



Lecture 09: 网络虚拟化

SSE316: 云计算技术
Cloud Computing Technologies

陈壮彬

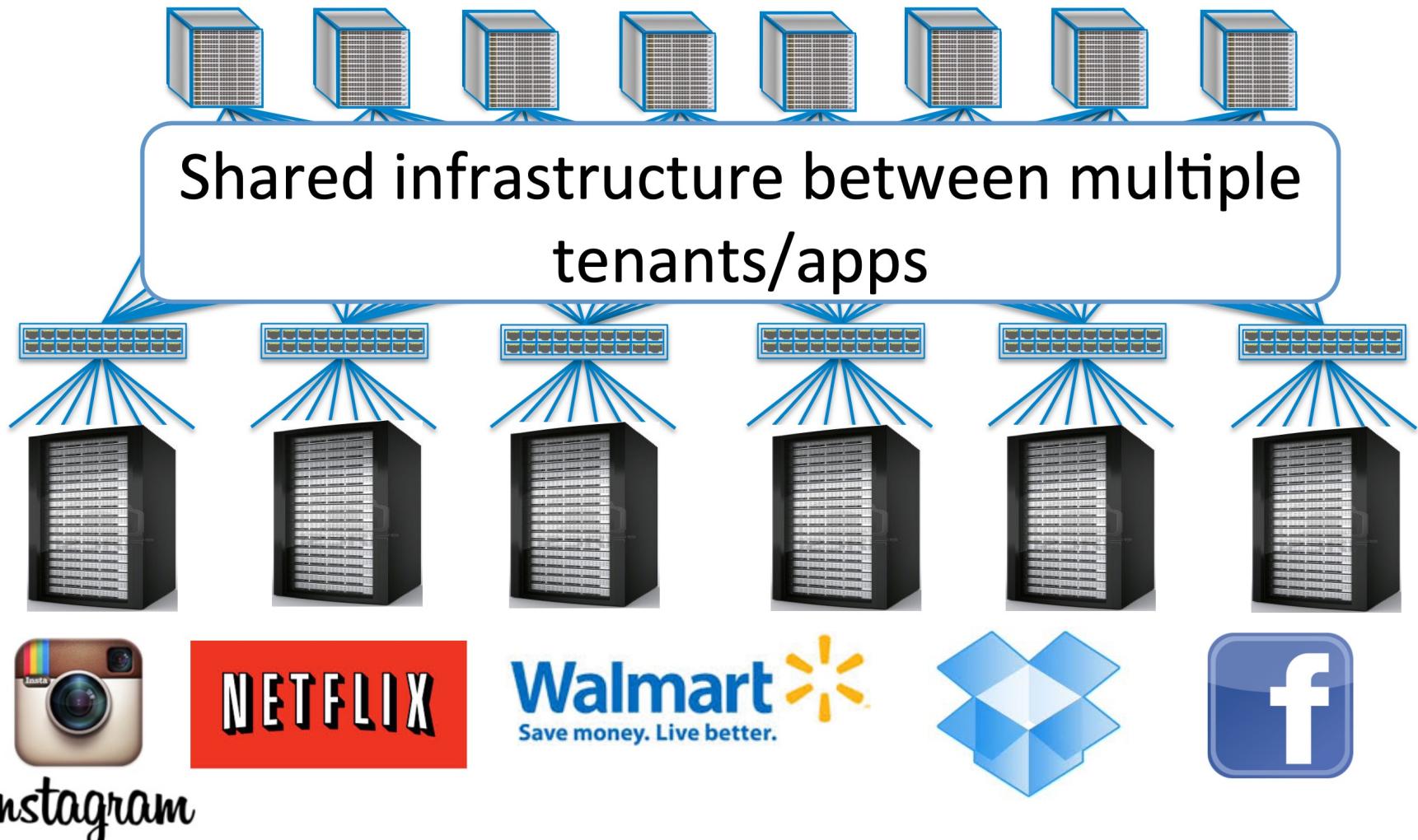
软件工程学院

chenzhb36@mail.sysu.edu.cn

Today's topics

- 网络虚拟化背景
- 网络中的节点通信
- 虚拟局域网 VLAN
- 虚拟可扩展局域网 VxLAN

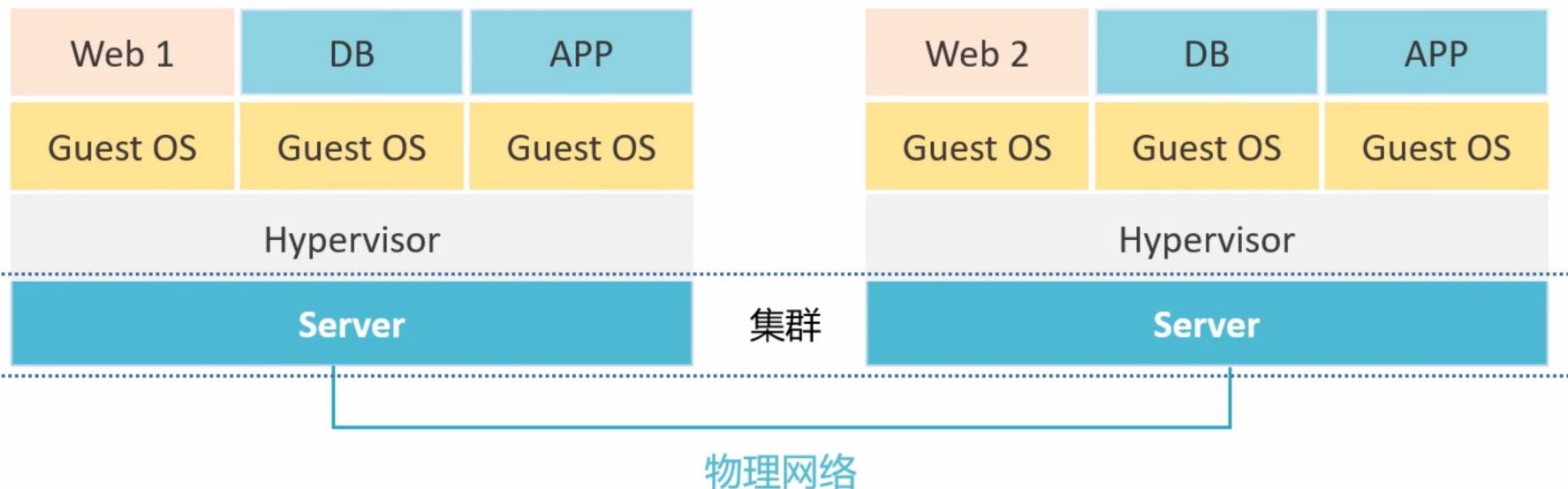
云数据中心多租户



计算虚拟化

- 不同用户/程序独立使用计算资源，如内存、IO、存储
- 降低 IT 成本，提高业务部署灵活度，降低运维复杂度

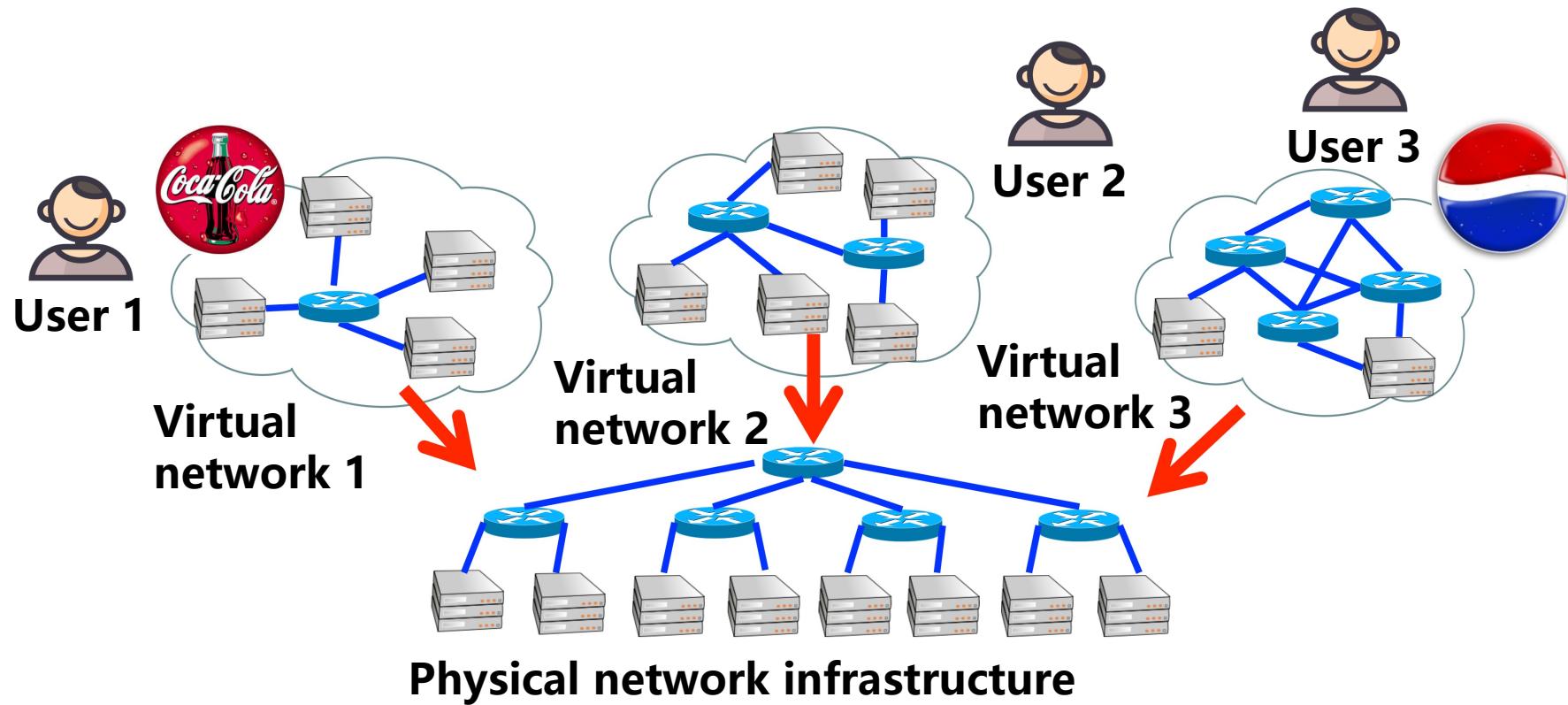
业务以虚拟机形式部署于服务器集群



网络虚拟化

口具有运行多个虚拟网络的能力，这些网络

- 每个网络具有独立的控制平面和数据平面
- 能够在一个物理网络的基础上共存
- 可以由可能互不信任的各方单独管理



网络虚拟化定义

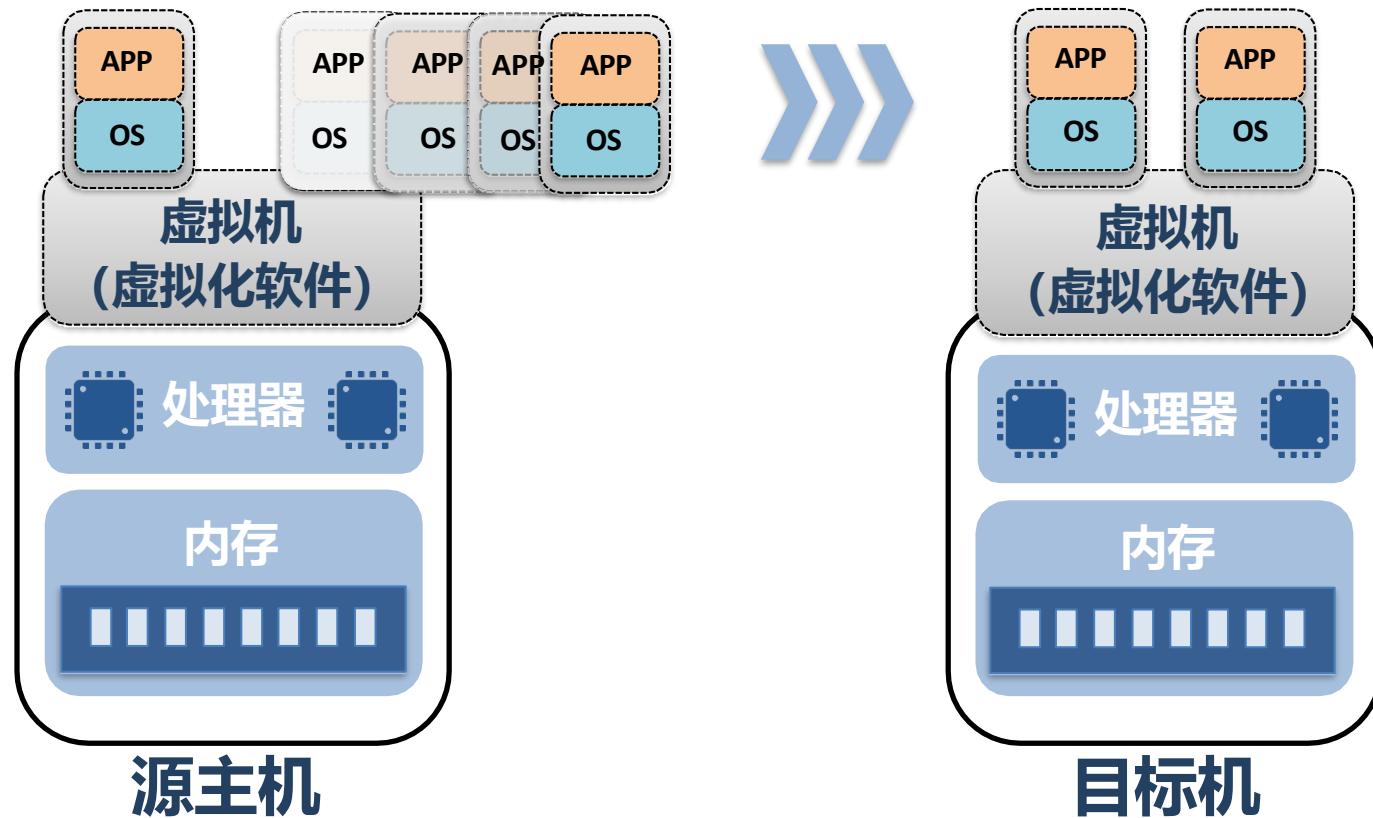
网络虚拟化是在单个共享的物理网络之上创建多个逻辑隔离的虚拟网络的过程

实现了网络资源的抽象，允许多个租户或应用共享相同的物理网络资源，且不会相互干扰

网络虚拟化简化了管理，提高了灵活性和资源利用率

虚拟机迁移

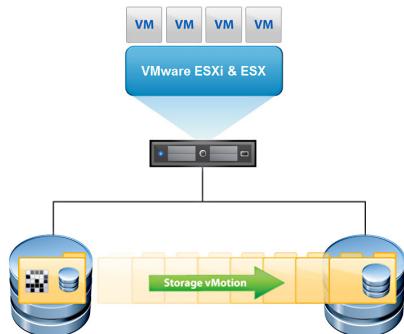
虚拟机灵活迁移是云提供高质量服务的基础



需要迁移的资源



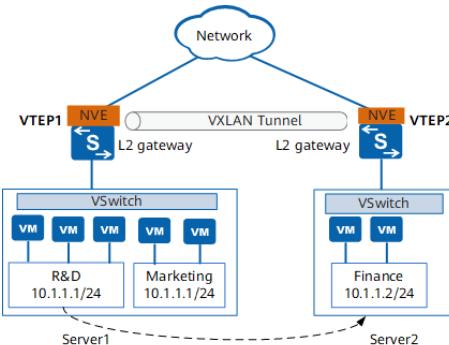
Storage



共享数据和
文件系统



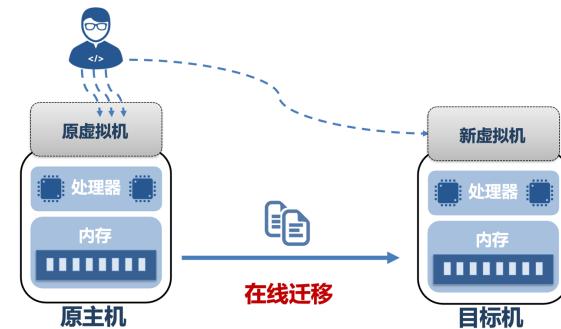
Network



虚拟网络技术
如VxLAN



Memory



预拷贝和
后拷贝

**网络中，不同节点是如何
传输数据的？**

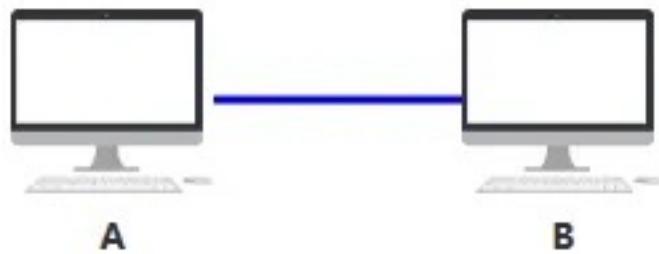
网络连接

很久之前，有一台孤苦伶仃的电脑 A，不与任何人连接



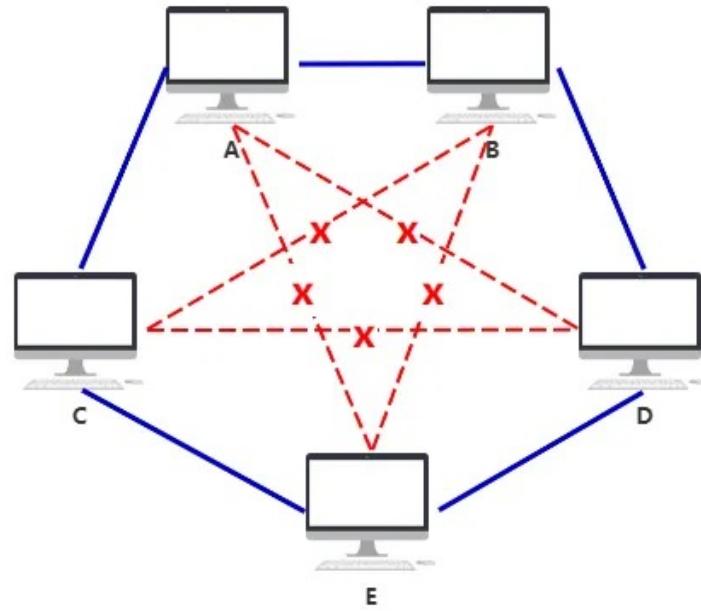
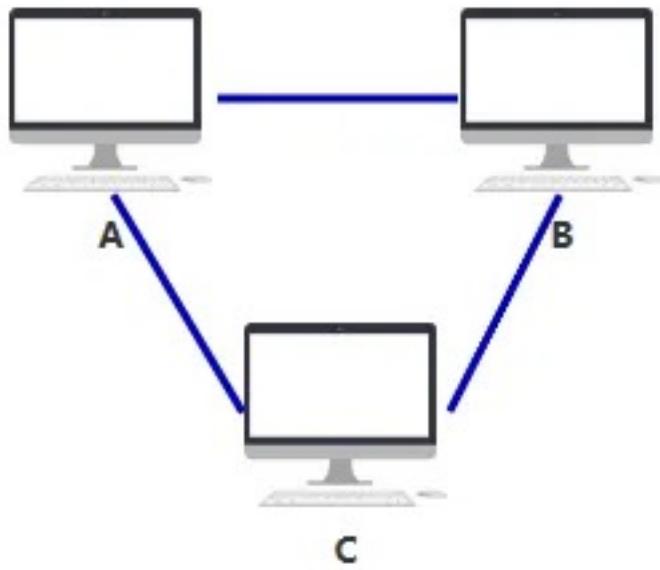
有一天，A 找到了朋友 B，希望与对方建立通信

于是它们各自开了一个网口，用一根网线相连，愉快地打游戏



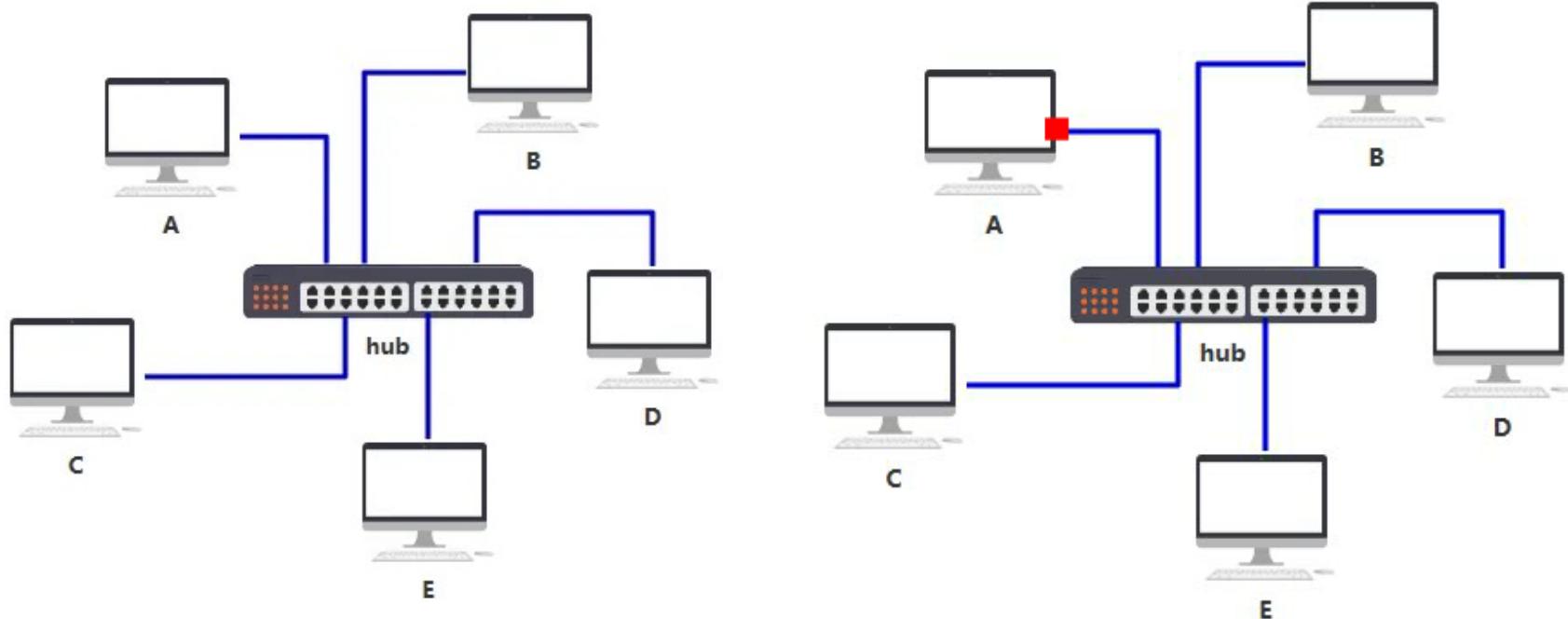
更多的主机相连

- 新朋友 C 也想加入进来！
- 于是，三个人各自开了两个网口，两两相连
- 但是，更多的朋友能一直加入进来吗？



集线器 Hub

- 一个中间设备，把所有网线插到这个设备，由它做转发
- 能实现彼此通信，且网口的数量和网线的数量减少了
- 该设备叫**集线器**，无差别地将电信号转发到所有出口（广播）

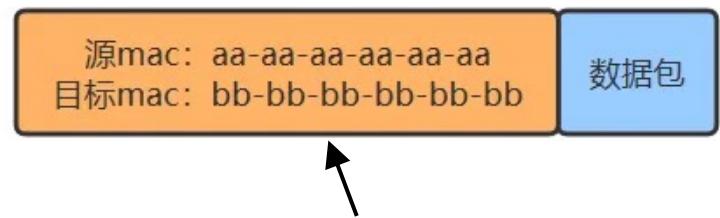


MAC 地址

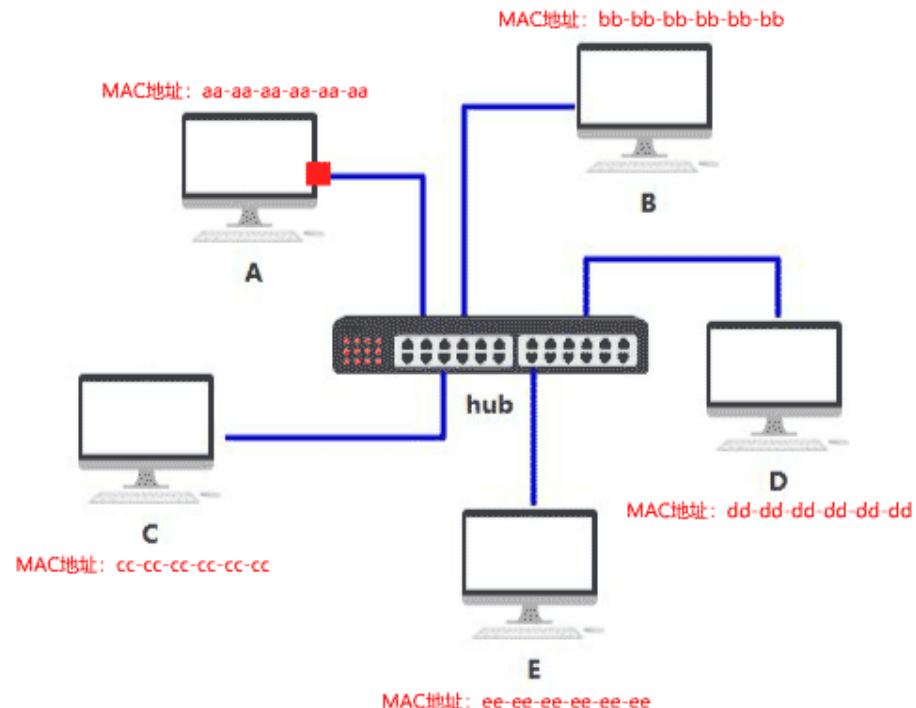
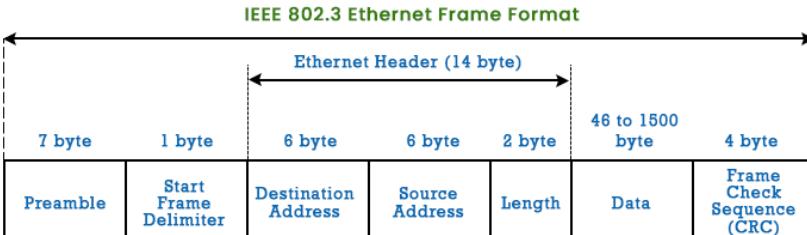
□设备怎么知道数据包是发给自己的呢?

□MAC地址 (媒体访问控制地址, Media Access Control Address)

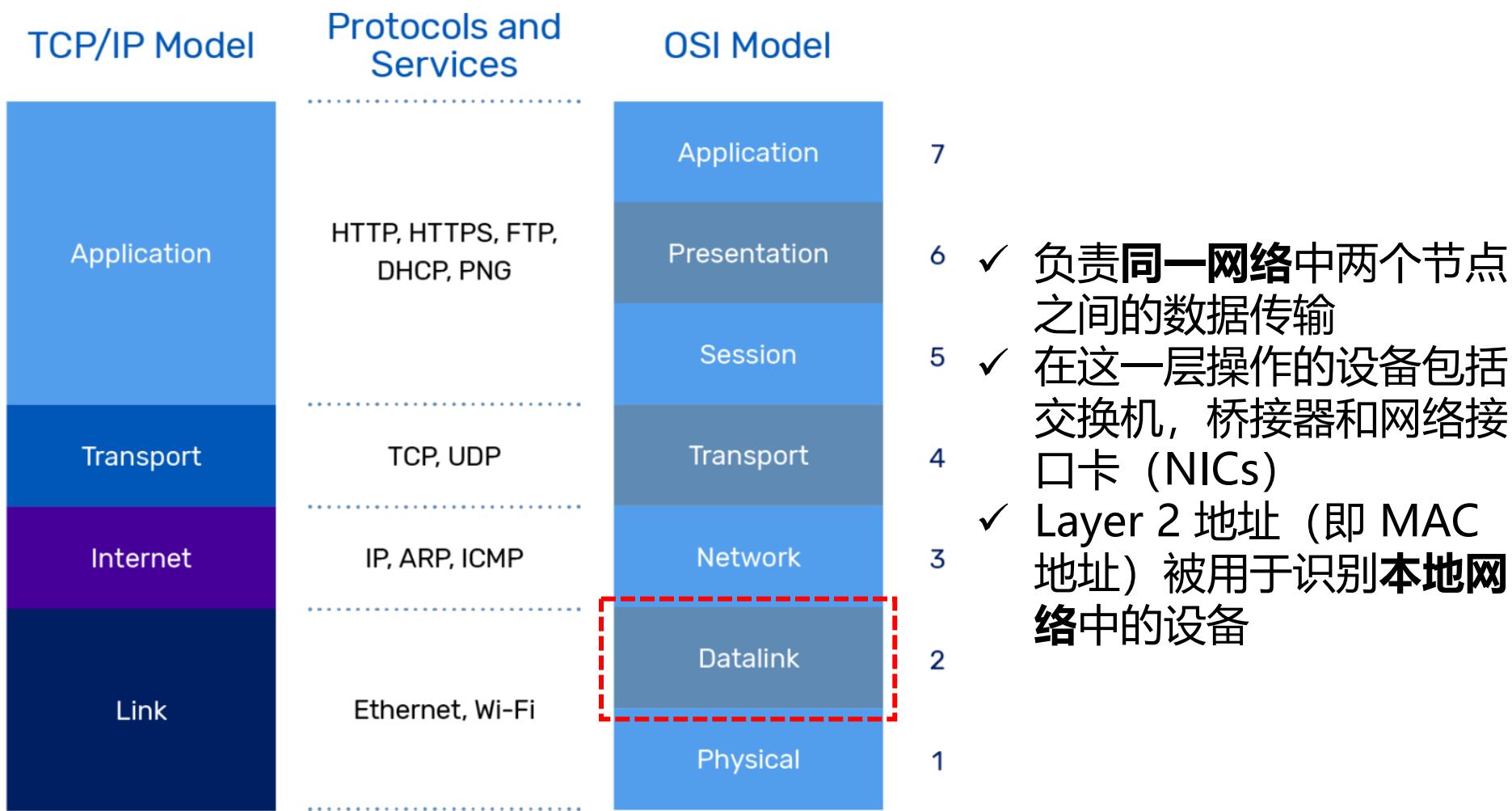
- 48位的二进制数, 常以十六进制表示, 如 “00:0A:95:B1:2C:EB”
- 每个网络接口卡 (NIC) 都有一个固定且唯一的 MAC 地址
- 设备局域网通信的基础, 工作在数据链路层 (二层网络 Layer 2)
- 如以太网和 Wi-Fi



把数据打包成一个数据帧 **frame**,
并在头部加上 MAC 地址

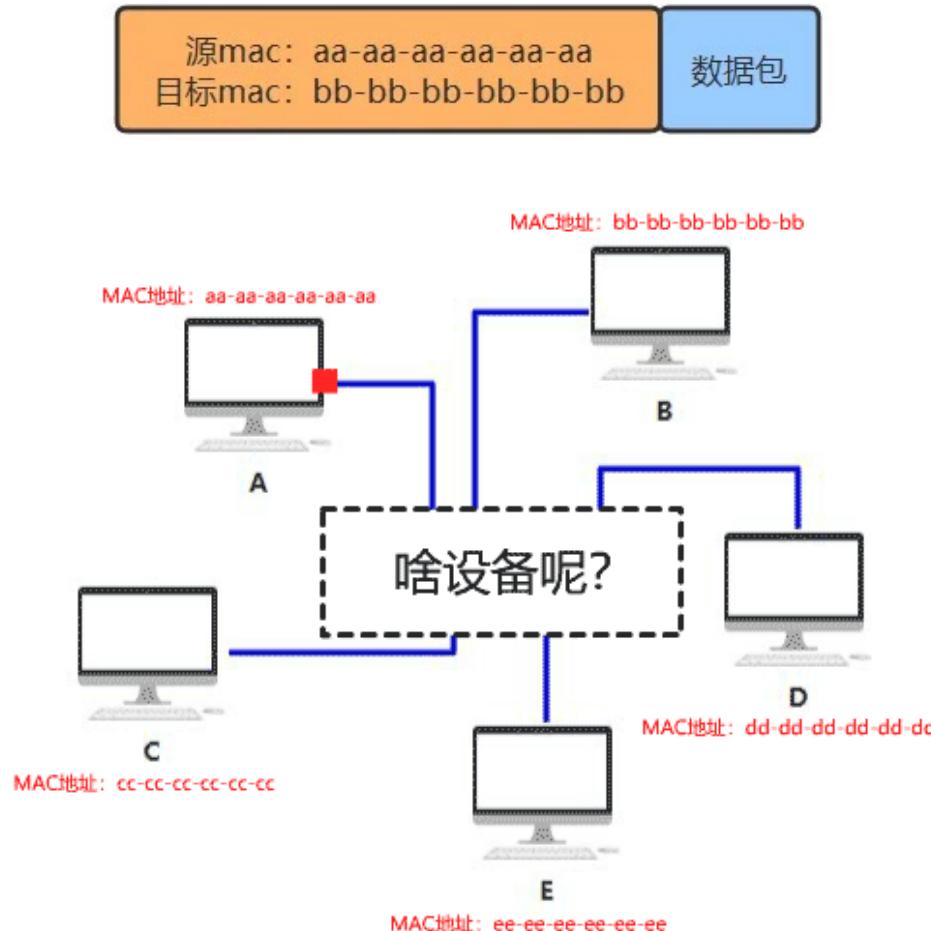


TCP/IP 模型 – 数据链路层（二层）



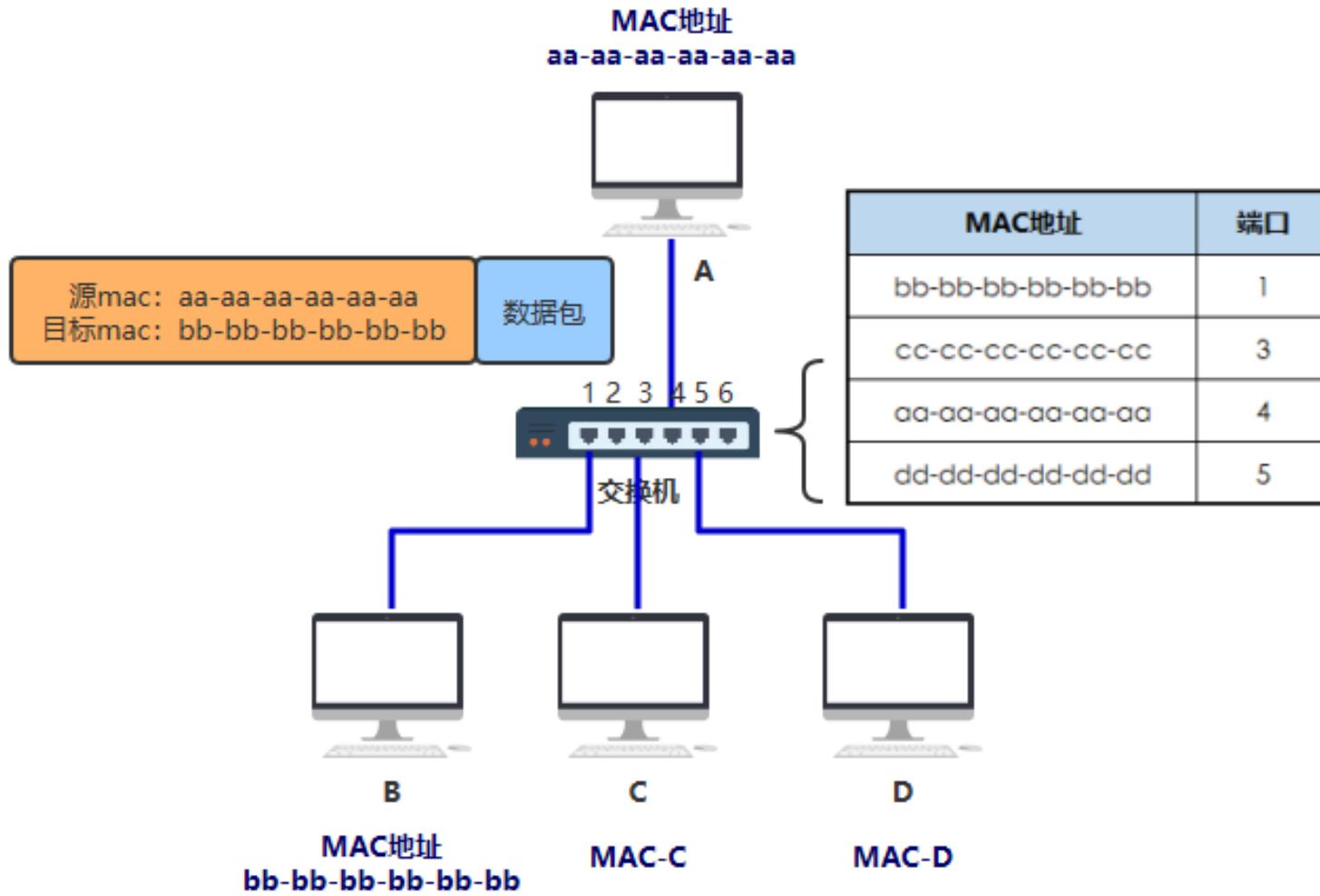
更智能的集线器？

□能否依据目的 MAC 地址精准发送数据帧？



交换机 Switch

口能依据 MAC 地址路由表，精准转发数据帧



MAC 地址表

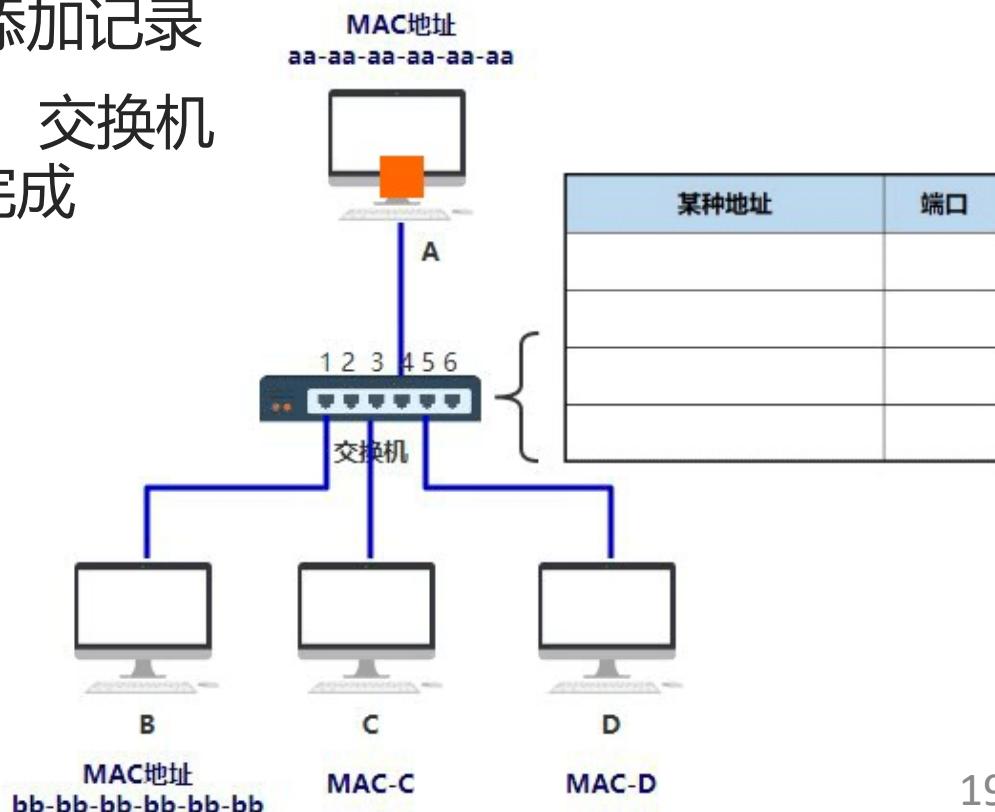
- 交换机内部维护一张 MAC 地址表
- 记录着每一个 MAC 地址的设备，连接在其哪一个端口上
- 以这种传输方式组成的小范围网络，叫做**以太网 (Ethernet)**

MAC 地址	端口
bb-bb-bb-bb-bb-bb	1
cc-cc-cc-cc-cc-cc	3
aa-aa-aa-aa-aa-aa	4
dd-dd-dd-dd-dd-dd	5

MAC 地址表

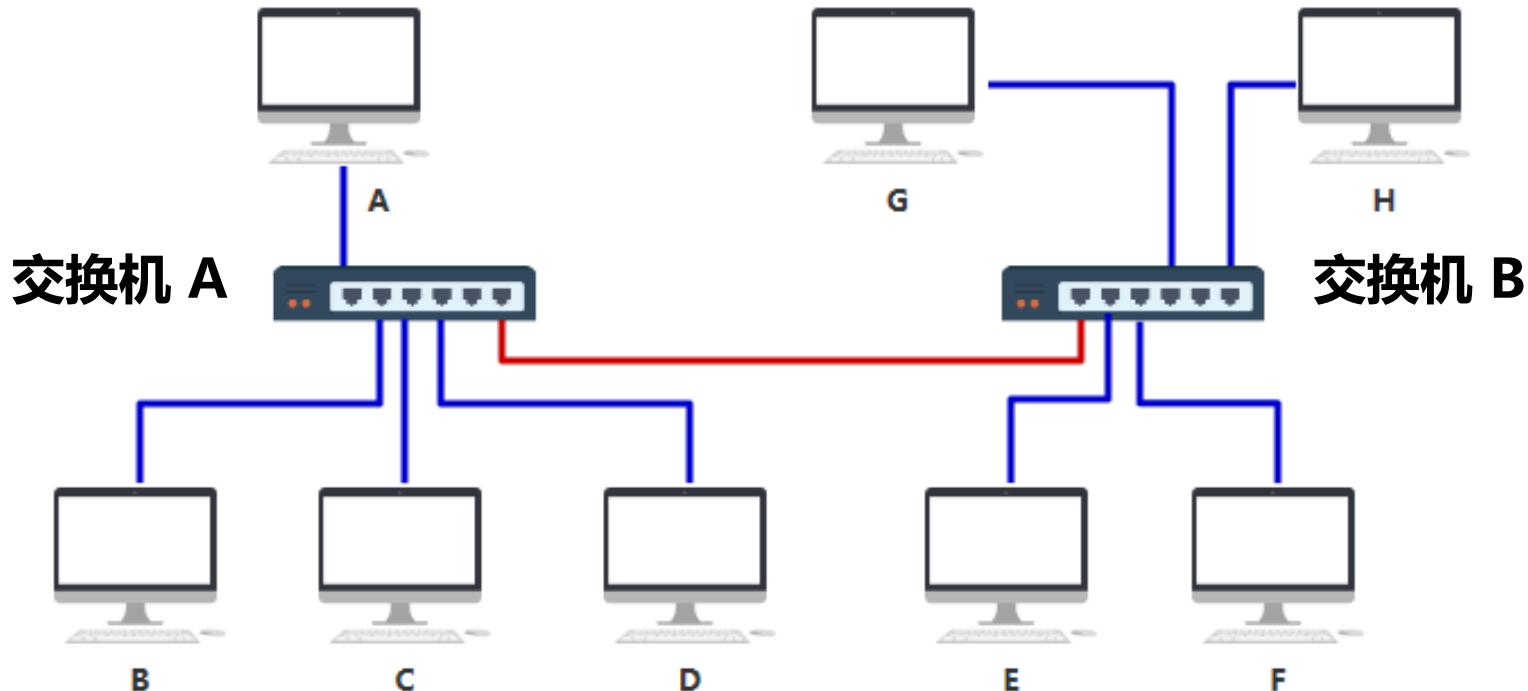
口最开始时，MAC 地址表为空，该如何建立？

- A 向 B 发送数据
- A 从端口 4 进入交换机，添加相应记录
- 交换机查表无记录，将数据帧发送给所有设备
- B 做出响应，端口为 1，添加记录
- 经过网络中设备不断通信，交换机最终将 MAC 地址表建立完成



多个交换机相连

- 随着机器数量增多，交换机的端口不够用了
- 最简单的方法 – 将多个交换机相连！



MAC 地址表怎么写？

冗余 MAC 地址项

交换机 A 的 MAC 地址表

MAC 地址	端口
bb-bb-bb-bb-bb-bb	1
cc-cc-cc-cc-cc-cc	3
aa-aa-aa-aa-aa-aa	4
dd-dd-dd-dd-dd-dd	5
ee-ee-ee-ee-ee-ee	6
ff-ff-ff-ff-ff-ff	6
gg-gg-gg-gg-gg-gg	6
hh-hh-hh-hh-hh-hh	6

交换机 B 的 MAC 地址表

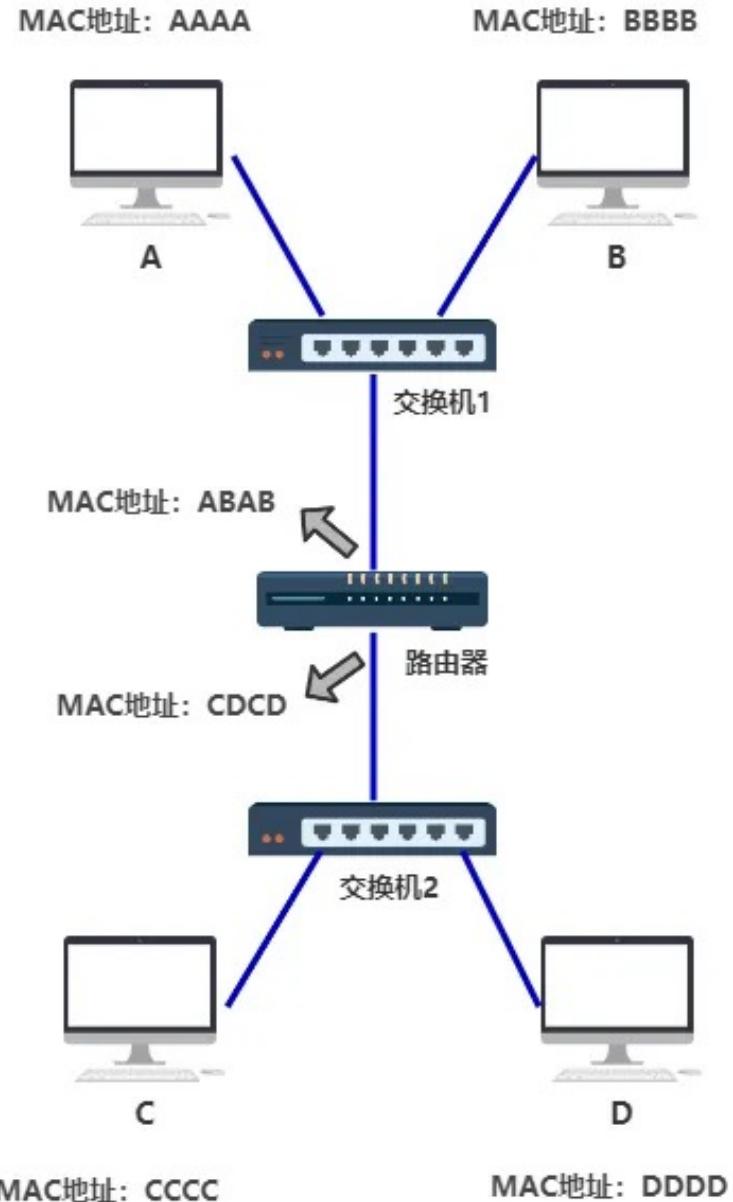
MAC 地址	端口
bb-bb-bb-bb-bb-bb	1
cc-cc-cc-cc-cc-cc	1
aa-aa-aa-aa-aa-aa	1
dd-dd-dd-dd-dd-dd	1
ee-ee-ee-ee-ee-ee	2
ff-ff-ff-ff-ff-ff	3
gg-gg-gg-gg-gg-gg	4
hh-hh-hh-hh-hh-hh	6

- 存在大量冗余
- 现代分布式系统包含几十万甚至上百万台物理 / 虚拟服务器
- 交换机无法记录如此庞大的映射关系

路由器 Router

□ 在红色的线中接入一个新的设备，
拥有独立的MAC地址，并把数据
帧做一次转发

□ 这个设备便是**路由器**，每一个端口
有独立的MAC地址



IP 地址

路由器如何知道，哪个集合的机器应该从某个端口统一转发出去呢？

IP 地址 (Internet Protocol Address)

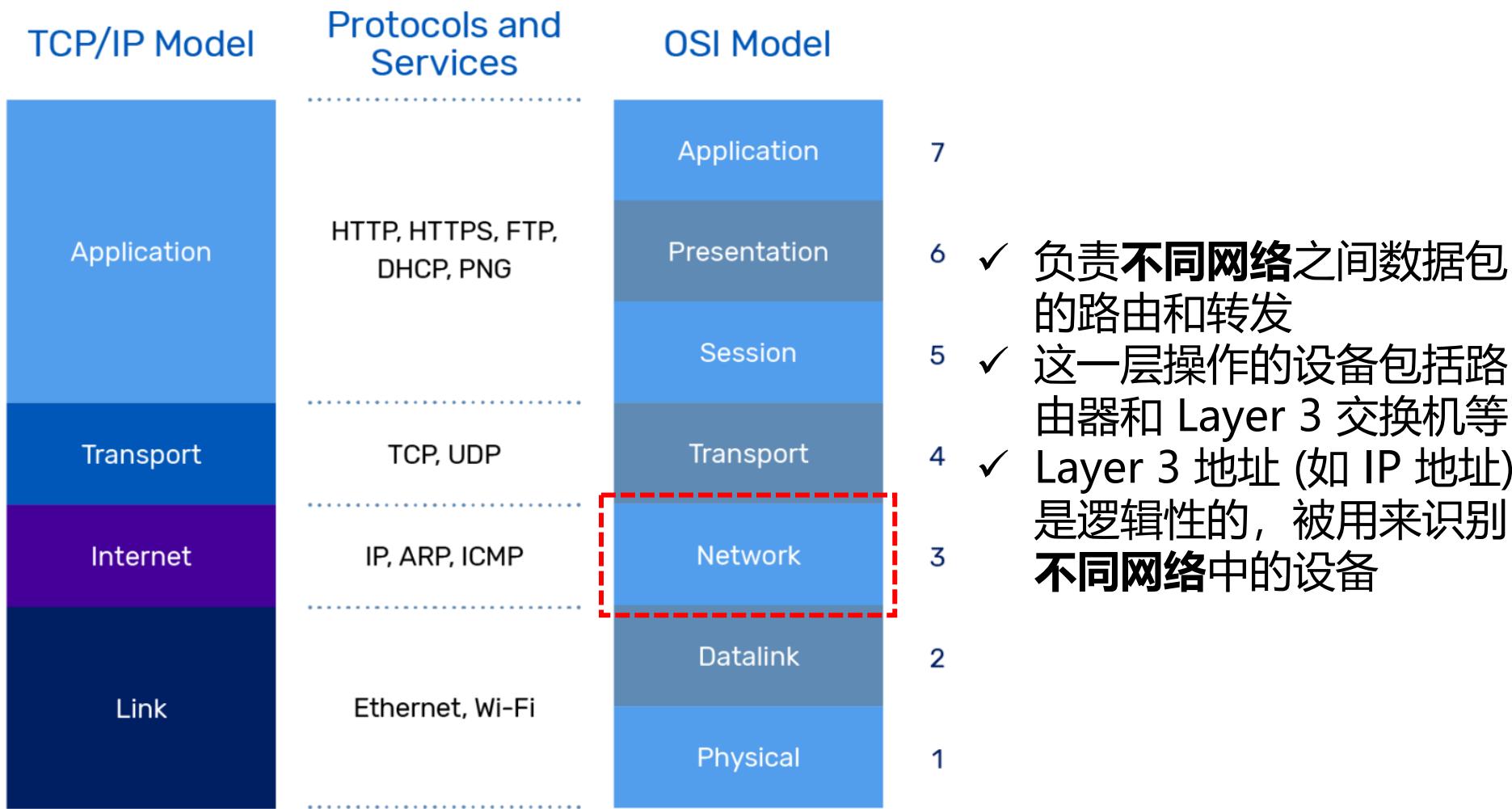
- 互联网协议中用于标识网络中每个设备的标识符，并确定路由
- 包括 IPv4 和 IPv6 (解决 IPv4 地址空间有限的问题)
- 可静态分配 (手动设置)，也可以动态分配，如 DHCP 协议
- **设备互联网通信的基础，工作在网络层 (三层网络，Layer 3)**

每个机器一个32位的编号 (IPv4)，如
11000000101010000000000000000001

分成四个部分，用点相连
11000000.10101000.00000000.00000001

换算成十进制
192.168.0.1

TCP/IP 模型 – 网络层（三层）



IP 地址

- 现在每个设备不仅拥有一个 MAC 地址，还拥有一个 IP 地址
- MAC 地址不可变，IP 地址可变



A

MAC: aa-aa-aa-aa-aa-aa
IP: 192.168.0.1



B

MAC: bb-bb-bb-bb-bb-bb
IP: 192.168.0.2



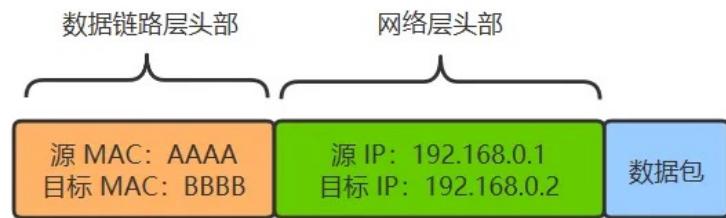
C

MAC: cc-cc-cc-cc-cc-cc
IP: 192.168.0.3



D

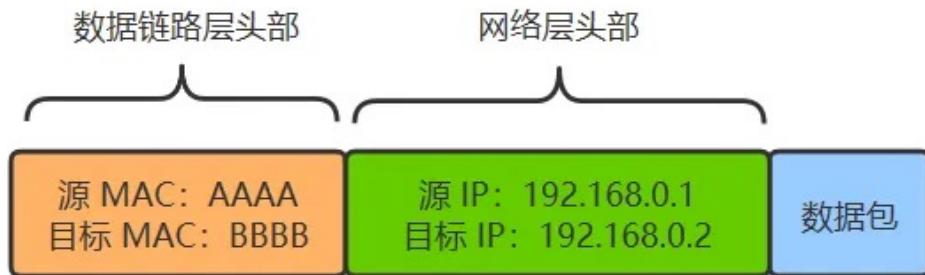
MAC: dd-dd-dd-dd-dd-dd
IP: 192.168.0.4



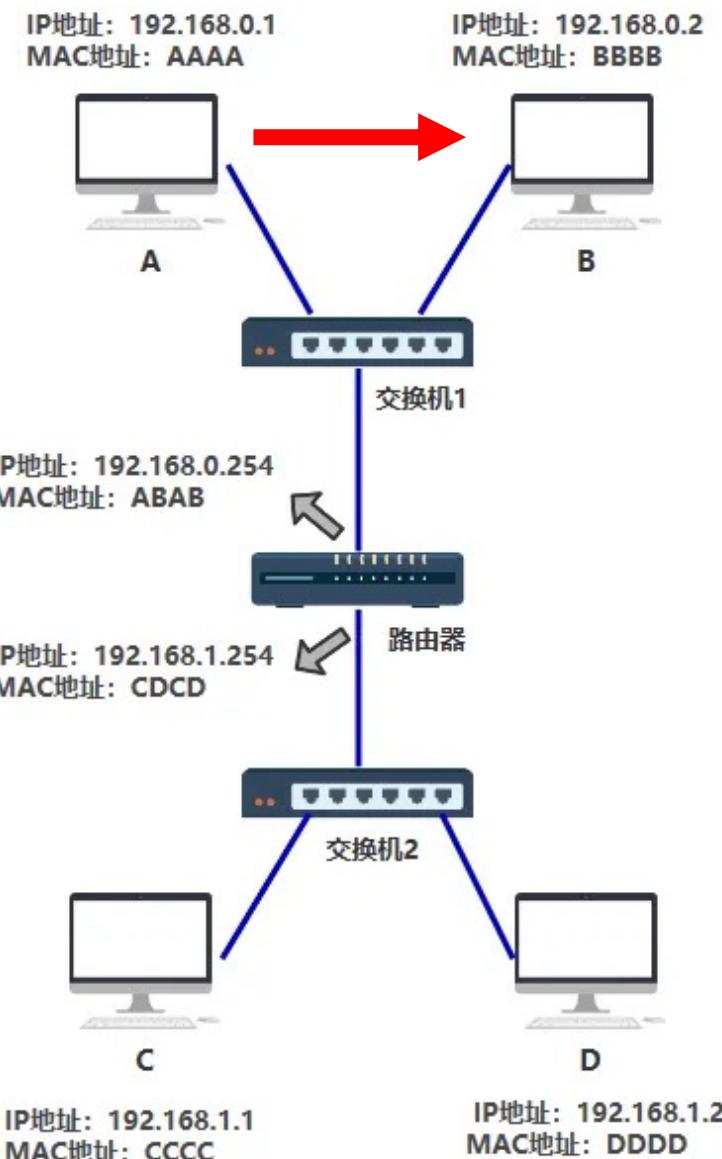
在数据帧的基础上加上 IP 地址，编程数据包 (packet)

基于 IP 地址的数据包转发

口 A 将数据包发给 B



A 和 B 连接着同一个交换机，直接利用 MAC 地址即可实现转发，网络层（IP 地址）的作用未体现



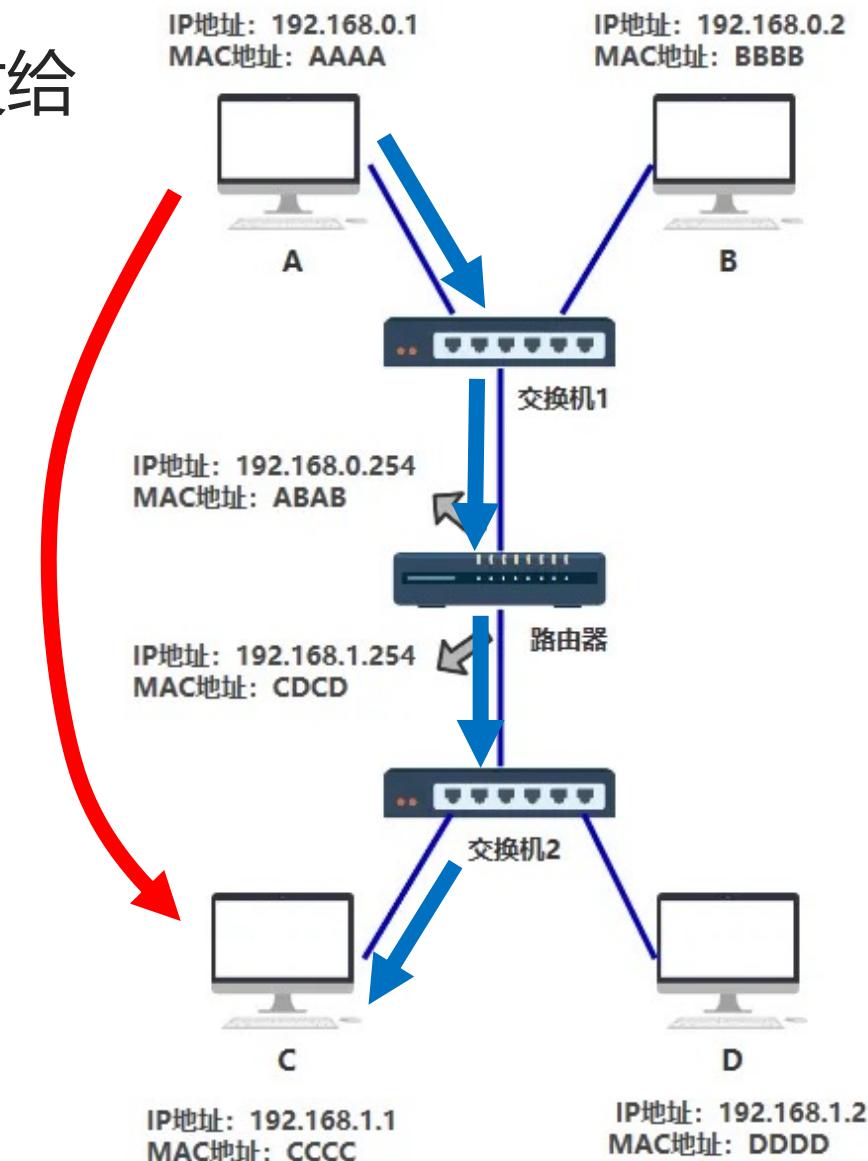
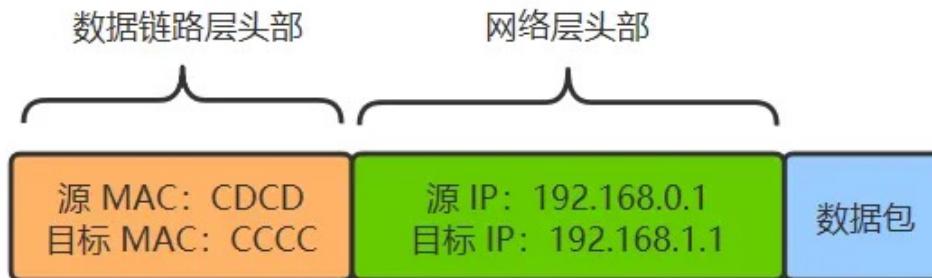
基于 IP 地址的数据包转发

口 A 将数据包发给 C，需要先发给路由器，再由路由器转交给 C

A 到路由器的数据包



路由器到 C 的数据包

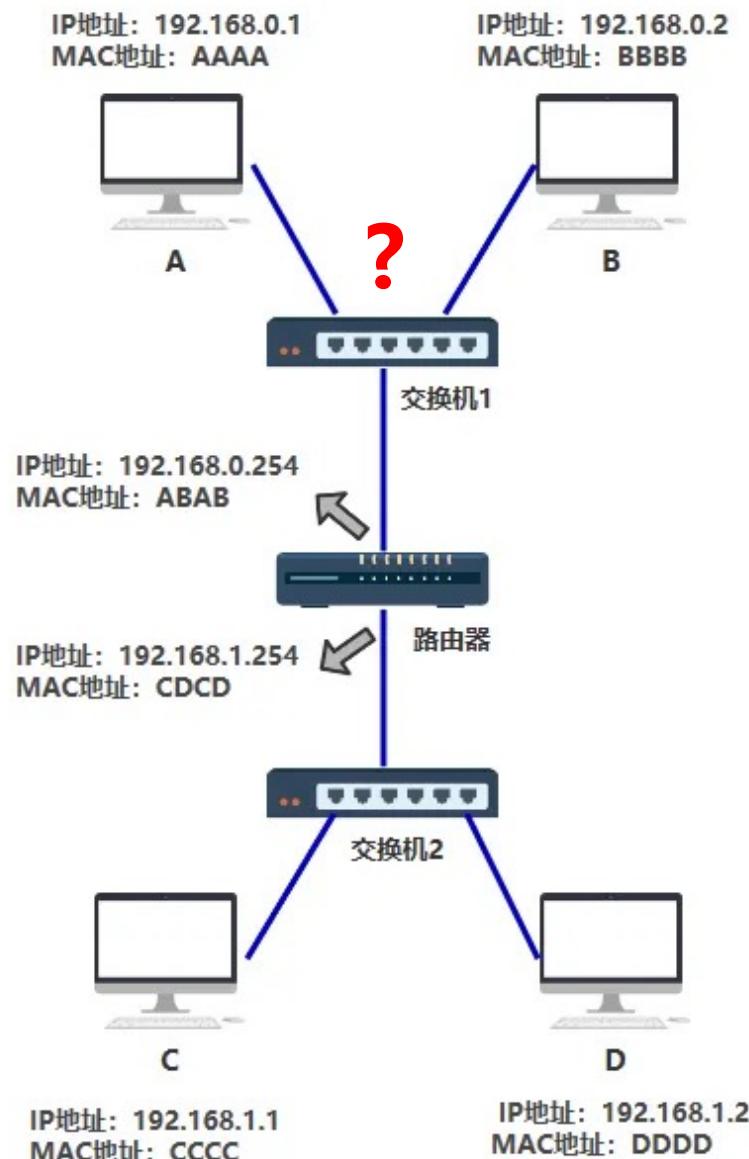


基于 IP 地址的数据包转发

□ 交换机如何知道，哪些数据包应该让路由器转发（如 A -> C），哪些由自己直接转发（如 A -> B）

答案：子网！

□ 是网络设计中的一种策略，将一个大型的 IP 网络划分为多个较小的网络，以提高网络管理效率和安全性

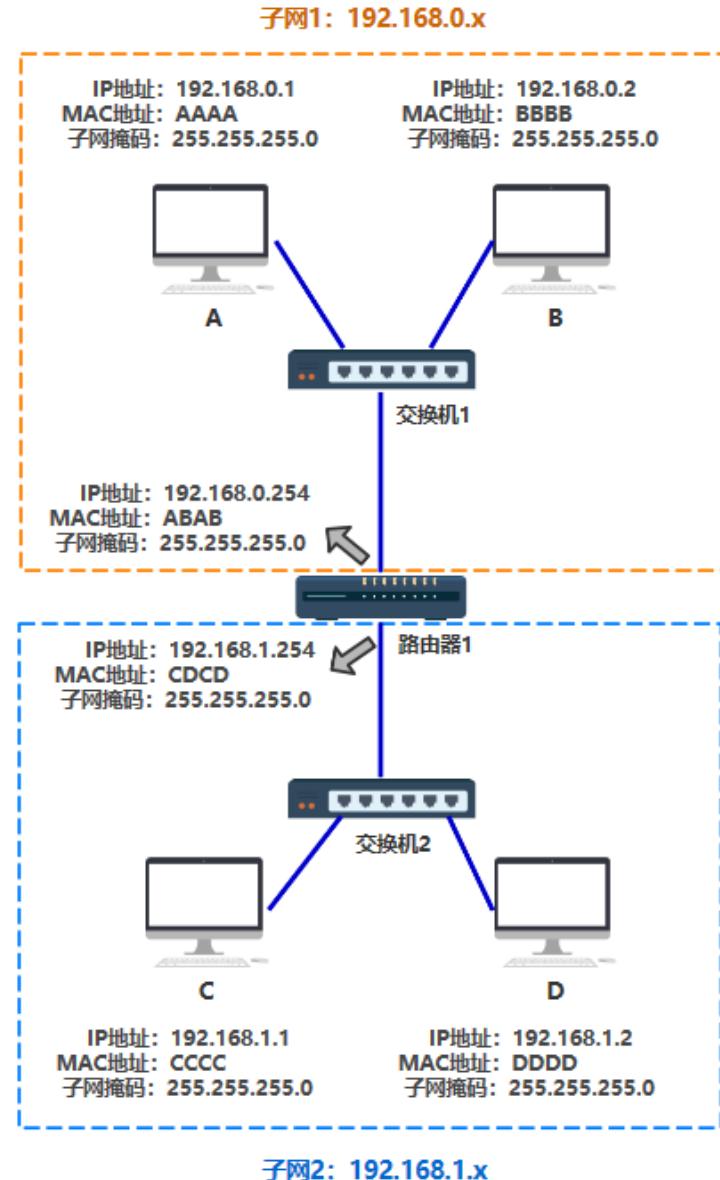


子网 Subnet

比如以 **192.168.0.xxx** 开头的，
就算是在一个子网，由子网掩码
(Subnet Mask) 实现

- A: $192.168.0.1 \& 255.255.255.0 = 192.168.0.0$
- B: $192.168.0.2 \& 255.255.255.0 = 192.168.0.0$
- C: $192.168.1.1 \& 255.255.255.0 = 192.168.1.0$
- D: $192.168.1.2 \& 255.255.255.0 = 192.168.1.0$

A 和 B 在同一个子网
C 和 D 在同一个子网



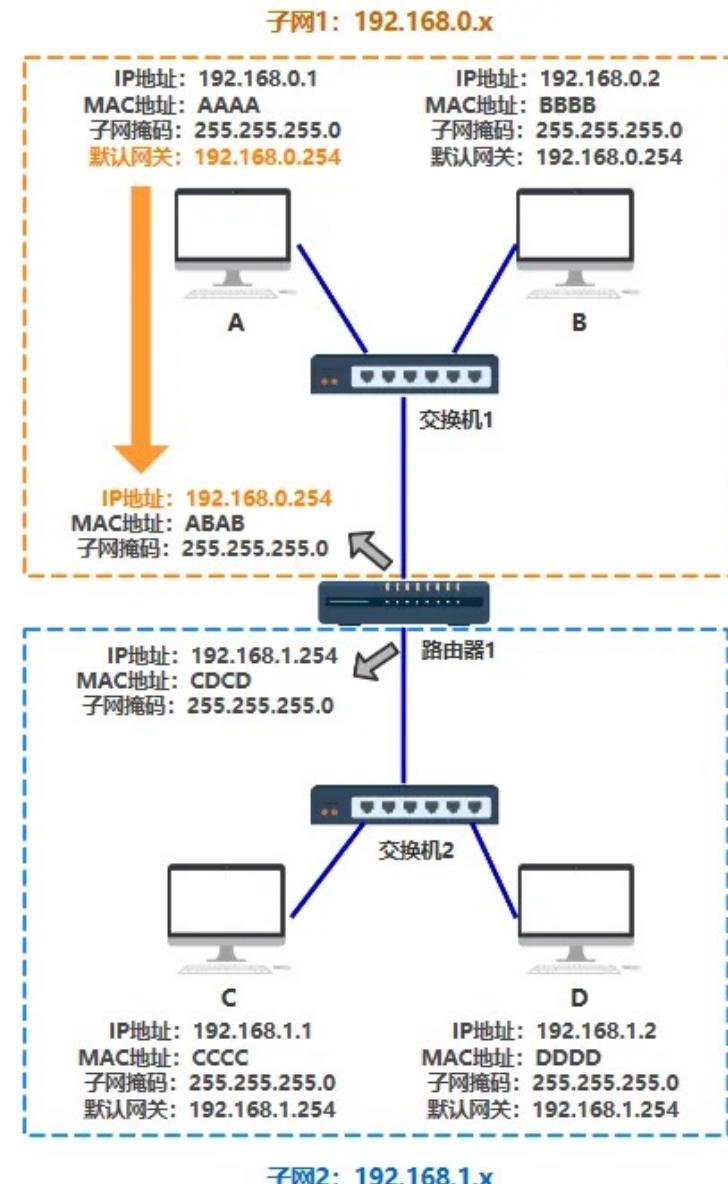
默认网关 Default Gateway

□ A 如何知道，哪个设备是路由器？

答案：默认网关！

□ 默认网关就是 A 在自己电脑里配置的一个 IP 地址，以便在发给不同子网的机器时，发给这个 IP 地址

□ 对 A 来说，A 只能直接把包发给同一个子网下的某个 IP，A 不关心是发给路由器还是其他设备，只要该设备有个 IP 地址即可



路由表 Routing Table

- 现在 A 要给 C 发数据包，已经可以先发给路由器了
- 路由器怎么知道，收到的数据包该从哪个端口发送出去？

答案：路由表！

目的地址	子网掩码	下一跳	端口
192.168.0.0	255.255.255.0		0
192.168.0.254	255.255.255.255		0
192.168.1.0	255.255.255.0		1
192.168.1.254	255.255.255.255		1

目的地址	下一跳	端口
192.168.0.0/24		0
192.168.0.254/32		0
192.168.1.0/24		1
192.168.1.254/32		1

IP地址: 192.168.0.1
MAC地址: AAAA
子网掩码: 255.255.255.0
默认网关: 192.168.0.254

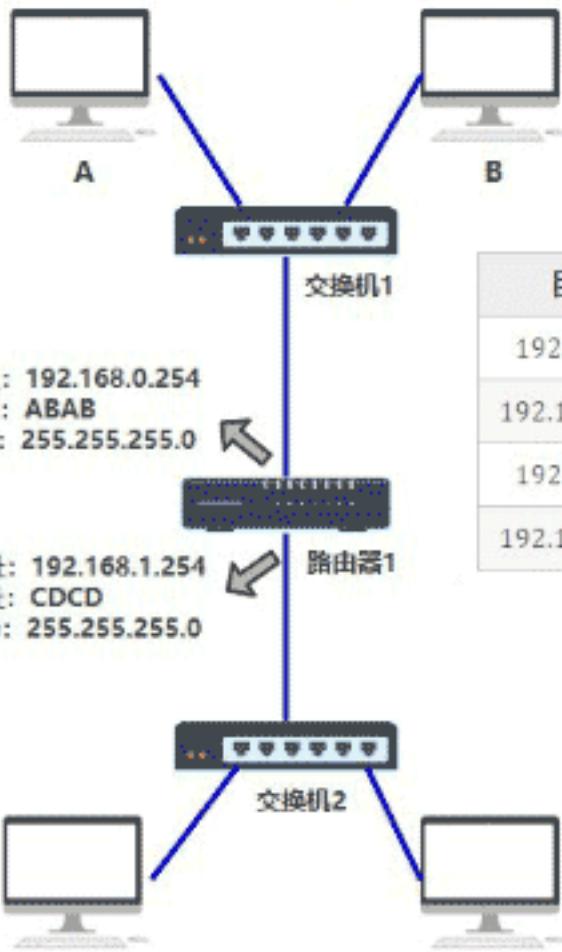
IP地址: 192.168.0.2
MAC地址: BBBB
子网掩码: 255.255.255.0
默认网关: 192.168.0.254

IP地址: 192.168.0.254
MAC地址: ABAB
子网掩码: 255.255.255.0

IP地址: 192.168.1.254
MAC地址: CDCD
子网掩码: 255.255.255.0

IP地址: 192.168.1.1
MAC地址: CCCC
子网掩码: 255.255.255.0
默认网关: 192.168.1.254

IP地址: 192.168.1.2
MAC地址: DDDD
子网掩码: 255.255.255.0
默认网关: 192.168.1.254



目的地址	端口
192.168.0.0/24	0
192.168.0.254/32	0
192.168.1.0/24	1
192.168.1.254/32	1

ARP 协议

口上述转发基于 IP 地址，数据在数据链路层转发需要 MAC 地址

答案： ARP 协议 (地址解析协议，Address Resolution Protocol)!

口IP地址和MAC地址的对应关系，不同设备有一张 ARP 缓存表

IP 地址	MAC 地址
192.168.0.2	BBBB

口一开始这个表是空的，A 为了知道 B 的 IP 地址对应的 MAC 地址，会广播一条 arp 请求，B 收到后，会带上自己的 MAC 地址给 A 相应，以此更新整个 ARP 表

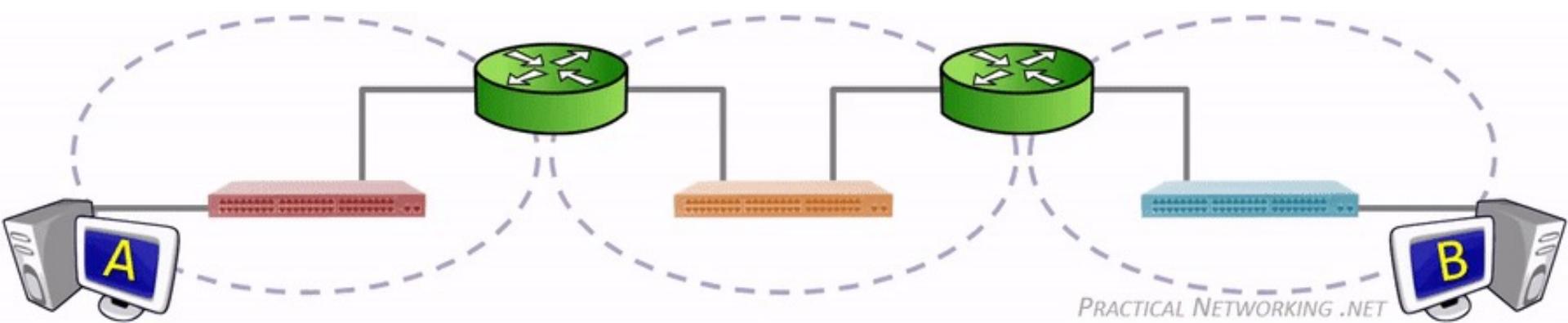
网络通信



路由器 (router), 三层网络, 基于 IP 地址



交换机 (switch), 二层网络, 基于 MAC 地址



局域网 LAN (Local Area Network)

局域网有什么问题?

局域网的问题

口广播风暴

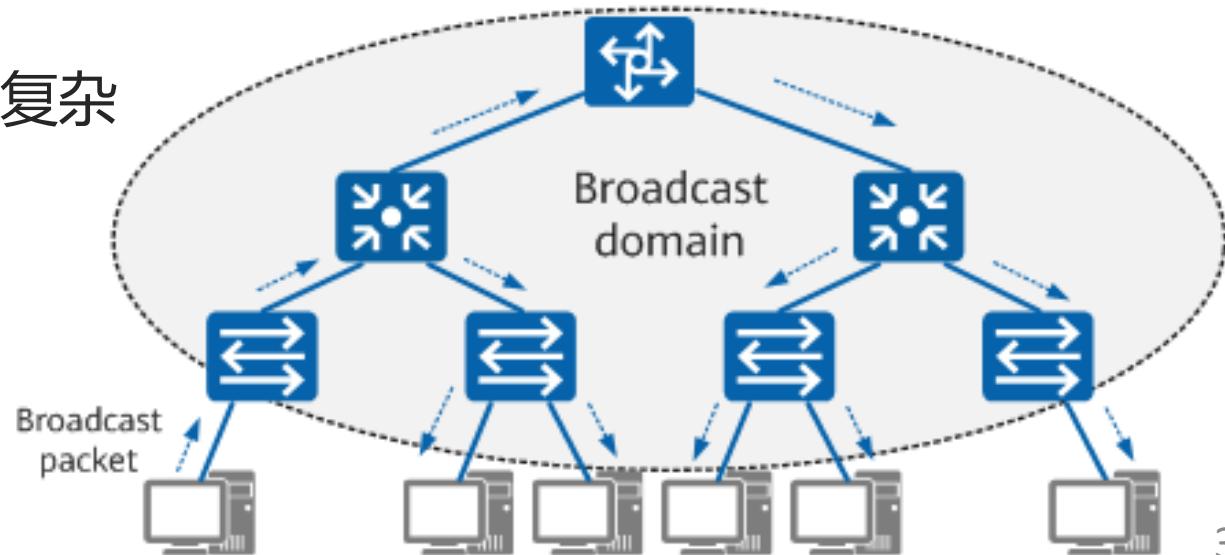
- 一个局域网为一个广播域，所有设备都可以收到广播消息
- 例如广播 ARP 包，进行服务发现等
- 当设备数量增多时，容易引发广播风暴

口安全和隐私

- 不同设备可以尝试访问或监听其他设备，存在安全隐患

口效率和管理问题

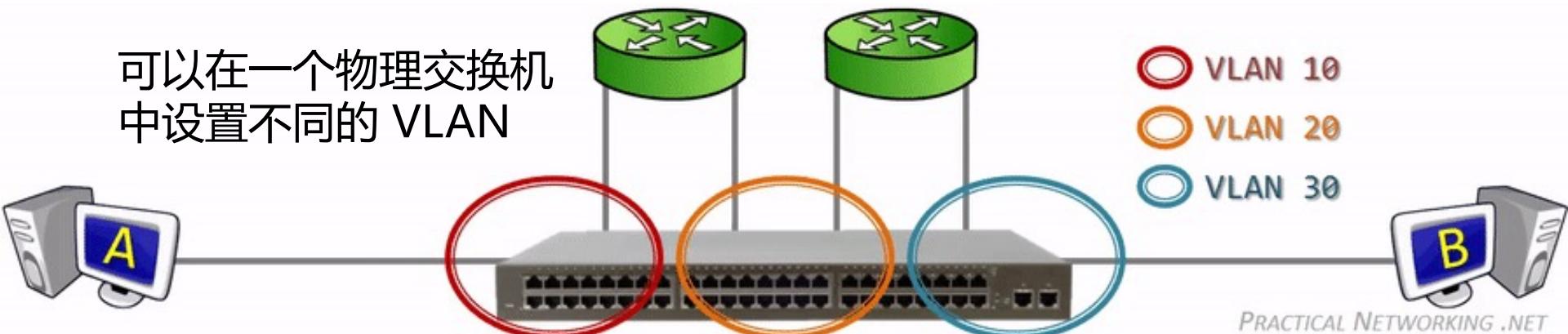
- 占用带宽且管理复杂



虚拟局域网 VLAN

口虚拟局域网 (Virtual Local Area Network)

- 将一个**物理二层网络**拆分为多个**逻辑虚拟网络**，相互独立
- 每个 VLAN 充当一个**单独的广播域**
- 同一 VLAN 中的主机能够直接相互通信
- 跨 VLAN 通信需要借助三层网络

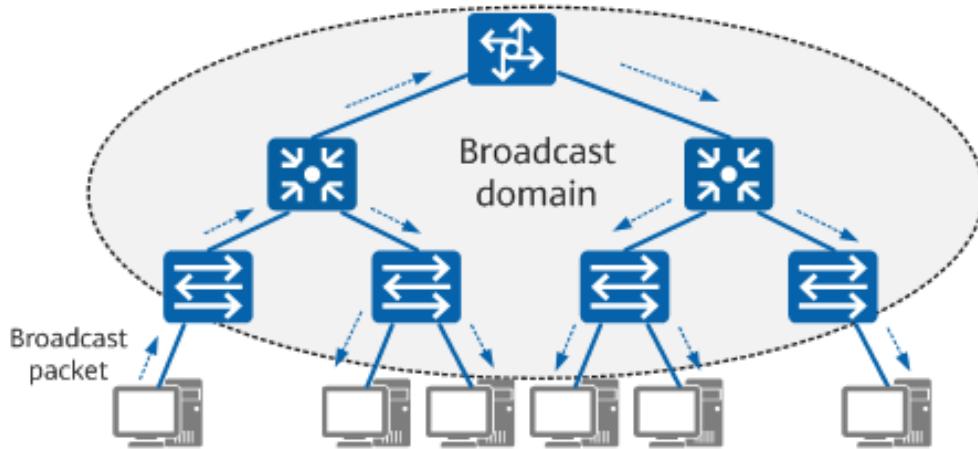


支持 VLAN 的交换机的 Mac 地址路由表格式：

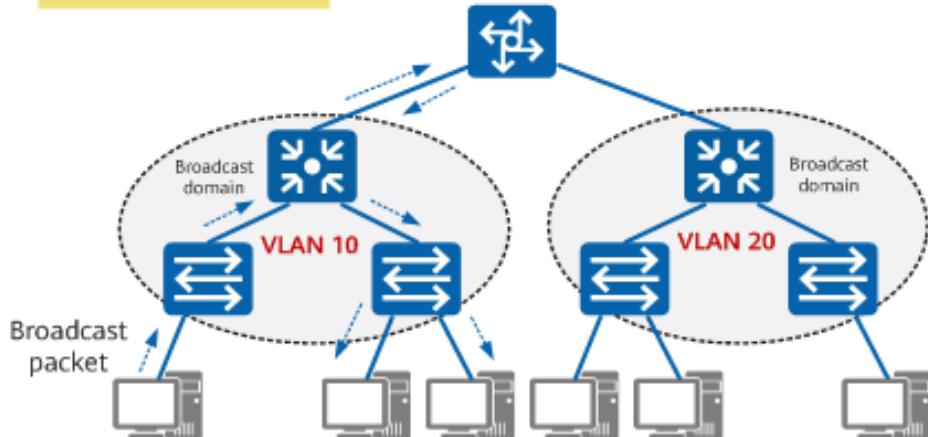
VLAN# | MAC Address | Port

VLAN 拆分广播域

Without VLANs

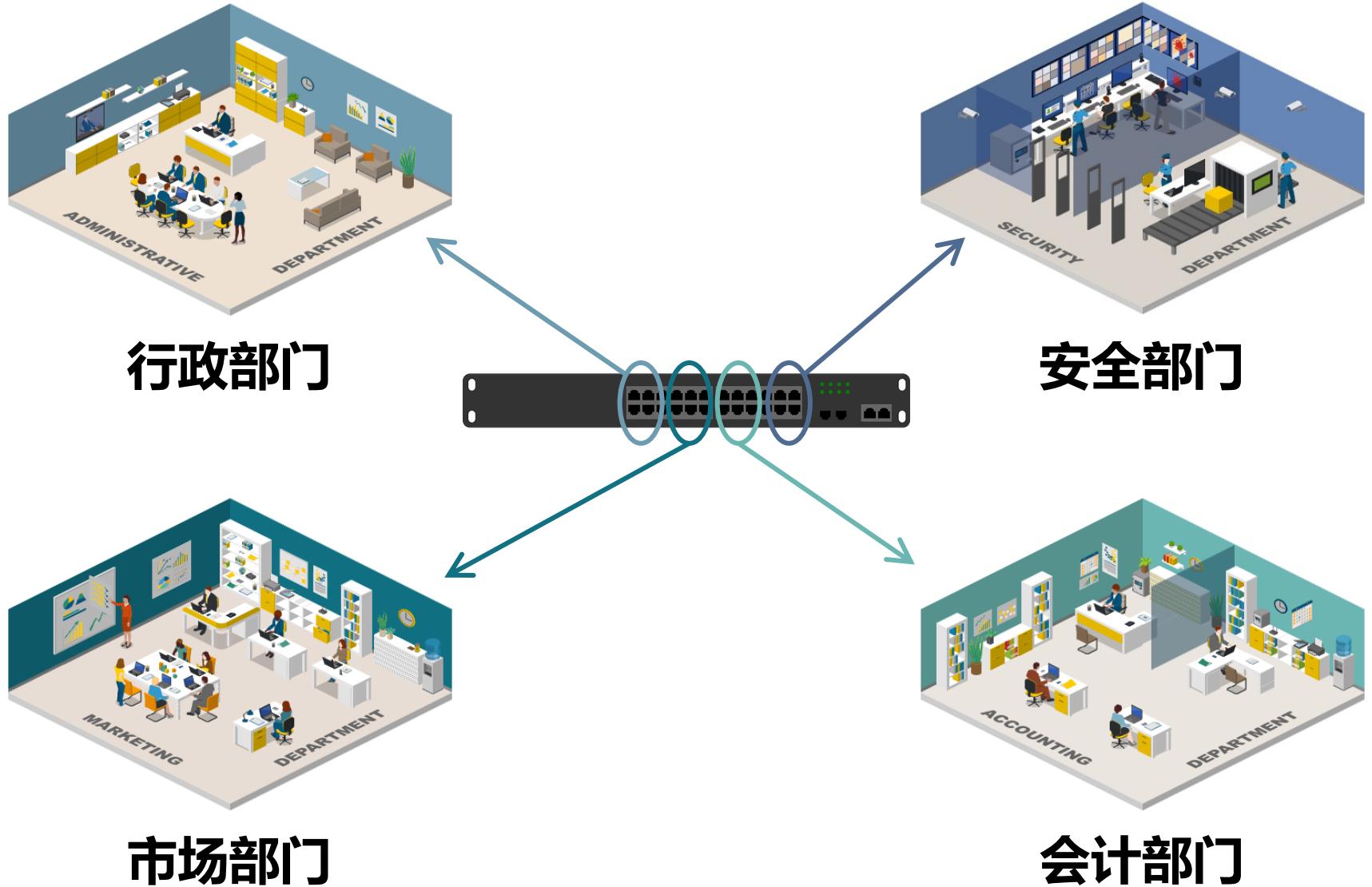


With VLANs

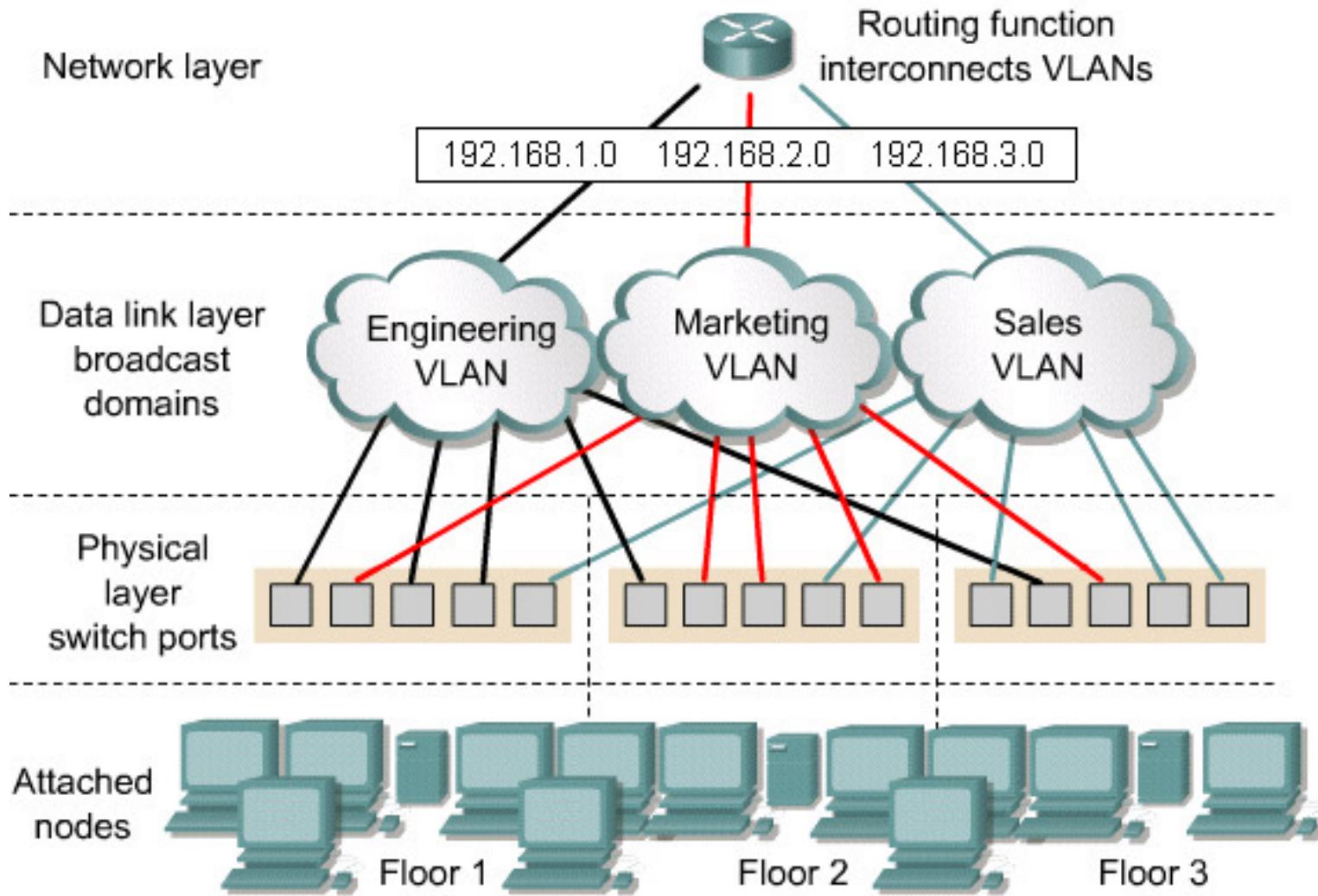


- ✓ 每个广播域被限制在单个 VLAN 中
- ✓ VLAN 之间互不干扰
- ✓ 提升了安全性和隐私性

多部门公司管理



VLAN 功能划分



在网络设计中，VLAN 和 subnet 通常是一一对应的关系

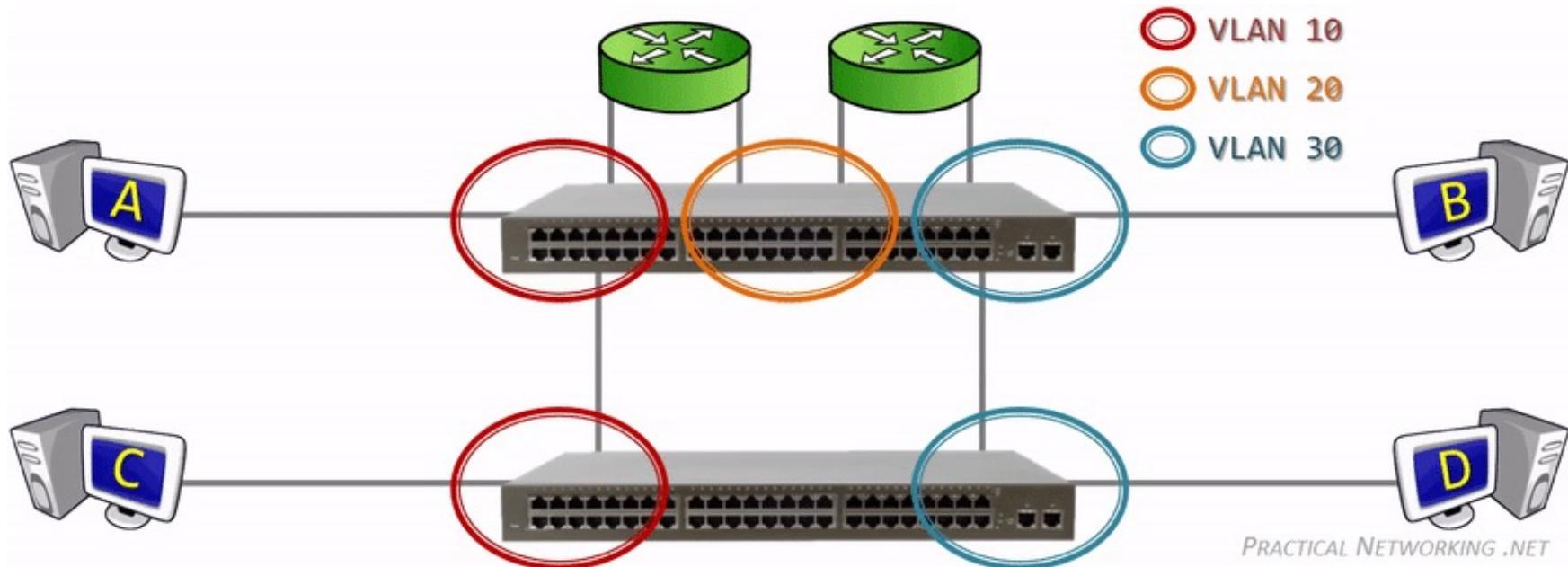
VLAN vs. Subnet

	VLAN	Subnet
Difference	<p>It is used to divide Layer 2 networks.</p> <p>Users in different VLANs can communicate with each other after VLANIF interfaces are configured for routing.</p> <p>A maximum of 4094 can be divided, and the number of devices in a VLAN is not limited.</p>	<p>It is used to divide Layer 3 networks.</p> <p>Users on different subnets can communicate with each other as long as they have reachable routes.</p> <p>The total number of subnets affects the maximum number of devices in each subnet.</p>
Relationship	<p>One or more subnets can be configured in a VLAN.</p>	<p>One or more VLANs can be configured on a subnet.</p>

VLAN 的另一个功能

□ 拓展虚拟交换机至多个物理交换机

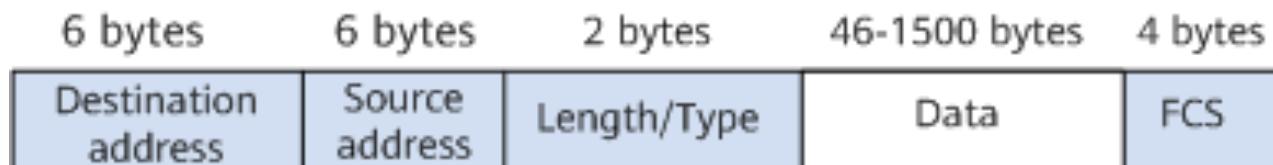
- 延伸二层网络的物理边界，跨房间、楼层、楼栋...
- **但物理边界仍然不会太大**
 - **性能问题**: 二层网络，过大的 VLAN 会导致广播流量非常大
 - **管理复杂性**: 更改 VLAN 成员需配置多个交换机，网络布线复杂
 - **故障隔离**: 过大的 VLAN 会导致故障印象范围变大
 - ...



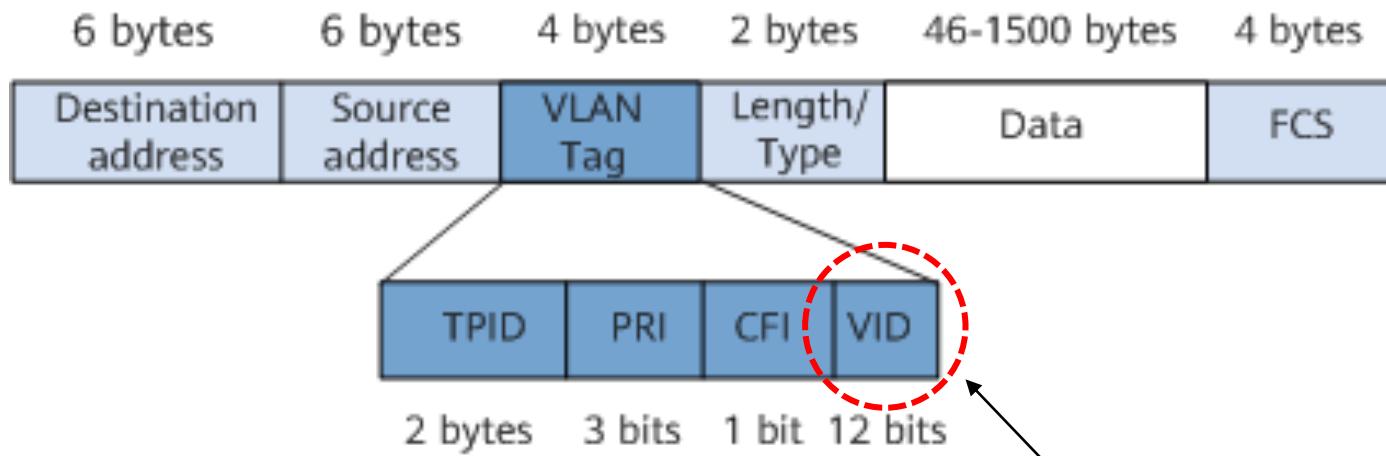
VLAN Tag

IEEE 802.1Q向以太网帧添加 4 字节的 VLAN 标签，使交换机能够识别接收到的帧所属的VLAN

Standard Ethernet frame



VLAN-tagged frame

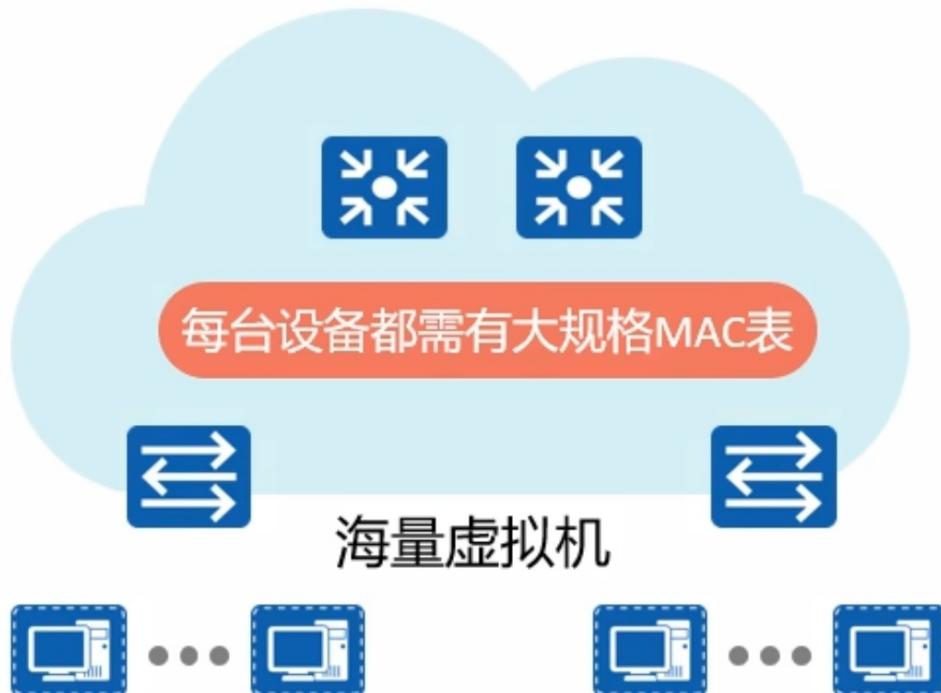


利用 VID (12 bits) 划分不同的 VLAN

VLAN 面临的问题

口虚拟机规模受设备表项规则限制

- 服务器虚拟化后，VM 的数量比原有物理机发生了巨大的增长
- 二层设备的 MAC 地址表规则较小，无法满足快速增长的 VM



Switch# show mac-address-table dynamic Mac Address Table			
Vlan	Mac Address	Type	Ports
10	0001.c7ad.e316	DYNAMIC	Fa0/24
10	0002.4ab7.1701	DYNAMIC	Fa0/24
11	0001.c7ad.e316	DYNAMIC	Fa0/24
11	0002.4ab7.1701	DYNAMIC	Fa0/24
12	0001.c7ad.e316	DYNAMIC	Fa0/24
12	0002.4ab7.1701	DYNAMIC	Fa0/24
12	000d.bd94.6b58	DYNAMIC	Fa0/1
12	0060.3e5c.bald	DYNAMIC	Fa0/2
12	00e0.f934.e3c4	DYNAMIC	Fa0/3
13	0001.c7ad.e316	DYNAMIC	Fa0/24
13	0002.4ab7.1701	DYNAMIC	Fa0/24

Switch#

VLAN 面临的问题

□ 网络隔离能力限制

- VLAN Tag 只有 12 bits, 4094 个不同的 VID (2 个保留 ID)
- 对于大型虚拟化云服务场景，租户数量远大于 VLAN 个数
- 传统二层网络中的 VLAN 无法满足网络动态调整的需求

802.1Q Tag帧：



VLAN ID长12 bit, 仅能表示4096个逻辑单元。

大型数据中心需支持的租户规模远大于该数目

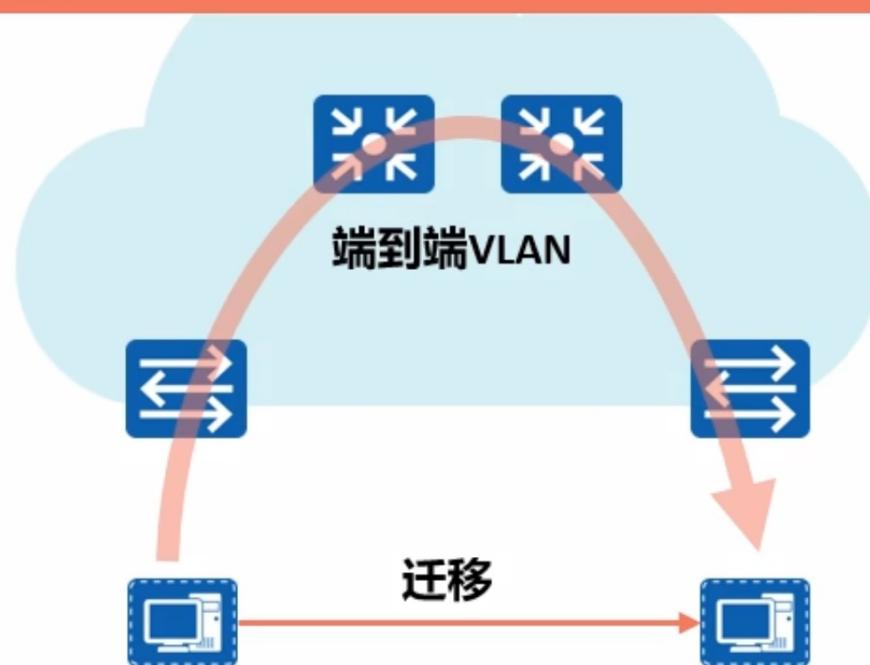
数据中心

VLAN 面临的问题

□ 虚拟机迁移范围受限

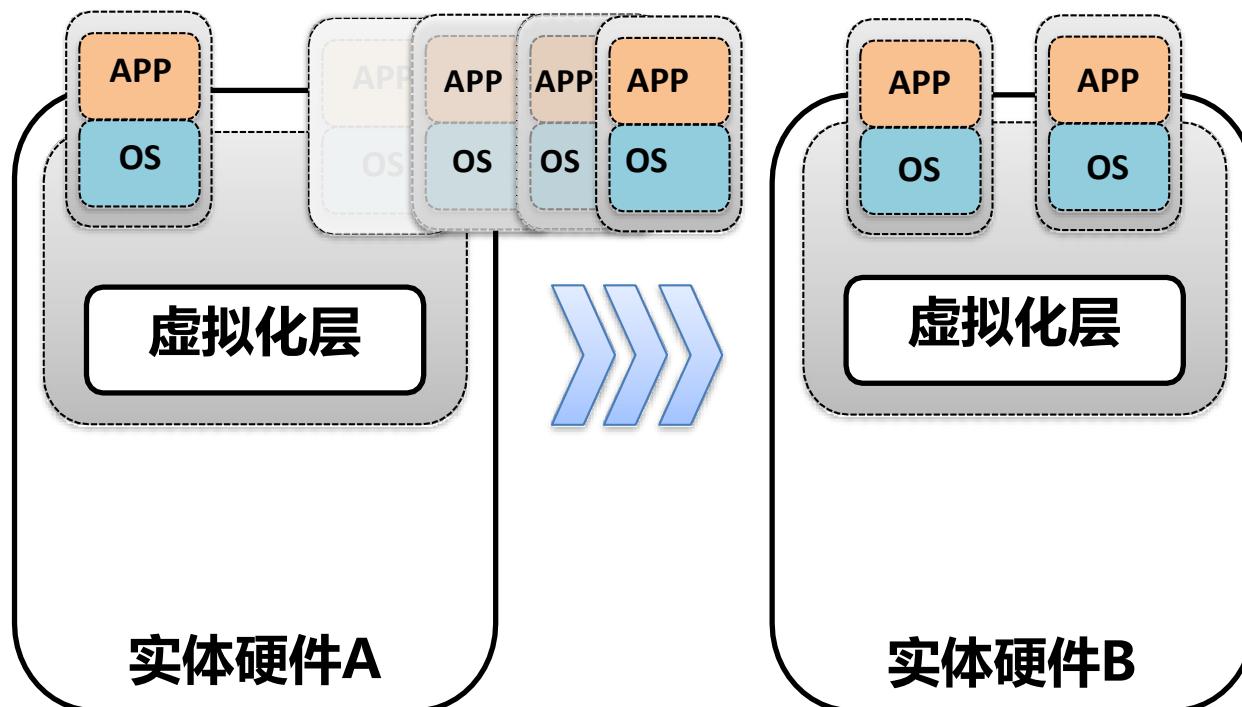
- 传统的二层网络将虚拟机迁移限制在较小的局部范围内
- 跨 VLAN 迁移会引起业务中断、信息丢失等

虚拟机只能在一个VLAN内迁移，VLAN范围存瓶颈



VM 迁移

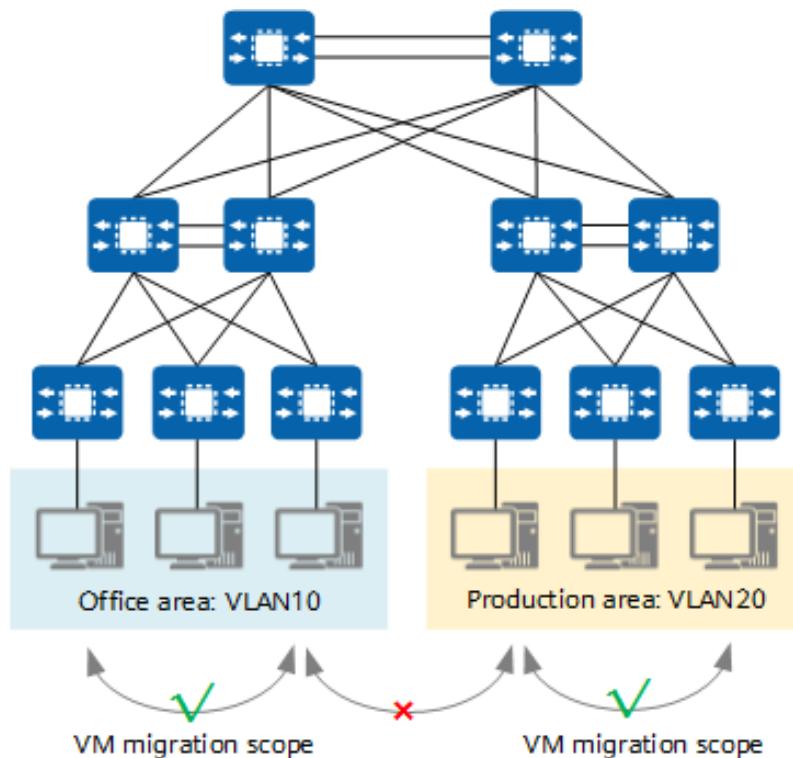
- 将 VM 从一个物理机迁移到另一个物理机
- 其 MAC 地址、IP 地址和运行状态（如 TCP 会话）要保持不变
- 因此，VM 迁移必须发生在二层网络！



VLAN 中的 VM 迁移

口在 VLAN 中，VM 迁移的范围仅限于 VLAN 内

- MAC 地址：通常在 VM 创建时分配，与 VM 关联，通常不会改变
- IP 地址：每个 VLAN 通常有自己的 IP 地址范围，跨 VLAN 迁移 VM 需要更改 IP 地址



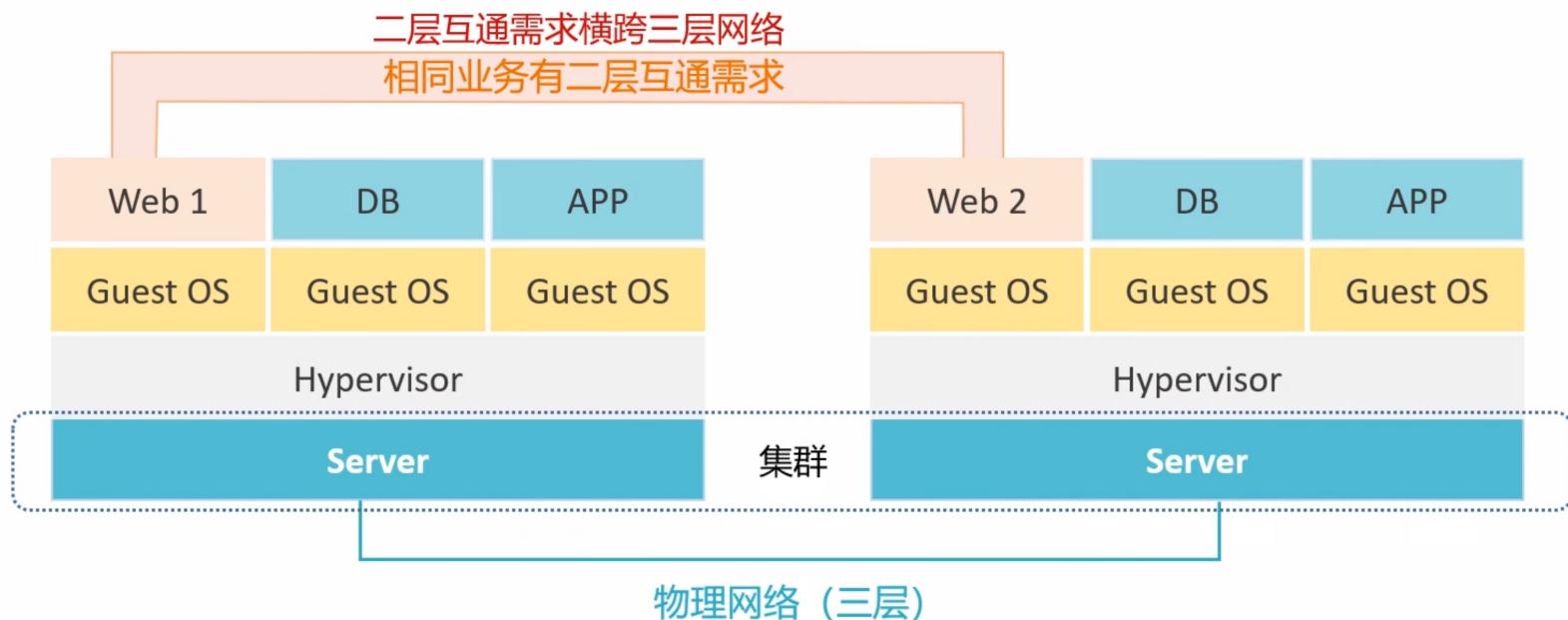
我们前面提到过，VLAN
的物理边界不会太大

云时代的网络新需求：二层扩展

□ 虚拟化/云计算集群内允许虚拟机任意迁移

- 相同业务（相同网段）虚拟机可能运行在不同的服务器
- 同一个虚拟机（相同 IP）先后运行在不同的服务器（物理位置）

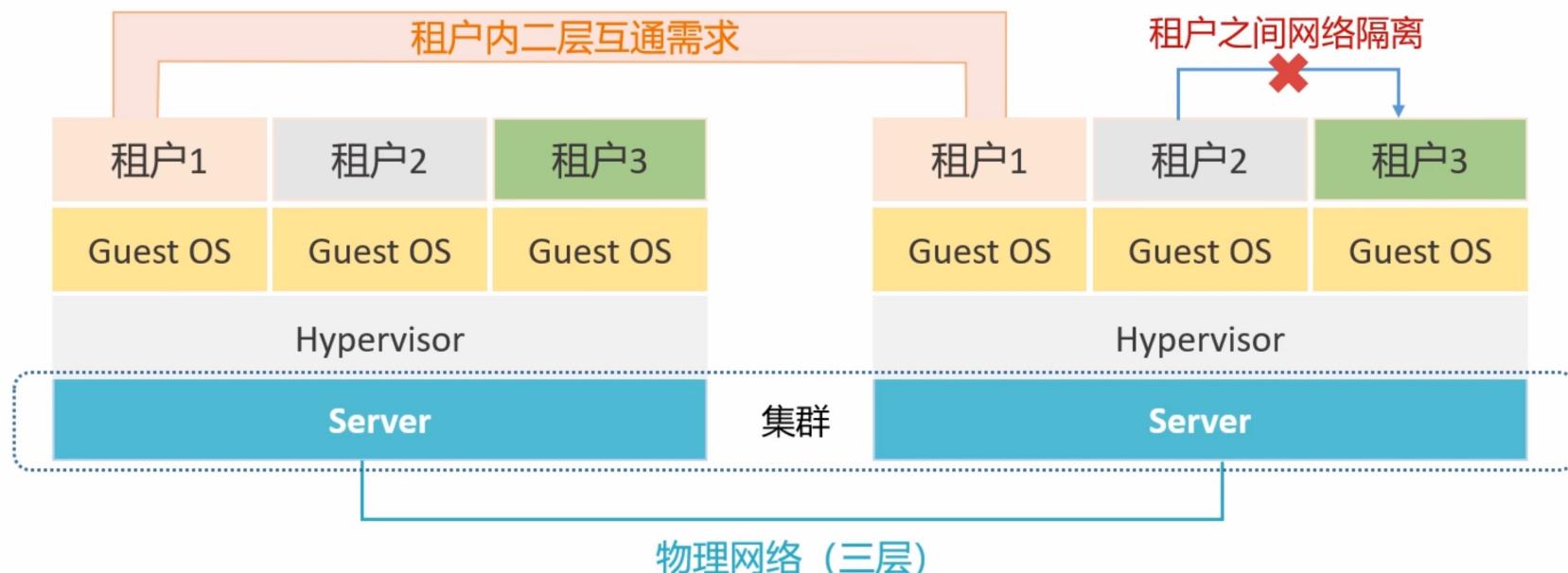
□ 物理服务器可能分布在地理位置跨度非常大的机房，因此需要使用三层进行互联



云时代的网络新需求：多租户隔离

口云化场景一般支持多租户，即不同用户共享物理资源

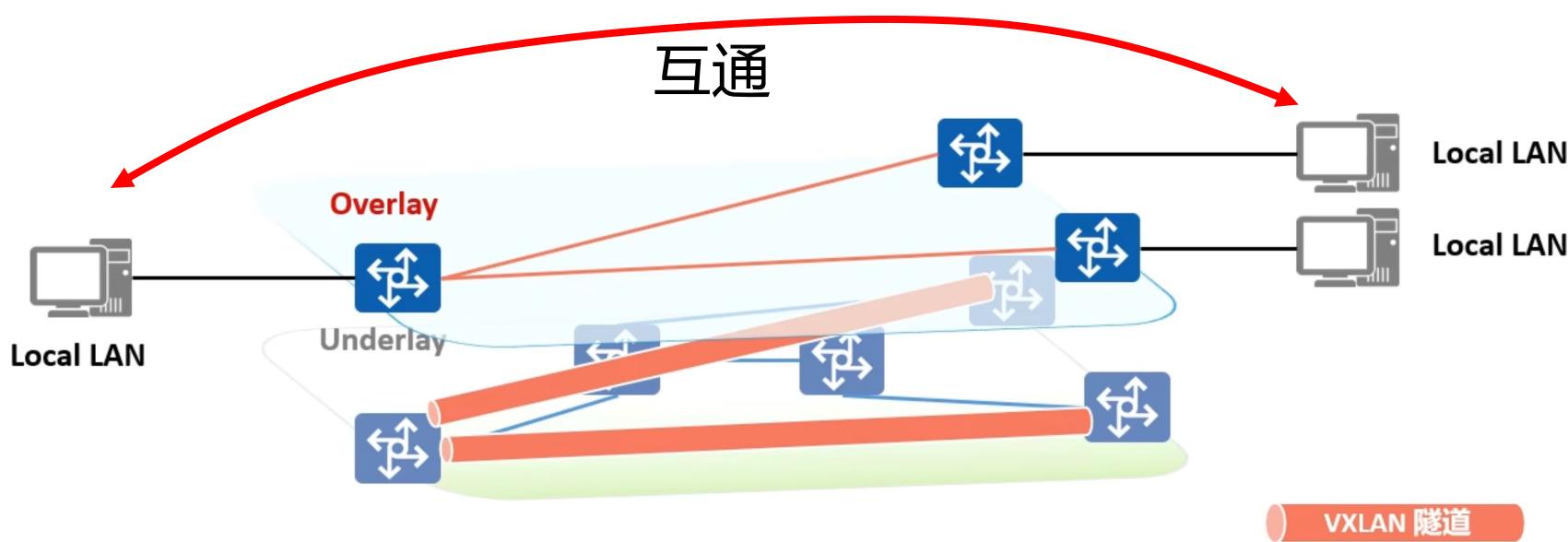
- **租户间隔离**：租户可能配置相同的 MAC 和 IP 地址
- **租户内互访**：租户内相同网段能够直接进行二层通信，即便处于不同物理位置的机房中



VxLAN 简介

□ VxLAN 本质上是一种 VPN (Virtual Private Network) 技术

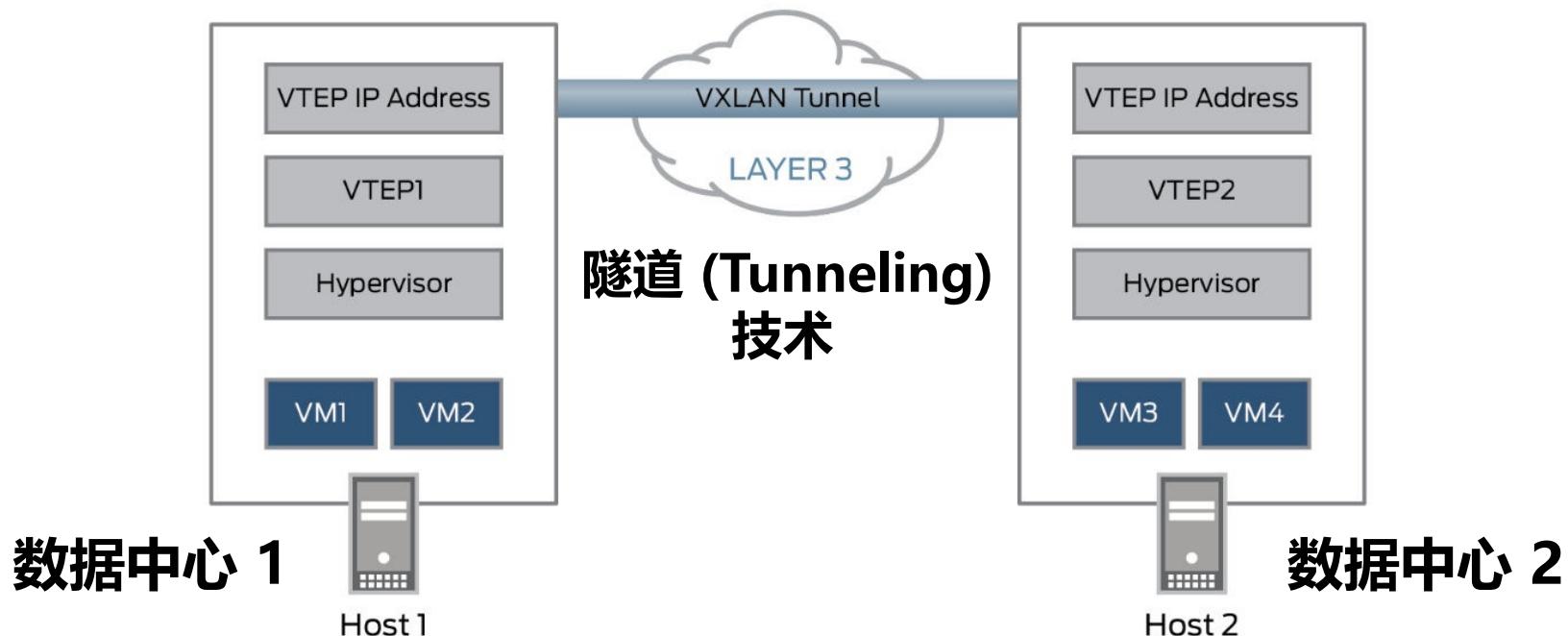
- 在路由可达的物理网络上叠加虚拟网络
- 通过VxLAN网关之间的VxLAN隧道实现VxLAN网络内部互通
- 也可以实现与传统的非 VxLAN 网络互通



VxLAN 简介

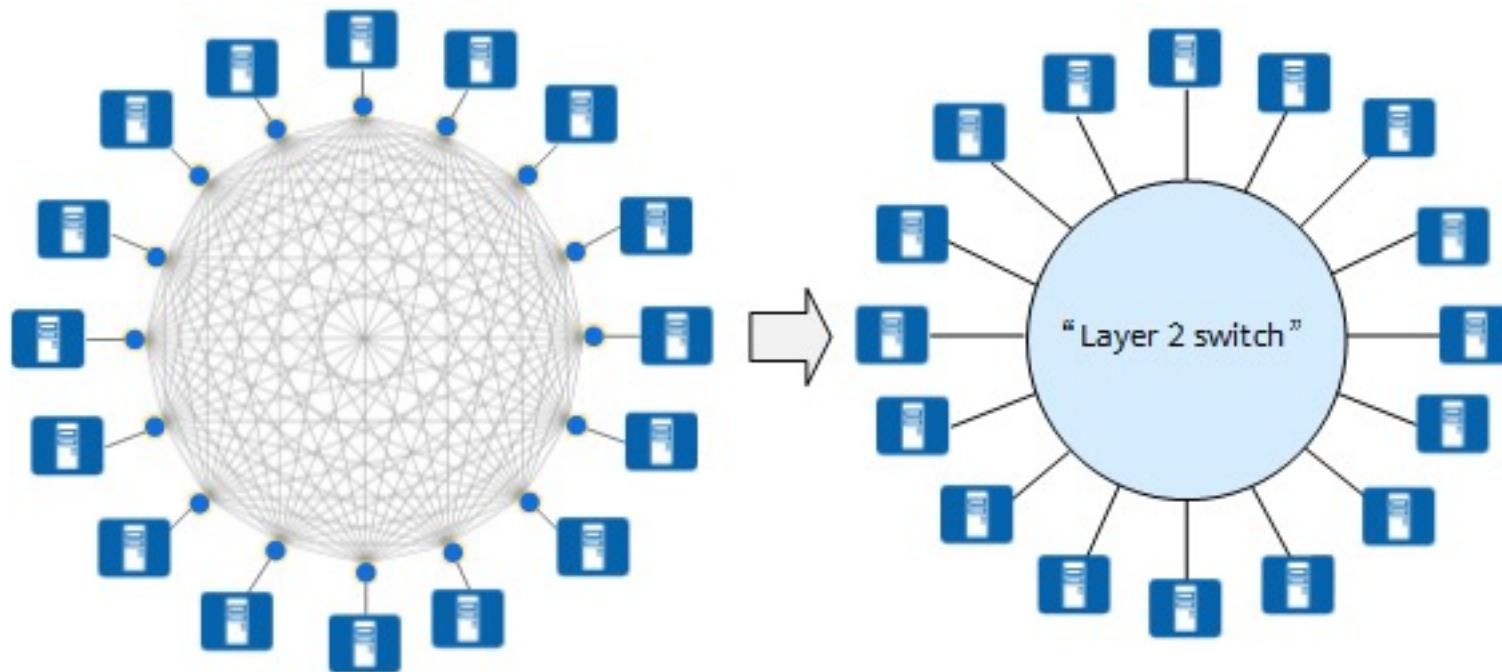
□ VxLAN 借助三层网络上扩展二层的连接

- 采用 **MAC in UDP 封装 (encapsulation)** 来延伸二层网络
- 将以太报文封装在 IP 报文之上，**像普通 IP 包一样通过路由在三层网络中传输**，无需关注虚拟机的 MAC 地址
- 通过路由网络，**虚拟机迁移不受网络架构限制**

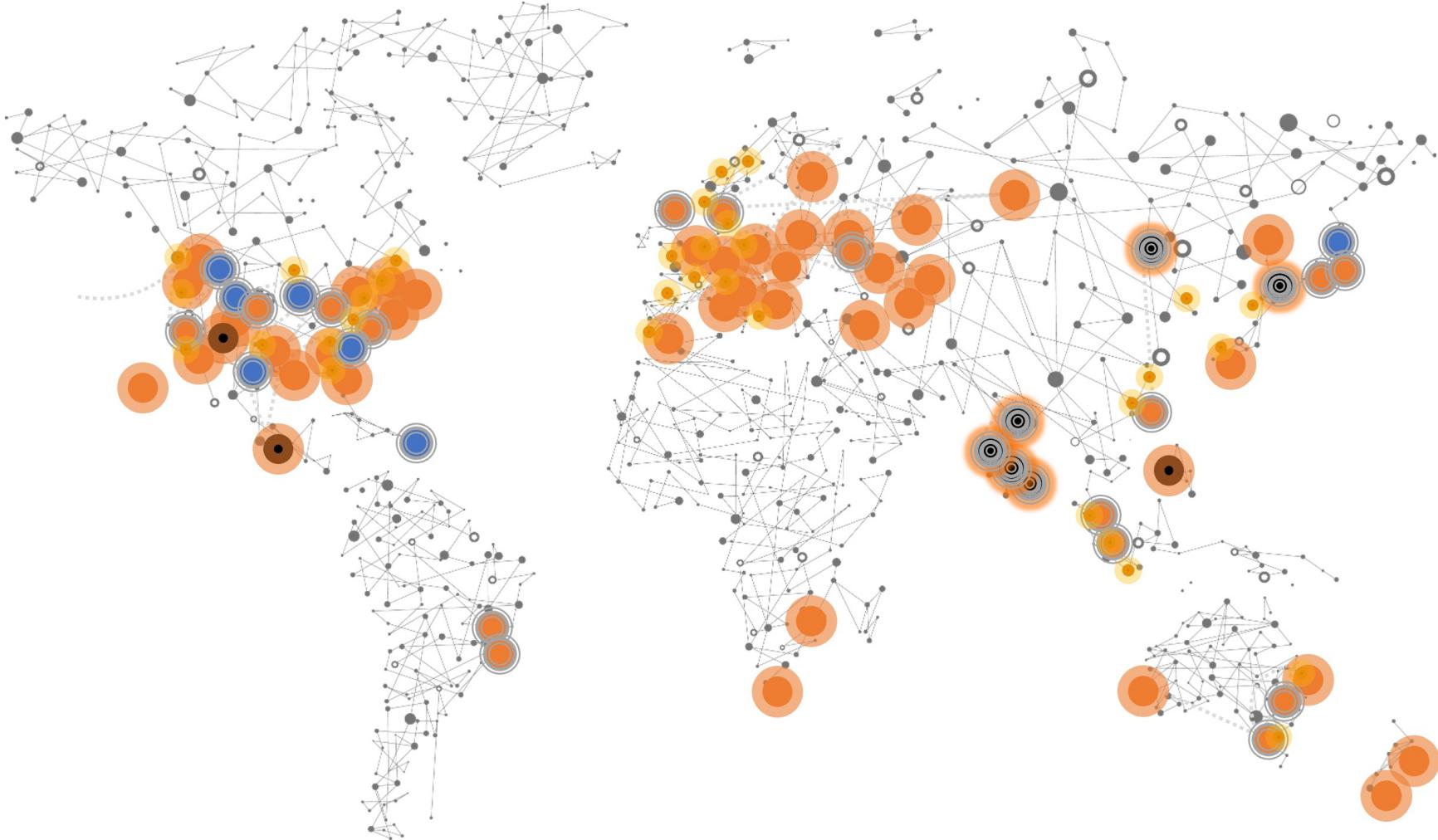


VxLAN 中的 VM 迁移

- VxLAN 使 VM 的 MAC 和 IP 地址封装在 UDP 包中，可传输到任意其他地理位置属于同一 VLAN 的物理机
- VxLAN 将整个基础设施网络虚拟化为一个大型的“2层虚拟交换机”

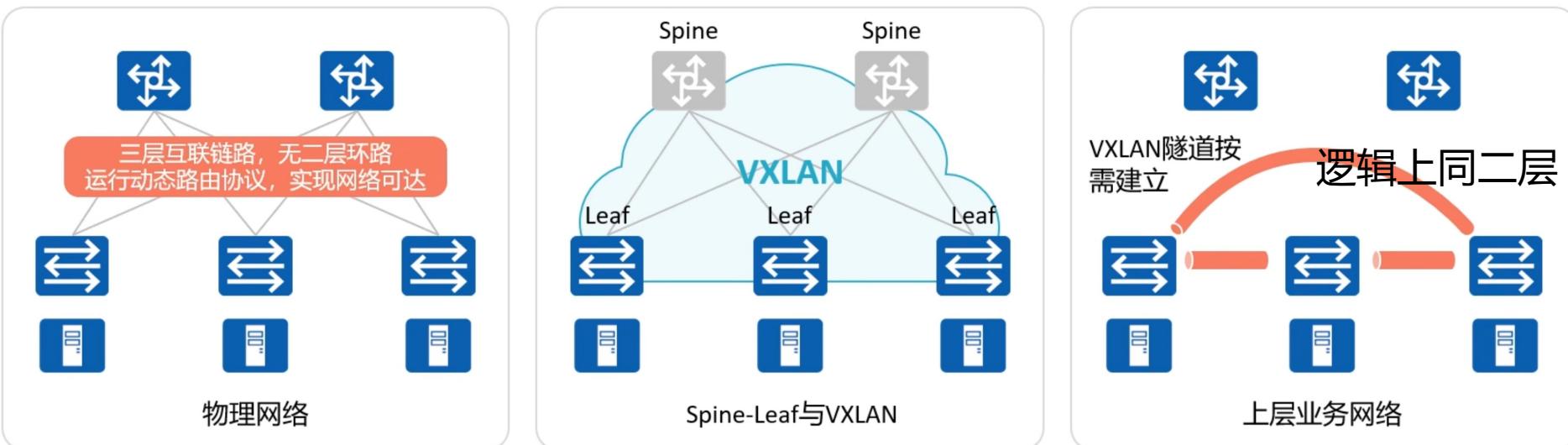


VxLAN 突破了二层网络的边界限制



VxLAN 在数据中心的应用

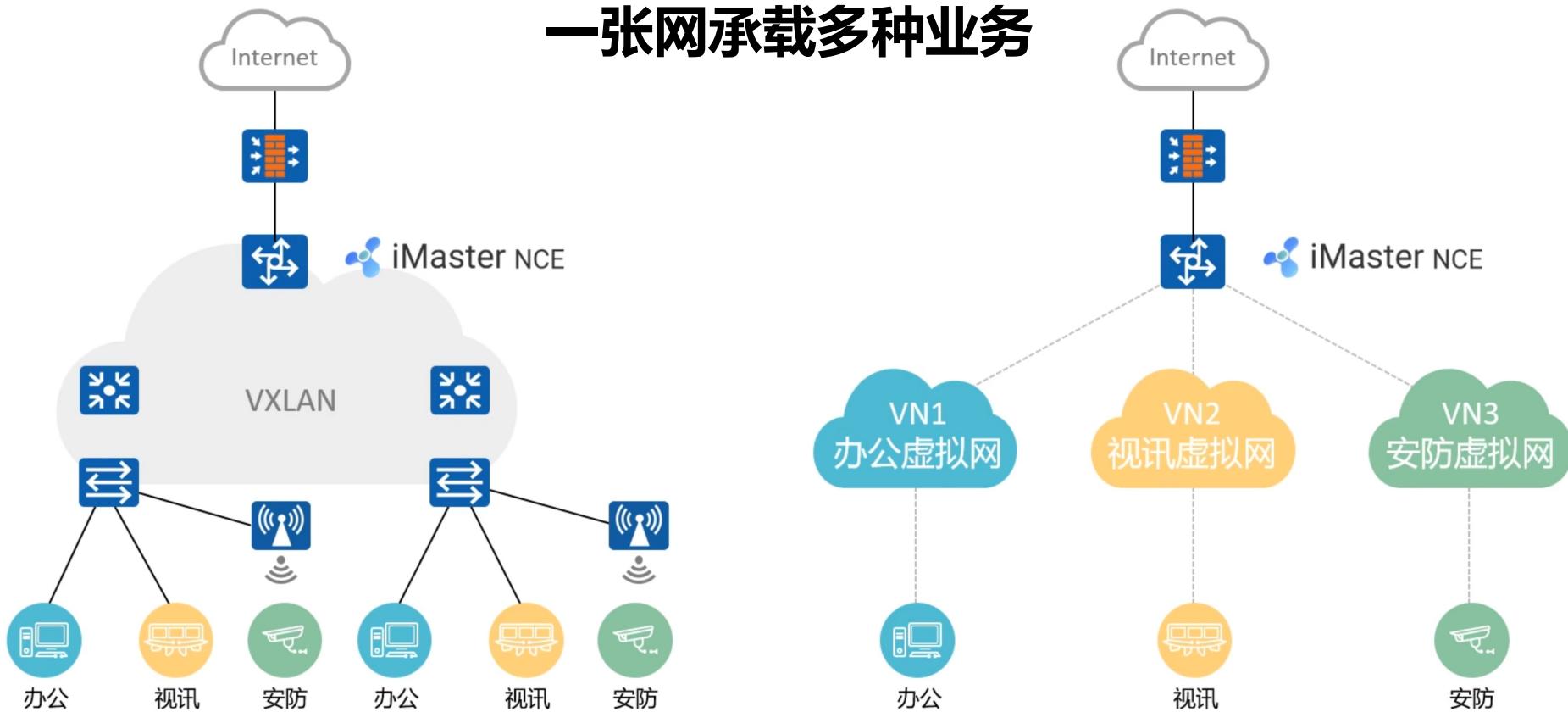
- 数据中心采用 Spine-Leaf 两层物理架构，结合 VxLAN 应用
- Spine 节点执行路由器转发，转发时不感知 VxLAN
- Leaf 节点负责资源接入，完成 VxLAN 封装及解封装
- 数据中心的业务均由 VxLAN 承载



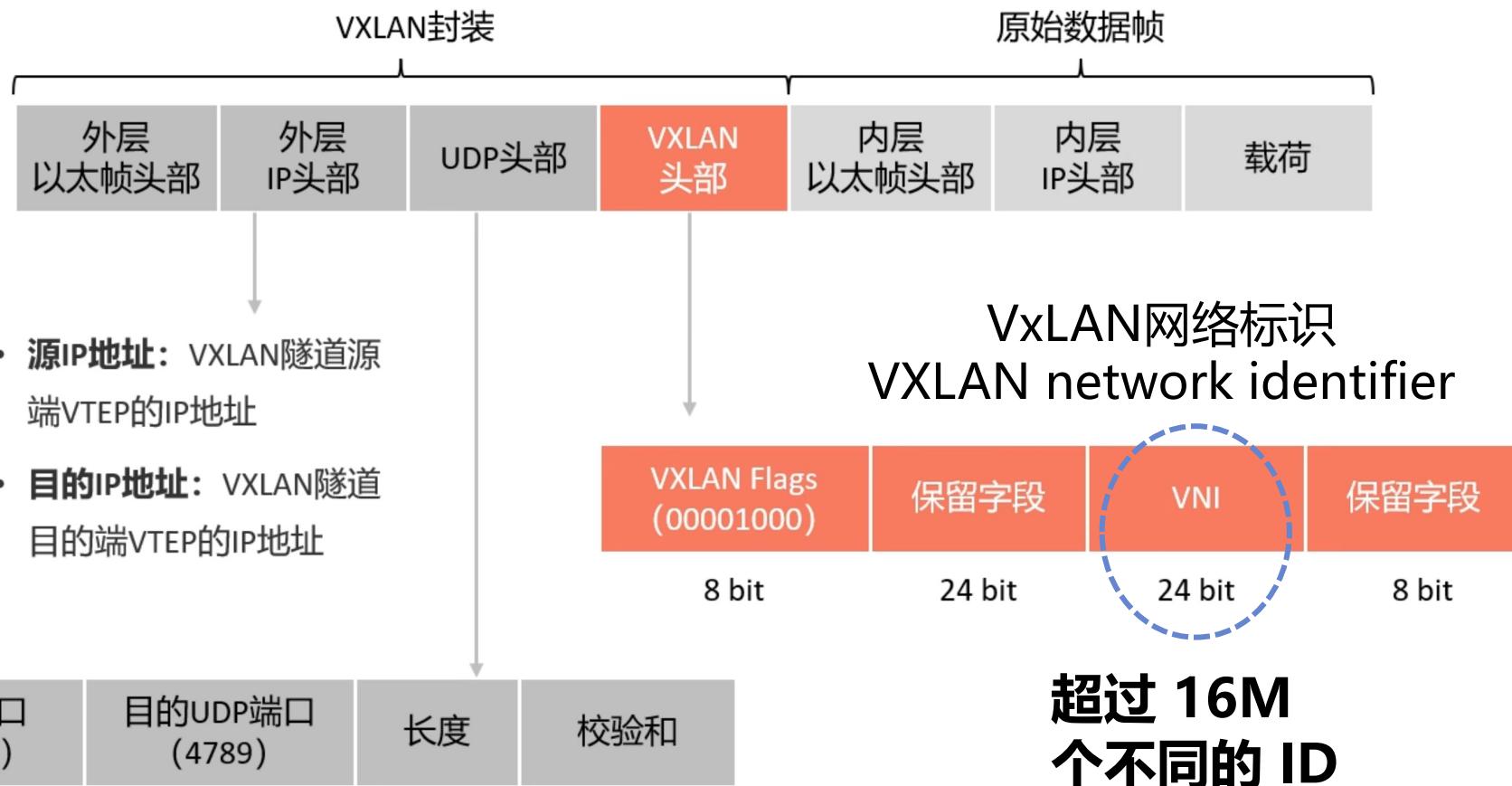
在园区网络中使用 VxLAN

实现 “一网多用”

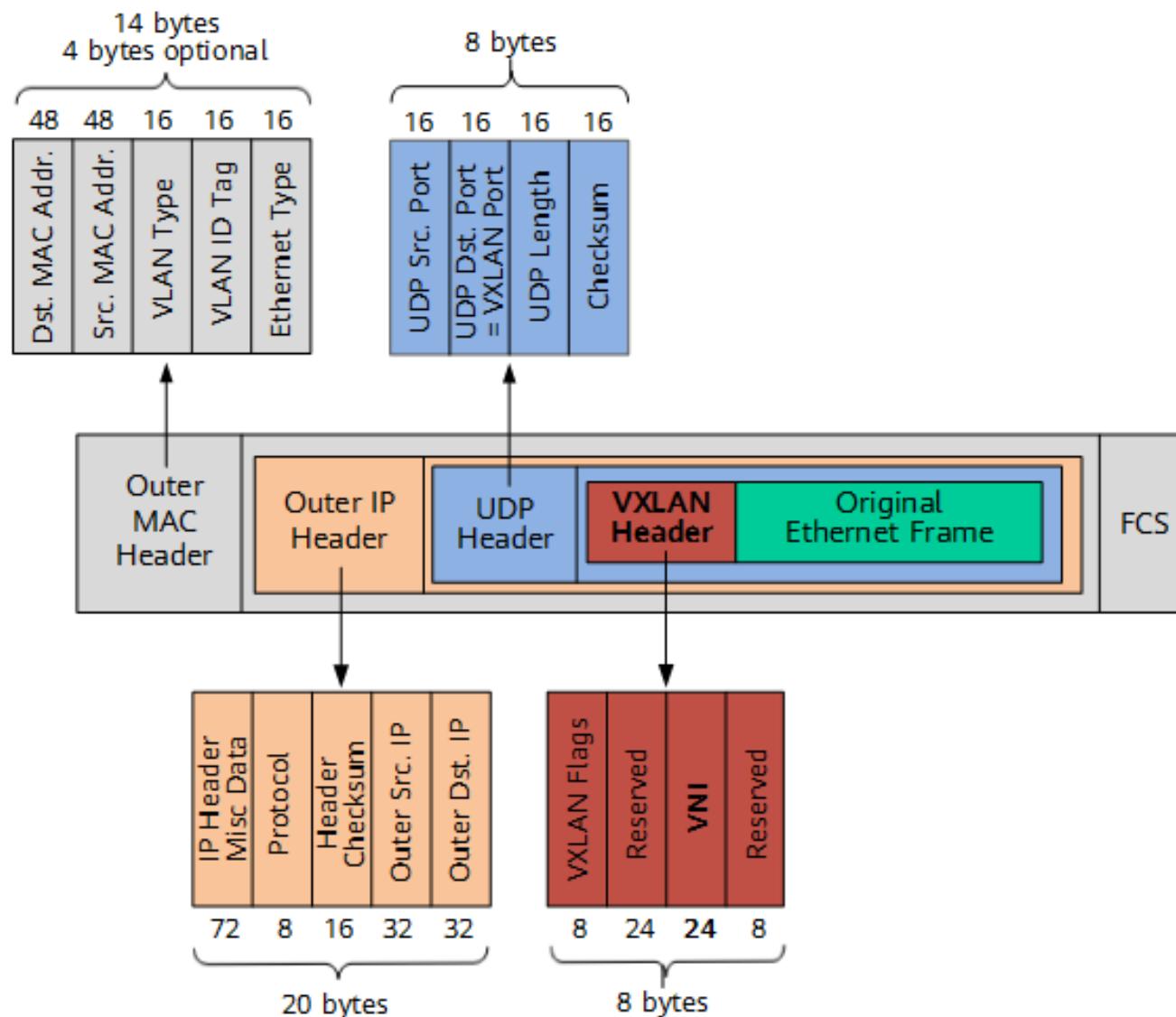
一张网承载多种业务



VxLAN 的报文格式



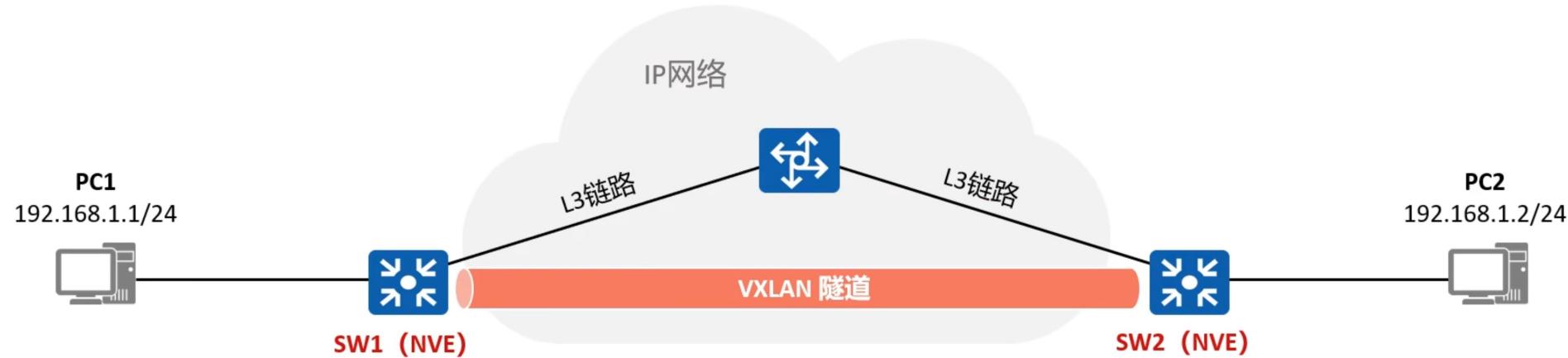
VxLAN 数据包



VxLAN 基本概念：NVE

□ NVE (Network Virtualization Edge, 网络虚拟边缘)

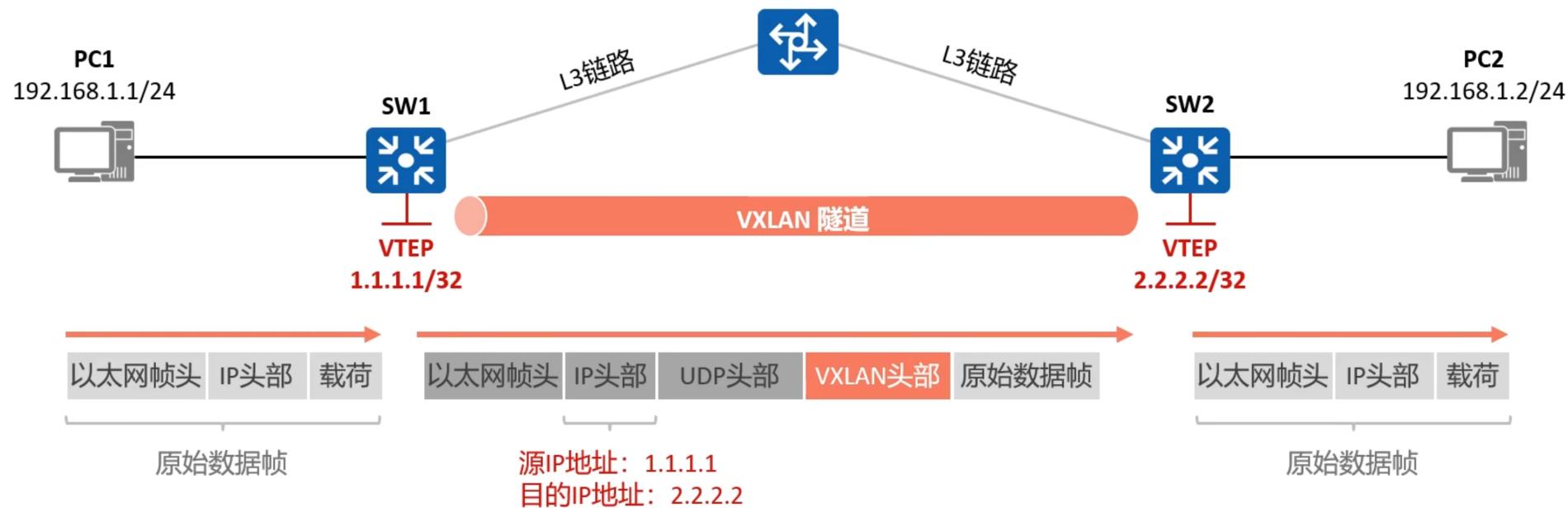
- 实现网络虚拟化功能的实体，软硬件交换机均可
- NVE 在三层网络上构建二层虚拟网络，是运行 VxLAN 的设备
- 把普通网络连接到 VxLAN 网络



VxLAN 基本概念：VTEP

□ VTEP (VxLAN Tunnel Endpoints, VxLAN 隧道端点)

- VTEP 位于 NVE 中，用于 VxLAN 报文的封装和解封装
- VxLAN 报文 (其外层 IP 头部) 中
 - 源 IP 地址为源端 VTEP 的 IP 地址
 - 目的 IP 地址为目的端 VTEP 的 IP 地址



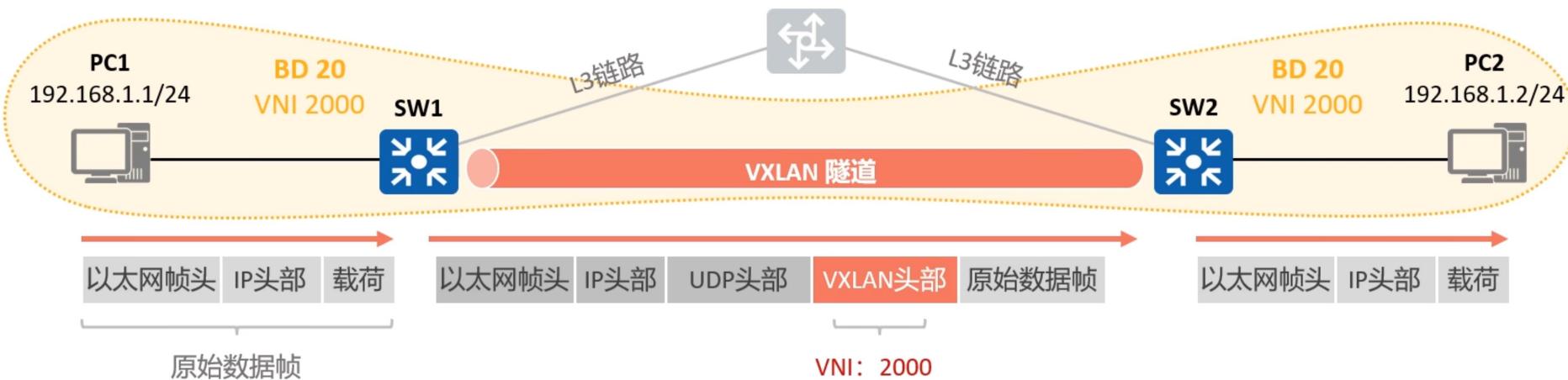
VxLAN 基本概念：VNI 与 BD

□ VNI (VxLAN Network Identifier, VxLAN 网络标识)

- 类似VLAN ID，用于区分VxLAN，不同 VxLAN 段不能二层互通
- 一个租户可拥有一个或多个 VNI，最多支持超 16M 租户

□ BD (Bridge Domain, 桥域)

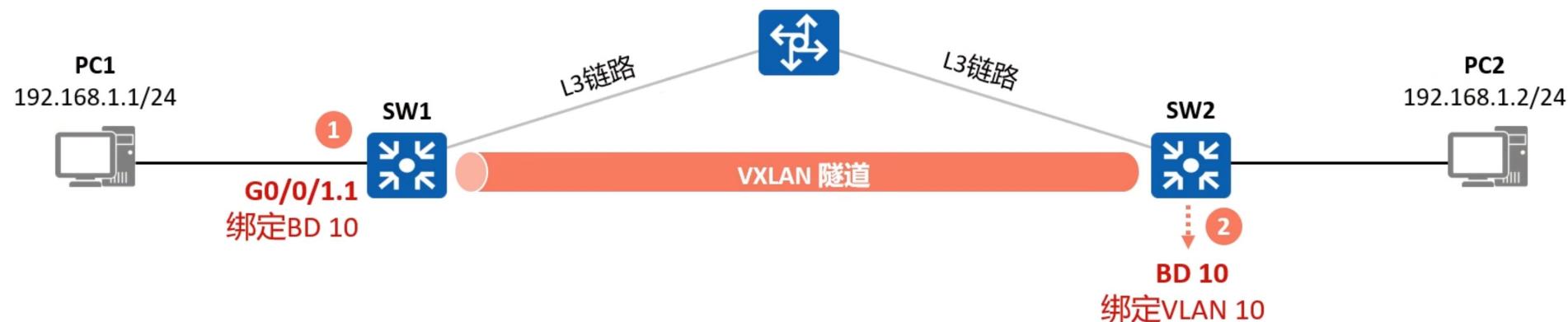
- 类似传统网络中采用 VLAN 划分广播域，在 VxLAN 网络中一个 BD 标识一个大二层广播域（逻辑上的广播域）
- VNI 以 1:1 方式映射到广播域 BD，同一个 DB 内可二层互通



VxLAN 基本概念：VAP

□ VAP (Virtual Access Point, 虚拟接入点)

- 实现 VxLAN 的业务接入
- VAP 有两种配置方式
 - 二层子接口接入：例如在 SW1 创建二层子接口关联 BD 10，则这个子接口下的特定流量会被注入到 BD 10
 - VLAN 绑定接入：例如在 SW2 配置 VLAN 10 与广播域 BD 10 关联，则所有 VLAN10 的流量会被注入到 BD 10



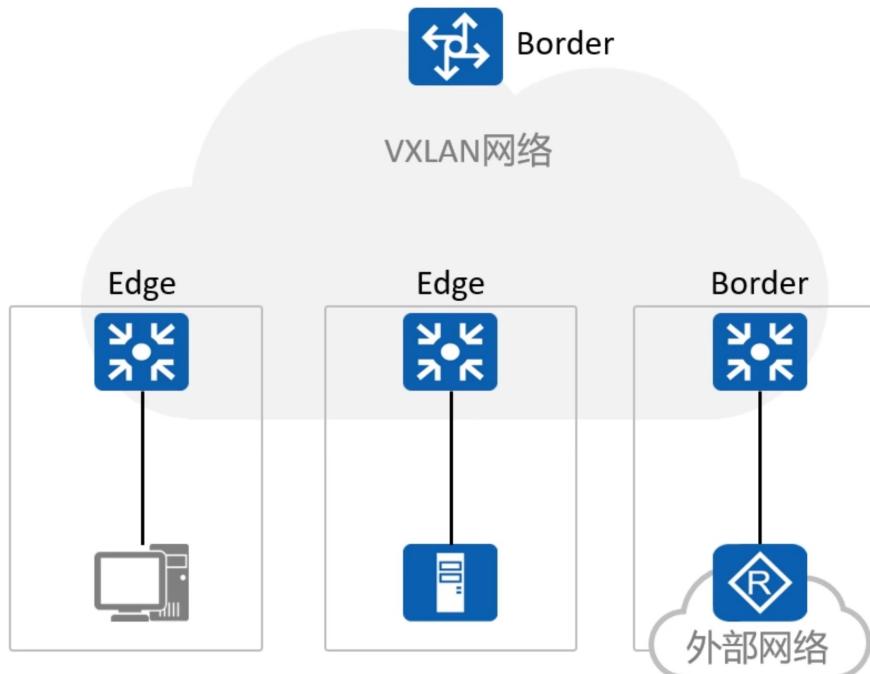
VxLAN 基本概念：Edge 和 Border

□ Edge

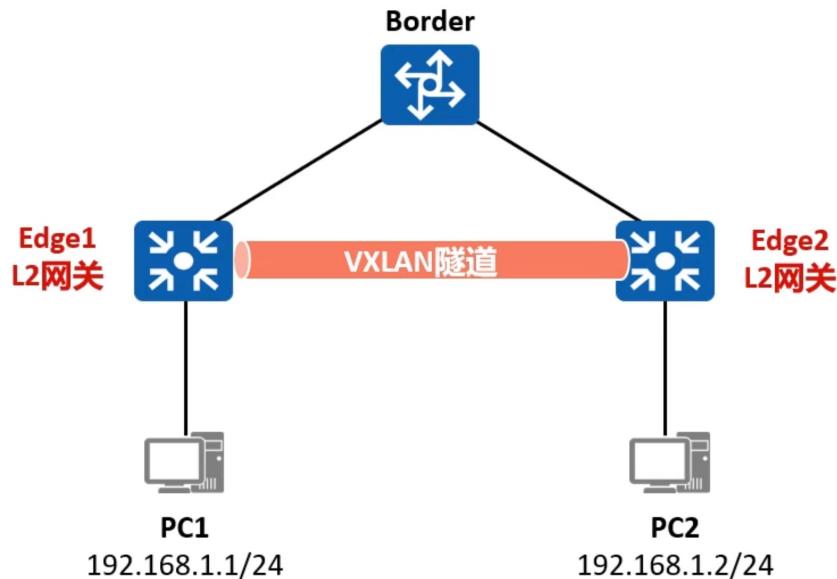
- VxLAN 网络的边缘接入设备，传统网络的流量由此进入 VxLAN

□ Border

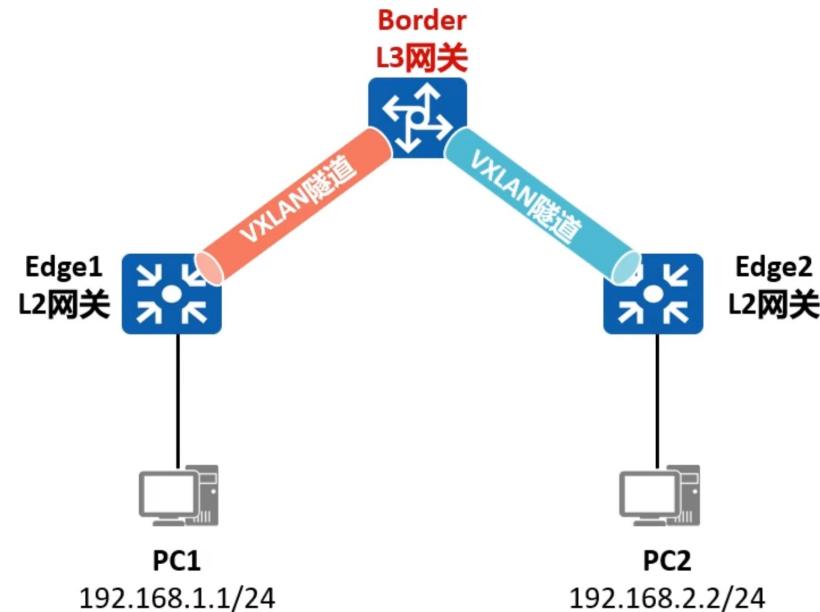
- VxLAN 网络和外部网络通信的结点，用于外部流量进入 VxLAN 网络或者 VxLAN 内部流量访问外部
- 一般具有三层转发能力的设备 (如 Router、Firewall)



VxLAN 基本概念：VxLAN 二、三层网关



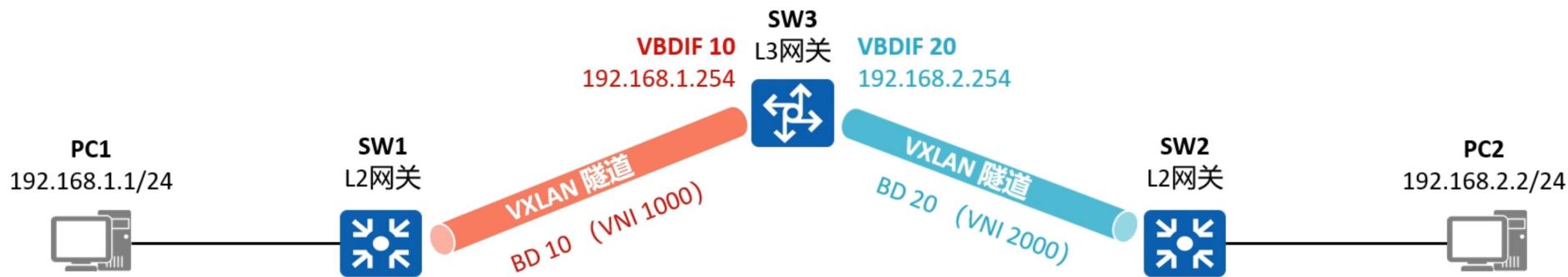
二层 (L2) 网关：实现流量进入 VxLAN 网络，也可用于同一 VxLAN 网络内终端的同子网通信



二层 (L3) 网关：用于 VxLAN 网络内终端的跨子网通信以及对外部 (非 VxLAN 网络) 的访问

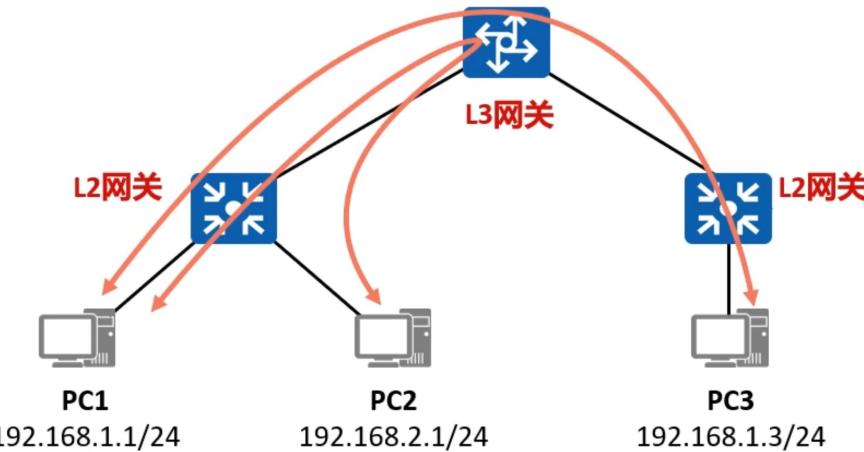
VxLAN 基本概念：VBDIF

- 类似传统网络采用VLANIF实现不同广播域互通，VxLAN引入了VBDIF
- VBDIF接口在VxLAN三层网关上配置，是基于BD创建的三层逻辑接口
- 通过VBDIF接口可实现不同网段用户通过VxLAN网络通信，及 VxLAN 网络和非VxLAN网络间的通信，也可实现二层网络接入三层网络

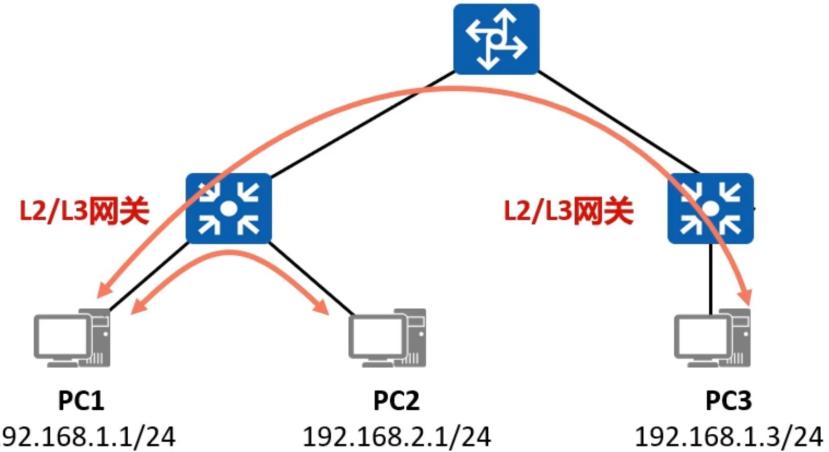


VxLAN 基本概念：分布与集中式网关

集中式网关



分布式网关



L3网关部署在一台设备上，所有跨子网的流量都通过该设备转发

优点：跨子网流量集中管理，简化网关部署和管理

缺点：转发路径并非最优。ARP表项规格瓶颈：由于采用集中式网关，网关上需要维护大量通过VxLAN接入网络的终端 ARP

VTEP设备既是L2又是L3网关，非网关节点对VxLAN隧道不感知，仅作为VxLAN报文的转发节点

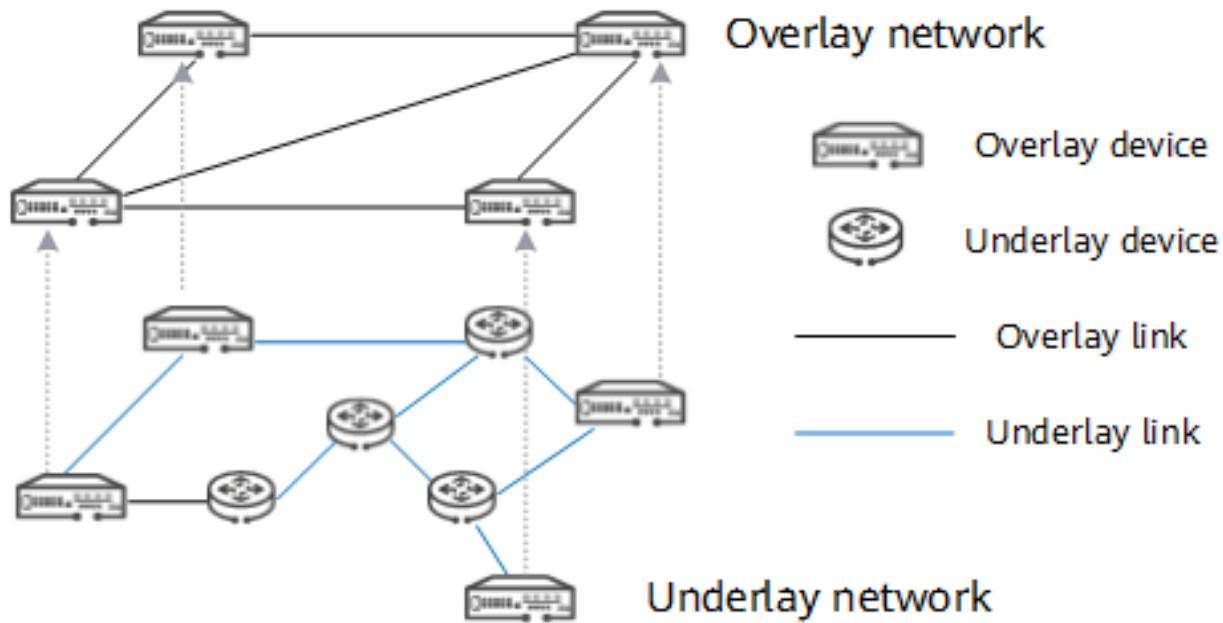
优点： VTEP节点只学习连接在本节点下终端的ARP表项，解决了集中式三层网关带来的ARP表项瓶颈问题，网络规模扩展能力强

缺点： 相对集中式部署配置、实现复杂，部署工程量大

覆盖网络

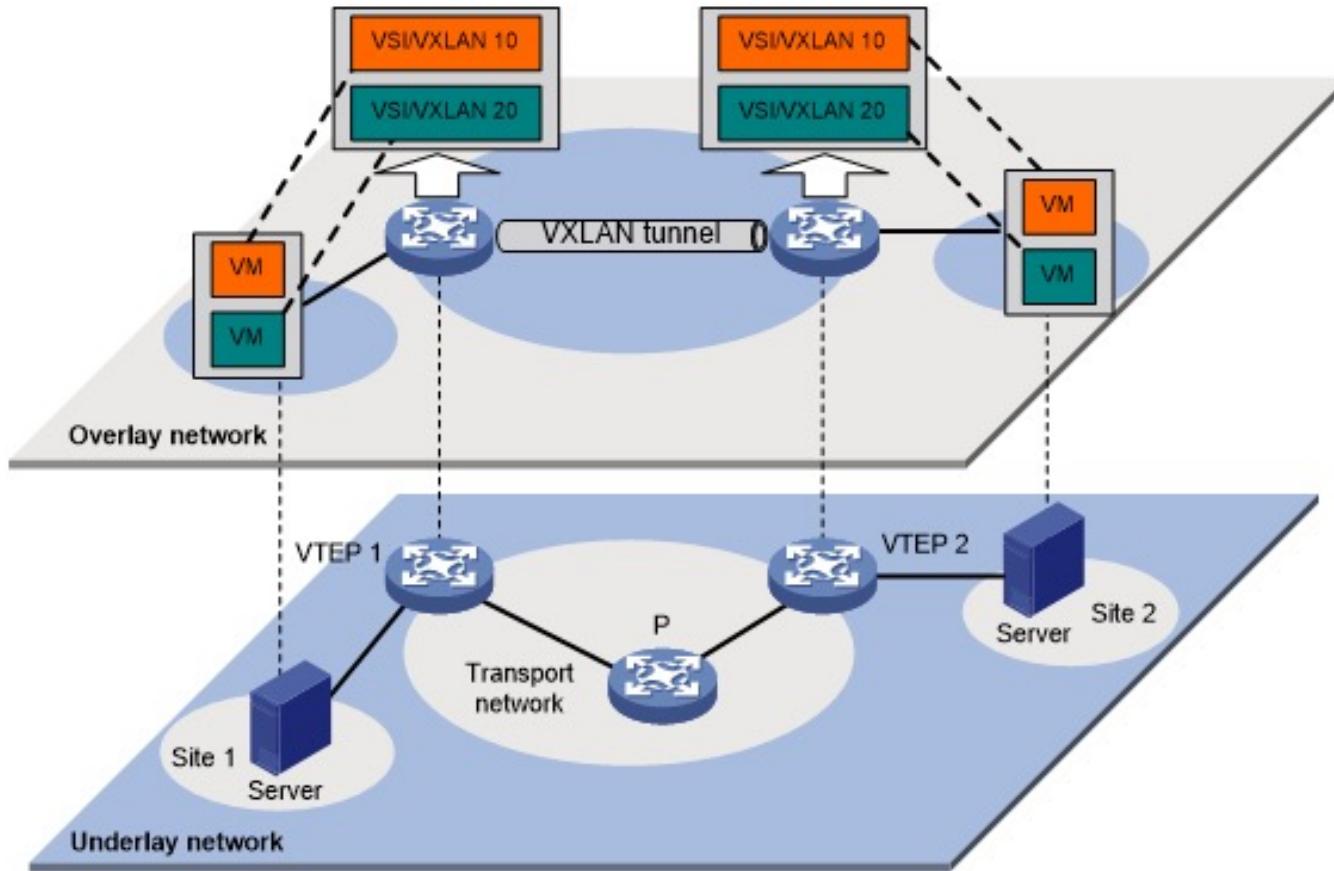
覆盖网络 (Overlay network)

- 口 覆盖网络是建立在现有物理网络基础设施之上的**虚拟网络**
- 口 允许网络管理员独立于底层物理网络拓扑配置，创建、管理和维护虚拟网络
- 口 使用各种**隧道和封装协议**，将虚拟网络的流量封装在可以通过物理网络传输的数据包中，如 VxLAN、NVGRE 和 GENEVE



覆盖网络

口通过 VxLAN 构建的 Overlay network



虚拟私有云 (Virtual Private Network)

是一个租户专用的虚拟网络，在逻辑上与云基础设施中的其他虚拟网络隔离

租户可以将实例、数据库和容器等云资源部署到虚拟私有云中

租户可以构建类似传统网络的虚拟网络环境，如定义自己的IP地址范围、创建子网以及配置路由表和网络网关



虚拟私有云 (Virtual Private Cloud)



HUAWEI CLOUD

Virtual Private Cloud (VPC)

Virtual Private Cloud (VPC) allows you to isolate online resources with virtual private networks. VPC enables your cloud resources to securely communicate with each other, the internet, and on-premises networks.



Alibaba Cloud

Virtual Private Cloud

A virtual private cloud service that provides an isolated cloud network to operate resources in a secure environment.



Amazon Virtual Private Cloud (Amazon VPC)

Define and launch AWS resources in a logically isolated virtual network



Azure Virtual Network

Create your own private network infrastructure in the cloud.



中山大學 软件工程学院
SUN YAT-SEN UNIVERSITY SCHOOL OF SOFTWARE ENGINEERING

谢谢

陈壮彬
软件工程学院
chenzhb36@mail.sysu.edu.cn