云网一体化数据中心网络关键技术

王江龙，雷波，解云鹏，何琪，李云鹤
（中国电信股份有限公司研究院，北京 102209）

摘　要：“新基建”带来数据中心发展的新契机，数据中心作为运营商打造云网一体化新型基础设施的重要组成部分，为 5G+AI 时代的新兴应用提供强大的技术底座。重点分析面向云网融合发展的数据中心网络的新需求，对超大规模组网、超低时延无损网络、业务端到端统一承载等关键技术进行探讨；最后，给出一种大规模云数据中心组网和承载方案，并对其中的关键技术领域进行验证。

关键词：大规模数据中心；云网融合；无损网络；EVPN/VxLAN；SRv6
中图分类号：TP393
文献标识码：A
doi: 10.11959/j.issn.1000−0801.2020118

Key technology of data center network for cloud network convergence

WANG Jianglong, LEI Bo, XIE Yunpeng, HE Qi, LI Yunhe
Research Institute of China Telecom Co., Ltd., Beijing 102209, China

Abstract: “New infrastructure” brings new opportunities for the development of data centers. Data center has become an important part of the operator’s cloud-network convergence infrastructure, providing a strong technical foundation for innovative applications in the 5G + AI era. The new requirements of the data center network in the cloud network convergence background were analyzed. The key technologies in ultra-large-scale networking, ultra-low latency lossless network, and unified service end-to-end bearer were given. Finally, a large-scale cloud data center networking solution were presented, and the key technologies were verified.

Key words: large-scale data center, cloud-network convergence, lossless network, EVPN/VxLAN, SRv6

1　引言

中共中央政治局常务委员会在2020年3月4日召开会议，明确指出，加快5G网络、数据中心（data center，DC）等新型基础设施建设进度（以下简称“新基建”）。“新基建”作为发力于科技端的基础设施建设，将成为新一轮数字化经济发展的催化剂，也将是未来科技创新竞争力的核心所在。数据中心首次被国家列入加快建设的条目，作为“新基建”七大领域中的一大亮点，将协同5G网

收稿日期：2020−03−10；修回日期：2020−03−26
基金项目：国家重点研发计划基金资助项目（No.2018YFB1800100）
Foundation Item: The National Key Research and Development Program of China（No.2018YFB1800100）

络、工业互联网等基础设施加速新一代信息技术发展。

数据中心作为数字经济领域基础设施，会随着“新基建”的推进而持续扩大规模。一方面，政务、金融等传统企业，通过建立数据中心或者使用公有云服务部署促进企业数字化转型，实现企业上云；另一方面，数据中心将成为新兴信息技术应用的核心载体，为 5G、人工智能、物联网、大数据等新兴产业提供强大的技术底座。

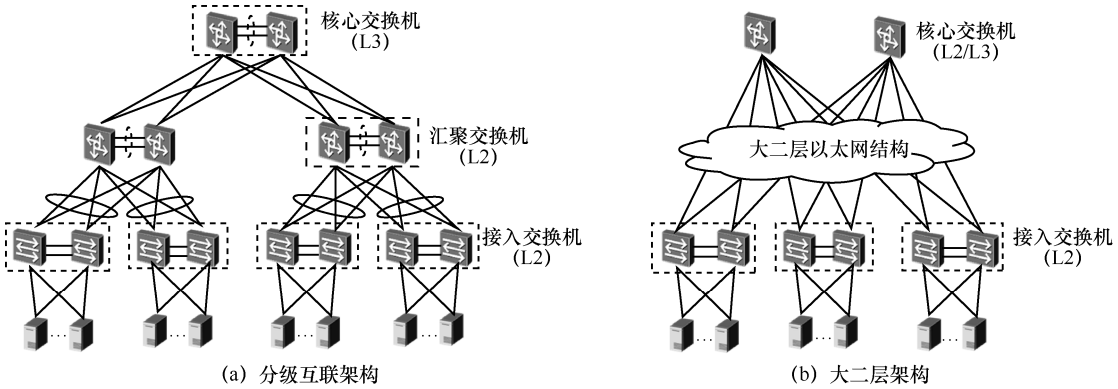
中国电信提出云改战略，全面推进网络上云、IT 上云、业务上云。同时结合网络优势，把云网融合作为一个战略性的、网络型的基础设施大力推进。数据中心作为运营商云网一体化基础设施，以“2+31+X”的数据中心/云布局为核心，构建高质量承载网络，为客户提供包含计算、存储和连接的整体服务。数据中心网络将成为运营商的“核心网络”，承载对外销售互联网数据中心（IDC）、公有云、私有云、电信云等多类型业务。同时，数据中心网络也将和城域网、骨干网高效协同，将应用、云计算、网络及客户灵活连接起来，提供一个完整的端到端业务方案。数据中心网络在运营商云网一体化规划布局中起着非常重要的作用。

2 传统数据中心网络架构

目前，传统的数据中心内部组网架构以分级互联架构和大二层架构为主，如图 1 所示。

分级互联架构，即传统的“接入—汇聚—核心”的三层组网结构。汇聚往下二层组网，为一个独立的 VLAN/STP 分区网络。不同的分区必须经过核心交换机通过三层网络互访。汇聚及核心层交换机横向采用堆叠等私有虚拟化技术，保证业务可靠性。DC 内以南北向流量为主，可承载上百台服务器规模，主要提供对外互联网访问的业务。该架构在目前运营商数据中心中仍是主要网络结构，对外销售型 IDC 业务、规模较小的资源池基本上还是以此架构部署为主。

大二层架构指资源池化逻辑大二层网络架构，是近些年来云计算、大数据等分布式技术在数据中心大规模部署后使用的一种网络架构。资源池虚拟化后，虚拟机管理迁移等需求使得数据中心内二层的东西向流量大量增长。因此，数据中心组网也出现了很多新的网络技术，比如多链接透明互联（TRILL）、最短路径桥接（SPB）等通过路由计算实现的大二层组网技术、可扩展虚拟局域网（virtual extensible local area network，VxLAN）、使用通用路由封装的网络虚拟化（NvGRE）等 overlay 技术。随着虚拟化数据中心规模的扩大，基于 VxLAN 的大二层网络架构成为主要部署方案，网络结构也趋向扁平化，在接入交换机和核心交换机之间 full-mesh 连接，三层路由应用于二层多路径，构建大二层网络。该架构主要面向计算资源虚拟化以及存储网络和 IP 网



(a) 分级互联架构　　(b) 大二层架构

图 1　传统数据中心网络架构

络融合的资源池，可承载上千台服务器。

随着数据中心承载业务类型的不断增加，传统数据中心网络出现了很多问题。一方面，网络架构、关键技术的选择导致规模可扩展性差；另一方面，数据中心内网络业务需求不再单一，不再仅仅以带宽保障为核心驱动，面向各种新兴业务的网络能力的差异化承载、高效智能的运维管理都面临着新的挑战。

3　未来数据中心网络的主要挑战及需求

（1）企业上云要求数据中心网络具备云网融合能力

云计算的发展加速了企业的数字化转型，企业上云是大势所趋。对于中小企业，不仅能减小成本投入，还能缩短业务上线周期。随着产业互联网的加速发展，政府、大中企业客户也成为云计算市场增长的主要拉动力量。对于政企大客户而言，更倾向于使用混合云成为上云的主要路径，一些敏感企业数据仍保留在私有云或者专属云中。目前，运营商纷纷开展云专网、云专线、SD-WAN 等云网融合业务，加强云网基础设施统一规划建设。与大网协同，公有云、私有云、混合云等连接对数据中心网络提出了更高的要求。随着云计算应用及混合云、多云的普及，要求数据中心网络具备云网融合能力，能够简化管理、自动开通、灵活部署，满足云网融合业务“一点受理、自动开通、统一运维、自助随选”等新型业务需求。

（2）5G 及 NFV 加速云化网元进入数据中心

5G 提出网络全面云化，采用以数据中心为基础的云化架构，承载在数据中心内独立的电信云网络上。一方面，5G 核心网云化部署，用户面功能（UPF）将下沉到边缘云。UPF 除了满足互联网、VoLTE 等传统业务需求外，还将分布在综合接入机房以及边缘数据中心，提供边缘计算（MEC）业务。另一方面，城域网网元功能逐步迁入 CT 云中，随着转控分离 vBRAS 部署，城域网业务控制面功能转移至数据中心进行云化部署。城域网业务从现在的网元通信也将演变成 CT 云资源池间的通信。可以预见，5G 云化网元进入数据中心后，城域网流量大幅增长，下沉节点也会成倍增加，用户流量就近入云，用户面流量的 mesh 化和本地化，会增加大量分布式东西向流量，云网一体化数据中心网络架构将考虑适应该流量模型的可扩展。

其次，5G 的部署还将催生车联网、人工智能、VR/AR 等大量新兴业务，其高带宽、低时延、多连接的特征也需要数据中心网络来满足。业务流量增长还将带动数据的指数增长，给数据中心带来海量计算、存储、智能分析、安全性等要求。

（3）AI 等新兴技术应用对数据中心提出超低时延无损需求

随着深度学习算法的突破，人工智能技术发展步入了快车道。深度学习依赖海量的样本数据和强大的计算能力，也推动高性能分布式存储和高性能计算的发展。高效的 AI 训练需要非常高的网络吞吐来处理大量的数据，大量的数据将会在计算、存储节点之间传送。通常情况，在低于 10% 链路带宽利用率的低负载流量环境下，流量突发引起的网络的分组丢失率也接近 1‰，而这 1‰的分组丢失在 AI 运算系统中直接带来的算力损失接近 50%。随着业务负载增加，数据中心分布式“多打一”流量逐步增多，网络分组丢失也越来越严重。分组丢失和时延引起的网络重传会进一步降低网络的吞吐量，使模型训练的效率大大下降，甚至导致训练的失败。

服务器技术的快速发展带动了数据中心计算、存储能力的飞速提高，随着存储介质读写速度和计算能力的提升，数据中心网络通信时延成为性能进一步提升的瓶颈。特别是面向 HPC、分布式存储、AI 应用等新型业务场景，传统数据中心以太网架构因拥塞易出现分组丢失带来的网络

传输瓶颈异常凸显。构建零分组丢失、超高吞吐、超低时延的无损网络，是未来数据中心网络的一大典型需求及特征[1]。

（4）边缘数据中心将成为数据中心新形态

边缘计算是近年的研究热点。在边缘计算技术逐步成熟后，数据中心的发展将呈现两极化，一方面资源逐步整合，云数据中心规模越来越大，对于大规模组网的性能要求越来越高；另一方面，将涌现大量边缘数据中心，以保障边缘实时性业务性能。云数据中心将时延敏感型业务卸载，交由边缘数据中心处理，减少网络传输带宽压力和往返时延。边缘数据中心负责实时性、热数据存储业务；云数据中心则负责非实时性、冷数据存储等业务。在网络中规模引入边缘计算后，如何进行云、网、边高效协同组网，合理分配全网算力资源，满足低时延业务需求，是边缘数据中心网络以及云边协同组网的一大挑战[2]。

总体上，云网一体化数据中心网络的主要需求集中在规模可扩展、超低时延无损特性、业务端到端承载、云网协同以及一体化管理和运维等方面。数据中心网络在应对云网一体化的挑战下，不仅要兼顾原传统 IDC 业务，还需考虑云网融合、5G、人工智能、边缘计算等新业务的综合承载能力。

4 关键技术

4.1 超大规模组网

云数据中心呈现超大规模发展趋势，其数据中心内组网通常需要实现几万台服务器集群作为单一资源池的通信。要满足大规模组网能力下的高可扩展、高可靠性，需要在架构设计、基础网络路由组织、自动化管控等维度来考虑。

（1）IP-CLOS 架构

CLOS（无阻塞多级交换网络）架构由美国加利福尼亚大学 Mohammad Al-Fares 提出[3]。IP-CLOS 脱胎于无阻塞的 CLOS 架构，在 Spine 交换机和 Leaf 交换机之间以 full-mesh 全三层连接，可以承载上万台服务器规模，理论上数据中心规模不再受限于网络，典型的 IP-CLOS 结构如图 2 所示。

Facebook 采用 IP-CLOS 构建其数据中心内网络，如图 3 所示。2014 年，Facebook 首次推出了 F4 数据中心架构，采用分层核心和 Pod 设计。基本构建块 Pod 包含 48 个 Leaf 交换机，通过 40 Gbit/s 链路汇聚到 4 台 Fabric 交换机。每个 Pod 通过 40 Gbit/s 上行链路连接到 4 个独立的 Spine 主干平面。Pod 和 Spine 平面（Spine plane）构成了一个模块化的网络拓扑，可以容纳几十万台服

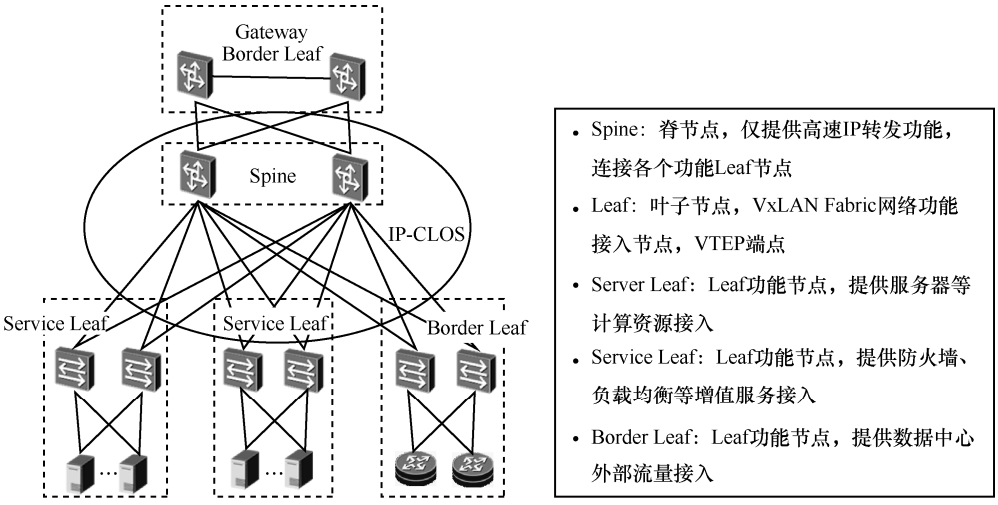
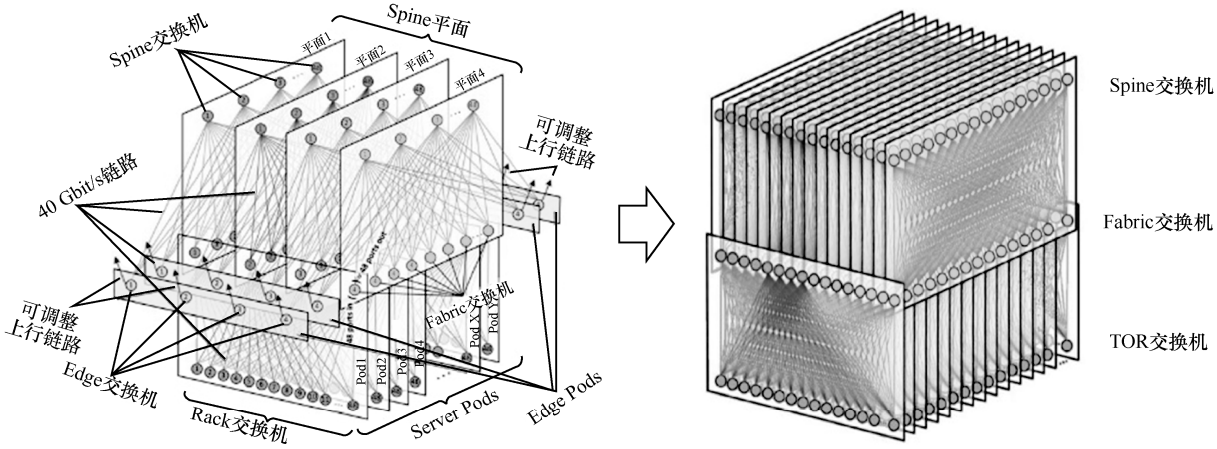


图 2 IP-CLOS 典型架构



（a）Facebook数据中心F4架构　　（b）Facebook数据中心F16架构

图 3　Facebook 大规模数据中心 F4 到 F16 架构演进

务器。这种架构设计的关键在于每台交换机有相同数量的 40 Gbit/s 下行链路和上行链路，能够实现真正的无阻塞。

2019 年，OCP 会议上 Facebook 发布 F16 架构，Fabric 交换机由原来的 4 平面变成了 16 平面。值得注意的是，Spine 平面的设计使用了 16 个 128 端口、100 Gbit/s 架构的交换机构建，而不是 4 个 128 端口、400 Gbit/s 交换机。这一选择使得网络更简单、扁平。每台 TOR 交换机达到 1.6 THz 的带宽，满足了架顶的带宽需求，同时大量减小了转发路径的跳数和芯片数量，大大节省了成本。

Facebook F4 到 F16 架构的演进，再一次验证了 IP-CLOS 网络在超大规模组网中的优势，在可扩展性、可靠性以及节省成本上都表现突出。

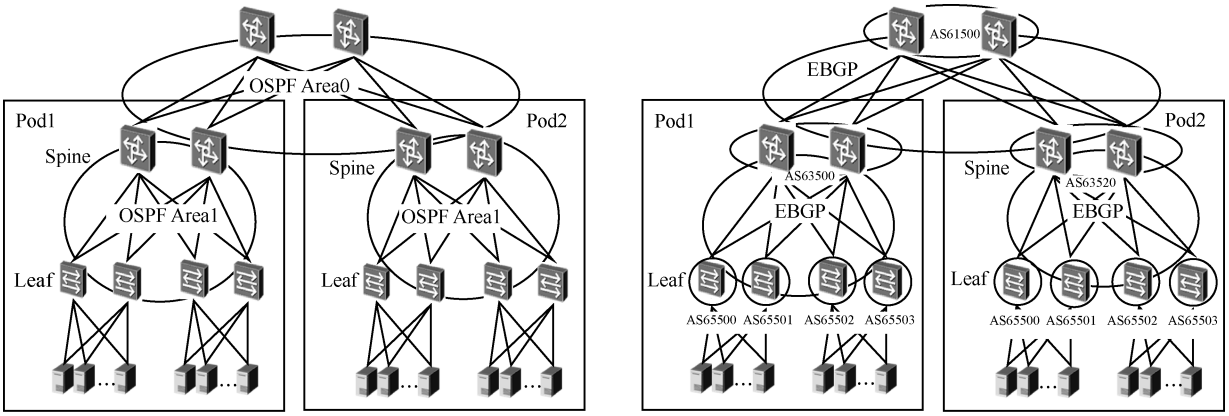
（2）基于 EBGP 的大规模路由组织

基于路由的三层寻址是实现网络可达的基础手段。目前数据中心内部网关协议（interior gateway protocol，IGP）路由主要以开放式最短路径优先（open shortest path first，OSPF）和中间系统到中间系统（intermediate system to intermediate system，ISIS）为主。OSPF 路由技术相对成熟，网络运维人员使用经验丰富。ISIS 相对于 OSPF，对大规模网络的支持能力和收敛性能更好。

RFC7938 提出将外部边界网关协议（external border gateway protocol，EBGP）应用于大规模数据中心的建议[4]，且目前在 Facebook、阿里巴巴等云数据中心内有广泛部署案例。有别于 OSPF、ISIS 等链路状态协议，BGP 是一种距离矢量路由协议，在路由控制、网络收敛时的网络稳定性更好。在中小型数据中心组网时，使用 BGP 和 ISIS、OSPF 协议性能相差不大。但是在超大规模数据中心组网中，BGP 的应用性能会更加优异。因为 OSPF、ISIS 等链路状态协议在网络域内任何节点发生故障时，会引起全网状态信息的泛洪和数据库信息更新，在此基础上收敛路由。而距离矢量路由协议只在节点间通告路由，通过增量刷新的方式更新路由信息。同时分区路由域独立，故障域可控，在超大规模组网中则能表现出更强的稳定性。EBGP 路由的规划和配置方式如图 4 所示。将 Pod 内的 Spine 设备规划为同一 AS 号，为每一组堆叠的 Leaf 规划一个单独 AS 号。Pod 内 Leaf 只和本 Pod 内 Spine 建立 EBGP 邻居，Leaf 间不建立 EBGP 邻居。规划配置相对于 OSPF 和 ISIS 会更复杂一些。

（3）“SDN+OpenStack”云网一体管控

数据中心内网络设备规模大，对自动化部署



(a) 数据中心组网OSPF组网规划　　(b) 数据中心组网EPGP组网规划

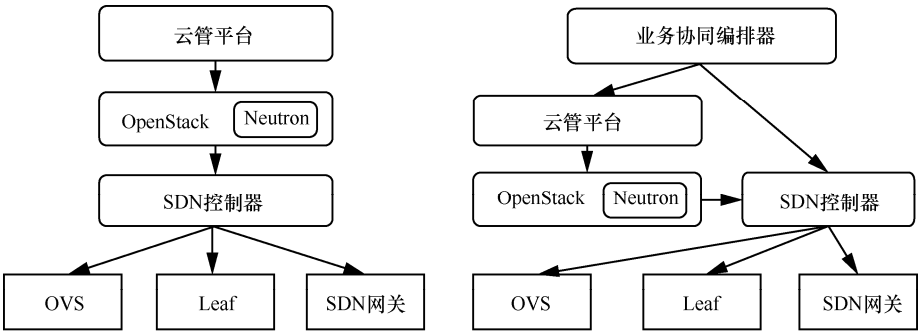
图 4　数据中心组网路由规划

的要求高。在云资源池内引入 SDN 技术，通过部署云管平台、SDN 控制器等，提供数据中心云网络自动开通、灵活部署、智能管控等能力。SDN 控制器可实现对 SDN TOR 交换机、SDN 网关（border Leaf）、vSwitch 等转发设备的统一管控，实现对网络资源的灵活调度。在部署方式上，SDN 控制器可以和 OpenStack Neutron 集成，通过云管平台统一管理云资源池内的计算、存储和网络资源，也可以通过业务协同编排器统一协调云管平台和 SDN 控制器，实现对数据中心内资源的统一编排和调度。基于 SDN 控制器的自动化管控方案如图 5 所示。

4.2　超低时延无损网络组网

数据中心网络的低时延，离不开 RDMA 技术的部署和应用。高性能计算、分布式存储以及 AI 应用系统等对低时延网络有强烈需求的分布式应用，服务器端已经多采用这种配置模型。因为传统的 TCP/IP 在收发报文时，要经过多次的内核处理，会导致多达数十微秒的固定时延。为将协议栈时延降至约 1 μs，同时减少 CPU 负担，远程内存直接访问（remote direct memory access，RDMA）成为主要部署技术。RDMA 可以从发送端地址空间中取出数据，直接传送到接收端的地址空间中，快速完成计算节点内存间数据的快速交换，而不需要内核内存参与，大大降低了服务器侧的处理时延。

同时，考虑到未来云网一体化下业务端到端需求，基于以太网的 RoCE（RDMA over converged ethernet）技术已经逐步替代无限带宽（infiniband，IB）等专用技术，成为主流技术。目前最新的



(a) 基于云管的自动化管控　　(b) 基于业务协同编排的统一管控

图 5　基于 SDN 控制器的自动化管控方案

RoCEv2 版本使用 IP/UDP 替代 IB 网络层，提供 IP 路由及 ECMP 能力，成为高性能数据中心网络主要部署协议[5]。

为了适配服务器传输 RoCE 流量的高效性，数据中心网络必须综合解决分组丢失、时延、吞吐等多方面的问题，构建零分组丢失、低时延、高吞吐的无损网络，主流技术包括流控制、拥塞控制和负载均衡等技术。

图 6 给出了典型的基于 ROCEv2 的高性能以太无损网络部署方案。

数据中心内超低时延无损网络的部署方案主要考虑以下几点。

（1）减小硬件转发时延

网络设备转发、转发跳数以及光纤距离都会影响网络时延。所以，在部署上一般以 Pod 为单位支持 RDMA，并使用 RoCEv2 协议规范。单 Pod 内使用两级 CLOS 结构减少转发层级。在硬件设备上，25 Gbit/s、100 Gbit/s 模型部署最为成熟。

（2）控制网络零分组丢失

当网络拥塞造成缓存溢出分组丢失时，会引发数据分组重传，从而进一步增加时延。一方面，可以通过增加交换机缓存、网络带宽来提高抗拥塞的能力；另一方面，可以进行应用层算法优化规避“多打一”流量，减少网络拥塞点，以及部署流控技术通知源端降速以消除拥塞等。

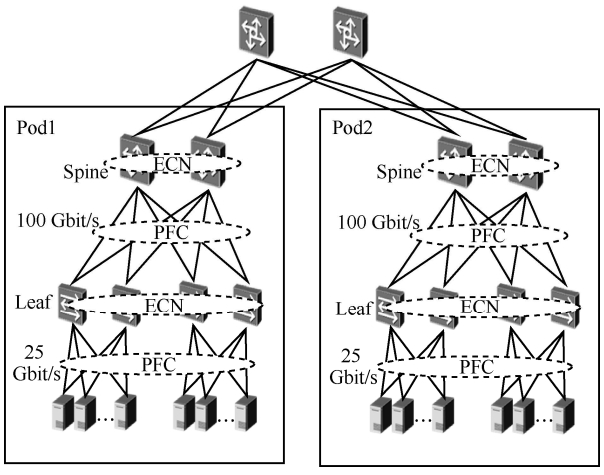


图 6　基于 RoCEv2 的高性能以太无损网络部署方案

目前，最基本的部署方式为 ECN 结合 PFC 来处理网络拥塞。

显示冲突通告（explicit congestion notification，ECN）是一种基于流的端到端流控技术，是在交换机出口发起的拥塞控制机制。通过 TCP 发送端和接收端以及中间路由器的配合，感知中间路径的拥塞，并主动减缓 TCP 的发送速率，避免拥塞而导致的分组丢失。

优先级的流量控制（priority-based flow control，PFC）是一种基于队列的反压技术，是在交换机入口发起的拥塞管理机制。在通常无拥塞情况下，交换机的入口缓冲器不需要存储数据。当交换机出口的缓冲器达到一定的阈值时，交换机的入口缓冲器开始积累，当入口缓冲器达到设定的阈值时，交换机入口开始主动迫使它的上级端口降速。

之所以采用 PFC 结合 ECN 的机制来实现网络拥塞控制，主要是因为 PFC 存在死锁等潜在问题，常用解决方法会优先触发 ECN 报文，用来减少网络中 PFC 的数量，在 PFC 生效前完成流量的降速。

现有方案虽然得到一定的应用部署，但是仍存在一定局限性。各参数配置基本为静态手工完成，且每次最优参数的调整都需要有经验的工程师持续遍历尝试。为此，产业界也提出一些无损网络增强特性，包括虚拟输入队列（virtual input queuing，VIQ）、动态虚拟通道（dynamic virtual lane，DVL）等技术，且相关技术已经在 IEEE 802.1 工作组 TSN 任务组进行标准化工作[6]。同时，业内也在积极探索基于专用芯片的动态拥塞调度机制，通过 AI 能力实现无损网络自调参、自优化，尝试解决手工配置静态参数导致网络无法动态适应负载流量模型变化的问题，实现智能无损网络。

4.3　云网业务统一承载

云网一体化架构下，数据中心内网络和外部网络的边界逐渐模糊，云内网络、云间网络、用户到

云网络构建成一个整体，通过资源端到端管控、业务端到端发放来提供服务的一体化。尤其面向云网融合业务，端到端统一承载技术显得尤为重要。

VxLAN 是目前云数据中心典型的网络虚拟化技术。本质上是一种大二层技术，通过 IP 网络进行二层隧道流量的传输。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，极大地提升了数据中心内业务承载的灵活性。

基于 IPv6 的分段路由（segment routing over IPv6，SRv6）是目前承载网关注度和讨论度极高的研究热点技术，是基于源路由理念而设计的在网络上转发 IPv6 数据分组的一种协议，具备可编程、易部署、易维护、协议简化的特点。它通过集中控制面可实现按需路径规划与调度，同时 SRv6 可以完全复用现有 IPv6 数据平面，满足网络灵活演进要求。

- 简化网络配置。SRv6 不使用 MPLS 技术，完全兼容现有 IPv6 网。
- 网络可编程：SRv6 报文中路由扩展头（segment routing header，SRH）编程能力强，既可以做路径控制，也可以做业务扩展[7]。
- 业务端到端建立简单：只要物理网络 IPv6 路由可达，就可实现网络端到端的无缝互通。在跨域网络扩展能力上表现出极强的灵活性。

以太虚拟私有网络（ethernet virtual private network，EVPN）定义了一套通用的控制层面协议，可被用来传递 MAC/ARP/主机路由表、网段路由等信息。并且支持 VxLAN、SR 等多数据平面的转发。同时，EVPN 实现控制平面和转发平面分离，设计思路和 SDN 相似，因此经常用在 SDN 架构的部署方案中。

云网一体化下网络协议发展方向如图7所示。

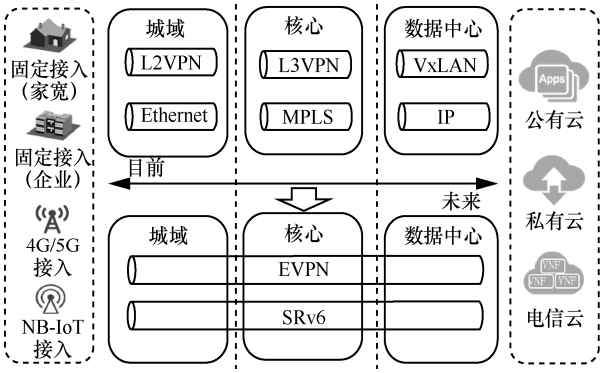


图 7　云网一体化下网络协议发展方向

采用 SRv6/EVPN 可有效统一“云内网络、云间网络、用户到云网络”承载协议，提供“固移融合、云网融合、虚实网元共存”的云网一体化网络的业务综合承载方案。SRv6 既符合国家的 IPv6 战略，又符合未来技术演进方向，标准化后有可能取代现有 VxLAN 等技术，成为承载层的统一隧道协议，应用领域从骨干网、城域网向数据中心网络逐步扩展。EVPN 基础标准已经完备，应用领域已经从数据中心走向广域网。SRv6/EVPN 能提供云网一体化环境下的 L2/L3 业务高效承载[8]。

5　面向云网融合的大规模数据中心组网方案

图 8 给出了一个面向云网融合的大规模云数据中心的网络架构设计，满足大规模组网的高可扩展、高效灵活的云网业务承载需求。通常情况下，一个 region 代表提供云服务的一个区域。一个 region 内至少包含 2 个或者 3 个可用区（available zone，AZ），用于搭建高可用组网架构。

5.1　物理网络组网

单可用区的物理组网如图 8 所示，整个组网设计遵照超大规模组网原则和技术，按照业务功能分 Pod 区规划，各区域可独立扩展。

核心层：设置一对核心交换机设备（Super-Spine），提供流量高速转发，与各 Pod 区的 Spine 交换机交叉互联。核心交换机采用去堆叠配置，可通过设备替换或者横向增加设备的方

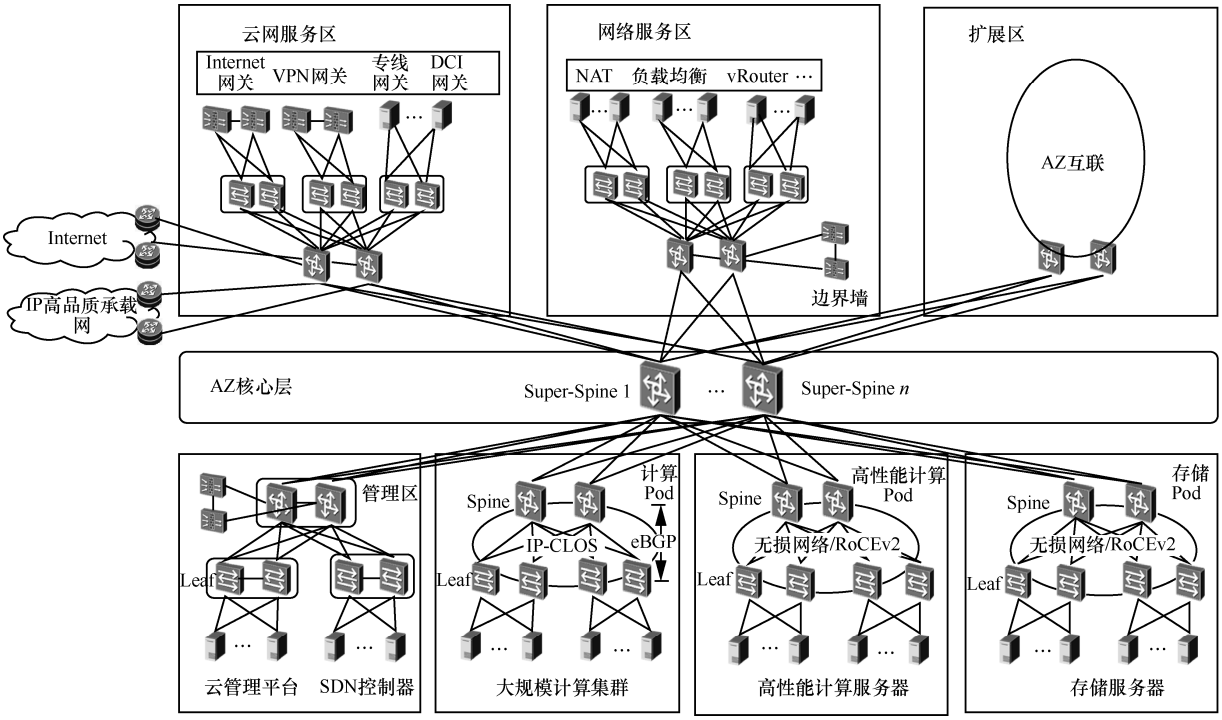


图 8　单可用区的物理组网

式升级到更高带宽和更大接入规模，支持业务进一步扩展。

计算 Pod：承载大规模计算集群来提供租户虚拟机资源，单 Pod 内采用基于 Spine-Leaf 的典型 IP-CLOS 结构，Spine 和 Leaf 节点设备均采用“去堆叠+eBGP”的组网方式，保障单 Pod 模块的高可扩展性。

高性能计算 Pod/存储 Pod：为了增强数据中心网络承载 AI、高性能计算、分布式存储等业务能力，基于 eBGP 的 IP-CLOS 组网，在高性能计算 Pod、存储 Pod 中部署 RoCEv2 的无损网络技术，保障单 Pod 内 RDMA 流量的超低时延特性。经实测验证，无损网络部署可提升分布式存储 IOPS 20%，单卷性能达到 35 万 IOPS，平均时延降低 12%，并且严格保障了读取数据汇聚时“多打一”流量的零分组丢失传输。

管理区：放置内部管理组件，不对外提供服务，为安全可靠区。主要承载云管理平台、SDN 控制器、OpenStack、级联节点等。在 Spine 设备旁挂内网管理防火墙设备，保证管理网安全。

扩展区：用于和多个可用区之间的互联。可用区之间利用低延迟光纤传输网络互联，要求时延小于 1 ms，保障多活业务的可靠性。

网络服务区：承载云内网络的服务，包括 vRouter 等。所有租户的业务流量直接走到这个分区，在分区内匹配相关业务对应处理。

云网服务区：主要用于对外接入云网融合业务的区域，包括 Internet 网关、VPN 网关、专线网关、DCI 网关等。其中，Internet 网关承载 Internet 流量访问，VPN 网关用于 IPSec-VPN 接入。同时，分别设立云专线接入网关、DCI 接入网关用于入云专线访问 VPC、云间跨 region VPC 互联的连接，满足云网融合业务的各类连接需求。

5.2　逻辑网络组网

数据中心云内网络业务承载方案如图 9 所示。业务承载依赖大二层网络，目前通用的做法是采用基于 MP-BGP 的 EVPN 承载的 VxLAN。硬件 VTEP 节点包括 Internet 网关、VPN 网关、专线接

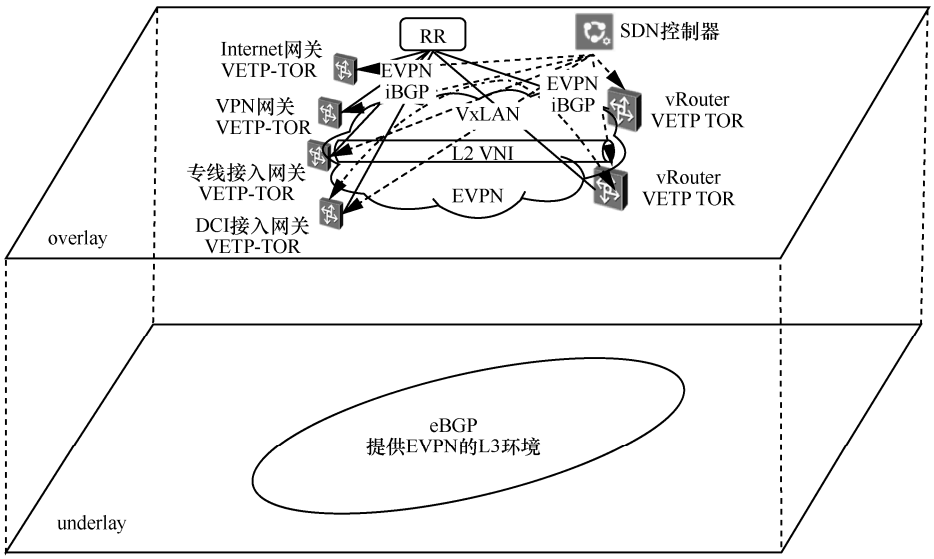


图 9　云内网络业务承载方案

入网关、DCI 接入网关的 VTEP TOR、网络服务区 TOR 等。各 VTEP 节点通过网络设备间的 eBGP 发布并学习 EVPN VxLAN 所需的 loopback IP。VTEP 使用 BGP 多实例功能组建 overlay 网络，管理服务区汇聚作为 EVPN BGP RR，与所有 VTEP 节点建立 iBGP 邻居。VTEP 节点创建二层 BD 域，不同的 VTEP 节点属于相同 VNI 的 BD 域，自动创建 VxLAN 隧道，实现业务流量转发。

以入云专线承载为例，客户使用云专线产品接入云内 VPC 网络时，流量从专线接入网关进入，通过 VTEP TOR 走 VxLAN 到网络服务区 TOR，进入 vRouter。vRouter 封装成 VxLAN 后，将报文路由到 Pod 内，通过多段 VxLAN 拼接和计算节点的虚拟交换机建立连接，VxLAN 报文在虚拟交换机上解除封装进入 VPC。

VxLAN/EVPN 技术是目前大规模云数据中心网络通用且高效的业务承载方案，能够实现云内业务快速发送和自动化配置。后续随着 SRv6 技术标准的成熟，SRv6/EVPN 的统一承载方案会逐渐向数据中心内网络演进。目前，Linux 已经支持大部分 SRv6 功能，Linux SRv6 提供一种整合 overlay 和 underlay 的承载方案，保证 underlay 网络和主机叠加网络（host overlay）SLA 的一致性。在数据中心中引入 SRv6 承载，还需进行大量的研究和实践。

6　结束语

运营商正在积极布局以数据中心为核心的云网一体化基础设施，加快云网应用创新孵化。超大规模组网可扩展、超低时延无损网络、高效灵活的云网承载成为了数据中心网络的新特征。当前很多关键技术仍需不断探索实践，如何构建高性能数据中心网络，统一融合承载多类型资源池、多业务场景，会是后续重点研究内容。

参考文献：

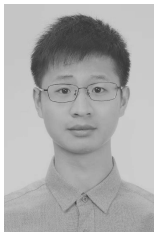
[1] 开放数据中心标准推进委员会. 无损网络技术及应用白皮书[R]. 2018.
ODCC. Lossless network technology and application white paper[R]. 2018.

[2] 雷波. 边缘计算中的网络建设需求分析[J]. 通信世界, 2019(23): 47-48.
LEI B. Analysis of network construction requirements in edge computing[J]. Communications World, 2019(23): 47-48.

[3] Al M, FARE S, LOUKISSAS A, et al. A scalable, commodity

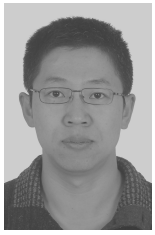
data center network architecture[J]. Computer Communication Review, 2008, 38(4): 63-74.
[4] PREMJI A, LAPUKHOV P, MITCHELL J. Use of BGP for routing in large-scale data centers[J]. Computer Science, 2016.
[5] ZHU Y, ZHANG M, ERAN H, et al. Congestion control for large-scale RDMA deployments[J]. ACM SIGCOMM Computer Communication Review, 2015, 45(5): 523-536.
[6] IEEE. The lossless network for data centers[R]. 2018.
[7] IETF. IPv6 segment routing header (SRH): RFC 8754[S]. 2020.
[8] 陈运清, 雷波, 解云鹏. 面向云网一体的新型城域网演进探讨[J]. 中兴通讯技术, 2019, 25(2): 2-8.
CHEN Y Q, LEI B, XIE Y P. Evolution of new metropolitan area network for cloud network convergence[J]. ZTE Technology Journal, 2019, 25(2): 2-8.

[作者简介]



王江龙（1993− ），男，中国电信股份有限公司研究院工程师，CCSA“网络 5.0 技术标准推进委员会”接口与协议组副组长，主要研究方向为未来网络架构、新型 IP 网络技术等。

雷波（1980− ），男，中国电信股份有限公司研究院高级工程师，边缘计算产业联盟 ECNI 工作组联席主席，CCSA“网络 5.0 技术标准推进委员会”管理与运营组组长，主要研究方向为未来网络架构、新型 IP 网络技术等。



解云鹏（1975— ），男，中国电信股份有限公司研究院高级工程师，CCSA“网络 5.0 技术标准推进委员会”架构组副组长，主要研究方向为未来网络架构、新型 IP 网络技术等。

何琪（1968− ），女，中国电信股份有限公司研究院高级工程师，CCSA“网络 5.0 技术标准推进委员会”需求组副组长，主要研究方向为未来网络架构、新型 IP 网络技术等。

李云鹤（1981− ），男，中国电信股份有限公司研究院工程师，主要研究方向为新型 IP 网络技术及设备等。