KubeCon | CloudNativeCon

Europe 2022

WELCOME TO VALENCIA

KubeCon CloudNativeCon Europe 2022

Supporting Long-Lived Pods Using a Simple Kubernetes Webhook

Clement Labbe, Slack

slack

Supporting Long-Lived Pods

Clem (he/him)
Software Engineer, *Slack*

KubeCon CloudNativeCon Europe 2022
PromCon North America 2021

Agenda

- Intro
- The Problem
  - Long Lived Pods
  - Nodes Get Killed
- Solution
  - Minimum Pod Lifespan
  - Node Taint Service
  - Teaching the Killers
  - Service Config to Pod Tolerations
  - Simple Admission Webhook
- Last Few Words
  - Min Lifespan not Max Lifespan
  - We Already Had An Admission Webhook

slack

Intro | Cloud Compute

16M Peak Concurrent Users

45k EC2 Instances

162 K8s Clusters

316 Services Deployed to K8s

1k Engineers

235 Chef Roles

slack

Agenda

Intro

The Problem

- Long Lived Pods
- Nodes Get Killed

Solution

- Minimum Pod Lifespan
- Node Taint Service
- Teaching the Killers
- Service Config to Pod Tolerations
- Simple Admission Webhook

Last Few Words

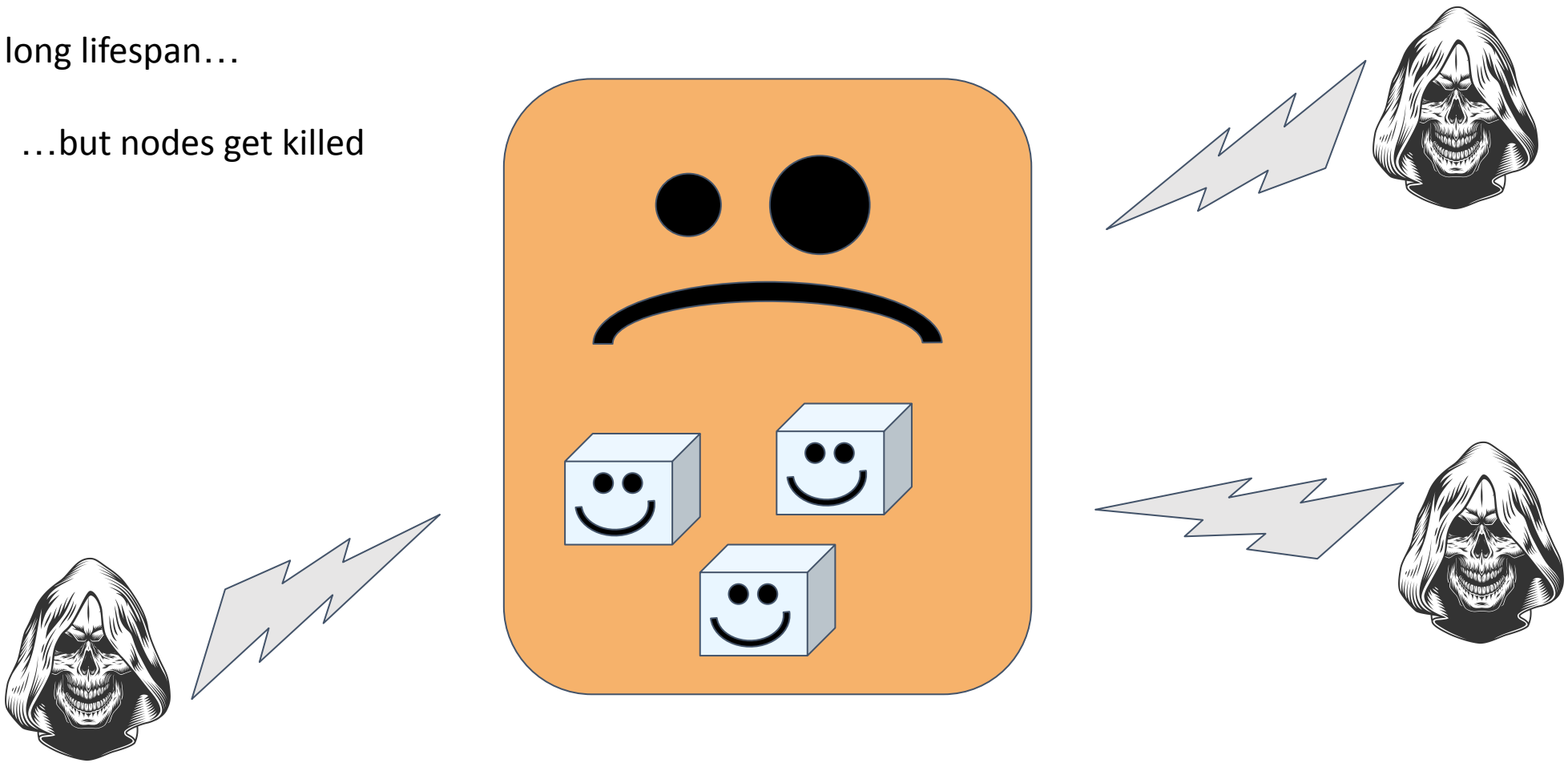
- Min Lifespan not Max Lifespan
- We Already Had An Admission Webhook

slack

The Problem

Some pods want a long lifespan…

…but nodes get killed



The Problem | Long Lived Pods

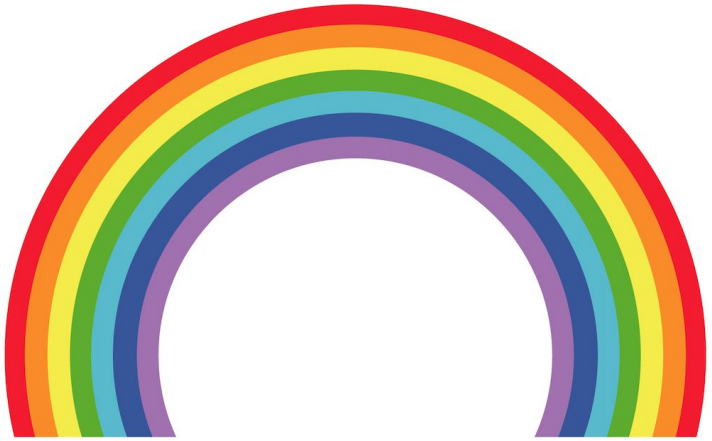
∞ scale-out

fun

graceful shutdown

insta-boot

stateless



Ideal World

VS

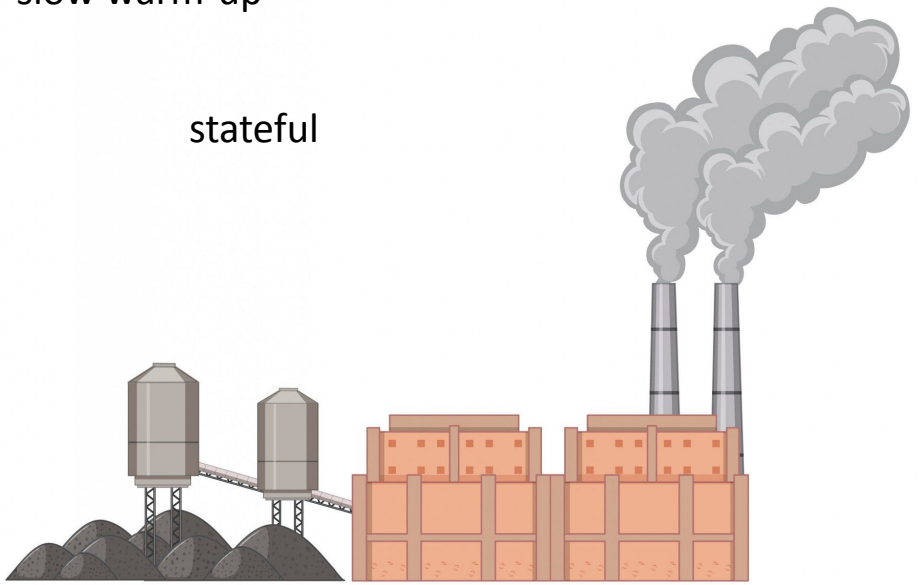
long-lived websockets

termination-phobic

intensive scale-out

slow warm-up

stateful



Real World

slack

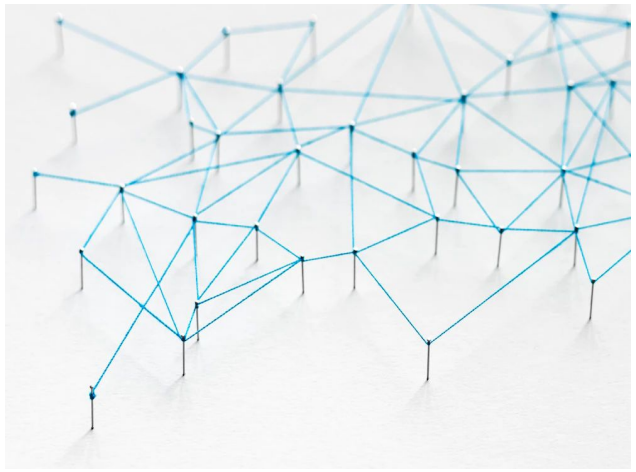
…we are left with a long tail of unruly ducks

The Problem | Long Lived Pods

Some examples…

Batch Jobs

- Re-start work from scratch when killed mid run



Distributed Caches

- New replicas are slow to warm up
- Pulling data from existing nodes is expensive

Jenkins Controller

- Singleton
- Cannot get killed when engineers need it

slack

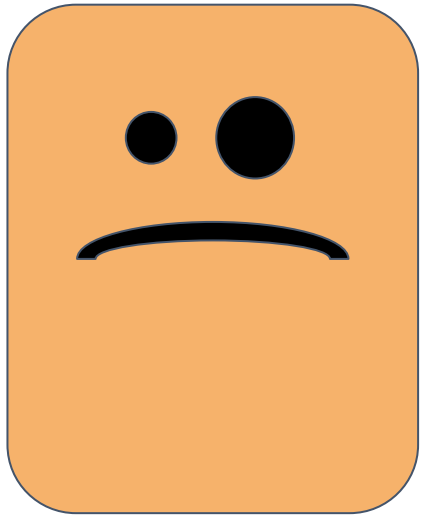
The Problem | Nodes Get Killed

We control:

ASG scale-in

Chaos Engineering

2 weeks anniversary stab in the back



We don't control:

Emergency patch rollout

AWS terminations

slack

Agenda

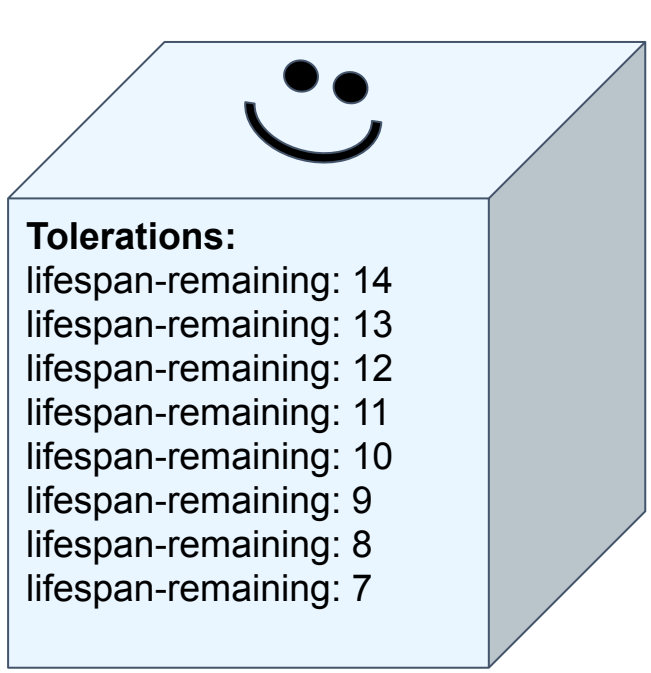
- Intro
- The Problem
  - Long Lived Pods
  - Nodes Get Killed
- **The Solution**
  - Minimum Pod Lifespan
  - Node Taint Service
  - Teaching the Killers
  - Service Config to Pod Tolerations
  - Simple Admission Webhook
- Last Few Words
  - Min Lifespan not Max Lifespan
  - We Already Had An Admission Webhook

slack

The Solution | Minimum Pod Lifespan

Matching Pods and Nodes with Tolerations and Taints:

- Nodes live for 14 days



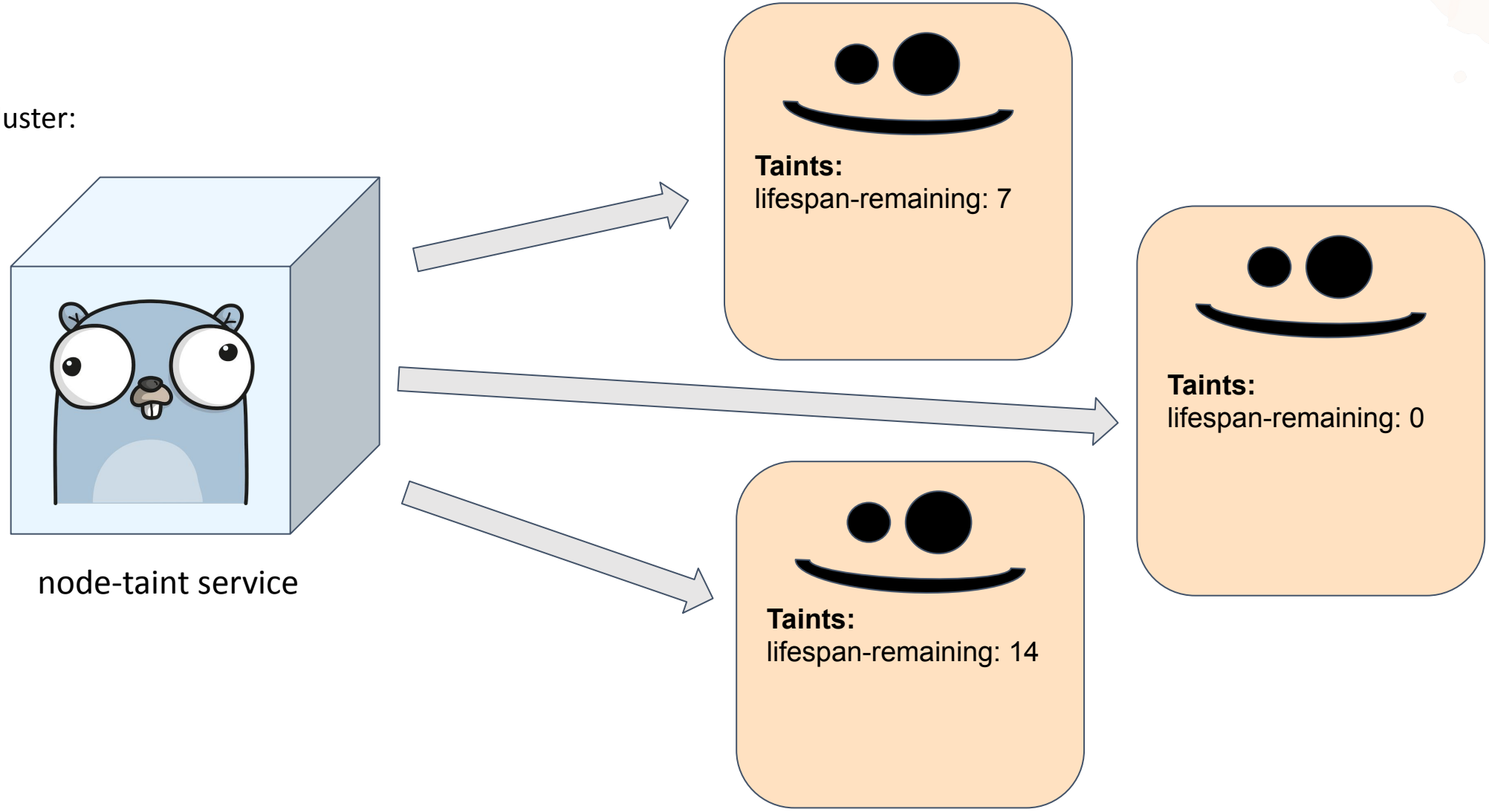
Pod wants a 7 days lifespan

Node is 4 days old:
14 - 4 = 10

slack

Node Taint Service

In each k8s cluster:



slack

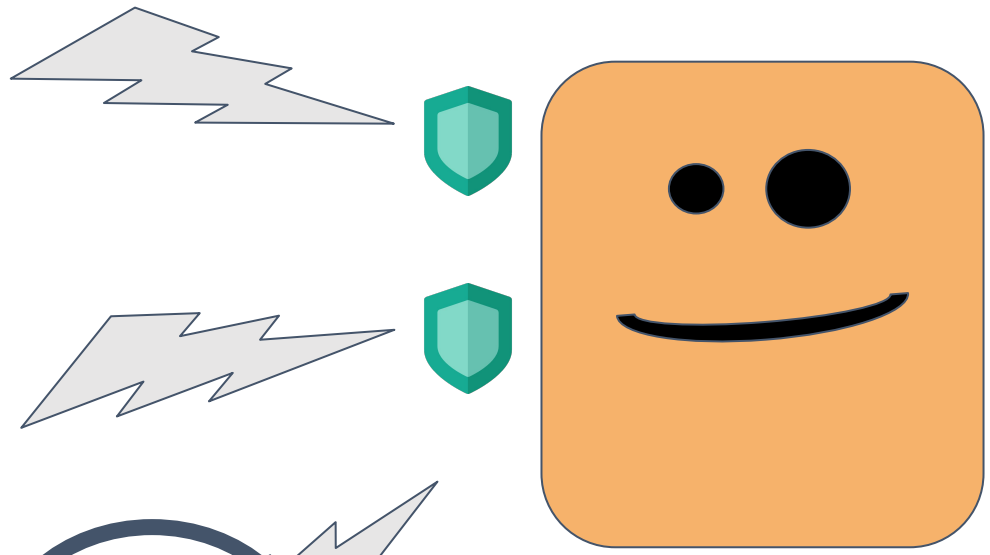
Teaching the Killers

We control:

ASG scale-in
(cluster-autoscaler)

Chaos Engineering
(kube-test-cluster)

2 weeks
anniversary stab
in the back



We don’t control:

Emergency patch
rollout

AWS terminations

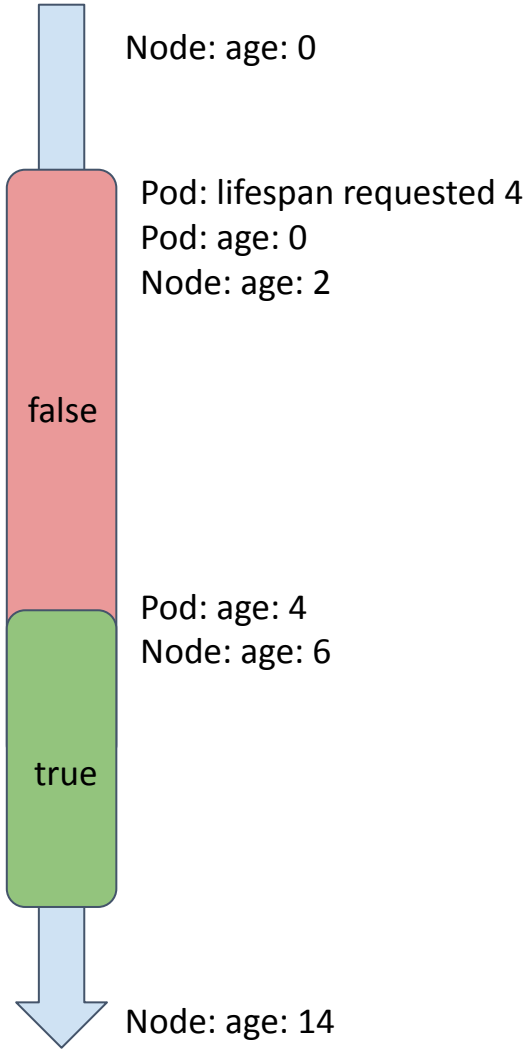
slack

ASG Scale-In | Cluster Autoscaler

kubernetes/autoscaler

```yaml
---
apiVersion: apps/v1
kind: Deployment
metadata:
 name: cluster-autoscaler
spec:
 template:
   command:
     - ./cluster-autoscaler
     - --v=4
     - --stderrthreshold=info
     - --cloud-provider=aws
     - --balance-similar-node-groups
     - --expendable-pods-priority-cutoff=-10
     - --scale-down-delay-after-add=7200s
     - --skip-nodes-with-local-storage=false
     - --ignore-taint=slack.com/lifespan-remaining
     - --max-total-unready-percentage=100
```

Chaos Engineering | kube-test-cluster




```go
// isLifespanEligible returns a boolean indicating whether the given node can
// be terminated without violating lifespan contracts. In other words, it
// looks at all pods scheduled on the node in question, and returns false if
// any of those pods are younger than their requested lifespan.
//
func isLifespanEligible(log logrus.FieldLogger, client *bedrockClient.Client,
  cluster string, node *corev1.Node) (bool, error) {

 pods, err := client.ListPodsWithLabels(cluster, "", nil)
 if err != nil {
   return false, fmt.Errorf("listing pods: %w", err)
 }

 for _, pod := range pods {
   lifespanStr := pod.Labels[bedrock.LifespanRequestLabel]
   lifespan, _ := strconv.Atoi(lifespanStr)

   podAge := time.Since(pod.CreationTimestamp.Time)
   if podAge < 24*time.Hour*time.Duration(lifespan) {
     log.Debugf("termination would violate min lifespan %d for pod %d days old",
       lifespan, int(podAge.Minutes()/60/24))
     return false, nil
   }
 }

 return true, nil
}
```

Agenda

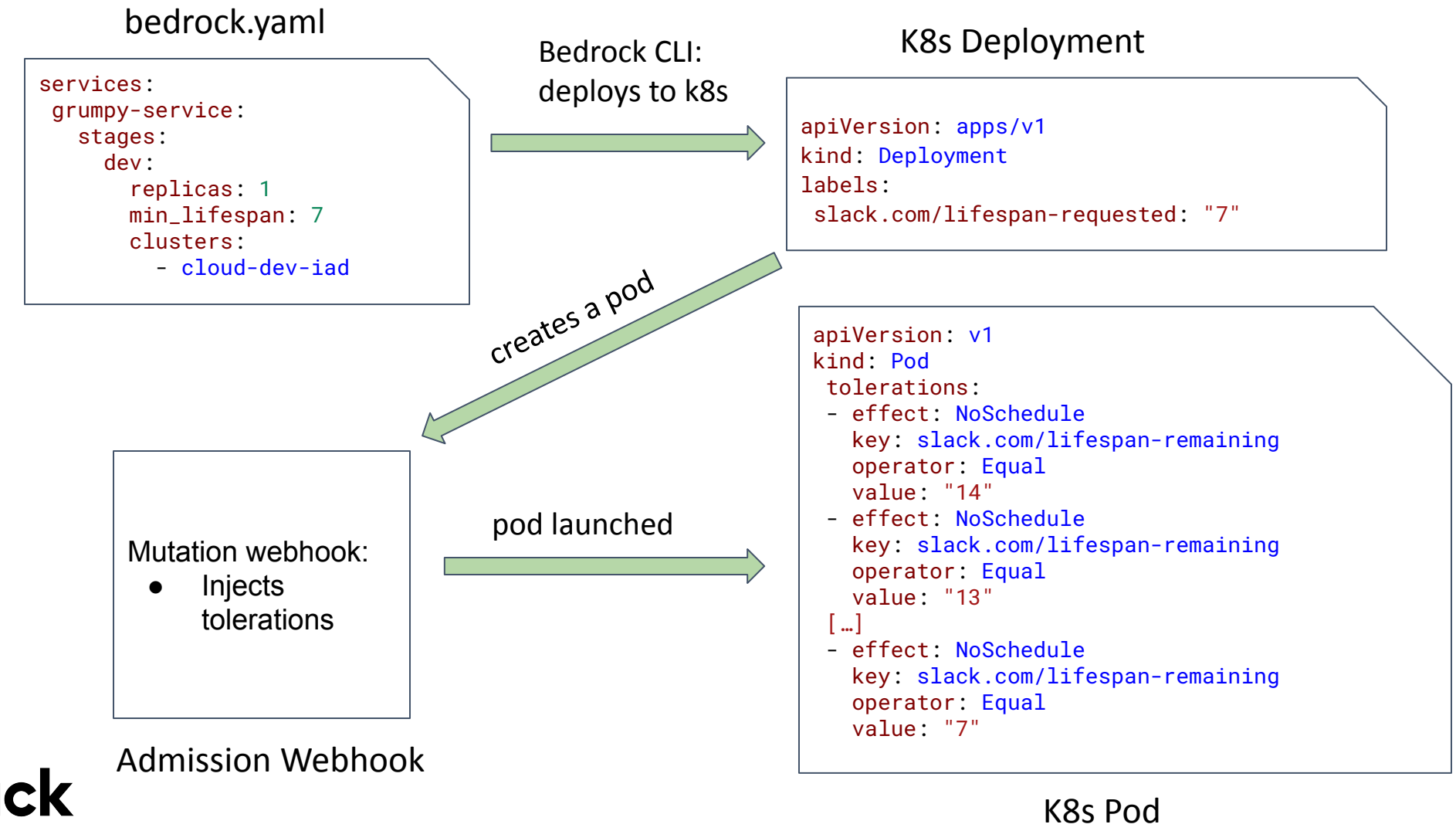
Intro
The Problem
- Long Lived Pods
- Nodes Get Killed

The Solution
- Minimum Pod Lifespan
- Node Taint Service
- Teaching the Killers
- **Service Config to Pod Tolerations**
- Simple Admission Webhook

Last Few Words
- Min Lifespan not Max Lifespan
- We Already Had An Admission Webhook

slack

Service Config to Pod Tolerations

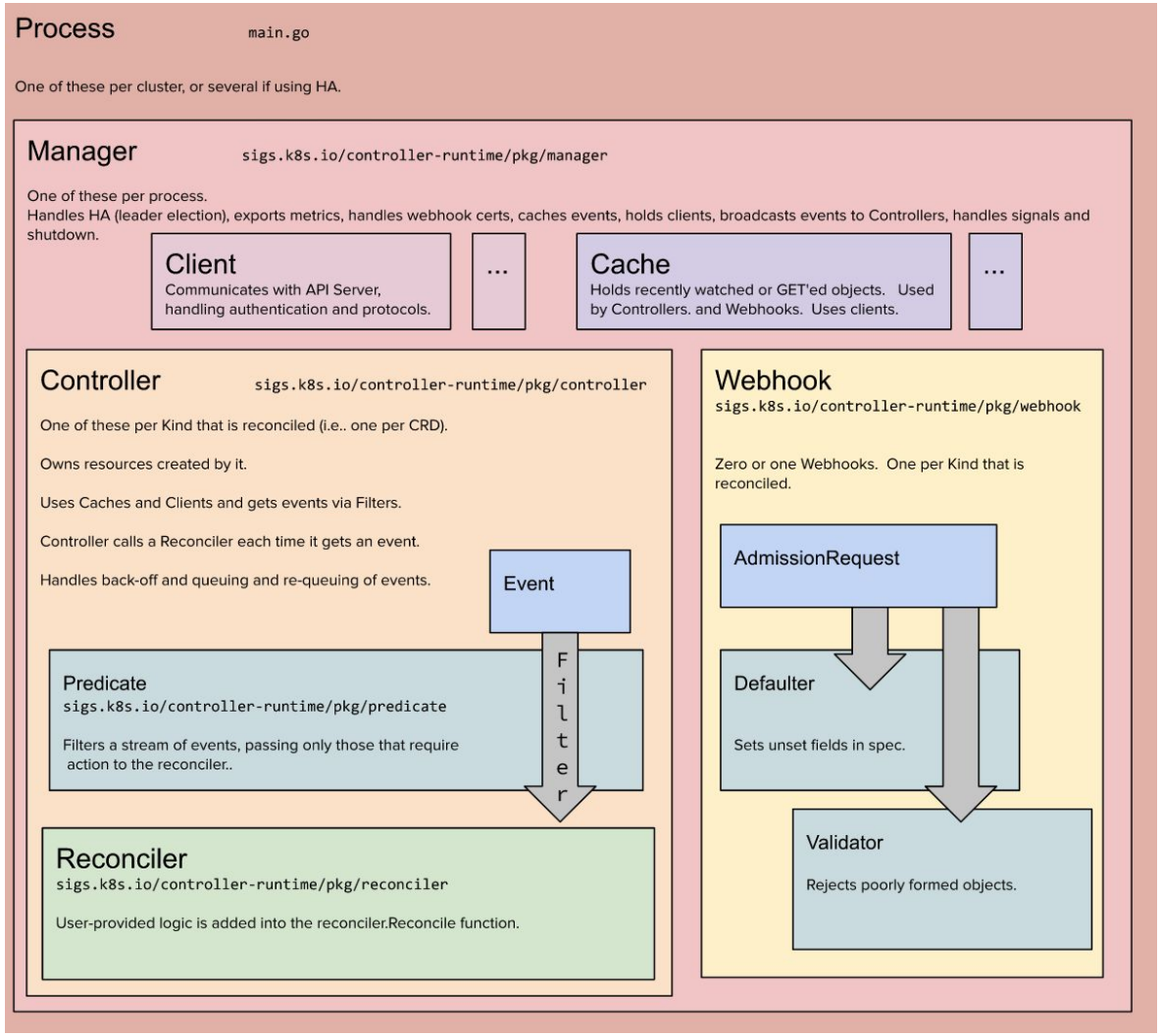


slack

Simple Admission Webhook

kubebuilder

Kubebuilder is a framework for building Kubernetes APIs using custom resource definitions (CRDs).

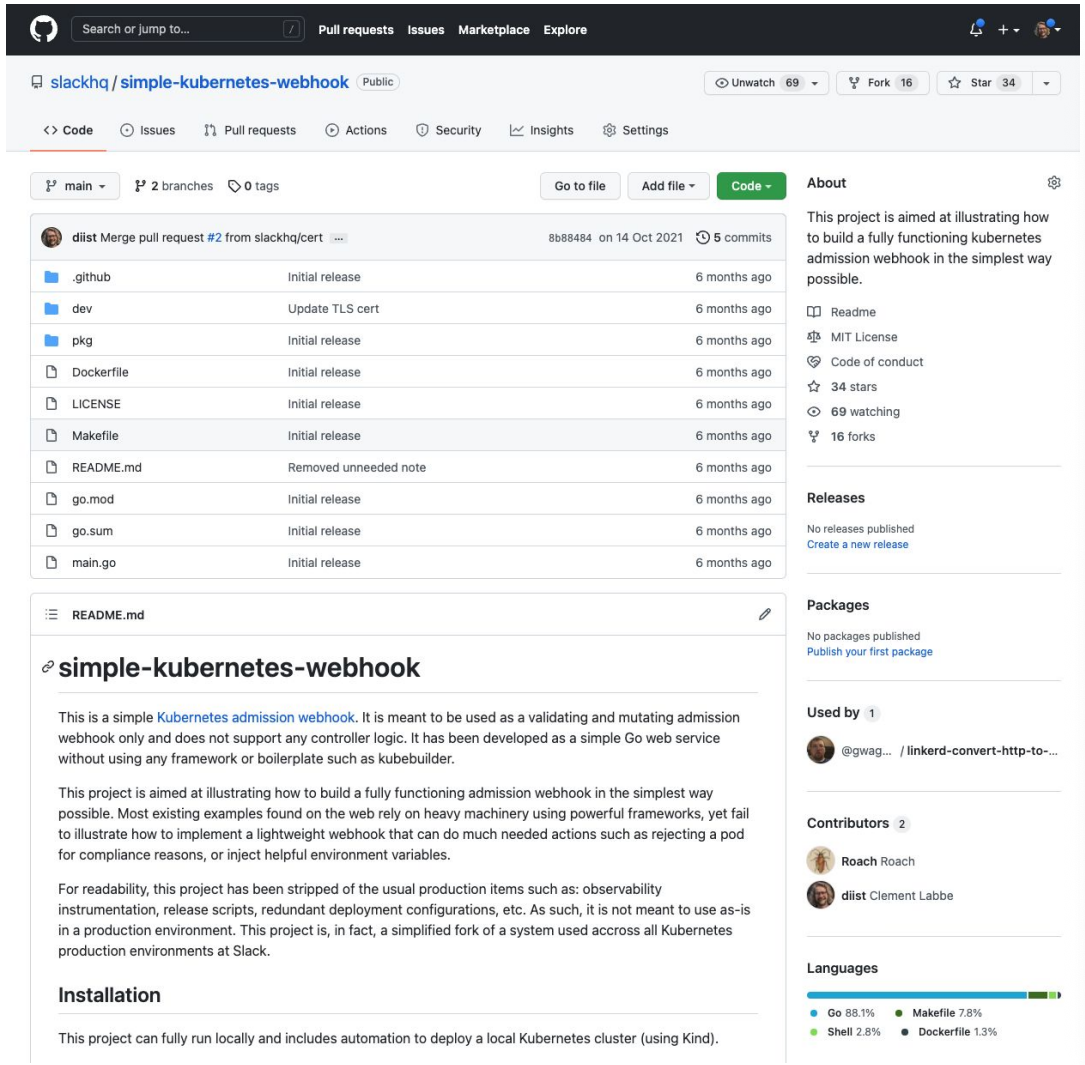


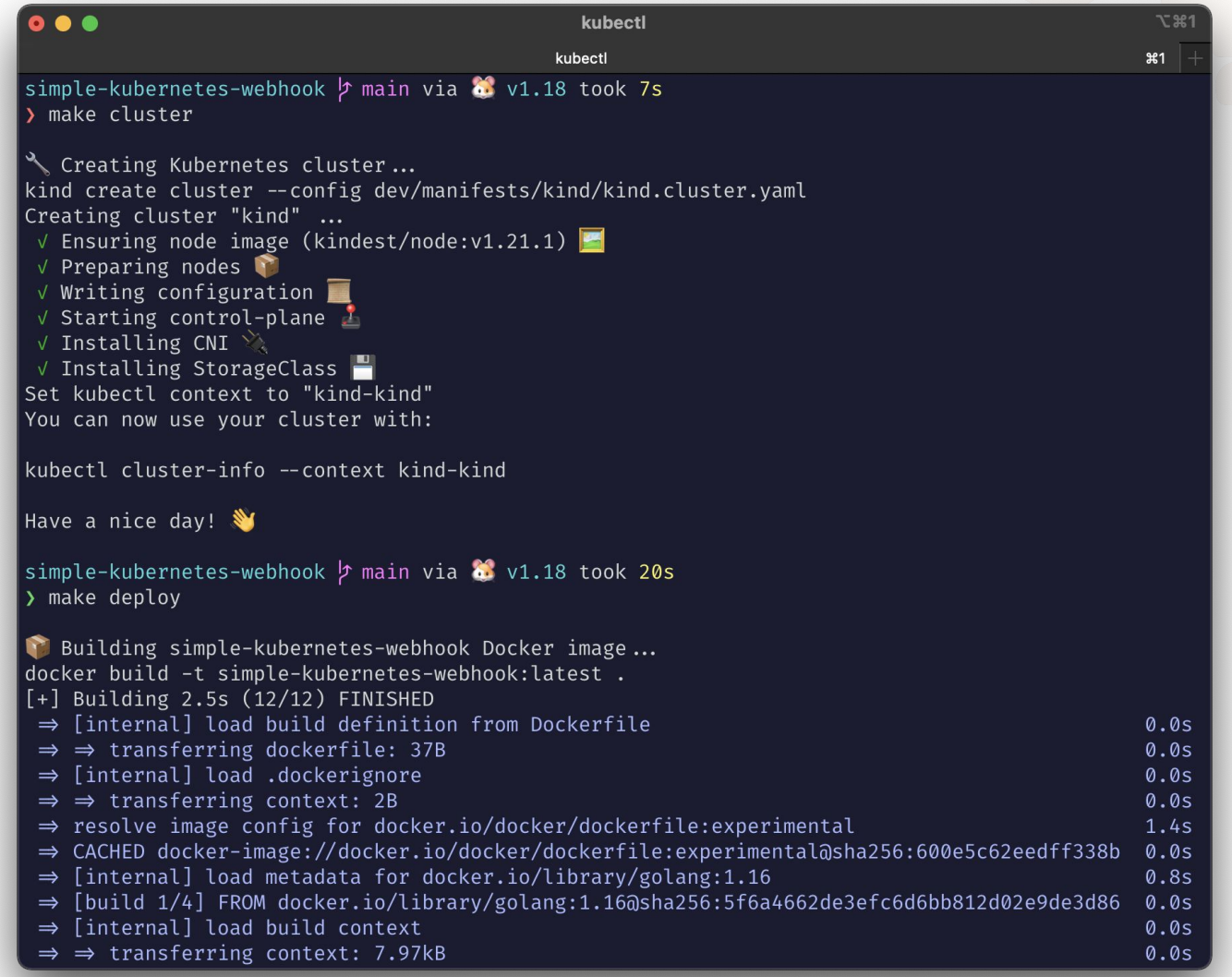
Architecture Concept Diagram

Simple Admission Webhook

KubeCon CloudNativeCon Europe 2022

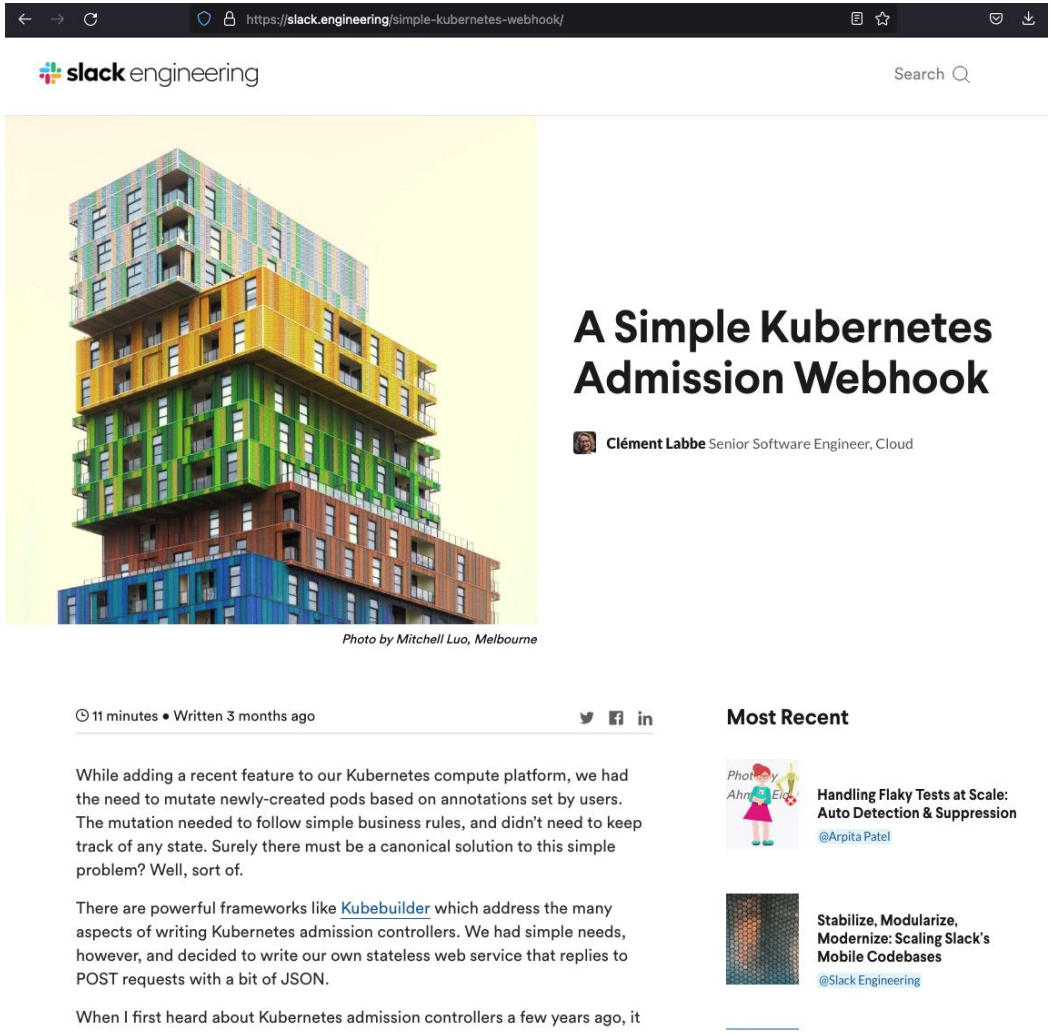
https://github.com/slackhq/simple-kubernetes-webhook





Simple Admission Webhook

https://slack.engineering/simple-kubernetes-webhook/

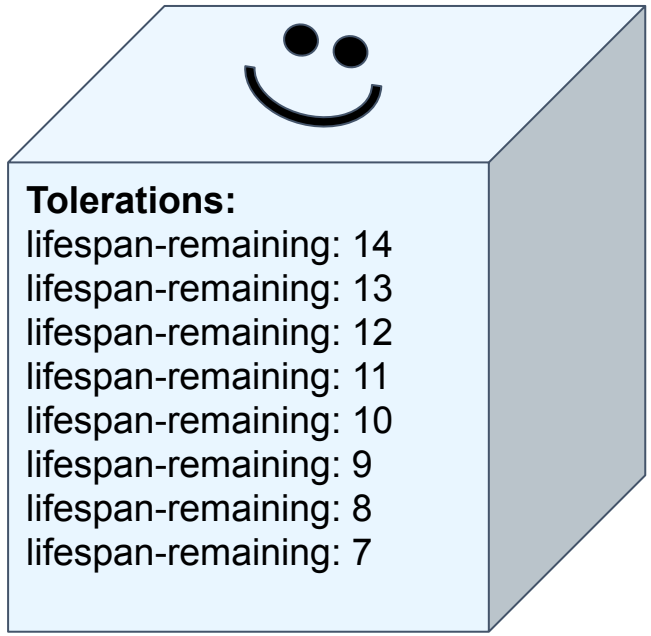


slack

The Solution | Minimum Pod Lifespan

Min Lifespan feature at Slack:

- Used by half a dozen services
- Everyone set max value of 7 days



Pod wants a 7 days lifespan

Node is 4 days old:
14 - 4 = 10

slack

Agenda

- Intro
- The Problem
  - Long Lived Pods
  - Nodes Get Killed
- The Solution
  - Minimum Pod Lifespan
  - Node Taint Service
  - Teaching the Killers
  - Service Config to Pod Tolerations
  - Simple Admission Webhook
- **Last Few Words**
  - Min Lifespan not Max Lifespan
  - We Already Had An Admission Webhook

slack

Min Lifespan not Max Lifespan

Min lifespan guarantee…

- But no guarantee that the pod will get killed at the expiry date
- Users want control over lifespan dynamics
- eg. Jenkins controller

We Already Had An Admission Webhook

Existing Admission Webhook:

- Using an ancient Kubebuilder version
- Breaking major version upgrade would require a lot of rework
- Mutates and Validates in-flight (no controllers needed)
- Decided to write the “Simple Webhook”, then migrate old to new
- Migration still ongoing

Thanks

Sean Waller

Tricia Bogen

Javier Turegano

Credits

Images:

- Human skull vector created by dgim-studio - www.freepik.com
- Rainbow cartoon vector created by brgfx - www.freepik.com
- Pollution vector created by brgfx - www.freepik.com
- Shield icons created by Freepik - Flaticon
- Speech icon vector created by zirconicusso - www.freepik.com
- Village house vector created by brgfx - www.freepik.com
- Zombie heads created by brgfx - www.freepik.com
- Appointment booking vector created by freepik - www.freepik.com

Thank you!
…Questions?

We’re Hiring!

slack