ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.21, No.7, July 2010, pp.1635−1645
doi: 10.3724/SP.J.1001.2010.03853

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

终端性能自适应传输协议*

王伟杭[1,2+], 任勇毛[1], 唐明洁[1,2], 李 俊[1], 钱华林[1]

[1](中国科学院 计算机网络信息中心,北京 100190)

[2](中国科学院 研究生院,北京 100049)

End-System Performance Aware Transport Protocols

WANG Wei-Hang[1,2+], REN Yong-Mao[1], TANG Ming-Jie[1,2], LI Jun[1], QIAN Hua-Lin[1]

[1](Computer Network Information Center, The Chinese Academy of Sciences, Beijing 100190, China)

[2](Graduate University, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: E-mail: wangweihang@cstnet.cn

Wang WH, Ren YM, Tang MJ, Li J, Qian HL. End-System performance aware transport protocols. *Journal of Software*, 2010,21(7):1635−1645. http://www.jos.org.cn/1000-9825/3853.htm

Abstract: With the rapid development of numerous applications and optical network, fast long distance optical network has emerged in the Internet field. Some novel studies indicated that, due to the exigent requirement of scientific applications and the rapid improvement of network bandwidth, the network transfer speed has outstripped the processing speed of the end systems. Therefore, the congestion has been pushed into the end nodes and the processing ability of the end system has become a bottleneck in such a network. Thus several novel end-system performance aware transport protocols have been proposed. The authors categorize and describe all these protocols, and emphatically analyze that congestion detection, rate adaption mechanisms, advantages and disadvantages of all kinds of schemes. Subsequently, the current existing open issues are summarized and some further interesting directions are pointed out.

Key words: fast long distance optical network; transport protocol; end-system performance aware; congestion detection; rate adaption

摘 要: 随着各种应用的需求和光网络技术的飞速发展,互联网领域出现了高速长距离光网络.最新研究发现,由于当前各种科学应用的迫切需求以及网络带宽的迅速提高,网络速率已经远远超出了终端系统的处理能力.在高速长距离光网络环境中,拥塞已经从网络转移到了终端,终端系统的处理能力逐渐成为传输速率的瓶颈.因此,各种终端性能自适应的高速传输协议应运而生.基于这一类协议改进的不同思路,对它们进行了分类描述,重点分析了这些协议的拥塞检测和速率适配机制以及各自的优缺点,在归纳和总结目前研究中仍然存在的开放性问题的同时,提出了进一步的研究方向.

关键词: 高速长距离光网络;传输协议;终端性能自适应;拥塞检测;速率适配

* Supported by the National High-Tech Research and Development Plan of China under Grant No.2007AA01Z214 (国家高技术研究发展计划(863)); the Knowledge Innovation Program of the Chinese Academy of Sciences under Grant No.CNIC_QN_08004 (中国科学院知识创新工程青年人才领域项目)

Received 2009-08-13; Accepted 2010-03-11

中图法分类号: TP393 文献标识码: A

随着信息技术的快速发展,科研活动的信息化水平日益提高,e-Science 应用不断扩展.分布于全球的各种科学观测仪器、传感器以及大规模实验装置通过高速网络连接进行数据传输和远程控制.在高能物理、天文学等领域的大科学工程尤其需要高性能网络的支持.例如,中国科学院高能物理研究所和欧洲粒子物理研究中心(CERN)之间有粒子物理方面的合作研究项目,位于 CERN 的大型粒子对撞机每年产生几十 PB 的数据,这些数据要采用网格传送到分布在全球的各个研究中心,以便处理和分析,这对网络有着非常高的要求.在天文学领域,天文学家采用 VLBI(very long baseline interferometry,甚长基线干涉测量法)来获得详细的图像,相关的实验由全球分布式的仪器采集数据并通过网络传送到一个中心点.高性能计算和可视化研究系统需要通过高带宽的链路(约 Gbps 或几百 Mbps)远程传送 TB 级数据[1].

e-Science 科研应用的海量科学数据传输对网络提出了很高的要求,传统的 IP 分组共享路由网络由于带宽、QoS 等方面的限制,无法满足这种需求.因此,近年来,国际科研组织纷纷研究部署光网络.典型的项目有美国 Internet2 的动态电路网(dynamic circuit network,简称 DCN[2])、ESnet 的科学数据网(science data network,简称 SDN[3])以及 CHEETAH[4],DRAGON[5],UltraScience Net[6],HOPI[7]等、欧洲的 PHOSPHORUS 项目[8]、加拿大 CANARIE 主导的 UCLP(user controlled lightpath)[9]、日本的 JGN2[10]、韩国的 KREONet2[11]等.由中国科技网(CSTNET)参与发起、多个国家参与的中美俄环球科教网络(global ring network for advanced applications development,简称 GLORIAD[12])就是为了满足各国间科学数据传输需求而建设的光网络.

光网络提供了巨大的带宽,但其传输性能却并没有得到同步的提升.其原因是,传统的 TCP 协议是针对低速、低时延的网络而设计的,其保守的拥塞控制和流量控制机制在高速长距离网络中性能很差[13].尽管研究人员提出了多种 TCP 改进协议,试图提高 TCP 协议在高速网络中的性能,但是带宽利用率仍然不高.因此,各种基于 UDP 的改进协议纷纷被提了出来,这类协议相对于 TCP 类改进协议传输性能有了显著提高[14].这类协议的典型特征是采用 UDP 协议传输数据,通过 TCP 传递控制信息,以提供高速可靠数据服务.目前,典型的基于 UDP 的改进协议有 RBUDP(reliable blast UDP)[15],SABUL(simple available bandwidth utilization library)[16],UDT (UDP-based data transfer protocol)[17],Tsunami[18]等.在基于 UDP 协议的改进协议中,有一部分协议在设计速率调整算法时没有考虑终端系统的处理能力,比如 RBUDP,SABUL/UDT,这类协议只适应于高性能的终端系统;还有一部分协议考虑到了终端的性能问题,它们依据终端系统的接收能力实时调整发送速率,从而获得更好的传输性能.本文主要讨论这些终端性能自适应的传输协议.

本文第 1 节概述终端性能自适应传输协议的基本原理.第 2 节对现有的终端性能自适应传输协议进行分类描述,重点分析每种协议的拥塞检测和速率适配机制.第 3 节对已有的各种传输方案进行摘要比较.第 4 节提出本领域尚待研究的开放性问题和进一步的研究方向.第 5 节总结全文.

1 终端性能自适应传输协议基本原理

光网络技术的飞速发展,尤其以密集波分多路复用技术(dense wavelength division multiplexing,简称 DWDM)为代表,网络带宽迅速提高,使得通过单根光纤实现 Tbps 数据传输成为可能[19].而终端系统由于受到网卡速率、I/O 总线带宽、CPU 芯片组处理速率、内存容量、前端总线等硬件速率限制以及 TCP/IP 协议处理速率限制[20]等,其性能已成为高速网络的传输瓶颈[21,22].此外,由于各种终端系统的性能本身存在很大的差异,在进行网络传输的同时,往往还要执行数据处理等其他任务,其传输能力也处于不断变化之中.因此,关注发生在终端系统中的拥塞,设计终端性能自适应的传输协议是一个非常有意义的研究课题.

这一类协议尝试利用各种拥塞检测参数对接收端性能进行检测,按照一定的拥塞检测(congestion detection)算法,一旦认为接收端即将发生拥塞或拥塞消除,便显式反馈给发送端,发送端根据反馈采取相应的速率调整(rate adaption)策略.

本文主要介绍 RBUDP+[23],RAPID[24],RAPID+[25],GTP(group transport protocol)[21],PA-UDP[26] 和

RTsunami[27]协议.它们的区别在于选用了不同的拥塞检测参数,而且设计的拥塞检测算法和速率适配策略各不相同.首先考虑拥塞检测算法,我们将接收端拥塞的加剧和减弱分别表示为拥塞↑(congestion aggravation)、拥塞↓(congestion alleviation).表 1 列举了上述几种协议所对应的拥塞检测参数以及拥塞检测算法.

Table 1 Congestion detection metrics and algorithms of transport protocols

表 1 各种协议的拥塞检测参数和拥塞检测算法

Protocols	Congestion detection metrics	Congestion aggravation	Congestion alleviation
RBUDP+	The time slice allocated to I/O process: $T_{I/O}$; sum of all other processes' time slices: $\sum_j T_j$	Once I/O process has been executed, other processes will be executed at: $t+T_{I/O}$	Other processes has been finished, the I/O process will be executed at: $t+T_{I/O}+\sum_j T_j$
RAPID	Dynamic priority of I/O process: *dynamicpriority*	*dynamicpriority*−1≤*ERROR_PRIORITY*	*dynamicpriority*−1>*ERROR_PRIORITY*
RAPID+	Packet loss ratio in the *n*th and (*n*−1)th cycle: α_n, α_{n-1}	$\alpha_n>\alpha_{n-1}$	$\alpha_n\leqslant\alpha_{n-1}$
GTP	Packet loss ratio of the *i*th flow: $loss_i$	$loss_i>0$	$loss_i=0$
PA-UDP	Remaining file size: *bitsLeft*; Left buffer: *memLeft*	*BitsLeft*>*memLeft*	*bitsLeft*≤*memLeft*
RTsunami	Congestion level: *CL*; dynamic threshold: *U*(*k*), *k* is congestion level	*CL*>*U*(*k*)	*CL*<*U*(*k*)

拥塞检测算法指示接收端的拥塞情况,速率适配据此采取一定的速率适配算法调整发送速率.速率适配算法可以广义地表示为

$$\text{拥塞}\uparrow: r\leftarrow r(1-g),$$

$$\text{拥塞}\downarrow: r\leftarrow r(1+f),$$

其中,r 表示发送速率,f 表示接收端拥塞缓解采用的调整因子,而 g 表示接收端拥塞加剧时采用的调整因子.表 2 总结了上述几种协议所对应的速率适配因子 f 和 g 的表达式.

Table 2 Rate adaption factors of transport protocols

表 2 各种协议所对应的速率适配因子

Protocols	f	g
RBUDP+	0	1
RAPID	0	1
RAPID+	$\begin{cases}0, & count<k\\ \gamma, & count=k\end{cases}$ *count* is counter of times that packet loss sustainably decreases; *k* is an empirical value; γ is increased factor of transfer speed	$1-\frac{\delta_{n-1}}{r}$ δ_{n-1} is receiver's processing speed in the (*n*−1)th period
GTP	$\min\left\{0.02, \frac{r_{i,target}}{r}-1\right\}$ $r_{i,target}$ is the target speed of *i*th flow allocated by centralized control	$\min\{0.5loss_i, 0.125\}$
PA-UDP	$\frac{R_{\max}}{r}-1$ $R_{\max}$ is the maximum speed set by user	$\frac{bitsLeft}{bitsLeft-memLeft}\times r(disk)$ $r(disk)$ is receiver's speed of moving data from buffer to disk
RTsunami	$\frac{1}{\lambda}-1$ λ is increased factor of inter-packet delay, $0<\lambda<1$, default is 5/6	$\frac{\alpha\times 2^k}{1+\alpha\times 2^k}$ α is adaptable factor, default is 0.125; *k* is congestion level

2 终端性能自适应传输协议分类描述

从对上述几种协议拥塞检测算法以及速率适配方案的比较我们可以看到,这些协议对各种终端性能参数

进行检测,比如有些协议检测网络传输进程丢包率,有些协议检测接收端缓冲区使用情况等.为了便于分析和研究,我们根据传输协议的拥塞检测参数将这些方案分为以下几类.

2.1 基于进程调度的拥塞检测机制

这一类方案通过监测接收端各个进程的交互特性提供反馈控制机制,例如进程动态优先级(dynamic priority)、时间片等参数.这类协议试图准确预测接收端即将发生拥塞和拥塞缓解的时间点,以决定速率调整的时刻.在本文中,我们主要介绍 RBUDP+和 RAPID 协议.这两种协议都对 RBUDP(reliable blast UDP)[15]协议进行了改进.RBUDP 协议非常简单,发送端利用 UDP 协议,以恒定速率持续发送至所有数据被发送,然后利用 TCP 收到接收端关于收到数据的反馈,发送端据此再通过 UDP 重传丢失的数据.此过程不断重复,直至所有数据被成功接收为止.研究人员发现,在高速长距离光网络环境中,RBUDP 以恒定速率持续发送的策略极易导致接收端缓冲区溢出[21,22],尤其是当初始速率设置过高时,这个问题尤为严重.研究人员便尝试修正 RBUDP 中的数据发送方案,以使得当接收端即将发生拥塞时,发送端不是保持原来的速率持续发送,而是停止发送一段时间.而当拥塞缓解时,发送端再继续以原来的速率发送.这两种协议的区别在于拥塞检测算法有所不同.

2.1.1 RBUDP+

RBUDP+协议采用接收端所有任务的时间片作为拥塞检测因子.其设计思想在于,出于交互性、任务间公平性的考虑,接收端 CPU 在网络传输进程与其他任务间不停地交替.每当接收端处理传输任务以外的其他任务时,如果发送端持续发送数据,接收端可能由于不能及时处理快速到来的数据而将其丢弃.因此,RBUDP+协议的核心思想即预测接收端处理其他任务的时间点,在接收端处理其他任务期间内停止发送,而在其他任务执行结束时再继续传输.

RBUDP+协议的通信过程如图 1 所示.其中,时刻 t,t'为接收端执行网络传输进程(I/O process)的时刻,时间片 $T_{I/O}$, $T'_{I/O}$ 为 CPU 分配给传输进程的时间片, $\sum_j T_j$, $\sum_j T'_j$ 为在 I/O process 执行结束后,其他任务的时间片之和.我们以第 1 次拥塞检测和速率适配为例,说明 RBUDP+的拥塞检测及速率适配算法.在 RBUDP+中,每当接收端开始执行 I/O process 时(图 1 中时刻 t),接收端便预测 CPU 分配给其他进程的时间片之和(图 1 中 $\sum_j T_j$),同时使用 TCP 协议将 $T_{I/O}$, $\sum_j T_j$ 显式反馈给发送端.发送端根据反馈认为,在网络传输时间片执行结束时(即时刻 $t+T_{I/O}$),如果继续发送将使接收端缓冲区溢出而丢包,因此,发送端在时刻 $t+T_{I/O}$ 停止发送(实际上,考虑到 RTT 的存在,发送端在时刻 $t+T_{I/O}-RTT/2$ 即停止发送),等其他进程任务执行结束时(时刻 $t+T_{I/O}+\sum_j T_j$),再继续以原来的速率发送.

与 RBUDP 相比,RBUDP+的性能只有在接收端需要频繁地多任务间切换、发送速率较高的情况下才有所提高.在接收端只有网络传输进程或发送速率较小时,接收端才有足够的能力处理持续到来的数据,RBUDP+频繁地停止发送反而降低了传输效率.文献[23]通过仿真指出,接收端只运行网络传输、以 3.4Gbps 传输 700MB 的文件,RBUDP+的用时比 RBUDP 要多 1.3s.对于典型的科学数据应用,可能需要远程传输 TB 级的数据,那么以 3.4Gbps 的速率、1TB 数据的传输时间差将达到 1947.4s.当接收端进行多个任务、以低于 2.4Gbps 的速率传输 700MB 数据时,RBUDP+比 RBUDP 需要更多的传输时间.当传输速率大于 2.4Gbps,RBUDP+的传输时间才开始小于 RBUDP.综合以上两种情况,RBUDP+协议的设计者提出,可以依据接收端运行任务以及传输可用带宽的不同情况来确定选择 RBUDP 或 RBUDP+传输.

虽然在接收端运行多个任务、传输带宽较高时,RBUDP+性能较 RBUDP 有所提高,但是 RBUDP+的拥塞检测和速率适配算法还都过于简单:首先,协议的设计者没有明确给出一个如何对时间片准确估计的方案.事实上,Pallab 等人已经从理论上分析证明,RBUDP+无法对进程时间片进行准确估计[25];其次,接收端每次处理 I/O process 时都要求发送方停止发送一段时间,带宽浪费严重.例如,接收端每次为 I/O process 分配的时间片 $T_{I/O}$ 平均为 10ms,为其他任务分配的时间片之和 $\sum_j T_j$ 为 100ms,取 RTT 值为 300ms.线路上有数据传输的时间为 $T_{I/O}$,

没有数据传输的时间为 $\sum_j T_j + RTT/2$,带宽利用率仅为 4%.另外,协议假设 $T_{I/O}$>*RTT*/2.对于典型的高速长距离传输网络,极有可能 *RTT*/2>$T_{I/O}$.这时,发送端根本无法在 $t+T_{I/O}$ 时刻停止,导致接收端来不及处理而丢弃数据.在之后的 UDP blast 中,更多的包被重传,从而降低了传输效率.更为严重的情况是 $RTT/2 > T_{I/O} + \sum_j T_j$,这种情况下,发送端根本就不会停下来,于是,RBUDP+退化为 RBUDP 协议.

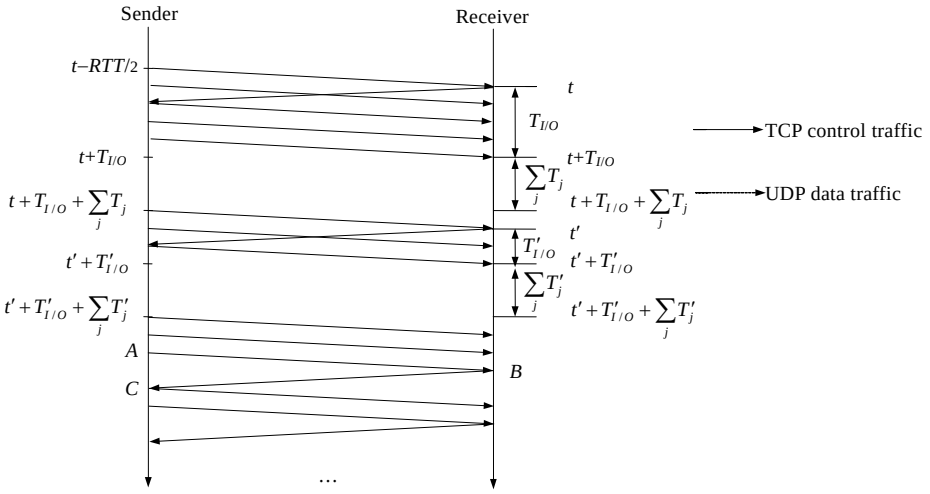

Fig.1 Time sequence diagram of RBUDP+ transmission

图 1 RBUDP+传输时序图

2.1.2 RAPID

由 Banerjee 等人提出的 RAPID,通过检测接收端网络传输进程的动态优先级(dynamic priority)来预测拥塞发生的时刻.RAPID 在接收端运行一个性能监测进程 PMP(performance monitoring process),以固定的时间间隔(polling interval)监测传输任务的动态优先级.同时,每当收到一个数据包,接收进程便向 PMP 发送信号.在一个 polling interval 中,如果 PMP 未接收到任何信号,则它认为有数据包被丢弃.但是,为了区分拥塞和错误,RAPID 对第 1 个丢包并不立即采取速率调整措施,而是记录下接收任务此时的动态优先级,这个值存于传输进程的 *ERROR_PRIORITY* 变量中.自此以后,PMP 会在每个 polling interval 监测接收进程的 dynamic priority,当接收进程的 dynamic priority 达到了比 *ERROR_PRIORITY* 高 1 个级别的时候,PMP 认为接收端即将发生丢包,进入拥塞状态.这时,计算出一个停止发送的持续时间(suspend interval)反馈给发送端.发送端收到反馈立即停止发送,等 suspend interval 结束后再以原来的速率继续发送.

不同于 RBUDP+每次处理 I/O process 发送端都要停下来,而无论接收端是否真的来不及处理来自 I/O process 的数据这一情况,RAPID 减少了发送端停止发送的次数,带宽利用率明显高于 RBUDP+.但是,RAPID 协议的拥塞检测算法也存在明显的缺陷.接收端对于第 1 个丢包不作处理,这时拥塞可能已经发生,继续发送将会浪费宝贵的带宽.另外文献[24]指出,进程优先级改变 1 个级别需要大概是 100ms 的时间.从接收端发出反馈到发送端做出反应需要 *RTT*/2,再到接收端真正接收不到数据还需要 *RTT*/2.因此,从接收端发出反馈到真正空闲(没有数据到来)实际上需要 1 个 *RTT* 时间.如果在 1 个 *RTT* 时间内优先级改变超过了 1,那么按照 RAPID 的拥塞检测算法,接收端会因为持续到来的数据而发生拥塞.事实上,依据文献[24]的仿真实验结果,当 *RTT* 为 120ms 时,RAPID 的传输时间就超过了 RBUDP.因此,对于 *RTT*≥300ms 的长距离光网络,算法中的 1 个优先级级别并不适用.

虽然 RBUDP+和 RAPID 与 RBUDP 相比在一定程度上缩短了数据传输时间,提高了传输效率,但是它们与 RBUDP 一样,都需要用户设定初始速率.用户在使用这两种协议传输数据之前,必须测量链路可用带宽,这大大

减小了可用性.同时,由于发送端都需要等到所有数据发送完才能收到接收端返回的确认信息,发送端必须保留所有已发送的数据,因此,这两种协议传输的文件大小受限于发送端的缓存大小.另外,这两种协议采用的停止-等待式(stop-and-go)速率适配算法会产生抖动,使得它们不适于对抖动敏感的应用.

2.2 基于丢包率的拥塞检测机制

这类协议通过监测网络传输应用的丢包率,从而调整发送速率.当所观测到的丢包率在增大或大于某一特定阈值时,发送端认为接收端拥塞程度有所加剧,需要降低数据发送流量以缓解拥塞;反之,则认为接收端的拥塞有所缓解,可以相应地增加数据发送流量.其中,速率适配算法的实现可以在发送端收到关于丢包率的反馈后进行,也可以是由接收端执行再反馈给发送端.其代表性成果有 RAPID+和 Group Transport Protocol(GTP)等等.

2.2.1 RAPID+

RAPID+协议设计的巧妙之处在于,当丢包率增加时,发送端将速率减为前一个持续周期内接收端从缓冲区读数据的速率δ_{n-1}.协议的设计者认为丢包率有所增加的原因是,在第 $n-1$ 个周期,缓冲区已被填满,新的发送速率设置为前一个周期读数据的速率,可以快速使丢包率最小化的同时保持高的吞吐量.

当丢包率减小时,发送端认为接收端拥塞程度有所缓解,发送端保持和上一周期相同的速率.同时,如果丢包率连续 k(经验值)个周期都在下降,那么发送端认为可以安全地根据丢包率来增加发送速率.RAPID+协议的拥塞检测和速率适配过程如图 2 所示,其中:α_n,α_{n-1} 为第 n 个、第 $n-1$ 个周期的丢包率;R_{n+1} 和 R_n 为第 $n+1$ 个、第 n 个周期的发送速率;$count$ 为丢包率连续减小次数的计数器,取值从 1~k;γ为速率增加的比例因子.

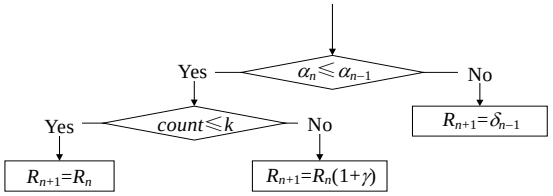

Fig.2 Congestion detection and rate adaption of RAPID+

图 2 RAPID+协议的拥塞检测和速率适配过程

RAPID+依据接收任务丢包率的变化调整发送速率,减小了丢包发生的次数.相对于持续发送的 RBUDP,传输相同大小的文件,RAPID+实际发送的数据包远远少于 RBUDP 实际发送的数据包.协议的设计者 Feng 等人通过实验证明:传输 1GB 文件时,RAPID+实际发送的数据比 RBUDP 减少 41.47%~111.86%;传输 2GB 数据时,减少值为 139.32%~387.91%[25].但是,RAPID+的拥塞检测机制过于简单,丢包率的相对变化并不能准确地反映接收端的拥塞状态.RAPID+的拥塞检测是相对地比较当前的丢包率和上一次速率调整后的丢包率,没有一个绝对值可供参考.考虑这种情况:α_{n-1} 已经很大,虽然有$\alpha_n \leqslant \alpha_{n-1}$,但$\alpha_n$ 仍然很大,保持原来的速率甚至增加速率可能会加剧接收端的拥塞,浪费带宽.因此,设定适当的、动态的门限值更能准确地反映接收端的拥塞情况.

2.2.2 GTP

考虑到运行相同协议的多个传输流之间的公平性问题,Ryan 等人提出了 GTP 协议.GTP 协议的创新点在于通过两个层面实现对多个流速率的公平分配:集中层面控制和单个流层面控制.单个流层面控制是对单个流进行拥塞控制,与 RAPID+类似,利用丢包率来调整发送速率.集中控制的目的是实现公平分配带宽资源.集中控制过程中,接收端周期性地(默认为 $3RTT_{\max}$,$RTT_{\max}$ 表示观测到的所有流中 RTT 的最大值)对每个流的容量进行估计,然后对所有流的容量之和利用 Max-Min[28]公平分配策略进行速率分配.在单个流层面的控制,接收端周期性地(单个流的 RTT)根据这个流的丢包情况更新为其分配的速率,反馈给发送端.这个速率的上限受制于集中控制使用 Max-Min 原则为其分配的速率($r_{i,target}$),如表 2 中 GTP 速率适配算法所示.

Ryan 等人针对 GTP 在多点到点的通信模式,以及同一链路运行 GTP 的多个并行流之间的公平性参数进行了详尽的仿真实验.文献[19]指出,在多点到点的通信模式下,3 种基于 UDP 的改进方案(RBUDP,UDT,GTP)都保

持了较高的吞吐量(>400Mbps),然而,只有 GTP 仍然保持极低的丢包率(GTP 为 0.06%,RBUDP 和 UDT 分别为 53.3%, 8.7%).另外,文献[21]的仿真结果表明,多个具有不同 *RTT* 的 GTP 并行流可以公平地分配带宽.

但是需要指出的是,在目前的 GTP 仿真实验中,运行 GTP 协议的多个流,*RTT* 均不超过 100ms,这并不满足跨越洲际的长距离光网络的 *RTT* 值的要求.因此,GTP 协议在长距离网络中的高效性值得进一步研究.另外,在仿真实验中,如表 2 所示,速率适配算法增加和减小的百分比被经验性地限定为 0.02 和 0.125,但却缺少理论分析来证明这种取值的合理性.如何在瞬息万变的实际网络中确定这些参数,以及是否需要采用动态变化的取值策略,协议的设计中没有提出一个明确的方案.

2.3 基于缓冲区利用情况的拥塞检测机制

通过对接收端缓冲区使用情况的测量,发送端可以了解接收端当前的负载状态,从而决定如何调整发送速率.考虑一个数据传输过程,如果接收端缓冲区无限大,那么即使处理数据的速率小于数据的到达速率,那么接收端也可以先将其保存下来,再进行处理.然而事实的情况是,接收端的缓冲区非常有限,如 Linux 2.6.9 默认的接收缓存为 131KB[26],并且数据处理速率远远小于到达速率.这样小的接收缓存极容易被高速到来的数据填满,从而丢弃后续到来的数据.这里,我们简单介绍两种对接收端缓冲区进行拥塞检测的传输协议:PA-UDP 和 RTsunami.

2.3.1 PA-UDP

当接收速率快于磁盘写速率时,快速到达的数据将使得接收端缓冲区以一定的速率增长,这个速率与接收速率和处理速率的差相关.Eckart 等人建立了一个关于缓冲区不会溢出、并且可以被充分利用时传输速率的数学模型.PA-UDP 的拥塞检测是通过简单地比较剩余文件大小(*bitsLeft*)和接收端缓冲的可用空间(*memLeft*)来实现的.当 $bitsLeft \leqslant memLeft$ 时,发送速率简单地设置为预先设定的最大值.而当 $bitsLeft > memLeft$ 时,发送速率设置为由数学模型得到的、能够使缓冲区被充分利用并避免溢出的速率大小.

PA-UDP 协议试图利用数学模型对传输速率进行准确估计,但是此协议所基于的数学模型假定磁盘写速率总是比接收速率要慢.虽然从整个传输过程来看,这是成立的,但是,由于缓冲区的存在,在具体的传输过程中,磁盘速率有时会比接收速率要快.比如,当发生网络拥塞或终端缓冲区占用率较高的情况时,发送方暂时减少发送速率,这时,接收速率可能比磁盘速率慢.文献[26]仅给出了在 LAN 中传输一个 5GB 文件的实验结果(在 LAN 中,1Gbps,*RTT*<1ms,无背景流量),平均吞吐量仅有 331.96Mbps.因此,PA-UDP 协议尚不成熟,传输速率不高,有待于进一步深入研究.另外,我们利用 FLDnet 实验床[27]发现,在一条 1Gbps 的端到端链路,PA-UDP 协议与 TCP 传输流同时传输 512MB 的文件,PA-UDP 抢占带宽的能力极强,TCP 传输流的吞吐率只有 PA-UDP 传输流的 1.55%[29].因此,两者并存时,利用 TCP 传输的应用不能与 PA-UDP 公平地分配带宽.

2.3.2 RTsunami

RTsunami 协议在 Tsunami[18]的基础上,对其拥塞控制算法进行了改进.Tsunami 协议采用的是简单的基于丢包率的拥塞检测机制.而在 RTsunami 中,其拥塞控制算法 CDRA(congestion detection and rate adaptation)[27]综合考虑了缓冲区利用率和拥塞的变化趋势.通过比较拥塞水平值 *CL* 与动态阈值 $U(k)$(k 表示拥塞级别),来判断终端的处理能力.如果 $CL<U(0)$,发送端认为接收端没有发生拥塞,发送方可以通过减小包间延迟以增大发送速率;而当 $CL>U(k)$时,发送端认为接收端已经处于拥塞状态,需要通过增大包间延迟来减小发送速率.

RTsunami 结合使用了当前的缓冲区占用率和队列长度的变化来定义拥塞水平 *CL* 值:

$$CL = \varphi BO + (1-\varphi)2^{que_i / que_{i-1} - 1 - \gamma}.$$

这里,各参数的含义分别为:*BO* 表示缓冲区占用率;*que* 是队列长度,i 为更新周期;参数 γ 保证了 *CT*(拥塞趋势)值在一定的范围内;φ 根据不同的缓冲区占用率 *BO*,在[0,1]范围内动态改变.CDRA 算法给 *BO* 设定 3 个阈值:δ_1,δ_2 和 δ_3,且满足 $\delta_1>\delta_2>\delta_3$.在 RTsunami 现阶段算法的具体实现中,采用静态映射表来确定 φ 参数的值,见表 3.

而动态阈值 $U(k)$的确定表示为

$$U(k)=U(k-1)+I/2^{k-1},$$

其中,k 表示拥塞级别,I 是拥塞水平 *CL* 的固定增量.

RTsunami 协议周期性地反馈重传队列中的块序号和接收端的拥塞水平 *CL* 值,消除了传输文件大小的限制.更为重要的是,RTsunami 协议的拥塞检测不仅考虑了当前磁盘空间的利用率,而且还考虑了磁盘空间的拥塞趋势,提出了更为准确地表现终端拥塞情况的 *CL* 值.同时,协议的设计者对动态阈值 $U(k)$的设计方案,使得随着拥塞级别 k 的增加,发生发送速率减小的情况会越来越频繁,并且减小的幅度也会越来越大.这些设计都有利于快速地缓解接收端的拥塞.协议的作者 Ren 等人利用 1Gbps 的 FLDnet 实验床进行仿真,在丢包率为 1%的链路环境中,RTsunami 的传输速率是 Tsunami 的 2.43 倍.另外,仿真结果表明,RTsunami 具有较强的带宽抢占能力,这会对其他应用造成一定影响,具体的影响程度需要通过包含各种网络背景流量的网络仿真来加以评估.

Table 3 Relationship between buffer occupancy and φ

表 3 缓冲区占用率 *BO* 与φ参数映射表

BO (buffer occupancy)	φ
$BO>\delta_1$	1
$\delta_1>BO>\delta_2$	σ, σ is a constant, $\sigma>0.5$
$\delta_2>BO>\delta_3$	$1-\sigma$
$\delta_3>BO$	1

3 终端性能自适应传输协议比较

以上总结了现有的几种终端性能自适应传输协议的拥塞检测和速率适配方案,这些协议目前尚不成熟,都或多或少存在着各种各样的问题.在表 4 中,我们从多个方面对上述各种协议的关键特征进行了综合对比.通过初始速率(initial rate)、建立连接(connection setup)、拥塞检测参数(congestion detection metrics)以及拥塞控制算法的运行端(algorithm executor)这几个特征,可以了解每种协议的工作过程.而协议内部公平性(intra-protocol fairness)和 TCP 友好性(TCP-friendliness)[30]反映了在并行流存在时,各种协议的设计思路.

Table 4 Summary comparison of end-system performance aware transport protocols

表 4 各种终端性能自适应传输协议的摘要比较

Protocols	Initial rate	Connection setup	Congestion detection	Algorithm executor	Intro-Protocol fairness	TCP-Friendliness
RBUDP+	Specified by user	No	Time slice of process	Receiver	Not considered	Require certification
RAPID	Specified by user	No	Dynamic priority of process	Receiver	Not considered	Require certification
RAPID+	Negotiated by sender and receiver	Yes	Variance in packet loss ratio	Sender	Not considered	Require certification
GTP	Negotiated by sender and receiver	Yes	Packet loss in one period	Sender	Max-Min fairness	No
PA-UDP	Negotiated by sender and receiver	Yes	Comparison of remaining file size and left buffer	Sender	Support	No
RTsunami	Negotiated by sender and receiver	Yes	Congestion level of buffer	Sender	Not considered	Require certification

其中,拥塞控制算法包括拥塞检测算法和速率适配算法两部分,这两部分算法通常是在同一终端运行.在接收端运行拥塞控制算法的协议,由接收端实现拥塞检测和速率适配算法,然后将结果反馈给发送端,发送端通过调整 UDP 包间延迟来改变发送速率;而发送端运行拥塞控制算法的情况是,由接收端将拥塞检测参数反馈给发送端,发送端运行拥塞检测和速率调整算法,并依据算法结果调整包间延迟.

Intra-Protocol fairness 指的是运行同一种协议的多个流能够得到相同级别的服务,这就要求多个流的速率分配满足特定的公平指标,如 Max-Min 公平标准.在现有的终端性能自适应传输协议中,除了 GTP 与 PA-UDP,其他 4 种协议在设计时并未考虑到协议内部的公平性问题.

而 TCP-Friendliness 则代表一种传输协议与 TCP 共享相同的链路时,抢占带宽资源的能力.一种协议抢占带宽的能力越强,我们称这种协议的 TCP 友好性越差.目前,大多数应用仍采用 TCP 协议进行数据传输.因此,选择一个 TCP 更为友好的高速传输协议显得尤为重要.在现有的终端性能自适应协议当中,只有 PA-UDP 是开源的,

我们在文献[29]中,基于 FLDnet testbed 进行模拟实验,当 PA-UDP 存在时,TCP 协议的吞吐率仅为 PA-UDP 协议吞吐率的 1.55%.因此,PA-UDP 的 TCP 友好性极差.另外,Ryan 等人也在文献[19]中指出,由于设计时未考虑到 TCP 友好性的问题,GTP 抢占带宽的能力很强.其他几种协议的 TCP 友好性尚需进一步加以研究.

4 开放性问题及进一步的研究方向

在对高速长距离网络传输协议的研究过程中研究人员发现,终端性能已成为高速数据传输的瓶颈,因此提出了各种终端性能自适应的高速传输协议,这些协议从不同的角度提出了多种拥塞检测机制和速率适配方法.但是,由于对终端性能自适应传输协议的研究本身是一个很新的研究领域,目前提出的方案尚不成熟,很多方面需要进一步加以研究.同时,由于高速长距离网络传输协议设计的复杂性,在终端性能自适应传输协议设计方面仍然存在很多开放性问题,主要体现在如下几个方面:

1) 拥塞检测机制的准确性问题.到目前为止,对于多种拥塞检测机制缺乏统一的认识和理论分析.同时,如何保证拥塞检测参数的准确度是这类协议研究的关键问题所在.例如,RBUDP+中基于时间片估计的拥塞检测机制与接收端发生拥塞并没有本质上的联系,这种拥塞检测机制缺乏准确性和有效性.另外,多种改进方案中共同存在的、依据固定阈值进行拥塞检测的设计方案有待进一步加以改进.因此,基于丢包率、基于接收端缓冲区利用情况等拥塞检测机制,其有效性以及准确度仍需进一步通过理论分析,并结合网络仿真和实际网络的测量进行验证.

2) 速率适配的效率问题.目前的终端性能自适应传输协议的多数研究方案,缺乏针对速率适配效率的总结性归纳与规范化的描述.虽然已有一些研究利用建模方法针对终端瓶颈速率进行分析[26,31],但针对各种方案的速率适配效率的比较、分析仍然太少.现有的速率适配算法仅依赖于经验性地设计,同时通过有限的、局部的仿真实验来证明设计方法的有效性.这样的算法往往是静态的,不能适应高速网络的快速变化.比如在 RBUDP+和 RAPID 中,只是通过停止发送和持续发送进行速率适配.显然,这种速率适配方案效率很低.

3) 网络中间的拥塞问题.关注发生在终端系统中的拥塞,是终端性能自适应传输协议的根本特点.然而,与传统的高速传输协议(例如 TCP 及其改进协议[32])只考虑网络中的拥塞、忽略终端系统瓶颈的局限性类似,目前的终端性能自适应传输协议只考虑了终端系统的性能对传输能力的影响,却缺乏当网络带宽成为瓶颈时的应对策略.同时,现有的实验结果都是基于网络传输瓶颈在终端的实验环境,而缺乏在路由网络中的验证.当应用于路由路径(route path)时,网络中间的拥塞是有可能发生的.如果发生,这些传输协议能否快速地做出反应?其传输性能如何?因此,这些传输协议能否适用于分组共享的路由网络仍有待进一步加以研究.

4) 性能评价问题.高速光网络传输协议的发展仍然处于起步阶段,对于各种终端性能自适应的高速传输改进协议,目前详尽的性能评价研究还很少,尤其是公平性等性能评价参数.基于电路交换的光网络技术可以提供端到端的光链路,应用流可以独享端到端的链路.如果端到端只开启一个传输流,那么公平性问题并不重要.但是,如果同时在两点间开启多个传输利用相同协议或不同协议的传输流,运行相同协议的多个流之间的公平性(intra-fairness)或运行不同协议的传输流之间的公平性(inter-fairness)就要被考虑.目前,光分组混合网络仍然是高速长距离网络的主要存在形式,光链路在很多地方还不能延伸到终端用户.即使将来光网络技术普及应用,小流量普通用户仍将是采用共享路由网络接入.而分组共享网络就必须考虑传输流之间的公平性问题,因此,对于终端性能自适应传输协议的公平性等指标的性能评价尚需更多的研究.

5 结束语

本文对多种新的基于终端性能检测的高速传输协议进行了分类描述,重点分析了高速长距离光网络这一热点领域中已有的策略和算法的优缺点.在归纳和总结目前研究中仍然存在的开放性问题的同时,提出了我们对于这一领域进行进一步研究的各种思路.

致谢 感谢中国科技网研发部唐海娜老师,她对本文工作给予了很大的支持;感谢 NGI 研究组一起从事高速长

距离网络传输协议研究的李一鸣、罗旋、王学智等其他同学,他们参与了部分协议原理的讨论.

References:

[1] Ren YM, Tang HN, Li J, Qian HL. Optical network control and management for grid applications. Journal of Software, 2008,19(6): 1481−1490 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/19/1481.htm [doi: 10.3724/SP.J.1001.2008.01481]

[2] Dynamic circuit network. 2008. http://www.internet2.edu/network/dc/

[3] ESnet, energy sciences network. 1994. http://www.es.net/

[4] Zheng X, Veeraraghavan M, Rao NSV, Wu QS, Zhu MX. CHEETAH: Circuit-Switched high-speed end-to-end transport architecture testbed. IEEE Communication Magazine, 2005,43(8):11−17. [doi: 10.1109/MCOM.2005.1497551]

[5] Lehman T, Sobieski J, Jabbari B. DRAGON: A framework for service provisioning in heterogeneous grid networks. IEEE Communications Magazine, 2006,44(3):84−90. [doi: 10.1109/MCOM.2006.1607870]

[6] Rao NSV, Wing WR, Carter SM, Wu Q. Ultrascience net: Network testbed for large-scale science applications. IEEE Communications Magazine, 2005,43(11):12−17. [doi: 10.1109/MCOM.2005.1541682]

[7] Summerhill R. Next generation networking and the HOPI testbed. In: Proc. of the CANS 2005. http://cans2005.cstnet.cn/down/1102/A/afternoon/hopi-cans-summerhill-2-nov-2005_4A.pdf

[8] PHOSPHORUS. 2008. http://www.phosphorus.pl/

[9] Communications Research Centre. User controlled light paths. 2005. http://uclp.uwaterloo.ca/

[10] JGN2. 2004. http://www.jgn.nict.go.jp/english/index.html

[11] KREONet2. 1988. http://www.kreonet.re.kr/english/

[12] GLORIAD. 2004. http://www.gloriad.org/gloriad/index.html

[13] Ren YM, Qin G, Tang HN, Li J, Qian HL. Performance analysis of transport protocol over fast long distance optical network. Chinese Journal of Computers, 2008,31(10):1679−1686 (in Chinese with English abstract).

[14] Ren YM, Tang HN, Li J, Qian HL. Performance comparison of udp-based protocols over fast long distance network. Information Technology Journal, 2009,8(4):600−604. [doi: 10.3923/itj.2009.600.604]

[15] He E, Leigh J, Yu O, DeFanti T. Reliable blast UDP: Predictable high performance bulk data transfer. In: Proc. of the IEEE Cluster Computing. Washington: IEEE Computer Society, 2002. 317−324. http://www.evl.uic.edu/cavern/papers/cluster2002.pdf

[16] Gu Y, Grossman RL. SABUL: A transport protocol for grid computing. Journal of Grid Computing, 2003,1(4):377−386. [doi: 10.1023/B:GRID.0000037553.18581.3b]

[17] Gu Y, Grossman RL. UDT: UDP-Based data transfer for high-speed wide area networks. Computer Networks, 2007,51(7): 1777−1799. [doi: 10.1016/j.comnet.2006.11.009]

[18] Meiss MR. Tsunami: A high-speed rate-controlled protocol for file transfer. 2004. http://steinbeck.ucs.indiana.edu/~mmeiss/papers/tsunami.pdf

[19] Wu XR, Chien AA. Evaluation of rate-based transport protocols for Lambda grids. In: Proc. of the IEEE Conf. on High-Performance Distributed Computing (HPDC-13). 2004. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1323499

[20] Wang S, Su JS. A survey of technology for TCP acceleration. Journal of Software, 2004,15(11):1689−1699 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/1689.htm

[21] Wu XR, Chien AA. GTP: Group transport protocol for Lambda grids. In: Proc. of the 4th Int'l Symp. on Cluster Computing and the Grid (CCGrid). Washington: IEEE Computer Society, 2004. 228−238. http://www-csag.ucsd.edu/papers/GTP_CCGrid2004.pdf

[22] Xiong C, Leigh J, He E, Vishwanath V, Murata T, Renambot L, Defanti T. Lambdastream—A data transport protocol for streaming network-intensive applications over photonic networks. In: Proc. of the 3rd Int'l Workshop on Protocols for Fast Long-Distance Networks (PFLDNet). 2005. http://www.evl.uic.edu/files/pdf/lambdaStream.pdf

[23] Datta P, Sharma S, Feng W. A feedback mechanism for network scheduling in LambdaGrids. In: Proc. of the 6th Int'l Symp. on Cluster Computing and the Grid (CCGrid). Washington: IEEE Computer Society, 2006. 584−591. http://synergy.cs.vt.edu/pubs/papers/datta-ccgrid2006-lambdagrid.pdf

[24] Banerjee A, Feng W, Mukherjee B, Ghosal D. RAPID: An end-system aware protocol for intelligent data transfer over lambda grids. In: Proc. of the 20th Int'l Parallel and Distributed Processing Symp. (IPDPS 2006). 2006. http://synergy.cs.vt.edu/pubs/papers/banerjee-ipdps2006-rapid.pdf

[25] Datta P, Feng W, Sharma S. End-System aware, rate-adaptive protocol for network transport in LambdaGrid environments. In: Proc. of the SC 2006. New York: ACM Press, 2006. http://sc06.supercomputing.org/schedule/pdf/pap229.pdf

[26] Eckart B, He XB, Wu QS. Performance adaptive UDP for high-speed bulk data transfer over dedicated links. In: Proc. of the IEEE Int'l Parallel and Distributed Processing. Cambridge: IEEE Computer Society, 2008. 1–10. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=4536281

[27] Ren YM, Tang HN, Li J, Qian HL. A novel congestion control algorithm for high performance bulk data transfer. In: Proc. of the IEEE Int'l Symp. on Network Computing and Applications (NCA 2009). Cambridge: IEEE Computer Society, 2009. 288–291. http://portal.acm.org/citation.cfm?id=1590957.1591218

[28] Bertsekas DP, Gallager R. Data Networks. 2nd ed., Prentice-Hall, 1992.

[29] Wang WH, Tang MJ, Ren YM, Li J. Characterization and evaluation of end-system performance aware transport schemes for fast long-distance optical networks. Information Technology Journal, 2010,9(4):766–773. http://scialert.net/abstract/?doi=itj.2010.766.773

[30] Mahdavi J, Floyd S. TCP-Friendly unicast rate-based flow control. In: Technical Note Sent to the End-to-end Interest Mailing List. 1997. http://www.psc.edu/networking/papers/tcp_friendly.html

[31] Banerjee A, Mukherjee B, Ghosal D. Modeling and analysis to estimate the end-system performance bottleneck for high-speed data transfer. In: Proc. of the 5th Int'l Workshop on Protocols for Fast Long-Distance Networks (PFLDNet). 2007. 55–60. http://wil.cs.caltech.edu/pfldnet2007/paper/EndSystems.pdf

[32] Huang XM, Lin C, Ren FY. Recent development of high speed transport protocols. Chinese Journal of Computers, 2006,29(11): 1901–1908 (in Chinese with English abstract).

附中文参考文献:

[1] 任勇毛,唐海娜,李俊,钱华林.支持网格应用的光网络控制和管理.软件学报,2008,19(6):1481–1490. http://www.jos.org.cn/1000-9825/19/1481.htm [doi: 10.3724/SP.J.1001.2008.01481]

[13] 任勇毛,秦刚,唐海娜,李俊,钱华林.高速长距离光网络传输协议性能分析.计算机学报,2008,31(10):1679–1686.

[20] 王圣,苏金树.TCP 加速技术研究综述.软件学报,2004,15(11):1689–1699. http://www.jos.org.cn/1000-9825/15/1689.htm

[32] 黄小猛,林闯,任丰源.高速传输协议研究进展.计算机学报,2006,29(11):1901–1908.

王伟杭(1985－),女,黑龙江大庆人,硕士生,主要研究领域为传输协议.

李俊(1968－),男,博士,研究员,博士生导师,主要研究领域为下一代互联网,高速网络.

任勇毛(1981－),男,博士,助理研究员,主要研究领域为高速网络,传输协议.

钱华林(1940－),男,研究员,博士生导师,主要研究领域为下一代网络体系结构.

唐明洁(1984－),男,硕士生,主要研究领域为数据挖掘.