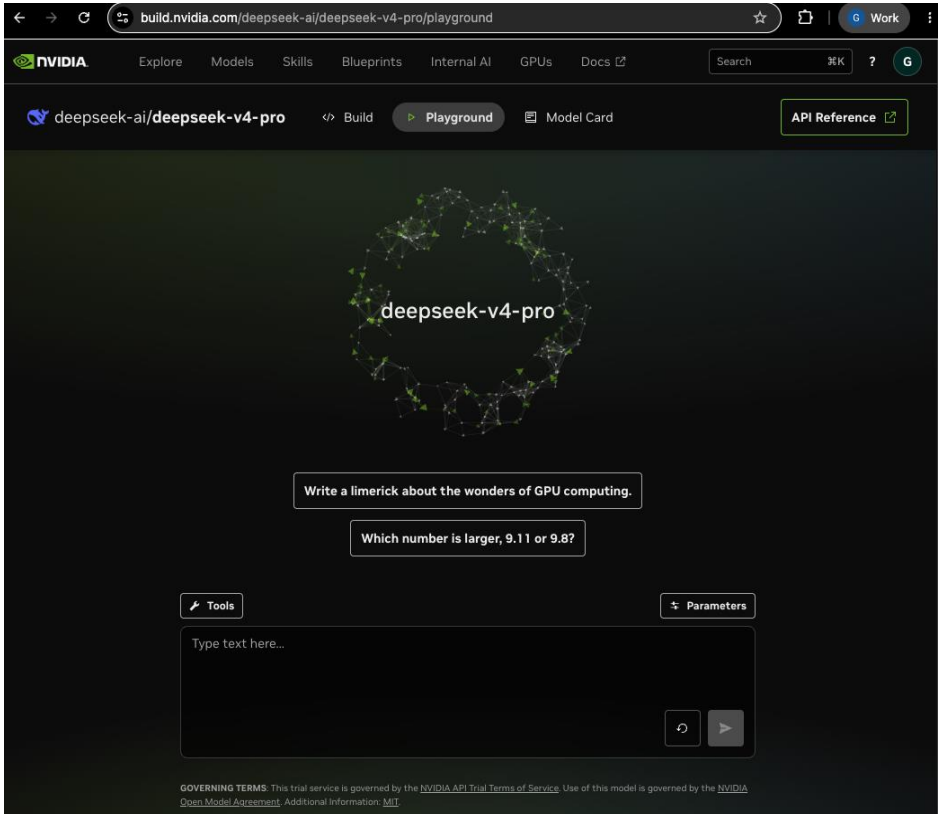
Inference in Progress... Please Monitor Responsibly

Observability Learnings for AI Inference Platforms at Scale

Gaurav Sharma
Nvidia

AI Inference Request Lifecyle



User Prompt

↓

Inference Platform (Control Plane)

↓

Model Endpoint (Data Plane)

↓

Model Inference

↓

Response Stream

↓

User

AI Inference Orchestration Challenges

Challenges	Orchestration Requirements
Multiple Consumers	Global Endpoint + Auth
Multiple Models	Model Abstraction
Multiple GPU Types	GPU Aware Placement
Multiple Clusters	Multi-cluster Deployment
Multiple Traffic Patterns	Routing + Autoscaling

AI Inference Platform

Population-Scale Inference

"... powers more than 500 government websites and processes over 15 million inferences daily, surpassing 6 billion total inferences leveraging NVIDIA's NVCF, all while maintaining strict sub-second response times..."

PIB, MeitY, Mar 2026

Control Plane

- NVCF APIs (Global Endpoint + Auth)
- Functions Definitions (Model Abstraction)
- Function Deployment (GPU Aware Placement)
- Secrets Management
- Rate Limiting
- Autoscaling

Invocation Plane

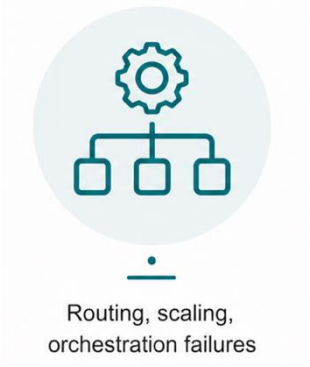
- Routing inference requests to function workloads

Compute Plane

- NVCF Cluster Agent (GPU Cluster Discovery for Platform)
- Worker Orchestration on GPU Clusters to execute Inference

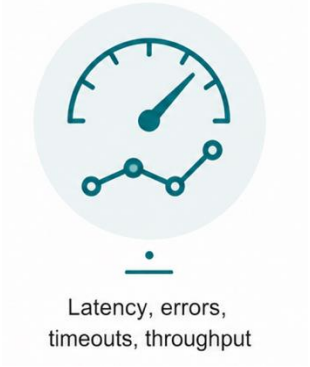
Inference Reliability Challenges

Control Plane



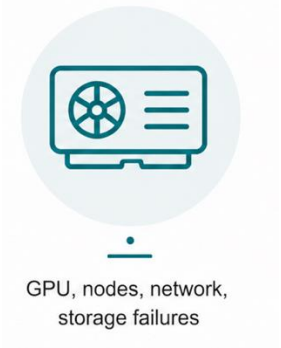
Routing, scaling, orchestration failures

Invocation Plane



Latency, errors, timeouts, throughput

Compute Plane



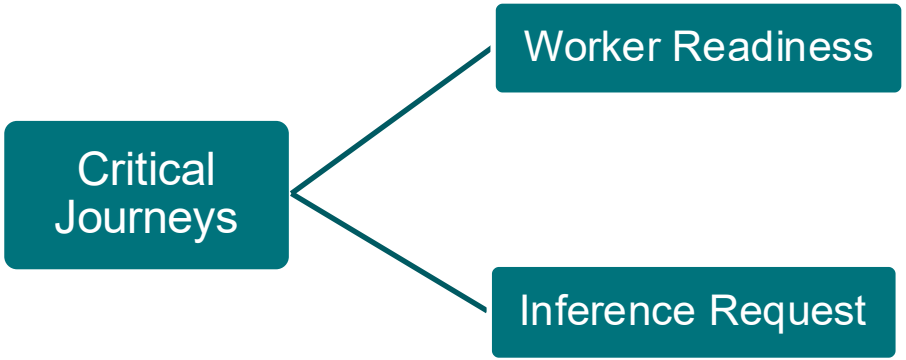
GPU, nodes, network, storage failures

Monitoring Inference

Alert via SLO

Correlate via Drill Downs

Inference Request Journeys



Metrics Instrumentation

Define SLIs

Set Targets & Budgets

Alert on Burn Rate

Worker Readiness

Kubernetes	GPU Software Stack	GPU Hardware	Custom Monitoring
Node Ready	GPU Operator	GPU Availability	Compute Plane Agent Checks
Pod Health	Driver	XID Errors	Model Downloads Checks
Resource Pressure	Container Toolkit	ECC Errors	Custom Network Monitoring
Kubelet	Device Plugin	Temperature	GPU Provisioning Workflows

Healthy GPU Node

Ready Pods

- GPU Operator components
- NVIDIA driver ready
- Device plugin ready

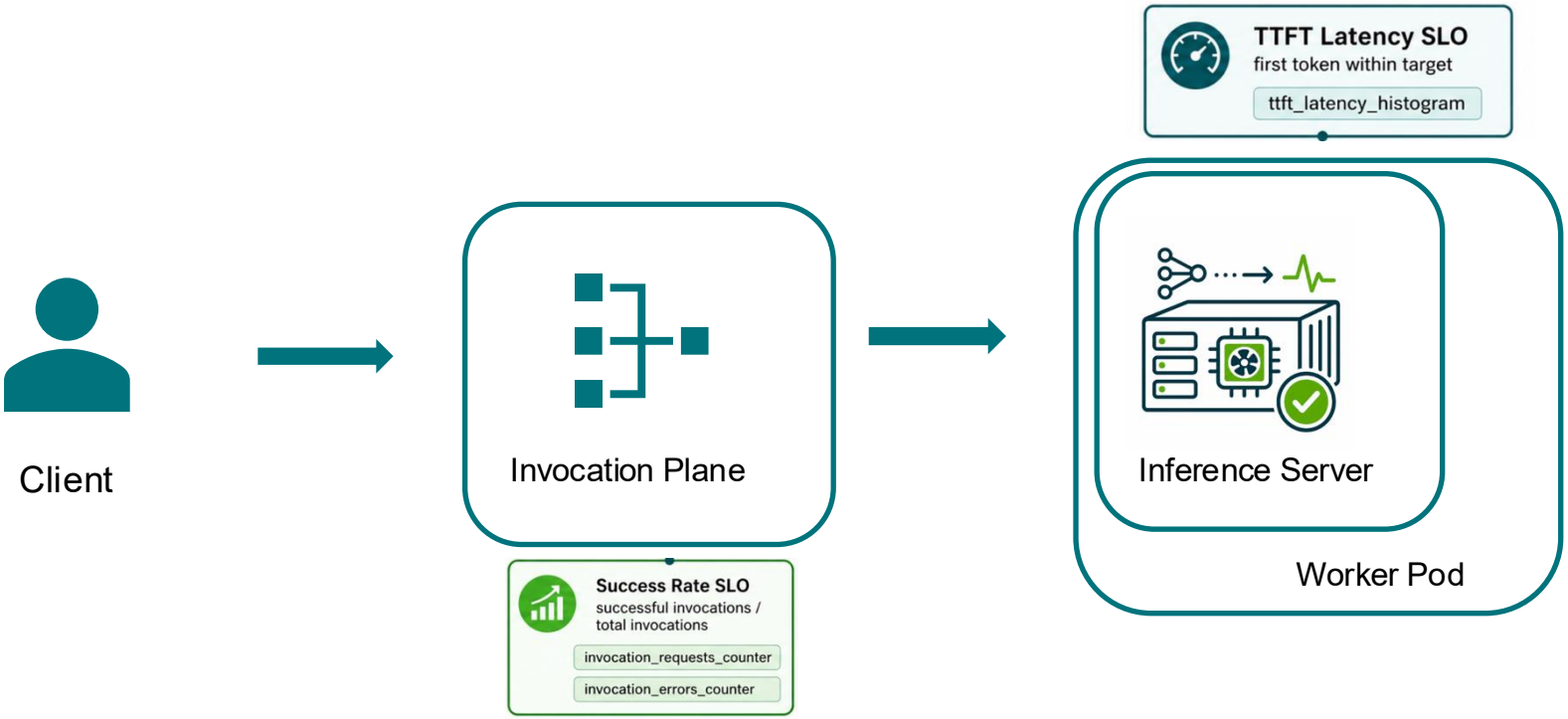
Clean Signals

- DCGM metrics
- XID / ECC Errors
- NVLink / thermal / power
- Resource pressure normal

Validation Checks

- Worker Pods Stable
- Network probes passing
- Test workload succeeds
- Inference Server Logs

Inference Request Health



NVIDIA OSS

NVIDIA Cloud Functions

GPU Operator

DCGM-Exporter

Dynamo

NVSentinel

KAI Scheduler

Q&A


KubeCon
CloudNativeCon
India 2026
Thank You