算网融合关键技术和发展路径研究

Research on Key Technologies and Development Paths of Computing and Network Convergence

李振文，李　芳，赵俊峰（中国信息通信研究院，北京 100191）
Li Zhenwen，Li Fang，Zhao Junfeng（China Academy of Information and Communications Technology，Beijing 100191，China）

摘　要：
为实现算力和网络资源的统一纳管和融合路由调度，业界已经进行了积极的研究和探索，并推动制定了算网融合的整体框架，具体的技术和标准也在研究和制定，但由于涉及到异构算力的统一度量和算力交易等流程，实现复杂度较高，建议基于现有云、算侧和网侧的调度系统以及SRv6、APN、CFN、RDMA等关键技术，采用边研究边实践的策略，分3个阶段逐步推进，最终实现算网融合的目标架构。

关键词：
算网融合；APN6；CFN
doi：10.12045/j.issn.1007-3043.2024.02.006
文章编号：1007-3043（2024）02-0031-05
中图分类号：TP391
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
In order to realize unified management of computing power and network resources and converged routing scheduling，the industry has carried out active research and exploration，and promoted the formulation of the overall framework of computing network convergence. Specific technologies and standards are also in the process of development. However，due to the unified measurement of heterogeneous computing power and the processing of computing power transaction，the implementation complexity is relatively high. It is recommended to adopt a strategy of researching while practicing based on existing cloud，computing side，and network side scheduling systems，as well as key technologies such as SRv6，APN，CFN，RDMA，etc.，step by step in three stages，and finally to achieve the target architecture of computing and network convergence.

Keywords：
Computing and network convergence；APN6；CFN

引用格式：李振文，李芳，赵俊峰. 算网融合关键技术和发展路径研究[J]. 邮电设计技术，2024(2)：31-35.

0 引言

“东数西算”工程是我国为促进信息基础设施优化升级、推动数字经济加速发展而提出的一项重大战略工程，而“东数西算”工程要实现算力全国调度，就需要算网融合的支撑。所谓算网融合，是以通信网络设施和计算设施的融合发展为基础，通过计算、存储及网络资源统一编排管控，满足业务对网络和算力灵活泛在、弹性敏捷需求的一种新型业务模式。在此背景下，算网融合的架构和技术成为业界研究热点。

收稿日期：2024-01-09

1 算网已有架构和调度技术分析

1.1 算网融合是实现云、算、网资源的统一管理和调度

算网融合本质上希望打破云计算、存储资源和网络资源各自独立、无法协同的现状。运营主体和服务方式方面，算网融合的运营者除电信运营商之外，还有云厂商和第三方企业；运营者可提供多样化网络接入，具备算力感知、一体化管理和编排调度能力，可实现算网服务的弹性供给、自主定制、随需交易；支撑技术方面，算网融合既需要SDN、NFV以及转发面的VxLAN、EVPN、SR/SRv6等现有技术的增强，也需要新技术如算网统一度量和交易、编排调度、算力资源发布

以及APN6、CFN、RDMA等技术的支撑。

1.2 云、算侧资源管理与调度架构

随着以容器和微服务为代表的云原生技术的发展，算力资源统一管理和调度技术成为目前行业研究热点，当前应用较多的算力调度系统以超算和HPC的资源调度为主，主要有IBM公司的LFS、Altair公司的PBS pro以及开源的Slurm等。面向大模型训练等智算场景，微软在其CycleCloud上将超算算力调度系统和云的Kubernetes进行结合，为用户提供可专用于AI大模型训练的环境。此外国内企业也已经开始了对算力调度系统的研究，并推出了如Quick Pool、SkyForm等产品。Slurm在科研机构和院校中应用较多，其架构如图1所示，采用Slurmctld服务监测资源和作业。各计算节点启动Slurmd守护进程，被作为远程shell使用（等待作业、执行作业、返回状态、再等待更多作业）。Slurmdbd（Slurm DataBase Daemon）数据库守护进程，将多个Slurm管理的集群的记账信息记录在同一个数据库中。用户可以使用一系列命令工具如Srun（运行作业）等对作业进行管理。另外还可以通过Slurmrestd（Slurm REST API Daemon）服务，使用REST API与Slurm进行交互。节点是Slurm调度的单位之一，每个节点都有自己的资源，如CPU、内存、GPU等。节点由Slurm自动分配给作业，通常只需要用户指定数量。但如果有特别的需要，用户也可以直接给定节点列表或者用参数排除一些节点。

Kubernetes也是一个开源平台，用于管理容器化的工作负载和服务，在大规模集群的资源管理中应用广泛（见图2）。Pod是在Kubernetes集群中运行部署应用或服务的最小单元，可支持多容器。Node节点主要作为计算节点，实现本地Pod的部署运行和相关计算、存储和网络资源的纳管[1]。在Kubernetes中，通过调度将Pod放置到合适的Node节点上，调度器通过Kubernetes的监测机制来发现集群中尚未被调度到节点上的Pod。它会依据提前设置的调度原则来做出调度选择。kube-scheduler是Kubernetes集群的默认调度器。kube-scheduler给一个Pod做调度选择时包含过滤和打分2个步骤，其中过滤阶段会过滤掉候选节点中不满足可用资源需求的节点，形成可调度节点列表，而打分阶段，调度器会根据预设的打分规则为每一个可调度节点打分，最终选出一个最合适的节点来运行Pod。在做调度决定时需要考虑的因素包括单独和整体的资源请求、硬件/软件/策略限制、亲和以及反亲和要求、数据局部性、负载间的干扰等。

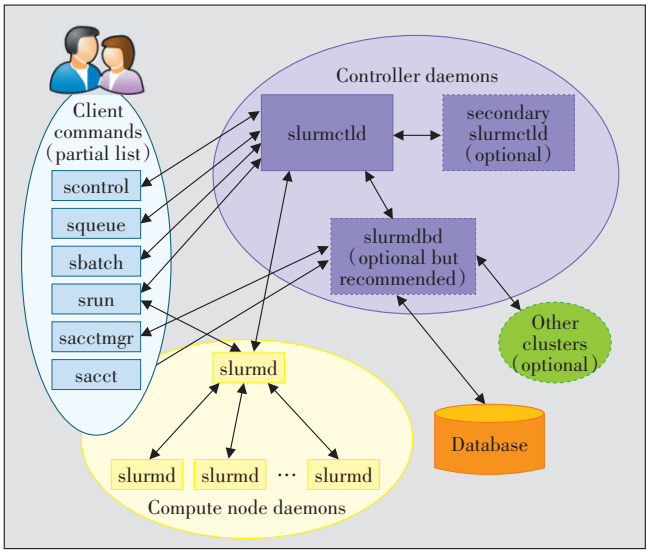

图1　Slurm架构

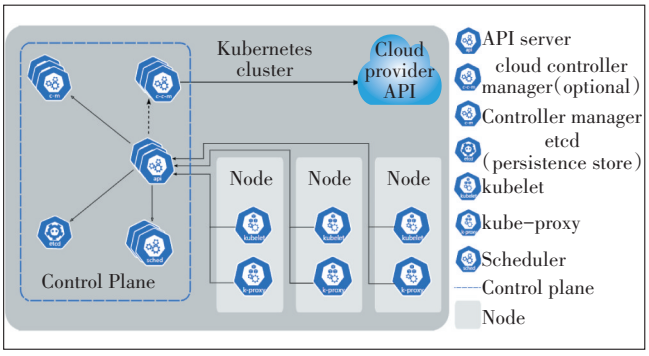

图2　Kubernetes集群的组件

1.3 网侧资源管理与调度架构

VxLAN+EVPN方案是数据中心网络的重要部署方案。VxLAN技术通过将原始报文封装在UDP报文中，可以将传统的二层网络扩展到三层网络，实现数据中心网络的虚拟化，提高网络的可扩展性和灵活性。EVPN技术则是一种基于BGP的以太网虚拟专用网技术，利用EVPN构建VxLAN的控制平面，解决VxLAN需要通过泛洪的方式学习终端主机地址的问题，从而提供跨数据中心的数据传输和VPN服务。

同时，VxLAN和SDN联合部署已经成为智能化云数据中心的必要组件，VxLAN作为数据平面解耦租户网络和物理网络，SDN将租户的控制能力集成到云管平台，与计算、存储资源联合调度，提升了数据中心内业务承载的灵活性（见图3）。

1.4 小结

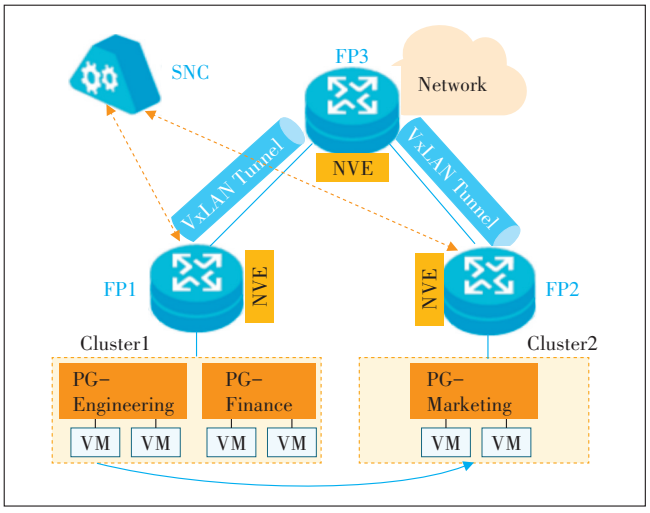

图3　SDN+VxLAN数据中心网络承载方案

云、算侧算力调度系统实现了集群内算力任务和容器化资源的调度管理，在进行负载均衡时可以考虑CPU、内存和网络带宽利用率等因素，并且通过调度算法的不断优化，使得集群内节点的利用率更高，但是这里的网络资源信息还相对粗放，没有精确的带宽、时延等信息，使得用户获取到的算力服务路径不一定是最优路径，这个问题同样存在于DNS域名解析服务器进行终端请求的应答过程中。

在网络侧，VxLAN+EVPN作为Overlay的方案，较好地解决了数据中心间虚机迁移的问题，但同时也存在无法支撑将Underlay网络资源的信息与算力资源信息融合到一起进行调度的问题，所以为了更好地支撑算网融合，需要SRv6等更具有潜力的网络技术。另外，针对AI分布式训练和HPC高性能计算场景，RDMA技术也已经被广泛应用于智算集群内的互联。

2 算网融合目标架构和关键技术分析

2.1 整体目标架构相关标准进展

中国三大运营商、设备商、服务器厂商等在CCSA立项了《算力网络总体技术要求》，目前已完成报批稿，主要规定了算力网络的总体技术架构和技术要求，包括算力网络的总体架构和接口描述，以及算力服务技术要求、算力路由技术要求、算网编排管理技术要求等，其中算力网络总体功能逻辑架构如图4所示。

为了实现对算力和网络的感知、互联和协同调度，算力网络架构体系从逻辑功能上划分为算力服务层、算力路由层、算网管理层、算网基础设施层四大功能模块，具体如下。

a）算力服务层。提供算力的各类能力及应用，并将用户对业务SLA的请求（包括算力请求等参数）传递给算力路由层。

b）算力路由层。基于抽象后的计算资源发现，实现对算力节点的资源信息感知；另一方面，通过在用

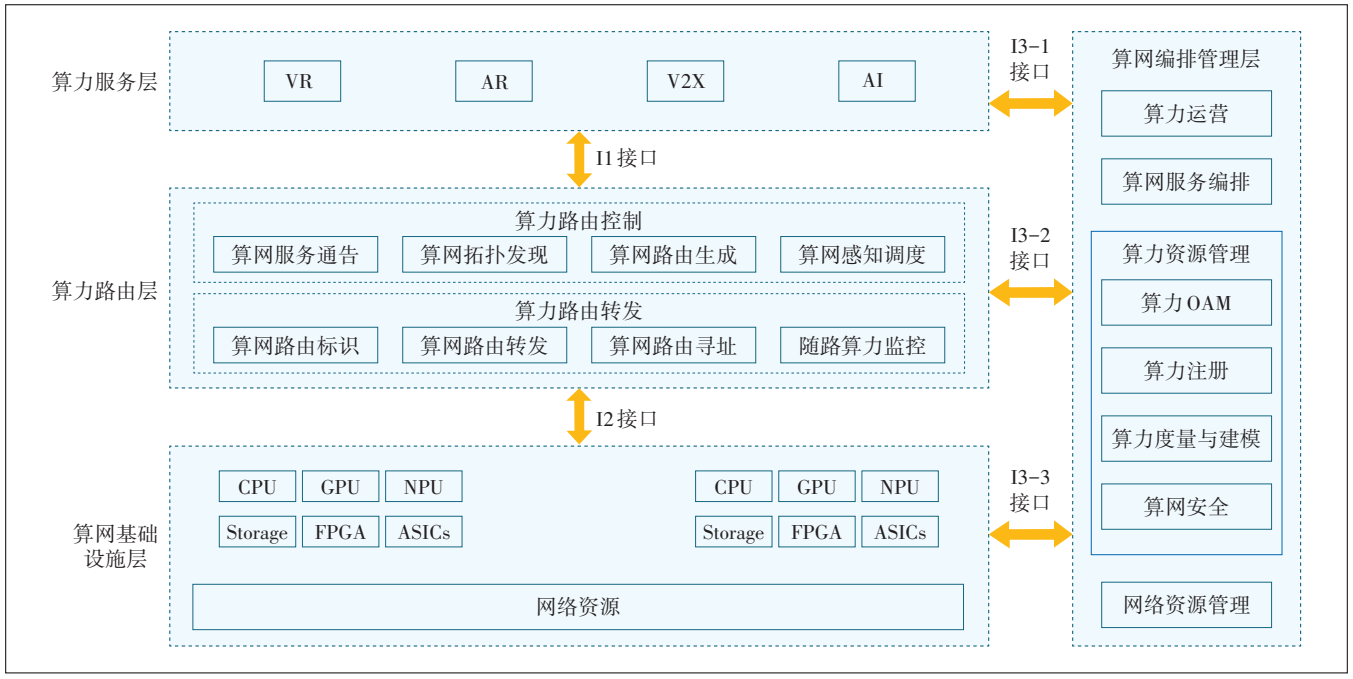

图4　算力网络总体功能逻辑架构

户请求中携带业务需求，实现对用户业务需求的感知。综合考虑用户业务请求、网络信息和算力资源信息，将业务灵活按需调度到不同的算力节点中，同时将计算结果反馈到算力服务层。算力路由层的部署实现支持集中式方式和分布式方式。

c）算网编排管理层。实现对算力服务的运营与编排管理、对算力路由的管理、对算力资源的管理以及对网络资源的管理，其中算力资源管理包括基于统一的算力度量衡体系，完成对算力资源的统一抽象描述，进而实现对算力资源的度量与建模、注册和OAM管理等功能；以支持网络对算力资源的可感知、可度量、可管理和可控制。

d）算网基础设施层。为满足新兴业务的多样性计算需求，基于提供信息传输的网络基础设施，在网络中提供泛在异构计算资源，包括单核CPU、多核CPU、CPU+GPU+FPGA等多种算力组合。其中算网基础设施层作为算力网络的新型基础设施层，算力服务层、算力路由层和算网编排管理层作为实现算力网络可感、可控、可管的三大核心功能模块，实现对算力和网络资源的感知、控制和管理[2]。

2.2 支撑算力运营和交易的关键技术

2.2.1 算力资源建模，包含算力度量、算力分级等

算力是设备或平台为完成某种业务所具备的处理业务信息的关键核心能力，根据所运行算法和所涉及的数据计算类型不同，可将算力分为逻辑运算能力、并行计算能力和神经网络计算能力。算力的统一量化是算力调度、使用的基础。对不同的计算类型，不同厂商的芯片有各自不同的设计，这就涉及异构算力的统一度量。不同芯片所提供的算力可通过度量函数映射到统一的量纲。

算力分级可以供算力提供者设计业务套餐时参考，也可作为算力平台设计者在设计算力网络平台时对算力资源的选型依据。智能应用对算力的诉求主要是浮点计算能力，因此业务所需浮点计算能力的大小可作为算力分级的依据。当前算力可分为超大型算力、大型算力、中型算力和小型算力4个等级。

2.2.2 算力交易

泛在计算的算力交易平台是一套基于区块链的去中心化、低成本、保护隐私的可信平台。平台的计算节点由多种形态的算力设备组成，包含大型GPU设备或FPGA服务器集群、中小型企业闲散的空余服务器及个人闲置的计算节点等。平台可以实现自动算力交易、自动算力匹配、费用结算功能。在算力卖家向算力买家提供服务的过程中，后者提出使用请求，算力交易平台根据用户需求自动寻找、匹配算力节点，并生成相应的账单；在得到买家认可后，平台调度相应的算力资源为买家提供服务，随后执行算力业务的节点根据提供的算力获得相应的报酬。

2.3 支撑算网资源融合管理调度的关键技术

2.3.1 算网转发技术——SRv6

SRv6是源路由技术的一种，它采用现有的IPv6转发技术，通过灵活的IPv6扩展头，实现网络可编程。为了实现SRv6转发，需要向IPv6报文中插入一个段路由头（Segment Routing Header，SRH）的扩展头，存储IPv6的Segment List信息。报文转发时，依靠Segments Left和Segment List字段共同决定IPv6目的地址（IPv6 DA）信息，从而指导报文的转发路径和行为。未经压缩的SRv6 SID是128位，主要由标识节点位置的LOC字段（IPv6前缀格式，可路由）、标识服务和功能的FUNC字段（本地识别）以及ARG字段3个部分组成。SRv6网络编程标准中，SRv6节点（Endpoint）通过本地定义的行为（Behavior）处理SRv6报文。SRv6定义了多种Endpoint Behavior，每个节点需要实例化它们并分配SID，同时通过路由协议发布，以通知其他SRv6节点本节点能提供的Behavior。常用的Endpoint Behavior有END、END.X、END.DT4、END.DT6等，实现Underlay选路、Overlay业务承载等功能[3]。

2.3.2 算网感知技术——APN6

APN6是在数据平面利用IPv6报文扩展头（Extension Headers），如逐跳选项头（Hop-by-Hop Options Header）、段路由头（Segment Routing Header）的可编程空间，携带应用的相关信息（标识和需求）到网络中，网络设备依据这些信息为其提供相应的网络服务，如将报文映射进相应的能够保障其SLA的SRv6路径等。应用感知信息可以由用户终端设备或应用直接生成，也可以由网络边缘设备生成，分别对应APN6的主机侧方案和网络侧方案[4]。

2.3.3 算网融合路由技术——CFN

为了解决边缘计算系统中网络信息和算力信息割裂，无法统一纳管和进行最优资源调度的问题，Yizhou Li等提出了CFN的概念，并在IETF提交了草案：Framework of Compute First Networking（CFN）[5]，架构和原理如图5所示。

CFN网络按角色分为服务器节点、CFN节点和客

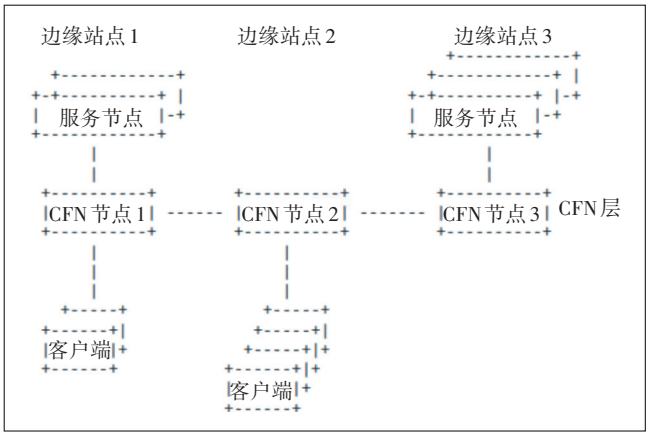

图5　CFN网络拓扑

户端。CFN通过控制面完成算力资源信息的全网同步。服务节点将本地服务状态注册到CFN节点的数据库表项中。本地服务状态一般包括服务的唯一标识(Service ID)、服务IP地址和计算资源情况等。CFN节点将本地服务状态封装到CFN路由协议报文中并扩散到其他CFN节点。CFN节点基于CFN路由协议将本地以及收到的其他CFN节点扩散的服务状态信息汇总生成服务信息路由表。CFN数据面完成客户端对服务节点Service ID请求的路由转发。与客户端距离最近的CFN节点收到请求后，根据网络资源、计算资源情况进行综合评估，选择一个服务节点以及相关联的CFN出口节点，将原请求数据包封装并发送。CFN Egress节点收到数据包，根据Service ID查找对应服务节点IP，将数据封装并发送。外层数据包源地址为客户端IP，目的地址为服务节点IP。报文封装的内层数据包源地址为客户端IP，目的地址为Service ID。服务节点收到数据包后在本地查询与Service ID绑定的服务地址，调用对应的服务，将结果返回给客户端[3]。

3 结束语

在我国提出“东数西算”的大背景下，我国电信运营商希望借助政策发展的契机，在售卖网络管道和出租数据中心基础资源的同时，释放更多的管道潜能，所以积极投入算力与网络相融合的研究中，并在国际、国内标准组织推动制定了一系列算网融合的标准架构，但要真正实现算网融合的规模商用，无论是商业模式还是技术实现细节上都还存在较大差距。上述标准框架中，目标架构和业务流程都比较完善，但同时这种非常完善的架构也会带来系统复杂度的大幅增加。由于要将CPU、GPU、FPGA以及内存和存储等异构算力资源进行归一化度量，需要研究算力的度量标准；另外，还需要建设算力交易平台，解决算力的交易问题并进行标准化。从实现路径上，建议基于现有云、算侧和网侧的调度系统和SRv6、APN和CFN、RDMA等关键技术，采用边研究边实践的策略，分3个阶段逐步推进。

第1阶段：单运营商场景。运营商内部负责云和网络的运营团队间不考虑算力资源交易和结算流程，这样一方面简化了算力运营和交易相关平台的实现，另一方面，从流程上简化了算力需求者提出需求后，在进行算力资源匹配后交易确认环节引入的处理时延。算力资源池也限制运营商的自有资源，减少资源种类，更易进行度量。

第2阶段：单运营商、单云场景。运营商内部负责云和网络的运营团队间，以及运营商和第三方云供应商之间基于算力运营和交易平台，实现了算力资源的交易和结算；算力资源池也拓展至本运营商的自有算力资源和第三方云供应商的算力资源。

第3阶段：多运营商、多云场景。不同运营商间、运营商与第三方云供应商间都实现了算力运营和交易，运营商既可以是算力资源的购买者，也可以是算力资源的售卖者；同时，一些企业和个人终端的零散算力资源也可以进行交易。

参考文献：

[1] 曹畅，唐雄燕，张帅，等. 算力网络[M]. 北京：电子工业出版社，2021：119.
[2] 中华人民共和国工业和信息化部. 算力网络 总体技术要求：YD/T 4255-2023[S]. 北京：人民邮电出版社，2023：6-7.
[3] 曹云飞，霍龙社，何涛. 基于SRv6的可编排计算优先网络实现方法[J]. 邮电设计技术，2022(4)：4-9.
[4] 感知应用的IPv6网络(APN6)架构研究：2020B84[S]. 北京：中国通信标准化协会，2021：10.
[5] LI Y. Framework of Compute First Networking(CFN)draft-li-rtgwg-cfn-framework-00[EB/OL]. [2023-10-22]. https://datatracker.ietf.org/doc/html/draft-li-rtgwg-cfn-framework-00.

作者简介：
李振文，工程师，主要从事5G承载、分组传送、算力网络等方面的研究工作；李芳，教授级高级工程师，主要从事5G承载、分组传送、算力网络等方面的技术与标准研究工作；赵俊峰，高级工程师，主要从事5G承载、分组传送、确定性网络、算力网络等方面的技术与标准研究工作。