基于多源NT的链路利用率推断

段 琪 蔡皖东 田广利

(西北工业大学计算机学院 西安 710072)

摘 要 链路利用率是网络运行状态的重要指标。目前基于NT技术的链路性能推断一般是采用单个源节点,但多源NT具有更多优点。研究了多源NT的链路利用率估计技术;提出汇合测量方法,并证明利用此测量方法,多源NT链路利用率是可辨识的,同时给出测量子图选取的充要条件;提出采用EM算法的链路利用率的极大似然估计方法;最后通过模型仿真和网络仿真对推断方法的有效性进行了验证。

关键词 NT,链路利用率,链路性能推断,EM算法

中图法分类号 TP393 **文献标识码** A

Research on Link Utilization Inference Technology Based on the Multiple Source Network Tomography

DUAN Qi CAI Wan-dong TIAN Guang-li

(School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China)

Abstract The link utilization is an important parameter to describe the running state of the network. At present the research on the link parameter inference technology of network tomography is based on the single source measurements. And the multiple source network tomography has many advantages. The link utilization estimation technology based on multiple source tomography was researched. The joining measurement method was proposed and multiple source link utilization is identifiable if the network is measured by the new method. Furthermore, the necessary and sufficient condition of the measurement sub-network selection to make the link identifiable was proposed. At last, the maximum likelihood estimation of link utilization computed by the EM algorithm was derived and the effectiveness of that was validated by the model simulation and network simulation results.

Keywords Network tomography, Link utilization, Link parameter inference, EM algorithm

1 引言

网络可管理性取决于网络可测量性。通过网络测量,能够对网络流量、故障、性能等情况进行分析和科学评估,为网络系统的资源配置、性能改进和系统维护提供科学决策依据。国内外学术界和研究机构对网络测量技术进行了大量的研究。总的来看,网络测量方法有两种,即网络内部直接测量和端到端测量。基于内部测量的链路信息获取需要内部节点配合,而各个ISP基于自身安全等因素考虑,使得普通网络用户无法获取测量数据。基于端到端的测量的链路性能推断成为研究热点,并取得了关键的进展。NT(network tomography)是近年来国际上提出的一种新的网络测量技术[1,2],在网络拓扑固定的条件下,根据网络外部的测量信息来分析和推断网络内部的性能以及网络拓扑。目前,NT技术研究主要采用单个测量源节点(简称单源断层扫描技术NT),如AT&T和马萨诸塞州立大学的MINC(Multicast-based Inference of Network-internal Characteristics)项目[3]以及莱斯大学研究的单播NT项目[4]等。MINC项目的Caceres首先提出了采用组播技术引入丢包率相关性[5],在拓扑已知的情况下进行丢包率推断的算法,随后N. G. Duffield, Lo Presti等人又提出了基于组播端到端测量的时延分布推断算法[6,7]。莱斯大学的Coates, Nowak等人在组播的基础上,提出了单播"包对"(Unicast Packet Pairs)模拟组播的测量方法来进行丢包率推断[8],并实现了单播NT技术。

多源测量是采用多个源节点对多个目的节点的端到端测量,是单源测量方法的扩展。设源节点和目的节点的个数分别为M,N,则测量覆盖的网络结构称为M-by-N网络结构。由于单源测量只能覆盖一个树型网络结构,往往不能满足测量的需求。首先,单源NT技术只能推断出树型拓扑结构,丢失了较多拓扑结构信息,即使采用多个源节点分别进行测量,由于各个逻辑拓扑中的节点和链路没有统一标识,因此这些独立的树型拓扑无法合并成图型结构。如图1所示,当网络的逻辑拓扑为图1(a)时,分别采用3个源节点进行测量和拓扑推断,则获取到3个单独拓扑结构,分别为图1(b)、图1(c)和图1(d),这不仅丢失节点i_3,i_7,i_8和相关链路的信息,而且图1(b)、图1(c)和图1(d)所示的网络结构无法合并到图1(a)所示的网络结构。采用多源断层扫描,则可以推断出图1(a)所示的网络结构。因此多源断层扫描比单源断层扫描推

到稿日期:2009-01-20 返修日期:2009-05-22 本文受教育部博士点基金(200806990030)资助。

段 琪(1983—),男,博士生,主要研究方向为网络安全及仿真,E-mail:duq0118@hotmail.com;蔡皖东(1955—),男,教授,博士生导师,主要研究方向为分布式计算、网络信息安全与对抗。

断出的拓扑结构更加接近真实网络结构。其次，多源断层扫描可以推断出更多的链路性能。图 1(a)中的链路 3,4,5,6,8,9,10,14,15,17 是无法通过单源断层扫描进行推断的。

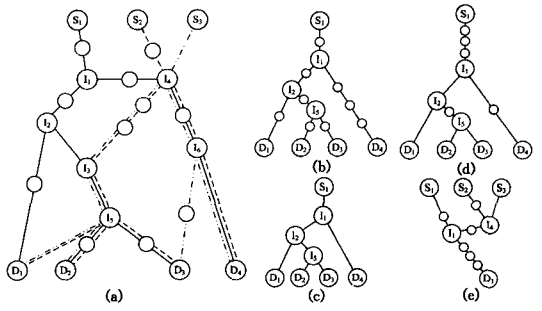

图 1　单源逻辑拓扑结构和多源逻辑拓扑结构

最近，多源 NT 技术也受到学者的关注，研究成果主要集中在网络拓扑推断上。文献[9]提出，如果网络中每个 1-by-2 子网和每个 2-by-1 子网的链路性能是可以辨识的，则可以推断出 M-by-N 网络的逻辑拓扑，但文献中没有提出如何辨识一个 2-by-1 子网的方法。如何辨识一个 2-by-1 网络的链路性能是多源 NT 的关键问题。如通过辨识图 1-2(e)子网，可以获取到链路 4 的性能。目前，多源 NT 技术尚未得到较好解决，特别是尚未见到基于多源 NT 的链路性能技术。因此，基于多源 NT 链路性能推断问题具有研究意义。

随着硬件的升级，Internet 中较大部分网络具有较轻的流量负载，丢包现象较少，链路利用率和时延分布成为主要的性能指标。同时，链路利用率与可用带宽和时延也具有相关性。链路利用率可以大致描述网络的负载情况。

本文研究了基于多源 NT 链路利用率推断技术。主要贡献包括：(1)提出汇合测量方法，使从两个源节点到一个目的节点探测包在共享路径上的链路利用率具有相关性；(2)证明利用包对测量方法和汇合测量方法，多源 NT 链路利用率是可辨识的，同时给出测量子图选取的充要条件；(3)提出采用 EM 算法的链路利用率的极大似然估计方法。最后通过模型仿真和网络仿真对推断方法的有效性进行了验证。

2　网络模型与假设

首先假设测量网络具有以下特性：(1)网络节点之间不存在多路径；(2)网络中路由器对发送到同一个目的节点的数据包转发采用先进先出策略；(3)测量过程中网络拓扑稳定。Internet 中基本上满足前两个假设，当网络测量时间比较短时，第 3 个假设也基本满足。

用 $G=(V,L)$来描述网络拓扑，其中 V 是节点集合，L 是连接节点的链路集合。节点集合 S,R 和 I 分别表示源节点集合、目的节点集合和中间节点集合，$V=S\cup R\cup I$。入度大于 1 的中间节点叫汇合节点，出度大于 1 的中间节点叫分叉节点。令 $M=|S|$，$N=|R|$，$G=(V,L)$称为 M-by-N 网络，简记为 G_D^S。节点 i 到节点 j 的路径用 $P[i,j]$表示。$P[k,i]$和 $P[k,j]$的共享路径用 $P[k;i,j]$表示，分叉节点用 $b(k;i,j)$表示；$P[i,k]$和 $P[j,k]$的共享路径用 $P[i,j;k]$表示，汇合节点用 $j(i,j;k)$表示。$T_k=(V(k),L(k))$表示以 k 为源节点的子树，子树中非发送节点 i 的父节点用 $f(i)$表示。$IT_k=(V(k),L(k))$表示以 k 为目的节点的倒子树，子树中非目的节点 i 的子节点用 $d(i)$表示。

探测包在链路上的延时包括传播时延(propagation delay)、传输时延(transmission delay)、处理时延(processing delay)和排队时延(queuing delay)。对于特定的数据包，前 3 项是固定的，因背景流量的存在和流量的突发特性使得在缓存中排队的数据包数目不确定，即队列中的等待时间是随机变化的，本文主要考虑报文在队列中的等待时间。用随机变量 $\lambda(p)\in[0,\infty]$表示探测报文在通过路径 p 时的队列等待时间。$\alpha(p)=P[\lambda(p)=0]$表示报文通过路径 p 时没有队列等待时间的概率，简称为链路空闲率。只有当链路空闲，报文通过该链路时才没有队列延时，本文称没有排队时延的探测包为空闲包。用 $\bar{\alpha}=1-\alpha$ 表示链路利用率。为了公式表达的方便，公式中主要采用链路空闲率表达，$\alpha(P[i,j])$简记作 $\alpha_{i,j}$。链路 l 上的链路是否空闲，用随机变量 $X(l)$表示。$X(l)=0$，就表示报文在链路 l 没有排队时延；如果 $X(l)=1$，就表示报文在链路 l 有排队时延。

3　测量方法

目前在单源 NT 中普遍采用的测量方法是，将 1-by-N 网络分解为多个 1-by-2 子网，分别采用包对进行测量，如图 2(a)所示。由于包对中两个报文间隔很小，可以认为在公共路径 $P[s_k,b]$上两个报文的时延相等，利用此相关性可以证明 1-by-2 网络链路利用率是可辨识的(identifiable)，进一步可以证明 1-by-N 网络链路利用率是可辨识的[10](本文称为定理 1)。对于 2-by-1 子网链路时延推断，本文提出汇合测量方法，并证明基于此测量方法，2-by-1 子网是可辨识的。

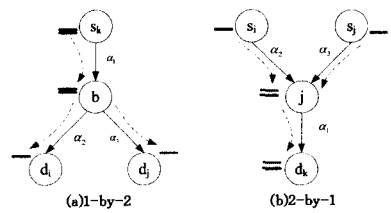

图 2　采用包对对 1-by-2 子网测量和采用汇合测量法对 2-by-1 子网测量

设 2-by-1 子网如图 2(b)所示，源节点 s_i 和 s_j 分别发送探测包 $pk_{i,k}$和 $pk_{j,k}$。由于 $pk_{i,k}$和 $pk_{j,k}$在路径 $P[s_i,j]$和 $P[s_j,j]$上的时延是随机的，无法保证 $pk_{i,k}$和 $pk_{j,k}$同时到达汇合点而形成包对。但是通过汇合测量方法，可以保证在链路空闲时 $pk_{i,k}$和 $pk_{j,k}$同时到达汇合点。原理是两个源节点向一个目的节点发送的两个大小相同的空闲探测包，如果(几乎)同时到达目的节点，由于两个探测包在共享路径上的固定时延相同，因此它们也必然(几乎)同时到达了汇合节点。使两个源节点向一个目的节点发送的两个相同大小的空闲探测包同步到达汇合节点的方法是，将后达到目的节点的探测包的发送时刻提前 δrm，δrm 表示两个包的到达时刻差。

设 s_i 和 s_j 到 d_k 的探测报文的发送时间和接收时间分别为 $ST_{i,k}$，$ST_{j,k}$和 $RT_{i,k}$，$RT_{j,k}$，进行 K 次测量，计算固定时延，分别为 $\varepsilon_{i,k}=\min\{RT_{i,k}(n)-ST_{i,k}(n)\}$和 $\varepsilon_{j,k}=\min\{RT_{j,k}(n)-ST_{j,k}(n)\}$。调整发送节点 s_j 的发送时间，使 $ST_{j,k}=ST_{i,k}+\varepsilon_{i,k}-\varepsilon_{j,k}$，对 2-by-1 子网进行测量。本文称此测量方法为汇合测量法。

4　链路利用率的可辨识性

两个空闲探测包由于同时到达汇合节点，构成包对，因此

在共享路径上的行为具有相关性。定理 2 表明该测量方法可以解决 2-by-1 子网链路时延分布推断问题。

引理 1 路径 $P[i,j]$包含 $P[i,k]$和路径 $P[k,j]$，已知其中两个路径的利用率，则可以推导另外一个路径的利用率。

证明：因为 $\alpha(P[i,j])=\alpha(P[i,k])\cdot\alpha(P[k,j])$，且 $\alpha(P[i,k])$和 $\alpha(P[k,j])$相互独立，已知其中两个路径利用率，可以求另外一个路径利用率。

定理 1 采用包对测量，对于任意的 1-by-2 网络或 1-by-N 网络，链路利用率是可辨识的。

定理 2 采用汇合测量法，对于任意的 2-by-1 子网，链路利用率是可辨识的。

证明：如图 2(b)所示，根据单个源的测量结果可以得到

$$P(X(P[s_i,d_k])=0)=\alpha_{s_i,j}\alpha_{j,d_k} \tag{1}$$

$$P(X(P[s_j,d_k])=0)=\alpha_{s_j,j}\alpha_{j,d_k} \tag{2}$$

根据汇合测量方法的测量结果可以得到

$$P(X(P[s_i,d_k])=0,X(P[s_j,d_k])=0)=\alpha_{s_i,j}\alpha_{s_j,j}\alpha_{j,d_k} \tag{3}$$

根据式(1)、式(2)和式(3)可以得到

$$\alpha_{j,d_k}=\frac{P(X(P[s_j,d_k])=0)\cdot P(X(P[s_i,d_k])=0)}{P(X(P[s_i,d_k])=0,X(P[s_j,d_k])=0)} \tag{4}$$

$$\alpha_{s_i,j}=\frac{P(X(P[s_i,d_k])=0,X(P[s_j,d_k])=0)}{P(X(P[s_j,d_k])=0)} \tag{5}$$

$$\alpha_{s_j,j}=\frac{P(X(P[s_i,d_k])=0,X(P[s_j,d_k])=0)}{P(X(P[s_i,d_k])=0)} \tag{6}$$

根据式(4)、式(5)和式(6)可以无偏估计链路中 3 个链路的链路利用率，因此 2-by-1 子网链路利用率是可辨识的。

需要注意的是，汇合测量方法并不是“逆向的包对测量”。采用包对测量 1-by-2 子网，可以得到更多的关于链路利用率的方程，分别为

$$P(X(P[s_k,d_i])=0)=\alpha_1\alpha_2 \tag{7}$$

$$P(X(P[s_k,d_j])=0)=\alpha_1\alpha_3 \tag{8}$$

$$P(X(P[s_k,d_i])=0,X(P[s_k,d_j])=0)=\alpha_1\alpha_2\alpha_3 \tag{9}$$

$$P(X(P[s_k,d_i])=0,X(P[s_k,d_j])=1)=\alpha_1\alpha_2\bar{\alpha}_3 \tag{10}$$

$$P(X(P[s_k,d_i])=1,X(P[s_k,d_j])=0)=\alpha_1\bar{\alpha}_2\alpha_3 \tag{11}$$

$$P(X(P[s_k,d_i])=1,X(P[s_k,d_j])=1)=\bar{\alpha}_1+\alpha_1\bar{\alpha}_2\bar{\alpha}_3 \tag{12}$$

本文在后面的推断算法中，采用式(9)、式(10)、式(11)。另一方面，在汇合测量方法中，则没有类似式(10)、式(11)和式(12)的公式成立。式(1)、式(2)和式(3)刚好满足对数方程组满秩的条件。

定理 3 采用汇合测量法，对于任意的 M-by-1 网络是可辨识的。

证明：用归纳法证明。设目的节点为 k，M-by-1 网络可以表示为 $IT_k=(V(k),L(k))$。首先，选取发送节点为 $i,j\in S_k$ 且 $j(i,j;k)=f(k)$，根据定理 2，可知 $X(P[f(k),k])$可辨识。假设 $v\in I_k$，$X(P[v,k])$可以辨识，选取发送节点为 $i,j\in S_v$ 且 $j(i,j;k)=v$，根据定理 2，可知 $X(P[f(v),k])$可辨识，进一步 $X(P[f(v),v])$可辨识，进行归纳可得任意汇合节点间的链路利用率可辨识。最后选取发送为任意的兄弟节点 $i,j\in S_k$ 且 $j(i,j;k)=d(i)$，根据定理 2，可知 $X(P[i,d(i)])$和 $X(P[j,d(j)])$可辨识，即任意发送节点和发送节点的子节点间的链路利用率可以辨识。综合以上两步可知 M-by-1 网络是可辨识的。

定理 4 采用包对测量和交叉汇合测量法，对于任意的 M-by-N 网络是可辨识的。

证明(简要)：设任意一条路径 $P[s,d]$上的中间节点(包括分叉节点和汇合节点)为 $i_k,k=1,2,\cdots,|P[s,d]|-1$，根据定理 1 和定理 3 可以得到 $X(P[s,i_k]),\forall k=1,2,\cdots,|P[s,d]|-1$，根据引理，利用 $X(P[s,i_k])$和 $X(P[s,i_{k+1}])$可以得到路径上任意一个链路的利用率 $X(P[i_k,i_{k+1}])$，所以整个 M-by-N 网络是可辨识的。

5 测量子图的选取

对 M-by-N 网络中每个 1-by-2 子网和 2-by-1 子网都进行测量，将使测量流量和推断算法复杂度随被测网络节点的增加呈指数倍增加。实际上，很多对 1-by-2 子网和 2-by-1 子网的测量是不必要的。定理 5 给出了选取测量子图的依据。

定理 5 任意的网络 G_N^S，测量子网集合记为 $\mathcal{G}$，链路时延分布是可辨识的充分必要条件是：(1)对于任意一个分叉节点 b，至少存在一个分叉节点为 b 的 1-by-2 子网属于 $\mathcal{G}$；(2)对于任意一个汇合节点 j，至少存在一个汇合节点为 j 的 2-by-1 子网属于 $\mathcal{G}$；(3)对于任意一个发送节点 s 至少存在一个发送节点为 s 的子网属于 $\mathcal{G}$；(4)对于任意一个目的节点 d 至少存在一个目的节点为 d 的子网属于 $\mathcal{G}$。

充分性的证明类似前面定理的证明。其必要性较明显，此处不再证明。

6 链路利用率推断算法

采用极大似然估计来推断 M-by-N 网络内部链路时延分布。对于任意 1-by-2 子网 $G_{i,j}^k\in\mathcal{G}$，采用包对进行测量，记 $\vec{Y}_{k;i,j}=(Y_{k,i},Y_{k,j})=(X(P[k,i]),X(P[k,j]))$，测量值记为 $\vec{y}_{k;i,j}$，取值空间为 $\{0,1\}^2$。对于 2-by-1 子网 $G_k^{i,j}\in\mathcal{G}$，在 $ST_{j,k}=ST_{i,k}+\epsilon_{i,k}-\epsilon_{j,k}$条件下，采用汇合测量方法，记 $Y_{i,k}=X(P[i,k]),Y_{j,k})=X(P[j,k])$，测量值记为 $y_{i,k}$和 $y_{j,k}$。引入变量 $Y_{i,j;k}=(Y_{i,k},Y_{j,k})$，记 $O=(0,0)$和 $\bar{O}=\{(0,1),(1,0),(1,1)\}$，$Y_{i,j;k}$的取值空间为 O 和 $\bar{O}$。

记 $N(y)$表示测量值为特定值 y 的次数。测量值概率记为 $g(y;\alpha)=P(Y=y,\alpha)$。所以所有的观测值对数似然函数为

$$\begin{aligned}\ell(y;\alpha)=\log g(y;\alpha)=&\sum_{\forall G_{i,j}^k\in\mathcal{G}}\sum_{\vec{y}_{k;i,j}\in(0,1)^2}N(\vec{y}_{k;i,j})\log g(\vec{y}_{k;i,j};\alpha)+\sum_{\forall G_k^{i,j}\in\mathcal{G}}\Big(\sum_{y_{k;i,k}\in\{0,1\}}N(y_{i,k})\log g(y_{i,k};\alpha)+\sum_{y_{j,k}\in(0,1)}N(y_{j,k})\log g(y_{j,k};\alpha)+\sum_{y_{i,j;k}\in\{O,\bar{O}\}}N(y_{i,j;k})\log g(y_{i,j;k};\alpha)\end{aligned} \tag{13}$$

在得到测量集合 y 后，利用 MLE 的方法求 α 的估计值为

$$\hat{\alpha}=\arg\max_\alpha\ell(y;\alpha) \tag{14}$$

上式无法直接求解。本文采用 EM 算法求解，为此引入测量包在各个链路的利用率作为不可见数据以简化上式。对于 $\forall G_{i,j}^k\in\mathcal{G}$ 用 $D_{k;i,j}$ 表示探测包每条链路所经历的时延集合，同样对于 $\forall G_k^{i,j}\in\mathcal{G}$ 用 $D_{i,j;k}$ 表示探测包每条链路所经历的时延集合，$D=\{D_{i,j;k},D_{k;i,j};\forall G_{i,j}^k\in\mathcal{G},\forall G_k^{i,j}\in\mathcal{G}\}$。因此完全数据的对数似然函数可以表示为

$$\ell(y,D;\alpha)=\log g(y,D;\alpha)=\log g(y|D;\alpha)+\log g(D;\alpha) \tag{15}$$

实际上，在已知探测包在所有链路上的利用率集合 D 的

条件下，必然获得相应的观测值的 y，即 $g(y|D;\alpha)=1$，$\log g(y|D;\alpha)=0$。因此

$$\ell(y,D;\alpha)=\log g(D;\alpha)=\sum_{\forall G_{i,j}^{k}\in\mathcal{G}}\log g(D_{ki,j};\alpha)+\sum_{\forall G_{k}^{i,j}\in\mathcal{G}}\log g(D_{i,j_1k};\alpha)$$
$$=\sum_{l\in L}N_l(0)\log\alpha_l+N_l(1)\log\bar{\alpha}_l \tag{16}$$

在等式中，$N_l(0)$（或 $N_l(1)$）表示链路 l 上的时延等于 0（或 1）的报文数量。在时延集合 D 已知的条件下，α_l 的 MLE 估计值 $\hat{\alpha}_l$ 可以通过上式进行求导来获得。

$$\hat{\alpha}_l=\frac{N_l(0)}{N_l(0)+N_l(1)} \tag{17}$$

时延集合 D 和 $N_l(d)$（$d\in\{0,1\}$）是不可见数据，利用 EM 算法进行求解，该算法使用 α_l 和 $N_l(d)$ 的前次估计值来推断它们当前的估计值。这样，通过有限步的迭代来得到收敛的 $\hat{\alpha}_l$ 值。为此，设 $\hat{\alpha}_l^{(e)}$ 表示 α_l 第 e 步的概率估计值，其中 e 为正整数。其过程如下。

(1)初始化

选择所有链路的初始时延分布 $\hat{\alpha}_l^{(0)}$。由于缺乏 α_l 的先验信息，可假设 $\hat{\alpha}_l^{(0)}=0.5$。

(2)E 步(Expectation)

在已知完整的测量值集合 y 和当前第 e 步估计值 $\hat{\alpha}_l^{(e)}$ 的条件下，计算对数似然函数的条件期望 $\hat{\alpha}_l''$ 为

$$Q(\hat{\alpha}_l'',\hat{\alpha}_l^{(e)})=E_{\hat{\alpha}_l^{(e)}}[\ell(y,D;\hat{\alpha}_l'')|y]=\sum_{l\in L}(\hat{N}_l(0)\log\hat{\alpha}_l''+\hat{N}_l(1)\log\hat{\bar{\alpha}}_l'') \tag{18}$$

式中，

$$\hat{N}_l(d)=E_{\hat{\alpha}_l^{(e)}}[N_l(d)|y]=\sum_{\forall G_{i,j}^{k}\in\mathcal{G}}\sum_{\vec{y}_{ki,j}\in Y_{ki,j}}N(\vec{y}_{ki,j})\log g(X(l)=d|\vec{Y}_{ki,j}=\vec{y}_{ki,j};\hat{\alpha}^{(e)})+\sum_{\forall G_{k}^{i,j}\in\mathcal{G}}\Big(\sum_{y_{k,i}\in Y_{k,i}}N(y_{k,i})\log g(X(l)=d|Y_{k,i}=y_{k,i};\hat{\alpha}^{(e)})+\sum_{y_{k,j}\in Y_{k,j}}N(y_{k,j})\log g(X(l)=d|Y_{k,j}=y_{k,j};\hat{\alpha}^{(e)})+\sum_{y_{i,j_1k}\in\{0,\bar{0}\}}N(y_{i,j_1k})\log g(X(l)=d|Y_{i,j_1k}=y_{i,j_1k};\hat{\alpha}^{(e)})\Big) \tag{19}$$

(3)M 步(Maximization)

根据上式估计的 $\hat{N}_l(d)$，从而获得第 $e+1$ 步时延分布概率的估计值为

$$\alpha_l^{(e+1)}(d)=\operatorname{argmax}_{\hat{\alpha}_l}F(\hat{\alpha}_l'',\hat{\alpha}_l^{(e)})=\frac{\hat{N}_l(d)}{\sum_{d\in Q}\hat{N}_l(d)} \tag{20}$$

(4)迭代

交替地使用 E 步和 M 步进行计算，直到估计的延迟分布概率达到收敛状态，即 $|\hat{\alpha}_l^{(e+1)}-\hat{\alpha}_l^{(e)}|\leqslant threshold$。其中，$threshold$ 表示预先设定的门限值。

利用 EM 算法，求解速度相对较快一些，虽然似然函数可能存在多个稳定点，但仅有一个是最大值点。使用 EM 算法可能收敛于一个局部最大值点，所以要求算法要选取合适的初始值。EM 算法的复杂性主要是由 E 步中计算条件期望 $\hat{N}_l(d)$ 所决定的，其复杂性随着探测报文对的数量和网络拓扑链路数的增加而增加，计算的复杂性明显增大。

7 仿真

7.1 模型仿真

首先采用模型仿真来验证算法的有效性。模型仿真采用 2 种 2-by-2 网络，结构如图 3 所示。为减少算法复杂度，同时验证定理 5 的正确性，实验采用尽量少的测量子网集。其中 a,b 的测量子网集 $\mathcal{G}$ 分别为 $\{G_{d_1,d_2}^{s_2},G_{d_1,d_2}^{s_2},G_{d_1}^{j_1,s_2},G_{d_2}^{j_1,s_2}\}$，$\{G_{d_1,d_2}^{s_2},G_{d_1,d_2}^{s_2},G_{d_1}^{j_1,s_2},G_{d_2}^{j_1,s_2}\}$。在 G 中，每个链路的利用率在仿真前随机产生。仿真时，测量值根据路径上各个链路的利用率随机产生时延。每个 1-by-2 和 2-by-1 测量子网产生 10000 测量值。设置推断算法初始链路为 0.5，门限 $threshold=0.001$，对两种网络结构进行 500 次仿真。仿真结果表明，493 次仿真中，推断算法可以无偏差地推断出各个链路的利用率，平均迭代次数为 26。其余 7 次，EM 算法收敛于局部最优点，平均误差值为 0.08。

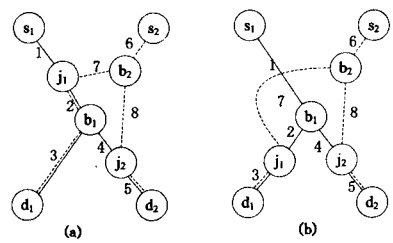

图 3 仿真使用的 2 种 2-by-2 网络结构

7.2 网络仿真

采用网络仿真工具 ns-2 对图 3 中 2 种结构进行仿真。设置每条逻辑链路上有 3 个物理链路，物理链路带宽为 2Mb/s 到 10Mb/s，每个物理链路背景流量为泊松流，链路利用率为 0.1～0.3，数据包的大小为 128byte。对于 1-by-2 测量子网，探测包对为两个间隔为 0.01ms、周期为 1s 的 CBR 数据流，每个 1-by-2 测量子网产生 400 个测量值。对于 2-by-1 测量子网，两个源节点分别发送周期为 1s 的 CBR 数据流。设置推断算法初始链路利用率为 0.5，门限 $threshold=0.01$。每个网络结构运行 200 次，得到两种网络结构平均相对误差分别为 2.4%和 2.1%。

结束语 本文研究了单播多源 NT 链路利用率估计技术，提出了新的测量方法，并证明了技术的可识辨性，同时给出了选取测量子网的充要条件，最后通过仿真验证了该技术的有效性。本文的研究扩展了单源断层扫描技术推断链路的能力和适用范围。

参考文献

[1] Coates M, Alfred III H, et al. Internet Tomography[J]. IEEE Signal Processing Magazine, 2002, 19(3): 47-65

[2] Castro R, Coates M, et al. Internet tomography: Recent developments[J]. Statistical Science, 2004, 52(3): 499-517

[3] http://www-net.cs.umass.edu/minc/

[4] Coates M, Nowak R. Unicast Network Tomography using EM Algorithms[R]. TR-0004. ECE Department, Rice University, Sep. 2000

(下转第 114 页)

注入 10,000 个事件。这 10,000 个事件中,有 50%的事件在系统中是有满足的订阅的,并且满足的订阅是随机的。订阅集合中订阅的取法如下:首先从 1000 棵订阅树的森林中随机选择出一棵树,然后从这个树中选出 1 到 100 之间随机数目的订阅,再继续从这个森林中随机选出另外一棵树并选出随机数目的订阅。这个过程一直持续下去,直到选出需要数目的订阅为止。这样,从一棵树中选出的订阅的平均数目为 50,并且这些订阅都是存在覆盖关系的,在实验中记录事件匹配的时间和系统内存的占用开销。

图 4 显示了不同系统在该实验中的时间和空间开销。其中,X 轴表示原子订阅的数目,原子订阅从 1,000 增长到 10,000。实验结果表明,与 SIENA 相比,OncePubSub 在原子订阅匹配上具有很大的优势。当系统中包含有 10,000 条订阅时,ONCE 比 SIENA 在时间开销上低了 87.1%,在空间开销上低了 21.0%。

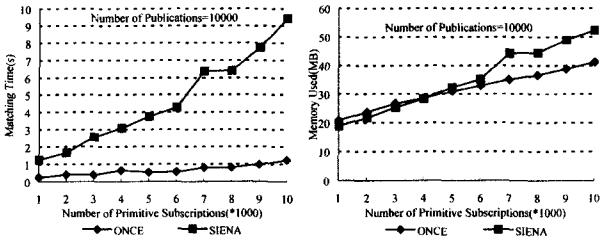

图 4 实验 1 记录的原子匹配时间和空间开销

在发布/订阅系统的实际运行中,订阅与事件匹配、订阅的添加、订阅取消经常会交织进行。实验 2 就是对这一场景进行模拟,进而记录系统在这些操作中的性能和开销。

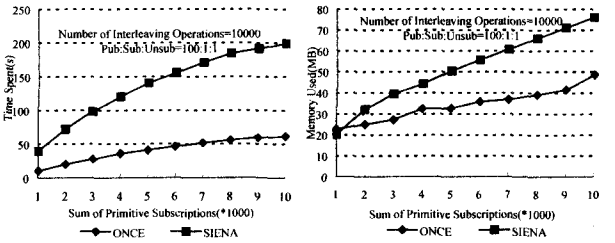

图 5 实验 2 记录的操作的时间和空间开销

首先向系统中注入一定数目的订阅,然后开始下面场景的模拟过程:系统接收到一系列订阅请求,然后又接收到相同数目的取消订阅请求,这个请求数目从{10,20,30,40,50}中随机选取。然后,系统将接收到数量为订阅请求数目 100 倍的事件,进行订阅匹配。订阅、取消的订阅和事件的格式都符合上一个实验所作的规定。3 种操作一直循环执行,直到操作总数目到达 10,000 为止。在这一过程中,系统中的订阅总数总是保持恒定。我们记录这 10,000 个操作的时间和空间开销。

如图 5 所示,OncePubSub 在这种交织操作中的性能也有很大的优势。例如,当系统中订阅数目维持在 10,000 时,OncePubSub 执行 10,000 次交织操作所用的时间仅为 SIENA 的 30.1%,所使用的内存仅为 SIENA 的 60.0%。通过观察图 4 和图 5 的实验数据,可以看出 OncePubSub 在大量的订阅和事件数据下具有良好的可伸缩性。

结束语 本文设计了一种高效的原子订阅管理和匹配算法,并在发布/订阅系统 OncePubSub 中实现。该算法是将订阅组织成订阅森林结构,这样可以迅速地将大量不匹配的订阅过滤出来,同时对谓词加以组织,对其进行有效的存储和高效匹配,从而达到快速匹配事件的目的。该算法也能够高效地支持大量订阅的动态添加和删除。实验验证了上述优点,并且与 SIENA 相比,我们的系统在时间和空间开销上具有很大的优势。

参 考 文 献

[1] Kale S, Hazan E, Cao F, et al. Analysis and Algorithms for Content-based Event Matching[C]// the 4th International Workshop on Distributed Event-based Systems. Columbus, Ohio, USA, June 2005

[2] Ashayer G, Leung H K Y, Jacobsen H A. Predicate Matching and Subscription Matching in Publish/Subscribe Systems[C]// the 22nd International Conference on Distributed Computing Systems Workshops. Vienna, Austria, July 2002

[3] Aguilera M K, Strom R E, Sturman D C, et al. Matching Events in a Content-based Subscription System[C]// the 18th ACM Symposium on Principles of Distributed Computing. Atlanta, GA, USA, May 1999

[4] Campailla A, Chaki S, Clarke E, et al. Efficient Filtering in Publish-Subscribe Systems Using Binary Decision Diagrams[C]// the 23rd International Conference on Software Engineering. Toronto, Ontario, Canada, May 2001

[5] Li G, Huo S, Jacobsen H. A Unified Approach to Routing, Covering and Merging in Publish/Subscribe Systems Based on Modified Binary Decision Diagrams[C]// the 25th International Conference on Distributed Computing Systems. Columbus, Ohio, USA, June 2005

[6] Zhao X C, Jin B H, Yu S, et al. Composite Subscription and Matching Algorithm for RFID Applications[C]// the 22nd IEEE International Conference on Advanced Information Networking and Applications. Ginowan, Okinawa, Japan, March 2008

[7] Carzaniga A, Rosenblum D S, Wolf A L. Design and Evaluation of a Wide-area Event Notification Service[J]. ACM Transaction on Computer Systems, 2001, 19(3): 332-383

(上接第 88 页)

[5] Caceres R, Duffield N G, Horowitz J, et al. Multicast-Based Inference of Network-Internal Characteristics Accuracy of Packet Loss Estimation[C]// IEEE INFOCOM. New York, 1999

[6] Duffield N G, Presti L F, Paxson V, et al. Inferring link loss using striped unicast probes[C]// IEEE INFOCOM 2001. Anchorage, Alaska, 2001, 2: 915-923

[7] Presti F L, Horowitz J, Towsley D, et al. Multicast-based inference of network-internal delay distributions[J]. IEEE/ACM Transactions on Networking, 2002, 10(6): 761-775

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

[9] Bestavros J, Byers W, Harfoush K A. Inference and labeling of metric-induced network topologies[J]. IEEE Transactions on Parallel and Distributed Systems, 2005, 16(11): 1053-1065

[10] Lawrence E, Michailidis G, Nair V. Network Delay Tomography Using Flexicast Experiments[J]. Journal of Royal Statistical Society Series B, 2006, 68(5): 785-813