ago' respectively). Both entries have a blue 'Archived' status tag. At the bottom right of the main panel, there is a pagination control showing '< 1 2'.

Name	Status
Llama-3.3-70B-Instruct vLLM-Public-20250220 Created by Demo DDD Last updated a month ago	Archived
QwQ-32B-Preview vLLM-Public-20250214 Created by Demo DDD Last updated 8 months ago	Archived

Level 4:

Optimized Model Layer

- **Inference Engine** offers a multimodal inference platform supporting text, image, and video generation
- Integrates and optimizes open-source models such as Llama, DeepSeek, Mistral, Sora, and Flux
- Supports quantization, multi-node inference, and cross-region caching with TeaCache
- Provides ready-to-use APIs, enabling rapid integration into enterprise AI applications

Models Library

Inference is the real-time phase of AI that processes data fast to make decisions. It powers things like autonomous driving, voice assistants, and recommendation systems, helping you achieve smarter, more efficient operations!

Model Type

LLM Video Image

Category

text-to-text llm glm

deepseek qwen3 thinking

qwen llama3 llama4

Show 64 more

Reset filters

LLM



OpenAI: GPT OSS 20b

Featured text-to-...

Input \$0.04/M · Output \$0.15/M

Serverless Dedicated



ZAI: GLM-4.6

Featured llm ...

Input \$0.60/M · Output \$2.00/M

Serverless Dedicated



DeepSeek: DeepSeek-V3.2-Exp

Featured llm ...

Input \$0.27/M · Output \$0.41/M

Serverless Dedicated



DeepSeek: DeepSeek-V3.1-Terminus

Featured llm ...

Input \$0.27/M · Output \$1.00/M

Serverless Dedicated



Qwen: Qwen3 Next 80B A3B Thinking

Featured llm ...

Input \$0.15/M · Output \$1.50/M

Dedicated



Qwen: Qwen3 Next 80B A3B Instruct

Featured llm ...

Input \$0.15/M · Output \$1.50/M

Serverless Dedicated

Video



veo-3.1-fast-generate-preview

Featured video-ga... ..

\$0.15/S



veo-3.1-generate-preview

Featured video-ga... ..

\$0.40/S



sora-2-pro

Featured video-ga... ..

\$0.50/S







Who we work with on Inferencing

 mirelo.ai

 Higgsfield

UTPAI

 OpenRouter

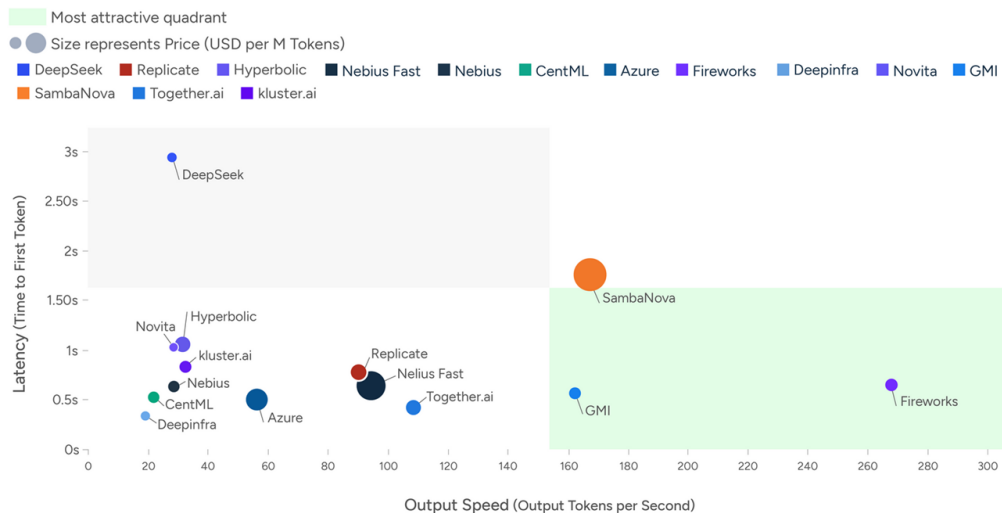
 FastVideo

 VLLM

 SGL

 LMCache

Industry-Leading Inference Performance



Latency vs. Output Speed:

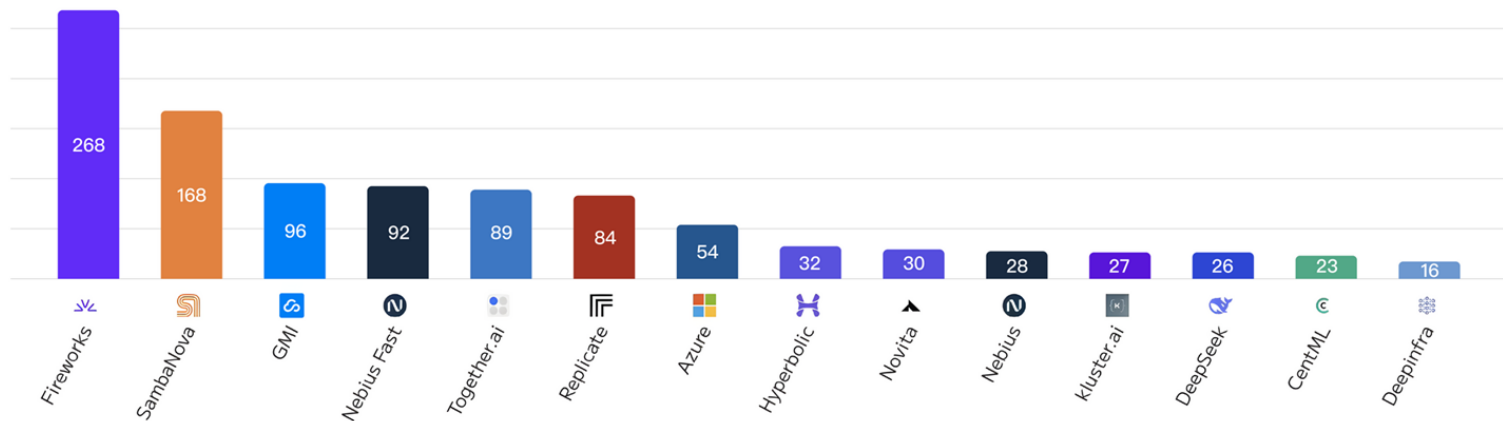
DeepSeek V3 0324 (Mar'25) Providers

Latency: Seconds to First Token Received; Output Speed: Output Tokens per Second; 1,000 Input Tokens

Industry-Leading Inference Performance

Output Speed

Output Tokens per Second, Higher is better



The Value of Efficient Inference (same hardware)

30x

Performance
boost

Due to bottlenecks in model
execution

71%

Cost Reduction

Possible with proper
optimization

11x

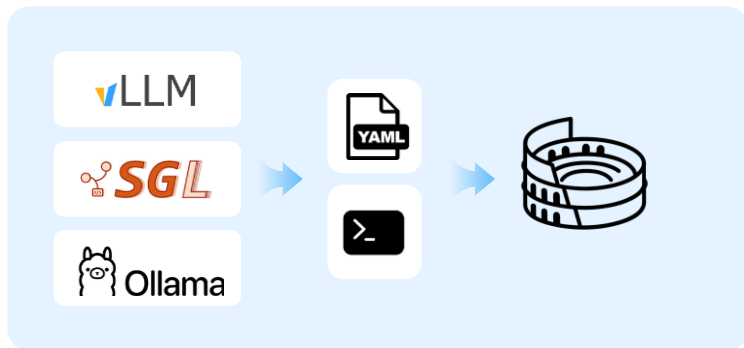
Inference speed

Due to challenges balancing
latency and throughput

Introducing GMI Inference Engine Benchmark tool

Open Source Community Release

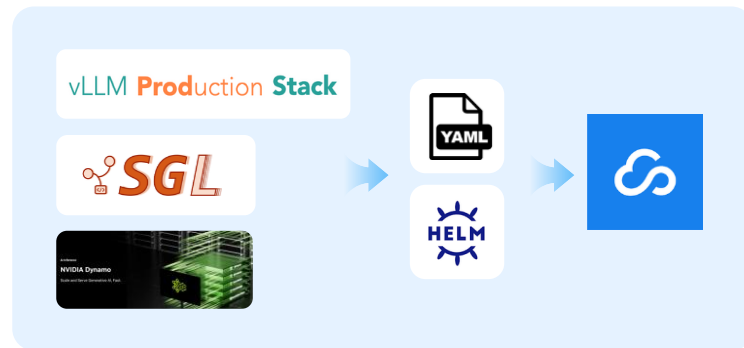
A framework for evaluating various large model inference engines (e.g., vLLM, SGLang, Ollama).



Deployment utilizing a single machine command line

GMI Cloud Edition

A service that facilitates benchmarking and assessment of diverse production-grade inference scenarios, including vLLM, SGLang, the vllm production stack, and Dynamo.



Online cluster deployment to guarantee optimal hardware and network performance

GMI Inference Engine Benchmark tool

Inference Engine Arena Local Leaderboard    Compare performance of different inference engines and models - v0.1.0

Model: GPU: Benchmark: Engine: 

☒ Output Throughput (tokens/s) ☐ TTFT (ms)

Model	Engine	GPU	Benchmark	Input Throughput (tokens/s)	Output Throughput (tokens/s)	Input \$/1M tokens	Output \$/1M tokens	TTFT
meta-llama/llama-3.3-70B-instruct	sglang	NVIDIA H100 80GB HBM3 (4x)	conversational_short	787.76	727.26	\$5.6419	\$6.1112	117.65
meta-llama/llama-3.3-70B-instruct	vllm	NVIDIA H100 80GB HBM3 (4x)	conversational_short	789.11	714.15	\$5.6322	\$6.2234	43.01
meta-llama/llama-3.3-70B-instruct	vllm	NVIDIA H100 80GB HBM3 (4x)	conversational_short	788.84	713.9	\$5.6342	\$6.2256	43.11

 Manage Columns

Initiate a dedicated inference endpoint utilizing artifact xyz...

Benchmark



Conduct benchmarks and incorporate new records.

Engage with the core of the GMI Cloud platform:

1. The inference configuration can be initiated or replicated with a single click via the GMI Cloud platform.
2. If data for the necessary configuration is not yet available, new benchmarks can be executed immediately via the GMI Cloud platform.



Thank you



Welcome to GMI Cloud

278 Castro St, Mountain View, CA 94041

gmicloud.ai