WwIT
驱动数据 成就未来

Run Your PG on ZFS

朱贤文

POSTGRESQL支持和服务，我们是专业的！

关于成都文武信息技术

- 成立于2014年
- 成都高新区天府软件园B7-611
- 专注于提供PostgreSQL商业支持和整体解决方案
- 业务范围
  - PostgreSQL 产品支持，技术服务
  - PostgreSQL 性能优化，紧急支援
  - PostgreSQL 远程运营和维护
  - PostgreSQL 咨询，培训及相关服务
  - 数据库迁移
  - 高可靠的高性能一体机

主题

ZFS及其特性介绍

ZFS运行数据库的优势

如何获取ZFS

ZFS的数据库应用：快照

ZFS的数据库应用：克隆

ZFS的数据库应用：日志和复制

Q&A

ZFS特新及其介绍

SUN出品，2005年，随Solaris发布，Enterprise Ready

数据自动动态分层

丰富完善的RAID级别

- RaidZ1 对应于RAID5，RaidZ2 对应于RAID6，RaidZ3 对应于RAID7，无写惩罚，无写漏洞
- 支持Mirror，Strip使用，可以做成传统的raid10,raid50/60/70等

重复数据删除

数据压缩

快照，克隆

远程复制

如何获取ZFS

IllumOS
- OmniOS
- OpenIndiana
- SmartOS
- HungHu

FreeBSD

Linux

OS X

Also, Oracle Solaris

运行和维护数据库的任务

不常见的情形
- 服务器故障
- 在同一个raid组内的多个硬盘同时顺坏
- 数据中心失效

常见的人为错误
- 删除数据
- 变更数据
- 删除表

大多数备份方案专注于最最有可能的失效的情形

ZFS能帮助保护更多的常见的问题

开发测试常见的情况

在一个多个T的数据库上测试一个升级的脚本

升级失败，快速回滚到升级前的版本

灵活地PITR支持可以快速恢复到指定的日志

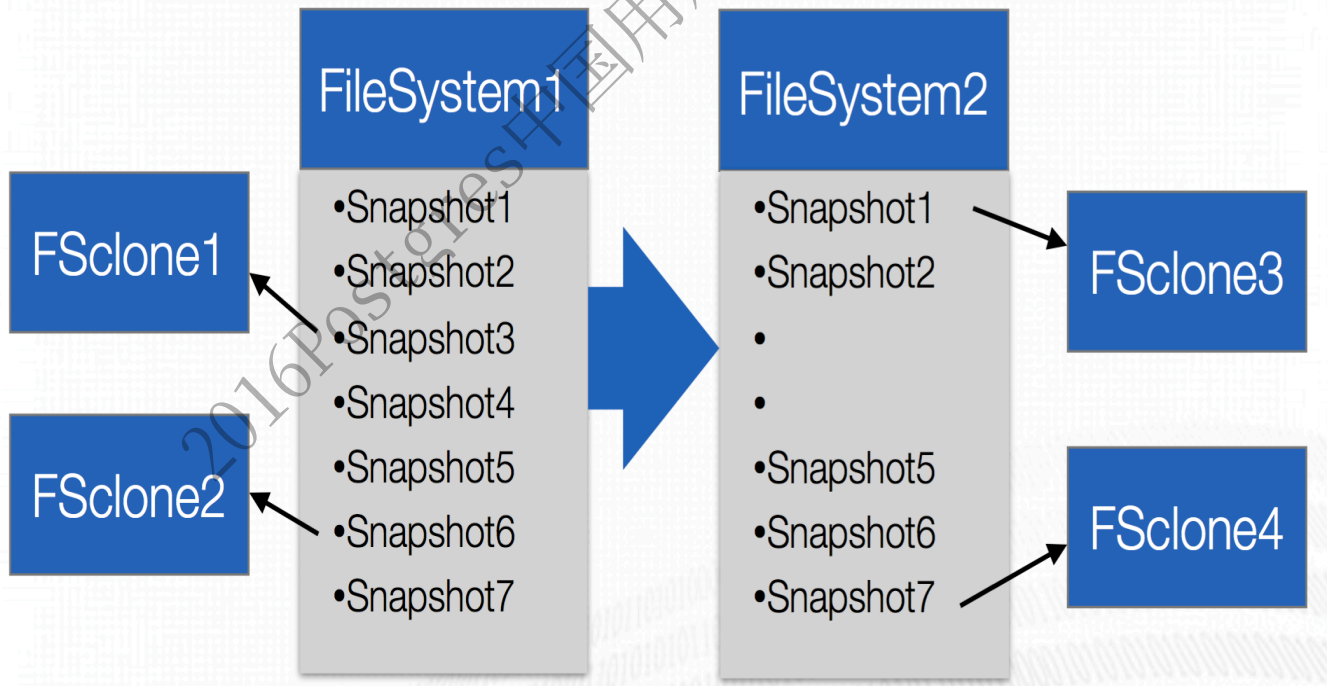
ZFS运行数据库的优势

- 快速有效的复制（单向周期性的更新）
- 低 / 没影响的快照
- 通过克隆读写快照
- 池化的物理设备
- 发送 / 接收快照
- 通过快照容易做增量复制
- 通过SSD做缓冲，提高性能, L2ARC
- 升级硬件时能做到零停机
- 连续的一致性检查和自动修复
- 内置压缩，加密等高级功能，提升数据库性能

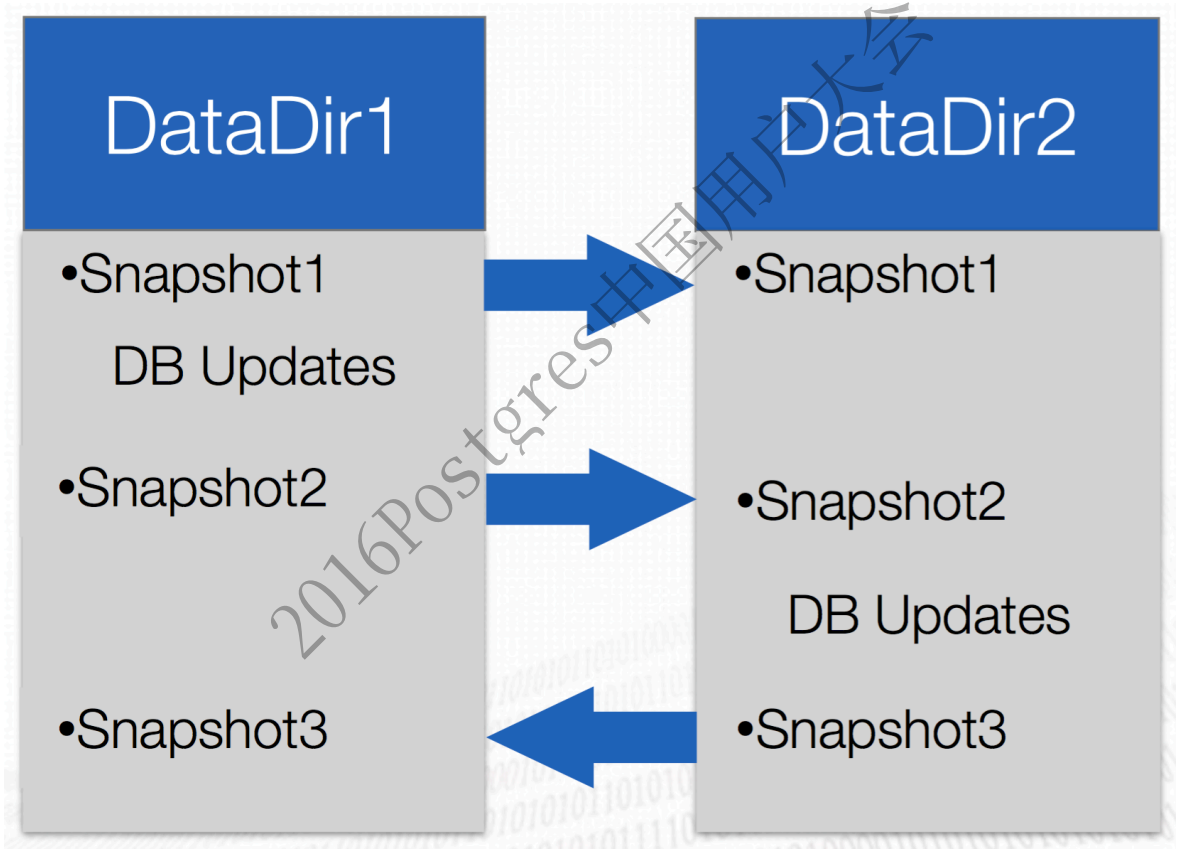
ZFS特性应用:Snapshot/Clone

zfs snapshot <fs|vol>@snap_name

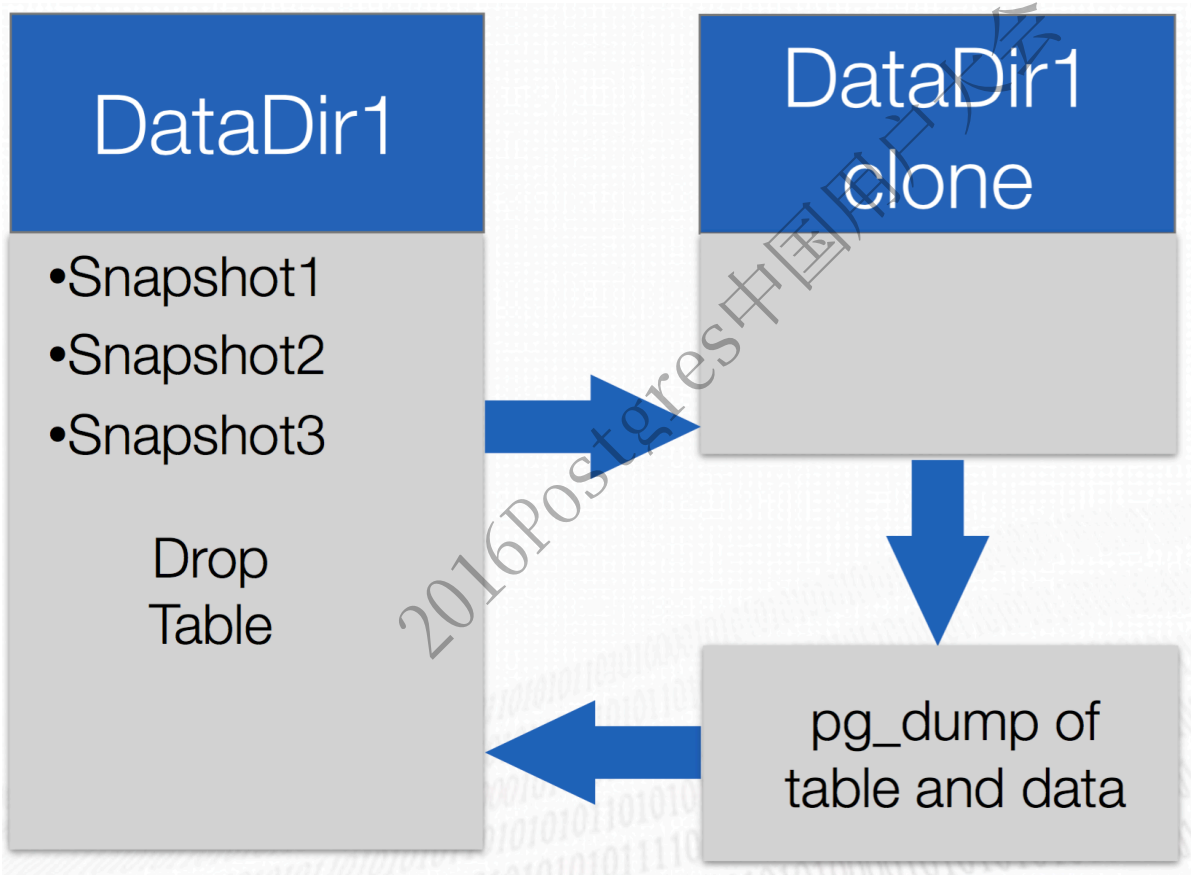
zfs snap_name <fs|vol>



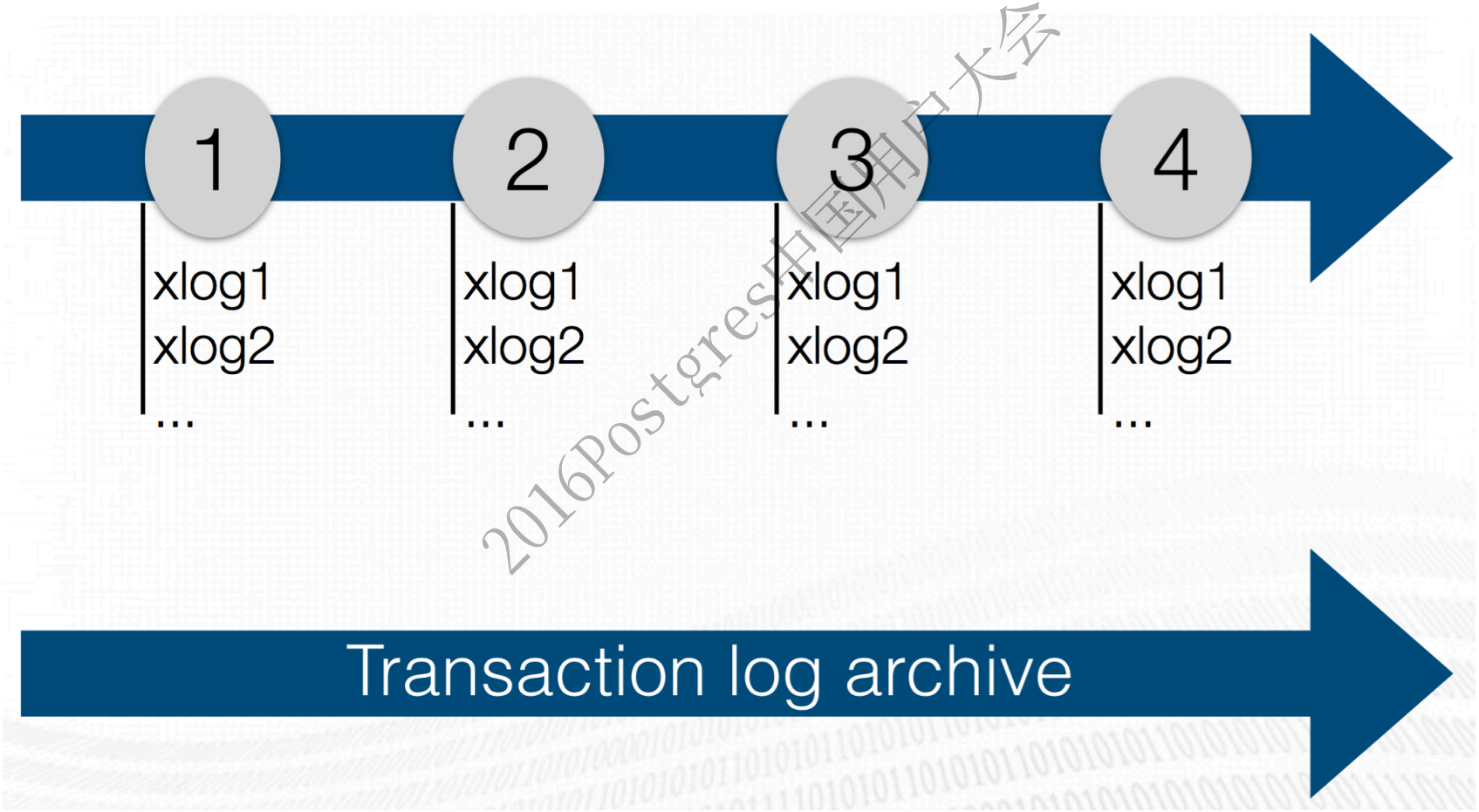
ZFS特性应用: Send/Receive



ZFS特性应用: Clone/Recover



ZFS特性应用: 日志归档



ZFS特性应用: 快照妙用

Tune to your needs

My preferences for transactional databases

- **Every 10-20 minutes, keep for 9 hours**
- **Daily, keep for 10 days**
- **Weekly, keep for 8 weeks**

cronable scripts (examples in the VM)

- **snapshots.sh creates snapshots with hour, day, or week name embeded**
- **snapshot-cleanup.sh removes snapshots older than a specified time**

ZFS特性应用: 压缩

Can improve performance especially for streaming data sets

Turning on only compresses subsequent writes

Check out test results from citusdata

- **http://citusdata.com/blog/64-zfs-compression**

Commands

- **zfs set compression=on fs/name (LZJB)**
- **zfs set compression=gzip fs/name**
- **zfs set compression=gzip-9 fs/name**

ZFS特性应用

For Solaris and its derivatives

- Replicate the OS drive
- Simple bare-metal restore

OS files with metadata in the DB

- Snapshot both for a full consistent test/recovery environment

Using Postgres FDW with non-DB files

- Same story—snapshots, replication, and clones

ZFS数据库常用参数

For Data File: recordsize=8k , primarycache left default

For Temp File and Log File: recordsize and primarycache left default

For Archived Log: enable compression , left recordsize and primarycache left default

logbias=throughput to avoid write twice

If use SSD, secondarycache will be set to improve OLTP read performance

有问必答

WwIT

驱动数据 成就未来