新一代高可用xenon在传统企业的落地与使用

浙报传媒　徐晨亮

公司简介

浙江日报报业集团
ZHEJIANG DAILY PRESS GROUP

- 2000年6月25日，成立浙江日报报业集团
- 2002年，成立浙江日报报业集团有限公司
- 拥有浙江日报、钱江晚报等报刊、出版社26家，发展了浙江新闻客户端、天目新闻客户端、小时新闻客户端等一批新媒体
- 浙江日报(浙江新闻客户端)、浙江在线、钱江晚报传播力位居国内同类媒体前列。
- 连续多年入选“中国500最具价值品牌”“亚洲品牌500强”“世界媒体500强”。

浙报集团官网
浙江在线
爱海宁城市门户
边锋网
东阳网
东阳新闻网
杭州购房宝
浩方电竞平台官网
虎山论坛
乐清城市网
蜜儿网
平安浙江网
钱报网
钱江潮
浙江新闻
小时新闻
爱海宁
东阳侬
名医在浙里
钱报178
世界浙商网
淘志愿
无线瑞安
永康新闻
乐清+
掌上诸暨
浙二好医生
浙江舆情

个人简介

- 浙报传媒MySQL DBA
- 知数堂第12期学员
- MySQL源码爱好者

scrounger

浙江 杭州

扫一扫上面的二维码图案，加我微信

内容

- MHA高可用时期
- xenon的落地
- xenon的高可用思路及原理
- xenon/MySQL Plus在浙江日报的使用概况
- xenon/MySQL Plus踩到的坑

MHA高可用时期

MHA高可用时期

- 1.历史包袱
- 2.未开启GTID（原因是某些排序操作使用了create temporary table ...操作不允许开启GTID）
- 3.业务通过vip方式连接
- 一主三从架构

xenon的落地

MHA VS xenon

MHA

- 支持非GTID/GTID
- 中心化部署
- 不管复制延迟
- 极端情况下可能丢失数据
- 支持5.0+
- 支持任何存储引擎

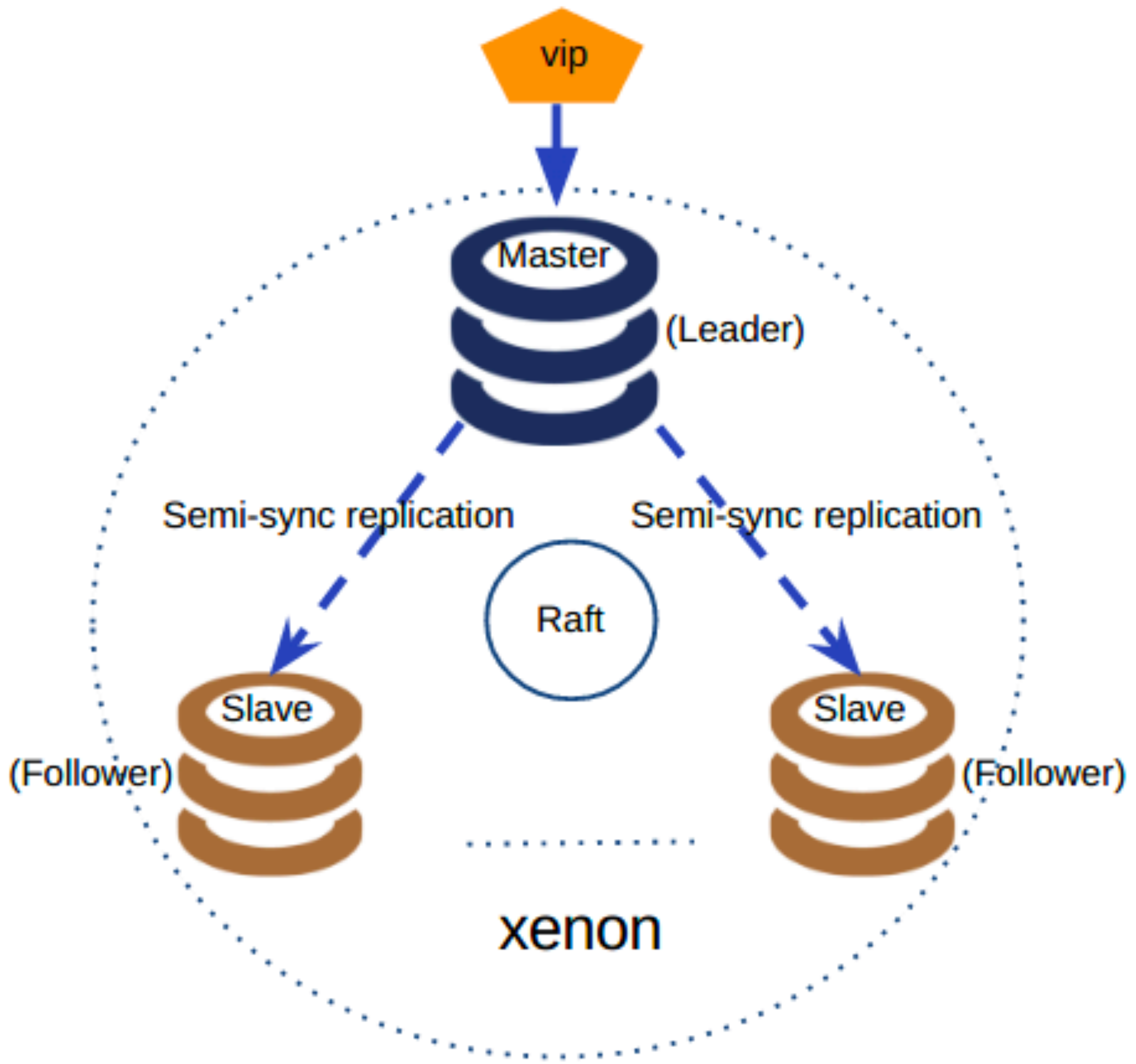
xenon

- GTID
- raft协议
- 去中心化
- 原生MySQL复制
- Zero data loss
- 集群管理/mysql管理/节点重建

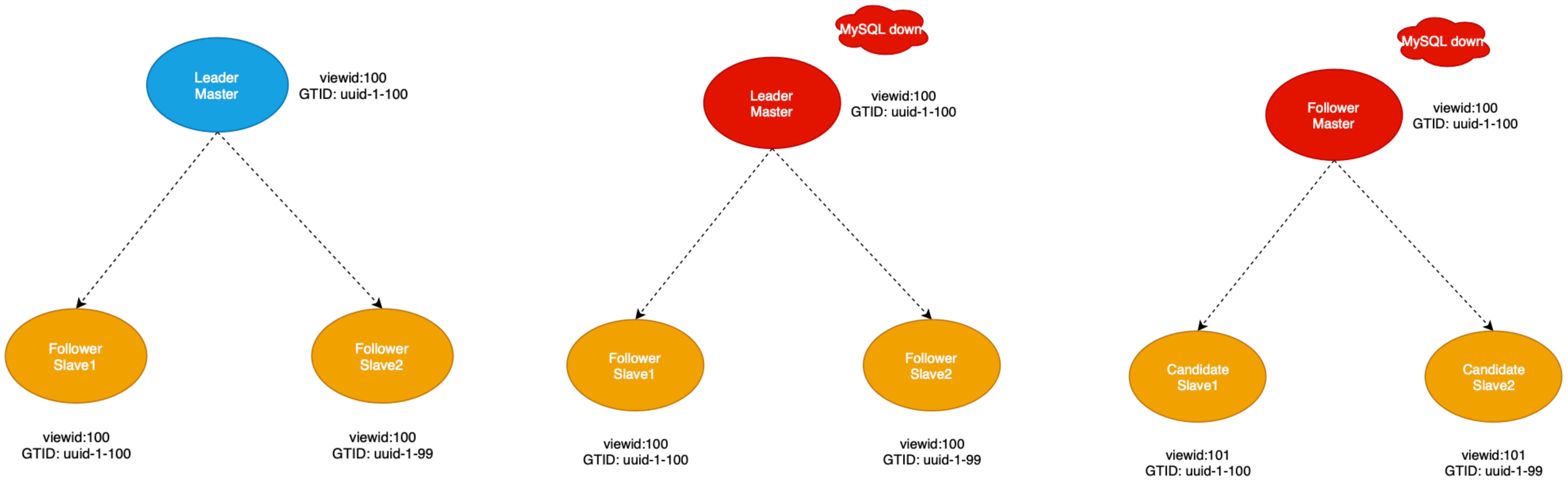
xenon的高可用思路及原理

xenon的高可用思路及原理

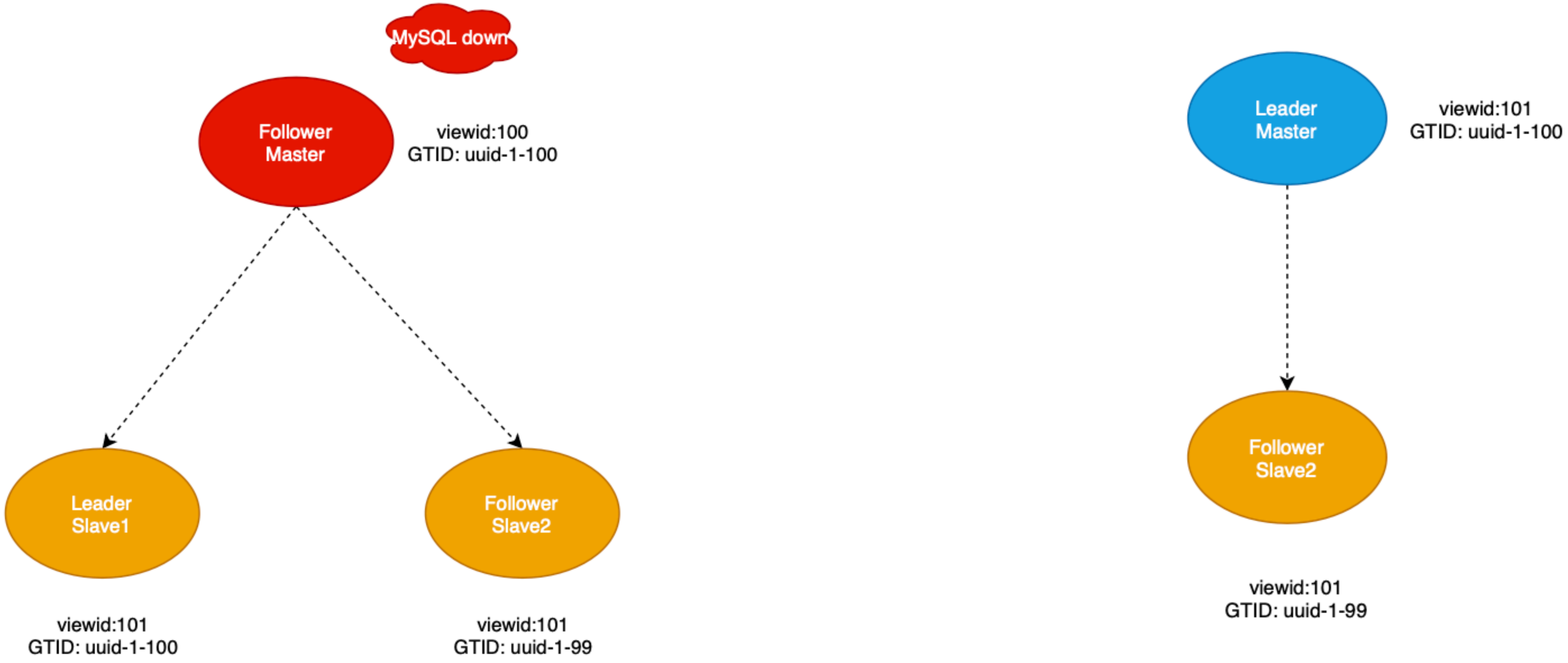
- raft+ protocol
- Semi-replication + after-sync
- HA via VIP
- …



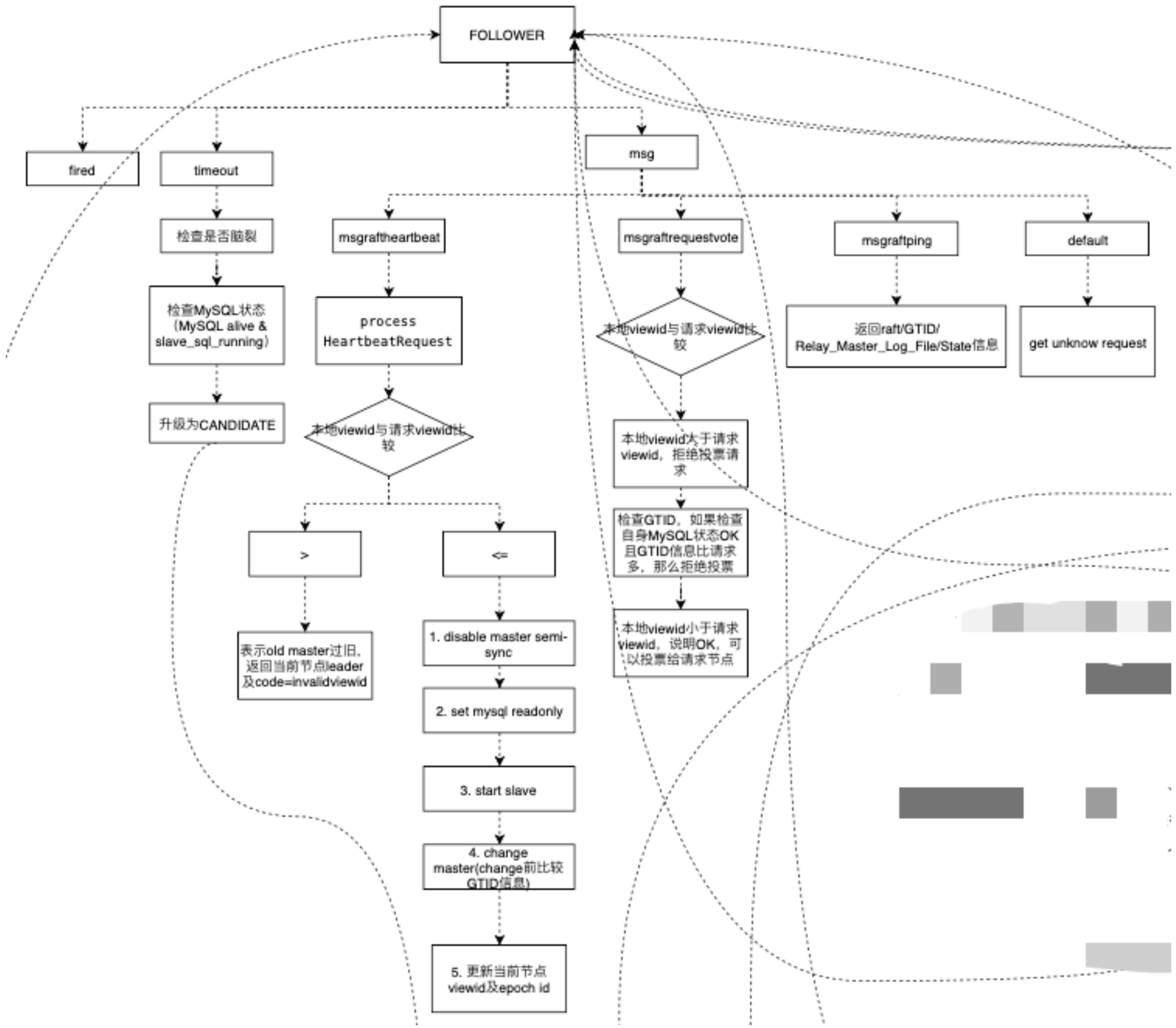
failover过程

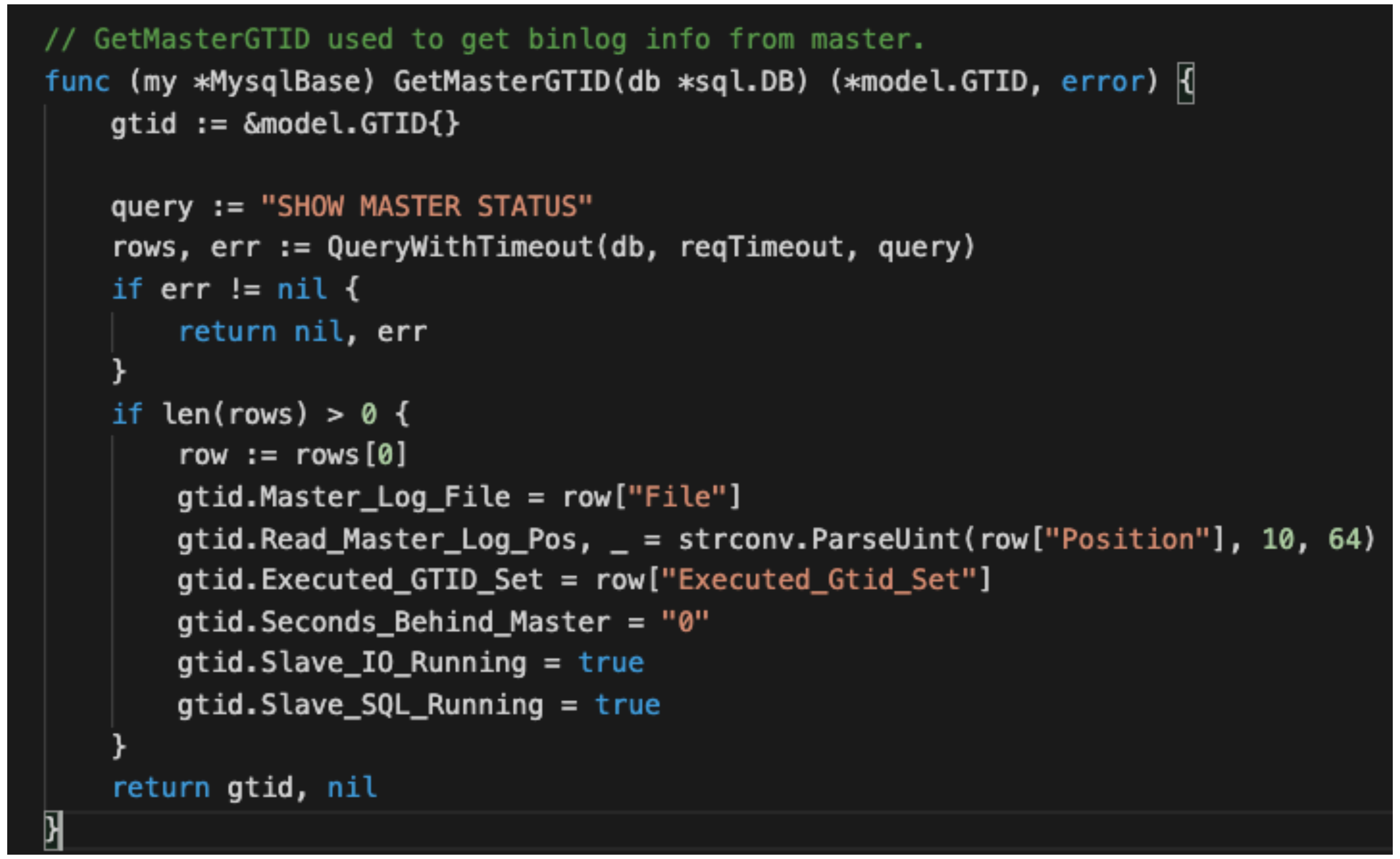

Leader
Master
viewid:100
GTID: uuid-1-100
Follower
Slave1
viewid:100
GTID: uuid-1-100
Follower
Slave2
viewid:100
GTID: uuid-1-99
MySQL down
Leader
Master
viewid:100
GTID: uuid-1-100
Follower
Slave1
viewid:100
GTID: uuid-1-100
Follower
Slave2
viewid:100
GTID: uuid-1-99
MySQL down
Follower
Master
viewid:100
GTID: uuid-1-100
Candidate
Slave1
viewid:101
GTID: uuid-1-100
Candidate
Slave2
viewid:101
GTID: uuid-1-99

failover过程


MySQL down
Follower
Master
viewid:100
GTID: uuid-1-100
Leader
Slave1
viewid:101
GTID: uuid-1-100
Follower
Slave2
viewid:101
GTID: uuid-1-99
Leader
Master
viewid:101
GTID: uuid-1-100
Follower
Slave2
viewid:101
GTID: uuid-1-99

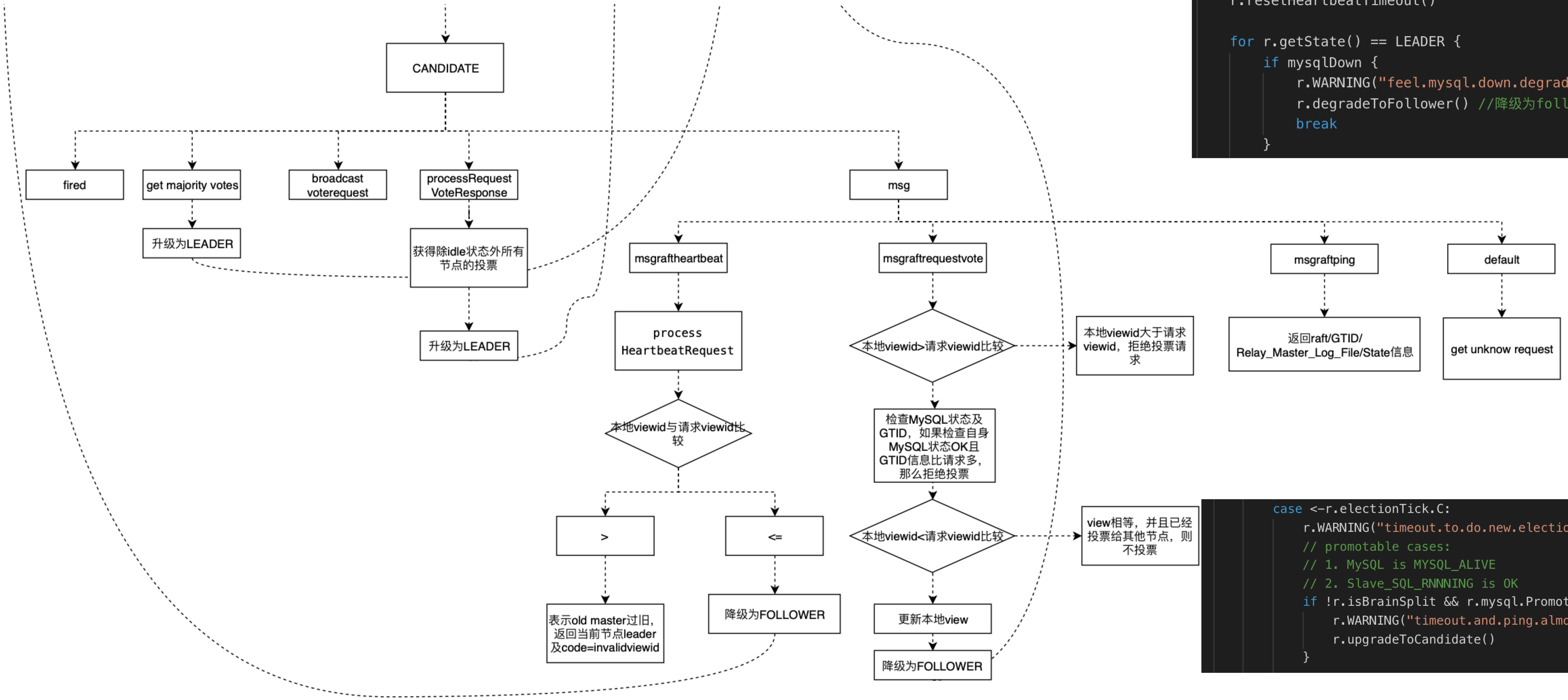
failover过程


FOLLOWER
fired
timeout
msg
检查是否脑裂
检查MySQL状态（MySQL alive & slave_sql_running）
升级为CANDIDATE
msgraftheartbeat
process HeartbeatRequest
本地viewid与请求viewid比较
>
<=
表示old master过旧，返回当前节点leader及code=invalidviewid
1. disable master semi-sync
2. set mysql readonly
3. start slave
4. change master(change前比较GTID信息)
5. 更新当前节点viewid及epoch id
msgraftrequestvote
本地viewid与请求viewid比较
本地viewid大于请求viewid，拒绝投票请求
检查GTID，如果检查自身MySQL状态OK且GTID信息比请求多，那么拒绝投票
本地viewid小于请求viewid，说明OK，可以投票给请求节点
msgraftping
返回raft/GTID/Relay_Master_Log_File/State信息
default
get unknow request



```
// GetMasterGTID used to get binlog info from master.
func (my *MysqlBase) GetMasterGTID(db *sql.DB) (*model.GTID, error) {
    gtid := &model.GTID{}

    query := "SHOW MASTER STATUS"
    rows, err := QueryWithTimeout(db, reqTimeout, query)
    if err != nil {
        return nil, err
    }
    if len(rows) > 0 {
        row := rows[0]
        gtid.Master_Log_File = row["File"]
        gtid.Read_Master_Log_Pos, _ = strconv.ParseUint(row["Position"], 10, 64)
        gtid.Executed_GTID_Set = row["Executed_Gtid_Set"]
        gtid.Seconds_Behind_Master = "0"
        gtid.Slave_IO_Running = true
        gtid.Slave_SQL_Running = true
    }
    return gtid, nil
}
```

failover过程


CANDIDATE
fired
get majority votes
broadcast voterequest
processRequest VoteResponse
msg
升级为LEADER
获得除idle状态外所有节点的投票
升级为LEADER
msgraftheartbeat
msgraftrequestvote
msgraftping
default
process HeartbeatRequest
本地viewid>请求viewid比较
本地viewid大于请求viewid，拒绝投票请求
返回raft/GTID/Relay_Master_Log_File/State信息
get unknow request
本地viewid与请求viewid比较
检查MySQL状态及GTID，如果检查自身MySQL状态OK且GTID信息比请求多，那么拒绝投票
>
&lt;=
本地viewid<请求viewid比较
view相等，并且已经投票给其他节点，则不投票
表示old master过旧，返回当前节点leader及code=invalidviewid
降级为FOLLOWER
更新本地view
降级为FOLLOWER

```
// send heartbeat
respChan := make(chan *model.RaftRPCResponse, r.getMembers())
r.sendHeartbeatHandler(&mysqlDown, respChan)
r.resetHeartbeatTimeout()

for r.getState() == LEADER {
    if mysqlDown {
        r.WARNING("feel.mysql.down.degrade.to.follower")
        r.degradeToFollower() //降级为follower
        break
    }
```

```
case <-r.electionTick.C:
    r.WARNING("timeout.to.do.new.election")
    // promotable cases:
    // 1. MySQL is MYSQL_ALIVE
    // 2. Slave_SQL_RNNNING is OK
    if !r.isBrainSplit && r.mysql.Promotable() {
        r.WARNING("timeout.and.ping.almost.node.successed.promote.
        r.upgradeToCandidate()
    }
```

failover过程

- **leader.go func (r *Leader) prepareSettingsAsync()**
- // MySQL1. wait relay log replay done
  - r.mysql.WaitUntilAfterGTID(gtid.Retrieved_GTID_Set) // 等待本地relay log执行完成
- // MySQL2. change to master
  - r.mysql.ChangeToMaster() //做reset slave all，然后其他节点change到该节点
- // MySQL3. enable semi-sync on master
  - r.mysql.EnableSemiSyncMaster() // 打开半同步
- // MySQL4. set mysql master system variables
  - r.mysql.SetMasterGlobalSysVar() // 设置master必要参数（可配置）
- // MySQL5. set mysql to read/write
  - r.mysql.SetReadWrite() // 打开读写模式
- // MySQL6. Start vip
  - r.leaderStartShellCommand() // call vip

rebuildme原理

手工创建一个备库的过程

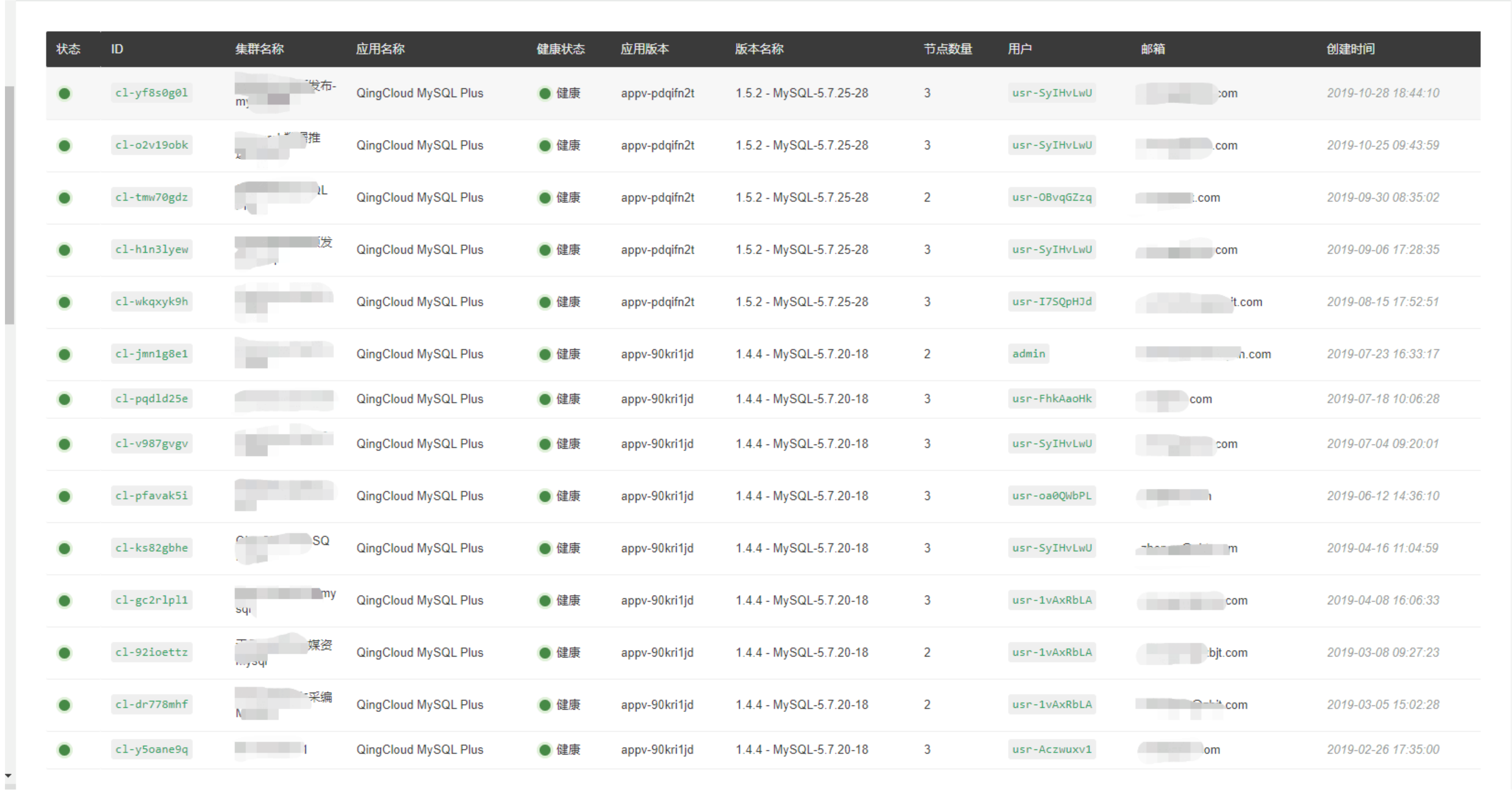
- 1. 登录备库
- 2. 执行xtrabackup物理备份
- 3. 拷贝到目标机器
- 4. apply redo log
- 5. copy back
- 6. 修改数据目录权限
- 7. 启动MySQL
- 8. set gtid purged
- 9. change master
- 10. start slave

rebuildme原理

- 1. first to check I am leader or not
- 2. find the best to backup
- 3. check bestone is not in BACKUPING(Slave_SQL_Running && Slave_IO_Running && Seconds_Behind_Master<100)
- 4. disable raft
- 5. stop monitor
- 6. force kill mysqld
- 7. check bestone is not in BACKUPING again
- 8. remove data files
- 9. do backup from bestone
- 10. do apply-log
- 11. start mysqld
- 12. wait mysqld running
- 13. wait mysql working
- 14. stop slave and reset slave all
- 15. set gtid_purged
- 16. enable raft
- 17. wait change to master
- 18. start slave
- 19. rebuildme ok

rebuildme

xenon/MySQL Plus在浙江日报的使用概况

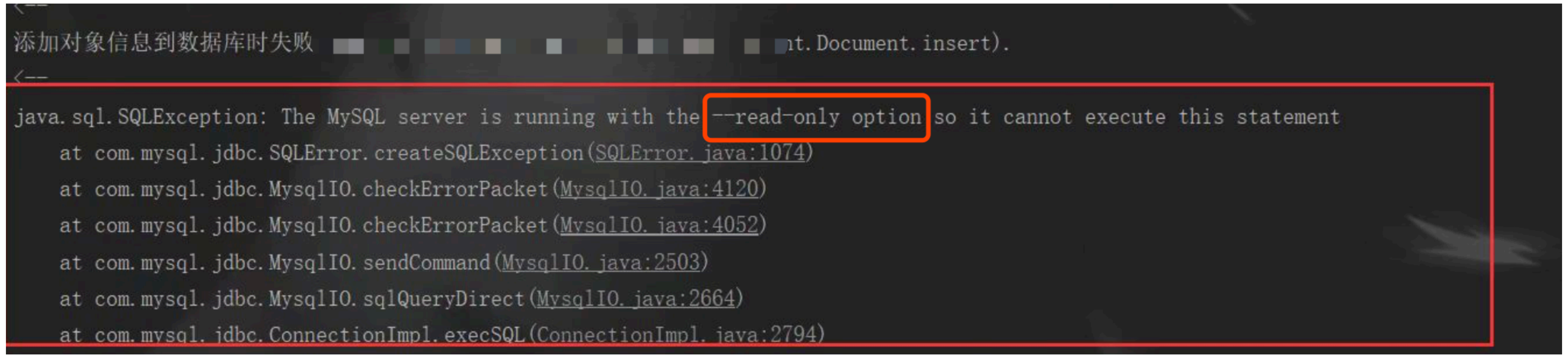


状态	ID	集群名称	应用名称	健康状态	应用版本	版本名称	节点数量	用户	邮箱	创建时间
●	cl-yf8s0g0l	[redacted]	QingCloud MySQL Plus	● 健康	appv-pdqifn2t	1.5.2 - MySQL-5.7.25-28	3	usr-SyIHvLwU	[redacted]	2019-10-28 18:44:10
●	cl-o2v19obk	[redacted]	QingCloud MySQL Plus	● 健康	appv-pdqifn2t	1.5.2 - MySQL-5.7.25-28	3	usr-SyIHvLwU	[redacted]	2019-10-25 09:43:59
●	cl-tmw70gdz	[redacted]	QingCloud MySQL Plus	● 健康	appv-pdqifn2t	1.5.2 - MySQL-5.7.25-28	2	usr-OBvqGZzq	[redacted]	2019-09-30 08:35:02
●	cl-h1n3lyew	[redacted]	QingCloud MySQL Plus	● 健康	appv-pdqifn2t	1.5.2 - MySQL-5.7.25-28	3	usr-SyIHvLwU	[redacted]	2019-09-06 17:28:35
●	cl-wkqxyk9h	[redacted]	QingCloud MySQL Plus	● 健康	appv-pdqifn2t	1.5.2 - MySQL-5.7.25-28	3	usr-I7SQpHJd	[redacted]	2019-08-15 17:52:51
●	cl-jmn1g8e1	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	2	admin	[redacted]	2019-07-23 16:33:17
●	cl-pqdld25e	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	3	usr-FhkAaoHk	[redacted]	2019-07-18 10:06:28
●	cl-v987gvgv	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	3	usr-SyIHvLwU	[redacted]	2019-07-04 09:20:01
●	cl-pfavak5i	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	3	usr-oa0QWbPL	[redacted]	2019-06-12 14:36:10
●	cl-ks82gbhe	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	3	usr-SyIHvLwU	[redacted]	2019-04-16 11:04:59
●	cl-gc2rlpl1	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	3	usr-1vAxRbLA	[redacted]	2019-04-08 16:06:33
●	cl-92ioettz	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	2	usr-1vAxRbLA	[redacted]	2019-03-08 09:27:23
●	cl-dr778mhf	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	2	usr-1vAxRbLA	[redacted]	2019-03-05 15:02:28
●	cl-y5oane9q	[redacted]	QingCloud MySQL Plus	● 健康	appv-90kri1jd	1.4.4 - MySQL-5.7.20-18	3	usr-Aczwuxv1	[redacted]	2019-02-26 17:35:00

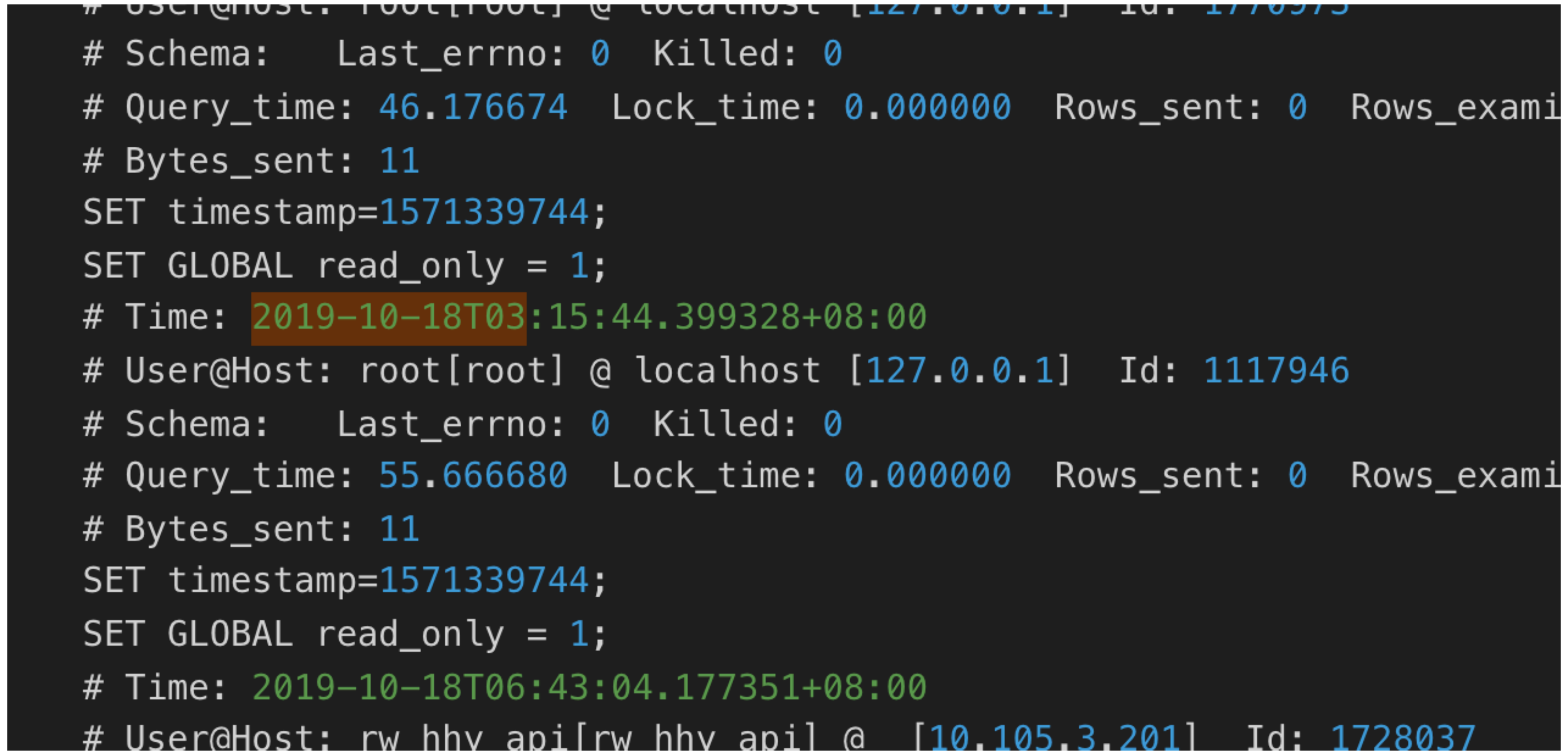
实例数 50+　业务数 40+　单节点QPS 20000+

xenon/MySQL Plus踩到的坑

复杂情况下的网络切换



```
<--
添加对象信息到数据库时失败 [redacted] nt.Document.insert).
<--
java.sql.SQLException: The MySQL server is running with the --read-only option so it cannot execute this statement
    at com.mysql.jdbc.SQLError.createSQLException(SQLError.java:1074)
    at com.mysql.jdbc.MysqlIO.checkErrorPacket(MysqlIO.java:4120)
    at com.mysql.jdbc.MysqlIO.checkErrorPacket(MysqlIO.java:4052)
    at com.mysql.jdbc.MysqlIO.sendCommand(MysqlIO.java:2503)
    at com.mysql.jdbc.MysqlIO.sqlQueryDirect(MysqlIO.java:2664)
    at com.mysql.jdbc.ConnectionImpl.execSQL(ConnectionImpl.java:2794)
```



```
# User@Host: root[root] @ localhost [127.0.0.1]  Id: 1770973
# Schema:   Last_errno: 0  Killed: 0
# Query_time: 46.176674  Lock_time: 0.000000  Rows_sent: 0  Rows_exami
# Bytes_sent: 11
SET timestamp=1571339744;
SET GLOBAL read_only = 1;
# Time: 2019-10-18T03:15:44.399328+08:00
# User@Host: root[root] @ localhost [127.0.0.1]  Id: 1117946
# Schema:   Last_errno: 0  Killed: 0
# Query_time: 55.666680  Lock_time: 0.000000  Rows_sent: 0  Rows_exami
# Bytes_sent: 11
SET timestamp=1571339744;
SET GLOBAL read_only = 1;
# Time: 2019-10-18T06:43:04.177351+08:00
# User@Host: rw_hhy_api[rw_hhy_api] @  [10.105.3.201]  Id: 1728037
```

Leader 降级设置global read_only超时

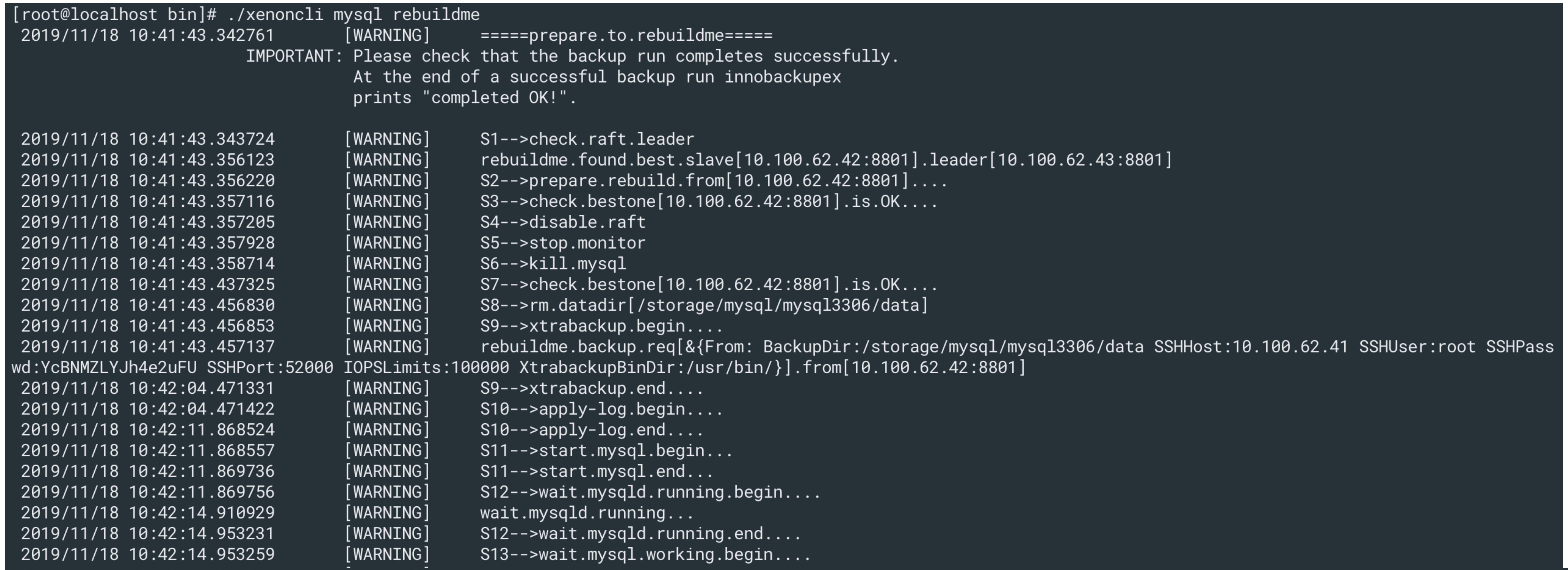
复杂情况下的网络切换

LinuxAIOHandler::collect:sync_array_free_cell:LinuxAIOHandler::find_completed_slot:
buf_page_io_complete:fil_aio_wait:io_handler_thread:start_thread

解决办法：

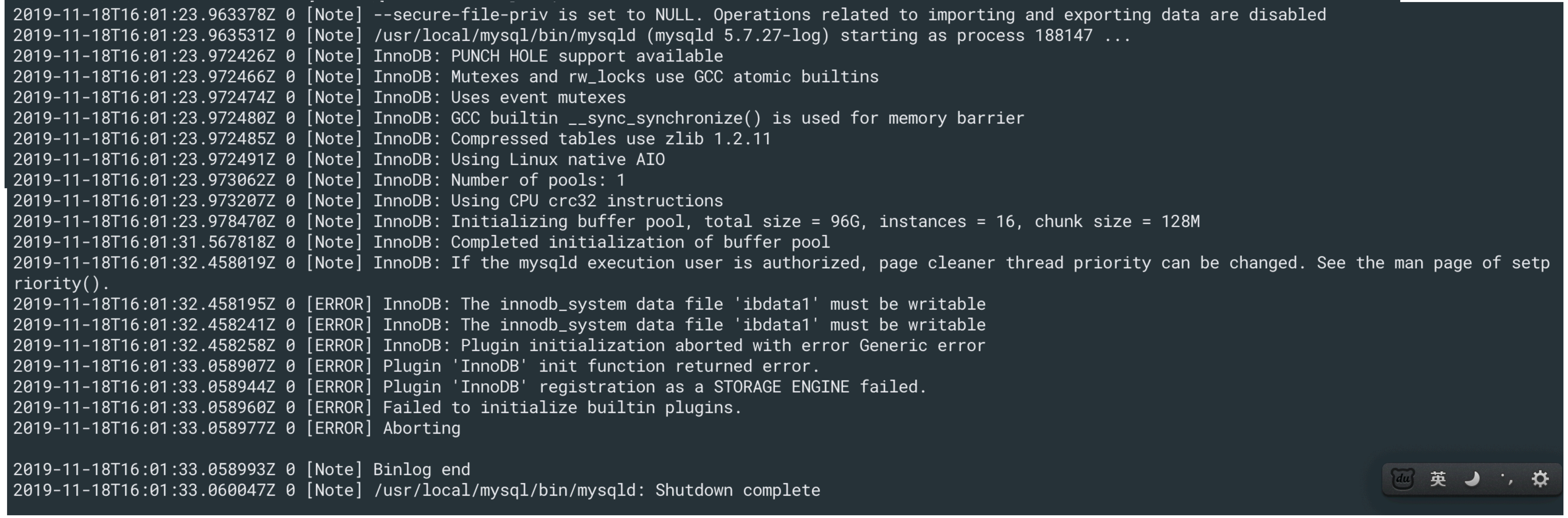
- innodb_use_native_aio = OFF
- 添加check read_only
- MySQL Plus 1.5.5修复bug

rebuildme卡住



```
[root@localhost bin]# ./xenoncli mysql rebuildme
 2019/11/18 10:41:43.342761       [WARNING]     =====prepare.to.rebuildme=====
                        IMPORTANT: Please check that the backup run completes successfully.
                                   At the end of a successful backup run innobackupex
                                   prints "completed OK!".

 2019/11/18 10:41:43.343724       [WARNING]     S1-->check.raft.leader
 2019/11/18 10:41:43.356123       [WARNING]     rebuildme.found.best.slave[10.100.62.42:8801].leader[10.100.62.43:8801]
 2019/11/18 10:41:43.356220       [WARNING]     S2-->prepare.rebuild.from[10.100.62.42:8801]....
 2019/11/18 10:41:43.357116       [WARNING]     S3-->check.bestone[10.100.62.42:8801].is.OK....
 2019/11/18 10:41:43.357205       [WARNING]     S4-->disable.raft
 2019/11/18 10:41:43.357928       [WARNING]     S5-->stop.monitor
 2019/11/18 10:41:43.358714       [WARNING]     S6-->kill.mysql
 2019/11/18 10:41:43.437325       [WARNING]     S7-->check.bestone[10.100.62.42:8801].is.OK....
 2019/11/18 10:41:43.456830       [WARNING]     S8-->rm.datadir[/storage/mysql/mysql3306/data]
 2019/11/18 10:41:43.456853       [WARNING]     S9-->xtrabackup.begin....
 2019/11/18 10:41:43.457137       [WARNING]     rebuildme.backup.req[&{From: BackupDir:/storage/mysql/mysql3306/data SSHHost:10.100.62.41 SSHUser:root SSHPass
wd:YcBNMZLYJh4e2uFU SSHPort:52000 IOPSLimits:100000 XtrabackupBinDir:/usr/bin/}].from[10.100.62.42:8801]
 2019/11/18 10:42:04.471331       [WARNING]     S9-->xtrabackup.end....
 2019/11/18 10:42:04.471422       [WARNING]     S10-->apply-log.begin....
 2019/11/18 10:42:11.868524       [WARNING]     S10-->apply-log.end....
 2019/11/18 10:42:11.868557       [WARNING]     S11-->start.mysql.begin...
 2019/11/18 10:42:11.869736       [WARNING]     S11-->start.mysql.end...
 2019/11/18 10:42:11.869756       [WARNING]     S12-->wait.mysqld.running.begin....
 2019/11/18 10:42:14.910929       [WARNING]     wait.mysqld.running...
 2019/11/18 10:42:14.953231       [WARNING]     S12-->wait.mysqld.running.end....
 2019/11/18 10:42:14.953259       [WARNING]     S13-->wait.mysql.working.begin....
```

xenon日志



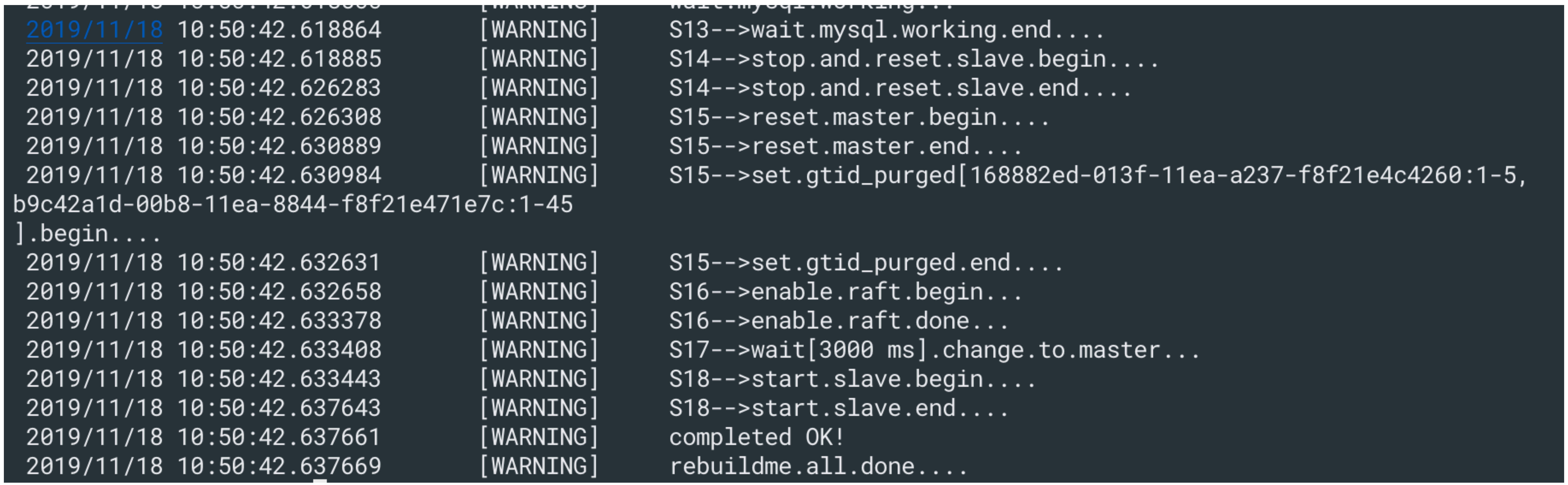
```
2019-11-18T16:01:23.963378Z 0 [Note] --secure-file-priv is set to NULL. Operations related to importing and exporting data are disabled
2019-11-18T16:01:23.963531Z 0 [Note] /usr/local/mysql/bin/mysqld (mysqld 5.7.27-log) starting as process 188147 ...
2019-11-18T16:01:23.972426Z 0 [Note] InnoDB: PUNCH HOLE support available
2019-11-18T16:01:23.972466Z 0 [Note] InnoDB: Mutexes and rw_locks use GCC atomic builtins
2019-11-18T16:01:23.972474Z 0 [Note] InnoDB: Uses event mutexes
2019-11-18T16:01:23.972480Z 0 [Note] InnoDB: GCC builtin __sync_synchronize() is used for memory barrier
2019-11-18T16:01:23.972485Z 0 [Note] InnoDB: Compressed tables use zlib 1.2.11
2019-11-18T16:01:23.972491Z 0 [Note] InnoDB: Using Linux native AIO
2019-11-18T16:01:23.973062Z 0 [Note] InnoDB: Number of pools: 1
2019-11-18T16:01:23.973207Z 0 [Note] InnoDB: Using CPU crc32 instructions
2019-11-18T16:01:23.978470Z 0 [Note] InnoDB: Initializing buffer pool, total size = 96G, instances = 16, chunk size = 128M
2019-11-18T16:01:31.567818Z 0 [Note] InnoDB: Completed initialization of buffer pool
2019-11-18T16:01:32.458019Z 0 [Note] InnoDB: If the mysqld execution user is authorized, page cleaner thread priority can be changed. See the man page of setp
riority().
2019-11-18T16:01:32.458195Z 0 [ERROR] InnoDB: The innodb_system data file 'ibdata1' must be writable
2019-11-18T16:01:32.458241Z 0 [ERROR] InnoDB: The innodb_system data file 'ibdata1' must be writable
2019-11-18T16:01:32.458258Z 0 [ERROR] InnoDB: Plugin initialization aborted with error Generic error
2019-11-18T16:01:33.058907Z 0 [ERROR] Plugin 'InnoDB' init function returned error.
2019-11-18T16:01:33.058944Z 0 [ERROR] Plugin 'InnoDB' registration as a STORAGE ENGINE failed.
2019-11-18T16:01:33.058960Z 0 [ERROR] Failed to initialize builtin plugins.
2019-11-18T16:01:33.058977Z 0 [ERROR] Aborting

2019-11-18T16:01:33.058993Z 0 [Note] Binlog end
2019-11-18T16:01:33.060047Z 0 [Note] /usr/local/mysql/bin/mysqld: Shutdown complete
```

mysql错误日志

解决办法1

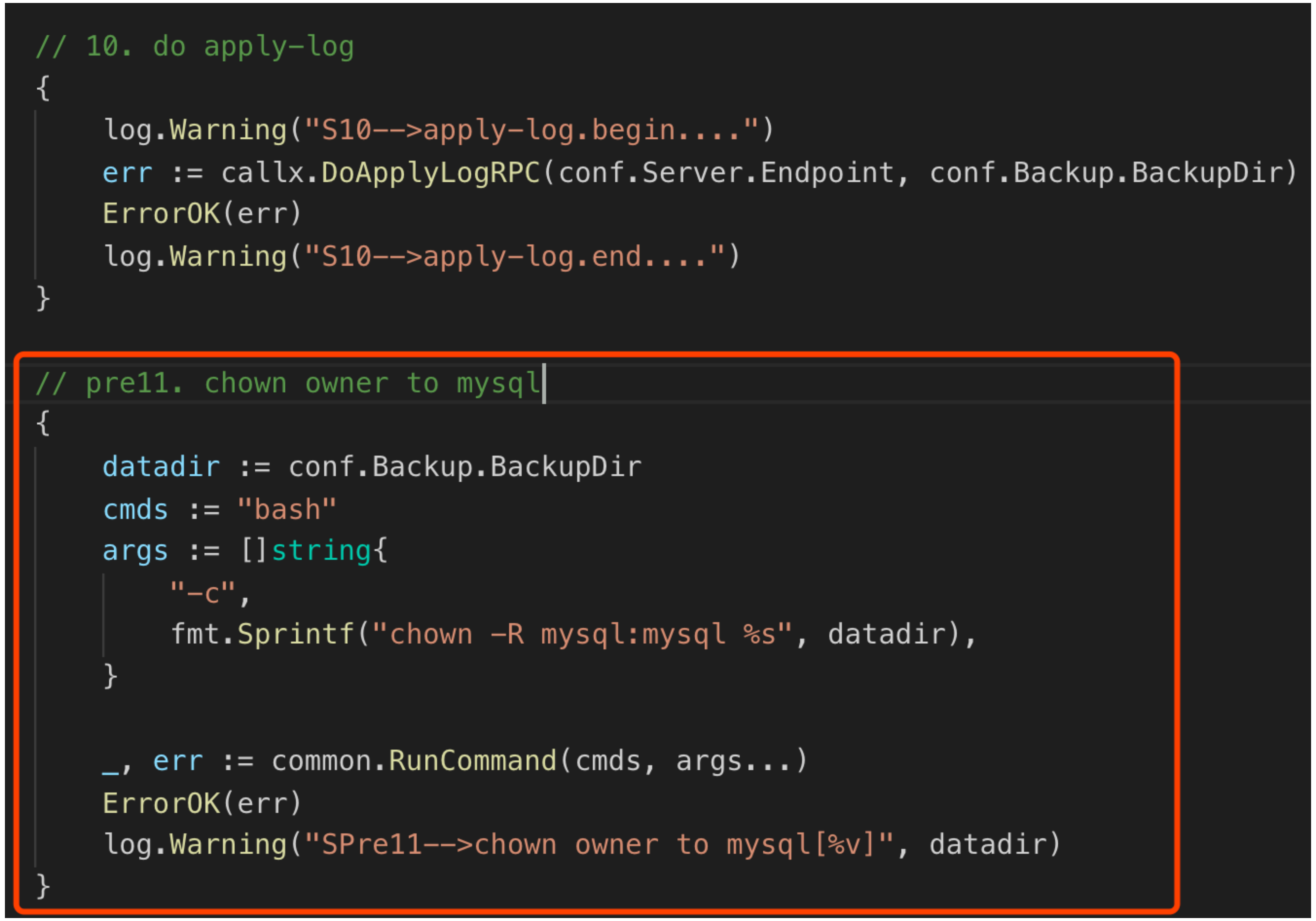
手动修改mysql datadir目录权限

chown -R mysql.mysql data/



```
 2019/11/18 10:50:42.618864        [WARNING]     S13-->wait.mysql.working.end....
 2019/11/18 10:50:42.618885        [WARNING]     S14-->stop.and.reset.slave.begin....
 2019/11/18 10:50:42.626283        [WARNING]     S14-->stop.and.reset.slave.end....
 2019/11/18 10:50:42.626308        [WARNING]     S15-->reset.master.begin....
 2019/11/18 10:50:42.630889        [WARNING]     S15-->reset.master.end....
 2019/11/18 10:50:42.630984        [WARNING]     S15-->set.gtid_purged[168882ed-013f-11ea-a237-f8f21e4c4260:1-5,
b9c42a1d-00b8-11ea-8844-f8f21e471e7c:1-45
].begin....
 2019/11/18 10:50:42.632631        [WARNING]     S15-->set.gtid_purged.end....
 2019/11/18 10:50:42.632658        [WARNING]     S16-->enable.raft.begin...
 2019/11/18 10:50:42.633378        [WARNING]     S16-->enable.raft.done...
 2019/11/18 10:50:42.633408        [WARNING]     S17-->wait[3000 ms].change.to.master...
 2019/11/18 10:50:42.633443        [WARNING]     S18-->start.slave.begin....
 2019/11/18 10:50:42.637643        [WARNING]     S18-->start.slave.end....
 2019/11/18 10:50:42.637661        [WARNING]     completed OK!
 2019/11/18 10:50:42.637669        [WARNING]     rebuildme.all.done....
```

解决方法2



```go
// 10. do apply-log
{
    log.Warning("S10-->apply-log.begin....")
    err := callx.DoApplyLogRPC(conf.Server.Endpoint, conf.Backup.BackupDir)
    ErrorOK(err)
    log.Warning("S10-->apply-log.end....")
}

// pre11. chown owner to mysql
{
    datadir := conf.Backup.BackupDir
    cmds := "bash"
    args := []string{
        "-c",
        fmt.Sprintf("chown -R mysql:mysql %s", datadir),
    }

    _, err := common.RunCommand(cmds, args...)
    ErrorOK(err)
    log.Warning("SPre11-->chown owner to mysql[%v]", datadir)
}
```

```
{

    datadir := conf.Backup.BackupDir

    cmds := "bash"

    args := []string{

        "-c",

        fmt.Sprintf("chown -R mysql:mysql %s", datadir),

        }

    _, err := common.RunCommand(cmds, args…)

    ErrorOK(err)

    log.Warning("SPre11-->chown owner to mysql[%v]", datadir)

}
```

xenon未来期望

- 支持MySQL 8.0
- 增加切换次数控制
- 一个”精美”的web界面

Thanks