网络存储系统 I/O 响应时间边界性能研究

崔宝江[1]，刘军[2,3]，王刚[2]，刘璟[2]

(1. 北京邮电大学 信息安全中心，北京 100876；2. 南开大学 计算机科学系，天津 300071；3. 天津财经大学 信息系，天津 300222)

摘 要：为了对网络存储系统性能进行预测和改进，利用定量分析法研究了系统 I/O 响应时间与各性能影响因素之间的关系。通过分析网络 RAID 存储系统的数据传输原理，建立了该系统的闭合排队网络模型，并研究了其 I/O 响应时间的性能边界。实验表明，理论性能边界反映了实际系统性能的变化趋势和性能边界。进一步分析发现，当并发任务数较低时，存储中心服务器 CPU 处理能力和缓冲区命中率是影响 I/O 响应时间的关键因素。

关键词：网络存储；性能建模；排队网络；I/O 响应时间

中图分类号：TP302 **文献标识码**：A **文章编号**：1000-436X(2006)01-0069-06

Study on I/O response time bounds of network storage systems

CUI Bao-jiang[1], LIU Jun[2,3], WANG Gang[2], LIU Jing[2]

(1. Information Security Centre, Beijing University of Posts and Telecommunications, Beijing 100876, China;

2. Dept. of Computer Science, Nankai University, Tianjin 300071, China;

3. Dept. of Information, Tianjin University of Finance and Economics, Tianjin 300222, China)

Abstract: In order to predict and improve the performance of networked storage systems, the relationship was explored between the system I/O response time and its performance factors by quantitative analytical method. Through analyzing data flow in networked RAID storage system, we established its analytical model utilizing closed queuing networks and studied on performance bounds of the system I/O response time. Experimental results show that the theoretical bounds are found to be in agreement with the actual performance bounds of the networked RAID storage system and reflect the dynamic trend of its actual performance. Furthermore, we conclude that the CPU processing power and cache hit rate of the server at the storage center are the key factors affecting the I/O response time, as the concurrent jobs are lower.

Key words: networked storage; performance modeling; queuing networks; I/O response time

1 引言

网络技术的飞速发展，促使信息量正以超乎人们想象的速度增长。随着海量数据处理需求的日益迫切，用户对数据存储系统在性能方面提出了更高的要求。现有对存储系统性能方面的研究主要集中在 DAS 存储系统[1]，随着网络存储技术的快速发展，针对网络存储系统性能方面的研究逐渐成为当前热点。

由于存储网络的引入，使影响网络存储系统性的因素更加复杂，不仅涉及到 DAS 存储系统，而且和网络、主机系统性能密切相关。对于网络存储系统性能方面的评价和研究，国内外都做了大量工作。目前，这些关于性能方面的研究大都是定性的[2~4]，

收稿日期：2005-10-10；**修回日期**：2005-11-10

基金项目：国家自然科学基金资助项目（60273031）；高校博士点科研基金资助项目（20020055021）；天津市科技发展计划重点基金资助项目（043800311）

Foundation Items: National Natural Science Foundation of China (60273031) ; Education Ministry Doctoral Research Foundation of China (20020055021); Technological Development Project Foundation of Tianjin(043800311)

定量研究模型仍然有限[5]。本文构建了基于网络RAID结构的网络存储系统，对其结构和数据处理流程进行了分析，建立了它的闭合排队网络模型CQNM(closed queueing networks model)，并在排队网络理论基础上提出了其性能定量分析模型。利用此模型可快速分析网络RAID存储系统I/O响应时间的性能边界，定量地分析影响网络存储系统性能的关键因素，并区分各种影响因素对系统性能的影响程度。

本文的结构如下：第2节描述网络RAID存储系统的结构；第3节建立了网络RAID存储系统的排队网络模型，并在此基础上提出了其I/O响应时间的边界性能分析模型；第4节对I/O响应时间的性能分析模型进行了验证，并利用模型分析了网络RAID存储系统的性能影响因素；第5节是结论。

2 网络RAID存储系统的结构

网络RAID存储系统的结构如图1所示，构建了基于两级RAID5冗余结构的网络存储系统。分布于网络中的存储设备，利用IP存储协议ENBD映射为存储中心服务器的虚拟存储设备。这些虚拟设备通过存储中心服务器中的软RAID设备驱动程序构建成不同级别的RAID存储空间。从而把在网络中分布的存储资源组织成存储中心服务器可利用的具有统一地址空间的虚拟存储空间。

在网络RAID存储系统中，存储中心服务器应用程序对本地虚拟RAID存储设备的读写请求，通过调用ENBD Client端程序，将数据由网络传送到远端的存储服务器。存储服务器中的ENBD Server端程序接收到数据包后，解析出原数据和命令，将读写请求通过设备文件系统或设备驱动程序对存储设备完成具体读写操作，最后将相应信息再反馈回存储中心服务器。在上面对网络存储系统的基本存储数据处理流程进行分析的基础上，可建立其抽象性能分析模型。

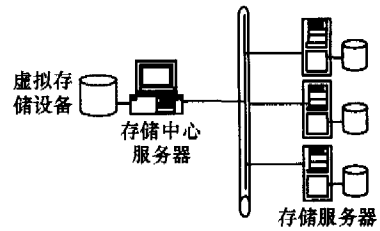


图1 网络RAID存储系统的网络拓扑图

3 网络RAID存储系统的性能分析模型

3.1 建立排队网络模型

基于以上对网络RAID存储系统基本存储流程的分析，我们将数据处理流程中的主要环节抽象为一个个服务节点，其中，存储中心服务器和存储服务器中的中央处理器抽象为CPU服务节点，网卡抽象为网卡服务节点，用于在存储中心服务器和存储服务器间传输数据的网络抽象为网络传输节点，存储服务器中的磁盘抽象为磁盘I/O节点。同时，将每一个服务节点抽象为M/M/1队列，在此基础上建立网络RAID存储系统的CQNM模型，见图2，图中左侧的存储中心服务器通过网络连接到右侧的存储服务器。CPU服务节点负责处理本地的应用程序和数据，网卡服务节点通过网卡向网络中发送或接收数据，网络传输节点通过网络传输数据，磁盘I/O节点负责对磁盘进行读写操作。

3.2 建立边界性能分析模型

建立网络RAID存储系统的CQNM模型后，我们利用BJB[6]（balanced job bounds）方法定量分析其响应时间的性能边界。假定D_i为节点i（$i\in\{1,\cdots K\}$）的服务需求，定义$D_{\max}=\max\{D_i,i\in(1,\cdots K)\}$，$D_{\text{sum}}=\sum_{i=1}^{K}D_i$，$i\in(1,\cdots K)$，$D_{\text{avg}}=D_{\text{sum}}/K$，则网络RAID存储系统I/O响应时间的性能边界为

$$\max(ND_{\max}, D_{\text{sum}}+(N-1)D_{\text{avg}})\leqslant R(N) \qquad (1)$$

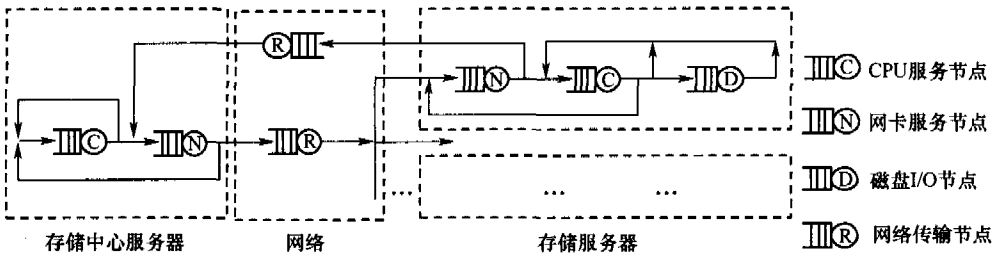


图2 网络RAID存储系统的排队网络模型

下面根据网络 RAID 存储系统中数据处理的具体流程，进一步建立每个服务节点服务需求 D_i 的分析模型。

CPU 服务节点服务需求的计算又分为存储中心服务器和存储服务器两类。存储中心服务器中 CPU 服务节点的服务需求 D_{Cm} 是指用于处理存储行为的时间，包括 CPU 服务节点对网络虚拟磁盘数据读写操作的时间，以及缓存未命中时 ENBD Client 对数据处理的时间和进行传输所消耗的 TCP/IP 协议处理时间，可表示为

$$D_{Cm}= T_{mpro}\cdot\frac{S_m}{STU_m}+(1-P_m)T_{mp}IP_{num} \quad (2)$$

其中，S_m 为单个任务操作所处理的字节数，T_{mpro} 为单个数据条纹单元的平均处理时间，包括内存拷贝和读写操作时间。STU_m 为中心服务器条纹单元的大小，P_m 为读写的缓存命中率，$1-P_m$ 为读写操作访问存储服务器磁盘的概率。在实验环境中，由于文件大小远大于系统缓存，假定写操作的缓存命中率为 0。T_{mp} 为 ENBD Client 对单个 IP 包包含的数据进行数据处理和传输所消耗的 TCP/IP 协议处理时间之和。IP_{num} 为单个任务被分片所得的 IP 数据包数，则

$$IP_{num}=\lceil S_{mRi}/MSS \rceil \quad (3)$$

MSS 为以太网中 TCP 最大分段大小。S_{mRi} 为 RAID*i* 时，存储中心服务器处理单个任务所实际操作的字节数。例如，对于 RAID5 系统的写操作，由于要同时写数据单元和校验单元，故

$$S_{mR5}=S_m\cdot STP_m/(STP_m-1) \quad (4)$$

STP_m 为存储中心服务器一个条纹包含的条纹单元数；对于读操作，则

$$S_{mR5}=S_m \quad (5)$$

同理，存储服务器中 CPU 服务节点的服务需求 D_{Csn} 可表示为

$$D_{Csn}=\frac{1-P_m}{STP_m}\cdot\left(T_{snpro}\cdot\frac{S_{sn}}{STU_m}+T_{snp}\cdot IP_{snnum}\right) \quad (6)$$

其中，T_{snpro} 为存储服务器单个数据条纹单元的平均处理时间，S_{sn} 为存储服务器对于存储中心服务器的一个 RAID5 任务所需要实际处理的字节数，可表示为 $S_{sn}=S_{mR5}/STP_m$。T_{snp} 为存储服务器处理单个 IP 包包含的数据所花费的处理时间。IP_{snnum} 为存储服务器对于中心服务器的一个 RAID5 任务所需要实际处理的 IP 数据包数，可表示为 $IP_{snnum}=IP_{num}/STP_m$。

存储中心服务器网卡服务节点的服务需求可以表示为每个任务通过网卡进行传输所用的平均时间

$$D_{Nm}=(1-P_m)\cdot\frac{IP_{num}\cdot Frames_e}{TRate_e} \quad (7)$$

其中，$TRate_e$ 为以太网的传输速率，$Frames_e$ 为以太帧的大小。

与之类似，存储服务器中网卡服务节点的服务需求表示为

$$D_{Nsn}=\frac{1-P_m}{STP_m}\cdot IP_{snnum}\cdot\frac{Frames_e}{TRate_e} \quad (8)$$

网络传输节点的服务需求可以表示为每个任务在网络传输过程中所用的平均时间

$$D_{Net}=(1-P_m)\cdot\frac{S_{mR5}}{TRate_e} \quad (9)$$

磁盘 I/O 节点的服务需求表示为每个任务在磁盘中进行操作所需时间

$$D_d=\frac{1-P_m}{STP_m}\cdot(1-P_{sn})\cdot\left(seek+latency+\frac{STU_m}{TRate_d}\right)\cdot\frac{S_{sn}}{STU_m} \quad (10)$$

其中，P_{sn} 为存储服务器的缓存读命中率，$1-P_{sn}$ 为读操作访问磁盘的概率，由于文件大小大于系统缓存，假定写操作的命中率为 0。*seek* 和 *latency* 分别为磁盘的平均寻道时间和平均延迟时间，$TRate_d$ 为磁盘的最大持续传输速率。

将上面各节点的服务需求模型代入式(1)中，即可得到基于网络 RAID 存储系统 I/O 响应时间的性能边界分析模型。

4　性能测试与分析

利用上述模型的基础上，对网络 RAID5 存储系统的响应时间和性能影响因素进行分析。网络存储系统实验环境是由 4 台运行 Linux7.3 操作系统的 PC 机组成的一个网络 RAID5 存储系统，其网络拓扑参见图 1。PC 机的配置为 P III 650 CPU，64MB 内存，10/100Mbit/s 网卡，36GB IBM DDYS-T36950 SCSI 磁盘，磁盘为 4MB 缓存，4.9ms 平均寻道时间，2.99ms 平均延迟，磁盘的最大持续传输速率为 35Mbit/s。存储中心服务器条纹单元的大小为 16KB，

一个条纹包含3个条纹单元，单个任务处理的数据量为200MB。以太网的带宽为100Mbit/s，TCP最大分段大小为1460B，以太帧的大小为1518B。此外，CPU平均处理时间 T_{mpro}、T_{mp}、T_{snpro}、T_{snp} 分别取实际测试的平均值0.027ms、0.045ms、0.033ms、0.028ms。

图3为上述网络存储系统实验环境中，在并发写操作时系统响应时间和并发任务数的实际测试值和采用性能分析模型所获得的理论边界值的对比图。从图中可以看出，从性能分析模型获得的理论边界值在不同的并发任务数时，都准确的反映了实际响应时间的变化趋势和其性能边界。在上面验证的基础上，下面利用所建立的性能分析模型，对网络RAID5存储系统响应时间的性能影响因素进行深入分析。

图4为网络RAID5存储系统写操作时响应时间、并发任务数与网络带宽之间的关系图。从图中可以看出，网络带宽在并发任务数较高时是影响网络存储系统响应时间的主要因素，但在并发任务数较低时，对响应时间的作用不明显。图5为网络RAID5存储系统写操作时响应时间、并发任务数与存储中心服务器缓存命中率之间的关系图。和网络带宽的作用不同，从图中可以看出存储中心服务器缓存命中率在不同并发任务数时，对系统响应时间的影响作用很明显，没有因并发任务数的不同而波动。图6为网络RAID5存储系统写操作时响应时间、并发任务数与存储中心服务器CPU单个任务处理时间的关系图。经过对图6的分析可以看出，在并发任务数较小时，随着CPU处理性能的提高，系统响应时间在降低；而当并发任务数较大时，CPU处理性能的高低对系统响应时间没有影响了。图7和图8分别为网络RAID5存储系统写操作时响应时间、并发任务数与存储服务器CPU单个任务处理时间的关系图，以及响应时间、并发任务数与存储服务器中磁盘最大持续传输速率的关系图。和前面几个影响因素不同，在不同并发任务数时存储服务器CPU性能和磁盘性能对系统I/O响应时间的影响很小。

通过上述分析可知，在并发任务数较低时，存储中心服务器CPU处理能力是系统响应时间的关键性能影响因素，通过提高存储中心服务器CPU处理能力的方法可改善网络RAID5存储系统的性能；而网络带宽则是非关键性能影响因素，它对系统响应时间的影响较小。而在并发任务数较高时，网络带宽则成为关键性能影响因素，通过提高网络带宽可有效改善系统的响应时间，而CPU处理能力则变为非关键性能影响因素，它对性能的影响变的很小。除此之外，存储中心服务器的缓冲区命中率在不同任务数时都是影响系统性能的关键因素，

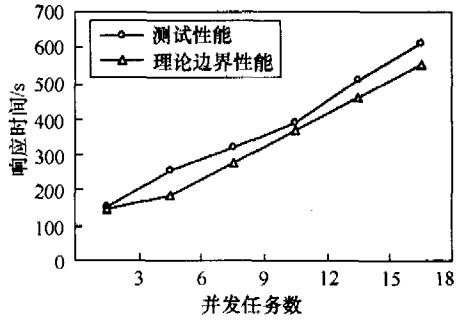


图3 网络存储系统的实测性能与理论性能

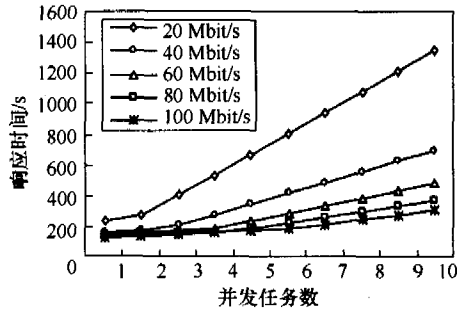


图4 响应时间、并发任务数与网络带宽的关系

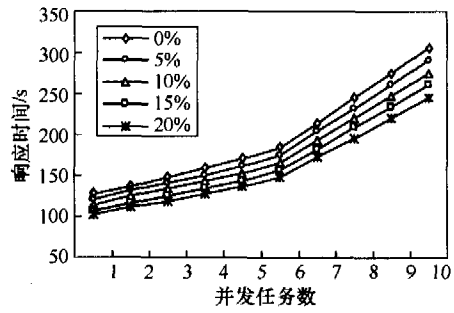


图5 响应时间、并发任务数与缓存命中率的关系

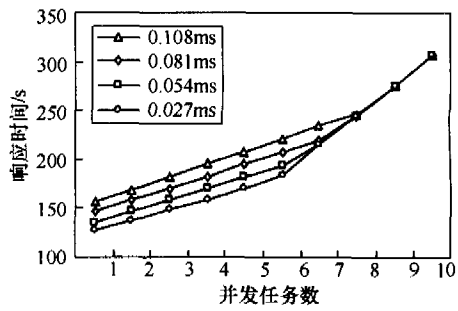


图6 响应时间、并发任务数与存储中心服务器CPU性能的关系

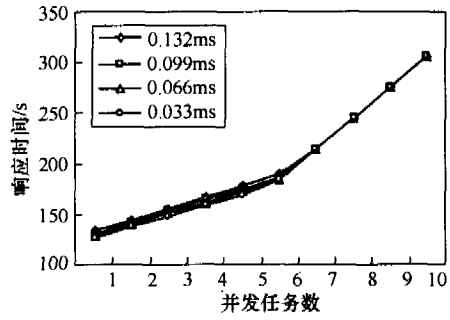


图 7　响应时间、并发任务数与存储服务器 CPU 性能的关系

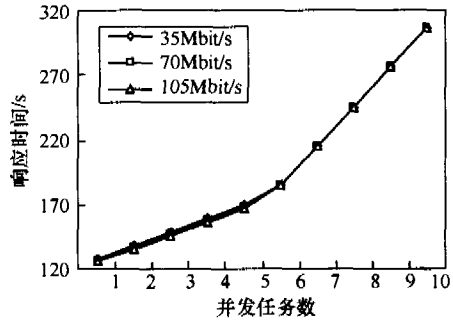


图 8　响应时间、并发任务数与磁盘最大持续传输速率的关系

而存储服务器 CPU 性能和磁盘性能则在不同任务数时对系统 I/O 响应时间的影响都很小，属于非关键性能影响因素。上述网络 RAID 存储系统性能边界分析模型提供了定量分析系统性能影响因素的一种有效手段，在优化系统性能时，可基于上述分析结果着重考虑改善关键性能影响因素，忽略非关键性能影响因素，即可达到事半功倍的效果。

5　结论

由于存储网络的引入，使影响网络存储系统性能的因素更加复杂。为了定量地分析网络 RAID 存储系统的 I/O 响应时间，并区分各种影响因素对系统性能的影响程度，本文通过对网络 RAID 存储系统数据传输过程的分析，建立了其排队网络模型，并在此基础上提出了网络 RAID 存储系统的 I/O 响应时间性能边界分析模型。经过与实测性能值的对比表明，由此模型得出的 I/O 响应时间的性能边界值，反映了网络 RAID 存储系统 I/O 响应时间的变化趋势和性能边界。此外，利用系统响应时间的性能分析模型可知，在并发任务数较低时，存储中心服务器 CPU 处理能力和缓冲区命中率是系统响应时间的关键性能影响因素；而在并发任务数较高时，网络带宽和存储中心服务器的缓冲区命中率则成为系统响应时间的关键性能影响因素。存储服务器的 CPU 处理能力和磁盘的最大持续传输速率对系统性能影响很小，是非关键性能影响因素。

参考文献：

[1] BARVE R, SHRIVER E, GIBBONS P B. Modeling and optimizing I/O throughput of multiple disks on a bus[A]. Proceedings of Sigmetrics '98/Performance '98[C]. New York: ACM Press,1998. 264-275.

[2] LU Y P, DAVID H C D. Performance study of iscsi-based storage subsystems[J]. IEEE Communications Magazine,2003,41(8): 76-82.

[3] HE X B, BEEDANAGARI P, ZHOU D. Performance evaluation of distributed ISCSI RAID[A]. Proceedings of the 2003 IEEE/ACM International Workshop on Storage Network Architecture and Parallel I/O (SNAPI'03)[C]. New Orleans, LA, USA, 2003.

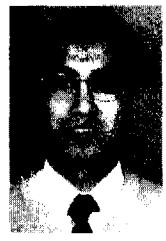
[4] WEE T N, HILLYER B K, SHRIVER E. Obtaining high performance for storage outsourcing[A]. Proceedings of Conference on File and Storage Technologies (FAST '02)[C]. Monterey, California, 2002. 145-158.

[5] ZHU Y L, ZHU S Y, XIONG H. Performance analysis and testing of the storage area network[A]. 19th IEEE Symposium on Mass Storage Systems and Technologies[C]. Maryland, USA, 2002.

[6] LAZOWSKA E D, ZAHORJAN J, GRAHAM G S, *et al*. Quantitative System Performance: Computer System Analysis Using Queueing Network Models[M]. Englewood Cliffs, NJ: Prentice-Hall, 1984.

作者简介：

崔宝江（1973-），男，山东东营人，博士，北京邮电大学讲师，主要研究方向为数据安全、网络攻防、网络存储。



刘军（1963-），男，河北香河人，南开大学博士生，天津财经大学副教授，主要研究方向为网络存储、并行与分布式系统等。

王刚（1974-），男，北京人，博士，南开大学副教授，主要研究方向为数据布局、并行与分布式系统。

刘璟（1942-），男，北京人，南开大学教授、博士生导师，主要研究方向为并行与分布式系统、海量存储、算法设计与分析。

（上接第35页）

[10] LI X, STOJMENOVIC I. Partial delaunay triangulation and degree limited localized bluetooth scatternet formation[J]. IEEE Transactions on Parallel and Distributed Systems, 2004, 15(4):350-361.

[11] FOO C C, CHUA K C. BlueRings- bluetooth scatternets with ring structures[A]. Proceeding of the IASTED International Conference on Wireless and Optical Communication (WOC 2002)[C]. Banff, Canada, 2002. 1563-1567.

[12] BARRIERE L, FRAIGNIAUD P, NARAYANAN L, OPATRNY J. Dynamic construction of bluetooth scatternets of fixed degree and low diameter[A]. Proc. of the fourteenth ACM-SIAM Symp. on Discrete Algorithms (SODA)[C]. 2003. 781-790.

[13] BIENKOWSKI M, BRINKMANN A, KORZENIOWSKI M. Orhan Orhan: cube connected cycles based bluetooth scatternet Formation[A]. Proc. of International Conference on Networking 2005[C]. LNCS, 2005. 413-420.

[14] CHANG C T, CHANG C Y, SHEU J P. Constructing a hypercube parallel computing and communication environment over bluetooth radio system[A]. Proc. of the IEEE International Conference on Parallel Processing[C]. 2003. 447-454.

[15] YANG S F, HUANG T C, YANG C S, BAI S W. A self-determinant scatternet formation algorithm for multi-hop bluetooth networks[A]. Proceedings of the 2003 International Conference on Parallel Processing Workshop (ICPP 2003 Workshops)[C]. Kaohsiung, Taiwan, ROC, 2003. 289-298.

[16] TAN G, MIU A, GUTTAG J, BALAKRISHNAN H. An efficient scatternet formation algorithm for dynamic environments[A]. IASTED Communications and Computer Networks (CCN)[C]. Cambridge, MA, 2002. 68-74.

[17] 郑少仁等. Ad hoc 网络技术[M]. 北京：人民邮电出版社, 2005.
ZHENG S R, *et al*. Ad Hoc Network Technology[M]. Beijing: post Telecom press, 2005.

[18] 马建仓等. 蓝牙核心技术及应用[M]. 北京：科学出版社, 2003.
MA C J, *et al*. Bluetooth Core Technology and Application[M]. Beijing: Science press, 2003.

作者简介：

杨帆（1977-），男，上海人，吉林大学博士生，主要研究方向为 ad hoc 网络、无线传感器网络、蓝牙技术的研究及应用系统设计等。

钱志鸿（1957-），男，吉林长春人，博士，吉林大学教授，主要研究方向为 ZigBee/UWB 无线通信技术的研究与应用、蓝牙技术的研究及应用系统设计、通信系统微弱信号检测理论与应用、通信系统集成电路的故障检测等。