大规模 ext4 文件系统部署的收获和挑战

马　涛
淘宝网高级技术专家

议程

- 我们的现状
- 为什么选择 Ext4
- 大规模 ext4 的部署问题和挑战
- 我们与社区的合作
- 对未来 ext4 新特性的展望

我们的现状

- 3.7 亿注册用户
- 6000 万固定用户访问
- 在线商品超过 8 亿件
- 平均每分钟出售商品 4.8 万件
- 截止 2010 年单日交易峰值 19.5 亿元

为什么选择 ext4

- 与现有系统的兼容，运维以及切换成本低
- ext4 被社区以及主流发行厂商支持
- 开发时间较长，相对成熟
- ext4 的一些新特性

ext4 的新特性

- extent
-
- no-journal
-
- Mballoc and delay allocation
-
- large volume and file size support, fallocate etc.

ext4 与 ext2 的对比测试

- 大文件的创建删除
-
- 随机读写
-
- 目录树文件随机访问
-
- 测试用例：
- http://code.taobao.org/p/dirbench/

大文件创建

文件系统	文件大小	创建命令	下层存储介质	消耗时间
Ext2	140GB	dd if=/dev/zero of=/mnt/img \ bs=4096 count=36700160	15KRPM SAS 6Gps	**15m22.850s**
Ext4	140GB	falloc –p /mnt/img –o 0 –l 140g	15KRPM SAS 6Gps	**0m0.136s**
Ext2	140GB	dd if=/dev/zero of=/mnt/img \ bs=4096 count=36700160	Intel X25-M	**35m4.727s**
Ext4	140GB	falloc –p /mnt/img –o 0 –l 140g	Intel X25-M	**0m0.089s**

大文件删除

文件系统	文件大小	删除命令	下层存储介质	时间
Ext2	140GB	rm /mnt/img	15KRPM SAS 6Gps	**1m28.686s**
Ext4	140GB	rm /mnt/img	15KRPM SAS 6Gps	**0m4.313s**
Ext2	140GB	rm /mnt/img	Intel X25-M	**0m8.978s**
Ext4	140GB	rm /mnt/img	Intel X25-M	**0m2.595s**

大文件随机访问

文件系统	文件大小	IO 类型	IO 大小	存储介质	时间
Ext2	140G	direct IO 读	512KB	15KRPM SAS 6Gbps	2m1.529s
Ext4	140G	direct IO 读	512KB	15KRPM SAS 6Gbps	1m43.647s
Ext2	140G	direct IO 写	512KB	15KRPM SAS 6Gbps	2m7.389s
Ext4	140G	direct IO 写	512KB	15KRPM SAS 6Gbps	1m41.701s
Ext2	140G	direct IO 读	512KB	Intel X-25M	0m39.412s
Ext4	140G	direct IO 读	512KB	Intel X-25M	0m0.977s
Ext2	140G	direct IO 写	512KB	Intel X-25M	1m12.680s
Ext4	140G	direct IO 写	512KB	Intel X-25M	1m26.927s

目录树随机访问

文件系统	文件 IO 操作	存储介质	时间
Ext2	创建	15KRPM SAS 6Gbps	13m52.492s
Ext4	创建	15KRPM SAS 6Gbps	**10m29.626s**
Ext2	创建	Intel X25-M	**24m35.620s**
Ext4	创建	Intel X25-M	**15m50.266s**
Ext2	读取	15KRPM SAS 6Gbps	**4m31.463s**
Ext4	读取	15KRPM SAS 6Gbps	**2m53.712s**
Ext2	读取	Intel X25-M	**2m32.163s**
Ext4	读取	Intel X25-M	**1m39.319s**
Ext2	更新	15KRPM SAS 6Gbps	**11m43.514s**
Ext4	更新	15KRPM SAS 6Gbps	**6m38.789s**
Ext2	更新	Intel X25-M	**33m8.519s**
Ext4	更新	Intel X25-M	**15m46.342s**

大规模 ext4 部署

- cdn 系统
- hadoop 集群
- 云计算平台
- 其他一些淘宝业务

我们遇到的问题与挑战

- ext2 向 ext4 的数据迁移
-
- ext4 不同特性的选择
-
- 针对 hadoop 应用的优化
-
- 针对 cdn 系统的优化
-
- 针对 ssd 的优化

ext4 新特性的展望

- snapshot
-
- checksum
-
- bigalloc
-
- inline data

Q & A