

基于分离映射的混合交换路由系统设计与实现

朱伟伟, 罗洪斌, 陈 哲, 苏 伟

(北京交通大学 电子信息工程学院 下一代互联网互联设备国家工程实验室, 北京 100044)

摘 要: 当今互联网采用“尽力而为”的 IP 分组转发模式进行数据包的传输, IP 分组转发模式对所有的数据包进行一视同仁的转发而不加以区分, 这种转发模式不能够满足当今网络服务多样性的要求。为此, 文中提出了一种基于分离映射的混合交换路由系统, 根据数据流的特性将某些数据流直接利用底层物理资源采用电路交换的方式进行数据包的传输, 而对其他流仍然用 IP 分组进行尽力而为的转发。文中详细地介绍了该方案的设计和实现, 并对其主体功能进行了测试, 验证了该方案的可行性。

关键词: 一体化网络; 身份与位置分离; 混合交换路由; 多协议标签交换

中图分类号: TP31

文献标识码: A

文章编号: 1673-629X(2013)02-0007-04

doi: 10.3969/j.issn.1673-629X.2013.02.002

Design and Realization of Hybrid Switching Routing System Based on Separation Map

ZHU Wei-wei, LUO Hong-bin, CHEN Zhe, SU Wei

(National Engineering Laboratory for Next Generation Internet Interconnection Devices, School of Electronic
and Information Engineering, Beijing Jiaotong University, Beijing 100044, China)

Abstract: The current Internet uses the best-effort IP packet mode to transmit, it forwards the packets in the same way without distinction and can not meet the diverse requirements of today's network services. To this end, propose a hybrid switching routing system based on separation map. Some of the data stream use the underlying physical resources to transmit the packets in circuit-switched mode directly, the other stream still use the best-effort IP packet forwarding mode, which is according to the characteristics of the data stream. It introduces the design and realization of the program in detail, tests its main function and verifies the feasibility of the program.

Key words: universal network; identifier/locator separation; hybrid switching routing; multi-protocol label switching

0 引言

当今互联网采用尽力而为的 IP 分组转发模式进行数据包的传输, 路由器对所有 IP 分组一视同仁地“尽力而为”进行转发。然而, 一些特殊服务(如语音、视频、设备监控、网络计算等)的分组希望可以进行特殊的处理, 以保证其服务质量要求(如最大时延、最小带宽等)。但是 IP 提供的“尽力而为”分组转发不能满足服务多样性的要求。

另一方面, 由于当今网络所有数据包都由 IP 协议处理, 进行分组交换, IP 层成为数据传输的瓶颈, 不仅丢包率高, 而且底层的物理资源没有得到充分利用, 耗能大。目前, 光传输技术已是一种稳定可靠的传输技

术, 它已拥有了一个庞大的网络体系, 随着进一步的发展, 近年来光传输带宽得到了成百倍、上千倍的提高, 其传输资源不可限量。然而其传输资源却没有得到充分的利用, 据统计, 互联网骨干网络忙时最大平均链路利用率不足 30%, 很多网络闲时链路利用率在 5% 以下^[1]。如果能够根据所传送的数据流的特性, 让一部分数据流直接通过底层的光纤进行电路交换, 而其他数据流进行 IP 分组交换, 不仅可以减少 IP 层处理时的丢包, 而且可以充分利用底层的物理资源, 对降低网络能耗也具有重要的意义。

1 混合交换路由系统方案设计

文中在一体化网络的基础上提出一种混合交换路由系统的设计方案, 在核心网的入口路由器上对数据流的某些特性(如流所传送文件的大小或流所要求的服务质量等)进行判定, 根据判定的结果使不同的数据流通过不同的平面传输。例如, 对于服务质量要求

收稿日期: 2012-05-15; 修回日期: 2012-08-22

基金项目: 国家自然科学基金资助项目(61100219, 60903150, 60870015); 中央高校基本科研业务费专项资金(2012JBM010)

作者简介: 朱伟伟(1987-), 男, 硕士研究生, 主要研究方向为下一代互联网; 罗洪斌, 教授, 主要研究方向为下一代互联网。

高或者传送大文件的流,使其通过电路交换传输,直接走底层的光纤链路,而其他的数据流则通过分组模式传输。文中利用 MPLS (Multi-protocol Label Switching 多协议标签交换)来模拟电路交换。

1.1 一体化网络概述

一体化网络^[2,3]是一种基于身份位置分离思想的信息网络体系架构,支持多元化的终端和异构网络的接入,是一种全新的未来网络体系架构。为了解决传统互联网 IP 地址同时携带终端身份信息和位置信息的问题,一体化网络原创性地引入接入标识 (Accessing Identifier: AID) 和交换路由标识 (Switching Routing Identifier: RID) 分离映射理论,在接入网中用接入标识代表用户的身份信息,核心网中用交换路由标识代表用户的位置信息。位于核心网边界的接入交换路由器完成两种标识的解析映射以及映射关系的维护等功能。

一体化网络中的标识映射协议栈 (Identifier Mapping Protocol: IDMP) 实现了标识映射^[4]的思想。位于核心网边界的接入交换路由器 (Access Switch Router: ASR) 实现了分离映射功能;同时,标识映射服务器 (Identifier Mapping Server: IDMS) 存储了 AID 和 RID 的映射表,而广义交换路由器 (Generic Switch Router: GSR) 对核心网中的数据包进行快速地路由和转发。一体化网络的优点使其有效克服了传统互联网面临的可扩展性、安全性、移动性以及可控可管性^[3,5]等方面的严重缺陷。

1.2 MPLS 概述

MPLS^[6]是一种用于快速数据包交换和路由的体系,它为网络数据流量提供了路由、转发和交换等能力。它位于 OSI 七层模型中的第 2 层和第 3 层之间的 2.5 层技术。它通过在 IP 包头添加 32 比特的“shim”标签,可使原来面向无连接的 IP 传输具有面向连接的特性,加快 IP 包的转发速度。

由于 MPLS 这种可以使面向无连接的 IP 传输具有面向连接的特性,文中用它来模拟实际传输中的电路交换。而且 MPLS 位于 TCP/IP 协议栈的 2.5 层,由 MPLS 转发的数据包不会向上面的 IP 层递交而直接进行转发,使得转发过程更为简单。

1.3 混合交换路由系统方案设计

对于进入 ASR 的数据流,不同数据流对服务质量的要求是不同的,而且不同的数据流所传送文件大小不同,传送所用时间也不相同。文中根据数据流的服务质量和所传送文件的大小来决定它所进入的转发平面,即 IP 分组转发平面或 MPLS 平面。当进入 ASR 的数据流对服务质量要求较高或所传送的文件较大

时,就对其添加 MPLS 标签,而对其他的数据包则不作任何处理,直接通过 IP 分组进行转发。其总体设计方案如图 1 所示。

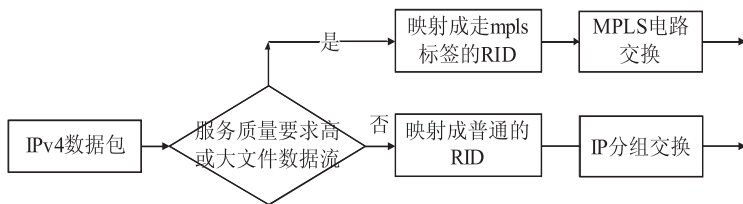


图 1 混合交换路由系统总体设计方案

在系统实现时,为方便起见,为不同的转发平面分配不同的 RID 池,例如,给 MPLS 电路交换平面分配的 RID 池为 10.1.1.0/24 网段,给 IP 分组转发平面分配的 RID 池为 10.2.1.0/24 网段。在数据流进入核心网边界接入交换路由器时,判定数据流的服务质量和数据流所传送文件的大小,如果服务质量要求高或是数据流所传送的文件较大,就将其 RID 映射到 10.1.1.0/24 网段内,对于其余的数据流就将其 RID 映射到 10.1.2.0/24 网段。

当数据包进入 MPLS 子系统时,对于 RID 在 10.1.2.0/24 网段的数据包,系统就认定这样的数据包要进行电路交换,它就会被打上 MPLS 标签,经 MPLS 进行模拟电路交换。而对于 RID 在 10.1.1.0/24 网段的数据包,系统会认为此数据包是要进行 IP 分组转发的数据,而后它会被传送到 IP 层进行处理,进行普通的分组转发。

2 混合交换路由系统各子系统的实现

当一个 IP 数据包进入到一体化网络时,ASR 完成的功能主要包括截获 IP 网络数据包,替换 IP 标识、添加 MPLS 标签、路由转发,以下为各子系统的设计流程。

2.1 内核映射子系统的设计

为实现分离映射的功能,需要对 Linux 内核的网络协议栈进行修改^[7],在协议栈的设计中,映射模块的入口在 Netfilter 的 PREROUTING 钩子之后,FORWARD 钩子之前,而且可以通过开关来控制映射的开启与关闭。

当打开映射开关时,进入 ASR 的数据包会经过上面的支路进行分离映射后进行转发,而将映射开关关闭时,进入 ASR 的数据包则不进行分离映射直接进行普通的 IP 分组转发。

如图 2 所示,在 ASR 接收到数据包后,数据包从虚拟设备接口层进入网络层协议栈,接收函数为 ip_rcv(),接下来判定映射子系统是否打开,如果没有打开,则不进行映射,直接做路由转发处理;如果映射子

系统已经打开,则对接收到的数据包进行分离映射,接下来将数据包交给 MPLS 子系统进行处理。

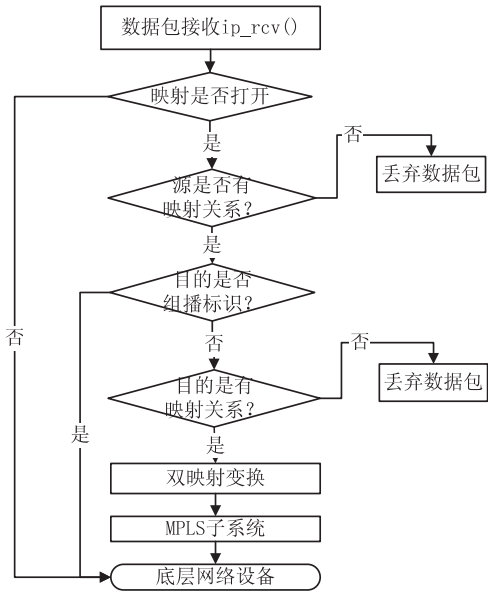


图2 网络层处理流程

2.2 MPLS 子系统的设计

对于 MPLS 网络^[8]而言,接入交换路由器 ASR 即为其边缘路由器(LER)。MPLS 层的数据处理流程如图3所示,当数据包进入 ASR 后,首先在 IP 层进行处理,在映射打开的情况下,分离映射子系统会根据数据包的服务质量和所传送文件的大小等性质对数据包路由标识进行分离映射。

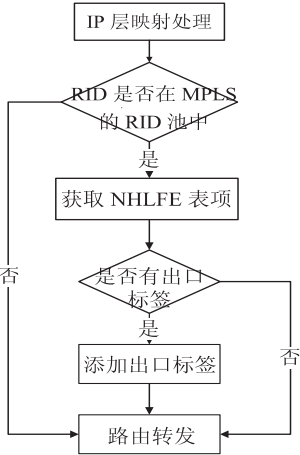


图3 MPLS 层处理流程

MPLS 子系统^[9,10]先对接收到数据包的 RID 进行判定,如果 RID 不在 MPLS 对应的 RID 池中,则直接进行路由转发;如果 RID 在 MPLS 对应的 RID 池中,则数据包会进入 MPLS 子系统进行处理,首先获取 RID 的 NHLFE 项,查看是否为其设定了出口标签,如果设定了出口标签,就对其添加出口标签,将其转发出去^[11]。

至此,对于要进行电路交换的数据流就会通过 MPLS 转发出去。

3 功能测试

3.1 测试平台和环境

测试平台如图4所示,终端 A1 和终端 A2 接入到接入交换路由器 ASR1 上,终端 B 接入到接入交换路由器 ASR2 上,ASR1 和 ASR2 位于核心网与接入网的边界,完成终端通信过程中的分离映射和路由平面映射功能。广义交换路由器 GSR1 和 GSR2 位于核心网,负责核心网中的路由以及一体化网络报文的转发。

本测试环境中的终端、路由器以及服务器都是基于普通 x86 架构的 PC 机或者工控机。ASR 和 GSR 为 Linux 平台,内核为修改后的同时具有 MPLS 和分离映射功能的 2.6.27-mpls-map 版本。终端可以使用 Linux 平台或者 Windows 平台,对版本无特殊要求。

终端 A1 的 eth0 口的 AID(即 IP 地址)为 192.168.1.1,终端 A2 的 eth0 口的 AID 为 192.168.2.1,终端 B 的 eth0 口的 AID 为 192.168.3.1。在这里,假定从终端 A2 发出的数据包对传输质量的没有要求,进行普通 IP 分组转发,而从终端 A2 上发出的数据包对传输质量要求较高,需要进行电路交换。在骨干网中,10.1.1.0/24 网段的地址进行普通 IP 分组交换,10.1.2.0/24 网段地址进行 MPLS 电路交换。

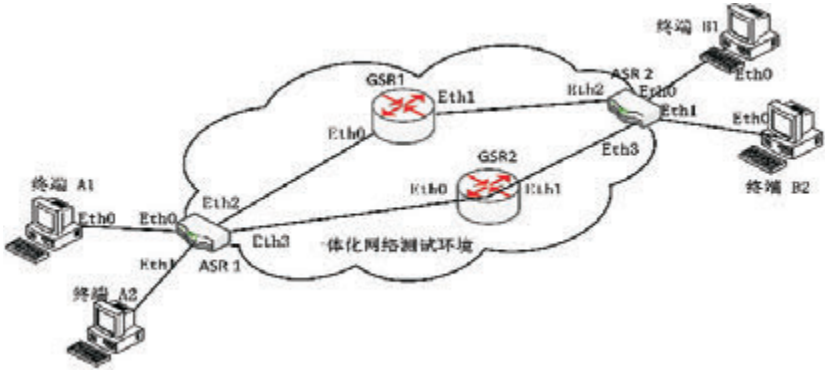


图4 测试平台网络拓扑

ASR1 为终端 A1 分配的 RID 为 10.1.1.1,为终端 A2 分配的 RID 为 10.1.2.1,ASR2 为终端 B 分配的 RID 为 10.1.3.1。

3.2 功能的测试

搭建好实验环境,在终端 A1 和终端 A2 上同时 ping 终端 B,分别在 ASR1 的 eth2 和 eth3 口上使用 wireshark 工具进行抓包,结果如图5和图6所示。

由图5和图6可知,从 ASR1 的 eth2 口发出的包的 RID 被映射成了 10.1.1.1,而且是在骨干网中通过普通 IP 分组模式进行传输。从 ASR2 的 eth3 口发出的包的 RID 被映射成了 10.1.2.1,在 ASR1 上数据包已经被添加上了 MPLS 标签,出口标签为 1000,在骨干网中以 MPLS 电路交换的方式进行传输。

以上的实验过程及测试结果说明,一体化网络的

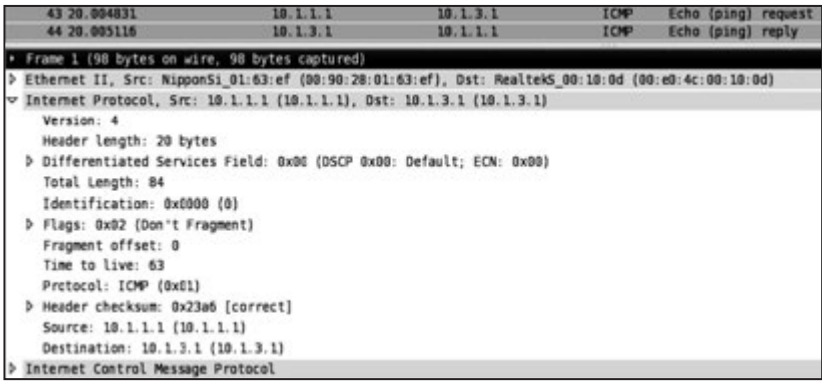


图 5 ASR1 的 eth2 上抓到的包

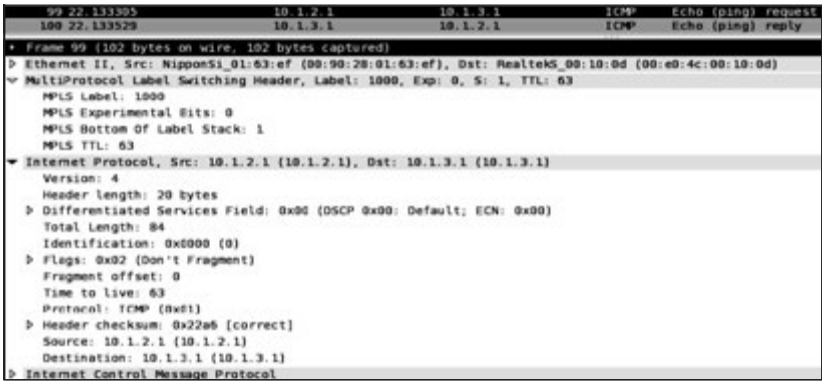


图 6 在 ASR1 的 eth3 上抓到的包

网络虚拟化基本功能测试通过,假定的不同性质的数据流在 ASR 处被分离开,服务质量要求较高或传送文件较大的数据流通过 MPLS 模拟电路交换被发送出去,而其余的数据流则通过 IP 分组交换被发送出去。

4 结束语

文中设计并实现了一种基于身份与位置分离映射的混合交换路由系统。在该系统中,接入路由器对数据流的某些特性(如流所传送文件的大小或流所要求的服务质量等)进行判定,并根据判定结果使不同的数据流通过不同的分组/电路平面传输。例如,对于服

务质量要求高或者传送大文件的流,使其通过电路交换传输,直接走底层的光纤链路,而其他的数据流则通过分组交换传输。文中设计了该系统的所有功能模块,并且进行了功能测试。

参考文献:

[1] Guichard J, Faucheur F L, Vasseur J P. Definitive MPLS network designs [M]. Indianapolis: Cisco Press, 2005.

[2] 张宏科, 苏伟. 新网络体系基础研究——一体化网络与普适服务[J]. 电子学报, 2007, 35(4): 593-598.

[3] 杨冬, 周华春, 张宏科. 基于一体化网络的普适服务研究[J]. 电子学报, 2007, 35(4): 607-613.

[4] 董平, 秦亚娟, 张宏科. 支持普适服务的一体化网络研究[J]. 电子学报, 2007, 35(4): 599-606.

[5] Farinacci D, Fuller V, Oran D. Locator/ID separation protocol (LISP) [M]. [s.l.]: IETF, 2007.

[6] Rosen E. Multiprotocol Label Switching Architecture [S]. RFC3031, 2001.

[7] Bovet D, Cesati M. Understanding The Linux Kernel [M]. [s.l.]: [s.n.], 2005.

[8] 陈启美, 吴政, 刘海. MPLS 组件与框架—MPLS 体系结构解析[J]. 电力自动化设备, 2002(2): 87-90.

[9] 肖宇峰, 李昕, 时岩. Linux 网络内核分析与开发 [M]. 北京: 电子工业出版社, 2010.

[10] DeGhein L. MPLS 技术架构 [M]. 陈麒帆译. 北京: 人民邮电出版社, 2008.

[11] 赵强, 鲁昆生. 多协议标记交换 (MPLS) 技术研究及应用 [J]. 武汉理工大学学报, 2004(3): 94-96.

+++++ (上接第 6 页)

Semi-supervised Clustering [C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle: ACM Inc, 2004: 59-68.

[10] Nigam K, McCallum A K, Thrun S, et al. Text Classification from Labeled and Unlabeled Documents Using EM [J]. Machine Learning, 2000, 39(1): 103-134.

[11] 张敏, 于剑. 基于划分的模糊聚类算法 [J]. 软件学报, 2004, 15(6): 858-869.

[12] Frey B J, Dueck D. Clustering by Passing Messages between

Data Points [J]. Science, 2007, 315(5814): 972-976.

[13] 肖宇, 于剑. 基于紧邻传播算法的半监督聚类 [J]. 软件学报, 2008, 19(11): 2803-2813.

[14] CLUTO Document Datasets Toolkit [EB/OL]. [2010-06-01]. <http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/datasets.tar.gz>.

[15] Luo Congnan, Li Yanjun, Chung S M. Text Document Clustering Based on Neighbors [J]. Data and Knowledge Engineering, 2009, 68(11): 1271-1288.

基于分离映射的混合交换路由系统设计与实现

作者: [朱伟伟](#), [罗洪斌](#), [陈哲](#), [苏伟](#)
作者单位: [北京交通大学 电子信息工程学院 下一代互联网互联设备国家工程实验室, 北京 100044](#)
刊名: [计算机技术与发展](#)
英文刊名: [Computer Technology and Development](#)
年, 卷(期): 2013 (2)

本文链接: http://d.g.wanfangdata.com.cn/Periodical_wjz201302004.aspx