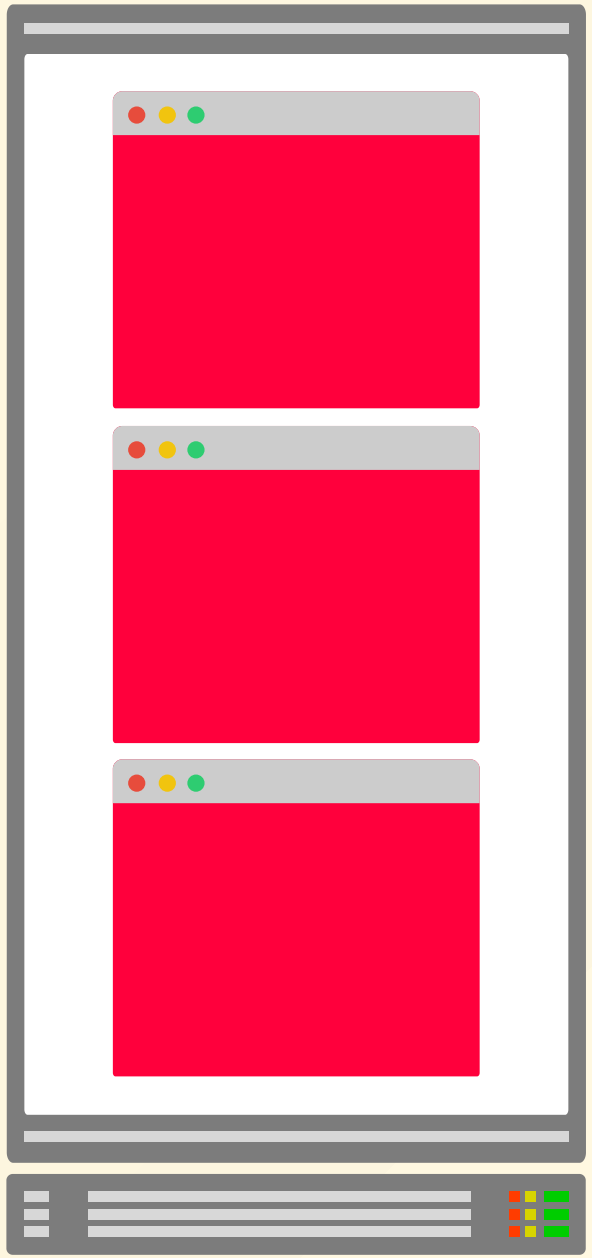
Proactive cluster autoscaling in Kubernetes

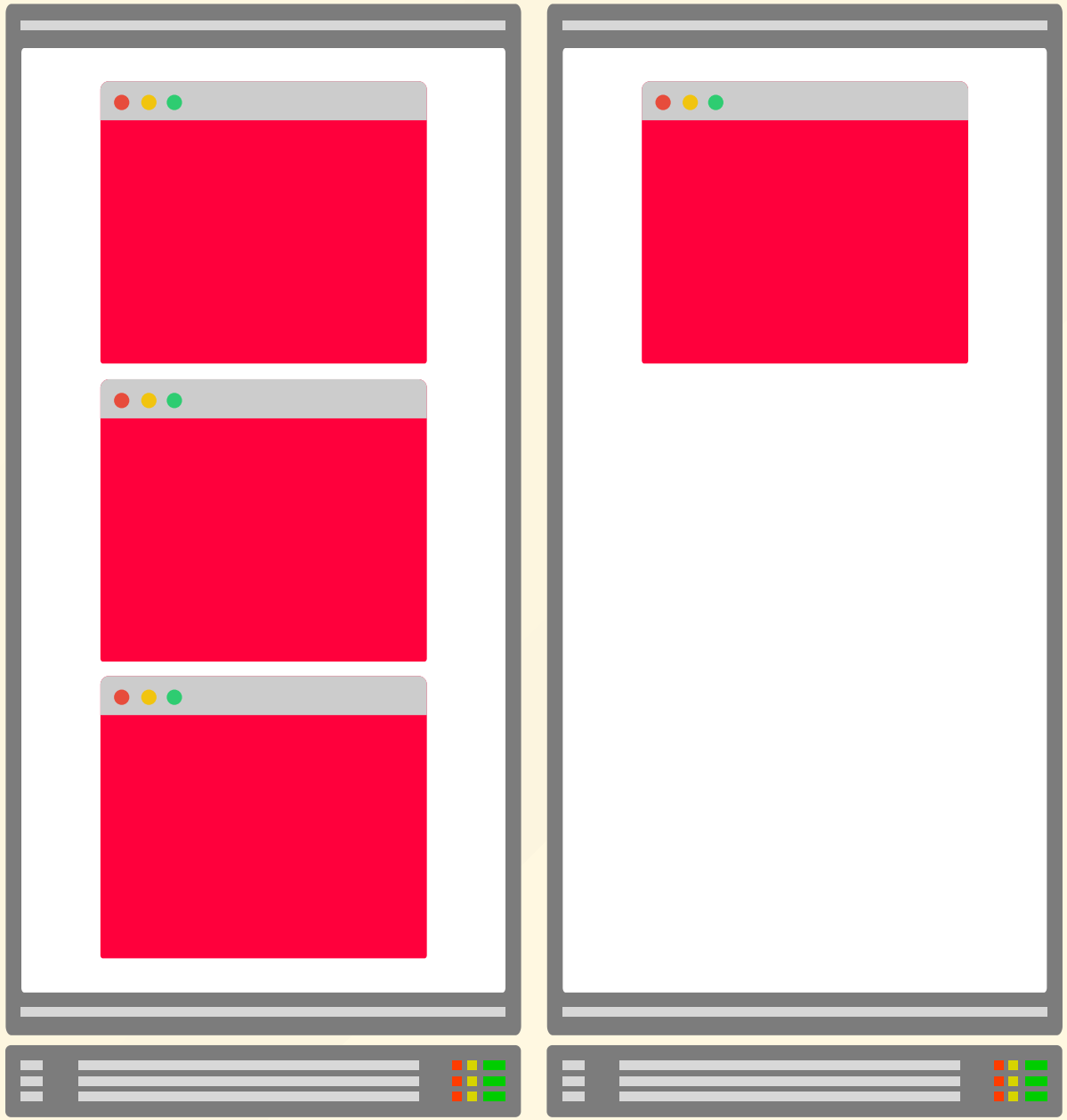
Chris Nesbitt-Smith

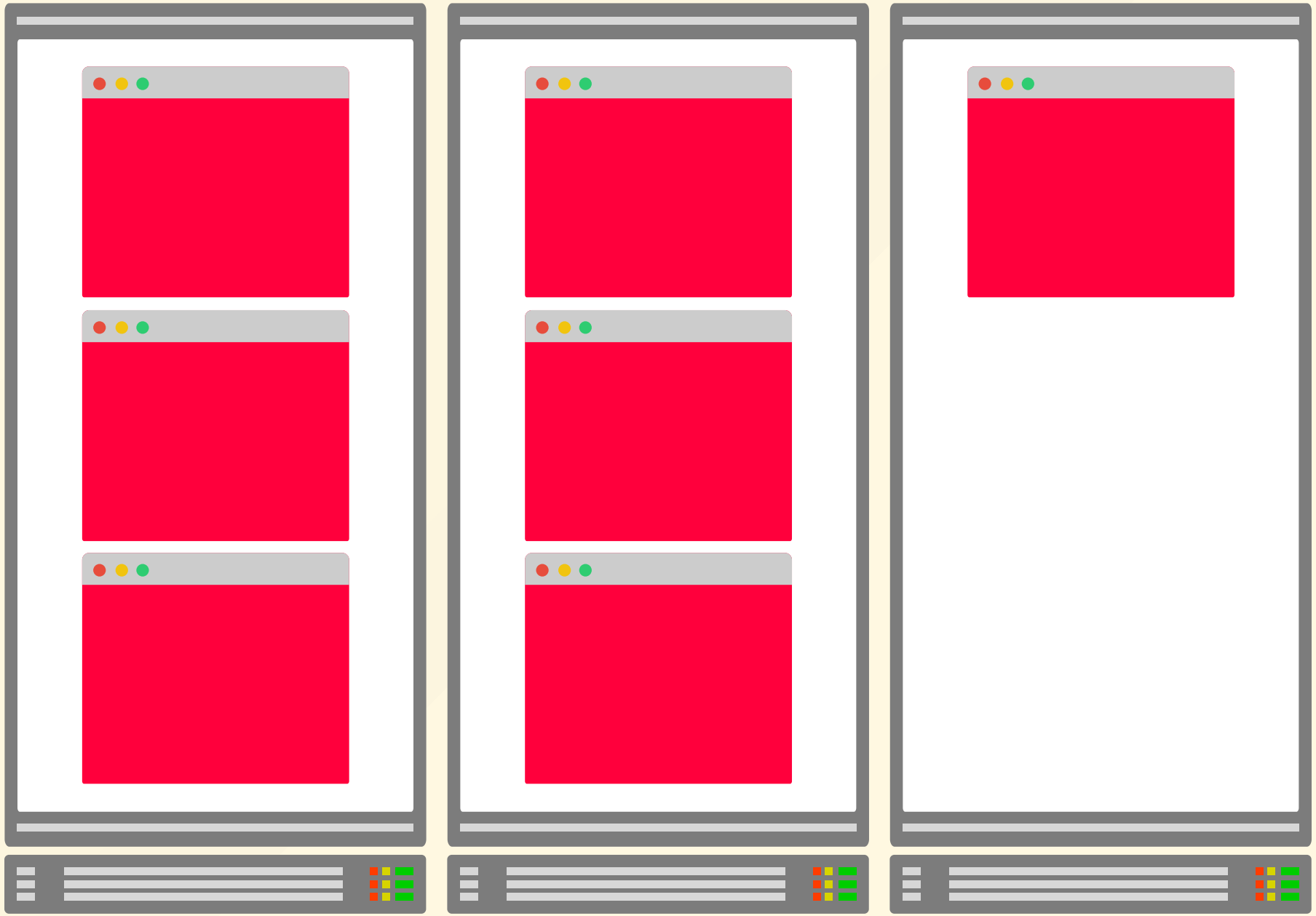
- Learnk8s - Instructor+consultant
- Crown Prosecution Service (UK gov) - Consultant
- Opensource



Cluster AutoScaler

github.com/kubernetes/autoscaler/tree/master/cluster-autoscaler



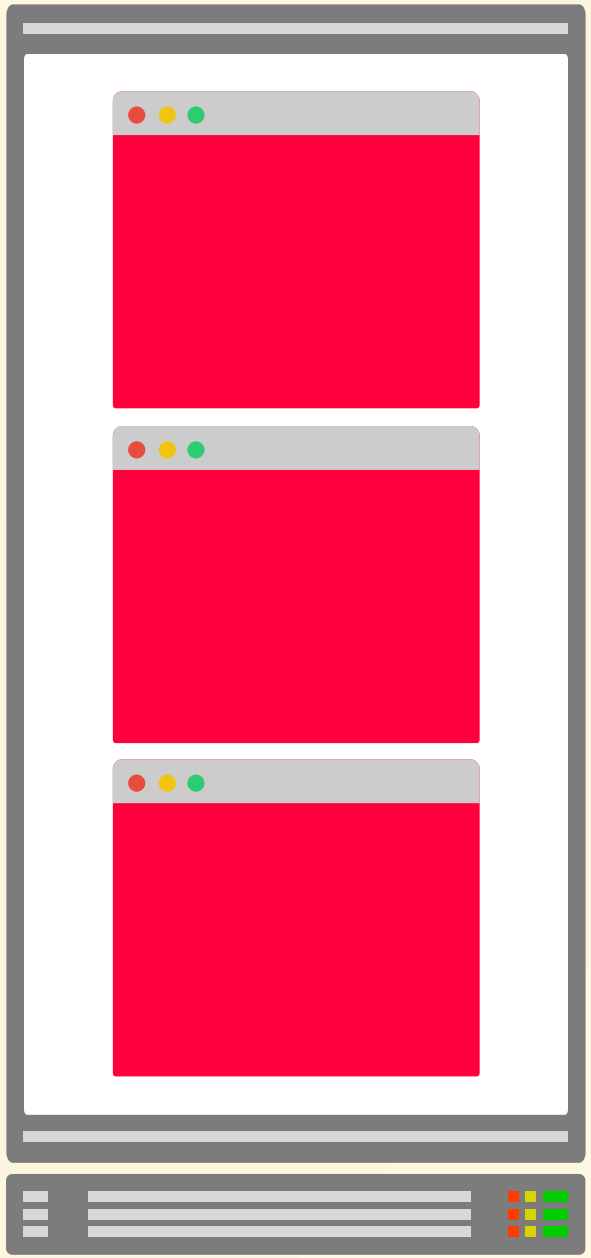


Cluster AutoScaler

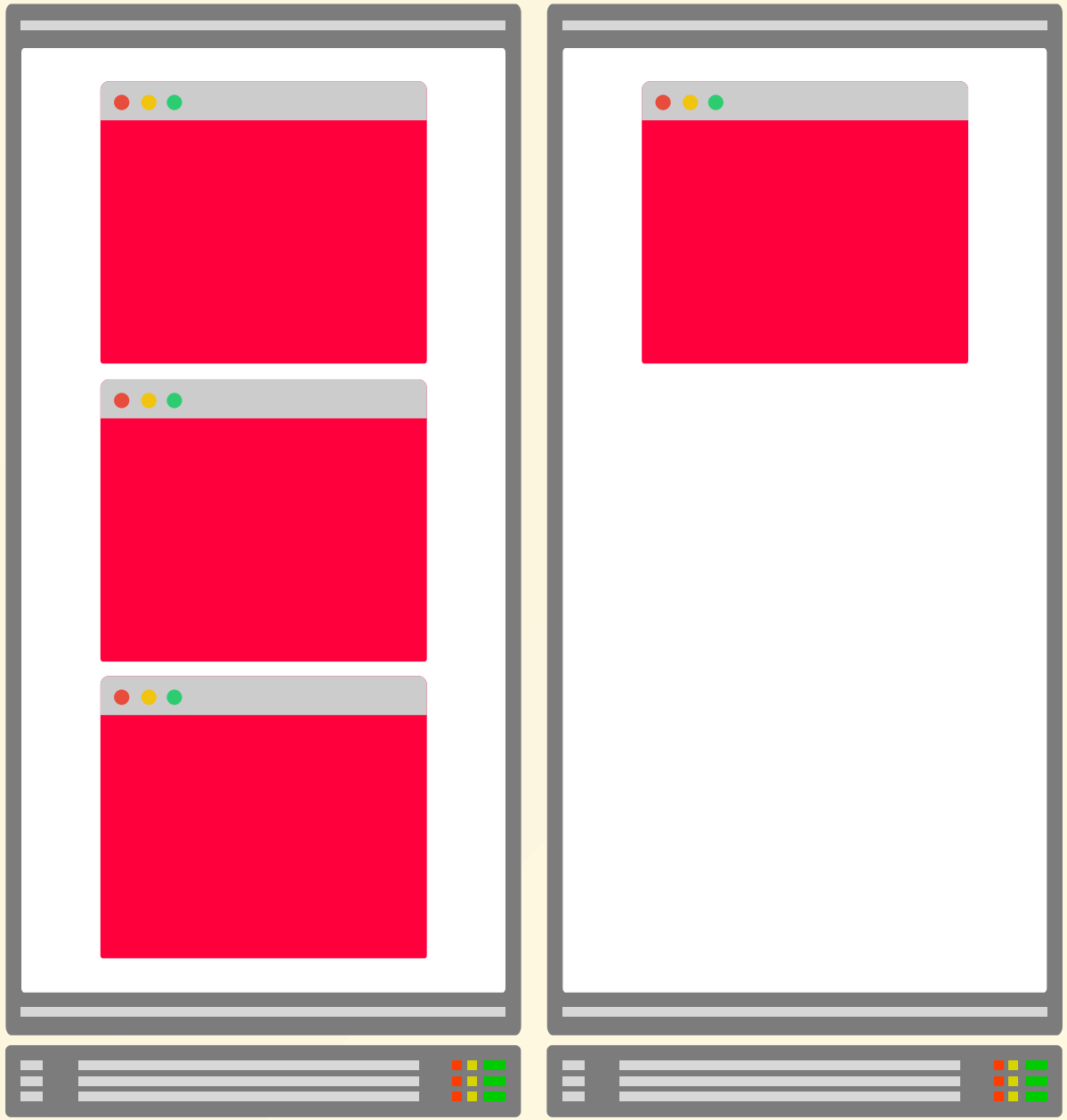
1. Memory Utilization
2. CPU Utilization
3. Pending pods

Cluster AutoScaler

1. ~~Memory Utilization~~
2. ~~CPU Utilization~~
3. Pending pods

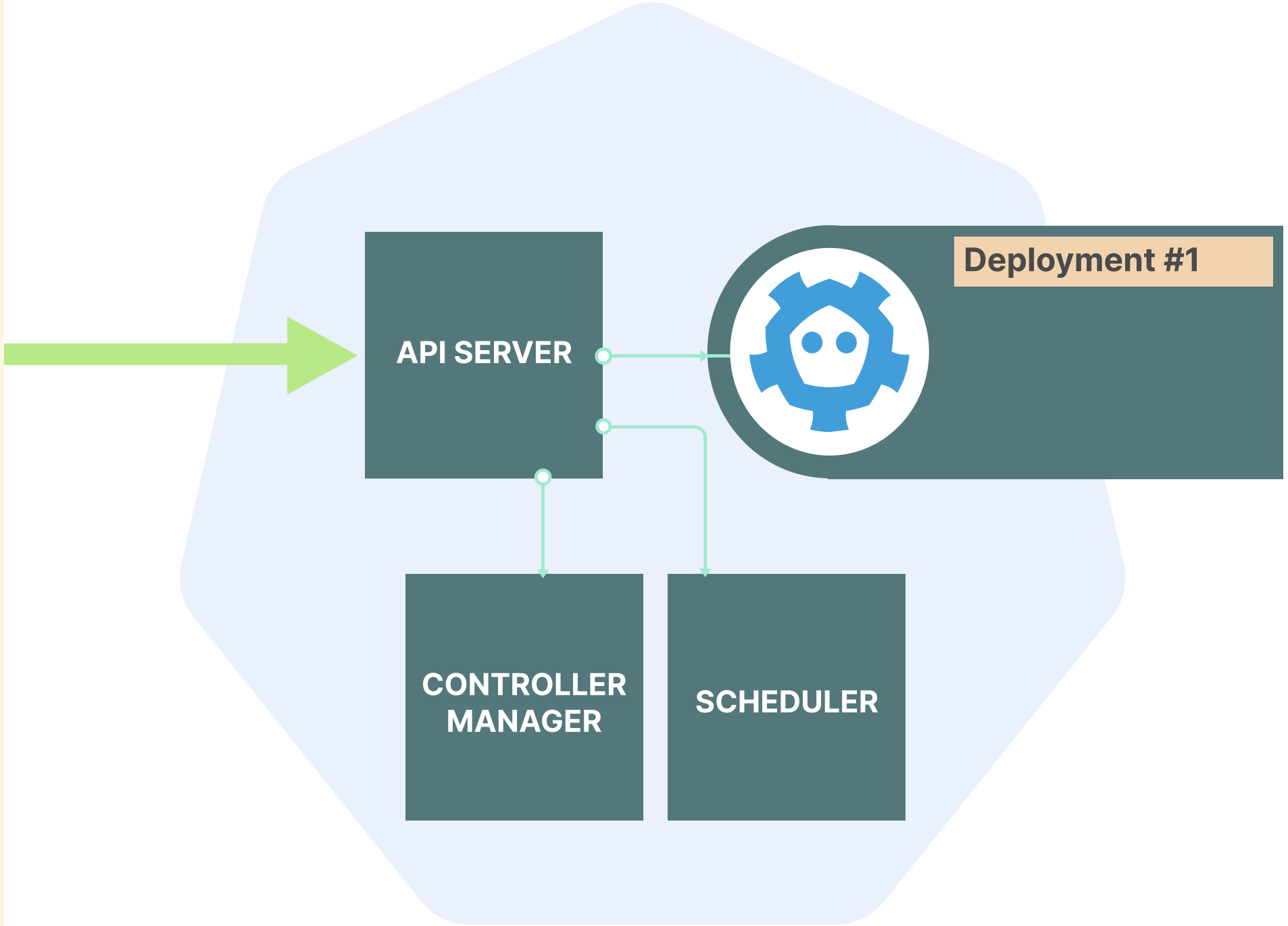


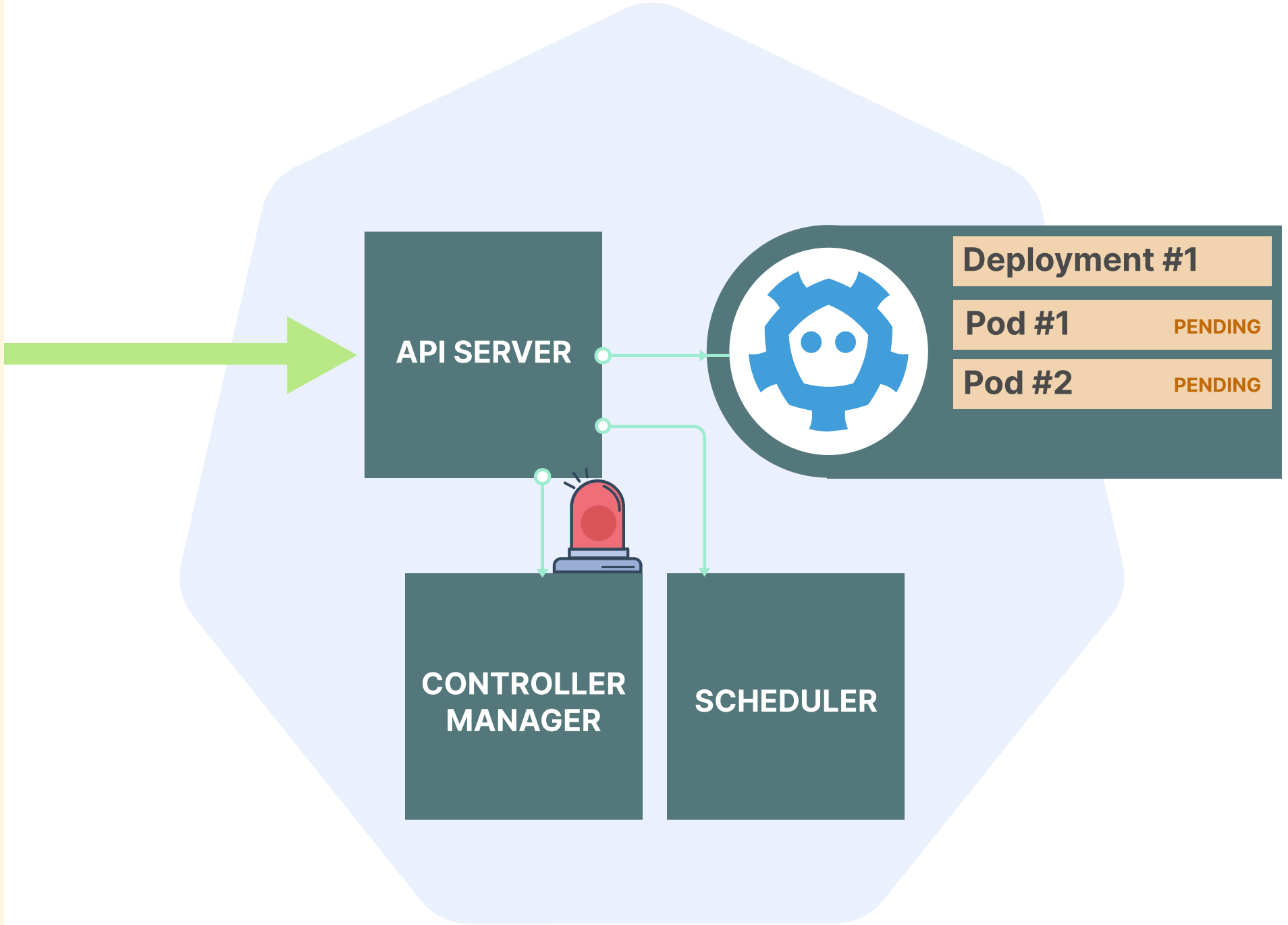
PENDING 😴

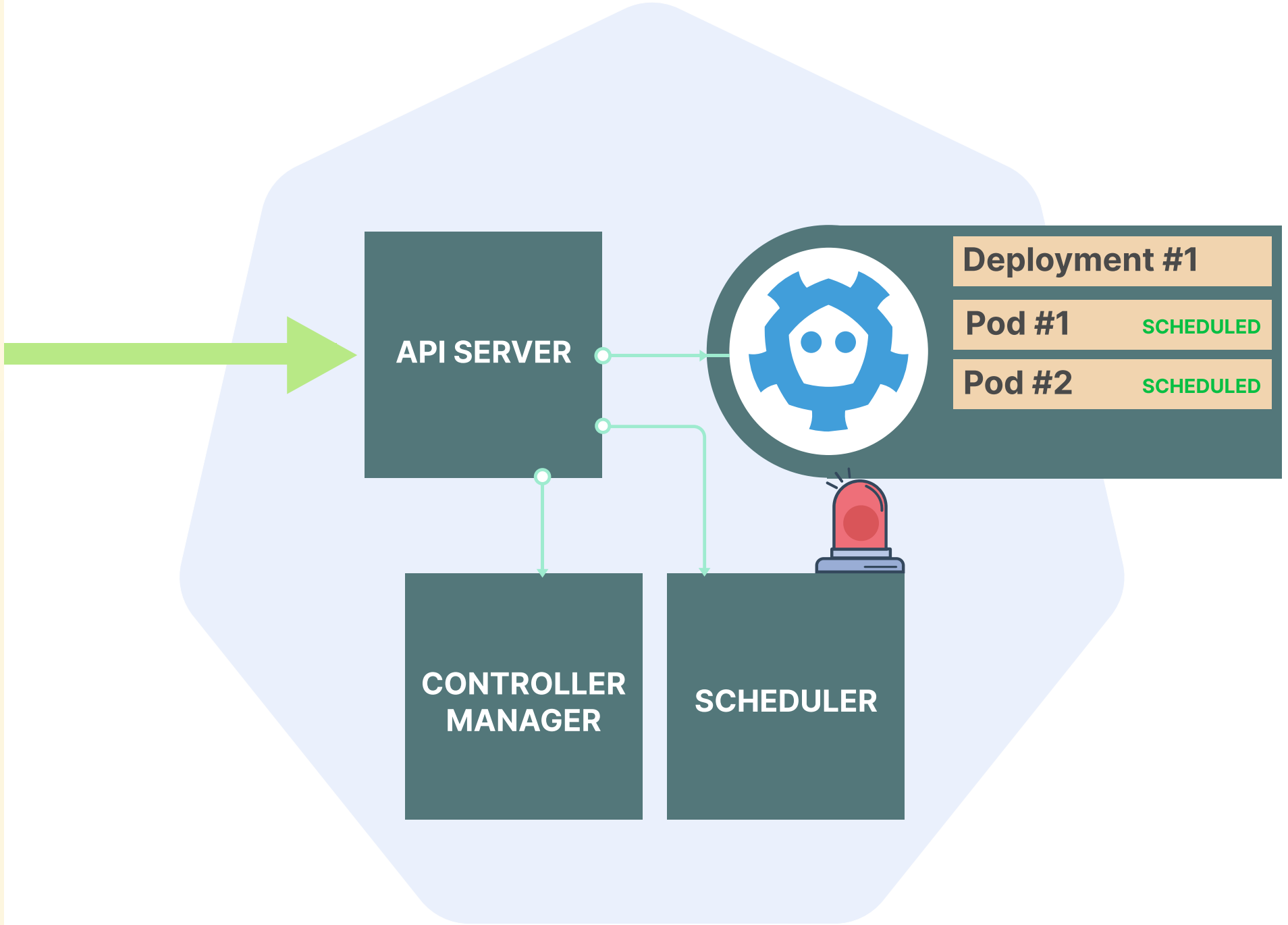


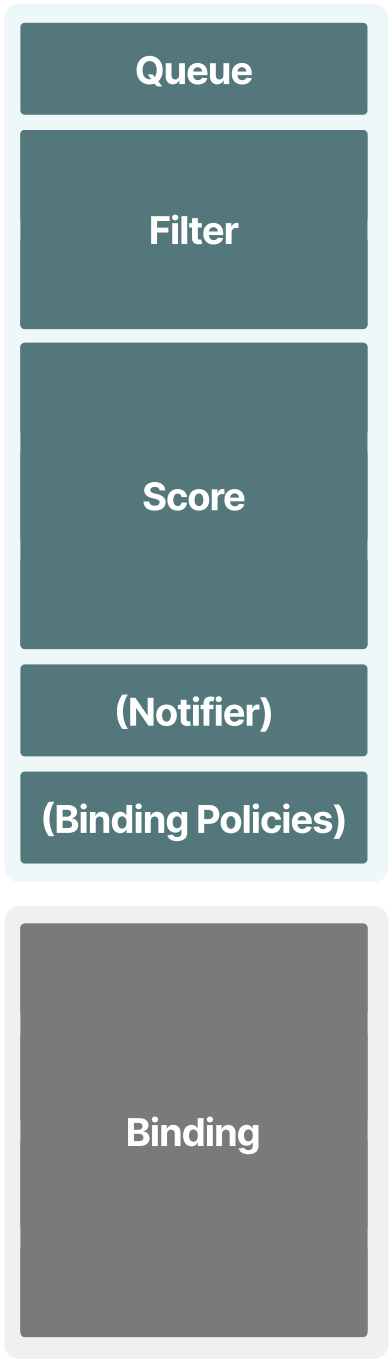
Kubernetes Scheduler

Create a Deployment with 2 replicas


API SERVER
Deployment #1
CONTROLLER MANAGER
SCHEDULER


API SERVER
Deployment #1
Pod #1
PENDING
Pod #2
PENDING
CONTROLLER MANAGER
SCHEDULER

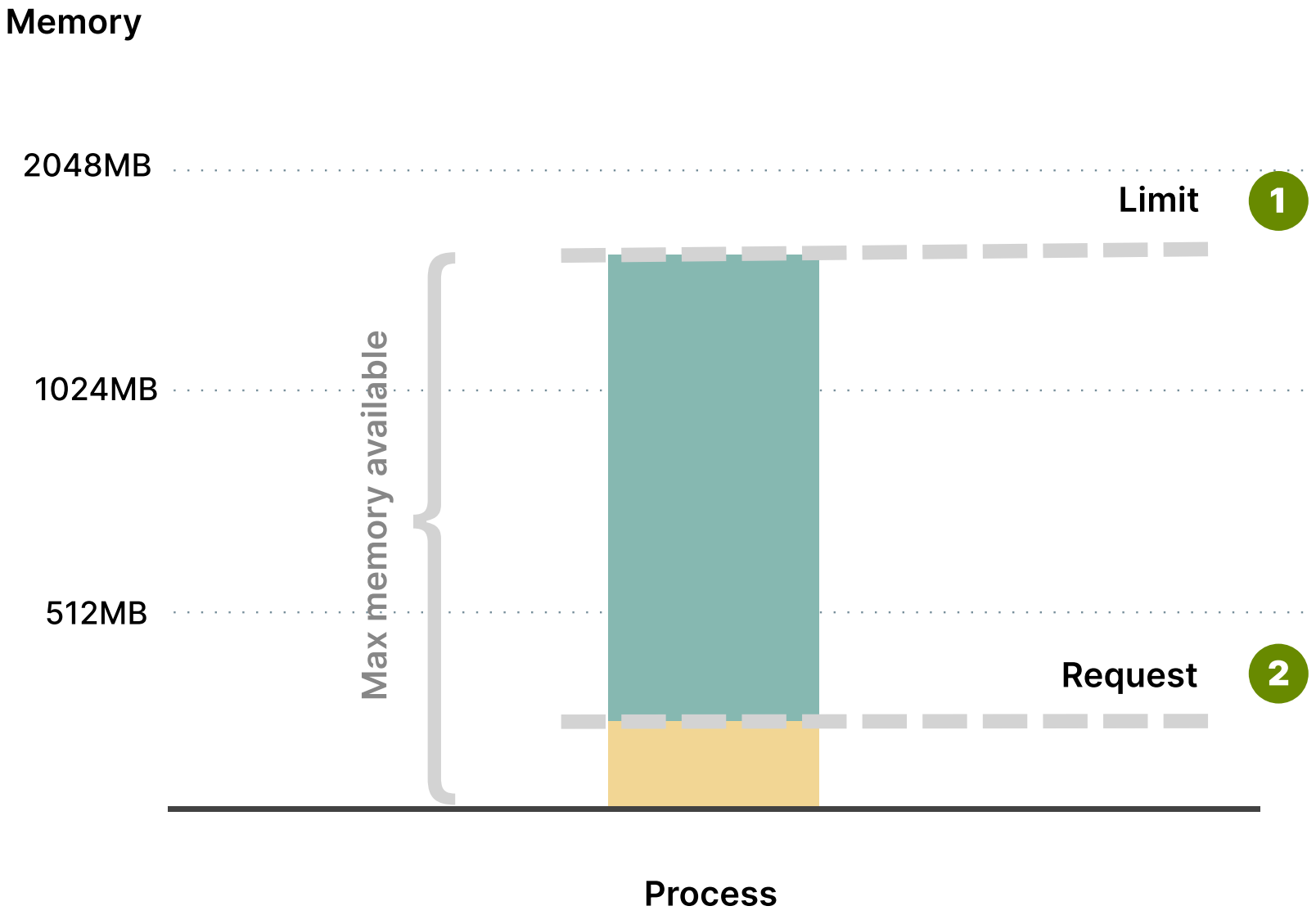

API SERVER
Deployment #1
Pod #1
SCHEDULED
Pod #2
SCHEDULED
CONTROLLER MANAGER
SCHEDULER



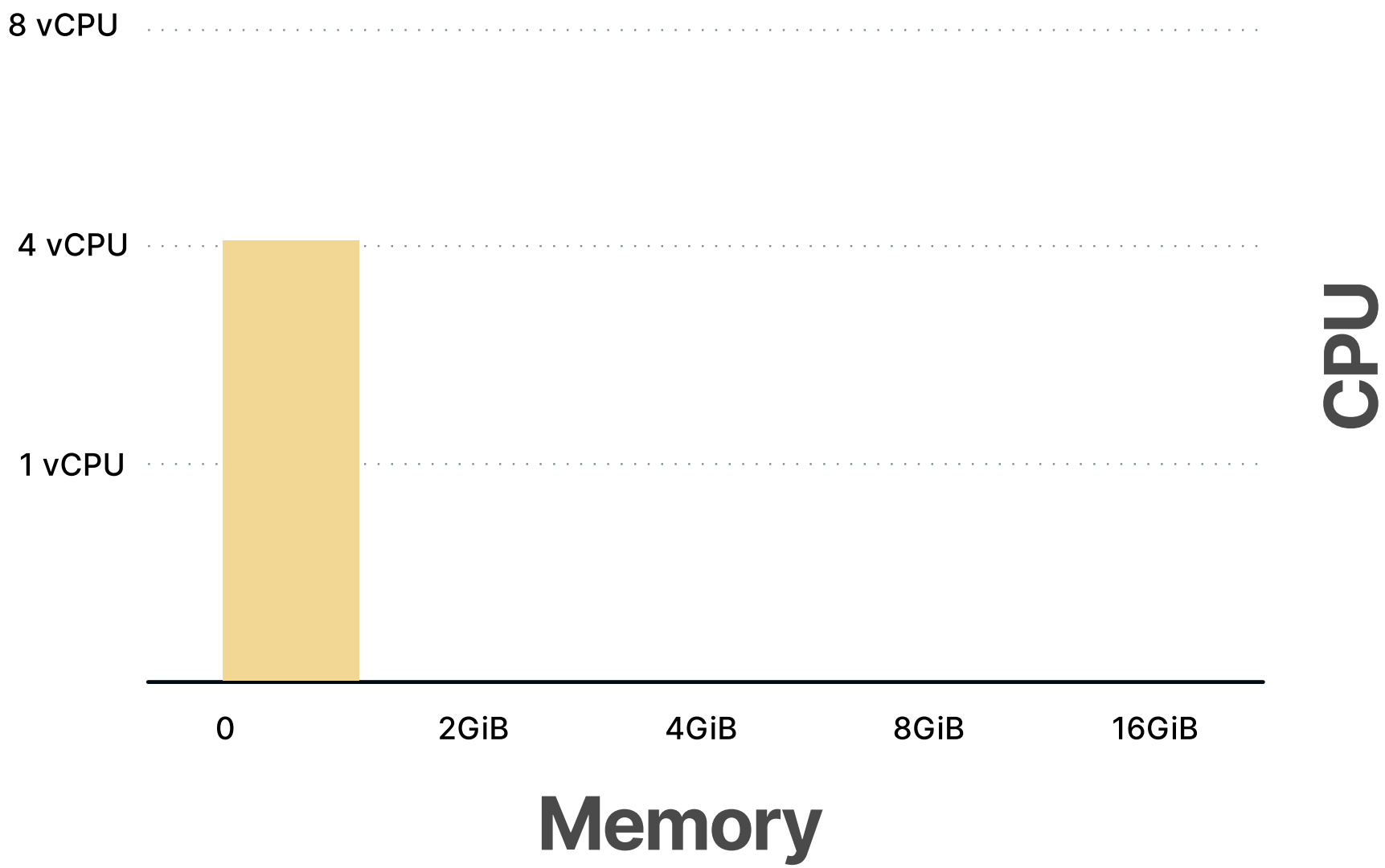
Scheduling

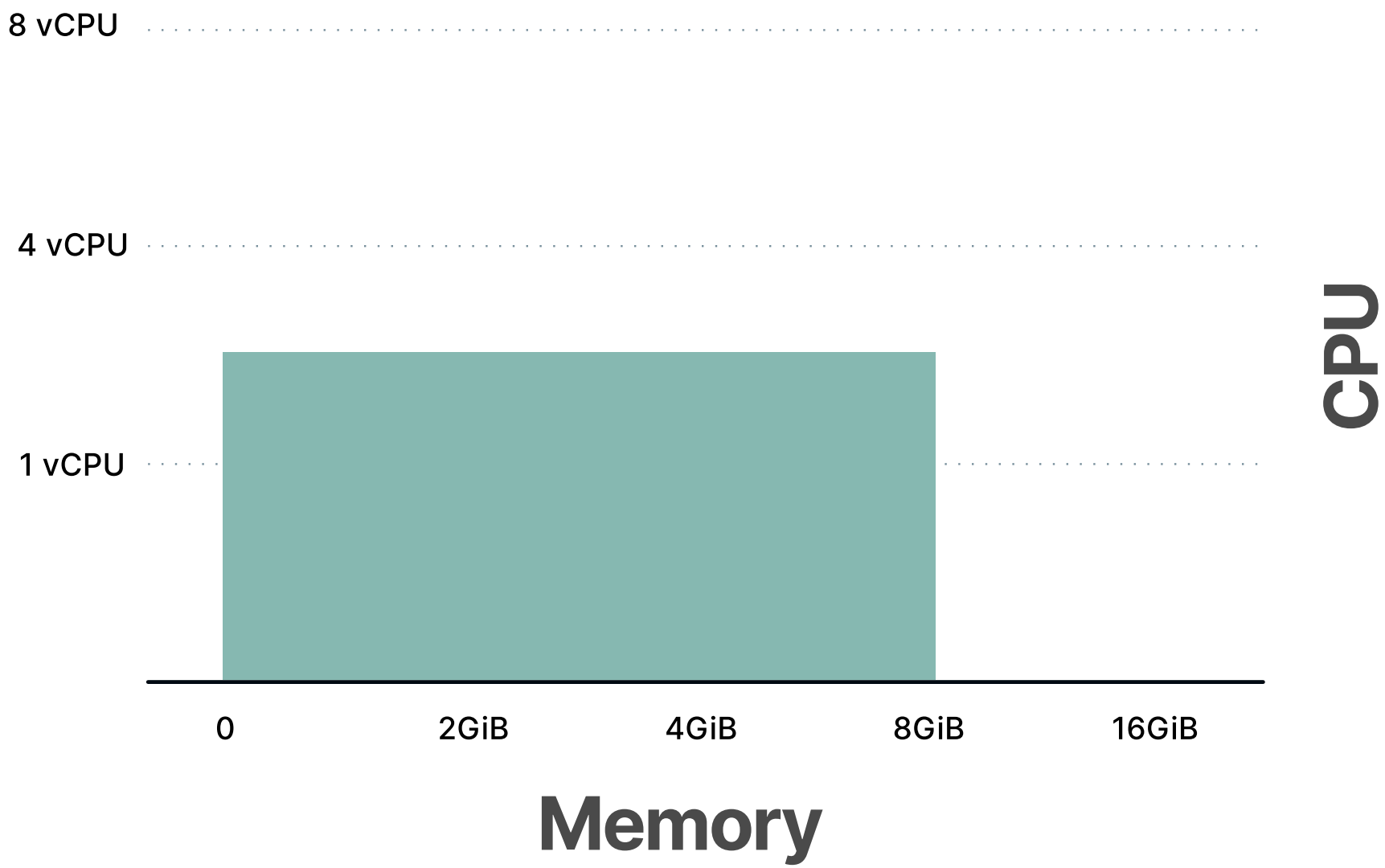
Binding

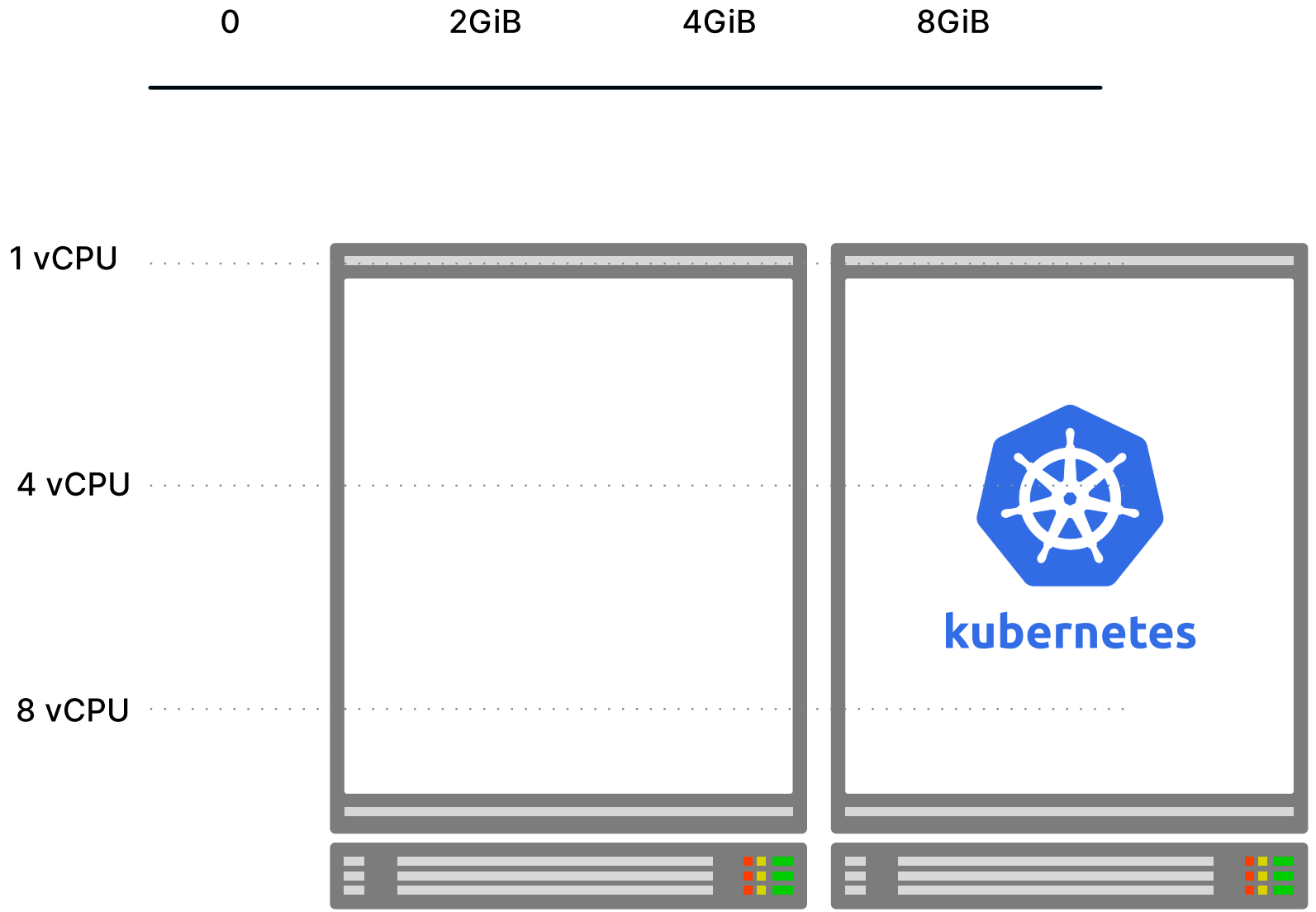
Requests & Scheduler


Memory
2048MB
1024MB
512MB
Max memory available
Limit
1
Request
2
Process

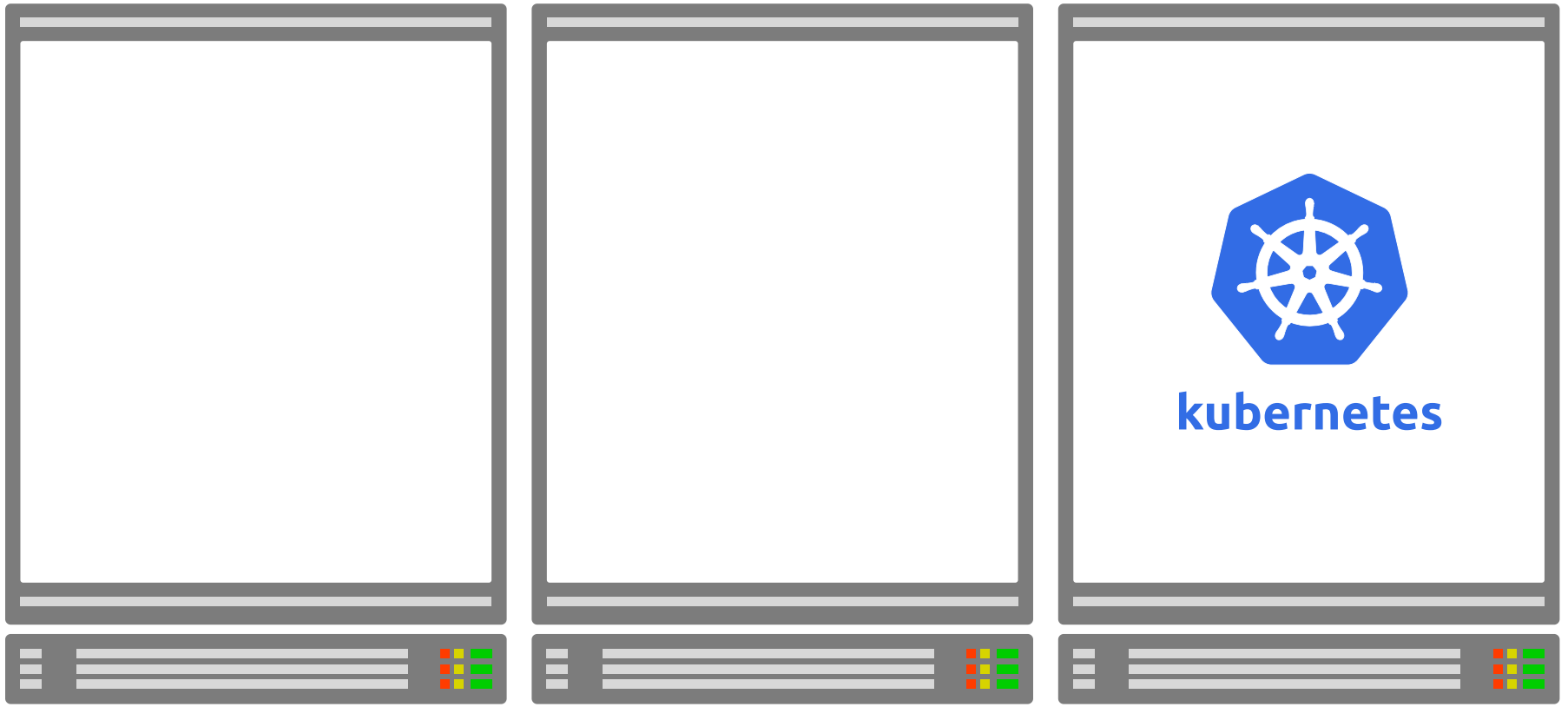
```yaml
---
resources:
  requests:
    memory: "64Mi"
    cpu: "250m"
  limits:
    memory: "128Mi"
    cpu: "500m"
```

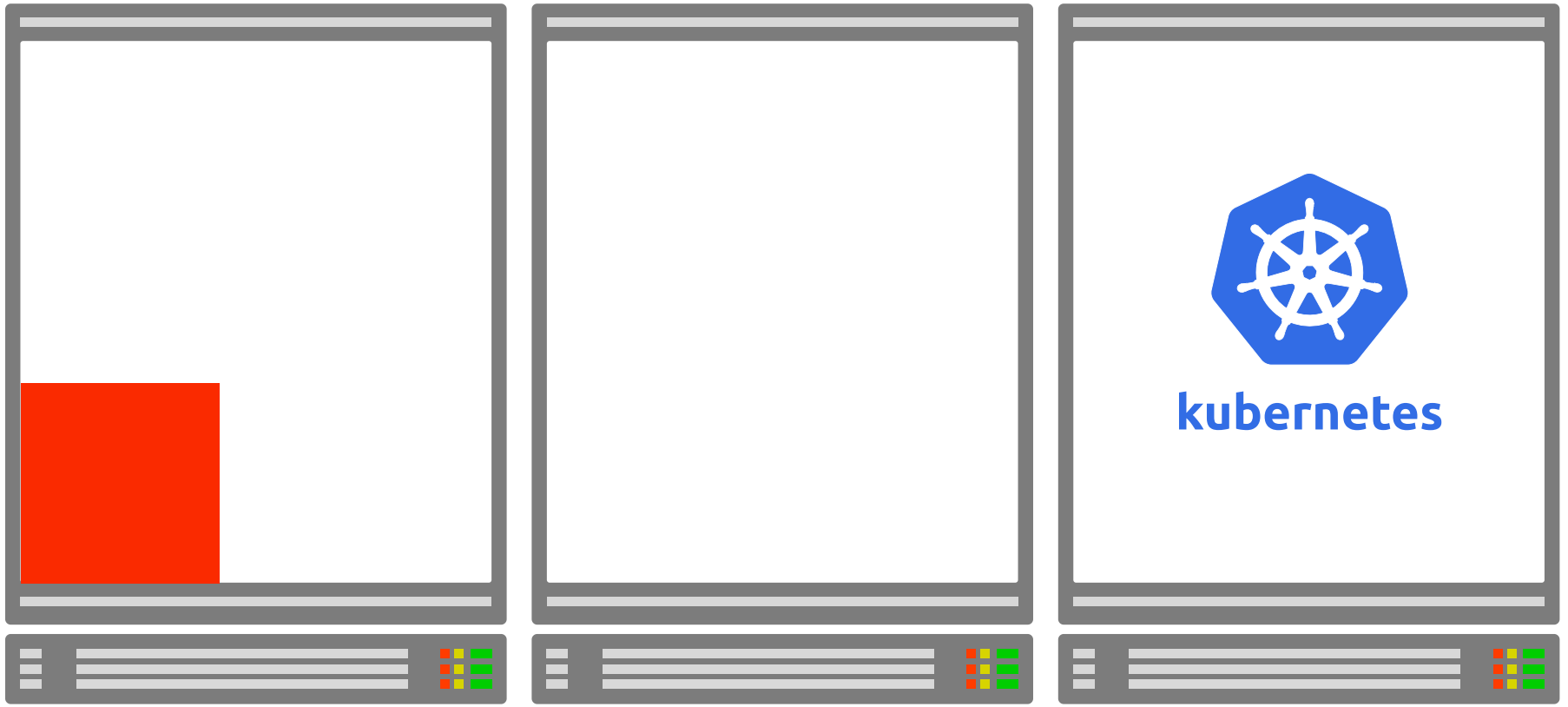

8 vCPU
4 vCPU
1 vCPU
CPU
0
2GiB
4GiB
8GiB
16GiB
Memory


8 vCPU
4 vCPU
1 vCPU
CPU
0
2GiB
4GiB
8GiB
16GiB
Memory

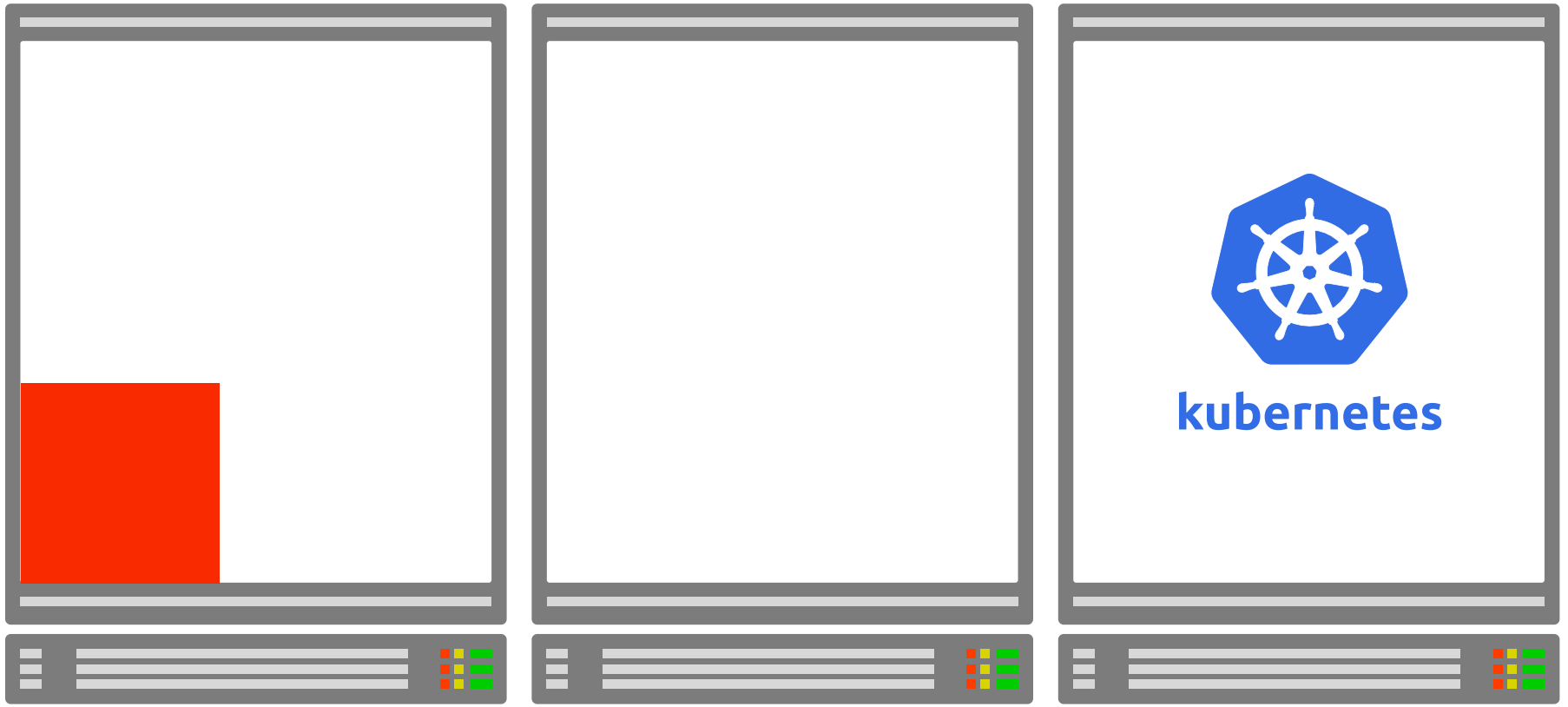

0
2GiB
4GiB
8GiB
1 vCPU
4 vCPU
8 vCPU
kubernetes

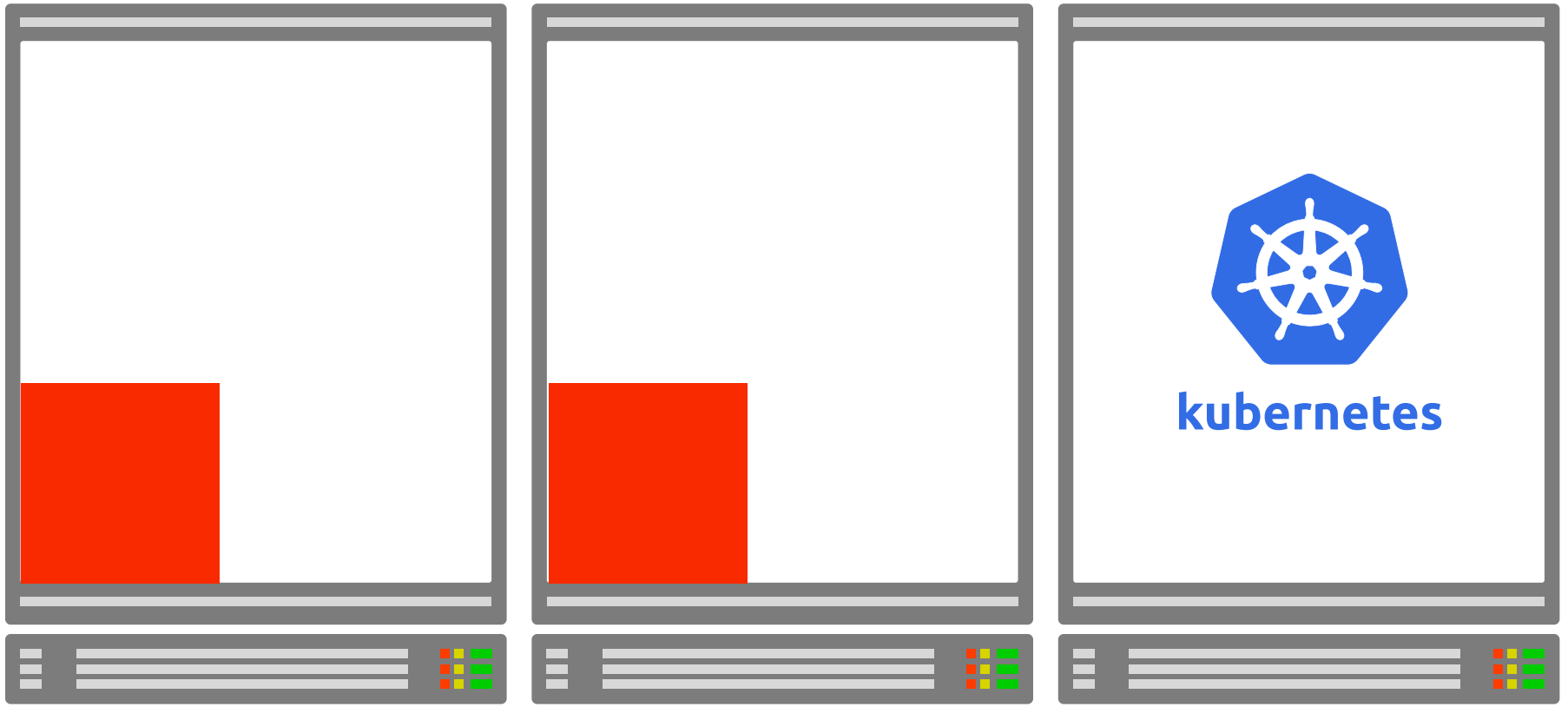
Next container



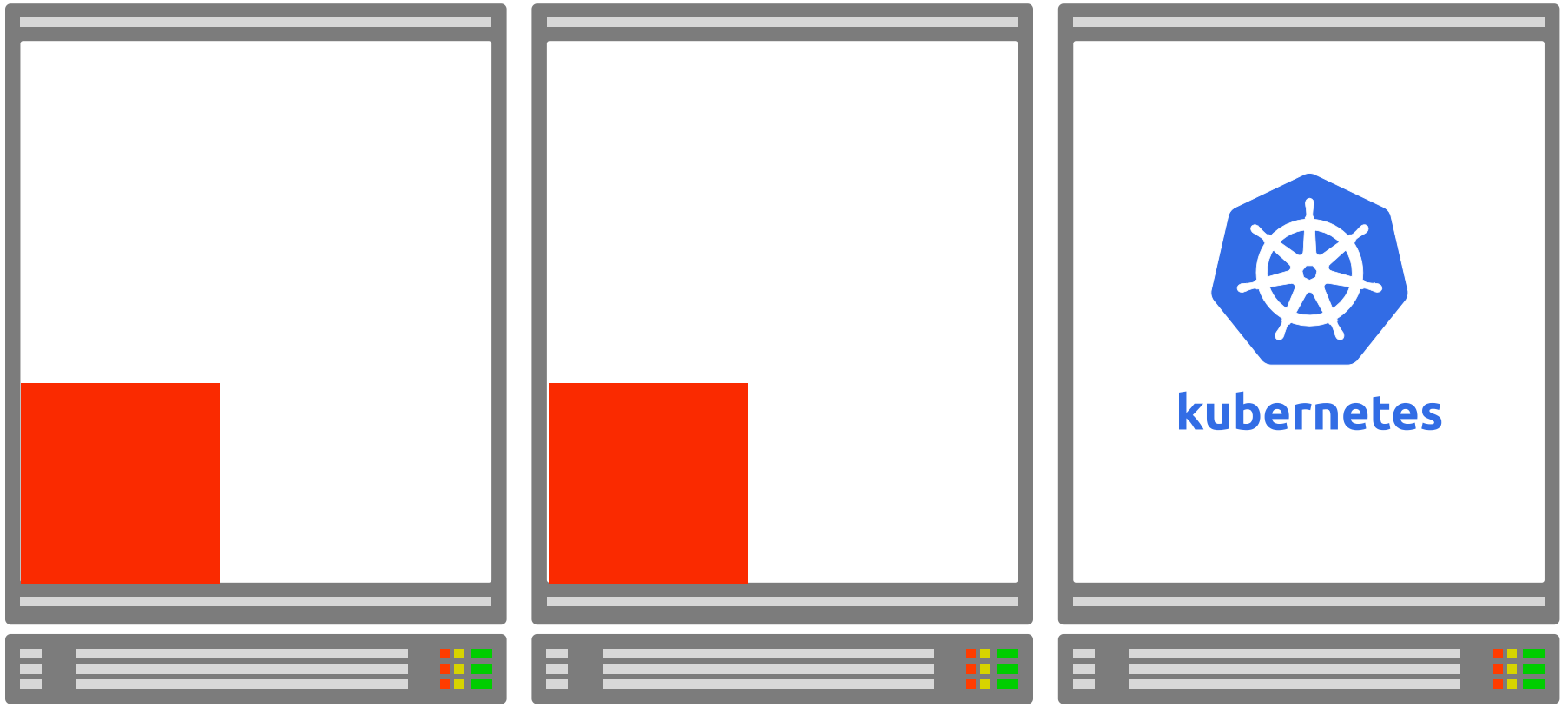

kubernetes

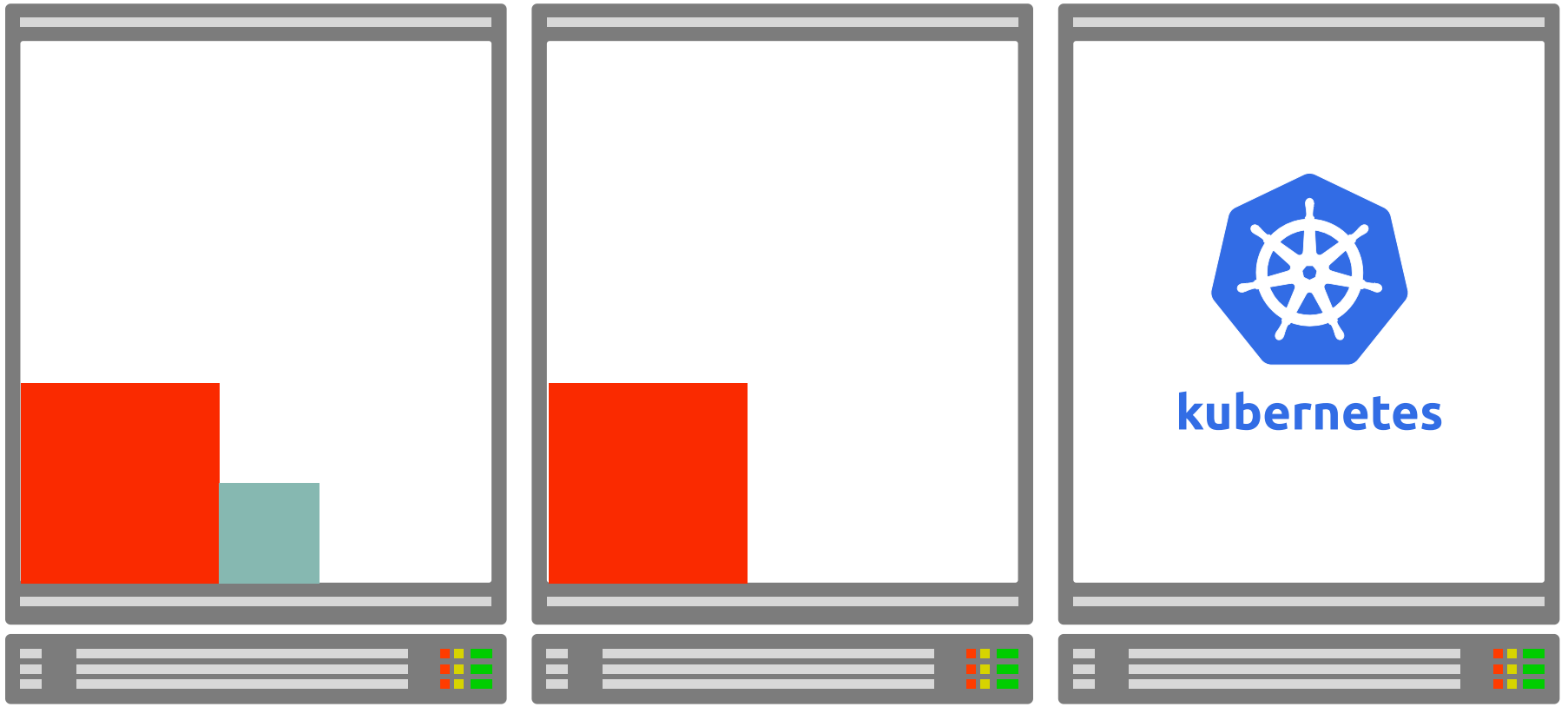
Next container



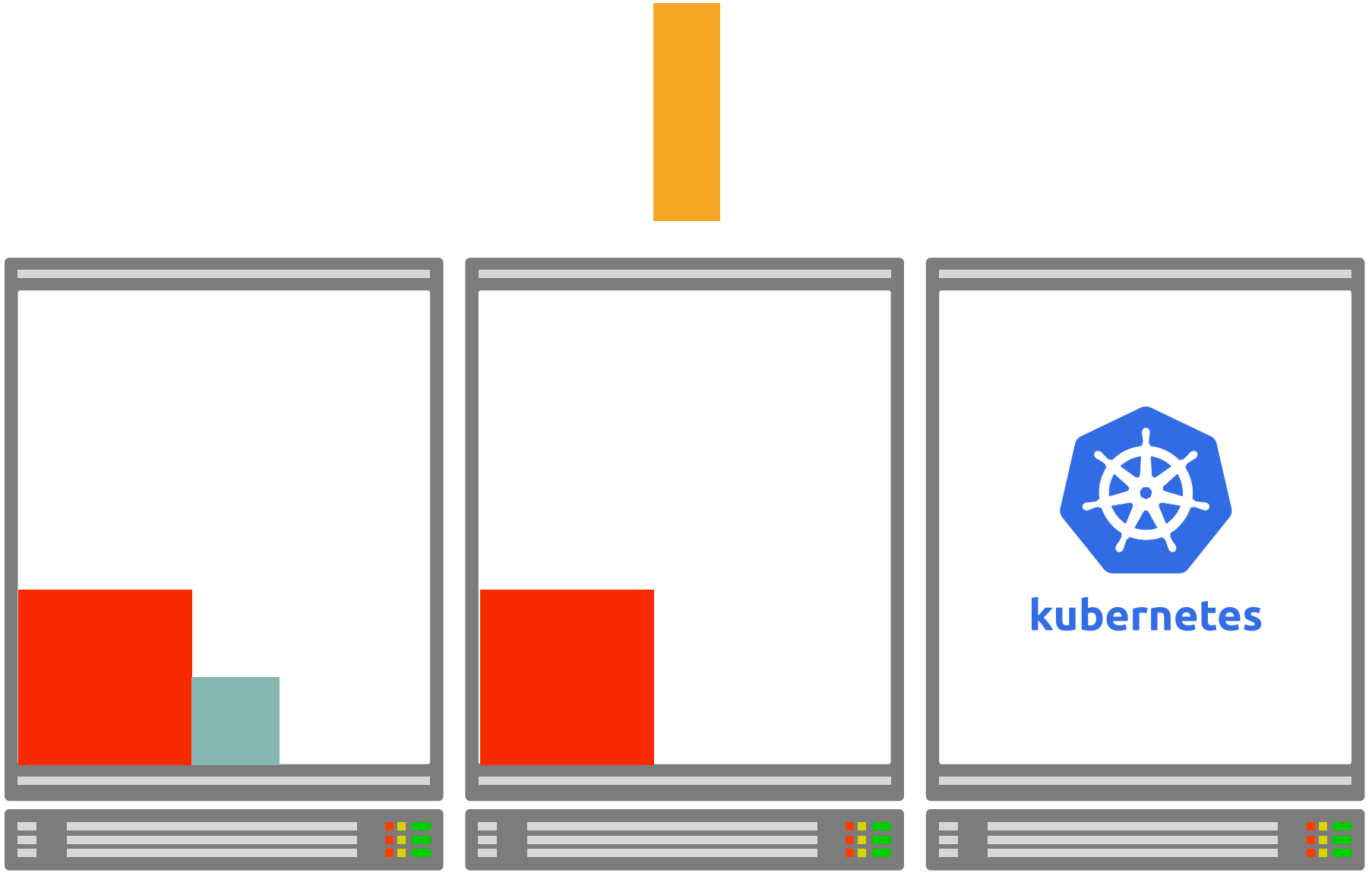

kubernetes

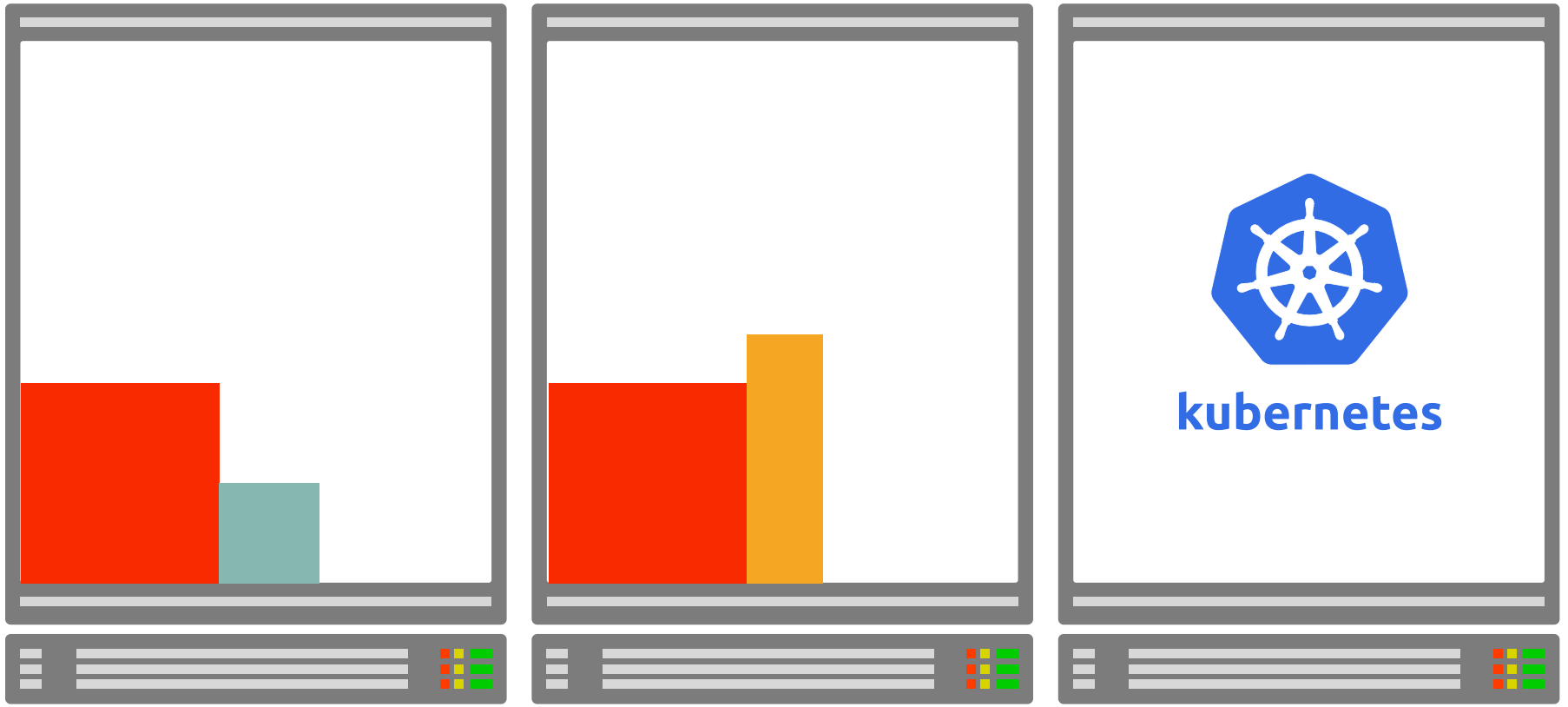
Next container



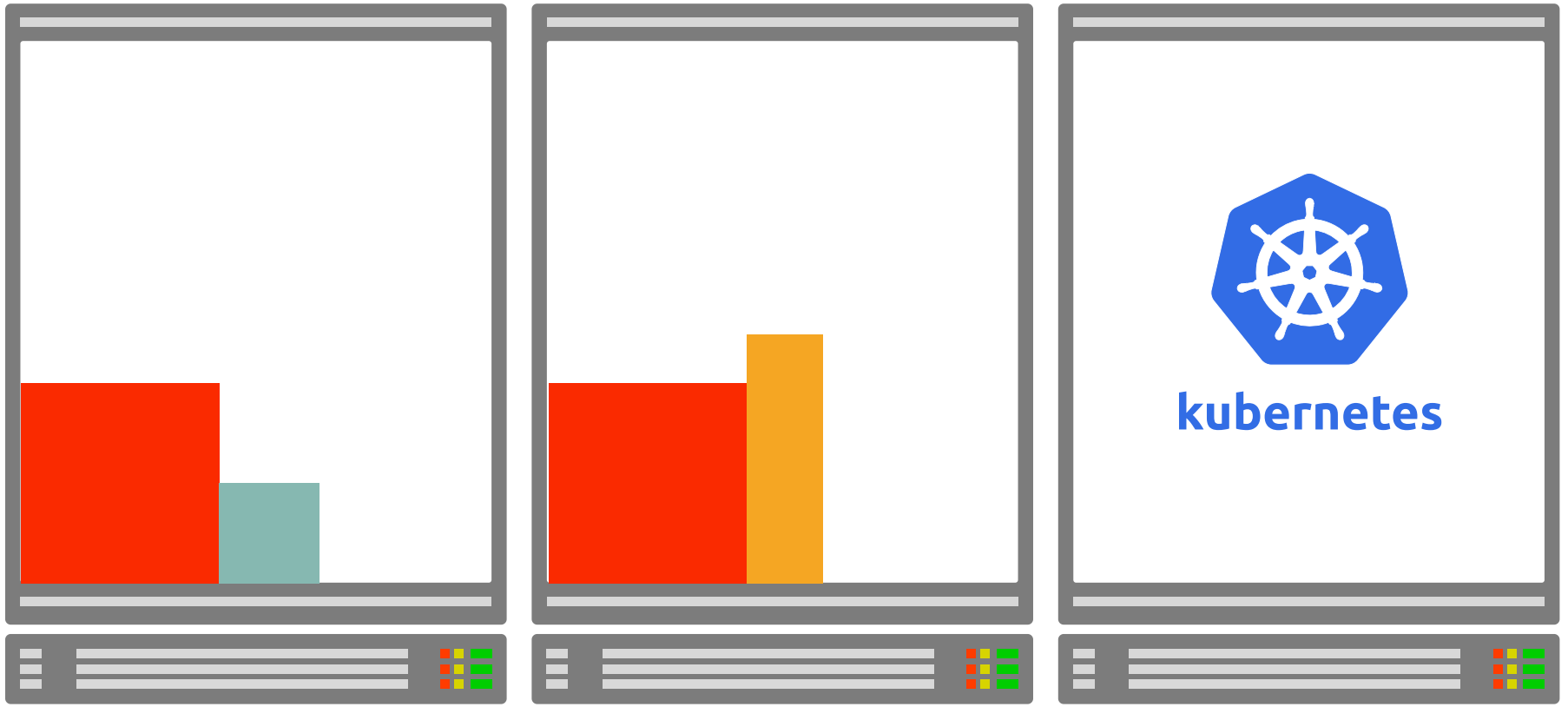

kubernetes

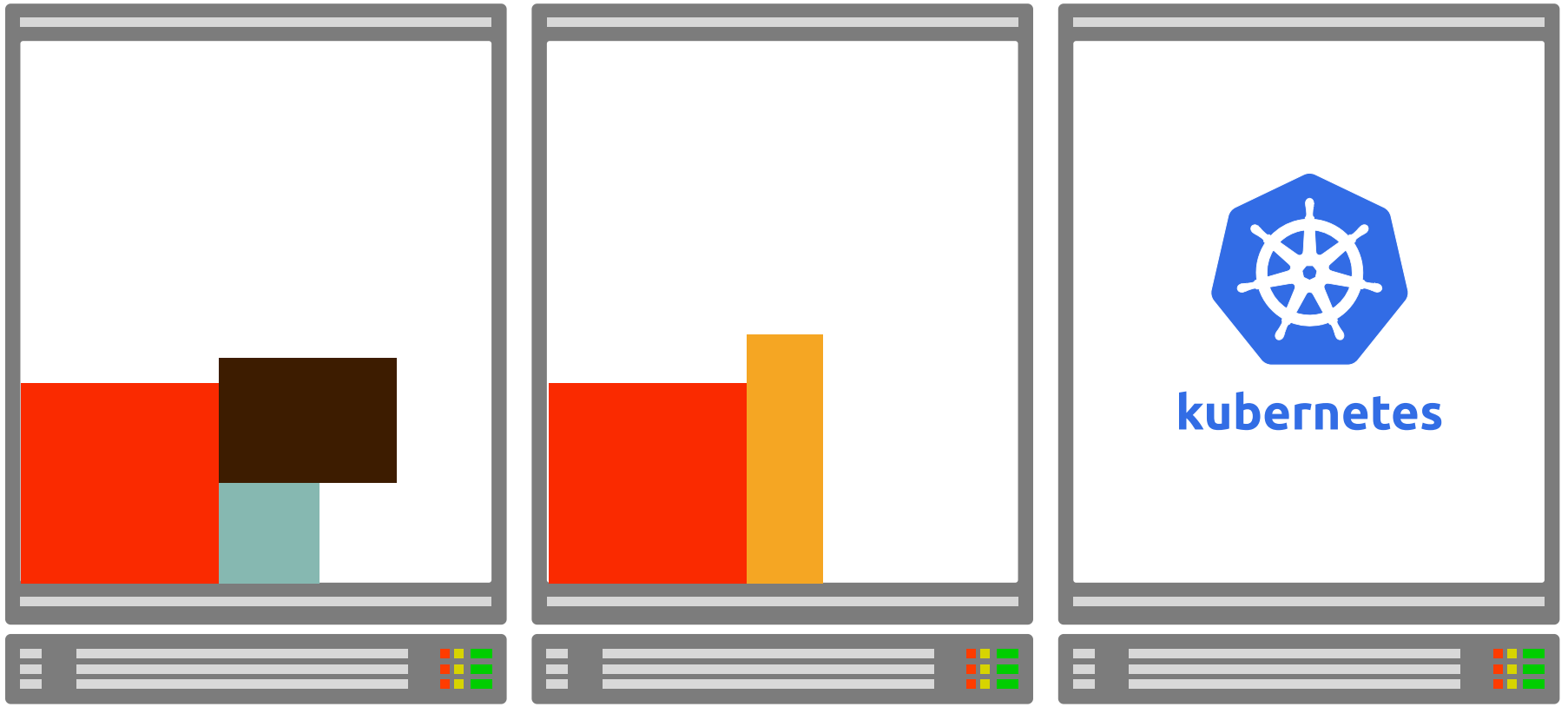
Next container

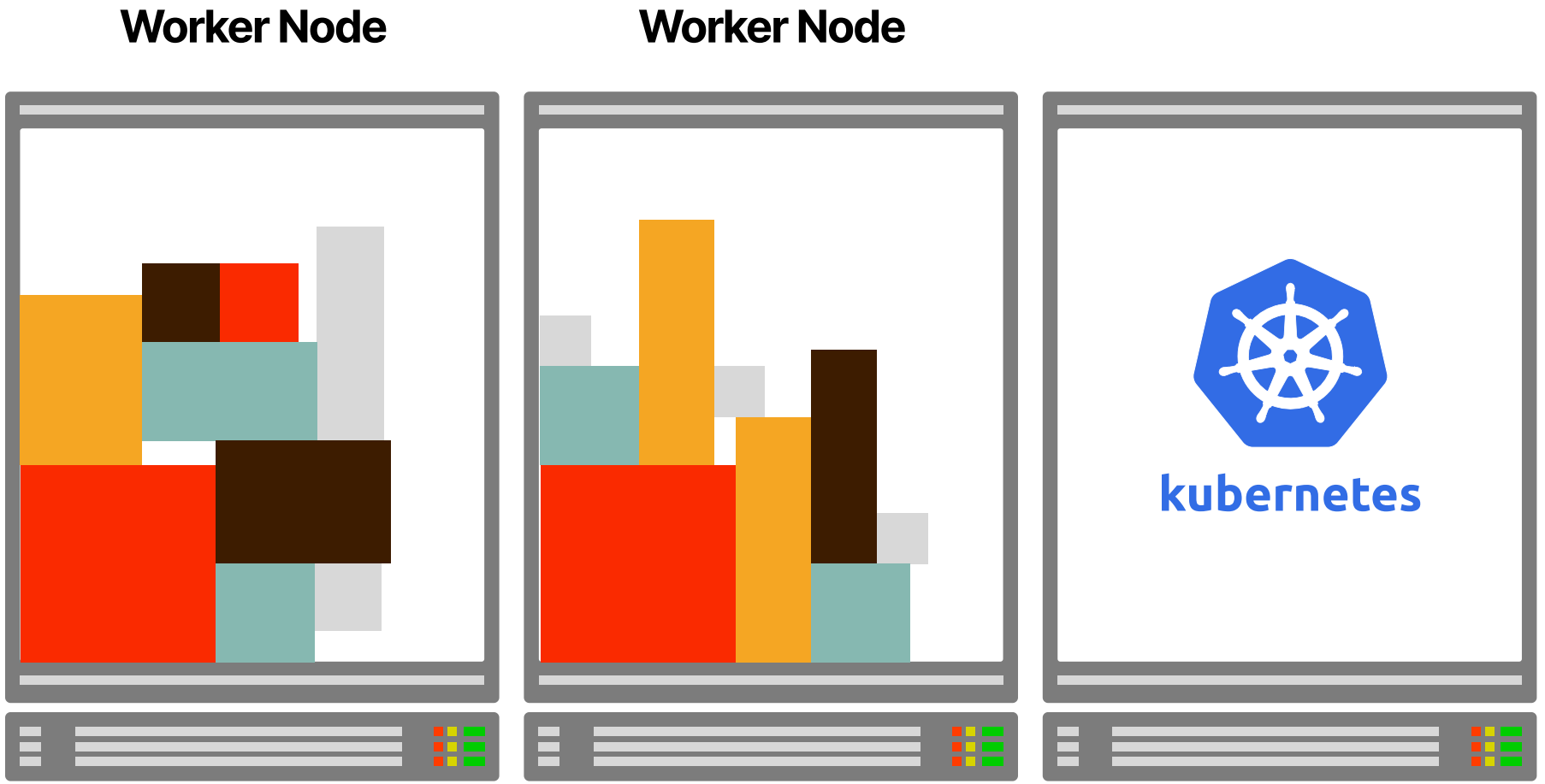



kubernetes

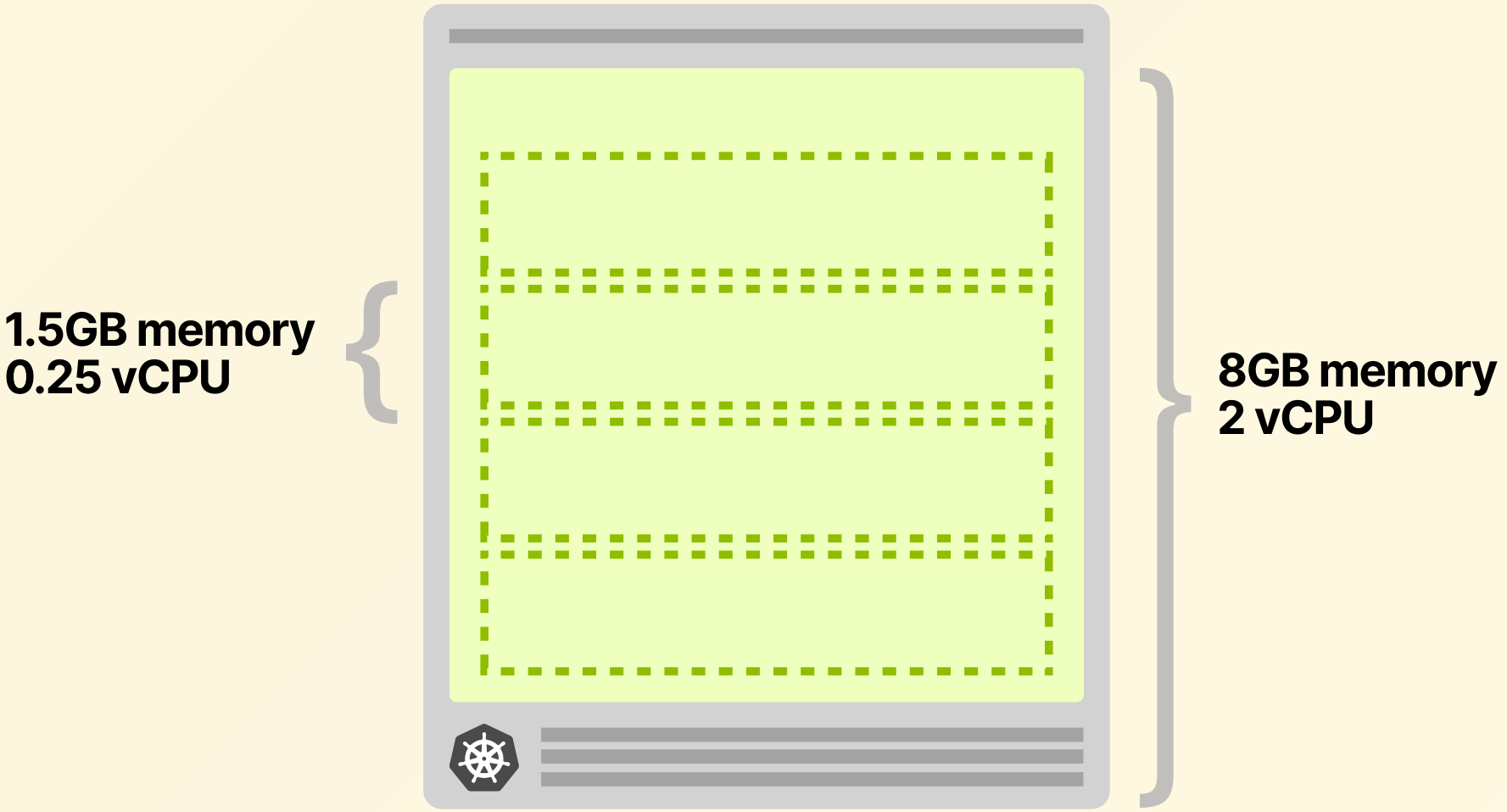
Next container

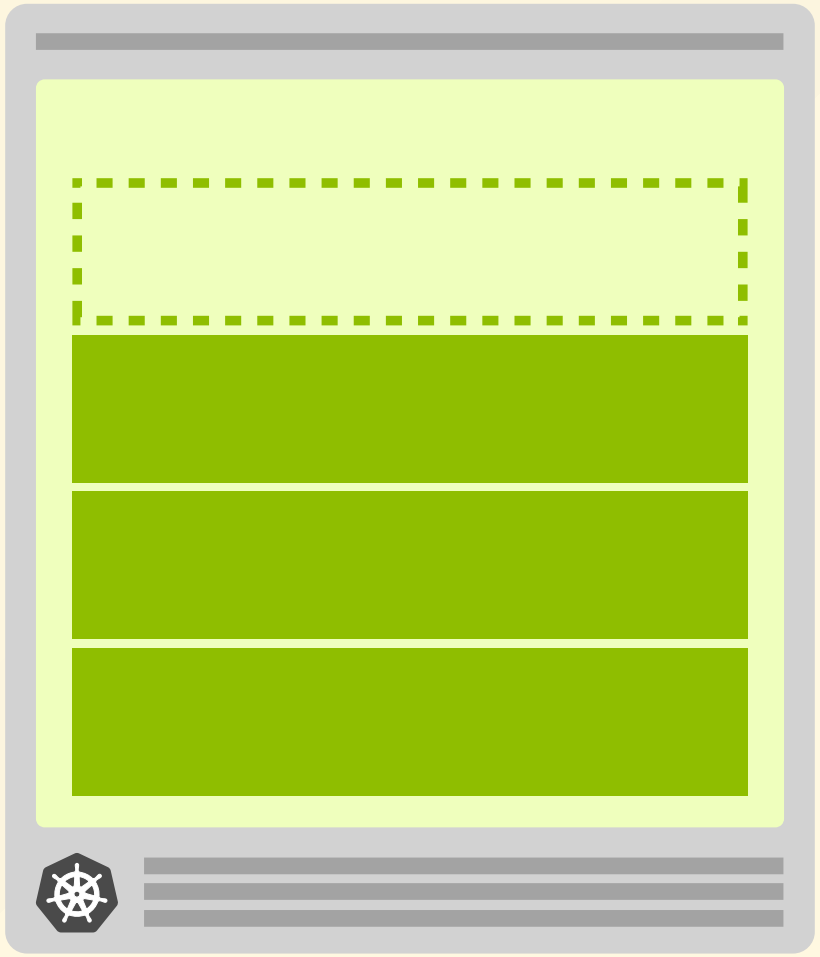


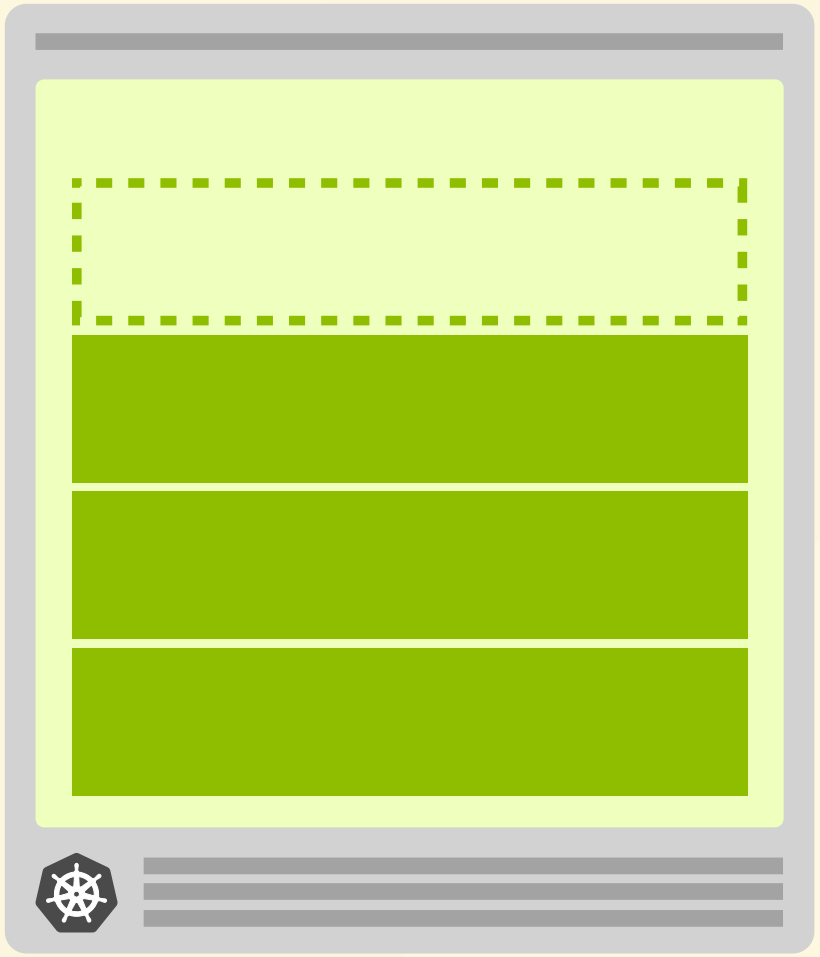

kubernetes


Worker Node
Worker Node
kubernetes

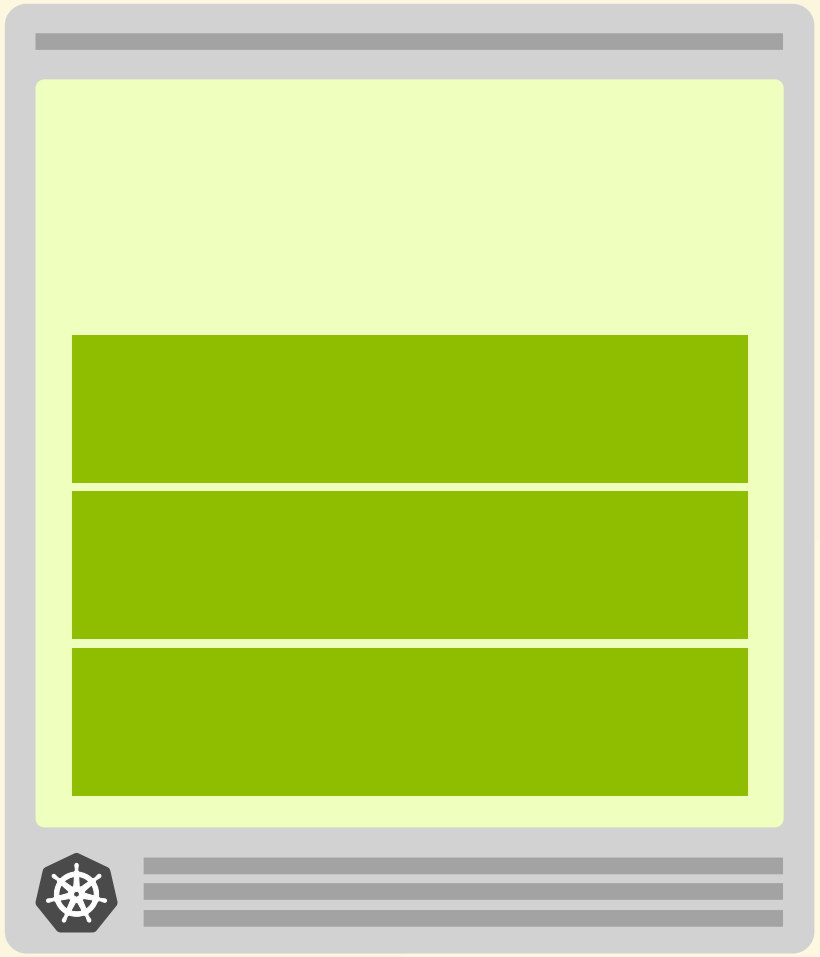
Requests & Limits


1.5GB memory
0.25 vCPU
8GB memory
2 vCPU

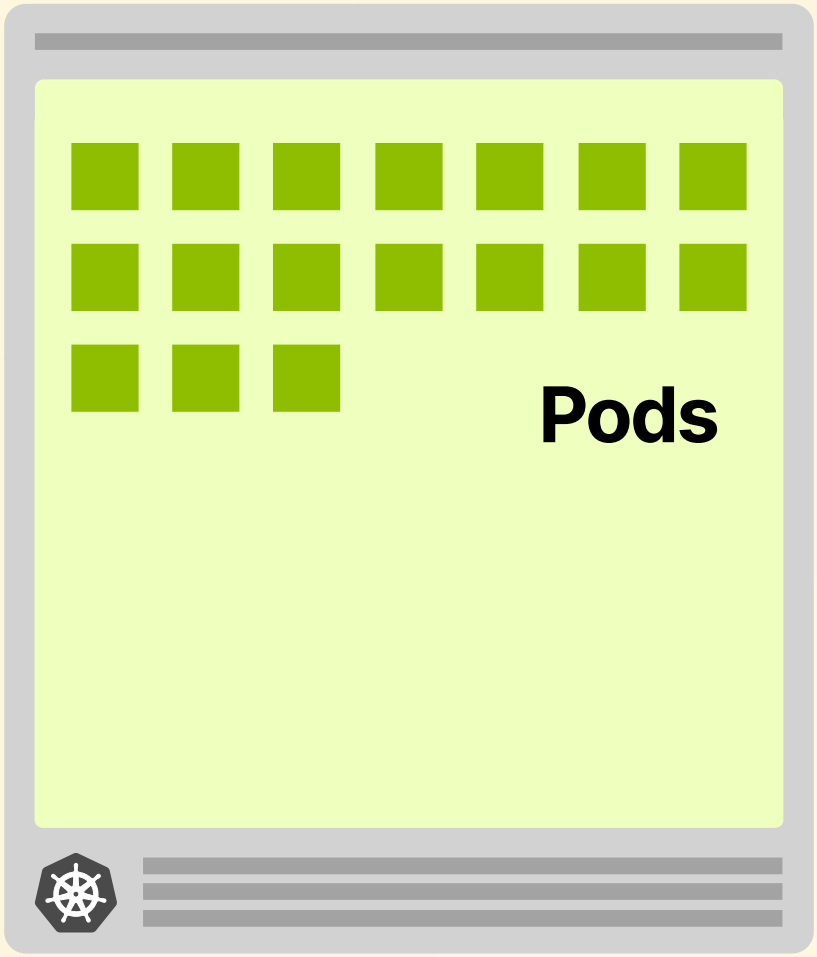


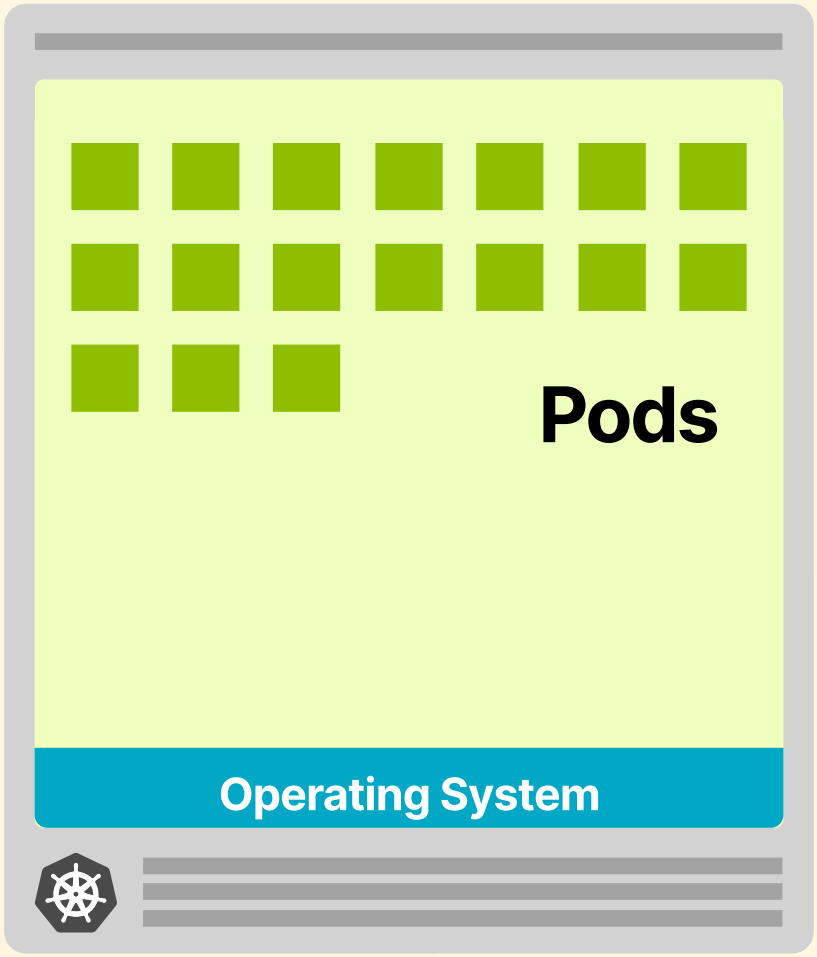


Pending

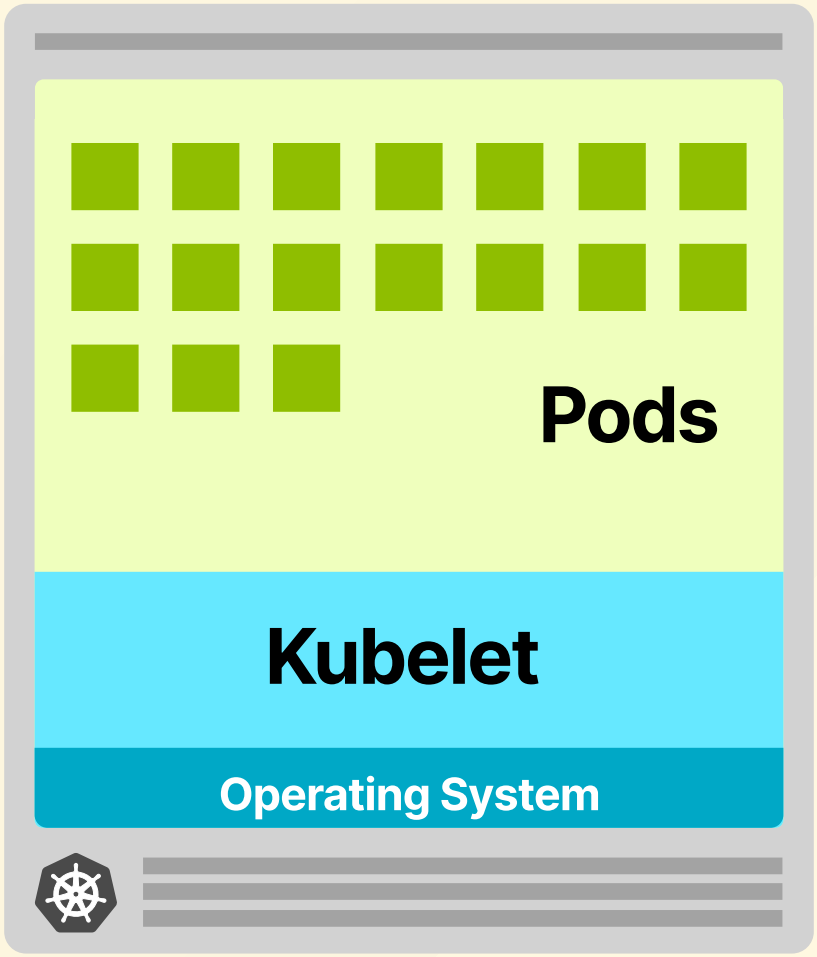


Node instances


Pods

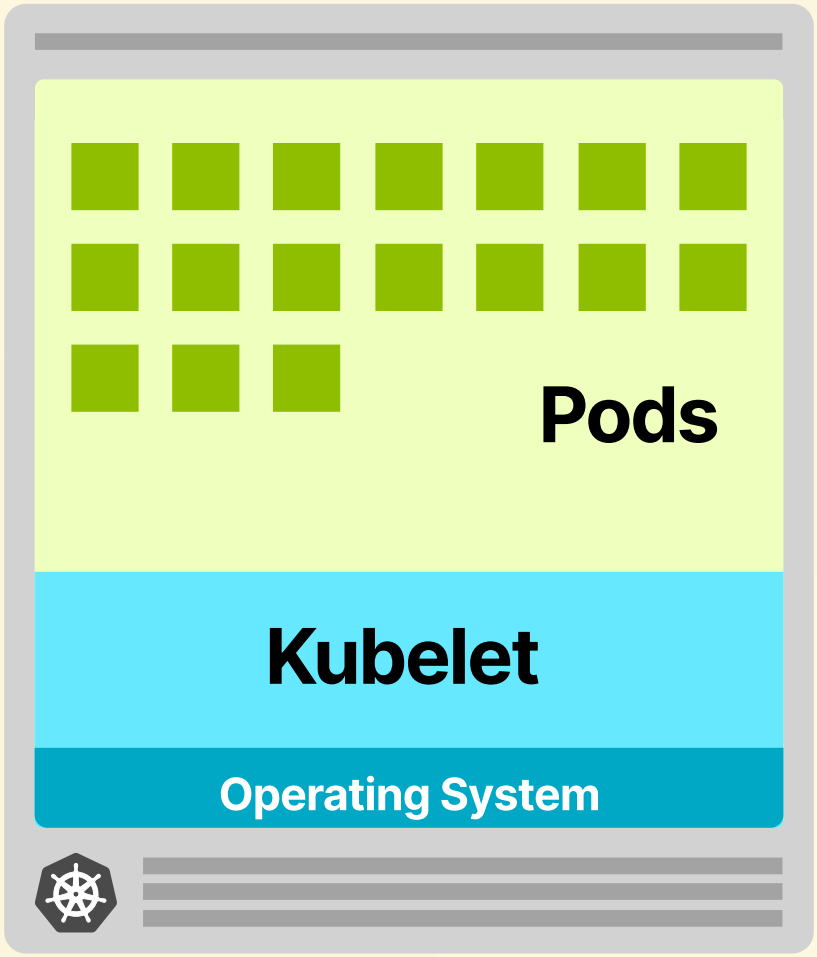


4 **Memory and CPU reserved to the OS**



3 Memory and CPU reserved to the kubelet

4 Memory and CPU reserved to the OS

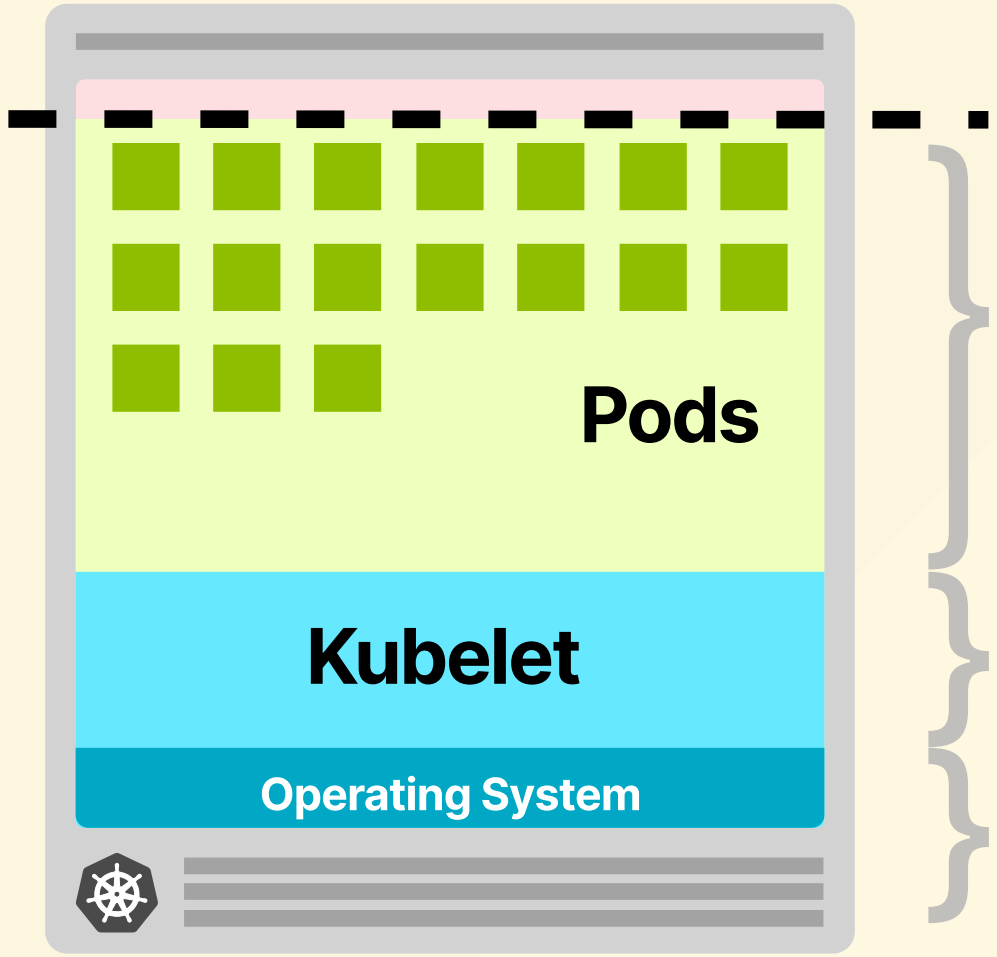


1

2 Memory and CPU left to Pods

3 Memory and CPU reserved to the kubelet

4 Memory and CPU reserved to the OS



1. Eviction threshold
2. Memory and CPU left to Pods
3. Memory and CPU reserved to the kubelet
4. Memory and CPU reserved to the OS

Kubelet reserved memory

+ 255MiB

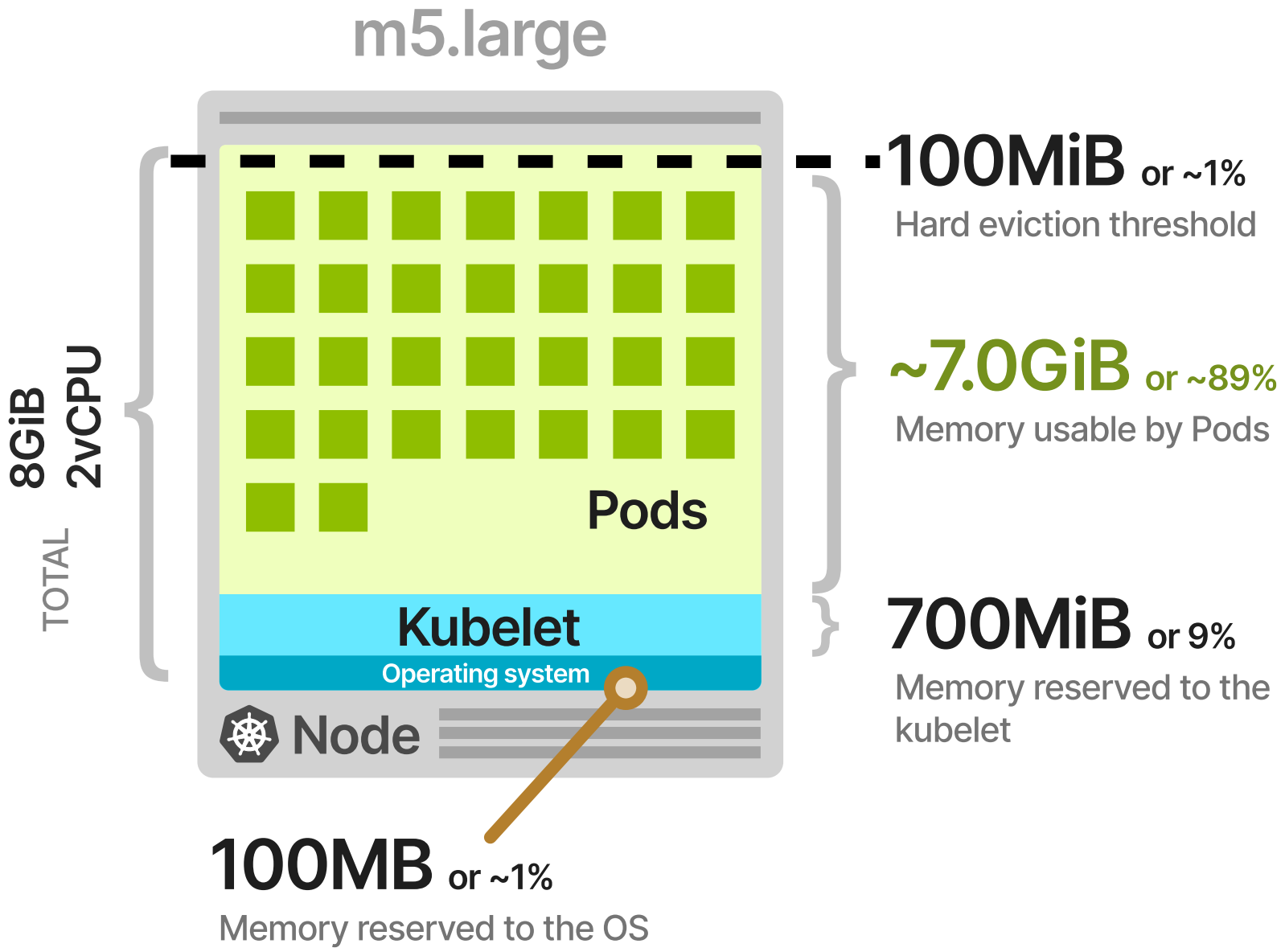
Kubelet reserved memory

+ 255MiB
+ 11MiB * MAX_PODS

Kubelet reserved memory

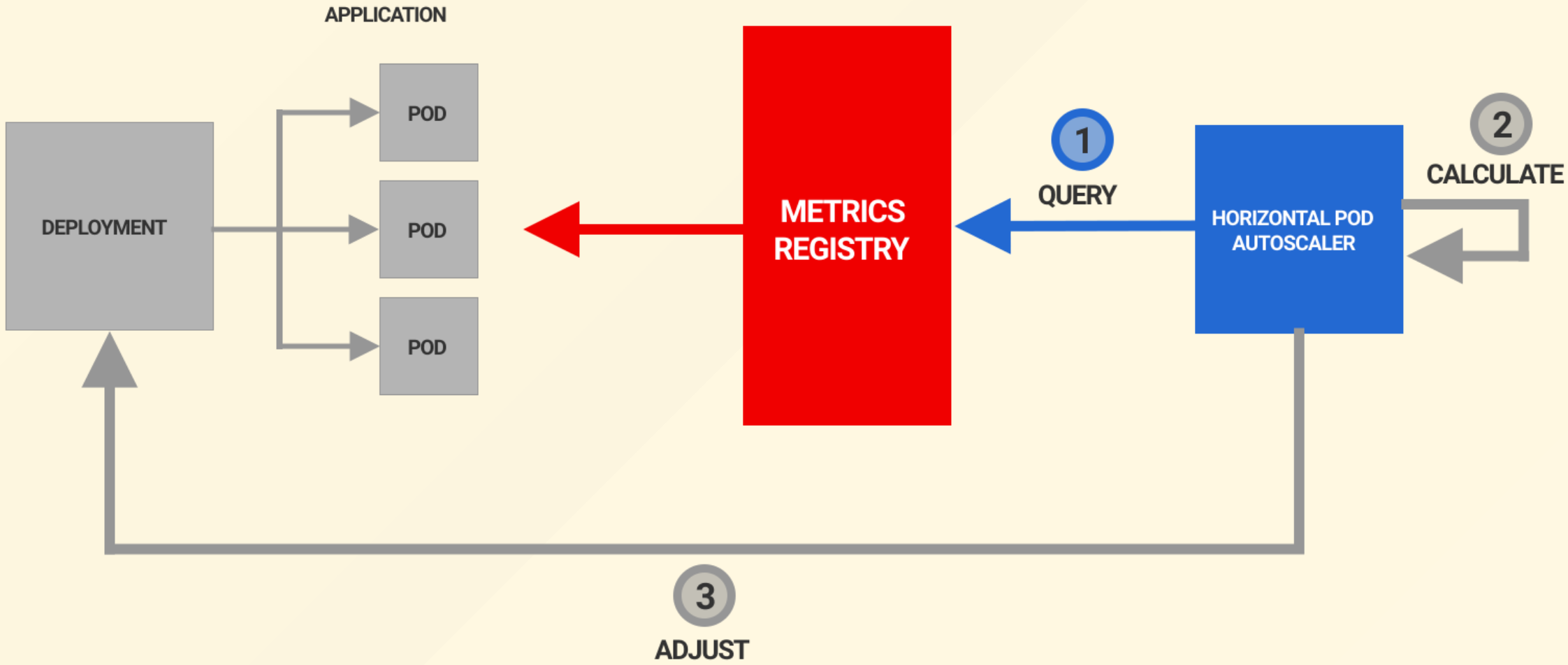
+ 255MiB
+ 11MiB * MAX_PODS

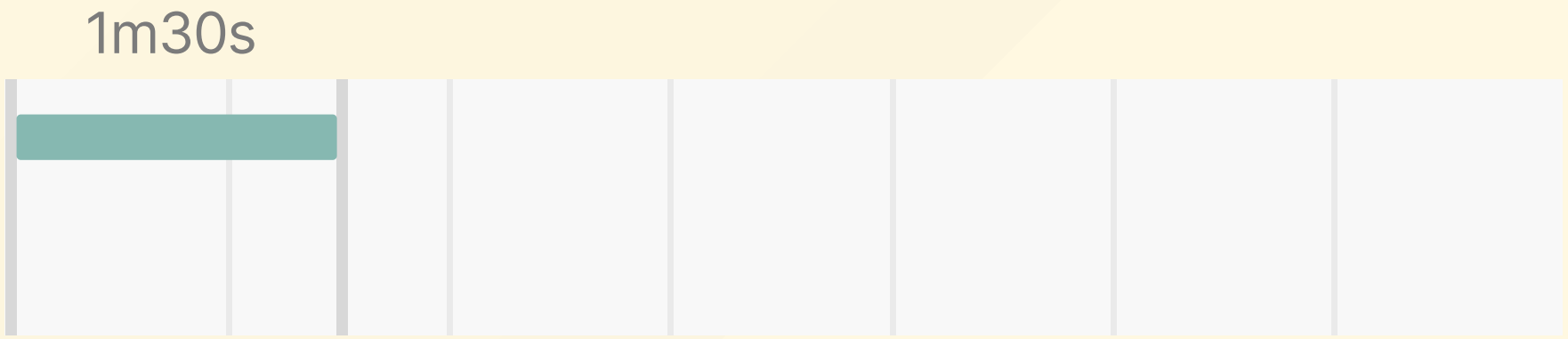
= Reserved memory

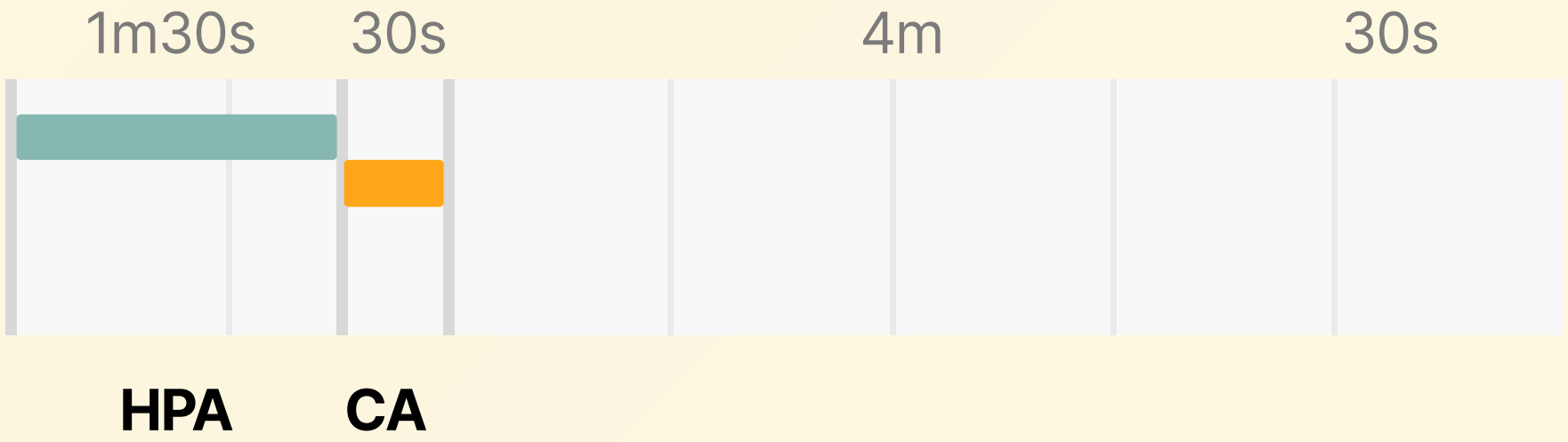

m5.large
8GiB
2vCPU
TOTAL
Pods
Kubelet
Operating system
Node
100MiB or ~1%
Hard eviction threshold
~7.0GiB or ~89%
Memory usable by Pods
700MiB or 9%
Memory reserved to the kubelet
100MB or ~1%
Memory reserved to the OS

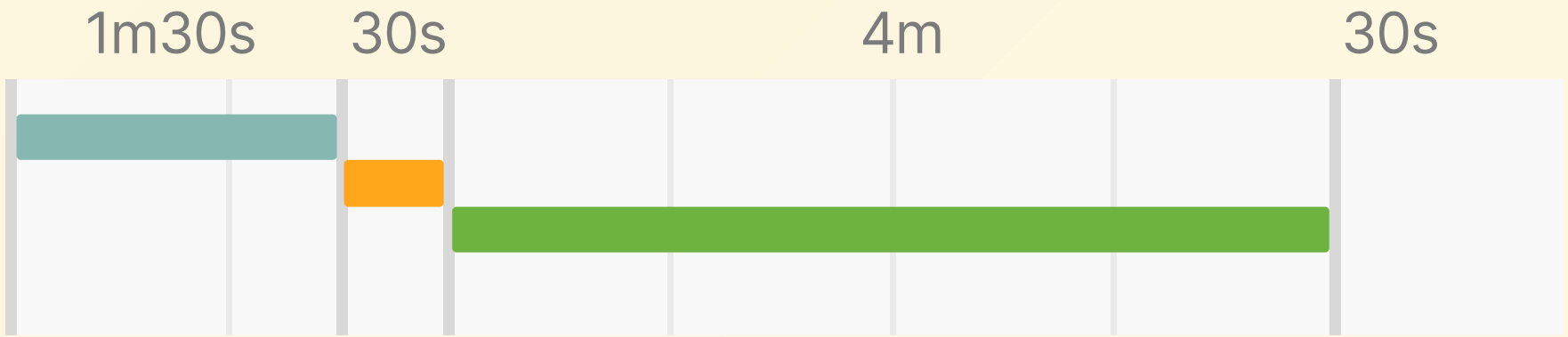
Cluster AutoScaler Lead Time

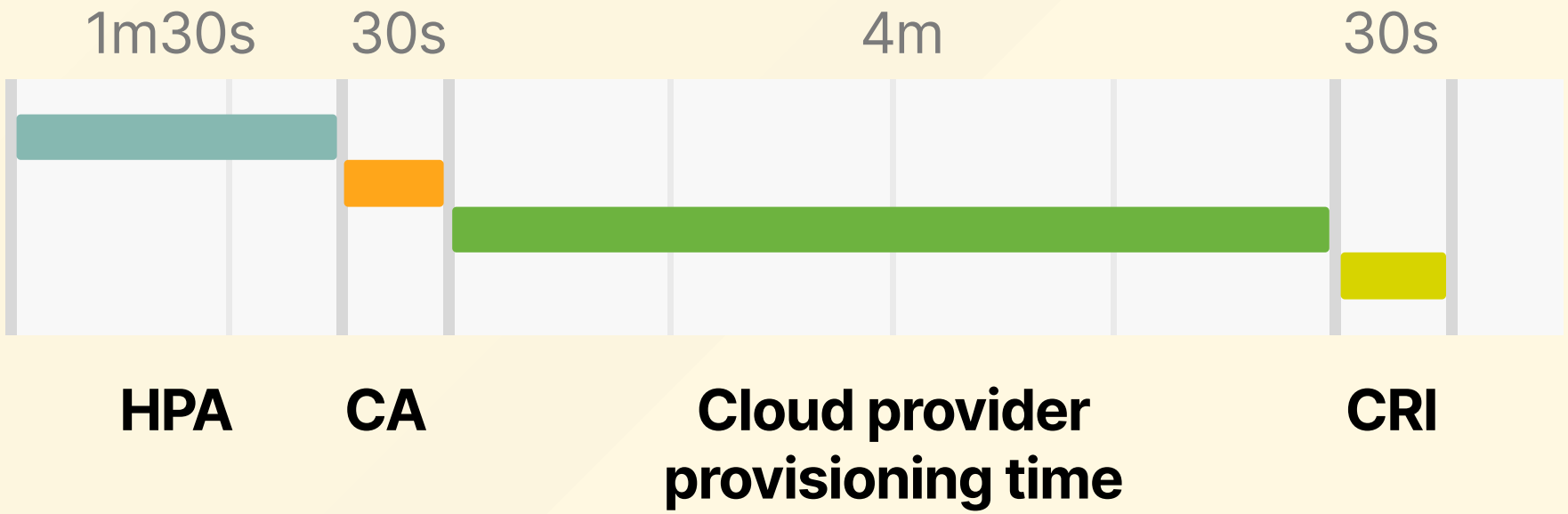
CLUSTER


APPLICATION
POD
POD
POD
DEPLOYMENT
METRICS REGISTRY
1
QUERY
HORIZONTAL POD AUTOSCALER
2
CALCULATE
3
ADJUST

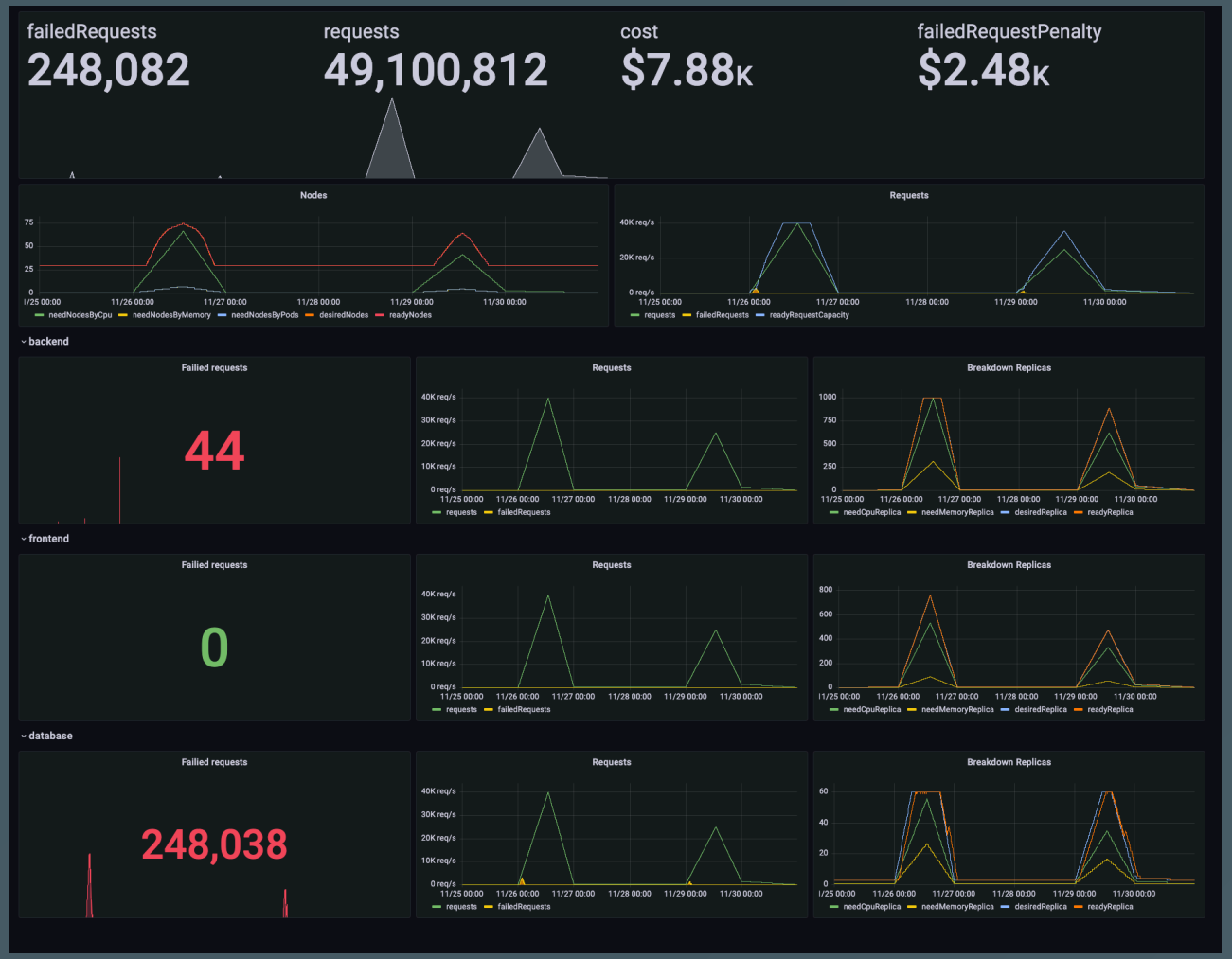

1m30s
HPA


1m30s
30s
4m
30s
HPA
CA


1m30s
30s
4m
30s
HPA
CA
Cloud provider
provisioning time


1m30s
30s
4m
30s
HPA
CA
Cloud provider provisioning time
CRI

github.com/appvia/cloud-spend-forecaster



Strategies for faster scaling

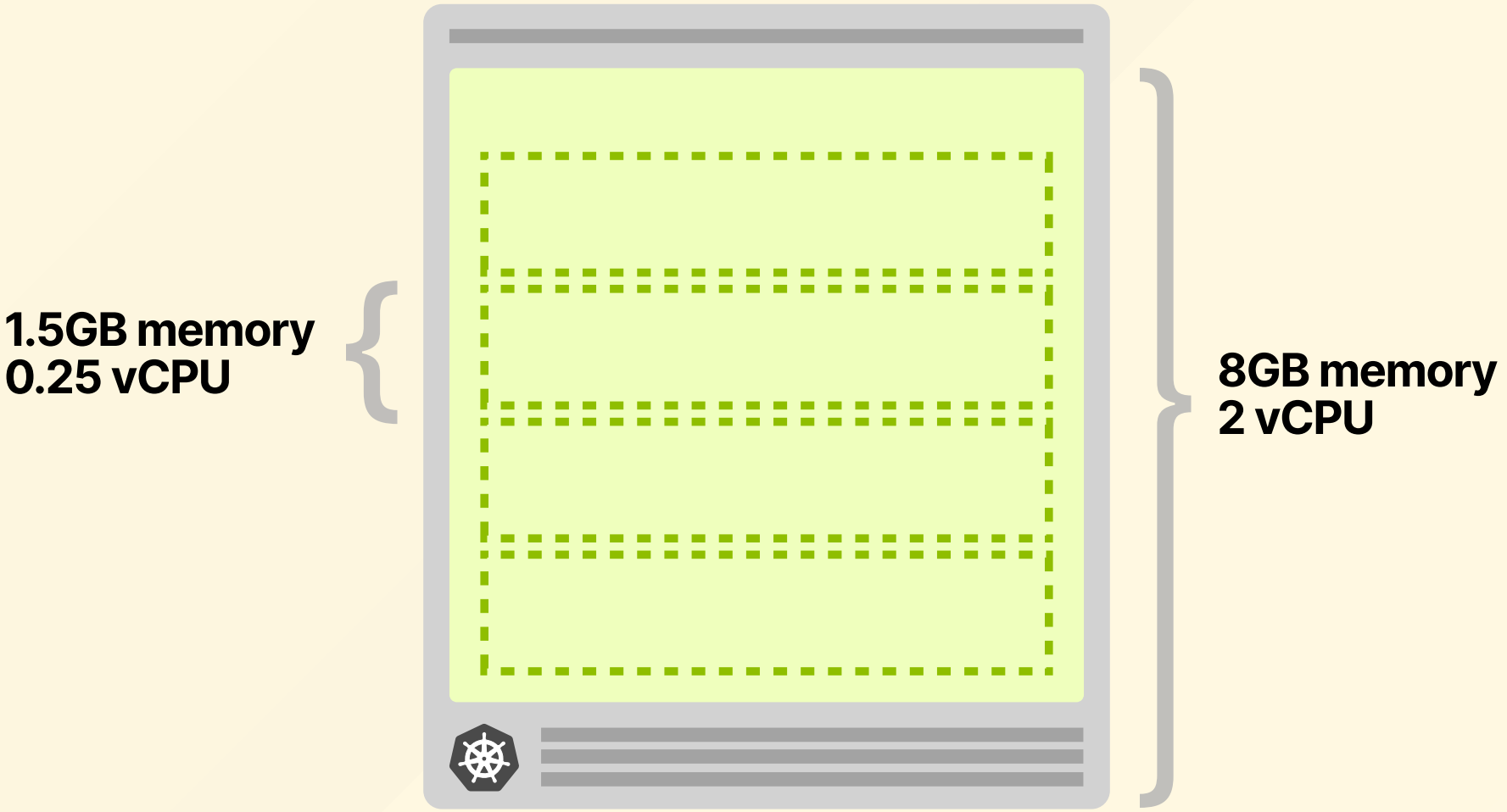
Strategies for faster scaling

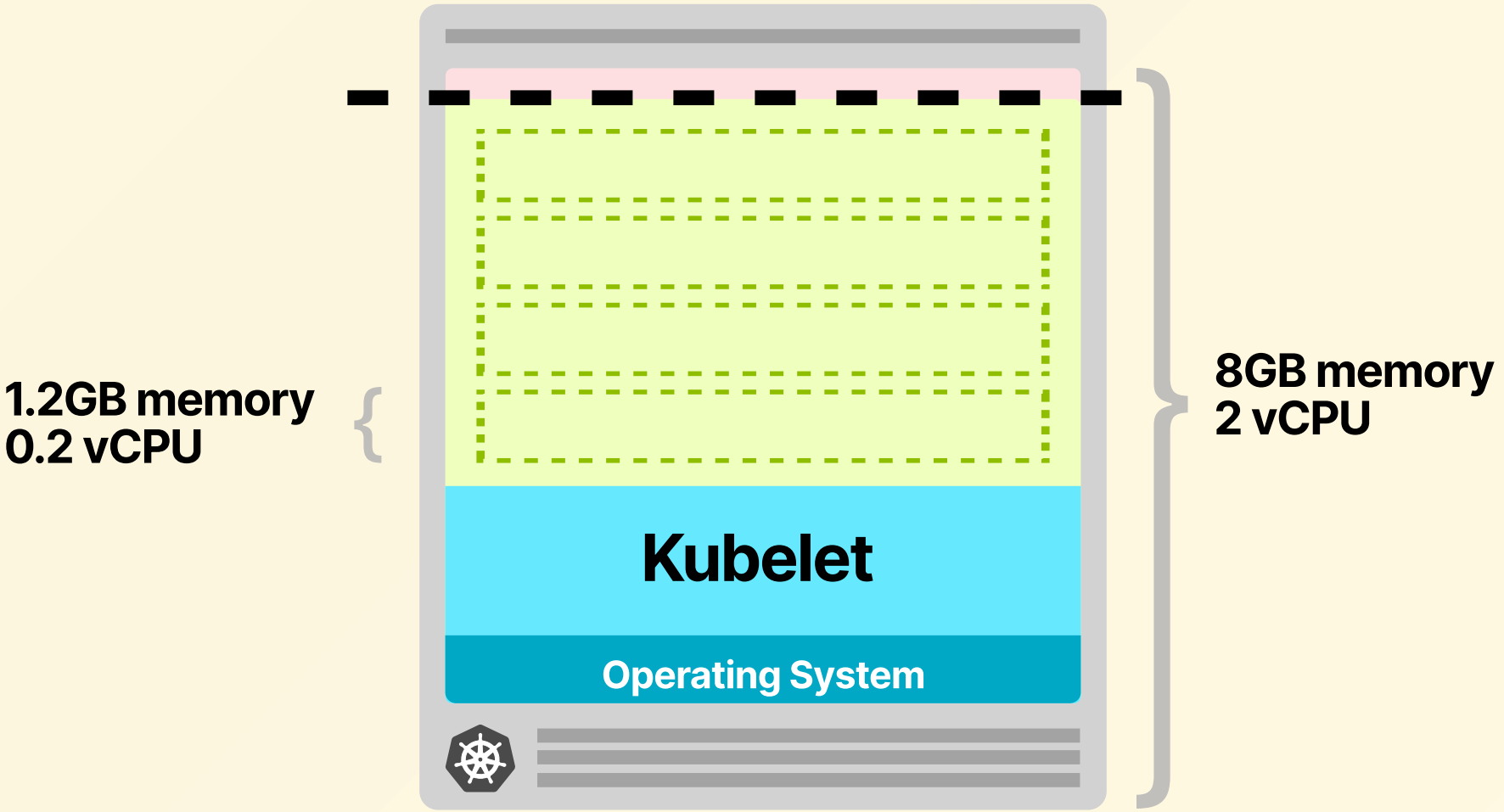
1. **Don't scale!**
2. (Pre) scale

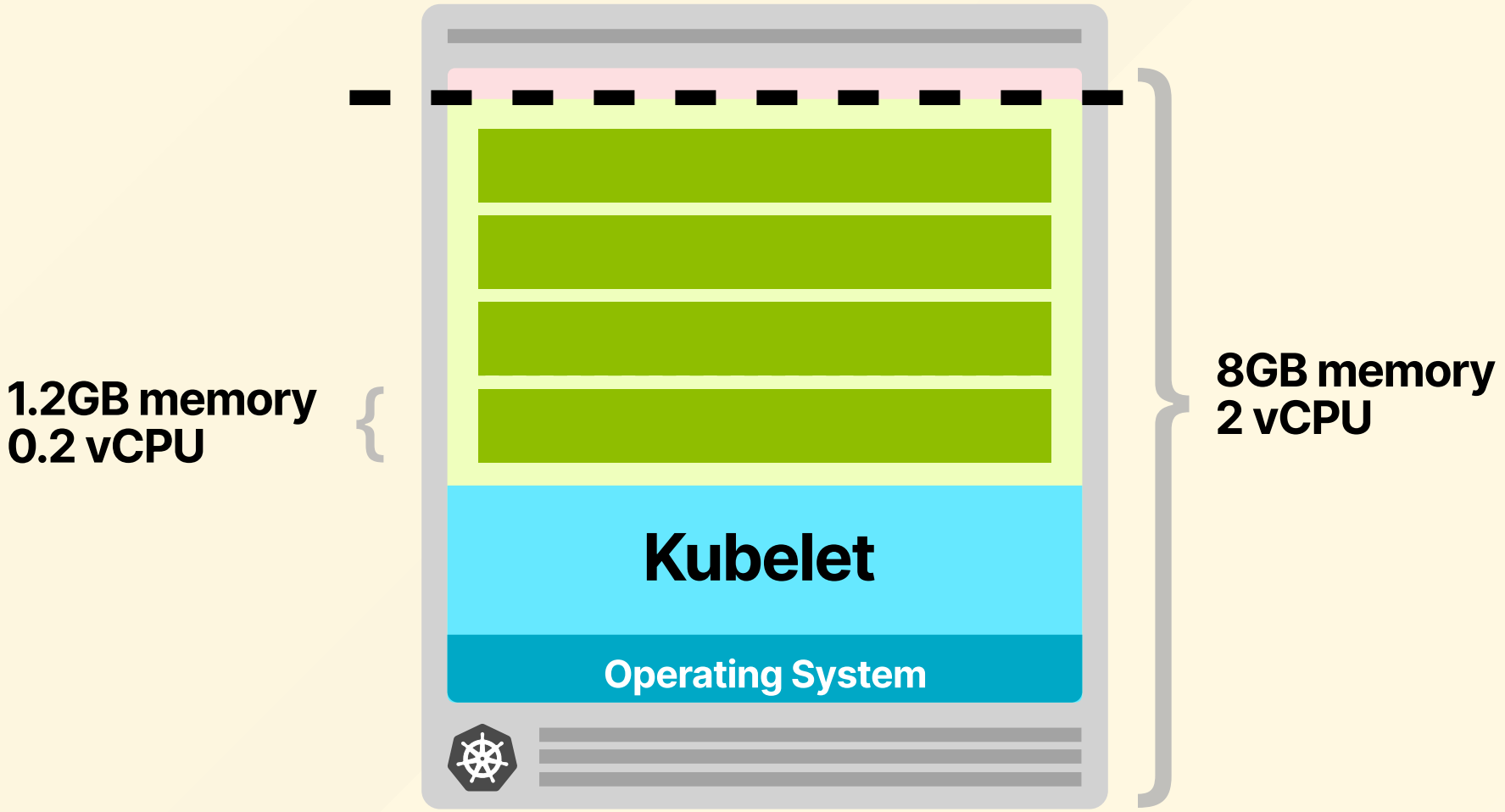
Strategies for faster scaling

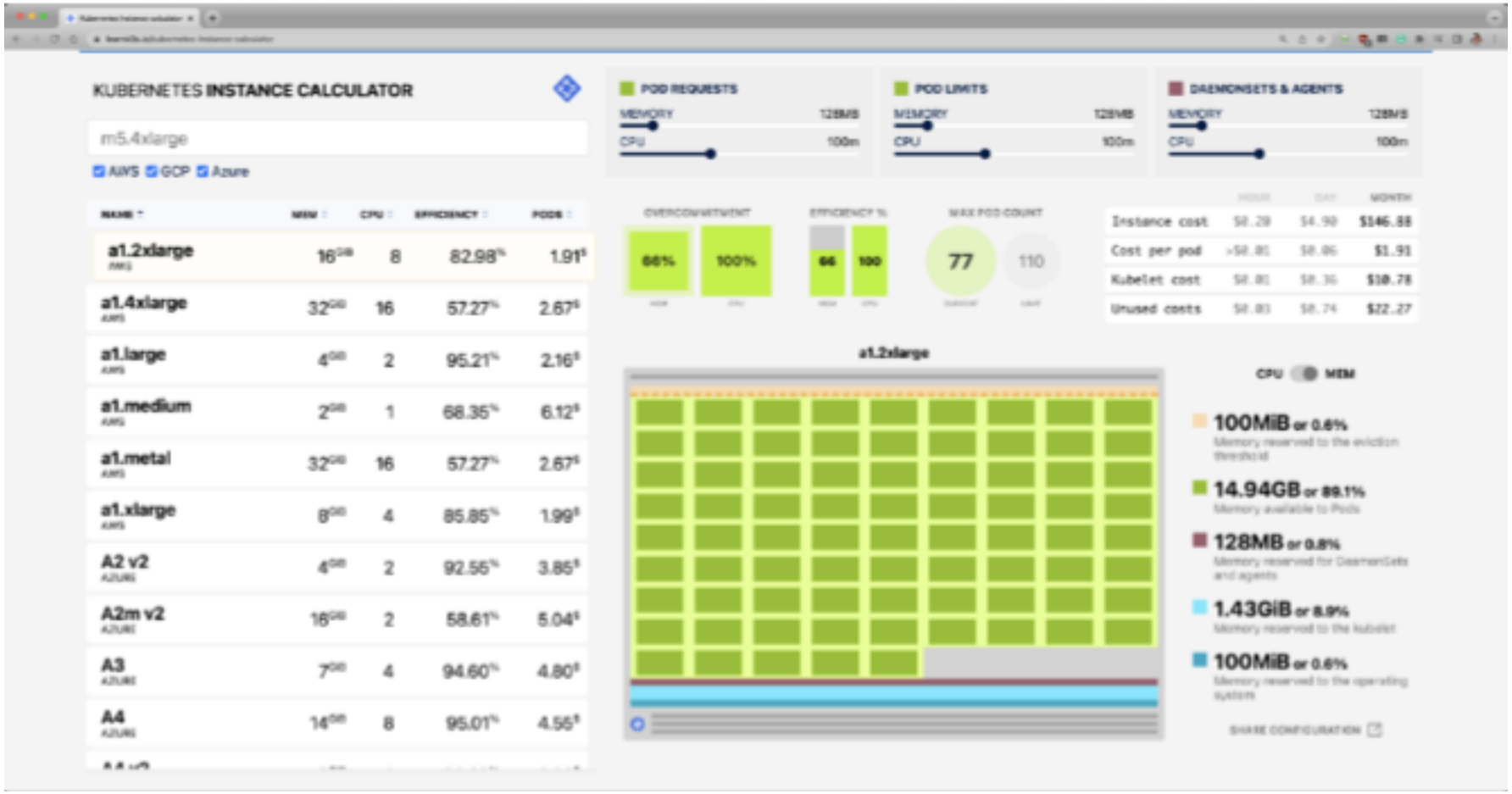
1. Don't scale!
2. **(Pre) scale**

Don't Scale!


1.5GB memory
0.25 vCPU
8GB memory
2 vCPU


1.2GB memory
0.2 vCPU
Kubelet
Operating System
8GB memory
2 vCPU


1.2GB memory
0.2 vCPU
Kubelet
Operating System
8GB memory
2 vCPU



learnk8s.io/kubernetes-instance-calculator

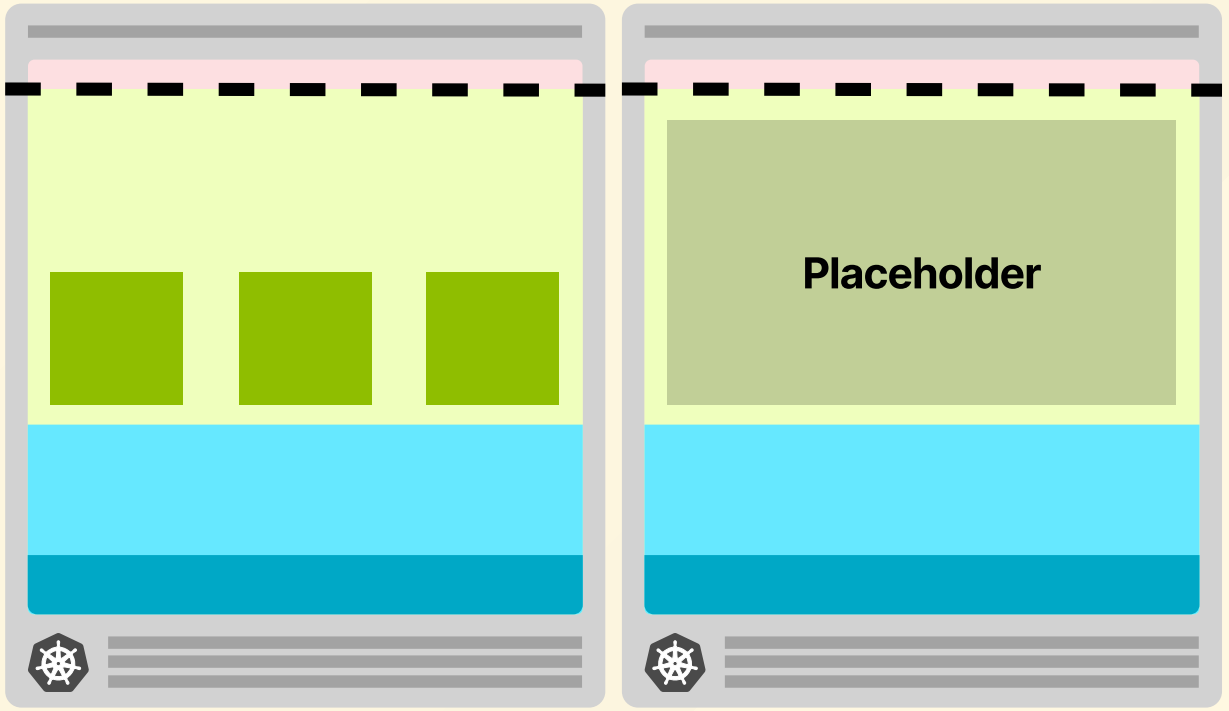
KUBERNETES **INSTANCE CALCULATOR**

m5.4xlarge

AWS GCP Azure

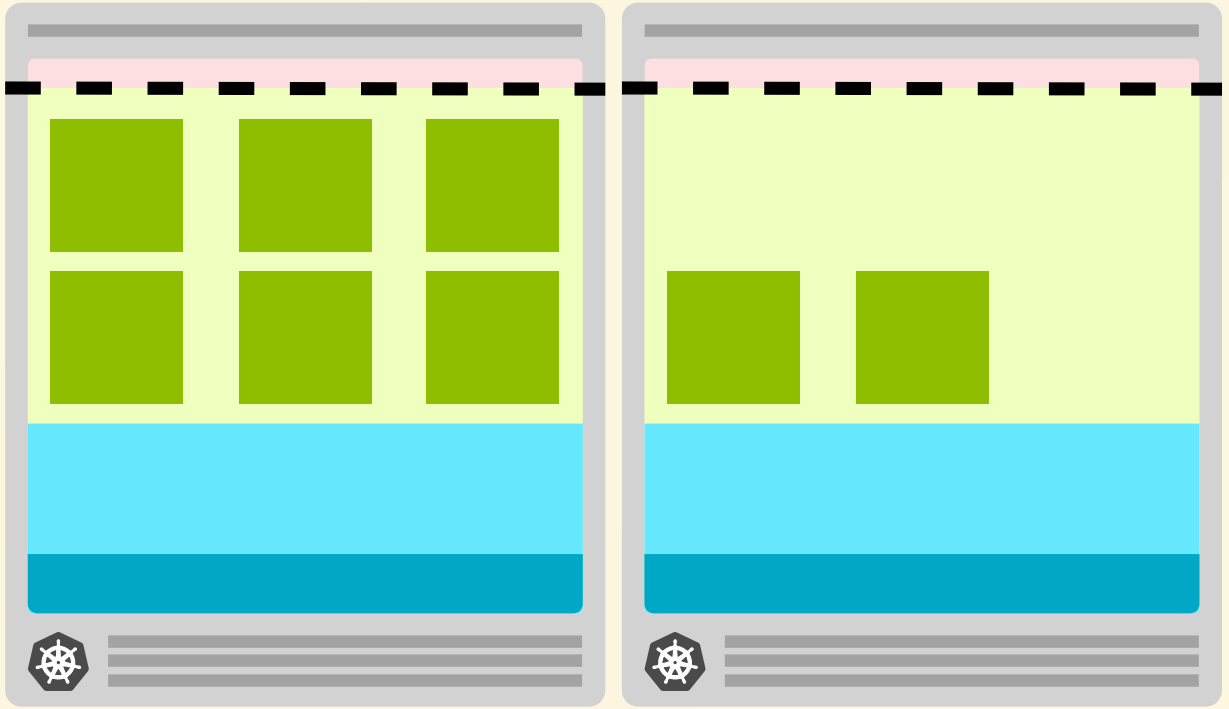
NAME	MEM	CPU
a1.2xlarge AWS	16GiB	8
a1.4xlarge AWS	32GiB	16
a1.large AWS	4GiB	2
a1.medium AWS	2GiB	1
a1.metal AWS	32GiB	16
a1.xlarge AWS	8GiB	4
A2 v2 AZURE	4GiB	2
A2m v2 AZURE	16GiB	2
A3 AZURE	7GiB	4

Proactive scaling

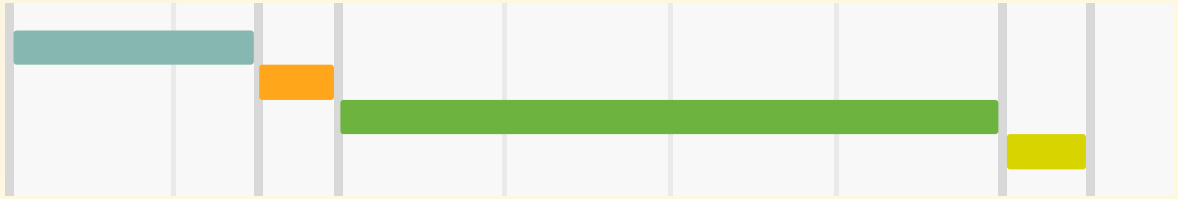

Placeholder

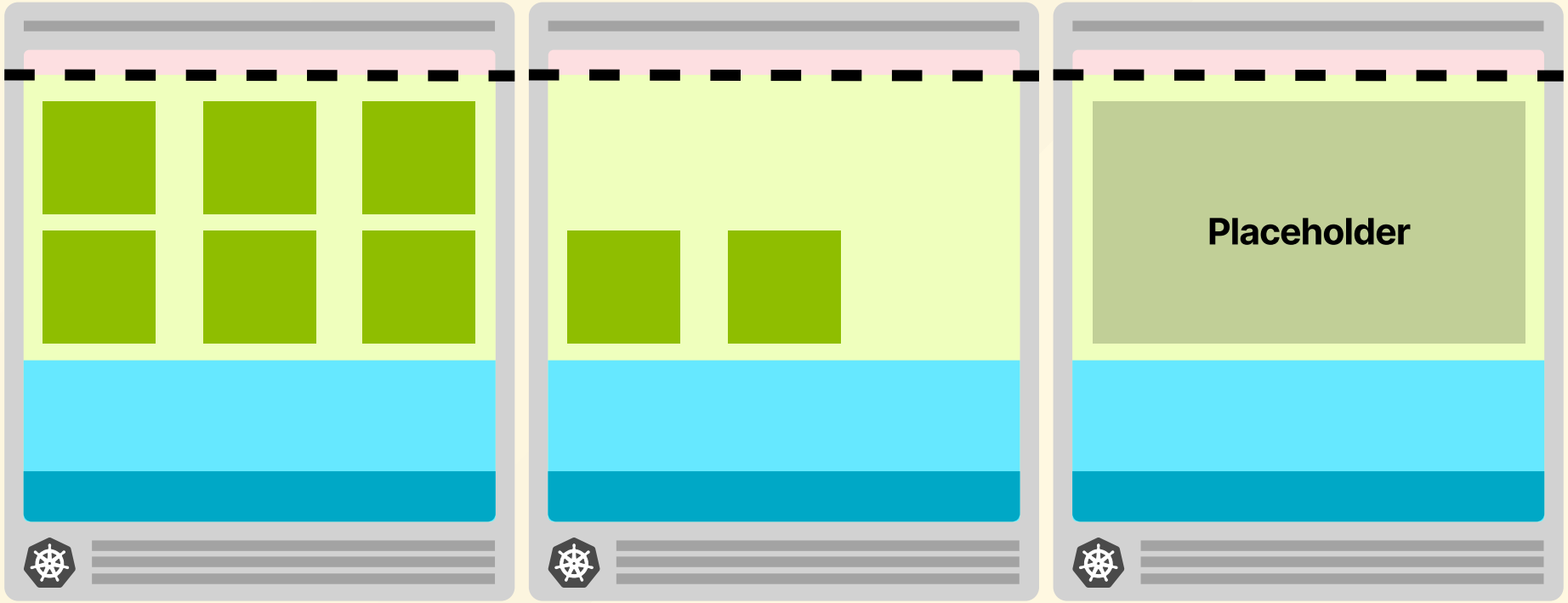
Placeholder

PENDING



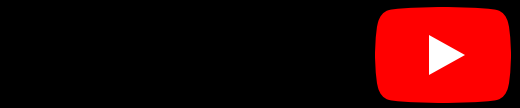
Provisioning a node in the background

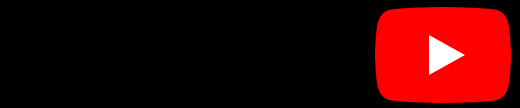



Placeholder

LIVE DEMO

```
default       0s          Warning   FailedScheduling          pod/podinfo-8558cfcd5d-5x2v7                0/2 nodes are available: 1 Insufficient memory, 1 node(s) didn't match Pod's node affinity/selector.
default       1s          Normal    NodeHasSufficientMemory   node/lke74013-115226-6332129f84dd           Node lke74013-115226-6332129f84dd status is now: NodeHasSufficientMemory
default       1s          Normal    NodeHasNoDiskPressure     node/lke74013-115226-6332129f84dd           Node lke74013-115226-6332129f84dd status is now: NodeHasNoDiskPressure
kube-system   0s          Normal    SuccessfulCreate          daemonset/calico-node                       Created pod: calico-node-tpxpn
kube-system   0s          Normal    Scheduled                 pod/csi-linode-node-fvk46                   Successfully assigned kube-system/csi-linode-node-fvk46 to lke74013-115226-6332129f84dd
kube-system   0s          Normal    Scheduled                 pod/calico-node-tpxpn                       Successfully assigned kube-system/calico-node-tpxpn to lke74013-115226-6332129f84dd
kube-system   0s          Normal    SuccessfulCreate          daemonset/csi-linode-node                   Created pod: csi-linode-node-fvk46
kube-system   0s          Normal    Scheduled                 pod/kube-proxy-ds24f                        Successfully assigned kube-system/kube-proxy-ds24f to lke74013-115226-6332129f84dd
kube-system   0s          Normal    SuccessfulCreate          daemonset/kube-proxy                        Created pod: kube-proxy-ds24f
default       0s          Normal    Synced                    node/lke74013-115226-6332129f84dd           Node synced successfully
default       0s          Normal    Starting                  node/lke74013-115226-6332129f84dd           Starting kubelet.
default       0s          Normal    NodeAllocatableEnforced   node/lke74013-115226-6332129f84dd           Updated Node Allocatable limit across pods
default       0s          Normal    NodeHasSufficientMemory   node/lke74013-115226-6332129f84dd           Node lke74013-115226-6332129f84dd status is now: NodeHasSufficientMemory
default       0s          Normal    NodeHasNoDiskPressure     node/lke74013-115226-6332129f84dd           Node lke74013-115226-6332129f84dd status is now: NodeHasNoDiskPressure
default       0s          Normal    NodeHasSufficientPID      node/lke74013-115226-6332129f84dd           Node lke74013-115226-6332129f84dd status is now: NodeHasSufficientPID
default       0s          Normal    RegisteredNode            node/lke74013-115226-6332129f84dd           Node lke74013-115226-6332129f84dd event: Registered Node lke74013-115226-6332129f84dd in Controller
kube-system   1s          Normal    Pulling                   pod/kube-proxy-ds24f                        Pulling image "linode/kube-proxy-amd64:v1.23.10"
kube-system   1s          Normal    Pulling                   pod/calico-node-tpxpn                       Pulling image "docker.io/calico/cni:v3.22.1"
kube-system   0s          Normal    Pulling                   pod/csi-linode-node-fvk46                   Pulling image "bitnami/kubectl:1.16.3-debian-10-r36"
kube-system   0s          Normal    Pulled                    pod/kube-proxy-ds24f                        Successfully pulled image "linode/kube-proxy-amd64:v1.23.10" in 3.544455967s
kube-system   0s          Normal    Created                   pod/kube-proxy-ds24f                        Created container kube-proxy
kube-system   0s          Normal    Started                   pod/kube-proxy-ds24f                        Started container kube-proxy
default       0s          Normal    Starting                  node/lke74013-115226-6332129f84dd
kube-system   0s          Normal    Pulled                    pod/calico-node-tpxpn                       Successfully pulled image "docker.io/calico/cni:v3.22.1" in 8.06147771s
kube-system   1s          Normal    Created                   pod/calico-node-tpxpn                       Created container upgrade-ipam
kube-system   0s          Normal    Started                   pod/calico-node-tpxpn                       Started container upgrade-ipam
kube-system   0s          Normal    Pulled                    pod/calico-node-tpxpn                       Container image "docker.io/calico/cni:v3.22.1" already present on machine
kube-system   0s          Normal    Created                   pod/calico-node-tpxpn                       Created container install-cni
kube-system   0s          Normal    Started                   pod/calico-node-tpxpn                       Started container install-cni
kube-system   0s          Normal    Pulling                   pod/calico-node-tpxpn                       Pulling image "docker.io/calico/pod2daemon-flexvol:v3.22.1"
kube-system   0s          Normal    Pulled                    pod/csi-linode-node-fvk46                   Successfully pulled image "bitnami/kubectl:1.16.3-debian-10-r36" in 11.657565568s
kube-system   0s          Normal    Created                   pod/csi-linode-node-fvk46                   Created container init
kube-system   0s          Normal    Started                   pod/csi-linode-node-fvk46                   Started container init
kube-system   0s          Normal    Pulling                   pod/csi-linode-node-fvk46                   Pulling image "linode/csi-node-driver-registrar:v1.3.0"
default       0s          Warning   FailedScheduling          pod/podinfo-8558cfcd5d-5x2v7                0/2 nodes are available: 1 Insufficient memory, 1 node(s) didn't match Pod's node affinity/selector.
kube-system   0s          Normal    Pulled                    pod/calico-node-tpxpn                       Successfully pulled image "docker.io/calico/pod2daemon-flexvol:v3.22.1" in 5.534993769s
kube-system   0s          Normal    Created                   pod/calico-node-tpxpn                       Created container flexvol-driver
kube-system   0s          Normal    Started                   pod/calico-node-tpxpn                       Started container flexvol-driver
kube-system   0s          Normal    Pulling                   pod/calico-node-tpxpn                       Pulling image "docker.io/calico/node:v3.22.1"
default       0s          Normal    NodeReady                 node/lke74013-115226-6332129f84dd           Node lke74013-115226-6332129f84dd status is now: NodeReady
kube-system   0s          Normal    Pulled                    pod/csi-linode-node-fvk46                   Successfully pulled image "linode/csi-node-driver-registrar:v1.3.0" in 4.741448054s
kube-system   0s          Normal    Created                   pod/csi-linode-node-fvk46                   Created container csi-node-driver-registrar
kube-system   0s          Normal    Started                   pod/csi-linode-node-fvk46                   Started container csi-node-driver-registrar
kube-system   0s          Normal    Pulling                   pod/csi-linode-node-fvk46                   Pulling image "linode/linode-blockstorage-csi-driver:v0.5.0"
default       0s          Normal    Scheduled                 pod/podinfo-8558cfcd5d-5x2v7                Successfully assigned default/podinfo-8558cfcd5d-5x2v7 to lke74013-115226-6332129f84dd
default       0s          Warning   FailedCreatePodSandBox    pod/podinfo-8558cfcd5d-5x2v7                Failed to create pod sandbox: rpc error: code = Unknown desc = failed to set up sandbox container "4d21d0493b5ac49eb7a173b9946fb379f9e0d98b1d55ab13547ad0826842bc6d" network for pod "podinfo-8558cfcd5d-5x2v7": networkPl
default       0s          Normal    TaintManagerEviction      pod/podinfo-8558cfcd5d-5x2v7                Cancelling deletion of Pod default/podinfo-8558cfcd5d-5x2v7
default       0s          Normal    SandboxChanged            pod/podinfo-8558cfcd5d-5x2v7                Pod sandbox changed, it will be killed and re-created.
default       0s          Warning   FailedCreatePodSandBox    pod/podinfo-8558cfcd5d-5x2v7                Failed to create pod sandbox: rpc error: code = Unknown desc = failed to set up sandbox container "5d02b7c46e44913d4fabb9e02dc77044a3163f976c0a46f47c398108ac3fa000" network for pod "podinfo-8558cfcd5d-5x2v7": networkPl
default       0s          Normal    SandboxChanged            pod/podinfo-8558cfcd5d-5x2v7                Pod sandbox changed, it will be killed and re-created.
default       0s          Warning   FailedCreatePodSandBox    pod/podinfo-8558cfcd5d-5x2v7                Failed to create pod sandbox: rpc error: code = Unknown desc = failed to set up sandbox container "3447a34aeda5501828a9c476e0d6d7077eb7f895c8fbcd7022af1643865241c6" network for pod "podinfo-8558cfcd5d-5x2v7": networkPl
default       0s          Normal    SandboxChanged            pod/podinfo-8558cfcd5d-5x2v7                Pod sandbox changed, it will be killed and re-created.
default       0s          Warning   FailedCreatePodSandBox    pod/podinfo-8558cfcd5d-5x2v7                Failed to create pod sandbox: rpc error: code = Unknown desc = failed to set up sandbox container "77a8281777afde238f38cddc9141ed2e52c78540ccb4ad7b296da3fc8c20a6ef" network for pod "podinfo-8558cfcd5d-5x2v7": networkPl
default       0s          Normal    SandboxChanged            pod/podinfo-8558cfcd5d-5x2v7                Pod sandbox changed, it will be killed and re-created.
kube-system   0s          Normal    Pulled                    pod/calico-node-tpxpn                       Successfully pulled image "docker.io/calico/node:v3.22.1" in 6.353089879s
kube-system   0s          Normal    Created                   pod/calico-node-tpxpn                       Created container calico-node
kube-system   0s          Normal    Started                   pod/calico-node-tpxpn                       Started container calico-node
default       0s          Warning   FailedCreatePodSandBox    pod/podinfo-8558cfcd5d-5x2v7                Failed to create pod sandbox: rpc error: code = Unknown desc = failed to set up sandbox container "0fc48fe077b786da531ebc656016468d3ae3f41e6cf520e27b638dbfbab580e2" network for pod "podinfo-8558cfcd5d-5x2v7": networkPl
default       0s          Normal    SandboxChanged            pod/podinfo-8558cfcd5d-5x2v7                Pod sandbox changed, it will be killed and re-created.
kube-system   0s          Warning   Unhealthy                 pod/calico-node-tpxpn                       Readiness probe failed: calico/node is not ready: BIRD is not ready: Error querying BIRD: unable to connect to BIRDv4 socket: dial unix /var/run/bird/bird.ctl: connect: no such file or directory
default       0s          Normal    Pulling                   pod/podinfo-8558cfcd5d-5x2v7                Pulling image "stefanprodan/podinfo"
kube-system   0s          Warning   Unhealthy                 pod/calico-node-tpxpn                       Readiness probe failed: calico/node is not ready: felix is not ready: readiness probe reporting 503
kube-system   0s          Normal    Pulled                    pod/csi-linode-node-fvk46                   Successfully pulled image "linode/linode-blockstorage-csi-driver:v0.5.0" in 6.754070918s
kube-system   0s          Normal    Created                   pod/csi-linode-node-fvk46                   Created container csi-linode-plugin
kube-system   0s          Normal    Started                   pod/csi-linode-node-fvk46                   Started container csi-linode-plugin
kube-system   0s          Warning   Unhealthy                 pod/calico-node-tpxpn                       Readiness probe failed: calico/node is not ready: BIRD is not ready: Error querying BIRD: unable to connect to BIRDv4 socket: dial unix /var/run/calico/bird.ctl: connect: connection refused
kube-system   0s          Warning   Unhealthy                 pod/calico-node-tpxpn                       Readiness probe failed: 2022-09-26 21:01:13.334 [INFO][249] confd/health.go 180: Number of node(s) with BGP peering established = 1...
default       0s          Normal    Pulled                    pod/podinfo-8558cfcd5d-5x2v7                Successfully pulled image "stefanprodan/podinfo" in 4.44920297s
default       0s          Normal    Created                   pod/podinfo-8558cfcd5d-5x2v7                Created container podinfo
default       0s          Normal    Started                   pod/podinfo-8558cfcd5d-5x2v7                Started container podinfo
default       0s          Normal    ScalingReplicaSet         deployment/podinfo                          Scaled down replica set podinfo-8558cfcd5d to 1
default       0s          Normal    Killing                   pod/podinfo-8558cfcd5d-d5khs                Stopping container podinfo
default       0s          Normal    Killing                   pod/podinfo-8558cfcd5d-z9hvq                Stopping container podinfo
default       0s          Normal    Killing                   pod/podinfo-8558cfcd5d-rxbmc                Stopping container podinfo
default       0s          Normal    Killing                   pod/podinfo-8558cfcd5d-h4c5j                Stopping container podinfo
default       0s          Normal    ScalingReplicaSet         deployment/overprovisioning                 Scaled up replica set overprovisioning-7d874998c to 1
default       0s          Normal    Scheduled                 pod/overprovisioning-7d874998c-rgxf9        Successfully assigned default/overprovisioning-7d874998c-rgxf9 to lke74013-115226-6331fbd57928
default       0s          Normal    SuccessfulCreate          replicaset/overprovisioning-7d874998c       Created pod: overprovisioning-7d874998c-rgxf9
default       0s          Normal    Pulling                   pod/overprovisioning-7d874998c-rgxf9        Pulling image "k8s.gcr.io/pause"
default       0s          Normal    Pulled                    pod/overprovisioning-7d874998c-rgxf9        Successfully pulled image "k8s.gcr.io/pause" in 11.25774753s
default       0s          Normal    Created                   pod/overprovisioning-7d874998c-rgxf9        Created container pause
default       0s          Normal    Started                   pod/overprovisioning-7d874998c-rgxf9        Started container pause
default       0s          Normal    ScalingReplicaSet         deployment/podinfo                          Scaled up replica set podinfo-8558cfcd5d to 5
default       0s          Normal    Scheduled                 pod/podinfo-8558cfcd5d-8m87n                Successfully assigned default/podinfo-8558cfcd5d-8m87n to lke74013-115226-6332129f84dd
default       0s          Normal    Scheduled                 pod/podinfo-8558cfcd5d-hlrd8                Successfully assigned default/podinfo-8558cfcd5d-hlrd8 to lke74013-115226-6332129f84dd
default       0s          Normal    Scheduled                 pod/podinfo-8558cfcd5d-4mvj4                Successfully assigned default/podinfo-8558cfcd5d-4mvj4 to lke74013-115226-6332129f84dd
default       0s          Warning   FailedScheduling          pod/podinfo-8558cfcd5d-mh96r                0/3 nodes are available: 1 node(s) didn't match Pod's node affinity/selector, 2 Insufficient memory.
default       0s          Normal    Preempted                 pod/overprovisioning-7d874998c-rgxf9        Preempted by default/podinfo-8558cfcd5d-mh96r on node lke74013-115226-6331fbd57928
default       0s          Normal    Killing                   pod/overprovisioning-7d874998c-rgxf9        Stopping container pause
```

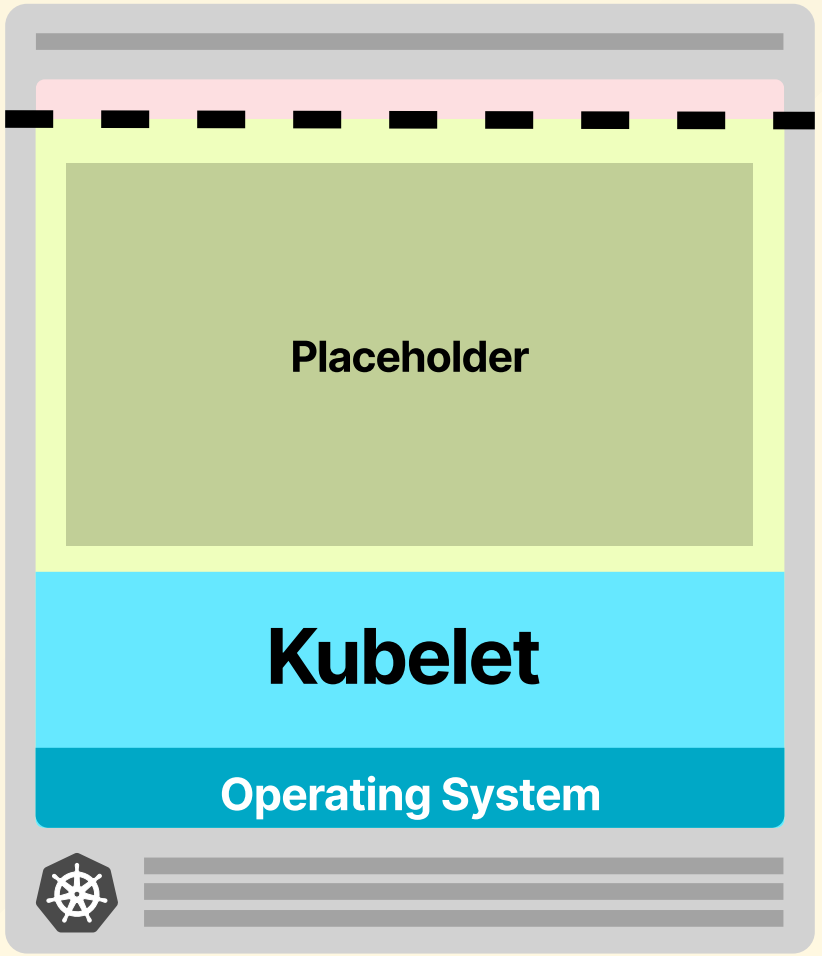




Designing the right placeholder

LESS THAN

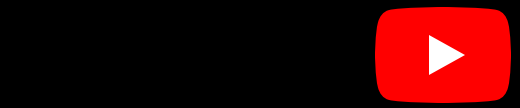
5.9GB memory
1.73 vCPU

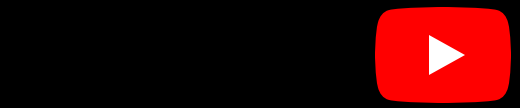


8GB memory
2 vCPU

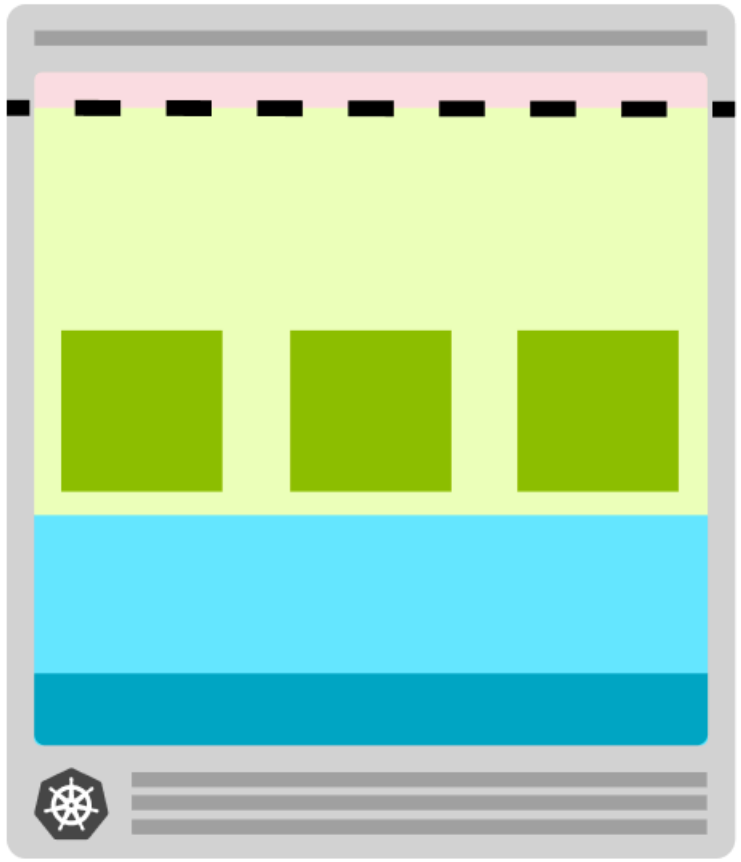
```yaml
apiVersion: scheduling.k8s.io/v1
kind: PriorityClass
metadata:
  name: overprovisioning
value: -1
globalDefault: false
```

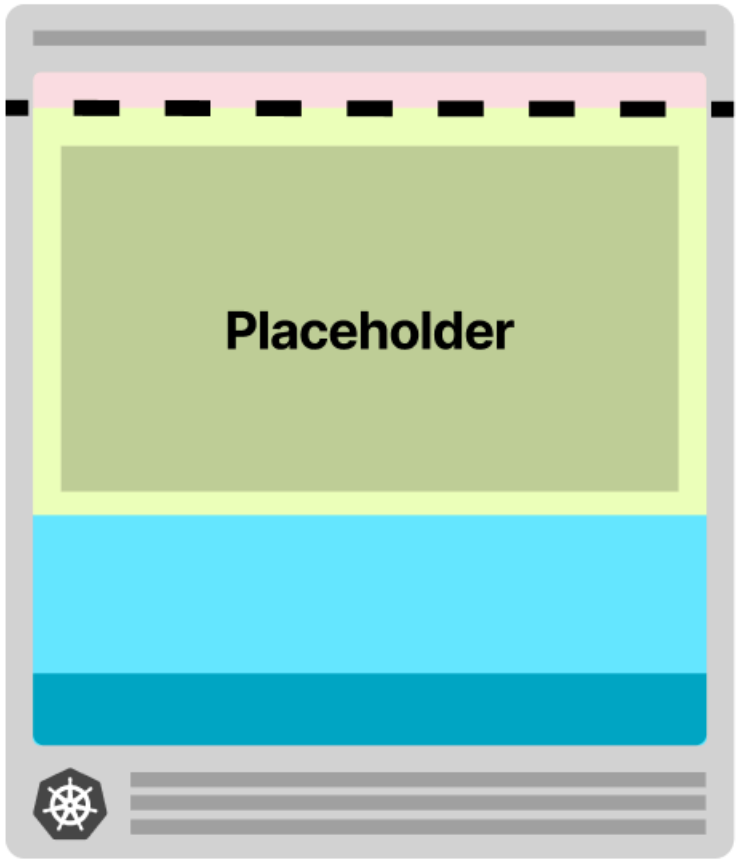
DEMO



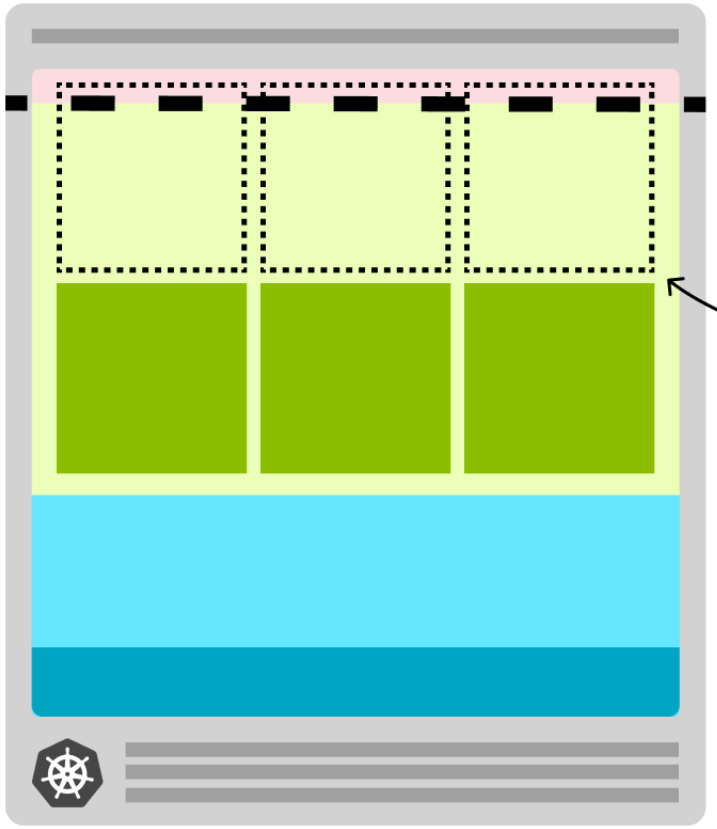


you are still billed for this node even if it does nothing



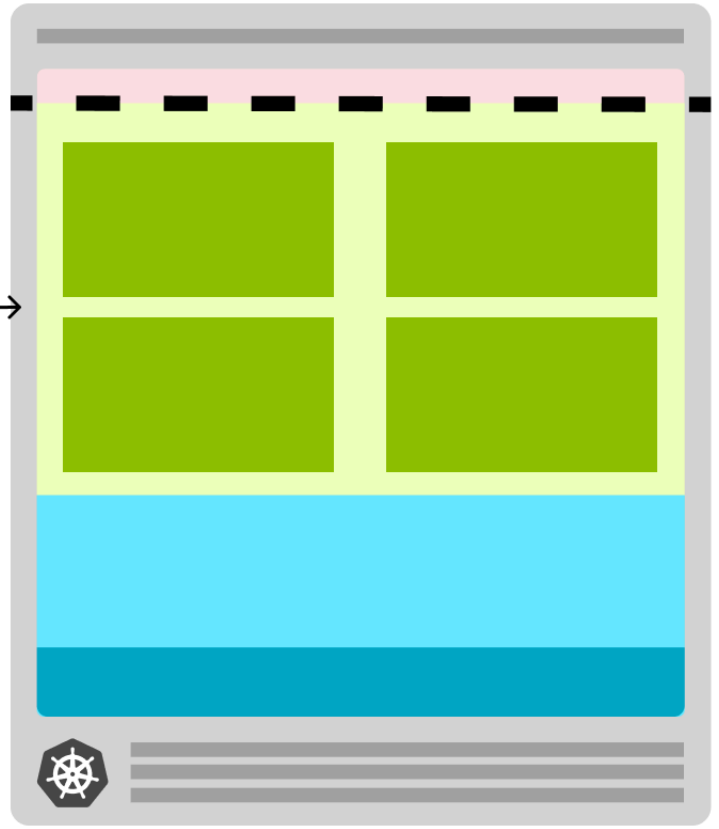


1



can't fit more pods.
memory and CPU
are not utilized in full

pods utilize all available
resources



2

```
apiVersion: scheduling.k8s.io/v1
kind: PriorityClass
```

- cns.me
- talks.cns.me
- github.com/chrisns
- learnk8s.io

Q&A

cns.me

Chris Nesbitt-Smith

instance-calculator