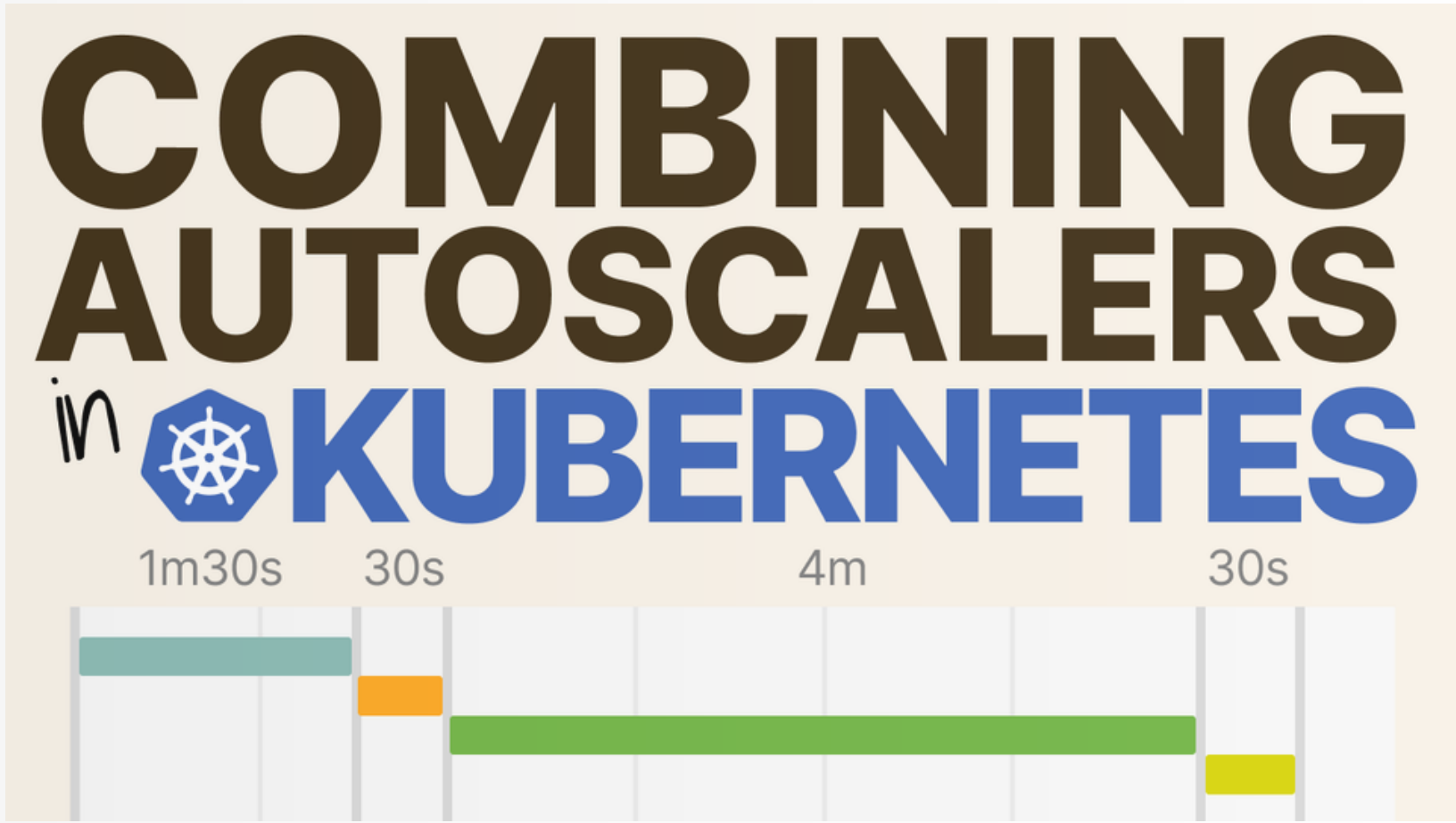
Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

When designing a Kubernetes cluster, you might need to answer questions such as:

How long does it take for the cluster to scale?

How long do I have to wait before a new Pod is created?

Let's explore...



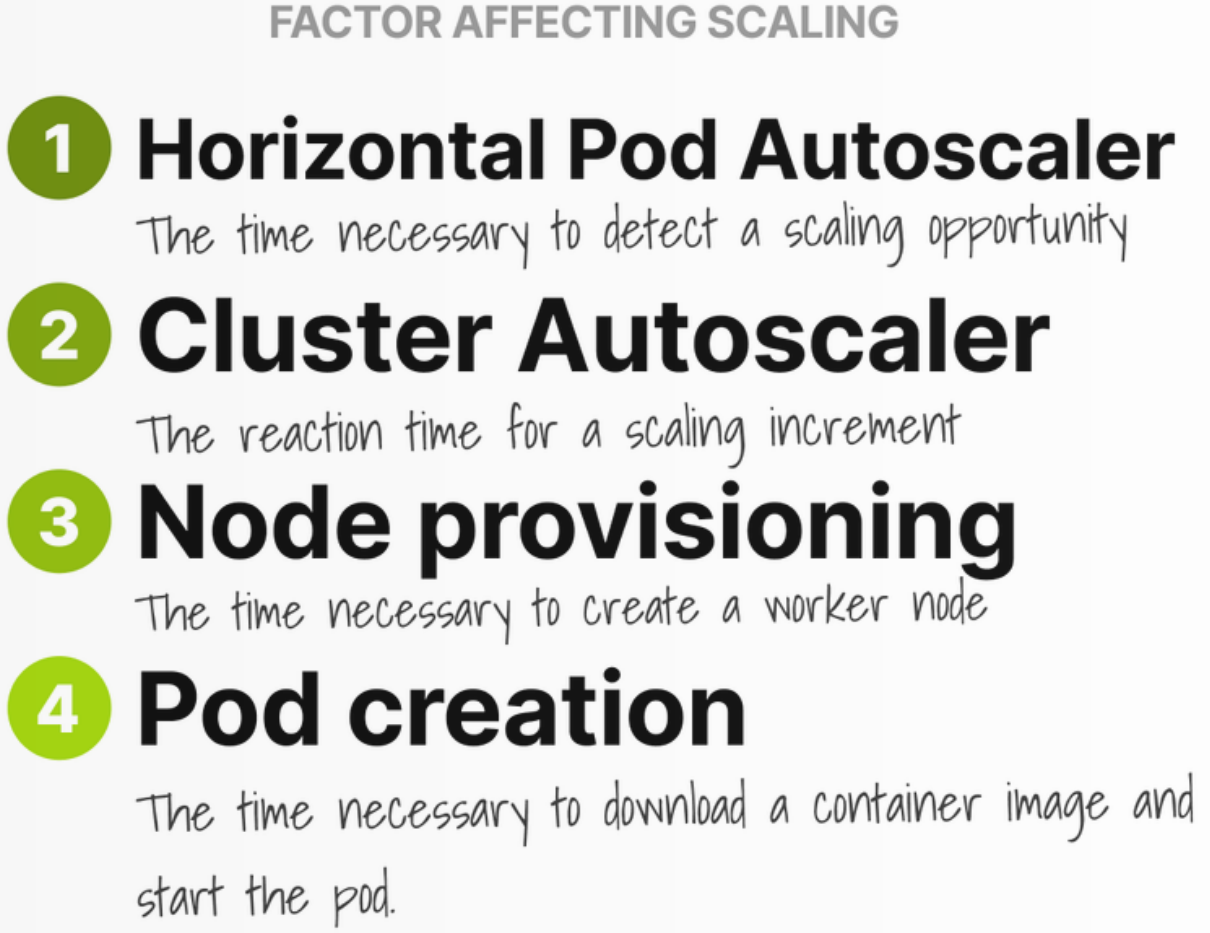
9:00 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

1/

Four significant factors affect scaling:

① Horizontal Pod Autoscaler reaction time
② Cluster Autoscaler reaction time
③ Node provisioning time
④ Pod creation time



9:00 PM · Sep 19, 2023

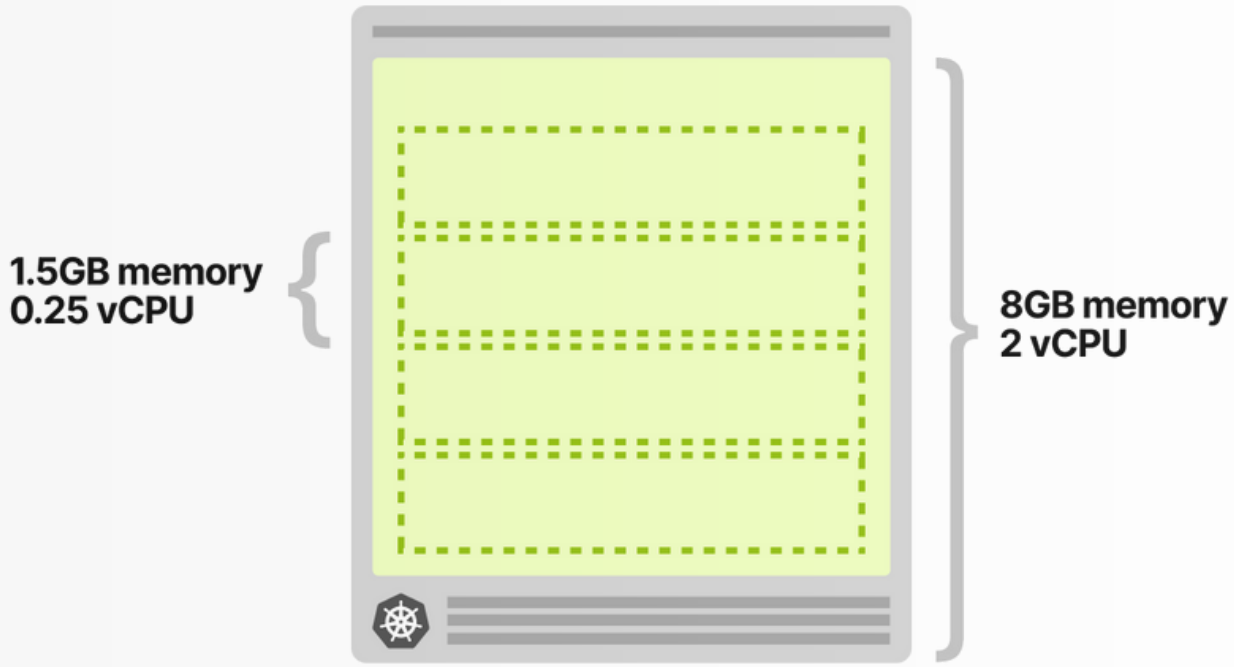
Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

2/

Let's have a look at an example

Imagine having an application that requires and uses 1.5GB of memory and 0.25 vCPU at all times

You provisioned a cluster with a single node of 8GB and 2 vCPU, and 4 replicas; what happens next?



9:01 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

3/

The pod is pending because you still need to consider reserved resources

Not all available memory and CPU are usable by pods

9:02 PM · Sep 19, 2023

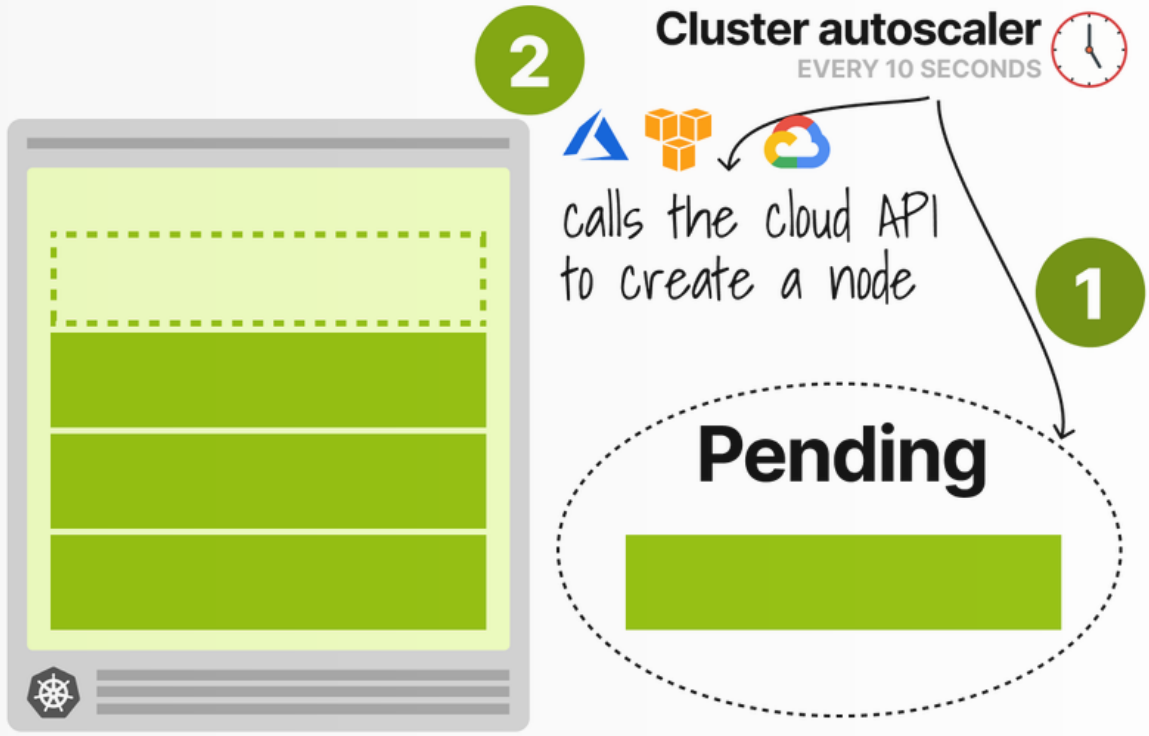
Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

4/

But why does the cluster autoscaler wait for the Pod to be pending before it triggers creating a node?

The CA doesn't measure CPU and memory

Instead, it checks for unschedulable Pods every 10 seconds; if there's at least one, it creates a node



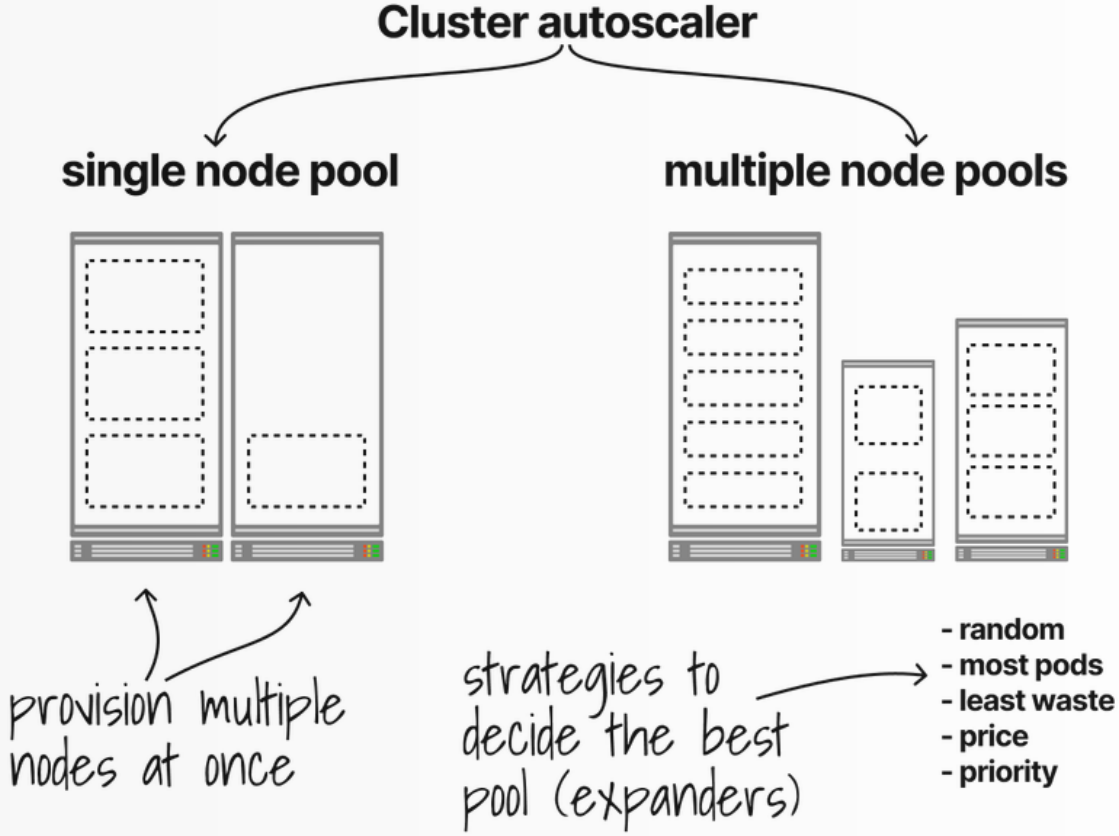
9:03 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

5/

It can create more than one node simultaneously if you have several pods

If you have several node pools, you can choose a strategy (expander) to decide which pool is selected

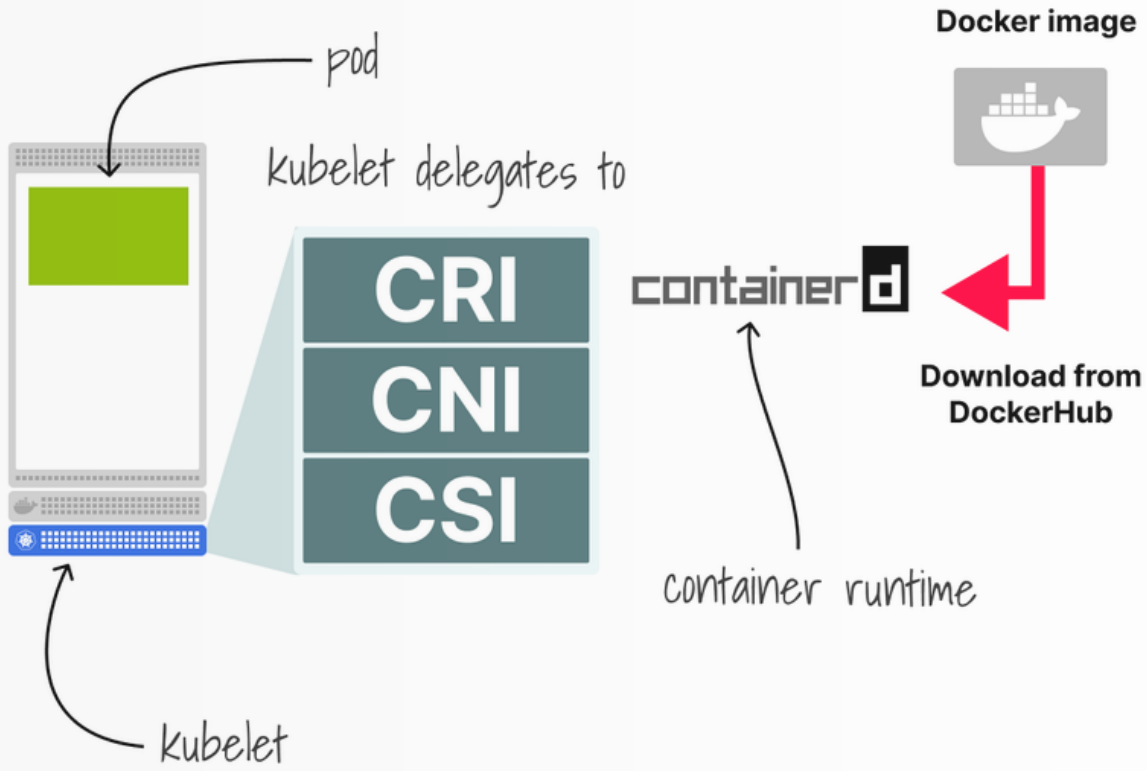


9:04 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

6/

As a final step, the node is provisioned, the container image is pulled from the registry and the pod is launched



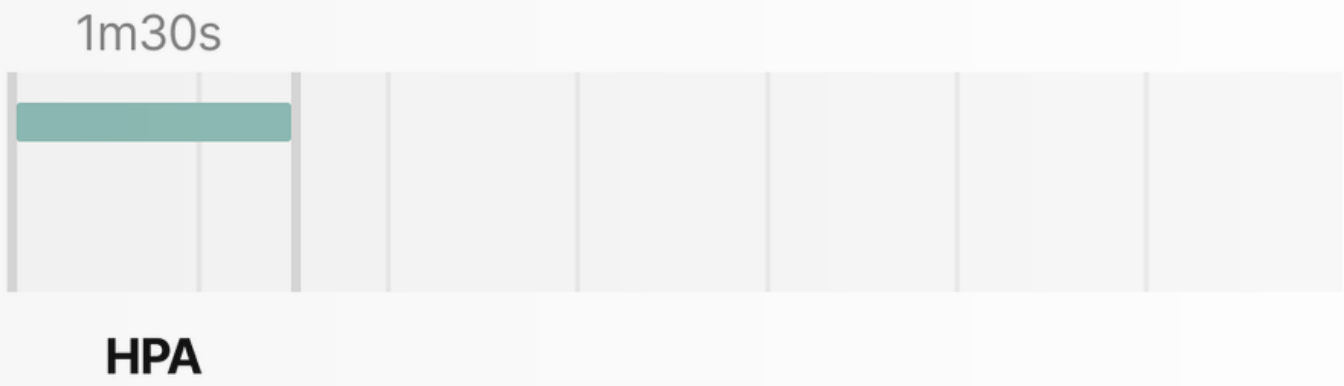
9:04 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

7/

Now that you're familiar with the process let's look at how this translates to timings

The Horizontal Pod Autoscaler might take up to 1m30s to increase the number of replicas

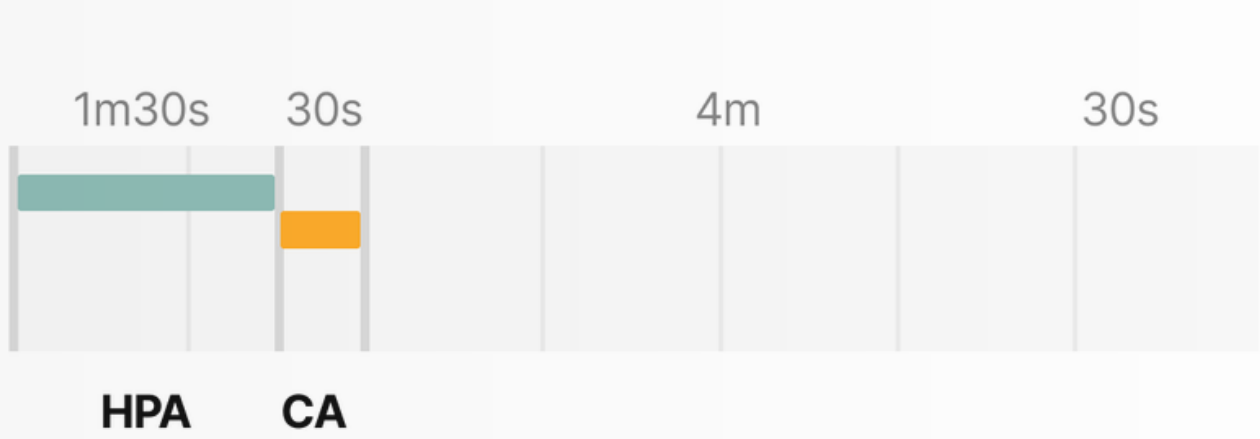


9:04 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

8/

The Cluster Autoscaler should take less than 30 seconds for a cluster with less than 100 nodes and less than a minute for a cluster with more than 100 nodes

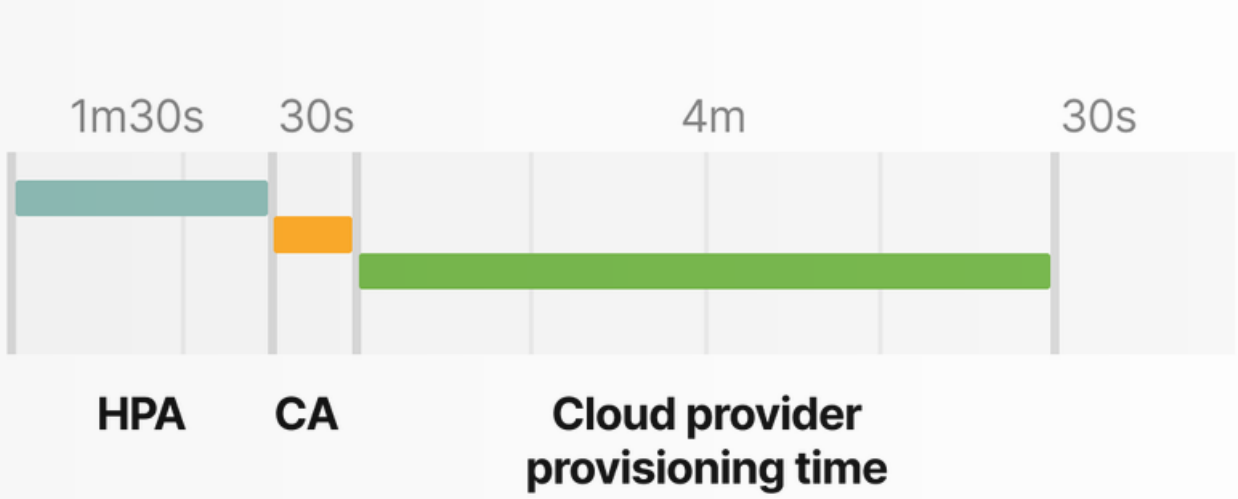


9:04 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

9/

The cloud provider might take 3 to 5 minutes to create the computer resource

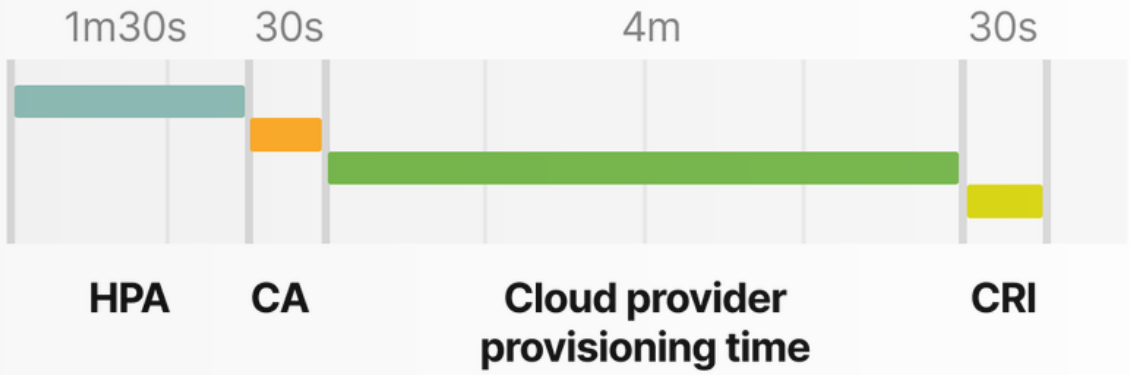


9:05 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

10/

The container runtime could take up to 30 seconds to download the container image



9:05 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

11/

In the worst case, with a small cluster, it takes 6m30s from when a pod is pending to launch it in a new node

If the cluster has more than 100 nodes, it's up to 7m

HPA delay	**1m30s**
CA delay	**30s**
Cloud provider delay	**4m**
CRI	**30s**
Total	**6m30s**

worst case lead time for scaling pods with a cluster <100 nodes

HPA delay	**1m30s**
CA delay	**1m**
Cloud provider delay	**4m**
CRI	**30s**
Total	**7m**

worst case lead time for scaling pods with a cluster 100+ nodes

9:05 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

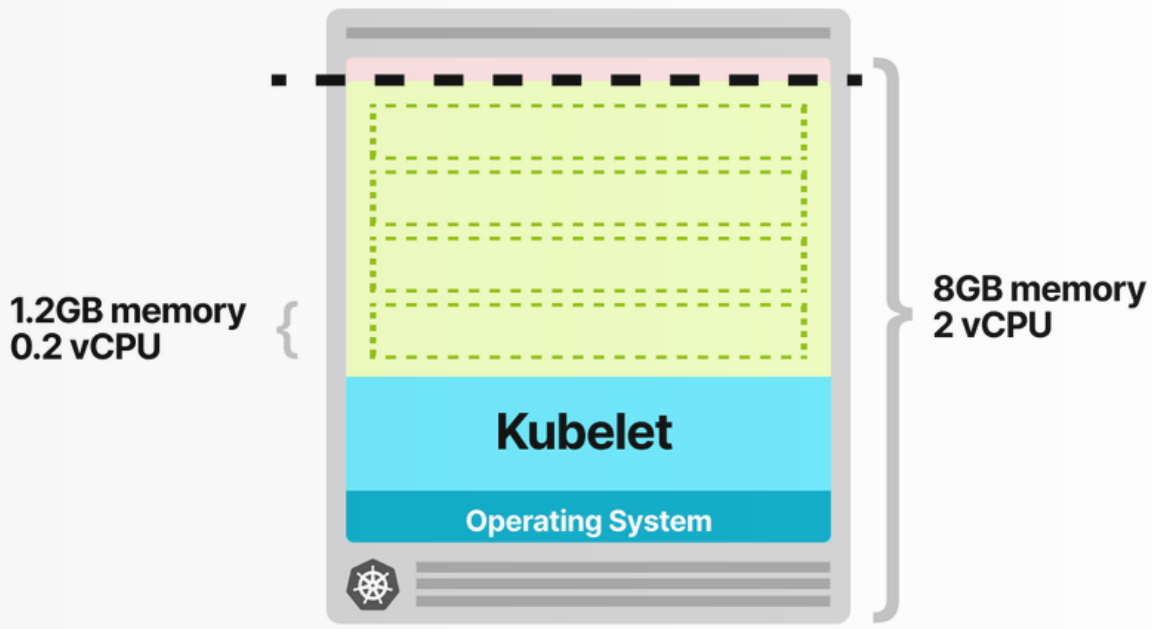
12/

How can you fix this?

If you can, don't scale

Select a node instance type that is optimized for your workload

You can use this calculator to help you with the task:
learnk8s.io/kubernetes-ins...



9:05 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

13/

If your traffic follows predictable patterns, scale your fleet up and down with the CronHPA autoscaler github.com/AliyunContaine... or KEDA keda.sh

With those, you can control the number of replicas with a cron expression

```yaml
apiVersion: autoscaling.alibabacloud.com/v1beta1
kind: CronHorizontalPodAutoscaler
metadata:
  labels:
    controller-tools.k8s.io: "1.0"
  name: cronhpa-sample
spec:
  scaleTargetRef:
    apiVersion: apps/v1beta2
    kind: Deployment
    name: my-deployment
  jobs:
    - name: "scale-down"
      schedule: "@date 2023-09-28 23:59:00"
      targetSize: 1
    - name: "scale-up"
      schedule: "@date 2023-09-30 00:00:00"
      targetSize: 3
```

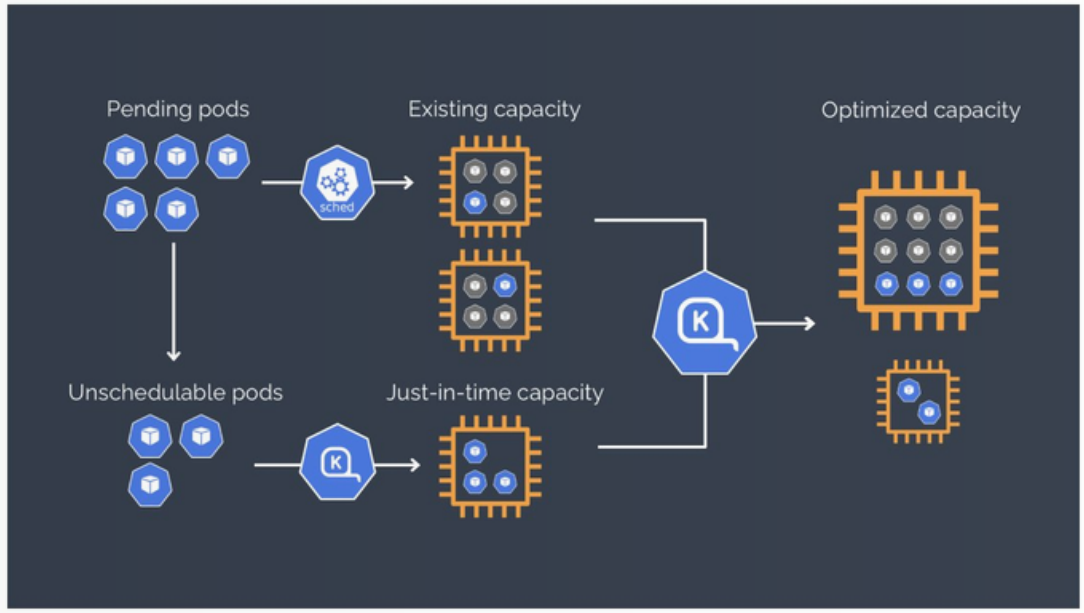
9:06 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

14/

If you are on AWS, use Karpenter github.com/aws/karpenter

Karpenter uses a more proactive model that can provision the correct instance and minimize provisioning lead time



9:06 PM · Sep 19, 2023

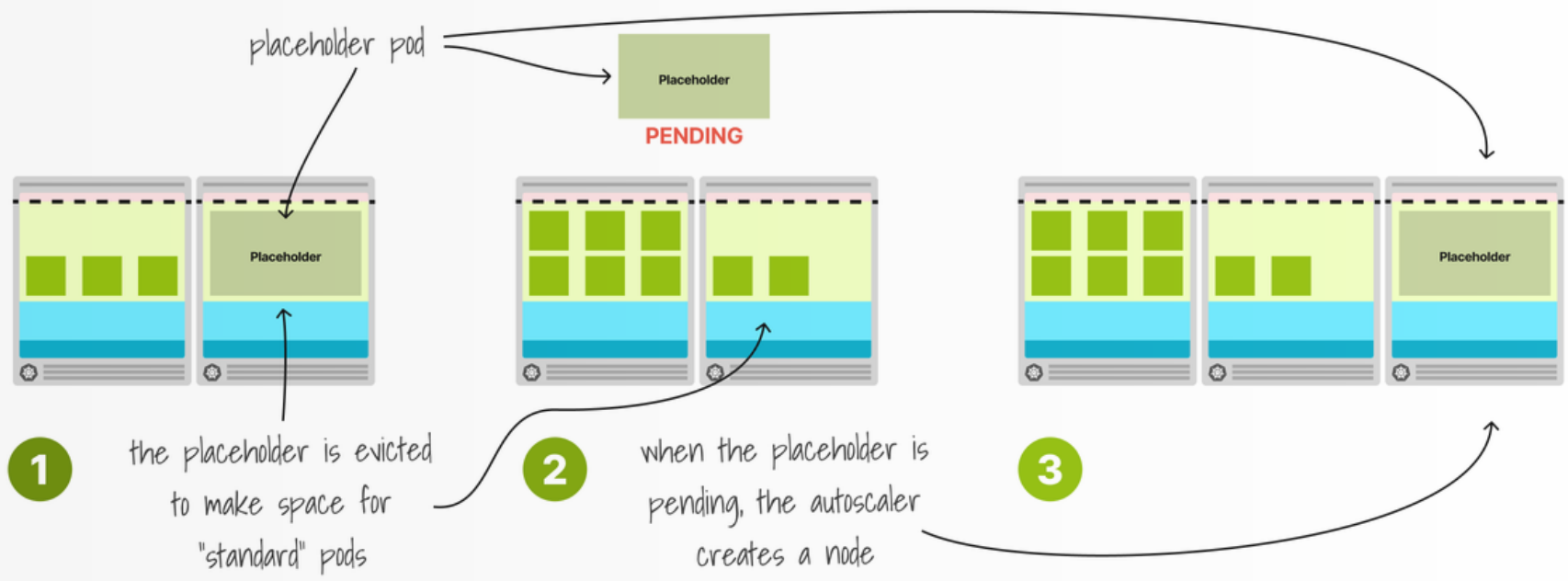
Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

15/

If everything else fails, you can use a placeholder pod always to keep a spare node in the cluster

The pod is configured with the lowest priority and evicted when another pod needs the space

github.com/kubernetes/aut...



9:06 PM · Sep 19, 2023

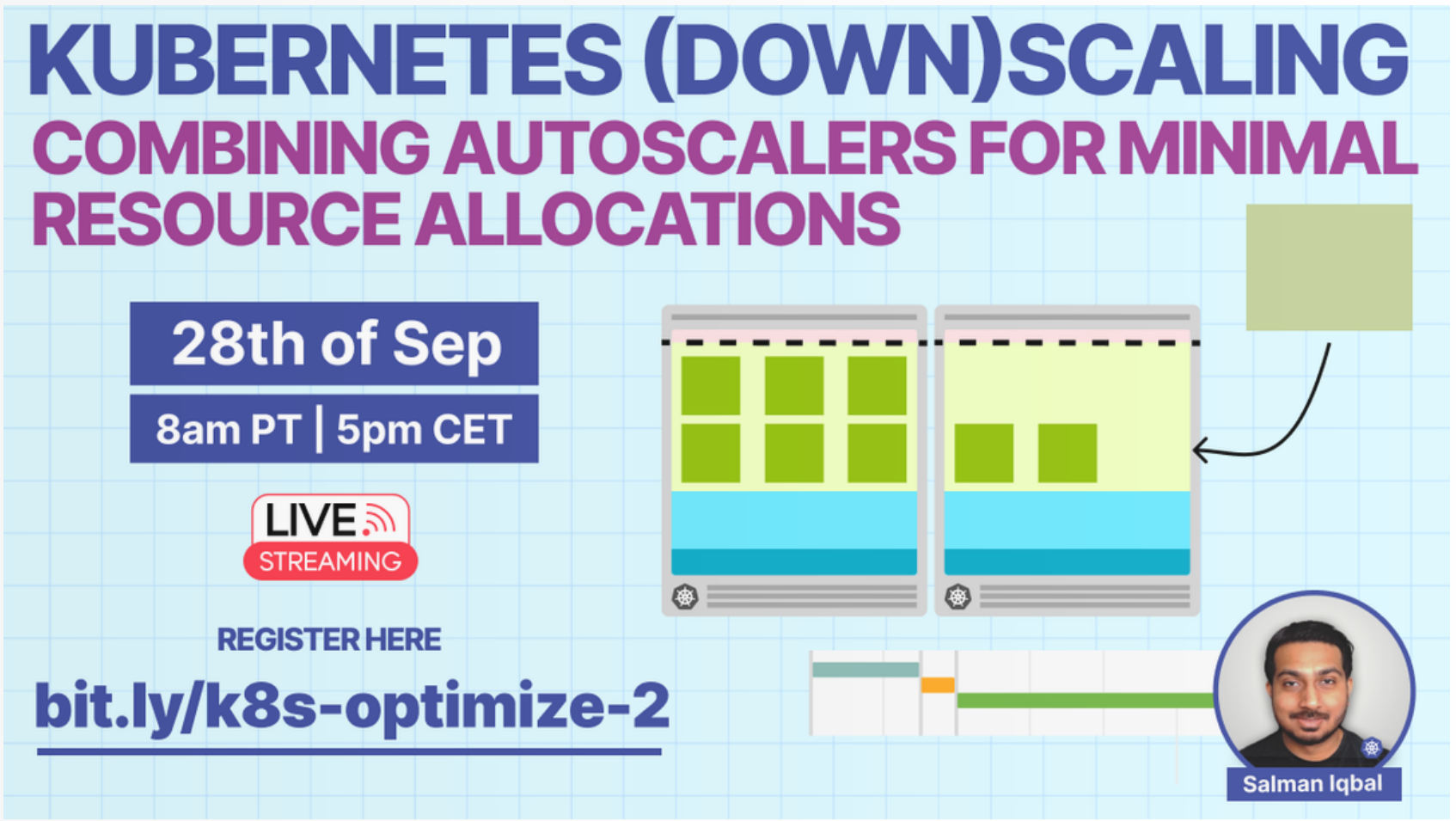
Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

16/

This thread is based on @SoulmanIqbal 's webinar on combining autoscalers

He will present it live next Thursday as the second episode of "Cost optimization and efficiency in Kubernetes" organized by @Akamai and @learnk8s

Sign up here (it's free): event.on24.com/eventRegistrat...



9:09 PM · Sep 19, 2023

Daniele Polencic — @danielepolencic@hachyderm.io
@danielepolencic

17/

And finally, if you've enjoyed this thread, you might also like:

- The Kubernetes workshops that we run at Learnk8s learnk8s.io/training
- This collection of past threads twitter.com/danielepolenci...
- The Kubernetes newsletter I publish every week learnk8s.io/learn-kubernet...

9:10 PM · Sep 19, 2023 · **646** Views