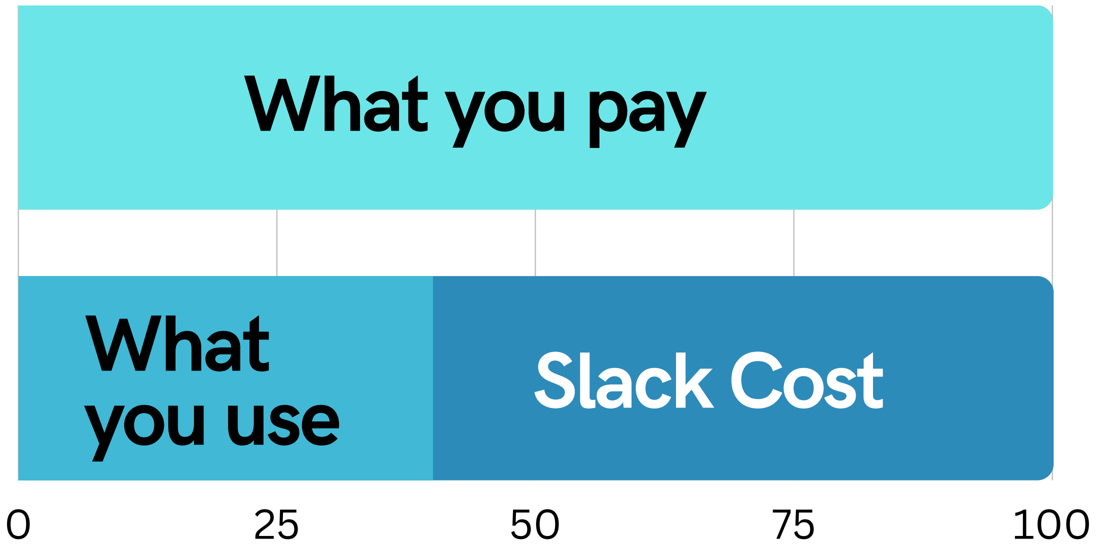
How to reduce Kubernetes Slack Cost ?



What is a Slack Cost ?

Slack cost = Requested Resources − Real Usage

Potential Waste:
~ $10/month/1 GiB

This can grow huge as you scale

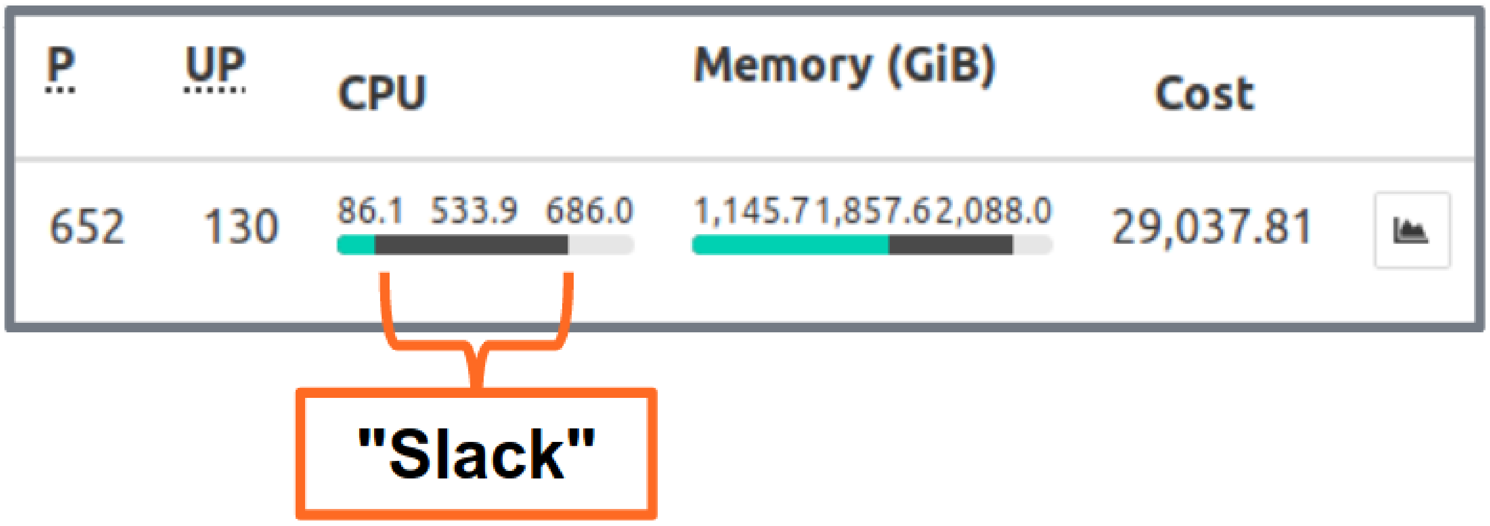
How to reduce Kubernetes Slack Cost ?

- Identify Slacks
- Lower Resource Requests
- Adapt Vertical Pod Auto Scaler

Identify Slacks

Kubernetes Resource Report (kube-resource-report) helps to visualize the slack cost

Helps to right size the requests for the containers in a pod.



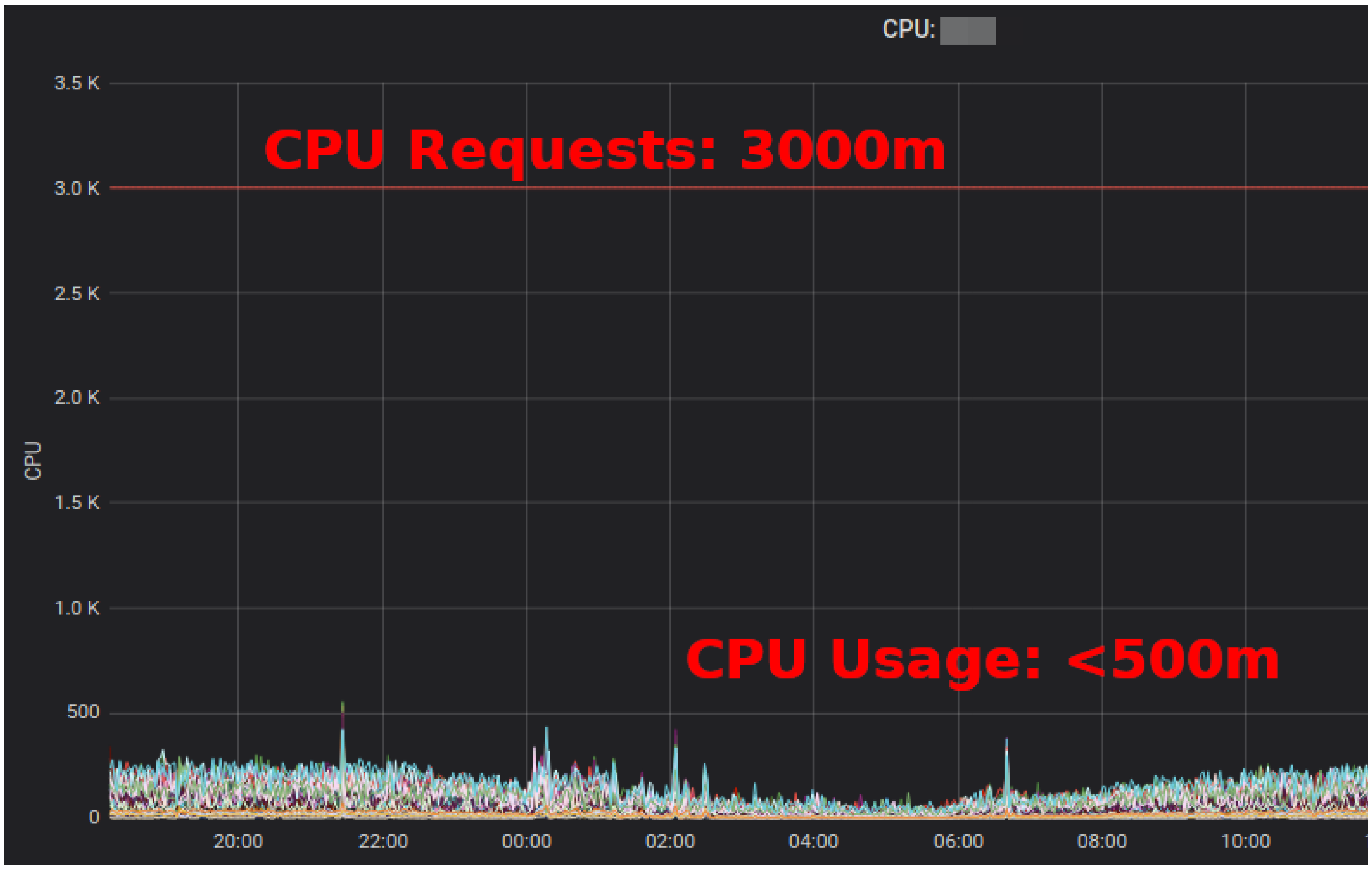
P	UP	CPU	Memory (GiB)	Cost	
652	130	86.1 533.9 686.0	1,145.7 1,857.6 2,088.0	29,037.81	

"Slack"

2

Lower Resource Requests

Leverage 'Grafhana' to look at CPU/memory usage over time to set the right resource requests



CPU Requests: 3000m

CPU Usage: <500m

Adapt Vertical Pod Auto Scaler

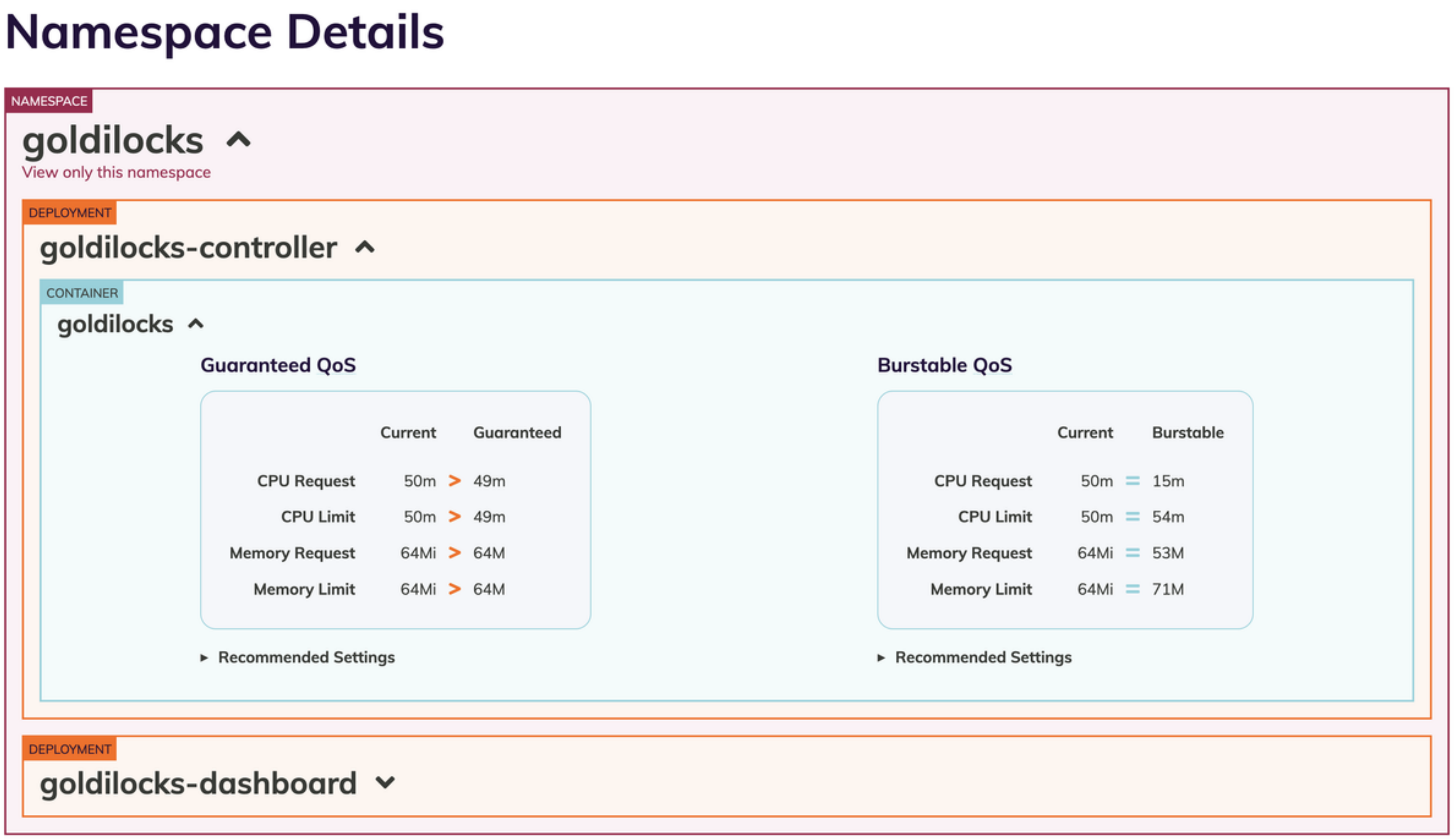
It is not ideal to modify Kubernetes resource requests in a YAML file manually.

Kubernetes Vertical Pod Autoscaler (VPA) Automagically:

- Adapts resource requests
- Limits to match the workload

Bonus Tip

Fairwind's Goldilocks is a tool that creates a VPA for each workload in a namespace and provides recommendations in a dashboard:



Namespace Details

NAMESPACE

goldilocks ^

View only this namespace

DEPLOYMENT

goldilocks-controller ^

CONTAINER

goldilocks ^

Guaranteed QoS

	Current		Guaranteed
CPU Request	50m	>	49m
CPU Limit	50m	>	49m
Memory Request	64Mi	>	64M
Memory Limit	64Mi	>	64M

▸ Recommended Settings

Burstable QoS

	Current		Burstable
CPU Request	50m	=	15m
CPU Limit	50m	=	54m
Memory Request	64Mi	=	53M
Memory Limit	64Mi	=	71M

▸ Recommended Settings

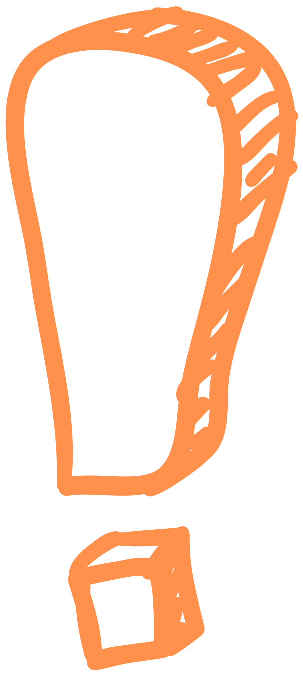
DEPLOYMENT

goldilocks-dashboard ⌄

There are two truths about cloud costs:

1. To save, you have to optimize Visible Costs.

2. To save maximum, you have to optimize Invisible Costs too



To get the best out of cloud, one should have the ability to mitigate both "visible and invisible costs"

TL;DR (tools links added in the post)

1. Slack cost = Requested resources - Real Usage

2. Kubernetes Resource Report helps to identify Slacks

3. Leverage 'Grafhana' to set the right resource requests

4. Adapt Vertical Pod Auto Scaler to automate

5. Use Goldilocks as a utilization recommendation tool

Reshare this Post

It's the best thing you can do to help others on LinkedIn

+ FOLLOW

for more valuable content