新手机 新应用 新娱乐

PostgreSQL Enteprise Appliance

Digoal.Zhou

7/6/2011

SKY mobi
新手机 新应用 新娱乐

Database LifeCycle

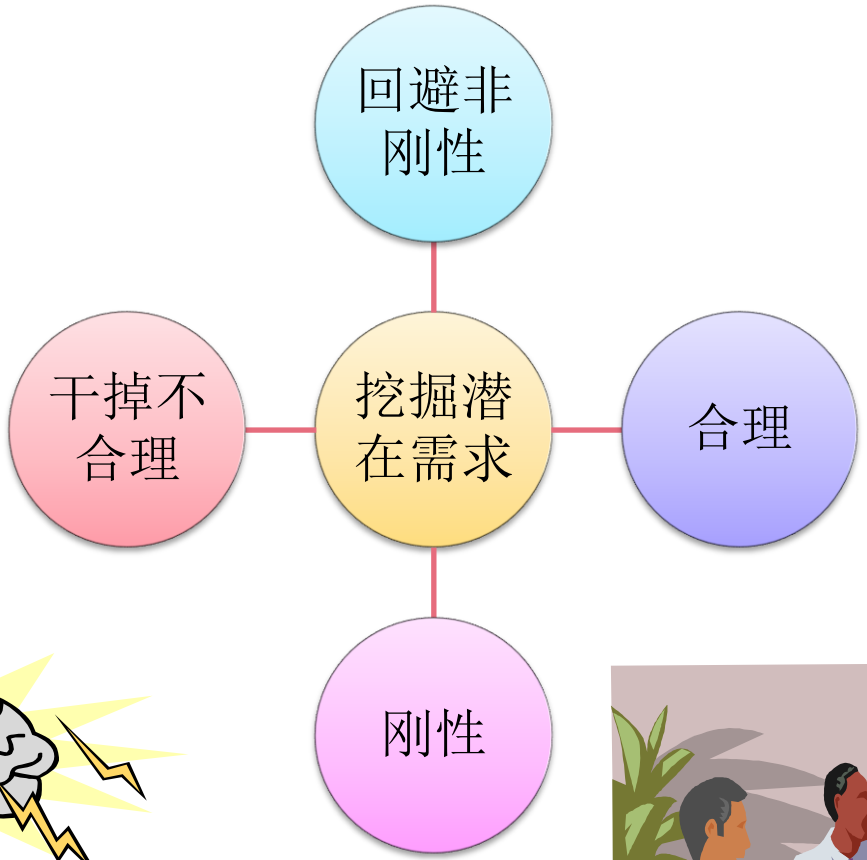
DBA

SKY mobi
新手机 新应用 新娱乐

用户就是上帝

SKY mobi
新手机 新应用 新娱乐

需求阶段



头脑风暴

需求评审

SKY mobi
新手机 新应用 新娱乐

设计阶段

- 业务类型例举浅析
  - APPStore型(静态)
    - 资源基础信息
    - 资源排行榜
    - 资源打分,评论等
    - 包月类信息
    - 活跃数据<内存大小
  - SNS型(动态)
    - 个人属性
    - 社会关系(好友，群组)
    - 分享(微博，即时消息，图片，视频，音乐，博文)
    - 磁性应用(社区网游)
    - 活跃数据>内存大小
  - 网游型(动态)
    - 个人属性(经验值,等级,装备,金钱…)
    - 关系(好友，群组)

SKY mobi
新手机 新应用 新娱乐

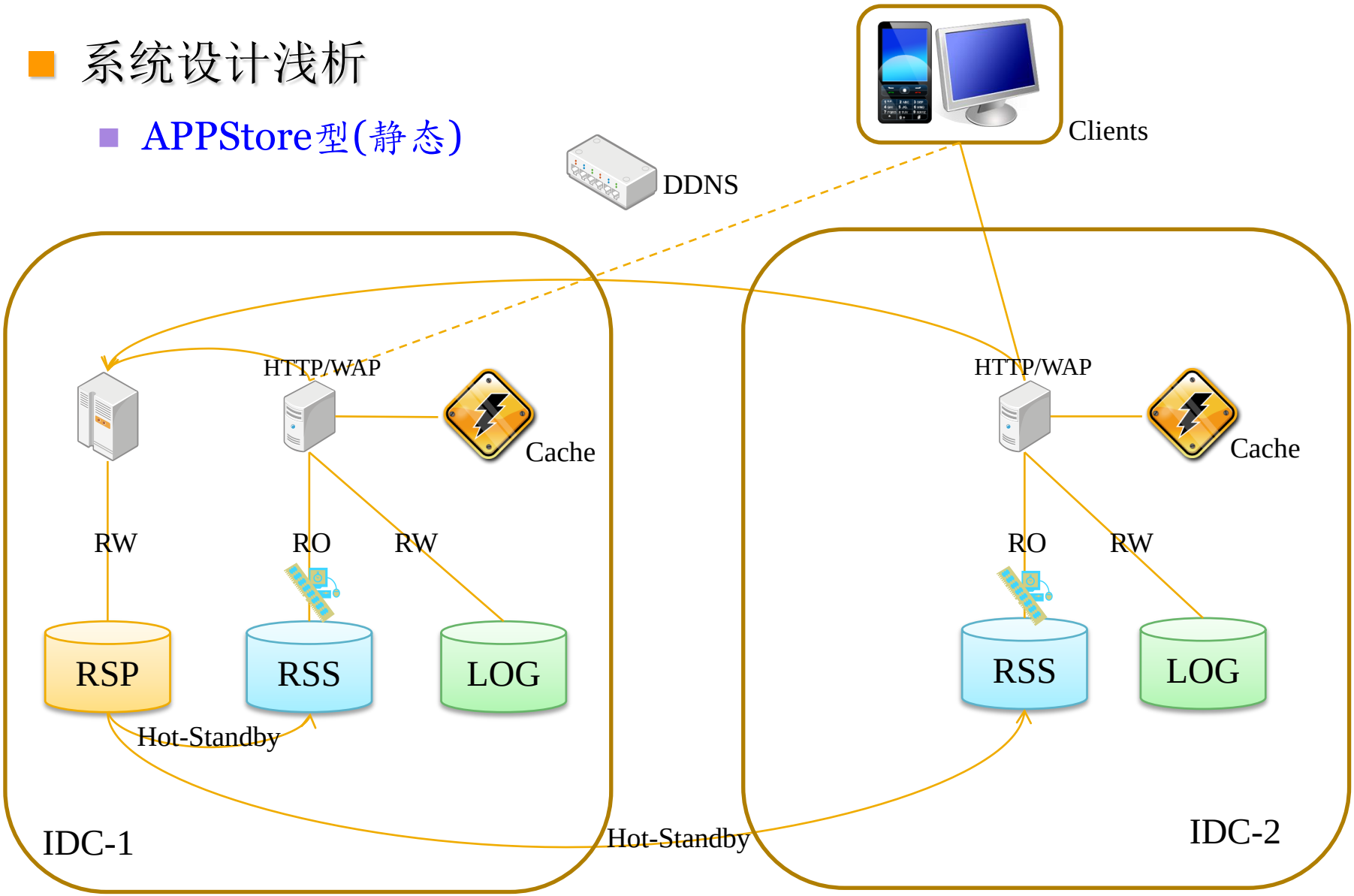
设计阶段

- 数据库挑战浅析
  - APPStore型(静态)
    - 读请求(70%)
      - 白名单，黑名单，资源信息，包月信息等
    - 写请求(30%)
      - 资源信息增删改，新增日志
    - 请求数到达数据库极限
    - 应用层缓存刷新或失效
    - 硬件故障
    - IDC故障
    - 用户体验

SKY mobi
新手机 新应用 新娱乐

设计阶段

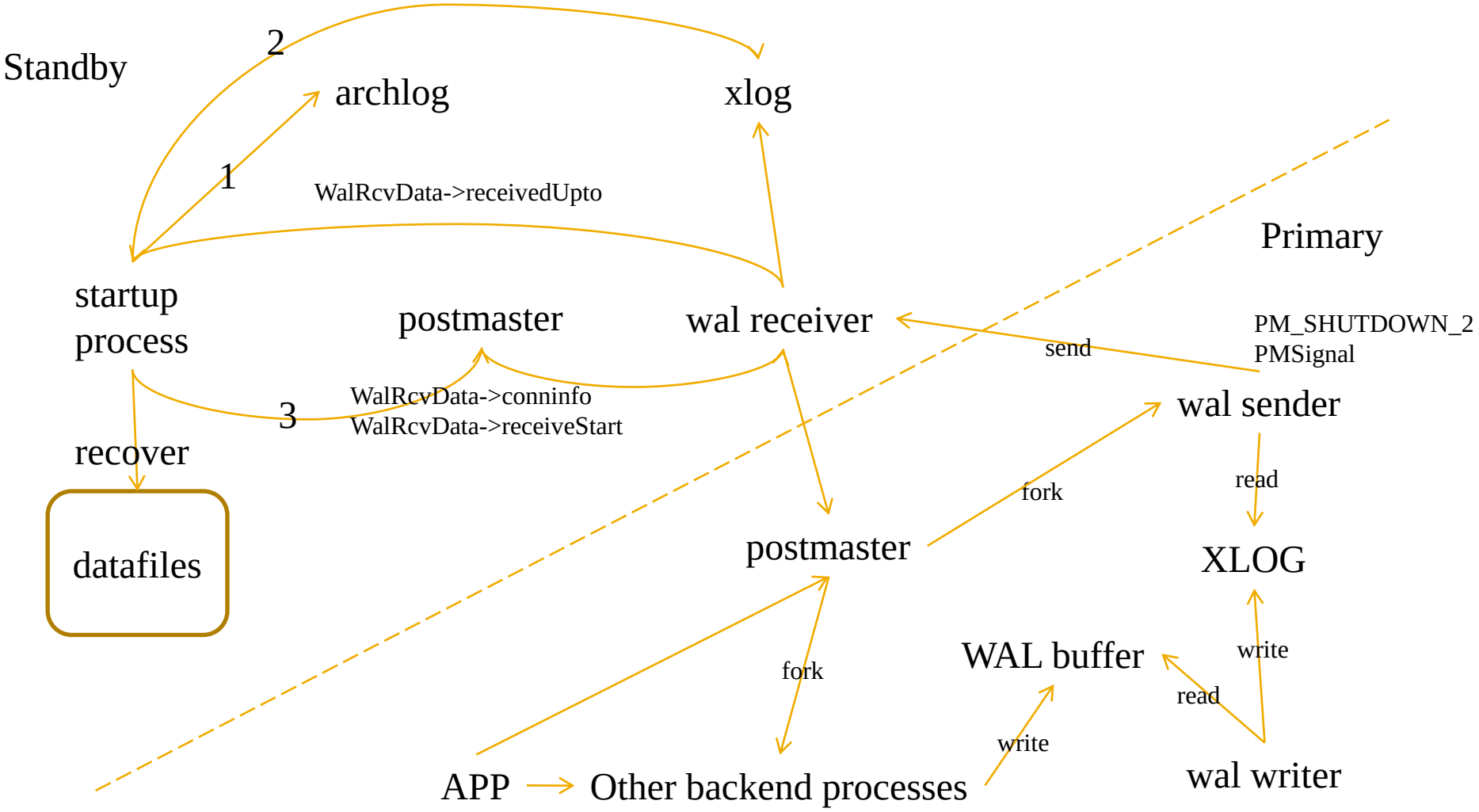
- 系统设计浅析
  - APPStore型(静态)



SKY mobi
新手机 新应用 新娱乐

设计阶段

系统设计浅析 – 异步流复制介绍



SKY mobi
新手机 新应用 新娱乐

设计阶段

Parameter Tuning :

Primary

max_wal_senders

wal_sender_delay (The sleep is interrupted by transaction commit)

wal_keep_segments

vacuum_defer_cleanup_age (the number of transactions by which VACUUM and HOT updates will defer cleanup of dead row versions.)

Standby

hot_standby

wal apply & SQL on standby conflict reference parameter

max_standby_archive_delay

(the maximum total time allowed to apply any one WAL segment's data.)

max_standby_streaming_delay

(the maximum total time allowed to apply WAL data once it has been received from the primary server)

wal_receiver_status_interval

(minimum frequency, The standby will report the last transaction log position it has written, the last position it has flushed to disk, and the last position it has applied.)

hot_standby_feedback

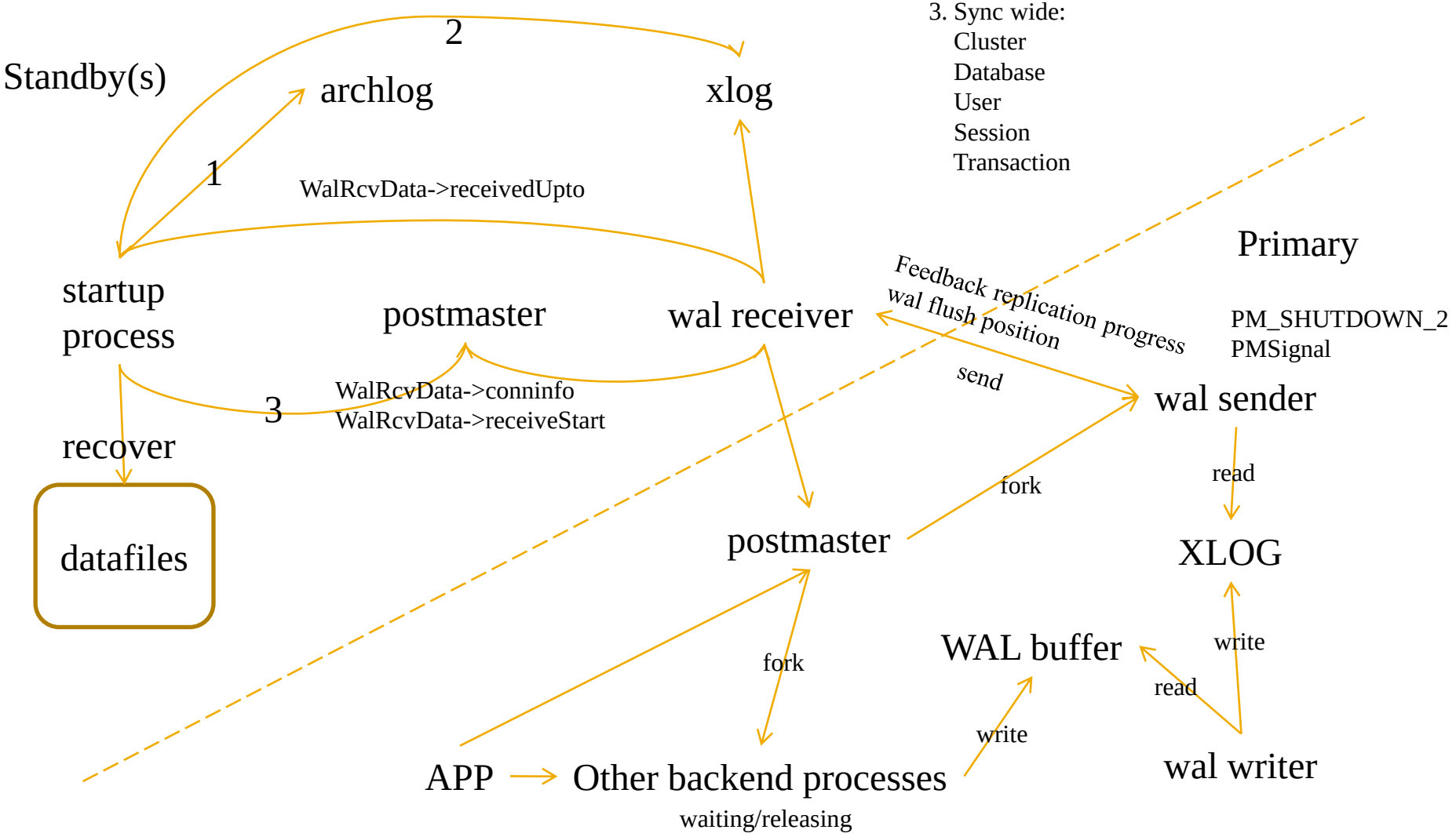
(send feedback to the primary about queries currently executing on the standby.)

SKY mobi
新手机 新应用 新娱乐

设计阶段

系统设计浅析 – 同步流复制介绍

1. Many standbys
2. Only one sync standby at the same time
3. Sync wide:
   Cluster
   Database
   User
   Session
   Transaction



SKY mobi
新手机 新应用 新娱乐

设计阶段

Parameter Tuning :

Primary

- max_wal_senders
- wal_sender_delay
- wal_keep_segments
- vacuum_defer_cleanup_age
- synchronous_replication
- synchronous_standby_names
  - (primary_conninfo in standby's primary_conninfo)

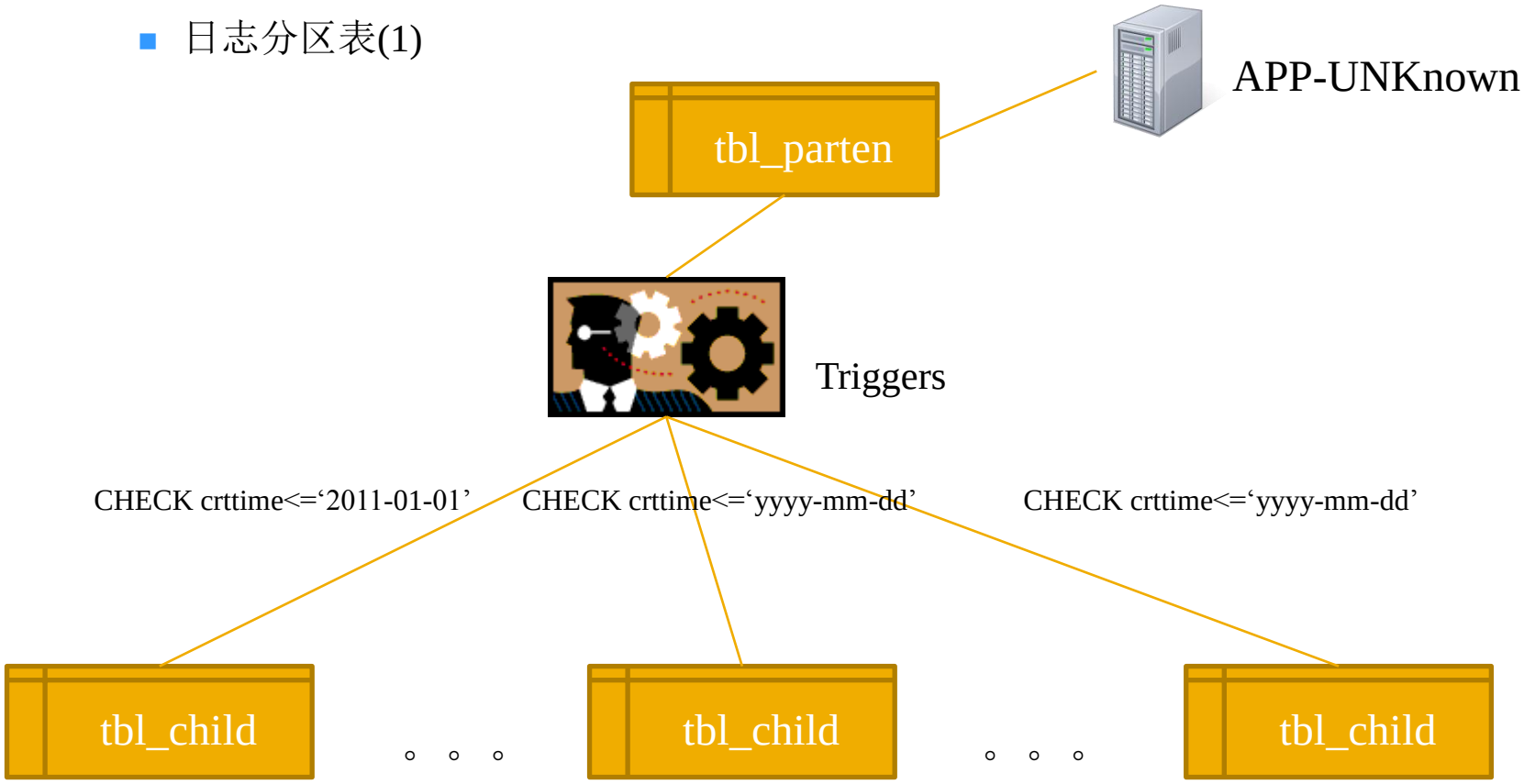
Standby

- hot_standby
- max_standby_archive_delay
- max_standby_streaming_delay
- wal_receiver_status_interval
- hot_standby_feedback

SKY mobi
新手机 新应用 新娱乐

设计阶段

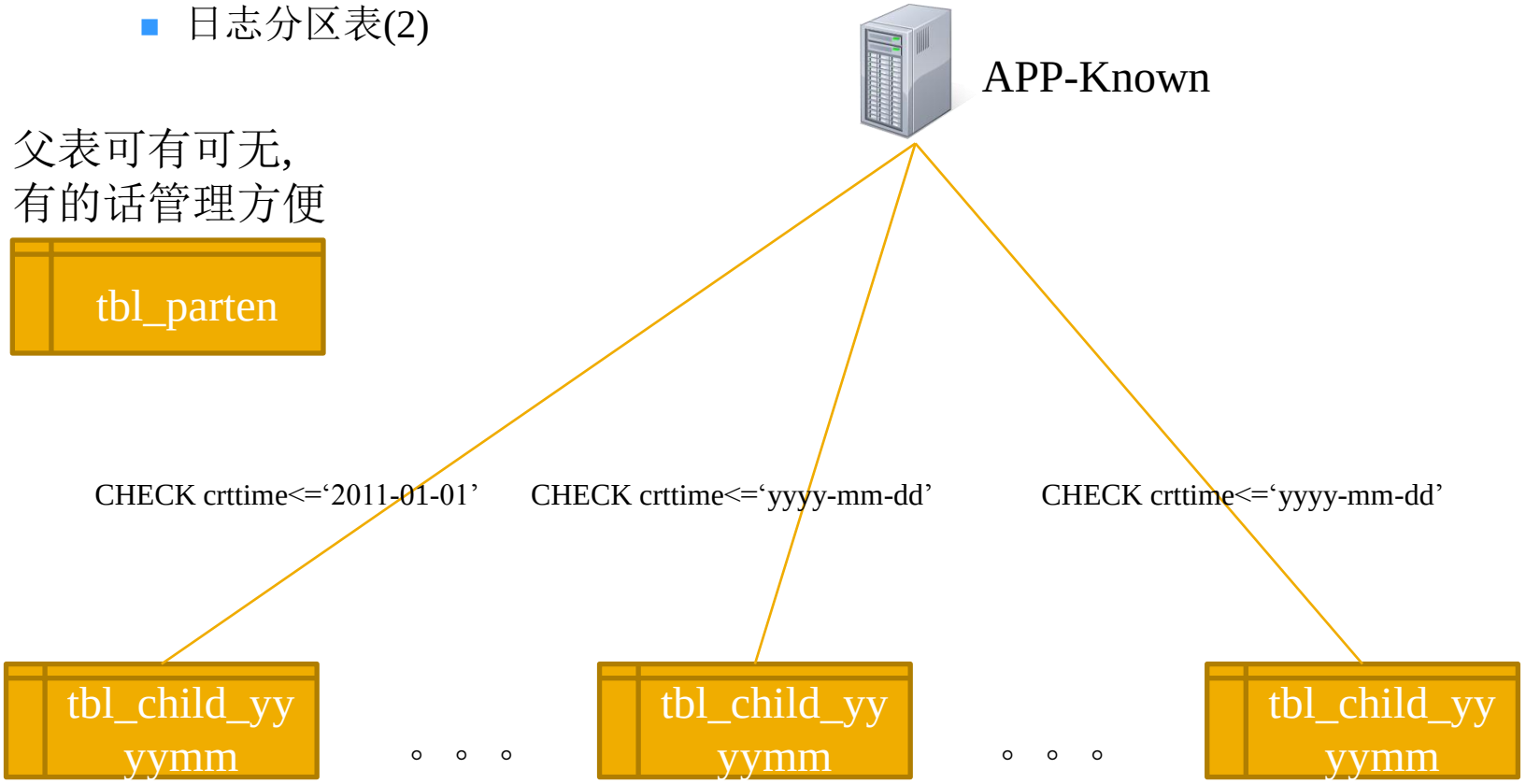
- 系统设计浅析
  - APPStore型(静态)
    - 日志分区表(1)



SKY mobi
新手机 新应用 新娱乐

设计阶段

- 系统设计浅析
  - APPStore型(静态)
    - 日志分区表(2)



SKY mobi
新手机 新应用 新娱乐

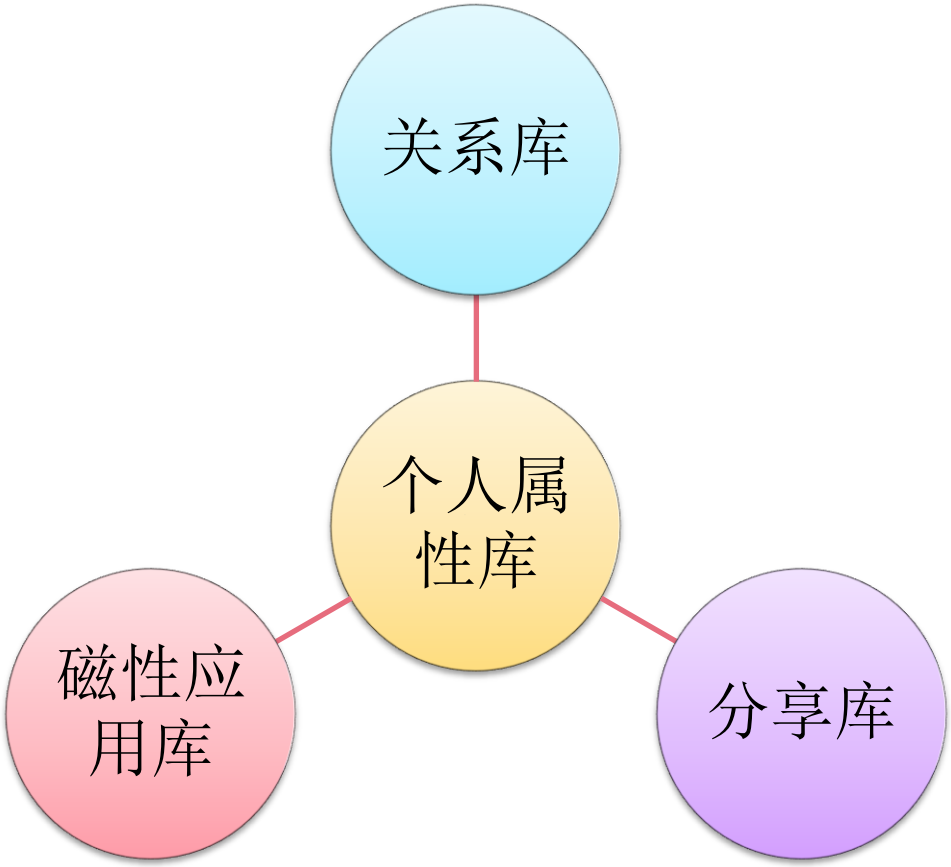
设计阶段

- 数据库挑战浅析
  - SNS型(动态)
    - 写请求(50%)
    - 读请求(50%)
    - 个人属性数据>内存大小
    - 社会关系数据>内存大小
    - 分享数据>内存大小
    - 磁性应用数据>内存大小
    - 活跃数据>内存大小

SKY mobi
新手机 新应用 新娱乐

设计阶段

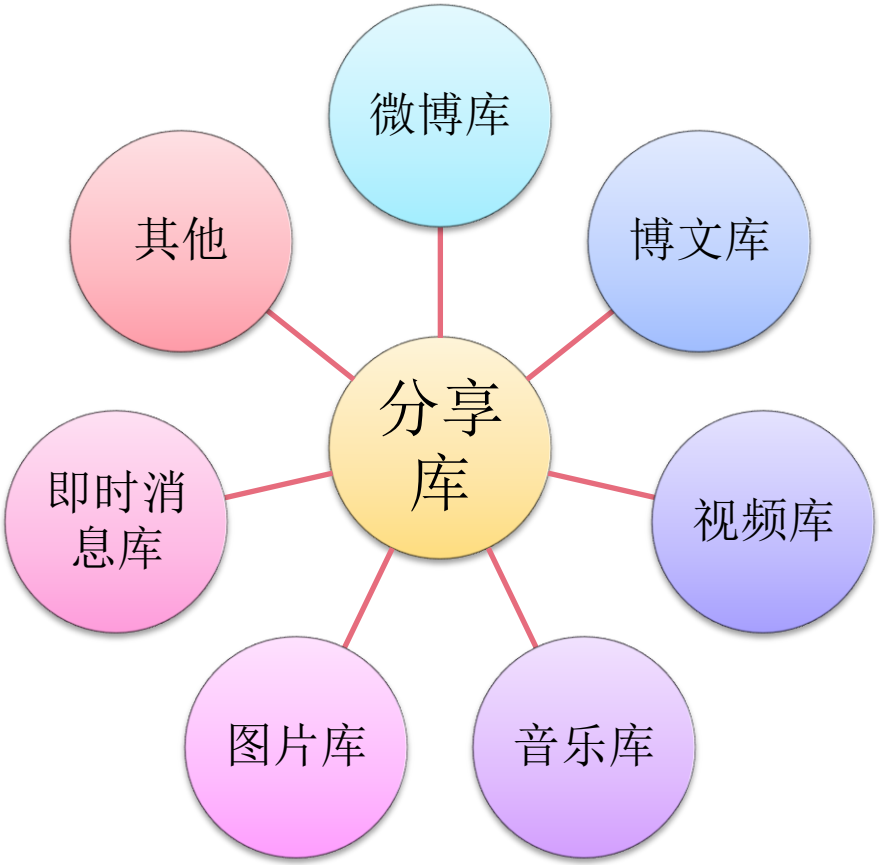
- 系统设计浅析
  - SNS型(动态)
    - 数据库功能性拆分



SKY mobi
新手机 新应用 新娱乐

设计阶段

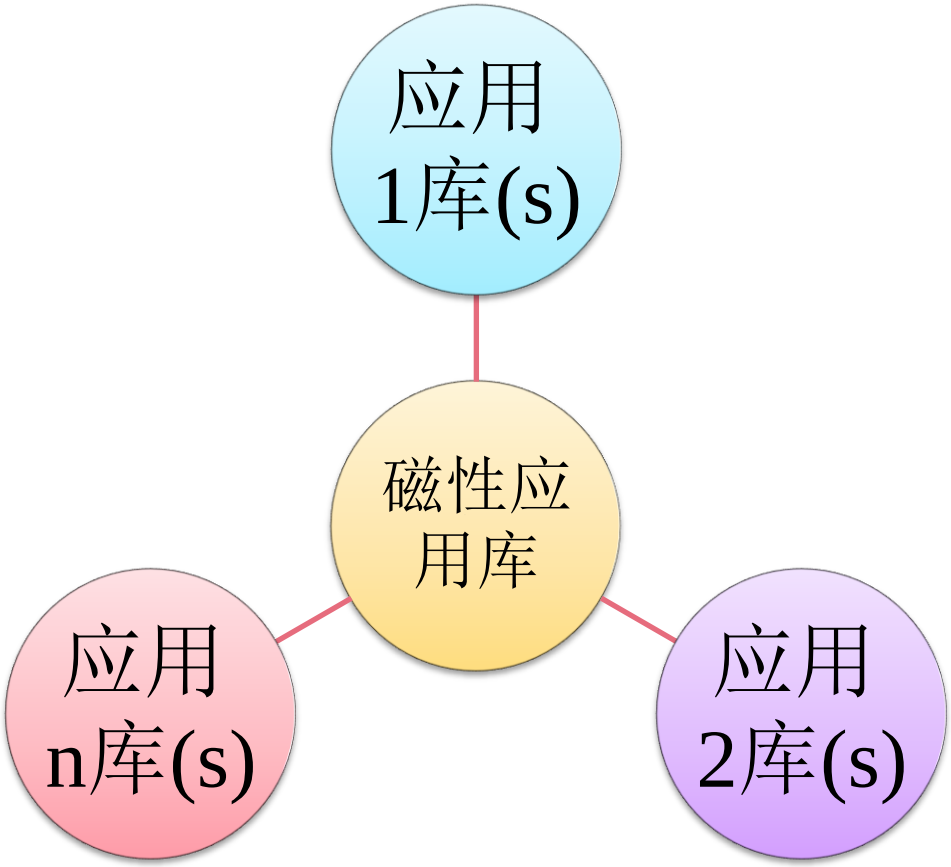
- 系统设计浅析
  - SNS型(动态)
    - 数据库功能性拆分



SKY mobi
新手机 新应用 新娱乐

设计阶段

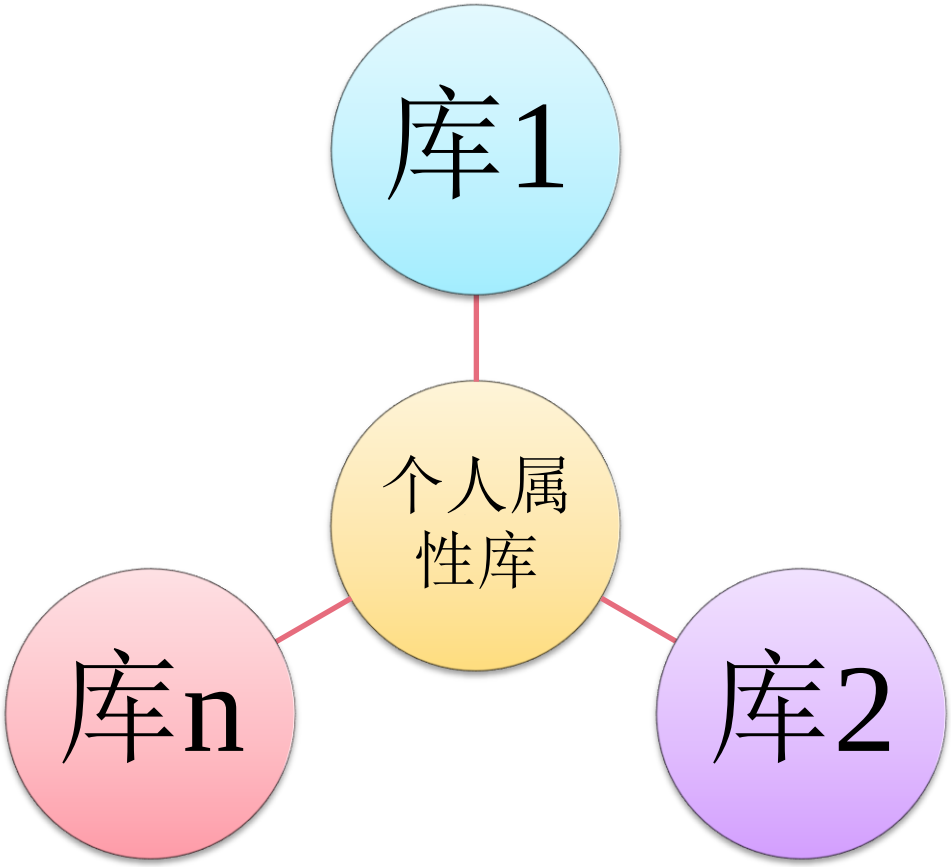
- 系统设计浅析
  - SNS型(动态)
    - 数据库功能性拆分
    - 一般游戏库还可分区域



SKY mobi
新手机 新应用 新娱乐

设计阶段

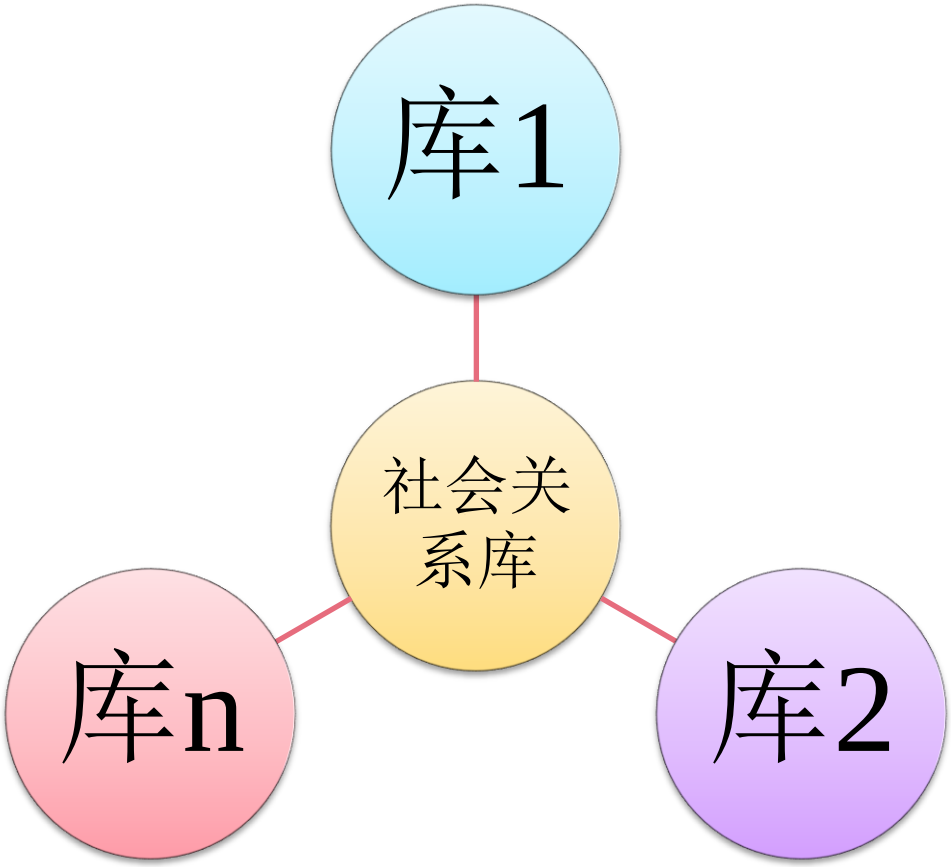
- 系统设计浅析
  - SNS型(动态)
    - 数据库水平拆分



SKY mobi
新手机 新应用 新娱乐

设计阶段

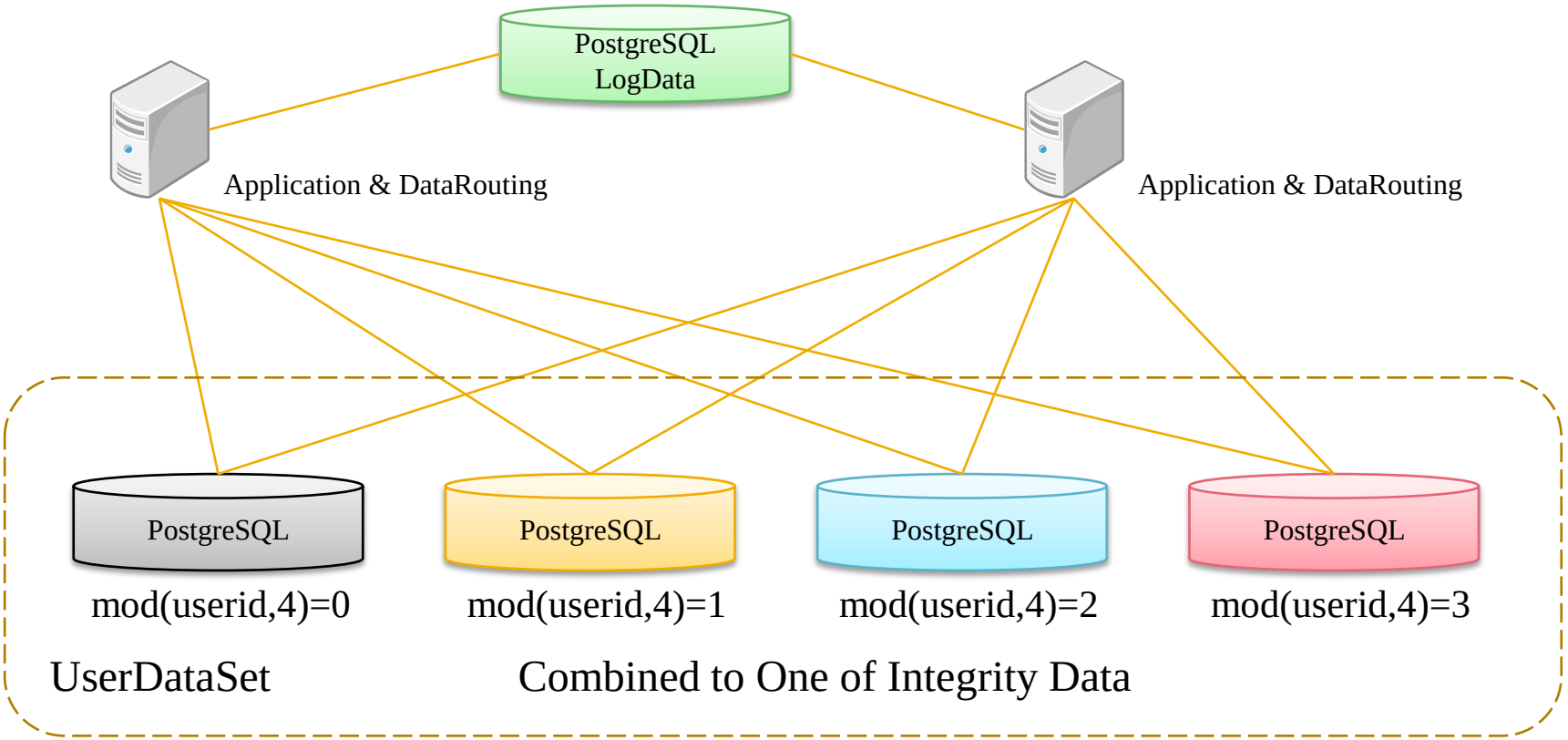
- 系统设计浅析
  - SNS型(动态)
    - 数据库水平拆分



SKY mobi
新手机 新应用 新娱乐

设计阶段

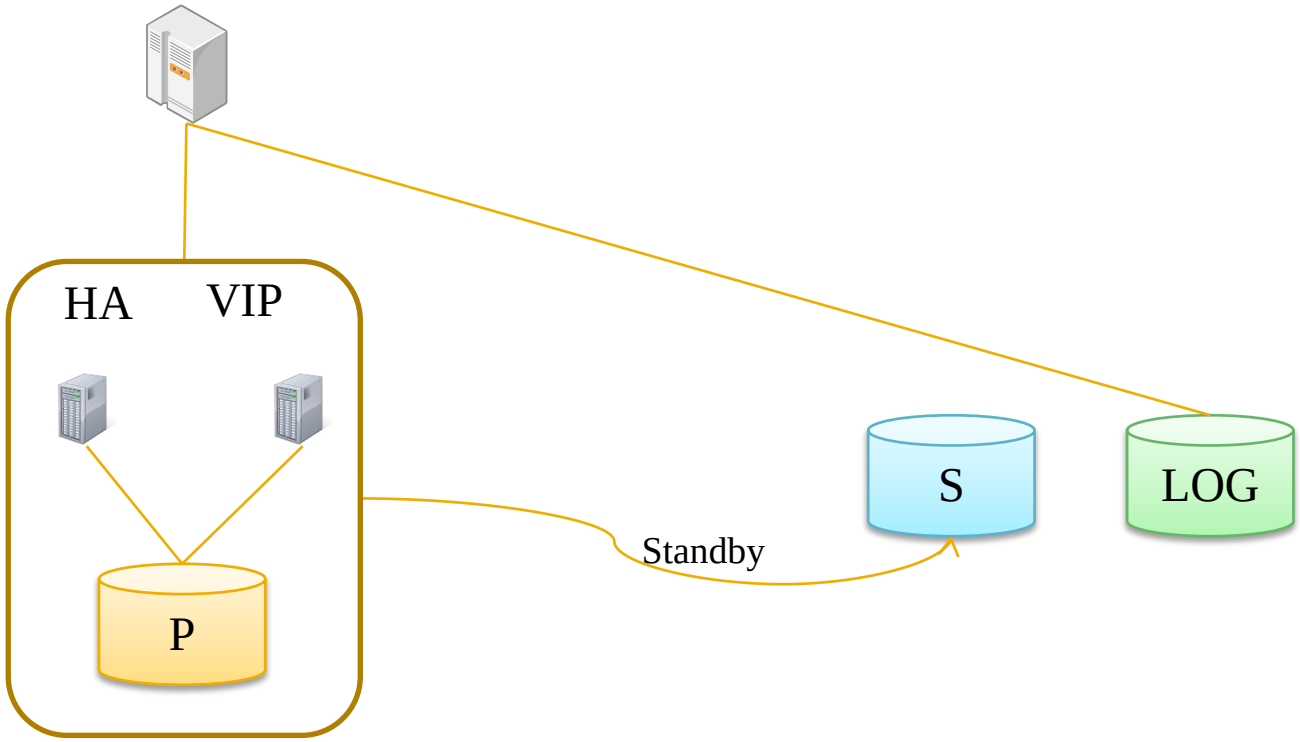
- 系统设计浅析
  - SNS型(动态)
    - 数据库水平拆分



SKY mobi
新手机 新应用 新娱乐

设计阶段

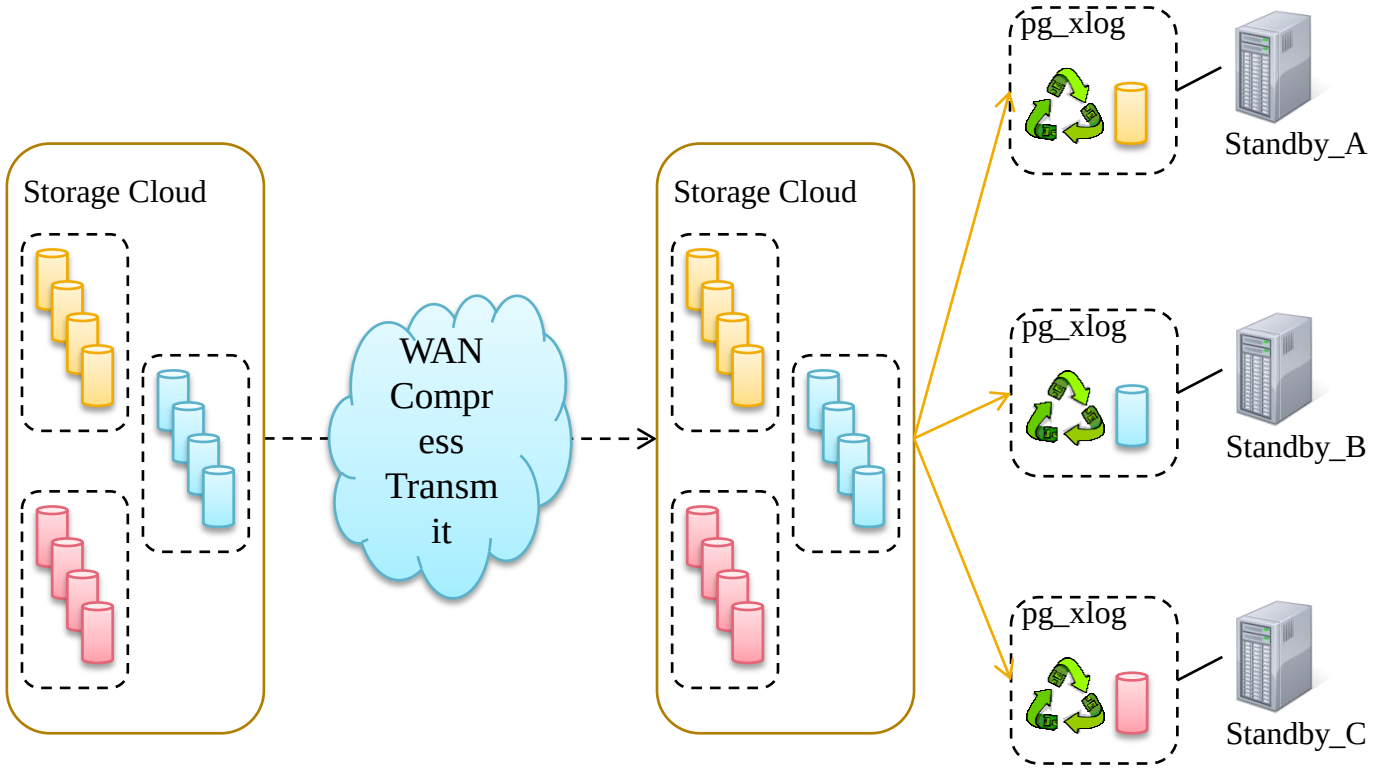
- 系统设计浅析
  - SNS型(动态)
    - IDC内部高可用设计



SKY mobi
新手机 新应用 新娱乐

设计阶段

- 系统设计浅析
  - SNS型(动态)
    - 异地容灾设计

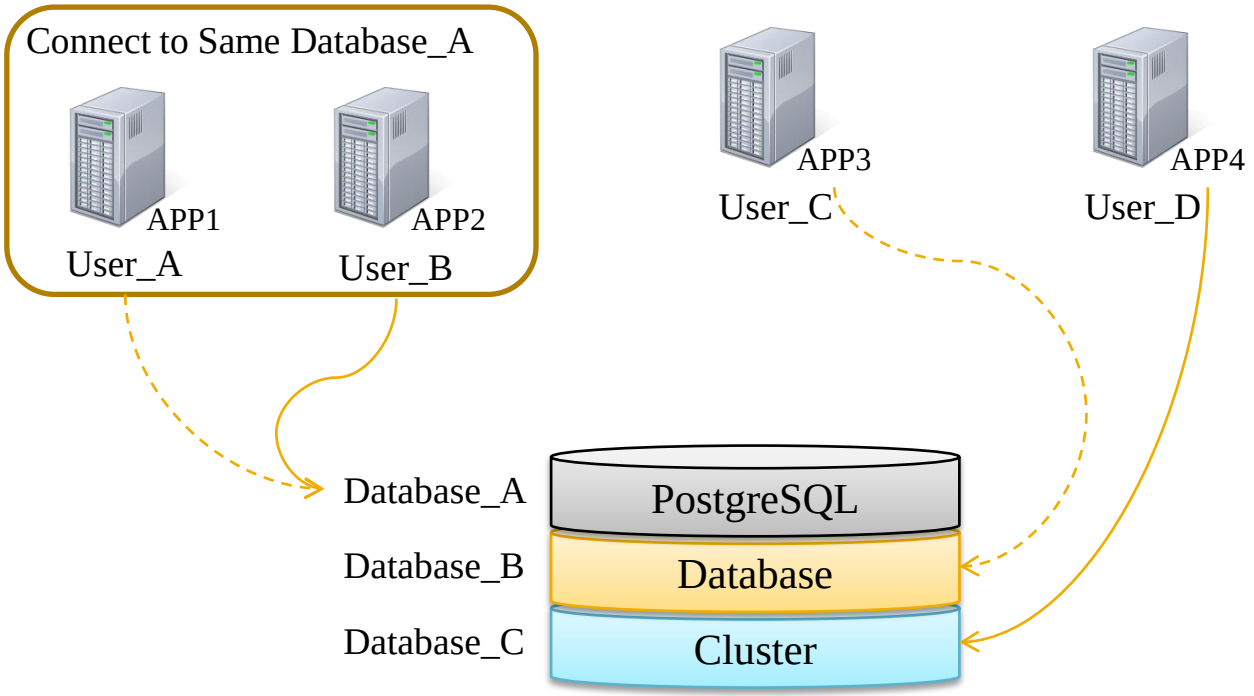


SKY mobi
新手机 新应用 新娱乐

设计阶段

- 系统设计浅析
  - SNS型(动态)
    - 混合场景设计

Tuning settable run-time parameter using ALTER DATABASE or ALTER ROLE.



postgresql.conf :
synchronous_commit = on

User_A :
synchronous_commit = off
User_B :
Inherit postgresql.conf value.

Database_B :
synchronous_commit = off
Database_C:
Inherit postgresql.conf value.

SKY mobi
新手机 新应用 新娱乐

开发阶段

- SCHEMA设计
  - 字段类型,长度,存储格式。
  - fillfactor , autovacuum_enabled , threshold , STATISTICS 相关参数。
  - 索引,约束,PK,UK,FK,CHECK。
- 函数开发
  - Estimate data exists use perform replace select count(*)
  - Eliminate long transaction
  - Eliminate complex operation
- SQL优化
  - Use prepared statement
  - 批量提交
  - Eliminate big resule set return at once结果集指针
  - Eliminate use database operation resource when application can do it
  - Eliminate full table scan, use index when need it

SKY mobi
新手机 新应用 新娱乐

测试阶段

- Loadrunner
- Pgbench
- 平均，最高，最低TPS，QPS，SQL响应时间.

SKY mobi
新手机 新应用 新娱乐

试运阶段

- 调整系统参数
- /etc/sysctl.conf
  - kernel.shmmni
  - kernel.shmall
  - kernel.sem
  - fs.file-max
  - fs.aio-max-nr
  - net.ipv4.ip_local_port_range
  - net.ipv4.tcp_tw_recycle
  - net.ipv4.tcp_max_syn_backlog
  - net.ipv4.ip_conntrack_max
  - net.ipv4.tcp_timestamps
  - net.core.rmem_default
  - net.core.rmem_max
  - net.core.wmem_default
  - net.core.wmem_max
  - vm.overcommit_memory
  - vm.overcommit_ratio
  - vm.lowmem_reserve_ratio
- /etc/security/limits.conf
  - nofile
  - nproc
  - core
  - memlock

SKY mobi
新手机 新应用 新娱乐

试运阶段

- 调整数据库参数
- # **Connection**
- listen_addresses
- max_connections
- superuser_reserved_connections
- unix_socket_directory
- unix_socket_permissions
- tcp_keepalives_idle
- tcp_keepalives_interval
- tcp_keepalives_count
- # **Memory**
- shared_buffers
- maintenance_work_mem
- max_stack_depth (ulimit -s)
- # **Kernel Resource Usage**
- max_files_per_process
- # **Cost-based Vacuum Delay**
- vacuum_cost_delay
- vacuum_cost_limit
- # **Background Writer**
- bgwriter_delay
- bgwriter_lru_maxpages
- bgwriter_lru_multiplier
- # **Asynchronous Behavior**
- effective_io_concurrency
  - (asynchronous I/O requests. Currently, this setting only affects bitmap heap scans)

SKY mobi
新手机 新应用 新娱乐

试运阶段

- # **WAL Settings**
- wal_level
- synchronous_commit
- wal_sync_method
  - (open_datasync , fdatasync , fsync , fsync_writethrough , open_sync)
- wal_buffers
- wal_writer_delay
- commit_delay
- commit_siblings
- # **WAL Checkpoints**
- checkpoint_segments
- checkpoint_timeout
- checkpoint_completion_target
- checkpoint_warning
- # **WAL Archiving**
- archive_mode
- archive_command
- archive_timeout
- # **Streaming Replication**
- # **Synchronous Replication**
- # **Standby Servers**
- # **Planner Method**
- # **Planner Cost Constants**
- random_page_cost
- effective_cache_size
- # **Genetic Query Optimizer**
- # **Other Planner Options**
- default_statistics_target
- constraint_exclusion

SKY mobi
新手机 新应用 新娱乐

试运阶段

- **# Log**
- log_destination
- logging_collector
- log_directory
- log_truncate_on_rotation
- log_rotation_age
- log_rotation_size
- log_min_duration_statement
- log_checkpoints
- log_lock_waits
- log_statement
- **# RUNTIME STATISTICS**
- **# AUTOVACUUM PARAMETERS**
- autovacuum
- log_autovacuum_min_duration
- autovacuum_vacuum_threshold
- autovacuum_analyze_threshold
- autovacuum_vacuum_scale_factor
- autovacuum_analyze_scale_factor
- autovacuum_freeze_max_age
- autovacuum_vacuum_cost_delay
- autovacuum_vacuum_cost_limit
- **# CLIENT CONNECTION DEFAULTS**
- deadlock_timeout
- statement_timeout

SKY mobi
新手机 新应用 新娱乐

运维阶段

- 监控
- 权限管理
- 审计
- 备份，恢复测试
- 性能报告
- 优化，与开发人员一起提高系统性能
- 扩容
- 资源调配(如业务量下滑，对应的数据库不再需要高端的硬件设备，涉及数据迁移，数据库合并等)
- 日志分析
- 空间回收(如删除归档,备份并删除不再需要的数据)

SKY mobi
新手机 新应用 新娱乐

运维阶段

- 监控
  - 监控要素
    - 系统：CPU，内存，交换分区，网络，IOPS，磁盘，Kernel报错，硬件故障等。
    - 数据库：慢查询，DEAD TUPLE比例，锁等待事件，buffer使用情况，严重异常日志等。
  - 常用软件
    - 实时告警 Nagios®
    - 趋势图 cacti
    - sendmail

SKY mobi
新手机 新应用 新娱乐

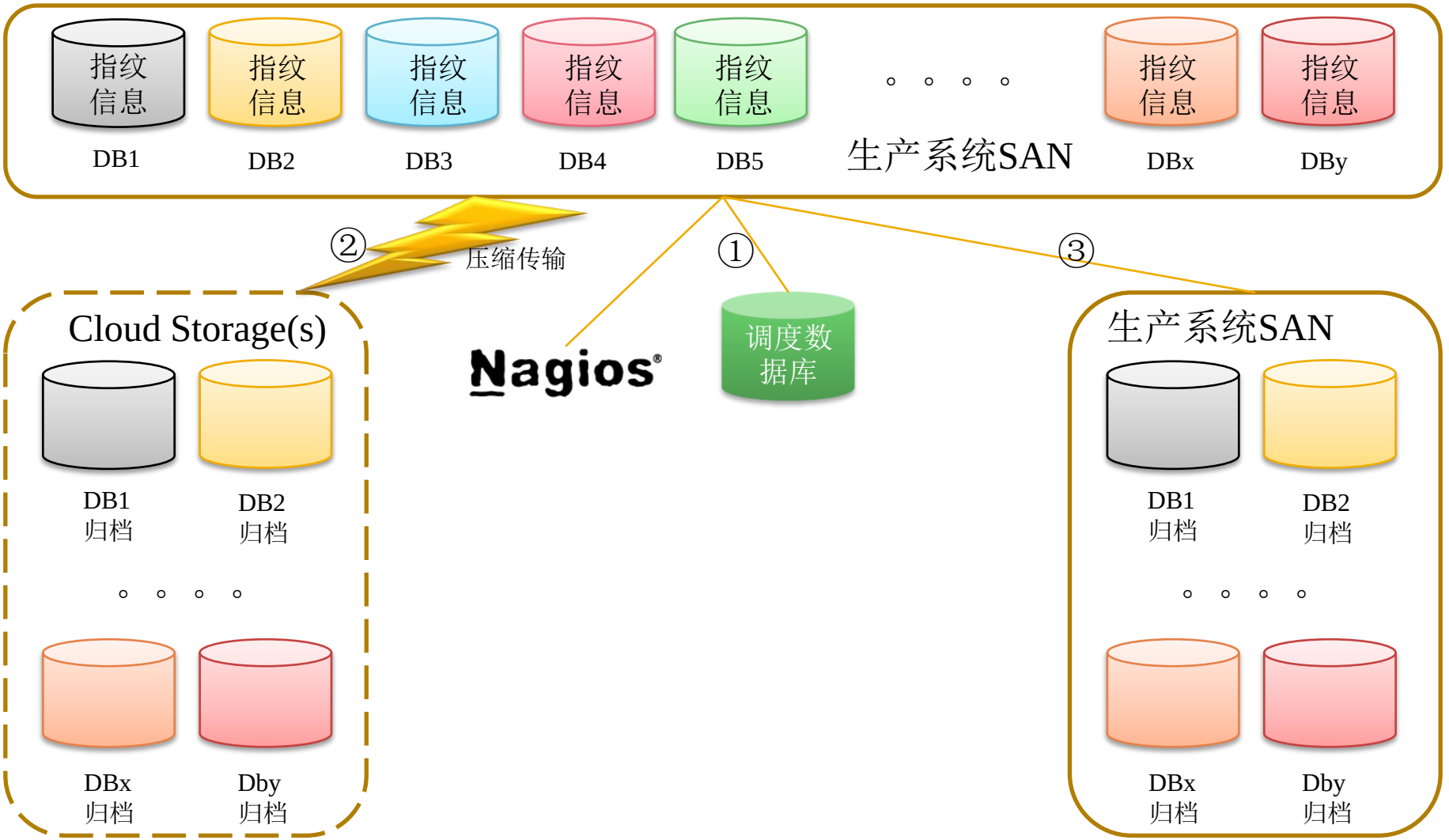
运维阶段

- 权限管理
  - 最小权限连接：pg_hba.conf
  - 最小权限对象：grant / revoke
- 审计
  - 登录审计
  - DDL操作审计
  - 按用户审计(目前PostgreSQL还不支持)

SKY mobi
新手机 新应用 新娱乐

运维阶段

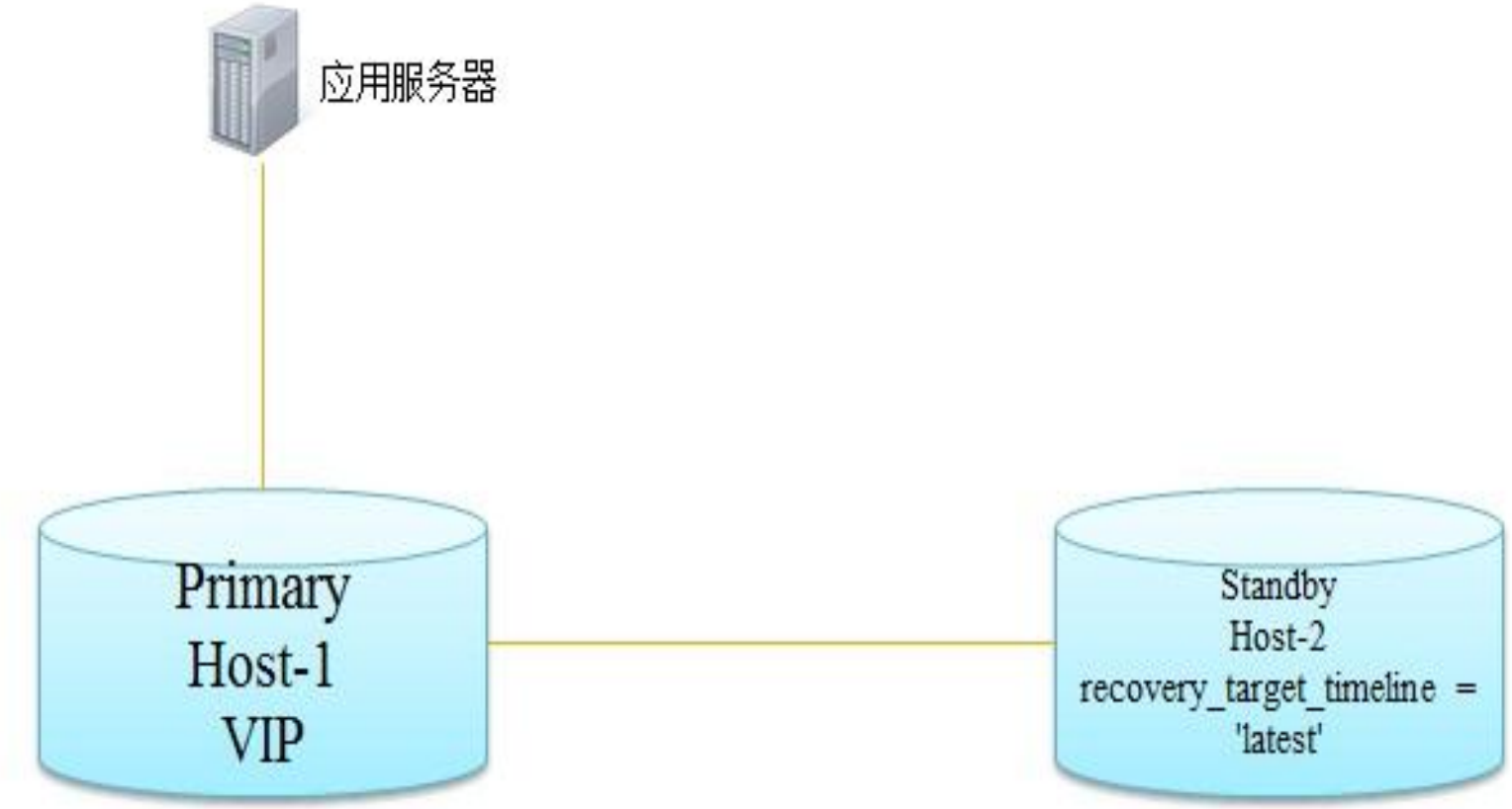
- 备份



SKY mobi
新手机 新应用 新娱乐

运维阶段

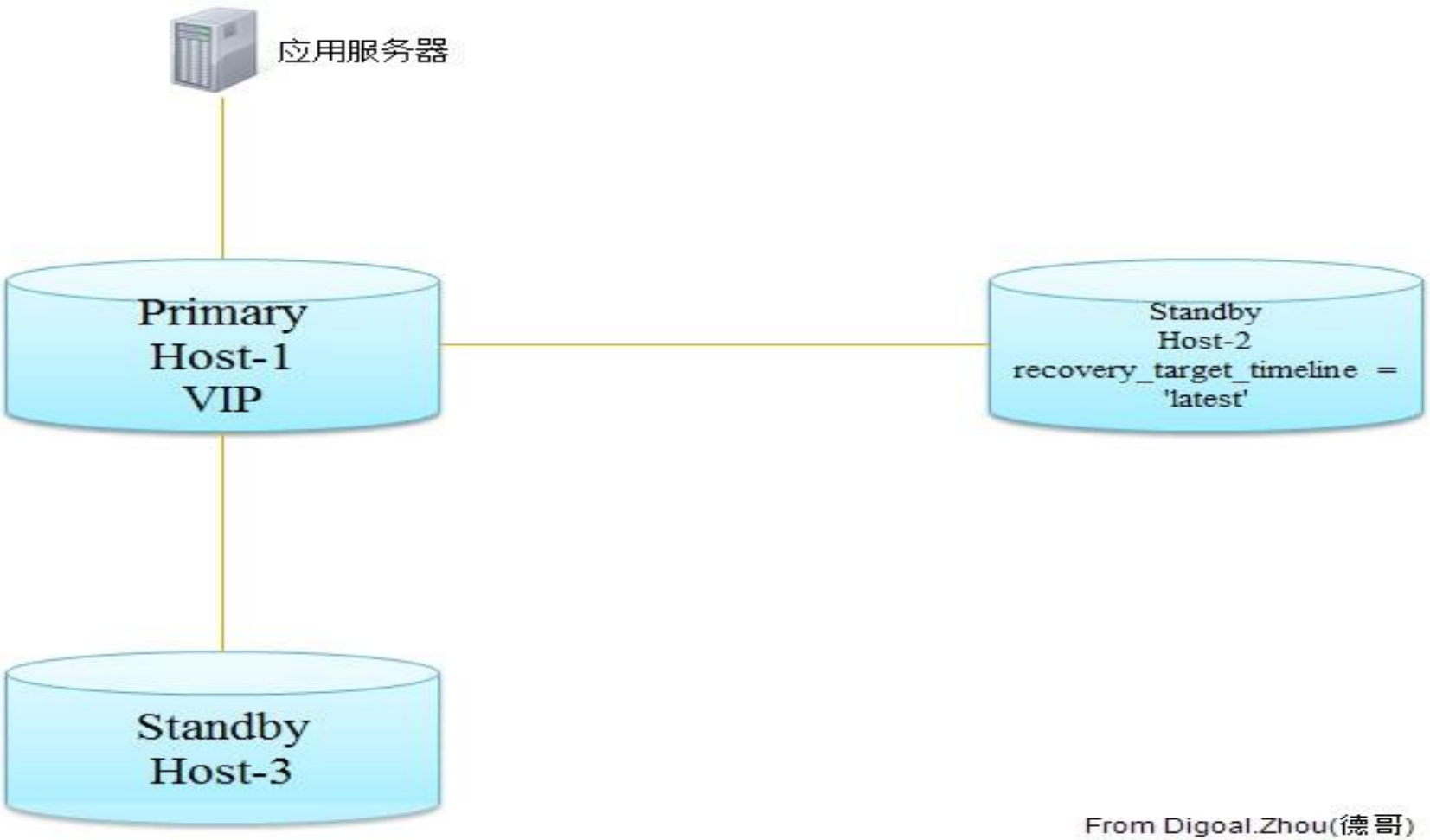
- 扩容(接近平滑)



From Digoal.Zhou(德哥)

SKY mobi
新手机 新应用 新娱乐

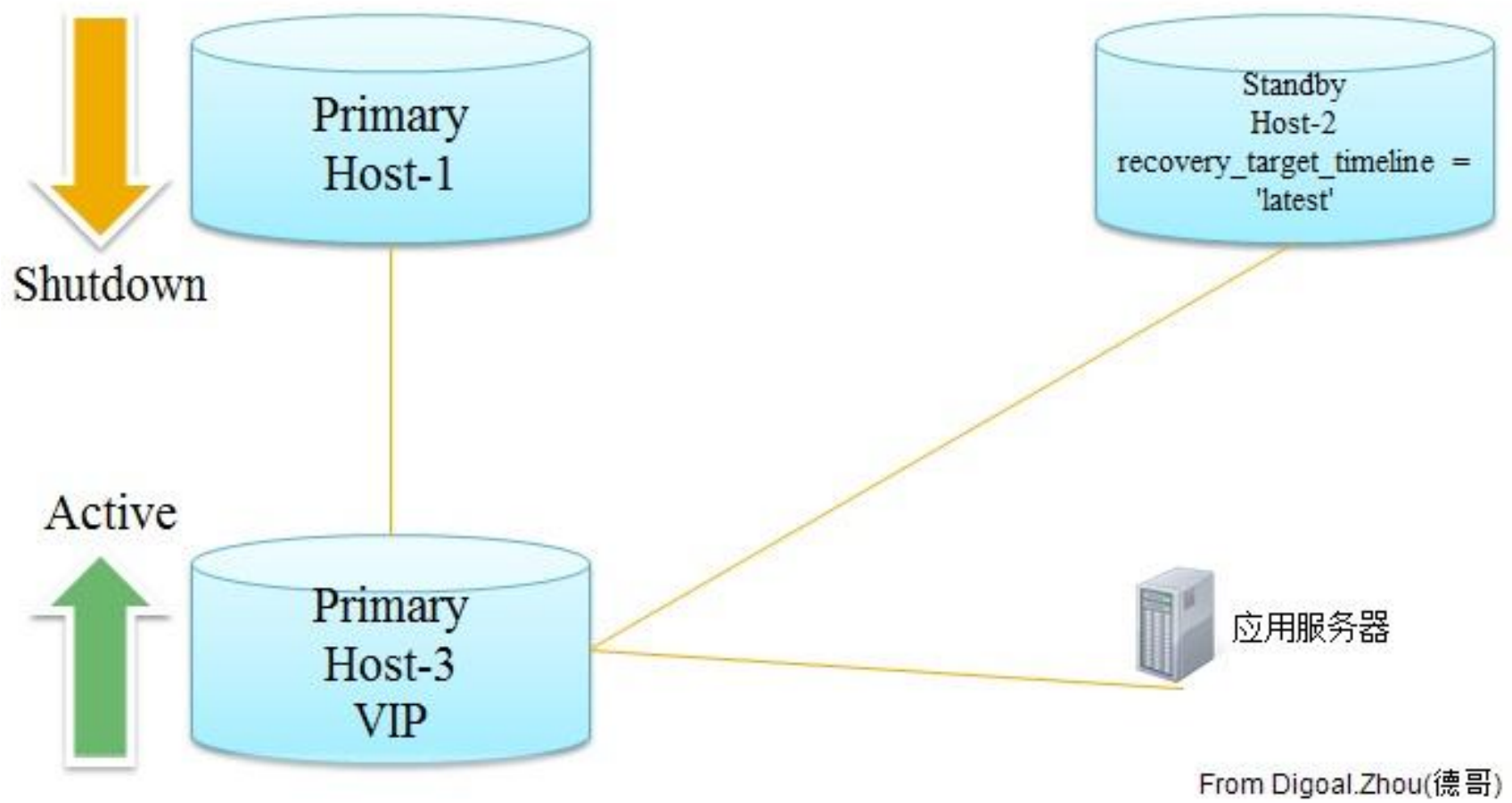
运维阶段



From Digoal.Zhou(德哥)

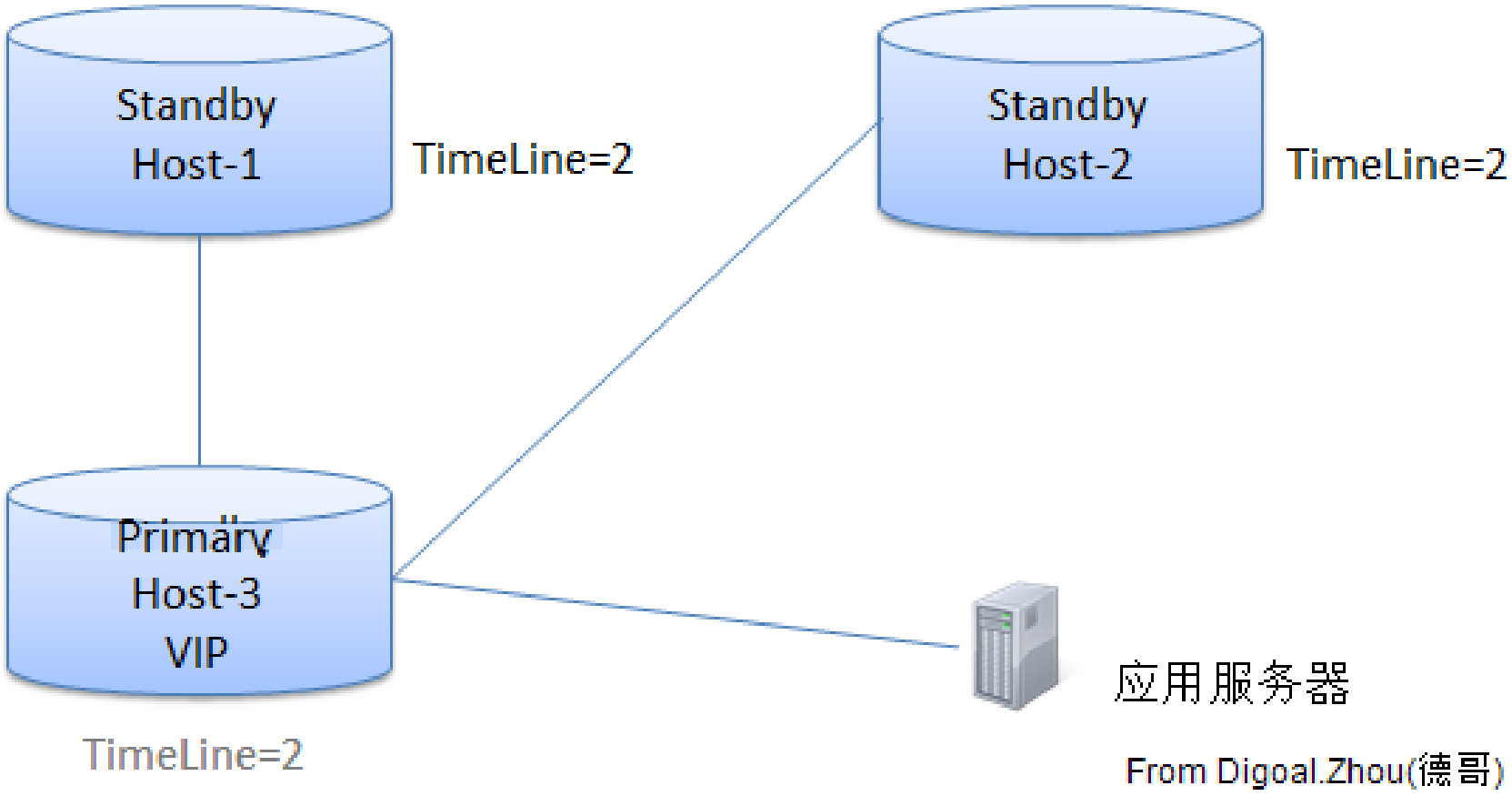
SKY mobi
新手机 新应用 新娱乐

运维阶段



SKY mobi
新手机 新应用 新娱乐

运维阶段



From Digoal.Zhou(德哥)

SKY mobi
新手机 新应用 新娱乐

Thanks

- Thanks all people contribute to PostgreSQL.

- Digoal.Zhou
- Blog
- http://blog.163.com/digoal@126