Qunar网高可用方案之QMHA

师-黄勇

qunar数据库架构

自我介绍

- Oracle DBA
  - 智联、淘宝
- 去IOE大潮下的改变
- MySQL DBA
  - 百度，去哪儿

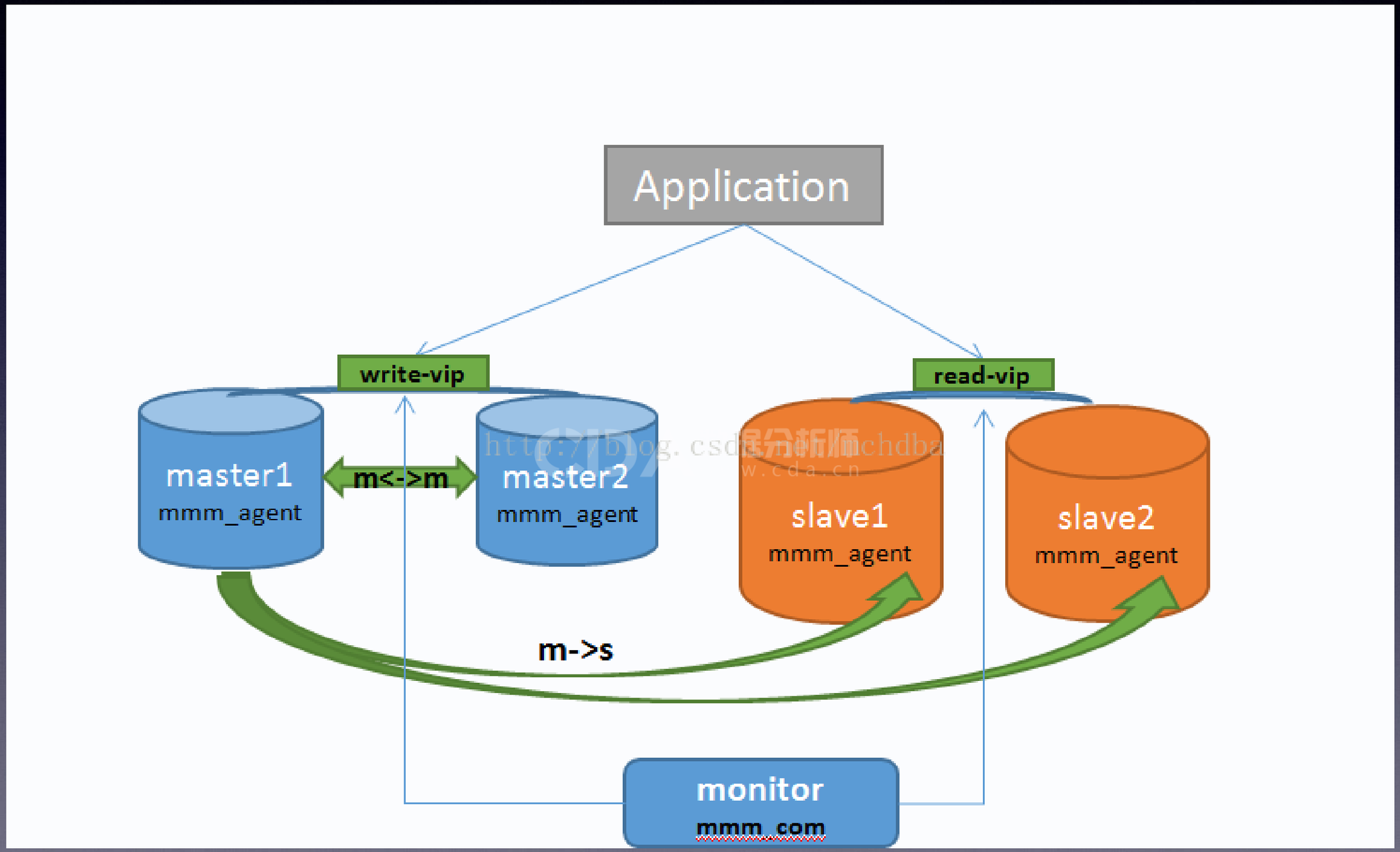
mail: thunderbird.huang@gmail.com

微信：elnino_1114

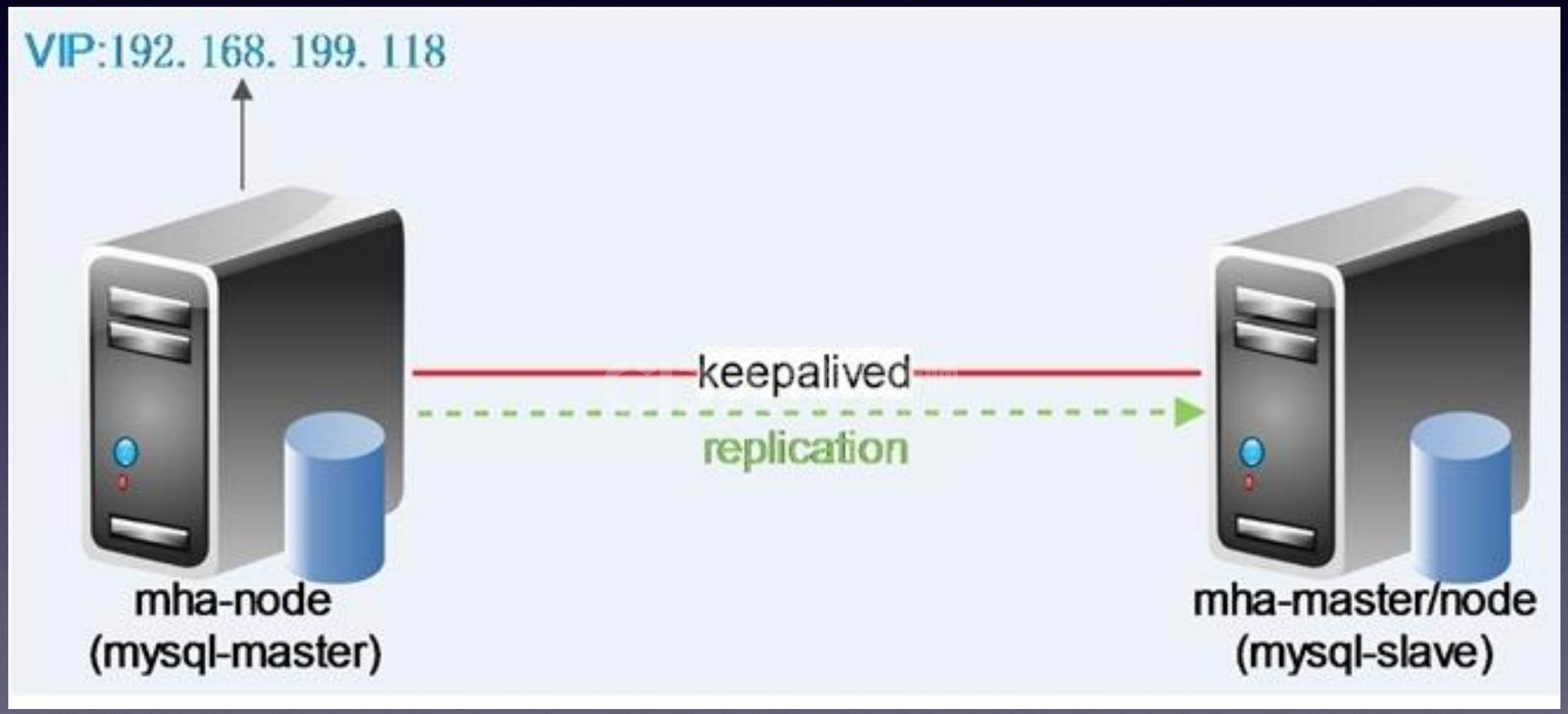
主题内容

- 常用的高可用方案
- QMHA的诞生
- GTID和semi-sync
- 分布式哨兵集群
- QMHA的高可用
- MariaDB MAXSACLE

常用高可用方案-MMM


Application
write-vip
read-vip
master1
mmm_agent
m<->m
master2
mmm_agent
slave1
mmm_agent
slave2
mmm_agent
m->s
monitor
mmm_com

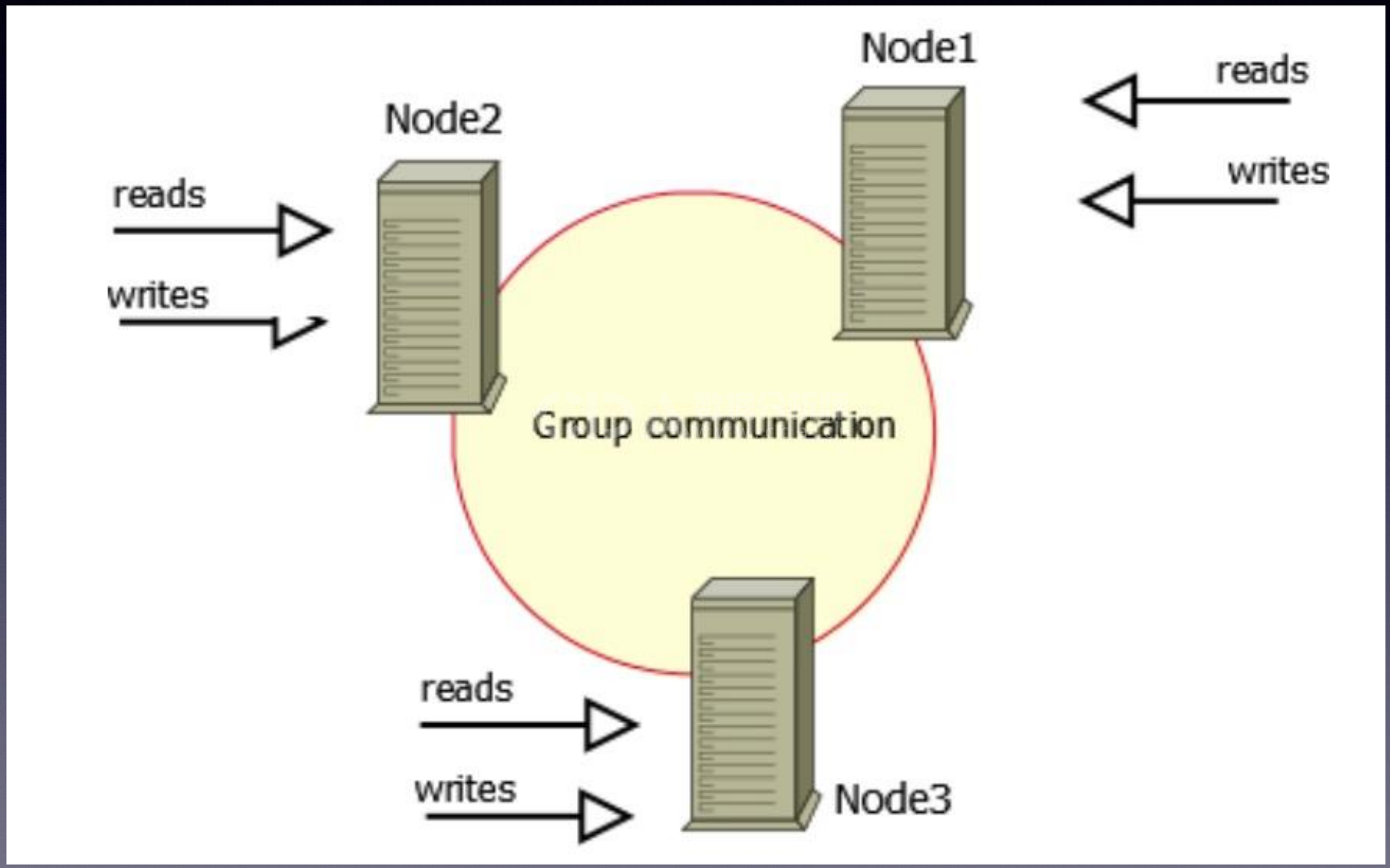
常用高可用方案-MHA


VIP:192.168.199.118
keepalived
replication
mha-node
(mysql-master)
mha-master/node
(mysql-slave)

MMM/MHA架构的问题

- 网络分区，导致数据库双写，数据冲突，需要业务修改数据或者重做
- DBA部署和运维不方便，容易出问题（绑定vip，配置文件等），编写合适的切换逻辑脚本
- MMM通过vip实现漂移，不能跨网段，更不能跨机房
- MHA需要各个节点间开通ssh信任，这是安全的漏洞
- MMM的版本已不更新，谈不上对mysql新特性的支持

常用高可用方案-PXC

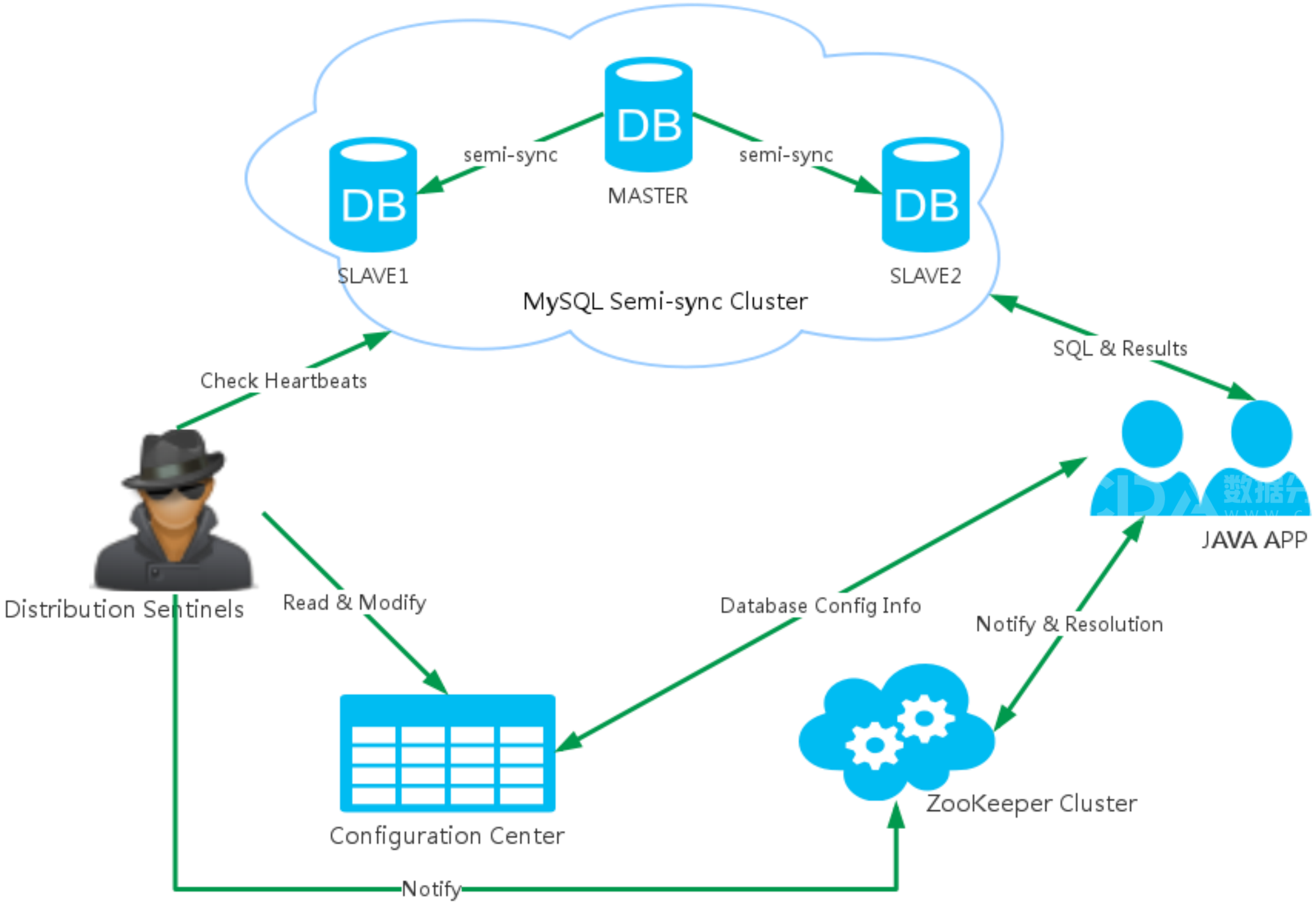

Node1
reads
Node2
reads
writes
writes
Group communication
reads
writes
Node3

PXC架构的问题

- PXC内部节点强一致性，既是优点也是缺点
- 不能跨机房，qps下降，性能损失严重
- 大事务/密集事务会对PXC造成影响
- 集群的写吞吐量取决于最差的一个节点
- flow control
- DBA在运维上需要有学习代价

QMHA的诞生

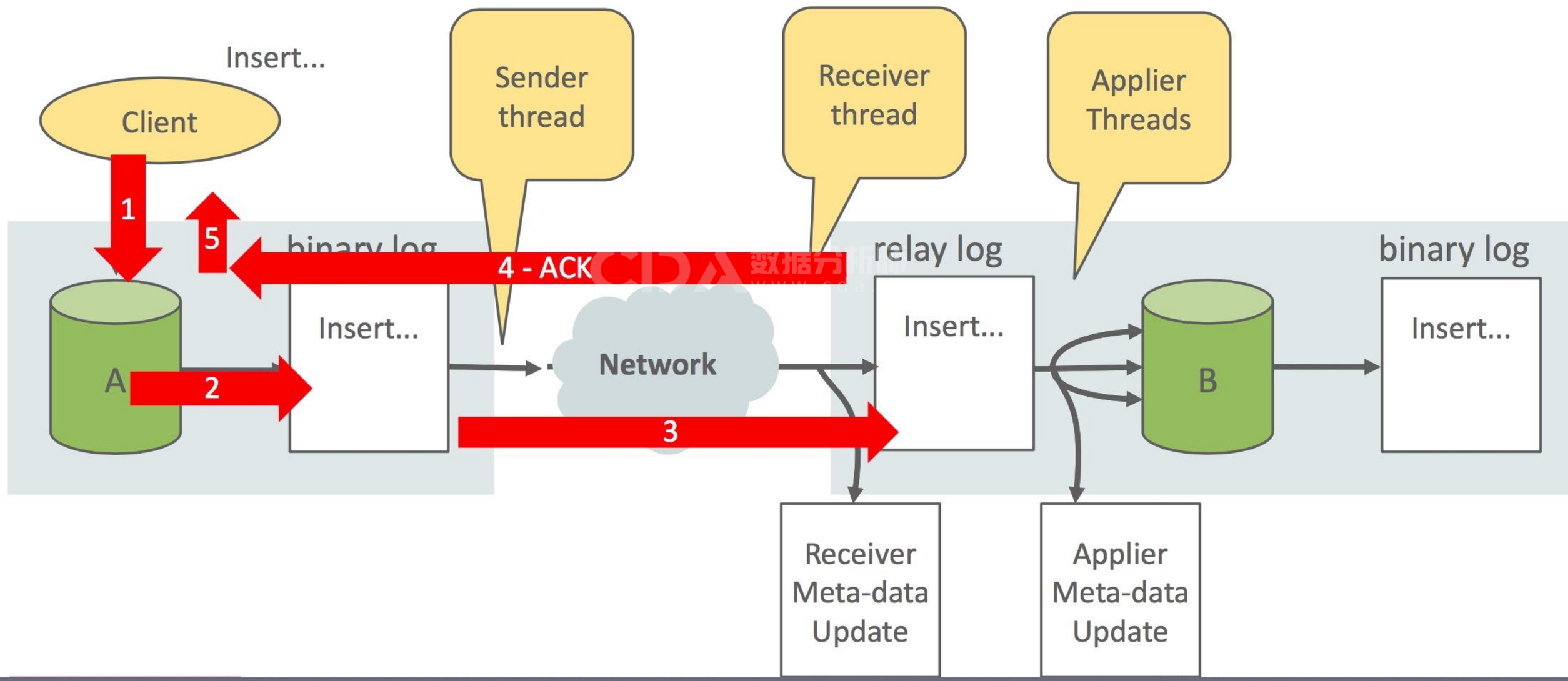
- QMHA架构
- master-slave+sentinel+zookeeper
- master-slave=semi-sync+gtid
- namespace,no vip


DB
semi-sync
semi-sync
DB
MASTER
DB
SLAVE1
SLAVE2
MySQL Semi-sync Cluster
Check Heartbeats
SQL & Results
JAVA APP
Distribution Sentinels
Read & Modify
Database Config Info
Notify & Resolution
Configuration Center
ZooKeeper Cluster
Notify

用到的技术

- GTID
  - uuid+xid，全局唯一性
- semi-sync
  - 至少一个slave接收到master的事务写入relay-log并刷盘
  - innodb_flush_log_at_trx_commit=1 & sync_binlog=1
- 分布式哨兵sentineld
- zookeeper，解析namespace

semi-sync replication

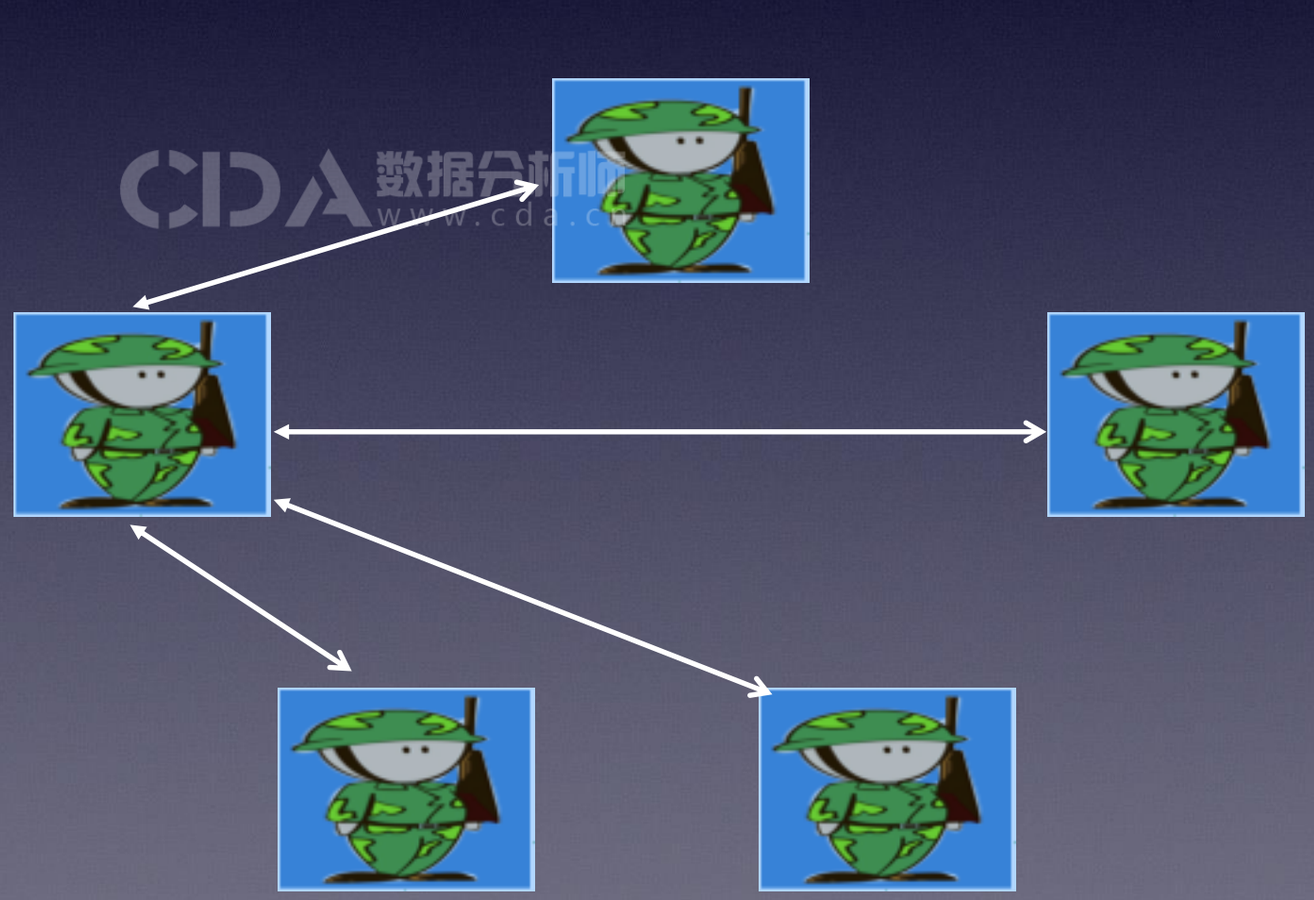


分布式哨兵集群

- 解决MMM/MHA由于网络导致的问题
- 哨兵集群基于多点判断
- 参考redis-sentineld
- sentineld的再实现

分布式哨兵集群

- sentineld: python
- SDOWN
- ODOWN
- campaign(vote)
- leader
- failover



QMHA的高可用

- failover
- switchover
- add a node
- delete a node

tc_order_refund	write	[redacted]	192.168.243.15_3306	online	normal	TC
	read	[redacted]	192.168.243.16_3306	online	normal	TC
		[redacted]	10.90.6.60_3307	online	normal	TC
dba_ccdb	write	[redacted]	10.90.4.207_3306	online	normal	呼叫中心
	read	[redacted]	192.168.39.227_3306	online	normal	呼叫中心
		[redacted]	10.90.4.209_3306	online	normal	呼叫中心
	statistic	[redacted]	192.168.242.11_3310	online	normal	呼叫中心

角色	主机名	实例	是否在线	状态说明	节点类型	操作
write	[redacted]	10.90.4.207_3306	online	normal	qmha	上线 删除 标记为维护 标记为正常 移动到读
read	[redacted]	192.168.39.227_3306	online	normal	qmha	上线 删除 标记为维护 标记为正常 提升为写
	[redacted]	10.90.4.209_3306	online	normal	qmha	上线 删除 标记为维护 标记为正常 提升为写
statistic	[redacted]	192.168.242.11_3310	online	normal	qmha	上线 删除 标记为维护 标记为正常 提升为写

QMHA解决的问题

- 无网络分区：分布式哨兵判断实例死活
- 跨机房：集群中的节点可以分布在多个机房，且建议这样
- 主从一致性：介于ms和pxc之间，性能高
- 0事务丢失：failover和switchover没有事务丢失
- 集中配置：后台配置中心(mysql)存储和维护集群的实时信息
- 集群维护：集群节点上下线、主从切换都对业务透明且操作简单
- 快速切换：测试结果显示failover需要8-16s；switchover需要2s
- 切换逻辑控制：大事务或者主从延迟将不提供切换或者线上服务等均可控

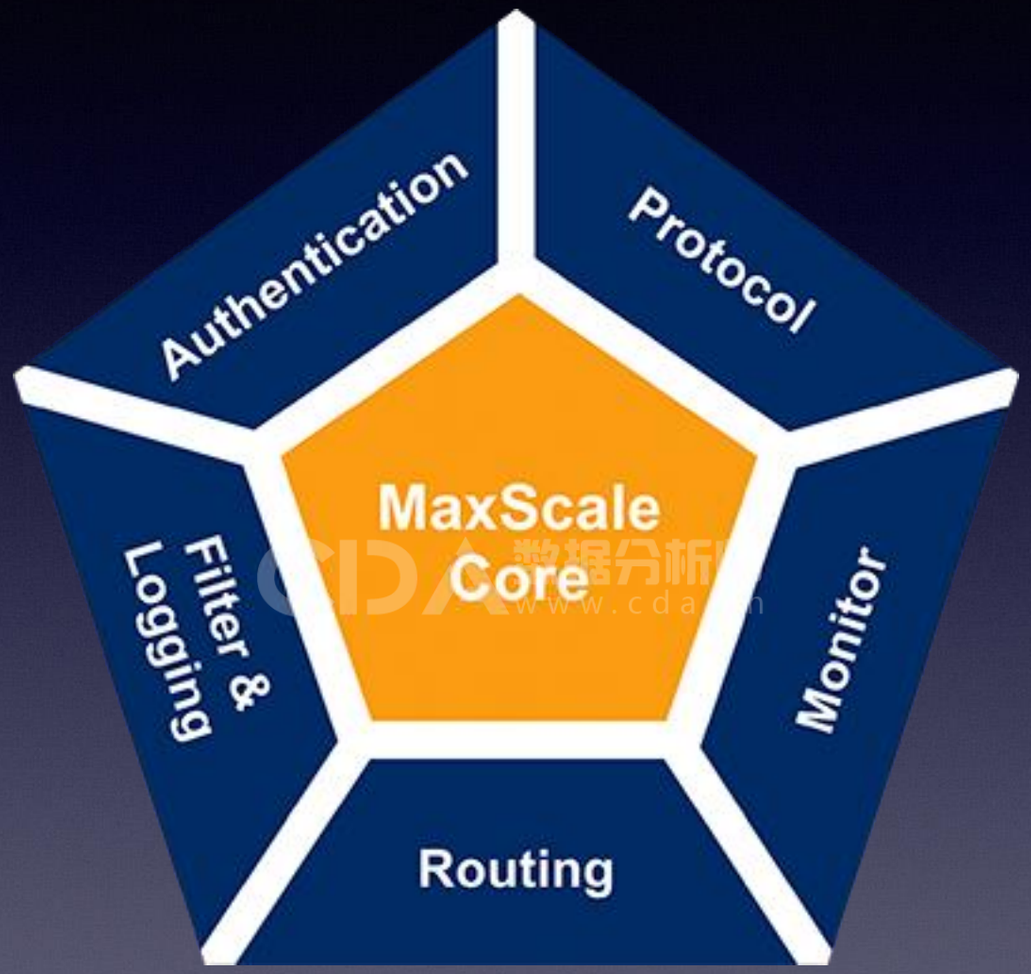
QMHA的对比

各个架构对比	MMM/MHA	PXC	QMHA
一致性	一般	强一致	较好
可用性	一般，受网络影响	一般，受网络影响	很好，网络影响小，可跨地域
数据丢失	主从切换可能会数据丢失	0数据丢失	semi-sync时0数据丢失
成本	至少2台，运维要求低	至少3台，PXC运维门槛较高	至少2台，运维要求低

QMHA缺点和要做的

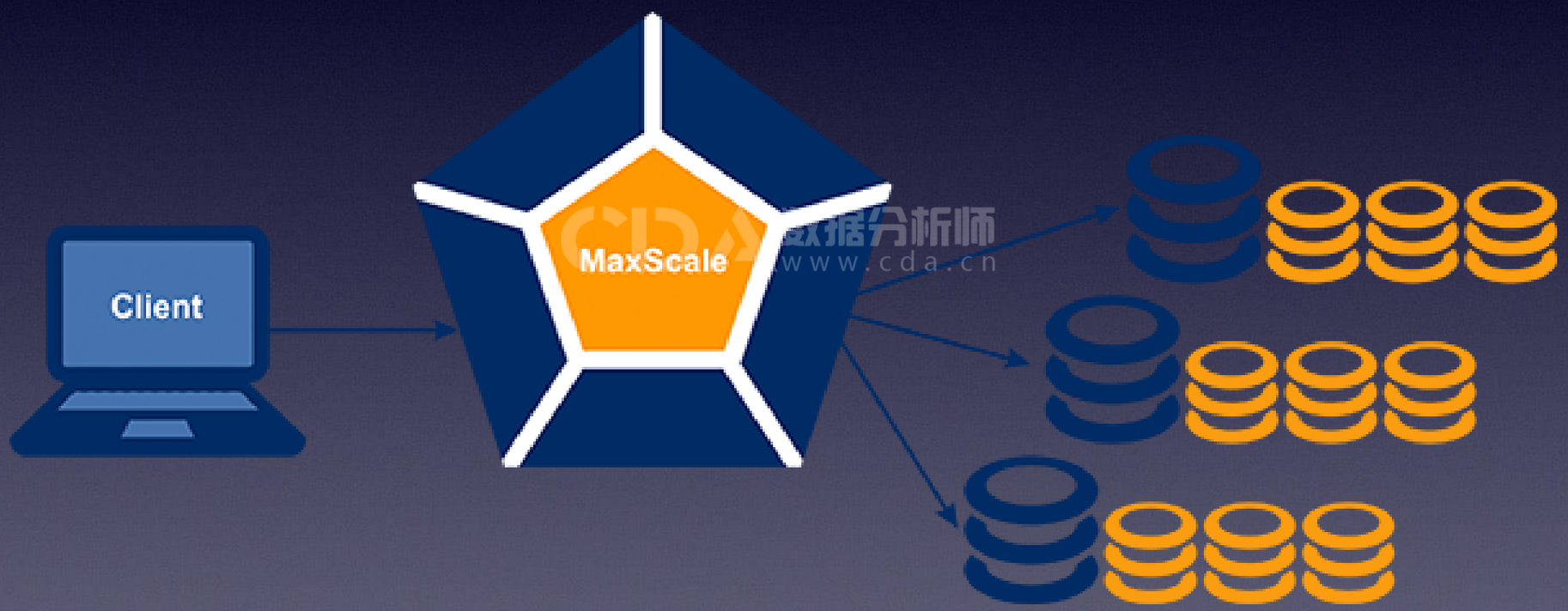
- MHA可以自动补binlog，PXC可以自动IST，QMHA呢？
- QMHA要能在failover之后自动补缺失的binlog给原节点
- 跨机房的从库会存在延迟，当出现机房故障时跨机房的从库可能会由于延迟而出现数据不一致
- 只读数据源的负载利用权重进行控制
- 目前只支持JAVA

MariaDB MAXSCALE



MariaDB MAXSCALE

As a Proxy



MariaDB MAXSCALE

As a Binlog Server