基于高速通信协议的 COSMOS 机群文件系统性能研究

贺 劲 徐志伟 孟 丹 马 捷 冯 军
(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)
(jhe@gatekeeper.ncic.ac.cn)

摘 要 作为曙光 3000 超级服务器的重要组成部分, COSMOS 机群文件系统对机群文件系统协议、结构及性能优化等问题进行全面深入的探讨. 首先描述了基于曙光 3000 机群高速通信协议 BCL-3 的 COSMOS 文件系统的实现, 然后引入并发带宽利用率, 描述了通信与 I/O 对机群文件系统性能影响程度; 最后介绍了有关性能实验并对实验结果作出解释.

关键词 机群文件系统, 高速通信协议, 异步 I/O

中图法分类号 TP302

PERFORMANCE ANALYSIS OF THE COSMOS CLUSTER FILE SYSTEM BASED ON HIGH-SPEED COMMUNICATION PROTOCOL

HE Jin, XU Zhi-Wei, MENG Dan, MA Jie, and FENG Jun
(*National Research Center of Intelligent Computing Systems, Institute of Computing Technology, Chinese Academy of Sciences, Beijing* 100080)

Abstract As an important component of Dawning 3000 super-server, the COSMOS cluster file system allows to explore many important fields in the cluster file system, including architecture, protocol and performance optimization. Described in this paper is the implementation of the COSMOS cluster file system based on BCL-3 high-speed communication protocol in Dawning 3000 cluster. Then parallel bandwidth utilization is introduced to describe the influence of communication and I/O subsystem on the performance of the cluster file system. Finally, the related experiments are reported and explained.

Key words cluster file system, high performance communication protocol, asynchronous I/O

1 引 言

随着 CPU 与互联网络技术的高速发展, 以商品化部件构造的机群系统已经成为建造高性能计算机与超级服务器的主流技术. 目前在国内外正在建造和已经完成的机群中, IBM 公司的 SP 机群[1]、加州大学伯克利分校的 NOW 系统[2]以及国家智能计算机研究开发中心的曙光系列超级服务器[3]是其中的典型代表. 根据 Amdahl 法则, 每 1 MIps 的计算性能需要 1 MBps 的 I/O 带宽与之匹配. 因此在机群系统中, 如何为用户提供高性能可扩展的 I/O 系统成为近年来研究的热点, 分布/机群文件系统正是其中主要技术之一.

原稿收到日期: 2001-04-03; 修改稿收到日期: 2001-08-28
本课题得到国家"八六三"高技术研究发展计划基金(863-306-ZD01-01)和国家杰出青年科学基金(69925205)资助

早期分布/机群文件系统中的典型代表有加州大学伯克利分校的 SpriteFS[4]与卡内基-梅隆大学的 AFS[5], 它们都基于客户-服务器模型, 文件协议也相对简单. 后续的分布/机群文件系统一般都非常复杂, 如加州大学伯克利分校的 xFS[6,7]机群文件系统, 它采用了无集中式服务器结构, 试图通过文件系统节点间复杂的协作缓存算法来提高机群文件系统性能; 另外 IBM 公司的 Calypso[8]与 GPFS[9]文件系统也使用了复杂的缓存与令牌管理算法来优化系统性能. 但目前分布文件系统的发展趋势[7,10,11]是根据实际应用需求, 简化文件系统协议并以提高读写性能及文件系统可用性为主要设计目标.

有研究[6,12]认为, 分布/机群文件系统的性能主要受其协议层开销、I/O 子系统与通信模块性能的影响. 本文作者基于曙光 3000 超级服务器, 使用高速通信协议 BCL-3[3], 设计与实现了机群文件系统原型 BCL-COSMOS, 对上述问题进行了深入研究, 得出了在系统不再受限于 I/O 带宽时, 机群文件系统底层通信服务的改善可以大幅度提高整体性能的结论.

本文后续内容的组织如下: 第 2 节描述基于 BCL-3 高速通信协议的 COSMOS 文件系统体系结构; 第 3 节给出 BCL-COSMOS 系统的性能模型; 第 4 节介绍实验平台及测试方法, 并对实验数据进行分析, 论述了影响文件系统性能的关键问题; 第 5 节总结全文, 阐述今后的工作.

2 系统设计与实现

曙光 3000 超级服务器是国家“八六三”高技术研究发展计划的重大研究成果, 它是一个基于 SMP 节点的机群系统. 系统由 70 个节点组成, 峰值运算速度为 403.2GFLOPS.

COSMOS 是曙光机群上使用的基于 AIX 文件系统 JFS 实现的, 具备单一系统映象、可扩展的分布式文件系统, COSMOS 提供 NFS 语义, 做到与 Unix 文件系统完全的二进制兼容, 并采用并行I/O 等措施来提高系统的整体 I/O 性能.

2.1 COSMOS 系统

组成 COSMOS 文件系统的两大部分是机群文件系统协议与底层通信模块. 文件系统协议处理用户应用对 COSMOS 文件系统的请求并通过通信模块与其它系统成员进行数据与元数据交换; 底层通信模块则负责转发与接收来自其它系统成员的数据.

2.1.1 系统概况

曙光 3000 超级服务器上的 COSMOS 机群文件系统具有理想的带宽和吞吐率性能, 以及良好的可扩展能力.

COSMOS 由 3 类用户空间进程及 VFS 核心扩展构成, 这 3 类用户空间进程分别是 COSMOS 客户进程、系统元数据服务器以及存储服务器.

COSMOS VFS 核心扩展利用标准的 Unix VFS 接口来支持普通应用在 COSMOS 文件系统上的二进制兼容[13]. COSMOS 客户进程与它的 VFS 核心扩展一起协同工作等, VFS 层的请求将被转发到 COSMOS 客户进程, 由它在用户空间与系统其它组成部分通信. 这样的体系结构有利于文件系统的可移植性[14]. 图 1 是 COSMOS 文件系统的一个示意图.

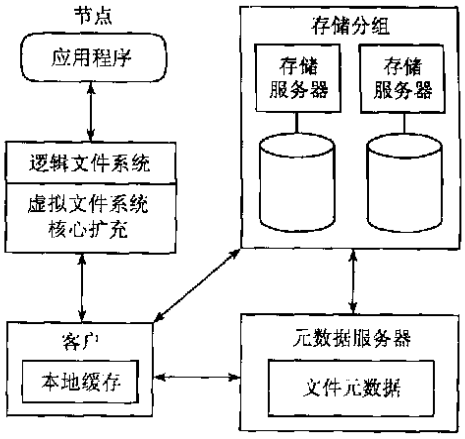

图 1 COSMOS 体系结构示意图

为了避免单服务器系统的可扩展瓶颈, COSMOS 文件系统中使用了多个元数据服务器和存储服务器, 这种结构可以增强系统的可用性与性能.

COSMOS 存储服务器支持网络磁盘分组, 在这种设置下, 一个 COSMOS 文件被分割成多个块, 分别存放在多个存储服务器中. 在文件读写发生时, 多个存储服务器并发执行 I/O 操作, 从而使最终用户获得良好的聚集带宽. 另外, 存储服务器通过多线程及异步 I/O 操作磁盘数据. COSMOS 存储服务器的配置非常灵活, 既可以配置为支持网络磁盘分组, 又可以配置为非网络磁盘分组模式. 前者试图通过多服务器聚集提高性能, 后者则试图通过减少每个服务器需要服务的客户个数来提高效率.

元数据服务器所管理的元数据有以下几类: *i* 节点, 目录和超级块. *i* 节点信息包含了磁盘 *i* 节点数据和动态系统中各结点上的 *i* 节点缓冲信息. 目录文件记录了各目录在各结点上的缓冲情况, 同时

也管理了一块目录缓冲区. 超级块除了静态的文件系统信息外, 还包括动态的系统空间信息, 系统维护不严格的空闲空间信息, 由主元数据服务器根据客户进程的请求向各个从元数据服务器和存储服务器查询获得.

在本文的后续部分中, 我们将把 COSMOS 存储服务器所在的机群节点简称为存储服务器, COSMOS 客户端接口所在机群节点为客户节点.

2.1.2　读写流程

当某个用户进程试图读取位于 COSMOS 文件系统的文件时, COSMOS 的 VFS 接口层将请求转发给此节点机器上的 COSMOS 客户进程, 客户进程首先在本地的缓存块中搜寻此文件对应的数据块. 如果缓存命中, 则直接将数据返回用户进程; 否则将向存储这些数据的目的存储服务器发出读文件请求, 存储服务器进程随后将启动线程取出相应的数据, 然后将数据返回给发出请求的 COSMOS 客户进程.

相比读过程, 写文件过程略微简单. 用户进程发出写文件请求后, COSMOS VFS 接口同样将请求转发给客户进程, COSMOS 客户进程将相应的缓存块状态标记为"脏". 在某些情况下, 客户进程将向相关的元数据服务器发出数据冲刷请求, 将所有缓存块中的数据传输到对应的存储服务器进程, 由这些存储服务器进程将数据写入到其本地存储子系统中.

2.2　BCL-3 高速通信协议

原有的 COSMOS 版本使用 TCP/IP 通信协议在多个进程之间进行通信, 而本文将要讨论的 COSMOS 文件系统使用的是为曙光 3000 机群设计的 BCL-3 高速通信协议.

2.2.1　协议特征

BCL-3 是一种基于高速网络的半用户级通信协议. 半用户级通信的定义是指综合了用户级通信与内核级通信的特点, 既保持了内核级通信协议高安全性, 又具有用户级通信的高性能, 并且协议实际开销比传统内核级通信要低得多.

BCL-3 具有高带宽与低延迟特性, 提供了基于共享内存的节点内通信协议, 对 SMP 节点有足够支持. 同时由于使用了基于 ACK/NAK 的流量控制协议, BCL-3 具有良好的可扩展能力. 此协议一个非常重要特性是提供了机群节点间可靠的数据传送服务. 因为使用了标准接口与不依赖特定系统的实现方法, BCL-3 具有非常好的可移植性. 图 2 是曙光 3000 机群上 BCL-3 高速通信协议支持的系统软件体系结构示意图. 它的设计目标是为多种系统软件, 包括 COSMOS 文件系统, 提供足够的通信支持.

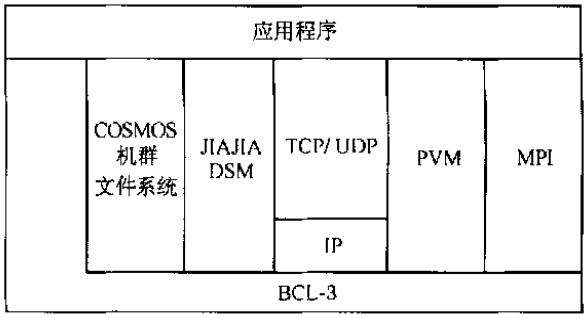

图 2　BCL-3 高速通信协议与通信软件层次关系示意图

2.2.2　应用程序接口 API

BCL-3 高速通信协议提供了灵活的应用程序调用接口(API), 所有的发送与接收原语都采用非阻塞工作方式, 通过另外的查询接口 API 来等待通信的完成. 这种设计方式方便了用户编程, 同时用户程序可以尽量做到通信与计算重叠, 从而提高应用的性能.

在 BCL-3 中, 每个被发送的消息在目的进程中必须有相应缓冲区进行接收, 我们将该方式称为 Rendezvous, 指通信双方必须进行必要的约定, 消息发送进程必须得到接收进程缓冲区的相关信息.

为了对机群文件系统提供支持, BCL-3 使用一种特殊的缓冲区并增加了写远程页面的 API. 当 COSMOS 进程需要对其它进程的缓冲区进行写操作时, 如 Cache 冲刷, 通信双方不必像使用普通缓冲区一样需要约定. 这种方式可以减少协议上的开销, 另外还可以减少内存拷贝, 这样对于提高系统带宽非常有利.

2.3　基于 BCL-3 的 COSMOS 实现

与使用标准 TCP/IP 通信接口的 COSMOS 文件系统相比, 基于 BCL-3 的 COSMOS 文件系统实现并非简单替换原有的通信原语, 因为毕竟 BCL-3 与 TCP/IP 具有不同的通信属性. 为了发挥 BCL-3 的性能, 我们对文件系统代码进行了适量调整, 在保证原有文件系统协议语义的基础上, 结合 BCL-3 特点来进行文件系统协议的设计. 在相关工作的基础上[15,16], 我们分析 COSMOS 的通信模式大致可划分为两类: 一类是单纯的数据传输, 主要与读写文件操作相关, 它经常涉及较大量的数据, 只会发生在 COSMOS 客户进程与存储服务器之间; 另外一类是小数据量的传送, 一般出现在元数据服务器与系统其它部分的信息交换流程中. 在我们以下介绍的内容中, 我们只针对前一类通信进行了优化, 此时元数据

服务器与系统其它部分通信仍然使用 TCP/IP 协议.

当 COSMOS 客户需要从存储服务器读取数据时, 它使用 TCP Socket 向存储服务器组发出请求, 而客户进程本身是单线程, 为了保证后续操作的正确性, 所以客户进程需要阻塞在等待这些存储服务器使用 BCL-3 协议传输的数据上, 一旦数据全部传输完成后, 此次 COSMOS 读操作才完全结束.

当 COSMOS 客户需要向存储服务器组写入数据时, 它首先还是使用 TCP Socket 向这些存储服务器发出请求, 然后等待它们的 BCL-3 应答消息, 然后客户使用 BCL-3 协议交替向这些服务器发出数据, 直到所有数据发送完毕.

目前 BCL-COSMOS 只是原型系统, 在现在的实现中还未使用远程写 API, 性能可能会受到一些影响.

3 性能模型

我们引入了并发带宽利用率来描述 COSMOS 文件系统性能. 并发带宽利用率 e 指 COSMOS 客户得到的实际聚集带宽与理论聚集带宽的比值, 它描述了 COSMOS 机群文件系统对于并发 I/O 负载的处理效率.

我们将分别讨论两种不同存储服务器配置下的 COSMOS 文件系统性能.

3.1 并发带宽利用率

网络不能提供足够的带宽为数据传输服务时, 机群文件系统的理论带宽上限为所有存储服务器网络的聚集带宽. 反之, 当文件系统的带宽不足以饱和网络能力时, 理论带宽上限为存储服务器文件系统的聚集带宽. 设系统实际可利用带宽为 BW_{real}, 存储服务器的网络聚集带宽与 I/O 聚集带宽分别为 $BW_{cluster\text{-}net}$和 $BW_{cluster\text{-}I/O}$, 则

$$e = \frac{BW_{real}}{BW_{ideal}} = \frac{BW_{real}}{\min(BW_{cluster\text{-}net}, BW_{cluster\text{-}I/O})}. \quad (1)$$

COSMOS 文件系统中文件读写的总时延由文件系统协议开销 $t_{protocoal}$, 存储服务器上文件读写时间 $t_{I/O}$以及文件的网络传输开销 t_{net}组成. 如下所示:

$$t = t_{net} + t_{I/O} + t_{protocol}.$$

设文件大小为 s, 单个 COSMOS 客户获得文件读写带宽为

$$bw = \frac{s}{t} = \frac{s}{t_{net} + t_{I/O} + t_{protocol}}. \quad (2)$$

设 COSMOS 文件系统由 m 个客户与 n 个存储服务器节点构成, 单个存储服务器提供的峰值文件读写带宽为 $bw_{I/O}$, 网络带宽为 bw_{net}.

3.1.1 网络磁盘分组方式

第 2.1.1 小节已经对网络磁盘分组方式进行了介绍. 设单个存储服务器为每个客户节点提供的实际文件平均读写带宽为 $bw_{I/O\text{-}real}$, 在客户端得到的实际聚集网络带宽为 $bw_{net\text{-}real}$, δ 为以时间为度量的协议开销, 则系统的实际聚集带宽 BW_{real}如下所示:

$$BW_{real} = m \times \frac{s}{\frac{s}{bw_{net\text{-}real}} + \frac{s}{n \times bw_{I/O\text{-}real}} + \delta}.$$

在理想情况下, $\delta \to 0$, 即协议开销可忽略. 设$b = bw_{I/O\text{-}real}$, $b' = bw_{net\text{-}real}$, $k = \frac{bw_{net\text{-}real}}{bw_{I/O\text{-}real}}$, 则 $BW_{real} = \frac{m \times b \times k}{n + k}$.

另设 $B = bw_{I/O}$, $B' = bw_{net}$. 当系统受限于 I/O 带宽时, 并发 I/O 带宽利用率为

$$e = \frac{m}{n} \times \frac{b}{B} \times \frac{k}{n + k}, \quad k > 1. \quad (3)$$

前两项与系统规模有关, 设 $\theta = \frac{m}{n} \times \frac{b}{B}$, θ 可视为描述子系统规模扩展能力的参数, 则式(3)可进一步简化为

$$e = \theta \times \frac{k}{n + k}. \quad (4)$$

同理可证, 当系统受限于网络带宽时, 并发 I/O 带宽利用率为式(5)所示, 其中 $\theta = \frac{m}{n} \times \frac{b'}{B'}$.

$$e = \theta \times \frac{k'}{n + k'}, \quad k \times k' = 1. \quad (5)$$

3.1.2 非网络磁盘分组方式

假定 $c = \frac{m}{n}$, $c \in N$, 此时单个 COSMOS 文件的数据块全部存放在一个存储服务器上. 考虑所有客户同时对此文件系统发出 I/O 请求的情形, 假定在理想情况下, 所有被访问文件正好均匀分布在存储服务器上, 则此时系统可视为 n 个相互独立的子系统, 每个子系统由 c 个客户节点与 1 个存储服务器组成, 整个系统的带宽利用率就是单个子系统的带宽利用率. 则每个 COSMOS 子系统的实际聚集带宽 BW_{real}如下所示:

$$BW_{real} = c \times \frac{s}{\frac{s}{bw_{net\text{-}real}} + \frac{s}{bw_{I/O\text{-}real}} + \delta}.$$

若忽略 δ, 则带宽利用率如下:

$$e = \theta \times \frac{k}{n + k}, \text{ 受限于 I/O}. \quad (6)$$

$$e = \theta \times \frac{k'}{n + k'}, \text{ 受限于网络.} \quad (7)$$

可见, 本小节结论与第 3.1.1 完全一致.

根据上述等式, 我们对影响并发带宽利用率的因素有如下结论:

① 系统规模一定时, 若系统受限于 I/O, e 随实际网络带宽与 I/O 带宽的比值的上升而上升.

② 系统规模一定时, 若系统受限于网络, e 随实际网络带宽与 I/O 带宽的比值的下降而上升.

③ 提供具有良好规模可扩展性能的 I/O 与网络接口可提高并发利用率.

4　性能测试

本章主要介绍相关实验的结果及分析.

4.1　实验平台及方法

我们的实验平台是曙光 3000 超级服务器. 所有机群节点通过高速网络与 100Mbps 以太网络连接. 在这个实验平台上由 4 个存储服务器、1 个元数据服务器以及 16 个客户节点组成.

测试中, 多个 COSMOS 客户节点上同时启动测试程序向机群文件系统提交多个读写请求, 由测试程序得出 COSMOS 文件系统的带宽性能.

测试所使用的工具是 Intel 公司的并行文件系统测试工具. 这个工具将读写的文件大小和每次读写的块大小作为参数. 这样可以测试文件系统对不同文件大小以及读写粒度的敏感程度.

在测试过程中, 我们都采取了一些措施以避免 COSMOS 客户节点与存储服务器两端的 Cache 影响.

4.2　性能分析

图 3 和图 4 是一组基于 BCL-3 高速通信协议和基于 TCP/IP 的 COSMOS 的性能测试对照示意图[①]. 在这两个图中, 纵轴表示 COSMOS 客户端得到的文件系统聚集带宽(单位为 MBps), 横轴表示 COSMOS 文件系统中并发访问的客户节点个数.

这组数据表明, 由于通信性能的提高, 网络带宽已经大大超过了存储设备提供的带宽, 系统瓶颈已经转移到磁盘 I/O, 而不在是网络接口. 因此, 图中

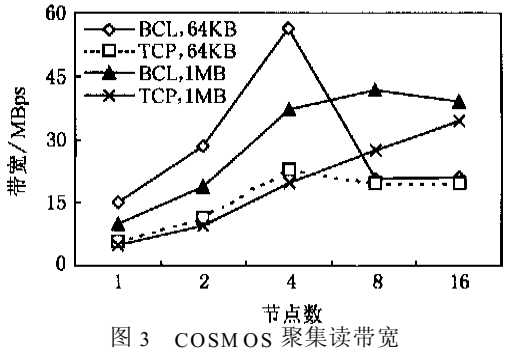

图 3　COSMOS 聚集读带宽

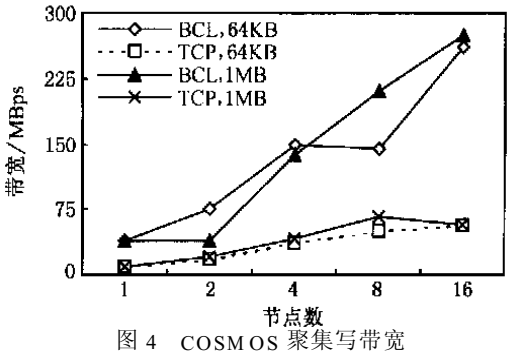

图 4　COSMOS 聚集写带宽

客户数从 1 增加到 4 时, 基于 BCL-3 通信协议的 COSMOS 文件系统带宽要明显优于基于 TCP/IP 的系统, 前者相对后者的带宽加速比最大, 达到了 2.63(读) 和 4.85(写).

但图 3 和图 4 同时还反映出, 当单个存储服务器上需要同时服务于两个或以上的 I/O 读请求时[②](对应于客户数达到 8 个以上时), 基于 BCL-3 相对基于 TCP/IP 的系统已经没有性能优势. 我们认为这是由于存储服务器上的 I/O 子系统本身在重负载条件下, 处理并发 I/O 请求的能力急剧下降, 由通信性能提高带来的带宽优化已经对最终性能影响微乎其微, 所以 BCL-COSMOS 的读文件带宽已经基本上与基于 TCP/IP 的系统持平.

为了证实以上的推论, 我们对曙光 3000 机群服务节点和计算节点上的本地硬盘进行了测试. 我们选择了两种文件读写粒度并变化 I/O 的并发度[③], 测试文件大小为 128MB. 图 5 的磁盘 1 和磁盘 2 系列分别对应计算节点和服务节点上的本地硬盘. 纵轴表示本地文件系统聚集带宽(单位为 MBps), 横轴表示此存储子系统上的并发的 I/O 读线程个数.

从图 5 可以看出, 单个硬盘的性能变化趋势基

① 此处的存储服务器仍然是普通的曙光 3000 机群计算节点, 存储子系统为单个磁盘.

② 写请求因为总是被缓存, 所以这里不讨论.

③ 由于 COSMOS 文件系统中存储服务器的写操作没有使用 sync 系统调用对其进行同步, 所以基本都被写入系统 File Buffer, 相当于写 Cache, 因此我们这里不再对写性能做单独测试.

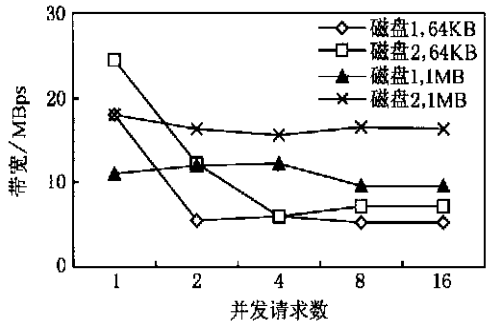

图 5 I/O 子系统聚集读带宽变化曲线

本保持一致, 当并发度由 1 增加到 2 时, 小粒度读操作的聚集带宽迅速下降; 而大粒度操作则基本不变. 我们认为这是合理的结果, 因为小粒度 I/O, 磁盘寻道与旋转延迟构成了主要开销, 而大粒度操作则以传输时间为主.

图 6 给出了根据实测值得到的并发 I/O 带宽利用率. 纵轴表示实际并发 I/O 带宽利用率, 横轴表示 COSMOS 文件

当客户数小于存储服务器个数时, 对于基于 TCP 的 COSMOS, 系统受限于网络带宽, 所以可以从图 6 看出由于通信带宽的提高产生的性能改善; 基于 BCL-3 高速通信协议的 COSMOS 的并发 I/O 带宽利用率要远高于基于 TCP/IP 的系统, 很快几乎就达到系统的最优性能.

根据式(6), 我们计算了 BCL-COSMOS 的理论并发 I/O 带宽利用率, 大约为 90%, 这与实际测量值非常接近. 对于基于 TCP 的 COSMOS, 由于此时系统受限于网络, 我们根据式(7)计算出理论值约为 60% (1MB 粒度) 与 70% (64KB 粒度), 与实际值相差较远, 我们认为是式(5)与式(7)中对协议层开销的省略导致, 因为此时 TCP 小消息延迟要远大于 BCL-3.

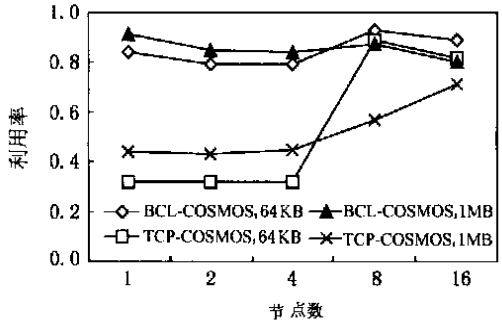

图 6 并发带宽利用率变化曲线

但当客户数量增加到一定程度时, 每个存储服务器需要支持两个以上客户时, COSMOS 系统已经严重受限于 I/O 时, 这时通信带宽的提高对于整个并发带宽利用率的影响已经无足轻重, 基于 TCP 与 BCL-3 协议的 COSMOS 的并发带宽利用率已经非常接近.

所以, 采用高性能的通信网络可以极大地提高 COSMOS 总的聚集带宽, 但随着规模扩大, 由通信性能改善而获得的好处已经因为 I/O 带宽远低于通信带宽而显现不出来. 因此, 我们还对基于高性能 I/O 子系统和 TCP/IP 的 COSMOS 文件系统进行了测试. 从表 1 可以看到, 在 I/O 带宽不成为系统瓶颈时, COSMOS 文件系统具有优异的可扩展能力. 这也说明, 对于类似 COSMOS 结构的机群文件系统, 提供高性能的存储服务节点是非常必要的.

表 1 基于 TCP/IP 的 COSMOS 可扩展性能测试结果

客户节点个数	读带宽/MBps	写带宽/MBps
1	4.55	8.98
2	4.56	9.09
4	4.47	8.32
8	4.24	8.32
16	3.58	6.65

由此我们可以推断, 如果存储服务器可以提供足够的 I/O 带宽, 则 BCL-COSMOS 文件系统可以获得比表 1 更优的性能.

5 结束语

本文描述了基于 BCL-3 高速通信协议的 COSMOS 机群文件系统的原型系统. 引入了并发带宽利用率来描述文件系统的性能模型并对此文件系统进行了相关测试. 数据分析的结果表明, 提高底层通信带宽可以极大地提高系统的聚集带宽并能提高并发带宽利用率. 在今后的工作中, 我们将继续完善此系统, 使通信与 I/O 系统更加均衡, 以取得更好的性能.

参 考 文 献

1 RS/6000 SP Researce Center. http://www.rs6000.ibm/support/sp/resctr

2 T E Anderson. A case for NOW. IEEE Micro, 1995, 16(1): 54~64

3 马捷. 基于 SMP 节点的机群通信系统关键技术的研究[博士论文]. 中国科学院计算技术研究所, 北京, 2001
(Ma Jie. Research on key issues of communication system on cluster of SMP's[Ph D dissertation](in Chinese). Institute of

Computing Technology, Chinese Academy of Sciences, Beijing, 2001)

4 Michael N Nelson, Brent B Welch, John K Ousterhout. Caching in the sprite network file system. ACM Trans on Computer Systems, 1988, 6(1): 134~154

5 John H Howard, Michael L Kazar, Sherri G Menees *et al*. Scale and performance in a distributed file system. ACM Trans on Computer Systems, 1988, 6(1): 51~81

6 Thomas E Anderson, Michael D Dahlin, Jeanna M Neefe *et al*. Serverless network file systems. In: The 15th ACM Symp on Operating System Principles. Copper Mountain, CO, 1995

7 Randolph Y Wang, Thomas E Anderson, Michael D Dahlin. Experience with a distributed file system implementation. University of California, Tech Rep: CSD-98-986, 1998

8 Ajay Mohindra, Murthy Devarakonda. Distrubted token management in calypso file system. In: Proc of the IEEE Symp on Parallel and Distributed Processing. New York, 1994

9 An Introduction to GPFS 1.2. 1998. http://www.rs6000.ibm.com/resource/technology/paperl.html

10 HaiFeng Yu, Amin Vahdat. Design and evaluation of a continuous consistency model for replicated services. In: Proc of the 4th Symp on Operating Systems Design and Implementation (OSDI). San Diego, 2000

11 Randal C Burns, Robert M Rees, Darrell D E Long. Consistency and locking for distributerly updates to web servers using a file system. In: Workshop on Performance and Architecture of Web Servers. Santa Clara, CA, 2000

12 Steven A Moyer, V S Sunderam. PIOUS: A scalable parallel I/O system for distributed computing environments. In: Proc of Scalable High-performance Computing Conf. Knoxville, 1994

13 S Kleiman. Vnodes: An architecture for multiple file system types in Sun Unix. In: Summer Usenix Conference Proceedings. Altanta, 1986

14 Peter J Braam, Philip A Nelson. Removing bottlenecks in distributed filesystems: Coda & interMezzo as examples. In: Proc of LinuxExpo 1999. San Jose, CA, 1999

15 史小东, 冯军. 基于 NOW 的机群文件系统合作式缓存技术研究. 见: 第六届计算机科学与技术研究生学术讨论会论文集. 大连, 2000. 321~326
(Shi Xiaodong, Feng Jun. Research on cluster file system cooperative cache technology based on NOW. In: Proc of the 6th Workshop on Computer Science and Technology for Postgradute(in Chinese). Dalian, 2000. 321~326)

16 王建勇. 可扩展的单一映象的文件系统[博士论文]. 中国科学院计算技术研究所, 北京, 1999
(Wang Jianyong. Scalable single-image file system [Ph D dissertation](in Chinese). Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 1999)

贺　劲　男, 1974 年生, 博士研究生, 主要研究方向为机群文件系统与通信系统.

徐志伟　男, 1956 年生, 研究员, 博士生导师, 主要研究方向为计算机体系结构、操作系统和分布式网络计算, 国家"八六三"重点项目曙光 2000 及曙光 3000 的首席设计师、IEEE Computer Socity 出版理事会理事、中国计算机学会理事、国际联络工委主任.

孟　丹　男, 1966 年生, 博士, 副研究员, 主要研究方向为高性能计算机体系结构.

马　捷　男, 1975 年生, 博士, 主要研究方向为机群通信系统.

冯　军　男, 1976 年生, 硕士, 主要研究方向为机群文件系统.