Kubernetes Cost Optimization: How to Actually Do It

Annie Talvasto

San Francisco, CA
July 2021

What to expect?

Kubernetes Cost Optimization: How to Actually Do It

Annie Talvasto - Product Marketing Manager at CAST AI

@annietalvasto
cast.ai

Cloud-native technology marketer & Kubernetes specialist

Kubernetes & CNCF meetup co-organizer
Co-host of the Cloud gossip podcast - cloudgossip.net

First:

Why does cost optimization matter?

Why is cost optimization complex?

a1.large	$0.051	2	4 GiB	EBS Only	Up to 10 Gigabit
a1.xlarge	$0.102	4	8 GiB	EBS Only	Up to 10 Gigabit
a1.2xlarge	$0.204	8	16 GiB	EBS Only	Up to 10 Gigabit
a1.4xlarge	$0.408	16	32 GiB	EBS Only	Up to 10 Gigabit
a1.metal	$0.408	16	32 GiB	EBS Only	Up to 10 Gigabit
t4g.nano	$0.0042	2	0.5 GiB	EBS Only	Up to 5 Gigabit
t4g.micro	$0.0084	2	1 GiB	EBS Only	Up to 5 Gigabit
t4g.small	$0.0168	2	2 GiB	EBS Only	Up to 5 Gigabit
t4g.medium	$0.0336	2	4 GiB	EBS Only	Up to 5 Gigabit
t4g.large	$0.0672	2	8 GiB	EBS Only	Up to 5 Gigabit
t4g.xlarge	$0.1344	4	16 GiB	EBS Only	Up to 5 Gigabit
t4g.2xlarge	$0.2688	8	32 GiB	EBS Only	Up to 5 Gigabit
t3.nano	$0.0052	2	0.5 GiB	EBS Only	Up to 5 Gigabit
t3.micro	$0.0104	2	1 GiB	EBS Only	Up to 5 Gigabit
t3.small	$0.0208	2	2 GiB	EBS Only	Up to 5 Gigabit
t3.medium	$0.0416	2	4 GiB	EBS Only	Up to 5 Gigabit
t3.large	$0.0832	2	8 GiB	EBS Only	Up to 5 Gigabit
t3.xlarge	$0.1664	4	16 GiB	EBS Only	Up to 5 Gigabit
t3.2xlarge	$0.3328	8	32 GiB	EBS Only	Up to 5 Gigabit

Cost optimization, FinOps, FinDev, Cloud Cost Management, FinDevSecOps, Cloud Economics, Cloud Financial Management, FinDev…

You can do many things:

Policies
Unused or unattached resources
Idle resources
Heat maps
Right sizing
Reserved instances
Spot instances
Multi cloud
Understand your bill
Native cost management tools

Choosing the right compute services
Autoscaling
Set budgets, limitations & alerts
Allocation
Structure of your resources
Tagging & labelling
Review & repeat
Goal setting
Dedicated host
Forecasting

Cost optimization, FinOps, FinDev, Cloud Cost Management, FinDevSecOps, Cloud Economics, Cloud Financial Management, FinDev…

You can do many things:

Policies
Unused or unattached resources
Idle resources
Heat maps
Right sizing
Reserved instances
Spot instances
Multi cloud
Understand your bill
Native cost management tools

Choosing the right compute services
Autoscaling
Set budgets, limitations & alerts
Allocation
Structure of your resources
Tagging & labelling
Review & repeat
Goal setting
Dedicated host
Forecasting

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. Check if your workload is spot-ready
5. Cherry-pick spot instances
6. Bid your price on spot
7. Use mixed instances
8. Make multiple regions work for you

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. Check if your workload is spot-ready
5. Cherry-pick spot instances
6. Bid your price on spot
7. Use mixed instances
8. Make multiple regions work for you

CPU
Memory
Storage
Network

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. Check if your workload is spot-ready
5. Cherry-pick spot instances
6. Bid your price on spot
7. Use mixed instances
8. Make regions work for you

AWS offers more than 150 different instance types.

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. **Verify storage transfer limitations**
4. Check if your workload is spot-ready
5. Cherry-pick spot instances
6. Bid your price on spot
7. Use mixed instances
8. Make multiple regions work for you

Each application out there comes with its unique storage needs. Make sure that the VM you pick has the storage throughput your workloads need.

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. **Check if your workload is spot-ready**
5. Cherry-pick spot instances
6. Bid your price on spot
7. Use mixed instances
8. Make multiple regions work for you

How much time does your workload need to finish the job?
Is it mission- and time-critical?
Can it handle interruptions?
Is it tightly coupled between instance nodes?
What tools are you going to use to move your workload when cloud provider pulls the plug?

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. Check if your workload is spot-ready
5. **Cherry-pick spot instances**
6. Bid your price on spot
7. Use mixed instances
8. Make multiple regions work for you

Once you pick an instance, check its frequency of interruption – the rate at which this instance reclaimed capacity during the trailing month.
You can see it the AWS Spot Instance Advisor in ranges of <5%, 5-10%,10-15%,15-20%, and >20%

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. Check if your workload is spot-ready
5. Cherry-pick spot instances
6. **Bid your price on spot**
7. Use mixed instances
8. Make multiple regions work for you

Here's a rule of thumb: Set the maximum price to one that equals the on-demand price.

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. Check if your workload is spot-ready
5. Cherry-pick spot instances
6. Bid your price on spot
7. **Use mixed instances**
8. Make multiple regions work for you

A mixed-instance strategy gets you great availability and performance at a reasonable cost.

Cost optimization:

8 best practices

1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. Check if your workload is spot-ready
5. Cherry-pick spot instances
6. Bid your price on spot
7. Use mixed instances
8. **Make multiple regions work for you**

Configure multiple node groups,
Scope each of them to a single cloud region,
automate your deployments to take advantage of price fluctuations between regions.

What is actually holding teams back?

State of FinOps 2021 report:

What challenges do FinOps teams face?
- 39% Getting engineers to take action
- 33% Dealing with shared costs
- 26% Accurate Forecasting
- 24% Reducing waste or unused resources

What tools do FinOps practitioners use?
- 26% native tools
- 24% 3rd party vendor platforms
- **32% Spreadsheets**

Nearly half of respondents say that they have "Little to no automation"

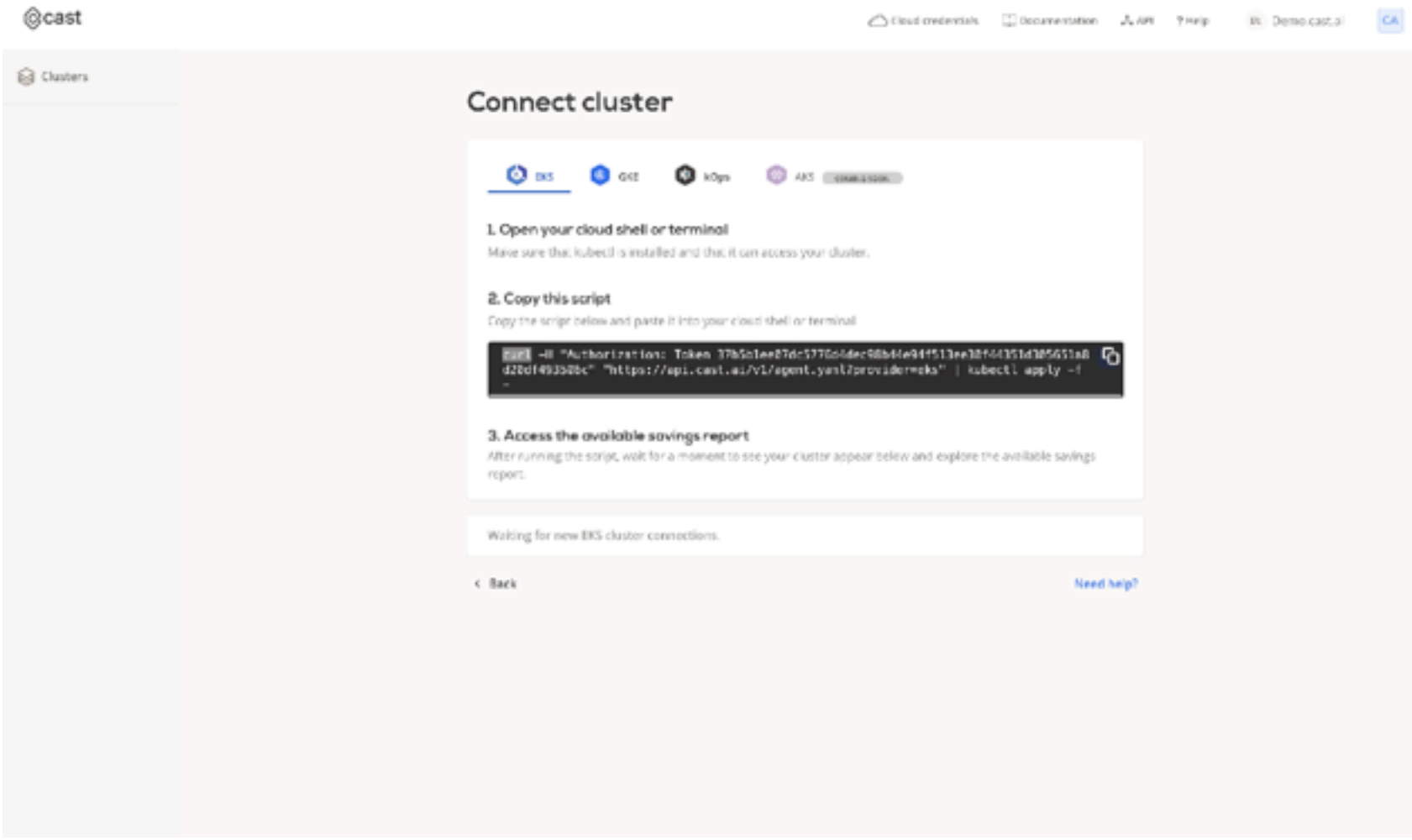
CNCF project solutions

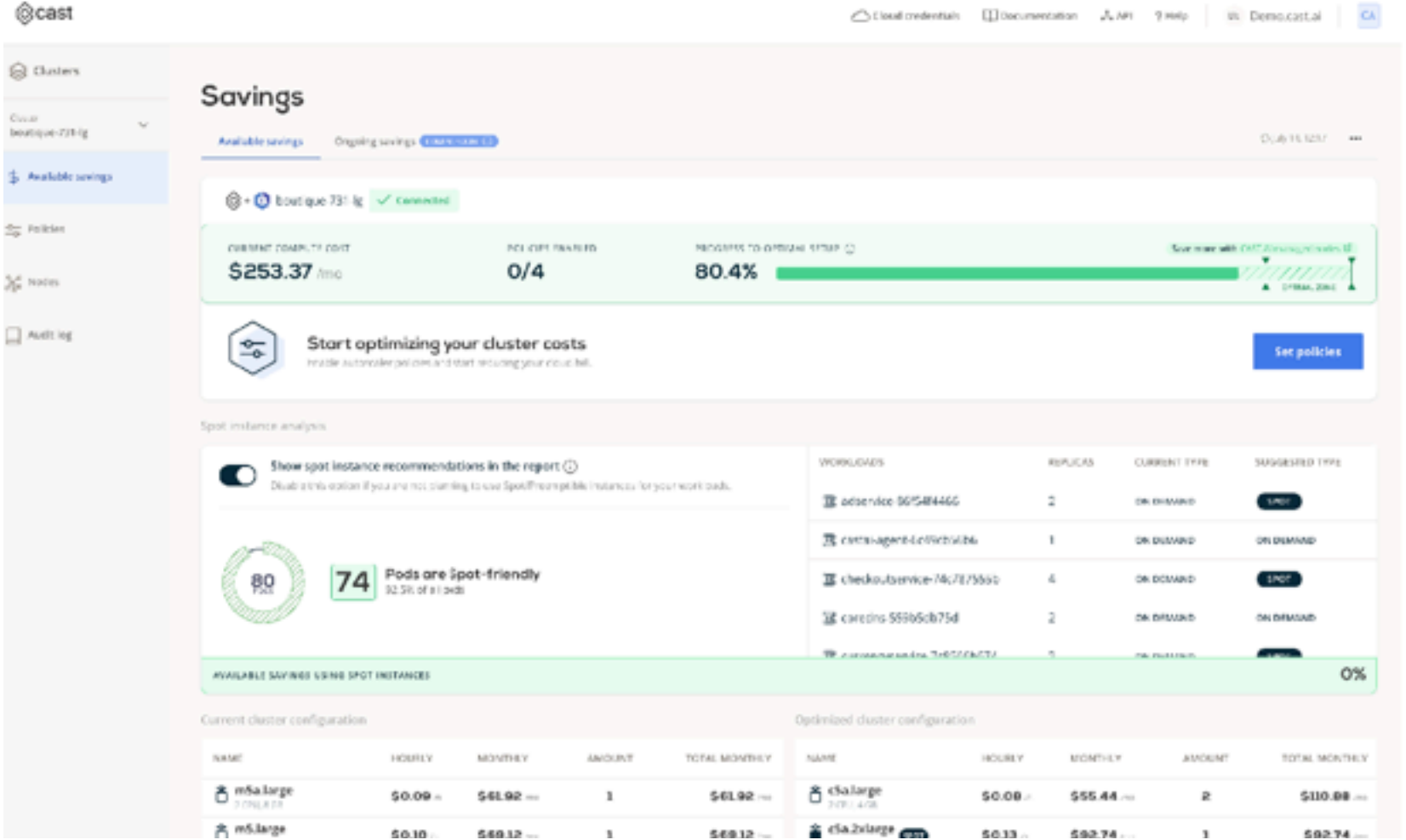
Event driven autoscaling: KEDA
Application Management: Helm
Monitoring: Prometheus
Container workflow: Flux
Service Mesh: Kuma

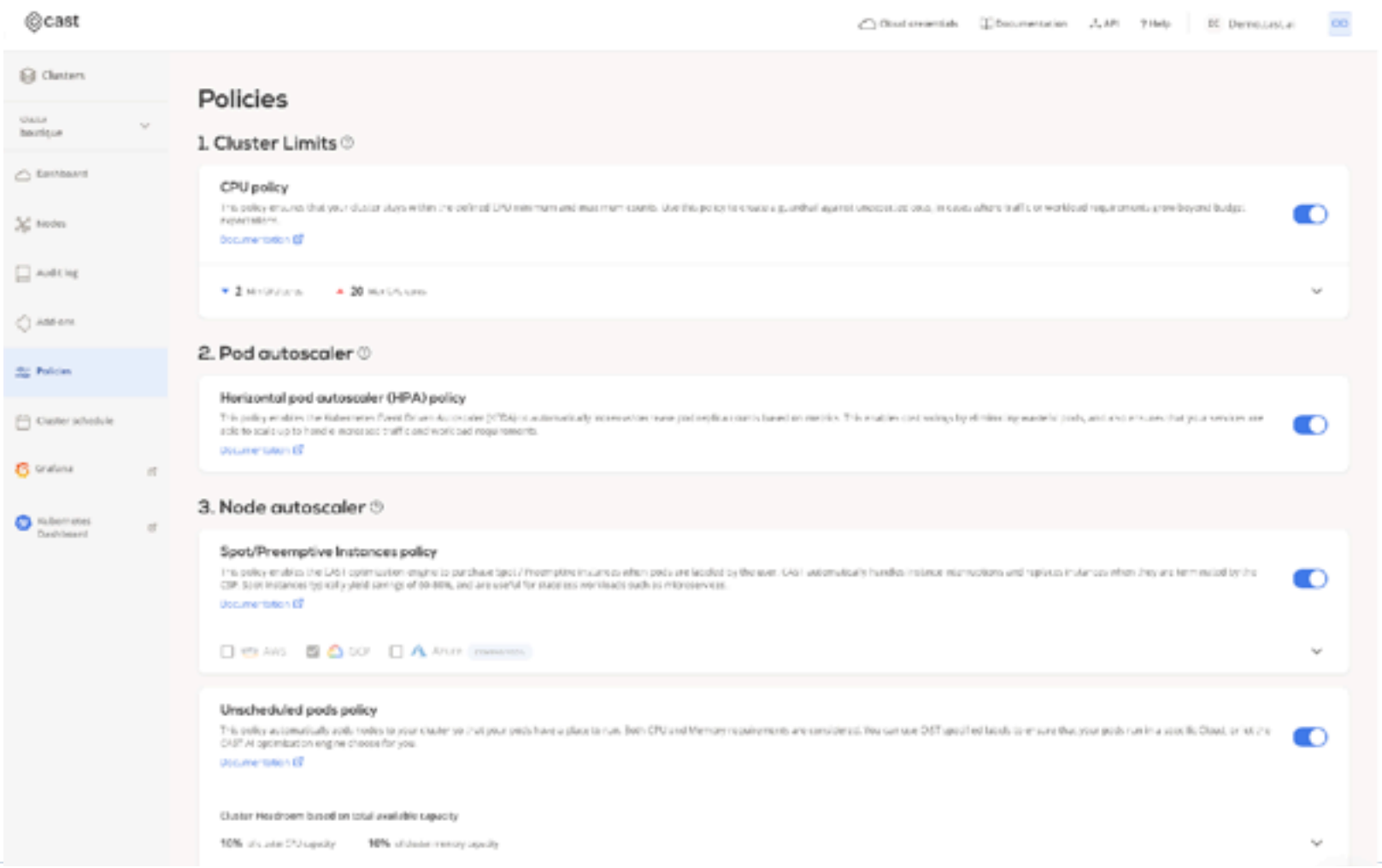

Comprehensive solution:
AI & automation driven cost optimization

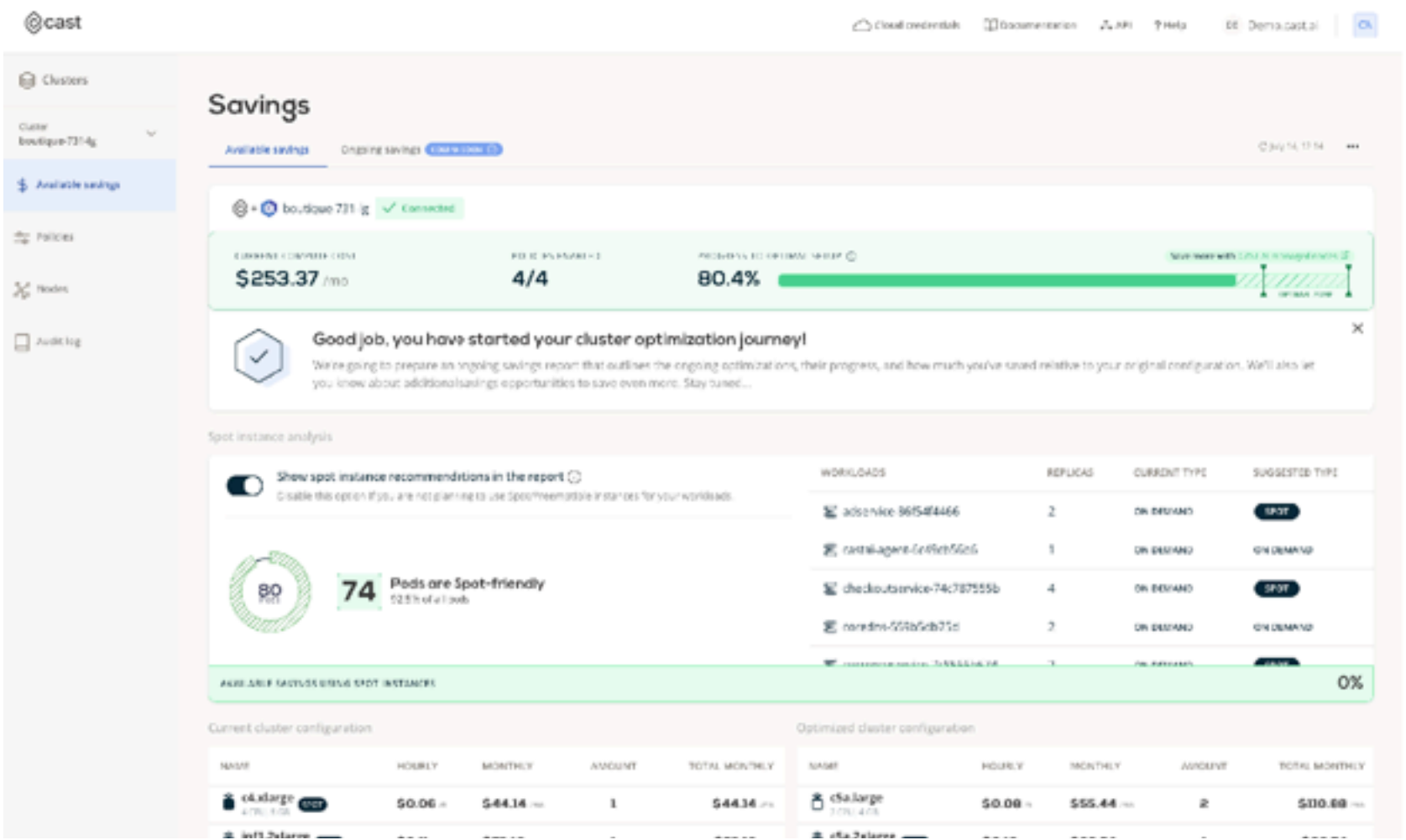
Comprehensive solution:

CAST AI demo


cast
Clusters
Connect cluster
EKS
GKE
kOps
AKS
1. Open your cloud shell or terminal
2. Copy this script
3. Access the available savings report
Waiting for new EKS cluster connections.
Back
Need help?


cast
Clusters
Available savings
Policies
Nodes
Audit log
Savings
Available savings
Ongoing savings
$253.37 /mo
0/4
80.4%
Start optimizing your cluster costs
See policies
Show spot instance recommendations in the report
80
74
Pods are Spot-friendly
AVAILABLE SAVINGS USING SPOT INSTANCES
0%
Current cluster configuration
Optimized cluster configuration


cast
Clusters
Dashboard
Nodes
Audit log
Add-ons
Policies
Cluster schedule
Grafana
Kubernetes Dashboard
Policies
1. Cluster Limits
CPU policy
Documentation
2. Pod autoscaler
Horizontal pod autoscaler (HPA) policy
Documentation
3. Node autoscaler
Spot/Preemptive Instances policy
Documentation
Unscheduled pods policy
Documentation
Cluster Headroom based on total available capacity


cast
Clusters
Savings
Available savings
Ongoing savings
$253.37 /mo
4/4
80.4%
Good job, you have started your cluster optimization journey!
Spot instance analysis
Show spot instance recommendations in the report
74
Pods are Spot-friendly
0%
Current cluster configuration
Optimized cluster configuration

Want to learn more & get a surprise gift? Go to:

cast.ai/annie

Recap

8 best practices

CNCF project solutions
Event driven autoscaling: KEDA
Application Management: Helm
Monitoring: Prometheus
Container workflow: Flux
Service Mesh: Kuma

Automation option - CAST AI

Resources:

FinOps Foundation principles:
https://www.finops.org/framework/principles/

8 best practices to reduce your AWS bill for Kubernetes:
https://cast.ai/blog/8-best-practices-to-reduce-your-aws-bill-for-kubernetes/

Slides: https://github.com/AnnieTalvasto/presentations

Thank you!