MantisGrid AI

SOLUTION BRIEF

Autonomous Reliability for AI Infrastructure

The Missing Layer in Modern AI Infrastructure

mantisgrid.ai | sales@mantisgrid.ai

1 THE MISSING LAYER IN MODERN AI INFRASTRUCTURE

AI infrastructure has evolved rapidly—from cloud to GPUs to large-scale distributed training and inference. While provisioning is well supported by cloud and neo-cloud providers, Day 1 and Day 2 operations have become significantly more complex.

KEY INSIGHT

Reliability is hard at any scale. At scale, it breaks.

A recent Meta study highlights the impact of failures on AI workloads:

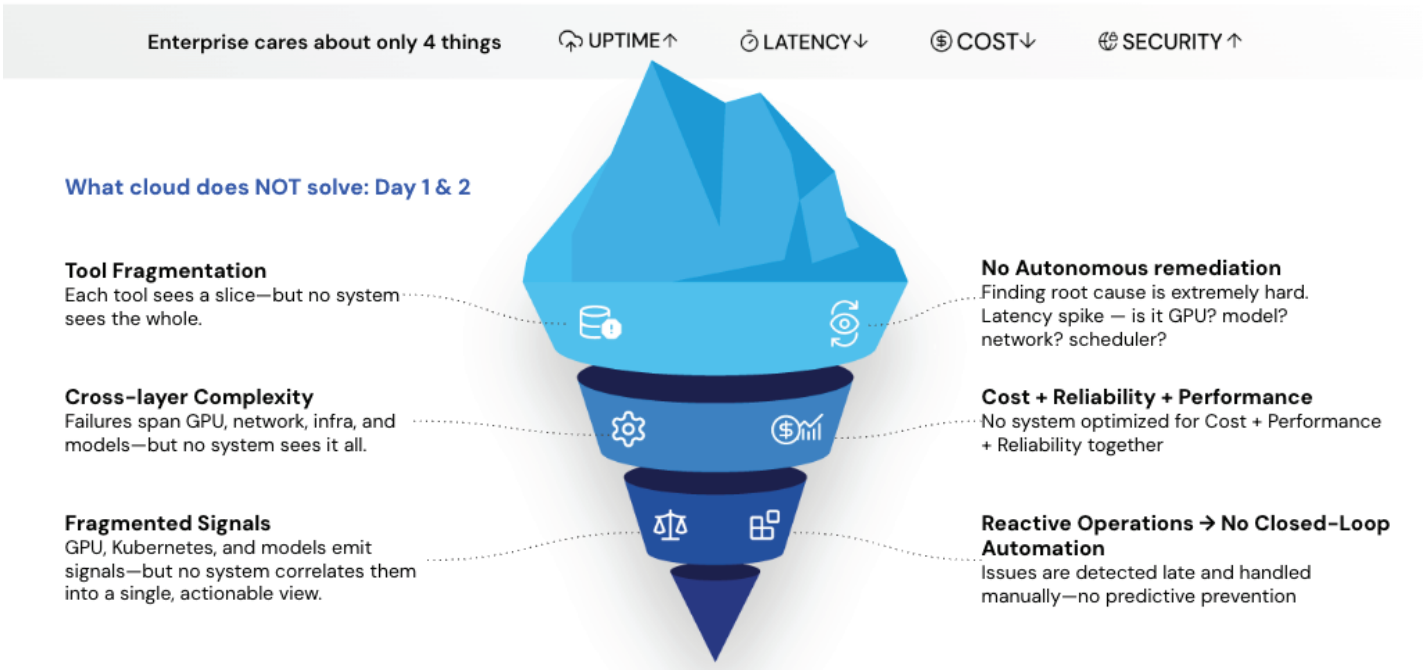
"The most commonly attributed failures are due to the backend network, the filesystem, and GPUs."

47.7 days	7.9 hrs	~19%	Every few hrs
Mean time to failure at 8 GPUs	Mean time to failure at 1,024 GPUs	Total GPU runtime impacted by hardware failures	Failure rate at 1,000+ GPUs

Failures rarely originate from a single layer. A performance issue may stem from GPU contention, orchestration behavior, model inefficiencies, or network instability. These issues are deeply interconnected, yet most tools observe them independently.

Large technology companies have built internal reliability platforms to manage this complexity. Most enterprises, however, lack the resources and expertise to replicate such systems.

Enterprises today operate across multi-cloud, edge, and on-prem environments, running highly dynamic AI workloads. However, the operational layer remains fragmented and reactive.



At the same time, organizations are measured on a small set of critical business outcomes—uptime, latency, cost, and security—yet lack a system to consistently enforce them.

Cloud providers are optimized for infrastructure provisioning, not for enforcing enterprise-wide operational outcomes across environments. They do not provide native mechanisms to enforce cost boundaries across environments or deliver unified reliability across multi-cloud deployments.

> **KEY INSIGHT**
>
> *What's missing is a system that continuously ensures business outcomes across the entire AI stack.*

2 THE MANTISGRID AI APPROACH

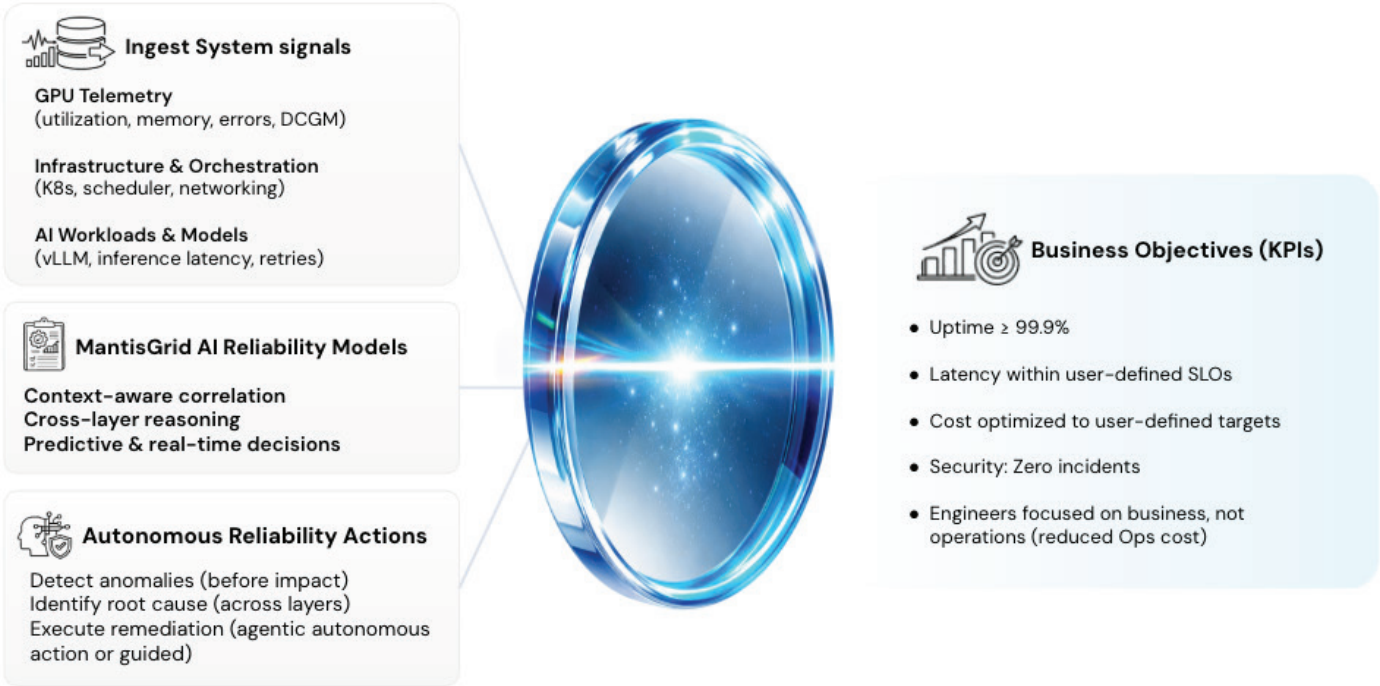
This is the gap MantisGrid Cloud is designed to solve.

MantisGrid Cloud is our SaaS platform that autonomously manages infrastructure, AI training and inference across any cloud to meet your business intent—uptime, performance, and cost. It senses signals and executes closed-loop control across GPUs, infrastructure, and workloads—with governed human approval.

Unlike traditional tools that optimize individual metrics, MantisGrid Cloud optimizes business outcomes.

Enterprises define intent in terms of availability, performance, cost efficiency, and security. MantisGrid Cloud continuously evaluates system behavior against these objectives and ensures alignment in real time.

By combining system-wide context with intelligent decisioning, MantisGrid Cloud reduces last-minute firefighting by proactively detecting and resolving issues—cutting operational churn and autonomously addressing, or alerting teams to, what matters most to keep systems running smoothly.



At scale, GPU and AI job failures are continuous—making reliability a system-level problem, not a tooling problem. MantisGrid Cloud solves this with a closed-loop, intelligent control system.

3 WHAT SETS MANTISGRID CLOUD APART

Unlike observability tools or automation scripts, MantisGrid Cloud acts as a real-time decision system for AI infrastructure.

AI-Native Reliability

Understands the unique failure patterns of training and inference workloads with purpose-built reliability models, going beyond traditional monitoring.

Autonomous Operations

Continuously senses, reasons and acts -- closing the loop from detection to remediation to enforcement, with optional human governance.

Cross-Layer Intelligence

Correlates signals across GPUs, infrastructure, and AI workloads to identify true root causes and optimize system-wide behavior—not isolated components.

Outcome-Driven Control

Aligns infrastructure to business intent—ensuring uptime, latency, cost, and security goals are continuously met without manual intervention.

4 BUSINESS IMPACT & DEPLOYMENT

MantisGrid Cloud enables organizations to reduce incidents, improve GPU utilization, and optimize infrastructure costs while minimizing manual operations. Delivered as a SaaS control layer, it integrates seamlessly with existing multi-cloud, hybrid, and edge environments without requiring infrastructure changes. As AI systems scale beyond human manageability, MantisGrid Cloud provides a self-managing, adaptive approach that continuously aligns infrastructure with business outcomes—bringing intelligence, autonomy, and predictable reliability to AI operations.

Contact for Demo

mantisgrid.ai | sales@mantisgrid.ai