



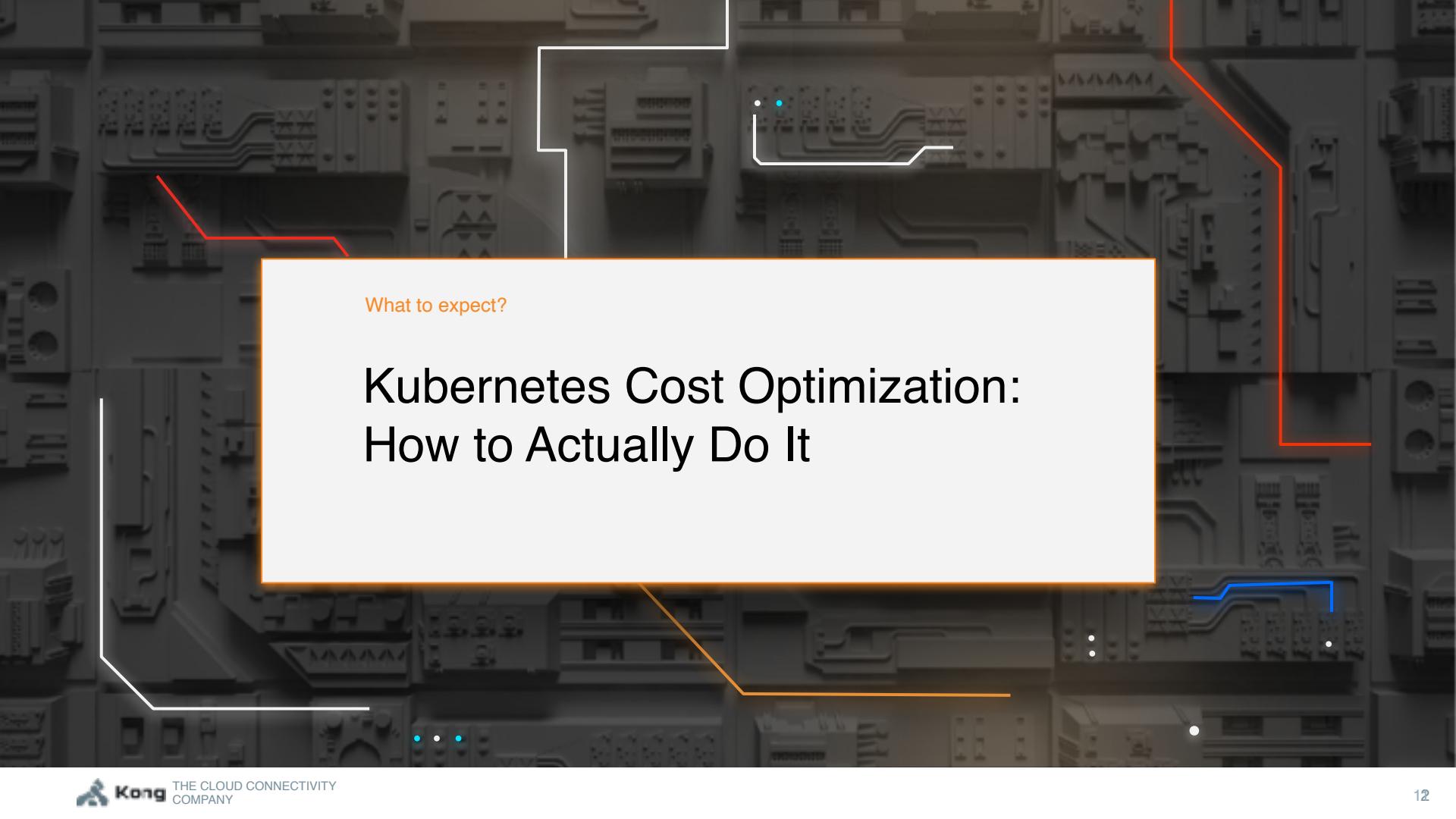
THE CLOUD
CONNECTIVITY
COMPANY

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?

t1.small	\$0.051	2	4 GiB	EBS Only	Up to 10 Gigabit
a1.large	\$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
t5.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
t5.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
t5.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

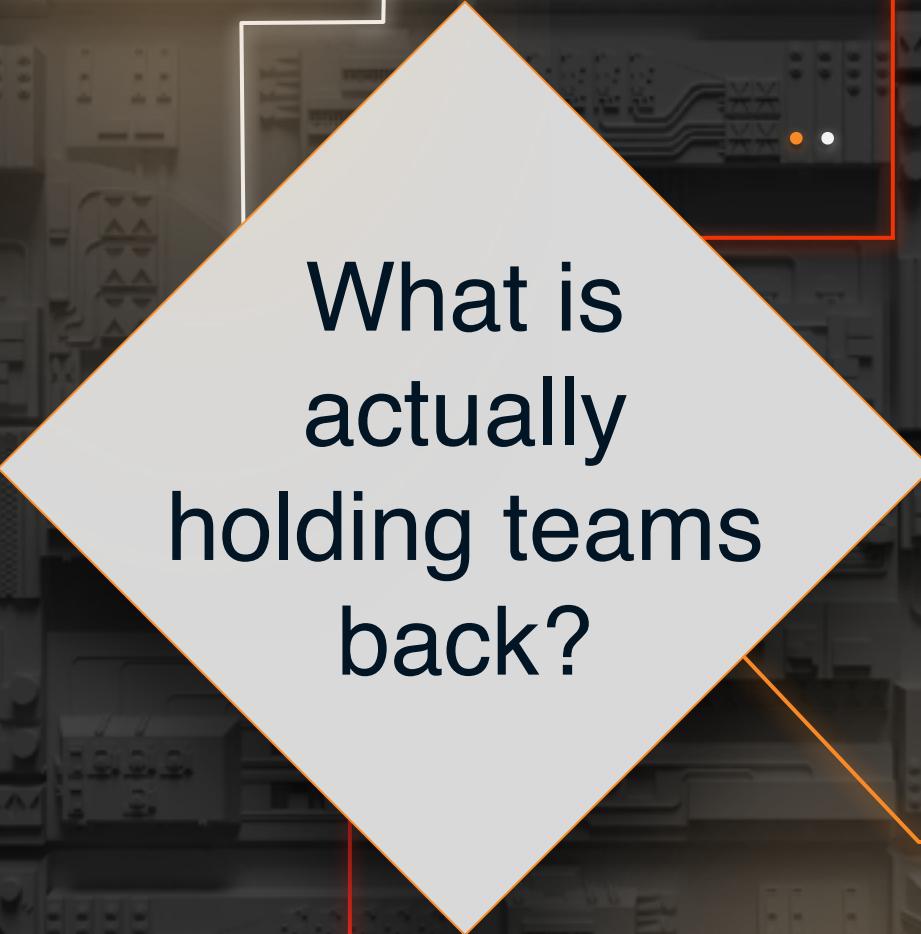
1. Define your requirements
2. Choose the right instance types
3. Verify storage transfer limitations
4. Check if your workload is spot-ready
5. Cherry-pick spot instances
6. Bid your price on spot
- 7. Use mixed instances**
8. Make multiple regions work for you

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

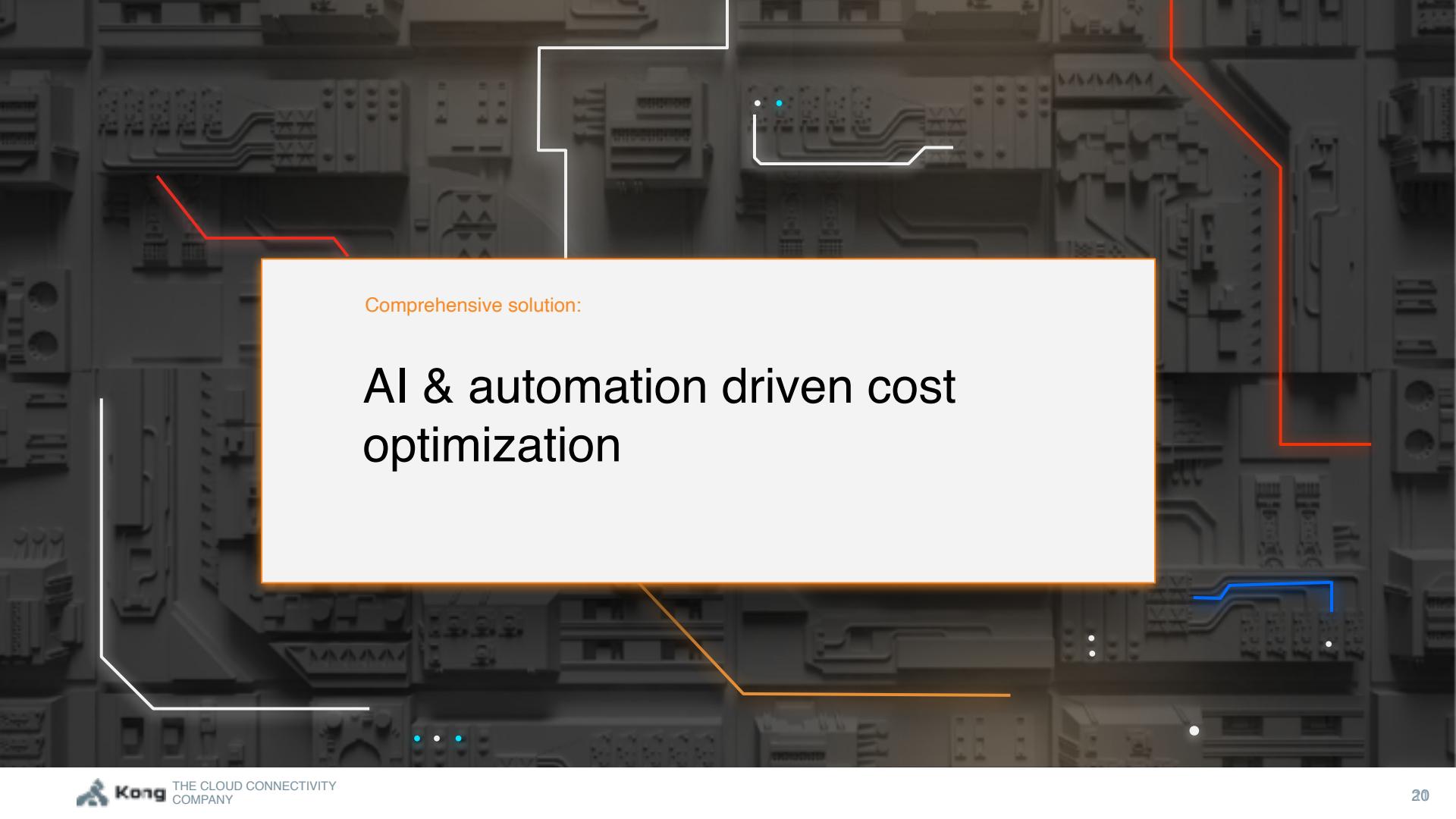
Application Management: Helm

Monitoring: Prometheus

Container workflow: Flux

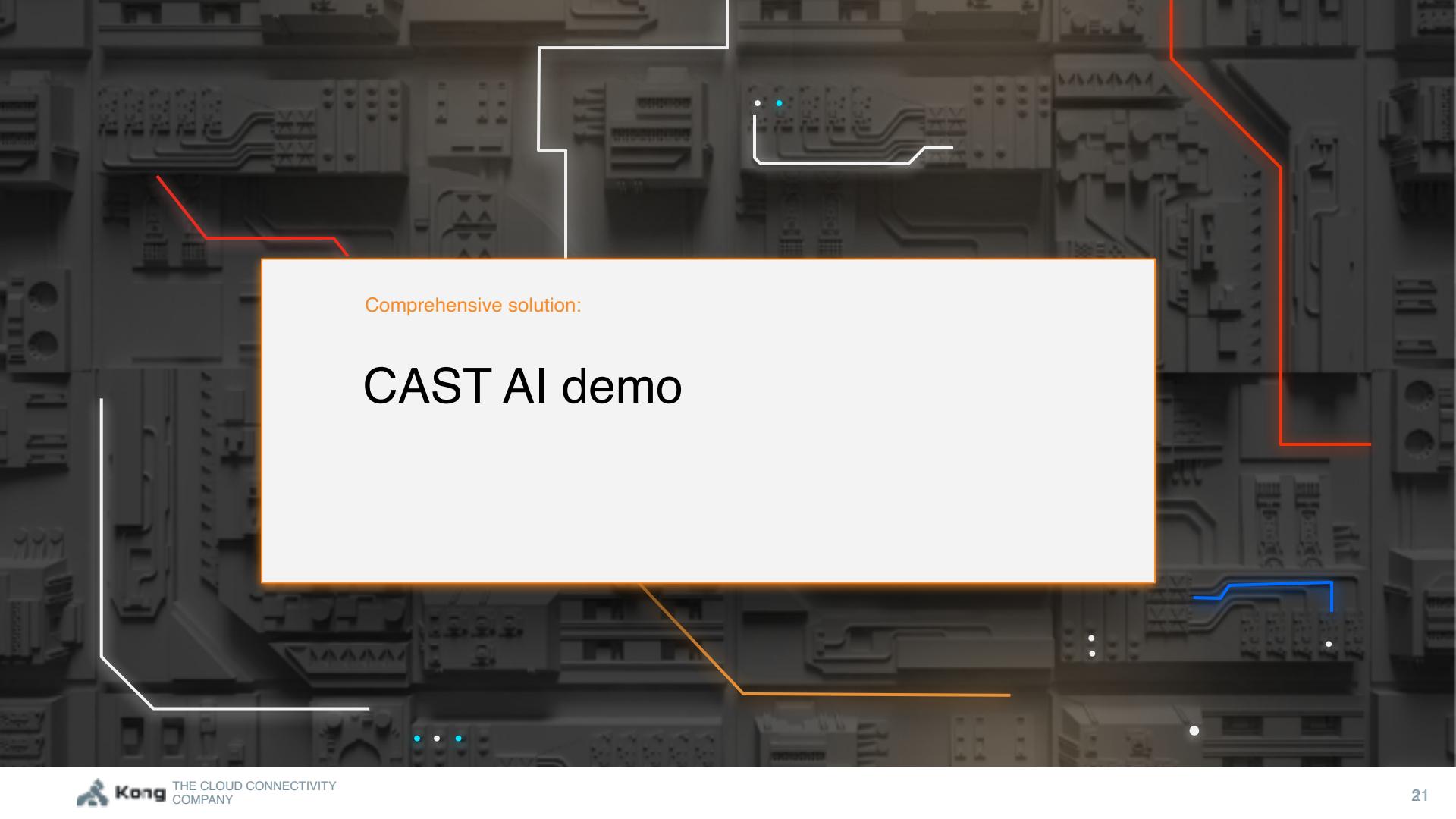
Service Mesh: Kuma





Comprehensive solution:

AI & automation driven cost optimization



Comprehensive solution:

CAST AI demo

Connect cluster



1. Open your cloud shell or terminal

Make sure that kubectl is installed and that it can access your cluster.

2. Copy this script

Copy the script below and paste it into your cloud shell or terminal

```
curl -H "Authorization: Token 37b501ee87de5776c4dec98b64e944f513fe38f64151d305651e8 d2bd149350cc" "https://api.cast.ai/v1/agent.yaml?provider=eks" | kubectl apply -f -
```

3. Access the available savings report

After running the script, wait for a moment to see your cluster appear below and explore the available savings report.

Waiting for new EKS cluster connections.

< Back

Need help?

[Clusters](#)Cluster
boutique-731-fg[Available savings](#)[Policies](#)[Nodes](#)[Audit log](#)

Savings

Available savings

Ongoing savings [Edit filters](#)

July 15, 2021

...

+ boutique-731-fg ✓ Connected

CURRENT COMPUTE COST

\$253.37 /mo

INCLUDED FREE

0/4

MOVING TO SPOT RATE ⓘ

80.4%

Save more with AWS Direct Connect, EC2

AWS Lambda, and Amazon VPC



Start optimizing your cluster costs

Create automation policies and start reducing your cloud bill.

[See policies](#)

Spot instance analysis

 Show spot instance recommendations in the region ⓘ

Disable this option if you are not planning to use SpotFrom public instances for your workloads.

80
PCT74 Pods are spot-friendly
92.5% of all pods

WORKLOADS	REPlicas	CURRENT TYPE	SUGGESTED TYPE
adservice-805484465	2	ON DEMAND	SPOT
cartagent-617f1fb6	1	ON DEMAND	ON DEMAND
checkoutservice-747875550	4	ON DEMAND	SPOT
cartcons-559b6cb75d	2	ON DEMAND	ON DEMAND
ecommerce-74020c7f2	5	ON DEMAND	ON DEMAND

AVAILABLE SAVINGS USING SPOT INSTANCES

0%

Current cluster configuration

NAME	HOURLY	MONTHLY	AMOUNT	TOTAL MONTHLY
m5large	\$0.09 ...	\$61.92 ...	1	\$61.92 ...
m5large	\$0.10 ...	\$68.12 ...	1	\$68.12 ...

Optimized cluster configuration

NAME	HOURLY	MONTHLY	AMOUNT	TOTAL MONTHLY
d1large	\$0.08 ...	\$55.44 ...	2	\$110.88 ...
d1.2xlarge	\$0.13 ...	\$92.74 ...	1	\$92.74 ...



Clusters

Usage
Budget

Dashboard

Nodes

Audit Log

Add-ons

Policies

Cluster schedule

Metrics

Well-Balanced
Dashboard

Policies

1. Cluster Limits

CPU policy

This policy ensures that your cluster stays within the defined CPU minimum and maximum bounds. Use this policy to create a guardrail against unnecessary waste, in cases where traffic or workload requirements grow beyond budget.


[Documentation](#)
▼ 2% utilization ▲ 20% utilization


2. Pod autoscaler

Horizontal pod autoscaler (HPA) policy

This policy enables the Horizontal Pod AutoScaler (HPA) to automatically increase or decrease pod replicas based on metrics. This enables cost savings by eliminating wasted capacity, and also ensures that your services are able to scale up to handle increased traffic and workload requirements.


[Documentation](#)

3. Node autoscaler

Spot/Preemptive Instances policy

This policy enables the GCP spot instance engine to purchase spot / preemptible instances when pods are located by the user. GCP automatically handles instance termination and replaces instances when they are terminated by the GCP spot instance engine (yield savings of 90-80%), and are useful for stateless workloads such as microservices.


[Documentation](#)
 AWS GCP Azure OpenShift


Unscheduled pods policy

This policy automatically adds nodes to your cluster so that your pods have a place to run. Both CPU and Memory requirements are considered. You can use GCP specified loads to ensure that your pods run in a specific Cloud, or let the GCP AI optimization engine choose for you.


[Documentation](#)
Cluster Headroom based on total available capacity
10% cluster CPU capacity
10% cluster memory capacity


[Clusters](#)Cluster
boutique-731-g[Available savings](#)[Policies](#)[Nodes](#)[Audit log](#)

Savings

[Available savings](#)[Ongoing savings](#) [Check now](#)

July 14, 17:54


 boutique-731-g ✓ Connected

Estimate compute cost

\$253.37/mo

Pods analyzed

4/4

Pods ready to optimize

80.4%

Save more with CloudWatch Metrics

**Good job, you have started your cluster optimization journey!**

We're going to prepare an ongoing savings report that outlines the ongoing optimizations, their progress, and how much you've saved relative to your original configuration. We'll also let you know about additional savings opportunities to save even more. Stay tuned...



Spot instance analysis

[Show spot instance recommendations in the report](#)

Disable this option if you are not planning to use spot/mixed instances for your workloads.



80

Pods are Spot-friendly

52.5% of all pods

WORKLOADS	REPLICAS	CURRENT TYPE	SUGGESTED TYPE
adservice-86f54ff466	2	ON DEMAND	SPOT
cartmanagement-6effef565	1	ON DEMAND	ON DEMAND
checkoutservice-74c787555b	4	ON DEMAND	SPOT
coredns-57fa5cb71d	2	ON DEMAND	ON DEMAND
kong-533a2581a7e	1	ON DEMAND	ON DEMAND

PERCENTAGE USING SPOT INSTANCES

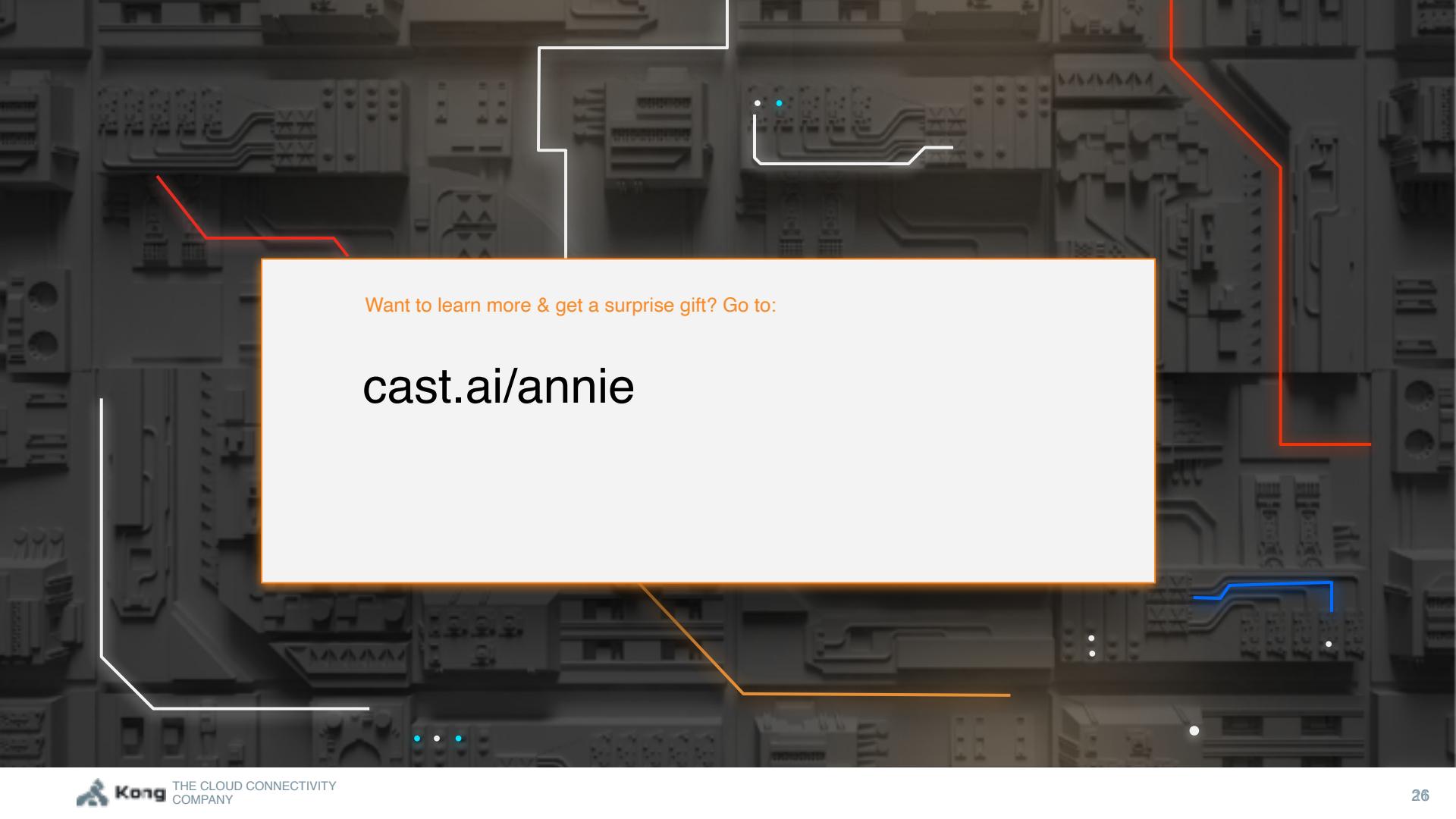
0%

Current cluster configuration

Optimized cluster configuration

NAME	HOURLY	MONTHLY	AMOUNT	TOTAL MONTHLY	NAME	HOURLY	MONTHLY	AMOUNT	TOTAL MONTHLY
cSalage-479c4ca	\$0.06/mo	\$44.14/mo	1	\$44.14/mo	cSalage-479c4ca	\$0.08/mo	\$55.44/mo	2	\$100.88/mo
mfa-balancer	---	---	-	---	mfa-balancer	---	---	-	---





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!

