Cast AI: Autonomous Cloud Optimization Agent

	Video
Company	**Cast AI**
Problem	Manual K8s tuning; high costs; slow ops; wasted time
Agent	Autonomous agent; optimizes clusters (telemetry, rightsizing, autoscale, rebalance)
Human Control	Engineers set policies; review/override actions; logs & transparency
Results	40–70% lower costs; faster ops; engineers freed for key work
Why It Worked	Repetitive, data-heavy infra; ideal for AI automation

Problem:
Manual and inefficient tuning of Kubernetes clusters caused wasted cost, slow operations, and headaches for DevOps engineers.

Agent:
Cast AI's autonomous agent optimizes Kubernetes clusters—analyzing telemetry, rightsizing workloads, choosing optimal instances, autoscaling, and rebalancing nodes automatically.

Human Control:
Engineers define policies, review actions, and can override or disable automation at any time. All actions are logged and reversible, ensuring transparency.

Results:

- 40–70% lower cloud costs
- Faster optimization
- Freed up engineers for higher-value work

Why it worked:
Infrastructure optimization is repetitive, measurable, and data-heavy—ideal for AI-driven automation.

Q1: What was the problem?

- Engineers spent lots of time fixing and managing computer clusters.
- Most computers were not used much (only 10%), so money was wasted.
- Work was slow and engineers couldn't focus on building new things.

Q2: What Agent did they build?

- Cast AI made a smart robot program to manage clusters automatically.
- It checks computer use, cloud prices, and makes things run better, without humans.
- The robot uses data about usage and costs to help pick cheaper and better options.

Q3: How do humans stay in control?

- Engineers set rules for the robot to follow.
- All robot actions are recorded and can be reversed.
- If anything goes wrong, humans can stop the robot right away.
- Humans make the important decisions.

Q4: What results did they get?

- The robot saved 40–70% of cloud money for companies.
- Example: A company's bill dropped from $414/mo to $138/mo in minutes.
- Engineers spent less time on boring work, so they could create and improve things.

Q5: Why did this work?

- Fixing clusters is boring, repetitive, and all about numbers—great for robots.
- The robot makes lots of small improvements faster than humans.
- Engineers still watch over everything to keep it safe and working.
- Other companies with cloud computers can use the same idea.

Sources:

- VentureBeat – "Akamai saves 70% using AI agents orchestrated by Kubernetes"

  https://venturebeat.com/data-infrastructure/cutting-cloud-waste-at-scale-akamai-saves-70-using-ai-agents-orchestrated-by-kubernetes

- Cast AI Blog – "Automate Kubernetes Deployment to Reduce Your Cloud Bill by 60%" https://cast.ai/blog/category/cloud-cost-optimization/
- Cast AI Official Website – cast.ai