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
Following
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 2016
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

- PV
  - 持久化
  - VolumePlugin
- PVC
  - 解耦存储细节
- StorageClass
  - 定义不同规格的存储池
- provisioner
  - 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
```

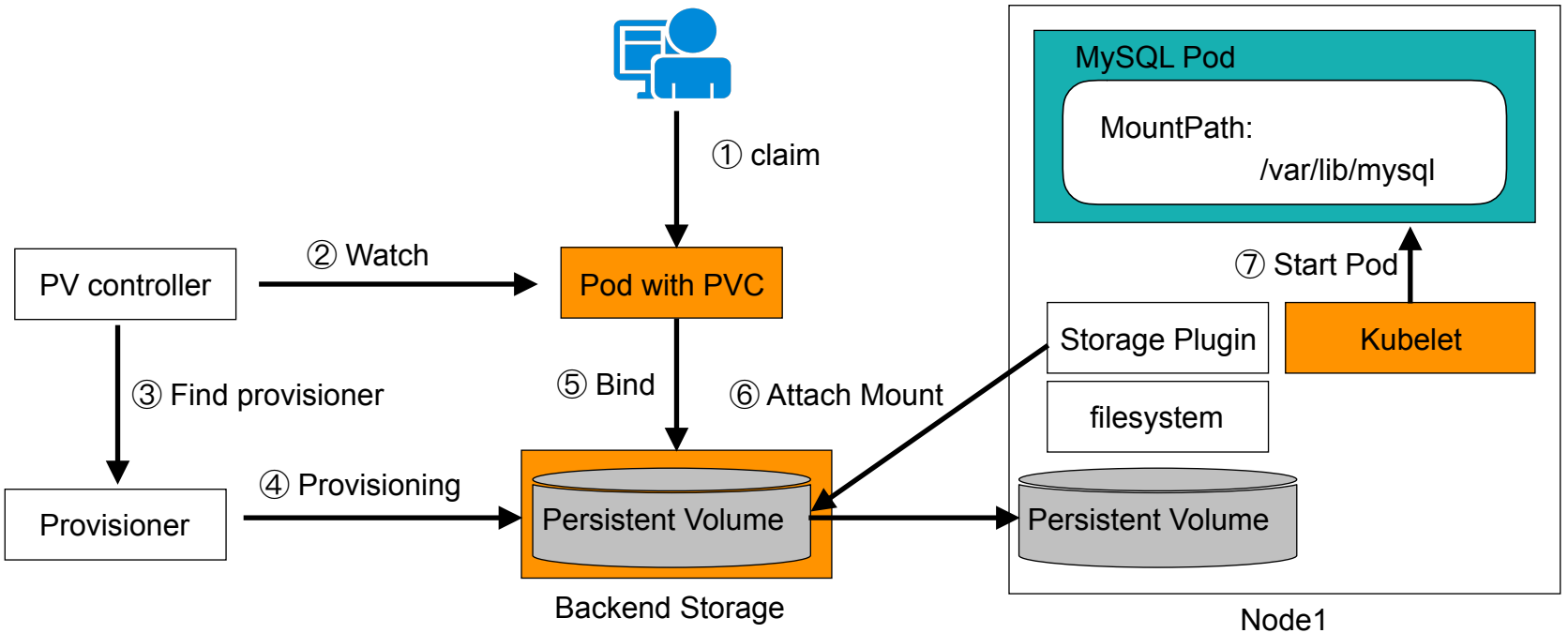
```
apiVersion: v1
kind: Pod
metadata:
  name: mysql
spec:
  volumes:
  - name: mysql-data
    persistentVolumeClaim:
      claimName: mysql-pvc
  containers:
  - image: mysql
    name: mysql
    volumeMounts:
    - name: mysql-data
      mountPath: /var/lib/mysql
```

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

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

What kubernetes offer?

Persistent Volume 挂载过程



What kubernetes offer?

Statefulset

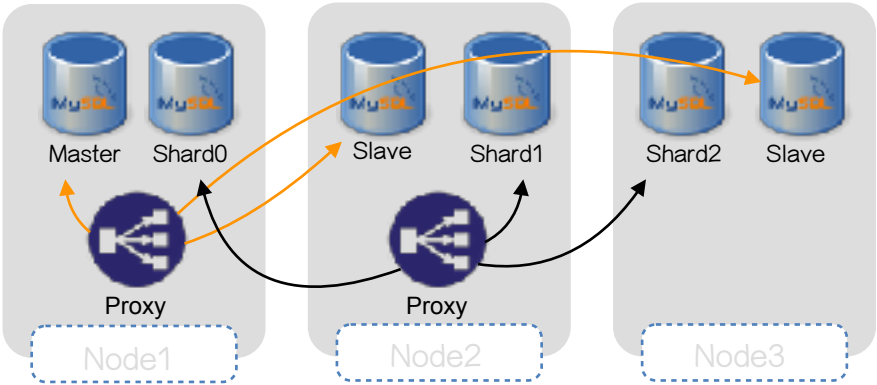
- Ordinal
- Stable Network ID
  - $(statefulset name)-$(ordinal)
- Storage
  - Pod with a single PersistentVolume
- Created sequentially, in order from {0..N-1}
- Terminated in reverse order, from {N-1..0}.

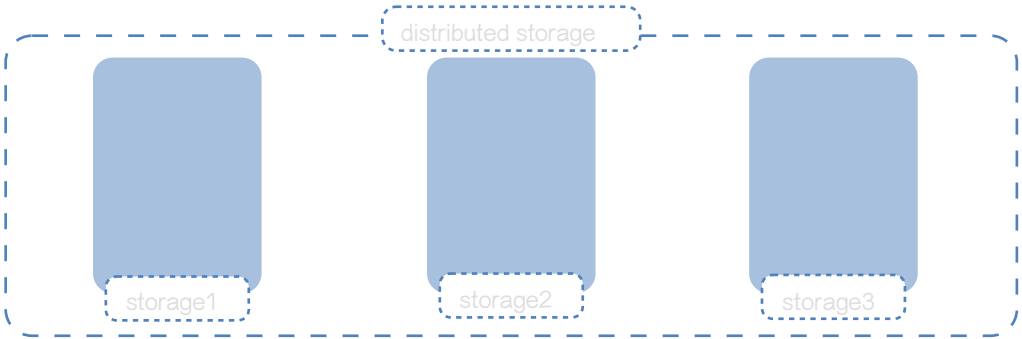
```
apiVersion: apps/v1beta1
kind: StatefulSet
…
spec:
  replicas: 3
  template:
    metadata:
…
    spec:
      containers:
      - name: mysql
        image: mysql
        volumeMounts:
        - name: data
          mountPath: /var/lib/mysql
  volumeClaimTemplates:
  - metadata:
      name: data
    spec:
      resources:
        requests:
          storage: 1Gi
      accessModes:
      - ReadWriteOnce
```

Build database workload

- 计算存储分离架构
  - 计算：worker node
  - 存储：分布式存储
- kuberentes
  - statefulset提供数据库集群
  - pvc提供持久存储
  - service提供资源访问

整体复杂度巨大

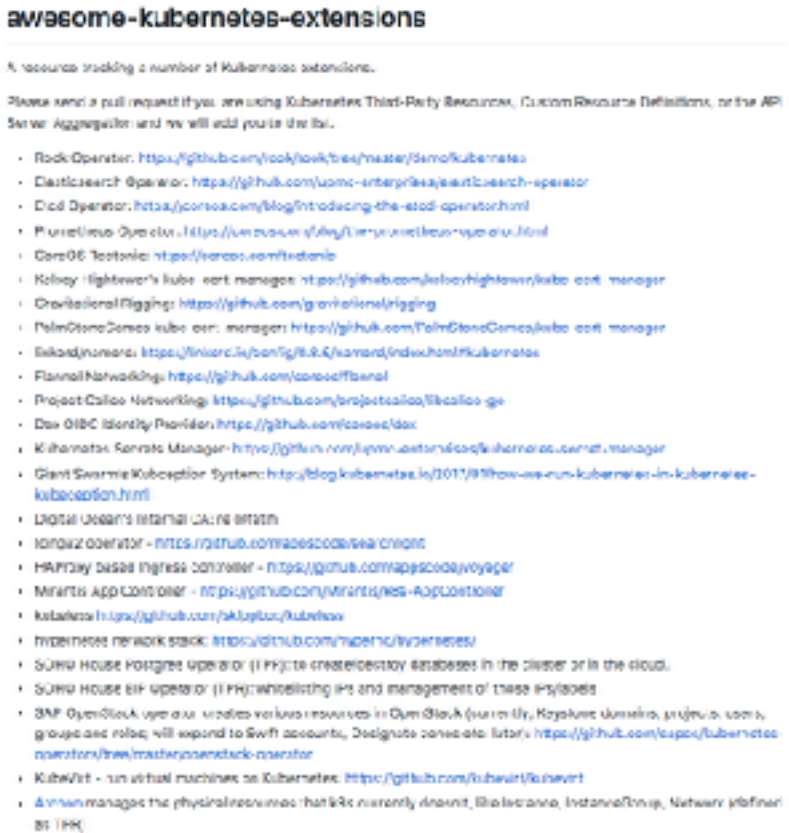




Build database workload

- 计算存储分离架构
  - 计算：worker node
  - 存储：分布式存储
- 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

整体复杂度巨大



Build database workload

- 计算存储分离架构
  - 计算：worker node
  - 存储：分布式存储
- 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

整体复杂度巨大

The problems

The problems

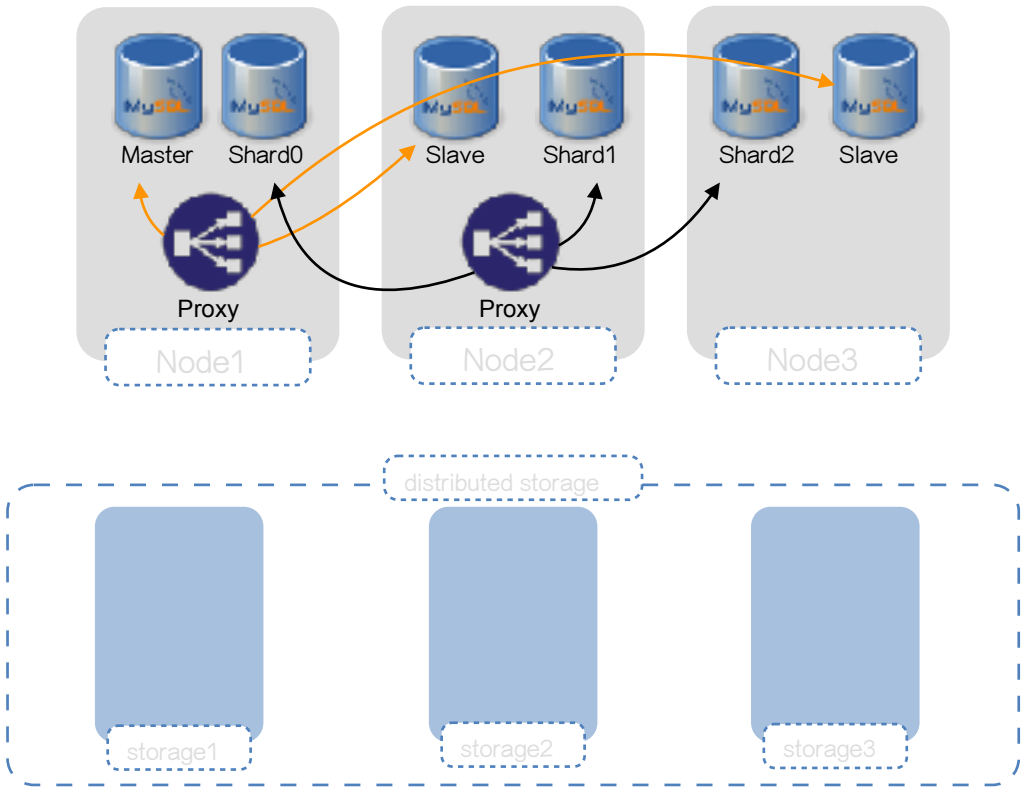
Performance

计算存储分离

- 提高了数据库实例的部署密度和计算资源利用率
- 架构清晰
- 扩展方便

问题

- IO路径更长，网络开销明显
- 无法针对Pod进行限流
- 数据库Latency Sensitive 型应用, 网络延时会极大影响数据库能力(QPS,TPS)

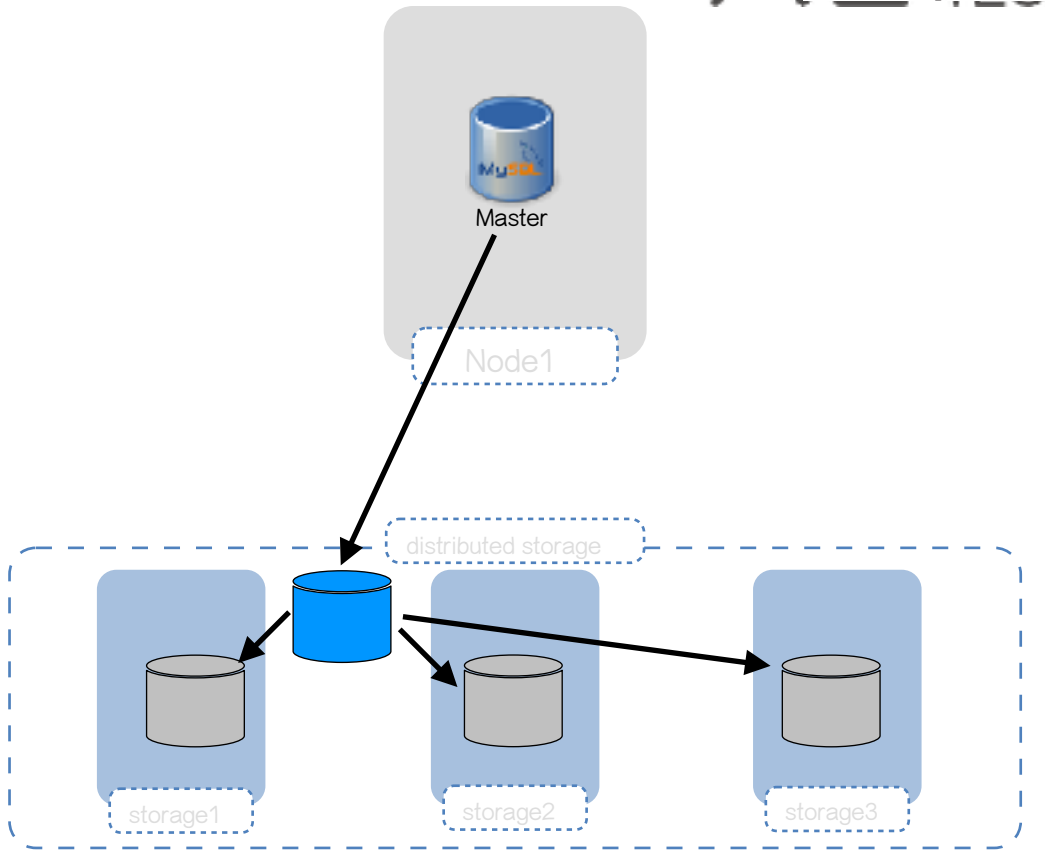


The problems

Performance

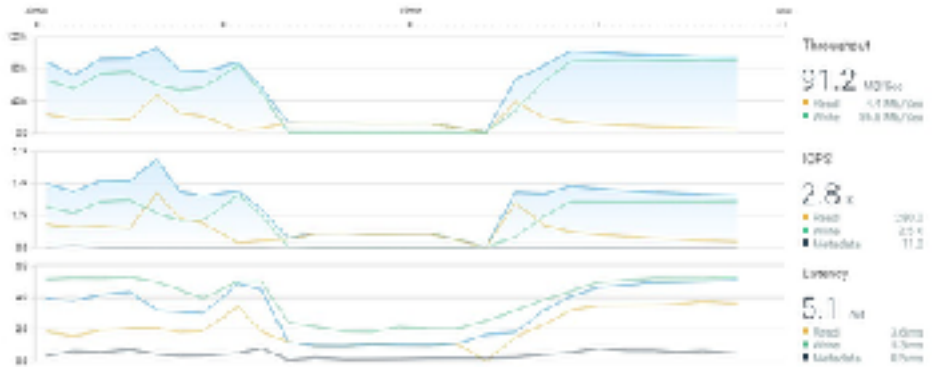
结合MySQL数据库特性的性能优化

- 减少IO发生次数
- 数据库层关闭DoubleWrite
- 文件系统层，支持 Atomic Write

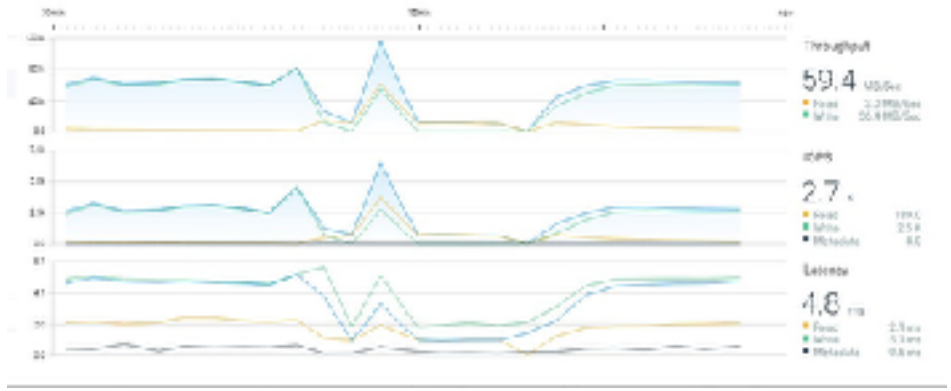


The problems

Performance



Double write on



Double write off

- Sysbench指标
  - TPS ↑28.0892%，QPS ↑28.0893%，RST ↓169.2033%
- 分布式文件系统指标
  - IOPS 提升22.3%
  - Latency 下降 39%
  - 在IOPS 提升22.3%的情况下, Throughput 仅多消耗 3.6%

The problems

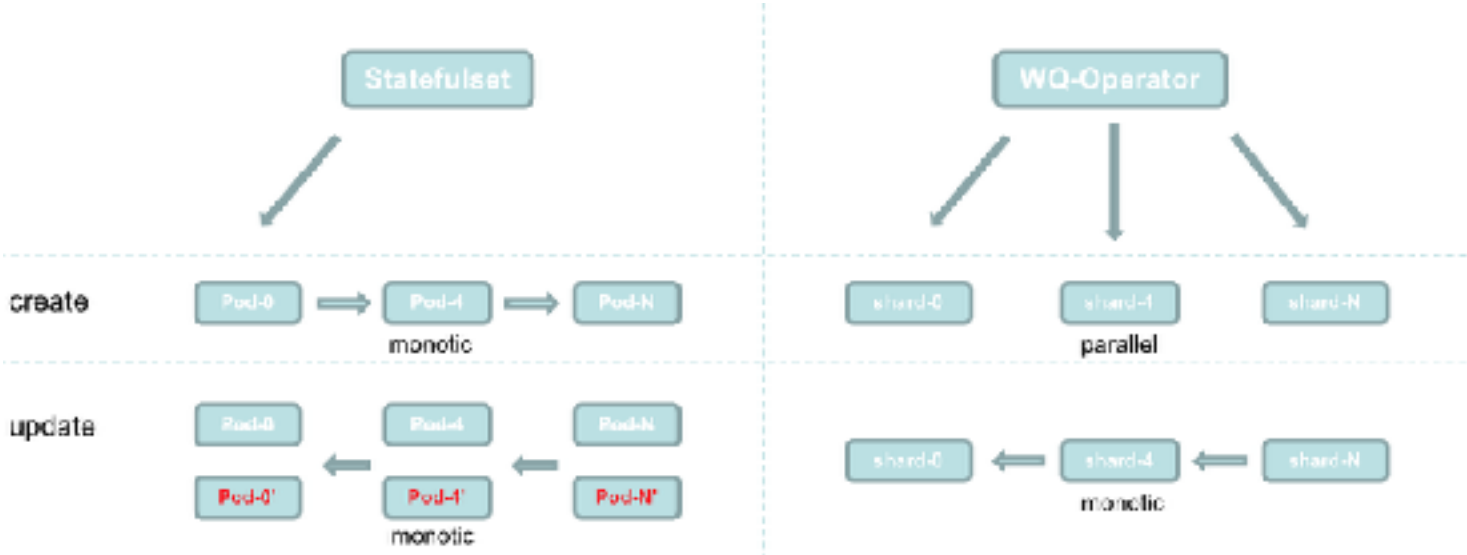
RollingUpdate

集群资源更新

- CPU、memory
- db/proxy config

异常处理

- node reboot
- instance restart
- human mistakes



CRD并没有提供RollingUpdate的机制

ControllerRevision

Summary

- Kubernetes
  - pv/pvc
  - statefulset
  - service
- Build database
- The problems
  - Performance
  - RollingUpdate

Thanks & QA