GMI

Optimizing the AI Stack for the Age of Inferencing

Maximizing Infrastructure Utility for Scalable AI

Speaker: Yujing Qian, VP of Engineering, GMI Cloud

World Summit AI, April 16, 2025

What to Expect:
Inference Optimization Key Strategies

Hardware Acceleration

P/D disaggregation

Performant inference engine

Distributed KV Cache

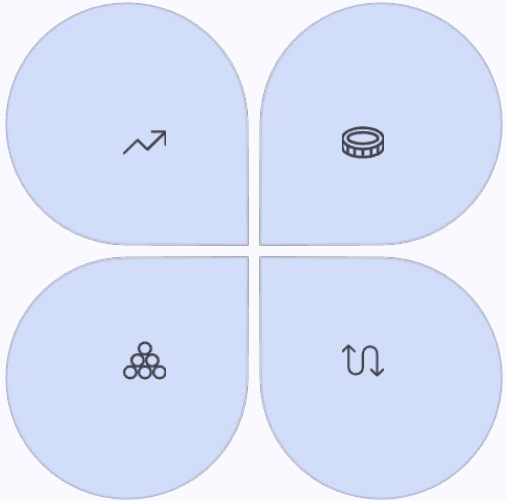
Why Inference Optimization Matters Today

Business Value

Inference drives real-world applications

Cost Control

Optimization essential for maximizing ROI



Efficiency

Balancing performance with resource usage

Adaptation

Ecosystems must evolve with technology

The Cost of Inefficient Inference

30 x

Performance Gap

Due to bottlenecks in model execution

71%

Cost Reduction

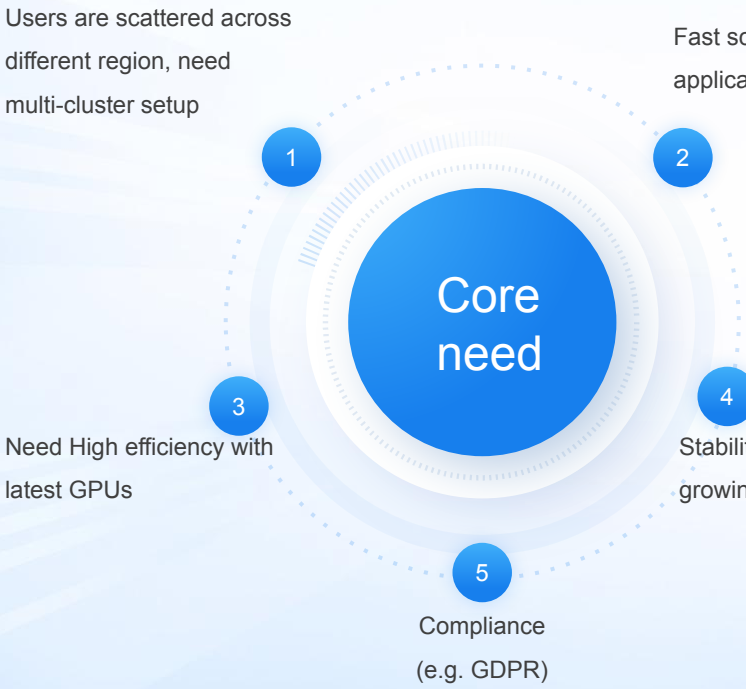
Possible with proper optimization

13 x

Latency Impact

Due to challenges balancing latency and throughput

Key of AI inference: efficiency, scalability and stability



- Global cluster coverage
- Fast & Flexible auto-scaling
- Efficient inference cluster
- High reliability and security

Scalable Inference System Best Practices

1

Performant inference engine

Minimum core component, squeeze every bit of performance out of your hardware

2

Workload Optimization

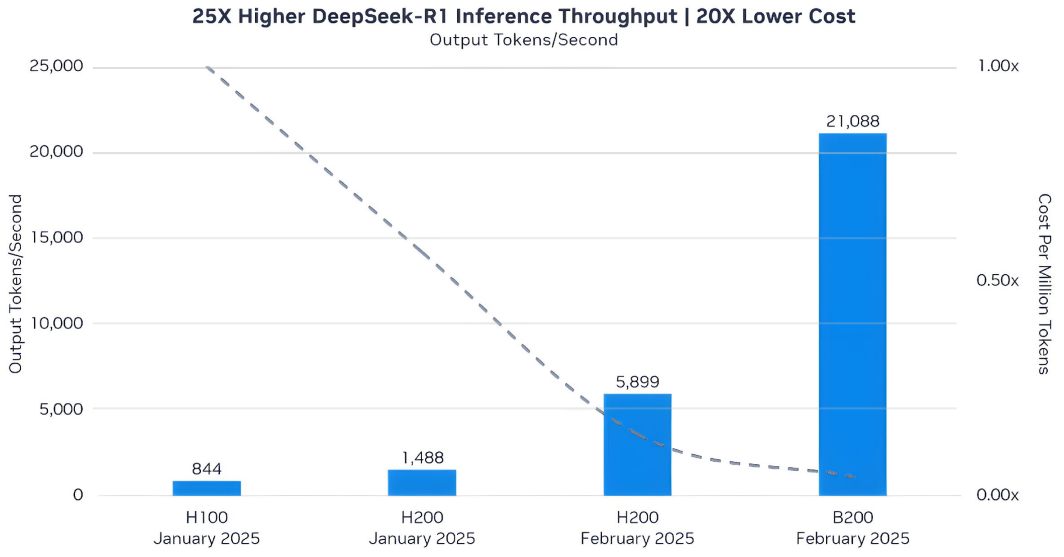
Intelligent distribution across hardware types and workload types

3

Global Autoscaling

Scale out our workload to hybrid cloud and multi-cluster

Always on the lookout for Best hardware



*data provided by NVIDIA

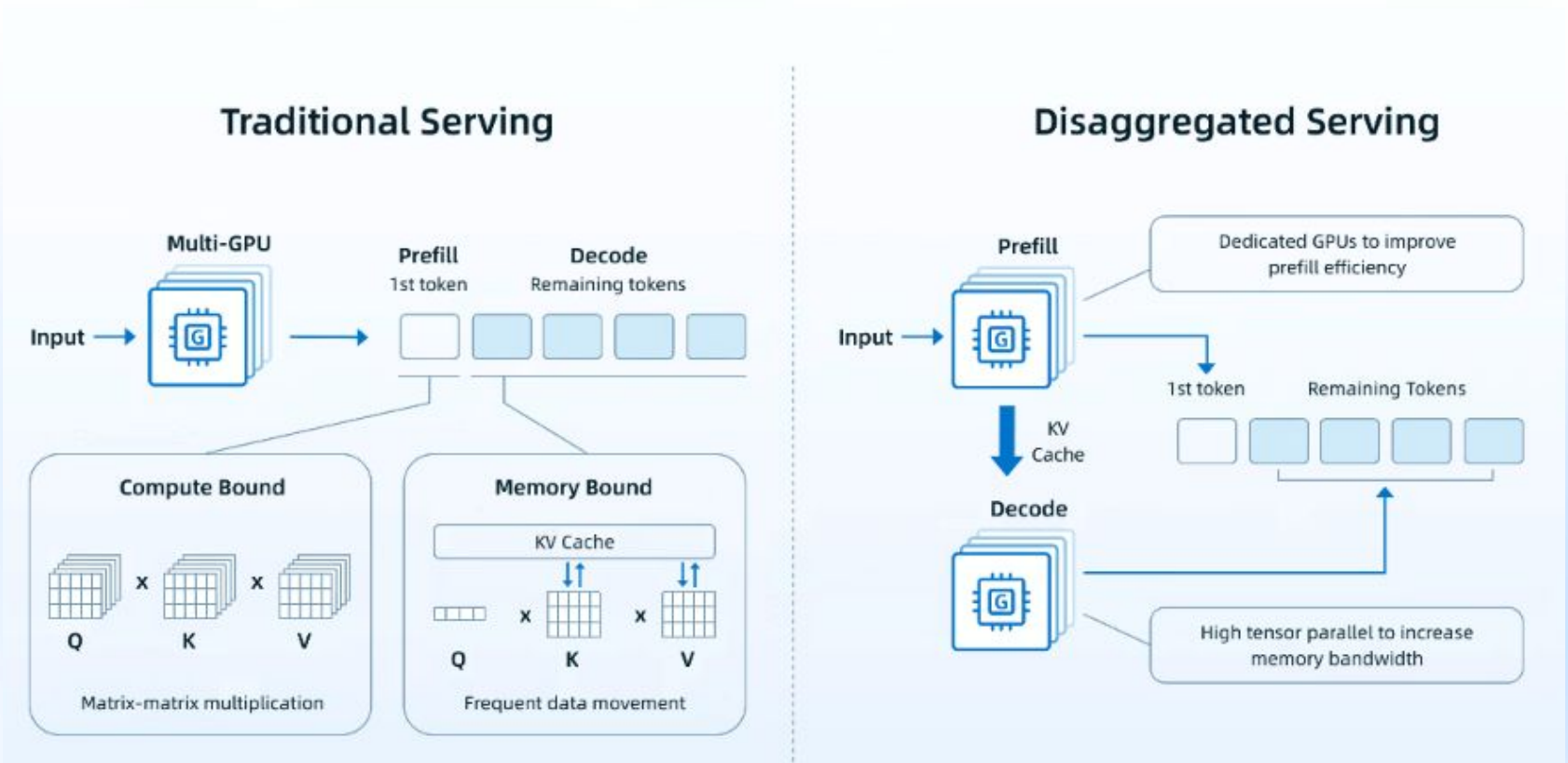
DeepSeek-FP4

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

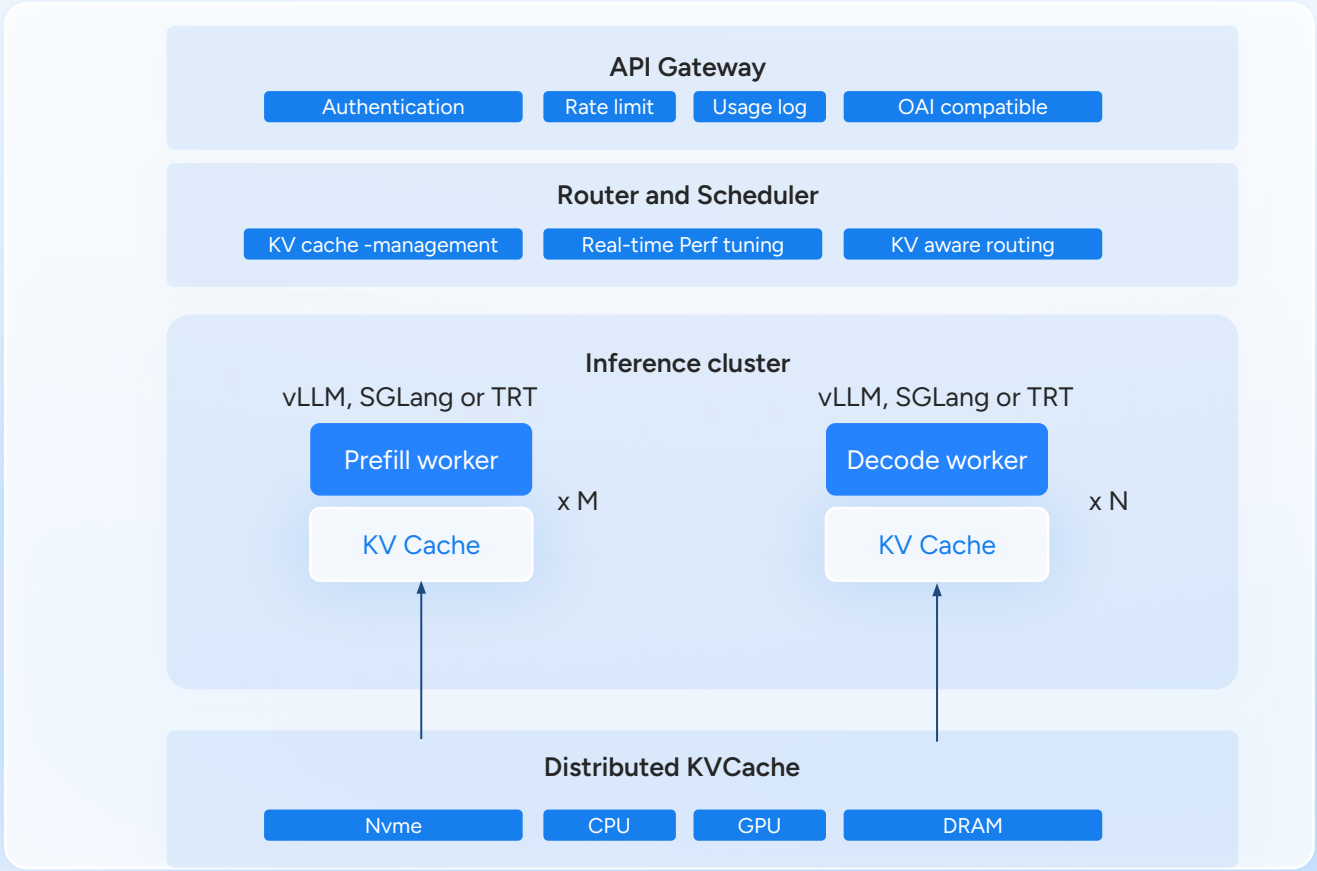
B 200 is more than 25x performant than H200. Latest hardware hosts latest model better

Hardware upgrades is one the easiest way to improve performance

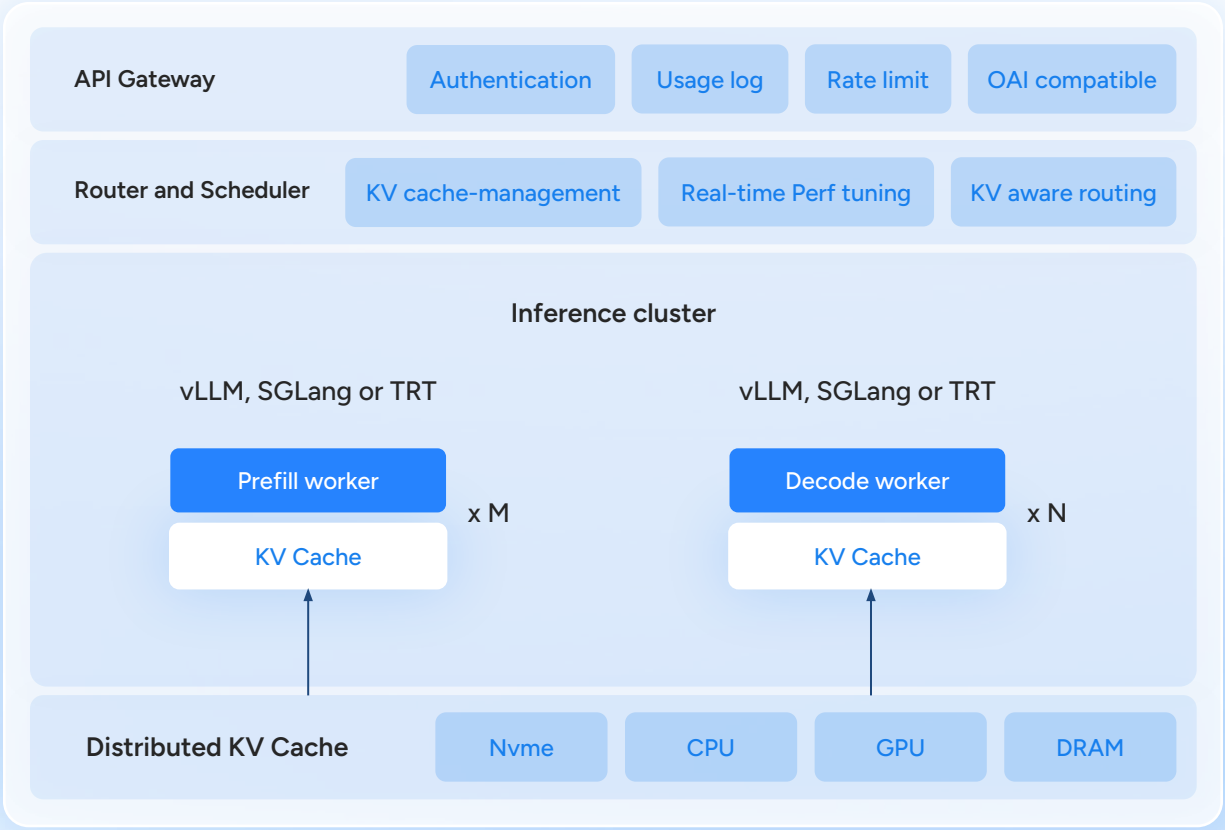
Prefill/Decode Disaggregation



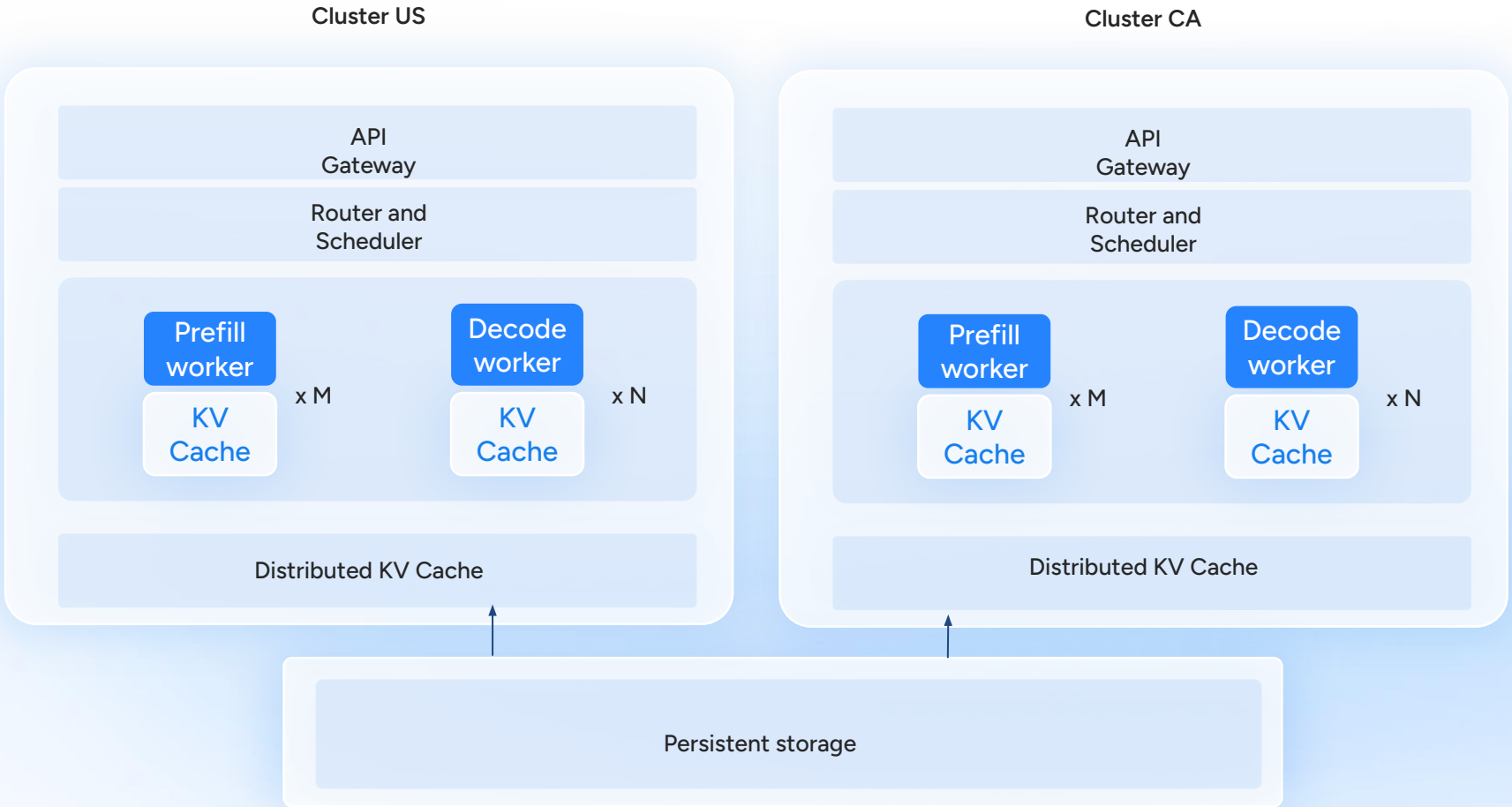
Inference cluster with Prefill/Decode Disaggregation



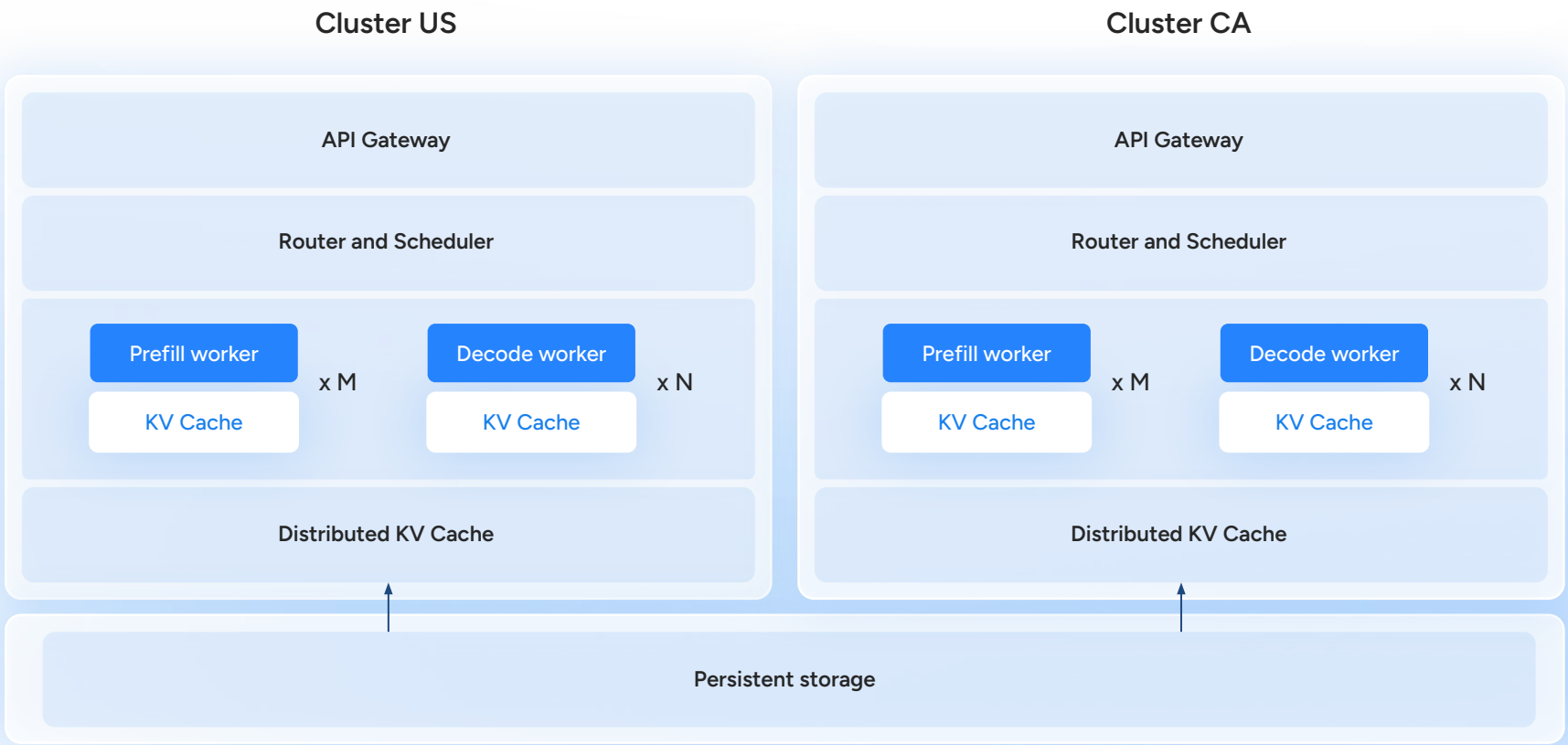
Inference cluster with Prefill/Decode Disaggregation



Inference engine with Distributed KV Cache + Persistent storage



Inference engine with Distributed KV Cache + Persistent storage



Boosting Cluster utilization

High speed autoscaling

Improve model loading time using techniques like RDMA

Fast LoRA switches

Use multi-LoRA to reuse base models

Offline queue

Queue up offline jobs and run them when cluster is idle

Average unattended private cluster

10+ min

Scaling time

Due to slow model loading and default scaling metric

40%

GPU utilization

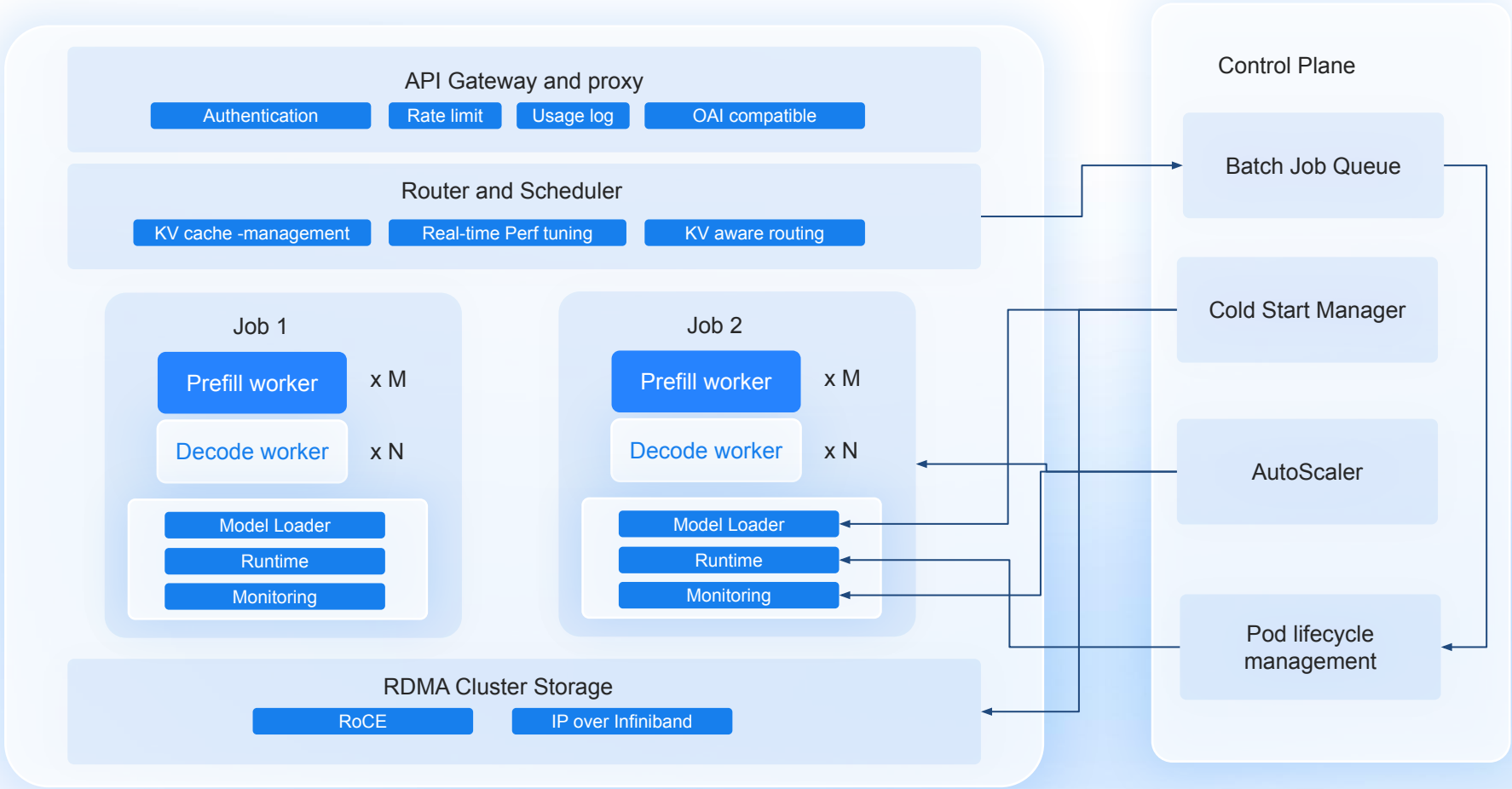
GPU bubbles everywhere

<5

Workload supported

Due to challenges managing multiple jobs at the same time

Cluster level Autoscaling



Cluster level Autoscaling

Global auto scaling

Traffic aware
Dynamic Autoscaling

Fast model spin up
with cluster preheat

Deploy same model in both Canada and US with one workflow.

Manual multi-cluster setup

days

Spin up time

Cluster, network, environment, file transfer… you name it

>5

Manual steps

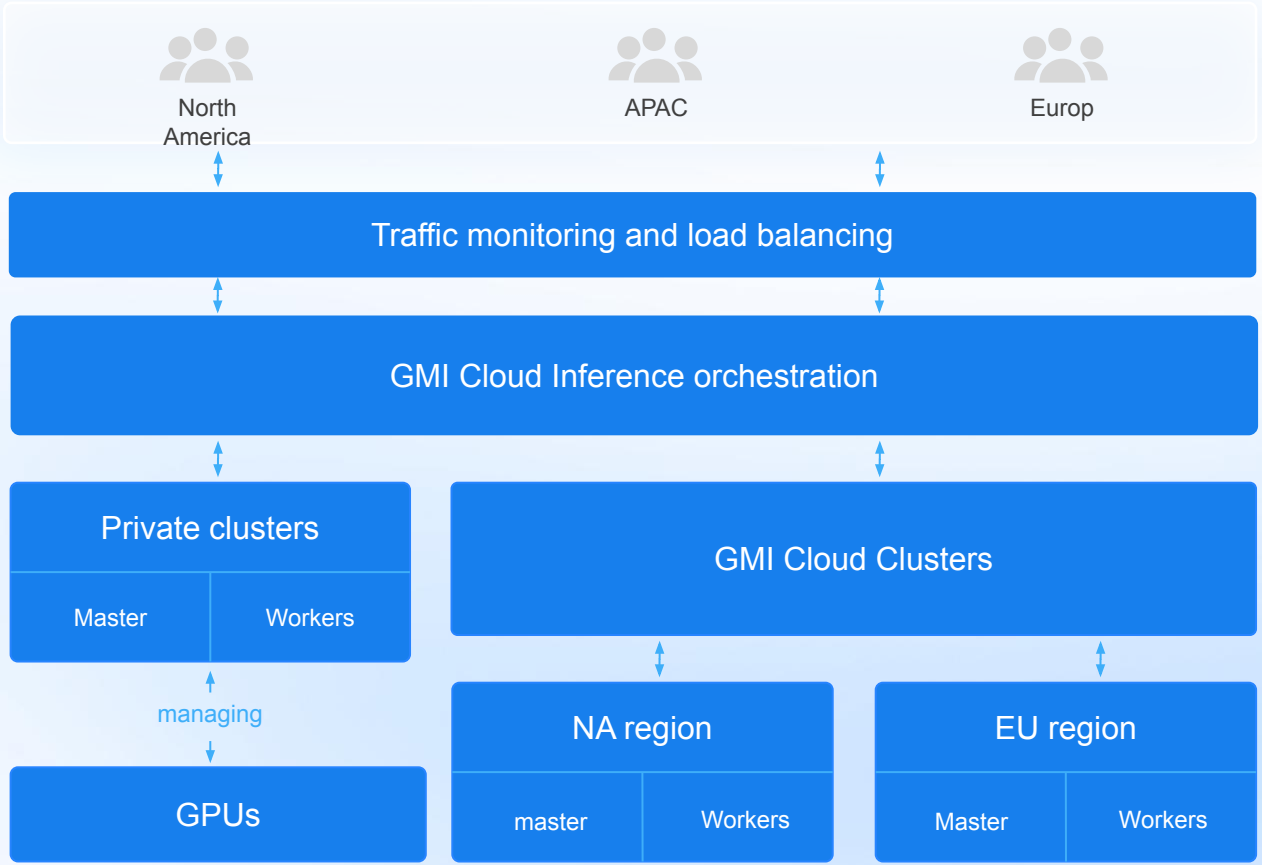
Just to get resource provisioned

no

Support functionality

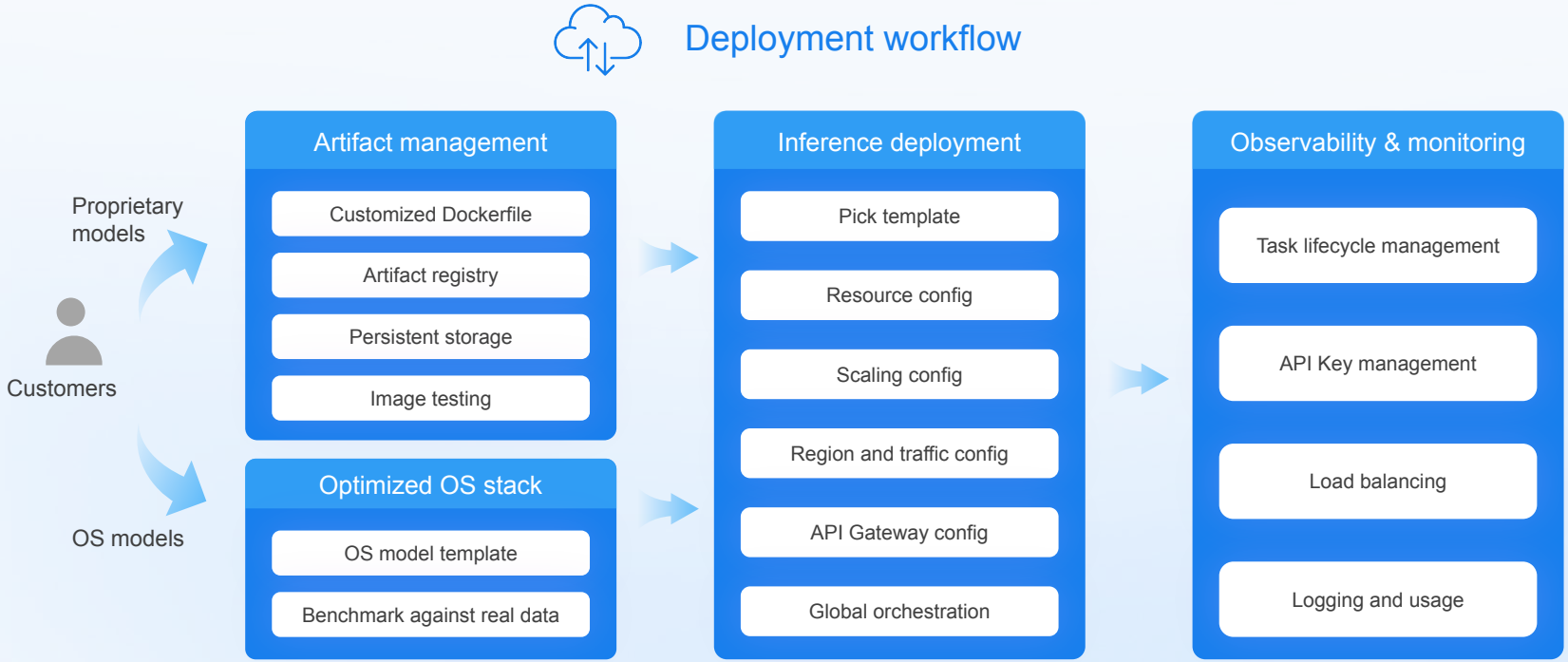
E.g. monitoring, endpoint control, etc

Inference Scaling Stack: Global Autoscaling



- Smart traffic routing: user location, traffic, availability
- Cross-cluster autoscaling based on user provided metrics
- Support hybrid-scaling and cluster management.

One-click Inference Engine Deployment



Global End to End monitoring

GMI Inference monitoring

Prometheus → Grafana

Hardware: DCGM → Metrics → http server

Service: Server status, Workload

K8s Cluster

K8s control plane

K8s control plane

K8s control plane

K8s Worker

GPU GPU GPU

GPU

Memory

Disk

K8s Worker

GPU GPU GPU

GPU

Memory

Disk

Storage

K8s Cluster

K8s control plane

K8s control plane

K8s Worker

K8s Worker

K8s Cluster

K8s control plane

K8s control plane

K8s Worker

K8s Worker

Key Takeaways Summary

1

Hardware Acceleration

Scale efficiently with dynamic resource management

2

Performant inference engine

Serve tokens with highest cost-efficiency

3

Workload Optimization

Balance online and offline jobs efficiently

4

Global Auto Scaling

Deploy and scale across multiple clusters and hybrid clouds

GMI

Get Started with Real-World AI Stacks

Try GMI Cloud Inference Engine
Get $100 of credits with code: **INFERENCE**

Contact us at **sales@gmicloud.ai**