第33卷 第19期 计 算 机 工 程 2007年10月
Vol.33 No.19 Computer Engineering October 2007
·网络与通信· 文章编号：1000—3428(2007)19—0125—03 文献标识码：A 中图分类号：TP393

集群环境中影响NFS文件系统带宽的测试与分析

曹立强，罗红兵，张晓霞
（北京应用物理与计算数学研究所，北京 100088）

摘 要：NFS是集群系统中提供全局文件共享的主要手段，研究影响NFS带宽的因素对于优化集群系统的性能十分重要。该文针对集群系统中I/O特征建立了NFS的带宽模型，设计和实现了基于MPI开发的并行文件系统测试工具Mpbonnie，在集群系统中测试和分析了多种因素对NFS性能的影响。结果表明，除已知的存储和网络因素外，与NFS带宽关系密切的因素还包括客户端数量、服务器文件系统类型、读写方式和服务器处理能力等。

关键词：NFS；带宽；性能分析

Measurement and Analysis of Factors Affecting NFS Bandwidth in Cluster

CAO Li-qiang, LUO Hong-bing, ZHANG Xiao-xia
(Institute of Applied Physics and Computational Mathematics, Beijing 100088)

【Abstract】NFS is a predominant distributed file system in the cluster. Besides network bandwidth and storage bandwidth, there are some other factors affecting NFS bandwidth as well. With a new MPI-based parallel I/O benchmark Mpbonnie, this paper measures and analyses NFS performance with different configurations in the cluster. It proves server side file system, I/O mode and server side processor have great impact on NFS performance.

【Key words】NFS; bandwidth; performance analysis

集群是当前高性能计算机的主流体系结构，使用工业化标准技术和部件是构建集群的主要形式。作为实现文件单一映像和全局I/O的主要手段，NFS被广泛地应用在集群系统中。例如，以联想深腾6800和曙光4000A为代表的国内高性能计算机系统均部署有NFS文件系统。高性能计算机服务于大规模并行计算，大规模并行计算对I/O性能有巨大的需求，因此，研究集群系统中影响NFS文件系统性能的因素十分重要。

已有研究表明，基于高带宽的网络与快速I/O设备建立的NFS文件系统具有较好性能，目前的工作大多局限于这两个方面[1~2]，但是上述因素并不是决定NFS性能的充分条件。经过分析发现，服务器文件系统类型、读写方式、服务器处理能力、缓存大小和同时读写的客户端数量也是影响NFS文件系统性能的重要因素。利用测试程序评价文件系统性能是一种重要的研究方法。已有的NFS测试程序面向单个节点或者事务处理的办公环境，为在集群环境下面向科学计算程序测试NFS文件系统性能造成了障碍。在分析集群环境NFS测试需求和已有测试程序不足的基础上，本文介绍了新研发的基于MPI通信与同步的读写密集型I/O测试程序Mpbonnie。在一个典型的集群环境中，分析测试了影响NFS性能的各个因素，并根据测试结果评价它们对NFS性能影响的大小。

1 NFS带宽模型

NFS接收用户发出的文件系统调用，通过RPC将请求转发到服务器。在服务器端，它调用服务器上的本地文件系统，完成调用请求。本文建立NFS的带宽模型如图1所示。NFS请求如果在客户端缓存中命中，则它的最高带宽为客户端系统的内存总线带宽b_1，如果未命中，则通过网络发送到服务器。如果请求服务能够在服务器缓存中完成，由于内存总线带宽远高于网络带宽，此时最高带宽为$\min(b_2,b_3)=b_2$；如果需要读取存储资源，则最高带宽为$\min(b_2,b_3,b_4)=\min(b_2,b_4)$。设$B_{NFS}$为NFS带宽，则

$$B_{NFS}=\begin{cases} b_1, \text{客户端缓存命中} \\ b_2, \text{服务器缓存命中} \\ \min(b_2,b_4), \text{客户端服务器缓存未命中} \end{cases}$$

网络带宽b_2和存储资源b_4是NFS带宽的决定性因素，提高这些资源的性能能够提高NFS的带宽。而客户端缓存和服务器缓存，尤其是前者，能够提高NFS文件系统的性能。

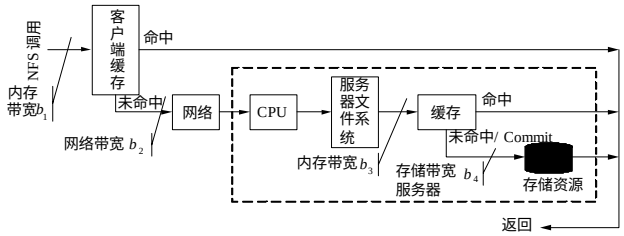

图1 NFS带宽模型

在实际的应用过程中，由于网络因素和存储因素相对固

基金项目：中国工程物理研究院科学技术基金资助项目(20060647)
作者简介：曹立强(1976－)，男，博士，主研方向：并行文件系统及性能优化；罗红兵，副研究员；张晓霞，高级工程师
收稿日期：2006-10-25 **E-mail：**clq@iapcm.ac.cn

定，因此影响 NFS 性能的因素主要有服务器处理能力、读写方式、缓存以及客户端数量等。

服务器处理能力不仅决定了服务器能够服务请求的数量，而且更多更快的处理器能够处理 NFS 请求。不同的服务器文件系统组织数据存储的方式不同，能够提供的读写带宽也不同。一些文件系统采取了延后写方式，将原本大量的小数据块读写合并为较少量大数据块的读写，期望提高读写带宽。

在集群系统环境相对固定的前提下，实际可以获得的网络带宽与参与读写的客户端数量相关。多个客户端同时读写数据导致网络中的数据包数量增加。在网络能够允许的范围内，更多的数据包能够充分地利用网络带宽，提高单位实际内传输数据量，有利于提高文件系统聚集读写性能。然而当数据包进一步增加，超过网络能够承受的范围时，会发生过多的网络冲突现象，使得网络带宽波动或者下降，进而影响到 NFS 带宽，因此，参与读写的客户端数量也是一个影响 NFS 性能的因素。

2 测试方法

常用的文件系统测试工具有开放源码的Bonnie、Iozone[3]和SPEC(standard performance evaluation corporation)公司维护的Specfs[4]等。Bonnie是一个用标准C库函数实现的文件系统存取速度测试工具。它测量文件系统在顺序读、顺序写和随机定位 3 种模式下的传输速率和CPU占用率，而其中顺序读写又分为字符读写和块读写两种方式。Iozone的测试手段与Bonnie类似，但是它能够每次读写步长，并测试读写不同大小文件的性能。Specfs是一种专用于测量NFS吞吐率和响应时间的工具，其中集成了NFS客户端，并通过RPC与服务器通信。Specfs提取了典型办公环境下的文件大小分布和各种文件操作的比例，抽象成为测试用负载。它测量NFS服务器在不同负载状态下的响应时间，并用这个值反映系统的性能。分析发现，这些工具并不完全适用于集群环境中NFS性能的测试。Bonnie和Iozone是运行于单个节点上的文件系统测试工具，使用它只能够获得一个NFS客户端的性能。Specfs的工作负载模拟了典型的办公环境下工作负载，其读写请求数量占总请求数量的 27%，然而通过在深腾 6800 上面搜集文件读写特征发现，95%以上的操作是读写操作。特征差异很大，参考Specfs测试结果不能如实地反映NFS在集群系统中的性能。

基于 MPI 消息通信环境，开发了一个集群环境下并行读写带宽测试程序 Mpbonnie，其主体结构如图 2 所示。

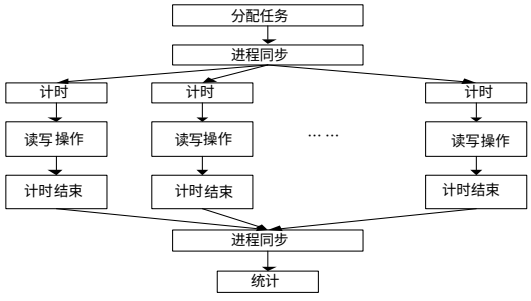

图 2 Mpbonnie 的结构

Mpbonnie 的设计目标在于获得多进程并同时读写全局文件系统的性能。程序的 0 号进程在程序初始化之后在各个进程之间均匀分配负载，然后调用 MPI_Barrier()同步；接下来各个进程依据 0 号进程分配的 I/O 负载操作并计时，待所有进程完成 I/O 操作后 0 号进程统计并行 I/O 的读、写带宽并输出结果。Mpbonnie 的输入参数中规定了读写文件的大小，各个进程的读写操作是它的主要流程，因此它是一个读写密集型并行 I/O 测试程序。由于利用了 MPI 通信库可扩展性好、同步效率高的优点，因此它能够精确地反映集群环境下应用程序所能够获得的并行 I/O 带宽。

在使用 Mpbonnie 的过程中，发现 NFS 读性能受缓存影响明显，而写操作受缓存影响小。其原因在于为了保证写入数据的安全，NFS 要求在关闭文件操作返回之前将所有修改后的数据写入存储资源。因此，将文件关闭操作置于计时结束之前，使得计时结束时所有数据已经写入存储资源，确保测试结果真实地反映文件系统性能。

3 NFS 性能分析

本文重点调试了 NFS 文件系统的写性能。这是由于受到文件系统缓存的影响，NFS 的读性能好，而写性能差，应用程序写数据的性能瓶颈制约更明显。其次，由于 NFS 文件系统持续写流程与持续读流程一致，因此对 NFS 写性能的调试手段也能够应用于 NFS 读性能调试中。

在本文的集群测试环境中，服务器通过千兆以太网接入网络，而客户端通过百兆以太网接入。测试调节服务器本地文件系统的类型、使用方式、服务器 CPU 的数量、改变缓存大小等条件后单个 NFS v3 客户端写数据所能够获得的性能，并与之前的性能比较；然后利用 bonnie 或者 Mpbonnie 测试程序测试了多客户端数量情况下的 NFS v3 性能，并比较不同客户端数量能够取得的聚集带宽。

在本文的试验环境和试验条件下可以得出如下的规律：服务器本地文件系统类型、读写方式和服务器的处理能力是影响 NFS 文件系统性能的关键因素；而 18 个以上客户端进程同时写 NFS 服务造成了性能下降并明显波动。

3.1 服务器文件系统对性能的影响

NFS 服务建立于服务器本地文件系统之上，不同的本地文件系统影响 NFS 的性能。利用单机的 Bonnie 测试程序本文分别测试了本地文件系统性能和以此为基础的 NFS 性能，每次读写文件大小 1GB，取它们的写性能。

在本地文件系统测试中，VxFS 文件系统比 UFS 文件系统写性能好，但是以 VxFS 为基础的 NFS 性能比以 UFS 为基础的 NFS 写性能差。两者之间的差距可以高达 30%，结果如表 1 所示。

表 1 NFS 在 VxFS 和 UFS 上的写性能比较

VxFS	UFS	NFS+VxFS	NFS+UFS
145 059	39 210	7 792	11 027

NFS 文件系统支持两种写模式：同步写模式和异步写模式。同步方式下的服务器依据数据包的接收顺序将数据写入存储资源中，而异步方式下的服务器可以调度接收到的数据包，合并小的操作成为大块数据操作，因此，可以期望异步写具有较好的性能。在安装 NFS 时增加 nosync 开关，可以使服务器缺省的同步写方式改为异步写方式，其他环境不变。利用 Mpbonnie 测试程序对上述两种写方式进行对比，每次每客户端写数据大小为 100MB，结果如图 3 所示。

可以看到大多数情况下异步写方式性能好于同步写方式，这种性能改善的代价在于略微增加了服务器崩溃时数据丢失的数量。

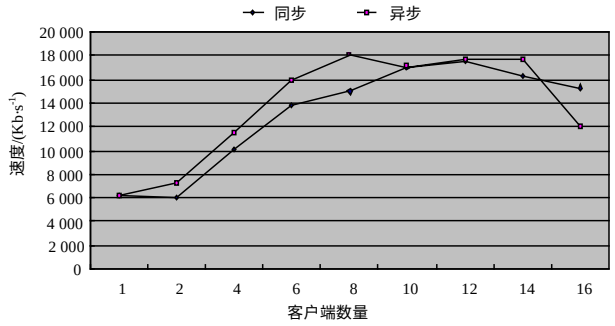

图 3　同步和异步写性能比较

3.2　服务器性能对文件系统性能的影响

本文在其他环境与上面测试相同的情况下利用 Mpbonnie 测试程序比较了双 CPU 系统与单 CPU 系统在客户端数量增长情况下的性能曲线，结果如图 4 所示。在客户端数量少的情况下，单 CPU 系统的性能略好，但是客户端数量持续增长之后的双 CPU 系统性能好于单 CPU 系统 NFS 写性能。

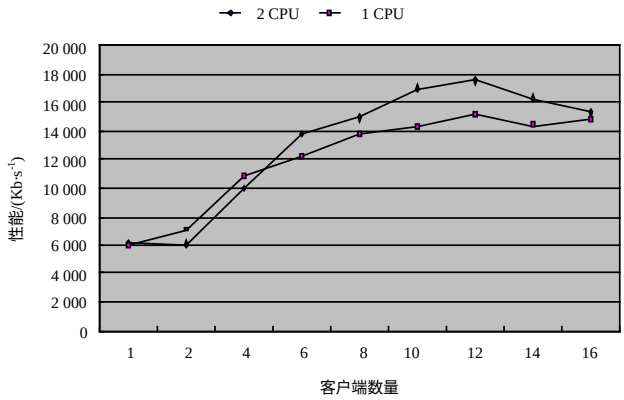

图 4　双 CPU 与单 CPU 写性能比较

双 CPU 系统的最高值可以达到 17 531 Kb/s，单 CPU 系统最高值为 15 170Kb/s。虽然上述两种测试过程中服务器的 CPU 资源都没有使用到饱满的程度(CPU%>90)，但是双 CPU 系统性能好于单 CPU 系统性能，可以期望，当 CPU 数量更多或者更快时，NFS 会有更好的性能。

3.3　NFS 缓存的大小对性的影响

而NFS v3 采取了初始化过程中协商机制确定缓存大小，不同系统间的缓存大小可以不同，并可以手工调节[5]。通过安装NFS文件系统时指定rsize与wsize选项客户端在 16KB与256KB之间调节NFS的读写缓存大小。使用Bonnie测试 10 次，取平均值作为结果，如表 2 所示。

表 2　缓存大小对 NFS 写性能的影响

缺省值	16KB	32 KB	64 KB	128 KB	256 KB
16 920	5 004	18 497	15 625	15 787	17 971
4 577	3 844	5 834	5 952	6 346	/

可以看出，使用 NFS v3 网络文件系统的缺省缓存大小能够取得较好的性能。手工调节其缓存大小，低于 32KB 的缓存使得 NFS 性能差；达到 32KB 时 NFS 系统性能较好，从 32KB 进一步增大时，性能变化并不显著。

3.4　多客户端情况下的性能下降与波动

使用 Mpbonnie 测试程序，每次每客户端写 10MB 大小的数据，本文测试了客户端线性增长到 22 个情况下 NFS 聚合性能的曲线。为了获得比较准确的数值，在每种条件下连续测量 10 次，取它们的平均值作为测试结果。为了表示多次测量结果之间的波动情况，取每次测试所获得的样本的平均方差作为参考值，结果如图 5 所示。

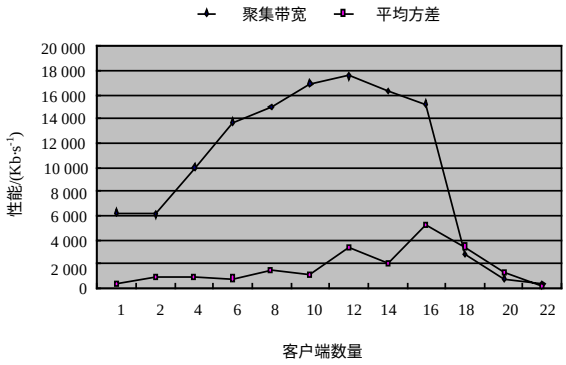

图 5　NFS 聚集写性能与平均方差随客户端变化曲线

可以发现，在客户端数量持续增长的过程中，NFS 性能曲线表现出首先持续增高，当客户端数量增长到 18 个后，系统性能有大幅度的下降，而 10 次测量的平均方差有随客户端数量增加而增长的趋势，在 18 到 20 个客户端的条件下，多次测量样本的标准方差甚至超过了带宽的平均值。这说明测量样本之间的差异较大，波动明显。

图 5 中只标明了 22 个客户端以内的性能曲线，当客户端数量持续增大时，NFS 的聚集带宽一直保持在较低的水平，例如，测试表明，32 个客户端同时写文件的聚集带宽最高为 3 779 Kb/s，尚不如单一读写进程所获得性能好。

4　小结

综上所述，在建立 NFS 的带宽模型，分析影响 NFS 性能因素的基础上，本文利用 Mpbonnie 及其他测试手段在集群系统中分析评价了它们对 NFS 带宽的影响。在网络与存储等硬件系统确定的集群环境中，得到如下结论：

(1)服务器文件系统类型、服务器处理能力和文件系统的读写方式是影响 NFS 文件系统性能的关键因素。

(2)保持同时读写的客户端数量在一定范围内是获得稳定高速并行 I/O 性能的关键。

(3)缺省值基础上进一步增大 NFS 缓存对优化性能意义不大。

参考文献

1 Martin R. Culler D. NFS Sensitivity to High Performance Networks[C]//Proc. of SIGMETRICS '99. 1999-05.

2 Lever C, Honeyman P. Linux NFS Client Write Performance[C] //Proceedings of the Usenix Technical Conference on FREENIX Track, Monterey. 2001-06.

3 Norcott W D, CappsD. IOZone Filesystem Benchmark[Z]. (2006-07). http:// www.iozone.org.

4 Standard Performance Evaluation Corporation (SPEC)[Z]. SFS 3.0. (2001-07). http://www.spec.org.

5 Sun Microsystems Inc. NFS Server Performance and Tunning Guide for Sun Hardware[Z]. 1998.