IPv6网络路径容量测量方法研究

唐 军 裴昌幸 苏 博

(西安电子科技大学综合业务网理论与关键技术国家重点实验室 西安 710071)

摘 要 分组离散技术是常用的路径容量测量方法之一,它通常采用连续发送策略来发送测量分组。然而该策略对发送带宽的利用率较低,影响了测量的准确性。针对IPv6网络环境,提出了一种新的基于分片机制的路径容量测量方法。该方法通过构造长度大于路径最大传输单元的原始测量分组,迫使协议栈对该分组进行分片操作,生成长度相等的测量分片序列。实验结果表明,与经典的pathrate工具相比,该方法能够获得更小的源端分组离散以及更好的背景流量抑制能力。

关键词 路径容量,分片机制,分组离散,背景流量,网络测量

中图法分类号 TP393 **文献标识码** A

Research on Path Capacity Estimation for IPv6 Networks

TANG Jun PEI Chang-xing SU Bo

(State Key Lab. of Integrated Services Networks, Xidian University, Xi'an 710071, China)

Abstract Packet-dispersion technique is one of the useful path capacity estimation methodology and usually employs continuous sending policy to send probing packets. But the accuracy of measurement is affected by the lower utilization of transmission rate at the sender. Based on fragmentation mechanism, a novel scheme of path capacity estimation for IPv6 networks was proposed. An original probing packet with its size larger than Path Maximum Transfer Unit is generated at the sender, and then be split into a serial of equal-sized fragments by the protocol stack. It is shown by extensive experimental results that the proposed scheme can lead to lower dispersion of the probing packets at the sender and a greater capability against the cross traffic, compared with the classical tool called pathrate.

Keywords Path capacity, Fragmentation, Packet dispersion, Cross traffic, Network measurements

1 引言

分组离散技术(packet-dispersion technique)被认为是目前最经典的网络带宽测量方法之一,它利用测量分组到达接收端时的时间间隔来估算网络路径带宽。根据测量指标的不同又可分为分组对(packet pair)技术和分组链(packet train)技术,前者主要用来估算路径容量(path capacity),即网络路径上最窄链路(narrow link)的物理带宽,其衍生算法包括Pathrate、bprobe、RateTracer等[1-8];后者则常被用来估算可用带宽(available bandwidth),即网络路径当前能够提供的最大传输速率,其衍生算法有cprobe等[9-12]。为了克服背景流量的干扰,分组离散技术通常采用连续发送策略发送测量分组,然而在实际的网络环境中,受源端操作系统及协议开销的影响,该策略并不能保证分组的背靠背(back to back)特性,在网络负载较重时会带来较大的测量误差。目前大多数文献主要侧重于研究如何改进测量分组的长度、数量、排列方式以及带宽统计算法来提高测量准确度,尚未有相关文献对测量分组的发送策略进行研究。

针对上述问题,本文提出了一种应用于IPv6网络、基于分片机制的网络路径容量测量方法。该方法利用IPv6源端分片特性产生测量分组序列,减小了源端协议开销对测量的影响,提高了分组离散技术测量的准确性。本文第2节分析了源端分组离散对带宽估计的影响,第3节介绍了分片发送策略的基本思想及性能分析,第4节是实验验证。

2 源端分组离散对路径容量的影响

考虑一条由 H 条存储转发(store-and-forward)链路组成的网络路径 P,假设第 i 条链路的容量为 C_i,可用带宽为 A_i,链路利用率为 u_i。所有网络设备按照先来先服务(first come first served)的原则处理分组。那么依照定义可以得到 P 的路径容量 C 与可用带宽 A 分别为:

$$C=\min_{i=0,1,\cdots,H} C_i \tag{1}$$

$$A=\min_{i=0,1,\cdots,H} A_i=\min_{i=0,1,\cdots,H} C_i(1-u_i) \tag{2}$$

式中,C_0 为源端的最大分组传输速率。由于在实践中往往很难得到待测路径上所有链路的容量,因此分组离散技术通过考察测量过程中分组离散(即相邻两个测量分组最后一个比

本文受国家自然科学基金资助项目(61072067),国家重点实验室专项基金(ISN1001004),高等学校学科创新引智计划资助(B08038),陕西省工业攻关计划(2009K01-46)资助。

唐 军(1981—),男,博士生,主要研究方向为下一代互联网关键技术、移动IPv6,E-mail:canarmy@qq.com;裴昌幸(1946—),男,博士,教授,博士生导师,主要研究方向为量子通信、网络测量;苏 博(1982—),男,博士生,主要研究方向为无线传感器网络。

特之间的时间间隔)的变化来估算带宽。以分组对技术为例,假设测量分组的大小为 L,其在链路 i 上所产生的传输时延为 $\tau_i=L/C$。若它们所经历的处理时延也相同,在不考虑背景流量的情况下,分组散 d_i 应为:

$$d_i=\max\{d_{i-1},\tau_i\}=\max\{d_{i-1},L/C_i\} \tag{3}$$

如果在整条路径上应用式(3)可知:分组离散从源端开始不断变大,在最窄链路处达到最大值并一直保持到接收端。令最窄链路的容量为 C_{nl},分组传输时延为 τ_{nl},那么分组在接收端的离散 d_H 应为:

$$d_H=\tau_{nl}=L/C_{nl}=L/C \tag{4}$$

通过观察分组对在接收端形成的离散值 d_H 就可以利用式(4)估算出路径容量。

然而在实际的网络中,背景流量的存在会给测量分组带来额外的排队时延,因而产生较大干扰。假设背景流量服从泊松分布,其在第 $i(1\leqslant i\leqslant H)$ 为 λ_i,那么在分组对在第 i 跳上实际的离散 D_i 应为:

$$D_i=\begin{cases}D_{i-1}, & L/D_{i-1}\leqslant A_i\\ (L+\lambda_i D_{i-1})/C_i=L/C_i+u_i D_{i-1}, & L/D_{i-1}>A_i\end{cases} \tag{5}$$

从式(5)可以看出,除非当前链路具有足够的可用带宽,否则背景流量会增加分组离散 D_i,且增量与上一跳的分组离散 D_{i-1} 相关。将上述结论应用在整条路径上可知:源端分组离散(initial packet dispersion)D_0 对测量起着至关重要的作用。若 D_0 较大,则算法受背景流量的影响也较大,从而导致对路径容量的低估。文献[13]的研究表明:在几乎所有的网络中,操作系统和协议的开销占据了实际数据传输时间的大部分。这个开销将会严重影响分组对的背靠背特性,从而产生分组间间隙(inter-packet gap,IPG),假设在源端 IPG 为 G_0,那么 D_0 应为:

$$D_0=L/C_0+G_0 \tag{6}$$

从式(6)可以看出,若要减小源端分组离散,可以使用分组发送速率较快的设备或减少源端协议开销,而后者即是本文研究的重点。

3 基于分片机制的路径容量测量方法

3.1 基本思想

为了减小源端协议开销、提高带宽估算的准确度,本文提出了一种适用于分组离散技术的分组发送策略。与传统的连续发送策略不同,该方法利用 IPv6 源端分片机制产生测量分组序列。分片机制是为了解决如何在网络上传输长度大于路径最大传输单元的分组而设计的,在 IPv6 中只有源端才能进行分片操作,中间路由器将片分组看作是普通的 IPv6 分组,不会对其进行重组,因此分片策略能够与分组离散技术相结合。

本文方法主要由以下几个步骤组成:

1)根据具体的网络环境确定原始测量分组的大小。若期望的测量分组长度为 L_{exp},单次测量所需分组数量为 N_{exp},分组中首部长度为 L_h,片首部长度为 L_{fh},不可分片扩展首部的长度为 L_{ufeh},那么原始测量分组的大小 L_{origin} 应为:

$$L_{origin}=\lceil L_{exp}/8\rceil\cdot 8N_{exp}-(N_{exp}-1)(L_h+L_{fh}+L_{ufeh}) \tag{7}$$

式中,L_h 和 L_{fh} 均为常数,其大小分别为 40Bytes 和 8Bytes,而 L_{exp} 一般受以太网最大及最小帧长的限制,其大小范围为[46,1500]Bytes,L_{origin} 受 IPv6 分组最大长度的限制,其大小不超过 216Bytes(不包含 IPv6 首部)。由此可以得到分片发送策略单次所能产生的最大测量分组数量为:

$$N_{max}=\left\lceil\frac{65536-L_{ufeh}}{L_{exp}-(L_h+L_{fh}+L_{ufeh})}\right\rceil \tag{8}$$

举例来说,若所需的测量分组长度为 1500Bytes 且 $L_{ufeh}=0$,那么分片发送策略最多可以产生 45 个测量分组。相同条件下,若分组长度为 100Bytes,则最多可产生 1260 个测量分组。对于大多数应用场景,分片发送策略所提供的单次测量分组数量能够满足需求。

2)将源端输出接口的链路最大传输单元的大小修改为 L_{exp}。目前绝大多数的操作系统均可以自由配置该值。

3)发送原始测量分组至接收端。

4)接收端对测量样本进行筛选过滤。若采用 UDP 作为测量分组的传输层协议,由于其不保证可靠交付且网络设备对其是单独路由的,因此在接收端就有可能出现分组丢失或分组乱序现象,这将对测量产生很大的影响。在分片发送策略中,每个测量片分组的片首部中都包含片偏移量(Fragment Offset)字段,该字段记录着紧跟在片首部之后的数据相对于原始分组可分片部分起始处的偏移量。在正常情况下,该值随着片分组到达接收端的先后顺序递增。在接收的片分组中检查该值的连续性可以迅速判断片分组是否乱序或丢失。

5)在接收端测得测量片分组的分组离散并根据式(4)估算路径容量。

3.2 性能分析

以开源的 FreeBSD 4.8-RELEASE 操作系统[14]为例,比较连续发送策略与分片发送策略在源端的协议开销。图 1 显示了该操作系统中 IPv6 分组的发送流程。为了便于分析,将忽略中断服务以及移动分组到存储器所需的时间,而只关注 CPU 的协议处理时间。

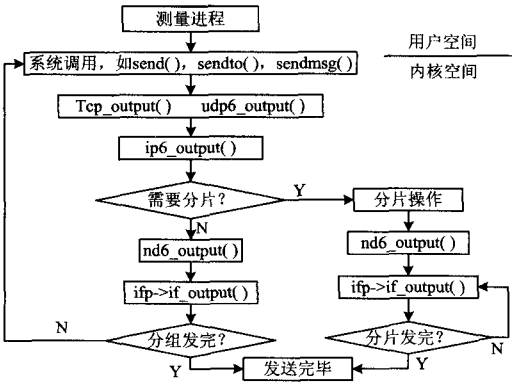


图 1 IPv6 分组发送流程(FreeBSD 4.8-RELEASE)

若采用连续发送策略,当一个测量分组发送完毕之后,测量进程会立刻从内核空间返回用户空间,然后重新调用 sendto()、udp6_output()、ip6_output()等函数来发送后一个测量分组,其中包含大量的诸如参数验证、首部填充、路由及输出接口确定等操作。文献[15]的研究表明,随着协议的不同,这些操作所需的指令条数约为 5000~10000 条。此外操作系统在用户模式与内核模式之间的环境切换(Context Switch)也会带来额外的开销。假设在传输层采用 UDP 协议,函数 f 处理所消耗的时间为 $T_{f()}$,环境切换开销为 T_{cs},则连续发送策

略所产生的 IPG 可表示为：

$$G_c \approx T_{sendto()} + T_{udp6\_output()} + T_{ip6\_output()} + T_{nd6\_output()} + T_{ifp->if\_output()} + T_{cs} \quad (9)$$

若采用分片发送策略，那么协议栈首先对原始测量分组进行分片操作，当所有片分组都生成完毕后，会调用一个循环将所有片分组通过 nd6_output()函数依次发送出去，在此期间无需进行环境切换，也无需重复调用 sendto()、udp6_output()、ip6_output()等函数。此时 IPG 可表示为：

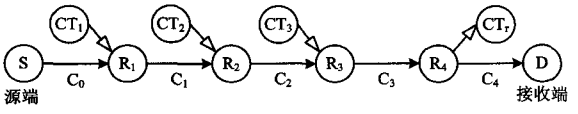
$$G_f \approx T_{nd6\_output()} + T_{ifp->if\_output()} \quad (10)$$

对比式(9)与式(10)可以得到：$G_f < G_c$，这是由于分片发送策略减少了函数的调用次数，规避了低效的环境切换。结合第 1 节的分析可以看出：在相同条件下，分片发送策略能够带来比连续发送策略更小的源端分组离散，从而使分组离散技术受到的背景流量干扰更小。

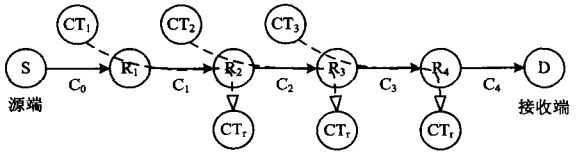
分片发送策略会给测量带来额外的开销，其主要产生于图 1 中的分片操作阶段。与连续发送策略相比，该操作需要更多的系统资源，并且会产生几百毫秒的延时。然而在第一个片分组发出之前分片操作已经结束，因此分片开销对分组离散的大小不会产生任何影响。

4 实验及分析

在实验室搭建的网络平台上对经典的 pathrate 工具及本文方法进行了对比测试，主要考察源端分组离散和路径容量。实验网络拓扑如图 2 所示。其中图 2(a)中背景流量为 p-持续背景流，图 2(b)中背景流量为 1-持续背景流。$R_1 - R_4$ 为中间路由器，C_0 为发送设备的最大传输速率，$C_1 - C_4$ 为各条链路的容量，其大小分别为 100、80、50、70 Mbps。实验采用文献[16]提出的基于伪 Pareto 分布的自相似流量生成算法来模拟背景流量，$CT_1 - CT_3$ 为 3 个干扰源集合，每个集合均包含 n 个服从 Pareto 分布的 ON/OFF 源，每个源发出的数据分组大小均为 1500Bytes，并且各个源的 ON 周期重尾特性均相同，OFF 周期重尾特性也相同，其 Pareto 分布的参数分别为 α_{ON} 和 α_{OFF}。CT_r 为背景流量接收端，所有网络设备均由千兆光纤连接。对于本文提出的策略，源端 S 根据测量工具的要求生成分片序列发往接收端 D，后者记录下分组离散并估算出路径容量值。测量间隔为 100ms，时间戳分辨率为 1μs。



(a) p-持续背景流



(b) 1-持续背景流

图 2 实验网络拓扑图

4.1 源端分组离散

图 3 显示了测量分组(分片)长度 L 分别为 1500 和 100Bytes、发送速率 C_0 分别为 100 和 1000Mbps 时，对连续发送策略与分片发送策略分别进行 50 次测量所产生的源端分组离散分布情况。图中横坐标表示测量序号，纵坐标表示源端分组离散。可以看出利用分片发送策略可以获得比连续发送策略更小的源端分组离散，该测量结果印证了第 2 节的分析结论。从图 3(d)中还可以看出，尽管较小的测量分组能够获得较小的源端分组离散，但是其发送带宽利用率却非常低，这主要是因为：(1)受时间戳分辨率的影响，过小的分组离散比较难测量，因此会产生较大误差[17]。(2)与大分组相比，小分组的每比特开销(per-bit overhead)相对较大[15]。

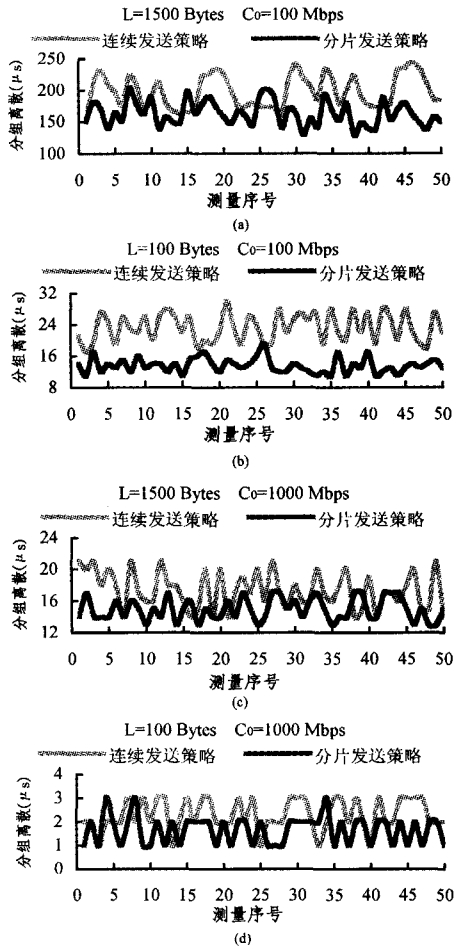


图 3 源端分组离散分布图

4.2 路径容量

设置干扰源的参数 $\alpha_{ON}=1.1$，$\alpha_{OFF}=1.9$，以此模拟轻负载网络场景。测量分组(分片)数量为 2，且大小服从[500，1500]Bytes 上的均匀分布。对两种方法分别进行 1000 次测量后所得的路径容量多峰分布如图 4(a)、(b)所示。设置干扰源的参数 $\alpha_{ON}=1.9$，$\alpha_{OFF}=1.1$，以此模拟重负载网络场景，其余参数不变，所得到的路径容量多峰分布如图 4(c)、(d)所示。可以看出在网络负载较轻的情况下，尽管本文方法的亚容量分散范围(sub-capacity dispersion range，SCDR)的强度较弱，但是采用方法、容量众数(capacity mode，CM)均为分布中的全局众数，也就是说两种策略均能估算出准确的路径容量值。在网络负载较重的情况下，受背景流量增加的影响，两种方法所产生的 SCDR 均大幅增强，但是本文方法依然能够获得比 pathrate 更弱的 SCDR，并且其 CM 依然能够保持为全局众数。无论哪种情况，窄链路后容量众数(post-narrow capacity modes，PNCMs)都较弱，不会对 CM 的识别产生影响。

(下转第 349 页)

marks for localization in wireless senor networks[J]. Computer Communications,2007(30):2577-2592

[2] Huang Gui,Zaruba G V. Static path planning for moblie beacons to localize sensor networks[C]//Proceedings of the Fifth Annual IEEE International Conference. 2007:323-330

[3] 黎作鹏.基于移动锚节点的无线传感器网络定位技术研究[D].哈尔滨:哈尔滨工程大学,2010

[4] Wang B,Yan B Y,Yuan D H. The basic Study of the Features of the Adhoc Nodes Mobility Mode[J]. Journal of Sichuan University,2006,42(1):68-72

[5] 梁甲金.基于移动锚节点的无线传感器网络定位技术研究[D].成都:西南交通大学,2010

[6] Jian L, Pmohapatra L. Location Aided Knowledge Extraction Routing for Mobile Adhoc Networks[J]. wireless Communications and Networking,2003,5(2):1180-1184

[7] 王继春.无线传感器网络节点定位若干问题研究[D].合肥:中国科学技术大学,2009

[8] Lee J,Chuang W,Kim E. A new range-free localization method using quadratic programing [J]. Computer Communications, 2011(34):998-1010

(上接第 314 页)

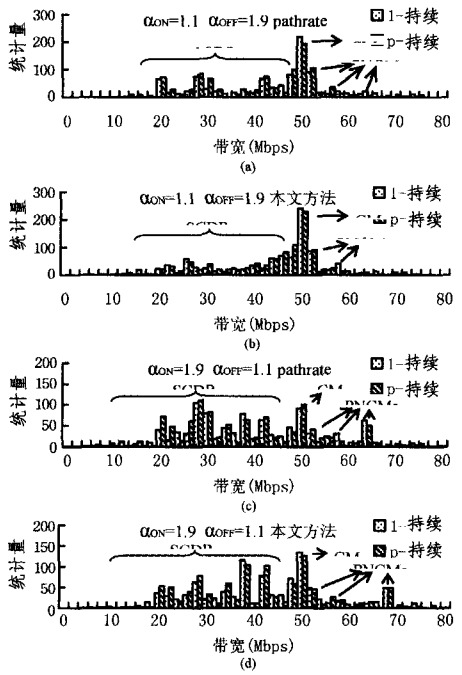


图 4 路径容量多峰分布图

需要指出的是:对于 IPv4 网络来说,由于中间路由器需要对分片进行重组,这样会扰乱测量分组的离散,因此本文方法并不适用于 IPv4 网络环境。

结束语 本文重点研究了源端分组离散对路径容量的影响,分析了 IPv6 协议栈中的分组发送机制,提出了一种新的 IPv6 网络路径容量测量方法。理论分析及实验表明:该方法克服了目前常见方法中分组间间隙较大、发送带宽利用率低的缺点,能够有效地降低背景流量对测量的影响,特别适合高速 IPv6 网络环境。在今后的工作中将考虑如何减小分片开销,以更好地满足实时测量的需求。

参 考 文 献

[1] Dovrolis C,Ramanathan P,Moore D. What do packet dispersion techniques measure [C]//Proceedings IEEE INFOCOM. 2001: 905- 914

[2] Carter R L,Crovella M E,Measuring bottleneck link speed in packet-switched networks[J]. Performance Evaluation, 1996, 27&28:297-318

[3] Shigeo S,Takahiro Y,Kenichi M. A New Approach to the Bottleneck Bandwidth Measurement for an End-to-End Network Path[C]//IEEE International Conference on Communications. 2005:59-64

[4] Lai K,Baker M. Nettimer:a tool for measuring bottleneck link bandwidth[C]//Proceedings of 3rd USENIX Symposium on Internet Technologies and Systems. 2001:122-133

[5] 张文杰,钱德沛,伍卫国,等.一种非均匀包对序列带宽测量方法[J].西安交通大学学报,2002,36(10):1045-1048

[6] 李智涛,徐雅静,刘利宏,等.一种新的 IPv6 网络带宽测量方法[J].电子与信息学报,2008,30(9):2283-2286

[7] Li Xing-feng, Luo Wan-ming, Yan Bao-ping. Study and Implementation of Bottleneck Bandwidth Measurement in IPv6 Networks[C]//ICCT International Conference on Communication Technology. 2006:1-4

[8] 李雯,潘乔,朱畅华,等.一种适用于 IPv6 的高效瓶颈带宽测量方法[J].计算机工程,2007,33(22):142-144

[9] Crocker M,Lazarou G,Baca J,et al. A Bandwidth Determination Method for IPv6-based Network[J]. The International Journal of Computers & Applications,2009,31(2):109-118

[10] Cabellos-Aparicio A,Garcia F J,Domingo-Pascual J. A novel available bandwidth estimation and tracking algorithm [C]// IEEE International Workshop on Bandwidth on Demand. 2008: 87-94

[11] Xu Da-wei,Qian De-wei. A bandwidth adaptive method for estimating end-to-end available bandwidth[C]//IEEE Singapore International Conference on Communication Systems. 2008: 543-548

[12] 邱全杰,吴中福.一种 IPv6 网络可用带宽测量方法及分析[J].计算机科学,2011,38(4):84-86

[13] Tanenbaum A S. Computer Networks[M]. Prentice Hall,2003: 476-484

[14] Li Q,Jinmei T,Shima K. IPv6 Core Protocols Implementation [M]. Morgan Kaufmann Publishers,2007:131-286

[15] Comer D E. Network Systems Design Using Network Processors [M]. Prentice Hall,2004:97-102

[16] Kramer G. On Generating Self-similar Traffic Using Pseudo- pareto Distribution[R]. Department of Computer Science,University of California,Davis,2000

[17] Dovrolis C, Ramanathan P, Moore D. Packet-dispersion techniques and a capacity-estimation methodology [J]. IEEE/ACM Transactions on Networking,2004,12(6):963-977