IP 网端到端性能测量技术研究的进展

林 宇,程时端,邬海涛,金跃辉,王文东
(北京邮电大学程控交换技术与通信网国家实验室,北京 100876)

摘 要: 本文介绍了 Internet 端到端测量技术的最新进展,包括性能拓扑推测、时延测量、丢包率测量. 最后讨论了未来 Internet 端到端测量的研究趋势和应用前景.

关键词: 网络测量; 时延; 丢包率; 拓扑推测; 基于性能的推测
中图分类号: TP393. 1 **文献标识码**: A **文章编号**: 0372-2112 (2003) 08-1227-07

The Achievement of End-to-End Performance Measurement Technologies in IP Networks

LIN Yu, CHENG Shi-duan, WU Hai-tao, JIN Yue-hui, WANG Wen-dong
(*National Lab of Switching Technology & Telecom Networks, Beijing University of Posts and Telecommunications, Beijing* 100876, *China*)

Abstract: This paper introduces the achievements of Internet end-to-end performance measurement technologies, including metric-induced topology inference, delay inference, loss probability inference and bandwidth measurement. Finally, we point out the application trend in this field and present some topics to be further studied.

Key words: network measurement; delay; loss probability; topology inference; metric-based inference

1 引言

近年来,Internet 测量受到了国际上的普遍关注. 从网络研究看,采集、分析、解释测量数据是网络流量、拓扑、行为建模分析的基础和验证手段,Internet 流量的自相似性[1]、Internet 拓扑的幂率分布[2,3]等重要规律都是通过网络测量发现的;从网络运营、维护和服务看,网络测量的结果是进行宏观网络控制和管理、业务计费的重要依据. 在服务质量研究领域,QoS 控制、管理、计费和测量是相互关联的整体. 许多 QoS 控制机制,如基于测量的连接接纳控制[4,5]、QoS 路由、流量工程[6]、拥塞/瓶颈检测等,需通过测量获取网络性能;QoS 管理需监测网络和用户的服务性能,进行动态资源管理、确认服务等级合约、执行入侵和攻击检测等;QoS 计费需要根据用户实际获得服务质量来收费;它们都需要 QoS 测量的大力支持. 从应用性能优化看,多媒体应用需要通过测量了解当前网络的性能信息来优化编码器/解码器,以获得更好的业务质量. 网络测量对于许多 Internet 应用和协议,特别是涉及大量数据传输和具有时延限制媒体流的应用至关重要. 内容分发网络[7]中的请求路由协议、对等网络[8]、网络缓存的位置选择和维护策略[9]、端系统的组播[10]、内容服务器中的流调度和接纳控制策略[11]、DNS 和 Web 性能检测[12]等都需要网络测量的有力支持.

网络测量总的目标是将 Internet 网络拓扑、带宽、性能等映射成随时间空间变化的函数. 但是,Internet 分布化、不协作(Uncooperative)、异质(Heterogeneous)的特点以及流量特征的复杂性,使得 Internet 测量研究是极具挑战性的工作[13].

根据测量技术获得网络节点支持的多少以及测量点的位置,测量可分为基于路由器的测量(Router-Based Measurement)、端到端测量(End-to-End Measurement)和路由器协作测量(Router-Aided Measurement). 基于路由器的测量主要由路由器中的管理软件来完成测量. ISP(Internet Service Provider)通常采用基于路由器的测量来监测其内部网络的拓扑、流量、时延、丢包率等. 由于各 ISP 之间的非协作性,这些数据对外保密;另外,将大量路由器统计的性能数据传递给中心网管系统,本身就需消耗大量带宽,增加网络负荷[13];这些原因使得在许多场合下不适合采用基于路由器的测量. 端到端测量的目标是在只有边缘主机参与下,无需路由器的配合,获取网络性能统计,并且尽可能减小对网络造成的负荷. 路由器协作测量在边缘主机上执行测量,但需要路由器的配合,这方面近来提出了一些新协议[14~16]. 由于 Internet 不同 ISP 之间不协作的特点,依赖于路由器配合的测量方法的能力将受限,因此,路由器协作的测量需要获得标准化组织和工业界的支持.

另外,根据是否发送主动探针(Active Probe),测量技术可分为主动测量和被动测量技术. 主动测量通过向网络发送探

收稿日期:2002-10-14;修回日期:2003-05-14
基金项目:自然基金项目(No. 90204003);863 项目(No. 2002AA103063,2001AA121052,2001AA112071);博士点基金项目(No. 20010013003)

针,并根据探针所携带的信息来推测网络的情况,主动测量将影响网络的负荷,产生大量探针的主动测量方法无法实用.探针是由同一源发送的数据包序列,根据发送探针数据包的类型,还可划分为组播探针、单播探针.根据探针结构的差异,常用的探针可分成:单数据包探针、等长和非等长数据包对探针、数据包串探针等[61].被动测量不向网络发送探针,而是监听网络中的分组流来推测网络的情况,被动测量不会对网络的负荷造成影响.

目前,Internet 测量研究工作大致可划分为三部分:端到端性能测量(主要包括性能拓扑推测、时延、丢包率测量、带宽测量等)、路由/路由器相关测量(包括流量抽样技术、根据路由器端口流量推测端到端流量特征、路由器参数推测、路由器协作测量协议、路由测量、网络距离推测等)、应用层测量(Web 测量、DNS 系统性能测量等).限于篇幅,本文主要介绍端到端性能测量,路由/路由器相关测量另文讨论.端到端性能测量的研究最初是由 MINC[17]项目中组播树丢包相关的研究引发,而后推广到单播网络以及其他性能的研究中.网络测量有关的工具和基础资料可参见 CAIDA[18]项目.

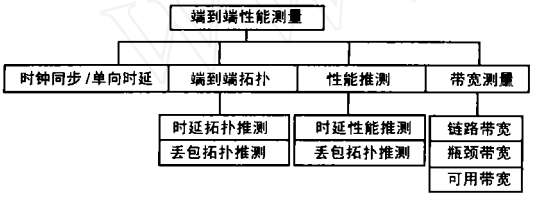

图 1 端到端性能测量的研究工作及其联系

在端到端测量技术中,时间测量是许多测量工作的必需因素,这使得测量主机间时钟同步是许多端到端测量技术的重要基础;在网络拓扑是不可知的情况下,端到端拓扑推测是端到端性能推测的基础,此外,它还是基于性能角度对复杂物理拓扑进行抽象简化的有效手段;性能测量(时延、丢包)的研究主要集中于树型拓扑的性能推测,对于一般网络,可将其划分成多个树型拓扑分别进行测量,而后将这些树型拓扑上获得的性能参数恢复成整个网络的性能;网络带宽测量的研究包括链路带宽、瓶颈带宽和可用带宽三部分,带宽测量中引入的一些探针技术,如 Packet Pair[55~60],也被运用在时延丢包测量中.

2 端到端单向时延测量和主机时钟同步

测量主机间时钟同步是端到端测量重要的技术基础.以往的时钟同步技术,大都利用 GPS 或 PSTN、CDMA 等网络借助外部时钟源来实现测量主机间同步.这种技术精度高,但费用昂贵且在测量主机数量大时难于实现,无法满足大规模端到端测量的需要.在广域网上,NTP[19]协议只能到达几十毫秒量级的精度,这个误差对于时延测量而言不可接受.

主机时钟同步与单向时延测量研究密切相关.端到端双向时延测量可通过环回时延 RTT(Round Trip Time)获得,但端到端单向时延测量却需要两端主机间时钟同步.单向时延测量有其特殊的意义:Internet 路径通常不对称,往返路径可能穿过不同的 ISP 甚至不同网络结构;双向链路性质不对称(如 ADSL,卫星链路),双向拥塞排队不同;某些应用如 FTP 的性能更依赖于单方向性能.

称测量获得的时延为测得时延(Measured Delay),称实际的时延为真实时延(True Delay).定义测量样点集为 $\Omega=\{v_i=(t_i,d_i),i=1,\cdots,N\}$,其中,$t_i$ 是发送主机在发送时刻为该数据包打上的时间戳,d_i 为测得时延,即接受者接受时刻减去发送时间戳,它不是真实时延.通常,两终端主机在测量初始时刻时间不同,且按照不同的频率运行,称频率之差为时钟频率偏差(Clock Shew).如果两主机的时钟完全同步(初始时刻和频率都相同),则测得时延就等于真实时延;如果时钟频率偏差为常数,则测量时延曲线类似于图 2(a),测得时延样点的下延近似于一条直线 $d(t_i)=\alpha t_i+\beta$,其中,β的含义是两时钟在测量初始时刻的时间偏差加上固定的传播和发送时延,α的含义是两时钟频率偏差导致测量时延(下延)随着测量时间的推移而线性的增加(或减小),测得时延的抖动是拥塞变化导致排队时延的不同.

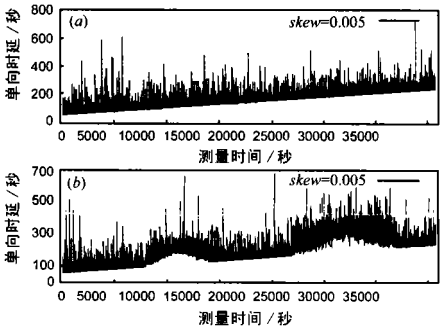

图 2 文[24]仿真获得的时延

实际上,时钟的行为非常复杂,时钟频率将随着时间、温度等发生漂移,这使得测得时延下延不再是线性变化,比如图 2(b)中下延变成曲线.另外,主机时钟可能被 cron 进程重新调整校准(一天可能发生几次),或者由于运行了 NTP 协议,定期与外部时钟进行校准(调整的时间粒度较小).由于这些原因,使得实际测量时延曲线类似于图 3(a)和 3(b)(出现突变)[24].单向时延测量的关键任务就是要消除时钟频率偏差以及外部或本地时钟校准带来的误差,且计算复杂度应足够小(通常,时钟频率漂移在较短的时间区段内给测量时延带来的误差不明显,因此可将测量数据分区段处理,以消除时钟频率漂移的影响.).

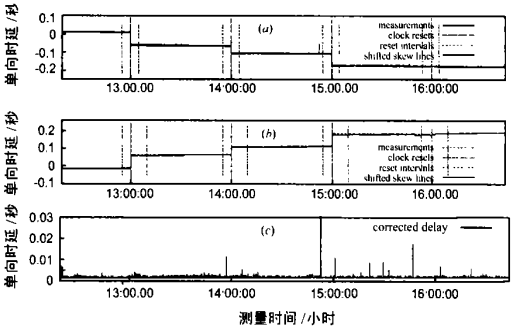

图 3 测量获得的时延及修正后的结果

要消除时钟行为对测量的影响,关键在于基于某种最优化目标来确定测量样点下延直线(即 α 和 β). 文[20]采用 Median Line Fitting 技术来校正时钟频率偏移,但该算法在数据抖动大时效果很差(单向时延受网络拥塞程度的影响常常变化很大). 文[20,21]采用在测量时延下延样点集上执行线性回归来确定 α 和 β,线性回归在数据分布服从正态分布时性能很好,但测量时延数据并不满足正态分布. 文[21]采用 Piecewise Minimum 算法得到一系列分段条件下的时钟频率偏差,但是这些偏差值并不一致,因此也无法正确估计频率偏差. 文[21]还采用"最小化所有样点到直线距离的和"为最优化目标,形式如下:

$$f_{obj1}(\alpha,\beta)=\sum_{i=1}^{N}(d_i-\alpha t_i-\beta)=\sum_{i=1}^{N}d_i-\sum_{i=1}^{N}t_i\alpha-N\beta \quad (1)$$

并运用 Linear Programming 算法[22,23]实现 $O(N)$ 计算复杂度.

以上算法仅针对频率偏差提出解决方法,但都无法解决时钟重置/调整问题. 文[24]定义了三种最优化目标:最小化曲线和直线构成区域的面积,形如

$$f_{obj2}(\alpha,\beta)=\sum_{i=1}^{N}(d_i-\alpha t_i-\beta+d_{i+1}-\alpha t_{i+1}-\beta)\frac{t_{i+1}-t_i}{2} \quad (2)$$

最大化恰好落在直线上的测量时延下延样点数目,形如

$$f_{obj3}(\alpha,\beta)=\sum_{i=1}^{N}1\{d_i=\alpha t_i+\beta\} \quad (3)$$

以及式(1),并采用凸分析理论寻找与测量时延下延重合程度最大的一系列直线段,解决了时钟重置/调整问题. 在没有时钟重置/调整问题时,其算法本质和精度与 Linear Programming Algorithm 相同. 图 3(a)和 3(b)中给出了采用文[24]算法确定的一系列时钟频率偏差直线以及时钟调整/重置的位置,利用这些信息就可以校正测得时延,获得类似图 3(c)的校正结果. 文[24]的算法复杂度为 $O(N)$,可实现在线估测.

3　性能相关的拓扑推测

为了推测和标记网络的性能,需了解网络的拓扑. 常用的网络拓扑测量方法是分析来自网络内部资源的数据(比如 BGP 路由表, ICMP Replies),生成 Internet 拓扑或性能的报告[25~27],这种方法是基于路由器或路由器协作的,它适合于在大时间尺度上进行宏观分析,但不适合小时间粒度的场合.

目前端到端拓扑推测的研究[28~33]主要针对连接一个发送者与多个接受者之间的树型拓扑,比如图 4(a)发送者 0 和叶节点之间的树,且假定发送者与接受者之间的路由固定. 端到端拓扑推测不同于物理拓扑测量问题,它利用端节点性能之间的相关性(时延/丢包相关性),来推测与某种性能(比如时延、丢包率)相关的逻辑拓扑[28~33]. 端到端拓扑推测最初在组播树条件下分析丢包率相关性引入[34],后又推广到单播情况以及其他性能(在第 4 节讨论),其基本思想是:将接受者任意两两分组,收集每一对接受者某种沿路径单调递增的性能数据,然后通过性能的相关性来分析拓扑. 比如,端到端时延等于路径上各链路时延之和,是随路径单调递增(至少不减)的函数,如果两个接受者共享的路径越长,则它们的时延相关性将越大(极端情况下,两个接受者路径相同,则时延特性应当一致). 由于端到端拓扑推测利用的是性能相关性,它不可能发现没有分叉物理路径上的中间节点,比如图 4(a)物理路径"0→1→2→3"上的节点 1 和 2,对于 4(a)这样的物理拓扑,拓扑推测的结果只可能类似于图 4(b).

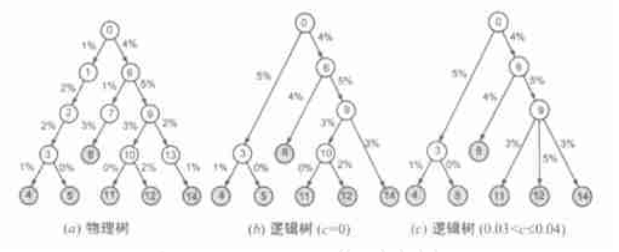

图 4　MINT:物理拓扑和逻辑丢包拓扑(不同敏感性参数 c)

以丢包相关性为例说明推测过程. 设源节点到任意一主机间的路径为 $path_i$,如果构成路径 $path_i$ 的边序列是路径 $path_j$ 的子序列,则称 $path_i$ 为 $path_j$ 的子路径(Subpath). 设 $path_{i,j}$ 表示路径 $path_i$ 和 $path_j$ 的最大共享子路径. 令 $f(\cdot)$ 为沿路径单增性能函数,比如时延,丢包率. 考虑一个叶节点集合 $Node=\{s_0,s_1,s_2,\cdots,s_n\}$,$s_0$ 为发送源,所有叶节点两两组合的集合为 $\Omega'=\{(s_i,s_j),1\le i<j<n\}$. 拓扑推测过程如下:(1)选择 Ω' 中最大共享子路径上丢包率最大的节点对,即 $(s_i,s_j)=\{\max_{(s_u,s_v)\in\Omega'}f(path_{s_u,s_v})\}$. 由于函数随路径单增,因此 s_0 不可能与其他节点也共享最大子路径 $path_{s_i,s_j}$,s_0 与节点 s_i,s_j 之间必有一内部节点 r. (2)生成一个新的节点 r,将 s_i,s_j 从节点集合 $Node$ 中去掉,将新节点 r 加入节点集合 $Node$,再重复(1),直到节点集合 $Node$ 中只剩下三个节点. 以图 4(a)为例(链路上标记的是各链路及节点的丢包率,由于节点 11 和 12 之间的最大子路径"0→6→9→10"上的丢包率最大,所以为节点 11/12 插入一个父节点 10,整个推测过程为:{0,4,5,8,11,12,14}→{0,4,5,8,10,14}("□"表示新加入节点,"."表示删除的节点)→{0,4,5,8,9}→{0,3,8,9}→{0,3,6},即推得图 4(b)的逻辑拓扑. 需要说明如何获得最大子路径的丢包率,比如 $f(path_{s_{11},s_{12}})$,测量可以获得端到端丢包率 $f(path_{s_{11}})$ 和 $f(path_{s_{12}})$,而链路"$s_{10}\to s_{11}$"的发包成功率等于在数据包到达节点 s_{12} 条件下该数据包也成功到达节点 s_{11} 的条件概率,链路"$s_{10}\to s_{12}$"的发包成功率也可类似获得,则 $f(path_{s_{11},s_{12}})$ 可知.

除了丢包相关性外,还可利用其他随路径单增函数,比如最小时延(即传播时延加发送时延,不计排队时延)事件次数、时延相关性、时延偏差[28~33]. 在组播条件下,去往两接受者的数据包在共享路径上时延特性完全相同,如果共享路径越长,则其时延相关系数越逼近于 1. 在单播条件下,去往两接受者的数据包在共享路径上时延存在差异,此时可利用 Packet Pair(即源端连续发送的两个长度相同背靠背的数据包)来近似,由于两个数据包靠得很近,它们在共享链路上的特性非常接近. 文[34,35]在 Internet 中的测量结果证实了这一点.

实际上,上述推测示例是在理想情况下进行的,由于测量获得的性能总是与真实性能存在差异,因此拓扑推测可能发

生误差. 拓扑推测问题可一般地描述为:设 $X=\{x_{ij}\}$ 为测量获得的接受者 i,j 之间某种性能,$\gamma=\{\gamma_{ij}\}$ 为接受者 i,j 之间真实的性能值,T 为节点间某种树型拓扑,则推测获得的拓扑是具有最大似然概率的拓扑树[13]:

$$T^{*}=\arg\max_{T\in F}\max_{\gamma\in G}p(X|\ \gamma,T) \tag{4}$$

其中,F 表示在节点间可能构成的所有树型拓扑的集合,G 为满足随路径单增性质的性能集合. 注意到当网络节点数目很大时,要寻找式(4)确定的全局最优树很困难,但某些方法获得的次优解也能到达较好效果. 文[29]基于决策二叉树,至底向上通过回归式地选择/聚合生成二叉树来确定次优拓扑,但决策二叉树算法的局部贪婪原则可能导致结果不够理想. 文[33]利用蒙特卡罗过程,从全局而非局部的角度考察搜索方向,将搜索集中在具有最高似然概率的区域,到达了较好效果.

文[31]进一步提出了一种框架性的基于性能的网络拓扑 MINT(Metric-Induced Network Topology),其思想是针对某种性能及其敏感性参数 c(Sensitivity Parameter),将物理拓扑转化为一系列不同敏感参数条件下的逻辑拓扑. 比如要针对丢包率进行拓扑推测,图 4(a)是物理拓扑,其逻辑拓扑为 4(b),设定敏感性参数 $c>0.03$,则只有丢包率高于 0.03 的中间链路(即连接非叶节点的链路)才在逻辑拓扑中保留,丢包率小(等)于 0.03 的链路将按一定规则同其父节点合并,即可得图 4(c). MINT 的意义在于将复杂的物理拓扑转化为相对简单的逻辑拓扑,它能很好地标识出那些性能恶化的逻辑链路或区域(如高丢包率、高时延),便于网络故障定位.

4 丢包率测量

本节讨论基于树型网络的丢包率测量[34,37,38,40,41,43~45]. 对于一树型网络,节点为 $j=0,\cdots,m$,设测量 n 条不同的路径 $i=0,\cdots,n$. 定义 a_{ij} 为第 i 次测量路径包含链路 j 的概率(定义为概率的原因是考虑可能出现随机路由情况,如多路径负荷分担),以 a_{ij} 为元素可构成路由矩阵 A. 考虑图 5 对应的树,如采用‘0→1→2’和‘0→1→3’两条测量路径,则矩阵 A 为

$$A=\begin{pmatrix}1&1&0\\1&0&1\end{pmatrix} \tag{5}$$

设 $\theta_1,\theta_2,\theta_3$ 为图 5 中各链路 1,2,3 数据发送成功率的对数,令 $\overline{p_i}$ 为由源(节点 0)到节点 i 的端到端丢包率,则有

$$\begin{pmatrix}\log\overline{p_2}\\ \log\overline{p_3}\end{pmatrix}\approx\begin{pmatrix}1&1&0\\1&0&1\end{pmatrix}\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\end{pmatrix} \tag{6}$$

由于 A 不是满秩的,所以无法解出 $\theta_1,\theta_2,\theta_3$. 可采用其他方式来补充 A. 令 $\overline{p_{2|3}}$ 为在节点 3 收到数据包的条件下,节点 2 收到数据包的条件概率,则 $\overline{p_{2|3}}\approx\theta_2$,类似定义 $\overline{p_{3|2}}$,则补充后的方程为:

$$\begin{pmatrix}\log\overline{p_2}\\ \log\overline{p_3}\\ \log\overline{p_{2|3}}\\ \log\overline{p_{3|2}}\end{pmatrix}\approx\begin{pmatrix}1&1&0\\1&0&1\\0&1&0\\0&0&1\end{pmatrix}\begin{pmatrix}\theta_1\\ \theta_2\\ \theta_3\end{pmatrix} \tag{7}$$

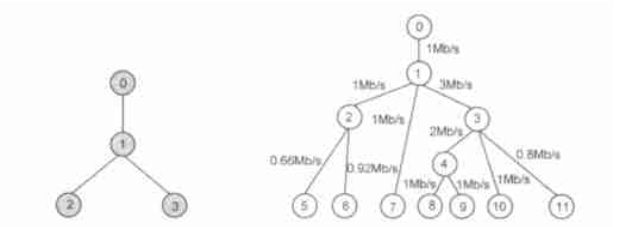

图 5 简单的树型拓扑　　图 6 仿真拓扑

这样可求解出各链路的丢包率. 由于方程是超定的,可采用最小平方误差来估计 $\{\theta_i\}$,文[34,37,38]讨论了更为复杂高效的算法以适应大规模网络推测的需要.

以上补充 A 的方法依赖于组播方式下链路间的丢包相关性. 由于并非所有网都支持组播,且单播业务与组播业务特征不同,因此有必要将这种方法推广到单播网络中. 对于单播,可采用 Packet-Pair 技术来近似. 对于 Internet 中广泛存在的 Drop Tail 队列管理方式,如果 Packet-Pair($packet_1$,$packet_2$)中后一个数据包 $packet_2$ 没有被丢弃(说明此时缓冲区还未满),则 Packet-Pair 前一数据包 $packet_1$ 未被丢弃的概率极大(接近 1),文[36]的工作证实了这一点. 如果 $packet_2$ 的目的地址为节点 3,而 $packet_1$ 的目的地址为节点 2,则在 $packet_2$ 正确到达节点 3 条件下 $packet_1$ 到达节点 2 的概率就近似等于 $\overline{p_{2|3}}$.

如果队列管理机制不是 Drop Tail,比如 RED[39](根据当前等价队长以一定概率丢包),则以上假定不一定成立. 设从源发送给节点 k 的总数据包数为 n_k,节点 k 成功接受到的数据包数为 m_k. 假定每一个数据包丢失为贝努利分布,则在已知 n_k 和 p_k 的前提下,m_k 的概率分布为

$$l(m_k|\ n_k,p_k)=\binom{m_k}{n_k}p_k^{m_k}(1-p_k)^{n_k-m_k} \tag{8}$$

其中 $p_k=\prod_{j\in Path(0,k)}\alpha_j$ 为路径 $Path(0,k)$ 发送成功的概率,α_j 为某链路发送成功的概率. 定义

$$\beta_j=Pr\{\text{第一数据包 }\rho(j)->j|\text{ 第二数据包 }\rho(j)->j\} \tag{9}$$

其中,$\rho(j)$ 表示节点 j 的父节点,‘$\rho(j)->j$’表示从 $\rho(j)$ 成功到达 j. 若队列管理为 Drop Tail,则 $\beta_j>1$.

设两个数据包分别去往 k,l 节点,设 $n_{k,l}$ 为 Packet-Pair 中第二数据包到达 k 节点的次数,$m_{k,l}$ 为两个数据包都成功到达目的地的次数,则在已知 $n_{k,l}$ 和 $p_{k,l}$ 时 $m_{k,l}$ 的概率分布为

$$l(m_{k,l}|\ n_{k,l},p_{k,l})=\binom{m_{k,l}}{n_{k,l}}p_{k,l}^{m_{k,l}}(1-p_{k,l}^{m_{k,l}})^{n_{k,l}-m_{k,l}} \tag{10}$$

其中,$p_{k,l}=\prod_{j\in Path(0,\langle k,l\rangle)}\beta_j\prod_{i\in Path(\langle k,l\rangle,k)}\alpha_i$,即共享路径上 β 的乘积与非共享路径上的 α 乘积. 则总的概率分布为

$$l(m|\ n,p)=\prod_k l(m_k|\ n_k,p_k)\times\prod_{k,l}l(m_{k,l}|\ n_{k,l},p_{k,l}) \tag{11}$$

文[37]采用 EM[54](Expectation Maximum)算法,利用最大似然估计来估测 $\{\alpha_i\}$ 和 $\{\beta_i\}$,该算法的复杂度随着网络规模线性增加. 图 6 和图 7 是文[40]在 ns[42] 上仿真的拓扑以及仿真的部分结果. 其中链路 2 和 5(即连接节点与其父节点对应的链路,比如连接节点 5 和它的父节点 2 的链路称为链路 5)具有较高的丢包率,图 7 上侧两子图分别对应于 Drop Tail 和

RED 缓存管理情况下各链路实际丢包率和估测丢包率,下侧子图为 Drop Tail 和 RED 下丢包率估测误差. 此外,文[38,43]给出了采用被动方式(比如利用 TCP 流)进行丢包性能推测的方法,文[41]还给出了基于丢包率估计的瓶颈链路检测方法.

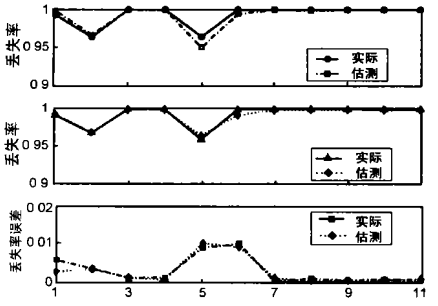

图 7　在 Link2 和 5 上高丢失率

5　时延性能推测

与丢包率测量类似,时延推测也是利用端到端时延与链路时延的关系. 时延推测的工作可分为 4 类:(1)文[43,44]将连续时延分布进行离散化(或量化),使得时延推测转化为便于计算的矩阵操作,这种方法简单,但由于对时延分布进行了量化操作,使得时延推测的精度有所损失;(2)文[48]提出非参数化的连续时延分布(Nonparametric Density Estimation)推测方法;(3)文[49]提出利用累计生成函数(Cumulate Generating Function)推测时延分布的方法,文[50]进一步研究了在时延非平稳情况下的时变时延分布推测方法;(4)文[51]通过端到端时延抖动来推测链路的时延抖动. 这里主要介绍文[49,50]的工作.

一条路径的端到端时延等于各链路时延之和.

$$Y_i = a_{i1}X_{i1} + \cdots + a_{im}X_{im},\ i = 1, \cdots, n \qquad (12)$$

其中:X_{i1}为第 i 次 Probe 通过链路 j 时的时延;$a_{ij} \in \{0,1\}$表示路径和链路间的拓扑关系. 假定$\{X_{i1}\}_{i=1}^{n}$为随机变量 X_j(链路时延)的 n 个独立同分布的样本. 为了整理成归整的线性模型形式,文[49]采用随机变量的累计生成函数,形式如下:

$$K_{Y_i}(t) = \log E[e^{tY_i}] = \log E[e^{t(a_{i1}X_{i1} + \cdots + a_{im}X_{im})}] = \sum_{j=1}^{m} a_{ij}K_{X_j}(t) \qquad (13)$$

其中,$K_{X_j}(t)$为向量,则有

$$K_Y(t) = AK_X(t) \qquad (14)$$

其中,$K_Y(t) = [K_{Y_1}(t), \cdots, K_{Y_n}(t)]^T$,$K_X(t) = [K_{X_i}(t), \cdots, K_{X_n}(t)]^T$. $K_{X_i}(t) = \log \int_{-\infty}^{\infty} e^{tx} p_{X_i}(x)\,dx$ 与随机变量 X_j 的概率密度分布 p_{X_i}一一对应,因此在 A 为满秩时,可通过 $K_Y(t)$来推测 $K_X(t)$,就可获得链路时延分布特征. 由于网络拥塞常常意味着在瓶颈链路(或区域)出现大时延、高丢包率,因此时延推测的重要应用是检测瓶颈. 如果定义‘瓶颈’为这样的事件——链路时延超过门限的概率大于某指定值,则 Chenoff Bound 公式给出了‘瓶颈’事件的概率上界:

$$P(X_j \geq \delta) \leq \min_{t>0}(e^{-t\delta} e^{tK_{X_j}(t)}) \qquad (15)$$

图 8[49]给出了一个例子,Probe1,…,5 用于构造一个满秩矩阵 A. 图 8 中各链路带宽为 1Mbps,传播时延 50ms,链路 3 的流量负荷被设置为高于其他链路. 表 1 给出了 Chenoff Bound 公式计算结果,如果定义‘链路时延超过 0.005s 的概率最少为 0.5’为瓶颈链路的判别标准,即 $P(X_j \geq 0.005s) \geq 0.5$,则瓶颈链路 3 被正确地检测.

表 1　利用时延性能进行瓶颈链路检测

Link	1	2	3	4
$P(X_i \geq 0.005s)$	0.439	0.415	0.964	0.392

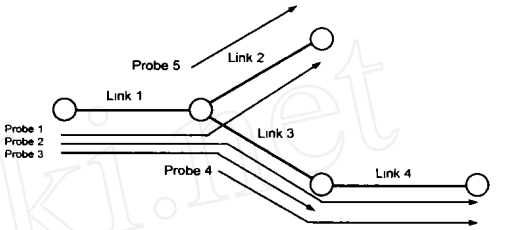

图 8　时延推测仿真拓扑

对于大规模网络测量来说,A 常常是不满秩的(需要很多测量路径和探针),此时只能获得一些链路时延性能的线性组合 $\sum_j \omega_j K_{X_j}(t)$,注意到每一个线性组合项对应着网络的某一个区域(或链路的组合),可利用上述瓶颈检测方法确定哪些区域是拥塞(高时延)常常发生的位置,然后再在该区域内补充发送一些测量探针,构造一个小的满秩矩阵,获取该区域内各个链路的时延特征. 这种逐步缩小搜索区域思想非常适合于大规模网络拥塞(瓶颈)定位.

上述时延推测技术假定时延是平稳分布的(Stationary),此时获得时延分布密度就描述了时延特性. 但实际的网络在不同时期其时延分布特性可能差异很大(比如工作时间和深夜),因此时延是非平稳的(Non-stationary). 文[53]研究了非平稳时延特性,它将时间划分成一系列时间窗口,它假定时延特征是时变的,但在一时间窗口内时延满足平稳分布,然后采用 Sequential Monte Carlo 过程获取各个时间窗口内的时延分布密度,这样就描述了非平稳情况下的时延特征.

6　一般网络拓扑下的性能推测

第 4 节和第 5 节中的性能推测研究都是基于树型结构,利用树型网络内在的性能相关性来实现性能测量,对于一般网络拓扑,需将其划分成若干个树. 一般网络推测需解决两个主要问题:(1)如何高效地将一个一般网络划分成一个树的集合,并且满足某些最优目标,比如测量点最少[52];(2)如何利用多个树测量获得性能信息来构造出完整网络性能信息.

文[53]引入了两种方法来构造全网的性能信息:基于 MVWA(Minimum Variance Weighted Algorithm)和 EM 的方法. MVWA 针对各个树单独进行推测,对于多棵树重叠的链路,从不同树上返回一个平均权重和一个推测值,将加权平均值作为推测结果. 基于 EM 的算法运用 EM 技术[54],综合利用多棵树获得的信息来推测链路的性能. 文[53]指出在测量次数少或多棵树测量结果偏差较大时,EM 算法的性能优于 MVWA. 文[53]的工作主要针对丢包测量,但它还将 EM 算法扩展运用到时延测量中.

7 端到端性能测量的研究趋势

端到端性能测量工作具有非常广泛的应用前景,表 2 中做了一小结. 在端到端性能测量中,许多问题都可以归结为形如

$$y_t = A_t\theta_t + \varepsilon_t \qquad (16)$$

的线性模型,其中,A_t 为与拓扑相关的矩阵,θ_t 为性能向量,ε_t 为端到端的性能指标,ε_t 为系统误差. 这个模型在信号处理、统计分析[62]等领域已经进行了大量研究,有许多类似方法和模型可以在端到端性能推测中借鉴. 从这个线性模型看,y_t, A_t, θ_t, ε_t,都是时变的随机向量和矩阵. 网络的拓扑、性能都是时间和空间的函数,现有的研究工作大都假定性能时空独立来简化问题,也使其无法完全解决网络中实际复杂问题. 在这方面有必要进一步深入地研究.

表 2 端到端网络测量工作的应用前景

端到端性能测量	端到端时钟同步	在不需要 GPS、PSTN、CDMA 等外部时钟同步机制的条件下,达到消除端到端主机频率偏差的目的
	端到端拓扑	发现与某种性能(时延丢包)及其敏感参数相关的逻辑拓扑,便于网络运维和扩容规划
	性能推测	发现网络中大时延/高丢失率的区域或链路;接纳控制;流量工程
	带宽测量	内容分发网络;应用层组播;对等网络;弹性重叠网;传输协议中速率控制;网络维护管理;网络扩容

参考文献:

[1] W Leland, M Taqqu, On the self-similar nature of ethernet traffic[J]. IEEE/ACM Trans on Networking, 1994, 2: 1 - 15.

[2] C Faloutsos, M Faloutsos. On Power-Law Relationships of the Internet Topology [A]. ACM SIGCOMM'99 [C]. Cambridge, USA, 1999. 251 - 260.

[3] Medina, I Matta, J Byers. On the origin of power laws in Internet topologies [J]. Computer Communication Review, 2000, 30(2): 18 - 28.

[4] R J Gibbens, F P Kelly. Measurement-based connection admission control [A]. ITC'97 [C]. Washington, USA, 1997. 879 - 888.

[5] G Bianchi, A Capone. Throughput analysis of end-to-end measurement-based admission control in IP [A]. IEEE INFOCOM'00 [C]. Tel Aviv, Israel, 2000, 3. 1461 - 1470.

[6] Feldmann. A, Greenberg C. NetScope: Traffic Engineering for IP networks [J]. IEEE Network Special Issue on Internet Traffic Engineering, 2000. 11 - 19.

[7] Barbir et al. Known CDN Request-Routing Mechanisms [OL]. http://www.globecom.net/ietf/draft/.

[8] STOICA I, MORRIS, R Chord. A scalable peer-to-peer lookup service for internet applications [A]. ACM SIGCOMM'01 [C]. San Diego, USA, 2001. 149 - 160.

[9] J Kangasharju, J Roberts. Object replication strategies in content distribution networks [J]. Computer Communications, March 2002, 25(4): 367 - 383.

[10] Chu Y, Rao S G, et al. A case for end system multicast [A]. ACM SIGMETRICS'00 [C]. Santa Clara, USA, 2000. 1 - 12.

[11] M E Crovella, R Frangioso. Connection Scheduling in Web Servers [OL]. http://ns.chejue.ac.kr/~jkim/web_os/.

[12] M Grossglauser, B Krishnamurthy, Looking for science in the art of network measurement [OL]. http://www.research.att.com/~mgross/Papers/iwdc2001.ps.

[13] Coates A O, Hero III, R Nowak. Internet tomography [J]. IEEE Signal Processing Magazine, 2002, 19(3): 47 - 65.

[14] S Shalunov, et al. A One-way Delay Measurement Protocol [OL]. http://moat.nlanr.net/AMP/AMP/IPMP/.

[15] A J McGregor. The IP measurement protocol [OL]. http://moat.nlanr.net/AMP/AMP/IPMP/.

[16] M J Luckie, A J McGregor. Towards Improving Packet Probing Techniques [OL]. http://www.acm.org/sigcomm/.

[17] R Caceres, N G Duffield. Multicast-based inference of network-internal characteristics [A]. IEEE Infocom'99 [C]. New York, USA, 1999, 1. 21 - 25.

[18] NetGeo-The Internet Geographic Database [OL]. http://www.caida.org/tools/utilities/netgeo/.

[19] RFC 1305, Network Time Protocol-Specification, Implementation and Analysis [S]. March, 1992.

[20] V Paxson. On calibrating measurements of packet transit times [A]. ACM Sigmetrics, Madison [C]. Wisconsin, USA, 1998. 11 - 21.

[21] Moon S B, Skelly P. Estimation and removal of clock skew from network delay measurements [A]. IEEE INFOCOM 1999 [C]. New York, USA, 1999. 227 - 234.

[22] Dyer M E. Linear time algorithms for two-and three-variable linear programs [J]. SIAM Journal on Computing, 1983, 13. 31 - 45.

[23] Megiddo N. Linear-time algorithms for linear programs in R^3 and related problems [J]. SIAM Journal on Computing, 1983, 2(4): 759 - 776.

[24] Li Zhang, Zhen Liu and Cathy Hong, hui Xia. Clock synchronization algorithms for network measurements [A]. IEEE INFOCOM'02 [C]. New York, USA, 2002, 1. 160 - 169.

[25] R Govindan, A Reddy. An analysis of internet inter-domain routing and route stability [A]. IEEE INFOCOM'97 [C]. Kobe, Japan, 1997. 850 - 857.

[26] T Griffin, G Wilfong. An Analysis of BGP convergence properties [A]. ACM SIGCOMM1999 [C]. Cambridge, MA, 1997. 277 - 288.

[27] J-J Pansiot, D Grad. On routes and multicast trees in the internet [J]. Computer Communication Review, 1998, 28(1): 41 - 50.

[28] S Ratnasamy, S McCanne. Inference of multicast routing trees and bottleneck bandwidths using end-to-end measurements [A]. IEEE INFOCOM 1999 [C]. New York, USA, 1999. 353 - 360.

[29] N G Duffield, J Horowitz. Multicast topology inference from end-to-end measurements [A]. in ITC2000 [C]. Monterey, CA, 2000, 27. 1 - 10.

[30] N G Duffield, J Horowitz F. Multicast topology inference from measured end-to-end loss[J]. IEEE Trans Inform Theory, 2002, 48: 26 - 45.

[31] Bestavros, J Byers, K Harfoush. Inference and labeling of metric-induced network topologies [A]. INFOCOM 2002 [C]. New York, USA, 2002, 2. 628 - 637.

[32] R Castro, M J Coates, M Gadhiok. Maximum likelihood network topology identification from edge-based unicast measurements [R]. USA:

Rice Univ,Oct 2001.

[33] R Castro, M Coates. Maximum likelihood identification of network topology from end-to-end measurement [R]. USA: Rice Univ, 2002.

[34] R Cáceres, N Duffield. Multicast-based inference of network-internal loss characteristics [J]. IEEE Trans Inform. Theory, 1999, 45: 2462 - 2480.

[35] N G Duffield, F Lo Presti. Inferring link loss using striped unicast probes [A]. IEEE INFOCOM 2001 [C]. Anchorage, Alaska, USA, 2001, 2. 915 - 923.

[36] V Paxson. End-to-end Internet packet dynamics [J]. IEEE/ACM Trans Networking, 1999, 7: 277 - 292.

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

[38] Y Tsang, M Coates. Passive network tomography using EM algorithms [A]. Proc 2001 IEEE Int Conf Acoust, Speech, and Signal Processing [C]. Salt Lake City: IEEE, 2001. 1469 - 1472.

[39] IETF RFC 2309, Recommendations on queue management and congestion avoidance in the Internet [S]. Apr 1998.

[40] R Nowak, M Coates. Unicast network tomography using the EM algorithm [R]. Houston, USA: Rice Univ, Dec 2001.

[41] A-G Ziotopolous, A O Hero. Estimation of network link loss rates via chaining in multicast trees [A]. IEEE Int Conf Acoust, Speech, and Signal Proc [C]. Salt Lake City, USA: IEEE, 2001. 2517 - 2520.

[42] UCB/LBNL/VINT network simulator ns (version 2) [OL]. http://www.isi.edu/nsnam/ns/.

[43] Y Tsang, M Coates. Passive unicast network tomography based on tcp monitoring [R]. USA: Rice University, 2000.

[44] M Coates, R Nowak. Networks for networks: Internet analysis using Bayesian graphical models [A]. Proc 2000 IEEE Neural Network for Signal Processing Workshop [C]. Sydney, Australia, 2000, 2. 755 - 764.

[45] Bestavros, K Harfoush. Robust identification of shared losses using end-to-end unicast probes [A]. Proc IEEE Int Conf Network Protocols 2000 [C]. Osaka, Japan, 2000. 22 - 33.

[46] M Coates, R Nowak. Network tomography for internal delay estimation [A]. Proc. 2001 IEEE Int Conf Acoust, Speech, and Signal [C]. Salt Lake City, USA, May 2001. 3409 - 3412.

[47] F Lo Presti, N G Duffield, et al. Multicast-based inference of network-internal delay distributions [J]. IEEE/ACM Tran on Networking, 2002, 10(6): 761 - 775.

[48] Y Tsang, M Coates, R Nowak. Nonparametric Internet tomography [A]. Proc 2002 IEEE Int. Conf. Acoust, Speech, and Signal Processing[C]. Orlando, USA, 2002 Vol. 2. 2045 - 2048.

[49] M F Shih, 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, USA, 2001. 3421 - 3424.

[50] M Coates. Sequential Monte Carlo inference of internal delays in non-stationary communication networks [J]. IEEE Trans Signal Processing 2002, 50: 366 - 376.

[51] N Duffield, F Lo Presti. Multicast inference of packet delay variance at interior network links [A]. IEEE INFOCOM 2000 [C]. Tel Aviv, Israel, 2000, 3. 1351 - 1360.

[52] M Adler, T Bu. Tree layout for internal network characterizations in multicast networks [OL]. http://www.cs.umass.edu/~micah/pubs/inference.ps.

[53] T Bu, N G Duffield, F Lo Presti, D Towsley. Network tomography on general topologies [A]. ACM SIGMETRICS 2002 [C]. Marina Del Rey, USA, 21 - 30.

[54] Geoffrey J McLachlan. The EM Algorithm and extensions [M]. New York, John Wiley, 1997. 120 - 211.

[55] V Jacobson. Congestion avoidance and control [A]. ACM SIGCOMM'88 [C]. Stanford, CA, 1988. 314 - 329.

[56] S Keshav. A Control-Theoretic Approach to Flow Control [A]. ACM SIGCOMM'91 [C]. Zurich, Switzerland, 1991. 3 - 15.

[57] R L Carter, M E Crovella. Server selection using dynamic path characterization in wide-area networks [A]. IEEE INFOCOM'97 [C]. Kobe, Japan, 1997, 3. 1014 - 1021.

[58] C Dovrolis, P Ramanathan, D Moore. What do packet dispersion techniques measure [A]. IEEE INFOCOM 2001 [C]. Anchorage, Alaska, USA, 2001, 2. 905 - 914.

[59] Vern Paxson. Measurements and Analysis of End-to-End Internet Dynamics [D]. Berkeley: University of California, April 1997.

[60] Attila P'asztor, Darryl Veitch, The packet size dependence of packet-pair like methods [A]. IWQoS'2002 [C]. Miami Beach, Florida, USA, 2000. 204 - 213.

[61] K Harfoush, A Bestavros. Measuring Bottleneck Bandwidth of Targeted Path Segments [R]. USA: Boston University, July 2001.

[62] F O'Sullivan. A statistical perspective on ill-posed inverse problems [J]. Statistical Science, 1986, 1(4): 502 - 527.

作者简介:

林　宇　男,1976 年 12 月生于福建省浦城县,1998 年毕业于北京邮电大学,获机械电子工程和通信工程双学士学位,现为北京邮电大学交换和网络国家重点实验室讲师,研究兴趣包括 Internet 测量、无线网络、移动组播、TCP 建模.

程时端　女,1940 年生于上海,北京邮电大学交换与网络国家重点实验室教授,博士生导师,主持过十余项重大科研项目,发表文章 200 余篇,专著 2 册,译著 3 册,1992-1999 年间,任 863 网络与交换专家组组长,目前研究方向包括 TCP/IP,协议工程,业务量工程,移动互联网性能分析和服务质量控制等.

邬海涛　男,1976 年 9 月生于江西南昌,北京邮电大学交换与网络国家重点实验室博士研究生,1998 年毕业于北京邮电大学电信工程工程系,目前研究方向为宽带网络服务质量,TCP/IP 改进,区分服务,流控,拥塞控制及无线分组网络性能.