基于 SRv6 的算力网络资源和服务编排调度

Computing Power Network Resources Based on SRv6 and Its Service Arrangement and Scheduling

黄光平 /HUANG Guangping，史伟强 /SHI Weiqiang，谭斌 /TAN Bin
（中兴通讯股份有限公司，中国 深圳 518057）
(ZTE Corporation, Shenzhen 518057, China)

摘要：提出一种以 IP 网络为中心的算力网络架构，即在网络域创建云池算力资源和服务的状态，从而实现网络层的算力编排和调度。算网一体编排和路由，是该算力网络架构的核心特征。针对算力网络中的服务多实例应用场景，所提架构方案对 SRv6 或基于 SRv6 的业务功能链（SFC）做功能增强和扩展，以满足单服务对应动态多实例的算力路由需求。控制面架构方案采取一种分级分层状态表的维护机制，将不同颗粒度的算力资源和服务状态在不同的网络域做同步通告，并创建对应的分级路由表，从而压缩节点的状态表和边界网关协议（BGP）的通告频率。转发面则执行算力服务标识语义封装，承载网骨干节点仍然保持无状态转发。

关键词：算力网络；SRv6；算力状态；分级路由

Abstract: An IP network-based architecture of computing power network is proposed, which creates the state of cloud pool computing power resources and services in the network domain to realize the computing power arrangement and scheduling of the network layer. Integrated computing network arrangement and routing are the core features of the computing power network architecture. For the service multi-instance application scenario in the computing power network, the proposed architecture scheme enhances and extends SRv6 or SRv6-based service function chaining (SFC) to support the single service routing requirements for dynamic multi-instances. The control surface architecture scheme adopts a maintenance mechanism of hierarchical state tables, which synchronously notifies the computing power resources and service states of different granularity in different network domains, and creates the corresponding hierarchical routing table, to compress the state table of the node and the notification frequency of the border gateway protocol (BGP). Accordingly, a dual-semantic encapsulation with IP topology and computing service identification in the forwarding plane would also be proposed, while the backbone network nodes would remain unaware of computing power metrics.

Keywords: computing power network; SRv6; computing status; classified routing

DOI：10.12142/ZTETJ.202103006
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.TN.20210623.1809.002.html

网络出版日期：2021–06–24
收稿日期：2021–05–10

在互联网协议（IP）承载网络域，通过精细化动态感知，网络控制器或网络节点可以创建基于多云池内算力资源及服务状态的算力路由表，并据此进行算力资源和服务的编排调度。这是以网络为基础平台的算力网络架构的核心要素。也就是说，在 IP 拓扑路由的基础上，新增算力资源和服务路由，使路由策略约束机制由当前的 IP 拓扑单约束演变为 IP 拓扑和算力双约束。这给网元控制面、转发面和管理面均带来新的挑战，也是算力网络为 IP 网络引入的全新议题。

当前主流的云侧应用级跨云池计算资源调度系统，如内容分发网络（CDN）、AWS（亚马逊公司的云计算服务）等，均与特定应用或应用集群硬绑定。除此之外的其他应用无法

接入该系统纳管的计算资源。此外，这种云侧算力调度系统纳管的云池资源是一种典型的封闭调度平台，仅限于在服务商自营的资源中，且从技术和运营模式上均不兼容多元云池计算资源。更重要的是，这类云侧调度系统与网络资源无关，即它的网络连接服务要么适用于公共网络的“尽力而为”服务，要么适用于专线租用或业务虚拟专用网络（VPN）的开通。网络与计算业务独立配置、独立编排、独立调度。以网络为基础平台的算力网络，构建的是一个开放平台，即与具体的应用和业务完全解耦，且兼容多元云池算力资源和服务。与云侧算力调度显著不同的是，在算力网络架构下，算力和网络的状态和路由表均由网络维护，因此这种算力网络架构内生支持算网一体编排和调度。

然而，一个开放的算力网络平台，可以创建多元云池算力资源、服务状态、路由表，其前提是算力资源和服务的标准化度量和标识。SRv6（基于IPv6 的源路由技术）中间转发节点无状态的优良特征，非常适合算网一体路由策略和路由转发，但是需要在转发面和控制面进行功能增强和扩展，以满足算力网络场景下的全新需求。同时，根据应用的算网服务级别协议（SLA）需求，网络需要进行精准灵活的资源匹配和编排，并需要对应用的算力 SLA 进行更细颗粒度的感知。

1 算力资源和服务的颗粒化度量

当前，云池算力资源和服务的运行模式是与业务强相关，并且高度本地化的，不存在互通和交易，因此尚无系统的度量和标识方案。但是，云池内的算力资源和服务在网络域进行应用流颗粒度的编排和调度，涉及算力资源和服务的跨池跨域调度，以及平台层面的多方资源交易。因此，对算力资源和服务进行层次化颗粒度的度量和标识，是算力网络架构的关键因素。如图 1 所示，从交付和执行模式来看，算力资源可以分为 3 个层次，或称为 3 种颗粒度。

1.1 算力资源和服务的层次化颗粒度

（1）基础设施即服务（IaaS）类型算力资源

该类型算力资源属于裸资源，包括中央处理器（CPU）、图形处理器（GPU）、现场可编程门阵列（FPGA）、专用集成电路（ASIC）等。当前这些资源的度量颗粒度，比如核数，无法满足算力网络精细颗粒度的资源调度。因此，需要针对各类异构的计算裸资源进行系统的标准度量。可服务计算资源的标准量化数据，是网络对算力资源感知并创建状态的数量依据。

（2）函数即服务（FaaS）类型算力服务

虚拟机、容器、微内核等更细颗粒度计算单元的出现，让一些基础计算功能或服务的驻留和运行模式发生根本性的变化。例如，分布式的微服务架构，将传统单一应用系统解耦成独立的微服务群组，应用层根据特定的业务逻辑调用不同的微服务，完成特定的业务功能。

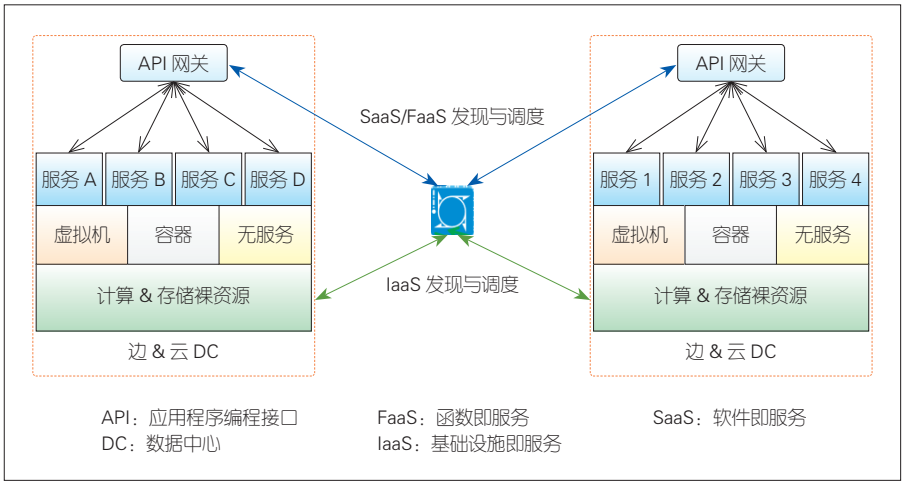
在这种架构下，一些与业务无关的基础计算功能或算法可以实现分布式灵活部署，更加快速地满足新型业务需求，缩短新业务上线周期，大幅降低部署成本。基础计算功能是算力裸资源的一种可服务形态，而算力网络需要创建基于其状态的路由表，并在网络域完成对这种计算功能服务的编排和调度。

（3）软件即服务（SaaS）类型算力服务

相对于当前增值业务的单站点资源部署和服务模式，在算力网络目标架构下，增值算力服务的驻留和服务将由单点变为全网虚拟 SaaS 池的模式。同一类增值算力服务资源，在上层交易系统的支撑下，可以在算力网络域完成跨池编排和调度。

1.2 算力资源和服务的度量和标识

如 1.1 所述，算力资源的标准化度量，需要针对上述 3 种颗粒度的资源和服务进行业务无关的通用度量，以及 CPU、GPU 等异构裸资源的度量。目前，学术界和信息技术（IT）界已经开始了一些有益的尝试。资源和服务标准化标识的实现，首先需要建立一个结构化的标识体系，对各种颗粒度的资源和服务进行收敛和标定。考虑到网络单元的存



▲图 1 层次化算力资源和服务颗粒度

储和处理容量限制，网络域可感知、可编排、可调度的资源和服务标识需要优选数字化标识机制[1]。

2 基于 SRv6 的算力网络增强控制面技术

在网络域创建、维护云池算力资源和服务的状态，也就是完成对多资源和服务颗粒度的精细化和动态感知，是控制面在算力网络架构下的首要功能。控制面有集中式和分布式两种通用架构技术。

2.1 集中式控制面架构增强

目前的控制器主要有 3 类。第 1 类是管理与编排（MANO）控制器，负责纳管移动边缘计算（MEC）内的计算和存储资源、侧重占用率之类的宏观数据，其颗粒度无法满足算力网络的精细化编排和调度需求。因此，可以基于上述算力资源的标准化度量，对 MANO 纳管的算力资源颗粒度进行扩展和增强。第 2 类是数据中心和边缘计算中心控制器，负责纳管云内网络拓扑资源。其颗粒度可达服务器对应的端口号，但无法纳管层次化的算力资源和服务。同样，它也可以进行扩展和增强，以涵盖对算力资源的精细化纳管。第 3 类是 IP 承载网控制器，负责纳管承载网络域的拓扑资源。

另一种可选方案则是新增算力资源编排器，可与上述 3 类控制器并列；但也可以居于更上一层，在纳管层次化算力资源的同时，统一纳管数据中心或边缘计算中心、IP 承载网的网络拓扑资源，可以实现单点算网全局资源视图。

2.2 分布式控制面架构增强

跨云池的算力资源和服务分布式路由协议，目前主要是基于边界网关协议（BGP）增强和扩展。BGP 在现网通告的对象主要是节点端口、链路等拓扑资源的状态。这些资源的变化周期通常为小时、天，甚至月的数量级，网络的高并发拓扑变更会造成路由震荡等严重后果。在算力资源和服务状态（尤其是 FaaS 级算力服务的状态）被通告的情景下，其资源标识种类和通告频率均远大于网络拓扑资源及其通告频率。例如，在一些通用计算功能实例中，一次服务执行的生命周期最短可达毫秒级。大规模的通告量和高通告频率，对算力路由表的稳定将造成严重的后果。因此，简单地扩展 BGP 通告的资源种类，无法解决路由表高度不稳定的问题。本文中，我们提出一种分级通告分级路由的机制，极大地压缩 BGP 通告的资源数据量和通告频率；还提出一种独立于 BGP 的全新算力路由协议雏形。

2.2.1 基于 BGP 的分级路由机制

分级分域路由通告的算力网络路由解决方案，旨在解决两个算力网络路由的问题：多种云内算力资源及服务在路由节点上引起的超大路由表项问题、算网端到端路由问题[2]。

我们将算力资源和服务划分为两种颗粒度：

（1）边缘计算节点或数据中心的粗颗粒度（颗粒度记为 1）算力资源，包括但不限于：

- 计算及存储资源的种类，如 CPU、GPU、嵌入式神经网络处理器（NPU）、ASIC 等；
- 上述资源种类的可用状态，包括但不限于量化空闲资源值，如使用率、可用核数目等；
- 提供的算力服务种类，包括 SaaS/FaaS 服务种类及标识，以及服务对应的忙闲状态属性，并且服务的忙闲状态阈值可配置，如 90% 及以上为忙的状态；

（2）边缘计算节点或数据中心的细颗粒度（颗粒度记为 2）算力服务，包括但不限于：

- 算力服务种类以及其所对应的可服务实例数；
- 每实例的处理容量；
- 算力服务与其实例之间的标识映射关系，如一个任播地址 Anycast 标识一个算力服务，关联的群组成员地址为实例地址。

粗颗粒的算力资源状态仅在边缘计算节点或数据中心节点之间通告，并维护对应的路由表项。首次上线的节点，通告上述粗颗粒度全集数据，此后根据可配置的变更门限值来触发变量更新通告和同步。通告可有两种方案：BGP 扩展方案，即将上述粗颗粒度算力资源信息，通过扩展 BGP 协议载荷，通告至邻居网络边缘节点；集中式控制器方案，包括但不限于通过路径计算单元通信协议（PCEP）、边界网关协议 – 链路状态（BGP–LS）等通告同步上述粗颗粒度算力资源相关信息。

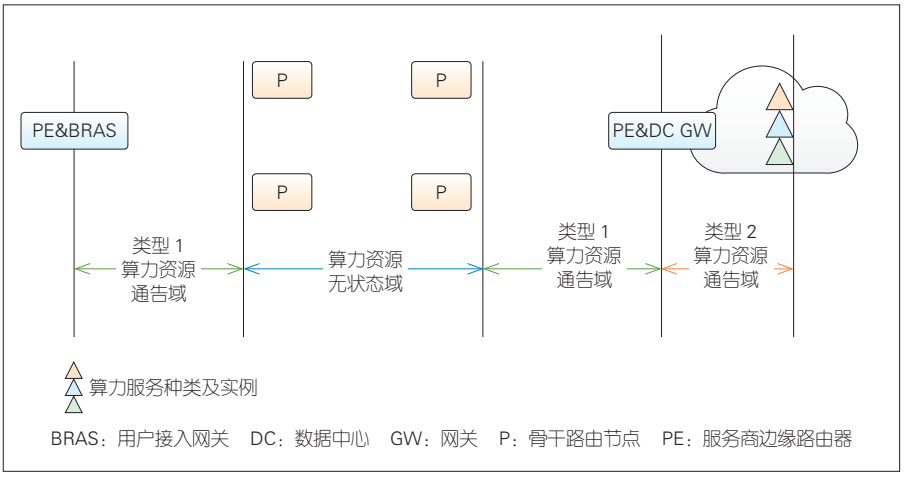
细颗粒度算力服务状态，仅在边缘计算或数据中心节点所归属的域内网络边缘节点进行维护，无须通告邻居网络边缘节点。首次上线的节点，通告或发布上述全集信息，此后根据可配置的变更门限值，触发变量更新通告和同步。细颗粒度的算力服务通过如下可选方案通告网络边缘路由节点：发布订阅的应用消息，并向网络边缘节点通告状态数据；通过内部网关协议（IGP）扩展通告，将上述细颗粒度算力服务信息通过扩展 IGP 协议载荷，向网络边缘节点通告。

2.2.2 基于 BGP 的地址路由和算力服务路由的两级路由表机制

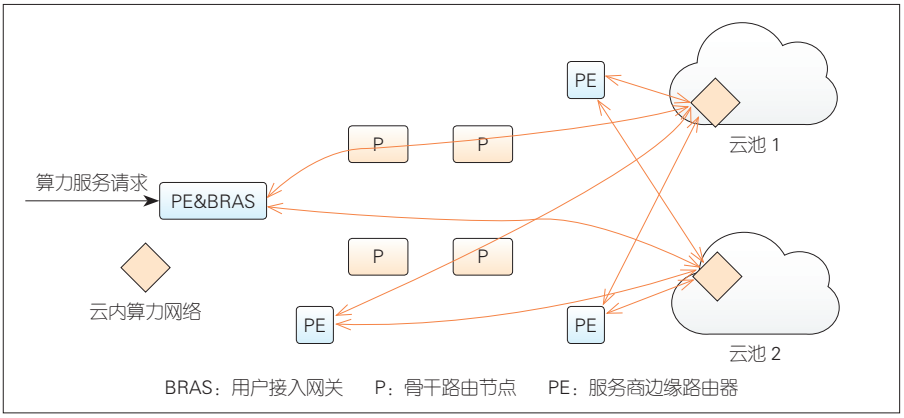
用户接入网络边缘节点维护类型 1 路由表，即路由节点仅感知边缘计

算或数据中心节点的粗颗粒度算力资源信息，并以此创建、维护对应的算力路由表。类型 1 的算力资源颗粒度较粗，变更频率较低，因此网络边缘节点维护的类型 1 路由表的大小与联动的边缘路由和数据中心节点数目成正比，路由表规模可以得到数量级的压缩。

边缘计算或数据中心节点归属的域内网关或网络边缘节点维护类型 2 算力服务路由表，即上述域内网关或网络边缘节点可以感知本边缘计算或数据中心节点内的算力服务状态，并以此创建、维护对应的算力服务路由表或映射表。类型 2 路由表的大小，与该网络边缘节点、网关归属的边缘计算或数据中心提供的算力服务规模成正比。由于仅做本地的或有限归属边缘计算的或数据中心节点的算力服务信息状态维护，类型 2 路由表规模得到极大的压缩。两级算力颗粒度类型路由及通告机制如图 2 所示。



▲图 2 两级算力颗粒度类型路由及通告机制



▲图 3 基于网络 L4 的新型算力路由协议通告

2.2.3 新型算力路由协议

云内算力资源和服务的种类以及状态变更频率均与现网 IP 拓扑通告有着显著区别。为了适应新型算网一体路由架构，我们提出一种全新的算力路由协议。该协议内生支持算力资源和服务的跨域通告，并将与 BGP 解耦，从而规避算力资源的动态对现网路由收敛的负面影响。网络和算力资源的融合路由策略通过算法优化解决。我们还提出了一种基于网络 L4 的新算力路由协议架构，其主要特征是算力资源和服务在云内直接发布，并由服务商边缘路由器（PE）为其创建算力路由表，如图 3 所示。

两种可能的协议模式为：发布订阅机制和定向通告机制。

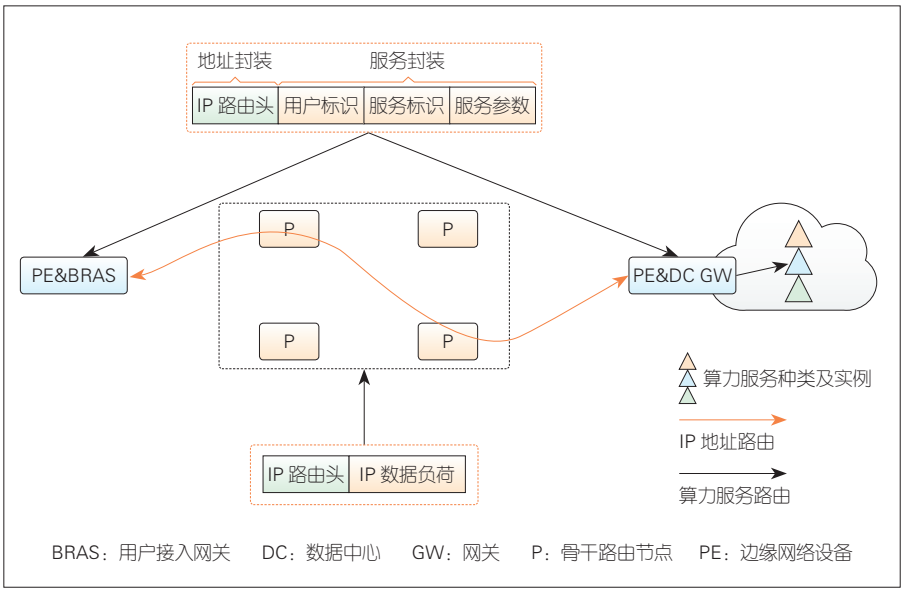
（1）发布订阅机制：作为发布主体，云池内算力网关对云内层次化算力资源进行发布，并对云池内算力资源状态信息进行结构化设计；支持增量发布，支持高频率动态更新；发布对象为网络边缘节点以及用户的接入网关。

（2）定向通告机制：云内算力网关向网络边缘节点以及用户接入网关主动发起面向连接的状态通告，网络边缘节点以及用户接入网关仅接收通告并据此创建和更新路由表；支持基于隧道的高频率更新通告。

3 基于 SRv6 的算力网络增强转发面技术

算力网络路由是一种集网、云、算为一体的综合路由。在网络入口节点，算力网络路由根据用户业务的算力和网络双 SLA 约束，制定算网路由策略。和当前 IP 拓扑路由显著不同的是，IP/ 多协议标签交换（MPLS）拓扑路由本质上解决的是“去哪里”，即明确路由的网络目的节点，在参数上体现为 IP 地址或标签。在算力网络架构下，网、云、算综合路由本质上解决的是“去哪里”+“干什么（执行何种计算服务）”，即在 IP 路由的基础上，叠加了算力服务路由。因此，转发面的报文头需要执行 IP 路由 + 算力服务路由双重封装。算力网络的 IP 和算力服务双重路由机制网络流程图，如图 4 所示。

如 2.2.2 节所述，在分级路由表的机制下，网络在入口和出口节点，维护有两种不同颗粒度的算力路由表，



▲图 4 算力网络 IP 和算力服务双重报文封装和路由机制

这对应转发面的 IP 拓扑和算力服务双重路由封装。在用户接入网关（如 BRAS）处，网络执行上述两级封装，并由用户接入网关根据 2.2.2 节所述本地维护的类型 1 路由表，计算生成到选定的边缘计算或数据中心节点的路由，并执行 IP 拓扑地址封装。我们有两种封装方案：（1）目的地址封装方案，即将选定的边缘计算或数据中心节点归属的网络边缘节点或网关地址，作为目的地址，封装在报文头对应的字段中，包括但不限于互联网协议第 4 版（IPv4）、互联网协议第 6 版（IPv6）、MPLS 等网络数据平面；（2）源路由地址方案，即以选定的边缘计算或数据中心节点归属的网络边缘节点或网关作为出节点，编排源路由路径，并封装在对应的报文头中，包括但不限于 SR-MPLS、SRv6 等网络数据平面[3]。

用户接入网关（如 BRAS）根据用户算力服务请求执行算力服务标识封装，这包括：单一算力服务标识封装、基于 SRv6 的业务功能链（SFC）、多算力服务标识链封装。算力服务标识的封装包括两种方案：（1）增强 SRv6 算力服务标识编程扩展方案，即在片段识别（SID）的 Locator + Function（定位器 + 功能）结构中，算力服务标识作为 Function 封装在 SID 中，并可选择扩展 Argument 来作为算力服务的必要输入参数；（2）算力服务标识封装在 IP 与 L4 传输层之间的 overlay 层中，如 SFC 架构下的网络业务报文头（NSH）、三层网络虚拟化 overlay（NVO3）的 Geneve 等，还可以在 IPv6 之上引入一个全新标识层，用于封装算力服务标识，从而实现与 IP 层完全解耦。在这种 IP 拓扑和算力服务双路由封装、点到点路由的机制支持下，网络中间转发节点无须识别算力服务标识，仅做普通路由转发，即平滑继承当前网络中间节点无状态的特征。

类型 1 路由的出节点执行算力服务标识解封装，并查找节点维护的所属边缘计算或数据中心算力服务的路由表或映射表，从而将用户数据路由至对应的服务实例，并终结全部端到端算网路由。

特别地，为了保持流粘性，即确保同一应用的数据流被路由至同一个算力服务实例，出节点维护应用数据流标识与算力服务实例的映射关系，并将后续应用数据流路由至同一算力服务实例。这种映射关系的维护方法包括但不限于 5 元组方案（源 IP 地址、目的 IP 地址、源端口、目的端口、传输层协议类型）。在 IP 拓扑和算力服务双重封装的机制下，算力服务标识仅仅体现了服务类型的抽象语义，而实际服务实例节点的映射关系被维护在 2.2.2 节所述的类型 2 路由表中。由于路由表具有与业务无关的中性特征，算力业务流粘性的维护保证，需要在出入口节点维护业务相关的状态。在两级路由、两级封装的全流程下，流粘性也需要维护对应的两个颗粒度的状态，即在入口节点维护业务标识和算力服务标识的状态，业务标识可通过类似前述 5 元组的模式实现。在出口节点维护业务标识、算力服务标识和服务标识实例的状态，服务标识实例可以是虚拟局域网（VLAN）/ 虚拟扩展局域网（VxLAN）号、端口号、IP 地址等。

4 网络对算力应用的感知

在当前数据网络的转发和路由机制中，网络资源和策略对应的最小颗粒度是流甚至报文。也就是说，从本质上看，网络路由策略是与业务无关的。在算力网络架构下，网络感知云池算力资源和服务，并根据应用的算力 SLA，在网络层对算力资源和服务进行编排和调度。与当前网络策略和路由机制不同的是，算力资源和服务对应的最小颗粒度是算力应用，且必须与业务相关。当前网络路由策略的聚合服务质量（QoS）机制，无法直接对标算力 QoS 的颗粒度。算力 QoS 更加灵活，不便于聚合，因此算力网络的另一个全新技术挑战是网络层（L3）对应用的算力 SLA 的感知。

由于 ISO 层级解耦的内生架构原

则，当前网络层没有感知接口，对应用无感知。算力网络架构下，应用的算力 SLA 的感知主要有两种方案：一种是控制面方案，即所谓的带外方案，通过类似接入控制信令扩展向网络入口网关通告特定算力应用的 SLA，网络入口网关据此创建算力应用颗粒度的会话。控制面方案的优势是安全、可信、与设备硬件无关。另一种方案是转发面方案，即所谓的带内方案，通过在 IPv6 或 SRv6 的扩展头中增强封装应用标识及其 SLA，网络节点解封装即可执行对应的路由策略。转发面应用感知方案的优势是网络每个节点均可做精细化策略和资源匹配，但这也引入了额外的安全问题，以及大量的冗余硬件设备处理负荷。

5 结束语

算力资源和服务的标准化度量和标识是算力网络中一个重要的支撑要素。层次化资源和服务颗粒度下的度量和标识，带来了精细化的可编排、可调度算力资源和服务体系。在网络域创建云池算力资源和服务的状态，给控制面尤其是路由协议如 BGP 等带来了挑战。本文中，我们提出了一种基于聚合原则的分级分层路由表机制，即将算力资源和服务分为粗和细两种颗粒度，极大地压缩了路由协议的通告频率和路由表尺寸。同样，在转发面引入基于 SRv6 可编程的增强功能，或扩展 overlay 层的 IP 拓扑和算力服务标识双重语义封装，都能较好地适应 IP 拓扑和算力服务双重路由的全新需求和场景。同样，当前网络 L3 不能感知应用的层级解耦模式，无法应对算力网络的资源匹配和调度需求。这需要通过带外模式，即控制面增强扩展方案来实现网络层对算力应用感知，对现网架构以及设备的影响最小。

参考文献

[1] 朱海东 . 云网一体使能网络即服务 [J]. 中兴通讯技术 , 2019, 25(2): 9–14. DOI: 10.12142/ZTETJ.201902002
[2] 刘铎 , 杨涓 , 谭玉娟 . 边缘存储的发展现状与挑战 [J]. 中兴通讯技术 , 2019(3): 15–22. DOI: 10.12142/ZTETJ.201903003
[3] 马洪源 . 面向 5G 的边缘计算及部署思考 [J]. 中兴通讯技术 , 2019(3): 77–81. DOI: 10.12142/ZTETJ.201903011

作 者 简 介

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；发表论文 3 篇，申请专利 20 余件。

史伟强，中兴通讯股份有限公司有线架构总经理；主要研究方向为 IP 网络、光网络和 SDN 系统架构与技术，先后从事网管、接入网和 SDN 控制器等产品的架构设计和研发管理工作；获 2012 年国家科学进步奖二等奖等奖项；发表论文多篇，申请专利 3 项。

谭斌，中兴通讯股份有限公司未来网络技术研究项目经理；主要研究方向为 IP 网络、SDN 系统架构与技术，先后从事有线路由器、接入产品开发、产品规划和市场等工作；申请专利 2 项。