第33卷 第10期 计 算 机 工 程 2007年5月
Vol.33 No.10 Computer Engineering May 2007
·开发研究与设计技术· 文章编号：1000—3428(2007)10—0271—03 文献标识码：A 中图分类号：N945.15

SAN存储系统的性能分析模型

余寅辉[1,2]，余镇危[1]，杨传栋[1,2]，张 英[2]
（1. 中国矿业大学(北京)计算机系，北京 100083；2. 中国矿业大学中国科学院计算技术研究所网络联合实验室，北京 100080）

摘 要：存储区域网(SAN)作为一种新的网络存储体系结构，已成为存储领域重要的研究方向。性能研究是网络存储研究中的一个重点内容。该文通过对SAN数据传输流程的分析，建立了存储区域网系统的排队网络模型，并求得了基于光纤通道的模型性能参数的定量结果。通过仿真实验发现，模型的理论结果和仿真实验性能测试结果十分接近。

关键词：存储区域网；排队网络；性能分析模型；光纤通道

Performance Analysis Model of SAN Storage System

YU Yinhui[1,2], YU Zhenwei[1], YANG Chuandong[1,2], ZHANG Ying[2]
(1. Dept. of Computer, China University of Mining & Technology (Beijing), Beijing, 100083; 2. United Network Laboratory, Institute of Computing Technology, Chinese Academy of Sciences, China University of Mining and Technology, Beijing 100080)

【Abstract】 Storage area network is a new network storage architecture and has become an important research aspect in storage field. Performance research is quite important in network storage research work. Based on the analysis of the transfer process of SAN, a queueing network model for SAN is established. On the basis of the model, it proposes a quantitative performance analysis model for the SAN based on fiber channel protocol. The simulation results show that the difference of SAN system and system models' performance between the analysis model and simulation result is small.

【Key words】 Storage area network(SAN); Queueing network; Performance analysis model; Fiber channel

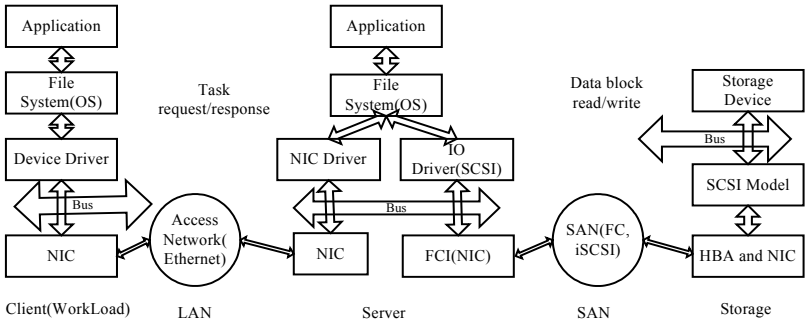
存储区域网络(Storage Area Network, SAN)是一种新的网络存储体系结构[10]。存储网络工业协会(SNIA)如此定义：“SAN是一个网络，其主要目的是在计算机和存储元素之间以及存储元素之间传输数据”。SAN独立于前端网络LAN，与传统的直接相联存储(Direct Access Storage, DAS)相比，具有服务器和网络的吞吐能力提高、服务器负担降低、可扩展性好及安全性好等优点。

网络存储系统的性能研究是网络存储研究领域的重要研究内容。研究报告[1]指出，存储系统的建设预算占IT信息系统建设的比重超过40%，但资源利用率通常不到70%。存储系统资源利用率和性能的分析评价对优化资源配置、提高资源利用率具有指导意义。目前国内外在该领域做了大量的工作，但大都是定性的研究[6~8]，定量的工作仍然有限。

1 SAN系统的I/O路径模型

典型的SAN系统由光纤通道(Fibre Channel, FC)来传输SCCI命令和数据。随着网络带宽的不断增加，SAN不再局限于LAN，已扩展到MAN和WAN环境中。基于SAN系统结构的多个IP存储标准已经陆续发布，包括iSCSI[2](Internet SCSI)、FCIP[3](Fibre Channel over TCP/IP)和iFCP[4](Internet Fibre Channel Protocol)。尽管如此，SAN系统的基本数据传输流程没有变化。其I/O流程结构见图1。Client为存储系统加载工作负载，以任务为单位向Server发送数据请求。任务请求经过IP封装发给Server；Server解析请求数据包，生成数据读写的SCSI命令和数据，通过FC(或TCP/IP)的封装，由SAN传入Storage，Storage通过相应协议解析出SCSI命令(或数据)，从磁盘读写数据；根据任务请求类型和数据量，Server可能通过SAN多次访问Storage，最终向Client传回请求的反馈。

图1 SAN系统I/O路径示意图



2 SAN系统的排队网络模型

2.1 排队网络模型

基于数据流程的分析，建立了SAN系统的开环排队网络模型。Client的数量及其任务请求数量具有随机性，系统中处理的任务个数在不断变化，该特点适合以开环排队网络建模。SAN系统I/O路径中的相应主要构成元素可以作为排队网络模型中的服务节点。模型示意如图2所示。

基金项目：教育部博士点基金资助项目(20030290003)
作者简介：余寅辉(1982－)，男，硕士生，主研方向：计算机网络仿真与性能评价；余镇危，教授、博导；杨传栋，博士生；张 英，研究员
收稿日期：2006-05-22 **E-mail**：dyh480@163.com

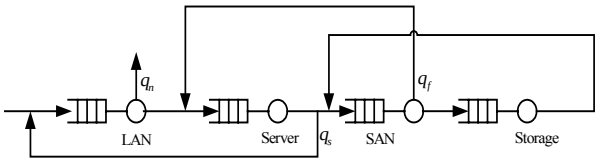


图 2 SAN 系统的开环排队网络模型

以N_t表示在[0, t) 内发生Client请求事件E的总次数，随机过程$\{N_t, t\geq 0\}$为计数过程，它在不相重叠的时间间隔内，事件E发生的次数相互独立。即若t1< t2< t3< t4，则在[t1, t2)内事件E发生的次数N_{t2}-N_{t1}与在[t3, t4)内事件E发生的次数N_{t4}-N_{t3}相互独立，此时计数过程N_t满足独立增量过程。由于计数过程N_t在[t, t+ s)内事件E发生的次数N_{t+s}-N_t仅与时间差s有关，而与t无关，即N_t也是平稳增量过程，因此N_t为泊松过程，可以假定SAN系统的任务到达符合泊松到达模式，各个节点处理时间符合负指数分布，为M/M/1 队列。

2.2 模型的分析和求解

根据模型，以k_i， μ_i， ρ_i分别表示节点i(i = n, s, m, f)，下脚标含义为：s表示Server；*n*表示LAN；*m*表示Storage；*f* 表示SAN)的队列在某一时刻的队长、节点任务处理速率和节点利用率。以$\eta(k_n,k_s,k_f,k_m)$表示稳态概率，根据Jackson[10]定理可以得到系统稳态概率的满足乘积型解，即

$$\eta(k_n,k_s,k_f,k_m)=\prod_{i=s,n,m,f}(1-\rho_i)\rho_i^{k_i} \quad (\text{其中，}\rho_i=\lambda_i/\mu_i) \tag{1}$$

以λ表示系统任务到达速率，q_n表示任务退出排队网络系统的概率；q_s表示任务从Server进入LAN的概率；q_f表示任务从SAN进入Server的概率。根据数据流程分析，$q_n=1/2$，$q_f=1/2$。假定平均每个任务在系统中需要访问Storage A_{io}次，任务到达Storage的平均速率为$A_{io}\times\lambda$。稳态下Server处理任务速率与系统的任务速率均为λ，从而有

$$q_s=1/(1+A_{io})$$

求得：

$$\rho_n=\frac{\lambda}{\mu_n q_n}=\frac{2\lambda}{\mu_n} \tag{2}$$

$$\rho_s=\frac{\lambda(1-q_n)}{q_n q_s\mu_s}=\frac{(1+A_{io})\lambda}{\mu_s} \tag{3}$$

$$\rho_f=\frac{\lambda(1-q_n)(1-q_s)}{q_n q_s q_f\mu_f}=\frac{2A_{io}\lambda}{\mu_f} \tag{4}$$

$$\rho_m=\frac{\lambda(1-q_n)(1-q_s)(1-q_f)}{q_n q_s q_f\mu_m}=\frac{A_{io}\lambda}{\mu_m} \tag{5}$$

由此得到系统的平均响应时间为

$$T_q=\sum_{x=n,s,f,m}\frac{\rho_x}{1-\rho_x}\Big/\lambda$$

系统的吞吐率(单位时间内响应的任务数量)为

$$G=\frac{1}{T_q}$$

同时，系统的性能瓶颈也可由式(2)~式(5)方便地得出，即为利用率最大的节点。

3 基于 FC 的 SAN 系统参数分析

结合数学模型，对基于 FC 协议的 SAN 系统的进行了定量的性能参数分析。

Client数量为N_c，单个Client端给系统发送请求的概率为F_{req}，系统的任务到达率可以表示为$T_q=N_cF_{req}$。对于LAN(忽略传播时延和网络回传时间RTT)，设链路带宽为B_{lan}，每个任务请求的分组个数为N_p，每个请求分组的大小为S_{req}，响应分组的大小为S_{resp}，当传输协议使用TCP，LAN的节点时延可表示为

$$\mu_n=(S_{req}N_p+3S_{resp})/B_{lan}+N_pRTT \tag{6}$$

如果使用 UDP 协议时，节点时延表示为

$$\mu_n=S_{req}N_p/B_{lan} \tag{7}$$

无论传输的数据块多大，最终封装成SCSI数据帧由FC协议传输。每个任务需要访存M_{task}次。假定链路带宽为B_{san}，最大帧长为Len，控制负载为Ctrl，任务访存的平均块大小为S_{block}，由于在FC-1 层采用了 8B/10B编码机制，因此任务在SAN节点的总的网络时延表示为

$$M_f=1.25M_{task}(Ctrl+Len)[S_{block}/Len]/B_{san}\ ([\ ]\text{表向上取整}) \tag{8}$$

对于Server，单个任务在Server(考虑单服务器)每次处理时间均值为Tdeal，每个任务需要访存Mtask次，Server节点针对单个任务处理次数为(Mtask +1)，q_s=1/(Mtask +1)；Server节点处响应时间可表示为

$$\mu_s=(M_{task}+1)T_{deal} \tag{9}$$

对Storage节点，磁盘cache的命中率为P_{hit}(假定读写命中率一致)，磁盘单次I/O的访存上限为B_{limit}，磁盘的平均寻道时间为T_{seek}，平均旋转时延为T_{rota}，磁盘数据传输速率为V_{trans}，任务访存的平均块大小为S_{block}，单个任务的磁盘平均访问时延表示为

$$\mu_m=M_{task}(S_{block}/V_{trans}+(1-P_{hit})(T_{seek}+T_{rota})S_{block}/B_{limit}) \tag{10}$$

由以上分析，易得基于 FC 协议的 SAN 系统的响应时间和各节点处的吞吐量、响应时间和利用率。

4 仿真

采用 OPNET 构建了一个的 FC-SAN 仿真存储系统。基本配置如下：单服务器(Sun Blade 1 000 Model 900 Cu)单CPU，Solaris 操作系统，磁盘类型为 Seagate Barracuda ATA III，磁盘接口类型为 ATA/UDMA-100，LAN 使用 TCP/IP，带宽为 100Mbps，SAN 为 1Gbps 带宽 8 端口的 FC 光纤交换机。客户端数量为 100，每个客户端任务请求频率为 1/100s(见表 1)。

表 1 参数配置

参数	值	参数	值
M_{task}	2(读写各 1)	T_{seek}	8.9ms
P_{hit}	0(变化)	T_{rota}	4.16ms
S_{block}	50KB	V_{trans}	30Mbps
T_{deal}	0.1s	Ctrl	42B
Len	2 120B	N_p	1
S_{resp}	60B	S_{req}	1 000B

图 3 显示了系统平均响应时间随着 Client 任务到达率增加而变化的理论值和计算值的变化曲线。图 4 列出了 LAN 节点和 SAN 节点处的吞吐量随着任务到达率增加而变化的仿真结果和理论值曲线。可以看出，随着外部任务到达的变化系统响应时间增加，网络的吞吐量也不断地增加，但是分析模型值和仿真值相差很小。

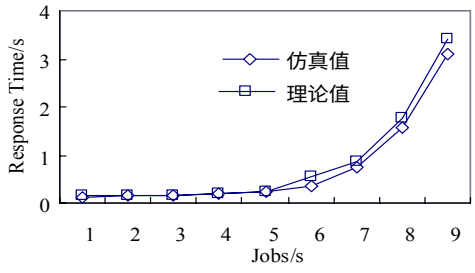


图 3 系统平均响应时间变化

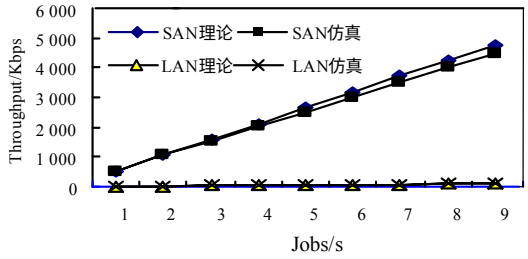


图 4 各节点吞吐率变化

磁盘响应时间与多个因素相关，图 5 说明了在任务到达率为 4 时，磁盘响应时间随着磁盘缓存命中率变化曲线和对应的理论分析值变化，由于磁盘仿真模型的不同和仿真精度的差别，与 OPNET 仿真结果相比，虽然和理论分析值有一定误差，但仍在可接受的范围内。在仿真实验中，逐步增加系统工作负载，通过比较各节点的利用率得出系统的性能瓶颈所在——Server 处(如图 6)，这与理论模型的分析是一致的。

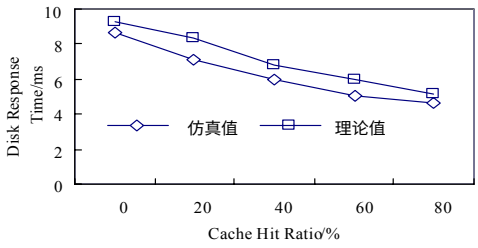


图 5 缓存命中率影响磁盘响应

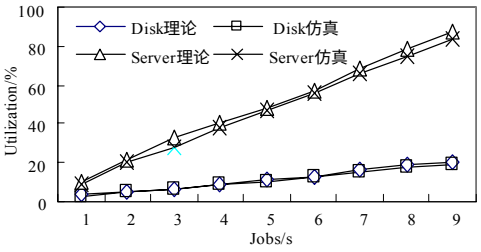


图 6 各节点利用率变化

5 总结

本文针对 SAN 系统的性能提出了一种基于开环排队网络的定量分析模型，求出基于 FC 协议的 SAN 系统性能参数表示，通过仿真实验说明了该分析模型的正确性。

参考文献

1 Lynch M. The Storage Report——Customer Perspectives & Industry Evolution[R]. 2001-06.

2 Rajagopal M, Rodriguez E, Weber R. Fiber Channel over TCP/IP [EB/OL]. 2004. http://www.ietf.org/rfc/rfc3821.txt.

3 Satran J, Meth K, Sapuntzakis C, et al. Internet Small Computer Systems Interface[EB/OL]. 2004. http://www.ietf.org/rfc/rfc3720.txt.

4 Monia C, Mullendore R, Travostino F, et al. A Protocol for Internet Fibre Channel Storage Networking[EB/OL]. http://www.ietf.org/rfc/rfc4172.txt.

5 Storage Networking Industry Association Technical Council. SNIA Shared Storage Model——A Framework for Describing Storage Architectures[EB/OL]. 2003-04. http://www.snia.org.

6 Wang Chaoyang, Zhen Feng, Zhu Yaolong, et al. Simulation of Fibre Channel Storage Area Network Using SANSim[C]//Proc. of the 11th IEEE International Conference on Networks. 2003: 349-354.

7 Bucy J, Ganger G. The Disksim Simulation Environment[R]. Carnegie Mellon University, 2003.

8 Zhang Ming, Qing Yang. Performability Evaluation of Networked Storage Systems Using N-SPEK[C]//Proceedings of the 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid. 2003.

9 Farley M. SAN 存储区域网络[M]. 孙功星, 译. 北京: 机械工业出版社, 2002.

10 林　闯. 计算机网络和计算机系统的性能评价[M]. 北京: 清华大学出版社, 2001.

11 崔宝江, 李　中, 刘　璟. IP 存储广域网性能分析模型[J]. 小型微型计算机系统, 2005, 26(9).

(上接第 262 页)

3.2 基于 IP 上的 BACnet 协议实现

BACnet 体系结构是在 OSI-RM 基础上，为满足控制的实时性要求，对其进行了裁剪，使得 BACnet 协议的体系结构更紧凑、更高效。BACnet 体系结构由 OSI-RM 的 7 层变成了 4 层：物理层，数据链路层，网络层和应用层。在研究中只用到一块 DDC，因此不存在 BACnet 设备互联问题，DDC 的地址由传输介质访问控制(MAC)层分配和寻址。在 IP 协议与 BACnet 协议的转换上，采用常用的异构网络互联技术——“隧道”技术。这种技术就是先将带有控制信息 BACnet 协议包封装在 IP 协议包中，在 Internet 中传输，封装 BACnet 协议包的 IP 包到达目的地时，再将 IP 协议包拆装，分离出有用的 BACnet 协议包。DDC 控制器起着协议包封装和拆装的功能。

4 结论

AmI 综合了多个领域技术发展的成果尤其下嵌入式计算技术的不断发展所取得的成就，展示了人类在不久将来生活的美好前景。智能家居控制系统作为 AmI 的主要子系统之一，实现对家居的成功控制与否直接关系到 AmI 整个系统的应用成败。

参考文献

1 王飞跃，吴朝晖.　ASOS：嵌入式操作系统的发展趋势[N]. 计算机世界, 2000-11-20: B6-B11.

2 He Dongzhi, Wang Zhixue, Li Wei. A Scheduling Algorithm for ASOS and Its Application to Traffic Control[C]//Proceedings of IEEE International Conference on Intelligent Transportation Systems. 2003-10.

3 Weiser M. The Computer for the Twenty-first Century[J]. Scientific American, 265(3): 94-10.

4 ISTAG. Scenarios for Ambient Intelligence in 2010[Z]. 2001-02. http://www.cordis.lu/ist/istag.htm.

5 Weber W, Braun C, Glaser R, et al. Ambient Intelligence——Key Technologies in the Information Age[C]//Proc. of IEEE International Conference on Electron Devices. 2003.