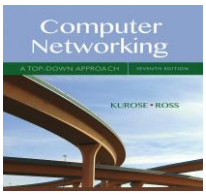
Chapter 6
链路层和局域网
The Link Layer and LANs



Computer Networking: A Top Down Approach

7th edition
Jim Kurose, Keith Ross
Pearson/Addison Wesley
April 2016

链路层和局域网：目标

- 理解链路层服务背后的原理：
  - 差错检测，纠正
  - 共享广播信道：多路访问
  - 链路层寻址
  - 局域网：Ethernet, VLANs
- 数据中心网络
- 各种链路层技术的实例化与实现

链路层, 局域网：路线图

- 概述
- 错误检测，纠正
- 多路访问协议
- 局域网
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化: MPLS
- 数据中心网络

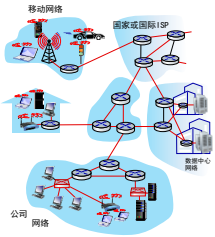
- Web页面请求的历程

链路层：概述

术语：

- 主机和路由器：节点
- 沿着通信路径连接相邻节点的通信信道：链路
  - 有线的
  - 无线的
  - LANs
- layer-2：*帧*, 封装了数据报

链路层负责将数据报从一个节点传输到链路上物理相邻的另一个节点



链路层：上下文

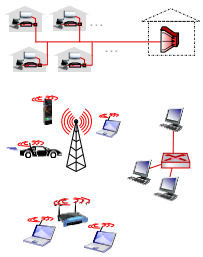
- 不同的链路用不同的链路层协议来传输数据报：
  - 例如，第一条链路是WiFi，下一条链路可以是Ethernet
- 每种链路层协议提供不同的服务
  - 例如，是否提供可靠的链路传输服务

交通运输的类比：

- 从普林斯顿到洛桑
  - 豪华大轿车：普林斯顿到JFK机场
  - 飞机：JFK机场到日内瓦
  - 火车：日内瓦到洛桑
- 游客 = 数据报
- 运输区段 = 通信链路
- 运输方式 = 链路层协议
- 旅行社 = 路由算法

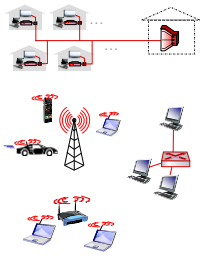
链路层：服务

- 成帧，链路接入：
  - 成帧：将数据报封装到帧中，添加首部，尾部
  - 链路接入: 媒体访问控制
  - 帧报头中的"MAC"地址标识源和目的(和IP地址不同！)
- 相邻节点之间的可靠交付
  - 我们已经知道如何完成这个任务！
  - 很少在低比特差错率的链路上使用
  - 无线链路：高差错率
    - Q: 为什么链路级和端到端都有可靠交付?



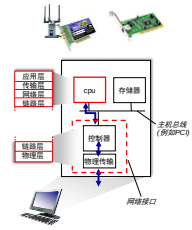
链路层：服务（其它）

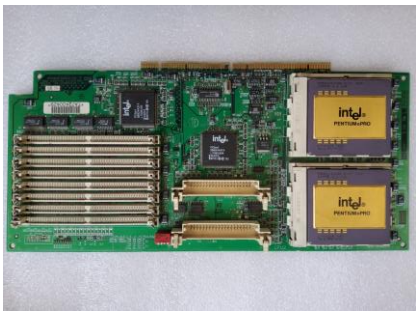
- 流量控制：
  - 相邻的发送和接收节点的步调同步
- 差错检测：
  - 信号衰减引起的差错，噪音.
  - 接收节点检测到差错，信号重传，或丢弃帧
- 纠错：
  - 接收节点识别并*纠正*比特差错而不需要重传
- 半双工和全双工：
  - 在半双工的情况下，链路两端的节点可以传输数据，但不能同时传输



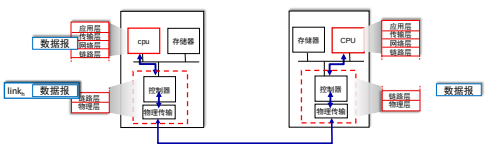
链路层在何处实现?

- 每一个主机中
- 链路层实现在*网络接口卡* (NIC) 或芯片上
  - Ethernet, WiFi卡 or 芯片
  - 实现链路层和物理层
- 连接到主机的系统总线
- 硬件、软件的结合





接口通信



发送方：
- 在帧中封装数据报
- 增加差错检测比特，可靠的数据传输，流量控制等

接收方：
- 查找差错、可靠的数据传输、流量控制等
- 提取数据报，传递到接收方的上层

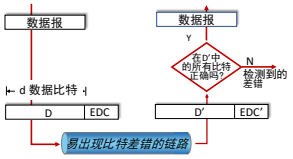
链路层, LANs: 路线图

- 概述
- 差错检测，纠正
- 多路访问协议
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化: MPLS
- 数据中心网络

- Web页面请求的历程

差错检测

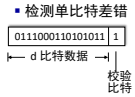
EDC（Error Detection and Correction）：差错检测和校正位（如冗余）
D：受差错检测保护的数据，包括首部字段



差错检测不是100%可靠！
- 协议可能会漏掉一些错误，但很少发生
- 更大的EDC域会有更好的检测和校正效果

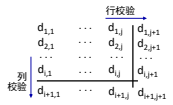
奇偶校验法

1比特奇偶校验：
- 检测单比特差错



偶校验：设置奇偶校验位，使1的个数为偶数

二维奇偶校验：
- 对D中的d个比特被划分为i行j列，产生i+j+1差错检测比特
- 检测*和纠正* 单比特差错，可检测（但不能纠正）2比特错误



无差错：
1 0 1 0 1 | 1
1 1 1 1 0 | 0
0 1 1 1 0 | 1
1 0 1 0 1 | 0

可纠正的单比特差错：
1 0 1 0 1 | 1
1 1 1 1 0 | 0 ← 校验错
0 1 1 1 0 | 1
1 0 1 0 1 | 0
↓ 校验错

* Check out the online interactive exercises for more examples: http://gaia.cs.umass.edu/kurose_ross/interactive/

因特网校验和（回顾）

目标： 在传输层检测差错（即翻转位）

发送者：
- 将UDP段的内容（包括UDP报头字段和IP地址）作为16位整数序列处理
- 校验和：片段内容的相加（二进制反码）
- 校验和的值放到校验和字段

接收者：
- 计算接收段的校验和
- 检查计算得到的检验和是否等于校验和字段的值：
  - 不相等 - 检测到差错
  - 相等 - 没有差错被检测到. *但也许差错还是存在?*

循环冗余检测(CRC，Cyclic Redundancy Check)

- 更强的差错检测编码
- D: *d* 比特数据位（给定的，可看成一个二进制数）
- G: 生成多项式，也称为 *r+1 比特模式*（给定的）

<D,R> = D·2^r XOR R

目标： 选择 *r* 个 附加比特, R, 使得 <D,R> 能够被 G (模 2 算术) 整除
- 接收方知道 G, 把 <D,R> 除以 G. 如果得到非零余数：检测到差错!
- 可以检测所有少于 r+1 位的突发差错
- 被广泛应用到实践中 (Ethernet, 802.11 WiFi)

循环冗余校验（CRC): 例子

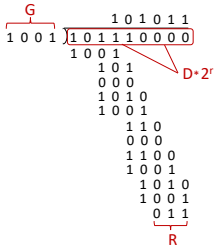
我们要求：
$D \cdot 2^r \ XOR \ R = nG$

等价于：
$D \cdot 2^r = nG \ XOR \ R$

等价于：
如果我们用 $D \cdot 2^r$ 除以 G，余数 R 满足：

$$R = remainder\left[\frac{D \cdot 2^r}{G}\right]$$

例子：
D=101110, *d*=6, G=1001, *r*=3



* Check out the online interactive exercises for more examples: http://gaia.cs.umass.edu/kurose_ross/interactive/

链路层, LANs: 路线图

- 引言
- 差错检测，纠正
- 多路访问协议
- LANs
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化: MPLS
- 数据中心网络

- Web页面请求的历程

多路访问链路, 协议

两种类型“链路”：

- 点对点
  - 以太网交换机、主机之间的点对点链路
  - PPP用于拨号访问
- 广播（共用电线或介质）
  - 传统 Ethernet
  - 有线接入网的上行HFC(混合光纤同轴电缆)
  - 802.11 无线 LAN, 4G/4G, 卫星

多路访问协议

- 单共享广播信道
- 节点同时进行两次或两次以上的传输：干扰
  - 如果节点同时接收到两个或多个信号就发生*碰撞*

多路访问协议(MAC, Multiple Access Protocol)

- 确定节点如何共享信道的分布式算法，即确定节点何时可以传输
- 信道共享的通信必须使用信道本身！
  - 没有带外信道用于协调

一种理想的多路访问协议

给定: 速率为 *R* bps的多路访问信道 (MAC)

希望:

1. 当仅有一个节点发送数据时，它可以以 *R* 的速率发送
2. 当有M个节点发送数据时，每个节点的平均发送速率为 *R/M*
3. 完全分散的:
   - 没有特殊的节点来协调传输
   - 没有时钟、时隙的同步
4. 简单

MAC 协议: 分类

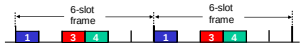
三大类:

- 信道划分协议
  - 将信道分成更小的“片piece”(时隙、频率、编码)
  - 将片分配给节点单独使用
- 随机接入协议
  - 信道不分割，允许碰撞
  - 从碰撞中“恢复”
- 轮流协议
  - 节点轮流发送数据，但可能有更多数据要发送得节点得到更长的时间段来发送其数据

信道划分协议: TDMA

TDMA: 时分多址(time division multiple access)

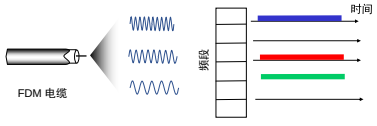
- “轮流”访问信道
- 将时间划分为时间帧，并进一步划分每个时间帧为N个时隙
- 每个节点在每轮得到固定长度的时隙（长度=单个分组传输时间）
- 未使用的时隙空闲
- 例如:6个节点的LAN，1,3,4节点有数据包要发送，时间片2,5,6空闲



信道划分协议: FDMA

FDMA: 频分多址(frequency division multiple access)

- 信道频谱划分为频段
- 每个节点被分配固定的频段
- 频段中未使用的传输时间空闲
- 例如:6个节点的LAN，1,3,4节点有数据要发送，频带2,5,6空闲



随机接入协议

- 当节点有数据包要发送时
  - 以信道的全部速率 R 传输
  - 节点之间不存在*事先*协调
- 两个或多个正在传输节点：“碰撞”
- 随机接入协议指定:
  - 如何检测碰撞
  - 如何从碰撞中恢复 (例如，通过延迟重传)
- 随机接入协议的例子:
  - ALOHA, 时隙 ALOHA
  - CSMA（载波侦听多路访问）, CSMA/CD， CSMA/CA

时隙 ALOHA

假定:

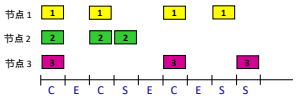
- 所有帧大小相同
- 时间被划分为大小相等的时隙(传输一帧的时间)
- 节点只在时隙的开始传输数据
- 节点同步
- 如果时隙中有2个或2个以上的节点传输数据，则所有节点都检测到碰撞

操作:

- 当节点获得新帧时，在下一个时隙传输
  - *如果没有碰撞*：节点可以在下一个时隙发送新的帧
  - *如果碰撞*：节点以p的概率在随后的每个时隙重新传输帧直到成功

随机 – *为什么*?

时隙 ALOHA



C: 碰撞
S: 成功
E: 空的

优点:

- 单个活动节点可以在全信道速率下连续传输
- 高度分散:每个节点检测碰撞并独立地决定何时重传
- 简单

缺点:

- 碰撞，浪费时隙
- 空闲时隙
- 时钟同步

时隙 ALOHA: 效率

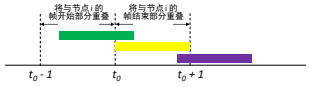
效率: 长期运行中成功时隙的份额（大量节点，所有节点都有许多帧要发送）

- *假设*: N 个节点有许多帧要发送，每个节点在时隙内传输的概率为 p
  - 一个给定节点在时隙内成功传送的概率 = $p(1-p)^{N-1}$
  - *任意*一个节点成功传送的概率 = $Np(1-p)^{N-1}$
  - 最大效率: 求出 p^*，使得 $Np(1-p)^{N-1}$ 最大化
  - 对于大量节点，取 $Np^*(1-p^*)^{N-1}$ 当 N 趋于无穷时的极限，得到：

  最大效率 = 1/e = .37
- *最多*: 信道有效传输时间仅有 37%！

纯 ALOHA

- 非时隙 Aloha: 更简单，无同步
  - 当一帧首次到达时：立即传输
- 在没有同步的情况下，碰撞概率会增加：
  - 在 t_0 发送的帧与在 $[t_0\text{-}1,t_0\text{+}1]$ 发送的其它帧发生冲突



- 纯 Aloha 的最大效率: 18%！

载波侦听多路访问
CSMA (Carrier Sense Multiple Access)

简单的 CSMA: 传输之前先听：

- 如果侦听到信道空闲：传输整个帧
- 如果侦听到信道繁忙：延迟传输

- 类比：谈话时不要打断别人！

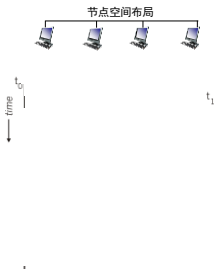
CSMA/CD:具有*碰撞检测*的CSMA

- 在短时间内*检测*到碰撞
- 碰撞后传输中止，减少信道浪费
- 在有线传输中容易做到碰撞检测，但在无线传输中难以做到

- 类比：有礼貌的健谈者

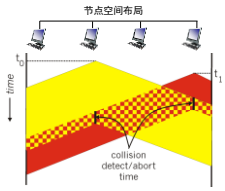
CSMA: 碰撞

- 在载波侦听的情况下仍可能发生碰撞：
  - 传播延迟意味着两个节点可能无法听到对方刚刚开始的传输
- 碰撞：整个数据包传输时间浪费
  - 距离和传播延迟是决定碰撞概率的重要因素



具有碰撞检测的载波侦听多路访问
CSMA/CD （CSMA with Collision Detection）：

- 减少在碰撞中浪费的时间
  - 在检测到碰撞时传输中止



以太网 CSMA/CD 算法

1. NIC从网络层接收数据报，创建帧
2. NIC 侦听信道：
   如果空闲：开始传输帧
   如果繁忙：等到信道空闲再传输
3. 如果 NIC 传输整个帧时没有碰撞，NIC 就完成了该帧的传输！
4. 如果 NIC 在发送时检测到另一个帧也在传输：中止传输
5. 中止之后，NIC 执行*二进制指数后退：*
   - 第 *m* 次碰撞后，NIC 从 *{0,1,2, …, 2^m-1}* 中随机选择 *K*。NIC 等待 *K*·512 比特时间（即发送512比特进入以太网所需时间的K倍），返回步骤 2
   - 更多碰撞：更长的后退间隔

CSMA/CD 效率

- t_{prop} = LAN 中 2 个节点之间的最大传播时延
- t_{trans} = 一个最大长度的帧的传输时间

$$efficiency = \frac{1}{1 + 5t_{prop}/t_{trans}}$$

- 效率接近于 1
  - 当 t_{prop} 接近 0
  - 当 t_{trans} 接近无穷
- 性能优于 ALOHA: 简单，廉价，分散！

轮流（Taking Turns）协议

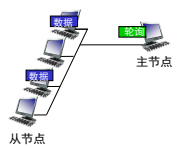
信道划分协议:
- 高负载时*有效*和*公平*地共享信道
- 低负载时效率低下：即使只有 1 个活跃节点也分配 1/N 的带宽

随机接入协议
- 低负载时高效：单节点可以充分利用信道
- 高负载时：碰撞开销

轮流协议
- 具备上述两种协议的优点

轮流协议

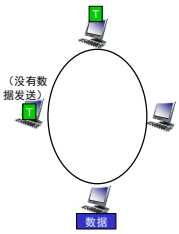
轮询：
- 主节点“邀请”其他节点依次进行传输
- 问题：
  - 轮询开销
  - 延迟
  - 单点故障（主）



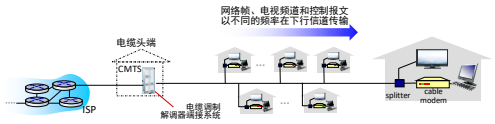
轮流协议

令牌传递协议：

- 控制 *令牌* 按顺序从一个节点传递到下一个节点。
- 令牌消息
- 问题：
  - 令牌开销
  - 延迟
  - 单点故障（令牌）

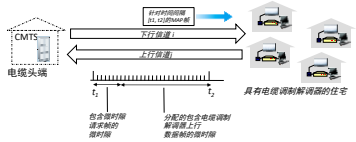


电缆接入网络：FDM, TDM 和 随机接入的例子



- 多个下行（广播）FDM 信道：最多 1.6 Gbps/信道
  - 仅有单一的 CMTS（Cable Modem Termination System） 在下行信道上传输
- 多个上行信道（最多 1 Gbps/信道）
  - 多路访问：所有用户争用(随机接入)一部分上行信道时隙；其他时隙被明确分配给用户，类似于TDM

电缆接入网络：



DOCSIS: 数据经电缆服务接口规范

- FDM 在上行，下行频率信道
- 上行：一些时隙被分配（TMD），一些有争用（随机接入）
  - CTMS通过在下行信道上发送 MAP 帧来制定哪个电缆调制解调器能够使用上行微时隙
  - 电缆调制解调器在专用的一组微时隙中向CTMS发送微时隙请求，当推断出有碰撞时使用二进制指数回退将其微时隙请求在以后的时隙中重新发送

MAC 协议的总结

- 信道划分协议，按时间、频率或编码
  - 分时，分频
- 随机接入协议（动态的），
  - ALOHA, S-ALOHA, CSMA, CSMA/CD
  - 载波侦听：有些技术容易(有线)，有些技术难(无线)
  - CSMA/CD used in Ethernet
  - CSMA/CA used in 802.11
- 轮流协议
  - 轮询协议,令牌传递协议
  - 蓝牙, FDDI（光纤分布式数据接口）,令牌环

链路层, LANs: 路线图

- 引言
- 差错检测，纠正
- 多路访问协议
- LANs
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化: MPLS
- 数据中心网络

- Web页面请求的历程

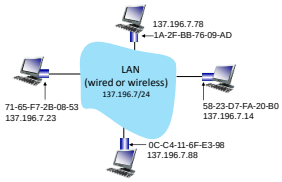
MAC地址

- 32 位 IP 地址:
  - *用于接口的网络层地址*
  - 用于第 3 层（网络层）转发
  - 例如： 128，119.40.136
- MAC（或局域网或物理或以太网）地址:
  - 功能:将“本地”的帧从一个接口连接到另一个物理连接的接口（IP寻址意义上的相同子网）
  - 48 位 MAC 地址（对于大多数 LAN）刻入在 NIC ROM ，有时也可通过软件设置
  - 例如: 1A-2F-BB-76-09-AD

十六进制（基础16）记法（每个“数字”表示4位）

MAC 地址

LAN中的每个接口
- 具有唯一的48位 MAC 地址
- 具有唯一的 32 位 本地IP 地址

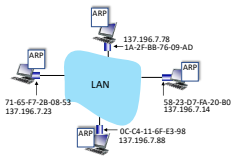


MAC 地址

- MAC 地址分配由 IEEE 管理
- 制造商购买 MAC 地址空间的一部分（以确保唯一性）
- 类比:
  - MAC 地址：如社会保险号码
  - IP地址：如邮政地址
- MAC 地址的可移植性
  - 可以将接口从一个局域网移动到另一个局域网
  - IP地址不可移植性：节点的IP地址取决于节点所在的IP子网

ARP：地址解析协议

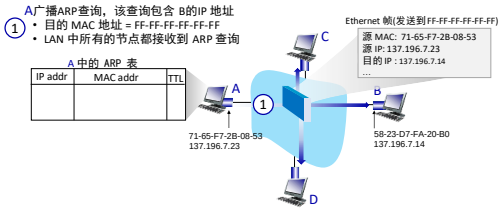
问题：已知接口的IP地址，如何确定接口的MAC地址？



ARP 表: LAN中的每个IP节点（主机，路由器）都有一个ARP表
- 存储某些LAN节点的IP/MAC地址映射:
  < IP address; MAC address; TTL>
- TTL (Time To Live):寿命值，从表中删除某个映射的时间（如，典型值为20分钟）

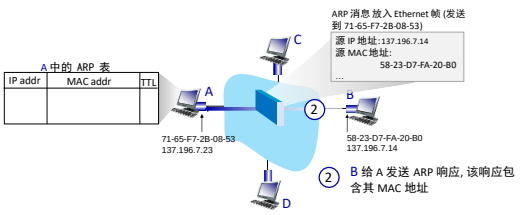
ARP 协议的执行过程

例子：A希望将数据报发送到B
- B的 MAC 地址不在A的 ARP表中，所以A使用ARP协议寻找B的MAC地址



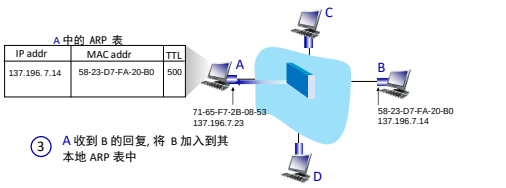
ARP 协议的执行过程

例子：A希望将数据报发送到B
- B的 MAC 地址不在A的 ARP表中，所以A使用ARP协议寻找B的MAC地址



ARP 协议的执行过程

例子：A希望将数据报发送到B
- B的 MAC 地址不在A的 ARP表中，所以A使用ARP协议寻找B的MAC地址

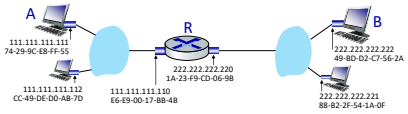


路由到另一个子网：寻址

示例:通过路由器R将数据报从A发送到B

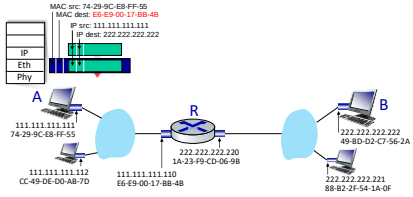
关注在IP（数据报）和MAC层（帧）级别的寻址

- 假设:
  - A 知道 B 的 IP 地址
  - 知道第一跳路由器R的IP地址(如何知道的?)
  - A 知道 R 的 MAC 地址 (如何知道的?)



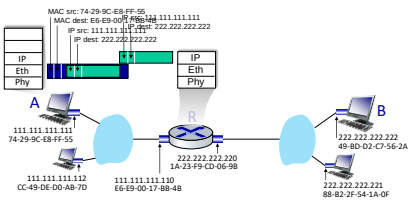
路由到另一个子网：寻址

- A 创建 IP 数据报，包含 IP 源 A 和目的 B
- A 创建 链路层帧，包含 A到B 的 IP数据报
  - R 的 MAC 地址是帧的目的地址



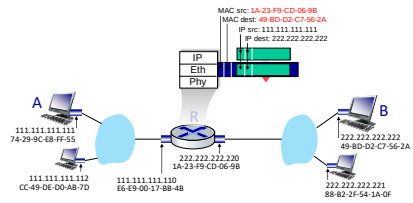
路由到另一个子网：寻址

- 帧从 A 发送到 B
- R 接收到帧，删除数据报，传送到 IP 层



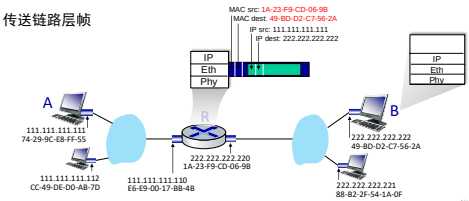
路由到另一个子网：寻址

- R 确定传出接口，将IP源为A，目的为B的数据报发送到链路层
- R 创建链路层帧，包含 A 到 B 的 IP数据报，帧的目的地址为 B 的 MAC地址



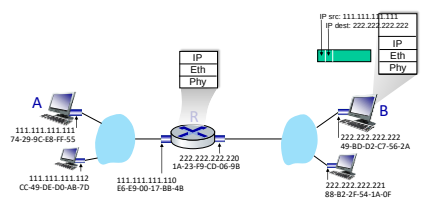
路由到另一个子网：寻址

- R 确定传出接口，将IP源为A，目的为B的数据报发送到链路层
- R 创建链路层帧，包含 A 到 B 的 IP数据报，帧的目的地址为 B 的 MAC地址
- 传送链路层帧



路由到另一个子网：寻址

- B 接收到帧，提取 IP数据报的目的地址 B
- B 将数据报发送到 IP 层



链路层, LANs: 路线图

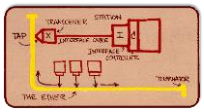
- 引言
- 差错检测，纠正
- 多路访问协议
- LANs
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化: MPLS
- 数据中心网络

- Web页面请求的历程

以太网 Ethernet

有线LAN技术:

- 几乎占领着现有的有线局域网市场
- 更简单，便宜
- 速率: 10 Mbps – 400 Gbps
- 单芯片，多速率(e.g., Broadcom BCM5761 博通)



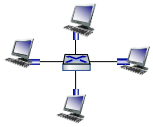
Metcalfe's Ethernet sketch

https://www.uspto.gov/learning-and-resources/journeys-innovation/audio-stories/defying-doubters

Ethernet:物理拓扑

- 总线: 90年代中期盛行
  - 所有节点在同一个冲突域中（可以发生冲突）
- 交换机:现在盛行
  - 中心是活跃的二层交换机
  - 交换机将链路彼此隔离，局域网中不同链路能够以不同速率在不同媒体上运行。

总线:同轴电缆



交换机

以太网帧结构

发送接口将IP 数据报（或其他网络层协议的数据包）封装为以太网帧

前同步码:

- 用于同步接收方和发送方的时钟速率。
- 前七个字节为10101010，最后一个字节为10101011

以太网帧结构

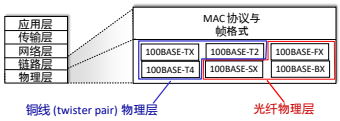
- 地址：源网络适配器或目的网络适配器的6字节MAC地址
  - 网络适配器收到一个帧时，若该帧的目的地址是适配器的MAC地址或是广播MAC地址，那么网络适配器将把该帧传递给本机的网络层协议。
  - 否则，适配器丢弃该帧。
- 类型：标识高层协议
  - 多数情况下是IP协议，但是也有可能是其他协议，如Novell IPX, AppleTalk
  - 用于复用多种网络层协议。
- CRC：在接收方进行循环冗余检测
  - 若检测到错误，则丢弃该帧

以太网：不可靠，无连接

- 无连接: 发送网卡和接收网卡之间无需握手
- 不可靠: 接收网卡不向发送网卡发送ACK或NAK
  - 只有当原始发送者使用了更高层的协议（如TCP)时，丢弃帧中的数据才会被重发，否则丢弃的数据就丢掉了。
- 以太网的MAC协议: 无时隙的、使用二进制指数回退的CSMA/CD

802.3 以太网标准：链路层 & 物理层

- 许多不同的以太网标准
  - 使用相同的MAC协议和帧格式
  - 不同的速度: 2 Mbps, 10 Mbps, 100 Mbps, 1Gbps, 10 Gbps, 40 Gbps
  - 不同的物理层介质：光纤，铜线



链路层, LANs: 路线图

- 引言
- 差错检测，纠正
- 多路访问协议
- LANs
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化: MPLS
- 数据中心网络

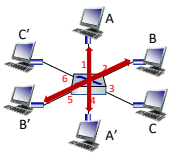
- Web页面请求的历程

以太网交换机

- 交换机是一种链路层设备: 扮演一个主动的角色
  - 存储/转发以太网帧
  - 检查传入帧的MAC地址, 选择性地将帧转发到一个或多个输出链路中，使用CSMA/CD 多路访问
- 透明的：主机并不知道交换机的存在
- 即插即用，自学习
  - 交换机无需配置

交换机: 多个同时传输

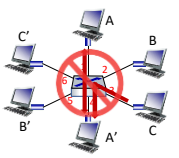
- 主机与交换机之间有专用的、直接的连接。
- 交换机缓存数据包
- 每个入链路上都使用了以太网协议，因此：
  - 无碰撞；全双工
  - 每条链路是它自己的冲突域。
- 交换: A-to-A′ 和B-to-B′ 可以同时传输，不会发生碰撞。



有6个接口(1,2,3,4,5,6)的交换机

交换机: 多个同时传输

- 主机与交换机之间有专用的、直接的连接。
- 交换机缓存数据包
- 每个入链路上都使用了以太网协议，因此：
  - 无碰撞；全双工
  - 每条链路是它自己的冲突域。
- 交换: A-to-A′ 和B-to-B′ 可以同时传输，不会发生碰撞。
  - 但是A-to-A′ 和C to A′ 不能同时传输



有6个接口(1,2,3,4,5,6)的交换机

交换机转发表

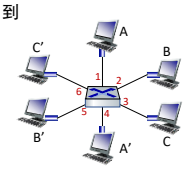
Q: 交换机是如何知道，A可以通过接口1到达，而B可以通过接口2到达呢？

A: 每个交换机都有一个交换机表，每个表项：

- 包含主机的MAC地址，通往主机的接口，时间戳
- 与路由表类似！

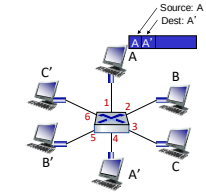
Q: 在交换机表中，表项是如何创建和维护的？

- 类似于路由协议？



交换机: 自学习

- 交换机学习“哪个接口可以到达哪个主机”，即主机与接口的对应关系。
  - 对交换机收到的每个入帧，交换机将学习发送者的“位置”，即其所在的局域网网段
  - 在交换机表中记录发送方MAC地址（帧的源地址）及对应的位置（接口）信息。



MAC addr	interface	TTL
A	1	60

交换机表
(初始为空)

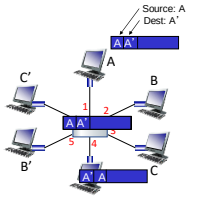
交换机: 帧过滤/转发

当帧到达交换机时:

1. 记录入帧的接口x以及发送主机的MAC地址
2. 使用目的MAC地址对交换表进行索引
3. If 使用目的MAC地址找到对应的表项
   then {
   if 目的MAC地址对应的接口为x
   then 丢弃该帧
   else 将该帧转发到表项所指示的接口
   }
   else 洪泛 /* 即将该帧转发到除x外的所有接口*/

自学习, 转发: 例子

- 目的主机A′ 的位置未知:洪泛
- 目的主机A的位置已知:
  选择相应的链路，进行发送

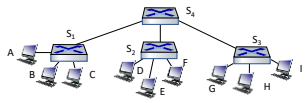


MAC addr	interface	TTL
A	1	60
A′	4	60

交换机表
(初始为空)

互连的交换机

自学习的交换机可以连接在一起:

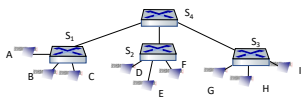


Q: 主机A向主机G发送数据时，S_1是如何知道数据包要通过S_4和S_3转发的呢？

- A: 自学习！（工作原理与单交换机的情况完全相同!）

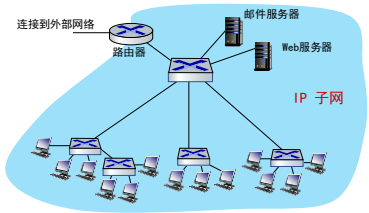
多交换机自学习的例子

假定主机C向I发送帧, I 向C发送响应



Q: 请给出S_1, S_2, S_3, S_4中的交换机表和数据包转发情况。

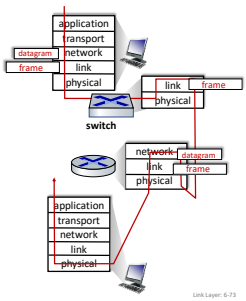
小型机构网络



交换机 vs. 路由器

都是存储转发设备:

- 路由器：网络层设备(检查网络层头)
- 交换机：链路层设备(检查链路层头)



都有转发表:

- 路由器: 使用路由算法和IP地址计算转发表
- 交换机: 使用洪泛（flooding）学习和MAC地址学习，得到交换机转发表

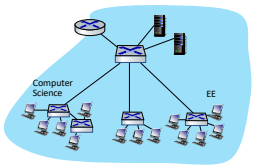
链路层, LANs: 路线图

- 引言
- 差错检测，纠正
- 多路访问协议
- LANs
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化: MPLS
- 数据中心网络

- Web页面请求的历程

虚拟局域网(VLANs): 动机

Q: 当局域网规模扩大时，用户的连接点发生了什么变化？

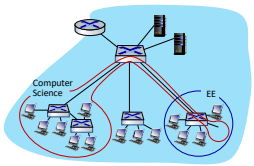


单一的广播域：

- *扩展：* 所有2层广播流量(ARP, DHCP, unknown MAC) 必须穿过整个局域网
- 效率，安全，隐私问题

虚拟局域网(VLANs): 动机

Q: 当局域网规模扩大时，用户的连接点位置发生变化时该如何处理？



单一的广播域：

- 所有2层广播流量(ARP, DHCP, unknown MAC) 必须穿过整个局域网，如何隔离机构中的N个组？买N个交换机？
- 效率，安全，隐私问题

管理问题：

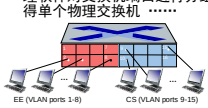
- CS的用户办公室搬到EE– *物理上* 连接点为EE的交换机，但是想在逻辑上保持CS交换机的连接

基于端口的虚拟局域网

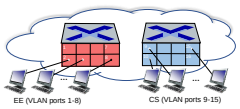
虚拟局域网(VLAN)

支持VLAN的交换机可以在单个物理局域网上配置多个虚拟局域网。

基于端口的 VLAN：通过交换机管理软件对交换机端口进行分组，使得单个物理交换机 ……

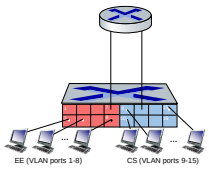


作为多个虚拟交换机运行。

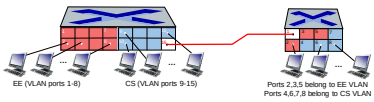


基于端口的虚拟局域网

- 流量隔离：发送到/来自端口1-8的数据帧只能到达端口1-8
  - 同样可以定义基于端口MAC地址的VLAN
- 动态成员配置：在VLAN间端口可以动态的配置
- VLANS间的转发：路由转发（类似于不同交换机间的转发）
  - 实际上，厂商销售的是交换机加路由器的组合



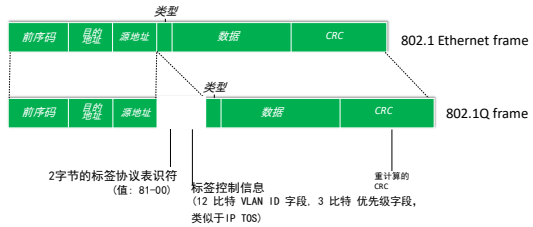
跨多交换机的VLANs



trunk port（干线端口）: 在跨多个交换机上的VLANS间传输数据帧

- VLAN内交换机之间转发的帧不是普通的802.1帧 (必须携带VLAN ID信息)
- 802.1q 协议为trunk ports 间的帧增加/移除额外的头部字段

802.1Q VLAN 帧格式



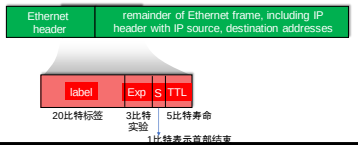
链路层, LANs: 路线图

- 引言
- 差错检测，纠正
- 多路访问协议
- LANs
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化: MPLS
- 数据中心网络

- Web页面请求的历程

多协议标签交换
Multi-Protocol Label Switching (MPLS)

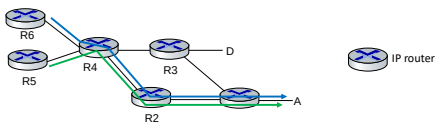
- 目标: 在支持MPLS的路由器之间进行高速IP数据报转发，使用固定长度标签 (而不是目的IP地址匹配)
  - 使用固定长度标识符更快的查找
  - 借鉴了虚拟电路(VC)方法
  - IP数据报仍然保留IP地址!



MPLS 路由器

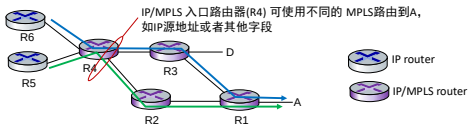
- 又称标签交换路由器
- 只根据标签值转发报文到出口(*不检查IP地址*)
  - MPLS转发表与IP转发表不同
- *灵活性*: MPLS转发决策可与IP转发策略不同
  - 使用目的地址和源地址—两个地址、将不同的流路由到相同的目的(流量工程)
  - 如果链路出现问题可以迅速地重新路由：备份路径实现就计算好了

MPLS 与 IP 路径



- IP routing: 到目的地的路径仅由目的地地址决定

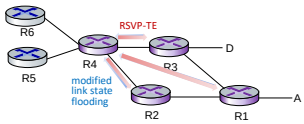
MPLS 与 IP 路径



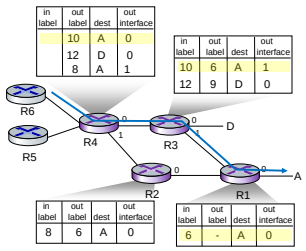
- IP routing： 到目的地的路径仅由目的地地址决定
- MPLS routing： 到目标的路径可以基于源地址和目标地址
  - 广义转发(10年前的MPLS)
  - *快速重路由*：提前计算好备份路由，以防链路故障

MPLS 信令

- 修改OSPF、IS-IS链路状态扩散协议，承载MPLS路由信息：
  - 如，链路带宽，预留链路带宽的数量
- 入口MPLS路由器使用RSVP-TE信令协议在下游路由器上建立MPLS转发



MPLS 转发表



链路层, LANs: 路线图

- 引言
- 差错检测，纠正
- 多路访问协议
- LANs
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化：MPLS
- 数据中心网络

- web请求生命中的一天

数据中心网络

成千上万的主机，在很近的距离上经常是紧密耦合的：

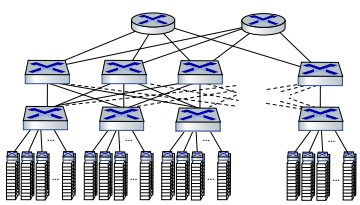
- 电子商务 (如， Amazon)
- 内容服务商 (如，YouTube, Akamai, Apple, Microsoft)
- 搜索引擎，数据挖掘(如，Google)

挑战：

- 多个应用程序，每个应用程序为大量客户端服务
- 可靠性
- 管理/平衡负载，避免处理、网络、数据瓶颈

在一个40英尺高的集装箱里的微软芝加哥数据中心

数据中心网络: 网络元素



边界路由器
- 连接外部数据中心

第一层交换机
- 连接~16 第二层交换机

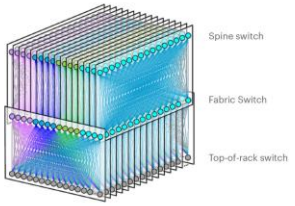
第二层交换机
- 连接~16个机架顶部交换机

机架顶部（TOR）交换机
- 每一个机架顶部
- 40-100Gbps 以太网连接

服务器机架
- 20- 40 服务器刀片：主机

数据中心网络：网络元素

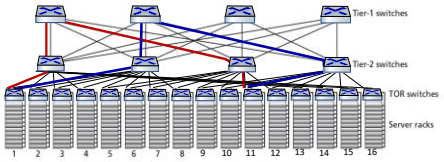
Facebook F16数据中心网络拓扑结构：



https://engineering.fb.com/data-center-engineering/f16-minipack/ (posted 3/2019)

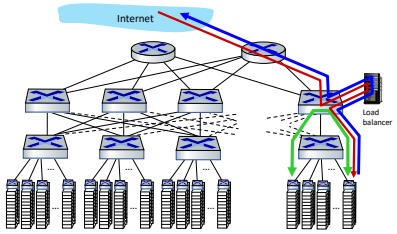
数据中心网络：多路径

- 交换机、机架之间存在丰富的连接:
  - 增加机架之间的吞吐量(因为存在多条路径)
  - 提高可靠性（通过冗余路径）



图上标红和标蓝的两条路径为机架1和机架11之间不相交的两条路径

数据中心网络： 应用层路由



负载均衡：应用层路由

- 接收外部客户端的请求
- 将请求分发给数据中心内不同的节点进行处理
- 返回数据给外部客户端 (对客户端隐藏数据中心内部)

数据中心网络: 协议创新

- 链路层:
  - RoCE: 基于融合以太网的远程内存直接访问
- 传输层:
  - 使用ECN(explicit congestion notification)的传输层拥塞控制算法(DCTCP, DCQCN)
  - hop-by-hop (backpressure) 拥塞控制实验
- 路由、管理:
  - SDN被广泛使用于各个组织的数据中心
  - 尽可能将相关服务，数据放置在尽可能近的位置（例如，在同一机架或附近机架中），以最大程度地减少2级，1级通信

链路层, LANs: 路线图

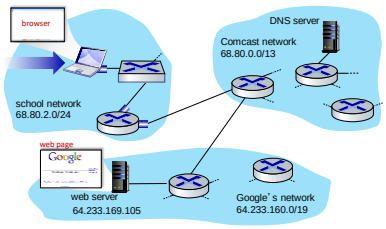
- 引言
- 差错检测，纠正
- 多路访问协议
- LANs
  - 寻址, ARP
  - Ethernet
  - 交换机
  - VLANs
- 链路虚拟化：MPLS
- 数据中心网络

- Web页面请求的历程

综合：Web页面请求的历程

- 我们关于协议的旅程完成啦!
  - 应用层，传输层，网络层，链路层
- 将学过的内容放在一起!
  - 目标：理解在一个看似简单的场景中涉及的协议（所有层）
  - 场景：学生将笔记本电脑连接到校园网络，请求www.google.com网页

Web页面请求的历程: **场景**

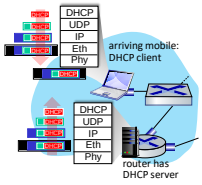


场景:

- 移动客户端连接到网络
- 请求web页面 www.google.com

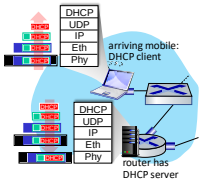
Sounds simple!

Web页面请求的历程: **连接到网络**



- 连接到网络的笔记本需要获取自己的IP地址，第一跳路由的IP地址，DNS服务器的地址：使用DHCP
- DHCP请求报文从上至下会被封装为UDP报文，IP报文，802.3以太网报文
- 以太网帧在LAN上广播（目的地址:FFFFFFFFFFFF），被路由器上的DHCP服务器接收到
- 以太网报文被分解为IP报文，UDP报文，DHCP报文

Web页面请求的历程: **连接到网络**

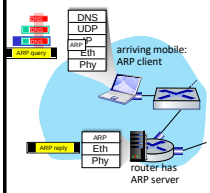


- DHCP服务器构造DHCP ACK报文，里面包含客户端的IP地址，第一跳路由的IP地址，DNS服务器的名字和IP地址
- 报文在DHCP服务器封装，通过LAN转发，客户端接收到后解封装
- DHCP客户端接收到DHCP ACK回应报文

现在客户端拥有IP地址，DNS服务器的名字和地址，第一跳路由的IP地址

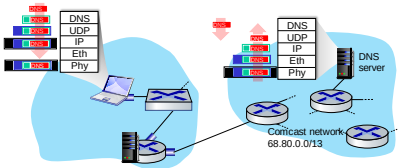
Web页面请求的历程...

ARP (before DNS, before HTTP)



- 在发送HTTP请求之前，需要www.google.com的IP地址: DNS
- DNS请求被依次封装到UDP、IP、Eth报文中。为了发送给路由器，需要知道路由器的MAC地址 : ARP
- 发送ARP 请求广播，被路由器接收到，返回带有路由器MAC地址的ARP回复报文
- 现在客户端知道第一跳路由的MAC地址，可以发送包含DNS请求的以太帧

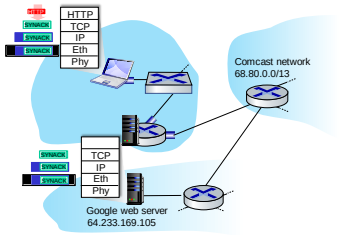
Web页面请求的历程... **使用DNS**



- IP数据报文解封装到DNS报文
- DNS服务器将www.google.com的IP地址回复给客户端

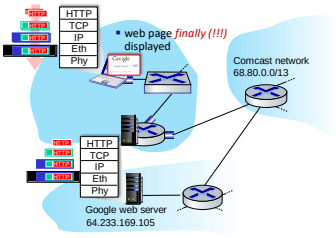
- 包含DNS请求的IP数据报文通过LAN交换机从客户端转发到第一跳路由
- IP数据报文从校园网络转发到运营商(Comcast)网络，被路由到DNS服务器（由RIP，OSPF，IS-IS和/或BGP路由协议创建的路由表转发)

Web页面请求的历程... TCP 和HTTP



- 为了发送HTTP请求，客户端首先创建到web server的TCP socket
- TCP SYN segment (tcp三次握手的第一步) 被路由到web服务器
- web 服务器回复 TCP SYNACK (tcp三次握手的第二步)
- TCP 连接已创建！

Web页面请求的历程… HTTP 请求和响应



- HTTP 请求通过TCP socket 发送
- 包含HTTP请求的IP 数据报文 被路由到 www.google.com
- web 服务器回复HTTP响应报文 (包含web页面)
- 包含HTTP响应报文的IP 数据报文被路由回客户端

Chapter 6: 总结

- 数据链路层服务背后的原理:
  - 错误检测，纠正
  - 多路访问控制
  - 链路层寻址
- 实例化，实现各种链路层技术
  - 以太网
  - 交换机、虚拟局域网VLANs
  - MPLS
- 综合: Web页面请求的历程

Chapter 6：总结

- *完成* 向下协议栈
- 扎实理解网络原理，练习!
- ..... 可以就此止步.... 但还有 *更多* 有趣的话题!
  - wireless
  - 安全

作 业

- 说明（举一个不同于图6-5的例子）二维奇偶校验能够纠正和检测单比特差错。说明（举一个例子）某些双比特差错能够被检测但不能纠正。
- 考虑5比特生成多项式，G = 10011， 并且假设D的值为1010101010。R的值是什么？

图6-5 二维偶校验

作 业

- 如图6-33所示，考虑通过两台路由器互联的3个局域网。
  - a. 对所有的接口分配IP地址。对于网1使用形式为192. 168. 1. xxx的地址，对于网2使用形式为 192. 168. 2. *xxx* 的地址，对于网3使用形式为192. 168. 3. xxx的地址。
  - b. 为所有的适配器分配MAC地址。
  - c. 考虑从主机E向主机B发送一个IP数据报。假设所有的ARP表都是最新的。就像在6.4.1节中对单路由器例子所做的那样，列举出所有步骤。
  - d. 重复（c），现在假设在发送主机中的ARP表为空（并且其他表都是最新的）。

图6-33 由路由器互联的3个子网

作 业

- 在某网络中标识为A到F的6个节点以星形与一台交换机连接，考虑在该网络环境中某个正在学习的交换机的运行情况。假定：
  - （i） B向E发送一个帧；
  - （ii） E向B回答一个帧；
  - （iii） A向B发送一个帧；
  - （iv） B向A回答一个帧。
- 该交换机表初始为空。显示在这些事件的前后该交换机表的状态。对于每个事件，指出在其上面转发传输的帧的链路，并简要地评价你的答案。

作业

- 考虑显示在图6-29中的MPLS网络，假定路由器R5和R6现在是MPLS使能的。假定我们要执行流量工程，使从R6发往A的分组要经R6-R4-R3-R1交换到A,从R5发向A的分组要过R5-R4-R2-R1交换。给出R5和R6中的MPLS表以及在R4中修改的表，使得这些成为可能。

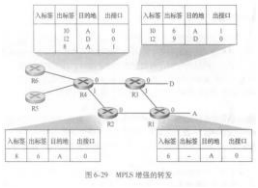


图 6-29 MPLS 增强的转发

作业

- 考虑在图6-30中具有等级拓扑的数据中心网络。假设现在有80对流，在第1和第9机架之间有10个流，在第2和第10机架之间有10个流，等等。进一步假设网络中的所有链路是10 Gbps,而主机和TOR交换机之间的链路是1 Gbps
  - a.每条流具有相同的数据率；确定一条流的最大速率。
  - b. 对于相同的流量模式，对于图6-31中高度互联的拓扑，确定一条流的最大速率。
  - c. 现在假设有类似的流量模式，但在每个机架上涉及20台主机和160对流。确定对这两个拓扑的最大流速率。

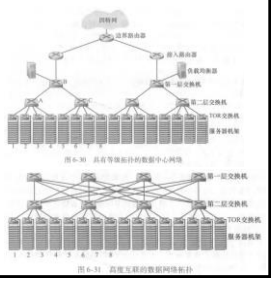


图 6-30 具有等级拓扑的数据中心网络

图 6-31 高度互联的数据网络拓扑

Additional Chapter 6 slides

Pure ALOHA efficiency

P(success by given node) = P(node transmits) *
P(no other node transmits in $[t_0\text{-}1,t_0]$ *
P(no other node transmits in $[t_0\text{-}1,t_0]$

$= p \cdot (1\text{-}p)^{N-1} \cdot (1\text{-}p)^{N-1}$

$= p \cdot (1\text{-}p)^{2(N-1)}$

… choosing optimum p and then letting n

$= 1/(2e) = .18 \rightarrow \infty$

even worse than slotted Aloha!