分布式并行服务器中的高性能通信研究

唐 续[1]，刘心松[2]，杨 峰[2]，刘 谐[2]

(1. 电子科技大学电子工程学院 成都 610054; 2. 电子科技大学计算机科学与工程学院 成都 610054)

【摘要】基于快速以太局网中分布式并行服务器通信特点和对网络协议栈的分析，设计并实现了一种分布式操作系统传输协议。该协议在维持通用编程接口的同时，精简网络协议栈底层的冗余操作，减少操作系统干预和协议自维护开销。在8个节点的服务器系统上的测试结果表明，分布式操作系统传输协议获得比用户数据报协议更小的往返延迟和系统开销，应用环境下有约20%的增益，能有力地支持系统扩展和大并发访问服务。

关 键 词 分布式; 并行; 通信协议; 性能

中图分类号 TP393 文献标识码 A

An Efficient Communication Protocol for Distributed Parallel Server

TANG Xu LIU Xin-song YANG Feng LIU Xie

(1. School of Electronic Engineering, UEST of China Chengdu 610054;

2. School of computer science and Engineering, UEST of china Chengdu 610054)

Abstract Aiming at the characteristics of the LAN of fast Ethernet and distributed parallel server system, a novel distributed operating system transport protocol (DOSTP) is proposed in this paper. The goal of this work is to study how to reduce the communication software overheads in the distributed parallel server system with off-shelf equipments, moreover to keep the common programming interfaces. Several measurements, such as decreasing redundant operations at underlying protocol stack, reducing OS interventions and the overheads of protocol self-maintenance, are used to improve the performance. Experiments have been done on the server of 8 PIII PC nodes and 100Mbps Ethernet with switches. Compared with UDP/IP, DOSTP decreases round-trip time of 20% for 64 bytes packets under the same application conditions.

Key words distributed; parallel; communication protocol; performance

分布式并行服务器系统以其高性价比[1]、高可用性、强扩展能力，广泛应用于通用网络服务。高性能网络通信机制被要求以支撑其整体的高性能。为此，国内外学者对通信软件协议进行了大量研究。研究热点在于通过旁路操作系统的通信机制而定制用户层的快速消息路径，并报告了较好的性能[2~4]。但该类方式通常需依靠Myrinet等高级网络接口技术的支持[5]，并且通用的编程接口和受保护的多用户网络并发访问等与性能增益的冲突难以解决。

本文的研究目标则是在一个对外提供TCP/IP协议族通信的通用网络服务系统中，提供服务器内部节点间的高效通信，以期提高整体的网络服务能力和系统可扩展性。

1 Linux通信性能分析

Liunx下基于以太网的网络通信中，通常采用分组交换的通信模式，可用如下简化公式其端端时延的通

收稿日期：2004 – 02 – 19
基金项目：四川省科技攻关项目(02GG006-018)
作者简介：唐 续(1975 –)，男，硕士，讲师，主要从事分布式并行系统，计算机网络方面的研究.

信软件性能模型表示为：$t = t_0 + t_c + m/r$。其中t_0称为启动时延，包括数据结构处理(查询、修改与互斥等)时间、协议维护(Address Resolution Protocol, ARP)操作、数据分段与重组等)时间、处理协议数据单元(Protocol Data Unit, PDU)本身的时间(字段解析、校验等)。t_c为数据拷贝时间，r为传输带宽据，m为消息长度。为此可以得出：对于短消息突发通信方式有$m<<r$，端端时延主要为数据拷贝时间t_c和启动时延t_0。而Linux通信数据拷贝次数已减少到传统受保护多用户通信模式的最少(用户空间→内核空间→网卡缓存)。所以，减小t_0是减小端端时延的主要方式。

2 设计原理与实现

分布式并行服务器系统的内部主要通信任务为4类：数据收集、数据发布、事务、实时传送。其各具有短时突发性、周期性和报文独立性等特点。为此，将分布式操作系统传输协议(Distributed Operationg System Transpon, DOSTP)设计为用户数据报类协议。经分析发现，Linux网络子系统采用数据抽象技术而具有开放性，新网络协议的开发只需面向抽象接口，就可专注于本协议专用操作的设计与实现[6]。如图1所示DOSTP注册为一类协议族AF_DOSTP，通过DOS_UDP操作子集向上挂接到BSD socket接口；通过向散列表ptype_base添加AF_DOSTP的数据包类型，以接收从网卡到来的本协议数据包。图2所示为DOSTP内部结构，其基于功能划分，强调以功能分割而不是协议层次来划分模块，使得协议函数具有简化的调用路径和参数传递。对于用户程序，只需替换创建socket时传入的参数AF_INET为AF_DOSTP就可完成原UDP/IP通信代码到本协议的移植。

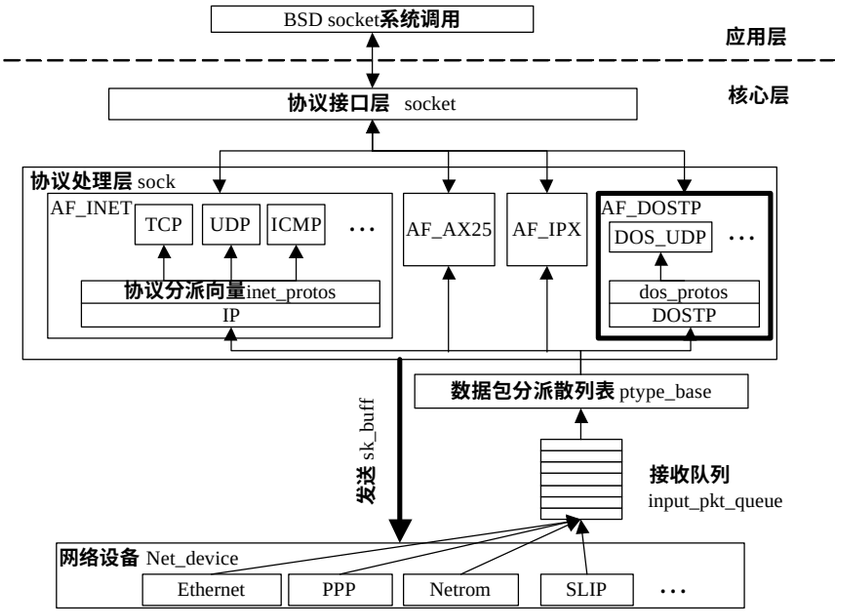

图1 Linux网络与DOSTP

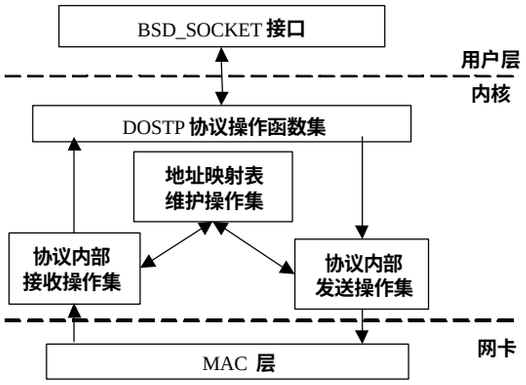

图2 DOSTP协议模块结构

2.1 简化寻址

服务器系统配置为单一子网，每个服务节点在系统内有单一IP地址，但32位IP地址的处理、存储和传输显然冗余。作为内部通信的DOSTP中改用16位的主机地址，其直接截取应用服务程序传入的IP地址的低16位进行处理和传输，接收方也只把还原的IP地址返回应用服务程序。应用程序对地址转换过程透明，使其能方便地在同一程序中交互使用两种协议栈，来处理系统内部以及系统与客户的不同数据。

2.2 数据报分段与重组

以太网中，有最大传输单元(Maxim Transport Unit, MTU)

的限制。如果数据报长度大于MTU，必须被按照MTU先分片后再发送。TCP/IP协议族中，这一功能由IP层提供，称为IP分片。对于用户数据报类型的协议，由于不能强制用户只使用小于MTU的数据报，很容易导致分片。

1) IP分片问题：运行于Internet环境中，IP数据报可能在发送主机乃至于在传送途中被路由器多次分片；分片可能因不同路由而错序或延迟到达；可能重复和丢失；目的主机上，属于不同数据报的分片也可能交错。为此，IP头部专门设置处理分片的字段。同样，在接收端重组分片报文并重新计算校验和，还要防止碎片缓存空间溢出，碎片等待队列、超时定时器等数据结构及相关操作被引入协议关键路径。IP分片成为最终通信用户额外且昂贵的开销。

2) DOSTP改进：在服务器系统中，各服务节点在同一子网内以交换机连接，报文有序传输。DOSTP中能简单采用标记加序号的分片方式。对于需分片的用户数据报，发送方操作只有分片序号的赋值和给头包/尾包设置First/Final标志。分片的报文顺序到达目的主机，并根据序号和标志直接进入碎片等待队列的队尾排队或新建队列，重组的操作得以简化。另外，各分片的报文将在极短的邻近时段到达，因此只需为分片等待队列设置很短的定时器超时值，如1 s(Linux内核实现中为30 s)，以此减小碎片缓存占用内核空间及其维护的开销。

2.3　协议头精简

分布式并行服务器系统中，通信报文长度通常较小，报文头部负载不可忽略。基于上文所述的改进，DOSTP有可能精简协议头。DOSTP协议头部字段的设置参考IP、UDP和TCP头部的基本字段。协议头部总共16字节，比UDP/IP的协议头28字节减小42%。传输效率得以提高，协议封装过程得以加快。

2.4　ARP开销

ARP协议用于解析通信双方网络地址到物理地址的映射，它是TCP/IP协议族中不可或缺的底层协议之一。该协议动态自维护一张地址映射表，其表项通过ARP报文交互而建立，对应当前直接可达的活动主机。Internet中的环境复杂，为了保证最新的地址映射信息，ARP表项将通过定时器而过时，然后引发新的ARP确认；闲置一段时间的表项将被清除。

基于上述原理分析并试验，发现在分布式并行服务器环境中，ARP协议及其实现存在如下的性能问题：

1) 在发起初次通信时解析对方主机地址的ARP报文交互才被触发。之后如果相关表项被清除，ARP报文解析又将再次进行。分布式并行服务器系统内80%以上的通信都是请求/响应或单向的发布和收集，常具有长间隙突发的特点。引入ARP将使这些通信过程附加一次报文往返，使通信延迟增加一倍以上。

2) Linux实现中，在非连接条件下，更新的ARP表项每30 s将过时。之后的通信报文会引发ARP重新确认。服务节点较多时，这将给网络引入较大的通信负载。如250个节点的系统中，任意节点间有间歇的网络通信，则平均约每0.96 ms将在网卡上各有一次ARP报文接收和发送。取系统中往返时延的典型值80 μs，可计算得ARP引入的网络系统负载达8.3%。可见，ARP开销对于系统可扩展性不可忽视。

针对上述问题，DOSTP做了两项改进：1) 系统启动时，各活动节点主动广播自己的地址映射信息，接收并保存其它节点的广播信息。通过这种系统预热的方式避免了数据通信过程中的ARP解析，缩短了通信的系统延迟；2) 系统运行中，地址映射表仍然动态维护，但使用更长的定时器时间(当前设为Linux实现的15倍)。这一简单处理可获得如下好处：1) 增大了地址映射表维护的时间间隔，系统操作对通信路径的阻碍减小；2) 增大了ARP的有效期，使ARP报文交互数减少，主机和网络的负载都得以降低；3) 同时也保持了节点间可达性动态自维护这一特点。

2.5　路　由

Linux网络系统中，路由操作函数处于网络的关键路径上。当中需考虑多地址、多路径、多路由策略的情况、需对多种目的/源地址的类型及其合法配合进行检查，还可能根据路由策略访问一到多张路由信息表。查表相关代码的执行对于协议处理来说相当费时。

以Hash链表组织的路由缓存能旁路路由信息表查找操作，相对缩短路由解析的时间。然而，每一路由缓存项是按主机地址而非网络地址维护，且为单向的访问信息(来回的路径为不同的缓存项)。对于大并发用户访问量的网络服务器(这里考虑短时间内将有更多交互的IP地址)，路由缓存项将会非常多。然而，Linux

系统所允许的缓存数量有限制(内核版本2.4.2中为256项)，路由缓存项在闲置一段时间后将被删除，过多的项也会按一定策略被删除。显然，路由缓存的命中率将随并发访问量的增加而降低。同时，新建的缓存项总在链表头，在应用中，新的客户IP不断引发新的缓存项，服务器系统节点间通信的缓存项将被压到各链表的最后，成为访问最慢和更可能被删除的项。此外，路由缓存作为进出网络路径中共享的数据结构，对其访问和动态维护须多次加/解锁，以维护数据一致性。这将影响Linux软中断系统的执行，阻碍协议栈中的数据流动。由于在同一子网内通信，DOSTP协议中合理精简了这部分路由处理，在关键路径中只保留了地址检查和广播报文分派等必须的简单操作，并建专用数据结构绑定和缓存通信双方的地址映射信息。以此缩短了DOSTP的通信路径，同时也实现了它与TCP/IP协议栈在处理逻辑上的完全隔离，如图1所示。

3 性能测试

3.1 测试环境

用于测试的服务器系统为8个节点，Intel PIII 1.13 G处理器、256 M内存、D-Link DEF-530TX网卡、10/100 M自适应交换机QS8224I。每节点安装DOSTP模块和Linux(版本2.4.2-2)提供的INET网络组件。

3.2 测试内容

1) 关键路径测试 如上所述，DOSTP由于在系统启动时就已完成地址解析而建立快速通道，其对比INET的UDP/IP在初次通信和突发通信时有明显的性能优势。表1所示仅给出通过乒乓程序获得的稳态通信对比测试。表中数据说明不同大小的报文，DOSTP对比UDP/IP都有一定程度的性能增益。由于UDP协议本身已很简化，同时，其在Linux实现中的也已相当优化，DOSTP在关键路径上的优势不显著。

2) 应用环境测试：测试中启动服务器的各类应用服务，并通过报文生成工具产生大量外部网络地址针对服务器的并发访问流。在此模拟的应用背景下，测试两节点间突发往返64字节数据报的平均时延。两种协议比较如表2所示。表中数据只具有相对意义(其值受当时主机和网络负载所影响)，但能够说明两协议在相同条件下的对比情况。并发访问前的数据表明，DOSTP在实际应用中表现更好，有约20%的增益。并发访问后的数据表明，在大并发流时，UDP/IP性能恶化严重，而DOSTP受到的影响很小。

表1 用户数据报往返延迟比较

数据/bytes	64	512	1 024	2 048	4 096	8 192	16k
UDP/IP	88	165	255	385	567	930	1 652
DOSTP/μs	76	152	240	368	547	903	1 610
增益/μs	12	13	15	17	20	27	42

表2 大并发流环境测试

协议	64字节数据报往返时间/μs	
	并发访问前	并发访问后
UDP/IP	约198	> 293
DOSTP	约165	173~179

4 结 束 语

分布式并行服务器系统往往需要高性能专用通信系统的支持。而应用服务商和上层应用开发者更关注系统软硬件的通用性。为此，本文研究了在通用构架下提高通信性能的可能性，深入挖掘通信底层中各种性能增长的可能措施，并提出了一种精简通信协议DOSTP。测试结果表明，其获得比传统协议更低的通信延迟和更小的网络负载；能有力地支持系统扩展和大并发访问服务。

参 考 文 献

[1] 刘心松. 具有分布式并行I/O接口的分布式并行服务器系统的性能研究[J]. 电子学报. 2002, 30(12): 1 806-1 810

[2] Gopalak rishnan R, Parulkar G M. Efficient user-space protocol implementations with QoS guarantees using real-time upcall[J]. IEEE/ACM Trans Networking, 1998, 6(4): 374-388

[3] 董春雷, 郑纬民. 基于Myrinet的用户空间精简协议[J]. 软件学报, 1999, 10(3): 300-303

[4] 周桂林, 戈 弋, 李三立, 等. 一种适用于机群系统的用户层消息传递机制[J]. 软件学报, 2001,12(5): 689-697

[5] Boden N J, Cohen D. Myrinet:A Gigabit-per-second local area network[J]. IEEE Micro,1995,15(1): 29-36

[6] 唐 续, 刘心松, 杨 峰. Linux网络协议栈分析及协议栈添加的实现[J]. 计算机科学, 2003, 30(2): 130-132

编 辑 孙晓丹