AD-Campus 解决方案

大中型园区网络设计指南(VXLAN 园区)

资料版本：5W102-20231124

目　录

摘要

本文档主要针对大中型园区场景，从 AD-Campus 解决方案架构、产品选型建议、网络资源规划、网络部署方案设计、智能运维等方面描述了方案的部署思路，指导用户进行网络部署方案的 HLD（High Level Design，概要设计）和 LLD（Low Level Design，详细设计）。

本设计指南作为园区网络解决方案设计阶段的参考资料，已经过 H3C 工程师充分验证，帮助客户快速构建一张极简、智能、可信的园区网络。

1 概述

1.1 大中型园区网络介绍

在企业或者组织内部部署的网络称为园区网，按组织的对网络承载能力的不同要求，可以将园区网络划分成三大类型：

- 大中型园区网络，通常指具备 2000 终端以上接入管理能力的园区网络。
- 中小型园区网络，通常指承载 2000 以下终端的园区网络。
- 总部-分支园区网络，通常指除了总部具有一定规模的园区网络外，在不同地域存在多个办公分支机构，每个分支机构网络可看做一个单园区网络，但规模相对较小，均要与总部有一定量的通信需求。

本文档针对大中型园区网络提供网络设计指导。

1.2 大中型园区网络的问题和挑战

现代 IT 系统由三部分组成：“云”、“管”和“端”。“云”通常指远端的公有云/私有云数据中心，是企业应用的承载中心；“端”就是需要接入网络，使用云服务的各种固定和移动终端；“管”就是连接“端”和“云”的传送通道，主要包括广域网和园区网两部分。这三者之间相互影响，相互适应。对于园区网来说，作为连接“端”和“云”的中间管道，经常受“端”和“云”两方面的影响，因为和终端的联系最紧密，因此受“端”的影响更大一些。

包括 Gartner 在内的主流研究机构都发布了对于“端”的趋势分析和预测。总结为两点：

(1) 终端的移动化和无线化成为趋势。每个人都有不止一个移动终端，移动互联越来越深入到人们的生活中。在企业中，BYOD 应用越来越广泛，移动办公成为主流。

(2) IOE 万物互联成为趋势。企业园区的物联网会带来百亿级别的智能设备接入。

“端”的变化对园区网带来的影响和冲击，有如下几点：

1. 移动化的冲击

- 策略如何跟随，体验如何保持不变？除了无线漫游，企业中还存在其他方式的移动，比如员工从办公位跑到会议室里边开会，需要共享胶片和资源，相关的权限需要继续保持；员工从总部出差到分支希望一些重要的业务权限不变；公司搬家后希望即插即用，业务不中断等等。但从目前网络现状看，很难做到或者说 IT 需要花费很大的代价进行网络配置调整，用户的体验也不够好。比如，某公司一部分员工分流搬家到新的办公大楼之后，预先做了大量的工作进行网

段重划和配置调整以保证业务连续性，但不幸的是搬家后 IP 电话/打印机等必须的办公设备依旧长达近两周时间不能使用，对正常办公影响较大。

- 如何方便的进行用户审计？传统网络状态下，用户移动，IP 地址发生变化，当审计的时候，由于用户的 IP 地址不停变化，需要结合不同时间段内用户的 IP 地址情况综合去查，查询虽然可以实现，但是比较麻烦，达不到所见即所得的状态，即见 IP 地址即见用户，单独审计 IP 的效果。

2. 无线化的冲击

企业中无线技术出现之后，网络中始终存在着有线和无线两张不同的网络，虽然无线接入逐渐占据主流，但是有线无线混跑的局面在未来一段时间还将长期存在，那么客户会面临有线无线两张网络的割裂问题，管理割裂：有线无线两组不同的人来同时管理网络；数据转发割裂：有线走交换机转发，无线走 capwap 的 AC 集中式转发；策略执行点割裂：有线的执行点在交换机上，无线的策略执行点在 AC 上。这样网络的运维往往需要两套人马，带来了成本的上升和管理复杂问题。虽然业界对于有线无线融合已经有呼声，并且也都提出了一些方案，但总是存在这样那样的问题，不能达到满意的效果。

3. IOE 的冲击

- IOE 带来了终端接入数量的爆炸式增长， IOE 接入园区网络的方式主要有两种，一种是直接接入园区网，比如监控摄像头，移动 POS 终端等；另外一种是通过物联网关接入，比如楼宇自动化系统中，使用物联网关路由器，一侧使用 ethernet 端口接园区网络，网关其他端口接各种现场控制总线。IOE 终端接入园区网络之后，往往和园区网现有业务有不同的属性，需要进行安全隔离和 QOS 保证；同时对于直接接入园区网的大量物联网终端，配置管理工作量较大，如何实现简单运维，快速部署也是园区网需要考虑的课题。
- 那么当前园区网存在的问题总结为两个词：僵化和复杂。
- 僵化就是不灵活，网络往往与物理位置紧耦合。网管人员往往根据地理位置，楼栋，划分若干个 L3 网段，终端基本不能大范围移动；如果要移动，策略不能自动跟随，体验难以保持不变。
- 复杂就是网络运维工作量巨大，网管人员 80%-90%的时间都陷在网络运维的泥潭里不能自拔，只有 10%-20%的时间才能用于创新。

1.3 AD-Campus解决方案概述

为了解决园区网存在的上述问题，H3C 提出了 AD-Campus 新园区架构。

1.3.1 方案介绍

H3C 应用驱动园区网解决方案（AD-Campus）创新地引入了云原生架构，既实现了网络控制、编排、管理的入口统一，又实现了园区、数据中心和广域网场景的融合，同时引入了分析组件，通过精细化的数据采集及大数据、AI 分析，为园区网络带来智能运维能力。在云原生架构之上，结合 SDN+VXLAN 的技术，通过构建基于 VXLAN 的新一代柔性园区基础网络，配合软件定义的相关技术，颠覆传统的园区网“人适应网”的现状，实现整个园区网范围内“网随人动”的效果，在不需要做任何网络配置调整、增加运维复杂度的基础上，让用户和终端可以在整个企业园区的任意角落移动，保持用户和终端始终处于既定的隔离网络、延续既定的网络策略，从而大大降低了园区运维的复杂度，满足智能化、移动化和物联网建设背景下对于园区网络的新诉求。

AD-Campus 方案的特征包括：

1. 极简

极简：指网络管理大幅简化，真正将网管员从低价值劳动中解放出来。AD-Campus 的目标是消灭命令行，从 controller 一点进入管理网络。极简包括如下四个方面：

(1) 网随人动

- 网随人动 2.0 实现用户管控极简，用户在园区内移动，地址不变，权限不变，位置和地址解耦，网络管理员零干预。
- 策略随行，指用户移动到哪里，策略就跟随到哪里，用户的体验不变，比如研发的员工从总部出差到各个办事处，依旧能获得研发的权限，访问研发的相关资源，和在总部一样。
- 新名址绑定。名址绑定就是用户/业务和 IP 地址绑定，IP 地址不光从技术层面承载连通性，支撑路由转发，而且还具有了直接标识业务/用户的功能，策略可以直接根据 5 元组来实施，大大简化了传统网络策略的部署。最终达到所见即所得的效果：IP 即用户，网段即业务。同时也大大简化了网络的审计。新名址绑定在上述能力基础上，实现复杂场景用户与地址的自动绑定，如公共机、基于 5W1H 场景接入策略变化，均可实现自动绑定和解绑。

(2) 智慧物联实现终端接入极简。终端接入网络，自动分拣，映射到对应虚拟专网，上线即隔离，高效安全，同时有鹰视探测保驾护航。

(3) 极简自动化

- 三份配置打天下：全网网络设备只需维护 3 份配置，简化业务部署。
- 弹性扩容：设备开箱上电后自动加载版本，加载业务策略配置，网络管理员零干预启动，实现即插即用。
- 故障替换即插即用：设备故障后，采用同型/异型设备替换，即插即用，策略自动复制。
- Qos 业务全网自动部署：基于应用分类定义业务的保障策略，三步完成全网 Qos 部署。
- 组播极简部署：图形化页面配置整网组播能力，支持按需复制减少冗余流量。

(4) 有线无线深度统一，将有线无线两张割裂的网络进行深度融合，一套网管，一种数据平面，一套策略。

(5) 业务按需交付：业务指 4-7 层服务，如防火墙，IPS，ACG 等，这里引入了服务链概念，把园区传统的通过策略路由方式的复杂引流策略转换为一种简单的按需使用，自由编排的引流方式来快速实现。

2. 智能

先知保障，将 AI（Artificial Intelligence，人工智能）应用于运维领域，基于已有的运维数据（日志、监控信息、应用信息等），借助大数据和 AI 技术，通过机器学习和深度学习算法，从应用的视角来观察网络，主动感知网络和应用存在的问题。并针对业务问题提供自动化排障能力，帮助用户快速进行故障定界和恢复，实现降低运维成本，提高企业产品的竞争力的目标。

3. 融合

- 园数 WAN 融合：实现多域场景的统一部署，统一策略。
- 管控析融合：管理、控制和分析组件融合部署，无缝连接，数据共享。
- 多园区融合：分权分域保证了多园区的管理权责清晰，安全又灵活。
- 网安融合：安全产品控制组件融合部署，业务部署和策略下发实现联动。
- PON 融合：PON 网络自动化部署，拓扑统一呈现。

- 开放融合：提供软件定义能力，通过 controller 上层 API 接口开放，允许第三方进行增值开发，快速融合新的应用。

4. 可信

- 出口安全
- 终端行为管理
- 终端数据管理
- 终端合规检查-鹰视

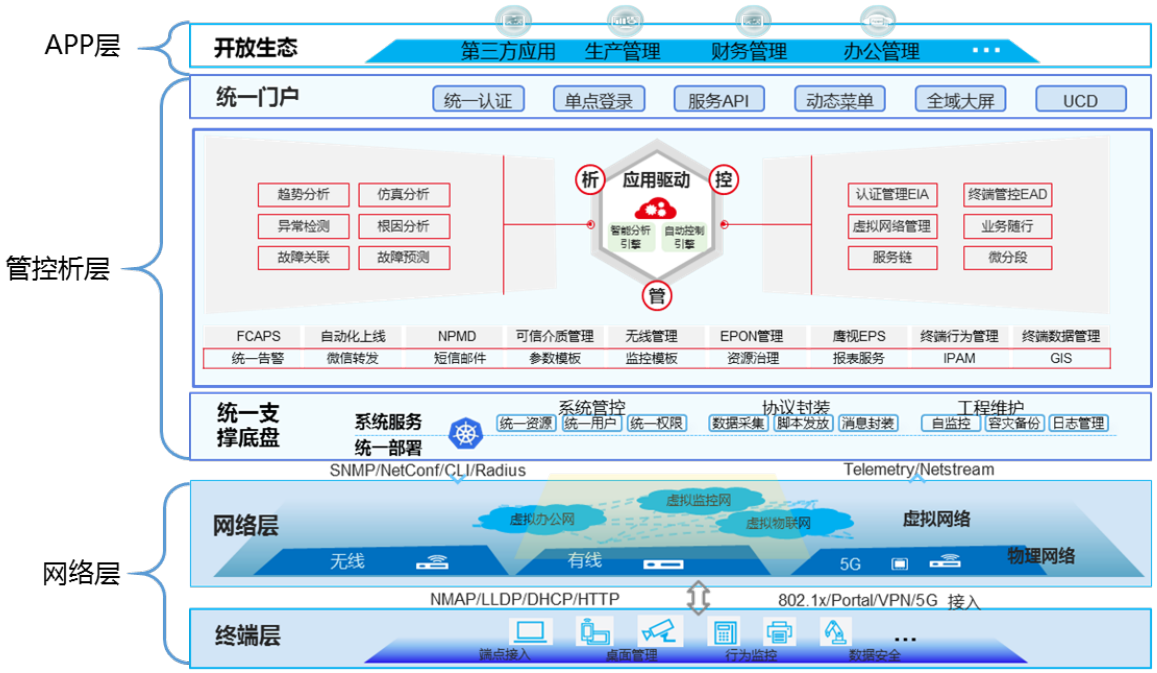
5. 超宽

- 25G/100G/400G 交换机
- WIFI6

1.3.2 方案架构

AD-Campus 方案总体架构如下图，划分三个层面：APP 层、管控析层和网络层。

图1 方案架构图



- APP 层

  AD-Campus 解决方案基于其管控析层的各组件可以提供标准化北向 API 接口，将网络的管理、控制和分析能力开放给第三方应用，使第三方应用可以根据自身业务的需要定制自己所需要的业务应用，快速满足在不同行业领域的业务需求。

- 管控析层

  管控析层是 AD-Campus 解决方案的主体部分，为网络提供自动化部署、业务管理、用户接入控制、智能运维、安全威胁防护等能力。对比传统园区网络使用网管系统进行管理和运维的方式，AD-Campus 解决方案使用 SeerEngine-Campus 实现网络的自动纳管和业务自动部署，更加灵活，效率更高。

SeerEngine-Campus 通过统一的操作界面，实现管理员意图到设备配置命令的业务编排，实现从抽象网络模型到设备和端口实体的映射，使管理员面对的不再是一台台网络设备和一条条命令行，而是完整的网络运行能力。

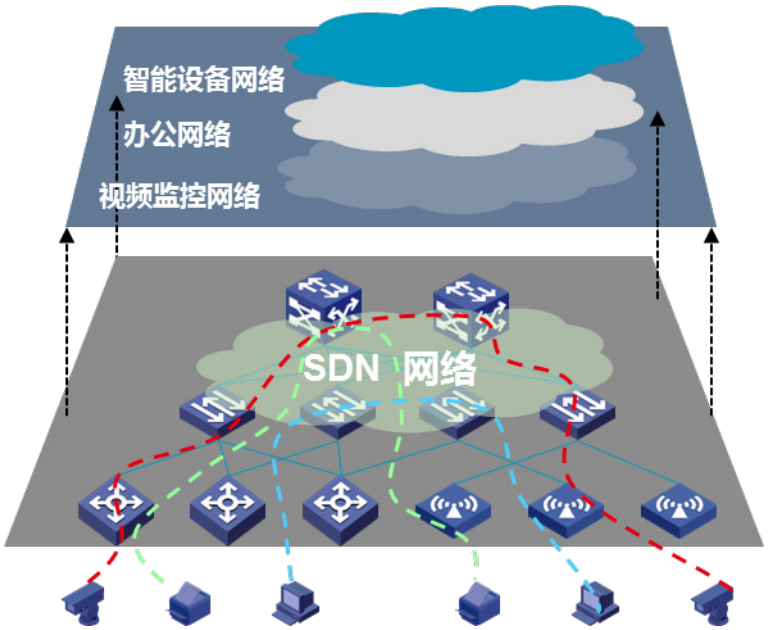
同时统一门户和统一支撑底盘使多个容器化组件以整体的方式呈现，单点登录，数据共享，为管理员应对超大网络和多园区管理提供了完整的端到端可视和运维体验。

- 网络层

  引入 EVPN 和 VXLAN 技术，网络层划分为物理网络和虚拟专网两层。其中终端本身不具有虚拟化能力，也不要求一定具有明显业务特征，仅在接入物理网络时，才会映射到对应的虚拟网络中，成为虚拟网络中的端设备。

  - 物理网络，也称为 underlay 网络，为园区网络提供基本的连接服务和转发能力。
  - 虚拟专网，也称为 overlay 网络，通过 VxLAN 技术在物理网络上构建多个相互隔离的虚拟专网，业务访问策略部署在虚拟专网上，服务于不同的业务需求。

  图2 虚拟专网图



1.3.3 方案涉及组件

方案涉及的组件在网络层和管控析层。

网络层组件主要是交换机、无线控制器（无线 AC）、无线接入点（无线 AP）、路由器和防火墙。

- 交换机，负责有线二三层数据报文转发，主要使用的是园区核心交换机和园区接入交换机。
- 无线控制器和无线接入点，负责无线服务发布、无线用户接入、无线报文转发等，无线控制器主要使用的是框式和盒式无线控制器，也有交换机无线插卡，无线接入点主要使用的是 fitAP。
- 路由器，负责出口路由和报文转发。
- 防火墙，负责园区网络安全，主要在出口或网络边界，提供访问控制和病毒防御。

管控析层较常用组件是 SeerEngine-Campus、SeerAnalyser-Campus、EIA/EAD、WSM、EPS、SMP。

- SeerEngine-Campus，是园区网控制组件，主要负责园区设备自动化上线和网络管理业务编排，同时将园区安全组推送给 EIA（园区准入 AAA）。SeerEngine-Campus 提供开放接口，支持与第三方平台对接。
- SeerAnalyser-Campus，是园区智能运维分析组件，主要负责园区设备的运维数据可视化，大数据分析和故障根因分析。
- EIA/EAD，是园区准入 AAA 认证服务和域控服务，主要负责用户准入控制，同时根据用户分组提供授权。
- WSM，是无线配置管理组件，主要负责无线网络的可视化配置编排。
- EPS，是鹰视系统，主要负责园区终端系统的安全防护，定期扫描终端状态，防 MAC 仿冒终端。
- SMP，是安全产品控制组件，主要负责防火墙产品的业务编排。

以上主要组件之间，均根据业务功能需要有一定业务联动，实现一个操作流程完成多组件配置，一个页面查看多种相关性数据。

1.3.4 方案价值

AD-Campus 方案的价值：

- 实用性和先进性

  采用先进成熟的技术满足办公网络各种应用系统的需求，兼顾其他相关的管理需求，在满足各种应用系统的同时，又体现出网络系统的先进性。

- 高可靠性

  网络必须具有高可靠性，以避免单点故障。在网络设计中特别是关键节点的设计中，选用电信级的网络设备，并采用网络冗余结构，设定可靠的网络备份策略，保证网络具有故障自愈的能力，最大限度地保证办公网络系统的高效运行。

- 标准性与开放性

  基于标准，坚持统一规范的原则，采用国际标准协议实现在网设备和新增设备的互连互通，真正实现开放，从而为未来的发展奠定基础。采用国际上通用的网络标准协议，不仅保证与其它网络(如公共数据网、Internet)之间的平滑连接和互通，还能适应未来若干年的网络发展趋势，便于将来网络自身的扩展。

- 高安全性

  安全体系应该是一个多层次、多方面的结构，需要从全方位四个层次进行安全防范：网络层安全、应用层安全、系统层安全和管理层安全。

- 高性能

  网络系统的性能是整个信息系统良好运行的基础，设计中必须保障网络及设备的高吞吐能力，保证各种信息（数据、语音、视频）的高质量传输，力争实现透明网络，网络不能成为业务数据交换的瓶颈。

- 灵活性及可扩展性

  网络是一个不断发展的系统，不仅需要保持对以前技术的兼容性，还必须具有良好的灵活性和可扩展性，具备支持多种应用系统的能力，提供技术升级、设备更新的灵活性，能够根据业务不断深入发展的需要，根据未来系统的增长和变化，平滑的扩充和升级现有的网络覆盖范围、扩大网络容量和提高网络的各层次节点的功能，最大程度的减少对网络架构和现有设备的调整。

- 易操作性和可管理性

  随着业务的不断发展，网络管理的任务必定会日益繁重。所以在网络设计中，必须建立一套全面的网络管理解决方案。网络设备必须采用智能化，可管理的设备，同时采用先进的网络管理软件，实现先进的管理。最终能够实现监控、监测整个网络的运行情况，合理分配网络资源、动态配置网络负载、可以迅速确定网络故障等。通过先进的管理策略、管理工具提高网络的运行性能、可靠性，简化网络的维护工作，从而为办公、管理提供最有力的保障。

2 网络架构设计

SDN 网络包括有线部分和无线部分，有线部分由接入，汇聚，核心三层设备组成，搭配园区网控制组件：SeerEngine-Campus。无线部分由无线 AC、无线 AP 组成，可选搭配 WSM。无线网络相对独立，单独介绍。

整体有线网络架构如下：

(1) 接入到汇聚使用 VLAN 进行联通，在汇聚层设备上，不同 VLAN 映射到不同 Vxlan 进行隔离，同时汇聚和核心设备之间运行 Vxlan 构建 overlay 网络，构建一个逻辑上的大二层网络，同时采用分布式 L3 网关并通过可靠的机制有效地抑制广播风暴。

(2) 策略管理上采用了面向业务的分组模式，将属性或者访问权限相近的用户分到一个安全组中，同时也将服务器侧的资源划分到安全组进行统一管理。策略定义时，基于矩阵表格的方式简单直观。具体策略可简单可复杂，实现各种高级复杂的策略控制功能。

(3) 基于 5W1H 的灵活的用户认证接入机制，根据 who（谁），whose（谁的设备），what（什么设备），when（什么时间），where（什么地点），how（什么方式）多个维度覆盖各种接入场景。用户可根据自己的需求，灵活定制场景，满足自己个性化的需求。

(4) 支持用户终端在整个生命周期中 mac 和 IP 的强绑定，终端不管移动到哪里，可以做到终端始终绑定唯一固定的 IP，满足某些企业强安全需求。

(5) 整网的核心是园区网控制组件：SeerEngine-Campus。所有对网络的自动化上线，接入管理，用户组/策略管理，业务配置管理全部在 SeerEngine-Campus 上通过直观的图形化界面完成。SeerEngine-Campus 将管理员的操作在后台转化为网络设备的具体命令进行下发给设备执行。

在此架构下，为了实现分层设计分层部署，需要将 Underlay 和 Overlay 的部分分开进行设计。

2.1 Underlay网络架构设计

2.1.1 物理网络架构设计

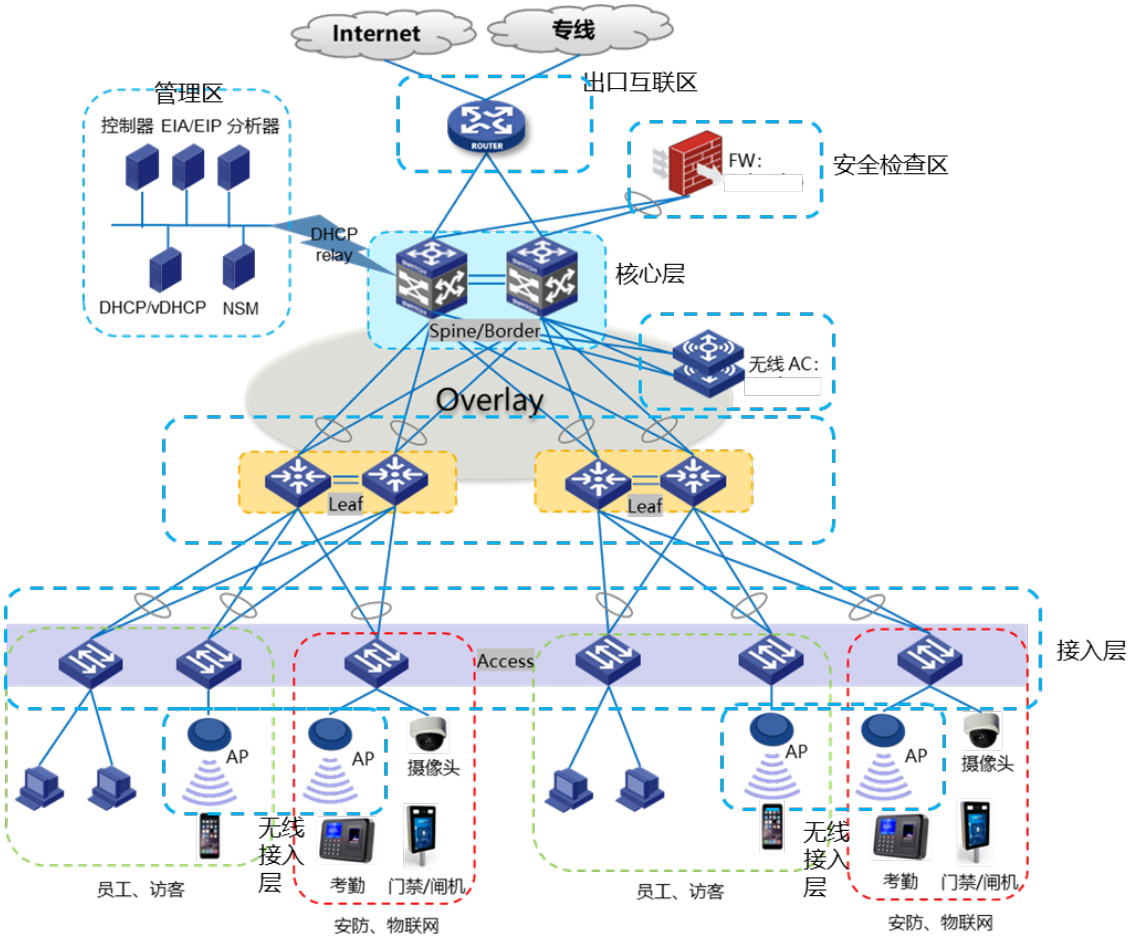
园区网络系统是一张物理网，一般采取树形结构部署，分为核心层、汇聚层、接入层，以及无线管理区、网络管理区、出口互联区、安全检查区。其中核心层、汇聚层和网络管理区是最重要的部分，设计中需要重点关注。特定的扁平化组网需求中，核心层和汇聚层可以合一。

各区/层有如下定义和作用：

- 核心层（Spine 层）：是园区数据交互的核心，连接园区各个部分，负责路由反射、数据高速转发等，通常要部署性能高、稳定性好的交换机，多采用框式中高端交换机。

- 汇聚层（Leaf 层）：是园区用户的分布式网关，负责用户接入、东西向流量转发、南北向流量转发。通常在部署的时候要兼顾成本和性能，根据用户数选用满足成本要求的中低端盒式交换机。在一些对性能和稳定性要求高的应用场景，也可以选中端框式交换机。
- 接入层（Access 层）：负责园区有线用户接入和无线 AP 接入，通常选用中低端盒式交换机，端口数量多，不需要支持太多三层功能。接入层区分有线接入层和无线接入层，无线接入层选用的交换机需要支持 POE 供电，成本会较普通交换机，所以有线接入层和无线接入层一般都会分开部署，避免浪费 POE 端口。接入层交换机支持通过级联扩展接入端口数量，但是为了支持自动化，对接入层数有一定限制，一般不超过三层。
- 无线管理区：是园区内部署无线控制器（后续简称无线 AC）的区域。大中型园区网络推荐使用独立 AC 部署，旁挂核心交换机，根据需要管理的无线 AP 数量和无线用户数量，可以选择一组或多组无线 AC 来进行管理。
- 网络管理区：用于部署网络管理服务器的区域，如网管系统，AAA 认证服务器，SDN 控制组件等。网络管理区与核心区之间需要采用三层交换机连接，确保 IP 可达。网络管理区的服务器不允许直连核心区交换机。
- 出口互联区：是园区内部网络的边界，一般需要部署路由器，路由器负责园区外部 WAN 网、专网或 Internet 与园区内部网络的互通，必要的时候增加防火墙进行防病毒检测等安全防护。如果园区路由器使用 BRAS，则 BRAS 还需要负责园区内部用户的准出认证和计费统计。
- 安全检查区：主要用于部署防火墙等安全检测设备，旁挂核心交换机，可以对南北向和东西向流量进行安全检测。

图3 园区网络物理架构图



组网描述：

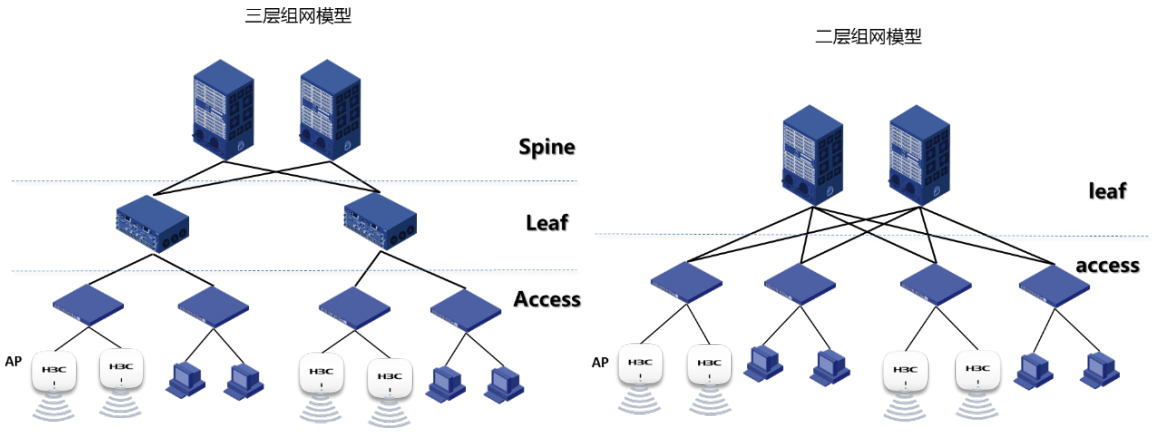
- 核心层交换机和汇聚层交换机采用 IRF2 堆叠或 MLAG 结构。核心交换机和汇聚交换机之间交叉互联，保证设备冗余和链路冗余，避免单点故障。核心交换机和汇聚交换机之间为标准 2 级 Clos 架构，采用 EVPN 协议构建 Overlay 逻辑组网。
- 接入层交换机双线双归属到对应的汇聚交换机组中，同时接入交换机还可以进行多级级联，以满足不同场景下的特殊需求。
- 多速率 PoE 接入层交换机支持 5G/2.5G/1G,可满足大功率高速 WiFi6 AP 的接入。
- 无线网络为集中式转发模型，无线控制器旁挂在 Spine 上，无线控制器完成无线终端的认证，无线终端的网关落在 Spine 交换机上。
- 防火墙旁挂在 Spine 交换机上，采用聚合连接到两台 Spine 上，单臂模式，提供跨 VRF 和内部访问外部网络的安全控制。
- 服务器区提供设备自动化上线、业务部署、认证、管理、运维服务，与 Spine 三层互通，可部署在本地园区，也可部署到远端 IDC，通过专线访问。

2.1.2 内部网络分层设计

园区内部网络主要是核心层、汇聚层和接入层。标准网络架构有三层组网和二层组网两种形式。

- 三层组网模型就是核心层（Spine）、汇聚层（Leaf）和接入层（Access）都有，大型复杂网络必须采用这种方式，通过多个 Leaf 做分布式网关，分担用户的压力，并可以在分布式网关上应用复杂的访问策略。
- 二层组网模型是核心层和汇聚层合一，Leaf 做核心，采用集中式网关，下面均是接入。这种网络一般有扁平化管理的需求，组网相对简单，但也意味着所有流量都要到核心来进行转发，下面接入层设备无法实现复杂的访问策略控制。

图4 组网模型



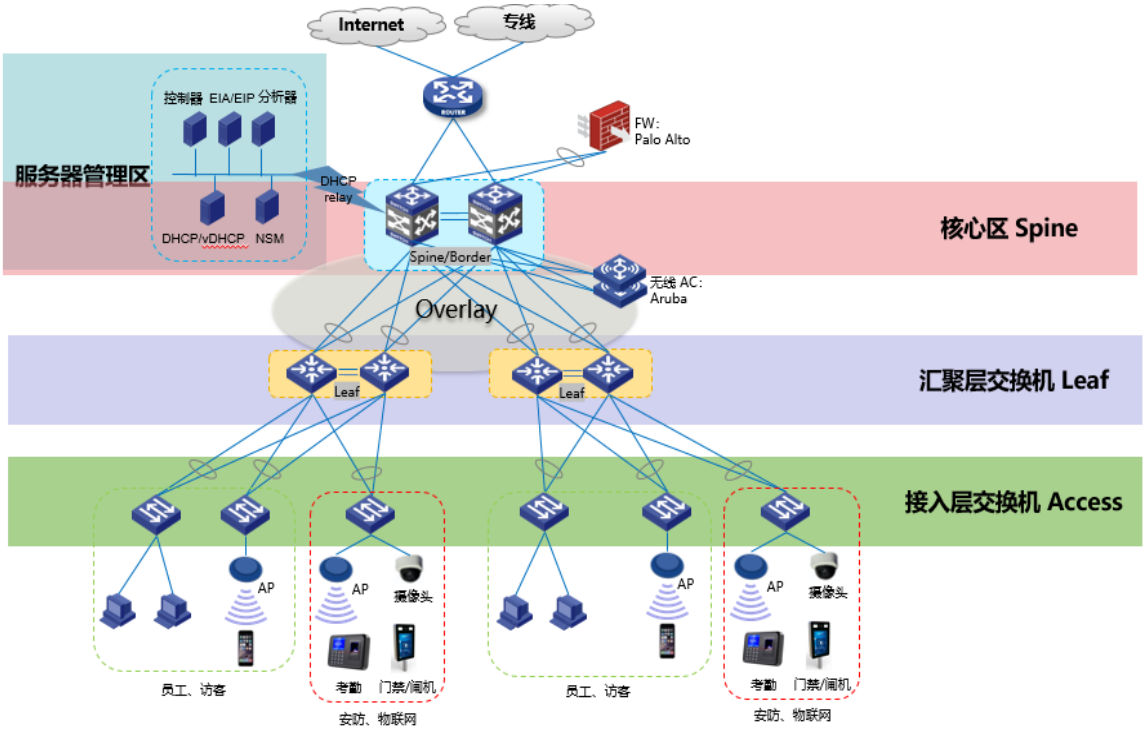
2.2 Overlay网络架构设计

面对承载业务系统的多样性，为了能够在网络层灵活支持业务系统的发展，精细化隔离不同系统及数据，使得网络更加智能、易管理，就需要打破传统的设计理念，在传统物理网络架构基础上构建一套以应用驱动为主的新网络，AD-Campus 解决方案就提供了这样一套新形态的网络平台。通过 SDN 控制组件除了可以实现自动化部署之外，还可以实现用户、终端的资源分配、用户组的策略定义，并且在定义资源分配和用户组策略的过程中采用向导式的配置模式以及可视化的策略定义界面，实现网络的智能化。

2.2.1 逻辑网络架构设计

基于传统物理网络架构基础上的逻辑网络拓扑如下：

图5 逻辑网络拓扑



如上图所示，整体网络物理架构由核心层、汇聚层以及接入层设备构成，不同点是在核心层与汇聚层设备上启用 Overlay 技术，同时在内部服务器区部署 SDN 控制组件，并由 SDN 控制组件控制整个网络的运行。具体设计如下。

- 核心层和汇聚层设备之间构建 overlay 网络，构建一个无状态网络，同时采用分布式 L3 网关并通过可靠的机制有效地抑制广播风暴，接入层设备采用动态 VLAN 接入，汇聚层再完成 VLAN 到 VxLAN 的映射。
- 策略管理上采用了面向用户的分组模式,将属性相同的设备或者访问权限相近的用户分到一个策略组中，同时也将服务器侧的资源划分到相应的策略组进行统一管理。
- 采用认证系统，根据用户登录的用户名或设备的 MAC 地址与 IP 地址绑定，实现用户或设备不管到任何地方，其 IP 地址不变，从而使得其的安全策略也不便，方便管理运维。
- 方案的核心是 SDN 网控制组件组件：SeerEngine-Campus。所有对网络的自动化上线，接入管理，用户组/策略管理，业务配置管理，网络运维管理全部在 SeerEngine-Campus 上通过直观的图形化界面完成。SeerEngine-Campus 将管理员的操作在后台转化为网络设备的具体命令进行下发给设备执行。
- 服务器管理区，SeerEngine-Campus /DHCP 服务器和网络设备之间三层互联，Spine 设备的上行链路配置 port trunk permit VLAN1 4094， EIA 三层接入；
- Spine/Border 合设，Spine 设备主要作为路由反射器 RR 设备。Border 与 Spine 合设，与服务器之间的互通；Spine 设备与 Leaf 设备之间连接的链路为 underlay 链路，配置 Spine 和 Leaf 设备之间路由可达；
- Leaf 下行口作为用户上线认证点

- Leaf 与 Access 设备连接的链路为 Leaf 下行接口，Leaf 下行接口配置为认证接口，用于用户认证；
- 当用户上线时，Leaf 设备通过“下行接口+VLAN ID”来识别不同的 Access 接口，并且根据不同的登录帐号进入到不同的安全组内；
- 认证用户通过 DHCP Relay 在 option82 中携带不同的安全组信息，向 DHCP Server 申请分配 ip 地址；
- DHCP Server 识别安全组信息，分配对应的 ip 地址；

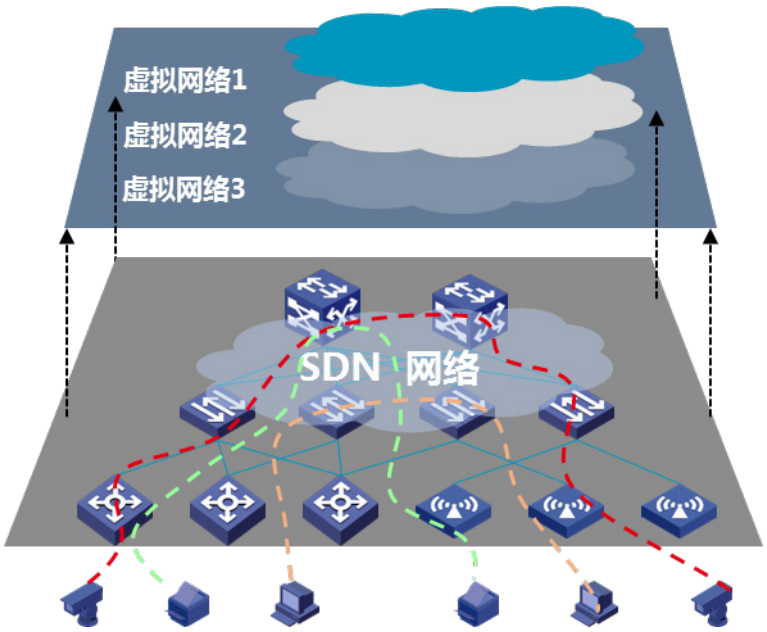
- Access 设备 VLAN 分配规则

  Access 设备作为二层接入设备，用于连接终端设备。Access 与 Leaf 设备连接的链路为 Access 上行接口，配置为 port trunk permit VLAN all；

  SeerEngine-Campus 控制组件会为 Access 设备的每个下行接口分配一个 VLAN ID，来标记每个终端的位置；从 VLAN 101 开始，同一台 Leaf 下，不同下行接口连接的 Access 设备，VLAN 都是从 VLAN 101 开始分配，可解决 VLAN 数量的限制问题。

2.2.2 业务逻辑网络架构设计

图6 园区网络业务逻辑划分图



网络按照业务需要划分多个虚拟网络，虚拟网络之间逻辑隔离。例如如下虚拟网络需求：

- 办公网络
  - 主要功能：园区内部人员办公所用网络，包括邮件、飞书等各种 OA 系统及互联网访问等
  - 主要终端：园区内部员工办公 PC、手机、PAD
  - 网络出口：专线和 internet
- 访客网络
  - 主要功能：访客上网
  - 主要终端：访客 PC、手机、PAD

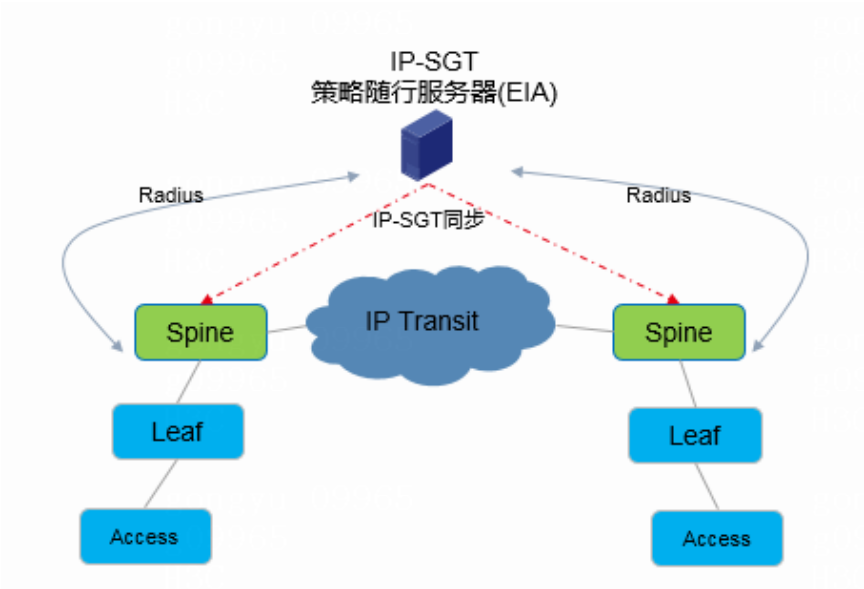
- 网络出口：internet
- IOT 物联网络
  - 主要功能：园区物业管理，资产定位，能耗管理，安防及部分亚终端网络
  - 主要终端：摄像头、考勤机、闸机、烟感、停车地磁、资产标签等
  - 网络出口：专线

2.2.3 IP-SGT 订阅实现策略随行

1. 场景 1：IP-Transit 场景

目前 VXLAN 微分段场景，EVPN 的主机路由中会携带 SGT 信息（即用户 IP 对应的角色）进行全网同步，实现 SGT 信息的全网同步。为实现该功能，多园区之间需要支持建立 VXLAN 隧道。但是有的局点园区之间因为各种原因不支持直接建立 VXLAN 隧道，园区之间只能通过原始 IP 报文互通，即 IP-Transit 场景。该场景无法通过 EVPN 全网同步 SGT 信息，导致无法实现全网的策略随行。为解决该问题，需要引入 IP-SGT 订阅功能，策略执行点从 EIA 订阅全网用户的 IP 和 SGT 的对应关系，以实现全网用户的策略随行。

图7 IP-Transit 场景



场景 1 组网说明：

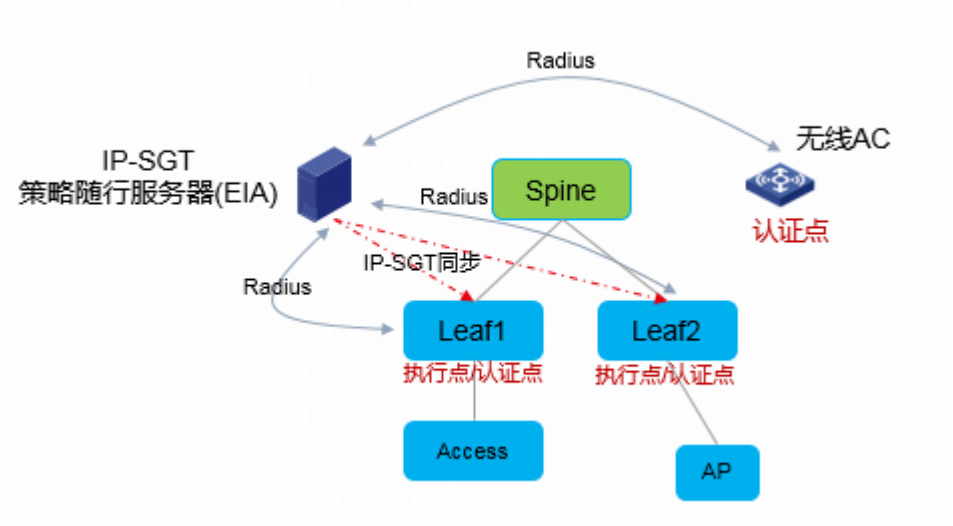
(1) 隔离域间不使用 EVPN 互联，即域间流量不通过 VXLAN 封装，直接使用 IP 报文承载，因此无法通过 EVPN 的主机路由传递 SGT 信息。

(2) 红线表示边界设备向策略随行服务器（EIA）订阅 IP-SGT 信息，用户认证成功后，EIA 会把 IP-SGT 对应关系推送到边界设备，这样边界设备会有其他隔离域的用户 IP 和权限的对应关系信息，跨隔离域互访可以在本隔离域的边界设备上执行策略控制。

2. 场景 2：安全组 VLAN 解耦场景

目前 VXLAN 组网中，创建安全组时缺省会关联一个 VLAN，用于无线业务授权（无线业务授权 VLAN 后，通过 Leaf 下行口的静态 AC 关联 SGT），在大规模安全组场景下，需要将安全组和 VLAN 解耦，无线用户的 SGT 信息通过 EIA 订阅获取。

图8 安全组 VLAN 解耦场景



场景 2 组网说明：

(1) 对应组策略模式且安全组数量较多场景，安全组需要与 VLAN 解耦。

(2) 解耦后，创建安全组时不关联 VLAN。即安全组的 ID 与 VLAN 无关。无线用户认证时，只返回认证结果，不需要授权 VLAN。

(3) 策略执行点通过 IP-SGT 订阅可以获取无线用户 IP 的 SGT 信息。有线业务用户仍然授权 SGT，IP 的 SGT 信息通过 EVPN 的主机路由全网同步。

2.3 软件部署方案设计

园区场景下需要部署 SeerEngine-Campus 与 vDHCP 服务器，它们的主要功能如下：

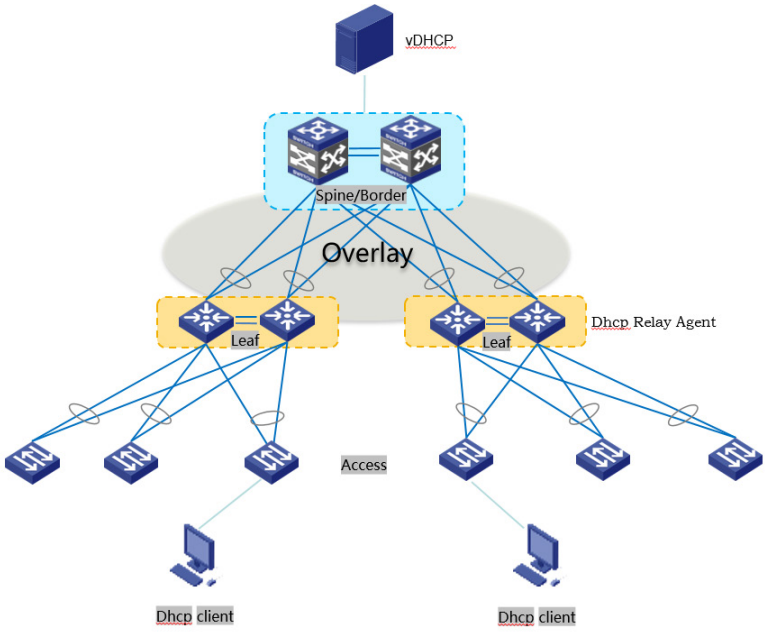
- SeerEngine-Campus：园区网控制组件，以单机或三机集群模式部署在统一数字底盘上；
- vDHCP 服务器：可作为用户上线时使用的 DHCP 服务器，为用户终端分配 IP 地址。同时用户使用 SeerEngine-Campus 支持的设备空配置自动化上线功能时，DHCP 服务器可作为给设备分配 IP 地址的 DHCP 服务器。vDHCP Server 为园区场景下常使用的 DHCP 服务器，以单机或双机主备模式部署在统一数字底盘上；
- 其他 DHCP 服务器：当园区用户数量大于 1.5w 时，需要额外部署其他 DHCP 服务器用于业务网段地址分配，推荐使用微软 DHCP 服务器。如果园区选用 BRAS 做准出认证和计费（非准入准出联动），则必须部署其他 DHCP 服务器，用于业务网段的地址分配，其他 DHCP 服务器仍时推荐微软 DHCP。使用 BRAS 做准入准出联动时，并不要求必须部署其他 DHCP 服务器。

下图为 AD-Campus 的典型组网，组网由 Spine、Leaf、Access 三层组成：

- Access 作为二层交换机，连接用户终端设备（DHCP Client）；

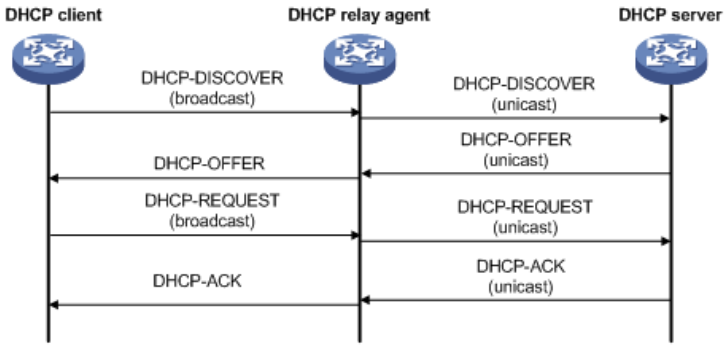
- Leaf 设备作为 DHCP Relay agent，开启 DHCP 中继功能；
- Spine 设备转发 DHCP 报文到 DHCP Server；

图9 Spine、Leaf、Access 三层组网



DHCP 客户端和 DHCP Server 处于不同物理网段时，客户端可以通过 DHCP 中继与 DHCP Server 通信，获取 IP 地址及其他配置信息。

图10 DHCP 中继的工作过程



- 具有 DHCP 中继功能的网络设备收到 DHCP 客户端以广播方式发送的 DHCP-DISCOVER 或 DHCP-REQUEST 报文后，将用户所在安全组对应的 vxlan id 封装到 Option 82 中，并且修改 DHCP Discovery 报文中的源地址为 vsi4094 的 IP 地址，以单播转发给指定的 DHCP Server。
- DHCP Server 识别 DHCP Discovery 报文中的 Option 82 携带的 vxlanid 为客户端分配 IP 地址等参数，并通过 DHCP 中继将配置信息转发给客户端，完成对客户端的动态配置。

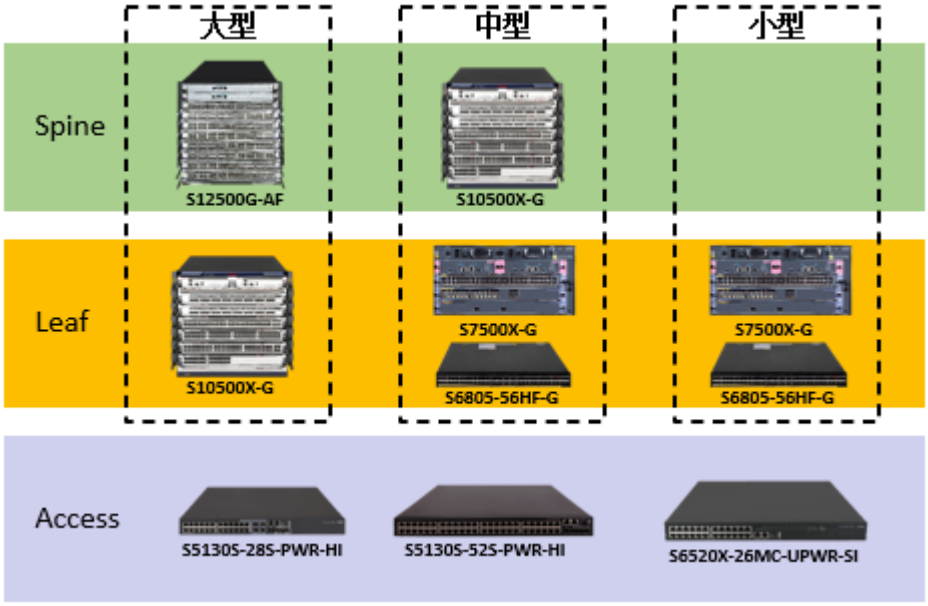
EIA：终端智能接入管理，用于终端用户的认证与接入。

上述软件都可以部署物理服务器或虚拟机。

2.4 网络设备选型

为了能够满足各种规模的网络建设需求，最大化投入产出比，在保证运营维护便利的前提下对组网建设进行标准化。

图11 网络设备选型



- 大型 SDN 园区设备选型：

  Spine 采用 S12500G-AF（可选 4/8/16 槽位），Leaf 采用 S10500X-G（可选 6/8/12 槽位），Access 采用 S5130S-28S-PWR-HI（24 口）、S5130S-52S-PWR-HI（48 口），Access 多速率采用 S6520X-26MC-UPWR-SI（5G/2.5G/1000/100BASE-T 自适应）
- 中型 SDN 园区设备选型：

  Spine 采用 S10500X-G（可选 6/8/12 槽位），Leaf 采用 S7500X-G、S6805-56HF-G，Access 采用 S5130S-28S-PWR-HI（24 口）、S5130S-52S-PWR-HI（48 口），Access 多速率采用 S6520X-26MC-UPWR-SI（5G/2.5G/1000/100BASE-T 自适应）
- 小型 SDN 园区设备选型：

  Leaf 采用 S7500X-G、S6805-56HF-G，Access 采用 S5130S-28S-PWR-HI（24 口）、S5130S-52S-PWR-HI（48 口），Access 多速率采用 S6520X-26MC-UPWR-SI（5G/2.5G/1000/100BASE-T 自适应）

2.5 网络规模评估

承载用户的规模（在线终端规模）是园区网建设非常重要的考量因素。在 AD-Campus 方案中，终端规模主要由 Spine 和 Leaf 控制，终端接入规模影响因素参考如下：

- Leaf 决定本机下的终端规模，受限于设备 AC 规格、主机路由、单板数量、内存、CPU 等因素。
- Spine 决定本 Fabric 下的终端规模，受限于主机路由，跨 VPN 路由引入方式等。

- Fabric 设备选型评估建议参考 PMO 上《AD-Campus 解决方案产品适配清单》里面提供的计算工具。

3 网络资源规划

3.1 隔离域规划

隔离域是指由一个或者多个 Fabric 组成一个网络连通域，每个隔离域可具有独立的认证系统、DHCP 服务器、无线 AC 控制组件等网络服务，实现网络服务的本地化部署，从而减少远端管理带来的网络带宽的消耗和降低网络延迟。多个隔离域间，可通过建立 EBGP 邻居的方式拉通多隔离域间的用户路由，从而达到隔离域间用户互访的目的。

通过评估不同园区之间是否有网随人动的需求，来判断需要如何划分隔离域。并不是隔离域越大越好，在隔离域内的园区之间，会同步用户的主机路由，多个隔离域之间是不会做主机路由同步的。

3.2 二层网络域规划

每个二层网络域只属于一个隔离域，每个二层网络域可以覆盖隔离域下的所有 Fabric，从而实现多 Fabric 的 IP 随行。

3.3 私有网络（VPN）规划

为用户定义一个逻辑上的私有隔离网络，每个私网为一个 VRF，不同私网之间默认网络隔离。

3.4 VLAN规划

VLAN(Virtual Local Area Network)又称虚拟局域网,是指在交换局域网的基础上,采用网络管理软件构建的可跨越不同网段、不同网络的端到端的逻辑网络。一个 VLAN 组成一个逻辑子网,即一个逻辑广播域，它可以覆盖多个网络设备,允许处于不同地理位置的网络用户加入到一个逻辑子网中。

域网内部 VLAN 划分如下表：

序号	VLAN 类型	VLAN 范围
1	设备自动化上线VLAN	系统默认：1
2	M-LAG VLAN	系统默认：2
3	BFD MAD检测VLAN	系统默认：100
4	有线业务VLAN	系统默认范围（可修改）：101-3000
5	自动化部署上线Underlay VLAN	系统默认范围（可修改）：3001-3500
6	无线业务VLAN	系统默认范围（可修改）：3501-4000
7	园区出口VLAN	用户自定义

序号	VLAN 类型	VLAN 范围
8	园区免认证VLAN	用户自定义
9	无线AP注册上线VLAN	系统默认：4093
10	接入设备纳管VLAN	系统默认：4094

3.5 IP地址规划

合理的 IP 地址规划，对于大型网络是至关重要的，很多 IP 地址在部署后很难变更，以北京办公楼网络改造为契机，可以重新梳理 IP 地址规划，对某集团网络未来扩展和保障网络稳定具有很大意义。

3.5.1 地址规划原则

为了保证地址规划的科学性和可行性，规划应遵循以下基本原则：

- 简单性：地址分配应清晰明了易于实施，降低网络扩展的复杂性，简化路由表的条目。
- 扩展性：地址分配在每一层次上都要留有余量，在网络规模扩展时能保证地址叠合所需的连续性。
- 灵活性：地址分配应具有灵活性，以满足多种路由策略的优化，充分利用地址空间。
- 寻迹性：规整目前现用地址段，使用较为整段的地址或便于记忆的地址范围。

3.5.2 总体设计思想

为了节省网络地址空间，同时考虑网络地址的统一管理和将来扩展的需要，局域网网络地址规划的总体设计思路是：

- 尽量采用 RFC 1918 规定的私有 IP 地址空间，即 10.0.0.0/8、172.16.0.0/16、192.168.0.0/24；
- 局域网内各区域在上述地址空间中分配连续的 IP 地址区域块；
- 为了将各类业务分开，将地址空间按业务类别分为几个部分；
- 各局域网的网关地址为该网地址空间的最后一个可用地址；
- 广域网 IP 地址考虑到点到点模式接入，将采用 30 位子网掩码；
- 各分支单位的网络互联将通过 NAT 转换实现。

3.5.3 IP 地址段分配

办公地址段				
部门	IPv4		IPv6	
访客				
员工-研发				

办公地址段					
员工-非研发					
IOT					
管理互联类地址					
VLAN 1		VLAN 4093		VLAN 4094	
lo 0		外联地址			
其他地址					

3.6 DHCP服务规划

园区网络常用 vDHCP Server 和微软 DHCP Server，并需要在 Leaf 为三层网络部署 DHCP relay，使用 BRAS 做准出认证时（非准入准出联动）,还需要将 BRAS 作为 DHCP 二级 relay。

- vDHCP Server 是必配组件，自动化使用 vDHCP 为设备分配地址。当用户组网为 IPv4 单栈时，接入终端最大规模为 5W，当用户组网为 IPv6 单栈时，接入终端最大规模为 1.5w，当用户组网为双栈混合使用时，接入终端最大的规模为 5W，此时要求 IPv6 不超过 1.5W，也可以使用 vDHCP Server 为业务 VLAN 分配地址。由于 vDHCP Server 必须与控制器融合部署，不能独立部署，则当集群出现故障时，vDHCP Server 将无法继续提供逃生服务。
- 微软 DHCP Server 独立部署，提供业务 VLAN 的地址分配，一般为双机热备。如果存在 BRAS 做准出认证（非准入准出联动），要求微软 DHCP Server 与控制器地址不同网段，在管理区 L3 交换机上为微软 DHCP Server 和控制器交互设置路由，并为微软 DHCP Server 和 BRAS 创建二层透传 VLAN，BRAS 连接 L3 交换机的子接口为微软 DHCP Server 的网关。
- Leaf 作为 DHCP relay，收到 DHCP 客户端以广播方式发送的 DHCP-DISCOVER 或 DHCP-REQUEST 报文后，将用户所在安全组对应的 vxlan id 封装到 Option 82 中，并且修改 DHCP Discovery 报文中的源地址为 vsi4094 的 IP 地址，以单播转发给指定的 DHCP Server。
- BRAS 作为 DHCP 二级 relay（仅在使用 BRAS 做准出认证，非准入准出联动时）。为了实现双栈用户一次认证双栈通过，需要 Leaf 将终端的 MAC 地址通过 option 携带在 DHCP 报文中，DHCPv4 使用 option61，DHCPv6 使用 option79，并将 Leaf 上 DHCP Server 地址设置为 BRAS 的地址，确保 DHCP 报文转发给 BRAS。当 DHCP 报文通过 BRAS 的时候，BRAS 作为 DHCP 二级 relay，可以记录终端的 MAC 地址信息，将 MAC 地址作为终端的唯一标识实现无感知。需要注意，为了确保 DHCP 回程报文也必须经过 BRAS，需要将 BRAS 连接 L3 管理交换机的接口设置为分配业务网段地址 DHCP Server 的网关，并开启非首跳 DHCP relay 功能。

3.7 认证方式规划

用户认证分为准入认证和准出认证，准出需要支持计费。

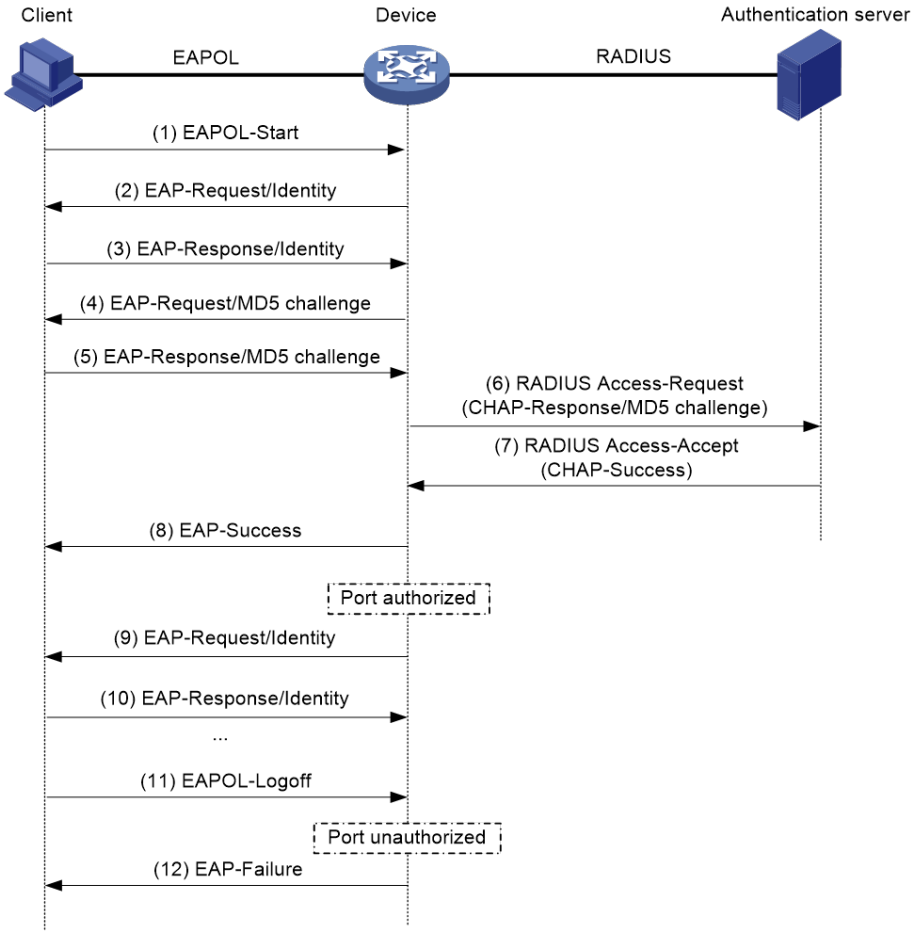
AD-Campus 方案用户准入认证支持 MAC 认证、802.1X 认证、MAC Portal+认证和 Web Portal+认证、访客认证，准出认证支持通过 SR88X 对接第三方 AAA 服务器进行 IPOE Web 认证。

认证使用的 AAA 服务器规划有以下原则：

(1) EIA 仅做准入认证，不支持做准出和计费。支持所有准入认证方式。使用名址绑定功能必须使用 EIA 做准入认证。

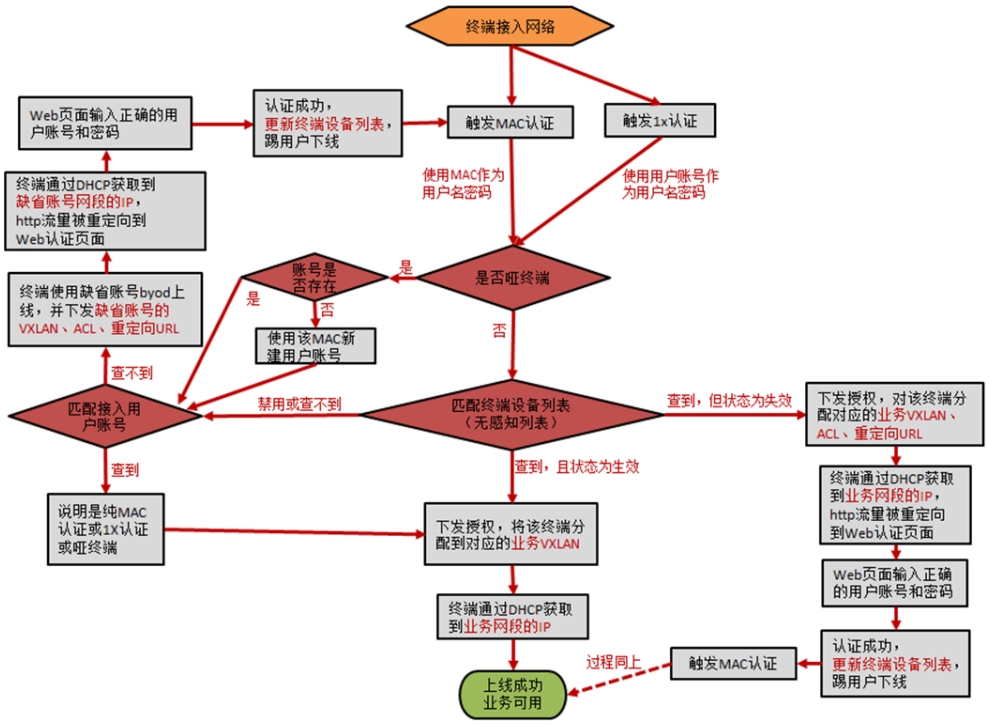
- MAC 认证：通常适用于物联终端，包括哑终端，这些终端多数没有操作页面，无法输入用户名和密码，则需要通过其 MAC 地址作为认证的凭据。
- 802.1X 认证：通常在安全性要求较高的场景下使用，需要在 EIA 上提前创建用户账号。使用 802.1X 的用户终端需要支持作为 802.1X 认证的客户端，一般的移动终端或 Windows 系统都内置了 802.1X 客户端，可以支持此认证方式。也可以选择配合 iNode 客户端使用，首次认证后由 iNode 客户端自动认证，实现更安全更方便的准入控制。

图12 流程图



- MAC Portal+认证：通常在安全性要求不太高，且不希望每次登陆都要做认证的场景。从用户体验来看，与 Portal 认证相同，只需要终端有浏览器，就可以输入用户名密码进行认证，首次认证后，可以在很长一段时间内实现无感知认证。同时 MAC Portal+认证还可以支持名址绑定功能，确保用户在更换接入位置后，IP 地址不变。

图13 流程图



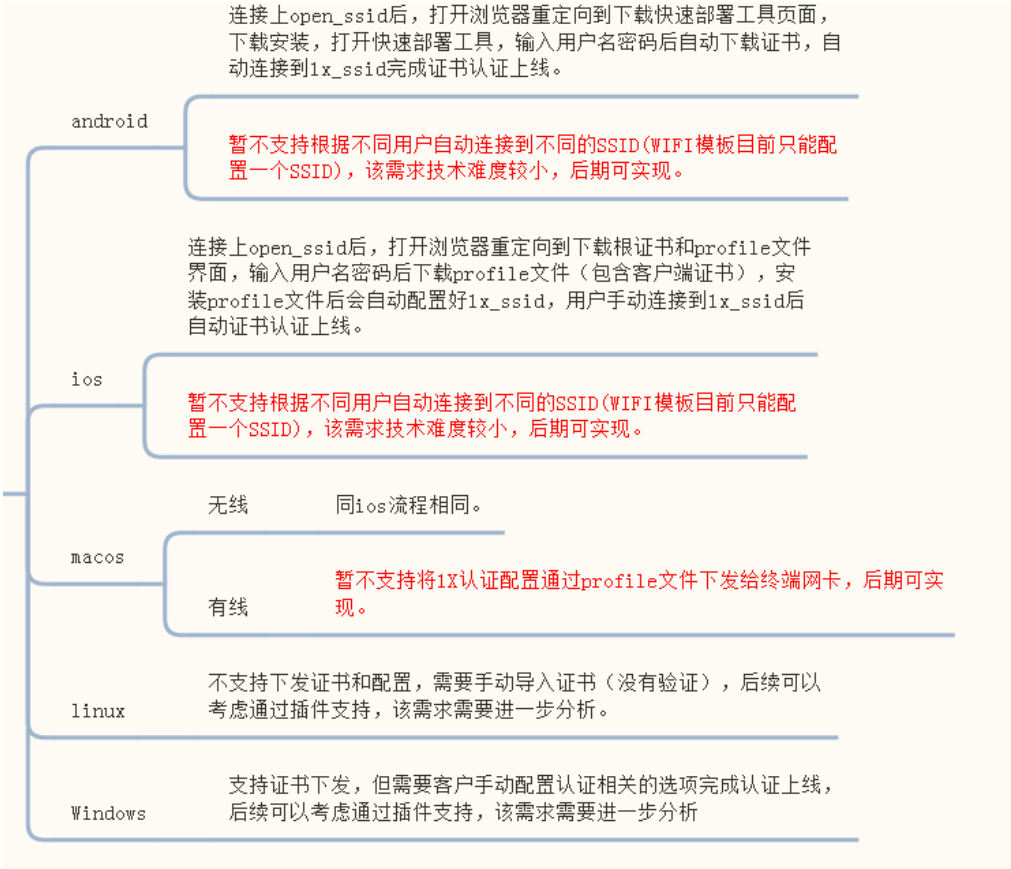
- 访客认证：支持短信认证、二维码认证、密码认证，Facebook 认证即将支持。需要注意，短信认证默认使用亿美短信平台，如果要使用其他短信平台，需要定制开发。
- 证书认证：

  采用 EAP-TIS 方式进行证书认证，TLS 认证是基于 client 和 server 双方互相验证数字证书，是双向验证方法。首先由 server 提供自己的证书给 client，client 验证 server 证书通过后，提交自己的数字证书给 server，交由 server 进行验证。

  证书快速部署：

  标红部分为目前暂不支持后期可实现需求，未标红部分为目前已支持。

图14 各操作系统支持情况



证书管理：

证书管理主要包括如下几个部分：

- 查看：EIA 终端设备管理中可以查看每个终端的证书详情，包括唯一序列号，证书有效期等信息。
- 过期提醒：可以做到用户证书过期前提醒，提醒方式为 iMC 平台告警，同时给用户发送邮件，短信。
- 吊销：管理员可以吊销终端的客户端证书，被吊销的证书无法接入网络。
- 过期重发：终端重新按照证书快速部署流程走一遍即可。

(2) 第三方 AAA 做准入认证，支持 MAC 认证、802.1X 认证、Web Portal+认证。

Web Portal+认证：在园区使用微分段时可以使用。在有线终端以我司交换机作为 NAS 认证点时使用此认证方式。主要支持的第三方 AAA 服务器有深澜和城市热点。从用户体验来看，与 Portal 认证相同，只需要终端有浏览器，就可以实现首次认证，后续很长一段时间内无感知接入。为了实现后续无感知，需要认证服务器支持 MAC 快速认证。主流的深澜和城市热点均支持。

第三方 AAA 做准入认证，需要第三方 AAA 支持授权 user-group 和授权 VLAN。此特性在大部分主流 AAA 上均支持。

(3) 第三方 AAA 做准出认证，有以下两种方式：

- EIA 准入后，对接第三方 AAA 服务器，将准入信息同步到第三方 AAA 服务器实现无感知准出，同时由第三方 AAA 服务器对出口流量进行计费。此方式下，AD-Campu 方案主要适配的 AAA 服务器是城市热点和深澜，其他第三方 AAA 服务器需要研发确认，可能需要第三方定制开发。
- 用户直接在第三方 AAA 做准出认证，可采用流量触发的方式，或者由 SR88X 作为 BRAS 对接第三方 AAA 做准出认证。这种情况下，一般准入会采用免认证或者无感知的认证方式，避免用户需要做两次认证，降低网络使用体验。

3.8 安全组资源规划

安全组分为两种，用户安全组和 IT 资源组。AD-Campus 解决方案中的策略控制就是通过设定用户安全组之间、用户安全组与 IT 资源组之间的访问策略来生成的。

- 用户安全组定义用户网络权限，通过与 EIA 的协同，完成用户到用户安全组的绑定。用户上线后，通过认证授权用户到不同安全组，为用户提供不同的网络权限。
- IT 资源组定义网络资源（如服务器）的网络权限，通过与 EIA 的协同或者静态映射，完成网络资源到 IT 资源组的绑定。

单个隔离域内支持的用户安全组数量和 IT 资源组数量有限制，请根据园区业务的需要，结合用户数量和使用的 Leaf 设备型号规划安全组数量。

安全组数量规模如下：

场景	设备类型	安全组数量
非微分段场景	纯有线用户场景（MARVEL) 注：有Marvel芯片产品情况下，整体规格按照Marvel评估。	每VRF最大50，单隔离域最大50，多隔离域之和500 若项目需超过推荐安全组数量，请联系研发单独评估
	纯有线用户场景（BCM)	每VRF最大500，单隔离域最大500，多隔离域之和500 若项目需超过推荐安全组数量，请联系研发单独评估
	有线无线一体化	需咨询研发； 安全组数与用户安全组、Leaf下行口数量、Leaf下行口是否聚合、组间策略数量、在线用户数、芯片类型、无线转发方式均有关，需通过公示进行计算，计算复杂，具体计算请咨询用服二线或产品部。
微分段场景	纯有线用户场景（MARVEL) 注：有Marvel芯片产品情况下，整体规格按照Marvel评估。	单个隔离域最大150个，所有隔离域之和最大150个 若项目需超过推荐安全组数量，请联系研发单独评估

场景	设备类型	安全组数量
	纯有线用户场景（BCM)	单个隔离域最大500个，所有隔离域之和最大500个 若项目需超过推荐安全组数量，请联系研发单独评估
	4K安全组场景（仅BCM支持）	可以配置4K安全组，同时需要开启相关功能（vpn-default和业务VPN静态路由互引、安全组VLAN解耦、无线业务IPSGT订阅）。 使用限制如下： 1）4K安全组要求全BCM组网，spine/leaf都是BCM才行。 2）要求微分段。 3）单VPN只支持创建1K个安全组，整网可以创建4K安全组。 4）目前对外开局，推荐模型为：4个VPN，每个VPN 125个二层网络域，每个二层网络域混杂8个sgt。如果二层网络域数量超过500个，请单独咨询研发确认。
	有线无线一体化	需咨询研发； 安全组数与用户安全组、Leaf下行口数量、Leaf下行口是否聚合、组间策略数量、在线用户数、芯片类型、无线转发方式均有关，，需通过公示进行计算，计算复杂，具体计算请咨询用服二线或产品部。

安全组资源占用情况估算：

策略矩阵每一行下发一个 Policy，每一列下发一个 node，每个 node 内可以添加多个规则条目。ACL 资源占用取决于策略矩阵中所有应用的规则条目总数（记为 N），各款型设备 ACL 占用数量：

BCM 系列 TD3 板卡/6550XE：占用 4N 个 ACL 资源。

4 Underlay 网络设计

4.1 管理网设计

Spine 和管理区内服务器之间的网络是管理网，至少包含一台三层交换机，需要保证 VLAN1 和 VLAN4094 的地址 IP 可达。还需要保证连接管理 SeerEngine-Campus 等软件的 VLAN 地址可以访问。

如果部署 BRAS 做准出认证，由于 BRAS 一般使用带外管理口纳管，则需要在管理区的控制器和 L3 管理交换机上为纳管 BRAS 设置路由，并为 BRAS 与管理区内各服务器之间设置路由，确保可以互访。

4.2 二层网络设计

4.2.1 生成树设计（防环网设计）

为了避免在二层网络中产生环路，选择在整网运行 STP 协议，通过设备间交互的 STP 报文，选择性的阻塞某些端口，最终将网络修剪为无环的树形结构，在保证不产生环路的前提下，还提供了链路备份的功能。最初的生成树协议为 STP（Spanning Tree Protocol，生成树协议），之后又发展出 RSTP（Rapid Spanning Tree Protocol，快速生成树协议）、PVST（Per-VLAN Spanning Tree，每 VLAN 生成树）和 MSTP（Multiple Spanning Tree Protocol，多生成树协议）。

自动化 1.0（老自动化）方案中，为防止网络中可能存在的环路，SDN 控制组件自动下发如下配置：

- Spine、Leaf、access 交换机全部开启 STP 功能，STP 模式置为 PVST；
- 只开启 VLAN 1 的 PVST，关闭其他 VLAN 的 PVST 功能，配置核心 Spine 设备为 VLAN 1 的根交换机；
- 连接服务器、PC、打印机等终端设备的接口开启边缘端口。

自动化 2.0（新自动化）方案中，为防止网络中可能存在的环路，SDN 控制组件自动下发如下配置：

- Spine、Leaf、access 交换机全部开启 STP 功能，STP 模式置为 MSTP；
- 将部分 vlan ignore（主要是 underlay vlan），避免不当阻塞引发 spine/leaf 间之间的路由组网链路被阻塞；
- Spine 设备为根交换机，配置最高优先级，同时配置根保护，避免被抢夺；
- 在核心配置 TC restriction，限制 TC 报文扩散，避免拓扑变化报文过度扩散引发的整网的表项刷新以及网络动荡。
- 将拓扑中级联链路 cost 优先级调高，降低拓扑变化带来的影响；
- 连接服务器、PC、打印机等终端设备的接口开启边缘端口；
- 为了避免网络中 HUB 等不支持 STP 协议的设备造成环网，按需可整网开启 STP 黑洞探测；
- 大型网络中，为避免底层 Access 出现错误连接到两组 Leaf 上，从而导致 Leaf 间环网，可按需开启 LLDP 跨域检测。

4.3 三层网络设计

4.3.1 路由设计

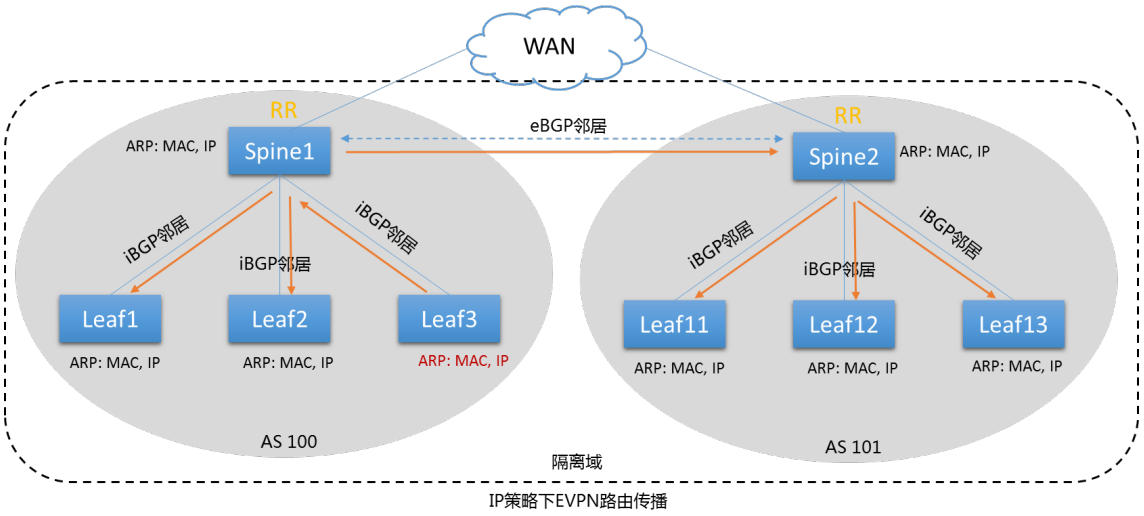
本方案中 Leaf 和 Spine 设备之间运行 OSPF、BGP 路由协议，完成 underlay 及 overlay 网络的构建，Spine 和 Leaf 使用一个 ospf area 0，并且在 Spine 和 Leaf 之间建立 BGP EVPN 邻居，BGP EVPN 是利用 MP-BGP 扩散 MAC 地址和主机路由（ARP 表项）。用户的主机路由会通过 BGP EVPN 扩散到隔离域内所有分布式网关 Leaf 和 Spine 上，用于对流量进行访问策略控制。

多 Fabric 场景下，Fabric 内 BGP 邻居均在一个 AS 号内，BGP 邻居均为 iBGP 邻居。由于 BGP 路由发布规则规定，从 iBGP 邻居收到的路由不可以再发布给 iBGP 邻居，则 AS 内所有 iBGP 邻居需要建立全连接，这种组网往往是不现实的。方案中将 Spine 交换机作为反射器，所有 Leaf 交换机作为非反射器，这样所有 Leaf 上学习到的 EVPN 主机路由就可以通过 Spine 反射到所有 Leaf 上。

不同 Fabric 需要使用不同 AS 号，则不同 Fabric 的 Spine 之间建立 eBGP 邻居。为了使 eBGP 邻居之间也可以自动建立 VXLAN 隧道，则 eBGP 邻居之间也需要建立全连接，确保多 Fabric 之间的流量不会绕行其他 Fabric。

对于使用 IP 策略的多个 Fabric，同一隔离域内的多个 Fabric 之间才有发布 EVPN 主机路由的需要，不同隔离域之间不需要主机路由扩散，仅发布网段路由即可。

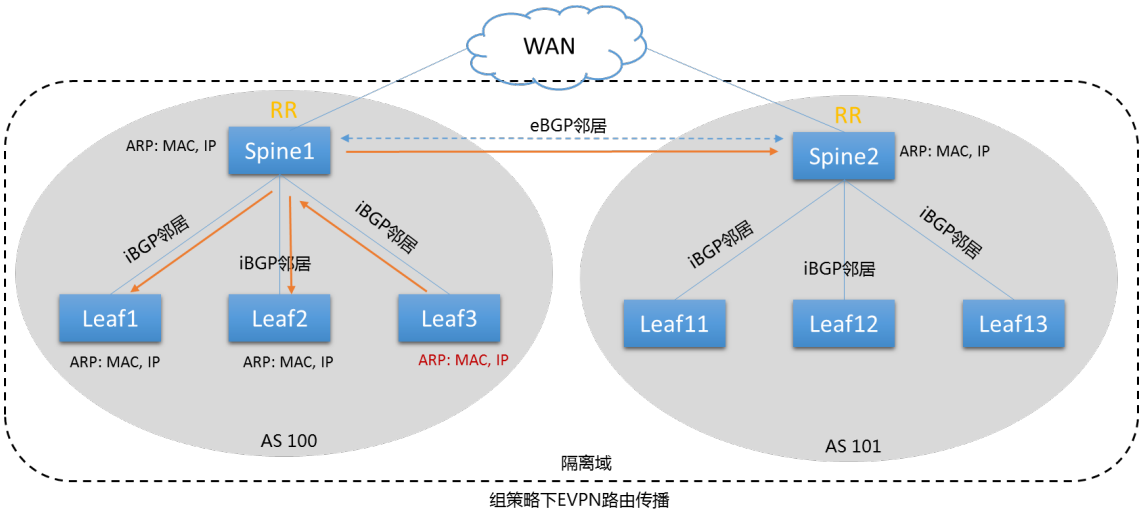
图15 IP 策略下 EVPN 路由传播



IP策略下EVPN路由传播

对于使用组策略的多个 Fabric，需要在多个 Fabric 之间发布 EVPN 主机路由，主机路由中携带其他隔离域的 SGT 信息。通过控制器上开启路由抑制功能，控制主机路由仅发布到各 Fabric 的 Spine，不会向下发布给 Leaf，会在 Spine 控制跨隔离域访问流量。

图16 组策略下 EVPN 路由传播



对于单个 Fabric，在 SDN 控制器上只需要在创建 Fabric 的时候，设置 BGP 的 AS 号，并设置 RR（新自动化方案中，不再需要单独设置 RR），即可自动下发 BGP 的邻居配置和 EVPN 地址族（l2vpn evpn）配置到 Leaf 和 Spine。BGP 会自动根据配置建立 BGP EVPN 邻居关系，同时根据 BGP EVPN 邻居自动建立 VXLAN 隧道，每个 BGP EVPN 邻居就是一台 VXLAN 网络中的 VTEP 节点。

对于多个 Fabric，在隔离域互联中可以增加 Fabric 连接，SDN 控制器可在建立了 Fabric 连接的多个 Fabric 之间自动下发 eBGP 邻居配置和所需的 EVPN 地址族配置到各 Fabric 的 Spine。单 Fabric 不需要此配置。

图17 增加 Fabric

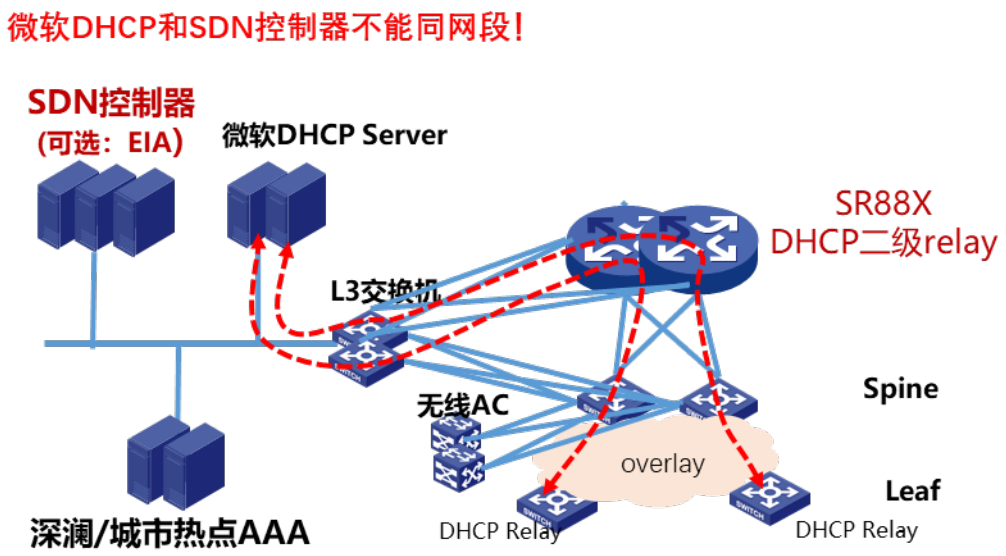
VTEP（Spine 和 Leaf）收到地址同步后，会检查收到的路由所属的私网，如果路由是允许接收的，则将路由中发布的主机路由表项放到 L3 路由表项中。为了避免大量垃圾 MAC 表项占用资源，默认

设置禁止 EVPN 从 ARP 表项学习 MAC 表项（arp mac-learning disable），则不会将学习到的 ARP 表项添加 MAC 表。

4.3.2 BRAS 路由设计

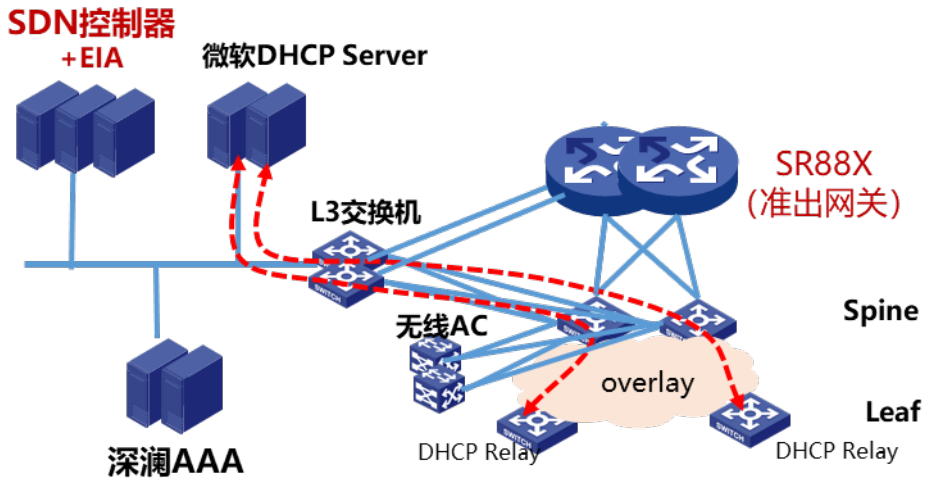
BRAS 做准出认证（非准入准出联动），与管理区控制器、AAA 服务器、DHCP 服务器的的交互需要通过管理区 L3 交换机转发，不通过核心交换机。BRAS 做为二级 DHCP relay，需要设置到 Leaf 一级 relay 的路由。BRAS 需要设置为共享出口，将目的为 BRAS 及访问外网的所有报文通过共享出口 VPN 转发给 BRAS。

图18 BRAS 做准出认证



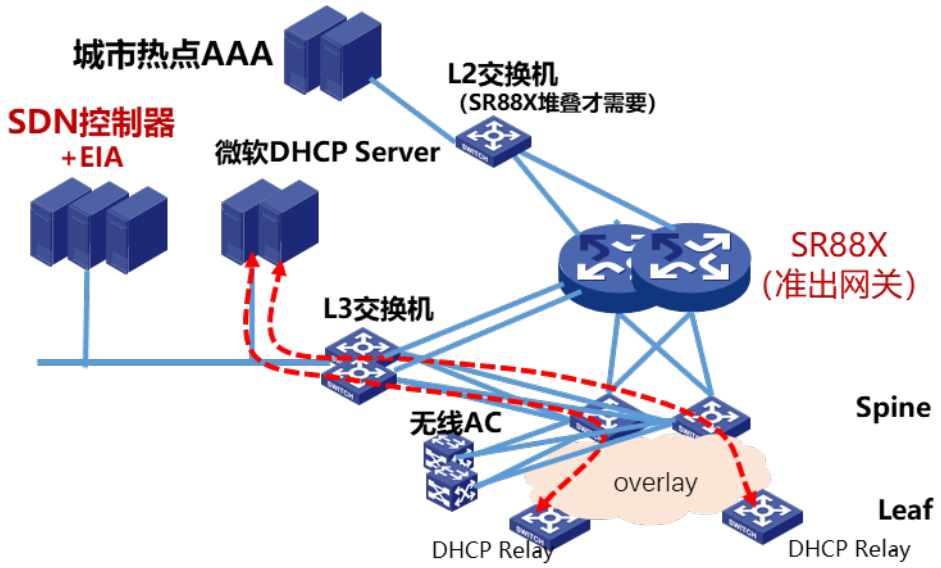
BRAS 做准入准出联动（第三方 AAA 服务器为深澜），与管理区控制器之间通过 L3 管理交换机与 BRAS 的管理口互访。BRAS 与第三方 AAA 服务器互访仍需要通过 Spine，需要为两者互访设置路由。BRAS 需要通过 vDHCP 服务器获取接口地址，需要在 BRAS 上设置路由访问 vDHCP 服务器。不需要在 BRAS 设置与微软 DHCP 的路由，两者没有互访需求。

图19 BRAS 做准入准出联动



BRAS 做准入准出联动（第三方 AAA 服务器为城市热点），与管理区控制器之间通过 L3 管理交换机与 BRAS 的管理口互访。BRAS 与第三方 AAA 服务器直连不需要通过 Spine，EIA 与第三方 AAA 服务器互访需要通过 Spine 和 BRAS，需要为两者互访设置路由。BRAS 需要通过 vDHCP 服务器获取接口地址，需要在 BRAS 上设置路由访问 vDHCP 服务器。不需要在 BRAS 设置与微软 DHCP 的路由，两者没有互访需求。

图20 BRAS 做准入准出联动（第三方 AAA 服务器为城市热点）



5 Overlay 网络设计

Overlay 网络需要管理员对网络业务策略进行规划，并在 SDN 控制组件上进行业务部署，通过一键下发，完成业务部署的自动化。整个过程，无需逐台进行设备配置，也不需要手工配置命令行。

5.1 Fabric业务部署设计

园区内需要物理隔离的业务用私网 VPN 隔离，私网内细分且也需要做隔离的业务可以划分为安全组来进行管理。安全组内用户之间默认是可以互相访问的。基于以上原则，一般可以按照如下原则划分安全组：

(1) 按照业务部门来划分，例如将一个工作组的人划分到一个安全组。

(2) 按照终端类型来划分，例如将打印机划分到一个安全组。

(3) 按照安全级别来划分，例如同一安全级别的人划分到一个安全组，具有相同权限。

总之，安全组的划分，要在明确园区业务的前提下，进行合理的划分，一方面要确保业务访问策略可以有效控制不同安全组内成员的互访关系，一方面要确保安全组划分尽量简单，数量可控，避免占用太多的设备资源。否则，安全组一旦划分有误，修改相对繁琐，还可能涉及要清除一些动态保存的表项，否则可能出现业务访问异常。

5.2 矩阵式策略规划设计

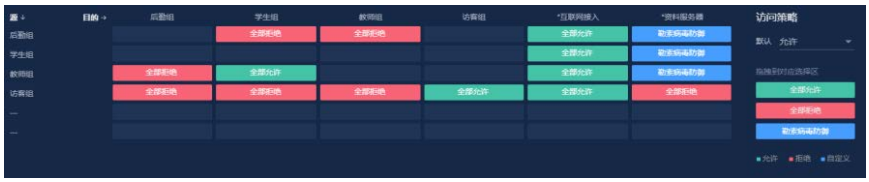
组间策略定义采用矩阵式，定义了安全组之间的访问策略。在组间策略定义好后，控制组件通过把组间策略转换成 PBR 配置下发到网络设备，实现网络访问权限的控制。

图21 定义组间策略

用户组	网络属性	IP组	虚拟专网
员工-研发	VXLAN1或SGT1	网段1	VRF1
员工-非研发	VXLAN2或SGT2	网段2	VRF1
访客	VXLAN3或SGT3	网段3	VRF2
IOT	VXLAN4或SGT4	网段4	VRF3
...			

如下图，控制组件支持拖拽式快速完成策略规划和部署。根据网络访问权限归纳，明确安全组之间的默认访问策略，通过设置合理的默认策略，减少访问策略的数量，可以减少对设备 ACL 资源的消耗。

图22 访问策略



5.3 静态IP场景设计

方案中 802.1x 和 MAC 认证是基于二层报文的，即认证通过后，用户的二层报文会被授权到对应的 VXLAN。后续用户可以根据需求使用动态或静态方式获取 IP。

如果使用静态 IP，不需要配置业务 DHCP 服务器（vDHCP 仍然需要，用于设备自动化上线），创建二层网络域时，“IPv4/IPv6 地址获取方式”选择“手动”。终端配置 IP 时，需要配置与二层网络域中子网相同的网段。

图23 手动获取

如果用户需要的网段较多，在创建二层域的子网时，可采用“一个主网段+多个从网段”的方式。其中主网段可以同时支持动态或静态 IP，从网段只支持静态 IP。对于静态 IP 终端，配置与主网段或从网段相同网段的 IP 即可（仅 IPv4 子网支持从网段，IPv6 子网不支持从网段）。

图24 子网

5.4 Fabric内部跨VPN互通设计

5.4.1 VPN 间互通场景

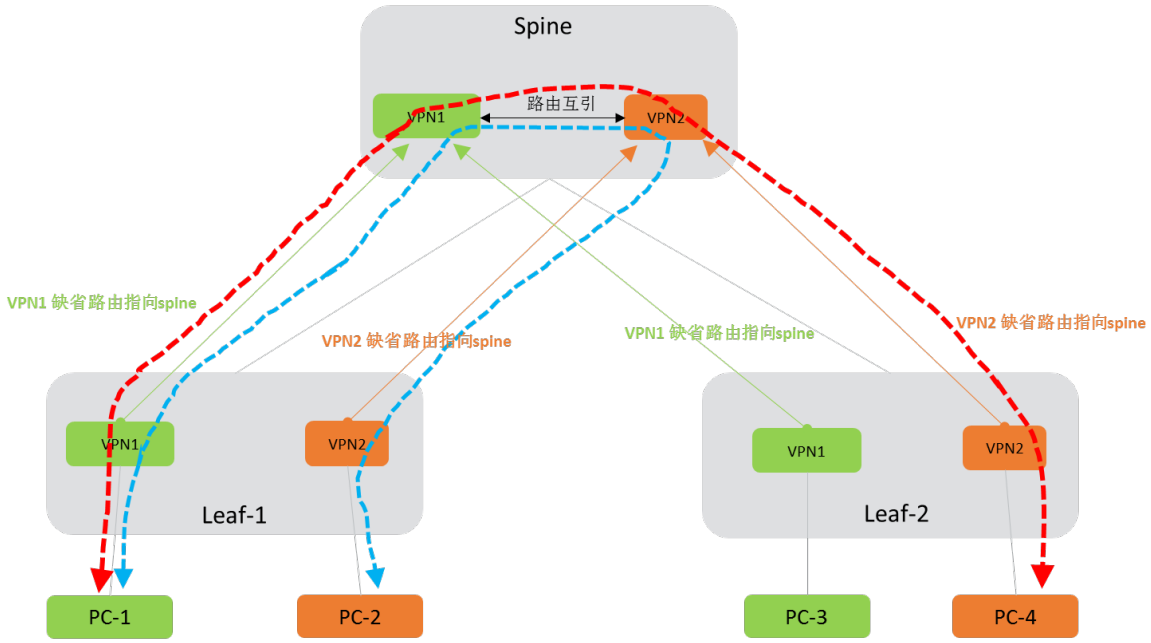
方案中，不同业务 VPN 间路由隔离。即业务 VPN 间的流量默认不通。如果有业务 VPN 互通的需求，需要在 Spine 上手工配置 VPN 间路由互引，方法如下：

(1) Spine 的各个业务 VPN 实例视图下手工配置 RT 互引（缺省不互引）。配置之后，Spine 上业务 VPN 之间可以互相学习路由。

(2) Spine 上各业务 VPN 配置各自的缺省路由。然后 BGP 视图的各个 VPN 子视图中引入缺省路由。该缺省路由会通过 BGP 发布给 Leaf 的对应业务 VPN。该 VPN 的用户访问其他 VPN 时，流量会命中缺省路由，转发给 Spine。由于 Spine 上配置了路由互引，因此流量会转发给对应的目的 VPN。

流量转发模型如下图所示，Leaf 上跨 VPN 的流量均先通过各自 VPN 的缺省路由转发给 Spine，在 Spine 上，由于配置了路由互引，因此可以根据路由找到对应的 VPN，然后流量经过 Spine 转发给对应的 Leaf 的对应 VPN。

图25 流量转发模型



5.4.2 VPN 内个别终端与其他 VPN 互通的设计

某些场景下，用户希望实现 VPN 下个别终端与其他 VPN 互通，而不是整个 VPN 间互通。

可通过路由复制+路由策略的方式进行路由互引。方法如下：

(1) Spine 的各个业务 VPN 实例视图下，手工配置路由复制，指定需要从其他 VPN 复制的明细路由。明细路由通过路由策略里匹配 32 位主机路由进行指定，命令如下：

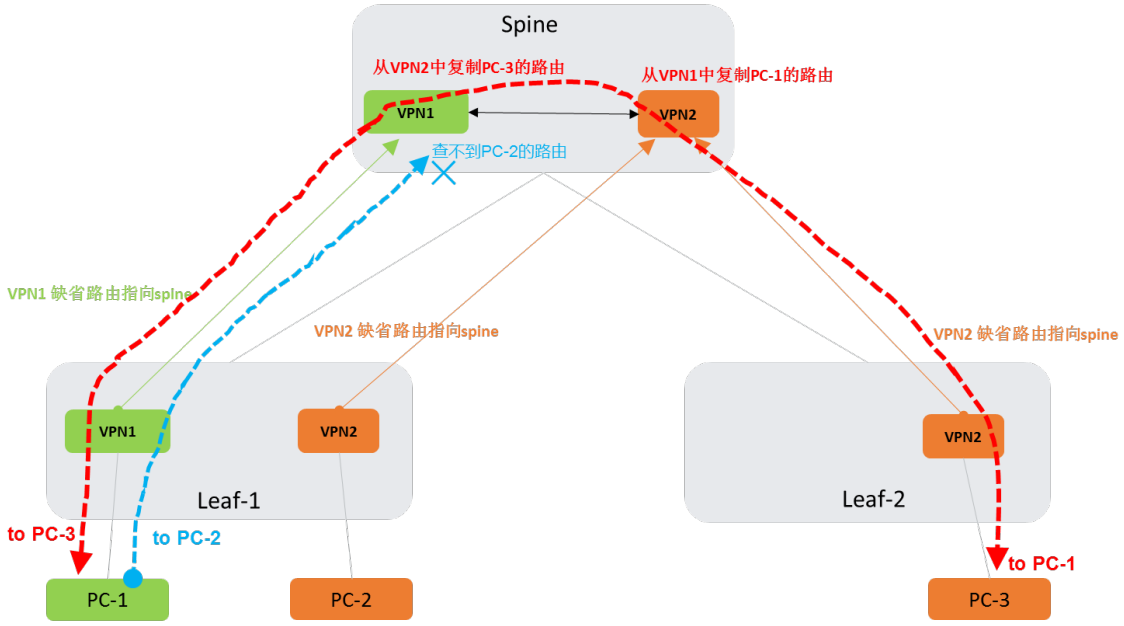
route-replicate from vpn-instance vpn2 protocol bgp 1 route-policy vpn1_vpn2

路由策略具体匹配的 32 位主机路由以实际组网为准。需要注意源 VPN 和目的 VPN 都要进行配置，确保双向路由可达。

(2) Spine 上各业务 VPN 配置各自的缺省路由。然后 BGP 视图的各个 VPN 子视图中引入缺省路由。该缺省路由会通过 BGP 发布给 Leaf 的对应业务 VPN。该 VPN 的用户访问其他 VPN 时，流量会命中缺省路由，转发给 Spine。由于 Spine 上配置了路由复制，该 VPN 可以学习到其他 VPN 的特定主机路由，因此流量会转发给对应的目的 VPN。

流量转发模型如下图所示，VPN1 的 PC-1 需要与 VPN2 的 PC-3 互通，因此需要在 Spine 的 VPN1 中从 VPN2 复制 PC-3 的路由，在 VPN2 中从 VPN1 复制 PC-1 的路由。PC-1 和 PC-3 的流量转发路径如红色虚线所示。对于 PC-1 到 PC-2 的流量，由于 VPN1 中没有从 VPN2 复制 PC-2 的路由，因此 Spine 的 VPN1 中没有 PC-2 的路由，PC1 到 PC-2 的流量到达 Spine 后，会查不到路由。因此 PC-1 和 PC-2 无法互通。

图26 流量转发模型

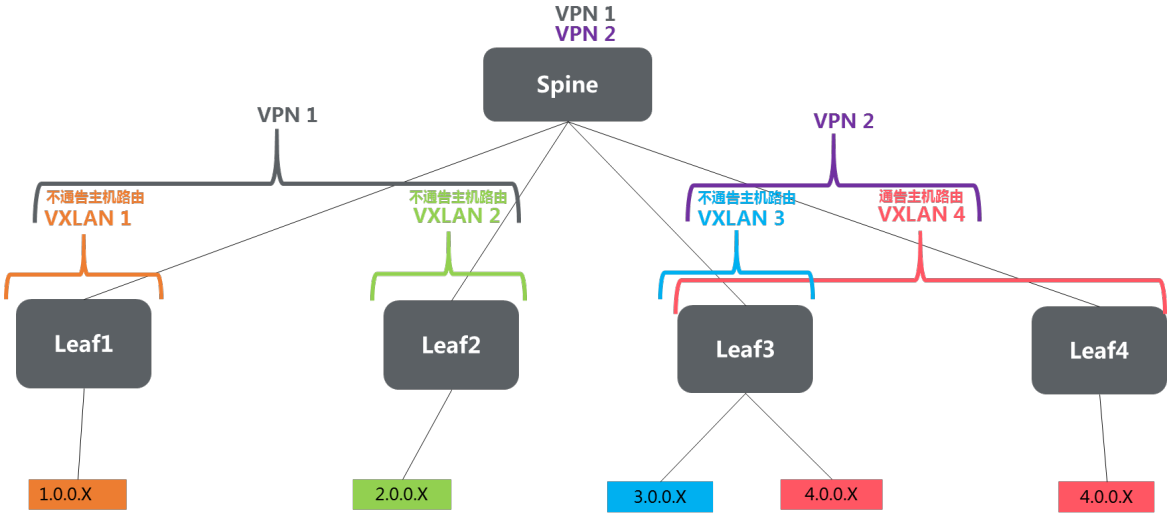


5.5 非分布式网关设计

在政务网等场景下，组网通常为单园区组网，Leaf 设备规模较大，且用户对于整网网随人动的使用需求较低，目前的 VXLAN 分布式网关组网对设备资源要求较高，通过实现二层网络域的按需下发，达到灵活组网、节约设备资源的目的。

方案支持在创建二层网络域（VXLAN）时，选择一个或多个网关设备，并可以选择开启或关闭主机路由通告。典型拓扑如下所示：

图27 典型拓扑



- VXLAN1，VXLAN2，VXLAN3 均为单台网关，只扩散网段路由，不扩散主机路由，降低了整网的路由规模，其他设备通过网段路由即可找到对应 Leaf。

- VXLAN4 为分布式网关，需要扩容主机路由。
- 组间策略 PBR 下发位置为源安全组对应二层网络域所在的网关设备的 VSI 接口。不存在该 VXLAN 的设备不需要下发策略。可以节约设备的 ACL 资源。

6 无线网络设计

6.1 无线工勘设计

6.1.1 AP 点位工勘原则

1. 勘测准备

(1) 确定覆盖区域的范围，搜集覆盖区域的平面图，包括覆盖区域大小、障碍物分布等，用以分析对信号的阻挡。

(2) 确认现场的覆盖要求，包括需要接入的用户数量、业务类型、带宽要求等。

(3) 确认允许安装设备的安装位置及安装方式。

(4) 确认安装现场可以提供的设备供电和走线方式。

(5) 确认现场组网情况、出口资源等。

(6) 根据组网方案和现场工勘情况，详细列出所需网络设备清单及辅料清单。

2. 现场勘测

(1) 现场工勘需要关注场地内现有无线网络的部署情况，确认用户自建和运营商代建的 WLAN 设备是否可以关闭。

(2) 如果条件允许，建议工勘时携带一个 Fat AP 到现场，并使用安装了 Cloudnet APP 的无线终端测试信号覆盖、衰减以及查看无线空口信道状态等情况，主要关注 2.4GHz 的 Channel 1、6、11 这几个常用信道情况。另外还可以通过绿洲上的“云工勘”导入待施工现场的平面图或者在云端手动绘制平面图，可对墙体材料等进行设定，通过放置不同型号的 AP，最终模拟无线体验效果。

(3) 操场体育馆等高密且存在人员流动的场景，部署时还需要考虑人流密集区域人体对于信号的遮挡造成的影响，有条件可以通过前期的现场测试进行评估。

3. 勘测整理

完成现场工勘后，还需要梳理点位并整理输出勘测报告，以备后续部署时使用。

6.1.2 产品推荐选型

- 宿舍场景推荐终结者、面板或放装 AP。该场景要求 AP 入室安装，每个房间放置一个 AP 保证信号覆盖。
- 普通教室场景推荐 2.4G 和 5G 双频或三频放装 AP。一般普通教室结构规则，部署放装型 AP 可满足该场景下用户同时进行网页浏览，社交通讯软件等业务使用的需求，可根据接入用户数情况选择双频或三频 AP。
- 室内高密场景推荐 2.4G 和 5G 三频放装 AP。该场景中接入终端密集，且可能存在部分设施遮挡的情况。三频放装 AP 可满足该场景对业务带宽和多终端高并发的要求。

- 室外场景可根据覆盖情况推荐选择双频或三频室外 AP。该场景可能需要将 AP 安装在室外，考虑到防水防尘接地等要求，需使用室外款型 AP，可根据覆盖区域终端接入密集程度选择双频或三频室外 AP。

6.1.3 室内 AP 安装规范

AP 安装于室内时，必须遵从以下原则：

(1) 安装位置必须保证没有强电、强磁和强腐蚀性设备，AP 应至少离开此类设备 2～3 米以避免干扰。

(2) 安装位置温度、湿度不能超过主机工作温度、湿度的范围。

(3) AP 安装时必须牢固固定，不允许悬空放置。

(4) AP 及天线要运营商基站天线或 4G/5G 基站天线，距离至少 5 米

(5) 在宿舍和教室内可采用壁挂和吸顶的方式，如下图所示。

图28 室内 AP 的一般安装方式

6.1.4 室外 AP 安装规范

AP 安装于室外时，必须遵从以下原则：

(1) 室外硬件安装所涉及的建筑墙体和放置装置需坚固完整。抱杆安装如下图所示。

图29 室外 AP 抱杆安装方式

(2) 室外设备安装必须做好防雷、防水处理。室外施工需具有附加的防雷装置，如避雷针、地桩、地网、接地排等。AP 表面要垂直于水平面，未接线的出线孔应用防水塞封堵，各接头处应做好严格的防水密封措施。

(3) AP 如放置于防水箱内，箱体必须固定牢固，保持垂直。箱体要保持通风以利于设备的散热。箱体可安装在楼顶墙体的内立面上，起到遮光挡雨的作用。进入防水箱的全部线缆需做防水弯，或采用下走线方式。防水箱也可抱箍固定在用于安装天线的抱杆上。

(4) AP 及天线要远离 3G/4G/LTE 的天线，距离至少 5 米。

6.1.5 AP 电源安装规范

AP 供电采用 PoE 和本地供电两种方式，优先采用本地供电。若本地供电存在困难，可以采用带 PoE 功能的以太网交换机进行供电；大功率 AP 应采用本地供电。

1. PoE 供电

如果上层交换机为以太网供电交换机，则不需增加 PoE 供电模块，直接用网线对 AP 设备进行远端供电。

如果上层交换机为普通交换机，则需增加 PoE 供电模块，对 AP 进行供电，原则上不允许串接 PoE 供电器。

2. 本地供电

AP 采用交流供电，电源要求为 220V±10V，50Hz±2.5Hz 波形失真小于 5%。对不满足要求的电源，应增加稳压设备。

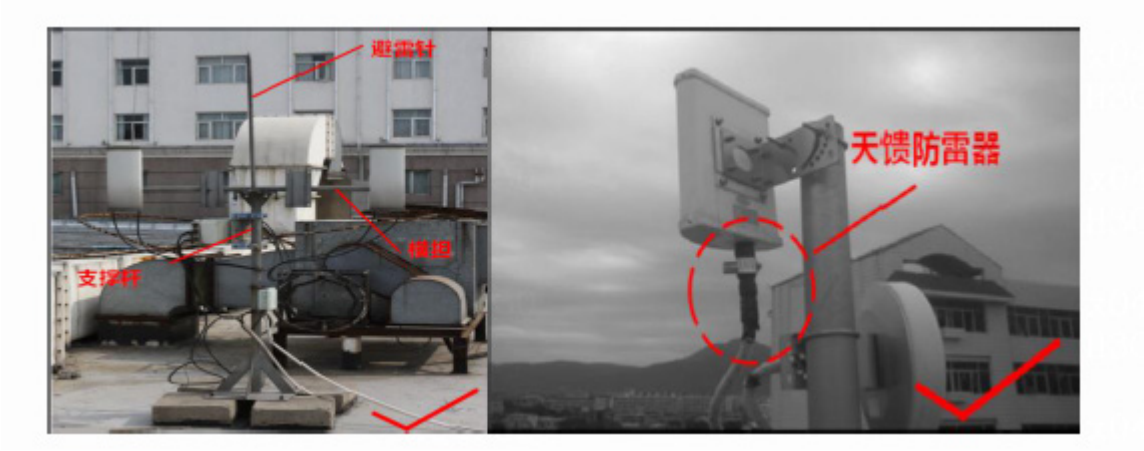
本地供电交流电源插座应采用有保护地线（PE）的单相三线电源插座，且保护地线（PE）可靠接地。

6.1.6 天线安装规范

1. 安装天线时应遵循以下原则

- 各类型天线支架应结实牢固，支撑杆要保持垂直，横担要保持水平，天线实际安装位置、型号应符合工程设计方案要求。
- 天线支架安装位置如高于楼顶，必须安装避雷针，避雷针长度符合避雷要求，并可靠接地。天线顶端要低于避雷针，且处于 45°避雷保护角范围之内。天线支架安装位置如在建筑物屋檐下或外墙低矮处时，天线支架不必安装避雷针。
- 室外天线必须安装天馈防雷器。定向天线的方位角和俯仰角可以根据覆盖目标进行微调，以满足信号覆盖的要求。

图30 室外天线安装示意图



2. 安装全向室外天线需要注意以下事项

(1) 全向室外天线抱杆直径要求 35mm～50mm，一般采用直径为 50mm 的圆钢制作抱杆。

(2) 在抱杆上安装全向室外天线后需保证抱杆顶端与天线下部的抱箍部分平齐，如图 31 所示。安装完成后天线高度需满足信号覆盖需求，并且天线顶端需处于避雷针 45 度防雷保护角之内。

(3) 安装全向天线时，一般不允许直接在抱杆上焊接避雷针（全向天线体的水平方向 1 米范围内不允许有金属体存在），而是在两根全向天线抱杆中间位置单独设置一根避雷针，避雷针的高度要使全向天线顶端处在其防护角之内。

(4) 由于环境限制使避雷针无法单独制作时，可以采用如图 32 所示的方法进行安装，但要求避雷针距天线抱杆 1 米以上。右图中的角钢用于固定天线抱杆，图示没有采用将天线抱杆固定在水泥墩上的方法，而是将角钢的一端焊接到避雷针的柱子上，另一端焊接到抱杆上来固定天线抱杆。

图31 全向天线安装示意图

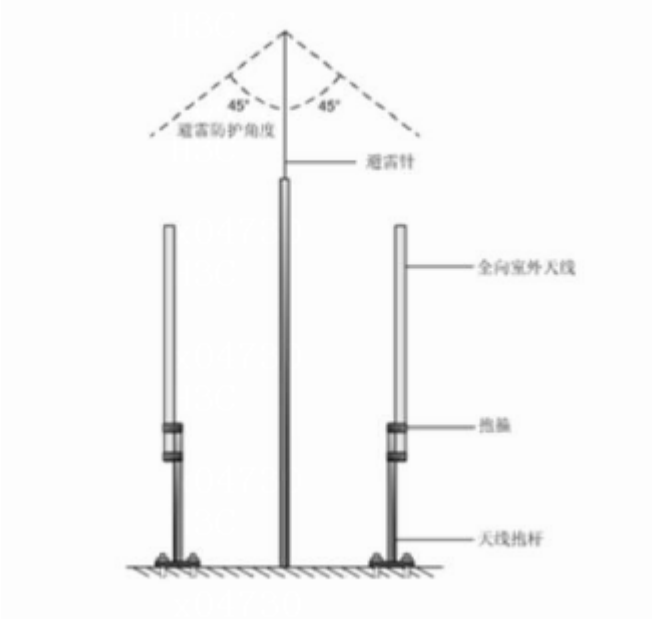
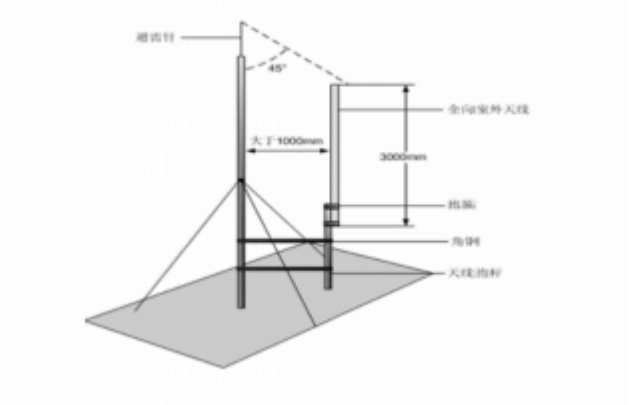


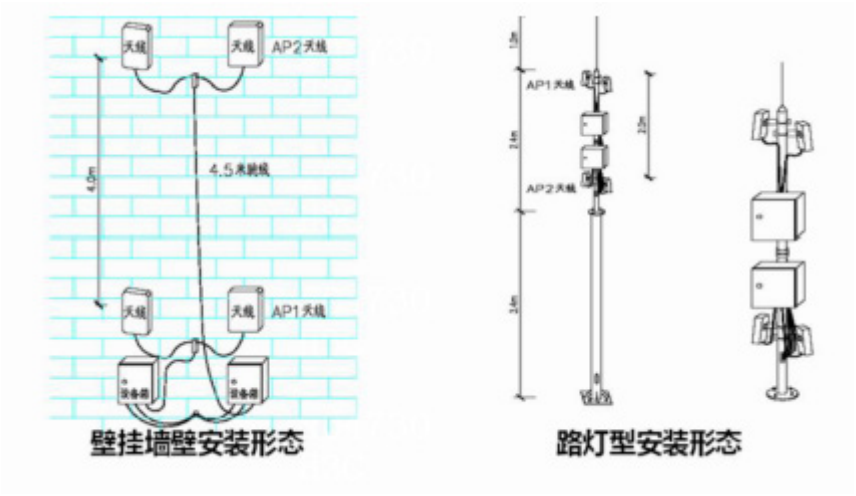
图32 特殊情况下全向天线安装示意图



在进行多个 AP 的安装时，各个 AP 的天线之间应该保持一定距离的间隔，以避免邻道干扰。

如图 33 所示，在壁挂墙壁的安装形态下，不同 AP 的天线位于上下两层，且上下层天线间垂直距离应隔开 4 米之外。在路灯型安装形态下，一根灯杆安装 2 台 AP，4 副天线，考虑到安装位置的限制，各 AP 天线间应保持垂直距离至少大于 2 米，可将机箱位于两 AP 天线之间以增大天线间的距离。

图33 不同类型天线安装形态



6.1.7 馈线安装规范

1. 馈线安装时应遵守以下规范：

(1) 馈线必须按照设计方案（文件）的要求布放，要求走线牢固、美观，不得有交叉、扭曲、裂损情况。

(2) 馈线的套管均推荐使用铁管、普利卡管、PVC 管。

(3) 加套 PVC 管或者（铁管）的馈线水平/垂直走线固定间距为 1 米，未加套管的馈线水平/垂直走线固定间距为 0.8 米，推荐使用镀锌铁管，拐弯处可以使用普利卡管。

(4) 馈线转弯半径：7/8 馈线大于 120mm，1/2 馈线大于 70mm，8D 馈线大于 50mm。

(5) 室外馈线加套 PVC 管，水平布线的 PVC 管每 6 米在 PVC 管下方必须切口，作漏水口。

(6) 馈线避免与消防管道及强电高压管道一起布放走线，确保无强电、强磁的干扰。

图34 馈线安装实景图

2. 馈线接头与 AP、天馈防雷器、天线、耦合器等连接口连接时

(1) 距离馈线接头必须保持 50mm 长的馈线为直出，方可转弯。

(2) 必须连接可靠，接头进丝顺畅，不得野蛮强扭。

(3) 室外馈线接头必须进行密封，步骤如下：用电工（绝缘）胶布包裹接头金属部分打底。用防水胶布包裹电工（绝缘）胶布，并保证完全密封。再用电工（绝缘）胶布严密包裹防水胶布。

(4) 室内馈线接头只需用电工胶布包裹作防尘处理。

3. 接地安装规范

(1) 室内 AP 应接地，室外 AP、天馈防雷器、PoE 模块必须接地，接地电阻小于 5 欧姆，接地线严禁超过 30 米。

(2) 多股地线（一般不使用单芯）与地排连接时，必须加装接地端子（铜鼻），接线端子尺寸应与线径吻合，压（焊）接牢固。

(3) 接地端子与地排的接触部分应平整、紧固，无锈蚀、氧化，不同材料连接时应涂凡士林或者黄油防锈。

(4) 加套 PVC 管的地线固定原则与射频走线相同。加装线槽时，线槽固定间距为 0.3 米。地线的曲率半径应大于 130mm。

6.1.8 AP 信息录入规范

(1) 在扫描 AP 序列号之前对 AP 进行命名，使用 EXCEL 表格保存，并将命名标签使用标签打印机打印。

(2) 使用扫码枪或 Cloudnet APP 对 AP 信息进行扫描录入，需做到名称和 MAC 地址，序列号一一对应，录入一个贴一个标签，避免出现贴错的情况。

(3) 贴好标签的 AP 再放入包装箱中，保证安装前标签不会被外力磨损或液体侵蚀。

(4) 将已经录入好的 AP 信息与之前的 AP 名称表格比对整理，保证录入正确。

(5) AP 安装后需要完善 AP 信息，AP 信息表中应有该 AP 的名称、序列号、MAC 地址、型号、连接的交换机名称与端口和所在具体的物理位置等信息。

6.2 WLAN网络架构设计

6.2.1 无线设备部署设计

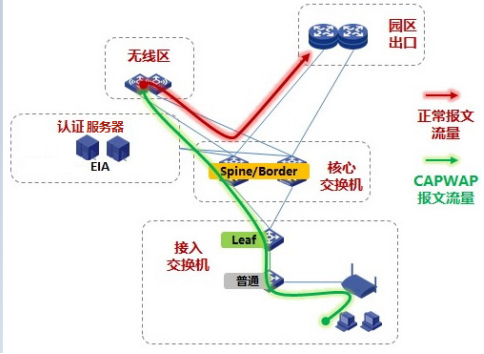
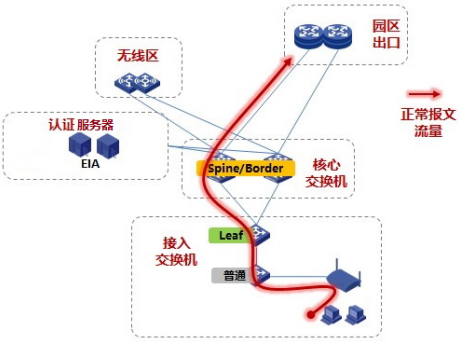
表1 无线设备部署位置

产品	部署位置说明
AC	无线AC一般旁挂在核心交换机上。 对于组网为“Spine+Leaf+Access交换机”的网络，AC旁挂在Spine交换机上；对于组网为“Leaf+Access交换机” 的网络，AC旁挂在Leaf交换机上。 无线AC N+1的环境，多台无线AC旁挂在Spine交换机上。
AP	无线AP可以挂接在Leaf交换机下，也可以挂接在Access交换机下。

6.2.2 无线转发模式设计

无线设备的组网规划分为“无线集中转发”和“无线本地转发”，一般推荐使用无线本地转发，可以充分利用交换机的转发能力。

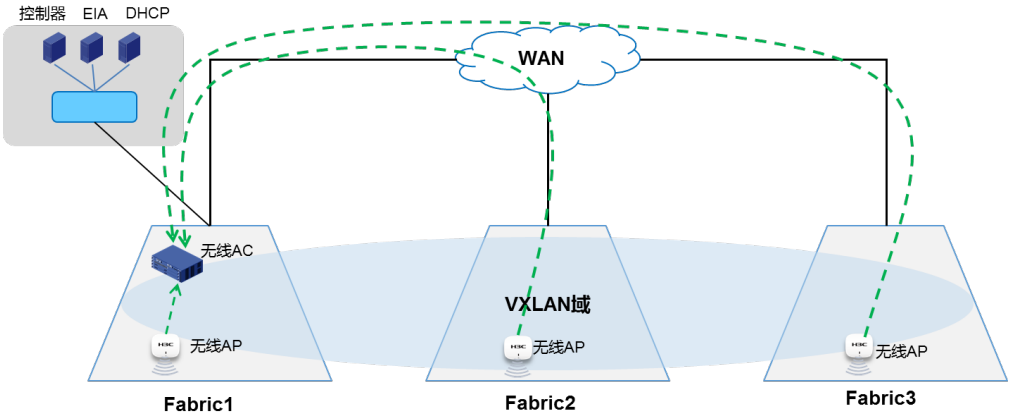
表2 无线设备组网规划说明

模式	无线集中转发	无线本地转发
规划说明	无线用户的流量都经过AC，无线用户的网关部署在连接AC的Spine交换机上	无线用户的流量不经过AC，无线用户的网关部署在连接着相关AP的Leaf交换机上
流量模型	 所有的无线用户的流量，都是先到AC，再到目标网络	 无线用户的流量，进入AP之后，就从AP所在的有线网络转发
用户的认证节点	用户的认证点在AC上	用户的认证点在AC上
策略下发位置	连接AP的Leaf交换机和连接AC的Spine交换机都下发策略，但匹配的是Spine交换机上的策略。	连接AP的Leaf交换机和连接AC的Spine交换机都下发策略，但匹配的是Leaf交换机上的策略。
注意事项	此时无线用户不占用Leaf和Spine交换机的认证用户数，但会占用Spine交换机的ARP/ND表项	此时无线用户不占用Leaf交换机的认证用户数，但会占用Leaf交换机的ARP/ND表项
AD-Campus的功能	支持网随人动 支持策略随行	支持网随人动 支持策略随行

对于多 Fabric 模型，存在两种情况：

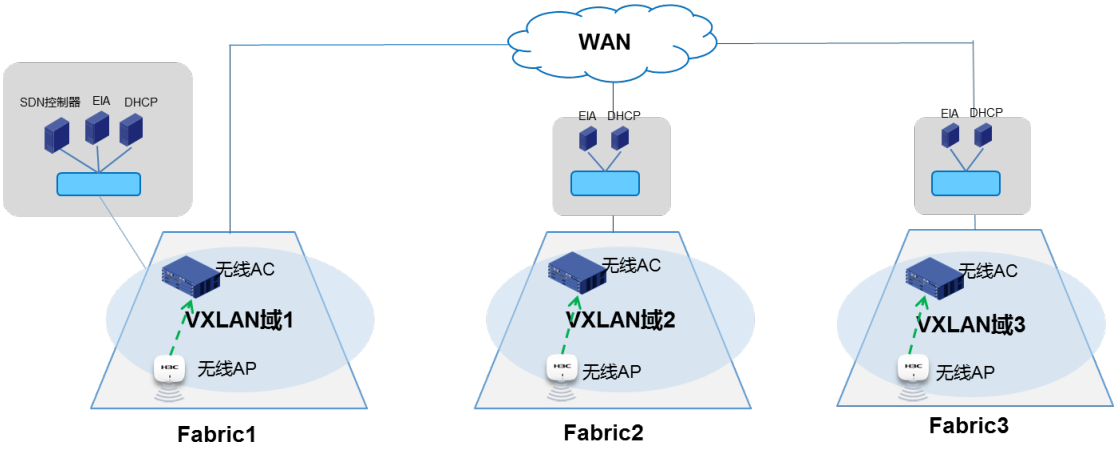
- 整网集中部署 AC 于某个 Fabric，其他 Fabric 的 AP 直接到该集中式 AC 注册和接受管理，整网一个无线管理域。这种情况推荐使用 AP 本地转发，可避免无线的流量跨广域到集中式 AC 来交换，属于方案的标准模型。

图35 整网集中部署 AC 于某个 Fabric



- 每个 Fabric 一套 AC/AP，要求各 Fabric 的 AP 只能向各 Fabric 的 AC 注册和接受管理，相当于多个无线管理域。

图36 每个 Fabric 一套 AC/AP



6.3 无线AP管理设计

无线 AC 最主要的功能之一就是 AP 管理。

无线 AC 和无线 AP 之间 IP 可达才能实现无线 AP 自动上线无线 AC，其中可以是二层网络，也可以是三层网络。在 AD-Campus 方案中，VXLAN 网络中默认使用 VLAN4093 进行 AP 管理，控制组件会进行无线管理网的部署，将无线 AC 和无线 AP 之间的网络打通。

无线 AP 自动化上线到无线 AC 是通过 DHCP option43 得到无线 AC 的地址，无线 AP 会向这个地址发送上线请求，最终通过这个地址与无线 AC 建立 CAPWAP 连接。这一系列自动化上线过程，需要先确保将无线 AC 和无线 AP 之间的链路放通，连接 AP 的 Access 下行口默认设置 VLAN4093，VLAN4093 的 AP 在 Leaf 上线免认证，直接进入 VXLAN4093。Spine 上旁挂无线 AC 的端口也会设置无线 AC 的 VLAN4093 进入 VXLAN4093。

由于无线 AP 数量庞大，可以通过划分 AP 组来管理 AP，比如同一楼层的 AP 为一个 AP 组。针对这一 AP 组,其中的 AP 可以继承 AP 组的配置，当某个 AP 比较特殊时，在 AP 上进行的配置，生效优先级高于 AP 组的配置，将会以 AP 上的配置为准执行。

在 AD-Campus 方案中，推荐使用 WSM 组件进行无线 AP 管理、无线服务的配置。

WSM 组件可以融合部署在统一数字底座上，实现统一登录，统一管理。

6.4 无线服务设计

SSID 需要根据业务来规划，一般分为内网用户 SSID、内网哑终端 SSID 和访客 SSID，这些不同无线服务之间用不同业务 VLAN 隔离开。

- 内网用户 SSID：一般针对内网使用无线移动终端的用户，均为智能移动终端，有 web 操作页面。多采用 mac portal 认证和 1x 认证。
- 内网哑终端 SSID：一般针对内网打印机、摄像头等无线终端，这类终端一般很少主动发包，只有在要提供服务的时候，或者 ARP 探测的时候，才会回应报文。多采用 MAC 认证或免认证。
- 访客 SSID：一般针对临时访客，可以通过扫描二维码、邮件审核等方式通过认证。多采用 portal 认证。

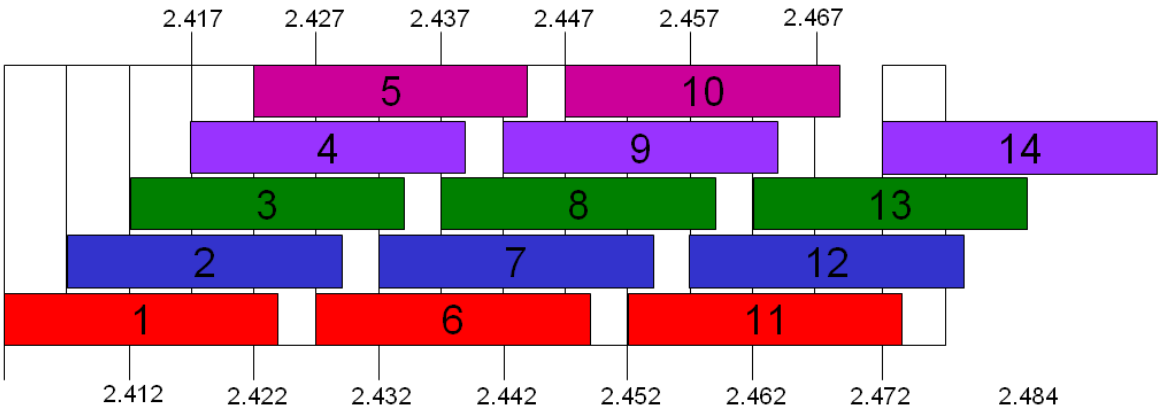
6.5 无线射频规划

无线是通过电磁波传输信号，存在干扰，因此对于射频的规划设计是尤其重要，需要实现对整网的射频进行规划，可以通过 WSM 管理软件来对整网的射频进行配置和管理。

6.5.1 2.4G 信道设计

802.11 协议在 2.4GHz 频段定义了 14 个信道，在北美地区（美国、加拿大）开放 1-11 信道，在欧洲开放 1-13 信道，如图。在中国，与欧洲一样，同样开放 1-13 信道。

图37 信道



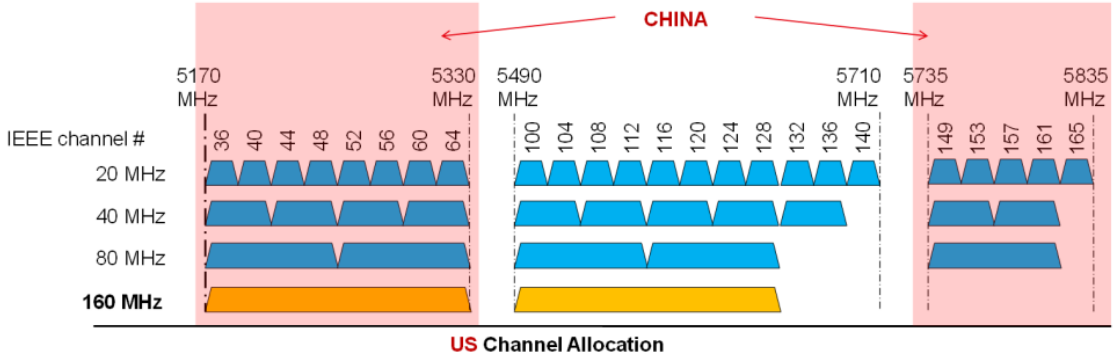
如图所示，信道 1 在频谱上和信道 2、3、4、5 都有交叠的地方，这就意味着：如果有两个无线设备同时工作，且它们工作的信道分别为 1 和 3，则它们发送出来的信号会互相干扰。

为了最大程度的利用频段资源，可以使用 1、6、11。2、7、12。3、8，13。4、9、14 这四组互相不干扰的信道来进行无线覆盖。由于只有部分国家开放了 12~14 信道频段，所以一般情况下，使用 1、6、11 三个信道。

6.5.2 5G 信道设计

随着技术的发展，11n 和 11ac 的设备和终端成为潮流，当今大多数终端都能支持 5G 频段。相比于 2.4G，5G 射频资源更加丰富，2012 年 10 月 31 日，中国无线电管理委员会正式发布了新的 WiFi 5GHz 授权频段 5150~5350MHz（Channel36~64），加上原有的 5750~5835MHz 频段（Channel149~165），5G 射频一共有 13 互不干扰的 20M 频宽信道。

图38 5G 工作频段划分图



6.5.3 频宽设置原则

对于 2.4G，建议保持默认频宽（20M），对于 5G 信道，对于独立密封环境，可以保持默认 80M 频宽，对于高密开放场景，推荐使用 40M 频宽或者 20M 频宽。如果采用雷达信道，需要关闭雷达信道避让功能。在 40M 或 80M 频宽模式下，165 信道无法向前绑定信道，因此如果信道配置为 165，频宽为 40 或 80，实际信道认为 20M。

表3 频宽设置

频宽	可配信道
20M	36,40,44,48,52,56,60,64
40M	36,44,52,60
80M	36,52,149

6.5.4 信道设置原则

5G 频段具有速率高、可用信道丰富（13 个互不干扰的 20M 频宽信道）、穿透能力相对 2.4G 弱等特点，相对应用更为广泛，部署时建议 5G 信道使用 40M 频宽，对于封闭较好场景可以采用 80M 频宽（注意高低频搭配使用）。具体信道规划原则如下：

任意相邻区域使用无频率交叉的频道，如：1、6、11 频道。

适当调整发射功率，避免跨区域同频干扰。

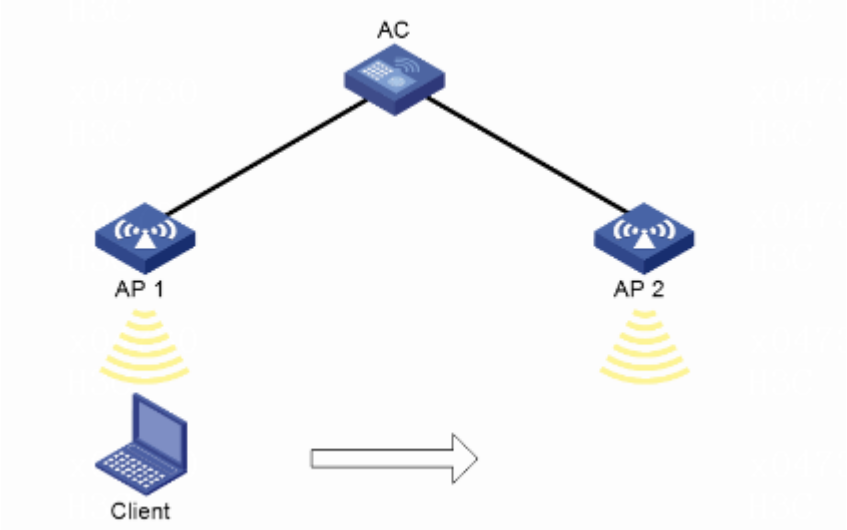
蜂窝式无线覆盖实现无交叉频率重复使用。

6.6 无线漫游规划

在同属于一个 ESS（Extended Service Set，拓展服务集）区域中的不同 AP 覆盖范围内，无线客户端从一个 AP 上接入转移到另一个 AP 上接入的过程称为漫游。在漫游期间，客户端的 IP 地址、授权信息等维持不变， 在 WLAN 网络中，无线终端具备有移动通信的能力，但由于单个 AP 设备的信号覆盖范围都是有限的，终端用户在移动过程中，往往会出现从一个 AP 覆盖范围跨越到另一个 AP 覆盖范围的情况。因此保障用户在不同 AP 间移动和切换接入时的良好的业务体验是 WLAN 网络质量的关键指标。

6.6.1 二层漫游

图39 二层漫游



如图所示，客户端漫游的具体过程如下：

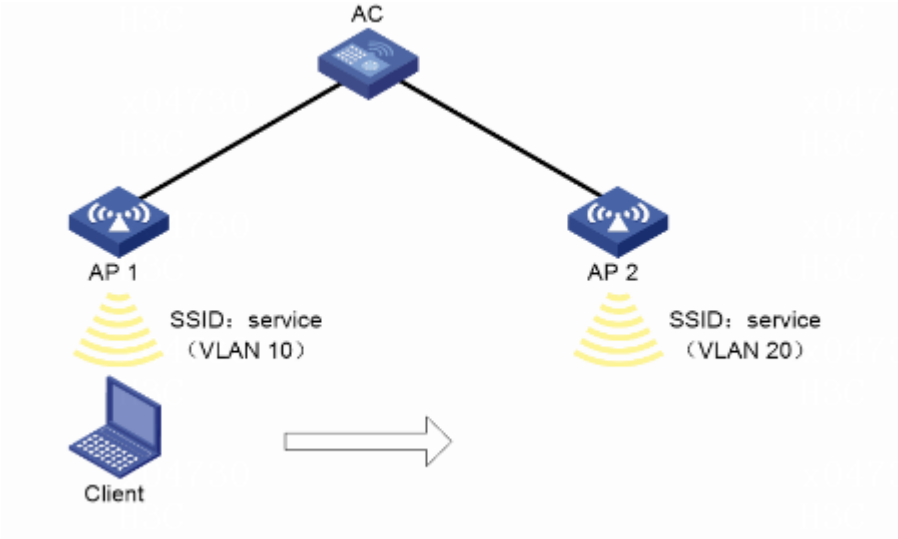
(1) 客户端在 AP 1 上初始上线，在 AC 上会创建该客户端的漫游表项信息（漫游表项信息主要包括客户端上线 SSID、PMKID、认证方式、安全认证模式以及漫游 VLAN 等）。

(2) 客户端漫游到 AP 2，AC 查找该客户端的漫游表项。

(3) 客户端重新认证，在 AP 2 上上线。

6.6.2 三层漫游

图40 三层漫游



当客户端从 AP 1 漫游到 AP 2 时，设备无需任何特殊配置，即可完成跨 VLAN 的漫游。

6.6.3 漫游增强技术

WLAN 漫游增强技术目前包括以下几种：

- 802.1X 快速漫游：当客户端认证方式为 RSN+802.1X 认证，可以进行 802.1X 快速漫游，客户端不需要再次认证即可完成漫游上线。
- MAC 快速漫游：当客户端认证方式为 MAC 地址认证时，可以进行 MAC 快速漫游，客户端不需要再次认证即可完成漫游上线。
- 802.11r：用来缩短无线客户端在漫游过程中的时间延迟，从而降低无线客户端连接中断率，提高漫游服务质量。
- 虚拟 BSS 漫游：AC 通过实时监控无线客户端信号强度，使客户端接入服务质量更好的 AP，实现客户端在一个 ESS（Extended Service Set，拓展服务集）区域中的无缝漫游。
- 协同漫游：结合 IEEE802.11k、IEEE802.11r 和 IEEE802.11v 协议实现无线客户端在一个 ESS 区域中的无缝漫游。

表4 不同方式下的漫游时延

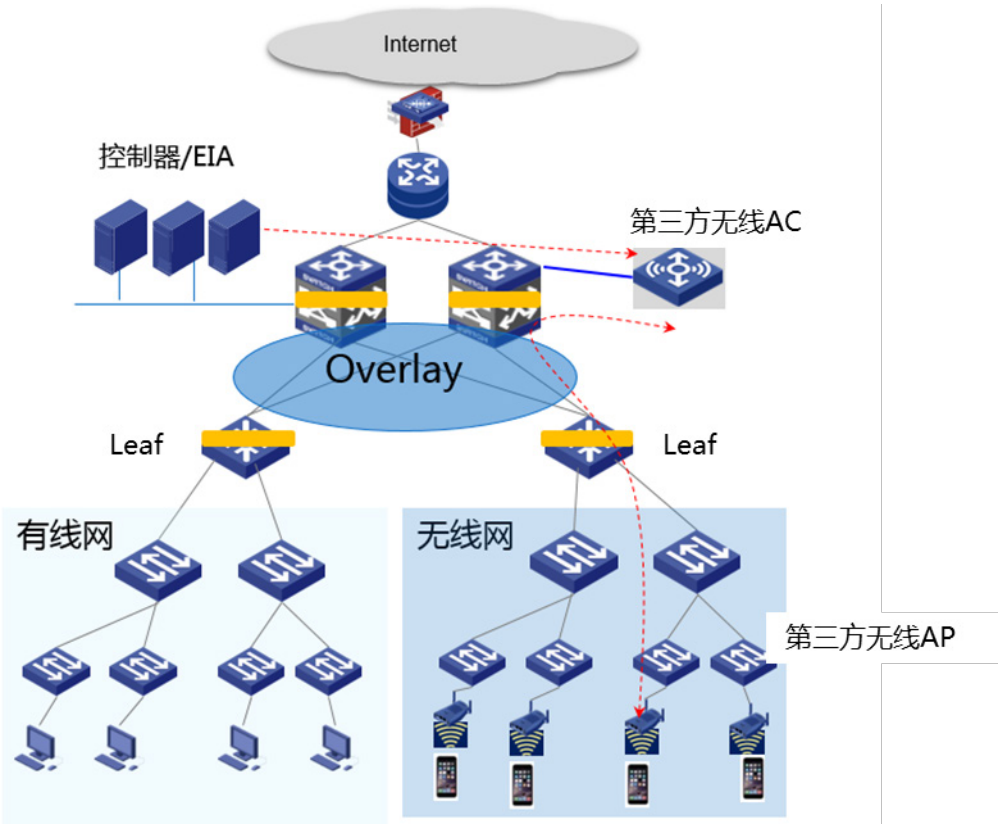
漫游方式	时间	建议	说明
802.1X快速漫游	<100ms	系统自动生效该功能，无需配置	PMK快速漫游需要终端也支持该能力，目前几乎所有终端都具备该能力，兼容性要好于802.11r快速漫游功能。
MAC快速漫游	<50ms	通过mac-authentication fast-connect enable	已经通过MAC地址认证的客户端在AC内漫游时，不需要再次进行MAC地址认证，可提高客户端AC内漫游的上线速度

		开启	
802.11r	<50ms	如果客户不需要启用PMF功能，建议开启802.11r快速漫游功	• 802.11r 快速漫游需要终端也支持该能力，目前主流机型支持较多。 • 802.11r 快速漫游与 PMF（Protected ManagementFrame：管理帧保护 ）功能互斥，即如果配置了 802.11r 快速漫游，则不能再配置 PMF 功能。
虚拟BSS漫游	<50ms	通过配置seamless-roaming enable开启虚拟BSS漫游	在医疗场景中使用
协同漫游	<50ms	IEEE802.11k、IEEE802.11r和IEEE802.11v协议的一种漫游技术，由AP和无线客户端协同引导，实现了无线客户端在一个ESS（Extended Service Set，拓展服务集）区域中的无缝漫游	协同漫游目前仅支持同一AC内的漫游组网方式，且协同漫游要求AP必须支持Wi-Fi 6标准

由于无线漫游的体验和终端强相关，在实际使用中，漫游和终端关系较多，为了更好的体验漫游效果，建议部署 AD-Campus 智能分析组件，利用 AD-Campus 智能分析组件的大数据智能分析系统，通过信息诱导，智能通知，主动诱导，弱信号拒绝等多种手段，来实现用户的无感知漫游体验。

6.7 第三方无线对接设计

图41 无线方案整体设计



(1) **组网说明：**采用 AD-Campus Spine-Leaf-Access 三层典型组网，第三方无线 AC 旁挂在 Spine，第三方无线 AP 连接到 Access 交换机（POE 款型），使用 EIA 做 Radius server。有线的认证点在 leaf，无线的认证点根据认证方式不同，要求不同：

- 802.1X 认证，要求认证点只能在第三方无线 AC，转发方式根据第三方无线的支持情况设置，建议使用 AP 本地转发。
- MAC 认证，认证点可以在第三方无线 AC，也可在 Leaf 交换机。如果在 Leaf 交换机认证，需要在第三方无线 AC 上设置无线服务免认证，并且采用 AP 本地转发方式。
- MAC Portal+认证，认证点只能在 Leaf 交换机，此时需要在第三方无线 AC 上设置无线服务免认证，并且采用 AP 本地转发方式。
- Portal 认证，认证点只能在第三方无线 AC，转发方式根据第三方无线的支持情况设置。访客一般使用此种认证方式。

重要：如果采用本地转发在 Leaf 认证，需要充分评估 Leaf 的认证承载能力，避免超过 Leaf 能力影响用户上线。

(2) **AP/AC 管理通道建立：**AP 和 AC 之间通过 vxlan4093 进行 capwap 管理通道建立，通过控制组件在 leaf 上进行创建 Vxlan4093 作为一个管理类二层网络域。Leaf 上通过将 vxlan4093 和 VLAN 4093 关联，POE access 交换机上，手工将连接第三方无线 AP 的端口配置 PVID 4093/Trunk all，这样 AP 的 capwap 管理报文就会携带 VLAN 4094 的 tag 到达 leaf，并映射

到 vxlan4093 中；第三方无线 AC 连接 spine 的端口配置 VLAN 4093，并配置管理 IP 地址；同时在 spine 入口关联到 vxlan4093。这样，AP 和 AC 之间的 vxlan4093 通道就建立起来。

(3) **AP 注册上线过程：**控制组件在创建 vxlan 4093 的二层网络域时，添加 option 43 选项，在选项中填入第三方无线 AC 的 VLAN 4093 的管理 IP 地址，这些操作最终会在外置的 windows DHCP server 上创建一个作用域，并在该作用域中携带 option 43 选项。AP 加电启动时，首先在 vxlan 4093 内通过 DHCP server 进行管理 IP 地址的申请，获得管理 Ip 地址的同时，也会从 option 43 中获得 AC 的管理 IP 地址。之后 AP 会使用 AC 的管理 IP 地址向 AC 发起注册，经过多次交互之后，双方建立 capwap 管理隧道，AP 上线完成。

(4) **AC 认证授权过程：**首先在第三方无线 AC 上创建 VLAN 4094，并配置另外一个业务 IP 地址，该地址用于 radius 报文的源地址，用于和 EIA 建立 radius 会话。第三方无线 AC 使用 1X/mac 认证时，可以直接使用 EIA 作为 radius server 进行兼容认证。目前已对接测试过 PEAP 方式。认证通过后，EIA 根据账号分组，给第三方无线 AC 授权业务 VLAN。

(5) **无线终端流量转发过程中：**在第三方 AC 对接模式下：

- 如果使用集中转发模式，无线终端流量通过 capwap 隧道到达第三方无线 AC，对 capwap 隧道解封装后，再打业务 VLAN 上行到 spine，在 spine 通过提前下发的业务 VLAN 到 vxlan 的映射，进入相应 vxlan 做 overlay 转发。同时 spine 上也会下发组间策略，对无线流量进行策略控制。
- 如果使用本地转发模式，无线终端流量在第三方无线 AP 打业务 VLAN 上行到 leaf，在 leaf 通过提前下发的业务 VLAN 到 vxlan 的映射，进入相应 vxlan 做 overlay 转发。同时 leaf 上也会下发组间策略，对无线流量进行策略控制。

(6) 对于有线终端的流量，是在 leaf 上进行转发和策略控制，由于组间策略是在 spine 和 leaf 同时下发，策略一致，所以保证了有线和第三方无线流量的策略统一控制。

备注：

- 第三方无线 AC 和 H3C 的 EIA 的对接测试考虑到每个用户的业务需求有所差异，在开局现场再根据细化的需求再做一次对接测试，以确保对接功能 OK。
- 第三方无线 AC/AP 的网管有各自的配置管理方式，具体请咨询第三方无线产品的技术支持工程师。

7 IPv6 部署设计

7.1 IPv6网络设计

IPv6 网络是在 IPv4 网络地址不足的情况下规划部署，则 IPv6 网络的部署一般考虑以下两种方式：

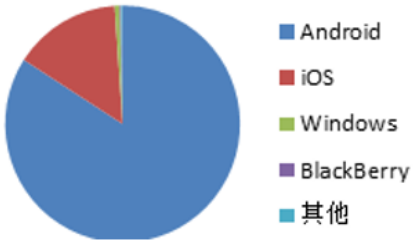
- 与 IPv4 共存一段时间作为过渡，双栈部署
- 去除 IPv4，仅保留 IPv6，为 IPv6 单栈网络

IPv6 地址的分配也存在多种方式，可以通过 DHCPv6 服务器自动分配，也可以通过网关无状态方式。多数终端两种方式都可以支持，但无线终端略有不同。

无线终端主流的操作系统是 IOS 系统和 Android 系统，这两种系统的终端获取 IPv6 地址方式见下图：

图42 获取方式

	IOS系统	Android系统
DHCPv6	支持	不支持
网关无状态	支持	支持
默认双栈	是	是
双栈下未获取IPv4地址时可获取IPv6地址	可获取	不可获取
协议栈选择	支持	支持



为兼容两种操作系统无线终端，无线 IPv6 部署限制：

- 需为无线终端同时提供 IPv4、IPv6 地址
- 无线终端通过网关无状态分配方式获取 IPv6 地址

对于采用 DHCPv6 服务器获取 IPv6 地址的终端来说，DHCPv6 服务器的部署也是网络设计重要的一部分。

7.2 双栈部署

7.2.1 网络设备上线

AD-Campus 方案 5 期 B03 及之前版本的设备自动化，只支持 underlay 为 IPv4 的情况，也就是设备上线后，必需要先获取 VLAN1 和 Loopback0 的 IPv4 地址，连接上控制组件之后，overlay 可以支持 IPv4/v6 部署，从而支持上层的 IPv4/v6 业务。

7.2.2 终端认证

AD-Campus 支持的认证方式包括：802.1x 认证、MAC 认证、MAC Portal+认证。

双栈场景下，通过 MAC 地址识别用户，当 IPv4 的认证通过，IPv6 的认证放行，即 IPv4 单栈认证，双栈放行。由于目前 V9 EIA 暂未支持 IPv6 认证，则不支持 IPv6 单栈认证通过，IPv4 放行的场景。

BRAS 上 IPoE 认证和 IPoE Web 认证支持单栈认证双栈通过。

1. 用户 802.1x 和 mac 认证

- 采用 IPv4 认证，设备配置 IPv4 和 IPv6 地址，用户双栈从 DHCP server 获取 IPv4 和 IPv6 地址，若 IPv6 获取方式为手动或 SLAAC（网关无状态自动配置）时，设备也可不配置 IPv6 地址。
- 采用 IPv6 地址，设备配置 IPv4 和 IPv6 地址，用户双栈从 DHCP server 获取 IPv4 和 IPv6 地址

2. 用户 mac-portal 认证方式

BYOD 二层网络域只能配置单栈，mac-portal 用户在 BYOD 中时只能是单栈的（IPv4 或 IPv6），后续进入业务安全组后可双栈或单栈。

表5 终端认证方式

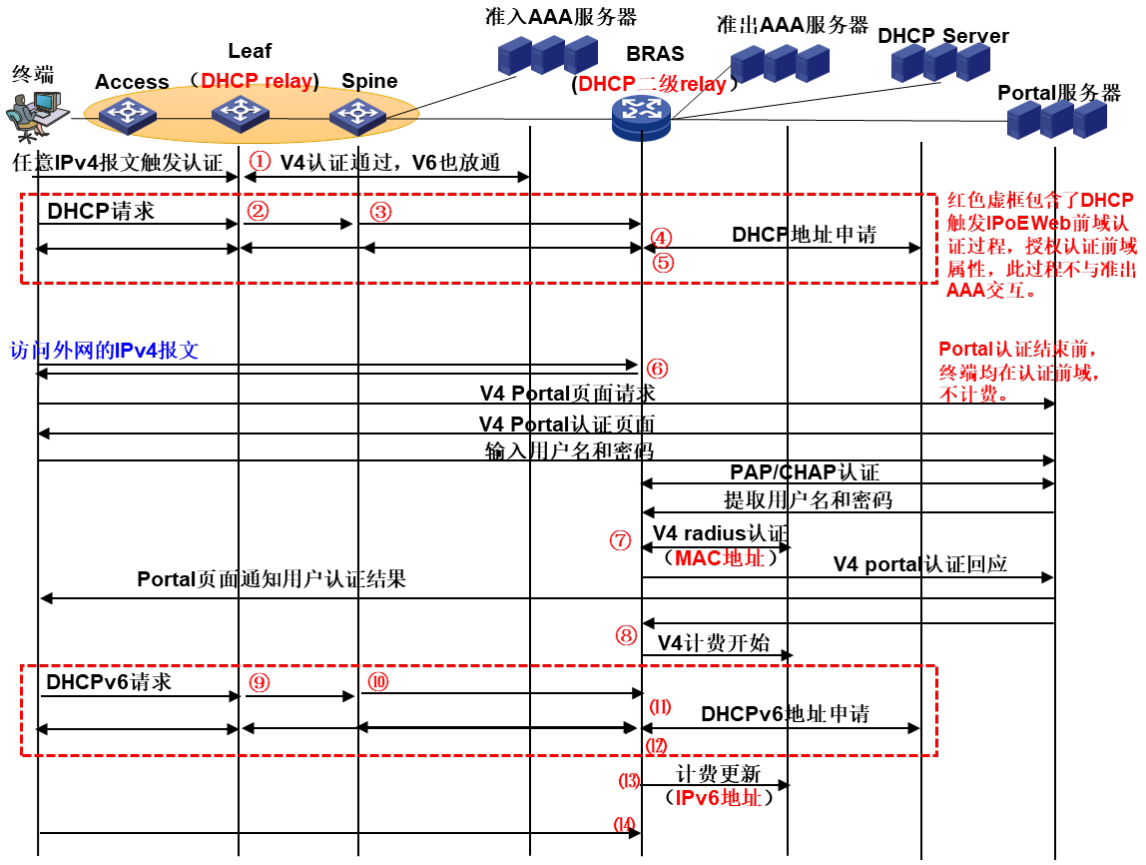
认证方式	实现	优势	适用场景
802.1x	不区分IPv4、IPv6地址	可通过客户端实现安全管理	对终端管理严格
MAC		简单，无需客户端	基于IP接入的物联网终端
MAC Portal	先通过IPv4认证，再申请IPv6地址 IPv6仅支持单IPv6协议栈	安全性适中，且无需安装客户端	通过网页进行准入认证，典型为手机终端

双栈场景下，暂不支持对用户的 IPv6 地址进行名址绑定。

3. 用户 IPoE Web 认证

用户 IPoE Web 认证适用于用户直接在准出 AAA 认证场景，单栈认证双栈通过。

图43 用户 IPoE Web 认证



以上图中过程为 MAC Portal+认证 BYOD 过程已经完成后的过程，BYOD 的 DHCP 报文不上送 BRAS，不参与 IPoE Web 认证过程。

以下为上述过程详细描述：

(1) 终端在 Leaf 完成 V4 准入认证（MAC 认证、1X 认证或 MAC Portal+认证），V4 认证通过，V6 也放通，无需 V6 再做准入认证。

(2) 终端 DHCP 申请 IPv4 地址，Leaf 添加 option61/79/82 将 DHCP 报文发送 DHCP Server。

(3) Spine 通过路由使 DHCPv4 请求报文转发到 BRAS，将 BRAS 发送的 DHCPv4 回应转发给 Leaf；

(4) BRAS 上 DHCP 请求触发 IPoE Web 认证前域认证，将 DHCP 报文路由转发给 DHCP Server，并通过 MQC 把 DHCPv4 回应报文重定向到 CPU 处理；

(5) DHCP 分配 IPv4 地址，IPoE Web 认证前域上线，授权认证前域 user-group。

(6) 匹配认证前域的 MQC，将 portal 服务器等服务器地址放行，http/https 报文重定向到 CPU 处理。portal 模块对用户访问的目的地址进行仿冒，应答并推送认证页面 url。

(7) 与准出 AAA 进行用户 portal 认证上线，携带用户名/密码给准出 AAA，将 MAC 地址做为 radius 属性带给准出 AAA，由准出 AAA 记录绑定关系。

(8) V4 Portal 认证通过后退出认证前域，准出 AAA 授权认证后域 user-group，生成后域用户表项。

(9) 终端 DHCP 申请 IPv6 地址，Leaf 添加 option79 将 DHCPv6 报文发送 DHCPv6 Server。

(10) Spine 通过路由使 DHCPv6 请求报文转发到 BRAS，将 BRAS 发送的 DHCPv6 回应报文转发给 Leaf；

(11) （如果 V6 免认证则没有后续认证过程）BRAS 上收到 DHCPv6 请求报文，记录 option79 中携带的 MAC 地址，触发 v6 协议栈补栈，将 DHCPv6 报文路由转发给 DHCPv6 Server，并通过 MQC 把 DHCPv6 重定向到 CPU 处理；

(12) DHCPv6 分配 IPv6 地址，BRAS 在用户表项中增加记录 ipv6 地址，ipv6 的权限继承 v4。

(13) 立刻发送计费更新，刷新 IPv6 地址和 IPv6 统计（一般 v6 与 v4 合并统计）。

(14) 客户端访问 IPv6 网页，bras 设备直接放行。（无需用户再输入用户名和密码）

7.3 纯IPv6单栈部署

7.3.1 网络设备上线

控制组件 6 期 B01 版本发布的支持 underlay ipv6 自动化（netconf）特性，为 UC 底座为 IPv6 的网络，可支持 underlay 为 IPv6 的自动化。限制：S5130-EI/HI 不支持。

7.3.2 终端认证

IPv6 单栈场景下，只能使用 V7 EIA 做认证，V9 EIA 暂未支持弹 IPv6 Portal 认证页面。

1. 用户 802.1x 和 mac 认证

采用 IPv6 地址，用户从 DHCP server 获取 IPv6 地址。

2. 用户 mac-portal 认证方式

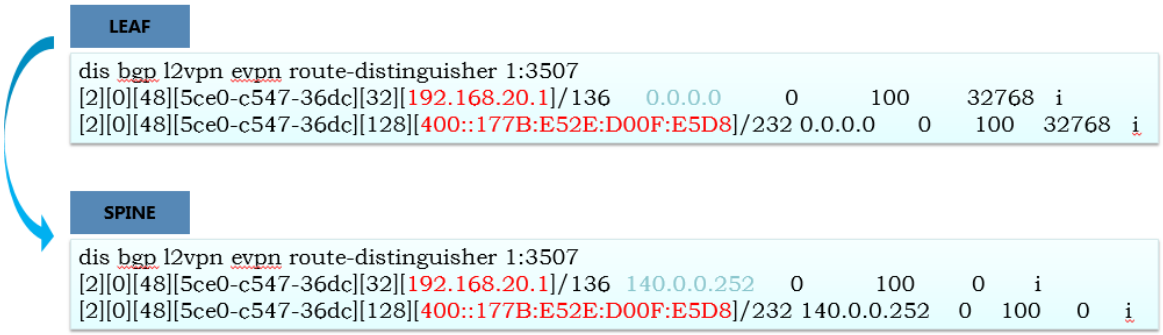
BYOD 二层网络域只能配置单栈，mac-portal 用户在 BYOD 中时只能是 IPv6 的，radius scheme 只能配一个 IPv6 的地址；acl 3001 只能配置 IPv6 规则。

IPv6 单栈场景下，暂不支持对用户的 IPv6 地址进行名址绑定。

7.4 IPv6路由学习

内网路由同步：园区网终端经过认证后，会在 LEAF 设备上上线，生成 IPv6 的 EVPN 路由信息。

图44 内网路由同步



Leaf 设备上终端的 128 位主机路由（ND）可以通过 EVPN 协议同步到 Spine 设备上，并加入到 IPv6 路由表中

7.5 DHCPv6 Server部署

DHCPv6 Server，可使用微软 DHCPv6 Server，也可以使用我司的 vDHCP，优缺点对比如下：

- vDHCP 要看用户数量，vDHCP 承载用户数量较少，且不能独立部署。
- Windows Server 做 DHCPv6 Server 时，不支持 HA。

其他 DHCPv6 Server 需要配套验证。

表6 终端 IPv6 支持度

类型	Windows 系统	MAC 系统	Andriod 系统	IOS 系统
DHCPv6	Y	Y	N	Y
网关无状态	Y	Y	Y	Y
默认双栈	Y	Y	Y	Y
ND RDNSS	Y	Y	Y	Y

备注：

- Android 系统需先获取 IPv4 地址，然后才能通过网关无状态方式获取 IPv6 地址
- ND RDNSS 指通过 ND 的 RA 报文通告 IPv6 DNS Server 对应 IPv6 地址
- 终端 IPv6 支持度来源于全球 IPv6 测试中心《2019 IPv6 支持度报告》

在交换芯片中，IPv6 主机路由表项占用资源是 IPv4 主机路由表项占用资源的 2 倍或 4 倍（取决于交换芯片），为了减少 IPv6 临时地址数量，建议网络 IPv6 地址分配方式：

- 有线业务，推荐 PC 采用 DHCPv6 申请地址
- 无线业务，推荐采用网关无状态地址分配方式：Android 手机不支持 DHCPv6

7.6 IPv6组间策略控制

IPv6 对硬件资源消耗几倍于 IPv4，在非微分段模式下启用 IPv6 的组间策略会让网络难以为继，所以建议在微分段模式下使用 IPv6 的组间策略能力。

策略矩阵的方式与 IPv4 没有区别（组策略模式）。策略生效范围可以选择“ipv4 用户”和“ipv6 用户”，根据采用双栈部署还是 IPv6 单栈部署来选择：

- 双栈部署时都选择，后续配置组间策略时，IPv4 和 IPv6 的组间策略都会下发设备并生效。
- IPv6 单栈部署时仅选择 IPv6 用户，后续配置组间策略时，只有 IPv6 的组间策略会下发设备并生效。

图45 配置组间策略

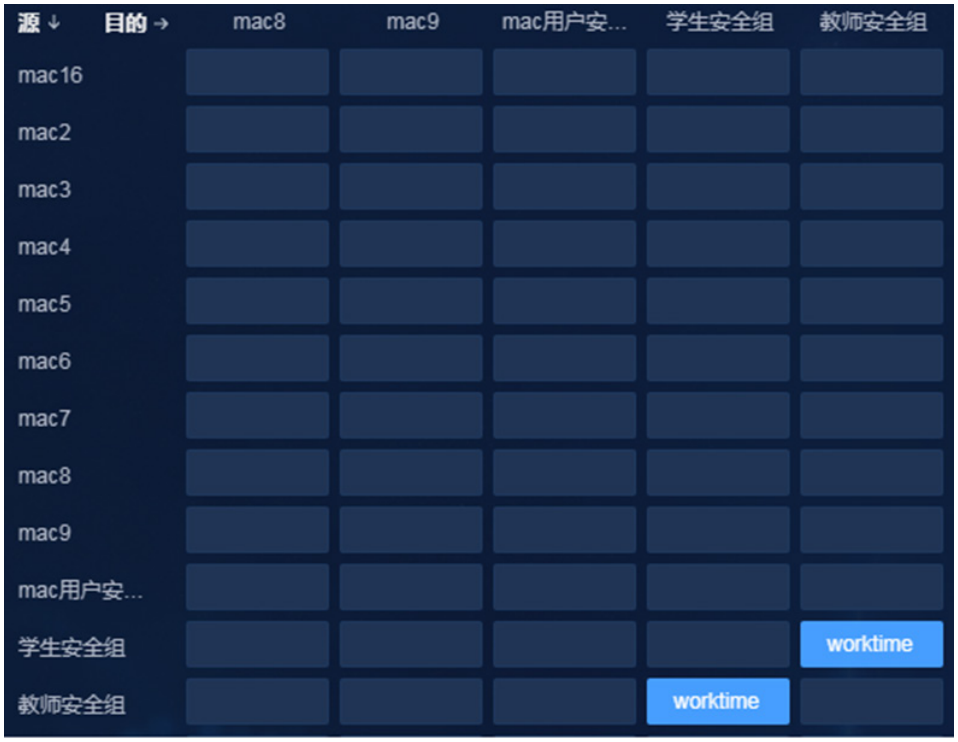


图46 策略详情

下发到设备上的 PBR 和 ACL。

图47 下发到设备上的 PBR 和 ACL 1

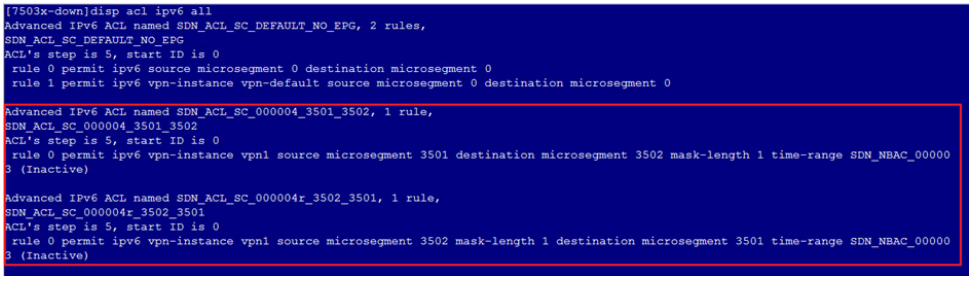


图48 下发到设备上的 PBR 和 ACL 2

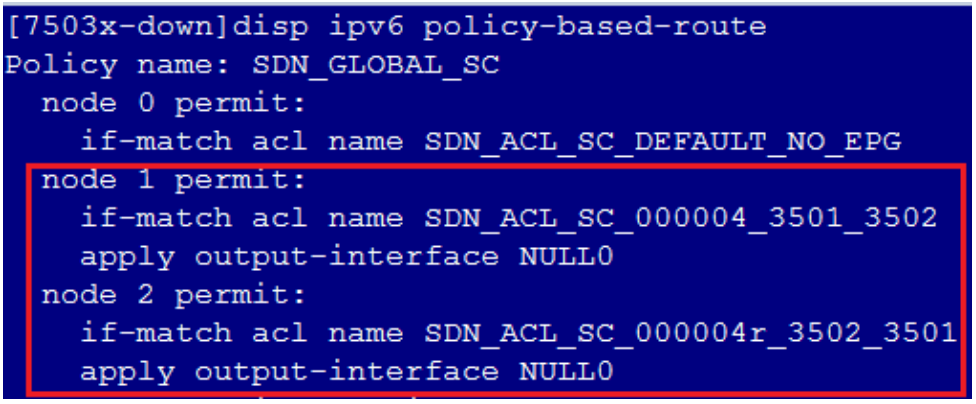
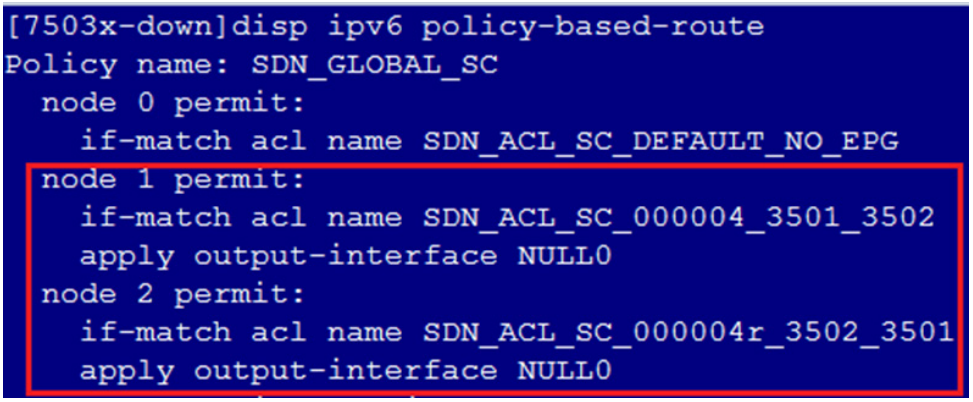


图49 下发到设备上的 PBR 和 ACL 3



7.7 IPv6网络智能运维

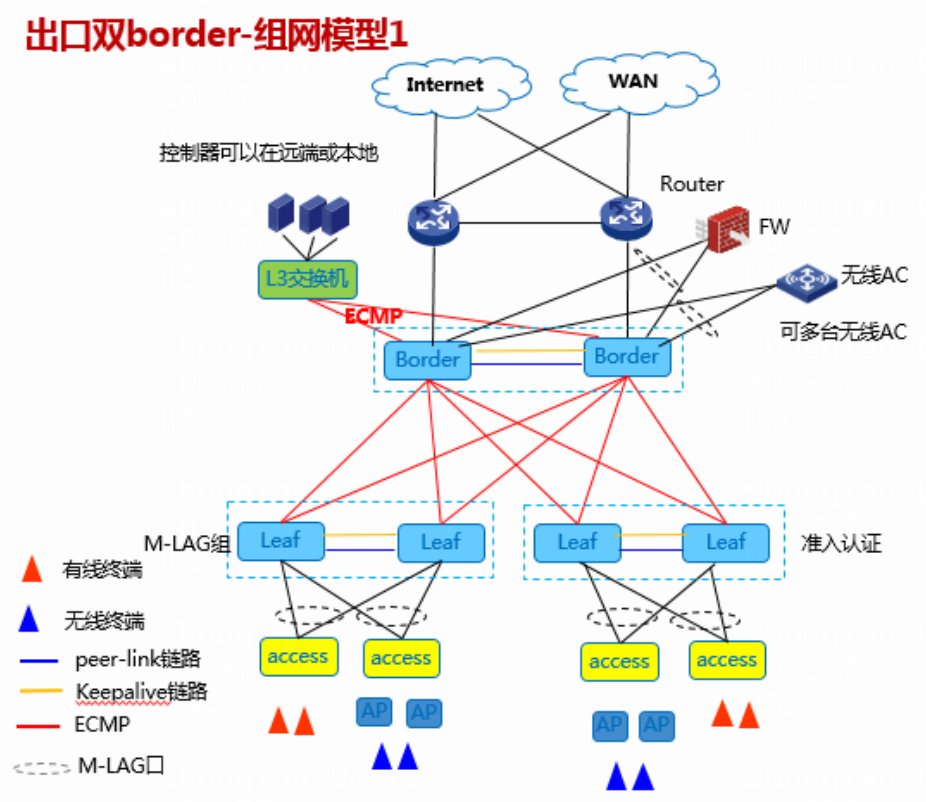
目前分析组件暂不支持 IPv6 单栈部署的智能运维。

双栈部署下，分析组件通过 IPv4 进行数据采集和分析。

8 出口网络设计

8.1 出口Border组网模型1

图50 出口双 Border 组网模型

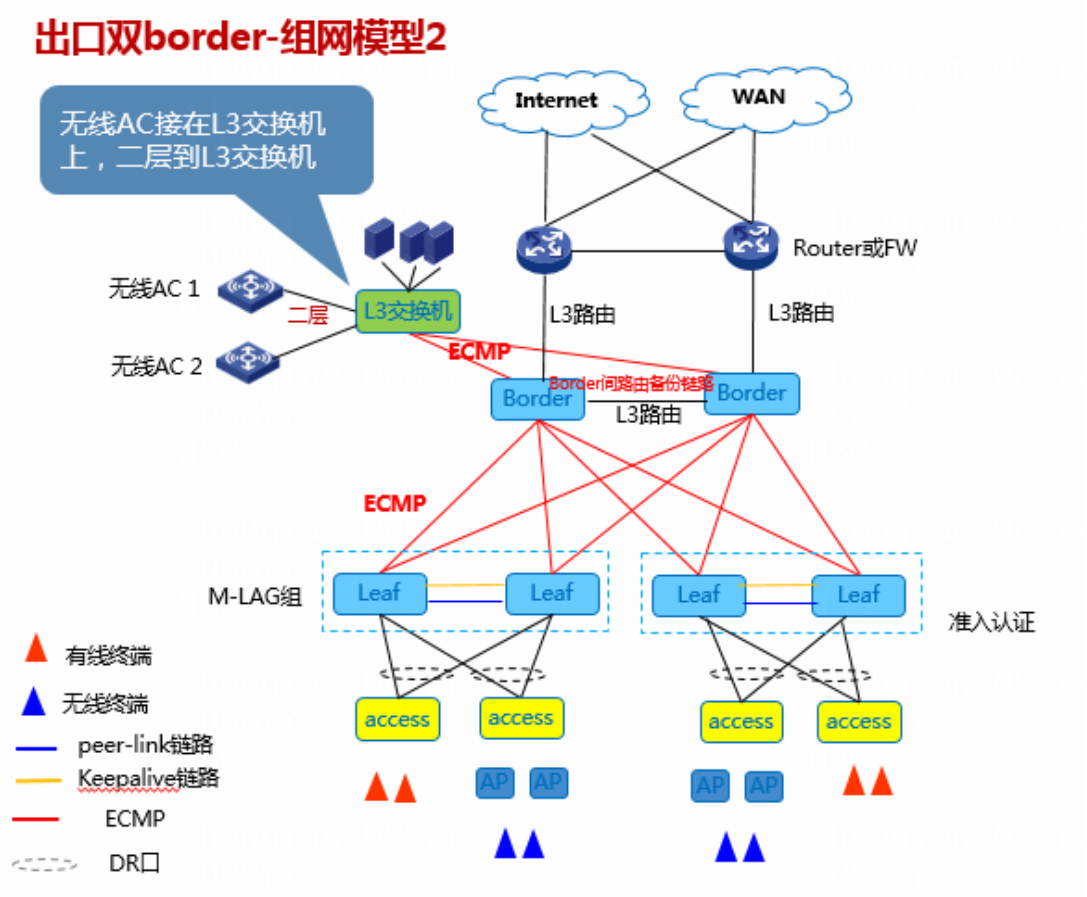


该方案 FW 采用单机旁挂部署，不串接。如果 Border 是双机堆叠方式，则 FW 使用链路聚合方式直接和 Border 互联；如果 Border 使用的是单机，则 Border 需要使能 M-LAG（Future），并通过链路聚合方式和 FW 互联，形成双臂连接方式。北向流量在 Border 上通过静态路由引流到 FW 的一个子接口，FW 处理完之后，再通过另外一个子接口送回到 Border。当 FW 出现故障时，可配置静态路由联动接口检测，实现快速 bypass，bypass 情况下，Border 查正常路由，直接从出口转发，不再绕行 FW。

也可以根据具体场景，考虑服务链配置方式，通过控制组件进行服务编排下发，提升使用体验。当前支持 Border 堆叠+FW 双臂部署模式（FW bypass 和 Border M-LAG 场景，future）

8.2 出口Border组网模型2

图51 组网模型



与模型 1 的主要不同点：

(1) **出口 FW/路由器可靠性设计**：采用串接/主备方式。FW 可采用 VRRP 方式或者主备路由方式和双 border 之间形成备份关系。当 FW 发生主备切换时，借助 VRRP 的主备切换或者路由备份机制，确保 Border 可以自动进行主备机切换。

(2) **无线 AC 可靠性设计**：目前首推采用 1+1 主备方式，AP 同时和两个 AC 之间建立主备连接，当主 AC down 掉，AP 重新连接备份 AC。避免 AC 之间产生耦合，提升整网的可靠性。此方案下，由于 AC 跨三层和 spine 互联，如果要实现有线无线策略一体化，只能采用 AP 本地转发。

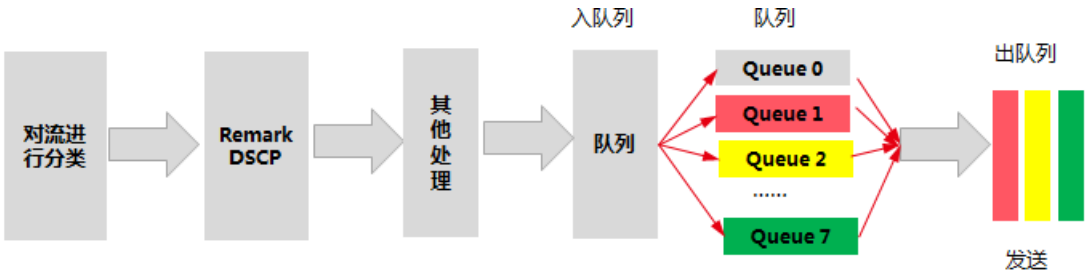
9 网络 Qos 设计

QoS 即服务质量，对于网络业务，影响服务质量的因素包括传输的带宽、传送的时延、数据的丢包率等。在网络中可以通过保证传输的带宽、降低传送的时延、降低数据的丢包率以及时延抖动等措施来提高服务质量。网络资源总是有限的，在保证某类业务的服务质量的同时，可能就是在损害其它业务的服务质量。因此，网络管理者需要根据各种业务的特点来对网络资源进行合理的规划和分配，从而使网络资源得到高效利用。

9.1 Qos设计和部署

9.1.1 QoS 技术简介

图52 QoS 技术对报文的处理流程图



(1) 流分类。配置 ACL 规则，匹配报文的五元组信息，对报文进行分类；

(2) 着色。报文进行分类后，修改报文的 DSCP 值；需要高优先级保证的报文，把其 DSCP 值改大；

(3) 其他处理。如：流量监管、流量整形。

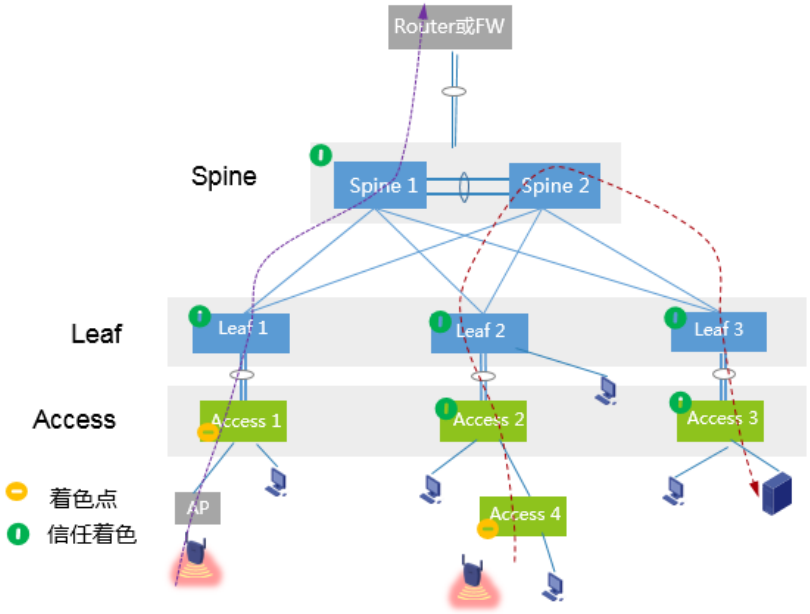
(4) 入队列。根据报文的 DSCP 值映射到不同的内部队列。交换机内部有 8 个队列，其中队列 7 的优先级最高，队列 0 的优先级最低，默认情况下，报文进队列 2 。

(5) 出队列。出方向上，根据队列调度算法对报文进行调度。常见的队列调度算法有 SP、WRR。交换机有默认的队列调度算法，当前控制组件下发的是 SP+WRR 方式。

9.1.2 QoS 方案原理

在边界设备对报文进行着色，网络内部的设备信任边界设备的着色结果，并按照着色结果进行队列调度。

图53 组网图



边界设备是指的流量接入 Fabric 的设备。比如：上图中，紫色虚线的流量是 Access 设备下挂的终端访问 Fabric 外的服务器；对这部分流量做高优先级保障，此时的边界设备就是终端直连的这台 Access 设备。在该台 Access 设备上对报文进行着色（匹配报文的五元组，修改报文的 DSCP 值），leaf 和 spine 设备只要信任 Access 设备的着色结果（端口和全局配置 qos trust dscp 和 qos trust tunnel-dscp），按照着色结果进行调度即可。如果 Fabric 外部访问 Fabric 内部的终端或服务器，此时的边界设备就是 Spine。上图中，红色虚线的流量是 Fabric 内部终端互访的流量，此时，着色点是末级 Access 设备，即：Access4 。转发路径上的其他设备（Access2、Leaf2、Spine、Leaf3 和 Access3）只要信任着色结果，按照着色结果调度即可。此外，为了避免终端私自修改 DSCP，在 Leaf 上的用户直连口和 Access 下行口，通过 QoS 配置，将这些端口上不做 QoS 保证的报文的 DSCP 设置为 BE（进队列 2）。

9.1.3 部署指导

(1) 明确业务类型、流量模型及每种业务对服务质量的需求。

关注点	目的
要保障业务的流量模型，明确端到端的转发路径	确定要部署QoS策略的位置（包括着色点和信任点）
网络中存在的带宽瓶颈点	确定是采用QoS策略还是网络扩容
业务类型及保障的优先级	梳理网络流量类型，如：语音、视频等；确定哪些流量要做保障，哪些不需要以及要保障流量的保障优先级。
IPoE Web认证前域的流量模型，明确允许访问的服务、重定向的流量特征	确定BRAS上部署IPoE Web认证前域的流量特征，以及部署Qos策略在全局还是接口

(2) 流量分类设计。根据网络中各类业务对丢包、时延等参数的要求，将业务服务等级划分如下。实际中，也可以根据需求，调整业务的服务等级。

应用名称	服务等级（队列）
网络控制（网络控制平面）	5
电话（IP电话业务）	5
信令（广播电视，视频监控）	5
多媒体会议（桌面多媒体会议）	4
实时互动（视频会议和高清视频）	4
多媒体流（VOD流媒体业务）	3
广播视频（IP语音和视频业务）	3
低延迟数据（客户/服务器应用程序）	1
OAM（网络运营、维护和管理类业务）	1
高通量数据（文件传输业务）	0
标准（其他应用）	2
低优先级数据（无需带宽保证的业务）	0

(3) 调度策略设计

- 有线网络的调度策略：

  在 Fabric 的边界处对报文进行着色（标记或重标记报文的 DSCP），并进行带宽控制；Fabric 内部的设备通常只需要信任边界设备的着色结果，并按照着色结果进行队列调度，我司方案中，0~3 队列使用 WRR（根据权重进行轮询调度），4~7 队列使用 SP（根据优先级进行调度）。

  对于南北向流量做 QoS 保障，Access 和 Spine 是边界设备，它们负责着色。

  对于东西向流量做 QoS 保障，通常 Access 是边界设备，负责着色。

  - 边界设备负责流量识别：

    边界设备作为着色点，负责对数据流进行识别、分类及流量标记。Access 作为边界设备时，通常建议 Access 上使用全局 QoS 策略，以便节省 ACL 资源，降低对 Access 设备的要求。

    需要注意，如果使用控制组件下发 QoS 策略，需要 Access 为我司方案配套。如果 Access 非我司方案配套，可以选择在该 access 上手工配置 QoS 策略，该 access 接入的 leaf 设备上只需要 trust 该 access 的着色结果；也可以选择将着色点放在 leaf 设备上，此时，就无法在 access 上做 QoS 保障。

  - 网络侧设备按照着色结果进行调度

    Leaf 和 Spine 作为网络侧设备，要在其端口信任 DSCP，基于边界设备着色后的优先级进行调度。

- 无线业务的调度策略

  当前有 2 种方式：

  方式一：着色点放在 Access 设备（无线本地转发场景）或 Spine 设备（无线集中转发场景）。即使无线 AP 上已经对报文进行着色，也要在 Access 或 leaf 上重新着色。该种方式可以由控制组件完成 QoS 相关配置，但是不能做到端到端的 QoS 保障。

  方式二：端到端的 QoS 保障。

  - 上行流量：如果 AP 不信任无线终端报文的优先级，则要通过授权 user-profile 重新设置优先级，user-profile 里配置 QoS 策略，定义流分类和修改优先级。（该方式仅适用于 AC 集中转发场景，不适用于 AP 本地转发场景。）
  - 下行流量：AP 的以太网端口配置 trust dscp，则下行报文达到 AP 后，AP 信任报文优先级。无线空口发送时，会按照报文携带的优先级映射为 802.11e 的优先级发送。

10 网络安全设计

防火墙有串接或旁挂两种部署方式，一般采用 2 台防火墙，做 RBM 方式（主备模式）部署，确保可靠性。

建议防火墙旁挂在 Border 设备，Border 设备做 M-LAG，通过 M-LAG 口分别跟 2 台 FW 互连。此方式可以便于防火墙单链路故障或单台防火墙故障时，业务流量可以通过另一台防火墙转发。

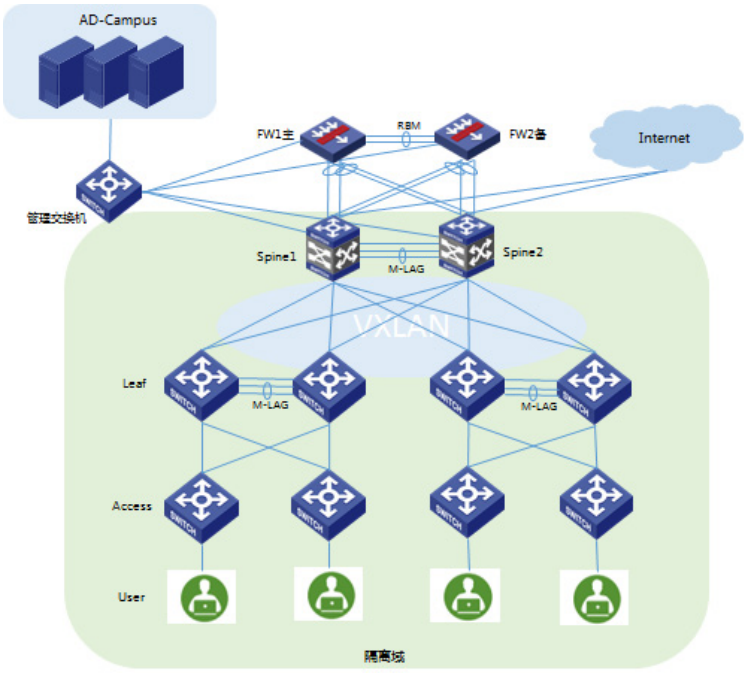
当前方案已支持单隔离域单 Fabric 内南北向流量以及跨私网流量过防火墙，该方案通过在 Border 上配置静态路由，将南北向流量（即：内网跟外网互访的流量）以及跨私网互访的流量引到防火墙。Border 上路由表项规格一般都很大，不会造成瓶颈，并且该方案中 Border 和防火墙对接使用普通 IP 方式，方便后续扩展到传统网中。

该方案对组网及板卡等要求如下：

- 支持 Spine-Leaf-Access 三层组网以及单 Leaf 组网
- 2 台 Spine 或单 Leaf 要组成 M-LAG 系统
- 仅支持盒式/框式防火墙，不支持防火墙插卡
- 2 台防火墙做 RBM，工作在主备模式，且防火墙要旁挂 Border 设备（这里的 Border 指的是三层组网中的 Spine 设备或者单 Leaf 组网的 Leaf 设备）
- 支持南北向流量和跨私网流量过墙，且过相同的防火墙
- 当前仅支持组策略，且仅支持共享出口

如图 54 所示，两台 Spine 设备组成 M-LAG 系统，两台防火墙组成 RBM 系统，旁挂在 Spine 设备。防火墙工作在主备模式，使用 VRRP 与 RBM 联动。

图54 业务引流防火墙的典型组网图

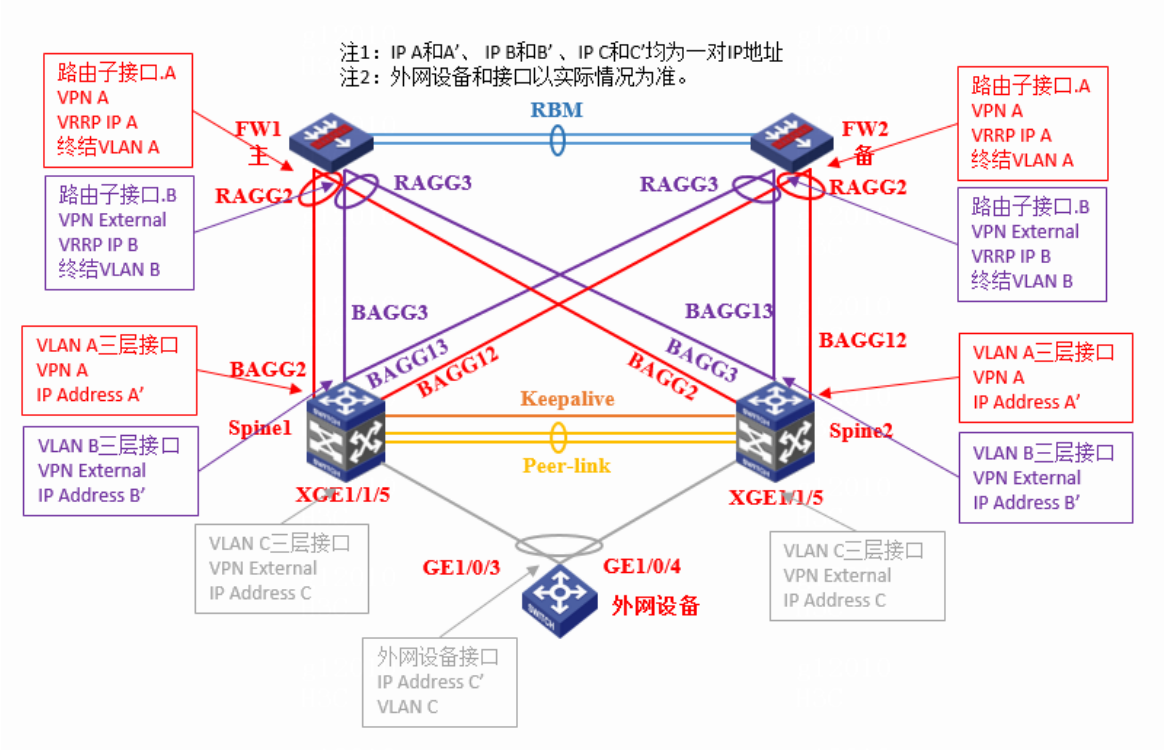


10.1 南北向流量引流防火墙设计

10.1.1 网络资源规划

如图 55 所示，用户 VPN A 访问外网时，分别为 Spine 设备、防火墙、外网设备分配相关的网络资源。

图55 VPN 访问外网资源规划



(1) 分配防火墙下行链路 VLAN A、上行链路 VLAN B、出口 VLAN C；

(2) 为防火墙下行链路分配一对 IP 地址 A 和 A’，掩码最小为 29 位。两台防火墙的下行链路分别创建路由子接口，绑定 VPN A，并组成 VRRP 系统，VRRP 虚 IP 地址配置为 IP A，并终结 VLAN A；接口的 VRRP 实 IP 地址与 IP A 配置为同网段 IP 地址，掩码相同。两台 Spine 设备的上行 M-LAG 接口分别允许通过 VLAN A，配置 VLAN A 的三层接口的 IP 地址为 IP A’，并绑定 VPN A；

(3) 为防火墙上行链路分配一对 IP 地址 B 和 B’，掩码最小为 29 位。两台防火墙的上行链路分别创建路由子接口，绑定 VPN External，并组成 VRRP 系统，VRRP 虚 IP 地址配置为 IP B，并终结 VLAN B；接口的 VRRP 实 IP 地址与 IP B 配置为同网段 IP 地址，掩码相同。两台 Spine 设备的下行 M-LAG 接口允许通过 VLAN B，配置 VLAN B 的三层接口 IP 地址为 IP B，并绑定 VPN External；

(4) Spine1 和 Spine2 设备的出接口组成 M-LAG 接口，分配一对 IP 地址 C 和 C’，掩码为 31 位。出接口允许通过 VLAN C，配置 VLAN C 的三层接口的 IP 地址为 IP C，并绑定 VPN External。配置外网设备互联接口的 IP 地址为 IP C’。

10.1.2 业务引流实现

通过配置静态路由的方式，实现用户 VPN A 通过 Spine 旁挂的防火墙跟外部网络互通的需求。

如图 56 所示，用户 VPN A 访问外网的去程流量：

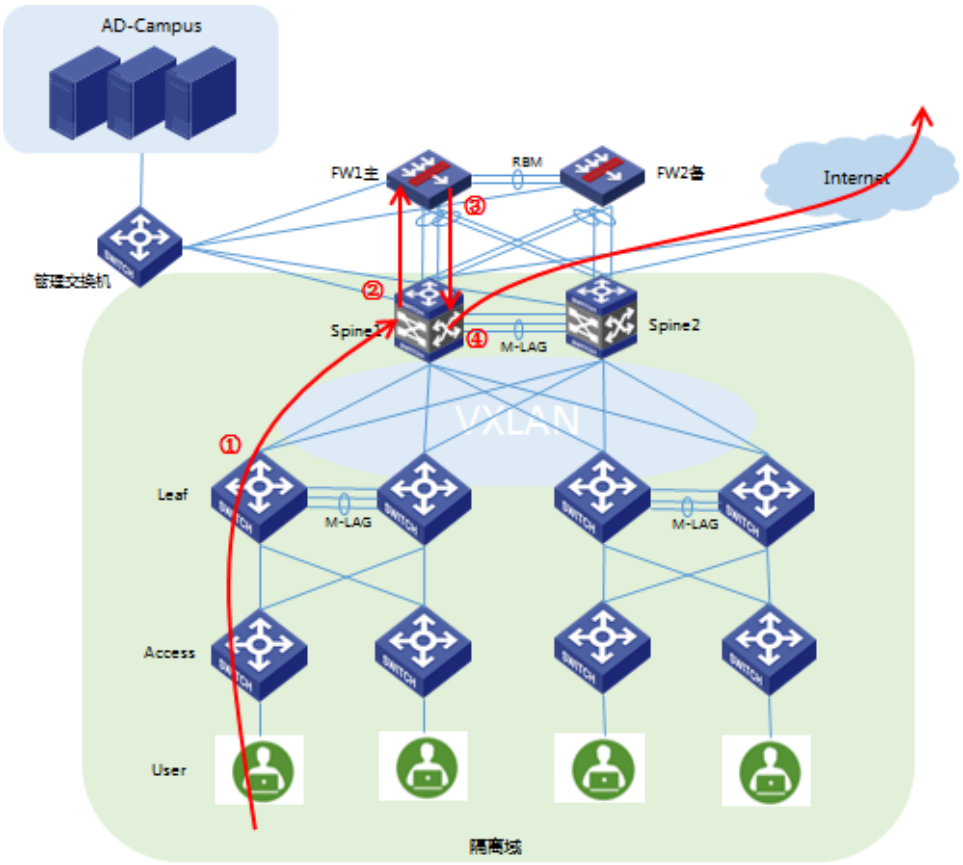
(1) Leaf 设备上，在 VPN A 实例内，匹配 BGP 路由，将流量从 Leaf 设备上送至 Spine 设备；

(2) Spine 设备上，在 VPN A 实例内，匹配默认路由，下一跳为 IP A，将流量转发至防火墙下行链路接口；

(3) 防火墙上，在 VPN A 实例内，匹配默认路由，目的 VPN 切换至外网 VPN，下一跳为 IP B’，通过防火墙的上行链路将流量转发至 Spine 设备；

(4) 根据流量到达的 Spine 设备，在外网 VPN External 实例内，匹配默认路由，下一跳为 IP C’，将流量转发至对应外网设备的互联接口。

图56 VPN 访问外网去程流量图



如图 57 所示，用户 VPN 访问外网的回程流量：

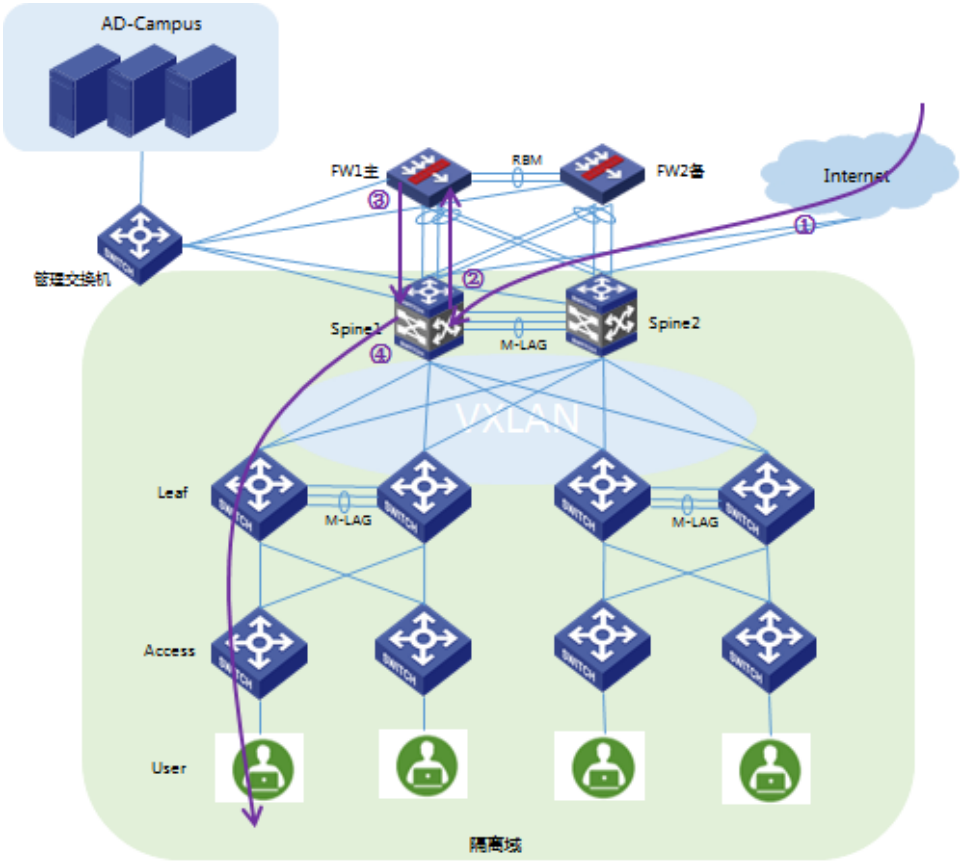
(1) 在外网设备上，通过相关路由协议将流量转发至 Spine1 或 Spine2 设备，下一跳为 IP C，此时流量属于外网 VPN External；

(2) Spine 设备上，在外网 VPN External 实例内，匹配目的为用户 VPN A 的内网网段的静态路由，下一跳为 IP B，将流量转发至防火墙的上行链路接口；

(3) 防火墙上，在外网 VPN External 实例内，匹配目的为用户 VPN A 的内网网段的静态路由，目的 VPN 切换至内网 VPN A，下一跳为 IP A’，通过防火墙的下行链路将流量转发至 Spine；

(4) Spine 设备匹配用户 VPN A 的主机路由，将流量下送至 Leaf 设备，最终转发至客户端。

图57 VPN 访问外网回程流量图



注意

（1）如果 FW 是单台，需要 2 台 Border 做 M-LAG，通过 M-LAG 口跟该 FW 互连。（Future）

10.2 跨私网流量引流防火墙设计

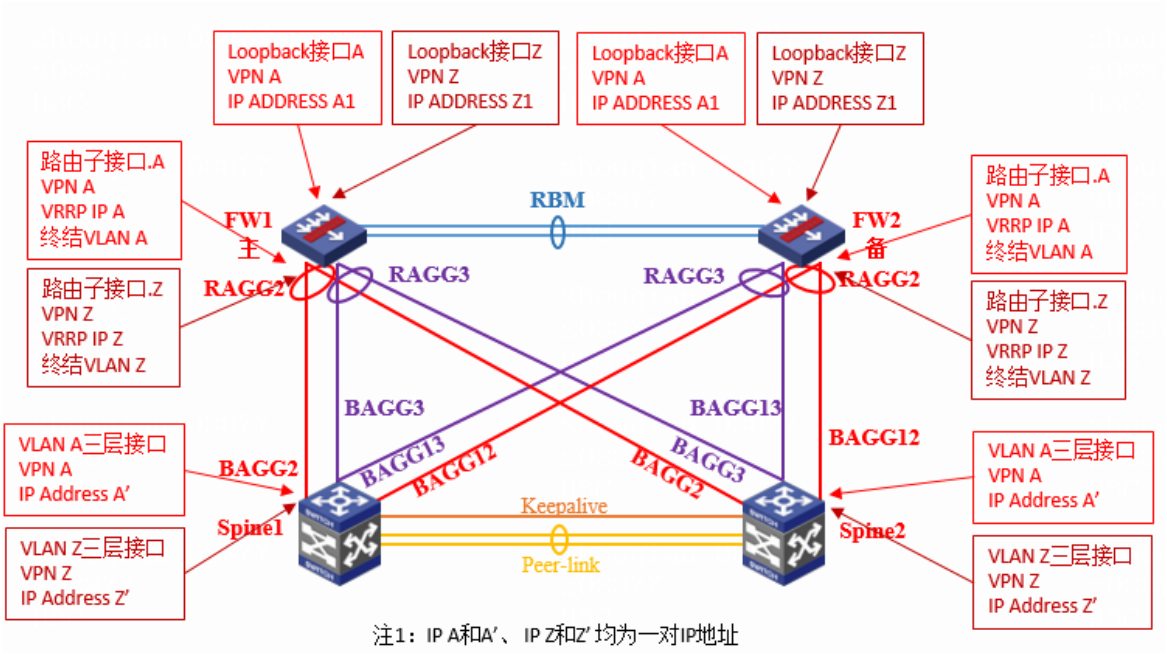
跨私网的流量默认是隔离的，如果有需要互通的需求，可以通过静态路由将跨私网流量引到防火墙，通过安全策略控制流量放通或不放通。

10.2.1 网络资源规划

如图 58 所示，用户 VPN A 与 VPN Z 互访时，分别为 Spine 设备和防火墙分配相关的网络资源，本场景所述流量只经过防火墙上行链路。

在防火墙上两个 VPN 之间的明细路由转发时采用 Loopback 口的方式，分别为 VPN A 和 VPN Z 创建对应的 Loopback 接口，并指定接口 IP 地址。配置跨 VPN 的静态路由时，下一跳指向目的 VPN 的 Loopback 接口 IP 地址。/

图58 VPN 互访网络资源规划



(1) 为 Spine 和防火墙的互连链路分配 VLAN A 和 VLAN Z；

(2) 为防火墙分配 VPN A 和 VPN Z 互访时所需的 Loopback 接口的 IP 地址，VPN A 对应 IP A1，VPN Z 对应 IP Z1，掩码均为 32 位；

(3) 为防火墙和 Spine 设备之间的链路分配一对 IP 地址 A 和 A'，掩码为 29 位，用于用户 VPN A。两台防火墙的下行链路分别创建路由子接口，绑定 VPN A，并组成 VRRP 系统，VRRP 虚 IP 地址配置为 IP A，终结 VLAN A。两台 Spine 设备跟防火墙下行链路对接的 M-LAG 接口分别允许通过 VLAN A，并配置 VLAN A 的三层接口的 IP 地址为 IP A'，绑定 VPN A；

(4) 为防火墙和 Spine 设备之间的链路分配一对 IP 地址 Z 和 Z'，掩码为 29 位，用于用户 VPN Z。两台防火墙的下行链路分别创建路由子接口，绑定 VPN Z，并组成 VRRP 系统，VRRP 虚 IP 地址配置为 IP Z，并终结 VLAN Z。两台 Spine 设备跟防火墙下行链路对接的 M-LAG 接口分别允许通过 VLAN Z，并配置 VLAN Z 的三层接口的 IP 地址为 IP Z'，绑定 VPN Z。

10.2.2 业务引流实现

通过配置静态路由的方式，实现用户 VPN A 与 VPN Z 通过 Spine 旁挂的防火墙互访的需求。如图 59 所示，用户 VPN A 访问 VPN Z 的去程流量：

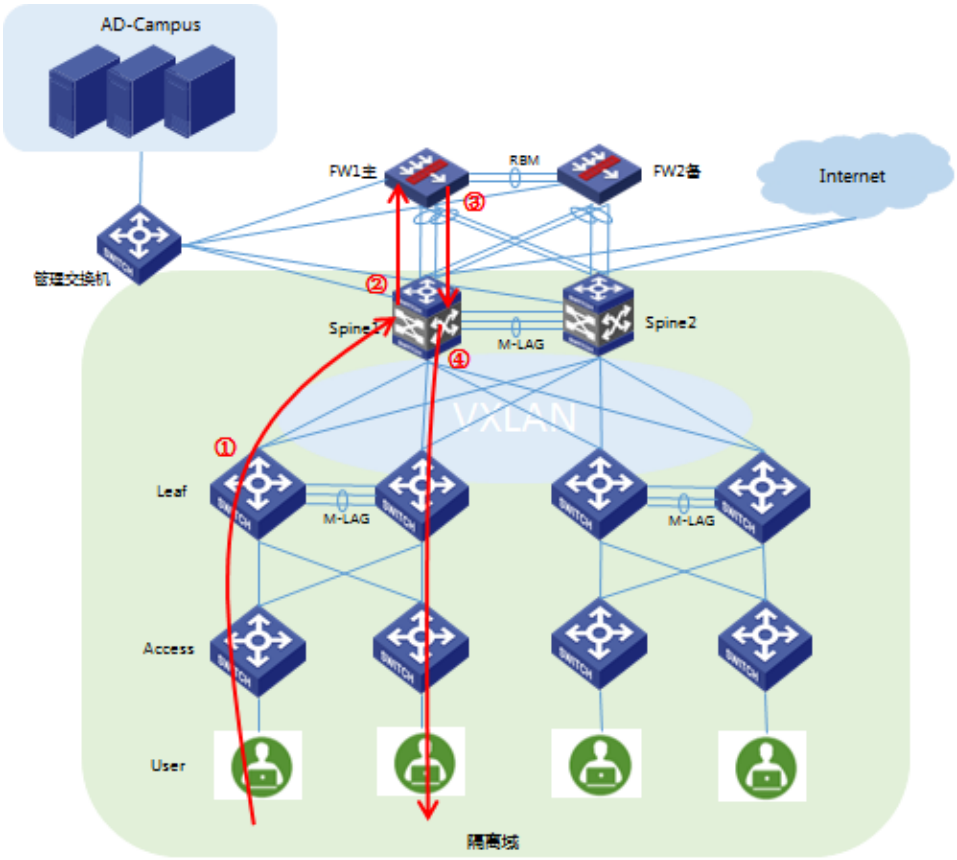
(1) 在 Leaf 设备上，在 VPN A 实例内，匹配目的地址为 VPN Z 的内网网段的 BGP 路由，将流量从 Leaf 设备上送至 Spine 设备；

(2) 在 Spine 设备上，在 VPN A 实例内，匹配目的地址为 VPN Z 的内网网段的静态路由，下一跳为 IP A，将流量转发至防火墙的下行链路接口；

(3) 在防火墙上，在 VPN A 实例内，匹配目的地址为 VPN Z 的内网网段的静态路由，出接口为 Loopback A，下一跳为 IP Z1，此时用户流量所在 VPN 切换至 VPN Z，将流量通过防火墙下行链路转发至 Spine 设备；

(4) 在 Spine 设备上，在 VPN Z 实例内，匹配目的主机的 BGP 路由，将流量送至 Leaf 设备，最终转发至目的主机。

图59 私网互访的去程流量图



如图 60 所示，用户 VPN A 访问 VPN Z 的回程流量：

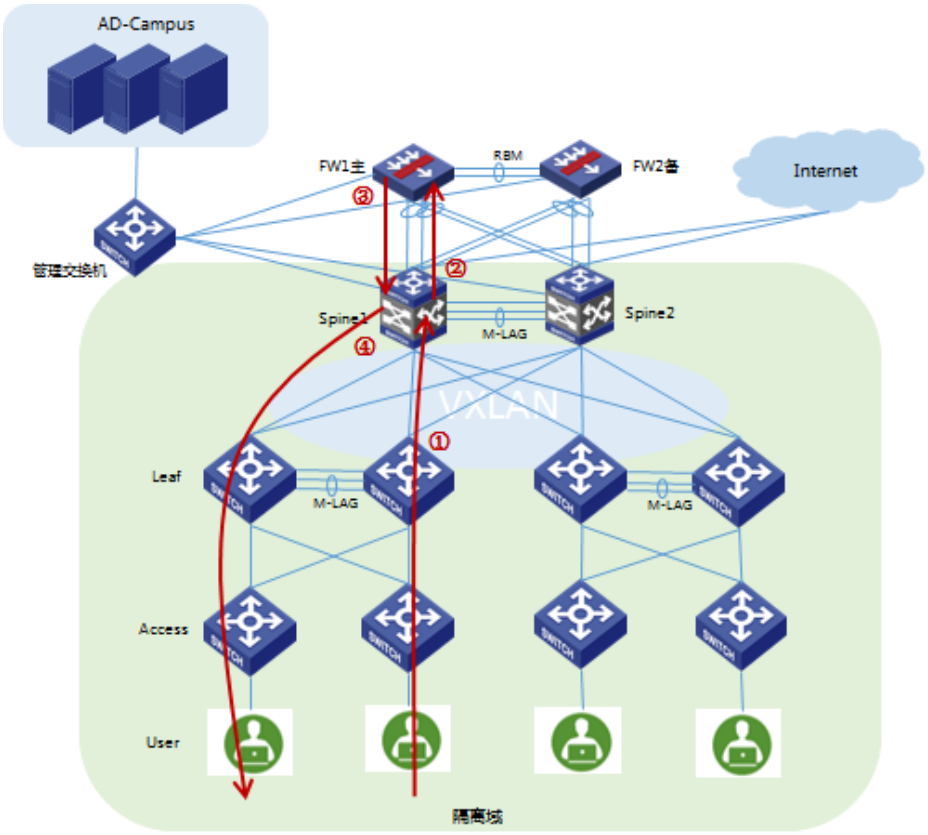
(1) 在 Leaf 设备上，在 VPN Z 实例内，匹配目的地址为 VPN A 的内网网段的 BGP 路由，将流量从 Leaf 设备上送至 Spine 设备；

(2) 在 Spine 设备上，在 VPN Z 实例内，匹配目的地址为 VPN A 的内网网段的静态路由，下一跳为 IP Z，将流量转发至防火墙的下行链路接口；

(3) 在防火墙上，在 VPN Z 实例内，匹配目的地址为 VPN A 的内网网段的静态路由，出接口为 Loopback Z，下一跳为 IP A1，此时用户流量所在 VPN 切换至 VPN A，将流量通过防火墙下行链路转发至 Spine 设备；

(4) 在 Spine 设备上，在 VPN A 实例内，匹配目的主机的 BGP 路由，将流量送至 Leaf 设备，最终转发至目的主机；

图60 私网互访的回程流量图



10.3 防火墙故障逃生

虽然当前组网已具备很高的可靠性，但是仍需要考虑极端场景，比如：防火墙设备全部故障的情况。防火墙故障逃生方式就是针对这类极端场景，使得用户私网能够继续访问外网或者其他私网，而无需绕行防火墙；当防火墙故障恢复后，流量能够重新绕行防火墙，并尽可能缩短故障-恢复过程带来的流量中断时间。当前防火墙故障逃生方案涉及的配置需要手配。

故障逃生方案分为两部分：

(1) 故障探测

两台 Spine 通过 NQA 探测防火墙的状态，将南北向流量以及跨私网所用的静态路由关联 track（该 track 关联故障探测的 NQA）。当 NQA 变成 Negative 时，引流到防火墙的静态路由则不生效。

(2) 故障逃生

两台 Spine 上配置自环链路和逃生路由（逃生路由的优先级低于引流 FW 的静态路由），用于防火墙故障后的流量逃生；配置逃生路由，下一跳指向自环链路对应的接口。当 FW 全部故障或多条链路同时故障时，则 NQA 探测不成功，路由会切到逃生路由，南北向流量和跨私网流量通过自环链路仍可以通。当 FW 或链路都恢复后，NQA 探测成功，南北向流量和跨私网流量可以仍走 FW。

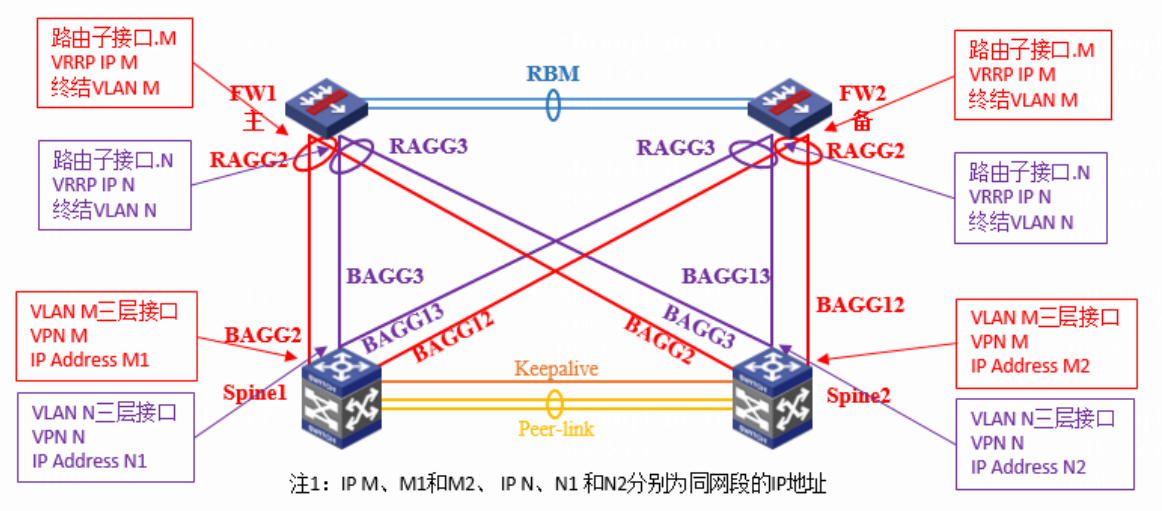
10.3.2 故障探测

对于南北向流量，故障探测需要在两台 Spine 创建上行口和下行口，分别绑定不同的私网，用下行口去探测上行口或者反之也可以，该探测报文的路径和南北向流量引流防火墙的路径是一致，所以 Spine 就可以探测防火墙的状态。对于跨私网流量，由于跨私网的流量仅经过防火墙下行链路，故只需要探测跨私网静态路由的下一跳。

如图 61 所示，上行探测链路规划 VPN M，两个防火墙链路接口的 VRRP 虚 IP 地址为 M （防火墙路由子接口可以绑定私网，也可以不绑定），两个 Spine 设备的链路接口地址分别为 M1 和 M2，三个 IP 地址属于同一个网段；下行探测链路规划 VPN N，两个防火墙链路接口的 VRRP 虚 IP 地址为 N，两个 Spine 设备的链路接口地址分别为 N1 和 N2，三个 IP 地址属于同一个网段。

请注意：故障探测涉及的配置可以配在根墙，也可以配置在虚墙，一般是配置在根墙上。FW 对接的路由子接口可以绑定私网，也可以不绑定私网。

图61 故障探测资源规划图



在 Spine1 设备上配置 NQA，探测报文的源 VPN 为 VPN M1，探测的目的地址为 IP N1，属于 VPN N。同理，在 Spine2 设备上配置 NQA，探测源 VPN 为 VPN M2，探测的目的地址为 IP N2，属于 VPN N。以 Spine1 为例，介绍下探测报文的转发过程：

探测链路的去程路由：

(1) 在 Spine1 设备上，在 VPN M 实例内，匹配目的网段为 N1/mask 的静态路由，下一跳为防火墙的下行链路接口地址 IP M；

(2) 在防火墙上，匹配目的网段为 N1/mask 的直连路由（如果防火墙的路由子接口绑定私网 M，此处匹配的是 VPN M 内的静态路由），下一跳为 IP N1，流量经过防火墙的上行链路转发至 Spine1 设备。

探测链路的回程路由：

(1) 在 Spine1 设备上，在 VPN N 实例内，匹配目的网段为 M1/mask 的静态路由，下一跳为防火墙的上行链路接口地址 IP N；

(2) 在防火墙上，匹配目的网段为 M1/mask 的直连路由（如果防火墙的路由子接口绑定私网 N，此处匹配的是 VPN N 内的静态路由），下一跳为 IP M1，流量经过防火墙下行链路转发至 Spine1 设备。

10.3.3 故障逃生

当 Spine 设备探测到防火墙故障后，原本绕行防火墙的流量在 Spine 设备上需要能够直接访问外网或者目的 VPN。因此需要在 Spine 设备上实现跨 VPN 之间的静态路由切换，采用自环链路的方式。在 Spine 设备上分别为源和目的 VPN 创建一个三层路由子接口，绑定对应 VPN，并配置对应的 IP 地址。配置跨 VPN 静态路由时，下一跳指向目的 VPN 绑定的子接口的 IP 地址。

在两台 Spine 设备上规划，VPN A 对应的自环链路接口 IP 地址为 IP A2，VPN Z 对应的自环链路接口 IP 地址为 IP Z2，VPN External 对应的自环链路接口 IP 地址为 IP B2。

1. VPN 访问外网场景

(1) 如图 62 所示，防火墙全故障后，Spine 设备上的去程路由 ii 设计如下：

- 路由 i（引流防火墙的路由）：Spine 设备到防火墙，在 VPN A 实例内，匹配默认路由，将流量转发至防火墙下行链路接口，下一跳为 IP A，优先级默认为 60，增加 TRACK 联动防火墙故障探测的 NQA；
- 路由 ii（逃生路由）：Spine 设备到外网设备，在 VPN A 实例内，匹配默认路由，将流量经过自环链路再转发至外网设备，该条默认路由的下一跳为 IP B2，流量所属 VPN 切换至 VPN External。该路由的优先级设置为 100（优先级只要大于 60）。该逃生的静态路由要和静态 ARP 配合使用，配置举例如下：

  逃生路由：

  ip route-static vpn-instance A 0.0.0.0 0 B2（下一跳）preference 100

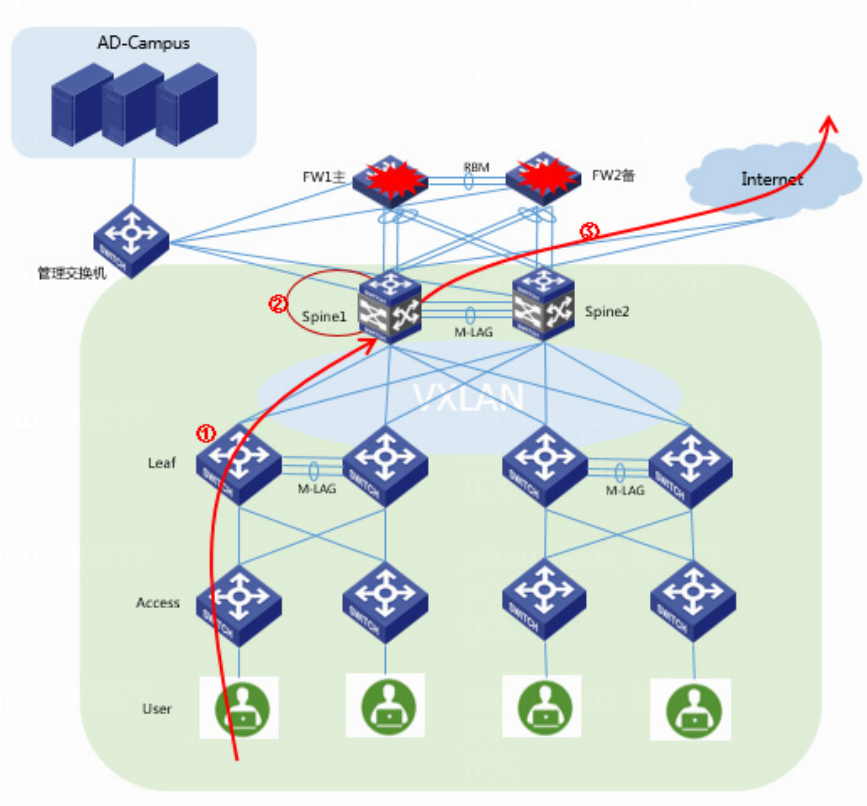
  静态 ARP：

  arp static B2 x-x-x-x vpn-instance A【注意：指定的所属私网是私网 A，而不是 B2 所在的私网】

  注：x-x-x-x 是 IP 地址 B2 所在接口的 MAC 地址，可以通过 display interface 查看接口对应的 MAC 地址。每条逃生路由都要配置对应的静态 ARP，方式类似。

  当防火墙全部故障后，路由 i 失效，路由 ii 生效；当防火墙故障恢复后，路由 i 生效，路由 ii 失效。

图62 防火墙故障时 VPN 访问外网去程流量图

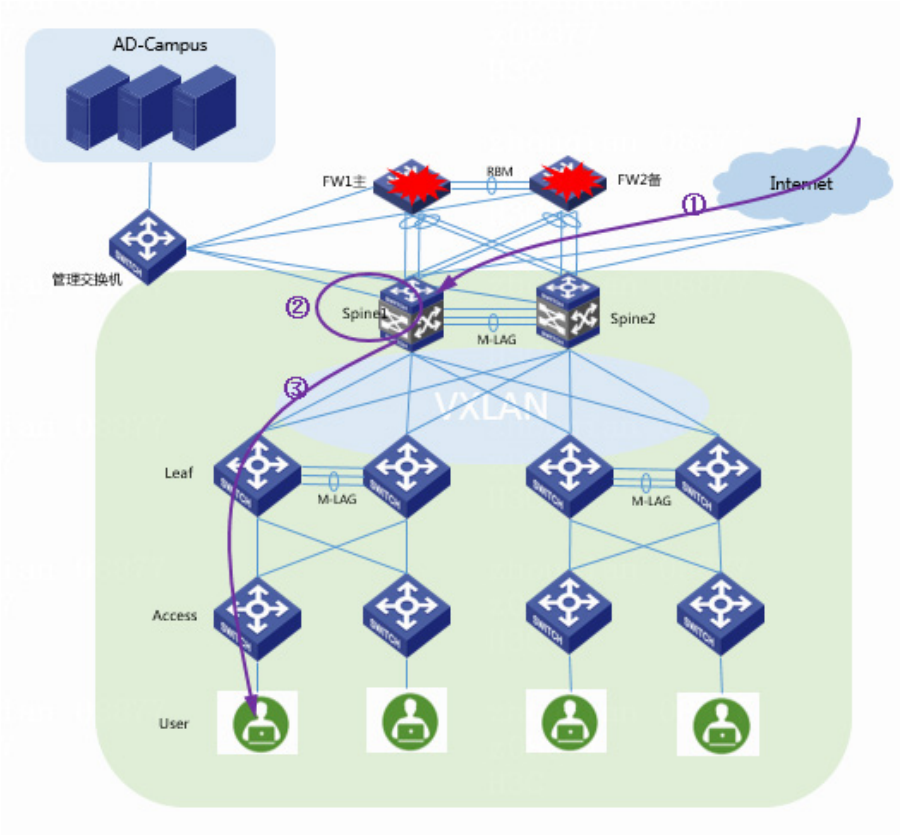


(2) 如图 63 所示，防火墙全故障后，Spine 设备上的回程路由 ii 设计如下：

- 路由 i（引流防火墙的路由）：Spine 设备到防火墙，在外网 VPN External 实例内，匹配目的为用户 VPN A 的内网网段的静态路由，将流量转发至防火墙的上行链路接口，下一跳为 IP B，优先级默认为 60，增加 TRACK 联动防火墙故障探测的 NQA；
- 路由 ii（逃生路由）：Spine 设备到内网，在外网 VPN External 实例内，匹配目的为用户 VPN A 的内网网段的静态路由，下一跳为 IP A2，流量所属 VPN 切换至 VPN A。该路由的优先级设置为 100（优先级只要大于 60）；

当防火墙全部故障后，路由 i 失效，路由 ii 生效；当防火墙故障恢复后，路由 i 生效，路由 ii 失效。

图63 防火墙故障时 VPN 访问外网回程流量图



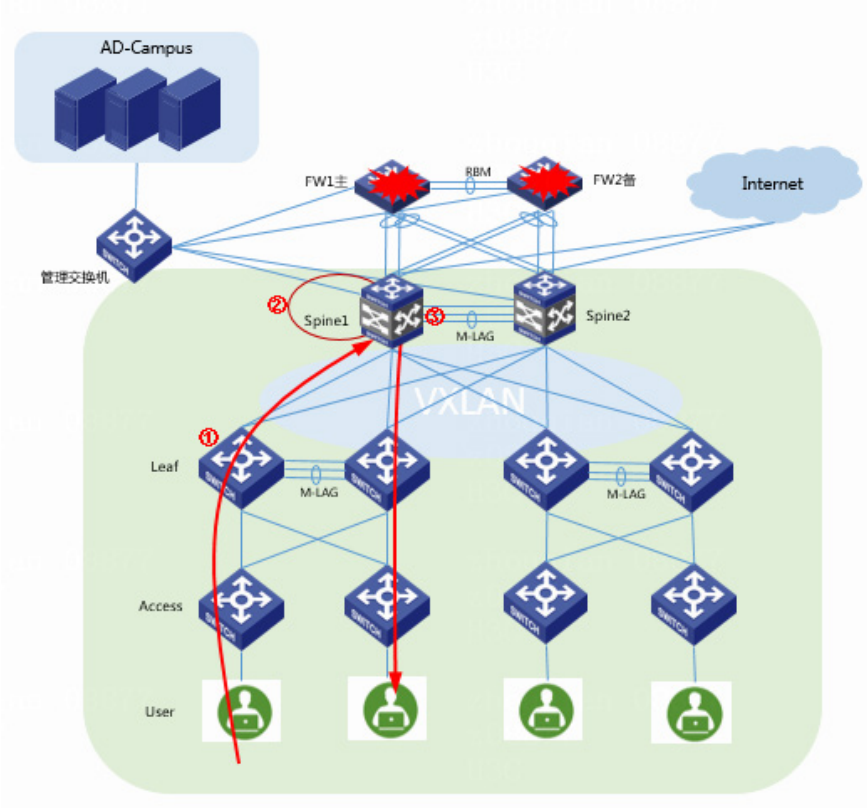
2. VPN 互访场景

(1) 如图 64 所示，防火墙全故障后，Spine 设备上的去程路由 ii 设计如下：

- 路由 i（引流防火墙的路由）：Spine 设备到防火墙，在 VPN A 实例内，匹配目的地址为 VPN Z 的内网网段的静态路由，下一跳为 IP A，将流量转发至防火墙下行链路接口，优先级默认为 60，增加 TRACK 联动防火墙故障探测的 NQA；
- 路由 ii（逃生路由）：Spine 设备上，在 VPN A 实例内，匹配目的地址为 VPN Z 的内网网段的静态路由，下一跳为 IP Z2，通过自环链路，流量环回到本设备，同时流量所属 VPN 切换至 VPN Z。该路由的优先级设置为 100（优先级只要大于 60）；

当防火墙全部故障后，路由 i 失效，路由 ii 生效；当防火墙故障恢复后，路由 i 生效，路由 ii 失效。

图64 防火墙故障时 VPN 互访去程流量图

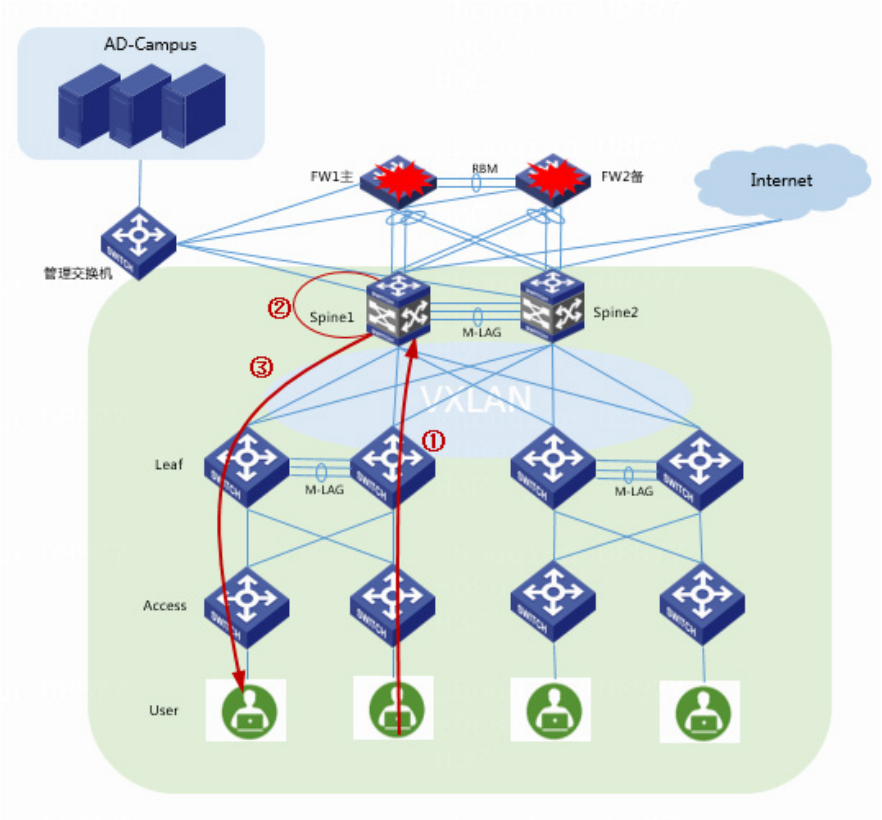


(2) 如图 65 所示，防火墙全故障后，Spine 设备上的回程路由 ii 设计如下：

- 路由 i（引流防火墙的路由）：Spine 设备到防火墙，在 VPN Z 实例内，匹配目的为用户 VPN A 的内网网段的静态路由，下一跳为 IP Z，将流量转发至 FW 设备下行链路接口，优先级默认为 60，增加 TRACK 联动防火墙故障探测的 NQA；
- 路由 ii（逃生路由）：Spine 设备上，在用户 VPN Z 实例内，匹配目的为用户 VPN A 的内网网段的静态路由，下一跳为 IP A2，通过自环链路，流量环回到本设备，流量所属 VPN 切换至 VPN A。该路由的优先级设置为 100（优先级只要大于 60）。

当防火墙全部故障后，路由 i 失效，路由 ii 生效；当防火墙故障恢复后，路由 i 生效，路由 ii 失效。

图65 防火墙故障时 VPN 互访回程流量图



3. 故障逃生方案使用限制

(1) 防火墙故障逃生当前只能手配

(2) 该方案要求 Spine 和 Leaf 互连的单板和自环链路都要用 TD3 芯片的单板（12500G-AF，10500X 的*SH 单板）、TM2 芯片（UNIS S12600-G/H3C S12500G-AF 的 SF 线卡； UNIS S10600X-G/H3C S10500X-G/的 SF 线卡； UNIS S8600X-G /H3C S7500X-G 的 SF 线卡；UNIS S7800XP/S9600XP /H3C S6800-G/9800-G）。

(3) 自环链路仅用于逃生时临时使用。

(4) 要根据流量大小，选择自环的物理链路数量，避免走逃生时，带宽不够。

(5) 逃生方案中，配置了静态 ARP，请不要 reset arp all，否则会将静态 ARP 表项删除。

11 网络可靠性设计

11.1 设备可靠性设计

11.1.1 IRF 可靠性设计

1. 智能弹性架构

IRF（Intelligent Resilient Framework，智能弹性架构）是 H3C 自主研发的软件虚拟化技术。它的核心思想是将多台设备连接在一起，进行必要的配置后，虚拟化成一台设备。使用这种虚拟化技术可以集合多台设备的硬件资源和软件处理能力，实现多台设备的协同工作、统一管理和不间断维护。

2. IRF 规划原则

- 本方案中，Spine 和 Leaf 交换机将采用自动化上线的方式，自动堆叠，服务器接入区的交换机手动堆叠，BRAS 手工堆叠；
- 手动配置时，Member 1 使用 irf-port1/2 与 Member2 使用 irf-port2/1 互联，且每个堆叠口包含 2 个 40G 物理接口，堆叠使用配备的 40G 线缆进行连接；
- 手动配置时为 Member 1 配置优先级为最大值 32，Member 2 优先级为 16；
- 每组堆叠设备配置不同的 IRF 域 ID，确保同一网络中不同 IRF 设备 MAD 检测不互相影响；
- 核心设备堆叠成员设备之间单独连线作为 BFD MAD 检测专用，且用于 MAD 检测的接口需关闭 STP 协议，每组堆叠设备规划一个 MAD 检测专用 VLAN，且该 VLAN 仅仅在 MAD 检测链路允许放行，并为每个成员配置 BFD MAD 检测的同网段的专用 IP 地址，成员 ID 小者取低位地址。接入设备堆叠使用 LACP MAD 检测。
- 堆叠组内配置聚合负载分担为优先本地转发，减少跨框流量对堆叠链路的压力；

3. IRF Domain-id

IRF Domain 是一个逻辑概念，一个 IRF 堆叠组对应一个 IRF 域。为了适应各种业务功能区的组网，同一个网络里部署了多个 IRF 堆叠组，IRF 堆叠组之间使用域编号（DomainID）来以示区别。

区域	设备型号	设备名	IRF Domain-id
核心区	S6525XE-54HF-HI	XXX	10
汇聚层	S6525XE-54HF-HI	XXX	20
	S6525XE-54HF-HI	XXX	30

4. 成员设备的角色

IRF 中每台设备都称为成员设备。成员设备按照功能不同，分为两种角色：

- Master：负责管理整个 IRF 堆叠组设备。
- Standby：作为 Master 的备份设备运行。当 Master 故障时，系统会自动从 Standby 中选举一个新的 Master 接替原 Master 工作。

Master 和 Standby 均由角色选举产生。一个 IRF 中同时只能存在一台 Master，其它成员设备都是 Standby。 确定成员设备角色为 Master 或 Standby 的过程称为角色选举。角色选举会在以下情况下进行：IRF 建立、Master 设备离开或者故障、IRF 合并等。角色选举规则如下：

- 当前 Master 优先，IRF 不会因为有新的成员设备/主控板加入而重新选举 Master。不过，当 IRF 形成时，因为没有 Master 设备，所有加入的设备都认为自己是 Master，则继续下一条规则的比较。
- 成员优先级大的优先。如果优先级相同，则继续下一条规则的比较。
- 系统运行时间长的优先。在 IRF 中，成员设备启动时间间隔精度为 10 分钟，即 10 分钟之内启动的设备，则认为它们是同时启动的，则继续下一条规则的比较。
- CPU MAC 地址小的设备。

5. IRF 端口

一种专用于 IRF 成员设备之间进行连接的逻辑接口，每台成员设备上可以配置两个 IRF 端口，分别为 IRF-Port1 和 IRF-Port2。它需要和物理端口绑定之后才能生效。

在 IRF 模式下，IRF 端口采用二维编号，分为 IRF-Portn/1 和 IRF-Portn/2，其中 n 为设备的成员编号。两个 IRF 端口之间的链路被称为 IRF Link。

与 IRF 端口绑定，用于 IRF 成员设备之间进行连接的物理接口。IRF 物理端口可能是以太网电口或者光口。通常情况下，电口或者光口负责向网络中转发业务报文，将它们与 IRF 端口绑定后就作为 IRF 物理端口，可转发的报文包括 IRF 相关协商报文以及需要跨成员设备转发的业务报文。在做 IRF 堆叠是 IRF 端口规划如下：

- IRF 端口主备对应关系为， Master 设备的 IRF-Port1/2 端口对应 Standby 设备的 IRF-Port2/1 端口；
- 堆叠设备均使用 2 条 40G 链路进行堆叠，以确保堆叠设备之间的带宽容量；

6. MAD（多 Active 检测）

IRF 链路故障会导致一个 IRF 变成两个新的 IRF。这两个 IRF 拥有相同的 IP 地址等三层配置，会引起地址冲突，导致故障在网络中扩大。为了提高系统的可用性，当 IRF 分裂时我们就需要一种机制，检测出网络中同时存在的两个 IRF，并进行相应的处理，尽量降低 IRF 分裂对业务的影响。MAD（Multi-Active Detection，多 Active 检测）就是这样一种检测和处理机制。它主要提供以下功能：

- 通过 LACP（Link Aggregation Control Protocol，链路聚合控制协议）、BFD（Bidirectional ForwardingDetection，双向转发检测）或者免费 ARP（Gratuitous Address Resolution Protocol）来检测网络中是否存在从同一个 IRF 系统分裂出去的且全局配置相同的 IRF；
- 分裂检测：IRF 分裂后，通过分裂检测机制 IRF 会检测到网络中存在其它处于 Active 状态（表示 IRF 处于正常工作状态）的 IRF。冲突处理会让 Master 成员编号最小的 IRF 继续正常工作（维持 Active 状态），其它 IRF 会迁移到 Recovery 状态（表示 IRF 处于禁用状态），并关闭 Recovery 状态 IRF 中所有成员设备上除保留端口以外的其它所有物理端口（通常为业务接口），以保证该 IRF 不能再转发业务报文。（缺省情况下，只有 IRF 物理端口是保留端口，如果要将其它端口，比如用于远程登录的端口，也作为保留端口，需要使用命令行进行手工配置。
- 冲突处理：IRF 链路故障导致 IRF 分裂，从而引起多 Active 冲突。因此修复故障的 IRF 链路，让冲突的 IRF 重新合并为一个 IRF，就能恢复 MAD 故障。如果在 MAD 故障恢复前，

处于 Recovery 状态的 IRF 也出现了故障，则需要将故障 IRF 和故障链路都修复后，才能让冲突的 IRF 重新合并为一个 IRF，恢复 MAD 故障；如果在 MAD 故障恢复前，故障的是 Active 状态的 IRF，则可以通过命令行先启用 Recovery 状态的 IRF，让它接替原 IRF 工作，以便保证业务尽量少受影响，再恢复 MAD 故障。

7. MAD 故障恢复：

由于 BFD MAD 技术可以保证在数百毫秒内检测到 IRF 分裂故障，所以在各业务功能区网络设计时，建议都采用 BFD MAD 技术。BFD MAD 检测链路是 IRF 堆叠组成员设备之间的一个三层心跳线，传输 BFD 检测报文，当 IRF 堆叠组成员设备发生分裂双活时，BFD 会检测到冲突并处理将其中一个台设备处于 Recovery 状态，处于 recovery 状态后会关闭该 IRF 中所有成员设备上除保留端口以外的其它所有物理端口（通常为业务接口），以保证该 IRF 成员设备不能再转发业务报文。

11.1.2 无线 AC 可靠性设计

无线 AC 控制组件均采用 1+1 双链路或 N+1 方式保证可靠性，多台无线 AC 旁挂 spine：

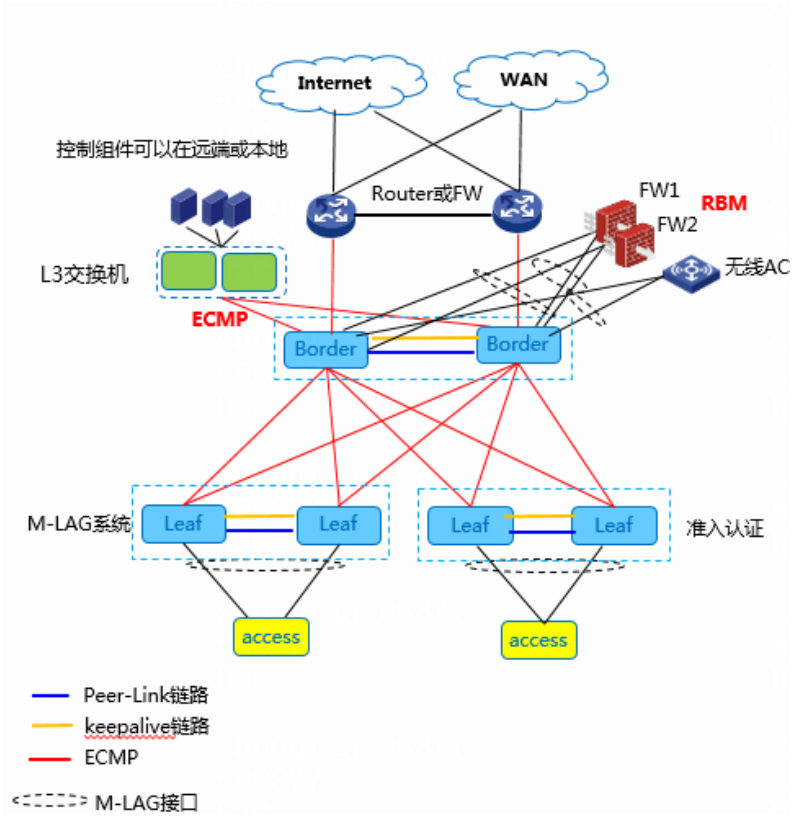
- 如果 Border 是 IRF 堆叠模式，则 AC 通过链路聚合方式和 Border 进行互联；
- 如果 Border 是双机 M-LAG 模式，则 2 台 Border 通过 M-LAG 口分别跟每台 AC 互联。
- 如果 Border 是双机负载分担，则每台 AC 与两台 Border 全连接，通过路由负载分担。

此时无论 AP 本地转发或者 AC 集中转发，都可以支持有线无线策略一体化。

11.1.3 Border/Spine 可靠性设计

园区方案中，大多数场景下 Border 和 Spine 合一，可以使用 IRF 堆叠或者去堆叠双机独立部署模型。在去堆叠方案成熟之前，可以继续使用堆叠方案部署；当去堆叠方案成熟之后，主推去堆叠方案，堆叠方案逐渐淡出，不再主推。

图66 去堆叠 M-LAG 的推荐组网



其中 spine 和 leaf 均为去堆叠 M-LAG 组网方式，spine 跟 L3SW/Router 通过 ECMP 方式互连；通过 M-LAG 口跟 FW 或无线 AC 设备。

双 border 方式下，每个 Border 都和外网的路由器或者 FW 建立单独的路由连接关系。Border 上行到控制组件区采用多条 ECMP 路径进行冗余，提升可靠性。

11.1.4 Leaf 可靠性设计

Leaf 的位置相对较低，可采用堆叠方式或者 M-LAG 方式（推荐）形成双机方式，提高可靠性。如果部署 M-LAG 的话，双机之间需要同时部署 keepalive 链路和 Peer-link 链路。Leaf 上行通过多链路 ECMP 和 Border 之间建立多条等价路径，以提升可靠性。

11.1.5 Access 可靠性设计

Access 可采用堆叠方式保证可靠性，并实现扩展接入端口数量；上行通过链路捆绑的方式提升上行链路的可靠性和带宽。

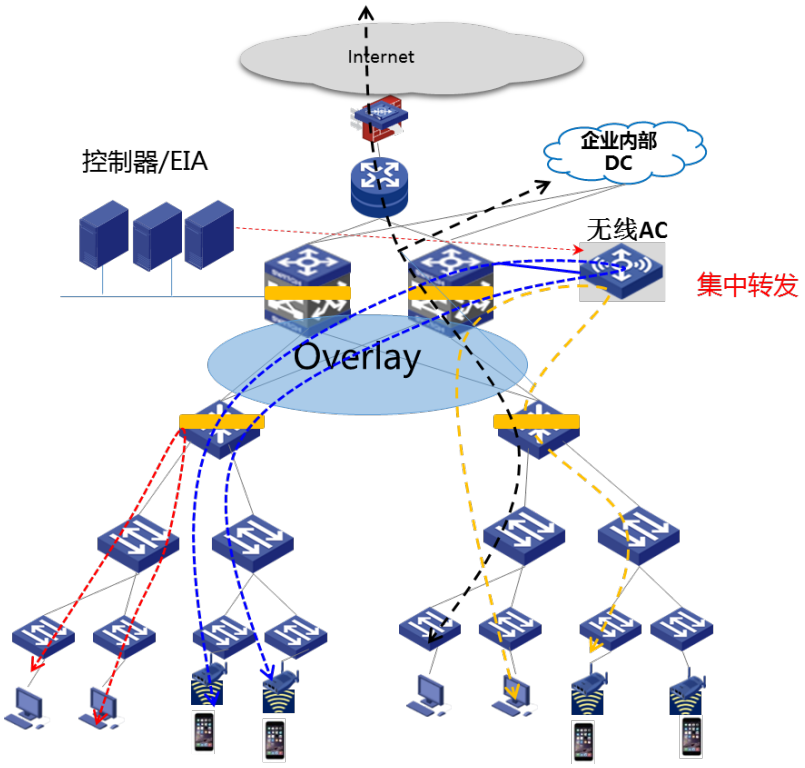
11.2 控制组件可靠性

目前采用 3 机热备方式，三机的数据内部通过集群方式实时同步，增加可靠性。如果想进一步增加可靠性，后续支持主备 3+3 集群方式（future），进一步增加控制组件的可靠性。

12 流量访问模型设计

12.1 网络访问流量模型1（无线AC集中转发）

图67 网络访问流量模型（无线 AC 集中转发）



网络（无线 AC 集中转发）中主要存在如下几种访问模型：

(1) **有线终端之间互访（红色）**：Access 端口之间 per port per VLAN 隔离，流量上行到 leaf，再下行到其他有线终端。或者跨越 spine，下行到其他 leaf 的有线终端。

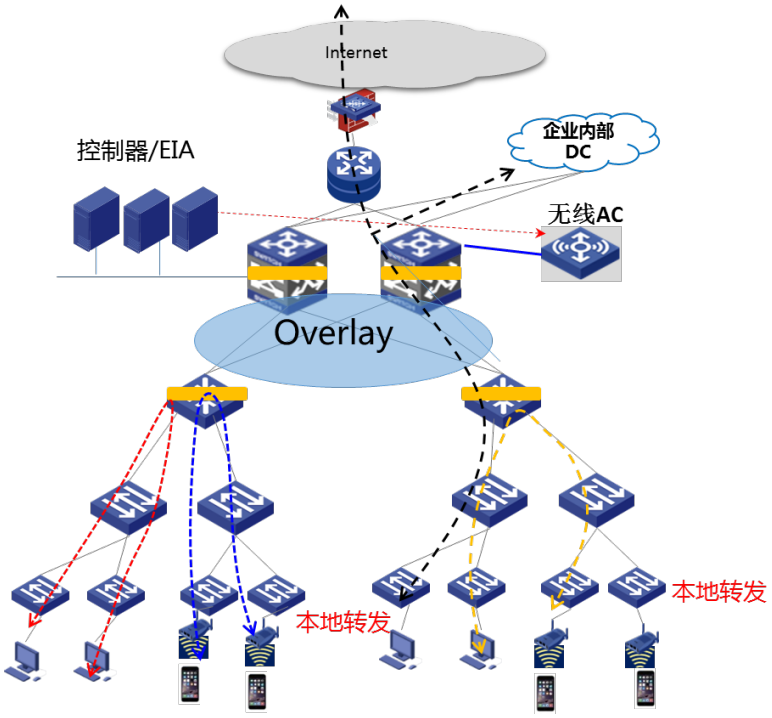
(2) **无线终端之间互访（蓝色）**：由于是 AC 集中转发，流量需要绕行 AC，经过 spine 之后再下行到同 leaf 的无线终端或者跨 leaf 的其他无线终端。

(3) **有线/无线终端之间互访（橙色）**：由于是 AC 集中转发，无线流量需要绕行 AC，经过 spine 之后再转发到跨 leaf 或者同 leaf 的有线终端。

(4) **终端互访 internet/内部 DC（黑色）**：有线终端通过 access 到 leaf 走二层转发，Leaf 走 vxlan 封装到 spine。Spine 解掉 vxlan 封装，走正常路由转发到出口或者内部 DC；无线终端绕行 AC，走正常路由转发到出口或者内部 DC。

12.2 网络访问流量模型2（无线AP本地转发）

图68 网络访问流量模型（无线 AP 本地转发）



网络（无线 AP 本地转发）中主要存在如下几种访问模型：

(1) **有线终端之间互访（红色）**：Access 端口之间 per port per VLAN 隔离，流量上行到 leaf，再下行到其他有线终端。或者跨越 spine，下行到其他 leaf 的有线终端。

(2) **无线终端之间互访（蓝色）**：由于是 AP 本地转发，流量不需要绕行 AC，流量上行到 leaf，再下行到其他无线终端。或者跨越 spine，下行到其他 leaf 的无线终端。

(3) **有线/无线终端之间互访（橙色）**：由于是 AP 本地转发，无线流量不需要绕行 AC，经过 spine 之后再转发到跨 leaf 的有线终端或者不上 spine 直接转发到同 leaf 的有线终端。

(4) **终端互访 internet/内部 DC（黑色）**：有线/无线终端通过 access 到 leaf 走二层转发，Leaf 走 vxlan 封装到 spine。Spine 解掉 vxlan 封装，走正常路由转发到出口或者内部 DC。

网络上除了上述的数据流之外，还有网络设备和控制组件之间交互的管理流和认证流。

13 逃生相关设计

13.1 逃生概述

AD-Campus 方案中，认证成功后，用户才能加入规划的业务安全组、进行互访及访问外部网络，因此，认证服务器（EIA）是 AD-Campus 方案正常工作的重要组件。

若认证服务器不可达，这将导致接入用户不能正常接入网络，进而不能访问网络资源，甚至不能实现用户间互访。为了解决认证服务器不可达导致的用户业务中断，AD-Campus 引入了逃生方案。

逃生是指认证点（NAS）的一种特殊状态：NAS 检测到 AAA 不可用时，用户不经过认证就可以获得缺省的权限（VXLAN 或 SGT），访问有限的资源。当 AAA 恢复后，NAS 会自动触发重认证，与 AAA 进行交互，将用户切换回正常授权。

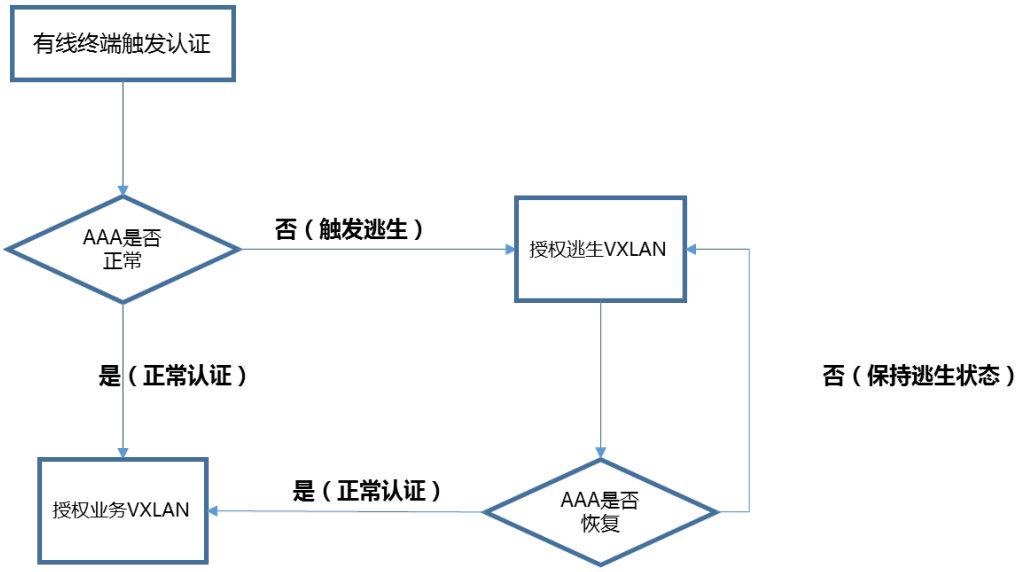
13.2 有线业务逃生

需要通过控制组件提前配置逃生类型的二层网络域和安全组，目前只允许在业务 VPN 下创建。

关于逃生安全组，IP 策略和组策略模式稍有区别，如下：

- IP 策略模式：每个隔离域创只允许创建一个逃生安全组和二层网络域，VPN 选择任意一个业务 VPN 即可。逃生时，接入用户自动获得逃生二层域对应的 VXLAN 授权。流程图如下：

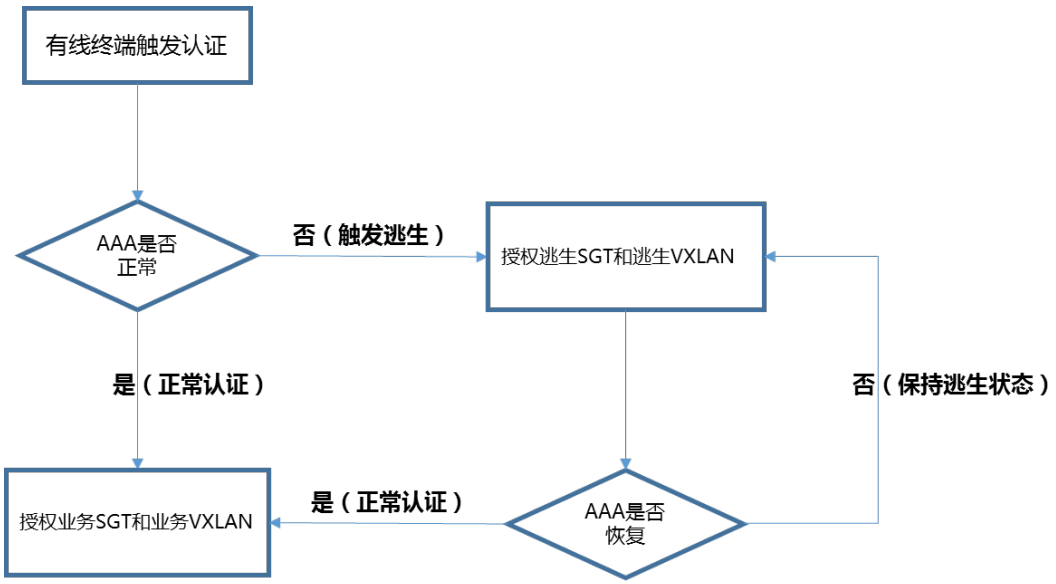
图69 流程图



- 组策略模式：每个隔离域创只允许创建一个逃生二层域，但是逃生安全组可以在每个 VPN 下创建一个。逃生过程分为动态接入用户和静态接入用户，如下：

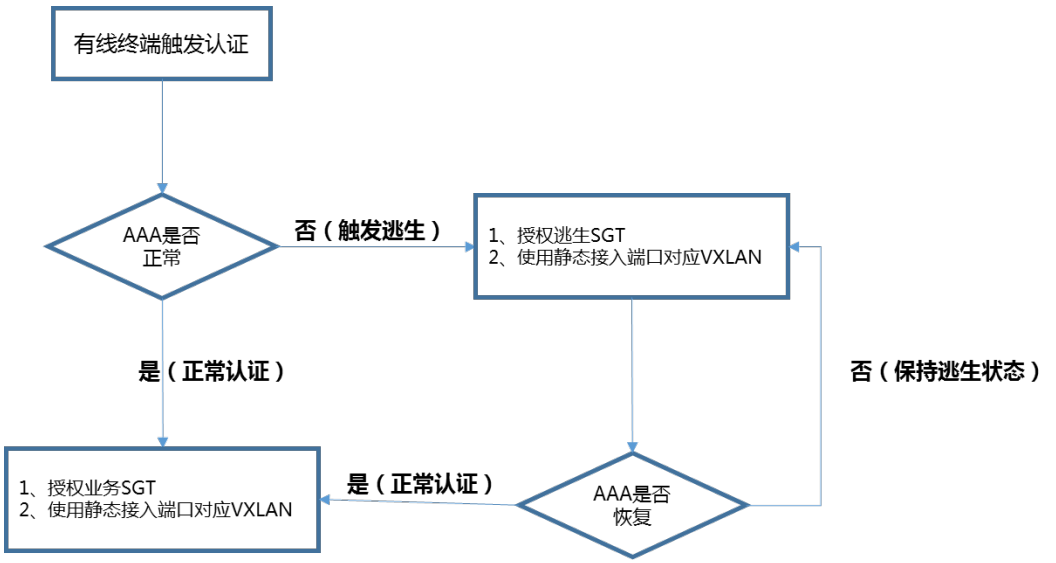
  a. 组策略模式逃生时，动态接入用户（动态授权 VXLAN 场景）自动获得逃生二层域和该二层域对应 VPN 下的逃生 SGT 授权。即逃生时，所有 VPN 的动态接入用户均进入唯一的逃生二层域（VXLAN），SGT 也使用该二层域对应的逃生 SGT。流程图如下：

图70 流程图



b. 组策略模式逃生时，静态接入用户（使用静态 AC 接入场景）自动获得静态 AC 对应 VPN 的逃生 SGT 授权，VXLAN 直接使用静态 AC 设置的 VXLAN。即逃生时，不同 VPN 的静态接入用户进入所属 VPN 的逃生 SGT 授权，二层域（VXLAN）保持不变，与不逃生时的业务二层域（VXLAN）相同。流程图如下：

图71 流程图



13.3 无线用户逃生

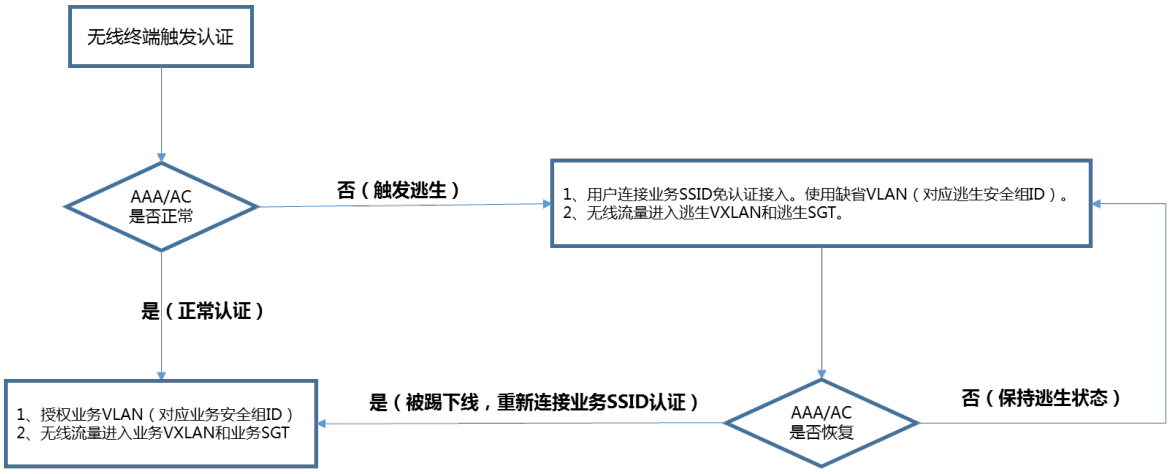
无线用户逃生的二层域、安全组业务配置，直接共用有线用户的即可。但是其逃生过程与有线不同。无线用户的认证点（NAS）为 AC，当 AC 与 AAA 不通或者 AC 与 AP 不通时，均会触发逃生。

逃生时，已在线用户保持不变，新用户连接 SSID 时可以免认证接入，并使用 SSID 缺省的 VLAN。

逃生用户的流量携带 SSID 缺省的 VLAN 到达 Leaf 下行口（AP 本地转发）或者 Spine 连 AC 的接口（AC 集中式转发），因此需要将 SSID 的缺省 VLAN 设置为逃生安全组对应的 ID。这样当逃生用户流量到达时，可以匹配逃生安全组静态服务实例的 VLAN，获得逃生 VXLAN 和 SGT 的权限。

当 AAA 恢复后，AC 会将逃生用户踢下线，用户重新连接 SSID 进行认证即可正常接入网络。流程图如下：

图72 流程图

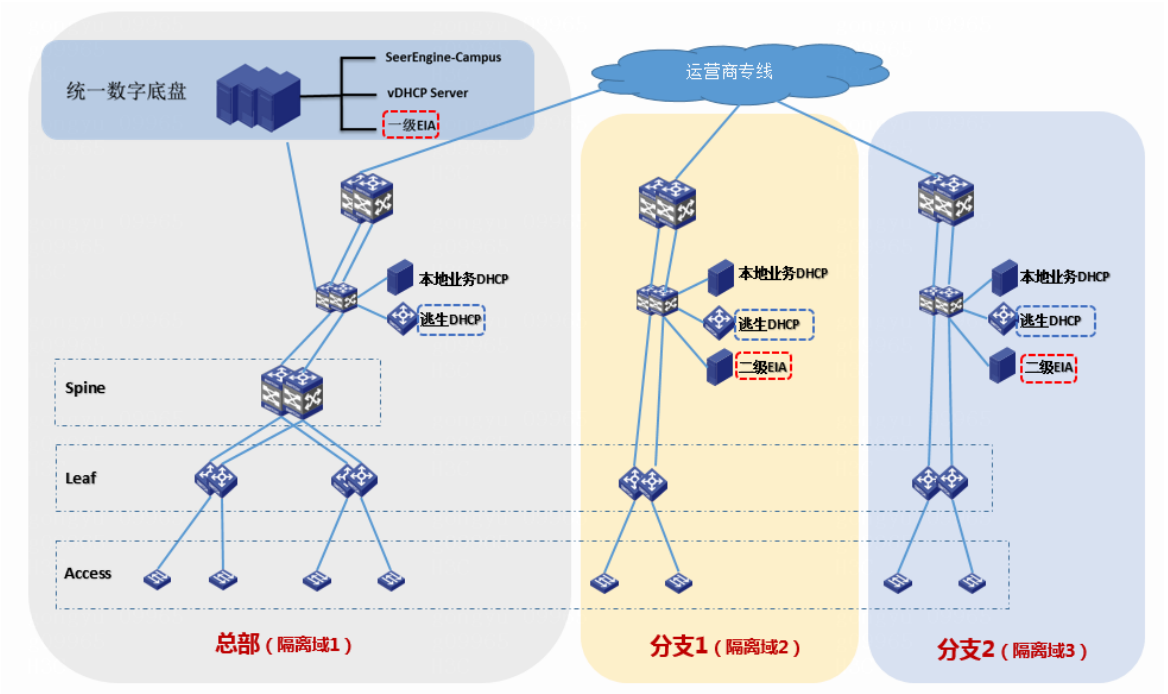


13.4 多园区的主备AAA和逃生DHCP设计

典型的多园区场景如下所示。

- 总部使用一级 EIA（所在底盘一般使用集群部署，可靠性较高），各分支使用二级 EIA（所在底盘可使用单机部署，可靠性稍弱）。
- 总部和各个分支均使用各自独立的业务 DHCP 和逃生 DHCP，且逃生 DHCP 建议靠近 Spine 部署，远离 EIA，以尽量避免两者同时故障。由于 vDHCP 和 EIA 部署在同一底盘，因此 vDHCP 不允许作为逃生 DHCP。因为 EIA 不可达的时候，vDHCP 很大概率也是不可达的。
- 分支的主用 AAA 使用本地的二级 EIA，备用 AAA 使用总部的一级 EIA。
- 总部的 AAA 使用一级 EIA 接口，由于一般是集群部署，可靠性较高，可暂不考虑备用 AAA。

图73 多园区场景



13.5 静态IP逃生

AAA 故障时，认证点会将新上线终端授权到逃生 VXLAN 中，即新终端会进入逃生网段。

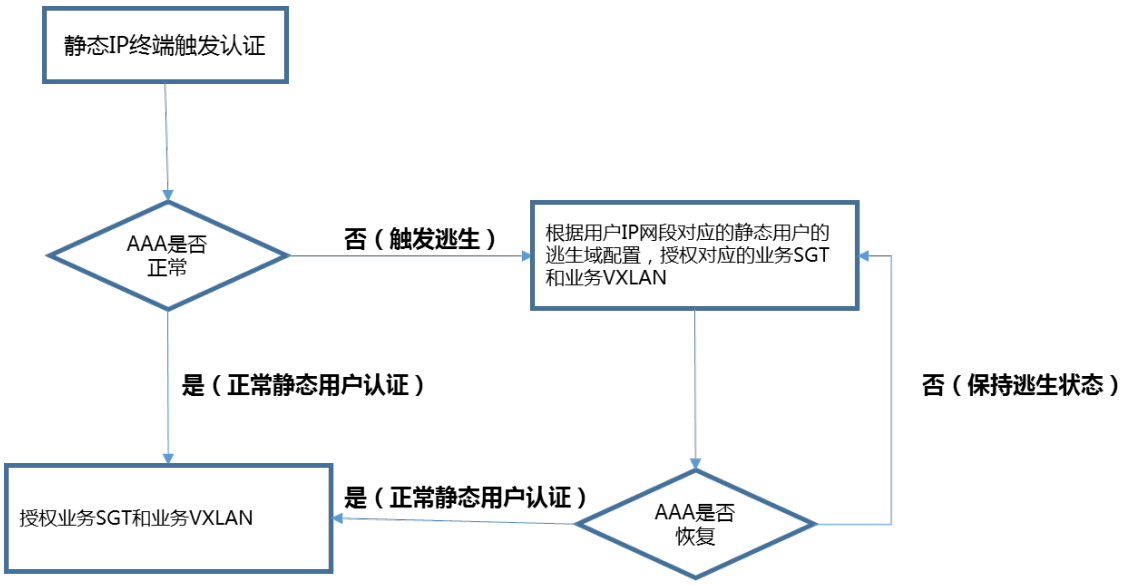
对于静态 IP 终端，由于之前已经配置了业务网段的 IP 和网关，逃生时授权进入逃生 VXLAN（即逃生网段）后，会导致终端的 IP 和逃生 VXLAN 网段不一致，流量不通。为了解决该问题，方案支持通过配置“静态用户”的方式实现静态 IP 终端的逃生，逃生时新终端可以进入对应的业务 VXLAN（即业务网段）。

认证点创建静态用户后，可以根据网段匹配认证域和逃生域，逃生域里可以定义授权 VXLAN 和微分段。

当 AAA 正常时，静态 IP 终端走正常认证流程，授权业务 VXLAN 和业务微分段。

当 AAA 故障时，静态 IP 终端匹配对应网段的逃生域，获得逃生域里定义的授权，仍然可以进入业务 VXLAN 和业务微分段，流量不受影响。流程图如下：

图74 流程图



14 网络运维设计

14.1 全网监控规划

AD-Campus 方案中，目前告警监控可以选择如下方案：

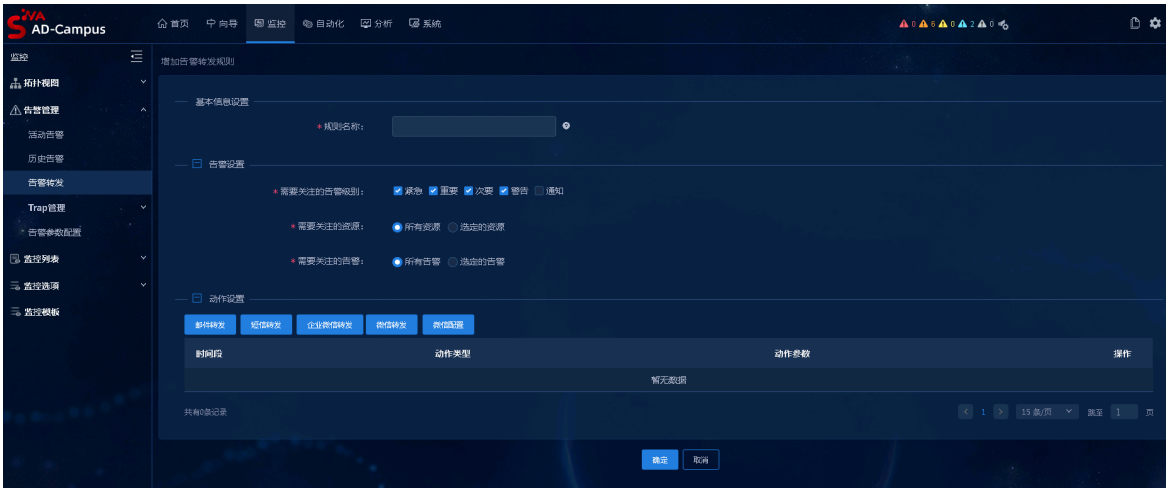
- AD-Campus 统一数字底盘作为网管监控平台。
- 各设备自行上报 Syslog 或 Trap 到客户第三方监控平台，由第三方监控平台实现告警功能。

实际项目开局中，可按现场实际情况选用上述方案中的一种进行部署。本文档仅对 AD-Campus 方案中各组件如何被监控作出说明。

14.1.1 统一数字底盘为监控告警平台

统一数字底盘作为监控告警的平台，目前统一数字底盘支持通过邮件、短信及微信的方式，将告警信息及时通知运维人员，方便系统维护。

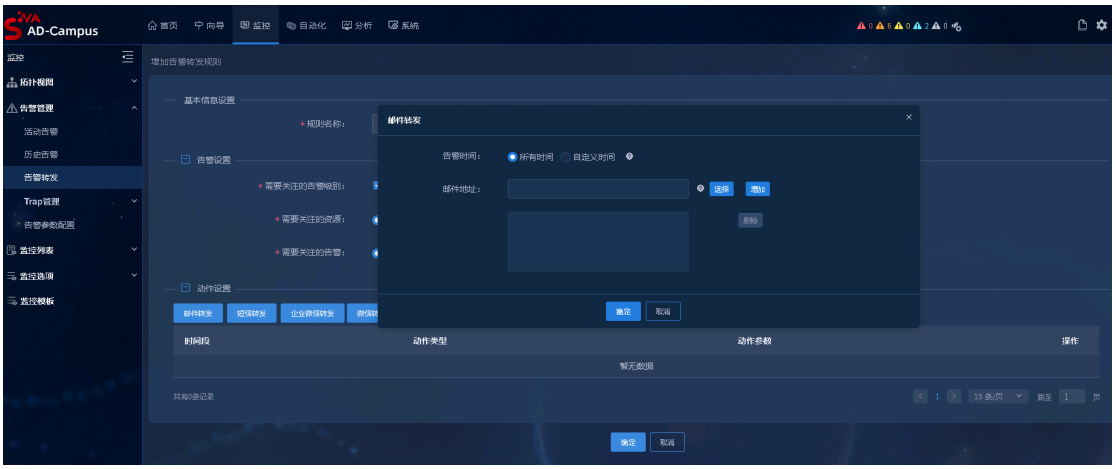
图75 告警转发



可根据需要的告警级别、关注的资源及关注告警进行邮件转发或者短信转发。

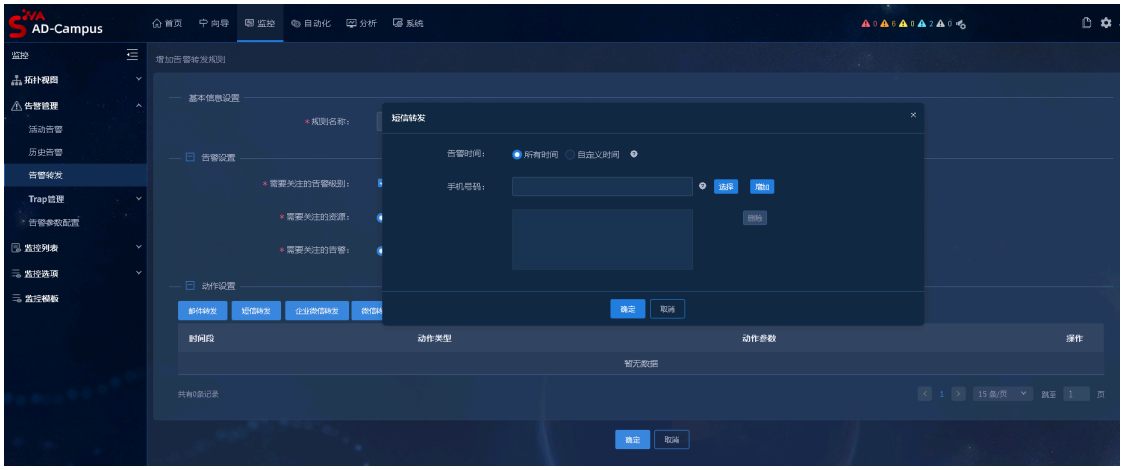
(1) 告警转发到邮件

图76 邮件转发



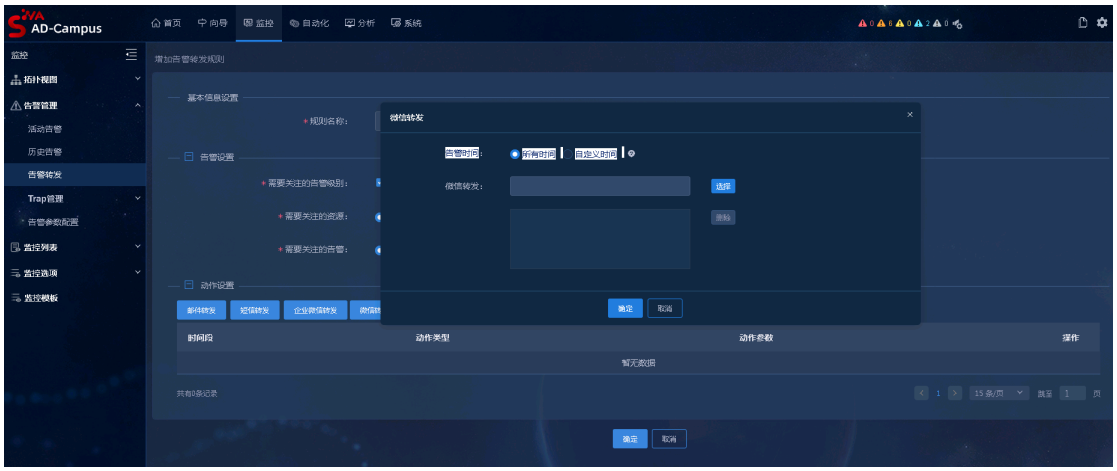
(2) 告警转发到短信

图77 短信转发



(3) 告警转发到微信

图78 微信转发



Trap 可根据需求升级为告警，监控平台也可以根据需求增加 Trap 升级为告警的规则。

14.1.2 网络设备的监控对接方式

网络设备一般通过 SNMP 与告警监控平台对接，由设备上报 Trap 主动显示告警，或由告警监控平台轮询设备 MIB 产生告警。

14.1.3 服务器产品的监控方式

服务器主要通过 iLO 口的 IPMI 协议与第三方告警平台对接，从而监控服务器的 CPU、内存、硬盘、风扇、温度等信息。可以在服务器操作系统上安装 SNMP 软件来进行对接。目前统一数字底盘安装在 CentOS 7.x 操作系统上，可以自行安装 SNMP 并进行相应配置，不影响统一数字底盘的正常功能。

对服务器要求能够监控内容有 CPU 状态、内存状态等，具体如下表所示，并能在状态异常时由告警平台自动产生告警。

监控项
CPU状态
内存状态
风扇状态
温度
电源状态
网卡状态
存储状态

14.1.4 设备监控，查看设备详情

主要是针对如下三类指标进行监控：

(1) 系统指标监控

(2) 接口指标监控

(3) IP 报文指标监控

图79 指标详情

模板类型	模板指标组	采集间隔	模板指标	指标参数
系统指标	CPU	5min	CPU利用率	告警级别、阈值、触发次数
	内存	5min	内存利用率	告警级别、阈值、触发次数
	设备缺陷	5min	设备响应时间	告警级别、阈值、触发次数
			设备不可达性比例	告警级别、阈值、触发次数
接口指标	接口统计	5min	接口接收速率	告警级别、阈值、触发次数
			接口发送速率	告警级别、阈值、触发次数
			接口输入带宽利用率	告警级别、阈值、触发次数
			接口输出带宽利用率	告警级别、阈值、触发次数
			接口接收广播包速率	告警级别、阈值、触发次数
			接口发送广播包速率	告警级别、阈值、触发次数
	接口告警统计	5min	接口输入包丢弃率	告警级别、阈值、触发次数
			接口输出包丢弃率	告警级别、阈值、触发次数
IP报文指标	IP报文统计	5min	接收IP报文速率	告警级别、阈值、触发次数
			转发IP报文速率	告警级别、阈值、触发次数
			输入IP报文丢弃率	告警级别、阈值、触发次数
			输出IP报文丢弃率	告警级别、阈值、触发次数

14.1.5 设备配置管理能力

园区配置管理能力包括设备版本升级、设备配置部署、设备配置备份和恢复。

- 设备版本升级
- 设备配置部署
- 设备配置备份
- 设备配置业务对比
- 设备配置恢复

统一管控平台：使用同一套管控平台实现传统园区网络和 VXLAN 园区网络的管控，实现传统园区的监控和配置管理。

传统园区自动化程度提高：将原本传统园区登录设备反复做的命令行工作，转变为监控任务、部署任务、备份任务，实现业务操作可定时、可反复执行，实现业务快速部署，简化运维工作量投入。

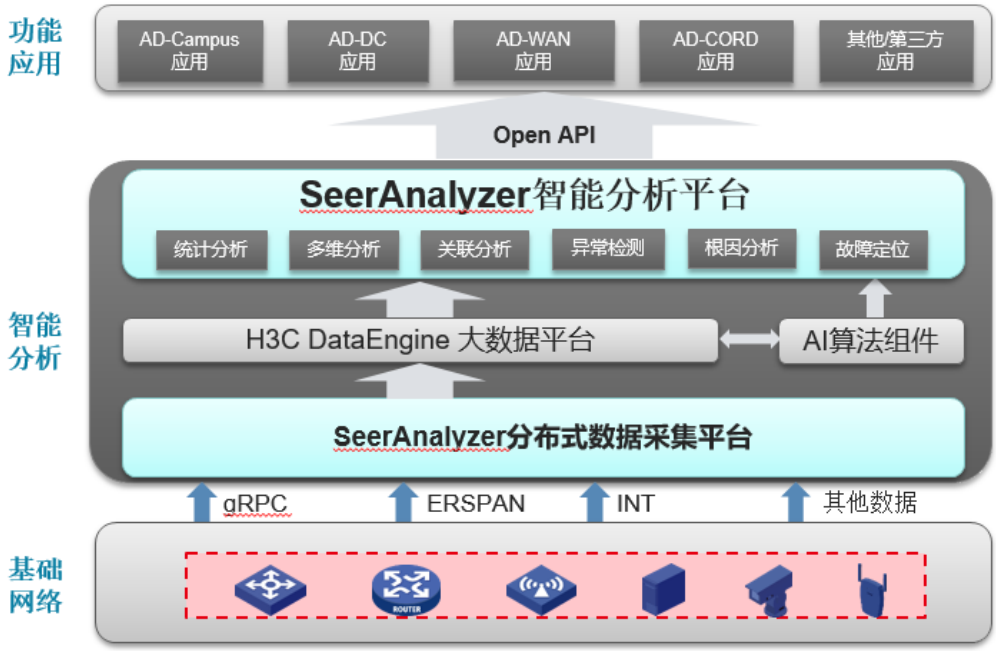
14.2 智能运维规划

14.2.1 SeerAnalyzer 分析组件可视化网络架构

SeerAnalyzer（先知分析组件）是新华三 AD-Campus 解决方案的核心组件，通过对设备性能、用户接入、业务流量的实时数据采集和状态感知，并通过大数据分析技术和 AI 算法，将网络的运行可视化，主动感知网络的潜在风险并自动预警。

SeerAnalyzer 针对不同业务场景采集不同的数据并进行针对性的分析，园区场景主要从对网络、用户、应用三个方面进行数据采集和分析。

图80 系统架构



- 智能应用：

SeerAnalyzer 提供北向开放 API，为 SeerAnalyzer 智能运维应用以及其他上层应用提供丰富的分析能力集。

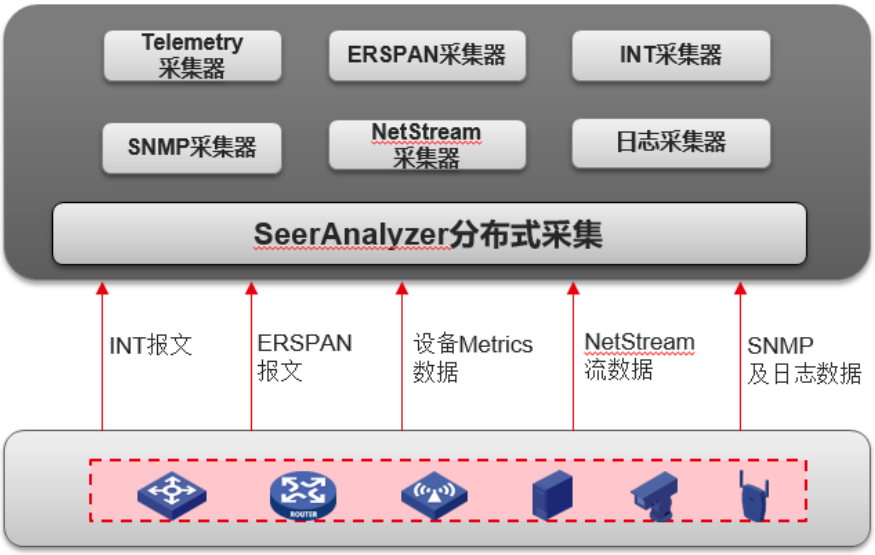
- 智能分析：

  SeerAnalyzer 采用 Spark、Flink 等分布式计算引擎以及 AI 人工智能模型库完成数据在线/离线分析任务，以满足智能运维应用及场景需求。

- 数据采集：

  SeerAnalyzer 采集器通过使用分布式部署架构，实现数据采集能力的横向扩展以满足不同网络规模的数据采集需求。

图81 数据采集

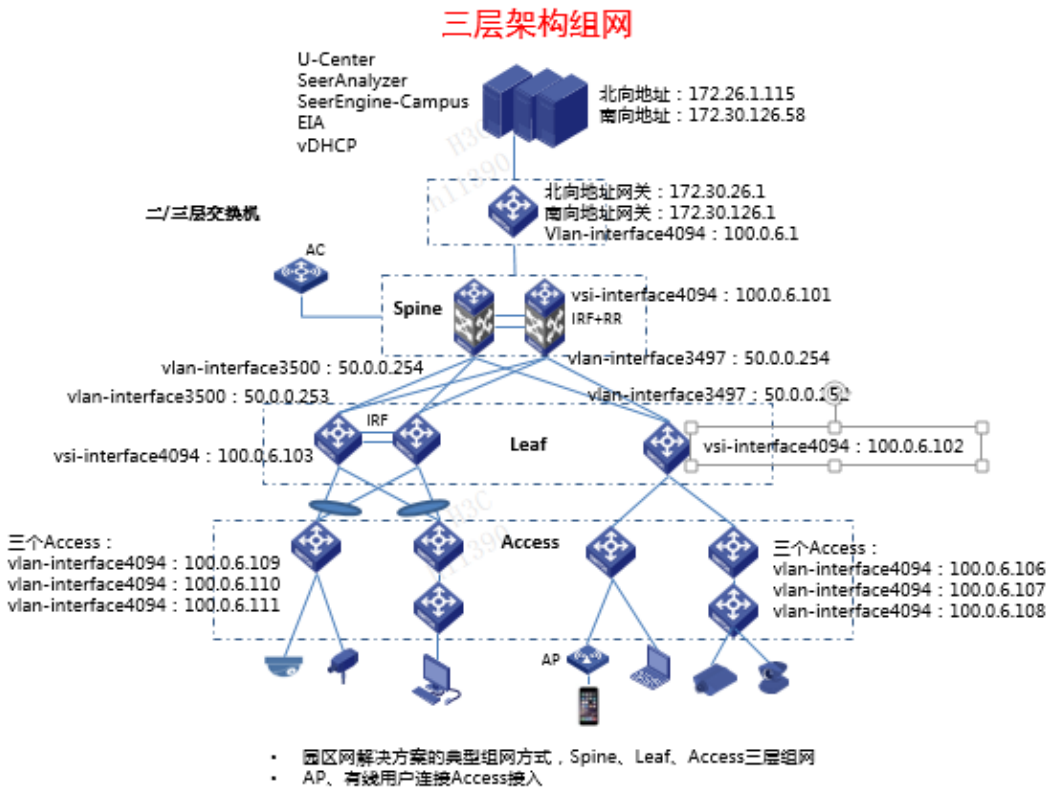


  - 支持 gRPC Telemetry 数据毫秒级采集，实时感知设备真实状态。
  - 支持 ERSPAN 和 INT 等最新的应用数据流采集技术，满足应用流量可视及应用质量分析等需求（园区后续支持 ERSPAN，INT 暂无计划）。
  - 分布式部署架构支持灵活的横向扩展，可根据网络规模按需部署采集器数量。
  - 全面支持各种数据采集技术，并可根据实际需要按需部署全部或部分采集器。

- 基础网络

  基础网络中的各种设备作为 SeerAnalyzer 的 Sensor 持续采集网络中的各种数据，以满足网络数据分析的要求。

图82 组网图



组网说明：

- 统一数字底盘上部署园区网络（SeerEngine-Campus、EIA、vDHCP），智能分析引擎（SeerAnalyzer ）。
- Spine、Leaf、Access 等交换机需要与 SeerAnalyzer 互通，AD-Campus 典型组网中在 vxlan4094 网络互通，管理交换机的 VLAN-interface4094 作为网关，其他组网，只需要设备与 SeerAnalyzer 可以互通即可。
- 在 Spine 和 Leaf 设备上开启 SNMP、Netconf、SSH、本地用户、syslog、gRPC。Access 设备除以上外，再开启 MAC information。
- 在分析组件上进行资产导入、配置采集模板和网络业务采集任务，完成设备故障采集、认证数据采集、健康度等分析任务。

14.2.2 数据采集规划

SeerAnalyzer 的数据采集分为三大部分：

- 设备侧数据采集；
- Cloudnet 无线数据采集；
- AAA 服务器、EPS 服务器、基础网管、控制组件等数据源的数据采集；

1. 设备侧数据采集

设备侧的采集需要消耗设备的CPU，多园区的话还需要消耗WAN网带宽，需要考虑以下几个方面：

(1) 客户需要的功能，不同的功能对应不同的数据采集项。只需要开启所需功能的数据采集，不要开启不需要的数据采集。

表7 数据采集项

功能模块	采集任务	采集项或 sensor path
网络健康度	SNMP	SNMP采集： nodeStatus 链路接口
	gRPC	nodeipindex gRPC采集： Device_Boards Device_PhysicalEntities Device_ExtPhysicalEntities
变更分析	netconf	获取设备配置 虚拟交换实例 ARP、ND、路由、MAC、L2VPN MAC等表项信息
	gRPC	ARP、ND、路由、MAC、L2VPN MAC等表项信息增量上报 mac/underlaymacevent mac/overlaymacevent arp_event/arptableevent nd/ndtableevent route_stream/ipv4routeevent route_stream/ipv6routeevent
链路分析	SNMP	光模块
	gRPC	设备侧配置sensor path： Ifmgr_Statistics Ifmgr_Interfaces
光模块诊断	SNMP	光模块
	gRPC	设备侧配置sensor path： Device/Transceivers Ifmgr_Statistics Ifmgr_Interfaces
网络拓扑	Netconf	获取LLDP邻居信息
端口指标监控	Netconf	fmgr_Interfaces

功能模块	采集任务	采集项或 sensor path
		Ifmgr_EthPortStatistics Ifmgr_Statistics
	gRPC	设备侧配置sensor path： Ifmgr/Interfaces Ifmgr/Statistics Ifmgr/EthPortStatistics
资产信息	从基础网管/控制组件同步	
容量管理	Netconf	获取表项资源
	gRPC	ResourceMonitor_Resources ResourceMonitor_ResourceEvent ResourceMonitor_Monitors
接口缓存监控	gRPC	BufferMonitor_IngressDrops BufferMonitor_EgressDrops BufferMonitor_PFCStatistics BufferMonitor_PFCSpeeds BufferMonitor_CommBufferUsages BufferMonitor_CommHeadroomUsages BufferMonitor_PortQueOverrunEvent BufferMonitor_PortQueDropEvent BufferMonitor_EcnAndWredStatistics
音视频质量分析	gRPC	设备侧配置sensor path： SQA/CallEvent SQA/BidirectionalCallEvent SQA/CallTrafficEvent
应用健康度iNQA	gRPC	设备侧配置sensor path： inqa/statisticses/statistics/amses/ams/losses/loss inqa/statisticses/statistics/losses/loss
有线用户旅程	gRPC	802.1X和MAC认证： DHCPSP/DhcpUserEvent MACA/MACAuthTrace Dot1X/Dot1XAuthTrace Portal认证： portal/portalusertraceevent
无线用户健康度及用	Cloudnet	

功能模块	采集任务	采集项或 sensor path
户旅程		
处置保障	Cloudnet	
无线诊断（一键诊断、doctor ap检测、无线安全检测）	Cloudnet	

数据采集的开启和关闭分两个部分：

- 在分析组件上需要设置对应的数据采集和分析任务。目前有一些重要的数据采集任务会提前预置，默认开启，比如设备基本信息的采集等。
- 在设备侧需要开启对应的采集，SNMP 和 Netconf 一般设备侧都已经设置好了建立连接的配置。GRPC 需要单独开启，并且通过设置 sensorpath 决定需要采集哪些数据。目前 GRPC 均需要在设备侧手工命令行配置。SNMP Trap 的采集也需要单独在设备侧开启。

(2) 根据设备能力选择数据采集方式和采集周期。目前已经采取增量上报、打包上报、超阈值暂停采集等方式减少数据采集对设备 CPU 的影响，但是对于一些低端交换机来说，本来 CPU 已经是业务争抢的资源，更要避免数据采集的消耗。选择数据采集方式时有如下建议：

- 对于 S5130S 这类 Access 设备来说，应该避免使用 GRPC 采集方式，仅适用 SNMP 和 Netconf 采集，并且要将 SNMP 和 Netconf 的采集周期设置较大值，减少对 CPU 的影响。
- 如果选用 S5560X 或 S6520X 设备作为 Leaf，由于 Leaf 同时承担用户接入认证和分布式网关的责任，本身 CPU 压力已经较大，建议仅对用户旅程采用 GRPC 方式采集，表项采集使用 GRPC 方式的话必须使用增量上报方式，同时对于 SNMP 和 Netconf 使用默认 5min 采集周期，不要改小采集周期。
- 如果需要做所有特性的数据采集，并且希望采用 GRPC 采集方式获得较高的数据采集精度，建议至少要选择 6550XE/6525XE 及以上级别的设备做 Leaf，并且需要为 GRPC 周期性全量采集配置合适的采集周期。一般除了端口 buffer 的性能监控数据采集外，其他特性 GRPC 数据采集周期建议不要小于 1 分钟。
- 如果设备采用堆叠的方式，，那么对于接口相关的数据采集，每次采集的数据量会随着堆叠设备数量增加而增加，则在超过 2 台堆叠的情况下，需要适当增加数据采集的周期时长。
- 对于传统组网的智能运维，有部分非 AD-Campus 方案适配的设备，需要跟分析组件产品研发确认数据采集的支持情况。
- 对于已有园区新增部署分析组件，需要考虑已有园区的网管或运维系统对设备侧已经有数据采集，比如 SNMP 采集，而分析组件需要多采集一份数据，对设备会叠加一些 CPU 压力，此时需要根据设备承受能力调整分析组件的采集周期，避免和已有系统的采集并发进行，造成设备侧 CPU 异常升高。

(3) 对于多园区场景，数据采集会产生一定量的 WAN 网流量，流量的大小与园区设备数量、用户数量、采集内容均有关系，需要确保数据采集和其他业务流量总和不要超过 WAN 网带宽。分为如下两种场景：

- 一套分析组件，跨 WAN 网进行设备侧数据采集，涉及两部分设备侧数据采集：

- 有线设备和用户的数据采集，分析组件已经提供了资源计算工具，输入有线设备数量和用户数即可自动计算有线侧数据采集所需带宽，建议采用资源计算工具来评估有线数据采集带宽。
- 无线设备和用户的数据采集，无线数据采集均是通过无线 AC 上报分析组件，则可以根据无线 AC 的数量评估无线数据采集量。每台无线 AC 数据上报不会超过 4M/s，多台无线 AC 进行叠加即可。如果还存在无线 AC 跨 WAN 网管理无线 AP，无线 AC 和 AP 之间主要是 CAPWAP 管理流量，需要根据 AP 的数量评估。

o 多套分析组件分级部署。一般在多个园区总规模超大，并且园区跨省距离较远时，考虑采用分级部署方式，一方面降低 WAN 网带宽，减少数据传输时延，一方面避免超出分析组件运维管理能力。分析组件分级部署情况下，主分析组件或子分析组件也可以同时负责本地多个园区的运维，数据采集规划可参考上一条，本地采集的数据已经在本地分析组件上完成处理，则主分析组件和子分析组件之间的数据同步对比直接采集设备侧数据减少了很多，对 WAN 网带宽的使用基本可以忽略不计。

2. Cloudnet 无线数据采集。

Cloudnet 是专门用来做无线数据采集和分析的组件，分析组件上的无线数据大部分来自 Cloudnet。

目前分析组件已经和 Cloudnet 做到了数据共享，Cloudnet 获取的数据经过初步解析后直接写入 kafka，而分析组件直接消费 kafka 中的数据。

对于分析组件来说，只要园区存在无线网络和用户需要智能运维，就必须配置 Cloudnet 组件（或可以配置打包了 Cloudnet 的 WSM 组件）。

3. 服务器侧数据源数据采集，

包括两种场景：

- 服务与分析组件可以融合部署统一数字地盘上，则分析组件与服务采用内部接口同步数据。
- 服务独立部署在其他服务器上，则需要确保分析组件北向 IP 地址与服务器之间 IP 可达。分析组件上可以将服务器添加为数据源，周期性从数据源同步数据。
- AAA 服务器目前支持 EIA、深澜、城市热点和天擎。

14.2.3 控制组件联动特性

分析组件上部分功能涉及与控制组件联动，在需要这些功能的时候，必须要部署了同一配套版本的控制组件。目前涉及控制组件联动的功能如下：

1. 无线 VIP 用户保障。

VIP 用户保障功能需要在分析组件创建 VIP 用户，还可以设置非 VIP 用户限速：

- 分析组件将非 VIP 用户限速的配置推送给控制组件，由控制组件下发非 VIP 限速配置给无线 AC。当 AP 上没有 VIP 用户存在时，非 VIP 用户限速配置不会生效，仅当 AP 上存在 VIP 用户时，才会对非 VIP 用户进行限速。
- 用户上线后，经过分析组件识别上线用户的用户名来判断是否是 VIP 用户，识别后，将 VIP 用户的 MAC 地址推送给控制组件，由控制组件将 VIP 用户的 MAC 地址下发给无线 AC，则无线 AP 可以对此 MAC 地址保障优先接入和高优先级转发。

2. 故障智能闭环

故障智能闭环分三种闭环方式：

- 预案类闭环，会将分析组件发现的问题和建议推送给控制组件，由控制组件来实施隔离 IP、重启单板等操作。这一过程需要控制组件根据问题描述进行业务编排，决定对哪台设备下发什么配置，来实现解决或隔离故障，由管理员审核控制组件下发的配置没有问题后，一键下发。
- 建议类闭环，会将分析组件发现的问题和建议推送给控制组件，但是不会形成解决问题或隔离故障的配置，需要管理员根据建议解决问题后，手工关闭问题。
- 通知类闭环，仅会将分析组件发现的问题通知给控制组件，不会给出建议。这类问题一般情况是可以自愈的，不需要管理员操作，只需要管理员可以查找到通知记录即可。

14.2.4 分析组件部署规划

分析组件支持物理服务器部署或者虚拟机部署，具体所需硬件以方案的硬件配置指导为准。

为分析组件提供 License 授权的 Licens server 服务器建议采用物理服务器部署，如果需要采用虚拟机部署，则需要勾选不允许虚拟机自动漂移。

分析组件可靠性规划：

- 如果要求具有可靠性，数据采集和分析不允许缺失或中断，则需要三机集群部署。如果同时有控制组件、EIA 部署，可融合部署在统一数字地盘。
- 如果对可靠性没有太高要求，可以选择单机部署。如果同时有控制组件、EIA 做三机集群部署，则需要单独准备服务器进行单机部署，此种部署方式可以降低控制组件三机集群所需硬件资源，也可以减小分析组件对控制组件可能的影响。

分析组件部署根据不同组网场景规划不同的部署方式：

1. 单园区场景

一般一套分析组件部署，根据设备数量和用户数量评估所需服务器硬件资源配置。

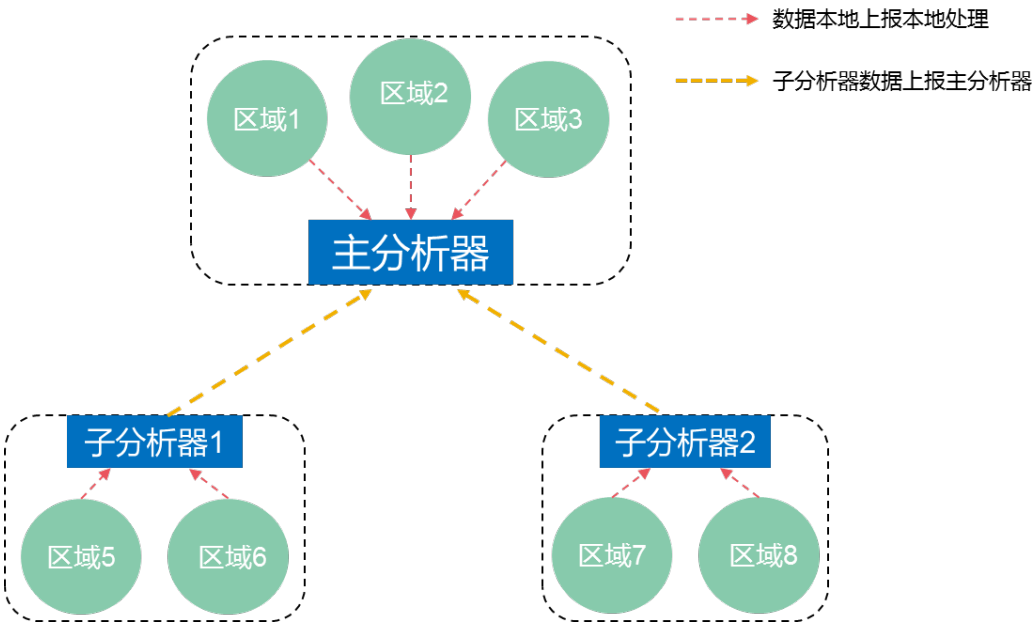
2. 多园区设备和用户规模在一套分析组件能力范围内

为了减少所需服务器资源，在 WAN 网带宽足够的情况下，这种场景可以采用多园区一套分析组件的方式部署。

3. 多园区设备和用户规模超出一套分析组件能力范围

这种场景需要部署多套分析组件，采用分级部署的方式。

图83 分级部署



根据多园区之间的 WAN 网带宽和每个园区的规模来评估如何分级，有以下几个原则：

- 将 WAN 网带宽较小的园区考虑在本地部署子分析组件，避免带宽小的 WAN 网承载设备数据采集。
- 根据多园区地域划分，将距离近的多个园区采用同一套分析组件运维，这也是考虑管理人员上，距离近的多园区可能是同一批人员运维。
- 存在多个园区一套分析组件时，需要确保不会超过分析组件承载能力。如果任意两个园区的规模总和均会超过一套分析组件的承载能力，就需要一个园区一套分析组件。

举例说明：比如园区 A、B、C，其中园区 C 出口带宽最小，园区 A 和园区 B 加起来的设备和用户数，并未超过一套分析组件承载能力，则在园区 C 部署一套子分析组件，在园区 A 或 B 部署一套主分析组件。

分级部署时，子分析组件可以根据承载园区的规模评估所需服务器硬件资源，主分析组件根据所承载的园区的规模评估所需服务器硬件资源，这里主分析组件承载园区不包含子分析组件承载的园区。主分析组件从子分析组件周期性同步的是子分析组件上各个区域的健康概要数据和问题，占用内存不会超过 1GB，所以不会提高主分析组件的硬件配置。

分级部署后，在主分析组件可以看到包含子分析组件所有分析组件管理区域的健康概要数据和未关闭问题列表，为统一评估多园区的健康质量提供帮助。如果要查看子分析组件管理园区的详细数据，可以一键免登录跳转到子分析组件对应页面查看。

4. 多域融合场景

当同时存在园区、数据中心、WAN 的多域运维需求时，分析组件可以通过部署一套分析组件加载不同授权的方式，满足园区、数据中心、WAN 的运维需求，而不需要分别部署分析组件。

分析组件会根据已经部署的控制组件所属领域，或者授权内容，呈现不同的综合菜单，体现为融合的数据呈现方式。例如同时有园区和 WAN 的运维需求，两者都有拓扑展示功能，分析组件将体现一个拓扑，将园区和 WAN 的拓扑统一呈现。

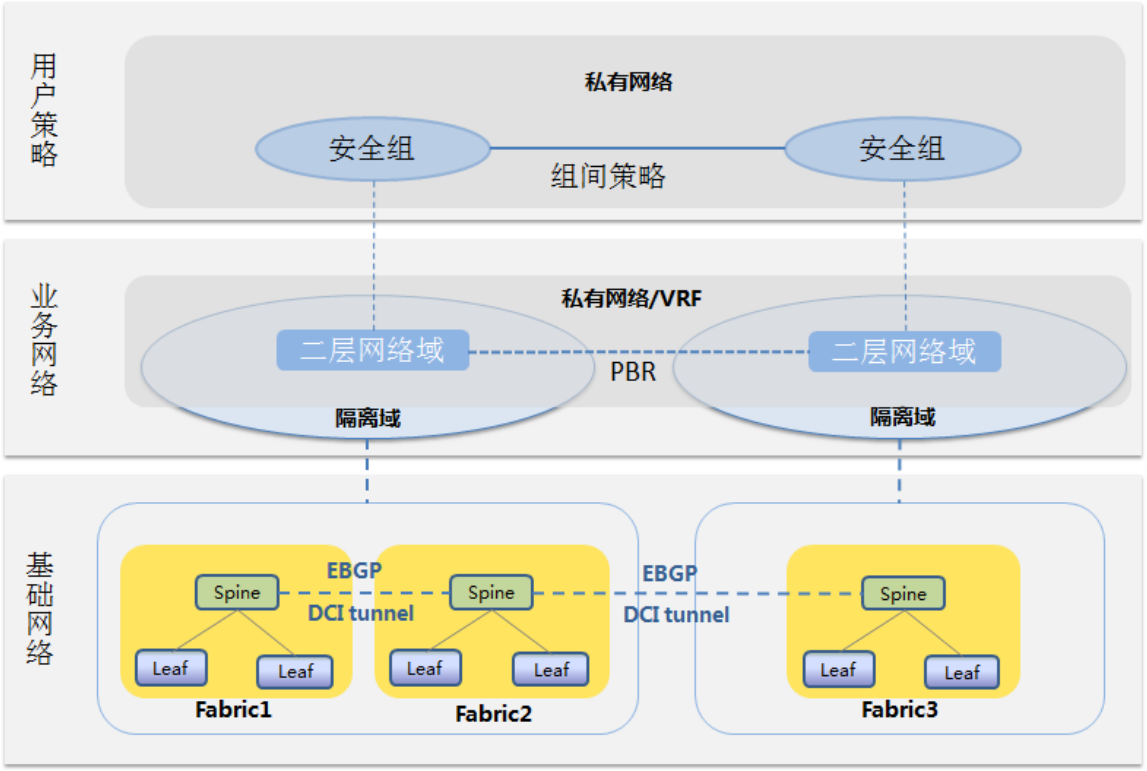
在多域融合场景，所需的服务器硬件资源，会比单域时所需资源高，但是也不是简单的叠加，可以根据资源计算工具来计算，并参考方案的硬件部署指导评估所需资源。

15 多园区设计

15.1 整体设计

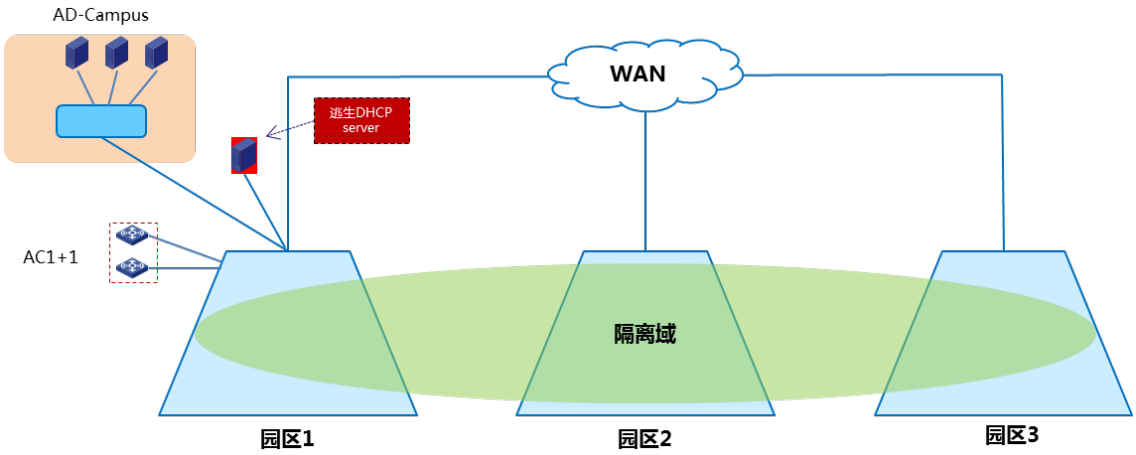
如下图，多园区采用分层设计思路，完成对用户业务的抽象。其中，基础网络层，通过对 Fabric 的管理，完成物理设备纳管及自动化部署；业务网络层，通过对隔离域、二层网络域的编排，完成用户网络的部署；用户策略层，通过定义私有网络、安全组、组间策略，完成用户策略编排。

图84 分层设计



15.1.2 典型组网 1：单隔离域，一套 AD-Campus

图85 组网图



组网说明：

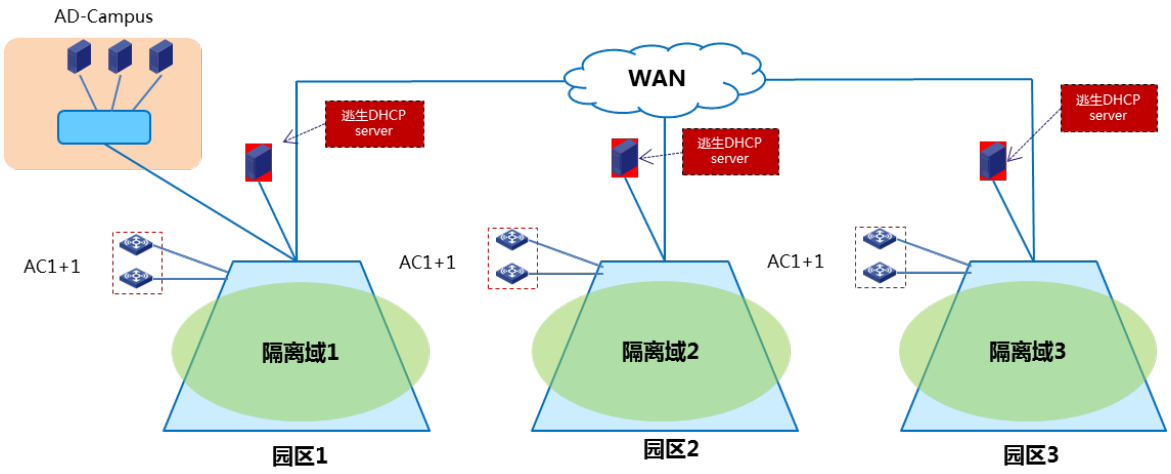
- 多个园区为同一个隔离域（传统 vlan 组网无此模型）；
- EIA/业务 DHCP 服务器集中部署（即和 AD-Campus 一起部署在主园区）
- AC：推荐集中式部署 AC
- 逃生 DHCP Server：推荐在总部的 spine 直接旁挂部署一台逃生 DHCP server（只有一个隔离域），推荐微软 DHCP server。考虑到此种场景的 HA 本身就不太高，如果有高 HA 需求建议选用典型组网 3 进一步提升 HA。

组网特点：

- 多个园区的网络策略（私网、VXLAN、组间策略）和用户策略（用户、角色(接入组)、场景、安全组）统一配置， 管理运维简单；
- 用户可在多园区之间移动， IP、权限均可保持不变；

15.1.3 典型组网 2：多隔离域，一套 AD-Campus

图86 组网图



组网说明：

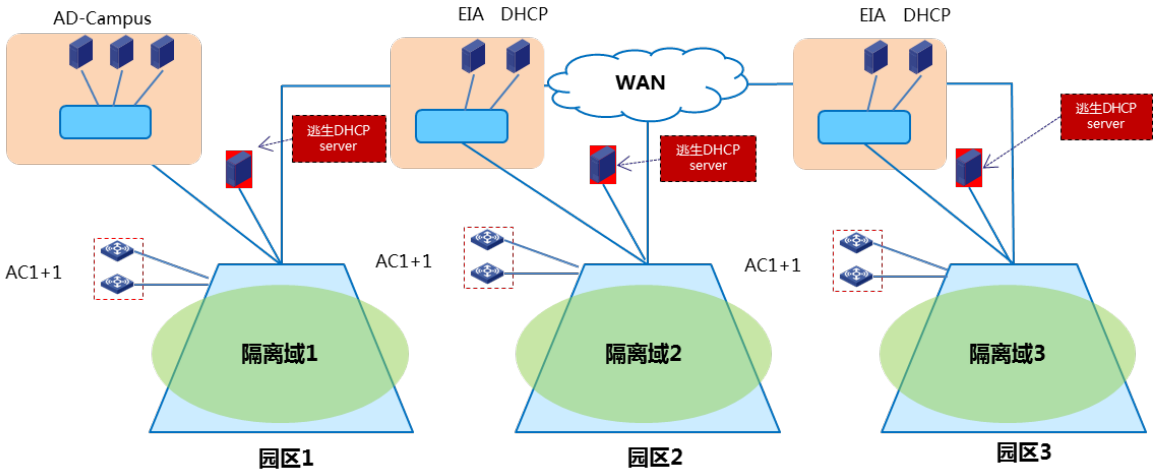
- 每个园区为一个隔离域；
- EIA/业务 DHCP 服务器集中部署（即和 AD-Campus 一起部署在主园区）
- AC：推荐每个园区部署一个（或一组）AC
- 逃生 DHCP Server：推荐每个隔离域部署独立部署一台逃生 DHCP server，推荐微软 DHCP server，旁挂 Spine 部署；此模型，若上行链路中断，业务地址会无法正常续约；若 HA 要求比较高，推荐典型组网 3 模型。

组网特点：

- 同一个用户组在不同的园区中对应不同的安全组（IP 网段）
- 用户在多园区间移动，IP 网段会发生变化

15.1.4 典型组网 3：多隔离域，一套 AD-Campus，多套 EIA/DHCP

图87 组网图



组网说明：

- 每个园区为一个隔离域；
- 每个园区部署一组 EIA/业务 DHCP 服务器（主 EIA 放在总部，二级 EIA 放在分园区，当分园区上行链路故障时，不影响分园区认证。）
- AC：推荐每个园区部署一个（或一组）AC
- 逃生 DHCP Server：推荐每个隔离域一个单独的逃生 DHCP server 旁挂 spine 独立部署，推荐微软 DHCP server。

组网特点：

- 该组网 HA 能力优域组网 1 和组网 2，可靠性较高。
- 同一个用户组在不同的园区中对应不同的安全组（IP 网段）
- 用户在多园区间移动，IP 网段会发生变化

15.2 多园区分层设计

1. Fabric

Fabric 是指由单个 Spine/Leaf/Access 或者 Leaf/Access 等标准模型构成的网络架构，具备独立的设备自动化能力。每一个 Fabric 属于一个网络物理连通域，多个 Fabric 之间可通过 WAN 或其他方式互联。随着企业规模的变大，一套 Fabric 架构已经无法满足用户正常的网络部署需求，用户可通过扩展多 Fabric 的方式进行网络的横向扩展。Fabric 可通过建立 EBGP 邻居的方式拉通多 Fabric 间的用户路由，从而达到 Fabric 间用户互访的目的。

2. 隔离域

隔离域是指由一个或者多个 Fabric 组成一个网络连通域，每个隔离域可具有独立的认证系统、DHCP 服务器、无线 AC 控制组件等网络服务，实现网络服务的本地化部署，从而减少远端管理带来的网

络带宽的消耗和降低网络延迟。用户可通过建立多个隔离域的方式实现多园区的统一运维和管理，单个隔离域中支持配置多个 Fabric 的设计增强了单园区的网络的扩展性，两个层面的设计，为网络的实施提供了更灵活的管理部署方案。多个隔离域间，可通过建立 EBGP 邻居的方式拉通多隔离域间的用户路由，从而达到隔离域间用户互访的目的。单个隔离域中支持配置多个 Fabric 的设计。

3. 二层网络域

定义用户 Overlay 网络，每个二层网络域只属于一个隔离域。每个二层网络域的都能够覆盖隔离域下的所有 Fabric，从而实现多 Fabric 的 IP 随行、名址绑定。

4. 私有网络

为用户定义一个逻辑上的私有隔离网络，不同私网之间默认网络隔离。实现上每个私网为一个 VRF。通过在不同隔离域下创建二层网络域，可实现用户私有网络跨隔离域部署。

5. 安全组

定义用户网络权限，通过与 EIA 的协同，完成用户到安全组的绑定。用户上线后，通过认证授权用户到不同安全组，为用户提供不同的网络权限。

6. 组间策略

定义安全组间的访问策略，控制组件通过把组间策略转换成 PBR 配置下发到网络设备，实现网络访问权限的控制。

7. 互联链路要求

园区 Fabric 之间如果是跨广域网，需要确保 MTU 不小于 1600；

园区之间链路时延的建议小于 100ms，如果 EIA 位于远端园区，链路时延建议小于 50ms。

8. 多园区 SA 的部署

多园区场景，建议总部园区 3 节点集群部署（和 Campus 控制组件融合部署,建议控制组件部署在 master 节点，分析组件部署在 slave 节点），分支园区小规模单机独立部署 （需要手工导入资产），总部园区分析组件可以汇总分支的健康 KPI 数据和问题 。

多园区场景，若是分支之间的 WAN 通道带宽满足，可以集中在总部部署 SA，分支无需部署 SA。通道带宽要求与设备数量、用户数有关。例如 5000 用户需要 64MB/S，1W 用户需要 128MB/S，2W 用户需要 256MB/S，通常情况下 5000 用户对应 1000 设备，1W 用户对应 2000 设备，2W 用户对应 4000 设备 。具体请联系研发评估。

9. 多园区 EIA、DHCP 和 AC 部署

- 目前多园区多 Fabric 建议的场景是总部部署 SeerEngine-Campus、vDHCP 和 EIA V9，各个园区分别部署本地业务 DHCP、分级 EIA 和逃生 DHCP。
- 多园区 EIA，建议总部部署上级 EIA，其他园区部署下级 EIA。
- 业务 DHCP 服务器各园区需要独立部署。
- 考虑到园区之间链路的中断，建议各个园区部署逃生方案。
- BYOD 和自动化相关业务，建议总部部署 vDHCP 来提供。
- 多园区，EIA、业务 DHCP 和 BYOD DHCP 可以复用，但是业务 DHCP 不推荐复用。
- 多园区多 Fabric 分别各自的 AC 控制组件，和单园区部署 AC 控制组件没有区别。单园区多 Fabric，只能配置一个无线二层网络域，添加多个 AC 控制组件，各自的 AC 控制组件配置各自的 AP 模板，AC 控制组件需要关闭自动注册。

15.3 多园区出口

15.3.1 共享出口

整个网络环境中只有一个出口,多个园区可通过网络规划实现多个园区共用一个出口访问 Internet。

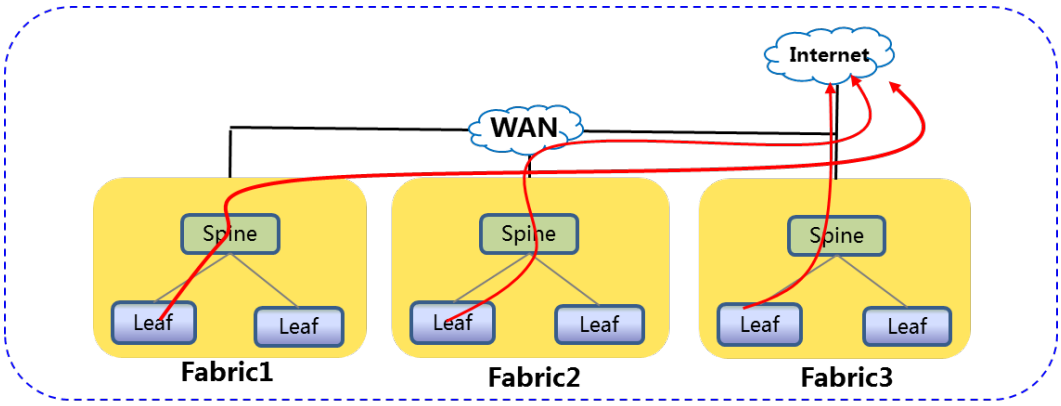
实现原理：

(1) Fabric 间通过建立 EBGP 邻居完成路由同步。

(2) Fabric 设立出口，通过发布静态默认路由方式，把默认流量引向出口。

(3) 默认路由通过 EBGP 邻居通告给其他 Fabric。

图88 组网规划



15.3.2 多出口备份

存在多个出口时，用户可通过对网络的规划实现出口备份。

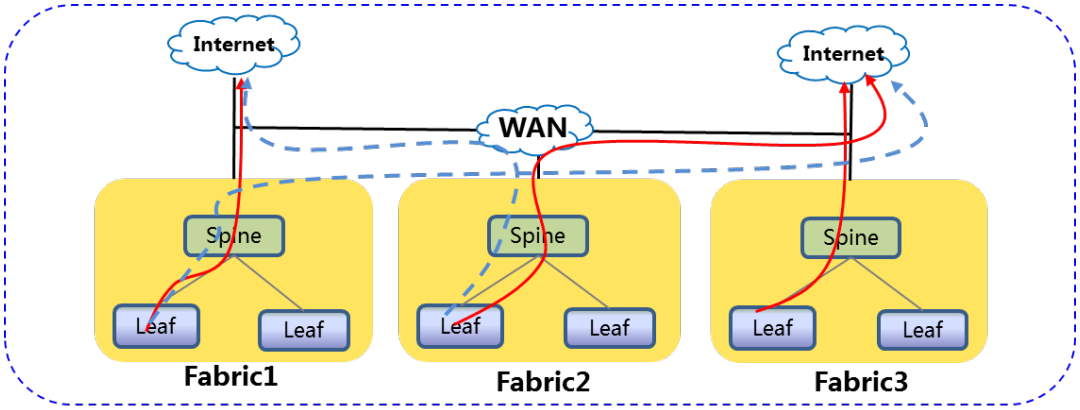
实现原理：

(1) Fabric 间通过建立 EBGP 邻居完成路由同步。

(2) Fabric 设立出口，通过发布静态默认路由方式，把默认流量引向出口。

(3) 默认路由通过 EBGP 邻居通告给其他 Fabric。

图89 组网规划



出口优选原则：

- Fabric 本地设有出口时，本地优先，如 Fabric1。
- 本地没有出口时，会根据 BGP 协议的路由优选原则进行优选，如 Fabric2。

BGP 协议的路由优选原则：

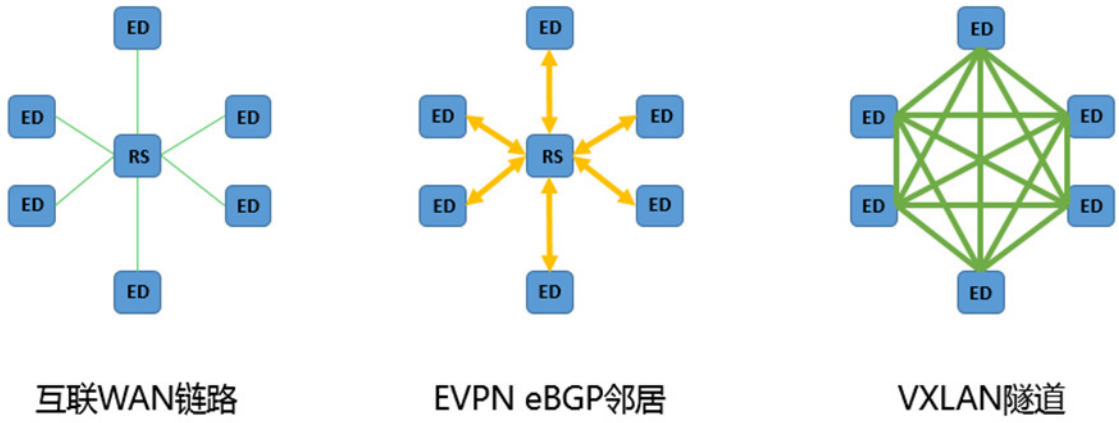
- 丢弃下一跳（NEXT_HOP）不可达的路由。
- 优选首选值（Preferred-value）最大的路由。
- 优选本地优先级（LOCAL_PREF）最高的路由。
- 依次选择 network 命令生成的路由、import-route 命令引入的路由、聚合路由。
- 优选 AS 路径（AS_PATH）最短的路由。
- 依次选择 ORIGIN 类型为 IGP、EGP、Incomplete 的路由。
- 优选 MED 值最低的路由。
- 依次选择从 EBGP、联盟 EBGP、联盟 IBGP、IBGP 学来的路由。
- 优选 IGP Metric 值最小的路由。
- 优选迭代深度值小的路由。
- 如果当前的最优路由为 EBGP 路由，则 BGP 路由器收到来自不同的 EBGP 邻居的路由后，不会改变最优路由。
- 优选 Router ID 最小的路由器发布的路由。如果路由包含 RR 属性，那么在路由选择过程中，就用 ORIGINATOR_ID 来替代 Router ID。
- 优选下一跳地址为 IPv4 地址的路由。
- 优选 CLUSTER_LIST 长度最短的路由。
- 优选 IP 地址最小的对等体发布的路由。

15.4 路由服务器

15.4.1 功能简介

在多园区场景下，需要在不同 AS 域中的边界设备间，建立 EBGP 对等体的全连接关系。当园区数量比较多时，对等体数目会很多，全连接对网络资源和设备性能的消耗很大。因此方案引入了路由服务器（Route Server）功能，采用类似 IBGP 路由反射的方案，在主园区配置一台设备作为 Route Server，需要建立全连接的边界设备与 Route Server 建立 eBGP 对等体，这些边界设备将作为 Route Server 的客户机，如图所示。Route Server 向客户机发送路由时，不修改路由的 AS_Path、Next_Hop 和 MED 属性，使客户机之间不需要建立 EBGP 全连接也能学习到彼此的路由，客户机之间的流量转发也不需要经过 Route Server。

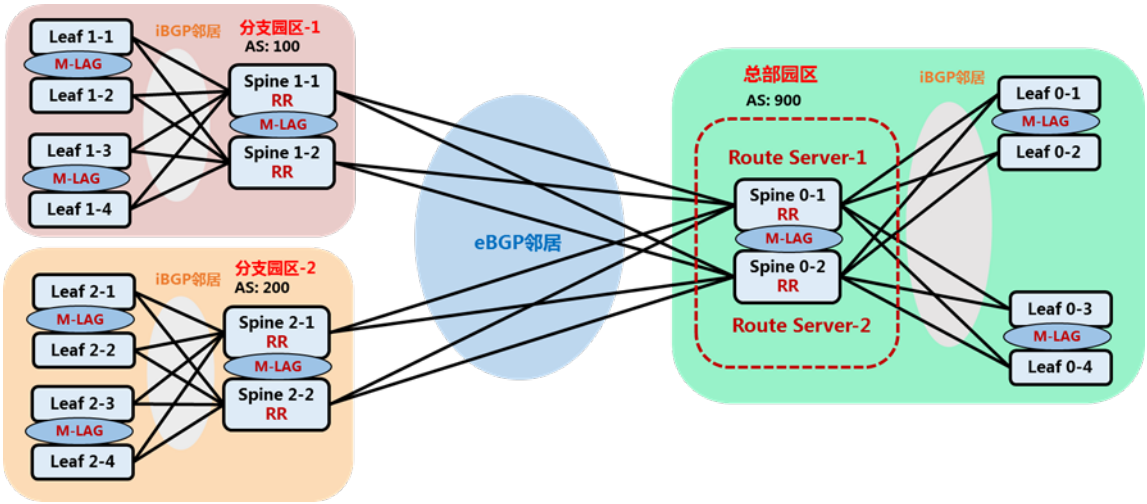
图90 路由服务器



15.4.2 M-LAG 场景

在 M-LAG 场景， ED 设备均为 M-LAG 组网，总部 ED 设备两个 M-LAG 成员设备均作为路由服务器（RS），分支 ED 设备的两个 M-LAG 成员设备均需与这两台 RS 使用 loopback 实地址建立 EBGP 邻居，通过路由扩散，最终通过 EVPN 在所有 ED 之间建立虚隧道。

图91 M-LAG 场景



16 组播设计

园区的组播场景可分为：

- 单 Fabric 组播：组播源和接收者属于同一 VPN，且分布在同一 Fabric。该场景下可以使用二层组播方案或 EVPN 三层组播方案。
- 多 Fabric 组播：组播源和接收者属于同一 VPN，可以分布在不同 Fabric，但 Fabric 之间走 DCI 隧道。该场景只能用 EVPN 三层组播方案。

- 跨域组播：组播源和接收者属于同一 VPN，可以分布在不同 Fabric，Fabric 之间跨 WAN，走的是普通 IP 转发。该场景下，除了要用到 EVPN 三层组播，还涉及其他协议，比较复杂，具体介绍下后面章节。

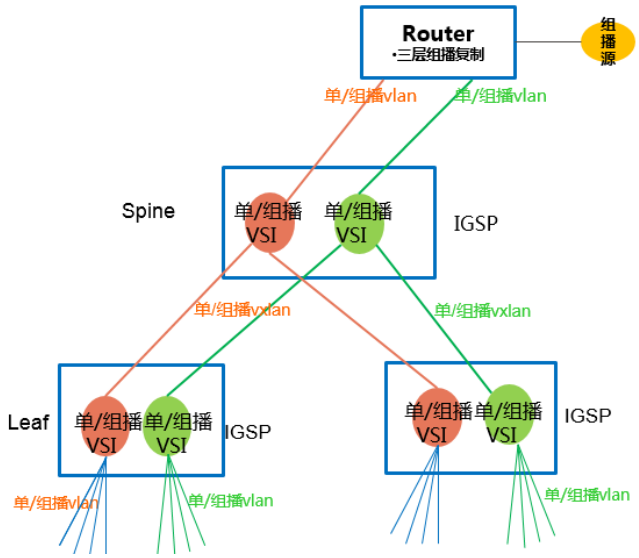
16.1 单Fabric组播方案

单 Fabric 组播方案有两种：二层组播方案和 EVPN 三层组播方案。

16.1.1 二层组播方案

二层组播方案可以解决同一 Fabric 内同 VPN 内的组播需求。它适用于组播接收者较少或组播流量较小的场景，支持 IPv4 和 IPv6 组播。

图92 二层组播方案



该方案要求组播源通过一台外部的 Router（Router 可以是组播路由器或支持组播的三层交换机），接入到 Spine 设备。接收者可以从 access 接入。

实现原理如下：

- Router 负责三层组播复制，即：将组播流从一个 VLAN 复制到其他 VLAN 中；
- Spine 负责将组播流量从 VLAN 映射到 VXLAN（VLAN 跟 VXLAN 是一一映射）；Spine 和 Leaf 的 VSI 实例运行二层组播；
- 当 Spine 收到来自 Router 的组播报文后，经过单播 VXLAN 隧道发给 Leaf；
- Leaf 收到报文后，解封装，再在相应的 VSI 实例内，根据二层组播表项转发到相应的 AC 口。

该方案的优点：

- 配置简单，仅需要 Spine 和 Leaf 在 VSI 实例内运行二层组播；
- 对 Access 交换机无要求。

缺点：

- 要求组播源通过 Router 接入 Spine；

- 适合组播接收者较少或组播流量较小的场景。

16.1.2 EVPN 三层组播方案

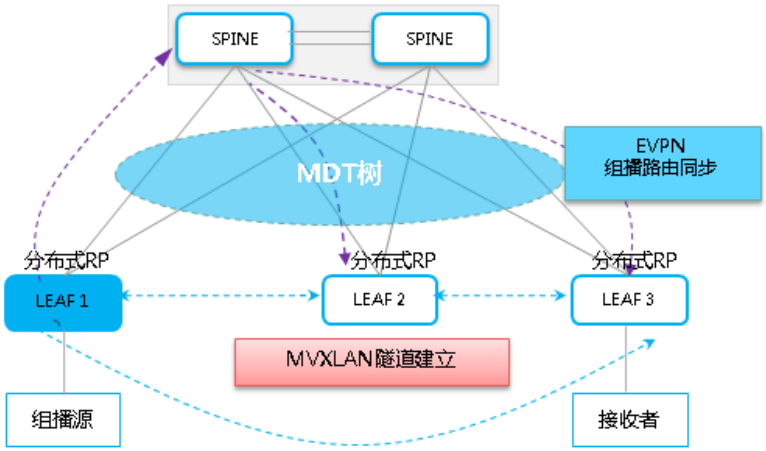
单 Fabric 的 EVPN 三层组播，组播源和接收者接入位置可以任意，但是如果组播源或接收者从 spine 以外网方式接入（即：以 vlan-interface 接入），需要特定的板卡。

1. EVPN 组播的控制平面

为了支持 EVPN 组播，MP-BGP 在 EVPN 地址族新增如下 2 类 EVPN 路由，用于创建 MDT 树。

- Supplementary Broadcast Domain Selective Multicast Ethernet Tag Route：增强型广播域选择性组播以太网标签路由，也叫 SBD-SMET 路由（6 类路由），包含私网组播源地址和组播组地址信息，用于接收者侧的 VTEP 通告希望接收某个(*，G)或(S，G)的组播流量。
- Selective Provider Multicast Service Interface Route：选择性组播业务接口路由，也叫 S-PMSI A-D 路由（10 类路由），包含私网组播源地址、私网组播组地址、Default-Group 或 Data-Group 地址及 MVXLAN 源接口地址。主要用于：
    - 组播源侧 VTEP 与其所有 BGP 邻居间建立 Default-MDT。
    - Default-MDT 向 Data-MDT 切换。

图93 Default-MDT 树的建立



SPINE 和 LEAF 组成 EVPN 组播的公网，在公网上运行 PIM-SM 协议；LEAF 连组播接收者侧组成 EVPN 组播的私网侧，私网侧运行 PIM-SM 或 PIM-SSM。

Default-MDT 树的建立过程：

(1) 组播源侧的 LEAF（LEAF1）向其所有 BGP 邻居（即：LEAF2 和 LEAF3），发送携带（*，*）信息的 10 类路由，开始创建 default-MDT。

(2) LEAF2 和 LEAF3 收到 10 类路由后，路由中携带的（*，*）信息触发 LEAF2 和 LEAF3 发送公网 PIM 加入信息，并在公网沿途建立组播表项，形成以 LEAF1 为根，LEAF2 和 LEAF3 为叶子的 SPT 树，也就是 default-MDT 树。

在公网中通过 default-MDT 传送组播数据时，组播报文被传输到支持同一 VPN 实例的所有 Leaf 上，无论该 Leaf 是否下挂接收者，故 default-MDT 不是按需转发。Data-MDT 树可以做到按需转发，原理如下：

当私网组播数据通过了由 Default-MDT 向 Data-MDT 切换的 ACL 规则的过滤时，由组播源侧 LEAF 发起 Default-MDT 向 Data-MDT 的切换，切换过程如下：

(1) 源端 LEAF 从配置的 Data-Group 范围中，选取一个引用次数最少的 Data-Group 地址，并将其通过 10 类路由发送至远端 LEAF，该路由中包含私网组播源地址、私网组播组地址、源端 VTEP 上 MVXLAN 源接口地址、Data-Group 地址。

(2) 远端 LEAF 收到 10 类路由消息后，检查本地是否有私网组播流量的接收者：如果有，则回复加入信息加入以组播源所在的 LEAF 为根的 Data-MDT；如果没有，则将该消息缓存起来，等待有接收者时，直接回复加入信息加入 Data-MDT。

(3) 当组播源端的 LEAF 发送 10 类路由消息一定时间后，该 LEAF 会停止使用 Default-Group 地址对私网组播数据进行封装，并改用 Data-Group 地址进行封装，组播数据沿 Data-MDT 向下分发。

(4) Default-MDT 切换到 Data-MDT 之后，当某下游 LEAF 不再连接接收者时，可以通过发送 PIM 剪枝消息退出 Data-MDT。

当情况变化导致不满足切换条件时，组播源侧的 LEAF 会从 Data-MDT 反切回 Default-MDT，反切过程与 Default-MDT 切换为 Data-MDT 相同。只要满足如下条件之一，就会进行反向切换。

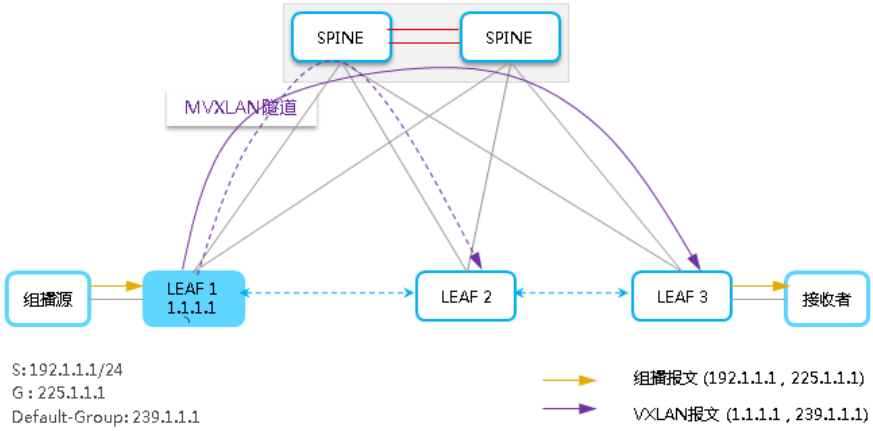
(1) 更改 Data-Group 范围后，用于私网组播数据封装的 Data-Group 不在新的范围之内。

(2) 控制私网组播数据由 Default-MDT 向 Data-MDT 切换的 ACL 规则发生变化，私网组播数据不能通过新 ACL 规则的过滤。

2. EVPN 组播的转发平面

当 default-MDT 建立完成后，组播源即可通过 default-MDT，将私网组播数据发送给组播接收者。以下图为例，介绍 EVPN 网络中，组播报文的传输过程。

图94 组播数据报文传输过程



如上图所示，LEAF1 下挂的组播源发送组播报文给 LEAF3 下挂的组播接收者。以先有接收者加入，后有流量打入这种场景，介绍组播报文跨越公网的传输过程。

(1) LEAF3 连的接收者发送 IGMP report 报文给 LEAF3，LEAF3 收到后，本地生成 6 类路由，并携带 L3VPN 实例的 RT，发给所有其他 LEAF。

(2) 其他 LEAF 收到（*，G）的 6 类路由后，更新本地组播路由表，由于是分布式 RP，故组播表项的入接口是本设备，出口是 MTunnel 口。

(3) 组播源发送私网组播报文（192.1.1.1,225.1.1.1）到 LEAF1 。

(4) LEAF1 查本地组播路由表，如果出接口是本地，则按照普通三层组播处理；如果出接口在远端，则将私网组播报文进行 MVXLAN 封装（报文外层源地址是 LEAF1 的 loopback0 的 IP 地址，报文外层组地址是 default-group 地址），沿着已经创建的 default-MDT 将封装后的组播报文发给其他 LEAF 设备。

(5) 组播报文到 SPINE 设备后，SPINE 根据本地公网组播转发表项，将报文转发给其他 LEAF 设备。

(6) 组播流量达到远端 LEAF 后，远端 LEAF 发现组播流量是从公网侧进来，将对应的组播表项的入接口修改为 MTunnel 口，出接口为本地 vsi-interface 接口。后续远端 LEAF 收到组播报文后，先解封装 VXLAN 报文，还原成原始的私网组播报文，再根据本地的组播转发表项，将报文发给接收者。LEAF2 收到组播报文后，发现本地没有接收者，丢弃组播报文。

注：如果根据 data-MDT 转发，则私网的组播报文只会发给下挂接收者的 leaf，也就是私网组播报文只发给 leaf3，不会发给 leaf2，即：按需复制。

EVPN 组播的优点：

- EVPN 组播可以实现按需复制，仅把组播流发给有接收者的 LEAF 设备；
- 实现同 VPN 内跨 VXLAN 的三层组播转发；
- EVPN 组播使用核心复制，减轻头端复制的压力；
- 组播源和接收者的位置任意。

EVPN 组播的缺点：

- 实现比较复杂；
- 暂不支持 IPv6 组播。

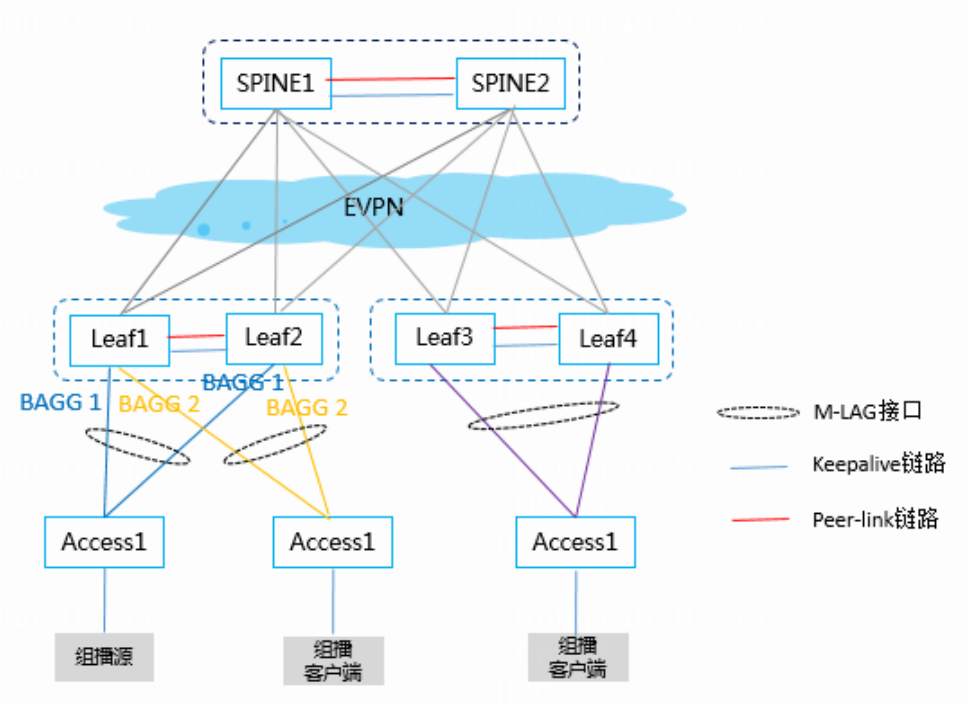
16.2 EVPN组播支持M-LAG

EVPN 组播利用 M-LAG（Multichassis link aggregation，跨设备链路聚合）将两台物理设备连接起来虚拟成一台设备，避免设备单点故障对网络造成影响，从而提高 EVPN 组播的可靠性。

如下图所示，在 EVPN 组播组网中，Leaf 和 Spine 设备均支持 M-LAG。M-LAG 组网下，不支持组播源和接收者单挂接入，单挂接入解释如下：

Access 设备仅接入 M-LAG 系统的其中一台设备，则该 access 设备称为单挂设备，这种接入方式称为单挂接入。组播源和接收者接入单挂 access 设备，则此时组播源和接收者是单挂接入。或者，组播源和接收者直接连接在 M-LAG 系统的一台设备，此时组播源和接收者也是单挂接入。

图95 EVPN 组播支持 M-LAG 组网



EVPN 组播支持 M-LAG 通过 peer-link 链路在组成 M-LAG 系统的成员设备间同步组播流量和组播接收者加入请求（IGMP 成员关系报告报文或者 PIM 加入报文），使成员设备上的组播源和组播接收者信息保持一致，形成设备级备份。当一台成员设备发生故障（设备故障、上下行链路故障等）时，组播流量可以由另一台成员设备进行转发，从而避免组播流量转发中断。

如图所示，以 Leaf 1 和 Leaf 2 组成的 M-LAG 系统为例，EVPN 组播支持 M-LAG 的工作机制为：

(1) Leaf 1 和 2 做 M-LAG，与其他 VTEP 建立 MVXLAN 隧道，组播源地址是虚拟地址、目的地址是相同的 Default-group 地址；

(2) Leaf 1 从 BAGG2 收到组播接收者发送的加入请求报文后，通过 peer-link 链路将加入请求报文同步到 Leaf 2。

(3) Leaf 1 和 Leaf 2 均根据加入请求建立相应的组播转发表项，并向组播源侧的 Leaf 发送 6 类路由。

(4) Leaf 1 从 BAGG1 接口接收到组播源发送的组播流量后，通过 peer-link 链路将组播流量转发至 Leaf 2。

(5) 如果开启 Underlay 组播负载分担，则组播流量在 Leaf3 和 4 做分担，发给 Spine。如果未开启，则根据组播选路原则（最长匹配/下一跳的 IP 地址大小等），由 Leaf3 和 4 中的一台将组播流量发给 Spine 设备。

(6) 流量到了 Spine，Spine 根据公网的组播表项将流量转发给 Leaf 设备。

(7) Leaf1 和 2 收到后，先解 MVXLAN 封装，然后再发给 Access 设备。组播流在 2 台设备间采用奇偶原则进行负载分担，即：M-LAG 系统编号为 1 的成员设备 转发 组播组地址为奇数的流量，M-LAG 系统编号为 2 的成员设备 转发 组播组地址为偶数的流量。

(8) 如果私网组播数据满足 Data-group 切换条件，则要由 Default-group 向 Data-group 切换。M-LAG 系统的主设备（假设为 Leaf 3）负责选取 Data-group，进行 Default-group 向

Data-group 的切换，并通过 10 类路由将选取的 Data-group 通告给 Leaf 4。Leaf 4 接收到 10 类路由后，Leaf 4 使用相同的 Data-group。

16.3 单隔离域多Fabric组播方案（Fabric之间走DCI隧道）

该场景下，组播源和接收者属于同一个 VPN，可以分布在不同的 fabric，但组播源和接收者不能接在 ED 设备上，并且每个 fabric 不能是单 leaf 模型。

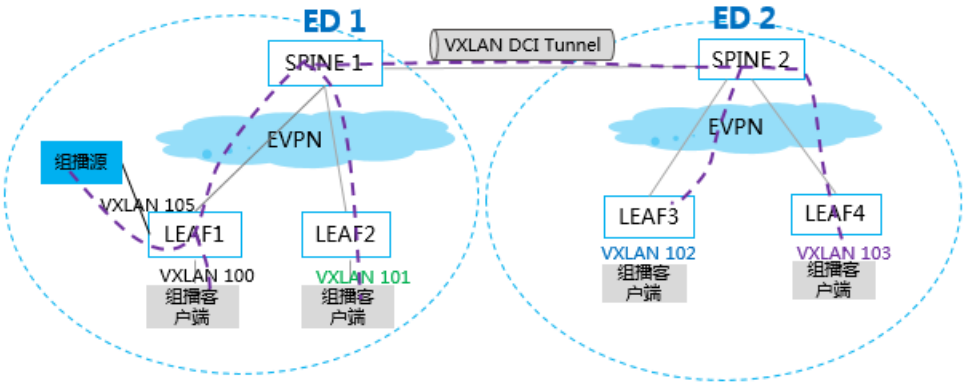
Fabric 内运行的仍是 EVPN 三层组播，再通过 BGP EVPN 路由感知 fabric 外是否存在组播接收者，以此来控制 ED 是否将组播流量经 VXLAN-DCI 隧道转发至其他 Fabric，实现按需转发组播流量。

跨 Fabric 的三层组播互通组网中，ED 之间需要建立 VXLAN-DCI 隧道，组播流量通过该隧道在 Fabric 之间转发。Fabric 内 ED 和 VTEP 之间的 BGP EVPN 路由发布和 MVXLAN 隧道建立过程与同 Fabric 内组网完全相同。

下面介绍 ED 从对端 ED 接收到 SBD-SMET 路由（6 类路由）和 S-PMSI A-D 路由（10 类）后，需要做的处理。

- 6 类路由：ED 根据 6 路由的下一跳地址查找 VXLAN-DCI 隧道，该隧道接口作为组播流量的出接口，以便将 Fabric 内的组播流量通过该隧道转发给对端 ED。
- 10 类路由：ED 将该路由的隧道属性中的组播源地址修改为本地 ED 的 MVXLAN 源接口地址，以便在 Fabric 内建立以该 ED 为组播源的 Default-MDT 或 Data-MDT。

图96 跨园区的 EVPN 三层组播场景



完成 BGP EVPN 路由发布和 VXLAN/MVXLAN 隧道建立后，跨 Fabric 的三层组播流量的转发过程如下：

(1) Leaf 1 接收到组播源发送的组播流量后，识别流量所属的 VPN，并在对应的 VPN 内通过组播隧道将流量转发给本 Fabric 内的 Leaf 2 和 ED 1。

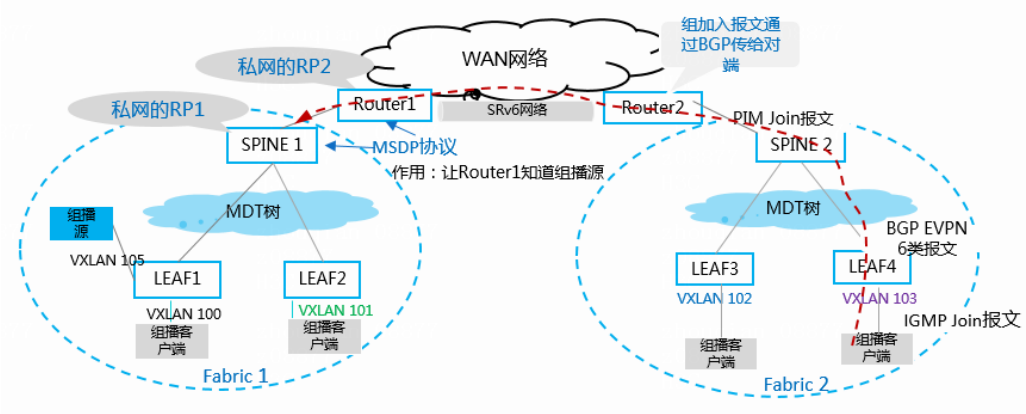
(2) Leaf 2 将组播流量转发给其下的组播接收者；ED 1 解 MVXLAN 封装后，再进行 VXLAN 封装，将组播流量经 VXLAN-DCI 隧道转发给 ED 2。

(3) ED 2 解 VXLAN 封装，再进行 MVXLAN 封装，将组播流量转发给 Leaf 3 和 4，Leaf3 和 4 再将组播流量转发给组播接收者。

16.4 跨域组播方案（Fabric之间跨WAN，走IP转发）（仅用于演示测试，需要wan团队一起配合测试，不推荐实际开局部署）

该场景下，组播源和接收者属于同一 VPN，可以分布在不同的 Fabirc，Fabric 间走的不是 DCI 隧道，而是普通 IP 转发。

图97 跨域组播

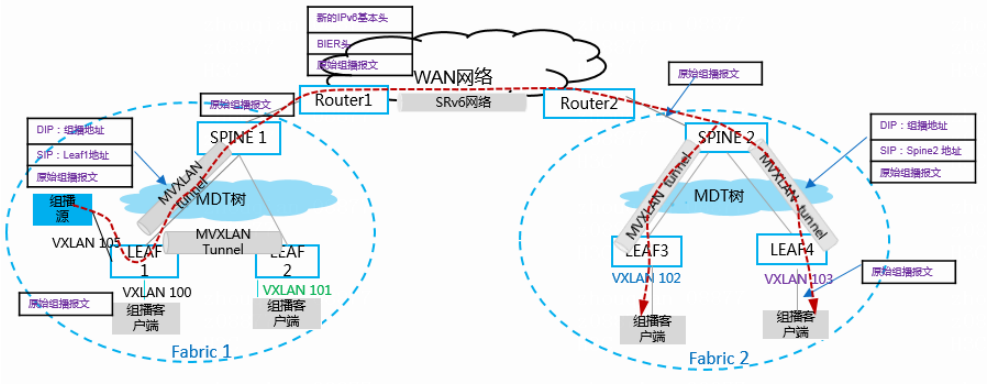


以上图为例介绍跨域组播转发原理，组播源在 Fabric1，接收者分布在 2 个 Fabric。

- 在 Fabric 内，仍运行 EVPN 三层组播。Fabric2 的组播接收者发送 IGMP 加入报文，该 Fabric 内的 LEAF4 会通过 BGP 的 6 类路由，通告给 Spine2。
- 由于私网的 RP 是在对端的 Router1 上，故 Spine2 收到后，向 Router2 发 PIM 加入报文。
- Router1 和 Router2 之间运行 BIER 协议。Router2 会通过 BGP 报文将该加入报文传给对端的 Router（即：Router1）。
- Router 1 和 Spine1 运行 MSDP，通告组播源信息。
- 这样就形成完成的组播转发路径。

组播表项建立后，组播报文只需按照表项转发即可。下图是组播转发过程中，报文封装示意图。

图98 跨域组播转发



在 Fabric1 内，私网的组播报文在 Leaf1 上进行 MVXLAN 封装，Spine1 将该报文转发到 Leaf2，同时解 MVXLAN 封装，转发给 Router1。

Router1 和 2 之间对组播报文进行封装，具体封装格式见上图。

Router2 收到报文，解封装后，发给 Spine2 。

Spine2 对报文进行 MVXLAN 封装后，发给相应的 Leaf，此后的过程同单 Fabric 的组播转发过程。

17 支持 Access 下发 voice VLAN 功能

园区网的 Access 设备会有下接 IP 电话的场景，且会存在人员变更或工位搬迁，IP 电话接入位置会发生变化。为了简化管理员的工作量，AD-Campus 方案支持向 Access 设备下发 voice VLAN ID，以便 IP 电话可以通过该 VLAN 接入网络。

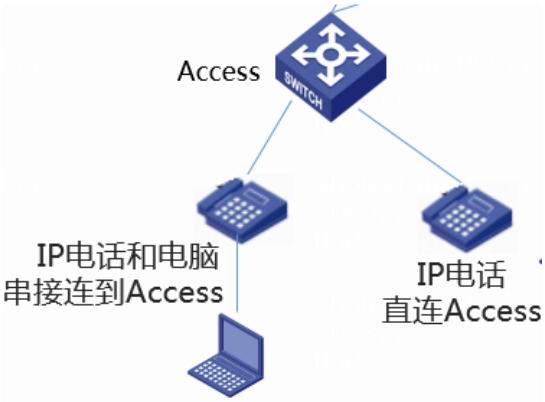
控制组件把 Access 设备的端口配置成 hybrid 口，每端口分 2 个不同的 VLAN，区分数据流和语音流，保证终端即插即用。

该功能的使用场景及要求：

(1) Access 下接 IP 电话，IP 电话的接入方式

- IP 电话单独接入 Access 设备
- 电脑和 IP 电话串接连到 Access 设备

图99 接入方式



(2) IP 电话和电脑都要认证

- 认证点是 Leaf 设备；
- 建议 IP 电话用 MAC 认证，电脑用 1x/MAC 认证/MAC Portal 认证；
- 为了方便管理，建议将 IP 电话规划到单独的 VXLAN。

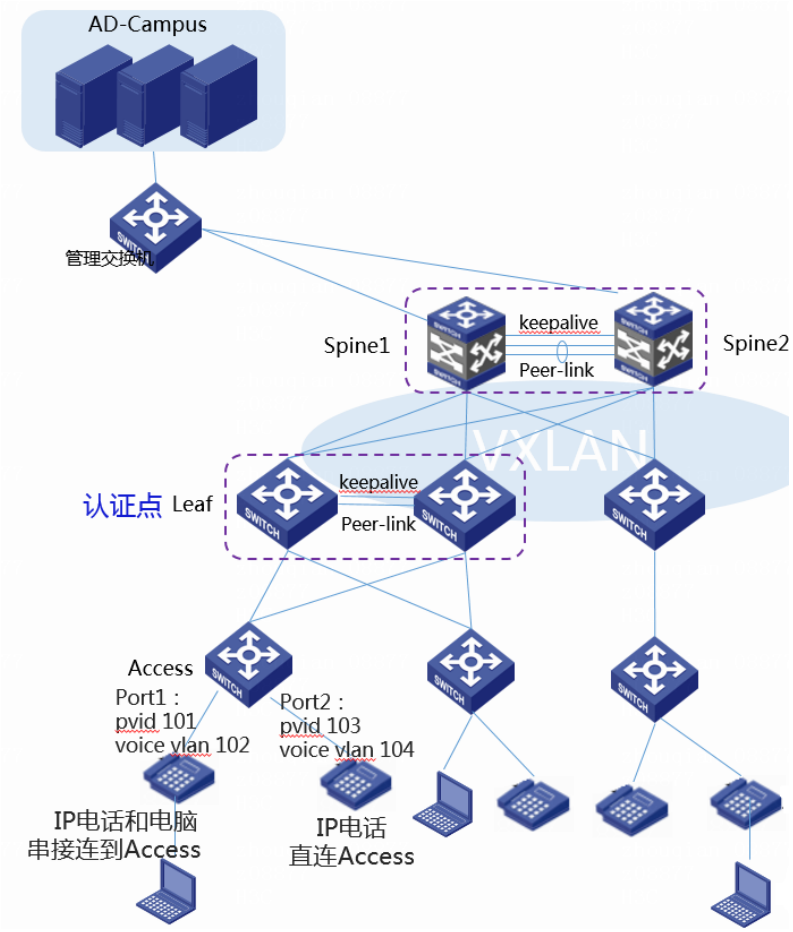
(3) IP 电话要通过 LLDP 协议跟 Access 设备协商出 Voice VLAN

(4) 控制器上要开启端到端保障

该功能开启后，access 和 leaf，leaf 和 spine 的互连端口会下发 qos trust dscp 。

下面介绍 voice VLAN 功能及 IP 电话工作过程：

图100 工作过程



(1) 开启 Voice VLAN 功能，Access 自动化上线或手动纳管且激活后，控制器将 access 端口配置成 hybrid 端口，每端口分两个 VLAN，分别给 IP 电话和电脑使用，端口配置参考如下：

port hybrid vlan 101 untagged //数据 VLAN

port hybrid pvid vlan 101

voice-vlan 102 enable //语音 VLAN

(2) IP 电话连到 Access 的 Port 1，该端口 UP，IP 电话发出的首个报文触发 MAC 认证（该报文不带 tag）；该报文到达认证点 Leaf 设备后，Leaf 会向 AAA 服务器发起认证，认证通过后，AAA 服务器授权 VXLAN（IP 电话所在的 VXLAN）；IP 电话向 DHCP 服务器申请 IP 地址，获取的 IP 地址是 IP 电话所在 VXLAN 网段。

(3) Access 通过 LLDP 报文与 IP 电话协商出 Voice VLAN，本例中是 VLAN102；此后，IP 电话发出的报文都带 VLAN 102；

(4) Access 收到 IP 电话发出的报文，发现报文是从语音 VLAN 102 进来的，会将报文的 DSCP 字段修改成 46（默认值），将语音报文优先处理，从而保证语音通信的质量。

(5) Leaf 收到 IP 电话发出的报文，会将认证表项的初始 VLAN 改成 102。

(6) Leaf 和 Access，Spine 和 Leaf 之间的互连端口都会配置 qos trust dscp，故语音报文到 Leaf 和 Spine 设备，Leaf 和 Spine 会信任语音报文的 DSCP 值，以保障语音报文优先处理。