新手机 新应用 新娱乐

PostgreSQL 经验谈

Digoal.zhou

6/16/2012

SKY mobi
新手机 新应用 新娱乐

温馨提醒 - Happy Father's Day

SKY mobi
新手机 新应用 新娱乐

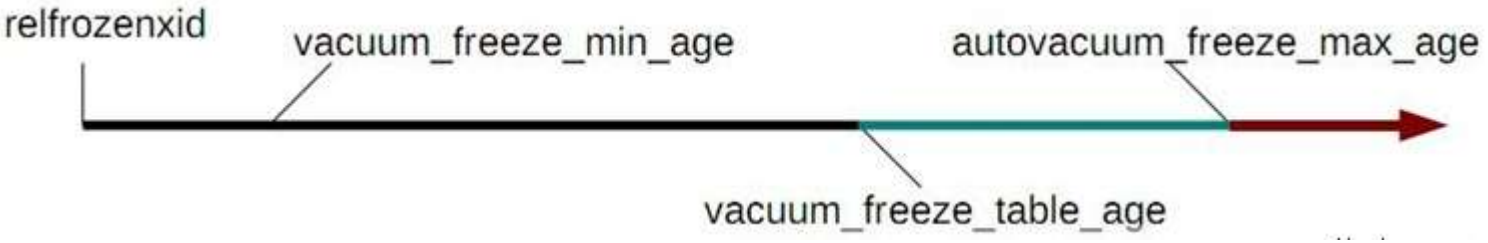
主要内容

- 浅析PostgreSQL数据库特点
- 构建高效,稳定,安全,易扩展的PostgreSQL数据库系统
- PostgreSQL容灾
- PostgreSQL备份, 还原.
- PostgreSQL优化
- PostgreSQL迁移
- PostgreSQL压力测试
- PostgreSQL监控
- PostgreSQL Caveat
- PostgreSQL展望

SKY mobi
新手机 新应用 新娱乐

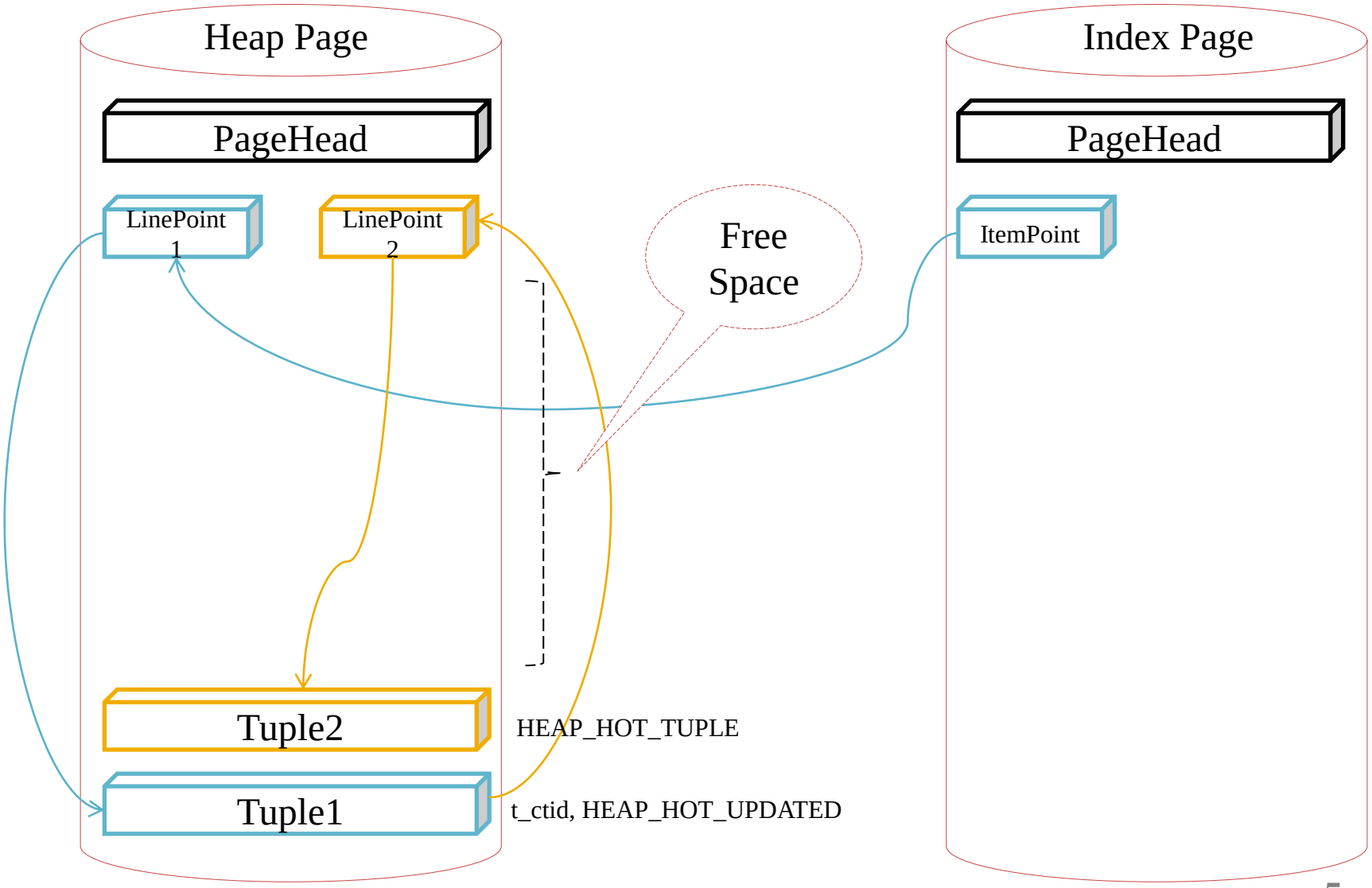
浅析 PostgreSQL 特点

- 功能
  - 严格遵循ACID(事务操作原子性(支持savepoint),状态一致性,隔离(read committed, repeatable read, serializable),持久化(fsync xlog))
    - 在保证ACI的前提下, 降低持久化标准可以提升百倍性能. (异步提交, 异常DOWN库或DOWN机可能导致wal_buffer里面未flush到xlogfile的信息丢失(最多wal_writer_delay*3时间的wal信息),不会导致数据库不一致或不可恢复.)
  - two phase commit(例如利用dblink做跨库事务)
  - MVCC, 8个锁级别(DML读写不冲突)
  - 更新优化机制(HOT, 降低索引需要更新的概率, 索引字段没有被更新且被更新的记录所在的BLOCK有足够的空间存下更新后的记录时符合HOT)
  - 垃圾数据通过VACUUM机制回收(Object's VM file中的块不需要扫描, 触发full scan除外.vacuum 与DML不冲突)



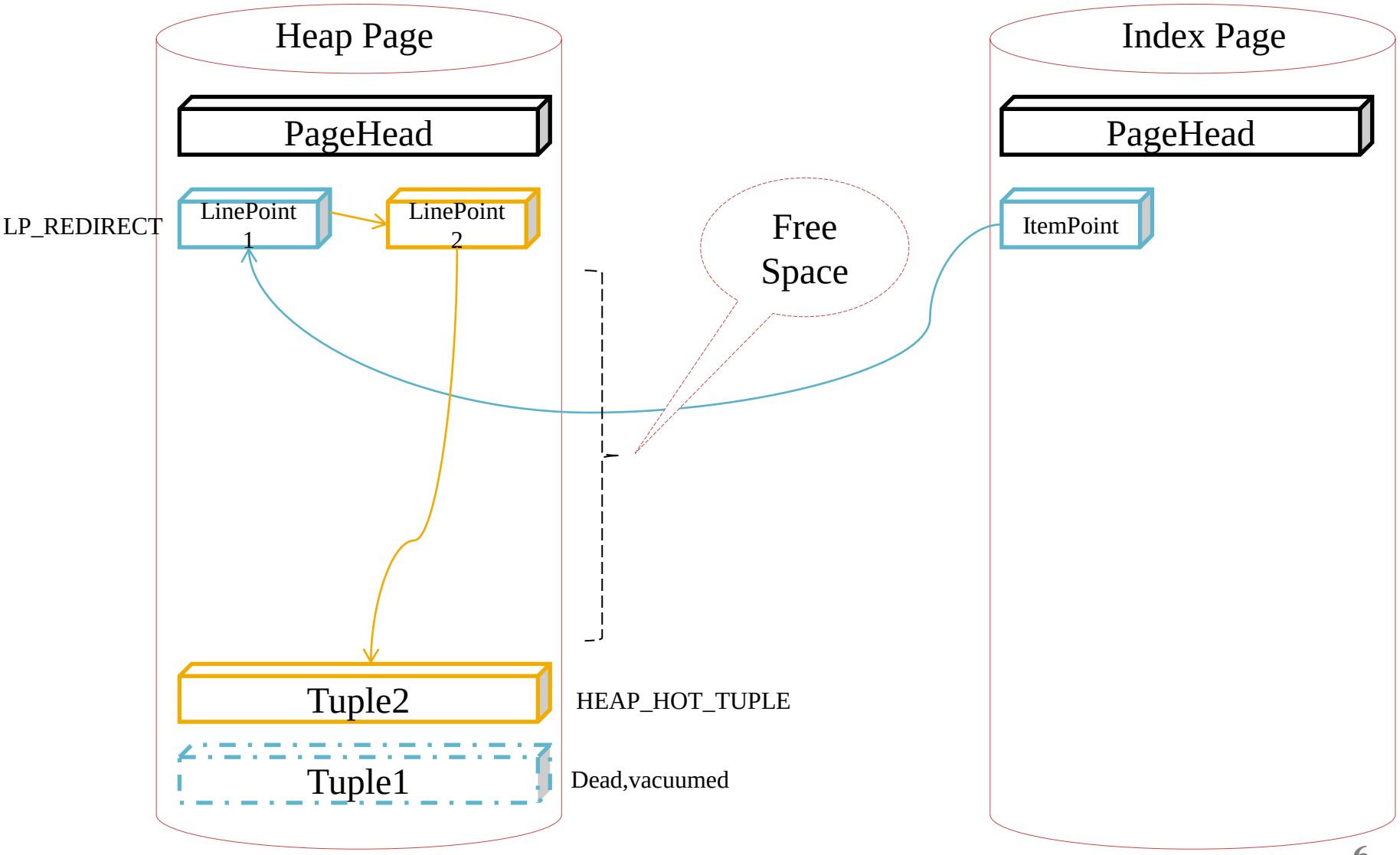
SKY mobi
新手机 新应用 新娱乐

HOT Update



SKY mobi
新手机 新应用 新娱乐

HOT Update



SKY mobi
新手机 新应用 新娱乐

浅析 PostgreSQL 特点

- 功能
  - 存储(表空间,临时表空间)
  - 字段存储选项(plain, main, external, extended)
  - 支持在线创建,删除索引(不和select,update,delete,insert冲突)
  - 索引维护(支持多个完全一致的索引,使得维护索引更加自由)
  - 索引类型(btree,hash,gin,gist,sp-gist,btree-gist,btree-gin)
  - 函数语言(C, SQL, plpgsql, plperl, pltcl, plpython, ...)
  - 支持plpgsql 函数debugger
  - 游标(支持[no]scroll, with[out] hold, 支持游标删除,更新(where current of))
  - 触发器([per row|per statement] INSERT, DELETE, UPDATE[of column,], TRUNCATE)
  - 规则(per statement, do also|instead|nothing)
  - 支持分区表(支持基于触发器,规则的分区表,多层分区支持,支持按照分区约束优化查询扫描)

SKY mobi
新手机 新应用 新娱乐

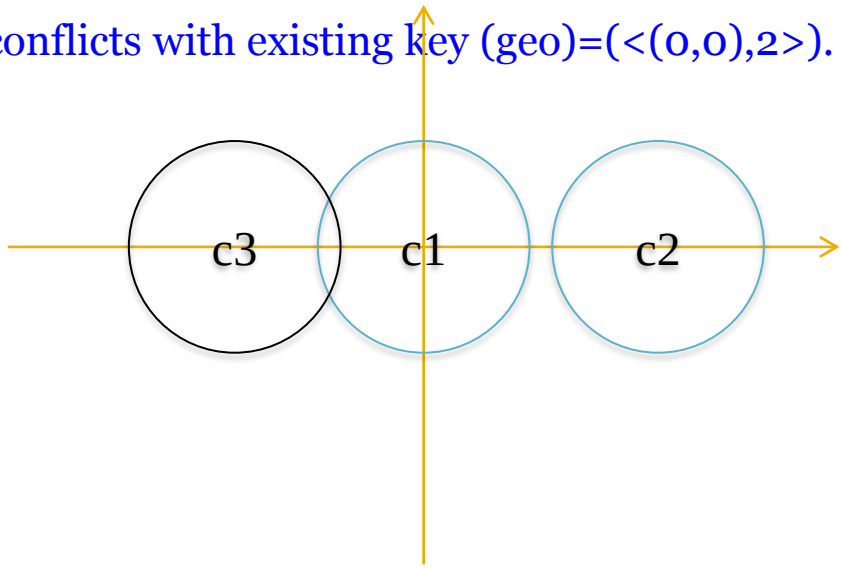
浅析 PostgreSQL 特点

- 功能
  - 继承(一个表可以继承多个表,也可以被多个表继承,支持多级继承,不允许环路)
  - 约束(not null, check, unique, primary key, exclude 约束)
  - 数据类型(数字, 货币, 字符, 比特流, 字节流, 时间, 布尔, 枚举, 几何, 网络, 全文检索(中文分词nlpbamboo), UUID, XML, JSON, ARRAY, 自定义复合, range, oid)
  - 内建函数(字符,时间,数字,聚集,数学,比特,字节流,规则表达式,枚举,几何,网络,全文检索,序列,JSON,XML,数组,范围,窗口,集合,触发器,系统管理等类型的函数)
  - 外部数据源表(PostgreSQL,文件,MySQL,Oracle,Sybase,ODBC,redis,couchDB,...), 支持外部表的JOIN, 支持外部表的统计信息收集, 目前仅支持外部表只读.(外部表在异构平台数据迁移中非常便利)
  - 支持的几种join方法(hashjoin, nestloop, merge)

SKY mobi
新手机 新应用 新娱乐

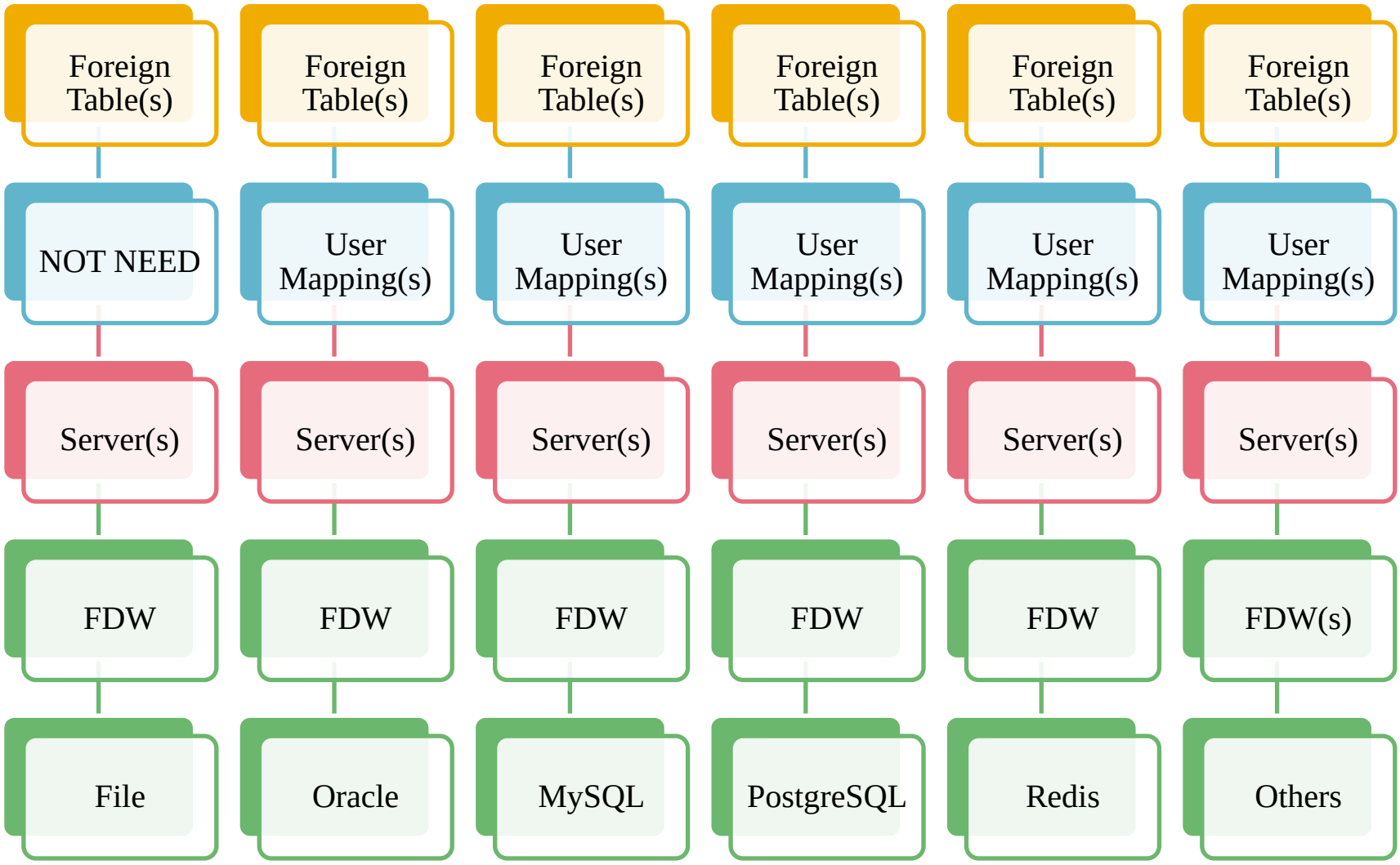
浅析 PostgreSQL 特点

- 举例
- exclusion约束
  - CREATE TABLE test(id int,geo circle,EXCLUDE USING GIST (geo WITH pg_catalog.&&));
  - INSERT INTO test values(1,'<(0,0),2>'::circle);
  - INSERT INTO test values(1,'<(4.1,0),2>'::circle);
  - INSERT INTO test values(1,'<(-1.9,0),2>'::circle);
  - ERROR: conflicting key value violates exclusion constraint "test_geo_excl"
  - DETAIL: Key (geo)=(<(-1.9,0),2>) conflicts with existing key (geo)=(<(0,0),2>).



SKY mobi
新手机 新应用 新娱乐

Foreign Data



SKY mobi
新手机 新应用 新娱乐

浅析 PostgreSQL 特点

- 功能
  - 权限管理(表空间,数据库,schema,表,视图,序列,fdw,域,函数,大对象,复合类型, 其中表支持列级权限)
  - 权限分配(系统(全局)配置, 数据库级配置, 用户级配置, 会话级, 事务级配置)
  - 在线备份, 还原(基于时间点/事务号/自定义恢复点的恢复)
  - standby(hot_standby(支持standby在恢复的同时提供查询), warm_standby(恢复时不提供查询), 同步/异步standby(支持事务粒度的), 支持级联的standby,支持暂停和继续恢复,)
  - 数据块复制(流复制, log shipping)
  - SQL复制(触发器(londiste3, slony-I), SQL分发(pgpool-II))
    - 基于SQL分发的复制(架设于客户端和数据库之间)需要注意, 必须所有节点都执行成功事务才能算成功. 否则数据就不一致了.
    - 基于触发器的复制(架设于主库和slave库之间)则不需要考虑跨库事务的问题.

SKY mobi
新手机 新应用 新娱乐

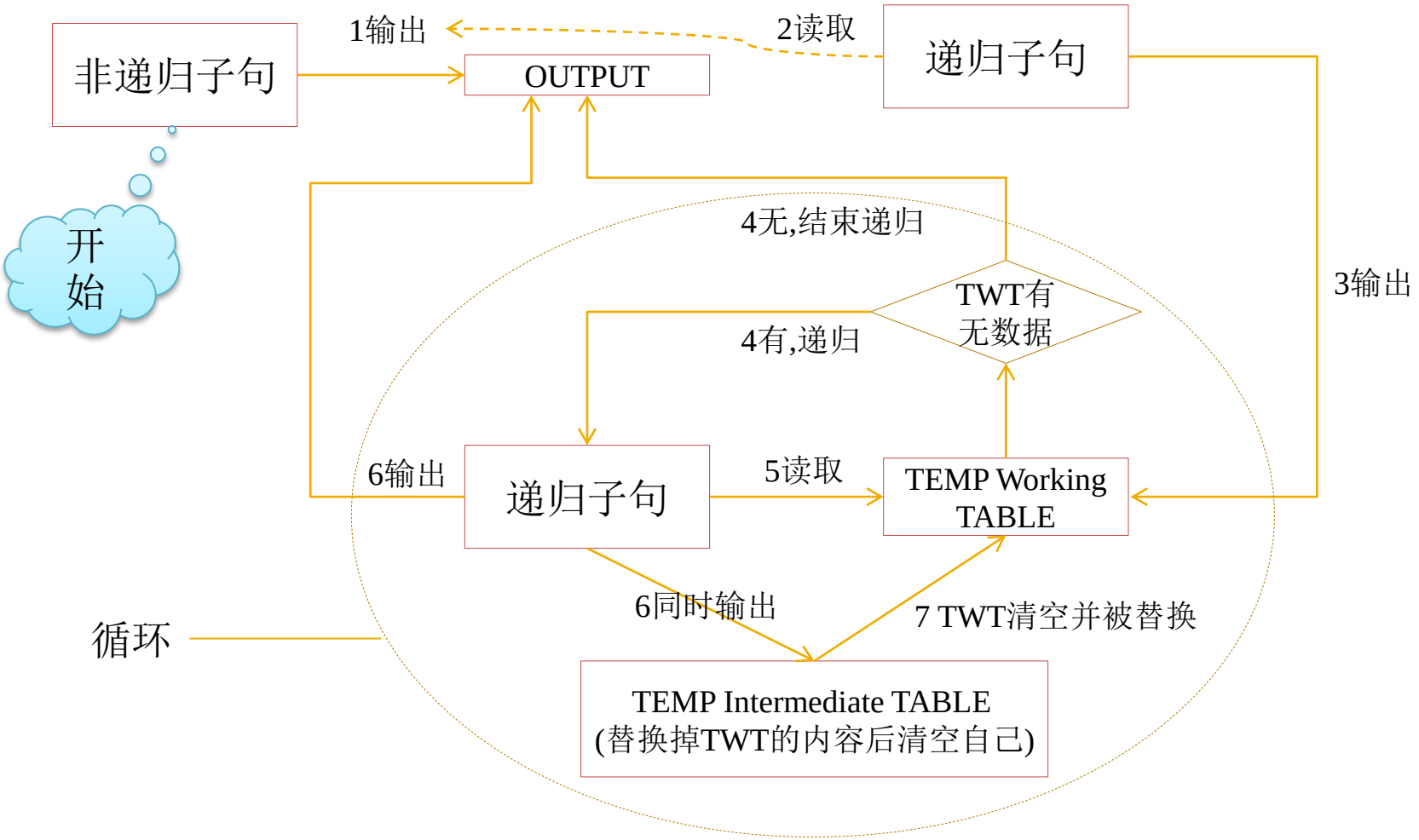
浅析 PostgreSQL 特点

- 功能
  - Common Table Expression(WITH语法,可以构造逻辑复杂的查询,如递归,树形查询,支持WITH中使用DELETE,update,select等.)
  - SQL支持insert,update,delete returning语法, 适用后续需要对先前记录进行更改的操作, 如二次确认.
  - 计划器(explain analyze verbose costs buffers timing format [text xml json yaml])

SKY mobi
新手机 新应用 新娱乐

WITH(Common Table Expressions)

- UNION ALL 去重复(去重复时NULL 视为等同)
- 图中所有输出都涉及UNION [ALL]的操作, 包含以往返回的记录和当前返回的记录



SKY mobi
新手机 新应用 新娱乐

浅析 PostgreSQL 特点

- 性能
  - 支持自动平滑调整的checkpoint
  - 全文检索类型以及索引(例如中文分词nlpbamboo)
  - gin,gist索引(有效提高范围查询,几何类型,多值类型如array等的查询性能,支持exclude 约束, 临近值查询加速, 排序等.)
  - INDEX ONLY SCAN (需要用到 visibility map, 并不是所有tuple都只需要扫描Index)
  - 支持表空间级COST值(支持不同的表空间配置不同的IO成本)
  - 支持多个临时表空间
  - 部分索引(create index on t(column) where ?)
  - 表JOIN的顺序和算法优化(可强制按SQL书写的顺序进行JOIN(join_collapse_limit=1), 也可以让数据库穷举选出最优JOIN顺序)
  - SQL子查询的算法优化(可强制按照子查询结构计算执行计划(from_collapse_limit=1), 或让数据库把子查询中的表提升到上一级的表中进行关联, 从而选择最优的关联顺序.)

SKY mobi
新手机 新应用 新娱乐

浅析 PostgreSQL 特点

- 性能
  - unlogged table
  - 支持并行还原(pg_restore -j)
  - (Cache预加载), 通过pgfincore加载数据至OS Cache, 同时支持查看数据表, 索引,toast表等当前的哪些数据块在OS Cache里面, 可以把这些数据记录到数据库中, 重启后按照这个记录加载这些数据块到OS Cache里面.
  - 动态性能视图(pg_stat_statements, pg_stats, pg_statio_xxx_tables|sequences|indexes, pg_stat_activity, pg_stat_xxx_tables|indexes|bgwriter|database|database_conflicts|replication|functions)
  - hint(enable_bitmapscan,enable_hashagg,enable_hashjoin,enable_indexscan,enable_indexonlyscan,enable_material,enable_mergejoin,enable_nestloop,enable_seqscan,enable_sort,enable_tidscan)
  - 优化器可用的成本维度(seq_page_cost, random_page_cost, cpu_tuple_cost, cpu_index_tuple_cost, cpu_operator_cost, effective_cache_size, 表空间级cost)

SKY mobi
新手机 新应用 新娱乐

浅析 PostgreSQL 特点

- 其他
  - 国际化(列级别的collation, 库级collation, 多国语言, 多字节编码)
  - 平台支持(Windows, unix, linux, bsd, solaris, HPUX, AIX, ...)
  - 函数(支持100个参数)
  - 模块(btree-gin, btree-gist, citext, dblink, file_fdw, hstore, isn, ltree, pageinspect, passwordcheck, pgbench, pg_buffercache, pg_freespacemap, pg_standby, pg_stat_statements, pgstattuple, pg_trgm, pg_upgrade, sepgsql, tablefunc, uuid-osp, ...)
  - more ([github, pgfoundry, pgxn], other [postgis] ...)
  - 审计([断开]连接,按用户或数据库维度的审计,客户端IP,端口,用户,数据库,SQL,执行时间...)
  - 连接鉴权(client_ip, user, dbname, 认证方法(reject, md5, trust, ldap, ...))

SKY mobi
新手机 新应用 新娱乐

浅析 PostgreSQL 特点

- 限制
  - Limit — Value
  - Maximum Database Size — Unlimited
  - Maximum Table Size — 32 TB
  - Maximum Row Size — 1.6 TB
  - Maximum Field Size — 1 GB
  - Maximum Rows per Table — Unlimited
  - Maximum Columns per Table — 250 - 1600 depending on column types
  - Maximum Indexes per Table — Unlimited

SKY mobi
新手机 新应用 新娱乐

浅析 PostgreSQL 特点

- 许可
  - 开源,BSD
- 可定制化
  - 例如,
  - 自建复合类型, 枚举类型
  - 自定义access method
  - 表名最大长度限制的修改,
    - src/include/pg_config_manual.h
    - #define NAMEDATALEN 64
    - re initdb
  - boolean值默认输出格式的修改,
    - src/backend/utils/adt/bool.c
    - change : result[0] = (b) ? 't' : 'f';
    - remake software

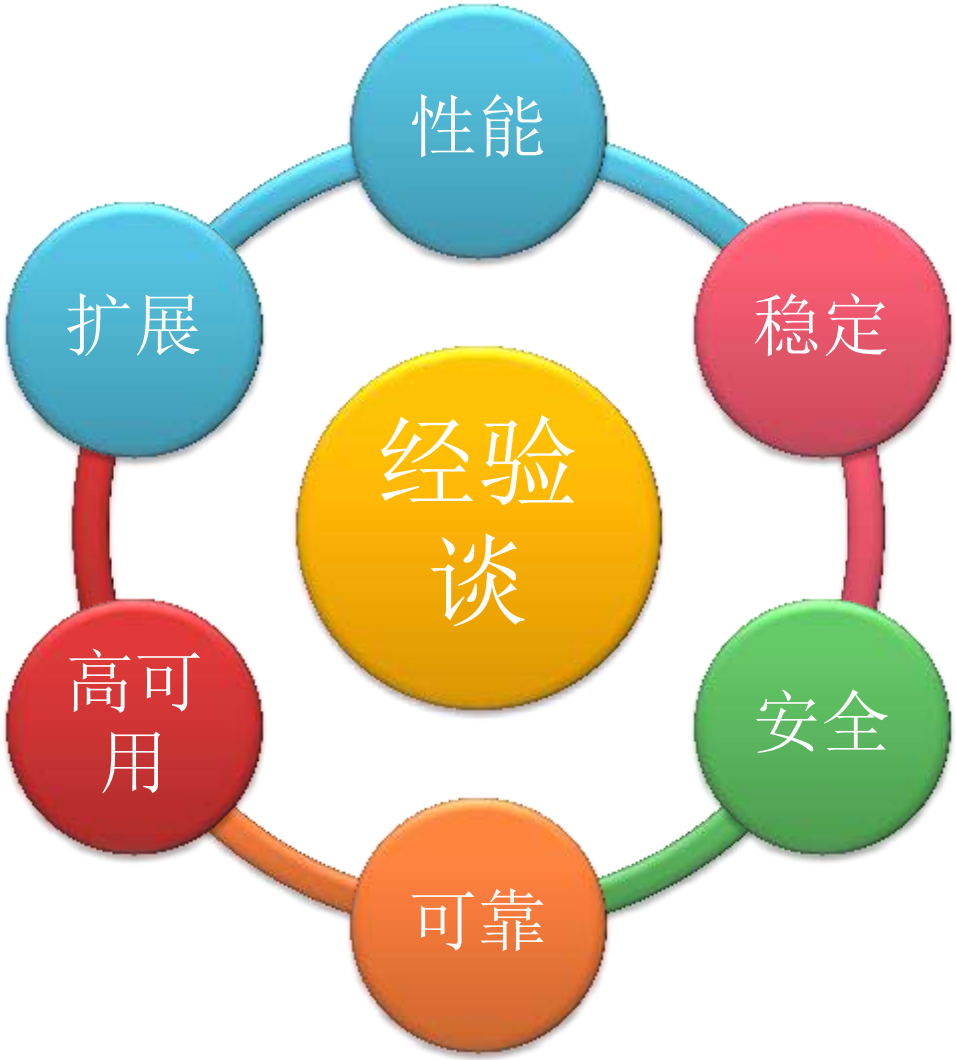
SKY mobi
新手机 新应用 新娱乐

浅析 PostgreSQL 特点

- PostgreSQL性能参考数据
  - HP DL360 G5(2*E5440, 14G MEM, 6*146G SAS 10K转)
  - CentOS 5.x x64, PostgreSQL 9.2
  - batch insert(20columns/tuple, 200bytes/tuple, 64tuples/batch , RESULT: 23Wtuples/s , 3616tps, 1.9ms/avgrequest)
  - insert(20columns/tuple, 200bytes/tuple, RESULT : 78579tps, 0.2ms/avgrequest)
  - select(5000Wtuples/table, 10Gbytes/table, 20columns/tuple, 200bytes/tuple, 按pk查询, RESULT : 72113tps, 0.17ms/avgrequest)
  - update(1000Wtuples/table, 2Gbytes/table, 20columns/tuple, 200bytes/tuple, 按pk更新, RESULT : 28474tps, 0.48ms/avgrequest)
  - delete(1000Wtuples/table, 2Gbytes/table, 20columns/tuple, 200bytes/tuple, 按pk删除, RESULT : 41372tps, 0.28ms/avgrequest)

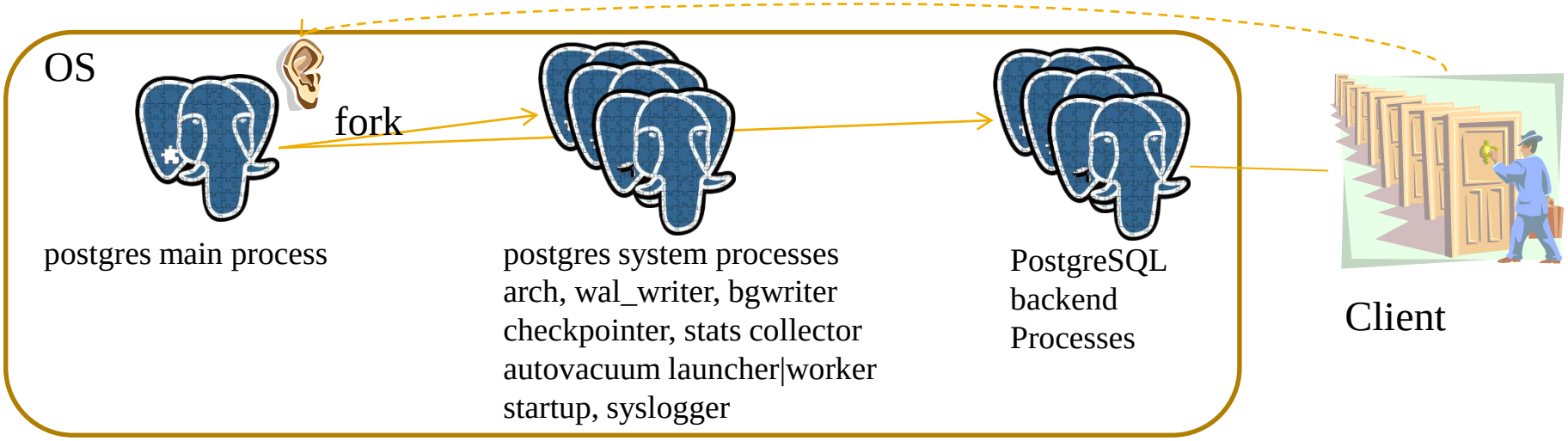
SKY mobi
新手机 新应用 新娱乐

如何构建一个高效,稳定,安全,易扩展的PostgreSQL数据库系统



SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 性能



- CPU cache
- shared buffer, wal_buffer
- OS cache
- Storage RAID Card cache
- hard disk buffer
- hard disk medium

Access Speed Down

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 性能

- CPU的考量要点
  - 每个postgresql进程在工作时最多只分配一个CPU核使用.
  - 高并发的场景同系列的多核CPU比较有优势
  - 低并发复杂运的算场景同系列的频率高的CPU比较有优势
- 内存的考量要点
  - 并发连接数, 活跃数据, shared_buffer的配置大小, 常用SQL中group by, distinct的比例和work_mem的配置大小, autovacuum work进程的配置数量以及maintenance_work_mem的配置大小.
- OS配置
  - 推荐 linux x64位
  - 内核参数 (kernel.shmmax , kernel.shmall , kernel.shmmni , kernel.sem , net.core.rmem_default , net.core.rmem_max , net.core.wmem_default , net.core.wmem_max , vm.overcommit_memory , fs.aio-max-nr)
  - limit (nofile , nproc , core , memlock)
  - blockdev --setra 扇区(512字节)

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 性能

- 存储配置
  - cache (打开带断电保护的读写cache)
  - raid (建议RAID10, 更新频繁的数据库不建议使用RAID5)
  - medium (ssd or 机械盘) (更新频繁的数据库建议使用满足IOPS需求的存储或SSD, 或将频繁更新的数据表空间放到SSD上)
- 网络的考量要素
  - 几个需要着重考虑网络的场景 (big result [set] , backup , 容灾)
- 连接池
  - PostgreSQL和客户端交互是一对一的进程模式, 因此不适合直接与高并发的短连接应用直接连接. 加连接池是比较好的解决办法. (如pgbouncer)
- 数据库编译时需要考量的要点
  - with-wal-blocksize(根据pg_test_fsync结果选择最优) ,
  - with-blocksize(OLAP系统或经常需要读取大量数据进行分析的系统可以选择较大的blocksize)

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 性能

- 文件系统考量
  - ext4, xfs, zfs
- 数据库集群初始化时需要考量的要点
  - --encoding (满足需求的情况下选择单字符占有字节数少的, 方便未来和其他业务线或国际化的话建议选择UTF8)
  - --locale (C)
  - --xlogdir (选择独立的存储)
  - --pgdata (选择独立的存储)
- 影响性能的postgresql.conf主要参数
  - shared_buffers
  - work_mem
  - maintenance_work_mem
  - vacuum_cost_delay , vacuum_cost_limit
  - synchronous_commit , wal_writer_delay , wal_buffers

SKY mobi
新手机 新应用 新娱乐

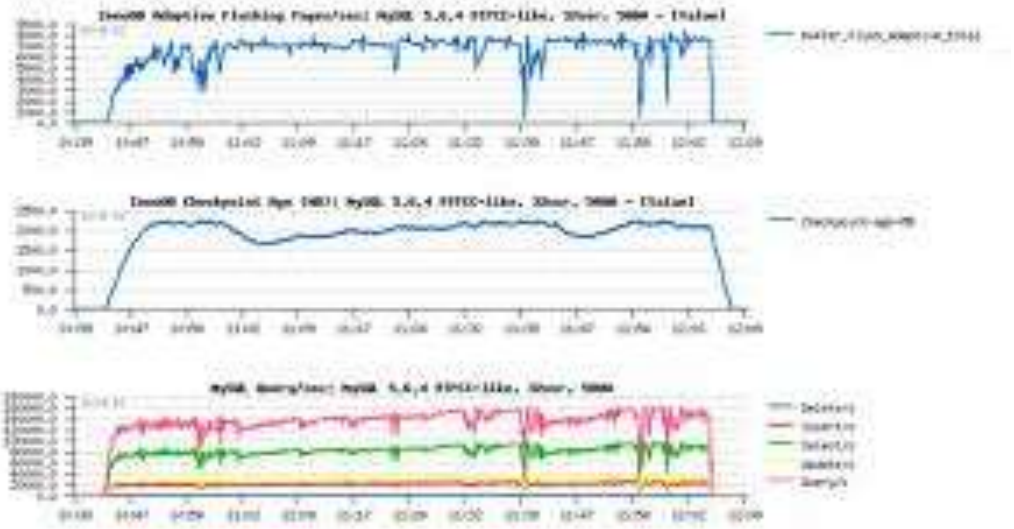
PostgreSQL经验谈- 性能

- 影响性能的主要参数 : postgresql.conf
  - wal_sync_method
  - commit_delay , commit_siblings
  - checkpoint_segments , checkpoint_completion_target
  - random_page_cost , effective_cache_size
  - vacuum_cost_delay, vacuum_cost_limit
  - autovacuum , autovacuum_vacuum_cost_delay, autovacuum_vacuum_cost_limit
  - log_lock_waits , deadlock_timeout
  - temp_tablespaces

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 稳定

- SQL平均响应时间抖动
- 每秒处理SQL请求数抖动
- 抖动和什么有关?
  - checkpoint
  - 突发性资源争抢
  - 锁等待
  - ...
- 稳定性目标
  - 不影响业务
    - 如数据库异步操作
    - 性能在业务接受范围内



SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 安全

- 权限最小化
  - pg_hba.conf
  - iptables
  - 角色权限
- 防止SQL注入
  - 应用层过滤器, 禁用简单查询协议(Simple Query protocol (PQexec))
- 密码复杂度策略
  - 长度, 组成(大小写,数字,特殊字符)
- 审计
  - 登陆,登出,DDL,
- 密码加密传输, 存储(encrypted), 数据加密传输, 存储
- 防火墙
  - 外网访问限制
- 不建议数据库直接暴露在外网上

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 可靠

- 单节点的PostgreSQL如何确保数据可靠性?
  - (XLOG fsync)确保成功提交的数据在发生异常后可恢复, 未提交的事务在发生异常后回退.
    - 注意这里指的异常不包括存储故障
  - fsync = on
  - full_page_writes = on
  - synchronous_commit = on
  - 关闭没有断电保护的持久化存储的cache(一般指硬盘), (存储,RAID卡一般都会带断掉保护电池)
- 一个降低可靠性, 不破坏数据一致性, 但是可以大大提高性能的参数
  - (D in ACID)
  - synchronous_commit = off
  - 关闭synchronous_commit后, 正常关闭数据库不会导致数据丢失. 异常DOWN机或DOWN库则最多可能丢失wal_writer_delay*3的最近时间的xlog信息. (也就是未从wal_buffers flush到xlog file的信息)

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 高可用

- 数据库高可用的显著特点
  - 要求数据可靠性
- 为什么要求数据可靠性?
  - A刷卡消费了1000元购买商品, 在数据库中完成了这笔事务.
  - 数据库DOWN机
  - Failover , 数据库从主库切换到备份库
  - A消费1000元购买商品的这个事务不能丢失.
- 久经考验的高可用方案
  - RHCS, 共享存储, VIP
    - 共享存储, 节点切换前要fence主库, 确保主库不再对数据发生操作. 同一时刻只有一个节点在对数据文件进行读写.
    - 缺陷
      - 不能低于存储故障的风险.
      - 如果配置的多播心跳, 网络抖动会造成节点切换. 网络情况不好可选仲裁磁盘.

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 高可用

- 久经考验的高可用方案
  - 同步流复制, 不共享存储, VIP
    - 优点是不需要采购昂贵的存储, 可以抵御存储单点故障.
    - 切换前fence主库
    - 缺点是同步流复制对性能有影响(下降5%左右).
      - http://blog.163.com/digoal@126/blog/static/16387704020119220345876 5/
  - 异步流复制不适合引入高可用方案, 原因是会导致切换后丢失事务信息. 如A消费的事务可能在切换后丢失. (当然如果业务系统容忍这种丢失, 也可以考虑的.)

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 扩展

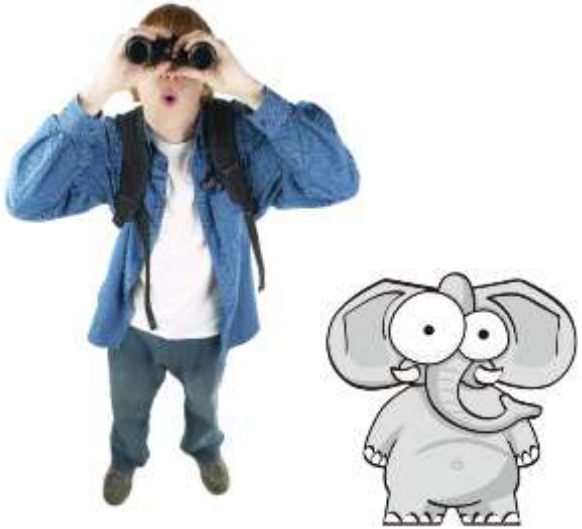
- 扩展应该和应用紧密结合
  - 和应用不相干的扩展举例
    - 中间件(pgpool-II), 读写分离, 分布式.
- 和应用结合数据库扩展
  - 读写分离(流复制), 分表
  - 分库
    - 跨库事务, 全局一致性备份和还原(目前只有PostgreSQL-XC做到, 因为它是有一台GTM统一管理事务ID的)
    - 跨库关联, 原来在一个库中可以通过JOIN来完成, 分库后需要分成多步操作.
    - 分库的原则是尽量避免跨库查询和跨库事务, 一般按照频繁作为条件出现的字段进行分库. 如userid取模.
- plproxy, 一个数据库代理, 部署分布式数据库比较方便
- OLTP speedup 见后面的优化章节

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 扩展

- 焦点
  - 让系统跑得好, 不能只关注数据库.
  - 做数据库优化, 也不能仅关注数据库

- 平衡
  - 让数据库各方面能力(CPU,内存,IO)发挥得当



SKY mobi
新手机 新应用 新娱乐

PostgreSQL in APP SYSTEM

扩展考量 :
1. 持久化MQ
减少同步的数据库操作.
2. 付费提速(水平扩展)

最终用户

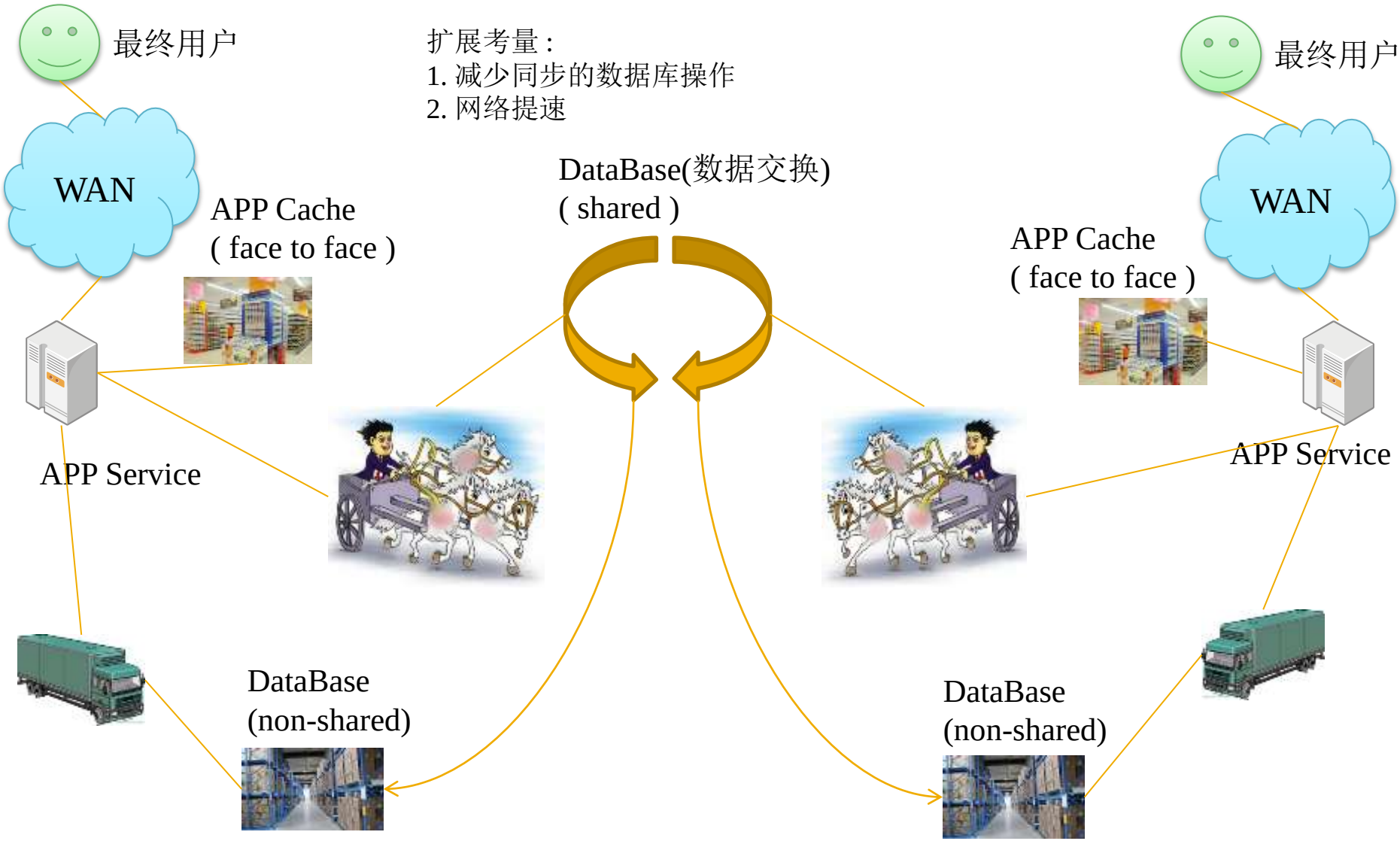
APP Cache (face to face)

APP Service

DataBase

SKY mobi
新手机 新应用 新娱乐

PostgreSQL in APP SYSTEM



SKY mobi
新手机 新应用 新娱乐

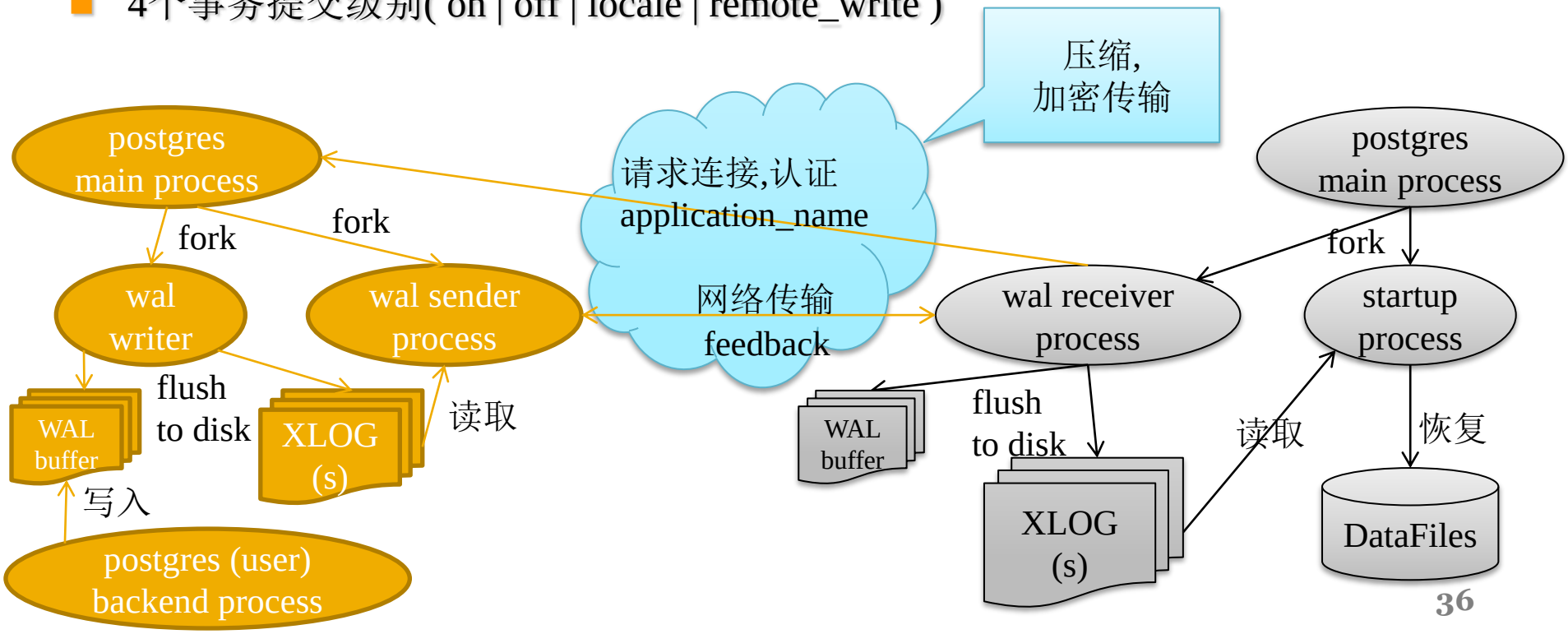
PostgreSQL经验谈- 容灾

- 容灾和备份的区别
  - 异地容灾的出发点是应对机房故障，注重异地镜像，实时性和性能影响.
  - 备份的出发点是应对逻辑错误和物理错误，注重历史数据和归档保留以及性能影响.
- 异地容灾考虑
  - 网络因素
  - 异步异地容灾
    - synchronous_commit=local
  - 同步异地容灾
    - synchronous_commit=remote_write

SKY mobi
新手机 新应用 新娱乐

流复制容灾

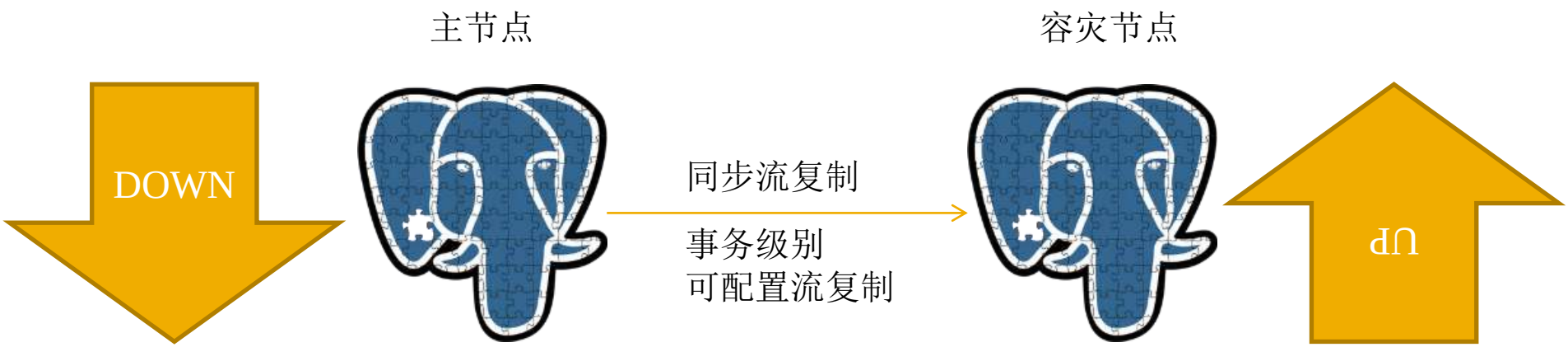
- 利用流复制容灾的好处
  - 对容灾节点到主节点的网络要求比较宽泛, 网络情况不好, 丢包, 中断都不好破坏容灾节点的数据一致性, 仅带来数据传输的延迟(可以想象为断点续传). 前提条件是确保主节点的xlog信息完整(主节点通过配置归档或者wal_keep_segments来保留xlog文件).
- 4个事务提交级别(on | off | locale | remote_write)



SKY mobi
新手机 新应用 新娱乐

同步流复制容灾

- synchronous_standby_names = comma-separated list of application_name from standby(s); '*' = all
- synchronous_commit = off, local, remote_write, or on
- 由于同步复制需要等待同步复制节点返回已将xlog记录写入buffer(remote_write)或者flush到磁盘(on)了,事务才算成功提交.所以对网络延时和standby节点的IO性能要求很高.



容灾节点与主节点保持一致.
UP后不会丢失最后一个成功提交的
同步提交事务以及此事务之前的数据.

SKY mobi
新手机 新应用 新娱乐

同步流复制容灾

- 主节点：
- success Transaction A synchronous_commit = locale | off
- success Transaction B synchronous_commit = remote_write | on
- success Transaction C synchronous_commit = locale | off
- 异常DOWN 机

- 同步流复制节点：
- 可以确保Transaction A and B的xlog数据已经到达, remote_write(到达wal buffer), on(到达xlog file). 因此数据不会丢失.
- Transaction C 如果没有来得及传输到standby节点, 则可能丢失. 主节点每次commit之后都会向standby节点发送xlog file里面新产生的数据. 实时性极高. 因此丢失的量非常少. 所以事务C在standby节点可能丢失也可能没丢失.

SKY mobi
新手机 新应用 新娱乐

同步流复制容灾

- 同步流复制容灾折中的方案, 选择不同的事务提交方法.
- 同步流复制都建议采用多个standby节点. 单个standby的情况下, standby DOWN机将导致同步提交的事务无法继续.
  - 如果发生这种情况可以在主节点设置synchronous_commit = locale或临时关闭同步复制, 并让app重新建立和数据库的连接.
- 关键业务, 关键事务, synchronous_commit = remote_write | on
  - 设置为remote_write指XLOG信息写入主节点的xlog file和standby节点将接收到的xlog写入buffer后. 事务才可以提交完成. 这个降低了对standby节点的IO要求, 缩短了同步事务在standby上的时间开销.
  - 设置为on指XLOG信息写入主节点的xlog file和standby节点将接收到的xlog写入xlog file后. 事务才可以提交完成. 对standby节点的IO能力要求较高, 否则会影响同步事务的处理耗时.
- 关键业务, 非关键事务, synchronous_commit = locale
  - 设置为remote_write指XLOG信息写入主节点的xlog file, 事务才可以提交完成.

SKY mobi
新手机 新应用 新娱乐

同步流复制容灾

- 非关键业务, 非关键事务, 对写性能要求很高, 可以选择synchronous_commit = off
  - 设置为off指XLOG信息写入wal buffer后, 事务才可提交完成, 在数据库DOWN机后, 在buffer中还未flush到磁盘的wal信息将丢失, 也就是说这部分XLOG信息如果要用于恢复数据库的话, 则无法恢复这部分信息, 最大丢失的数量是wal_write_delay*3的时间范围. 注意正常的关库不需要recover, 所以不会丢失数据.

- synchronous_commit = on 事务保护最高级别, 一般只建议用在局域网内的容灾

- 事务级的synchronous_commit设置
- BEGIN;
- SET LOCAL synchronous_commit TO on | local | remote_write | off;
- Other SQLs;
- END;

SKY mobi
新手机 新应用 新娱乐

异步流复制容灾

- 与同步流复制容灾不同之处
  - 不需要配置synchronous_standby_names
  - 主节点事务提交成功返回前不需要等待standby节点接收到的xlog数据写入wal buffer或xlog file.
  - 性能方面优于同步流复制
  - 数据保护级别低于同步流复制

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 备份

- 备份分类
  - 逻辑备份
    - 一致性备份, 可以恢复到备份开始的时间点.
    - 缺点, 还原时间长, 需要重新收集统计信息. 优点, 可以选择表, schema, database进行备份, 支持跨平台数据还原.
  - 数据文件(物理)备份
    - 一致性备份, 备份数据文件和XLOG归档文件, 可以恢复到备份结束后的任意时间点.
    - 缺点, 文件较大. 优点, 还原时间和XLOG数量有关, 一般建议勤备份基础数据文件. 缩短恢复耗时.
    - 《PostgreSQL Selectivity Tablespace PITR》

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 备份

- 典型的物理备份
  - digoal=# select * from pg_start_backup(now()::text);
  - rsync -acvz --exclude=pg_xlog $PGDATA/* backup_host::backup_pgdata/$DATE
  - digoal=# select * from pg_stop_backup();

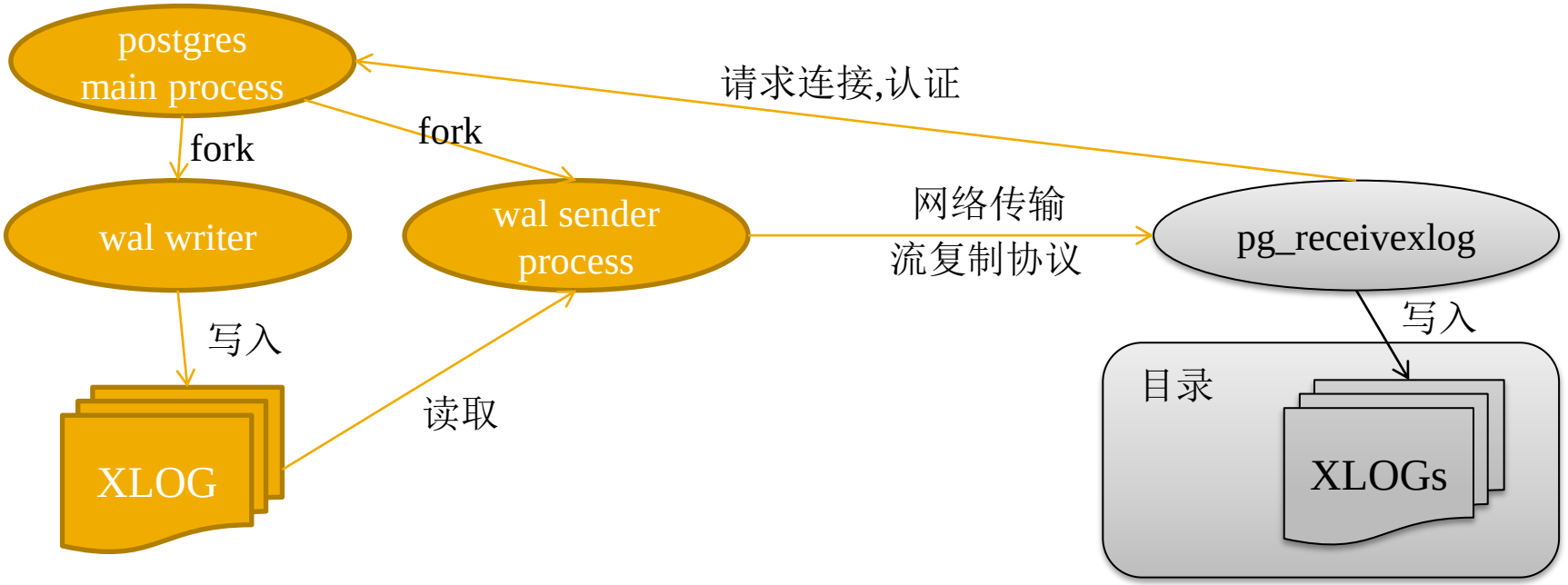
  - 或者使用pg_basebackup代替以上步骤
  - pg_basebackup -D $backup_dir/$DATE -f p -x stream -h $PRIMARY -p $PGPORT -U $replica

  - XLOG归档文件的备份可以在archive_command里面就做成同时写原程和本地. 因此计划好基础数据文件备份就好了.

SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 备份

pg_receivexlog, 利用流复制协议进行**实时归档.** 增强数据保护级别.
PREPARE
1.主节点数据库上创建一个superuser或replication角色用于流复制
2.配置主节点(或级联standby节点)的pg_hba.conf允许客户端和主节点建立流复制的连接
host replication postgres ip地址 trust
3.配置主节点(或级联standby节点)wal_level(archive/hot_standby), max_wal_senders(>0)
4.确保远程归档节点和主节点(或级联standby节点)网络(数据库监听端口)可达



SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 还原

- 平常在重大操作前勤创建还原点(用于物理还原)
  - pg_create_restore_point
- 逻辑还原
  - custom格式的备份文件可以使用pg_restore -j 并行还原.
  - 如果备份文件比较大, 需要做全还原的话, 想调整还原的表的顺序怎么办?
    - PostgreSQL Logical Backup's TOC File
- 物理还原
  - PITR还原目标
    - recovery_target_name
    - recovery_target_time
    - recovery_target_xid
  - PostgreSQL datafile backup and recovery case

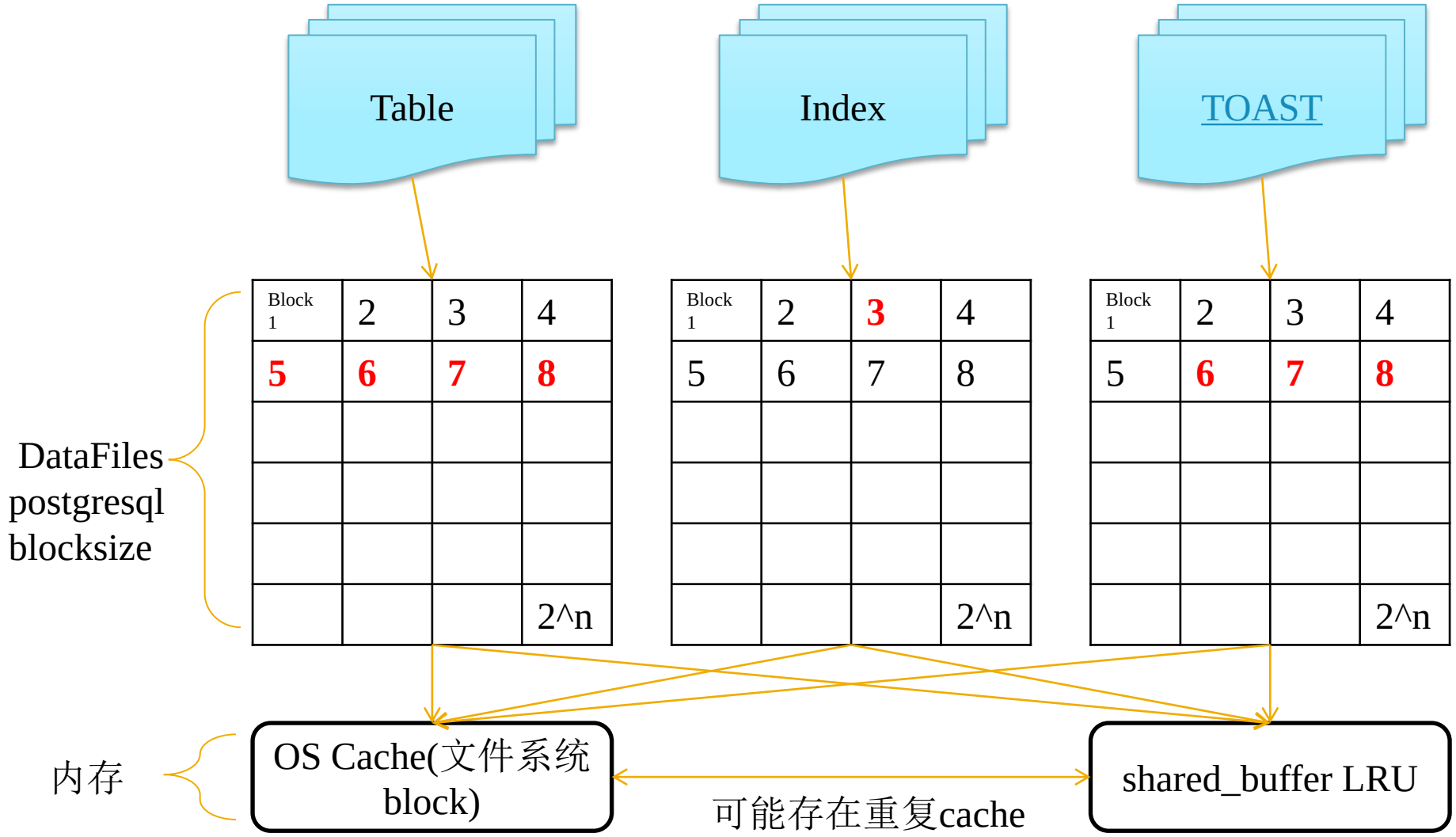
SKY mobi
新手机 新应用 新娱乐

PostgreSQL经验谈- 优化

- OLTP优化
- OLTP系统特点
  - 高并发
  - 小事务, 查询(多), 新增(多), 更新(多), 删除(较少)
  - 备份
- Speedup
  - 查询
  - 新增
  - 更新, 删除
  - 备份

SKY mobi
新手机 新应用 新娱乐

speedup select



SKY mobi
新手机 新应用 新娱乐

TOAST

- TOAST和HEAP TABLE存储不太一样，TOAST表一般包含如下字段：

  tableoid -- TOAST表的OID

  cmax

  xmax

  cmin

  xmin

  ctid

  chunk_id -- HEAP表通过TOAST pointer把这行中一个被TOAST的列关联到这里

  chunk_seq -- 同一个chunk_id如果大于TOAST_MAX_CHUNK_SIZE，将被切片存储。这里存储切片后的序号。

  chunk_data -- 真实的数据

  chunk_id + chunk_seq = primary key

SKY mobi
新手机 新应用 新娱乐

speedup select

- SQL层面的优化
  - 索引
    - 选择合适的索引访问方法 (btree, gist, gin, btree-gist, sp-gist, hash),
    - 选择合适的列, (驱动列),
    - 选择合适的条件(partial index),
  - 选择合适的数据类型
    - range类型的例子.
      - **PostgreSQL range gist index 20x+ speedup than Mysql index combine query**
      - IP地址匹配的应用案例, 原来使用的是两个int字段进行范围匹配. btree索引. 3600 tps. 修改成int8range类型, gist索引, 8Wtps.
  - 合适的JOIN算法
    - nestloop, hashjoin, mergejoin
  - 合适的成本
    - random_page_cost

SKY mobi
新手机 新应用 新娱乐

speedup select

- IOPS优化
- OS Cache
  - 数据库关闭后还可能存在
  - posix_fadvise
- shared_buffer
  - 数据库关闭后清空, LRU算法

- 利用pgfincore speedup select
  - 持久化 POSIX_FADV_WILLNEED
  - 抛弃 POSIX_FADV_DONTNEED
  - 镜像, 预读 至内存(-- 保存表,索引,TOAST对应的OS PAGE CACHE状态表的snapshot, -- 通过状态表预加载OS PAGE CACHE)

SKY mobi
新手机 新应用 新娱乐

extend select

- 利用流复制或其他复制技术扩展查询
  - hot standby
    - 平台必须一致, 操作系统种类一致, 数据库版本要求一致
  - londiste3
    - 跨平台, 跨数据库版本, 跨操作系统的复制
    - 基于触发器, 延时相比hot standby高, 性能影响相比hot standby高
  - 其他复制软件
    - slony-I
    - pgpool-II
    - ...
  - 应用层缓存
    - memcache
    - redis

SKY mobi
新手机 新应用 新娱乐

speedup complex select

- 将运算迁往APP层, 简化数据库SQL
- 子查询过多或JOIN表数量过多的情况下可以考虑让优化器按照SQL书写顺序进行关联, 杜绝优化器选择不优执行计划的情况.(from_collapse_limit=1 | join_collapse_limit=1)
- 数据冗余, 减少JOIN
  - 适当的数据冗余减少跨表查询
- 宽表
  - 把窄表返回大量条记录改为宽表返回少量记录可以降低查询的IO请求, 对内存小且存储IO能力不强的场景比较有效. 但是注意, 宽表可能带来一个问题(热点更新)
- 事务级work_mem调整, 优化GROUP BY, 排序.
  - 如 set local work_mem to '512MB';
- 应用层缓存, 将实时查询转非实时性查询
- 分区表查询优化, 1. constraint_exclusion, 2. 封装SQL时直接使用子表, 3. 分区字段上使用约束和索引, 4. SQL在分区字段带过滤条件时, 使用常数.

SKY mobi
新手机 新应用 新娱乐

speedup insert

- batch insert
  - insert into table (columns) values (),(),...,();
    - (20W+ tuples inserted per second)
  - copy
    - (40W+ tuples inserted per second)
  - begin; sqls; end;
    - (10W+ tuples inserted per second)
- 分组提交, 减少IO请求次数.
  - commit_delay
  - commit_siblings
- 异步提交
  - wal_writer_delay
  - synchronous_commit
- 选择合适的wal_sync_method (pg_test_fsync)
- 选择合适的wal_blocksize (pg_test_fsync)
- xlogdir与其他目录分开物理存储

SKY mobi
新手机 新应用 新娱乐

speedup update

- 使用消息队列(如redis的list数据类型), 异步更新或批量更新
- 消除热点更新行
  - 假设应用程序需要并发的频繁的更新同一行记录, 由于行锁的原因, 同时只有一个客户端获得这个锁, 其他客户端都在等待, 所以对这条记录的更新就是串行的, 假设更新耗时1毫秒, 那么一秒钟就只能处理1000次这样的更新请求.
  - 优化方法是修改业务逻辑, 例如把更新1行变成更新多行.
    - http://blog.163.com/digoal@126/blog/static/163877040201111154255438/
- 频繁更新的数据放到内存或SSD
- 拆分表(表拆分的好处, 可以分布到不同的表空间, autovacuum颗粒度更小, prevent wrap的概率变小, 维护更加方便, 如添加默认值字段时间缩短, 拆成多表后还为以后数据库服务器扩展提供方便了)
- 拆分库, 逻辑规则(业务数据和日志数据分离, 各业务子模块拆分, 按用户ID 2^n取模拆分)
  - 分布式数据库模块 plproxy

SKY mobi
新手机 新应用 新娱乐

speedup update

- 利用内存加速update
- [参考数据]
  - http://en.wikipedia.org/wiki/IOPS

内存
iops :
几十万

io Memory
iops :
几十万

SATA3 SSD
iops :
几千到几万

SATA3 DISK
iops :
100多

speedup update

- 利用Linux tmpfs加速update (解决表空间的 io 瓶颈, 当XLOG部署瓶颈时)
- 单节点部署(流/log shipping复制不需要两边目录结构一致, pg_tblspc,pg_xlog都可以用软链接解决, 因此可以在单节点部署流复制), 如图

synchronous_commit = locale

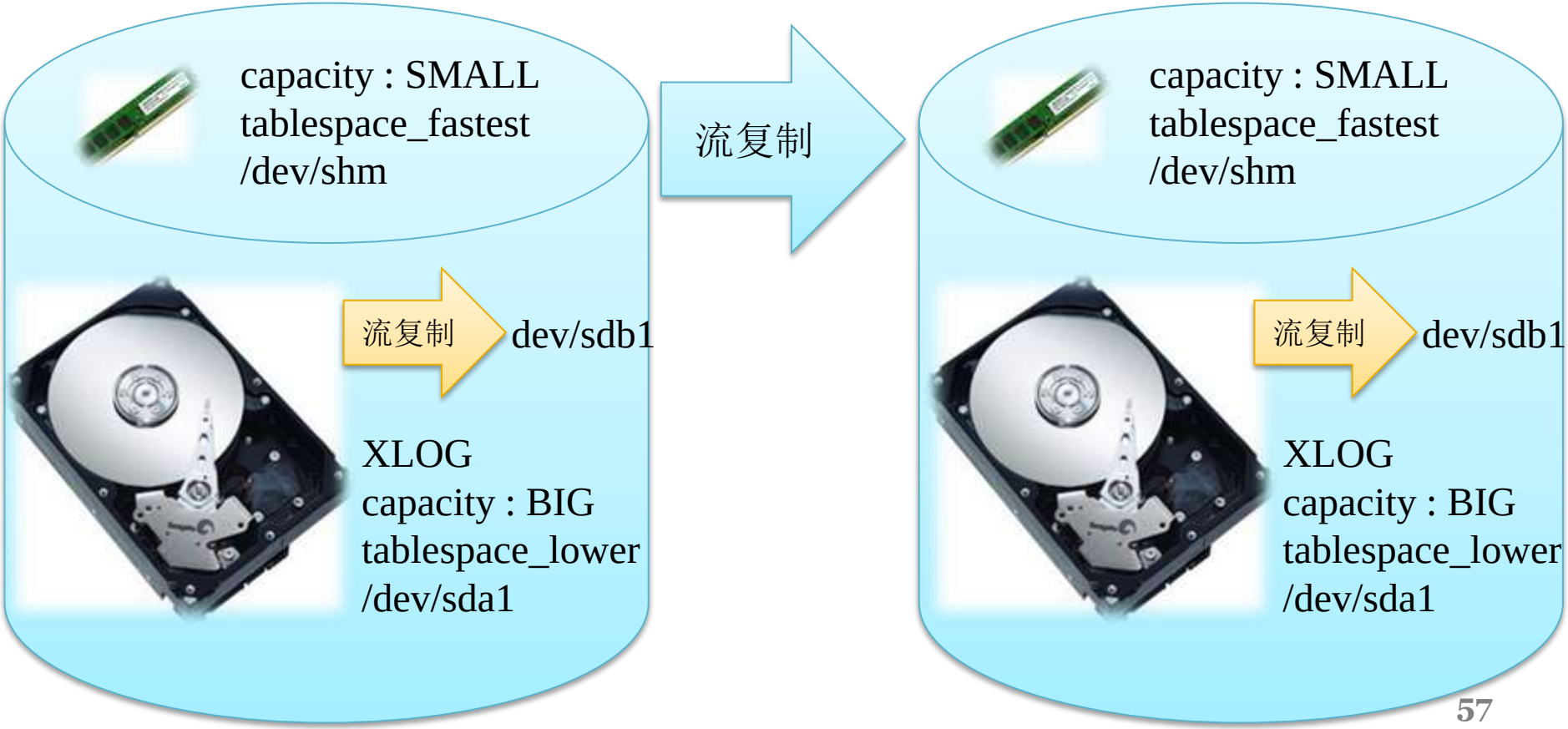
Primary :
$PGDATA : /pgdata01/pg_root
tbs_fastest : $PGDATA/pg_tblspc/16384
-> /dev/shm/tbs_fastest
tbs_lower : $PGDATA/pg_tblspc/16385
-> /pgdata02/tbs_lower
pg_xlog : $PGDATA/pg_xlog
-> /pgdata03/pg_xlog

Standby :
$PGDATA : /pgdata_std/pg_root
tbs_fastest : $PGDATA/pg_tblspc/16384
-> /pgdata_std/tbs_fastest
tbs_lower : $PGDATA/pg_tblspc/16385
-> /pgdata_std/tbs_lower
pg_xlog : $PGDATA/pg_xlog
-> /pgdata_std/pg_xlog

speedup update

- 利用Linux tmpfs加速update (解决表空间的 io 瓶颈, 当XLOG部署瓶颈时)
- 多节点部署(安全起见, 确保至少有一份tbs_fastest的数据存在持久化存储里面)
  如图 :

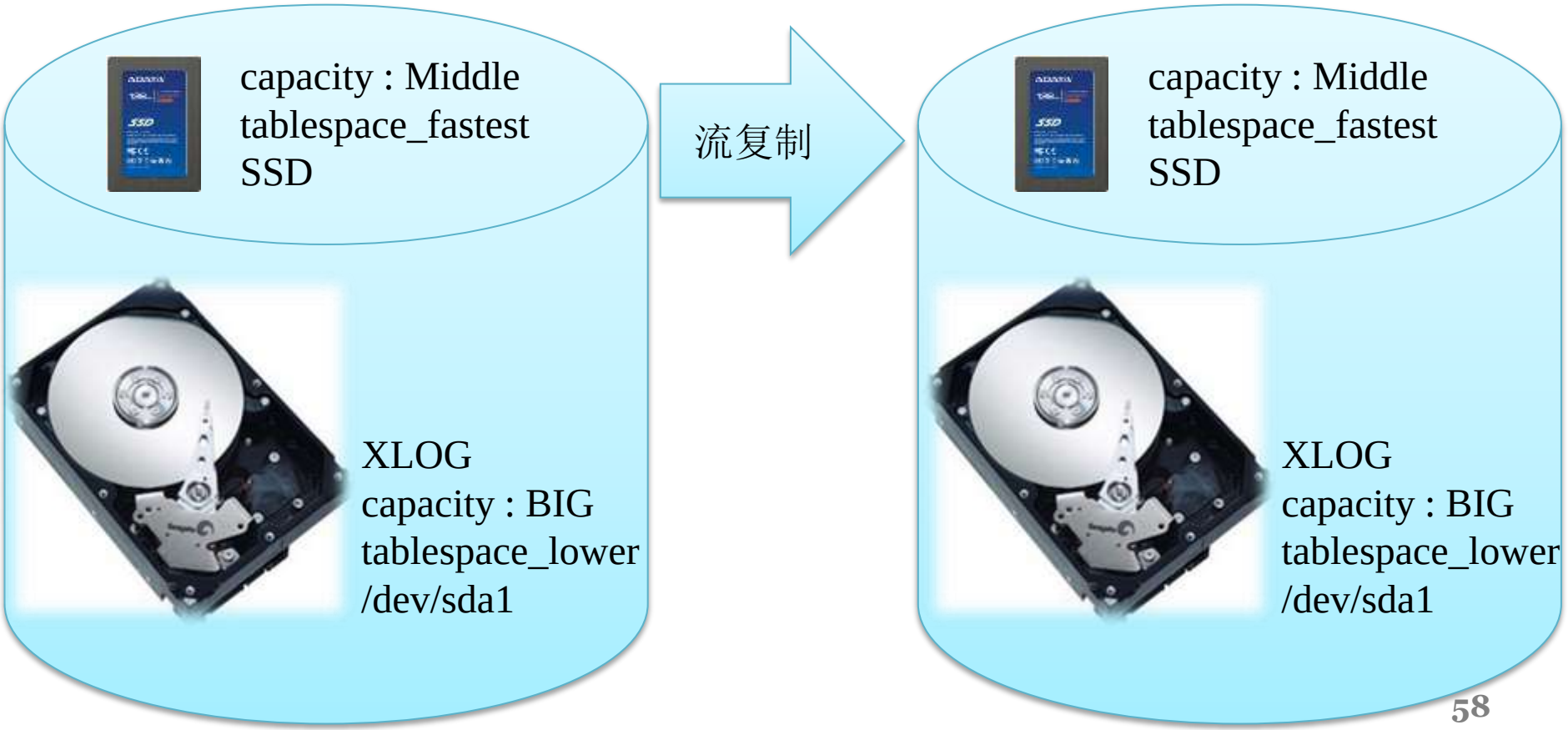
synchronous_commit = remote_write | on



speedup update

- 比较靠谱的方案, SSD, 容量越来越大, 趋于成熟, 成本越来越低.
- 多节点部署如图 :

synchronous_commit = remote_write | on



SKY mobi
新手机 新应用 新娱乐

speedup delete

- 队列, 异步删除
- 频繁删除的数据放到内存或SSD
- 放内存的方案, 同步流复制
- 按业务规则拆表
  - 例如按历史时间清理的数据库, 表按时间分区

SKY mobi
新手机 新应用 新娱乐

speedup backup

- 备份分类
  - 在线数据文件备份
  - 在线逻辑备份
- 备份操作的特征
  - 文件全扫描, 连续读取
- 优化方法
  - 降低对生产的影响(IO争抢)
    - 备份放在standby节点进行.
    - max_standby_archive_delay
    - max_standby_streaming_delay
  - 备份提速
    - RAID卡或存储控制器的读缓存
    - RAID卡或存储控制器的读预取
    - OS层面通过blockdev 设置块设备的readahead

SKY mobi
新手机 新应用 新娱乐

max_standby_archive/streaming_delay

xlog 接收到的位置

假设在这段时间xlog的恢复一直没有赶上XLOG的接收位置, 当这个时间范围超过设置的delay时, 执行的SQL将被强制cancel掉.

xlog 已恢复到的位置(当已恢复位置赶上接收到的位置时开始重新计时)

SKY mobi
新手机 新应用 新娱乐

PostgreSQL 迁移

- 异构平台迁移(如 Oracle -> PostgreSQL, MySQL -> PostgreSQL)
- 跨版本, 跨平台迁移
  - Linux -> AIX
  - Linux 32bit -> Linux 64bit
  - 同平台跨版本迁移 8.3 -> 9.1
    - 迁移方法, pg_upgrade
  - common迁移方法
    - 慢, 一致性迁移的话影响业务久, pg_dump, pg_restore
    - 复杂, 一致性迁移电话影响业务短, londiste3, 增量迁移数据(需主键)
    - 两者结合
- 相同大版本迁移
  - 8.3 -> 8.3
    - log-shipping warm-standby
  - 9.1.3 -> 9.1.3, 9.1.4
    - stream replication standby
    - 小版本不一致, 迁移前先升级到一致版本.

SKY mobi
新手机 新应用 新娱乐

PostgreSQL 压力测试

- 重在建模
- PostgreSQL压力测试
  - pgbench
    - Use pgbench test Your PostgreSQL DBSystem performace
    - http://blog.163.com/digoal@126/blog/static/16387704020115153463131 3/
    - PostgreSQL性能优化综合案例讲解
    - http://blog.163.com/digoal@126/blog/static/16387704020122138215085 8/
  - pg_test_fsync
    - 测试调用OS各种同步写接口写wal-blocksize的性能
  - pg_test_timing
    - 测试系统时钟是否正常(不会回退), 以及取时overhead
- 数据库产品横向压力测试
  - python, ruby, java, ...

SKY mobi
新手机 新应用 新娱乐

PostgreSQL监控

- 实时监控
  - nagios监控
    - check_postgres
      - https://github.com/bucardo/check_postgres
    - 自定义nagios监控脚本
      - 返回值(0正常, 1告警, 2严重错误)
- 非实时监控
  - sar
    - Use PostgreSQL collect and analyze Operation System statistics
  - pg_stat_statements
    - select query,calls,1000*(total_time/calls) from pg_stat_statements order by calls desc limit 10;
    - select query,calls,1000*(total_time/calls) from pg_stat_statements order by total_time desc limit 10;
    - select query,calls,1000*(total_time/calls) from pg_stat_statements order by total_time/calls desc limit 10;

SKY mobi
新手机 新应用 新娱乐

PostgreSQL监控

- Use PostgreSQL collect and analyze Operation System statistics
- 邮件内容图例：
- 1. ldavg_15 TOP10 :
- 2. rtps TOP10 :
- 3. wtps TOP10 :
- 4. iowait TOP10 :
- 5. swap_page_in_out TOP10 :
- 6. swap_usage TOP10 :
- 7. newproc_p_s TOP10 :

SKY mobi
新手机 新应用 新娱乐

PostgreSQL监控

- 非实时监控
  - 慢查询 (log_min_duration_statement)
  - stat动态视图快照, 分析数据量增长趋势, 内存的使用情况等.(如, pgstatspack)
  - sar 信息分析简介 :
  - # (min,max,avg) 负载排行, 例如是否需要建索引, 是否需要使用绑定变量等.
  - # (min,max,avg)每秒读请求排行, 例如是否需要建索引, 是否需要加内存, 是否需要对存储性能扩容等.
  - # (min,max,avg)每秒写请求排行, 是否需要减少索引, 是否需要对存储性能进行扩容等.
  - # (min,max,avg)iowait排行, 判断是读还是写居多之后, 例如是否需要加内存, 是否需要使用异步IO, 是否需要将常用数据放入内存, 是否需要对存储性能进行扩容等.
  - # (min,max,avg)swap in,out 排行, 例如是否需要加内存, 是否需要将常用数据放入内存等.
  - # (min,max,avg)swap使用比例排行, 例如是否需要加内存, 是否需要调整数据库参数, 是否需要使用大页等.
  - # (min,max,avg)每秒新建进程排行, 例如是否需要加个数据库连接池使用长连接, Oracle是否需要使用共享连接, 应用程序是否可以将短连接改成长连接的模式等.

SKY mobi
新手机 新应用 新娱乐

PostgreSQL Caveat

- 危险操作

- -- postgresql.conf
- fsync = off
- full_page_writes = off
  - 当文件系统不能防止partial page writes (zfs可以)
- autovacuum = off
  - 动态库(update, delete), 且没有规划的vacuum.

- -- 权限放大镜
- -- pg_hba.conf
- 使用trust认证

- 滥用超级用户(删库删表删表空间,删用户, 直接操作文件系统上的文件等)

- -- SCHEMA
- 不规划字段类型和长度, 将来扩字段将rewrite全表(9.2 有大幅改进, 大部分操作不需要rewrite表)
- 该用索引的地方不使用索引, 导致查询变慢, 影响业务
- 滥用索引, 导致UPDATE,DELETE变慢,同时浪费空间

SKY mobi
新手机 新应用 新娱乐

PostgreSQL Caveat

- 危险操作

- -- SQL
- 不使用绑定变量, 导致SQL硬解析, CPU杀手
- 原本可以直接查询子表的, 分区表中滥用主表进行查询, 导致SQL解析耗时与子表个数成正比, CPU杀手
- 业务逻辑造成的死锁, 业务杀手
- 分区表场景中不开启constraint_exclusion, 查主表时任意条件都将导致全子表扫描, CPU杀手

- -- 其他
- Planner Cost Constants配置不当, 执行计划毒药
- 表空间设计不合理, IO杀手
- 开启没有断电保护的或非企业级的硬件cache, 数据杀手

SKY mobi
新手机 新应用 新娱乐

PostgreSQL 展望

- 期待的
  - 单条SQL使用多核(类似Oracle的HINT), 这可以大大提高数据库逻辑恢复时创建索引的速度.
  - 可靠,稳定,易开发,易维护,易扩展的分布式OLTP集群, 目前PostgreSQL-XC刚刚release1.0.0 . pl/proxy比较成熟(但是pl/proxy无法做到在线的全局数据库一致性备份).
  - 降低checkpoint带来的性能影响. checkpoint时IO请求非常多, 当数据文件存储性能低下时, 影响较大. (IO的改进是否可以考虑降低离散IO的概率, 能组合成连续IO的flush请求尽量组合成连续IO. 其他手段, 增加增量checkpoint操作类型.)
  - 索引表, 目前的表是heap存储的. 索引表在某些场景下使用时比较高效, 如经常要按某列的顺序检索(使用CLUSTER适合静态表, 不适合动态表).
  - count()的改进, 9.2的index only scan已经是一个改进.

SKY mobi
新手机 新应用 新娱乐

PostgreSQL 展望

- 期待的
  - pgbouncer增加类似pg_hba.conf功能, 因为使用pgbouncer后, PostgreSQL根据客户端的连接用户名,库名,IP地址的过滤没有办法实现. PostgreSQL看到的IP是pgbouncer主机的, 而不是客户端主机的IP.
  - flash back query.
  - 函数封装(目前EnterpriseDB实现的wrap)
  - 二级缓存(如SSD硬盘作为除内存以外的二级缓存, 或使用SSD加速checkpoint)
  - 只读表空间, 表
  - 基于表空间, 表, 数据库, schema 的存储配额
  - 基于表空间或数据库的数据文件备份,恢复与传输
  - 逻辑备份或物理备份(基于数据文件的备份)的增量备份和还原的支持(如pg_basebackup增量备份, pg_dump增量备份, 增量还原.)

SKY mobi
新手机 新应用 新娱乐

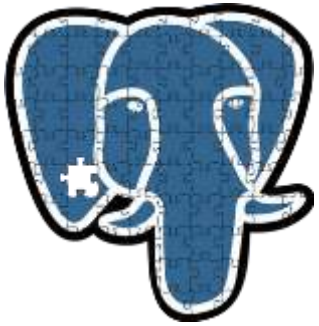
PostgreSQL 展望

- 期待的
  - 分区表性能优化(高效算法触发器), 简化分区表的创建/维护操作
  - real_testing数据采样, 模糊敏感数据.
  - 降低begin;end;的overhead
  - master-master复制.
  - 审计功能丰富(如基于库,用户,schema,表,列,定制化where过滤条件等的审计)

SKYMOBI 新手机 新应用 新娱乐

谢谢

PostgreSQL 2012 China Conference



- 【参考】
- http://www.postgresql.org/about/featurematrix/
- http://www.postgresql.org/docs/9.2/static/index.html

- 关于本PPT有问题请发邮件至 digoal@126.com
- QQ: 276732431
- 群: 3336901
- 【更多内容请关注】
- http://blog.163.com/digoal@126/