基于可编程硬件的虚拟路由器数据平面设计与实现

刘中金,李 勇,杨 懋,苏 厉,金德鹏,曾烈光
(清华大学电子工程系,北京 100084)

摘 要: 未来网络体系结构创新和验证亟需建设虚拟化网络实验平台,虚拟路由器作为其中的核心组网设备,其结构和性能决定了实验平台的灵活性和承载能力.本文提出基于并行流水线的虚拟路由器数据平面结构,结合并行包分类和异步多指针轮询调度机制,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.本设计在可编程硬件上进行了原型实现,并结合商用及软件路由器在真实的网络环境中部署、测试与实验.实验结果表明与传统单流水线结构相比,本设计能以更高灵活性和并行性支持异构的路由器实例独立运行;在逻辑资源开销和延时特性未显著增加的情况下,并行虚拟路由器可以达到与硬件可比的线速转发能力.

关键词: 虚拟化; 数据平面; 并行流水线; 可编程硬件
中图分类号: TP393.1 **文献标识码:** A **文章编号:** 0372-2112 (2013) 07-1268-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2013.07.004

Design on Data Plane of Programmable Hardware-Based Virtual Router

LIU Zhong-jin, LI Yong, YANG Mao, SU Li, JIN De-peng, ZENG Lie-guang
(*Department of Electronic Engineering, Tsinghua University, Beijing* 100084, *China*)

Abstract: Building virtualized network experiment platform is considered to be an effective method for network architecture innovation and validation. The structure and performance of the virtual router determines the capacity and flexibility of network experiment platform. In this article, the virtual router's data-plane architecture with parallel pipelines is presented. Combined with parallel packet classification and asynchronous pointer polling scheduling mechanisms, we implement isolated heterogeneous router instances on the same physical underlying. Prototype system is deployed on programmable hardware which is tested with software routers in real network environment. Experimental results show that compared with traditional single-pipeline architecture, our design get greater flexibility and parallelism and supports heterogeneous router instances operating independently; logic resources overhead and delay characteristics are not significantly increased while each router instance achieves wire-speed forwarding which is comparable with that hardware.

Key words: virtualization; data plane; parallel pipeline; programmable hardware

1 引言

随着互联网中新业务的不断涌现与兴起,TCP/IP 架构的缺陷日益显现,如地址不足,安全性差,服务质量难以保障等问题暴露出来[1].为了解决这些问题,研究人员提出了 Clean-Slate 的网络创新思路,目的是通过网络体系结构的创新来从根本上克服 TCP/IP 模式的不足.在这一范畴内,NDN[2]、MobilityFirst[3]等体系结构已经得到了广泛的关注和深入的研究.为了验证新型网络体系结构的可行性和有效性,研究人员迫切需要高性能的网络创新实验平台来进行实际部署和评估[4].网络创新实验平台需要允许多样化的体系架构在其上并行部署,并且保证这些实验网络相互隔离,互不影响,以满足不同的研究人员的使用需求[5].

网络虚拟化是指在同一网络设备上同时运行多个逻辑上相互隔离的网络,它被认为是搭建网络实验平台的核心手段.虚拟路由器[4]是指可以同时运行多个独立虚拟路由器实例的网络核心设备,它是网络实验平台的重要组成部分,能够为网络实验平台提供虚拟化的节点和链路资源.虚拟路由器的灵活性、隔离性、转发和扩展能力决定了运行在实验平台上虚拟网络的性能.

E Kohler 等人提出通过软件来实现路由器 Click[6],

收稿日期:2012-08-08;修回日期:2013-02-22;责任编辑:李勇锋
基金项目:国家 973 重点基础研究发展计划(No.2013CB3291005);国家自然科学基金(No.61171065,No.61021001,No.61133015);国家 863 高技术研究发展计划(No.2013AA010601,No.2013AA010605)

并通过操作系统虚拟化来实现虚拟路由器,由于通用处理器在包转发速度上难以与专用设备相比,Click 面临转发速率慢的瓶颈;为了打破这一瓶颈,Han 等人利用 GPU 的高速并行计算能力来加速软件路由器[7],实验结果表明它可以达到 40Gbps 的转发速率.与软件实现的思路有所不同,Anwer 等人通过使用可编程硬件来构建多个路由表[8],采用动态选择路由表实现多个并行路由器实例.上述方案只能并行运行同构的路由器实例,难以支持异构网络和虚拟化能力的进一步扩展.

本文提出一种基于可编程硬件的并行流水线结构虚拟路由器,详细介绍了数据平面的设计和实现.通过在实际硬件和真实的网络环境下进行测试,验证了本设计在同一硬件底层中对异构的转发结构和并行虚拟网络的支持,且并行独立的虚拟网络都可以达到与硬件可比的峰值转发速率.

2　设计目标

本文设计目标是在数据平面中实现高效的虚拟化结构,使得异构网络体系结构在虚拟路由器中并行运行,相互隔离,在此基础上还要满足灵活性的要求以及高性能的转发能力和扩展性.

(1)高效的虚拟化能力及隔离性

实验平台中会承载大量并行网络实验,实验用户希望自己的实验过程和结果尽可能接近真实,在带宽占用、数据包处理等方面不能受到其他用户实验的影响.因此,虚拟路由器数据平面需要有一种高效的虚拟化方式,将属于不同网络的数据包进行分类、隔离和独立处理.

(2)异构网络支持能力

实验平台中的网络实验可能基于完全不同的体系结构,而这些新型体系结构之间并不类似,具有明显的异构特征,如 NDN 和 MobilityFirst 的命名、编址、路由等方式都不尽相同,它们具有不同的数据包处理过程.在网络体系结构的研究和实验过程中,研究人员会不断有新的想法和思路需要在平台中验证,实验环境和条件也会随之发生变化.虚拟路由器需要能够快速灵活地配置各个路由器实例的转发功能.

(3)高性能的转发能力

网络实验会经常运行实际的应用来测试算法或协议的可行性,它们使用不同的流量模型,在流量大小和模式上都有不同.为了满足应用的流量需求,虚拟路由器中运行的实例的性能需要能够达到与硬件转发速度相匹配的能力.

(4)具有良好的扩展性

实验平台规模扩展需要路由器的快速增量部署,路由器需要在转发性能上满足扩展性的要求,还要在虚拟路由器的虚拟化能力上具备动态扩展的能力,通过逻辑资源的增加或者物理设备的简单连接可以实现虚拟化个数的增加.

3　虚拟路由器及其数据平面设计与实现

本设计中虚拟路由器结构采用了数据平面和控制平面相分离的思想[9,15,16],它的结构主要包括两个部分:基于硬件的数据平面及基于通用处理器的控制平面.

数据平面主要实现虚拟化并根据控制平面配置的规则进行数据包的高速转发和处理,数据包由物理端口输入到数据平面,高速的包分类模块根据分类规则将各种数据包分离到不同的处理流水线中进行转发处理,最终确定输出端口,并从相应的物理端口发出.数据平面采用高效的包分类机制来实现虚拟化和隔离特性;通过多流水线设计来实现异构网络的支持能力,并提供可扩展的能力;在各个流水线中用户可以利用高速的查找处理算法来实现高速的包转发能力.

控制平面建立在通用处理器平台上,如 x86 或 ARM 平台等,通过操作系统虚拟化实现控制平面的虚拟化与隔离,控制平面负责路由协议和控制协议的处理.控制平面和数据平面包通过高速的总线来实现数据交互.

3.1　数据平面结构

虚拟路由器数据平面的整体结构如图 1 所示.

(1)输入队列:从物理端口,如以太网或光纤端口输入的数据包经过 PHY 芯片处理后由 MAC 协议解析为硬件可以处理的数据包,如 NDN、IP 等包格式.每个端口到来的数据包都在单独的队列中排队等待处理.

(2)并行包分类:在同一队列中排队的数据包可能从属于不同的虚拟网络,基于不同的体系结构,需要高速的包分类机制来决定各个数据包的目的流水线,以便于分别进行处理和查找操作.

假设每个端口速率为 R_i,所有队列的数据包都由统一的包分类模块进行处理,那么包分类器需要同时处理的速率为 ΣR_i 的数据包.以 10Gbps 的端口速率计算,如果有 4 个端口,那么包分类的速率就需要达到 40Gbps 的吞吐能力.这对于传统的包分类算法是一个很大的挑战,无法采用传统的基于 TCAM、Trie 或者空间分割的算法,并且需要较高的功耗和成本.

为了降低包分类的复杂度,充分利用可编程硬件所提供的并行性特征.本设计不使用单独的包分类模块,而是将包分类分散到各个队列上,如图 2 所示,数据包在输入调度之前有两条路径,一条是采用并行的包分类机制进行包分类,每个包分类模块负责单个队列的包分类,可以将包分类的速率降低到单队列的速度

上,这样可以利用传统的包分类算法.另一条路径用于缓冲包分类完成前的包头数据,因此,每个包分类模块可以最多处理 FIFO 深度的包头数据.以 IP 包为例,我们如果采用五元组做包分类依据,FIFO 采用 64-bit 的位宽,那么使用深度为 6 的 FIFO 就可以完成包头缓冲.

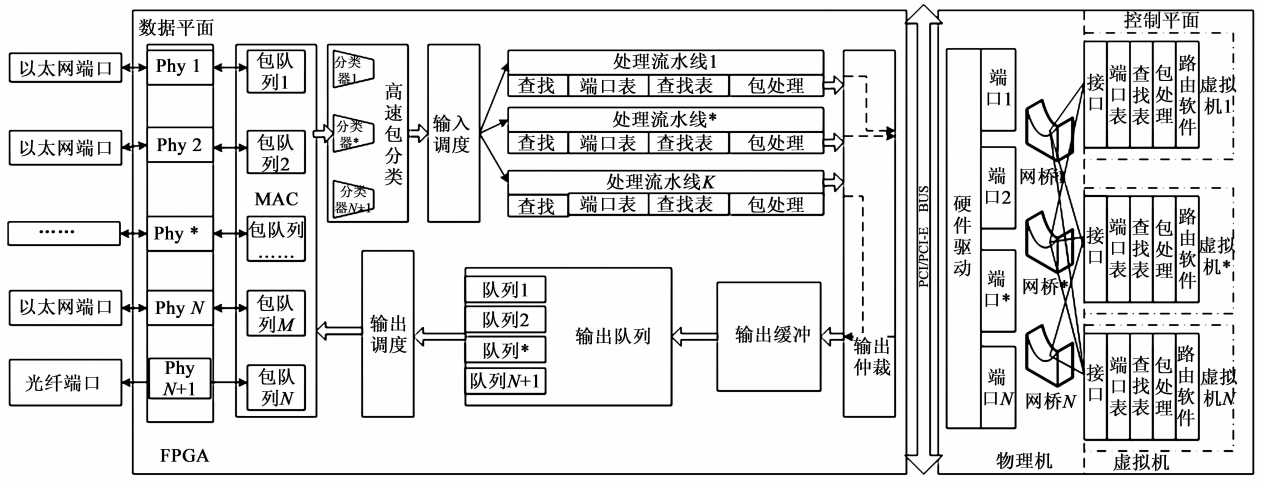

图1 虚拟路由器整体结构

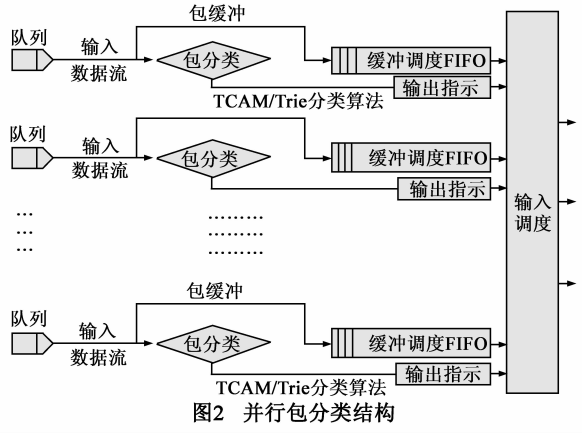

图2 并行包分类结构

(3)输入调度:数据包根据包分类规则确定处理流水线后,需要被高效地调度到处理流水线中.传统的方案有轮询、加权轮询等,这些调度算法可以保证输入的公平性.在本设计中,有多条流水线作为目标调度,因此需要对传统轮询方案进行改进,在未采用 VOQ 排队策略时,令轮询的指针数目与流水线数目一致.具体算法描述如下:

Step 1 设输入端口数为 N,处理流水线数为 $M(N>M)$.输入端口维护状态信息,$S_j(1\leqslant j\leqslant N)$来指示目的流水线状态,每条流水线维护一个输入指针 $P_i(1\leqslant i\leqslant M)$用来指示当前流水线对应的输入队列.初始状态下 $P_i=i, S_j=0$.

Step 2 如果输入队列对应缓存是空的,且输出流水线处于空闲状态下,那么每个输入指针 P_i 从当前位置开始,轮询选择第 1 个非空队列 k,如果 $S_k=i$,则 $P_i=k$.

(4)处理流水线:处理流水线是路由器中进行包转发和处理的关键部分.数据包处理的方式和灵活性也在于此,不同的流水线承载着不同的体系结构,这些应用对应的包处理流程不尽相同,在特定的处理流水线中,数据包的各个数据区域被解析出来,按照各体系结构的处理方法进行转发,查表处理.如基于 TCAM,Trie 或 Bloom-filter 等的查找方案都可在其中使用.

在单一流水线结构中,采用多表查询的方案需要在同构的查找表中选择目的表项进行查找,这些表的查找模块都必须是同样结构的或者相互串联的.如果需要改变虚拟路由器的功能,必须对这些结构的关键模块进行修改,这些修改必然会影响其他虚拟路由器的功能和性能,影响虚拟路由器的隔离性能.在我们的结构中,可以在不同的流水线中部署不同的应用方案和体系结构,如 IPv4 路由器可以和 OpenFlow[10] 的方案并存.

(5)输出仲裁:不同流水线中的数据包在转发查找完后需要统一地进行缓冲以及交换处理.本设计采用无状态信息的调度算法:通过轮询确定各流水线数据包的缓冲顺序;输出缓冲队列也是按照轮询的调度算法输出到各端口.

(6)输出缓冲:本设计采用输出排队的交换机制,它可以提供最理想的性能.为了达到 100% 的吞吐率,需要外围存储器达到所有端口的链路速率总和.

3.2 原型实现

NetFPGA[11]是一个线速、开源、可灵活配置的网络开发平台,被研究人员广泛应用来搭建和测试网络相关设计.但是,基于 NetFPGA 的设计经常受限于 FPGA

的容量,而且它必须结合总线与上位机进行通信,通信速率与灵活性都受到限制.为了克服这一问题,作者基于高性能的 Virtex-5 和 ARM 设计构建了一套可以独立运行的新的网络测试板卡.在无上位机支持的条件下,提供大规模的逻辑资源和通用处理器的灵活性.本文在 NetFPGA 和基于 Virtex-5 的板卡上都做了原型实现.

控制平面利用 OPENVZ 虚拟机实现.OPENVZ 是一种轻量级的虚拟化方案,它基于操作系统虚拟化技术,对比于全虚拟化和半虚拟化技术,它对内核的改动较小,具有更高的灵活性.路由协议软件选择开源的 zebra[12]来实现.控制平面的功能在文献[13]中有详细阐述.

4　性能分析

4.1　资源开销比较

并行流水线的设计必然会带来额外的硬件开销.为了测试本设计的资源利用率,分别在 NetFPGA 和 Virtex-5 的板卡上构建了 IPv4 的路由器,表 1 列出了不同路由器实例数目对应的逻辑资源利用率.

表 1　不同虚拟路由器数目的资源使用率比较

虚拟化数目	资源使用比例(NetFPGA/Virtex-5)		
	Slices	LUT	BRAM
1	55%/23%	49%/33%	53%/31%
2	79%/24%	64%/35%	53%/47%
3	91%/27%	77%/37%	60%/53%
4	96%/30%	88%/42%	67%/59%
10	不支持/41%	不支持/53%	不支持/66%

从结果可以看出,NetFPGA 平台最多可以支持 4 个路由器实例,Virtex-5 板卡则最少可以支持 10 个路由器实例,并可继续扩展.与未虚拟化(虚拟数目为 1)的情况相比,随着虚拟化数目增加,虚拟路由器的资源开销并不是倍增,而是以一个较小的比例递增的;而且资源开销的增长速度随数目增加逐渐降低.

4.2　异构支持及隔离性比较

为了测试虚拟路由器的隔离特性,采用图 3 中的实验拓扑.实验使用 4 个虚拟路由器节点,其中有两个 NetFPGA 节点,一个 Virtex-5 板卡节点和基于软件搭建的 Click 节点.每个节点通过虚拟化虚拟出两个虚拟路由器实例,其中一个路由器实例基于 IPv4 协议,另一个路由器实例运行 802.1Q 协议.IPv4 路由器与无线路由器连接,运行 RIP 协议,无线手持设备通过无线路由器接入该网络;在 802.1Q 网络中运行 OSPF 协议.

实验表明:两个虚拟网可以独立运行各自的协议和转发功能,生成相应的路由表.手持终端可访问 IPv4 中的软硬件网络节点,但是无法访问 802.1Q 网络中节点.结果证明本设计实了对异构体系结构的支持和完全的隔离特性.

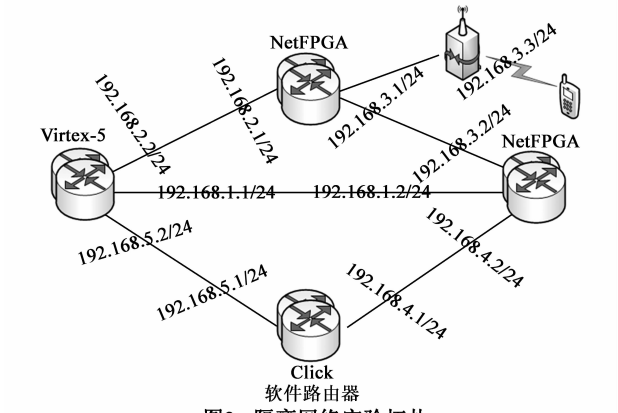

图3　隔离网络实验拓扑

4.3　转发性能

为了测试多个流水线的转发性能,我们在 NetFPGA 中分别虚拟化一个 IPv4 路由器和一个基于 802.1Q 的路由器,并利用基于 NetFPGA 的 Packet Generator[14]作为流量产生工具,分别向两个端口分别发送 1Gbps 的 IPv4 包和 802.1Q 包.测试结果如图 4 所示.

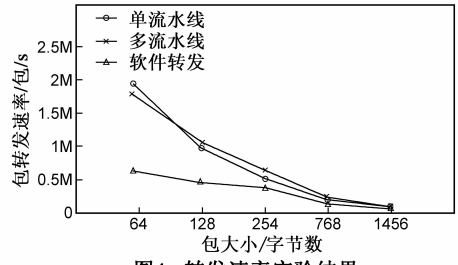

图4　转发速率实验结果

从图中来看,虚拟化后的路由器转发速率(多流水线)与未虚拟化(单流水线)的数据转发速率基本是可比的,都可以达到 1Gbps 线速转发速率.当包长较小时(<254 字节),硬件的转发速率要明显高于软件的转发速率.

包分类和调度的过程必然会引入额外的延时,我们以 ICMP 协议的延时来比较,从表 2 来看,虚拟化后的转发延时增加了 44%;为了测试转发速率对网络应用的影响,我们还测试了图 3 拓扑中无线手持终端通过两个硬件节点转发到 Click 路由器端口的延时.无线手持终端与路由器的距离在 10m – 40m 的范围变化,相应的转发延时在 25ms – 90ms 这一范围内变化;比较虚拟化前后的数值变化,可以看出虚拟化造成的转发延时的增加并不能影响到全局的延时.这也说明虚拟化所增加的延时对网络性能或应用的影响十分有限.

表 2 延时比较(5 次平均)

类型	回环延时	转发延时	无线转发延时
单流水线	0.03ms	0.25ms	25ms – 90ms
多流水线	0.03ms	0.36ms	25ms – 90ms

5 结论

本文提出了一种基于可编程硬件的虚拟路由器数据平面结构,并详细阐述了其实现方法.实验结果表明本设计可支持多个异构的虚拟路由器实例,它们之间具有完全的隔离特性,每个虚拟路由器实例可以达到线速的转发速率.

参考文献

[1] N M Mosharaf, K Chowdhury, R Boutaba. A survey of network virtualization[J]. Computer Network, 2010, 54(5): 862 – 876.

[2] Named Data Networking [OL]. http://www.named-data.net, 2012 – 02 – 22.

[3] MobilityFirst Project [OL]. http://mobilityfirst.winlab.rutgers.edu, 2012 – 02 – 22.

[4] T Anderson, L Peterson, S Shenker, J Turner. Overcoming the Internet impasse through virtualization[J]. Computer, 2005, 38(4): 34 – 41.

[5] 周烨,李勇,苏厉,金德鹏,曾烈光.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, Jin De-peng, Zeng Lie-guang. Research of network innovation experimental environment based on network virtualization[J]. Acta Electronica Sinica, 2012, 40(11): 2152 – 2157. (in Chinese)

[6] E Kohler, R Morris, B Chen, J Jannotti, M F Kaashoek. The click modular router[J]. ACM Transactions on Computer Systems, 2000, 18(3): 263 – 297.

[7] S Han, K Jang, K S Park, S Moon. PacketShader: a GPU-accelerated software router[A]. Proceedings of ACM SIGCOMM[C]. New Delhi: ACM, 2010. 195 – 206.

[8] M B Anwer, N Feamster. Building a fast, virtualized data plane with programmable hardware[A]. Proceedings of the 1st ACM Workshop on Virtualized Infrastructure Systems and Architectures[C]. Bacelona: ACM, 2009. 1 – 8.

[9] Forwarding and Control Element Separation (ForCES) Framework[OL]. http://www.ietf.org/rfc/rfc3746.txt, 2012 – 02 – 22.

[10] OpenFlow Switch[OL]. http://www.openflowswitch.org, 2012 – 02 – 22.

[11] NetFPGA Project[OL]. http://www.netfpga.org, 2012 – 02 – 22.

[12] Zebra Project[OL]. http://www.zebra.org, 2012 – 02 – 22.

[13] 杨懋,刘中金,李勇,曾烈光,金德鹏,苏厉.基于可编程硬件的虚拟路由器控制平面[J].清华大学学报(自然科学版),2012,52(5):586 – 591.
Yang Mao, Liu Zhong-jin, Li Yong, Zeng Lieguang, Jin Depeng, Su Li. Control plane of a programmable hardware-based virtual router[J]. Tsinghua Science and Technology, 2012, 52(5): 586 – 591. (in Chinese)

[14] G A Covington, G Gibb, J W Lockwood, N McKeown. A packet generator on the NetFPGA platform[A]. Proceedings of the 17th IEEE Symposium on Field Programmable Custom Computing Machines[C]. Napa: IEEE, 2009. 235 – 238.

[15] 徐明伟,江学智,陈文龙.路由器分布式控制研究综述[J].电子学报,2010,38(8):1892 – 1899.
Xu Ming-wei, Jiang Xue-zhi, Chen Wen-long. Survey on distributed control in a router[J]. Acta Electronica Sinica, 2010, 38(8): 1892 – 1899. (in Chinese)

[16] 彭来献,田畅,路 欣,郑少仁.一种支持多优先级的高速 Crossbar 调度算法[J].电子学报,2004,32(8):1305 – 1309.
Peng Lai-xian, Tian Chang, Lu Xin, Zheng Shao-ren. A new scheduling algorithm supporting multi-priorities for high-speed crossbars[J]. Acta Electronica Sinica, 2004, 32(8): 1305 – 1309. (in Chinese)

作者简介

刘中金 男,1988 年 12 月出生,山东聊城人.2009 年毕业于北京理工大学电子工程系,其后进入清华大学电子工程系攻读博士学位,从事可编程虚拟化路由器方面的有关研究.
E-mail: zj-liu09@mails.tsinghua.edu.cn

李 勇 男,1985 年出生,湖南长沙人,2007 年于华中科技大学获得工学学士学位,2012 年在清华大学电子工程系获得工学博士学位.主要研究领域为未来网络、下一代 IP 网络体系结构、移动管理、移动容迟网络、网络虚拟化等.