基于叶脊架构的云资源池构建

Cloud Resource Pool Construction Based on Spine-leaf Architecture

文 娅[1]，戚金园[2]（1. 中国人民解放军31401部队；2. 上海邮电设计咨询研究院有限公司，上海 200092）
Wen Ya[1]，Qi Jinyuan[2]（1. Chinese People's Liberation Army 31401 Unit；2. Shanghai Posts & Telecommunications Designing Consulting Institute Co.，Ltd.，Shanghai 200092，China）

摘 要：
随着云计算的迅猛发展，云资源池主要流量方式发生变化，传统的三层网络架构的局限性日益凸显。具有扁平化设计的叶脊（Spine-leaf）二层架构更能适应大量东西向流量的需求，并具有高可靠、低时延、无阻塞等优点。以云资源池的构建为研究对象，对传统的三层网络架构与新型的叶脊架构进行对比分析。并提出了使用叶脊架构构建云资源池的具体实施方案。

关键词：
云资源池；三层网络架构；叶脊架构；Overlay
doi：10.12045/j.issn.1007-3043.2024.06.017
文章编号：1007-3043（2024）06-0084-06
中图分类号：TN919
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
With the rapid development of cloud computing, the main traffic mode of cloud resource pools has changed, and the limitations of three-tier network architecture are becoming increasingly prominent. In contrast, the two-tier architecture of Spine-Leaf with flat design is more adaptable to a large number of east-west traffic, and has the advantages of high reliability, low latency, non blocking, etc. It focuses on the construction of cloud resource pool and analyzes the traditional three-tier network architecture and the new Spine-Leaf architecture. Finally, a specific solution for constructing cloud resourcing pool using Spine-Leaf architecture is proposed.

Keywords：
Cloud resource pool；Three-tier network architecture；Spine-leaf architecture；Overlay

引用格式：文娅，戚金园. 基于叶脊架构的云资源池构建[J]. 邮电设计技术，2024（6）：84-89.

0 引言

云资源池统筹云数据中心各类基础设施，提升了计算机资源的利用率，具有资源利用率高、容灾能力强、扩展性好，按需租用、安全稳定等特点。用户可以按需租用计算资源、存储资源、网络资源等。客户能够通过自助服务方便地进行资源的租用、监控与管理。构建云资源池需要满足超大规模、高性能、安全稳定的要求。传统构建云资源池使用的是三层架构，但随着云资源池的演变，已经越来越不适合使用[1]。新型的叶脊架构采用扁平化二层设计，更能适应云资源池的构建。

收稿日期：2024-05-06

1 传统三层架构

传统的三层网络架构如图1所示，包含以下3层。

a）接入层。接入层主要面向终端用户，负责将工作站接入网络。接入层设备一般不需要较高的成本。

b）汇聚层。汇聚层是中间层，向下连接接入层，向上连接核心层。它先把接入层传输的数据进行汇聚、分发等处理，再传输到核心层，这样可以大大降低核心层设备的负担。此外，汇聚层还具有安全、地址过滤、网络分析等功能。汇聚层的设备需要具备较高

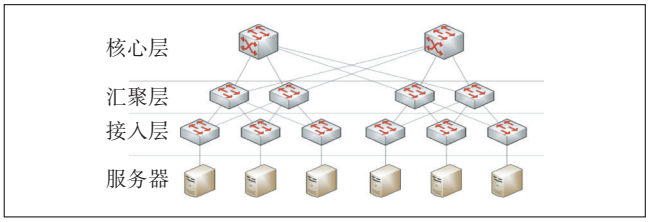

图1　传统三层网络架构

的性能。

c）核心层。核心层是网络的枢纽中心，起到至关重要的作用。核心层交换机主要负责高速转发骨干网络之间的流量，具有可靠性高、适应性强等优点。核心层设备具有强大的性能，同时设备价格也十分高昂[2]。

在构建云资源池时，传统的三层架构会有许多的不足之处。首先，它不能适应流量流向的变化，这也是其主要的问题[3]。传统三层架构在设计时主要考虑了南北向流量，即外部终端与数据中心的流量交互。然而，随着网络技术的发展，特别是虚拟化、微服务等技术的广泛应用，一次任务不再由某一个服务器完成，而是需要几个服务器通过分布式计算协同完成。于是云资源池数据中心出现越来越多的东西向流量（即平级设备之间的流量）。当东西向流量进行传输时，需要依次经过接入层、汇聚层、核心层、汇聚层、接入层，这给核心交换机和汇聚交换机带来了巨大的挑战，进一步造成设备成本高、效率低、通信时延长的问题[4]。其次，传统三层架构中，每个下层交换机会通过2条上行链路分别与2台上层交换机连接，但由于三层架构使用的是生成树协议，实际传输流量时只用到一条链路，而另一条链路作为备份。备用线路大部分时候处于闲置状态，从而造成宽带的浪费，降低了网络的使用率[5]。最后，由于不同服务器之间的通信路径是不确定的，产生的时延也是无法预测的，这对于云上大数据等业务来说是无法接受的。

2 叶脊架构

叶脊架构是一种大二层的扁平化架构，由叶交换机和脊交换机构成，每个叶交换机都与上行的脊交换机全部连接。同样地，每个脊交换机也与下行的叶交换机全部连接，是一种全网状连接方式[6]。与传统网络的三层架构不同，叶交换机之间或者脊交换机之间不需要同步数据，而传统三层网络中的汇聚层和核心层交换机需要同步数据。叶脊架构示意如图2所示。

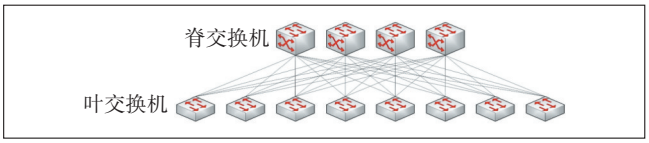

图2　叶脊架构

在叶脊架构中，叶交换机的功能与传统三层网络架构中的接入交换机类似，它直接连接物理服务器、边界路由器、防火墙、存储设备、负载平衡器等终端设备。脊交换机的功能与核心交换机相似，负责叶交换机之间的通信。叶交换机和脊交换机之间的连接通过等价多路径路由协议（ECMP）实现，完成多路径转发。与传统三层网络架构的核心交换机不同，流量可以不经由脊交换机发送至外部网络[7]，而是通过叶交换机来发送。这样脊交换机主要进行流量的转发，不必再承担其他的辅助功能。

叶脊架构的二层设计极大地提升了数据传输的效率，相比于传统的三层网络架构有以下几个方面的优势：一是支持任意类型流量转发，可扩展性高。叶脊架构有很好的横向扩展能力[8]。当把叶交换机和脊交换机控制在一定的比例范围内时，只需复制原有的结构即可实现扩展，十分方便。另外，这种二层设计十分灵活，在数据中心流量较小时，可以适当减少脊交换机的数量；当流量增加时，再相应地增加脊交换机即可。目前，大二层的叶脊架构能够满足大部分数据中心的宽带需求，且对于南北向流量和东西向流量的处理方式完全一样，任何类型的流量均能通过Leaf-Spine-Leaf，3跳可达。针对超大型的数据中心，可以再增加一层Core交换机，跨集群流量可以通过Leaf-Spine-Core-Spine-Leaf，5跳可达[9]。二是网络延迟可预测。在传输流量时，由于2个叶交换机之间的通信路径数目是已知的，且都只通过一台脊交换机，因此网络延迟是可以预测的。三是宽带利用率高。每条叶交换机到脊交换机的上行链路都具有负载均衡功能，这能够充分利用带宽。四是安全性和可用性高。传统三层网络使用的是生成树协议，当一台交换机出现问题或者网络拓扑发生变更时，将会出现重新收敛的情况，导致流量传输效率降低。在叶脊架构中，即使出现类似情况，则不会出现重新收敛的情况，数据依然会继续在其他无故障路径上传输，只是降低了该故障路径的带宽，网络性能基本没有变化[10]。

Overlay网络技术是一种与叶脊网络架构相配合的技术，两者相辅相成。叶脊网络架构有效解决了网

络延迟等问题，但当2个处于不同VLAN具有相同配置的服务器进行通信时，脊交换机不知道应该如何转发，这在一定程度上限制了服务器的灵活部署。Overlay网络技术面向应用层，通过网络虚拟化技术，在现有的IP网络基础上搭建一个虚拟网络。这个虚拟网络可以使服务器之间的通信畅通无阻，从而解决上述问题。目前，Overlay网络常用的技术是VxLAN技术[11]，它既能实现服务器的任意部署，也能解决VLAN ID资源不足的问题。

3 基于叶脊架构的云资源池构建方案

3.1 构建方案

大型云资源池的网络架构更适合采用叶脊架构，该架构支持南北向流量与东西向流量的高速转发，具有扩展性强、高宽带、低时延等优点。在链路的带宽设计上，既要满足目前业务流量的转发需求，也要考虑未来转发速率的提升和业务的进一步扩展[12]。

叶交换机主要负责服务器等各种终端设备的接入。比较好的部署方式是M-LAG双活方式，该方式可以保证业务流量的稳定性、可靠性。每个叶交换机都与脊交换机相连，实现全连接的网络架构[13]。叶交换机连接服务器等终端的下行链路带宽一般为10GE和25GE，根据业务需要，有时也可以达到40GE。叶交换机到脊交换机的上行链路带宽一般为40GE和100GE。此外，Service Leaf和Border Leaf是2种比较特殊的叶节点，它们不连接业务服务器。Service Leaf用于连接防火墙、负载均衡等设备；Border Leaf一般作为数据中心网络的南北向网关，用于连接外部路由器，提供对端PE流量的发送和接收功能。

脊交换机主要负责叶交换机之间流量的高速转发。具体数量根据叶交换机到脊交换机的收敛比来确定，不同业务场景根据具体需求确定[14]。叶交换机和脊交换机之间一般使用三层路由接口互联。物理网络的路由协议首选开放式最短路径优先协议（OSPF），当云资源池规模较大、并进行分区部署时，可以选择边界网关协议（EBGP），同时通过等价多路由协议（ECMP）连接，实现多路径转发、负载均衡、备份[15]。

一种基于叶脊架构云资源池的网络拓扑如图3所示。具体的配置与数量如表1所示。

方案定义了综合接入区、高速转发区、业务资源区和带外管理区4个逻辑区。其中，综合接入区负责内外联功能，可以根据实际需要进行变更和剪裁，高速转发区是连接各个分区的转发平面，业务资源区是物理网络中承载业务服务器的统一资源池[16]。

a）综合接入区。综合接入区作为业务服务区的外延网络，提供云资源池网络、互联网访问、客户专线接入、DCN接入、DCI互联等网络通道，由外网接入模块和骨干网接入模块组成。此部分可根据实际部署需求进行裁剪或添加。

b）高速转发区。承担高速数据交换的任务，作为一个集中的转发交换区域，提供东西向和南北向流量，为各功能区节点提供最佳传输通道。

c）业务资源区。提供所有云业务的底层网络承载，各个云业务系统的内部流量交互在该区域内完成，此部分是云网络的核心部分，其专区包括业务资源模块和云功能模块。

d）（带外）管理区。承载集群管理流量，通过CN2 VPN 1107打通带外管理模块和业务管理模块。

云资源池业务流量如图4所示。

3.2 架构要点

3.2.1 宿主机服务器建设方案

宿主机有3种不同的配置：单台可提供100vCPU计算能力、单台可提供56vCPU计算能力、单台可提供84vCPU计算能力。根据具体需求，这些配置可以进行灵活搭配。本方案中，云资源池节点的本地盘宿主机合计可提供1 368vCPU计算能力。

3.2.2 弹性裸金属建设方案

本方案计划建设5台弹性裸金属服务器。为满足弹性裸金属服务器的使用需求，需配置相应的存储网关服务器，且存储网关服务器的数量不应小于弹性裸金属服务器数量的1/4。

3.2.3 管理服务器建设方案

为满足全栈云资源池的管理需求，需部署云管平台，因此需要为云资源池节点配置管理服务器。在具体的方案中，可以把多种不同的配置的管理服务器搭配使用。管理服务器所纳管的设备主要包括宿主机、本地盘宿主机、GPU服务器、弹性裸金属服务器。

3.2.4 块存储服务器建设方案

根据部署模式和提供方式不同，存储资源可分为集中式存储和分布式存储两大类。

集中式存储主要基于集中式部署的磁盘阵列/磁带库来提供存储，可提供块存储和文件存储，其主流技术包括FC-SAN、IP-SAN、NAS等。它通过硬件来保障性能和可靠性，技术相对成熟，是目前资源池的主

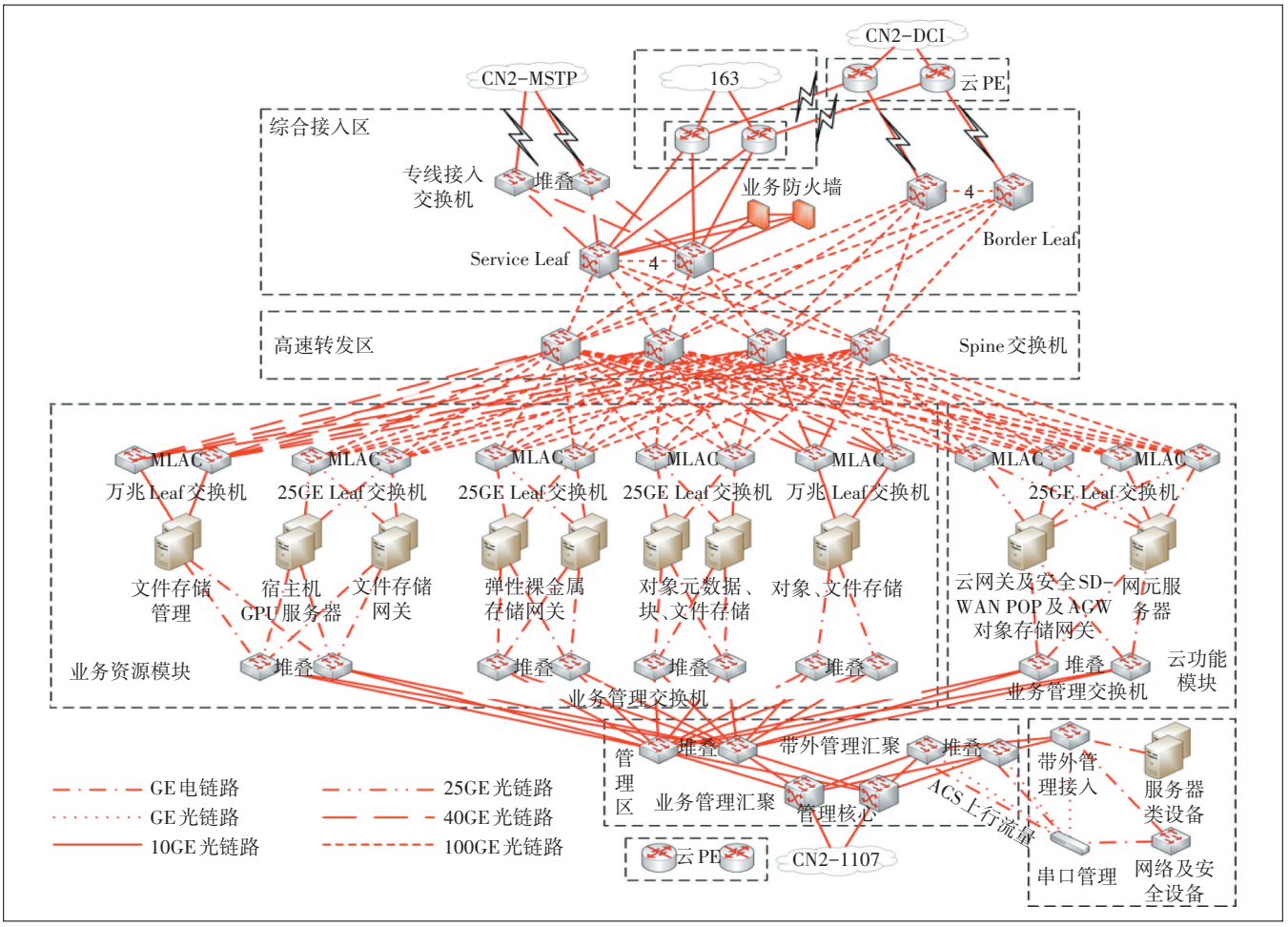

图3　云资源池网络拓扑

要存储资源提供方式，但部署成本较高，扩容不灵活。

分布式存储基于通用或定制化的x86服务器集群提供存储，可提供对象、文件和块存储，具备低成本、灵活扩容、高并发访问等优势，通过软件实现可靠性。它可作为资源池的分级存储手段，满足中低端存储、数据归档备份、大数据存储等需求。其中，对象存储的可用性高、适应度好（可存储各类大小数据）、存储数量大；分布式块存储可满足块存储和卷管理需求；分布式共享文件存储满足大容量文件的共享存储访问需求；大数据文件系统主要存储大数据文件。

本方案的全栈云资源池块存储采用了高性能可扩展的弹性分布式架构设计，硬件支持x86架构的通用服务器和ARM架构国产化服务器，软件采用自研的分布式存储引擎，借助数据冗余和缓存加速等多项技术，提供高可用性、持久性以及稳定时延的块存储服务，并通过OpenStack的管理平台进行统一管理。目前，有普通IO、高IO、通用型SSD、超高IO等多种不同规格的块存储云硬盘，通过采用常规型分布式存储资源池和全闪型分布式存储池，为云主机或裸金属等计算服务器提供持久性块存储的服务。

3.2.5　网络功能服务器建设方案

根据具体业务及数据模型，按照不同的宿主机建设规模，同时考虑性能和冗余需求，搭建网元节点服务器集群。每个集群分别配置不同的接入网络，而整体网络结构配置统一的出口网络设备、客户专线网络设备、云网关网络设备。

4　结语

随着云计算的快速发展、云资源池的规模越来越大，传统的三层网络架构已不能满足大型云资源池数据中心对流量高速转发的要求，其他相关问题也日益凸显。叶脊架构的扁平化二层设计很好地满足了大型云资源池数据中心的建设需求，更能适应东西向流量，具有高稳定、高传输效率、低时延等优点。因此在

表1 设备配置与数量

设备名称	数量	配置
宿主机	273	2路38核(CPU-1C 38 Cores,2.8 GHz),1 024 GB内存,2×480 GB SSD系统盘,2×GE电+4×25GE(按NUMA平衡配置),带ilo管理口,支持IPv6,1×RAID卡
GPU服务器	10	2路28核(6348,2.6 GHz),512 GB内存,2×480 GB SSD系统盘,6×NVIDIA A10(24 GB)GPU显卡,1×双口GE网卡,2×双口25GE网卡
本地盘宿主机	17	2路32核(CPU-1A,3.0 GHz),768 GB内存,系统盘:2×480 GB SSD,数据盘:10×3.2 TB PCIe NVMe SSD,1×双口GE网卡,2×双口25GE网卡(按NUMA平衡配置),1块RAID卡
弹性裸金属服务器	5	2路28核(6 348,2.6 GHz),512 GB内存,2×480 GB SSD系统盘,1×双口GE网卡,1×双口25GE智能网卡(商用卡),1块独立RAID卡。
弹性裸金属存储网关	2	2路32核(CPU-1A,3.0 GHz),512 GB内存,2×480 GB SSD系统盘,1×双口GE网卡,4×双口25GE网卡(按NUMA平衡配置,CX5或CX6),1块独立RAID卡
管理服务器	8	2路38核(CPU-1C 2.8 GHz),1 024 GB内存,2×480 GB SSD系统盘,2×GE+4×25GE(按NUMA平衡配置),1×RAID卡,带ilo管理口,支持IPv6
块存储服务器(HDD)	102	DELL 2路20核(5 218R,2.10 GHz),256 GB内存,2×480 GB SSD系统盘,12×8 TB(SAS直通)+2×3.2 TB PCI-E硬盘,1×RAID卡+1×SAS直通卡,2×GE电+4×25GE网卡,带ilo管理口
块存储服务器(SSD)	140	2路28核(6348,2.6 GHz),512 GB内存,1个480 GB硬盘,8个7.68 TB硬盘,1个双口1GE类型局域网接口,2个双口25GE类型局域网接口(含SFP28光模块)。
SD-WAN POP及AGW	1	2路28核(6 330N,2.8 GHz),512 GB内存,2×480 GB SSD系统盘,1个RAID卡,1×双口GE网卡,4×双口25GE网卡(按NUMA平衡配置)
云网关安全服务器	3	2路28核(6330N,2.8 GHz),512 GB内存,2×480 GB SSD系统盘,1个RAID卡,1×双口GE网卡,4×双口25GE网卡(按NUMA平衡配置)
网元服务器	30	2路32核(CPU-1A,3.0 GHz),512 GB内存,2×480 GB SSD系统盘,1个RAID卡,1×双口GE网卡,3×双口25GE网卡(备注:要求25GE型号为Mellanox CX5或CX6 2块网卡配置在NUMA0,1块网卡配置在NUMA1)
对象存储元数据服务器	5	2路26核(5320,2.2 GHz),384 GB内存,1×480 GB SSD系统盘,24×960 GB SATA SSD数据盘,2SAS卡,2×GE电+4×25GE光(按NUMA平衡配置),带ilo管理口,支持IPv6
对象存储网关服务器	14	2路28核(6348,2.6 GHz),512 GB内存,2×480 GB SSD系统盘,1×RAID卡,2×GE电+8×25GE(按NUMA平衡配置),带ilo管理口,支持IPv6
对象存储节点服务器	28	2路26核(5320,2.2 GHz),384 GB内存,1×480 GB SSD系统盘,24×12 TB SATA数据盘,2×3.2 TB PCIE SSD卡(按NUMA平衡配置),2SAS卡,1×双口GE卡,2×10GE双口卡(按NUMA平衡配置)
文件存储管理服务器	4	2路20核(4316,2.3 GHz),384 GB内存,2×480 GB SSD系统盘,2×960 GB SSD数据盘,1×RAID卡,2×GE电+4×10GE(按NUMA平衡配置),带ilo管理口,支持IPv6
文件存储网关服务器	4	2路28核(6 348,2.6 GHz),512 GB内存,2×480 GB SSD系统盘,2×1.8 TB SAS数据盘,1×RAID卡,2×GE电+8×25GE(按NUMA平衡配置),带ilo管理口,支持IPv6
文件存储节点服务器	17	2路26核(5 320,2.2 GHz),384 GB内存,1×480 GB SSD系统盘,24×12 TB SATA数据盘,2×3.2 TB PCIE SSD卡(按NUMA平衡配置),2SAS卡,1×双口GE卡,2×10GE双口卡(按NUMA平衡配置)
25GE leaf接入交换机	72	48×25GE光口+6×100GE光口,支持堆叠,支持IPv6
万兆leaf接入交换机	6	48×10GE光口+6×40GE光口,支持堆叠,支持IPv6
业务管理交换机	38	48×GE电口+4×10GE光口,支持堆叠,支持IPv6
专线接入交换机	2	48×10GE光口+6×40GE光口,支持堆叠,支持IPv6
带外管理接入交换机	25	48×GE电口+4×10GE光口,支持堆叠,支持IPv6
串口管理交换机	6	支持console口管理,48个console端口(含端口转换模块),至少2个GE光口,支持IPv6
带外管理汇聚交换机	2	48×10GE光口+6×40GE光口,支持堆叠,支持IPv6
SPINE交换机	4	5块18口100GE板卡,1块36口40GE板卡,支持IPv6
Service Leaf交换机	2	2块18口100GE板卡,1块36口40GE板卡,1块18口10GE板卡,支持IPv6
业务管理汇聚交换机	2	48×10GE光口+6×40GE光口,支持堆叠,支持IPv6
管理核心	2	不少于20个10GE端口

构建大型云资源池时,可以首选叶脊架构。

参考文献:

[1] 谢羽成,孙健,丁宁,等.一种基于叶脊网络架构云资源池的优化方法[J].江苏通信,2022,38(3):34-37,49.

[2] 李猛.大规模数据中心内云计算网络演变的研究及分析[J].江苏通信,2021,37(1):93-96.

[3] 乔爱锋.云网融合体系架构及关键技术研究[J].邮电设计技术,

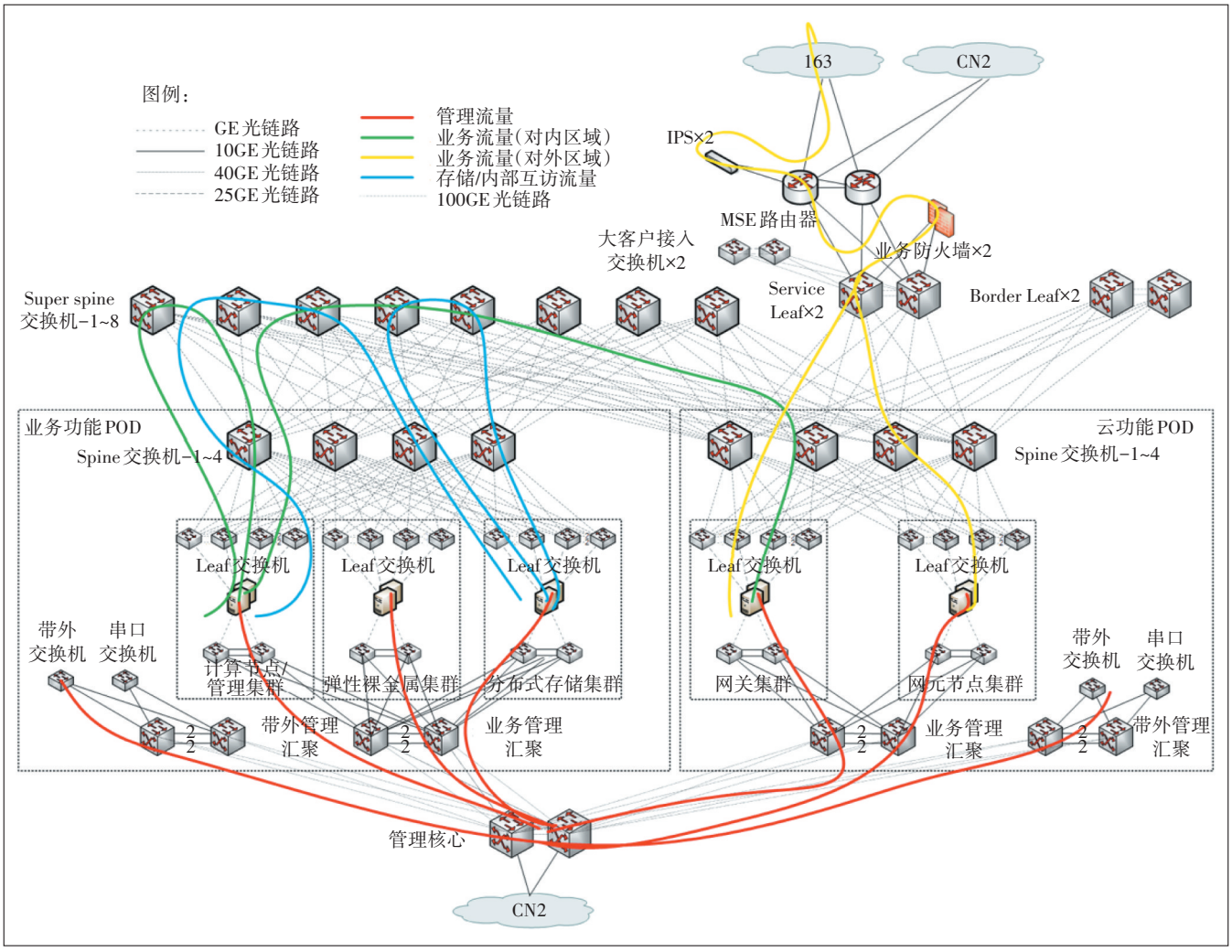

图4　云资源池业务流量示意

2022,2022(6):14-18.
[4] 余嘉历.应用叶脊网络架构改造企业数据中心网络[J].数字化用户,2019,25(29):68.
[5] 安志聪.数据中心叶脊网络架构及预端接布线设计探讨[J].建筑电气,2020,39(11):54-59.
[6] 吴佳騄,刘恒.云网融合场景下的传输网络架构探讨[J].通信技术,2022,55(1):70-76.
[7] 包琅允.叶脊架构在数据中心的应用[J].邮电设计技术,2021,2021(2):74-77.
[8] 马季春,孟丽珠.面向云网协同的新型 城域网[J].中兴通讯技术,2019,25(2):37-40.
[9] 陆锋.数据中心网络的Spine-Leaf架构[J].IT经理世界,2020,23(2):82.
[10] 栾琳琳,杨占胜,刘海涛,等.面向边缘集群的软件定义流量控制[J].电脑编程技巧与维护,2022,2022(8):16-19,59.
[11] 唐玉涛.云计算资源池数据中心的网络架构[J].门窗,2020(4):292-293.
[12] 刘茂.基础设施云资源池的设计参考[J].科技与企业,2015,2015(15):101-101.
[13] 闫岩,姜海洋.面向算力下沉的新型城域网一体化演进的研究[J].电信工程技术与标准化,2022,35(8):80-82,92.
[14] 宫良.SDN/NFV技术在未来城域网中的应用[J].广播与电视技术,2018,45(1):16-20.
[15] 李彦刚,邓文平,王宏,等.域内路由协议OSPF与IS-IS差异性的研究与分析[J].计算机科学,2015,42(z1):256-259.
[16] 辛笛,徐海宁.业务平台云资源池建设方案设计[J].科技创新导报,2017,14(17):157-158.

作者简介：
文娅,中国人民解放军31401部队;戚金园,学士,主要从事IT支撑系统以及数据网络和业务平台设计工作。