东数西算场景下的算力网关研发及应用

Research and Application of Computing Power Gateway in East-Data-West-Computing Project

马思聪/MA Sicong，孙吉斌/SUN Jibin，孙一豪/SUN Yihao
（中国电信股份有限公司研究院，中国 北京 102209）
(China Telecom Corporation Research Institute, Beijing 102209, China)

DOI：10.12142/ZTETJ.202304002
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20230724.1524.004.html
网络出版日期：2023-07-25
收稿日期：2023-06-15

摘要：提出了一种面向算力网络场景的新型网络设备——算力网关。认为算力网关是实现算力网络一体化调度的基础，通过感知业务应用需求，结合当前的算力状况和网络状况，生成路由信息并发布到网络，将计算任务报文路由到合适的计算节点，以实现算力资源的最优调度。现网实践验证了算力网关技术方案的有效性。

关键词：算力网关；算力感知；算力路由；东数西算

Abstract: The computing power gateway, a new network device, is proposed for computing power network scenarios, which is the foundation for achieving integrated scheduling of computing power networks. By perceiving the information of services and combining the current computing power and network performance, the routing information will be published to the network to route computing task packets to suitable computing nodes, in order to achieve optimal scheduling of computing resources. Finally, the effectiveness of the computing power gateway technology solution is verified through practical verification in the current network.

Keywords: computing power gateway; perception of computing power; routing of computing power; east-data-west-computing

引用格式：马思聪，孙吉斌，孙一豪 . 东数西算场景下的算力网关研发及应用 [J]. 中兴通讯技术，2023, 29(4): 2-7. DOI: 10.12142/ZTETJ.202304002

Citation: MA S C, SUN J B, SUN Y H. Research and application of computing power gateway in east-data-west-computing project [J]. ZTE technology journal, 2023, 29(4): 2-7. DOI: 10.12142/ZTETJ.202304002

随着数字经济进入新发展阶段，业务数字化、技术融合化和数据价值化等加速演进，开启数字经济引领高质量发展新征程。在此发展过程中，算力作为数字时代核心资源的作用日益突出，以算力为核心的数字信息基础设施建设被提到前所未有的高度[1]。中国相继出台一系列围绕算力基础设施的政策文件，如《全国一体化大数据中心协同创新体系算力枢纽实施方案》《新型数据中心发展三年行动计划》等[2]，并加快实施“新基建”“东数西算”等工程，加强区域协同联动，推进热点区域与中西部地区、一线城市与周边地区的数据中心协调发展，引导算力的集群化发展。

为了实现算力像电力、热力、水一样，由统一的社会基础设施进行供应，真正地服务于社会经济的各行各业，需要在算力基础设施的供给模式方面进行创新，算力网络应运而生。算力网络是通过网络分发算力节点的计算、存储、算法等资源信息，并结合网络信息和用户需求，提供最佳的计算、存储、网络等资源的分发、关联、交易与调配，从而实现各类资源最优化配置使用的新型网络技术。作为算力网络中的核心网元设备，算力网关以算力度量、算力标识为依据，通过算力路由、算力感知等核心功能，传输发布相关算力策略与数据转发，是实现算力网络一体化调度的基础。

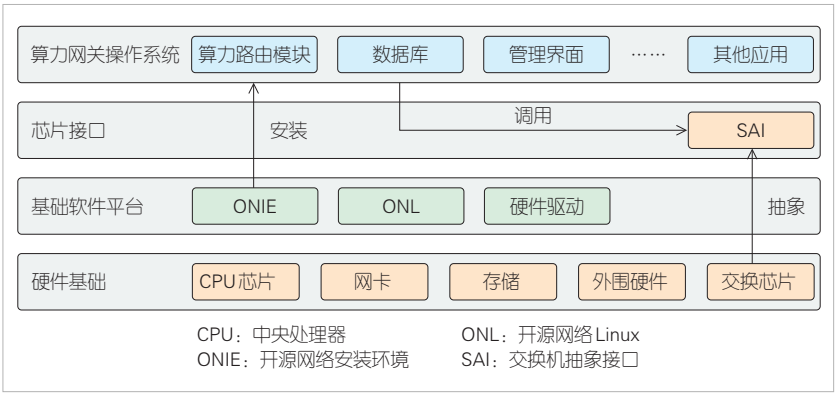
1 算力网关架构及组网方案

1.1 总体架构

算力网关基于开放的白盒网络设备架构，将网络中的物理硬件和节点操作系统（NOS）进行解耦，使标准化的硬件

基金项目：国家科技重大专项（2022YFB2901400）

配置与算力网络相关协议进行组合匹配，具有灵活、高效、可编程等特点，有助于算力网络相关协议的制定。算力网关整体架构如图1所示，主要分为硬件基础、基础软件平台、芯片接口和操作系统4个部分[3]：



▲图1 算力网关整体架构

1）硬件基础

硬件是算力网关系统运行的物理基础，主要由CPU、交换芯片、网卡、存储和外围硬件等构成。其中，CPU是对计算机的所有硬件资源（如存储器、输入输出单元）进行控制调配并执行通用运算的核心硬件单元，主要管控系统运作；交换芯片主要提供高性能和低延时的交换能力，是算力网关的核心芯片；网卡分为用于设备管理的管理网卡和用于网关与网络中其他设备通信的业务网卡，业务网卡与交换芯片共同决定了算力网关的转发性能；存储主要包括内存和硬盘，用于设备应用数据的存储和保存；外围硬件主要包括风扇、电源等用于维持设备正常运行的其他基础硬件。

2）基础软件平台

基础软件平台由开源网络安装环境（ONIE）、开源网络Linux（ONL）以及硬件驱动构成。其中，ONIE为算力网关提供一个开放的安装环境，可实现网关硬件和网络操作系统系统的解耦，支持在不同厂商的硬件上引导启动算力网关操作系统；ONL建立在开放网络硬件上，向网关系统提供基础操作系统，为交换硬件提供管理接口，使用ONIE来安装到板载闪存中。

3）芯片接口

交换机抽象接口（SAI）是一种标准化的应用程序编程接口（API），可以看作是一个用户级的驱动。在不同的专用集成电路（ASIC）芯片上，SAI为上层应用提供了统一的API。SAI的具体实现由不同ASIC芯片提供商负责，使用者不需要关心网络硬件供应商的硬件体系结构的开发和革新，通过始终一致的编程接口就可以很容易地应用最新、最好的硬件。

SAI本质就是在各ASIC的软件开发套件（SDK）之上抽象出统一接口。芯片厂商研发的ASIC的SDK需要与这层抽象适配，使得转发应用能够在不同的ASIC上运行。SAI向上为操作系统提供统一的API，向下可以对接不同的ASIC。

4）操作系统

算力网关操作系统基于社区版本的云开发网络软件（SONiC）开发，通过拓展协议和网关接口等能力实现了算力网络所需的相应功能。

算力网关操作系统由多个功能模块组成，这些模块通过集中式和可扩展的基础架构相互交互。本系统模块间交互依赖于redis数据库引擎（一个键值数据库，提供独立于语言的接口，可以在所有子系统之间进行数据持久化、复制和多进程通信）。通过依赖redis引擎基础架构提供的发布者/订阅者消息传递模型，应用程序可以仅订阅它们需要的数据，并避免与其功能无关的实现细节[4]。

算力网关操作系统将每个模块放置在独立的docker（容器）中，以保持语义相似组件之间的高内聚性，同时减少不相关组件之间的耦合。每个组件都被设计得相对独立，摆脱了平台和底层交互的限制。当前，算力网关操作系统主要包含以下几个dockers：Bgp、Web、Database、SwSS、Syncd、Teamd、Pmon、DHCP-relay等。

1.2 组网方案

算力网络目前在技术路线上可以分为集中式、分布式和混合式3种。在算力网关应用部署中，我们主要考虑混合式和分布式两种组网方案[5]。

1）混合式方案

在混合式的方案中，算网编排系统依靠云/算管控模块通过算力网关收集来自每个资源池的算力信息，通过网络管控模块收集网络拓扑信息。算网编排系统确定最优算力资源节点和网络路径。云/算管控模块与网络管控模块分别下发算力资源分配指令和路由策略，如图2所示。

在此架构下，算力网关主要功能包括：获取算力节点的算力信息及链路信息，接收网络管控模块下发的路径策略信息等。

2）分布式方案

在分布式方案中，算力网关需要实现算力资源感知、算力路由分发、资源表项生成、策略定制等全部功能，如图3所示。除支持算力信息发布和通告外，分布式方案还需通过算力和路径计算生成路由策略，并依据用户和应用感知对路由策略进行绑定，进而实现对算力资源和网络资源的信息同步与统一调度。

2 算力网关核心技术实现

算力网关通过感知算力和网络信息，将当前的计算能力状况和网络状况作为路由信息发布到网络，并将计算任务报文路由到合适的计算节点，以实现整体系统最优和用户体验最优。其中，算力感知和算力路由是算力网关的两大核心技术能力。

2.1 算力感知能力

算力感知是对算力资源的性能、实时负载、网络状况以及业务需求的全面感知，主要是需要明确网络中有多少算力资源，用户有怎样的算力需求。算力感知包括算力信息感知、网络状况感知、业务需求感知。

1）算力信息感知

算力信息的感知通常包括算力资源池的IP地址、计算能力、存储能力等内容。

如图4所示，云资源池一般由云管平台集中纳管，算力网关可以与云管平台通过RESTful API等接口交互，获取算力资源池的IP地址、计算能力、存储能力等，并最终把感知的算力信息上报到算力交易平台。

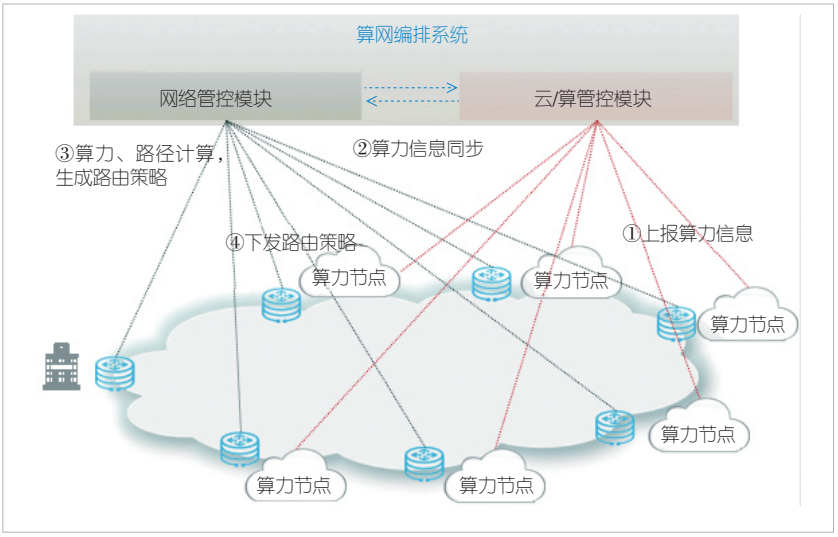
2）网络状态感知

网络信息的感知通常包括时延、带宽、丢包率、抖动等内容。以网络时延为例，由于算力资源池分布在不同位置，用户到资源池的网络路径也会根据网络拥塞状态发生变化，因此需要探测用户与各个算力资源之间的时延信息。算力网关的时延探测分为两部分：一是算力网关到算力资源池之间的时延探测；二是算力网关与算力网关之间的时延探测，如图5所示。

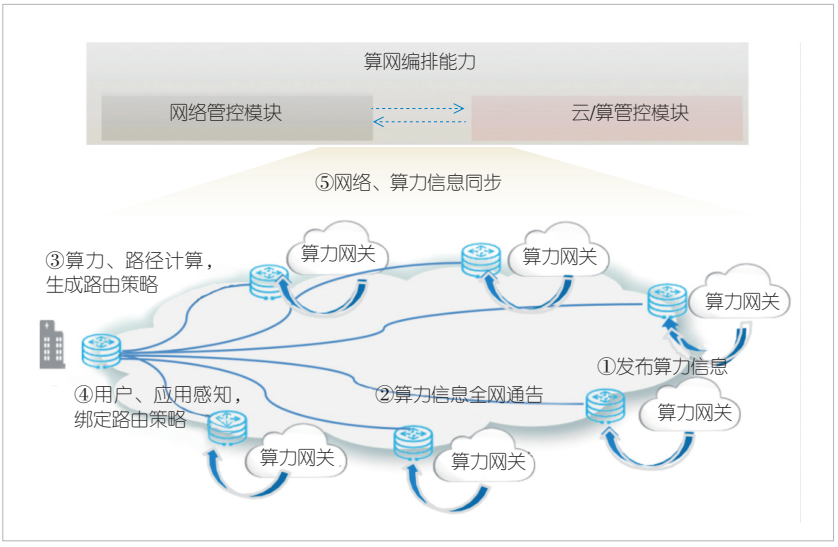
3）业务需求感知

除了对算力资源和网络状况的感知外，算力感知还应具备感知用户业务需求的能力，以实现更优的算力调度。

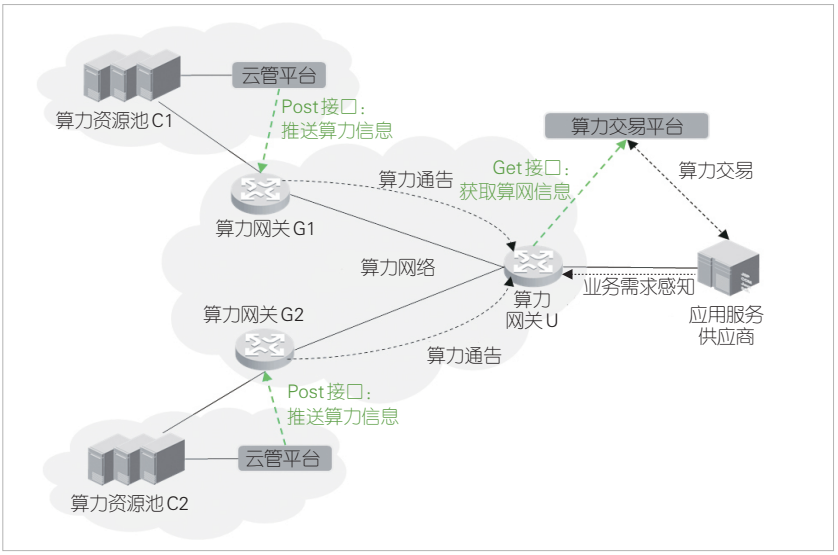
业务需求感知可以在用户入口的算力网关接收业务请求并感知业务需求，包括网络需求（时延、抖动等）和算力需求（算力请求类型、算力需求参数等），依据算力度量标准和特定的算法匹配可用算力。这样不仅能够精确匹配具体应用的业务需求，还能动态



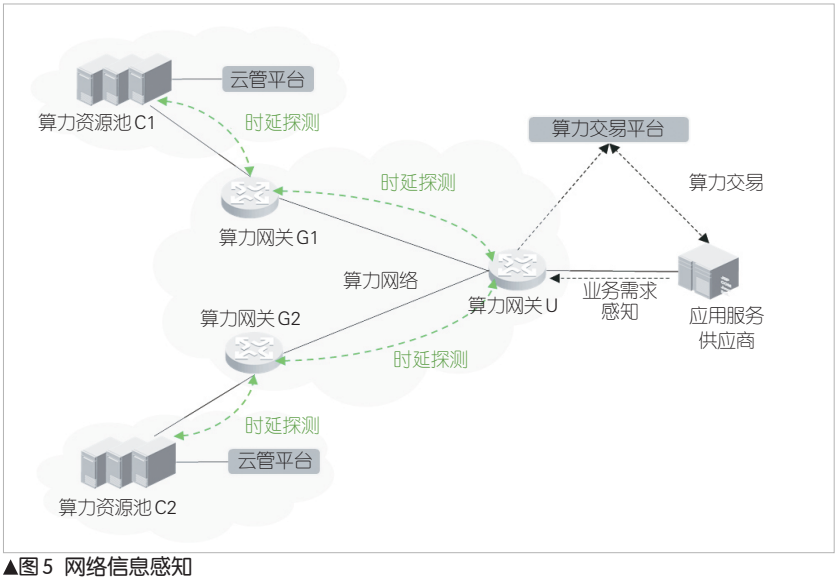
▲图2 算力网关混合式组网



▲图3 算力网关分布式组网



▲图4 算力信息感知



▲图5 网络信息感知

和实时地对算网进行调度，达到算力和网络的最优化。

业务应用需求可通过特定的协议或字段来与算力网关交互，从而实现算力网络对用户业务需求的感知。IPv6协议增加了扩展头部，具有很强的扩展性，可以在用户侧数据包头采用IPv6标准头+目的选项报头（DOH）扩展头的方式，利用扩展的字段携带应用的需求信息，包括带宽需求、时延需求、抖动需求、丢包率需求、计算和存储需求等[7]。

2.2 算力路由能力

算力路由是将算力信息引入路由域，通过对用户的业务需求、算力资源和网络资源的信息感知，动态选择满足业务需求的“转发路径+目的服务节点”，将业务沿指定路径调度至服务节点，从而实现算力和网络资源的全局优化。

算力路由技术可分为算力路由控制技术和算力路由转发技术。根据实现方式不同，算力路由控制技术又可以分为集中式控制和分布式控制。算力路由转发技术需要支持根据算力路由生成的“转发路径+目的节点”来指导业务转发，并且能够根据算力资源和网络状况的变化，动态调整控制面信息。

1）集中式算力路由控制

集中式的算力路由控制主要依托于上层算力交易平台及软件定义网络（SDN）控制器协同：先通过SDN控制器采集算力网络拓扑，再根据算力平台的算力匹配结果向算力网关下发SRv6 Policy，通过网络编排的方式形成算力与用户之间的路由控制，如图6所示。

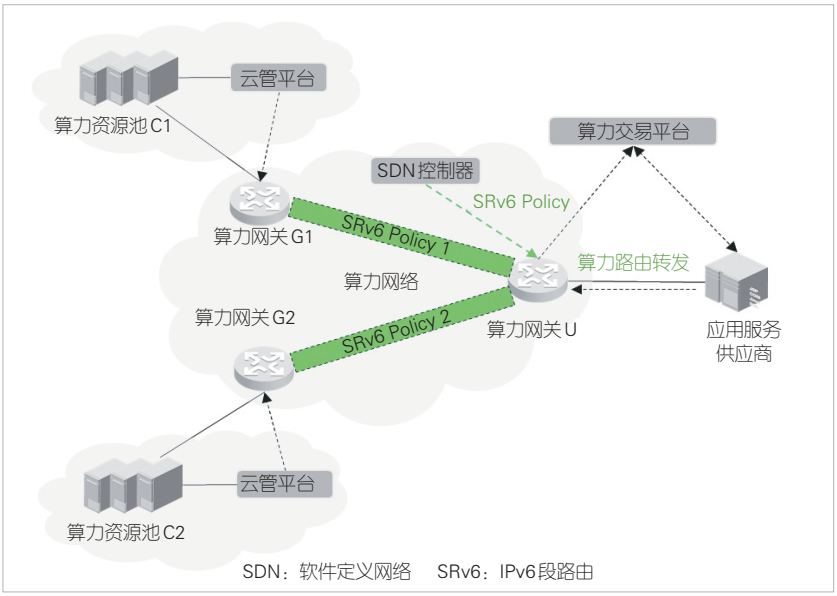
在SRv6网络里，业务需求可以被翻译成有序的指令列表，由沿途的网络设备去执行，以实现网络业务的灵活编排和按需定制。SRv6网络主要有SRv6 BE（指用最短路径算法计算得到的最优SRv6路径）和SRv6 Policy两种引流技术。SRv6 BE通过内部网关协议（IGP）收敛得出最短路径，业务无法按照指定的路径转发。SRv6 Policy可以在网络中任意节点之间规划路径。因此，使用SRv6 Policy不仅能够满足用户网络在时延、带宽、抖动和可靠性等各方面的差异化需求，还能够基于确定性路径的精细化控制来提高网络带宽的利用率[6]。

在集中式路由控制场景中，通过SRv6 Policy技术既能够实现算力网络的编排，保障算力资源与用户之间的路径确定性，又可以根据算力的实时变化实现算力的控制与调度。

2）分布式算力路由控制

分布式控制需要算力网关将感知的算力信息和网络信息进行通告，并且在用户入口的网关生成算力路由表项，形成用户业务需求与算力资源的协商和映射。而这种机制需要依靠特定的算力路由协议。

算力路由协议是实现算力路由控制和调度的关键技术。算力路由协议需要支持将感知的算力信息和网络信息在算力网关之间通告，并且在用户入口网关支持算力路由表的生成与更新，即基于通告的算力节点信息生成算力状态拓扑，进

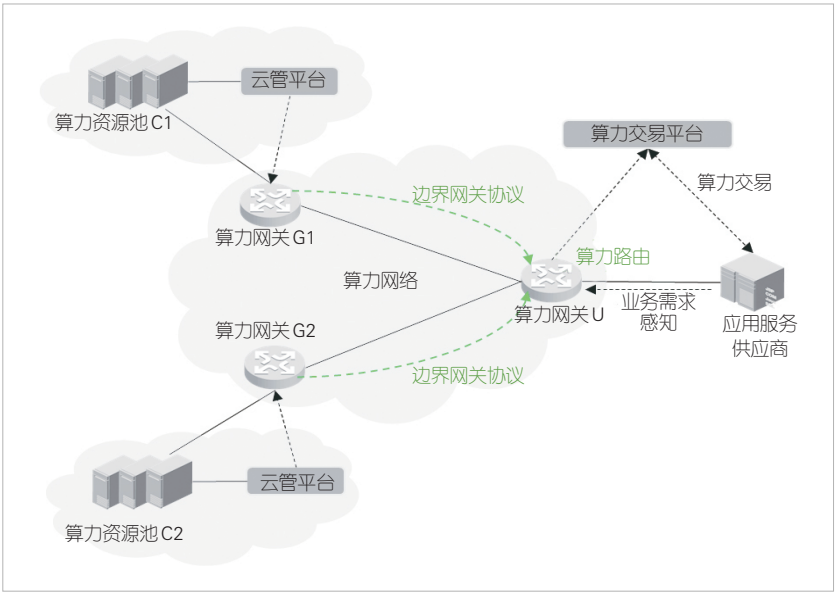


▲图6 支持SRv6的集中式算力路由控制

一步生成算力感知的新型路由表，用于支持后续业务转发。算力路由协议可以通过扩展基础路由协议来实现算网信息的通告。

以边界网关协议（BGP）为例，基于BGP的多协议扩展可利用BGP update消息中的路径属性预留字段TLV（一种可变格式）来扩展传递算力信息和网络信息。这种扩展的BGP协议就是算力边界网关协议（CP-BGP）。使用CP-BGP协议的算力路由控制如图7所示。

在算力网络中，算力资源池侧的算力网关可以感知算力节点的算力信息和网络信息，将相应信息填充到扩展的BGP update报文中，并通告用户侧的算力网关。用户侧算力网关可以接收扩展的BGP协议报文，解析算力信息和网络信息并生成BGP算力路由表。



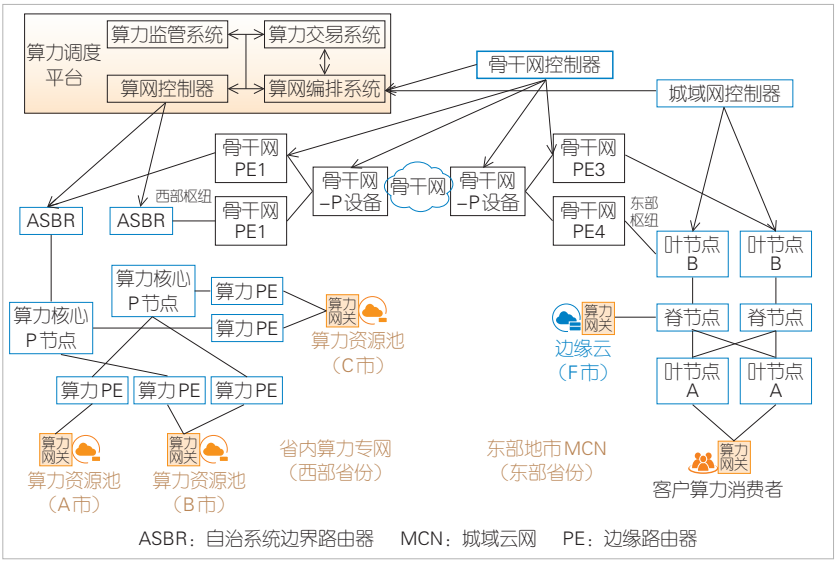
▲图7 支持算力边界网关协议的分布式算力路由控制

3 面向东数西算场景的实践

算力网关设备目前已经在“东数西算”业务场景成功落地应用，可以将东部算力需求有序引导到西部，促进东西部协同联动。

3.1 建设实施方案

算力网关的实践方案主要包括网络层面、管控层面两大部分，如图8所示。其中，网络层面包括西部的省内算力调度专网、运营商骨干网络以及东部的城域网络，管控层面包括西部省内的算网调度平台、骨干网控制器以及城域网控制器。各资源域网络控制器对接算网调度平台中的算网编排系统，同时基于部署在各资源池节点的算力网关设备，获取算力感知信息和算力路由信息，实现对云网资源的全局统一管控和调度。



▲图8 面向东数西算场景的算力网关实践方案

网络方案设计采用核心层和接入层两层架构，全路由器组网，如图9所示。

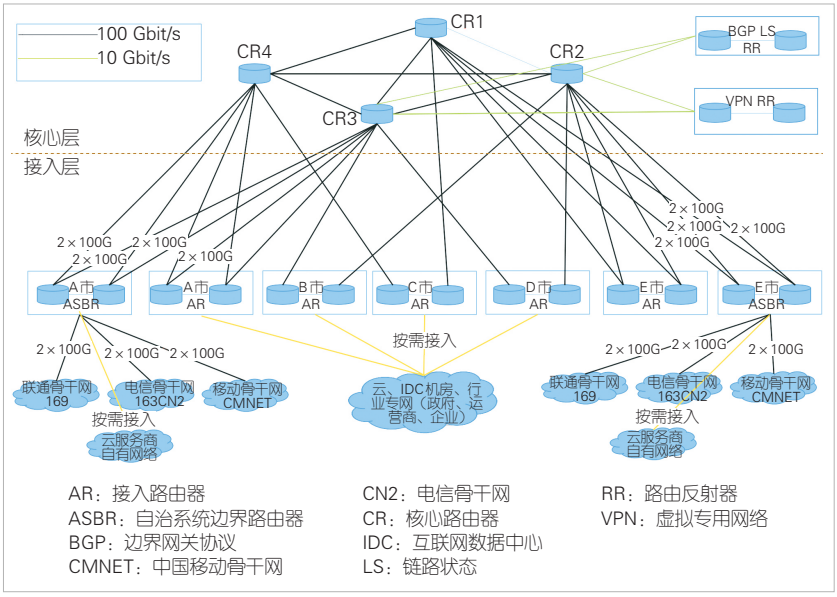
核心层核心路由器（CR）互联各市的接入路由器（AR）节点，虚拟专用网络（VPN）路由反射器（RR）负责VPN业务路由反射，BGP LS RR负责上送SR-TE信息。接入层每个地市部署2台AR，对接行业专网、IDC网络，并互联各云资源池。A地市和B地市各部署2台ASBR，对接各运营商骨干网络及云服务商自有网络。

核心层CR路由器之间采用Full Mesh互联。VPN RR和BGP-LS RR接入CR2和CR3。对于A地市和B地市的AR及ASBR，每组通过8条100GE线路交叉互联至属地CR路由器。对于其他地市AR，每组通过2条100 GE线路上联至核心层路由器。其中，一条上联A地市，另一条上联B地市。

路由协议设计采用公有AS号，并配有相应的IPv4/IPv6地址。网络架构采用SRv6技术路线，通过以太网虚拟专用网络（EVPN）统一业务面协议，并部署SRv6流量工程（SRv6-TE）。所有设备通过OpenAPI接口与控制器对接，并通过Telemetry上送网络运行数据。

3.2 实践成效

本次实践基于算力网关和算力调度平台，通过中国电信主导的西部多云协同算力调度专网、东部城域云网、骨干网



▲图9 面向东数西算场景的算力网络架构设计

以及各资源域网络控制器，对接统一云网编排系统，实现东西部之间的三维空间重构、实时云渲染等业务场景的全局可视调度。

本次算力网关的落地实践具有重大意义：一方面，中国电信在东数西算领域开展了创新尝试。省内算力专网及算力资源调度的实践充分验证了算力网关落地的可行性。另一方面，借助算力调度平台能够实现西部省份算力资源的统筹调度，打造全栈算力服务，全面提升信息技术（IT）资源利用率，助力产业数字化及数字产业化的发展[8]。

本次实践表明，算力网络和算力网关有助于实现算力设施由东向西布局，未来将带动相关产业有效转移，促进东西部数据流通、价值传递，延展东部发展空间，推进西部大开发形成新格局，提升国家整体算力水平。

4 结束语

算力网关通过网络控制面分发服务节点的计算能力、存储、算法等资源信息，力图打破传统网络的界限，将网络传送能力与IT的计算、存储等基础能力更好地结合起来，实现整网资源的最优化配置和使用，推动网络从“泛在连接能力平台”向“融合资源供给平台”升级演进。

算力网关的落地应用有助于实现算网一体化服务，有效提升资源利用率，减少网络资源和计算资源的浪费，降低整体能耗，助力东数西算战略落地[9]。

致谢

本文相关技术应用由中国电信股份有限公司、中电万维信息技术有限责任公司、中兴通讯股份有限公司、英特尔（中国）有限公司等单位共同完成。解云鹏、高守纪、乔建、田毅、何秀文、段敏等人承担了大量研发和试验工作。在此，向他们表示感谢！

参考文献

[1] 李正茂, 雷波, 孙震强, 等. 云网融合: 算力时代的数字信息基础设施 [M]. 北京: 中信出版集团, 2022
[2] 中国工信产业网. “四力”汇聚，算力网络发展迈入快车道 [EB/OL]. [2023-06-10]. https://www.cnii.com.cn/tx/202303/t20230320_455966.html
[3] 网络通信与安全紫金山实验室. 白盒交换机技术白皮书 [R]. 2021
[4] Github. SONiC system architecture [EB/OL]. [2023-06-10]. https://github.com/sonic-net/SONiC/wiki/Architecture
[5] 赵倩颖, 邢文娟, 雷波, 等. 一种基于域名解析机制的算力网络实现方案 [J]. 电信科学, 2021, 37(10): 86-92. DOI: 10.11959/j.issn.1000-0801.2021233
[6] 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
[7] DEERING S, HINDEN R. Internet protocol, version 6 (IPv6) specification [J]. RFC, 1995, 2460: 1-39. DOI: 10.17487/rfc8200
[8] 解云鹏, 马思聪, 田毅, 等. 从“东数西算”甘肃节点 看中国电信的算力调度探索与实践 [J]. 通信世界, 2022(22): 34-37
[9] 国家发展改革委, 中央网信办, 工业和信息化部, 等. 关于加快构建全国一体化大数据中心协同创新体系的指导意见 [EB/OL]. [2023-06-10]. https://www.gov.cn/zhengce/zhengceku/2020-12/28/content_5574288.htm

作 者 简 介

马思聪，中国电信股份有限公司研究院高级工程师；主要研究领域为未来网络、云网融合下的算力网络技术；主持和参与算力网关试点验证与研发工作；发表论文3篇。

孙吉斌，中国电信股份有限公司研究院研发工程师；主要研究领域为未来网络关键技术、算力路由协议等；先后参与“东数西算”场景下的算力网关试点部署与研发工作，以及算力路由协议标准的制定工作。

孙一豪，中国电信股份有限公司研究院研发工程师；主要研究领域为算力网络、IPv6标准和关键技术等。