微服务架构下的算力路由技术

Computing-Power Routing Technologies Under Architecture of Micro-Service

陈晓/CHEN Xiao[1,2]， 黄光平/HUANG Guangping[1,2]

（1. 中兴通讯股份有限公司，中国 深圳 518057；
2. 移动网络和移动多媒体技术国家重点实验室，中国 深圳 518055）
(1. ZTE Corporation, Shenzhen 518057, China;
2. State Key Laboratory of Mobile Network and Mobile Multimedia Technology, Shenzhen 518055, China;)

DOI：10.12142/ZTETJ.202201014
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.tn.20220217.1717.006.html
网络出版日期：2022-02-18
收稿日期：2021-12-26

摘要：从算力网络的目标架构和愿景出发，研究微服务集群架构下的端到端路由技术解决方案，聚焦算力路由穿透集群L4-L7代理节点并进入服务级颗粒度的场景。在确保与现网平滑兼容前提下，从协议转控面角度分析IPv6段路由（SRv6）和虚拟可扩展局域网（VxLAN）的增强算力路由解决方案。

关键词：算力路由；微服务；SRv6；VxLAN

Abstract: From the viewpoint of the designed and envisioned computing power networking architecture, an end-to-end routing solution under the architecture of micro-service is proposed, which focuses on extending the L3 routing to the computing service within the micro-service cluster. Enhanced segment routing IPv6 (SRv6) and virtual extensible local area network (VxLAN) computing-power networking solutions have been analyzed and presented in detail with the principle of smooth compatibility with the ongoing commercial network architecture.

Keywords: computing-power routing; micro-service; SRv6; VxLAN

在全行业数字化转型升级的宏观背景下，继通信网络之后，算力成为至关重要的数字化产业基础设施。在5G及后5G时代，算力向边缘乃至超边缘下沉，已经成为行业趋势。同时，随着行业算力需求的多样化和终端算力的增强，算力的泛在化将成为新的行业形态。与通信网络不同，各种算力（尤其是异构算力）之间并无统一的架构和体系，缺乏协同机制。这导致算力成为“孤岛”，泛而不强，强而不专。因此，基于公共通信网络的泛在连接，将端、边、云的泛在算力有效协同起来，使之形成统一、动态、智能的算力资源池，成为算力网络的重要目标之一[1]。

算力可分为两类：一类为基础算力，如中央处理器（CPU）、图形处理器（GPU）、数据处理器（DPU）、专用集成电路（ASIC）等，属于静态算力资源；另一类为服务算力，如算法、功能等通用服务级算力，这类算力直接面向业务数据，属于动态算力资源。算力、算法和数据构建了有机的整体。为了发挥异构算力的最大算力效能，不同的算力将采用不同的算法来处理不同类型的数据。在算力网络目标架构下，上述两类算力均被统一感知、统一调度、统一路由，从而连算成网，形成一个层次化的统一算力资源池。

在当前云网业务模式下，算力和网络独立部署、独立运营、独立服务。用户分别向云服务和网络服务提供商提交服务申请，构建服务合同，组合实现完整的应用服务。这种算网分离模式催生了互联网的繁荣，导致算力和网络服务粗放式交付模式的产生，造成巨大的资源浪费。因此，算力网络的另一个重要目标是算网深度融合，即算力和网络服务在一个平面、一个接口、一个路由策略中进行。

综上所述，网络需要将传统的感知和路由向层次化算力方向延伸，从而构建一个基于通信网络的算力和网络资源全网视图，并以此作为全新的业务交付平台，在大幅提升算力和网络资源效率的同时，为行业提供更加丰富、高效的算网融合业务能力，进而赋能信息通信技术（ICT）深度融合的全行业数字化转型升级。

1 微服务架构下的服务路由和寻址机制现状

将应用程序解耦成独立的子服务集群，并分别开发、测试、维护和交付，是微服务架构及其部署和运营模式。微服务架构是基于以更加灵活、更易扩展为主要原则的全新应用部署模式的，服务网格内部的交互和通信由应用网关在L7

层统一执行。应用仍然是最小可访问、可调度的资源颗粒度。微服务以及服务网格用户均不可见，网络也无法感知和路由。应用网关将作为应用代理终结L3层路由流量。

基于层次化算力资源感知和路由的算网一体路由机制，是算力网络架构的重要特征。而应用是算力资源的服务对象，并不是算力网络本身调度和路由的对象。因此，以应用为颗粒度的微服务部署架构，对网络屏蔽了算力服务并终结了L3层网络流量。这是端到端算力路由面临的一个行业现实问题。算力网络的资源调度，改变了传统网络以端口地址为对象的数据通信转发，它是以应用和资源为服务对象的资源匹配。应用从传统的向平台要资源转变为向网络要资源。作为应用从单体到微服务解耦的架构演进，微服务也使得网络为分布式算力协同提供网络能力支撑成为可能。因此，核心问题就是，如何打破以数据通信转发为主的网络功能与以应用为主的微服务协同功能之间的界限，让网络能够看到应用内部，从而更好地为应用服务[2]。

1.1 端到端算力路由面临的分段微服务路由挑战

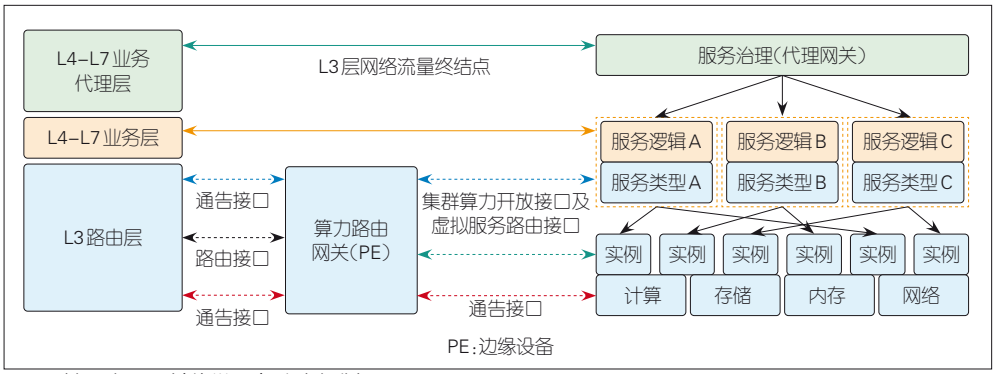
端到端微服务路由被应用网关分隔为独立的两段：一段为终端到应用网关的L3层路由，一段为应用网关到微服务实例的局部服务路由。其中，后者往往是L7层流量路由。如图1所示，应用代理网关终结L3层流量路由，即全网算力路由的最小颗粒度将被作为应用，或者微服务集群被作为应用单位。集群内的微服务仅限于局部算力调度和路由，无法执行全局算力资源协同。虽然如此，微服务集群内的基础算力资源、微服务种类及其实例状态，仍然有可能被外部网络感知。外部网络基于这类算力资源状态执行跨微服务集群路由。网络虽然可感知微服务并基于动态状态执行微服务集群路由，但是无法执行微服务的调度和路由本身。当然，这类粗颗粒度的微服务集群间路由机制也可以在集群资源调度和管理中心完成，并由后者感知和维护微服务集群内的算力服务和资源状态，从而实现基于L7层的端到端微服务路由。

1.2 算力路由面临的跨池虚机组网和寻址机制挑战

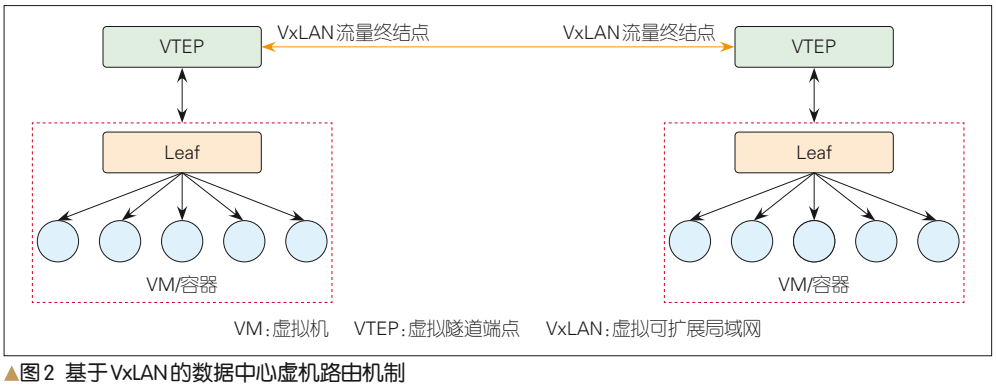
相对于虚拟局域网（VLAN），虚拟可扩展局域网（VxLAN）拥有更加庞大的寻址空间和更加灵活的组网机制，已经成为数据中心内部及数据中心之间虚机组网和寻址的主流机制。VxLAN是基于用户数据报协议（UDP）的L2层模拟网络技术。尤其是在跨数据中心虚拟网络组网场景中，虚拟隧道端点（VTEP）作为数据中心虚机集群代理，在UDP层即L4层终结了L3层网络路由流量。如1.1节所述，在跨数据中心虚机集群之间的微服务路由场景下，VTEP同样将端到端L3层微服务路由分隔成内外相互独立的两段：一段为数据中心之间的L3层外网路由，一段为数据中心内的L3层内网路由。因此，L3层端到端算力调度和路由面临着又一个现网部署的挑战。如图2所示，跨数据中心的微服务调度和路由终结于VTEP。基于L3层网络之上的跨数据中心虚机、容器及微服务集群资源感知、调度和路由虽然不失为一种可行的方案，但是缺少了与网络深度融合的算网一体调度和路由的综合优势[3]。

2 分布式微服务治理架构下的IPv6段路由（SRv6）算网端到端路由方案

在微服务架构下，各个微服务节点涉及服务治理的基础



▲图1 基于应用网关的微服务路由机制



▲图2 基于VxLAN的数据中心虚机路由机制

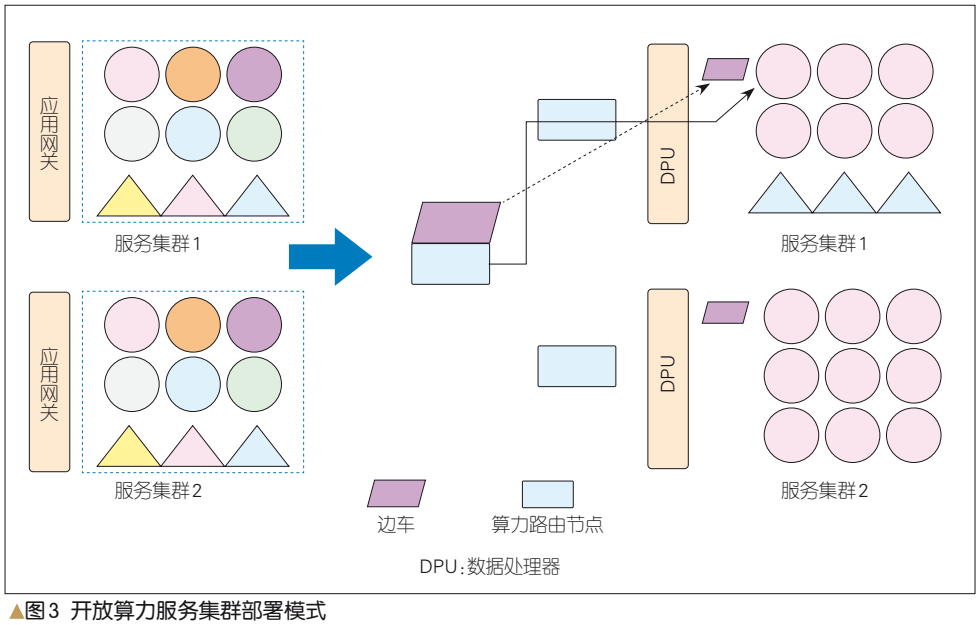
功能模块，如通信、安全、服务熔断、负载均衡等。这些模块往往被解耦成单独的模块并作为统一代理，以执行相应的功能[4]。目前行业内有两种主流的公共微服务治理模块部署模式，即应用网关模式和边车模式。其中，应用网关部署场景已在1.1节中阐述，它是一种集中式的代理入口模式；边车则是分布式模式，与微服务同节点部署。虽然如此，边车模式并不意味着微服务本身可被端到端L3层网络路由。在微服务集群的前端，系统往往通过统一的负载均衡接口对多个微服务集群执行应用代理接入。但是在分布式边车模式下，服务治理代理模块可分别在微服务的远端和近端部署。这为微服务作为一种公共算力资源对外开放和路由提供了技术支持。

2.1 基于微服务路由的服务集群部署模式变迁方案

如前所述，在当前的微服务集群治理模式下，各个集群的服务种类和服务能力是等价均衡的。这就相当于多个对等的应用服务实例以多个对等的微服务集群进行部署，从而实现资源的均衡利用。如1.1节所述，对于这种算力资源的部署模式，从全网算力资源协同和调度的视角看，应用为最小颗粒度，这跟算力网络的目标架构相比还存在较大的差距。从算力资源的部署和交付时间来看，这也是非常粗放和低效的一种模式。算力资源的部署，应根据用户发起请求的动态位置、节点的基础算力资源种类与能力，以及节点所在区域的业务需求等多因素，进行灵活编排，并据此执行最优的算网服务策略，在同等资源约束下为用户交付最优的算力服务质量。同时，算网资源的使用效率也会得到极大提升。如图3所示，对于不同的算力服务集群，服务部署的维度不再以特定应用为聚合颗粒度，而是根据不同的特征原则进行集群部署。比如，GPU算力池将部署图像处理算力服务，靠近工业控制现场的算力节点将部署工业数据采集和控制类算力服务。服务集群按照服务本身的应用场景和需求进行部署，而不以应用为颗粒度组织集群。在这种架构下，服务集群不再需要应用网关的统一入口逻辑功能，服务将可以被外部用户直接调度和路由。服务本身将成为一种公共服务算力资源颗粒度。这是与当前行业中微服务架构最根本的区别，也是全网算力网络的架构特征和关键内涵。特别地，在无需统一的对外网关接口的全新场景下，结合当前CPU算力卸载到专用加速硬件的最新发展动态，外部L3层路由流量直接对口服务集群的DPU模块，并通过DPU模块无缝路由至集群中的算力服务。

在图3服务集群部署模式中，边车可以进行远端部署。比如，在靠近算力服务请求方的网络边缘或入口处，在代理远端服务执行L7层的服务治理功能后，L3层将执行端到端算力服务路由，实现全网异构，以及跨池算力资源的灵活智能协同和调度。由于边车在当前微服务架构中主要负责执行微服务集群中的东西向微服务流量通信，在其实际的功能清单中，涉及集群以及微服务本地状态的一部分功能，并不适合直接从微服务集群中全部迁移出并在远端部署，比如鉴权、业务熔断等。因此，在这种微服务公共治理代理远端部署模式中，远端和本地模块会同时存在、同时部署，并形成互相补充和联动的关系，二者配合完成泛在微服务的端到端公共功能和治理。另外，在算力网络整体架构中，算网大脑也将执行一部分微服务治理的功能。比如，微服务提供方通过算网大脑注册本地服务种类、虚机及容器实例资源状态、微服务本身的认证等，微服务使用方则通过算网大脑完成接入认证、服务熔断等。

总的来说，在算力网络的目标架构下，算力服务的部署将呈现真正的泛在、异构、多样和层次化颗粒度的特征。通过网络统一调度和路由，算力和网络资源将成为一种深度融合、动态联动的公共基础能力，将为千行百业的业务应用提供高效、便捷、优质的算网服务。



▲图3 开放算力服务集群部署模式

2.2 微服务架构下的SRv6算网路由

基于SRv6技术拉通网络和云内业务的端到端路由，是近年来行业研究和实践的重点方向。SRv6基于网络和业务灵活可编程的功能特征[5]，同样也为算力网络提供一种优质的支撑技术。算网一体编排的核心要素是算力、网络资源和策略在统一的转发平面执行。这就意味着，SRv6的编程功能由网络向云池内的算力服务做深度延伸。如2.1节所述，算力网络架构下的云内微服务集群部署模式将使得集群内的算力服务全网路由可达。因此，算力服务将被作为SRv6端到端路由中的一个段路由，并被编排到统一的算网路由策略中。特别地，在应用需要多个集群内或集群之间的算力服务按照一定的时序组合完成服务的场景下，SRv6照样可以进行业务功能链的路由编排和策略执行。

2.3 SRv6算力路由在微服务架构下的终结模式

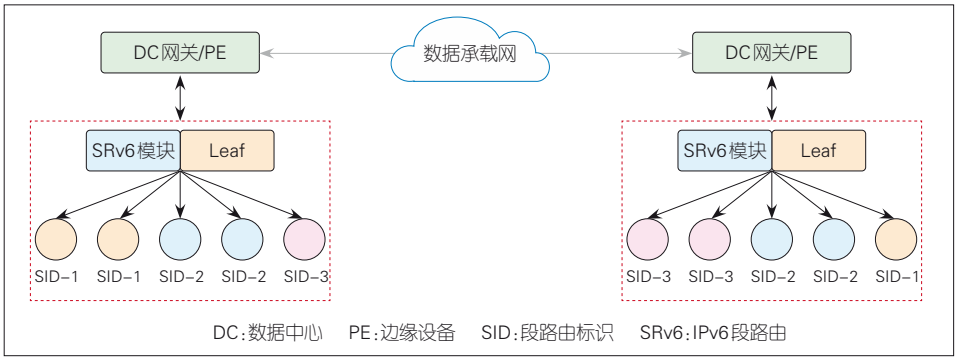
从算网端到端路由的全景视角看，云池外网络和云池内网络大多是基于两套架构、两套体系，甚至两套协议的异构网络的。近年来，两者呈现出互相渗透、互相影响的趋势。一方面，云池外网基于IPv6技术逐渐向云池内延伸；另一方面，云池内组网技术逐渐成熟，且更新迭代的周期快于外部网络，这些技术开始向外网渗透，如脊-叶（Spine-Leaf）组网架构[6]。具体到本文所述的SRv6算力路由[7]，SRv6逐渐深入到云池内网，如微服务集群内，为端到端网络+算力的综合路由提供网络基础设施条件。SRv6算力路由流量在云池内微服务架构下有两种终结模式：

• SRv6路由流量终结于数据中心（DC）网关，即SRv6轻度入云。在这种模式下，路由策略无法纳入云内业务，SRv6路由仅涉及网络侧端到端连接隧道。

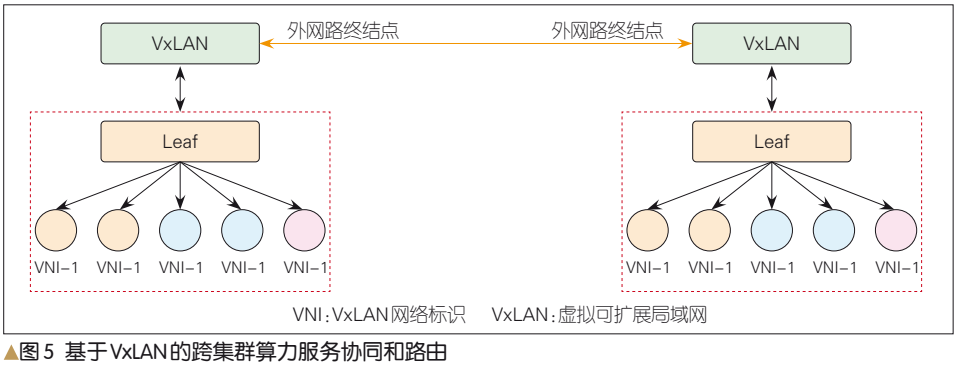
• SRv6路由流量终结于云池内Leaf节点的SRv6服务链（SFC）模块，即SRv6将深度入云，如图4所示。在这种模式下，SRv6端到端路由将执行策略编排并将自身纳入云池内的业务功能，以形成算网一体路由。但是，集群内被编排的业务仅作为一个服务节点被代理访问，并非一个独立的段路由。业务功能本身作为一个段路由被纳入算力路由策略，是最完备的SRv6算力路由场景。在该场景下，Leaf节点的SRv6模块有可能被跳过。但在实际部署中，考虑到成本和业务功能的复杂度，这种完备的SRv6算力路由方案并非最优选项。

3 VxLAN架构下跨微服务集群组网和寻址方案

VxLAN是当前云池内虚机及容器组网和寻址的主流方式。微服务集群之间的L3层寻址流量终结于VTEP，微服务集群内以及跨集群多种微服务之间的协同处理和路由面临巨大障碍。即便如此，VxLAN的网络标识空间仍非常庞大，并且部署模式也非常灵活。因此，跨集群的微服务协同和路由场景，可根据应用需求将关联微服务进行虚拟组网，即赋予同样的VxLAN网络标识（VNI），从而完成一组微服务跨集群的协同和组网路由。如图5所示，分布在两个集群资源池内的3种不同服务被赋予同样的VNI，不同服务之间在一个VNI标识的虚拟二层网络内灵活地完成数据协同处理。这种模式重用VxLAN的底层组网机制，虽然在算力服务方面实现了服务之间的灵活路由，但是在网络侧（尤其是云池外网侧），VxLAN是L3层路由之上的隧道通路，同时网络本身未被纳入端到端算力路由，应用在网络维度的SLA（服务等级协议）需求无法体现在路由策略中。这是这种方案的不足之处。当然，算力路由的主流场景应该是单算力服务的路由寻址。跨集群场景下的多服务协同组网路由，仍然可能通过SFC机制实现算网统一路由编排。



▲图4 SRv6入云模式下的算力路由



▲图5 基于VxLAN的跨集群算力服务协同和路由

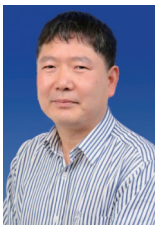
4 结束语

在算力网络架构下，端、边、云的全颗粒度算力成为全网可见、可调度、可路由的资源。网络由传统的拓扑路由进一步转变为算力路由，从而使能全新的网络架构和业务部署及交付模式，助力全行业数字化转型。其中，云内算力资源由当前的封闭模式转变为算力网络架构下的开放模式，对网络路由的颗粒度提出全新要求，在已经趋于成熟稳定的微服务架构下，增强SRv6、VxLAN等现网技术，并提供端到端算力路由解决方案，成为行业的一种优选路线。本文结合微服务架构及其部署和交付模式，对L3算网路由技术方案进行多维度的分析和探讨，为算力网络架构下端到端算力路由方案提供有益参考。

参考文献

[1] 李少鹤, 李泰新, 周旭. 算力网络: 以网络为中心的融合资源供给 [J]. 中兴通讯技术, 2021, 27(3): 29–34. DOI:10.12142/ZTETJ.202103007
[2] 兰巨龙,胡宇翔,张震, 等. 未来网络体系与核心技术 [M]. 北京: 人民邮电出版社, 2017
[3] SCHOLL B, SWANSON T, JAUSOVEC P. 云原生: 运用容器、函数计算和数据构建下一代应用 [M]. 北京: 机械工业出版社, 2020
[4] RICHARDSON C. 微服务架构设计模式 [M]. 北京: 机械工业出版社, 2019
[5] 李铭轩, 曹畅, 杨建军. 基于可编程网络的算力调度机制研究 [J]. 中兴通讯技术, 2021, 27(3): 18–22. DOI:10.12142/ZTETJ.202103005
[6] 魏月华,陈晓,张征. 数据中心网络架构和协议演进分析 [J]. 中兴通讯技术, 2021, 27(3): 51–55. DOI:10.12142/ZTETJ.202103011
[7] 黄光平, 史伟强, 谭斌. 基于SRv6的算力网络资源和服务编排调度 [J]. 中兴通讯技术, 2021, 27(3): 23–28. DOI:10.12142/ZTETJ.202103006

作 者 简 介



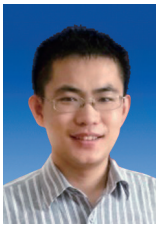
陈晓，中兴通讯股份有限公司有线架构部部长；长期从事电信产品和相关技术的研究规划。

黄光平，中兴通讯股份有限公司资深架构师；主要研究方向为下一代IP网络架构及关键技术，先后从事增值业务消息系统设计和开发、确定性网络以及远程宽带接入网关全球标准工作，近年聚焦算力网络架构、路由协议、算力标识等技术研究；发表论文6篇，申请专利20余项。

➡上接第20页

ACM, 2013: 127–132. DOI: 10.1145/2491185.2491190
[5] BOSSHART P, DALY D, GIBB G, et al. P4: programming protocol–independent packet processors [J]. ACM SIGCOMM computer communication review, 2014, 44(3): 87–95. DOI: 10.1145/2656877.2656890
[6] P4 Language Consortium. P4 [EB/OL]. (2003–07–31) [2021–12–15]. https://p4.org/
[7] WANG S H, MENG Z L, SUN C, et al. SmartChain: enabling high–performance service chain partition between SmartNIC and CPU [C]// Proceedings of ICC 2020 – 2020 IEEE International Conference on Communications. IEEE, 2020: 1–7. DOI: 10.1109/ICC40277.2020.9149136
[8] SCANO D, GIORGETTI A, SGAMBELLURI A, et al. Hierarchical control of SONiC–based packet–optical nodes encompassing coherent pluggable modules [C]//2021 European Conference on Optical Communication (ECOC). USA: IEEE, 2021: 1–3. DOI: 10.1109/ECOC52684.2021.9605850
[9] CONNOR O B, GHAFFARKHAH A, PUDELKO M, et al. Enabling the era of next generation SDN [EB/OL]. [2021–12–10]. https://opennetworking. org/stratum
[10] SANTIAGO DA SILVA J, STIMPFLING T, LUINAUD T, et al. One for all, all for one: a heterogeneous data plane for flexible P4 processing [C]// Proceedings of 2018 IEEE 26th international conference on network protocols. IEEE, 2018: 440–441. DOI: 10.1109/ICNP.2018.00063
[11] AGRAWAL A, KIM C. Intel Tofino2 – A 12.9 Tbps P4–programmable ethernet switch [C]//Proceedings of 2020 IEEE Hot Chips 32 Symposium (HCS). IEEE, 2020: 18–22. DOI: 10.1109/hcs49909.2020.9220636
[12] LUINAUD T, SANTIAGO DA SILVA J, LANGLOIS J M P, et al. Design principles for packet deparsers on FPGAs [C]//Proceedings of 2021 ACM/SIGDA International Symposium on Field–Programmable Gate Arrays. ACM, 2021: 280 – 286. DOI: 10.1145/3431920.3439303
[13] CAO Z, SU H Y, YANG Q M, et al. P4 to FPGA–A fast approach for generating efficient network processors [J]. IEEE access, 2020, 8: 23440–23456. DOI: 10.1109/ACCESS.2020.2970683
[14] LAKI S, HORPÁCSI D, VÖRÖS P, et al. High speed packet forwarding compiled from protocol independent data plane specifications [C]// Proceedings of the 2016 ACM SIGCOMM conference. ACM, 2016: 629–630. DOI: 10.1145/2934872.2959080

作 者 简 介



董永吉，解放军战略支援部队信息工程大学副研究员；长期从事路由与交换技术、网络安全和新型网络体系结构方面的研究工作；先后主持了1项国家重点研发课题，参与多项国家重点研发计划、“863”“973”项目，获得3项科研成果奖；发表论文10余篇，申请国家发明专利14项，出版专著2部。

胡宇翔，解放军战略支援部队信息工程大学教授、博士生导师；主要研究方向为新型网络体系结构、路由与交换技术。

崔鹏帅（通信作者），解放军战略支援部队信息工程大学副研究员；主要研究方向为新型网络体系结构、可编程数据平面。