Scale Your Service on What Matters: Autoscaling on Latency

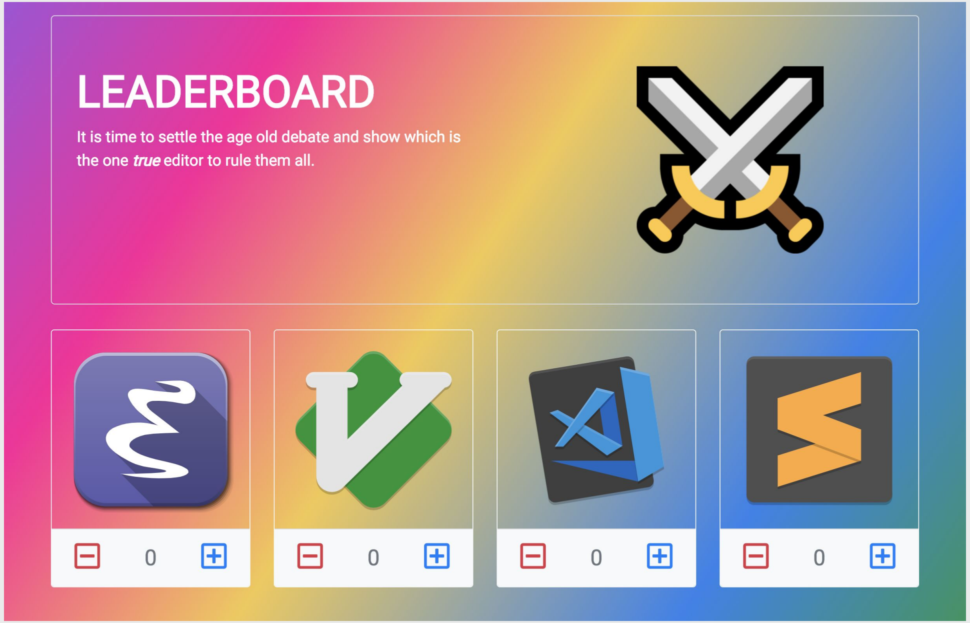
👋 Hello KubeCon!

Thomas Rampelberg

Software Engineer @ Buoyant

Twitter: @grampelberg

Get your votes in



LEADERBOARD

It is time to settle the age old debate and show which is the one ***true*** editor to rule them all.

http://kc.l5d.io

↕️ Autoscaling in Kubernetes

- Cluster Autoscaler

➡ Horizontal Pod Autoscaler

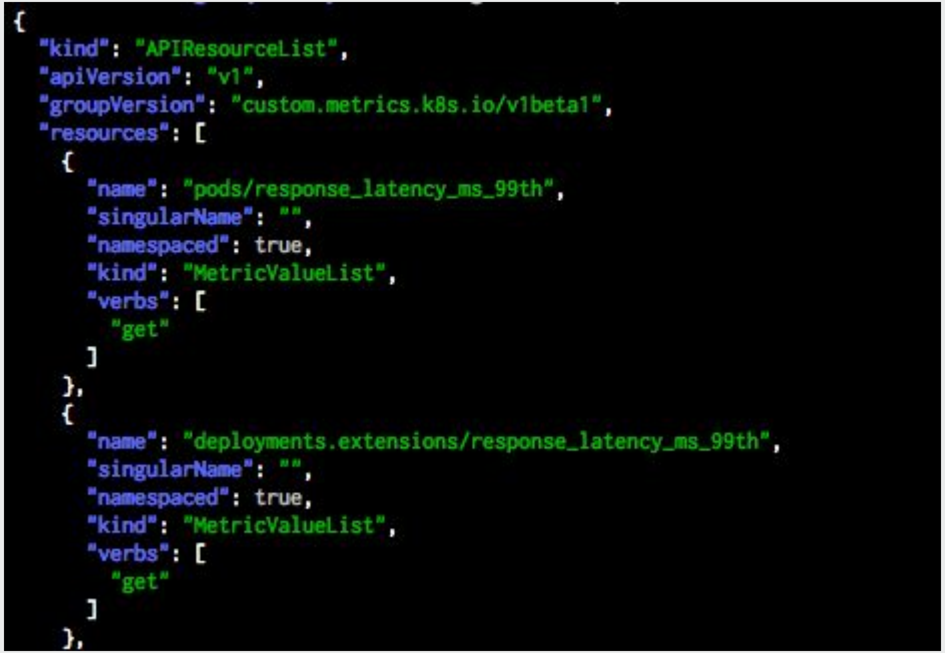
- Vertical Pod Autoscaler

↔ Horizontal Pod Autoscaler

- metrics.k8s.io

➜ custom.metrics.k8s.io

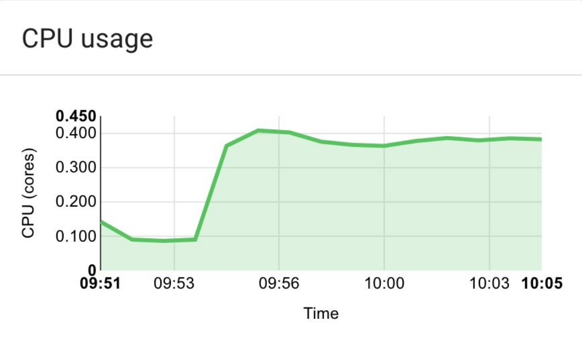
- external.metrics.k8s.io



```
{
  "kind": "APIResourceList",
  "apiVersion": "v1",
  "groupVersion": "custom.metrics.k8s.io/v1beta1",
  "resources": [
    {
      "name": "pods/response_latency_ms_99th",
      "singularName": "",
      "namespaced": true,
      "kind": "MetricValueList",
      "verbs": [
        "get"
      ]
    },
    {
      "name": "deployments.extensions/response_latency_ms_99th",
      "singularName": "",
      "namespaced": true,
      "kind": "MetricValueList",
      "verbs": [
        "get"
      ]
    },
```

CPU is an approximation

- Not every workload is CPU bound
- Isn’t 100% utilization good?
- CPUs are different in the cloud
- Orchestrated environments are complex



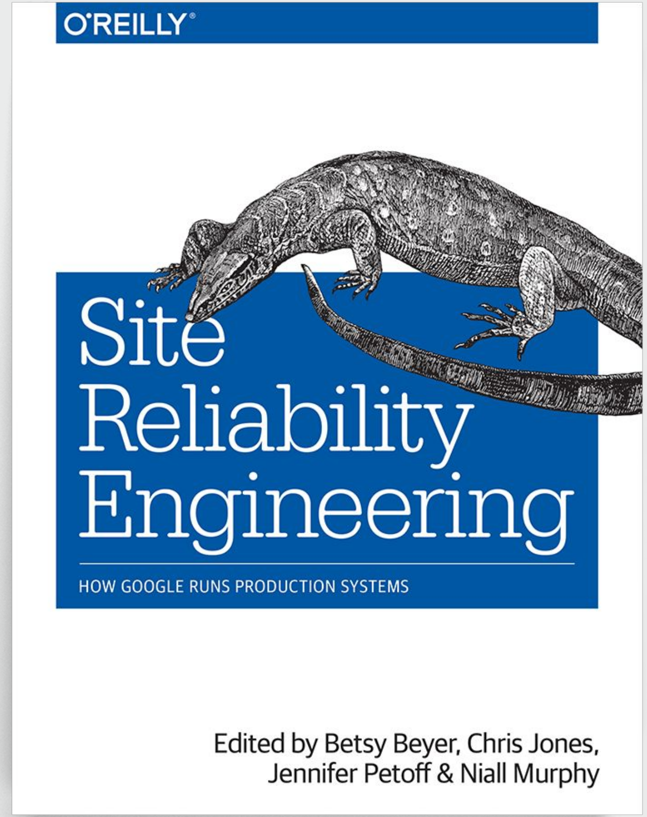
CPU utilization is wrong: http://bit.ly/cpu-wrong
Utilization is useless as a metric: http://bit.ly/useless-metric

Memory is workload specific

What can you scale on?

Golden Signals

- Latency
- Traffic
- Errors
- Saturation



O'REILLY®

Site Reliability Engineering

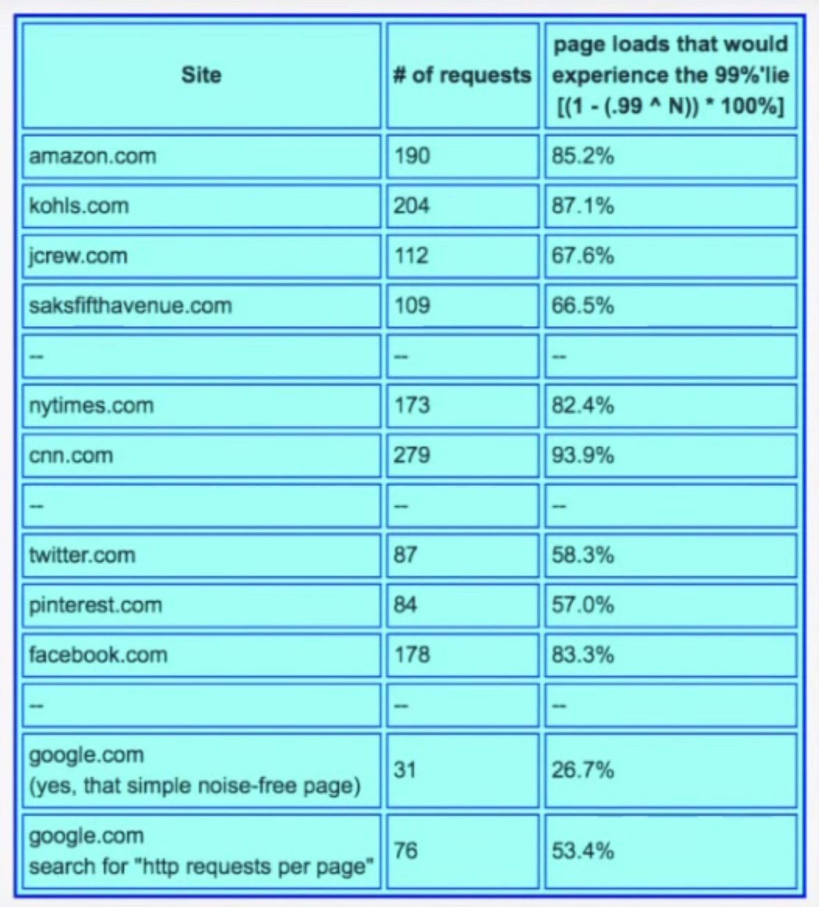
HOW GOOGLE RUNS PRODUCTION SYSTEMS

Edited by Betsy Beyer, Chris Jones,
Jennifer Petoff & Niall Murphy

http://bit.ly/golden-signals

🤗 Every request matters

- Tail latency is important
- Users see responses
- Latency is not normally distributed



Site	# of requests	page loads that would experience the 99%'lie [(1 - (.99 ^ N)) * 100%]
amazon.com	190	85.2%
kohls.com	204	87.1%
jcrew.com	112	67.6%
saksfifthavenue.com	109	66.5%
--	--	--
nytimes.com	173	82.4%
cnn.com	279	93.9%
--	--	--
twitter.com	87	58.3%
pinterest.com	84	57.0%
facebook.com	178	83.3%
--	--	--
google.com (yes, that simple noise-free page)	31	26.7%
google.com search for "http requests per page"	76	53.4%

http://bit.ly/latency-wrong

🌡️ What is required?

- □ Measure the latency of a service
- □ Expose custom metrics
- □ Autoscale!

LINKERD

An open source *service mesh* and CNCF member project.

- ➢ 24+ months in production
- ➢ 2,500+ Slack members
- ➢ 7,500+ GitHub stars
- ➢ 40m+ DockerHub pulls
- ➢ 100+ contributors
- ➢ 400b+ requests/mo

salesforce OfferUp CHASE

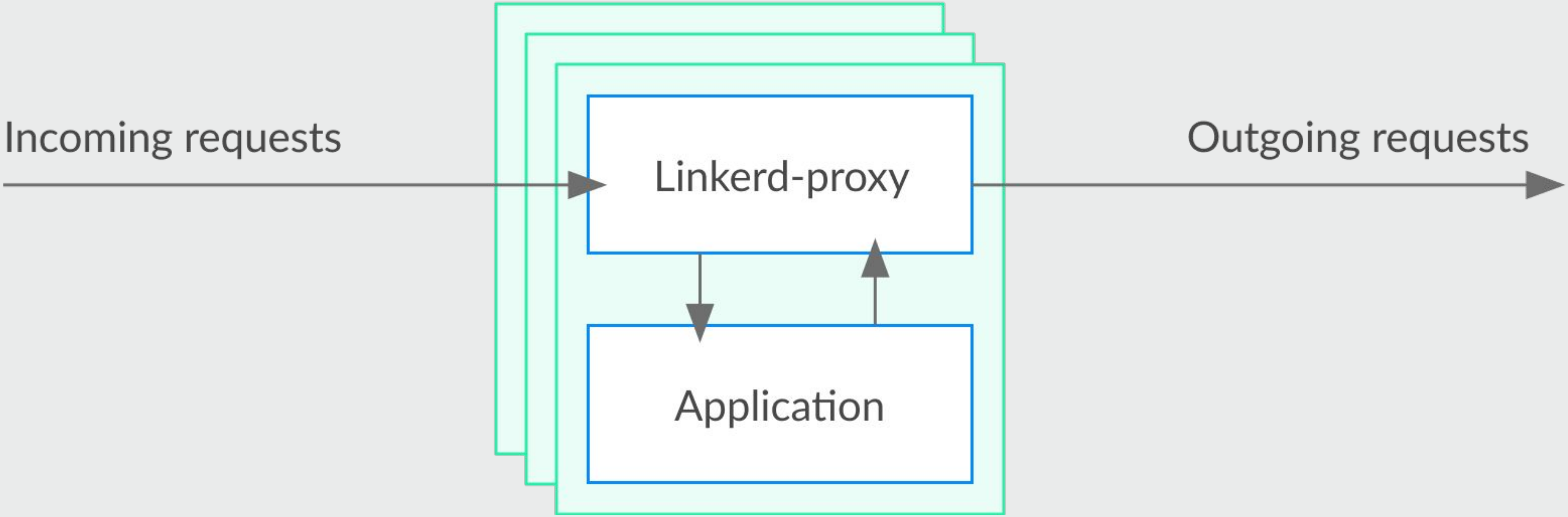
COMCAST monzo Expedia

planet. STRAVA credit karma

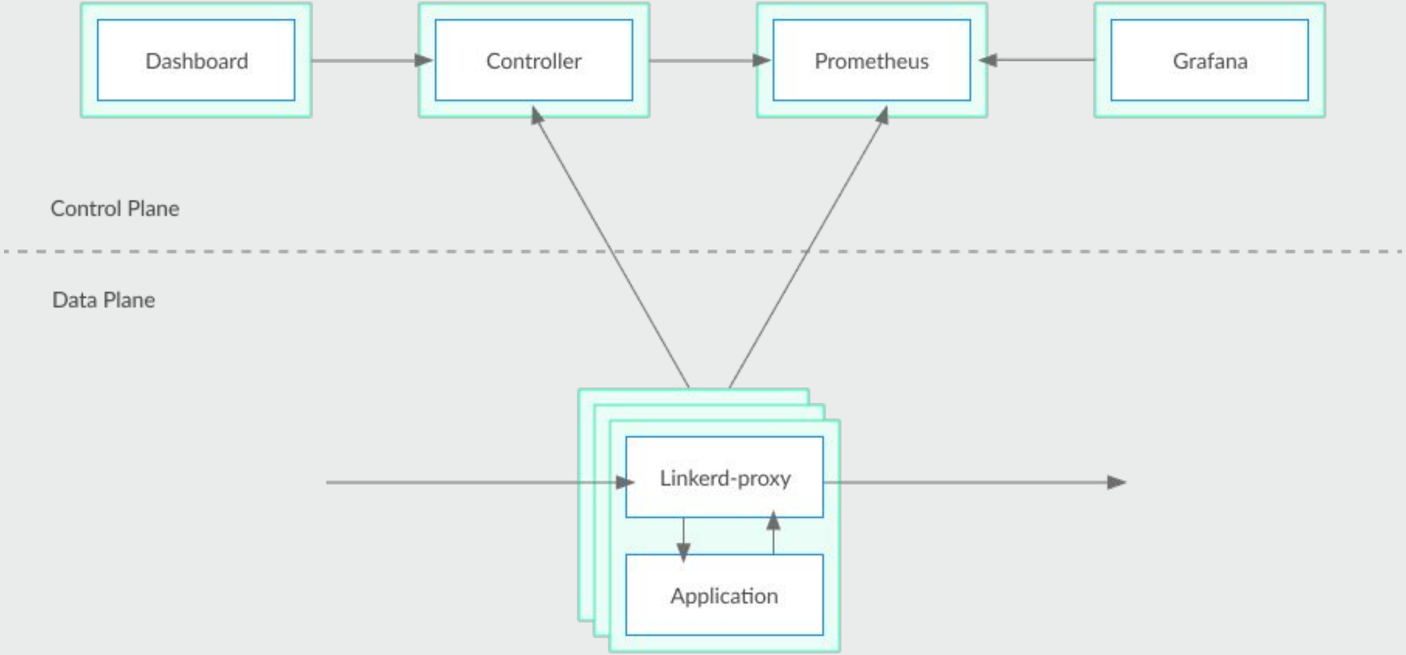
BIGCOMMERCE Olark FOX

CLOUD NATIVE
COMPUTING
FOUNDATION

What is Linkerd?



Architecture



📈

Measure the latency of a service

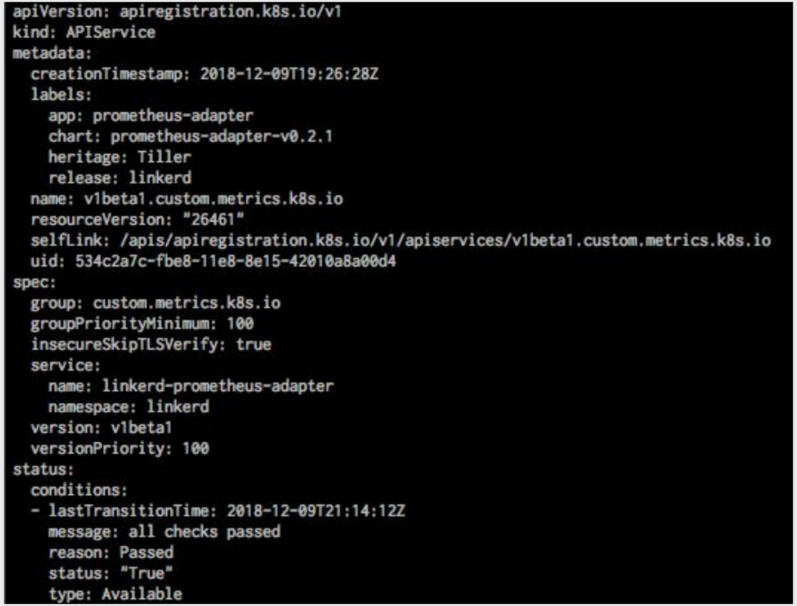
🌡 What is required?

☑ Measure the latency of a service

□ Expose custom metrics

□ Autoscale!

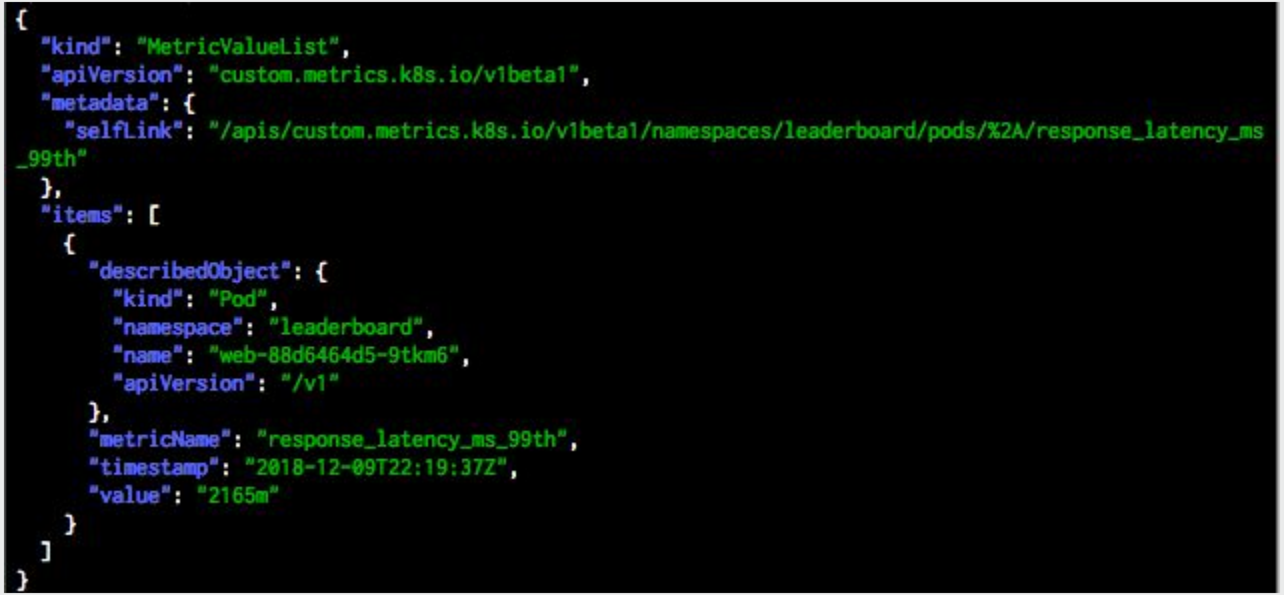
🙋‍♀️ What are custom metrics?



```
apiVersion: apiregistration.k8s.io/v1
kind: APIService
metadata:
  creationTimestamp: 2018-12-09T19:26:28Z
  labels:
    app: prometheus-adapter
    chart: prometheus-adapter-v0.2.1
    heritage: Tiller
    release: linkerd
  name: v1beta1.custom.metrics.k8s.io
  resourceVersion: "26461"
  selfLink: /apis/apiregistration.k8s.io/v1/apiservices/v1beta1.custom.metrics.k8s.io
  uid: 534c2a7c-fbe8-11e8-8e15-42010a8a00d4
spec:
  group: custom.metrics.k8s.io
  groupPriorityMinimum: 100
  insecureSkipTLSVerify: true
  service:
    name: linkerd-prometheus-adapter
    namespace: linkerd
  version: v1beta1
  versionPriority: 100
status:
  conditions:
  - lastTransitionTime: 2018-12-09T21:14:12Z
    message: all checks passed
    reason: Passed
    status: "True"
    type: Available
```

http://bit.ly/k8s-aggregation

Prometheus Adapter

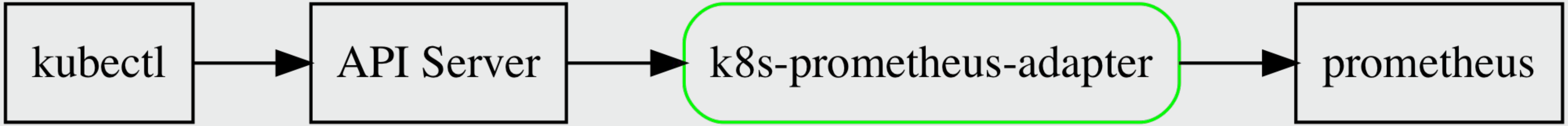


```
{
  "kind": "MetricValueList",
  "apiVersion": "custom.metrics.k8s.io/v1beta1",
  "metadata": {
    "selfLink": "/apis/custom.metrics.k8s.io/v1beta1/namespaces/leaderboard/pods/%2A/response_latency_ms
_99th"
  },
  "items": [
    {
      "describedObject": {
        "kind": "Pod",
        "namespace": "leaderboard",
        "name": "web-88d6464d5-9tkm6",
        "apiVersion": "/v1"
      },
      "metricName": "response_latency_ms_99th",
      "timestamp": "2018-12-09T22:19:37Z",
      "value": "2165m"
    }
  ]
}
```

histogram_quantile(0.99, sum(irate(<<.Series>>{<<.LabelMatchers>>, direction="inbound"}[5m])) by (le, <<.GroupBy>>))

http://bit.ly/k8s-adapter

🏢 Architecture



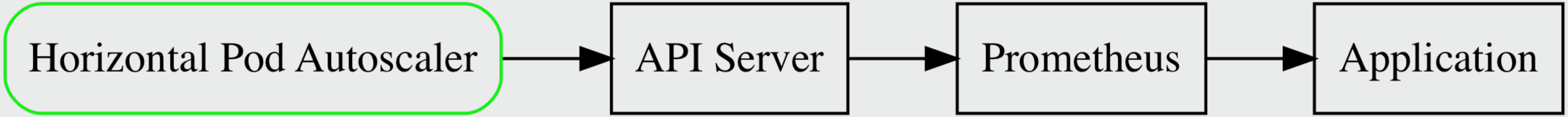
💪

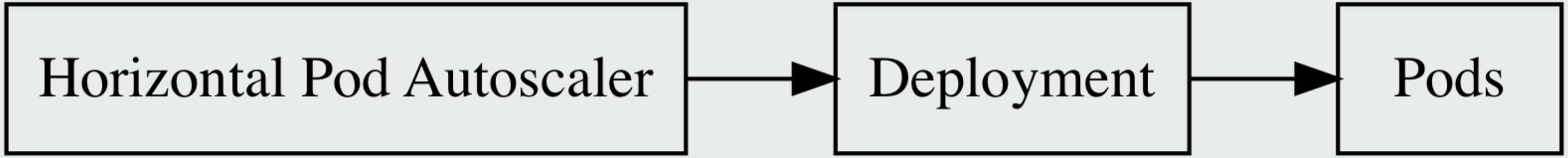
Expose custom metrics

🌡️ What is required?

- ☑ Measure the latency of a service
- ☑ Expose custom metrics
- □ Autoscale!

🏢 Architecture





Autoscale!

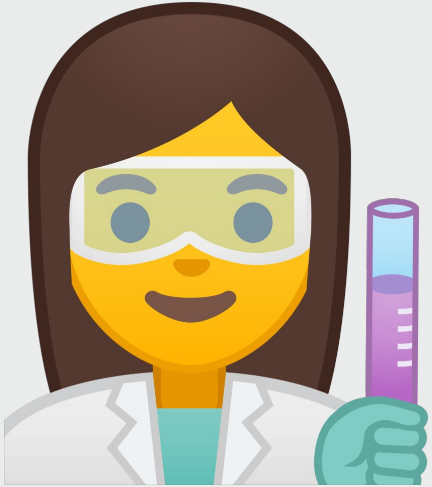
🌡️ What is required?

- ✅ Measure the latency of a service
- ✅ Expose custom metrics
- ✅ Autoscale!

🚏Route Based Scaling

- /
- /vote
- /vote/{editor}/minus
- /vote/{editor}/plus
- /health

🧪 Predictive Scaling



predict_linear(v range-vector, t scalar)

Slides	http://bit.ly/l5d-autoscale
Code	http://bit.ly/kubecon-auto
Get Started!	https://bit.ly/linkerd-get-started
Prometheus Adapter	http://bit.ly/k8s-adapter