
KubeCon
CloudNativeCon
OPEN SOURCE SUMMIT
China 2019

Understanding Scalability and Performance in the Kubernetes Master

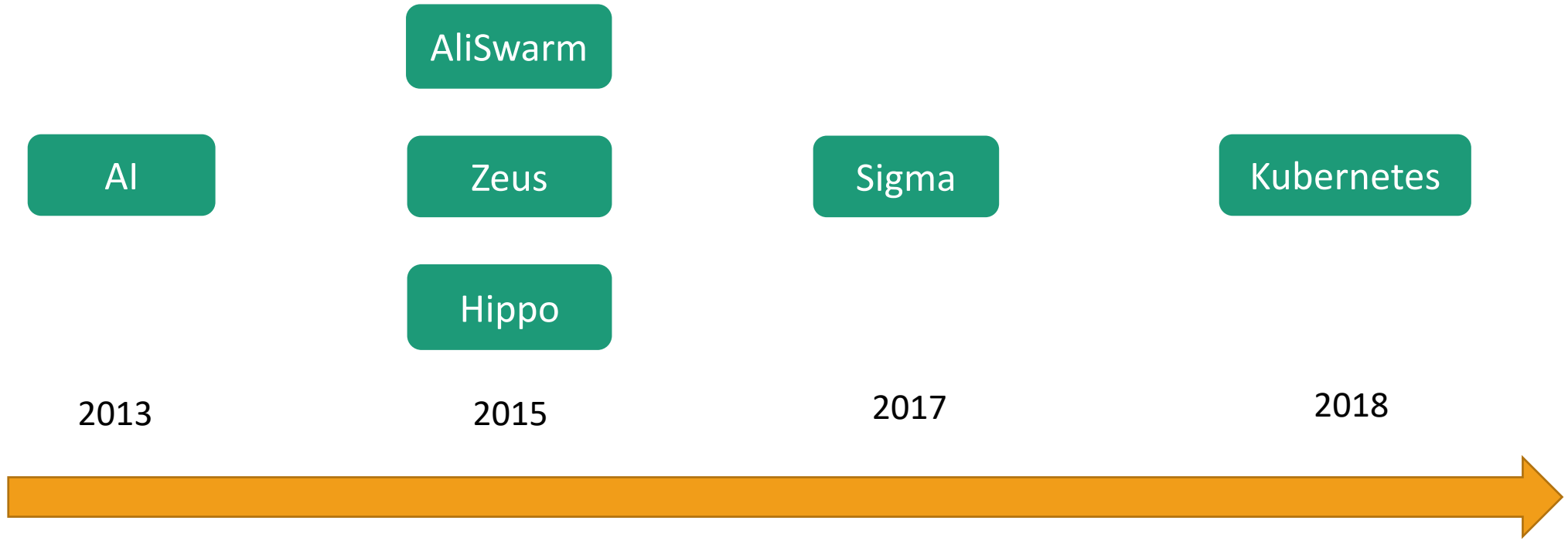
Xingyu Chen, Fansong Zeng Alibaba Cloud

Agenda

- Background
- Kubernetes in Alibaba
- Experience on scalability
  - etcd/apiserver/controller
- Q&A

Background

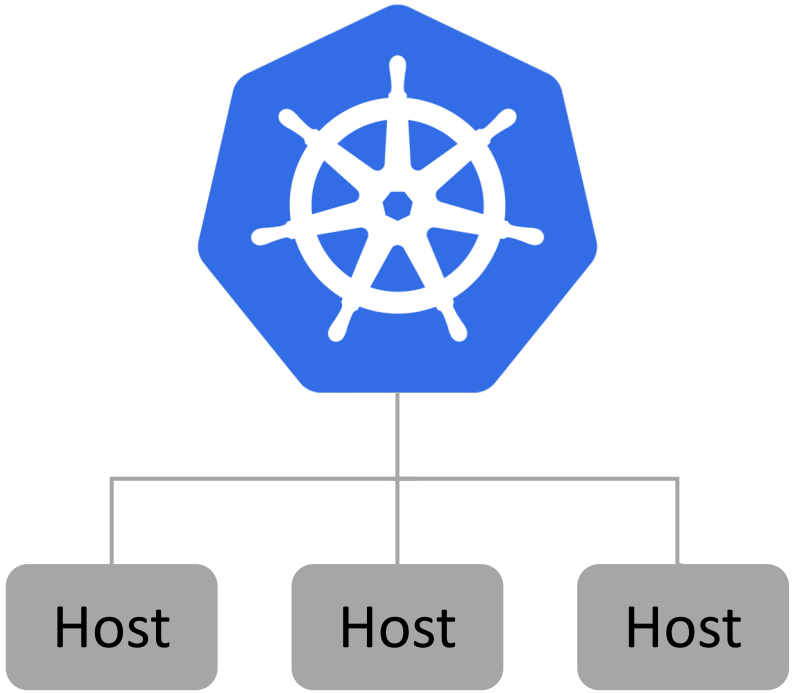
It is a long story…



Kubernetes in Alibaba

- Production environment
  - 10, 000s of applications
  - 1, 000, 000s of containers

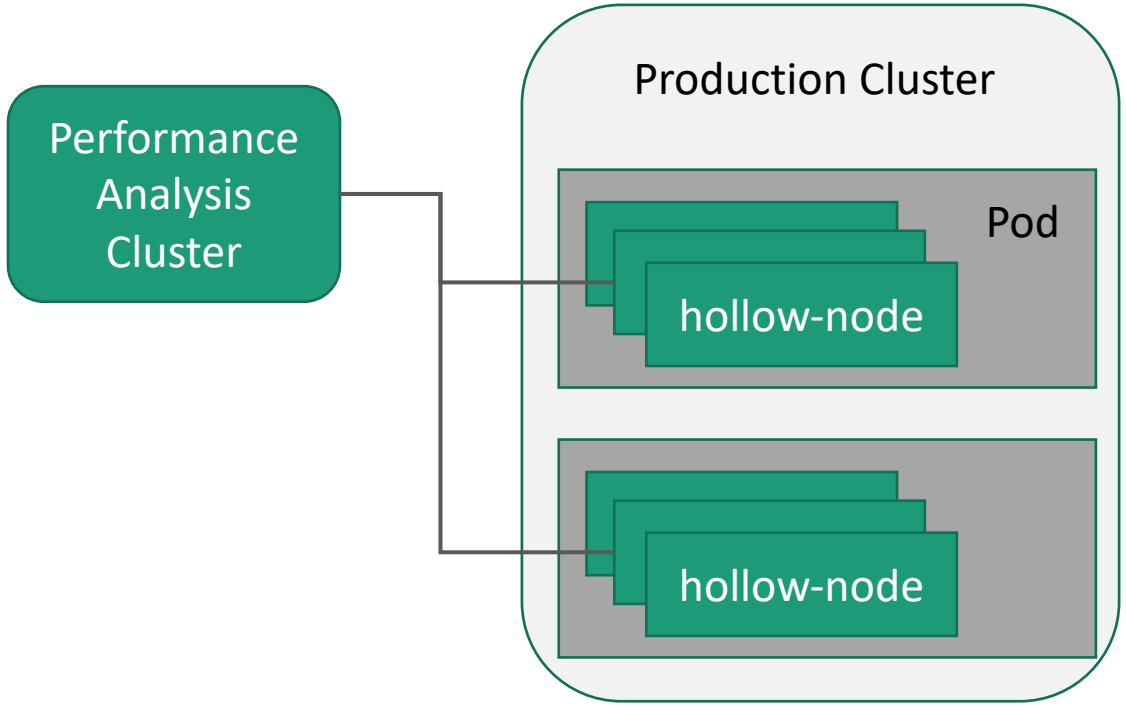
  - 10s of clusters
  - 100, 000s of nodes
  - 10, 000 nodes / largest cluster



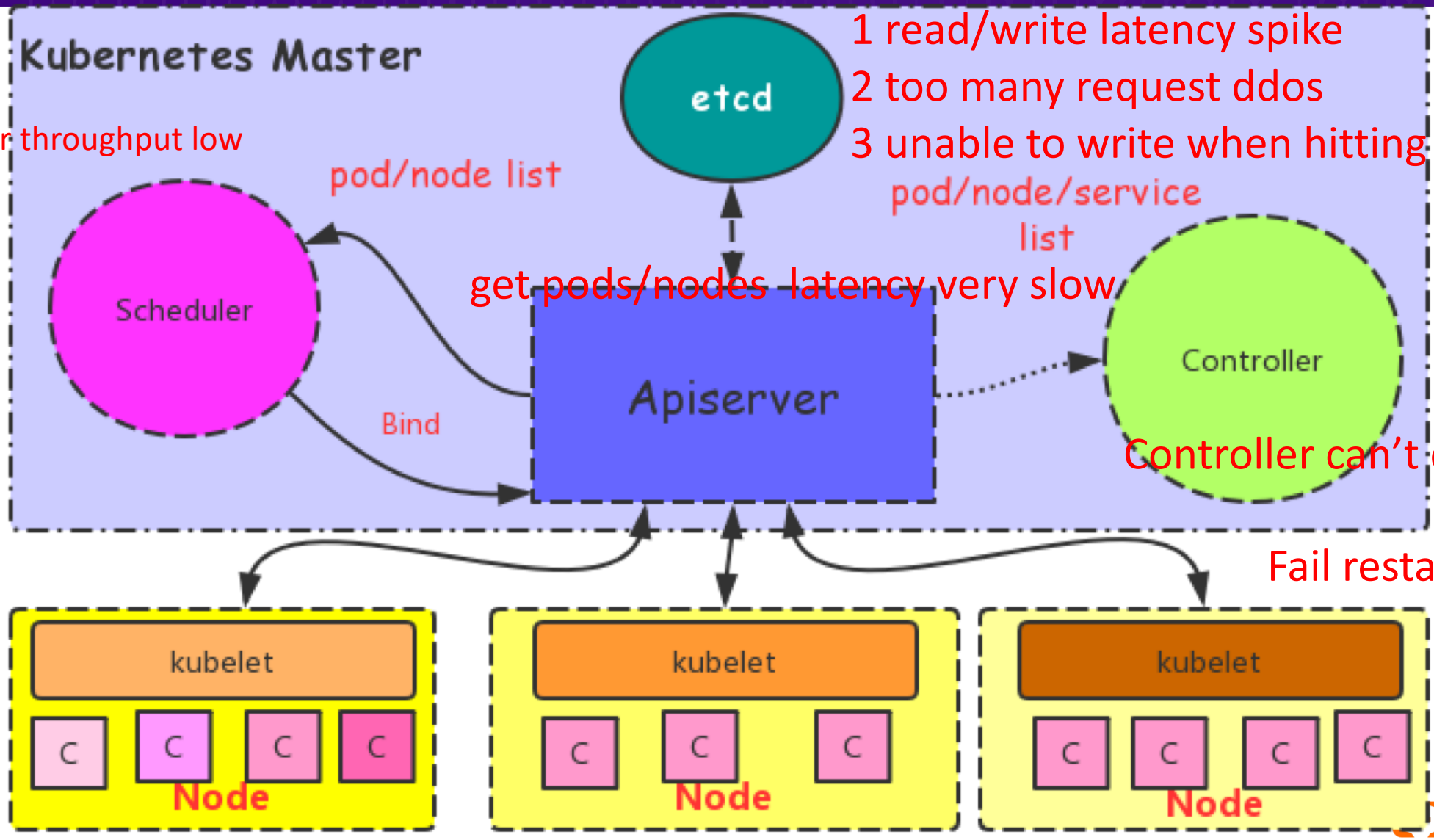
Experience on scalability

A 10k nodes cluster

# Pods	~200k
# Objects	~1000k
# Latency	~10s



Experience on scalability



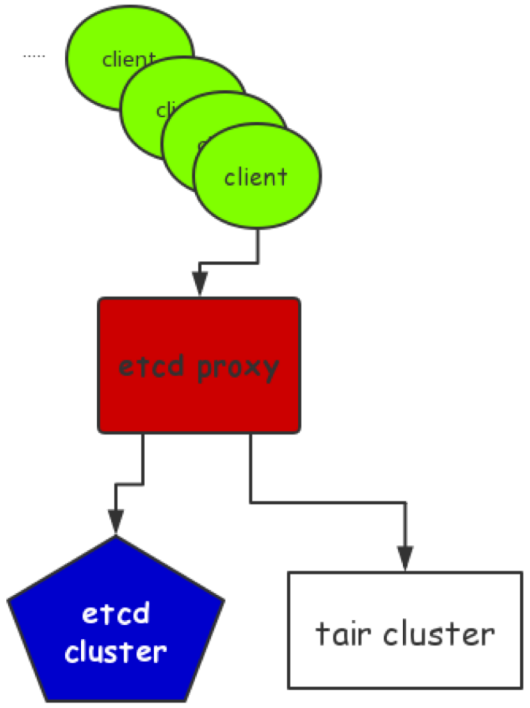
1 read/write latency spike
2 too many request ddos
3 unable to write when hitting storage limit

Scheduler throughput low

get pods/nodes latency very slow

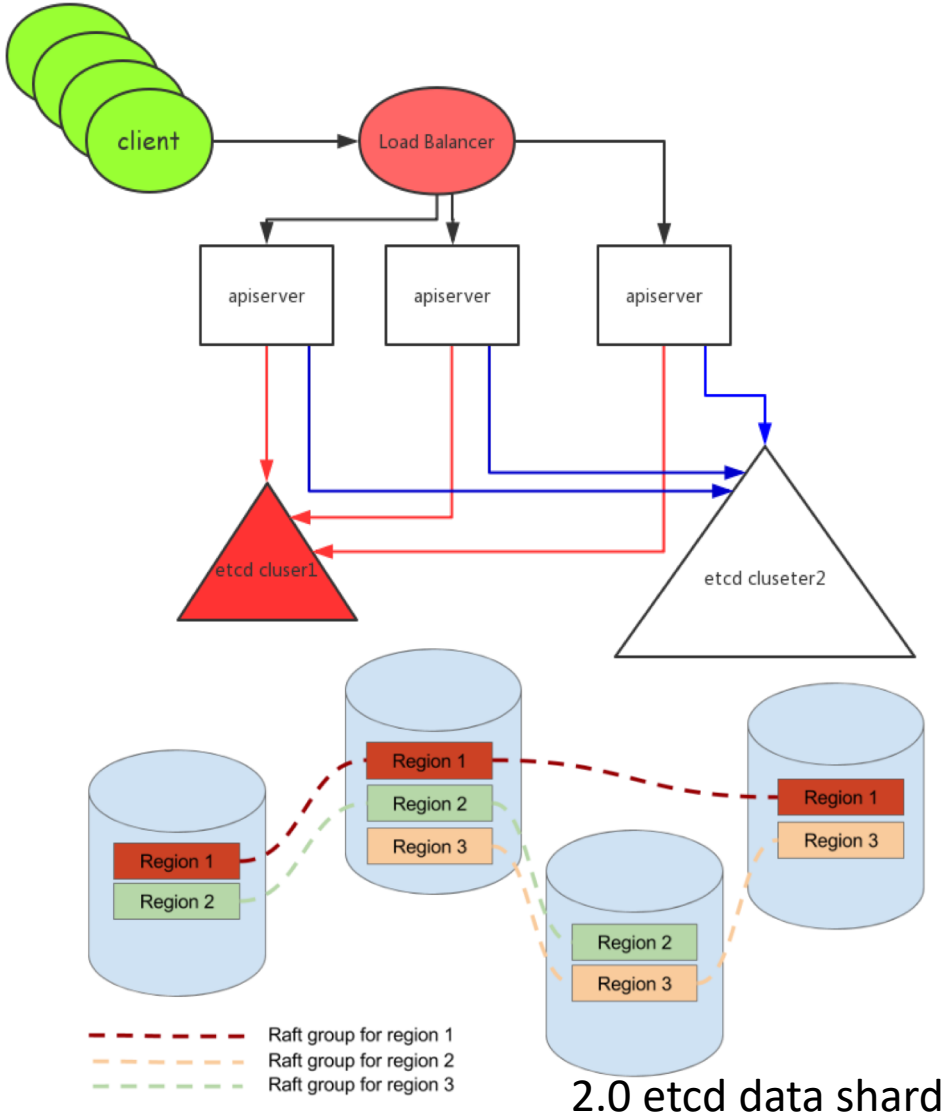
Controller can't catch up

Fail restart slowly

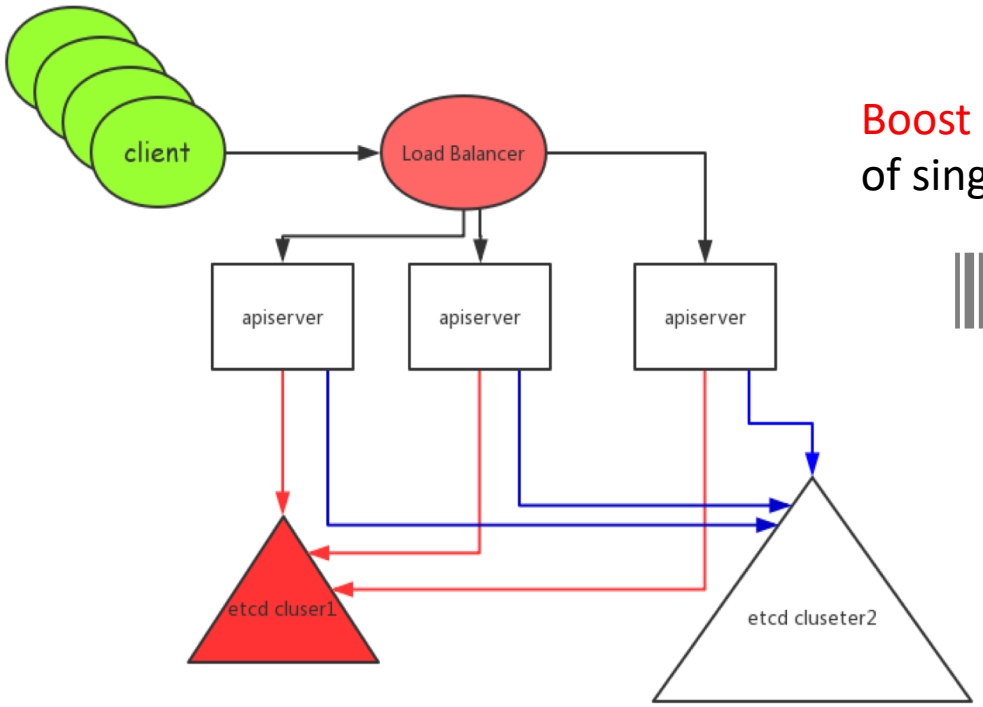


1.0 etcd over tair

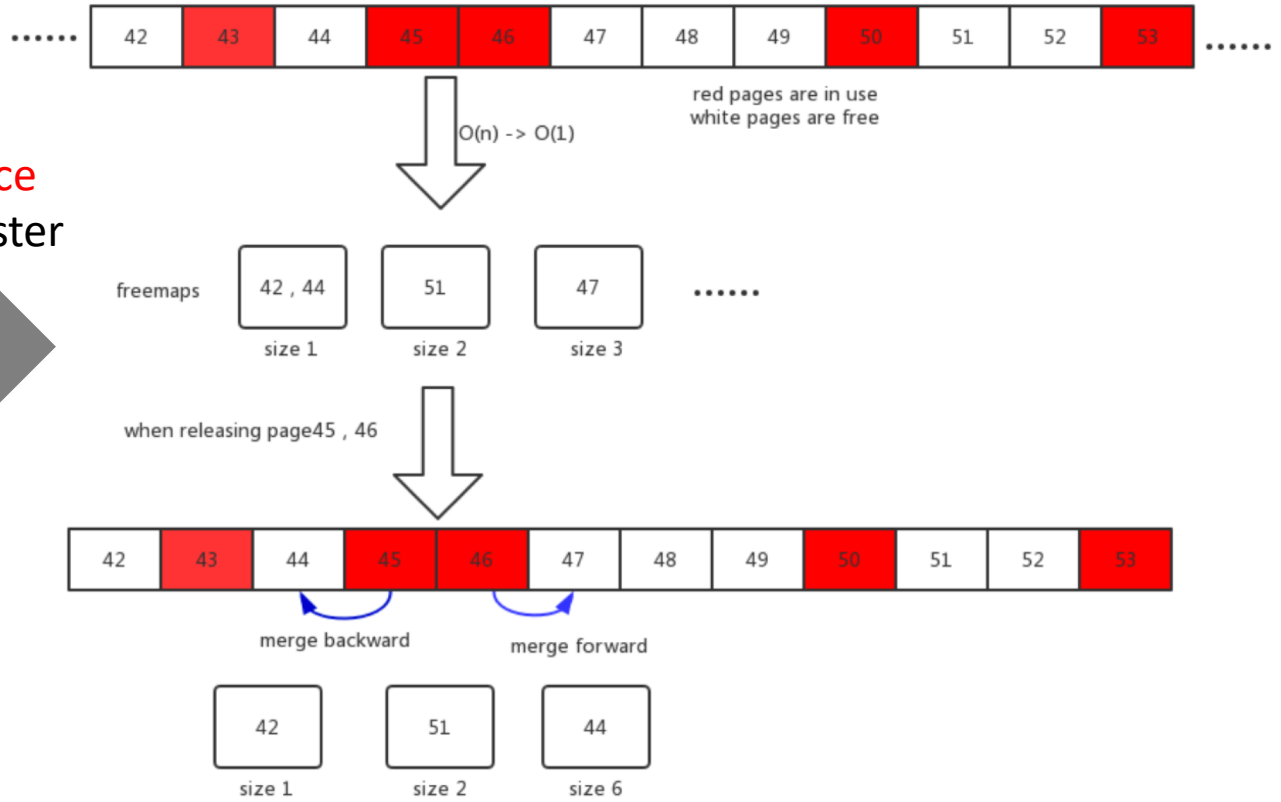
- Boost storage limit
- Level up data isolation



2.0 etcd data sharding



Boost performance
of single etcd cluster

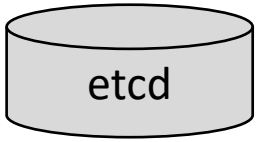


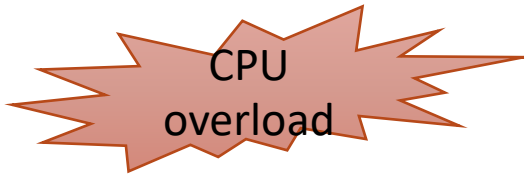
Segregated hashmap algorithm

Efficient node heartbeats

- Kubelet send hearbeats every 10s
- 15kb with tens of images and volumes

transaction log **~1GB/min**

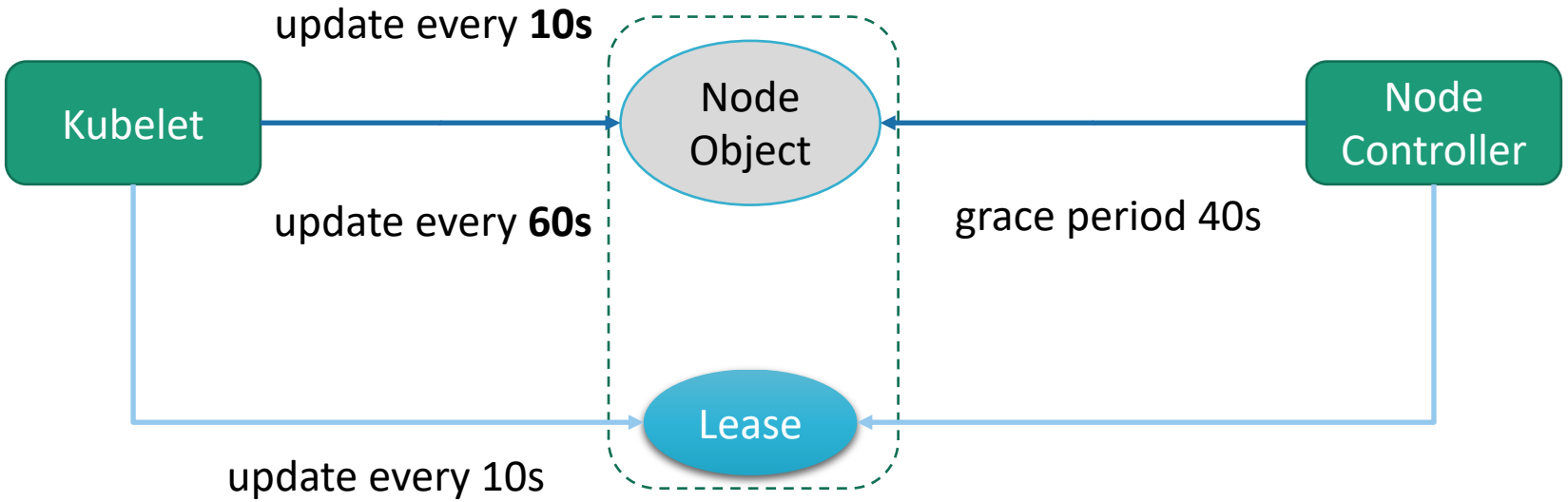




larger than **80%**

Efficient node heartbeats

- Add a new `Lease` build-in API

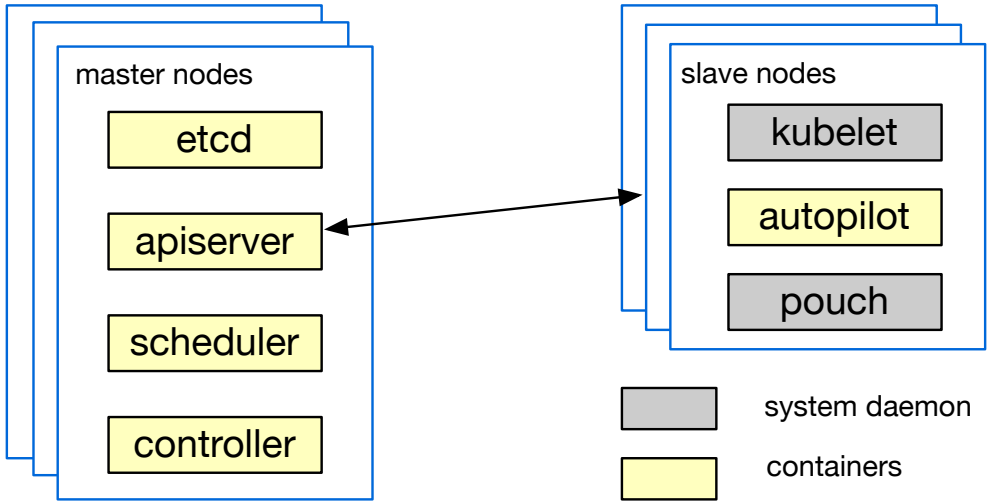


Lease objects ~100 bytes

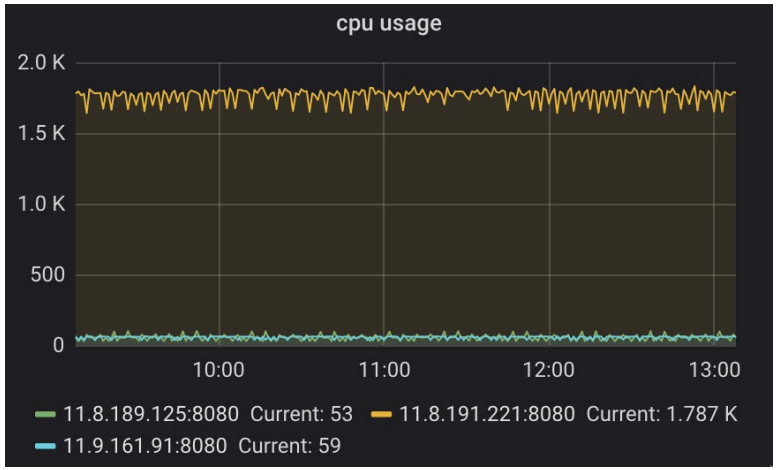
Enabled @ 1.14

API Server load balancing

- 3 nodes HA cluster

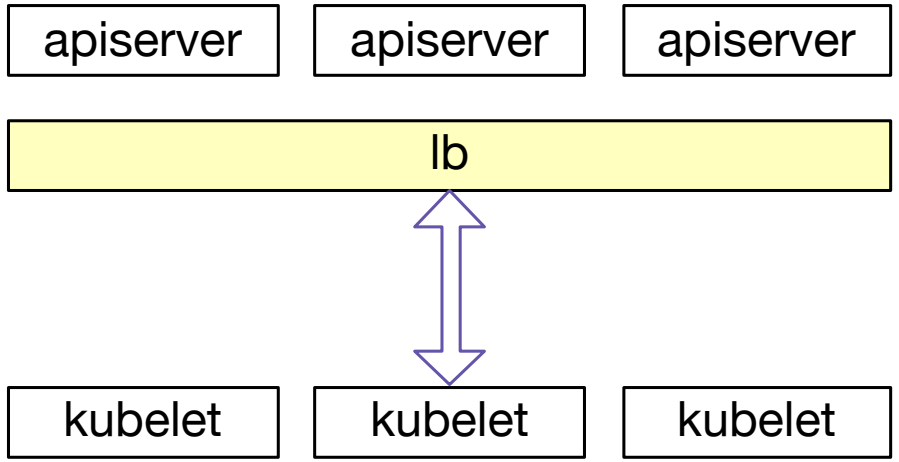


upgrades
failures

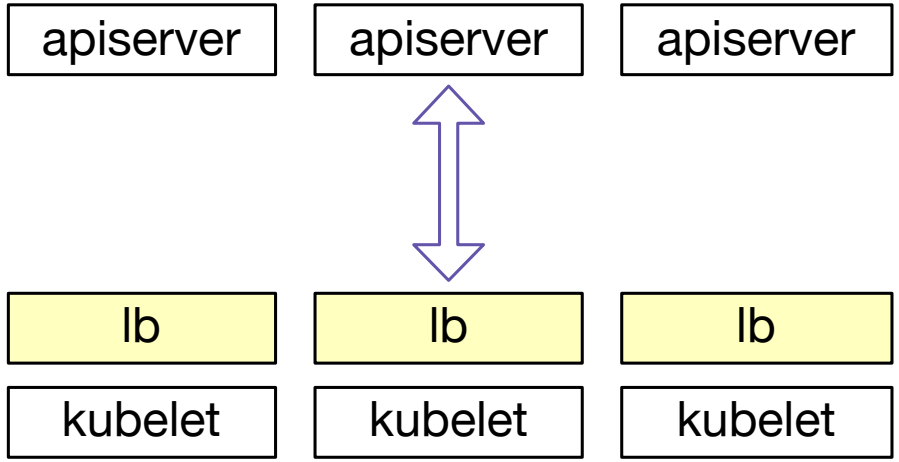


API Server load balancing

- Add a load balancer ?



(1)



(2)

API Server load balancing

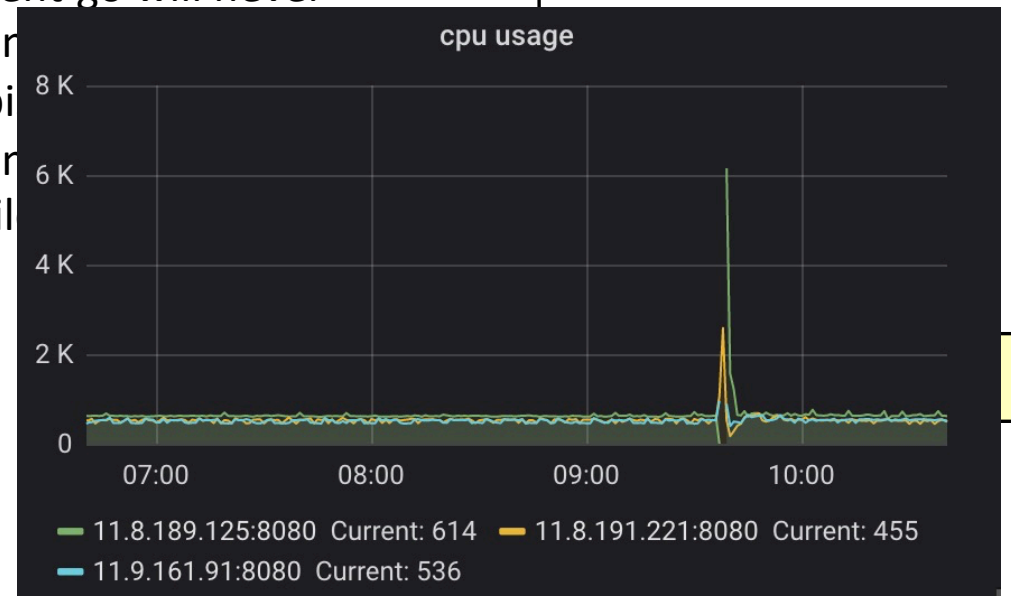
- t… nec…

Upgrade with
{maxSurge=3}

Throttling:
1. send ‘too many requests ’ when exceeded the low
2. send ‘Connection: close’ when exceeded the high

apiserver

client-go will never
cor…
api…
cor…
fail…



apiserver

watch server

persistent
connection

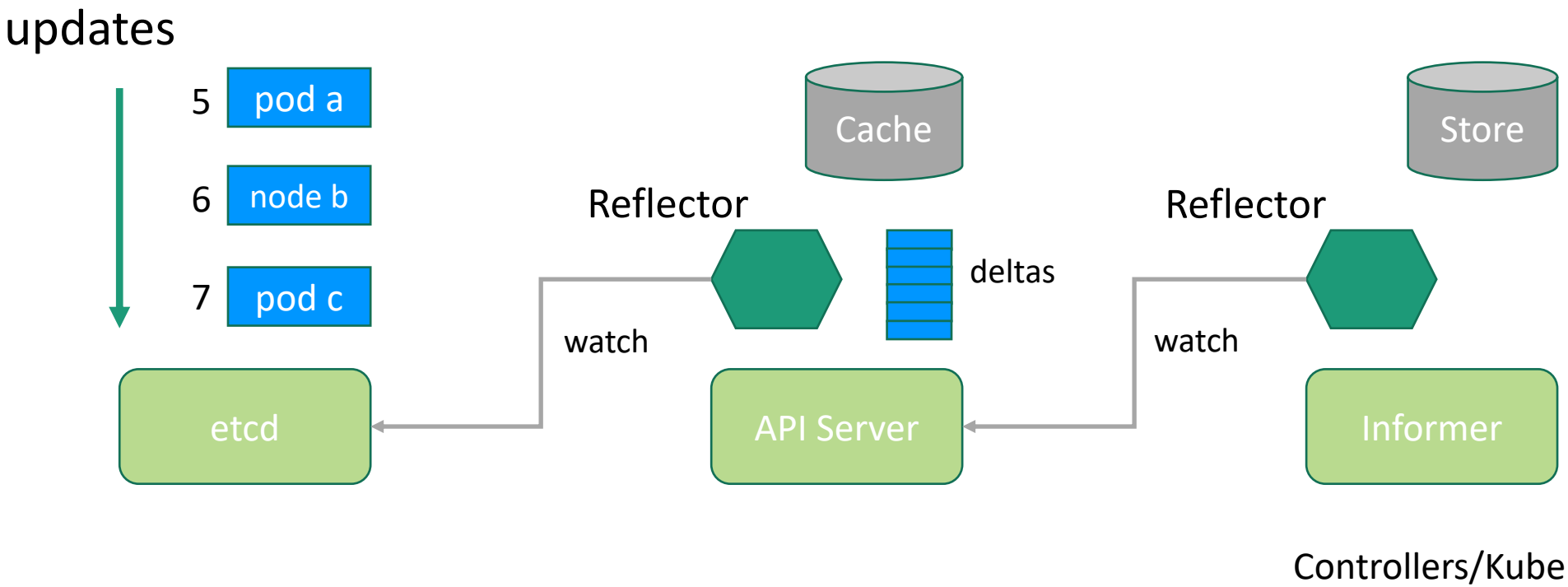
root cause

watch channel

Retry:
1. try another server if there are too many ‘too many requests ’
2. try another server after several minutes

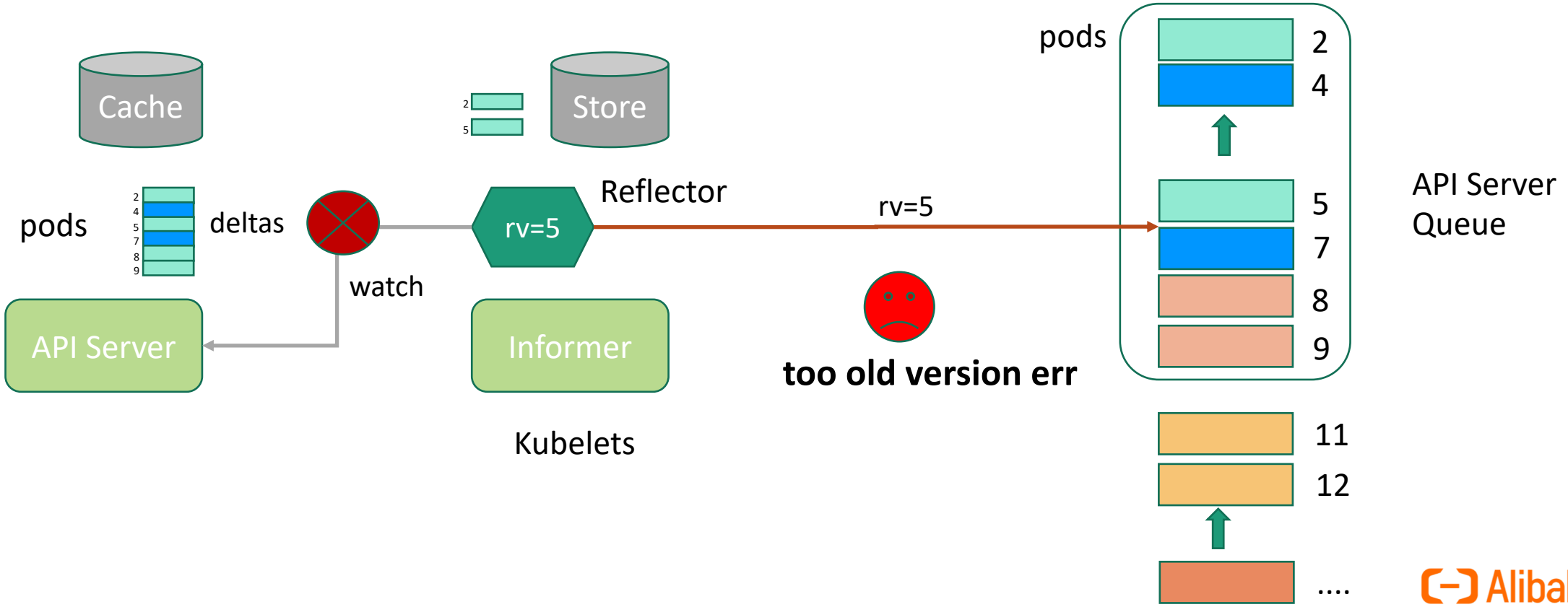
List-Watch & Cacher

- Key communication mechanisms between client and server



List-Watch & Cacher

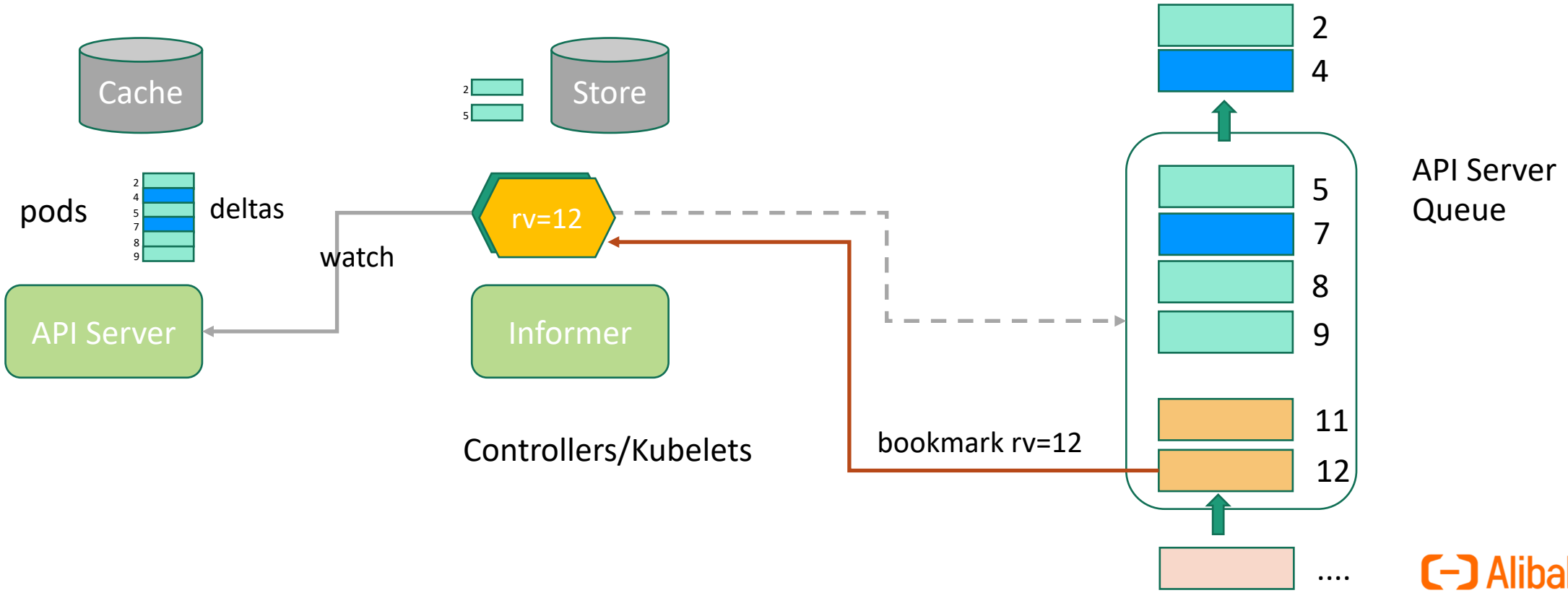
- What happens if the connection is broken ?



List-Watch & Cacher

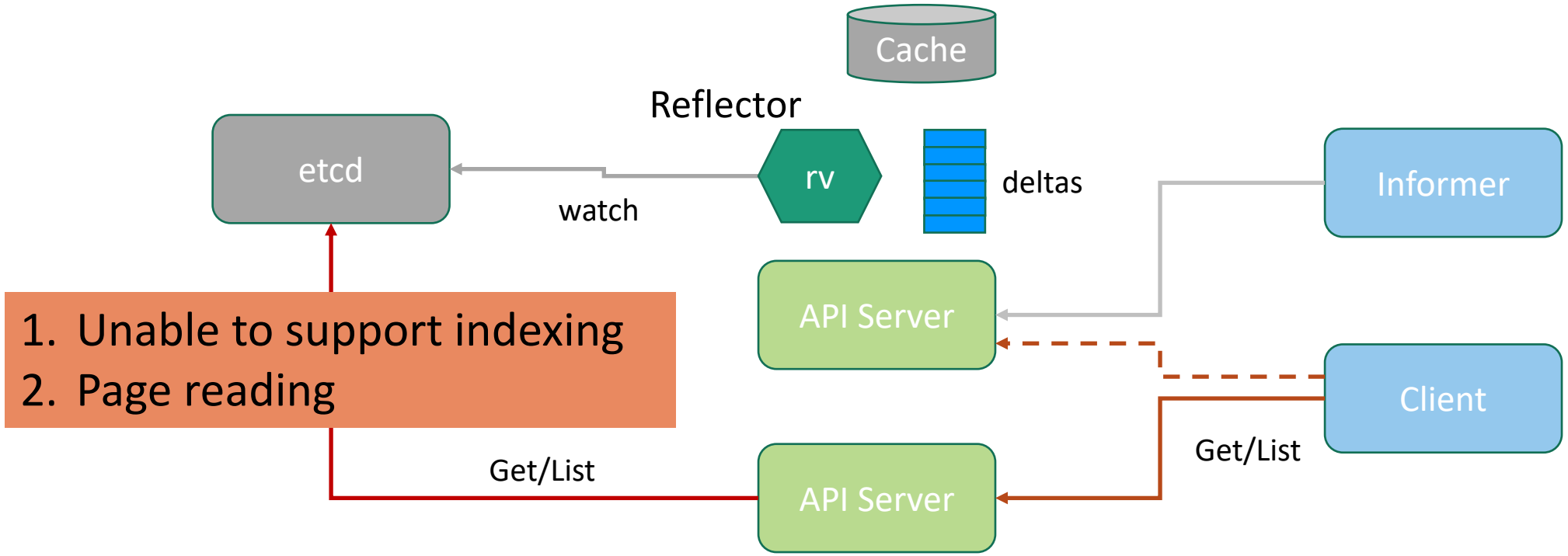
- Watch `bookmark`

Reduce to 3% released at 1.15



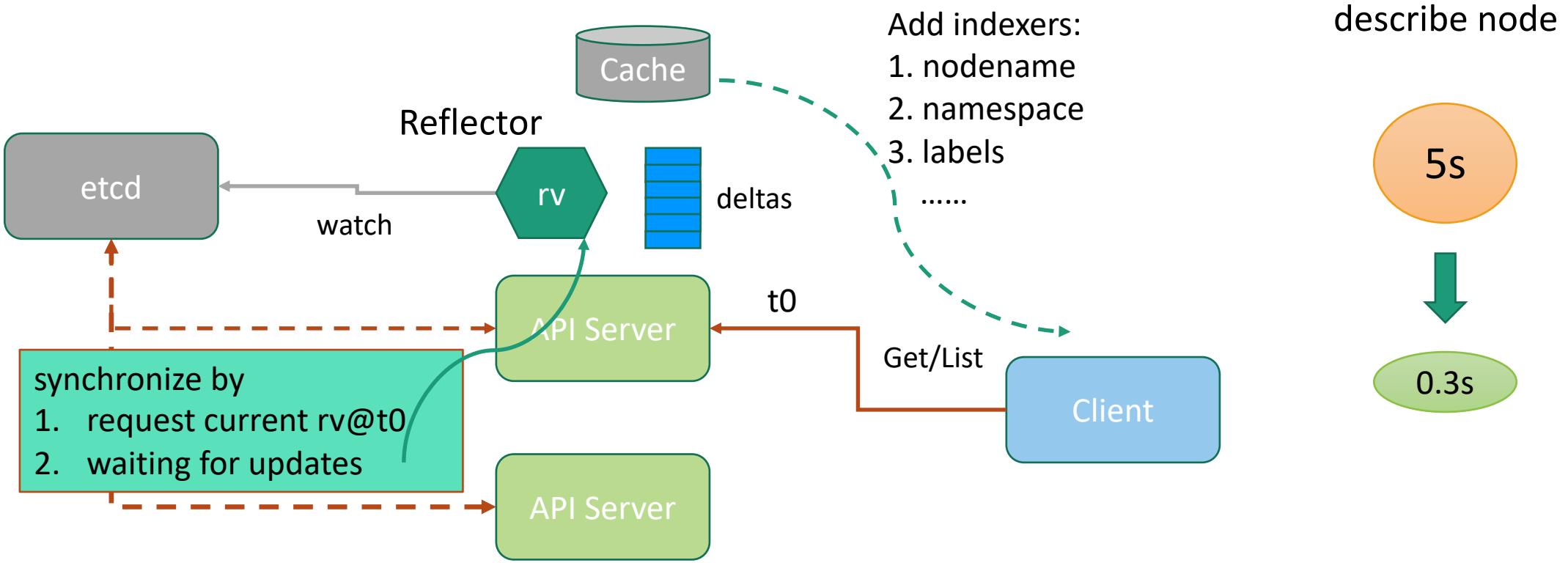
List-Watch & Cacher

- Improve cache to serve read requests



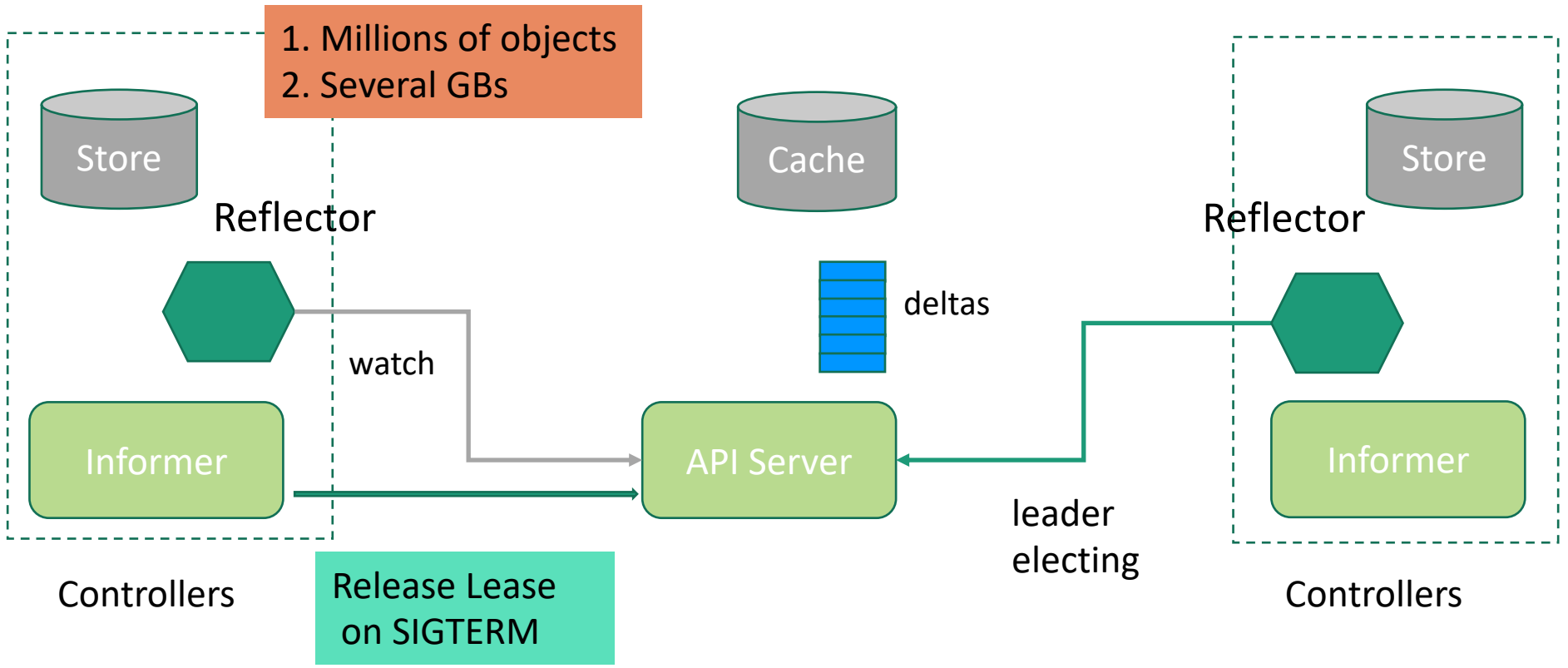
List-Watch & Cacher

- Improve cache to serve read requests



Controller failover

- Restarting controllers takes several minutes ⬅ MTTR ➡ Reduce to several seconds



Q&A

KubeCon CloudNativeCon OPEN SOURCE SUMMIT China 2019

THANK YOU

免登陆听课

动手实践课后自测

CKA课程内容同步

阿里云原生最佳实践

立即听课

Alibaba Cloud

Customized scheduler

- Equivalence classes

- Relaxed randomization