Cloud_Native
Rejekts [EU'19]

Knowing what your Kubernetes cluster is doing

Federico Hernandez

Sr. Software Engineer

Meltwater (Gothenburg, SE)

@recollir

Cloud_Native
Rejekts [EU'19]

Knowing what your Kubernetes cluster is doing

Federico Hernandez

~~Sr. Software~~ **YAML** Engineer

Meltwater (Gothenburg, SE)

@recollir

Meltwater
Outside Insight

30000+ customers

1500+ employees

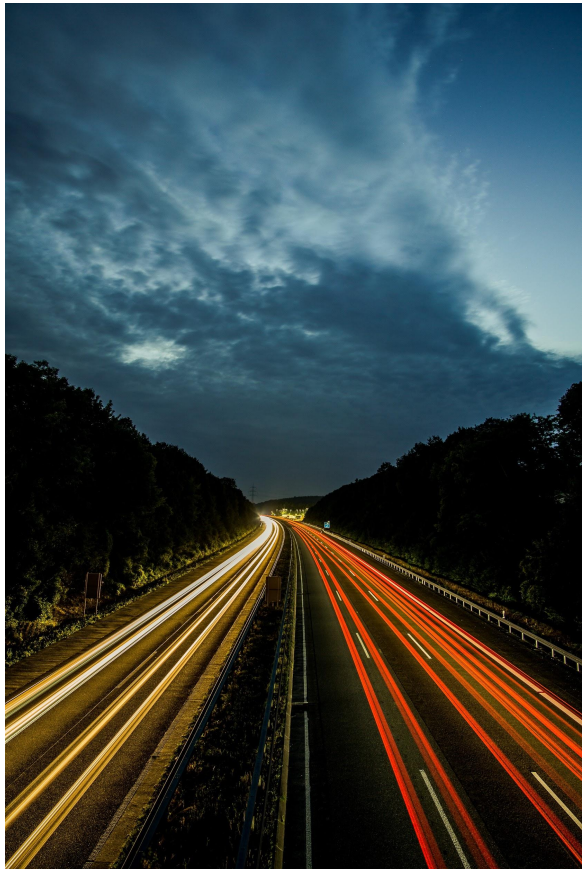
55 global offices

Founded 2001 in Oslo

9+ engineering offices (7+ countries)

330+ engineers in 50+ teams

Kubernetes in prod since early 2018



Foundation Kubernetes Service

kops

AWS

dex

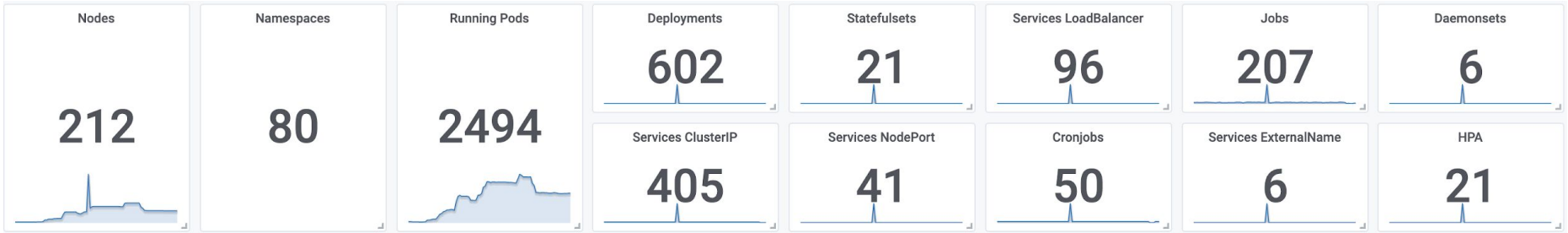
CNI

EXTERNAL DNS FOR kubernetes

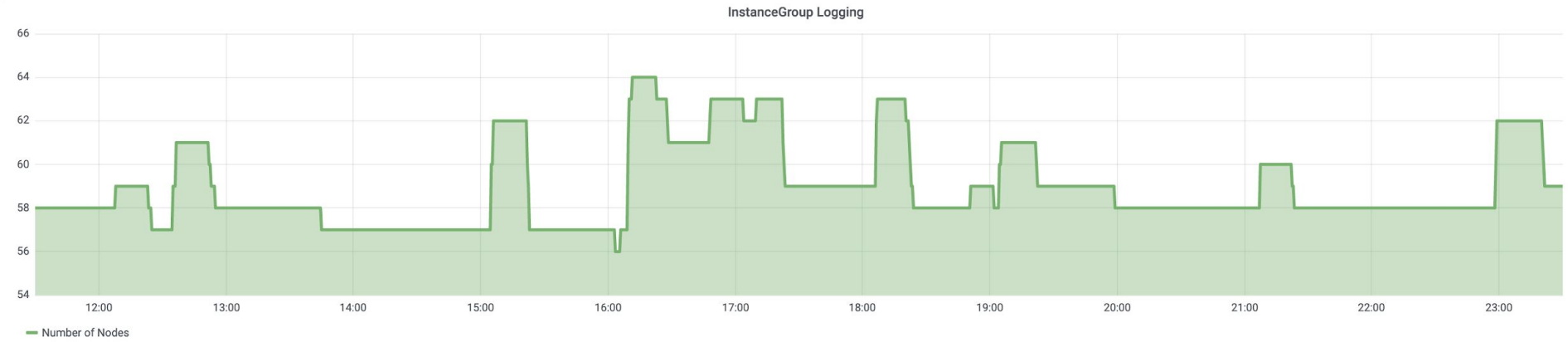
Thanos


Nodes
149
Namespaces
77
Running Pods
2102
Deployments
561
Statefulsets
21
Services LoadBalancer
93
Jobs
182
Daemonsets
6
Services ClusterIP
394
Services NodePort
41
Cronjobs
50
Services ExternalName
6
HPA
21

@recollir

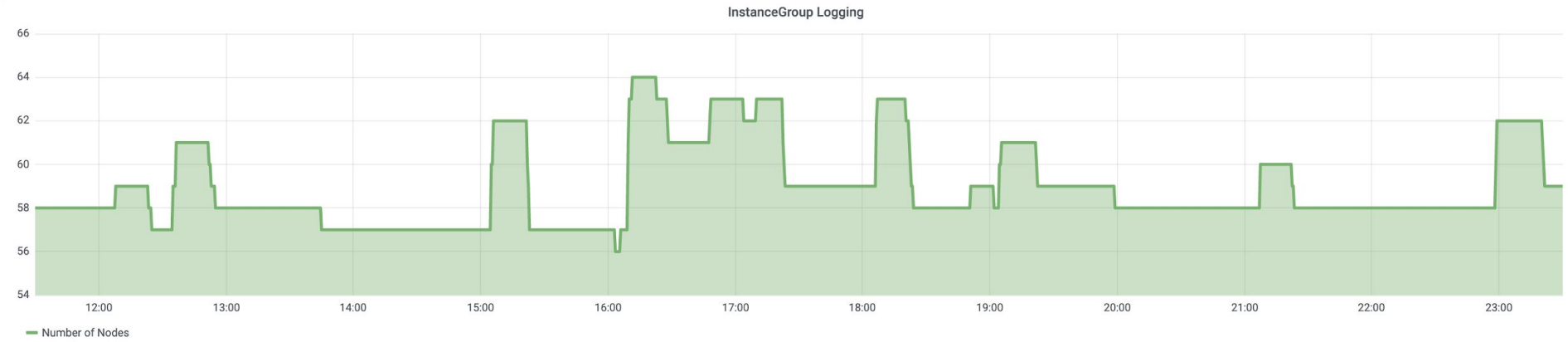

Nodes
212
Namespaces
80
Running Pods
2494
Deployments
602
Statefulsets
21
Services LoadBalancer
96
Jobs
207
Daemonsets
6
Services ClusterIP
405
Services NodePort
41
Cronjobs
50
Services ExternalName
6
HPA
21

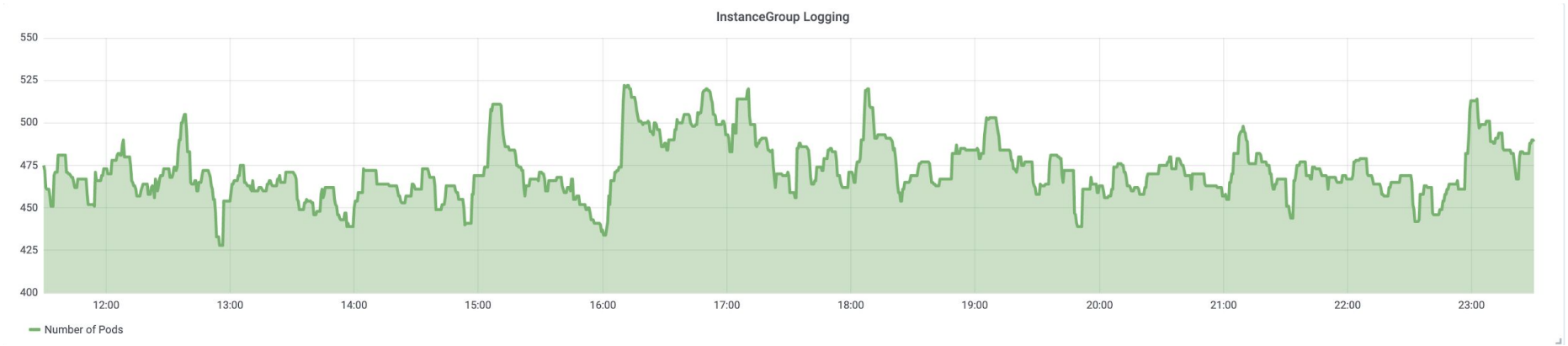
@recollir


InstanceGroup Logging
66
64
62
60
58
56
54
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Number of Nodes

@recollir

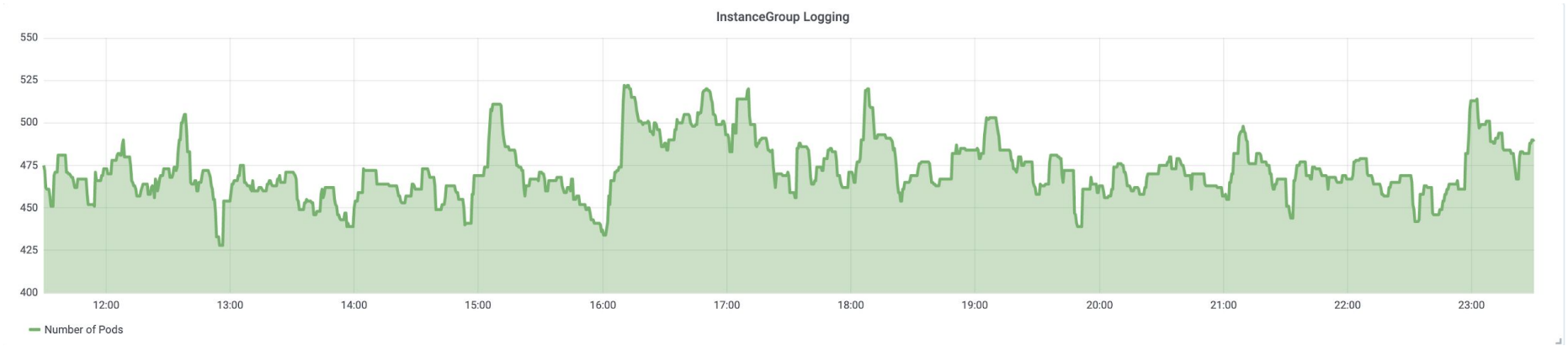
```
sum(kube_node_labels{label_foundation_meltwater_io_instance_class="logging"})
```




InstanceGroup Logging
550
525
500
475
450
425
400
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
Number of Pods

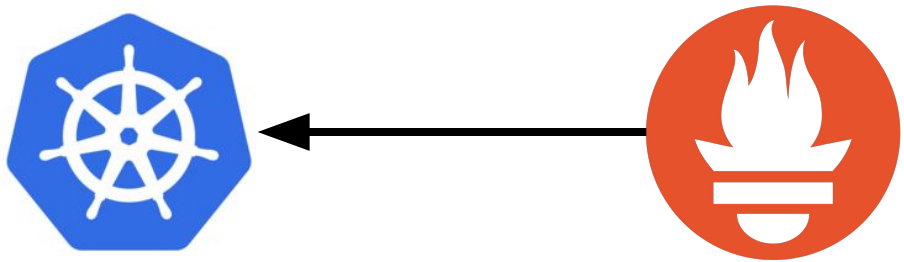
@recollir

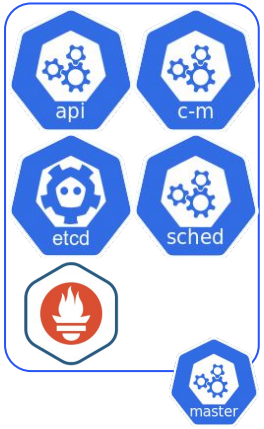
```
sum(kube_pod_info * on(node) group_left(label_kubernetes_io_role) kube_node_labels{label_foundation_meltwater_io_instance_class="logging"})
```

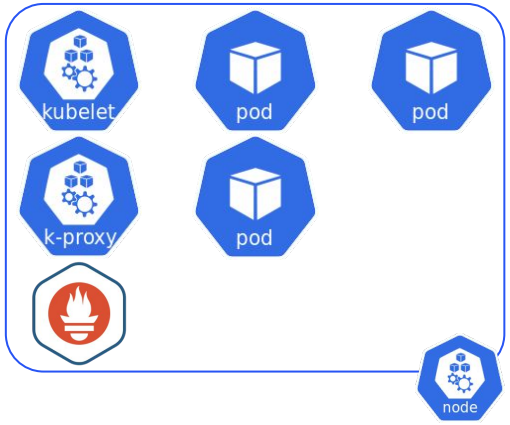


@recollir

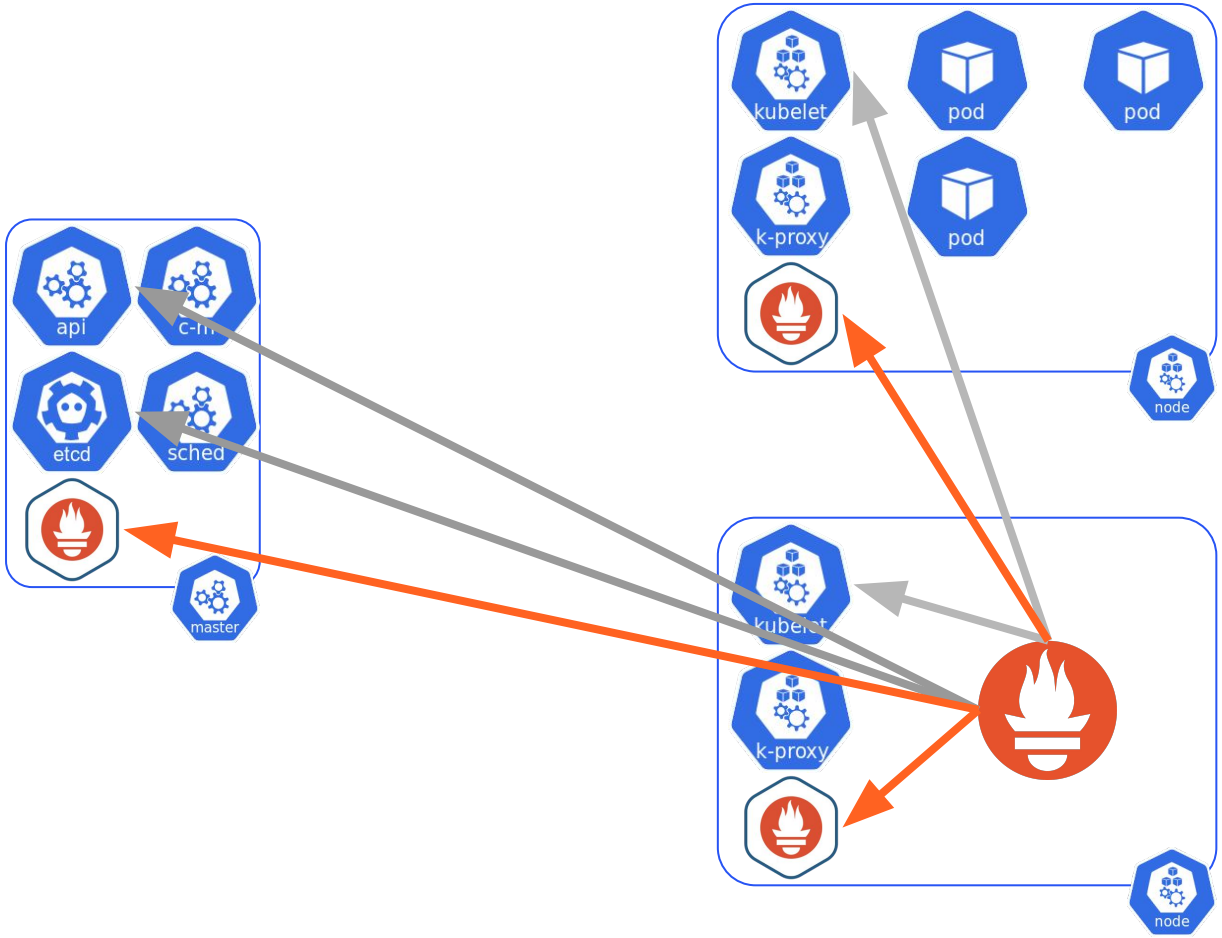
Kubernetes Monitoring

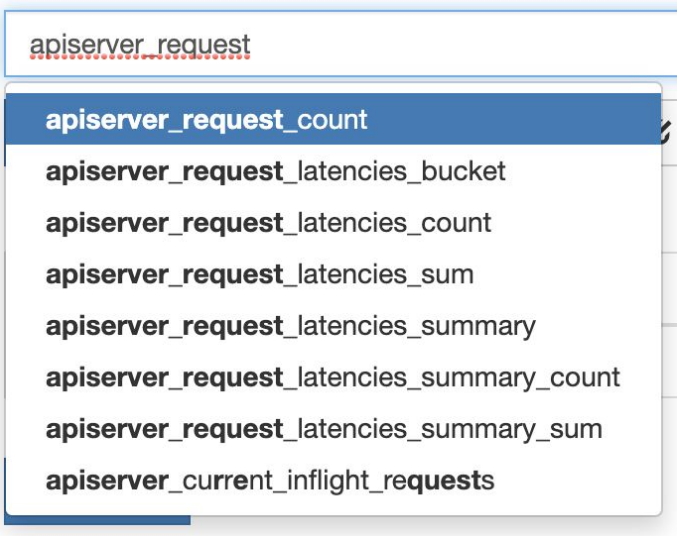



api
c-m
etcd
sched
master


kubelet
pod
pod
k-proxy
pod
node

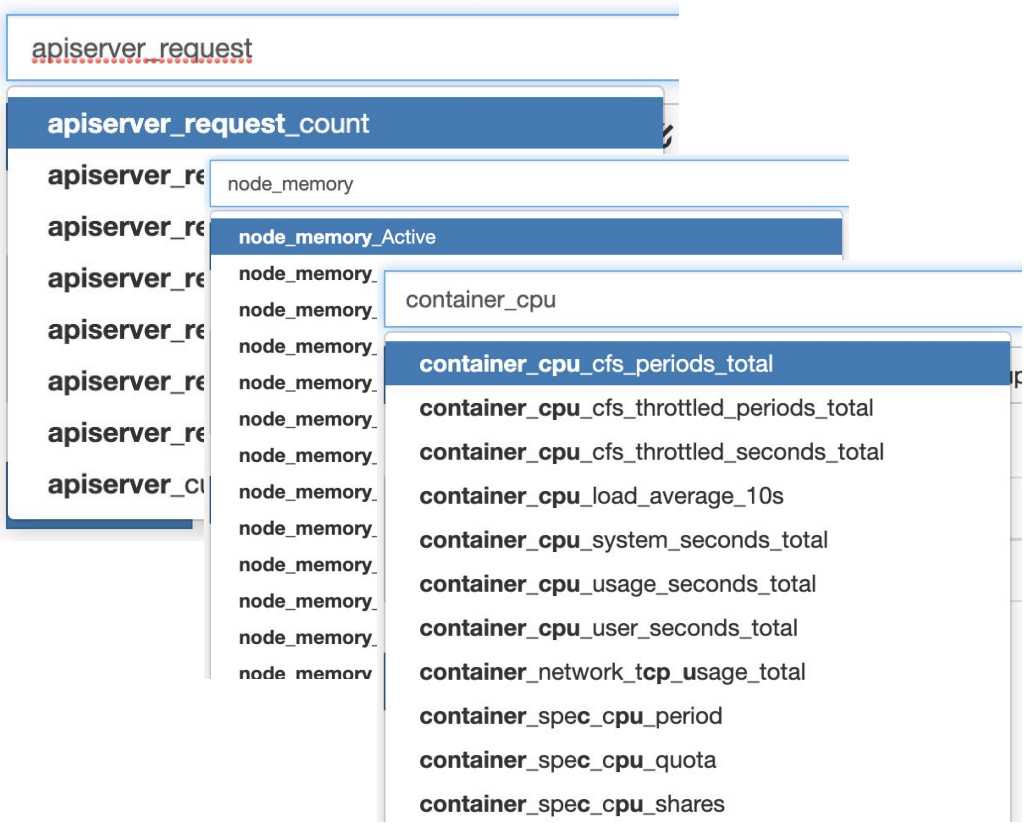

kubelet
k-proxy
node


kubelet
pod
pod
k-proxy
pod
node
api
etcd
sched
master
k-proxy
node



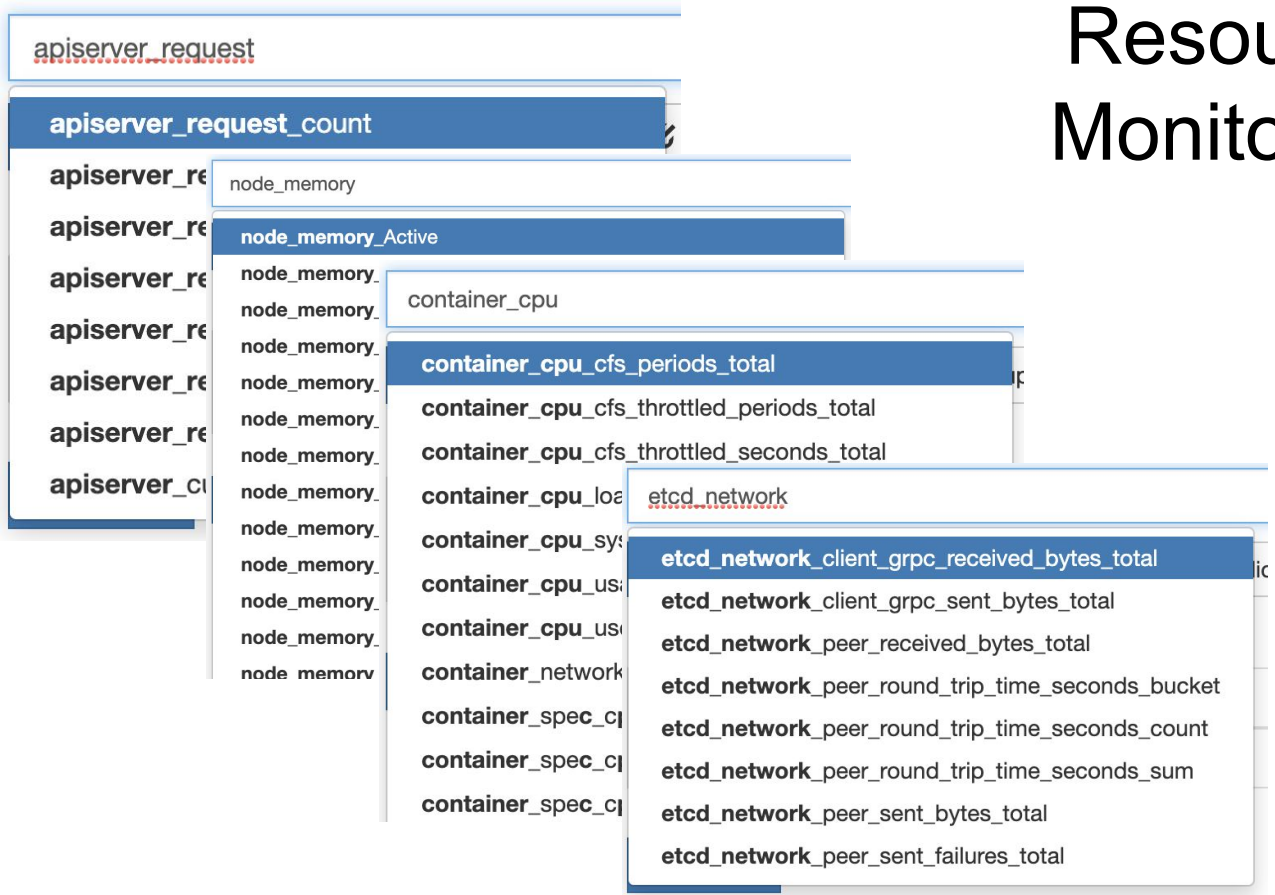
Resource Monitoring

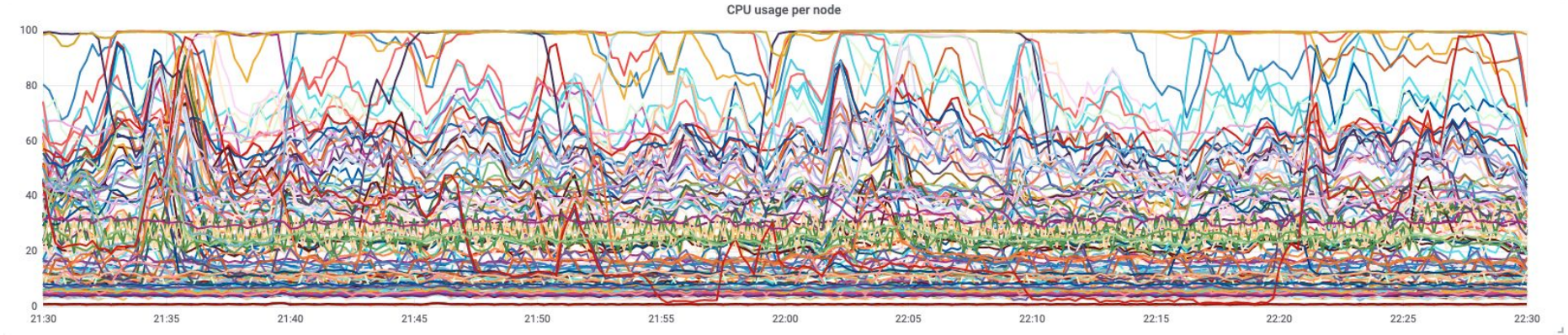
Resource Monitoring



Resource Monitoring

Resource Monitoring




CPU usage per node
100
80
60
40
20
0
21:30
21:35
21:40
21:45
21:50
21:55
22:00
22:05
22:10
22:15
22:20
22:25
22:30

@recollir

But what about the state in the cluster?

Photo from unplash.com

But what about the state in the cluster?

PODS

NODES

DEPLOY MENTS

CRON JOBS

SERVICES

HPA

```
$ kubectl get nodes
NAME                                           STATUS   ROLES    AGE
ip-10-105-131-35.eu-west-1.compute.internal    Ready    node     15m
ip-10-105-132-141.eu-west-1.compute.internal   Ready    node     8m
ip-10-105-132-204.eu-west-1.compute.internal   Ready    node     15m
ip-10-105-139-144.eu-west-1.compute.internal   Ready    node     3m
ip-10-105-140-160.eu-west-1.compute.internal   Ready    master   44m
ip-10-105-153-189.eu-west-1.compute.internal   Ready    master   30m
ip-10-105-154-225.eu-west-1.compute.internal   Ready    node     16m
ip-10-105-157-252.eu-west-1.compute.internal   Ready    node     8m
ip-10-105-158-44.eu-west-1.compute.internal    Ready    node     16m
ip-10-105-159-165.eu-west-1.compute.internal   Ready    node     3m
ip-10-105-160-153.eu-west-1.compute.internal   Ready    node     15m
ip-10-105-161-207.eu-west-1.compute.internal   Ready    node     3m
ip-10-105-161-30.eu-west-1.compute.internal    Ready    node     7m
ip-10-105-163-248.eu-west-1.compute.internal   Ready    node     15m
ip-10-105-164-167.eu-west-1.compute.internal   Ready    node     5m
ip-10-105-166-77.eu-west-1.compute.internal    Ready    master   21m
```

```
$ kubectl get deployment
NAME                          DESIRED   CURRENT   UP-TO-DATE   AVAILABLE
cluster-autoscaler-general    1         1         1            1
cni-metrics-helper            1         1         1            1
dns-controller                1         1         1            1
heapster                      1         1         1            1
kube-dns                      10        10        10           10
kube-dns-autoscaler           1         1         1            1
kubernetes-dashboard          1         1         1            1
metrics-server                1         1         1            1
```

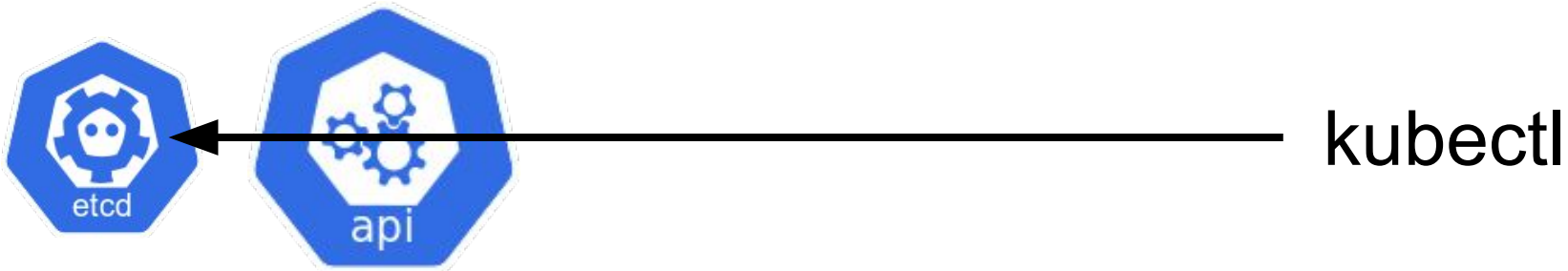
```
$ kubectl get pods --all-namespaces --no-headers | wc -l
    2349
```

```
$ kubectl describe node ip-10-105-211-13.eu-west-1.compute.internal
Labels:                 beta.kubernetes.io/arch=amd64
                        beta.kubernetes.io/instance-type=r4.2xlarge
                        beta.kubernetes.io/os=linux
                        kops.k8s.io/instance-class=general
                        kops.k8s.io/instancegroup=general-eu-west-1
                        kubernetes.io/role=node
Conditions:
  Type                  Status    Reason
  ----                  ------    ------
  OutOfDisk             False     KubeletHasSufficientDisk
  MemoryPressure        False     KubeletHasSufficientMemory
  DiskPressure          False     KubeletHasNoDiskPressure
  PIDPressure           False     KubeletHasSufficientPID
  Ready                 True      KubeletReady
```

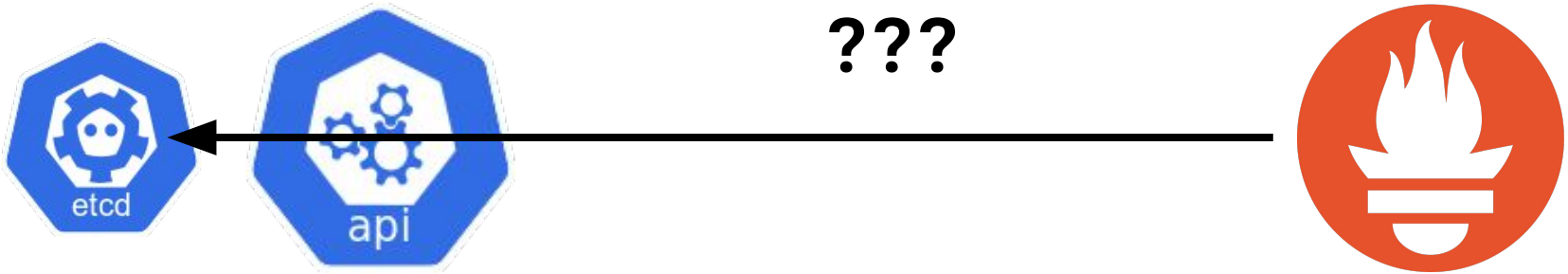
```
$ kubectl describe node ip-10-105-211-13.eu-west-1.compute.internal
Capacity:
  cpu:                8
  ephemeral-storage:  125684580Ki
  hugepages-2Mi:      0
  memory:             62884252Ki
  pods:               58
Allocatable:
  cpu:                8
  ephemeral-storage:  115830908737
  hugepages-2Mi:      0
  memory:             62781852Ki
  pods:               58
Non-terminated Pods:          (27 in total)
  Namespace                   Name                                                                CPU Requests  CPU Limits
  ---------                   ----                                                                ------------  ----------
  xxxxxxxxxxxxxxx             xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx      1 (12%)       0 (0%)
  xxxxxxxxxx                  xxxxxxxxxxxxxxxxxxxx                                                0 (0%)        0 (0%)
  xxxxxxxxxx                  xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx                                    500m (6%)     0 (0%)
Allocated resources:
  (Total limits may be over 100 percent, i.e., overcommitted.)
  CPU Requests  CPU Limits   Memory Requests  Memory Limits
  ------------  ----------   ---------------  -------------
  6380m (79%)   2970m (37%)  10564Mi (17%)    10756Mi (17%)
```

@recollir

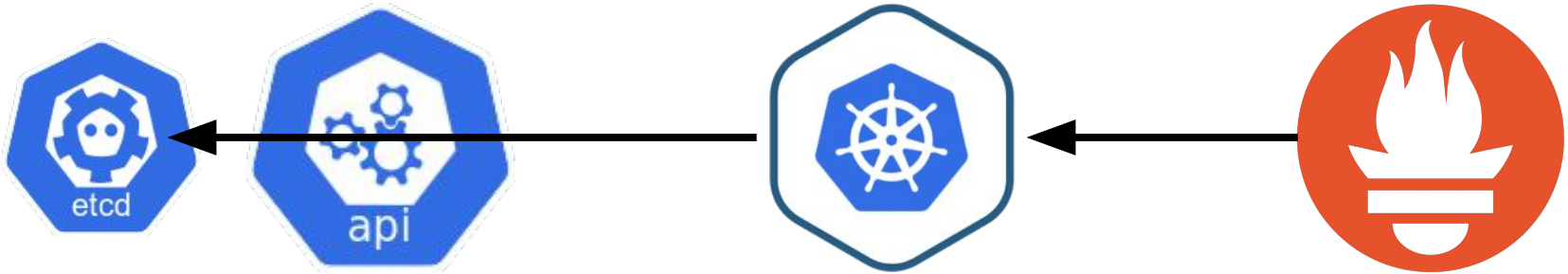
Kubernetes Monitoring



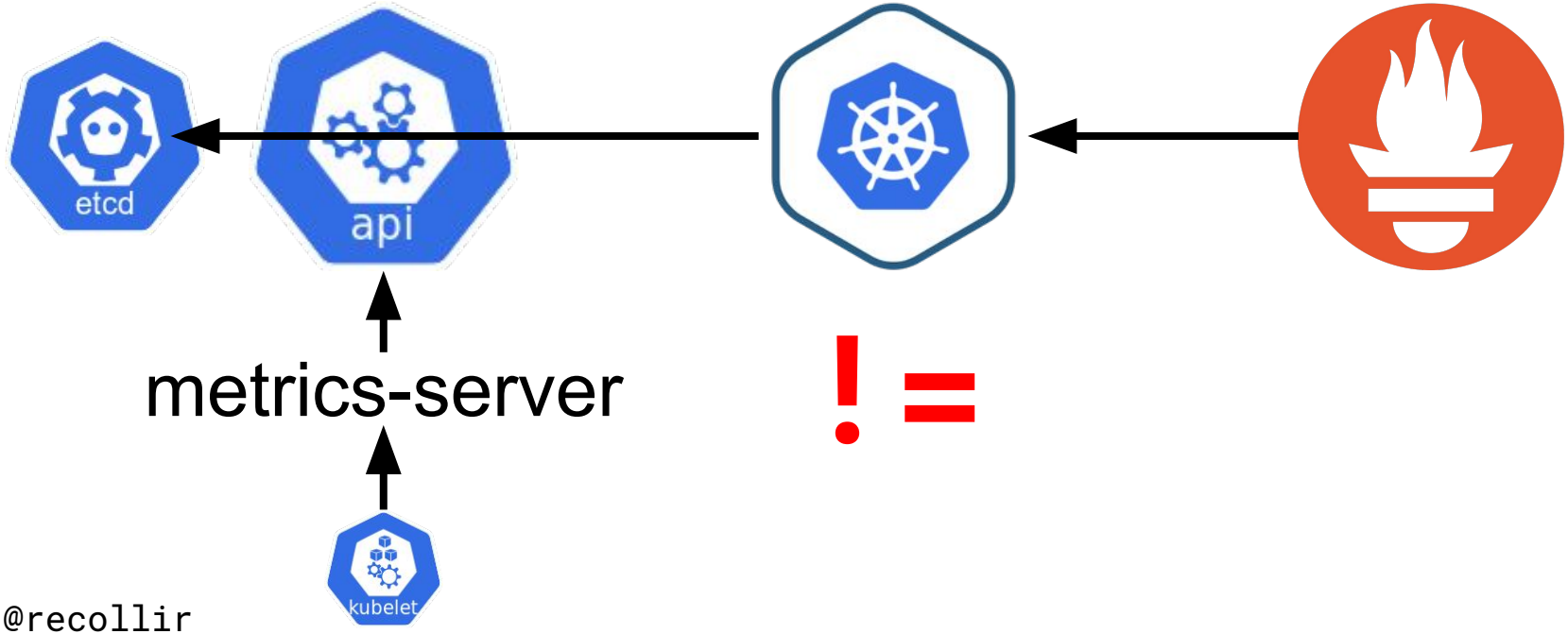
Kubernetes Monitoring

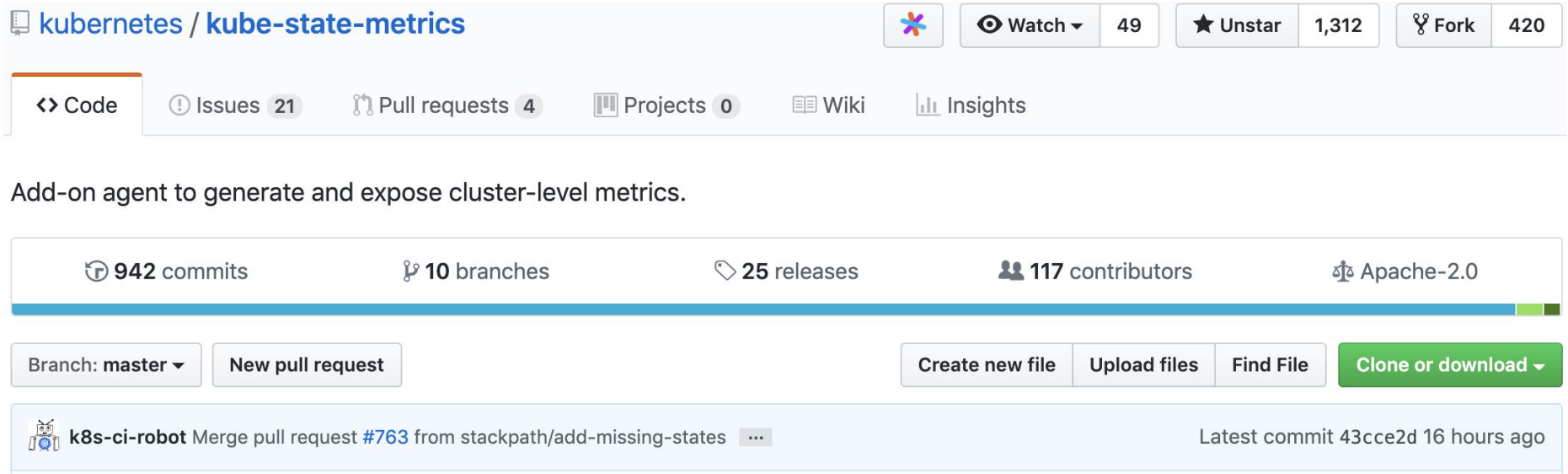


kube-state-metrics

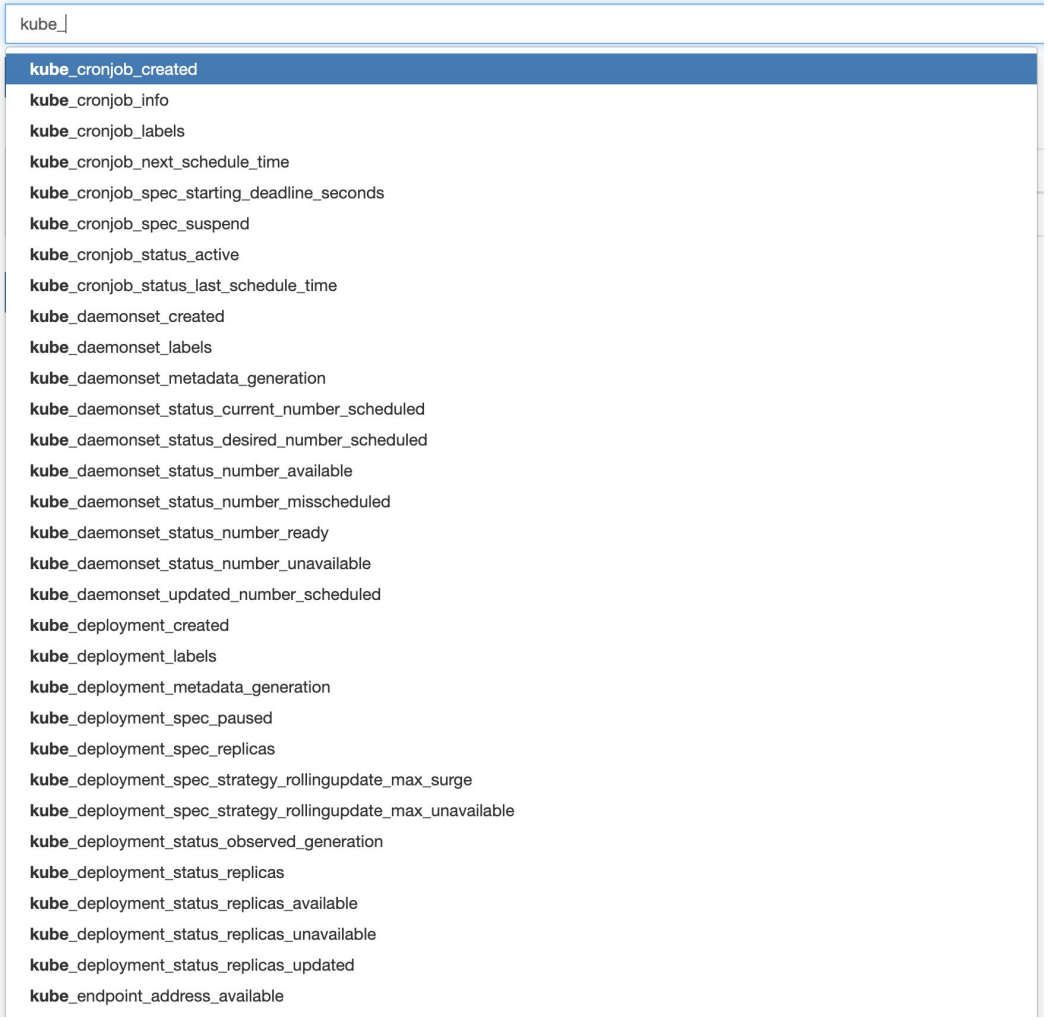


kube-state-metrics




kubernetes / kube-state-metrics
Watch 49
Unstar 1,312
Fork 420
Code
Issues 21
Pull requests 4
Projects 0
Wiki
Insights
Add-on agent to generate and expose cluster-level metrics.
942 commits
10 branches
25 releases
117 contributors
Apache-2.0
Branch: master
New pull request
Create new file
Upload files
Find File
Clone or download
k8s-ci-robot Merge pull request #763 from stackpath/add-missing-states
Latest commit 43cce2d 16 hours ago

@recollir


kube_
kube_cronjob_created
kube_cronjob_info
kube_cronjob_labels
kube_cronjob_next_schedule_time
kube_cronjob_spec_starting_deadline_seconds
kube_cronjob_spec_suspend
kube_cronjob_status_active
kube_cronjob_status_last_schedule_time
kube_daemonset_created
kube_daemonset_labels
kube_daemonset_metadata_generation
kube_daemonset_status_current_number_scheduled
kube_daemonset_status_desired_number_scheduled
kube_daemonset_status_number_available
kube_daemonset_status_number_misscheduled
kube_daemonset_status_number_ready
kube_daemonset_status_number_unavailable
kube_daemonset_updated_number_scheduled
kube_deployment_created
kube_deployment_labels
kube_deployment_metadata_generation
kube_deployment_spec_paused
kube_deployment_spec_replicas
kube_deployment_spec_strategy_rollingupdate_max_surge
kube_deployment_spec_strategy_rollingupdate_max_unavailable
kube_deployment_status_observed_generation
kube_deployment_status_replicas
kube_deployment_status_replicas_available
kube_deployment_status_replicas_unavailable
kube_deployment_status_replicas_updated
kube_endpoint_address_available

```
kube_kind_metric{label1="value1",
                 label2="value2"} value
```

kube_

kube_service_spec_type

info		info	info				info
labels	labels	labels	labels	labels	labels	labels	labels
created	created	created	created	created	created		created
start_time	status_replicas	spec_type	metadata_resource_version	status_replicas	status_current_number_scheduled	metadata_generation	spec_unschedulable
completion_time	status_replicas_available	spec_external_ip		status_replicas_current	status_desired_number_scheduled	spec_max_replicas	spec_taint
owner	status_replicas_unavailable	status_load_balancer_ingress		status_replicas_ready	status_number_available	spec_min_replicas	status_phase
status_phase	status_replicas_updated			status_replicas_updated	status_number_misscheduled	status_current_replicas	status_capacity
status_ready	status_observed_generation			status_observed_generation	status_number_ready	status_desired_replicas	status_capacity_cpu_cores
status_scheduled	spec_replicas			replicas	status_number_unavailable	status_condition	status_capacity_memory_bytes
container_info	spec_paused			metadata_generation	updated_number_scheduled		status_capacity_pods
container_status_waiting	spec_strategy_rollingupdate_max_unavailable			status_current_revision	metadata_generation		status_allocatable

```
$ kubectl port-forward service/kube-state-metrics 8080:8080 &
$ curl -s http://localhost:8080/metrics | fgrep HELP > ksm.txt
$ fgrep kube_node_   ksm.txt
```

```
# HELP kube_node_info                          Information about a cluster node.
# HELP kube_node_created                       Unix creation timestamp
# HELP kube_node_labels                        Kubernetes labels converted to Prometheus labels.
# HELP kube_node_spec_unschedulable            Whether a node can schedule new pods.
# HELP kube_node_spec_taint                    The taint of a cluster node.
# HELP kube_node_status_condition              The condition of a cluster node.
# HELP kube_node_status_phase                  The phase the node is currently in.
# HELP kube_node_status_capacity               The capacity for different resources of a node.
# HELP kube_node_status_capacity_pods          The total pod resources of the node.
# HELP kube_node_status_capacity_cpu_cores     The total CPU resources of the node.
# HELP kube_node_status_capacity_memory_bytes  The total memory resources of the node.
# HELP kube_node_status_allocatable            The allocatable for diff. resources of a node that are available.
# HELP kube_node_status_allocatable_pods       The pod resources of a node that are available for scheduling.
# HELP kube_node_status_allocatable_cpu_cores  The CPU resources of a node that are available for scheduling.
# HELP kube_node_status_allocatable_memory_bytes The memory resources of a node that are available for scheduling.
```

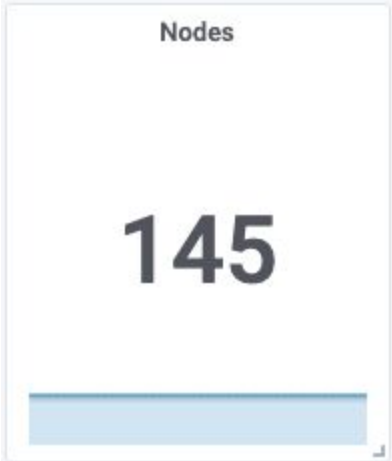
kube_node_labels

Node Labels

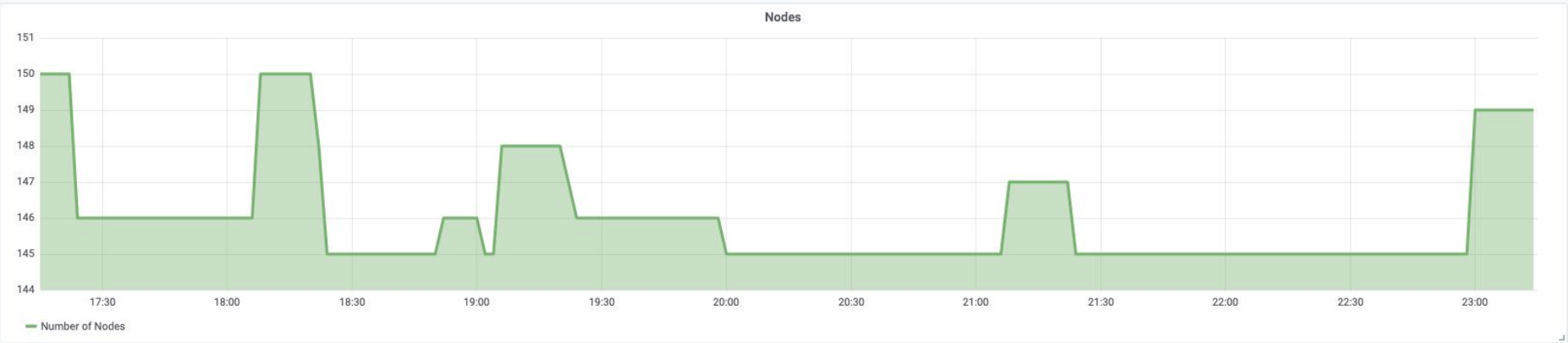
label_beta_kubernetes_io_instance_type	label_foundation_meltwater_io_ec2_class	label_foundation_meltwater_io_instance_class	label_kubernetes_io_role	Value
r4.2xlarge	r4	logging	node	1
r4.2xlarge	r4	logging	node	1
r4.2xlarge	r4	logging	node	1
r4.2xlarge	r4	logging	node	1
r4.2xlarge	r4	general	node	1
r4.2xlarge	r4	general	node	1
r4.2xlarge	r4	general	node	1
r4.2xlarge	r4	general	node	1
r4.2xlarge	r4	general	node	1

1 2

```
sum(kube_node_labels)
```



```
sum(kube_node_labels)
```

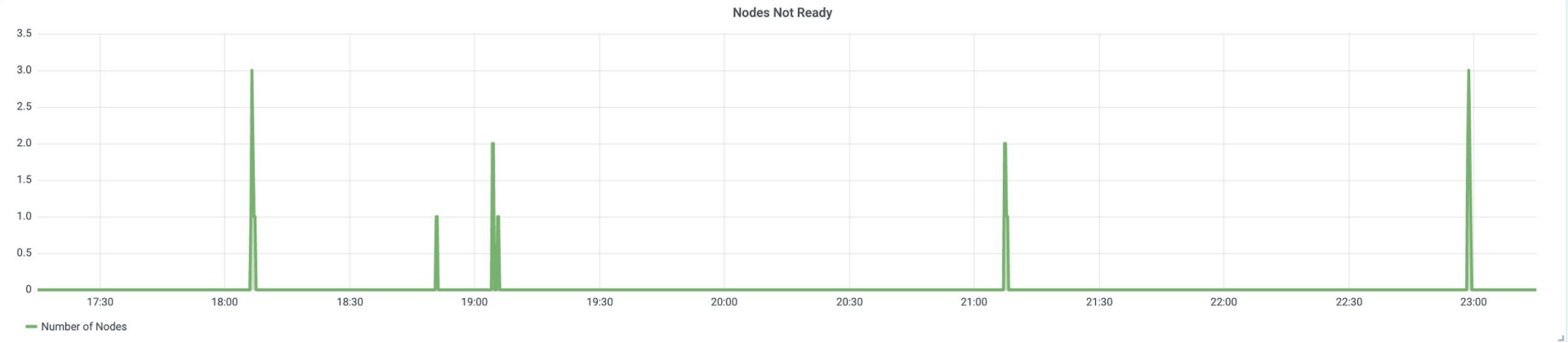


```
sum(kube_node_labels) by (label_foundation_meltwater_io_instance_class, label_beta_kubernetes_io_instance_type)
```

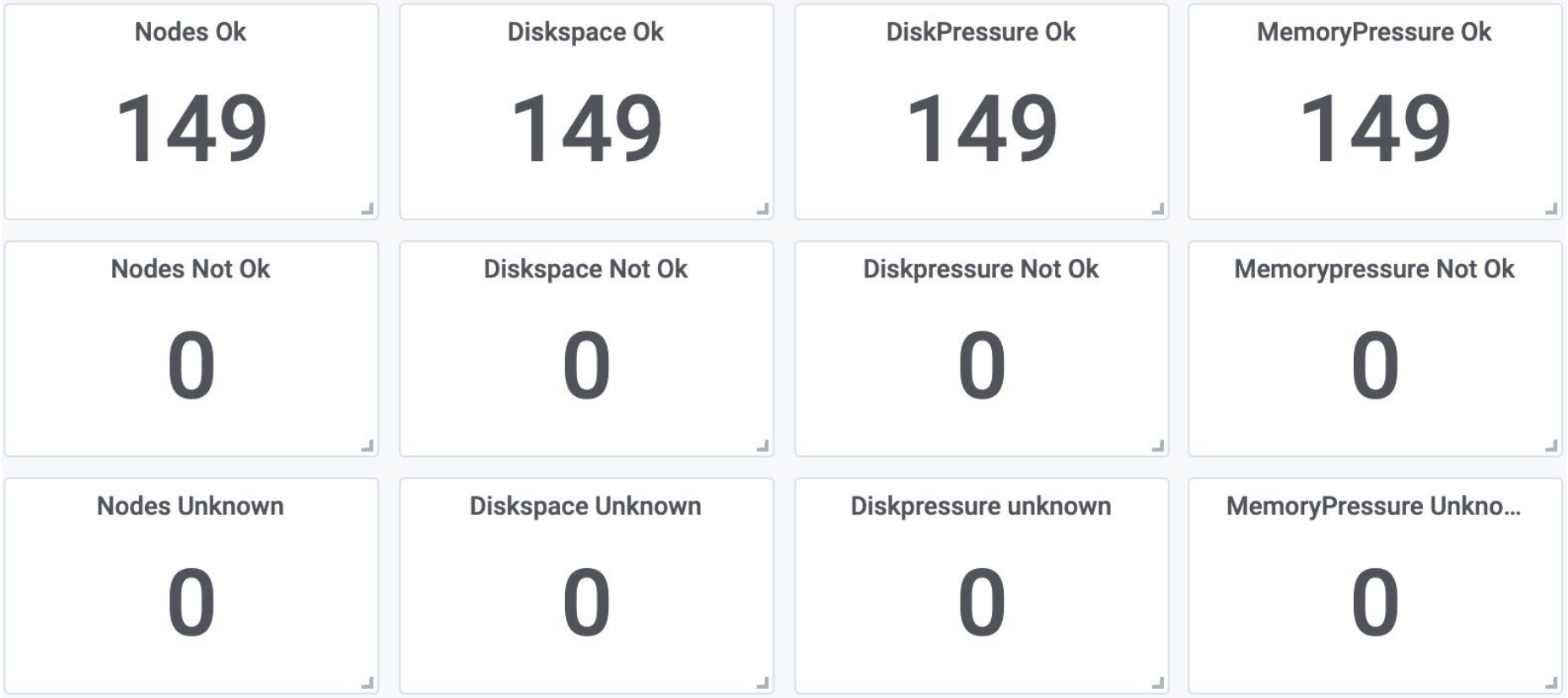
Node Labels

Instance Type	Instance Group	Value
r4.xlarge	master	3
r4.2xlarge	general	77
c5.2xlarge	logging	120
r4.4xlarge	monitoring	9

```
sum(kube_node_status_condition{condition="Ready",status="false"})
```

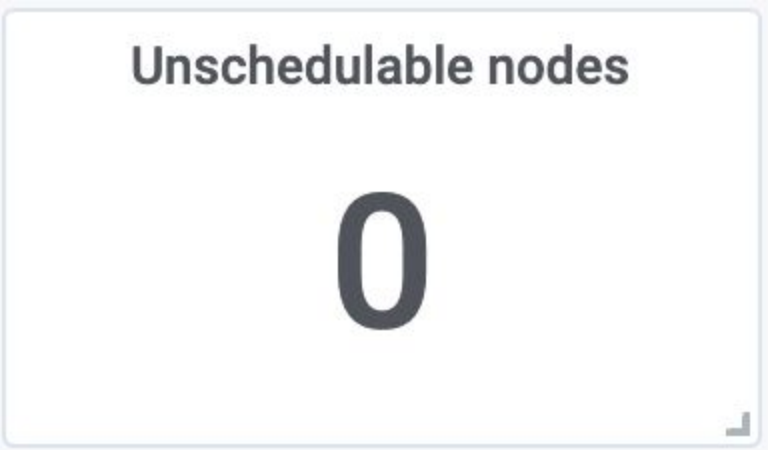


@recollir


Nodes Ok
149
Diskspace Ok
149
DiskPressure Ok
149
MemoryPressure Ok
149
Nodes Not Ok
0
Diskspace Not Ok
0
Diskpressure Not Ok
0
Memorypressure Not Ok
0
Nodes Unknown
0
Diskspace Unknown
0
Diskpressure unknown
0
MemoryPressure Unkno...
0

@recollir

```
sum(kube_node_spec_unschedulable)
```

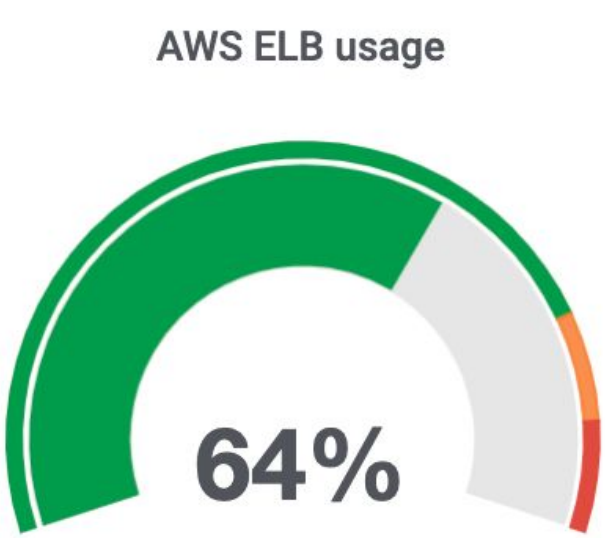


```
sum(kube_service_spec_type) by (type)
```

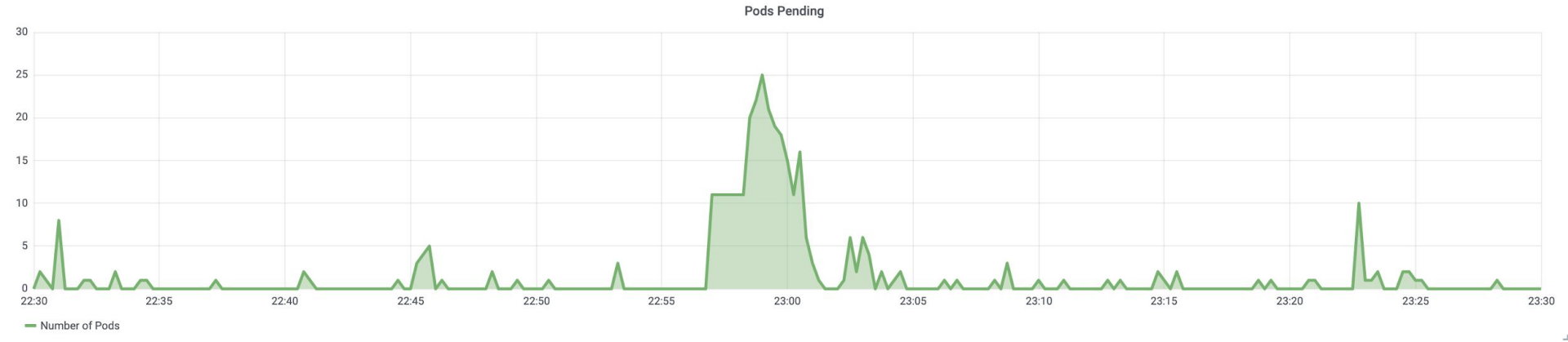
Kind: Service

Type	Number
ClusterIP	394
LoadBalancer	93
NodePort	41
ExternalName	6

```
sum(kube_service_spec_type{type="LoadBalancer"}) / aws_elb_limit_total
```



```
sum(kube_pod_status_phase{phase="Pending"})
```



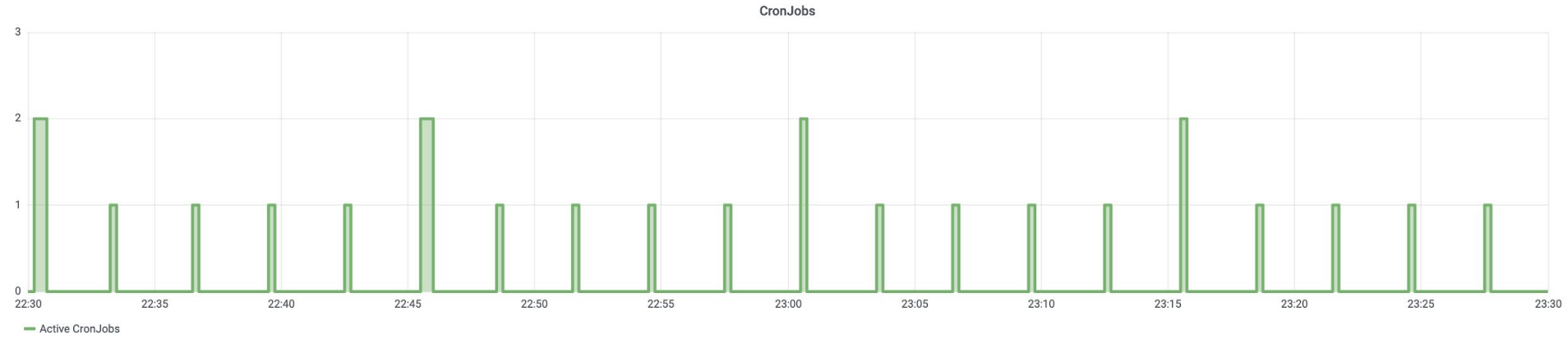
@recollir

```
sum(kube_pod_container_status_last_terminated_reason) by (reason)
```

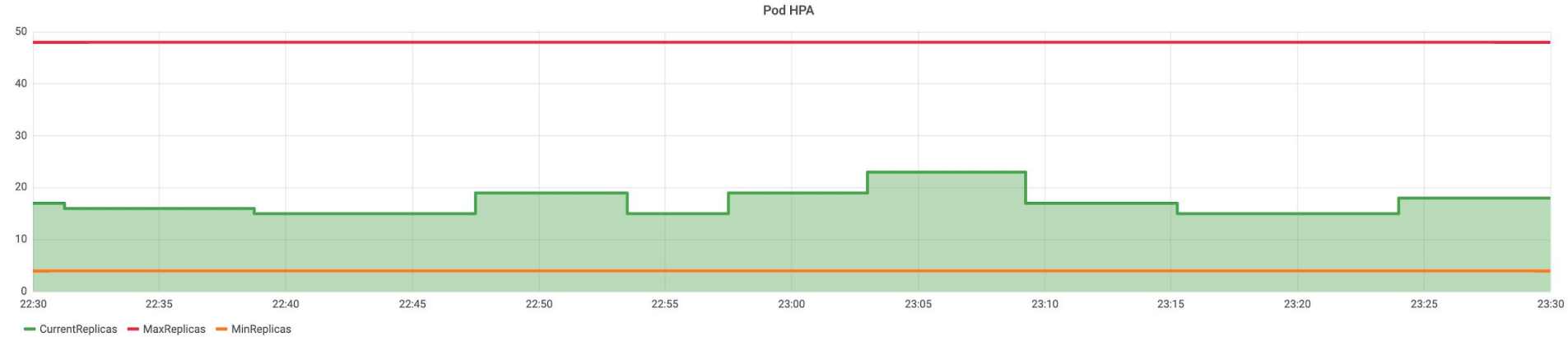
Pod: Termination Reason

Reason ▲	Number
Completed	13
ContainerCannotRun	1
Error	109
OOMKilled	39

```
sum(kube_cronjob_status_active{namespace="foo"})
```

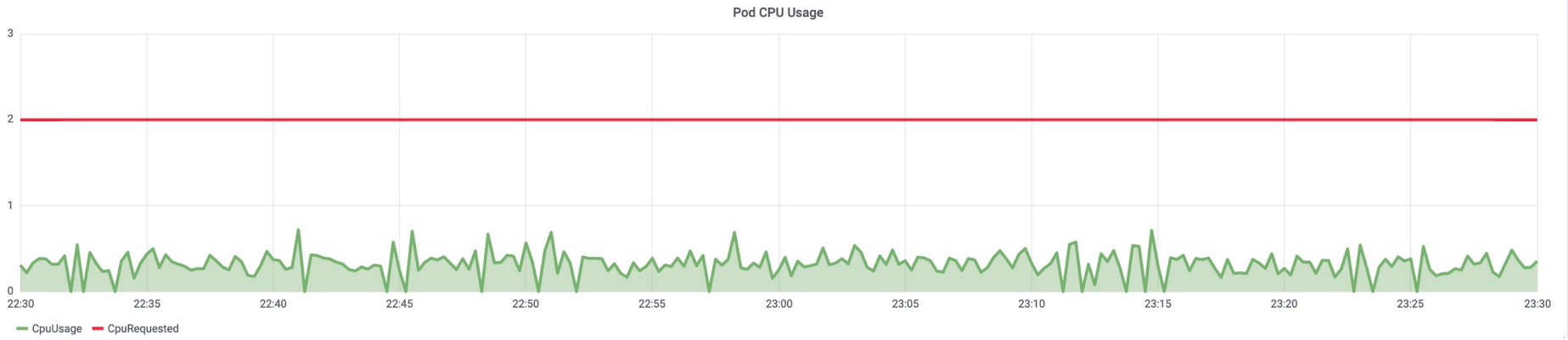


```
kube_hpa_status_current_replicas{hpa="indexer-prod"}
kube_hpa_spec_max_replicas{hpa="indexer-prod"}
kube_hpa_spec_min_replicas{hpa="indexer-prod"}
```

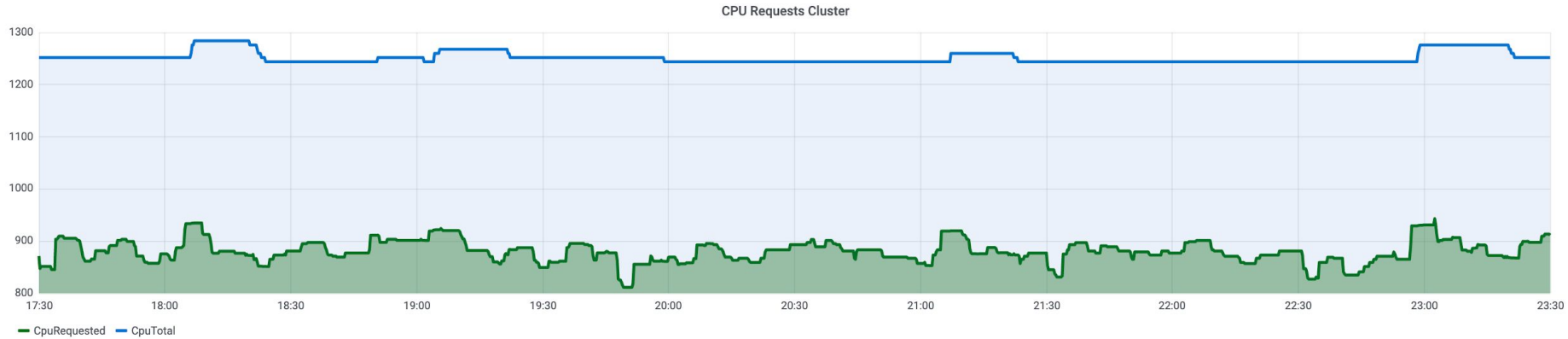


```
kube_pod_container_resource_requests_cpu_cores{namespace="foo",container="foo"}
```

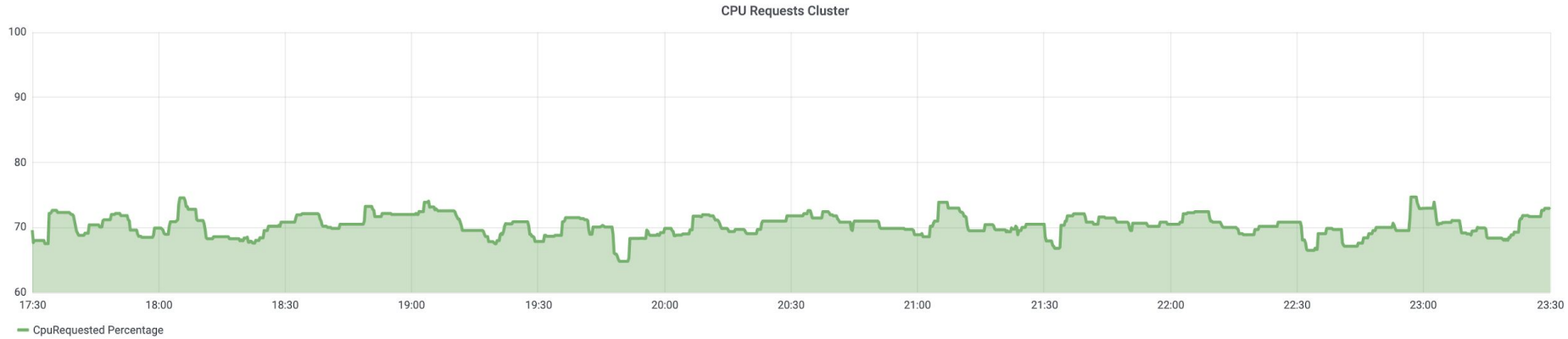
```
irate(container_cpu_usage_seconds_total{namespace="foo", container_name="foo"}[30s])
```



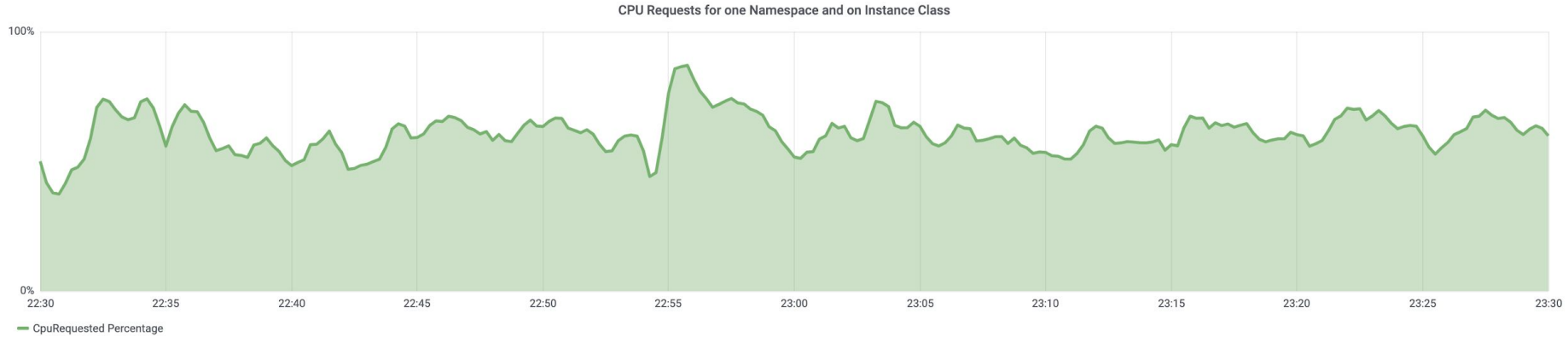
```
sum(kube_node_status_allocatable_cpu_cores)
sum(kube_pod_container_resource_requests_cpu_cores)
```



```
sum(kube_pod_container_resource_requests_cpu_cores) / sum(kube_node_status_allocatable_cpu_cores)
```



```
sum(rate(container_cpu_usage_seconds_total{namespace="foo", foundation_meltwater_io_instance_class="bar"}[1m]))
/
sum(kube_pod_container_resource_requests_cpu_cores{namespace="foo"}
    * on(node) group_left(label_foundation_meltwater_io_instance_class)
    kube_node_labels{label_foundation_meltwater_io_instance_class="bar"})
```



@recollir

- kube-state-metrics - Prometheus exporter for cluster state
- github.com/kubernetes/kube-state-metrics
- Mostly binary value (0/1) metrics for Kubernetes objects (kind)
- Usage:
  - With sum() or count() for overall information
  - Filter with labels
  - *Join* with resource metrics

```
resourcemetric * on(label) group_left(label) kube_kind_metric
```

Federico Hernandez

@recollir