新手机 新应用 新娱乐

PostgreSQL PL/Proxy 原理与实践

Digoal.Zhou

10/27/2011

SKY mobi
新手机 新应用 新娱乐

目录

- 背景
- PL/Proxy运行原理
- 语法
- 应用场景设计
- 性能测试
- HA探讨
- 讨论

SKY mobi
新手机 新应用 新娱乐

背景

- 大数据量
- 大并发读写
- 响应速度要求高

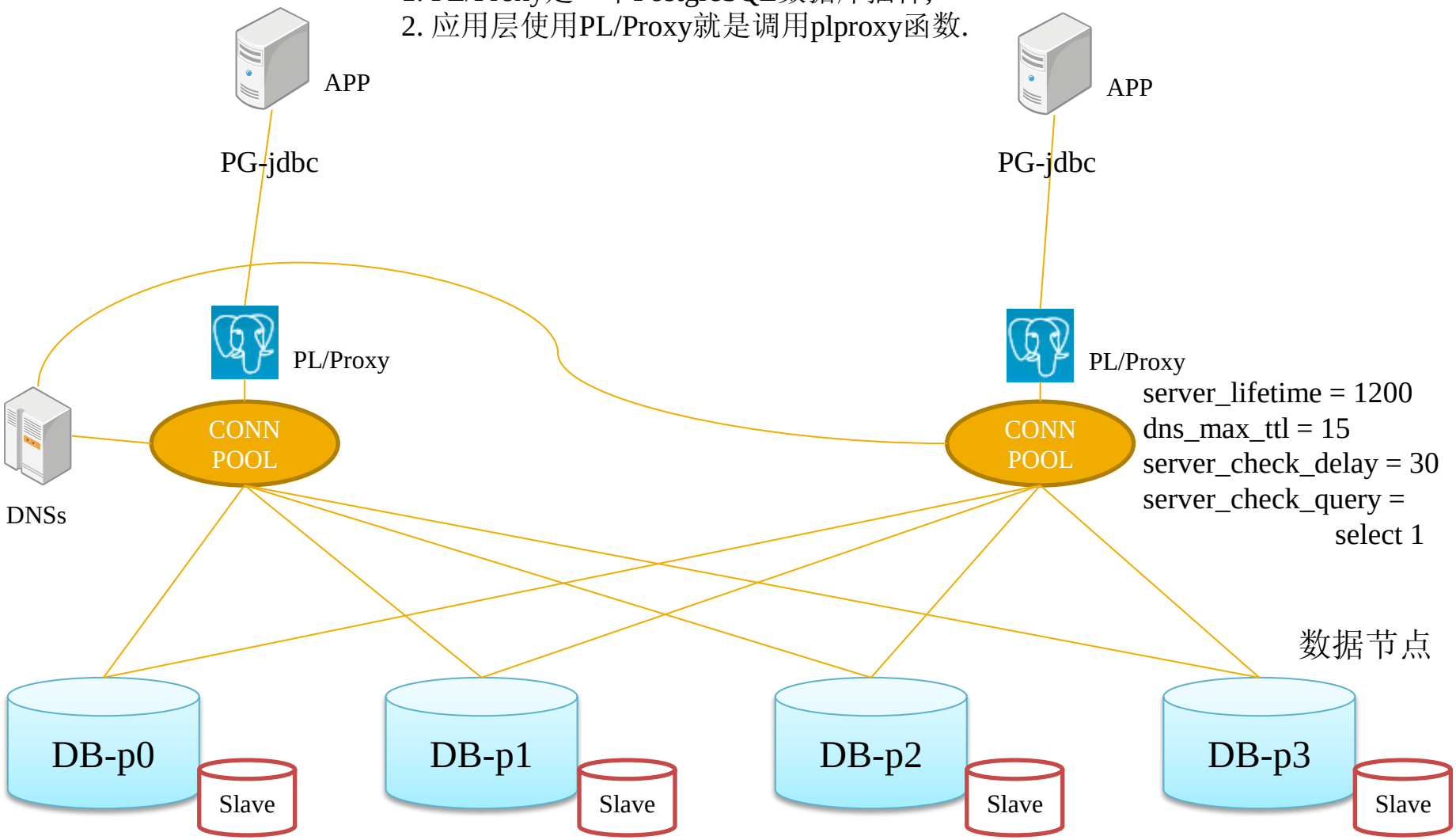
- 传统解决方案
  - 硬件扩容
  - Key-Value 数据最终一致
  - 应用层缓存，数据库复制，对读有效.
- 缺点
  - 成本昂贵，投入产出比不高。
  - 丢失了事务，适应范围小。
  - 无法解决写需求。

- PL/Proxy
  - 水平扩展，对硬件要求不高
  - PostgreSQL函数操作具有原子性，不违背事务原则
  - PL/Proxy路由选择非常灵活，非常容易做到读写的负载均衡。

SKY mobi
新手机 新应用 新娱乐

PL/Proxy运行原理

1. PL/Proxy是一个PostgreSQL数据库插件,
2. 应用层使用PL/Proxy就是调用plproxy函数.



SKY mobi
新手机 新应用 新娱乐

PL/Proxy运行原理

- PL/Proxy
  - 接收应用程序发起的SQL请求(调用plproxy函数),
  - 解析为提交给数据节点的SQL,
  - 旁路(CONNECT模式)或者选择数据节点(CLUSTER模式),
  - (CLUSTER模式1)查询SQL/MED配置的集群信息,选择数据节点通过libpq async API发送解析的SQL给数据节点(多个则并行),等待所有数据节点返回结果,返回结果给应用程序.
  - (CLUSTER模式2)查询集群配置版本,是否更新集群配置缓存,选择数据节点通过libpq async API发送解析的SQL给数据节点(多个则并行),等待所有数据节点返回结果,返回结果给应用程序.
- 连接池
  - 提高连接效率,复用连接.
- 数据节点
  - 存放实体数据,接收并执行plproxy发送的SQL请求,将执行结果返回给plproxy.

SKY mobi
新手机 新应用 新娱乐

PL/Proxy语法

- plproxy函数的用法。
  - 应用程序传递参数，选择远程数据库节点，参数传递给远程数据库相同名字和参数类型的函数执行，收集数据节点返回结果发送给应用程序。
  - 在PLProxy函数中直接写SELECT查询，收集返回结果发送给应用程序。
- PL/Proxy 函数支持的命令.
- CONNECT
  - CONNECT 'libpq connstr' ; | connect_func(...) | argname
- CLUSTER, [RUN ON ALL|ANY|int2,4,8]
  - CLUSTER 'cluster_name'; | cluster_func(..)
- SELECT
- SPLIT
- TARGET

SKY mobi
新手机 新应用 新娱乐

cluster数据节点选择算法

```
tag_run_on_partitions(ProxyFunction *func, FunctionCallInfo fcinfo, int tag,
                                        DatumArray **array_params, int array_row)
{
        ProxyCluster    *cluster = func->cur_cluster;
        int                             i;

        switch (func->run_type)
        {
                case R_HASH:
                        tag_hash_partitions(func, fcinfo, tag, array_params, array_row);
                        break;
                case R_ALL:
                        for (i = 0; i < cluster->part_count; i++)
                                cluster->part_map[i]->run_tag = tag;
                        break;
                case R_EXACT:
                        i = func->exact_nr;
                        if (i < 0 || i >= cluster->part_count)
                                plproxy_error(func, "part number out of range");
                        cluster->part_map[i]->run_tag = tag;
                        break;
                case R_ANY:
                        i = random() & cluster->part_mask;
                        cluster->part_map[i]->run_tag = tag;
                        break;
                default:
                        plproxy_error(func, "uninitialized run_type");
        }
}
```

SKY mobi
新手机 新应用 新娱乐

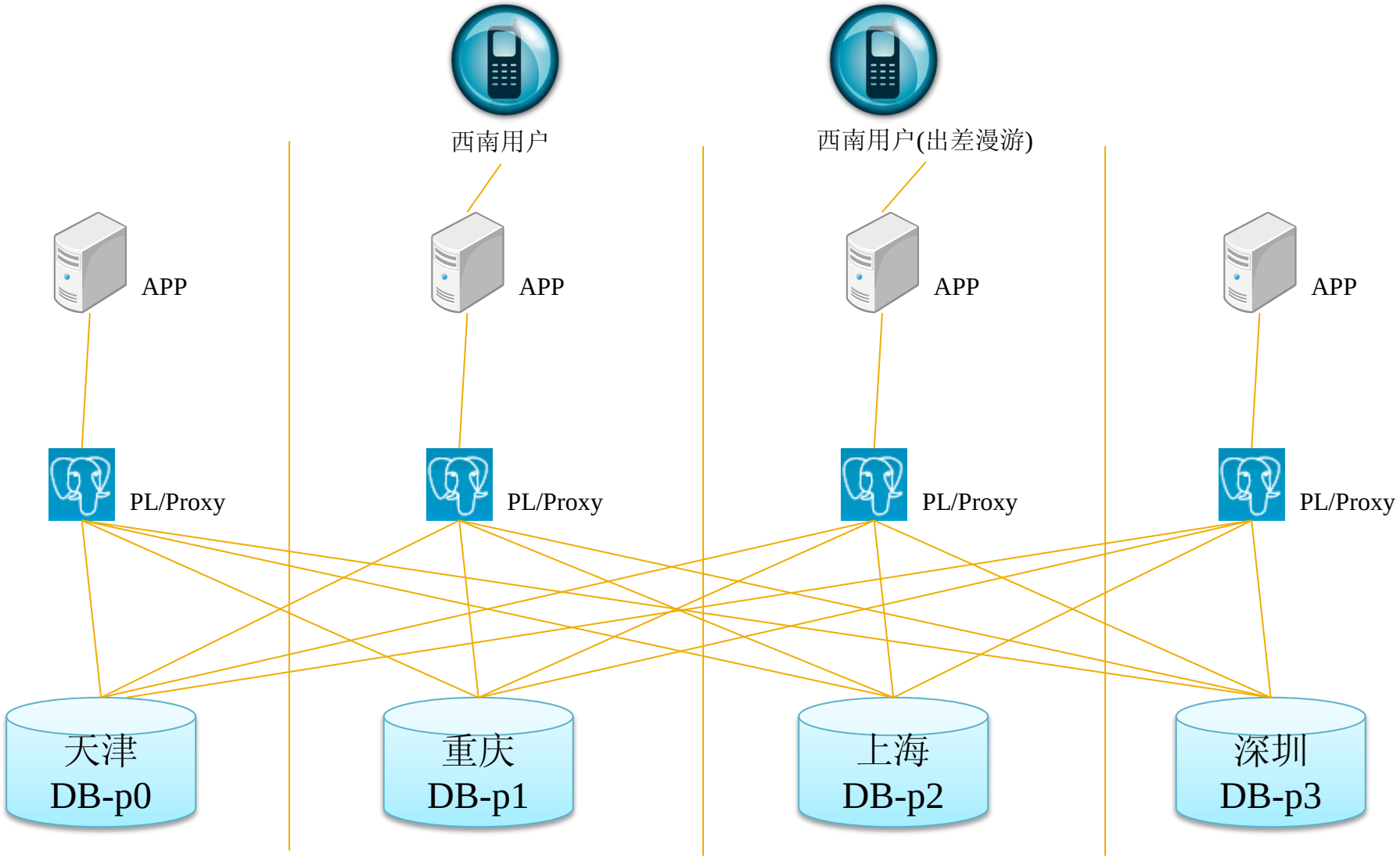
PL/Proxy使用注意

- 跨数据节点的操作需要应用层来处理
  - 如，跨数据节点Foreign Key，跨数据节点的JOIN。
- 跨数据节点的写操作要保持事务一致性需要特殊处理
  - 如，PgQ
  - http://wiki.postgresql.org/wiki/PGQ_Tutorial

SKY mobi
新手机 新应用 新娱乐

应用场景1

- 地域分布,用户信息就近存储

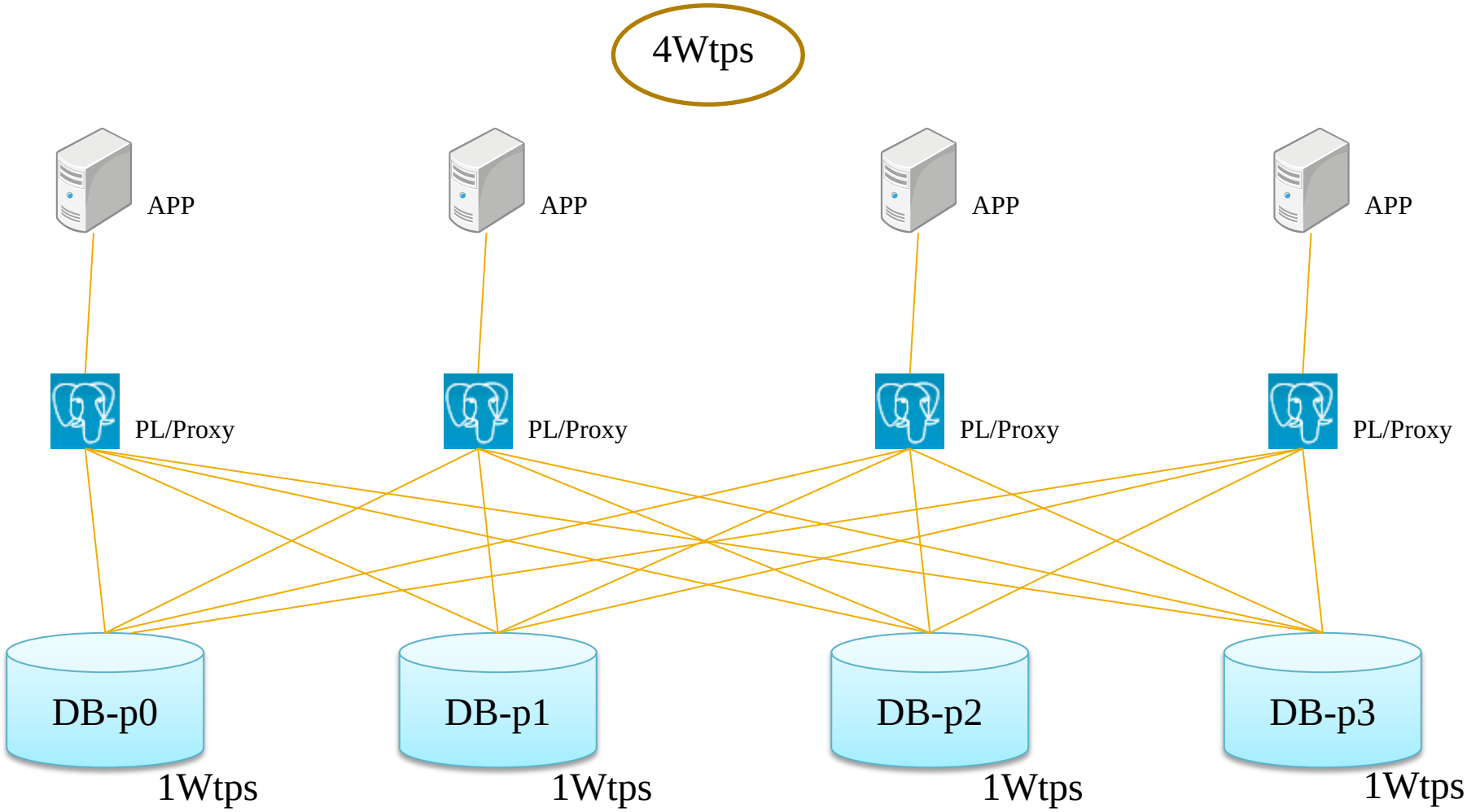


SKY mobi
新手机 新应用 新娱乐

应用场景2

- 用户ID分布

1. USERID : (或者其他能转换成int2,4,8的类型)
2. 与cluster进行 bit & 运算得到运行节点

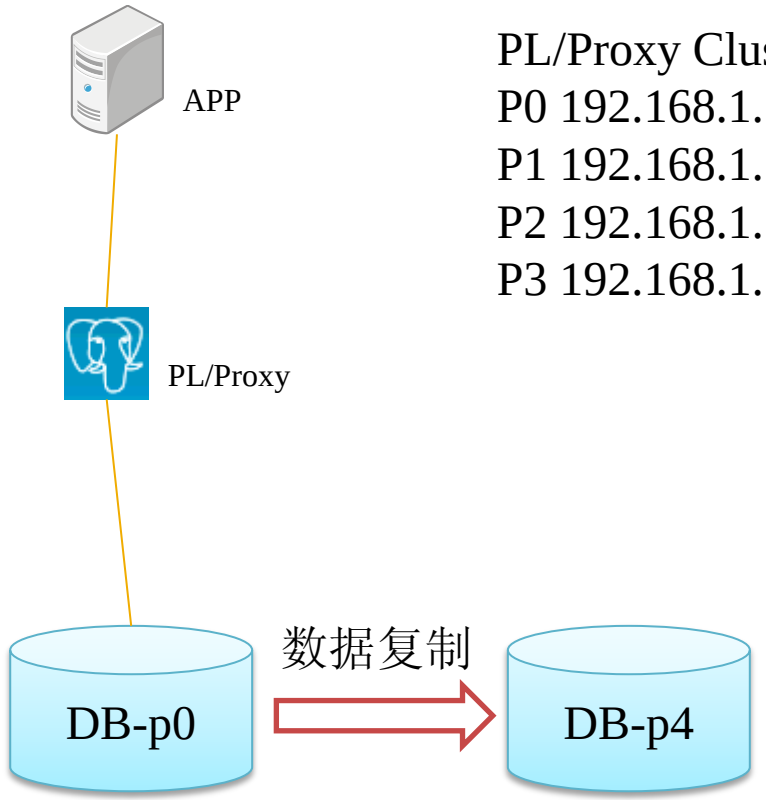


SKY mobi
新手机 新应用 新娱乐

应用场景3

- 扩容. 如, 扩其中一台.

0	1	2	3
4	5	6	7



PL/Proxy Cluster(old)
P0 192.168.1.10
P1 192.168.1.11
P2 192.168.1.12
P3 192.168.1.13

PL/Proxy Cluster(new)
P0 192.168.1.10
P1 192.168.1.11
P2 192.168.1.12
P3 192.168.1.13
P4 192.168.1.14
P5 192.168.1.11
P6 192.168.1.12
P7 192.168.1.13

SKY mobi
新手机 新应用 新娱乐

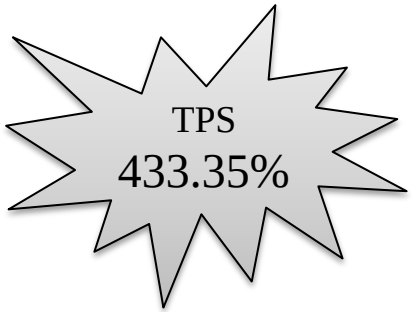
性能测试

- 数据模型
  - 用户账户信息，5000W
  - 用户游戏账户信息，5000W
- 操作
  - 根据PK更新用户账户信息，根据索引更新用户游戏账户信息
- 单节点压力测试结果

	更新用户账户	更新用户游戏账户	总计
TPS	3017	4371	7388
平均响应时间(毫秒)	2.3	1.4	1.85

- 4节点PL/Proxy压力测试结果

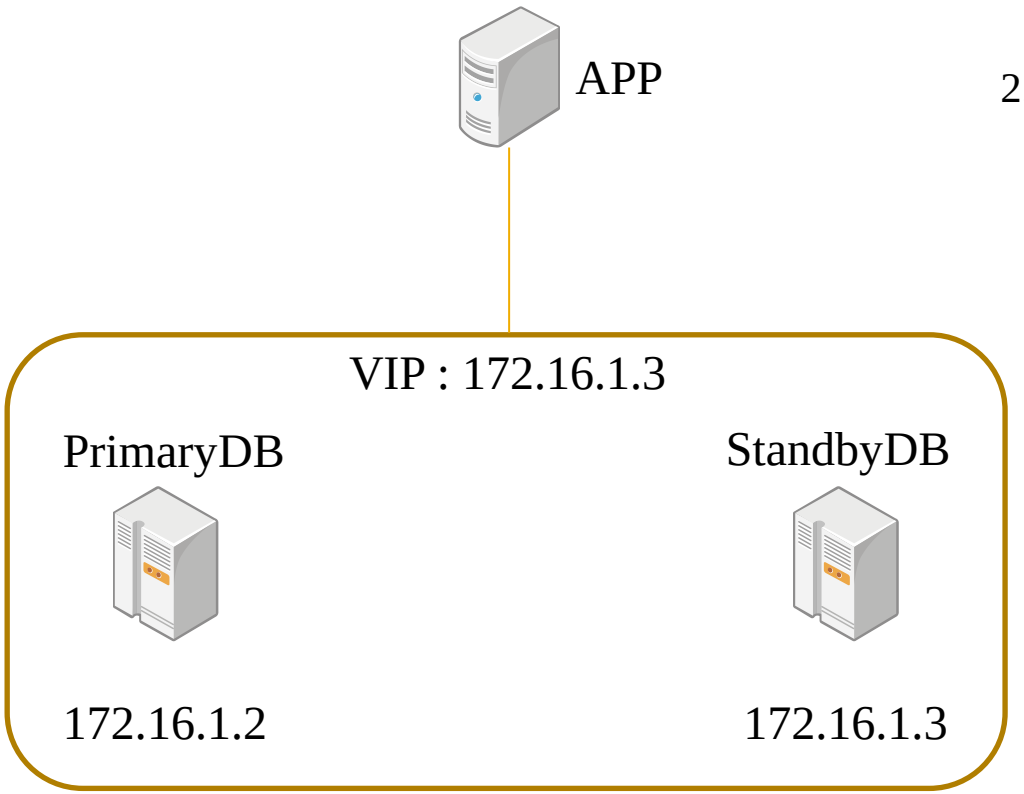
	更新用户账户	更新用户游戏账户	总计
TPS	15226	16790	32016
平均响应时间(毫秒)	2.1	1.9	2



- 参考
- http://blog.163.com/digoal@126/blog/static/16387704020119253563089 5/

HA探讨

- IP漂移



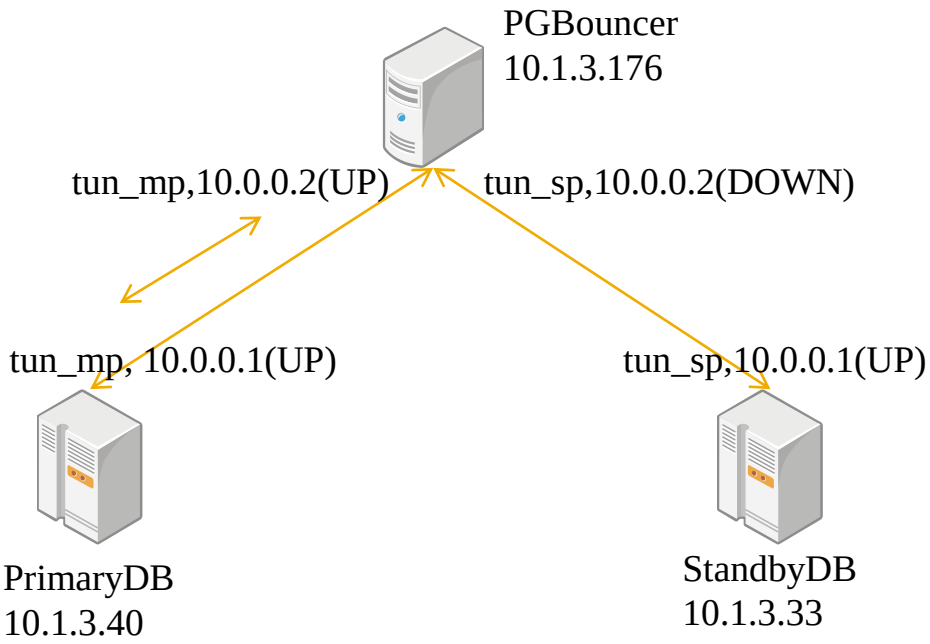
1. 有可能出现VIP冲突,
   如Standby认为PrimaryDOWN了
   实际上没DOWN, 只是心跳不通。
2. 主库和备库必须在同一网段

HA探讨

- 隧道

建立隧道：
PrimaryDB
ip tunnel add tun_mp mode ipip remote 10.1.3.176 local 10.1.3.40
StandbyDB
ip tunnel add tun_sp mode ipip remote 10.1.3.176 local 10.1.3.33
PGBouncer
ip tunnel add tun_mp mode ipip remote 10.1.3.40 local 10.1.3.176
ip tunnel add tun_sp mode ipip remote 10.1.3.33 local 10.1.3.176

隧道IP配置：
PrimaryDB
ifconfig tun_mp up 10.0.0.1/24
StandbyDB
ifconfig tun_sp up 10.0.0.1/24
PGBouncer
ifconfig tun_mp up 10.0.0.2/24



Failover
ifconfig tun_mp down
ifconfig tun_sp up 10.0.0.2/24

Failback
ifconfig tun_sp down
ifconfig tun_mp up 10.0.0.2/24

SKY mobi
新手机 新应用 新娱乐

HA探讨

- 隧道(续)

```
[root@db5 ~]# route
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
10.0.0.0        *               255.255.255.0   U     0      0        0 tun_mp
```

```
[root@db5 ~]# ifconfig tun_sp up 10.0.0.2/24
[root@db5 ~]# route -n
Kernel IP routing table
Destination     Gateway         Genmask         Flags Metric Ref    Use Iface
10.0.0.0        0.0.0.0         255.255.255.0   U     0      0        0 tun_sp
```

测试隧道切换

```
#!/bin/bash

for ((i=0;i<1;i=0))
do
sleep 2
ifconfig tun_mp down
ifconfig tun_sp 10.0.0.2/24
sleep 2
ifconfig tun_sp down
ifconfig tun_mp 10.0.0.2/24
done
```

SKY mobi
新手机 新应用 新娱乐

HA探讨

- 隧道切换测试(续)

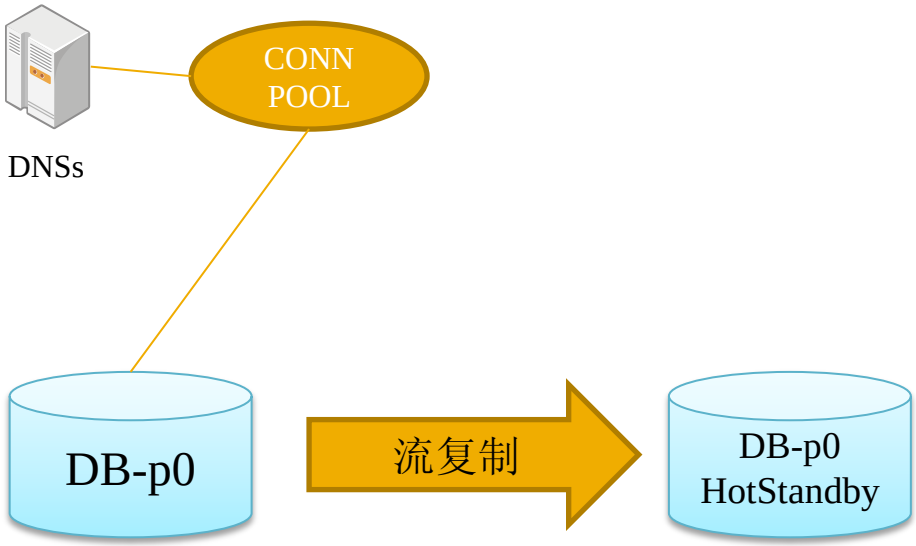
```
digoal=> select count(*) from pg_stat_activity;
ERROR:  server conn crashed?
ERROR:  server conn crashed?
The connection to the server was lost. Attempting reset: Succeeded.
digoal=> select count(*) from pg_stat_activity;
 count
-------
    21
(1 row)
```

SKY mobi
新手机 新应用 新娱乐

HA探讨

- DNS

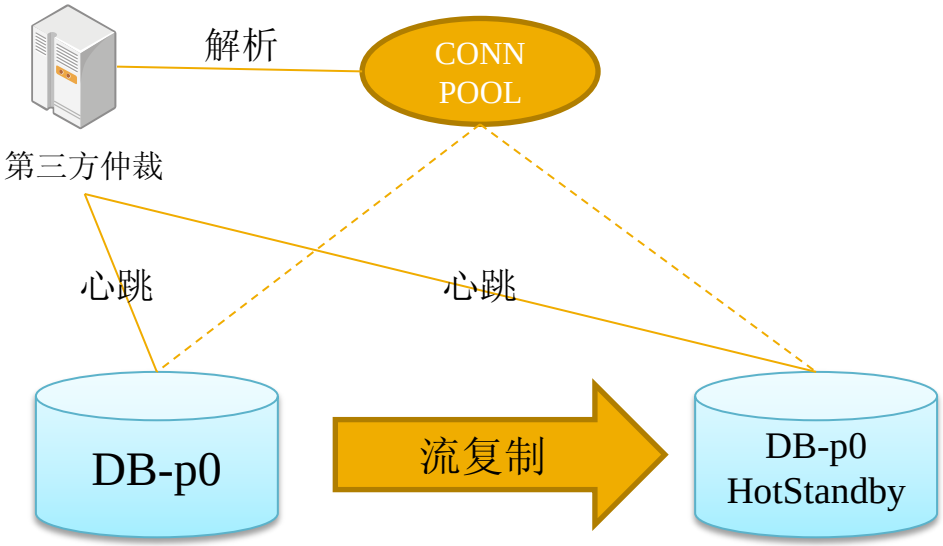
```
[databases]
digoal = host=$HOSTNAME dbname=digoal port=1921 pool_size=16
server_lifetime = 1200
dns_max_ttl = 15
server_check_delay = 30
server_check_query =
                    select 1
```



SKY mobi
新手机 新应用 新娱乐

HA探讨

- 第三方仲裁, 如vFabric Database Direct



SKY mobi
新手机 新应用 新娱乐

讨论

- Thanks !

- 参考
  - http://pgfoundry.org/projects/plproxy

- Author : Digoal.Zhou
- MSN : zzzqware@hotmail.com
- QQ : 276732431
- Email : digoal@126.com
- BLOG : http://blog.163.com/digoal@126/