

基于通用 PC 的千兆流归并系统

申文超, 杨波, 吕国晗, 严程, 李星

(清华大学电子工程系 NGN 实验室, 北京 100084)

摘要: Cisco Netflow 在流量监控、网络安全等许多方面得到了广泛的应用。目前高端路由器大都支持 Netflow 数据输出, 但使用路由器来产生 Netflow 存在输出字段不够丰富、无法定制等弊端。本文设计并实现了一种基于通用 PC 和普通千兆网卡的高速数据包采集与流归并系统, 实时采集千兆链路流量并输出 Netflow 数据。通过修改网卡驱动, 实现了数据包在内存中的零拷贝。通过多索引队列实现负载均衡, 系统可将一路网卡流量分配到多个 CPU 并行处理, 有效利用了多 CPU 的计算资源, 大大提高了系统处理能力。系统的流归并模块使用 Netflow v9 格式, 可输出丰富的流信息。测试表明, 该系统能够实现单路千兆链路环境下 100 万 pps 的数据包采集和流归并。

关键词: Cisco Netflow; 数据包采集; 流归并; 千兆链路

中图分类号: TP 393.07

文献标识码: A

文章编号: 0438-0479(2007)S2-0126-03

近年来, Cisco Netflow 在流量监控、网络安全等许多方面得到了广泛的应用^[1]。虽然目前高端路由器大都支持 Netflow 数据输出, 但其输出字段还不够丰富, 缺少如流 TTL、流应用层类型等一些有价值的信息。这使得对这些数据的分析和应用受到一定的局限。另一方面, 由于开启 Netflow 功能会使路由器负载加重, 因此那些负载已经较重的路由器无法开启 Netflow 功能。

我们设计并实现了一个基于通用 PC 和普通网卡的千兆数据包采集和流归并系统。首先, 通过修改网卡驱动, 我们实现了数据包在内存中的零拷贝, 大大提高了数据包的采集效率。其次, 我们使用流量均衡器将数据包分载到多个 CPU 进行处理, 充分利用了多 CPU 的并行计算资源。最后, 我们在修改 nProbe^[2] 的基础上实现了基于本采集模块的流归并系统, 它支持 Netflow v9 格式输出, 能够输出丰富的流信息。

与本文最相关的工作是^[3], 与文献描述的系统相比, 本系统的采集模块实现了软件流量均衡, 可将一路网卡流量送到多个 CPU 上进行并行处理。文献^[4]提出了一种软件流量均衡方法, 但是他们是在专用采集卡上实现的, 而我们则是在通用网卡上实现的。

1 系统结构

1.1 硬件平台选择

数据包采集和分析可以基于 ASIC、网络处理器和通用 PC 其中 ASIC 用硬件实现, 处理速度快, 但实现复杂, 而且不容易实现复杂的功能。网络处理器的特点是使用多个 CPU 并行处理数据包, 效率较高。但近年来, 多核多 CPU 也已成为通用 PC 体系结构的发展趋势, PC IE 也极大地提高了网卡到主存的传输带宽。这些发展使得通用 PC 体系结构和网络处理器之间的差距越来越小。而另一方面, 通用 PC 在编程上的灵活性以及价格上的优势则是网络处理器无法比拟的。因此, 我们选择通用 PC 作为数据采集和分析的硬件平台。

数据采集可以使用专用的数据采集卡, 比如 Endace 的 DAG 卡, 或者使用通用网卡。文献表明^[3,5]通过修改网卡驱动, 通用网卡也能达到很高的数据包采集性能。因此, 我们选择使用通用网卡进行数据采集。

1.2 系统软件结构

如图 1 所示, 系统功能的实现分成两个主要模块, 一个是数据包采集, 一个是流归并。

数据包采集模块有两种工作模式, 一种是单轮询模式, 一种是双轮询模式。其中, 单轮询模式的实现方式和 nCap^[3] 类似, 其性能也与 nCap 非常接近, 在此不再赘述。在双轮询模式中, 我们采用了 3 种关键技术来提高数据包采集性能。

首先是从网卡读取数据包时, 采用了轮询和中断相结合^[5]的方式。

第 2 种是数据包的内存零拷贝。初始化时, 我们在内存中分配一个循环数组队列 rx_ring 以存储数据包, 数据包通过 DMA 方式由网卡传送到 rx_ring 通过

收稿日期: 2007-08-16

基金项目: 国家 863 项目 (2006AA01Z201130) 资助

作者简介: 申文超, 男, 硕士研究生。

Email: shenwenchao@gmail.com

mmap 将 rx_ring 映射到用户进程空间. 这样,流归并程序在读取数据包时就无需进行额外的拷贝.

第 3 种是构建多索引队列实现负载均衡. 为此,我们在内核中实现了一个流量负载均衡模块 mPF_RING 均衡模块在 rx_ring 的基础上构建了几个索引队列 rx_index,用于存储 rx_ring 中对应数据包的序号. 均衡器按照一定规则分配数据包索引到不同队列中,该规则基于数据包的源地址、目的地址、源端口、目的端口等信息,从而可以保证属于同一条流的多个数据包一定会分配到同一个队列中去. 每个索引队列通过 mmap 映射到用户进程空间,由不同的数据包采集线程通过轮询方式处理. 线程在得到索引之后,根据索引值直接从 rx_ring 中读取数据包. 该均衡模块实现了多个用户态线程对数据包的并行访问,通过线程和 CPU 之间的绑定,可以将数据包分载到不同的 CPU 并行处理,从而充分利用了多 CPU 的并行计算资源.

流归并模块在修改 nprobe^[2]的基础上实现. 使用一个 Hash 表结构存储当前流状态信息,Hash 表的大小是可调的. 一个线程对采集到的数据包进行分析并据此更新 Hash 表项,另外一个线程周期性扫描 Hash 表,发现过期的流并将其输出. 流输出支持 Netflow v9 格式.

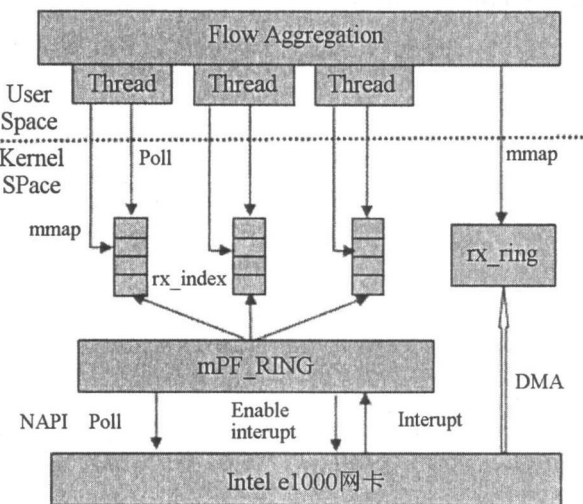


图 1 数据包采集和流归并框图

2 性能测试

系统测试试验床使用一台 Dell PE850 作为数据采集器,该主机配有一个双核 Pentium 4 3.0 GHz 的 CPU,1G 内存,一个 Intel e1000 系列网卡, Linux 2.6.18 内核和修改的 e1000 网卡驱动. 我们使用 Compass CN-100 流量发生仪发送一路千兆数据到采集器.

如前所述,数据包采集模块有单轮询和双轮询两种工作模式,我们分别使用这两种模式进行测试. 在双轮询模式中,我们在两个 CPU 上分别运行两个抓包线程, mPF_RING 将采集到的数据包交给这两个线程处理.

表 1 比较了两种模式下系统性能测试结果. 由于单轮询模式性能和 nCap 相近,因此该表也可以看成是 nCap 和双轮询模式进行比较. 表中丢包率是通过比较流量发生仪显示的数据包发送数量和采集器的数据包接收数量得到的. 我们测试了不同包速率和并行流数目的结果,发现流数目对性能几乎没有影响,因此表中只给出了 60 000 条并行流的情况. 可以看出,在 100 万 pps 的时候两种情况的丢包率都很小,但对于双轮询模式,由于流量被均匀分配到两个 CPU 进行处理,单个 CPU 的利用率均较低,多余的计算能力可以用来对数据进行更复杂的分析. 根据我们的测量,清华大学校园出口链路的包速率在 40 万 pps 左右,因此该系统能够对该链路进行实时流归并. 需要指出的是,双轮询模式中 CPU0 的利用率要高于 CPU1,高出的部分来自于 Linux 内核软中断,它是由 mPF_RING 内核模块产生的.

3 总结

本文介绍了一种基于通用 PC 和普通千兆网卡的高速数据包采集与流归并系统,它使用数据包零拷贝和软件负载均衡技术,大大提高了数据包采集效率,并能够充分利用多 CPU 进行并行处理. 测试表明,系统能够达到近 100 万 pps 的流归并能力,这远远超过一般千兆链路的数据包速率. 今后,我们拟在实际网络环

表 1 流归并性能测试结果

数据包大小和速率	单轮询模式			双轮询模式			
	CPU / %	软中断 / %	丢包率 / %	CPU 0 / %	CPU 1 / %	软中断 / %	丢包率 / %
100 (997)	100	0.0	0.37	72	44	37	0.46
300 (386)	91	0.5	0.26	32	19	17	0.04
64 ~ 1500 (157)	17	1.3	0.08	18	10	8	0.00

境下进行系统性能测试,同时对流归并模块进行扩展,比如添加对 P2P、VOIP 等流量的识别功能.随着多核 CPU 的广泛应用,该系统能够处理更多的链路以及对数据包进行更复杂的处理.

致谢 感谢清华大学网络中心王继龙、张千里提出了本项目中最初的想法,同时感谢 DragonLab 项目的支持.

参考文献:

- [1] 李卫. 细粒度访问流及其应用 [C]// CERNET 年会. 昆明, 2006.
- [2] Deri L. nProbe: an open source Netflow probe for gigabit networks[C]// Proc of Terena TNC 2003. Zagreb, Croatia, 2003.
- [3] Deri L. nCap: wire-speed packet capture and transmission [C]// Proc of E2EMON. 2005.
- [4] Loris Degioanni, Gianluca Varenni. Introducing scalability in network measurement: toward 10 Gbps with commodity hardware[C]// MC 2004. 2004.
- [5] Rizzo L. Device polling support for FreeBSD [C]// BSDConEurope Conference. 2001.
- [6] Deri L. Improving passive packet capture: beyond device polling[C]// Proc of SANE 2004. 2004.

Commodity PC Based on Flow Aggregation System for Gigabit Networks

SHEN Wen-chao, YANG Bo, LÜ Guo-han, YAN Cheng, LI Xing

(NGN Laboratory, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China)

Abstract: Cisco Netflow is widely used in many areas such as traffic monitoring, network security, etc. Most of high-end routers today can export Netflow, but their fixed format lacks some valuable information and cannot be customized. This paper covers the design and implementation of a high speed flow aggregation system based on commodity PC and general purpose network cards. By modifying the network card driver, no packet copy in the RAM is achieved. By implementing a software based traffic scheduler, traffic from one network card is assigned to several CPUs for parallel process. This improves the system performance drastically. Netflow v9 format is used. According to the performance test, the system can process traffic from a single Gigabit link with 1 000 000 pps very well.

Key words: Cisco Netflow; packet capture; flow aggregation; Gigabit link