doi:10.3969/j.issn.1000-436x.2014.z1.010

面向应急响应的高速网络流量采集设计与实现

马亚洲，龚俭，杨望
（东南大学 计算机科学与工程学院，江苏 南京 211189）

摘 要：网络安全应急响应在网络分析和追踪时需要应急采集，即捕获特定 IP、端口、协议的原始分组。基于高速网络分组捕获工具 PF_RING DNA，利用多核多线程并发采集与规则匹配的网络分组，并分配共享缓冲区提高分组的磁盘存储性能，同时通过对采集规则设置不同的状态，实现动态添加采集规则和人为干预采集过程。实验结果表明，在双万兆网卡的环境下，应急采集系统可以捕获并处理 19.98 bit/s(3.5 Mpacket/s)的网络流量，最大应急采集速率为 1 297 Mbit/s（204.9 kpacket/s）。

关键词：应急响应；PF_RING DNA；分组采集；动态规则
中图分类号：TP309 **文献标识码**：A **文章编号**：1000-436X(2014)Z1-0046-06

Design and implementation of high-speed network traffic sensor for emergency response

MA Ya-zhou, GONG Jian, YANG Wang
(School of Computer Science and Engineering, Southeast University, Nanjing 211189, China)

Abstract: In the network analysis and tracking, network security emergency response needs a emsrgency sensor that captures saw packets of specific IP, port, protocol. Base on the high-speed packet capture tool PF_RING DNA, it uses mutil-thread to capture network packets that match sensor rules, and allocates the shared buffer to improve the performance of the disk storage of packets, at the same time through setting different states for the packet sensor rule, impliments adding sensor rules and human intervention dynamically. The experimental results show that in the dual 10 Gigabit NICs environment, emergency sensor can capture and handle network traffic of 19.98 Gbit/s(3.5 Mpacket/s), and the maximum rate of emergency sensor is 1 297 Mbit/s(204.9 kpacket/s).

Key words: emergency response; PF_RING DNA; packet capture; dynamic rule

1 引言

“十一五”211 工程在 CERNET（china education and research network）网络中心和 38 个核心节点上建设了高性能网络管理与安全保障系统[1]，内容包括网络流量实时监控、网络安全异常检测和网络安全事件应急响应等。CHAIRS（cooperative hybrid aided incidence response system）系统是该项目的应急响应协同服务系统，为各节点的安全管理人员提供应急响应管理功能，提高 CERNET 安全事件响应的效率[2]。

面向应急响应的高速网络流量采集，即应急采集，是网络安全事件应急响应的重要内容，指在网络主节点检测到网络安全事件之后，为获取进一步的信息以对事件的发展进行跟踪和原因分析等响应，需要在接入网网络边界针对特定的通信对象和通信特征进行实时的网络流量采集，具体指捕获特定 IP、端口、协议的原始分组，获得分析样本，以便对该安全事件做出进一步的分析判断。面向应急响应的网络流量采集的需求，有针对性：应急采集的数据源应能为应急响应提供关键、有价值的分组交互信息；高效性：应急采集能够高效、及时响应应急采集任务，并能同时进行多个应急采集任务；准确性：应急采集在采集分组时不漏采、不多采、不错采；可控性：人为可以控制应急采集任务的开始和结束；通用性：

收稿日期：2014-10-14

应急采集应具有通用性，方便系统的移植和复用。

使用分组采集工具 PF_RING DNA 采集高速网络流量，充分利用多核多线程并行处理并筛选出所需要的网络流量，并分配分组共享缓冲区来提高分组写入磁盘的性能。实验表明，应急采集系统在多核多线程并发下能够高速捕获并处理网络流量、动态添加和控制应急采集任务，并将采集到的分组分类存储管理。

2　应急采集研究现状及面临问题

高速网络流量采集可分为基于硬件和基于软件[3]。

基于软件的分组采集工具主要基于网卡的零拷贝思想，在 Linux 环境下，常用的分组采集工具为 Tcpdump，虽然 Tcpdump 具有基于内核的 BPF 分组过滤的优点[4]，但其使用的流量捕获工具为 libpcap，libpcap 的零拷贝版本在实际利用万兆网卡的测试中，最大分组捕获速率仅为 845.68 Mbit/s（1.52 Mpacket/s），且每个进程只能使用一个分组过滤表达式。在使用 512 字长度的分组测试 PF_RING TNAPI 时[5]，最高分组捕获速率达到了 8.2 Gbit/s（2 Mpacket/s）。

基于硬件的分组采集工具主要利用硬件高速处理的优点。基于 ASIP（application specific instruction processor）技术的 NP[6]（network processor）和基于 FPGA 的网络捕获工具[3]虽然都能达到物理最大速率，但其灵活性和通用性差、成本高的缺点也是明显的。

根据应急采集的需求和现有工具的不足，高速网络流量下的应急采集面临如下问题：1）高性能捕获网络流量，为获得所需的分析样本，应急采集需要首先捕获主节点边界的全部流量，而 CERNET 主节点边界的流量通常已经超出现有软件工具的抓包能力，例如 CERNET 南京主节点边界的峰值流量已经达到 17 Gbit/s（3.17 Mpacket/s）；2）高性能分组分类筛选，高速网络流量捕获后需要处理的性能压力大，且网络流量中含有与应急采集任务无关的分组，需要从捕获的流量中筛选出与应急采集任务相关的分组，同时由于应急采集任务因时间变化、人为干预等因素而开始或结束，需要动态调整应急采集任务；3）高性能分组存储，在应急采集时可能遇到流量峰值，导致因磁盘写入性能而出现分组丢失现象，同时为了高效分析和管理分组数据，需要对应急采集到的分组分类存储、文件切分等处理。

3　应急采集解决方案和总体设计

分组捕获，首先考虑分组数据源的选择，Scott Campbell 和 Jim Mellander[7]提出了“尽可能多地收集原始数据，关注数据丰富的区域”的思想，同时指出丰富的数据位于组织内信任层的边界，例如外部边界的接口、防火墙的边界等，因此应急采集的数据源选择在 CERNET 南京主节点的边界。

应急采集的拓扑结构如图 1 所示，由于 CERNET 南京主节点的边界流量超过单个万兆网卡的捕获能力，所以使用分光器分别复制 CERNET 南京主节点与 CERNET 主干网的 2 个上联信道的进出网络流量，并分别传输给应急采集主机配置的 2 个万兆网卡，在应急采集主机上对网络流量进行筛选和分类。

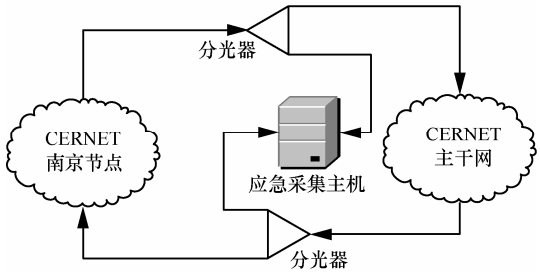

图 1　应急采集网络拓扑

为了更好地表达应急采集任务，将应急采集任务转化为应急采集规则，在应急采集主机上的总体架构如图 2 所示（分组分类线程数目和应急采集规则数目可配置，这里以 2 个分组分类线程和 3 个应急采集规则为例），针对所面临的问题，应急采集系统分为分组捕获、分组分类和分组存储 3 个模块。

针对高性能捕获网络流量，同时考虑系统的通用性，方便系统的扩展和移植，分组捕获模块使用比 PF_RING TNAPI 更高采集效率的 PF_RING DNA。PF_RING DNA[4]使用零拷贝技术同时捕获多个网卡上的分组，并使用用户定制的负载均衡算法，把分组分发给各个分组分类线程。

针对高性能分组分类筛选问题，分组分类模块充分利用 CPU 多核环境运行多个分组分类线程并对线程进行优化处理，根据动态分组分类算法对流量进行筛选和分类，将与应急采集规则匹配的分组写入到相应的共享缓冲区中。

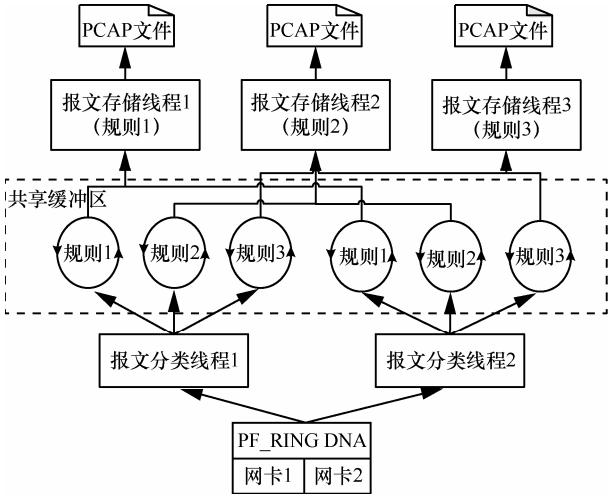

图 2　应急采集总体架构

针对分组存储的问题，为了平缓应急采集规则所对应的流量峰值，降低因磁盘性能而分组丢失的可能性，分组存储模块为每个分组分类线程分配与应急采集规则数目相等的分组缓冲区，每个共享缓冲区都对应一个应急采集规则。同时，为了按照应急采集规则对分组进行分类存储，每个分组存储线程都对应一个应急采集规则，从共享缓冲区中读取同一类的分组并保存在多个、固定大小的文件中。

4　分组捕获

为了采集 CERNET 南京主节点边界的所有网络流量（17 Gbit/s，3.17 Mpacket/s），并充分利用现有的 2 个万兆网卡，使用基于软件的网络分组捕获工具 PF_RING DNA。

PF_RING DNA 其采用零拷贝的方式，避免了 CPU 的中断参与，同时，PF_RING DNA 的 libzero 库提供了灵活、高效的分组处理机制，例如，libzero 库提供了分组汇聚的功能，如图 3 所示，通过汇聚各个网卡上的分组，再利用用户定制的分组负载均衡算法，灵活地分发给多个线程或应用。既有软件的灵活性和通用性，又有硬件高速网络流量采集的能力。

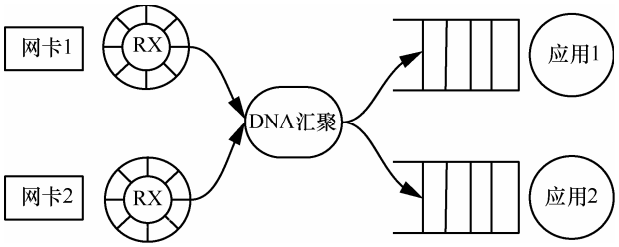

图 3　DNA 汇聚

5　分组分类

分组分类线程根据应急采集规则和动态分组分类算法对网络流量进行筛选和分类，将与应急采集规则匹配的分组写入到相应共享缓冲区中，否则丢弃分组。

5.1　应急采集规则

在网络安全应急响应中，为了能够根据通信对象或通信特征进行应急采集，分组采集规则内容如表 1 所示，包括 IP、协议、端口和时间等内容，源 IP 和宿 IP 为某个 IP 或某个 IP 地址段。源宿 IP 关系指源 IP 和宿 IP 之间“且”、“或”关系，例如，采集 58.192.112.0/20 的所有流量，则配置源宿 IP 均为 58.192.112.0/20，源宿 IP 关系为“或”。协议为运输层的协议编号，例如 TCP 协议编号为 6。源端口和宿端口为源宿端口号。源宿端口关系，为源端口与宿端口的“且”、“或”关系。与入侵检测规则不同的是，应急采集规则加入了应急采集的开始时间和结束时间。因为系统时间只有在开始时间和结束时间之间时，此规则才进行应急采集，所以应急采集规则存在着状态变化。

表 1　　应急采集规则内容

采集规则内容	说明
源 IP	源 IP 或源 IP 地址段
宿 IP	宿 IP 或宿 IP 地址段
源宿 IP 关系	源宿 IP“且”、“或”关系
协议	运输层协议
源端口	源端口号
宿端口	宿端口号
源宿端口关系	源宿端口“且”、“或”关系
采集开始时间	单位为 s
采集结束时间	单位为 s

在进行应急采集时，有时需要人为干预和控制应急采集过程，包括对未开始的应急采集任务人为强制开始，或对正在进行的应急采集任务强制结束。为了区分应急采集规则的状态变化和人为干预应急采集过程，对应急采集规则设置了不同的状态，应急采集规则的状态（如图 4 所示）包括开始、准备、强制采集、采集、强制结束、结束，每种状态代表不同的含义，并且实时更新状态。

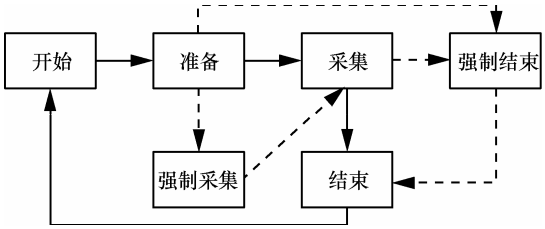

图 4　采集规则状态自动机

应急采集规则的开始状态为起始状态，应急采集规则初始化之后，则由开始状态转变为准备状态。若系统时间大于应急采集规则的开始时间，则由准备状态转变为采集状态，只有处于采集状态的规则才参与分组分类。若系统时间大于应急采集规则的结束时间，则由采集状态转变为结束状态。采集规则在结束状态时，则对相关数据清零和资源关闭，转变为开始状态，等待下一条采集规则的初始化。特别指出的是，规则的状态处于准备状态时，若设置了强制采集状态，规则状态转变为采集状态；同理，若设置强制结束状态，规则状态从准备或采集状态转变为结束状态，实现人为对应急采集过程的干预和控制。

5.2　动态分组分类算法和多线程优化

由于应急采集规则的状态随时间和人为干预不断变化，处于采集状态的应急采集规则也需要实时更新变化，考虑到应急采集规则数目相对较少（一般应急采集规则数目为 1~5）、规则内容逻辑复杂、规则更新较频繁，因此使用了 Linear 分组分类算法[8]。

分组分类线程每采集到一个分组，只与应急采集规则数组中处于采集状态的规则进行匹配，若匹配则将分组写入到与此规则对应的共享缓冲区中，若与所有处于采集状态的规则都不匹配，则丢弃分组。

为了充分利用现有系统 CPU 多核（16 个物理核，32 个逻辑核）的优势和提高线程的并行处理能力，对分组分类线程进行如下优化处理：1）将多线程绑定在不同的物理核上，避免了多线程在同一核上的切换所带来的性能压力；2）提高分组分类线程的优先级，确保线程在核竞争中处于优势，增加线程在核上运行时间和几率。

6　分组存储

PF_RING DNA 使用可定制的负载均衡函数将分组分发给各个分组分类线程。常用的负载均衡算法根据分组 MAC 地址、四元组（源宿 IP、源宿端口）或五元组（源宿 IP、源宿端口、协议）的散列结果来实现。不失一般性，同一个 IP、端口或协议的分组可能分散在不同的分组分类线程中，同时有的分组可能匹配多个应急采集规则，所以需要考虑线程间的数据同步问题，采用基于缓冲区分片算法[2]，对共享缓冲区进行分片划分，确保每个分片只有一个分组分类写线程和一个分组存储读线程，具体算法如下。

1) 对申请的共享缓冲区按图 5 所示进行划分，平均分配给各个分组分类线程。若共享缓冲区大小为 *total* 个单位，分组分类线程数为 *n*，则每个分组分类线程所分配的空间为 *total*/*n* 个单位。

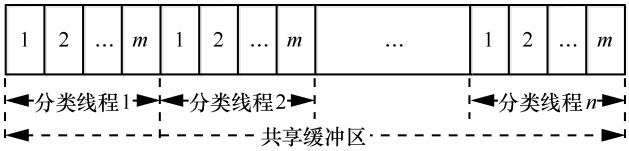

图 5　共享缓冲区分片

2) 对每个分组分类线程分配的空间进行划分，平均分配给各个应急采集规则，作为存储与应急采集规则匹配的分组队列，若最大采集规则数为 *m*，对于每个分组分类线程来说，每个应急采集规则对应的共享缓冲区大小为 *total*/*n*/*m* 个单位。

分组分类线程和分组存储线程对共享缓冲区的操作如图 2 所示，对于每个分类线程来说，在其分配的共享缓冲区中，不同采集规则匹配的分组存储在共享缓存区的不同区域中，并且每个区域都维持着一个写指针和一个读指针，读写指针每到队列末尾时的下一个位置为队列的开头，组成一个环形队列。

分组存储线程随应急采集规则的采集状态而创建，随应急采集规则的结束状态而销毁，每个分组分类线程都对应一个应急采集规则，将匹配此规则的所有分组保存在 PCAP 文件中。由于与同一规则匹配的分组分散在不同的分组分类线程中，而不同的分组分类线程将分组写入到不同的环形队列中，但是每个队列中的分组在时间上是有序的，分组存储线程将多个有序的队列合并成为一个有序的队列。若所读取的队列为空或不存在分组，则此队列不参与分组时间排序。排序后的分组为了区分所对应的应急采集规则，不同应急采集规则匹配的分组存放在不同的目录下，并且生成多个文件大小可配置的 PCAP 文件，所以与同一应急采集规则匹配的分组存储在同一目录下的不同文件中，便于 PCAP 文件分析和管理。

7 性能测试

系统的配置环境如表 2 所示，测试的网络流量来自 CERNET 南京节点边界。

表 2 系统环境配置

配置	规格
CPU	Intel(R) Xeon(R) CPU E5-2650 0 @ 2.00 GHz
内存	DDR3 32 G
操作系统	CentOS release 6.5 (Final) Linux fire 2.6.32-431.el6.x86_64
网卡类型	Intel(R) 10 Gigabit Network Connection
网卡驱动	ixgbe 3.10.16-DNA
分组捕获工具	PF_RING DNA v.5.6.1
磁盘	SCSI 299 GB

7.1 分组捕获性能测试

为了测试 PF_RING DNA 对 2 个万兆网卡的分组捕获能力，图 6 是在 19.98 Gbit/s（3.5Mpacket/s）流量和分组分类线程采集分组后立即丢弃的情况下，测试分组分类并发线程数与分组捕获分组丢失率的关系。从图中可以看出，分组分类并发线程数在大于 2 个时，PF_RING DNA 能够在不分组丢失的情况下捕获 2 个万兆网卡上的所有流量，表明在使用分组捕获工具 PF_RING DNA 和 2 个万兆网卡的情况下即可捕获 CERNET 南京节点边界的所有流量（17 Gbit/s，3.17 Mpacket/s）。

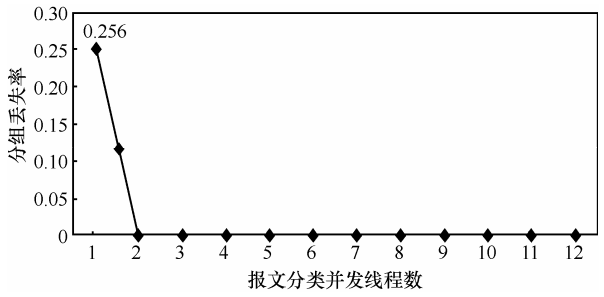

图 6 分组分类并发线程数与分组丢失率关系

7.2 分组分类性能测试

为了测试分组分类线程的分组分类性能，选择了常用端口 80 作为应急采集规则的测试用例，分组分类线程进行分组分类后，立即将所有的分组丢弃。在 19.98 Gbit/s（3.5Mpacket/s）流量下，配置 4 个分组分类线程，通过配置不同数目的测试用例来测试分组分类线程总体 CPU 的变化（如图 7 所示），从图中可以看出，随着应急采集规则数目的增加，分组分类线程的总体 CPU 也随之增加，在配置 12 个应急采集规则时，分组分类线程的总体 CPU 利用率为 294%，平均每个分组分类线程的 CPU 为 73.5%，能够满足一般应急采集 1~5 个规则的需求。

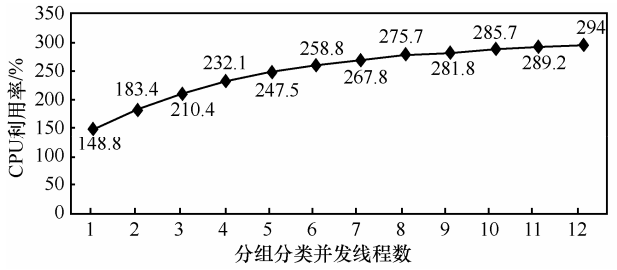

图 7 分组分类并发线程数与 CPU 利用率关系

7.3 分组存储性能测试

为了测试分组存储性能，选取流量大小约为 107.3 Mbit/s（34.4 kpacket/s）的端口 443 作为应急采集规则的测试用例。

在 19.98 Gbit/s（3.5 Mpacket/s）流量下，配置 4 个分组分类线程，通过配置不同数目的测试用例来测试应急采集系统内存利用率的变化，也即分组共享缓冲区利用率的变化，如图 8 所示，应急采集系统的内存利用率随应急采集规则数的增加而增加，在配置 9 个应急采集规则时，偶尔出现因分组缓冲区过满而分组丢失的现象，系统内存使用率为 48.1%时才出现分组丢失，提高了系统的分组存储性能。所以，在配置 8 个应急采集规则时的应急采集速率即为最大分组存储速率约为 858.4 Mbit/s（275.2 kpacket/s）。

7.4 系统总体性能测试

为了测试应急采集系统的总体性能，系统配置了 2 个万兆网卡、4 个分组分类线程，在最大的网络流量 19.98 Gbit/s(3.5 Mpacket/s)下，选取流量大小约为 129.7 Mbit/s（20.49 kpacket/s）的端口 8080 作为应急采集规则的测试用例，通过配置不同数目的测试用例来测试应急采集系统的 CPU 利用率（如图 9 所示）和内存利用率（如图 10 所示）。

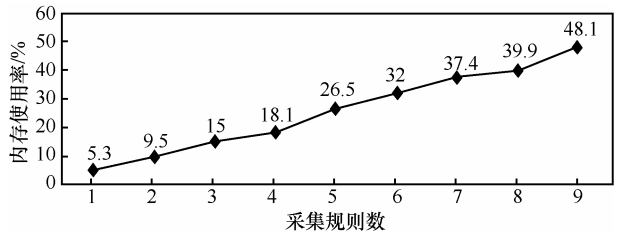

图 8 采集规则数与内存使用率关系

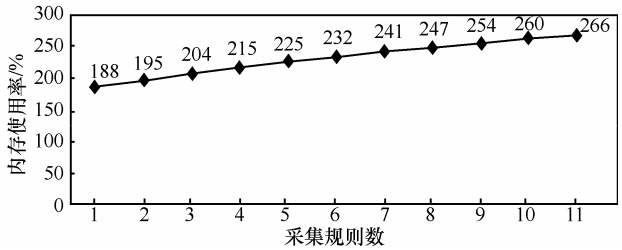

图 9　采集规则数与 CPU 利用率关系

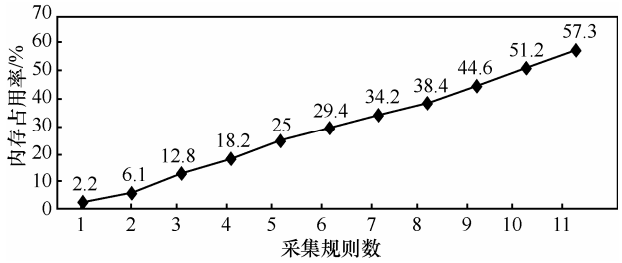

图 10　采集规则数与内存占用率关系

在动态配置第 11 个测试用例时，出现了因环形共享缓冲区队列过满而出现分组写入失败（分组丢失）的现象，所以应急采集系统的最大应急采集速率为配置 10 个测试用例时的应急采集速率，约为 1 297 Mbit/s（204.9 kpacket/s）。

出现分组丢失的原因是分组存储线程未能及时读取并处理分组，而分组存储线程的 CPU 利用率只有 3.0%。通过查看系统 IO 状况，发现磁盘每秒 IO 操作值有时达到 100%，说明系统的磁盘请求过多，不能及时将分组写入到磁盘中，因此磁盘的分组写入性能为系统的瓶颈。

8　结束语

通过使用分组采集工具 PF_RING DNA 捕获 2 个万兆网卡上的分组、多个分组分类线程并发分类筛选分组以及分配分组共享缓冲区提高分组存储性能，同时，为应急采集规则赋予不同的状态，应急采集系统能够在分组丢失的情况下采集并处理 19.98 Gbit/s(3.5 Mpacket/s)大小的流量，最大应急采集速率为 1 297 Mbit/s(204.9 kpacket/s)，并能够动态添加应急采集规则和人为控制应急采集过程，同时，对应急采集到的分组进行分类存储和文件切分，方便分组分析和管理。所以，应急采集系统能在零分组丢失的情况下对 CERNET 南京节点边界的所有流量进行应急采集，为应急响应提供重要的分组信息。

参考文献：

[1] 孙成峰.面向万兆网络的滥用入侵检测系统改进[D]. 东南大学, 2013.
SUN C F. The Improvement of Misuse Intrusion Detection System in 10 Gbps Ethernet[D]. Southeast University, 2013.

[2] 吕少阳. CHAIRS 系统运行管理与离线检测的设计与实现[D]. 东南大学, 2013.
LV S Y. The Research and Implementation of Operation Management System and Offline Detection System of CHAIRS[D]. Southeast University, 2013.

[3] 林洪周.万兆网络数据包捕获系统的研究与开发[D].华中科技大学, 2008.
LIN H Z. The Research and Development of 10 Gbps Network Packet Capture System[D]. Huazhong University of Science & Technology, 2008.

[4] 张显,黎文伟.基于多核平台的数据包捕获方法性能评估[J].计算机应用研究, 2011.
ZHANG X, LI W W. Performance evaluation of packet capture methods based on multi-core platform[J]. Application Research of Computer, 2011,28(7).

[5] Packet Capture Performance at 10 Gbit: PF_RING vs TNAPI[EB/OL]. http://www.ntop.org/pf_ring, 2014.8.

[6] 钟婷,刘勇,耿技.基于 IXP2400 网络处理器的高速包过滤的研究[J]. 计算机应用, 2005,25(11).
ZHONG T, LIU Y, GENG J. Study on fast packet filter under network processor IXP2400[J]. Computer Application, 2005,25(11).

[7] CAMPBELL S, MELLANDER J. Experiences with intrusion detection in high performance computing[J]. CUG, 2011.

[8] 王韬.高速网络环境下的报文监测[D].东南大学, 2004.
WANG T. Packet Sensor on High Speed Network[D]. Southeast University, 2004.

作者简介：

马亚洲（1990-），男，河南周口人，东南大学硕士生，主要研究方向为网络入侵检测。

龚俭（1957-），男，上海人，东南大学教授、博士生导师，主要研究方向为网络安全、网络行为、网络体系结构。

杨望（1979-），男，安徽宣城人，东南大学讲师，主要研究方向为网络安全、网络管理。