ICS 33.040.40
CCS I631

T/NIDA

全 球 固 定 网 络 创 新 联 盟

T/NIDA 006-2025

通算数据中心网络技术要求

Technology Requirements for Network Construction of General Computing Data Center Network

2025-09-26 发布 2025-09-26 施行

全球固定网络创新联盟（NIDA）发布

目录

前　言

本文件按照 GB/T 1.1-2020《标准化工作导则 第 1 部分：标准化文件的结构和起草规则》的规定起草。

请注意本文件的某些内容可能涉及专利权和著作权。本文件的发布机构不承担识别专利和著作权的责任。全球固定网络创新联盟不对标准涉及专利的真实性、有效性和范围持有任何立场；不涉足评估专利对标准的相关性或必要性；不参与解决有关标准中所涉及专利的使用许可纠纷等。

本文件由全球固定网络创新联盟技术委员会提出并归口。

本文件起草单位：中国信息通信研究院云计算与大数据研究所、中国工商银行数据中心、华为技术有限公司、平安科技（ 深圳 ）有限公司

本文件主要起草人：郭亮、王少鹏、余学山、张力、李久勇、李晨飞、蒙祖瑞

通算数据中心网络建设技术要求

1 范围

本文件规定了通算数据中心网络建设的技术要求，内容涵盖通算数据中心网络架构，组网场景及关键技术。

本文件适用于通算数据中心网络的建网规划和技术评估，并对网络测试提供技术依据。

2 规范性引用文件

下列文件中的内容通过文中的规范性引用而构成本文件必不可少的条款。其中，注日期的引用文件，仅该日期对应的版本适用于本文件；不注日期的引用文件，其最新版本（包括所有的修改单）适用于本文件。

3 术语和定义

4 缩略语

下列缩略语适用于本文件。

BGP：边界网关协议（Border Gateway Protocol）

CI/CD：持续集成与支持交付（Continuous Integration and Continuous Delivery）

DC：数据中心（Data Center)

ECN：显式拥塞通知（Explicit Congestion Notification)

FC：光纤通道 （Fibre Channel）

FCT：流完成时间（Flow Completion Time)

FET：流有效吞吐（Flow Efficient Throughput)

FNR：流重传率（Flow NAK Rate)

HDD：机械硬盘（Hard Disk Drive)

IOPS：每秒进行读写操作的次数（Input/output Operations Per Second）

MDC：多数据中心控制器（Multi-Datacenter-Controller）

M-LAG：跨设备链路聚合组（Multi-chassis Link Aggregation Group)

NOF：网络承载的NVMe (NVMe over Fabric)

NVMe：非易失性高速传输总线（Non-Volatile Memory express)

PFC：基于优先级的流控（Priority-based Flow Control)

POD：交付点（Point of Delivery）

RDMA：远程直接内存访问（Remote Direct Memory Access)

RoCE：基于融合以太的远程内存直接访问协议（RDMA over Converged Ethernet）

SLA：服务质量协议（Service Level Agreement）

SSD：固态硬盘 （Solid-State Drive）

SZTP：安全零接触配置（Secure Zero Touch Provisioning）

5 通算数据中心网络建设概述

5.1 业务演进趋势

在数字化转型背景下，基于5G、大数据、人工智能、物联网等技术的行业数字化升级已经覆盖各行各业。万物智联的新业务对企业数据中心产生了新的诉求。以金融行业为例，数字金融已成为一种利用数字技术和互联网平台以及服务和产品的数字化、智能化和在线化，为客户提供更加便捷、高效、安全和个性化的金融服务和产品的新业务形式。随着AI技术的强化，数字金融将提供更加精准的金融服务，提升运营效率的同时改变商业模式，如图1所示。

图 1 数字金融场景和需求

数据中心在数字化时代的新型诉求包括业务自动化和高效开放、超高可靠、高性能、智能运维以及全生命周期安全。这些要求具有以下内涵：

- 企业通过在数据中心构建平台+APP的服务模式向客户提供全方位的服务。APP的迭代更新频率高，业务必须尽快部署。传统的手动配置模式耗时且容易出错，且跨多厂商设备和多业务技术的业务部署对运维人员提出了更高要求。数据中心网络需要提供高效、开放的自动化业务部署能力，构建包含运营运维系统、管控系统以及多厂商设备互联的端到端业务自动化架构，并在underlay网络、overlay业务、网安一体化等业务场景中实现端到端的全流程业务部署自动化。
- 在5G移动业务时代，全天24小时都面临客户访问请求。每日客户访问量和日均服务调用次数都在亿次级别以上。这要求数据中心网络能提供7*24小时超高可靠和超快的响应速率。
- 企业对大数据和AI技术有广泛的需求的情况下，对算力、存储的规模会不断增长。随着大语言模型和存储技术的演进，数据中心要求构建大带宽无损网络，以保障智能计算、模型训练和推理及高速存储的无丢包和低时延传输。
- 企业客户在数字化时代对体验提出了更高的要求。持续且快速的服务需求对网络故障处理时效性提出挑战。未来，随着数据中心规模的进一步扩展，需要借助AI、数字孪生、专家经验等技术实现网络的自感知、自分析、自修复、自优化，从而推动数据、技术、流程和组织的智能协作、动态优化和互动创新。

- 随着数据中心业务深入生活场景，安全防护链也日趋复杂。与此同时，外部信息安全形势严峻，面临的安全威胁愈发复杂。针对高安全保障要求，数据中心网络应构建全方面立体式的安全保障体系，以防止数据泄露，保护客户资产和隐私。

5.2 技术演进趋势

5.2.1 自动化业务部署

提升数据中心业务部署效率和准确率，有效增强客户体验并降低网络运维OPEX。构建一个集资源系统、业务编排系统、工单系统、业务协同与部署系统以及业务检测系统于一体的自动化交互式ICT系统，实现业务规划、资源分配、配置变更审批、业务部署、业务检测及业务可视化等全流程的自动化，已成为数据中心网络运维管理的关键需求。

数据中心业务通常跨POD甚至跨DC部署，涉及多厂家设备及频繁的业务变更。业务自动化要具备敏捷开发、持续演进、开放可编程、易用性高等能力。当前随着CI/CD相关技术的演进，如工作流编排、多厂商设备即插即用等，为融合构建自动化ICT系统提供技术支撑。

5.2.2 智能计算

随着AI大语言模型的突破性发展，大模型训练和强化训练也不断演进。高性能计算是智算数据中心的核心需求，专业高性能芯片如GPU、FPGA、ASIC 等被广泛应用于支持大量的浮点运算和并行计算需求。尽管AI训练算法的提升在一定程度上降低了对智算集群规模的依赖，但单芯片算力提升速率远不能满足大模型训练参数量增长对算力的需求，scaling-law原则依然适用，智算集群的规模仍然在不断扩大。

与此同时，模型推理能力的增强也推动了基于推理的应用部署，模型推理对算力资源的需求也在不断扩大。高投资的智算集群，提升算力效率是关键。RoCEv2已成为智算网络的首选。它有效降低了网络延迟，但也对丢包极其敏感，这对智算网络提出了更高的诉求，需要大带宽、无损、稳定时延以及更高的网络吞吐来提升通信效率

5.2.3 存储

数字化将使得数据中心面临并行且快速的数据处理能力。作为人工智能应用和服务的基础设施，储系统需要提供50GB以上带宽和100万以上IOPS的极致性能。基于闪存技术的成熟，固态硬盘（SSD）提高了存储密度和性能，且成本也逐渐降低，逐渐取代HDD获得广泛的应用。非易失性内存主机控制器接口规范（NVMe）为SSD提供了高效的通信机制，能充分发挥SSD的性能优势，使其能够充分利用SSD的性能优势，提升读写速度、降低延迟等。

与SSD和NVMe配套的存储网络也需要支持更高性能和无损转发。传统的FC网络面临带宽增长的技术瓶颈，且与数据中心其他区域的技术割裂，增加了运维的复杂度。以太网在大带宽、无损传输等方面有良好的技术积累，且具备完整的供应链。基于高速以太网的NVMe全闪存存储和NoF存储网络将是未来数据中心的主要发展方向。

5.2.4 AI for Network

数据中心业务的互访关系错综复杂，完成一个业务可能要涉及多个业务系统协同。端到端实时感知业务质量以及对各业务系统的质量进行分段定界和可视化呈现，已成为数据中心可视化运维的新诉求。随着分布式和大数据业务的兴起，数据中心东西向流量占比不断增加，分布式流量极易引起微突发等问题，传统监测手段难以察觉此类问题。此外，分布式架构和智能计算带来设备规模的倍增，业务稳定运行要求提前识别潜在的网络隐患并快速定位故障，这也给网络运维带来了新的挑战。

为了应对上述挑战，数据中心运维迫切需要从传统的手动运维向智能网络运维转型。随着Telemetry秒级采集、数字孪生、知识图谱、AI、Co-pilot和Agent等技术能力的发展，使得更加高效、自动化、智能化的网络运维成为可能。

5.2.5 组网演进

为满足数字化转型所带来的业务发展需求，需要部署多个数据中心来提供服务。同时，数据安全、业务的可靠性和连续性也越来越受到重视，备份和容灾成为普遍需求。需要建设多个数据中心来解决容灾备份问题。

伴随着5G，云计算和大数据的发展，虚拟化和资源池化成为主流需求，需要整合跨地域、跨DC资源，形成统一资源池。此外，业务系统采用多DC分布式部署，形成多活，为用户提供就近服务，提高用户体验。分布式多数据中心成为当前的主流解决方案。

为满足数据中心持续创新的诉求，数据中心架构需要具备灵活可扩展、可演进性。数据中心网络的设计需要满足计算、存储、大数据中心等的发展需求，并能兼顾一定周期内的技术演进，并支撑分布式和云化数据中心的演进。

5.2.5.1 Multi-POD

单 DC 内采用多 POD 的建设方式，通过合理的业务 POD 划分，既可提高计算、存储资源池化共享能力，同时也保障 DC 的可扩展性和各 POD 区的弹性扩缩容能力。在 POD 间部署适当的安全策略，实现不同层次的安全防护以和故障隔离。业务支持在 POD 间进行灵活备份和迁移，从而实现负载均衡和高可靠性。未满足未来业务容量和性能的持续提升需求，数据中心应预留合理的带宽容量和可靠的接入设置。

5.2.5.2 Multi-DC

随着数字化程度的不断提高，数据中心对安全性和可靠性提出了更高的要求。数据中心必须持续确保其可靠性，满足 7*24 小时网络运行的需求，并满足在同一数据中心或跨数据中心部署的系统和应用的高可靠性要求。因此数据中心建设需要支持多 DC 场景，包括:

a) 同城双活

在同城两个 DC 部署同时运行的两套业务系统，相同子系统的安全策略保持一致，对外提供相同服务，形成双活。该场景提供双倍服务能力同时可互相实时灾备接管。在应用处理层面上实现完全冗余，极大地提升了服务的连续性和可靠性，使用户无法感知故障。

b) 异地容灾

异地的灾备中心是同城双活的两个主数据中心的备份数据中心，用于备份主数据中心的数据、配置、业务等。当主用双活数据中心因自然灾害等原因发生故障时，异地灾备中心可以快速恢复数据和应用，以保证业务正常运行，并将灾害造成的损失降至最低。

5.3 政策约束

数据中心在不同行业有特定的政策约束。以金融行业为例，金融行业对国家具有重要的意义，在促进经济发展、优化资源配置、稳定经济、推动创新和合作方面发挥重要的作用。各国也出台对应的政策，保障数据中心的建设。例如，应建立并完善数据中心的智能运维机制，加强多场景协作与多站点集成管理与控制，提升站点感知、异常检测和故障预测能力，降低人工操作风险。

6 通算数据中心建网架构

6.1 单 DC 多 POD 建网

6.1.1 总体架构

在单DC内划分多POD区域，用于承载不同的业务。POD间由交换核心区连接，如图2所示。

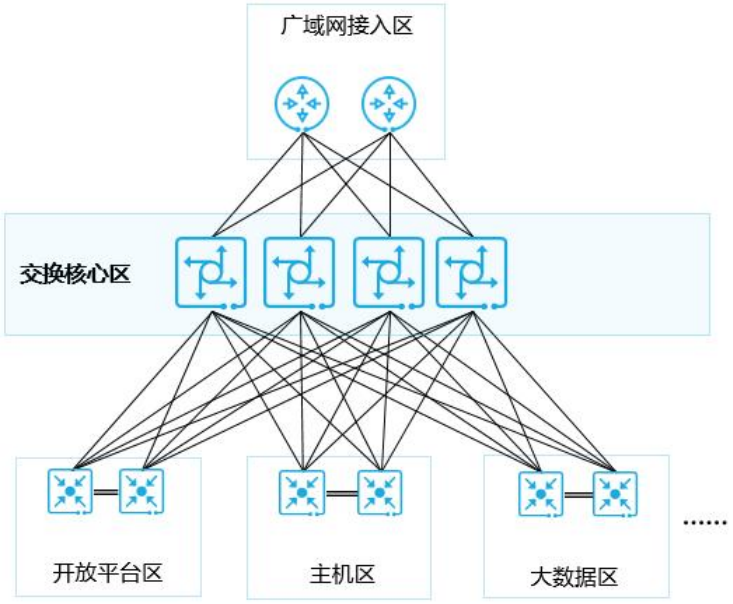

图 2 单 DC 建网架构

6.1.2 交换核心 POD

交换核心POD是数据中心网络的核心，用于连接数据中心内部各个功能分区，是数据中心的交换总线。

- 交换核心POD由核心设备独立部署。采用Spine-Leaf架构确保了交换核心POD的可扩展性，并允许灵活地添加或删除功能POD。
- 交换核心POD的带宽要求具备可扩展性，并对未来业务发展预留带宽。当前200G/400G是主流，而800G则成为大型数据中心的新趋势。
- 交换核心POD需支持建立三层路由控制平面，通过标准路由协议的快速收敛来保证网络可靠性。

6.1.3 智算 POD

智算POD指执行AI模型训练或推理的分区，通常划分为3个平面，即业务平面、参数平面和管理平面，如图3所示。

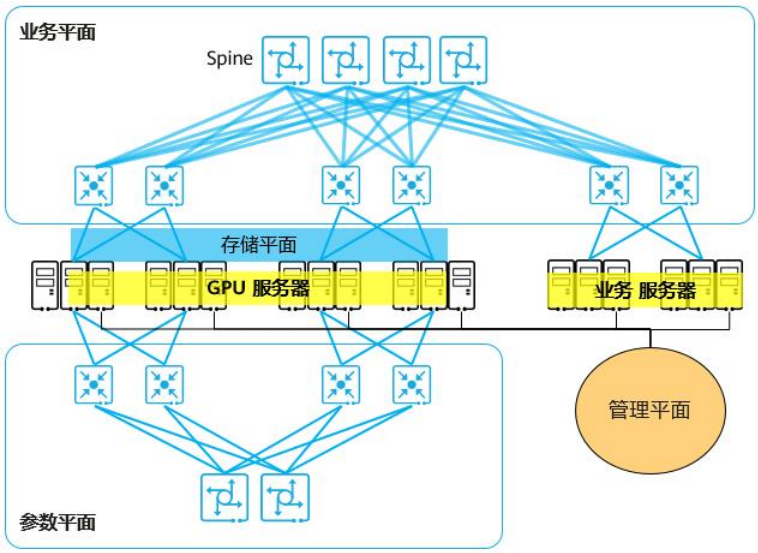

图 3 智算 POD 建网架构

在智能计算 POD 中，GPU 之间的通信效率至关重要。RDMA 被广泛用于数据传输和内存访问，智能计算网络通常基于 RoCEv2 协议构建。基于 RoCEv2 的 AI 训练对网络传输指标（如网络延迟、丢包率和抖动）非常敏感，1‰的丢包率会使训练效率降低 50%。RoCEv2 并未提供全面的可靠性保护机制。为了提高智能计算集群的算力可用率，需要高性能、大带宽且无丢包的网络。构建智能计算 POD 的需求如下：

- 智能计算POD的规模可根据服务需求灵活调整，需具备高网络可用性和可扩展性，采用CLOS spine-leaf架构。
- 随着AI模型规模和GPU带宽的增加，智能计算POD的网络带宽需求也在提升。目前，200G/400G为主流，800G则成为支持大规模AI训练的新趋势。
- 智能计算POD需要非阻塞网络，传统ECMP无法支持无损网络传输。对于少量、大带宽AI流量的调度，需要更高效的负载均衡机制，同时还需要更高效的流量控制和拥塞控制。
- 随着网络规模的扩大，智能计算POD的网络运维管理复杂性增加。需提供基于RDMA流量的可视化功能，以简化并缩短故障检测与分析时间。

智能计算POD的关键技术如下：

高性能	流量控制	PFC 优先级流量控制
	拥塞控制	ECN
	负载均衡	网络级负载均衡/包级负载均衡
高可靠性	设备级高可靠	光模块级故障检测
		数据面故障感知快速收敛
智能运维	数字地图	多维可视/路径导航/存储即插即用
	网络故障定界定位	故障感知和预测/定界定位
	应用故障定界定位	应用故障感知/定界定位

6.1.4 大数据 POD

大数据POD用于处理大规模的分布式数据计算任务。例如在金融行业数据中心中，大数据POD用于计算和分析客户信息，包括开户、交易和转账的金融行为、偏好和风险分析。它还可以用于计算和分析金融组织信息，如业务处理流程的改进点、资源分配优化以及金融数据。图4展示了大数据POD的网络架构。

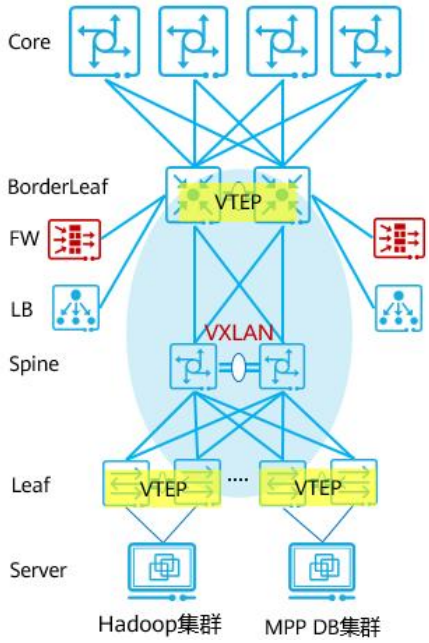

图 4 大数据 POD 建网架构

大数据 POD 构建有如下要求：

- 大数据POD的规模会因业务需要灵活调整，要求网络具备高可用和可扩展能力，通常采用spine-leaf架构以支持ECMP；
- 需要支持灵活的服务器迁移，并且网络应支持业务支持自动化配置和删除；
- 大数据POD处理大量数据输入输出，要求网络有足够的接入带宽，如10G接入及100G/200G汇聚；
- 大数据计算业务，尤其是与客户移动应用相关的业务，要求快速响应，因此网络要保证低延迟和无拥塞；
- 大数据计算涉及关键数据处理，可靠性要求更高。因此网络要支持故障快速倒换、故障快速定位与恢复等以及故障预测。 同时，网络还应该提供业务级的可视化运维能力，以可实时呈现业务的性能指标和路径变更等信息。

大数据 POD 的关键技术包括：

自动化部署	业务自动化	业务自动化
	业务流编排	业务流编排
	配置仿真	配置仿真
开放	北向开放	北向开放
	南向开放	多设备管控
高性能	拥塞控制	ECN/基于 VXLAN 的 Overlay ECN

高可靠性	链路级高可靠	M-LAG 可靠性
智能运维	应用故障定界定位	应用故障感知
		应用故障定界定位
安全	设备层安全	设备层安全
	网络层安全	网络层安全
	管控层安全	管控层安全

6.1.5 存储 POD

存储 POD 用于数据存储和管理。数据中心应提供高性能、高可靠、高可用的数据存储服务，以确保数据的安全性和持久性，并满足不同应用的数据存储需求。图 5 展示了存储 POD 的网络架构：

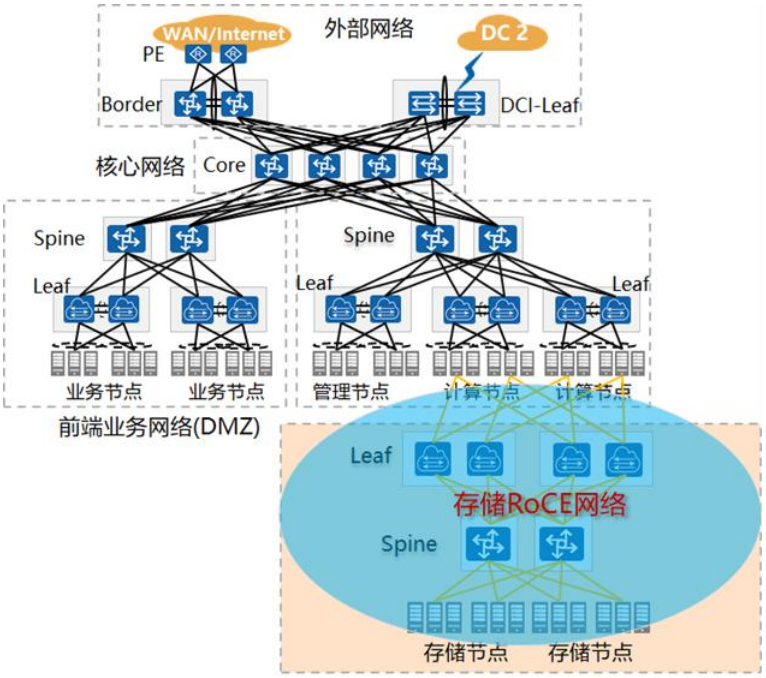

图 5 存储 POD 建网架构

存储网络是指计算服务器用于访问存储数据的通信网络。通常，该网络是一个独立的物理网络。随着全闪存存储器的普及，基于 NVMe over fabric 的方案成为主要选择。此外，为了支持 TCP 和 RoCE 协议的混合应用，并满足统一数据中心网络技术的要求，基于以太网的 NOF 解决方案更受青睐。NVMe over Ethernet 需要高性能、无损、高可靠性、易用且统一运维的网络服务。构建存储 POD 的需求如下：

- 存储POD的规模可根据业务需求灵活调整，需具备高网络可用性和可扩展性，广泛采用Spine-Leaf架构。
- 存储网络承载对丢包敏感的RoCE数据，必须支持高效的流量控制和拥塞控制能力，以确保无损传输。
- 鉴于数据中心关键信息存储的重要性，存储网络的可靠性要求极高。需实施多层次可靠性保障。在链路级可靠性方面，基于本地冗余链路，快速链路故障检测与切换是关键要求；在设备级可靠

性方面，要求在本地冗余链路及远端设备上均能快速检测并执行故障切换；在网络级可靠性方面，需检测静默故障并支持网络级收敛。

- 为存储海量数据，存储系统通常需管理大量主机，并允许新主机动态接入网络。存储网络需快速检测并管理新增主机，智能调整存储网络配置。
- 随着存储网络规模的扩大，网络运维日益复杂，智能化网络运维变得不可或缺。

存储 POD 网络的关键技术包括：

高性能	拥塞控制	ECN/AI ECN
	流控控制	PFC 优先级流量控制
高可靠性	链路级高可靠	M-LAG 可靠性
	设备级可靠性	数据面故障感知快速收敛
	网络级可靠性	静默故障快速收敛
智能运维	数字地图	多维可视/路径导航/存储即插即用
	网络故障定界定位	故障感知和预测/定界定位
	应用故障定界定位	应用故障感知/定界定位

6.2 多 DC 建网

6.2.1 总体架构

为满足数据中心超高可靠性诉求，多数据中心部署方案，如同城数据中心的主备或双活部署，以及异地三数据中心部署方案，成为必然选择。图 6 展示了多数据中心部署架构。

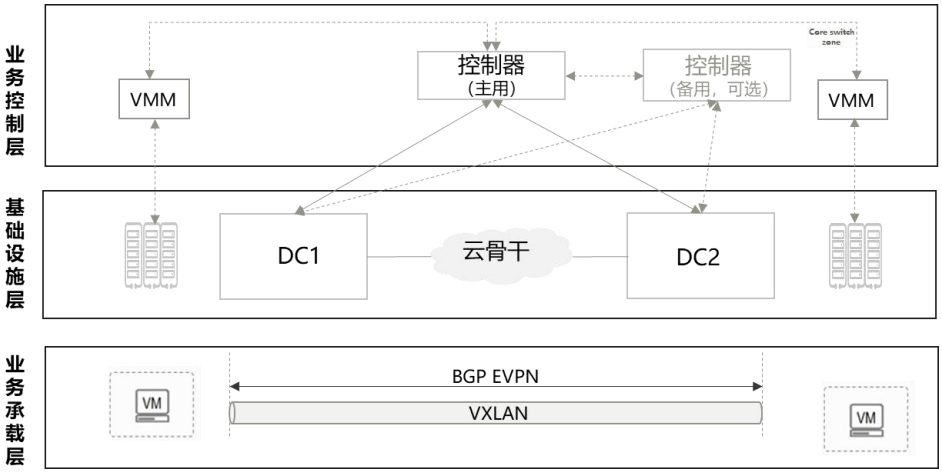

图 6 多 DC 建网架构

多DC间通过MDC (Multi-Datacenter-Controller)连接各DC的控制器，以实现跨DC的业务自动化部署、灵活的安全策略控制以及跨DC网络运维。

6.2.2 部署方案

多 DC 部署通常包含两重场景：

a) 同城双活：通常在同一城市的主备数据中心部署相同的业务系统，并同时对外提供业务。

b) 异地容灾：在不同地域的数据中心部署相同的业务系统。正常情况下，异地数据中心的业务系统不提供对外服务，仅用于备份。下图展示了异地容灾数据中心的服务互联需求：

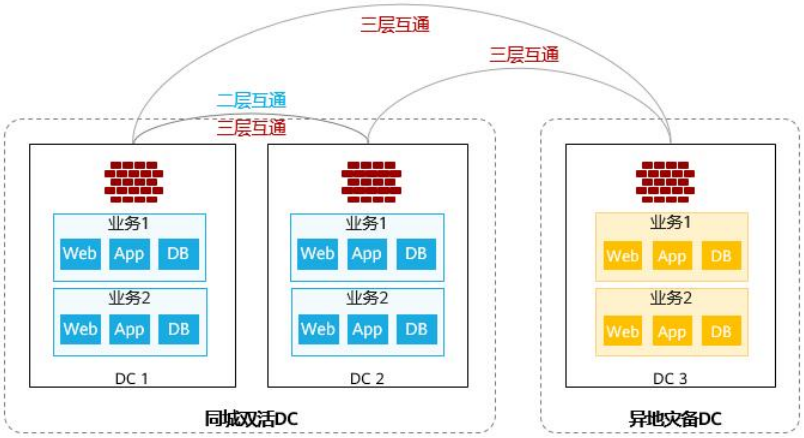

图 7 异地灾备通信流程

为了支持数据中心的高可靠性，需要采用多数据中心（Multi-DC）架构。为了实现应用和数据的备份与恢复目标，多数据中心应具备以下能力：

- 多数据中心状态检测与切换机制需提供更高的可靠性。
- 为提升服务部署效率并降低配置错误率，需实现跨数据中心的自动服务部署。
- 不同数据中心间的存储需进行数据复制以备备份之需。鉴于数据中心间通常距离较远，需实施支持长距离传输的流量控制与拥塞控制能力，以确保数据无损传输。
- 在跨数据中心大数据协作处理服务的场景下，高带宽成为必需。
- 跨数据中心的安全防护措施不可或缺，以保障不同安全域间数据传输的安全性。

多 DC 的关键技术包括：

跨 DC 可靠性	同城双活
	异地容灾
跨 DC 高性能	长距无损传输
跨 DC 运维	增强远距离 PFC
跨 DC 安全	多 DC 运维
	多 DC 安全

6.3 建网要求

基于通算数据中心网络内单 DC 和跨 DC 各类场景和业务对网络的诉求，通算数据中心网络建设需要支持如下建网能力：

- 单 DC 内需要支持：

a) 自动化部署：包括业务自动化部署、业务流编排能力和配置仿真校验

b) 开放：包括北向开放对接 ICT 系统及南向开放对接多厂家设备

c) 高性能：包括流量控制、拥塞控制、负载均衡技术
d) 高可靠性：包括基于链路级、设备级、网络级可靠性
e) 智能运维：包括数字地图、网络故障定界定位、应用故障定位能力
f) 安全：包括设备层安全、网络层安全、管控层安全

- 跨 DC 需要支持：

a) 跨 DC 可靠性：包括同城双活和异地容灾可靠性
b) 跨 DC 高性能：包括长距离 PFC 能力
c) 跨 DC 运维：包括跨 DC 的智能运维和故障定界定位
d) 跨 DC 安全：包括跨 DC 的数据加密和管理安全等

7 通算数据中心单 DC 网络建设关键技术能力要求

7.1 自动化部署

7.1.1 业务自动化部署

网络自动化技术用于提供网络业务的自动化部署能力，以提高业务部署效率。业务自动化的技术要求包括：

a) 支持underlay网络的业务自动化部署，常见的网络配置包括但不限于IGP、BGP、Eth-trunk、DHCP、VLANIF、STP实例等；
b) 支持overlay网络的业务自动化部署，包括但不限于VXLAN、EVPN等；
c) 支持网安融合业务的自动化部署，包括网络配置和安全配置的联合自动化部署；
d) 支持异构网络业务的自动化部署，包括多协议互通配置自动化部署。

7.1.2 业务流编排

业务流编排技术是一种网络服务化的能力。通过业务自定义编程功能，使不具备专业开发能力的客户可以通快速灵活地编排网络操作，将业务知识应用于网络，以适应快速变化的业务需求。服务流编排需要简化系统之间的边界，使服务对象能够在系统之间流动。服务流编排的技术要求包：

a) 业务流编排支持设计态功能，包括：
   1) 支持动作管理功能，动作是组成流程的基本元素，要包含动作模型描述和动作设计描述，分别描述一个动作在执行过程中需要的输入和输出属性数据模型，并描述动作的执行方式，如do、retry、rollback和dryrun等；
   2) 支持流程设计与管理功能，支持将多个原子接口编排成一个流程，流程可独立使用或以嵌套模式运行；
   3) 支持可拖拽的UI界面能力以简化设计；
   4) 支持端到端网络能力建模，而非单一站点功能与特性，并支持将模型化的网络能力编排为业务工作流；
   5) 支持模型驱动，对网络能力进行抽象建模，屏蔽网元具体实现，适配不用转发协议和厂商间实现差异；
   6) 支持必要的业务资源管理、自动分配及回收。
b) 业务流编排支持运行态功能，包括：
   1) 支持自动生成API接口，通过描述网络服务的功能而非具体实现，实现与其他ITSM平台系统集成；
   2) 支持工作流统计管理，包括执行记录、状态维护等；
   3) 支持业务配置自动分解，并支持分解后dryrun能力，并可支持配置变更自动仿真校验；

4) 支持业务配置自动下发，并支持配置下发后的一致性校验；
5) 支持业务部署失败时自动回滚，并自动检测回滚后配置一致性。

7.1.3 配置仿真校验

仿真校验技术用于事先对配置变更的影响进行评估验证，拦截错误的网络配置，以辅助专家验证策略的有效性并降低现网实施风险。仿真校验要求支持如下功能：

a) 支持多分支仿真验证，可根据各分支不同配置变更的影响选择最优方案；
b) 支持配置变更风险分析，并提供从多个角度（包括但不限于路由、隧道、转发）的风险评估；
c) 支持多场景模拟，包括但不限于协议模拟、网络连接模拟和可靠性模拟；
d) 支持配置变更优化建议，可基于风险评估结果提供配置优化建议。

7.2 开放

7.2.1 北向开放

数据中心的业务涉及多系统间继承和互通，且不同系统通常跨多厂商设备构建。数据中心网络管理与控制系统必须在北向和南向都具备开放性和可编程能力。数据中心网络管控系统北向开放要求支持如下功能：

a) 支持北向图形用户界面（GUI）功能，以实现人机交互；
b) 支持北向API网关，以实现机器间交互及不同系统之间的调用；
c) 支持北向接口协议，包括但不限于RESTFUL、SNMP、FTP。

7.2.2 南向开放

数据中心网络管控系统支持南向开放能力以实现多厂商设备对接，实现跨设备网络和业务管理和运维。南向开放要求支持如下功能：

a) 支持多厂商设备极简适配能力，以快速支持多厂家设备管理和运维；
b) 支持多厂商设备开箱即用；
c) 支持南向接口协议，包括但不限于RESTCONF、SNMP、Telemetry、FTP

7.3 高性能

基于 RoCEv2 协议，智能无损网络旨在实现智能计算与存储网络的融合，应整合先进的流量控制、拥塞控制和负载均衡技术，以在以太网上传输流量，实现零丢包、低延迟和高吞吐量。

流量控制与拥塞控制必须协同工作，以解决网络拥塞问题。两者之间的区别在于：流量控制是端到端的，需要抑制发送端的传输速率，以便接收端能够及时接收数据；而拥塞控制是一个全局过程，涉及所有主机、网络设备以及所有影响网络传输性能的因素。

7.3.1 流控技术

流量控制技术通过抑制发送端的发送速率，使接收端能够及时接收到数据。流控技术需要支持如下功能：

a) 支持PFC技术，包括：
    1) 支持基于优先级的流量控制。PFC是IEEE 802.1Qbb标准中定义的一种流控机制，旨在通过实现按优先级执行暂停控制来防止以太网中的帧丢失。在AI或存储场景中，如RDMA等对丢包敏感的流量，可设定特定的服务类别并设定特定的服务优先级，以进行流量控制；
    2) 支持PFC死锁预防，以降低网络出现PFC风暴的概率，PFC风暴会阻塞流量；
    3) 支持PFC死锁自动恢复，破除PFC死锁环路，释放缓存依赖，解决PFC风暴类问题；

4) 支持PFC门限设置。

7.3.2 拥塞技术

拥塞控制是一个全局性的过程，目的是让网络能承受现有的负荷。拥塞控制通常需要转发设备、流量发送端、流量接收端协同作用，并结合网络中的拥塞反馈机制来调节整网流量以缓解拥塞。拥塞控制需要支持如下功能：

a) 支持ECN技术，包括：

1) 支持ECN与PFC协同工作。应合理设置ECN阈值与PFC阈值，确保ECN阈值与PFC阈值之间的缓冲区能够容纳从ECN拥塞标记时刻到应用降速时刻之间的时间段内应用发送的流量，从而避免触发网络PFC流控；

2) 支持ECN门限设置要兼顾网络中同时存在的时延敏感小流和吞吐敏感大流。应合理设置ECN阈值，以满足高流量发送速率的吞吐量敏感大流的带宽需求，同时满足延迟敏感小流量的延迟要求，即时触发ECN拥塞标志以通知应用降低速率；

3) 支持ECN机制。转发设备可在IP数据包中标记ECN拥塞。流量接收端根据IP数据包中的ECN字段检测网络拥塞。若检测到网络拥塞，则发送带有ECN拥塞信息的数据包，通知发送端流量降低发送速率。

b) 支持AI ECN技术，包括：

1) 支持智能地根据实时网络流量模型调整无损队列的ECN阈值，确保在零丢包情况下实现低延迟和高吞吐量，为无损服务提供最佳性能；

2) 支持AI ECN智能算法，基于实时网络流量模型进行AI训练，预测网络流量变化，并及时推断出最优的ECN阈值。同时支持ECN阈值可根据实时网络流量变化实时调整，精确管理和控制无损队列缓冲区，确保整个网络的最佳性能

3) 支持设备收集实时网络流量特征，并将收集的数据发送至AI ECN组件；

4) 支持AI ECN与队列调度技术协同使用。无损队列的AI ECN功能能够实现网络中TCP流量与RoCEv2流量的混合调度，确保RoCEv2流量的无损传输，实现低延迟和高吞吐量；

c) 支持Overlay ECN技术，包括：

1) 支持在VXLAN网络中应用overlay ECN技术，支持IP报文中的ECN字段映射到VXLAN报文中，将拥塞状态传递到流量接收端，及时缓解VXLAN网络的拥塞，实现网络性能的最大利用；

2) 支持硬件网络overlay、软件vSwitch overlay及软硬混合overlay场景中的AI ECN功能；

3) 支持在入VXLAN隧道的设备发生拥塞的overlay ECN处理；

4) 支持在VXLAN隧道的转发设备发生拥塞的overlay ECN处理；

5) 支持在出VXLAN隧道的设备发生拥塞的overlay ECN处理。

7.3.3 负载均衡

负载均衡是实现网络高吞吐的前提。传统的基于五元组（源IP、目的IP、源端口、目的端口和协议）的静态负载均衡方式存在明显不足，会导致链路间流量分布不均，进而降低网络吞吐量。先进的负载均衡应支持网络规模的负载均衡，并确保网络吞吐量超过90%。负载均衡的技术要求包括：

a) 支持网络级负载均衡功能，包括：

1) 支持感知业务流量模型，如智算网络AI训练任务所涉及的接入设备及流量模型；

2) 基于流量模型和网络资源状态的统一路径计算，以解决由局部和全局流量负载不平衡引起的流量拥塞问题，并提高网络吞吐量。

b) 支持包级负载均衡技术，通过包喷洒实现数据包均匀分散到网络链路，并支持解决数据包乱序重组的问题。

7.4 高可靠性

在数据中心中，恢复点目标（RPO）和恢复时间目标（RTO）有着严格的要求。网络故障检测的速度和类型以及快速网络故障切换对于提高网络可靠性至关重要。应实施多层次的可靠性技术，以构建高可靠性的网络系统。

7.4.1 链路级可靠性

链路级可靠性旨在实现链路级保护以及快速链路故障检测与切换。链路级可靠性的技术要求包括：

a) 支持M-LAG（Multi-chassis Link Aggregation Group）技术，包括：

1) 支持M-LAG，将不同设备上的端口聚合为一个逻辑接口。即使某台设备故障或聚合链路中的某条链路故障，聚合链路也不会完全失效，从而确保数据传输的可靠性；

2) 支持M-LAG主备模式，实现负载均衡流量转发和双链路备份保护。M-LAG设备可以配对，协商并管理M-LAG设备及M-LAG成员接口的主/备状态；

3) 支持通过M-LAG同步报文进行信息同步，如MAC地址表项、ARP表项、ND表项以及M-LAG成员接口的状态；

4) 支持M-LAG主备模式，实现双链路备份保护，仅主M-LAG设备发送和接收流量。M-LAG成员接口的主/备状态可根据接入设备发送的协议报文（如ARP、ND、IGMP、DHCP和MLD消息）进行选举，涵盖初始场景、故障切换场景和故障恢复场景；

5) 支持M-LAG在主备模式和主备模式下的快速故障切换和切换回；

6) 支持心跳检测和双主检测。双主检测故障不得影响M-LAG的正常运行。

7.4.2 设备级可靠性

大规模DCN网络中，光模块数量激增导致光模块故障概率增大，需要支持基于光模块的故障检测。同时当网络链路故障发生时，大规模DCN网络中的路由收敛时间无法满足高性能存储服务或高性能数据库访问服务的延迟要求。需要增强的故障快速恢复能力，支持亚毫秒级的收敛。设备级可靠性的技术要求包括：

a) 支持基于光模块的故障检测，包括：

1) 光模块脏污和松动检测；

2) 光通道级检测，在单通道故障时，端口状态保持UP，模型训练不中断；

b) 支持基于数据面故障感知快速收敛能力，包括：

1) 支持通过转发平面检测故障端口；

2) 支持在本地备份路径可用时，基于检测到的故障进行转发平面路径切换；

3) 支持将故障检测传播到远程网络；

4) 支持远程设备根据故障通知快速切换备份路径。

7.4.3 网络级可靠性

网络级可靠性指的是针对网络静默故障导致的业务会话级异常，能够实现快速故障识别和秒级故障收敛。网络级可靠性要求支持如下功能：

a) 支持静默故障检测，例如链路故障、异常转发表项、异常转发组件、物理端口处于Up状态但无法转发流量以及配置错误等；

b) 支持网络故障恢复，在识别到网络中的故障流后，重新进行哈希选路，从而实现流量的快速切换。

7.5 智能运维

智能运维是提升数据中心服务能力的关键。基于运维大数据平台和智能算法，智能运维构建了多种场景下的运维能力，如事前预测、事中监控、事后分析、按需扩容和及时应急响应，从而提升数据中心的服务能力。

7.5.1 网络数字地图

基于数字孪生技术，网络数字地图通过大数据平台形成数字副本，并支持基于数据的分析、计算和展示能力。

7.5.1.1 多维可视和分析

a) 支持多维数据可视能力，包括：

1) 支持网络实时数据采集，包括但不限于网络设备配置、网络业务配置、网络与服务状态及性能数据；

2) 支持多种类型数据的关联存储，并提供智能搜索功能；

3) 支持多维数据可视化及深入查看；

4) 支持历史数据的关联回放、基于指定时间段的变更趋势分析以及多维数据的深入可视化。

b) 支持多维数据分析能力，包括：

1) 支持统一收集、清洗运维数据，并将运维数据存储于统一平台；

2) 支持对多源数据进行关联分析，以辅助故障定位；

3) 支持对多源数据的关联分析结果进行模型训练，将运维数据转化为知识和洞察，为智能运维提供支持。

7.5.1.2 网络路径导航

支持网络路径导航能力，准确感知业务质量的变化，并识别网络质量的劣化，通过智能网络路径选择，为不同业务提供带宽、时延和可用性的差异化与确定性服务质量保障能力，包括：

a) 网络拓扑和状态感知，基于BGP-LS快速感知网络拓扑变化， 包括节点、链路故障，链路带 宽、时延变化等；

b) 业务SLA感知，基于iFIT的业务随流检测，通过Telemetry秒级上报机制，精确感知业务SLA，分层显示网络和业务质量；

c) SLA质差定界定位：业务质量劣化自动触发iFIT逐跳检测，基于业务转发路径发现质差点，并可与网络拓扑结合，直观可视定位定界结果；

d) SLA质差恢复：基于业务的SLA质差定位结果，重新计算网络路径，使用SR-TE、 SR-Policy等技术对网络路径进行重优化，引导流量避开质差点，使得的业务SLA 持续得到保障。

7.5.1.3 存储主机即插即用

支持存储主机即插即用，对接入主机实施快速管控，并智能的调整智能无损网络的相关配置以达到低时延、无丢包和高吞吐的性能。具体能力要求包括：

a) 支持主机接入和离开感知，并支持主机接入和离开的信息在网络内的扩散；

b) 支持针对主机接入和离开场景智能调整智能无损网络的相关配置，以支持主机的有效传输；

c) 支持网络设备因PFC死锁、CRC错误报文达到告警阈值等问题触发接口Error-Down后，向网络其他设备扩散该接口Error-Down信息，接受该信息设备及时调整路径信息。

7.5.2 网络故障定界定位

7.5.2.1 故障感知和预测

快速故障检测是网络运维能力的基础。故障检测要求对网络状态和服务状态进行实时检测、上报和集中分析，以提高实时性和准确性。故障检测的技术要求包括：

a) 支持秒级数据采集，如使用Telemetry技术秒级采集业务、设备、链路、端口、光模块等多类指标数据，实时监测网络设备运行状况；

b) 支持通过端口镜像捕获网络流量数据，从传统、虚拟、云和容器环境中实时获取完整的服务流数据，解析并构建实时统一的服务视图。直观展示应用服务的业务逻辑、依赖关系、服务运行质量及告警信息，提供自动故障分析功能，显示每个业务的服务路径并区分各应用节点的时间消耗，发现性能瓶颈并在整个服务提供过程中进行运维保障；

c) 支持智能计算场景下的无损转发。少量数据包会大幅降低总通信量并减慢整个集群的训练速度，因此智能计算网络必须能够以超过1‰的精度检测NPU与GPU间通信流量的丢包率，并能够识别丢包位置，以支持智能计算网络在丢包发生时能及时介入，防止单点丢包拖慢整个智能计算集群。

7.5.2.2 故障定界定位

故障定界与定位是指基于网络和业务劣化及故障情况，分析和确认故障位置及原因。故障定界与定位的技术要求包括：

a) 支持基于网络状态数据、告警数据、性能数据、日志数据等的多个网络指标进行关联分析，如基于知识图谱或AI能力提前识别网络可靠性、容量、性能、稳定性等隐患，统一评估全网潜在风险，降低故障发生概率；

b) 支持光模块脏污检测，解决光纤端面如果进灰或脏污容易导致链路闪断，从而影响集群持续训练或导致网络通信效率降低的问题；

c) 支持光模块松动检测，解决在集群经过长时间运行之后，光模块松动的自动检测能力；

d) 支持网络丢包原因定位：智算网络出现丢包时要根据丢弃原因快速闭环以便恢复训练业务，因此智算网络需要能直接查询、定位因端口拥塞、ACL 丢弃、路由查表失败等原因导致的丢包计数；

e) 支持RDMA 通信性能监控：AI 训练过程中卡间通信通常为毫秒级的突发流量，通过端口带宽利用率无法有效监控 RDMA 流级 (IP 对) 通信性能，因此智算网络需要能够监控 RDMA 通信的流完成时间、流有效吞吐等指标；

f) 支持网络拥塞监控：网络中存在拥塞时会影响集合通信性能，但是端口 / 队列的 PFC 报文计数受 xon/xoff 水线、流量模型等影响无法量化评估网络拥塞程度，因此智算网络需要具备端口 / 队列反压时长 / 反压 Pause 监控和统计能力，量化拥塞程度。

7.5.2.3 无损升级

支持设备的无损升级，以减少业务中断时间。无损升级的技术要求包括：

a) 支持基于M-LAG等技术的设备升级，通过备份设备的独立升级，提高升级效率并实现秒级业务中断

7.5.3 应用故障定界定位

7.5.3.1 应用故障感知

应支持应用故障感知，感知业务真实的 SLA。应用故障感知的技术要求包括：

a) 支持应用级随流检测，如IFIT随流检测技术，通过在业务报文中插入检测参数，并通过分布式数据上送和集中式数据分析，感知业务真实的SLA；
b) 支持基于IFIT的端到端和逐跳业务质量检测。

7.5.3.2 应用故障定界定位

应用故障定界定位的技术要求包括：

a) 支持应用级故障自动定界定位，如通过IFIT逐跳检测，确认业务劣化的具体网元和链路；

b) 支持智算POD中对AI训练或推理任务的监控及故障自动化分析排障功能

7.6 安全

7.6.1 设备层安全

设备层安全技术要求包括：

a) 支持出厂可信。出厂软件包不包含硬编码的认证凭证，需提供内部账户和通信矩阵提供了清单，且该清单需公开透明且安全无后门；

b) 支持开局安全，性基于证书和签名等安全技术实施SZTP（安全零接触配置），确保设备的安全网络接入；

c) 支持基于主机的入侵防御，通过监控本地系统是否被入侵或感染。一旦检测到疑似入侵或感染事件，系统会发送日志提示管理员隔离并保护系统，防止进一步入侵甚至危及其他设备的安全；

d) 支持退网安全，所有需要从网络中移除的设备只有在通过安全检查后才能被释放。

7.6.2 网络层安全

网络层安全技术要求包括：

a) 支持协议安全。所有网络协议默认采用高强度安全加密算法，与协议中的不安全算法兼容并提示风险；

b) 支持链路安全。支持基于硬件的MacSEC加密，防止数据在传输过程中被窃取或篡改，确保数据完整性。同时应支持加密数据包线路速度转发，实现加密数据的安全传输。

7.6.3 管控层安全

管控层安全技术要求包括：

a) 支持安全配置检查，如不安全的协议和算法、异常端口、账号和密码策略、密码存储模式等，以便及时发现不安全的配置风险，并提前在生产网络中降低风险；

b) 支持管理安全。通过首次登录、身份认证、最小权限原则、外部入侵检测和数据加密等措施构建安全管理系统，防止身份冒充、否认、篡改、信息泄露、拒绝服务（DoS）和权限提升等安全攻击。支持日志安全。为安全日志提供独立的存储空间，防止日志被篡改

8 数据中心跨 DC 网络关键技术要求

在多 DC 场景中，数据中心除要满足单 DC 内的技术要求外，还要满足跨 DC 间的关键技术要求。

8.1 跨 DC 可靠性

8.1.1 同城双活 DC 可靠性

在同城双活数据中心场景中，应部署并管理两个数据中心为双活模式。同城双活数据中心的技术要求包括：

a) 支持数据中心（DC）的主备状态检测。正常情况下，两个数据中心均处理服务。若主备检测失败，服务不受影响；

b) 当单个数据中心发生故障时，支持快速数据中心切换，以减少服务中断时间；

c) 当单个数据中心从故障中恢复时，支持快速的主备数据中心恢复及服务回切；

d) 支持同城主备场景下的数据备份，要求链路往返时间（RTT）和带宽满足相应需求。

8.1.2 异地容灾 DC 可靠性

异地冗余数据中心（Geo-redundant DCs）指的是在多个不同区域部署数据中心，以进一步提升数据中心的容灾能力。异地蓉欧能够在数据中心的技术要求包括：

a) 支持通过两个主用站点和一个容灾站点实现地理冗余。除了市内的主用数据中心外，还增加了远程数据中心备份；

b) 支持数据中心间的心跳检测。正常情况下，所有数据中心各自履行职责。如果心跳检测失败，服务不受影响；

c) 支持异地冗余数据中心的切换和回切机制。当其中一个主用数据中心发生故障时，快速进行主用数据中心切换。当市内的主用数据中心同时发生故障时，服务应切换到备份数据中心。当数据中心故障恢复后，市内的主用数据中心应切换回原状态；

d) 支持数据中心间的数据备份。链路往返时间（RTT）和带宽应满足要求。

8.2 跨 DC 高性能

跨DC业务要求长距离无损传输，网络需要支持增强长距离PFC技术，包括：

a) 支持依靠短周期、高频率、持续少量调节流量发送与暂停的机制，在缓冲区大小和带宽固定的情况下，实现与PFC相比的长距离无损传输；

b) 支持周期性扫描接口优先级队列的缓存占用情况，并通过向上游设备发送PFC反压帧的方式来控制每个周期内需要上游设备停止发送流量的时长，以持续调整流量的发送与暂停。

8.3 跨 DC 运维

跨 DC 运维的技术要求包括：

a) 同城双活场景支持一个控制器管理同城内的双活数据中心；

b) 异地容灾场景支持多数据中心管理。每个数据中心可由独立的控制器管理，跨数据中心管理通过跨数据中心控制器实现；

c) 支持跨数据中心网络安全业务自动部署，以保持安全策略的一致性并减少配置错误；

d) 支持跨数据中心故障定界与定位，并支持以可视化方式呈现端到端故障定位结果，从而缩短故障定界与定位时间；

e) 支持跨数据中心配置仿真，以提升跨数据中心业务部署的可靠性。

8.4 跨 DC 安全

跨 DC 安全技术要求包括：

a) 支持数据中心之间的数据安全。应使用如MacSEC等技术进行数据加密；

b) 支持不同数据中心控制器之间的管理安全。应使用如SSH、TLS等技术进行管理通道加密。