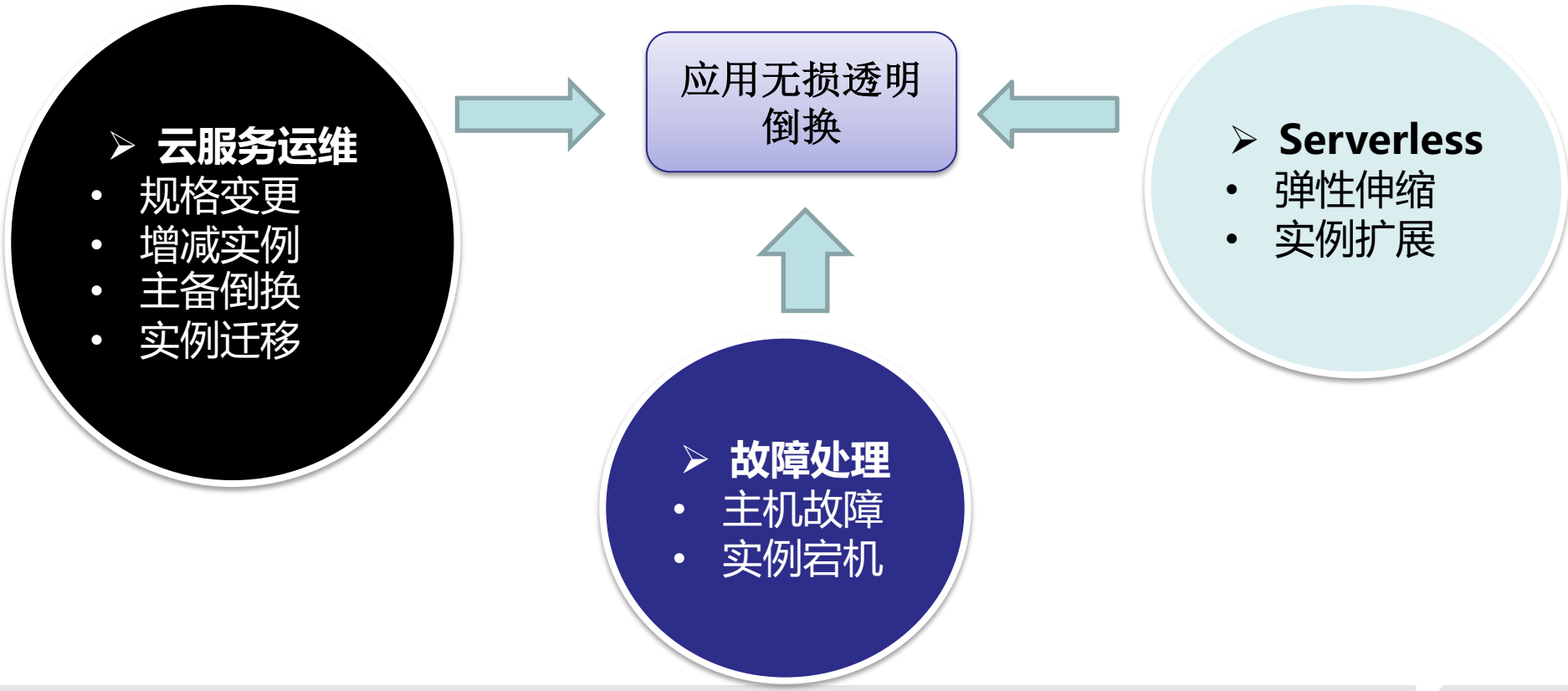
# 面向互联网场景的云原生高可用

华为云数据库 彭立勋

# 云原生数据库在互联网场景的挑战

云服务不可避免的需要做一些日常维护操作，以及异常故障处理，以及云原生数据库向Serverless演进时，都会带来数据库实例的倒换操作，通常这会引起用户的应用程序处理中断，使得应用不得不做复杂的处理逻辑来处理异常。
将数据库的倒换操作做到**对应用无损透明**，可以极大的降低应用复杂度。



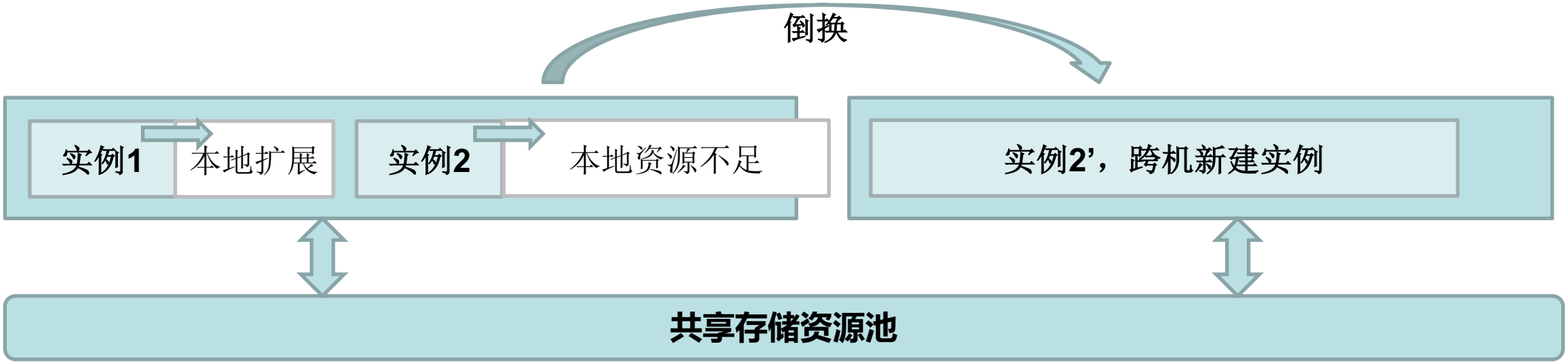
# 应用对 SwitchOver 和 FailOver 的处理

当前，数据库系统切换与故障转移下，用户的应用程序需感知系统的变化并提供复杂的应对措施：

- **连接是否中断**
- **事务是否中断**
- **如何进行事务补偿**
- **如何重建 Session 上下文**

# 云原生数据库 Serverless 的挑战

- 云原生开始全面进入 Serverless 化，资源都要做到弹性可伸缩
- 云原生存算分离架构，共享存储池可以实现存储资源的在线弹性伸缩
- 计算资源目前普遍与物理主机资源绑定，跨机调度会产生实例倒换



# 应用无损透明倒换要解决的目标

- 倒换时避免连接和事务中断
- 无需应用对事务进行补偿
- 无需恢复和重建 Session 上下文

- ALT (Application Lossless and Transparent) 框架

# ALT 技术架构

- **Drain Session**

排干正在进行的事务或查询。

- **Transaction Guard**

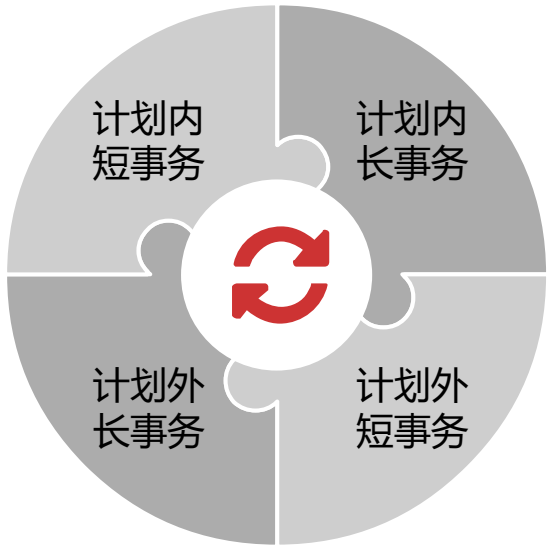
由 Sync MSG 和 Redo Applying 在 Node B 中保持事务状态(包括锁)

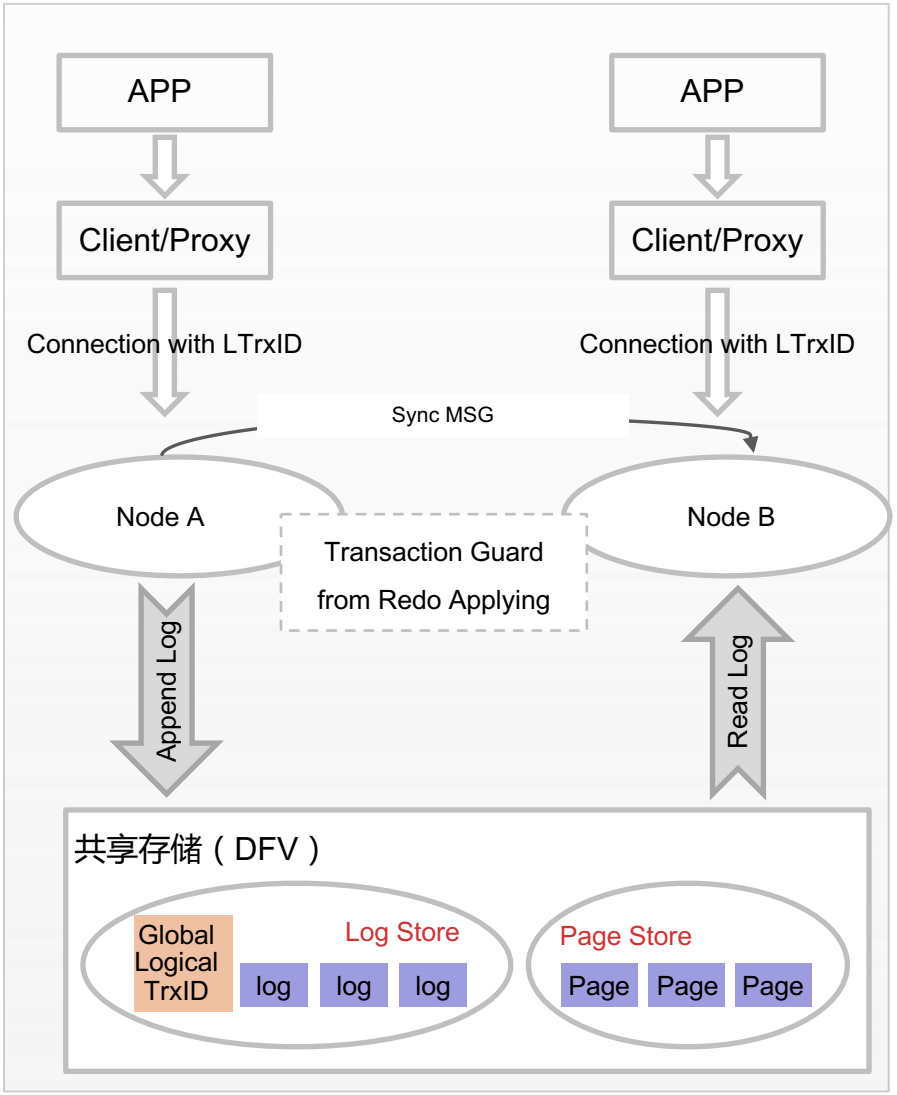
- **Logical Transaction ID**

由Redirect的Connection携带，定位到Transaction Guard的In-flight transaction ，Connection Resubmit SQL语句，事务继续幂等执行，满足一致性。

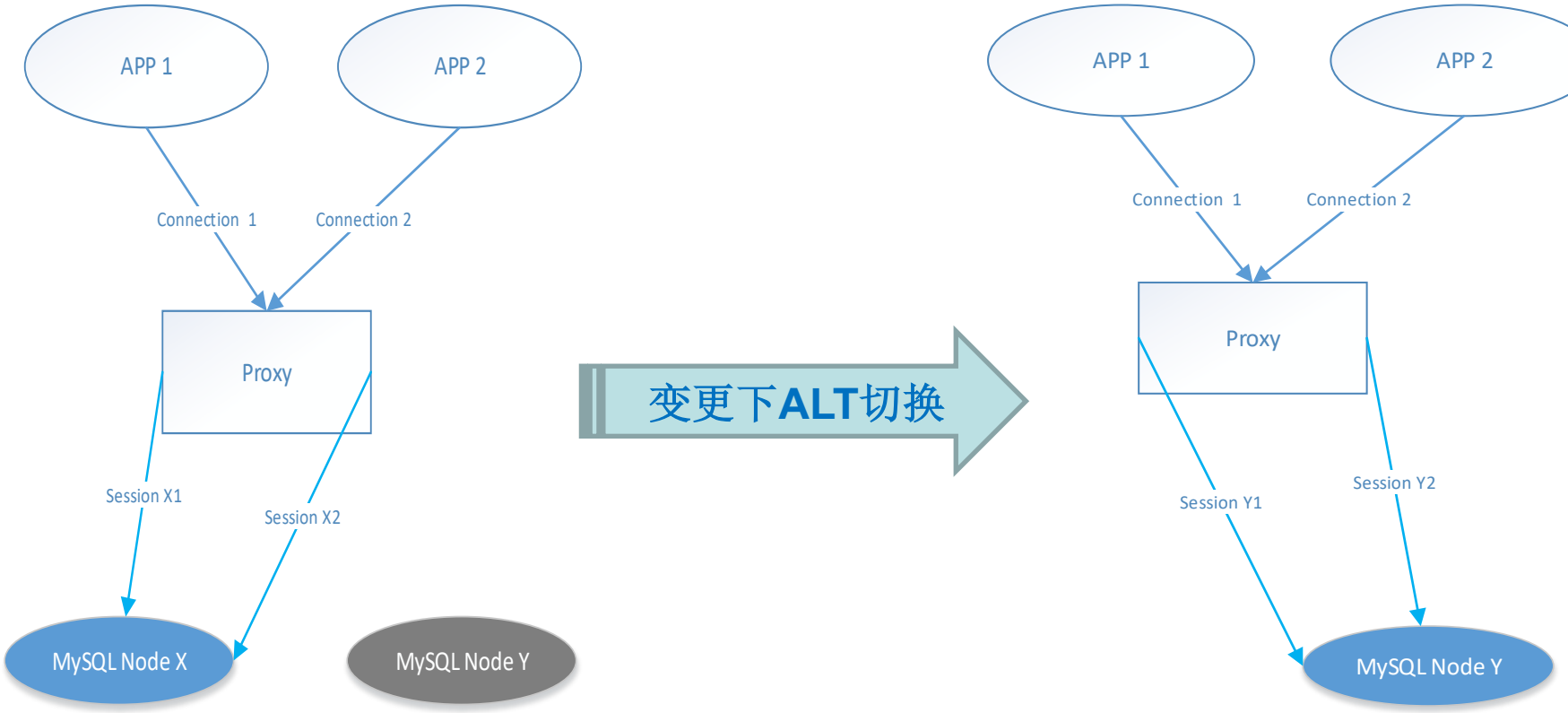
- **Last B-Tree Cursor**

查询继续进行，在新的节点访问BTREE，从上次Scan的位置开始。确保结果集不重复不丢失。为此每次返回结果集都携带Last B-Tree Cursor到Client/Proxy。

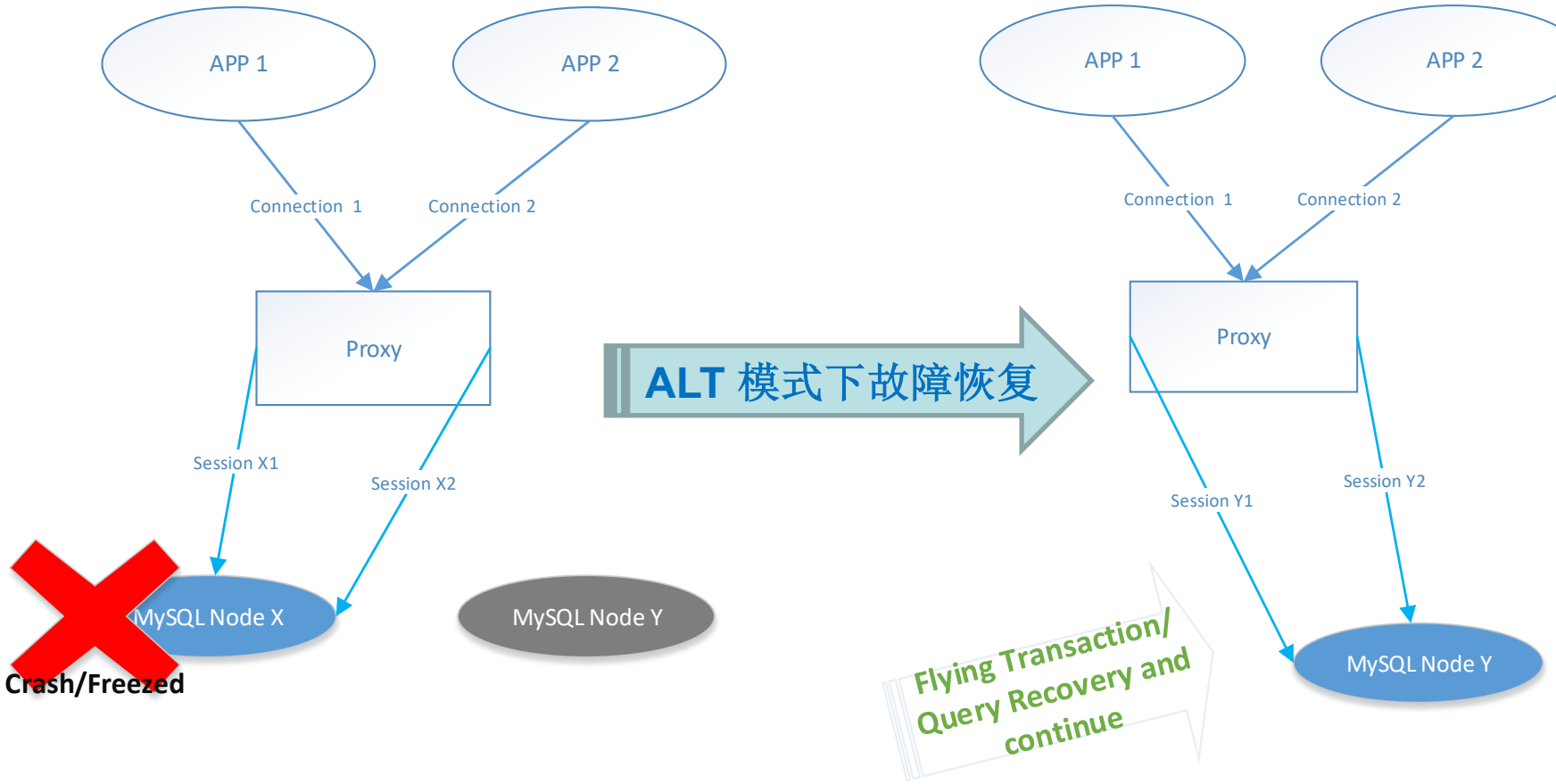


# ALT 架构 -- 切换


APP 1
APP 2
Connection 1
Connection 2
Proxy
Session X1
Session X2
MySQL Node X
MySQL Node Y
变更下ALT切换
APP 1
APP 2
Connection 1
Connection 2
Proxy
Session Y1
Session Y2
MySQL Node Y

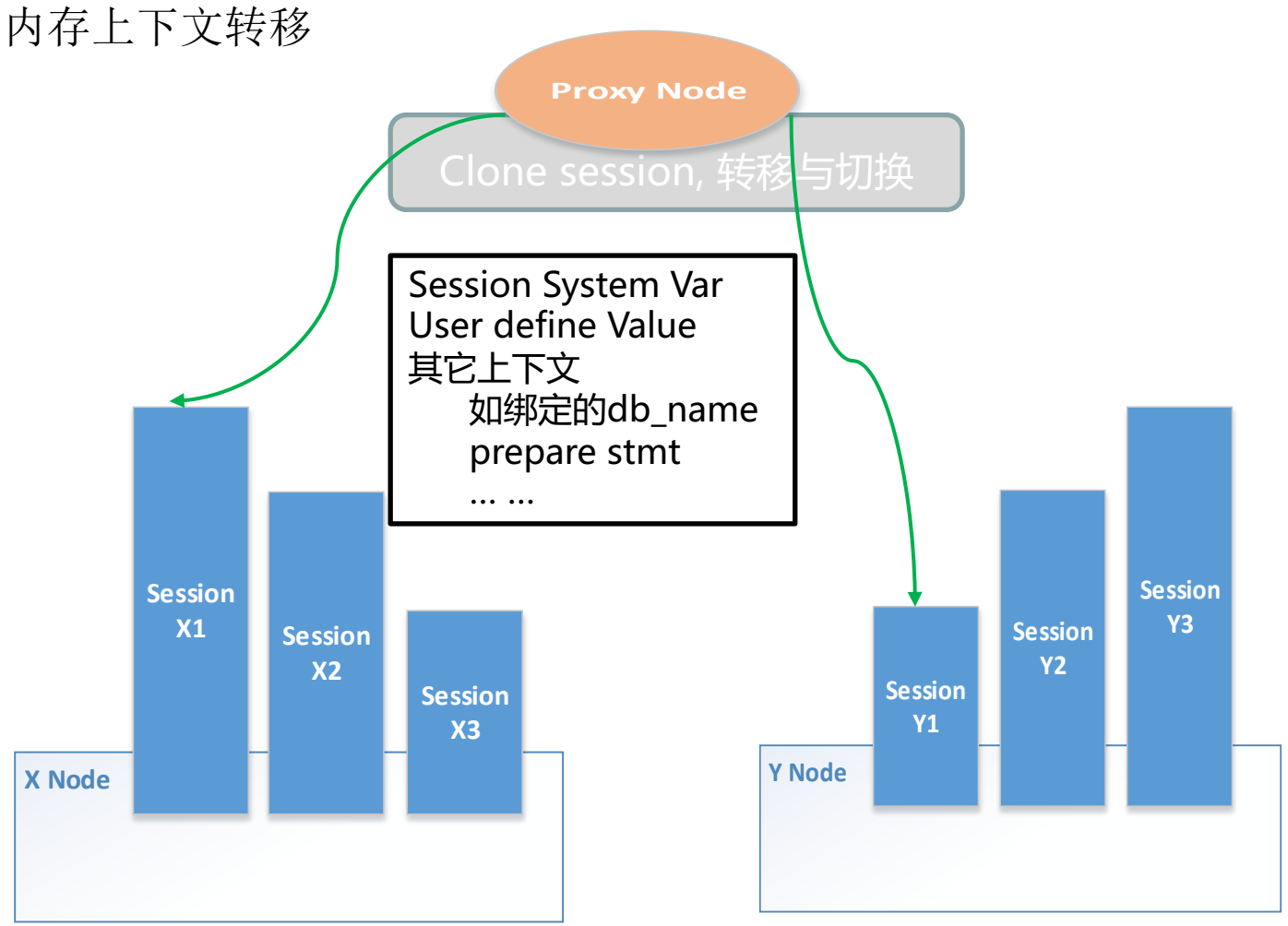
# ALT 架构 -- 故障转移


APP 1
APP 2
Connection 1
Connection 2
Proxy
Session X1
Session X2
MySQL Node X
Crash/Freezed
MySQL Node Y
ALT 模式下故障恢复
APP 1
APP 2
Connection 1
Connection 2
Proxy
Session Y1
Session Y2
MySQL Node Y
Flying Transaction/
Query Recovery and
continue

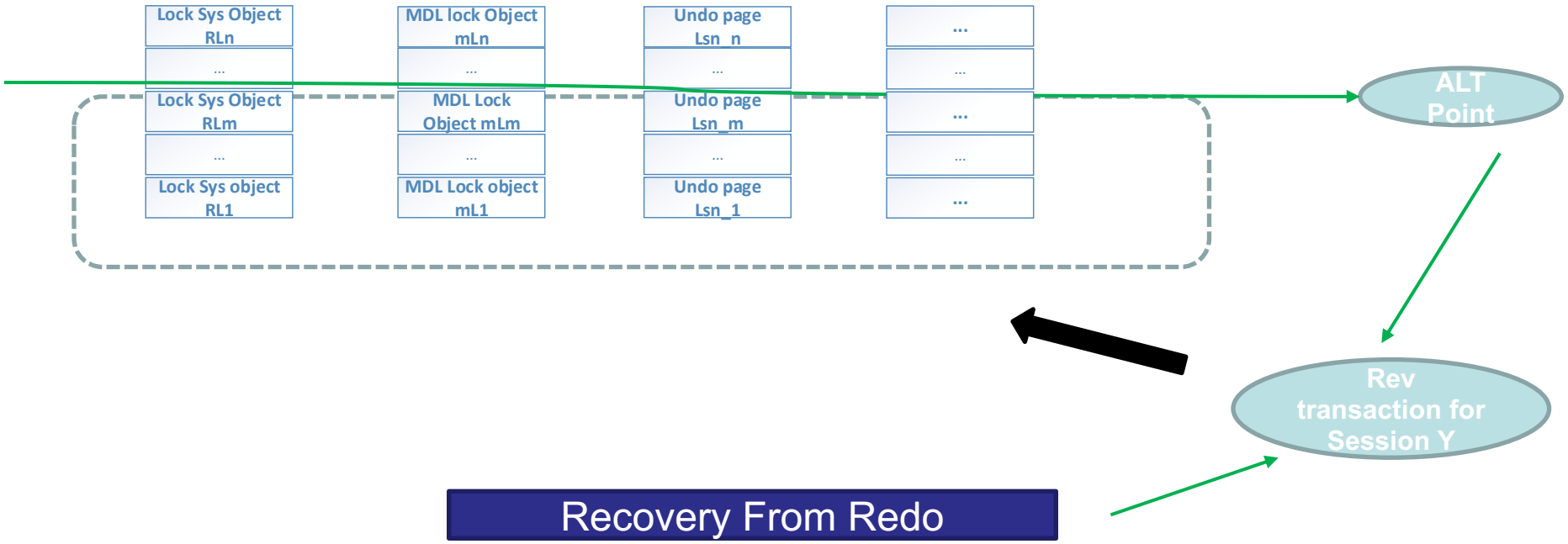
# Drain Session -- 事务排干

- **MySQL主节点上事务排干机制确保达到逻辑的事务安全边界，用户设置 drain_timeout 限定事务排干的超时时间**

- **逻辑的安全事务边界**
  - InnoDB事务块
  - 直接加表锁
  - DDL
  - 用户锁
  - XA事务
  - Lock for backup/binlog
  - ……
  - 最小单位为一个command

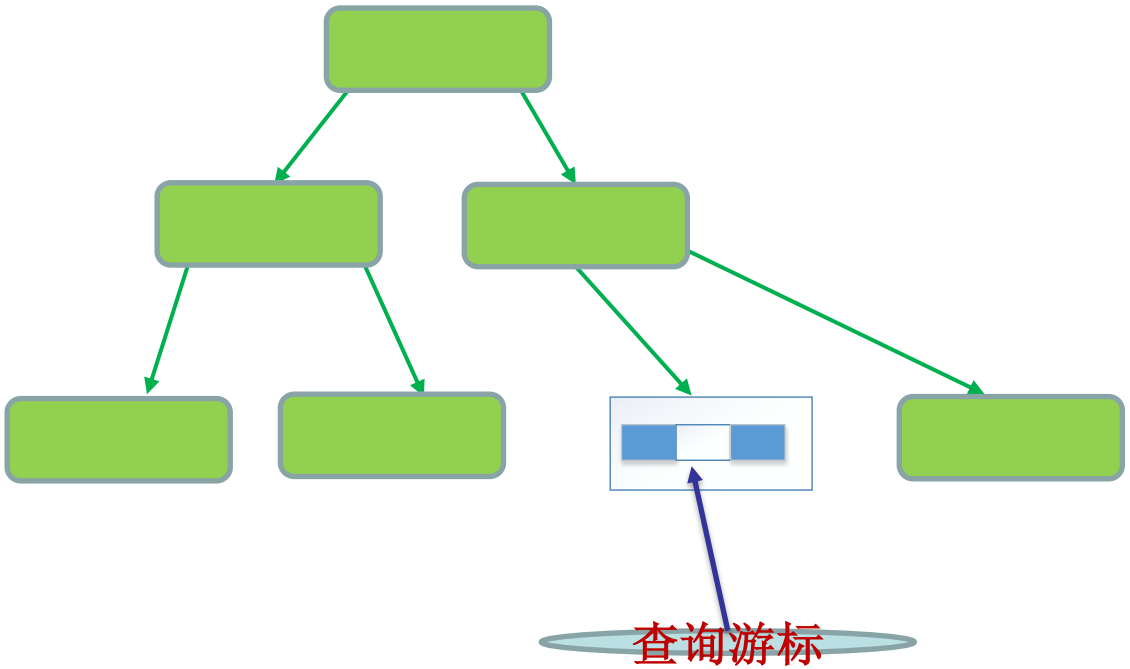


# 事务断点续作


Lock Sys Object RLn
...
Lock Sys Object RLm
...
Lock Sys object RL1
MDL lock Object mLn
...
MDL Lock Object mLm
...
MDL Lock object mL1
Undo page Lsn_n
...
Undo page Lsn_m
...
Undo page Lsn_1
...
...
...
...
ALT Point
Rev transaction for Session Y
Recovery From Redo

# 查询透明连续

- Proxy fetch 数据，记录ALT游标集
- 从上次ALT查询游标集开始扫描数据生成结果集



# TwinSession 切换

- **Strong ALT  Transform**

**应用完全无感， 满足L4**

- **Weak ALT  Transform**

**L1 ~ L3**

- **Normal HA Transform**

**L0**

# 普通模式切换

```
sysbench 1.1.0 (using bundled LuaJIT 2.1.0-beta3)

Running the test with following options:
Number of threads: 16
Report intermediate results every 3 second(s)
Initializing random number generator from current time


Initializing worker threads...

Threads started!

[ 3s ] thds: 16 tps: 21.30 qps: 476.52 (r/w/o: 343.41/85.19/47.92) lat (ms,95%): 831.46 err/s: 0.00 reconn/s: 0.00
[ 6s ] thds: 16 tps: 26.34 qps: 527.17 (r/w/o: 367.12/107.70/52.35) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 9s ] thds: 16 tps: 26.67 qps: 528.68 (r/w/o: 370.01/105.67/53.00) lat (ms,95%): 634.66 err/s: 0.00 reconn/s: 0.00
[ 12s ] thds: 16 tps: 26.00 qps: 524.27 (r/w/o: 365.95/105.99/52.33) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 15s ] thds: 16 tps: 26.67 qps: 531.73 (r/w/o: 372.38/106.01/53.34) lat (ms,95%): 634.66 err/s: 0.00 reconn/s: 0.00
[ 18s ] thds: 16 tps: 26.00 qps: 528.33 (r/w/o: 372.33/103.67/52.33) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 21s ] thds: 16 tps: 26.67 qps: 525.02 (r/w/o: 365.68/106.34/53.00) lat (ms,95%): 657.93 err/s: 0.00 reconn/s: 0.00
[ 24s ] thds: 16 tps: 26.33 qps: 528.62 (r/w/o: 367.97/107.99/52.66) lat (ms,95%): 634.66 err/s: 0.00 reconn/s: 0.00
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest10 SET k=k+1 WHERE id=617'
FATAL: `thread_run' function failed: ./oltp_common.lua:465: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'COMMIT'
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'INSERT INTO sbtest4 (id, k, c, pad) VALUES (504, 531, '75441643858-32061664403-54315910126-01229990993-05871574797-83278955170-38932605958-56572776247-46579995
092619', '61551499371-56857136246-56266348994-57560586781-75384316516')'
FATAL: `thread_run' function failed: ./oltp_common.lua:416: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: `thread_run' function failed: ./oltp_common.lua:495: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'COMMIT'
FATAL: `thread_run' function failed: ./oltp_common.lua:416: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest11 SET c='54173116347-89113967422-01529493787-96832329439-13455350158-56215181069-76371726401-67668291351-52026572528-77749824980' WHERE id=580'
FATAL: `thread_run' function failed: ./oltp_common.lua:476: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest15 SET c='36329479782-30211309926-84814013391-95333779200-20556665876-72393460953-15548184141-28513044642-72025053584-27928387359' WHERE id=500'
FATAL: `thread_run' function failed: ./oltp_common.lua:476: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest13 SET k=k+1 WHERE id=498'
FATAL: `thread_run' function failed: ./oltp_common.lua:465: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest12 SET k=k+1 WHERE id=500'
FATAL: `thread_run' function failed: ./oltp_common.lua:465: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest6 SET k=k+1 WHERE id=518'
FATAL: `thread_run' function failed: ./oltp_common.lua:465: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest11 SET k=k+1 WHERE id=670'
FATAL: `thread_run' function failed: ./oltp_common.lua:465: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest11 SET k=k+1 WHERE id=514'
FATAL: `thread_run' function failed: ./oltp_common.lua:465: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest15 SET k=k+1 WHERE id=450'
FATAL: `thread_run' function failed: ./oltp_common.lua:465: SQL error, errno = 1815, state = 'HY000': refCnt: 0
FATAL: mysql_drv_query() returned error 1815 (refCnt: 0) for query 'UPDATE sbtest3 SET k=k+1 WHERE id=505'
```

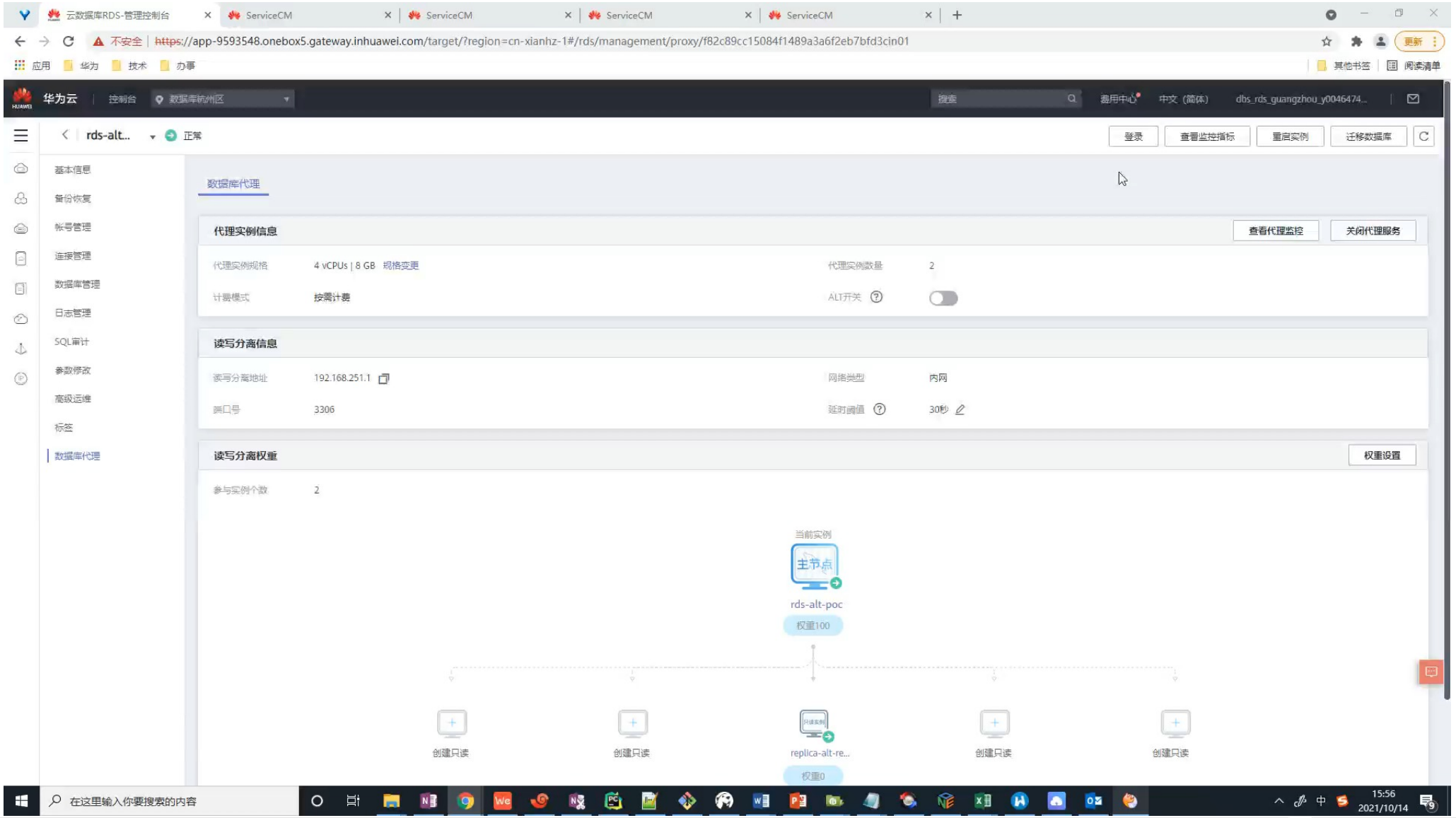
进行普通主备切换

# ALT模式切换

```
sysbench 1.1.0 (using bundled LuaJIT 2.1.0-beta3)

Running the test with following options:
Number of threads: 16
Report intermediate results every 3 second(s)
Initializing random number generator from current time


Initializing worker threads...

Threads started!

[ 3s ] thds: 16 tps: 21.63 qps: 490.13 (r/w/o: 353.70/87.84/48.58) lat (ms,95%): 787.74 err/s: 0.00 reconn/s: 0.00
[ 6s ] thds: 16 tps: 26.70 qps: 519.06 (r/w/o: 359.84/106.15/53.07) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 9s ] thds: 16 tps: 25.31 qps: 514.22 (r/w/o: 359.69/103.58/50.96) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 12s ] thds: 16 tps: 26.00 qps: 514.00 (r/w/o: 360.00/102.00/52.00) lat (ms,95%): 657.93 err/s: 0.00 reconn/s: 0.00
[ 15s ] thds: 16 tps: 26.01 qps: 516.17 (r/w/o: 359.12/105.37/51.68) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 18s ] thds: 16 tps: 24.35 qps: 514.27 (r/w/o: 364.19/101.05/49.03) lat (ms,95%): 657.93 err/s: 0.00 reconn/s: 0.00
[ 21s ] thds: 16 tps: 26.31 qps: 514.24 (r/w/o: 360.03/101.58/52.62) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 24s ] thds: 16 tps: 26.34 qps: 514.38 (r/w/o: 356.03/105.68/52.67) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 27s ] thds: 16 tps: 25.66 qps: 512.22 (r/w/o: 359.92/101.64/50.66) lat (ms,95%): 657.93 err/s: 0.00 reconn/s: 0.00
[ 30s ] thds: 16 tps: 25.33 qps: 517.33 (r/w/o: 363.33/102.67/51.33) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 33s ] thds: 16 tps: 25.34 qps: 516.38 (r/w/o: 362.03/103.68/50.67) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 36s ] thds: 16 tps: 26.00 qps: 515.02 (r/w/o: 360.68/102.34/52.00) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 39s ] thds: 16 tps: 26.33 qps: 513.59 (r/w/o: 358.28/102.98/52.33) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 42s ] thds: 16 tps: 25.67 qps: 510.09 (r/w/o: 355.73/103.02/51.34) lat (ms,95%): 694.45 err/s: 0.00 reconn/s: 0.00
[ 45s ] thds: 16 tps: 25.67 qps: 515.31 (r/w/o: 361.32/102.66/51.33) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 48s ] thds: 16 tps: 25.67 qps: 512.36 (r/w/o: 359.02/102.34/51.00) lat (ms,95%): 657.93 err/s: 0.00 reconn/s: 0.00
[ 51s ] thds: 16 tps: 25.00 qps: 513.99 (r/w/o: 359.99/103.33/50.67) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 54s ] thds: 16 tps: 26.67 qps: 515.99 (r/w/o: 360.66/102.66/52.67) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 57s ] thds: 16 tps: 25.00 qps: 512.36 (r/w/o: 359.69/102.01/50.67) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 60s ] thds: 16 tps: 26.33 qps: 516.63 (r/w/o: 359.64/104.66/52.33) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 63s ] thds: 16 tps: 25.67 qps: 513.35 (r/w/o: 359.68/102.00/51.67) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 66s ] thds: 16 tps: 25.36 qps: 515.12 (r/w/o: 362.65/102.76/49.71) lat (ms,95%): 669.89 err/s: 0.00 reconn/s: 0.00
[ 69s ] thds: 16 tps: 25.31 qps: 512.88 (r/w/o: 361.02/100.25/51.62) lat (ms,95%): 669.89 err/s: 0.00 reconn/s: 0.00
[ 72s ] thds: 16 tps: 26.67 qps: 515.65 (r/w/o: 356.65/106.00/53.00) lat (ms,95%): 646.19 err/s: 0.00 reconn/s: 0.00
[ 75s ] thds: 16 tps: 25.66 qps: 512.63 (r/w/o: 356.97/104.33/51.33) lat (ms,95%): 657.93 err/s: 0.00 reconn/s: 0.00
[ 78s ] thds: 16 tps: 25.00 qps: 512.73 (r/w/o: 362.71/99.68/50.34) lat (ms,95%): 657.93 err/s: 0.00 reconn/s: 0.00
[ 81s ] thds: 16 tps: 26.00 qps: 512.61 (r/w/o: 357.96/102.66/51.99) lat (ms,95%): 669.89 err/s: 0.00 reconn/s: 0.00
^C
```

进行ALT切换

# 演示


云数据库RDS-管理控制台
ServiceCM
ServiceCM
ServiceCM
ServiceCM
不安全 | https://app-9593548.onebox5.gateway.inhuawei.com/target/?region=cn-xianhz-1#/rds/management/proxy/f82c89cc15084f1489a3a6f2eb7bfd3cin01
更新
应用
华为
技术
办事
其他书签
阅读清单
华为云
控制台
数据库杭州区
搜索
费用中心
中文（简体）
dbs_rds_guangzhou_y0046474...
rds-alt...
正常
登录
查看监控指标
重启实例
迁移数据库
基本信息
备份恢复
帐号管理
连接管理
数据库管理
日志管理
SQL审计
参数修改
高级运维
标签
数据库代理
数据库代理
代理实例信息
查看代理监控
关闭代理服务
代理实例规格
4 vCPUs | 8 GB 规格变更
代理实例数量
2
计费模式
按需计费
ALT开关
读写分离信息
读写分离地址
192.168.251.1
网络类型
内网
端口号
3306
延时阈值
30秒
读写分离权重
权重设置
参与实例个数
2
当前实例
主节点
rds-alt-poc
权重100
创建只读
创建只读
replica-alt-re...
权重0
创建只读
创建只读
在这里输入你要搜索的内容
15:56
2021/10/14

# 重新定义HA

- 传统HA衡量指标 RTO，只是衡量数据库系统提供服务的间隔时间，但无法衡量真实的业务影响

- HA应当体现对用户的受损程度，以及针对各个连接，多久能正常使用

- 站在用户的角度重新定义高可用性指标

1. (AIR)Application Impairment Ratio 应用受损比率

能满足ALT条件切换的connection的比率， 0% <= ALR <= 100%

2. (AFD)Application Freeze Duration 应用卡顿时间

应用一条命令执行在ALT过程中，被延长的执行时间

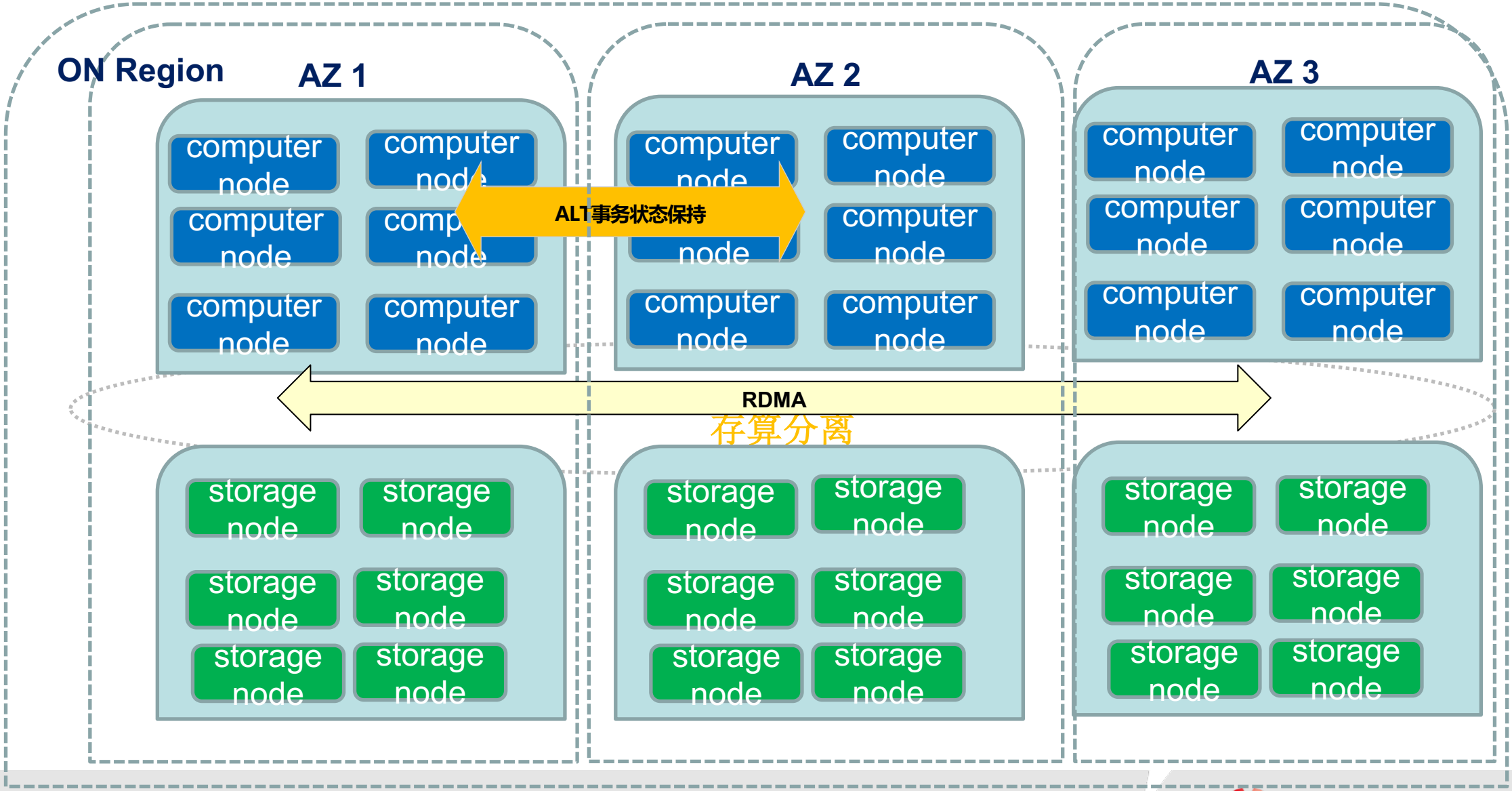
--- Max AFD

--- Average AFD

# ALT支持MySQL云原生ServerLess

MySQL实例通过存算分离技术做到存储弹性可扩展；

ServerLess下MySQL实例计算资源弹性可扩展，**通过ALT事务状态保持。**



# ALT软硬件结合能力

- Optane memory
- RDMA and Global memory
- 实时生成 ALT Point 快照成为可能
- 支持FailOver下全量ALT能力