可扩展高速网络流量被动测量平台的设计与实现

徐加羚,龚 俭
(东南大学 计算机科学与工程系,江苏 南京 210096;
江苏省计算机网络技术重点实验室,江苏 南京 210096)

摘 要:随着高速网络测量研究内容的扩展,网络测量设施在提高性能的同时须支持测量的可扩展性以适应不同网络环境和添加新测量研究的需要。针对已有网络测量软件可扩展性的不足,分析了高速多链路逻辑信道的特点及对被动测量方式的影响,设计和实现了一套适用于高速网络环境的可扩展被动流量测量平台系统。该系统基于多机协同数据流模型,采用分层耦合设计结构、对象化抽象和XML格式的交互描述,支持对高速多链路逻辑信道测量与新测量功能的可扩展性。

关键词:被动测量;多链路逻辑信道测量;网络测量;网络管理
中图分类号:TP393.06 **文献标识码**:A **文章编号**:1673-629X(2006)09-0132-04

An Expansible High-Speed Network Traffic Passive Measurement Platform

XU Jia-ling, GONG Jian
(Computer Science and Engineering Dept., Southeast Univ., Nanjing 210096, China;
Provincial Key Laboratory of Computer Network Technology, Nanjing 210096, China)

Abstract: As the expanding of high-speed network measurement research, network measurement infrastructure, besides performance improvement, should provide expansibility to support kinds of high network link types and new measurement needs. However the existing software-based network measurement methods today are short of expansibility and unfit for the high-speed network measurement. Based on the analysis of the multi-link logic channel, a typical kind of high-speed network connection, and its impact on passive measurement methodology, this paper gives out an expansible high-speed network traffic passive measurement platform. This platform's design uses data flow model of multi-processor system, layered coupling framework, object-oriented abstraction and XML formatted intercommunication to make it much easier to support high speed multi-link logic channel measurement, and to add new measuring function when needed.

Key words: passive measurement; multi-link logic channel measurement; network measurement; network management

0 引 言

随着近年来高速网络技术的迅速发展,高速吉比特和10吉比特光纤传输技术应用已经普及,同时高速网络的管理和研究也成为当前网络研究的热点之一。作为网络管理和研究的支撑技术的网络测量技术,特别是被动测量方式,必须能够适应高速网络研究的需要。由于高速网络研究的快速发展带来的研究内容的不断扩展和高速网络环境的多样性,适用于高速网络环境的测量系统在提高性能的同时,必须具备测量功能的可扩展性和支持多种高速网络环境的适应性。

然而,目前常用的网络测量软件手段在设计上缺乏可扩展性和灵活性,因而不适应高速网络环境下的测量需要。常用基于软件的网络测量方式主要包括开发库(包)和专用系统两类。开发库以函数调用集合方式提供对流量采集和分析的静态代码支持,例如广泛使用的Libpcap[1],该方式在为测量软件开发提供很大的灵活性的同时不可避免地需要引入大量的重复开发,但性能比较低下,不能适应高速测量的需要,且缺乏可扩展性。专用的测量软件系统软件针对具体的测量需求给出了定制解决方案,如Netflow[2],NeTraMet[3]为代表的网络测量与监控系统。此类系统一般也缺乏测量功能的可扩展性,无法直接增加对新的测量需求的支持。

根据上述存在的问题,笔者分析了高速网络环境中典型的多链路逻辑信道特点及对被动测量系统的影响和被动测量方式自身特点,在此基础上结合CERNET华东北地区网的实践工程和科研项目需要,设计和实现了一套适用于高速网络环境的可扩展被动流量测量平台系统。系统以多机协同结构下的数据流模型为依照,设计上采用了

收稿日期:2006-02-14
基金项目:国家973计划课题(2003CB314803)
作者简介:徐加羚(1979-),男,浙江杭州人,硕士研究生,主要研究方向为网络测量和网络行为;龚 俭,博士,教授,博士生导师,主要从事入侵检测、网络行为学、网络体系结构的研究。

分层耦合系统结构、系统单元和功能对象抽象模型以及XML格式的交互描述,支持对高速网络多链路逻辑信道测量与新测量功能添加的可扩展性。

1　高速网络环境下的被动测量

网络测量的手段包括主动测量和被动测量两类方式。主动测量是指通过向网络中主动注入报文数据后根据注入报文在网络中传输状况来测试被测网络的相应测度指标。主动测量方式主要用于端至端信息测量。由于其引入额外测试流量,因此主动测量方式存在着影响被测网络的行为(例如IPERF[4]),或测量结果会受网络旁路流量(cross traffic)影响(例如pchar,pathchar和pathrate等)[5],以及其算法模型对实际网络环境而言过于简化等缺陷。由于主动测量方式在高速高负载网络环境中的测量误差很大,从而其在高速网络环境中的使用效果并不理想。

被动测量指通过被动采集被测网络中正常业务流量数据来测量网络当前状况。由于该方式不影响被测网络行为,且从应用业务流量中能获取丰富信息和用于多点协同测量,被动测量从而在高速网络环境中得到了越来越多的应用,例如目前的IPTraf[6],CoralReef[7],Scampi[8],Perme[9]和Watch[10]系统等实用和试验软件系统。

对多链路逻辑信道(Multi-link Logical Channel)是常用的高速主干网络信道组织形式之一。它指主干路由设备间通过多条并行的物理链路组成更高(或更可靠)传输吞吐的逻辑传输信道。在多链路逻辑信道上,所传输的业务数据根据负载均衡策略(或链路备份策略)动态随机分配到各条物理链路上完成实际传输。因而,用于高速网络被动测量的测量系统必须考虑对多链路逻辑信道测量的扩展性的支持。

然而传统基于单信道方式的被动测量系统不再适用于多链路逻辑信道环境的测量工作。首先,多链路逻辑信道测量改变了测量系统的系统模型。由于通用计算机平台的硬件限制,单机测量模式无法负担多链路测量所需的处理器处理能力、内存存贮能力和对多链路的扩展性支持。因此测量系统需采用多机协同的测量系统结构来实现。其次,多机系统结构引入了时钟同步的需要。时戳信息是网络测度计算的基本数据。在单机系统中依靠本机时钟足以提供极高的短期时戳精度;而在多机系统中流量数据来自多个数据采集机,必须引入时钟同步机制来控制多个数据采集点的时钟误差。最后多链路逻辑信道带来的多机协作系统结构改变了测量算法的数据流模型。如图1(a)所示,单机系统中只有一个数据源,测量算法只需按采集-处理-输出惟一数据路径进行数据处理;而在多机测量系统中,如图1(b)所示,信道测量所需的数据可能来自不同的链路,须根据测量需要进行多数据路径归并,因此存在数据流的重组。

此外在高速网络环境下,基于被动方式的不同测量应用有不同的数据输出要求。测量应用根据输出数据量和输出速度可分为少量消息低速输出和大数据量密集输出。如流量统计、IDS和病毒过滤等应用产生的安全事件信息和统计信息的生成速度慢且数据量少;而流测度相关的测量应用所需的报文传输信息和时间信息以及用于离线分析的TRACE数据生成将高速产生大数据流输出。因而高速网络环境下的测量系统需要对不同的数据输出进行不同的优化传输方式。

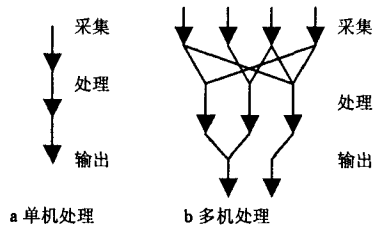

图1　被动测量的数据流模型

2　可扩展被动测量平台系统设计

2.1　多机测量数据流模型

针对高速网络多链路逻辑信道下被动测量算法模型特点和测量应用的可扩展性要求,笔者对图1(b)进行细化,建立了如图2所示的数据流模型。根据数据对象的不同,该数据流模型包含4个数据处理阶段和2次数据流重组。

该模型中,链路数据采集层获得链路数据流量数据。经过第一次数据重组后,链路数据集分成子集后送到链路数据处理层进行处理。经过链路数据处理层处理后的数据经过第二次数据重组,根据信道综合数据处理层各处理程序的需要进行归并生成信道综合数据的子集,最后经过信道综合数据处理层生成结果数据集后输出。

图2中,数据流模型中两次数据流重组采取不同的组合模式。链路流量数据到链路数据子集的重组采用了1:N的重组模式,将有限的链路数据根据可扩展的链路相关测量需求进行分配。链路数据子集到信道综合数据子集的重组采用N:1和1:N的信道测量与前级链路测量之间的数据组合关系。N:1代表了信道测量对多个链路测量的数据依赖性;而1:N代表了链路数据在多个信道测量中的复用。

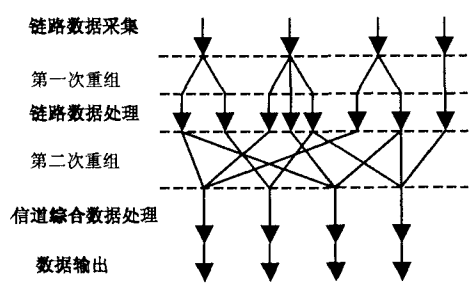

图2　被动测量的数据处理模型

2.2　测量平台系统部署

根据数据流特点,测量系统采用如图3所示的部署方案。该方案采用了测量系统内应用分层耦合的系统部署

结构并用 NTP[11]作为系统内多机间时钟同步方式。

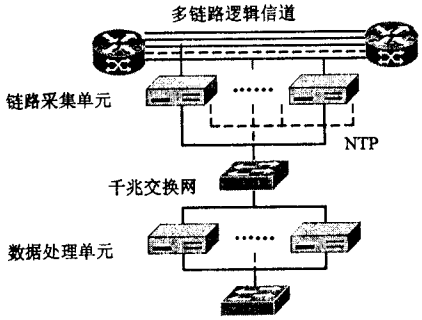

图 3 高速网络可扩展被动流量测量系统部署

系统为两层多机结构,包括链路采集单元和数据处理单元。单元是系统中担任特定任务的服务器实体。前级链路采集单元为采集链路流量、处理链路数据以及向后级单元进行数据投递。后级数据处理单元的主要工作为接收来自一个和多个前级单元的链路相关数据,计算信道流量测度并输出最终测量结果。两层结构提供了系统扩展的灵活性:一方面可根据链路数量扩充采集单元的数量;另一方面前后级单元可根据测量需要和系统负载情况进行组合。各链路采集单元通过分光器对高速通信光纤进行物理分光获得当前链路中双向流量的镜像作为系统数据输入源。在系统前后级单元之间,系统采用千兆以太交换方式(G-bit switch)互联,利用以太链路层单播和组播方式实现级间投递:以单播方式进行 N:1 的常规数据投递;以组播方式按 1:N 将相同数据投递给多个后级接收单元,减少了采集单元在进行相同数据投递时的数据复制消耗。

测量系统采用 NTP 方式实现系统内多机间时钟同步。测量系统中的链路采集单元负责对采集到的报文数据标记时间戳信息,是系统时间信息的引入点,因而需要在采集单元间使用时钟同步机制。系统链路采集单元通过独立网络连接用 NTP 方式进行时钟同步。在同一局域网内部,测量系统各采集单元设置成 NTP 同级 peers,并采用广播方式进行多方对时,同步精度可达到毫秒级。

2.3 测量平台系统框架

可扩展流量测量系统的分层框架如图 4 所示。单元由多个功能模块和控制模块按照构架设计结构组成。同级单元采用相同的结构。

前级采集单元中报文采集模块屏蔽了不同底层数据采集的具体差异,提供了统一的数据输出接口。报文分发模块负责报文数据的分发,该模块根据各处理引擎的报文地址规则和匹配规则优化报文的分发和传输路径以减少数据的复制量。

前后级单元采取相似的控制模块和控制接口。控制模块根据控制接口的规范定义接收外部控制命令,控制单元其它模块的配置和协同运行并维护单元的状态转换。

系统前后两层间的数据投递和接受模块封装了前后级单元间的数据投递。针对采集单元不同测量需求输出数据量和输出速度的不同,投递和接收模块对低速消息型数据投递和高速海量数据投递予以区分用不同的投递接口和投递方式来优化投递效率。

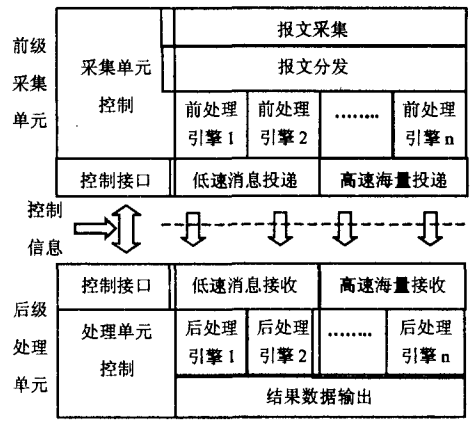

图 4 测量系统总体框架

前级采集单元中的前处理引擎模块和后级处理单元中的后处理引擎模块是系统中的可扩展部分。测量系统可根据测量需要增加多个处理引擎模块以实现测量功能的可扩展性。引擎模块通过统一的数据接口与所在模块的其他模块进行数据和控制的交互。

2.4 系统单元和测量功能的对象模型

测量系统的实现采用了面向对象的设计模型,其 UML 表示如图 5 所示。系统中所有实现实际功能的类从图 5 中的功能模块类和单元类派生。采集单元和处理单元由单元类派生并包含控制模块和其他功能模块类。

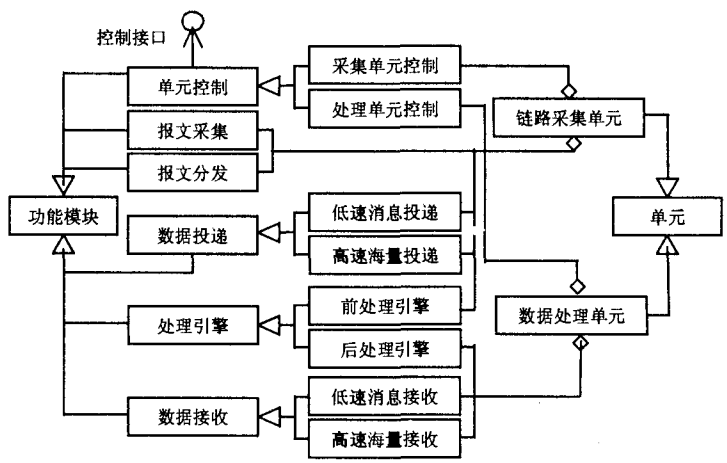

图 5 测量平台系统对象模型

在对象模型中,从功能模块类派生的各功能实现模块子类从父类功能模块继承统一的配置和控制接口。单元类中各实际功能子类从功能模块类继承的统一的接口调用实现各子功能模块对象间通信。通过对实时预处理引擎和后级处理引擎模块类的再次派生可以在系统中方便添加多个新测量功能。

在动态行为上,系统设计采用统一的状态转换。所有功能模块类采用如图 6 所示的状态转换图,当单元的单元控制模块收到配置和控制请求后就分发给相应功能模块

执行并改变各功能模块对象的状态。当单元内部所有模块的状态达成一致则该单元也进入相同状态,并由状态决定单元可执行的操作类型。

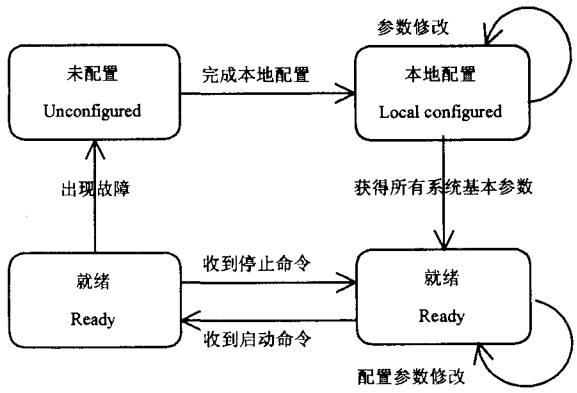

图 6　各功能模块子类的状态图

系统的配置和控制使用 XML 格式作为交互的描述方式。来自用户的 XML 格式的命令由控制模块解析后将其中的各个子元素分发给各个模块各自解析和执行。使用 XML 格式的交换形式和模块各自解析为系统提供了控制和配置的可扩展性。新添加的测量功能不需要对系统的控制和配置总体做改动,只需要在 XML 格式中添加新的元素让新模块处理即可。

3　系统原型和运行试验

根据华东北地区网络中心实际网络流量监测需要,笔者对该高速流量测量平台的测量功能进行扩充,实现了网络安全检测系统原型。原型基于 Linux 平台用 C/C++ 和 Java 语言混合开发实现,数据采集使用基于 Linux 2.4 内核修改零拷贝技术。报文分发模块采用平衡二叉树数据结构进行地址前缀匹配。实际运行中共使用 6 台服务器对 CERNET 华东北地区网络与 CERNET 国家主干之间的三条千兆光纤链路构成的逻辑信道进行监测。原型系统实现包括了传输层端口统计、传输层协议统计、特征报文统计、组流及流测度统计和应用层协议识别 6 个功能模块。原型系统在连续一周试验运行中,被测逻辑信道各条物理链路往返流量总和约 600Mbps 到 1200Mbps 之间。通过从系统网卡统计信息得到的测试结果,系统丢包率最大峰值低于 0.05%。

4　结　论

根据高速网络环境的特点以及被动测量应用对被动测量设施的可扩展性要求,分析了高速网络常用的多链路逻辑信道的特点以及被动测量方式自身的特点,提出和实现了用于高速网络环境下可扩展被动流量测量平台。相对目前常用的网络被动测量软件方式对多链路测量和测量任务扩展的支持不足,该测量平台采用了分层耦合系统结构、系统单元和功能对象抽象模型以及 XML 格式的交互描述方式,支持对高速网络环境中的多链路逻辑信道的测量,能根据需要增加测量功能。在实际主干网络环境中,该高速网络可扩展被动流量测量平台取得了很好的实用效果。

参考文献:

[1] Ranum M J, Landfield K. Implementing {A} Generalized Tool For Network Monitoring[A]. Proceedings of the Eleventh Systems Administration Conference ({LISA}'97)[C]. San Diego, CA, USA: [s. n.], 1977.

[2] Cisco Ltd. NetFlow[EB/OL]. http://www.cisco/com/warp/public/732/Tech/netflow, 2005.

[3] Brownlee N. Using NeTraMet for Production Traffic Measurement[A]. Integrated Management Strategies for the New Millenium[C]. Seattle, WA: [s. n.], 2001.

[4] Tirumala A, Ferguson J. IPERF[EB/OL]. http://dast.nlanr.net/Projects/Iperf/index.html, 2001.

[5] Shriram A, Murray M, Hyun Y, et al. Comparison of Public End-to-End Bandwidth Estimation Tools on High-Speed Links[A]. PAM 2005[C]. Boston, MA: [s. n.], 2005.

[6] Paul G. Java IPTraf[EB/OL]. http://cebu.mozcom.com/riker/iptraf/index.html, 2002-03.

[7] Keys K, Moore D, Koga R, et al. The architecture of the CoralReef Internet Traffic monitoring software suite[A]. PAM, 2001[C]. [s. l.]: [s. n.], 2001.

[8] Imec J C. SCAMPI - A Scaleable Monitoring Platform for the Internet[A]. Proceedings of the 2nd International Workshop on Inter-Domain Performance and Simulation (IPS 2004)[C]. Budapest, Hungary: [s. n.]. 2004.

[9] 程　光,丁　伟,龚　俭. 基于高速 IP 网络测量平台研究[J]. CSIT, 2004, 1(2): 122-129.

[10] 周明中,丁　伟. 高速网络测量平台 WATCH 1.0 处理器的结构设计[J]. 计算机时代, 2004(150): 40-43.

[11] Mills D L. The network computer as precision timekeeper[A]. Proc Precision Time and Time Interval (PTTI) Applications and Planning Meeting[C]. Reston VA: [s. n.], 1996. 96-108.

(上接第 131 页)

参考文献:

[1] 林　强,林英鸿. 电子商务的物流配送研究[J]. 计算机科学, 2002, 28(7): 49-52.

[2] 张　玲,左　春. 基于 J2EE 标准开发保险企业服务软件[J]. 计算机工程与应用, 2001, 37(20): 137-140.

[3] 张计龙,张成洪,张　凯. 基于改进 MVC 的高校人事管理系统[J]. 计算机工程, 2004(5): 67-70.

[4] 张宇峰,曹广益. 用 EJB 开发网上 DIY 交易系统[J]. 微型电脑应用, 2003, 17(4): 35-38.

[5] 飞思科技产品研发中心. 精通 EJB(第 2 版)[M]. 北京: 电子工业出版社, 2003.