ISSN 1000-9825, CODEN RUXUEW
Journal of Software, Vol.17, No.1, January 2006, pp.108−116
DOI: 10.1360/jos170108
© 2006 by *Journal of Software*. All rights reserved.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

端到端的可用带宽测量方法*

刘 敏[1,4+], 李忠诚[1], 过晓冰[2], 邓 辉[3]

[1](中国科学院 计算技术研究所,北京 100080)

[2](IBM 中国研究中心,北京 100085)

[3](日立(中国)有限公司研究开发中心,北京 100004)

[4](中国科学院 研究生院,北京 100049)

An End-to-End Available Bandwidth Estimation Methodology

LIU Min[1,4+], LI Zhong-Cheng[1], GUO Xiao-Bing[2], DENG Hui[3]

[1](Institute of Computing Technology, The Chinese Academy of Sciences, Beijing 100080, China)

[2](IBM China Research Laboratory, Beijing 100085, China)

[3](Hitachi (China) Ltd. R&D Center, Beijing 100004, China)

[4](Graduate School, The Chinese Academy of Sciences, Beijing 100049, China)

+ Corresponding author: Phn: +86-10-62565533 ext 9240, E-mail: liumin@ict.ac.cn, http://www.ict.ac.cn

Liu M, Li ZC, Guo XB, Deng H. An end-to-end available bandwidth estimation methodology. *Journal of Software*, 2006,17(1):108−116. http://www.jos.org.cn/1000-9825/17/108.htm

Abstract: A majority of current bandwidth estimation methodologies rely on the principle of the bottleneck spacing effect. Based on the concept of packet dispersion, many packet pair/packet train techniques were presented to estimate capacity/available bandwidth. However, these methods failed for measurement on high capacity path, because they could not measure bandwidth beyond the source node's maximum sending rate. In addition, current methodologies do not consider the effect of cross traffic routing on available bandwidth estimation. This paper analyzes the effect of the routing of cross traffic packets on available bandwidth measurement in detail. Then based on Monte Carlo Method, a novel methodology fundamentally different in the basic idea from the previous methods is presented to measure end-to-end available bandwidth. This method sends small single packet randomly instead of sending packet pair/train back-to-back. It could work on network whose capacity is far beyond the maximum sending rate of the sender. Analysis and simulations show that besides end-to-end available bandwidth, this method could measure the capacity and idle ratio of targeted link, and then calculate the change of traffic flow on each node and the percentage of different cross traffic on each link.

Key words: available bandwidth; bottleneck bandwidth; capacity; Monte Carlo; bandwidth measurement

摘 要: 目前绝大多数带宽测量方法都是基于网络瓶颈分隔原理的,在此基础上形成了基于包对/包队列的各种容量/可用带宽探测方法.但是,这类方法的测量结果不能超过源节点的最大发送速率,因此无法在高带宽环境中使用.另外,目前的可用带宽测量理论均没有考虑背景流的不同路由对测量方法所产生的影响.全面分析了背

* Supported by the National Natural Science Foundation of China under Grant No.60273021 (国家自然科学基金)
Received 2004-10-14; Accepted 2005-02-03

© 中国科学院软件研究所 http://www.jos.org.cn

景流的路由对可用带宽测量的影响.在此基础上,基于蒙特卡洛(Monte Carlo)随机抽样的思想,提出了一种与现有测量方法截然不同的探测理论.该方法用随机发送单个小探测报文取代了目前的探测理论所依赖的包对/包队列,其测量范围不受源节点最大发送速率的限制.分析及实验表明,该方法不仅可以计算整条路径的可用带宽,也可以计算各段链路的容量和空闲率,进而分析得到各路由节点上的流量变化,以及各链路上对应的不同类型的背景流的分布.

关键词: 可用带宽;瓶颈带宽;容量;蒙特卡洛;带宽测量

中图法分类号: TP393 文献标识码: A

在过去的近 20 年里,可用带宽(available bandwidth)作为网络路由、流量工程、QoS 控制等方面的一个关键参数,引起了学术界和工业界的广泛关注.然而到目前为止,还没有一个很好的可用带宽测量方法能够满足实际应用环境中的测量要求.目前,甚至连如何定义可用带宽都尚未达成共识.

目前,绝大多数带宽测量方法都是基于 1988 年文献[1]中提出的网络瓶颈分隔原理的,在此基础上形成了基于包对(packet pair)/包队列(packet train)的各种容量(capacity)/可用带宽探测方法.但是,这类方法测量的容量/可用带宽不能超过源节点的最大发送速率,因此无法在高带宽环境中使用.另外,目前的可用带宽测量理论均没有考虑背景流(cross traffic)的不同路由对测量方法所产生的影响.本文在全面分析了背景流的路由对可用带宽测量的影响的基础上,基于蒙特卡洛(Monte Carlo)随机抽样思想,提出了一种与以往测量方法截然不同的探测理论.该方法用随机发送单个小探测报文取代了目前的探测理论所依赖的包对/包队列,其测量范围不受源节点的最大发送速率的限制.该方法不仅可以计算整条路径的可用带宽,也可以计算各段链路的容量和空闲率,进而分析得到各路由节点上的流量变化,以及各链路上对应的不同类型的背景流的分布.

1 基本概念

网络中的传输路径由从数据源到目的地的一系列存储转发链路所组成.链路的带宽或者容量是指该链路上数据报文的最大传输速率.网络路径的瓶颈带宽(bottleneck bandwidth)或者容量是指源节点到目的节点之间处理能力最低的链路所能达到的最大的数据传输速率,也就是说,当传输路径上没有其他业务流时,该路径所能提供给一个业务流的最大传输速率.传输路径上带宽/容量最小的链路称为该路径的瓶颈链路.然而,业务流很少是单独存在于一条路径上的,而经常是和其他业务流共享网络.网络路径的可用带宽是指网络在不降低其他业务流的传输速率的情况下,所能提供给一个业务流的最大传输速率.

如果用 N 表示一条路径的跳数,用 C_i 表示链路 $i(1\le i\le N)$的容量或传输速率,那么该路径的瓶颈带宽或者容量可以表示为

$$C=\min_{i=1..N}C_i \tag{1}$$

另外,如果用 u_i 来表示链路 i 的利用率$(0\le u_i\le 1)$,那么链路 i 上的剩余带宽就可以表示为 $C_i(1-u_i)$,这样,整条路径的可用带宽就可以表示为

$$A=\min_{i=1..N}[C_i(1-u_i)] \tag{2}$$

式(2)中对于链路利用率的计算没有一个很精确的定义或表达.另外,也没有对某一时刻的可用带宽和某一时段的可用带宽加以区分.我们在文献[2]中给出了可用带宽的数学定义,本文则基于链路空闲率重新修正了可用带宽的定义.

首先,给出链路可用带宽的“可用”定义:

定义 **1**. 论文所说的某一时刻单一链路“可用”,含义就是该链路的开始节点处于空闲状态.

我们可以借助一个状态阶跃函数来描述节点在某一时刻所处的状态.

定义 **2**. 定义 t 时刻 $Node_{i-1}(1\le i\le N)$的状态函数如下:

$$Node\_free_{i-1}(t)=\begin{cases}1, & \text{当节点空闲时}\\ 0, & \text{当节点忙时}\end{cases} \tag{3}$$

© 中国科学院软件研究所 http://www.jos.org.cn

相应的 $Link_i$(表示 $Node_{i-1}$ 和 $Node_i$ 之间的链路,$1\le i\le N$)在某一时段$[t_1,t_2]$的链路空闲率可以表示为

$$Free_i(t_1,t_2)=\frac{1}{t_2-t_1}\int_{t_1}^{t_2}Node\_free_{i-1}(t)\mathrm{d}t \tag{4}$$

而 Link_i 在某一时段$[t_1,t_2]$的可用带宽就可以表示为(假定该链路的容量为 C_i)

$$Avail\_bw_i(t_1,t_2)=C_i\cdot Free_i(t_1,t_2) \tag{5}$$

显然,当 $Link_i$ 处于空闲状态时,对于一个新到来的报文,$Node_{i-1}$ 可以立即转发该报文.也就是说,当 $Link_i$ 处于空闲状态时,报文在 $Node_{i-1}$ 所经历的排队时延为 0.考虑到实际应用中,不存在"同时发生"的网络事件,因此有如下推论:对于一个新到来的报文,若该报文在 $Node_{i-1}$ 所经历的排队时延为 0,则说明在该报文被 $Node_{i-1}$ 接收的时刻,$Link_i$ 处于空闲状态.

假设一条网络路径由 N 段首尾相连的存储转发链路组成(从 $Node_0$ 到 $Node_N$),可用带宽最少的链路出现在网络链路的第 b 段,$Link_b$ 的容量为 C_b,则该网络路径在某一时段$[t_1,t_2]$的可用带宽为

$$Avail\_bw_{0\sim N}(t_1,t_2)=C_b\cdot Free_b(t_1,t_2) \tag{6}$$

2 现有技术分析

现有的可用带宽测量方法基本上都是基于文献[1]中所提到的网络瓶颈的分隔原理的.众所周知的包对法[3]就是对该原理的应用.其具体做法是,首尾相连地发送两个长度为 L 的数据包,如果该路径中瓶颈链路的处理能力为 C,在没有其他流量的情况下,这两个包到达目的地的时间间隔应为$\Delta=L/C$.那么,接收方就可以根据测得的时间间隔Δ来计算该路径中瓶颈链路的处理能力,也就是瓶颈带宽.包对法是非常经典的一种测量方法,它对后来的可用带宽测量方法产生了深远的影响.可以说,目前绝大多数可用带宽的测量方法都是对包对法的改进.但是遗憾的是,包对法本身测量的却不是可用带宽,而是瓶颈带宽.原因在于,在文献[1]中,Jacobson 没有考虑到网络中的其他流量对测试包的影响.在文献[3]中,Keshav 也在拥塞控制的背景下对此进行了研究,结论是,只有当路由器采用一种公平排队(fair queuing)策略时,包对法才能用于测量可用带宽.而在先进先出(FCFS)的排队策略下,包对法无法测出可用带宽.

文献[4]中所提到的 Cprobe 可以说是第 1 个用于测量可用带宽的工具.Cprobe 用包队列代替了包对,通过一组首尾相连的包队列中第 1 个和最后一个报文的到达时间以及报文长度来计算可用带宽.文献[5]中所提到的 pipechar 采用的也是相似的方法.然而,文献[6]中的研究表明,Cprobe 测量的并不是可用带宽,而是另外一种完全不同的参数,目前称为 ADR(asymptotic dispersion rate).

文献[7]中提出的 TOPP(trains of packet pairs)的基本思想是:以递增的速率向目的节点发送包对组成的队列,根据不同包对的输入输出速率之间的关系来判断可用带宽.文献[8]中所提到的 SLoPS(self loading periodic streams)基于的原理是:当发送速率高于可用带宽时,报文的单向延迟会有上升的趋势.基于 SLoPS 的思想实现了可用带宽测量工具 pathload.虽然 TOPP 和 pathload 可以测出相对比较准确的可用带宽,但是它们都需要发送大量的探测报文.在无线网络等带宽非常有限的网络环境中,这么多的探测报文完全可能造成网络的阻塞.另外,文献[9]中的研究表明,pathload 会高估可用带宽.

文献[10]中提出的 IGI(initial gap increasing)和 PTR(packet transmission rate)通过发送一系列报文间隔逐渐增大的包队列,并探测输出间隔和初始间隔的差别来确定可以用来计算可用带宽的包队列.然而,文献[9]中的研究表明,该算法在网络的利用率较大时对带宽变化的反应非常迟钝.文献[9]中提出的 Spruce 和 IGI 一样,都是基于探测间隔模型的,但其测量所需的探测流量要高于 IGI.

目前,国内带宽测量的研究主要针对瓶颈带宽的测量.文献[11]分析了瓶颈带宽测量中的噪声特性,提出了一种基于信号模式的滤波算法.文献[12]则提出了一种非均匀包对序列带宽测量方法来测量瓶颈带宽.

上述方法所测量的带宽都不能超过源节点的最大发送速率,因此无法在高带宽环境中使用.另外,目前的可用带宽测量理论均没有考虑背景流的不同路由对测量结果所产生的影响.

© 中国科学院软件研究所 http://www.jos.org.cn

3 端到端的可用带宽测量

与文献[1,4,5,7−10]一样,本文假定所有的路由器均采用常见的先进先出(FCFS)的排队策略,并采用存储转发方式传输报文.

3.1 单段链路可用带宽测量理论方法

从式(5)中不难看出,单段链路可用带宽测量的重点是如何确定该链路的链路空闲率.$Link_i$ 的链路空闲率可以由 $Node_{i-1}$ 的状态函数的积分值计算得到.根据 Monte Carlo 方法,函数的积分值可以由大量随机抽样结果得到.因此,本文提出了以随机发送小报文、探测链路可用状态为基础的单段链路可用带宽测量理论.链路可用状态主要借助计算随机报文在该链路上的排队时延是否为 0 来判断.具体步骤如下:

单段链路可用带宽的测量方法($Link_1$ 的容量为 C_1):

1) $Node_0$ 随机发送长度为 L 的小报文给 $Node_1$;
2) 计算每个探测报文在 $Link_1$ 上的排队时延;
3) $Free_1=X/T$(其中 X 为在 $Link_1$ 上排队时延为 0 的报文个数,T 为总的探测报文个数);
4) $Avail\text{-}bw_1=C_1\cdot Free_1$.

我们在可用带宽测量中所关注的是探测报文的排队时延.而我们在测量中所能得到的只是某段链路或者某几段链路上的传输时延,该时延包括报文的发送时延、传播时延和排队时延.如果要得到报文的排队时延,则需要事先预知网络的最小传输时延(即在排队时延为 0 的情况下探测报文传输需要的时间).我们假定在所测得的传输时延中,最小的即为“排队时延为 0”的传输时延.需要指出的是,仿真环境中由于时间计量精度高而容易得到该值;但在实际环境中,操作系统实时性能的差异会对该值造成较大的误差.经测试,当发送及接收节点的实时性较高时(例如 Linux 系统),即在保证时间计量的精度为微秒级时,传输时延的测量可满足要求.

3.2 多段链路可用带宽测量理论方法

假设一条网络路径由 N 段首尾相连的存储转发链路组成(从 $Node_0$ 到 $Node_N$),可用带宽最少的链路出现在网络链路的第 b 段.根据式(6),计算该路径的可用带宽的关键是计算 $Link_b$ 的链路空闲率.如果我们简单地套用单段链路的可用带宽测量方法,将得到如下多段链路端到端可用带宽测量方法.

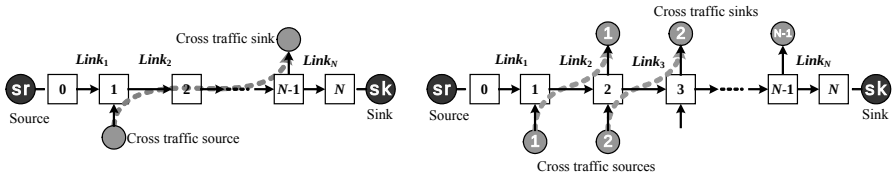
Avail-bw Method I($Link_b$ 为可用带宽最少的链路,其容量为 C_b):

1) $Node_0$ 随机发送长度为 L 的小报文给 $Node_N$;
2) 计算每个探测报文在 $Link_b$ 上的排队时延;
3) $Free_b=X/T$(其中 X 为在 $Link_b$ 上排队时延为 0 的报文个数,T 为总的探测报文个数);
4) $Avail\text{-}bw_b=C_b\cdot Free_b$.

然而在实际测量中,此方法将会产生较大的误差.Monte Carlo 方法的基本要求是测量点为随机抽样.然而经过实验我们发现,在多段链路的网络中,当测量该路径的源节点发送的探测报文在某段链路上的排队时延时,不仅该段链路上的背景流会影响测量结果,在该段之前的链路上的背景流也可能会对当前链路的排队时延产生影响,使得探测报文在流经 $Link_b$ 时,其部分探测点已不再是随机分布的.

为了说明这种情况,我们根据路由情况,将链路上的背景流分成两种基本类型.不妨设当前链路(需要测量空闲率的链路)为 $Link_x$,对应的两个网络节点分别为 $Node_{x-1}$ 和 $Node_x$,其前趋链路为 $Link_{x-1}$($x>1$ 时).若 $Link_x$ 上的背景流不存在于 $Link_{x-1}$,也即该背景流在到达 $Node_{x-1}$ 之前的途径链路并非 $Link_{x-1}$,则称该背景流是 $Link_x$ 上的 One-Hop Persistent 背景流;与之相对应,若该背景流不仅在 $Link_x$ 上出现,而且正是途径 $Link_{x-1}$ 而到达节点 $Node_{x-1}$ 的,则称该背景流是 $Link_x$ 上的 Path Persistent 背景流.图 1 给出了两种典型的背景流的网络拓扑图.图 1(a)中的背景流基本上贯穿整个测量路径,根据前面的定义,$Link_2$ 上存在 One-Hop Persistent 背景流,而从 $Link_3$ 到 $Link_{N-1}$,其背景流均为 Path Persistent 背景流.图 1(b)中各段链路上的背景流均为 One-Hop Persistent 背景流.真实环境中的背景流必然是 One-Hop Persistent 和 Path Persistent 这两种基本类型的组合.

© 中国科学院软件研究所 http://www.jos.org.cn



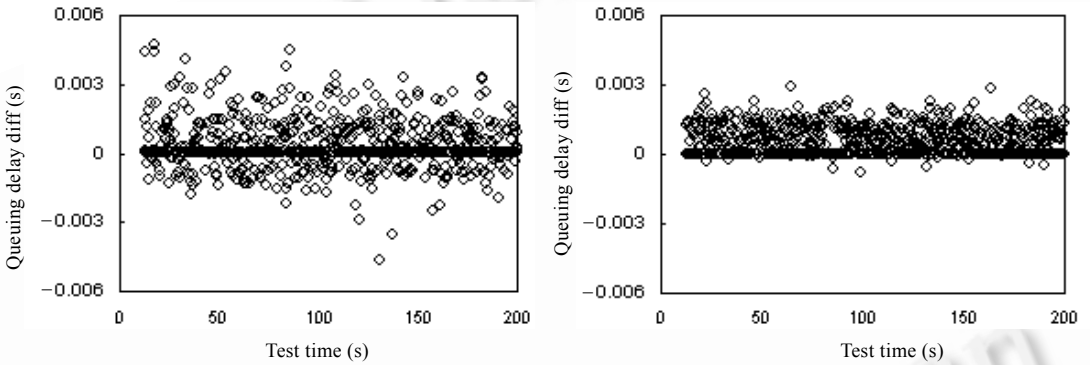
(a) One case of Path Persistent cross traffic (b) One case of One-Hop Persistent cross traffic

(a) Path Persistent 背景流的一个例子 (b) One-Hop Persistent 背景流的一个例子

Fig.1

图 1

为了简化问题,我们假定多段网络由两段链路(N=2)构成.不妨设两段链路的容量分别为 C_1=10M,C_2=8M,可用带宽最少的链路出现在 $Link_2$ 上.背景流的产生方式与文献[8]相同.$Node_0$ 随机向 $Node_2$ 发送探测报文.分别考虑 $Link_2$ 上的背景流为 One-Hop Persistent($Link_1$ 上的背景流为 5M,$Link_2$ 上的背景流为 4M)以及 Path Persistent(4M 的背景流先后流经 $Link_1$ 和 $Link_2$)时的情况,测量同一探测报文在 $Link_1$ 和 $Link_2$ 上的排队时延之差($Link_2$ 上的排队时延减去 $Link_1$ 上的排队时延)的分布,如图 2 所示.



(a) Queuing delay difference under One-Hop Persistent cross traffic

(a) One-Hop Persistent 背景流下的排队时延之差

(b) Queuing delay difference under Path Persistent cross traffic

(b) Path Persistent 背景流下的排队时延之差

Fig.2

图 2

从图 2(a)可以看到,当背景流为 One-Hop Persistent 时,前后两段相邻链路的排队时延基本上是相互独立的.这样,当一个探测报文到达当前链路时,无论其在前趋链路的排队状况如何,对当前链路而言仍属于一次合格的随机抽样探测.与之对应的是图 2(b),当背景流为 Path Persistent 时,同一报文在两段相邻链路上的排队时延之差的分布说明,当前链路的排队时延和前趋链路是相关的.事实上,因为 $C_1>C_2$,如果一个探测报文在 $Link_1$ 上经过了排队等待,其在 $Link_2$ 上必然也要进行排队.在这种情况下,该探测报文会得到必然的探测结果,而并非一次随机抽样,这违背了 Monte Carlo 方法在模拟积分结果时,测量点需随机抽样的基本要求.因此,如何在多段链路中寻找真正的随机抽样样本,成为在多段链路中测量链路利用率的关键问题.为此,将多段链路可用带宽的测量方法调整如下:

Avail-bw Method II($Link_b$ 为可用带宽最少的链路,其容量为 C_b):

1) $Node_0$ 随机发送长度为 L 的小报文给 $Node_N$;

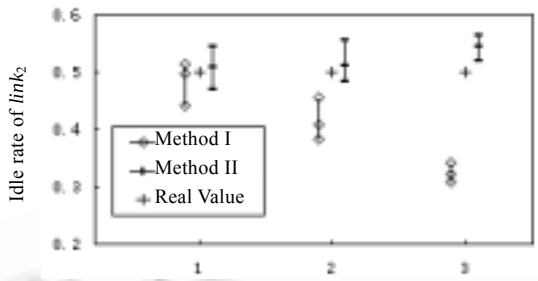
2) 计算每个探测报文在 $Link_{b-1}$ 以及在 $Link_b$ 上的排队时延;

3) $Free_b=Z/W$(其中 Z 为在 $Link_{b-1}$ 上以及 $Link_b$ 上排队时延均为 0 的报文个数,W 为在 $Link_{b-1}$ 上排队时延为 0 的报文个数);

4) $Avail\text{-}bw_b=C_b\cdot Free_b$.

© 中国科学院软件研究所 http://www.jos.org.cn

图 3 显示了在 NS2 仿真中分别采用 Avail-bw Method I 和 Avail-bw Method II 在 3 种背景流下的测量结果对比.在这 3 种背景流下可用带宽最少的链路都出现在 $Link_2$,且 $Link_2$ 的容量一样,因此我们只比较了测量得到的 $Link_2$ 的链路空闲率.仿真中两段链路的容量分别为 C_1=10M,C_2=8M,背景流的产生方式与文献[8]相同.各次测量的背景流用三元组{hop1,(path2,hop2)}(单位为 M)来表示.例如,某种背景流表示为{x,(y,z)},其含义是 $Link_1$ 有 x 流量的 One-Hop Persistent 背景流,而 $Link_2$ 上分别有 y 流量的 Path Persistent 背景流(y<=x)和 z 流量的 One-Hop Persistent 背景流.在每种背景流下,我们分别采用 Avail-bw Method I 和 Avail-bw Method II 各测量 10 次,然后计算 10 次结果的上下限和平均值(针对图中一条线的 3 个点).



No.1 {5,(0,4)}, No.2 {2,(2,2)}, No.3 {4,(4,0)}

Fig.3 Measurement result of Avail-bw Method I & Avail-bw Method II in three different cross traffic scenarios

图 3 Avail-bw Method I 和 Avail-bw Method II 在 3 种背景流下的测量结果对比

从图 3 中可以看出,随着 $Link_2$ 上的背景流中 Path Persistent 类型比例的不断提高,利用 Avail-bw Method I 方法得到的测量结果与真实值相比的误差越来越大.当 Path Persistent 类型的背景流为 0 时,Method I 和 Method II 的结果比较一致,但在第 2 组实验数据中(Path Persistent 和 One-Hop Persistent 流量各占 50%),Avail-bw Method I 得到的 10 次测量的平均值为 40%(真实值为 50%),相对误差为 20%,而 Avail-bw Method II 的测量结果为 51%,相对误差仅为 2%.在背景流全部为 Path Persistent 类型时,Avail-bw Method I 测量结果的相对误差为 34%,而 Avail-bw Method II 的相对误差仅为 8%.

以上所述的可用带宽测量方法,其核心在于计算可用带宽最小的链路的空闲率.该值的计算与探测报文本身的发送速率没有必然联系.若探测时间较长,其测量表征的是较长时间内的空闲率,而若探测时间较短,则反映了短时间内的空闲率.该方法在本质上与发送节点的最大报文发送速率无关(即源节点的接入带宽),这是对现有带宽测量方法的极大改进(基于包对法的算法为了保证精度,必须配置较高的接入带宽).而且,该方法可以检测特定链路的运行状况,而基于包对法的带宽测量只能测量整条路径的带宽变化.

另外,我们在测量中可以假定已知各段链路的容量.因为在一般情况下,各段链路的容量是常数,往往可以从网络服务供应商(ISP)那里获得,或者是利用某些路由器查询软件得到.当无背景流或者背景流在较长时间内保持稳定时,也可以凭借上述可用带宽测量方法计算得到链路容量.具体方法是:首先在源、目的节点之间启动一个业务流,测量当该业务流的发送速率分别为 a_1 和 a_2(a_1≠a_2)时,$Link_i$ 的链路空闲率 $Free'_i$ 和 $Free''_i$,由此得到 $Link_i$ 的容量为 C_i=(a_1−a_2)/($Free''_i$−$Free'_i$).与常见的带宽测量方法不同,这种基于链路空闲率的带宽测量方法不要求探测报文的发送速率大于或等于被测链路的容量.

3.3 多段链路流量分布的相关分析

上节所述的方法,不仅可以计算整条路径的可用带宽,还可以计算各段链路的容量和空闲率.而在已知各段链路的容量和空闲率的情况下,我们可以对各段链路的流量变化作简单建模,进而分析得到各路由节点上的流量变化,以及各链路上对应的 One-Hop Persistent,Path Persistent 类型的背景流的分布.注意:本文所述的背景流为除探测报文之外的所有报文,因此对背景流的分析即为对链路本身现有流量的分析.具体方法如下:不妨设待求路由节点为 $Node_X$,现在分析其后继链路 $Link_{X+1}$($Node_X$ 和 $Node_{X+1}$ 之间的链路)的流量变化.而其前趋链路为

© 中国科学院软件研究所 http://www.jos.org.cn

$Node_{X-1}$~$Node_X$ 之间的 $Link_X$.$Link_{X+1}$ 和 $Link_X$ 两段链路的容量分别为 C_{X+1} 和 C_X,利用第 3.2 节中的 Avail-bw Method II 测量得到的链路空闲率分别为 $Free_{X+1}$ 和 $Free_X$.

(1) 当 C_X<=C_{X+1} 时

$Node_X$ 上的流量变化为 C_X×(1−$Free_X$)−C_{X+1}×(1−$Free_{X+1}$),该值大于 0 说明背景流减少;小于 0 说明背景流增加,因此 $Link_{X+1}$ 上必然有 One-Hop Persistent 的背景流.

(2) 当 C_X>C_{X+1} 时

注意,此时(1)中的结论仍然适用,此外,我们还可以进一步详细分析在 $Link_{X+1}$ 上的各种背景流的分布比例.一般而言,瓶颈链路及其之前的链路往往满足 C_X>C_{X+1},因此对这种情况的研究对于分析网络瓶颈有着直接的帮助.

我们假定探测报文总共为 S 个,在 $Link_X$ 上满足排队时延为 0 的报文数为 Z_X.因此,第 3.2 节中的 Avail-bw Method II 的实质就是在这 Z_X 个有效的探测报文中,统计其在 $Link_{X+1}$ 上排队时延也为 0 的报文个数,然后计算得到 $Free_{X+1}$.

假定 $Link_{X+1}$ 上的 Path Persistent 背景流为 M(说明这些报文均也存在于 $Link_X$),那么 $Link_X$ 上的其他背景流 O 就满足 O=C_X(1−$Free_X$)−M,并且这些报文将会从 $Node_X$ 转发到非测量路径上去.为了方便说明,我们设 $\sigma=\frac{M}{C_X(1-Free_X)}$,即 M 在 $Link_X$ 的背景流中所占的比例,那么 O=C_X(1−$Free_X$)(1−σ).

在 $Link_X$ 上,S 个探测报文中只有 Z_X 个不受背景流的影响,其余 S–Z_X 个探测报文因为背景流的存在导致其排队时延大于 0.一般来说,可以假定这些在 $Link_X$ 上受背景流阻挡的探测报文,其被 M 和被 O 阻挡的概率一样,因此有 $(S-Z_X)\cdot\sigma$ 的探测报文受 M 的阻挡,而有 $(S-Z_X)(1-\sigma)$ 的探测报文是受 O 的阻挡.

如第 3.2 节所述,利用 Avail-bw Method I 对 $Link_{X+1}$ 进行测量之所以不准确,就是因为 $Link_X$ 上 M 的存在.因此,若能去掉受 M 阻挡的探测报文,就能修正 Avail-bw Method I 的测量结果.显然,对于探测 $Link_{X+1}$ 而言,此时有效的探测报文总数是 $S-(S-Z_X)\cdot\sigma$,而满足在 $Link_{X+1}$ 上排队时间为 0 的报文个数为 Z_{X+1}(因为 C_X>C_{X+1},在 $Link_X$ 上被 M 阻挡的探测报文在 $Link_{X+1}$ 上肯定会被同一报文再度阻挡,其排队时延必然大于 0,因此必不属于 Z_{X+1}).所以,根据第 3.2 节的分析结果,可以得到:

$$\frac{Z_{X+1}}{S-(S-Z_X)\cdot\sigma}=Free_{X+1},$$

计算可以得到 $\sigma=\frac{1}{S-Z_X}\left(S-\frac{Z_{X+1}}{Free_{X+1}}\right)$,然后可以分别计算得到:

$$M=\frac{C_X(1-Free_X)}{S-Z_X}\left(S-\frac{Z_{X+1}}{Free_{X+1}}\right),\quad O=\frac{C_X(1-Free_X)}{S-Z_X}\left(\frac{Z_{X+1}}{Free_{X+1}}-Z_X\right).$$

对于 $Node_X$ 而言,其接收到 $Node_{X-1}$ 转发的 C_x(1−$Free_x$)的背景流,其中 M 的背景流继续向 $Node_{X+1}$ 转发,而 O 的背景流则向非测量路径转发.同时,有 C_{X+1}(1−$Free_{X+1}$)−M 的新背景流进入测量路径,由 $Node_X$ 向 $Node_{X+1}$ 转发.对于 $Link_{X+1}$ 而言,其背景流为 C_{X+1}(1−$Free_{X+1}$),其中 M 为 Path Persistent 背景流,而 C_{X+1}(1−$Free_{X+1}$)−M 为 One-Hop Persistent 背景流.

4 仿真测试

4.1 可用带宽测量方法的测试

我们基于 Avail-bw Method II 实现了可用带宽测量工具 Idlerate,并在 NS2 中进行了仿真测试.本文选取文献[8]中的可用带宽测量工具 pathload 作为参考对象.仿真模型和文献[8]完全相同:测试路径为 5 跳,瓶颈链路(同时也是可用带宽最少的链路)出现在第 3 跳,其利用率为 u_t.每一跳的背景流由 10 个 Pareto 源产生,α=1.5.

Idlerate 和 pathload 分别在 4 种瓶颈链路负载(u_t=30%,60%,75%,90%)下各测试 50 组,其结果对比如图 4 所

© 中国科学院软件研究所 http://www.jos.org.cn

示.Idlerate 测量的结果是一个值,因此图 4 中给出的是 50 次测量值的分布区间.而 pathload 的测量仅能返回一个区间,因此图 4 中给出的是 50 次测量区间的平均值(为 50 次测量区间上限的平均值和 50 次测量区间下限的平均值所构成).分别测试了 Idlerate 在探测报文数为 100 和 200 时的测量情况.

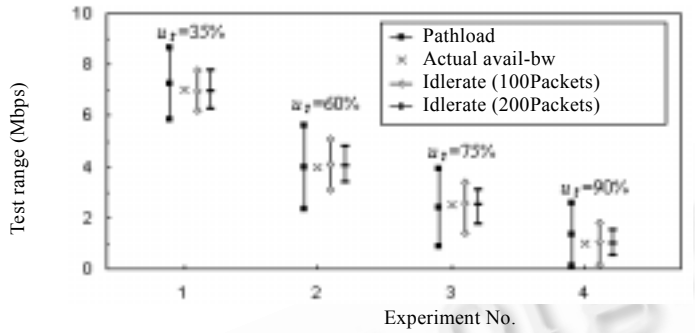


Fig.4 Contrast between Idlerate and pathload

图 4 Idlerate 与 pathload 的对比

如图 4 所示,Idlerate 的测量结果比 pathload 要准确:Idlerate 的 50 次测量值所得到的分布区间甚至小于 pathload 的 50 次测量区间的平均值.另外,在测量中,当探测报文数为 200 时,Idlerate 的探测流量仅为 11.2kbyte(每个报文为 56byte).而 pathload 中每个流(stream)的默认大小为 80kbyte,*pathload* 的每次测量(fleet)通常需要 12 个流.

4.2 多段链路流量分析模型的测试

我们针对图 5 所示的典型的网络拓扑环境测试了第 3.3 节中的多段链路流量分析模型.

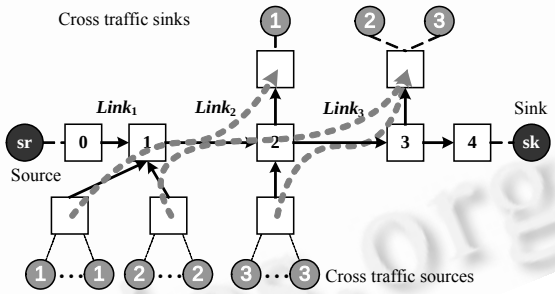


Fig.5 Test topology used in the NS simulations

图 5 NS 仿真中所采用的拓扑环境

图 5 中各链路的容量分别为 C_1=100M,C_2=10M,C_3=8M.背景流的产生方式与文献[8]相同.背景流 1、背景流 3 各自流经 $Link_2$ 和 $Link_3$,其流量均为 2M;背景流 2 贯穿 $Link_2$ 和 $Link_3$,流量也为 2M.实际探测流的流量仅为 1.3kbps,对已有网络背景流的影响可以忽略不计.考虑到本例中可用带宽最小的链路为 $Link_3$,选择的测试目标是分析 $Node_2$ 和 $Link_3$ 上的背景流分布.测试每次 10 分钟,重复测试 5 次,结果见表 1.

Table 1 Test result of cross traffic on $Node_2$ and $Link_3$

表 1 $Node_2$ 和 $Link_3$ 上背景流分布的测试结果

	Actual value	Measured value	
		Average	Stdev.
σ on $Link_2$	0.5	0.49	0.0319
M: Path persistent cross traffic on $Link_3$	2M	1.75M	0.126M
O on $Link_2$; exit measured path from $Node_2$	2M	1.83M	0.127M
One-Hop Persistent cross traffic on $Link_3$	2M	1.63M	0.236M

© 中国科学院软件研究所 http://www.jos.org.cn

5 结束语

本文分析了背景流的路由对可用带宽测量的影响.在此基础上,基于蒙特卡洛随机抽样的思想提出了一种与以往测量方法截然不同的可用带宽探测理论.该方法用随机发送单个小探测报文取代了目前的探测理论所依赖的包对/包队列,其测量范围不受源节点最大发送速率的限制.该方法不仅可以计算整条路径的可用带宽,也可以计算各段链路的容量和空闲率,进而分析得到各路由节点上的流量变化,以及各链路上对应的不同类型的背景流的分布.

致谢 作者感谢 M. Jain 和 C. Dovrolis 向我们提供了文献[8]中的详细仿真环境以及原始测量数据.

References:

[1] Jacobson V. Congestion avoidance and control. ACM SIGCOMM Computer Communication Review, 1988,18(4):314−329.
[2] Liu M, Shi JL, Li ZC, Kan ZG, Ma J. A new end-to-end measurement method for estimating available bandwidth. In: Bilof R, ed. Proc. of the 8th IEEE Symp. on Computers and Communications. San Francisco: IEEE Press, 2003. 1393−1400.
[3] Keshav S. A control-theoretic approach to flow control. ACM SIGCOMM Computer Communication Review, 1991,21(4):3−15.
[4] Carter RL, Crovella ME. Measuring bottleneck link speed in packet-switched networks. Perform. Eval., 1996,27(28):297−318.
[5] Jin GJ, Yang G, Crowley BR, Agarwal DA. Network characterization service (NCS). In: Williams AD, ed. Proc. of the 10th IEEE Symp. on High Performance Distributed Computing. San Francisco: IEEE Press, 2001. 289−299.
[6] Dovrolis C, Ramanathan P, Moore D. What do packet dispersion techniques measure? In: Proc. of the IEEE INFOCOM. Anchorage: IEEE Press, 2001. 905−914.
[7] Melander B, Björkman M, Gunningberg P. A new end-to-end probing and analysis method for estimating bandwidth bottlenecks. In: Proc. of the Global Internet Symp. San Francisco: IEEE Press, 2000. 415−420.
[8] Jain M, Dovrolis C. End-to-End available bandwidth: Measurement methodology, dynamics, and relation with TCP throughput. IEEE/ACM Trans. on Networking (TON), 2003,11(4):537−549.
[9] Strauss J, Katabi D, Kaashoek F. A measurement study of available bandwidth estimation tools. In: Proc. of the 3rd ACM SIGCOMM Conf. On Internet Measurement. New York: ACM Press, 2003. 39−44.
[10] Hu NN, Steenkiste P. Evaluation and characterization of available bandwidth probing techniques. IEEE Journal on Selected Areas in Communications, 2003,21(6):879−894.
[11] Lin Y, Wu HT, Cheng SD, Wang CG, Jin YH, Wang WD. A noise analysis for measuring network capacity bottleneck. Chinese Journal of Electronics, 2004,32(4):552−556 (in Chinese with English abstract).
[12] Zhang WJ, Qian DP, Wu WG, Luan ZZ, Xu DW. Non-Uniform packet pair sequence bandwidth measurement method. Journal of Xi'an Jiaotong University, 2002,36(10):1045−1048 (in Chinese with English abstract).

附中文参考文献:

[11] 林宇,邬海涛,程时端,王重钢,金跃辉,王文东.网络瓶颈带宽测量的噪声分析.电子学报,2004,32(4):552−556.
[12] 张文杰,钱德沛,伍卫国,栾钟治,许大炜.一种非均匀包对序列带宽测量方法.西安交通大学学报,2002,36(10):1045−1048.


刘敏(1976－),女,河南郑州人,博士生,副研究员,主要研究领域为网络测量和移动切换.


过晓冰(1977－),男,高级研究员,主要研究领域为计算机网络,系统管理.


李忠诚(1962－),男,博士,研究员,博士生导师,CCF高级会员,主要研究领域为计算机网络.


邓辉(1972－),男,博士,研究员,主要研究领域为3G,NGN,QoS,IMS.

© 中国科学院软件研究所 http://www.jos.org.cn