基于 Linux 内核高速 IP 网络测量器的研究[1]

徐加羚，程光，丁伟
（东南大学计算机系，南京 210096）
E-mail: glacier_xu@hotmail.com

摘要 随着网络技术和应用快速发展，高速 IP 网络报文的采样测量技术成为大规模网络的行为分析的基础组成部分。文章分析 Linux 系统内核的网络体系结构，研究基于 Linux 内核，应用于高速网络被动抽样测量环境下的 IP 报文采样测量器所必须的性能要求和体系结构特点。

关键词 Linux 内核 采样技术 被动测量 抽样测量

1. 引言

当今的信息化社会中，网络的应用几乎覆盖所有的经济和科研领域。近年来网络技术发展的突飞猛进，网络硬件技术的飞跃和大众对网络需求的不断提高，造成网络发展趋势为规模越来越大，拓扑结构越来越复杂。同时带来网络技术发展两级化，在对网络性能指标多维分析上，出现网络行为学[1]；另一方面，为行为分析提供抽样数据的网络测量技术也开始独立飞速发展。由于网络测量环境的不断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，测量技术成为当前网络技术的热点之一。

网络测量技术从测量手段上可分为主动测量和被动测量，由于适合高速大规模网络环境测量，被动抽样测量技术开始崭露头角，并逐渐开始引起业界的广泛关注[2]。当前被动抽样测量技术需要解决的问题是：高速、高效——高处理性能，简单算法设计，快速完成过滤和匹配。采用抽样方式——采集部分网络数据流，基于统计的分析。因此，针对以上要求，高速网络流采样测量系统的体系结构必须有自己的特点。

文章对基于 Linux 内核的高速分类进行探讨，对多个方案进行可行分析和性能差异的比较。在此之上，针对建立在 Linux 内核中的测量器，文章分析 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技术

网络测量技术从测量方式，被分为主动和被动测量两种；而按照测量数据的不同，又分为抽样测量和非抽样测量[3]，其中被动式抽样测量的研究是网络行为研究的最基本问题[4]。被动抽样测量是指利用网络本身的业务流量作为测量的对象，采用抽样的测量方式，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，简单算法，快速处理过滤和匹配；用抽样的方式——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。

在用途方面，用于测量目的的高速采样系统，面向的是与网络传输相关信息的过滤处理任务。而用于入侵监测等目的的测量系统，多侧重于对安全事件的侦测——这类的采样系统基于对报文中含有的具有攻击特征的网络信息进行识别。对于前者，报文中用于网络传输信息的维度和组合是在可处理的范围内（如：源 IP 地址、宿 IP 地址、端口等内容的组合），因而可以使用类似直接匹配的规则表数据结构；而后者对于报文中关于安全事件的信息的识别，需要有复杂的识别行为，采样往往使用基于动作的规则表[6]。

被动测量主要用于网络行为分析和流量行为分析[1, 2]，网络行为研究需要保证不同测

[1]基金项目：本文受国家 863 课题“2001AA112060”资助

点的测量样本具有一致性要求，流量行为研究要求保证测量样本的随机性。当进行流量行为测量时，测量数据样本将用于分析有关网络业务流的指标（例如，流量大小统计，各流量业务类型的百分比），测量采样使用单点测量方式，仅需网络关键节点的单点数据既可。网络行为分析提供测量时，测量目的是为了最终生成关于整个网络状况的综合报告，包括网络延时、丢包、抖动等等指标，以及路由跟踪等。因此，需要在网络边界进行多点协同测量[2]，各测量点的数据只有汇总做比较分析才能得出相关的结果。同时，被动测量的采样系统中的规则必须是原子的，即采样系统所使用的规则的粒度能足够细小，使上层数据分析需求能够最终分解，能以多个测量点的采样规则来表达，根据采样规则的数据也最终能够保证得到后台分析系统需要的数据。如果规则的粒度太粗糙，则各采样点得到的数据集就会缺乏做比较和综合分析时的可比性，即数据集之间可比的交集范围过小。

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，应该具备以下的特征：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

网络采样测量系统，从整体结构上来看，除了其上层的过滤和结果处理机制外，还必须包含具备下层的网络协议层机制。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，如对网络层进行过滤的采样系统至少具备物理层、链路层。而传统的流量测量工具，如 Tcpdump，是通过操作系统的网络协议栈实现，本身不是一个真正意义上的完整采样系统，也无性能保证。

Linux 是开放源代码的操作系统，可通过分析其网络协议栈来探讨在采样测量器中的网络协议部分的实现。Linux 的网络体系本身具有非常清晰和经典的 unix 网络协议栈层次结构[9]，因此可在适合的协议层次上添加测量器代码，将其改造为高性能的测量器系统。整体上，Linux 网络系统可分成五个层次结构，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]。

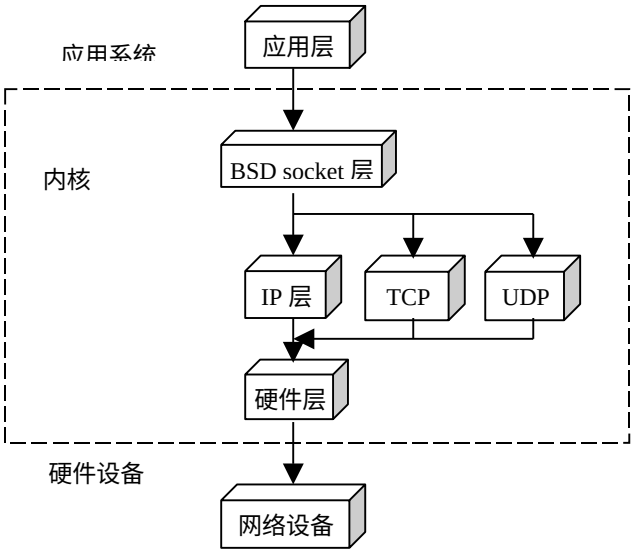

图 1　Linux 网络协议栈结构

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。INET Socket 层实现比 IP 协议层次更高，实现对 IP 分组排序、控制网络系统效率等功能。IP 层就是在 TCP/IP 网络协议栈中心的互联网层实现。硬件层在 TCP/IP 协议栈本身中就和数据链路层区分不明确，所以将硬件驱动和硬件发送组织工作的层次合称为硬件层[5]。

不同网络层次的采样器的过滤，应建立在相应的不同网络层次之上。因此文章研究的 IP 报文采样测量器的网络协议栈机制至少要包含物理层、链路层的实现以及在 IP 层同等的高度对 IP 栈进行接管。以下的分析 Linux 系统底层 IP 层一下的网络接收机制[5]。

图 2 所示以太帧的接收过程。当网卡感知线路上有报文经过的时候，就会调用硬件中断通知系统准备接收数据，Linux 系统下不同的网卡配置了对应的中断处理程序。中断处理程序来响应中断，当确定中断原因是收到报文数据，则调用函数 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_interrupt()和 ei_receive()两个函数是属于网卡驱动程序的一部分，不同的网卡硬件之间没有可通用行，所以如将采样程序嵌入该处，虽可能会有很高的效率，但是没有很好的可移值性，且造成硬件中断不能实时响应。因而不适合用来嵌入测量器代码。[5][8]

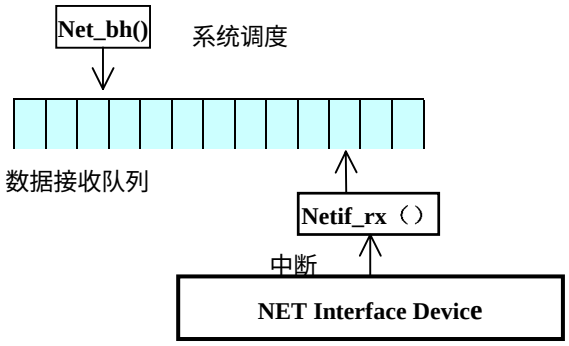

图 2 以太帧的接收过程

retif_rx()的作用是将接收到的报文拷贝到内存中，建立数据包封装报文，然后插入到一个叫 backlog 的接收队列中去。Backlog 接收队列是一个全局的队列，存放着从网络硬件层来的还未被处理的报文。对于底层不同的驱动程序，netif_rx 是一个到达上层协议栈的一个通用的入口。netif_rx 在中断程序中直接被驱动程序调用，和驱动程序拥有同样的实时性。但如需要在此对报文做过滤，在高速网络环境下会削弱硬件响应的实时性。其次，在中断程序中编写代码，要考虑到程序的可重入性，以及被中断现场保护。所以该处也不适于插入测量器代码。

图 2 中 net_bh()来处理 backlog 队列中的报文，是一个系统软中断。所谓软中断是对于系统上层来说，表现的像一个普通的系统中断，而与硬件中断不同的是，它是非实时的，只有等到操作系统有足够的空闲时，这个软中断才会得到执行。net_bh()作为软中断，在每次系统调用结束的时候会被调用一次，来检查一下 backlog 队列中是否还有未被处理的报文，有的话就做相应的处理。netif_rx 函数每次将报文放入队列都会对 net_bh()唤醒一次。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]

这种 Linux 网络底层的网络接收机制，对于建立高速采样器提供了一个非常类似的构建蓝本。文章通过借用这一机制，成功实现在 Linux 内核中嵌入高速采样测量器，取得了很好的实际效果。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，在系统改造前，报文按照原路径通过协议栈进入 Linux 系统，到达上层用户层，而修改后，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，图 3 为 Linux 系统协议栈的转化。

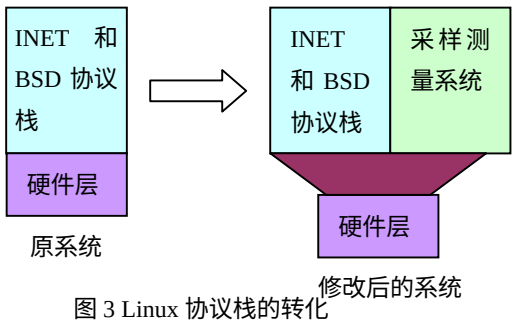

图 3 Linux 协议栈的转化

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；而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；而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。

规则集合的数据结构，对采样系统的性能起决定性作用，采样系统的主要任务就是使用规则集进行匹配。对用于被动测量的高速网络环境下的报文采样测量器特点分析时，已经得出结论：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。

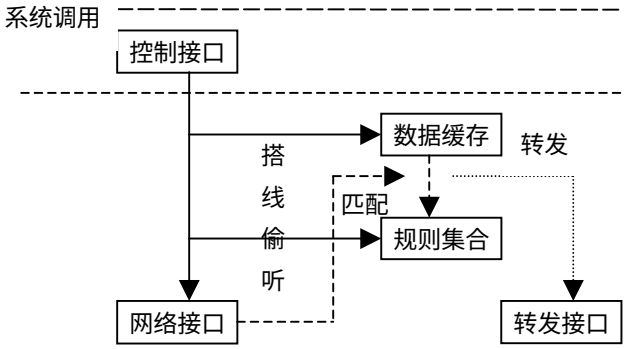

图 4　采样测量器体系结构

图 5 是方案一的数据结构。该方案建立在哈希表结构上。表头包括 hash 表部分，以及所需要的控制信息。如，读写控制，当前的表是否禁止使用，是否启用全局掩码。Hash 表将 4 段网络 ip 地址形式分为三部分，前两段地址段作为一部分 A，后两段地址分成两部分 B 和 C；其中 C 为域内主机号，作为掩码匹配部分。因此，A 部分的大小为 2 的 16 次方，作为第一阶散列，B 部分的大小为 2 的 8 次方，作为第二阶散列。采用散列可以加快搜索速度，并且简化查找，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。

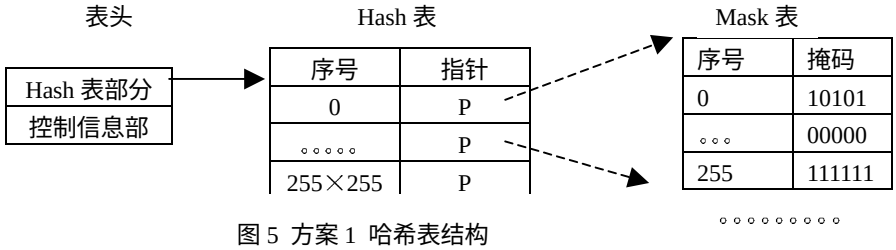

图 5　方案 1　哈希表结构

图 6 是方案 2 数据结构，该方案采用静态稠密二叉树。 静态稠密二叉树具有很好性质：可以采用用连续空间存放，树节点之间的父子关系无需用指针就可以确定；任何一种可折半查找的规则都可以化为对一棵二叉树的遍历；任何一棵二叉树都可以利用平衡算法换算成一棵稠密平衡的二叉树。

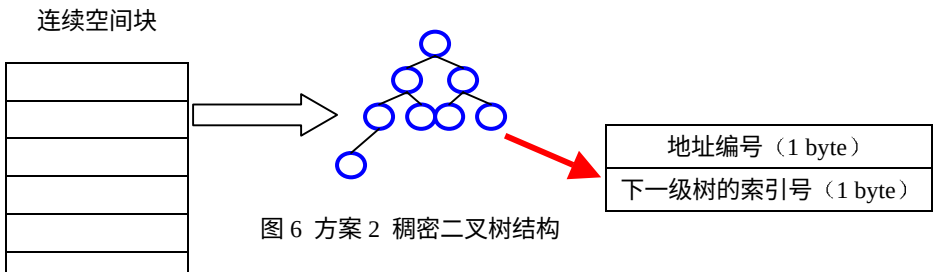

图 6　方案 2　稠密二叉树结构

利用静态稠密二叉树将规则表表达为一个森林模型形式。首先先将 ip 地址四段式结构表示为 **A.B.C.D** 四段——每段，只有 256 种可能。对于 A 段，将我们希望过滤的所有 IP 地址的 A 段部分按照大小顺序，构造静态稠密二叉树，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稠密二叉树结构。（由此也可见该方法的可扩展余地很大）。对于基于端口或其它的匹配规则都可以依次挂接上去，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。

使用森林结构的规则集合比前一种方案有更为优化。它完全没有冗余，不需要为任何不使用值预留空间；取消了指针；由于稠密二叉树的特点，所有指针均可由简单计算得出，无需空间保留绝对指针的值，节省了大量空间；规则集可很好的可扩展性，对应端口和报文特征的识别也可以采用该方法，并很好地挂接到规则森林中去，只是增加森林的维度而已，其结构没有本质变化。

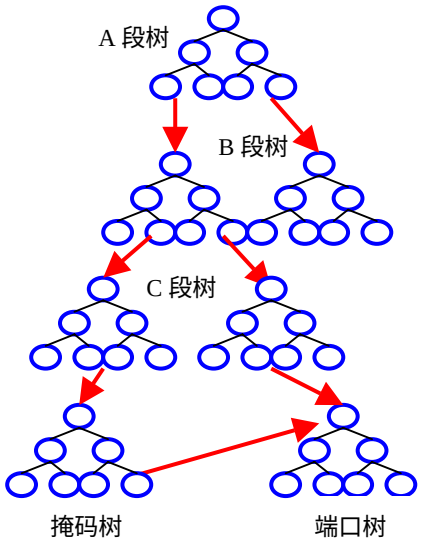

图 7 方案 2 森林结构的规则模型

缺点：

1. 牺牲了部分的规则查找的绝对效率，但相对于规则量增长而言，效率下降率变化呈现对数效应，效率还是稳定的。
2. 数据结构复杂，维护难度增加。加入了大量的空间申请，维护和释放工作。
3. 需要引进规则预处理系统，来生成规则树和森林，粗规则不能直接使用。
4. 规则不能动态改变。一旦需要对规则有所改变，必须重建整个规则森林。

前两种方案的不足在于：第一种方案的 mask 表的空间使用率不高，当 mask 表数量增加时，该效应尤其显著，其优点是散列方法的查找速度快而其中的 hash 表，其虽然空间占用大但仍在可承受范围内且不动态增长。第二种方法的缺陷在于查找效率低——例如在 4 层二叉树结构中，每棵树至少 3 层，也就是说，在规则集中查询一次，需要至少跳转 9 次。 实际情况下，用到的 A 和 B 段地址部分往往相对密集，也就是说，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。

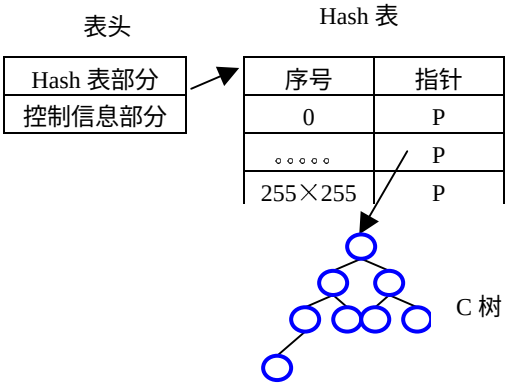

图 8 方案 3 哈希散列与稠密二叉树混合规则模型

结合上述方案的优点可生成更优化的第三种方案，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。再综合方案二中的稠密二叉树的想法，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（2^{21}）个 C 网地址段的过滤。由此可见，将哈希表和稠密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。

5. 结束语

文章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须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，针对上述的性能特点和功能要求，以及 Linux 内核特殊的编程环境，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。

文章对用于此领域的基于 Linux 内核的 IP 报文采样测量器的可行的体系结构特点做了初步的探讨和描述。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。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，计算机科学，第 27 卷第 10 期 p51-p54 ，2000

2．龚俭、程光，Distributed Sampling Measurement Model in a Large-Scale. Journal of Southeast University，Vol.18 No.1 P40，2002-6-11

3．程光、龚俭，Traffic Behavior Analysis with Poisson Sampling on High-speed Network. 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，计算机工程与应用，2002，Vol 38:5(17-19)

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

6．Network Working Group . Traffic Flow Measurement: Architecture. RFC2063

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

8．(美)Scott Maxwell. Linux Core Kernel Commentary，中文版. 机械工业出版社，2000

9．lDavid A Rusling . Linux Kernel. “The Linux Document Project “(LDP) http://www.linuxdoc.org/

Research of High Speed IP Network Meter on Linux Kernel

Xu Jialing Cheng Guang Ding Wei

(Department of Computer Science and Technology, Southeast University, Nanjing 210096)

Abstract: With the quick development in network technology field and fast widespread in WAN application, the network IP pocket sampling and measuring technology has become basis of the large-scale network’s behavior analysis. Besides brief introduction of network architecture in Linux kernel, the paper analyzes the performance need and proper architecture framework of high speed sampling system for passive network measure based on Linux kernel modify.

Keywords: Linux Kernel, sampling technology, passive measure, sampling measure