一种分布式的宽带网络测量系统①

马维旻，曾宇胸，杨永平

(北京师范大学 珠海分校 信息技术学院，珠海 519085)

摘 要：开发了一套跨平台、分布式的网络流量测量系统。可起到测量网络业务流、评估网络性能、规划网络设计等作用。该系统由控制中心、测量探针和数据存储服务器三个模块组成，实现了对网络流量的主动测量和被动测量。主要采用了 IPFIX 标准、FLEX web 技术，实现了多任务调度、RIA 应用，具有部署灵活、扩展性好、兼容性强、数据呈现清晰直观等特性。

关键词：网络测量；测量探针；IPFIX；RIA

Distributed Broadband Network Measurement System

MA Wei-Min, ZENG Yu-Xiong, YANG Yong-Ping

(Department of Information Technology, Beijing Normal University, Zhuhai Campus, Zhuhai 519085, China)

Abstract: A cross-platform, distributive network measure system is introduced which is helpful for traffic measurement, performance evaluation and network planning and designing. The system is composed of three modules: controlling center, measurement probe and data storage server. It realizes both active measurement and passive measurement of network traffic. It realizes multi-task controlling and RIA application by adopting IPFIX standard and FLEX web technology. It is of good deployment, extensibility, compatibility, and data presentability.

Keywords: network measurement; measurement probe; IPFIX; RIA

1 引言

经过几十年的发展，Internet 上新应用不断出现。网络已从单纯的数据传输发展到同时支持多种信息类型的传输，网络流量模型从基于泊松分布转变为具有自相似的特性[1,2]，动摇了互联网的传统理论基础。对网络协议的分布、网络流量特征、用户与网络行为缺乏准确的理解与精确的描述，严重影响了对网络资源的有效利用与网络自身的发展。如何正确和准确地测量网络业务流，对实时有效地评估网络性能、预测现实网络行为和科学规划网络设计等起到了至关重要的作用。因此研究高效的网络测量技术成为当前网络研究领域热点和所需迫切解决的问题之一。

网络测量技术总体上可分为主动测量、被动测量和基于 SNMP 的网络测量三种。

(1) 主动测量是通过向网络发送一些探测包，观察其在网络中实际传输的路径及延迟和排队情况，从而推出网络当前的工作状况。

(2) 被动测量技术是对通过测量点的流量进行记录和分析，并对网络的正常运行几乎没有影响[3]。可以得到有关网络端到端性能指标，如延迟、吞吐量、延迟抖动和丢包率等，而且还可以得到被测链路上的网络流量特征，如流量的数字特征、流量的分布、协议分布、链路利用率和用户使用网络的信息等。

(3) 基于 SNMP[4]的测量是通过读取相关网络设备的管理信息库(MIB)[5]来得到反映网络状况的性能指标，读取到的 MIB 中的信息传递到测量点时将占用网络的带宽。

网络流量测量技术研究是一项系统和复杂的工作。当前，世界上有许多研究机构和项目都已加入到了从事网络测量技术研究这一行列中，比如 IPPM[6]、CAIDA[7]、NIMI[8]、DIMES[9]、CONMI[10]等，提出了很多研究成果[11-14]，对维护网络正常运行和提高网络

① 基金项目:珠海市科学计划(PC20051027)
收稿时间:2010-05-21;收到修改稿时间:2010-06-29

服务性能具有非常重要的作用和现实意义。随着网络链路速度呈指数级的增长,数据包的捕获与分析已经遇到严重的性能问题。Weigle 研究表明[15],尽管处理器的处理速度仍然以摩尔定律增长，但其相对于通信链路速度的增长仍显得微不足道，传统的基于 Libpcap 包的软件捕获工具，如 Tcpdump[16]最高速率只能达到 250Mbps。因此研究和开发具有多通道、高速、大容量数据捕获处理分析能力的网络测量系统，以满足在高速链路下的数据包的抽样、存储、实时分析以及灵活精确地控制过滤器是目前网络流量测量系统的研究重点。

本文研究开发了一套分布式网络流量测量系统——APMMP，系统实现了主动测量和被动测量两种测量技术。系统满足测量规模扩展和灵活配置的要求，并考虑到随着应用系统数量的急剧增长，那种面向特定对象的性能检测手段在准确性和扩展性上越来越不能满足要求的情况，在设计上增加了分布式和模块化的思想，使系统成为一个可扩展的测量平台。

2　系统结构与功能说明

APMMP 系统结构如图 1 所示。图中，控制中心、测量探针和数据存储服务器视具体应用环境，可以设置多台。也可以将三个模块集成在一台主机上。可同时设置多个控制中心和数据存储服务器，方便系统部署使用，提高了系统的强壮性和可靠性，实现了资源共享及可扩性，提高了系统的可维护性。

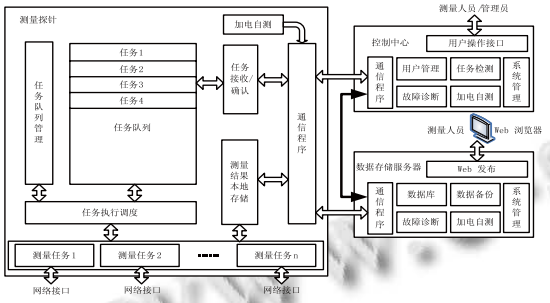

图 1　系统体系结构

系统功能主要分为如图 2 所示以下 4 个功能模块：

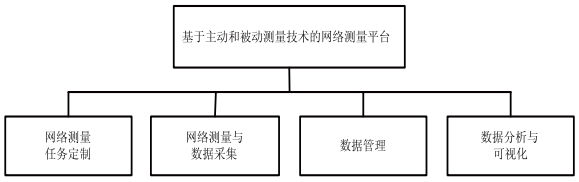

图 2　系统主要功能模块

(1) 网络测量任务定制、下发

为支持跨平台可安装要求，控制中心程序系统采用了 Java 语言开发。通过用户操作接口可定制网络测量任务和系统管理。如图 1 所示，通过测量探针上的相应网络端口，控制中心向测量探针下发测量任务，并可控制测量任务的开始和停止。如果有多个控制中心，可以在测量探针上设置相应数量的网络端口。控制中心与测量探针之间的通信采用 TCP 协议，保证测量任务的可靠传输。

测量任务可动态配置。通过用户操作接口上主动和被动测量任务模板，定制不同的网络测量任务，用户可以选择不同的测量指标，动态地构造测量任务，控制中心在每次测量时对测量探针也可进行动态选配。

控制中心可同时启动多个测量。在测量设施不互相冲突的情况下，控制端可以同时启动探针执行多个网络测量任务。系统也可同时为多个用户服务。这些任务将在测量探针上的任务队列中进行排队，来避免测量任务之间操作冲突。

(2) 网络测量与数据采集

测量探针的功能是接收、管理和运行测量任务。测量探针受控于控制中心。

测量任务队列记录和调度控制中心下发的测量任务，在测量任务需要启动测量时，提供给测量程序模块相应的测量任务参数。控制中心可以下发不同的测量任务，这些测量任务按照在队列中的排队顺序依次被执行。通过添加、删除队列中的任务来实现测量任务队列管理。

测量结果数据采集：如图 1 所示，测量探针通过不同的网络接口可以同时发起多个测量。测量数据通过相应网络接口在测量探针的本地数据库中进行存储。对性能指标(如单向时延、可用带宽、丢包率等)的测量主要采用主动测量方法。主动测量易于控制，这里要根据测量要求定制探测分组序列(协议类型、分组序号、发送间隔、分组大小等)。对网络流量的测量主要靠被动测量，通过 PF_RING 技术[17]来捕获数据包。被动测量可以根据测试需求定制数据采集的粒度和需要监控的信息，以减轻网络的开销。

(3) 数据管理

对于主动测量得到的原始数据返回数据存储服务器之前需要进行压缩以减轻网络传输负荷，为保证数据传输安全，在测量探针和数据存储服务器之间的通信采用了 SCTP 协议[18]，减少大量数据对网络带宽的占用，保证传输的效率。对于被动测量得到的业务量数据在传向数据库前已基于 IPFIX[19]模板进行了预处理，去掉了一些冗余信息。数据存储服务器根据数据

的类型采取相应的数据存贮、维护和检索策略。

(4) 数据分析与可视化

在数据存储服务器上实现了测量数据的分析、统计功能，并用 Web Service 技术进行发布。测量人员通过本地的浏览器即可打开数据显示界面，从所监测的网段导出网络流量信息，并查看每一次被动测量的统计数据，还可以查看某个 IP 地址或通过某个特定网络接口发送和接收的包、字节的数量等等。

3 系统主要技术说明

3.1 测量探针的多任务调度方法

如果要求一台测量探针上同时执行多个指定任务，比如要同时对多个网段进行流量采集与统计分析，或者同时对网络中某些服务器进行性能测试，然后把这些测试结果发送到另外指定的数据存储服务器上。就要实现同一台测量探针上执行多个不同的任务，通常采用的方法是，先执行完一个任务，再接收另外的任务，这样循环下去。这种方法的效率不高，尤其是在执行被动测试任务的时候，更需要一段长的测试时间，就不能保证响应任务的实时性。

系统使用多线程技术实现了多任务协同通信方法，测量线程间通过互斥锁和条件变量进行通讯，这样可以实时接收并且响应来自不同管理端发送过来的任务。多线程包括：接收任务线程、读取任务线程和任务执行线程，由接收任务线程负责监听网络上的控制中心发送过来的测试任务，并且把任务插入到任务队列中，完成对测试任务的接收。任务队列是一串保存任务信息的链表，线程通过互斥锁对任务队列进行添加或者删除。互斥锁用来保证一段时间内只有一个线程在执行一段代码。在多任务调度时，对互斥锁执行创建锁、上锁、解锁、查看锁状态和删除锁等操作。

互斥锁一个明显缺点是它只有两种状态：锁定和非锁定。而条件变量通过允许线程阻塞和等待另一个线程发送信号的方法弥补了互斥锁的不足，它常和互斥锁一起使用。使用时，条件变量被用来阻塞一个任务执行线程，当条件不满足时，任务线程往往解开相应的互斥锁，并等待条件发生变化。一旦其它的某个任务执行线程改变了条件变量，它将通知相应的条件变量唤醒一个或多个正被此条件变量阻塞的线程。这些线程将重新锁定互斥锁并重新测试条件是否满足。结合具体步骤说明如下：

任务接收包括如下步骤：

(1) 接收到任务信息后，先判断互斥锁是否打开，如果是锁定的状态，则休眠，如果是非锁定状态，则先设为锁定状态，然后把任务信息插到任务队列尾中，然后发个信号唤醒执行任务线程；

(2) 当退出程序时，需要将互斥锁释放处理。

执行任务线程是负责在任务队列中取出任务然后执行。任务的执行包括如下步骤：

(1) 当任务对列中没有任务的时候，线程处于阻塞状态，当收到接收任务线程的信号时，线程会被唤醒。

(2) 线程唤醒后，先判断互斥锁是否打开，如果是锁定的状态，则休眠，如果是非锁定状态，则先设为锁定状态，然后把头一个任务取出，然后把互斥锁设为非锁定状态，然后再开个进程执行任务。

3.2 基于流的被动测量技术

基于流的技术已逐渐被用于网络传输流量的分析与计费设施中，另外在设置 QoS 策略、部署网络应用和进行容量规划等方面都有更新的应用。但是，由于缺少一种输出传输流的标准格式，不利于互联网上不同厂商和运营商的设备或系统之间的互通。而 IETF 正在开展的流输出的标准化工作，一个典型的结果就是 RFC 3917 - IPFIX 标准的推出。IPFIX 标准以思科 IOS NetFlow 第九版为基础，利用以下“特征”定义流：源 IP 地址、目的 IP 地址、源端口、目的端口、3 层协议类型、服务类型字节、输入逻辑接口。

基于以上特征定义的流被用于网络设备，如路由器或交换机等向网络报告应用等输出一种标准的流信息。IPFIX 标准中还加入了对于描述流量输出至关重要的众多规则，包括时间戳、时间同步、流终止、数据包分段等信息。

APMMP 系统为实现 IPFIX 标准流输出功能，设置了一些数据结构，由于篇幅限制，只列出其中的几个主要数据结构，如表 1、表 2 和表 3 所示。

表 1 测量探针数据结构表：

表项名称	数据类型	说明
probe _id	unsigned32	用于标识测量探针
probe _name	varchar(50)	测量探针的别名
probe _ipv4address	binary(32)	探针 IPv4 地址
probe _ipv6address	binary(128)	探针 IPv6 地址
probe _interface	unsigned32	探针的接口编号
probe_protocol_version	unsigned8	协议版本（v4/v6）
probe _transport_protocol	unsigned8	传输层协议（目前：SCTP/UDP/TCP）
probe _transport_port	unsigned16	测量探针的传输层端口号

表 2 IPFIX 流表：(部分)

表项名称	数据类型	说明
flow_id	unsigned32	用于标识数据流
ip_version	unsigned8	IP 数据分组的协议版本（v4/v6）
ip_ttl	unsigned8	IP 数据分组的存活时间
protocol_identifier	unsigned8	IP 数据分组的上层协议标识
ip_class_of_service	unsigned8	IP 数据分组的服务类别
is_multicast	unsigned8	IP 数据分组的组播标识
fragment_identification	unsigned32	IP 数据分组的分片标识
fragment_offset	unsigned16	IP 数据分组的分片位移
fragment_flags	unsigned8	IP 数据分组的分片指示
ip_header_length	unsigned16	IP 数据分组的头部长度
ip_payload_length	unsigned64	IP 数据分组的负载长度
source_transport_port	unsigned16	IP 数据分组的源传输端口号
destination_transport_port	unsigned16	IP 数据分组的目的传输端口号

表 3 IPFIX 流记录表(部分)

表项名称	数据类型	说明
flow_record_id	unsigned32	用于标识每个流记录
metering_process_id	unsigned32	监测此数据流的测量进程标识
flow_id	unsigned32	该流记录的原始流标识
observation_domain_id	unsigned32	发送该流的测量探针标识
minimum_packet_length	unsigned16	最小包长度
maximum_packet_length	unsigned16	最大包长度
minimum_ttl	unsigned8	最小 TTL 值
maximum_ttl	unsigned8	最大 TTL 值
flow_start_time	datetime	该流的第一个数据包监测日期
flow_start_nano	unsigned32	该流的第一个数据包监测时刻
flow_end_time	datetime	该流的最后一个数据包监测日期
flow_end_nano	unsigned32	该流的最后一个数据包监测时刻
flow_duration_nano	unsigned64	流持续时间
flow_active_timeout	unsigned16	判断流为结束的溢出时间
flow_idle_timeout	unsigned16	判断流为空闲的溢出时间
flow_end_reason	unsigned8	判断流为结束的原因标识
octet_total_count	unsigned64	该流记录中记录的总字节数
packet_total_count	unsigned64	该流记录中记录的总包数
dropped_octet_total_count	unsigned64	该流记录中记录的被丢弃总字节数
dropped_packet_total_count	unsigned64	该流记录中记录的被丢弃总包数

3.3 基于 Adobe FLEX 的 web Service 表示层应用程序开发

APMMP 系统数据分析与可视化部分采用了目前主流的 B/S 结构。但是，浏览器是瘦客户机，对终端用户体验和扩展带来一定的局限性。传统的 HTML 应用程序功能单一、人机交互性差、安全性能不高。随着 RIA(Rich Internet Application)技术的不断发展，突破了体统 HTML 应用程序的设计限制和互动约束。对本系统来说，RIA 的最大优点体现在，只传输页面上变化的部分数据，有效减少访问数据存储服务器带来的网络流量开销，而且由于页面是 flash 格式，流量的动态显示效果十分突出。如图 3 显示了 APMMP 系统采用 FLEX[20]开发环境开发的流量显示界面。该实时监测界面可以根据用户的要求入指定协议(如 TCP/UDP 或 HTTP/FTP 等)，还可以指定包含或不包含源 IP、目的 IP 及端口的特定的网络流，并以图形的形式直观显示给用户。默认显示日流量图，即一天内的系统的流量信息，时间间隔为 5 分钟。

3.4 APMMP 实际应用

现将 APMMP 系统接入本校主教学楼出口的交换机镜像端口上，观察上午忙时网络流数量变化情况。测量结果如图 3 所示。上图直观地显示了所有网络流总数的变化曲线，其中纵坐标为流的总数，横坐标为时间。下图显示了其中一个网络流的字节数随时间变化的情况。

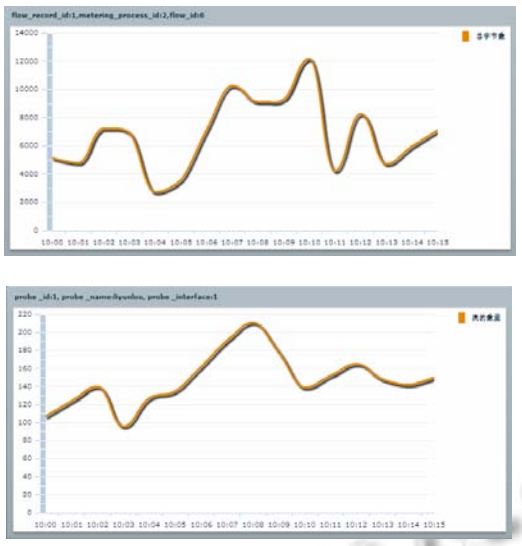

图 3 网络流信息实时变化监测结果显示

4 总结

论文论述了一个宽带网络测量系统 APPMP 的体系结构、功能模块及一些关键技术的实现方法。该系统的特点是部署灵活，有很好的结构扩展和功能扩展特性。并采用了多线程和 IPFIX 标准，保证了系统的性能和兼容性。系统支持 Web Service 的数据发布，降低了使用复杂性，提供良好的数据分析与现实功能。但是随着网络规模、流量的增加，对宽带网络测量系统的要求会显著提高，如何将取样和测量数据压缩技术运用到测量系统中将是主要研究问题之一。除此之外，还存在以下问题，一是被动测量存在的安全和隐私问题。由于被动测量可以将被测链路的传输的数据包进行拆解和解码，因此对网络用户而言，存在安全的隐患。虽然当前的被动测量系统一般只对数据包中与用户实际负载内容不相关的部分字节进行记录和分析，保护了用户的部分隐私，但是通过 IP 地址和端口号信息仍然可以对用户进行攻击；二是端口的标识问题。虽然一些常见的应用都大多数固定在某一端口，如 WWW 用 80 端口，FTP 用 21 端口，但是在实际的网络环境中，有一些应用未采用这些固定的端口，这就给应用类型的区分带来了不便。因此安全问题和应用类型的区分问题也需要进一步的研究。

参考文献

1 Leland WE, Taqqu MS, Willinger W, et al. On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Trans. on Networking (TON), 1994,2(1):1－15.

2 Paxson V, Floyd S. Wide area traffic: the failure of Poisson modeling. IEEE/ACM Trans. on Networking (TON), 1995, 3(3):226－244.

3 马维旻,王俊峰,叶晨.一种分布式的被动测量系统设计.计算机应用研究,2004:282－285.

4 Case J, et al. A Simple Network Management Protocol (SNMP). RFC1157, IETF, 1990.

5 McCloghrie K, Rose M. Management Information Base for Network Management of TCP/IP-based Internets: MIB-II. RFC1213, IETF, 1991.

6 2008CAIDA: The Cooperative Association for Internet Data Analysis.[2010-4-22]. http://www.caida.org/

7 IPPM(IP Performance Metrics).[2010-4-22]. http://datatracker.ietf.org/wg/ippm/charter/

8 Paxson V, Adams AK, Mathis M. Experiences with NIMI. Proc. of Symposium on Applications and Internet (SAINT) Workshops, 2002. 108－118.

9 Shavitt Y, Shir E. DIMES: Let the internet measure itself. ACM SIGCOMM Computer Communication Review, 2005, 35(5):71－74.

10 Claffy K, Crovella M, Friedman T, et al. Community-oriented network measurement infrastructure (CONMI) workshop report. ACM SIGCOMM Computer Communication Review, 2006, 36(2):41－48.

11 Song HH, Qiu L, Zhang Y. NetQuest: A flexible framework for large-scale network measurement. Proc. of ACM SIGCOMM Conference, June 2006. 121－132.

12 Machiraju S, Veitch D. A measurement-friendly network (MFN) architecture. Proc. of ACM SIGCOMM Workshop. November 2006. 53－58.

13 Soule A, Nucci A, Cruz RL, et al. Estimating dynamic traffic matrices by using viable routing changes. IEEE/ACM Trans. on Networking (TON), 2007,15(3):485－498.

14 Ringberg H, Soule A, Rexford J, et al. Sensitivity of PCA for traffic anomaly detection. ACM SIGMETRICS Performance Evaluation Review, 2007,35(1):109－120.

15 Weigle E, Feng WC.TICKETing High-Speed Traffic with Commodity Hardware and Software. Proc. Passive and Active Measurement, 2002.

16 Jacobson V, et al. Tcpdump. [2010-4-22]. ftp:// ftp .ee.lbl.gov, 1989.

17 PF_RING. [2010-4-22]. http://www.ntop.org/PF_RING.html.

18 Stream Control Transmission Protocol (SCTP).[2010-4-22]. http://www.sctp.org/

19 IP Flow Information Export (ipfix). [2010-4-22]. http://datatracker.ietf.org/wg/ipfix/charter/

20 Adobe FLEX. [2010-4-22]. http://www.adobe.com/products/flex/