大数据的信息基础设施

网络虚拟化

陈一帅

yschen@bjtu.edu.cn

北京交通大学电子信息工程学院

内容

- 背景
- 网络的分区扩展和虚拟化
- VLAN 技术
- VXLAN 技术

网络解聚合

- 硬件和软件（OS）分离，可分别买
- 硬件
  - Bare-metal 网络交换机（或白盒交换机）
  - 商用交换芯片（Broadcom Trident 芯片，Barefoot）
- 软件
  - 用自己的 NOS 和 App

OS 的网络栈

- OS 是资源和应用之间的 moderator
- 本地网络状态，路由表，ARP 表，VXLAN 隧道，ACL，Bridge 表，计数器，都在 NOS 的 Kernel 里
- 用“用户空间”的设备驱动，写进交换芯片中

内容

- 背景
- 网络的分区扩展和虚拟化
- VLAN 技术
- VXLAN 技术

网络分区

- 分区，为了可扩展
- 一个设计概念，IT 架构师在遇到诸如
  - 主机组的流量隔离
  - 不同的安全区域
  - 具有重叠 IP 地址的不同设备组
  - 不同的路径行为
  - 共享故障域
- 最常见的
  - 交换式以太网使用 VLAN 来分区
- 网络虚拟化

以太网 VLAN 分区扩展

- 桥非常流行
  - 硬交换机、上层协议（IP、IPX）对桥是透明的
  - 0 配置，桥会自学习，上来就用
- 基于桥的虚拟网：VLAN
  - 限制 flooding 只对有自己虚拟网节点的端口
  - 一个桥或交换机，只支持一个广播域。
  - VLAN 提供多个广播域，提高网络利用率

网络虚拟化

虚拟化的定义

虚拟化意味着应用程序可以使用资源，而无需考虑资源的位置，技术接口，实现方式，使用的平台以及可用资源的数量。

虚拟化的好处

- 共享：分解大量资源，大容量或高速的服务器
- 隔离：免受其他租户的保护，例如虚拟专用网
- 聚合：将许多资源合并到一个资源中，例如存储
- 动态：快速分配，更改/移动，负载平衡（例如虚拟机）
- 易于管理：⇒ 简单（分发，部署，测试）

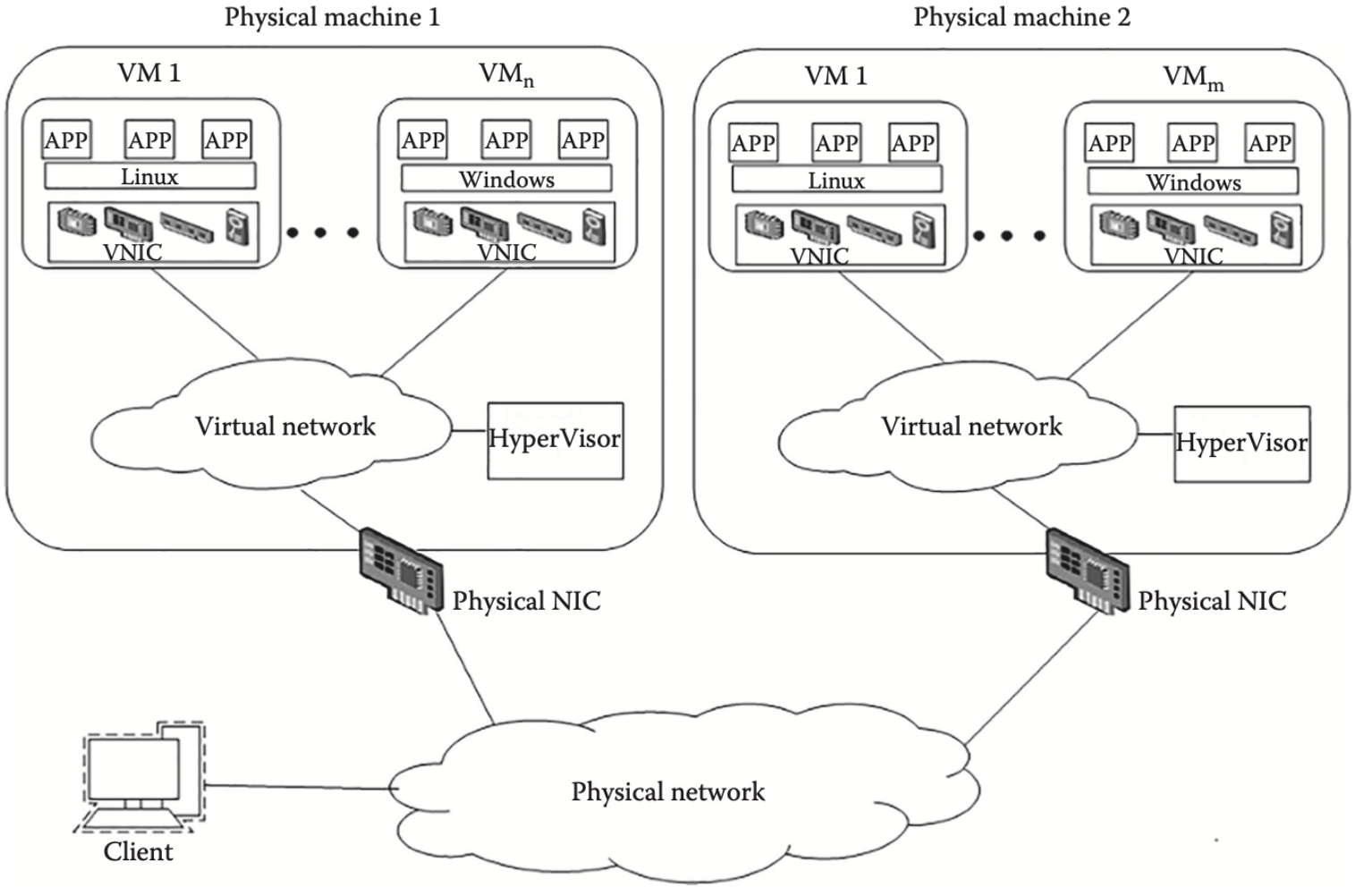
网络虚拟化

- 接口
- 链路
- 前向表
- 策略表（接入控制）
- NAT
- 缓存
- 队列
- 整个网络

网络虚拟化效果

- 允许租户在多租户网络中形成覆盖网络，租户自己可以控制
  - 连接层：租户网络可以是 L2，而提供者是 L3，反之亦然
  - 地址：MAC 地址和 IP 地址
  - 网络分区：VLAN 和子网
  - 节点位置：自由移动节点
- 使提供商可以为大量租户提供服务，无需担心
  - 客户网络中使用的内部地址
  - 客户节点数
  - 客户节点位置
  - 客户分区（VLAN 和子网）的数量和设置

基于虚拟网络的云计算平台



虚拟网络类型

- 虚拟的网络
  - L2
  - L3
- 实现方式
  - Inline
  - Overlay

虚拟的 L2 网络

- VLAN
  - L2 using L2
  - STP（Spanning Tree）广播管理
  - 用得最广
  - 将相互直接通信的网络设备和主机组合在一起
- VXLAN
  - L2 using L3
  - 在 UDP 之上

虚拟的 L3 网络

- VRF
  - 虚拟路由和前向
  - 为虚拟专用网（VPN）提供路由和路径隔离
  - Router 实现
  - 每个虚拟网络都有一个分别的路由表
  - 路由表查找时有一个 VRFID
  - 物理网卡带一个虚拟网络标签，指示逻辑接口
- MPLS
  - L3 using L3

实现方式

- Inline
  - 每个路由器或交换机都设置
  - VLAN，VRF
  - VRF 的每个虚拟网络都有一个分别的路由表
- Overlay
  - 基于隧道
  - 只有边缘路由器才设置，里面的路由器不知道
  - MPLS，VXLAN，IP-based VPN

内容

- 背景
- 网络的分区扩展和虚拟化
- VLAN 技术
- VXLAN 技术

VLAN 分配方法

- 可以基于下面的方法分配 VLAN 帧
  - 源接口
  - 源 MAC 地址
  - 源 IP 地址
  - 应用程序（由 TCP 或 UDP 目标端口定义）
- 最常见的 VLAN 分配方法是基于源接口

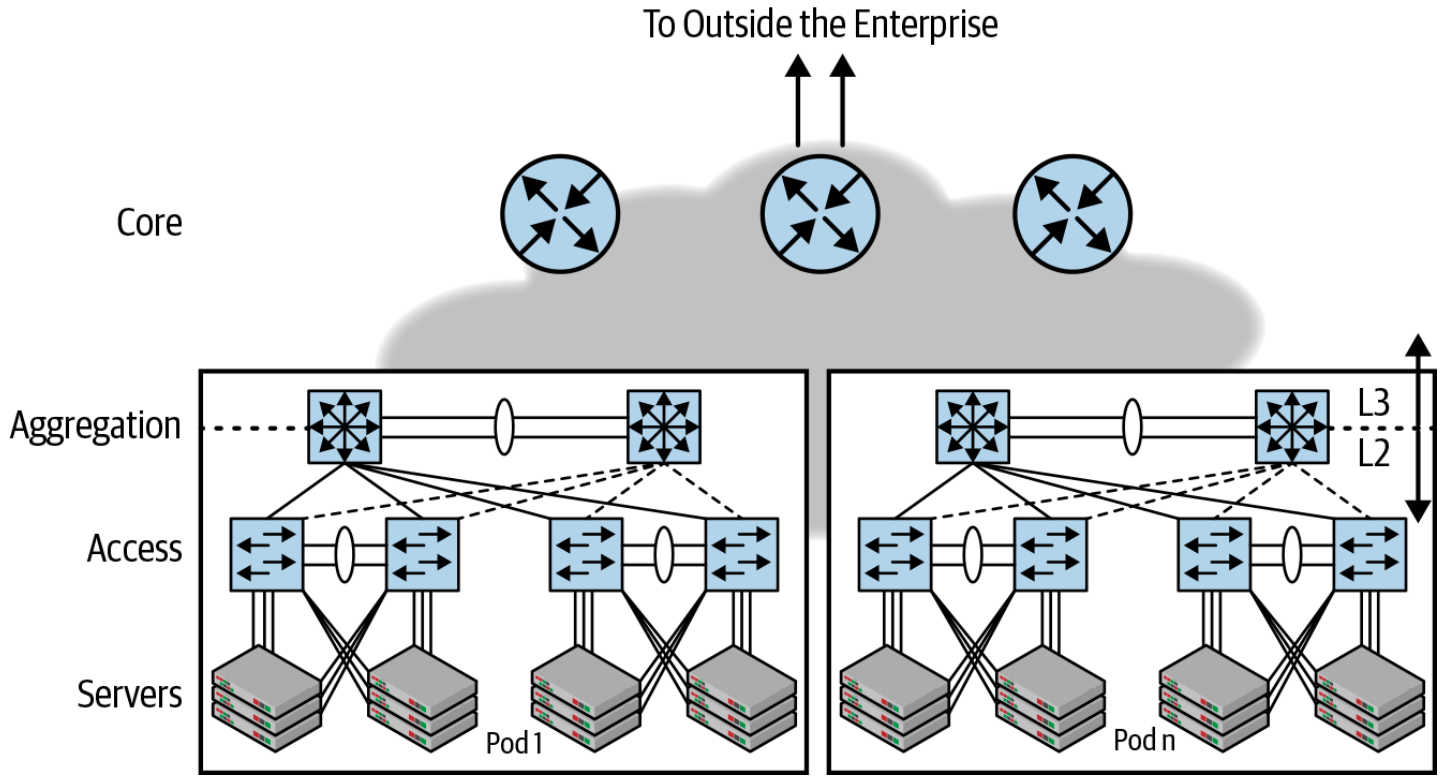
示例：VLAN 创建

- 在 Cisco NX-OS 网络操作系统中创建 VLAN 和配置访问端口
  - 创建名为“VLAN-Example”的 VLAN 101，
  - 将以太网接口（来自模块 3 的端口 11）配置为一个桥接接口，一个访问端口，将其加入 VLAN 101 中，并启用

```
! Creating VLAN 101
N7K-Switch1(config)# vlan 101
N7K-Switch1(config-vlan)# name VLAN-Example
! Including port 11 from module 3 in VLAN 101
N7K-Switch1(config-vlan)# interface Ethernet 3/11
N7K-Switch1(config-if)# switchport
N7K-Switch1(config-if)# switchport mode access
N7K-Switch1(config-if)# switchport access vlan 101
! Enabling the interface
N7K-Switch1(config-if)# no shutdown
```

VLAN 局限

- 在传统的 Access-Agg-Core 网络设计中，VLAN 在聚合交换机处终止
- 两个聚合交换机不能存在相同的 VLAN



VLAN 局限

- VLAN ID 为 12 位长，导致网络中最多 4,096 个单独的 VLAN
  - 在多租户数据中心中，4096 个 VLAN 是不够的
- 租户需要控制其 MAC，VLAN 和 IP 地址分配
  - 重叠的 MAC，VLAN 和 IP 地址
- STP 生成树的交换机数量众多，效率低下
  - 禁用了太多链接
  - 通过 IP equal cost multipath（ECMP）能够获得更好的吞吐量

桥接网络的配置难题

- 桥接网络需要很多协议。包括 STP 及其变体，FHRP，链路故障检测以及特定于供应商的协议，例如 VLAN 中继协议（VTP）
- 所有这些协议都大大增加了桥接解决方案的复杂性。这意味着当网络出现故障时，必须检查多个不同的运动部件以识别故障原因

VLAN 配置的难题

- 在云中，租户来来往往非常快。因此，快速配置虚拟网络至关重要。
  - VLAN 要求网络中的每个节点都配置有 VLAN 信息才能正常运行。但是添加 VLAN 也会增加控制平面上的负载。这是因为使用 PVST，要发送的 STP hello 数据包数等于端口数乘以 VLAN 数
- VLAN 要求路径中的每个节点都能够识别 VLAN。如果配置失败导致传输设备无法识别 VLAN，则网络将变得分区，从而导致复杂且难以固定的问题

VLAN 配置的难题

- 单个不堪重负的控制平面可以轻松关闭整个网络。因此，添加和删除 VLAN 是一个手动，费力的过程，通常需要几天的时间。
- 添加新节点也需要仔细计划。添加新节点会导致该节点需要生成的 STP 数据包数量发生变化，从而有可能使其超出其扩展限制的边缘。
- 因此，即使设置一个新节点也可能是一个漫长的过程，涉及许多人退出，每个人都伸出手指并希望一切顺利

VLAN 小结

- 虚拟的网络类型
  - L2
- Inline
  - 每个交换机都设置

解决办法

基于路由的更简洁的方案

内容

- 背景
- 网络的分区扩展和虚拟化
- VLAN 技术
- VXLAN 技术

VXLAN

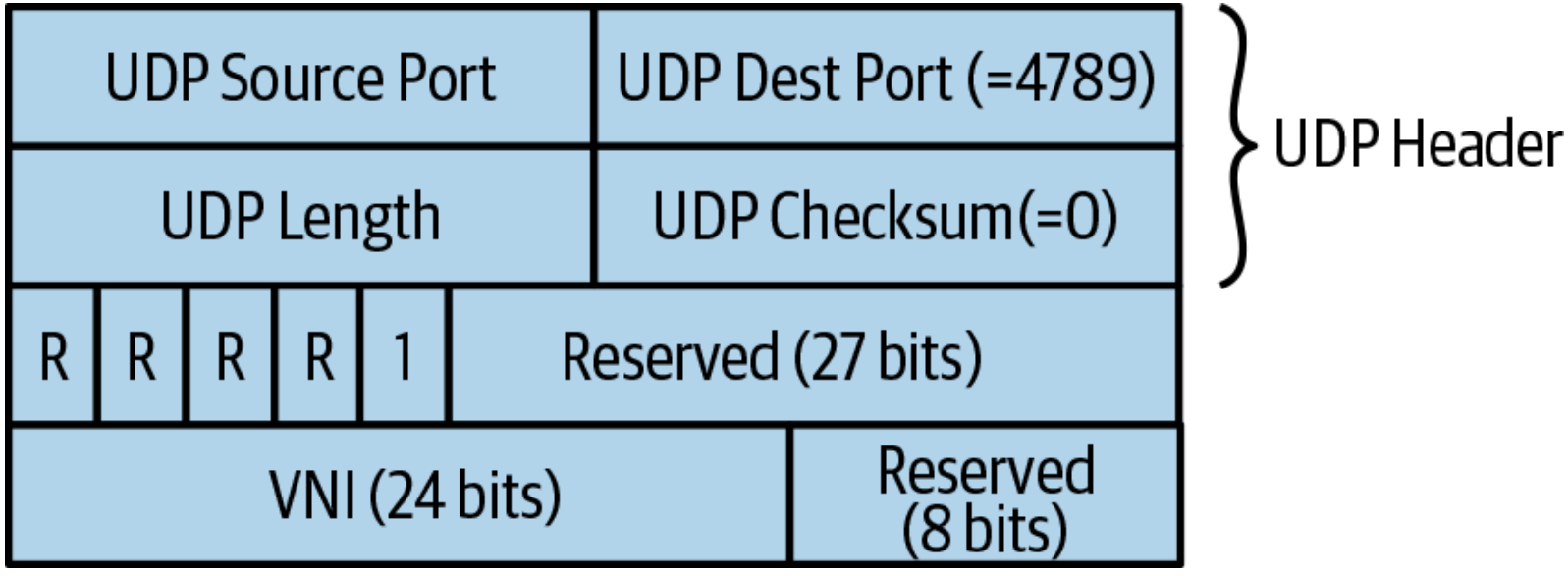
- Overlay 方式
  - 基于隧道
  - vSwitch 用作 VTEP（VXLAN 隧道端点）
- Ethernet over UDP over IP
  - 将 L2 帧封装在基于 IP 的 UDP 中，然后发送到目标 VTEP。
  - L2 using L3

VXLAN

- 隔离多个 tenant
  - 仅同一 VXLAN 中的 VM 可以通信
  - 解决了云环境中多个租户的 MAC 地址，VLAN 和 IP 地址重叠的问题
- 无需更改 VM
  - Hypervisors 管理程序负责所有细节

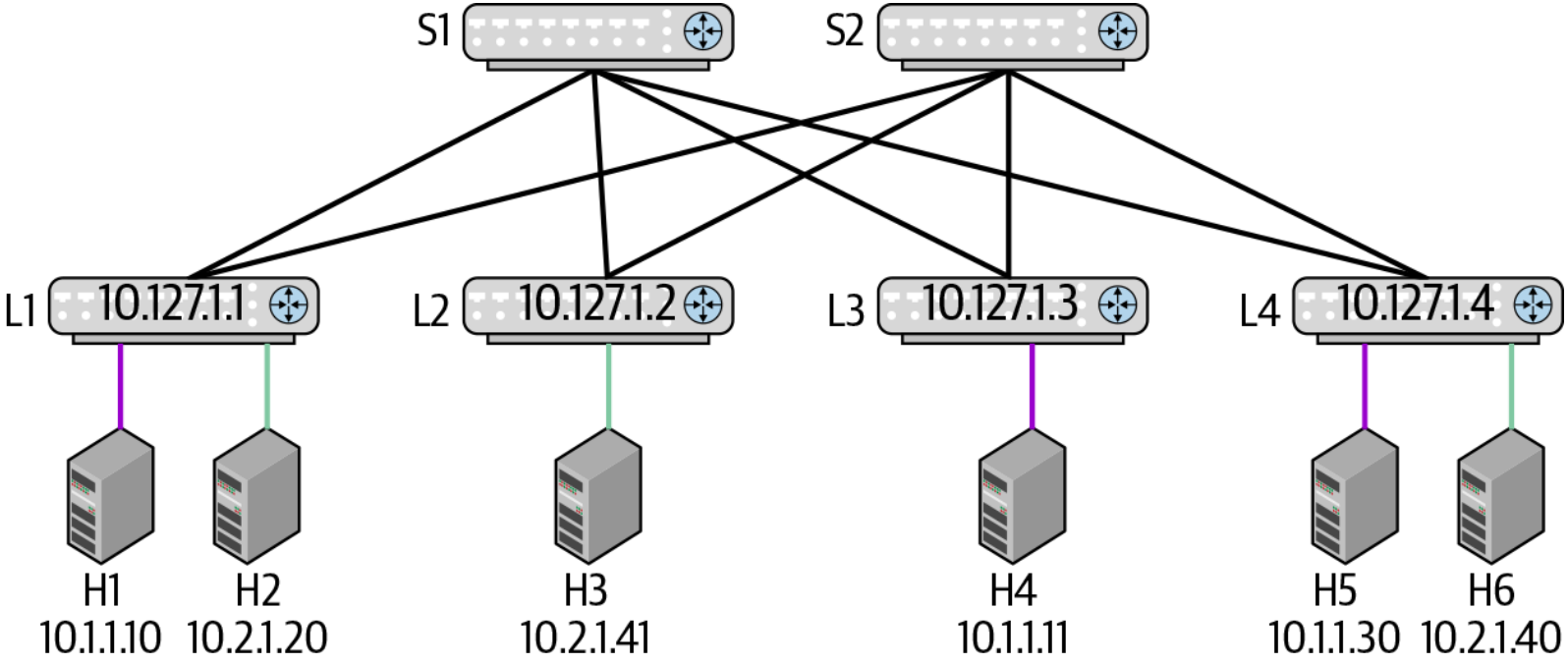
例：VXLAN

- 数据包头

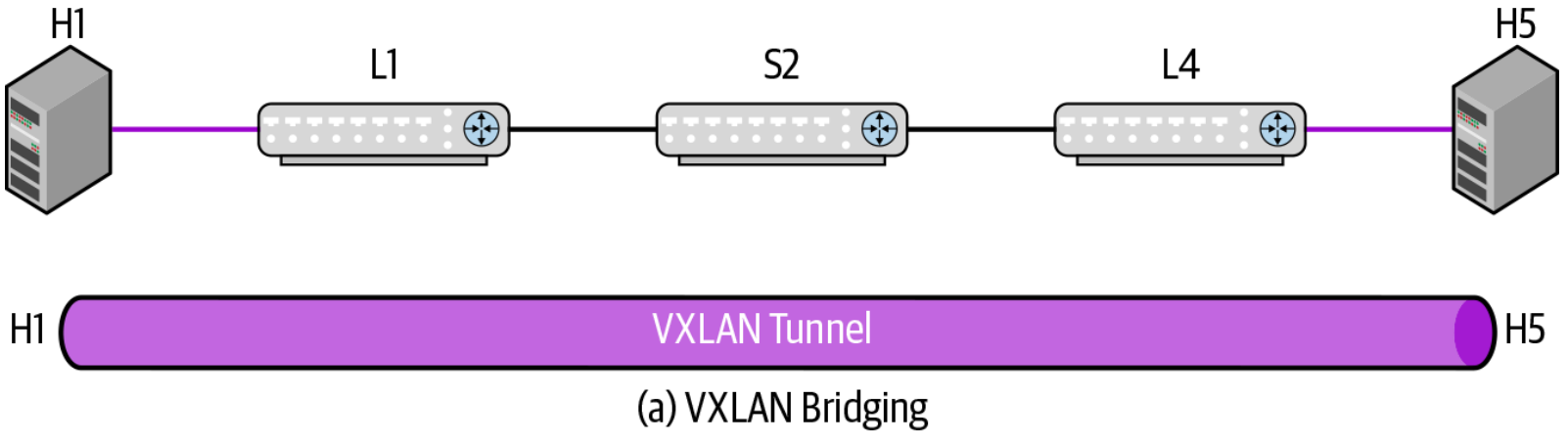

UDP Source Port
UDP Dest Port (=4789)
UDP Length
UDP Checksum(=0)
R
R
R
R
1
Reserved (27 bits)
VNI (24 bits)
Reserved (8 bits)
UDP Header

例：VXLAN

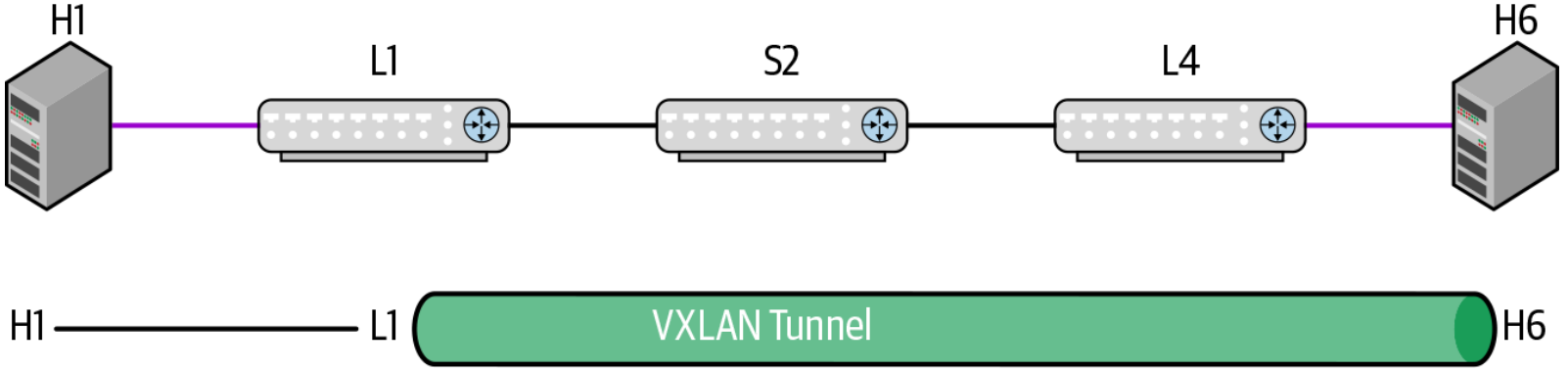
- 紫线、绿线，两个虚拟网络
- 讨论两个例子
  - 本网：H1 -> H5
  - 跨网：H1 -> H6



例：VXLAN



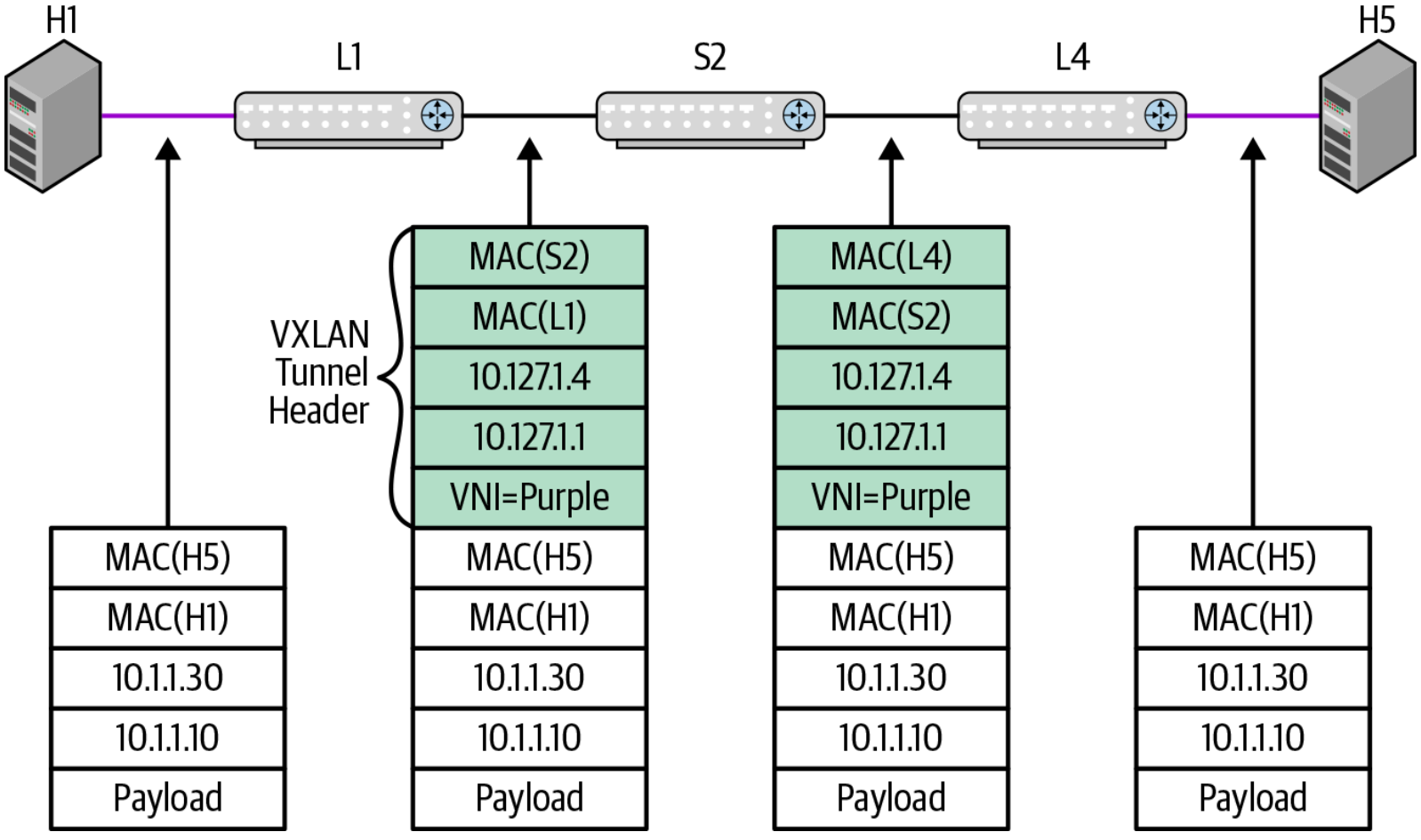
(a) VXLAN Bridging



(b) VXLAN Aymmetric Routing

(c) VXLAN Symmetric Routing

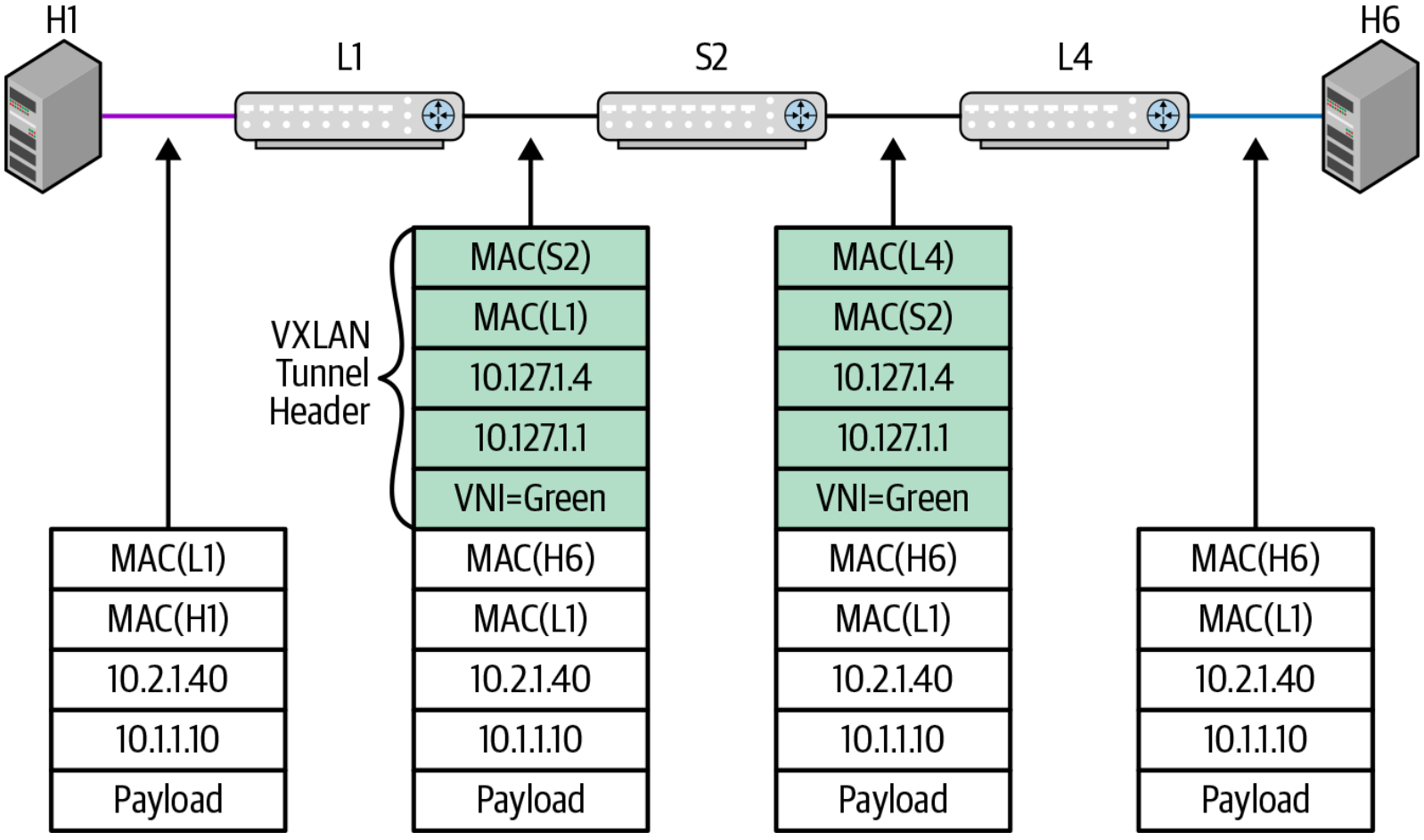
本网：VXLAN Bridge



- H1 发出的包一点也没有改变

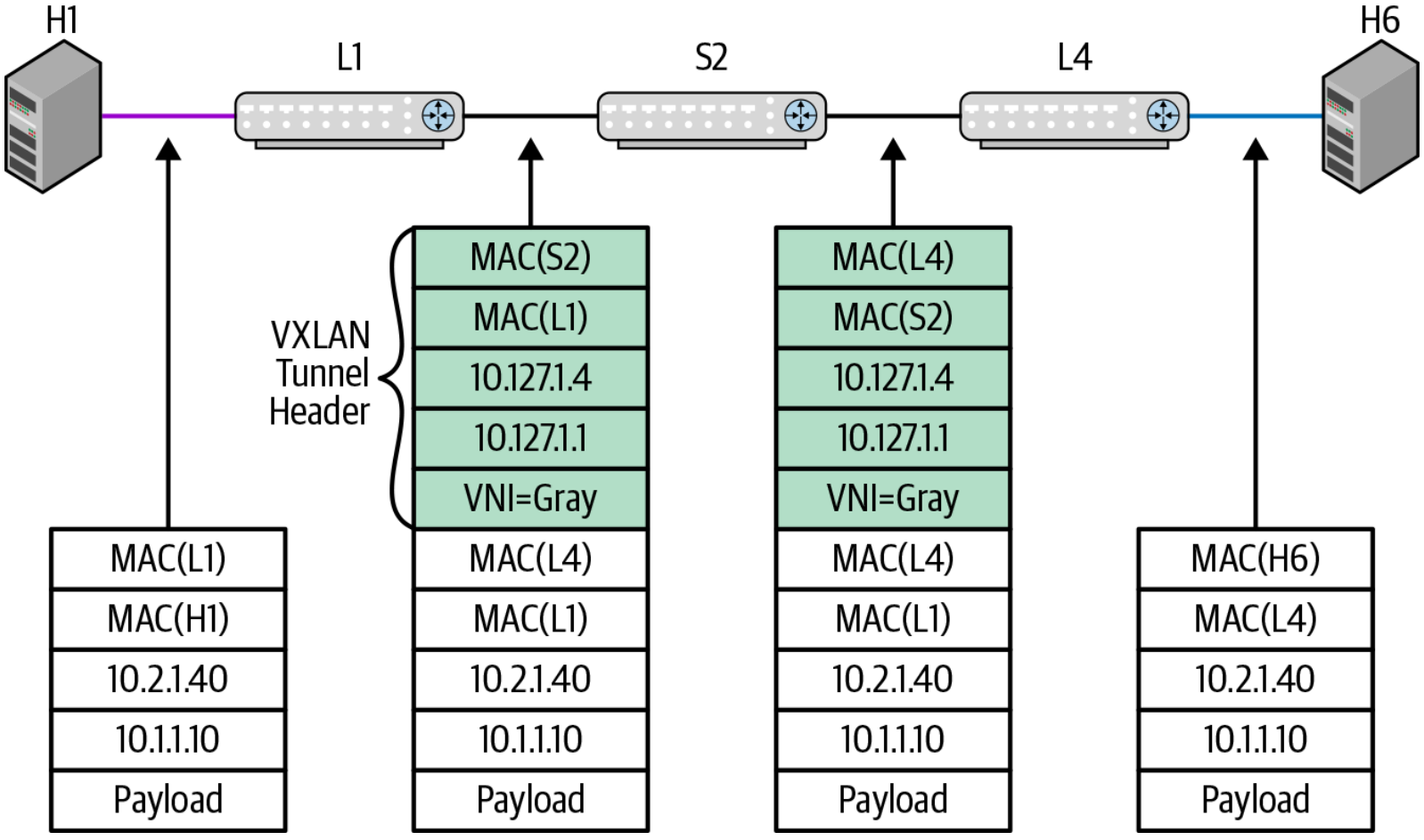
跨网：VXLAN routing

- 不对称
- L1 找到 H6 的 MAC，进入绿色的 VN



跨网：VXLAN routing

- 对称
- L1 找到 L4 的 MAC，用一个新的 VN



失败

- L2 基于 STP，特别容易失败，会导致 loop
- L3 可以应对，把节点移除
- 建议直接在 L3 上
- AWS 就拒绝 L2 和组播

小结

- 背景
- 网络的分区扩展和虚拟化
- VLAN 技术
- VXLAN 技术

练习

- 调研云计算平台支持的虚拟网络技术