



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

魏国飞 DBA负责人



GOPS2017  
Shanghai

# 目录

- ➔ 1 多活难点
- 2 多活架构
- 3 数据库改造
- 4 DBA挑战
- 5 收益与展望

# 多活难点-时延



GOPS2017  
Shanghai

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 $\mu$ s
Send 2K bytes over 1 Gbps network .....	20,000 ns	= 20 $\mu$ s
SSD random read .....	150,000 ns	= 150 $\mu$ s
Read 1 MB sequentially from memory .....	250,000 ns	= 250 $\mu$ s
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



GOPS2017  
Shanghai

	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



GOPS2017  
Shanghai



GOPS2017  
Shanghai

# 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



GOPS2017  
Shanghai

# 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





GOPS2017  
Shanghai

# 目录

1 多活难点

➔ 2 多活架构

3 数据库改造

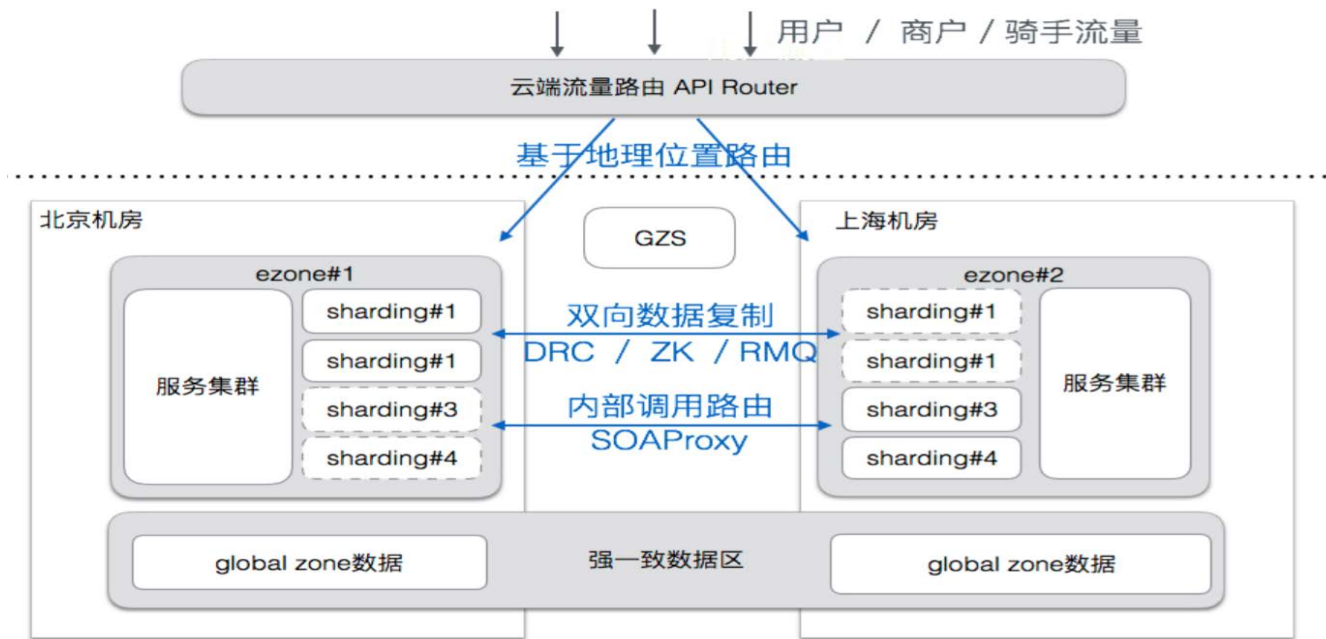
4 DBA挑战

5 收益与展望

# 多活架构-Overview



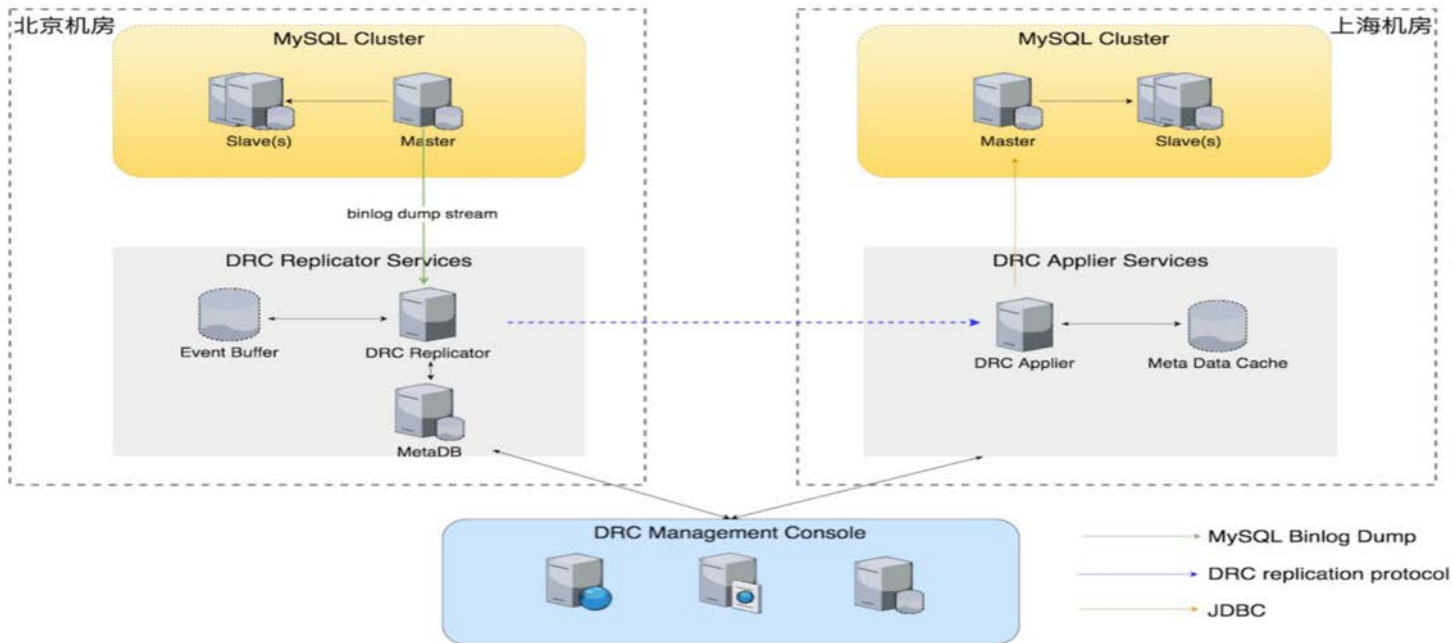
GOPS2017  
Shanghai



# 多活架构-DRC



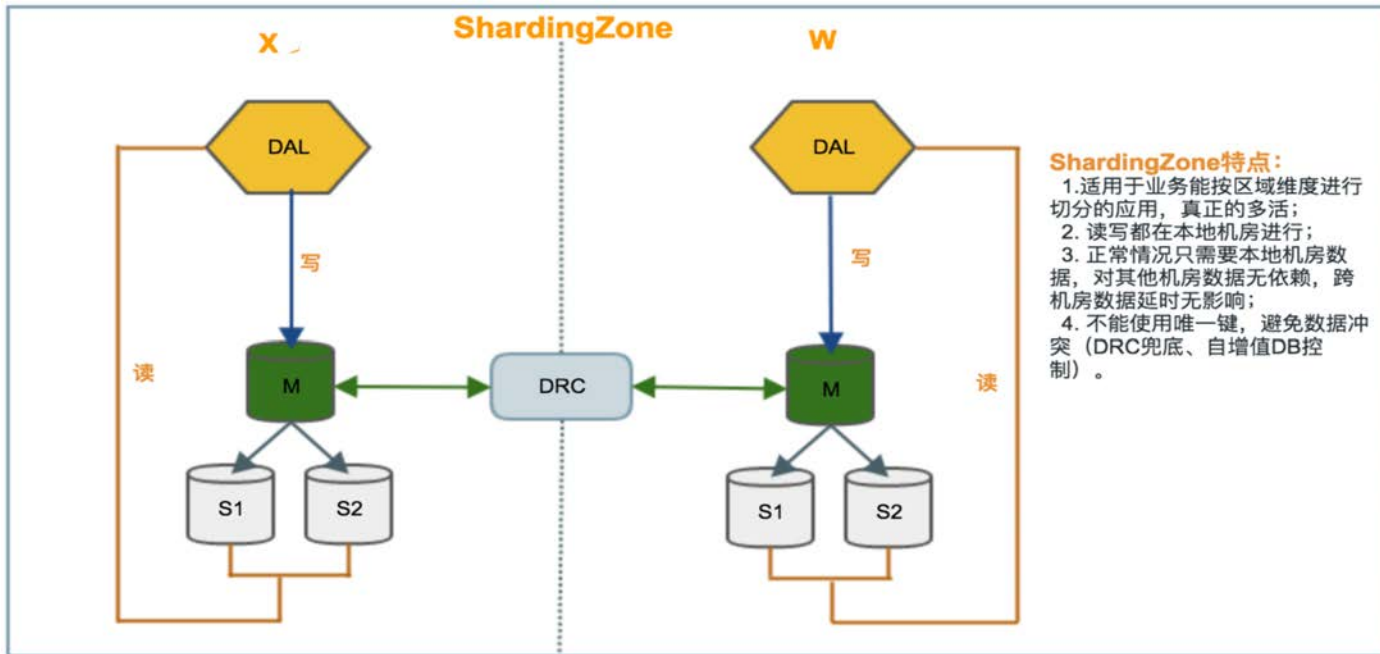
GOPS2017  
Shanghai



# 多活架构-DB



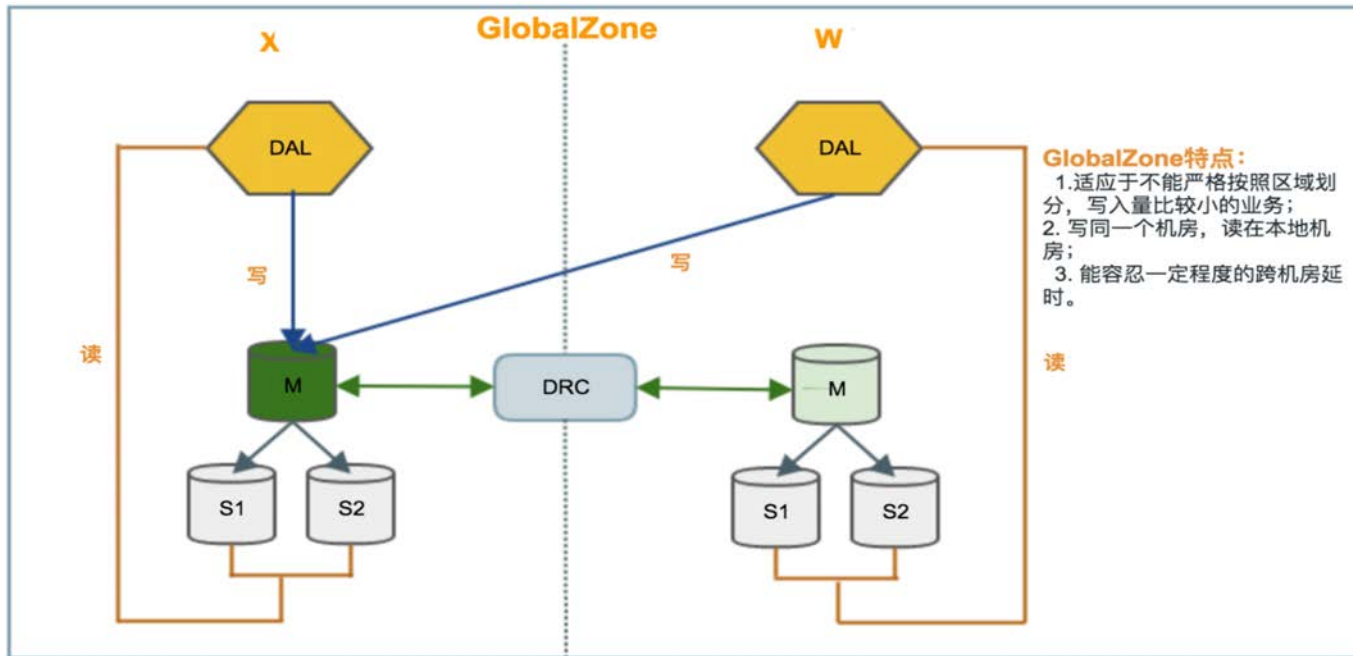
GOPS2017  
Shanghai



# 多活架构-DB



GOPS2017  
Shanghai





GOPS2017  
Shanghai

# 目录

1 多活难点

2 多活架构

➔ 3 数据库改造

4 DBA挑战

5 收益与展望

# 数据库改造-项目



GOPS2017  
Shanghai

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





GOPS2017  
Shanghai

# 目录

1 多活难点

2 多活架构

3 数据库改造

➔ 4 DBA挑战

5 收益与展望

# DBA挑战



GOPS2017  
Shanghai

- 数据
- HA
- 配置
- 容量
- DDL

# DBA挑战-数据



GOPS2017  
Shanghai

- DAL-Reject
- DRC-冲突
- DCP-校验





GOPS2017  
Shanghai

# 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



GOPS2017  
Shanghai

# 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	10.1.1.47	table_r_flow_monitor.py	1 21 6 ***
2	MHA每日巡检	10.1.1.47	mha_check_action.py	1 0 10 ***
3	报警汇总任务	10.1.1.47	alarm_summary.py	0 */1 ***
4	基础数据收集	10.1.1.47	base_collect.py	1 0 0 ***
5	数据库自动发布	10.0.1.47	db_release.py	0 05 21 *** 1,2,4
6	配置管理	10.1.1.47	db_config_manager.py	0 0 2 ***
7	配置管理	10.1.1.80,53	db_config_wg.sh	0 */1 ***
8	配置管理	10.1.1.13	table_collect.py	0 */6 ***
9	配置管理	10.0.1.47	db_release.py	0 1 2 *** 0,1,4,6
10	配置管理	10.0.1.76	capacity_avg_day.py	0 0 2 ***
11	Project参数一致性检测	10.0.1.127	check_project_param_cons.py	0 10 10 ***
12	MySQL配置巡检	10.0.1.127	check_mysql_param.py	0 20 9 ***
13	DAL配置巡检	10.0.1.76	dalconfig_check.py	0 45 9 ***
14	DAL配置巡检	10.0.1.127	analyze_dal.py	0 0 6 ***
15	DRC表字段巡检	10.0.1.127	check_drc_need_col.py	0 40 10 ***
16	SlowSQL每天统计	10.0.1.76	slow_query_count_day.py	0 10 7 ***
17	集群节点修复	10.0.1.47	repair_node.py	1 0 3 ***
18	集群节点修复	10.0.1.76	capacity_avg_hour.py	0 0 ***
19	更新DRC	10.0.1.76	drc_refresh.py	0 */6 ***
20	dal配置更新	10.0.1.76	reload_dal_conf	0 30 3,16 ***
21	集群节点拓扑更新	10.0.1.76	reload_master	0 12 ***
22	dal访问流量信息	10.0.1.47	dal_access.py	0 0 4 ***

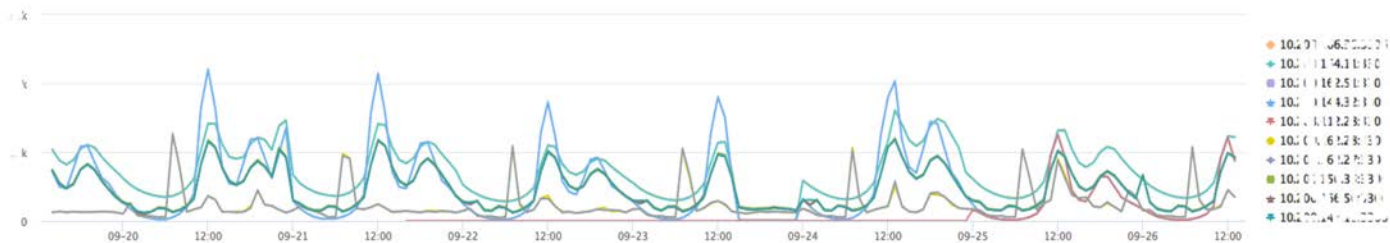
type	checkList	Count
生产环境所有机房dalgroup	生产环境所有机房dalgroup	5
生产/非生产slave混用	配置生产slave的抽数dalgroup	
生产/非生产slave混用	配置生产slave的dbadmin_dalgroup	
生产/非生产slave混用	配置dtslave的生产dbgroup	
单点配置	配置单点master或单点slave的生产dbgroup	4
单点配置	配置1个master和N个slave的生产dbgroup	
单点配置	配置N个master和1个slave的生产dbgroup	
单点配置	配置1个master和1个slave的生产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



GOPS2017  
Shanghai

- 多集群、多机房、不同流量



#	Host	Port	Project	Role	DBA	QPS	TPS	Proc	CPU	Net	Delay	IOPS	Disk	更新时间
1	10.10.1.1.0	30	B...der	slave	...	3071.79	2420.02	4.38	7.89	32.01	7.42	6619.3	58.88%	09-26 02:00
2	10.10.1.1.1	30	B...der	slave	...	2964.45	2418.31	4.39	7.5	31.93	7.34	6660.14	61.78%	09-26 02:00
3	10.10.1.1.1	30	E...ISO	slave	...	1702.89	967.19	2.58	5.6	22.42	1	6638.7	66.57%	09-26 02:00
4	10.10.1.1.2	30	E...ISO	slave	...	1752.02	970.42	2.68	5.77	22.39	1	6795.41	74.17%	09-26 02:00
5	10.10.1.1.3	30	E...ISO	slave	...	1733.4	961.51	2.62	5.66	22.36	1	6687.16	64.05%	09-26 02:00
6	10.10.1.2.1	30	Ele...arch	slave	...	575.95	5105.98	2.19	5.89	38.11	75.55	27789.52	52.67%	09-26 02:00
7	10.10.1.2.2	30	Ele...arch	slave	...	535.09	5014.78	2.27	5.93	37.38	75.6	27700.32	52.67%	09-26 02:00
8	10.10.1.3	30	L...aku	slave	...	216.88	702.14	4.17	7.87	41.92	1.3	5985.29	15.15%	09-26 02:00
9	10.10.1.3	30	L...aku	slave	...	094.85	701.43	4	7.57	41.97	1.26	5889.99	15.23%	09-26 02:00
10	10.10.1.4.5	30	L...aku	slave	...	067.3	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

序号	ddl_id	Host	Port	Sharding_Name	Project	DB	Tables	Rows	状态	一键执行	日志
1	23721	192.168.1.101	3306	sharding_test	-test	_test	metric_0,metric_1,metric_10,metric_101	1,044,750	执行中	一键重生Alter执行	日志
2	23729	192.168.1.102	3306	sharding_test	-test	_test	metric_0,metric_1,metric_10,metric_101	1,044,750	执行中	一键PT执行	日志

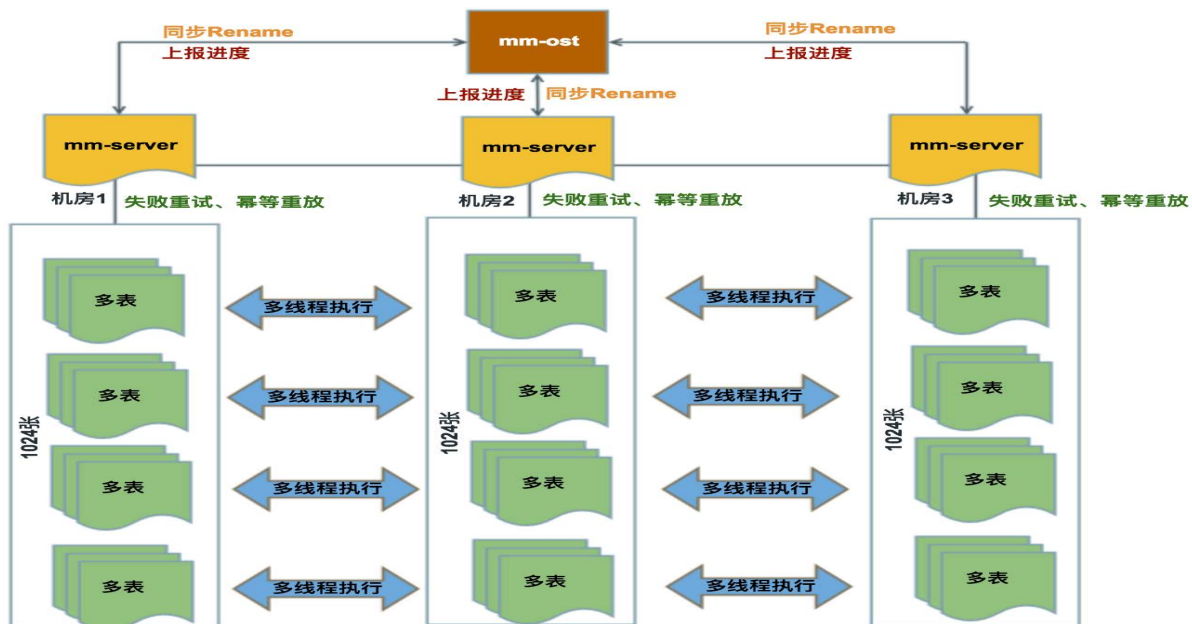
执行结果:

metric_101 100	metric_106 100	metric_107 100	metric_104 100	metric_105 100	metric_14 100	metric_19 100	metric_102 100
metric_78 0	metric_16 100	metric_102 100	metric_72 0	metric_71 0	metric_70 0	metric_77 0	metric_76 0
metric_75 0	metric_39 100	metric_39 100	metric_38 100	metric_37 100	metric_36 100	metric_35 100	metric_34 100
metric_33 100	metric_32 100	metric_31 100	metric_30 100	metric_82 0	metric_73 0	metric_61 100	metric_103 100
metric_119 100	metric_118 0	metric_68 0	metric_69 0	metric_111 100	metric_110 100	metric_66 100	metric_67 100
metric_115 100	metric_114 100	metric_62 100	metric_63 100	metric_20 100	metric_21 100	metric_22 100	metric_23 100
metric_24 100	metric_25 100	metric_26 100	metric_27 100	metric_28 0	metric_29 100	metric_113 100	metric_9 0
metric_8 0	metric_112 100	metric_1 100	metric_0 100	metric_3 100	metric_2 100	metric_5 100	metric_4 100
metric_7 0	metric_6 100	metric_60 100	metric_117 100	metric_116 100	metric_91 0	metric_90 0	metric_93 0
metric_92 0	metric_95 0	metric_94 0	metric_97 0	metric_96 0	metric_124 100	metric_125 100	metric_126 100
metric_127 100	metric_120 100	metric_121 100	metric_122 100	metric_123 100	metric_55 100	metric_54 100	metric_57 100
metric_56 100	metric_51 100	metric_50 100	metric_53 100	metric_52 100	metric_11 100	metric_10 100	metric_13 100
metric_12 100	metric_59 100	metric_58 100	metric_17 100	metric_16 100	metric_64 100	metric_47 100	metric_41 100
metric_108 100	metric_109 100	metric_65 100	metric_15 100	metric_63 0	metric_80 0	metric_81 0	metric_86 0
metric_87 0	metric_84 0	metric_85 0	metric_88 0	metric_89 0	metric_46 100	metric_99 0	metric_44 100
metric_45 100	metric_42 100	metric_43 100	metric_40 100	metric_98 0	metric_100 100	metric_48 100	metric_49 100



# DBA挑战-DDL

- 延时：多机房延时控制3-5s



# DBA挑战-DDL



GOPS2017  
Shanghai

- 控制：空间&延时&锁&定时&低峰&风险&时长

序号	定时任务	Host:Port	DB	TableName	Rows	Project	JobID	SQL	发布时间	预计时间	执行	状态
1	时:分	10.10.10.10			0		100	create table cyl_tes		3秒	原生执行	未执行
2	时:分	10.10.10.10			-297		29842	update im_aq set rea		3秒	原生执行	未执行
3	时:分	10.10.10.10			0		31026	create table `commod		3秒	原生执行	未执行
4	时:分	10.10.10.10			0		31026	create table `commod		3秒	原生执行	未执行
5	时:分	10.10.10.10						TABLE `dal_se		3秒	集群执行	未执行
6	时:分	10.10.10.10						TABLE `goods_e		3秒	执行	未执行
7	时:分	10.10.10.10						TABLE `goods_e		3秒	执行	未执行
8	时:分	10.10.10.10						TABLE `user` A		2分11秒	执行	未执行
9	时:分	10.10.10.10						TABLE `user` A		2分11秒	执行	未执行
10	时:分	10.10.10.10						TABLE `user_ad		23秒	执行	未执行
11	时:分	10.10.10.10						TABLE `user_ad		23秒	执行	未执行
12	时:分	10.10.10.10						TABLE `user_ev		1秒	执行	未执行

逻辑读  
逻辑写

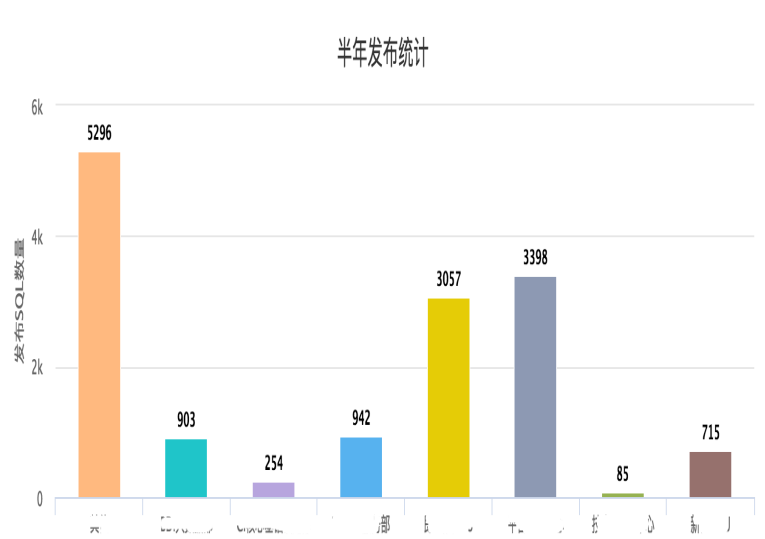
# DBA挑战-DDL



GOPS2017  
Shanghai

数量：

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





GOPS2017  
Shanghai

# 目录

1 多活难点

2 多活架构

3 数据库改造

4 DBA挑战

➔ 5 收益与展望



GOPS2017  
Shanghai

# 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



GOPS2017  
Shanghai

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

# Q&A



GOPS2017  
Shanghai



飞扬过海 

上海 长宁



扫一扫上面的二维码图案，加我微信





GOPS2017  
Shanghai



# Thanks

高效运维社区  
开放运维联盟

荣誉出品



GOPS2017  
Shanghai



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

