



# 第十一届中国数据库技术大会

DATABASE TECHNOLOGY CONFERENCE CHINA 2020

## 架构革新 高效可控



北京国际会议中心 | 2020/12/21-12/23

# Dmall OS数字化零售数据库多活架构演进

冯光普 @ 多点Dmall



架构革新 11<sup>th</sup>  
高效可控





# About Me



目前，多点Dmall，数据库负责人

- 数据库运维、架构、平台建设

更早，阿里巴巴，MySQL内核开发

- AliSQL分支特性开发、维护



# Dmall OS简介

架构革新 © 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020

## 3 个 场 景



**商圈**

综合体商超



**社区**

社区便利店  
拼团



**到家**

全球精选O2O

## 2 大 原 则



**助力 B**



**服务 C**

## 2 套 系 统



**Dmall OS**



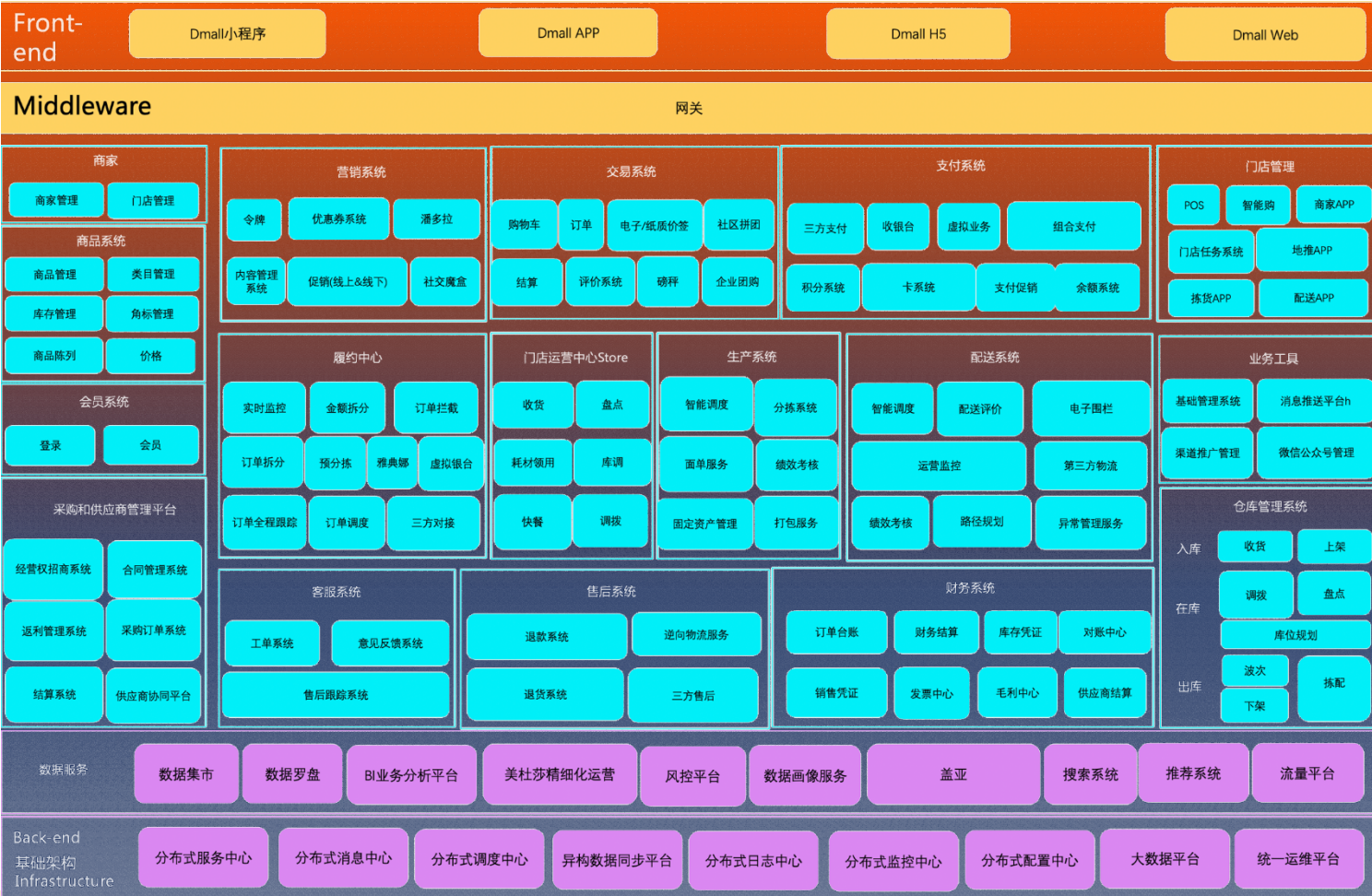
**Mini OS**

### 商 家

- 全面系统的数字化解决方案
- 到店到家一体化全渠道模式
- 降本增效、增加用户和收入

### 用 户

- 方便：到店到家全场景一体化
- 安全：专业供应链，品牌熟悉
- 丰富：满足个性化、一站购齐



-  30+省
-  120+连锁商超
-  13000+门店
-  1.6亿+用户
-  1800万+ MAU



## 单机房



鸡蛋放在1个篮子里

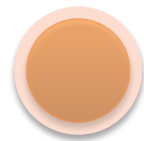


## 同城双活

混合云架构

+ 机房级故障

\* 难度适中



## 异地多活

单元化架构

+ 故障域隔离

\* 难度大，成本高



# 同城双活

混合云架构



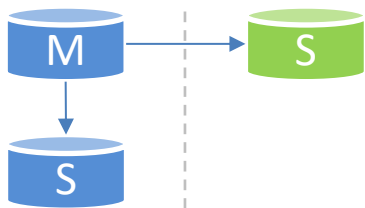


# 同城双活 - 数据库架构选型



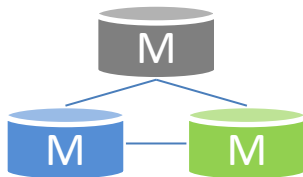
网络延迟 ~3ms

**目标：**单个机房不可用，其他机房可快速承接流量，恢复核心业务



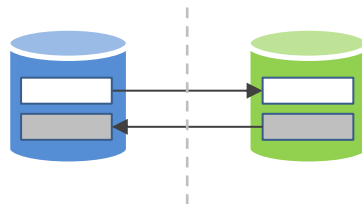
## 1. 原生复制/半同步

- 跨机房写延迟



## 2. MGR / PXC

- 要求3机房



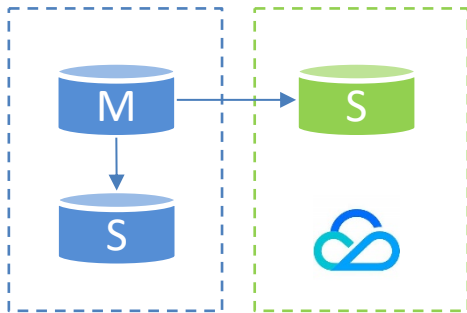
## 3. Sharding + 双向同步

- 同步中间件，架构/业务改造

# 同城双活 – 整体方案

## 同城三机房，Master/Slave + MGR，业务改造小

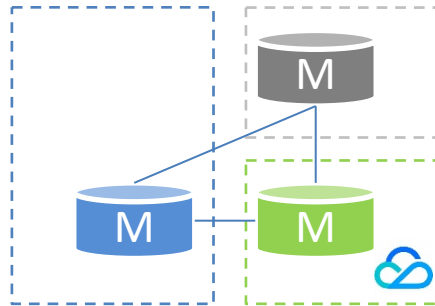
混合云



业务数据

读写性能、业务连续性

Master/Slave

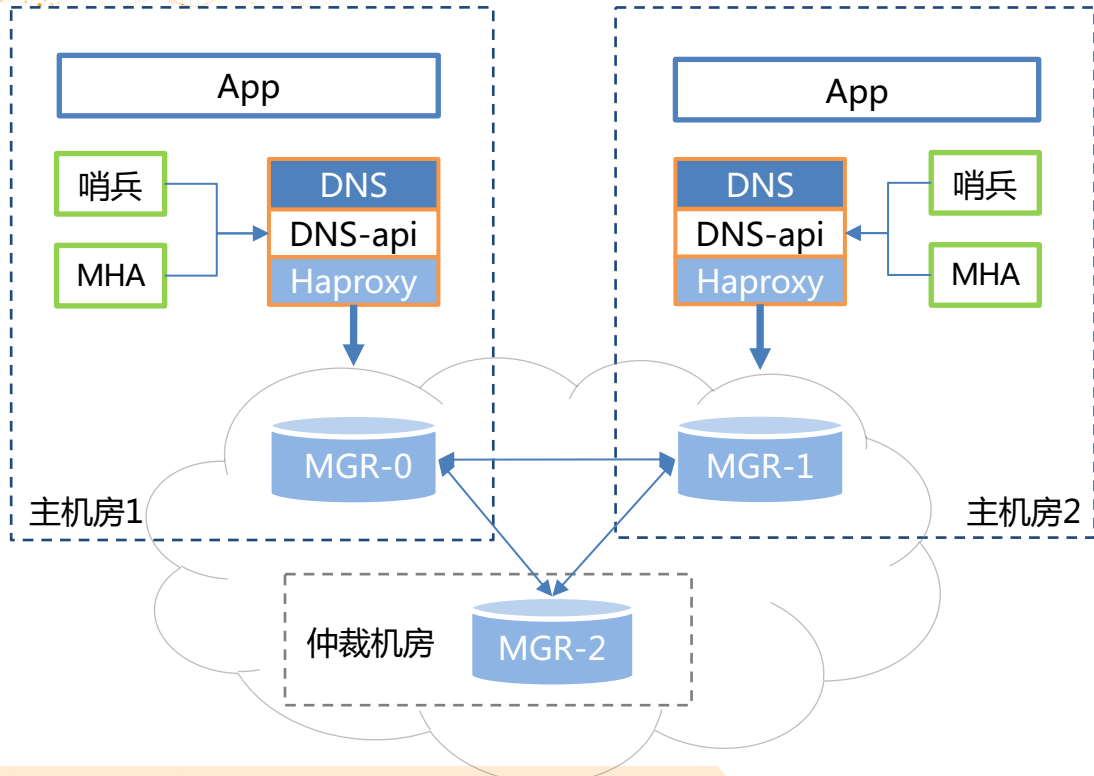


DNS、配置数据

强一致性

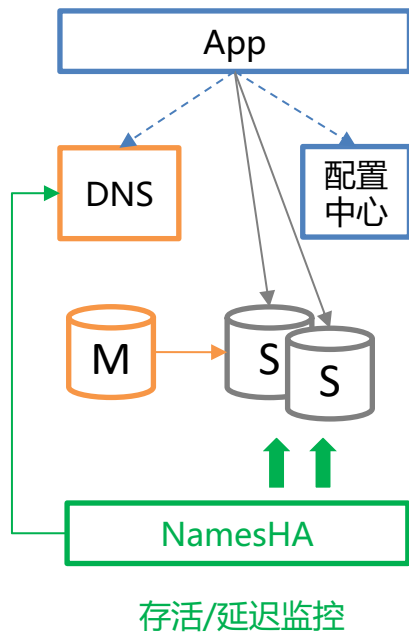
MGR (multi-master)

# 同城双活 – MGR-based DNS



- 数据强一致场景
- 2个主机房，1个仲裁机房
- MGR高可用：Haproxy
- Multi-Master模式，节点都可写
- MGR版本：8.0.18
  - 1) auto rejoin
  - 2) expel timeout
  - 3) OFFLINE\_MODE

# 同城双活 - 数据库接入层



读流量机房内闭环：配置中心 + 智能DNS

db.mysql.read-only.url:

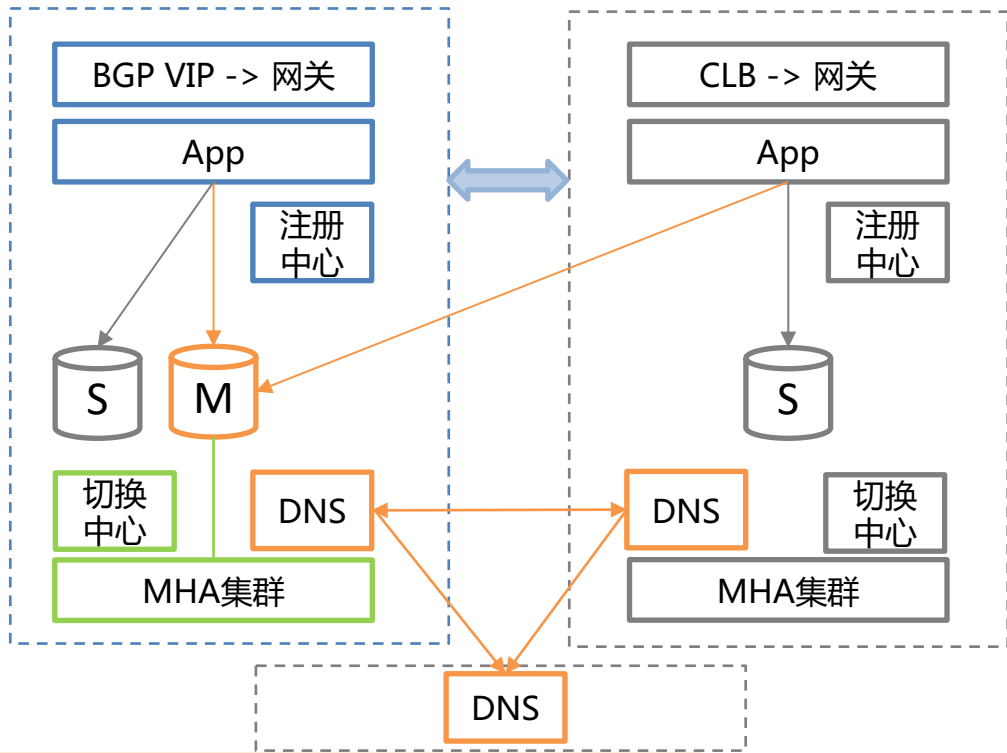
- dc1:  
- slave.3306.dc1.mysql.dmall.com:3306/xx\_db
- dc2:  
- slave.3306.dc2.mysql.dmall.com:3306/xx\_db

db.mysql.read-write.url:

- dc1:  
- master.3306.dc1.mysql.dmall.com:3306/xx\_db
- dc2:  
- master.3306.dc2.mysql.dmall.com:3306/xx\_db

# 同城双活 - MySQL高可用架构

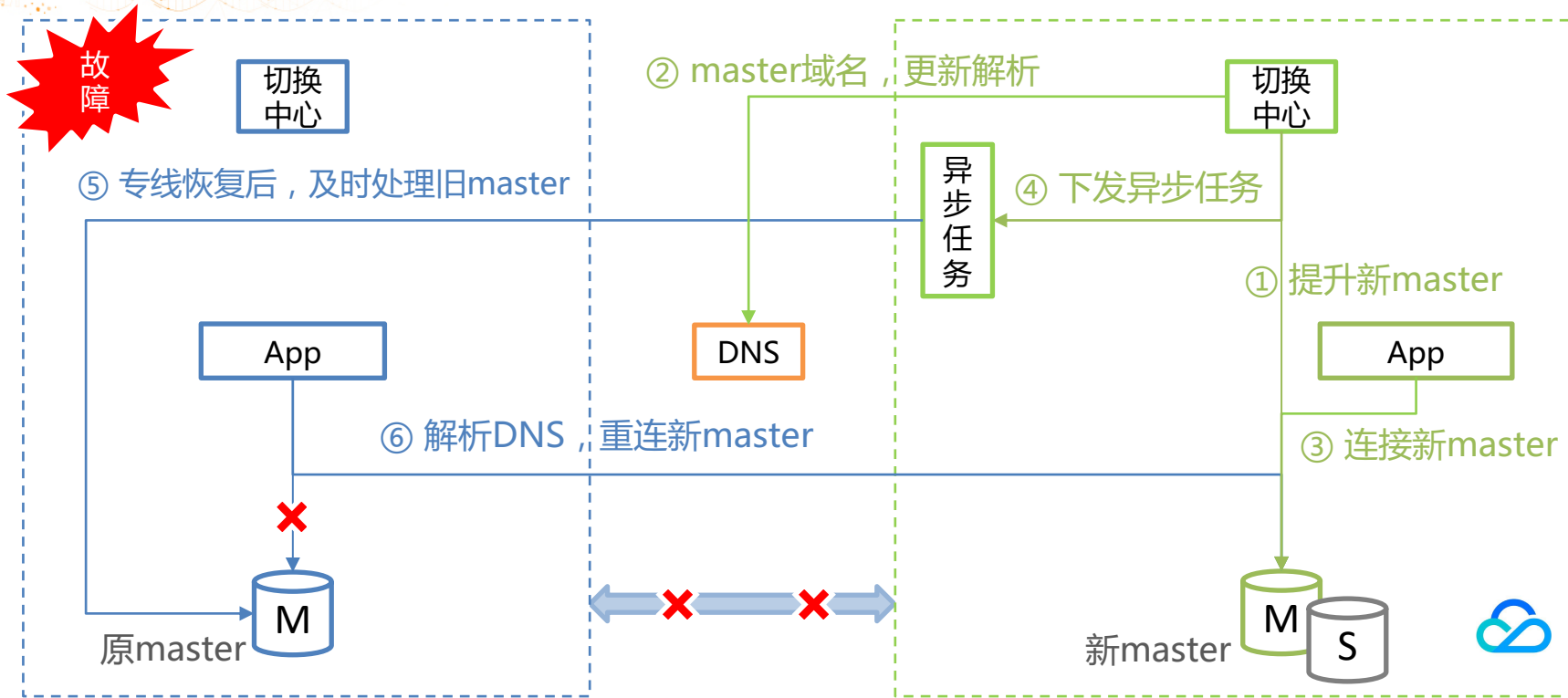
架构革新 ◎ 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020



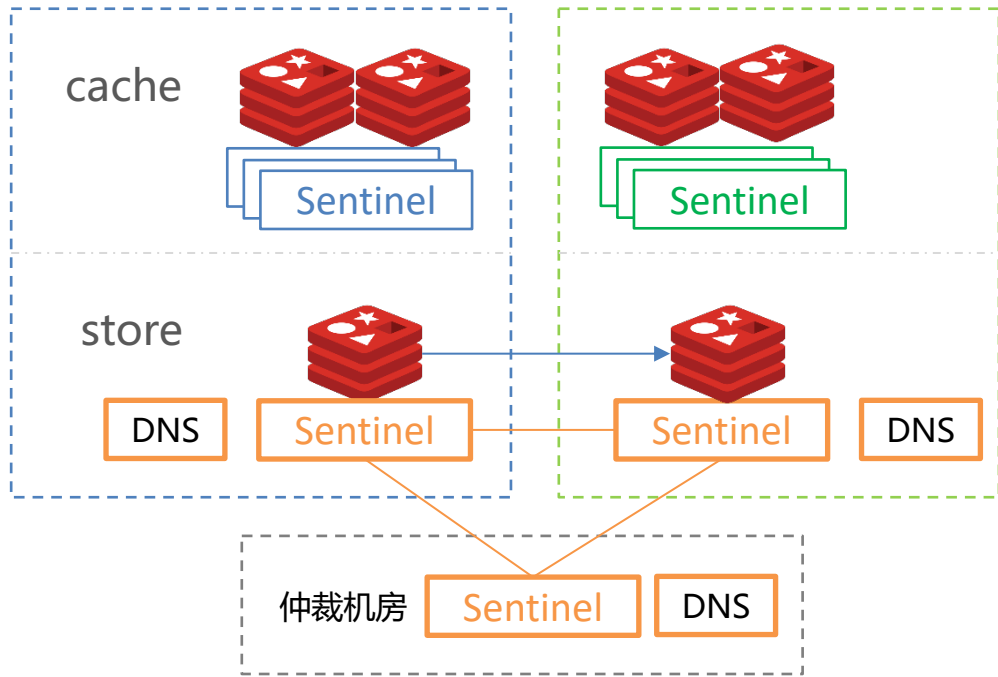
- 混合云架构，自建机房\*2 + 腾讯云
- 应用双活，RPC调用机房内闭环
- 数据库跨机房同步，读写分离
- MGR-based DNS ( multi-master )
- MHA双活部署，仅监控本机房master
- 切换中心，双活部署

# 同城双活 – MySQL容灾切换

架构革新 ◎ 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020



# 同城双活 – Redis高可用架构



- 2个主机房，1个仲裁机房
- Cache集群：机房内闭环，低延迟高性能，本机房哨兵监控
- Store集群：跨机房部署，可接受同城专线延迟，三机房哨兵监控
- 自动切换



# 同城双活 – DB PaaS

- 元数据存储：基于MGR（3机房集群）
- HA组件：2个主机房，双活部署
  - MHA，Sentinel，NamesHA，切换中心，...
- 平台改造难度/工作量：中等
  - 多机房支持（比如：multi salt-master，auto\_increment）
  - 批量容灾切换

# 同城双活 - 数据库架构总结

架构革新 ◎ 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020

## 方案

**3机房**：2主机  
房+1仲裁机房

强一致场景：  
**MGR架构**

高性能场景：  
**原生复制**

## pros

业务改造小，  
不折腾研发

读流量机房内  
**闭环**

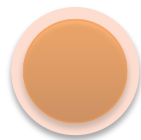
架构**简单**，成  
本可控

## cons

单机房写

跨机房延迟

非全自动切换



# 异地多活

单元化架构

# 单元化 - 整体架构方案

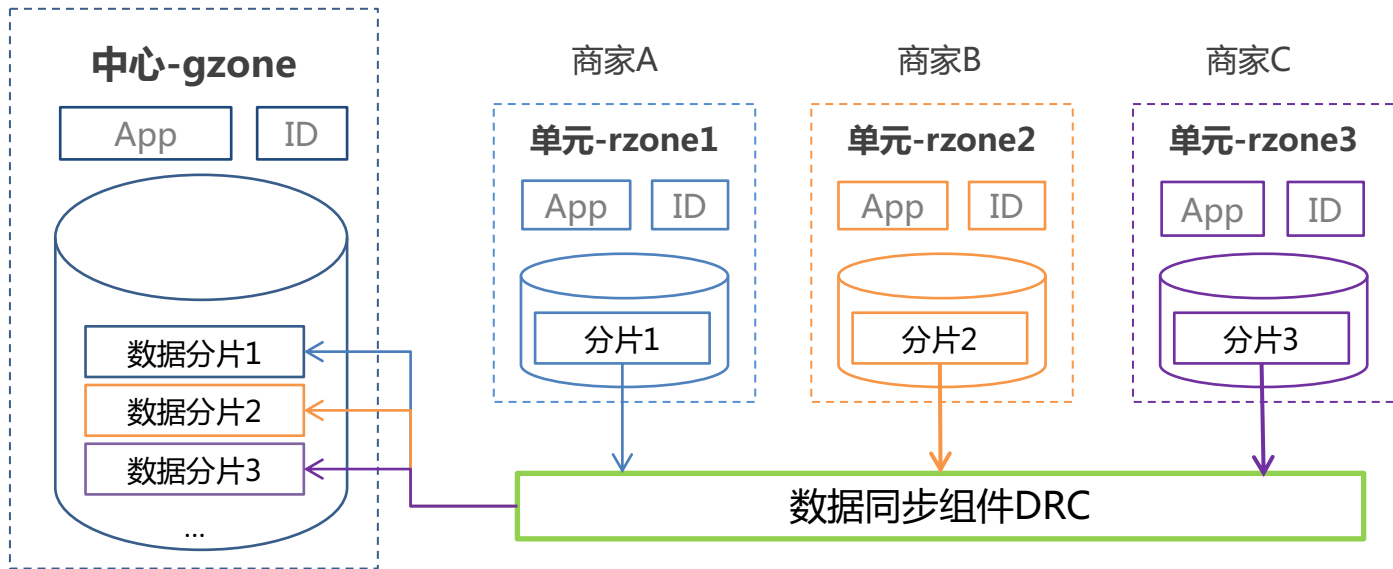
架构革新 ◎ 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020

HttpDNS + 接入网关

- 流量路由，基于商家ID

配置中心

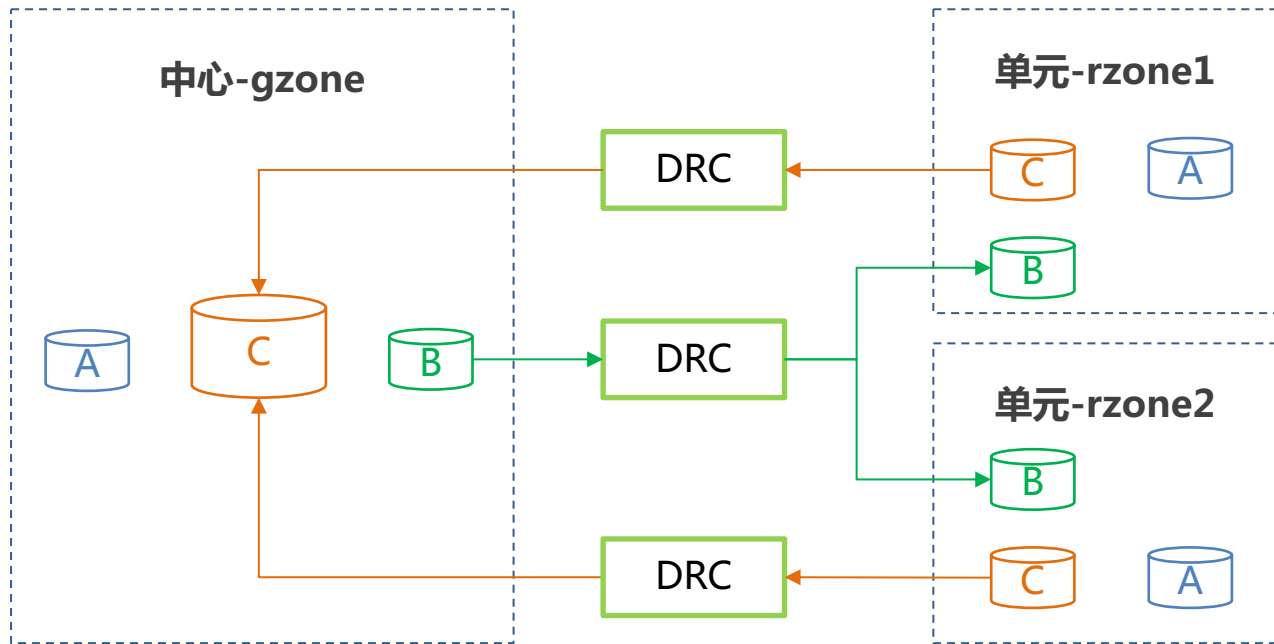
- 应用读写数据库单元内闭环



- 专线延迟大，**30~100ms**  
- 单元只包含**部分**数据  
- 单元/中心之间数据**同步**  
- **分布式ID**，**配置中心**

- 异步复制，**最终一致**

# 单元化 - 数据库架构



## 数据分类

- **A类**：zone内闭环，不同步
- **B类**：中心->单元，配置下发
- **C类**：单元->中心，数据汇总

## 同步原则

- 不做行级双向同步

# 单元化 - 分布式ID

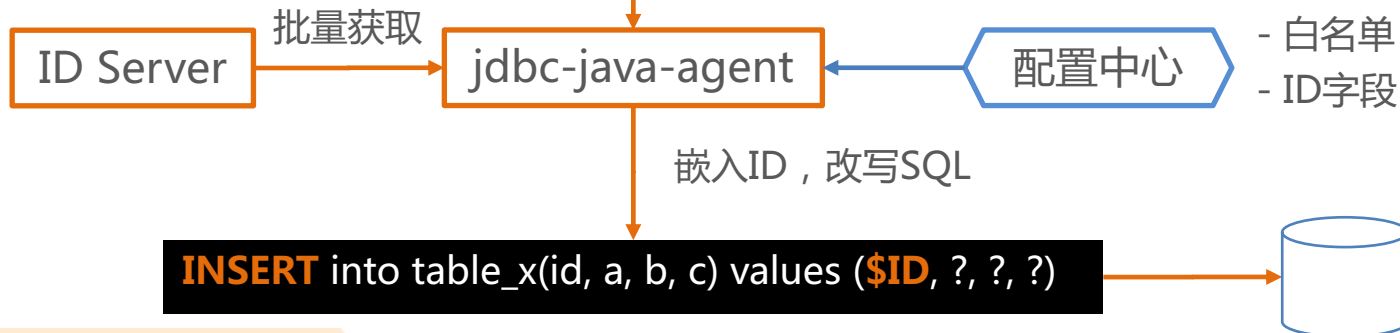
|   |                        |                              |                    |
|---|------------------------|------------------------------|--------------------|
| 0 | timestamp_ms (41 bits) | worker_id, zone_id (10 bits) | sequence (12 bits) |
|---|------------------------|------------------------------|--------------------|

各zone之间ID不重复

**INSERT** into table\_x(a, b, c) values (?, ?, ?)

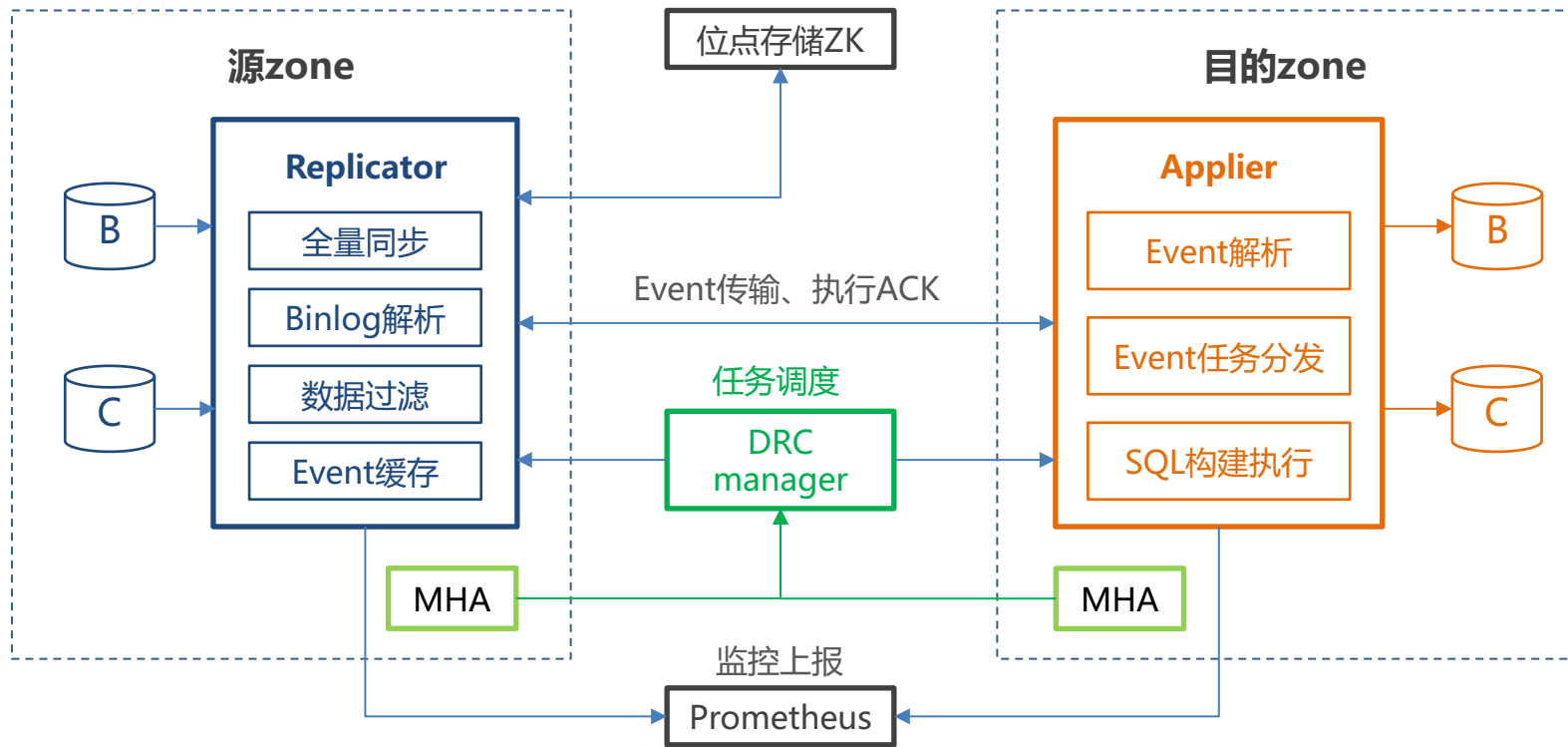
拦截INSERT

- 兼容自增ID
- 研发无需改代码



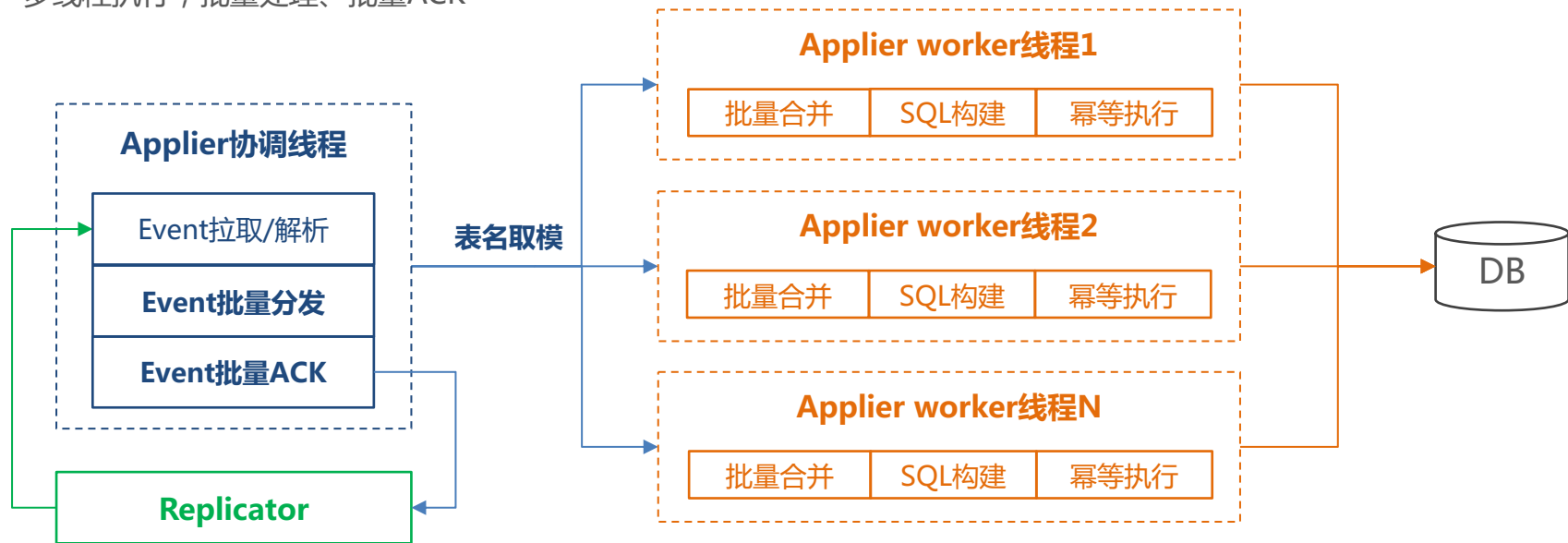
# 单元化 – MySQL同步组件

架构革新 ◎ 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020



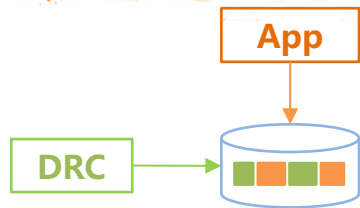
# 单元化 – MySQL同步性能优化

- Applier与目标DB同机房，Replicator与源DB同机房
- 发送端binlog过滤，压缩传输protobuf
- 多线程执行，批量处理、批量ACK





架构革新 © 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020



**MySQL参数：**  
**binlog\_rows\_query\_log\_events**

**SQL hint：**  
**/\*+ DMALL\_SQL\_HINT\_DRC \*/**

**Binlog Events序列：**

- BEGIN
- Rows\_query
- Table\_map
- Delete/Update/Write\_rows
- COMMIT

# 单元化 – MySQL同步binlog打标



```
BEGIN
/*!*/;
# at 1049
#201209 16:05:41 server id 1  end_log_pos 1128 CRC32 0x12f40744
# /*+ DMALL_SQL_HINT_DRC */ delete from test.t where id=7
# at 1128
#201209 16:05:41 server id 1  end_log_pos 1178 CRC32 0x16f2637e
# at 1178
#201209 16:05:41 server id 1  end_log_pos 1227 CRC32 0x25452d7f
```

Rows\_query

Table\_map: `test`.`t` mapped to number 78

Delete\_rows: table id 78 flags: STMT\_END\_F

```
BINLOG '
VYXQXx0BAAAATwAAAGgEAACAADcvKisgRE1BTExfU1FMX0hJTLrFRFJDICovIGRlbGV0ZSBmcm9t
IHRlc3QudCB3aGVyZSBpZD03RAf0Eg==
VYXQXxMBAAAAMgAAAJoEAAAAAE4AAAAAAEABHRLc3QAAXQAawMSAwEAAwEBAH5j8hY=
VYXQXyABAAAAMQAAAMsEAAAAAE4AAAAAAEAAgAD/wBcAAAAmafLCVkhAAAAfy1FJQ==
/*!*/;
### DELETE FROM `test`.`t`
### WHERE
### @1=92 /* INT meta=0 nullable=1 is_null=0 */
### @2='2020-11-18 16:37:25' /* DATETIME(0) meta=0 nullable=1 is_null=0 */
### @3=7 /* INT meta=0 nullable=0 is_null=0 */
# at 1227
#201209 16:05:41 server id 1  end_log_pos 1258 CRC32 0x26c17653
COMMIT/*!*/;
```

Xid = 35

DTCC  
2020



北京国际会议中心

2020/12/21-12/23

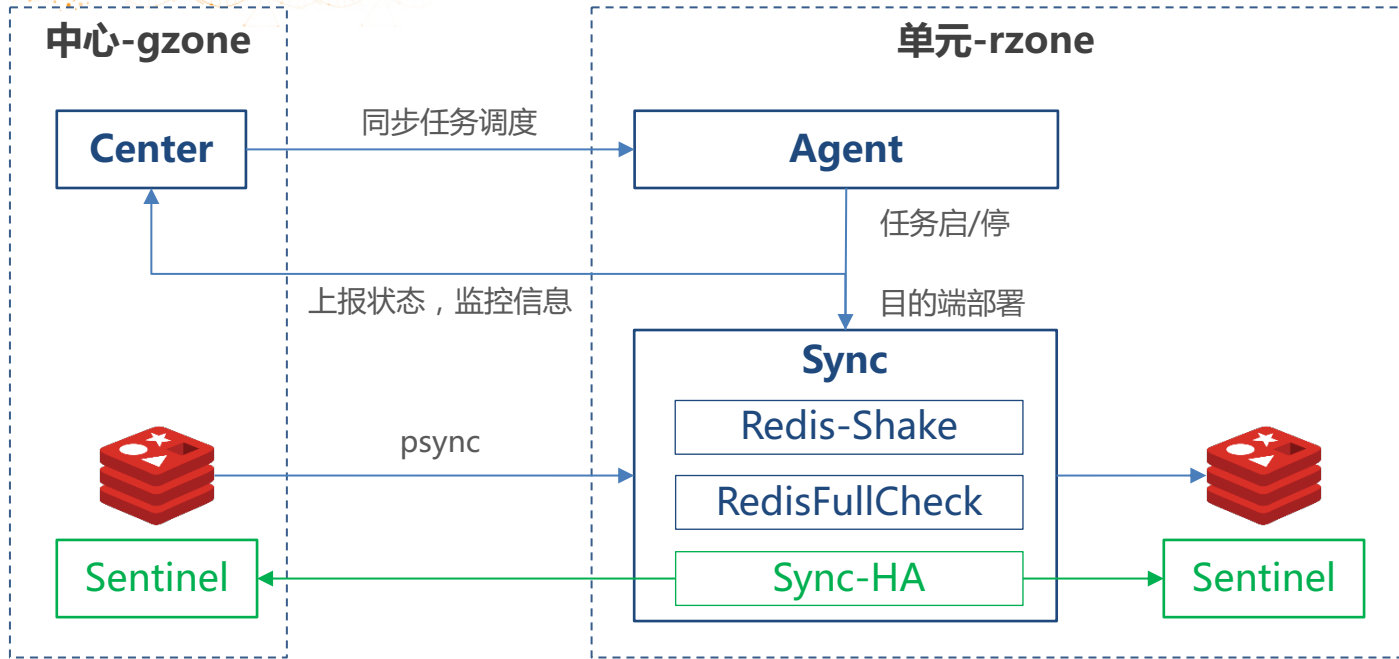


ChinaUnix

ITPUB

# 单元化 – Redis数据同步组件

架构革新 ◎ 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020



## 增强特性

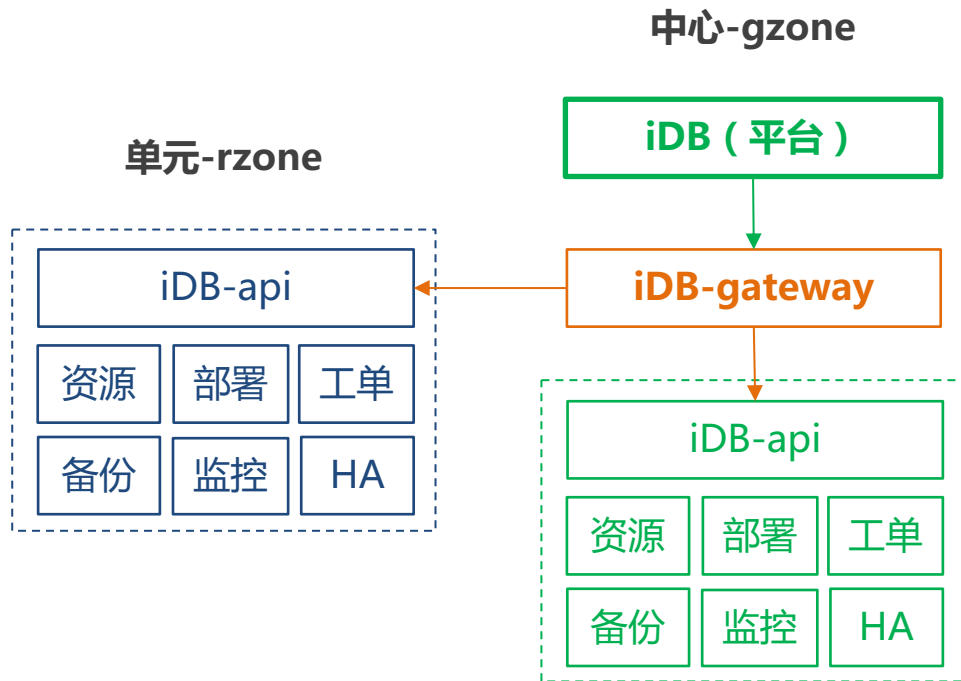
- 网络波动，断点续传
- 同步进程高可用调度
- 源/目的Redis faiover，Sentinel联动

- `SUBSCRIBE __sentinel__:hello`，更新哨兵集群地址、master-name
- 同步故障时，根据master-name从哨兵获取Redis集群新拓扑结构

# 单元化 – DB PaaS

## 平台改造工作量：大

- 各zone独立部署自动化平台组件
- 新增gateway层，实现api聚合
- DDL工单，需联动DRC，关注执行顺序



# 单元化 - 数据库架构总结

架构革新 ◎ 高效可控  
第十一届中国数据库技术大会  
DATABASE TECHNOLOGY CONFERENCE CHINA 2020

## 方案

流量拆分、数  
据分类

**DRC同步组件**

分布式ID、配  
置中心

## pros

数据域隔离

故障域隔离

弹性架构

## cons

基础组件开发  
难度大

理解/实施/运维  
成本高

折腾研发，业  
务改造量大，  
周期长



# Q & A



# THANKS

