基于主机的P2P流量检测与控制方案

吴 敏[1],王汝传[1,2]
(1.南京邮电大学 计算机学院,江苏 南京 210003;
2.南京大学 计算机软件新技术国家重点实验室,江苏 南京 210093)

摘 要:P2P流量逐渐成为了互联网流量的重要组成部分,在对Internet起巨大推动作用的同时,也带来了因资源过度占用而引起的网络拥塞以及安全隐患等问题,妨碍了正常的网络业务的开展。由于P2P采用了动态端口等技术,使得传统的端口映射的方法对P2P流量的识别不再有效,如何有效地监测和控制P2P流量是网络测量领域一个重要的研究课题。介绍了各种P2P流量识别方法及优缺点,然后提出一种新型的检测与控制策略——基于主机的P2P流量检测与控制。实验结果及分析表明该方法能比较有效地检测P2P流量并具有更好的检测精度和控制效果。

关键词:P2P;流量检测;流量控制
中图分类号:TP393 **文献标识码**:A **文章编号**:1673-629X(2009)10-0026-04

A Host-Based P2P-Traffic Detection and Control Scheme

WU Min[1], WANG Ru-chuan[1,2]
(1. College of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210093, China)

Abstract: P2P traffic has taken great portions in the network traffic. While having a significant impact on the Internet, it brings serious problems such as network congestion and traffic hindrance caused by the excessive occupation in the bandwidth. Since more and more P2P applications are using dynamic random port numbers, the traditional port matching technology is becoming useless, therefore how to effectively identify and control the P2P applications has been a very important research work. Introduces methods in identifying P2P traffic and their characters, then puts forwards a host-based strategy of detection and control of P2P traffic. Experimental results show that the method is efficient for traffic identification and has a more accurate pricision and control effects.

Key words: P2P; traffic identification; traffic control

0 引 言

随着P2P网络技术[1]在20世纪90年代后期的兴起,P2P流量逐渐成为了互联网流量的重要组成部分。精确地识别P2P流量对于有效地管理网络和合理地利用网络资源都具有重要意义[2,3]。

目前P2P流量检测技术大致有以下三类:基于端口的检测技术,深层数据包检侧技术和基于流量特征的检测技术。

基于端口的分析方法是在网络流量中探测P2P用户最基本、最直接的方法。但由于现在大多数P2P应用允许用户手动选择随意的端口号来设置默认的端口号或使用随机的端口号,从而使得端口号不可预测,还有一些P2P应用使用默认端口号(例如80端口)来伪装自己的功能端口,因此基于端口号的分析方法的效率变得很差。

深层数据包检测技术[4~6],通过深入检测其数据包中的有效载荷来进行检测,即通过应用层数据包的正则表达式的匹配来完成探测工作,以确定特定的P2P应用。当前许多商业P2P应用识别方案基于该方案,如L7-filter,Casco's和PDMLk等。该方法识别准确度可以达到95%,实现简单,维护方便。但该方法是高资源消耗的,在带宽越高的网络,检查时候所需要的开销和资源就越多,而且由于必须读取处理所有

收稿日期:2009-01-31;**修回日期**:2009-05-02
基金项目:国家自然科学基金(60573141,60773041);国家高科技863项目(2007AA01Z404,2007AA01Z478);江苏省自然科学基金(BK2008451);南京市高科技项目(2007软资127);现代通信国家重点实验室基金(9140C1105040805);江苏高校科技创新计划项目(CX08B-085Z,CX08B-086Z)
作者简介:吴 敏(1976-),女,江苏泰州人,讲师,博士研究生,研究方向为P2P技术、分布式计算、计算机密码学和网格计算等;王汝传,教授,博士生导师,研究方向为计算机软件、计算机网络和网格、对等计算、信息安全、无线传感器网络、移动代理和虚拟现实技术等。

网络流量,会严重地增加网络设备负担甚至会导致网络的崩溃,因而不适合大型网络。另外该方法对加密 P2P 流量捕获能力弱,对新的 P2P 应用必须升级后才能检测且该方法容易和隐私保护法律条款产生冲突。照检测技术发展的基于流量特征的检测技术[7,8]是利用 P2P 在传输层表现出来的流量特征来发现 P2P 应用。这类方法借用了统计学领域通用的一些概念,分析传输层的信息,不需要任何关于应用层协议的信息,几乎不需要任何额外的软件或者硬件并具有较强的加密和未知 P2P 流量的捕获能力,因而近年来关于流统计方式测量 P2P 流量得到了国内外广泛的关注,被认为是最有前途的一种方法。

目前主要包括以下几种识别方式:

{IP,port}识别、TCP/UDP 端口识别、Block Size 识别、基于会话(session)分类的识别、双向识别、流统计状态的识别等等。

实际上,从模式识别的角度而言,P2P 流量的识别过程看作是一个二分类问题:即对流量数据进行分类,分为 P2P 流和非 P2P 流。文中采用该思想提出一种基于主机协议特征的 P2P 流量识别与控制的方法。

1　基于主机的 P2P 流量识别方案

1.1　系统架构

在该方案中,可以考虑将系统分客户端和服务器两个部分(如图 1 所示)。其中客户端负责流量检测、向服务器发送流量信息以及控制策略的具体执行。服务器端负责 P2P 流量分析和策略管理。

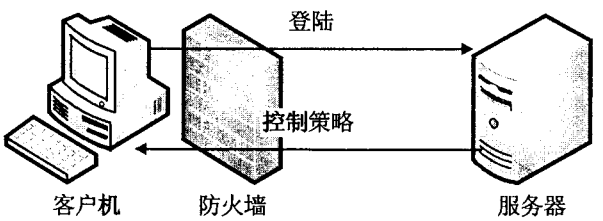

图 1　软件模型

该方案的设计思路是客户机向服务器发送流量信息,服务器向客户机发送控制策略和协议规则。NDIS 驱动根据获得的 P2P 协议特征字,匹配所有经过本机的数据包的包头,判断其是否为 P2P 流量;并且根据控制策略对 P2P 流量进行相应的控制。

整个流程如下:

(1)先启动服务器。

(2)客户机登入服务器,并从服务器端获取最新版本的协议规则。

(3)客户机将获得到的协议规则传给 NDIS 驱动。驱动获得规则后,开始启动控制。

(4)客户机定时地从驱动中获得本机的流量信息,并将它实时汇报给服务器。

(5)管理员通过观察服务器端上显示的各个客户机的流量信息,做出控制策略,并将策略发送给客户机(策略也包括取消控制)。

(6)客户机获得控制策略后,立即将它转发给驱动。驱动获得策略后就开始根据情况拦截 P2P 数据包。

1.2　P2P 流量识别方式

所有的网络应用程序都是基于 TCP/IP 协议的框架才能实现通信,由 TCP/IP 的体系结构可知,机器上的每一层都假设它在直接与另一台机器的同一层"交谈"。因此,每一层服务都有自己独特的协议来实现通信。应用层也不例外,不过应用层的协议是有应用程序开发者规定的,用于区别其他应用协议。因此,可以通过该特征来唯一地识别数据包的所属协议,即可以基于应用层协议特征识别 P2P 流量。下文以 PPLIVE 为例说明该识别过程。

PPLIVE 的传输层为 UDP 时,它的应用层协议特征字是:

BYTE1:=11101001

BYTE2:=00000011

从网卡上获得一个数据包后,层层剥掉其包头,对其进行分析。如果分析得出该数据包的传输层协议是用 UDP 时,则继续匹配该数据包的应用层数据。图 2 显示了对数据包进行层层分析的过程。

若应用层数据的前两个字节为 0XE9 和 0X03,则该数据包是由 PPLIVE 应用程序收发的,即它属于 P2P 流量;否则,不属于 P2P 流量。

由于要对原始数据包进行分析,所以该识别策略应由网卡驱动来实现。对于程序开发者来说,只要在 NDIS 的中间层驱动中添加一些必要的代码,向底层的驱动添加协议特征即可。如识别 BT 软件,则添加识别 BT 特征字的代码;如识别 eMULE 软件,则添加识别 eMULE 特征字的代码。

2　基于协议特征的 P2P 流量控制方案

完成 P2P 流量的识别后,可根据网络实际运行情况进行流量控制,主要有两种流量控制策略:带宽控制和流量控制。

2.1　带宽控制

带宽控制,就是每秒通过主机的 P2P 流量不能超过事先规定好的一个值。

具体的实现分析如下:

首先定义三个变量 T_0,B_0, Band:

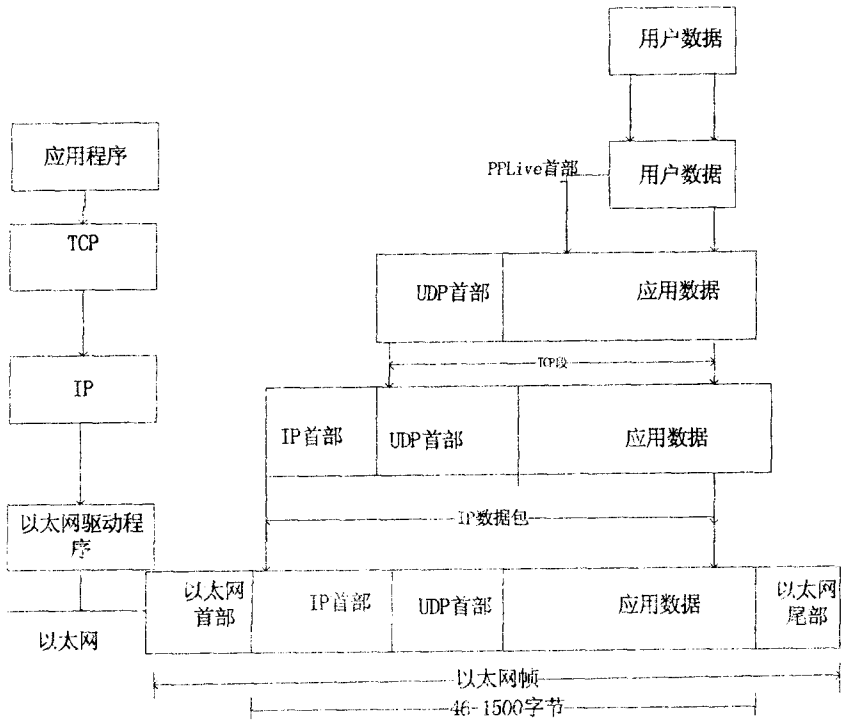

图 2 数据包分析过程

T_0 - 用来计算时间差

B_0 - 用来统计流量

Band - 要限制的带宽

每当来一个 P2P 数据包时，就获取一次系统时间 T_{now}。然后 T_{now} 和 T_0 相减获得时间差 T。

若 $T \leq 1$s，则判断 B_0 + sizeof(packet) 是否大于 Band。若大，则丢包，否则，放行并将数据包的大小加到 B_0；

若 $T \geq 1$s，则数据包放行。并且将 T_{now} 赋给 T_0，重新计时；B_0 置为 sizeof(packet)，重新计数。

通过这种方式可以实现带宽控制，从而保证每秒的流量不会超过规定的数值。

2.2 流量控制

流量控制，就是从控制开始，通过主机的 P2P 总流量不能超过规定的值。

具体的实现分析如下：

首先定义两个变量：

$Traffic_0$ - 统计 P2P 流量

Traffic - 要控制的流量值

每当来一个 P2P 数据包时，就判断 $Traffic_0$ + sizeof(packet) 是否大于 Traffic。若大，则丢掉该数据包；否则放行，并将 $Traffic_0$ 增加 sizeof(packet)。

3 实验结果分析

为了验证提出的基于协议特征的 P2P 流量检测与控制的可行性和准确性，设计了一个基于主机的 P2P 流量检测与控制原型系统，其软件模型如图 1 所示。

服务器先发送协议特征给客户机，然后客户机将特征传给 NDIS 中间层驱动并启动 P2P 流量检测与控制。客户机实时地向服务器汇报本机的 P2P 流量信息。服务器可以针对客户机的流量信息，发送控制策略给客户机。客户机的 NDIS 驱动就根据该控制策略来实行控制。

实际运行检测结果如图 3～5 所示。

其中图 3 是表示未实现 P2P 流量控制之前客户机显示的流量信息。

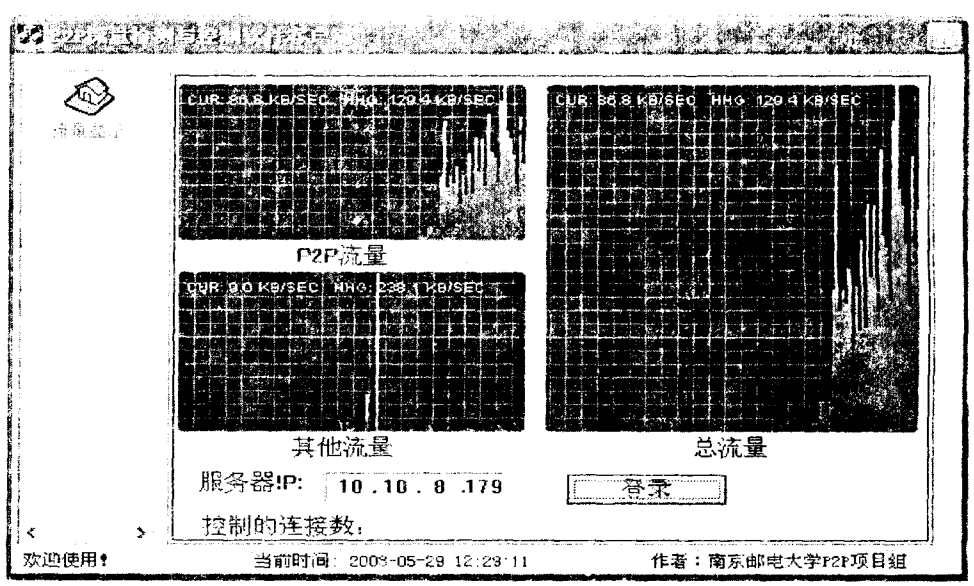

图 3 未控制前客户机流量信息

图 4 表示当服务器发送带宽控制 20kB/s 后，客户机的流量信息。

图 5 显示了当服务器发送流量控制 10M 后，客户机的流量信息。

实验中运行的 P2P 应用包括：BitTorrent、eMule、

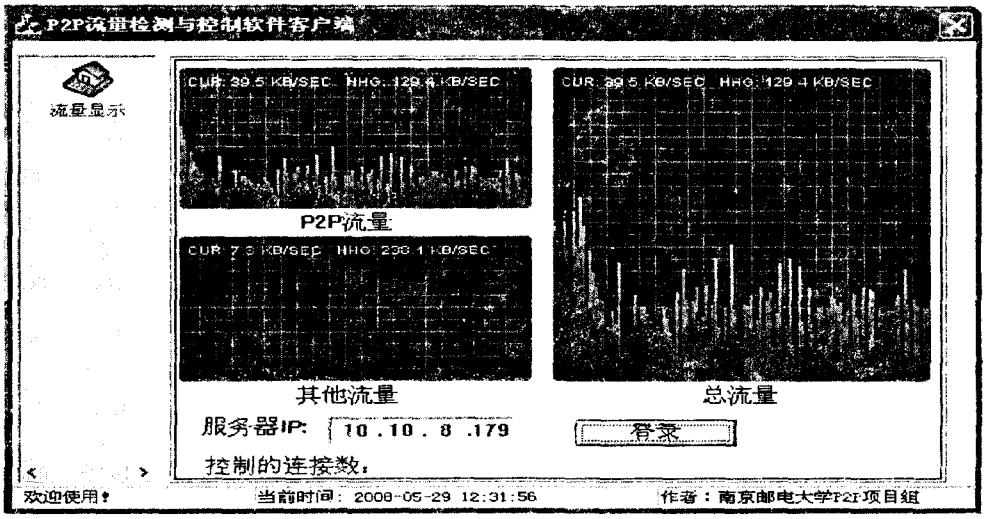

图 4　控制带宽后客户机流量信息

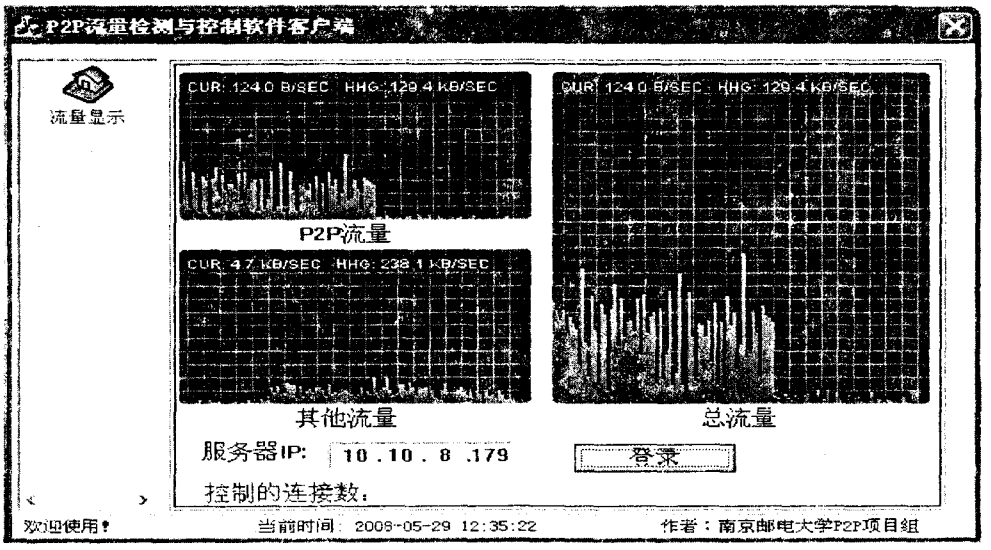

图 5　控制流量后客户机流量信息

迅雷、PPlive、Maze 及 Web 应用与 FTP 服务。从实验结果可以看出,该策略能够准确地识别出 P2P 流量并对其进行控制。

4　结束语

文中提出了基于协议特征的 P2P 流量检测与控制,并对其进行详细并系统的分析。同时为测试其可行性和准确性,还设计了一个测试系统来运行该算法。结果表明该算法是有效的并且有较高的准确性。

下一步的研究目标首先要完善该方案,如让数据发送端来控制速度,这样能够更有效地解决网络拥塞问题。然后是实现该方案和基于 P2P 流量特征的检测技术相融合,能够通过机器学习的方法提取 P2P 流量的特征后通过基于特征字的方法进行 P2P 应用的分类。

参考文献:

[1] 吴国庆.对等网络技术研究[J].计算机技术与发展,2008,18(7):100－104.

[2] Bartlett G, Heidemann J, Papadopoulos C. Inherent Behaviors for On－line Detection of Peer－to－Peer File Sharing[C]//In Proceedings of 10th IEEE Global Internet Symposium (GI'07) in conjunction with IEEE INFOCOM 2007. An－chorage, AK, USA: [s.n.],2007:55－60.

[3] Pereira R L, Vazao T, Rodrigues R. Adaptive Search Radius－Lowering Internet P2P File－Sharing Traffic through Self－Restraint [C]//in The 6th IEEE International Symposium on Network Computing and Applications (IEEE NCA07). [s.l.]: [s.n.],2007:253－256.

[4] 蒋海明,张剑英,王青青,等.P2P 流量检测与分析[J].计算机技术与发展,2008,18(7):74－76.

[5] 乐艳辉,李之棠,柳　斌.基于 Netfilter 的 P2P 流量测量系统[J].计算机应用与研究,2008,25(4):1224－1226.

[6] 陈海军,王四春,叶　晖.Linux 内核扩展模块的 P2P 流量控制方法与研究[J].计算机工程与设计,2007,28(16):3912－3914.

[7] 宫　婧,孙知信,顾　强.基于行为特征描述的 P2P 流识别方法的研究[J].小型微型计算机系统,2007,28(1):48－53.

[8] Zander S, Nguyen T, Armitage G. Automated Traffic Classification and Application Identification using Machine Learning [C]//In:Proceedings of the IEEE Conference on Local Computer Networks (LCN 2005). Sydney, Australia: [s.n.], 2005:250－257.