Solution Brief

virtana

# Virtana AI Factory Observability

## Complete Observability for AI Infrastructure—from GPU to Model Outcomes

As enterprises scale their AI initiatives, the infrastructure complexity behind these workloads grows exponentially. Traditional monitoring tools fail to capture the nuanced interdependencies across GPU clusters, distributed training jobs, storage, and network systems. This leads to blind spots, inefficiencies, and failures that derail AI performance and inflate costs.

Virtana AI Factory Observability (AIFO) delivers full-stack observability tailored to AI data centers. From real-time GPU metrics to infrastructure-aware model performance diagnostics, AIFO empowers IT and platform teams to optimize every layer of the AI stack—across cloud, on-premises, and hybrid environments.

## Key Capabilities

### GPU Cluster Monitoring

Track utilization, thermal conditions, memory health, and error rates across multi-vendor GPU environments.

- Reduce idle GPU time and increase training efficiency
- Identify throttling and thermal issues that hinder performance
- Correlate GPU activity to specific models or jobs

### Distributed AI Workload Observability

Visualize model behavior and infrastructure dependencies across multi-node jobs.

- Detect bottlenecks in training and inference workflows
- Identify stragglers, failed ports, or resource contention
- Analyze traces from pod to GPU to storage in real time

### Infrastructure-to-Model Correlation

Bridge the gap between physical resources and logical workloads.

- See where models are running—and why they're slowing down
- Trace application latency back to GPU contention or network congestion
- Map AI workloads to compute, storage, and network dependencies

### Storage & Network Telemetry

Monitor IOPS, throughput, latency, and bandwidth in context with AI pipeline health.

- Uncover silent storage failures or network misconfigurations
- Ensure high-throughput performance for large-scale data ingest
- Reduce training cancellations due to hidden infrastructure faults

### AI-Specific Root Cause Insights

Leverage AI/ML-powered diagnostics to pinpoint infrastructure causes of AI degradation.

- Surface correlated metrics tied to training or inference failures
- Diagnose port resets, faulty GPU topology settings, or inefficient GPU configurations
- Improve RCA speed and accuracy with contextual intelligence

## Capacity & Cost Optimization for AI

Make smarter decisions about scaling and spending.

- Reduce GPU underutilization by 40% within 6 months
- Right-size infrastructure across clouds and data centers
- Track kilowatt-hours for energy efficiency and sustainability reporting

# Powered by the Virtana Platform

AIFO is part of the broader Virtana Platform, integrating with modules like Global View, Container Observability, Infrastructure Observability, and Cost & Capacity Management.

## Built-in Integrations

- OpenTelemetry support for traces, logs, and metrics
- Seamless connection with Kubernetes environments
- Compatibility with ServiceNow, Slack, email, and other ITSM tools

## Flexible Deployment

- SaaS on AWS
- On-premises via Kubernetes

# Results that Matter

## 40% Reduction in Idle GPU Time

Global FSI Customer

## 60% Decrease in MTTR

AI-specific root cause insights speed up recovery. Healthcare Provider

## 15% Lower Power Usage

Energy-aware optimization of GPU workload placement. AI Lab – USA

# Why Virtana

- End-to-End AI Visibility: From model to hardware in a single view
- Predictive Analytics: Act before latency, cost, or power issues derail your pipeline
- Hybrid Coverage: Support AI workloads across public cloud, private cloud, and on-prem
- Security & Control: Designed for regulated industries and mission-critical environments

# Ready to Optimize Your AI Infrastructure?

With Virtana AIFO, you gain clarity, control, and confidence across every layer of your AI factory.
Request a Demo and see how Virtana accelerates AI outcomes—without wasting time or budget.

info@virtana.com | virtana.com | +1-408-579-4000