淘宝网
Taobao.com

MySQL和IO（下）

核心系统数据库组 褚霸
http://blog.yufeng.info
2011-12-23

淘宝网 Taobao.com

提纲

- 硬件
- 操作系统
- InnoDB引擎
- MySQL
- Flash存储设备选择
- 讨论时间

大话硬件

新硬件趋向并行化

软件需要提高并行度

淘宝网 Taobao.com

CPU和大内存

- NUMA架构
  - CPU 直连内存
- 更大的独立L1,L2和共享的L3
  - 缓存践踏
- 大内存管理开销
  - 大页
- 趋势整合IO控制器
- 计算力过剩问题

淘宝网 Taobao.com

Raid卡

- Raid卡喜爱大负载
  - 条带和多盘
- 内置缓存
  - WB 还是 WT
  - 读写缓存比例
  - 预读
  - BBU放电问题
- 逻辑卷
  - 分类不同用途的磁盘

淘宝网 Taobao.com

PCIe Flash卡和SSD

- Flash设备不担心并发IO请求
  - NCQ
  - IOPS多剩问题
- 中断亲缘和平衡问题
  - 对应用的影响
- 供应商和产品特点
  - Fusionio
  - Virident
  - 华为等国内厂商

大话操作系统

操作系统不应该再是黑盒子

追新无罪

淘宝网 Taobao.com

内核

- RHEL 5U4 还是 6U1?
- 超大2M页面
- 页面回写per设备
- 页面回收split lru

文件系统

- ext3 和 xfs
  - 数据根据底层设备智能对对齐，对SSD友好
  - 单个文件可并行dio
  - 更快的文件追加操作
  - 更少的锁冲突
- mount选项
  - nobarrier
  - data=ordered,writeback
- 预读
  - 真的有用吗?

淘宝网 Taobao.com

页面缓存

- vm.swappiness=0
- vm.dirty*
- vm.pagecache
- posix_fadvise清理buffered io引起的垃圾页面
- sync_file_range强制页面回写
- fsync天花板
  - 如何计算,测量

淘宝网 Taobao.com

块设备

- 队列调度算法
  - deadline或者noop
  - cfq害人不浅，顺序变离散
  - 请求队列长度和latency
- Flashcache
  - 担心uncached IO
  - 以2M大小的set为单位进行脏页回写
  - dev.flashcache.skip_seq_thresh_kb
- 软raid
  - 开销不大，整合设备

InnoDB
又一个操作系统
各种微调

淘宝网 Taobao.com

引擎如何工作

- 引擎核心部件
  - 库，表，行，列，事务概念，各种统计实现
  - 核心数据结构Btree
  - 内存管理， buffer和page
  - 逻辑文件空间管理，IO线程池管理读写请求，同步和异步
  - Cache淘汰，回写
- 和操作系统职能冲突
  - 谁更明白用户的需求
- 如何调优
  - BP是核心

大话MySQL

- MySQL架构
  - 网络层太老，HS这样的新架构出现。
  - SQL层和存储层分离，层间数据变换开销大。
  - 锁开销太大。
  - 对NUMA架构不友好。
- 网卡
  - 网络也是特殊的IO。
  - 担心你的千M网卡跑满。
  - 网络丢包重传问题（thin tcp)。

淘宝网 Taobao.com

Flash存储设备选择

- 用途划分
  - 读密集：SSD + Raid
  - 写密集：Fio(+Flashcache)
  - 容量密集：Fio+Flashcache
- 优缺点
  - 成本和性能
- 对未来SSD方案的展望
  - 设备层次无论如何都会存在,Cache不会灭亡。
  - 更快的存储介质如PCM会很快出现。
  - 系统架构和数据结构需要有大的突破，传统数据库系统需要跟进。

淘宝网 Taobao.com

谢谢大家

讨论时间～