主要内容

- 数据中心网络 (DCN)
  - 二层网络技术复习（交换，VLAN，STP）
  - 体系结构
  - 通信协议
    - 大二层技术：VXLAN （选）
    - 分段路由：Segment Routing（选）
- 论文讨论

下次课准备

- 抽查中期报告

数据中心网络

数据中心网络的需求

数据中心网络 Data Center Networking （DCN）

- 主机数量庞大的集群
- 多种应用服务
  - 电子商务 e-business (e.g. Amazon)
  - 内容服务器 content-servers (e.g., YouTube, Akamai, Apple, Microsoft)
  - 搜索引擎，数据挖掘 search engines, data mining (e.g., Google)
- 冗余性
- 容错性
- 高可用性
- 安全性……

数据中心网络的应用特点

- 服务器：1万~10万台服务器
- 存储：PB（Pbetabytes）级数据存储
- 负载：单一“应用”会跨越大量服务器 (e.g., Amazon.com)
  - 各种应用组件：caches, web servers, data bases, distributed file servers, …
  - 可扩展：Each component is “scaled” to meet needs of millions of users

Computers + Net + Storage + Power + Cooling

硬件环境

- 服务器集群，存储设备
  - 前端Web服务器，数据库等后端服务器。保证冗余性，容错性，高可用性
- 网络设备
  - 交换机、路由器；防火墙、安全网关；上网行为管理系统；网络负载均衡器等；冗余链路
- 电力供应系统
- 空调、冷却系统
- 气体灭火系统
- 其他防护措施

Data Centers with 100,000+ Servers

Google Oregon Datacenter

数据中心网络的特点

You Build It & You Control It

- 单独的管理域
- 极小的RTT（Round Trip Time）时间
- 大量的多路径
- 充足和一致的带宽
- 仅需要小规模的缓存
- 延迟/tail latency和带宽同等重要
- 同质化，协议修改更容易
- 从客户端-服务器模式到超大规模，大规模的并行计算
- 在一栋建筑中汇聚所有互联网的带宽

拓扑结构：模块化

- Containers

- Racks
  - Multiple servers
  - Top-of-rack switches

  架顶式交换机TOR

Internet vs. DCN

- 传播时延
  - Internet : 10-100 ms
  - DCN: 0.1 ms
- 传输时延
  - packet size 1 KB, link capacity 1 Gbps
    - → packet transmission time is 0.01 ms

比较数据中心网络与Internet

The Internet	Data Center Network (DCN)
Many autonomous systems (ASes)	One administrative domain
Distributed control/routing	Centralized control and route selection
Single shortest-path routing	Many paths from source to destination
Hard to measure	Easy to measure, but lots of data…
Standardized transport (TCP and UDP)	Many transports (DCTCP, pFabric, …)
Innovation requires consensus (IETF)	Single company can innovate
"Network of networks"	"Backplane of giant supercomputer"

数据中心的开销

Amortized Cost*	Component	Sub-Components
~45%	Servers	CPU, memory, disk
~25%	Power infrastructure	UPS, cooling, power distribution
~15%	Power draw	Electrical utility costs
~15%	Network	Switches, links, transit

The Cost of a Cloud: Research Problems in Data Center Networks. Sigcomm CCR 2009. Greenberg, Hamilton, Maltz, Patel.

*3 yr amortization for servers, 15 yr for infrastructure. 5% cost of money

交换机：Commodity Switches

- Low-cost switches
  - Especially for top-of-rack switches
- Simple memory architecture
  - Small packet-buffer space
  - Shared buffer over all input ports
  - Simple drop-tail queues

复习

- 以太网交换机的工作原理
- 生成树协议 STP
- 虚拟局域网 VLAN

复习：交换机的自学习（Self Learning）

- 以太网交换机
- 后向学习法（Backward learning ）选择输出端口
  - 关联源IP和入端口
    - Associates source address on frame with Input port
  - 已知端口：转发
    - Frame with destination address sent to learned port
  - 未知端口：洪泛
    - Unlearned destinations are sent to all other ports
    - 洪泛算法 flooding algorithm
- 不需要配置：软状态

交换机的过滤和转发机制

当交换机收到一帧：

index switch table using MAC dest address

if entry found for destination

then {

if dest on segment from which frame arrived

then drop the frame （过滤）

else forward the frame on interface indicated （转发）

}

else flood

forward on all but the interface on which the frame arrived

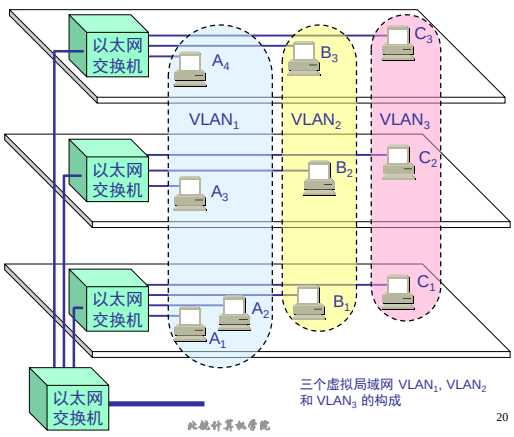
复习：虚拟局域网（Virtual LANs）

- 物理位置的限制
  - 早期的粗缆网络部署在建筑物中，计算机直接接入
  - 物理位置临近的人位于一个LAN中
  - 不依赖于组织机构
- 组网模式
  - 星型布线：基于Hubs和switches的
  - 一个交换机上可以连接多个LAN hub
  - LAN管理的灵活性
- 基于组织结构管理网络

Group users based on organizational structure, rather than the physical layout of the building.

为什么需要按组织结构管理?

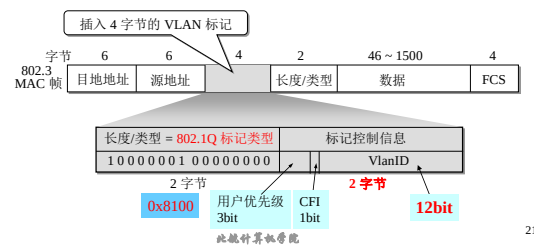
- 安全性（Security）：隔离流量
  - 以太网：共享传输介质
  - 网卡的混杂模式（“promiscuous” mode）
- 负载（Load）：限制本地流量
  - 负载不均衡
    - 不同应用，不同用户
  - 通信的本地性特点
    - E.g., traffic between people in the same research group
- 限制广播流量
- 用户移动和角色变化



三个虚拟局域网 $VLAN_1$, $VLAN_2$ 和 $VLAN_3$ 的构成

Virtual LANs – IEEE 802.1Q

- IEEE 802.1Q：允许在以太网的帧格式中插入一个 4 字节的标识符，称为 VLAN 标记(tag)，用来指明发送该帧的工作站属于哪一个虚拟局域网。
  - 帧最大长度：1522B

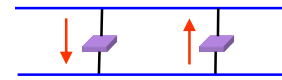


VLANs特点

- LAN和VLAN的主要区别：
  - VLAN工作在ISO模型的第2层
  - 不同VLAN之间的连接靠第3层交换
  - VLAN提供控制网络广播的方法
  - 网络管理员可以把用户划入VLAN
  - VLAN可以通过隔离通信域而提高网络的安全性
- 通过使用VLAN技术，可以按交换机端口以及所连接的用户进行逻辑分组
- 可以把一个交换机或多个相连的交换机的端口和用户划分为不同的组，通过这种方式，VLAN可以跨越一栋建筑、多个互连的建筑，甚至城域网
- 支持VLAN的交换机：透明性？
- 类似于面向连接的机制
  - VLAN标识符
  - VLAN数量只有4096个，无法满足大规模云计算IDC的需求

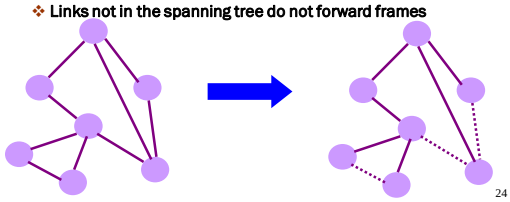
扩散（flooding）可能导致回路

- 交换过程有时需要广播帧：
  - unfamiliar destination or broadcast destination
- 广播的实现方式：扩散flooding
  - Transmitting frame out every interface
  - … except the one where the frame arrived
- 回路的产生
  - E.g., if the network contains a cycle of switches
  - Either accidentally, or by design for higher reliability



生成树（Spanning Trees）

- 确保无环的拓扑结构
  - Avoid using some of the links when flooding
  - … to avoid forming a loop
- 生成树 Spanning tree
  - Sub-graph that covers all vertices but contains no cycles
  - Links not in the spanning tree do not forward frames



生成树算法

- Radia Perlman（DEC公司）
  - 1983年，开发生成树协议，IEEE 802.1 D 规范
  - IS-IS路由协议的主要发明者
- 生成树协议
  - A protocol used by a set of bridges to agree upon a spanning tree for a particular extended LAN（或VLAN）
- 每个网桥选择其转发帧的端口
  - 阻塞某些端口，构造一个无环树
  - 可能阻塞某个网桥的所有端口

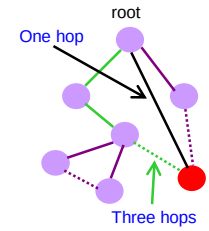
生成树算法（续）

- 互连的LAN抽象图结构
  - 节点：LAN，Bridges
  - 边：连接LAN's到bridges的端口
  - 简化：只考虑对应于网络的节点
- 生成树目标
  - 连接所有 LAN's
  - 无环树
  - 某些节点可能不转发帧
  - 支持动态重新形成新的生成树

26

如何构造生成树

- 分布式算法
  - Switches cooperate to build the spanning tree
  - ... and adapt automatically when failures occur
- 算法要点
  - 选择根节点
  - 确定到根节点的最短路径
    - Each switch identifies if its interface is on the shortest path from the root
    - And it exclude from the tree if not
  - Messages (Y, d, X)
    - 消息源：X
    - 声明根节点：Y
    - 距离：d



28

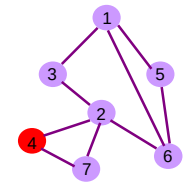
生成树算法的步骤

- 初始，每个交换机声明自己是根节点root
  - Switch sends a message out every interface
  - Example: switch X announces (X, 0, X)
- 交换机更新各自的表，确定根节点root
  - Upon receiving a message, check the root id
  - If the new id is smaller, start viewing that switch as root
- 计算到root的距离
  - Add 1 to the distance received from a neighbor
  - Identify interfaces not on a shortest path to the root
  - ... and exclude them from the spanning tree

29

例： Switch #4的视图

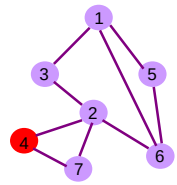
- Switch #4 声明自己是根节点 root
  - Sends (4, 0, 4) message to 2 and 7
- 接着， switch #4收到从 #2来的消息
  - Receives (2, 0, 2) message from 2
  - ... and thinks that #2 is the root
  - And realizes it is just one hop away
- 然后，switch #4又收到从 #7来的消息
  - Receives (2, 1, 7) from 7
  - And realizes this is a longer path
  - So, prefers its own one-hop path
  - And removes 4-7 link from the tree



30

例： Switch #4的视图（续）

- Switch #2 收到switch #3 的消息
  - Switch 2 hears (1, 1, 3) from 3
  - Switch 2 starts treating 1 as root
  - And sends (1, 2, 2) to neighbors
- Switch #4 收到switch #2的消息
  - Switch 4 starts treating 1 as root
  - And sends (1, 3, 4) to neighbors
- Switch #4 收到switch #7的消息
  - Switch 4 receives (1, 3, 7) from 7
  - And realizes this is a longer path
  - So, prefers its own three-hop path
  - And removes 4-7 link from the tree



生成树算法的健壮性

- 算法对链路失败的响应
  - 根节点失败
    - Need to elect a new root, with the next lowest identifier
  - 其他交换机或链路失败
    - Need to recompute the spanning tree
- 根节点周期性发送消息
  - Periodically reannouncing itself as the root (1, 0, 1)
  - Other switches continue forwarding messages
- 通过超时(soft state)检测失败
  - Switch waits to hear from others
  - Eventually times out and claims to be the root

STP的局限性

- 尽管某个网桥失败后，算法可以重新配置生成树，但对于拥塞的网桥，可能无法转发帧。
- 收敛性能
  - 如果局域网中节点过多，那么整网的收敛速度会下降
  - 一般情况下网络规模不超过100台交换机
- 降低了网络资源的带宽利用率
- 可扩展性
  - 生成树算法是线性扩展的，没有采用分层结构
  - 广播范围不能太大

数据中心（DCN）体系结构

概述

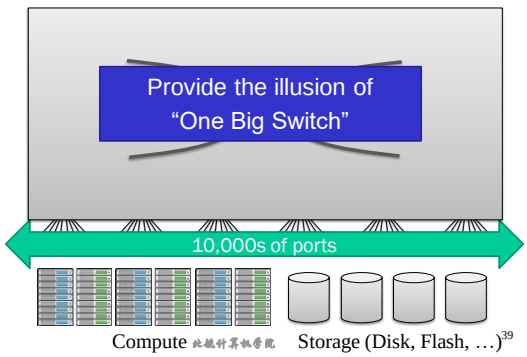
DCN的设计需求

- 针对不同类型的应用
  - 面向用户，例如： serving web pages to users
  - 内部计算，例如： MapReduce for web indexing
- 工作负载难以预测
  - 数据中心的多种服务
  - 新的服务需求
- 服务器会产生故障
  - 如MapReduce，故障处理，重新动态分配工作/任务（工作服务器）；数据复制， ...
  - 服务器间的流量矩阵“Traffic matrix”经常发生变化

灵活性Agility

- 负载管理 Workload management
  - Means for rapidly installing a service's code on a server
  - Virtual machines, disk images, containers
- 存储管理 Storage Management
  - Means for a server to access persistent data
  - Distributed filesystems (e.g., HDFS, blob stores)
- 网络Network
  - Means for communicating with other servers, regardless of where they are in the data center

数据中心网络（DataCenter Networks）



传统数据中心网络的设计

- 承载的大多是数据中心提供对外访问的业务
  - 流量分布符合80/20模型，且以南北向为主，东西向流量小
  - .COM应用模式
- 协议
  - 二层以太网交换机制
    - 生成树协议（Spanning Tree Protocol）
    - 二层转发
  - 三层路由机制
    - 动态路由协议

二层 vs. 三层?

- 以太网交换 (layer 2)
  - 交换机：Cheaper switch equipment
  - 配置：Fixed addresses and auto-configuration
  - 迁移：Seamless mobility, migration, and failover
- IP 路由 (layer 3)
  - 路由器：Scalability through hierarchical addressing
  - 路由协议：Efficiency through shortest-path routing
  - 多路径：Multipath routing through equal-cost multipath

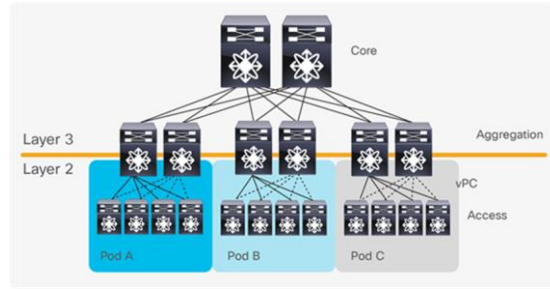
二层 vs. 三层?

Technique	Plug and play	Scalability	Small Switch State	Seamless VM Migration
Layer 2: Flat MAC Addresses	+	-	-	+
Layer 3: IP Addresses	-	+	+	-

传统数据中心网络三层架构

- Access Layer（接入层），或Edge Layer
  - 接入交换机通常位于机架顶部，也被称为ToR（Top of Rack）交换机，它们物理连接服务器
- Aggregation Layer（汇聚层），或Distribution Layer
  - 汇聚交换机则将多个接入层交换机互连在一起，所有汇聚层交换机通过核心层交换机相互连接。
  - 可提供其他的服务，例如防火墙，SSL offload，入侵检测，网络分析等
- Core Layer（核心层）
  - 核心交换机为进出数据中心的包提供高速的转发，为多个汇聚层提供连接性
  - 核心交换机为通常为整个网络提供一个弹性的L3路由网络

Traditional three-tier data center design



核心-汇聚-接入三层架构

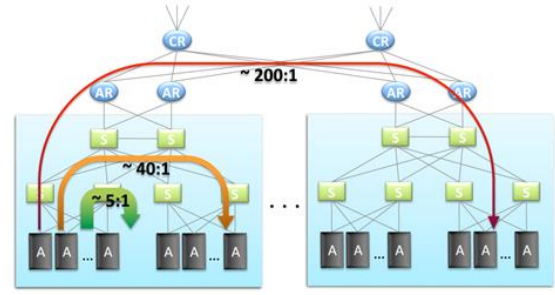
- 汇聚交换机是L2和L3网络的分界点
  - 汇聚交换机以下的是L2网络，以上是L3网络
  - 每组汇聚交换机管理一个POD（Point Of Delivery），每个POD内都是独立的VLAN网络
  - 服务器在POD内迁移不必修改IP地址和默认网关，因为一个POD对应一个L2广播域
- 汇聚交换机和接入交换机之间通常使用STP（Spanning Tree Protocol）
  - 对于一个VLAN网络只有一个汇聚层交换机可用
  - 一些私有的协议，例如Cisco的vPC（Virtual Port Channel）可以提升汇聚层交换机的利用率

传统网络架构的局限性

- 随着云计算、大数据等业务的兴起，分布式计算、分布式存储等技术开始在IDC内部大规模部署。
  - IDC内部的东西向流量急剧上升，流量的80/20模型转变成以东西向流量为主
- 传统网络架构的局限性
  - 扩展能力差：网络规模受限于核心交换机端口数量，无法平滑Scale-out(横向扩展)；
  - 收敛比过高：为南北向流量设计的流量模型，收敛模型呈三角型，越往上性能越低，东西向带宽严重不足

存在问题

- 服务器到服务器之间的带宽有限
  - 服务器到交换机之间的链路带宽通常为1Gbps，交换机之间的链路带宽通常为10Gbps
  - 若每个交换机下有50个服务器，那么服务器到交换机的总带宽为50Gbps, 远远大于交换机之间的带宽。那么服务器与交换机之间的收敛比（over-subscription）比例为5:1
  - 层次越高收敛比的情况越严重，服务器与路由器的收敛比甚至达到200:1。
- 一个子网内的服务器与另一个子网内的服务器进行通信会受到上层链路带宽的限制，未能抢占到带宽的服务器只能等待，浪费了服务器资源



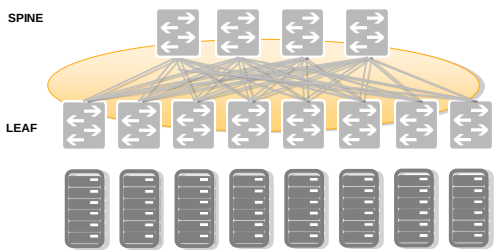
存在问题（续）

- 服务器虚拟化
- 资源分散
  - 通常，同一服务的服务器尽量部署在一个服务器集合内。为了服务的可扩展性和稳定性，就需要增加部分冗余资源，但这部分资源可能造成了资源浪费
- 不同服务之间存在影响
  - 在同一子网内的不同服务可能存在相互影响。例如一个服务发生流量泛洪时，在同一子树内的其他服务器也会受到影响。

叶脊网络架构（Leaf-Spine）

- Leaf-Spine架构主要为满足数据中心内部高速互连的需求，解决数据中心内流量的快速增长和数据中心规模的不断扩大
- Leaf-Spine结构也称为Clos结构
  - 两层的Fat-tree网络，leaf和spine之间为全网状连接（Full Mesh）
- Leaf（叶）交换机
  - leaf交换机是接入层，作为TOR（Top Of Rack）直接连接物理服务器，同时上联spine交换机。leaf交换机之上是三层网络
- Spine(脊)交换机
  - spine交换机可以当做传统三层架构中的核心交换机
  - 高端口密度的交换机
  - spine和leaf交换机之间通过ECMP（Equal Cost Multi Path）动态选择多条路径。

叶脊网络架构（Leaf-Spine）



Leaf 交换机

- Leaf 层由接入交换机组成，用于接入服务器
- 每个低层级的 Leaf 交换机都会连接到每个高层级的 Spine 交换机上，即每个 Leaf 交换机的上行链路数等于 Spine 交换机数量，同样，每个 Spine 交换机的下行链路数等于 Leaf 交换机的数量，形成一个 Full-Mesh 拓扑。
- 当 Leaf 层的接入端口和上行链路都没有瓶颈时，这个架构就实现了无阻塞（Nonblocking）。
- 一个包只需要经过一个 Spine 和另一个 Leaf 就可以到达目的端，延迟是可预测的

Spine 交换机

- Spine 层是网络的骨干（Backbone），负责将所有的 Leaf 连接起来。Spine Switch相当于核心交换机
- Spine 和 Leaf 交换机之间通过 ECMP（Equal Cost Multi Path）动态选择多条路径。
- Spine 交换机为 Leaf 交换机提供一个弹性的 L3 路由网络
- 数据中心的南北流量可以不用直接从 Spine 交换机发出，一般来说，南北流量可以从与 Leaf 交换机并行的交换机（edge switch）再接到 WAN router 出去。

寻址和路由

- Leaf-Spine架构使用定制的寻址方案和路由算法，采用第2层或第3层技术
  - 第3层：要求每个链路都被路由，通常使用开放最短路径优先（OSPF）或等价多路径路由（ECMP）来实现的边界网关协议（BGP）动态路由
  - 第2层：采用loop-free的以太网fabric技术，例如多链接透明互联（TRILL）或最短路径桥接（SPB, IEEE 802.1aq）
  - 核心网络还可使用带有ECMP的动态路由协议通过第3层连接到主干网。

Leaf-Spine 网络架构的部署

- 模块化
  - 出厂时预装配成四种类型的标准工程系统：Transit 机柜, Spine 机柜, Fabric 机柜, 和 Server 机柜。
- 公有云采用Leaf-Spine 网络架构
  - 各家公有云服务商Leaf-Spine网络中的Underlay Network和Overlay Network使用的协议和方案有很大区别
  - 例如，VXLAN+SDN解决方案

Leaf-Spine 架构优势

- 扁平化
  - 扁平化设计缩短服务器之间的通信路径，从而降低延迟，可以显著提高应用程序和服务性能。
- 易扩展
  - 如果Spine交换机的带宽不足，只需要增加Spine的节点数，也可以提供路径上的负载均衡；接入连接不足，则只需增加leaf节点数。
- 低收敛比
  - 容易实现1：X甚至是无阻塞的1：1的收敛比，而且通过增加Spine和Leaf设备间的链路带宽也可以降低链路收敛比。

Leaf-Spine 架构优势(续)

- 简化管理
  - 叶脊结构可以在无环路环境中使用全网格中的每个链路并进行负载平衡
- 边缘流量处理
  - 在物联网（IoT）业务中，可能有数千个传感器和设备在网络边缘连接并产生大量流量。Leaf可以在接入层处理连接，spine保证节点内的任意两个端口之间提供延迟非常低的无阻塞性能，从而实现从接入到云平台的敏捷服务。
- 多云管理
  - 数据中心或云之间通过spine leaf架构仍可以实现高性能、高容错等优势

局限性

- 网络交换机的数量远远大于三层网络架构。
- 扩展新的Leaf时需要大量的线缆、并占用大量Spine交换机端口。
- Spine交换机端口数量决定了最大可联接的Leaf交换机数量，也就决定了最大主机总数量。
- 叶子节点网络设备性能要求和功能要求较高
- 私有协议：各个厂商封装骨干节点与叶子节点间的转发
- 独立的 L2 Domain 限制了依赖 L2 Domain 应用程序的部署
  - 要求部署在一个二层网络的应用程序，现在只能部署下一个机架下
  - 独立的 L2 Domain 限制了服务器的迁移。迁移到不同机架之后，网关和 IP 地址都要变
  - 子网数量大大增加。每个子网对应数据中心一条路由，相当于每个机架都有一个子网，对应于整个数据中心的路由条数大大增加

数据中心网络中的Overlay

- Underlay网络
  - 数据中心网络的**物理基础层**：基础转发架构，只要数据中心网络上任意两点路由可达即可
  - 可以通过物理网络设备本身的技术改良、扩大设备数量、带宽规模等完善Underlay网络，其包含了一切现有的传统网络技术。
- Overlay网络
  - 建立在Underlay网络之上的网络
  - 用**逻辑节点和逻辑链路**构成了Overlay网络

Overlay网络

- 虚拟化技术：Overlay网络是在现有的网络（Underlay网络）基础上构建的一个虚拟网络。
  - 点到多点的隧道封装协议
  - L2 over L3 ：一个在L3之上的L2网络
  - 通过用**隧道封装**的方式，将源主机发出的原始二层报文封装后在现有网络中进行透明传输，到达目的地之后再解封装得到原始报文，转发给目标主机，从而实现主机之间的二层通信。
  - 通过封装和解封装，相当于一个大二层网络叠加在现有的基础网络之上，所以称为Overlay，也叫NVo3

数据中心内的大二层技术

- 随着云业务的运营，租户数量剧增。传统交换网络用VLAN来隔离用户和虚拟机
  - 理论上只支持最多4094个标签的VLAN，无法满足需求
- 物理二层→ 逻辑二层
  - Overlay解决方案
    - 网络设备厂商
    - 虚拟化软件厂商

数据中心内的大二层技术（续）

- 网络设备厂商，基于硬件设备开发大二层技术
  - 借鉴路由协议实现
  - IETF的TRILL (Transparent Interconnection of Lots of Links，多链接透明互联)
  - CISCO的FabricPath，与TRILL兼容
  - IEEE的SPB(Shortest Path Bridging,最短桥接路径)
    - IEEE802.1aq

大二层网络技术

- 通过网络边缘设备对流量进行封装/解封装，构造一个逻辑的二层拓扑
- 通过路由计算方式（IS-IS）进行二层报文转发
  - 在二层报文前插入额外的帧头，采用路由计算方式控制整网数据的转发
  - 在冗余链路下防止广播风暴，实现ECMP（等价链路）
  - 将二层网络的规模扩展到整个网络，不受核心交换机数量的限制。
  - 网络边缘设备必须支持相应的协议；
  - 硬件设备表项容量大、转发速度快。

数据中心内的大二层技术（续）

- 虚拟化软件厂商利用主机上的虚拟交换机（vSwitch）作为网络边缘设备，对流量进行封装/解封装。对网络硬件设备没有过多要求。
  - 虚实结合的Overlay技术
  - VXLAN（Virtual eXtensible LANs）是由VMWare和CISCO提出的Overlay技术方案，2014年8月RFC7348
  - NVGRE是由HP和微软等公司提出的标准，RFC 2784，RFC 2890
  - Nicira的STT（A Stateless Transport Tunneling Protocol for Network Virtualization）

各种Overlay技术比较

	VLAN	网络设备厂商提出的Overlay			虚拟化软件厂商提出的Overlay		
		EVI	Trill/SPB	VPLS	VXLAN	NVGRE	STT
网络虚拟化	O	O	O	O	O		
二层互联技术	O	O	O	O	O		
大型扩展	X	O	O	O	O		
简单控制平面	X	O	O	X	O		
自动化部署	X	O	O	X	O		
无需设备升级	O	X	X	X	O		
vSwitch支持	O	X	X	X	O		
厂商支持情况	-	硬件厂商			虚拟化及硬件厂商		

	VXLAN	NVGRE	STT
方案简述	L2 over UDP	L2 over GRE	无状态TCP
网络虚拟化方式	VXLAN报文 24bit VNI	NVGRE报文 24bit VSI	STT报文 64bit context ID
数据新增报文长度	50 Byte	42 Byte	58~76 Byte
技术特点	不改变L2~L4报文结构，现有网络设备即可支持多路径负载均衡	改变了GRE报文头，需要升级网络设备才能支持多路径负载均衡	改变了TCP报文头且无商用芯片支持，仅VMvare纯虚拟化环境可用
支持厂商	硬件厂商、VMware、HP、Citrix、RedHat、Broadcom	Microsoft、HP、Broadcom、Dell、Emulex、Intel	Nicira（VMware）

VXLAN

VxLAN的需求

虚拟扩展局域网（VxLAN，Virtual extensible Local ）

- **服务器的虚拟化极大的增加了数据中心的计算密度**
- **为了实现业务的灵活变更部署，虚拟机在二层网络中迁移需求越来越迫切**
- **传统网络无法满足：容量、灵活性及扩展性**

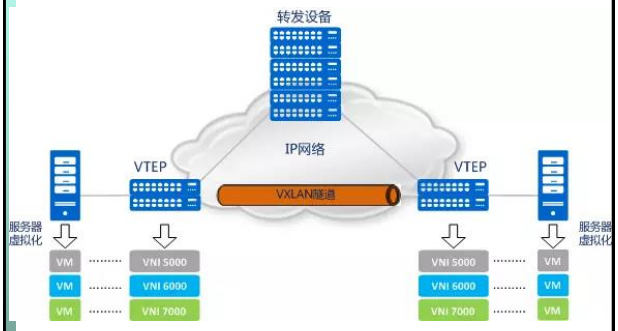
VXLAN： RFC7348

- RFC7348定义了VLAN扩展方案VXLAN（Virtual eXtensible Local Area Network，虚拟扩展局域网）
- VXLAN采用MAC-in-UDP（User Datagram Protocol）封装方式，是IETF定义的NVO3（Network Virtualization over Layer 3）中的一种网络虚拟化技术。
  - 将以太网报文封装在UDP协议的一种隧道转发模式（MAC-in-UDP），目的UDP端口号为4798
  - 实现二层网络在三层范围内进行扩展，满足数据中心大二层虚拟机迁移的需求。
  - 在VXLAN网络中，属于相同VXLAN的虚拟机处于同一个逻辑二层网络，彼此之间二层互通；属于不同VXLAN的虚拟机之间二层隔离。

VXLAN交换机

- 虚拟交换机
  - VXLAN最初只在虚拟交换机中实现
  - 局限性：虚拟交换机转发性能较低，不适合大流量的网络环境。
- 硬件交换机
  - 各硬件厂商推出支持VXLAN的硬件产品，与虚拟交换机一起，共同成为网络边缘设备
- 最终使VXLAN技术能够适应各种网络

VXLAN网络模型



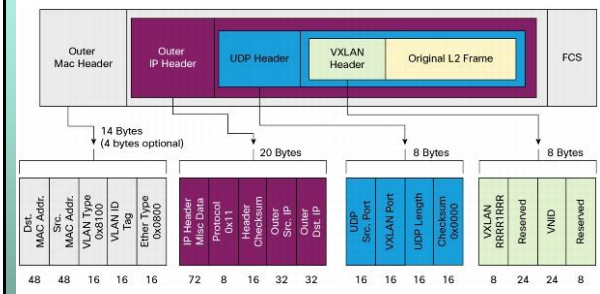
VXLAN网络模型

- VTEP（VXLAN Tunnel Endpoints，VXLAN隧道端点）
  - VXLAN网络的边缘设备，是VXLAN隧道的起点和终点，负责处理VXLAN报文(可以是独立网络设备或虚拟机服务器)
- VNI（VXLAN Network Identifier，VXLAN 网络标识符）
  - VNI是租户标识（类似VLAN ID），属于不同VNI的虚拟机之间不能直接进行二层通信。
  - 采用24比特标识二层网络分段，支持大规模租户隔离（16M个标签）
- VXLAN Tunnel隧道
  - 建立在两个VTEP之间的一条虚拟通道，传输经过VXLAN封装的报文。
  - VTEP为数据帧封装VXLAN头、UDP头、IP头后，通过VXLAN隧道将封装后的报文转发给远端VTEP，远端VTEP对其进行解封装。

VXLAN网络模型（续）

- VSI（Virtual Switching Instance，虚拟交换实例）
  - VTEP上为一个VXLAN提供二层交换服务的虚拟交换实例
  - VSI可以看作是VTEP上的一台基于VXLAN进行二层转发的虚拟交换机，它具有传统以太网交换机的所有功能，包括源MAC地址学习、MAC地址老化、泛洪等。VSI与VXLAN一一对应。
- VSI-Interface（VSI的虚拟三层接口）
  - 类似于Vlan-Interface，用来处理跨VNI即跨VXLAN的流量
  - VSI-Interface与VSI一一对应，在没有跨VNI流量时可以没有VSI-Interface

VXLAN报文格式



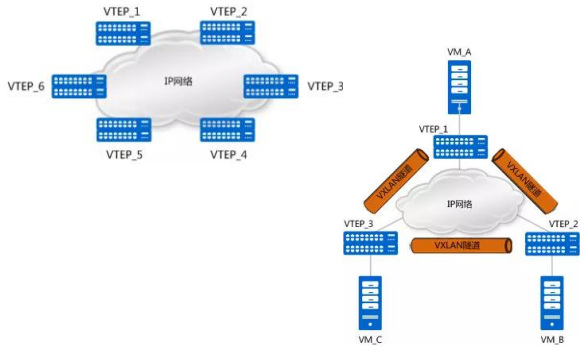
VXLAN报文头部字段

- VXLAN Header
  - 增加VXLAN头（8字节），其中包含24比特的VNI字段，用来定义VXLAN网络中不同的租户。此外，还包含VXLAN Flags（8比特，取值为00001000）和两个保留字段（分别为24比特和8比特）。
- UDP Header
  - VXLAN头和原始以太帧一起作为UDP的数据。UDP头中，目的端口号（VXLAN Port）固定为4789，源端口号（UDP Src. Port）是原始以太帧通过哈希算法计算后的值。

VXLAN报文头部字段（续）

- Outer IP Header
  - 封装外层IP头。其中，源IP地址（Outer Src. IP）为源VM所属VTEP的IP地址，目的IP地址（Outer Dst. IP）为目的VM所属VTEP的IP地址。
- Outer MAC Header
  - 封装外层以太头部。其中，源MAC地址（Src. MAC Addr.）为源VM所属VTEP的MAC地址，目的MAC地址（Dst. MAC Addr.）为到达目的VTEP的路径上下一跳设备的MAC地址。（隧道出入口地址）

建立VXLAN隧道



建立VXLAN隧道

- **大二层网络中虚拟机VM之间的通信**
  - **同一大二层域内的VTEP之间都需要建立VXLAN隧道。通过VXLAN隧道，“二层域”可以突破物理上的界限**
- **例如**
  - **VTEP_1连接的VM、VTEP_2连接的VM以及VTEP_3连接的VM之间需要“大二层”互通，那VTEP_1、VTEP_2和VTEP_3之间就需要两两建立VXLAN隧道**

Bridge-Domain

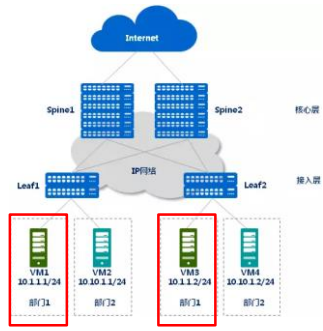
- VXLAN网络中，“大二层域”类似传统网络中VLAN（虚拟局域网）的概念，称为Bridge-Domain（BD）
  - 通过VNI来区分的不同的BD
  - VTEP生成BD与VNI的映射关系表
  - 进入VTEP的报文就可以根据自己所属的BD来确定报文封装时该添加哪个VNI

建立VXLAN隧道

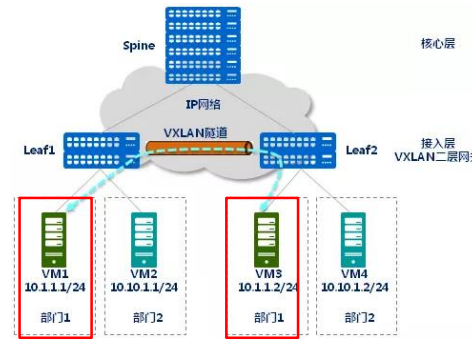
- **手工方式**
  - **在本端VTEP和对端VTEP之间建立静态VXLAN隧道**
  - **用户手动指定VXLAN隧道的源和目的IP地址分别为本端和对端VTEP的IP地址**
- **自动方式**
  - **需要借助其他协议如BGP建立VXLAN隧道**
  - **主要应用在EVN（Ethernet Virtual Network）和VXLAN的分布式网关场景中**

VXLAN应用部署方式

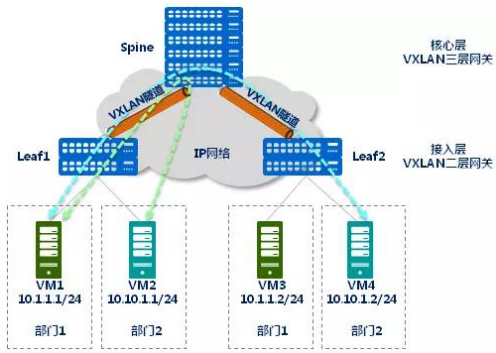
- **典型的“Spine-Leaf”数据中心组网**
  - 企业用户拥有多个部门（部门1和部门2），每个部门中拥有多个VM（VM1和VM3，VM2和VM4）。
  - 同部门的VM属于同一个网段，不同部门的VM属于不同的网段。
  - 用户希望同一部门VM之间、不同部门VM之间，VM与Internet之间均可相互访问。



相同子网互通



不同子网互通（集中式网关）



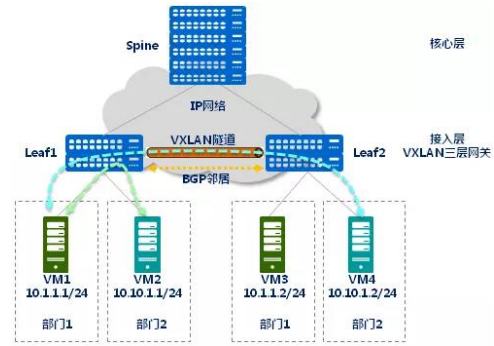
集中式网关的问题

- **Spine链路带宽限制**
  - 在不同子网互通（集中式网关）中，同一Leaf（Leaf1）下挂的不同网段VM（VM1和VM2）之间的通信，都需要在Spine上进行绕行，这样就导致Leaf与Spine之间的链路上，存在冗余的报文，额外占用了大量的带宽。
- **Spine的表项限制**
  - Spine作为VXLAN三层网关时，所有通过三层转发的终端租户的表项都需要在该设备上生成。但是，Spine的表项规格有限，当终端租户的数量越来越多时，容易成为网络瓶颈。

不同子网互通（分布式网关）

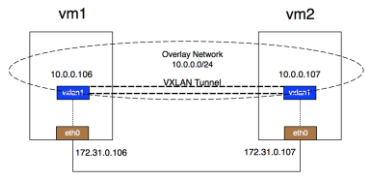
- 同Leaf节点下不同部门VM之间的通信
  - 在Leaf节点上部署VXLAN三层网关，即可实现同Leaf下不同部门VM之间的相互通信，不再需要经过Spine节点
- 跨Leaf节点不同部门VM之间的通信
  - 在不同Leaf节点上部署VXLAN三层网关。
  - 两个VXLAN三层网关之间通过BGP动态建立VXLAN隧道（BGP EVPN），并通过BGP的remote-nexthop属性发布本网关下挂的主机路由信息给其他BGP邻居，实现跨Leaf节点不同部门VM之间的相互通信

不同子网互通（分布式网关）



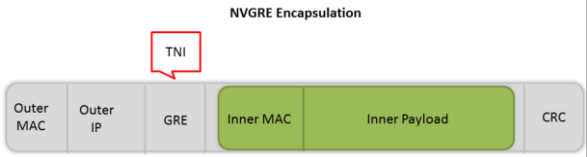
课后练习（选）

- 在Linux中建立VXLAN隧道
- 基于Open vSwitch建立VXLAN隧道



NVGRE

- NVGRE是将以太网报文封装在GRE内的一种隧道转发模式
- 采用24比特标识二层网络分段，称为VSI(Virtual Subnet Identifier)，类似于VLAN ID作用；
- 为了使NVGRE利用承载网络路由的均衡性，NVGRE在GRE扩展字段flow ID，这就要求物理网络能够识别到GRE隧道的扩展信息，并以flow ID进行流量分担；
- 未知目的、广播、组播等网络流量均被封装为组播转发。

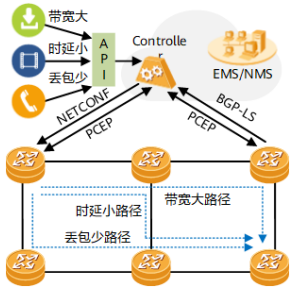


分段路由 Segment Routing

（选）

业务驱动网络

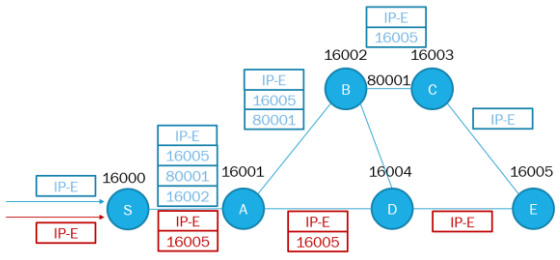
- 由业务来定义网络的架构
  - 由应用提出需求（时延、带宽、丢包率等），控制器收集网络拓扑、带宽利用率、时延等信息，根据业务需求计算显式路径。



基本概念

- 分段路由 Segment Routing（SR）
  - 源路由：节点（路由器、主机或设备）选择路径，并且引导数据包沿着该路径通过网络
  - 在数据报头中插入带顺序的段列表（segment list），以指示收到这些数据包的节点怎么去转发和处理这些数据包
- 更简单的控制平面
  - 在MPLS网络中，不再需要部署复杂的LDP/RSVP-TE协议
  - 通过IGP路由协议（ISIS/OSPF）或BGP对SR的扩展来实现标签分发和同步
- 易扩展的数据平面
  - 复用已有的MPLS和IPv6转发平面
- SR-MPLS和SRv6

例：SR转发过程



例：SR转发过程（续）

- 在头端节点控制流量要转发的路径，实现基于源的路由
- 通过封装三层标签规划路径
- 例
  - 管理员想采用一条经由节点B和节点C到节点E的路径，此时头节点S把节点B的标签16002封装到最外层，为了确保流量从节点B出来后去往节点C，把节点B的到节点C的链路标签80001封装到第二层，最后把16005封装到最内层。

Segment

- 标签Segment，标签栈Segment列表
  - 前缀（Prefix） Segment
  - 邻接（Adjacency）Segment
- 路由协议把这些标签通告到整个网络，网络中每个节点都知道区域中所有的Segment
- 控制数据的中间转发路径

前缀（Prefix） Segment

- 路由协议为前缀Segment分配标签Prefix-SID
  - 与IP地址相关联，也是全局唯一的。
- 节点（Node）Segment
  - 通常是节点的环回接口的主机前缀，这个环回接口通常也作为路由器ID
- Anycast Segment
  - 多个成员共用一个Anycast-SID，之间可以负载均衡并互相提供故障保护。

邻接（Adjacency）Segment

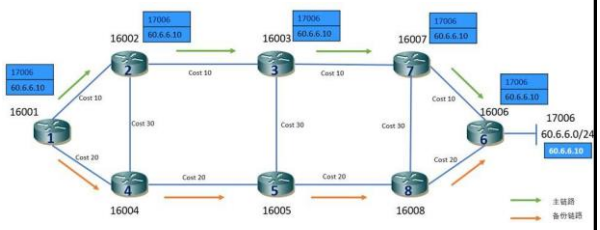
- Adjacency Segment或Adjacent-SID
  - 与某个邻居关联的Segment，标识一条出链路，引导流量从相关联的链路转发出去
- SR既可以控制流量从某个指定的节点转发，也可以控制从指定邻居的哪条链路转发
- Adjacency-SID是通告它的节点的本地Segment，而Prefix-SID是全局的
  - 节点会向网络中通告Adjacent-SID，但只有这个节点会把Adjacency-SID安装到转发表里，其他节点仅仅是知道这个Adjacency-SID，并不为它安装转发表

Segment 转发

- 标签转发
  - 动作：压入（PUSH）、继续（CONTINUE）、下一个（NEXT）分别对应MPLS转发的压入（PUSH）、交换（SWAP）、弹出（POP），使用传统MPLS的报文头
  - 头端节点判断一个IP前缀如果有SR出标签则执行压入动作，中间节点执行继续动作，也是根据标签查SR转发表，封装指定标签，多数时候出入标签值相同
- SR节点通过控制平面知道它自己是倒数第二跳
  - 弹出：SR节点能够通过控制平面知道它自己是倒数第二跳，此时如果此前缀通告的关闭倒数第二跳弹出标志没有置位，则它弹出这个前缀的Prefix-SID
  - 正常带标签转发
  - 更换显示空标签

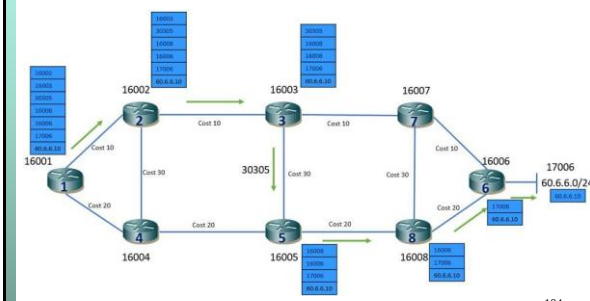
例子

- ST-BE （Best Effort）：最短路径转发
- 从节点 1 到前缀为 60.6.6.0/24 的网段（Prefix-SID 17006）。所有节点通过 IGP 学习到 Prefix-SID 17006，之后 IGP 通过 SFP 算法得到一条去往 60.6.6.10 的最短路径



说明

- 严格显示路径 SR-TE （Traffic Engineering）：基于 Adj-SID，头节点指定严格显示路径（Strict Explicit）



作业：提交大作业中期报告

小组提交大作业中期报告：

每组提交一个word文档。主要内容包括：

1、研究内容(如有变化请说明)

2、已完成工作（模块设计与实现，实验设计与结果分析等）

3、关键技术与难点

4、进度安排