高速主干网被动测量数据采集归并系统的设计与实现

李志春，张辉，有悦，李子木
清华大学信息网络工程研究中心

摘要　网络测量是监控管理网络的基础　高速网流量采集和数据归并是它的难点问题之一　文章以 CERNET 网络管理与流量计费为背景，实现了一种高性能可扩展的网络被动测量系统 Linuxflow

关键字　网络测量；流量采集；高速互连网络；flow；Linux

Design and Implementation of A High Speed Backbone Measurement System

LI Zhichun, ZHANG Hui, YOU Yue, LI Zimu
(Network Research Center of Tsinghua University, CERNET)

[Abstract]: Traffic metering and data aggregating of high speed network are one of the difficult problems for network measurement. For the needs of network management and IP accounting of CERNET, this paper designs and implements a highly scalable performance measurement system, Linuxflow.

[Key Word]: network measurement; traffic meter; high speed networking; flow; Linux

1. 引言

大型高速 IP 主干网络的运行管理、监控分析是当前 Internet 领域内最重要的研究课题之一　通过测量网络掌握网络运行状态是网络管理研究的基础　ISO 定义的网管 5 大功能无一不和网络测量密切相关　目前国际上涉及网络测量的研究及其相应的项目部署工作正在如火如荼地展开　IETF 的 RTFM 工作组制定了实时 flow 测量的系统框架[RFC2722]，并讨论了相关的网络行为、性能分析等内容[1,2]；IPPM 工作组制定了网络性能评价的标准[RFC2330][3]　网络测量研究项目主要有　美国的 NLANR(MOAT)AMP 和 PMA、CAIDA、IEPM、Surveyor，英国的 PPNCG，日本的 MAWI 等　其参与者既有 NSF、DARPA、HPC 等科研机构和美国多所大学，也有国家政府机构、知名的 IT 企业和大型网络运营商，并在 vBNS、Internet2、NGI 等高速互连网络中得到实施，成为研究 Internet 未来发展的重要手段

CERNET 作为国内最大的研究性互连网络，很早就开展了以网络测量为基础的诸多研究工作，并将研究成果应用于流量计费、性能测量与管理以及网络安全检测等众多的网管分析课题中

网络测量有主动和被动两种手段　主动测量如 ping、traceroute 可以对特定目标进行分析，过多使用会产生大量的非正常流量，影响网络的正常使用，但因其简单易用，目前应用稍广　被动测量监听网络上特定线路的流量并进行分析，较主动测量而言，其优点在于不会对网络有任何干扰　但对于高速网络，海量数据的采集、分析处理、压缩、存储等一系列难题是目前面临的主要困难 考虑到被动测量的诸多优势，我们在 CERNET 网络环境下，研制了可用于千兆网络环境的基于 Linux 的 Linuxflow 网络被动测量系统

2. Linuxflow 设计与实现

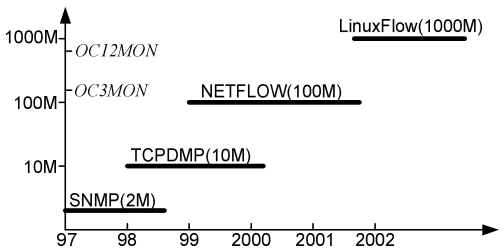

图1 网络被动测量系统在CERNET中使用的历史

基金项目　国家自然科学基金资助项目，编号　60103005
作者简介　李志春(1979　)，男，硕士研究生，主研方向　计算机网络管理与网络行为学分析；张辉，助理研究员；有悦，高工；李子木，博士后

从 1997 年至今，CERNET 主干网从 64K 升级到现在的 OC3 和 OC48，活动 IP 数量也增加了数十倍 相应地，CERNET 网络被动测量技术的发展过程及其适用的最高网络带宽如图 1 所示

SNMP 是标准的网管协议，开放性好、支持广泛、实现简单，多数网络设备对 SNMP MIB 进行扩展，比如 CISCO 的 IP accounting table (1.3.6.1.4.9.2.4.7.1)，来提供流量信息，但这种方式适用带宽低并会增加网络设备负载，影响其正常性能 因此只在 CERNET 早期测量系统采用过 2M 带宽以下 RMON 是专为远程监控服务的 SNMP MIB，不适合提供流量分析所需要的细粒度信息 Libpcap 库是 UNIX 上通用的流量监听函数库，以它为基础的监听程序 TCPDUMP 应用较广但它在超出 10M 带宽的情况下因效率低而丢失数据，早已被 CERNET 网络被动测量系统淘汰 Netflow 是 CISCO 基于 flow 测量的专有技术，其性能较高，理论上可适用 100M 带宽 但 CERNET 用户规模日趋庞大，使用 Netflow 本已略显勉强，一旦遭遇网络蠕虫泛滥的情况，其 NDE 设备负载严重恶化甚至停机，而 NFC 端则难以受理激增数十倍的垃圾 flow 此外，Netflow 也无法应用于千兆环境 因此 CERNET 已停止在主干网使用此方法 OC*X*MON 是 vBNS 开发的基于 IP over ATM 网络的专门性硬件流量采集系统 目前已有 OC3/12MON，OC48MON 正在研制过程中 OC*X*MON 由于其基于 ATM 的特殊性和高额的成本，仅在 CERNET 的科研项目中使用过 OC3MON

鉴于已有的多种技术不能很好的满足 CERNET 网管计费方面网络测量的需要，我们在权衡软硬件实现难度、成本、效果的基础上，以 CERNET 主干网络环境为研究背景，借鉴 Linux 内核网络协议栈的设计思想，实现了一套专用于流量采集的协议栈，并配合 packet 到 flow 的流量归并算法 以下称为 packet-to-flow），完成了基于 Linux 的高性能可扩展的 Linuxflow 网络测量系统 独立协议栈避免了采用内核通用网络协议栈会降低流量采集效率的缺点 在 packet-to-flow 中参考 Netflow 和 IETF RTFM 定义的基于 flow 的流量归并思路，在保留用户行为信息的同时减少了上层处理程序的数据量 目前该系统已在 CERNET 计费系统的流量采集模块中投入试运行，取得了良好效果

基于 GNU 开放源代码的 Linux 操作系统，是支持多 CPU 的强大网络操作平台 由于其内核源代码全部公开，特别是提供了易于扩展的内核 module 机制，可以方便地分析、修改、扩展内核中的功能，达到定制系统的目的和较高性能 基于以上考虑我们选用 Linux 来开发高性能的网络测量系统 目前系统的运行平台为多 CPU 高档微机工作站，Linux 内核版本为 2.4.17

本文首先介绍 Linuxflow 网络被动测量系统的总体结构，然后分别阐述系统的两个关键部分 内核空间独立流量采集协议栈的实现，用户空间调用接口和基于 Flow 的多线程归并处理实现 最后说明测量系统的分布式应用方式，并讨论其可扩展性

2.1 Linuxflow 网络测量系统的系统结构

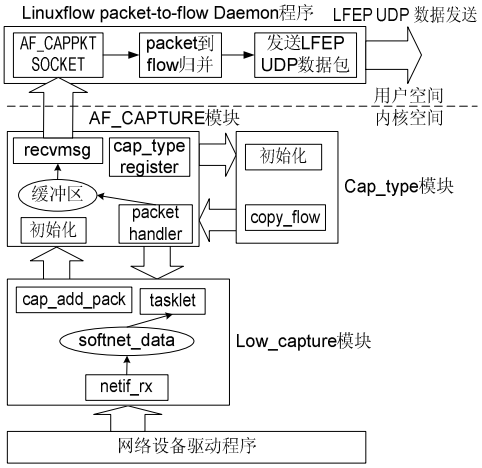

图2 Linuxflow流量采集系统的基本结构

图 2 为基于 Linux 2.4 系列内核实现的流量采集协议栈(内核空间)和用户空间的 packet-to-flow 的流量归并程序的基本结构 内核空间的流量采集协议栈程序主要通过几个内核 module 实现了协议栈的层次结构 接管内核对网卡送上来的 skbuff 的控制并通过在内核中注册的特定的 AF_CAPPKT 协议族 SOCKET 接口，实现用户空间程序对内核的系统调用，完成数据的传递机制 在用户空间通过一个多线程 Daemon 程序 packet-to-flow 实现采集的网络数据从 packet 到 flow 的归并，并通过基于 UDP 的 LFEP 协议完成 flow 数据向上层程序的递送

2.2 内核中的独立流量采集协议栈的实现

在分析 Linux 内核中通用网络协议栈的 packet 到达网络设备向高层协议传递机制的基础上，我们设计实现了专用于网络测量的独立协议栈，该协议栈主要分为三部分，如图 2 所示

- 底层包捕获部分

通过分析内核网络协议栈可知，在有 packet 到来的时候，网卡会触发硬件中断，在网卡的中断处理程序中生成 skbuff 并调用内核中的 netif_rx 例程，完成 packet 向协议栈的递送 我们通过在 module 中重新定义 netif_rx 内核符号，使网卡接收的数据被发送到我

们特殊的流量采集协议栈，而不是 Linux 的通用网络协议栈　在这一部分参照 Linux 网络协议栈实现自己的 tasklet[6]，完成接收到 packet 的处理工作　并提供 cap_add_pack 和 cap_remove_pack 接口，这样完成流量捕获功能的 SOCKET 接口就能将不同的协议处理程序 packet_type handler 链接到 tasklet 上

● 流量捕获 SOCKET 接口

此部分向内核注册 AF_CAPPKT 协议族，实现 AF_CAPPKT SOCKET 接口，提供与用户空间程序的交互，并向下层注册 packet 处理例程　程序参考了 Linux 中 AF_PACKET 和 BSD 中 BPF 的实现，采用了类似 BPF 的双缓冲结构提高效率　实现了流量采集的上层协议栈，完成了对网卡采集的 packet 进行特征抽取和数据处理、生成一定结构的记录、并通过用户空间程序的 SOCKET 调用将数据拷贝到用户空间缓冲区的工作，且提供了 cap_type 接口的可扩展挂钩

● cap_type 接口

此部分实现具体的流量特征提取、过滤和数据处理任务，根据不同流量采集任务的不同需要可以编写相应的 cap_type 接口实现　比如，可以编写程序返回完整的 packet　包含首部和数据　到用户空间，也可以只提取包头，或者只提取 packet 中感兴趣的特定成分　如计费只需要提取 packet 的源目的地址、源目的端口号、协议类型和 packet 大小即可，因此能减少内核空间向用户空间拷贝的数据量，从而提高效率　针对不同的网络测量目的，只要修改这一部分程序即可实现，保证了整个系统结构的稳定性和可扩展性

2.3 用户空间程序调用和基于 Flow 的多线程归并原理

对于用户空间的处理程序只需要通过类似代码

```
sk = socket(AF_CAPPKT,CAP_COPY_FLOW,
    ntohs(ETH_P_IP));
```

打开 AF_CAPPKT 结构的 SOCKET 接口，通过标准的 READ 或 RECV 系统调用即可获得采集到的流量数据　其中 CAP_COPY_FLOW 为 cap_type 的类型标识，对于不同的需要可以定义不同的 cap_type 和不同的返回数据格式，完成不同的流量采集功能，实现了通用流量采集系统

在大多数网络被动测量的流量采集和分析中，我们往往关心的是用户的网络行为信息　基于 flow 的流量统计分析就是针对某个域内用户产生的网络会话和 SOCKET 的分析和统计，以便掌握网络访问状态的宏观与微观信息　具体讨论可参考[1,2]

基于以上的考虑，我们参考 Netflow 的实现并以 IETF RTFM 工作组定义的系统框架[RFC2722]为基础，完成了一种基于 flow 的流量归并处理实现　由于从内核空间提取上行 packet 的流量信息，每个 packet 对应一条记录，所以这些记录数量巨大　在经过基于 flow 的归并操作后，记录数据得到大大压缩，减轻了上层程序的处理负担

我们引入一种专门针对网络流量测量的 flow 定义　具有相同源目的 IP 地址和源目的端口号、IP 分组特征类型的端到端的网络 packet 流　当每个 flow 中的第一个 packet 到达的时候，程序为它分配一条 flow record 记录，这个 flow 后续 packet 的信息不断更新 flow record 记录　基于需提取的特征在 flow record 上建立相应的 hash 表并配合匹配算法，来完成新到 packet 的匹配工作

对于每个 flow 定义最长不活动时间和最长活动时间　最长不活动时间为，如果经过一段时间没有新的 packet 到达 flow 则认为 flow 结束，准备导出；最长活动时间为，为了实时测量持续时间很长的 flow，在其持续时间超过最长活动时间后将其导出并重置　对于结束的 flow 将其导出到发送缓冲区，我们采用一种自定义的 LFEP　LinuxFlow Export Protocol　协议发送到目的主机做进一步的分析处理，如图 3 所示

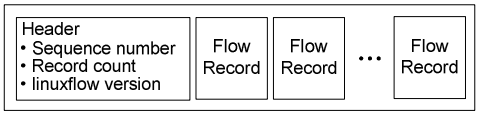

图3 Linuxflow Export UDP Datagram Format

2.4 测量系统的搭建和系统扩展性分析

对于低带宽网络(≤100M)，测量环境的搭建比较简单，仅需串接一个共享式 Hub 即可监听流量　对于高速城域网或主干网则需利用高档交换机上的二层镜像功能实现流量采集系统　如图 4 所示

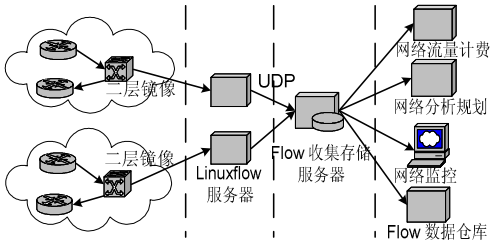

图4 Linuxflow流量采集与分析环境

图中可以配置多个 Linuxflow 流量采集服务器，分别针对网络的不同测量点测量，然后将结果通过 LFEP 协议发送到一个中心的 flow 数据收集和数据库存储服务器，从而实现分布式网络测量，并为以后的多种网络监视和分析处理提供量化数据支持　对于用单个

Linuxflow 采集服务器无法处理的高速链路，可以采用在交换机上做负载均衡将其分成带宽较低的多条线路分别测量再做集中的方法，但此种方案代价较高

3. Linuxflow 系统测试

3.1 系统性能与准确性测试

测试环境为 PIII XEON 700Mhz ×4、16GB 内存、70GB SCSI 硬盘、Intel 1000BaseSX 千兆网卡×2 的高性能微机工作站，Linuxflow 流量采集系统运行在 Red Hat 6.2 Kernel 2.4.17 系统上 测试线路为 CERNET 与 CHINANET 的千兆互连链路

性能测试主要研究系统 CPU 负载分别与网络 packet 速率和网络带宽的关系 经测试可知系统负载主要和 packet 速率相关，packet 速率相同而带宽不同时系统负载基本相等，两者间的关系如图 5 所示 准确性测试采用抽样检测的方法，在不同网络带宽下，产生确定大小的流量，和监听采集到的结果相比对，得到流量采集率 网络带宽与流量采集率的关系如图 5 所示 注 图中 packet 速率和网络带宽皆为测试线路双向带宽之和

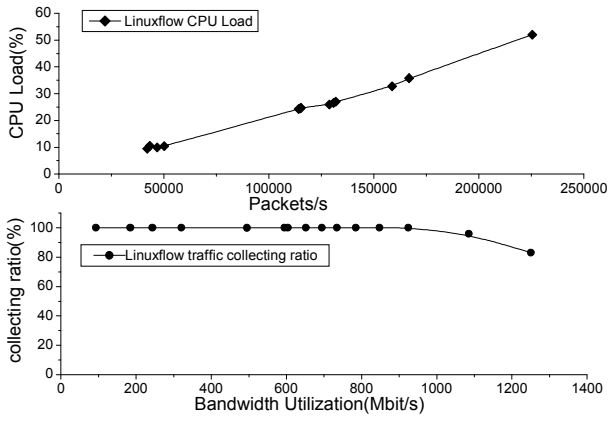

图 5 Linuxflow 性能曲线与准确性曲线

3.2 与 TCPDUMP 和针对 LIBPCAP 编程比较

测试环境为 PIII 933 Mhz×2、2G 内存、3COM 100BaseTX 网卡微机工作站 测试线路为 CERNET 100Mbps 国际链路

由于 TCPDUMP 或直接基于 LIBPCAP 的编程并没有基于 flow 的归并功能，故只与 AF_CAPPKT 作比较 TCPDUMP 丢包率在 50%以上无法通过此测试 鉴于 TCPDUMP 是基于 Libpcap 库的程序，采用直接对 Libpcap 编程进行流量采集来测试 Libpcap 的效率 编写测试程序 pcaptest 和基于 AF_CAPPKT 测试程序 lpftest 都完成对监听线路抓取每个 packet 前 60 个字节的相同操作 并且考虑 Libpcap 基于不同操作系统的差异，同机还采用 FreeBSD 测试了数据进行比较

各种采集程序和操作系统环境采集同一小时数据的结果如下

测试程序	**lpftest**	pcaptest	pcaptest
运行环境	**Linux 2.4.17**	Linux 2.4.17	FreeBSD4.4
数据采集率	**99.4%**	99.3%	99.9%
CPU 占用率	**1%**	16%	15.8%

4. 结论

本文在参考国内外网络测量领域相关技术的基础上，结合对 CERNET 高速主干网络管理和流量计费的需求特性分析，设计实现了基于 flow 的高性能可扩展被动网络测量系统 Linuxflow 该系统的特点在于

1 独立地设计并实现了 Linux 内核中专用的流量采集协议栈，保证流量采集的高效性；
2 实现了符合 IETF RTFM 实时 flow 测量理论框架的系统原型；
3 在千兆高速链路以及全网用户量大、统计复杂度高的情况下表现出良好的性能；
4 系统实施简单 一或几台主机即可完成大多数测量任务；
5 通用性好 根据不同需要可以方便地抽取用户行为的不同特征信息，完成不同的测量要求

该系统已在 CERNET 网管计费工作中得到实际验证，特别是在 RedCode 等网络蠕虫病毒泛滥时依然保证了 CERNET 计费系统网络测量功能的稳定，取得了良好的社会和经济效益

参考文献

[1] Nevil Brownlee, Margaret Murray. Streams, Flows and Torrents. PAM2001 workshop paper 2001

[2] Nevil Brownlee. Network Management and realtime Traffic Flow Measurement. Journal of Network and Systems Management, 1998, Vol 6, No 2: pp 223-227

[3]V. Paxson. Towards a Framework for Defining Internet Performance Metrics. Proc of INET ' 96, Montreal, June 1996

[4] White Paper NetFlow Services and Applications Cisco Corp. 2000 http://www.cisco.com/warp/public/cc /pd/iosw/ioft/neflct/tech/napps_wp.htm

[5] Daniel P.Bovet, Marco Cesati. Understanding the Linux Kernel. America:O'Reilly Press,2000

[6] Alessandro Rubini. Linux Device Drivers 2nd Edition.
America:O’Reilly Press,2001