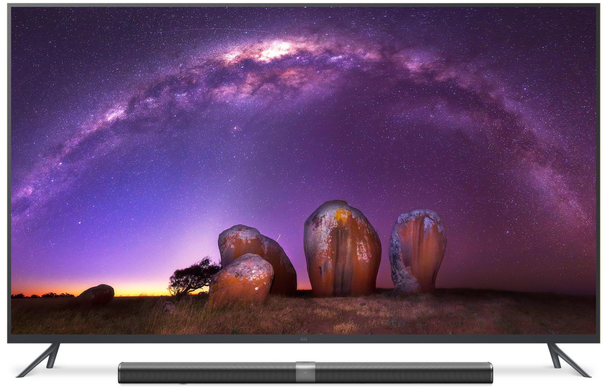
hosted by Alibaba Group 阿里巴巴集团 APACHE HBASE

HBase at Xiaomi

Guanghao Zhang
zghao@apache.org

About Xiaomi

- Founded in 2010
- Worldwide smartphone No.4 shipments, Q2 2018
- 300+ million global users
- Products: smart phone, TV, AI speaker, smart band…



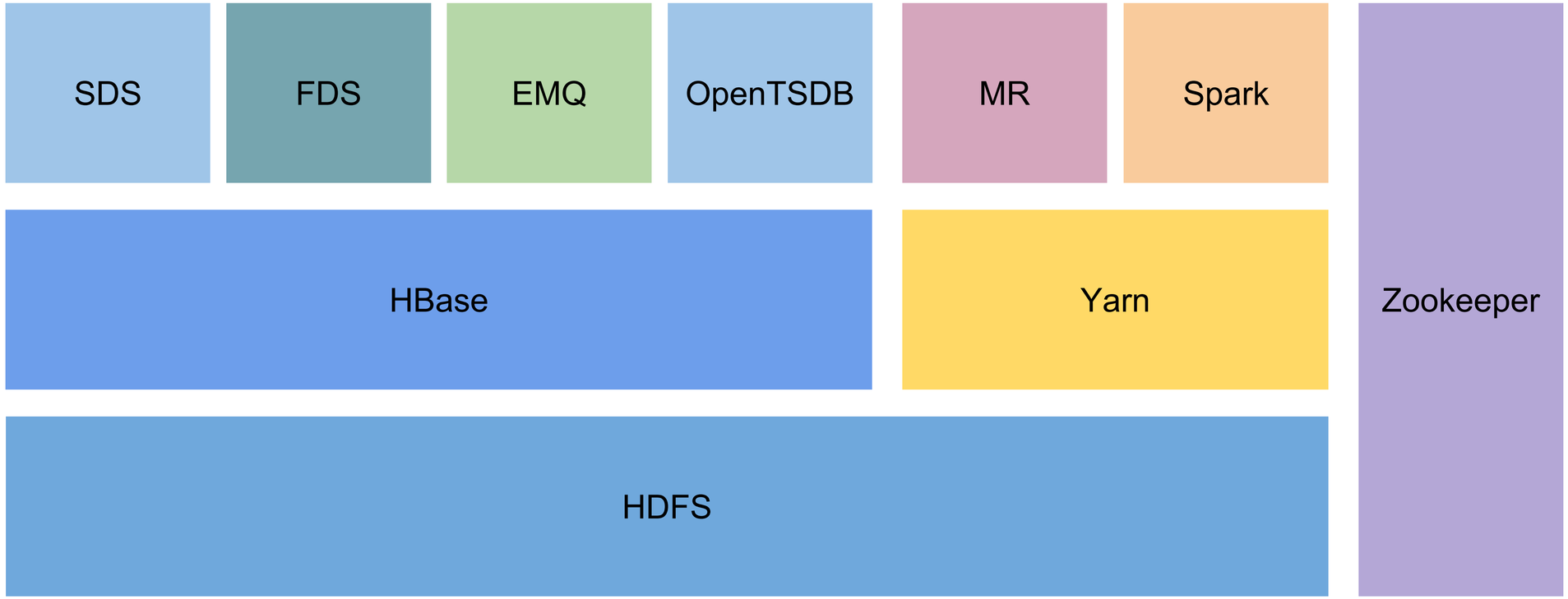
Our HBase Team

- 2 PMC members
- 1 Committer
- 3 HBase Contributers

Content

01 Xiaomi HBase

Architecture



Clusters

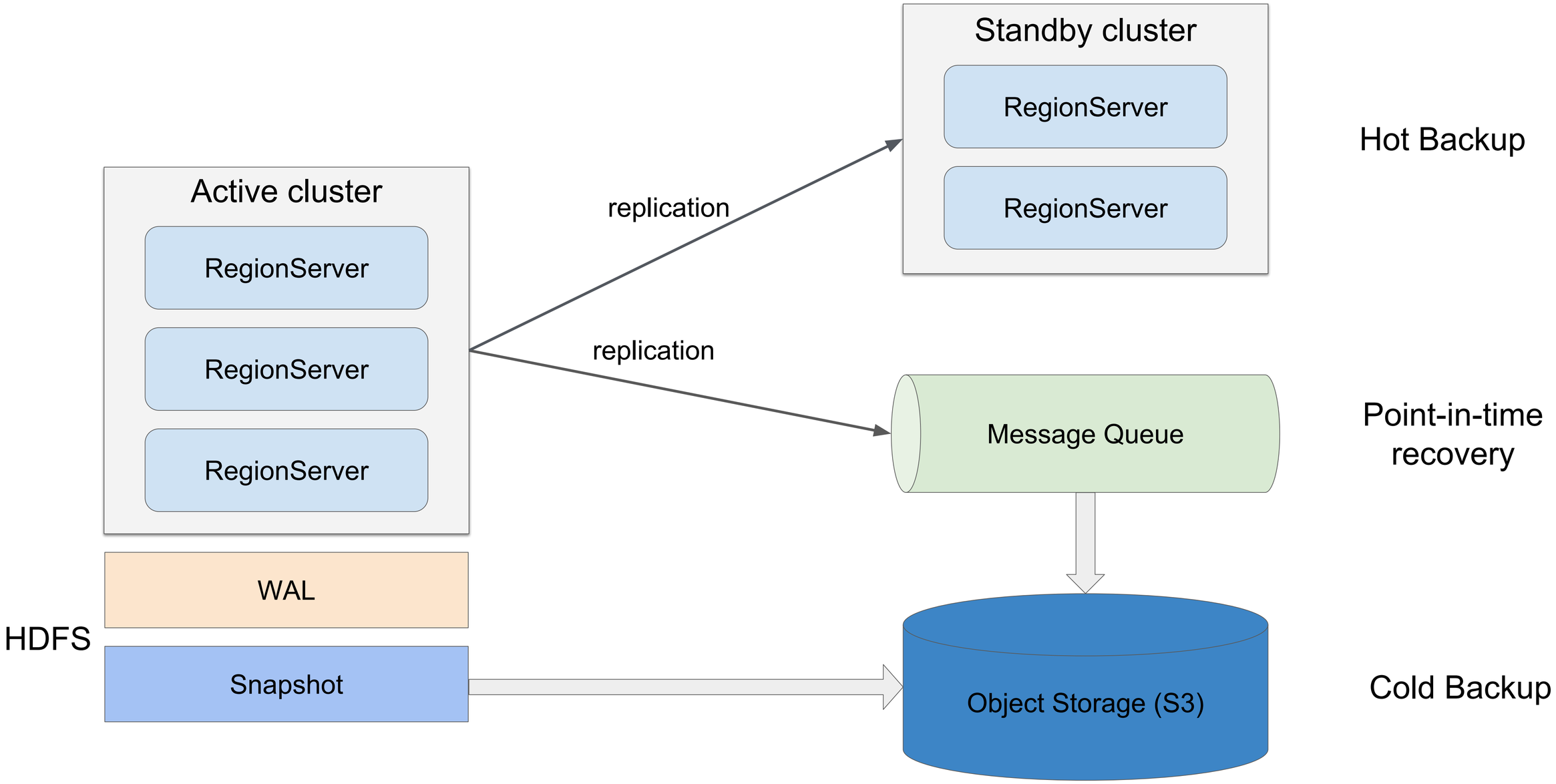
IDC

- 5 data centers (China), 30+ online clusters / 3 offline clusters

AWS / Alibaba Cloud / KS Cloud / Azure

- 16 clusters (China/Singapore/US/Europe/India/Russia)

Backup


Standby cluster
RegionServer
RegionServer
Hot Backup
Active cluster
RegionServer
RegionServer
RegionServer
replication
replication
Message Queue
Point-in-time recovery
WAL
HDFS
Snapshot
Object Storage (S3)
Cold Backup

Multi-tenancy

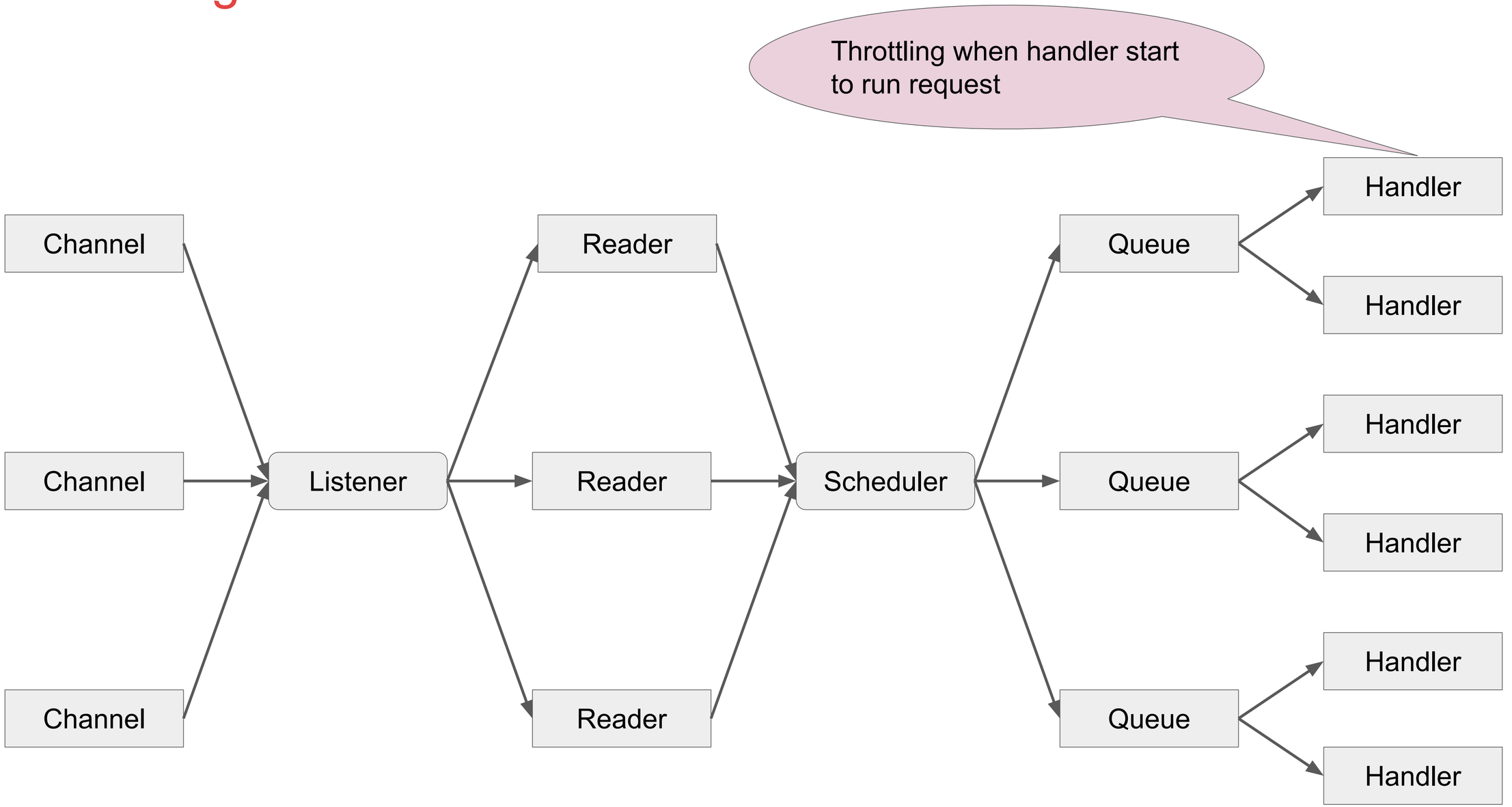
Physical isolation

- Independent HDFS cluster for HBase
- Independent HBase cluster for important business
- Online serving cluster: SSD vs HDD
- Offline processing cluster

Quota and throttling

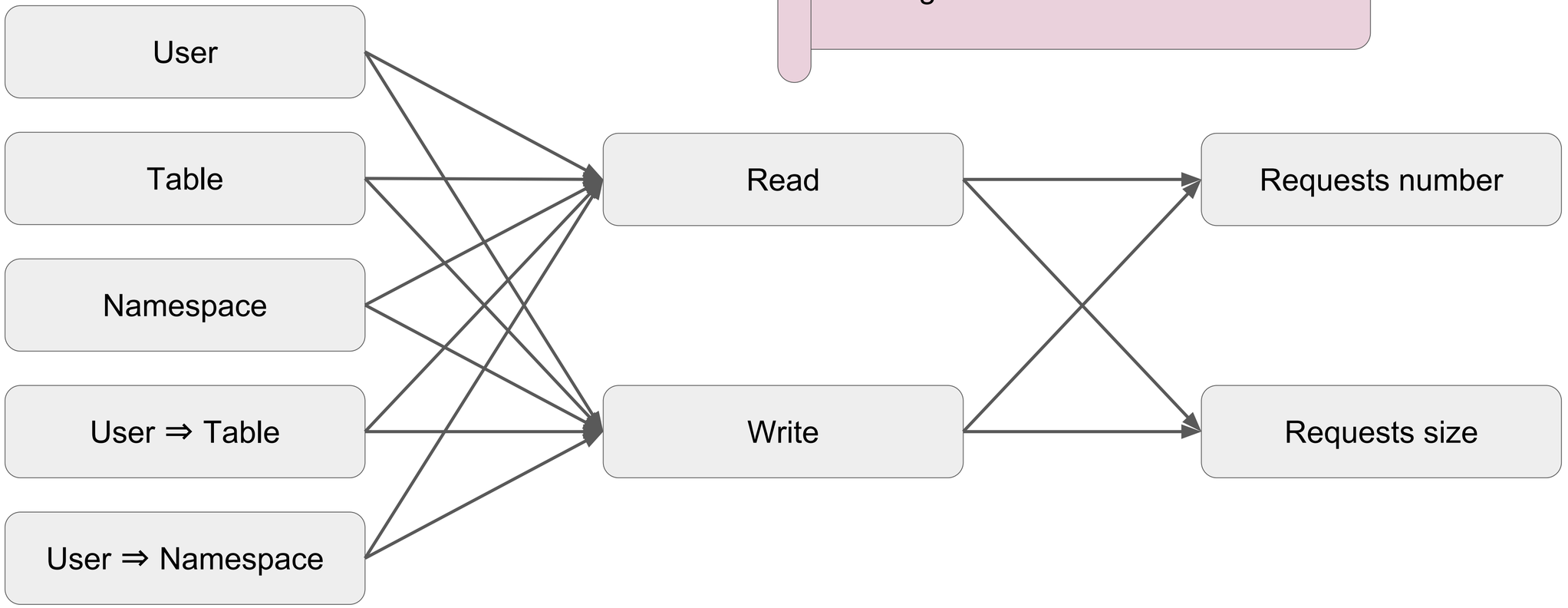
02 Quota and throttling

Rpc Throttling



Quota Type

1. Hard limit
2. If any one quota exceed, then throttling



More Quota Types at Xiaomi

RegionServer Quota: hard limit

Request Unit: calculate both number and size

- Read capacity unit: 1KB/sec
- Write capacity unit: 1KB/sec

Soft limit: allow to exceed user’s quota when regionserver quota not exceed

Other Improvements

Switch to start/stop throttling

Metrics and UI support

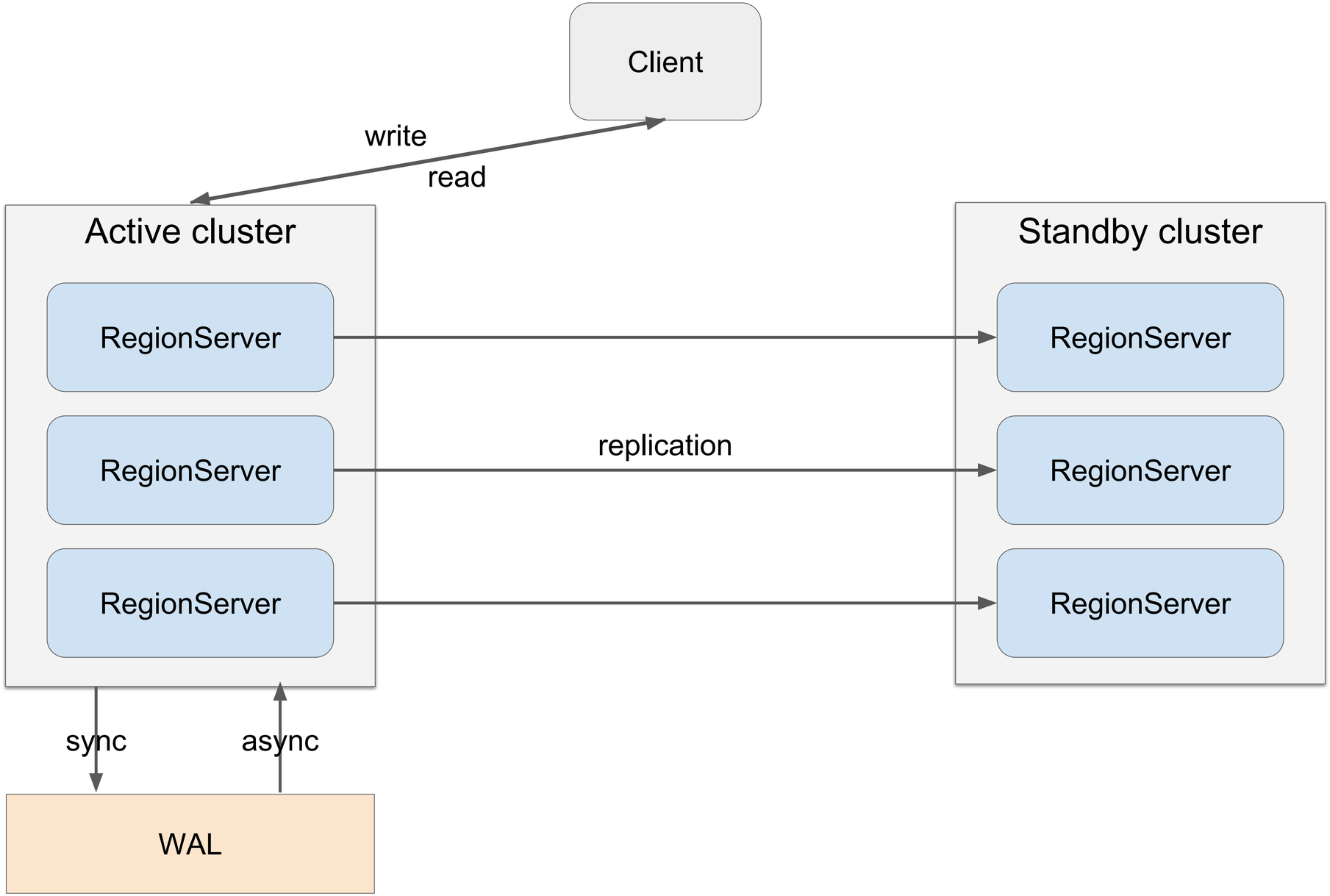
Handle ThrottlingException in client

- DoNotRetryNowException
- Avoid MR/Spark job failed by throttling

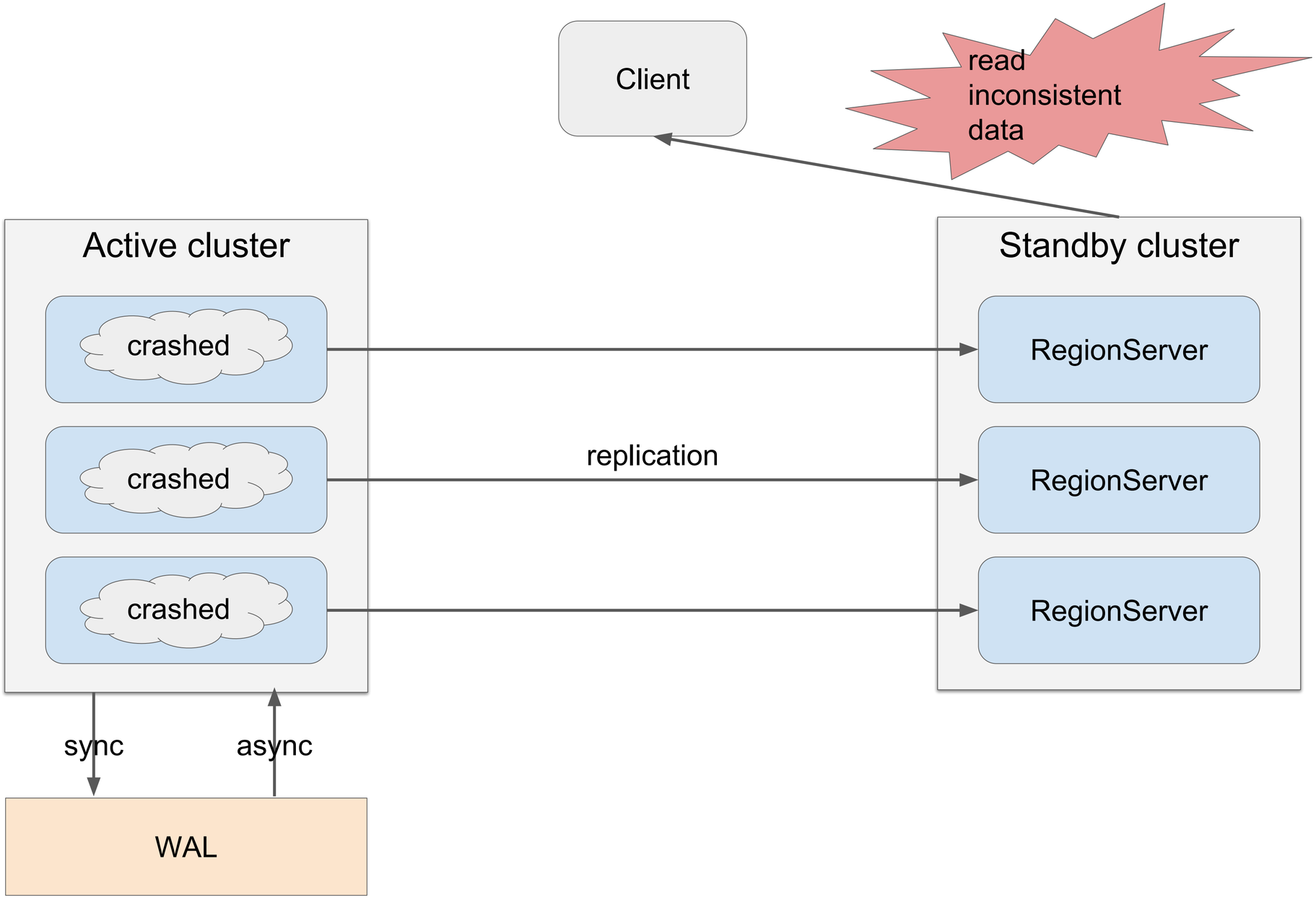
Punishment mechanism for huge requests

03 Synchronous Replication

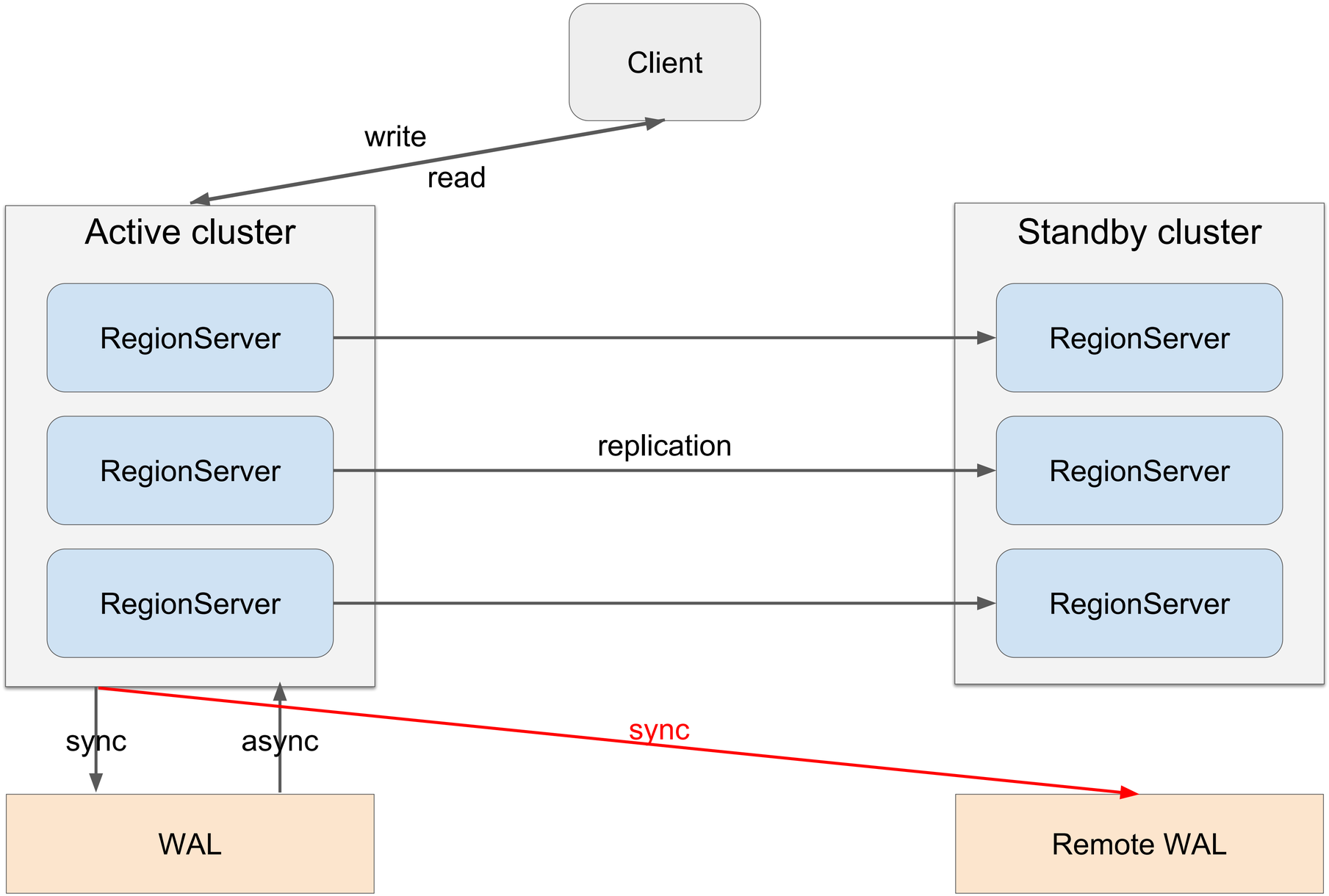
Async Replication



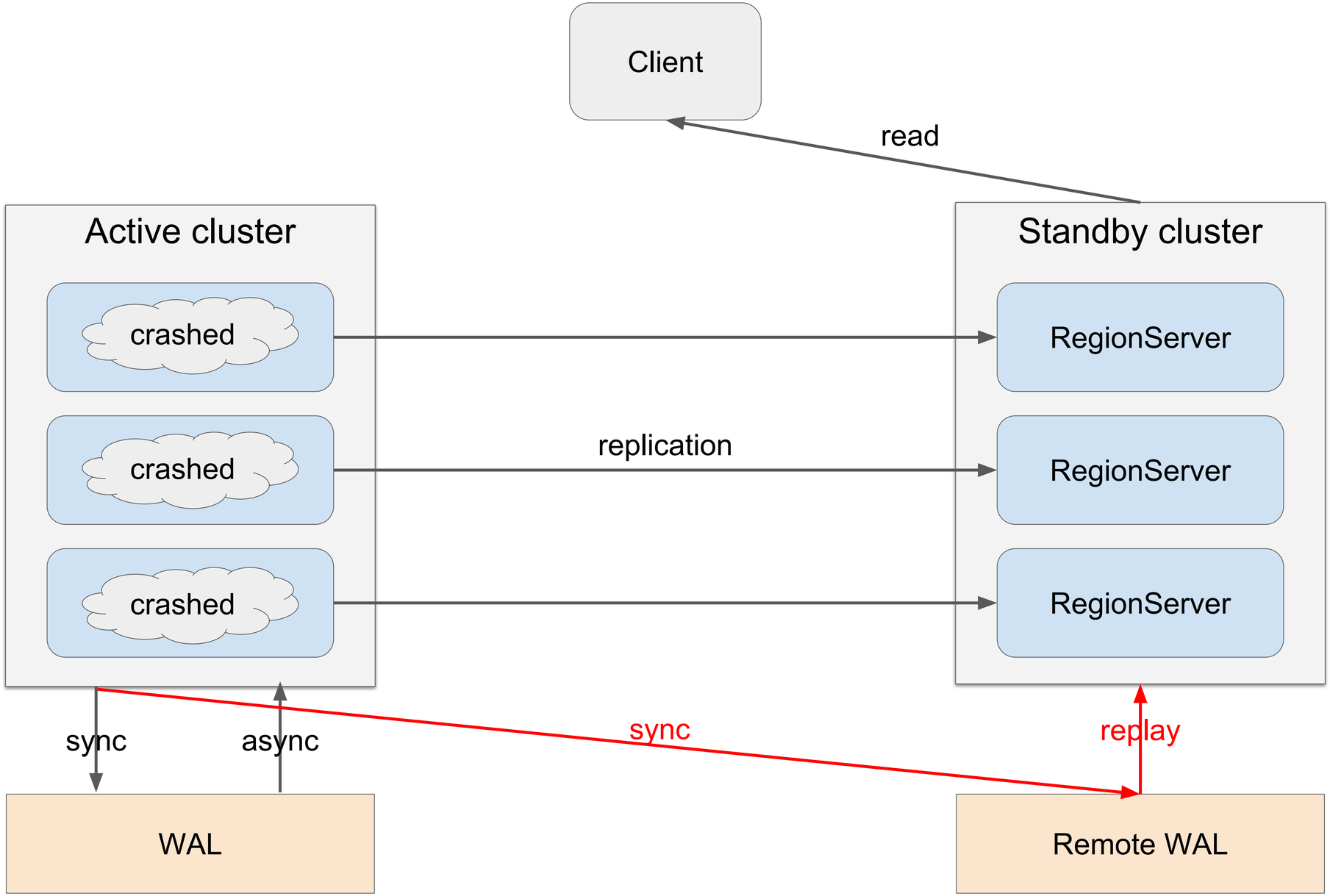
Async Replication



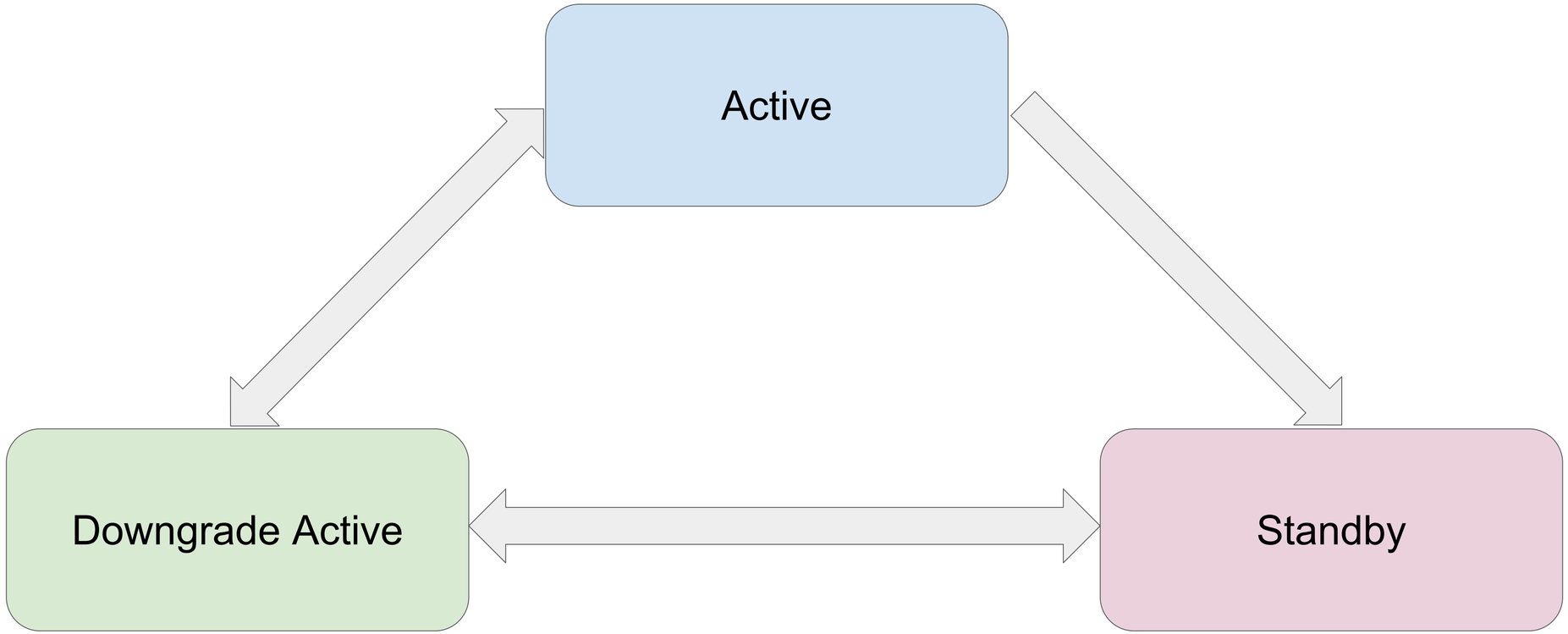
Sync Replication



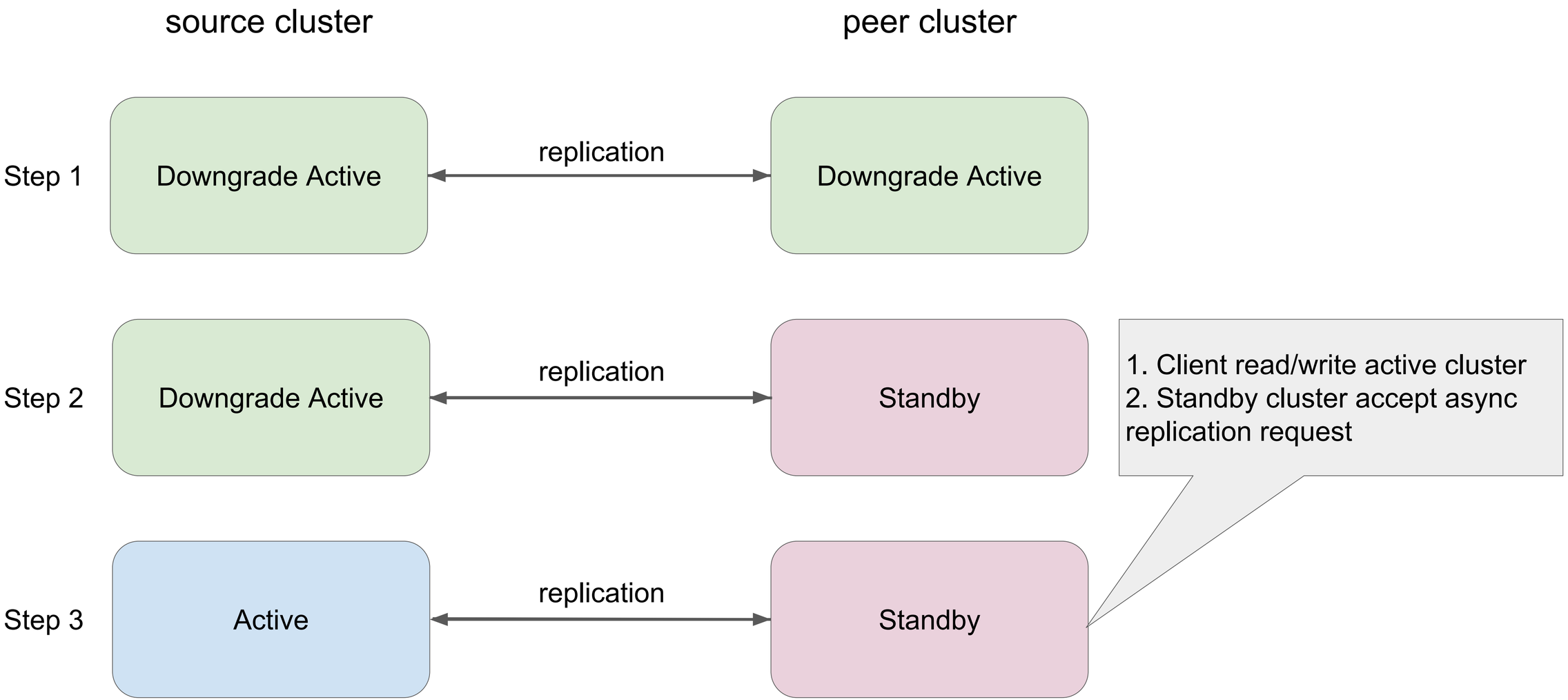
Sync Replication



Sync Replication State



Setup Sync Replication



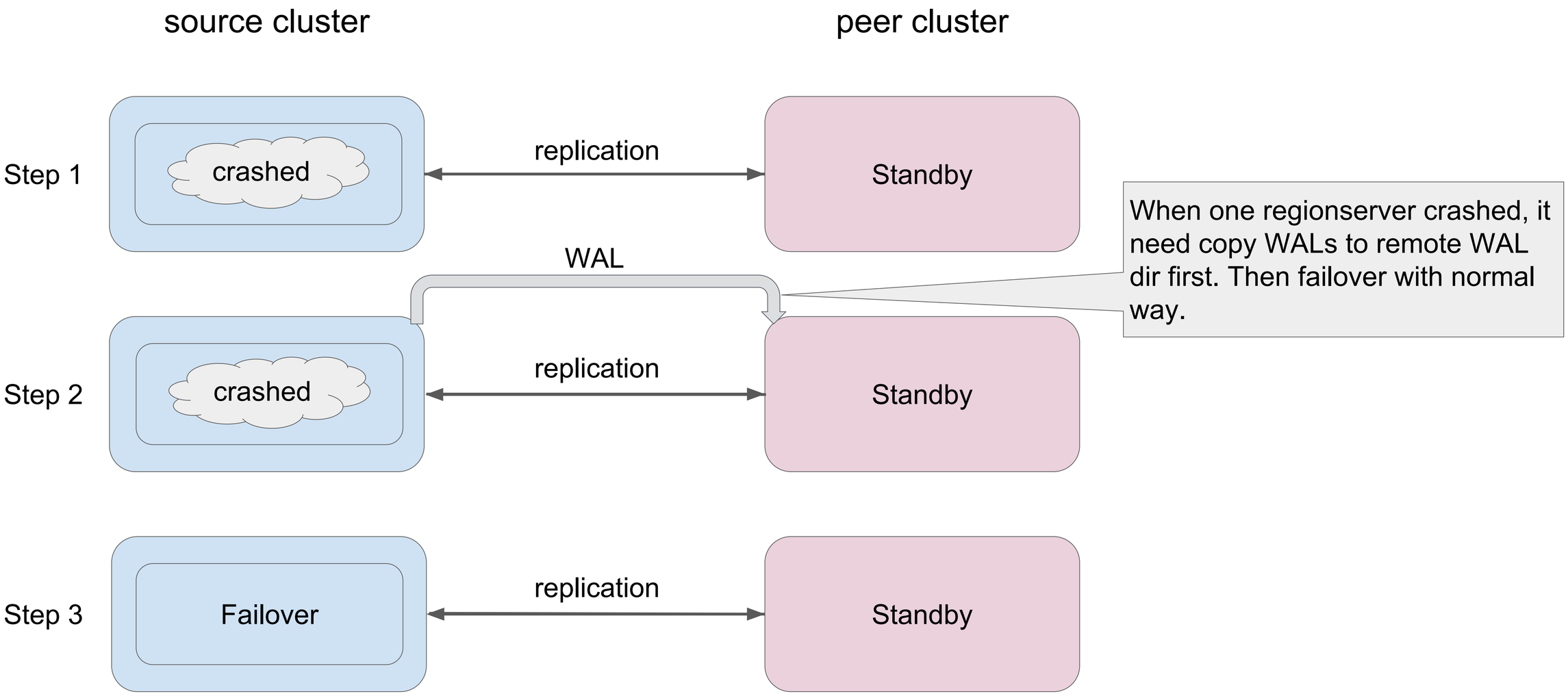
DualAsyncFSWAL

Add a remoteWALDir to ReplicationPeerConfig

Concurrent write WAL and remote WAL

Write WAL	Write remote WAL	Client	
Success	Success	Success	Same data
Success	Fail/Timeout	Timeout	Source cluster may has more data
Fail/Timeout	Success	Timeout	Peer cluster may has more data
Fail	Fail	Fail	Same data
Timeout	Timeout	Timeout	Source or peer cluster may has more data

One RegionServer Crashed



One RegionServer Crashed

source cluster

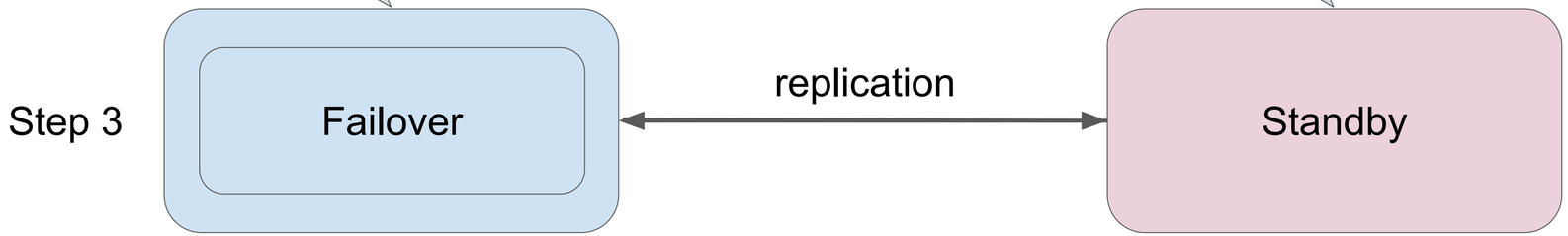
peer cluster

Case 1: source cluster has more data

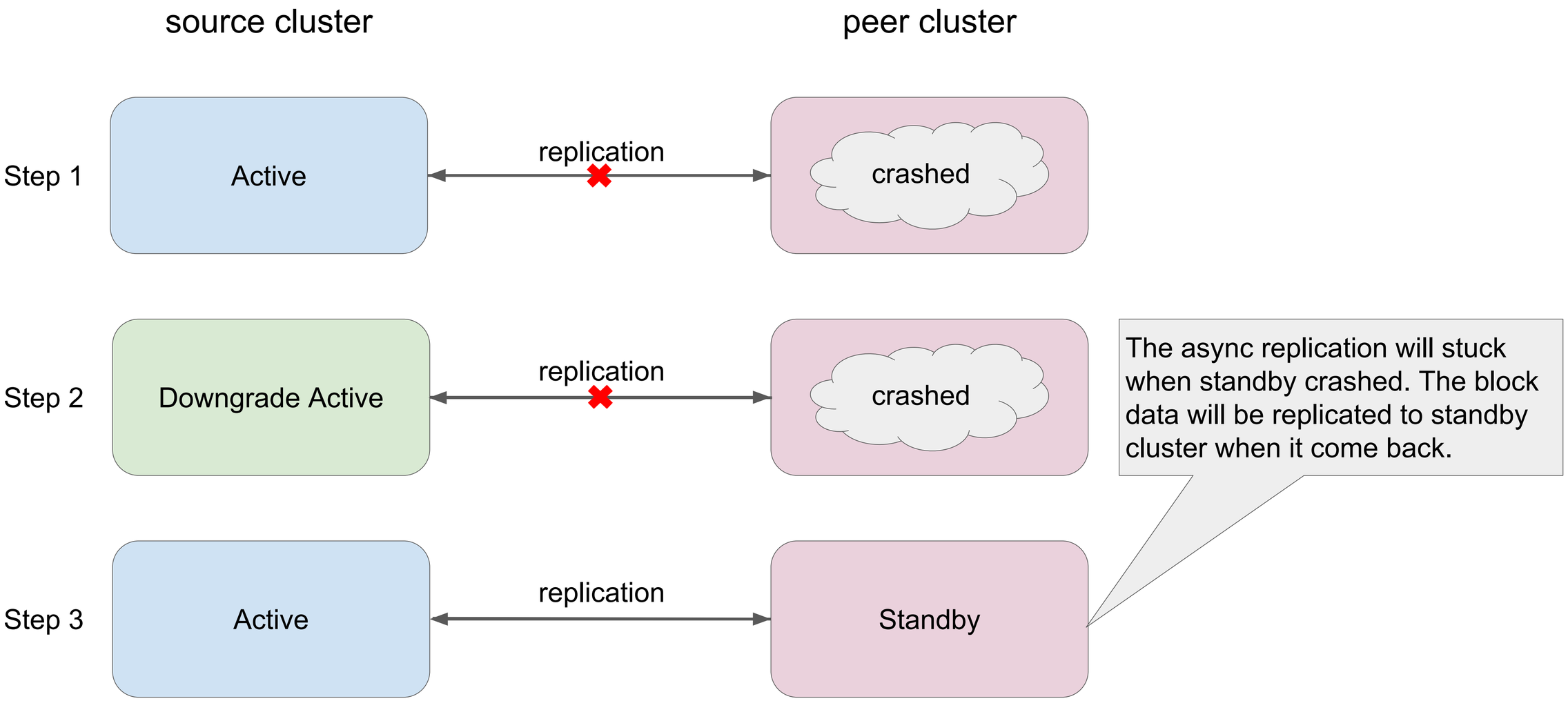
Solution: It will replicate data to peer cluster, so the data will eventually consistent.

Case 2: peer cluster has more data

Solution: Active cluster will copy WAL (which need to split) to remote WAL dir first, and the original remote WAL will be replaced. So the data will eventually consistent.



Standby Cluster Crashed



Standby Cluster Crashed

source cluster

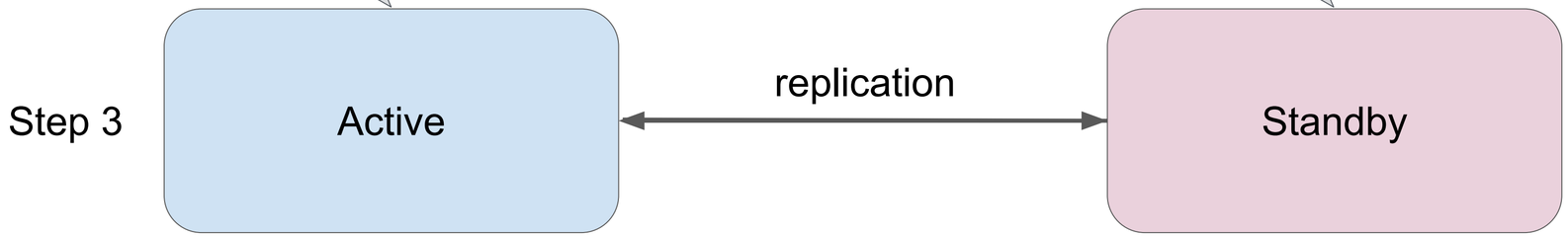
peer cluster

Case 1: source cluster has more data

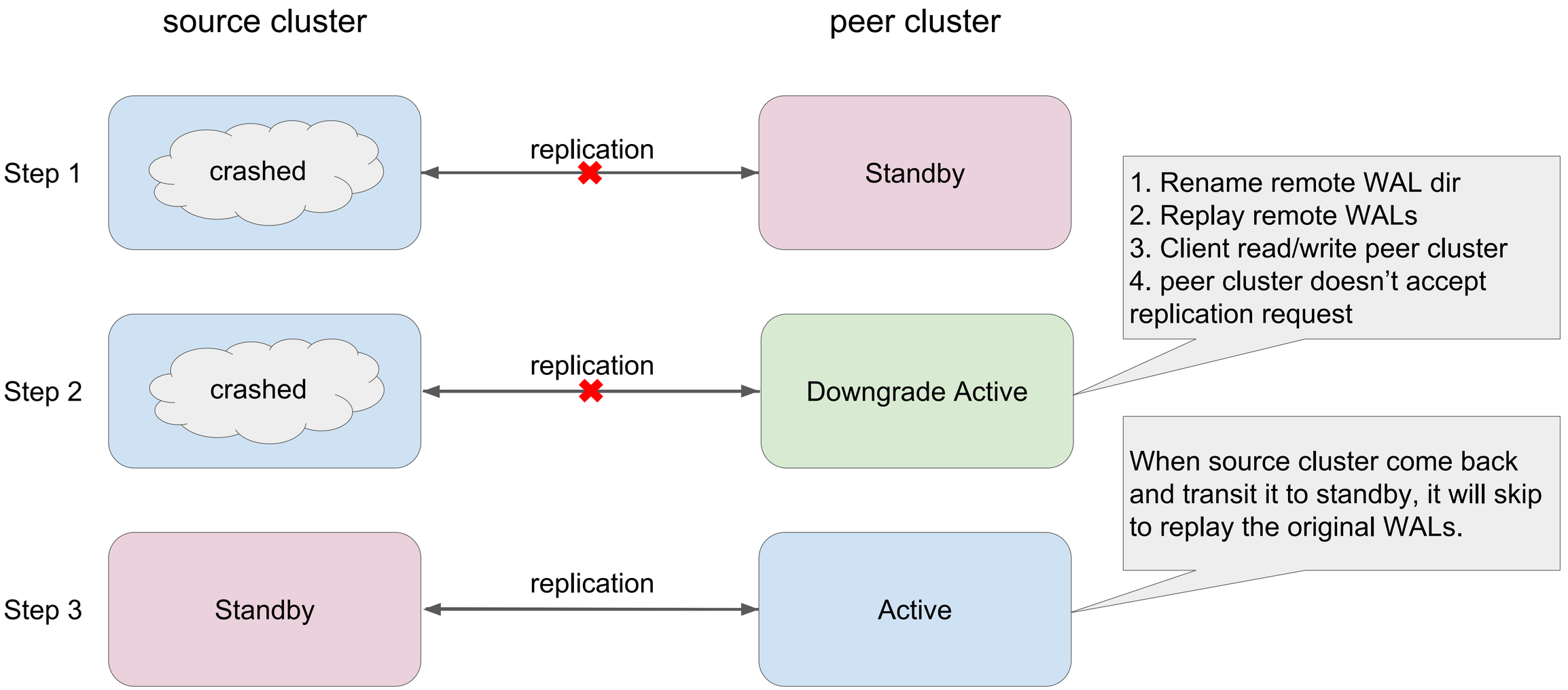
Solution: It will replicate data to peer cluster, so the data will eventually consistent.

Case 2: peer cluster has more data

Solution: The remote WALs not used anymore and will be deleted when the async replication replicate the original WALs to peer cluster. So the data will eventually consistent.



Active Cluster Crashed



Active Cluster Crashed

source cluster

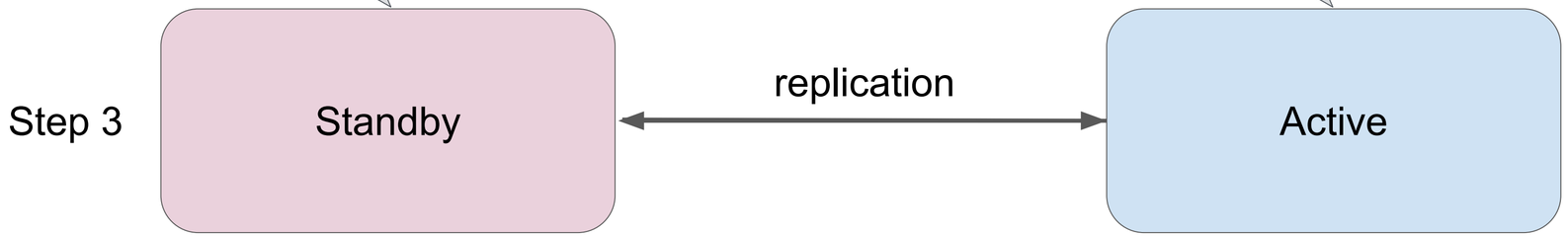
peer cluster

Case 1: source cluster has more data

Solution: It will skip to replay the original WALs(which has been replayed in peer cluster) and accept replication request which from peer cluster. So the data will eventually consistent.

Case 2: peer cluster has more data

Solution: It will replay the remote WALs and replicate it to source cluster. So the data will eventually consistent.



Sync Replication State Constraint

	Active	Downgrade Active	Standby
Write remote WAL	Yes	No	No
Accept client’s read/write request	Yes	Yes	No
Accept async replication request	No	No	Yes

Replication: Async *vs* Sync

	Async Replication	Sync Replication
Read Path	No affect	No affect
Write Path	No affect	Write extra remote WAL
Network bandwidth	100% for async replication.	100% for async + 100% for remote WAL.
Storage space	No affect	Need extra space for remote WAL
Eventual Consistency	No if active cluster crashed	Yes
Availability	Unavailable when master crash	Few time for waiting replay remote log.
Operational Complexity	Simple	More complex, Need to transition cluster state by hand or your services.

Performance

YCSB 10^8 operations, 100% Put

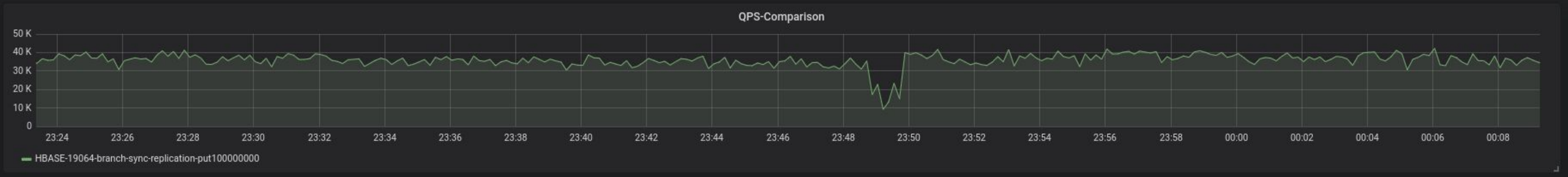
	Qps	Avg latency	P99 latency	P999 latency
Async replication	43K	3ms	<49.6ms	<567ms
Sync replication	37K	3.5ms	<74ms	<558ms

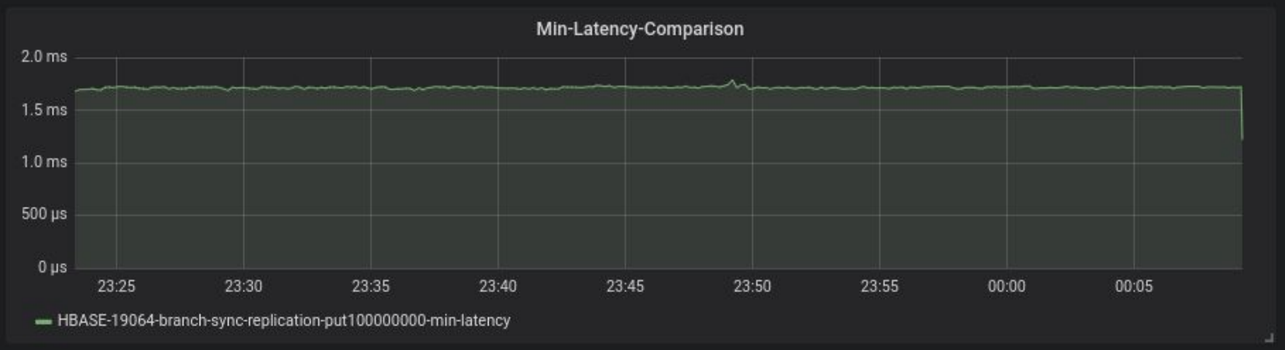
Qps: Almost 14% decline compared with async replication

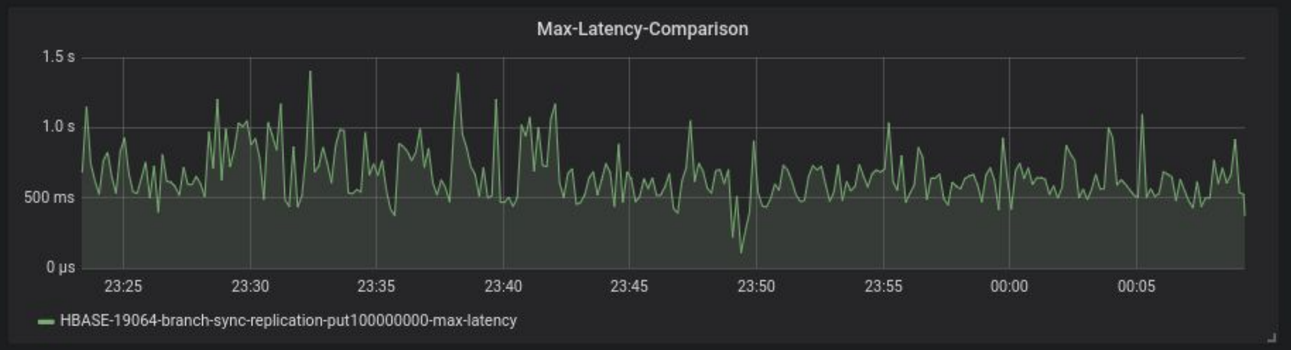
regionserver.heapsize=8g, datanode.heapsize=4g, CMS-gc, disk=1*4T(HDD), cpu=24core

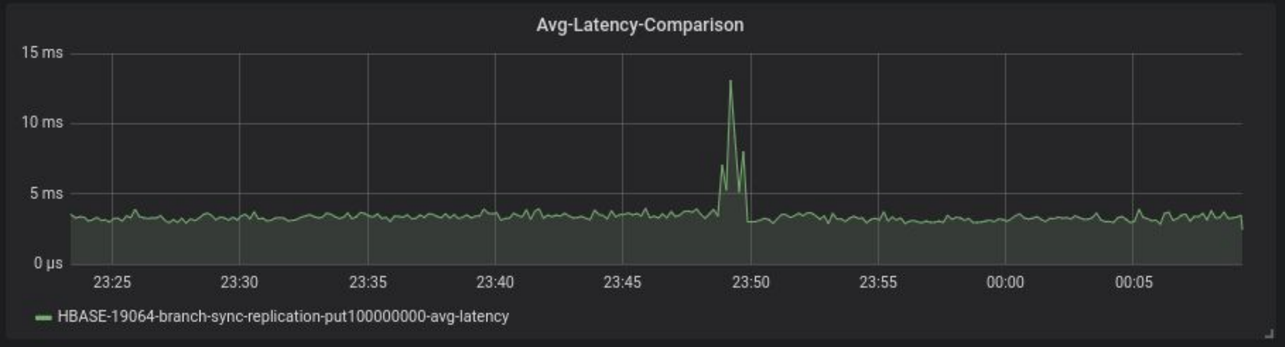
HBASE-19064-branch-sync-replication

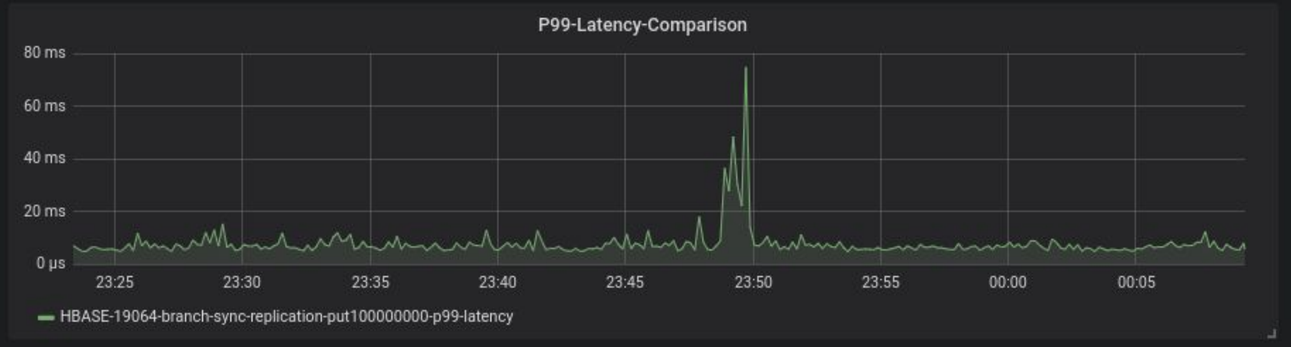
6月 25, 2018 23:23:21 to 6月 26, 2018 00:09:15

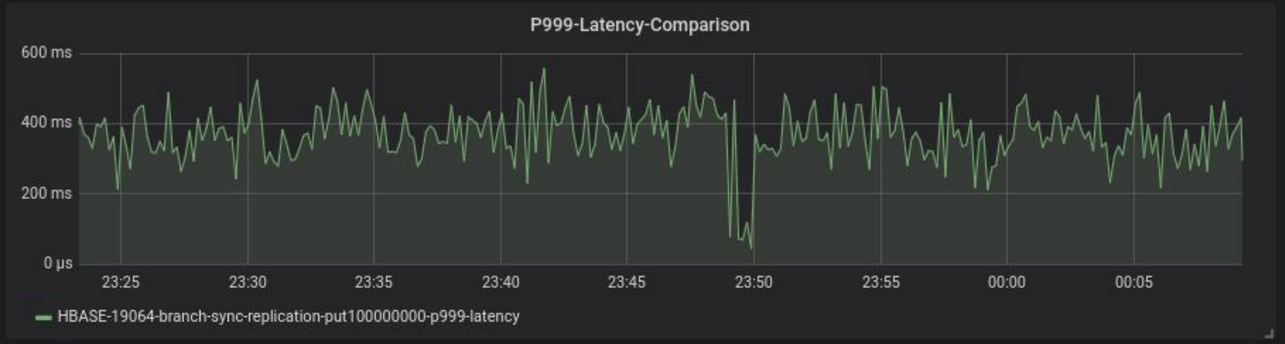


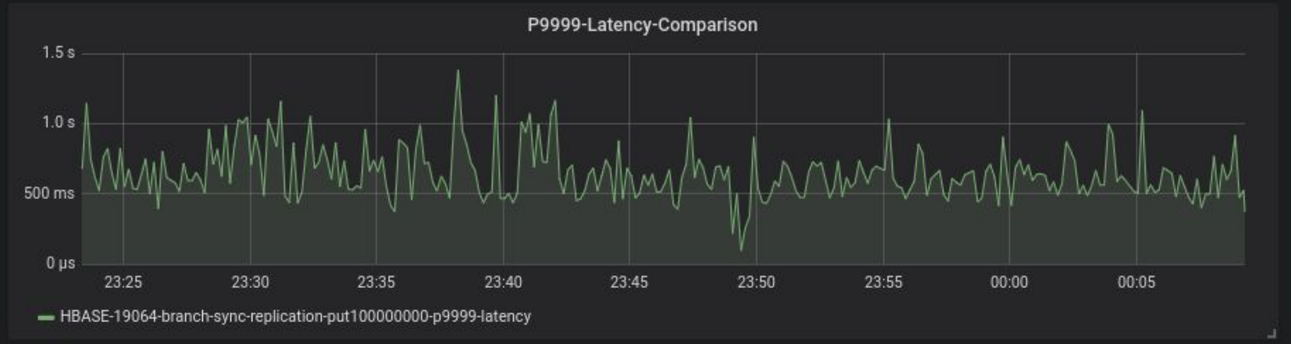












Sync Replication

HBASE-19064: The idea comes from Alibaba's presentation on HBaseCon Asia 2017

Sync Replication	Xiaomi solution	Alibaba solution
Base branch	master	0.94
Open source	will be released in hbase 3.0	internal solution
Qps	14% decline compared with async replication	2% decline compared with async replication

HBASE-20422: Synchronous replication for HBase phase II (OPEN)

Thanks

Guanghao Zhang
zghao@apache.org