第 33 卷 第 1 期 计 算 机 工 程 2007 年 1 月
Vol.33 No.1 Computer Engineering January 2007
·安全技术· 文章编号：1000—3428(2007)01—0176—02 文献标识码：A 中图分类号：TP393

基于 Linux 内核扩展模块的 P2P 流量控制

陈海军[1]，李仁发[2]，杨 磊[2]
（1. 湖南大学软件学院，长沙 410082；2. 湖南大学计算机与通信学院，长沙 410082）

摘 要：分析了 Linux Netfilter/Iptables 架构的实现机制和扩展技术，分析了 P2P 协议的特征，通过扩展 Linux 内核库，利用共享库实现用户数据空间与内核空间的数据交互，扩展防火墙的规则集，从而实现 P2P 流量控制的方法，而且可以根据不断出现的 P2P 业务更新规则集，具有很好的扩充性能。

关键词：防火墙；内核空间；用户空间；扩展

P2P Flow Control Based on Linux Kernel Expanded Model

CHEN Haijun[1], LI Renfa[2], YANG Lei[2]
(1. Software School, Hunan University, Changsha 410082; 2. School of Computer and Communication, Hunan University, Changsha 410082)

【Abstract】 After analyzing the mechanism and the expansion of technology for Linux Netfilter/Iptables, the characteristics of the P2P protocol, through expanding kernel lib based on the Linux system, this paper implements user data space and kernel space data interchange through sharing lib, expands firewall rule, implements P2P flow control, and according to emerging P2P business to update the rule set, with the good performance of expansion.

【Key words】 Firewall; Kernel space; User space; Expansion

P2P 改变了互联网以大网站为中心的状态，重返“非中心化”，使得网络资源共享变得更加容易、直接。在 P2P 应用逐渐普及的过程中，传统网络设备处理 P2P 带宽问题，逐渐暴露出以下缺点：

(1)阻塞 P2P 常用端口。此方法一方面有可能拒绝正常通信，另一方面导致 P2P 应用转向使用随机端口和专用端口躲避检查。

(2)阻塞 P2P 对等体向 P2P 信息服务节点的通信。这种方法导致了 P2P 对等体使用代理服务器的方法躲避检测，也导致 P2P 信息服务节点向随机分布和隐藏的方向发展。

(3)限制用户的上行带宽。这种方法违反了服务条约，而且导致了向公网用户的数据请求量增大。

长时间高度拥塞的网络也带来网络管理困难和功能失效的危险，此外各种对实时性要求较高的服务如 VoIP、Streaming Video 和 Audio 也遭遇了前所未有的不确定的网络运行环境。

Linux 内核中已经提供了对 P2P 流量感知模块 IPP2P，可以结合 IPTables 防火墙以及 tc 带宽限制功能来识别、控制 BT、EDonkey 等 P2P 协议流量。

本文将介绍基于 Linux Netfilter/IPTables 架构实现机制和扩展技术，在此基础上提出了扩展匹配选项实现防火墙的 P2P 流量控制功能，扩充后的 IPP2P 可以根据 P2P 协议特征进行流量控制，并且可将其转化为防火墙规则实现规则集的有效扩充。

1 Linux 防火墙的扩展 Netfilter/IPtables 的技术

Linux 中防火墙 Netfilter/IPTables 系统主要包括两个基本组件：定义在内核空间(Kernel Space)中的通用框架 Netfilter，即：包分割框架，数据包选择系统(PacketSelection)。其中后者又由两部分构成：在 Netfilter 框架上定义的数据结构“IP 表(IPtables)”、“链(chains)”和“规则(rule)”，在用户空间实现的应用程序 IPtables 用于插入、修改和删除信息包过滤表中的规则。具体防火墙工作流程见文献[2]。

由于 Netfilter 架构的加入，可以通过简单的内核模块化来实现新功能的扩展，在现有的 (Netfilter/ IPTables)中可以通过两种方式对现有的防火墙进行扩充：(1)扩展 Netfilter 通过编写相关内核模块调用 nf_register_hook ()直接在相关的钩子上注册从而获得新特性；(2)扩展 IP 表通过编写相关的匹配标准和目标来实现新特性，扩展 IP 表方式是对现有表的匹配规则的扩充与具体表无关。扩展 IP 表需要编写内核和用户两方的代码，内核模块提供了实际的数据包匹配规则代码，用户方代码提供了 IPTables 新的命令行选项的共享库。

2 Linux 防火墙 P2P 流量控制扩展匹配设计

新的匹配函数常常被当成一个独立的模块来写。用这些模块具有可扩展性，有两种方法可以实现：(1)用 Netfilter 框架的“nf_register_sockopt”函数让用户直接和你的扩展模块对话；(2)输出新模块的标记来让别的模块注册它们自己，Netfilter 和 IPTables 都是这么做的。

这种两种方式简单而有效，可以借鉴这种思想在防火墙的匹配选项中加入匹配选项来检测数据包中的内容，由于扩展 IP 表具有很好的灵活性，为此可以选用这种方式扩充匹配标准来实现扩展 P2P 流量感知模块。这种方式需要编写内核和用户空间代码，Netfilter/IPTables 的标准化提供了两种使用的重要数据结构，在实现这两部分代码时主要是填充相应的

基金项目：湖南省自然科学基金资助项目(05JJ40118)；湖南省教育厅基金资助项目(04C313)
作者简介：陈海军(1974－)，男，工程师、硕士，主研方向：计算机网络管理及网络安全；李仁发，教授；杨 磊，博士生
收稿日期：2006-08-30 **E-mail：**chj_yh@hnbc.com.cn

数据结构内容，然后将它们注册，从而扩展功能。

2.1 内核模块数据结构

新的 MATCH 功能可作为一个独立的模块，为了能使新模块被别的模块使用，可以使用 IPTables 提供的 ipt-register-match()将该模块进行注册，新的 MATCH 模块核心是 ipt-match 结构，它将作为 ipt-register-match()的参数注册到 MATCH 链表中备用，用来注册一个新的匹配类型，从而增加新的规则匹配选项。

构造新的匹配函数，核心是 struct ipt_match 结构，它传递参数给“ ipt_register_match() ”，这个结构有下面这些成员：

(1)struct list_head list

这个成员常设置为空“ {NULL,NULL} ”，由核心使用。

(2)const char name[]

这个成员是匹配函数的名称，被用户来指定。为了自动加载工作，这个名称必须和模块的名称相匹配。

(3)int (*match)()

一个指向 MTACH 功能函数的指针，返回非 0 表示匹配

(4)int (*check entry)()

该成员是一个指向一个函数的指针，该函数为一个控制检查其规范；如果它返回 0，这个控制就不被用户接受。

(5)iptables void (*destroy)()

该成员是一个指向函数的指针，当一个使用该匹配的入口被注销的时候就调用该函数。

这样就允许在用 checkentry 时能动态地分配资源并且在这里把它们完全地清理掉。

(6)Struct module me

该成员被标志为“ THIS_MODULES ”，它给出一个指向你的模块的指针。当该种类型的控制被创建后销毁，它使得计数器呈现出增长或者降低。这能防止某个用户在控制指向某个模块的时候移除该模块(并且因此 cleanup_module()被调用)。

2.2 用户空间数据结构

在内核中加入相关的内核模块后，为了在用户空间使用 IPTables 提供相关的规则，必须为该应用提供相关的命令行选项，为了使各个扩展模块使用一个版本的 IPTables 而不必编写相关扩展的特定软件版本，采用共享库可以解决该问题。共享库应该具有_init()函数，它的功能和内核模块功能相似，在装载时被自动调用，根据共享库是提供一个新的匹配或者一个新的对象，分别调用 register-match()或 register-target，共享库可以提供初始化数据结构和提供相关选项的功能。编写共享库中使用的重要数据结构是 iptables_match，它作为参数传递给 register-match () 注册相关的命令行匹配选项让 IPTables 识别该新匹配。代码说明如下：

```
Struct iptables_match
{structiptables_match *next
/*用于形成一个 MATCH 列表的指针，初始化为 NULL*/
ipt_chainlabel name
/*MATCH 功能的名字，必须与库函数名相同便于主程序根据 MATCH 名加载相应的动态连接库*/
const char *version
/*版本信息通常被设置 IPTABLES_ver-sion 宏 size_t size；该 MATCH 的数据大小为 size_t userspacesize；由于内核可能修改某些域，在这里填写被改变数据区大小，它一般和 size 大小相同*/
void *init (); //初始化 ipt-entry-match 结构
int parse();
//扫描并接收本 MATCH 的命令行参数，正确接受返回非 0
void *final_check ()
//检查是否强制选项，如 P2P 描述，如果不正确退出
void *print ()
/*查询当前表中的规则时，显示使用了当前 match 规则的额外信息*/
void (*save) ()
//PARSE 的反转，被 iptable-save 调用再生 match 的命令行参数
const struct option *extra_opts
//NULL 结尾的参数列表，提供命令行其余选项
//以下参数由 iptables 内部使用，用户不必填写。
Unsigned int option_offset;
Struct ipt_entry_match *m;
Unsigned int mflags;
Unsigned int used;
}
```

2.3 P2P协议数据识别

Netfilter 的 pom 中提供了 ipp2p 匹配，ipt_ipp2p.c 中定义了很多 P2P 软件的数据的判断模式，可以对其进行有效的扩展，便可以实现不同的 P2P 协议流量控制的需要。

(1)eMule/eDonkey/Kad

这几个协议用二进制数进行协商，见表 1。

表 1 P2P 协议特征

第 1 字节	第 2 字节	第 3 字节	其他字节	UDP 长度	类型
0xe3	0x9a	any	any	26	edonkey
0xc5	0x91	!0	any	12	emule
0xc5	0x90	!0	any	26	emule
0xc5	0x92	any	any	10	emule
0xc5	0x93	any	any	10	emule
0xe4	0x50	any	any	12	kad
0xe4	0x58	!0	any	14	kad
0xe4	0x59	any	any	10	kad
…	…	…	…	…	…

(2)Gnutella

明文检查起始数据是否为“ GNUTELLA ”或“ GND ”。

(3)KaZaA

UDP 数据部分的结尾 6 个字节是: “ KaZaA\0 ”。

(4)BitTorrent

UDP 长度 24 字节(含 UDP 头)，起始 8 个字节为：00 00 04 17 27 10 19 80。

(5)SoulSeek

前 8 个字节格式为: xx xx 00 00 yy zz 00 00, 其中 xx xx 为 16 位负载长度-4，yy!=0, zz 任意或者数据长度 8 字节，全 0 或者数据格式为：01 xx 00 00 00 yy .. zz 00 00 00 .., 其中负载长度大于 xx+6，负载第 xx+4+1 字节(zz)不为 0，而负载第 xx+5+1 字节，第 xx+6+1 字节为 0。

(6)BitTorrent 负载第 1 字节为 0x13, 而且后续数据为：“ BitTorrent protocol ”。表示的长度大于负载长度减 5，而且第 4、第 5 字节为 0，第 6 字节为 0x01 或 0x4c；如果第 2、第 3 字节等表示的长度小于负载长度减 5，则负载偏移该长度字节后的第 6 字节为 0xe3/0xc5。

2.4 编写新的Netfilter模块

为了在内核里接收/割裂(mangle)信息包，关键在特殊的协议和特殊的协议中 netfilter 头文件中的定义，如“ netfilter_ipv4.h ”，用函数“ nf_register_hook ”和“ nf_unregister_hook ”注册和注销 netfilter 钩子，如下写一个注册到“ netfilter 钩子 ”的模块。

(1)list：用来加入到链表中：设置为“ { NULL, NULL }”

(2)hook：当一个信息包找到这个钩子的点(point)这个函数就被调用。函数必须返回 NF_ACCEPT, NF_DROP or NF_QUEUE。如果是 NF_ACCEPT，下一个钩子到达那个点的时候将会被调用。如果是 NF_DROP，信息包被丢弃。如果是 NF_QUEUE，它是队列。

(下转第 183 页)