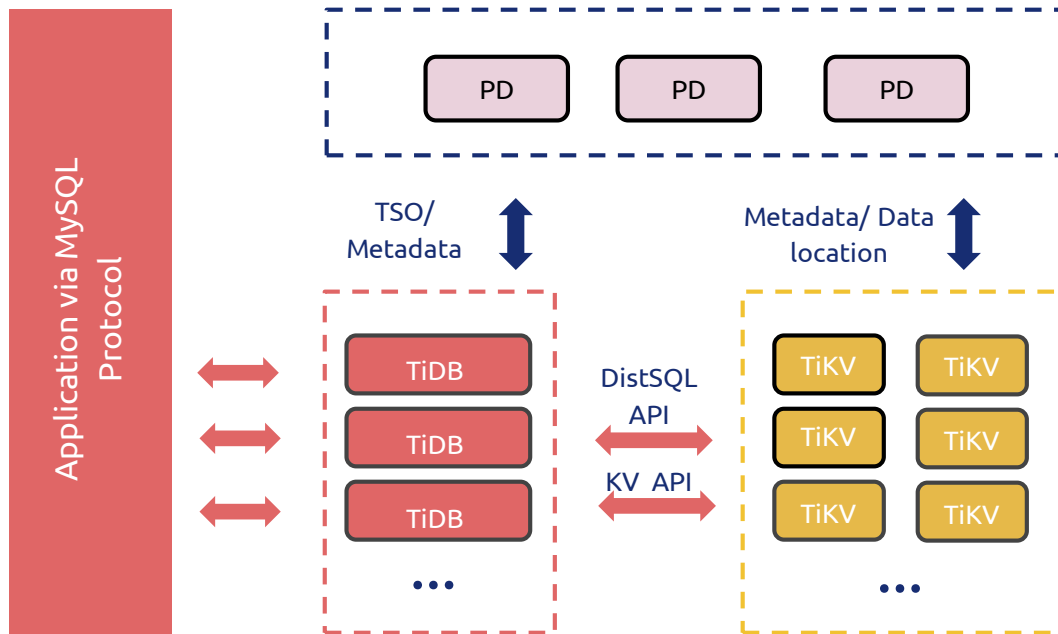# TiDB Operator 实现原理解析

weekface@PingCAP

About me

weekface (常军昌), software engineer, PingCAP

Focus on Cloud TiDB

changjunchang@pingcap.com

About TiDB

Traditional Deployment

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

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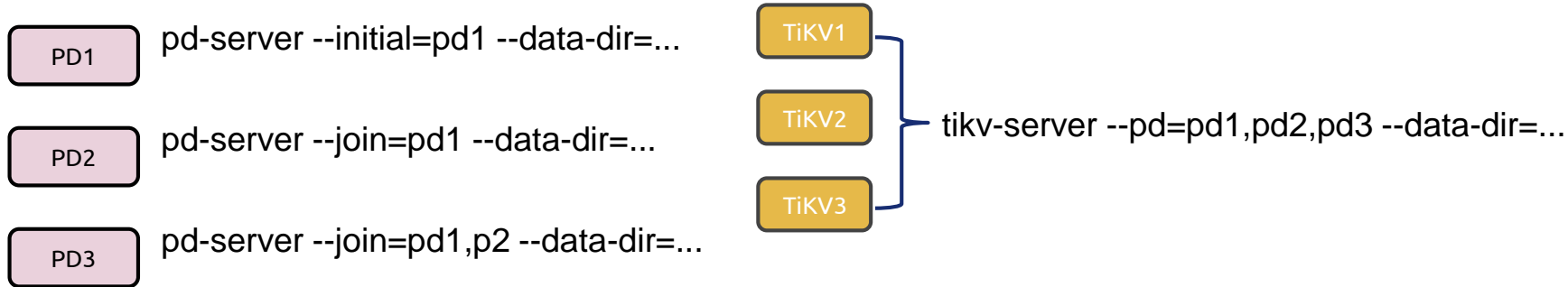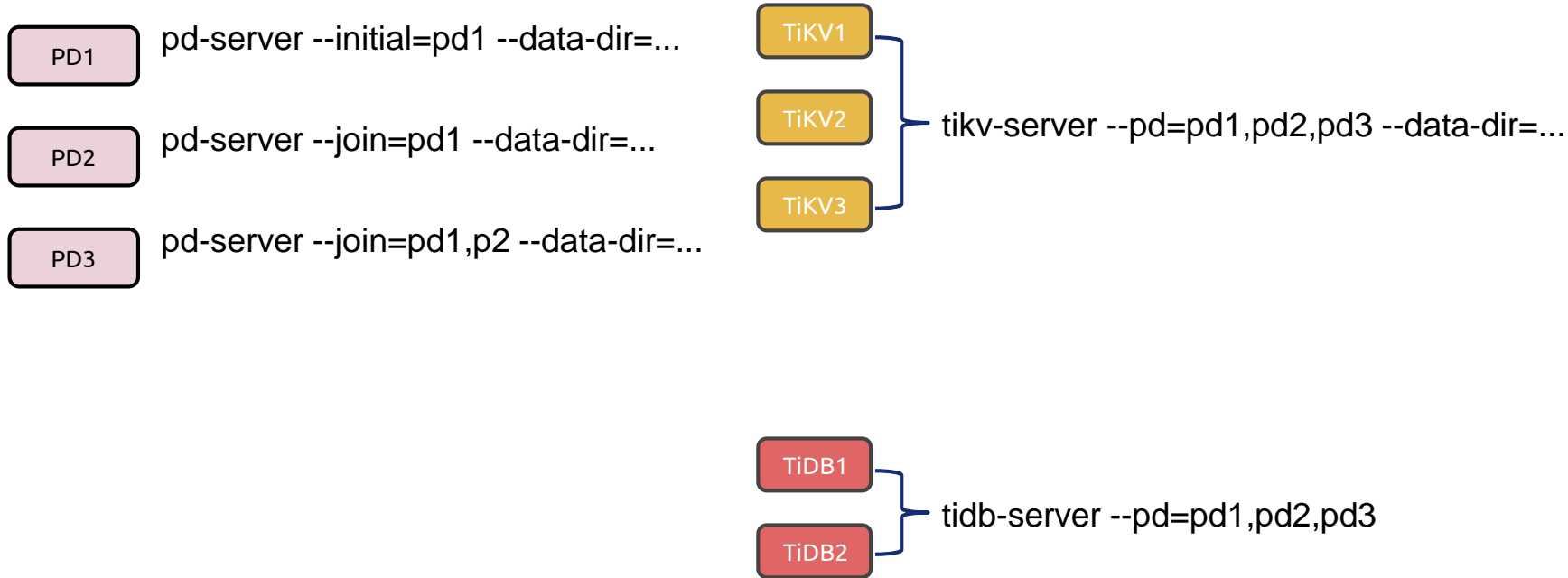
PD2 pd-server --join=pd1 --data-dir=...

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

TiKV1
TiKV2 tikv-server --pd=pd1,pd2,pd3 --data-dir=...
TiKV3

Traditional Deployment

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

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

PD3   pd-server --join=pd1,p2 --data-dir=...
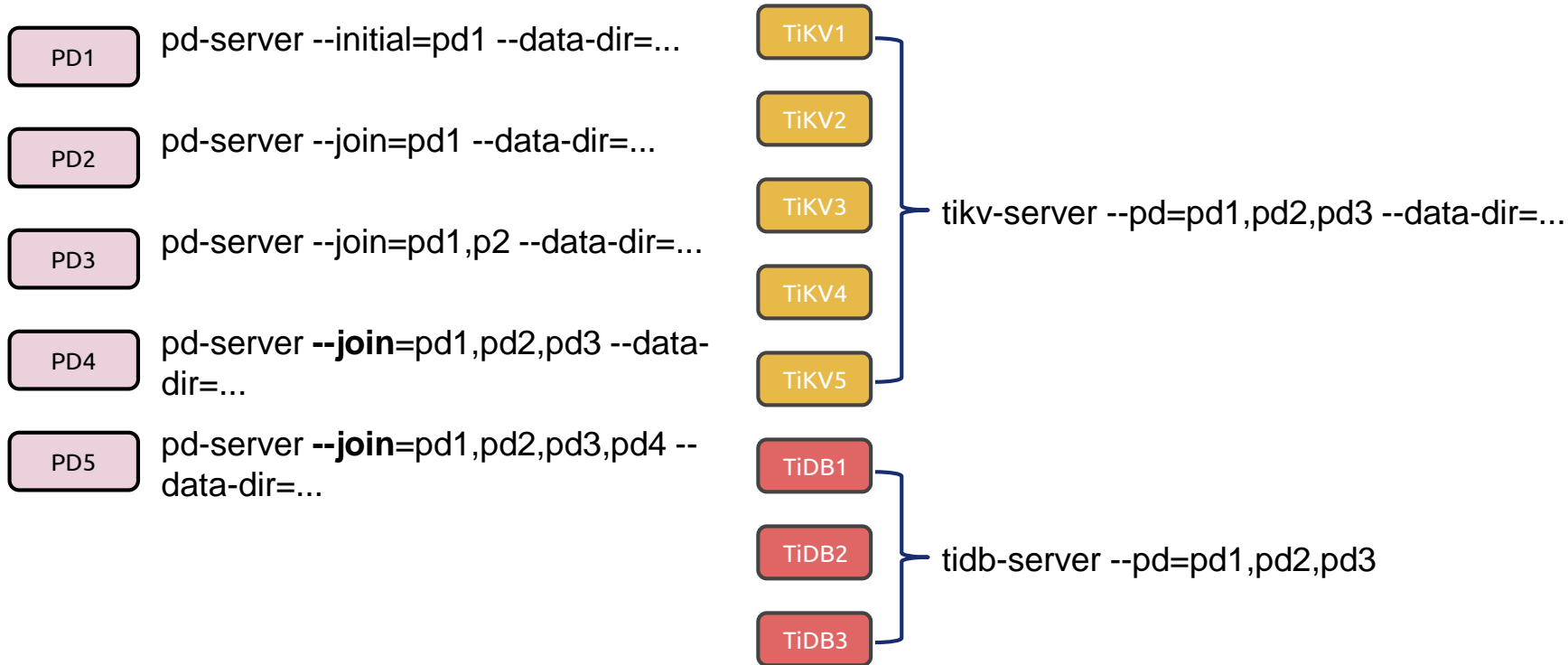
TiKV1
TiKV2
TiKV3   tikv-server --pd=pd1,pd2,pd3 --data-dir=...

TiDB1
TiDB2   tidb-server --pd=pd1,pd2,pd3

Traditional Deployment

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

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

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

PD4 pd-server **--join**=pd1,pd2,pd3 --data-dir=...

PD5 pd-server **--join**=pd1,pd2,pd3,pd4 --data-dir=...

TiKV1
TiKV2
TiKV3 tikv-server --pd=pd1,pd2,pd3 --data-dir=...
TiKV4
TiKV5

TiDB1
TiDB2 tidb-server --pd=pd1,pd2,pd3
TiDB3

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.
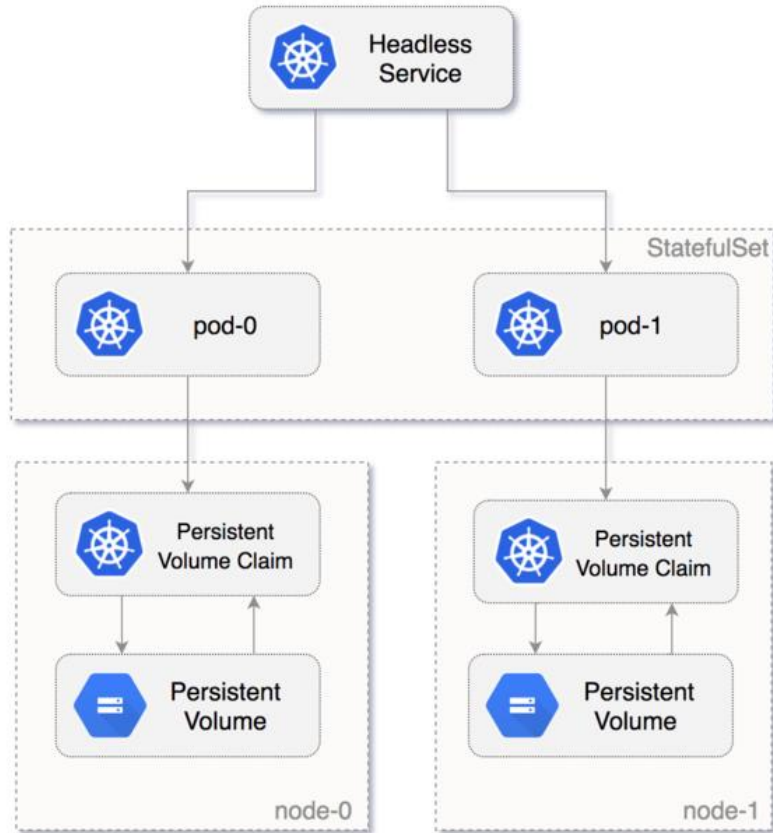
Based on StatefulSet - Persistent storage

PersistentVolumes associated with the Pods' PersistentVolumeClaims

Based on StatefulSet - Create/Upgrade Order

Create and manage three StatefulSet objects: PD StatefulSet, TiKV StatefulSet, TiDB StatefulSet
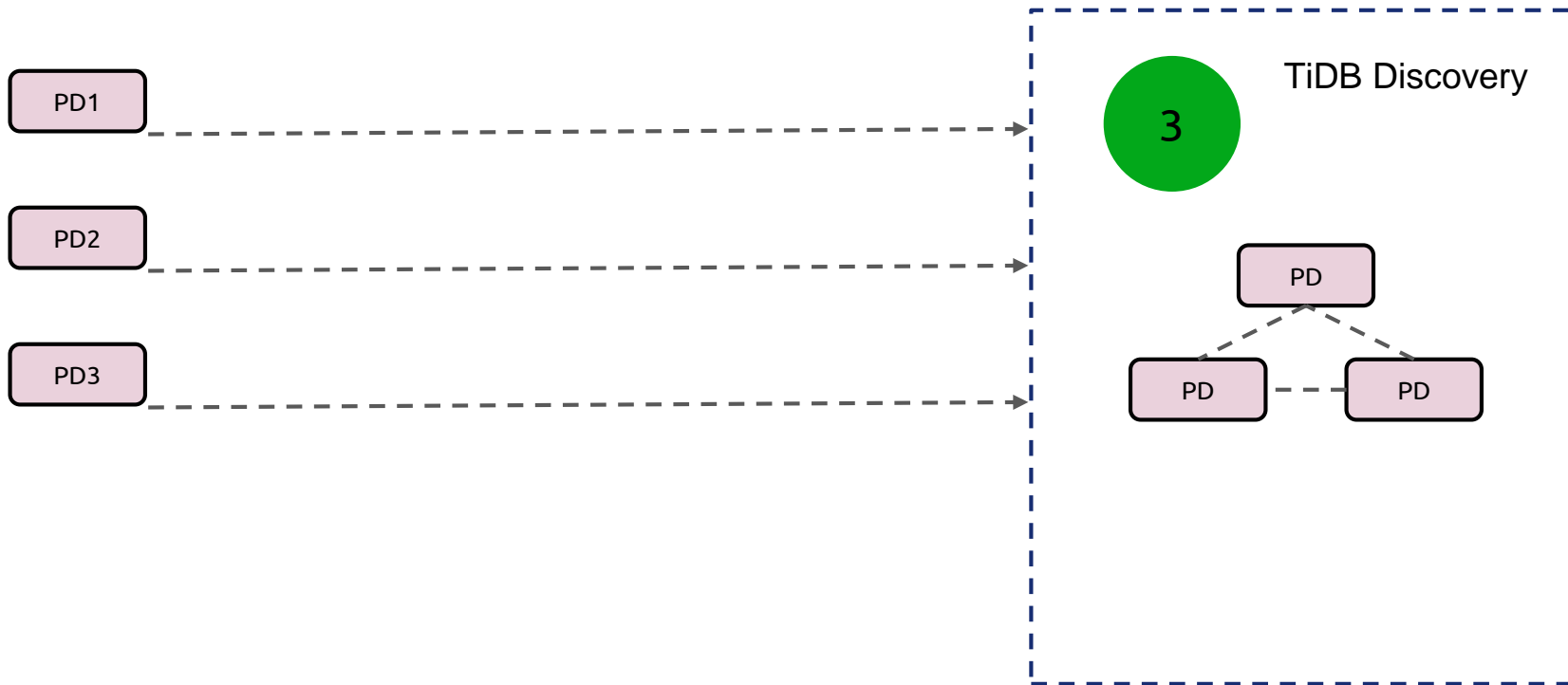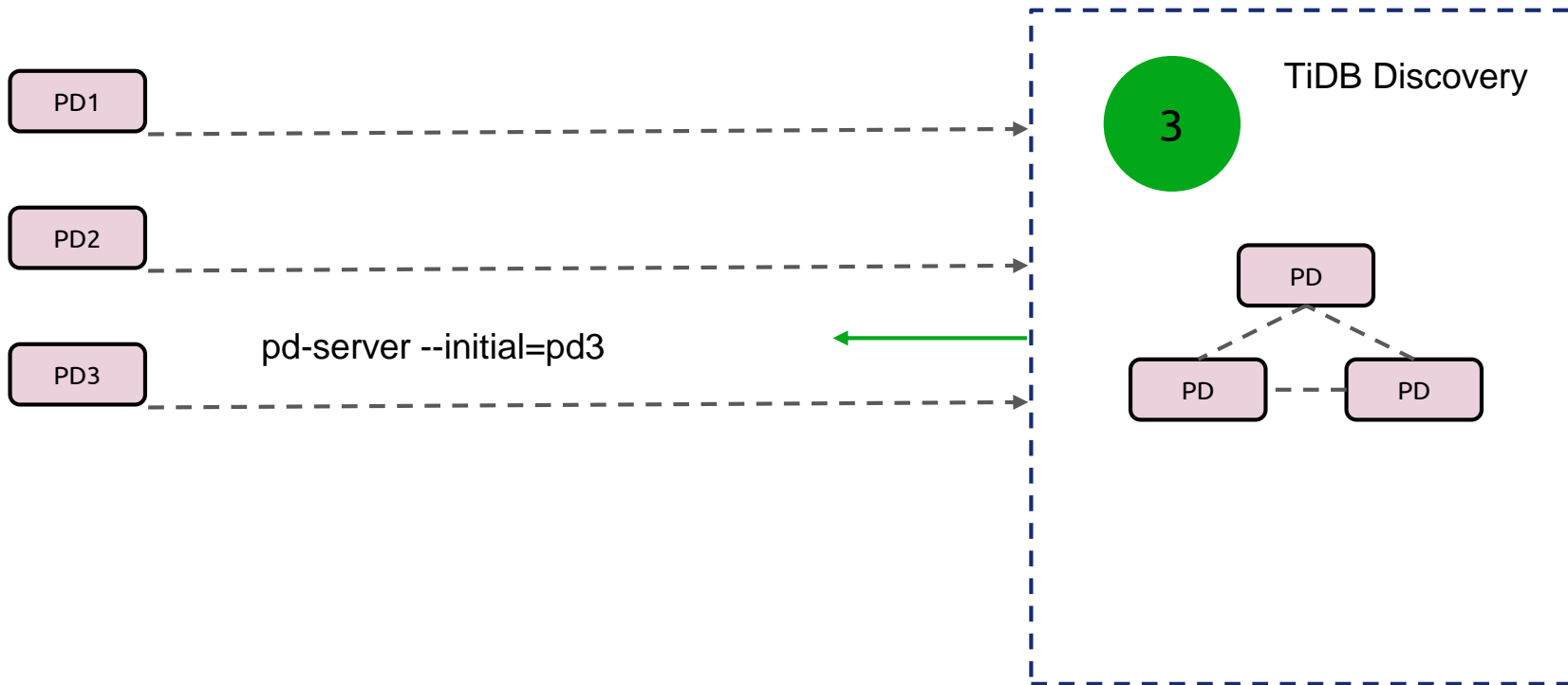
Based on StatefulSet - PD Startup Args
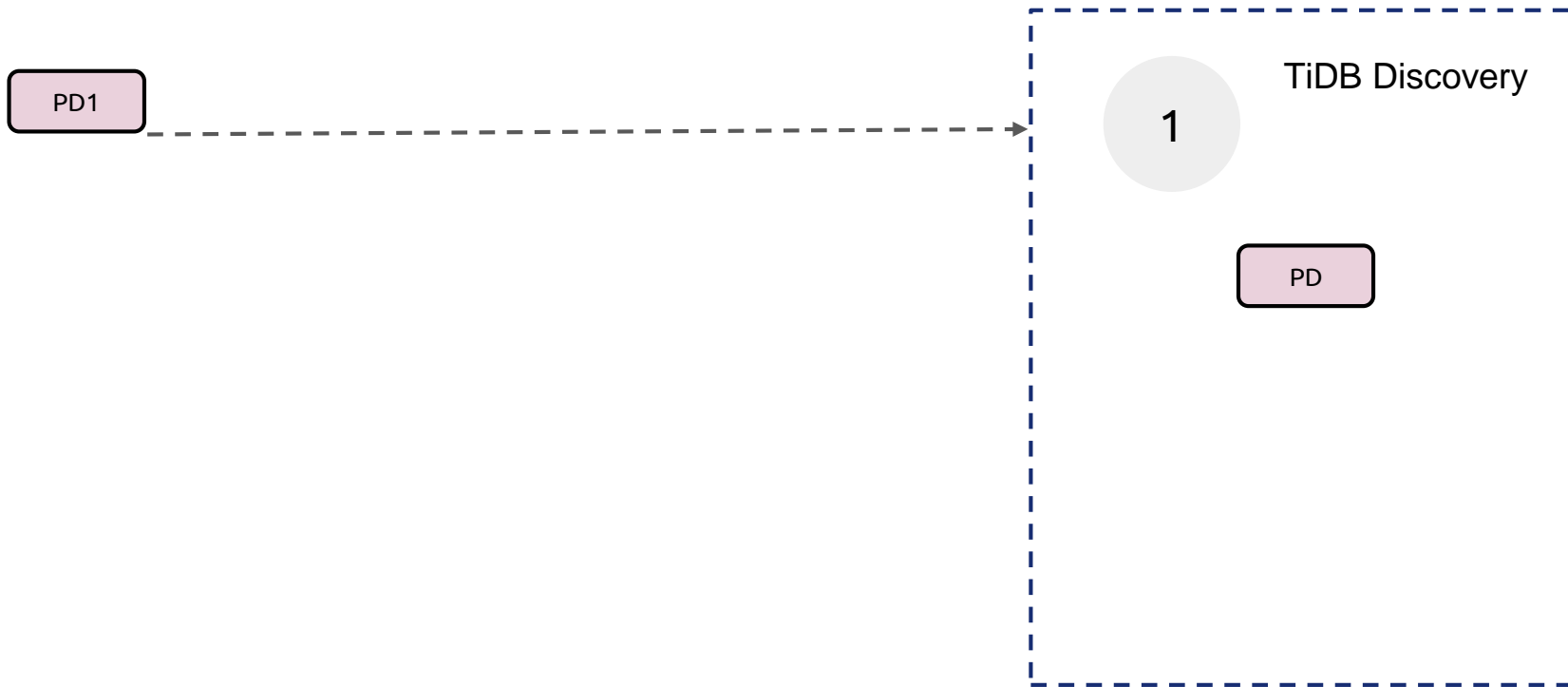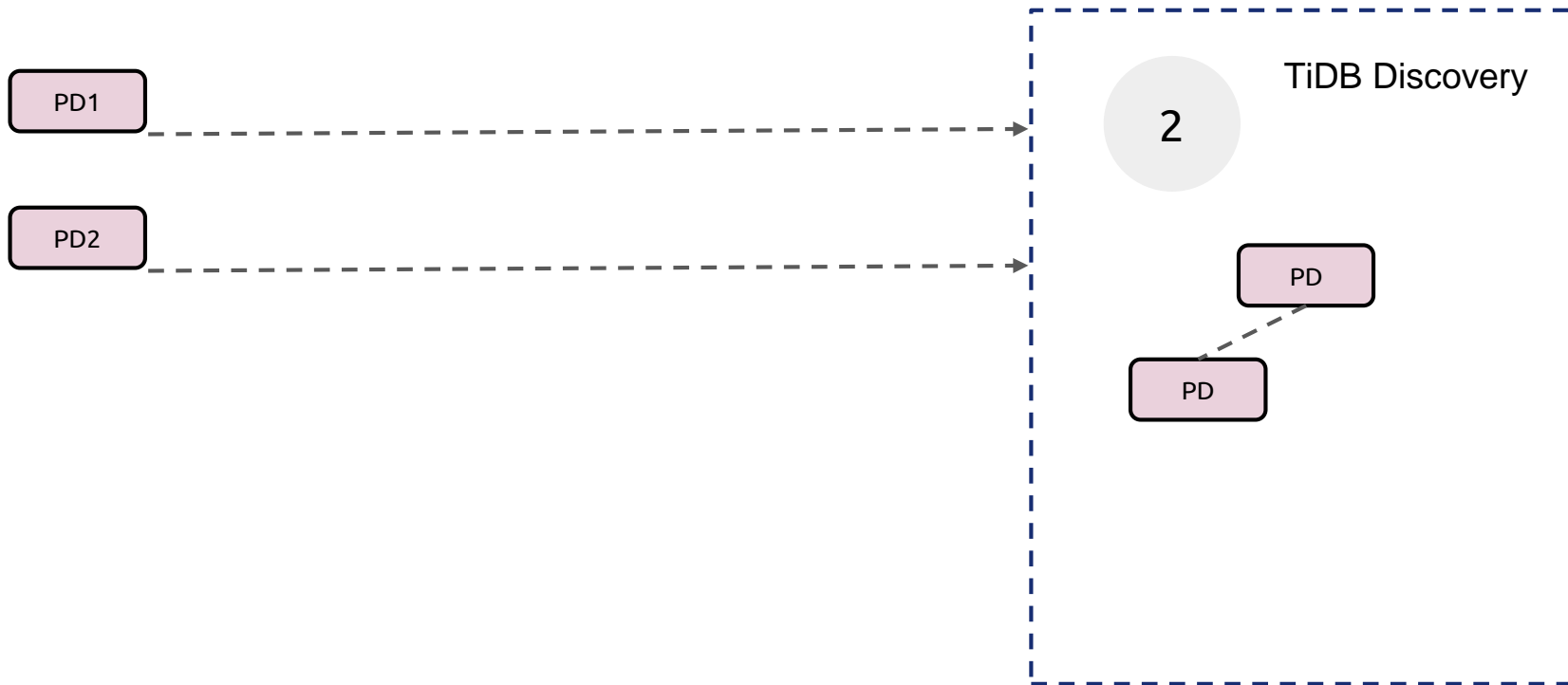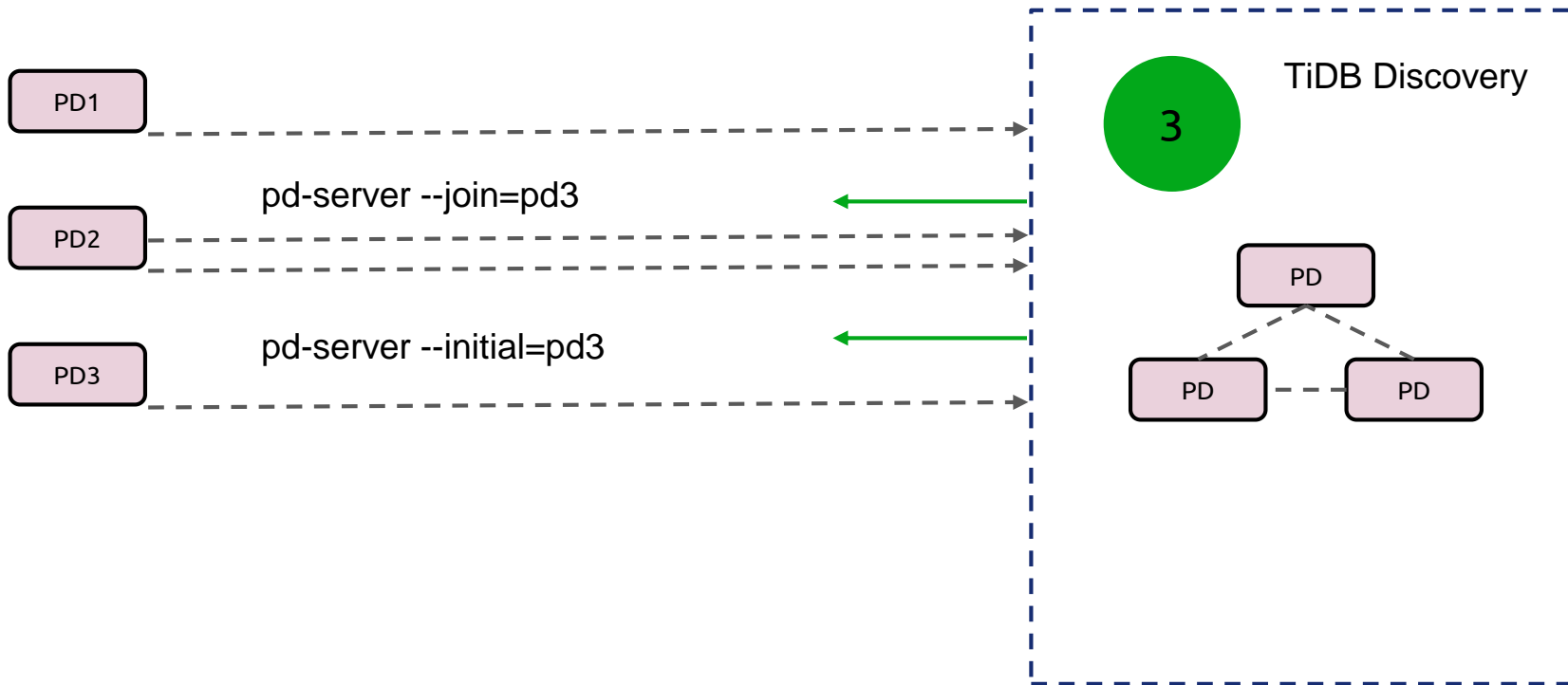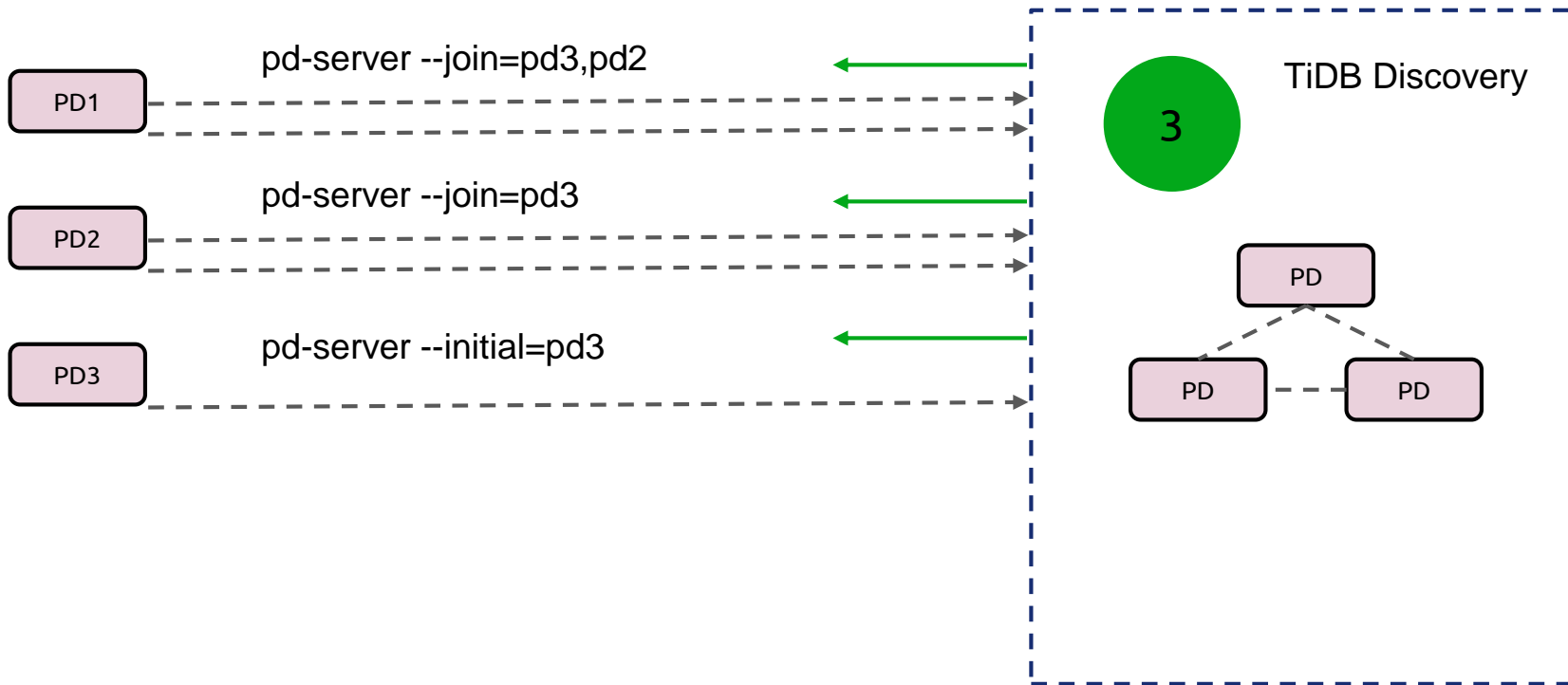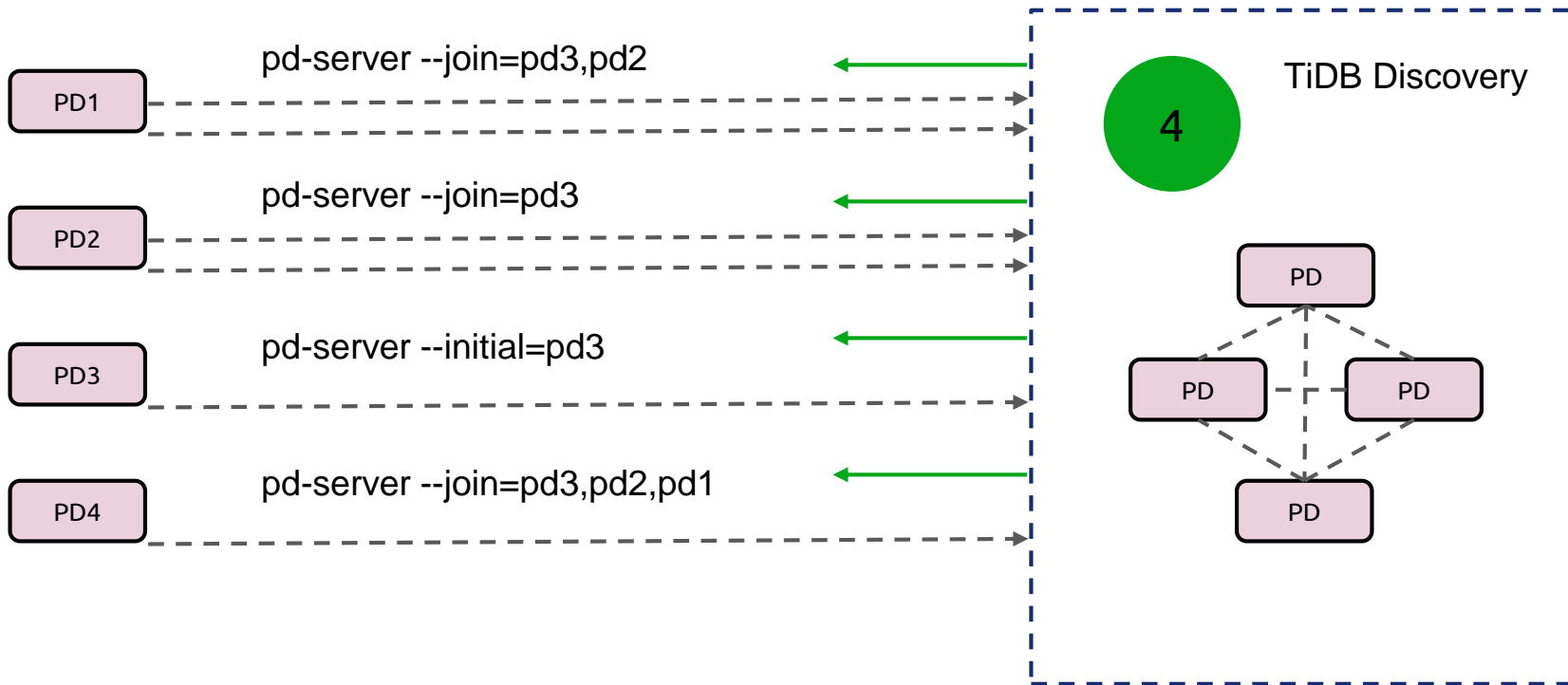
Based on StatefulSet - PD Startup Args

Based on StatefulSet - PD Startup Args

Based on StatefulSet - PD Startup Args

Based on StatefulSet - PD Startup Args

PD1

PD2

pd-server --join=pd3

PD3

pd-server --initial=pd3

TiDB Discovery

3

PD

PD   PD

Based on StatefulSet - PD Startup Args

pd-server --join=pd3,pd2

PD1

pd-server --join=pd3

PD2

pd-server --initial=pd3

PD3

TiDB Discovery

3

PD

PD   PD

Based on StatefulSet - PD Startup Args
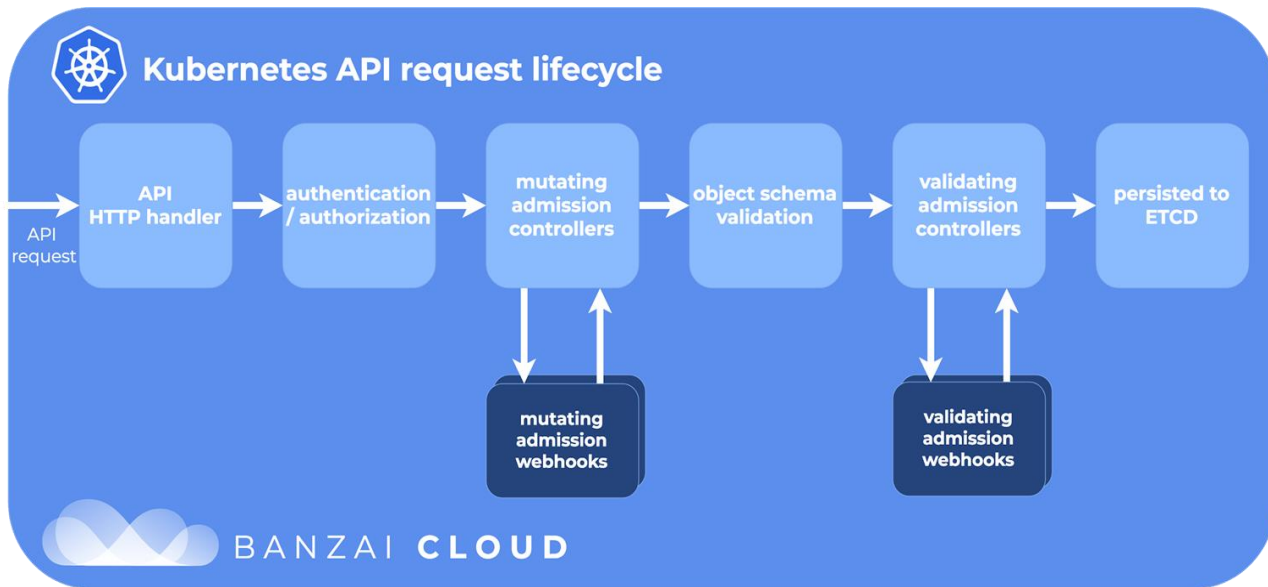
Based on StatefulSet - Before Stop/Offline

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
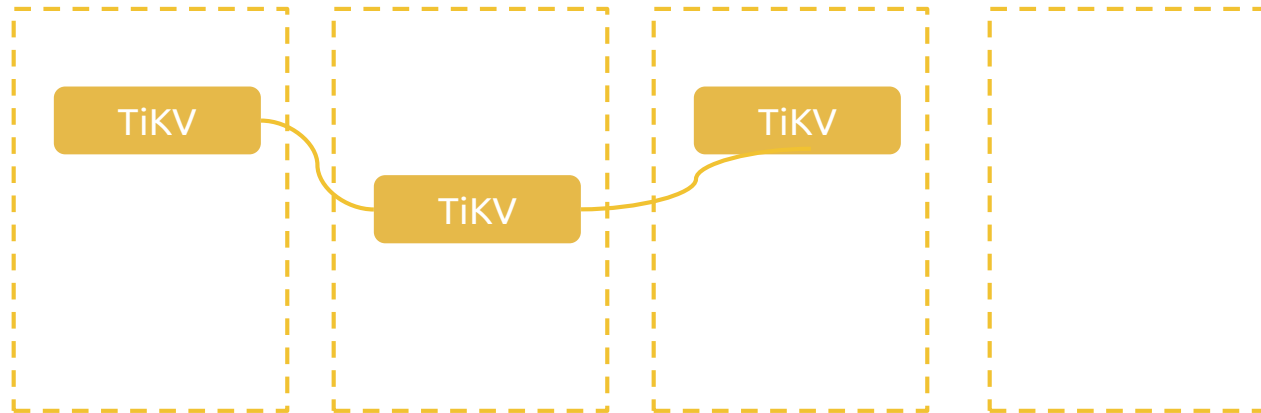
TiDB Scheduler

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:
...
```
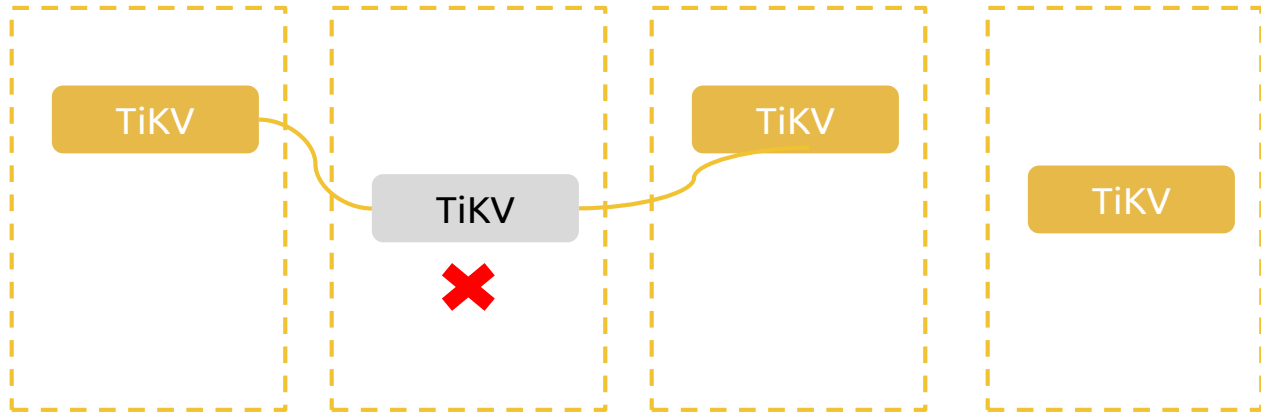
TiDB Sheduler

Kube Scheduler

Pod

TiDB Scheduler
Extender

predicates: HA ...
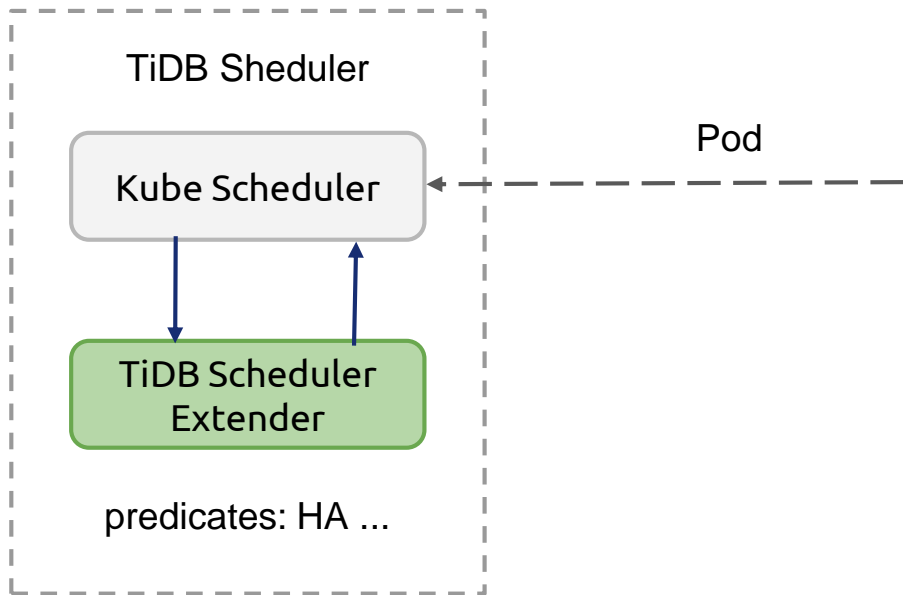
Failover

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
   ...
```
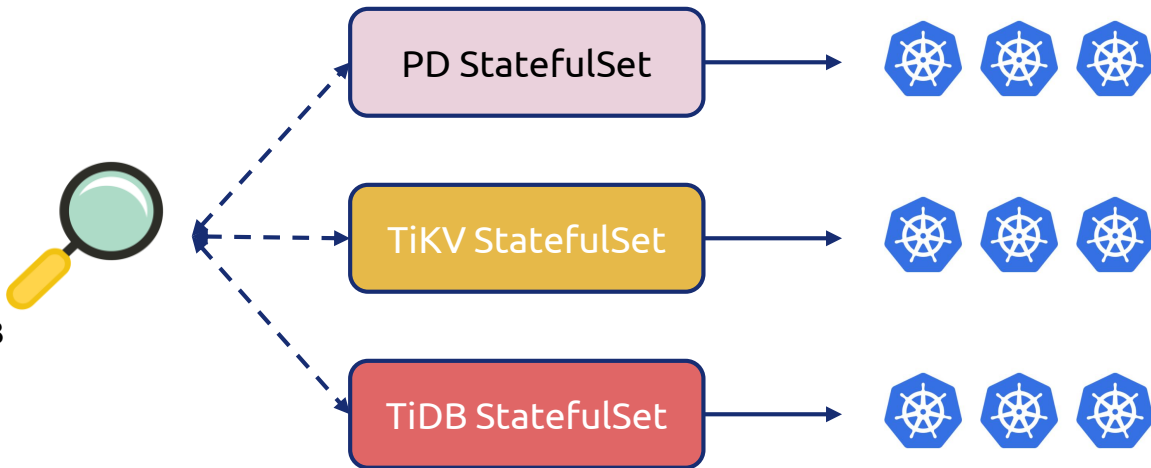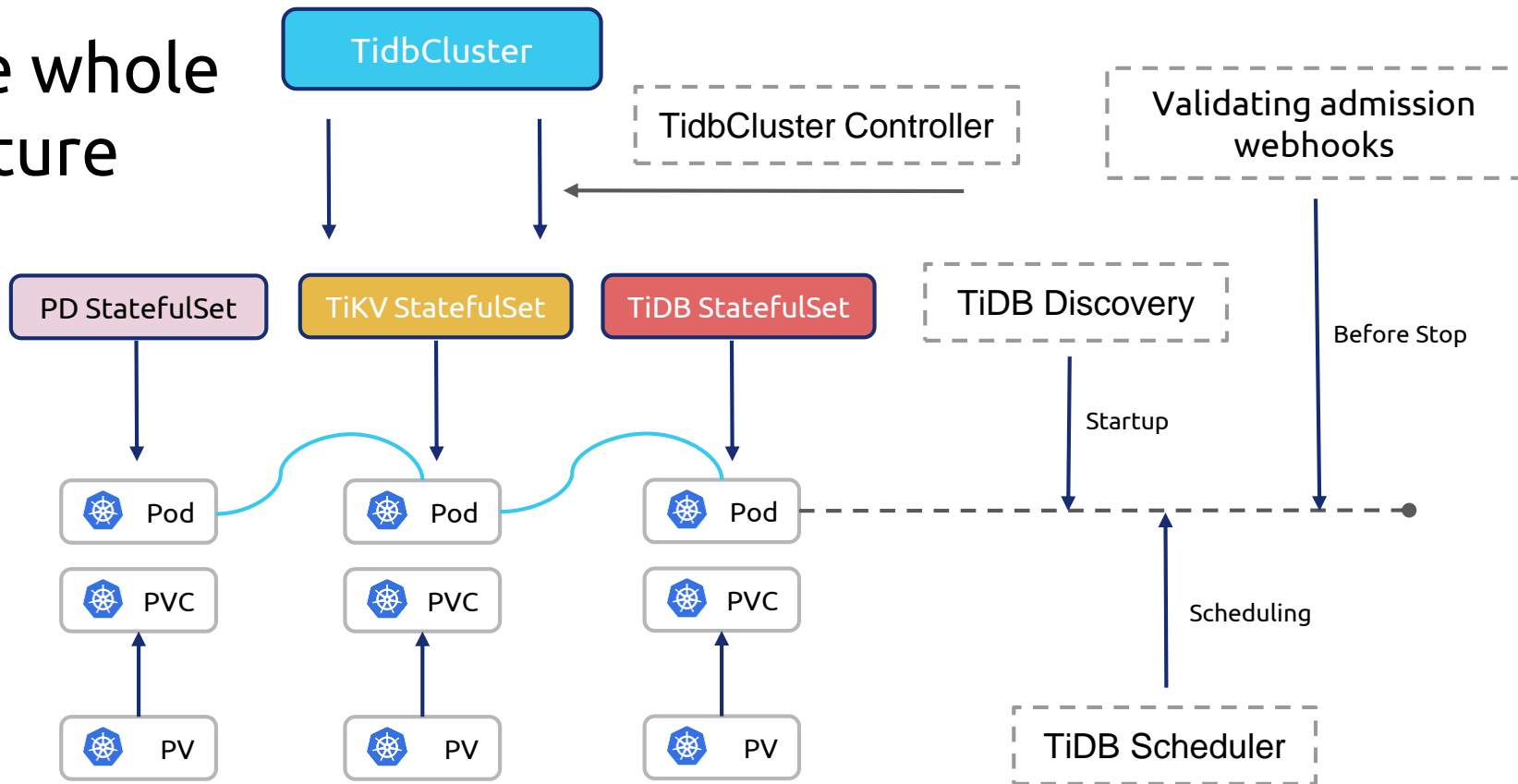
The whole picture

TidbCluster

TidbCluster Controller

Validating admission webhooks

PD StatefulSet

TiKV StatefulSet

TiDB StatefulSet

TiDB Discovery

Before Stop

Pod

Pod

Pod

Startup

PVC

PVC

PVC

Scheduling

PV

PV

PV

TiDB Scheduler

Thank you!

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