高速网络 运输层协议 网络协议

计算机科学1998 Vol. 25 №. 1

18-27

高速网络新型运输层协议研究概述*)

Survey on Transport Protocols for High-speed Computer Networks

潘建平 顾冠群 沈苏彬

(东南大学计算机系 南京210096)

TP393

摘 要 With the analysis on the problems for traditional transport protocols in the new environment of high-speed networks and novel applications, we have learned some experiences and lessons for revision and optimization on traditional protocols, and also realized the significance of new protocol design. After that, we emphasize our focus on the research and comparison of the new protocol families to demonstrate their flexibility and efficiency under the new environment; and also present the methodology, components and achievements of our research work. The international standardization process on transport protocols for high-speed computer networks ends the entire paper.

关键词 Computer networks, Network protocol, Transport protocol, OSI/TP, TCP/IP, XTP

1 背景和动机

传统运输层协议(如 DoD 的 TCP/IP 和 ISO 的 OSI/TP)适应的网络环境和面向的应用背景在最近十年里发生了巨大变化。光纤传输技术及相应媒体访问控制的发展,使网络环境由原来的低带宽高差错转变为现在的高带宽低差错;高性能多媒体计算机及相关技术的出现,将信息共享的传统应用拓展到知识共享的新型应用[1,2],相应的通信需求变得日益复杂。因此,传统运输层协议既没有满足新型应用需要的功能多样和性能灵活,又没有反映高速网络提供的新型传输服务,从而证实了高速网络运输层协议研究的必要性。

许多研究者从多个角度提出各种解决方案,较具理论意义和实践价值的有[4]:协议机制的改进;实现的优化;研究、设计和标准化新的协议。我们简单介绍前两者方案,并说明改进和优化通常只能弥补协议的性能缺陷,性能提高还受限于协议机制本身。总结协议设计实现几十年的经验教训,结合现有的协议工程和软件工程技术,完全能重新设计和实现更好适应高速网络环境和面向新型应用背景的新型协议。这不仅代表我们在内的许多研究者的观点,也得到国际性标准化机构(如 ISO/IEC JTC1和 ITU-T)的关注和认同。

我们选取快捷运输协议[5](Xpress Transport Protocol, XTP)作为高速网络协议研究的出发点和核心内容。另外我们还分别研究运输层协议与新型应用(如多媒体计算机会议)及与高速传输服务(如 ATM)的接口。XTP 是学术界提出的新型运输协议,OSI 和 Internet 的下一代运输层协议都透出 XTP 的影子,在工业界也很有影响。简单描述 XTP 协议功能和性能特点之后,介绍我们研究 XTP 协议的方法、路线和所获成果,最后简述新型运输层协议的国际标准化近况。

2 传统运输层协议面临的问题

2.1 新型应用通信需求

2.1.1 分布多媒体应用。异步多媒体电子邮件和同步计算机会议[3]是典型的分布多媒体应用。前者覆盖各种类型的媒体数据,在发送方构造、编码和压缩之后夹带在电子邮件进行传输,在接收方相应解压,解码和重现;后者通常具有音频、视频通道和白板、指针等共享工具,不同地理范围的用户通过网

*)国家自然科学基金,国家863高科技计划和国家教委博士点基金资助项目。潘建平 博士,研究方向为高速、高性能网络及协议。顾冠群 教授,博士导师,研究领域涉及计算机网络,分布式处理和 CIMS 等。沈苏彬 博士,副教授,主要研究计算机网络及应用。

络和计算机的支持进行协同工作。多点通信是计算机会议的主要特征，用户数和网络跨度使得多点通信效率极大地影响整体性能。传统协议的连接双方维持两个相同的数据流，任意一方的离开就意味着整个连接结束，连接依靠后向反馈和同步握手维护；多点通信具有多个资格能够动态调整的成员（迟后加入或提前离开），需要维持多个交错的数据流，全局同步相当困难，过多反馈将阻塞发送方，数据交替也使得传统的正确和顺序语义不再适用。

分布多媒体涉及文本、图形图案、静态图像、音频和视频等离散和连续媒体，特定媒体还能继续细分，如音质不同的语音、CD和高保真音源及不同时间或空间解析度的视频信息，媒体的自然性质不同，相应的通信需求也各异。如文本传输，特别是分布操作系统内存调度，要求百分之百准确；图形图案和静态图像依据编码格式和质量因子不同，能容忍特定类型的差错；连续媒体具有相当的时间和空间冗余性，可较大地容忍数据丢失。媒体特性决定文本和图形的数据量较小，静态图像和音频较为适中，活动视频较为巨大。另外离散媒体能容忍较大的传输延时和抖动，连续媒体通常涉及人的交互对延时有着严格的要求；特别是语音为维持其可理解性，对延时抖动相当敏感。总之分布多媒体应用有着复杂的质量需求，在性能方面表现为吞吐、差错、延时等的多样性，并要求网络协议以指定的语义予以保证。

2.1.2 分布式事务处理。分布式处理，特别是事务处理，具有连接短数据少的特点[2]。事务请求的结果返回以后，整个连接就结束，通信需求为低延时高吞吐，通信模式可分为单向或双向，有应答或无应答，单点或多点传输。传统协议仅支持双向的点到点模式，对单向事务显得多余，对多点事务则显得支持不足。特别是传统协议依靠多次握手建立连接之后才能开始数据传输；之后还需多次握手才能释放连接，连接代价甚至超过数据传输本身。此外，传统运输协议仅支持完全正确和顺序的数据投递，这种投递语义无法适合不同可靠性需求的事务处理。

总之，传统运输层协议在功能和性能两个方面没有支持多点投递，缺乏相应的群组管理，没有适应多样的通信需求的服务质量控制；性能方面的缺陷涉及固定的连接管理、差错控制和流量控制等。

2.2 高速网络传输服务

以光导纤维为传输媒体的高速网络，具有高带宽和低差错的特征[4]。对于网络协议高带宽意味着协议处理和控制决策必须更加迅速；光速限制使得带宽和延时的乘积变得很大，就是数据注入能力增强，这对强烈依赖反馈控制的传统协议是不利的。控制信息生成到参与决策的间隔会有更多的数据注入网络，使得控制原有的正确性和适用性不再成立。此外，网络数据通道将远远超过端系统和中继系统的缓存能力，传统协议对于性能和效率的折衷就不再成立。低差错指的是光纤网络的物理差错将大大减小，对于网络协议差错主要来源于端系统或中继系统由于缓存容量和协议处理限制造成数据丢失。这不仅破坏传统协议的某些假设，还影响传统协议的许多控制策略。如传统协议依赖数据重发进行差错恢复，但对高速网络差错率的提高往往反映网络某些区域处于拥挤状况，盲目的数据重发不仅不能完成原有的控制目的，反而会造成更大的网络拥挤和更多的数据丢失，甚至会造成反馈控制的振荡。

除了传统网络的异步传输，高速网络还有延时有界的同步和延时固定的等时传输以及多点投递和广播投递能力。出于最初设计的原因，传统协议没有也不可能反映新型传输服务，然而这又是新型应用所迫切需要的，必然迫使高层协议低效地构造和模拟所需的服务。

2.3 新型的网络层协议

TCP依赖的IPv4仅支持数据分组的单点投递，S. Deering提出IPv4多点投递主机扩展，占用D类IP地址作为多点投递群组标识，也定义D类地址直接映射到IEEE 802物理地址的方法。IPv4多点投递是基于接收方的协议，发送方无法了解多点投递组的具体成员，接收方自由加入和离开多点投递组。IGMP协议提供简单的组管理，如成员资格的探询和汇报等。

IPv4仅具有寻址路由和分段合段的功能，提供没有质量保证的投递服务。Internet IETF正在致力于资源预留协议RSVP的研究和标准化工作，使得运输层协议通过资源预留的方法保证所需的投递质量。RSVP是基于接收方的协议，能更好地适应多点投递成员的异构特性，还允许接收方在不同的多点投递通道之间进行切换。RSVP的PATH和RESV报文建立初始路径和资源预留，沿途路由器的软状态通过定时刷新以适应路由和群组的动态变化。

1995年12月IETF正式发布下一代IP（IPv6）协议规范，相关的寻址和路由协议及对具体网络子层的映射也随后公布。与IPv4相比IPv6能够适应高速ATM到低速无线网络，更重要的是IPv6具有极大的地址空间，全新的寻址方式，简化的报文格式，多样的选项设置，支持多点投递和具有流描述规范以提供具有服务质量要求的分组数据投递服务。可以看出，传统运输层协议依赖的网络层也在发生变化以适应高速传输服务和新型网络应用的发展。

3 可能的解决方案

3.1 传统协议机制的改进

3.1.1 DoD的TCP协议。TCP是面向连接的运输协议，提供完全可靠的顺序字节流服务。TCP通过多次握手建立及释放含有两个相向数据流的点到点连接。TCP固定地使用覆盖整个分组和伪报头的检查和进行差错检测；使用每个数据字节的序列编号检测分组重复、丢失和错序。TCP没有显式的差错通知，接收方仅累积返回分组应答；发送方根据定时器超时认为差错发生，后退到最后确认分组开始重发。TCP采用可变滑动窗口避免发送方的数据溢出接收方的缓存，接收方对分组的应答前移窗口下界和调整窗口大小。握手建立连接的方式将会降低分布式处理效率；固定的差错策略无法适应分布多媒体多样的通信需求。此外，TCP依赖的反馈式流量控制，及定时器触发的后退重发不适合高速网络。TCP也没能反映多点投递、同步和等时的新型传输能力。从TCP诞生起研究者就对其进行各种改进，这里列出的只是其中较具有影响的几个。

Jocobson Slow-Start 拥挤避免机制[6] 这种机制使用应答时间和丢失率估计分组是否经过网络的拥挤区域。如果发送方发现应答时间（实际上是测试的样本RTT值）增大，丢失率上升，就主动减小滑动窗口大小，限制发送量以期待能减缓拥挤。等到RTT和丢失率恢复，就逐步恢复原有的窗口大小。拥挤避免的目的是通过主动流量控制的方法部分实现速率控制。

大带宽延时乘积网络扩展[7] 大带宽延时乘积扩展有：使用窗口调节因子，使得原来16位窗口能表示64位字节空间；采用选择应答方法，发送方可以显式要求接收方返回指定分组的应答以降低对反馈的依赖；加入回应选项，能够更加精确估计RTT值。

大窗口和否定应答[8] 通过将窗口值从绝对值改为相对值的方法，TCP的窗口空间可以从16位扩展到32位，将能够提高TCP协议的吞吐率。另外，反向否定应答使得接收方及时返回差错信息，发送方能够提前准确获悉数据的差错发生，而不必过渡依赖定时器的溢出。

3.1.2 ISO的OSI/TP4协议。OSI/TP4依据DoD TCP为模型设计，两者在许多方面类似，当然也有相异之处。如TP具有不同类型的变长TPDU；在握手同步建立连接时，交换的信息有连接建立、数据传输、连接释放的延时（吞吐）和失败概率等。OSI认为完整的连接释放在会话层完成，TP4仅提供断开连接。此外，TP4能选择是否使用检查和，序列编号和流量控制基于TPDU而非字节编号。ISO在1982年发布OSI/TP草案以后，也在进行TP的改进工作，如1986年版，1988年版和1992年版及其它大量的增补、附录和修正等。对于TP4的重要修改有连接管理和差错控制，及请求应答和选择应答等。接收方获得发送方的请求应答后立即返回数据接收和应答信息。选择应答可以使得发送方更加详细地了解接收方的数据接收状态以提高差错恢复的效率。

3.2 传统协议实现的优化

3.2.1 通用的协议机制优化。一是**减少冗余的协议操作**。传统协议体系经常在不同层次出现相同或类似的协议功能，因此通过层次移动和相互组合可以简化整个协议栈的操作。如ISO/OSI协议栈的差错控制出现在数据链路层到运输层，复用解除、流量控制出现在数据链路层到应用层，分段重组出现在网络层到应用层等。此外传统协议没有很好地区分数据传输和控制传输，因此端系统协议处理变得更为复杂。如果能够尽早区别不同的数据和控制信息，协议处理的效率将会提高。降低协议控制的强度，如减少对于接收数据的应答频率，也会有利于数据吞吐的提高。

二是**增强必要的协议操作**。减少冗余操作能提高协议性能，但更重要的是增强必要的操作，如状态记录的检索及定时器和数据缓存的维护是必不可少的。如使用高速缓存保存最近使用状态记录，配合分组报头的预测，可以显著降低检索的负担；更进一步通过键标识减小检索空间也能加快检索速度。定时器的设置、运转、报警和撤除对任何操作系统都是相当的负担。为此可以降低定时器使用数量和频度，如每个连接使用单个定时器，以及通过硬件方法维护定时器以降低系统开销等。数据管道的增大使端系统缓存管理更为重要，更多的数据需留在缓存等待接收方的应答。避免过多的数据拷贝、移动和比较操作能够显著提高协议的性能。显然记录检索和定时器维护是对每个分组的优化，而缓存则是对每个数据字节的优化。

3.2.2 专门的协议机制优化。其它的实现优化集中在协议并行处理，专用多处理机和硬件硅编译等方面。经验表明，并行处理将是最终大规模提高协议效率的必经之路，然而这种潜力受到协议机制和所处环境的限制。许多研究者对协议并行化提出各自的方法，如协议层次之间、层次内部、协议结构、功能模块等不同的水平、垂直以及混合划分方式。在并行粒度、处理负担和可达比率方面，不同研究者存在各异的见解，这依赖所处环境的进程结构、事件管理以及内存映射等因素。有的研究者建议专用处理机实现网络协议，与主机结构和操作系统脱离。协议并

行化还导致在多个专用处理机并行实现的研究。通过硬件软件功能转移和分配，及专用大规模集成电路技术的发展，还有研究者考虑通过硬件芯片实现高层次协议，如硬件硅编译技术等；但硬件实现也对传统网络复杂固定的协议机制提出更加苛刻的要求。

3.3 设计和实现新的协议

尽管机制改进和实现优化能弥补某些性能缺陷，但是性能的改善往往受限于协议机制，对于原有的功能缺陷仍然无能为力。越来越多的研究者开始考虑重新设计新的运输层协议。按照出现的次序，我们介绍其中某些较具价值的代表，及各自的功能和性能特点。

Datakit 通用接收协议 URP 是1976由 AT&T 的 Bell 实验室设计的字节流传输协议，提供两类传输模式：字符模式和块模式，前者不要求恢复丢失数据；后者细分为报文块和流块，分别有完整和有限的差错控制和流量控制。URP 依靠后退重发恢复差错，采用滑动窗口的流量控制，握手连接建立时连接双方协商连接模式和流量窗口的大小。

Delta-t 是美国 LLNL 实验室在1978年设计的支持字符和块数据流以及请求应答事务处理的协议。基于定时器的连接管理是 Delta-t 的特色。发起方创建连接记录，启动定时器并发出数据；分组到达后启动对应的定时器，定时器溢出则认为连接释放。Delta-t 的差错和流量控制使用比特作为单位，比特编号用来监测分组的丢失、重复和失序，可选的检查和用来监测数据差错，使用后退重发进行差错恢复。Delta-t 使用滑动窗口进行流量控制。

NETBLT MIT 在1986年发布的 NETBLT 是适用于大带宽延时网络批量数据传输的协议。应用进程提供含有所需发送数据的缓存，NETBLT 分组化后作为一次数据突发发出。分组握手的连接建立将会协商缓存、分组和突发的大小。NETBLT 的流量控制有两个层次：应用层是客户进程的缓存大小，协议层是协商的突发大小和速率。每个缓存发送后，发送方等待接收方的响应。为了提高协议效率，NETBLT 也可以对几个缓存累积进行应答控制，NETBLT 使用序列编号和检查和来进行差错监测，如有差错通过选择重发的方法恢复。

VMTP 是1986年斯坦福大学 V 分布式系统设计的分布操作系统通信协议，提供可靠的实时事务处理服务，也是首次支持多点投递的运输层协议。V 系统中事务处理开始于客户发出请求，结束于多个服务器的应答，没有显式的连接过程。事务报文分割成分组块，每个分组块包括32个数据分组，分组块是执行协议控制的单位。VMTP 使用选择重发的差错恢复方法和基于速率的流量控制；采用全局的事务标识，可以适合移动性的事务处理。

SNR 1990年出现在 IEEE 通信学报，协议规范涉及数据交换和控制交换，后者是 SNR 的特色部分。接收方周期性地发出控制信息，其中的位图用来指示需要重新发送的分组。SNR 通过选择重发进行差错恢复；使用滑动窗口进行流量控制。协议具有3种模式：模式0无差错和流量控制；模式1无差错控制但具有流量控制；模式2既具有差错控制又具有流量控制。连接模式和窗口大小将在 SNR 连接建立时进行协商确定。

TP^{++} 是 Bellcore 在1991年提出的适合吉比特网络和多媒体应用的运输层协议，特别适合高带宽延迟网络。为了简化协议机制和方便质量控制，TP^{++}抛弃传统运输层协议的连接复用功能。TP^{++}的连接管理类似 Delta-t 的定时器连接模式；另外的特点就是使用前向差错校正码进行差错恢复。TP^{++}的流量控制同样依赖可变滑动窗口的抑制。

TP5 1991年由法国信息自动化所提出，是适应实时多媒体数据的运输层协议，也适合普通数据的传输，是传统 OSI/TP4的扩展。TP5给出新的 TPDU 格式 RT 表示实时多媒体数据。TP5不对 RTPDU 进行运输层意义的差错和流量控制，仅使用运输层速率和差错监测。对于特定连接的 RT 和 DT，TP5定义两类顺序关系：相对和绝对关系。前者相对于 DT 而言，先发送的 RT 只要先被接收就可接受。后者 RT 和 DT 之间的接受次序要与发送次序相同。

4 我们的工作

快捷运输协议是最为令人注目的新型协议，由国际 XTP 论坛标准化。XTP 的主要目标是在 VLSI 芯片实现100Mbps 以上的吞吐，适应数据流、数据报、批量数据及实时事务等不同类型的通信模式。XTP 是通用运输协议，没有其它新型协议特定场合的约束；在吸取传统和其它新型协议经验和教训的同时又具有自己鲜明的特色和优势，成为集大成的高速轻型协议。

东南大学计算机系网络研究室早在90年代初就开始高速网络新型协议的研究，并在1994年春成为 XTP 论坛在中国大陆地区的唯一成员。1996年初完成 XTP4.0中文版[5]标准化，获得 XTP 论坛认可（标准文档编号为 XTP-96-08）。我们选择 XTP 协议作为高速计算机网络协议研究的出发点和核心研究内容，现有的工作涉及深入探讨 XTP 协议机制，XTP

对下层高速传输网络接口(如 ATM 网络)和 XTP 对上层应用接口(如计算机会议系统)等诸多方面。在协议机制方面,研究重点分为三个方面[4]:自上而下的新型应用通信需求分析;自下而上的高速传输服务分析;高速计算机网络运输层协议机制的研究。我们选取分布多媒体和分布式处理研究新型应用的通信需求;选取 FDDI-Ⅱ,DQDB 和 ATM/S-ISDN 研究异步、同步和等时新型传输服务。我们选取 DoD 的 TCP 和 ISO 的 TP4以及 XTP 进行协议比较研究,分析它们在连接管理、差错控制、流量控制、速率控制和质量控制方面的功能和性能差异。

我们的研究发现 XTP 在功能和性能两个方面均有明显的优势。功能方面表现在:适合高速网络的大数据管道(64比特字节序列空间和流量控制空间),具有独立的流量和速率控制(方便连续媒体固定速率的数据传输),具有丰富的优先级(16位优先空间),具有带外数据传输能力(每个数据分组可以夹带8字节的带外数据),具有多样的差错控制模式(从无差错控制到类似 TCP 协议的完全正确顺序),具有快速数据连接(第1次分组交换将数据投递给接收方)和多样连接释放的功能以及显式可靠的多点投递支持(单个发送方多个接收方的多点模式,通过构造能够支持更复杂的多点到多点模式)。性能方面表现在:固定的分组头部格式设计(任何控制的解释不依赖之前的数值),分组头部控制域边界对齐(处理器在1次总线周期访问整个域),基于关键字和关键字交换的状态检索(仅在第1次分组投递需要携带显式地址,之后通过数组键检索的方式),仅选择丢失和差错数据重发(避免不必要的重发)以及基于发送方的状态获取(简化定时器的使用并减少接收方的处理负担)。

5 高速计算机网络运输层协议的标准化工作

研究和设计新型协议在研究界被广泛接受,这也受到国际性、地域性和国家性信息技术领域标准化机构的重视。OSI/IEC JTC1在对 OSI/TP 修订的同时,也正在进行新的运输层协议研究和标准化工作,如增强通信功能和设施 ECFF[9]及增强通信运输服务和协议 ECTS/ECTP[10],其中某些标准草案的概念和机制来自 XTP 协议。ECFF 的目标有两个:扩大 OSI 协议栈的应用,覆盖多点、实时及语音视频数据传输;提高运输层协议的性能和灵活性,允许应用选择协议以获得确保质量的运输服务。ECTS 有面向连接和无连接两类服务,其中面向连接的模式可以分为单向、双向和 N-向多点投递,无连接服务仅具有单向模式;并不再区分单点和多点投递,认为前者只是后者的特例。面向连接的服务原语有连接建立/终止、中止/恢复、加人/离开、数据传输和状态汇报;并提出新的 QoS 控制策略,有尽力而为、强制和确保三类。面向连接的 ECTP 规范12种不同类型的 TPDU 完成 ECTS 的服务定义。与传统 TP4不同的是,ECTP 不再支持紧急数据的传输和应答,但仍然支持 OSI/TP4协议改进的选择应答和否定应答。

ITU-T 考虑到 JTC1的 ECFF 和 ECTS/ECTP 标准化活动,将与之合作开展相应的 X.tms 研究。Internet 的标准者 IETF 也正在开展下一代 TCP/IPng 的标准化工作,其中 IPng(即 IPv6)已经发布,ECTS 也为 IETF 承认,TCPng 可能成为第一个与未来国际标准 ECTS/ECTP 一致的工业标准。

主要参考文献

[1] Gu Guanqun, Pan Jianping et. al, Communication Services and Protocols to Support Cooperative Work over High Speed Packet Networks, in Proc. of 1st Int'l Conf. on CSCW in Design. Beijing, May, 1996

[2] 顾冠群、潘建平,一种适合事务处理的高速计算机网络运输层协议,CIMS-China'96,96年8月

[3] 顾冠群、潘建平,高速网络支持多媒体应用的通信服务和协议,第九界中国网络与数据通信学术年会,西安,1996年10月

[4] 潘建平、高速计算机网络新型运输层协议机制的研究和实现的设计,[学位论文],南京:东南大学,1996年9月

[5] 顾冠群、潘建平 等译,快捷运输协议规范4.0(中文版),XTP 论坛,XTP-96-08,1996年6月

[6] Jocobson V., Congestion avoidance and control, Computer Comm. review. 18(4)1988

[7] Jocobson V., TCP extension of long delay path, Internet Request for Comment 1072, Oct. 1988

[8] Fox, R., TCP big windows and NAK options, Internet RFC-1106, June, 1989

[9] ISO/IEC, Enhanced communication functionality and facility, JTC1/SC6/WG4 Working draft, July 1995

[10] ISO/IEC, Enhanced communication transport services and protocols, JTC1/SC6/WG4 Working draft, Dec. 1995