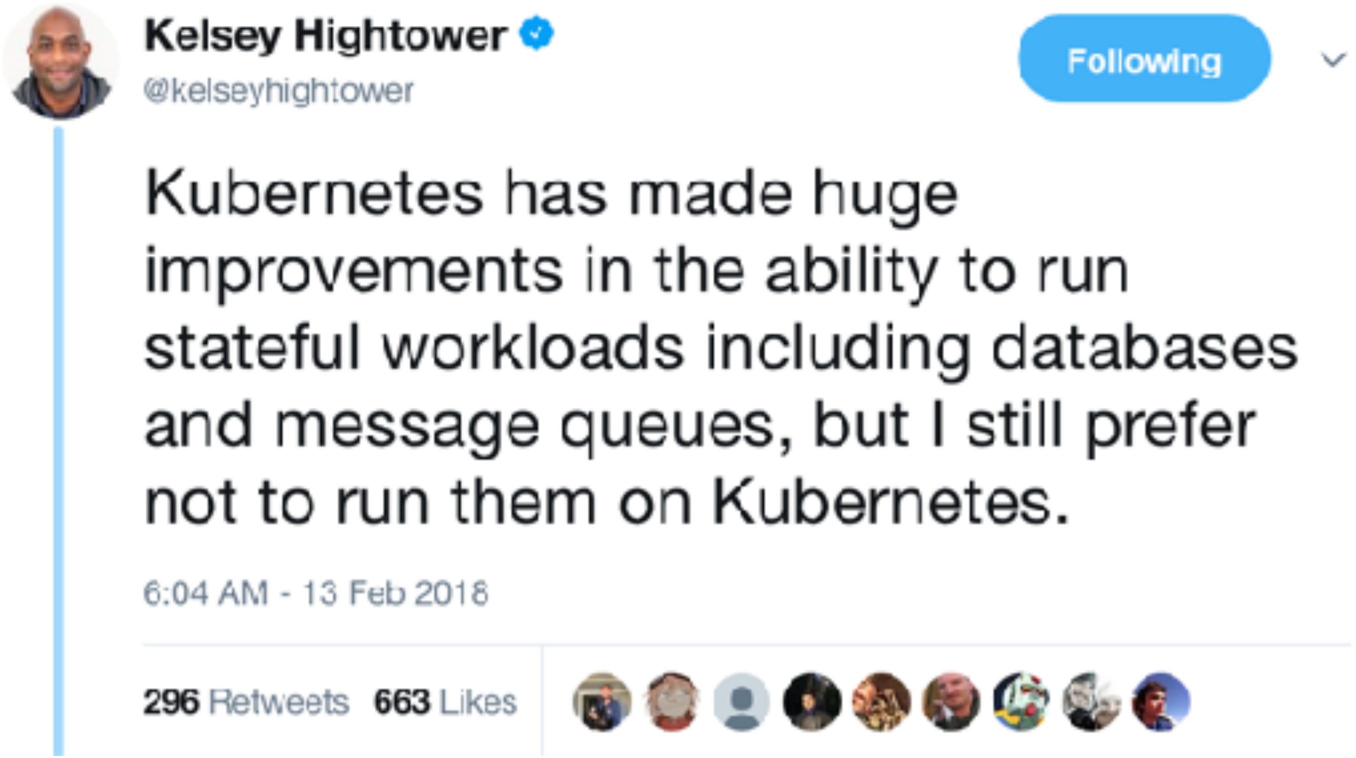
沃趣 WOQU TECH

Is Kubernetes ready for statefulset workloads?

杭州沃趣科技股份有限公司
Hangzhou WOQU Technology Co., Ltd.

Let Data Drive!



Kelsey Hightower
@kelseyhightower

Kubernetes has made huge improvements in the ability to run stateful workloads including databases and message queues, but I still prefer not to run them on Kubernetes.

6:04 AM - 13 Feb 2018

296 Retweets **663** Likes

思考

- Kubernetes为构建有状态的应用提供了哪些资源?
- 基于这些资源到底能不能将有状态应用（如数据库）的运行到kubernetes?

Contents

- What kubernetes offer?
- How to build statefulset workloads like database?
- The problems we are facing and thinking

What kubernetes offer?

Persistent Volume

- Pod volatile
- PV
  - 持久化
  - VolumePlugin
- PVC
  - 解耦存储细节
- StorageClass
  - 定义不同类型不同规格的存储类
- Provisioner
  - 动态提供存储资源
  - 扩展存储类型

```
apiVersion: v1
kind: Pod
metadata:
  name: mysql
spec:
  volumes:
  - name: mysql-data
    gcePersistentDisk:
      pdName: mysql
      fsType: nfs4
  containers:
  - image: mysql
    name: mysql
    volumeMounts:
    - name: mysql-data
      mountPath: /var/lib/mysql
```

```
apiVersion: v1
kind: Pod
…
  volumes:
  - name: mysql-data
    persistentVolumeClaim:
      claimName: mysql-pvc
  containers:
  - image: mysql
    name: mysql
    volumeMounts:
    - name: mysql-data
      mountPath: /var/lib/mysql
```

```
apiVersion: v1
kind: PersistentVolumeClaim
metadata:
  name: mysql-pvc
spec:
  resources:
    requests:
      storage: 1Gi
…
```

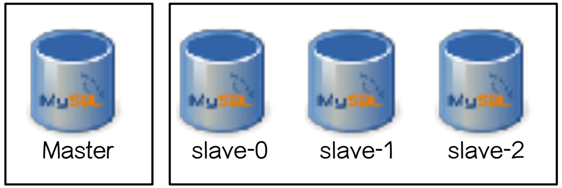
```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: production
provisioner: kubernetes.io/gce-pd
parameters:
  type: pd-ssd
  zone: europe-west1-b
```

```
apiVersion: storage.k8s.io/v1
kind: StorageClass
metadata:
  name: development
provisioner: kubernetes.io/gce-pd
parameters:
  type: pd-hdd
  zone: europe-west1-a
```

What kubernetes offer?

Statefulset

- Ordinal server-id 监控 日志
- Each Pod with a single PersistentVolume
- Stable Network ID
  - slave-0.slave.default.svc.cluster.local
  - slave-1.slave.default.svc.cluster.local
  - slave-2.slave.default.svc.cluster.local
- Created sequentially, in order from {0..N-1}
- Terminated in reverse order, from {N-1..0}.

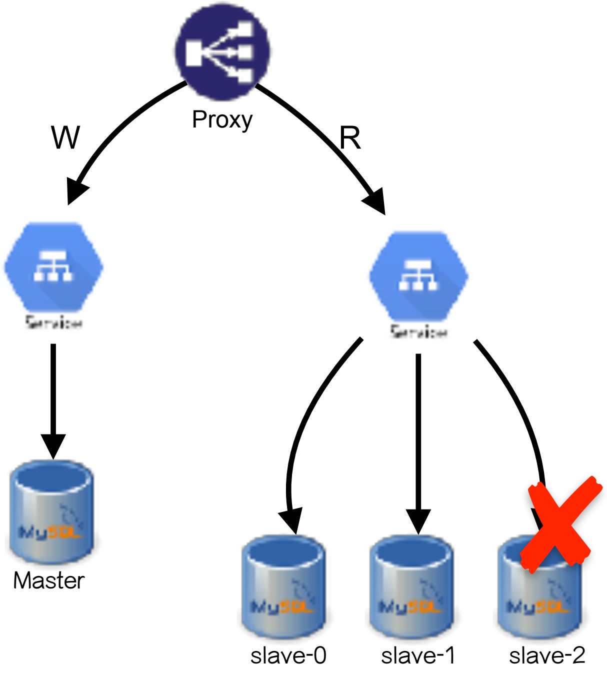


```
apiVersion: apps/v1beta1
kind: StatefulSet
…
spec:
  replicas: 3
  template:
    metadata:
…
    spec:
      containers:
      - name: mysql
        image: mysql
        volumeMounts:
        - name: data
          mountPath: /var/lib/mysql
  volumeClaimTemplates:
  - metadata:
      name: data
    spec:
      resources:
        requests:
          storage: 2Gi
      accessModes:
      - ReadWriteOnce
```

What kubernetes offer?

Service

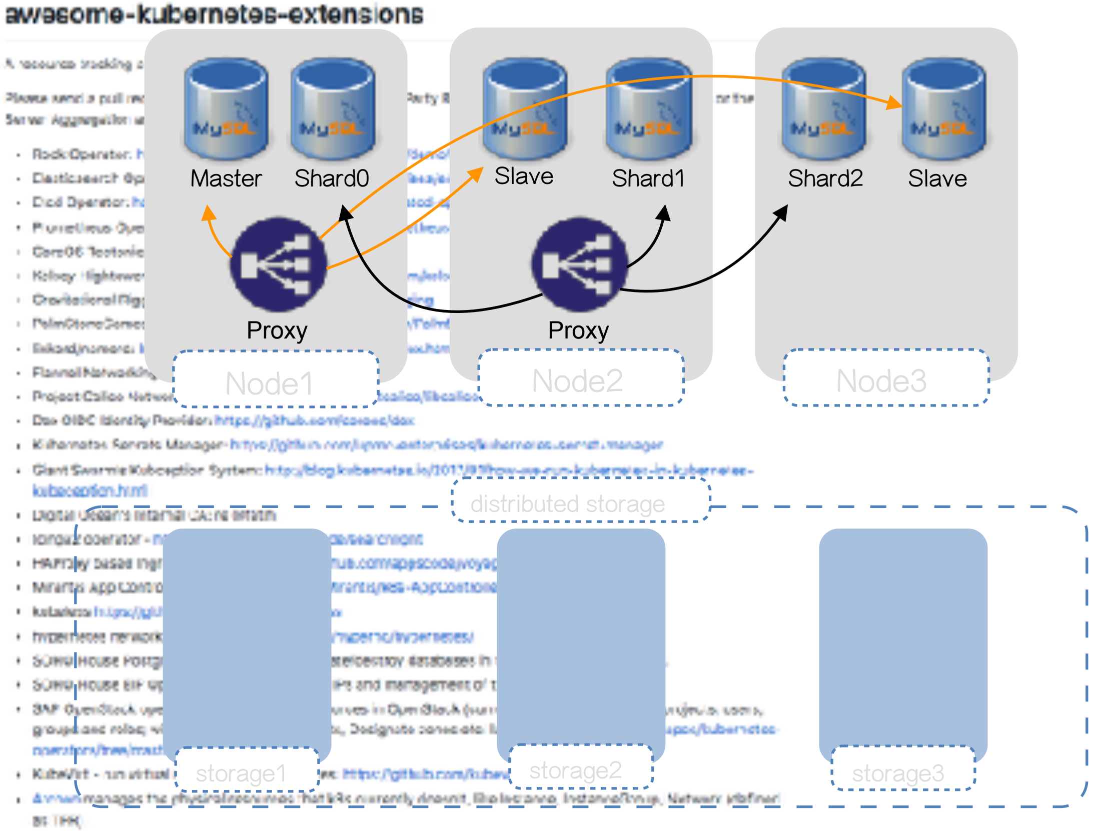
- DNS/IP
- LoadBalance
- ReadinessProbe



Build database workload

- kuberentes
  - statefulset提供数据库集群
  - pvc提供持久存储
  - service提供资源访问
- CRD：抽象资源对象MySQL+Proxy
  - Read/write split
  - sharding
- operator

整体复杂度巨大



Build database workload

- kuberentes
  - statefulset提供数据库集群
  - pvc提供持久存储
  - service提供资源访问
- CRD：抽象资源对象MySQL+Proxy
  - Read/write split
  - sharding
- operator

整体复杂度巨大

Projects

We host and nurture components of cloud-native software stacks, including Kubernetes, Prometheus and Envoy. Kubernetes and other CNCF projects are some of the highest velocity projects in the history of open source. We are regularly adding new projects to better support a full stack cloud-native environment.

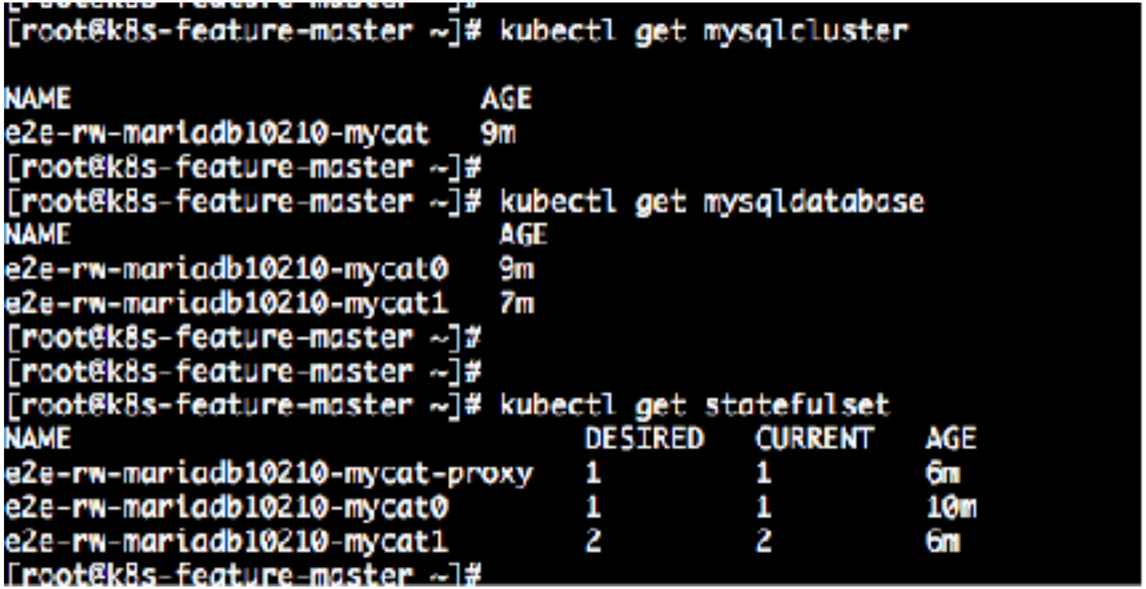
Vitess

Vitess is a database clustering system for horizontal scaling of MySQL through generalized sharding. By encapsulating shard-routing logic, Vitess allows application code and database queries to remain agnostic to the distribution of data onto multiple shards. With Vitess, you can even split and merge shards as your needs grow, with an atomic cutover step that takes only a few seconds. Vitess has been a core component of YouTube's database infrastructure since 2011, and has grown to encompass tens of thousands of MySQL nodes. It's architected to run as effectively in a public or private cloud architecture as it does on dedicated hardware. It combines and extends many important MySQL features with the scalability of a NoSQL database.

Build database workload

- kuberentes
  - statefulset提供数据库集群
  - pvc提供持久存储
  - service提供资源访问
- CRD：抽象资源对象MySQL+Proxy
  - Read/write split
  - sharding
- operator

整体复杂度巨大



```
[root@k8s-feature-master ~]#
[root@k8s-feature-master ~]# kubectl get mysqlcluster

NAME                         AGE
e2e-rw-mariadb10210-mycat    9m
[root@k8s-feature-master ~]#
[root@k8s-feature-master ~]# kubectl get mysqldatabase
NAME                          AGE
e2e-rw-mariadb10210-mycat0    9m
e2e-rw-mariadb10210-mycat1    7m
[root@k8s-feature-master ~]#
[root@k8s-feature-master ~]#
[root@k8s-feature-master ~]# kubectl get statefulset
NAME                               DESIRED   CURRENT   AGE
e2e-rw-mariadb10210-mycat-proxy    1         1         6m
e2e-rw-mariadb10210-mycat0         1         1         10m
e2e-rw-mariadb10210-mycat1         2         2         6m
[root@k8s-feature-master ~]#
```

The problems

RollingUpdate

集群资源更新

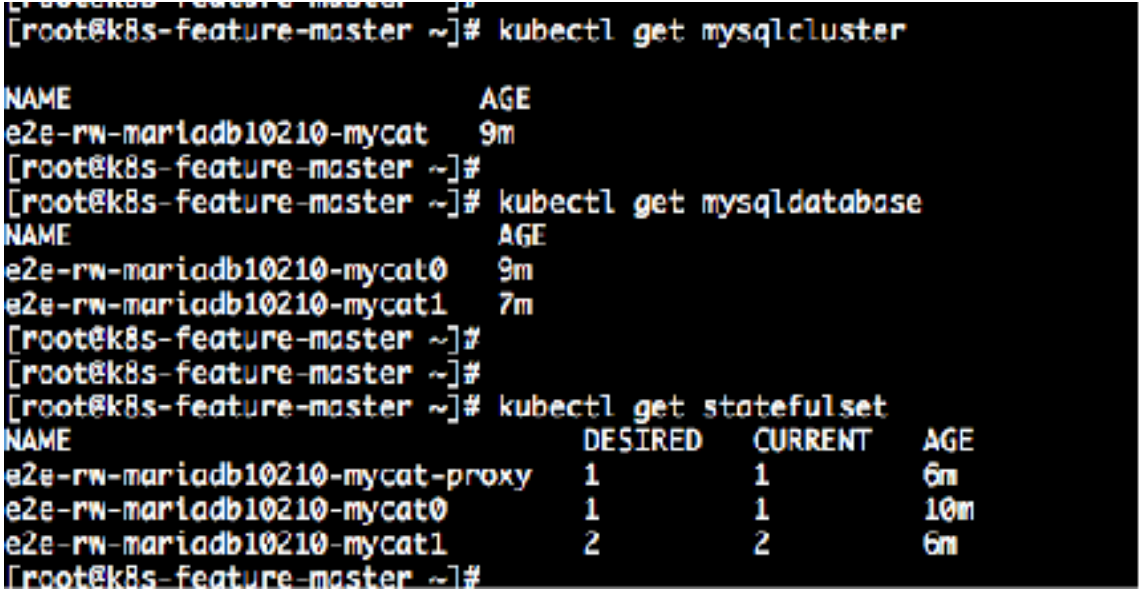
- CPU、memory

问题

- 升级过程中，operator重启
- 如何确认哪些资源已经更新

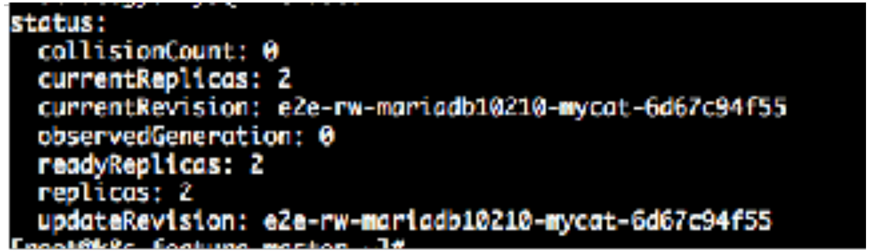
CRD并没有直接提供RollingUpdate的机制

ControllerRevision



```
[root@k8s-feature-master ~]# kubectl get mysqlcluster

NAME                        AGE
e2e-rw-mariadb10210-mycat   9m
[root@k8s-feature-master ~]#
[root@k8s-feature-master ~]# kubectl get mysqldatabase
NAME                         AGE
e2e-rw-mariadb10210-mycat0   9m
e2e-rw-mariadb10210-mycat1   7m
[root@k8s-feature-master ~]#
[root@k8s-feature-master ~]#
[root@k8s-feature-master ~]# kubectl get statefulset
NAME                              DESIRED   CURRENT   AGE
e2e-rw-mariadb10210-mycat-proxy   1         1         6m
e2e-rw-mariadb10210-mycat0        1         1         10m
e2e-rw-mariadb10210-mycat1        2         2         6m
[root@k8s-feature-master ~]#
```



```
status:
  collisionCount: 0
  currentReplicas: 2
  currentRevision: e2e-rw-mariadb10210-mycat-6d67c94f55
  observedGeneration: 0
  readyReplicas: 2
  replicas: 2
  updateRevision: e2e-rw-mariadb10210-mycat-6d67c94f55
[root@k8s-feature-master ~]#
```

沃趣 WOQU TECH

Thanks & QA

杭州沃趣科技股份有限公司
Hangzhou WOQU Technology Co., Ltd.

Let Data Drive!