IP 存储广域网的性能研究*)

崔宝江 熊 伟 王 刚 刘 璟
(南开大学计算机科学系 天津300071)

摘 要 由于光纤的传输距离有限,使 FC SAN 的连接距离受到限制。IP 存储广域网利用 IP 网络存储技术扩展了主机与存储设备的连接距离,提供了广域网范围内块级数据的访问。文中对 IP 存储广域网的结构、性能进行了研究,提出了 IP 存储广域网的系统性能评价模型,为评价和改进 IP 存储广域网的系统性能提供了理论依据。理论分析和测试结果表明,虽然影响 IP 存储广域网性能的因素有很多,但当网络带宽低于存储设备吞吐率时,IP 广域网网络带宽成为决定 IP 存储广域网最大系统吞吐率的决定因素。

关键词 IP 网络存储,存储广域网,性能评价

Research on Performance of IP Storage Wide Area Network

CUI Bao-Jiang XIONG Wei WANG Gang LIU Jing
(Department of Computer Science, Nankai University, Tianjin 300071)

Abstract Since the transmission distance of optical fiber is not very long, FC with optical fiber limits SAN to a small region. IP storage wide area network(SWAN)connect geographically separated host and storage device over long distances, and it transports block storage traffic in WAN range using IP-based storage technology. The architecture and performance of IP SWAN are researched in this paper, and a performance evaluation model for IP SWAN is presented. The model offers a theoretic foundation to evaluate and improve IP SWAN's performance. Also there are many factors which impact the throughput of IP SWAN, the result of both the model and experiment shows that IP WAN's bandwidth is the major factor which impacts on the max throughput of IP SWAN, when IP WAN's bandwidth is lower than storage device's throughput.

Keywords IP network storage, Storage wide area network, Performance evaluation

1 引言

存储区域网 SAN 利用光纤通道为分布于区域网中的主机和存储系统提供了高速的数据连接和传输通道。然而,光纤的传输距离有限,即使通过链路扩展器,FC 节点间的距离也只能到达100km。这使得主机和存储系统被限制在一个有限的范围内。而很多基于 C/S 或 B/S 结构的应用,其作用距离从局域网常常跨越到广域网。此外,远程容灾[1]、内容分发、数据网格 datagrid[2]等的不断应用,也对数据存储的范围提出了更高的要求,在更远的距离上拓展存储成为必要。

IP 存储广域网 SWAN[3](Storage Wide-area Network)使跨越广域网范围的数据存储成为可能,它是一种利用 IP 网络存储技术实现全球范围内块级数据访问的存储网络结构。它可以在现有广域网环境中使用户利用 IP 存储协议实现对全球范围内网络存储环境的访问。本文分析了 IP 存储广域网的结构,并在此基础上提出了 IP 存储广域网的性能评价模型。最后,通过实际测试分析和验证了性能评价模型。

2 IP 存储广域网的结构模型

为了说明各种网络存储系统结构上的差别,SNIA 定义了共享存储模型[4],参见图1 a)。存储系统的各部分被抽象到一个统一的存储域,按照数据不同的组织和处理形式,存储域又细分为文件/记录层、块聚集层、block 层。其中文件/记录层对应为文件系统和数据库管理系统,block 层对应着具体的数据存储空间,数据块的组织和处理在块聚集层进行。块聚集层又细分为在主机层进行的块聚集、在网络层进行的块聚集以及在存储设备层进行的块聚集。

目前的 IP 存储广域网主要分为两大类:一类是基于光纤通道的 SAN 在广域网范围的扩展:两个远程的 FC SAN 利用 FCIP 网桥或 mFCP、iFCP 网关,通过 IP 广域网连接起来。另一类是广域网内的主机和存储设备之间完全基于 IP 网络存储协议进行访问的存储网络,支持此功能的 IP 网络存储协议,包括 iSCSI、HyperSCSI、ENBD 等。

IP 存储广域网和 DAS、SAN 在结构上的区别可以用共享存储模型框架加以说明。在 DAS 块聚集层中,不存在网络层的块聚集,主机层和存储设备层的块聚集操作直接通过电缆传输。SAN 和 IP 存储广域网存在网络层的块聚集,SAN 通过光纤和光纤连接设备实现网络层的块聚集;IP 存储广域网则通过 IP 广域网实现了服务器和存储设备间的访问,完成了网络层的块聚集。

基于 iFCP、FCIP 以及基于 iSCSI、ENBD 等 IP 存储传输协议构建的两大类 IP 存储广域网也可以通过共享存储模型

*)本文得到国家自然科学基金(60273031)、高校博士点科研基金(20020055021)和南开大学科研启动基金的资助。崔宝江 博士研究生,主要研究方向:网络存储、并行与分布式系统。熊 伟 博士研究生,主要研究方向:海量存储与分布式系统。王 刚 讲师,主要研究方向:数据布局、并行与分布式系统。刘 璟,教授,博士生导师,主要研究方向:分布式系统、并行计算。

加以区分，参见图1 b)，c)，d)。基于 iFCP 构建的 IP 存储广域网，它在主机层和存储设备层的块聚集操作通过光纤通道协议实现，光纤通道协议在网络层通过 iFCP 网关实现了和 IP 协议的转换。基于 ENBD 和 iSCSI 的 IP 存储广域网都是通过 IP 广域网实现网络层的块聚集操作，但 iSCSI 在 IP 网络中传输的是 SCSI 设备驱动之下的 SCSI 命令集和数据；而 ENBD 在 IP 网络中传输的是在 SCSI、ATA/IDE 等设备驱动之上的操作命令和数据。

IP 存储广域网在存储域之外的服务子系统，提供了发现、监视、资源管理和配置等辅助功能。例如：iSCSI 和 iFCP 等都是通过 iSNS 服务提供了存储广域网内设备的自动发现和配置管理。ENBD 通过带外机制实现存储设备和主机间的配置管理。

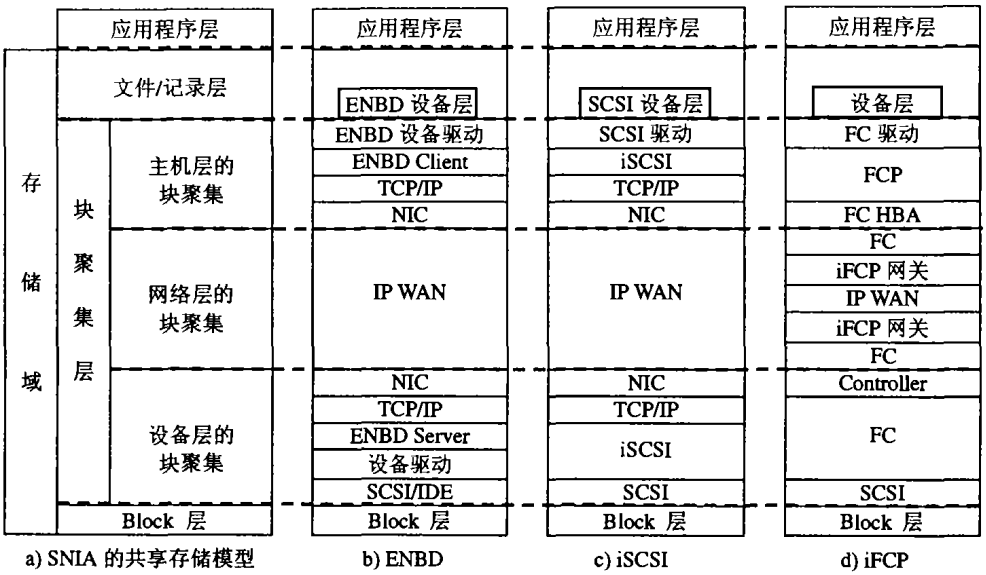

图1 SNIA 的共享存储模型和 IP 存储广域网结构模型

3 系统性能评价模型

从 DAS、SAN 和 IP-SWAN 的共享存储模型可知，三者的主要区别在于主机层和设备层之间的连接方式不同。连接方式的不同导致三者之间性能的差别，DAS 和 SAN 的连接通道传输速率远远高于 IP-SWAN 中 IP 广域网的传输速率。

下面通过引入几组参数来评价 DAS、SAN 和 IP-SWAN 存储系统的性能差别：

T_{hc}，T_{sc}：分别为客户端主机和存储服务器中 CPU 处理数据占用的时间；

T_{hm}，T_{sm}：分别为客户端主机和存储服务器中数据由于内存拷贝及上下文交换所消耗的时间；

T_{si}：存储设备与服务器之间通过存储接口界面传输数据消耗的时间；

T_d：磁盘读写数据的时间开销；

T_{mc}：通过网络接口界面发送和接收数据消耗的时间；

T_n：数据在 IP 广域网中传输所用的时间；

P_h，P_s：分别为客户端主机和存储服务器中处理一个字节平均占用的 CPU 周期；

D_r，D_n：分别为用户读写的原始数据的字节数，以及数据在 IP 广域网中传输所增加的 TCP/IP 报头的字节数；

F_h，F_s：分别为客户端主机和存储服务器中 CPU 的主频；

B_{hm}，B_{sm}：分别为客户端主机和存储服务器中系统总线带宽；

B_{si}：存储设备接口带宽；

B_d：磁盘持续数据传输速率；

B_n：IP 广域网的网络带宽；

B_{mc}：网络适配器的最大传输速率；

C_{hm}，C_{sm}：分别为客户端主机和存储服务器中内存拷贝次数；

N：数据并发读写的磁盘数；

$Thr_{DAS/SAN}$：DAS 和 SAN 的最大系统吞吐率；

$Thr_{IP\text{-}SWAN}$：IP 存储广域网的最大系统吞吐率。

对于 DAS 和 SAN，用户发出的读写请求由客户端主机的 CPU 进行处理，然后经过主机系统内部 I/O 总线，以及主机和存储设备之间的连接通道传输到存储设备，在磁盘中完成对数据的读写操作后反馈给用户。在传统的计算机系统中，上述处理流程都采用流水方式，则 DAS 或 SAN 的系统性能由各个独立的流水线阶段中最费时的阶段决定。这些阶段包括 CPU 处理时间 T_{hc}、内部 I/O 传输时间 T_{hm}、存储接口界面传输时间 T_{si}和磁盘读写操作时间 T_d。

对于 DAS 和 SAN，客户端主机的 CPU 处理时间 T_{hc}、内部 I/O 传输时间 T_{hm}、存储接口界面传输时间 T_{si}和磁盘读写操作时间 T_d，可以分别表示为：

$$T_{hc}=\frac{P_h * D_r}{F_h} \tag{1}$$

$$T_{hm}=\frac{C_{hm} * D_r}{B_{hm}} \tag{2}$$

$$T_{si}=\frac{D_r}{B_{si}} \tag{3}$$

$$T_d=\frac{D_r}{N * B_d} \tag{4}$$

当上述处理过程流水化较理想的情况下，存储系统的性能由处理速度最低的阶段决定，则：

$$Thr_{DAS/SAN}=\frac{D_r}{\max(T_{hc},T_{hm},T_{si},T_d)} \tag{5}$$

将(1)至(4)式代入上式可得：

$$Thr_{DAS/SAN}=\min\left(\frac{F_h}{P_h},\frac{B_{hm}}{C_{hm}},B_{si},N * B_d\right) \tag{6}$$

IP-SWAN 通过 IP 广域网连接客户端主机和存储服务器，存储服务器用于 IP 协议和存储协议的转换，存储设备直接相连到存储服务器上。因此，考虑 IP-SWAN 的系统性能

时，除了 T_{hc}、T_{hm}、T_{si}和 T_d 需要考虑之外，还包括额外需要考虑的存储服务器 CPU 处理时间 T_{sc}、存储服务器内部 I/O 传输时间 T_{sm}、网络接口界面传输时间 T_{nic}和 IP 广域网传输时间 T_n 的影响。它们可以分别表示为：

$$T_{hc}=\frac{P_h*(D_r+D_n)}{F_h} \tag{7}$$

$$T_{sc}=\frac{P_s*(D_r+D_n)}{F_s} \tag{8}$$

$$T_{hm}=\frac{C_{hm}*(D_r+D_n)}{B_{hm}} \tag{9}$$

$$T_{sm}=\frac{C_{sm}*(D_r+D_n)}{B_{sm}} \tag{10}$$

$$T_{nic}=\frac{D_r+D_n}{B_{nic}} \tag{11}$$

$$T_n=\frac{D_r+D_n}{B_n} \tag{12}$$

$$T_{si}=\frac{D_r}{B_{si}} \tag{13}$$

$$T_d=\frac{D_r}{N*B_d} \tag{14}$$

由 IP-SWAN 构建的存储系统，当处理过程流水化较理想的情况下，其性能可由式(15)表示：

$$Thr_{IP\text{-}SWAN}=\min\left(\frac{D_r+D_n}{\max(T_{hc},T_{hm},T_{nic},T_n,T_{sc},T_{sm})},\ \frac{D_r}{\max(T_{si},T_d)}\right) \tag{15}$$

将(7)至(14)式代入上式可得：

$$Thr_{IP\text{-}SWAN}=\min\left(\frac{F_h}{P_h},\frac{B_{hm}}{C_{hm}},B_{nic},B_n,\frac{F_s}{P_s},\frac{B_{sm}}{C_{sm}},B_{si},N*B_d\right) \tag{16}$$

式(16)给出了评价 IP-SWAN 最大系统吞吐率的评价模型。它综合考虑了 IP-SWAN 各个数据处理阶段对系统性能的影响，为评价和改进系统性能提供了理论依据。通过上面的评估模型，可以方便地分析和评估影响 IP-SWAN 性能的瓶颈，为改进和提高 IP-SWAN 性能提供了一种有效的手段。

在影响 IP-SWAN 性能的各个数据处理阶段中，CPU 处理效率和内部 I/O 总线的带宽都较高，假设数据处理复杂度较低同时占用系统资源较小的情况下，它们不会成为影响性能的瓶颈[5]。通过 DAS 或 SAN 和主机连接的存储设备，由于它和主机之间传输数据的存储接口界面带宽也较高(Ultra 160 SCSI 为160MB/s，Fiber Channel 单向传输速率为1Gb/s)，因此只有较慢的存储设备成为影响 DAS 和 SAN 系统性能的主要因素。而对于 IP-SWAN，由于数据在 IP 广域网中进行传输，因此 IP 广域网的网络带宽和存储设备性能都是制约其性能的主要因素。大多数情况下，IP 广域网网络带宽远低于存储设备性能[6]，这种条件下，IP 广域网网络带宽就成为决定 IP-SWAN 最大系统吞吐率的主要因素。

4 系统性能评测

为了评测 IP 存储广域网的性能，采用3台运行 Linux7.3 操作系统的 PC 机构建了基于 ENBD 协议的 IP-SWAN 测试平台。一台 PC 机配置为路由器，连接本地和远端的局域网，并安装 Nistnet 软件模拟 IP 广域网的性能变化。本地局域网中的 PC 机作为客户端服务器，远端局域网中的 PC 机作为存储服务器。PC 机的配置为 AMD 600 CPU，64MB 内存，seagate 20GB IDE 硬盘，10/100M 自适应网卡。采用的测试工具为 Bonnie，测试的性能为大数据量读写，测试的文件大小为 600MB。

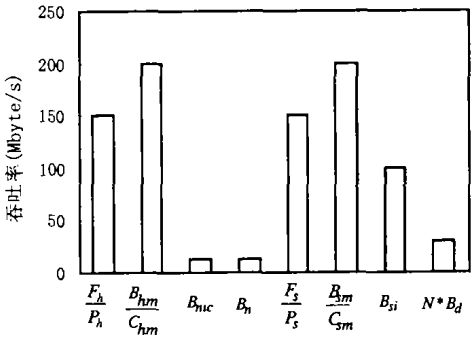

图2 IP-SWAN 测试平台中各数据处理阶段的理论性能

根据式(16)，可以得到在 IP-SWAN 测试平台中影响 IP-SWAN 性能的各数据处理阶段的理论性能值，参见图2。在这个 IP-SWAN 测试平台中，由于受10/100Mbps 网卡传输速率的限制，最大网络带宽为100Mbps。文[6]对广域网存储的网络连接条件进行了统计，发现90%的存储系统之间的网络带宽处于100kbps 到100Mbps 的范围内。从图2可以看出，在 IP-SWAN 测试平台中，即使处于此范围内最高点的100Mbps 的 IP 网络带宽，在所有影响 IP-SWAN 性能的各数据处理阶段中其性能仍然是最低，远小于测试平台中磁盘的持续数据传输速率(30Mbyte/s)，因而较低的 IP 广域网网络带宽成为制约 IP-SWAN 性能的瓶颈。根据公式(16)可知，此时 IP-SWAN 的最大系统吞吐率由 IP 广域网网络带宽决定，即 IP-SWAN 的理论性能为：

$Thr_{IP\text{-}SWAN}=B_n$。

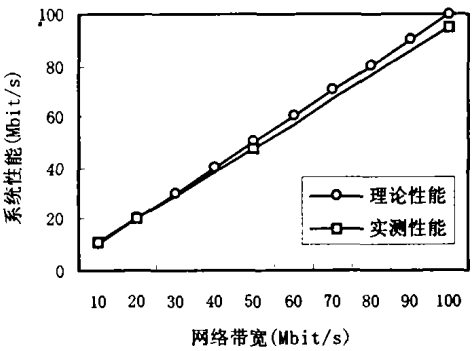

图3 IP-SWAN 的理论性能和实测性能

为了进一步验证性能评价模型，我们在 IP 广域网网络带宽分别为10Mbit/s，20 Mbit/s，50 Mbit/s 和100 Mbit/s 的条件下，对 IP-SWAN 的性能进行了实际测试。测试过程中，在 ENBD 的 client 和 server 端之间启动6个连接进程，此时通过网络监测工具 iptraf 监测到网络带宽利用率已接近饱和。图3是实际测试性能和理论性能的比较结果。由图3可以看出，实际测试性能和由式(16)得到的理论性能非常吻合，这进一步验证了式(16)性能评价模型的合理性。

对于网络存储系统，影响系统性能的因素较多，包括数据布局、磁盘数量、条纹单元大小、块大小等[7,8]。这些影响性能的因素在网络带宽仍有较大空余时能够影响系统性能，但由图3可以看出，当网络带宽的利用率接近饱和时，网络带宽就成为影响 IP 存储广域网系统性能的决定因素。

结论 本文对 IP 存储广域网的类型、结构和性能进行了分析和研究，提出了 IP 存储广域网的系统性能评价模型，并通过实际测试进行了验证。性能评价模型为评价和改进 IP 存

(下转第60页)

拥有根用户(系统管理员)权限的木马对 HIDS 及木马检测系统本身也有着巨大的威胁,因此我们的模型需要特殊的安全策略:

·自我保护机制:数字签名技术对保护数据完整性起着重要的作用,我们的模型将引入该技术对本系统的程序和关键数据进行保护。同样,多实例的技术也可以用于保护本系统,我们采用启动多个监视进程只激活其中一个的策略,一旦其中一个进程被木马删除,其他的进程可以激活和恢复被删除的进程。

·分布式数据存储:安全日志和检测数据对于本系统的检测和分析极为重要,不仅需要检测其是否受到破坏,还要进行恢复。因此我们采用分布式的数据存储方法,将数据的副本传送给其他的 HIDS,保存在其他的主机上。

4.6 模型的总体设计

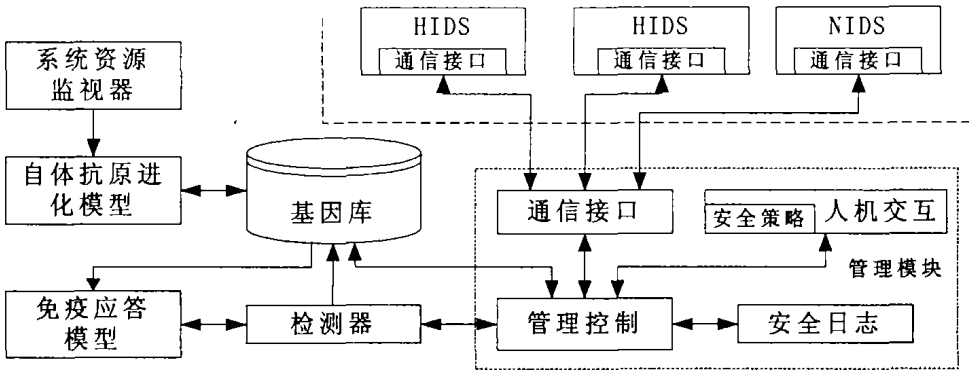

图4 基于人工免疫机制的木马检测系统模型

本系统作为免疫型 IDS 的子系统,必须和 IDS 中的其他模块紧密结合,相互通信。因此本系统设计了一个管理模块,分别由管理控制、通信接口、安全日志和人机交互四个子模块组成。在管理模块的控制下,检测器执行检测并经过管理控制模块审查提交记忆检测体;管理控制模块根据需要访问基因库,和其他 HIDS 交换基因信息(如自体抗原和记忆检测体等)。管理控制模块把对检测数据的分析结果和入侵检测的警告信息保存到安全日志中,同时和其他 HIDS 和 NIDS 交换日志以实现分布式数据存储;而所有的外部通信都由通信接口模块负责。人机交互模块是和管理员进行交互的接口,它包含一个安全策略子模块,其中定义了一些默认的安全策略,在管理员不在的时候指导管理控制处理相应的事务。

结束语 本论文首先指出了当前反病毒软件在检测未知木马方面的不足,介绍了人工免疫系统在自适应性方面的优点,并指出了人工免疫机制在木马检测方面的可行性;然后通过对一些木马新技术的分析,举出一个木马模型证明了现在的计算机安全体系的不足,并提出将木马检测从反病毒软件中迁移到免疫型 IDS 中作为子系统,利用其免疫机制来提高木马检测的自适应能力;最后在 Forrest 研究的基础上,提出了依据进程的系统资源使用状况来映射进程的系统调用的行为模式,并以此建立了基于人工免疫机制的木马检测模型。本系统是基于主机的检测模型,而且提出了自体抗原的进化问题,无疑会增加主机的负担,因此如何提高自体抗原的进化算法和其它免疫算法的效率,如何确定模型中定义的耐受时间、采样间隔、激活阈值等是下一步重点研究的方向。而如何和 IDS 实现良好的联动,如何防御和清除木马、恢复被感染的系统也有待我们进一步探索和研究。

参 考 文 献

1 赵俊忠,黄厚宽,田盛丰. 免疫机制在计算机网络入侵检测中的应用研究[J]. 计算机研究与发展,2003(9):1293～1299
2 焦李成,杜海峰. 人工免疫系统进展与展望[J]. 电子学报, 2003(10):1540～1548
3 左兴权,李士勇,李远贵. 人工免疫系统研究的新进展[J]. 计算机测量与控制, 2002(10):701～705
4 莫宏伟. 人工免疫系统原理与应用[M]. 哈尔滨:哈尔滨工业大学出版社, 2002
5 Dasgupta D. Immunity-based intrusion detection system: a general framework [C]. In: Proc. of the 22nd National Information Systems Security Conf. Crystal City: NIST Publishers, 1999
6 Kim J, Bentley P. The artificial immune model for network intrusion detection [C]. In: 7th European Conf. on Intelligent Techniques and Soft Computing. Aachen, Germany, 1999. 13～19
7 Kim J, Bentley P. Negative selection and niching by an artificial immune system for network intrusion detection [C]. GECCO'99, Orlando, Florida, 1999
8 王锐,等译. 网络最高安全技术指南[M]. 北京:机械工业出版社, 1998
9 朱明,徐骞,刘春明. 木马病毒分析及其检测方法研究[J]. 计算机工程与应用,2003,28:176
10 Stephanie F, Lawrence A, Alan P, Rajesh C. Self-nonself discrimination in a computer [C]. In: Proc. of the IEEE Computer Society Symposium on Research in Security and Privacy, 1994. 202～212
11 Stephanie F, Steven H, Anil S, Thomas L. A sense of self for unix processes [C]. In: Proc. of the IEEE Computer Society Symposium on Research in Security and Privacy, 1996. 120～128

(上接第52页)

储广域网的系统性能提供了理论依据。理论分析和测试结果表明,虽然影响 IP 存储广域网性能的因素有很多,当网络带宽小于存储设备性能时,IP 广域网网络带宽成为决定 IP 存储广域网最大系统吞吐率的决定因素。

参 考 文 献

1 Chang F, et al. Myriad: cost-effective disaster tolerance. Conference on File and Storage Technologies. Monterey, CA, Jan. 2002
2 Cancio G, et al. Tierney, The DataGrid Architecture: [Data Grid TechReport DataGrid-ATF-01]. July 2001
3 Yoshida H. SWAN: Storage Wide Area Network: An Overview of Today's Capabilities and Futrue Directions, Hitachi Data Systems Corp. technical paper, Aug. 2001
4 SNIA Technical Council. Shared Storage Model: A framework for describing storage architectures. draft SNIA TC Proposal document . June , 2001
5 Markatos E P. Speeding up TCP/IP: faster processors are not enough. In: the 21st IEEE Intl. Performance, Computing, and Communication Conf. (IPCCC'02), Phoenix, Arizona, April, 2002
6 Saroiu S, Gummadi P, Gribble S. A Measurement Study of Peer-to-Peer File Sharing Systems. In: Multimedia Computing and Networking Conf. San Jose, CA, Jan. 2002
7 Ng W T, et al. Obtaining High Performance for Storage Outsourcing, FAST 2002, Jan. 2002
8 Hwang K, Jin H, Ho R S C. Orthogonal Striping and Mirroring in Distributed RAID for I/O-Centric Cluster Computing. IEEE Transactions on parallel and distributed systems, 2002, 13(1)