TCP/IP 协议对快速网络的作用分析

曹秉超 张 媛 都志辉
(清华大学计算机科学与技术系 北京100084)

An Analysis of the Influence of TCP/IP Protocol on Fast Network

CAO Bing-Chao ZHANG Yuan DU Zhi-Hui
(Department of Computer Science and Technology, Tsinghua University, Beijing 100084)
Email: caobingchao99@sina.com zhangyuan99@mails.tsinghua.edu.cn

Abstract Among the network devices used for high performance parallel computing, Myrinet and Gigabit-Ethernet are relatively more widely applied. The modification in architecture and communication method has brought to great improvement of network based on Myrinet and Gigabit Ethernet. In this paper, we analyze major features of TCP/IP protocol, point out the restrictions of TCP/IP protocol over high speed network and propose some methods to enhance the performance.

Keywords TCP/IP protocol, Myrinet, Gigabit Ethernet

1 引文

Myrinet[1]和 Gigabit-Ethernet[2]是当今世界上性能最好的两种可以运用于局域网的高性能并行计算的网络系统。它们以其高效、高速的传输特点在广泛的领域里得到了好评。但是，这两种同样是高端的网络产品，其实现的技术在很多方面却是大相径庭的。而在当今的网络协议中，TCP/IP 协议是当然的主角。无论是万维网(WWW)，还是一般的局域网(LAN)，我们都曾经，或者正在使用 TCP/IP 协议。

对于 TCP/IP 协议，Myrinet 和 Gigabit-Ethernet 有着不同选择。虽然对于 Myrinet 来说，用户能够直接挂接 TCP/IP 协议[3]，并且以此在 Myrinet 网络系统上毫不费力地运行所有使用 TCP/IP 协议的应用程序；但是目前在基于"Myrinet"的网络上已经鲜有继续使用 TCP/IP 协议的例子了。不过，对于 Gigabit-Ethernet，TCP/IP 到目前为止一直是其最为经常使用的协议。本文将要探讨以下问题：为什么现在的 Myrinet 网络系统很少采用 TCP/IP 协议的形式？为什么 Gigabit-Ethernet 没有抛弃 TCP/IP 协议？TCP/IP 对 Myrinet 和 Gigabit-Ethernet 究竟各有什么样的制约？

2 Myrinet 和 Gigabit-Ethernet 各自的结构及技术现状

千兆以太网是对10Mbps 和100Mbps 以太网标准的一个扩展，它提供了1000Mbps 的原始数据带宽，同时和现有的以太网保持完全兼容。而 Myrinet 则是一种完全不同于一般以太网的网络设备。它们二者之间的区别表现在：

在网线的结构上 ·Myrinet 采用了独特的双向传输技术[1]，无论是铜线还是光纤，Myrinet 的网络连接都由8根内芯，每个方向通道上各有9根。其中，8根是数据线，1根是控制线。以同步发送异步接收的全双工方式，Myrinet 的通信速率可达1.28Gbps，现在64位66M 的情况下可以达到2.56Gbps。

·Gigabit-Ethernet 能够有效地支持非屏蔽双绞线上的千兆以太网操作，并且以1000Mbps 的速率在多模、单模光纤或屏蔽平衡型铜缆上实现半双工或者全双工的操作。

在网卡的结构上 ·Gigabit-Ethernet 的网卡结构基本类似于我们现在使用的一般的以太网卡。

·Myrinet 的网卡则有其独特的结构。Myrinet 的专用网卡有自己的 CPU 和 SRAM 存储器。CPU 的可编程性很好，用户可以编写自己的控制程序 MCP(Myrinet Control Program)，这样，在 Myrinet 网卡的 LANai 芯片上运行 MCP，承担了大部分的通信处理工作，大大减轻了主机的通信开销，同时也使得通信的速度大有提高。

在通信机制上 ·Gigabit-Ethernet 使用802.3以太帧格式，在半双工模式下使用 CSMA/CD 介质访问控制方法，并在协议上增加一些新的特性，以解决在高速环境里与标准的以太网帧结构相一致的物理特性问题，比如载体扩展和分组猝发传输[9]。

·Myrinet 在路由机制上采用了一种名为"切通路由(Cut-Through)"的技术[3]。简单说来，当交换机收到一个数据包的包头后，立即查看其目的通道，而不用在交换机中缓存，如果目的通道已经被其他数据包占用，则此数据包被阻塞，直到目的通道空闲为止。切通路由机制比通常以太网所使用的"存储转发寻径方式"效率要高。也体现了"零拷贝(Zero-Copy)"这一思想。

当然，作为快速网络的两个代表，Myrinet 和 Gigabit-Ethernet 都完全兼容 TCP/IP 协议；而传统的 TCP/IP 协议也的确对这两种网络设备的性能有着一定程度上(有时甚至是相当大)的制约。

3 TCP/IP 协议对 Myrinet 性能的影响

目前 Myrinet 在计算机行业中的使用已经是很普遍了，据2002年6月公布的"TOP 500"中，使用 Myrinet 作为集群式

曹秉超 本科生，主要研究兴趣为并行计算与网络协议。张 媛 本科生，主要研究兴趣为并行计算语言与高性能计算模型。都志辉 副教授，博士后，研究方向为并行编译、并行算法及并行程序设计。

计算机系统网络的有16家之多[5]，如下表所示。但是，在这些“超级计算机(Super Computer)”中，在通信机制上，基本没有采用 TCP/IP 的，它们大多数运用的都是诸如“内存通道”、“虚拟内存映像通信”、“SP2用户空间协议”等避开操作系统方法。

表1 2002年6月 TOP500中使用 Myrinet 的集群计算机

排位	35	47	50	53	56	57	58	64
速度	825 Gflops	715.1 Gflops	706.7 Gflops	677.9 Gflops	654 Gflops	617.3 Gflops	594 Gflops	564 Gflops
排位	79	96	106	107	166	229	238	243
速度	526 Gflops	470.9 Gflops	442.7 Gflops	442.5 Gflops	285 Gflops	237 Gflops	232 Gflops	226 Gflops

原本对于 Myrinet 来说，TCP/IP 协议是完全能够兼容的。因为在 Myrinet 的专用的管理软件 GM 中的 Driver 部分上，用户能够直接挂接 TCP/IP 协议，这样就可在 Myrinet 网络系统上毫不费力地运行所有使用 TCP/IP 协议的应用程序了。但是，我们知道，TCP/IP 协议的开发和研究就是基于不可靠的网络的。也就是说，TCP/IP 协议所最适用的是不稳定的网络系统[6]。于是，为了克服网络的不稳定性，在 TCP/IP 协议中就有了相当多的层，用来确保数据包在传输的过程中的正确性。但是 Myrinet 是一种什么样的网络呢?Myrinet 在用电缆(cable)作为传输介质的情况下，在25米的距离之内，其传输的错误率及丢包率在10^{-15}范围内，可见 Myrinet 是一种可靠性相当强的网络系统。所以如果在 Myrinet 上的信息传输是基于 TCP/IP 协议的话，那么势必降低 Myrinet 的传输效率，其传输的延迟时间也会大大加长。Myrinet 的链路带宽为1.28Gbps，而 TCP/IP 协议层的带宽仅有42Mbps!有实验表明，在传输的数据中，以小数据包为主的情况下，Myrinet 的传输延迟甚至大于普通的10M 以太网!如果 Myrinet 的使用是基于 TCP/IP 协议，那么就会在如下部分产生瓶颈(实验和数据的测量是在基于 PCI 总线和以 Pentuim-Pro200为 CPU 的工作站机群上进行的)[7]。

表2 延迟比较

网络类型	TCP 往返延迟(us)	UDP 往返延迟(us)
普通 Ethernet	14	1146
Myrinet 网络	1406	1370

在 TCP/IP 协议中共有四个层面[8]：链路层(data-link)，网络层(network)，传输层(transport)和应用层(application)。这几个层面之间本身的关系就极为复杂，数据的打包操作和解包操作都要在这几个层面里进行，而且又涉及到大量的数据冗余拷贝(用户层到核心层的拷贝，备份拷贝，到网络接口的缓冲区的拷贝等等)。此外，从链路层到网络层，TCP/IP 协议都要进行差错控制；从链路层到应用层，都要进行连接的建立和释放；从网络层到应用层，都要进行协议处理的调度。所以，数据传输速度的下降就不难理解了。实验数据表明，在网络层，由于路由和打包解包的操作，数据传输的速率已经下降到了560Mbps。在接下来的传输层，因为 Multiplexing 操作，对丢失的数据包的“再传输”，以及 CRC 码的检查所造成的影响，数据传输的速率进一步下降为410Mbps。再接下来，在应用层和协议核心层之间的 Socket 层中，虽然这个部分是独立于协议的，但是这个层面负责的是用户空间和核心空间之间的数据复制，以及数据流的控制，这时候，8K 大小的数据包的传输速率，在 TCP/IP 协议下竟然只有250Mbps 了。我们看到，正是由于 TCP/IP 协议的冗余性，导致了 Myrinet 强大的性能在逐步削减之后只剩下了原来的十分之一。当然，这些数据，都是在模拟理想状态的情况下测得的，而且传输速率大都以大小为8K 的数据包为标准。在实际情况下，网络状态可能没有那么好，数据包的平均尺寸也远远小于8K。那么在实际情况下，基于 TCP/IP 协议下的 Myrinet 网络系统的性能还要低，传输速率一般只能达到数十兆，这种情况下，比之于100M 以太网的用户，Myrinet 的使用者基本上没有任何网络传输速度上的优势了[8]。

表3 传输性能的逐步下降

所在层面	传输速率
单纯 Myrinet 网络系统	2.56Gbps
网络(Network)层	0.56Gbps
传输(Transprot)层	0.41Gbps
应用(Application)层	0.25Gbps

4 TCP/IP 协议对 Gigabit-Ethernet 性能的影响

千兆位以太网支持使用交换技术时全双工工作模式下的网络连接，能在共享式网络中以半双工工作模式操作。以太网介质无关接口是 GMII，其发送和接收数据通路扩宽到了8位，而100兆以太网介质无关接口是 MII，只提供4位。另外，在局域网方面，千兆以太网还具有以下明显的优点：

从十兆、百兆以太网可以平滑过渡到千兆以太网。升级时，只需换用千兆位网卡和交换机等网络设备即可。

网络拓扑结构多样，能适应不同层次的需要。既可组建成为共享式网络，也可实现交换式的网络环境。

千兆位以太网支持一些新的协议和标准，支持 MPEG-2 等多媒体压缩功能，便利了多媒体数据对象的传输。

对于20世纪90年代出现的千兆位的 Gigabit-Ethernet，人们最初的想法是在上面使用旧的 TCP/IP 协议。但存在以下一些问题：

由于长距离线路是受延迟时间限制而非带宽限制的，因此当传输速度到达1Gb/s 时，双向传输延迟远远超过了将数据发送到光纤上所用的时间，进一步增加带宽也没有用。

TCP/IP 协议采用“回到第 n 个分组重发协议”会使具有大的带宽延迟乘积的线路性能变差。

通信速度提高很快，协议得到处理的时间比以前减少，因此必须使得协议更加简单。

由此我们可以看出，虽然使用 TCP/IP 协议的1000兆以太网比起以往的10兆以太网、100兆以太网在性能上有着比较大的提高，不过在 TCP/IP 协议的限制之下，有一些1000兆以太网的一些增加传输性能的潜力就难以发挥出来了。

5 Myrinet 和 Gigabit-Ethernet 提高性能的途径

通过我们上面分别对 TCP/IP 作用于 Myrinet 和 Gigabit-Ethernet 这两种网络的影响，我们大致了解了究竟 TCP/IP 协议是如何阻碍高性能网络速度提高的。目前通常的做法是彻底抛弃传统的 TCP/IP，另起炉灶实现诸如“VIA”等能够“避开操作系统(OS-bypass)”的通信机制，这当然是一条比较彻底的解决方法；但同时考虑到 TCP/IP 协议使用的广泛性，以及在一些情况下只能基于 TCP/IP 协议(比如，要在以太网上运行 Argonne National Laboratory 开发的 MPICH，就只能

建立在 TCP/IP 的基础上),对传统的 TCP/IP 协议进行精简和修改也不失为一条途径。

对于 Myrinet。

方法一:对现有协议的精简。

有一些学术科研机构已经通过各种办法,从不同的方面对用于 Myrinet 的 TCP/IP 协议进行了修改、完善和发展。清华大学计算机系和以色列的 Hebrew 大学计算机系等,把克服 TCP/IP 协议冗余性的重点放在了对 TCP/IP 协议的精简上。

清华大学计算机系提出了"基于 Myrinet 的用户精简 RCP(Reduced Communication Protocol)协议"。RCP 协议(如图1)是一个建立在 Myrinet 的 API 层上的用户空间协议,它在两个用户进程之间提供一个顺序的可靠的虚拟通道。它是一个平衡协议,发送方和接收方完全对等,无论是先执行发送例程,还是先执行接收例程,都会在同一点得到同步并交换数据。虽然 Myrinet 层所允许的数据包的最大尺寸为8K,但是 RCP 协议的传输的数据的长度只受可分配的内存空间的限制,数据的打包和解包都直接在用户提供的原缓冲区进行,这样就避免了像在 TCP/IP 协议中那样冗余的多次数据的拷贝。RCP 协议在清华大学计算机科学与技术系的并行工作站机群系统上进行了测试,回路延迟为200us,是原 TCP/IP 协议的1/7,应用程序可见带宽达168Mbps,是在相同条件下使用 TCP/IP 协议带宽的4倍。

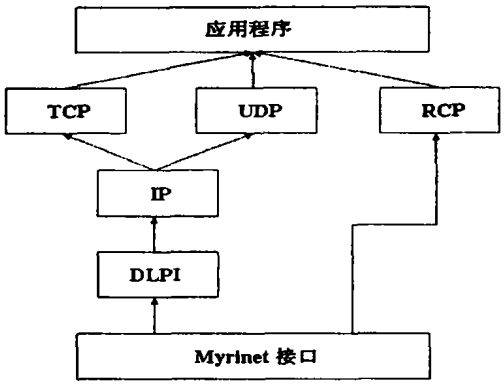

图1 清华大学 RCP 协议结构图

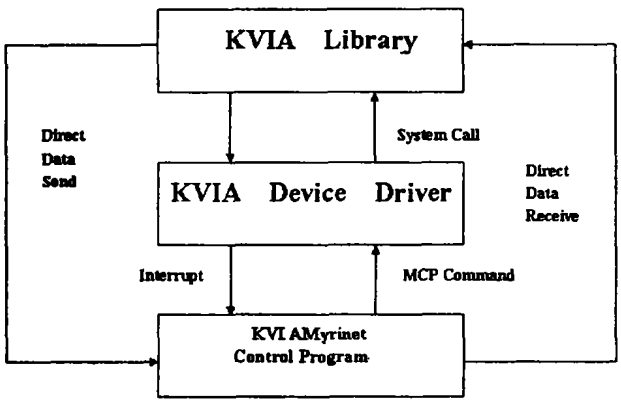

图2 韩国的 KVIA 通信标准结构

以色列的希伯莱大学(Hebrew University of Jerusalem)研制的 MNP 协议也是一种精简了的新型协议。在该协议中,原 TCP/IP 协议中的一些与数据包相关的操作被取消了,这样的操作主要有 checksum calculation 和数据备份拷贝。同时,MNP 保留了 Socket API 接口,并且能够和 TCP/IP 协议并存,支持所有 TCP/IP 协议支持的应用。在性能测试的结果上,NMP 的性能比 TCP/IP 高出了近一半。

方法二:完全放弃 TCP/IP 协议。

因为 Myrinet 本身并不对 TCP/IP 协议有着很大的依赖,所以在通过对其本身进行精简和修改的同时,我们也完全可以彻底地抛弃对 TCP/IP 协议的支持。这样反而更能解决一些本质性的、根源于 TCP/IP 协议内部的问题和缺陷。像早些年 UCBerkeley 的 AM(Active Message),UIUC 的 FM(Fast Message),还有稍后的 Princeton 的 SHRIMP(Scalable High Performance Really Inexpensive Multiple-Process),Cornell 的 U-Net,这些基于用户层的通信方式,都取得了相当好的效果。现有的比较成熟而且效果也比较好的一种 TCP/IP 的替代是"VIA(Virtual Interface Architecture)"。VIA 是一种基于用户层通信(user-level communication)的通信机制,主要由四部分构成:VI(虚拟接口)、VI Provider(虚拟通道提供者)、VI Consumer(虚拟通道消费者)、Completion Queue(完成队列)。它在传输数据的时候并不需要通过核心态,避免了在 TCP/IP 协议中由于上下文切换、内存拷贝、缓冲区管理所带来的巨大开销。

韩国高等科学与技术研究所的电子工程和计算机科学部开发研制的 Kaist VIA(KVIA)是一种高性能的 VIA 协议(结构如图2)[11]。KVIA 不同于普通的 VIA,它是基于描述子和通信数据的大小的,所以比普通的 VIA 有着更高的效率。该通信标准主要由函数库、器件驱动、Myrinet 控制程序这三部分构成。KVIA 用户函数库根据用户的请求调用系统函数,并且通过 MCP(Myrinet Control Program)直接执行数据的接收和发送。KVIA 的核心代理初始化内部的数据结构,同时通过 MCP 控制命令,来通知用户请求的 MCP。至于核心代理和 MCP 的通信,则是经由中断,或者是 I/O 映射缓冲区实现的。在 KVIA 中,虚拟通道位于主机的内存中,它能够起到降低 I/O 的负担,以及节省 Myrinet 内部的 SRAM 等作用。

在 CPU 为 Pentium-III 500MHz,内存为256M,Myrinet 的处理器为66MHz,64bit 的 LANai 芯片的配置下,KVIA 的"往返延迟(round-trip latency)"为40毫秒;"单向传输带宽(one-way bandwidth)"经测量为950Mbps,这已经达到了 Myrinet 系统理想情况下最高传输率的74%左右。

对于 Gigabit-Ethernet 要提高千兆以太网的性能,可以从硬件和软件角度分别考虑:

方向一:从硬件角度入手。

瑞典皇家科技学院提出了通过"协议预测"来改善千兆以太网的性能[12]。为了实现真正意义上的零拷贝,他们在网络适配器中加入了诸如"MCAM"(Match Content Addressable Memory)的硬件部件(如图3)。MCAM 是用户可编程的,具有很大的灵活性,利用它可以巧妙地避开 TCP/IP 协议中增加系统通信开销的部分。

值得一提的是,该方法只需要对操作系统的核心部分做少量修改,并不需要修改 TCP/IP 协议中链路层的内容。由于仅在传输层和以太 MAC 层间加入接口,实现对数据包流量的控制,因此易于实现。同时,MCAM 部件在 Warp 操作系统的集群计算机中也被大量地用于将分属不同协议族的消息检测分类。

方向二:从软件角度入手。

密西根大学提出了一种方案[13],试图通过编程实现一个

"自适应通信平滑器(如图4)"(Adaptive Traffic Smoother),使得集群系统对参加通信的实时和非实时数据包区别对待,即给予不同的响应速度,从而使系统的吞吐量和响应时间大大优于仅采用标准 TCP/IP 协议时的性能。它采用一个优先级队列,表征不同数据包的优先级大小、允许的最大传输延迟等。

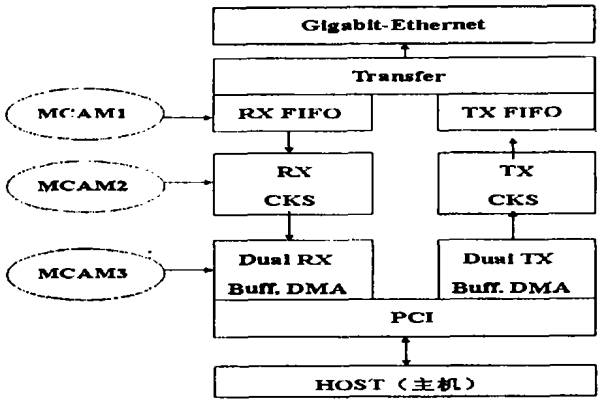

图3 MCAM 所处的位置结构图

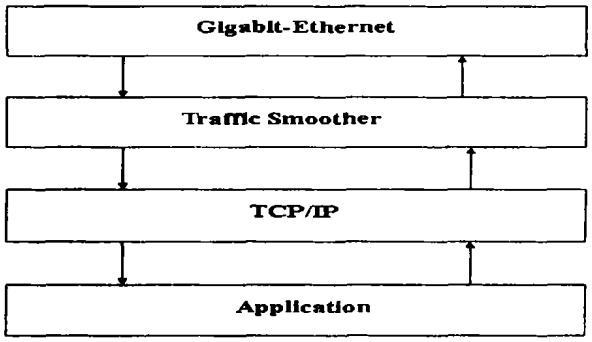

图4 密歇根大学的平滑器位置结构图

方法三:采用新的高速网络协议。

TCP/IP 这类传统网络协议过分依赖数据流发送方和接收方之间的反馈控制,在高速网络环境下,会出现控制信息大大滞后网络的动态变化的现象,甚至原有的控制信息已经不再适用或者会给网络造成更大的波动,这是网络协议控制所不希望的。无疑设计新的网络协议是解决传统网络协议功能和性能两个方面缺陷的彻底方案。快捷运输协议 XTP (Xpress Transport Protocol)作为新型网络协议进行研究的起点。XTP 协议本身仅仅是网络协议研究工作者在吸收传统运输层协议,比如 TCP/IP 协议和新型运输层协议(如 Stanford 的 VMTP,MIT 的 NETBLT 等)基础之上提出的一种通用的高速网络运输层协议。

面对网络协议,特别是运输层协议在高速网络环境和新型应用背景下功能和性能两个方面出现的问题,解决方案可以分为三类:改进传统网络协议机制[10],优化传统网络协议实现[15,16],以及设计新的网络协议[4]。这三种解决方案可以说是各有千秋,改进协议机制可以最大程度保证与原有协议的兼容性,保证平滑顺利的过渡,同时也能提高协议的性能;优化协议实现则不对传统协议作出大的改动,而是在协议实现方面利用现有的软件技术配合现有的硬件条件,从而能够发挥传统网络协议的潜在性能。

结论 传统的 TCP/IP 协议为连接管理、查错控制和流量控制提供了复杂的控制机制,它为大量应用提供了标准的点到点传输服务,因此被广泛接受,具有很高的市场占有率。但是它对于现代宽带高速通信网络来说,存在着一定的制约作用,在当今高可靠、高传输速率的网络产品已经相当成熟的背景下,TCP/IP 协议也迫切需要做出改进。在改进的过程中,我们要明确究竟是针对 Myrinet,还是 Gigabit-Ethernet。因为 Myrinet 有可以自编程的逻辑部件,所以,我们可以对 TCP/IP 协议本身做较大的改动,甚至可以彻底放弃 TCP/IP 协议;但是,对于 Gigabit-Ethernet 网络系统,其自身的结构性质决定了不可以对 TCP/IP 协议"动大手术",采用"协议预测"和"通信器平滑"是两种提高千兆以太网性能的方案。

参考文献

1 Boden N J, et al. Myrinet: Gigabit-per-Second Local Area Network. IEEE Micro, 1995, 14(1): 29～36

2 Mache J. An Assessment of Gigabit Ethernet as Cluster Interconnect. IEEE Computer Society International Workshop on, 1999. 36～42

3 Dubnicki C, et al. Myrinet communication. IEEE Micro, 1998, 18(1): 50～52

4 Cheriton D. VMTP as the transport layer for high- performance distributed systems. IEEE Communications Magazine, 1989, 27(6): 37～44

5 来自于 Myricom 公司的主页. Web 地址: http://www.myri.com/news/02620/index.html, July. 2002

6 郭庆平, 叶俊全. TCP/IP 协议结构分析. Computer And Communication, 1997, 14(2): 34～39

7 董春雷, 郑纬民. 基于 Myrinet 的用户空间精简协议. 软件学报, 1999, 10(3): 299～304

8 Barak A, Gilderman I, Metrik I. Performance of the Communication Layers of TCP/IP with the Myrinet Gigabit LAN. Elsevier Science B. V., Jun. 1999. 1～17

9 Gigabit Ethernet Alliance, Gigabit Ethernet Whitepaper. Web 地址: http://www.gigabit-ethernet.org 1999. 1214

10 Jocobson V. Congestion Avoidance and Control. ACM SIGCOMM88, Aug. 1988

11 Yu J-L, Lee M-S, Maeng S-R. An efficient implementation of Virtual Interface Architecture using adaptive transfer mechanism on Myrinet. Parallel and Distributed Systems, 2001. ICPADS 2001. In: Proc. Eighth Intl. Conf. on, 2001. 741～747

12 Kurmann C, Muller M, Rauch F Thomas M. Stricker Laboratory for Cumputer Systems Swiss Federal Institute of Technologh CH-8092 Zurich, Switzerla《Improving the Network Interfaces for Gigabit Ethernet in Clusters of PCs by Protocol Speculation》. Web 地址: http://www.inf.ethz.ch/

13 Seok-Kyu, Kweon, Kang G, Shin. Achieving Real-Time Communication over Ethernet with Adaptive Traffic Smoothing. Real-Time Technology and Applications Symposium, 2000. RTAS 2000. Proceedings. Sixth IEEE, 2000. 90～100