Kubernetes

Architecture & Introduction

BERLIN

Dr. Stefan Schimanski, Sergiusz Urbaniak





Who are we?

Engineers @ Mesosphere

Working on Kubernetes-on-Mesos

https://github.com/kubernetes/kubernetes/tree/master/contrib/mesos



Stefan @the1stein sttts@github



Sergiusz @mieszkoman s-urbaniak@github

Who/What is Mesosphere?

The company behind:

- Apache Mesos (https://mesos.apache.org/)
 A distributed systems kernel.
- Marathon (https://mesosphere.github.io/marathon/)
 An Apache Mesos framework
 for long-running applications.
- Chronos (http://mesos.github.io/chronos/)
 A distributed cron replacement.

DCOS

(https://mesosphere.com/)

A data center operating system.

Next 45 Minutes

- Archeology: before and without Kubernetes
- Deployment: kube-up, DCOS, GKE
- Core Architecture: the apiserver, the kubelet and the scheduler
- Compute Model: the pod, the service and the controller

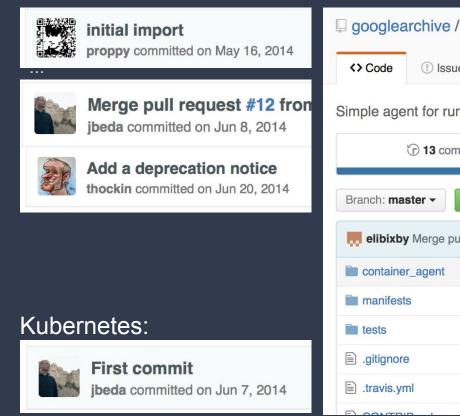
No whitepaper talk. We dive deep!

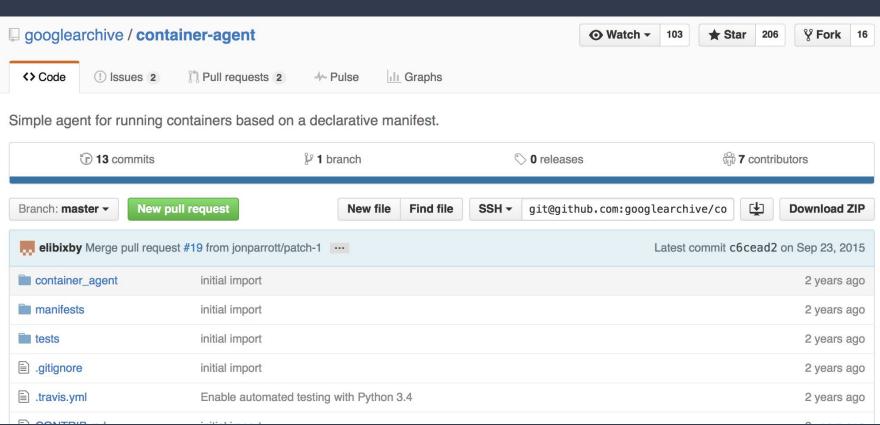


googlearchive / container-agent

How it all started

Python based agent to deploy declarative pods using Docker.





googlearchive / container-agent

A Pod was described as a ...

[...] container group defined by the manifest to share:

- Network Namespaces
- Volumes

This creates a runtime environment where:

- Containers can connect to a service running in other containers of the same group using localhost and a fixed port.
- Containers of the same group can't run services on the same ports.
- Containers of the same group can mount shared volumes defined in the manifest.

A Pod was described as a ...



[...] container group defined by the manifest to share:

- Network Namespaces
- Volumes

This creates a runtime environment where:

- Containers can connect to a service running in other containers of the same group using localhost and a fixed port.
- Containers of the same group can't run services on the same ports.
- Containers of the same group can mount shared volumes defined in the manifest.

Back to the roots

A Pod in Docker terms:

\$ docker run -it --name mypod busybox sleep 9999999

\$ docker run -it --ipc=container:mypod --net=container:mypod --v /tmp/mypod:/tmp/mypod

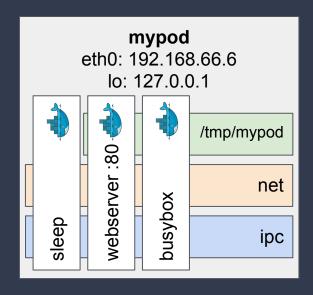
-w /tmp/mypod ubuntu python3 -m http.server 80

"root", "pause" or "infrastructure" container

"Webserver"

\$ docker run -it --ipc=container:mypod --net=container:mypod -v /tmp/mypod:/tmp/mypod sttts/busybox-curl /bin/sh "busybox", our worker

A Pod in Docker terms - visually



• **Note:** different OSes, resource constraints, restart policies

A little Twitter Mirror Pod

```
http 192.168.66.6:80
                                            mypod
                                        eth0: 192.168.66.6
                                           /tmp/mypod
                                          webserver:80
                                                        net
                                      sleep
                                                         ipc
curl -k 'https://mobile.twitter.com/search?q=%23kubernetes%20%23berlin&s=typd' \
```

```
done
```

> /tmp/mypod/index.html

while true; do

sleep 10

Get a Kubernetes cluster: mesos/docker with kube-up

- \$ git clone git@github.com:kubernetes/kubernetes.git
- \$ cd kubernetes
- \$ build/run.sh hack/build-go.sh && make
- \$ KUBERNETES_PROVIDER=mesos/docker cluster/kube-up.sh
- \$ alias kubectl=_output/local/bin/darwin/amd64/kubectl
- \$ kubectl get pods

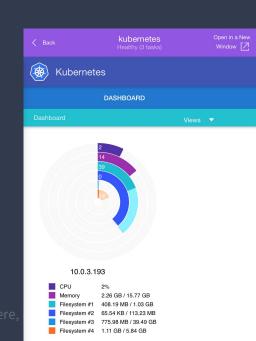
Package for Mesosphere's DCOS

\$ dcos config prepend package.sources \

https://github.com/mesosphere/multiverse/archive/version-1.x.zip

- \$ dcos package install kubernetes
- \$ dcos kubectl create -f nginx.yml pods/nginx
- \$ dcos kubectl get pods

NAME READY STATUS RESTARTS AGE nginx 1/1 Running 0 1m



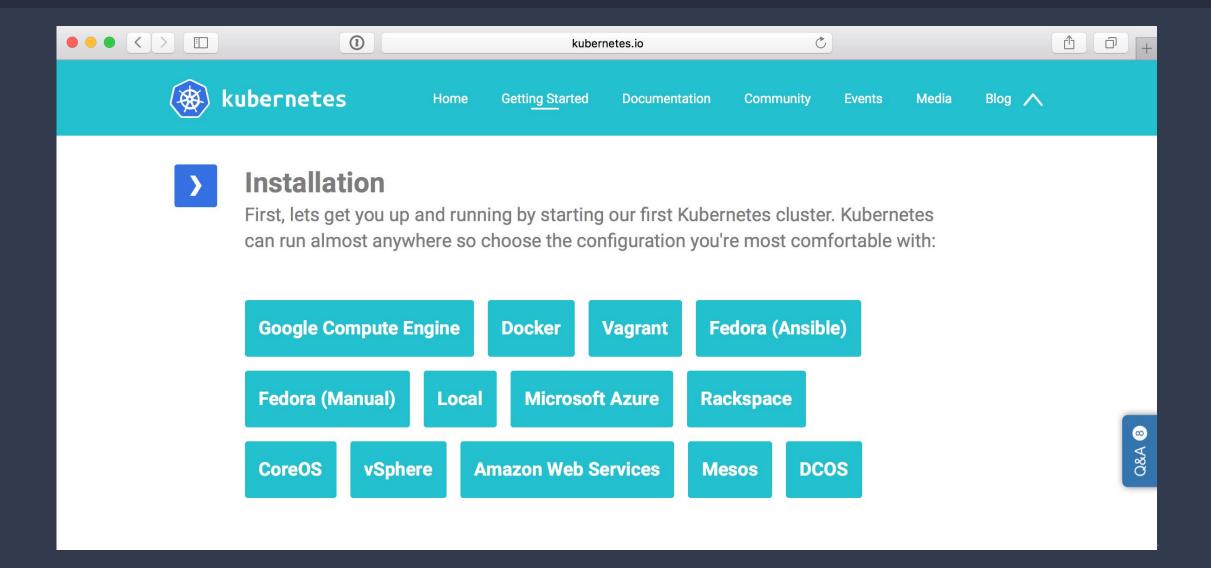
Google CKontainer Engine

- install the gcloud cli
- prepare your Google Cloud account with a project with enabled GCE APIs, e.g. "kube-test-1154"

- \$ gcloud components install kubectl
- \$ gcloud container clusters create demo-cluster --num-nodes=2
- \$ gcloud config set container/cluster demo-cluster
- \$ kubectl get pods

```
NAME READY STATUS RESTARTS AGE nginx 1/1 Running 0 1m
```

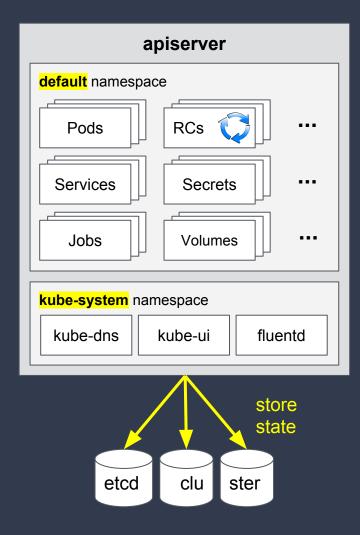
Getting Started Guides



Get a Kubernetes cluster: mesos/docker with kube-up

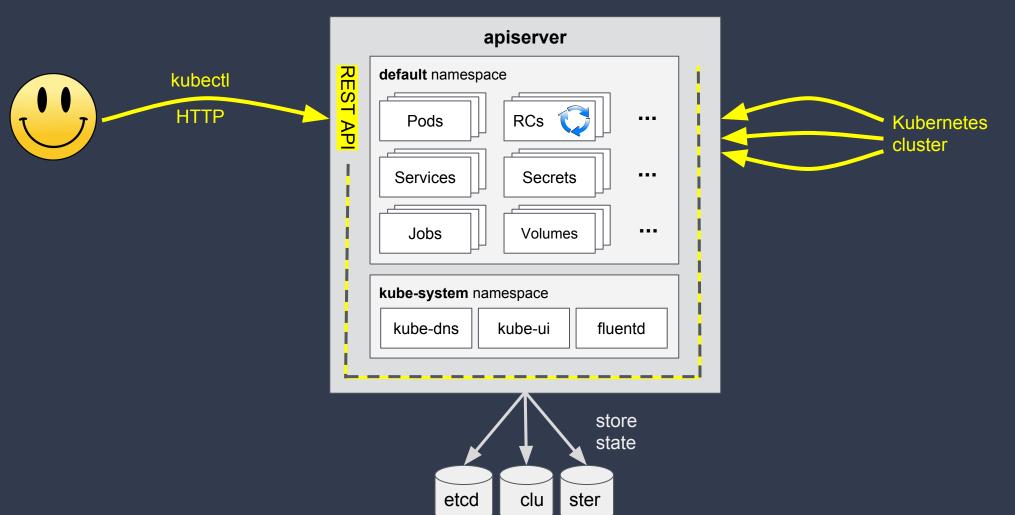
- \$ git clone git@github.com:kubernetes/kubernetes.git
- \$ cd kubernetes
- \$ build/run.sh hack/build-go.sh && make
- \$ KUBERNETES_PROVIDER=mesos/docker cluster/kube-up.sh
- \$ alias kubectl=_output/local/bin/darwin/amd64/kubectl
- \$ kubectl get pods

Kubernetes Core Architecture - the apiserver

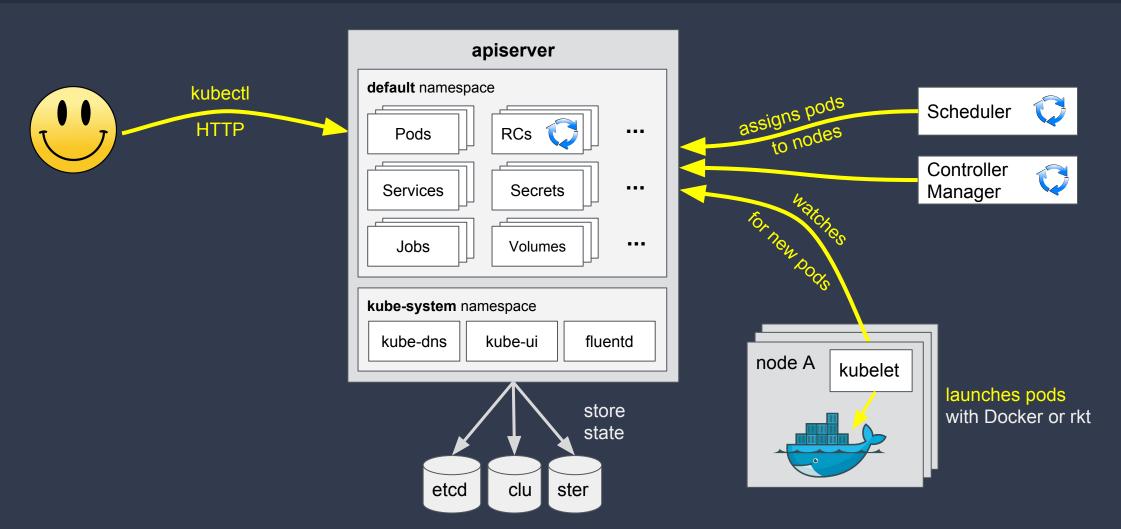


- all the cluster state
- no logic, only storage and API

Kubernetes Core Architecture - the apiserver



Kubernetes Core Architecture - the components



kubectl - interacting with a Kubernetes cluster

```
Usage:
 kubectl [flags]
 kubectl [command]
Available Commands:
 get
                 Display one or many resources
                 Show details of a specific resource or group of resources
 describe
                 Create a resource by filename or stdin
 create
 replace
                 Replace a resource by filename or stdin.
 patch
                 Update field(s) of a resource by stdin.
 delete
                 Delete resources by filenames, stdin, resources and names, or by resources and label selector.
                 Edit a resource on the server
 edit
                 Apply a configuration to a resource by filename or stdin
  apply
                 SUPERSEDED: Set and view the current Kubernetes namespace
 namespace
                 Print the logs for a container in a pod.
  logs
 rolling-update Perform a rolling update of the given ReplicationController.
 scale
                 Set a new size for a Replication Controller.
                 Attach to a running container.
  attach
                 Execute a command in a container.
  exec
                 Forward one or more local ports to a pod.
 port-forward
                 Run a proxy to the Kubernetes API server
  proxy
                 Run a particular image on the cluster.
  run
                 Deprecated: Gracefully shut down a resource by name or filename.
  stop
 #exposehelstein@miesTakema replication controller, service or pod and expose it as ao newMKubernetesWSerwicerved.
  autoscale
                 Auto-scale a replication controller
```

19

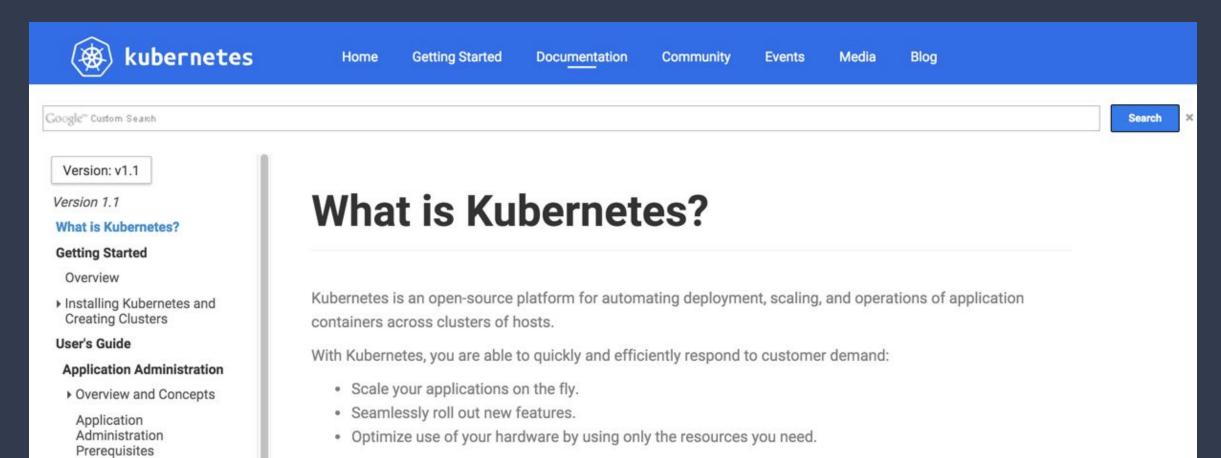
Same Pod in Kubernetes

```
apiVersion: v1
kind: Pod
metadata:
  name: mypod
spec:
  containers:
  - name: webserver
    image: ubuntu
    workingDir: /tmp/mypod
    command: ["python3", "-m", "http.server", "80"]
    volumeMounts:
    - name: htdocs
      mountPath: /tmp/mypod
  - name: busybox
    image: sttts/busybox-curl
    tty: true
    stdin: true
    volumeMounts:
    - name: htdocs
      mountPath: /tmp/mypod
  volumes:
  - name: htdocs
    hostPath:
 path: /tmp/mypod
#k8sber@the1stein@mieszkoman
```

\$ kubectl create -f mypod.yaml pod "mypod" created

Documentation

Documentation landing page: http://kubernetes.io/v1.1



Our goal is to foster an acceptant of companents and tools that relieve the hurden of running applications in

Documentation - Schema

JSON/YAML/REST schema: http://kubernetes.io/v1.1/docs/api-reference/v1/definitions.html

Persistent Volumes

Multi-tier Applications

Examples

▶ Setup and Configuration Updating Live Pods

Reference

API Reference

- ▶ API Basics
- ▶ Kubernetes API Reference

Operations

Definitions

Extensions API:

Extensions:

Operations

Extensions:

Definitions

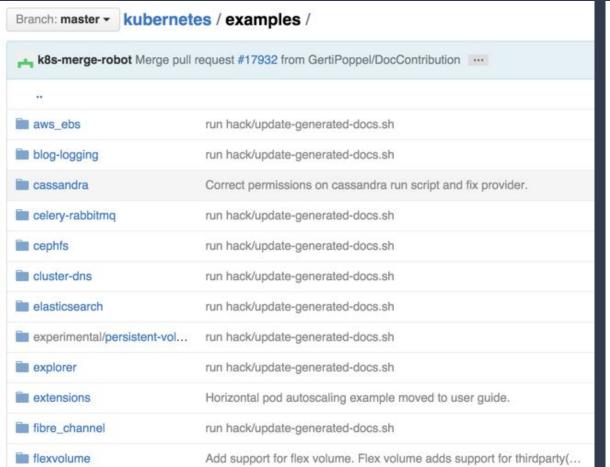
v1.Pod

Pod is a collection of containers that can run on a host. This resource is created by clients and scheduled onto hosts.

Name	Description	Required	Schema	Default
kind	Kind is a string value representing the REST resource this object represents. Servers may infer this from the endpoint the client submits requests to. Cannot be updated. In CamelCase. More info: http://kubernetes.io/v1.1/docs/devel/apiconventions.html#types-kinds	false	string	
apiVersion	APIVersion defines the versioned schema of this representation of an object.	false	string	

Examples - learn from them

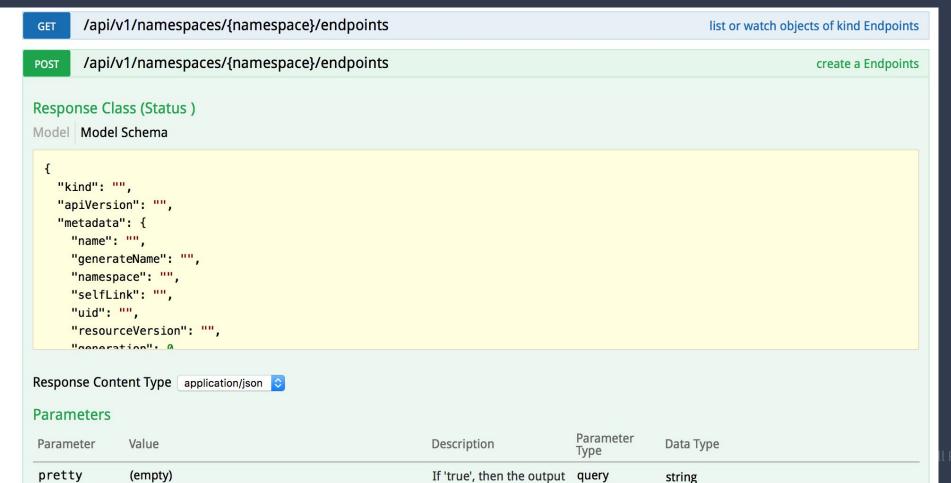
Deployment examples: https://github.com/kubernetes/kubernetes/tree/master/examples



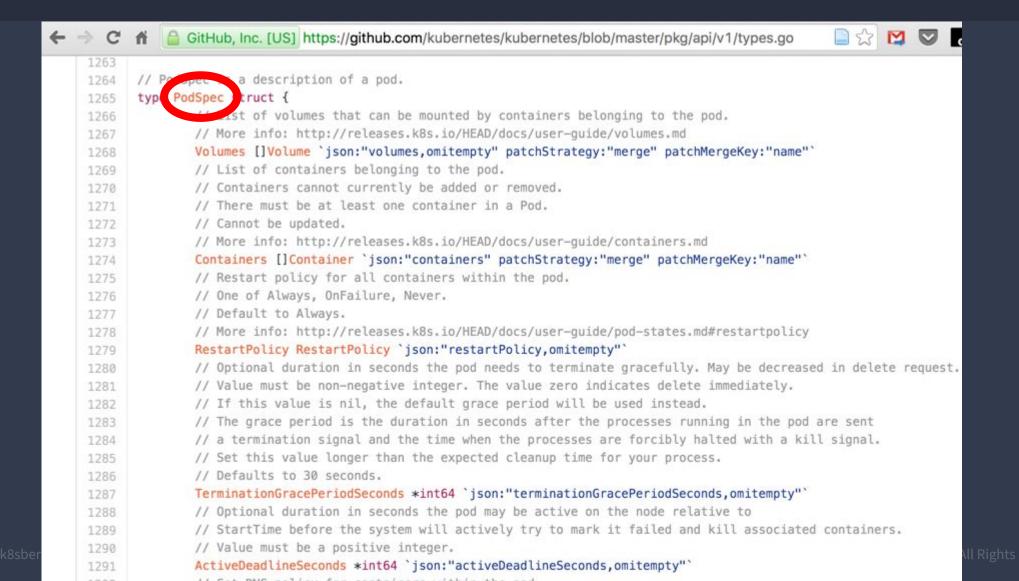
```
Branch: master - kubernetes / examples / guestbook / all-in-one / guestbook-all-in-one.yaml
Inzichang Update guestbook example according to config best practices
1 contributor
177 lines (176 sloc) 3.93 KB
                                                                                                  Raw
       apiVersion: v1
       kind: Service
       metadata:
         name: redis-master
         labels:
           app: redis
           tier: backend
           role: master
  10
           # the port that this service should serve on
  11
  12
         - port: 6379
  13
           targetPort: 6379
         selector:
  15
           app: redis
           tier: backend
  16
  17
           role: master
       apiVersion: v1
       kind: ReplicationController
       metadata:
         name: redis-master
```

Browsing the API

Open https://192.168.65.65:6443/swagger-ui/



The source of truth: api/v1/types.go



On the wire (JSON only)

"secret": {

```
$ curl -k -u admin:admin \
"kind": "Pod",
"apiVersion": "v1",
                                                         https://192.168.65.39:6443/api/v1/namespaces/default/pods/mypod
  "name": "mypod",
  "namespace": "default",
  "selfLink": "/api/v1/namespaces/default/pods/mypod"
  "uid": "37d657dc-b47a-11e5-938a-0242c0a84127",
  "resourceVersion": "5792",
  "creationTimestamp": "2016-01-06T13:34:32Z",
    "k8s.mesosphere.io/bindingHost": "192.168.65.44",
    "k8s.mesosphere.io/executorId": "cf3cd4adb282b475_k8sm-executor",
    "k8s.mesosphere.io/offerId": "20160106-081826-641837248-5050-1-03680",
    "k8s.mesosphere.io/slaveId": "20160106-081826-641837248-5050-1-S1",
    "k8s.mesosphere.io/taskId": "pod.380df08e-b47a-11e5-8201-0242c0a8412b"
},
"spec": {
      "name": "htdocs",
        "path": "/tmp/mypod"
#k8sber"@ame":te"defaults-tokennb3a3b",
```

Labels & Annotations

• in metadata block of every API resource

```
"labels": {
    "git": "cb52d79578b1379e32b7e6a119ed16232ef1b13b",
    "env": "prod",
    "sla": "super-premium",
    "app": "webshop",
    "tier": "frontend"
}
```

```
"annotations": {
    "k8s.mesosphere.io/bindingHost": "192.168.65.44",
    "k8s.mesosphere.io/executorId": "k8sm-executor",
    ...
}
```

labels can be used to filter objects serverside (annotations can't)

\$ kubectl get nodes -l gen=2011

Getting the Pod IP

```
$ kubectl get pod mypod -o yaml
```

. . .

hostIP: 192.168.65.68

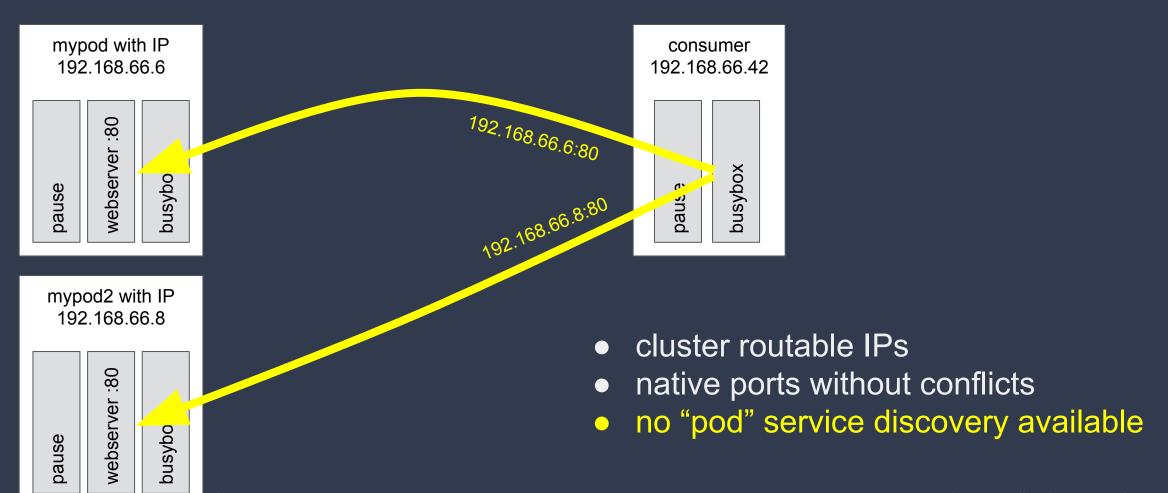
phase: Running

podIP: 192.168.66.11

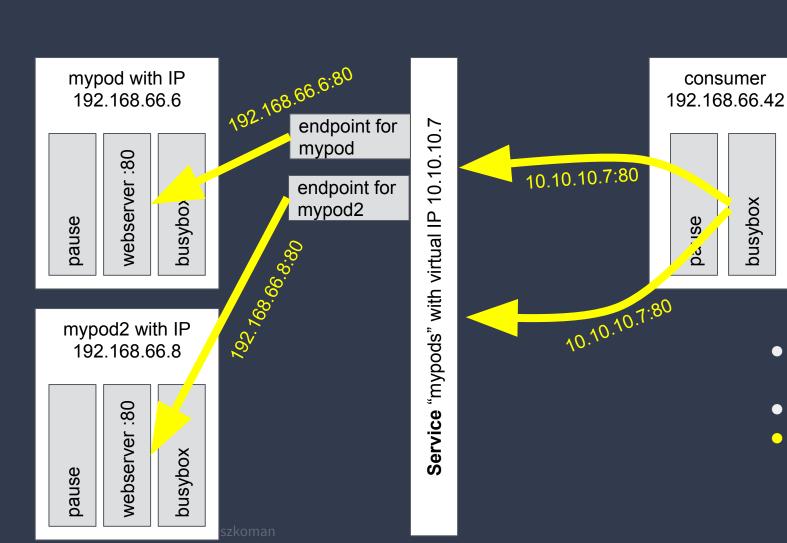
startTime: 2016-01-07T08:33:55Z

\$ kubectl get pod mypod -o template --template='{{.status.podIP}}' 192.168.66.12

From Pod to Pod



A Service

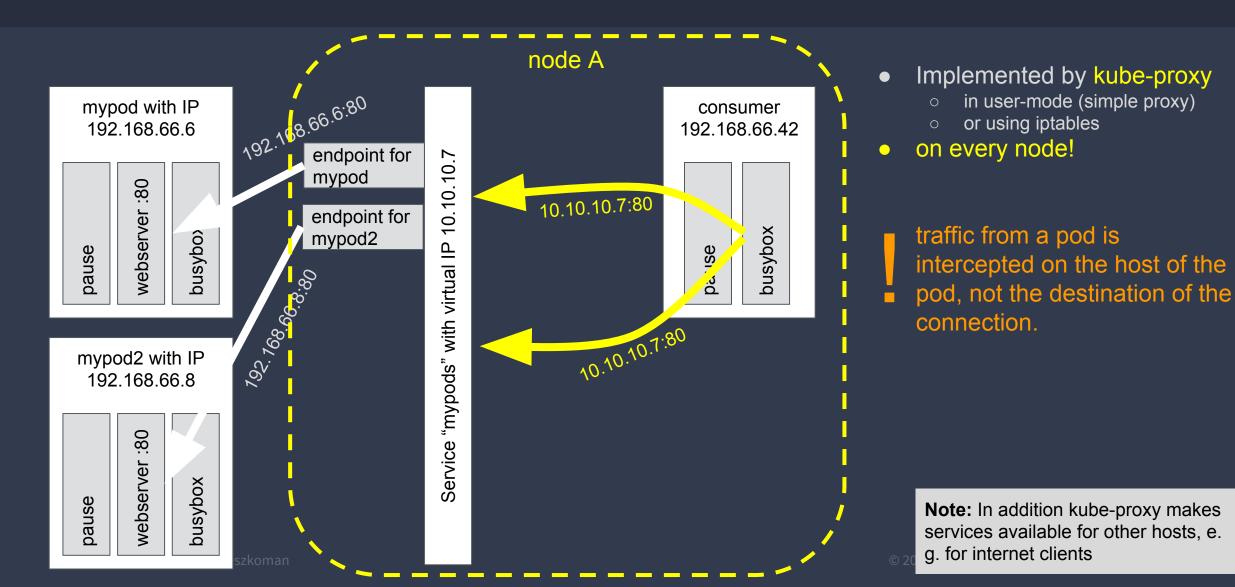


- a service as a static API object
 \$ kubectl create -f service.yaml
- virtual, but static IP (usually 10.x.y.z)
- no service discovery necessary

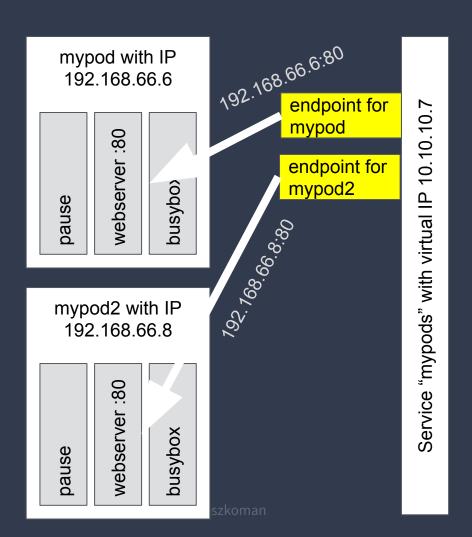
Defining a Service

```
$ cat mypods-service.yaml
apiVersion: v1
kind: Service
metadata:
                                                            match pods with these labels
  name: mypods
                                                            a service without any pod is completely valid
spec:
  ports:
                                                            services exist on their own
  - port: 80
  selector:
    name: mypod
$ kubectl create -f mypods-service.yaml
service "mypods" created
$ kubectl get services
                               EXTERNAL IP
NAME
                 CLUSTER IP
                                             PORT(S)
                                                          SELECTOR
                                                                        AGE
k8sm-scheduler
                 10.10.10.34
                                             10251/TCP
                                                                        6h
                               <none>
                                                          <none>
kubernetes
                 10.10.10.1
                                             443/TCP
                                                                        6h
                               <none>
                                                          <none>
                 10.10.10.32
                                              80/TCP
mypods
                                                          name=mypods
                                                                        5m
                               <none>
```

A Service, available on each node



Service Endpoints



- 1 endpoint for each matching pod of a service
- endpoint = pod-IP:port, e.g. 192.168.66.6:80

\$ kubectl get endpoints kube-ui

NAME ENDPOINTS AGE mypods 192.168.65.68:8001 2s

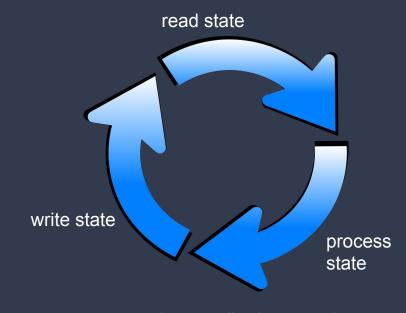
Special in Kuberentes-Mesos because it runs without overlay network here. Otherwise, :80 would be visible here.

Behind the scenes - Linux iptables

```
$ iptables -t nat -n -L
Chain KUBE-PORTALS-CONTAINER (1 references)
                              destination
target prot opt source
REDIRECT tcp -- 0.0.0.0/0
                                10.10.10.3
                                               /* kube-system/kube-ui: */ tcp dpt:80 redir ports 45123
                                                /* default/k8sm-scheduler: */ tcp dpt:10251 redir ports 59722
REDIRECT tcp -- 0.0.0.0/0
                               10.10.10.181
REDIRECT tcp -- 0.0.0.0/0
                               10.10.10.1
                                               /* default/kubernetes: */ tcp dpt:443 redir ports 55836
REDIRECT udp -- 0.0.0.0/0
                                10.10.10.10
                                                /* kube-system/kube-dns:dns */ udp dpt:53 redir ports 57016
REDIRECT tcp -- 0.0.0.0/0
                                10.10.10.10
                                               /* kube-system/kube-dns:dns-tcp */ tcp dpt:53 redir ports 42340
```

State plus Control

- state consists of all API objects, stored on the API-Server
- core Kubernetes architecture principle: logic is in control loops ("controllers") which
 - are state-less (can recover from failure)
 - and decoupled (communication via API-Server)



A Control Loop to run Pods

INPUT:

REPLICAS: number of mypods that should be running

ALGORITHM

- 1. running:= number of running pods with label name=mypod
- 2. If running > REPLICAS => delete some mypods
- 3. If REPLICAS > running => launch some mypods
- 4. Goto 1

Getting # of running mypods

```
REPLICAS = int(sys.argv[1])
master = os.getenv("KUBERNETES_MASTER", "http://localhost:8080")
kubeclient = toolkit.KubeHTTPClient(api_server=master, debug=False)
while True:
    # count mypods
    response = kubeclient.execute_operation(method="GET",
       ops path="/api/v1/namespaces/default/pods?labelSelector=name%3Dmypod")
    mypods = response.json()['items']
    running_mypods = len(mypods)
    print "{} running".format(running mypods)
    . . .
```

If running > REPLICAS => delete some mypods

```
if running_mypods > REPLICAS:
    to_delete = running_mypods - REPLICAS
    print " Too many are running. Deleting {} pods:".format(to_delete)
    for pod in mypods[:to_delete]:
        print " Deleting pod {}".format(pod['metadata']['name'])
        kubeclient.delete_resource(pod['metadata']['selfLink'])
```

If REPLICAS > running => launch some mypods

```
elif REPLICAS > running_mypods:
       to launch = REPLICAS - running mypods
       for n in range(0, to_launch):
           mypod_spec, _ = util.load_yaml(filename="mypod.yaml")
            mypod_spec["metadata"]["name"] += "-" + shortuuid.uuid()[:4].lower()
            print "Launching pod {}".format(mypod_spec["metadata"]["name"])
            response = kubeclient.execute operation(method='POST',
                ops_path = "/api/v1/namespaces/default/pods",
                payload = util.serialize_tojson(mypod_spec))
```

Kubernetes Replication Controllers - "RCs"

```
apiVersion: v1
kind: ReplicationController
metadata:
  name: mypod
spec:
  replicas: 10
  selector:
    name: mypod
  template:
    metadata:
      labels:
        name: mypod
    spec:
      containers:
      - name: webserver
        image: ubuntu
```

- \$ kubectl create -f mypod-rc.yaml
- \$ kubectl scale --replicas=5 rc/mypod

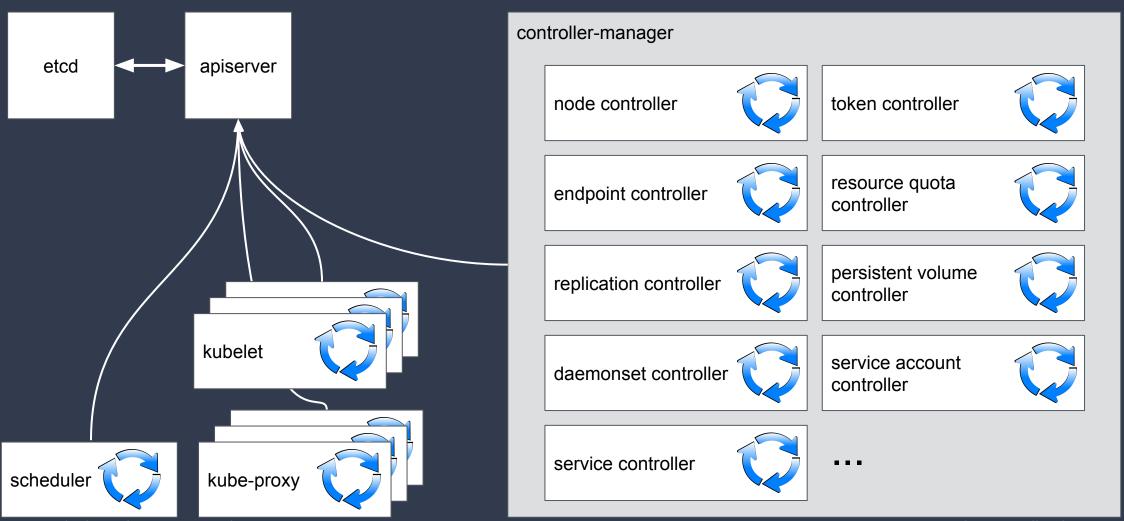
Other Controllers

```
apiVersion: extensions/v1beta1
kind: DaemonSet
metadata:
  name: mypod
spec:
  selector:
    name: mypod
  template:
    metadata:
      labels:
        name: mypod
    spec:
      nodeSelector:
        env=dmz
      containers:
      - name: webserver
        image: ubuntu
```

DaemonSets run pods once per node

- MyCassandraController might run Cassandra instances
- MyGaleraController might run MySQL Galera instances
- MyHaproxyUpdateController might write haproxy.conf
- MyNodeUpdateController might do rolling security updates on nodes or of containers
- MyBitcoinScheduler might schedule bitcoin pods for spare resources

Controllers, controllers everywhere ...



Of course: we are hiring in Hamburg, Berlin & San Francisco!



Backup

Watches

```
$ kubectl get pods -l "name=mypod" -w --v=9
=> GET <u>https://192.168.65.39:6443/api/v1/watch/namespaces/default/pods?</u>
<u>labelSelector=name%3Dmypod</u>
{"type":"ADDED","object":{"kind":"Pod","apiVersion":"v1", ...}}
{"type":"MODIFIED","object":{"kind":"Pod","apiVersion":"v1", ...}}
{"type":"DELETED","object":{"kind":"Pod","apiVersion":"v1",...}}
```