淘宝网
Taobao.com

Flash存储设备
应用实践

核心系统数据库组 褚霸
http://blog.yufeng.info
2011-12-24

淘宝网 Taobao.com

提纲

- 背景
- 历程
- 经验教训
- 提问时间

淘宝网 Taobao.com

背景

- 满足淘宝业务发展的需求。
- 避免对硬件厂商各类黑盒技术的依赖。
- 软硬件结合持续优化，大幅提高性能，节省机器成本。

淘宝网 Taobao.com

目前应用的领域

- 核心数据库
- CDN
- NOSQL数据库
- 搜索
- …

淘宝网 Taobao.com

数据库硬件选型

- 2010年7月开始软硬件选型到方案成熟历时8个月
- 确定大内存PC服务器机型。
- 高强度测试了Intel SSD, SAS+RAID卡, Fusionio, Virident,华为等存储设备，包括性能和数据安全性方面的十几个指标。

淘宝网 Taobao.com

软件和系统优化

- 优化了从Raid卡，块设备，DM层，文件系统，InnoDB引擎到MySQL数据库整个存储体系链的Cache和安全性。
- 引入Flashcache混合存储架构, 稳定性bugfix和添加数据预热等关键特性。
- Linux操作系统当成数据库部件调优，解决大内存下numa,swap,缓存高效使用，资源预留等大量棘手问题。

2010年7月商品库去IOE项目发布

- 混合存储，容量2.xT, 读多写小。
- 性能：
  - 单机QPS: 36000, 其中读32800/写3200。
  - 请求平均延时: 260us。
  - IO util: Flash存储卡<20%， 磁盘<10%。
- 跨入PCIe Flash存储卡时代,为后续同类项目打下坚实基础。

淘宝网 Taobao.com

2010年7月交易读库去O项目发布

- 纯SSD存储，容量1.xT，读为主。
- Intel SSD盘性能挖掘：
  - H700+4片Intel 320=15万IOPS（4K）
- RAID卡下SSD盘寿命测量。
- 标志SSD盘解决方案成熟。

淘宝网 Taobao.com

2010年9月交易主库去IOE项目发布

- 纯PCIe存储，1.28T，写多。
- 成本: 二千多万->三百万。
- 性能:平均请求延时<1ms，2倍余量->10倍余量。
- 扩展: 4台小机->32普通PC服务器。
- 标志PCIe Flash存储卡技术全面成熟。

淘宝网 Taobao.com

目前关注点

- 软件如何适应Flash设备带来的IOPS巨大变化。
- Flash设备在代替内存方面的努力。
- 精耕细作，进一步提高Flash设备效能。
- 下一代PCM存储介质的关注。

淘宝网 Taobao.com

经验教训

- 方案经过1111，1212大促考验，可否证明成熟?
- 单机性能过于强劲，失效对业务的影响。
- 寿命需要真正时间的考验。
- IOPS和Latency不可兼得，如何取舍?
  - 垃圾回收对性能的影响。
  - 抖动如何克服?
- 驱动程序对主机的影响。
- 设备厂家选择
  - 成本，性能，信誉等

淘宝网 Taobao.com

谢谢大家

提问时间～