

存转算一体的多模态网络共性平台技术研究



PINet Data Plane Platform Technology for Storage, Forwarding and Computing Integration

董永吉/DONG Yongji, 胡宇翔/HU Yuxiang, 崔鹏帅/CUI Pengshuai

(解放军战略支援部队信息工程大学, 中国 郑州 450002)
(PLA Strategic Support Force Information Engineering University, Zhengzhou 450002, China)

DOI: 10.12142/ZTETJ.202201006

网络出版地址: <https://kns.cnki.net/kcms/detail/34.1228.TN.20220223.2024.004.html>

网络出版日期: 2022-02-25

收稿日期: 2021-12-19

摘要: 面向多模态网络需求, 提出了一种存转算一体化的数据平面共性网络平台, 并介绍了该平台的框架组成及关键技术。通过软硬件协同、异构融合的设计优势, 该平台充分发挥了数据平面的灵活性与可扩展性, 支持多样化业务可定义的转发和处理, 满足多模态网络发展中存储、转发与计算融合的需求。

关键词: 多模态网络; 可编程数据平面; 异构融合

Abstract: For the multimodal network requirements, a data plane platform integrating storage, forwarding, and computing is proposed, and the framework composition and key technology are introduced. The platform further releases the flexibility and scalability of the data plane through the coordination of software processing and hardware processing. The advantages of heterogeneousness and integration support diversified network applications with definable forwarding and processing, which can cope with the urgent needs of integration of storage, forwarding, and computing in the polymorphic networks.

Keywords: polymorphic network; programmable data plane; heterogeneous fusion

当前互联网发展进入了新时代, 随路计算、确定性网络、低时延网络等新型网络技术的出现加速了工业生产、社会生活与网络的深层次融合。当前互联网的基础架构与技术体系, 在智慧化、多元化、个性化、高鲁棒、高效能等方面面临重大挑战, 亟须变革网络基础架构并构建全维可定义的多模态智慧网络^[1]。

然而, 传统单一面向尽力而为转发功能的网元设备结构, 既无法满足 5G、物联网等的业务流量多样化处理需求, 也无法满足低时延网络、确定时延网络等细粒度服务需求, 以及网络新业务的快速部署, 更无法支持网络体系创新, 所以网元设备的困局成为网络发展的瓶颈^[2]。

为了应对这一挑战, 学术界和工业界不断推出新技术和新理念。2008 年, 斯坦福大学提出了软件定义网络 (SDN) 及 OpenFlow 技术^[3], 将传统刚性封闭网元设备结构中的数据平面与控制平面解耦。分离出的应用平面和控制平面通过编

程化来应对不同场景的需求, 并采用标准化的 OpenFlow 协议对网元设备进行配置和管理, 提升了网络管控的灵活性, 一定程度上增强了部署新业务的能力。但由于 OpenFlow 协议定义能力有限, 导致为了支持新协议或新业务, 该协议需要向下兼容地扩展内容。于是该协议从最早的 1.0 版本的 12 个匹配域, 不断扩充到 1.5 版本的 45 个匹配域。这种刚性补丁扩容式的协议支持方式, 无法满足灵活可定义的新协议需求。同时, 每一次协议版本的升级和扩展, 都会导致数据平面和控制平面设备的重新研制。这样一来, 新技术的应用时间和开发成本则无法有效缩减。

针对 SDN 数据平面网元设备协议处理可扩展性差的问题, 协议无关转发 (POF) 研究引起业内的广泛关注, POF^[4]、P4 语言及协议无关的交换机架构 (PISA)^[5]先后被提出。其中, P4 具有与协议无关、可重配置性和平台无关性三大特点, 缓解了 SDN 编程能力不足以及可扩展性差的状况。尤其是数据平面的可编程性, 推动了虚拟扩展局域网 (VxLAN)、基于用户数据报协议 (UDP) 的低时延的互联网

基金项目: 国家重点研发计划项目 (2019YFB1802502)

传输层协议（QUIC）、网络带内遥测等新协议与功能的快速定制与部署，但P4技术规范^[6]仅仅描述了数据转发的模型。虽然支持面向转发的功能可定义，但仍无法满足数据包的存储、转发、计算协同处理复杂场景应用需求。因此，为了提供更好的网络传输服务，业界开始在数据平面上整合现场可编程门阵列（FPGA）、图形处理器（GPU）等异构加速单元^[7]，通过本地的硬件加速技术实现业务功能的卸载，通过网元实现数据的加速，从而实现整个网络性能和服务质量。

虽然数据平面在硬件可编程方面不断发展，但网元设备的设计却依旧停留在封闭结构上，没有发挥可编程硬件灵活可扩展的能力。针对此，底层可编程硬件与上层软件功能和管理配置解耦的白盒结构被广泛采纳。白盒设备的硬件兼容多种转发芯片的接口，并遵循开放计算项目（OCP）标准化规范设计。同时，软件的平台操作系统及承载的各种网络应用也采用规范化的接口，实现底层硬件与软件的独立设计。例如，微软打造的开源网络交换机操作系统SONIC^[8]，采用的交换机抽象接口（SAI），解耦了软硬件的实现细节，但是该系统接口与协议紧耦合，无法适配和部署新的业务。为了解决这个问题，谷歌提出了Stratum系统^[9]，旨在配合硬件的可编程特性，实现一个完全可编程的数据平面网元设备。为此，该系统集成了一个协议无关转发接口P4Runtime，在开放网络操作系统（ONOS）控制器的管理下，可以灵活转发部署新型协议；但是当前网元操作系统仅可以实现对交换芯片的抽象，硬件异构的加速单元没有纳入其管理范畴。

与此同时，为了更好地支持数据平面的可编程，业界开展了多种编译器的研究工作，尤其针对交换芯片、FPGA或软件交换机的P4语言编程，以及多种异构芯片的整体硬件编译方法设计^[10]。Intel公司的tonifo芯片^[11]支持基于商用的软件开发环境（SDE）开发，但是由于其不开源的特性，无法支持用户可扩展的、定制化的功能编译。在基于FPGA的编译器及结构设计方面，业界提出了多种解决方案^[12-13]，但这些方案主要集中在转发功能的实现上，没有充分利用FPGA的可编程特性。另外，面向交换机软件模型（BMv2）的软件编译结果限于性能瓶颈^[14]，无法直接应用于高速流量场景。而当一个网元设备同时包含多样的可编程目标器件时，由于缺乏一个统一调度的协同编译环境，编译器无法有效糅合异构资源的优势，因此不能达到高效、灵活的编译效果。

1 多模态网络的提出

针对现有网络架构存在的结构僵化、IP单一承载、难以抑制未知威胁等问题，我们从网络构造的角度来提升网络的

功能、性能、效能、安全等，将“结构可定义”贯穿网络的各个层面。邬江兴院士提出了一种网络各层功能多模态呈现的网络架构——全维可定义的多模态智慧网络（PINet）^[2]。PINet是一种技术体制与物理平台分离的网络发展范式，如图1所示，PINet将各种网络技术体制以模态的形式在多模态网络环境上动态加载并运行，按照模态自定义的报文格式、路由协议、交换方式、转发逻辑等进行处理，实现多种模态在同一物理网络平台上的共存、演进或变革发展。

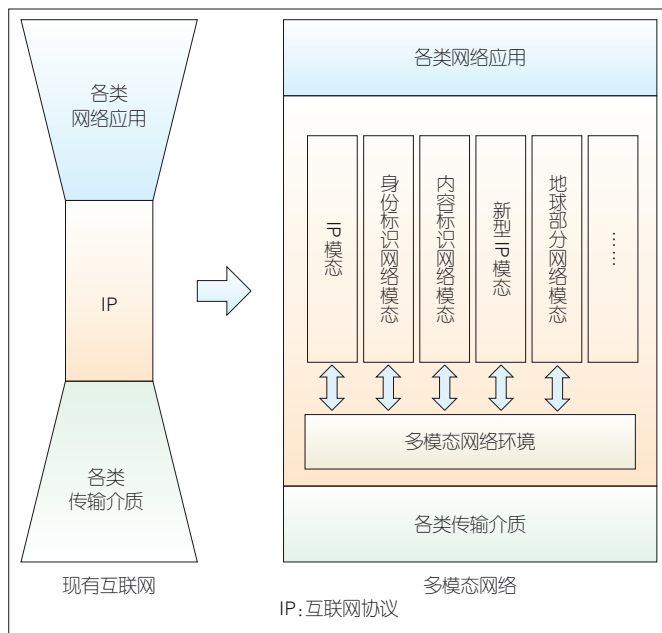
在多模态网络逻辑框架中，多模态网络环境是多模态网络的核心，而网元设备是支撑多模态网络环境的基石。然而，当前网元设备却无法满足不同模态网络环境的构建需求。因此，我们需要一个共性平台来重新整合数据平面中存储、转发和计算，支持自定义模态报文的解析、转发及交换处理。

2 多模态网络共性平台的技术特征

为应对新协议和应用部署的挑战，推动新型网络结构的演进发展，网元设备需要提供灵活可扩展、高性能和安全的支撑。本文梳理出面向存转算一体化的数据平面共性网络平台的技术特征，如图2所示。

（1）新型标识和报文的自定义解析与处理

面向定制化、个性化服务承载的需求，共性网络平台需要支持用户自定义接入标识结构和数据报文结构，并按照自定义逻辑进行报文处理，进而支撑传输协议、寻址路由等的全维度可定义。其中，寻址路由体现为基于互联网协议



▲图1 多模态网络逻辑框图



▲图2 共性网络平台的技术特征

(IP)、身份、内容、地理空间等标识的多种寻址路由方式与机制，传输协议体现为面向功能、场景、业务等需求的各种网络协议。通过支持各种新型网络机制的互联互通、协同组合，可以提高网络服务的多元化能力和对于用户需求的个性化适应能力。

(2) 存储、转发与计算一体设计

网络超融合、边缘计算、随路计算等日益多样化的业务功能需求对共性网络平台提出了更大的挑战。在满足共性网络平台支持灵活转发能力的基础上，在数据平面引入存储、转发与计算一体的、可定义的复合流水线结构设计，实现存储、转发、计算3种资源在数据转发的过程中灵活组合、调用，适配不同应用场景下特质化的处理需求。

(3) 内生安全构造

在由内生安全构造的共性网络平台上，“动态异构冗余”的设计思想已被引入数据平面。该平台以异构处理组件构成的元功能池为基础，生成多种等价异构执行体，并通过多模裁决和负反馈调度机制，实现内生防御的安全机制，以应对平台在软硬件设计过程中不可避免的后门及漏洞等安全威胁。该平台能够有效抵御各种病毒/木马等已知或未知的威胁，并通过将网络空间安全能力由“外挂”转变为“内生”，从而实现“高可信、高可用、高可靠”三位一体的网络安全服务。

(4) 面向异构软硬件资源的协同编译

为了提供共性网络平台上多种异构资源的可编程性和通用性，降低设备开发的难度，缩短面向特定场景的应用部署

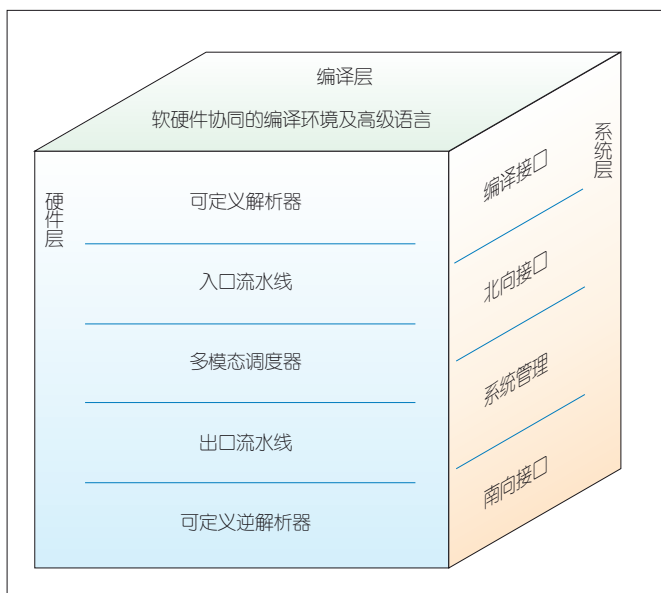
时间，我们需要提供一个“自顶向下”的由高级编程语言和编译器组成的编译环境。该编译环境能够面向多种异构软硬件功能组件构建统一编程模型，提供用户高级语言编程接口，屏蔽底层硬件支持协同调度各类型软硬件资源的细节，并通过单纯的数据包存储、转发、计算等处理逻辑的形式化描述，实现网络处理性能和功能灵活性之间的平衡。

(5) 多模态混合交换调度

在差异化的应用场景下，特定业务流的突发性、包长度、流量大小和速率特性有所不同。与此同时，不同的确定时延网络、低时延网络等新型网络技术体制的数据报文交换指标也不尽相同。每一种新型网络技术体制可被视为一种新型的模态。多模态混合调度技术可以为每一种模态提供定制化服务与质量保障，还可以均衡多种模态间的公平/优先级交换的策略，提升交换带宽资源利用率，保障网络交换服务质量。

3 共性网络平台的分层结构

共性网络平台的分层结构模型如图3所示，整体上分为硬件层、系统层和编译层。3个层面相互依存、相互支撑。其中，硬件层是整个平台的底层支撑，通过多种异构资源的叠加，支撑数据包存储、转发和计算的一体化处理；系统层实现异构资源的协同，以及各种网络功能的控制和管理；编译层实现对整个共性网络平台功能的编译，通过将高级语言描述的自定义功能映射到平台上，实现多样化业务的异构接入、交换和传输。



▲图3 共性网络平台架构图

3.1 硬件层逻辑设计

硬件层的处理逻辑如图4所示，主要分为5个组件：可定义解析、入口流水线、调度器、出口流水线和可定义逆解析。5个组件都引入可编程能力，可根据高级语言形式化描述的编译结果，灵活重构出定制化的服务功能。

可定义解析/逆解析组件可通过编译器接口进行配置。可定义解析组件支持按照用户定义的任意协议字段进行解析，并提取对应的字段作为入口/出口流水线中流表的关键词进行多域匹配；可定义逆解析组件支持按照用户自定义的任意数据格式进行封装，实现协议包头内容的增添、修改和转换功能。

入口/出口流水线组件采用融合异构处理资源的多级流水线设计，在支持类似P4可定义的转发流水线的基础上，将计算和存储功能挂载到并行的流水线上，再通过流水线的调用支持多种处理能力的融合，并采用协议无关的配置接口转发流表和动作信息，为不同的模态和协议流分配不同的流水线资源。这样可使得不同的模态和协议相对独立、并行地在平台上运行，而每种流表都可以基于用户定义的关键词进行构建，其匹配的动作集为存储、转发以及计算3种类别的操作集合。

调度器可以通过调度接口配置，实现面向流表的定制化的调度，并通过配置平台流水线上各个队列的属性、带宽保障、优先级和调度策略，实现精细化、差异化的交换能力服务。

3.2 系统层逻辑设计

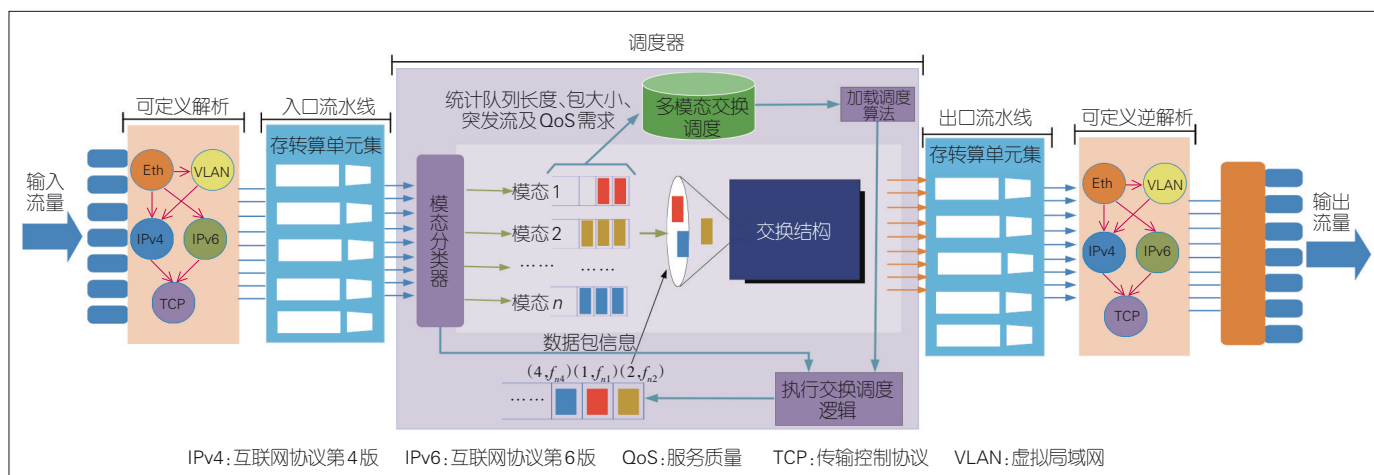
系统层实现对硬件层异构资源的管理，以及与控制器间的互通，系统逻辑如图5所示，主要包括编译接口、北向接口、系统管理和南向接口4个部分。编译接口为编译层与系

统层的接口，支持编译结果的配置与下发，实现异构资源的可定义重构。北向接口为共性网络平台与控制器之间的配置管理接口，用于接收控制平面的控制器生成的各种流表信息，引导和决定共性网络平台的数据处理行为，并采用协议无关的协议交互机制，支持用户自定义结构和流表信息的传递。系统管理是形成共性网络平台能力的功能集，采用内生安全构造核心节点功能，以动态、随机改变核心功能的静态性、确定性，进而提供内生安全的防护能力。系统管理由运维管理和异构多维资源管理两部分功能组成，通过容器化构建、模块间联动解耦的方式，实现了对硬件层异构多维资源的管理，以及控制器配置信息的维护。南向接口为硬件层和系统层的接口，是对交换芯片、FPGA、多核等异构芯片接口的统一抽象，对上提供一系列标准化的应用程序编程接口(API)，使系统层功能不再关心硬件层中异构芯片的硬件细节，并采用统一的方式接口进行管理和配置。

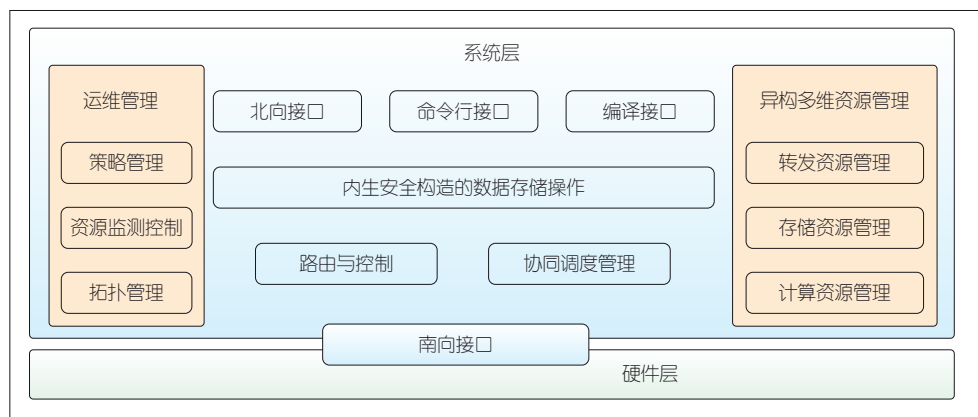
3.3 编译层逻辑设计

编译层逻辑组成如图6所示，主要包括存储、转发、计算一体化的编译器框架和高级可编程语言设计两部分。针对存储、转发、计算一体化的异构特点，编译层对数据平面关键要素进行提取，抽象出一套融合存储、转发和计算的数据平面操作指令集，用于定义和描述数据平面的行为（包括控制原语、存储原语、计算原语、转发原语），更加灵活地描述数据平面，实现数据平面与控制平面的解耦和接口的标准化。

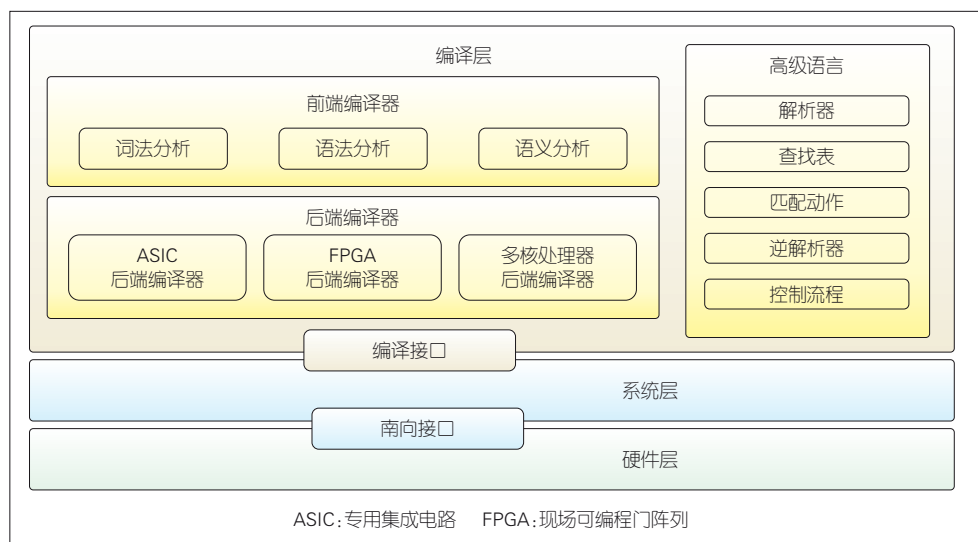
基于开放式结构设计，编译器框架参考了P4编译器前端解耦分离的设计。前端编译器扩展支持存储、转发和计算功能的统一编译；而后端编译器支持多样化的编译目标，以及可扩展的目标结构，并能设计编译流程与仿真验证环



▲图4 硬件层逻辑结构图



▲图5 系统层逻辑组成图



▲图6 编译层逻辑组成图

境，更好地支持编译器的发展与演进。前端编译器侧重于通用基础的编译功能，主要实现编译过程的结构词法分析、文件语法分析以及段落语义分析3个功能：词法分析主要将待编译的源文件按照语法分割为独立的标记和单词，而源文件中的制表符、空格等编码无效字符会被替换并删除，并根据注释中的相关辅助类语法将语法标记或词组分类；根据语法模板，语法分析可以从语法角度判断不同分组间代码结构的正确性，并生成抽象表达；语义分析针对整个源文件的含义进行分析，排查逻辑漏洞，展开嵌套循环，并生成中间表达形式（IR）。后端编译器面向独立目标器件构建，基于前端生成的中间表达形式，结合具体的芯片属性生成最终的目标文件。后端编译器支持多种芯片类型，如ASIC交换芯片、FPGA芯片、x86多核芯片、ARM多核芯片等。所有类型的后端编译器采用共享通用可扩展的接口，支持新型编程器件的可扩展和现有功能器件的可演进。

面向该框架设计，高级可编程语言统一抽象描述异构编

程对象的功能特征，主要包含5个要素：解析器、查找表、匹配动作、逆解析器和控制流程。其中，解析器通过定义并有序描述协议的特征，指导硬件层中可定义解析组件按照一定的逻辑解析数据报文；查找表用于构建硬件层中入口/出口流水线组件流表的匹配关键词和掩码的长度以及匹配模式方法，如最长匹配、精确匹配范围匹配等；匹配动作定义查找表匹配之后的动作模式，分别包括面向数学运算的计算类动作、面向本地存储的存储类动作、面向可定义流转的转发类动作；逆解析器通过定义有序的协议字段组合，指导可定义逆解析组件实现数据报文的协议的重组和封装；控制流程定义数据报文在硬件层的顺序及判断跳转执行逻辑，进而指导硬件层完成对进入系统中数据报文的处理流程。

4 结束语

本文提出了一种支持存储、转发、计算一体的多模态共性网络平台结构。该结构通过融合异构芯片功能构建数据平面的处理能力，采用可扩展的操作系统管理异构资源，从而提供统一的软硬件协同编译环境，形成存储、转发、计算一体的系统平台能力，支撑多模态网络中新型模式、协议和业务的融合、演进与发展。

参考文献

- [1] 李军飞, 胡宇翔, 伊鹏, 等. 面向2035的多模态智慧网络技术发展路线图[J]. 中国工程科学, 2020, 22(3): 141-147
- [2] 胡宇翔, 伊鹏, 孙鹏浩, 等. 全维可定义的多模态智慧网络体系研究[J]. 通信学报, 2019, 40(8): 1-12
- [3] MCKEOWN N, ANDERSON T, BALAKRISHNAN H, et al. OpenFlow: enabling innovation in campus networks [J]. ACM SIGCOMM computer communication review, 2008, 38(2): 69-74. DOI: 10.1145/1355734.1355746
- [4] SONG H. Protocol-oblivious forwarding: unleash the power of SDN through a future-proof forwarding plane [C]//Proceedings of the Second ACM SIGCOMM Workshop on Hot Topics in Software Defined Networking.

下转第74页➡