高速传输协议研究进展

黄小猛 林 闯 任丰源
(清华大学计算机科学与技术系 北京 100084)

摘 要 随着互联网的蓬勃发展,大规模的高速下一代互联网试验环境已经形成.而最新的研究发现:在流稀疏的吉比特级高速网络试验环境中,因为 TCP 使用了保守的加性增加和激进的乘性减少策略来调整拥塞窗口,使得 TCP 协议无法充分利用丰富的带宽资源.因此各种新的高速传输协议应运而生.文章基于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 下一代互联网;高速网络;传输协议
中图法分类号 TP393

Recent Development of High Speed Transport Protocols

HUANG Xiao-Meng LIN Chuang REN Feng-Yuan
(*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract With the rapid development of Internet, testbed of next generation Internet characterized by large scale and high speed have been constructed. Whereas some novel studies indicated that since the traditional TCP uses the conservative Additive Increase(AI) and aggressive Multiplicative Decrease(MD) mechanism to update the congestion window, it could not make full use of the abundant bandwidth resource of Gigabit high speed testbed shared by small scale flows. Thus several novel transport protocols have been proposed. The authors categorize and describe all these protocols, and emphatically analyzed the advantages and disadvantages of all kinds of mechanisms. Subsequently, the authors summarize the current existing open issues, and provide some further interesting directions.

Keywords next generation Internet; high speed network; transport protocol

1 引 言

美国一些科研机构和 34 所大学的代表于 1996 年在芝加哥提出开发新一代互联网 Internet2,以提供高速互联网服务的设想, Internet2 计划的最初目的是实现远程医疗、数字图书馆、虚拟实验室等资源共享,到 2004 年 2 月,Internet2 所建立的独立高速网络试验床 Abilene 的骨干带宽已经从 2.5Gbps 全面升级到 10Gbps.与此同时,在其他国家和地区也相继开展了下一代高速互联网络研究,英、德、法、日、加等发达国家目前除了拥有政府投资建设和运行的大规模教育和科研网络以外,也都建立了研究高速计算机网络及其典型应用技术的高速网试验环境.中国大陆相关机构也在积极开展新一代互联网发展战略研究,NSFCNET,中文全称是中国高速互

收稿日期:2005-11-02;修改稿收到日期:2006-01-10.本课题得到国家"九七三"重点基础研究发展规划项目基金(2003CB314804)和国家自然科学基金(60573122,90412012,60372019,60373013)资助.**黄小猛**,男,1980 年生,博士研究生,主要研究方向为高速网络拥塞控制、系统性能评价等. E-mail: xmhuang@csnet1.cs.tsinghua.edu.cn. **林 闯**,男,1948 年生,教授,博士生导师,主要研究领域为系统性能评价、计算机网络、随机 Petri 网、逻辑推理模型等. **任丰源**,男,1970 年生,博士,副教授,主要研究领域为网络流量的控制与管理、测控网络、传感器网络等.

联研究试验网络，即中国的下一代互联网，采用200Gbps密集波分复用DWDM光传输技术，在北京建立了连接6个节点的2.5～10Gbps高速计算机互连研究试验网，分别与中国教育和科研计算机网CERNET、中国科技网CSTNET互联，同时还分别与Internet2，GEANT，APAN，GTRN等国际组织连接.

从下一代互联网建设与发展的各种趋势表明：大规模的高速下一代互联网试验环境已经形成，未来的几年里，骨干带宽为622Mbps(OC-12)，2.5Gbps(OC-48)和10Gbps(OC-192)的高速长距离网络将成为下一代互联网中的主流网络.同时随着高速网络的发展以及各种新型应用的产生，对网络的数据传输要求也将不断提高，现在已经有越来越多的研究人员开始经常利用这些高速网络传输10Gbps～1Tbps的数据.从这种海量数据传输中受益的应用领域包括涉及量子物理学、地球观察、生物信息科学和射电天文学等方面的各种数据集中网格应用以及Web站点的镜像(例如在电子商务方面)和基于push的Web高速缓存更新应用等等[1,2].

2 TCP拥塞控制机制在高速网络中的局限

虽然下一代互联网络的高速网试验床已经建成，而且仍在不断地改进与升级.但是研究人员却发现在使用TCP协议的流稀疏高速网络测试环境中，网络无法保证100%的带宽利用率.文献[1]指出这一问题之所以产生是因为TCP使用了保守的加性增加和激进的乘性减少策略来调整拥塞窗口.举例来说，假设高速链路带宽是10Gbps，分组大小为1500bytes，回路时延为100ms，则在到达稳定传输时，TCP发送端的拥塞窗口会达到83333个分组大小，如图1所示，依据TCP协议在拥塞避免阶段的加性增加、乘性减少(AIMD)[3]的窗口调整策略计算出TCP拥塞避免阶段所经历的时间将会是4167s，约1.2h，这表示丰富的带宽资源在长时间内都无法得到充分利用.同时，文献[1]还指出TCP所面临的另一个问题，根据TCP窗口大小w与丢包率p之间的约束关系$w=1.22/p^{0.5}$[4]，要使窗口大小稳定在83333个分组大小，丢包率必须等于2×10^{-10}，这意味着每5×10^{9}个分组中只允许1个分组丢失，也就是说在将近1.7h内只允许发生一个丢包事件发生，但是即使是当前误码率最低的光通信技术也很难达到这样的要求.

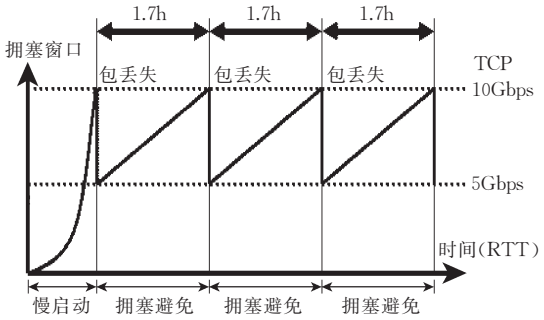

图1 TCP在10Gbps高速网络中的拥塞避免曲线

3 高速传输协议的分类描述

由于当前各种应用的迫切需求以及传统TCP协议本身所存在的缺陷，研究适应于下一代互联网的高速传输协议成为了网络研究中的一个新的热点问题.近年来，研究人员改进或提出了一些新的方案，为了便于分析和研究，我们从传输协议改进的不同思路可以将这些方案分为以下几类.

3.1 基于类TCP的隐式拥塞反馈改进方案

在传统TCP中，发送端每收到一个ACK包，将拥塞窗口增加$1/w$(w为拥塞窗口大小，以包为单位)，若收到一个丢包信号，譬如超时或三次重复确认，则将拥塞窗口减半.当研究人员发现这种保守的加性增加和激进的乘性减少策略不适用于高速网络时，便尝试修正TCP中的拥塞窗口调整方案，以使得拥塞窗口的增加过程更快速，而减少过程要更缓和一些.这一类协议的窗口调整算法可以广义地表示为

$$ACK:\ w\leftarrow w+\frac{f(x)}{w},$$

$$DROP:\ w\leftarrow(1-g(x))w.$$

其中，w表示拥塞窗口大小，$f(x)$表示发送端在收到ACK包采用的加性函数，而$g(x)$表示在收到丢包信号时采用的乘性函数，而变量x表示拥塞调整函数所可能选取的隐式拥塞反馈因子，譬如当前拥塞窗口大小、排队延时以及丢包时间间隔.发送端在不需要路由器提供任何的显示反馈的情况下，能够通过局部观察这些隐式参考因子来推断网络的拥塞程度.在本文中我们主要介绍High Speed TCP(HSTCP)[1]，Scalable TCP[2]，Binary Increase Congestion TCP(BICTCP)[7]，Hamilton TCP(H-TCP)[5]和TCP FAST[6]协议.它们的区别在于选

用了不同的反馈因子，而且设计 $f(w)$ 和 $g(w)$ 的方法各不相同. 表 1 列举了上述几种协议所对应的参考因子 x 以及相应的 $f(x)$ 和 $g(x)$ 表达式.

表 1　各种协议所对应的参考因子和窗口调整函数

传输协议	参考因子 x	$f(x)$	$g(x)$
TCP	拥塞窗口大小 w	1	0.5
HSTCP	拥塞窗口大小 w	$0.16w^{0.8}g(w)/(2-g(w))$	$0.69-0.12\log w$
Scalable TCP	拥塞窗口大小 w	$0.01w$	0.125
BICTCP	拥塞窗口大小 w	$\begin{cases} S_{\max}, & tw-w\geq S_{\max} \\ tw-w, & S_{\max}>tw-w>S_{\min} \\ 0, & tw-w\leq S_{\min} \end{cases}$ 其中 $tw=(maxwin+minwin)/2$	0.125
H-TCP	相邻丢包事件的时间间隔 t	$\begin{cases} 2(1-g(t)), & t\leq t_0 \\ 2(1-g(t))\cdot\left[1+10(t-t_0)+\left(\frac{t-t_0}{2}\right)^2\right], & t>t_0 \end{cases}$ 其中，t_0 为时间阈值	$\frac{RTT_{\min}}{RTT_{\max}}$
FAST	排队延时 t	$\begin{cases} w, & w_{\text{new}}\geq 2w \\ w(w_{\text{new}}-w), & w\leq w_{\text{new}}<2w \\ w(w-w_{\text{new}}), & w/2\leq w_{\text{new}}<w \\ -w, & w_{\text{new}}<w/2 \end{cases}$ 其中，$w_{\text{new}}=w\times RTT_{\text{base}}/RTT_{\text{avg}}+\alpha$，$t=RTT_{\text{avg}}-RTT_{\text{base}}$	0.5

3.1.1　HSTCP 和 Scalable TCP

在 HSTCP 和 Scalable TCP 中，当拥塞窗口发生变化时，其 $f(w)$ 和 $g(w)$ 能够随着当前的窗口大小自适应地动态变化.

我们在第 2 节提到了 TCP 窗口大小 w 与丢包率 p 之间存在约束关系，由 Sally Floyd 等人提出的 HSTCP 的核心思想正是从修改这一约束关系为出发点，考虑到光纤链路通常的丢包率为 10^{-7}，为了在这一丢包率级别上能够充分利用 10Gbps 的带宽，Floyd 将现有的约束关系由 $w=1.22/p^{0.5}$ 修改为 $w=0.12/p^{0.84}$，从而推导出相应的 $f(w)$ 和 $g(w)$ 表达式. 同时，为了保证与传统 TCP 的兼容性，HSTCP 中指定当拥塞窗口大小小于 38 个分组时，HSTCP 退化为传统的 TCP 协议.

而由 Kelly 等人提出的 Scalable TCP 是通过乘性增加、乘性减少(MIMD)来调整拥塞窗口大小. 作者通过实验对收敛速度、丢包恢复时间、速率折半所需时间和速率翻倍所需时间四项指标进行权衡，最终确定 $f(w)$ 取值为 $0.01w$，$g(w)$ 取值为 0.125. 与 HSTCP 类似，Scalable TCP 中也指定当拥塞窗口大小小于 32 个分组时，Scalable TCP 退化为传统的 TCP 协议.

HSTCP 和 Scalable TCP 的设计目标均是为了达到高吞吐量，但是真实网络的测试结果表明 HSTCP 和 Scalable TCP 在 TCP 友好性和 RTT 公平性方面存在严重的问题[7]. 由于 HSTCP 和 Scalable TCP 流抢占可用带宽的能力要远远高于 TCP，从而严重影响了 TCP 流对带宽的正常使用，甚至于会出现 TCP 流饿死的情况. 同时同构的 HSTCP 流或 Scalable TCP 流由于 RTT 的不同会产生流间不公平，而且当 RTT 的差距越大时，这种 RTT 不公平性就越突出.

3.1.2　BICTCP

基于对高速协议的高吞吐量、公平性和 TCP 友好性方面的考虑，Rhee 等人提出了 BICTCP 协议. 其核心思想与折半搜索算法非常相似，首先假设拥塞窗口的最小值 $minwin$(初始为 1)和最大值 $maxwin$(极大常量)，并取其中值作为目标窗口 tw 大小，即 $tw=(maxwin+minwin)/2$，当拥塞窗口更新为目标窗口大小之后，如果传输过程中没有发生丢包，则更改最小窗口值 $minwin$ 为当前拥塞窗口大小，并再次执行搜索算法寻找目标窗口，否则更改最大窗口值 $maxwin$ 为当前拥塞窗口大小，并再次执行搜索算法寻找目标窗口，通过这样多次迭代直到当前拥塞窗口的大小与目标窗口的差距小于某一阈值 $Smin$ 为止，此时拥塞窗口的大小即为稳定的拥塞窗口大小，这一过程被作者称之为“Binary Search Increase”.

在协议的具体实现中，由于目标窗口与当前拥塞窗口大小可能存在很大差距，过快地增加拥塞窗口大小会给网络带来巨大的压力，所以作者提出了一辅助性的策略“Additive Increase”，即当目标窗口大小与当前拥塞窗口大小的差距大于某一阈值 $Smax$ 时，拥塞窗口大小并不直接更新为目标窗口大小，而是增加 $Smax$ 个分组，然后更改最小窗口值 $minwin$ 为当前拥塞窗口大小，再继续目标窗口的搜索过程，如果目标窗口大小与当前拥塞窗口大小的差距小于阈值 $Smax$ 时，才真正执行“Binary Search

Increase"过程.

在表 1 中列出了这两个过程所对应的 $f(w)$表达式.对于丢包事件的处理,BICTCP 使用与 STCP 一样的倍乘因子,即 $g(w)=0.125$.

与 HSTCP 和 Scalable TCP 类似,在 BICTCP 中规定当拥塞窗口大小小于 31 个分组时,退化为传统的 TCP 协议.这种窗口门限的设计目的实际上是为了保证在非高速网络环境时的 TCP 友好性,但是这种方法存在一个明显的缺陷,即当带宽很高而 RTT 很小时(在高速局域网中这种情况很常见),拥塞窗口大小很容易大于这些窗口门限,虽然 TCP 本身依然能够达到比较理想的传输效果,但是此时 HSTCP,Scalable TCP 和 BICTCP 已经不再执行传统的 TCP 协议,所以 TCP 友好性无法得到很好的保障.

另外还需要指出,在目前的 BICTCP 仿真实验中,$Smax$ 和 $Smin$ 被经验性的取值为 32 和 0.01,而缺少理论分析证明这种取值的合理性,如何在瞬息万变的实际高速网络中确定这些参数以及是否需要采用动态变化的取值方案,文献[7]中并没有给提出一个明确的方案.

3.1.3 H-TCP

H-TCP 协议的主要创新点在于以相邻丢包事件的时间间隔 t 来作为反馈因子. Leith 等人认为如果 t 比较小,则表示丢包事件连续发生,网络拥塞严重,反之 t 越大则表示网络拥塞程度越轻.在 H-TCP 中,$g(t)=RTT_{\min}/RTT_{\max}$,其中 $RTT_{\min}$ 和 $RTT_{\max}$分别表示发送端在传输过程中所观测到的最小回路延时和最大回路延时.作者通过分析 TCP 友好性,提出如表 1 所示的 $f(t)$表达式,其中,t_0 表示的是一时间阈值,当观测到的 t 小于 t_0 时,执行保守的加性增加机制;当 t 大于 t_0 时,其加性函数能够对于可用带宽进行快速响应.仿真试验结果表明 H-TCP 在高带宽情况下的收敛速度、公平性要优于 HSTCP 和 Scalable TCP. 但是关于 $f(t)$和$g(t)$的设计完全是一种经验型的启发式设计,而且以相邻丢包事件的时间间隔作为控制因子也是一种全新的拥塞控制机制,目前对于这一机制的评价还鲜有相关的研究发表,所以 H-TCP 的性能还有待于进一步深入研究.

3.1.4 FAST

FAST 的基本原理与 TCP Vegas[10]类似,以排队延时作为反馈因子,当所观测到的 RTT 开始变大时,源端认为数据分组在路由器中产生了排队,网络中拥塞程度有所加剧,需要相应地降低数据发送流量以缓解拥塞,而当观测到的 RTT 变小时,则认为网络中的拥塞有所缓解,可以相应地增加数据发送流量.

FAST 结合使用了延时和丢包来判断拥塞,它根据如下窗口计算公式来计算目标拥塞窗口的大小:

$$w_{\text{new}}=\frac{w\times RTT_{\text{base}}}{RTT_{\text{avg}}}+\alpha.$$

这里,各参数的含义分别为:w 表示当前的拥塞窗口大小;w_{new} 表示预期的拥塞窗口大小;RTT_{base} 表示所观测到的最小 RTT 时间,在所有观测到的 RTT 样本中,取最小的 RTT 为 RTT_{base}; RTT_{avg} 表示使用指数加权滑动平均算法计算出来的平均 RTT 时间;α 表示一个非负的修正因子,对于不同的窗口大小都对应了不同的 α 值,在 FAST 现阶段算法的具体实现中,采用静态映射表来确定 α 参数的值,如表 2 所示.

表 2 吞吐量与 α 参数映射表[17]

吞吐量(Gbps)	α
<0.1	20
1	200
2.5	500
10	2000

为了避免过快地增加拥塞窗口大小给网络带来巨大的负荷,FAST 中使用表 1 所示的分阶段策略来调整拥塞窗口,对于丢包事件,FAST 使用与 TCP 一样的倍乘因子,即 $g(t)=0.5$.

在 FAST 协议中,如何进行 RTT_{base}参数和 α 参数的选取仍然是开放问题.当路由发生振荡时,例如当网络在数据传输过程中选取了更短的路径,选取观测到的最小的 RTT 作为 RTT_{base}的策略,实际上并不能反应网络的拥塞程度有所改善;同时,FAST 中对于 α 参数的映射策略是一种经验性的取值策略,当网络环境发生变化时,这种方法缺乏自适应性,如何动态选取 α 参数仍然是一个值得深入研究的问题.

3.2 基于由路由器提供显式拥塞反馈的改进方案

这一类方案是通过路由器提供显式反馈控制机制,即确切地以显式的方式通知终端用户网络的拥塞情况,终端用户据此调整发送速率,从而用户可以最大限度地利用网络资源,又避免网络发生拥塞.其代表性成果有 eXplicit Control Protocol(XCP)[8], Exponential Max-min Kelly Control (EMKC)[9]和 Variable-structure congestion Control Protocol (VCP)[10]等等.

3.2.1 XCP

如图 2 所示,XCP 借鉴了 ATM ABR 的设计思

想，并结合了 ECN[11] 的优点，在 XCP 中为数据包增加了一个拥塞报头. 其中，H_cwnd 和 H_rtt 都是由发送端写入，为路由器计算可分配带宽提供必要的信息. 而 $H\_feedback$ 是在发送端初始化后，由经过的路由器依次进行修改，路由器将计算出的吞吐量反馈信息写入，这一值既可以是正值也可以是负值，最后接收端收到数据包之后，将这些信息通过 ACK 包捎回. 发送端根据如下的策略来调整拥塞窗口大小：

$$\text{ACK}: w \leftarrow w + H\_feedback$$

H_*cwnd*(当前发送端的拥塞窗口大小)
H_rtt(当前发送端的 *RTT*)
H_feedback(根据发送端的需求作初始化)
H_cwnd(当前发送端的拥塞窗口大小)

图 2　XCP 的拥塞报头

在路由器中计算 H_*feedback* 时，XCP 协议采用效率和公平相分离的原则，即有效性和公平性分别由效率控制器(EC)和公平控制器(FC)进行独立控制. 其原理如图 3 所示. EC 的目标是实现链路的最大利用率，同时最小化丢包率和队列长度. ϕ 的计算公式可以描述为

$$\phi = \alpha \times d \times S - \beta \times Q.$$

其中 α 和 β 都是常数，根据对协议稳定性的分析，将 α 和 β 分别取值为 0.4 和 0.226. 参数 d 表示平均回路延时，参数 S 表示剩余的可用带宽. 参数 Q 表示稳定的队列长度，它的值可以根据平均回路延时来计算. 很显然，EC 使用的是一种 MIMD 模式.

图 3　EC 与 FC 的原理框图

FC 的目标是为每个流合理地分配反馈，FC 使用与传统 TCP 相同的 AIMD 模式以达到公平性，FC 中的反馈分配策略可以表述为

如果 $\phi > 0$，将正反馈平均分配到每个业务流.

如果 $\phi < 0$，将根据每个业务流当前的吞吐量成比例地分配负反馈.

如果 $\phi \approx 0$ 时，虽然有效性达到最优，但是对于公平性的收敛会进入停顿状态，所以 FC 中引入一个振荡带宽的概念，其计算公式可以描述为

$$h = \max(0, \gamma \cdot y - |\phi|).$$

其中，y 表示在一个 RTT 时间内的输入流量，γ 是一个值为 0.1 的常数. 这个公式的意义在于，每一个 RTT_{avg} 时间内，保证至少有 10% 的输入流量可以用来针对公平控制的再分配，振荡带宽的引入对于 EC 会产生负面的影响，但是考虑到 EC 和 FC 的折中，文中认为取 $\gamma = 0.1$ 可以达到一个较为理想的平衡.

通过将拥塞状态信息放入数据包报头中，使得 XCP 无需路由器维持每流状态，从而降低了路由器的开销，而且仿真试验表明，与传统的 TCP 拥塞控制机制相比，XCP 具有链路利用效率高、收敛速度快、公平性好、排队时延小的优点.

3.2.2　EMKC

EMKC 协议是从 Kelly 提出的 primal 算法改进而来，其创新点在于引入负丢包率的概念. 在 EMKC 中，路由器计算丢包率的表达式为 $p(t) = (X(t) - C)/X(t)$，其中 $X(t)$ 为路由器的聚合输入流量，C 为路由器带宽. 端主机在收到包含丢包率的扩展 EMKC 报头后采用如下的控制算法：

$$x_i(t) = \alpha + (1 - \beta p(t - RTT)) x_i(t - RTT).$$

其中 α 和 β 为正常数. 显然，当网络处于欠载时，丢包率为负值，发送端的流量既使用加性增加策略又使用乘性增加策略，发送端的流量将呈指数上升；当网络处于轻度过载时，丢包率变为正值，发送端的流量将既使用加性增加策略又使用乘性减少策略，当加性增加与乘性减少的影响相抵消时，发送端进入稳态传输；如果网络严重过载，乘性减少策略对流量的影响要远大于加性增加策略，此时发送端的流量将呈指数下降.

EMKC 与 XCP 一样，具有高效、公平性好、排队时延小等优点，而且与 XCP 相比较，处理 EMKC 报文为路由器带来的负载要更小. 但是当多个采用 EMKC 的流从最不公平的状态开始竞争带宽资源时，其收敛到公平的时间要远远大于采用 XCP 协议的流. 另外，只有当网络处于轻度过载状态时，EMKC 算法才能到达稳态点，所以使用 EMCK 协议的网络不可避免地存在一定的丢包率，而且网络中的流越多，丢包率越大，而在采用 XCP 的网络中，丢包率几乎为 0.

3.2.3　VCP

VCP 采用了滑模控制机制来调整拥塞窗口，当路由器显式通告的瓶颈链路的利用率低于 80% 时，发送端采用乘性增加策略；当利用率在[80%，100%]之间时，发送端采用加性增加策略；如果链路过载，发送端则使用乘性减少的策略. 但是 VCP 与 EMKC 和 XCP 不同的是，它的优点在于不需要特定的扩展报头，而是直接利用 IP 报头中的冗余位来向端主机通告链路利用率当前对应的区间，从而 VCP 比 EMKC 和 XCP 更容易在实际网络中部署. 但是由于使用了加性策略，VCP 在收敛到公平的指

标方面和 EMKC 一样,也要远远低于实时分配带宽的 XCP.

总的来说,通过路由器提供显式反馈,传输协议的各项指标都能够得到大幅度的提高.但是,毕竟这一类协议需要路由器的支持,这直接影响到协议的可扩展性,改变传统的"端到端"的传输服务模式[13]对于互联网来说是一个巨大的挑战,这一类协议最终能否被作为下一代互联网的标准化传输协议,我们仍将拭目以待.

3.3 基于带宽测量的改进方案

通过对网络带宽的估测,发送端可以了解网络当前的负载状态,从而决定如何调整拥塞窗口.这一类协议既可以通过隐式方式实现也可以通过显式方式实现,但是其核心思想在于对网络状态的测量,所以我们将其作为单独的一类.这里我们简单介绍端到端的 TCP Westwood(TCPW)[14]和需要路由器支持的 Congestion Avoidance with Distributed Proportional Control(CADPC)[15].

3.3.1 TCPW

在 TCPW 中,发送端通过观测返回 ACK 的时间间隔来估计端到端链路上的带宽信息,当网络发生拥塞时就使用估计值来调整拥塞窗口和慢启动门限.发送端通过计算采样时间间隔内发送端成功发送的数据量作为链路的可用带宽 Bwe,当发送端收到三个重复 ACK 时,设置慢启动阈值 $ssthresh = Bwe \times RTT_{\min}$(带宽延时积)且当前拥塞窗口 $w = ssthresh$. 如果发生超时,则令 $ssthresh = Bwe \times RTT_{\min}$且 $w=1$.仿真试验表明 TCPW 可以显著地提高传输效率和公平性,大幅度减少不必要的窗口重传,最重要的是 TCPW 在高速网络、无线网络以及有线和无线混合网络中都能保持有效.

3.3.2 CADPC

在 CADPC 中,其测量过程由辅助的 Performance Transparency Protocol (PTP)协议实现.在数据传输过程中,PTP 测量数据包从发送端发出,并在每经过一个路由器时都由路由器写入路由器标识以及对应的时戳和瓶颈带宽,在经过接收端计算之后,将最终得出的瓶颈负载流量 X 和瓶颈带宽 C 通过 PTP 协议发回到发送端. 发送端最终根据如下的 Logistic 方程来调整发送端流量:

$$x_i(t+1) = x_i(t)\left(2 - \frac{X}{C}\right).$$

由于是通过路由器显示通告带宽信息,CADPC 测量精度要高于 TCPW,但是也像 XCP 协议一样带来了一定的实现复杂度和可扩展性问题.

基于测量实现的传输协议具有对于由于网络误码产生的随机丢包不敏感的优点,避免了传统 TCP 因为频繁进入慢启动以及保守的加性策略而导致的网络传输性能下降.这一类协议研究的关键问题在于如何提高测量精度.在目前的 TCPW 版本中,使用了简单的指数加权滑动平均算法来进行滤波,其测量精度仍然容易受到网络噪声的影响.随着研究的深入,更多的滤波算法已经逐渐被应用到这一领域[16].

3.4 基于应用层的改进方案

前面几类方案都是在传输层进行改进,而且其中一些还需要路由器的支持,所以一直没有得到大规模的实际部署. 因此,研究人员提出了应用层传输协议,其核心思想不在于设计新的应用层流量调整算法,而是通过将数据信息和控制信息分开传输来提高传输效率.譬如在 Simple available bandwidth utilization library(SABUL)[17]中,数据通过 UDP 协议来传输,而确认包通过 TCP 协议来传输,然后通过这些确认信息来控制 UDP 协议的发送流量.而在 UDP-based Data Transfer Protocol(UDT)[18]中,数据和控制信息都通过 UDP 来传输,并通过合理设置定时器来避免当发送端收不到控制信息时陷入死锁.

虽然在应用层改进传输协议比在传输层改进其传输效率可能有所降低,而且因为使用更多的连接而增加了 CPU 的开销,但其可部署性却大大提高了,同时还有利于减少高速网络中确认分组的流量,提高网络带宽资源的利用率. 假设分组大小为 1000Bytes,确认分组大小为 40Bytes,如果传输流的吞吐量为 10Mbps 时,那么使用传统 TCP 所产生确认分组的吞吐量为 400Kbps;如果传输流的吞吐量为 10Gbps 时,那么它所产生确认分组的吞吐量将为 400Mbps;如果采用应用层传输协议中的数据信息和控制信息分离原则,假设每 10ms 发送一个控制分组,控制分组大小为 100Bytes,那么它所产生确认分组的流量仅为 80Kbps,显然这类方案明显地减少了网络中 ACK 的分组流量,并降低了端系统对于 ACK 分组的处理开销.但是这类方案打破了传统 TCP 协议的自同步机制,无疑会增加算法的复杂性,关于应用层的改进方案的性能评价还有待于更深入的研究.

4 开放性问题及进一步的研究方向

在下一代互联网高速传输协议的研究过程中,

研究人员发现了 TCP 协议在进行高速数据传输的缺陷，并提出了各种改进或全新的传输协议，从最初单纯解决 TCP 协议的低效问题，到围绕 TCP 友好性、RTT 公平性以及稳定性开展了一系列更深入的研究，但是到目前为此，在高速网络拥塞控制研究领域仍然存在很多开放性问题，主要体现在：

(1) 高速网络拥塞控制研究仍处在起步阶段，对于多种拥塞判定机制缺乏足够的有效性分析. 例如在文献[19]中就提出了在实际网络中基于排队延时的拥塞判定机制与丢包事件之间没有本质联系的观点，这种现象在所观测流的流量在瓶颈链路中所占比例较低时尤为明显. 基于丢包、基于延时或者其它有效的拥塞判定机制在高速网络传输控制系统中的表现还有待于进一步通过理论分析与验证，并结合实际网络的测量和网络仿真进行更深入的研究.

(2) 目前的高速网络拥塞控制中的多数研究没有充分强调模型分析的重要性，缺乏具有总结性结论和定律的归纳与描述. 在传输机制和控制算法的设计上仅采用依赖于经验的启发式设计加典型、有限和局部的仿真试验验证的设计方法，得到的算法往往是静态的和准静态的，不能适应快速变化的动态网络化环境，在 FAST 中通过参数对照表选取 α 参数的方法就充分暴露了这种设计缺陷，在多种协议中普遍存在的固定门限设计方案也有待进一步改进.

(3) 高速传输协议的设计已经成为一个复杂的多目标优化问题，目前还没有很好的综合解决方案与评价标准. 具体表现为：在保证传输协议具有高效率的同时，还需要传输协议具备良好的快速收敛性、TCP 友好性、RTT 公平性、稳定性以及能够逐步部署的工程可实践性，而目前高速网络传输协议对于这一多目标优化问题没有提出综合的解决方案；在算法性能的评价和验证方面，仅仅依赖典型、有限、局部的仿真试验往往无法科学与本质地刻画高速传输协议的特性，无法得出系统、科学、可信的判断，所以理论分析与证明的过程也不可缺少.

我们认为就网络传输控制而言，模型化工作中的突破，即便是很小的进展，往往会给该领域其他相关方向的研究带来意想不到的启发和促进. 数学模型对于传输控制之所以重要，是因为只有基于一定的模型，我们才有手段通过一定的观测变量较为准确地综合出网络的当前状态，甚至有可能预测其变化趋势，这样采取的控制策略才会有的放矢. 凭借局部经验和启发式算法，对运动规律和行为状态浑然不知的复杂系统进行控制的结果是可以想象的. 当然，建立精确的网络流量模型决非易事，甚至完全不可能，但已有的研究表明：粗线条的近似模型对于某些传输控制问题往往是足够充分的，模型应该是未知的网络行为与确定的传输控制目标之间建立联系的纽带.

因为模型的不完备和有效理论分析方法的欠缺，目前大多数研究采用不依赖模型的启发式算法设计，配合典型仿真实验加以验证的方法，取得了一些局部性的研究成果，但我们认为随着网络规模和复杂性的日益扩大，此方法将会越来越力不从心，其所得结果的局限性也将越发明显. 为寻找可能的、较为彻底的解决方案，在将来的研究中，我们将从体系结构、协议机制和算法实现等各个层次上强调模型和理论分析的重要性，抓住业务流量的突发性、网络状态的时变性和控制作用的滞后性等网络传输控制中的鲜明特征，依赖成熟且可行的理论方法譬如控制论、优化理论和博弈论等来设计高效的传输机制与算法，并使它们具有良好的动态自适应特性，从而克服目前多数启发式静态和准动态算法适应性差的缺陷.

在算法性能的评价和验证方面，仅仅依赖典型、有限、局部的仿真试验往往无法得出系统、科学、可信的判断，理论分析与证明的过程不可缺少，在研究过程中，我们认为应该综合应用流体流、时间序列分析和系统辨识等理论推导、归纳和拟合可以准确或近似描述网络传输控制中本质特性的分析模型，并借助网络测量的结果进行有效性验证.

5　总　结

本文对多种新的高速传输协议进行了分类描述，重点分析了高速网络传输这一热点领域中已有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.

参　考　文　献

1　Floyd S.. HighSpeed TCP for large congestion windows, RFC3649, December 2003

2　Kelly T.. Scalable TCP: Improving performance in highspeed wide area networks. Computer Communication Review, 2003, 33(2): 83～91

3　Allman M., Paxson V., Stevens W.. TCP Congestion Control, RFC 2581, April 1999

4　Padhye J., Firoiu V., Towsley D., Kurose J.. Modeling TCP throughput: A simple model and its empirical validation.

IEEE/ACM Transactions on Networking, 2000, 8(2): 133～145
5 Shorten R. N., Leith D. J.. H-TCP: TCP for high-speed and long-distance networks. In: Proceedings of PFLDnet'04, Chicago, IL, USA, 2004
6 Jin C., Wei D. X., Low S. H.. FAST TCP: Motivation, architecture, algorithms, performance. In: Proceedings of the IEEE INFOCOM, Hong Kong, 2004, 2490～2501
7 Xu L., Harfoush K., Rhee I.. Binary increase congestion control for fast long-distance networks. In: Proceedings of the IEEE INFOCOM, Hong Kong, 2004, 2514 ～2524
8 Katabi D., Handley M., Rohrs C.. Congestion control for high bandwidth-delay product networks. In: Proceedings of the ACM SIGCOMM, Pittsburgh, PA, USA, 2002, 19～23
9 Zhang Y., Kang S., Loguinov D.. Delayed stability and performance of distributed congestion control. In: Proceedings of the ACM SIGCOMM, Portland, Oregon, USA, 2004, 307～318
10 Xia Y., Subramanian L., Stoica I., Kalyanaraman S.. One more bit is enough. ACM SIGCOMM Computer Communication Review, 2005, 35(4): 37～48
11 Ramakrishnan K. K., Floyd S.. Proposal to add explicit congestion notification (ECN) to IP. RFC 2481, 1999
12 Brakmo L. S., Perterson L. L.. TCP Vegas: End-to-end congestion avoidance on a global Internet. IEEE Journal on Selected Areas in Communication, 1995, 13(8): 1465～1480
13 Paxson V.. End-to-end Internet packet dynamics. IEEE/ACM Transactions on Networking, 1999, 7(3): 277～292
14 Gerla M., Sanadidi M. Y., Wang R., Zanella A., Casetti C., Mascolo S.. TCP Westwood: congestion window control using bandwidth estimation. In: Proceedings of the IEEE Globecom 2001, San Antonio, Texas, USA, 2001, 1698～1702
15 Welzl M.. Scalable Performance Signalling and Congestion avoidance. Kluwer Academic Publishers, 2003, 91～137
16 Gerla M., Ng Bryan K. F., Sanadidi M. Y., Valla M., Wang R.. TCP Westwood with adaptive bandwidth estimation to improve efficiency/friendliness tradeoffs. Journal of Computer Communications, 2004, 27(1): 41～58
17 Gu Y., Grossman R. L.. SABUL: A transport protocol for grid computing. Journal of Grid Computing, 2003, 1(4): 377～386
18 Gu Y., Grossman R. L.. UDT: UDP-based data transfer. In: Proceedings of the PFLDnet'04, Chicago, IL, USA, 2004
19 Martin J., Nilsson A., Rhee I.. Delay-Based congestion avoidance for TCP. IEEE/ACM Transactions on Networking, 2003, 11(3): 356～368
20 Widmer J., Denda R., Mauve M.. A survey on TCP-friendly congestion control. IEEE Network, 2001, 15(3): 28～37
21 Ren Feng-Yuan, Lin Chuang, Liu Wei-Dong. Congestion control in IP network. Chinese Journal of Computers, 2003, 26(9): 1025～1034(in Chinese)
(任丰原,林 闯,刘卫东. IP 网络中的拥塞控制. 计算机学报, 2003, 26(9): 1025～1034)
22 Luo Wan-Ming, Lin Chuang, Yan Bao-Ping. A survey of congest ion control in the Internet. Chinese Journal of Computers, 2001, 24(1): 1～18(in Chinese)
(罗万明,林 闯,阎保平. TCP/IP 拥塞控制研究. 计算机学报, 2001, 24(1): 1～18)

HUANG Xiao-Meng, born in 1980, Ph. D. candidate. He is especially interested in congestion and flow control in high speed networks, computer networks performance evaluation, etc.

LIN Chuang, born in 1948, professor, Ph. D. supervisor. His current research interests include computer networks, performance evaluation, logic reasoning, and Petri net theory together with its applications.

REN Feng-Yuan, born in 1970, Ph. D., assistant professor. His main research interests include congestion and flow control in computer networks, robust control theory, mobile code technology and wireless sensor network etc.

Background

This work is supported in part by a grant from the National Grand Fundamental Research 973 Program of China (No. 2003CB314804), and the National Natural Science Foundation of China (No. 60573122, 90412012, 60372019 and 60373013).

Fast long-distance networks are now becoming commonplace. Increasing numbers of researchers now routinely transfer between 10 GB and multi-TB datasets over gigabit networks. Application domains for such massive transfers include data-intensive Grids, database mirroring for Web sites.

Although the connectivity infrastructure is now in place, or will soon be, the transport and application protocols available to date are proving inadequate for fast transfer of large volumes of data over such networks. Current versions of TCP cannot fully exploit the network capacity. For instance, recovery time from a congestion event grows at a super-linear rate, and can easily exceed 10 minutes in very high bandwidth-delay product networks. It also requires a large congestion window for high throughput, consuming valuable system resources. A number of research teams have begun investigating advanced protocols for domain-specific and general applications.