实验物理垃圾佬之乐 – PB 级磁盘阵列演进

续本达

清华大学 工程物理系

~~2022-11-26~~ 2023-03-25 Tunight

Outline

磁盘阵列不是备份

背景与动机

1. 把 700TB 数据从日本传到清华。
2. 数据回来了，要放在磁盘上访问，得有 PB 级的磁盘存储。
3. PB 级存储 ~~30–50 万~~

背景与动机

1. 把 700TB 数据从日本传到清华。
2. 数据回来了，要放在磁盘上访问，得有 PB 级的磁盘存储。
3. PB 级存储 ~~30–50 万~~

有经费为什么要当垃圾佬?

- 乐趣：有钱买新电脑为什么要超频?
  - 理解计算依赖的底层设备运行规律，是理解世界的一部分。
- 依附的自由：采购简单灵活
  - 当所有人都在关注如何在主流配置上塞下 10 个 chromium 时，如果你只用一个 chromium 就可以把手里的笔记本用上 10 年
- 懒：花钱是费心费力的事情
- 人生起起伏伏：有时有大钱钱，有时要捡垃圾，有时在有大钱钱的同时也要捡垃圾。

IT 服务陷阱

氪金无法换来省心，氪金无法节省时间，氪金会得到：

- 32 盘的 RAID 5
- 跨阵列（RAID 5）与服务器的 LVM
- 有 $4 \times 16\,\text{Gbps}$ 的 FC 接口，但“我配不通”，用了 $1\,\text{Gbps}$ 的 iSCSI
- 硬盘读不出来了，返厂修理换新的
- 一块硬盘/控制器/电源坏了，硬盘和盘柜一起报废换全新的
- 上云 $\rightarrow$ 忘了充值 $\rightarrow$ 责任不清没人管或没钱了 $\rightarrow$ 数据没了

目标效果

1. 数据本身远远比设备重要。
   - 做实验花一千万，避免数据管理用临时工
   - → 避免 IT 服务陷阱，自己管理。
2. 人类时间远重要于机器时间。
   - 要直通本质，将人工维护限制在不能全自动的部分。
   - → 珍爱生命，用自由软件。
3. 硬盘接口的发展十分缓慢，单盘在 150 MB/s
   - 垃圾佬用 6 Gbps 的接口已经够了。

年份	SATA	速度 Gbps
2003	1.0	1.5
2004	2.0	3
2008	3.0	6

年份	SAS	速度 Gbps	注解
2004	1	3	
2009	2	6	
2013	3	12	必须有风扇
2017	4	22.5	尚未普及

硬 RAID 为什么不行

- 硬 RAID 的固件不透明，导致
  - 换公司就要重新学 → 学不会找人 → IT 服务陷阱
  - 过于智能容易误操作 → 请“专业人士”操作 → IT 服务陷阱
  - 方便的功能都需要另外氪金
- 10 年来，硬盘容量乘 10 ，接口速度不变，硬 RAID 固件基本功能不变 → 盘阵重建时间乘 10。
  - 重建过程中，几乎会再坏一盘，再重建再坏一盘 → 无法成功重建。
  - 重建时无校验，可导致数据损坏。

硬 RAID 总是会坏定理

1. 100 万存储，“专业人员”配好，能用即走人
2. 用户无法理解或无法支付日常维护费用，不做日常维护
3. 盘阵在机房里，放进去不再有人管
4. RAID 6 一个硬盘坏了，大家都没有发现；一个硬盘坏了，大家还都没有发现；又一个硬盘坏了，盘阵无法访问，氪金修理 → IT 服务陷阱

硬 RAID 为什么是销售的最爱?

1. 一个块设备傻瓜式使用，与客户的操作系统解耦，降低客服成本。
2. 控制器复杂，价格高，利润高。
   - 控制器挑硬盘 → OEM 硬盘溢价高，利润高。
3. 固件操作复杂，需要“专业人士”，创造新的客户需求。

硬 RAID 在 10 年前是合理的方案

- CPU 太慢，省着用，专用硬件更好
- 硬盘容量约 1TB，管理起来不费事

块设备级操作与硬 RAID 对应

以 Linux 为例：

- LVM Logical Volume Manager `lvm(8)`
- md Multiple Device (Linux Software RAID) `md(4)`
- dm Device Mapper `dmsetup(8)`

前两者都基于 `dm` 。

文件系统级操作

- 与硬 RAID 模式没有一一对应关系。
- 块设备与文件系统统一考虑，功能更丰富同时操作更简单。

ZFS Zettabyte File System， 2006 年起源于 Solaris，被移植到 BSD、Linux 和 MacOS、NT。

Btrfs “Better FS”（than ext4），2009 年起源于 Linux

文件系统级操作

- 与硬 RAID 模式没有一一对应关系。
- 块设备与文件系统统一考虑，功能更丰富同时操作更简单。

ZFS Zettabyte File System， 2006 年起源于 Solaris，被移植到 BSD、Linux 和 MacOS、NT。

Btrfs “Better FS”（than ext4），2009 年起源于 Linux

其它“多盘文件系统”

暂不作详细讨论

HAMMER 2008 年起源于 DragonFly BSD

ReFS 2012 年起源于 Windows NT（王宇逸吃了螃蟹）

bcachefs 基于 Linux 的 `bcache`，β 态（武益阳吃了螃蟹）

场景分类

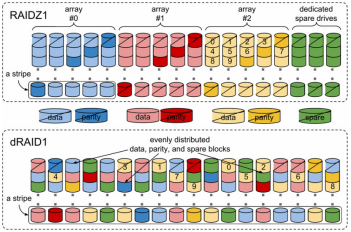
- 人类无法有效预计自己的需求。
  - “你的实验需要多大存储空间?”是伪命题。因为回答是“有多少用多少”。
  - 推论：不论多大的硬盘空间，最终都会被塞满。
  - 推论：要有渐进的增长方案。

种类	要求单盘容量相同	每次加盘数量	来源	尺度 TiB	盘位
Btrfs RAID 10	否	1–2	捡垃圾	30	服务器
ZFS RAID 10	否	4–8	小钱钱	100	服务器
ZFS declustered RAID	是	32–64	大钱钱	500	外接

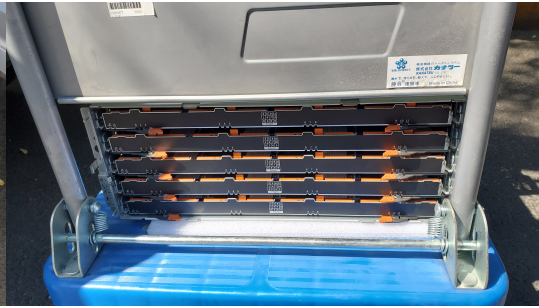
- 人生起起伏伏，有时有大钱钱，有时要捡垃圾，有时在有大钱钱的同时也要捡垃圾。

场景解说

- 不同容量硬盘 Btrfs RAID 1+0
  - Btrfs RAID 1：任何数据都至少在两个设备上
  - Btrfs RAID 1+0：两组到任意组 striping（分条）
    - 同一文件系统，大分条优先
- 不同容量硬盘 ZFS RAID 1+0
  - ZFS vdev mirror，不同 vdev 间平衡写
- 同容量硬盘 ZFS declustered RAID
  - 相比于 RAIDz（类似 RAID5），把 parity（奇偶校验）分到所有盘上



DELL 3060e

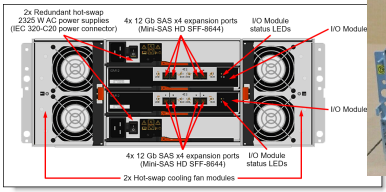


- 自费二手（6500 元，卖方没有公司无法开发票），冲动消费

是不是很眼熟

- IBM DCS3700

- Lenovo Thinksystem DE600S/DE6600H



- NetApp DE6600

- 关键字：4U 60 bay。推论：创新的机械设计非常珍贵
- 差别只是商标和固件

迁移策略

- 把硬 RAID，改造成 JBOD（just a bunch of disks）；
  - 换控制器，SAS 直通 + 交换扩展（SAS expander）
- 使用 `dm` 拼装硬 RAID 配置，恢复虚拟块设备和文件系统；
- 把数据复制到 Btrfs 或 ZFS。

致谢

1. 坏人、蛤蜊挽救“续老师的硬盘”。
2. 武益阳（+刷 host bus adapter）、张爱强、翁俊、刘学伟、王宇逸、徐闯重新规划机架。
3. TUNA 友人贡献的光辉案例、血泪案例和奇葩案例。
4. Chia 提供优质矿渣。