第 33 卷 第 8 期 计 算 机 工 程 2007 年 4 月
Vol.33 No.8 Computer Engineering April 2007
·软件技术与数据库· 文章编号：1000—3428(2007)08—0074—03 文献标识码：A 中图分类号：TP311

影响文件系统性能的若干因素的实验

陈 欢[1,2]，范志华[1,2]，熊 劲[1,2]，孙凝晖[1]
(1. 中国科学院计算技术研究所，国家智能计算机研究开发中心，北京 100080；2. 中国科学院研究生院，北京 100039)

摘 要：影响文件系统的因素有很多，该文从不同访问模式、多通道硬件配置、文件系统老化等方面设计试验，利用多种性能评测工具，对 EXT3、XFS 两种本地文件系统进行试验，分析这些因素对本地文件系统性能的影响以及作为 NFS 服务器端的文件系统时，对 NFS 性能的影响。在试验过程中，编写了测试文件系统元数据性能的 Thput benchmark 以及使文件系统能够快速老化的工具 FastAging。通过分析试验结果总结出针对应用中不同的访问模式配置最优的文件系统的解决方案，使用户能够获得最大的 I/O 性能。

关键词：文件系统；性能评价；老化

Experiments on Factors of Affecting File Systems Performance

CHEN Huan[1,2], FAN Zhihua[1,2], XIONG Jin[1,2], SUN Ninghui[1]
(1. National Research Center for Intelligent Computing System, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080;
2. Graduate School, Chinese Academy of Sciences, Beijing 100039)

【Abstract】This paper designs and carries out some experiments on access mode, multiple I/O channel, and system aging problem, etc. It makes use of many kinds of benchmarks to evaluate the performance of the EXT3 and the XFS file system to find how much those factors affect the file system performance. In the experiments, it develops the Thput benchmark which measures metadata performance of the file system and the FastAging tool which can make a file system old very fast. From the conclusion of the experiments, it can choose the most optimal file system according to the application characterization and the hardware configurations to let users get maximum I/O performance.

【Key words】 File system; Performance evaluation; Aging

在 Linux 环境下有很多种文件系统可供用户选择。比如在 Linux 核心 2.4.21 版本中就有 EXT2、EXT3、XFS、ReiserFS、JFS 等 20 多种文件系统。面对如此众多的文件系统，用户应该如何选择？而文件系统的性能通常是用户需要考虑的重要因素之一。事实上，影响这些文件系统性能的因素有很多，在不同的硬件配置及访问模式下文件系统的性能差异非常大。并且目前针对文件系统的性能评测大都在未经使用过的干净的文件系统上进行，其结果与真实的使用环境下的性能有一定差异。

本文对普遍使用的 Linux 环境下的 EXT3、XFS 文件系统进行了试验，找出了影响文件系统的诸多因素，比较这些因素对文件系统性能影响的程度。我们将着重考虑如下 4 个因素：文件系统的种类，硬件的配置，单任务或多任务，老化的程度等。通过分析试验结果，可以对应用的特征进行分析归类，映射到影响文件系统性能的这些因素上，并根据我们的试验结论为文件服务器配置相应的硬件环境及最合适的文件系统，达到给用户提供最佳文件访问性能的目的。

在试验过程中，开发了 Thput benchmark，用来测试文件系统元数据性能，还开发了用于快速老化文件系统性能的工具 FastAging Tool，即将干净的文件系统快速达到长时间使用的效果后，再进行性能评价。

1 文件系统简介

在 Linux 环境下，目前已经诞生了很多种本地文件系统。然而不同的文件系统有着各自的特点，对不同的应用和需求都从设计上做了不同的优化。首先以 EXT3、XFS、ReiserFS 为例，从文件系统设计的角度来看几种文件系统的特点。

Linux2.4 版本内核广泛使用 EXT2 文件系统作为它的本地文件系统。该文件系统在磁盘布局以及空闲块分配方法上吸取了快速文件系统 FFS 的许多重要优点。EXT3[1]文件系统在硬盘上的布局与 EXT2 文件系统是完全一样的，但 EXT3 是日志文件系统。EXT2 存在的潜在问题是：一旦系统突然因为断电或其它原因关机的时候，EXT2 文件系统就不能正常使用，直到它完成检查文件系统一致性的操作。做这种检查(fsck)所耗费的时间依赖于文件系统的大小。EXT3 文件系统在硬盘上多了一个特殊的日志区，用来记录文件系统的日志。它把跟踪到的磁盘内容的变化记录到日志中，使得当文件系统出现故障时，EXT3 文件系统可以不依赖于整个文件系统的大小而根据日志信息得到快速地恢复，其恢复速度要比 EXT2 快很多。

XFS[2]文件系统也是一种日志文件系统，它支持元数据的日志功能。XFS 的优势在于：提供快速恢复磁盘内容；64 位的文件系统，对大文件有很好的支持；对文件元数据的组织采用平衡 B＋树，可以快速搜索、快速分配空间；延迟磁盘空间分配；动态分配磁盘 inode 等。它的不足是由于受到同步磁盘写操作的限制，使它删除文件的性能比较低。

ReiserFS 已经被 Linux 内核所支持，它的优势在于对小文件能够提供较高的性能。ReiserFS 采用 B*树来组织目录、文件等信息。这使得创建、删除文件的性能较高。它还支持

作者简介：陈 欢(1980－)，女，硕士生，主研方向：分布式文件系统；范志华，博士生；熊 劲，副研究员、博士生；孙凝晖，研究员

收稿日期：2006-05-11 **E-mail：**huanchen@ncic.ac.cn

磁盘 inode 的动态分配，使得 inode 的数目可变而不像 EXT3 那样在文件系统创建之初就确定下来。

2 测试工具

2.1 IOzone

IOzone[3]是用于测试文件系统带宽、响应时间等性能指标的 benchmark。它提供的与文件读写相关的测试功能主要有 4 个：write, re-write, read, re-read。

(1)Write 提供向磁盘写一个新文件的性能测试，当写一个新的文件的时候，不仅仅要在磁盘上记录下要存储的数据，还要记录一些用于跟踪定位文件的元数据信息。元数据包括目录信息，空间分配信息，以及其它与文件位置相关的信息等等。由于首次写文件要记录这些元数据，因此其写的性能要比写一个已经存在的文件要低。

(2)Re-write 提供向一个已经存在的文件写的性能测试。由于再次写文件时所要记录的元数据比首次写要少，所以再写的性能要比首次写的性能好一些。

(3)Read 用于测试读一个已经存在的文件的性能。

(4)Re-Read 用于测试重读一个刚刚被读入的文件的性能。重读的性能要比首次读要高，是应为操作系统一般会维护一个用于存放刚刚读入的数据的 cache。当再次读这些数据的时候，就会从 cache 中读取数据，而不用再访问磁盘。

2.2 SPEC sfs

我们在测试中用的是 SPECsfs3.0[4]，它是 Standard Performance Evaluation 公司开发的最新的 benchmark，用于测试 NFS 文件服务器的吞吐率和响应时间。它支持 NFS 协议的第 2 版和第 3 版，支持用 TCP 或 UDP 协议进行通信的网络，它按照 NFS 实际负载的文件操作类型比例来提供与真实负载非常相近的综合的文件操作，从而获得真实环境下的性能数据，如表 1 所示。

表 1 SPECsfs 测试集文件操作比例

LOOKUP	27%	FSSTAT	1%
READ	18%	GETATTR	11%
WRITE	9%	SETATTR	1%
READLINK	7%	READDIRPLUS	9%
READDIR	2%	ACCESS	7%
CREATE	1%	COMMIT	5%
REMOVE	1%		

2.3 Thput Benchmark

Thput Benchmark 是我们自己编写的用来测试文件系统吞吐率的 benchmark，它支持单节点及多节点的测试。在单节点上，它也可以以单进程或多进程的方式进行文件系统吞吐率的测试；在多节点上，可以利用 mpi 将它布置到多个节点进行多进程的吞吐率测试。用户可以自己指定要创建的文件目录树结构、文件大小及文件的访问模式(如顺序或随机等)。Thput 程序根据用户的要求在指定的文件系统上对文件进行创建、删除、读、写等操作，统计出该文件系统的吞吐率。主要用 Thput 分析多进程任务对文件系统元数据性能的影响。所谓元数据性是能为文件系统创建删除文件的性能。通过在测试目录构造一组空文件及目录来测试文件系统的创建和删除的性能。

2.4 FastAging Tool

FastAging Tool 是我们自行开发的使文件系统快速老化的工具。所谓老化[5]，是指由于频繁的创建删除文件导致磁盘上的外部碎片及内部碎片很多，且对于某特定文件配不到连续的块导致属于一个文件的数据块分散的程度很高，目前对文件系统的评测都是在干净的，未经使用的文件系统上作性能评价，因此创建的文件一般都是连续的，可以顺序读取，能够达到文件系统的最大吞吐率。然而实际上，使用的文件系统一般都经过了一段时间的使用。那么在这种环境下是否还能够得到相应的性能？为此做了这个使文件系统快速老化的工具。

测试工具是根据已有的对一个真实文件系统的访问操作进行跟踪得到统计信息，然后对一个空文件系统做相应的访问操作，使得该文件系统快速达到长期使用得效果。所采用的操作类型比例是依照 SPECsfs 的综合的文件操作，并加以简化调整，对我们的文件系统加以快速老化，在老化过程中，监控文件系统的磁盘使用率，每增长一定比例我们就做一次性能测试，从而得到文件系统在老化过程中的性能变化趋势。

3 影响文件系统性能因素的分析

影响文件系统的因素很多，我们从多 I/O 通道、多任务、老化等方面设计试验，深入分析影响文件系统性能的因素。

3.1 多任务对文件系统性能的影响

当有一个用户对文件访问时，文件系统一般可以做到顺序读写；然而当有多个用户同对多个文件进行读写操作时，由于不同文件分布在磁盘的不同位置，而且不同用户之间根据调度频繁切换，使得对磁盘的访问不再是顺序的。为此通过扩展用户数来说明多任务对不同文件系统的影响。

首先利用 IOzone 测试多进程对文件系统的读写性能的影响。所选的参数是分别启用 1、2、4、6、8、10 个进程，所用进程的 I/O 总量，即各个进程的 I/O 量之和为 10GB，文件大小是由 I/O 总量平均到每个进程决定的。图 1 是多任务对 EXT3、XFS 两种文件系统的性能影响曲线图。结果表明，读写性能 XFS 文件系统均好于 EXT3 文件系统；随着读写进程数的增加，XFS 文件系统的读写性能变化比较平缓，受到的影响较小，而对 EXT3 文件系统的读写性能影响较大。原因在于两种文件系统分配数据块的策略不同。EXT3 采用查询 bitmap 的方法以 block 为单位分配数据块，而 XFS 文件系统以 extent 结构，采用 B+树的方法查询，分配数据块空间。

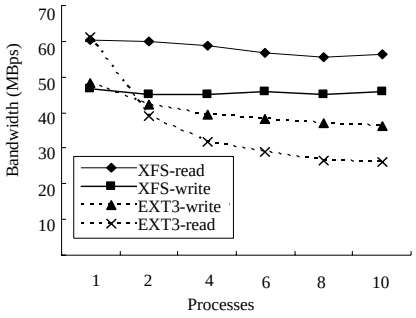

图 1 一块盘读写带宽

用 Thput 分析多进程任务对文件系统元数据性能的影响。通过在测试目录构造一组空文件及目录来测试文件系统的创建、删除文件的性能；通过创建、删除 4KB 大小的文件来反映不同文件系统对小文件的支持性能。表 2 的结果表明，EXT3 文件系统的对小文件的读写性能远远好于 XFS 文件系统。因为 XFS 采用磁盘同步写，导致小文件的读写性能较差。

表 2 一块盘小文件吞吐率

	Thr	Files	0K create	0K delete	4K create	4K delete
EXT3	1	5 000	1 283	44 132	350	23 926
EXT3	6	30 000	1 966	1 966	819	8 015
XFS	1	5 000	1 480	1 535	59	1 404
XFS	6	30 000	1 397	1 242	314	1 236

3.2 I/O 通道对文件系性能的影响

为测试多通道对文件系统的影响，设计了 5 种测试模式：单盘，一条 SCSI 总线两个磁盘，2 条 SCSI 总线 2 个磁盘，一条 SCSI 总线 3 个磁盘，2 条 SCSI 总线 4 个磁盘。当挂多个磁盘时，利用 LVM 将多块盘虚拟成一块磁盘，在其上创建 EXT3 和 XFS，分别测试它们带宽和吞吐率。在带宽测试中，我们用的测试工具是 IOzone，吞吐率用的是 thput 测试程序。

图 2 说明的是一根 SCSI 线挂 2 块磁盘及 2 根 SCSI 线各挂 1 块磁盘在一个进程进行读写的性能。这一组测试的结果表明：在扩展性上，磁盘的扩展对性能的提高有较大的影响，其中 XFS 的存储可扩展性较好，特别是对进程数较少时的读写性能提高很有利；磁盘通道的扩展对性能的影响不大。对于应用来讲，要是对本地文件系统的应用是读写密集型，那么推荐用 XFS 文件系统可以获得更好的读写带宽性能。

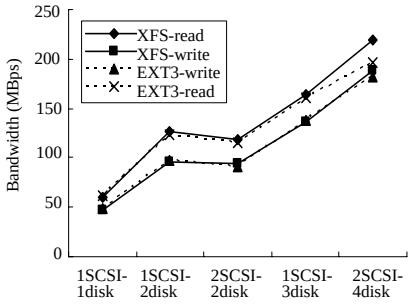

图 2 磁盘及通道的扩展对读写性能的影响

3.3 老化对文件系统性能的影响

为了获得在已经老化的文件系统的性能，使用了自己开发的一个快速老化文件系统的工具 FastAging。它通过随机、快速地在磁盘上创建删除不同大小的文件来达到老化的目的。在设计的试验中，在磁盘空间占用率每增长 5%就进行一次带宽和吞吐率的测试。

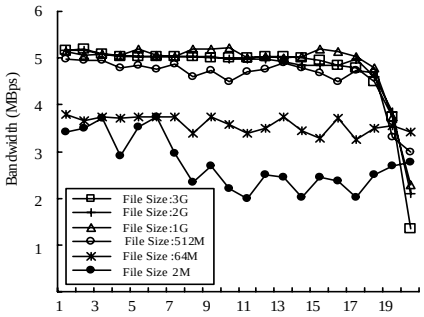

图 3 EXT3 文件系统老化对性能的影响

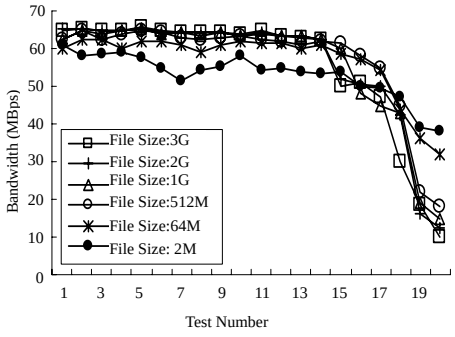

图 4 XFS 文件系统老化对性能的影响

图 3、图 4 分别表示的是 EXT3、XFS 文件系统老化过程中的写性能变化情况。其中横坐标表示每增长 5%的磁盘空间占用率就测一次性能。从图上看出，随着磁盘的老化，当磁盘尚未满时，读写的带宽性能都比较平稳；当磁盘利用率达到 90%，2 种文件系统的性能下降的均很快。原因在于分配块时，找空闲块会耗费大量的时间，且很难找到连续的块分配，不能连续写导致性能急速下降。

3.4 本地文件系统对 NFS 的支持

在分布式环境下，NFS 是使用最广泛的文件共享协议。在试验中还测试了 5 种模式下 EXT3、XFS 两种文件系统对 NFS 的支持的性能。用 SPECsfs 测 NFS 在采用不同本地文件系统时的性能，如前所述，SPECsfs 是测试当客户端生成对服务器文件系统各种类型的文件操作时服务器端的性能。这种测试更接近于真实系统，可以通过这测试来找到提高 NFS 性能的途径。测试结果表明，对于 mix 操作的应用模式来说，EXT3 的吞吐率性能要好于 XFS。原因在于 EXT3 对元数据操作的性能及小文件的存取性能很高，而 XFS 更适合于大文件的连续访问的模式。而在 SPECsfs 的测试中，对大文件读写所占的比例较小，如图 5 所示。

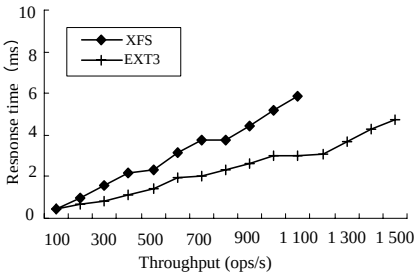

图 5 XFS、EXT3 的 NFS 性能

4 结论

通过我们设计的试验可以总结出不同文件系统各自的特点以及影响这些文件系统性能的主要因素。比如 XFS 文件系统对大文件的支持及多用户的支持更好，而 EXT3 对于小文件的支持更好。利用这些结论，可以根据用户应用的特点及硬件配置，定制最优的文件系统获得最佳的性能。比如，对于邮件服务器这样大量频繁操作小文件的数据的应用，就推荐使用 EXT3、ReiserFS 这样的文件系统；对于需要读写大文件的科学计算，可以选择使用 XFS 文件系统等。目前，实验结论已经运用到正在开发的项目上，即让 NFS 服务器能够同时利用多种本地文件系统，并为如何选择 NFS 服务器的本地文件系统和如何构建 NFS 服务器的硬件配置以获取最大的 I/O 带宽提供了重要依据。

参考文献

1 Cao Mingming. State of the Art, Where We Are with the EXT3 File System[C]//Proc. of Linux Symposium, Ottawa, Canada. 2005.

2 Sweeney A. Scalability in the XFS File System[C]//Proceedings of the Winter 1996 USENIX Conference, San Diego, CA. 1996-01: 33-44.

3 Iozone Filesystem Benchmark[Z]. 1998. http://www.iozone.org.

4 SPEC SFS97_R1 V3.0 Documentation[Z]. 2006. http://www.spec.org/sfs97r1/index.html.

5 Smith K. File System Aging——Increasing the Relevance of File System Benchmarks[C]//Proceedings of the 1997 SIGMETRICS Conference , Seattle, WA. 1997-06: 203-213.