MPLS 网络主动式流量和拥塞控制机制及性能分析 [1]

张志群 丁 炜 邵 旭

(北京邮电大学培训中心 北京 100876)

摘 要 MPLS 是具有大带宽-时延积的网络，用传统的 TCP 解决 MPLS 拥塞问题显得十分困难. 该文结合 MPLS 的网络特点，提出了一种适合 MPLS 网络的主动式流量和拥塞控制机制，在网络边缘节点引入拥塞反馈处理. 对实验模型进行了性能仿真分析，实验证明，与传统的 TCP 协议相比，该机制将流量和拥塞控制从用户端点扩展到 MPLS 边缘路由器，能够更及时地检测和控制网络拥塞，缩短了控制时延，可以进行较精确的流量调节，实现了提高吞吐量和改善缓冲区利用率的目的.

关键词 MPLS, 主动式流量和拥塞控制，带宽-时延积，边缘智能，反向通知树

中图号 TN919.3

1 引 言

流量和拥塞控制机制的目的是实现高吞吐量和低缓冲区容量. 目前 Internet 上广泛使用的拥塞控制协议是 Tahoe TCP[1]，改进协议主要有 Reno TCP[2], NewReno TCP[3] 以及 SACK TCP[4-6]. 这些协议本质上都使用 Jacobson 的拥塞避免机制 [1]，端点发送速率根据包确认由滑动窗口来控制. 窗口的大小根据连接路径中的拥塞情况来调整. 这些协议存在共同的问题：网络拥塞的检测和控制不是由发生拥塞的网络节点及时主动地完成，而是基于端到端的反馈由端点通过各种隐含信号推测出来的. 该方法不但延缓网络拥塞的检测和控制，而且在反馈过程中，仍有大量数据包发向网络，造成更严重的网络拥塞. 该问题的症结在于： IP 网络节点采用“存储-转发”模式，只能完成简单的分组转发，复杂计算处理在端点完成，即“端点智能，节点转发”.

面对 IP 网络对 QoS 的要求，在网络节点适度引入计算能力已显得十分必要，并且网络处理器在节点对协议的处理能力使之成为可能. 多协议标签交换 (Multi-Protocol Label Switch , MPLS) 正是体现了这种技术的变迁，在无连接的 IP 网络引入面向连接的机制，形成 MPLS 域，标签边缘路由器 (Label Edge Router , LER) 具有计算能力，完成分类、调度、 QoS 映射等处理，标签交换路由器 (Label Switch Router , LSR) 只完成简单转发，即“边缘智能，核心交换”.

再则，作为 IP 骨干网技术， MPLS 是具有大带宽-时延积 (Bandwidth-delay product) 的网络. Bolot 和 Shankar 指出，在基于反馈的拥塞控制系统中，链路瓶颈的拥塞持续时间与带宽-时延积直接相关 [7]. 网络端到端的时延越大，端点能够检测到网络发生拥塞的时间就越长. 网络带宽越大，在端点检测到网络拥塞之前，端点发送到拥塞网络中的数据量就越大，网络拥塞进一步恶化. 所以，拥塞控制问题可以归结为减小带宽-时延积问题，甚至是减少反馈时延问题. 因此，用基于端到端反馈的 TCP 来解决 MPLS 拥塞问题，获得大吞吐量，显得十分困难.

基于 MPLS 边缘智能和面向连接的特性，我们借鉴主动式流量和拥塞控制 (Active Traffic & Congestion Control , ATCC) 的思想，利用 MPLS LER 的计算能力，将反馈拥塞算法从端点扩展到了网络边缘节点，缩短拥塞反馈时延，减小带宽-时延积，以提高吞吐量，改善缓冲区利用率.

本文第 2 节，扼要介绍了 ATCC 基本原理；第 3 节，解释了将 ATCC 引入 MPLS 的理由，并阐述了 MPLS ATCC 的基本方法，设计了 MPLS LER 操作策略模型. 第 4 节，对 MPLS

[1] 2001-06-27 收到， 2002-02-18 定稿

863 项目 (863-317-9601-02)， Intel IXA 大学计划资助

ATCC 进行了仿真实验和性能分析。最后，对本文做了小结，并指出下一步的工作方向和研究热点。

2 主动式流量和拥塞控制

主动网也可理解为“IP 的智能网”，它支持网络行为的动态控制和修改。显然，可以利用节点的智能来进行主动式的流量和拥塞控制。

在传统的 TCP 反馈系统中，拥塞释放是随着端点发送速率的降低从端点传播到拥塞节点。在 ATCC 中，拥塞释放是从拥塞节点开始并传播到端点，并且流量的改变也是从节点传播到端点，所以系统是稳定的。在 ATCC 中，端点和网络节点都参与流量和拥塞控制。节点检测到拥塞，立即对拥塞做出反应，改变进入网络的流量，由节点完成拥塞检测与恢复部分工作，可以减少控制时延，从而减少在此时间内端点发出的无用数据包，以免造成进一步拥塞。

研究现有流量和拥塞控制机制，可以发现，这些机制都是由两部分组成——适应性拥塞控制算法和独立的拥塞通知方法。针对不同网络的特点，这两部分的实现位置、形式、内容和处理流程也各不相同。ATCC 算法也分成拥塞控制和通知两部分。其中，拥塞控制算法包含 ATCC Filter 和 ATCC SlowStart 功能模块；拥塞通知包括消息单元 ATCC Indication 和功能模块 ATCC QueueSpy 。

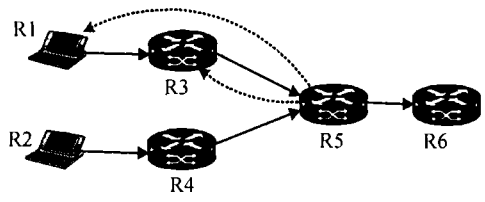

图 1 具有 ATCC 算法的网络拓扑图

图 1 为一个简单的 ATCC 控制算法的网络拓扑图 [8] 。R1 和 R2 为端点，具有 ATCC SlowStart 功能。R3–R6 为网络节点，其中，R3，R4 带有 ATCC Filter，R5 具有 ATCC QueueSpy 功能。图中，实线表示数据流的流向，虚线表示反馈的 ATCC Indication 消息。实际应用中，网络节点应同时具备 ATCC Filter 和 ATCC QueueSpy，这里简化结构是为了讨论方便。

ATCC Indication 消息单元用于在 ATCC 各功能模块间传送有关网络拥塞指示消息 [9,10] 。它包括两个部分：数据包序号和数据流标识。数据包序号用于表示在拥塞节点中被丢弃的数据包在所属的数据流中的 TCP 顺序号；数据流标识用于表示在拥塞节点被丢弃数据包所属的数据流，ATCC 利用被丢弃包 IP 的源 / 目的地址和 TCP 的源 / 目的端口号来唯一地标识网络中的数据流。与 ICMP 的源抑制策略不同的是，ATCC Indication 将要求上游节点主动进行流量控制，以尽快地减轻拥塞节点的负荷。若 ATCC Indication 在传输过程中丢失，则用户终端仍可按传统 TCP 进行拥塞控制。

ATCC QueueSpy 和 ATCC Filter 都位于网络节点，由用户或网络业务提供商通过主动式网络的业务动态加载服务将这两个功能模块加载至网络节点上 [11] 。ATCC QueueSpy 主要功能为网络节点中的队列监视，当通过节点的数据流负载达到一定程度，可能或已经导致节点拥塞时，ATCC QueueSpy 结合节点中的队列管理算法对数据流进行相应的控制。当 ATCC QueueSpy 决定对造成网络节点拥塞的数据流进行控制时，它向数据流的上游节点和源端发送 ATCC Indication 消息，要求它们分别进行相应的数据流控制。

ATCC Filter 主要功能为响应数据流下游节点发出的 ATCC Indication，主动地对造成下游节点拥塞的数据流进行控制。ATCC Filter 可采用的控制策略包括丢弃、重定向和缓存等方法。ATCC SlowStart 位于产生数据流的端点，它结合传统的 TCP 对 ATCC Indication 消息进行响应，重发在拥塞节点丢失的数据包，并进入 TCP 的 SlowStart 状态。

显然，在现有的无连接 IP 网络中引入 ATCC 机制，要在所有网络节点引入复杂的路由和流量计算，即“网络智能”，这与 IP 简单实用的原则是矛盾的。但是，在 MPLS 中，边缘节点

具有一定的计算能力，能够进行拥塞控制和流量编辑；同时，其面向连接的特性也可以迅速反馈控制信息，避免复杂的路由计算，所以，在 MPLS 域内引入 ATCC 机制是可行的。

3 MPLS 网络中主动式流量和拥塞控制机制

3.1 MPLS ATCC 机制

具有 ATCC 机制的 MPLS 网络拓扑如图 2 所示。MPLS 域中网络节点由 LERs 和 LSRs 组成。LERs 主要完成 MPLS 域与非 MPLS 域连接，以及不同 MPLS 域之间连接的功能，实现业务分类，分发 / 剥离标签，确定业务类型、策略管理以及接入流量工程控制等工作，具有较强的计算能力。ATCC Filter 等主要流量控制处理功能在此实现。LSRs 位于 MPLS 域内部，主要运行 MPLS 控制协议和第三层路由协议，完成标签交换和数据转发。ATCC QueueSpy 和 ATCC Indication 在此实现。

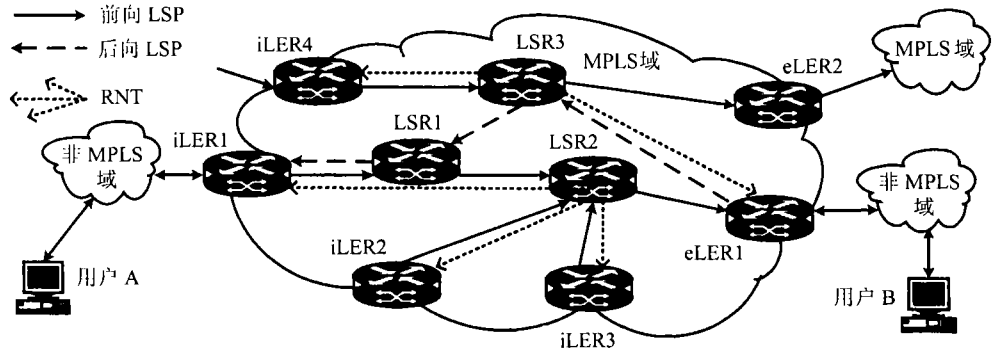

图 2　MPLS ATCC 网络拓扑图

在 MPLS 域内，入口 LERs 根据地址和 QoS 信息，将数据流映射成相应的转发等价类 (FEC)，并与固定长度的短标签 (Label) 绑定起来，将之插入数据分组头。数据流传送路径则由第三层路由协议、用户需求以及网络状态共同决定。通常，各个 MPLS 设备运行路由算法，根据计算得到的路由，在逻辑相邻的对等体之间进行标签分配，通过入口 (Ingress)LER，LSRs，出口 (Egress) LER 的标签链接建立起面向连接的标签交换路径 LSP(Label Switch Path)。LSP 是单向的。即两个端点建立的前向 LSP 和后向 LSP 在 MPLS 域内可能经过不同的 LSRs，但是，在 MPLS 边缘都要经过相同的 Ingress LER 和 Egress LER。在 MPLS 域内，LSRs 只根据标签进行交换，数据流沿着 LSP 交换到出口 LERs，将标签剥离，按照传统的 IP 协议转发 [12]。

图 2 描述了 MPLS ATCC 流程。当 LSR2 流量负载达到一定程度，LSR2 的 ATCC QueueSpy 监测到网络可能或者已经发生拥塞时，它立即发送 ATCC Indication 消息给向 LSR2 输入数据流的上游节点 iLER1，iLER2 和 iLER3。这些 Ingress LERs 收到下游 LSR2 发来的拥塞通知消息后，启动 ATCC Filter，主动地对影响下游节点拥塞的数据流进行编辑或丢弃，避免网络拥塞进一步加剧，直到端点对拥塞做出反应；同时，Ingress LERs 向端点发出拥塞通知，端点收到通知后，结合传统的 TCP 协议，重发在拥塞节点丢失的包，并进入 TCP 的 SlowStart 状态。从而，及时地预测和缓解拥塞状况，精确地调节流量。在此过程中，充分利用 LERs 的计算能力，可对流量进行较复杂的编辑处理，LSRs 只需增加拥塞监测和通知模块，不会造成过多网络负荷。

建立快速的信息反馈路径是实现 MPLS ATCC 的关键之一. 如前所述, LSP 是单向的, 所以在发生拥塞的 LSRs 与 Ingress LERs 间必须建立面向连接的反向路径, 用于传递消息. 由于 MPLS 具有标签合并功能, 多条前向 LSPs 的汇聚点称之为 PML(Path Merge LSR), 这样就形成了一棵以 PML 为根, Ingress LERs 为叶的多点到点的反向通知树 (RNT), 消息可以沿 RNT 从根节点反馈到末梢 [13]. 实际上, 每个 LSR 都可以拥有一棵以自己为根, Ingress LERs 为叶的 RNT, 这样, 以流量关系将 LSRs 与 LERs 的 Ingress 流向脉络组织起来, 一定程度上避免了 "令出多门" 现象. 图 2 中, 分别以 LSR2、LSR3 为根形成两棵 RNTs. RNT 的建立有以下优势:

(1) MPLS 是面向连接的, 只要求在前向 LSP 的每个 LSRs 中保存它的上游节点, 就可以沿原路径建立 RNT, 实现简单.

(2) 对于会聚到同一点 LSR 的所有 LSP, 只需要一棵 RNT, 在树丫点的 LSRs 的子树也依附在主树上, 从而减少了拥塞通知的信令开销.

(3) 通过 RNT, 控制消息从根部沿着原来的工作路径反向传送到每一个可能造成拥塞的 Ingress LERs 和端点, 而不会波及到其他未向拥塞点产生流量的 LERs 和端点, 即谁造成拥塞谁负责, 公平性更好.

(4) 拥塞反馈信息更加有序、显式、确定性和可靠, 避免 ATCC 中反馈的不确定性和无序.

3.2 MPLS LER 的策略模型

MPLS LERs 的操作策略模型如图 3 所示. 操作模型的组成部件包括: 触发器, 负责激活和初始化拥塞控制; 流状态, 分析流的目的地址及特定的 LSP 信息; 滤波器, 根据流状态信息对数据流进行编辑; 通知器, 向用户发送拥塞通知信息 [14].

在 MPLS 域入口 LERs, 根据 FEC 等价映射具有相同标签的分组序列被称之为流. 为了与 IP 遗留系统的传统拥塞控制方法兼容, 可以设定策略选择, 对数据流选择不同的操作策略. 选择 ATCC 策略的数据流, 将流的目的地址和 LSP 信息写入流状态. 若域内 LSRs 发生拥塞, 拥塞通知消息会经 RNT 触发该 LER 的触发器. LER 被触发后, 设定滤波器, 根据反馈信息检查流状态, 对影响拥塞节点的流量进行编辑, 对进入网络的流量进行编辑, 直至收到拥塞释放消息; 同时, 滤波器根据流量编辑情况通知用户, LEP 作为对等实体与用户端点按照 TCP 方式进行流量控制.

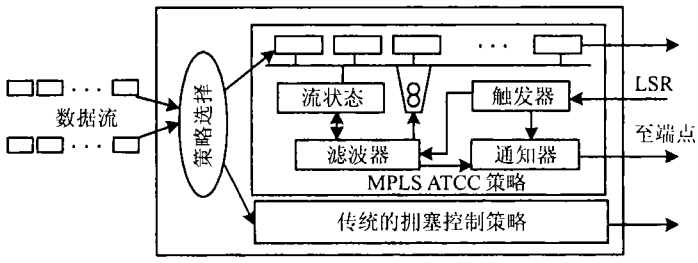

图 3 MPLS LER 操作策略模型

4 算法仿真与性能分析

4.1 仿真配置

下面, 我们通过仿真实验来比较 MPLS ATCC 与 TCP 在 MPLS 网络中的性能差异. 仿真实验网络拓扑如图 4 所示. 图中, n 个端点产生 Cross Traffic 数据流, 流过中间网络节点到达

对应的 Cross Traffic Sinks。m 个端点发送 Bulk Traffic 数据，流过中间网络节点到达对应的 Bulk Traffic Sinks。实验中，设 m，n=20，所有的端点均使用 1000byte 的包。中间节点的 Buffer 可存 25 个包，使用 FIFO Drop Tail 排队 [15]。

每次仿真从 R2 到 R3 的链路上都用不同的时延。链路时延在 20ms 到 300ms 范围内变化，这个范围涉及的带宽-时延积模型可以模拟从小型的局域网、宽带宽的广域网到窄带宽、长时延的卫星网络。基于此，我们通过仿真实验来证明 MPLS ATCC 在较大带宽 - 时延积范围的网络内都是有效的。

这里，利用 Cross Traffic 来模拟非受控数据流，如对拥塞无反应的突发视频流。Cross Traffic 为服从指数分布的通断型数据流，其通断状态平均维持时间为 2s，通状态时，其数据速率为 200kb/s。

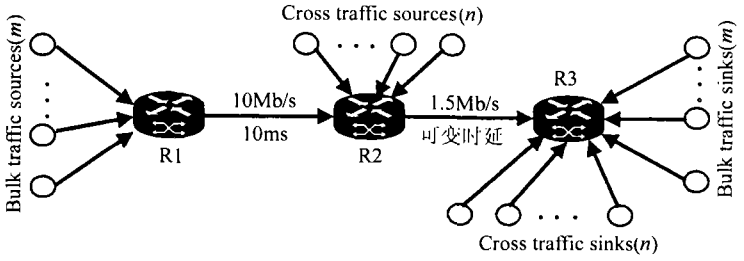

图 4　仿真实验网络拓扑图

4.2 MPLS ATCC 的性能分析：吞吐量

解决拥塞的最终目的是为了提高吞吐量，所有的吞吐量都是参照目的端点收到的有用包来计算的，重传的包不被考虑为有用。在不同时延下（直接对应着带宽-时延积的变化），只存在 bulk traffic sources 所经历的平均吞吐量如图 5 所示。仿真结果是 20 个 bulk traffic sources 吞吐量的均值。因为 MPLS ATCC 和 TCP 都依赖于网络节点和端点的反馈通信，故两者的性能都随着带宽-时延积的增大而下降。由于 MPLS ATCC 是中间节点进行反馈，其性能优于传统 TCP 性能，能提高约 25% 的吞吐量。两者在较小的带宽-时延积情况下表现出相似的性能，这是因为当网络反馈时延较小时，MPLS ATCC 节省的反馈时延被其网络节点处理所消耗。

在大带宽-时延积情况下，MPLS ATCC 减少了拥塞持续时间和受影响的端点。当瓶颈节点变得拥塞，MPLS ATCC 流量立即减少。结果，拥塞的节点在更短的持续时间内不加选择地丢包，这也就意味着更少的端点体验到拥塞。通过限制端点减少的最终数目，总的吞吐量仍保持很高。

在图 5 条件上，叠加 Cross Traffic Source 之后，端点的平均吞吐量如图 6 所示，MPLS ATCC 性能受 Cross Traffic 影响较小，TCP 协议性能则受影响较大。这是因为当网络节点发生拥塞时，由于 Cross Traffic 对拥塞无反应，其发送速率不会降低。MPLS ATCC 能够确保端点对拥塞迅速做出反应，进入 SlowStart 状态，并且 Filter 主动丢弃可能进一步造成拥塞的数据包，及早缓解网络拥塞，提高吞吐量。随着链路时延的加大，MPLS ATCC 也会出现性能急剧下降，接近 TCP 协议，这是由于 MPLS 域内拥塞通知的时延已接近较小局域网端点的反馈时延。

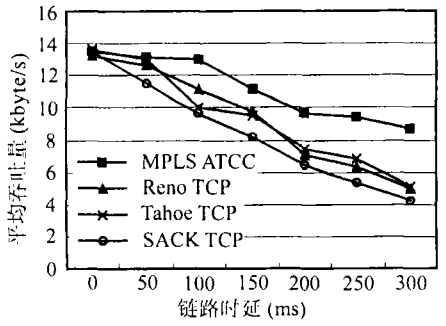

图 5 只存在 Bulk traffic 的平均吞吐量

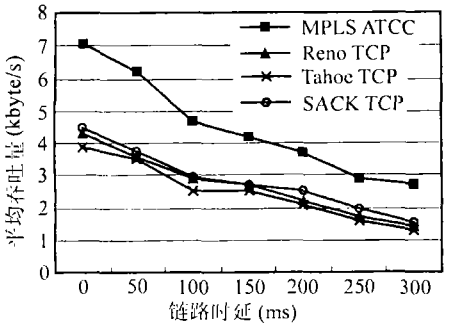

图 6 叠加 Cross traffic 后的平均吞吐量

4.3 MPLS ATCC 的性能分析：缓冲区利用率

在保证高吞吐量的同时，提高缓冲区利用率也是衡量算法性能的重要指标。假定 TCP 协议缓冲区空间有限，网络存在突发业务流， Finite-Buf-TCP 和 MPLS ATCC 的缓冲容量都为 496， 080bit，使用 FIFO Drop Tail 排队。其中，图 7 对应链路时延为 20ms，图 8 对应链路时延为 100ms 。

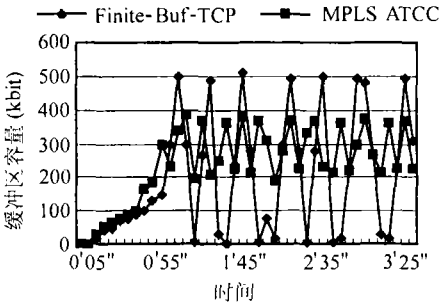

图 7 数据缓冲区占有量，时延 =20ms

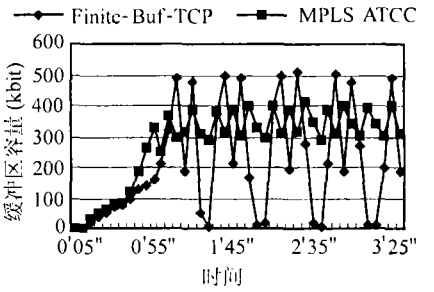

图 8 数据缓冲区占有量，时延 =100ms

从图 7 可看出，控制时延短的 MPLS ATCC 的队列抖动范围在 300kbit 到 400kbit 之间，控制时延长的 TCP 的队列抖动范围在 5kbit 到 450kbit 之间。控制时延越长，缓冲区抖动幅度就越大，并且在同一门限下不丢包所需的缓冲区容量也越大。这是由于 MPLS ATCC 可以根据拥塞情况，在 LERs 设立 Filters 调节进入网络的流量，避免拥塞的加剧，减少了窗口的降低程度，从而提高了缓冲区的利用率。而 Finite-Buf-TCP 通过 ACK 中的窗口通告信息来控制发送端的滑动窗口，由于丢包及窗口的降低需要较长时间才能恢复，增加了缓冲区为空的时间，造成缓冲区利用率下降。

比较图 7 和图 8 ，链路时延为 20ms 的 MPLS ATCC 的队列抖动幅度小，抖动频率为 12.5s ；链路时延为 100ms 的 MPLS ATCC 的队列抖动幅度大，抖动频率为 13.5s 。即控制时延越短，缓冲区的抖动幅度就越小，并且抖动频率越快，缓冲区利用率越高。这是由于反馈时延越短， MPLS ATCC 调整流量越频繁，滑动窗口的变化越快，缓冲区适应窗口变化也越快。Finite-Buf-TCP 在两种情况下也符合这种规律，从另一个侧面证明了控制时延对缓冲区利用率的影响。

5 结 束 语

传统的流量和拥塞控制机制在高速网络中存在诸多问题：低吞吐量，控制时延大，公平性问题，缺乏对 QoS 的支持等。

MPLS ATCC 充分利用 MPLS 网络边缘智能和面向连接的特点，缩短控制时延，提高了吞吐量，改善缓存区的利用率，在大带宽-时延积网络中实现了低缓存容量和大吞吐量的目标。

MPLS ATCC 机制还有亟待发展的地方，如利用 LERs 统计预测网络流量，精确调节流量等。MPLS 面向连接的特性，使得资源预留和显式速率成为可行。因此利用 MPLS 技术来解决 IP 网络的流量工程和 QoS 问题，将是未来研究的热点。

参 考 文 献

[1] V. Jacobson, Congestion avoidance and control, Proc. SIGCOMM Symposium on Communications Architectures and Protocols, ACM SIGCOMM, Stanford, CA. Aug. 16-19, 1988, 314-329, available from ftp://ftp.ee.lbl.gov/papers/congavoid.ps.Z.

[2] V. Jacobson, Modified TCP congestion avoidance algorithm, end2end-interest mailing list, April 30, 1990.

[3] S. Floyd. ACIRI, T. Henderson, The New Reno modification to TCP's fast recovery algorithm, RFC2582 Apr.1999.

[4] S. Floyd, Issues of TCP with SACK. Technical report, Mar. 1996, URL ftp://ftp.ee.ibl.gov/papers/issues sa.ps.Z

[5] S. Floyd, SACK TCP: The sender's congestion control algorithms for the implementation "sack1" in LBNL's "ns" simulator (viewgraphs), Technical Report, Mar. 1996. URL ftp://ftp.ee.lbl.bov/talks/sacks.ps

[6] M. Mathis, J. Mahdavi, S. Floyd, A. Romanow, TCP selective acknowledgment Options, RFC 2018, 1996.

[7] J. C. Bolot, A. Shankar, Dynamical behavior of rate based flow control systems, Comp. Commun. Rev., ACM SIGCOMM, 1996, 26(2), 5-18.

[8] 王斌，刘增基等，前向主动拥塞控制算法及性能分析，电子学报，2001, 29(4), 483-386.

[9] D. Murphy, Building an active node on the Internet, [M. Eng. Thesis], MIT, June 1997.

[10] D. J. Wetherall, D. L. Tennengouse, The Active IP Option. Proc. 7th SIGOPS Euro. Workshop. 1996, 86-95.

[11] D. Wetherall, Ulana Legedza, J. Guttag, ANTS: A toolkit for building and dynamically deploying network protocols, IEEE Network, 1998, 12(3), 12-20.

[12] 石晶林，丁炜等，MPLS 宽带网络互联技术，北京，人民邮电出版社，2001.3, 第 3 章.

[13] Fast Restoration of MPLS Label Switched Paths: draft-shew-lsp-restoration-00.txt.

[14] 赵键等，一类基于主动网络的网络拥塞控制策略，通信学报，2000, 21(7), 33-38.

[15] Theodore Faber, ACC: Using active networking to enhance feedback congestion control mechanisms, IEEE Network, 1998, 12(3), 61-65.

AN ACTIVE TRAFFIC AND CONGESTION CONTROL MECHANISM IN MPLS

Zhang Zhiqun Ding Wei Shao Xu

(*Beijing University of Posts & Telecommunications, Beijing* 100876, *China*)

Abstract Bandwidth-delay product in MPLS is so high that TCP protocols hardly solve the congestion in MPLS. This paper proposes an Active Traffic and Congestion Control (ATCC) mechanism in MPLS and its model, which utilities the characters of MPLS. MPLS ATCC moves the endpoint congestion control algorithms into the network. The comparison with traditional TCP shows that MPLS ATCC can find and control the congestion in time, reduce the round-trip delay, and edit the ingress traffic. So MPLS ATCC improves the total throughput and the utility of buffer.

Key words MPLS, Active traffic and congestion control, Bandwidth-delay product, Edge intelligence, Reverse notification tree

张志群： 男，1973 年生，博士生，从事 ATM, IP, MPLS 的交换结构和 QoS 研究.

丁 炜： 男，1935 年生，博士生导师，教授，从事 ATM, IP, MPLS 研究.

邵 旭： 男，1973 年生，博士生，从事 ATM, IP, MPLS 的网络管理和 QoS 研究.