KubeCon CloudNativeCon

THE LINUX FOUNDATION

OPEN SOURCE SUMMIT

AI_dev

Open Source GenAI & ML Summit

China 2024

Kuaishou's 100% Resource Utilization Boost: 100K Redis Migration from Bare Metal to Kubernetes

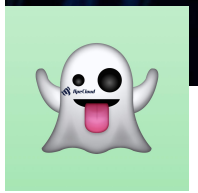
Yuxing Liu, Senior Software Engineer, Kuaishou
Xueqiang Wu, Director of R&D, ApeCloud | KubeBlocks Maintainer

KUAISHOU

KubeBlocks

About Me

- Former tech leader at Alibaba Cloud PolarDB-X
- Maintainer of open source PolarDB-X
- Alibaba 2021 Open Source Pioneer of the Year
- Currently serves as a maintainer for KubeBlocks
- Interested in operating systems, distributed systems, databases and more



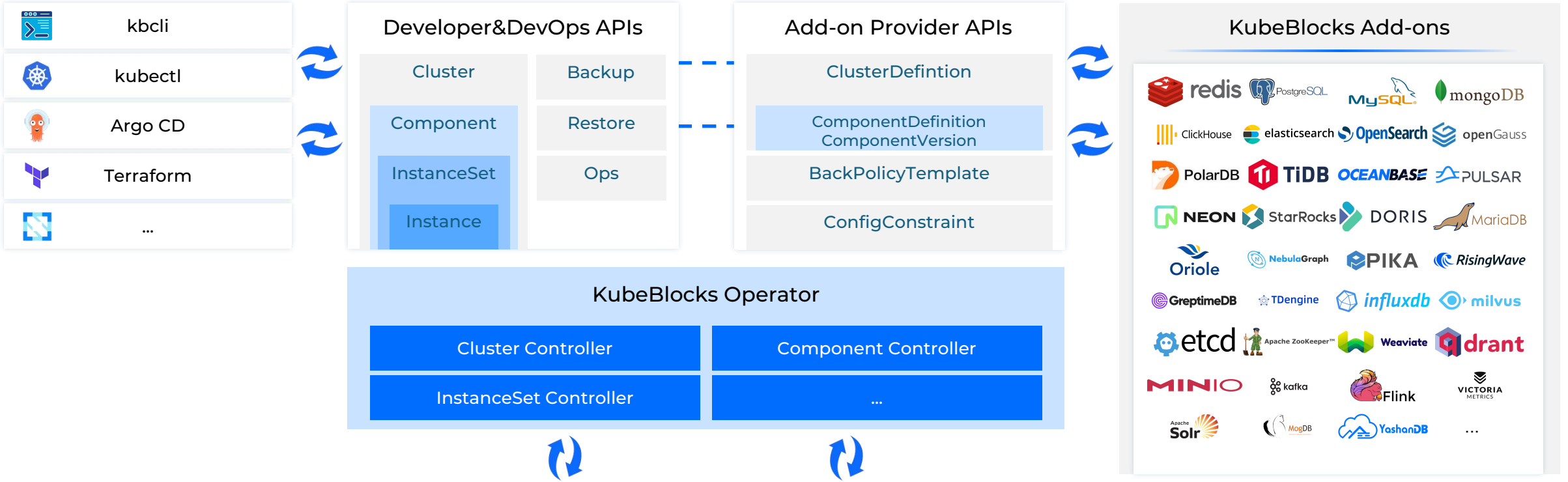
Xueqiang Wu
GitHub ID: free6om

Agenda

- What is KubeBlocks
- Problems KubeBlocks solved (single Redis cluster perspective)
- Multi Redis clusters and multi Kubernetes clusters (large-scale perspective)
- Q&A

Overview: What is KubeBlocks?

What is KubeBlocks



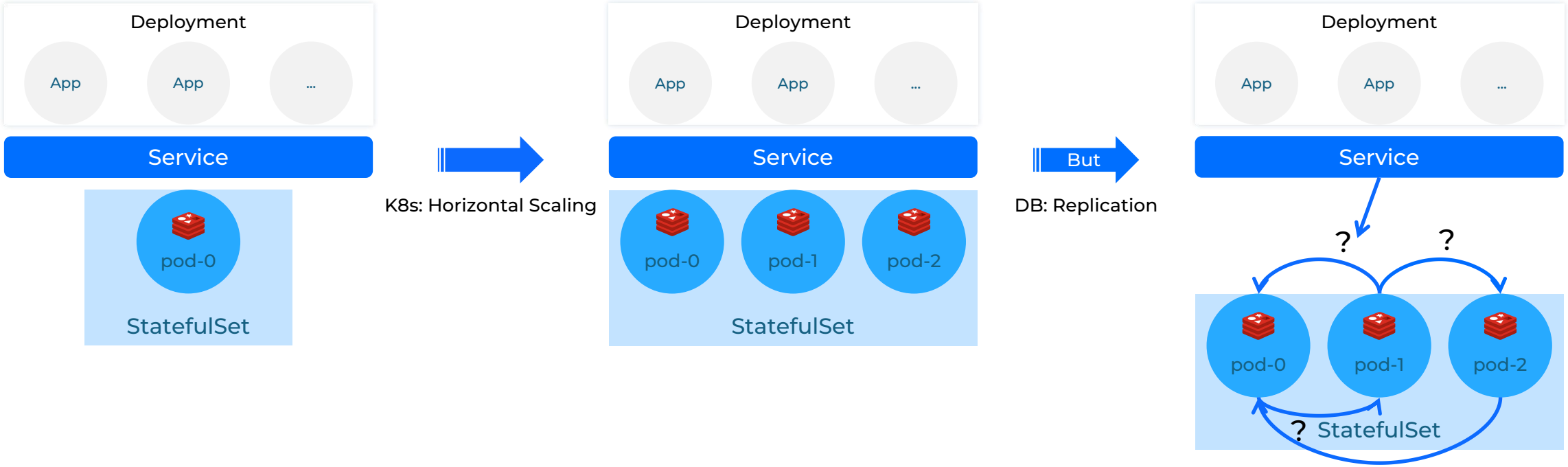
An open source Operator specially designed for running databases on Kubernetes

- 35 databases supported
- Extendable with add-on APIs
- Unified developer&DevOps APIs
- Support lifecycle management
- Support backup&recovery (with PITR)
- Integrated with Prometheus and Grafana
- More

Single Redis cluster perspective: what problems KubeBlocks solved to make run databases on K8s better?

How to handle data replication

Running Databases on Kubernetes is Challenging



- Single point of failure
- Throughput bottleneck
- High risk of data loss

- How to select the pod with ReadWrite ability
- How to know the right replication relations
- How to do an Update to minimize the service outage time
- ...

How to handle data replication

Role: A New Abstract Layer Above Stateful

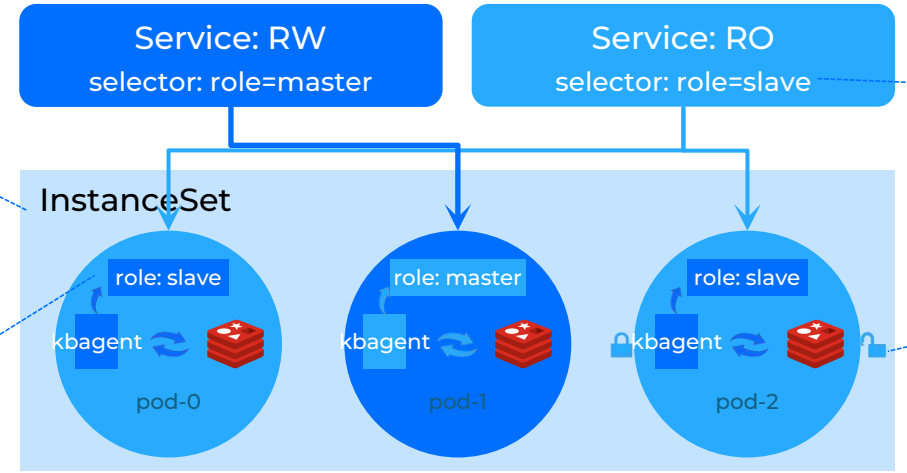
1. InstanceSet API

- A new workload API dedicated for databases
- Same sticky identifier for each Pod like StatefulSet
- Add a Role layer above stateful

2. Each Pod has a role label

① Role probe
② Trigger role event
③ Watch role event
④ Update Pod role label

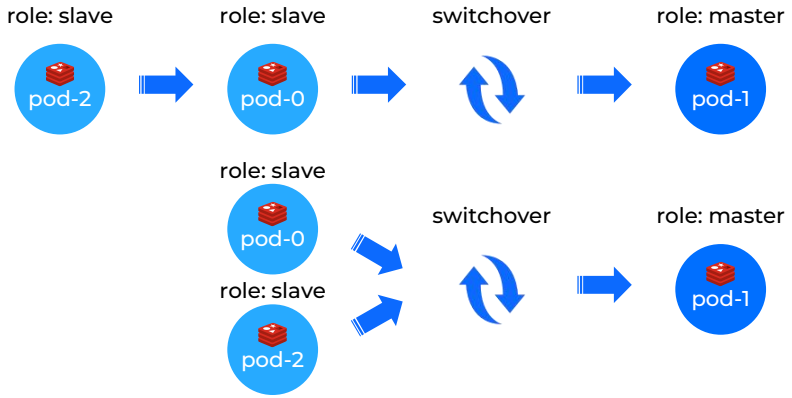
- A general role probe framework
- Multiple role event mechanism
- Network partition
- Event lag



3. Role-based service selector

4. Role-based membership reconfiguration

- member join
- member leave
- switchover



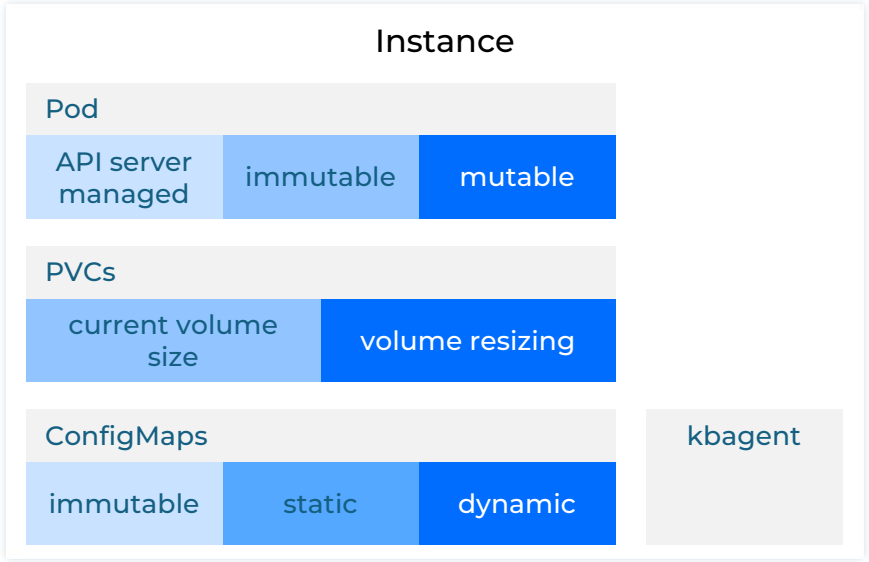
5. Role-based update strategy

- Serial, Parallel, BestEffortParallel
- Planned switchover

RTO from minutes to seconds

How to achieve high availability

In-place Instance Updates

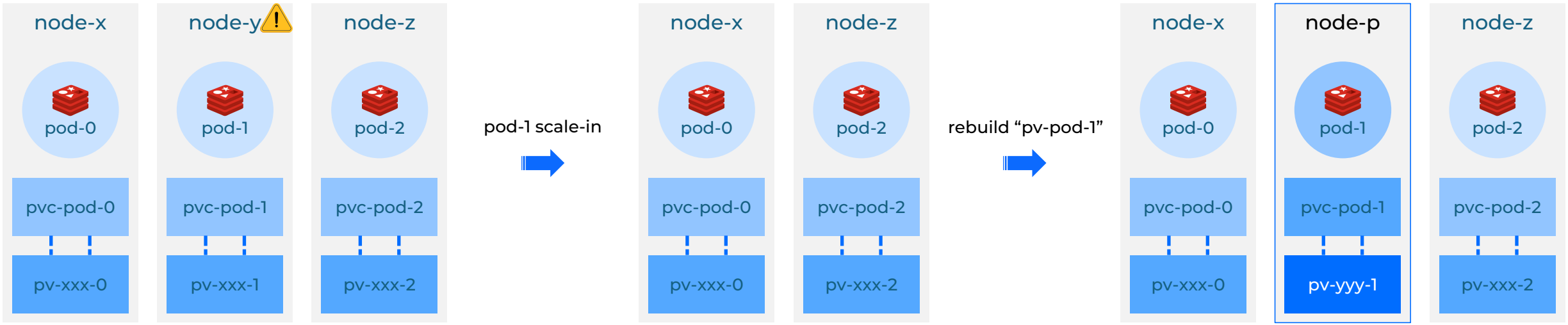


- In-place Pod update
- Volume resizing
- Configuration dynamic reloading

Uptime & MTBF increase
RTO decreases from seconds to near-zero

How to achieve high availability

Instance Rebuild via Specified Instance Scale-in



Node with name node-y down due to disk failure

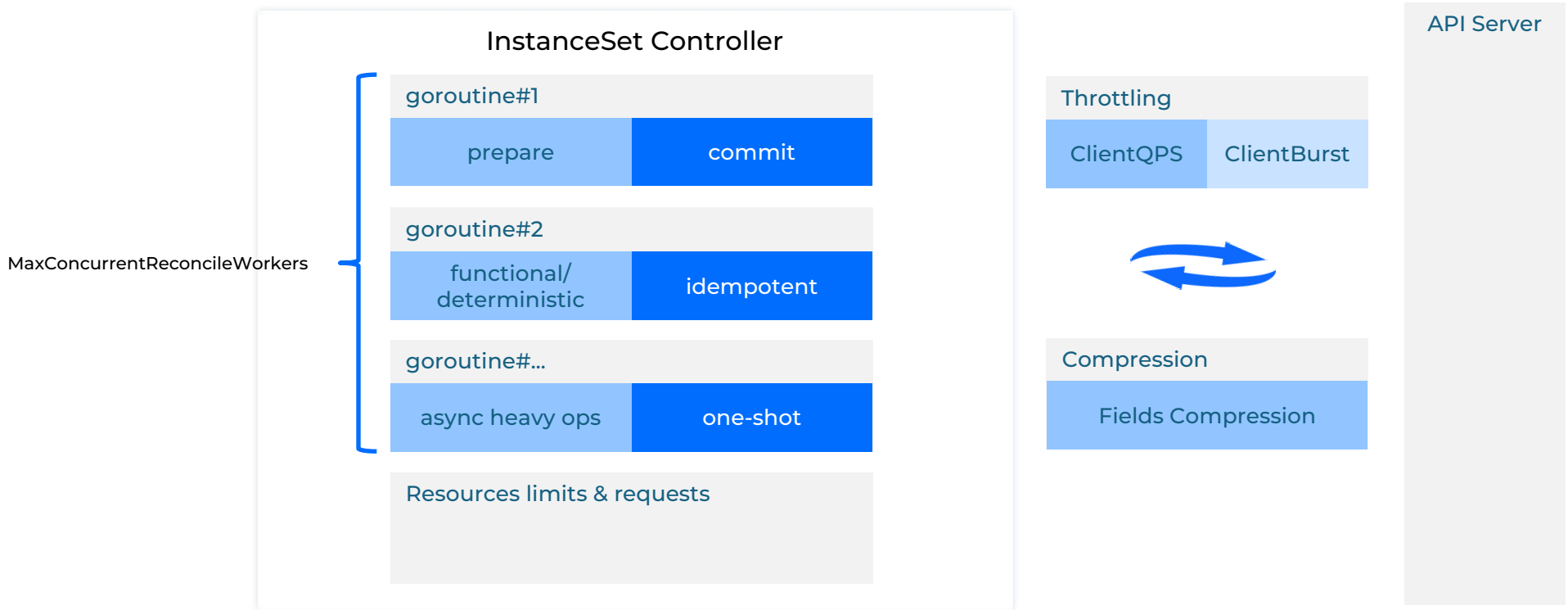
node-y is taken offline to repair, using the ‘specified instance scale-in’ feature to remove pod-1 and pvc-pod-1.

A new node with name “node-p” joins the K8s cluster, using the ‘instance rebuild’ feature to populate the same data as pv-xxx-1 to pv-yyy-1.

RTO from hours to minutes

How to handle large-scale cluster

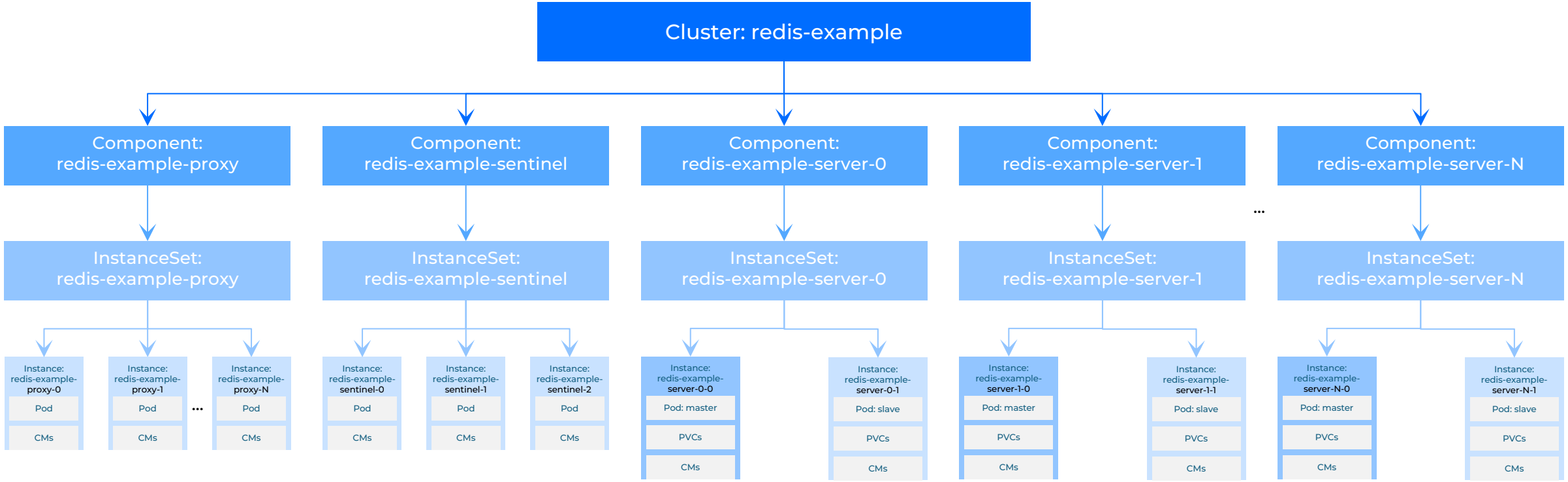
Operator P10K Problem



Definition:
Similar to the C10K problem, the P10K problem is the problem of optimizing Operator to handle a large number of Pods owned by a single CR at the same time.

Use case:
A single Redis cluster with close to 10,000 Pods

Recap



- Role
  - Role-based service selector
  - Role-based membership reconfiguration
  - Role-based update strategy
- High availability
  - Pod in-place update
  - Volume resizing
  - Configuration dynamic reloading
  - Instance rebuilding

- Operator P10K problem
  - Throttling
  - Object compression
  - Resources limitation
  - MaxConcurrenctReconcileWorkers
  - Two-phase reconciliation
  - Asynchonize heavy operations

Large-scale perspective: how Kuaishou uses KubeBlocks to run multi Redis instances on multi Kubernetes clusters?

About Me - Yuxing Liu

- Have worked at Alibaba Cloud, support the Commercialization of Cloud-Native Application Delivery
- Maintainer of the open source projects CNCF/Dragonfly and CNCF/Sealer
- Currently focused on the cloud-native transformation of Kuaishou's database business
- Github ID: starnop

Kuaishou - Redis Introduction

Feature? - Extremely Large Scale：

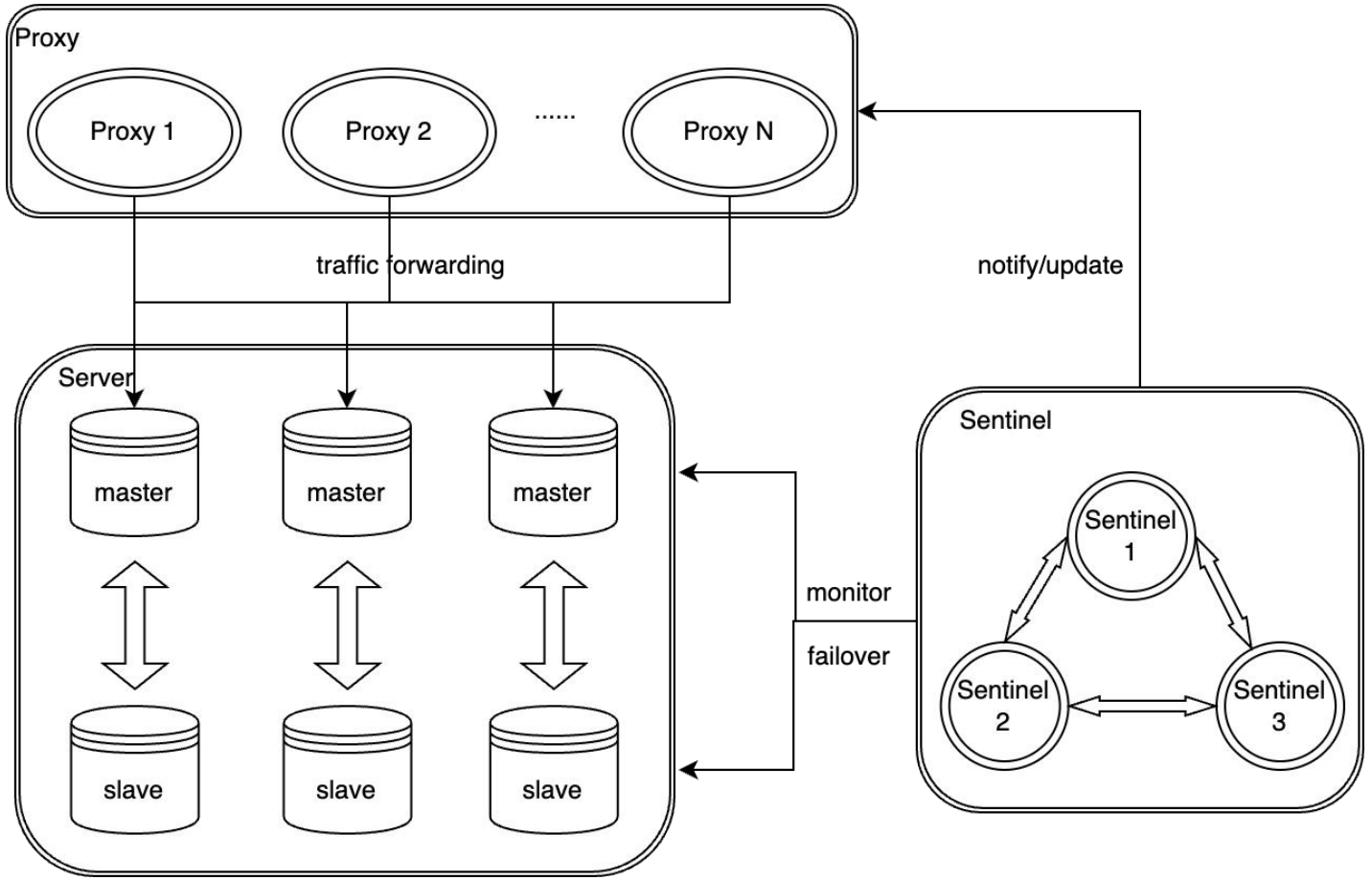
1. Extremely Large Total Instance Count
2. Single Cluster Instance Count Exceeding 10,000

Why Cloud-Native?

1. Improve Resource Utilization and Reduce Resource Costs
2. Decouple infrastructure and improve business agility

Organization Structure？

1. The Redis DBA Team
2. The Cloud-Native team

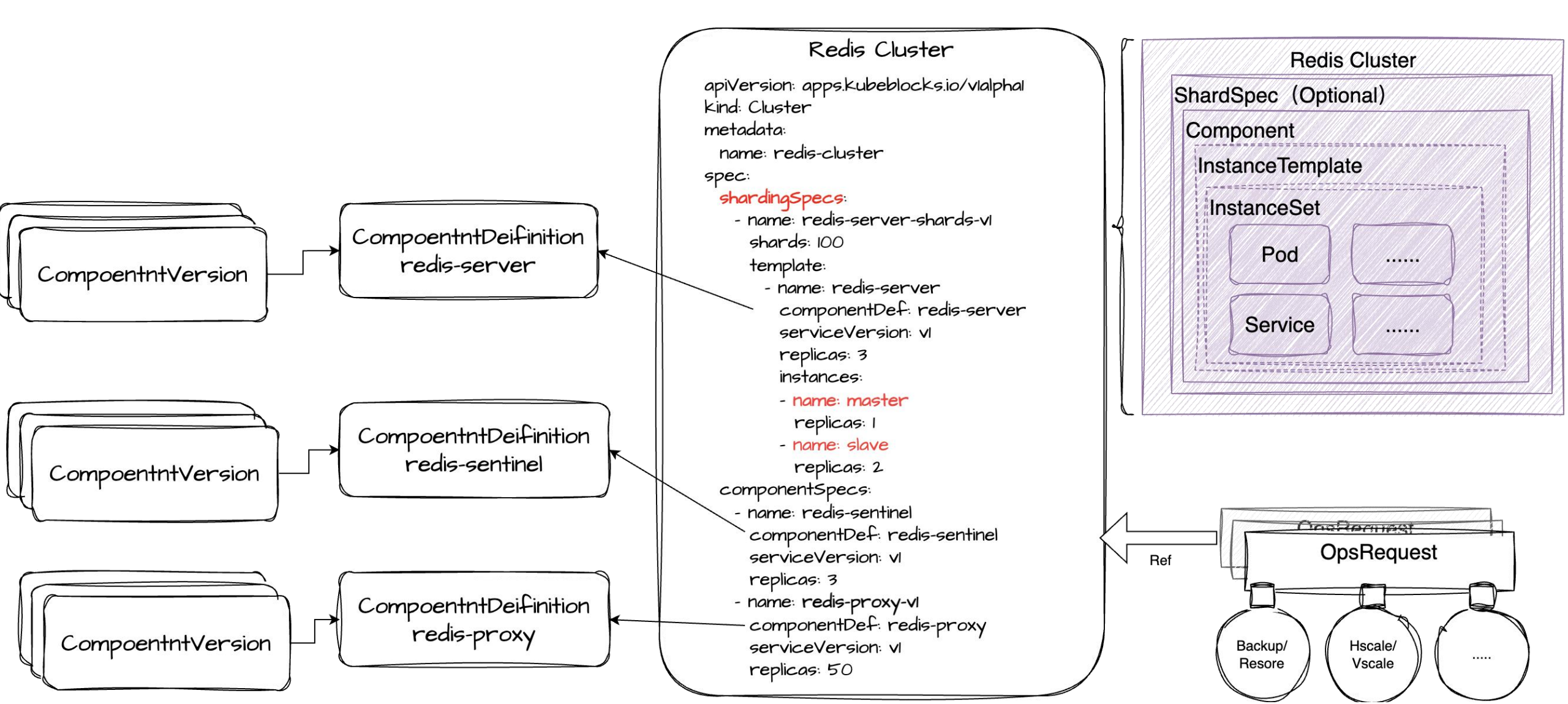


Why KubeBlocks

KubeBlocks - APIs that are more oriented towards stateful services

1. Powerful abilities: role-based management capabilities, etc.
2. High technical reusability： supports multiple database services
3. Low cloud-native transformation cost:
    1. Separates instance lifecycle management and O&M logic
    2. Provides process-oriented APIs through the OpsRequest Object

Redis Cluster Orchestration Definition

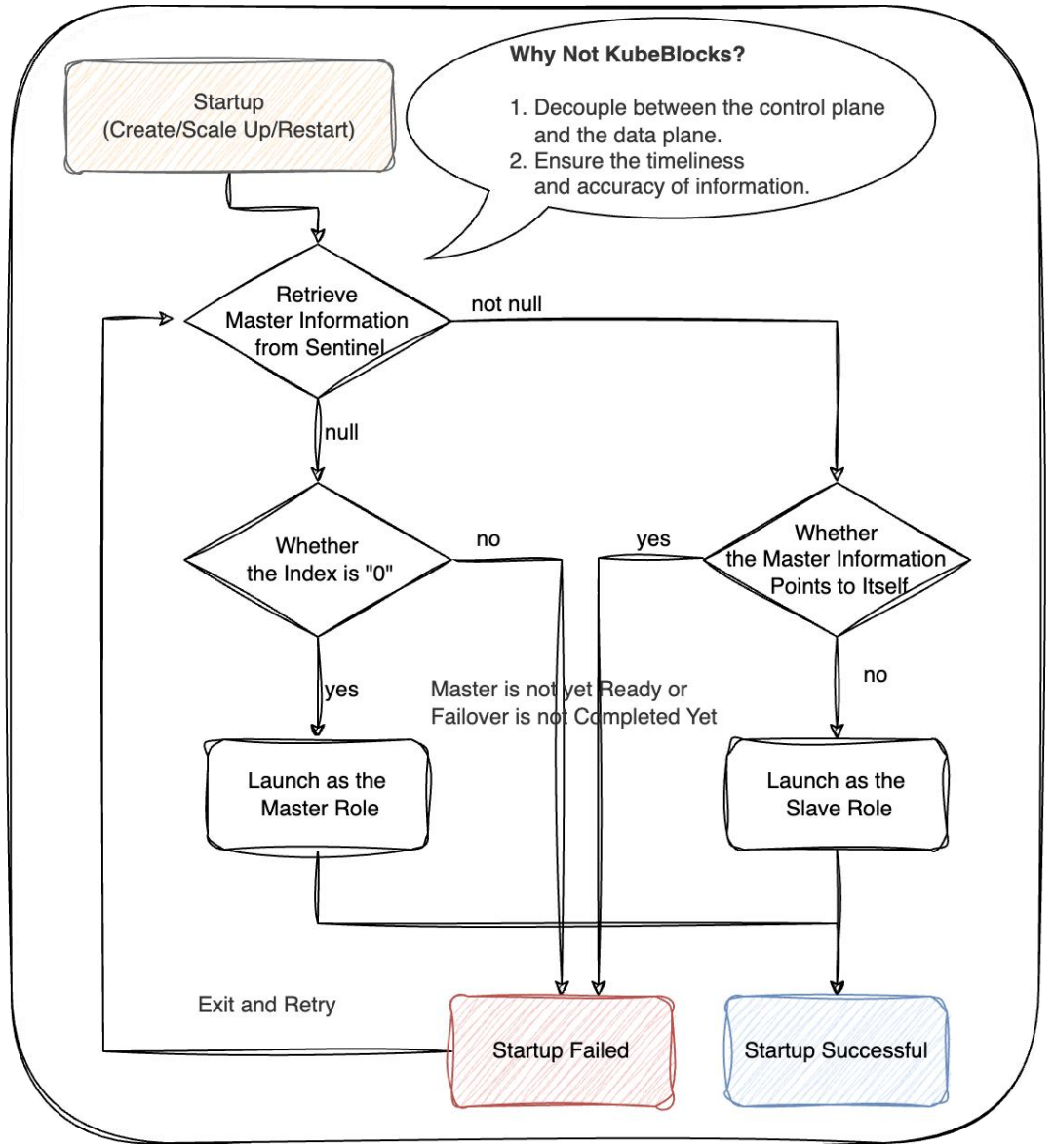


Role Management

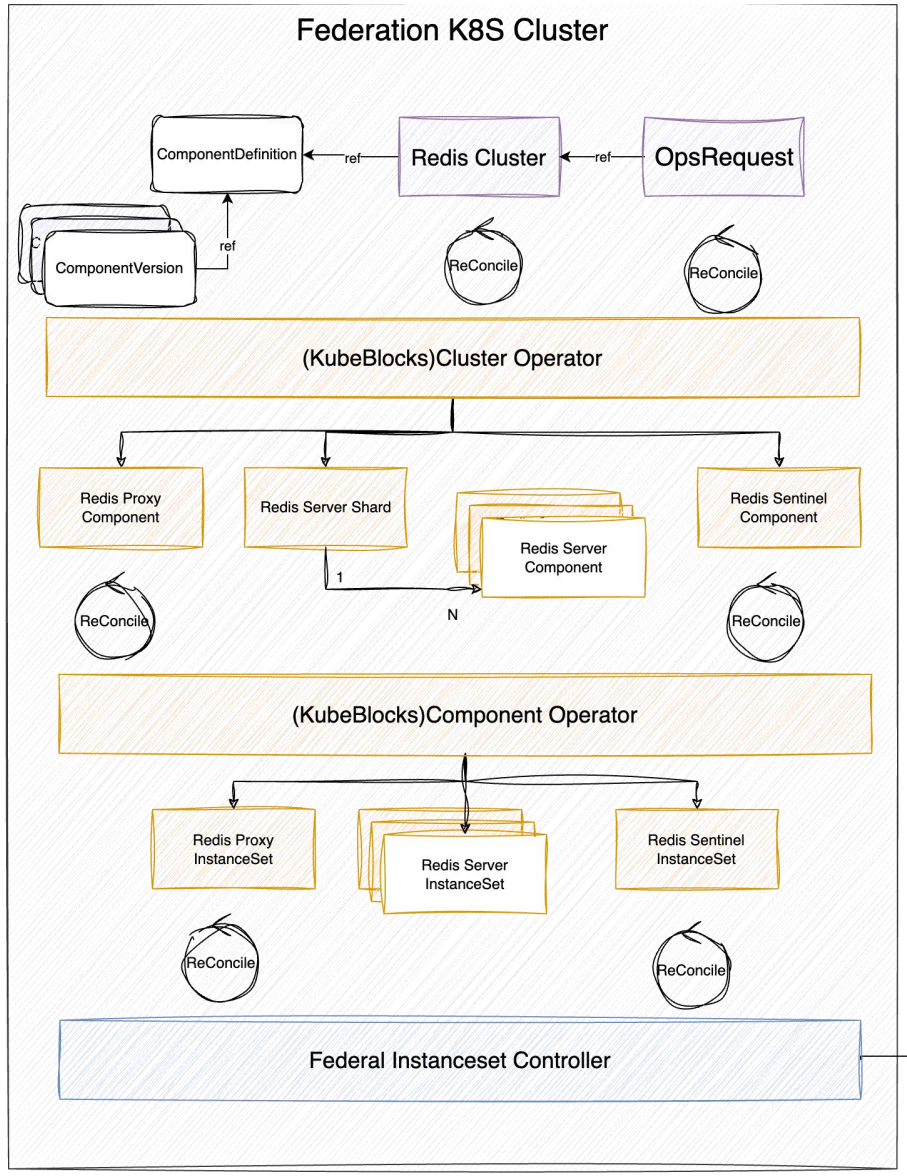
1. How to maintain the correct role relationships?

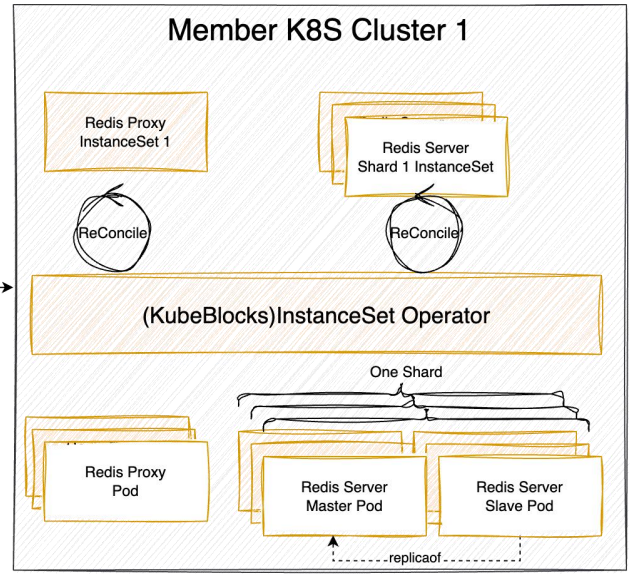
2. How to implement Role-Based O&M Management?

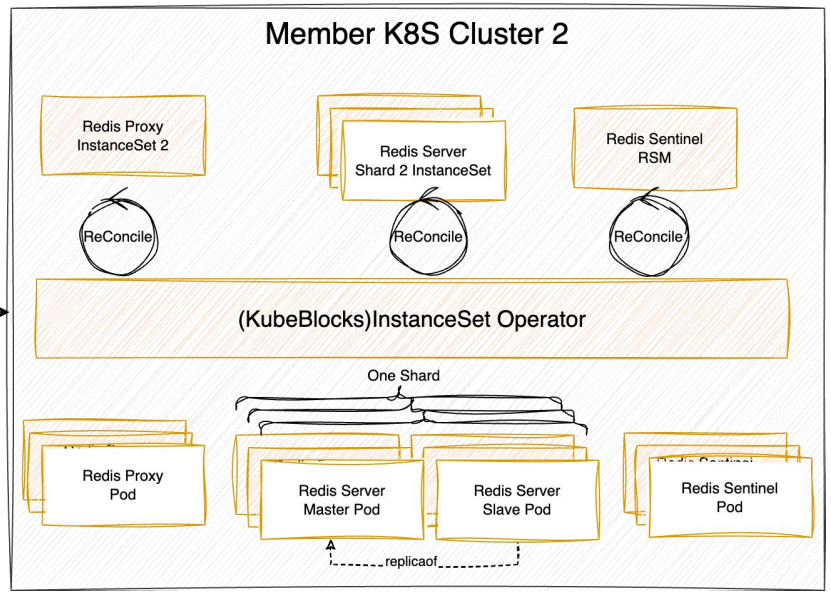
KubeBlocks has implemented it 👍



Architecture

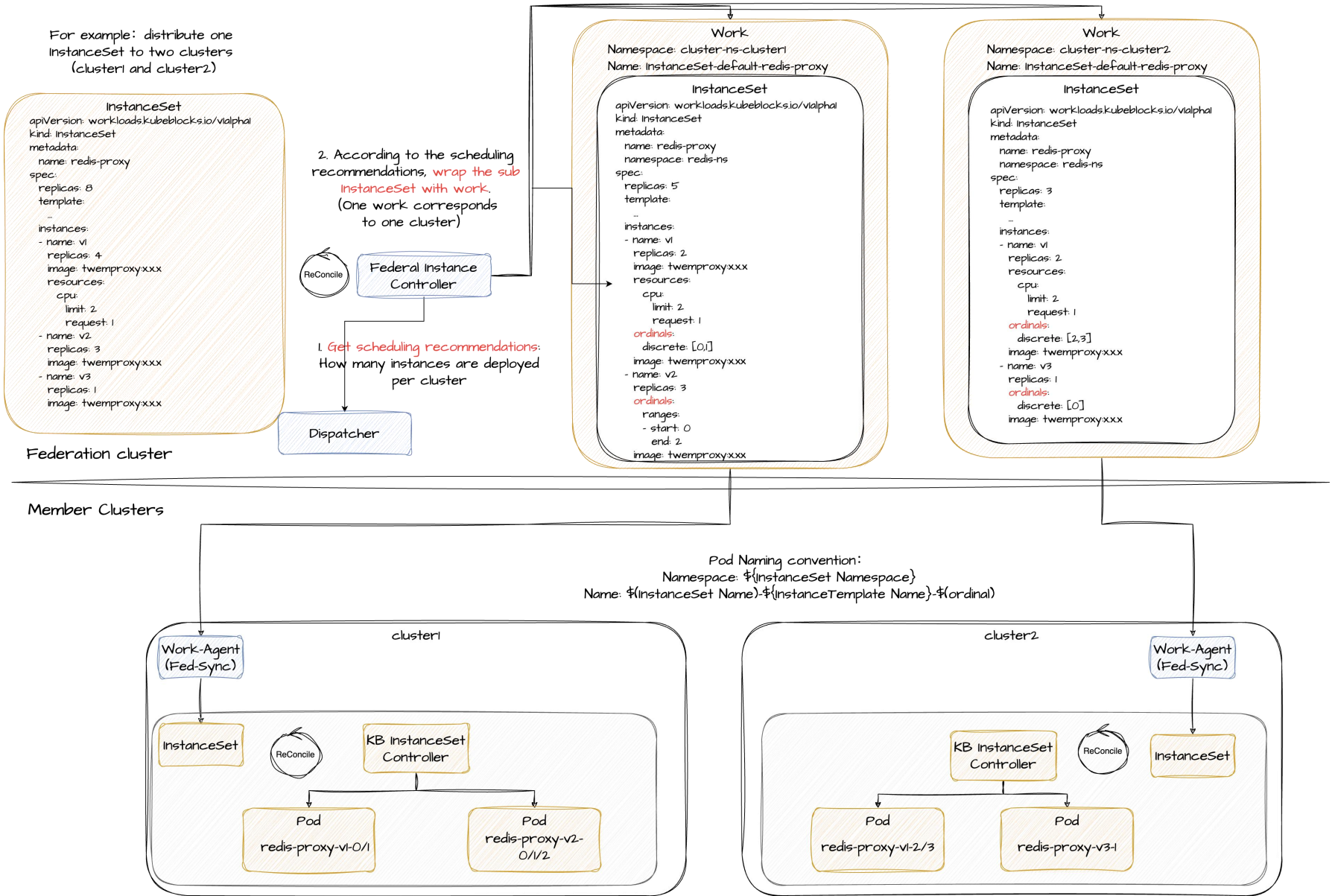




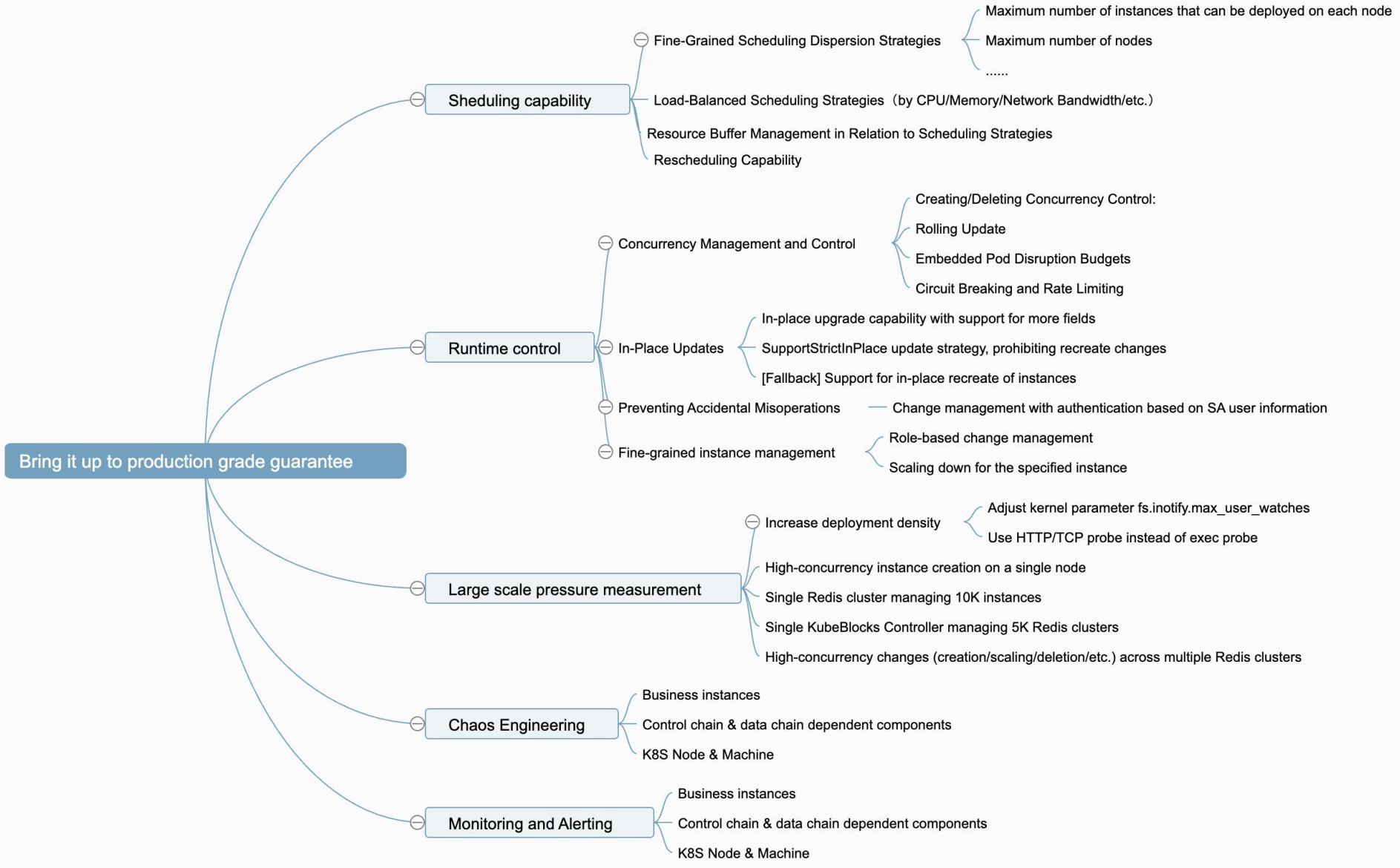


- Cross-Cluster Instance Distribution and Management
- Cross-Cluster Dependency Resource Distribution and Management
- Status Aggregation
- Concurrency Control
- .....

Multi-Cluster Distribution Management



Stability guarantee



Thoughts on Kubeblocks

The control plane for your cloud-native data infrastructure

Install, create, connect, and you have it all.

Collaborate with Kubeblocks

- Managing pods and PVCs directly by InstanceSet
- Instance Template(former heterogeneous pod)
- Integrated with the Federation cluster
- Parallel concurrency policy
- Restrict update policy
- In-place update and in-place vertical scaling
- Specify instances to scale down
- Performance optimization in large-scale/high-concurrency scenarios
- ...

WeChat

Slack

Join the KubeBlocks Community

https://github.com/apecloud/kubeblocks

WeChat
Official Account

Scan「KuaiShou Tech」QR code,
Get more technical articles