通信协议性能测量与分析①

孙忠义, 金同标, 殷进勇

(江苏自动化研究所, 连云港 222006)

摘 要: 为探索和证明透明进程间通信协议 TIPC(Transparent Inter Process Communication Protocol) [1]在进程间数据传输的优势, 首先简单介绍了透明通信协议 TIPC 的结构和优点, 然后在节点间和节点内分别对基于 TCP 协议和 TIPC 协议通信性能进行全面的测量, 并针对测量数据给出详细的分析. 所得出的结果不仅对并行程序设计中数据包大小的选取具有很强的指导意义, 而且对并行程序设计环境底层通信协议的选取也给出了科学的依据, 从而达到优化并行程序设计、提高应用程序执行效率的目的.

关键词: 透明进程间通信协议; TCP 协议; 通信性能测量; 数据包大小; 并行程序优化;

Performance Measurement and Analysis of Communication Protocols

SUN Zhong-Yi, JIN Tong-Biao, YIN Jin-Yong

(JiangSu Automation Research Institute, Lianyungan 222006, China)

Abstract: In order to explore and demonstrate the advantage of TIPC in communication between processes, the TIPC was briefly introduced first, then the communication performance of TCP and TIPC was measured in full-scale. In the same time, detailed analysis was applied to the measurement data. The result got here could instruct the message package size choosing in parallel program. what's more, the data got also indicate that the TIPC could get higher bandwidth than TCP. In this way Parallel program can be optimized and efficiency improved with communication in TIPC way.

Key words: TIPC Protocol; TCP Protocol; communication performance measurement;message package size; parallel program optimization

运行在机群系统上的应用程序由计算(Computation)和通讯(Communication)两部分组成. 目前最常用的应用程序性能评价模型[2]往往假定计算和通讯在时间上没有重叠(Overlap), 也就是说, 整个应用程序的运行时间是计算时间 Tcomputation 与通讯时间 Tcommunication 之和, 用公式可表示为 T =Tcomputation+Tcommunication. 而通信时间主要由消息启动时间 Tstart、消息包大小 TC、数据传输带宽 Tw 等多个参数决定, 而且每个参数之间又相互影响, 所以通过减少通信时间来优化并行程序设计是一项非常复杂的内容. 虽然目前已开发出许多面向机群的高性能用户级传输协议[3,4], 但一般都是针对特定的高性能计算机平台, 如中科院计算所开发的 BCL 系列用户级通信协议, 除了曙光系列计算机之外几乎没有其它的应用实例. 这主要是用户级通信协议脱离操作系统, 与具体平台相关度大、难以移植等缺点造成的. 另一方面, 现在的普通机群中运行的并行程序底层仍采用 TCP/IP 协议进行消息传递, 而经测量 TCP 在数据传输中开销很大[5]. 综上所述, 高性能的用户级传输协议难以普遍适用, 而广泛适用的 TCP 协议传输效率又很低, 严重影响机群中并行程序的执行效率. 提高机群通信性能需要一种高效易用的系统传输协议, TIPC 就是在这种情况下应运而生, 它是充分考虑普通机群环境而编写的系统级传输协议, 理论上在机群中可以取得更高

① 收稿时间:2012-02-04;收到修改稿时间:2012-03-27

的通信带宽. 但由于推广时间较晚, 现在很少在机群中使用 TIPC, 也一直没有可靠的数据证明 TIPC 传输速率比 TCP 高多少, 所以在工程中 TIPC 的价值还需要实验测量进行验证.

1 TIPC透明进程间通信协议简介

TIPC(Transparent Inter Process Communication Protocol)是由爱立信公司的 Maloy J. 专门为机群开发的处理较简单网络环境的通信协议, 在 2005 年完成, 现在由风河公司(Wind River Corp.)维护, 在 Linux2.6.30 以后的内核版本已经加入了这个通信协议. 在开发 TIPC 之前(实际上直至现在也是如此), 在普通机群节点间通信普遍使用 TCP/IP 协议, 但 TCP/IP 协议强大的功能主要体现在处理复杂网络中的点对点通信中, 并不非常适合普通的机群环境, 一方面 TCP/IP 协议中处理复杂网络环境的部分会带来较大的开销, 而这些开销在普通机群的简单网络中是完全没有必要的; 另一方面, 机群是一个随时可能变化的群体, 比如增加或减少一个节点, 进程可能会考虑负载平衡而迁移到另一个节点等, TCP/IP 并不具备处理这些变化的功能. 正是这种情况促进了 TIPC 协议的开发, 它能很好地处理普通机群中的各种变化.

TIPC 内部结构如图 1[6]所示, 此协议跨越了传输层和网络层, 所以节点间采用 TIPC 通信并不需要 IP 协议层, 在 TIPC 中采用一种特殊的 TIPC 网络地址来标识每个节点, 而通信中并不需要像 TCP/IP 协议族那样明确指出消息的目的节点, 因为在每个程序中都有一个端口来标识本进程, TIPC 可以通过这个端口号自动找到相应的节点, 从而实现通信的透明性. 而传统的 TCP/IP 协议族则要首先通过 IP 找到相应的节点, 然后在特定节点上找到相应的进程, 如果进程迁移到另外一个节点上, 通信就会出现连接错误, 所以这种通信是不透明的.

除了上面所提到的透明性, TIPC 的另外一些突出的优点如优化处理节点内消息和短消息的传递、简化连接建立与断开、支持真正的多播等. 本文以下内容将在相同的环境下针对各种消息类型对 TCP 和 TIPC 的通信性能做出全面的测量, 并针对结果给出深入的分析.

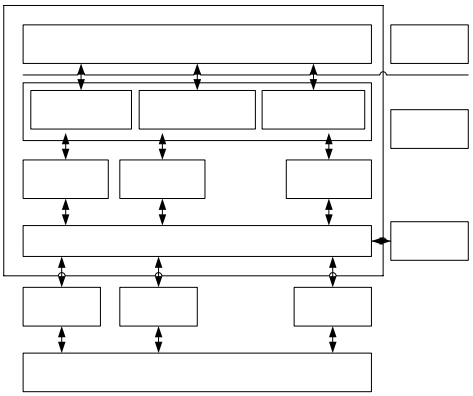

图 1 TIPC 内部结构图

2 通信协议性能的测试与分析

以下的全部测量是在曙光国产刀片服务器 TC2600 上完成的, 每个刀片上配置两个 Intel Xeon E5410 四核处理器, 主频 2.33GHz, 系统网卡为 Intel 1000Base-T. 为了更精确地反映出普通小型机群环境中的通信速率, 本测试采用典型的单一系统映像操作系统 Kerrighed-3.0.0, 它采用 Linux2.6.30 内核. 实验共使用 3 个节点, 一个主节点, 16G 内存, 两个从节点, 8G 内存, 所有节点间的距离在 2 米以内.

2.1 利用专用软件 IPERF 测量与分析

影响网络性能的因素很多, 不仅不同的网络硬件、通信协议会影响应用的通信性能, 同一环境相同通信协议下因采用的通信模式不同也会造成不同的通信性能, 如 TCP 通信中就包括最小延迟、最大吞吐量、最大可靠性、最小代价四种模式, 在不同模式中所获得的带宽也是不同的. 正是因为影响通信性能的因素太多, 所以很难提出一种通用测量网络协议性能的方法, 相应的软件也非常少. IPERF[7]是极少数测量通信协议性能应用软件中的一款, 它专门用于测试 TCP 和 UDP 的通信性能, 由于所面向应用的复杂性, 它的原理和测量结果也存在很大的片面性.

在 IPERF 客户端分别使用选项-l 和-w 选项控制发送消息 L 和套接字缓冲区 W 的大小, 用测得 TCP 协议的传输带宽绘制的曲线如图 2 所示.

图 2 中的五条曲线是所发送消息长度 L 分别为 1K、2K、4K、8K 和 16K 时所测得的, 为了易于插值得到平滑曲线, 坐标横轴刻度取套接字缓冲区 W 的对数, 也就是说第一个刻度 11 指的是 W=211=2K Byte,

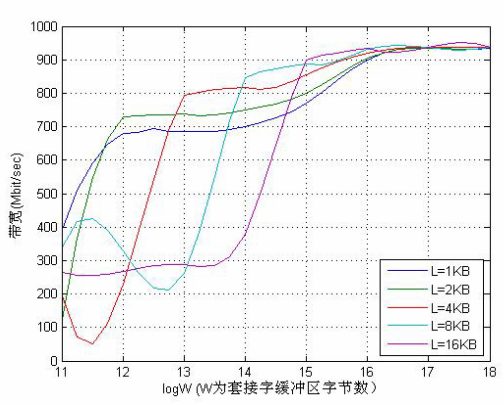

图 2　不同大小的消息在 TCP 协议下的传输带宽

而刻度 12 则为 4K 字节, 以下的坐标轴也是采用这种方式. 从坐标图中可以看出三个主要特点: 第一, 套接字缓冲区 W 越大, 消息的传递带宽越大, 这点在原理上也是显而易见的, 每次发送的数据多获得的带宽自然会增加. 第二, 套接字缓冲区 W 从小到大变化时, 发送缓冲区 L 的大小是一个转折点, 当 W<L 时, 消息传输带宽较小, 当 W=L 时消息传输带宽迅速达到一个比较大的值, 这是因为在 TCP 协议中, 要把发送缓冲区中的数据拷贝到套接字缓冲区后才能进行打包处理, 所以当 W<L 时, 每一个消息要分多次向套接字缓冲区中拷贝, 在网络中的传输层(TCP)中消耗了更多的时间, 所以带宽明显降低, 而当 W>L 时, 拷贝工作在网络中的传输层(指到套接字缓冲区的拷贝工作)可以一步完成, 从而从而缩短了协议的打包时间. 第三, 本实验所使用的是普通千兆网卡, 而试验中所测得的 TCP 协议最高传输带宽为 937Mbit/sec, 这还仅仅是数据带宽, 不包括消息首部和尾部, 按照工程经验来说, TCP 协议的传输带宽不会达到这么高. 经过深入了解 IPERF 源代码发现, IPERF 所测量的数据并不是模拟实际工程应用中所获得的带宽, 而是特定协议的极限带宽, 整个测试过程共用时 10s, 在这 10s 内仅仅建立一次连接, 建立连接后就按照指定的消息大小不停的发送数据, 直到 10s 后才断开连接连接, 这样 IPERF 没有考虑 TCP 收发过程中的握手信号的耗时, 而这部分时间恰恰是 TCP 协议中最耗时的部分.

从上面的解释得知, IPERF 所测得的数据的意义在于可以通过比较来指导根据不同的消息大小来选取合适的缓冲区, 而所测量的一个单独的数据没有实际的意义, 因为实际应用中远远达不到图 2 所示的带宽.

2.2 节点间通信性能测量与分析

为了测量实际应用中不同大小消息在 TCP 和 TIPC 协议下所能达到的带宽, 笔者做了以下测量: 在两个节点上做乒乓收发测试, 每一次连接建立后, 客户端和服务器各完成一次收发, 完成后即断开连接, 这样测得的数据是消息传输中双方的握手时间和数据传输时间的总和. 而所有的协议参数如滑动窗口的大小, 套接字缓冲区的大小都采用协议默认值. 对不同大小的数据连续发送 1e5 次所用的时间和达到的带宽分别如表 1 和图 3 所示.

表 1　TCP 和 TIPC 协议下数据包传输用时

数据(B) / 用时(s) / 协议	128	256	512	1024	2048	4096
TCP	26.0	26.8	27.8	30.0	37.5	43.3
TIPC	11.1	10.2	15.0	26.2	28.7	30.0

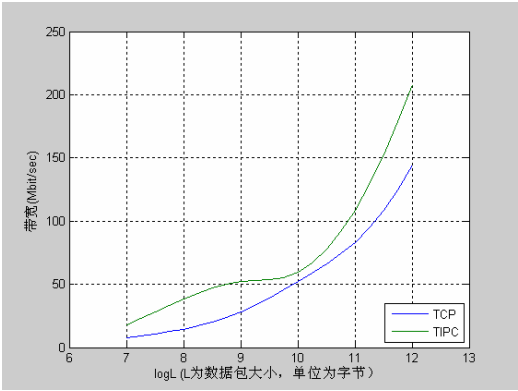

图 3　节点间 TCP 和 TIPC 协议下的传输带宽比较

本次仅测量了 4K 以下消息的传输时间, 因为在测量中发现系统中提供的网络接口 recv()每次最多仅能稳定接收 4344Byte 的数据, 而 4K 以上的数据在普通小型机群中是很少采用一次性发送的. 在确实需要发送大数据包中, 一般的程序设计思想是用 recv 的返回值与实际待发送的数据包进行比较, 如果发送的字节数不够则要用循环来处理, 直到发送的数据包大小合适为止. 从图 3 中可以看出 TIPC 的传输带宽明显比 TCP 要宽, 最少多出 15 个百分点, 而最多是 TCP 的两倍多. 同样是面向连接的数据流传输, TIPC 带宽更宽是因为 TCP 协议是面向复杂的 Internet 网络环境的, 它对消息的处理要比 TIPC 要复杂得多, 所以 TIPC 对消息的处理要简单的多. 例如每一次消息传输, TCP

协议在收发双方中至少要需要完成七次握手, 而 TIPC 协议只需要两次, 这也就是为什么数据包越小 TIPC 协议的传输带宽优势越大. 另一方面, 从图中可以看出两个协议的传输带宽都随着消息的增大而提高, 一般情况下是这样, 但也不尽然如此, 比如 TCP 消息包最大为 65536 字节(包括消息首部), 因为在 TCP 协议中, TCP 消息包的大小由 16 位数据表示. 大于 16K 字节的 TCP 消息要经过拆分成多个消息, 然后再接收端重新组装, 需要消耗很大一部分时间. 所以这时的传输带宽不仅不会继续增长, 反而会降低. 而 TIPC 协议规定每个消息包的数据上限为 66000 字节, 比 66000 更大的消息包是无法在一次系统调用中完成的.

2.3 节点内通信性能测试与分析

在实际的并行计算中除了节点间的消息传输外, 还存在单个节点内的消息传输, 特别是在多核节点中这种情况更为普遍. 在单个节点上的两个进程做上节所述的消息传输测量, 测量的结果如表 2 和图 4 所示.

表 2 节点内 TCP 和 TIPC 协议下数据包传输用时

用时(s) / 数据(B) / 协议	128	256	512	1024	2048	4096
TCP	8.2	8.3	8.6	9.3	9.7	10.2
TIPC	3.3	3.3	3.5	3.7	4.1	4.3

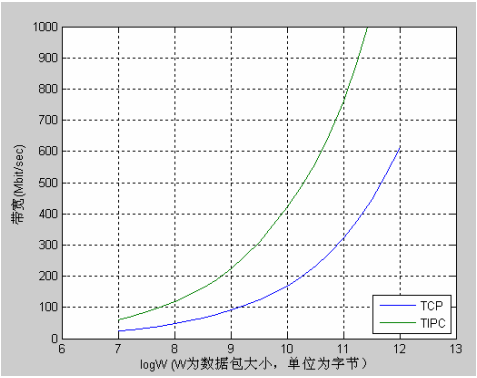

图 4 单节点 TCP 和 TIPC 协议下的传输带宽比较

从图 4 中可以看出单个节点的数据传输带宽比节点间的传输带宽要宽得多, 在千兆网卡上甚至能突破 1000Mbit/sec 的数据传输速率. 比起 TCP 协议, TIPC 协议在单节点上的数据传输优势更加明显, 比 TCP 的传输速率高出一倍以上. 另一点需要说明的是, 单节点上的进程间消息传输通过管道、共享内存等方式会更快, 之所以仍然要考虑使用网络协议进行消息传输是因为有时候机群系统会因负载均衡等原因自动实现进程迁移, 这样就不能确定两个进程是否在一个节点上, 或者说两个进程在一个时刻在同一个节点上, 在另一个时刻却有可能在两个节点上.

3 结语

从以上测量数据和分析可以得出在机群并行程序通信中, TIPC 协议具有比 TCP 协议更高的带宽, 尤其对节点间小消息和节点内的数据传输带宽优势更为明显. 如果当前的并行程序设计环境如 MPICH2、OPENMPI 和 PVM 等, 底层的通信协议中采用 TIPC 协议进行数据传输, 那么就会使得应用程序获得更高的执行效率, 这也是提高并行程序执行效率的一种行之有效的方法

参考文献

1 Maloy J. TIPC: Transparent Inter Process Communication Protocol. 2010.[2011-12-20] http://tipc.sourceforge.net/

2 李贵明,俞国扬,罗家融.基于 Linux 的 Beouwulf 集群的实现.计算机工程,2003,29(11):49–51.

3 Eiken V, Basu A, Buch V, Vogels W. U-Net: A User-level Network Interface for Parallel and Distributed Computing. the 15th ACM Symposium on Operating Systems Principles. 1995. Copper Mountain, Colorado. 1995. 1–14.

4 Chen Y, Jiao ZQ, Xie J, Du ZH. A Design and Implementation of High Performance Virtual Interface Architecture. Proc.of IEEE International Conference on Cluster Computing. 2003. Springer-Verlag Berlin Heidelberg. 2003. 125–135.

5 虞岩松.机群环境中高效 socket 研究[硕士学位论文].北京:中国科学院计算技术研究所,2005.

6 冀映辉,蔡炜,蔡惠智.TIPC 透明进程间通信协议研究和应用.计算机系统应用,2010,19(3):76–79.

7 http://iperf.sourceforge.net/ [br] [2011-12-15].