



# TiDB在Rancher中基于Longhorn块存储快速部署和数据备份还原

Rancher 中国

张文华

2020/7/4

# Agenda

- 使用效果
- Longhorn核心功能介绍
- TiDB Cluster 快速部署
- 数据快照及还原

# TiDB+ Rancher+ Longhorn结合后的使用效果

- 节约资源

在采用 Longhorn 之前，每个物理机平时大概有 70% 的磁盘 IO 能力处于闲置状态。采用 Longhorn 之后，Longhorn 把多余的物理机存储都利用了起来。TiDB 托管到 rancher 容器平台后结合 Longhorn 实现了弹性可扩展。

- 管理和运维更加简单便捷

采用 Longhorn 之后，我们不用再担心某个机器的硬盘故障，或者某个机器的读写负载不均衡的问题。

- 提高写入性能

由于 Longhorn 底层对文件作了副本支持，业务层 TiDB 将副本置为1，把 “max-replicas” 由默认的 3改为1 后由于不用再同步数据了所以写入性能大大提高。如果是Kafka、Elasticsearch 降低副本数，Java的GC 频率明显降低，同时也降低了 FullGC 现象。

- 数据备份恢复更便捷

使用Longhorn UI基于Longhorn的快照技术快速备份和还原测试数据，甚至可以把数据还原到某个时间点。还可以把快照数据增量备份到S3对象存储上，实现跨集群数据迁移。

# Agenda

- 使用效果
- Longhorn核心功能介绍
- TiDB Cluster 快速部署
- 数据快照及还原

# Longhorn V1.0 功能概述

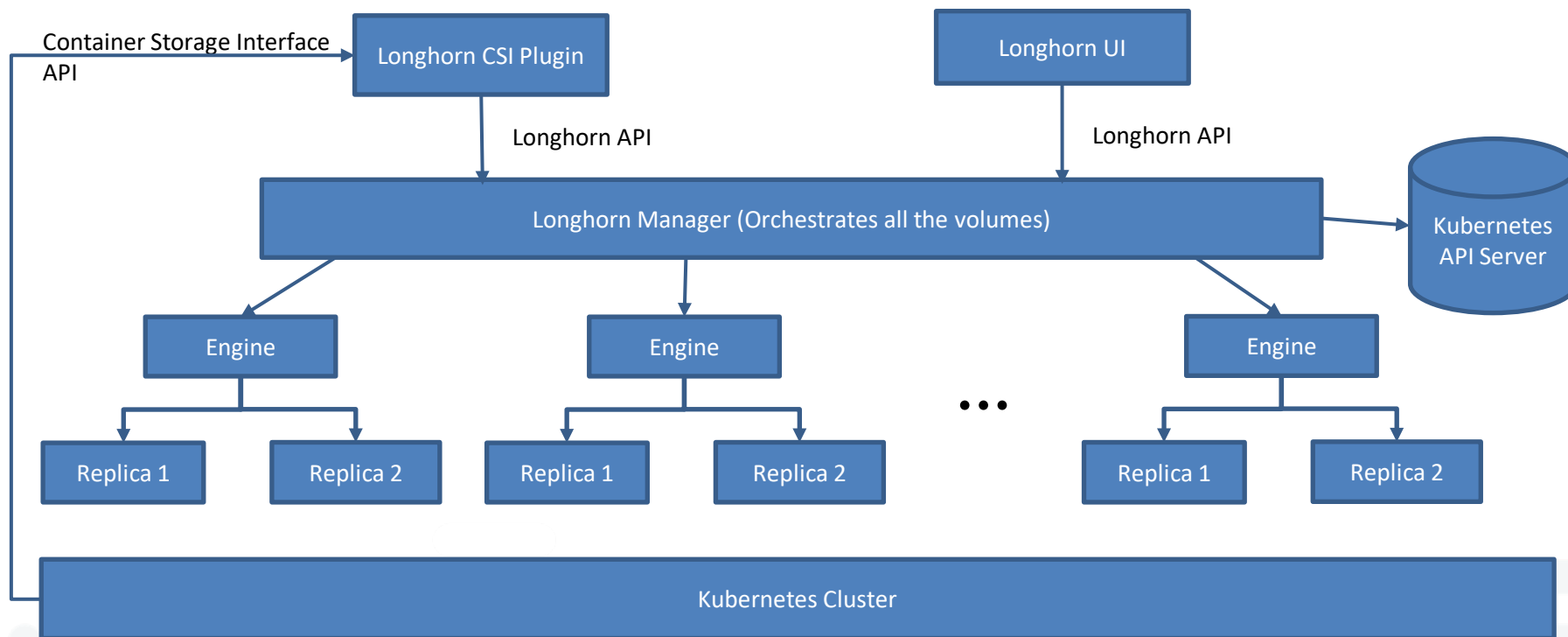


- ✓ 云原生存储，支持一键部署到k8s，可将本地磁盘或网盘设置为共享资源池。
- ✓ 管理轻量级，专注块存储、运维简单。为容器或虚拟机创建块存储卷。可以指定volume的大小，以及想要跨主机的同步replica的数量（这里的主机是指那些为volume提供存储资源的主机）。
- ✓ 为每个volume创建一个专用的存储控制器。这可能是与大多数现有的分布式存储系统相比，Longhorn比较有特色的地方。大多数现有的分布式存储系统通常采用复杂的控制器软件来服务于从数百到数千个不等的volume。但Longhorn不同，每个控制器上只有一个volume，Longhorn将每个volume都转变成了微服务。以Docker容器的形式操作存储控制器和replica。
- ✓ 跨存储主机调度多个replica保证高可用。Longhorn会监测每一个replica的健康状况，对问题进行维修，并在必要时重新生成replica。
- ✓ 可以创建volume的快照（snapshot）和快照的备份。这些快照可以备份到远程NFS或S3兼容的辅助存储中。只有更改的字节会在备份期间被复制和存储。
- ✓ 可定时做快照和备份。指定这些操作的频率（每小时，每天，每周，每月）配置Cron表达式、以及执行这些操作的确切时间（例如，每个星期日凌晨3:00），保留多少个循环快照和备份集。

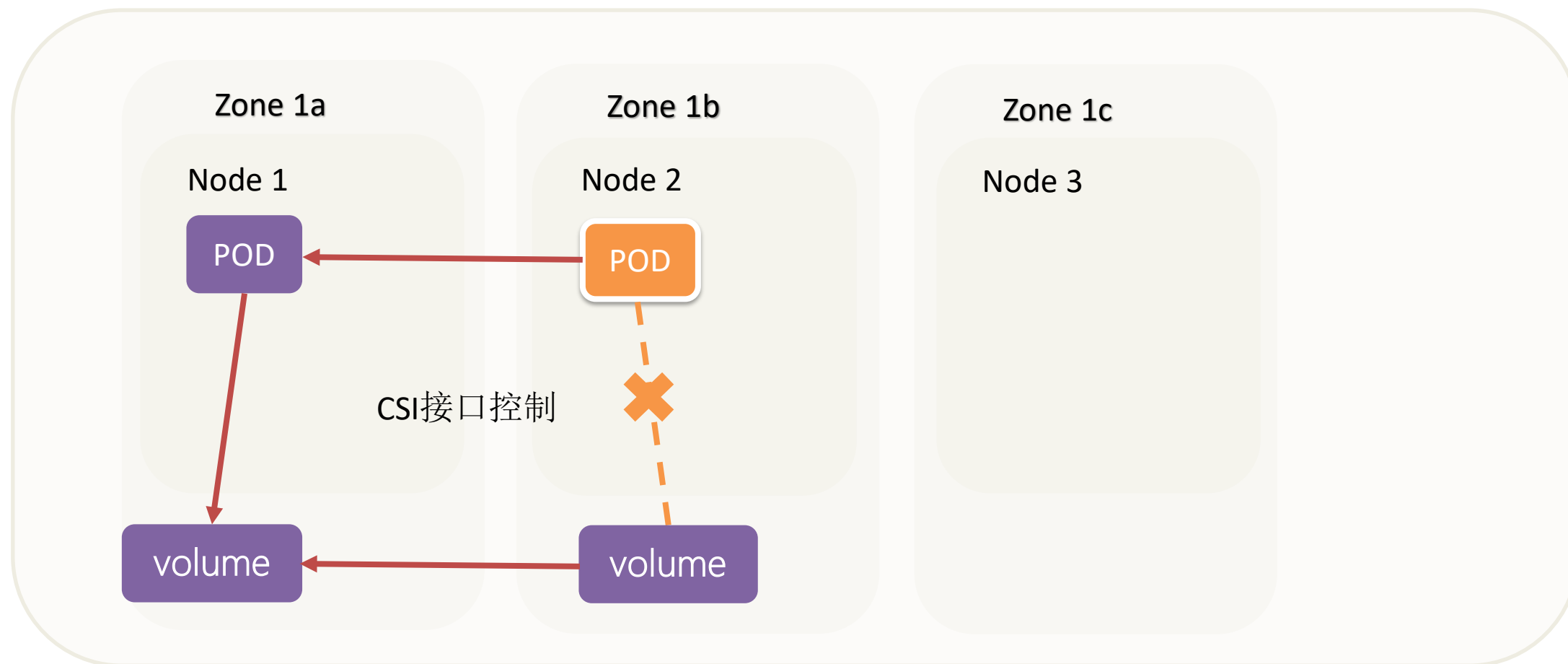
# Longhorn Architecture Overview

<https://github.com/longhorn/longhorn/wiki/Architecture-Overview-For-Developers>

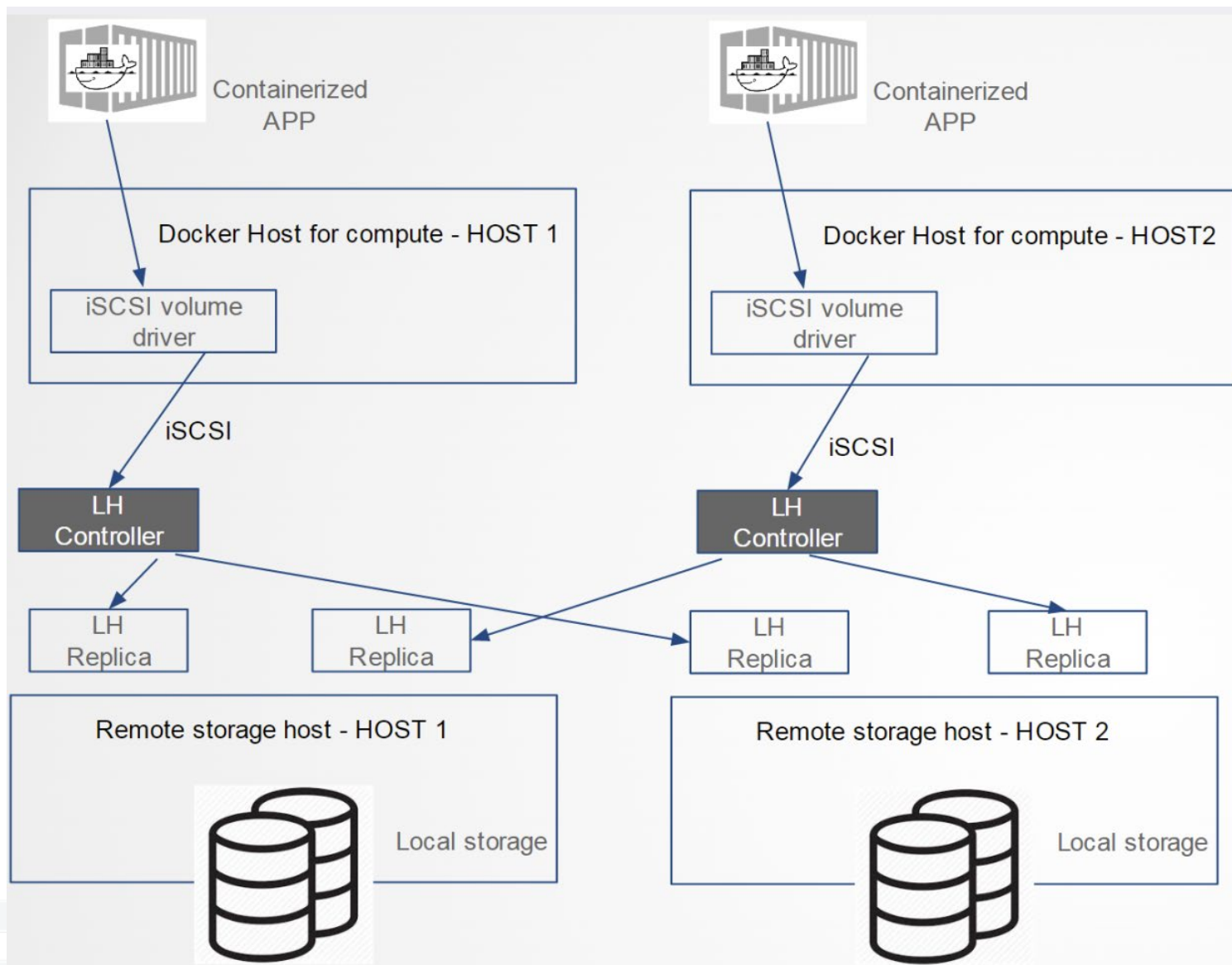
- Longhorn Manager: 控制平面
- Longhorn Engine: 数据平面



# 块存储随容器漂移



# 数据高可用



Compute container and longhorn controller on different hosts (Network storage model)



# Longhorn Volume – 快照

Volume Details

State: Attached

Health: Healthy

Ready for workload: Ready

Conditions: 🟢 restore 🟢 scheduled

Frontend: Block Device

Attached Node & Endpoint:

longhorn0002


/dev/longhorn/pvc-f6bfb898-b0aa-4b09-945c-1301e8bbb9e5

Size: 10 Gi

Actual Size: 228 Mi

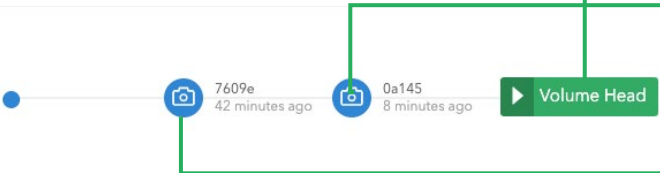
Base Image:

Replicas

 1301e8bbb9e5-r-57351e97  
Healthy

longhorn0001  
Node  
instance-manager-r-be3363ec  
/var/lib/longhorn/  
Running

Snapshots

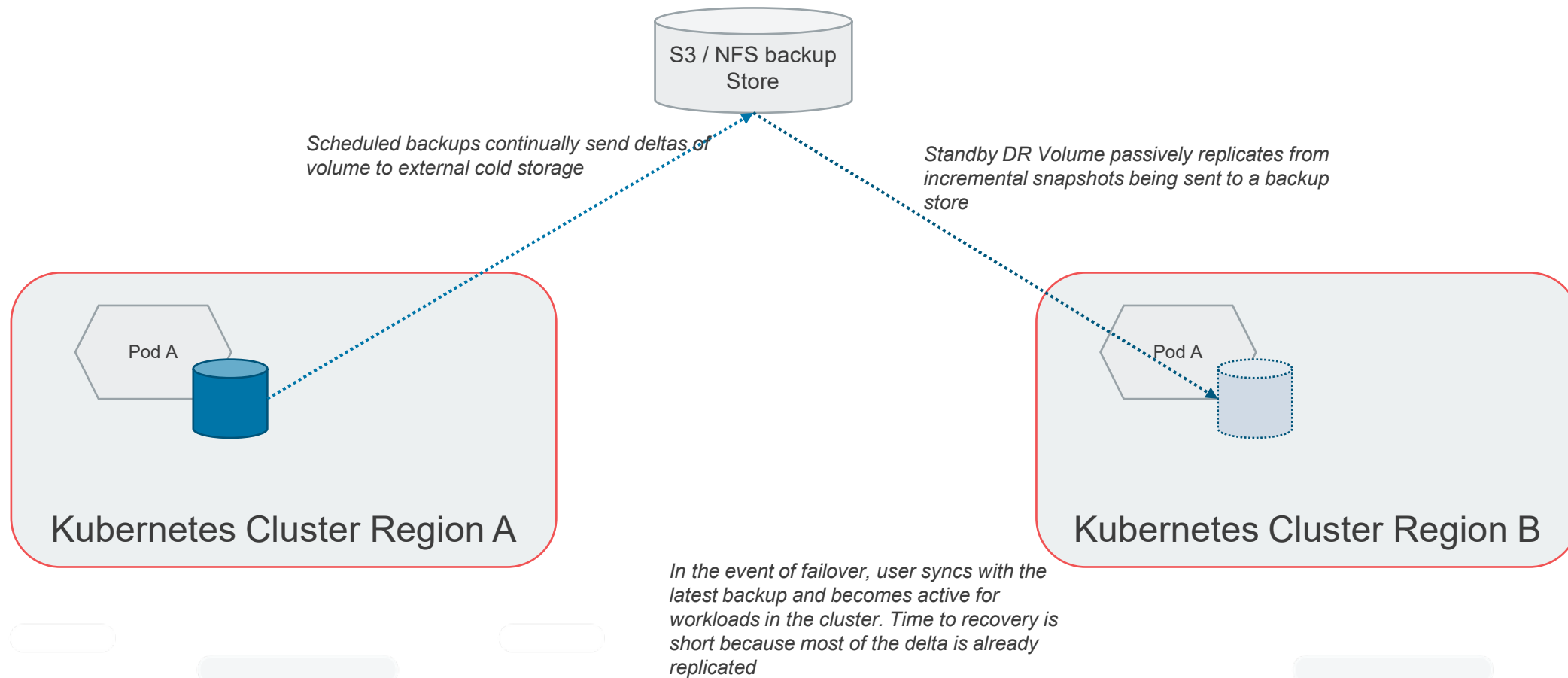


Timeline showing snapshots: 7609e (42 minutes ago) and 0a145 (8 minutes ago). A green arrow points from the 'Volume Head' to the 0a145 snapshot.

```
root@longhorn0001:/var/lib/longhorn# tree .
.
├── engine-binaries
│   └── longhornio-longhorn-engine-v1.0.0
│       └── longhorn
├── longhorn-disk.cfg
├── replicas
│   └── pvc-f6bfb898-b0aa-4b09-945c-1301e8bbb9e5-3b278f3a
│       ├── revision.counter
│       ├── volume-head-002.img
│       ├── volume-head-002.img.meta
│       ├── volume.meta
│       ├── volume-snap-0a145b62-ef5d-4280-907f-a19faf9bbe7b.img
│       ├── volume-snap-0a145b62-ef5d-4280-907f-a19faf9bbe7b.img.meta
│       ├── volume-snap-7609e14e-0f4e-488d-b631-7c411a8c5d1f.img
│       └── volume-snap-7609e14e-0f4e-488d-b631-7c411a8c5d1f.img.meta
```

# 备份的快照可在本数据中心或跨数据中心还原

*Multi-Cluster, Multi-site Disaster Recovery 可以还原出PV、PVC*



# Agenda

- 使用效果
- Longhorn核心功能介绍
- TiDB Cluster 快速部署
- 数据快照及还原

# 部署TiDB Cluster

1. `git clone https://github.com/pingcap/tidb-operator`
2. `vi tidb-operator/charts/tidb-operator/values.yaml`
3. `helm install tidb ./charts/tidb-operator --namespace rancher-operator`
4. `vi cluster.yaml` 定义集群规模

```
vi cluster.yaml
apiVersion: pingcap.com/v1alpha1
kind: TidbCluster
metadata:
  name: tidbcluster1
spec:
  timezone: UTC
  version: v4.0.1
  pd:
    affinity: {}
    enableDashboardInternalProxy: true
    baseImage: pingcap/pd
    config:
      log:
        level: info
      podSecurityContext: {}
    replicas: 3
    requests:
      cpu: "1"
      memory: 2000Mi
      storage: 5Gi
    storageClassName: longhorn
  pvReclaimPolicy: Delete
  schedulerName: tidb-scheduler
```

```
tidb:
  affinity: {}
  annotations:
    tidb.pingcap.com/sysctl-init: "true"
  baseImage: pingcap/tidb
  config:
    log:
      level: info
    performance:
      max-procs: 0
      tcp-keep-alive: true
  enableTLSClient: false
  maxFailoverCount: 3
  podSecurityContext:
    sysctls:
      - name: net.ipv4.tcp_keepalive_time
        value: "300"
      - name: net.ipv4.tcp_keepalive_intvl
        value: "75"
      - name: net.core.somaxconn
        value: "32768"
  replicas: 1
  requests:
    cpu: "1"
    memory: 2000Mi
  separateSlowLog: true
  service:
    type: NodePort
  slowLogTailer:
    limits:
      cpu: 100m
      memory: 150Mi
    requests:
      cpu: 20m
      memory: 50Mi
```

# 部署TiDB Cluster

tikv:

affinity: {}

annotations:

tidb.pingcap.com/sysctl-init: "true"

baseImage: pingcap/tikv

config:

log-level: info

hostNetwork: false

maxFailoverCount: 3

podSecurityContext:

sysctls:

- name: net.core.somaxconn

value: "32768"

privileged: false

**replicas: 3**

**requests:**

**cpu: "1"**

**memory: 4Gi**

**storage: 10Gi**

**storageClassName: longhorn**

kubectl create namespace tidb

kubectl apply -f cluster.yaml -n tidb


# 部署TiDB Cluster

[http://xxx.xxx.xxx:32335/dashboard/#/cluster\\_info/instance](http://xxx.xxx.xxx:32335/dashboard/#/cluster_info/instance)

命名空间: rancher-operator				
<input type="checkbox"/>	▶	Active	tidb-controller-manager	pingcap/tidb-operator:v1.1.1 0个Pods / 创建时间: 10 days ago / Pod 重启次数: 0
<input type="checkbox"/>	▶	Active	tidb-scheduler	pingcap/tidb-operator:v1.1.1 + 其他1个image 1个Pod / 创建时间: 10 days ago / Pod 重启次数: 0

命名空间: tidb				
<input type="checkbox"/>	▶	Active	tidbcluster1-discovery	pingcap/tidb-operator:v1.1.1 1个Pod / 创建时间: 6 days ago / Pod 重启次数: 0
<input type="checkbox"/>	▶	Active	tidbcluster1-pd	pingcap/pd:v4.0.1 3个Pods / 创建时间: 6 days ago / Pod 重启次数: 0
<input type="checkbox"/>	▶	Active	tidbcluster1-tidb	busybox:1.31.1 + 其他2个images 1个Pod / 创建时间: 6 days ago / Pod 重启次数: 0
<input type="checkbox"/>	▶	Active	tidbcluster1-tikv	pingcap/tikv:v4.0.1 + 其他1个image 3个Pods / 创建时间: 6 days ago / Pod 重启次数: 0

<input type="checkbox"/> 状态	PVC名称	大小	持久卷(PV)	存储类
命名空间: tidb				
<input type="checkbox"/> <span>Bound</span>	pd-tidbcluster1-pd-0	5 GiB	pvc-5751c90a-526f-4cc6-a1d0-096b72c3617c	longhorn
<input type="checkbox"/> <span>Bound</span>	pd-tidbcluster1-pd-1	5 GiB	pvc-ba677c89-51c2-49c8-b71c-7d9db9b77f2b	longhorn
<input type="checkbox"/> <span>Bound</span>	pd-tidbcluster1-pd-2	5 GiB	pvc-a9adf07f-5532-43f5-ac9b-345f540cc969	longhorn
<input type="checkbox"/> <span>Bound</span>	tikv-tidbcluster1-tikv-0	10 GiB	pvc-94c52f07-de9e-41f7-b2a1-7c4d6bd78af7	longhorn
<input type="checkbox"/> <span>Bound</span>	tikv-tidbcluster1-tikv-1	10 GiB	pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c	longhorn
<input type="checkbox"/> <span>Bound</span>	tikv-tidbcluster1-tikv-2	10 GiB	pvc-354c2af4-a4c6-43f8-87a7-061e8a5c8b62	longhorn

 TiDB Dashboard  
Dashboard version 4.0.1

概况

集群信息

流量可视化

SQL 语句分析

慢查询

集群诊断

日志搜索

高级调试

实例 主机

地址	状态	启动时间	版本
▼ tidb (1)			
tidbcluster1-tidb-0.tidb...	● 在线	上星期六 21:13	v4.0.1
▼ tikv (3)			
tidbcluster1-tikv-1.tidb...	● 在线	上星期六 21:13	v4.0.1
tidbcluster1-tikv-0.tidb...	● 在线	上星期六 21:13	v4.0.1
tidbcluster1-tikv-2.tidb...	● 在线	上星期六 21:13	v4.0.1
▼ pd (3)			
tidbcluster1-pd-1.tidb...	● 在线	上星期六 21:11	v4.0.1
tidbcluster1-pd-0.tidb...	● 在线	上星期六 21:11	v4.0.1
tidbcluster1-pd-2.tidb...	● 在线	上星期六 21:11	v4.0.1

# Agenda

- 使用效果
- Longhorn核心功能介绍
- TiDB Cluster 快速部署
- 数据快照及还原

# 数据volume批量备份

Delete Attach Detach **Create Backup** ⋮

✓	State ▾	Name ▾	Size ▾	Created ▾	PV/PVC ▾	Namespace ▾	Attached To ▾
✓	Healthy	pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c	10 Gi	5 days ago	Bound	tidb	tidbcluster1-tikv-1 on master1
✓	Healthy	pvc-354c2af4-a4c6-43f8-87a7-061e8a5c8b62	10 Gi	5 days ago	Bound	tidb	tidbcluster1-tikv-2 on work2
✓	Healthy	pvc-94c52f07-de9e-41f7-b2a1-7c4d6bd78af7	10 Gi	5 days ago	Bound	tidb	tidbcluster1-tikv-0 on work1
✓	Healthy	pvc-5751c90a-526f-4cc6-a1d0-096b72c3617c	5 Gi	5 days ago	Bound	tidb	tidbcluster1-pd-0 on work2
✓	Healthy	pvc-a9adf07f-5532-43f5-ac9b-345f540cc969	5 Gi	5 days ago	Bound	tidb	tidbcluster1-pd-2 on master1
✓	Healthy	pvc-ba677c89-51c2-49c8-b71c-7d9db9b77f2b	5 Gi	5 days ago	Bound	tidb	tidbcluster1-pd-1 on work1

通过多选后点击CreateBackup 完成批量快照的远程备份同时volume数据自动会做一次快照

Dashboard / volume / pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c

Volume Details

State: Attached

Health: Healthy

Ready for workload: Ready

Conditions: restore scheduled

Frontend: Block Device

Attached Node & Endpoint:  
master1  
/dev/longhorn/pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c

Size: 10 Gi

Actual Size: 346 Mi

Base Image:

Engine Image: longhorn/longhorn-engine-v1.0.0

Replicas

73d3348bfd7c-r-45643b4f  
Healthy

master1  
Node  
instance-manager-r-26ec3183  
/mnt/disk2/  
Running

73d3348bfd7c-r-5374b2c6  
Healthy

work2  
Node  
instance-manager-r-83745174  
/mnt/disk2/  
Running

73d3348bfd7c-r-bd915224  
Healthy

work1  
Node  
instance-manager-r-0fdf1d49  
/mnt/disk2/  
Running

Snapshots

9444e  
5 days ago

Volume Head

Take Snapshot Create Backup

Show System Hidden: ☐

16



# 数据volume批量备份



Dashboard / backup

Restore Latest Backup    Create Disaster Recovery Volume

	Name	Size	Last Backup At
<input type="checkbox"/>	 backups	0 Bi	
<input type="checkbox"/>	pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c	10 Gi	5 days ago
<input type="checkbox"/>	pvc-354c2af4-a4c6-43f8-87a7-061e8a5c8b62	10 Gi	5 days ago
<input type="checkbox"/>	pvc-5751c90a-526f-4cc6-a1d0-096b72c3617c	5 Gi	5 days ago
<input type="checkbox"/>	pvc-94c52f07-de9e-41f7-b2a1-7c4d6bd78af7	10 Gi	5 days ago
<input type="checkbox"/>	pvc-a9adf07f-5532-43f5-ac9b-345f540cc969	5 Gi	5 days ago
<input type="checkbox"/>	pvc-ba677c89-51c2-49c8-b71c-7d9db9b77f2b	5 Gi	5 days ago

Dashboard / backup / pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c



ID	Volume	Snapshot Name	Size	PV/PVC	Workload/Pod	Snapshot Created	Labels	Operation
backup-02fad0c463224c35	pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c	9444edfa-5be0-419a-bcd7-286f743ea8d4	278 Mi	Bound	tidbcluster1-tikv-1	5 days ago		

- Delete
- Restore
- Get URL



Backup URL:

s3://backupbucket@us-east-1/?backup=backup-02fad0c463224c35&volume=pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c



# 数据volume 定时自动备份、快照

Recurring Snapshot and Backup Schedule

Type	Schedule	Labels	Retain
<div><div>Snapshot ^</div><div>Backup</div><div>Snapshot</div></div>	Every 5 and 55th minute past every hour		<div>20</div> <div></div>

+ New

Cancel Save

- 每个数据卷可以设定自动备份或快照的Cron表达式，可以设定备份循环Retain的数量

# 基于快照的数据还原

- 恢复数据前先要把 TiKV、PD、TiDB 副本数设置为0，Longhorn中看到volume已经Detached，然后开始还原数据

工作负载: tidbcluster1-pd

命名空间: tidb	镜像名: pingcap/pd:v4.0.1
访问端口: n/a	Pod配置副本数: 0 Pod可用副本数: 0

Dashboard / volume

Delete Attach Detach Create Backup

	State	Name	Size	Created	PV/PVC	Namespace	Attached To
<input type="checkbox"/>	Detached	pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c	10 Gi	5 days ago	Bound	tidb	tidbcluster1-tikv-1
<input type="checkbox"/>	Detached	pvc-354c2af4-a4c6-43f8-87a7-061e8a5c8b62	10 Gi	5 days ago	Bound	tidb	tidbcluster1-tikv-2
<input type="checkbox"/>	Detached	pvc-94c52f07-de9e-41f7-b2a1-7c4d6bd78af7	10 Gi	5 days ago	Bound	tidb	tidbcluster1-tikv-0
<input type="checkbox"/>	Detached	pvc-5751c90a-526f-4cc6-a1d0-096b72c3617c	5 Gi	5 days ago	Bound	tidb	tidbcluster1-pd-0
<input type="checkbox"/>	Detached	pvc-a9adf07f-5532-43f5-ac9b-345f540cc969	5 Gi	5 days ago	Bound	tidb	tidbcluster1-pd-2
<input type="checkbox"/>	Detached	pvc-ba677c89-51c2-49c8-b71c-7d9db9b77f2b	5 Gi	5 days ago	Bound	tidb	tidbcluster1-pd-1

# 基于快照的数据还原

- 先把要还原的volume Attach 到一台宿主机进入维护模式

The screenshot shows a storage management dashboard with a table of volumes and an 'Attach to host' dialog box.

**Dashboard / volume**

Buttons: Delete, Attach, Detach, Create Backup

Table Headers: State, Name

State	Name
Detached	pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c
Detached	pvc-354c2af4-a4c6-43f8-87a7-061e8a5c8b62
Detached	pvc-94c52f07-de9e-41f7-b2a1-7c4d6bd78af7
Detached	pvc-5751c90a-526f-4cc6-a1d0-096b72c3617c
Detached	pvc-a9adf07f-5532-43f5-ac9b-345f540cc969
Detached	pvc-ba677c89-51c2-49c8-b71c-7d9db9b77f2b

**Attach to host**

\* Host: work1

Maintenance: ☒

Buttons: Cancel, OK

# 基于快照的数据还原

Dashboard / volume / pvc-ba677c89-51c2-49c8-b71c-7d9db9b77f2b

Volume Details

Last Backup At: 5 days ago

Instance Manager:  
instance-manager-e-d81a2d5d

Last time used by Pod: 29 minutes ago

Namespace: tidb

PVC Name: pd-tidbcluster1-pd-1

PV Name: pvc-ba677c89-51c2-49c8-b71c-7d9db9b77f2b

PV Status: Bound

Last Pod Name: tidbcluster1-pd-1

Last Workload Name: tidbcluster1-pd

Last Workload Type: StatefulSet

Replicas

7d9db9b77f2b-r-75d26867

Healthy

work1

Node

instance-manager-r-0fdf1d49

/mnt/disk2/

Running

Snapshots

c30a5

5 days ago

Volume Head

Revert

Backup

Delete

Dashboard / volume

Delete

Attach

Detach

Create Backup

	State	Name	Size
<input checked="" type="checkbox"/>	Healthy	pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c	10 Gi
<input checked="" type="checkbox"/>	Healthy	pvc-354c2af4-a4c6-43f8-87a7-061e8a5c8b62	10 Gi
<input checked="" type="checkbox"/>	Healthy	pvc-94c52f07-de9e-41f7-b2a1-7c4d6bd78af7	10 Gi
<input checked="" type="checkbox"/>	Healthy	pvc-5751c90a-526f-4cc6-a1d0-096b72c3617c	5 Gi
<input checked="" type="checkbox"/>	Healthy	pvc-a9adf07f-5532-43f5-ac9b-345f540cc969	5 Gi
<input checked="" type="checkbox"/>	Healthy	pvc-ba677c89-51c2-49c8-b71c-7d9db9b77f2b	5 Gi

- volume 进入维护模式后点击Revert即可恢复快照数据
- 进入所有要恢复数据的volume界面依次点击Revert还原数据
- Revert完成后批量把volume Detach ， 再依次恢复TiKV、PD、TiDB模块的实例数即可完成数据还原

RANCHER

© Copyright 2020 Rancher Labs. All Rights Reserved. Confidential

21

# 灾难/跨集群恢复





- 当三个副本的数据都损坏后，可以用之前对象存储备份的数据恢复，直接还原出多副本的 Volume、PV和PVC
- 其它k8s集群只要能访问到备份的数据的存储，就可以跨集群还原多副本的 Volume、PV和PVC

Restore Latest Backup

Create Disaster Recovery Volume

Name

Go

	Name	Size	Last Backup At	Created At	Operation
<input type="checkbox"/>	 backups	0 Bi		Invalid date	
<input checked="" type="checkbox"/>	pvc-336c5856-1ccb-4f06-b752-73d3348bfd7c	10 Gi	5 days ago	5 days ago	
<input type="checkbox"/>	pvc-354c2af4-a4c6-43f8-87a7-061e8a5c8b62	10 Gi	5 days ago	5 days ago	


Create Disaster Recovery Volume

Restore Latest Backup

Delete All Backups

## Create Disaster Recovery Volume

\* Name: pvc-336c5856-1ccb-4f06-b752-73c 

\* Size: 10 Gi 

\* Number of Replicas: 3 

Cancel

OK

## \* Volume Attachment Recovery Policy

wait

Required. Defines the Longhorn action when a Volume is stuck with a Deployment Pod on a failed node. pods: 'immediate' leads to the deletion of the volume attachment as soon as all workload pods are pending.

## Backup

Backup Target

s3://backupbucket@us-east-1/


The endpoint used to access the backupstore. NFS and S3 are supported.

Backup Target Credential Secret

minio-secret

The name of the Kubernetes secret associated with the backup target.


## Create PV/PVC

\* PV Name: testvol004 

File System: ☒ Ext4 ☐ XFS

Create PVC: ☒ Use Previous PVC: ☐

\* PVC Name: testvol004-renamed 

\* Namespace: default 

Cancel

OK



# 谢谢!

欢迎扫码关注Rancher微信公众号，  
获取更多干货信息

本次演示的内容建议应用在测试环境下