zookeeper多活實踐

主講人：趙子明

2017.08.11

拉扎斯网络科技（上海）有限公司
Rajax Network & Technology Co

餓了麼簡介

願景：以創新科技打造全球領先的本地生活平台

主營業務：外賣、商超、即時配送、餐飲供應鏈等

業務覆蓋城市/地區：2000+
C端用戶: 2.6億
B端商家: 130萬
註冊配送員：300萬
訂單量：千萬級

e饿了么

Make Everything 30’

目錄:

zookeeper簡介

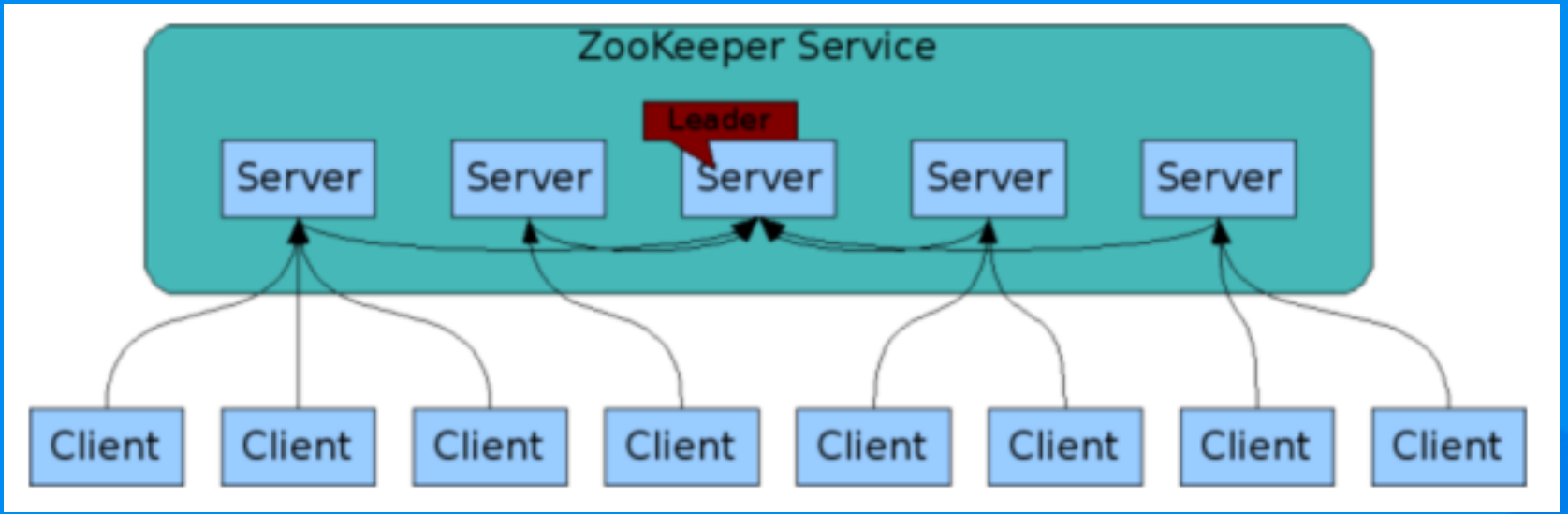
高可靠的分佈式協調服務，為分佈式而生

使用zab協議（改進版/簡化版的Paxos協議）

使用場景：數據發布訂閱（配置管理），命名服務，集群管理，分佈式鎖等

廣泛應用在大數據，微服務等場景下 hadoop kafka storm dubbo等

一個帶通知功能的存儲！

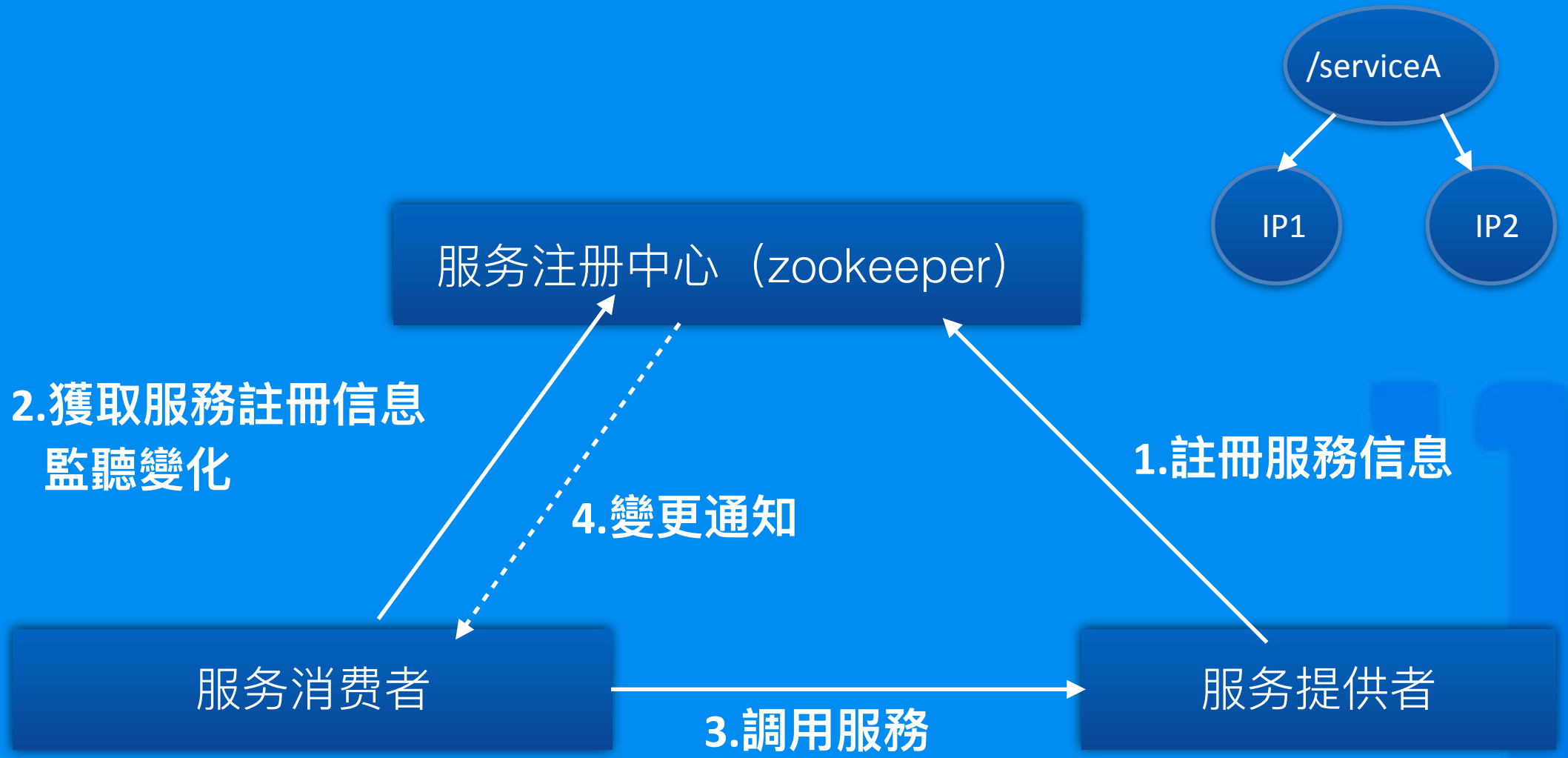


集群需要奇數節點

一個事務需n/2+1節點投票

必須超過一半多機器可用，集群才可用

Quorum



服務註冊發現：餓了麼zookeeper最重要的使用場景

餓了麼多活

異地多機房數據分片+按POI做流量分發，隨時可容災切換

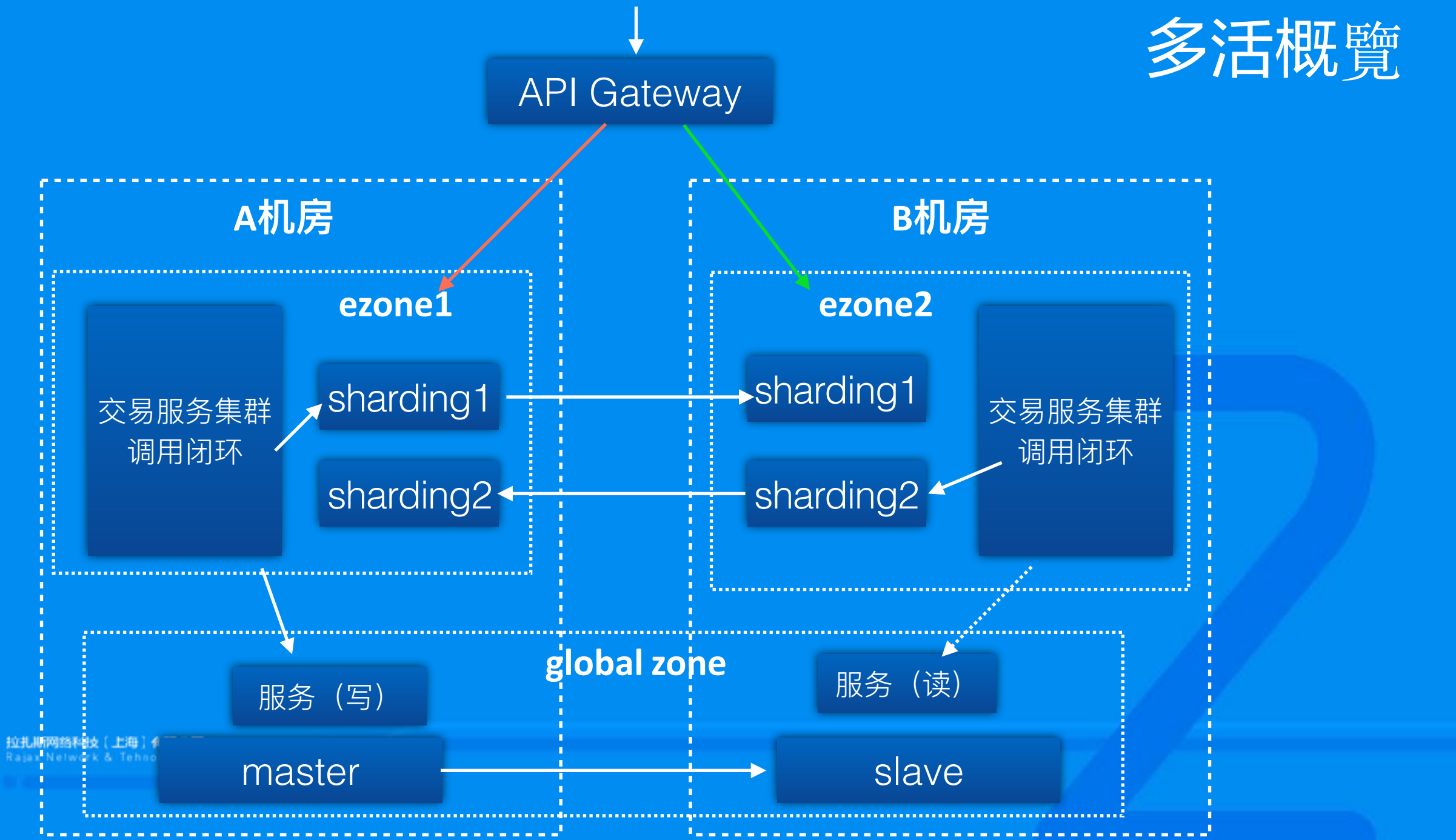
數據：多寫+實時雙向複製

流量：按POI做流量分發+可全局切換

服務：每個機房提供完整交易服務鏈路，機房內調用閉環

萬台服務器，千個工程師

多活概覽


API Gateway
A机房
B机房
ezone1
ezone2
交易服务集群
调用闭环
sharding1
sharding2
sharding1
sharding2
交易服务集群
调用闭环
global zone
服务（写）
服务（读）
master
slave

多活下的服務調用

核心交易鏈路服務調用閉環(下單)

其他服務跨機房調用（用戶積分）

多活對zookeeper的要求

任何一个机房出问题，其他机房内zookeeper可用（超過一半多機器可用連接），服务可调用！

正常情况下可以跨機房調用

Zookeeper常用部署方案在多活中有什麼問題?

常用部署方案：

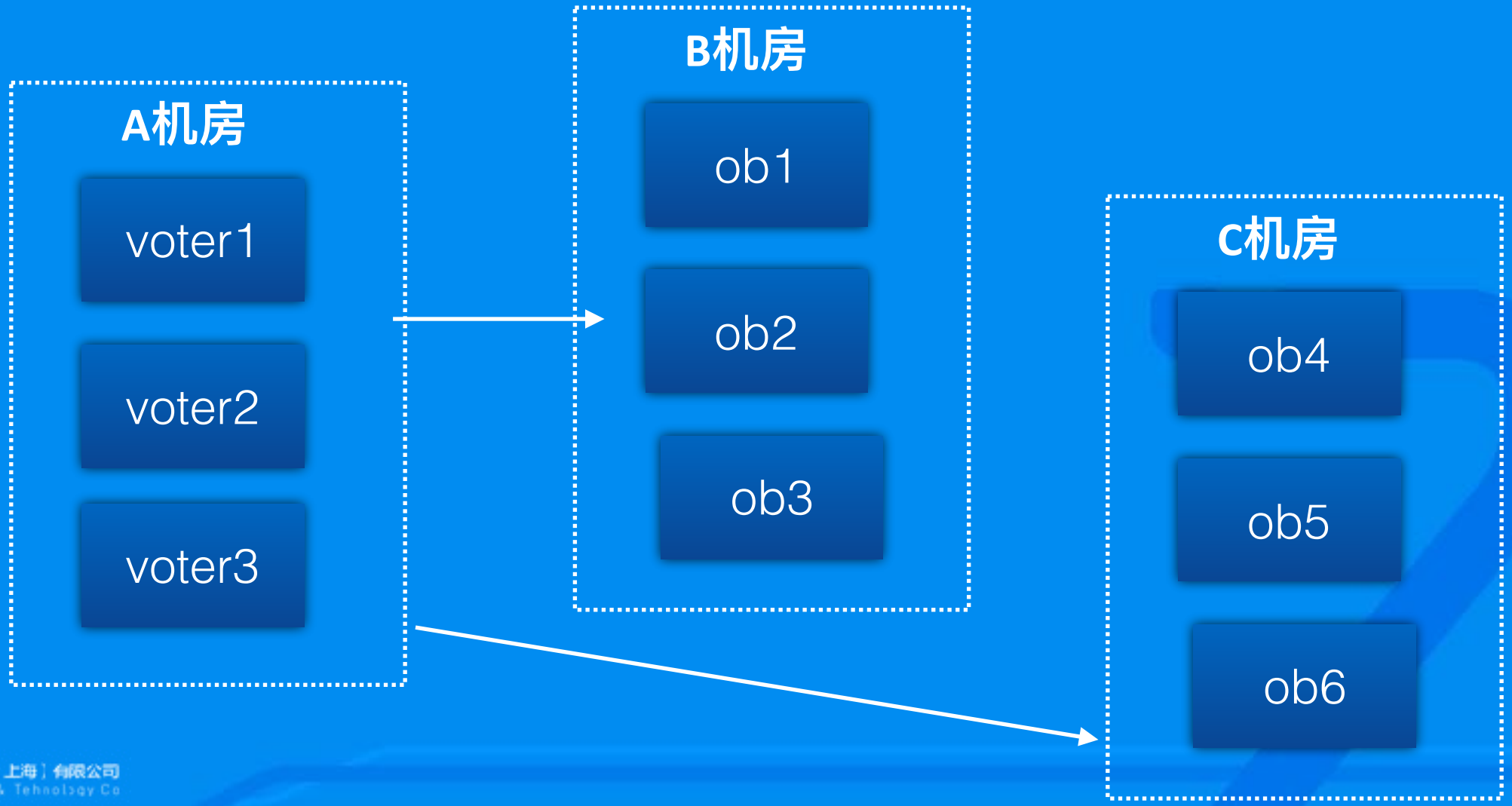
1.主機房voter 其他機房observer

2.多機房voter

3.多機房多集群

注： voter是參與投票的節點（leader/follower）

voter+observer部署方案



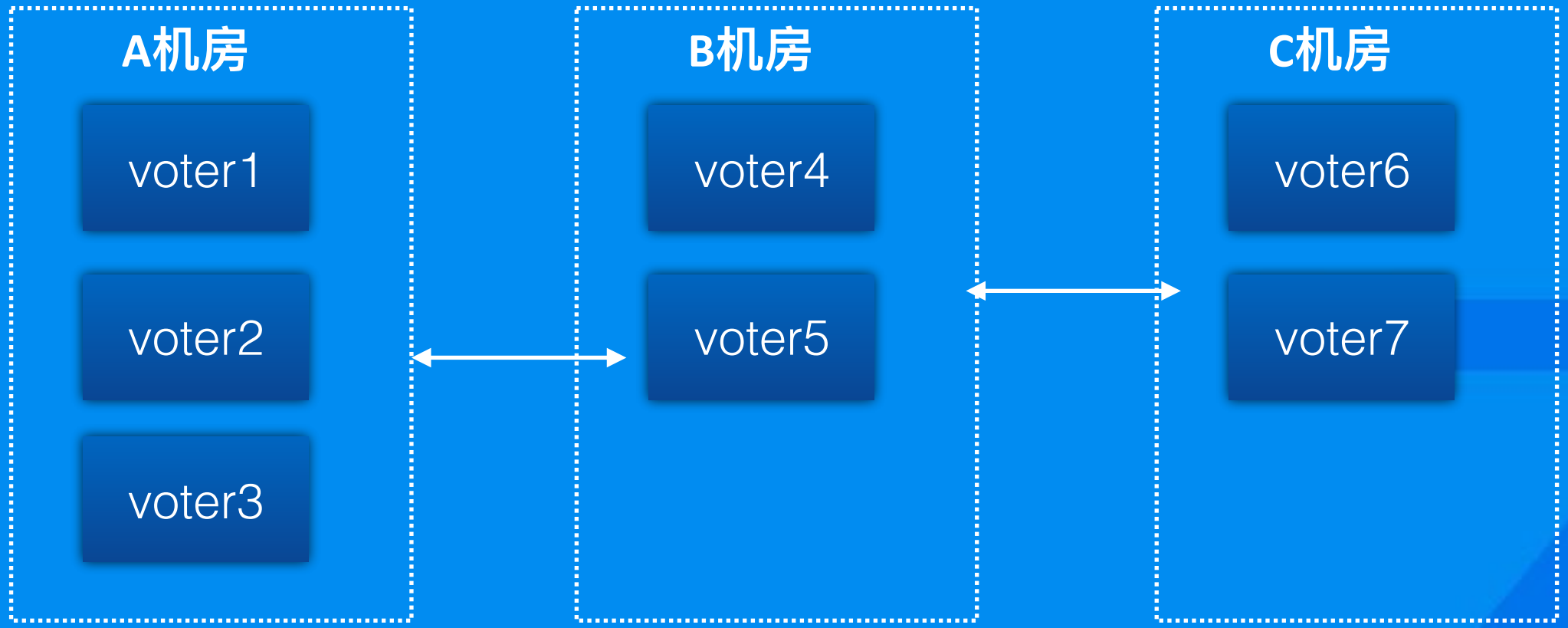
問題：

寫延遲適中（事務只在一個機房投票完成）

observer機器讀延時大

voter所在機房有問題，其他所有機房zookeeper 都不可用

多voter部署方案



N1=(N-1)/2

3

N2= 1~(N-N1)/2

2

N3=N-N1-N2

2

一個機房有問題，其他兩個機房超過一半機器連接，服務可用

問題：

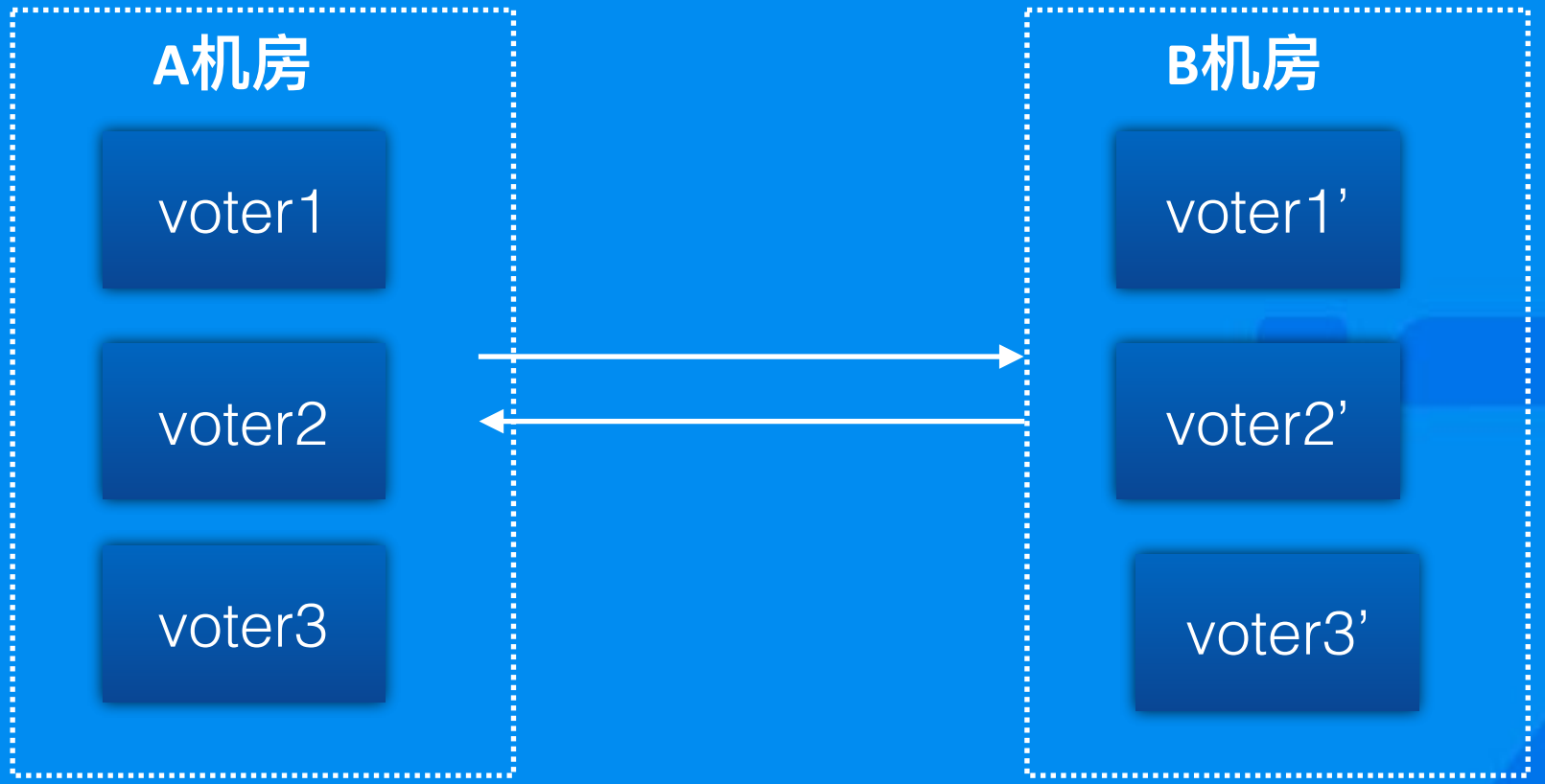
寫延遲大（事務需在多個機房間投票完成）

個別機房讀延遲大

一個機房掛掉或網絡問題，其他兩個機房可用

我們需要可用性更高的方案！

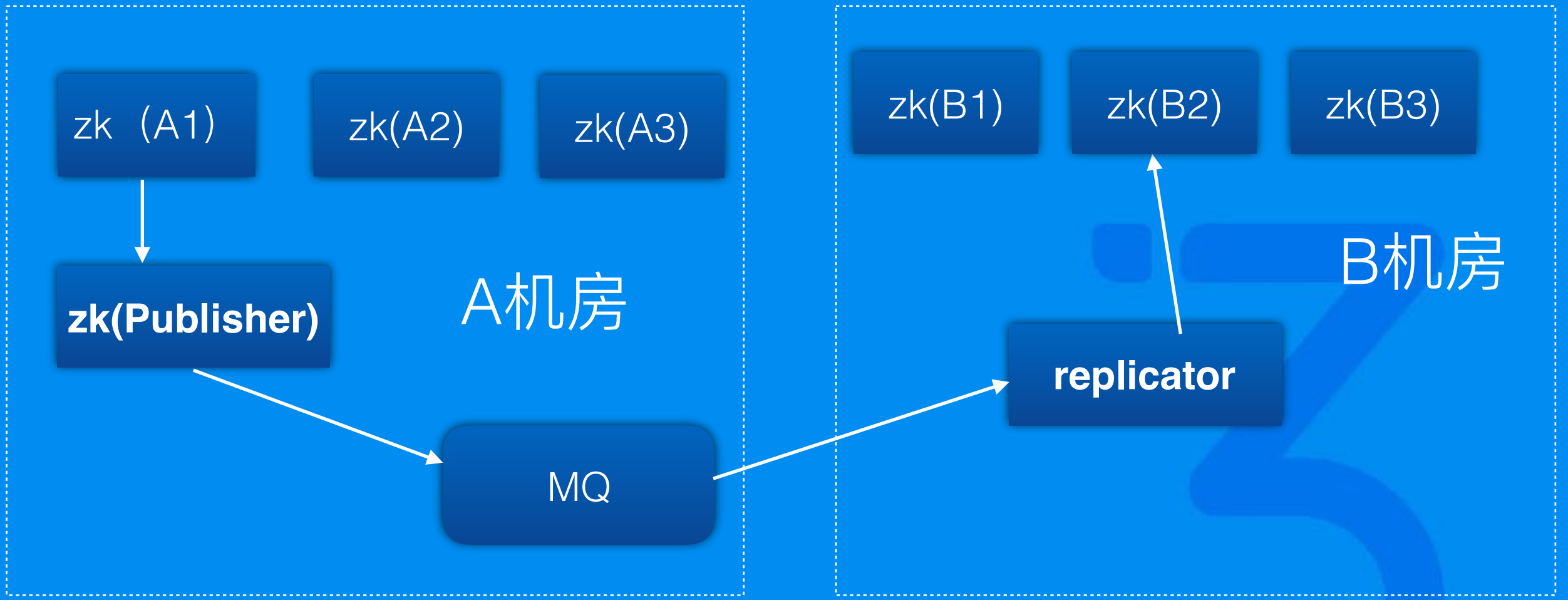
zookeeper多活方案



優先保證機房內部可用

多機房獨立集群+數據雙向同步

组件简介


zk （A1）
zk(A2)
zk(A3)
zk(Publisher)
A机房
MQ
zk(B1)
zk(B2)
zk(B3)
replicator
B机房

publisher

在zookeeper observer基礎上擴展

接收leader發來的事務並發送到mq中

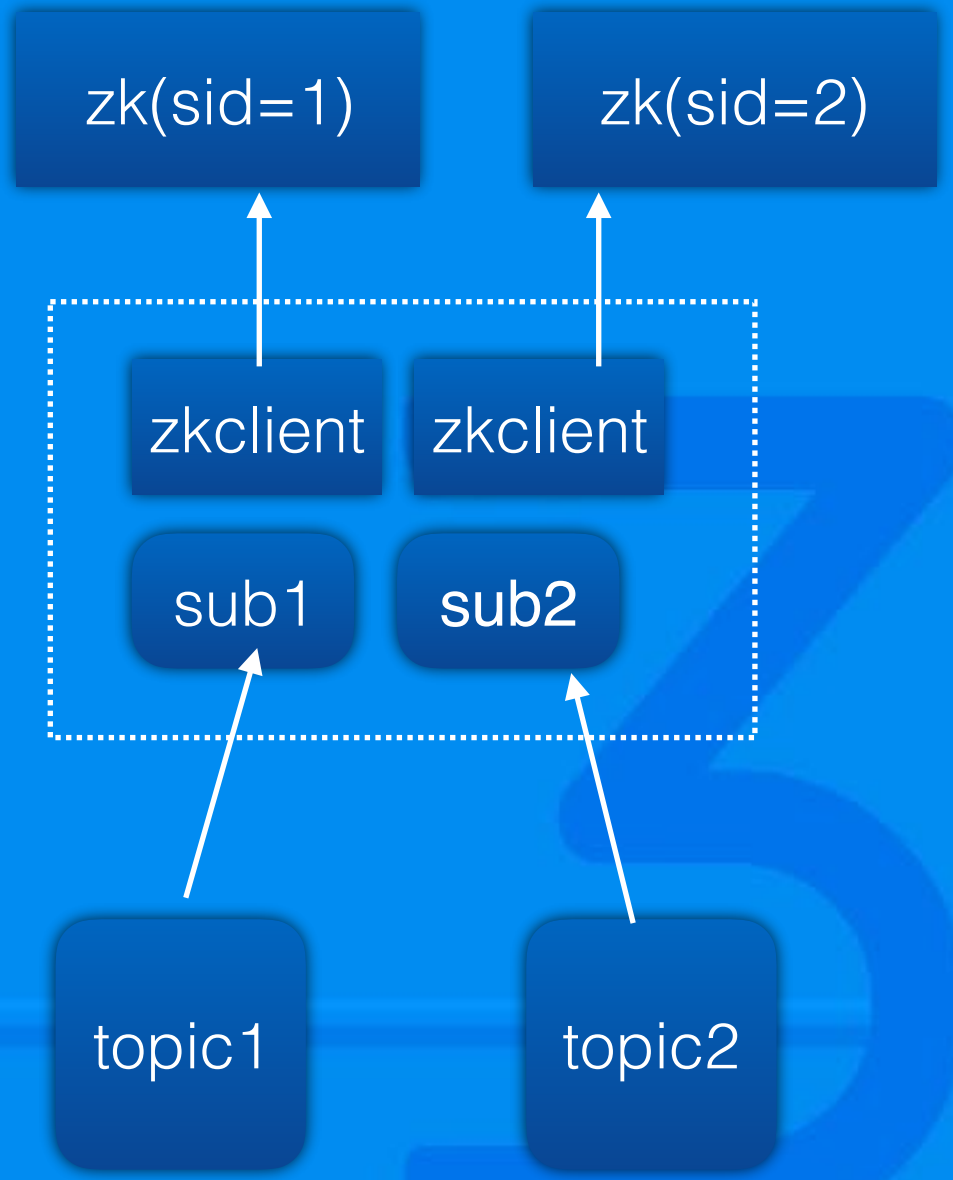
replicator

消費消息+寫入zookeeper

可訂閱多個topic

只寫到固定機器上

目標zookeeper需要配置超級權限

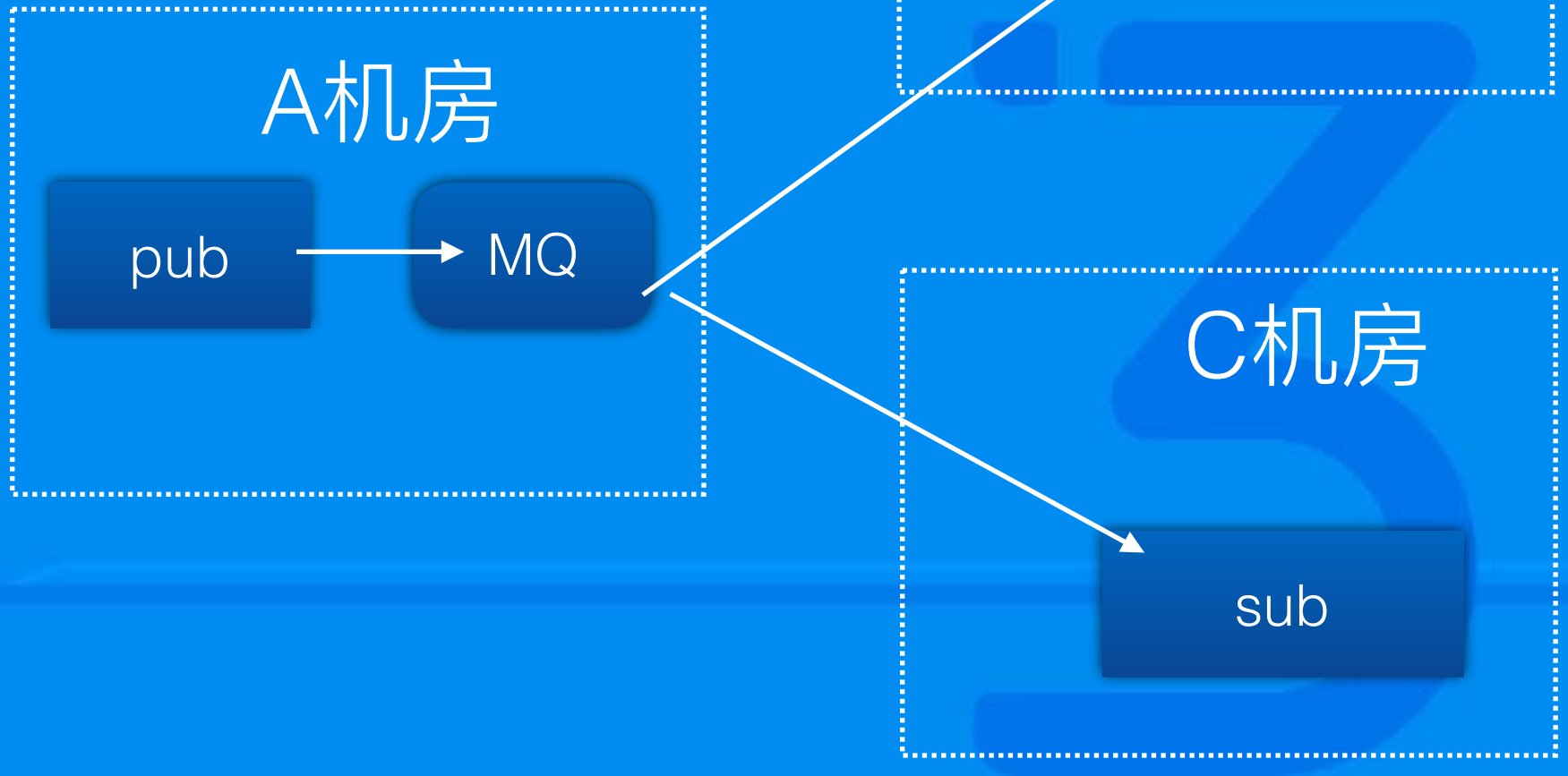


MQ

和publisher同機房部署,跨机房sub

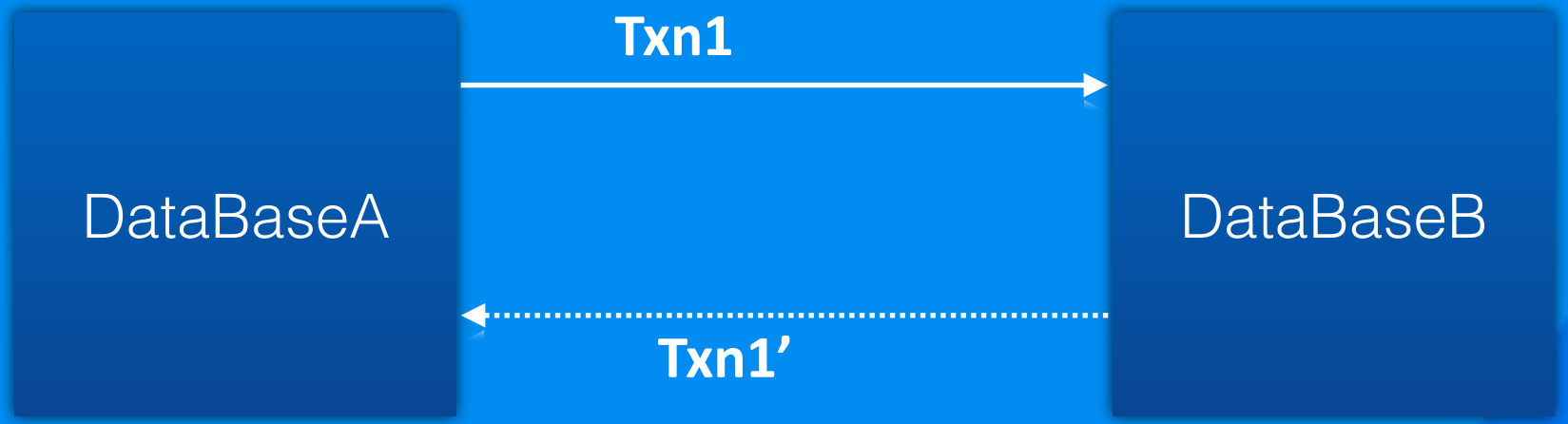
選運維同學熟悉的mq

支持持久化

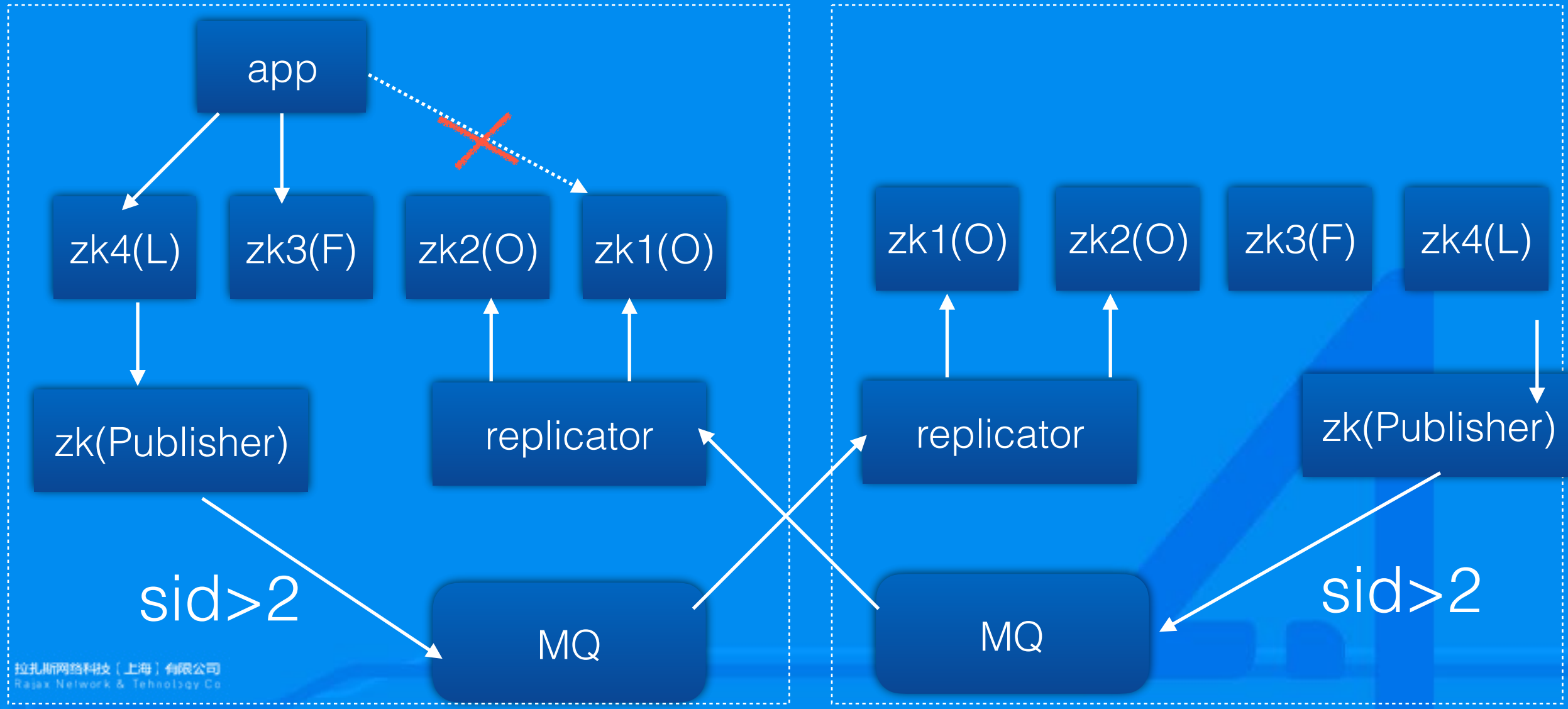


這個方案（組件）第一大難點：

循環複製


Txn1
DataBaseA
DataBaseB
Txn1’

午休帶來了靈感：找到標識過濾



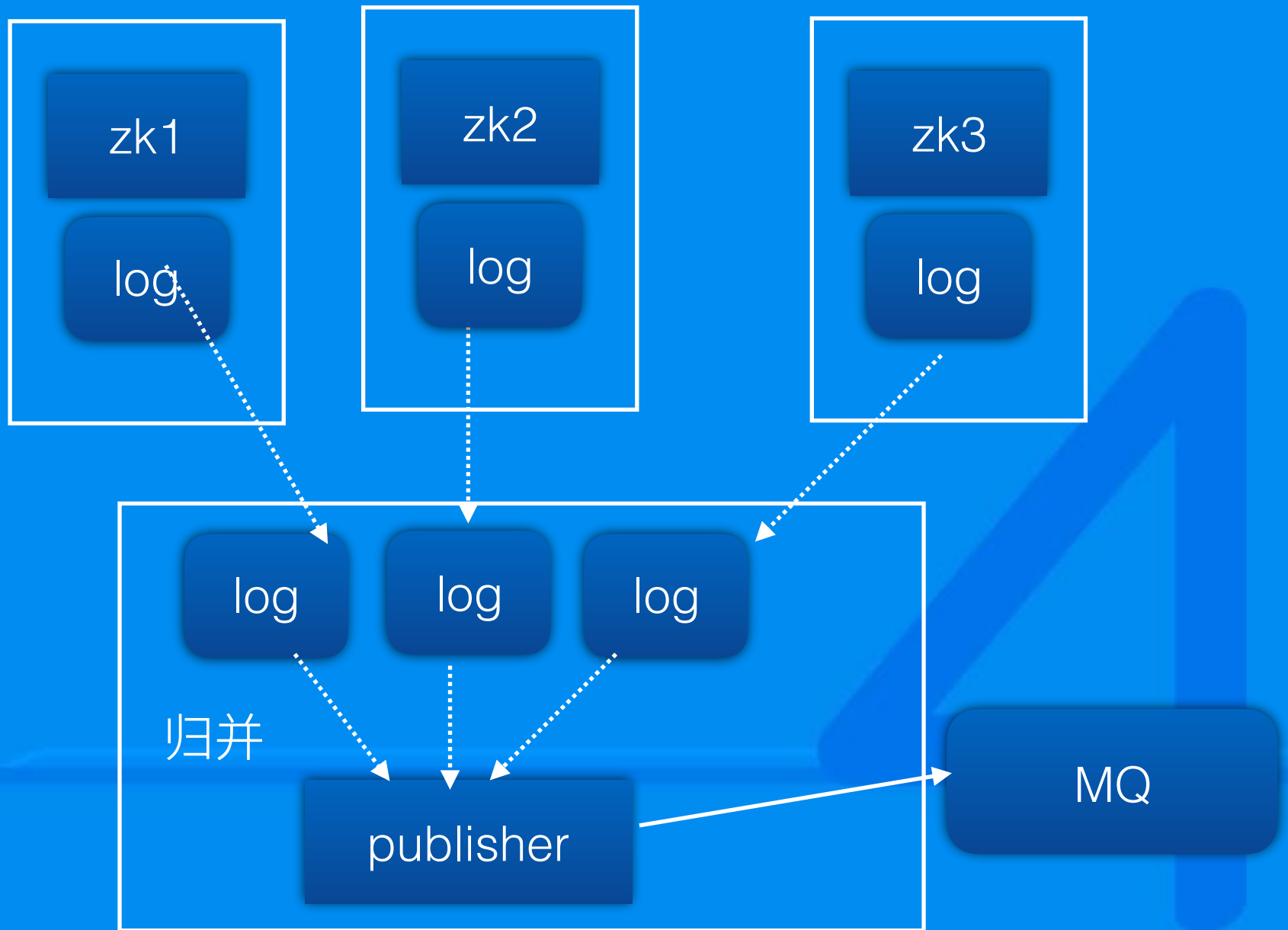
根據sessionId中的機器號區分內外數據

第二大難點：

事務連續性

超過500條，觸發leader -> publisher之間內存鏡像全量同步,會丟失中間狀態！

Quorum歸併方案



運維技巧

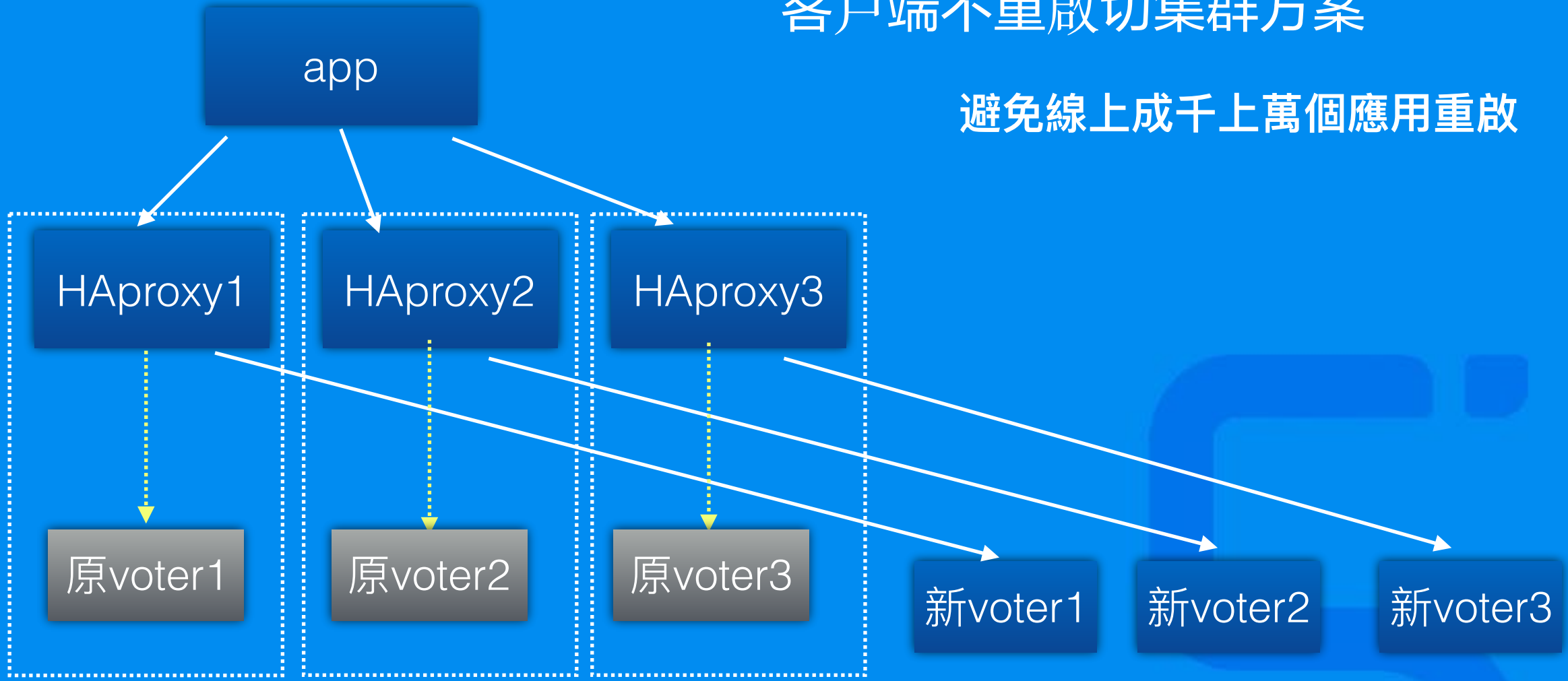
連接串 Domain > IP（切換方便）

最多5個節點，擴容用observer

無影響集群擴容：一台台增加myid 高於voter集群的observer

客戶端不重啟切集群方案

避免線上成千上萬個應用重啟



新集群事務ID > 就集群事務ID

The End

ming

欢迎关注饿了么技术社区

拉扎斯网络科技
Rajax Network &