基于Linux内核高速IP网络测量器的研究*)

徐加羚 程 光 丁 伟
(东南大学计算机科学与工程系 南京210096)

Research of High Speed IP Network Meter on Linux Kernel

XU Jia-Ling CHENG Guang DING Wei
(Department of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

Abstract With the rapid development in network technology field, the network IP packet measuring technology has become basis of the large-scale network's behavior analysis. The paper analyzes the function and performance requirement to high-speed network measure, and the construction of Linux's network subsystem. Based on these, the paper shows a proper framework design for a high-speed meter on Linux kernel and gives out an optimized quick sort arithmetic design, which gets good balance between space and time complication. At the end of the paper, a meter archetype according to the design is introduced. And any more there are several experiment data and figures of it under high-speed CERNET network environment, which show the correctness of the framework design and the high performance the arithmetic brings.

Keywords Linux kernel, Sort-arithmetic, Hash-complete binary tree, Network packet sampling, High-speed network measure

1. 引言

近年来随网络规模日益庞大，拓扑结构越来越复杂，同时网络技术也出现两级化发展，出现网络行为学[1]和为行为分析提供抽样数据的网络测量技术的发展。由于网络测量环境的提速给网络测量带来更大困难，加上网络分析技术对测量的要求越来越高，因此测量技术成为当前网络技术的热点之一。

网络测量可分为主动测量和被动测量，被动抽样测量技术由于适合高速大规模网络环境测量，逐渐开始引起业界的广泛关注[2]。针对高速网络环境的被动抽样测量系统的体系结构必须有自己的特点。

本文对基于Linux内核的高速分类进行探讨，对多个算法方案进行可行分析和性能比较。在此基础上，针对建立在Linux内核中的测量器，分析了Linux内核的网络体系结构设计，并修改Linux内核代码，开发出实用的高速抽样测量器。

2. 高速网络被动抽样测量的技术和性能要求分析

被动式抽样测量是网络行为研究的基础[3,4]，是指利用网络的日常流量作为被测对象，用抽样测量方式，利用统计概率来实现对流量总体特性的估计。大规模网络的被动测量需要满足：高速、高效——高处理性能，算法简单，快速处理过滤和匹配；用抽样的测量方式——只采集网络部分数据流，通过基于概率统计原理来分析网络状况[2]。这些特点决定高速网络被动测量技术的独特之处。

高速采样系统捕捉网络传输信息。其维度和组合简单有限，如源IP地址、宿IP地址、端口等的组合，因而适用类似直接匹配的规则表数据结构，不同于检测安全事件的IDS系统，需要复杂的动作表支持[6]。

被动测量主要用于网络行为分析和流量行为分析[1,2]，前者需协同测量，各测点数据需满足一致性；后者要求数据有随机性。流量行为分析用于分析网络业务流的指标(例如，流量大小统计、各流量业务类型的百分比)，只需网络关键节点的单点数据；网络行为分析测量关心的是整个网络的综合状况(网络延时、丢包、抖动等等指标)，因此要在网络边界进行多点协同测量[2]，故测量采样系统中的规则定义必须是原子的，即采样规则的粒度必须足够细小，使上层分析需求能最终分解用多个测量点的采样规则来表达，由采样规则得到的数据也必须保证能得到后台分析系统需要的数据。

综上所述，用于被动抽样测量的高速网络报文采样测量系统，应具备以下特征：基于直接匹配的规则表结构——针对简单匹配组合的需求，提供高速的匹配效率；具有原子性的过滤规则定义——利于简单匹配的效率提供，面向系统协同的数据综合分析；对抽样测量的支持——符合被动抽样测量要求，面向高速大型网络的宏观指标分析。

3. Linux内核的网络体系结构分析

一个完整的网络采样测量系统，和路由器相似，采样测量系统本身也是建立在一定层次的网络层次基础上，必须包含网络协议层的实现。

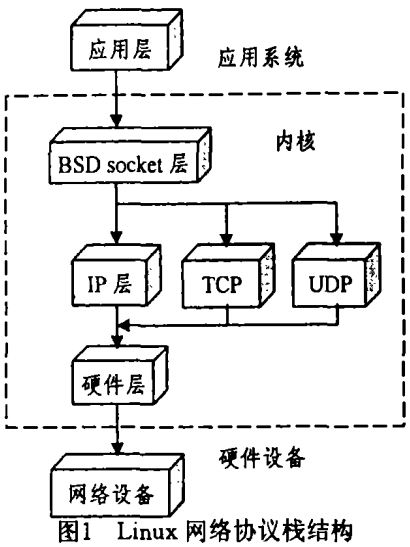

图1 Linux网络协议栈结构

*)本文受国家863课题"2001AA112060"资助。

Linux是开放源代码的操作系统，测量器的实现借鉴了其网络实现来探讨采样测量器的实现。Linux具有非常清晰和经典的Unix网络协议栈层次结构[9]，Linux网络系统可分成五个层次结构，如图1所示：硬件层/数据链路层、IP层、INET socket层、BSD socket层和应用层。除了应用层处于操作系统的用户层之外，其它四层处于内核层中，自下而上组成了Linux内核的网络系统体系[5]。

其中，在应用层和BSD socket层之间的应用程序接口以4.4BSD为模板。硬件驱动和硬件发送组织工作的层次合称为硬件层[5]。

图2为以太帧的接收过程。当线路上有报文经过时，网卡就会调用硬件中断通知系统准备接收数据。ei_receive()从网卡的接口中读出以太网帧数据。然后将数据通过netif_rx()拷贝到内存中等待处理。其中ei_receive()函数属于网卡驱动程序，无通用性不适合嵌入报文捕捉代码，且影响系统的实时性[5,8]。netif_rx()函数将收到的数据建立报文数据包，然后插入到名叫backlog的全局接收队列中去。netif_rx是从驱动程序到达上层协议栈之间的通用入口。netif_rx在硬件中断被调用，具有实时性。在此插入测量器功能代码，会影响硬件响应的实时性，且中断代码必须具有可重入性和现场保护机制，不适合复杂代码。

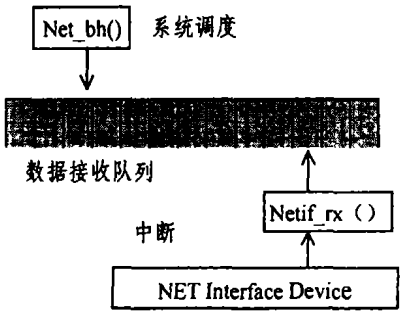

图2 以太帧的接收过程

图2中net_bh()处理backlog队列中的报文，是个非实时的系统软中断，只当系统空闲时，这个软中断才被执行。netif_rx函数每次将报文放入队列都会唤醒net_bh()。等到系统空闲时，net_bh()将被真正执行，一次处理完队列里所有的报文[8]。该位置是添加测量器代码最合适的地方。

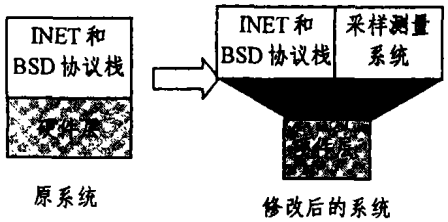

图3 Linux协议栈的转化

如图3所示，借助Linux网络底层的网络接收机制，在IP层以下，硬件层以上的net_bh()位置，构建一个于平行INET和BSD协议栈的测量采样系统，在测量器工作前，报文按照原路径通过协议栈进入Linux系统，到达上层用户层，而启动后报文被采样系统接管，不进入原有的Linux系统，直接由采样系统处理。

4. 采样测量器的体系结构分析

整个系统在功能上分为5个功能模块。控制接口——通过新增加一个专用的系统调用API来控制该系统；网络接口模块部分封装了在net_bh()函数内部对报文走向的控制；数据缓存模块用于管理采样得到的结果数据集；规则集模块封装了规则集合的数据结构和对规则的操作；转发接口模块将数据缓存中的采样结果数据通过网络接口转发到其它的主机保存或处理。图4为基于内核的网络报文高速采样测量器的结构图。

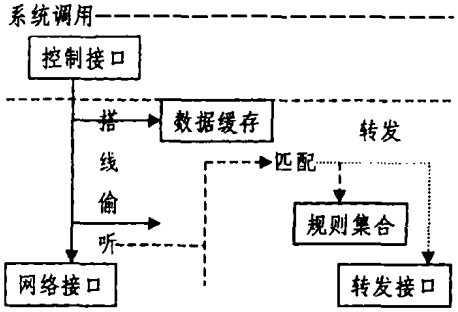

图4 采样测量器体系结构

规则集合的数据结构，对采样系统的性能起决定性作用。根据前述，已得出结论：用于被动抽样的高速测量器应基于直接匹配的规则表结构；具有原子性的过滤规则定义；对抽样测量的支持。此外，在内核中进行实现采样器同时需要尽量小的空间的占用量。在以上要求的基础上，提出了以下多种可行的数据结构和算法方案来进行分析对比。

图5所示的方案建立在哈希表结构上。其中的表头包括Hash表部分，以及必要的控制信息(读写控制等)。Hash表将4段网络IP地址形式分为三段，地址段前16位作为A段，地址后8位分成BC两段；其中C为域内主机号，作为掩码匹配部分。因此，A段的大小为2的16次方，作为第一级散列，B部分的大小为2的8次方，作为第二级散列。散列可加速搜索和简化查找，用空间换时间。将哈希查询分为两级散列是在不影响效率的基础上尽量减少空间占用的冗余。该方案的特点查找速度快，空间比较节省，实现容易，5M内存可以支持8k个c类段的过滤，但冗余大，很可能引起空间占用的几何爆炸增长，且规则几乎没有扩展的余地。

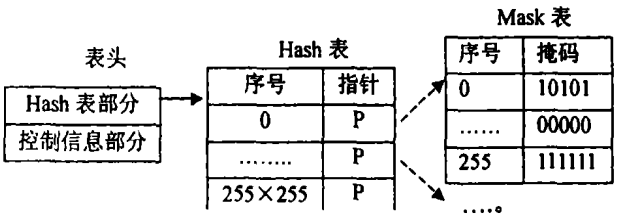

图5 方案1哈希表结构

图6的方案采用静态完全二叉树，具有以下优点：可用连续空间存放，节点间的父子关系无需用指针就可以确定；任何可折半查找的规则都可以化为对二叉树的搜索；任何二叉树都可用平衡算法换算成完全二叉树。

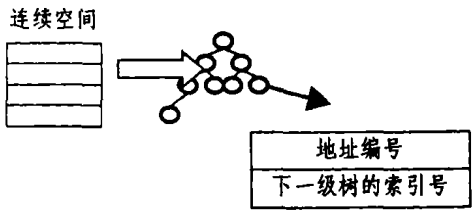

图6 方案2完全二叉结构

用静态完全二叉树可将规则表表达为森林模型形式。首先将IP地址四段式结构表示为A.B.C.D四段个8位。对于A

段，将希望过滤的所有IP地址的A段部分按照大小顺序，构造静态完全二叉树，在每个节点中填入对应A的值以及对B段树的索引。用同样方法构造的IP地址B段和C段的静态二叉树。而C的二叉树节点中可以指向掩码规则或端口规则的完全二叉树结构。对于基于端口或其它的匹配规则都可以类似挂接，使整个IP匹配规则集形成一个多维的森林结构，如图7所示。

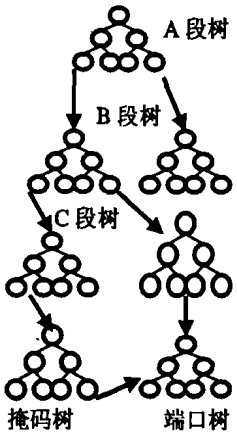

图7　方案2森林结构模型

森林形式的规则结构完全没有冗余，不需要为任何不使用值预留空间；取消了指针；所有指针均可由简单计算得出，节省了指针值空间；规则集具有可扩展性，对应端口和报文特征的匹配也可挂接到规则森林中去。

但该方案的缺点是：牺牲规则查找的绝对效率，相对于规则量增长而言，效率下降率呈现对数变化，效率还是稳定的；数据结构复杂，维护难度增加；需要规则预处理机制来生成规则树和森林；原始规则不能直接使用；规则不能动态改变。若需要对规则有所改变，必须重建整个规则森林。

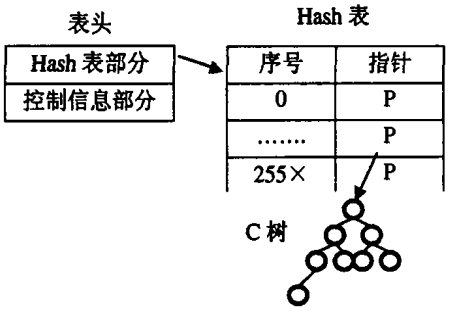

图8　方案3哈希散列—完全二叉树模型

上述两方案各有所优劣。方案一的Mask表的空间使用率不高，当Mask表数量增加时，尤其显著，其优点是散列查找速度快，而Hash表虽然空间占用大但仍可承受且不动态增长。第二种方法的缺陷在于查找效率低，实际情况下，由于用到的A和B段地址部分往往相对密集，森林结构体现不出查询的快速性，所以对A和B段不适合用二叉树结构。

结合上述方法的优点提出了第三种方案，如图8所示。使用方案中的Hash表，将AB段至少6次跳转查询缩减到1次，空间占用变大但无空间动态增长带来的几何爆炸问题。再综合完全二叉树思想，消除了原来Mask表的空间冗余问题，而查询效率损失不大。此外，使规则扩展成为可行。

Hash表占用为256k，树中的空间占用完全取决于C段地址数量。4M内存可以支持2M个C网地址段的过滤。由此可见，将哈希表和完全二叉树结构相结合的结构是测量用采样系统的规则数据结构实现比较优化的方案。

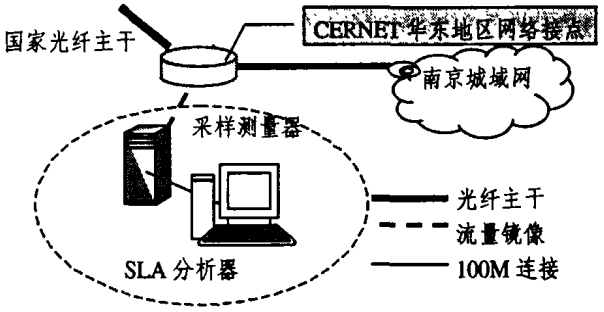

图9　测试试验的网络部署

5. 测量器功能与性能测试

采用上述的设计思想，实现了该高速IP报文采样测量器的原型系统。对该测量器的对报文采样能力、报文捕捉能力和最大负载能力已经做了正确性和性能测试。

测量器的原型系统是基于对RedHat 6.2 Linux发行版的2.2.19版本系统内核的修改。在测试中，测量器使用的硬件配置是普通的PC级服务器×86平台，使用的单颗INTEL PIII 1G Hz CPU、256M的内存。目的是在测试中在弱化硬件配置不同带来的性能差异（故选用当前主流低端硬件平台），该测量器在高速网络环境下能够达到性能表现。

测试建立在现实网络环境中的。部署如图9所示：测量器挂接在CERNET（中国教育科研网）华东地区的主干交换结点上；该结点连接华东地区网和国家光纤主干网（千兆网）。通过分光技术，采样测量器从该结点获得主干网流量的镜像流（日平均流量达到500Mb/s到800Mb/s之间）过滤采样后发往后台。

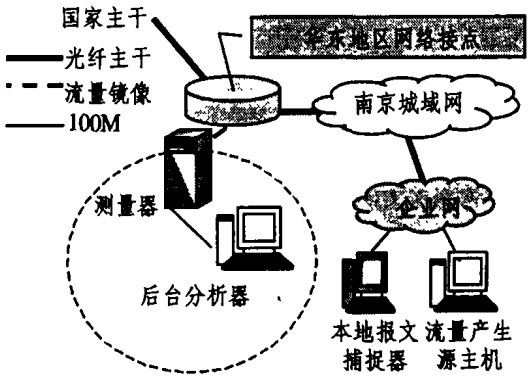

图10　测量器捕捉测验环境

在高速网络情况下对指定网络流事件的捕捉能力，是该高速IP网络报文采样测量器最重要的性能指标之一。该指标直接反应了测量器的设计思想是否正确和合理。

在图10中，设置两个测量点：一个架设在华东地区主交换接点上，测量器在该交换接点处进行过滤、采样。另一个点设在企业低速网络中，采用tcpdump进行采样。在企业的网络中通过对互联网进行不同类型的访问，产生各种应用类型流量事件。从两点捕捉到的流结果对比来判断测量器的捕捉是否正确。两个测量点对应的流事件的捕捉的结果如表1所示。

经验证，测量器捕获所有的报文与本地报文捕获器主机上捕捉到的报文在数量上和内容上均吻合。测试中，被捕捉流事件中报文的数量只占网络流量的十万分之几，测量器能够完全正确捕捉，实践检验证明了在600Mb/s以上的网络流速下，该测量器在“概率”上能做到采样零丢失率。

表1 对不同应用类型的流事件进行捕捉的对比结果

测试试验	CERNERT 总流量	测量器捕捉到报文数量	本地捕获报文数量
HTTP 应用协议	23575438	90(*含噪声报文8个)	82
ICMP(ping 应用)	28318811	49	49
FTP 应用协议	7770200	44	44
TELNET 应用协议	12319139	99	99
SMTP 应用协议	13208073	31	31
POP3应用协议	2118311	23	23
UDP 协议	11434920	19	19

注1:噪声 www 由于 http 访问中网页含有指向其他网址链接带来的流量。

注2:测试时网络主干流速为600Mb/s 到670Mb/s 之间。(2002年10月14日 晚上19:00～21:00)

在图9的网络环境下,该测量器对 CERNET 华东(北)地区网络的流量进行了实际的测量。在对指定 IP 网络流量进行捕捉采样的测试中,对 www.163.com 门户网站进行了捕捉采样 。每次的测量时间均为10个小时,时间段从当天的0:00到早上10:00,每5分钟粒度产生一次对于该5分钟内的网络流量和被捕捉流量以及丢包率的报告。在两次的测试期间(每次持续时间10小时),其丢报率均为0。

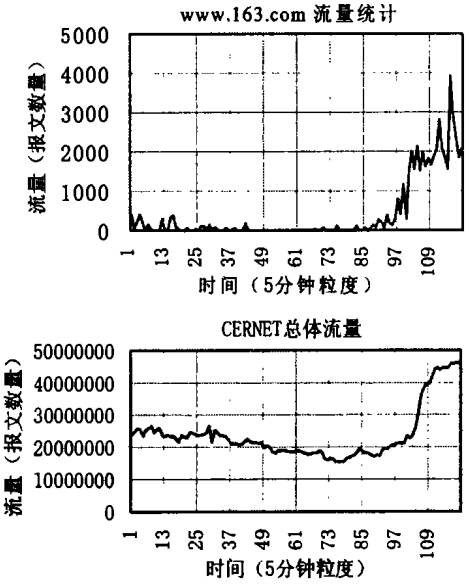

图11 www.163.com 的流量情况

对江苏某高校的7个 C 类网段的网络流量进行了测试(如图12所示)。测量时间依然为10个小时(12:00－22:00),每5分钟粒度产生一次网络流量、采样量及丢包率的报告。被测量的网络仍为 CERNET 华东北地区光纤主干网。在这10个小时的测试中,测量器的丢包数量依然为零。

此外,测量器的一项重要特性是报文采样转发能力,即测量器将所捕捉到的报文无丢失地发送到后台的能力。该能力反应了测量器的数据吞吐能力。测试中,将测量器设置成采样捕捉全网流量。在2小时连续的测试时间段中,被采样的网络的平均网络流量在533Mb/s 到600Mb/s 之间,采样测量器未出现任何丢包。

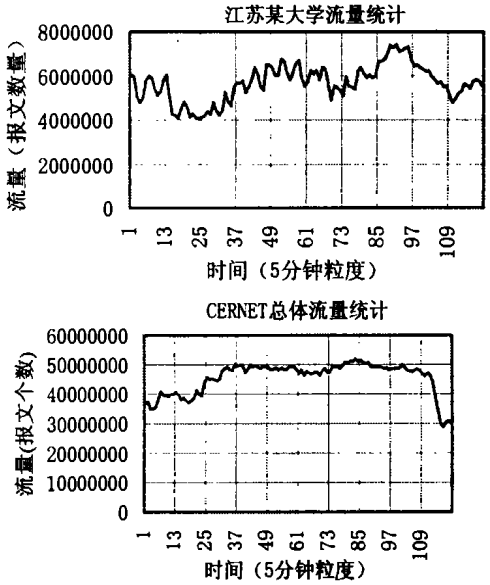

图12 江苏某高校的流量情况

上述各项测试给出了基于 Linux 内核的高速 IP 网络报文采样测量器的功能和性能的大致评价。在功能上,该测量器能够在高速网络上对指定的 IP 地址和 IP 网段的网络流报文进行准确的过滤捕捉。在性能上,该测量器能在600Mb/s 以上的网络流量环境下以零"概率"丢报率捕捉指定的网络流;在500Mb/s 以上的网络环境下达到全网采样零丢报率。

结束语 本文从高速网络环境的特点,被动抽样测量技术的特点几个方面出发,探讨和分析在高速网络环境下,对被动抽样采样测量器所必需的性能特点以及功能特点提出合理的推断分析。在结合对 Linux 内核源代码中网络协议机制分析的基础上,针对上述要求,结合 Linux 底层网络结构,提出对在高速网络环境下的被动抽样测量 IP 报文测量系统的框架构想。在此框架构想下,着重对多种可行的测量规则的数据结构和算法的时间和空间复杂度做详细的分析对比。

在实际应用中,已经建立了相关的测量器的实体原型系统并通过了实际使用测试。通过试验,从实践的角度检验了此种设计思想的理论正确性和相对于其他类型在设计思想的优越性。

高速网络环境下的被动抽样测量技术目前还处于探讨和研究中,相关的技术尚在研究阶段,大量相关的理论和实际问题仍有待探讨和解决。

参 考 文 献

1 龚俭,吴桦.网络的行为观测[J].计算机科学,2000,27(10):51～54

2 龚俭,程光.Distributed Sampling Measurement Model in a Large-Scale [J].Journal of Southeast University,2002,18(1):6～11

3 程光,龚俭.Traffic Behavior Analysis with Poisson Sampling on High-speed Network [J].信息技术与信息网络国际会议,2001

4 程光,龚俭.大规模高速网络流量测量研究[J].计算机工程与应用,2002,38:17～19

5 李善平,流文峰,李远程,王焕龙.Linux 内核2.4版源代码分析大全[M].机械工业出版社,2002

6 Network Working Group. Huston G. Traffic Flow Measurement [M]. Architecture. RFC2063,Oct. 1996

7 王学龙.嵌入式 Linux 系统设计与应用[M].北京:清华大学出版社,2001

8 (美)Scott Maxwell. Linux Core Kernel Commentary [M].机械工业出版社,2000

9 Rusling D A. Linux Kernel [EB/OL]. The Linux Document Project (TLDP). http://www.tldp.org/