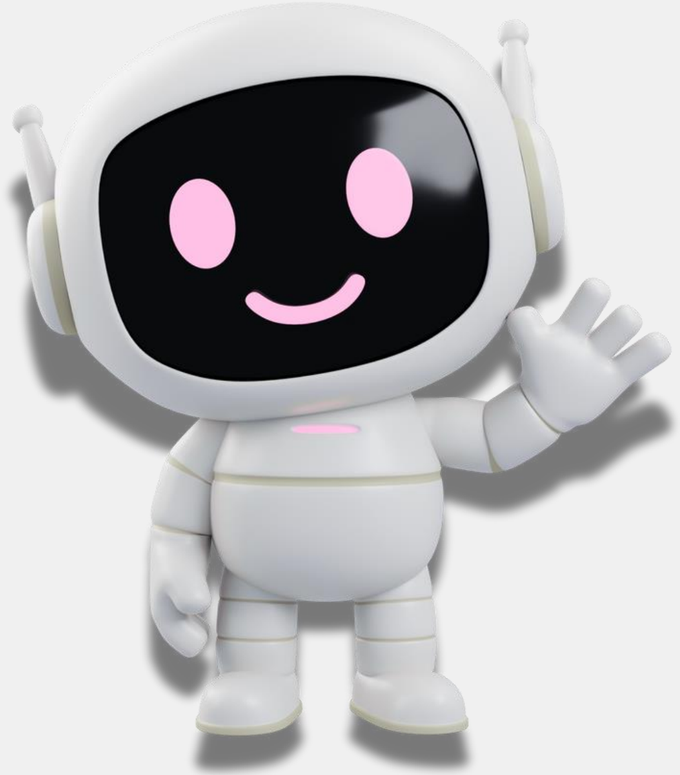
The Future of Kubernetes

AI-Driven Management

Kunal Das



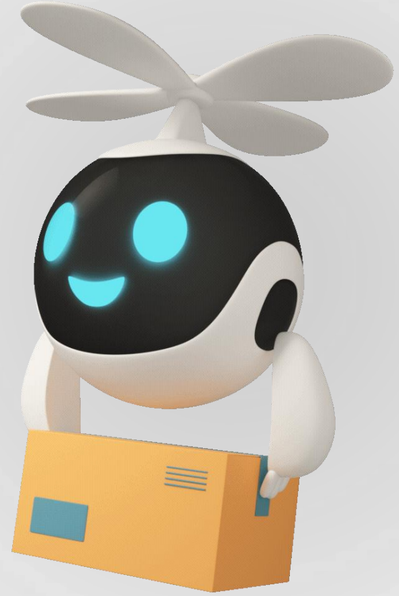
ABOUT ME

KUNAL DAS

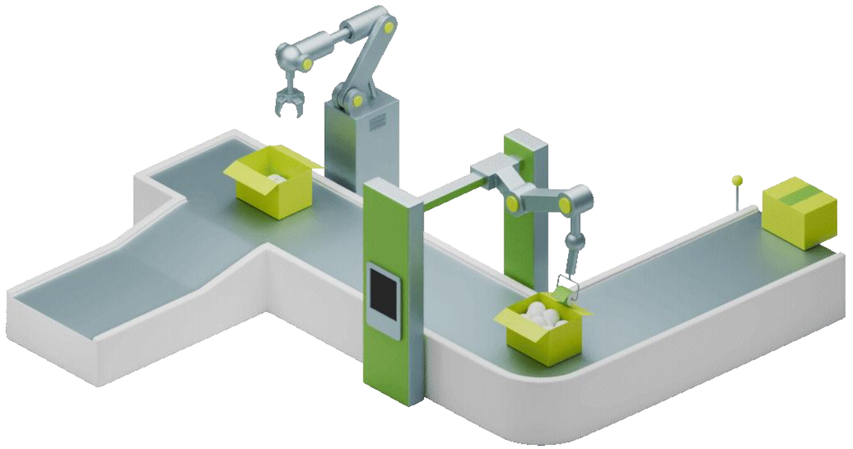
DEVELOPER ADVOCATE APAC, CASTAI

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

Contents



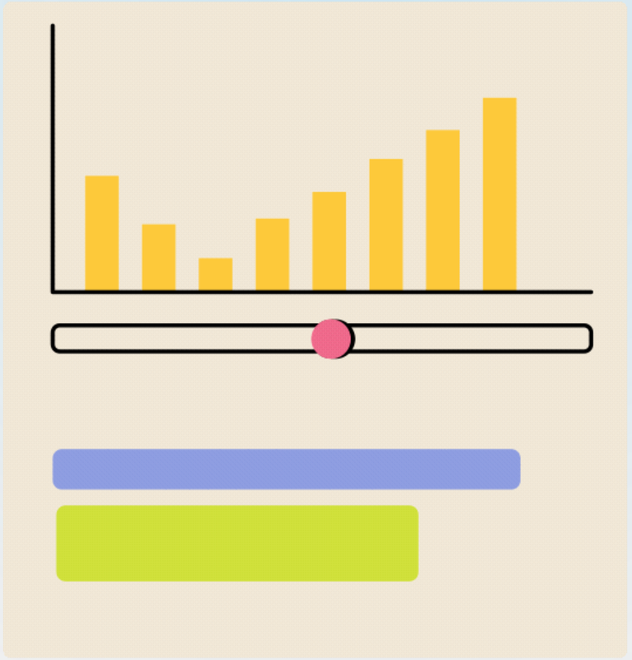
Containerized Workloads with Docker



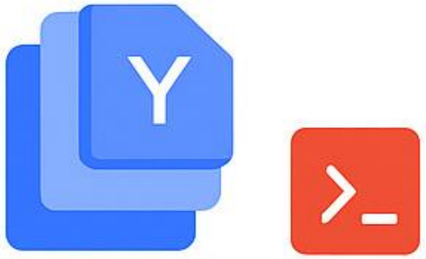
1. **Fast Deployment & Scaling**
   Lightweight containers → rapid launch & scale
2. **Efficient Resource Usage**
   Run more with less → lower costs
3. **Consistent Environments**
   No more “works on my machine” issues
4. **Easy Updates & Rollbacks**
   Redeploy quickly → less downtime

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.



Issue with Kubernetes Management



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 Scaling

Application and node scaling are managed based on current Context

Incident Response

Alerts guide teams to investigate and fix issues with standard procedures

Managing K8S Smartly

01

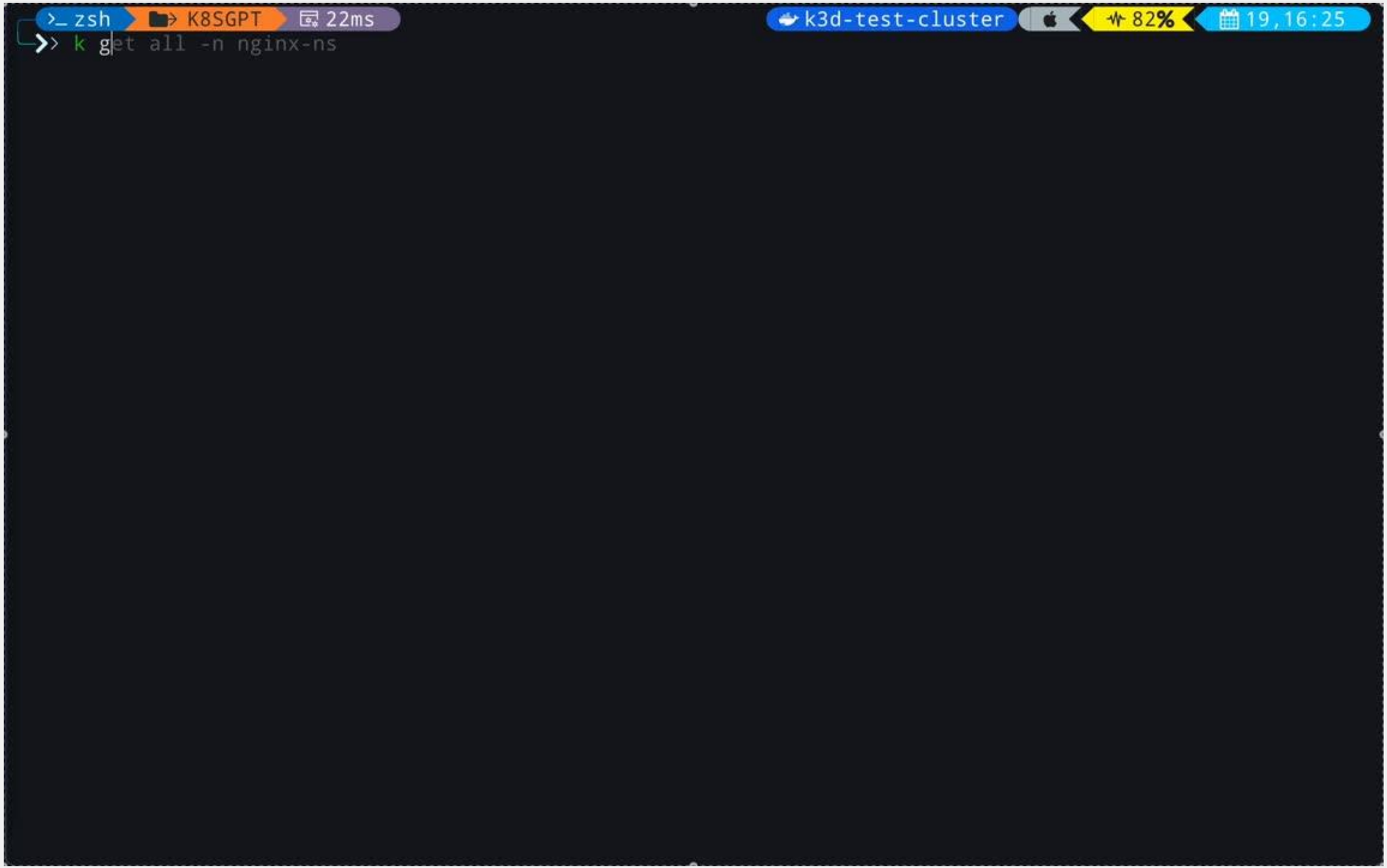
WORKLOADS & TROUBLESHOOTING

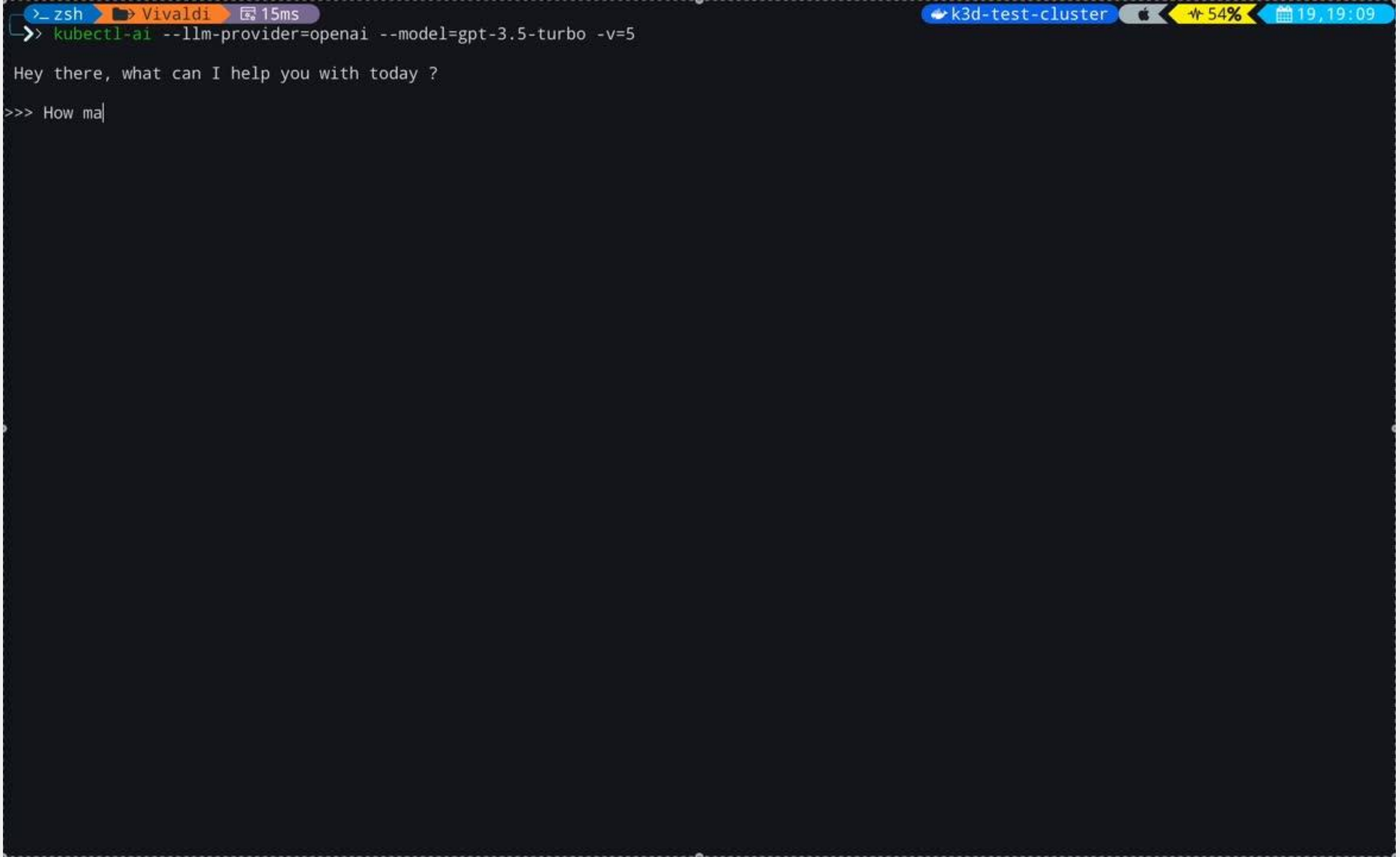
K8sGPT uses AI to quickly diagnose, explain, and resolve Kubernetes workload issues, making operations smoother and more efficient

02

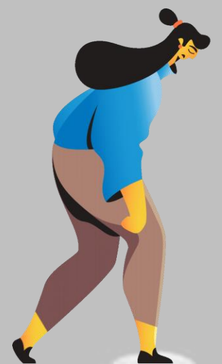
SCALLING AUTOMATION

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

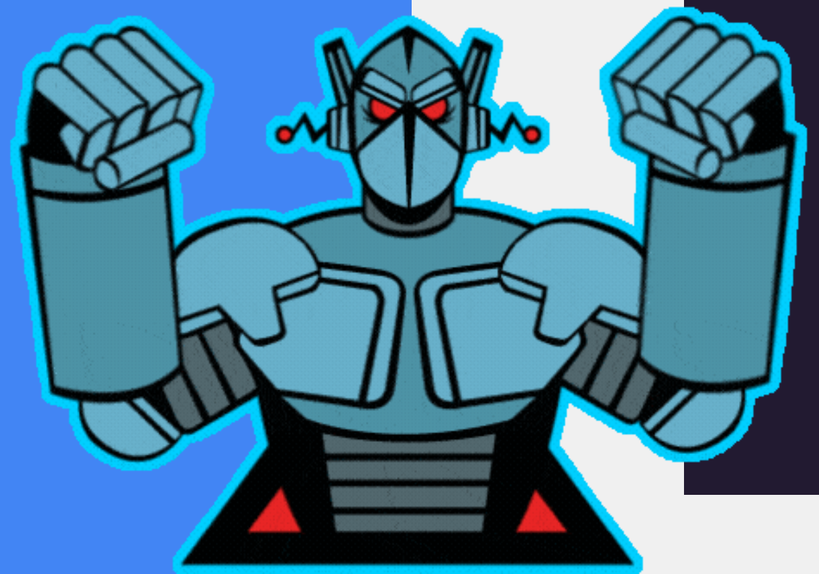

zsh
K8SGPT
22ms
k3d-test-cluster
82%
19,16:25
k get all -n nginx-ns


zsh
Vivaldi
15ms
k3d-test-cluster
54%
19, 19:09
kubectl-ai --llm-provider=openai --model=gpt-3.5-turbo -v=5
Hey there, what can I help you with today ?
>>> How ma

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

Security at Light Speed

2025 Threat Landscape

CVE-2025-23266: NVIDIA CONTAINER ESCAPE

SUPPLY CHAIN ATTACKS: 188% INCREASE IN MALICIOUS PACKAGES TARGETING K8S

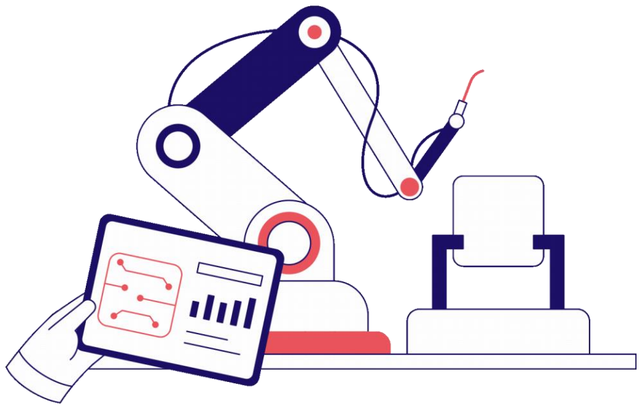
AVERAGE BREACH COST: $4.88M WITH 287 DAYS TO IDENTIFY AND CONTAIN

Why Individual Tools Fail at Scale?

Delayed Correlation

Incomplete Context

Human Bottleneck

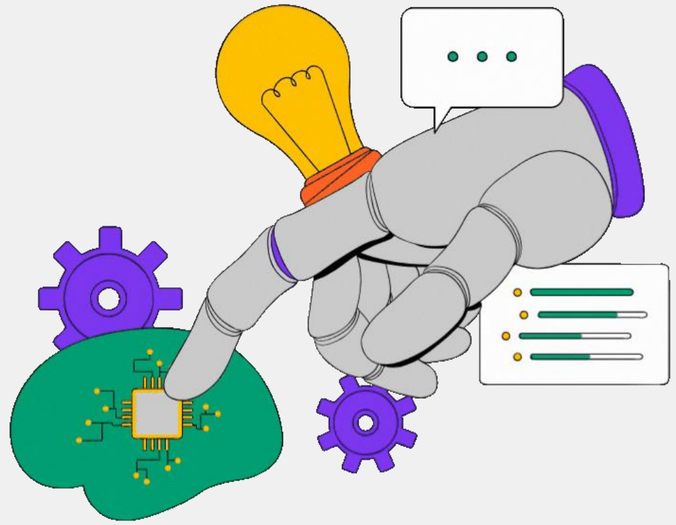


No Automated Response

Orchestrated Intelligence Platform

From Tool Collection to Unified AI Brain.

- **RESOURCE MANAGEMENT**
- **SECURITY**
- **COST OPTIMIZATION**



Question For the Audience

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

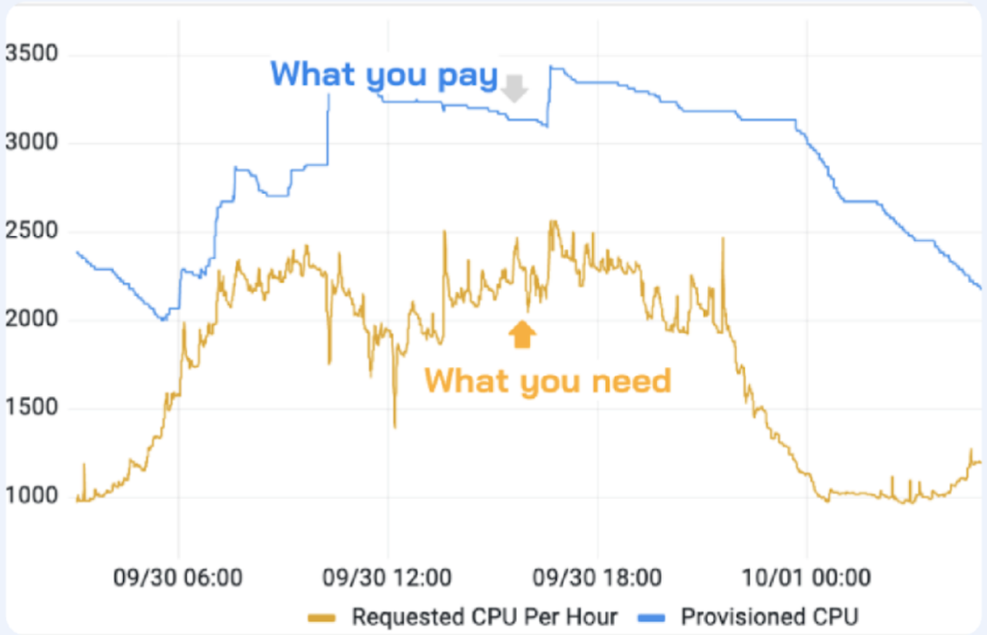
23%

AVERAGE MEMORY UTILIZATION

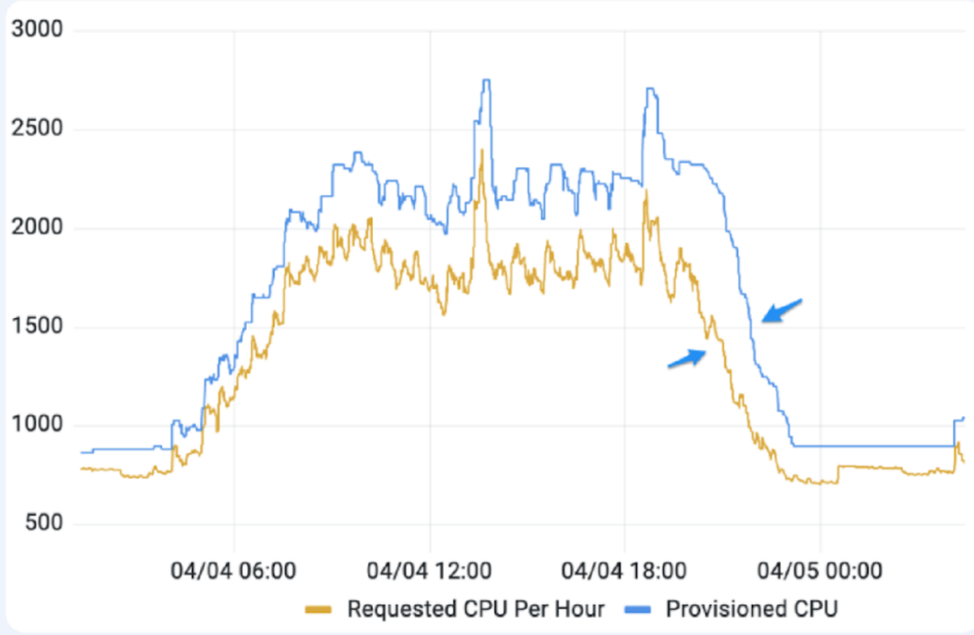
WORKLOAD MANAGEMENT

How automation enables extreme efficiency for Kubernetes Optimization

Before Optimization



After Optimization



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:

13 INSTANCES | **392** CPU | **421** GIB

$4,342.64 /mo

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:

8 INSTANCES | **276** CPU | **352** GIB

$1,430.41 /mo

CLUSTER: **$1,430.41** /mo

13 clusters

ID	NAME	PRO.	REGI...	NODES	CPU	MEMORY	CPU COST	COMPUTE COST	STATUS
	eks-demo-lio-04241039 445567108000		US East ...	4	14 CPU	44 GiB	$0.0126 /h	$180.65 /mo	Discovered READ ONLY
	eks-demo-jer-03281205 445567108000		US East ...	1	2 CPU	8 GiB	$0.0278 /h	$61.92 /mo	Discovered READ ONLY
	eks-demo-dan-0520072 445567108000		US East ...	2	4 CPU	16 GiB	$0.0311 /h	$138.24 /mo	Discovered READ ONLY
	kim-live-demo 445567108000		EU (Fran...	5	16 CPU	58 GiB	$0.0256 /h	$451.01 /mo	Discovered READ ONLY
	eks-demo-vkl-03071520 445567108000		US East ...	1	2 CPU	8 GiB	$0.0278 /h	$61.92 /mo	Discovered READ ONLY
	eks-demo-tud-04251122 445567108000		US East ...	1	2 CPU	8 GiB	$0.0278 /h	$61.92 /mo	Discovered READ ONLY
	eks-demo-ami-0523135 445567108000		US East ...	0	0 CPU	0 Bytes	$-.-- /h	$-.-- /mo	Discovered READ ONLY
	eks-sf-cod-02112018 445567108000		US East ...	0	0 CPU	0 Bytes	$-.-- /h	$-.-- /mo	Discovered READ ONLY
	phil-livedemo-0402 445567108000		EU (Fran...	1	4 CPU	8 GiB	$0.038 /h	$139.33 /mo	Discovered READ ONLY
	phil-demo-0305 445567108000		EU (Fran...	1	4 CPU	16 GiB	$0.0372 /h	$165.60 /mo	Discovered READ ONLY
	live-demo-gs-4 445567108000		EU (Fran...	3	10 CPU	36 GiB	$0.0329 /h	$352.66 /mo	Discovered READ ONLY
	eks-demo-lui-04020814 445567108000		US East ...	0	0 CPU	0 Bytes	$-.-- /h	$-.-- /mo	Discovered READ ONLY
	laurent-livedemo-0304 445567108000	+	EU (Fran...	4	16 CPU	32 GiB	$0.038 /h	$557.68 /mo	Discovered

Recap

Connected experience

Cost visibility

Multi-cloud coverage

Real-time rightsizing

Chat & Connect

Meet me at the CastAI Booth for further questions

KUNAL

THANK YOU