计算机研究与发展
Journal of Computer Research and Development
ISSN 1000-1239/ CN 11-1777/ TP
42(6): 1039～1046, 2005

网络 RAID 存储系统边界性能研究

崔宝江[1,2] 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与技术系 天津 300071)
[3](天津财经大学信息科学与技术系 天津 300222)
(cui-bj@sina.com.cn)

Research on Performance Bounds of Networked RAID Storage Systems

Cui Baojiang[1,2], Liu Jun[2,3], Wang Gang[2], and Liu Jing[2]
[1](*Information Security Centre, Beijing University of Posts and Telecommunications, Beijing* 100876)
[2](*Department of Computer Science and Technology, Nankai University, Tianjin* 300071)
[3](*Department of Information Science and Technology, Tianjin University of Finance and Economics, Tianjin* 300222)

Abstract Previous work on performance evaluation of networked storage systems has been mostly qualitative, and the quantitative analytical method and model are still limited. A quantitative analytical model based on CQN-FC (closed queueing networks with finite capacity) is presented according to the data flow of distributed networked software RAID (dns-RAID). In order to cope with the state space explosion problem of CQN-FC solution, a novel approximate performance bounds analysis (APBA) method is proposed, which has lower computational complexity than other approximate analytical methods in the literature. Experimental testing results show that, the performance bounds of dns-RAID based on CQN-FC calculated by APBA method can reflect the actual throughput and IPO response time bounds in light load, heavy load and over load respectively, and can offer the maximal system load as well.

Key words networked storage; performance bounds; queueing networks; performance analysis

摘 要 目前针对网络存储系统性能的研究大都集中在定性研究方面,缺乏有效的定量分析方法和模型.在有限容量闭合排队网络理论的基础上,提出了网络 RAID 存储系统性能的定量分析模型.并提出了一种新的计算有限容量闭合排队网络系统边界性能的分析方法-APBA 法,和其他近似分析方法相比,APBA 法的计算时间复杂度更低.测试结果表明,通过利用 APBA 方法,由网络 RAID 存储系统的性能定量分析模型获得的系统性能值,可以有效反映网络 RAID 存储系统在轻载区、重载区和过载区的性能边界,以及系统的最大负载量.

关键词 网络存储;边界性能;排队网络;性能分析

中图法分类号 TP302

1 引 言

目前,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.文献[2,3]研究了不同以太网环境和数据访问层对网络存储系统性能的影响,文献[4]分析了引入

收稿日期:2004-03-01;**修回日期**:2004-08-23
基金项目:国家自然科学基金项目(60273031);高校博士点科研基金项目(20020055021);天津市科技发展计划重点基金项目(043800311)

RAID 条纹化技术对 iSCSI 网络存储系统吞吐量和可靠性的影响,文献[5]讨论了基于 IP 广域网环境下网络延迟与网络存储系统的关系,这些关于性能方面的研究都是定性的. 文献[6]虽采用排队论方法定量分析了 SAN 的性能,但采用的是单点服务器模型,忽略了节点容量因素的影响以及各性能影响因素之间的关系. 目前针对网络存储系统性能方面的分析仍然缺少有效的定量分析模型和分析方法.

在排队网络理论基础上,本文建立了网络软 RAID 存储系统的有限容量闭合排队网络模型(也称阻塞的闭合排队网络),为定量分析网络存储系统的性能提供了一种有效的手段. 在对有限容量闭合排队网络系统性能的研究工作中,国内外的学者提出了多种近似分析方法和仿真方法,包括[7,8] throughput approximation, network decomposition, approximate MVA, matching state space 等方法. 对于闭合的阻塞排队网络,当节点规模较大时,由于状态空间随着节点数和任务数的增加而快速增大,上述方法都具有较高的计算成本[7]. 为解决这种状态空间爆炸问题,本文提出了一种在 BAS 阻塞机制下基于 CQN-FC 的 APBA (approximate performance bounds analysis) 方法. 经过数值结果和实际测试结果的验证,APBA 方法可以显著简化模型求解的时间复杂度,其计算结果准确反映了边界性能值以及与任务数之间的关系.

本文的结构如下:第 2 节描述网络 RAID 存储系统的有限容量闭合排队网络模型;第 3 节提出并分析了边界性能近似分析方法,第 4 节采用 APBA 法分析了网络 RAID 存储系统的边界性能,第 5 节为结论.

2 网络 RAID 存储系统排队网络模型

2.1 系统结构和数据处理流程

网络 RAID 存储系统是一种采用基于 Linux PC Cluster 结构构建的基于设备级的具有统一地址空间的大容量存储系统. 在网络 RAID 存储系统中,连接于存储服务器的 RAID 存储设备,通过 IP 存储协议 ENBD 映射为中心服务器的统一虚拟存储设备,然后利用中心服务器中的软 RAID 设备驱动程序,将多个虚拟的存储设备构建第 2 级 RAID 冗余结构. 在本文中构建了基于两级 RAID5 冗余结构的网络存储系统,其系统结构图如图 1 所示.

我们首先介绍网络 RAID 存储系统的软件环境和实现原理,在此基础上进一步分析网络 RAID 存储系统的数据传输流程,最终建立其排队网络模型. 整个系统构建在 Linux 环境中,在存储服务器中安装 ENBD 的 server 模块,在中心服务器安装 ENBD 的设备驱动模块和 client 模块. 中心服务器的 ENBD client 模块和存储服务器的 ENBD server 模块之间建立网络 Socket 连接,存储服务器的硬盘设备通过 Socket 连接映射为中心服务器的一个虚拟存储设备. 由 ENBD 设备驱动模块处理对此虚拟存储设备的操作.

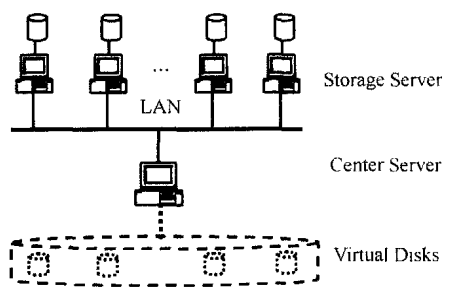

Fig. 1 NS-RAID storage system architecture.

图 1 网络 RAID 存储系统结构图

当中心服务器一侧的用户程序发出 IPO 请求后,由文件系统将其转换为对块设备的请求,转给函数 *ll_rw_block* 处理. 如果是对 RAID5 设备的操作,则调用 RAID5 的请求处理函数 *raid5_make_request* 把 IPO 请求分解成为若干个对硬盘设备的 IPO 操作. ENBD 设备驱动在内核空间,并处于 RAID 驱动之下,如果 IPO 操作处理的硬盘设备是 ENBD 设备,则 ENBD 设备驱动模块将用户对 ENBD 设备的读写请求连接到 ENBD 设备的请求队列中,由 client 模块将请求队列中的请求复制到用户空间中,再利用建立的网络 Socket 接口,传递给远端存储服务器的 ENBD server 模块. ENBD 的 client 和 server 之间提供多端口、多进程的连接,并具有负载均衡的功能.

2.2 排队网络模型

基于以上对网络 RAID 存储系统的分析,我们建立了网络 RAID 存储系统的 CQN-FC 模型,见图 2. 左侧的中心服务器通过 LAN 连接到右侧的多个存储服务器. 网络 RAID 存储系统的所有主要构成元素可以作为排队网络模型中服务节点,包括 CPU 服务节点、网络传输服务节点、磁盘 IPO 节点等 3 类节点. CPU 服务节点包括中心服务器和存储服务器的 CPU 服务节点,负责处理本地的应用程序和数据. 网络传输节点包括中心服务器和存储服务器的网络传输节点,其功能包括数据在接收和发送缓冲

区和网卡的缓冲区之间以 DMA 方式传送,以及从网络中接收数据或者发送数据. 磁盘 I/O 服务节点负责对磁盘进行读写操作. 假设图中各服务节点均为 M/M/1 队列,且每个服务节点的容量均有限.

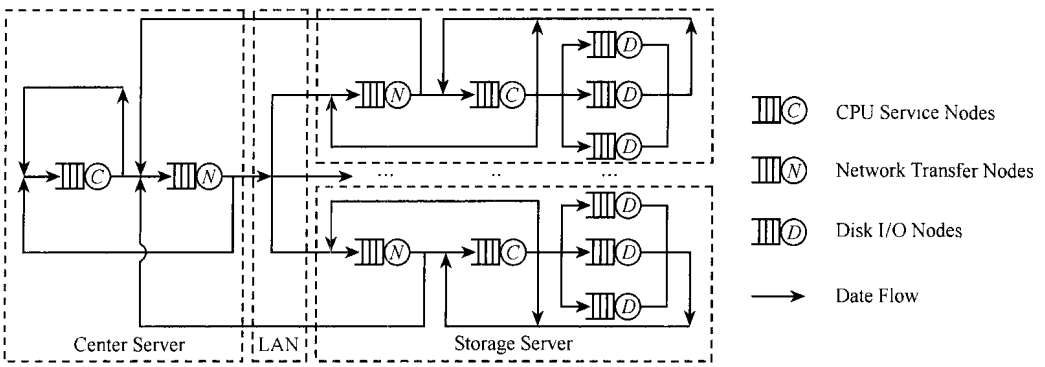

Fig. 2 Queueing network model for NS-RAID storage system.

图 2 网络 RAID 系统的排队网络模型

2.3 服务需求的计算

服务需求是针对某个服务节点而言的. 它定义为在任务处理过程中,一个服务节点被访问的次数与此节点平均每次的服务时间两者的乘积[9].

中心服务器中 CPU 服务节点的服务需求 D_{Cm} 包括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.

$$D_{Cm} = T_{mpro} \times \frac{S_m}{STU_m} + (1 - P_m) \times N_{mrw} \times T_{mp} \times IP_{num}, \quad (1)$$

其中,S_m 为单个任务操作所处理的字节数;T_{mpro} 为中心服务器单个数据条纹单元的平均处理时间;STU_m 为中心服务器条纹单元的大小;P_m 为读和写的缓存命中率;$1 - P_m$ 为任务访问存储服务器磁盘的概率;系数 N_{mrw} 表示中心服务器和存储服务器中 CPU 处理节点和网络传输节点针对中心服务器中的每次 RAID5 操作需要进行的数据处理和传输次数. 对于中心服务器中每次 RAID5 写操作,上述 CPU 处理节点和网络传输节点需要完成 RAID5 写逻辑中一次读和一次写两次数据处理和传输,N_{mrw} 为 2,对应读操作 N_{mrw} 为 1. T_{mp} 为 ENBD Client 对单个 IP 包包含的数据进行数据处理的时间和进行传输所消耗的 TCP/IP 协议处理时间. IP_{num} 为单个任务被分片所得的 IP 数据包数,$IP_{num} = \left\lceil \frac{S_{mR5}}{MSS} \right\rceil$,$S_{mR5}$ 为对于 RAID5 单个任务实际处理的字节数,包括数据单元和校验单元两部分数据,$S_{mR5} = S_m \times \frac{STP_m}{STP_m - 1}$,$STP_m$ 为中心服务器端一个数据条纹包含的数据单元数. MSS 为以太网中 TCP 最大分段大小,$\lceil \ \rceil$ 代表向上取整.

同理,存储服务器中 CPU 服务节点的服务需求 D_{Csn} 可表示为

$$D_{Csn} = \frac{1 - P_m}{STP_m} \times N_{mrw} \times \left(T_{snpro} \times \frac{S_{mR5}}{STU_{sn}} + T_{snp} \times IP_{num}\right), \quad (2)$$

T_{snpro} 为存储服务器单个数据条纹单元的平均处理时间,STU_{sn} 为存储服务器条纹单元的大小,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.

中心服务器端网络传输服务节点的服务需求可以表示为

$$D_{Nm} = (1 - P_m) \times N_{mrw} \times \frac{IP_{num} \times Frames_e}{TRate_e}, \quad (3)$$

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

存储服务器中网络传输服务节点的服务需求表示为

$$D_{Nsn} = \frac{1 - P_m}{STP_m} \times N_{mrw} \times IP_{num} \times \frac{Frames_e}{TRate_e}. \quad (4)$$

磁盘 I/O 节点的服务需求表示为

$$D_d = \frac{1 - P_m}{STP_m} \times \frac{1 - P_{sn}}{STP_{sn}} \times N_{snrw} \times \left(seek + latency + \frac{STU_{sn}}{TRate_d}\right) \times \frac{S_{mR5}}{STU_{sn}}, \quad (5)$$

P_{sn} 为存储服务器的缓存读写命中率;$1 - P_{sn}$ 为任务访问磁盘的概率;$seek$ 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;系数 N_{snrw}表示存储服务器中的磁盘 IPO 节点针对中心服务器中的每次 RAID5 操作需要进行的数据操作次数,对于中心服务器中的一次 RAID5 写操作,由于网络 RAID 存储系统由两级条纹组成,处于第 2 级的磁盘 IPO 节点需要完成 RAID5 写逻辑中两次读和一次写操作; N_{snrw}为 3; 相应的读操作 N_{snrw}为 1; STP_{sn}为存储服务器端一个数据条纹包含的数据单元数.

2.4 访问率的计算

节点 i 的访问率(visit ratio) v_i 是指一个任务处理过程中节点被访问的次数. 对于闭合排队网络,各节点的访问率满足:

$$v_i = \sum_{j=1}^{K} v_j p_{ji}, \ i \in \{1, \cdots, K\}. \tag{6}$$

式中, K 为排队网络的节点数, p_{ji}为从节点 j 到节点 i 的转移概率. 根据网络 RAID 存储系统的 CQN-FC 模型,可得到节点间的转移概率,如表 1 所示:

Table 1 Transition Probabilities for Write Operation of NS-RAID's Nodes

表 1 网络 RAID 存储系统写操作时各节点间的转移概率

Transition Probability	Cm	Nm	Nsn_i	Csn_i	Dsn_{ij}
Cm	P_m	$1-P_m$	0	0	0
Nm	0	0	$1/STP_m$	0	0
Nsn_i	0	0	0	1	0
Csn_i	0	0	0	P_{sn}	$(1-P_{sn})/STP_{sn}$
Dsn_{ij}	1	0	0	0	0

表 1 中, Cm, Nm 分别为中心服务器中的 CPU 处理服务节点和网络传输节点; Nsn_i, Csn_i 为第 i 个存储服务器中的网络传输节点和 CPU 处理服务节点, $i \in (1, \cdots, STP_m)$; Dsn_{ij}为第 i 个存储服务器中第 j 个磁盘 IPO 节点, $j \in (1, \cdots, STP_{sn})$. 根据表 1 中各节点的转移概率和式(6),并设定 v_{Cm}为 1,可以计算得到网络 RAID 存储系统中各节点惟一的访问率.

3 CQN-FC 边界性能分析方法

3.1 模型特征与参数

在介绍边界性能分析方法之前,首先引入 CQN-FC 的特征参数.

假定 CQN-FC 由 K 个单一服务节点构成,每一个服务节点的服务时间满足服务速率为 μ_i, $i \in \{1, \cdots, K\}$的指数分布,服务规则为 FCFS. 网络中的任务为单一类型,任务数量为 N. p_{ji}为任务从节点 j 到节点 i 的转移概率, v_i 为服务节点 i 的访问率, p_{ji}和 v_i 满足 $v_i = \sum_{j=1}^{k} v_j p_{ji}$, $i \in \{1, \cdots, K\}$, D_i 为节点 i 的服务需求, $i \in \{1, \cdots, K\}$, $D_i = v_i/\mu_i$, $D_{max} = \max\{D_i, i \in (1, \cdots, K)\}$, $D_{sum} = \sum_{i=1}^{K} D_i$, $i \in (1, \cdots, K)$, $D_{avg} = D_{sum}/K$. B_i 为节点 i 的容量(包括排队空间的容量和节点前端服务空间的容量), $i \in (1, \cdots, K)$, $B_{min} = \min\{B_i, i \in (1, \cdots, K)\}$, $M = \sum_{i=1}^{k} B_i$,并且单个任务的大小小于 B_{min}. 假定网络基于 BAS 阻塞机制,对于网络中的每一个环型结构 C,都不存在死锁,即 $N < \sum_{i \in C} B_i$.

3.2 CQN-FC 边界性能近似分析方法

对于具有第 3.1 节所描述的特征参数的 CQN-FC,有引理 1.

引理 1[10]. 对于容量有限的具有指数分布的闭合排队网络,在 BAS 阻塞机制下,如果排队网络中的任务数 $N \leq B_{min}$, B_{min}为节点容量的最小值,则此容量有限的闭合排队网络的队列长度分布与相应容量无限的具有指数分布的闭合排队网络相同,具有乘积形式解.

由于有限容量的闭合排队网络在不同的任务数量区间具有不同的阻塞情况,下面将总的任务数量区间 $[1, M)$ 分为 $[1, B_{min}]$, $(B_{min}, M-P+1)$, $[M-P+1, M)$三个区间,其中 $P = \min\{K, B_{min}\}$,分别分析不同区间中 CQN-FC 的边界性能.

(1) 当 $N \in [1, B_{min}]$

当 $1 \leq N \leq B_{min}$时,根据引理 1 及 Little 定律,闭合排队网络的吞吐量 $X_B(N)$等同于相应容量无限的闭合排队网络的吞吐量 $X(N)$,则

$$X_B(N) = X(N) = \frac{N}{\sum_{i=1}^{K} v_i TR_i(N)}. \tag{7}$$

$TR_i(N)$为当排队网络中有 N 个任务时一个任务在第 i 个设备的平均逗留时间,根据到达定理[11]和 MVA 算法[11]有:

$$TR_i(N) = \frac{1}{\mu_i}[1 + \bar{n}_i(N-1)], \tag{8}$$

其中, $\bar{n}_i(N-1)$为排队网络中有 $N-1$ 个任务时,

节点 i 的平均任务数.

因为[10]

$$\sum_{i=1}^{K} v_i TR_i(N) = \sum_{i=1}^{K} D_i[1 + \bar{n}_i(N-1)] = D_{\text{sum}} + \sum_{i=1}^{K} D_i \bar{n}_i(N-1) \geq D_{\text{sum}} + (N-1) D_{\text{avg}}, \quad (9)$$

故 $X_B(N) \leq \frac{N}{D_{\text{sum}} + (N-1) D_{\text{avg}}}$. (10)

同时,考虑到重载情况下每个服务节点的利用率 $U_i(N)$ 不大于 1: $U_i(N) = X_{Bi}(N) D_i \leq 1$,则排队网络最大的系统吞吐量受具有最大服务需求 $D_{\max}$ 的节点的限制.

$$X_B(N) \leq \frac{1}{D_{\max}} = \frac{N}{N D_{\max}}. \quad (11)$$

当 $X_{Bi}(N)$ 满足式(10)和式(11)时,则

$$X_B(N) \leq \frac{N}{\max\{N D_{\max}, D_{\text{sum}} + (N-1) D_{\text{avg}}\}}. \quad (12)$$

(2) 当 $N \in (B_{\min}, M - P + 1)$

在此区间内,由于排队网络中的任务数大于最小的服务节点容量,在上游节点中的任务将会由于被阻塞而使其平均逗留时间增加.

当网络中任务数为 N 时,服务节点 i 中的任务因节点 j 的队列达到最大容量而被阻塞,任务在服务节点 i 的平均逗留时间可表示为[8]

$$TR_i(N) = \frac{1}{\mu_i}[1 + \bar{n}_i(N-1)] + BT_i, \quad (13)$$

其中,BT_i 为节点 i 被阻塞的平均时间. 由于 $BT_i \geq 0$,则

$$X_B(N) = \frac{N}{\sum_{i=1}^{K} v_i\{\frac{1}{\mu_i}[1 + \bar{n}_i(N-1)] + BT_i\}} \leq \frac{N}{\sum_{i=1}^{K} \frac{v_i}{\mu_i}[1 + \bar{n}_i(N-1)]} \leq \frac{N}{D_{\text{sum}} + (N-1) D_{\text{avg}}} \quad (14)$$

与式(1)中相同,考虑到服务节点的利用率,则 $X_B(N)$ 可以表示为

$$X_B(N) \leq \frac{N}{\max\{N D_{\max}, D_{\text{sum}} + (N-1) D_{\text{avg}}\}}. \quad (15)$$

(3) 当 $N \in [M - P + 1, M)$

令 $e = N - (M - K)$,当 $N \in [M - P + 1, M)$ 时,$1 \leq e < K$,e 说明排队网络中至少有 e 个节点的容量已满. 由于在此区间排队网络中必然存在 e 个服务节点被阻塞的情况,因此这个区间内随着任务的增加,排队网络的性能将随任务的增加而急剧下降,系统处于满负荷状态,称此区间为过载区,其吞吐量可表示为

$$X_B(N) = \frac{N}{\sum_{i=1}^{K} \frac{v_i}{\mu_i}[1 + \bar{n}_i(N-1)] + \sum_{i=1}^{K} v_i BT_i}, \quad (16)$$

其中,$\sum_{i=1}^{K} v_i BT_i$ 为所有节点被阻塞的时间之和. 在过载区,闭合排队网络中被阻塞的服务节点数量至少为 e 个,则所有节点被阻塞时间之和必然大于等于其中 $v_i BT_i$ 最小的 e 个服务节点的被阻塞时间之和 $S_b(e)$,即

$$\sum_{i=1}^{K} v_i BT_i \geq S_b(e), e = N - (M - K), \quad (17)$$

$S_b(e)$ 可表示为

$S_b(e) = S_b(e-1) + BT_{be}$,其中

$$BT_{be} = \min_{i \neq b1, \cdots, be-1}\{v_i BT_i, i = 1, \cdots, K\},$$
$$e = N - (M - K). \quad (18)$$

当服务节点 i 中的任务因节点 j 的队列达到最大容量而被阻塞,则每次由节点 j 引起的平均阻塞时间是 $1P\mu_j$[12]. 由于节点 i 在排队网络中可能存在多个下游节点,所以有

$$BT_i \geq \min\left(\frac{1}{\mu_j}\right) = BTMIN_i,$$
$$j \in \{j \mid p_{ij} \neq 0, \bar{n}_j(N) = B_j\}. \quad (19)$$

将式(19)代入 $S_b(e)$ 后得到 $S_{\min b}(e)$:

$S_{\min b}(e) = S_{\min b}(e-1) + BT_{\min be}$,其中,

$$BT_{\min be} = \min_{i \neq \min b1, \cdots, \min be-1}\{v_i BTMIN_i, i = 1, \cdots, K\},$$
$$e = N - (M - K). \quad (20)$$

由式(17)至式(20)可知,$\sum_{i=1}^{K} v_i BT_i \geq S_{\min b}(e)$,$e = N - (M - K)$. 代入式(16)中,同时考虑服务节点的利用率,可得

$$X_B(N) \leq N \Big/ \{\max\{N D_{\max}, D_{\text{sum}} + (N-1) D_{\text{avg}}\} + S_{\min b}(N - (M - K))\}. \quad (21)$$

根据以上对有限容量闭合排队网络 3 个不同区间的分析,通过式(12),(15)和(21)可以计算得到 CQN-FC 吞吐量的边界上限值.

3.3 数值结果分析

下面我们采用 APBA 方法对 CQN-FC 进行计算并分析. 由于篇幅限制,我们仅表现了两组不同性能模型的计算结果.

例 1. 假定一有限容量闭合排队网络的节点数为 4,单类任务数为 13,各节点的容量分别为 $B_1 = 4, B_2 = 5, B_3 = 3, B_4 = 2$,服务速率分别为 $\mu_1 = 1, \mu_2 = 4, \mu_3 = 2, \mu_4 = 2$,访问率分别为 $v_1 = 1, v_2 = 1, v_3 = 1, v_4 = 1$. 此排队网络吞吐量的精确解[13]以及采用上述 APBA 方法获得的性能边界解参见图 3:

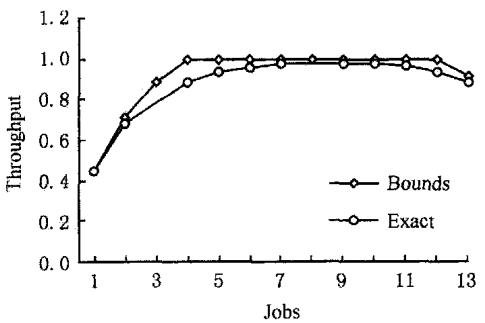

Fig. 3 Throughputs of CQN-FC.

图 3 CQN-FC 的吞吐量

例 2. 假定一有限容量闭合排队网络的节点数为 5,单类任务数为 14,各节点的容量分别为 $B_1 = 2, B_2 = 4, B_3 = 3, B_4 = 2, B_5 = 4$,服务速率分别为 $\mu_1 = 4, \mu_2 = 3, \mu_3 = 2, \mu_4 = 1, \mu_5 = 2$,访问率分别为 $v_1 = 1, v_2 = 1, v_3 = 1, v_4 = 1, v_5 = 1$. 此排队网络吞吐量的精确解[13],以及采用上述 APBA 方法获得的性能边界解参见图 4:

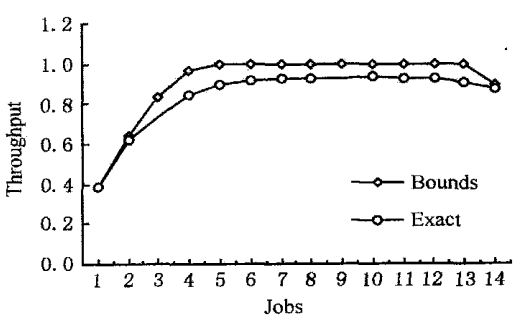

Fig. 4 Throughputs of CQN-FC.

图 4 CQN-FC 的吞吐量

从上面两个例子的分析可以看出,由 APBA 方法获得的性能值准确地反映了所求解 CQN-FC 的性能边界. 并且,由 APBA 方法的具体实现步骤可以看出,其实现过程中最复杂的是对 v_iBTMIN_i 的排序过程,此方法的时间复杂度为 $O(K\lg K)$,远低于其他 CQN-FC 的近似计算方法[7],参见表 2. 由于计算复杂度低,它更适用于多节点多状态容量有限闭合排队网络性能的快速分析.

Table 2 Time Complexity Comparisons Between APBA Method and Others

表 2 APBA 法与其他 CQN-FC 近似求解法的时间复杂度比较

Approximate Analytical Method	Time Complexity
Network Decomposition	$O(rK^4B_{\max}^3)$
Variable Queue Capacity Decomposition	$O(KN^3)$
Matching State Space	$O(K^3 + KN^2)$
Approximate MVA	$O(K^3 + rKN)$
APBA	$O(K\lg K)$

Annotate: iterative number $r, B_{\max} = \max\{B_i, i \in (1, \cdots, K)\}$.

4 网络 RAID 存储系统的边界性能分析和测试

网络 RAID 存储系统的实验环境由 4 台运行 Linux7.3 操作系统的 PC 机组成,网络拓扑参见图 1. 中心服务器的配置为 PIII800CPU,128MB 内存,100M D-Link 网卡,远端存储服务器的配置为 AMD600CPU,64MB 内存,100M D-Link 网卡,每台存储服务器配置 3 块 36GB 的 SCSI 磁盘,4MB 缓存,4.9ms 平均 seek 时间,2.99ms 平均延迟,磁盘最大持续传输速率为 35MB/s. 网络 RAID 存储系统测试的 I/O 操作类型为数据写操作,每次写的 I/O 数据量为 2MB. 存储服务器的存储设备和中心服务器中虚拟存储设备均配置为 RAID5,组成二级条纹 RAID 冗余结构. 其中,中心服务器的条纹单元大小配置为 16KB,存储服务器的条纹单元大小配置为 8KB,两种服务器的单个数据条纹都包含 3 个数据单元. 以太网中 TCP 最大分段大小为 1460B.

此外,在表 3 中给出了上述实验环境中其他的参数值. $B_{Cm}, B_{Csn}, B_{Nm}, B_{Nsn}, B_D$ 分别为中心服务器中 CPU 处理节点、网络传输节点容量,以及存储服务器中 CPU 处理节点、网络传输节点和磁盘 I/O 节点容量. CPU 处理时间、缓冲区命中率和节点的缓冲区容量均采用实际测试的平均值.

Table 3 Model Configurations

表 3 参数值

Parameters	Values	Parameters	Values
T_{mpro}	0.078ms	$Frames_e$	1518B
T_{mp}	0.034ms	B_{Cm}	64MB
T_{snpro}	0.047ms	B_{Csn}	30MB
T_{snp}	0.021ms	B_{Nm}	4MB
P_m	0.431	B_{Nsn}	4MB
P_{sn}	0.677	B_D	4MB

图 5 为网络 RAID 存储系统吞吐量的实际测试值和采用 APBA 方法获得的理论边界值. 从图 5 中可以看出,在轻载区系统吞吐量随任务数的增加而快速增加. 进入重载区后,由于少量节点的容量和处理能力达到最大值,少量任务在部分区间出现被阻塞或丢弃的情况,使总体性能不再增加,实际测试性能曲线呈现出波动的特征. 当进入过载区后,由于较多节点的容量和处理能力达到最大值,大量任务被阻塞或被丢弃,系统性能呈现快速下降的特征,当接近系统最大容量 M 时,由于任务被拒绝率过高,系统性能几近崩溃,实际测试由于难以获得有效数据而终止. APBA 方法获得的理论边界值准确的反映了网络 RAID 存储系统的实际性能变化情况,以及系统性能和任务数之间的关系. 过载区对应的任务数即反映了系统支持的最大任务数.

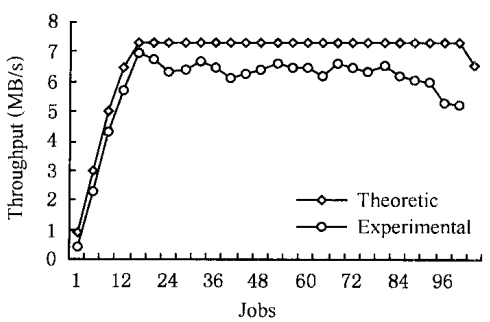

Fig. 5 Comparison of the theoretic and experimental throughput.

图 5 吞吐量的测试值和边界值比较

5 结 论

本文通过对网络 RAID 存储系统结构和存储过程的研究,提出了网络 RAID 存储系统性能的有限容量闭合排队网络模型,提供了定量分析网络 RAID 存储系统性能的手段. 同时,提出了一种边界性能近似分析方法 APBA 法,这种方法可以显著简化模型求解的时间复杂度. 上述模型和分析方法通过数值结果及实际测试结果进行了验证. 测试结果表明,通过利用 APBA 方法,由网络 RAID 存储系统的性能定量分析模型获得的系统性能值,可以有效反映网络 RAID 存储系统在轻载区、重载区和过载区的性能边界,以及系统的最大负载量.

参 考 文 献

1 Rakesh Barve, Elizabeth Shriver, Phillip B. Gibbons. Modeling and optimizing IRO throughput of multiple disks on a bus. In: Proc. Sigmetrics '98PPerformance '98. New York: ACM Press, 1998. 264～275

2 Yingping Lu, David H. C. Du. Performance study of iSCSI-based storage subsystems. IEEE Communications Magazine, 2003, 41(8): 76～82

3 Stephen Aiken, Dirk Grunwald, Andrew R. Pleszkun, *et al*. A performance analysis of the iSCSI protocol. In: Proc. MSS '03. Washington: IEEE Computer Society Press, 2003. 123～134

4 Xubin He, Praveen Beedanagari, Dan ZHou. Performance evaluation of distributed iSCSI raid. The 12th Int'l Conf. PACT2003, New Orleans, 2003

5 Wee Teck Ng, Bruce Hillyer, Elizabeth Shriver, *et al*. Obtaining high performance for storage outsourcing. Conference on File and Storage Technologies(FSAT '02), Monterey, California, 2002

6 Yao-Long Zhu, Shun-Yu Zhu, Hui Xiong. Performance analysis and testing of the storage area network. The 19th IEEE Symposium on Mass Storage Systems and Technologies, Maryland, USA, 2002

7 S. Balsamo, A. Rainero. Closed queueing networks with finite capacity queues: Approximate analysis. The 14th European Simulation Multiconference-Simulation and Modeling: Enablers for a Better Quality of Life(ESM2000), Ghent, Belgium, 2000

8 I. F. Akyildiz. Mean value analysis of blocking queueing networks. IEEE Trans. on Soft. Eng., 1988, 14(4): 418～428

9 E. D. Lazowska, J. Zahorjan, G. S. Graham, *et al*. Quantitative System Performance: Computer System Analysis Using Queueing Network Models. New Jersey: Prentice-Hall, 1984

10 Raif O. Onvural, Survey of closed queueing networks with blocking. ACM Comput. Surv., 1990, 22(2): 83～121

11 M. Reiser, S. S. Lavenberg. Mean value analysis of closed multichain queueing networks. J. ACM, 1980, 27(2): 313～322

12 C. H. Sauer, K. M. Chandy. Computer Systems Performance Modeling. Englewood Cliffs, NJ: Prentice-Hall, 1981

13 R. O. Onvural, H. G. Perrors. Approximate throughput analysis in cyclic queueing networks with finite buffers. IEEE Trans. Software Engineering, 1989, 15(6): 800～808

Cui Baojiang, born in 1973. Lecturer, received the Ph. D. degree from the Nankai University in 2004, His main research area are information security, remote disaster tolerance and network storage.

崔宝江,1973 年生,讲师,博士,主要研究方向为信息安全、远程容灾、网络存储.

Liu Jun, born in 1963. Associate professor, Ph. D. candidate, His main research area are network storage, parallel and distributed systems.

刘军,1963 年生,副教授,博士研究生,主要研究方向为网络存储、并行与分布式系统等.

Wang Gang, born in 1974. Associate professor, received the Ph. D. degree from the Nankai University in 2002, His main research area are data layout, parallel and distributed systems.

王刚,1974 年生,副教授,博士,主要研究方向为数据布局、并行与分布式系统.

Liu Jing, born in 1942. Professor and Ph. D. supervisor. His main research area are parallel and distributed systems, mass storage, design and analysis of algorithms.

刘璟,1942 年生,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.

Research Background

Up to now most studies on the performance evaluation of networked storage systems are qualitative, and the quantitative analytical method and model are still limited, especially for distributed networked RAID (DN-RAID) storage system. In order to analyze the system performance, a CQN-FC (closed queueing networks with finite capacity) quantitative analytical model of DN-RAID storage system is presented. Meanwhile, we put forward an APBA (approximate performance bounds analysis) method to work out the performance bounds of CQN-FC model, which has less time complexity than other methods. The performance bounds of DN-RAID based on CQN-FC calculated by APBA are compared with the exact solutions and experimental results, and found to be in agreement.

This paper is sponsored by the National Natural Science Foundation of China (Grant No. 60273031), Education Ministry Doctoral Research Foundation of China (Grant No. 2002005021) and Tianjin Municipal Science and Technology Development Foundation (Grant No. 043800311).