cloud computing circle

The Future of Kubernetes

AI-Driven Management

Santanu Kumar Das
Kunal Das

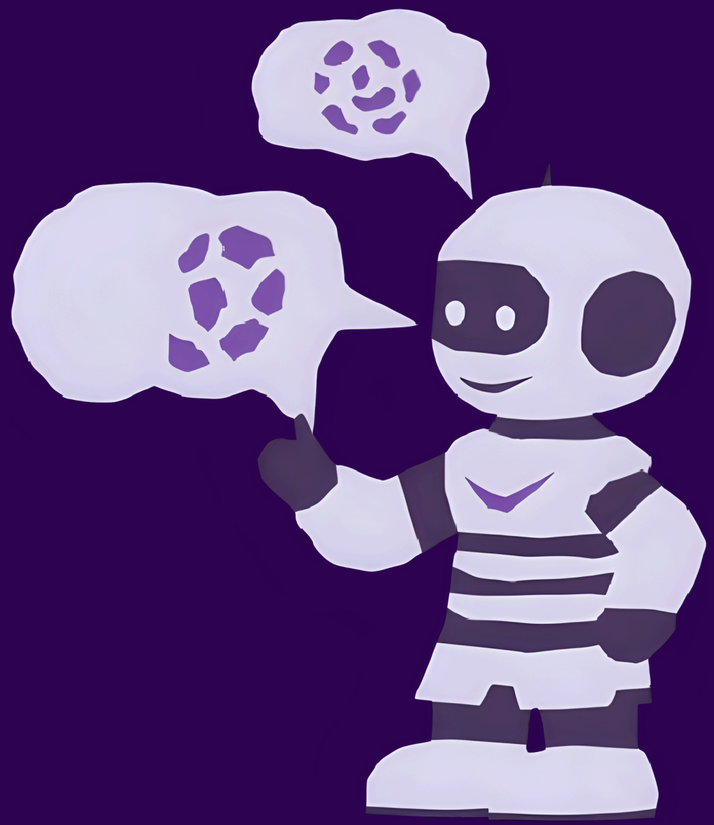
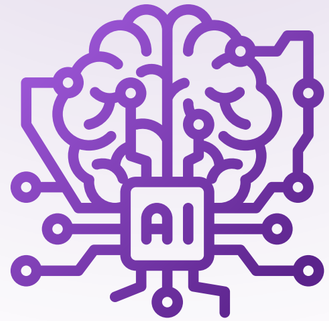


Table of Contents



About Us

Kunal Das, 👋

Developer Advocate APAC, CASTAI

Organizer of CNCF Kolkata, Cloud Computing Circle.
7x Azure,1x Hashicorp Certified

Santanu Kumar Das, 🙌

Sr DevOps Engineer, Optym

Azure & Hashicorp certified

Cloud, Compute & AI

Today, AI is at the core of how we manage, enhance, and get the most out of cloud and compute infrastructure for running critical workloads

Containerized workloads with Docker

01 Faster Deployment and Scaling

Applications and their dependencies are packaged into lightweight containers, enabling rapid deployment and seamless scaling across environments.

02 Improved Resource Efficiency

Unlike VMs, Docker containers share the host OS kernel and require less overhead, allowing you to run more workloads using fewer resources and reducing infrastructure costs.

03 Consistent Environments

Guarantees consistent environments from development through production, eliminating the “it works on my machine” problem commonly faced with VM-based deployments.

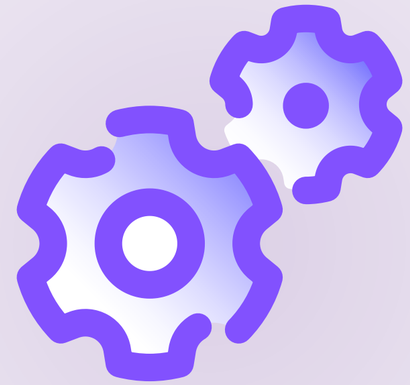
04 Simplified Updates and Rollbacks

Makes it easy to update or roll back workloads by redeploying container images, streamlining application maintenance and minimizing downtime compared to traditional VMs.

enter Kubernetes

- Automates container deployment, scaling, and management across the cluster.
- Delivers self-healing, rolling updates, and built-in service discovery.
- Manages complex, production-grade systems beyond Docker Compose’s limits.
- Offers a rich ecosystem of tools for monitoring, security, and integrations.

Traditional way of working with Kubernetes

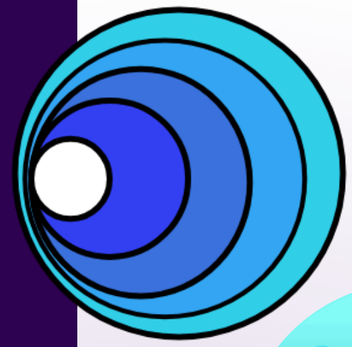


Deployment: Engineers use YAML files with kubectl or helm to deploy, update, and scale applications.

Monitor/Troubleshoot: Teams check logs, events and resource metrics using built-in tools and dashboards.

Resource Scalling: Application and node scaling are managed based on current demand and policies

Incident Response: Alerts guide teams to investigate and fix issues with standard procedures.



working with k8s in the era of AI

01

workloads & troubleshooting

K8sGPT uses AI to quickly diagnose, explain, and resolve Kubernetes workload issues, making operations smoother and more efficient

02

security posture

Falco actively monitors cluster activity, leveraging rules engine to detect threats and provide real-time alerts for enhanced security

03

scalling automation

PredictKube intelligently predicts demand and automates scaling decisions, optimizing resource usage while maintaining application reliability

capabilities

K8sGPT

What
AI-powered tool for Kubernetes troubleshooting

How
Scans cluster resources and explains issues in plain language using AI; run as a CLI or in-cluster service

Why
Saves time and reduces complexity by quickly pinpointing and describing Kubernetes problems

Falco

What
Open-source runtime security tool for Kubernetes

How
Uses rules engine to monitor system calls in real-time and detect anomalies or threats; deployed as a DaemonSet

Why
Enhances security posture by leveraging advanced rules for faster, smarter detection of suspicious activity

PredictKube

What
AI-driven resource prediction and scaling solution

How
Analyzes usage trends and predicts resource needs; integrates with Kubernetes autoscaler such as KEDA

Why
Ensures optimal scaling, prevents resource waste, and maintains application performance automatically

PROBLEMS

The Enterprise Scale

INDIVIDUAL DEVELOPER	ENTERPRISE CHALLENGE
• 1–3 clusters	500+ clusters across regions
• 10-50 nodes	50,000+ nodes (multi-cloud)
• 100s of pods	1M+ pods with complex dependencies
• Manual oversight	24/7 autonomous operations required
• Single team decisions	100+ teams, governance required

The Coordination Nightmare

01

Tool Conflicts

Multiple tools multiple decisions

02

No Global View

Each tool optimizes locally, creating global inefficiencies

03

Manual Orchestration

Humans become bottlenecks at enterprise scale

04

Governance Gaps

No unified policy enforcement across all clusters

Security at Light Speed

2025 Threat Landscape

CVE-2025-23266: *NVIDIA container escape*

Supply Chain Attacks: *188% increase in malicious packages targeting K8s*

Average Breach Cost: *$4.88M with 287 days to identify and contain*

Security at Light Speed

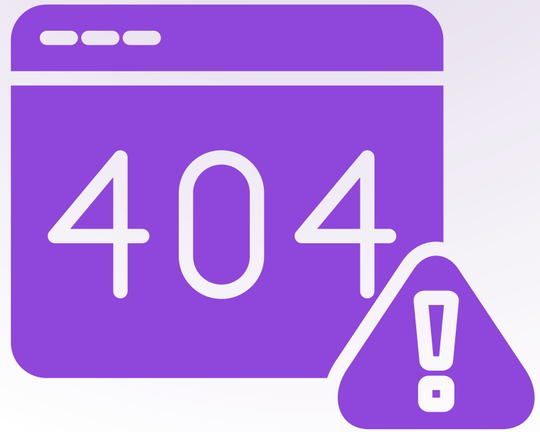
Why Individual Tools Fail at Scale?

Delayed Correlation: Falco detects threat in Cluster A, but doesn't alert Cluster B

Human Bottleneck: Security team gets 47 alerts/hour, misses critical ones

Incomplete Context: Each tool sees 20% of the attack surface

No Automated Response: Detection without coordinated remediation = useless



- Security Agent
- Open Source
- eBPF Monitoring

https://github.com/castai/kvisor

Orchestrated Intelligence Platform

From Tool Collection to Unified AI Brain.

- RESOURCE MANAGEMENT
- SECURITY
- COST OPTIMIZATION

Resource Utilization

CPU and memory utilization

The average CPU utilization across clusters remained low at 10% (-23% YoY), while average memory utilization was marginally better at 23% (+15% YoY), indicating no significant year-over-year improvement in resource efficiency across cloud platforms compared to our previous report from 2024.

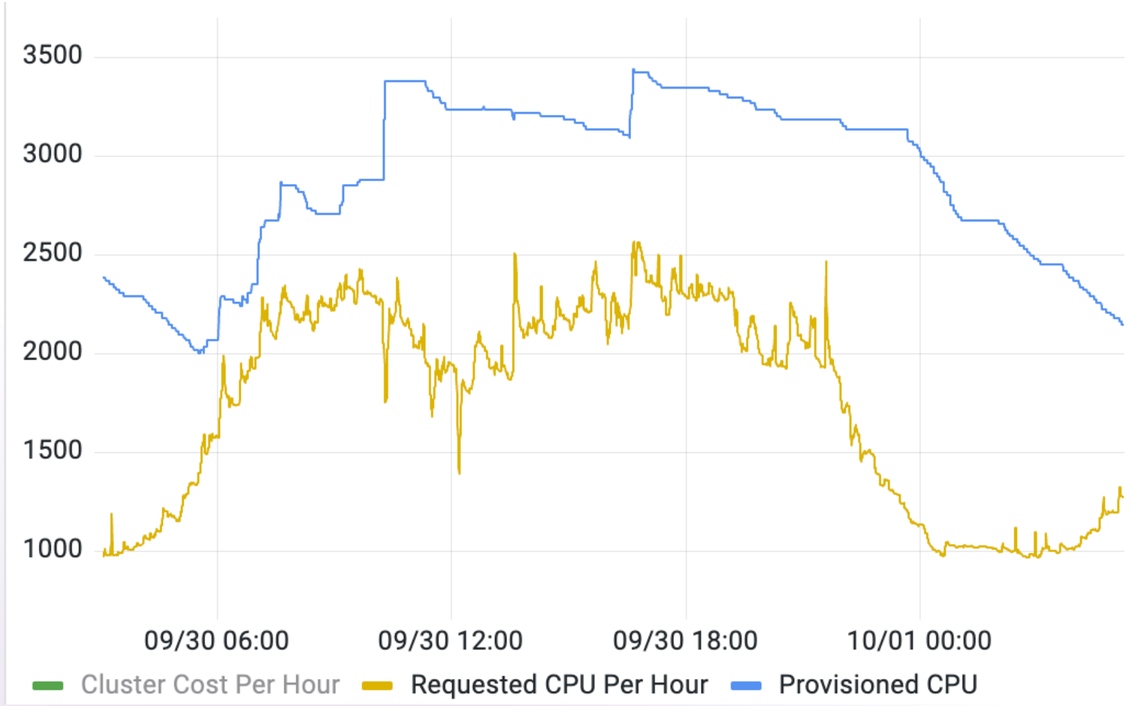
10%

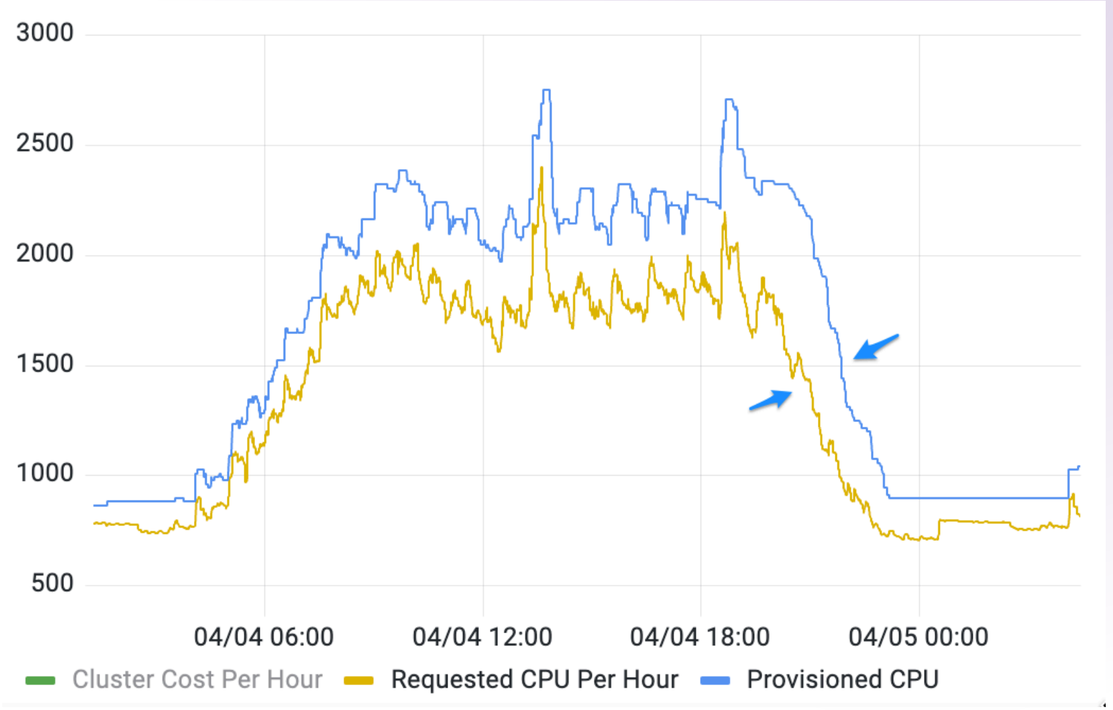
AVERAGE CPU UTILIZATION

23%

AVERAGE MEMORY UTILIZATION

WORKLOAD MANAGEMENT





RESOURCE MANAGEMENT

Rebalance: d8e9-8390

Completed

DURATION

18 min 31 sec

TIME

Start: 2024-04-16 12:00 AM
Finish: 2024-04-16 12:18 AM

SAVING ACHIEVED

67.1% 67.1% predicted

OPTIMIZED COST

$1,430.41 /mo $1,430.41 /mo predicted

NODES REPLACED

13 / 13

Current configuration

RESOURCES

Q.	NAME	CPU	GIB	CPU/H	TOTAL/MO
1 x	e2-custom-32-2… 32 CPU, 29 GiB	32 CPU	29 GiB	$0.027	$628.57
2 x	e2-custom-32-2… 32 CPU, 22.5 GiB	64 CPU	45 GiB	$0.026	$1,224.27
2 x	e2-custom-32-2… 32 CPU, 20.25 GiB	64 CPU	40.5 GiB	$0.026	$1,214.88
1 x	n2d-highcpu-80 80 CPU, 80 GiB	80 CPU	80 GiB	$0.009	$553.57
1 x	c2d-highcpu-56 56 CPU, 112 GiB	56 CPU	112 GiB	$0.009	$349.20
1 x	n2-custom-56-… 56 CPU, 34.5 GiB	56 CPU	34.5 GiB	$0.005	$202.49
5 x	n2-custom-8-16… 8 CPU, 16 GiB	40 CPU	80 GiB	$0.006	$169.55

INITIAL COMPUTE COST: $4,342.64 /mo

13 INSTANCES | 392 CPU | 421 GIB

CLUSTER: $4,342.64 /mo

Rebalanced configuration

RESOURCES

Q.	NAME	CPU	GIB	CPU/H	TOTAL/MO
2 x	n2-highcpu-96 96 CPU, 96 GiB	192 CPU	192 GiB	$0.005	$694.49
1 x	c2d-highcpu-32 32 CPU, 64 GiB	32 CPU	64 GiB	$0.009	$199.54
2 x	e2-custom-8-16… 8 CPU, 16 GiB	16 CPU	32 GiB	$0.03	$355.37
1 x	c2d-highcpu-16 16 CPU, 32 GiB	16 CPU	32 GiB	$0.009	$99.77
2 x	n2-custom-10-1… 10 CPU, 16 GiB	20 CPU	32 GiB	$0.006	$81.23

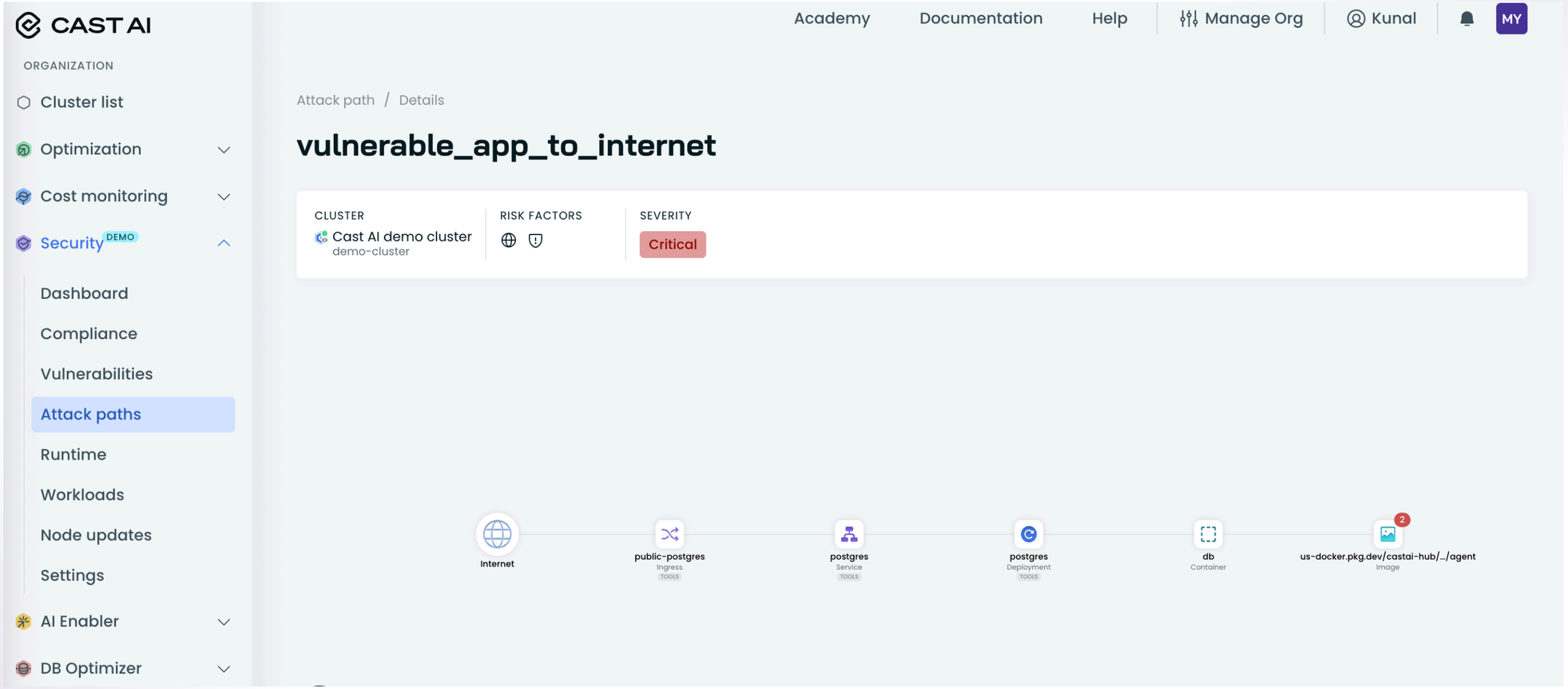
PREDICTED OPTIMIZED COMPUTE COST: $1,430.41 /mo

8 INSTANCES | 276 CPU | 352 GIB

CLUSTER: $1,430.41 /mo

SECURITY

Vulnerability scanning every 30 seconds



Chat & Connect

For any discussion feel free to connect

Kunal Das

Santanu Kumar Das

THANK YOU