TiDB Operator 实现原理解析

weekface@PingCAP

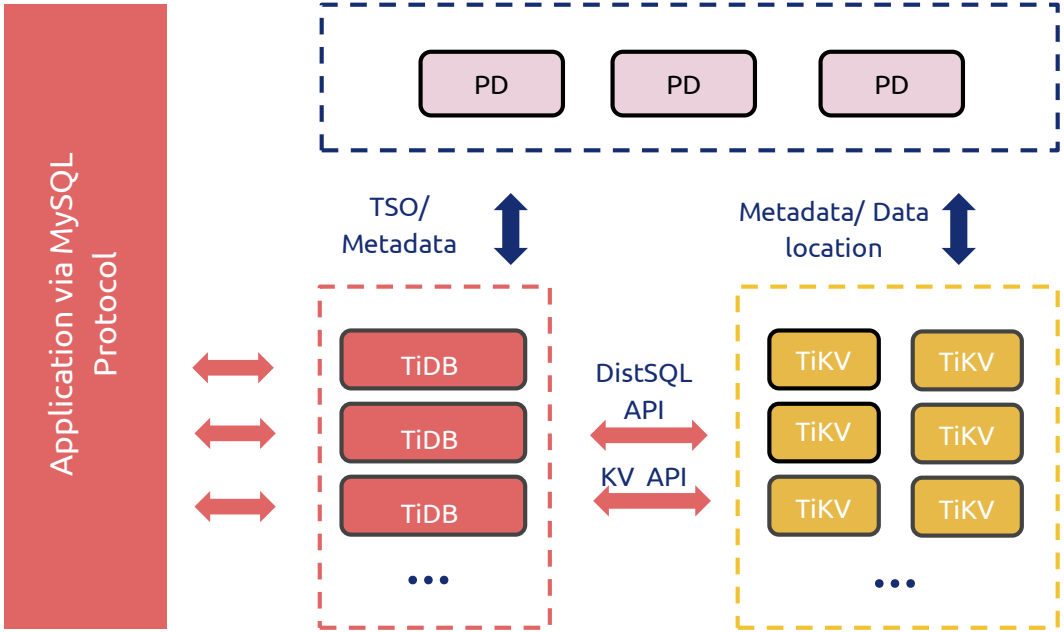
About me

weekface (常军昌), software engineer, PingCAP

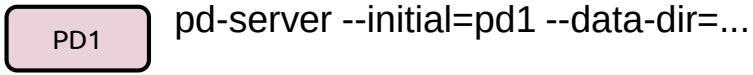
Focus on Cloud TiDB

changjunchang@pingcap.com

About TiDB


Application via MySQL Protocol
PD
PD
PD
TSO/ Metadata
Metadata/ Data location
TiDB
TiDB
TiDB
DistSQL API
KV API
TiKV
TiKV
TiKV
TiKV
TiKV
TiKV

Traditional Deployment



Traditional Deployment

PD1

pd-server --initial=pd1 --data-dir=...

PD2

pd-server --join=pd1 --data-dir=...

Traditional Deployment

PD1 pd-server --initial=pd1 --data-dir=...

PD2 pd-server --join=pd1 --data-dir=...

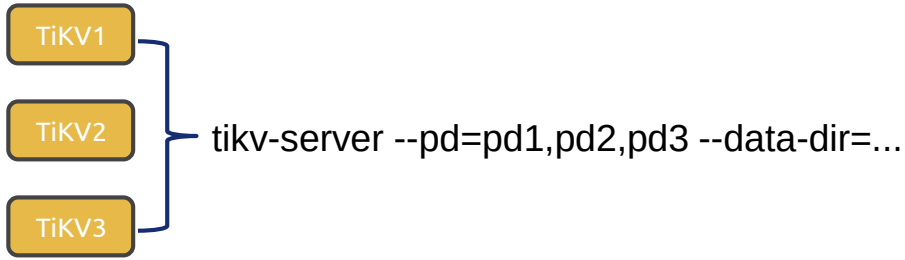
PD3 pd-server --join=pd1,p2 --data-dir=...

Traditional Deployment

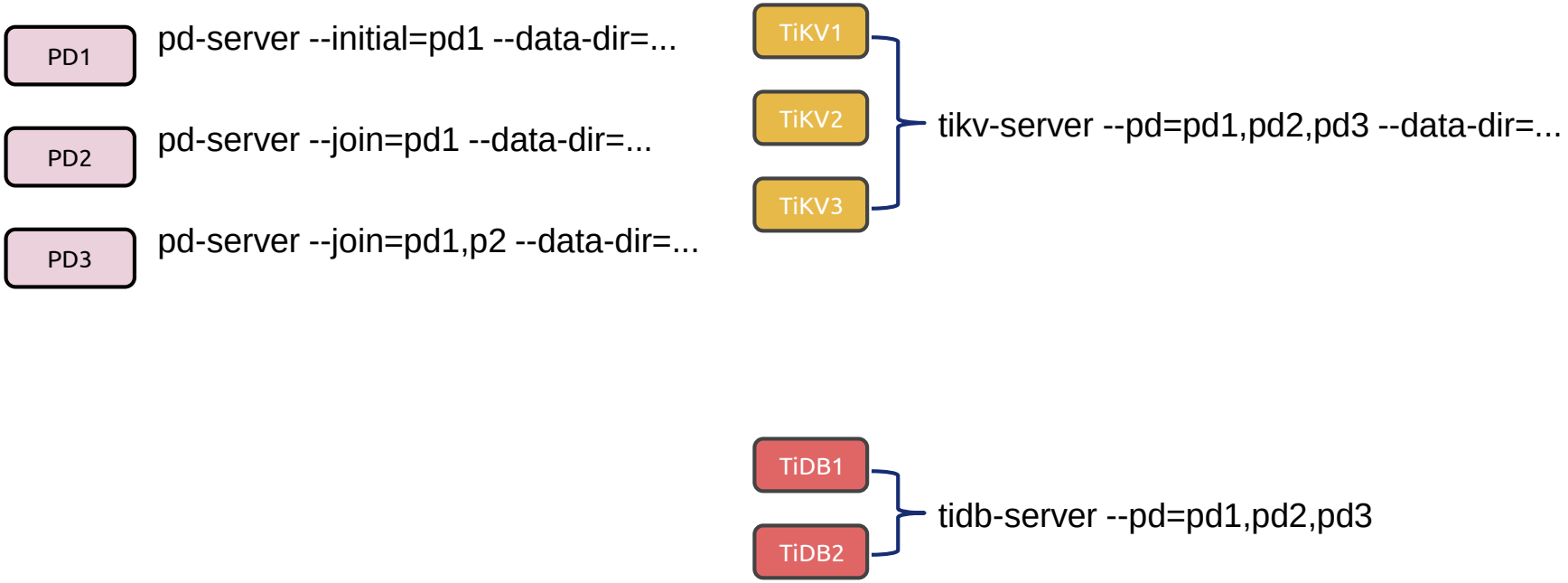
PD1 pd-server --initial=pd1 --data-dir=...

PD2 pd-server --join=pd1 --data-dir=...

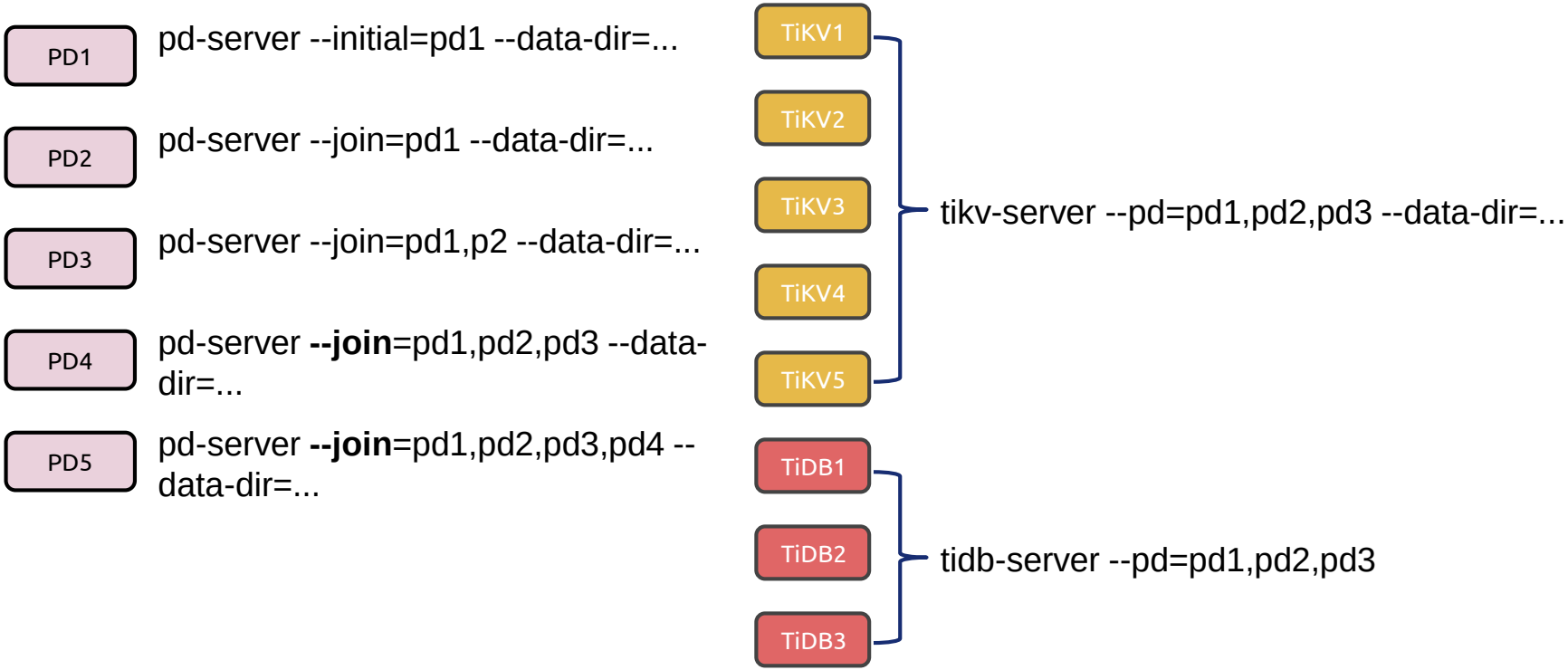
PD3 pd-server --join=pd1,p2 --data-dir=...



Traditional Deployment



Traditional Deployment



Summary

Operation	**PD**	**TiKV**	**TiDB**
Persistent storage	**Yes**	**Yes**	-
Create/Upgrade Order	1	2	3
Startup args	**Different**	Same	Same
Before Stop/Offline	Resign Leader/ Delete Member	Evict Leader/ Delete Store	Resign DDL Owner
Failover			

What’s TiDB Operator?

TiDB operator creates and manages TiDB clusters running in Kubernetes

Standing on the shoulders of giants

Based on Raw **Pod**, we can do all the things. But we don’t want to recreate a wheel

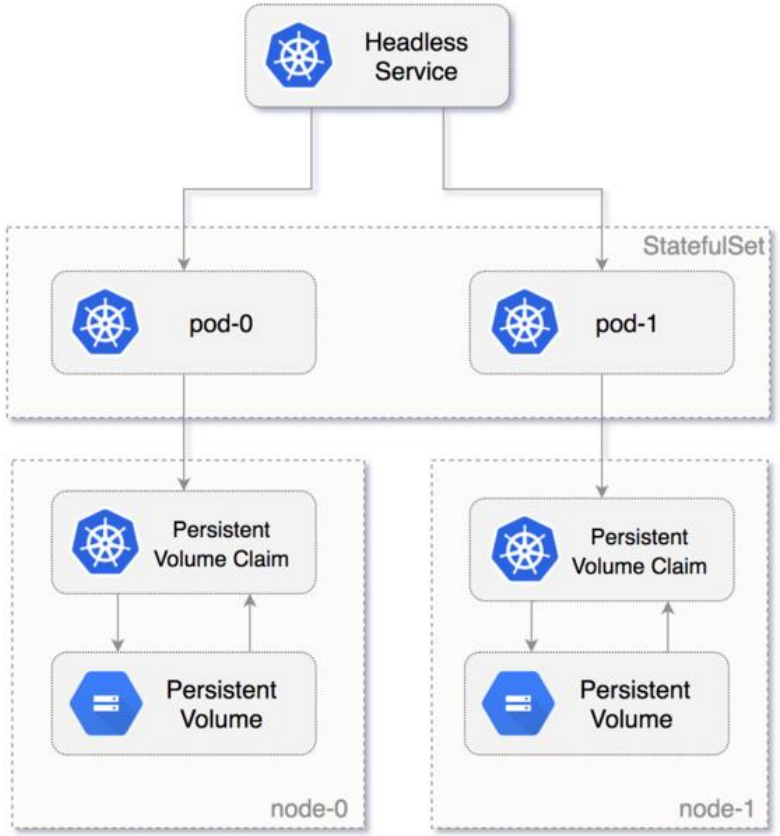
Standing on the shoulders of giants

ReplicaSet，**Deployment** and other workloads are used to manage **stateless** applications

Standing on the shoulders of giants

StatefulSet is the workload API object used to manage **stateful** applications.

- Stable, unique network identifiers.
- Stable, persistent storage.
- Ordered, graceful deployment and scaling.
- Ordered, automated rolling updates.



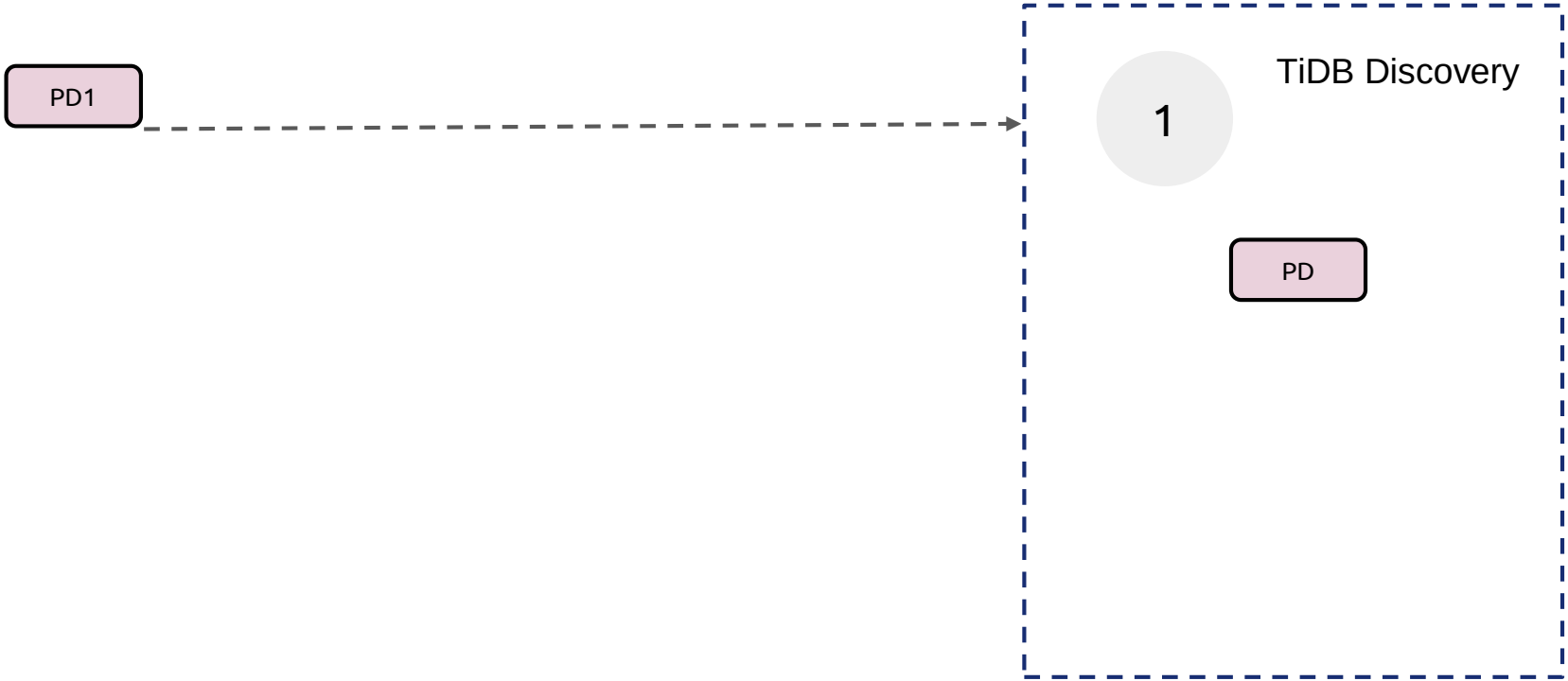
Based on StatefulSet - Persistent storage

PersistentVolumes associated with the Pods’ PersistentVolumeClaims

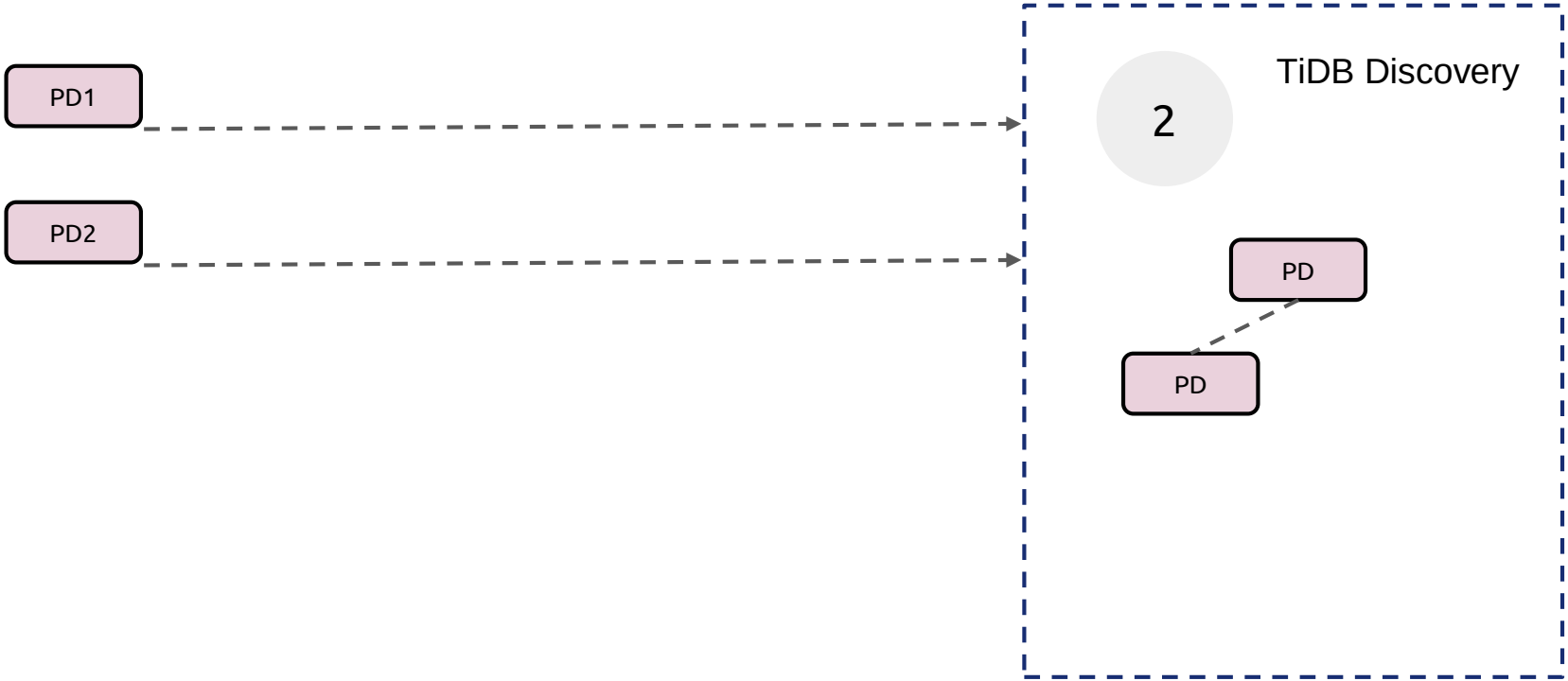
Based on StatefulSet - Create/Upgrade Order

Create and manage three StatefulSet objects: PD StatefulSet, TiKV StatefulSet, TiDB StatefulSet

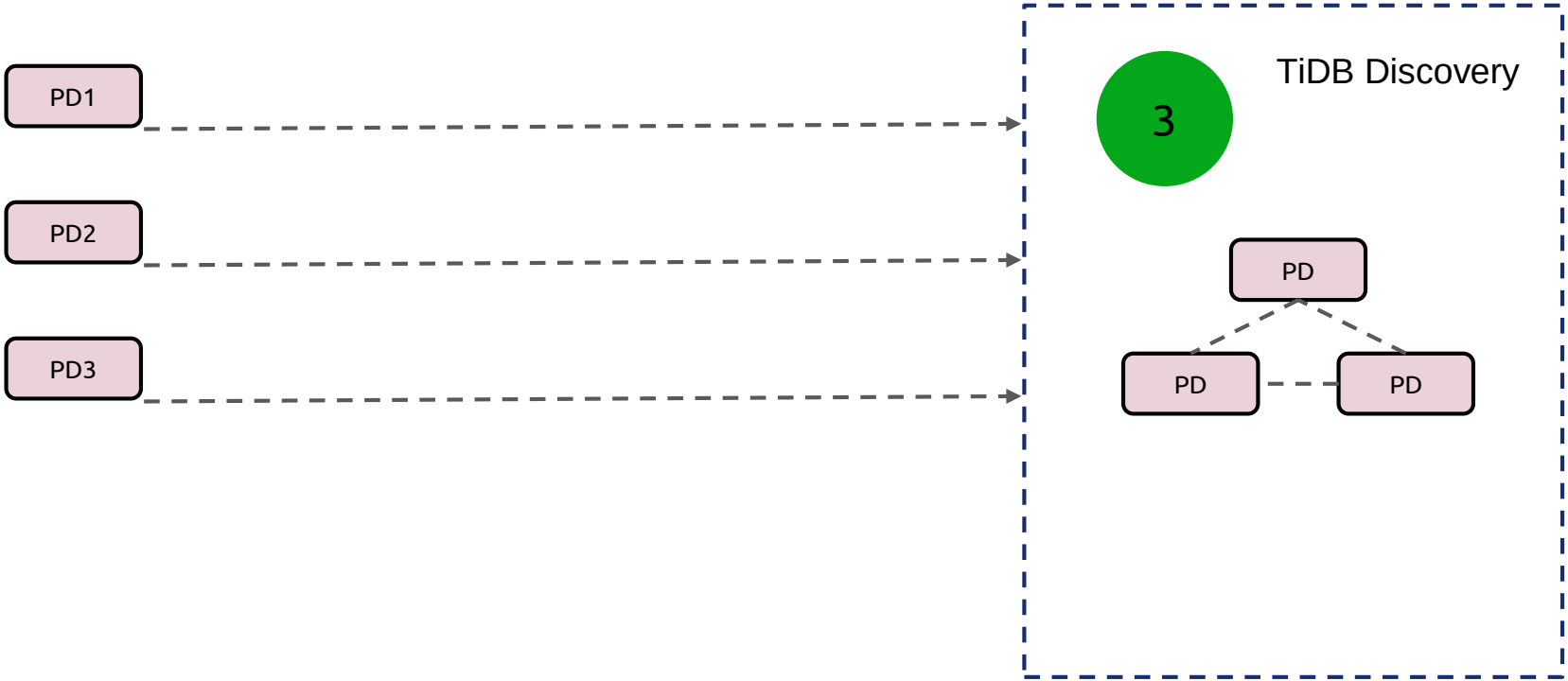
Based on StatefulSet - PD Startup Args



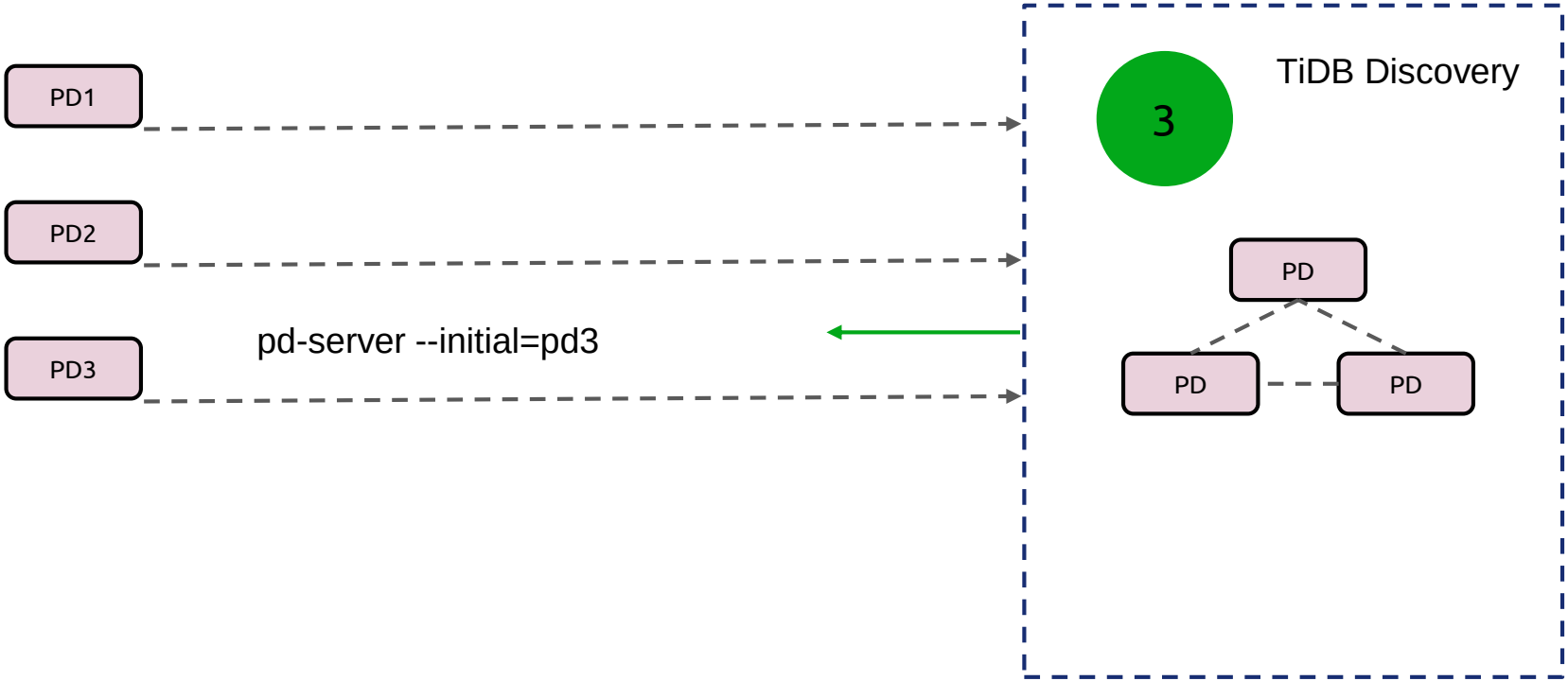
Based on StatefulSet - PD Startup Args



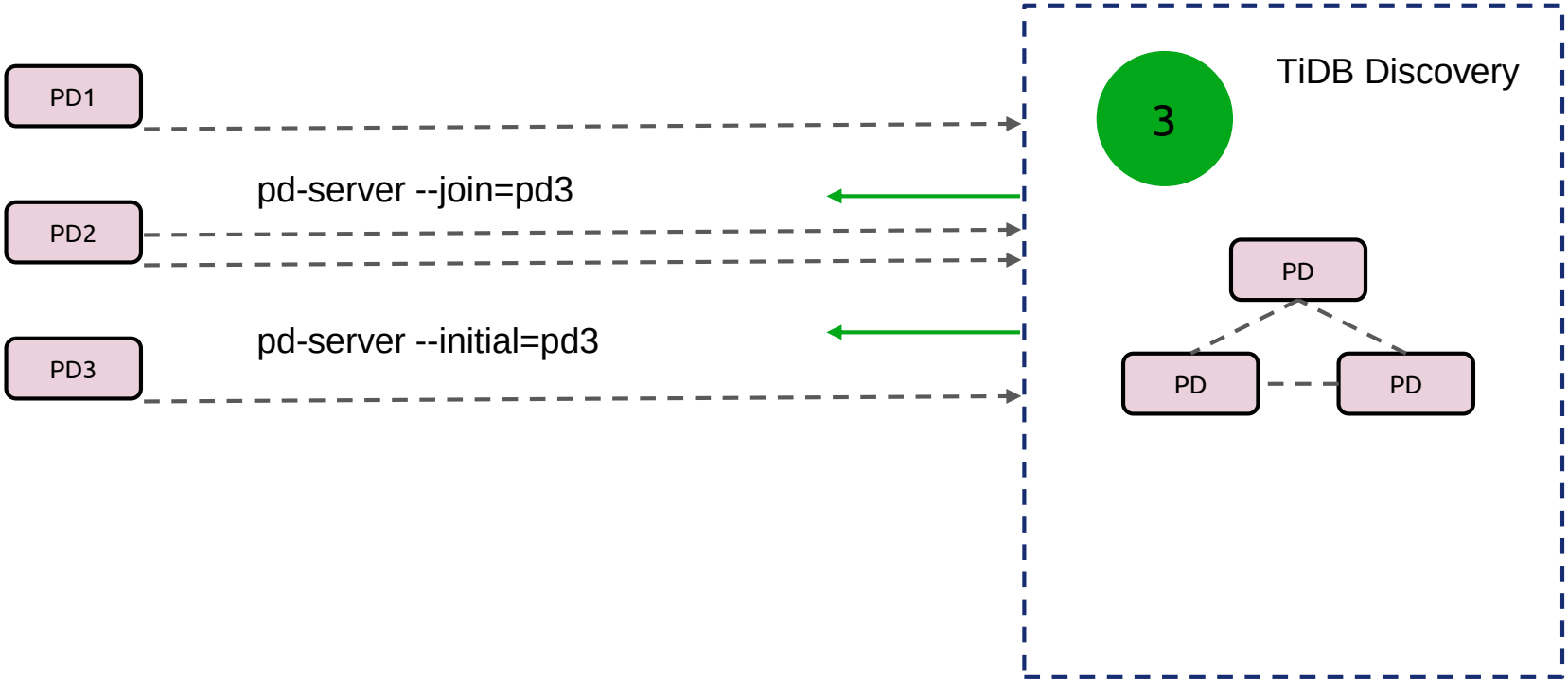
Based on StatefulSet - PD Startup Args



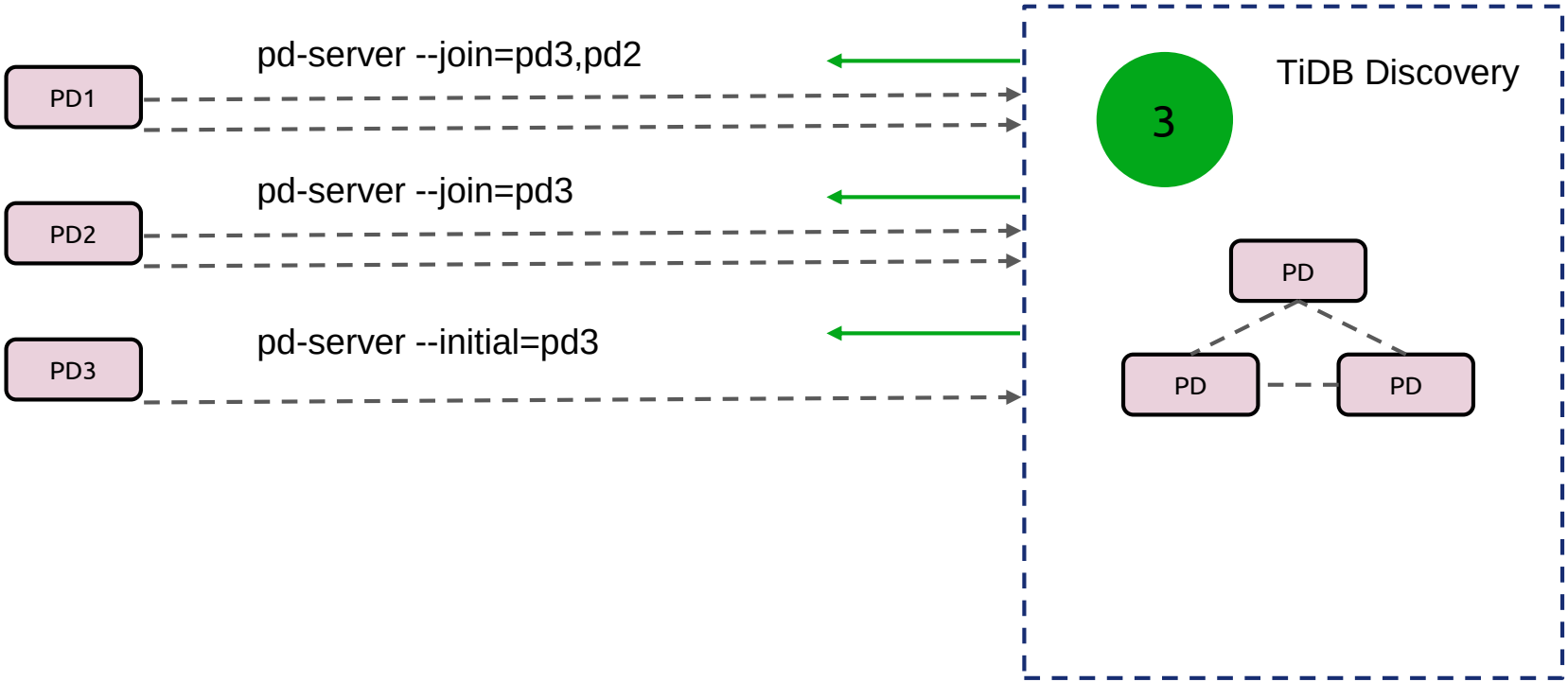
Based on StatefulSet - PD Startup Args



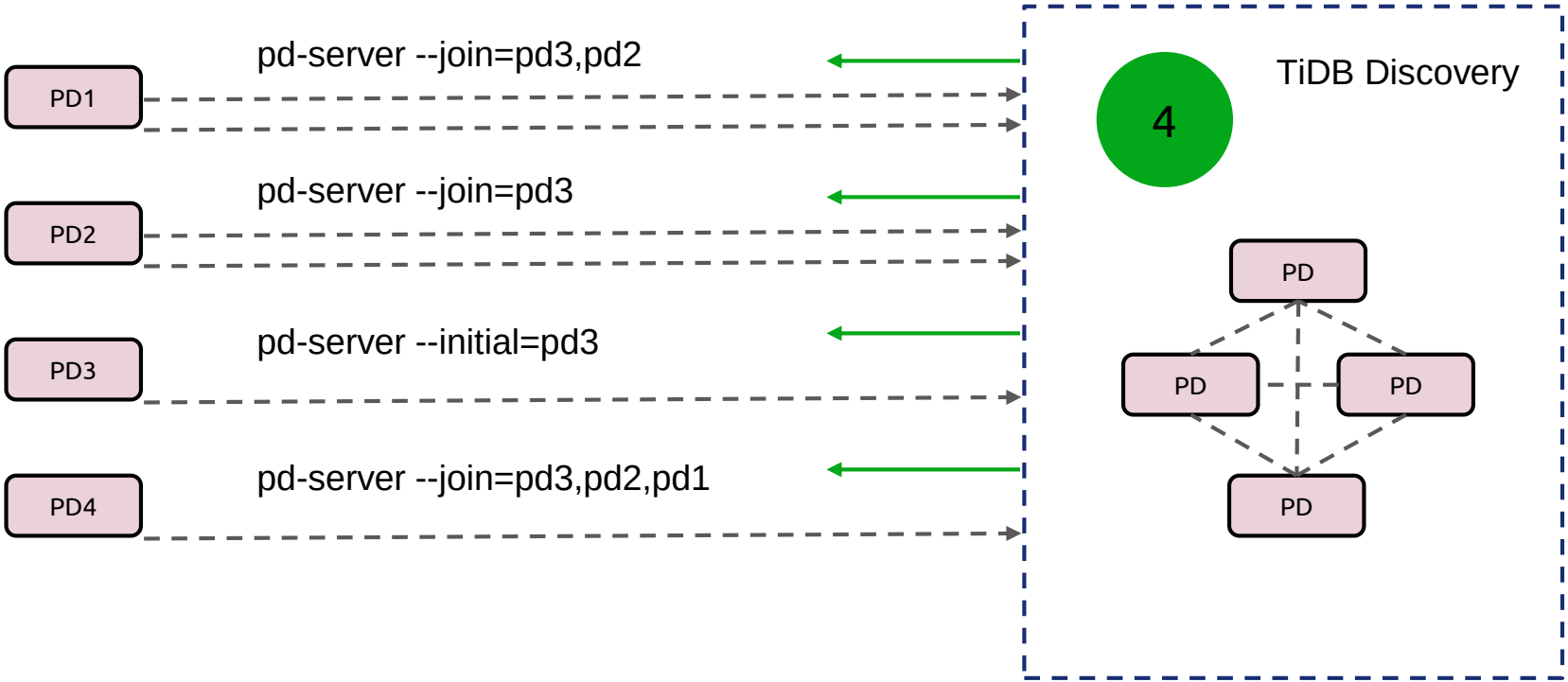
Based on StatefulSet - PD Startup Args



Based on StatefulSet - PD Startup Args



Based on StatefulSet - PD Startup Args



Based on StatefulSet - Before Stop/Offline

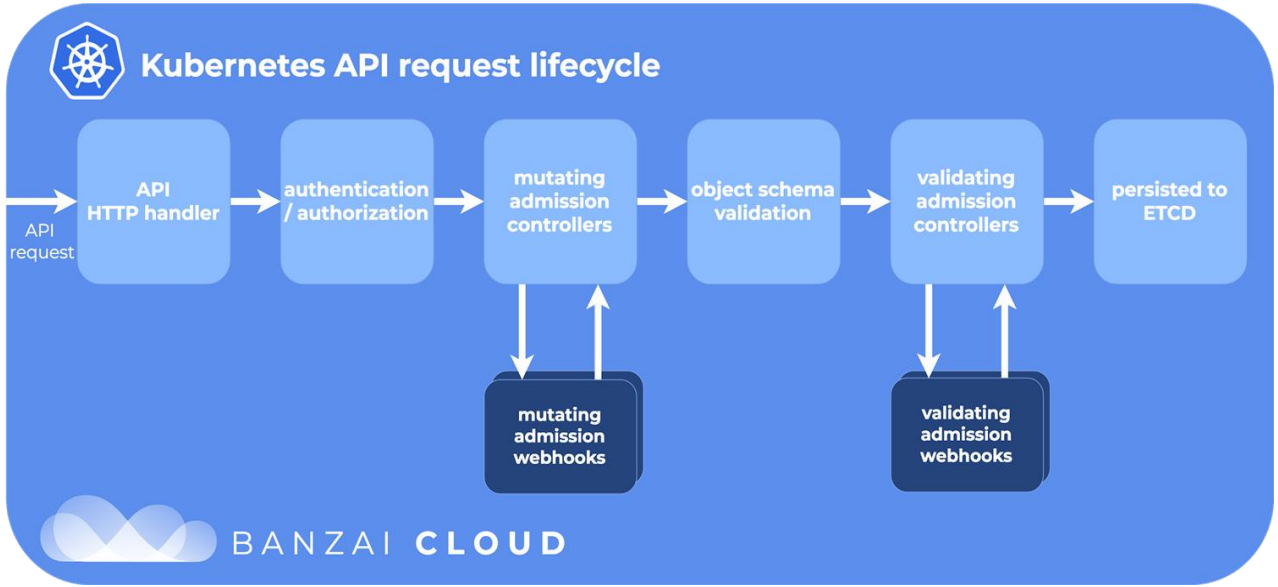
We should do something before the Pod terminated.

- K8s PreStop probe
- StatefulSet partition
- Validating admission webhooks

Based on StatefulSet - Before Stop/Offline

We should do something before the Pod terminated.

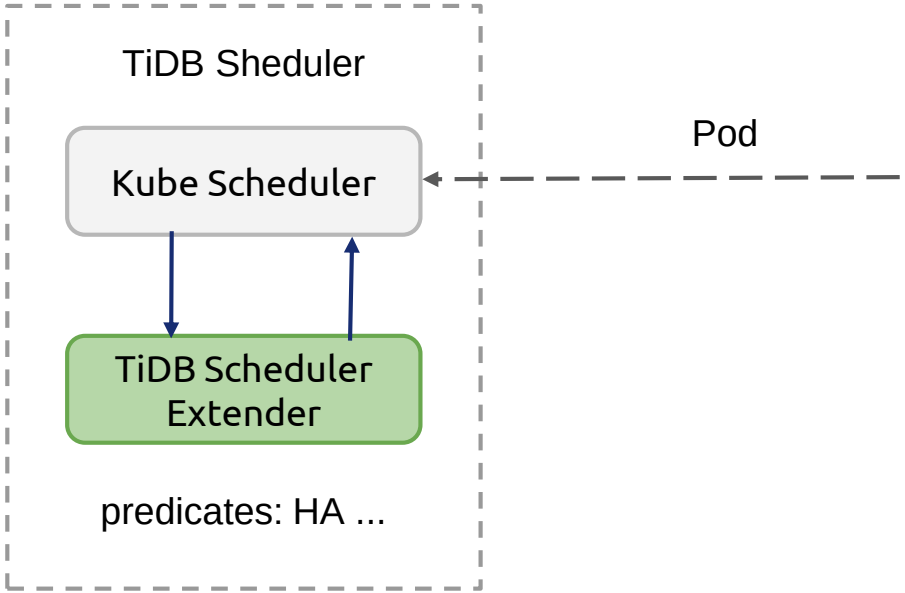
- K8s PreStop probe
- StatefulSet partition
- Validating admission webhooks



TiDB Scheduler

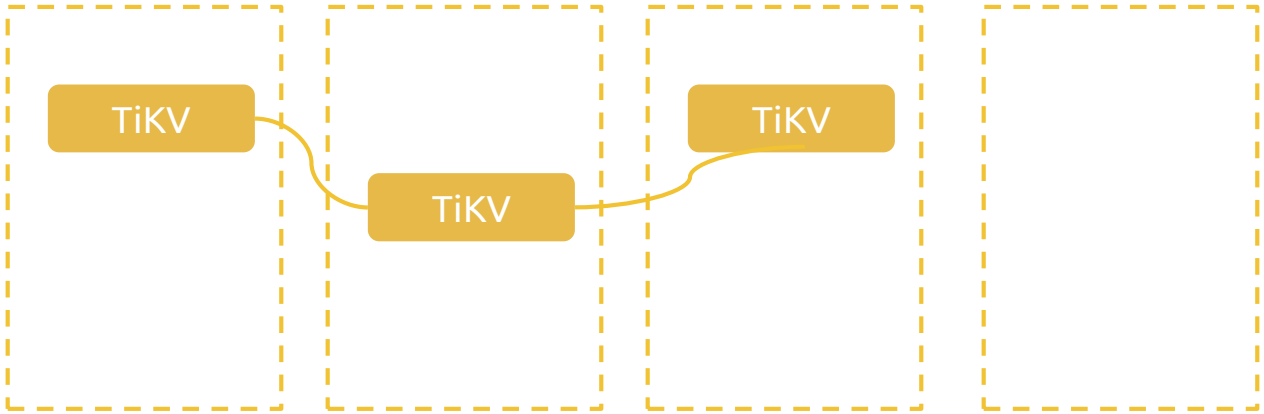
For data safety, we don’t permit two TiKV instances scheduled into the same node

```
apiVersion: apps/v1
kind: StatefulSet
…
spec:
  template:
    spec:
      schedulerName: tidb-scheduler
      containers:
...
```



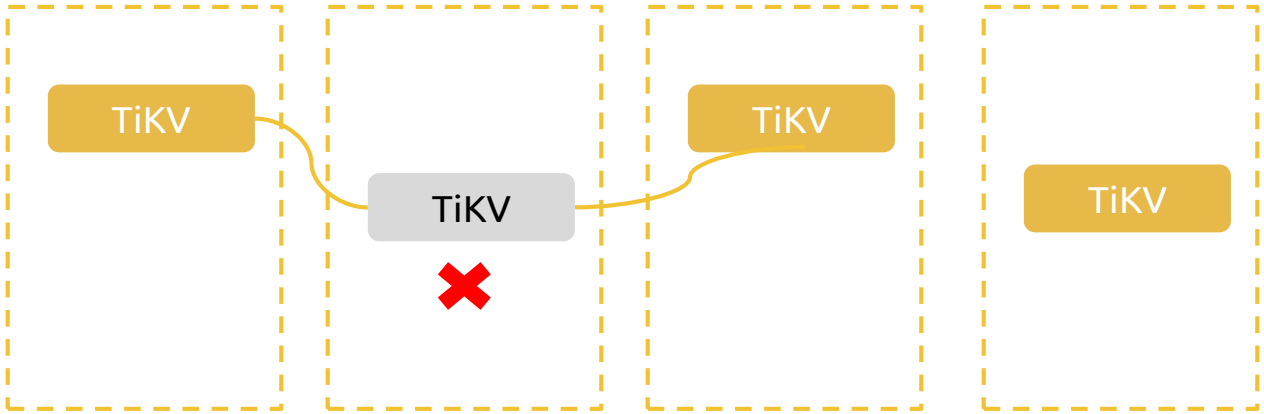
Failover

StatefulSet don’t failover automatically, we need this feature for TiDB cluster.



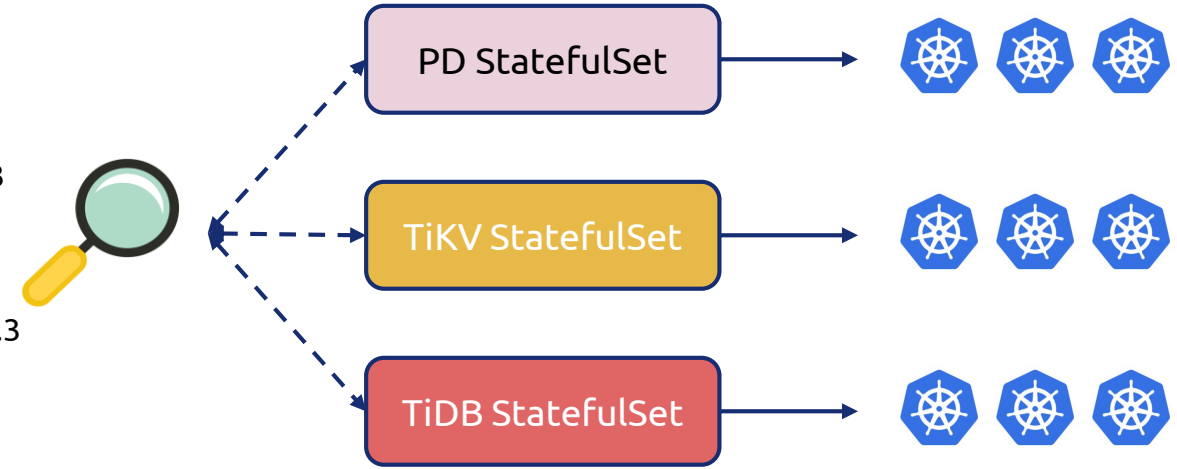
Failover

StatefulSet don’t failover automatically, we need this feature for TiDB cluster.

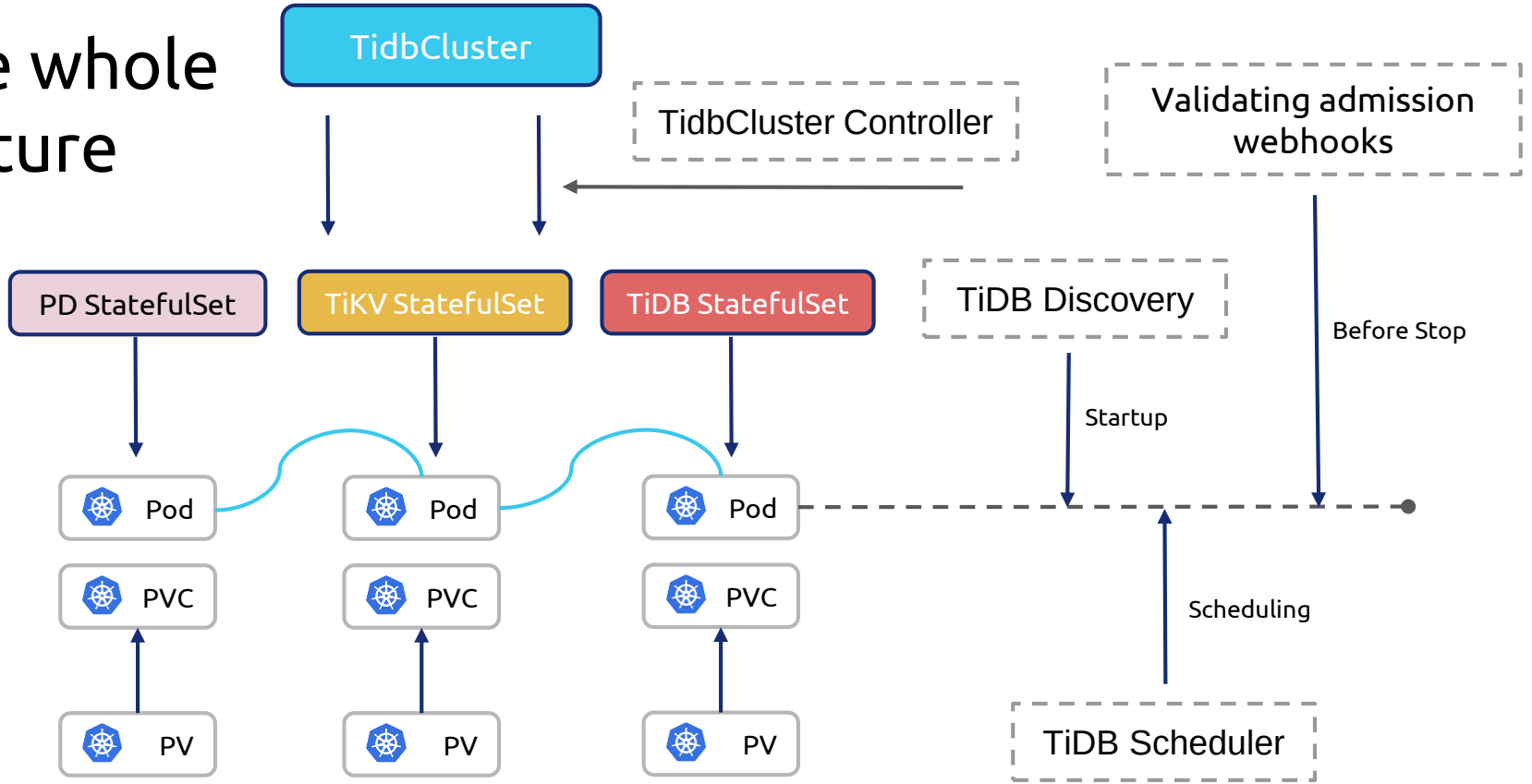


So we need a new API Object(CRD): TidbCluster

```
apiVersion: pingcap.com/v1alpha1
kind: TidbCluster
metadata:
  name: demo
spec:
  pd:
    image: pingcap/pd:v2.1.3
    replicas: 3
    ...
  tikv:
    image: pingcap/tikv:v2.1.3
    replicas: 5
    ...
  tidb:
    image: pingcap/tidb:v2.1.3
    replicas: 2
    ...
```



The whole picture



Thank you!

https://github.com/pingcap/tidb-operator