GOPS2017 Shanghai

GOPS 全球运维大会 2017 上海站

指导单位：数据中心联盟 Data Center Alliance

主办单位：高效运维社区 GreatOPS Community　开放运维联盟 OOPSA Open OPS Alliance

大会时间：2017年11月17日-18日

大会地点：上海光大会展中心国际大酒店（上海徐汇区漕宝路67号）

GOPS2017
Shanghai

饿了么异地双活数据库实战

虢国飞 DBA负责人

目录

多活难点-时延


```
Branch mispredict ............................ 5 ns
L2 cache reference ........................... 7 ns
Mutex lock/unlock ........................... 25 ns
Main memory reference ...................... 100 ns
Compress 1K bytes with Zippy ............. 3,000 ns  =   3 µs
Send 2K bytes over 1 Gbps network ....... 20,000 ns  =  20 µs
SSD random read ........................ 150,000 ns  = 150 µs
Read 1 MB sequentially from memory ..... 250,000 ns  = 250 µs
Round trip within same datacenter ...... 500,000 ns  = 0.5 ms
Read 1 MB sequentially from SSD* ..... 1,000,000 ns  =   1 ms
Disk seek ........................... 10,000,000 ns  =  10 ms
Read 1 MB sequentially from disk .... 20,000,000 ns  =  20 ms
北京到上海的网络延时..................... 30,000,000 ns  =  30 ms
```

多活难点-异地

	同城多活	异地多活
整体投入	高（机房投入 + 同城专线）	很高（机房投入 + 异地专线）
实现复杂度	低（依赖垮机房调用）	高（需要减少机房间的交互，清理调用边界）
可以扩展到多机房	中（只能在同城增加机房）	高（可以在全国选择机房，甚至扩展到全球）
服务可用性	低（降低现有可用性）	高（可以应对机房级故障）
对现有架构的影响	低（跨机房调用）	高（业务需要改造）
对服务质量的影响	无影响	无影响

多活难点-数据

- 错乱
- 冲突
- 环路
- 一致性

多活难点

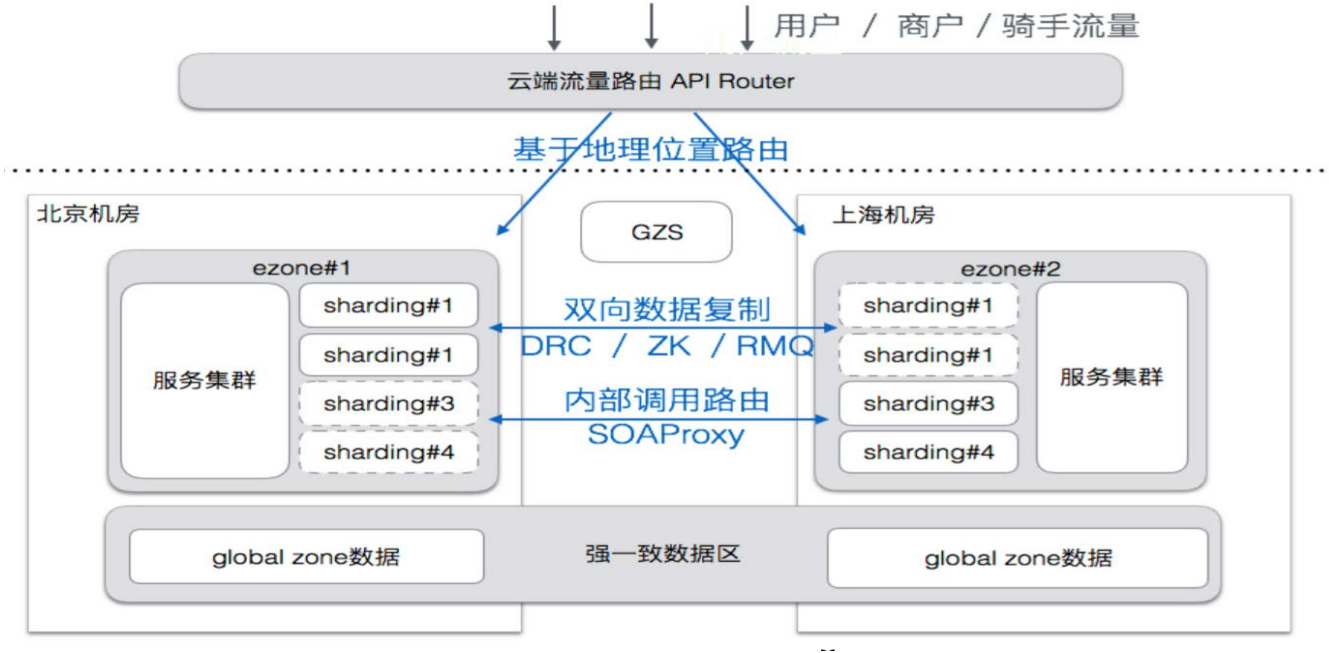
- 如何解决跨机房延时对业务影响（延时、抖动、断网）
- 如何分区访问流量，保障用户访问落到正确的机房
- 如何防止数据错乱，如何保障数据（最终）一致性

多活难点

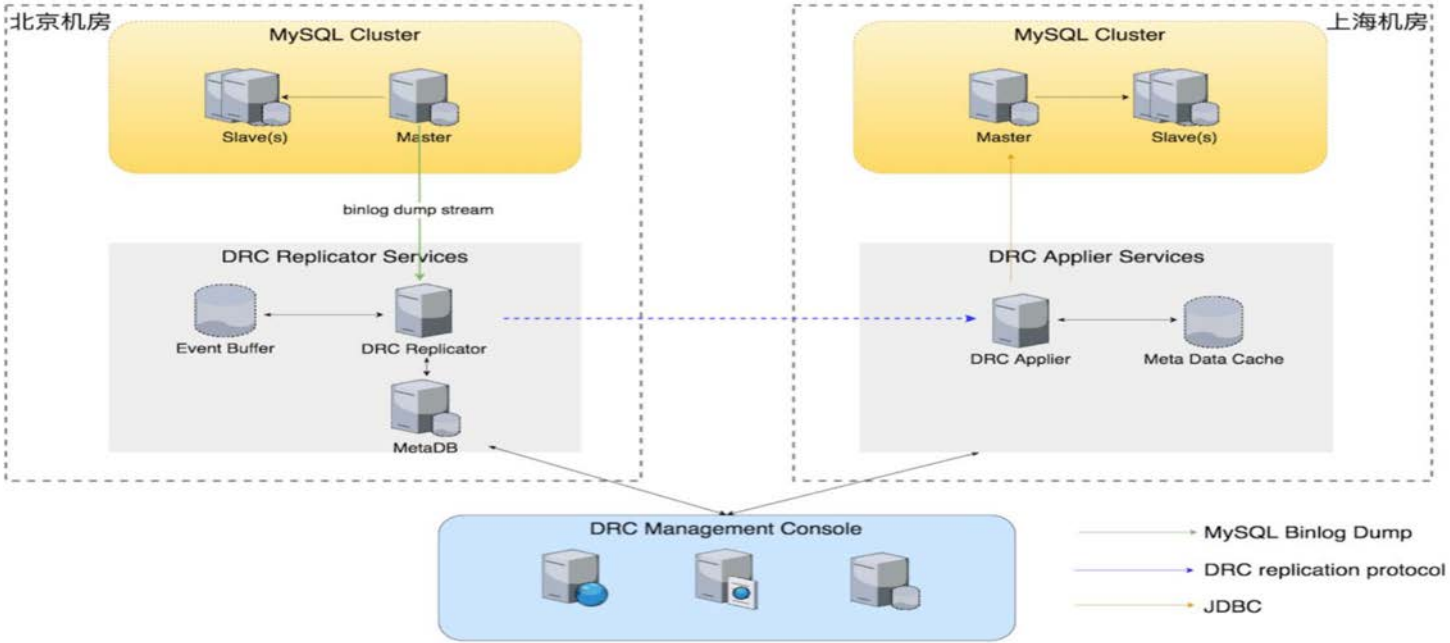
- 服务划分：类聚、划分、围栏（POI）
- 路由控制：SKey、APIRouter、SOA（内部调用）
- 脏写预防：冲突改造、SOA-Route、DAL-Reject
- 数据一致：DRC冲突、自增控制、数据校验

目录

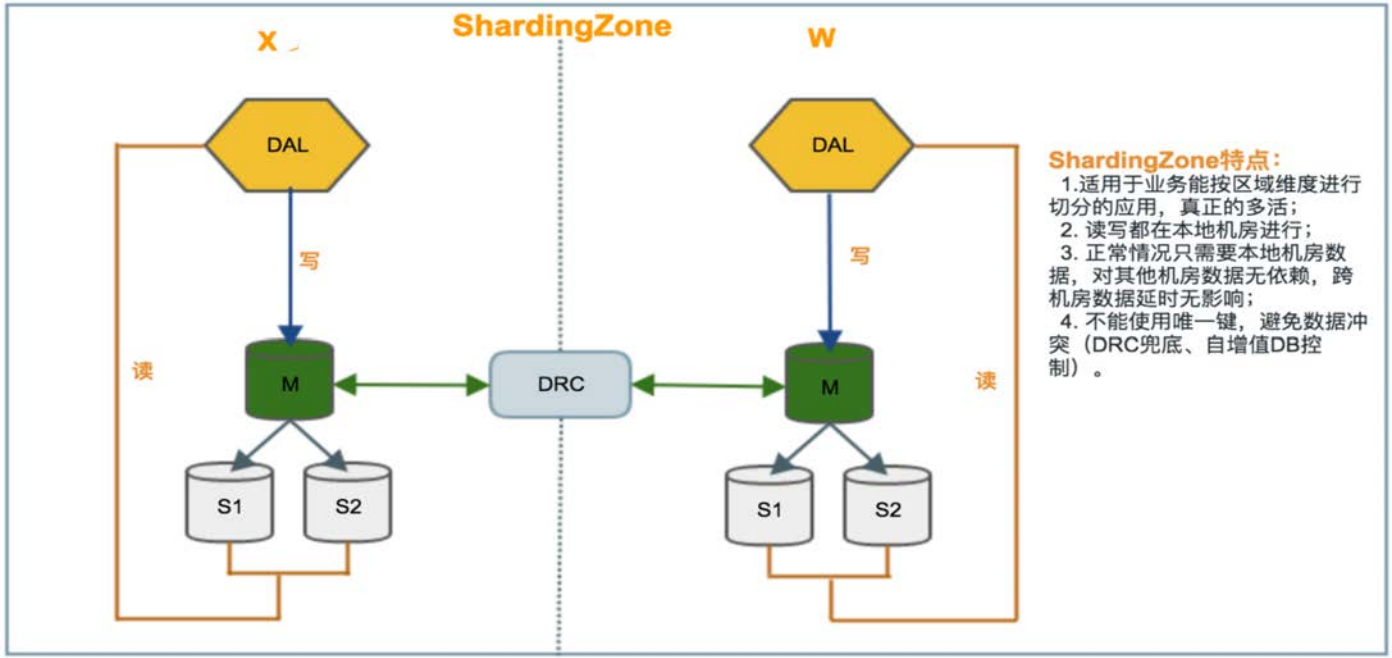
多活架构-Overview



多活架构-DRC


北京机房
MySQL Cluster
Slave(s)
Master
binlog dump stream
DRC Replicator Services
Event Buffer
DRC Replicator
MetaDB
上海机房
MySQL Cluster
Master
Slave(s)
DRC Applier Services
DRC Applier
Meta Data Cache
DRC Management Console
MySQL Binlog Dump
DRC replication protocol
JDBC

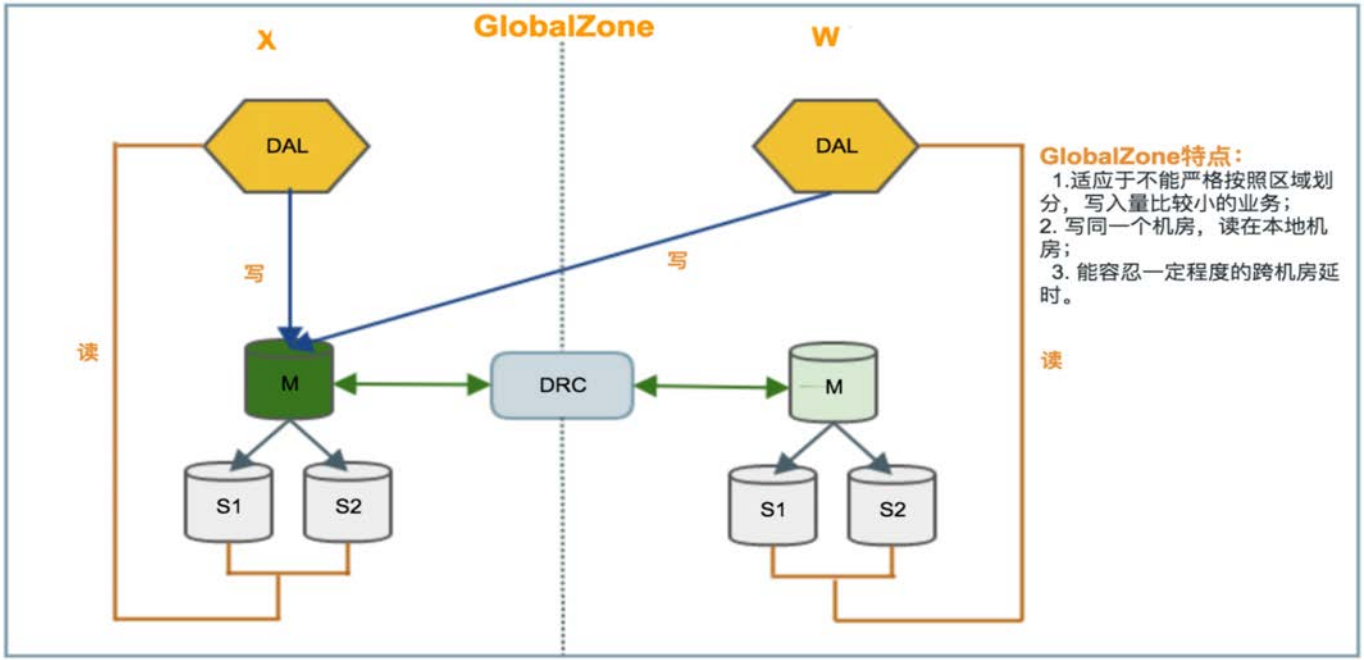
多活架构-DB



ShardingZone特点：

1.适用于业务能按区域维度进行切分的应用，真正的多活；
2. 读写都在本地机房进行；
3. 正常情况只需要本地机房数据，对其他机房数据无依赖，跨机房数据延时无影响；
4. 不能使用唯一键，避免数据冲突（DRC兜底、自增值DB控制）。

多活架构-DB


GlobalZone
X
W
DAL
DAL
写
写
读
读
M
DRC
M
S1
S2
S1
S2
GlobalZone特点：
1.适应于不能严格按照区域划分，写入量比较小的业务；
2. 写同一个机房，读在本地机房；
3. 能容忍一定程度的跨机房延时。

目录

数据库改造-项目

项目	改造原因	数据量	周期
全量数据导入	测试环境、生产环境数据全量同步	几百TB	两周
表增加DRC字段	表增加毫秒级别的时间戳，方便判断数据有效性	十几万表	五周
PK int改bigint	自增调整防止溢出	十几万表	
FK改bigint	Pk调整防止溢出	几万表	
业务分类迁移	不同类型业务要求放入不同集群	50+ DB	
自增调整	防止自增冲突，每个zone起始值错开	几百套	三周
原生改DRC	原生复制改成DRC复制，支持多写	几百套	
账号网段调整	原来账号限制在一个机房，现在需要支持多个机房	数千账号	
全量参数一致性	各个集群参数必须一致	几百套	
HA全量部署改造	按集群类型调整HA配置	几百套	

数据库改造-前后比较

GOPS2017
Shanghai

项目	改造前	改造后
实例	1200+	2000+
集群	400+	800+
Proxy	800+	1600+
HA	400+	800+
数据量	几百TB	翻倍
DDL	3位数/周	翻倍
DML	2位数/周	不变
机器故障	0.5台/周	2台/周
DBA	?	+2

目录

DBA挑战

- 数据
- HA
- 配置
- 容量
- DDL

DBA挑战-数据

- DAL-Reject
- DRC-冲突
- DCP-校验

DBA挑战-DCP

- 变动无需人工干预
- 全量、增量、延时校验、手动校验
- 黑白名单, 自定义规则
- 数据、结构、多维
- 延迟、并发、时长
- 修复SQL、配套工具

DBA挑战-DCP

- 每日几百套集群数据校验
- 日均校验数据60亿+
- 分钟级别校验频率
- 发现和修复数据一致性问题50+

DBA挑战-HA

- EMHA配置保持，切换同步，代理层（DAL）联动

DBA挑战-配置

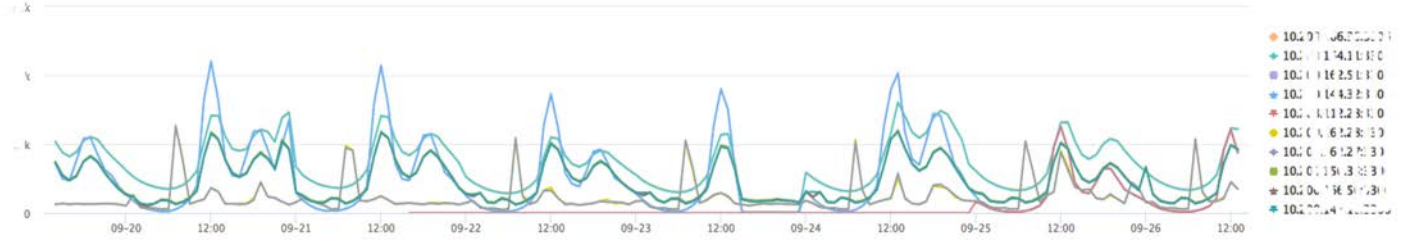
- DB配置一致，DAL配置一致

#	任务名称	任务服务器	任务脚本	任务计划
1	自增id溢出巡检	10.[illegible]7	table_r[illegible]flow_monitor.py	1 21 8 * * *
2	MHA每日巡检	10.[illegible]47	mha_[illegible]tion.py	1 0 10 * * *
3	报警汇总任务	10.[illegible]47	alarm[illegible]ary.py	0 */1 * * * *
4	基础数据收集	10.[illegible]47	baseC[illegible]ollection.py	1 0 0 * * *
5	数据库自动发布	10.0[illegible]47	db_re[illegible]e.py	0 05 21 * * 1,2,4
6	[illegible]	10.[illegible]47	dt_pa[illegible]n_manager.py	0 0 2 * * *
7	[illegible]	10.[illegible]80.53	db_b[illegible]p_wg.sh	1 */1 * * * *
8	[illegible]	10.[illegible]5	table_[illegible]collection.py	0 */5 * * * *
9	[illegible]	10.[illegible]7	db_rel[illegible]e.py	0 1 2 * * 0,1,4,6
10	[illegible]天	10.0[illegible]76	capaci[illegible]op_avg_day.py	0 0 2 * * *
11	Project参数一致性检测	10.0[illegible].127	check_[illegible]j_param_cons.py	0 10 10 * * *
12	MySQL配置巡检	10.0[illegible].127	check_[illegible]ql_param.py	0 20 9 * * *
13	DAL配置巡检	10.0[illegible]76	dalcor[illegible]heck.py	0 45 9 * * *
14	[illegible]	10.0[illegible].127	analyz[illegible]i_ba.py	0 0 6 * * *
15	DRC表字段巡检	10.0.[illegible].127	check_[illegible]_need_col.py	0 40 10 * * *
16	SlowSQL每天统计	10.0.[illegible]76	slow_q[illegible]ount_day.py	0 10 7 * * *
17	[illegible]	10.0.[illegible]47	pang[illegible]_repair.py	1 0 3 * * *
18	[illegible]	10.0.[illegible]76	capa[illegible]p_avg_hour.py	0 0 * * * *
19	更新DRC	10.0.[illegible]76	drcC[illegible]	0 */5 * * * *
20	dal配置更新	10.0.[illegible]76	reloa[illegible]Conf	0 30 3,16 * * *
21	集群节点拓扑更新	10.0.[illegible]76	reloa[illegible]ster	0 12 * * * *
22	dal访问流量信息	10.0.[illegible]	dalAc[illegible]	0 0 4 * * *

type	checkList	Count
生产环境所有机房dalgroup	生产环境所有机房dalgroup	35
生产/非生产slave混用	配置生产slave的抽数dalgroup	
生产/非生产slave混用	配置生产slave的dbadmin_dalgroup	
生产/非生产slave混用	配置dtslave的生产dbgroup	
单点配置	配置单点master或单点slave的生产dbgroup	4
单点配置	配置1个master和N个slave的生产dbgroup	
单点配置	配置N个master和1个slave的生产dbgroup	
单点配置	配置1个master和1个slave的生产dbgroup	
不标准配置	配置0个master和N个slave的生产dbgroup	8
不标准配置	配置N个master和0个slave的生产dbgroup	3
不标准配置	配置slave的推数dalgroup	

DBA挑战-容量

- 多集群、多机房、不同流量

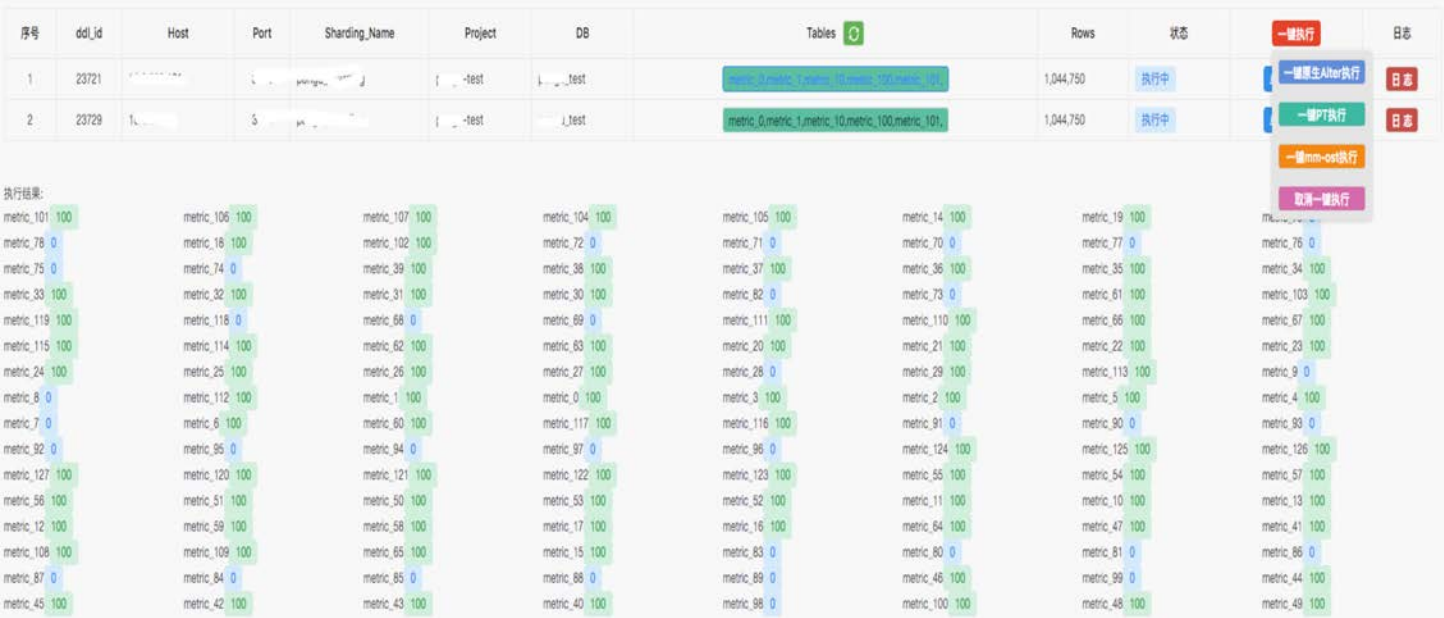


Project

#	Host	Port	Project	Role	DBA	QPS	TPS	Proc	CPU	Net	Delay	IOPS	Disk	更新时间
1	10.[illegible]	3[illegible]	B[illegible]der	slave	[illegible]	3071.79	2420.02	4.38	7.89	32.01	7.42	6619.3	58.88%	09-26 02:00
2	10.[illegible]	3[illegible]	B[illegible]der	slave	[illegible]	2964.45	2418.31	4.39	7.5	31.93	7.34	6660.14	61.78%	09-26 02:00
3	10.[illegible]	3[illegible]	E[illegible]SO	slave	[illegible]	1762.89	967.19	2.58	5.8	22.42	1	6638.7	66.57%	09-26 02:00
4	10.[illegible]	3[illegible]	E[illegible]SO	slave	[illegible]	1752.02	970.42	2.68	5.77	22.39	1	6795.41	74.17%	09-26 02:00
5	10.[illegible]	3[illegible]	E[illegible]SO	slave	[illegible]	1733.4	961.51	2.62	5.66	22.36	1	6687.16	64.05%	09-26 02:00
6	10.[illegible]	3[illegible]	Ele[illegible]arch	slave	[illegible]	1575.95	5105.98	2.19	5.89	38.11	75.55	27789.52	52.67%	09-26 02:00
7	10.[illegible]	3[illegible]	El[illegible]arch	slave	[illegible]	1535.09	5014.78	2.27	5.93	37.38	75.8	27700.32	52.67%	09-26 02:00
8	10.[illegible]	3[illegible]	L[illegible]aku	slave	[illegible]	1216.88	702.14	4.17	7.87	41.92	1.3	5985.29	15.15%	09-26 02:00
9	10.[illegible]	3[illegible]	l[illegible]aku	slave	[illegible]	1094.85	701.43	4	7.57	41.97	1.26	5889.99	15.23%	09-26 02:00
10	10.[illegible]	3[illegible]	L[illegible]aku	slave	[illegible]	1067.3	706.3	4	7.71	41.66	1.29	6051.8	15.21%	09-26 02:00

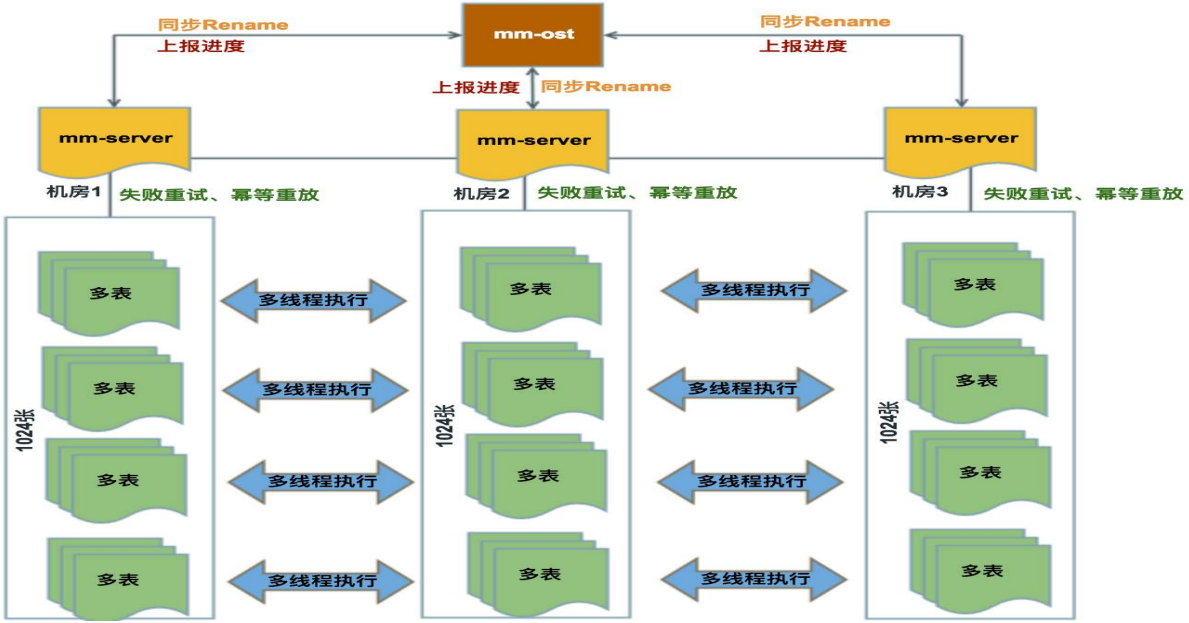
DBA挑战-DDL

- 类型：多活、非多活 、GlobalZone、多推、Sharding



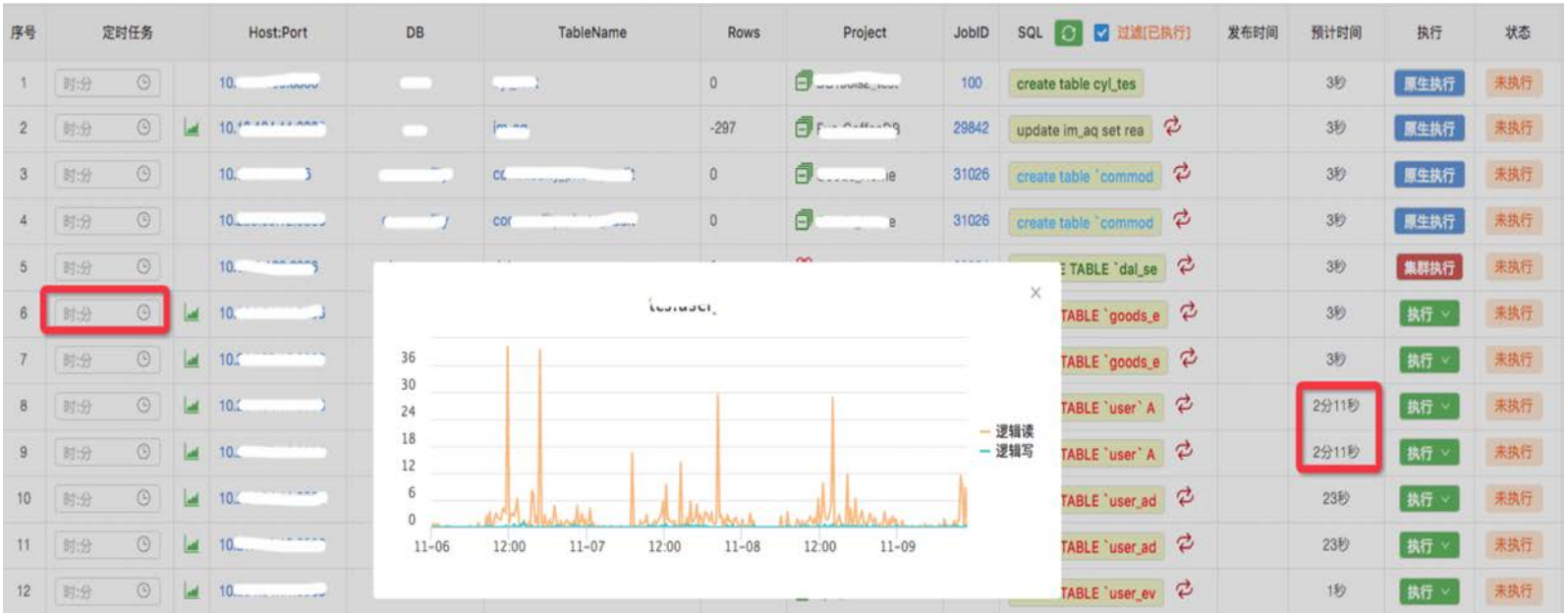
DBA挑战-DDL

- 延时：多机房延时控制3-5s



DBA挑战-DDL

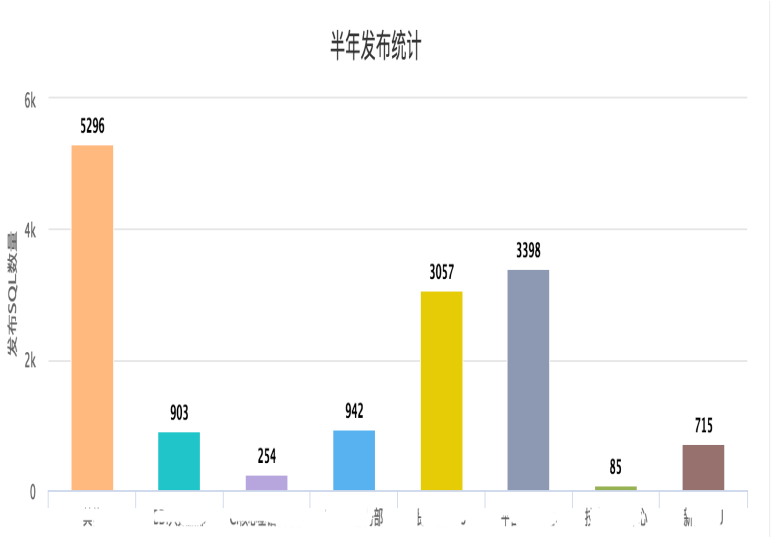
- 控制：空间&延时&锁&定时&低峰&风险&时长



DBA挑战-DDL

数量：

- 多活工单：4位数/周
- DDL表：4~5位数/周
- 自动/人工：8：2



目录

收益

- 打破单机房（地域）容量瓶颈
- 不受单机房（地域）故障影响
- 故障兜底
- 动态调整各机房流量

展望

- 多个机房
- Data-Sharding
- 自动动态扩缩容
- 多机房强一致

Q&A

GOPS2017
Shanghai

Thanks

高效运维社区

开放运维联盟

荣誉出品

想第一时间看到
高效运维社区公众号
的好文章吗？

请打开高效运维社区公众号，点击右上角小人，如右侧所示设置就好