基于 XTP 的 ITER 广域网链路优化方案研究

马宗萼[1,2]，谭海波[1]，李晓风[1]

（1. 中国科学院 合肥物质科学研究院信息中心，安徽 合肥 230031；
2. 中国科学院研究生院，北京 100049）

摘 要：为了满足国际热核聚变实验堆（ITER）计划全球网络不断增长的性能需求，提出了一种结合高速协议和改进传输机制的广域网链路优化解决方案。通过分析和研究快速传输协议 XTP，根据 XTP 提供的丰富功能接口，调整 burst/rate 模型有效避免拥塞产生，标记分组序号改进 go-back-n 重传机制，同时采用协议欺骗技术减少响应延迟，分析相关协议字段实施精确 QoS 策略，充分利用协议的特性对网络传输实现了灵活改进。测试结果表明该方案能够显著提高 ITER 中法国际链路的性能。

关键词：XTP；ITER；QoS；国际网络；优化；协议改进
中图法分类号：TP393.03 **文献标识号**：A **文章编号**：1000-7024（2012）01-0017-05

Research on optimization of global network for ITER based on XTP

MA Zong-e[1,2], TAN Hai-bo[1], LI Xiao-feng[1]

(1. Information Center, Hefei Institutes of Physical Science, Chinese Academy of Sciences, Hefei 230031, China;
2. Graduate University, Chinese Academy of Sciences, Beijing 100049, China)

Abstract: To meet the growing performance challenge of international thermonuclear experimental reactor (ITER) program's global network, a WAN optimization scheme is proposed, combining high-speed protocol and improved transport mechanisms. Based on deep research for the Xpress transport protocol (XTP), made use of feature-rich interface provided by XTP, adjust the burst/rate model to avoid congestion effectively, marked sequence numbers of packet to improve go-back-n retransmission algorithm, also employed protocol spoofing to reduce the latency, some relevant protocol fields are analyzed to implement accurate QoS policy, and the characteristics of XTP are used for a flexible network improvement. The experimental results show that the solution can significantly improve the performance of China-France global network for ITER.

Key words: XTP; ITER; QoS; global network; optimization; protocol improvement

0 引 言

国际热核聚变反应实验堆（以下简称 ITER）是当今世界最大的大科学工程国际科技合作计划之一[1]，ITER 计划吸引了包括中国、欧盟、印度、日本、韩国、俄罗斯和美国等世界主要核国家和科技强国共同参与，也是迄今我国参加的规模最大的国际科技合作计划。为了实现全球科学家的同步设计、制造和未来的远程实验，ITER 通过 Internet 建立了连接全球各主要参与国的业务协作网。目前 ITER 中法业务协作网内每天产生的数据量约为 200GB，预计未来一年内会增长到 500GB。除了需要保证 Oracle、CATIA、FTP、Web 等大量网络应用的服务质量，ITER 的“日落计划”在每天的固定时间对数据进行全球同步，这对网络性能提出了非常高的要求。

经流量分析发现，ITER 中法业务协作网络上的应用具有多突发流量、高带宽要求的特点，而传统的 TCP 传输协议在 ITER 广域网环境中的性能远远不能满足需要。考虑到升级物理设备的巨大代价以及更好地兼容当前网络应用，必须在尽可能少影响现有网络架构的前提下对链路进行合理优化，以满足未来的网络服务质量需求。

链路优化通常可以通过采用专有的高速传输协议，或者改进已有的传输控制策略，但是前者不一定能得到链路

收稿日期：2011-01-21；**修订日期**：2011-03-20
基金项目：国家科技计划基金项目（2008GB111000）
作者简介：马宗萼（1986-），男，广西合浦人，硕士研究生，研究方向为网络应用开发；谭海波（1976-），男，安徽泾县人，副研究员，硕士生导师，研究方向为计算机网络应用、工程数据库；李晓风（1966-），男，安徽砀山人，研究员，博士生导师，研究方向为软件工程、网络管理、计算机自动控制。E-mail: mazonge@hfcas.ac.cn

上现有网络设备的支持，也不一定适用于特定的网络应用场合；后者改进效果通常有限，可能达不到性能要求。下面分别从传输协议和控制策略两方面进行考虑，分析轻量级传输协议的功能和接口，同时结合协议特性对传输机制进行改进，然后对链路优化的效果进行测试和对比，最后给出结论和展望。

1 研究现状

广域网链路上的各类应用在传输层可划分为 TCP 应用、UDP 应用，TCP 协议最初的设计面向于低速、不可靠的网络，如今它在低带宽、低延迟的 LAN 环境中工作得很好，但是对于高带宽、高延迟、低丢包率的长距离链路来说，TCP 的慢启动策略、拥塞判断和避免机制不仅不再适用，反而阻碍了网络吞吐量的提高，导致带宽利用率低下；UDP 协议没有差错控制和拥塞避免机制，可以达到很高的传输速率，但是却没有可靠性保证。

目前的改进协议都是以 TCP 或 UDP 为基础[2]，TCP 改进协议主要对拥塞控制机制或差错控制机制进行修改，UDP 改进协议主要增加可靠性保证机制，使得改进后的协议适用于广域网通信[3-4]。但是这些改进协议通常只在特定场合中能够达到较高的性能，没有充分考虑不同的网络环境和应用需求，某些情况下的表现甚至比 TCP 协议更差[5]。

由于承载 ITER 业务协作网的中美俄环球科教网络（GLORIAD）是一条包丢失率不可预知，具有高延迟、高带宽特点的“长肥管道”，其上不同的网络链路间存在资源竞争关系，而现有的 TCP 协议及其特定的改进协议均不能保证为处于变化广域网环境中的链路提供足够的传输性能。为了对 ITER 业务协作网链路进行合理优化，必须在改进传输协议的同时制定符合需求的策略，保障链路的可靠性和稳定性。

2 XTP 协议

XTP 是一种可靠的轻量级传输层协议，由 PEI（protocol engines incorporated）提出[3]。XTP 拥有传统 TCP 协议的所有特征，同时又可以灵活定制，它为 TCP 协议的连接管理、差错控制、数据流控制、重传和确认等机制提供了更多选择，并且引入了许多新功能，包括优先级子系统、突发流量控制、MTU 检测等。由于 XTP 在设计上协议和策略是相分离的，因此在采用 XTP 协议的链路上，可以根据需要选择合适策略以更好地优化链路传输[6-7]。

XTP 协议帧有固定的 32 字节头部，其余为可变长度的用户数据或控制信息。头部的控制标志域 cmd 中包含 15 个可选的标志位字段，用于。除此之外，XTP 头部还包含对数据包有效信息的定位和描述。XTP 协议帧格式如图 1 所示。

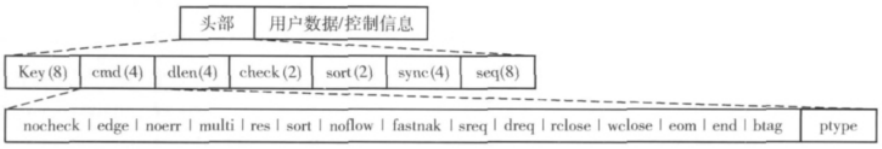

图 1 XTP 协议帧格式

XTP 本身性能不突出，但是它具有丰富的功能，通过修改相应的传输策略，可以将 XTP 定制为适用于特定广域网环境的链路传输协议。下面以 XTP 协议为基础，结合实际网络应用的需求，设计 ITER 广域网链路优化的解决方案。

3 链路优化方案和具体实现

3.1 改进流量和拥塞避免机制

通过图 1 可以观察到 ITER 网络链路中存在大量突发流量，这些突发流量会导致网络服务不稳定，甚至造成拥塞的发生，严重降低网络质量，因此需要对网络流量进行控制和整形。

对于需要大窗口来提高速率的长距离链路来说，由于 TCP 在滑动窗口允许的情况下不受控制地发送数据，容易出现拥塞和丢包的情况，而 TCP 的 AIMD 拥塞避免机制过于保守，发生丢包后窗口恢复时间非常长，这就导致大量带宽在长时间内不能得到有效利用[8]。目前出现了许多改进拥塞机制的方案[9-11]，例如可以通过运行在 TCP 接收方的算法分析当前流量趋势，动态修改通告窗口大小，达到控制流量的目的[12]。但是修改滑动窗口会造成网络吞吐量的震荡，而且实时运算网络流量模型，以及捕获和修改数据包的行为，会给网络带来额外的负载。

利用 XTP 的速率控制机制可实现简单高效的流量管理，除了使用滑动窗口外，XTP 还提供了 RATE 和 BURST 两个参数，其中 RATE 表示单位时间内可发送的最大字节数，BURST 表示单位时间内突发流量的最大字节数，发送端在 BURST/RATE 时间间隔内不传输超过 BURST 的数据量。通过调整 RATE 和 BURST 两个参数，可以避免流量突增的情况，同时根据需要对流量进行整形，保证网络数据的平稳传输。XTP 的速率控制机制如图 2 所示。

对于链路中的一般网络应用，在握手期间设定 BURST 为链路 MTU 的大小，RATE 为接收端窗口大小，在会话的整个生命周期内对这两个值动态修改。当检测到 XTP 会话不断增加，链路处于繁忙状态时，发送方将减少 RATE

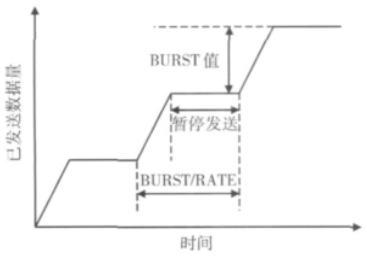

图 2　使用 rate 和 burst 避免突发流量

和 BURST 的值，避免拥塞的发生；而随着会话数量的减少，链路变得空闲，发送方将增加 RATE 和 BURST 的值，提高带宽利用率。

可见，改进的流量和拥塞控制机制能够根据接收端返回的状态，在链路繁忙阶段通过减少会话的可用流量有效避免拥塞发生，同时由于它没有改变滑动窗口的大小，避免了 TCP 漫长的窗口恢复阶段，因此在链路空闲时能够充分分配带宽资源。

3.2　改进差错和重传机制

TCP 的 go-back-n 重传策略在低延迟的 LAN 中没有问题，但它通常会重传未被接收方确认但已成功传输的数据，在高延迟网络中带来非常大的额外代价，尤其对于延迟敏感的应用，如多媒体应用，这种机制更显得低效。

XTP 的每个分组均有对应的标记序列号，接收端使用 dseq、rseq、hseq 这 3 个序号标记分组接收队列的状态。dseq 表示准备确认的下一个分组的序列号，rseq 表示已接收分组中最大的连续序列号，hseq 表示已接收分组中的最大序列号。如果 rseq 不等于 hseq 则说明出现了丢包，接收端可根据这些参数构造 CNTL 分组，用 span 字段标记 rseq 和 hseq 之间已正常接收的分组区间，通知发送方重传那些序列号在 rseq 和 hsesq 之间但不在正常接收区间内的分组，而不需要重传那些已接收但未进行确认的分组，如图 3 所示。相比于 TCP 的 go-back-n 重传算法，在高延迟的广域网链路中，改进后的重传机制更为精确和高效。

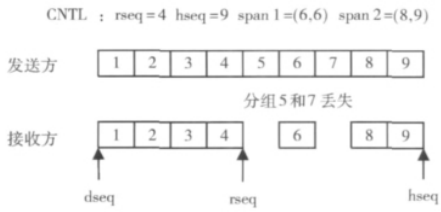

图 3　重传机制

3.3　协议欺骗

TCP 对每个分组都要等待确认以决定是否重传，在高延迟的长距离网络中这个响应时间非常浪费带宽资源。这里通过实现 TCP 协议代理监听 TCP ACK 消息[13]，并及时构造和回应发送分组，为发送方虚拟出一个低延迟的网络，可以在提高发送速率的同时有效利用带宽资源。协议欺骗设计如图 4 所示。

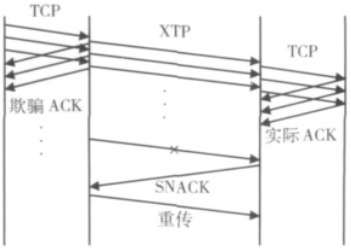

图 4　协议欺骗

XTP 设备的发送端将 TCP 分组转换为 XTP 分组，在转发分组的同时，向发送方回应自己构造 TCP ACK 信息，发送方在收到确认信息后，会认为前面发送的分组已经顺利到达目的主机，随后立即发送后续分组，对端的 XTP 设备将 XTP 分组转换为 TCP 分组，并收集本地的 ACK 信息。除此之外，设计 XTP 发送端使用快速负确认机制(SNACK)[14-15]，当 XTP 接收端检测到有分组丢失的情况出现时，发送 CNTL 报文通知发送方重传。

在协议欺骗的设计中实际回应方是 XTP 设备，响应时间很短，在发送方看来自己与接收方位于一个局域网内，继而提高分组发送速率。协议欺骗技术能够降低广域网的长延迟所带来的影响，减少响应等待时间，进而提高链路利用率。

在不考虑丢包的情况下，假设局域网平均 RTT 为 10ms，广域网平均 RTT 为 300ms，在通常情况下，完成 100 个分组的发送所需时间为 100×300ms=30s；在使用协议欺骗技术的情况下，发送同样数量的分组在理论上只需要 100×10ms+300ms=1.3s 就可以完成。

3.4　应用层优先级和 QoS

XTP 提供了 TCP 所不具备的丰富优先级系统和 QoS 信息域，XTP 设备通过程序自动策略调整优先级和网络参数，实现对网络应用多粒度多层次的管理功能。

优先级技术并不能真正提高网络带宽，而是根据链路繁忙情况以及网络应用的需求，合理调度紧缺资源，避免资源竞争。通过完善优先级策略，分析链路状态灵活调节网络应用的优先级，确保核心应用获得足够链路资源。

FIRST、TCNTL、JCNTL 类型的 XTP 分组中包含可以与路由设备协商 QoS 信息的 Traffic Specification 字段，通过在网络应用程序中分析这个字段获取反馈的网络状态信息，修正 RATE、BURST、滑动窗口等相关协议参数，同时根据不同的应用场合设置可选策略，从而在 XTP 会话建立阶段及其整个生命周期内实施精确的 QoS 策略。

4 实验结果与分析

在 ITER 中法业务协作网链路两端部署设计好的 XTP 实验设备，在 Internet 上建立一条 XTP 通道，对采用优化方案前后的端到端性能进行测试。实验环境如图 5 所示。

图 5 实验网络环境

通常在局域网中的网络应用所能获取到的带宽上限为 100Mbps，ITER 中法业务协作网链路的 RTT 约为 245ms。为了考察优化方案的性能，首先采用 TCP 的几种方案以及 XTP 优化方案后的链路传输性能进行分析，然后使用 Iperf 网络性能测试工具对这 3 种条件下的带宽情况进行测试，测试结果如图 6 所示。

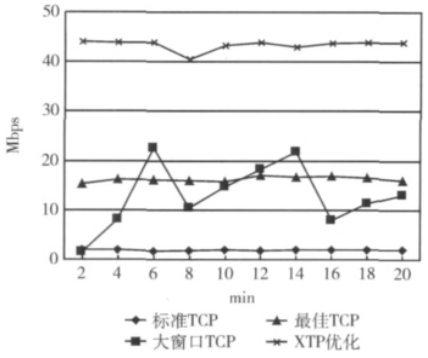

图 6 带宽利用率测试结果

根据公式：滑动窗口大小＝带宽×RTT，可以从中得出如下的结论：

（1）当使用标准 TCP 传输时，网络带宽利用率很低下，可用带宽稳定在 2Mbps 左右。这是由于标准 TCP 滑动窗口的最大尺寸仅为 64KB，通过计算得出单个 TCP 应用可使用的带宽最多只能达到 64KB×8/245ms＝2.09Mbps，与实验结果基本吻合。

（2）为了尽可能提高带宽利用率，应该根据公式将 TCP 窗口大小设置为 100Mbps×245ms/8＝ 3062.5KB，但通过实验发现没有达到理想的结果，可用带宽曲线呈锯齿状。这是由于单位时间内发送的 TCP 分组超过一定数目后，将耗尽中间路由器为这条信道所提供的缓冲区资源，出现丢包，进而导致 TCP 采用慢启动和 AIMD 拥塞避免算法管理流量造成的。

（3）为了在提高可用带宽的同时避免发生丢包，必须为滑动窗口寻找一个最佳大小。通过实验发现当滑动窗口大小为 512KB 时，可用带宽稳定在 16.7Mbps，这是一个最佳值，同时也是一个临界点，如果滑动窗口进一步增大超过 512KB，就可能导致丢包，使得可用带宽处于震荡状态；而如果减少到 512KB 以下，可用带宽就会平稳下降。

（4）使用基于 XTP 的链路优化方案，可用带宽能够维持在 40～50Mbps，在设备上对流量压缩后可稳定在 90Mbps 以上。由于优化方案在链路资源充裕时能够对网络流量进行有效管理，避免拥塞等不良网络状态的出现，充分利用带宽资源；而在链路资源紧张时，优化方案对重传精度和确认速度的提高有效提升了传输效率，从而保证网络服务平稳高效运行。

采用优化方案对 ITER 中法业务协作网链路进行改造后，在半年多的运行期内有效承载了 ITER 大科学工程项目在中法之间的数据通信，完全满足了网络链路的性能需求。图 7 所示为 ITER 中法国际链路一周内的网络流量图。

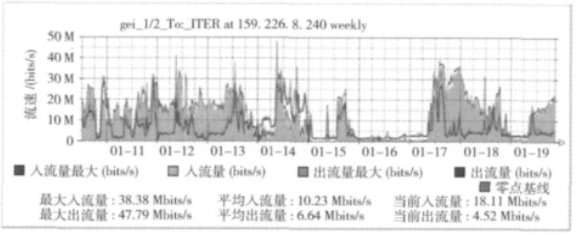

图 7 ITER 国际链路合肥端出口周网络流量

由此可见，基于 XTP 的链路优化方案通过实施改进的流量、差错控制以及协议欺骗等机制，为网络应用合理分配链路资源，相比于各类使用 TCP 协议的方案，它更有效提高了链路的带宽利用率，同时也符合广域网链路的优化需求。

5 结束语

针对日益增长的 ITER 中法业务协作网应用需求，本文以 XTP 协议为基础，提出一种广域网优化解决方案，完成对传统 TCP 流量控制、差错控制等机制的改进，采用协议欺骗技术提高分组确认速度，并辅以 QoS 策略对网络应用进行管理，经实验成功应用于实际链路，使网络性能获得较大提升。下一步的工作，将分析能够进一步优化链路传输的压缩和缓存等技术，同时完善网络状态分析模型，为链路实施更为精确的 QoS 策略。

参考文献：

[1] FENG Kaiming. Controlled nuclear fusion and ITER project [J]. China Nuclear Power, 2009, 2 (3): 212-219 (in Chinese). [冯开明. 可控核聚变和国际热核实验堆（ITER）计划 [J]. 中国核电，2009，2（3）：212-219.]

[2] REN Yongmao, TANG Haina, LI Jun, et al. Transport pro-

tocols for fast long distance networks [J]. Journal of Software, 2010, 21 (7): 1576-1588 (in Chinese). [任勇毛，唐海娜，李俊，等. 高速长距离网络传输协议 [J]. 软件学报，2010，21 (7)：1576-1588.]

[3] ZHOU Zhaoqing, CHEN Lijun. Application research on TCP/IP over satellite link [J]. Radio Engineering of China, 2006, 36 (1): 47-50 (in Chinese). [周兆清，陈立军. TCP/IP 协议在卫星链路上的应用研究 [J]. 无线电工程，2006，36 (1)：47-50.]

[4] Caini C, Firrincieli R, Marchese M, et al. Transport layer protocols and architectures for satellite networks [J]. International Journal of Satellite Communications and Networking, 2007, 25 (1): 1-26.

[5] REN Yongmao, TANG Haina, LI Jun, et al. Performance comparison of TCP variants for high-speed network by NS2 simulation [J]. Computer Engineering, 2009, 35 (2): 6-9 (in Chinese). [任勇毛，唐海娜，李俊，等. 高速网络 TCP 改进协议 NS2 仿真性能比较 [J]. 计算机工程，2009，35 (2)：6-9.]

[6] ZHANG Yasheng, PENG Hua, GU Jujuan. Research on TCP performance improvement over satellite [J]. Radio Communications Technology, 2010, 36 (5): 29-31 (in Chinese). [张亚生，彭华，谷聚娟. 卫星 TCP 加速技术研究 [J]. 无线电通信技术，2010，36 (5)：29-31.]

[7] WU Jie, GAO Suixiang. Performance improvement of TCP over satellite link based on the express transport protocol [J]. Journal of Computer Applications, 2006, 26 (7): 1563-1566 (in Chinese). [吴结，高随祥. 基于快速传输协议实现卫星 TCP 性能的改善 [J]. 计算机应用，2006，26 (7)：1563-1566.]

[8] XING Guowen, YU Zheming. Comparative analysis network congestion control algorithms of XCP and TCP [J]. Computer Engineering and Design, 2008, 29 (6): 1339-1341 (in Chinese). [邢国稳，虞哲明. XCP 与 TCP 的拥塞控制算法比较分析 [J]. 计算机工程与设计，2008，29 (6)：1339-1341.]

[9] XU Jing, SUN Zhen, LI Shiyin, et al. TCP congestion control algorithm based on self-similar flow prediction [J]. Computer Engineering and Design, 2010, 31 (12): 2713-2715 (in Chinese). [徐京，孙珍，李世银，等. 基于自相似流量预测的 TCP 拥塞控制算法 [J]. 计算机工程与设计，2010，31 (12)：2713-2715.]

[10] GAO Wenyu, LI Shaohua. Survey of congestion control in TCP protocol [J]. Information Technology, 2009, 33 (3): 11-15 (in Chinese). [高文宇，李绍华. TCP 拥塞控制研究综述 [J]. 信息技术，2009，33 (3)：11-15.]

[11] YANG Xiaoping, SHI Shuai, CHEN Hong. Improved algorithm for TCP congestion control [J]. Journal of Jilin University (Engineering and Technology Edition), 2006, 36 (3): 433-437 (in Chinese). [杨晓萍，史帅，陈虹. 一种改进的 TCP 拥塞控制算法 [J]. 吉林大学学报：工学版，2006，36 (3)：433-437.]

[12] YANG Hu, ZHANG Dafang, XIE Kun, et al. A serial traffic control algorithm based on the TCP sliding window in the Netfilter/Iptables framework [J]. Computer Engineering and Science, 2009, 31 (10): 8-11 (in Chinese). [杨虎，张大方，谢鲲，等. Netfilter/Iptables 框架下基于 TCP 滑动窗口的串行流量控制算法 [J]. 计算机工程与科学，2009，31 (10)：8-11.]

[13] GUO Wei. Research on application of Spoofing technology in satellite link TCP transmission [J]. Coal Technology, 2010, 29 (11): 146-147 (in Chinese). [郭伟. Spoofing 技术在卫星链路 TCP 传输运用研究 [J]. 煤炭技术，2010，29 (11)：146-147.]

[14] JIAO Chengbo, DOU Ruiyu, LAN Julong. Analysing and improving TCP SACK mechanism in wireless networks [J]. Application Research of Computers, 2007, 24 (3): 238-240 (in Chinese). [焦程波，窦睿彧，兰巨龙. 无线网络中选择性重传机制性能分析与改进 [J]. 计算机应用研究，2007，24 (3)：238-240.]

[15] LIU Liqiang, ZHOU Xiyi, ZHANG Ge. Research on improving TCP performance over wireless network [J]. Radio Communications Technology, 2008, 34 (1): 4-5 (in Chinese). [刘俐强，周细义，张舸. 改进无线网络 TCP 性能的研究 [J]. 无线电通信技术，2008，34 (1)：4-5.]