

一种基于流量行为分析的 P2P 流媒体识别方法

宫 婧¹, 孙知信², 陈二运³

(1. 南京邮电大学 理学院, 江苏 南京 210003;

2. 南京邮电大学 计算机技术研究所, 江苏 南京 210003;

3. 南京中兴软创公司, 江苏 南京 210012)

摘 要: P2P 流媒体的识别问题已经成为 P2P 研究领域中的热点问题。以目前主流的 P2P 流媒体软件: PPLive、PPStream、QQLive 为蓝本, 进行流量行为特征分析, 总结出相应的流量特征和特征字符串, 提出一种基于流量行为分析的 P2P 流媒体识别方法。该方法的核心思想是: 首先通过流量的统计算法将捕获到的流媒体数据包进行分类统计, 再运用特征字匹配算法进行二次识别匹配, 从而得出结果。随后, 开发出基于流量行为分析的检测系统, 试验结果证明该方法能够有效地识别 P2P 流媒体。

关键词: P2P; 流媒体; 行为分析; 流量检测

中图分类号: TP393

文献标识码: A

文章编号: 1673-629X(2009)09-0128-04

A Kind of P2P Streaming Media Identification Method Based on Traffic Behavior Analysis

GONG Jing¹, SUN Zhi-xin², CHEN Er-yun³

(1. College of Mathematics & Physics of Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

2. Institute of Computer of Nanjing University of Posts and Telecommunications, Nanjing 210003, China;

3. ZTESOFT Technology Co., Ltd. Nanjing 210012, China)

Abstract: How to identify P2P streaming media has become a hot issue in the field of P2P. Taking several most popular P2P streaming applications like PPLive, PPStream, and QQLive as the main research objects, analyzes their traffic behavior and generalized related traffic profiles and payload signatures. Then a novel P2P streaming media identification method based on traffic behavior analysis is proposed. The core of this method is two aspects. First, using traffic statistics algorithm, classify and statistics captured streaming media packets. Then, using features of characteristic matching algorithm, can identify and matching these packets, so result can come out. Subsequently, a detection system adopting this traffic behavior analysis based method has been developed to evaluate it, and the experiment results show that the method proposed in this paper can identify P2P streaming media traffic effectively.

Key words: Peer-to-Peer; streaming media; behavior analysis; traffic detection

0 引 言

随着需求量以及用户人数的不断增加, P2P 流媒体(P2P Streaming Media, PMS)已在因特网中得到了广泛的应用^[1], 基于 P2P 流媒体的内容分发已经成为流媒体研究领域中热点问题^[2,3]。但是目前关于 PMS

的研究主要集中在 PMS 原型系统的设计和 PMS 内容分发网络的架构上^[4,5], 专门针对 PMS 流量特征以及 PMS 流量识别技术的研究, 还是处于起步阶段^[6]。

现有的流媒体发现技术主要是针对 Web 流媒体的, 也就是基于 C/S 模式的流媒体。它是通过对流媒体协议和文件特征的提取和分析, 提取行之有效的网上流媒体发现技术, 使之能够及时有效地发现网上存在的流媒体服务器、节目, 以及正在实施的流媒体访问活动。

基于现有的 Web 流媒体都采用固定的流媒体协议, 而这些协议都有对应的默认端口, 文献[7]最早提出通过端口扫描技术来发现连接到 Internet 上的流媒体服务器, 但是这种方法只能进行“粗粒度”的流媒体

收稿日期: 2009-01-05; 修回日期: 2009-03-09

基金项目: 国家自然科学基金(60572131); 科技型中小企业创新基金(08C26213200495); 江苏省科技攻关项目(BE2007058); 江苏省高校自然科学基金研究项目(08KJB520005); 南京邮电大学基金项目(NY206050)

作者简介: 宫 婧(1977-), 女, 江苏南京人, 讲师, 硕士研究生, 研究方向为计算机网络安全; 孙知信, 教授, 研究方向为计算机网络安全。

发现,具有很大的不可信性。基于搜索引擎的流媒体发现技术^[8,9]由信息采集、索引数据库和查询接口三部分组成,通过捕获、分析网络流媒体信息,对数据信息进行提取并组织建立索引库,用户通过索引库可以找到网络中对应的流媒体信息。IBM Almaden Research Center^[10]对流媒体页面链接的一些元信息进行分析,找出它们的相关性,采用基于页面向量的方法发现网络中的 Web 流媒体,并且正确率比较高。

虽然流媒体发现技术主要是针对 Web 流媒体的,而 Web 流媒体和 P2P 流媒体又有着较大的差别,但是它们都是流媒体,都有流媒体的一些特性,因此,现有的 Web 流媒体发现技术的研究将有助于实现对 PMS 流量的识别。

文中以目前主流的 P2P 流媒体软件:PPLive、PP-Stream、QQLive 为蓝本,进行流量行为特征和负载层的深入分析,总结出相应的流量特征和特征字符串,提出一种基于流量行为分析的 P2P 流媒体识别方法,并根据该特征开发出相应的检测系统。检测系统主要分为三个模块:网络流量抽样模块,流量特征提取、统计、过滤模块和流量特征字匹配分类模块。系统的核心是流量的统计算法和特征字匹配算法。统计算法首先对获取的特征包链表按照和本地 IP 的交互情况进行分类,然后分别统计每个本地 IP 的源端口数、目的端口数、目的 IP 数等基本流量信息,随之使用判定函数对该统计结果进行判定,得出一全局的哈希映射特征表。其次,根据特征表对特征包链表进行过滤,将过滤剩下的可疑的 P2P 流媒体特征包链表采用特征字匹配算法进行二次识别,并分类得出结果。最后,在局域网环境中对系统进行测试,测试的结果表明,该系统能达到很好的效果。

1 PMS 流量行为分析系统

下列的叙述将用到以下概念,现定义之:

定义 1 PMS 可信度:即某 IP 对之间存在 PMS 通信可能性的大小,值越大,可能性就越大。

将抽样得到的一组数据包(抽样的多少,根据网络的具体情况而定)或者已经获取的数据包文件,进行协议的逐层分析统计,得到动态的一组特征向量 <IP-Pair, PMS 可信度>,将抽样的数据包利用该特征向量进行匹配过滤,然后将得到的过滤流,利用静态的 <Keywords, PMS 软件类型>特征向量进行分类,得到具体的 PMS 应用。

(1)文中对 PMS 数据包,从端口特性、数据包特性和通信协议三个方面,进行对比统计分析,提出如下判定函数:

$$\left\{ \begin{array}{l} \sum_{i=0}^{\beta} \text{本地某主机不同的源端口} > \Phi \\ \sum_{i=0}^{\beta} \text{本地某主机不同的目的端口} > \Phi \\ \sum_{i=0}^{\beta} \text{本地某主机不同的目的 IP} > \Phi \end{array} \right. \quad (1)$$

本地某主机不同的源端口数
 \approx 不同的目的端口数
 本地某主机不同的目的 IP 数
 \approx 不同的目的端口数

公式(1)中 β 为抽样数据包个数, Φ 为判断阈值。满足该式的认为是可疑的 PMS 流量。

本系统依据如下原理:

原理 1:如果一组抽样包,经过算法统计,某主机满足公式(1),则认为该主机存在可疑的 PMS 通信。

原理 2:如果某主机和某一目的 IP 存在 PMS 通信,则以后的该 IP 对之间,皆采用该类 PMS 通信。

本系统主要划分为三个相对独立的模块:抽样模块,流量特征统计、过滤模块和特征字分类模块。其逻辑互关系图如图 1 所示。

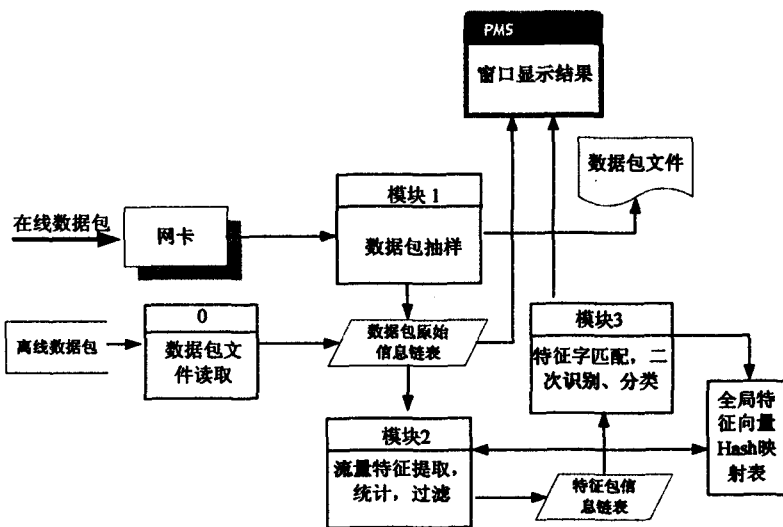


图1 系统逻辑关系示意图

2 关键算法设计

2.1 流量特征统计算法

流量统计算法是本系统中比较核心的算法,如图2所示,该算法的具体实现用到的是 STL 里的算法包括 find, find_if, 数据结构也采用 STL 提供的容器: List, Map。这样做的目的是减少算法的复杂度,同时提高算法的执行效率。

2.2 特征字符串匹配算法

特征字符串匹配算法是本程序的另一个核心,如图 3 所示。该算法主要对根据流量特征过滤的可疑的 PMS 流量进行二次识别,同时根据 IP 通信的特点,即和某个主机的一次通信被识别出来以后,以后和该主机的通信皆为该类通信,记录 IP 对之间的通信协议,以后该 IP 对之间,皆用该特征进行识别。图 3 中的全局特征向量 Hash 映射表即充当这个作用。

3 实验分析

在 VC6.0 环境下开发了 PMS 流量行为分析系统,并对 PMS 检测软件进行测试,验证其准确性和命中率。

* 测试环境:实验室局域网(PC>10)

* 本机 IP: 10. 10. 80. 131

* 测试时间: 10: 30am (高峰期)

* 抽样数据包个数: 300 个

(1)对传统的流量运用本软件进行检测。检测的目的是判读该程序统计算法的检测准确度,以误判率作为衡量的依据。通过对 HTTP 流量的检测,可以看出对于 HTTP 包的误判率很低($<0.5\%$),而且随着捕获个数的增加,误判率成下降趋势,同时也发现,误判的数据包多为一些采用非常规端口的 HTTP 数据包。通过对 FTP 流量的检测,可以发现误判率几乎为 0。这说明本算法很好地区别了常规流量与 P2P 类流量,这对本算法是一个很好的验证。

(2)对特征 PMS 流量测试分析。对 PMS 软件(PPLIVE, PPstream, QQlive)进行检测,主要运用统计

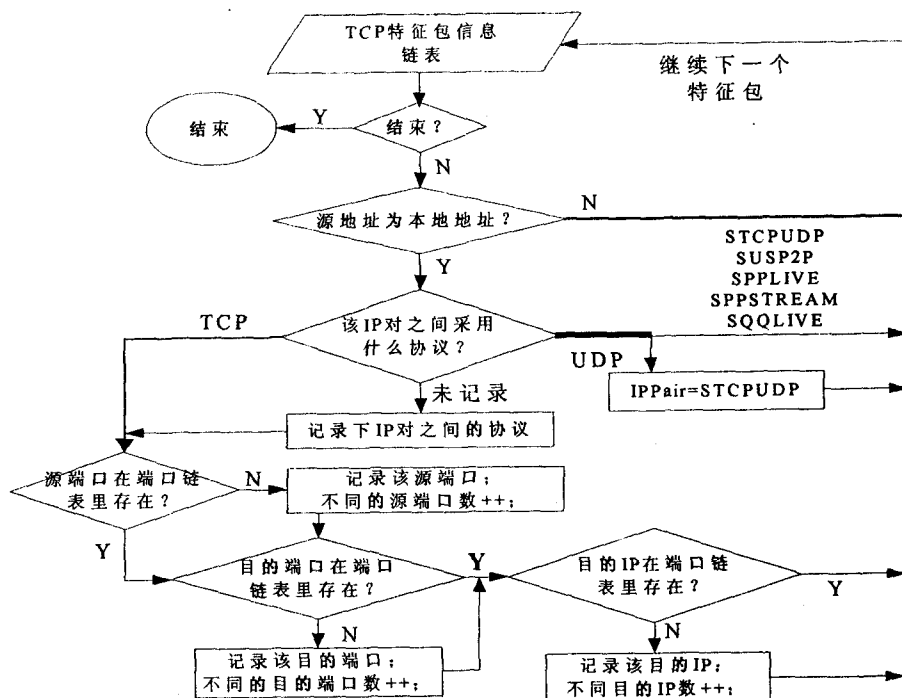


图 2 流量特征统计算法流程图

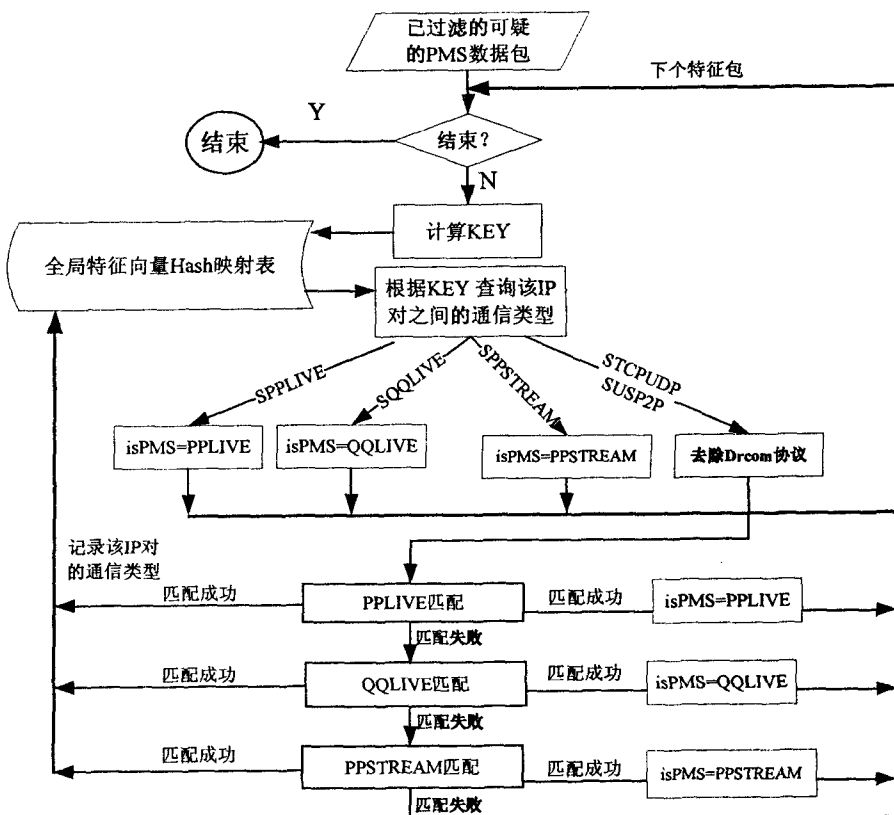


图 3 特征字符串匹配算法流程图

算法,及数据包匹配算法,故以漏判率作为衡量标准。

从图 4 中可以看出, PPLIVE 的包的漏判率在 20% 以下,而且随之捕获包数的增加,漏判率越来越接近于 5%。PPStream 的漏判率和 PPLive 类似。QQLive 的漏判率很低,初始即小于 8%,并随着数据

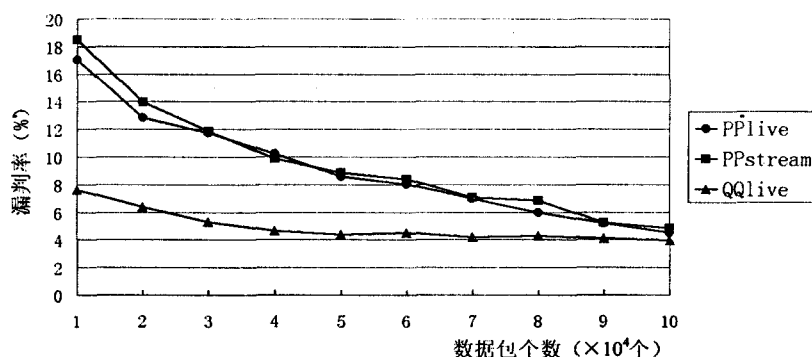


图 4 PMS 流量测试图

包数的增长,接近4%。也就是说,QQLive的流量96%可以检测出来。实验结果表明,该PMS流量行为分析系统,对PMS检测时,其准确性和命中率较好。

PMS流量行为分析系统采用VC++6.0的MFC进行开发,每个子窗口的状态由主窗口控制,如显示、隐藏。为了方便用户的使用,采用标签页的方式进行组合。其软件的运行效果如图5和图6所示。

通过该软件,用户可以在捕获窗口决定捕获数据的时间等,之后可以在数据包抽样界面得到所捕获的数据包的具体信息,如源、目IP地址,源、目端口号等。

4 结束语

文中主要是针对目前PMS流量占据大量网络带宽,造成网络拥塞,影响正常业务进行的比较严重的问题,综合PMS流量特征识别及特征字符串识别的优点,提出的一套基于流量行为分析的检测方法,并开发相应的测试系统,对笔者文中的想法进行验证。

参考文献:

- [1] Zhou LiJuan, Li ZhiTong, Liu Bin. P2P Traffic Identification on by TCP Flow Analysis[C]//Proceedings of the 2006 International Workshop on Networking, Architecture, and Storages. New York, NY, USA: [s. n.], 2006: 47 - 50.
- [2] 陈 刚,张伟文,吴国新. P2P 流媒体 Cache 的置换算法[J]. 计算机研究与发展, 2007, 44(11): 1857 - 1865.
- [3] 郑常熠,王 新,赵 进,等. P2P 视频点播内容分发策略[J]. 软

(下转第214页)

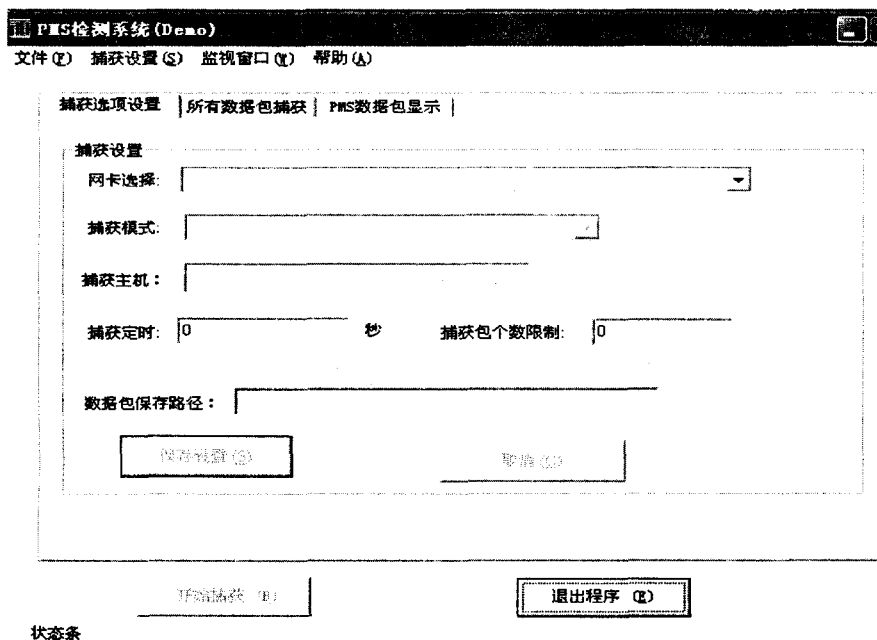
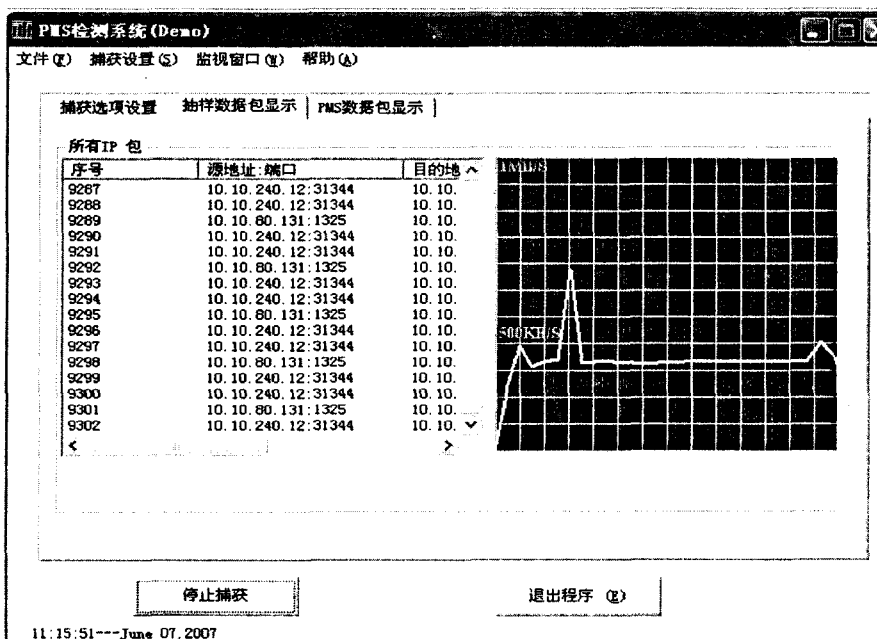


图 5 捕获设置窗口效果图



突发建立时间,OPB 从设备需要三个时钟周期读取数据之外,其他地址周期都可以在一个时钟周期完成读操作。

OPB PCI 桥接器支持任何突发大小的 OPB 写操作,突发大小只受制于 OPB 主设备写到 OPB PCI 桥的数据的大小。OPB 主设备读 PCI 从设备,OPB PCI 桥提供数据的速度和从 PCI 从设备中读取数据的速度一致。FIFO 中有效数据仅用一个时钟周期就可传输。

5 系统综合

系统综合时,除了要考虑全局的时钟、面积、管脚,以及功耗的约束之外,PCI 核的综合还需要添加特别的约束来满足 PCI 规范的要求。PCI 信号需要满足 PCI_33_3 的电平标准,尤其是使用 Virtex4 体系结构的器件时,为了满足 PCI 规范中建立和保持时间的要求,需要在部分 PCI 信号的 pad 和 I/O 缓冲器之间插入 IDELAY 单元。使用两个以上的 IDELAY 原语,且将 IOBDELAY_TYPE 设置为 Variable 时,必须使用 IDELAYCTRL 模块对其区域内独立的延时单元 (IDELAY) 进行连续的校正,来减小工艺、电压和温度变化对延时精度的影响。IDELAYCTRL 模块使用用户提供的参考时钟对 IDELAY 进行校准。IDELAYCTRL 原语在设计的流程中不能自动地产生,需要用户例化设计所需要的 IDELAYCTRL 原语,Xilinx 推荐使用两个 IDELAYCTRL 模块,使用 3 个以上的 IDELAYCTRL 会对时序产生影响。为了使用尽可能少的 IDELAYCTRL 单元,这就需要在对 PCI 信号的管脚进行约束时,尽可能将所有 PCI 的信号放在 FPGA 的同一逻辑 bank 中。以上 IDELAYCTRL 模块多于一个的情况,需要在 UCF 文件中使用如下语句对例化的延时控制器模块进行位置锁定^[7]:

```
INST "instance_name" LOC = IDELAYCTRL_X#Y#;
```

X,Y 指出了 IDELAYCTRL 单元的行和列的位置,以上约束语句指定使用和延时单元在同一逻辑 bank 中的延时控制器,对插入的延时单元进行控制,可以使时序更加容易满足设计的要求。

6 结束语

给出了一个基于 Xilinx Virtex4 器件的 SOPC 设计,对该系统中的 OPB PCI 桥接器进行了详细的介绍,并对其功能进行了仿真和测试,证明了设计功能的正确性。由于嵌入 CPU 的 FPGA 芯片具有软硬件可编程的特性,该设计实现的通用 SOPC 系统,为后续功能的开发和数据的处理提供了很大的灵活性。

参考文献:

- [1] 钟辉捷,雷航. 基于 Virtex4 的 SOPC 系统设计[J]. 航空计算技术,2007,37(3):83-85.
- [2] 齐利芳,贺占庄. SOPC 设计中的两种片上总线分析[J]. 计算机技术与发展,2006,16(1):179-181.
- [3] 田耘,胡彬,徐文波,等. Xilinx ISE Design Suite 10. x FPGA 开发指南. DSP、嵌入式与高速传输篇[M]. 北京:人民邮电出版社,2008.
- [4] Xilinx Corp. Embedded development kit(EDK)reference guide [EB/OL]. 2004. http://www.xilinx2china.com/ise/embedded/edk_docs.html.
- [5] 李贵山,陈金鹏. PCI 局部总线及其应用[M]. 西安:西安电子科技大学出版社,2003.
- [6] 林伟铭,黄联芬. 基于 PCI 的双向高速传输系统[J]. 现代电子技术,2007,30(17):41-46.
- [7] OPB IP IF/LogiCore v3 PCI core bridge user guide[EB/OL]. 2006-07-26. http://www.xilinx.com/bvdocs/ip-center/data_sheet/opb_pci.pdf.
- [8] Xilinx Corp. LogiCORE PCI version 3.0 user guide[EB/OL]. 2006-01-18. http://www.xilinx.com/products/logicore/pci/docs/pci_ug159.pdf.

(上接第 131 页)

件学报,2007,18(11):2942-2954.

- [4] 胡迎松,李强. 基于服务窗口的 P2P 视频点播模型[J]. 小型微型计算机系统,2007,28(12):8421-8425.
- [5] 姚源,褚伟. P2P 和 CDN 中 MDC 流媒体的性能对比[J]. 计算机技术与发展,2007,17(9):178-180.
- [6] 周红敏,孙名松,唐亮. 基于网络编码的 P2P 流媒体直播系统研究[J]. 计算机技术与发展,2008,18(6):225-227.
- [7] Karagiannis T, Broido A, Faloutsos M, et al. Transport Layer Identification of P2P Traffic[C]//Proceedings of the 2004 ACM SIGCOMM Internet Measurement Conference. New York, NY, USA: ACM Press, 2004:121-134.

- [8] Acharya S, Smith B. An experiment to characterize videos stored on the Web[C]//Proc. ACM/SPIE Multimedia Comput. Netw. (MMCN) 1998. New York, NY, USA: ACM Press, 1998:166-178.
- [9] Brin S, Page L. The Anatomy of a Large-scale Hypertextual Web Media Streaming Search Engine[C]//Proceedings of the Seventh International World Wide Web Conference. Brisbane, Australia: Elsevier Scienc, 2004:1-10.
- [10] Pieper J, Savitba, Dom B. Streaming - Media Knowledge Discovery[J]. Computer, 2001(8):68-74.