高技术通讯 2009 年 第 19 卷 第 8 期:811 ~ 816

doi:10.3772/j.issn.1002-0470.2009.08.008

一种 P2P 文件共享系统汇聚拥塞控制机制①

李 伟②* 陈山枝* **

(*北京邮电大学网络与交换国家重点实验室 北京 100876)

(**电信科学技术研究院 北京 100083)

摘 要 针对现有 P2P 文件共享系统采用并发多连接的文件传输方式,过分占用网络带宽资源,导致其它传统互联网业务性能低下的问题,提出了一种 P2P 文件共享系统汇聚拥塞控制机制(ACCM)。ACCM 采用应用层网络测量技术感知节点接入网链路拥塞状况,依据网络拥塞状况动态地调整 P2P 文件共享系统并发文件传输连接窗口,在最大化网络带宽利用率的基础上实现对传统互联网应用的友好性。网络实验结果表明,在网络拥塞发生时,ACCM 能够促使 P2P 文件共享系统并发连接窗口主动退避,实现和传统互联网应用的和平共处;在网络空闲时,ACCM 能够促使 P2P 文件共享系统扩大并发连接窗口,提高网络带宽资源的利用率。

关键词 P2P, 文件共享系统, 汇聚拥塞, 拥塞控制

0 引 言

在 P2P(peer-to-peer)文件共享系统如 BitTorrent[1]、KaZaA[2]中,节点建立多个并发传输控制协议(TCP)连接,同时下载或上传不同的文件分片。这种多点对多点的文件传输方式极大地提高了 P2P 文件共享系统的文件传输效率,同时也提高了网络带宽的利用率,但对传统的拥塞控制(congestion control)机制带来了挑战。传统拥塞控制机制的目标只在于让每个文件传输连接 TCP 友好,认为如果 N 个会话共享瓶颈链路,每个应该分得 $1/N$ 链路传输能力。现有的 P2P 文件共享系统为了加快文件下载速度,通过不断增加连接数来抢占大量带宽资源,极大地增加了底层传输网络的负担,使网络拥塞状况日益严重[3],导致其它传统互联网应用性能和服务质量急剧下降。

为了解决 P2P 文件共享业务所引起的公平性问题,本文提出了一种 P2P 文件共享系统汇聚拥塞控制机制(aggregate congestion control mechanism, ACCM)。传统的传输层拥塞控制机制关注单个连接之间的公平性,而 ACCM 强调互联网应用之间的友好性。ACCM 采用应用层网络测量技术感知节点接入网链路的共享拥塞状况,基于不同的网络拥塞状况动态地调整 P2P 文件共享系统并发文件传输连接窗口大小,在最大化网络带宽利用率的基础上实现对其它传统互联网应用的友好性。

1 相关工作

互联网服务供应商(ISPs)为了缓解 P2P 流量对网络的影响,主要采用了以下三种策略:(1)端口封锁。早期 P2P 软件采用固定端口进行通信,很容易控制其流量;(2)拦截 P2P 连接。通过特征码识别出 P2P 连接后,发送复位(RST)包阻断部分属于 P2P 的 TCP 连接;(3)利用 TCP 协议内在机制,对 P2P 连接传送的 IP 包进行概率丢弃,人为地制造网络拥塞现象,触发 TCP 拥塞避免机制,使之自动缩减 TCP 发送窗口,从而使连接速率降低,达到减少 P2P 流量的目的。

上述策略能够有效控制由 P2P 文件共享系统所带来的网络拥塞,但需添置新设备或升级已有设备,对 ISPs 而言,意味着巨大的网络升级开销。因此,针对此类采用并发多连接的业务,很多研究者已经提出了许多基于用户端的汇聚拥塞控制策略。拥塞管理器(congestion manager, CM)[4]对汇聚流使用了一种加性增加乘性减少(additive increase multiplicative decrease, AIMD)的拥塞窗口调整循环来实现与

① 国家自然科学基金(60672086,60502037)和 863 计划(2006AA01Z229)资助项目。

② 男,1983 年生,博士生;研究方向:P2P 网络,IP 网络路由技术;联系人,E-mail: Python.Gozap@gmail.com

(收稿日期:2008-05-27)

公平相结合的吞吐量。协作协议(coordination protocol,CP)[5,6]采用了速率随数据包二次抽样进行调整的函数方法,实现带宽的公平共享。多服务接入平台(multiservice access platform,MPAT)[7]保存了多个带宽评估循环,允许应用将带宽分配到不同的流,同时确保总吞吐量的公平性。Hacker 等人提出了一种方案[8]试图在获取高输出的同时获取好的公平性。这篇论文中使用的一种方法是将长虚拟路径往返时间(RTT)同并行 TCP[8]流相结合,从而在未充分利用带宽的网络中提高了带宽利用率,同时获取了公平性。竞争且周到的拥塞控制协议(competitive and considerate congestion control protocol,4CP)[9]利用效用函数计算丢包率的目标最优值,通过调整拥塞窗口来实现当前丢包率和目标丢包率的均衡,通过目标丢包率的设置来改变对网络拥塞状况的敏感度。基于内嵌测量背景模式的传输控制协议(inline measurement TCP background mode,ImTCP-bg)[10]使用了一种内嵌网络测量技术,这种技术使用修改的 TCP 协议来测量网络的可用带宽。基于测量得出的可用带宽,ImTCP-bg 对数据包的传输速率进行调整。

与 ISPs 所采用的控制方法相比,上述汇聚拥塞控制策略都在用户端实现,不会给 ISPs 带来开销。但这些策略的根本缺点是采用跨层设计思想,需要对 TCP 协议进行修改,尽管大量实验结果证明了上述控制策略的有效性,但在实际应用中却难以部署。因此,迫切需要一种新的汇聚拥塞控制机制。

2 汇聚拥塞控制机制

如图 1 所示,本文提出的 ACCM 机制由两部分组成:接入网链路拥塞检测机制和并发连接窗口控制机制。

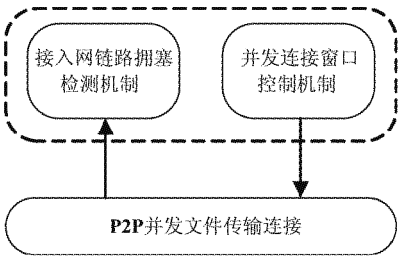

图 1 ACCM 机制框架

接入网链路拥塞检测机制基于应用层测量技术,通过检测并发连接所流经的物理网络路径(*Path*)之间是否共享拥塞来判断节点接入网链路拥塞状况。并发连接窗口控制机制类似于 TCP 滑动窗口机制,依据接入网链路拥塞状况在应用层动态调整并发连接窗口。当节点接入网链路发生拥塞时,不仅底层的 TCP 传输窗口要回退,应用层的并发文件传输连接窗口也要随加性增加适应性减少(additive increase adaptive decrease,AIAD)策略减小。因此,采用 ACCM 机制的 P2P 文件共享系统能够保证对其它传统互联网应用的公平性。

以下将详细介绍接入网链路拥塞检测机制及并发连接窗口控制机制。

2.1 接入网链路拥塞检测

共享拥塞检测[11-14]是现在的研究热点。其基本技术基于对以下事实的观察:如果路径之间共享一条或多条拥塞链路,那么两条路径的测量时延之间呈强相关性;如果它们不共享任何一条拥塞链路,则呈现弱相关性。与其它技术相比,它使用更少的包却提供了更快的收敛性及更高的精确性。

本文用 $XOCR(x, y)$ 来表示路径 $Path_x$ 和路径 $Path_y$ 之间单向延迟序列之间的协相关系数,其计算公式如下:

$$XOCR(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \times \sum_{i=1}^{n}(y_i - \bar{y})^2}} \quad (1)$$

当 $XOCR(x, y)$ 值为 1 时,路径 $Path_x$ 和路径 $Path_y$ 之间存在共享拥塞链路;当 $XOCR(x, y)$ 值为 0 时,路径 $Path_x$ 和路径 $Path_y$ 之间不存在共享拥塞链路。

但是,现有的共享拥塞检测算法只能检测两条路径之间是否共享拥塞链路,却无法精确定位拥塞所处的链路。因此本文首先分析 P2P 文件共享系统的文件传输方式,然后给出基于共享拥塞检测的接入网拥塞检测算法。

在 P2P 文件共享系统如 BitTorrent[1]、eMule[2]中,传输文件通常被分割为大小固定的分片(BitTorrent 和 eMule 均为 256K)分布在 P2P 网络中。当节点下载某个文件时,它将与多个拥有该文件不同分片的其它节点建立并发 TCP 连接,同时下载或上传不同的文件分片。

如图 2 所示,节点 P 的不同 TCP 连接(分别为 $TCP_1, TCP_2, \cdots, TCP_M$,共 M 条 TCP 连接)所流经的物理路径(分别为 $Path_1, Path_2, \cdots, Path_N, N \leqslant M$),具有相同的源点 P 和不同的目的节点(分别为 $D_1, D_2, \cdots, D_N$)。当某条物理路径($Path_y$)发生丢包事

件时,意味着 $Path_y$ 上的某段链路出现拥塞状况。此时如果 $Path_y$ 与经过节点 P 的其它 $N-1$ 条物理路径彼此之间都存在共享拥塞链路,那么此拥塞链路必定是节点 P 的接入链路,因为 P 的接入链路是它们唯一彼此之间都共享的链路。但是如果此时 $Path_y$ 仅和另外 $K(K < N-1)$ 条路径存在共享拥塞链路,那么拥塞链路不是节点 P 的接入网链路,网络拥塞应该发生在 $Path_y$ 和其它 K 条路径之间的共享链路。如果这些路径彼此之间都不存在共享拥塞链路,那么 P 的接入链路必定处于空闲状态。基于上述推理,本文给出接入网拥塞状况标识(flag)的定义如下:

$$\text{flag} = \sum_{x=1}^{N} XCOR(x, y) \tag{2}$$

其中,y 表示在检测过程中出现丢包事件的路径 $Path_y$,x 表示其它路径编号。

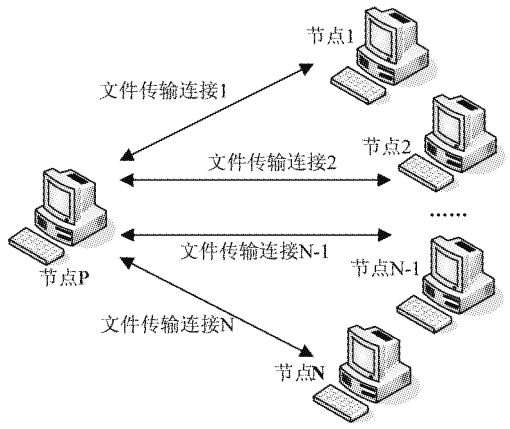

图 2　P2P 文件共享系统并发下载

接入网链路拥塞检测算法包括两个部分:采样和处理。在采样阶段,节点 P 将从 N 个连接中随机选择 M 条目的节点不同的路径,向每个目的节点发送一组(K 个)带时间戳的用户数据报协议(user datagram protocol,UDP)包序列,起始时间为发送节点的时钟时间 T_0。这样的 UDP 包我们称为探测包(probe packet)。在从 T_0 到 $T_0 + T$ 这段时间内,探测包以固定速率发送,T 为探测间隔。目的节点一旦接收到探测包,立刻计算单向时延,并将计算结果和原始时间戳一起发送回节点 P。节点 P 将此单向时延和原始时间戳一起,记录为一次时延采样。当收到来自目的节点的最后一个时延采样时(或最后一个探测包超时,或应答丢失),采样过程结束。最后,节点 P 记录每条物理路径的探测包丢包率及每个探测包的往返时延。

在处理阶段,若这 M 条路径探测过程中没有探测包丢失,说明节点 P 的接入网链路处于空闲状态。否则,则意味着网络中某段链路出现拥塞。假设此时路径 $Path_y$ 丢失探测包,则依据公式(1)计算路径 $Path_y$ 与其它路径之间的协相关系数,获得总计结果(公式(2))。若总计结果为 0,说明 $Path_y$ 与其它路径之间不存在共享拥塞,此时节点 P 的接入链路处于空闲状态。若总计结果为 M,说明 $Path_y$ 与其它 $M-1$ 条路径之间全部存在共享拥塞链路,也就意味着节点 P 的接入链路发生了拥塞。如果总计结果为 $k(0 < k < M)$,$Path_y$ 和其它 $M-1$ 条路径中某些路径之间存在共享拥塞链路,网络拥塞发生在其它链路但不是节点 P 的接入网链路,此时节点 P 的接入链路处于相对稳定的状态。

2.2　并发连接窗口控制

P2P 文件共享系统采用单一 ACCM 实例来控制它所有的数据传输连接。ACCM 为 P2P 文件共享系统维护一个并发数据传输连接窗口,只有当网络处于空闲状态时,它才允许 P2P 文件共享系统建立新的数据传输连接。

ACCM 并发连接初始窗口大小为 $W_0(W_0 > 1$,为保证其传输效率高于其它采用单 TCP 连接的业务,一般取 2),最小并发连接窗口数为 $W_{\min}$(一般 $W_{\min}$ 等于 W_0),最大并发连接窗口数为 $W_{\max}$($W_{\max}$ 一般小于 64),当前并发连接窗口数为 W_{curr}。在感知节点 P 的接入网链路拥塞状况后,类似于 TCP 窗口控制[15],ACCM 采用 AIMD 策略动态调整当前连接的窗口 W_{curr}。当节点的接入网链路空闲时,ACCM 将 W_{curr} 增加 Δ,允许 P2P 文件共享系统建立更多的文件传输连接,以提高数据传输效率以及接入网链路带宽利用率。当节点的接入网链路拥塞时,ACCM 便将 W_{curr} 减小 Δ,以阻止更多的 P2P 文件传输连接进入共享拥塞链路,以保证对其它互联网业务的公平性。在其它情况下,节点接入网链路处于稳定状态,ACCM 并发连接窗口 W_{curr} 将保持不变。

3　网络实验和性能分析

本节将对 ACCM 汇聚拥塞控制机制进行性能评价。本实验中,我们基于 XBT[16](一种开源 BitTorrent 协议实现,用 C++ 语言编写)实现了 ACCM 机制,并对标准 BitTorrent 系统和采用 ACCM 机制的 BitTorrent 系统就对传统互联网业务的公平性和网络带宽资源利用率两方面进行了对比分析。

3.1 实验环境

实验网络由 12 台 PC 机组成，为了网络实验能够运行更多的节点，我们为每台 PC 配置两个 IP 地址，并将每个 IP 地址和一个节点相绑定。因此，实验网络中便存在 24 个节点，它们之间由三台 Cisco 路由器相连，组成了如图 3 所示的拓扑结构。每个节点的上行链路带宽设置为 500kbps，下行链路带宽设置为 1Mbps，核心网链路传输时延设置为 30ms。BitTorrent 系统的 Tracker 服务器部署在节点 $Peer_9$，种子节点为 $Peer_{10}$，种子文件大小为 500MB。文件传输协议(FTP)服务器部署在节点 $Peer_{20}$，下载文件大小为 500MB。每个节点分别部署标准 BitTorrent 客户端和 BitTorrent + ACCM 客户端，FTP 客户端只部署在节点 $Peer_1$。

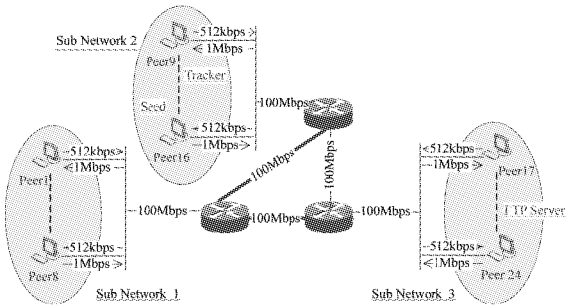

图 3 网络实验拓扑

ACCM 协议中探测报文探测间隔设置为 1 次/s，探测报文数量设置为 10 个/次。P2P 并发连接窗口初始值(W_0)设置为 2，窗口最小值(W_{min})设置为 2，窗口最大值(W_{max})设置为 32，AIAD 策略控制参数 Δ 设置为 2。

3.2 公平性分析

图 4 显示了节点 $Peer_1$ 的 FTP 业务与标准 BitTorrent 业务吞吐量。在 200s 之前，网络中只有 FTP 流量，节点 $Peer_1$ 的 FTP 业务吞吐量非常稳定，占据了一半左右接入网链路带宽。当标准 BitTorrent 业务启动后，由于多点对多点传输方式的侵略性，其吞吐量增长迅速，且随着并发连接的不断增长，标准 BitTorrent 业务抢占了节点 $Peer_1$ 大部分接入网带宽资源。相比标准 BitTorrent 业务，FTP 业务的吞吐量一直下降，到最后 FTP 业务已经基本不可用。

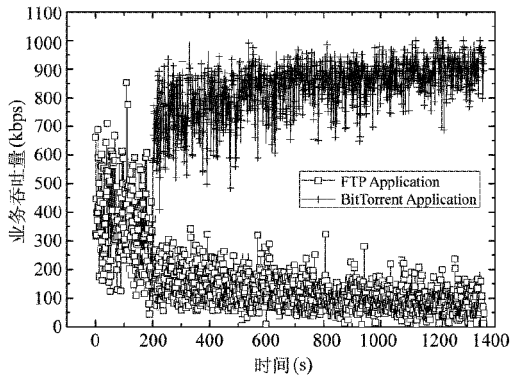

图 4 FTP 和标准 BT 吞吐量对比

图 5 显示的是节点 $Peer_1$ 的 FTP 业务与 BitTorrent + ACCM 业务的吞吐量。在 200s 之前，实验网络中只有 FTP 流量。此时节点 $Peer_1$ 的 FTP 业务的吞吐量非常稳定。在启动 BitTorrent + ACCM 业务之后，FTP 业务的吞吐量虽然同样有所减少，但是依然十分稳定。这意味着 ACCM 机制在一定程度上保证了 BitTorrent 业务对 FTP 业务之间的公平性。

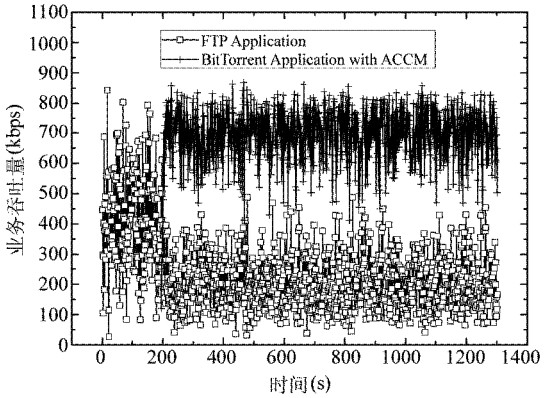

图 5 FTP 和 BT + ACCM 吞吐量对比

3.3 网络利用率

在使用标准 BitTorrent 系统与 BitTorrent + ACCM 系统时，节点 $Peer_1$ 的接入网链路带宽利用率对比结果如图 6 所示。

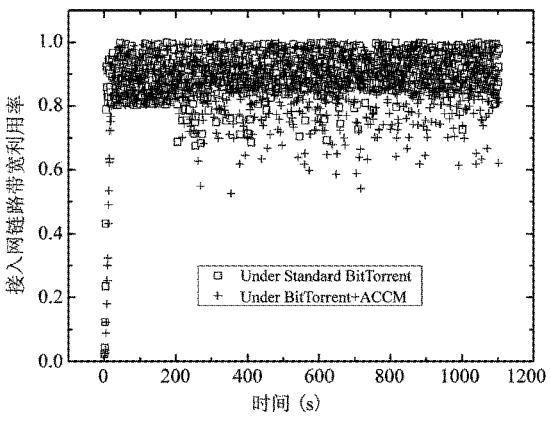

图 6 标准 BT 和 BT + ACCM 的接入网带宽利用率对比

在网络实验开始阶段,我们只启动了 BitTorrent 业务。由图 6 可以看到,在 30s 前,节点 $Peer_1$ 运行标准 BitTorrent 系统和运行 BitTorrent + ACCM 系统时,其接入网链路带宽利用率都在不断增长,但运行标准 BitTorrent 系统时的接入网链路带宽利用率增长更为迅速。这是因为标准 BitTorrent 业务是在启动之后立即增大并发连接窗口,迅速抢占节点的接入网带宽资源;而 ACCM + BitTorrent 系统在启动之后逐步扩大其并发连接窗口,以减轻对互联网络及传统网络流量的影响。在节点 $Peer_1$ 接入网稳定后(图中 30s 之后到 200s 阶段),使用 BitTorrent + ACCM 和标准 BitTorrent 系统时节点 $Peer_1$ 的接入网链路带宽利用率基本一致,这表明引入 ACCM 机制之后,BitTorrent 业务同样能够保证较高的带宽利用率。

在网络实验进行 200s 之后,我们向网络中注入一定数量的 FTP 流量。从图 6 中可以看到,在使用 ACCM + BitTorrent 系统时,节点 $Peer_1$ 的接入网链路带宽利用率变化更为剧烈,这表明一旦节点 $Peer_1$ 的接入网链路发生拥塞,ACCM 拥塞控制机制便开始工作,迫使 BitTorrent 业务出让一部分带宽资源给 FTP 业务。

在本实验的原型系统中,接入链路拥塞推断算法会导致额外的开销,这也是本系统一个较大的缺点。在实验中,探测包数量设置为 10 个/次,探测周期设为 1 次/秒,此时的额外开销为 6.4kbps。通过图 4、图 5 和图 6 的实验结果对比,我们认为接入链路拥塞推断算法所引起的额外开销处于可接受的范围之内,不会降低 BitTorrent 系统的文件传输性能。

ACCM 中 AIAD 策略的控制参数对 P2P 文件共享系统的文件传输速率和稳定性有较大的影响。较小的控制参数 Δ(如 $\Delta = 1$)将导致 P2P 文件共享系统启动时及网络由拥塞转为稳定时,其传输速率增长较为缓慢。但是较大的控制参数 Δ(如 $\Delta = 8$)将导致 P2P 文件共享系统在网络由稳定转为拥塞时,其传输速率将大幅度下降。因此,参考图 6 的实验结果,本文认为在实际的应用中,ACCM 中 AIAD 策略的控制参数 Δ 设为 2 是比较合理的选择。

4　结　论

本文提出并设计了一种汇聚拥塞控制机制——ACCM,它强迫 TCP 友好性基于 P2P 文件共享系统的所有连接,而不是单个连接。ACCM 采用应用层测量技术来检测节点接入网链路拥塞状况,在此基础上,使用 AIAD 策略来控制并发数据传输连接窗口,在最大化节点接入网链路带宽利用率的基础上实现了与其它传统互联网业务的和平共处。网络实验结果表明,在拥塞存在的情况下,ACCM 能够实现一定的公平性及拥塞避免;在不存在拥塞的情况下,ACCM 能使网络有效地利用带宽资源。此外,ACCM 既不需要网络节点的支持,也无需对网络协议栈进行修改,因此,它非常容易被整合到现有 P2P 文件共享系统之中。

参考文献

[1] BitTorrent, Inc. BitTorrent. http://www.bittorrent.com: BitTorrent, Inc, 2007

[2] Brilliant Digital Entertainment, Inc. KaZaA. http://www.kazaa.com: Brilliant Digital Entertainment, Inc, 2007

[3] Weblogic Research Centre. The true pictures of P2P file sharing. http://cachelogic.com/research/slide1.php: Weblogic, Inc, 2007

[4] Balakrishnan H, Rahul H, Seshan S. An integrated congestion management architecture for Internet hosts. In: Proceedings of ACM Special Interest Group on Data Communication, Cambridge, MA, USA, 1999. 175-187

[5] Ott D E, Sparks T, Mayer-Patel K. Aggregate congestion control for distributed multimedia applications. In: Proceedings of IEEE Conference on Computer Communications, Hong Kong, China, 2004. 10-23

[6] Floyd S, Handley M, Padhye J, et al. Equation-based congestion control for unicast applications. In: Proceedings of ACM Special Interest Group on Data Communication, Stockholm, Sweden, 2000. 43-56

[7] Singh M, Pradhan P, Francis P. MPAT: aggregate TCP congestion management as a building block for Internet QoS. In: Proceedings of IEEE International Conference on Network Protocols, Berlin, Germeny, 2004. 129-138

[8] Hacker T J, Noble B D, Athey B D. Improving Throughput and Maintaining Fairness Using Parallel TCP. In: Proceedings of IEEE Conference on Computer Communications, Hong Kong, China, 2004. 2480-2489

[9] Liu S, Vojnovi'c M, Gunawardena D. Competitive and considerate congestion control for Bulk Data Transfers. In: Proceedings of IEEE International Workshop on Quality of Service, Evanston, IL, USA, 2007.1-9

[10] Tsugawa T, Hasegawa G, Murata M. Background TCP data transfer with inline network measurement. In: Proceedings of IEEE Asia-Pacific Conference on Communications, Perth, Western Australia, 2005. 459-463

[11] Rubenstein D, Kurose J, Towsley D. Detecting shared con-

gestion of flows via end-to-end measurement. IEEE/ACM Transactions on Networking, 2002,10 (3): 381-395

[12] Katabi D, Bazzi I, Yang X. A passive approach for detecting shared bottlenecks. In: Proceedings of IEEE Conference on Computer Communications and Networks, Arizona, USA, 2001. 174-181

[13] Cui W, Machiraju S, Katz R, et al. SCONE: A tool to estimate shared congestion among internet paths: [technical report]. Berkeley: University of California, 2004

[14] Kim M S, Kim T, Shin Y J, et al. A wavelet-based approach to detect shared congestion. In: Proceedings of ACM Special Interest Group on Data Communication, Portland, Oregon, USA, 2004. 293-306

[15] Jacobson V, Karels M J. Congestion avoidance and control. In: Proceedings of ACM Special Interest Group on Data Communication, Stanford, CA, USA, 1988. 314-319

[16] Olaf V D S. Extended BitTorrent. http://xbtt.sourceforge.net: SourceForge, Inc, 2007

An aggregate congestion control mechanism for P2P file sharing systems

Li Wei*, Chen Shanzhi* **

(* State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876)

(** China Academy of Telecommunication Technology, Beijing 100083)

Abstract

In consideration of existing P2P file sharing systems' performance deterioration in traditional Internet applications due to the usage of parallel connections to transfer data and the occupation of too much bandwidth resource, the paper proposes an aggregate congestion control mechanism(ACCM) for P2P file sharing systems. By observing the share congestion in access links through the application-level measurement technology, the ACCM dynamically adjusts the window size of parallel transmission control protocol (TCP) connections, and achieves friendliness to tradition Internet applications on the basis of maximizing the utilization of network bandwidth. The experiments demonstrate that the ACCM can decrease the size of the parallel connection window in the presence of congestion to achieve certain fairness with traditional Internet applications, and when the network is idle, the ACCM can increase the size of the parallel connection window to improve the utilization of network bandwidth resources.

Key words: P2P, file-sharing systems, aggregate congestion, congestion control