基于 LRU 的大流检测算法

王洪波　裴育杰　林 宇　程时端　金跃辉
(北京邮电大学网络与交换技术国家重点实验室　北京　100876)

摘　要：高速网络中，检测大流是进行准确流量测量的一种重要可扩展解决方案。该文提出了一种新的基于 LRU 大流检测算法。它通过引入“小流早期丢弃”和“大流预保护”机制以提高测量准确性。算法分析表明：新算法具有 10Gbps 线速处理能力。该文基于实际互联网数据进行了实验比较，结果显示：与已有算法相比，新算法具有更高的测量准确性和实用性。

关键词：流量测量；大流；重尾分布；最近最久未使用
中图分类号：TP393.06　　**文献标识码**：A　　**文章编号**：1009-5896(2007)10-2487-06

A LRU Based Algorithm for Identifying and Measuring Large Flows

Wang Hong-bo　Pei Yu-jie　Lin Yu　Cheng Shi-duan　Jin Yue-hui
(*State Key Laboratory of Networking and Switching*, *Beijing University of Posts & Telecommunications*, *Beijing* 100876, *China*)

Abstract: Identifying and measuring large flows is an important scalable solution for traffic measuring accurately on high-speed networks. A new algorithm based on LRU replacement scheme is proposed, which uses the policies of “early dropping small flows” and “preparatively protecting large flows” to enhance the accuracy of traffic measurement. An analysis demonstrates that the new algorithm can support the 10Gbps line-speed processing. Experiments are also conducted based on real network traces. Results show that the proposed method is more accurate and practicable than existing algorithms.

Key words: Traffic measurement; Large flows; Heavy tailed distribution; Least Recently Used (LRU)

1　引言

互联网中，流量测量(traffic measurement)是网络监测、控制和管理的基础。测得的流量信息可用于网络计费(network accounting)、流量工程(traffic engineering)、拒绝服务攻击(DoS)检测等应用。当前，流量测量往往以流(flow)为单位来进行。IETF推荐的流量测量方法[1]是在路由器中维护一个流缓存区，在其中为每个流保存一个流记录，流记录包含流ID(流定义的5元组或目的IP等)及其流量信息(如字节数)。但是，随着网络带宽及网络流量的快速增长，网络中流数量每小时已经超过百万[2]；而目前半导体工业不能提供维护每流状态所需的大量高速存储器[3]，因此维护每流状态方法不能应用于高速网络。引入分组抽样是一种可扩展解决方案，但是抽样的代价是降低测量准确度，不适合用于对单个流的测量准确度要求高的应用。

已有研究[2,4]发现：网络中流的总数虽大，但是按字节总数计算的流大小符合重尾分布(heavy tailed distribution)；即少数大流占据大部分的流量，而大部分小流只占据小部分的流量。例如，文献[2]中结果显示：9%的自治域(AS)间的字节流量占所有自治域间流量的90%。事实上，对许多应用来说只获取大流的流量信息就能满足需求。例如，在计费应用中，基于使用量计费和基于时间计费是常用的两种计费方式；基于时间计费虽然简单但是它并不能反映用户对网络资源的使用情况，往往会导致用户对网络资源的过度使用；基于使用量计费的方法虽有利于提高网络的使用效率，但是由于要记录所有流的信息从而缺乏可扩展性。因此人们[5]提出了可扩展的门限计费方式：对超过一定量的大流按使用量计费，其它流则按时间计费，这种方法既有利于提高网络资源的使用效率又具有可扩展性。又如，在流量监控应用中，ISP需要发现骨干网中流量拥塞链路，通过对大流的重路由来降低或消除拥塞[6]。这些应用的关注对象都是大流。因此，检测大流作为一种可扩展解决方案已经成为高速网络中流量测量的一个重要方面，是近年来的研究热点。

Estan及Varghese[3]首先把大流检测问题引入网络测量领域，并给出了两个算法：“sample and hold”及“Multistage filters”。前者算法简单、易于实现，但是其误差偏高。后者对大流的测量有较高的准确性，但是它会把小流误检为大流；另外，它很难在通用的网络处理器中实现而必须使用专

2006-05-08 收到，2006-09-29改回
国家自然科学基金(90604019，60472067，60502037)，国家 973 (2006CB701306)和 CNGI(CNGI-04-8-1D)资助课题

用的ASIC(Application-Specific Integrated Circuit)设计，因而实现代价较高。文献[7]从队列管理需求出发提出一种使用最近最久未使用(Least Recently Used, LRU)缓存的方法用来鉴别长时高带宽流，此方法的优点是简单直接，但是其大流漏检率高。文献[8]利用贝叶斯理论提出了一种从周期性抽样的分组中推测大流的方法。由于它基于抽样和估计，因而测量误差较大。Papagiannakit等人[9]根据流量工程的需求考察了流的动态特性对流量工程的影响，他们关注的是大流检测的长期一致性而不是测量的准确性。最近，文献[10,11]分别提出了每流测量的新方法。但是它们只能用于测量每个流的分组个数而不能用于测量字节数。

本文提出了一种新的基于最近最久未使用(LRU)大流检测算法，它引入“小流早期丢弃”和“大流预保护”两个机制；使用实际互联网数据进行的实验证明：与已有算法相比，新算法具有更高的测量准确性和实用性，特别适合于流量计费等应用。

2 基于 LRU 算法

与文献[3]中一样，本文定义的“大流”为在一定测量时间段(measurement interval)内字节总数超过给定值(例如 0.1%的链路容量)的流。

2.1 简单算法及其缺点

针对队列管理应用，Smitha[7]曾提出使用LRU缓存来检测大流。其基本思想是：维护一个固定大小的LRU缓存，到达的新流依次记录到缓存中；始终保持最新到达分组所属流的流记录位于缓存的最顶部，而最久未到达分组所属流的流记录位于缓存的最底部；当有新流到达而缓存已满时，把缓存最底部的流替换出去为新流腾出空间。

由于小流持续时间短或者分组到达速率低，因而总有可能被替换出LRU缓存。而大流往往持续时间长且分组到达速率高，往往会排在LRU缓存的上部从而以较大的概率保留在缓存中。但是当小流过多且突发到达时会造成某些大流被替换出LRU缓存。文献[7]的实验结果中有 10%至 20%的大流没有被检测到。为此，文献[7]又提出一种改进方案：对所有到达的分组进行抽样，只有抽样分组才对LRU缓存进行更新。这种方案虽然可以减小进入LRU缓存的小流数量从而能提高大流检测率，但它与其它使用分组抽样的解决方案[8]相同，由于要对流的字节数进行估计会造成较高流量测量误差，不适用于计费等对准确度要求高的应用中。

2.2 新的基于 LRU 算法

本文的新算法仍然以LRU为基础，其基本思想是限制小流进入LRU缓存，对有可能成为大流的流进行保护，同时提供高准确度的字节数测量。实际上，已有研究[4]指出大流的大小和其速率是强相关的，即流越大其速率越高。本文利用这一特性限制小流进入LRU缓存。

首先给出如下记号：记 R 为链路速率(为方便起见，单位为字节每秒)；C 为测量时间段内的链路容量(即测量时间段内链路满负载下所传输的字节数)；F 为大流字节数门限(即测量时间段内字节数超过此值为大流)；记大流流量所占链路容量的比例为 $p = F/C$ 。

如图 1 所示，整个算法使用 3 个缓存区：流测速缓存、LRU 缓存、大流缓存。当分组到达时，同时在 3 个缓存区中查找(下节详细讨论其可实现性)；(1)若该分组所属的流没有在任何一个缓存区中记录，则在流测速缓存中创建一个新的流记录并初始化其流大小为到达分组的字节数；每个新到达的流首先在流测速缓存中保存固定的时间 t，如果在 t 时间内流的平均速率超过 pR，则把该流的流记录移至 LRU 缓存中并使用 LRU 替换策略替换 LRU 缓存中的相应记录；如果平均速率小于 pR 则认为此流为小流，把该流的流记录从流测速缓存中删除，本文称这种主动丢弃小流的方法为“小流早期丢弃”机制。(2)若该分组所属流在 LRU 缓存中，则累加分组字节数到相应流记录的字节测量值上，如果字节总数超过 βF，则把该流记录移至大流缓存中，否则把相应流记录移到 LRU 缓存的最顶部；其中的 β ($0 < \beta \le 1$，如取 β 为 0.85)为大流保护因子，作用是把有希望成为大流的流提前移到大流缓存中以避免被小流替换出 LRU 缓存，本文称之为“大流预保护”机制。(3)若该分组所属流在大流缓存中，则按照分组字节数更新相应流记录。

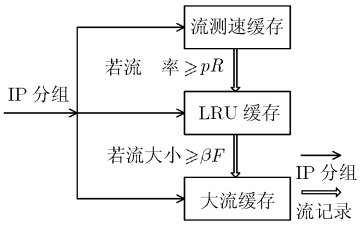

图 1 基于 LRU 大流检测算法模块图

2.3 算法分析及实现考虑

2.3.1 空间复杂度 假设 b 为分组平均大小；n 为每流平均分组数，则在时间 t 内到达流测速缓存的流个数，也即流测速缓存所需的空间(以流记录为单位)为

$$M_1 = \left\lceil t \Big/ (bn) \right\rceil \cdot R \tag{1}$$

由于每个大流的流量占链路总容量的比例为 p，所以最多会有 $1/p$ 个大流，LRU 缓存的空间可设为 $1/p$ 。为了容纳一些由于高突发速率而进入 LRU 缓存的小流，可以设定一个比例因子 $\alpha(\alpha \ge 1)$，使得 LRU 缓存的空间为

$$M_2 = \alpha \cdot 1 \Big/ p \tag{2}$$

对于大流缓存区，由于使用了“大流预保护”机制，凡是流大小大于 βF 的流都会保存在大流缓存区中，因而大流

缓存区空间应为

$$M_3 = 1/(\beta \cdot p) \tag{3}$$

因此整个算法需要的缓存空间为

$$M = M_1 + M_2 + M_3 = \left\lceil t/(bn) \right\rceil \cdot R + (\alpha + 1/\beta) \cdot 1/p \tag{4}$$

2.3.2 参数选取　算法中与应用无关的有 3 个参数：t，α，β。下面讨论它们的选取问题。

参数 t 是流测速缓存保存一个流的时间。一方面，根据式(1) 流测速缓存区的大小与 t 成正比，因而需要 t 越小越好；另一方面，它又不能太小以至于测到的流速率总是突发速率，而造成太多的小流进入 LRU 缓存，例如：若在 1Gbps 链路下需要检测超过 0.01%链路容量的大流，则进入 LRU 缓存速率门限应为 100kbps；如果一个小流仅有一个 IP 分组且字节数为最大值 1500，如果 t 小于 120ms，即使是这样的单个分组小流也会进入 LRU 缓存；同样，在 10Gbps 下，若 t 小于 12ms，也会出现这样的情况。又因为我们希望在 t 时间内收集到高速流较多的分组以尽可能平滑其突发，我们认为选取 40ms(10Gbps 链路下)至 120ms(1Gpbs 链路下)是合适的。

当然，由于 t 较小，一部分具有高突发速率的小流也会进入 LRU 缓存，因此使用 α 因子以增大 LRU 缓存以容忍这种小流的进入，由 LRU 替换策略可知，这种小流会在一定时间后被替换出 LRU 缓存。由式(2)可知，LRU 缓存大小与 α 成正比，因此 α 不易过大。参数 β 用于提前使大流进入大流缓存区以防止大流被替换出 LRU 缓存，β 越小大流被漏检的概率越低；但是由式(3)可知：β 越小大流缓存所需的空间越大。由第 4 节的实验可知：α 取 1.2，β 取 0.85 时，算法即能得到理想的准确率。

2.3.3 实现考虑及处理速度　算法的实现复杂度主要在于各个缓存的实现。流测速缓存及大流缓存可使用链式队列，LRU 缓存可使用双向链表分别实现。每个链表节点可使用静态随机访问存储器(SRAM)以提供高速访问。当分组到达时，为了能同时在 3 个缓存区中快速查找，可以把 3 个缓存区中的链表节点通过哈希表进行索引，或者直接使用基于 CAM (Content Addressable Memory)的相关存储机制提供 $O(1)$ 时间复杂度的查询功能。

根据现有研究，每分组平均 400 字节[12]，每流平均 10 个分组[13]。若 t 取 40ms 则在 10Gbps 链路满负载下，根据式(1)流测速缓存空间大小应为 12.5k流记录(在DoS等异常事件发生时，流个数会急剧增加从而需要更大的流测速空间，本文假设有其它的异常事件检测模块存在，当它检测到异常事件发生时，或者丢弃异常分组阻止进入流测速缓存，或者通知大流检测模块不进行大流检测)。若取 $p = 0.01\%$，$\alpha = 1.2$，$\beta = 0.85$，则根据式(4)算法共需缓存空间为 36.3k流记录。每个流ID最多需 104bit，若字节总数、指针各需 32 位，则每个流记录最多需 168bit。因此，在使用哈希表的情况下算法总共需约 6.1Mbit的SRAM。当前的半导体技术已经可以提供最大 64Mbit单模块的SRAM，最大可提供 1500k个表项的CAM，因此算法的实现是可行的。

算法处理每个分组的总时间为一次哈希表查询时间和相应的缓存区链表维护时间。缓存区链表维护中耗时最长操作的是从双向链表中删除节点并插入到头节点，此操作需要 7 次 SRAM 访问时间。因此最多需要 8 次 SRAM 访问时间，根据当前的半导体技术，SRAM 访问速度可达 2 至 5ns。若使用 4nsSRAM，则算法处理一个分组的总时间为 32ns，可以支持 10Gbps 链路的线速处理。

3　实验

本节使用互联网实际数据对算法进行实验验证并与已有算法进行比较。

实验数据采用互联网数据分析合作协会(CAIDA[14])及美国应用网络研究国家实验室(NLANR[15])提供的实际互联网数据，这些数据保存了一定时间内链路上经过的所有IP分组的分组头部内容，实验所用数据相关信息见表 1。数据集 OC48A-02 及OC48A-03 分别是CAIDA于 2002 年和 2003 年在美国某ISP的OC-48 骨干链路上采集的数据。实验中各采用其中的 3600s的数据。数据集UFL及PUR是NLANR分别在佛罗里达大学(University of Florida at Gainesville)和普渡大学(Purdue University)采集的数据，实验中各使用其中 180s作为实验使用。本文使用不同组织在不同网络、不同时期采集的多组数据以保证实验数据具有较好的代表性和多样性。

所有实验都采用两种流定义：5 元组流定义、目的IP地址流定义。与文献[3]相同，本文使用 5s作为一个测量时间段。

由于本文算法的基本假设是流大小满足重尾分布，因此首先对实验所使用的数据进行了统计，图 2 是实验数据在不同流定义下流大小的累积分布(Cumulative distribution)。为了清晰，在目的 IP 地址流定义下只画出了 OC48A-02 及 OC48A-03 流大小累计分布，对于 5 元组流定义只画出 OC48A-03、UFL、PUR 的流大小累计分布。由图可见，3%至 8%的流的总流量占用了链路流量的 90%以上，这是符合算法假设和前人的研究结果的。

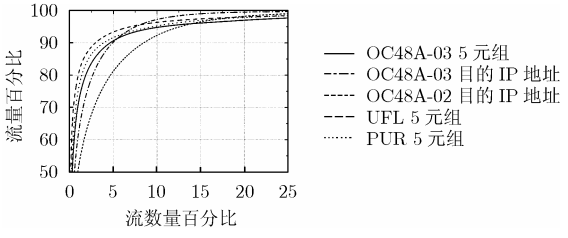

图 2　流大小的累积分布

表 1 实验数据描述

实验数据	日期	链路速率	平均每测量时间段内的流数量		平均每测量时间段内的数据量
			5 元组	目的 IP 地址	(Mbyte)
OC48A-02(CAIDA)	2002-8-14	2.5Gbps	173062	29529	385.28
OC48A-03(CAIDA)	2003-1-15	2.5Gbps	44987	14314	185.90
UFL (NLANR)	2004-6-16	622Mbps	25940	10843	196.13
PUR (NLANR)	2005-11-30	1Gbps	57756	6248	260.06

3.1 评价指标

从应用层角度看，合理的大流检测算法应该满足两方面的需求。一方面，从网络管理者角度看：大流检测算法应尽可能检测到所有大流，而且对所有大流字节数的平均估计误差尽可能小。例如，对于计费应用，平均误差衡量的是收费者进行收费的准确性，过大的误差或者会造成收费者较大的经济损失或者会造成被收费者的不满和投诉。因此本文使用大流漏检率和平均误差两个指标来评价这方面的性能要求，前者定义为算法检测到的大流与实际大流个数的比值，后者定义为所有大流的字节测量误差绝对值总和与所有大流总字节数的比值。

另一方面，从被管对象(即单个流)角度看：小流不应该被误判为大流；大流的字节数不应该被过多地测量。例如，小流若被误判为大流，对于计费应用将意味着以更高的费率对客户收取过多的费用进而引起客户的投诉，对于主动队列管理或安全应用将意味着小流被限速或被断开连接从而影响小流的QoS。同样，如果大流被过多地测量，在计费应用中也将引起客户的投诉。因此，实验中使用两个指标来评价大流检测算法在这方面的性能：小流被误检为大流时相对估计误差的最大值、大流相对估计误差的最大值。相对估计误差定义为误差与实际值的比值。

3.2 实验评价及比较结果

由于文献[3]中的“Sample and hold”及“Multistage filters”算法是目前测量准确率最高的两个算法，所以选用它们作为本文算法(简记为LRU$^+$)的比较对象。本文算法，取 $\alpha = 1.2$ ，$\beta = 0.85$ ，$t = 40$ms (2.5Gbps时)或 $t = 120$ms (622Mbps及 1Gbps时)。为了保持公平性，3 个算法对每个实验数据集采用相同的存储空间，其大小按式(4)计算。

本文分别定义大流字节数门限 F 为 0.1%，0.05%，0.01%链路容量，对这 3 种情况使用 3 个算法对不同的数据集进行大流检测，最后统计各个算法在所有数据集上的漏检率及平均误差。表 2 及表 3 分别时是流定义为五元组和目的 IP 地址时的实验结果。

从结果可见，“Multistage filters”的大流漏检率均为 0，LRU$^+$只有在 0.01%情况下才有较小的漏检率，而“Sample and hold”的漏检率最高。对于平均误差，LRU$^+$最低，Multistage filters次之，对于计费等对准确度要求较高的应用来说，这两个算法的误差率都能满足网络管理者的要求；而对于“Sample and hold”，平均误差偏高(5%至 10%)，由于“Sample and hold”的测量值总是比实际值要小，这意味着在计费应用中网络管理者将会总是少收较多的费用，因此这种算法较难被网络运营者接受而用于计费。

由于“Sample and hold”和LRU$^+$算法对每个进入大流缓存区中的流都进行每分组测量，即使小流进入大流缓存区，算法也可以通过最后的字节测量值来进一步区分出小流，因此这两个算法都不会存在小流被误判为大流的错误。LRU$^+$算法由于对每个大流都进行每分组测量，因此可以保证大流的字节测量值不会比实际值大。“Sample and hold”也可以保证测量值总比实际值小。这种特性适合于计费应用，因为它从不会对客户多计费从而可以保证“零投诉率”。但是“Multistage filters”是一种近似算法，在检测大流时，多个流的字节数会被保存在同一个记录单元。因此会出现小流被误判为大流、大流字节数被过高测量的错误。表 4 是大流字节数门限为 0.05%链路容量时，“Multistage filters”在不同数据集上产生误差的实验结果统计：小流被误检为大流时相对估计误差的最大值、大流相对估计误差的最大值。由结果可见：对于小流，误差最大可达 1689 倍于实际值；对于大流，误差最大可达 91%实际值，在计费应用中意味着对某个大流多收 91%的费用，这对用户是不可忍受的。

表 2 漏检率及平均误差实验结果(流定义为五元组时)

大流字节数门限	漏检率(%)			平均误差(%)		
	Sample and hold	Multistage filters	LRU$^+$	Sample and hold	Multistage filters	LRU$^+$
$> 0.1\%$	0	0	0	0.18	0.039	0.005
$> 0.05\%$	0.96	0	0	5.89	0.594	0.018
$> 0.01\%$	2.7	0	0.002	10.27	1.282	0.151

表 3　漏检率及平均误差实验结果(流定义为目的 IP 地址时)

大流字节数门限	漏检率(%)			平均误差(%)		
	Sample and hold	Multistage filters	LRU$^+$	Sample and hold	Multistage filters	LRU$^+$
$> 0.1\%$	0	0	0	0.072	0.018	0.002
$> 0.05\%$	0.47	0	0	2.813	0.424	0.010
$> 0.01\%$	1.62	0	0.001	6.752	1.087	0.139

表 4　最大相对误差实验结果(Multistage filters 算法)

实验数据	5 元组		目的 IP 地址	
	小流最大相对误差(%)	大流最大相对误差(%)	小流最大相对误差(%)	大流最大相对误差(%)
OC48A-02	86529	91.72	14421	24.83
OC48A-03	75410	28.51	25136	10.32
UFL	168937	27.01	112625	15.29
PUR	168156	84.18	0	5.94

因此，相对于“Sample and hold”，LRU$^+$具有更小的平均误差，易于被网络管理者接受；相对于“Multistage filters”，LRU 没有小流被误检为大流和大流被过量测量的错误，有利于被管对象。因此本文的算法具有更高的实用性，特别是对于计费应用如此。

4　结束语

高速网络中，传统维护每流状态的流量测量方法存在可扩展性问题。现有网络中流大小的重尾分布特性使得大流检测成为流量测量中重要的可扩展解决方案。本文提出了一种新的基于 LRU 大流检测算法，它引入“小流早期丢弃”和“大流预保护”机制以提高准确度。在分析算法的空间复杂度后，本文结合现有半导体技术分析了算法的可实现性及算法处理速度。最后，本文使用实际互联网数据进行了实验验证和比较。结果显示：与已有算法相比，新算法具有更高的测量准确性和实用性，特别适合于流量计费等应用。

参 考 文 献

[1] Brownlee N, Mills C and Ruth G. Traffic flow measurement: Architecture. RFC 2722, Oct. 1999.

[2] Fang W and Peterson L. Inter-as traffic patterns and their implications. In IEEE GLOBECOM, Rio de Janeireo, Brazil, Dec. 1999, 3: 1859–1868.

[3] Estan C and Varghese G. New directions in traffic measurement and accounting. In ACM SIGCOMM, Pittsburgh, PA, USA, August 2002: 323–336.

[4] Zhang Y, Breslau L, Paxson V, and Shenker S. On the characteristics and origins of internet flow rates. In ACM SIGCOMM, Pittsburgh, PA, USA, August 2002: 309–322.

[5] Altman J and Chu K. A proposal for a flexible service plan that is attractive to users and internet service providers. In IEEE INFOCOM, Anchorage, AK, USA, April 2001, 2: 953–958.

[6] Shaikh A, Rexford J and Shin K. Load-sensitive routing of long-lived IP flows. In ACM SIGCOMM, Cambridge, Massachusetts, United States, September 1999: 215–226.

[7] Smitha Kim I and Reddy A. Identifying long term high rate flows at a router. In High Performance Computing, Hyderabad, India, Dec. 2001: 361–371.

[8] Mori T, Uchida M, Kawahara R, Pan J and Goto S. Identifying elephant flows trough periodically sampled packets. Proceedings of the 4th ACM SIGCOMM conference on Internet measurement, Taormina, Sicily, Italy, Oct. 2004: 115–120.

[9] Papagiannakit K, Taft N and Diot C. Impact of flow dynamics on traffic engineering design principles. In INFOCOM, Hong Kong, China, March 2004, 4: 2295–2306.

[10] Kumar A, Xu J, Wang J, Spatschek O and Li L. Space-code bloom filter for efficient per-flow traffic measurement. In IEEE INFOCOM, Hong Kong, China, March 2004, 3: 1762–1773.

[11] Hao F, Kodialam M, Lakshman T V, and Zhang H. Fast, memory-efficient traffic estimation by coincidence counting. In IEEE INFOCOM, March 2005: 2080–2090.

[12] McCreary S and Claffy K C. Trends in wide area ip traffic patterns. In ITC Specialist Seminar, Monterey, California, May 2000: 1–25.

[13] Kumar A, Sung M, Xu J, and Wang J. Data streaming algorithms for efficient and accurate estimation of flow size distribution. In ACM SIGMETRICS, New York, NY, USA, June, 2004: 177–188.

[14] Cooperative Association for Internet Data Analysis (CAIDA). http://www.caida.org/

[15] National Laboratory for Applied Network Research (NLANR). http://www.nlanr.net/

王洪波： 男，1975年生, 博士，讲师，研究方向为IP网网络测量、管理与安全、下一代互联网体系结构等.

裴育杰： 男，1977年生，博士生，研究方向为IP网网络测量与管理、流量工程.

林 宇： 男，1976年生，博士，副教授，研究方向为互联网服务质量测量、P2P技术等.

程时端： 女，1940年生，教授，博士生导师，研究方向为宽带网交换与路由、IP网QoS控制、管理、测量等.

金跃辉： 女，1965 年生，副教授，研究方向为网络测量技术与体系结构、互联网传输层协议及优化机制等.