数据中心网络架构和底层协议演进

Data Center Infrastructure and Underlay Protocol Evolution

魏月华 /WEI Yuehua
陈晓 /CHEN Xiao
张征 /ZHANG Zheng

（中兴通讯股份有限公司，中国 深圳 518057）
(ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.202103011
网络出版地址：https://kns.cnki.net/kcms/detail/34.1228.tn.20210617.0922.002.html

网络出版日期：2021-06-17
收稿日期：2021-05-10

摘要：受计算规模的驱动，数据中心物理拓扑从接入 – 汇聚 – 核心三级网络架构演进到基于 Clos 的 Spine–and–Leaf 架构。计算资源的基本单位经历了物理服务器、虚拟机、容器化 3 个阶段。数据中心底层（underlay）连接协议逐步从以二层协议为主演进到以 IP 路由协议为主。但传统路由协议存在可扩展性、拓扑可见性、自动化部署能力等诸多问题。结合链路状态和距离矢量的胖树路由协议，解决了超大规模数据中心部署的痛点问题，有望逐渐成为超大规模数据中心底层网络的主流技术。

关键词：Spine–and–Leaf；路由；数据中心

Abstract: Driven by the scale of computing, the physical topology of the data center has evolved from an access–aggregation–core three–level network architecture to a Clos–based Spine–and–Leaf architecture. The basic unit of computing resources has gone through three stages: physical server, virtual machine, and containerization. The underlay connection protocol of the data center has gradually evolved from layer 2 protocol to IP routing protocol. However, traditional routing protocols have many problems, such as scalability, topology visibility, and automated provision capabilities. The fat–tree routing protocol, which combines link state and distance vector, solves the pain points of ultra–large–scale data center deployment, and is expected to gradually become the mainstream technology for ultra–large–scale data center underlay networks.

Keywords: Spine–and–Leaf; routing; data center

1 接入 – 汇聚 – 核心三级网络架构协议方案演进

受计算规模的驱动，数据中心的网络架构和解决方案，在过去 20 年里发生了很大变化。总的来说，数据中心物理拓扑从接入 – 汇聚 – 核心三级网络架构演进到基于 Clos 的 Spine–and–Leaf 架构。计算资源的基本单位经历了从物理服务器到虚拟机再到容器化 3 个阶段。

在物理服务器阶段，应用直接在物理服务器上运行，数据中心物理拓扑为经典的接入 – 汇聚 – 核心三级网络架构，整张网络采用二层协议互联，应用访问模式为客户端–服务器模式，并且南北向流量远大于东西向流量。其中，南北向流量在核心交换机处理，数据中心内跨网段需要经过核心交换机，内部子网的网关一般也配置在核心。在这种模型中，由于节点之间的通信都可能经过核心，因此核心交换机需要记录所有节点的互联网协议（IP）和介质访问控制（MAC）地址信息。在这种网络方案中，与计算节点规模相关的瓶颈最可能出现在核心交换机中。

2008 年，传统的数据中心逐步演进到云计算时代的数据中心。云计算时代计算资源的基本单位从物理机变成了虚拟机。计算资源的数量和密度都有数量级的提高。应用广泛采用微服务访问模式。这种模式带来的网络变化是：东西向流量超过南北向流量，成为数据中心的主要流量。

随后，网络虚拟化应运而生。数据中心网络中的每个宿主机都运行一个虚拟交换机（vSwitch）。虚拟交换机向上连接物理交换机，向下连接多个虚拟机。网络的边界从原来的接入交换机（置顶交换机）层，下沉到宿主机内部。这使得整张网络变成一个大的二层网络。在这个大二层网络内，虚拟机生命周期内的 IP 地址和 MAC 地址均保持不变。对于同网段的虚拟机，不管它们是否在同一台宿主机上，彼此都能够通过二层（MAC 地址）访问对方。此时，核心交换机不仅需要

记录宿主机的 IP/MAC 信息，还需要记录所有虚拟机的 IP/MAC 信息，以便支持虚拟机全网可迁移。

2016 年以后，数据中心进入大规模容器时代。容器也被称为轻量级虚拟机，可进一步提高部署密度。虚拟机与容器的最大区别在于：虚拟机平台交付的是虚拟机实例，抽象的是计算资源，而容器平台交付的是服务，访问入口为服务的 IP 地址，同时服务屏蔽了计算资源的细节（如虚拟机实例的 IP 地址或 MAC 地址）。

当把虚拟机换成容器后，考虑到容器的部署密度，如果继续采用大二层模型，交换机转发表容量将会成为网络瓶颈。为此，在每个服务器节点内可用虚拟路由器（vRouter）替换虚拟交换机。一个虚拟路由器管理一个网段。服务器域内是一个二层网络。服务器节点运行边界网关协议（BGP）代理，并负责节点之间或者节点和数据中心网络之间的路由同步。核心交换机只需要记录服务器节点本身的 IP 和它所管理的网段。表项与服务器的数量保持同一量级，但与容器的数量没有关系。

因此，数据中心网络拥有一个在三层网络下有无数个小二层网络的架构，如图 1 所示。这种以三层路由为主的数据中心协议架构，可以满足现代数据中心规模不断扩大和服务器数量不断增加的需求。

2 带宽与流量模型的变化

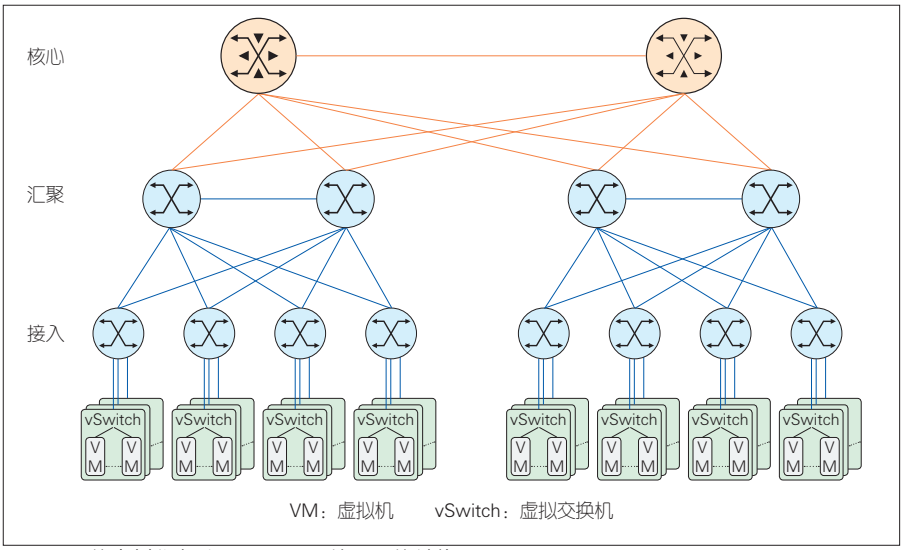
传统数据中心的流量主要是进出数据中心的流量，通常被称为南北向流量。即使在网络层之间存在很高的收敛比，传统的“树”拓扑也足以容纳这样的流量。如果需要更多的带宽，则可以通过“扩展”网络元素来增加带宽。例如，升级设备的线路板，或者采用端口密度更高的设备。

如今，许多大型数据中心承载着大量服务器到服务器的流量。这些流量并不会离开数据中心，通常被称为东西向流量。例如，某些应用程序需要集群之间的海量数据进行复制，或者需要虚拟机进行迁移。由于受到物理限制（例如交换机的端口密度低），采用扩展传统的树形拓扑来满足带宽需求的方式，不仅成本很高，而且难以实现。

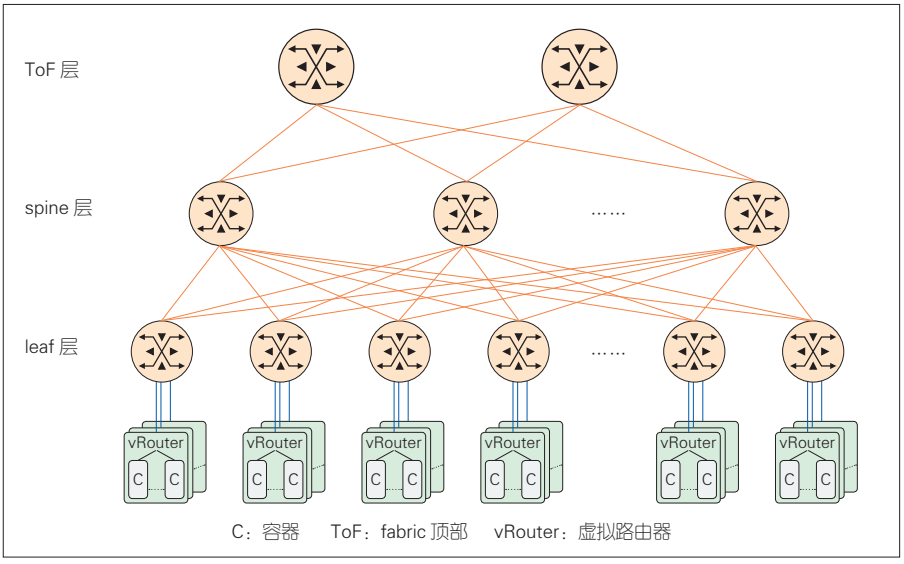
3 基于 Clos 的 Spine–and–Leaf 结构演进

东西向流量的增加使三层数据中心架构中的带宽成为瓶颈。此外，服务器到服务器的延迟会随着流量路径的不同而不同。为了解决这两个问题，基于 Clos 网络的 Spine–and–Leaf 架构被提出。

在如图 2 所示的三级 Clos 架构中，每个低层级的 leaf 交换机都与所有高层级的 spine 交换机相连，并形成全网状连接拓扑。leaf 交换机用于连接服务器等设备，spine 层则负责将所有的 leaf 连接起来。当 leaf 层的接入端口和上行链路都没有瓶颈时，这个架构就实现了无阻塞连接。



▲图 1 网络虚拟化与接入 – 汇聚 – 核心网络结构



▲图 2 典型的 Spine–and–Leaf 拓扑

在 Spine–and–Leaf 架构中，任意一个服务器到另一个服务器的连接，都需要相同数量的设备（除非这两个服务器都在同一个 leaf 下）。这使得延迟可以被预测。由于东西向带宽更高，因此它更适合现代微服务的场景。

当 Spine–and–Leaf 中任意一层存在带宽瓶颈时，只需要添加一台新设备，并将其和另外一层的所有设备相连即可。这种横向扩展的方法比较容易实施。

4 数据中心协议的选择与设计

4.1 选择三层路由的 Spine–and–Leaf 架构

Spine–and–Leaf 结构相当于传统网络架构中的“接入层 – 汇聚层”。如果采用二层交换技术，则生成树协议（STP）生成的无环树形结构会大大减少活跃可用的链路。

如果采用三层路由，Spine–and–Leaf 则可以充分利用 spine 和 leaf 之间的全网状连接，并选择最短路径。如果为了获得更高的整体利用率，该架构也可以选择特定的路径。

4.2 BGP 路由协议部署技术与特征 [1–2]

BGP 在应用于数据中心之前，主要用于运营商网络。BGP 数据中心与运营商网络最大的区别在于连接的密度：超大型数据中心的连接密度远大于运营商网络的连接密度。因此，BGP 协议在应用于数据中心之前需要经过适当的“改造”。

BGP 协议具有一些突出优势，主要包括：

（1）作为距离矢量协议，BGP 采用传输控制协议（TCP），互操作性好，总体上很成熟，目前已经获得广泛应用。设备商和各种开源平台都实现了 BGP 部署，并获得了良好的测试结果。

（2）由于 BGP 本身在广域通信网络上是一个广泛部署的路由协议，因此，从技术和运维的角度上看，将 BGP 应用于超大规模数据中心网络具有很高的接受度；

（3）相比于其他内部网关路由协议，BGP 具有较高的可扩展性；

（4）BGP 协议有诸多前缀过滤、路由标记和流量工程的能力选项，在过滤、修改路由参数和控制流量方面具有优势；

（5）BGP 可以同时用于底层（underlay）网络和叠加（overlay）网络。通常在这种情况下，底层网络使用外部 BGP（eBGP）对等体，叠加网络使用内部 BGP（iBGP）对等体。这使得网络的整体配置变得更简单。

BGP 协议作为数据中心的底层也面临一些挑战，具体包括：

（1）由于 BGP 协议具有易于扩展的特性，BGP 上逐步增加的多地址族、以太网虚拟专用网（EVPN）、虚拟专用局域网业务（VPLS）、BGP 链路状态（BGP–LS）等能力，使得 BGP 协议变得非常复杂。虽然可以通过一些开关来关闭这些功能，但是实际上仍无法避免实现 BGP 功能的软件代码漏洞和错误配置等问题；

（2）BGP 协议在自动化能力方面不足以满足大规模数据中心的需求；

（3）在数据中心 fabric 中的高密度拓扑中，需要大量专业的手动配置来使 BGP 快速收敛。例如，当流量从 fabric 上的一个位置移动到另一位置，或者当由 anycast 地址代表的一个服务实例从 fabric 上被删除时，BGP 收敛时间会很长。这将影响在 fabric 上正常运行的应用。

4.3 链路状态路由协议的演进 [3]

自 RFC 7938（在大规模数据中心路由中使用 BGP 的标准）发布起，BGP 几乎成了大规模数据中心的缺省选择。考虑到标准和部署的多种因素（如收敛速度、数据遥测等），业界提出在数据中心 fabric 中采用链路状态路由协议来代替 BGP 协议。

在超大规模数据中心采用链路状态路由协议的最大的挑战是，存在用于可达性计算和拓扑计算的路由信息洪泛问题。目前，国际互联网工程任务组（IETF）正在针对中间系统到中间系统（IS–IS）开展洪泛优化和集中计算优化泛洪树的工作。

在数据中心 fabric 中，与 BGP 协议相比，链路状态协议具有收敛速度快的优点。当一个可达目的地在 fabric 中从一个地方移动到另一个地方，或者完全从 fabric 上断开时，链路状态协议的收敛速度将远快于 BGP 的收敛速度。从 IS–IS 的角度来看，任何可达目标的更改都只是叶子连接的更改。这意味着系统无须运行最短路径优先（SPF）算法。这种方法被称为部分 SPF。它的速度非常快，并且每个交换矩阵设备只需要进行最少量的处理。

与数据中心结构中的 BGP 相比，链路状态协议的第二个优势是拓扑可见性。链路状态协议要求每个设备都拥有维护拓扑的完整视图。该拓扑（称为链接状态数据库）必须与网络洪泛域中的每个路由器同步。在使用控制器时，为了获得链路状态数据库的副本，链路状态协议仅需要连接光纤网络中的一个路由器。链接状态数据库对于流量工程和流量导流很有用，也有利于做数据遥测。

数据中心结构中链路状态协议面临的第一个挑战是扩展问题，这主要与消息洪泛有关。由于消息量大，链路状态协议会在大型结构中造成严重的洪泛。

此外，链路状态协议还面临另外

两个挑战：存在可达目的地数量的扩展性问题和计算无环路径集 SPF 算法所需的时间较长的问题。通过更快的处理器和 SPF 优化，虽然不能使链路状态协议的扩展性达到 BGP 的级别，但是足以支持运营商构建大部分的数据中心结构。

4.4 胖树路由协议特征分析 [4–6]

业界对数据中心 fabric 中路由技术的探索从未停止。针对基于 Clos 网络的 Spine–and–Leaf 结构，IETF 启动了结合距离矢量路由与链路状态路由的胖树路由协议的标准化工作。

胖树路由协议可将链路状态协议和距离矢量协议的优点结合起来，以最大程度地实现网络路由配置自动化和故障管理自动化，并用于 Spine–and–Leaf 结构的大规模数据中心中。胖树路由协议支持多线程，可匹配多核 CPU 的处理能力。因此，胖树路由协议可以极大地节省操作和运维成本，并减少人为错误。

4.4.1 拓扑适用性分析

如前所述，在数据中心进入云计算时代以后，东西向流量就超过了南北向流量，成为数据中心的主要流量。东西向流量在虚拟服务器与虚拟服务器之间，以及容器与容器之间的转发，本质上还是在胖树的北向与南向运动。只不过东西向流量的转发是最大程度的就近转发。

流量从 Spine–and–Leaf 结构底部的 leaf 节点向北到达结构的顶部，然后向南回到 leaf 节点。从所需的可达性信息角度来看，这种服务器到服务器的流量模式，所需的可达信息很少。例如，在三级 Clos 中，leaf 节点流量仅需要默认路由即可到达 spine 节点。同时 spine 节点流量不需要整个路由表即可到达 leaf 节点，只需要向南一级的节点可达信息。因此，胖树路由协议具有方向特性，具体表现为：向北为链路状态协议，向南则为距离矢量协议。

胖树结构（Spine–and–Leaf 结构）天然分层：结构顶部的节点保持在最高级别，而底部节点（leaf 节点）保持在最低级别。胖树路由协议用方向性来描述拓扑中不同级别之间的关系，并利用拓扑的这种特性，通过零接触部署（ZTP）功能进行错误布线检测。另外，这种协议在设计时也考虑了容错性，因此能够应对胖树结构的变异，比如同一层节点之间的水平链路或跨层的垂直直连链路。

4.4.2 拓扑发现

胖树路由协议通过交换链路信元（LIE）自动发现邻居，协商 ZTP，并检测错误布线。LIE 交换采用用户数据报协议（UDP），并且将互联网协议第 4 版（IPv4）报文中的生存时间值（TTL）（或互联网协议第 6 版报文中的 Hoplimit）设置为 1。LIE 包含的关键信息有本地链路 ID、SystemID、最大传输单元（MTU）、本地节点的交付点（PoD）值、所属层值等。

胖树路由协议通过交换拓扑信元来携带一个节点连接的邻居、前缀和能力等信息。由于胖树路由协议具有方向特性，拓扑信元可分为北拓扑信元和南拓扑信元。

无论是南拓扑信元还是北拓扑信元，拓扑信元都包括 6 种类别：节点拓扑信元、前缀拓扑信元、积极解聚合拓扑信元、消极解聚合拓扑信元、外部前缀拓扑信元和键值拓扑信元。

拓扑信元交换（洪泛）采用 UDP 协议，具有方向性。所有的北拓扑信元都是向北洪泛的，目的在于为更高层提供以南网络的完整拓扑视图。这可以保证从特定层节点（或低于特定层节点）收到的流量始终采用最具体的路由来到达目的节点。

所有南节点拓扑信元都被往南泛洪，而其他类型的南拓扑信元仅往南泛洪本节点为发起者的拓扑信元。这样，低一级的节点就会拥有去往上层节点所需要的路由信息。这些信息也可以到达 fabric 的其他地方。

胖树路由协议采用类似 IS–IS 协议的方式来保持链路状态数据库的同步。在计算最短路径时，胖树路由协议也是基于南向或北向的。两个方向的最短路径算法都不会产生环路：往北向的最短路径算法只利用北向（和东西向）邻居来计算“北拓扑信元”，往南向的最短路径算法只利用南向邻居来计算“南拓扑信元”

4.4.3 负载均衡

IP 网络中的负载均衡一直是个难题。BGP 负载均衡实施困难，而内部网关协议（IGP）仅能做到等价路径负载均衡。在胖树路由协议中，负载均衡只需要在北向的缺省路由上来实现（也可以在解聚合前缀和南向路由上实现）。胖树路由协议自动计算并继续使用所有可用最短路径上的可用带宽，使流量不会在 fabric 中迂回打转。

在正常情况下，每个前缀都带有一个关联的距离值（相当于典型的度量值）。当链路发生故障时，SPF 计算必须考虑当前不可用的带宽，并计算带宽调整后的距离（BAD），然后使用 BAD 值来代替初始距离值，以评估可用链接上的流量。

4.4.4 南向反射与路由解聚合

这种反射机制是指，只有节点的南向拓扑信元会被往北反射到上一层。因此，同一层的所有节点都能够相互感知对方。

反射机制可以触发积极解聚合。

为了解决流量黑洞问题，路由解聚合在发布缺省路由的基础上，会再发布一个更详细的路由。

解聚合包括两种类型：积极的解聚合和消极的解聚合。节点发布积极路由表示它可以到达某个前缀。而当节点不能到达某个前缀时，则通告消极路由。不管是哪种情况，解聚合的路由总是被通告为前缀或外部南拓扑信元，并且永远不会被重发。同时，其他节点不需要知道哪个节点正在发布解聚合的路由。

积极解聚合很简单。它是一种额外的路由通告。这样，南方的节点可以根据典型的最长匹配原则来进行路由布置，即胖树路由在默认路由中为部分连接的前缀打一个洞。

积极解聚合是非传递性的，以免给节点增加无用的路由信息。对于未解聚合的前缀，默认路由将为其提供可达性。

消极解聚合相对比较复杂。当fabric包含多个平面时，消极解聚合就是必需的。当某个节点失去某前缀的可达性时，该平面中所有上一层的节点都会触发消极解聚合。与积极路由不同，消极路由是可传递的。消极路由可以一直向南广播，直到解除流量黑洞。

4.4.5 零接触部署

胖树路由协议内置了零接触部署模式。除了ToF节点之外（ToF节点需要预先设定一个层值），其他节点无需任何初始化配置就可以自动接入fabric中。每个节点都以竞争在fabric中的最高点为原则。层决策算法利用相邻节点之间的位置信息进行运算，以确保所有节点找到在fabric中的稳定位置，从而自动完成一个稳定的胖树拓扑构建，并自动实现南向和北向路由策略。零接触部署能力能够有效消除可能的由错误布线对fabric构建产生的干扰。

零接触部署是胖树路由协议最突出的特性之一，对于提升超大规模数据中心网络构建的效率意义重大。

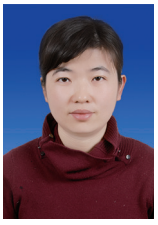
5 结束语

在未来，BGP将继续成为数据中心架构底层的重要选择。它最终会具备一些链路状态协议功能，例如更快的收敛和更接近自动化的部署。然而，BGP很难复制链路状态协议的某些功能，例如从一个位置获取整个拓扑的完整视图。同时，BGP的收敛速度很可能总是落后于链路状态协议。对此，IETF已经启动改进链路状态协议的标准化工作。但由于改动较大，同时协议复杂度较高，因此协议应用前景不明。胖树路由协议可将链路状态和距离矢量相结合：当数据报文沿fabric向上传递到ToF时，可采用类似链路状态的操作；当数据报文向fabric的边缘传递可达性和拓扑信息时，可采用类似距离矢量的操作。胖树路由协议解决了现有路由协议在Spine-and-Leaf IP结构中面临的诸多问题，具有扩展性好、运维简单的优点，可有效节省部署开销。

中兴通讯在IETF深入参与了胖树路由协议的标准化工作。我们认为，胖树路由协议有望成为超大规模数据中心底层网络的主流技术。

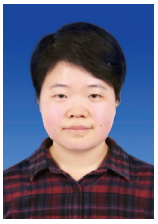
参考文献
[1] IETF. Use of BGP for routing in large-scale data centers: RFC 7938 [S]. 2016
[2] Dinesh G D. BGP in the data center [M]. California: O’Reilly Media, Inc. 2017
[3] IETF. Dynamic flooding on dense graphs: draft-ietf-lsr-dynamic-flooding-08 [S]. 2020
[4] IETF. RIFT: routing in fat trees: draft-ietf-rift-rift-12 [S]. 2021
[5] IETF. RIFT applicability: draft-ietf-rift-applicability-06 [S]. 2021
[6] IETF. A YANG data model for Routing in Fat Trees (RIFT): draft-ietf-rtgwg-policy-model-27 [S]. 2021

作 者 简 介



魏月华，中兴通讯股份有限公司承载网标准预研总工；拥有15年以上数据网络产品研发、设计及新技术预研经验；从事以太网、IP路由、云计算数据中心网络、SDN等技术和标准研究；发表论文3篇，获授权专利40余项。

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。



张征，中兴通讯股份有限公司标准专家；拥有20年的数据网络产品研发与设计经验；从事IP单播/组播路由、数据中心网络、SDN等技术与标准研究；主持多个IETF工作组标准的制定和RFC的发布；申请发明专利40余项。