P2P 流媒体系统中层次化网络拓扑推断技术的研究

雍兴辉,黄永峰
(清华大学电子工程系,北京 100084)

摘 要: 目前基于 P2P 技术的应用已经远远超过了传统互联网应用,成为占用互联网带宽最多的网络应用. 因此,如何减少 P2P 应用、特别是 P2P 流媒体播放系统对骨干网带宽的过度消耗,成为 P2P 技术中一个急需解决的问题. 本文提出利用一种基于网络断层扫描的分层推断方法,以将 P2P 流媒体系统中的流量限制在较小的网络范围内,从而减少跨网流量、同网之间的骨干网带宽消耗,缓解网络拥塞. 在 OPNET 上的仿真表明,该方法能够适应 P2P 流媒体的高动态性,有效降低跨网流量以及对骨干网的带宽消耗速度.

关键词: P2P; 流媒体系统; 网络断层扫描; 拓扑推断
中图分类号: TP393 **文献标识码:** A **文章编号:** 0372-2112 (2010) 01-0001-05

Network Topology Hierarchy Inference on P2P Media Streaming System

YONG Xing-hui, HUANG Yong-feng
(*Electronic Engineering Department*, *Tsinghua University*, *Beijing* 100084, *China*)

Abstract: Currently, P2P-based applications, prevailing over internet, become one of the greatest internet network-bandwidth consumers. So, It's indispensable to reduce the network bandwidth consume of P2P application, especially of P2P media streaming. This paper proposes a fast-inferring method based on network tomography in order to limit the traffic in smaller network, to decrease network-cross traffic and backbone network-bandwidth consume internally, and to suppress the congestions. The simulation shows that such technique can adapt the high dynamics of P2P media streaming system, and effectively slow down the network-cross traffic and backbone network-bandwidth consume.

Key words: P2P; media streaming system; network tomography; topology inference

1 引言

近年来,P2P(Peer-to-Peer,对等计算)技术得到了迅猛发展,各种 P2P 应用如 BT[1],Thunder[2],PPLive[3],PPStream[4]等都拥有极其庞大的用户数目.基于 P2P 系统的应用已经远远超过了其它的传统应用,成为占用互联网带宽最多的网络应用技术[5,6].目前的 P2P 系统大多是建立在底层物理网络之上的覆盖网络,一般其路由策略和物理网络结构几乎无关,在覆盖网上的相邻节点可能在物理网络上距离很远,由此带来的低效的数据转发策略和严重的数据冗余容易导致对网络带宽的过高消耗、使得网络拥塞严重.因此网络服务供应商(ISP)都不同程度的对 P2P 应用加以限制以解决网络拥塞情况.在这种情况下,实时性强、对带宽需求高的 P2P 流媒体系统需要知道更多物理网络结构信息,以实现类似于 IP 层多播的高效分发模式,减少对网络带宽的冲击.同时,相对于文件共享系统,P2P 流媒体系统中的节点变动(加入、退出、网络带宽变化等)更加频繁,那么对物理网络结构的获取就需要更加快捷.然而,目前绝大多数网络拓扑测量工具如 traceroute,大多基于 ICMP 协议,往往需要网络中各个中间路由器或者节点主机都支持 ICMP 才能正常工作,但 ISP 内部网络结构往往被视为机密,加上防火墙等网络安全技术的广泛使用,各路由器或者节点将不对 ICMP 报文做出响应——据参考文献[7]统计,超过 50% 的探测点不对 ICMP 报文做出响应,从而严重影响了这类工具的准确性和实时性.因此,基于网络层析成像(Network Tomography, NT)[8]技术(以下简称 NT)的拓扑测量技术应运而生.该技术是将节点之间的网络看成"黑盒",利用各种数据传输协议进行主机之间端之间的测量,并对测量数据运用统计学理论与推演算法,从而对网络"黑盒"的内部结构进行推断以获得确定或者近似的网络拓扑结构,以便用于网络监控及优化.目前基于 NT 的方法很多,其性能指标如数据获取时间、网络拓扑推断准确度、统计推演算法计算复杂度等

收稿日期:2008-09-28;修回日期:2009-02-23
基金项目:国家自然科学基金(No.60703053);国家 973 重点基础研究发展规划(No.2007CB310806);国家 863 高科技发展计划(No.2007AA010306)

各异,往往需要针对实际的应用背景进行优化.

2 NT 技术的相关研究

NT 将网络内部物理网络视为"黑盒",仅通过对端节点之间的测量数据进行统计分析,得到内部物理网络的确定或者近似网络拓扑结构.

由于网络推断和医学扫描的相似性,文献[9]首先提出 Network Tomography 的概念;文献[10~12]等基于端到端的 Path 测量,解决 Path 中 Link 的拓扑性能等推断问题;文献[9,13,14]等则是根据 Link 的测量来推断 Path 的性能.两个对称问题中前者的解并不唯一,大多是对实际物理网络的近似,即使这样,也能够对网络结构及网络传输优化等起到指导性的作用,这也是本文的关心点.

对于 Link 推断,涉及到单播、多播,度量选择,推断算法等关键问题.

文献[10,15,16]等给出了基于多播的方法;文献[12,17]分别提出了基于单播的、包对的方法;文献[18]提出了基于单播并合理利用 IP 包中的 TTL 域的方法;文献[19,18]给出了被动测量(监视 TCP 会话)的方法.文献[20]给出了多源测量的方法.

在 NT 技术中,为了便于进行统计分析,通常要求反映网络特性的度量具有单调性,其值一般随着共享路径(Path)的增长而增加;可测量性,在可接受的时间内能够收集足够的测量数据.文献[21]等给出了基于丢包率和传输延迟的测量方法,但是丢包在轻负载网络中极少发生,而延迟又需要较好的时钟同步,因此文献[22]提出了称为"Sandwich"基于延迟差的方法,避开了时钟同步问题.

在推断算法方面,文献[22]给出了极大似然估计法,文献[23]文献提出了期望最大值法、文献[24]则给出了伪最大似然估计法.

NT 技术,要么由于测量过程需要的时间较长、要么由于推断算法具有较高的复杂度,现阶段还处于线下处理,不能够做到在线推测,很少看见 NT 技术真正的运用到实际中.本文尝试将 NT 技术运用到 P2P 流媒体传输中,优化 P2P 的节点选择,从而实现较高效率的数据转发.

3 P2P 流媒体播放系统的 NT 方法

目前绝大部分 P2P 覆盖网络系统,都使用的基于延时最小的节点选择机制,不考虑真实的物理网络结构,总是优选选择延时较小的节点,当使用者逐渐增多时,与 IP 组播技术相比,在骨干网上将会出现很多重复的 P2P 流量,加速了骨干网的拥塞.利用 NT 技术得到物理网络的精确或者近似拓扑结构,就能合理安排数据传输和转发策略,建立起类似于 IP 组播的高效模式.另一方面为了保证播放流畅,P2P 流媒体系统中的每个节点需要维持稳定的数据接受率,但是网络中的节点变动(加入、退出、网络带宽变化等)频繁,这些特点决定了在 P2P 流媒体中需要采用一种快速 NT 技术.

考虑到各个网络的负载不同,同时为了绕开时钟同步问题,在 P2P 流媒体中,选择主动单播的"Sandwich"方法获取到的延时差作为度量.

3.1 "Sandwich"探测方法

如图 1 所示,每三个包一组,完成一次测量.包 p_1、p_2 发往节点 3,包 q 发往节点 2,p_1、p_2 的大小比 q 小的多.p_1、p_2 的发包间隔为 d.由于小包 $p2$ 位于大包 q 后,p_1、p_2 的接收间隔将随着共享路径的传输而不断增大,在非共享路径以及标准丢包策略下,由于 p_1、p_2 是连续的两个小包,p_1、p_2被连续转发并且经历的延时大致相等[13,14].考虑到网络环境传输流量的影响,接受时间间隔变化量将是一个随机变量 X_{ij}(图 1 中的 Δd).如果每次测量间隔足够大时,可以认为每次测量是相互独立的.设测量的样本均值 λ_{ij},样本方差为 σ_{ij},X_{ij}可以简单的认为近似服从参数为$(\lambda_{ij},\sigma_{ij})$的正态分布.

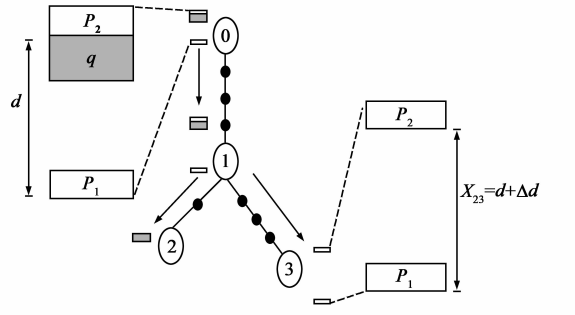

图1 Sandwich测量法

在选择拓扑推测算法时,出于对 P2P 流媒体系统实时性的考虑,我们选择了基于极大似然估计法的合并似然树算法[22].

3.2 极大似然估计法

对于真实网络拓扑的近似 T^*,设其根为探测包发送者 S,叶子为探测包接受者 $p_1,p_2,\cdots,p_n\in \boldsymbol{P}$,拓扑推断的任务就是通过对探测包的统计分析确定 S 到 R_i 之间的树型连接关系.设从 S 到 R_i 之间的所有可能的树 $T\in \boldsymbol{F}$,$x_{ij}\in \boldsymbol{x}$ 和 $\gamma_{ij}\in \boldsymbol{\gamma}$ 分别为测得的和实际的 p_i,p_j 之间的延迟.设 T 中的每一个连接(link)都会引入延时差 U_{ij},S_{ij}表示 p_i,p_j 共享连接的集合.通常 $\gamma_{ij}=\Sigma_{l\in S_{ij}}$,$x_{ij}$也必须潜在的满足单调性的关系.不妨设 G 满足单调性的性能集合.则得到的最大似然概率拓扑树为

$$T' = \arg\max_{T\in F}\sup_{x\in G} p(\boldsymbol{x}\mid\boldsymbol{\gamma}) \quad (1)$$

由于 $\boldsymbol{F}$ 空间极其庞大,当叶子数目为 N 时,其最小的 $\boldsymbol{F}$ 空间大小为 $N!/2$.因此要寻找到满足式(1)的

最优化很困难.为此,文献[22]提出了简化的算法如合并相似树算法、Markov 链 Monte Carlo 迁移的极大似然树算法等,我们采用较为简单快速的合并相似树算法(ALT)求得式(1)的近似解.ALT 算法基于贪婪算法的原理,通过每次合并具有最大相似度的节点逐步合并出一个二叉树拓扑结构.

设探测包接受者集合为 $\boldsymbol{P}$,$\boldsymbol{R}$ 为实数集合,对于 $\boldsymbol{P}$ 每对叶结点的相似度 γ_{ij} 估计如下:

$$\gamma'_{ij} = \arg\max_{\gamma\in R}(h_{ij}(x_{ij},\sigma_{ij}|\gamma) + h_{ji}(x_{ji},\sigma_{ji}|\gamma)) \quad i,j\in \boldsymbol{P};i\neq j \tag{2}$$

其中 $h_{ij}(x_{ij},\sigma_{ij}|\gamma)$ 是 γ 为均值、σ_{ij} 为方差的概率密度函数的适当形式.如果存在 i、j 使得

$$\gamma'_{ij} \geqslant \gamma'_{lm}\ \forall\ l,m = \boldsymbol{P} \tag{3}$$

我们可以认为探测包接收者、j 具有最大的相似度,i、j 具有相同的父节点 k,将 i、j 用其父节点 k 进行替代,就有 $\boldsymbol{P'} = \boldsymbol{P}\cup\{k\}\setminus\{i,j\}$,设 $p(k) = \{i,j\}$,节点 k 和其它节点之间的相似度计算如下:

$$\gamma'_{kl} = \arg\max_{\gamma\in R}(\Sigma_{r\in p(k)}h_{rl}(x_{rl},\sigma_{rl}|\gamma) + h_{lr}(x_{lr},\sigma_{lr}|\gamma)) \tag{4}$$

重复以上过程,直到只剩下一个节点为止.

由于 ALT 算法最终生成的是一个二叉树拓扑结构,其与实际网络结构可能差异较大,为了补偿 ALT 算法带来的误差,并考虑到 P2P 流媒体系统的及时性需求,这里引入了分层 NT 方法.

3.3　P2P 流媒体系统的分层 NT 方法

在 P2P 系统流媒体中,往往存在一个引导服务器向新加入节点给出系统中已经存在的部分活动节点,让新加入节点能够很容易的进入该系统中.

设新加入节点为 k,引导服务器返回的部分活动节点集合为 A,令集合 $B = A\cup\{k\}$,设 B 的大小为 N,$i\in B$,$B' = B\setminus\{j\}$,将 i 看成是数据发送者,B' 看成是数据接收者,进行上述的"Sandwich"探测和 ALT 推断.这是分层 NT 方法的第一层.

当我们遍历所有的 i 后,总共得到了 N 个二叉树拓扑结构,每个二叉树都包含了实际网络拓扑结构的部分信息如图 2(a).分层算法第二层的主要任务就是将这些二叉树拓扑进行合并(BTM),以得到一个更能真实反映物理网络的网状拓扑结构,集合 A 中的节点将他们的二叉树结构传回给 k,BTM 总是在 k 上进行.BTM 大致分为三个步骤.

第一步:二叉树合并,每个二叉树除了叶子节点和根节点(N_k)外,其它的节点均被视为不同的内部节点 M_k,构建一个以 B 中节点为外部节点的多连接网络 T,如图 2(b)所示.

第二步:内部节点合并,对于第一步中的 M_k,连接相同的外部节点可能存在多个内部节点,多个重复的内部节点(如 M_2,M_4)只保留一个,从而得到网络 T',如图 2(c)所示.

第三步:内部优化,如果两个外部节点之间存在多条路径,在不影响其它节点的连接性同时,选取最短的路径(如 N_3 与 M_1 之间).如图 2(d)所示.

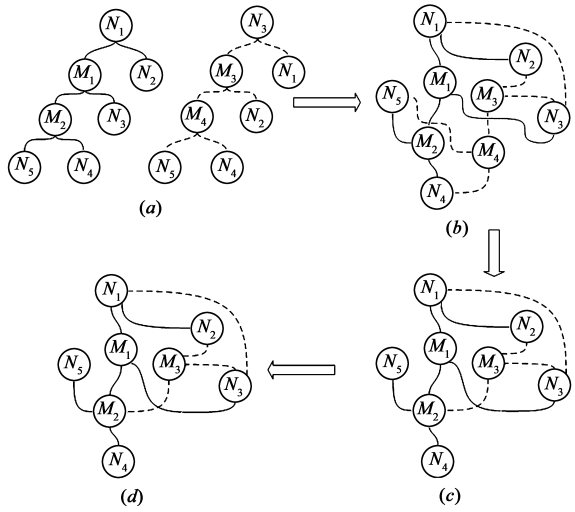

图2　两个ALT树的BTM过程示意图

我们将 BTA 所得的网络拓扑结构称为局部网络拓扑结构.节点 k 就可以优先向距其最短的其它节点索要数据从而完成传输直到符合数据需求.

为了便于引导服务器更好的完成初始节点的选择,引入第三层:节点 k 将第二层产生的网络拓扑结构传给引导服务器或超级节点,引导服务器或超级节点将把这些节点传入的网络拓扑结构做为子图按照上述方法进一步合成全局网络拓扑.当新节点加入 P2P 流媒体系统中,引导服务器将根据新节点的信息,在已有的网络结构中进行搜索,返回离新节点较近的部分引入节点.

除了这三层,还可以有更高的层在第三层的拓扑结构网络基础上(全局网络拓扑就变成了第二级局部网络拓扑)继续完成网络拓扑合并.分层 NT 的结构如

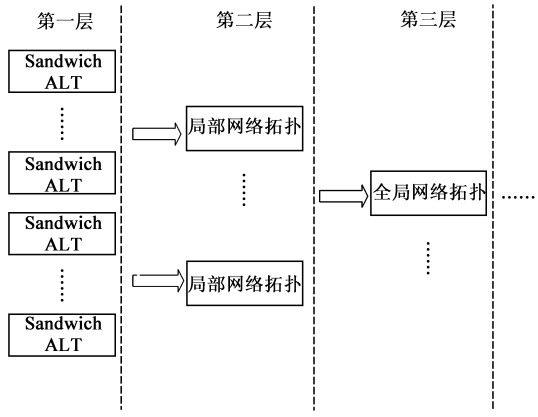

图3　分层NT示意图

图 3.

4 仿真结果与分析

为了说明上述方法的有效性,在 OPNET 之上分别就基于延时和基于网络感知的 P2P 流媒体系统的实际性能进行了对比仿真.

网络环境部署如图 4:一个媒体流服务器、一个引导服务器,8 个子网构成,子网内部是随机生成的树形结构(树的深度不超过 3),如图 2(a)所示.为了在模拟中加快网络资源的消耗,让媒体流的码率为 1Mbps,子网之间(模拟骨干网)的可用带宽为 50Mbps;子网内可用 link 带宽为 5Mbps.每个子网内的节点以指数随机间隔加入,并且间隔均值相同.仿真中假设所有节点进入该网络后一直具有提供服务的能力.

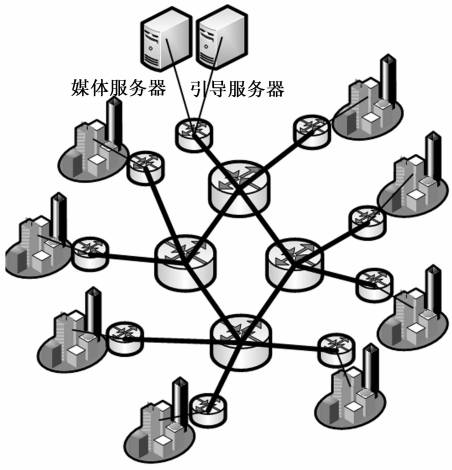

图4 仿真网络示意图

在基于网络感知的测量中,设两次"Sandwich"间隔时间为 10ms,每个"Sandwich"中两个小包之间的时间间隔为 2ms,小包为 54 字节,大包为 1550 字节.对于每个节点对进行 50 次"Sandwich"测量,同一组中的"Sandwich"不能同时进行测量.部分活动节点集合大小不超过六个.

在基于延时的测量中,每次测量之间的间隔为 10ms,同样进行 50 次测量.

我们着重考察这两种传输方案中,随着节点数目增多骨干网、子网内的带宽平均消耗的增长情况.子网内的带宽消耗是指各个子网内所有已使用带宽之和;骨干网的平均带宽消耗是所有非子网内部的 link(粗线所示)已使用带宽总和;同时也考察随着节点数的增多,得到的逻辑网络结构和实际网络的拟和程度和网络的收敛速度(以引导服务器在每次新节点加入后进入稳定状态的时间为准).

通过仿真 7 次,每次仿真 12 个小时,对如上网络的仿真结果如下:

从图 5 可以知道,使用基于延时的方法时,当子网内节点数目不多,带宽消耗较少的时候,大部分流量能够被限制在子网内部,这是由于这种情况下,网络延时主要由传播延迟组成,子网内的节点间的延时往往比子网间的延时小,新加入节点将会优先使用子网内节点.随着节点数目的逐渐增多,网络延时受到已有流量的影响,利用延时的方法,新加入节点将不能够区分网内或者网外节点,因此子网内和子网间的带宽接近与同步增长.而在基于 NT 的方法中,即使节点数目较多,新加入节点也能够区分较好的区分子网内和子网间节点,仍然能够较好的将流量限制在子网内,极大减缓了对骨干网的带宽消耗速.

但是,我们也可以看到基于 NT 相对于 IP 组播来说,其骨干网带宽消耗仍然较大,一方面因为上述 NT 方法得到的网络结构并不能完全反映实际物理网络的结构,另一方面,NT 方法中进行测量需要消耗一定的带宽,当网络变化异常激烈的时候,会带来较大的额外带宽消耗.

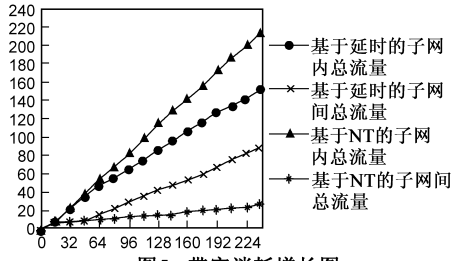

图5 带宽消耗增长图

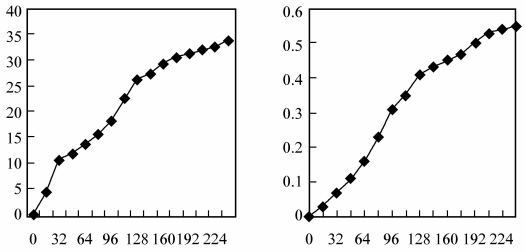

图6 收敛速度左(s)和拟合程度右

从图 6 可以看出,随着节点数的增多,收敛所需要时间越来越长,这是由于一方面随着网络的拥塞,探测包本身在传输中失效,一方面新加入节点需要尝试和更多节点之间的探测,以最终决定选用的节点.而拟和程度将会随着节点数目的增加得到提高,但是其提高的速度越来越慢,这是由于在本仿真中,当节点数目达到一定规模后,新加入的节点不会影响到以前节点的拓扑推断结果.结合图 6 和图 5 可以看出,随着拟合程度的不断提高,虽然该算法实际使用的收敛时间不断增加,但是骨干网的带宽消耗增加的速度减慢.

5 结束语

本文讨论了在 P2P 流媒体系统中如何使用基于端到端的网络拓扑推断技术.提出了一个适合 P2P 流媒

体系统应用的分层 NT 方法,并在 OPNET 上对该方法的应用进行了仿真,仿真结果说明相对于传统的 P2P 基于延时的节点选择方法,该方法能够大大减缓子网之间的网络消耗,使得数据传输效率更高.但是在实际的网络情况下,测量结果会受到网络噪声和背景流量的影响,同时在仿真网络环境里设置的网络拓扑主要以树型为主,该算法对更复杂的网络环境的适应性是下一步研究的内容之一.因此将上述分层 NT 技术融合到笔者所在实验室开发的 P2P 流媒体系统 Cool TV 中在实际网络中对上述进行验证是进一步的工作.

参考文献:

[1] BitTorrent[OL]. http://bitconjurer.org/BitTorrent

[2] Thunder(迅雷)[OL]. http://www.xunlei.com

[3] PPLive[OL]. http://www.pplive.com

[4] PPStream[OL]. www.ppstream.com

[5] Fraleigh C, Moon S, Lyles B, et al. Packet-level traffic measurements from the Sprint IP back-bone[J]. IEEE Network, 2003,17(6):6 – 16.

[6] Chujo T. Modeling internet traffic for network planning and provisioning[A]. In Proceedings of The 7th Asia-Pacific Network Operations and Management Symposium (APNOMS'03)[C]. Kyushu, Japan, 2003.

[7] K G Anagnostakis, M B Greenwald, and R S Ryger. Cing: measuring network-internal delays using only existing infrastructure[A]. Proc. IEEE INFOCOM[C]. SanFrancisco, CA, Apr. 2003.2112 – 2121.

[8] Coates, A Hero III, A O Nowak, R Bin Yu, McGill Univ., Montreal, Que. Internet tomography[J]. IEEE Signal Process. Mag., May 2002.47 – 65.

[9] Y Vardi. Network tomography: estimating source-destination traffic intensities from link data[J]. J. Amer. Stat. Assoc., 1996,91(433):365 – 377.

[10] Multicast based inference of network-internal characteristics (MINC)[OL] http://gaia.cs.umass.edu/minc.

[11] R Caceres, N Duffield, J Horowitz, and D Towsley. Multicast based inference of network-internal loss characteristics[J]. IEEE Trans. Inform. Theory, Nov. 1999,45:2462 – 2480.

[12] M Coates, R Nowak. Network loss inference using unicast end-to-end measurement[A]. ITC Seminar on IP Traffic, Measurement and Modeling[C]. Monterey, CA, Sep. 2000, 28:1 – 9.

[13] J Cao, S VanderWiel, B Yu, and Z Zhu. A scalable method for estimating network traffic matrices from link counts[OL]. http://www. stat. berkeley. edu/binyu/publ-ications. html, 2000.

[14] R J Vanderbei, J Iannone. An EM approach to OD matrix estimation[R]. Princeton University, Princeton, NJ, Tech. Rep. SOR94-04, 1994.

[15] Lopresti F, Duffield N, Horowitz J, Towsley D. Multicast-based inference of network-internal delay distributions[J]. IEEE/ACM Transactions on Networking, 2002, 10(6):761 – 775.

[16] M F Shihand, A O Hero. Unicast inference of network link delay distributions from edge measurements[A]. Proc. IEEE Int. Conf. Acoust., Speech, and Signal Processing[C]. Salt Lake City, UT, May 2001.3421 – 3424.

[17] A Bestavros, K Harfoush, and J Byers. Robust identification of share losses using end-to-end unicast probes[A]. Proc. IEEE Int. Conf. Network Protocols[C]. Osaka, Japan, Nov. 2000.22 – 33.

[18] Y Tsang, M Coates, R Nowak. Passive unicast network tomography based on tcp monitoring[R]. Rice Univ., Houston, TX, Tech. Rep TREE-05, 2000.

[19] Y Tsang, M Coates, R Nowak. Passive network tomography using EM algorithms[A]. Proc. 2001 IEEE Int. Conf. Acoust., Speech, and Signal Processing, vol. 3[C]. May 2001. 1469 – 1472.

[20] Michael Rabbat, Robert Nowak, Mark Coates. Multiple sources, multiple destination network tomography[J]. IEEE Infocom 2004.

[21] S Ratnasamyand S McCanne. Inference of multicast routing trees an bottleneck bandwidths using end-to-end measurements[A]. Proc. IEEE INFOCOM1999[C]. New York, vol. 1, Mar. 1999. 353 – 360.

[22] M Coates, R Casrto, R Nowak, M Gadhiok, R King, Y Tsang. Maximum likelihood network topology identification from edge-based unicast measurements[A]. SIGMETRICS2002[C]. Marina, Del Rey, California, 2002. 11 – 20.

[23] Fu Shih Meng, Hero III A O. Unicast-based inference of network link delay distributions with finite mixture models[J]. IEEE Transactions on Signal Processing, 2003, 51(8):2219 – 2228.

[24] Liang Gang, Yu Bin. Maximum Pseudo Likelihood Estimation in Network Tomography[J]. IEEE Transactions on Signal Processing, 2003, 51(8):2043 – 2053.

作者简介:

雍兴辉　男,1982 年 8 月出生于四川省广元市,研究兴趣是 P2P 流媒体网络传输,P2P 流媒体编码.

E-mail: loneystar19830818@163.com

黄永峰　男,1967 年 12 月出生湖北省赤壁市,研究兴趣为计算机网络,信息安全,多媒体通信. E-mail: yfhuang@tsinghua.edu.cn