阿里中间件
Alibaba
Middleware

HBase Synchronous Replication

Meng Qingyi
Shen Chunhui

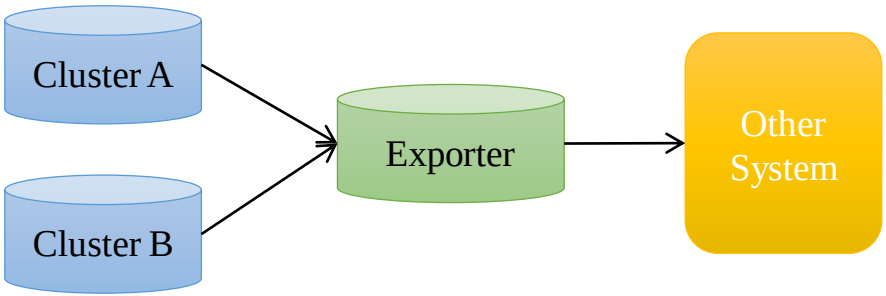
Alibaba Group

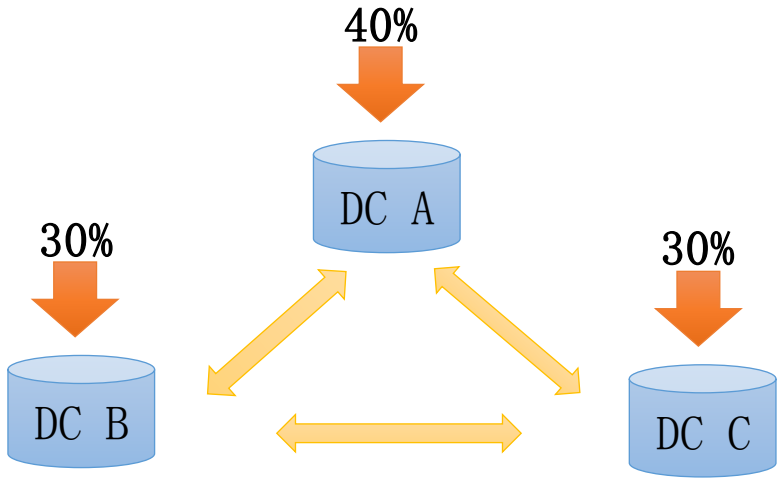
Outline

- Where to use replication
- Asynchronous replication
- Synchronous replication

Where to use replication

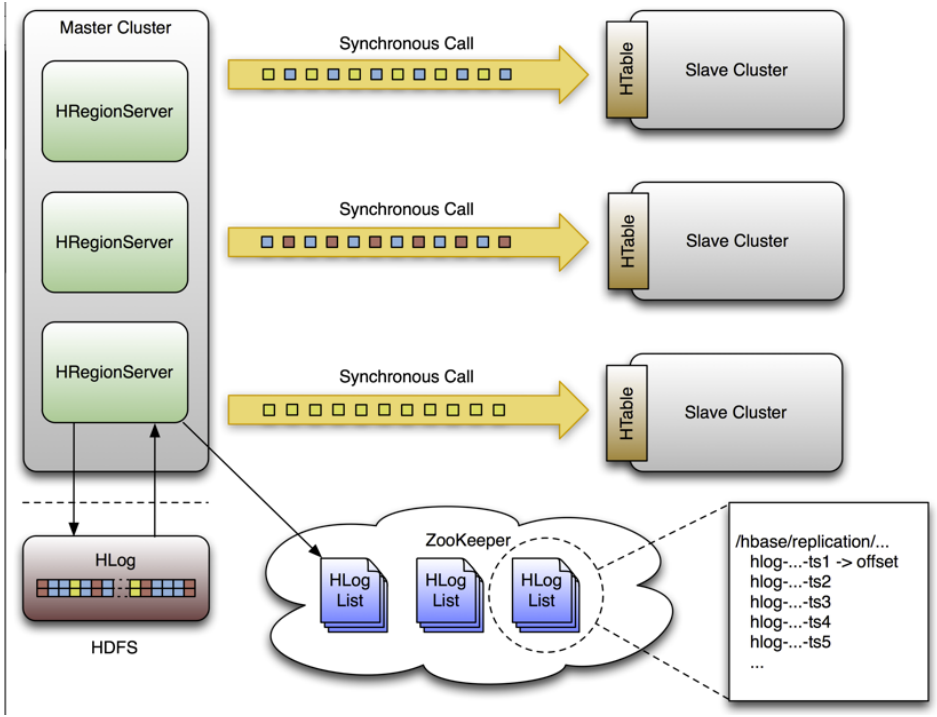
- Master-Slave
- Multi Datacenters
- Data Export





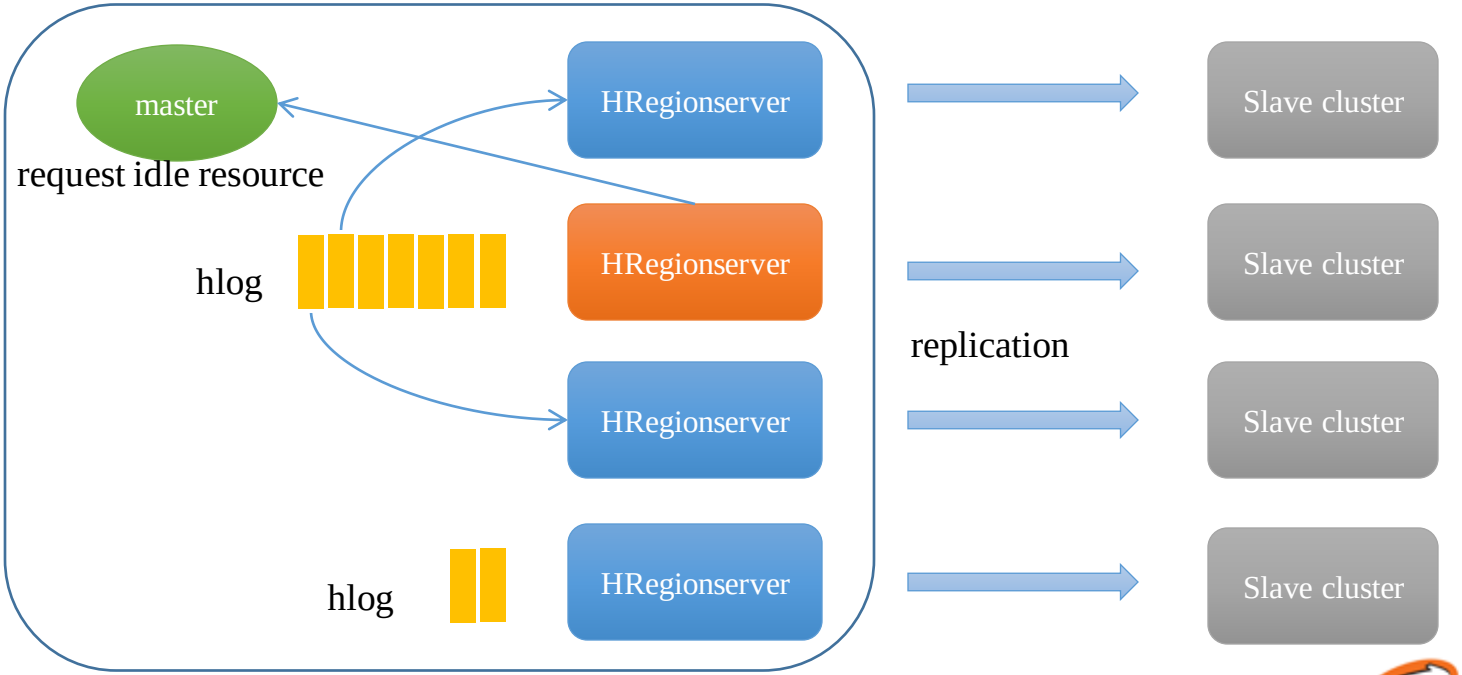
Asynchronous replication

- Improve
  - ✓ Enhance parallel on send
  - ✓ Enhance batch on sink
  - ✓ Use idle resources to reduce hotspot
  - ✓ Online configuration change
  - ✓ Replication failover isolation



Asynchronous replication

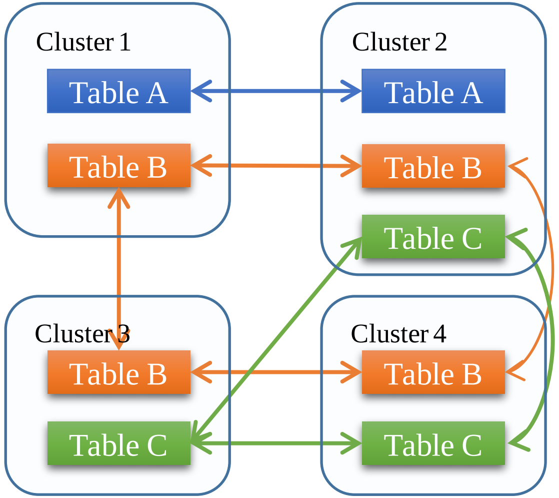
- Reduce hotspot

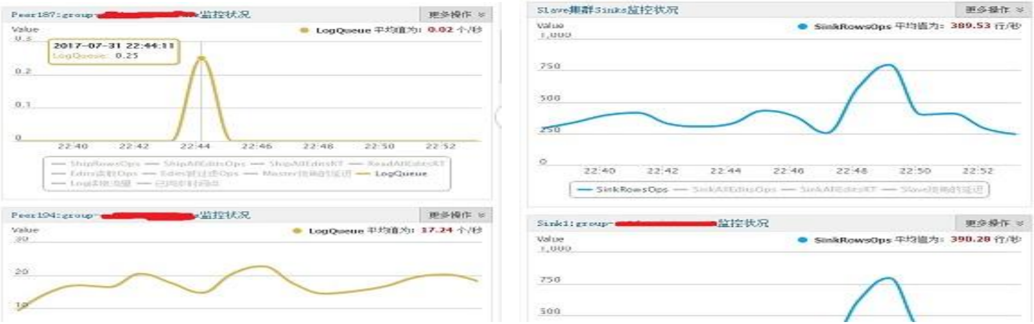


Asynchronous replication

- Replication topology
  - ✓ Table scope replication
  - ✓ Replication topology monitor
  - ✓ Replication cycle

Master -> Slave	ReplSyncedTime	M-View Delay	LogQueue	RS-LEVEL-LOAD	M-REPL-STATUS
group-hbase-[redacted]-mix -> group-[redacted]-mix	2017-07-31 22:43:43 (1501512223276)	MAX: 0.00 ms AVG: 0.00 ms	MAX: 0 AVG: 0	Detail	group-hbase-[redacted]-mix
group-hbase-[redacted]-mix -> group-[redacted]-mix	2017-07-31 22:43:43 (1501512223275)	MAX: 0.00 ms AVG: 0.00 ms	MAX: 0 AVG: 0	Detail	group-hbase-[redacted]-mix
hbase-[redacted]-core -> group-[redacted]-mix	2017-07-31 22:43:52 (1501512232240)	MAX: 0.00 ms AVG: 0.00 ms	MAX: 0 AVG: 0	Detail	hbase-[redacted]-core
hbase-[redacted]-core -> group-[redacted]-mix	2017-07-31 22:43:52 (1501512232168)	MAX: 214.00 ms AVG: 113.00 ms	MAX: 0 AVG: 0	Detail	hbase-[redacted]-core
hbase-[redacted]-core -> hbase-[redacted]-mix	2017-07-31 22:43:51 (1501512231730)	MAX: 555.00 ms AVG: 319.00 ms	MAX: 0 AVG: 0	Detail	hbase-[redacted]-core
hbase-[redacted]-core -> group-[redacted]-mix	2017-07-31 22:40:53 (1501512053767)	MAX: 4.16 s AVG: 922.00 ms	MAX: 1 AVG: 0	Detail	hbase-[redacted]-core
group-[redacted]-ssd -> group-[redacted]-ssd	2017-07-31 22:43:43 (1501512223426)	MAX: 197.00 ms AVG: 40.00 ms	MAX: 0 AVG: 0	Detail	group-[redacted]-ssd





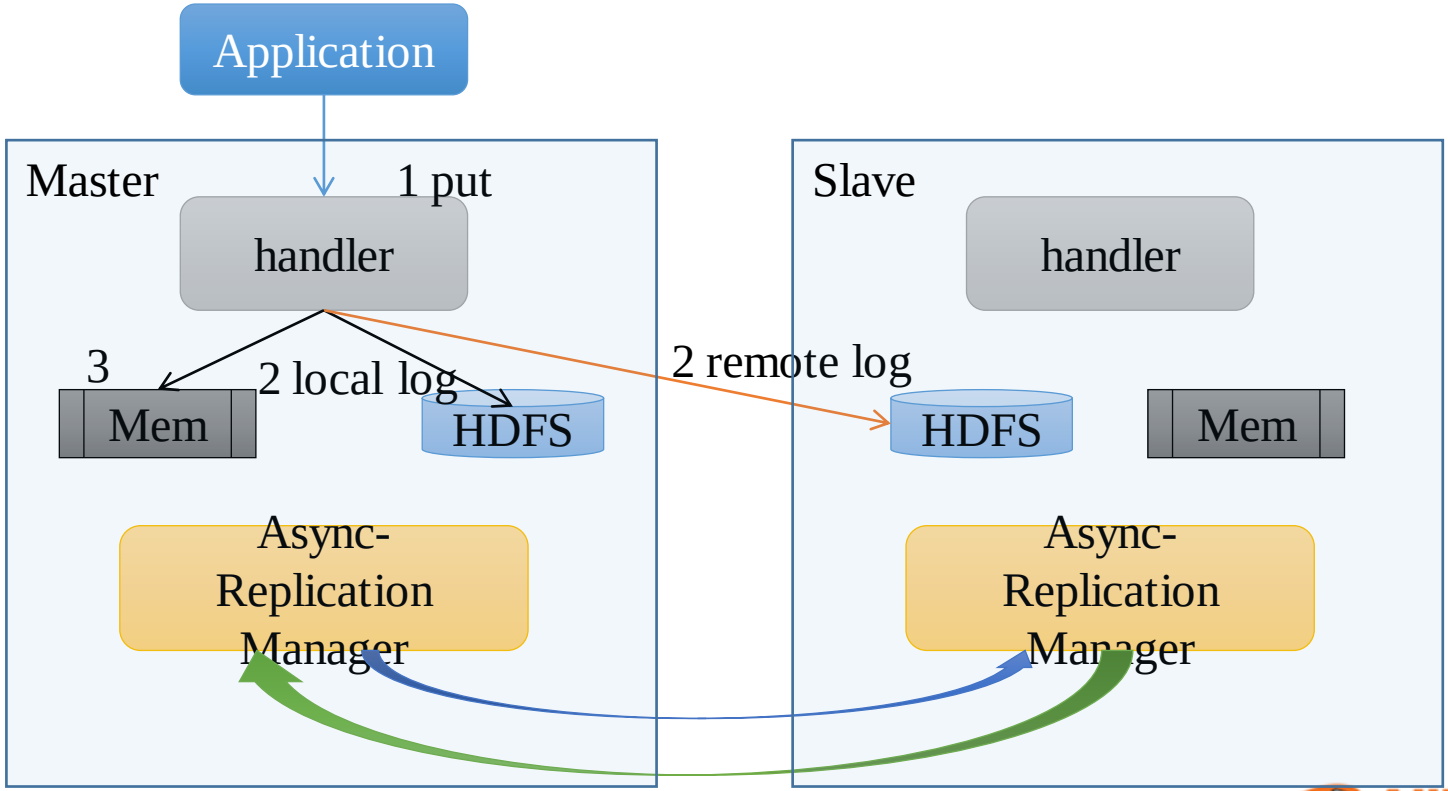
Synchronous replication

- Motivation
  - ✓ Replication within two datacenter
  - ✓ Access master on normal
  - ✓ Switch to slave when master down
  - ✓ Strong consistency on access

Synchronous replication

- Consistency semantic
  - ✓ Write
    - -Success when response is “success”
    - -Unknown when response is “failure”
  - ✓ Read
    - -Data is always readable after it is written successfully
  - ✓ In any circumstances, data remain eventual consistency between master and slave

Synchronous replication



Remote log

- Log content
  - ✓ Data not yet replicated by asynchronous replication
- File format
  - ✓ Same as hlog, collection of entries
- Log organization
  - ✓ remote log and hlog is many to one relationship
  - ✓ Use same prefix for file name
  - ✓ Store on slave hdfs

/hf-A/.logs/[redacted],64020,1467366444864/[redacted]%2C64020%2C1467366444864.1467366450457
/hf-B/.remotelogs/[redacted],64020,1467366444864/[redacted]%2C64020%2C1467366444864.1467366450457.1

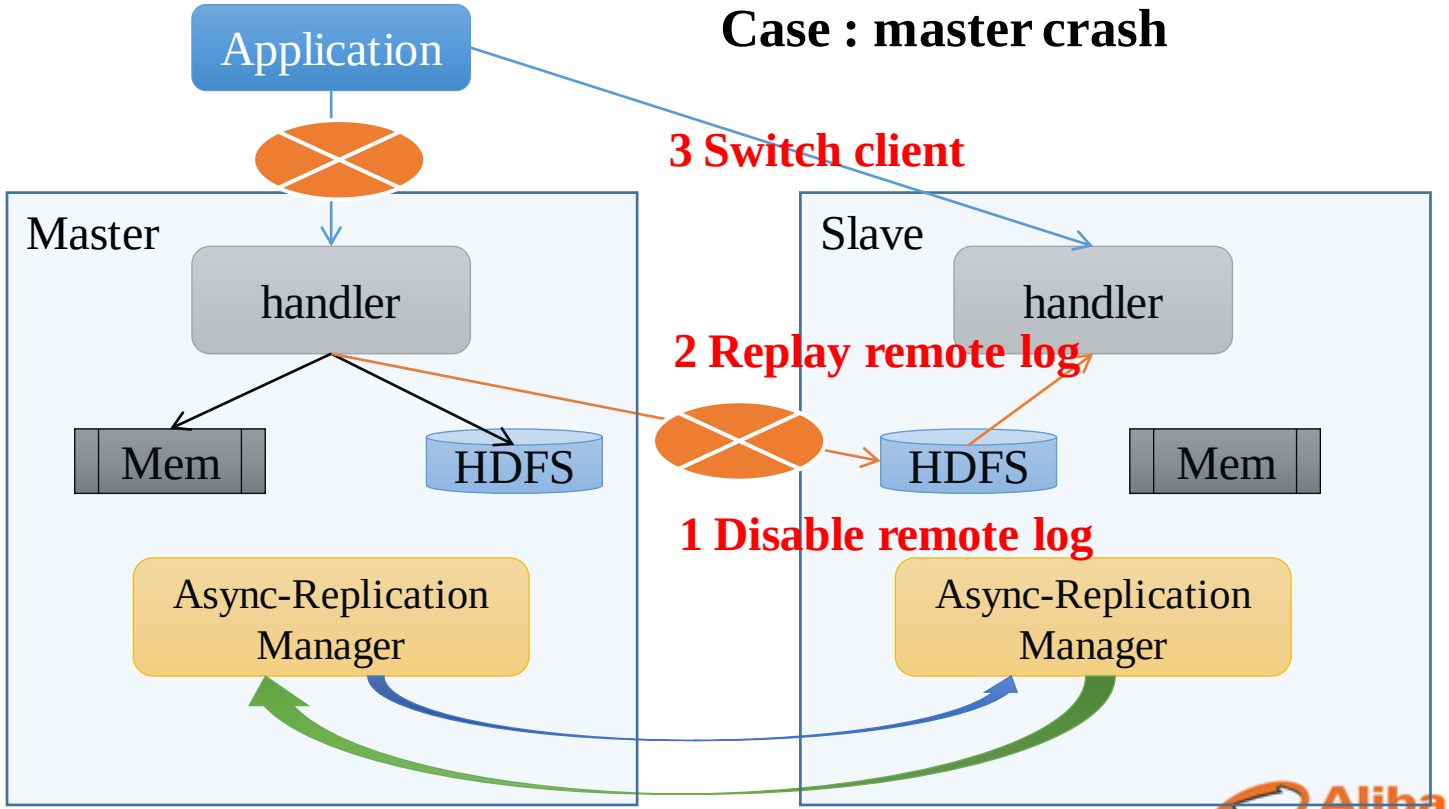
Remote log clean

- When to clean remote log?
  - ✓ when the corresponding hlog is replicated by asynchronous replication
- Who clean remote log?
  - ✓ Master cluster

Remote log

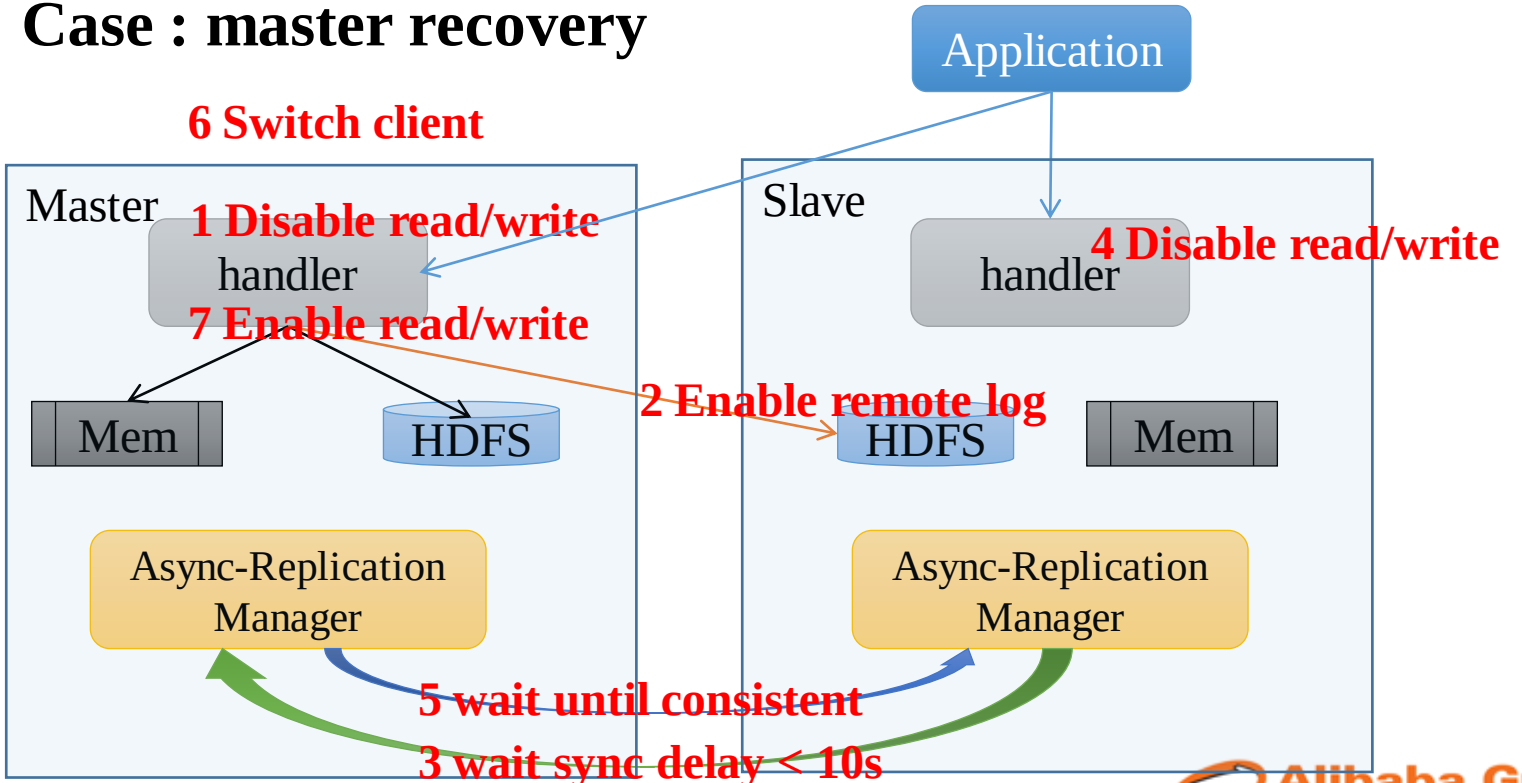
- When need disable Remote log
  - ✓ Before switch. There may be some client still accessing master.
- How to disable Remote log
  - ✓ Create lock file
  - ✓ recover lease for current remote logs

Failure scenarios

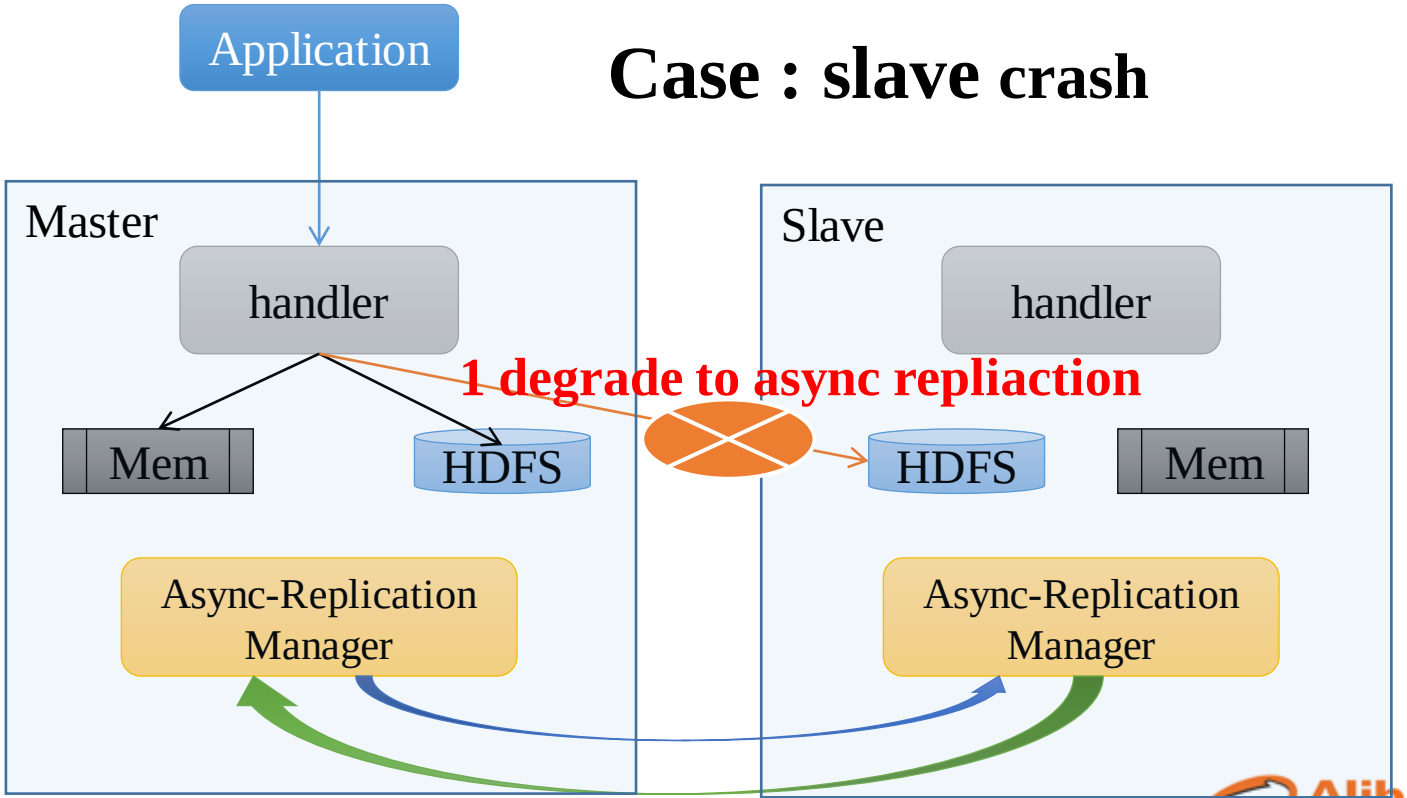


Failure scenarios

Case : master recovery



Failure scenarios



Consistency

Case	Action	Consistency
Local log success Remote log fail	1 Block and retry forever 2 if server crash, write remote log again on replay	Keep consistence when retry success
Local log fail Remote log success	Return fail to client	1 if client keep accessing master, remote log will be delete and never replay on slave 2 before remote log is delete, client switch to slave. Remote log will be replay and seen by client, async-replication will deliver this log back to master
Local log fail Remote log fail	Return fail to client	Remain consistence
Local log success Remote log success	Return success to client	Remain consistence

Switch support

- Availability monitor
  - ✓ Network partition
  - ✓ Node crash
  - ✓ Error rate
- Switch API
  - ✓ Define active and backup
    - -Active cluster is the one access by clients
    - -Backup cluster is disabled for access
  - ✓ Define switch process from cluster A to cluster B
    - -Switch A from active to backup
    - -Switch B from backup active
  - ✓ Unify synchronous and asynchronous
- Client switch
  - ✓ Logical cluster address
  - ✓ Push new cluster address

Synchronous replication

- Use case
  - ✓ Internal state for stream processing
  - ✓ Sequential access: pub/sub system
  - ✓ CheckAndPut operation
- Performance
  - ✓ 2% throughput decline than async replication (network delay = 0.5ms)

Synchronous vs. Asynchronous

	Asynchronous	Synchronous
Read Path	No affect	No affect
Write Path	No affect	~2% throughput decline
Network	100% for asynchronous replication	200% for asynchronous replication and remote log
Eventual consistency	No if master crash and can not recover	Yes
Availability	Blocking until master replication recovery which may take hours on massive crash	Block few minus waiting remote log replay
Storage space	2 copy	2 copy + remote log(small)

Thanks

tianwu.sch@alibaba-inc.com
qingyi.mqy@alibaba-inc.com

Alibaba Group