支持业务需求灵活定制的多态路由系统

张 岩,兰巨龙,胡宇翔,王 鹏,段 通
(国家数字交换系统工程技术研究中心,河南郑州 450002)

摘 要: 传统僵化单一的路由机制已经无法适应未来多样化的业务需求和各种新型网络体系结构的试验与部署.针对此问题,本文基于路由功能与业务需求自适配的思想提出了多态路由模型,并设计实现了多态路由原型系统.该系统通过虚拟化技术以及灵活可编程的数据平面结构,实现了同构和异构网络中多种路由协议的共存,完成了基于路由服务描述的路由协议个性化定制和数据平面的多表选择查询与转发处理.最后,基于NetFPGA-10G平台设计实现了多态路由原型系统.相较于现有路由试验系统,多态路由系统在实现路由协议定制化及异构网络共存的同时,更好地保证了业务的服务质量,具有更高的转发速率以及可扩展性.

关键词: 路由;多态;虚拟化技术;数据平面;NetFPGA
中图分类号: TP393.1 **文献标识码:** A **文章编号:** 0372-2112 (2016)04-0988-07
电子学报URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2016.04.033

A Polymorphic Routing System Providing Flexible Customization for Service

ZHANG Yan, LAN Ju-long, HU Yu-xiang, WANG Peng, DUAN Tong
(*National Digital Switching System Engineering & Technological R&D Center, Zhengzhou, Henan* 450002, *China*)

Abstract: The traditional rigid routing mechanism has been unable to meet the various service demands and the deployment of new network architectures in future. Aimed at this problem, this paper proposes the polymorphic routing model which is based on the adaptation between routing function and service requirements, and implements the polymorphic routing prototype. The prototype supports the coexistence and customization of a variety of routing protocols in both homogeneous and heterogeneous networks through the virtualization technology and the selective data plane for query and forwarding. Finally, the prototype was implemented based on the NetFPGA-10G platform. Compared with the existing routing systems, the polymorphic routing system realizes the coexistence and customization of various routing protocols, and achieves higher forwarding rate, better scalability and insurance of the quality of service at the same time.

Key words: routing; polymorphic; virtualization technology; dataplane; NetFPGA

1 引言

目前,传统路由传输方式已无法满足各种业务对网络传输质量的不同要求[1].因此,为了支持未来多样化的网络体系结构[2,3]与现有TCP/IP网络长期并存以及未来网络体系结构的渐进式试验与部署[4],设计新型路由系统[5],便成为解决此问题的关键点和突破口.Eddie Kohler等人[6]设计实现了模块化的软件路由器Click,但是由于采用纯软件实现,转发性能较差.为了提高转发速率,Han等人[7]利用通用GPU提出了高速软件路由器PacketShader,然而,PacketShader由于使用专用GPU实现数据平面核心的转发引擎,其可编程性有限,难以满足未来网络研究中非IP协议的转发需求.Dobrescu等人[8]利用集群技术提出了高性能软件路由器RouteBricks.但是其数据通路完全通过软件方式构建,因而修改难度较大.基于专用硬件方面,Anwer等人[9]设计了一套基于FPGA硬件的可编程虚拟路由器框架SwitchBlade.但是受限于FPGA资源限制,系统最多可支持四个同构的硬件虚拟数据平面,可扩展性较差.刘中金等人[10]基于并行流水线的虚拟路由器数据平面结构,在同一物理底层上实现了多个相互隔离的并行异构路由器.但是该设计针对普通以太网数据包仅有一个数据处理平面,不能实现基于业务分类的数据平面选择.

为了支持后续不断涌现的新兴路由需求,未来网络需要通过网络路由结构的自组织、功能的自调节和业务的自适配来最大程度地弥合网络路由服务能力与业务需求之间的时变鸿沟[11],支持多种网络结构并存

收稿日期:2015-07-16;修回日期:2015-09-18;责任编辑:孙瑶
基金项目:国家973重点基础研究发展计划(No.2012CB315901,No.2013CB329104);国家863高技术研究发展计划(No.2013AA013505);国家自然科学基金(No.61309019,No.61372121)

的网络寻址及路由,使网络具有动态调整适应多样化业务需求的路由传输能力. 基于此思想,本文设计了支持业务定制以及异构网络共存多态路由系统,主要贡献包括:(1)提出多态路由模型,设计协议需求匹配算法实现需求与协议之间的映射;(2)提出多态路由系统总体架构,设计基于控制平面的多态路由派生算法以及基于数据平面的多态路由决策算法,实现基于需求的协议定制以及路由选择;(3)设计实现多态路由原型系统;(4)对原型系统的功能、转发性能以及服务质量保障性能进行了测试验证与分析.

2　多态路由系统

2.1　多态路由运行实例

由于未来网络可能是多态的[12],即支持多种网络体系结构共存以及现有 IP 网络到未来网络的不断演进,处于网络结构核心的路由功能需要进行同步革新. 因此,本文提出了多态路由模型,即在多样化业务需求和网络动态行为驱动下,多态路由可以针对不同业务及网络类型特定的需求,为其选择派生相应的路由协议,从而构建出以满足具体应用需求的各种约束属性服务路径. 多态路由协议运行实例具体如图 1 所示.

在图 1 所示的场景中,当三种不同类型的业务到达路由器 R1 时,路由器 R1 通过下文设计的协议需求匹配算法分别为三种业务匹配到最佳的路由协议 A、B、C. 如果当前路由器中相应路由协议并未运行,则通过下文设计的多态路由派生算法启动相应路由协议. 由此,在多态路由协议中,不同的业务根据需求实现了基于定制化路由协议的多样化路径传输,更好地保障了业务服务质量.

2.2　多态路由模型

为了更好地描述多态路由模型,本文引入路由服务描述(Routing Service Description,RSD)的概念. 路由服务描述是对业务需求的刻画和表征,为了提高匹配效率,路由服务描述采用类似 NDN 中扁平式的命名方案. 定义如下:

$$RSD = D^{A}_{T_1}/D^{B}_{T_2}/\cdots/D^{K}_{T_i} \tag{1}$$
$$T_k = 1,2,3,\cdots$$

其中,上标 A,B,K 代表路由服务描述的不同层级,下标 T_k 表示相应层级下具体的类型.

基态路由(Base state Routing,BR)是具备满足所有路由服务描述的路由协议集合,可视为一个资源池. 使用向量空间定义如下:

$$\boldsymbol{BR}^K = \{\boldsymbol{\beta} = \sum_{i=1}^{K} \lambda_i \partial_i \mid \lambda_i \in R, i = 1,2,\cdots,K\} \tag{2}$$

$\boldsymbol{BR}^K$ 表示 K 维向量空间,$(\partial_1,\partial_2,\cdots,\partial_k)$ 是一组标准正交基底,即 $\partial_i = (\underbrace{0,0,\cdots,1,\cdots,0}_{K})$. K 对应于 RSD 中的层级总数,基态路由 $\boldsymbol{BR}^K$ 中的每一维向量 ∂_k 对应于 RSD 中的第 k 层需求.

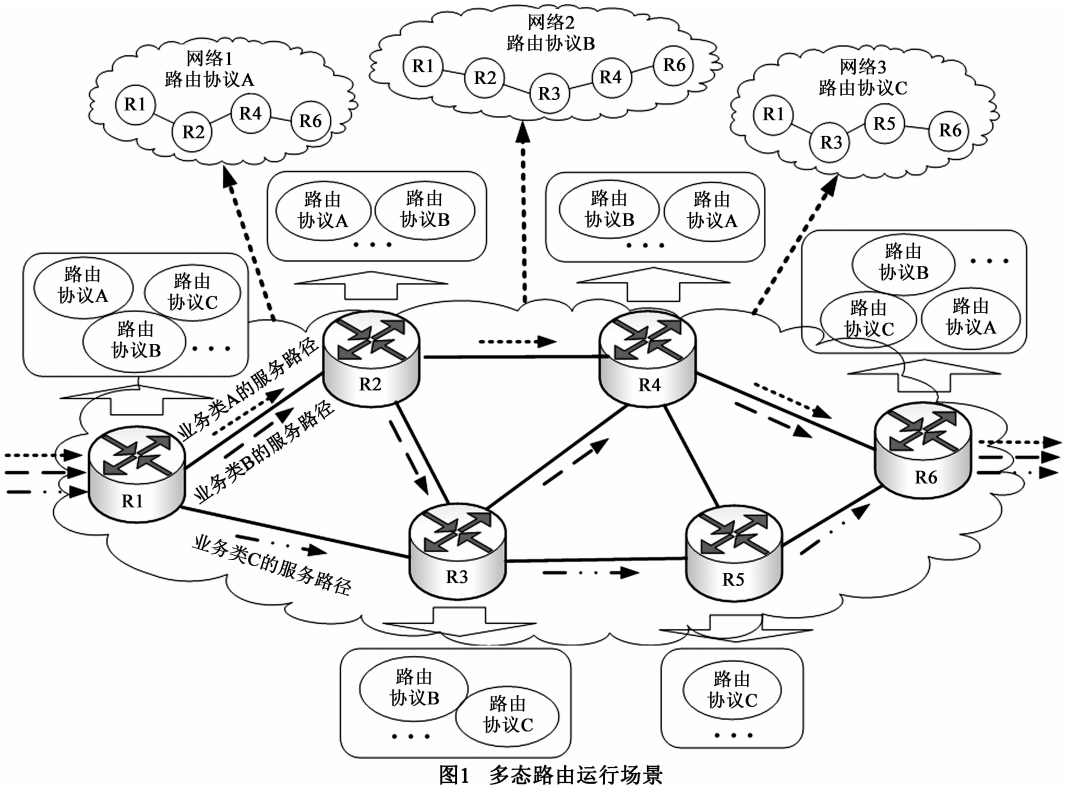

图1　多态路由运行场景

多态路由(Polymorphic Routing,PR)是基于特定路由服务描述,由基态路由进行实例特化以建立满足具体应用所需的各种约束属性服务路径的路由机制. 路由服务描述 RSD 可以视作基态路由向量空间的一组坐标,唯一确定了多态路由实例.

那么由基态到满足具体 RSD 需求的多态路由实例的派生过程即可描述为

$$\boldsymbol{PR} = T_1 \cdot \boldsymbol{SA} + T_2 \cdot \boldsymbol{SB} + \cdots + T_k \cdot \boldsymbol{SK} \quad (3)$$

其中,($\boldsymbol{SA},\boldsymbol{SB},\cdots,\boldsymbol{SK}$)为基态路由向量空间的一组标准正交基,与 RSD 需求层级一一对应,T_k 对应于 RSD 中第 K 层需求的类型.

为了建立路由服务描述与路由协议的映射关系,本文采用基于特里树的方法实现. 由于路由服务描述是层次化表示的,且不同的 RSD 具有共同的前缀表示,因此采用特里树结构可以节省大量存储空间,同时由于减少了无谓的前缀比较可以大大提高查询匹配效率.

基于具有高效查询效率的特里树结构,多态路由模型能够根据特定 RSD 快速查找派生出相应编号的路由协议. 针对此过程,本文设计了协议需求匹配算法进行实现,算法复杂度为 $O(N)$,N 为 RSD 的层级数目,如算法 1 所示.

算法 1 协议需求匹配算法

输入:路由服务描述 RSD

输出:路由协议编号 ID

```
0.  T = createTRIE(RSD)   //根据 RSD 建立 Trie 树
1.  (D_{T_1}^{A}, D_{T_2}^{B}, ..., D_{T_k}^{K}) (Decompose(RSD)   //解析 RSD 需求
2.  node = root
3.  Pre_Order(T)      //前序遍历 Trie 树
4.  for  i = A:K
5.    branches←get_branch(node)   //查找节点分支
6.    choose T_ith branch based on D_{T_i}^{L}//根据 RSD 确定分支
7.    branch←branches[T_i]
8.    node = node[branch].child
9.    if node is leaf            //找到叶子节点
10.     ID = node.id             //返回节点编号
11.     return
12.   end if
13. end for
```

2.3 总体架构

基于多态路由模型的实现机理,本文以可编程硬件和虚拟化技术为核心,设计了多态路由系统的总体架构. 该系统能够支持多协议共存、协议灵活可编程以及数据报文的高速转发. 如图 2 所示,总体架构分为控制平面和数据平面两大部分.

2.3.1 控制平面结构

控制平面的主要功能是根据需求完成对多态路由协议实例的派生以及相应配置. 为了实现多种路由协议实例的有效隔离,虚拟容器与物理主机的数据通信链路方式如图 3 所示.

图 3 中,虚拟容器创建虚拟网卡 0 ~ 3(与物理网卡数量一致),然后基于端口划分 VLAN,最后通过网桥 br0 ~ br3 与对应于物理板卡的 nf0 ~ nf3 分别连接. 网桥实现了虚拟容器与物理板卡间的报文通信,VLAN 实现

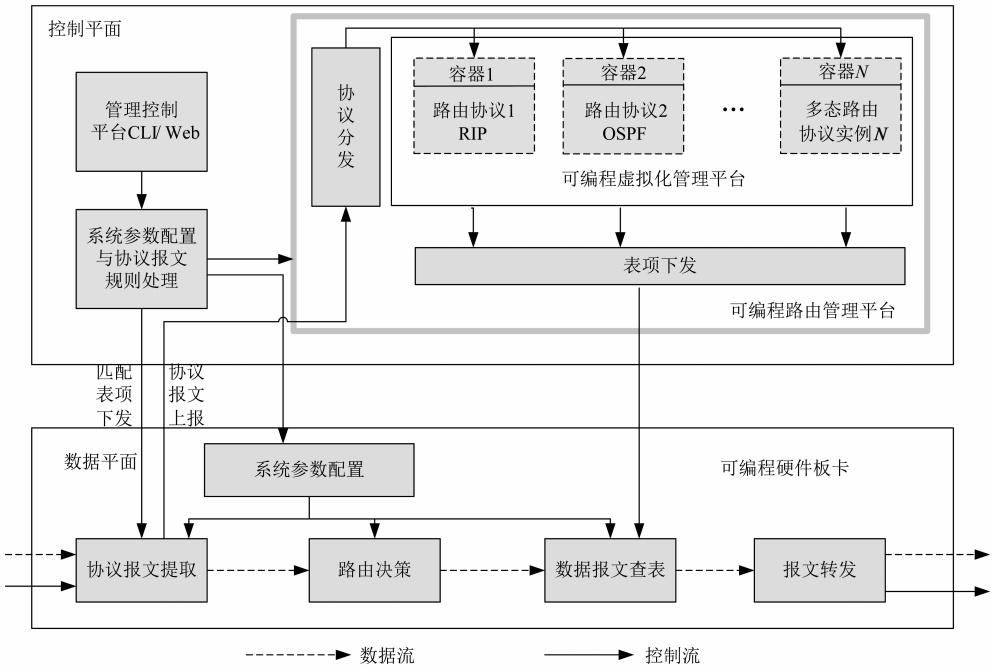

图2 多态路由系统总体架构

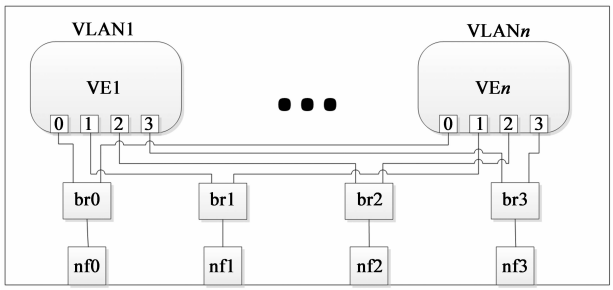

图3　网卡映射方式

了不同路由协议报文交互的有效隔离.

多态路由系统需要根据业务需求对路由协议报文进行灵活控制,以实现路由协议实例的按需派生. 因此,本文设计了多态路由派生算法,如算法 2 所示.

算法 2　多态路由派生算法

```
0.  for each packet arrived
1.    set vid←get VLAN tag      //获得 VLAN 标签
2.      for  i = 1:N
3.        if ∃i, Container[i] = = vid      //匹配 VLAN 标签
4.          strip VLAN tag, sent packet to container i   //选择相应虚拟容器
5.          rtable[i]←calculate routing table
          else
6.          start a new container N      //增加虚拟容器
7.          set Container[N]←vid
8.          Protocol[N]←Configure a new routing protocol   //配置新路由协议
          end if
      end for
9.  end for
```

当业务需求改变时,多态路由系统可以通过修改 VLAN-ID 映射表来修改协议分发规则,从而更新不同路由协议的交互对象,建立新的路由转发表,这为底层硬件路由决策模块提供判断依据. 同时,这也为用户编写复杂的路由控制策略提供了灵活可编程接口. 最后,路由管理平台还可根据上层不断涌现的新兴业务需求,为其开发并部署运行满足需求的路由协议,从而实现异构网络的路由机制共存.

2.3.2　数据平面结构

数据平面主要根据控制平面配置生成的规则进行数据包的高速转发与处理. 包头分类模块根据分类规则区分不同业务类别的数据包,然后多态路由决策模块根据 TOS—RID 映射表将数据包送交不同的路由转发平面进行处理. 因此,多态路由系统设计实现了基于可编程硬件平台的多表存储与查询机制,以及支持用户定制的表项更新策略,从而实现根据业务需求定制路由的目的. 图 4 是硬件数据平面的结构设计图.

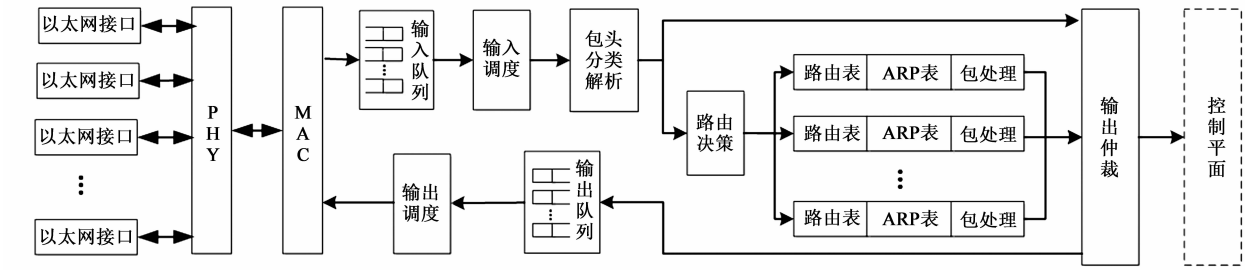

图4　硬件数据平面结构

现有设计中使用报文头部八位 TOS 字段对上层业务类型进行标识,具体标识方案可参考 DiffServ 中的相关设计. 本设计中,高速包头分类解析模块主要完成对路由协议控制报文以及普通数据包进行区分,路由决策模块通过在硬件 TCAM 中存储并维护一张 TOS—RID 映射表,实现业务类型与硬件路由表之间的高效查找匹配. 路由表以及 ARP 表均使用最长前缀匹配算法在硬件 TCAM 中完成处理,以此保证多态路由系统的高速转发功能. 系统结合包头分类解析模块以及路由决策模块,运行多态路由决策算法,从而实现基于业务类型匹配的定制化路由选择转发功能,如算法 3 所示.

算法 3　多态路由决策算法

```
0.  for each packet arrived
1.    set MACn←get ingress port
2.    if it is VLAN packet      /* 如果是协议报文,送交软件控制平面处理 */
3.      sent packet to CPUn
4.    else
5.      set tos←get TOS tag
6.      rid←Lookup TOS--RID table   //查表进行路由决策
7.      if rtable[rid] exist in hardware   //表项存在硬件中
8.        nexthop←Lookup routing table rtable[rid]
9.        MAC address←Lookup arp table atable[rid]
10.         sent packet to egress port MACx
11.     else   //表项不在硬件中,自定义表项更新策略下发至硬件
12.       write rtable[rid] in hardware based on user-defined strategy
```

```
13.          update TOS—RID table
14.            goto (8) - - - (10)
15.        end if
16.      end if
17.   end for
```

算法 3 中,数据平面由于资源限制导致路由转发平面数量不能完全满足控制层路由协议需要,但表项更新策略支持用户自定义编程实现,从而支持周期性选择更新硬件路由转发表来满足数据转发需要,从而提升了系统可扩展性. 针对异构网络共存时,本设计方案中的路由表以及 ARP 表可以进行相应泛化.

2.4 原型实现

多态路由原型系统控制平面采用 OpenVZ + Quagga 的架构,其中,OpenVZ[13] 属于操作系统级虚拟化软件. Quagga[14] 属于开源的路由协议配置平台,具有灵活性好、可编程性强的特点. 数据平面采用 NetFPGA-10G[15] 可编程板卡实现.

在原型系统中,目前数据平面支持四张路由表以及 ARP 表的并行查询,表项条目数量均为 32,可根据需要进行扩展.

3 性能测试与分析

为了对多态路由系统原型进行测试和验证,本节搭建了一个测试场景,然后分别从功能和性能的角度对多态路由系统原型进行测试和验证.

3.1 测试场景

如图 5 所示,多态路由系统原型测试场景由五个运行多态路由系统的节点组成,节点之间通过 10GE 的光纤接口进行互联. 在每个节点中通过 OpenVZ 虚拟化技术分别运行 RIP、OSPF 以及 ISIS 路由协议. 网络中分别部署 VOIP、FTP 以及视频三种类型业务,三种业务的 TOS 字段分别设为 0x00,0x01,0x02,在 TOS—RID 映射表中分别对应路由协议 RIP,OSPF,ISIS 生成的路由表.

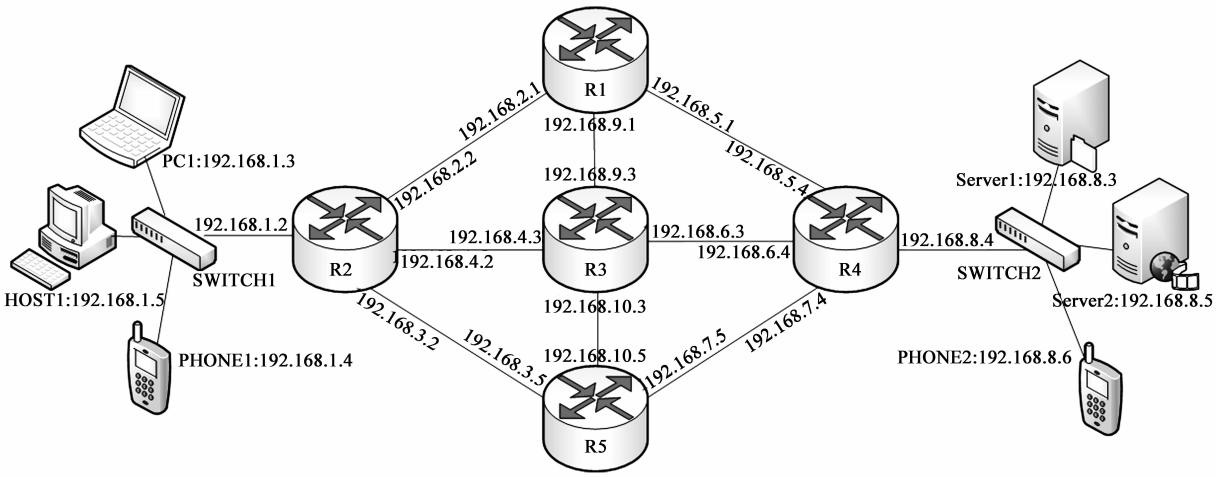

图5 多态路由原型系统测试场景

3.2 多态路由功能验证

基于上图所示的测试场景,使用 Traceroute 命令指定数据包 TOS 字段查看不同业务数据包转发路径.

结果表明,VOIP 业务的传输路径为 R2→R3→R4,FTP 业务的传输路径为 R2→R5→R4,视频业务的实际传输路径为 R2→R1→R4,这恰好完全对应于三种路由协议计算生成的路由表. 同时,不同路由协议之间相互独立运行,互不影响. 由此表明,多态路由原型系统实现了良好的隔离性和多态性,支持在同一个物理网络之上同时运行多种路由协议,确保了多态路由功能的实现.

3.3 转发性能分析

转发速率是衡量路由器系统的重要性能指标,在图 5 所示的测试场景的基础上,使用 Iperf3 网络性能测试软件来测试多态路由原型系统的转发速率,结果如图 6 所示.

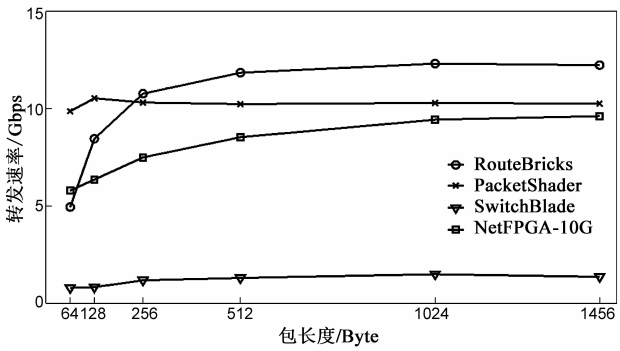

图6 不同长度数据包的转发速率

可以看出,多态路由原型系统的单端口最大转发

速率为9.6Gbps,最大转发速率接近于与理论最大值10Gbps,相较于基于 NetFPGA-1G 板卡实现的 SwitchBlade,系统转发性能有了较大提升. 同时,与采用 GPU 加速技术的 PacketShader 以及采用集群技术的 RouteBricks 相比,本文设计的多态路由原型系统与其具有接近的最高转发速率,这表明原型系统的设计在转发速率方面处于当前研究领域的前列,很好满足了未来网络对路由转发性能的要求.

3.4 QoS 保障性能分析

基于图5所示的测试场景,测试了三种路由方式下不同平均包长的业务流的 QoS 指标,结果如图7 图8所示.

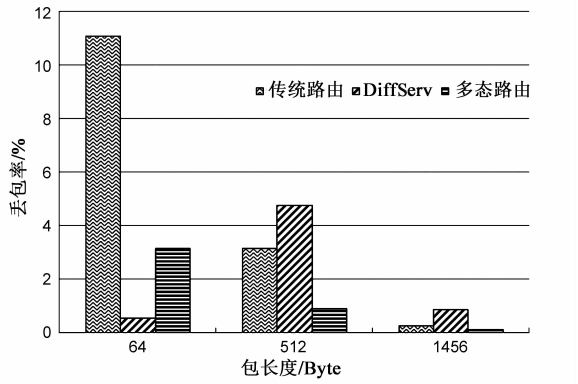

图7　三种路由方式下每种业务的丢包率

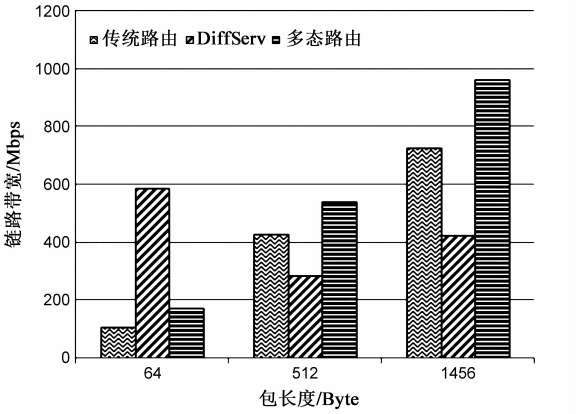

图8　三种路由方式下每种业务的实际传输带宽

实验结果表明,DiffServ 方式只保障了优先级高业务(64Byte)的传输带宽,其余两种业务的传输带宽低于传统路由方式,丢包率较高,只能保障单一业务的 QoS 需求. 而采用多态路由系统实现区分定制化多路径并行传输,减少了拥塞,相较于传统路由方式,业务流传输丢包率平均降低了73%,链路传输带宽平均提高了28%. 由此可见,多态路由系统通过对业务流进行区分,采用定制化路由传输方式,相较于 DiffServ 方式很好地提升了多业务共存时的服务质量.

4　结束语

针对当前路由机制由于结构僵化和功能单一,无法满足未来多样化业务需求的问题,本文提出了基于业务需求与路由协议自适配思想的多态路由模型,并采用 NetFPGA-10G 开放可编程硬件平台,以操作系统级虚拟化技术 OpenVZ 和可编程路由控制平台 Quagga 为核心,设计实现了一种支持多样化业务需求定制以及异构网络共存的多态路由原型系统. 相比于现有系统,该系统在支持业务定制路由服务路径的同时,在转发性能,服务质量保障方面均有提升,对于满足未来多样化的业务路由服务描述以及支持未来新型网络的渐进式试验与部署具有重要意义.

参考文献

[1] 张明川,许长桥,关建峰,等. 一种面向智慧协同网络的自适配路由策略研究[J]. 电子学报,2015,43(7):1249-1256.
Zhang Ming-chuan, Xu Chang-qiao, Guan Jian-feng, et al. Adaptive allocation routing scheme for smart and cooperative networks[J]. Acta Electronica Sinica, 2015, 43(7): 1249-1256. (in Chinese)

[2] McKeown N. Software-defined networking[J]. INFOCOM Keynote Talk, 2009, 17(2): 30-32.

[3] Van Jacobson, Diana K Smetters, James D Thornton, et al. Networking named content[J]. Communications of the ACM, 2012, 55(1): 117-124.

[4] 周烨,李勇,苏厉,等. 基于虚拟化的网络创新实验环境研究[J]. 电子学报,2012,40(11):2152-2157.
Zhou Ye, Li Yong, Su Li, et al. Research of network innovation experimental environment based on network virtualization[J]. Acta Electronica Sinica, 2012, 40(11): 2152-2157. (in Chinese)

[5] 罗腊咏,贺鹏,关洪涛,等. 可编程虚拟路由器关键技术与原型系统[J]. 计算机学报,2013,36(7):1350-1363.
Luo La-yong, He Peng, Guan Hong-tao, et al. Key technologies and prototype systems of programmable virtual routers[J]. Chinese Journal of Computers, 2013, 36(7): 1350-1363. (in Chinese)

[6] Kohler E, Morris R, Chen B, et al. The click modular router[J]. ACM Transactions on Computer Systems, 2000, 18(3): 263-297.

[7] Han Sangjin, Jang Keon, Papk Kyoung Soo, et al. PacketShader: a GPU-accelerated software router[J]. ACM SIGCOMM Computer Communication Review, 2010, 40(4): 195-206.

[8] Dobrescu M, Egi N, Katerina J, et al. RouteBricks: Exploiting parallelism to scale software routers[A]. Proceedings

of the 22nd ACM Symposium on Operating Systems Principles (SOSP'09)[C]. Big Sky, MT, New York, USA: ACM Press, 2009. 15 - 28.

[9] Anwer M B, Motiwala M, Tariq M B, et al. SwitchBlade: A platform for rapid deployment of network protocols on programmable hardware[A]. Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM'10)[C]. New Delhi, New York, USA: ACM Press, 2010. 183 - 194.

[10] 刘中金,李勇,杨懋,等. 基于可编程硬件的虚拟路由器数据平面设计与实现[J]. 电子学报,2013,41(7):1268 - 1272.
Liu Zhong-jin, Li Yong, Yang Mao, et al. Design on data plane of programmable hardware-based virtual router[J]. Acta Electronica Sinica, 2013, 41(7): 1268 - 1272. (in Chinese)

[11] 兰巨龙,程东年,胡宇翔. 可重构信息通信基础网络体系研究[J]. 通信学报,2014,35(1):187 - 198.
Lan Ju-long, Cheng Dong-nian, Hu Yu-xiang. Research on reconfigurable information communication basal network architecture[J]. Journal of Communications, 2014, 35(1): 187 - 198. (in Chinese)

[12] Kav'e Salamatian. Toward a polymophic future Internet: a networking science approach[A]. ITU-T Kaleidoscope 2010, Beyond the Internet? - Innovations for Future Networks and Services[C]. Pune, India: IEEE Press, 2010. 1 - 6.

[13] OpenVZ: Server Virtualization Open Source Project[OL]. http://www.openvz.org. 2007.

[14] Kunihiro Ishiguro. Quagga-A Routing Software Package for TCP/IP Networks[OL]. http://www.nongnu.org/quagga. 2012.

[15] NetFPGA-10G Project[OL]. https://github.com/NetFPGA/NetFPGA-public/wiki. 2014.

作者简介

张 岩 男,1991 年 1 月出生,河南舞钢人. 2013 年毕业于华中科技大学电子与信息工程系,其后进入国家数字交换系统工程技术研究中心攻读硕士学位,主要研究方向为新型网络体系结构、路由与交换技术.
E-mail: zy24987624@163.com

兰巨龙 男,1962 年出生,河北张北人. 国家数字交换系统工程技术研究中心总工程师、教授、博士生导师,主要从事新一代信息网络关键理论与技术的研究工作,目前作为首席科学家主持国家"973"项目"可重构信息通信基础网络体系研究".