ZFS基础

作者	时间	QQ技术交流群
perrynzhou@gmail.com	2022/09/02	672152841

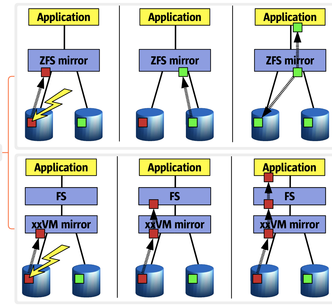
ZFS RaidZ vs Traditional Raid

- 核心区别
  - 1.传统RAID可以在任何文件系统部署，RAIDZ仅仅支持zfs
  - 2.传统RAID是基于硬件实现，而zfs是基于软件层面实现
  - 3.zfs通过checksum套测数据块是否完整性，传统RAID不会标记数据块的完整性。zfs会自动探测数据块完整性，如果有冗余会自动把完整的数据返回给应用
- Raid级别等价
  - 1.zfs的RAIDZ和传统RAID5等价(单个数据校验盘)
  - 2.zfs的RAIDZ2和传统RAID6等价(2个数据校验盘)
  - 3.zfs的RAIDZ3和RAID7等价(3个数据校验盘)
- Layout区别

RAID5			
Disk 1	Disk 2	Disk 3	Disk 4
1	2	3	P
5	6	P	4
9	P	7	8
P	10	11	12

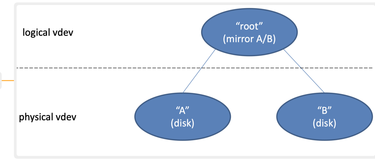
RAIDZ			
Disk 1	Disk 2	Disk 3	Disk 4
P0	1	P1	2
P2	3	4	X
P3	5	6	7
8	P4	9	10

- Write Hole
  - 传统RAID存在Write Hole问题
  - zfs的RAID通过transactions、cow、checksums等方式保证不会存在Write Hole的问题
- Rebuild速度
  - 相同数据量的情况下,传统RAID恢复速度快；zfs的RAIDZ恢复速度慢
- Read Data区别



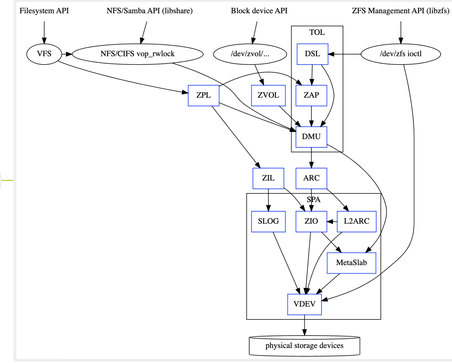
ZFS存储池

- 功能
  - 1.用来管理物理设备，管理方式类似于虚拟内存
  - 2.poll内文件系统共享存储空间
  - 3.poll结构是一棵树，叶子节点是物理磁盘，非叶子节点是逻辑设备(minor/raid-1 这样叶子节点物理磁盘按照存储模式构建的)
- 呈现



ZFS的系统架构

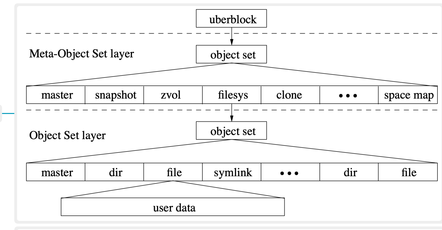
- 架构概览



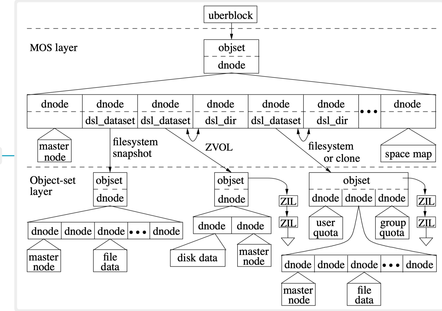
- 模块说明
  - VFS — Linux内核的虚拟文件系统
  - SPA — 从内核中多个设备抽象出来的存储池
  - ZPL — ZFS的Posix层
  - ZVOL — 基于DMU层的提供块设备接口的抽象
  - ZAP — 基于DMU提供的对象抽象构建name/value的键值对存储
  - DMU — 基于块基础上提供对象管理的抽象
  - ZIL — 记录zfs的事务的日志抽象
  - ARC — ZFS基于内存的数据缓存
  - L2ARC — ZFS基于高速设备的二次数据缓存
  - SLOG — ZFS的日志存储模块
  - ZIO — 基于pipeline和事件驱动机制的ZFS IO调度器
  - MeataSlab — ZFS的块分配器
  - VDEV — 基于多个磁盘设备并且为Stripe/Mirror/RaidZ多种存储模式的存储池管理和抽象
  - DSL — ZFS的数据集和快照管理的抽象

ZFS数据组织

- 组织



- 关系



ZFS刷脏流程