

基于 InfiniBand 协议的云存储关键技术研究^①杨东日^② 王 颖 储 浩

(中国科学院大学工程管理与信息技术学院 北京 100049)

摘 要 为了将网络中大量异构的计算设备、存储设备和网络交换设备整合起来协同工作,实现共同对外提供数据存储和业务访问的功能,给出了基于 InfiniBand 承载协议的云存储架构,并以小型计算机系统接口(SCSI)的存储设备接入 InfiniBand 网络作为研究对象,对涉及的主要关键技术如 SCSI 远程直接内存访问(RDMA)协议(SRP)技术、RDMA 的小型计算机系统接口扩展(iSER)技术和协议转换技术进行了分析和评价。分析表明,InfiniBand 协议不仅适合云存储网络的组建,同时能承载其他通信协议的协议转换技术,能很好地解决 InfiniBand 网络的扩展性问题。通过实验的方法,定量评价了 InfiniBand SRP 协议的性能、InfiniBand 与以太网互联和转换的性能。

关键词 大数据,云存储,InfiniBand,协议转换,SRP

0 引 言

近年来,随着互联网的发展和大量网络计算及存储设备的投入使用,每年新产生的数据以 60% 的速度爆炸式增长,传统的高成本、低效率的存储方式已无法满足大数据量存储的需求。云存储(cloud storage)通过集群计算、网格技术或分布式文件系统等技术,将网络中大量异构的存储设备、计算设备等通过应用软件集合起来协同工作,共同对外提供数据存储和业务访问的功能,具有存储量大、成本低廉、管理方便、数据安全性高等优势,已经成为海量数据存储的发展方向^[1-4]。云存储要求计算节点和存储设备之间的通信具备高速、稳定、可靠、低延时等特点,其主流的传输带宽要求达到 10Gbps 以上,目前应用较广的千兆以太网技术已不能满足云存储应用需求演进的速度。于是产生了带宽更高、开销更小的新型 I/O 结构 InfiniBand,用于高性能计算机内部或集群之间的互联。InfiniBand 协议^[5,6]是较新的网络交换协议,采用一种点到点的通信机制,主要用于将计算设备、网络设备和存储设备高效、稳定地连接在一起。InfiniBand 的规格及标准规范 2000 年正式发表。在 2006 年 11 月 13 日召开的巨型计

算机博览会(SC06)上公布的数据显示,世界前 500 名巨型机的排名中用于服务器互联的公开标准协议中 InfiniBand 的占有率首次超过了私有协议 Myrinet,成为主流数据交换技术。在 2010 年 6 月超级计算机 TOP500 排名中 41.6% 的系统采用了该交换技术,其中,前 10 名中有 50% 的系统采用了该交换技术,前 100 名中有 64% 的系统采用了该交换技术。我国曙光星云、天河 1 号等超级计算系统均采用了 InfiniBand 交换技术。当前,在高性能计算领域服务器互联网络的首选协议已经明确为 InfiniBand。其主要特点是高带宽、低延时,目前其单端口带宽最高可达 56Gbps,端到端延迟到达 200ns 以下。相对于以太网来说,基于 InfiniBand 协议的存储网络具有更高的性能、效率和灵活性。

为实现跨协议通信,发挥 InfiniBand 网络性能及带宽优势,同时兼容其它协议的网络设备,需要采取协议扩展技术,即使用 InfiniBand 协议承载其他的存储协议,如光纤通道(FC)、小型计算机系统接口(small computer system interface,SCSI)、互联网小型计算机系统接口(Internet SCSI,iSCSI)^[7]等。本文以 SCSI 接口的存储设备接入 InfiniBand 网络作为研究对象,研究了海量数据存储技术和协议转换的相关技术,给出了基于 InfiniBand 承载协议的云存

① 863 计划(2012ZX03002003)资助项目。

② 男,1973 年生,博士生;研究方向:计算机虚拟化技术,海量数据存储,云计算等;联系人,E-mail:yang_dr@163.com
(收稿日期:2013-04-15)

储架构,并测试了其相关技术的性能。

1 基于 InfiniBand 承载协议的云存储架构

InfiniBand 设计目标是把服务器中的总线网络化,因此 InfiniBand 兼具网络高性能和总线高带宽、低延时。总线技术采用了直接内存访问 (direct memory access, DMA) 技术,在 InfiniBand 中以远程直接内存访问 (remote direct memory access, RDMA) 形式继承,从而降低了 CPU 在处理通信时的资源消耗,使 InfiniBand 在 CPU、内存、外存等设备之间的通信效率优于以太网和光纤网。另外,InfiniBand 协议使用基于信任的、流控制的机制来确保连接的完整性,数据包极少丢失。InfiniBand 能够在同一个物理网络上支持多个计算机系统之间的 I/O 通信和内部处理器之间的通信。

图1给出了基于 InfiniBand 协议栈的系统体系结构。InfiniBand 协议标准定义了涉及系统通信的多种设备,包括信道适配器、交换机和路由器。信道适配器用于同其它设备的连接,其中主机通道适配器(host channel adapter, HCA)用于连接处理节点,目标通道适配器(target channel adapter, TCA)用于连接 I/O 控制器、交换机和路由。

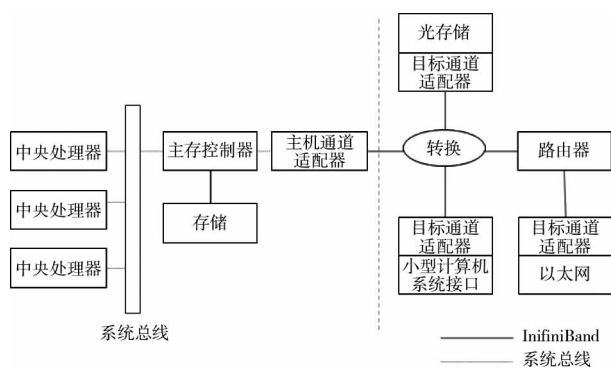


图 1 基于 InfiniBand 协议栈的系统体系结构

在此基础上,本研究构建了基于 InfiniBand 承载协议的云存储系统。学术界给出了很多解决方案^[8-10]。本研究基于既有的研究,结合当前网络存储的实际部署环境,给出了由多种类型的存储集群、存储网络管理设备、桥接服务器、传统的存储服务器及各种联网设备组成的基于 InfiniBand 承载协议的云存储系统架构。其网络架构如图 2 所示。

存储集群包括多种类型的存储方式,例如用于

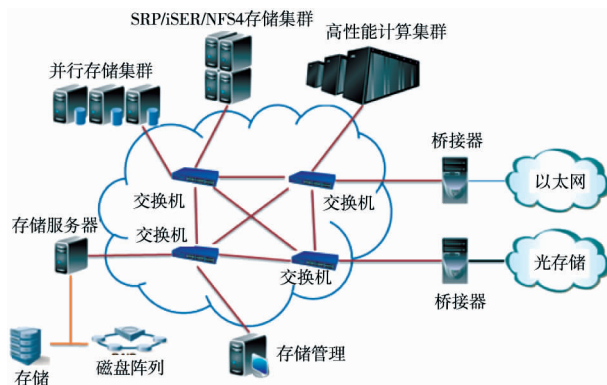


图 2 基于 InfiniBand 承载协议的云存储系统架构

远程直接内存访问的互联网小型计算机系统接口扩展 (iSCSI Extensions for RDMA, iSER) 存储集群、小型计算机系统接口远程直接内存访问协议 (SCSI RDMA protocol, SRP) 存储集群、网络文件系统 4.0 (Net File System 4.0, NFS4) 存储集群、并行存储集群 (Lustre, GFS 等)。存储网络管理设备用于监视 InfiniBand 云存储网络的运行状况, 记录网络故障信息, 管理员可以方便地查找存储服务器的故障磁盘。桥接服务器用于跨接 InfiniBand 协议与其他网络协议, 进行通信协议转换, 作为以太网或者其它类型网络的客户端访问 InfiniBand 云存储网络的桥梁。

2 InfiniBand 云存储的关键技术

2.1 SRP/iSER 与 RDMA 技术

SRP^[11,12]即 SCSI 远程内存访问协议,是 IB SAN 的一种协议,其主要作用是把 SCSI 协议的命令和数据通过 RDMA 的方式发送到 InfiniBand 网络上。iSER^[13,14]类似于 SRP,也是 IB SAN 的一种协议,其主要作用是把 iSCSI 协议的命令和数据通过 RDMA 的方式发送 InfiniBand 网络上,iSER 作为存储协议已被 IETF 标准化。

SRP 由 T10 技术委员会于 2002 年发表,其目标是使用 RDMA 协议完成主机到存储设备的块级访问。该标准草案定义了 SRP 的发起端、接收端在使用 RDMA 时的命令格式与数据格式,符合 InfiniBand 体系结构。iSER 已经成为 IETF 中 iSCSI 协议的正式组成部分,它将基于 TCP/IP 的 iSCSI 数据传输改为使用 InfiniBand RDMA 的方式实现。这两种存储协议都使用了 RDMA 技术。典型的存储网络协议栈如图 3 所示。

传统 TCP/IP 数据传输时会占用大量 CPU 资

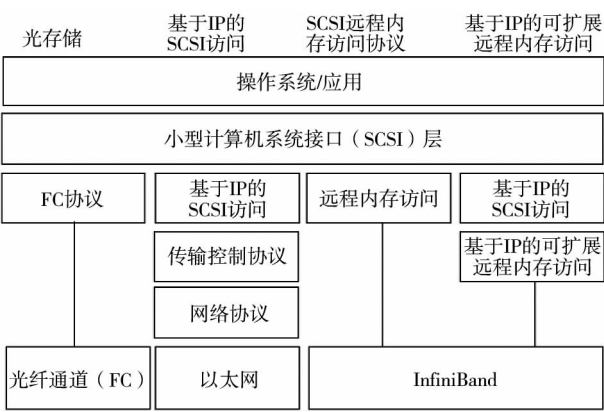


图3 典型存储网络协议栈

源。RDMA 技术通过网络把数据直接传入计算机的存储区(零拷贝技术^[15]),将数据从源数据系统快速移动到远程目标系统存储区,由于避免了操作系统拷贝过程,所以降低了对计算机处理能力的需求。为了达到这一目标,RDMA 技术将可靠传输协议固化于硬件,以支持绕过内核的零拷贝网络。

绕过内核技术使应用程序通过内核旁路向网卡发出指令,这就减少了处理网络通信时在内核空间和用户空间上下文切换的次数,如图4所示。

2.2 协议转换

为了增强 InfiniBand 云存储的适用性,方便以

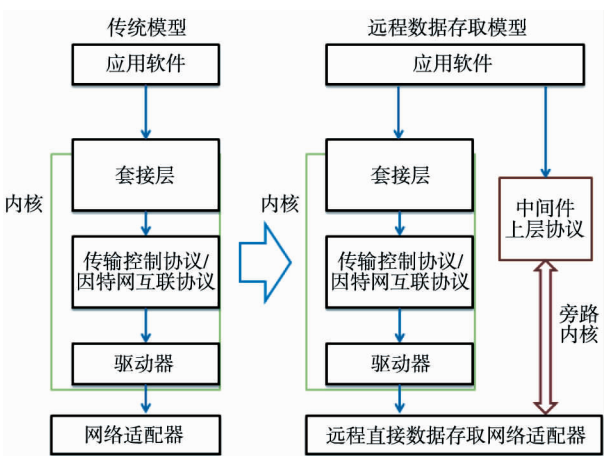


图4 传统模型与 RDMA 模型对比

太网或者其它网络类型的用户的访问,使用桥接服务器(Bridge)提供协议转换功能,实现跨协议的云存储服务。

存储服务器通过 SCSI 协议连接实际物理磁盘组。SRP Target 模块初始化,SRPT 模块获取系统中的 SCSI 存储设备信息并注册,以使模块具有处理 SCSI 命令的能力。InfiniBand 协议承载 SRP Target 及封装的 SCSI 信息进入 InfiniBand 网络,如图5所示。

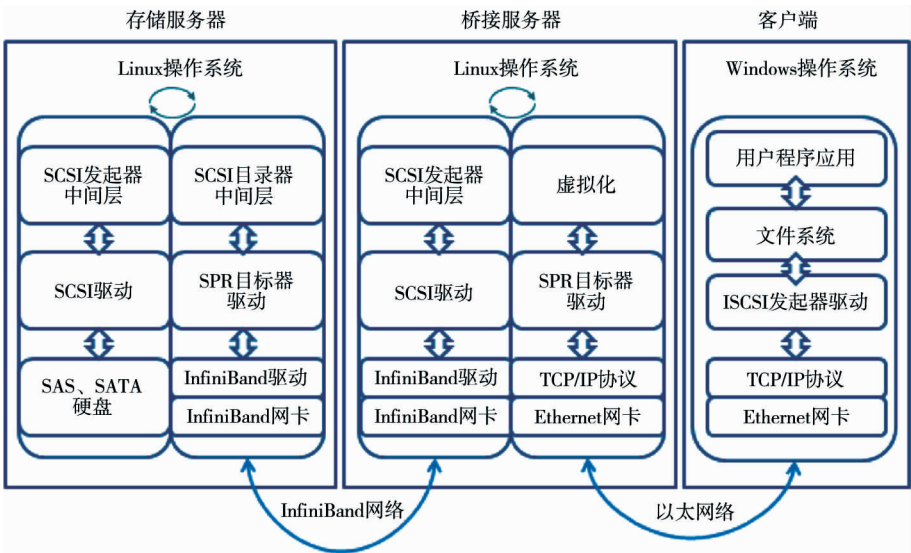


图5 协议转换的基本架构

桥接服务器通过 InfiniBand 网卡连接到 InfiniBand 网络和桥接服务器的 SRP initiator 发起与存储服务器之间的通信,并将存储服务器传输过来的 InfiniBand 数据包解开,获取 SRP Target 及封装的 SCSI 信息。桥接服务器在虚拟化模块中处理解析到

的 SCSI 信息,之后映射为新的 SCSI 存储信息并注册到 iSCSI Target 模块上,由 TCP/IP 协议承载 iSCSI Target 及封装的 SCSI 信息进入以太网。按照数据存储流程(图6),依次分为以下几项操作:

(1) 存储集群/存储服务器映射磁盘:存储集群/

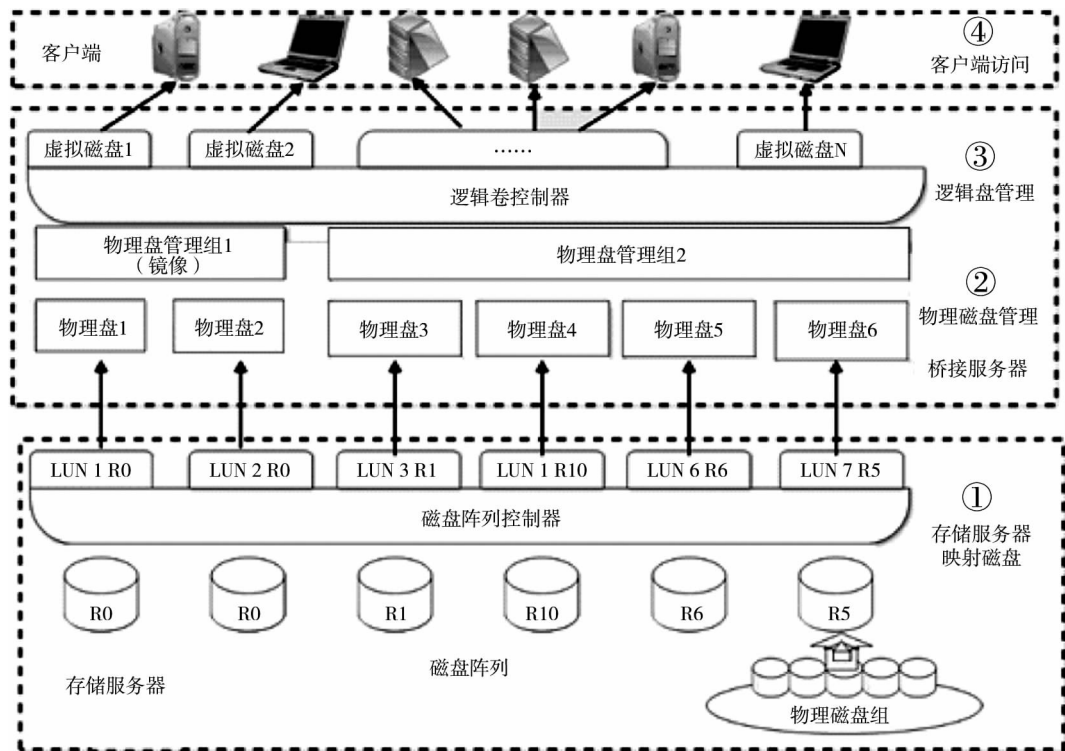


图 6 通过协议转换的数据存储流程

存储服务器使用独立磁盘冗余阵列(RAID)控制器整合多个存储单元,通过 SRP 协议、iSER 协议向 InfiniBand 网络映射磁盘 LUN(逻辑单元号)。

(2)桥接服务器物理磁盘管理:桥接服务器使用 InfiniBand 网卡接入 InfiniBand 网络。桥接服务器的磁盘控制器模块将存储服务器映射过来的磁盘 LUN 封装成系统识别的物理磁盘对象,多个物理磁盘对象可以放置在磁盘组中统一管理。

(3)桥接服务器逻辑卷管理:桥接服务器使用逻辑卷控制器,将多个磁盘组为用户划分多个虚拟磁盘 LUN,映射给客户端使用。

(4)客户端访问:客户端使用 iSCSI 访问桥接服务器的虚拟磁盘 LUN,实现数据存储。

3 测试实验

为了测试 InfiniBand 云存储网络在使用 SRP、iSER 和协议转换技术时的存储读写性能,分别进行两个测试实验:SRP 协议测试;协议转换测试。以太网技术的 iSCSI 协议的存储读写性能可以参照参考文献[16,17]中的测试结论,可以作为此测试实验的比较对象。

测试的硬件设备的主要指标见表 1。

表 1 硬件设备主要指标

类别	存储服务器	存储扩展单元	桥接服务器
型号	YQ-STD-H2000	YQ-STD-S45	YQ-STD-B2000
磁盘数量	2	45(满额)	2
处理器	Intel Xeon 5606	—	Intel 至强 5606
内存	24G	—	24G
主机接口	1 个 40Gbps IB 接口, 2 个 SAS 扩展口	4 个 SAS 扩展口	1 个 40Gbps IB 接口
IB 带宽	40Gbps	—	40Gbps
延时	< 100ns	—	< 100ns

3.1 SRP 协议性能测试

测试的对象为存储服务器 YQ-STD-H2000、存储扩展单元 YQ-STD-S45 和存储桥接服务器 YQ-STD-B2000,存储服务器和桥接服务器使用 InfiniBand 交换机的 40Gbps 端口互联,存储扩展单元使用 SAS 扩展端口连接到存储服务器,存储服务器和桥接服务器的操作系统都是 CentOS 6.2 版本的 Linux,网卡都是 Mellanox QDR 40Gbps 的 InfiniBand 网卡,测试工具为 iозone。如图 7 所示。

在桥接服务器上安装 iозone 测试软件,通过脚本文件执行 iозone 测试软件,设置吞吐量测试传输块大小分别为 4kB、8kB、16kB、32kB、64kB、128kB、



图 7 测试环境

256kB、512kB、1024kB, 测试文件大小设定为 1GB, 测试一个 SRP 客户端在访问 SRP 存储时实际的吞吐量。其中桥接服务器上配置 SRP initiator, 存储服务器上配置 SRP target。

结果如图 8 所示, 由该系统 SRP 吞吐量测试可知: (1) 数据块越大, 吞吐能力越强; (2) 写测试中, 随机写要强于顺序写; (3) 读测试中, 在数据块大于 64kB 的测试中, 随机读和顺序读能力均超过 2.5GB/s, 特别是在 128kB 之后, 读能力基本保持在 3GB/s 以上, 存储较好地利用了内存的缓存效应, 发挥了 InfiniBand 的带宽优势; (4) 基于以太网技术的 iSCSI 协议在万兆/千兆以太网下的读写速度在 400MB/s 和 100MB/s 左右, 因此基于 InfiniBand 协议的 SRP 优势明显。

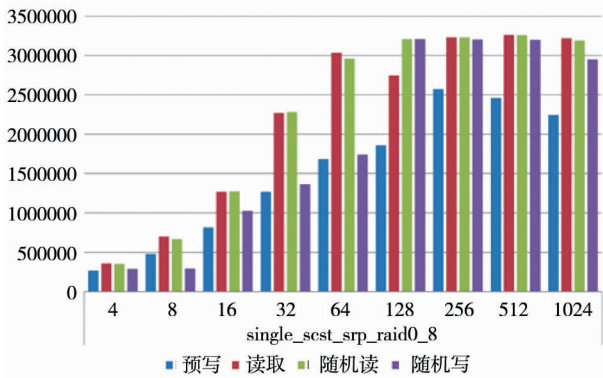


图 8 系统吞吐量测试结果

3.2 协议转换性能测试

协议转换性能测试选择万兆以太网与 InfiniBand 之间转换性能作为测试对象。其测试环境如图 9 所示。



图 9 协议转换测试环境

桥接服务器一端通过 IB Switch 连接 SRP 存储服务器, 一端通过 Ethernet Switch 连接测试服务器,

以太网端口速率为千兆。桥接服务器配置 SRP initiator 和 iSCSI target, 存储服务器上配置 SRP target, 测试服务器上配置 iSCSI initiator。在测试服务器上配置 iозone 脚本文件, 设置存储吞吐量测试的传输块大小分别为 4kB、8kB、16kB、32kB、64kB、128kB、256kB、512kB、1024kB, 测试文件大小设定为 100MB。

由协议转换性能测试结果(图 10)可知: (1) 随着数据块变大, 吞吐能力也随着变强; (2) 在数据块大于 64kB 的测试中, 顺序读写能力均超过 100MB/s, 随机读写能力也能保持在 100MB/s 左右; (3) 相对于 InfiniBand 下的 SRP 协议, 协议转换的数据传输效率要低很多, 一方面是桥接服务器的中转导致传输效率的略微降低, 一方面是 TCP/IP 协议下 iSCSI 的传输效率受制于千兆以太网的速率(千兆以太网的吞吐量约百兆字节)。

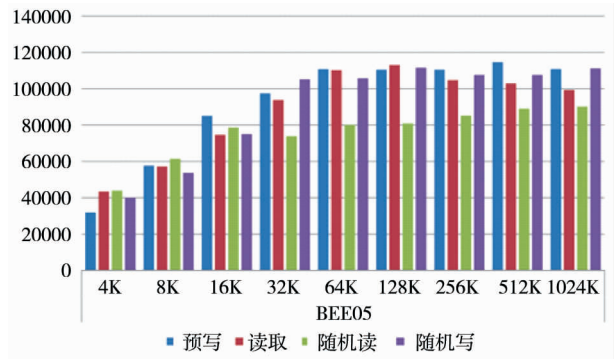


图 10 协议转换性能测试结果

4 结 论

随着云存储技术的发展, 云存储网络的性能也日益升高, 相应地对组网设备的性能要求也不断提高, InfiniBand 协议以其高带宽、低延时、低 CPU 占用等特性很好地适应了云存储技术的发展趋势, 具备多种优势的 InfiniBand 协议应该成为云存储的主流协议。本文通过实验取得了一些有意义的结论:

(1) InfiniBand 协议的特性适合于云存储网络的组建, 实验分析表明 InfiniBand 协议的确能够发挥出其技术优势。

(2) InfiniBand 协议承载其他通信协议的协议转换技术能够很好地解决 InfiniBand 网络的扩展性问题, 能够将更多的不同类型存储协议的存储设备加入到云存储网络中, 实验结果表明该方法可行。

参考文献

- [1] Leung A W, Shao M L, Bisson T, et al. Spyglass: fast, scalable metadata search for large-scale storage systems. In: Proceedings of the 7th USENIX Conference on File and Storage Technologies, San Francisco, USA, 2009. 9: 153-166
- [2] Sage W, Andrew L, Scott A B, et al. RADOS: a fast, scalable, and reliable storage service for Petabyte-scale storage clusters. In: Proceedings of the 2nd International Workshop on Petascale Data Storage, New York, USA, 2007. 35-44
- [3] Andrade A, Mun U, Chung D H, et al. Hybrid host/network topologies for massive storage clusters. In: The 24th IEEE Conference on IEEE, Washington DC, USA, 2007. 287-292
- [4] Weil S A, Brandt S A, Miller E L, et al. Ceph: a scalable, high-performance distributed file system. In: Proceedings of the 7th Conference on Operating Systems Design and Implementation, Seattle, USA, 2006. 307-320
- [5] Shanley T. InfiniBand Network Architecture. USA Boston: Addison-Wesley Professional, 2002
- [6] The InfiniBand Trade Association (IBTA). InfiniBand Architecture Specification. [http://www. InfiniBandta. org](http://www.InfiniBandta.org), 2012
- [7] Pentakalos O. An introduction to the infiniband architecture. In: International Computer Measurement Group Conference, Anaheim, USA, 2002. 425-432
- [8] 姜志颖, 韩冀中, 王锡贵. 基于 SRP 的 IB-SAN 设计与实现, 计算机应用研究, 2008, 25(4): 1168-1170
- [9] 盛月峰, 孙志刚. 基于 infiniband 的网络存储研究. 高性能计算技术, 2010, 1: 29-32
- [10] 邓里文, 余少华. InfiniBand 体系结构和协议规范. 光通信研究, 2001, 4: 4
- [11] InterNational Committee for Information Technology Standards. Information Technology - SCSI RDMA Protocol (SRP). USA Oregon: Cris Simpson Intel Corporation, 2002. 16-18
- [12] Mellanox Technologies: Mellanox OFED for Linux User Manual. Rev 2. 0-2. 0. 5. USA California: Mellanox Technologies, Ltd. , 2013
- [13] InfiniBand™ Trade Association. Supplement to InfiniBand Architecture Specification— Annex A12: Support for iSCSI Extensions for RDMA. 2006, 1. 1. 2
- [14] SCSI RDMA Protocol. [http://www. cisco. com/en/US/docs/server_nw_virtual/open_fabrics_enterprise_distribution/ofed_host_driver/release1. 1/user/guide](http://www.cisco.com/en/US/docs/server_nw_virtual/open_fabrics_enterprise_distribution/ofed_host_driver/release1.1/user/guide), 2012
- [15] 徐胜生, 初莹莹, 周敬利等. 基于 RDMA 协议的零拷贝技术研究. 计算机工程与应用, 2004, 3: 126-128
- [16] 张继征, 李栋, 郭劲等. 基于 iSCSI 的存储设备性能测试与分析. 计算机工程, 2006, 32(16): 264-266
- [17] 周敬利, 徐峰, 余胜生. Linux 下基于 iSCSI 存储系统的实现与性能评测. 计算机工程与科学, 2004, 26(6): 1-3

Research on key technologies of cloud storage based on InfiniBand

Yang Dongri, Wang Ying, Chu Hao

(College of Engineering and Information Technology, Graduate University of Chinese Academy of Sciences, Beijing 100049)

Abstract

In order to integrate the large number of heterogeneous services in networks respectively for computing, storage, and network switching to realize their effective coordination so as to jointly provide external data storage and service access functions, an InfiniBand host agreement based cloud storage architecture was presented. The connection of the storage devices with small computer system interface (SCSI) into InfiniBand networks was studied, and the key technologies relative to the cloud storage architecture, such as the SCSI RDMA (remote direct memory access) protocol (SRP), the Internet SCSI extensions for RDMA (iSER), and the protocol conversion, were analyzed and evaluated. The results show that the InfiniBand protocol is not only suitable for the construction of cloud storage networks, but also hosts protocol conversion technologies for other communication protocols, leading a good solution to the InfiniBand network scalability issue. Also, by using an experimental method, the performance of the InfiniBand SRP protocol, and the connectivity and conversion performance between the InfiniBand and the Ethernet were quantitatively assessed.

Keywords: big data, cloud storage, InfiniBand, protocol conversion, SRP