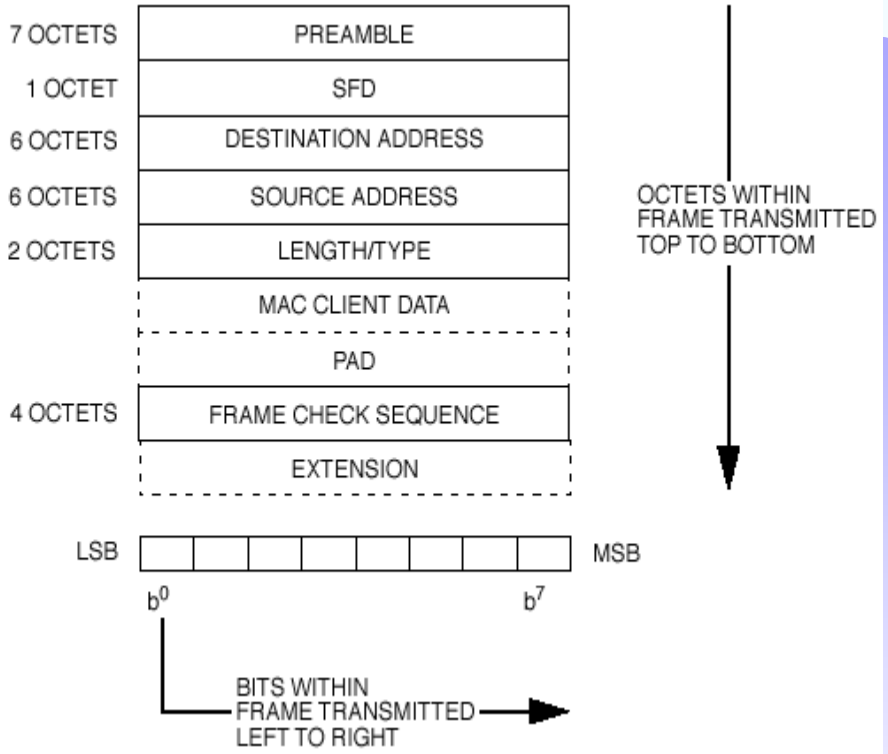
1.3 计算机网络的基本理论与技术

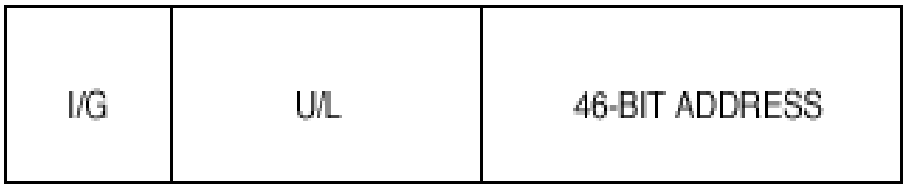
- 基本网络
- 命名与寻名
- 编址与路由
- 网络互连(IP)
- 数据运输
- 拥塞控制(Ch2)
- 组播

1.3.1 基本网络：之一以太网

- 1975年纯ALOHA原始ethernet：**单工竞争**系统，基本思想：
  - 无连接，先说后听，想发就发，错了重发；
  - 对数据帧不编号，不要求对方发回确认；**不可靠交付**，尽力而为
  - 建立在近距离、信道出错概率小-->局域网，出错由**高层重发**
- Time sloted ALHOA；
- CSMA（载波侦听多路访问）：先听后说+指数退避
  - 1持续CSMA、非持续CSMA、P持续CSMA
- CSMA/CD：多点接入、载波监听、碰撞检测
  - 信道效率（纯0.18，分片0.368）？**帧时T_0**的概念！
- 以太网**优势**
  - 可扩展性（10M—10G），灵活（多种媒介、全/半双工、共享/交换），易于安装使用、稳健性好。



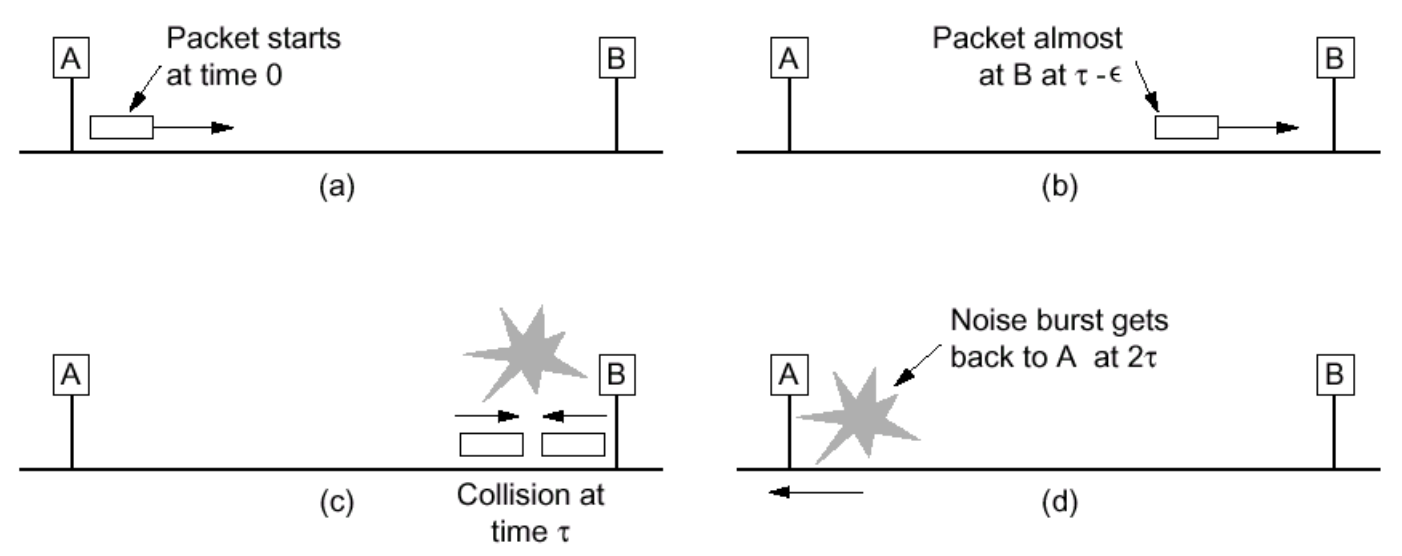
以太网的帧格式



I/G = 0 INDIVIDUAL ADDRESS
I/G = 1 GROUP ADDRESS
U/L = 0 GLOBALLY ADMINISTERED ADDRESS
U/L = 1 LOCALLY ADMINISTERED ADDRESS

以太网的地址格式

- 最短帧长
  - 避免帧的第一个比特到达电缆的远端前帧已经发完，帧发送时间应该大于 2τ；

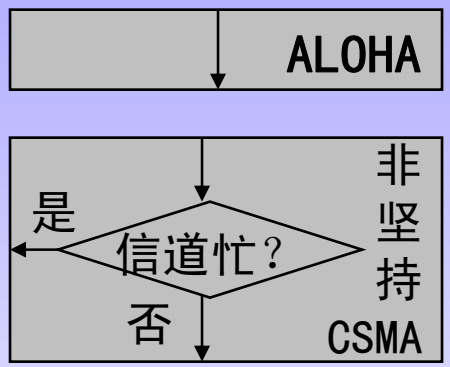


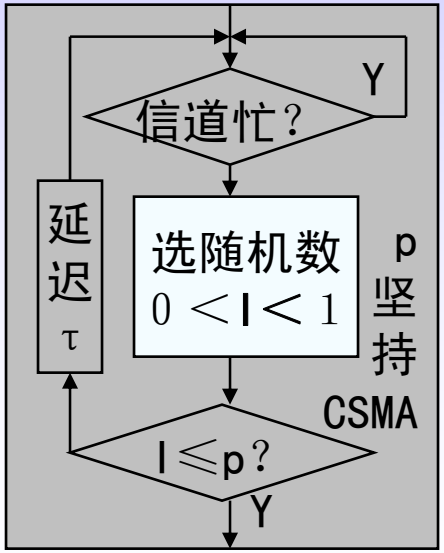
  - 10Mbps LAN，最大冲突检测时间为51.2微秒，最短帧长为64字节；
  - 网络速度提高，最短帧长也应该增大或者站点间的距离要减小。

传输算法

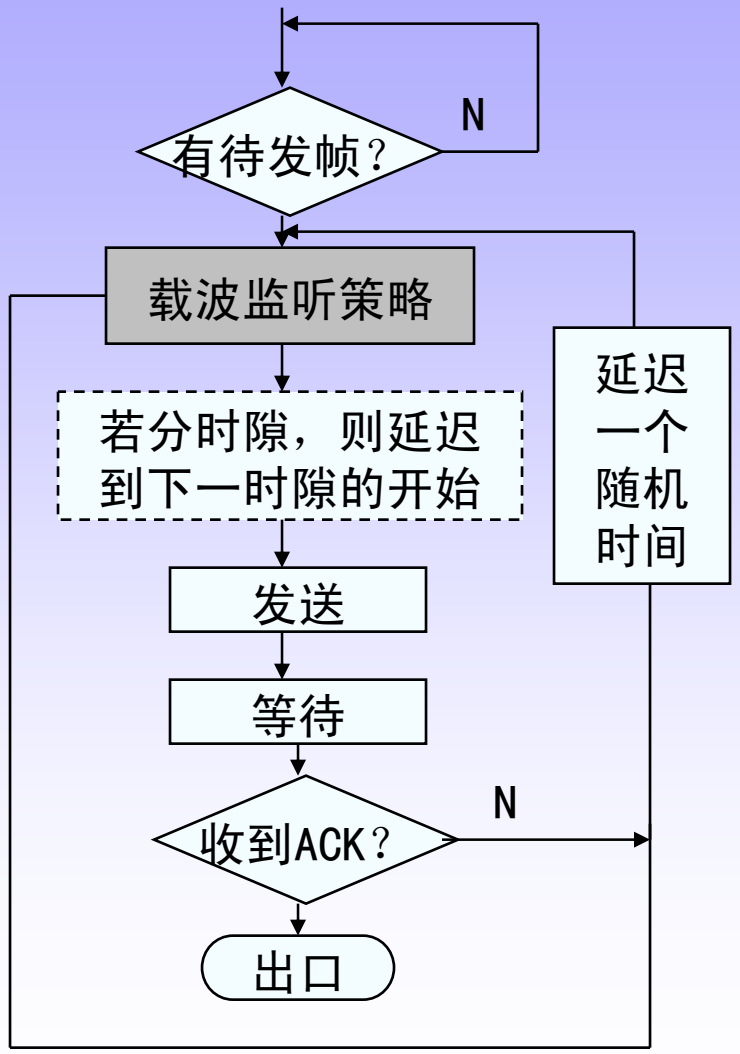
- ALOHA(无载波监听)
- 非坚持CSMA（Carrier Sense Multiple Access ）
  - 不能利用信道刚刚转入空闲这段时间
- p坚持CSMA：很难选择p值
- 1 坚持CSMA:坚持时容易产生冲突
- CSMA/CD(Collision Detect)

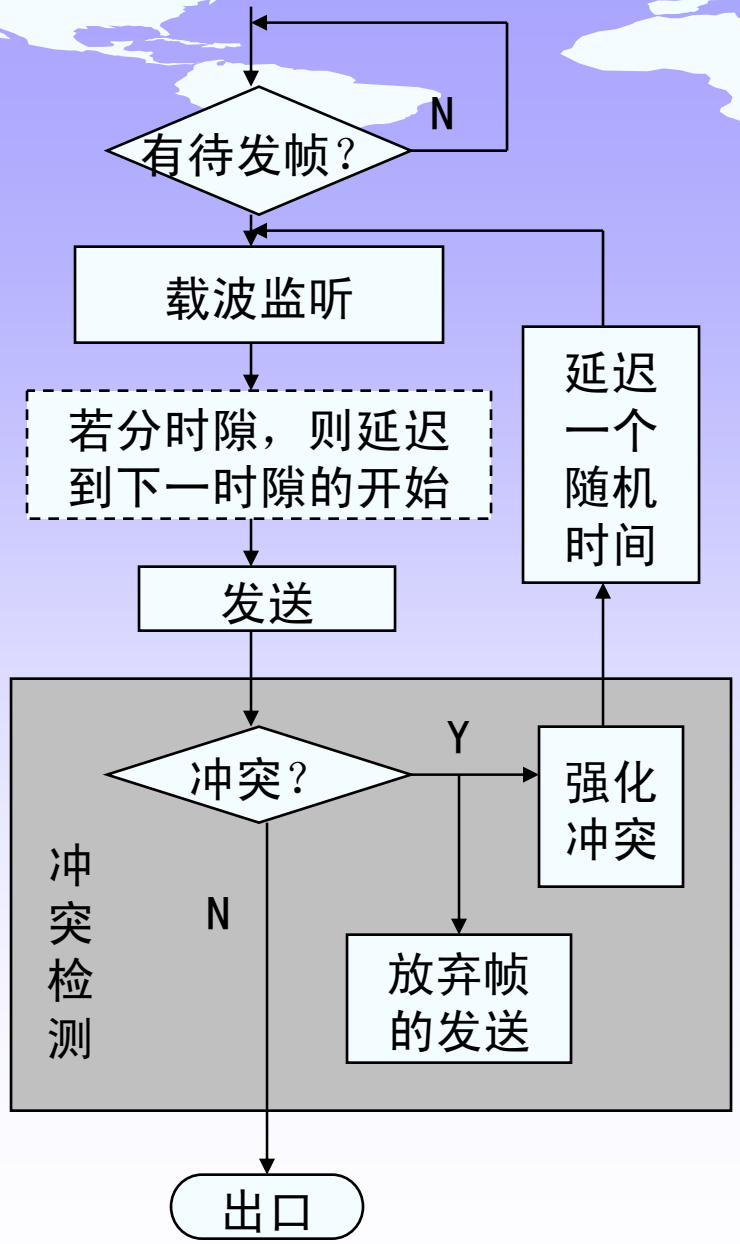
载波监听策略与CSMA接入流程





载波监听 3 策略





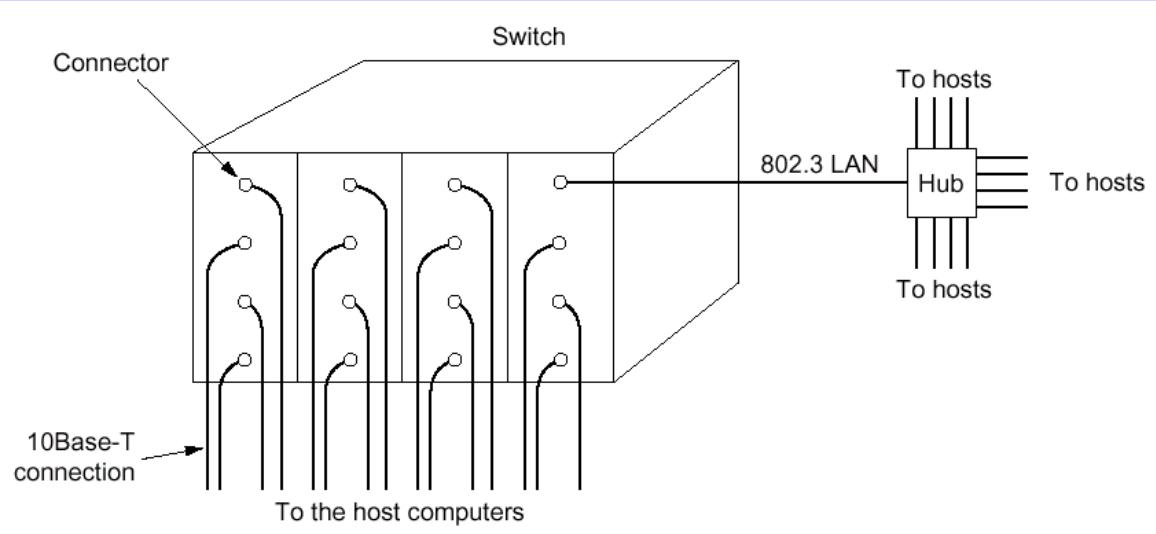
- CSMA/CD的流程图
- CD的方法：
  - 曼码的过零点在比特的正中央，冲突发生时，过零点的位置将发生变化
  - 发送帧的同时也进行接收，并逐比特比较，若不符则有冲突
- 也存在非坚持、p坚持和 1 坚持

截断二进制指数退避算法

- 算法决定重发帧所需的延迟时间
- 从离散的整数集合{0, 1, 2, 4...2^{k-1}}中随机选择一个数，设为r
- k=Min[重发次数, 10]:10前为冲突次数
- 所需的时延＝r×2τ（τ=51.2us）
- 重发16次不成功则丢弃该帧
- 时延随重发次数二增大--动态退避

- 二进制指数后退算法（binary exponential backoff）
  - 算法
    - 将冲突发生后的时间划分为长度为51.2微秒的时槽
    - 发生第一次冲突后，各个站点等待 0 或 1 个时槽再开始重传；
    - 发生第二次冲突后，各个站点随机地选择等待0，1，2或3个时槽再开始重传；
    - 第 i 次冲突后，在 0 至 2^i-1 间随机地选择一个等待的时槽数，再开始重传；
    - 10次冲突后，选择等待的时槽数固定在0至$2^{10}-1$间；
    - 16次冲突后，发送失败，报告上层。

- 交换式802.3 LAN
  - 目的：减少冲突；
  - 两种实现方法
    - 同一卡内是一个802.3LAN，构成自己的冲突域，卡间并行；
    - 使用端口缓存，无冲突发生。



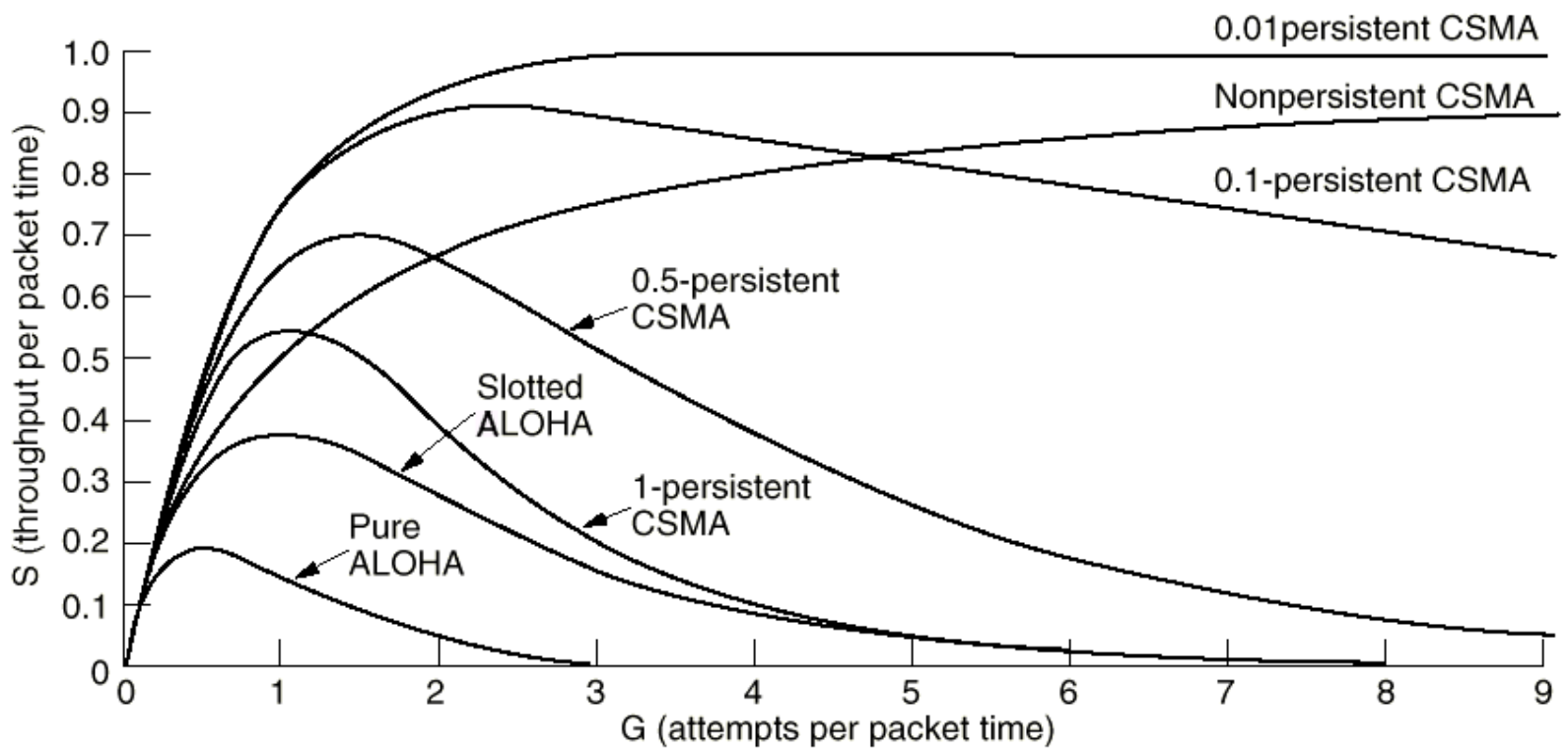
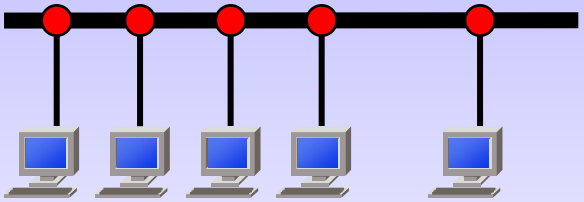


Fig. 4-4. Comparison of the channel utilization versus load for various random access protocols.

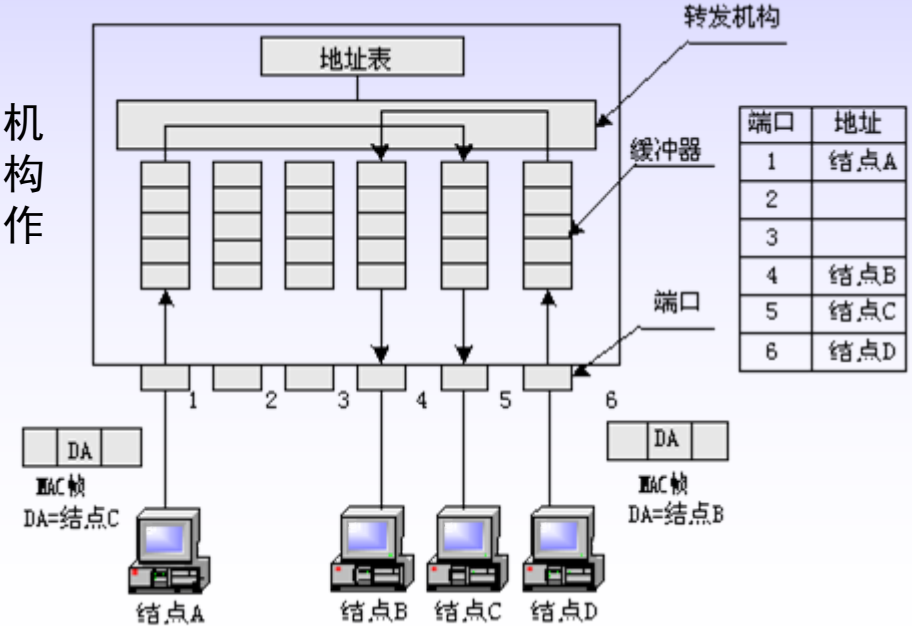
以太网的基本设备

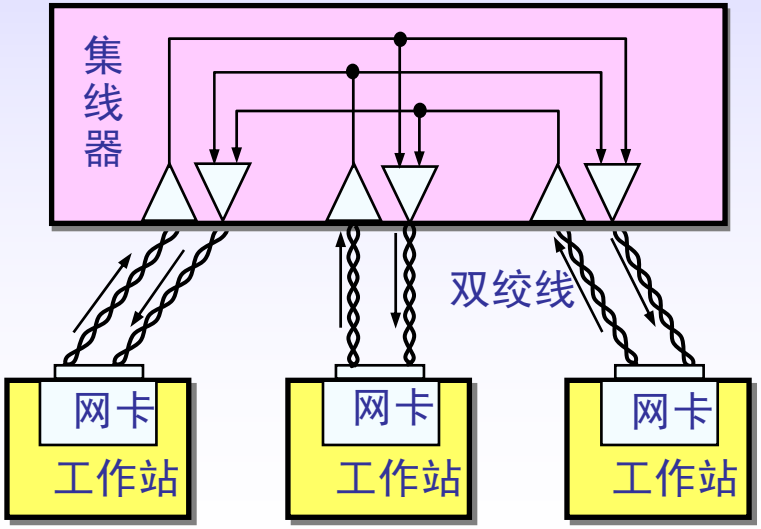
- 集线器（HUB）：**物理层**互连设备
  - 总线共享，线障隔离，**使用方便**
  - 带宽受限 ，广播风暴 ，单工传输，**通信效率低**
- 交换机（Switch）：**链路层**互连设备
  - 目的：减少冲突；隔离广播;构成VLAN；独立带宽
  - 实现方法
    - **直接交换**方式
    - **存储转发**方式
    - **改进直接交换**方式。

共享总线的集线器

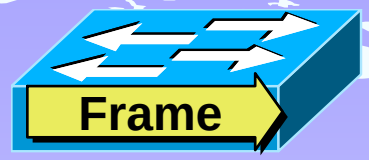


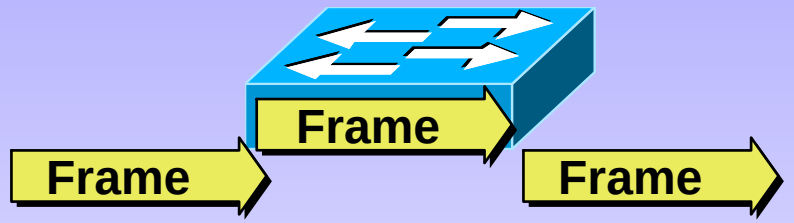
交换机的结构与工作过程

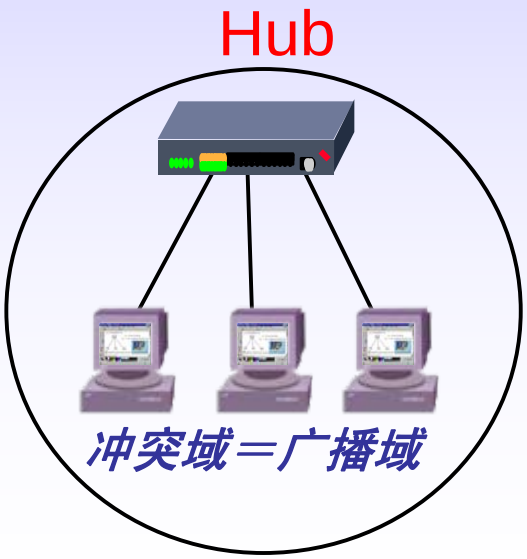


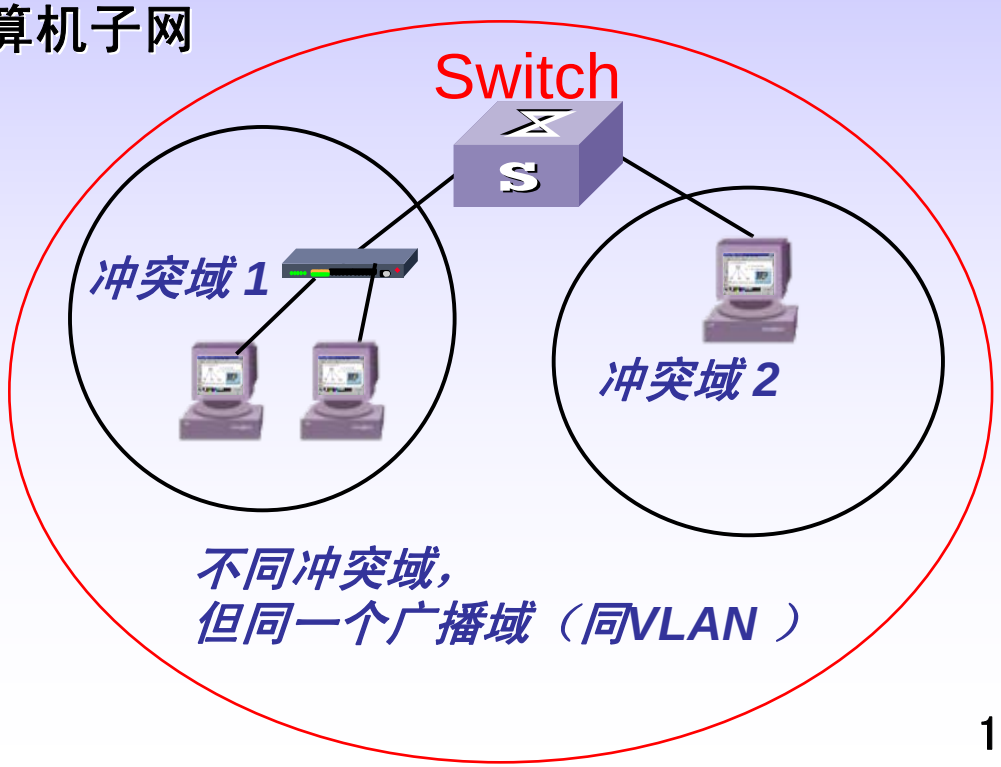


- 直通转发（cut-through）：
  - 交换机检测到**目标地址后即转发帧**
  - 优点－转发延迟小；缺点－错误率高
- 存贮转发(store and forword)
  - **完整地收到帧**并检查无错后才转发
  - 优点－错误率低；缺点－转发延迟大
- 广播域与冲突域
  - 同时**共享同一广播帧**的计算机子网
  - 同时**共享同一传播媒体**的计算机子网









网卡与MAC地址模式

- 网卡功能
  - 数据的封装与解封
  - 链路管理：CSMA/CD
  - 编码与解码
- MAC地址
  - 单播帧地址：仅对某个网卡
  - 广播帧地址：仅对某个子网
  - 多播帧地址：组地址
  - 杂收模式：Promiscuous mode:接收所有的可能接收的帧

高速局域网:快速以太网

- 对10 Mbps 802.3 LAN的改进
  - 10Base-T，使用HUB
  - 局域网发展史上重要里程碑
- Fast Ethernet标准
  - 1995年，IEEE通过802.3u标准，实际上是802.3的一个补充。原有的帧格式、接口、规程不变，只是将比特时间从100ns缩短为10ns。

Name	Cable	Max. segment	Advantages
100Base-T4	Twisted pair	100 m	Uses category 3 UTP
100Base-TX	Twisted pair	100 m	Full duplex at 100 Mbps
100Base-FX	Fiber optics	2000 m	Full duplex at 100 Mbps; long runs

CSMA

CSMA

CSMA

高速局域网:100Base-TX/F

- 100Base-TX
  - 使用2对5类平衡双绞线或150Ω屏蔽平衡电缆，1对 to the hub，1对 from the hub， 全双工；
  - 5类双绞线使用125 MHz的信号；
  - 4B/5B编码，5个时钟周期发送4个比特，物理层与FDDI兼容，比特率为 125 * 4/5 = 100 Mbps；
- 100Base-FX
  - 使用2根多模光纤，全双工
- 100Base-T4 和 100Base-TX 统称 100Base-T
- 两种类型的HUB
  - 共享式 HUB，一个冲突域，工作方式与802.3相同，CSMA/CD，二进制指数后退算法，半双工 ...
  - 交换式HUB，输入帧被缓存，一个端口构成一个冲突域。

高速局域网:千兆以太网

- 工作方式
  - IEEE 802.3定义的10M/100M以太网一致的CSMA/CD帧格式和MAC层协议
  - 以太网交换机（全双工模式）中的千兆端口不能采用共享信道方式访问介质，而只能采用专用信道方式，
  - 在专用信道方式下，数据的收/发能够不受干扰地同步进行
  - 物理层采用已有光纤通道技术；
- PAUSE协议
  - 规范发展完善了PAUSE协议，不采用CSMA/CD协议完成全双工操作。
  - 该协议采用不均匀流量控制方法最先应用于100M以太网中。
- 流控
  - 利用802.3定义的Pause控制帧进行流量控制，要求发送数据节点暂停数据发送，避免缓冲区溢出造成的丢包。
  - 只有在全双工时，才支持Pause流控，半双工时不支持流控。

万兆（10Gbps）以太网

- 2002.6月正式发布802.3ae 10GE标准
  - 只全双工，不支持单工和半双工，也不采用CSMA/CD
  - 不持自协商；提供广域网物理层接口。
- 长距离(40-50KM)网络
  - 扩展了网络的覆盖区域，且标准简化。
  - 支持现存的大量SONET网络兼容
- 两种物理层技术：
  - 局域网物理层LAN PHY；10.000Gbps精确10G；
  - 广域网物理层WAN PHY；入OC-192，异步SONET/SDH
  - 与10M/100M/1000Mbps帧格式完全相同；

之二： POS/DPT+SDH/WDM网络主干

- POS技术-Packet Over Sonet/SDH
  - 采用高速光纤传输，以点对点方式提供从STM1(155.520)到STM64(OC-192:9953.280=10Gbps)甚至更高的传输速率
  - 将IP包直接封装到SDH帧中，提高了传输的效率。

DPT-Dynamic Packet Transport

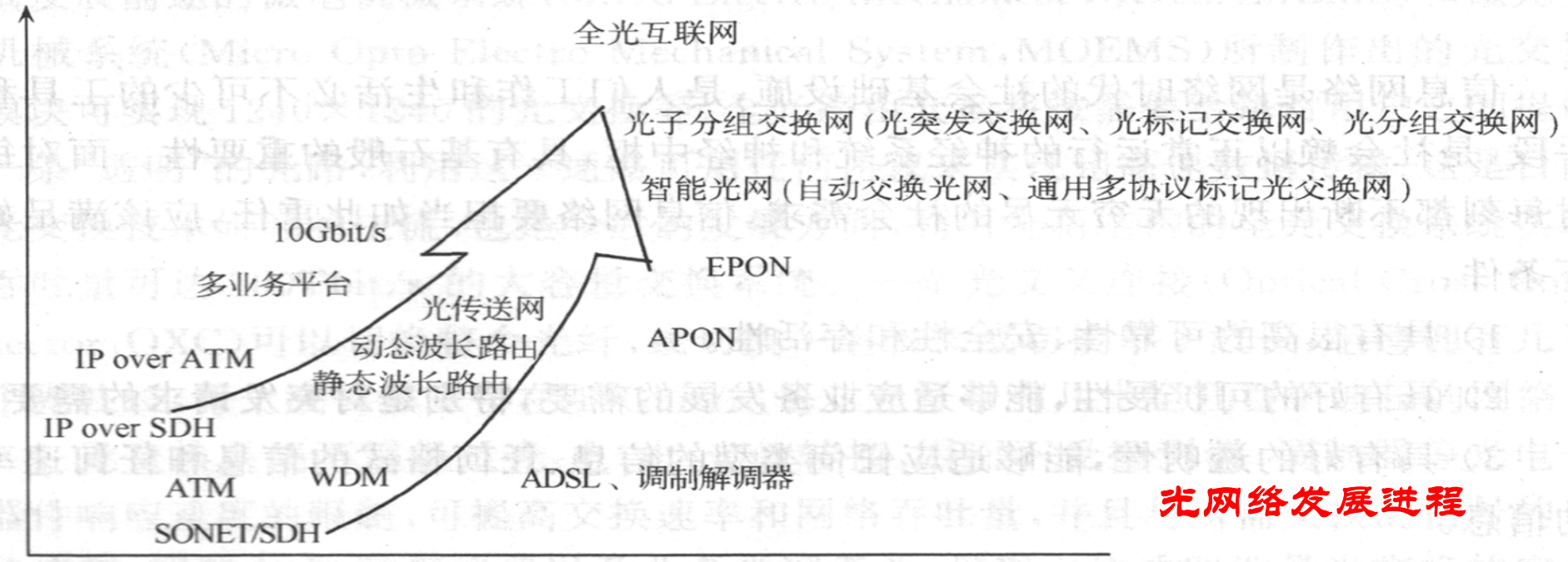
- 动态IP光纤传输技术
  - CISCO定义的一种全新的传输方法-IP优化的光学传输技术
  - 吸取POS技术的精华(可以识别SDH 帧中的K1/K2 比特从而保证快速的通道切换，可以基于原有的SDH线路进行传输等)
  - 克服成本较高的缺点。DPT 技术的关键在于提供了一种对带宽的空间复用(SRP: Spatial Reuse Protocol)机制，使多点可以共享一个光纤环、带宽可以进行动态分配
  - 一个SRP环上的每个设备永远只需要一对SRP端口(而点对点网状网中，每节点需N2个端口)，从而使网络扩容时不再需要增加端口，大大降低了网络成本。

之三：EOS+SDH/WDM网络主干

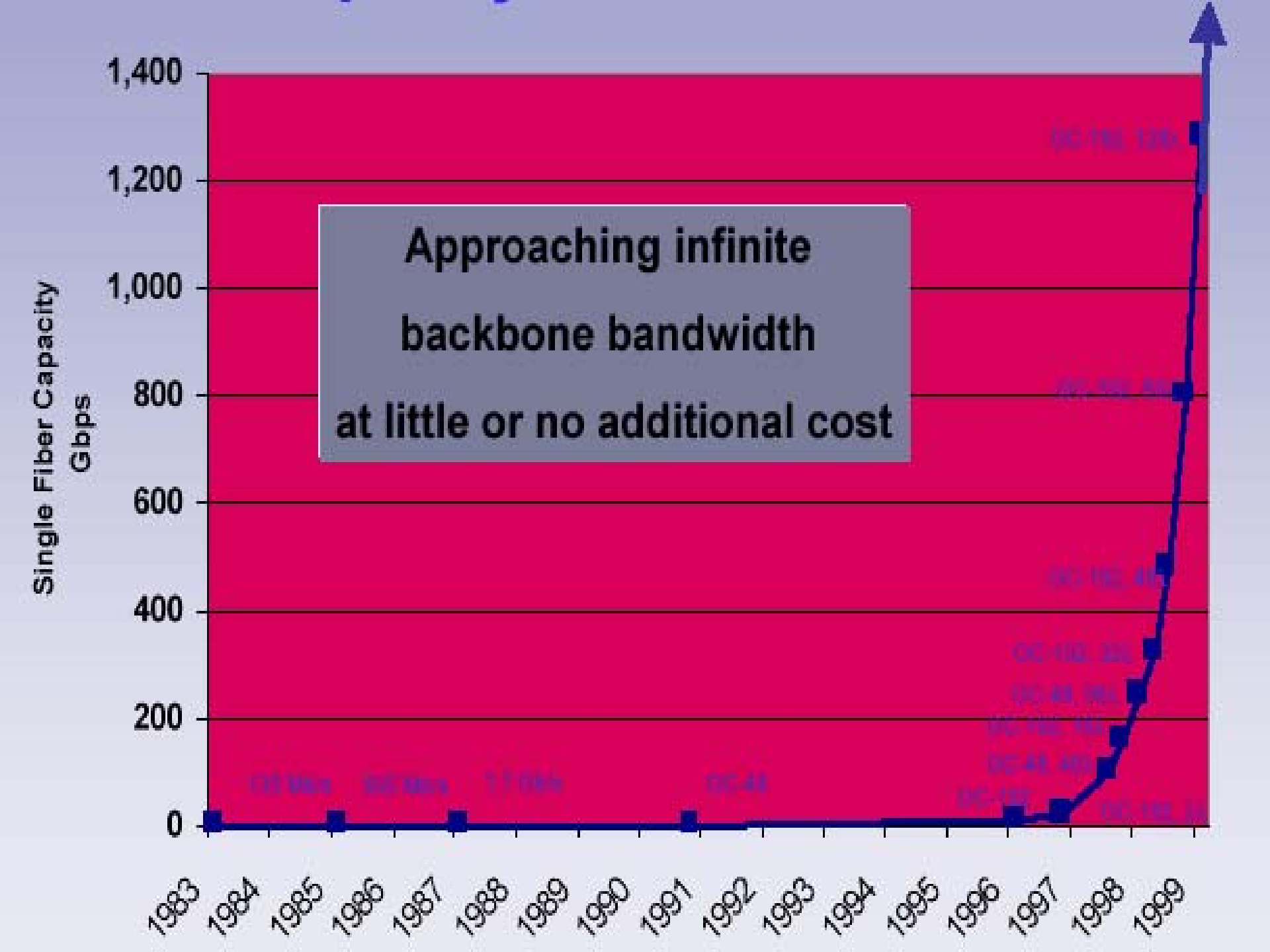
- Ethernet over SONET/SDH
  - 在SDH/SONET或DWDM上增加以太网的二层或三层交换板，这样在传统的SDH或WDM设备中可提供以太网接口100/1G/10Gbps
  - 采用1套EOS设备，放在现有端到端的SDH/SONET(或DWDM)与以太网二层或三层交换机之间，实现端到端的传输连接
  - 在现有的以太网二层或三层交换机上增加EOS(STM-1/STM-16)接口

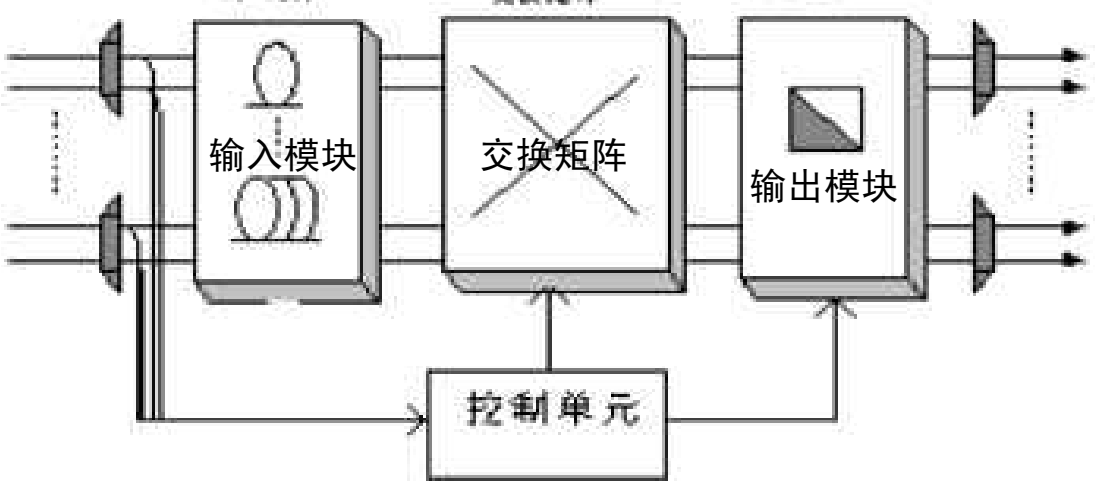
之四：全光网

- 全光网络：指光信息流在网络中的传输及交换始终以光的形式实现，而不需要经过光/电、电/光变换。
- 1980来近30年间，随着光器件的发展和光系统的演进，光传输系统的容量已从 Mbit/s发展到Tbit/s，提高了近10万倍。200光纤/光缆；10T/光纤
- 波分复用使波长本身成为组网（分插、交换、路由）的资源。逐步成熟的光分插复用（OADM）和光交叉联接（OXC）技术，只提供带宽传输的光层开始有组网能力。

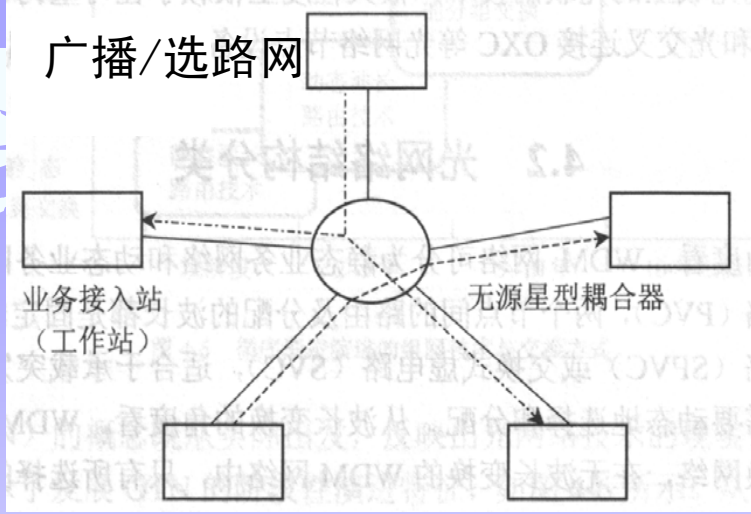


光网络发展进程

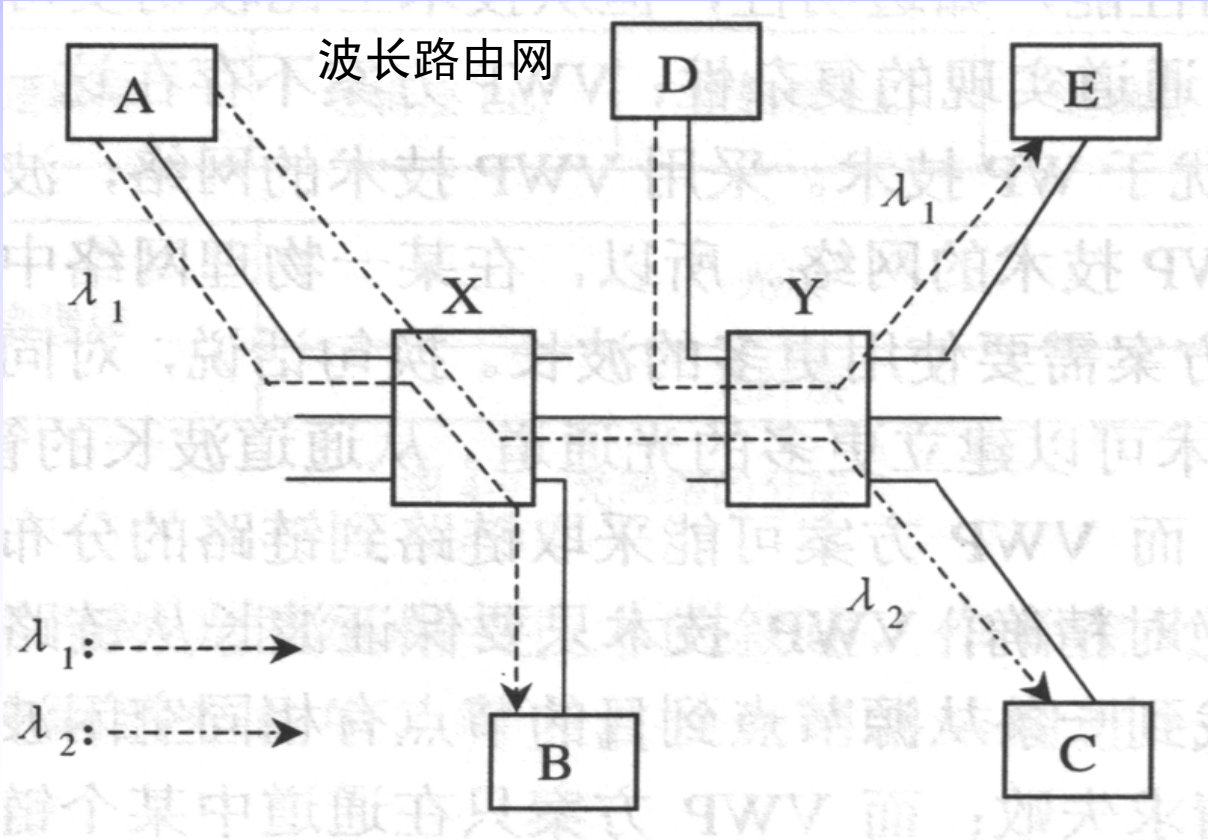

Single Fiber Capacity
Gbps
1,400
1,200
1,000
800
600
400
200
0
Approaching infinite
backbone bandwidth
at little or no additional cost
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999

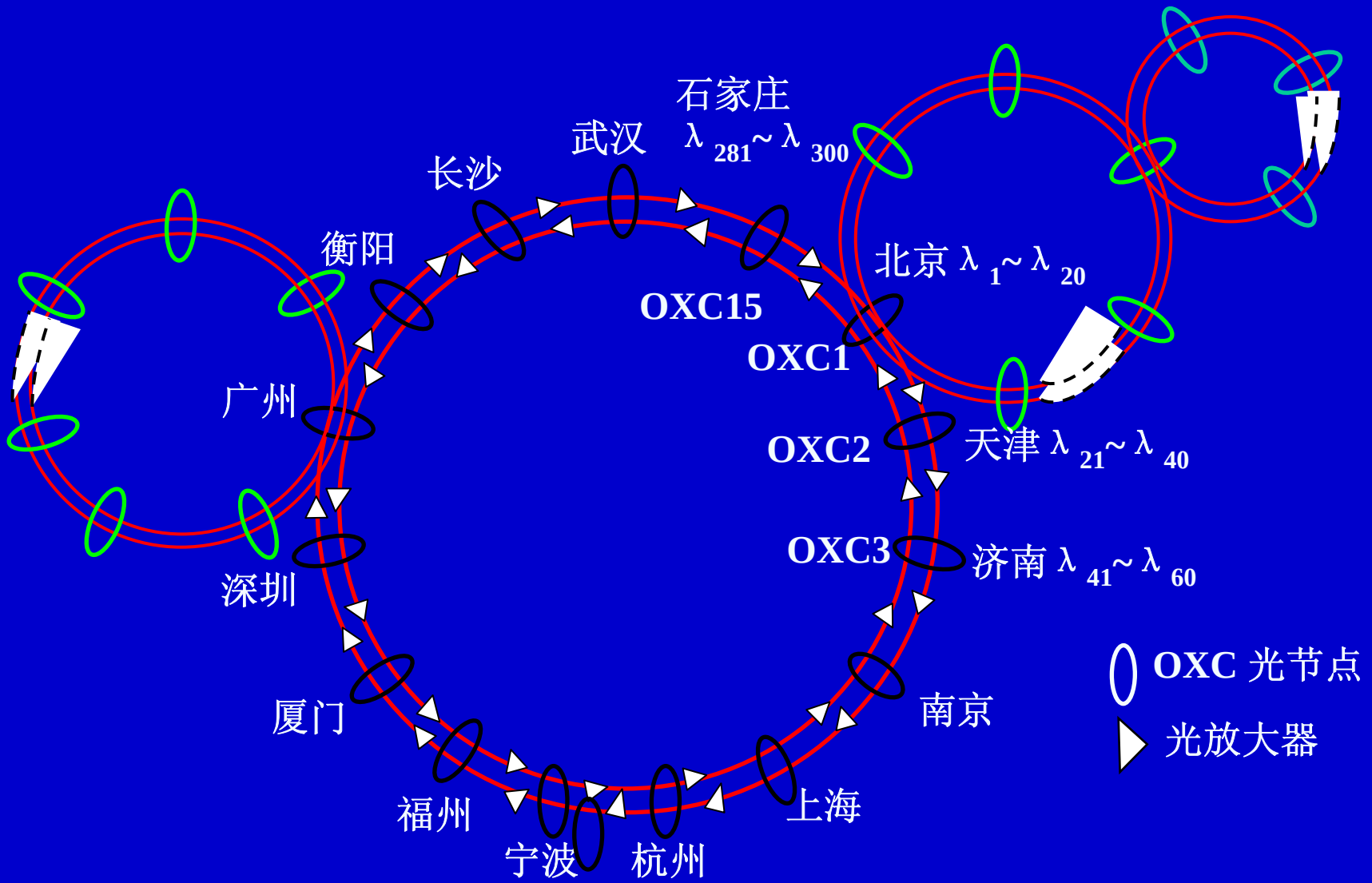


光交换系统构成



- 空分光交换网络
  - 光纤型空分光交换
  - 自由空间光交换
- 时分光交换网络
- 波分光交换网络
- 复合光交换网络
- 混合型光交换网
- 热光交换技术
- 液晶光交换技术
- 声光交换技术
- 微机电光交换技术





设想可变波长全光交换网

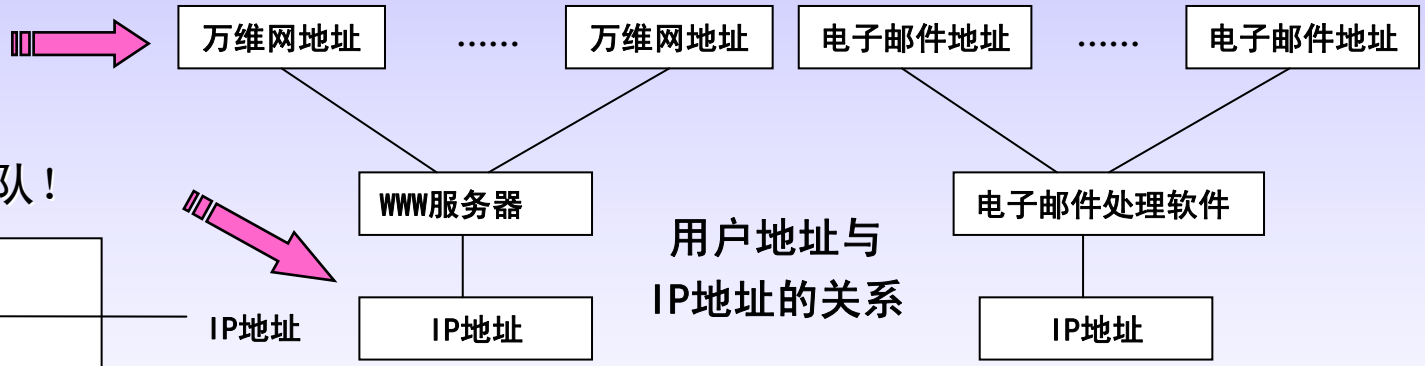
1.3.2 计算机的编址与解析

- 因特网是全球单一可寻址的抽象网络
  - 必须连接全球所有计算机或其它设备
  - 必须给每个计算机或设备（路由器等）一个全球唯一ID
  - IP协议希望这样，故有IPV4 = 2^{32}个地址，A/D/C/E类
  - IP 地址结构 = Net号 + Host号
- ICANN（Internet Corporation for Assigned Names and Numbers）因特网名字与号码指派机构
  - 中国向APNIC(Asia Pacific Network Information Center)申请IP地址
- IP地址引发问题（优点：统一互联、点分10进制、书写方便）
  - **不便记忆：→域名产生→解析需要**
  - 局域网早先出现：导致 Mac←→IP变换→ARP协议
  - IP地址分配不均不够：NAT，三个段公用私网地址10/8:1A+172.16.—172.31/16:15B+192.168.—192.168/16:1B = 1762.1775万个

因特网的三地址

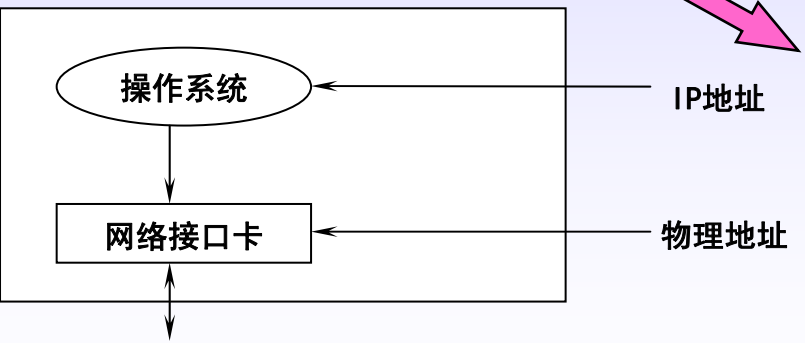
- 中国古人：姓+名+字+号
- 用户、网络及物理地址
  - 用户地址：公司/机关/团体/个人注册的因特网可访问的世界唯一的ID:URL地址
  - 网络(接口)地址：同一体系结构中的可访问的计算机ID:IP地址
  - 物理地址：同一体系结构中物理媒体可访问的计算机某端口的唯一ID:MAC地址

URL:亲爱的姚明，我是你的Fans？！



IP：NBA 火箭队！

用户地址与IP地址的关系



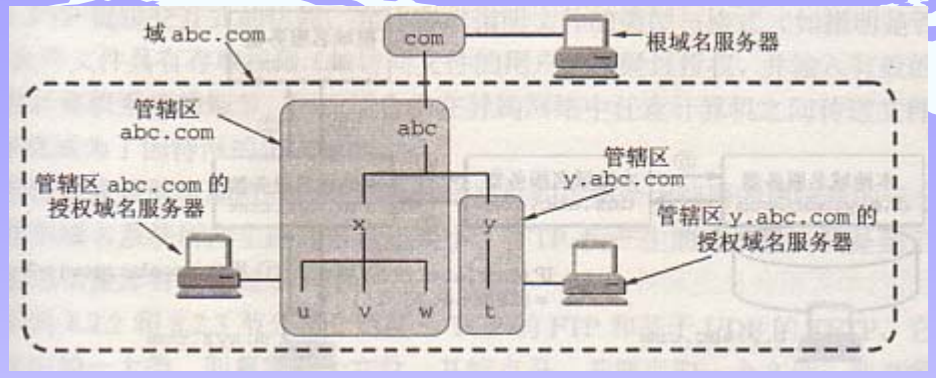
MAC：美国休斯敦亚历山大大街888号

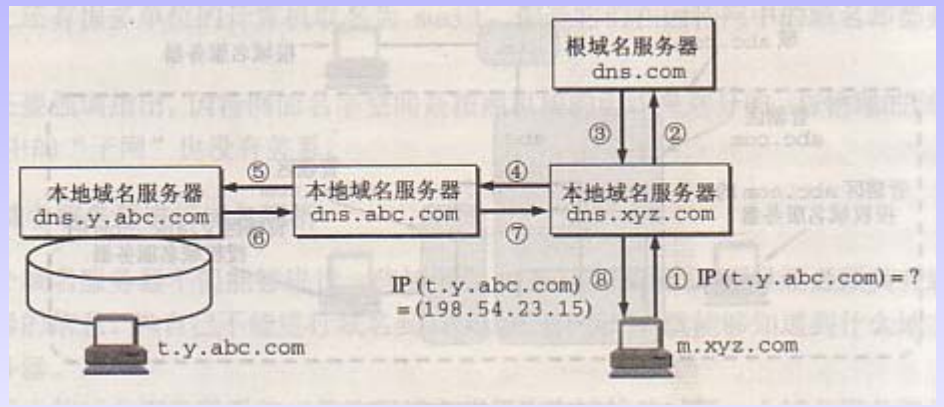
IP地址与物理地址的关系

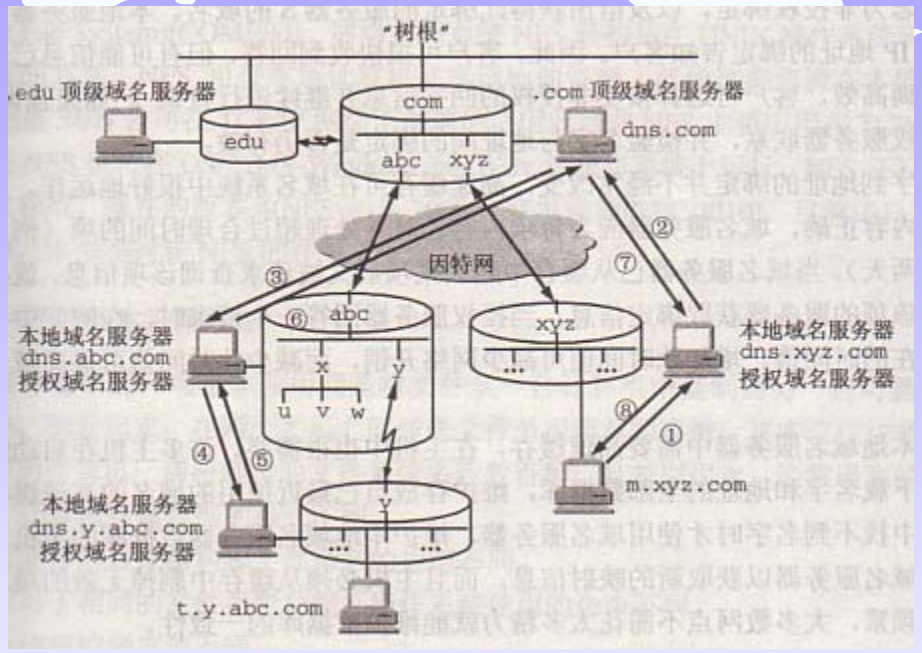
物理地址的配置和作用

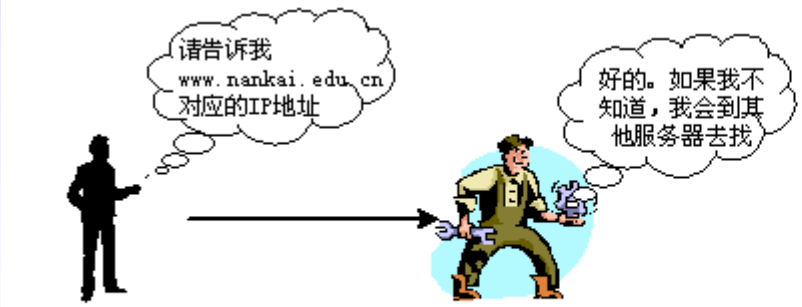
- 作用：
  - 过滤：指明网卡，过滤不属于本主机接收的数据
  - 广播：识别广播地址标志，接收广播报文并送主机
- 地址配置：
  - 固定方式：物理地址由厂商决定。不可改变
  - 可配置方式：通过EPROM内程序进行交互命令配置
  - 动态配置方式：可自动修改和管理物理地址，启动网络时，通过程序配置一个不冲突的物理地址

域名解析

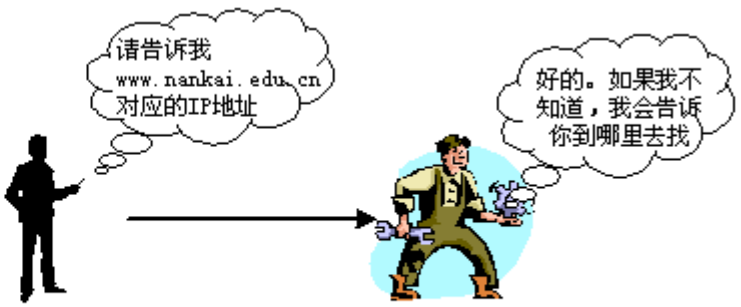








递归查询



迭代查询

域名技术

- 全球共有13台根域名服务器。这13台根域名服务器中名字分别为“A”至“M”，其中10台设置在美国，另外各有一台设置于英国、瑞典和日本。下表是这些机器的管理单位、设置地点及最新的IP地址：

名称 管理单位及设置地点 IP地址

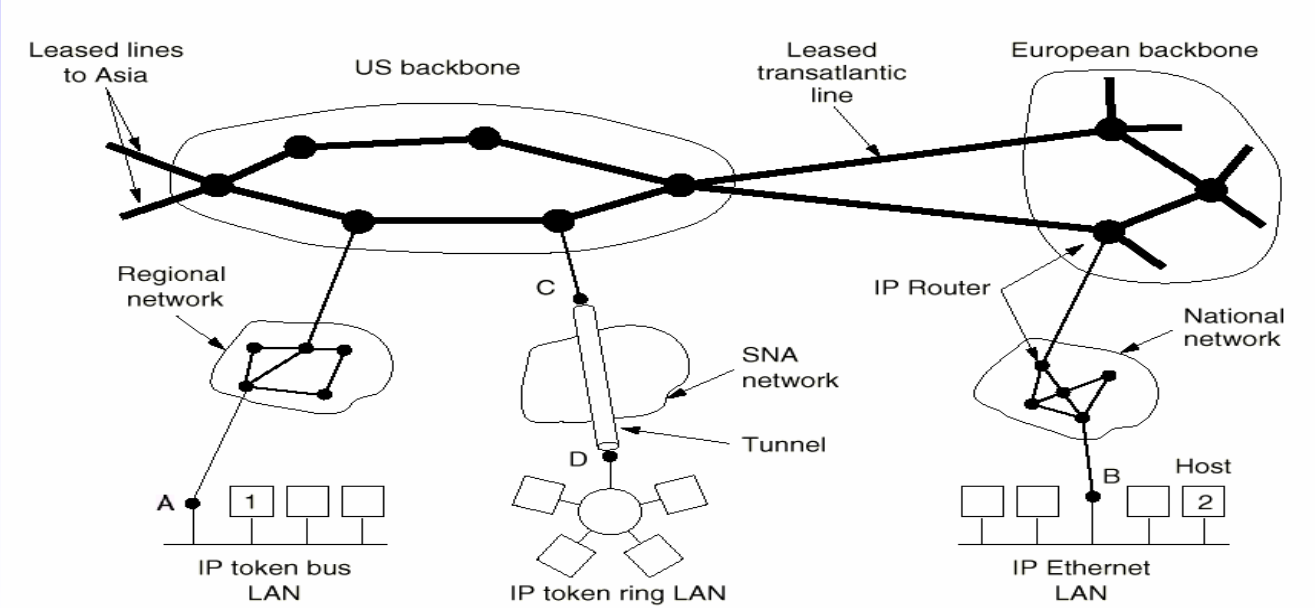
A INTERNIC.NET（美国，弗吉尼亚州） 198.41.0.4
B 美国信息科学研究所（美国，加利弗尼亚州） 128.9.0.107
C PSINet公司（美国，弗吉尼亚州） 192.33.4.12
D 马里兰大学（美国马里兰州） 128.8.10.90
E 美国航空航天管理局（美国加利弗尼亚州） 192.203.230.10
F 因特网软件联盟（美国加利弗尼亚州） 192.5.5.241
G 美国国防部网络信息中心（美国弗吉尼亚州） 192.112.36.4
H 美国陆军研究所（美国马里兰州） 128.63.2.53
I Autonomica公司（瑞典，斯德哥尔摩） 192.36.148.17
J VeriSign公司（美国，弗吉尼亚州） 192.58.128.30
K RIPE NCC（英国，伦敦） 193.0.14.129
L IANA （美国，弗吉尼亚州） 198.32.64.12
M WIDE Project（日本，东京） 202.12.27.33

1.3.3 网络互连

- 问题：多个独立网络怎么办？
- 多个网络互连的2个主要问题：
  - 异构：
    - 以太网、令牌环、点-点链路及各种交换网络，每个都有自己的地址模式、媒体访问协议及服务模式等
    - 传输介质不同/网络拓扑结构不同
    - 介质访问方式不同/网络编址方式不同
    - 分组长度/有连接/无连接服务的区别
    - 传输控制方式不同
    - 各层协议的功能定义、格式、接口与调用方式不同
  - 规模：
    - 因特网不断扩大，面临许多挑战：路由（100万--10亿个节点）、寻址、编址

Internet网络层协议

- 在网络层，Internet可以看成是**自治系统**的集合，是由**网络组成的网络**。
- 网络之间互连的纽带是**IP**（Internet Protocol）协议。



网络互连层的功能

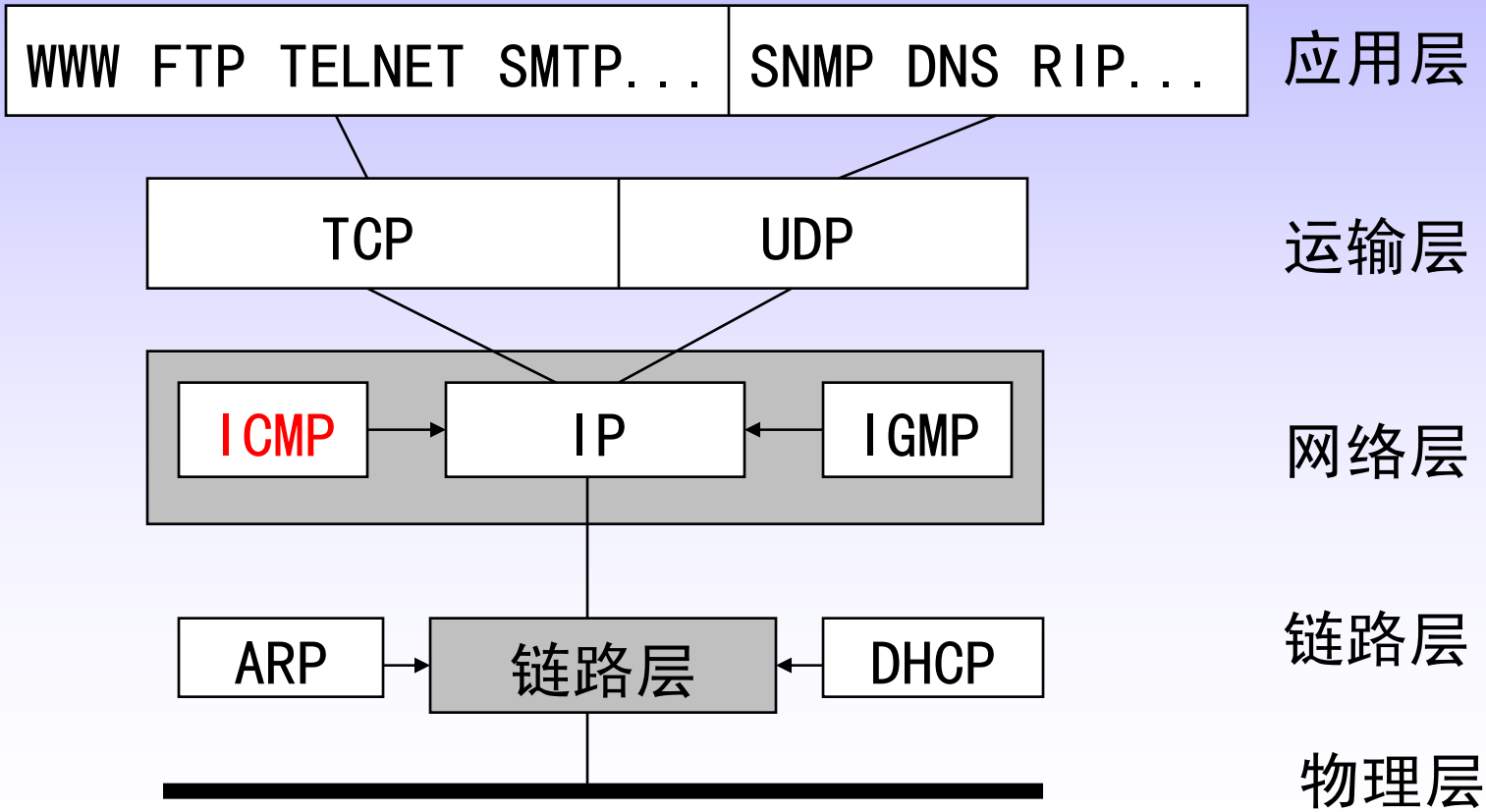
- 功能目标
  - 把多个异构或同构的网络**连接**成更大网络
  - 直接支持传输层的**端到端**服务。
- 关键问题
  - **了解**通信子网的拓扑**结构**，
  - 选择**路由**。
- 为传输层提供何种服务？
  - 面向连接服务：
    - 将复杂的功能放在**网络层**
    - 传统电信观点：通信子网应该提供可靠的、面向连接的服务
  - 无连接服务：将复杂的功能放在**传输层**
    - Internet的观点：通信子网无论怎么设计**都是不可靠**的，因此网络层**只需提供无连接**服务。

- 网络层的内部组织
  - 虚电路（virtual circuit）
  - 数据报（datagram）
- 虚电路子网与数据报子网的比较
  - 路由器内存空间与带宽的权衡
    - 虚电路方式，路由器需要维护虚电路的状态信息；
    - 数据报方式，每个数据报都携带完整的目的/源地址，浪费带宽
  - 连接建立时间与地址查找时间的权衡
    - 虚电路需要在建立连接时花费时间
    - 数据报则在每次路由时过程复杂
  - 服务质量QoS（Quality of Service）权衡
    - 电路方式很容易保证服务质量适用于实时操作，但比较脆弱。
    - 数据报不太容易保证服务质量，但是对于通信线路的故障，适应性很强。

- 网络互连设备
  - 中继器（repeater）
    - **物理层**互连，在电缆段之间拷贝比特；
    - 对弱信号进行放大或再生，以便延长传输距离。
  - 网桥（bridge）
    - **数据链路层**互连，在局域网之间存储转发帧；
    - 网桥可以改变帧格式。
  - 多协议路由器（multiprotocol router）
    - **网络层**互连，在网络之间存储转发包；
    - 必要时，做网络层协议转换。
  - 传输网关（transport gateway）
    - **传输层**互连，在传输层转发字节流。
  - 应用网关（application gateway）
    - **应用层**互连，在应用层实现；
    - half-gateway：满足不同国家、组织的管理需要

错误报告协议(ICMP)

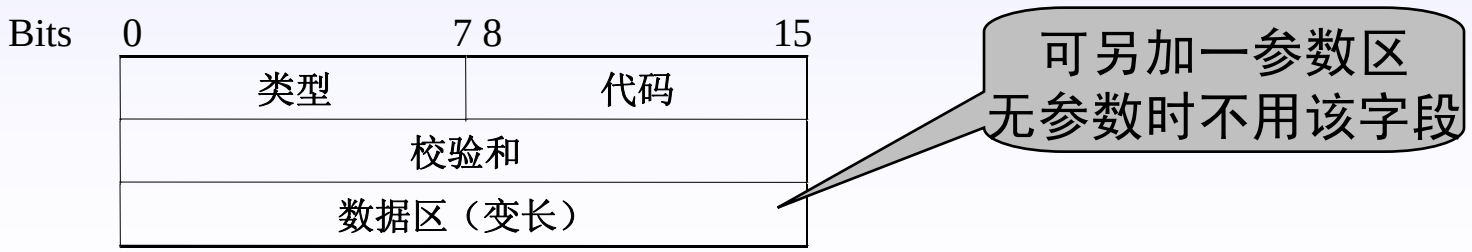
- 协议层次的回顾



ICMP协议

ICMP 报文分类

- 差错和控制报文协议：功能=报告IP传输中发生的（差错报文+控制报文+测试报文）
- ICMP报文=头部+数据部分
- ICMP**报文封装在IP数据包中进行传输**，IP头中的包类型=1。ICMP并不是IP的上一层协议，仅用IP的转发功能



ICMP报文格式

1.3.4 路由选择与算法

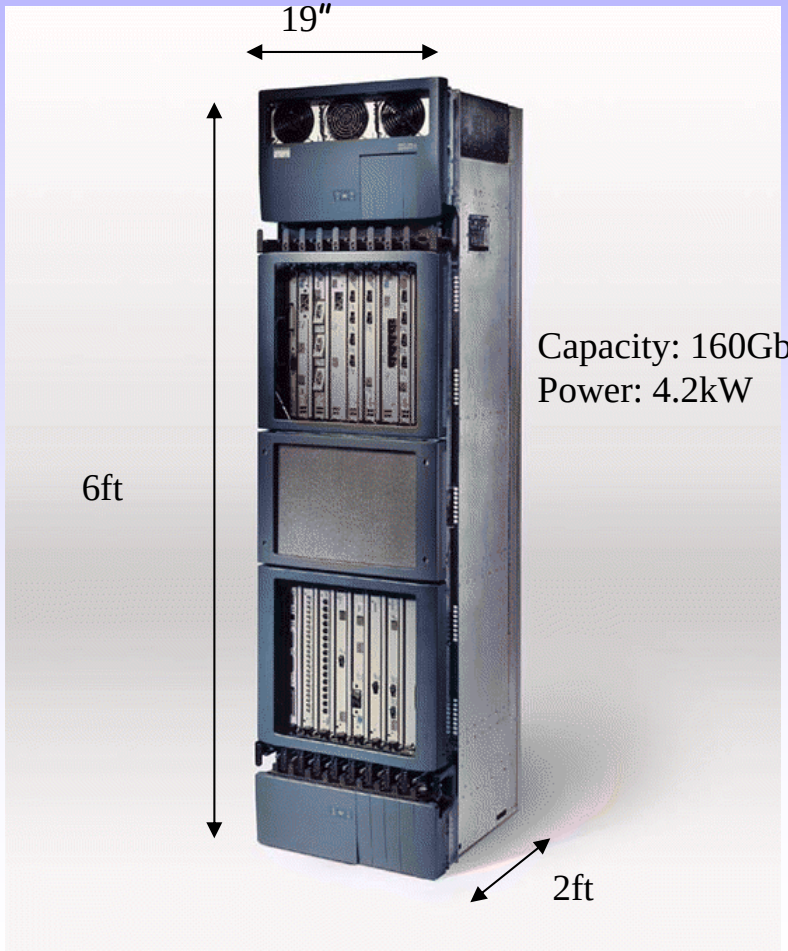
- So far 假设交换机或路由器完全知道网络拓扑，并能选择每个包出端口
  - 对虚电路网：仅连接请求包需要路由，所有后继包走连接请求相同的通道
  - 对数据报网（含IP网）：每个包都需要路由
  - 不论那种情况，交换机或路由器都需要查看和比较转发表和包的目的地址，并选择最佳出端口
- 路由的基本问题：**交换机和路由器怎样获得转发表中的信息**

路由器和路由协议

- 路由器：网络层互连设备/丰富灵活
- 静态路由：手工配置/简单直观，但网络规模大时，管理负担过重
- 动态路由：由路由协议动态更新
- R上的路由协议相互主动交换路由信息建立完整的路由表，并据此转发数据包

What a Router Looks Like

Cisco GSR 12416



Capacity: 160Gb/s
Power: 4.2kW

Juniper M160

Capacity: 80Gb/s
Power: 2.6kW

区别转发表和路由表

- 转发:取一个包,查看其目地址,参考转发表,把包发送到表指定的方向.是一个相对容易定义在局部节点上运行的进程
- 路由:是建立转发表的一个进程,它取决于复杂的分布式、连续评估网络全部历史的算法

转发表和路由表的定义

- 转发表和路由表有时互通，这里加以界定并定义
- 转发表：转发包时使用，它必须包含完成转发功能所需的足够信息，意味着，表中的每行由网号到输出接口和一些MAC地址的影射信息构成。以网号为优化目的，以利快速查询。可能由特殊硬件来实现
- 路由表：由路由算法建立，是转发表的前身，包含由网号到下一跳的影射，可还含怎样学习获得这些信息，使能决定何时丢掉某些信息。以拓扑变化后的优化计算为目的。

转发表和路由表例子

路由表

Network Number	Next Hop
10	171.69.245.10

转发表

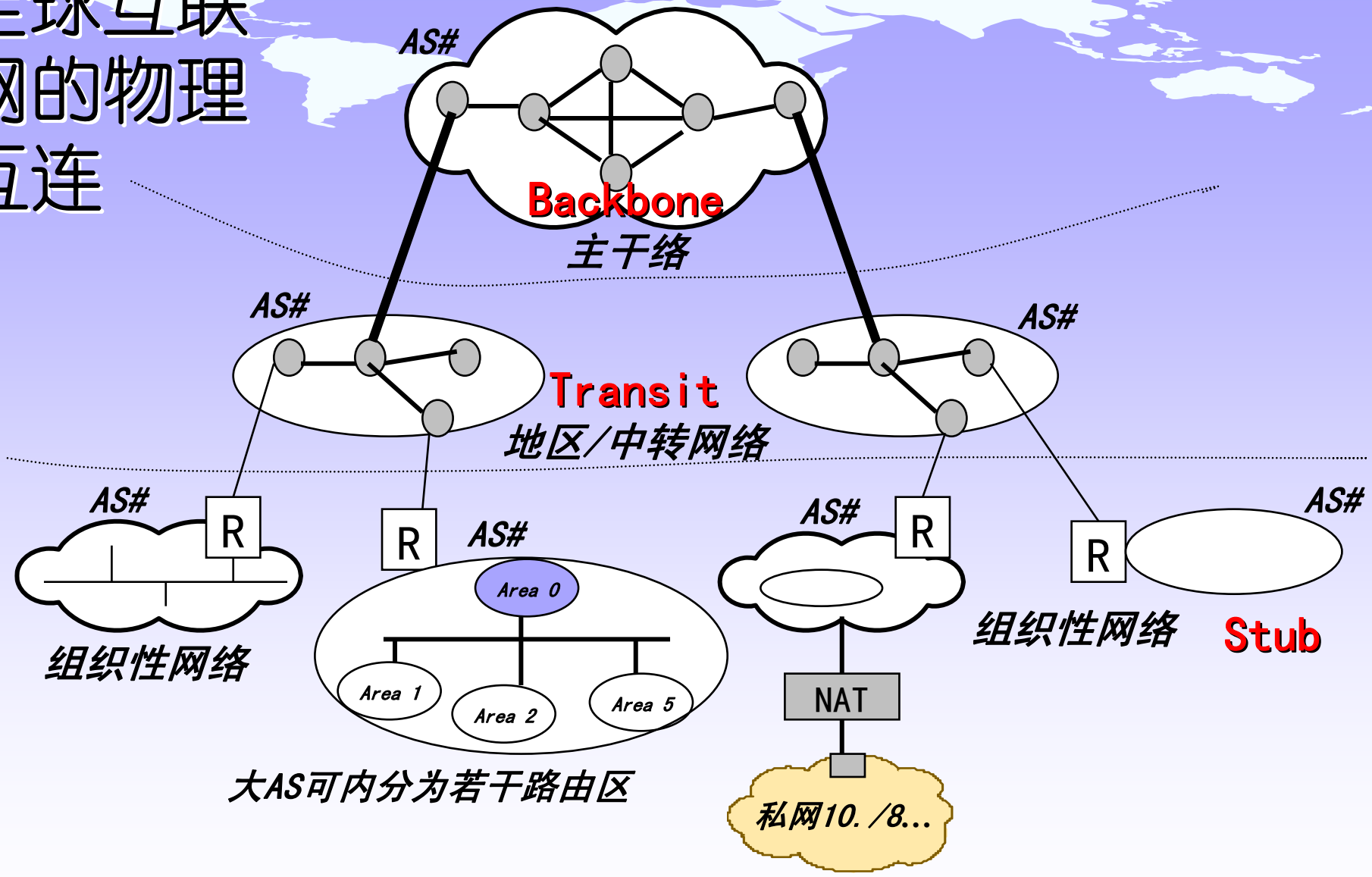
Network Number	Interface	MAC Address
10	ifo	8:0:2b:e4:b:1

- 路由的关键问题：
  - 任何时候都要为因特网建立机制，解的规模有多大？实践中不大于一百个节点
- 因特网采用分层结构，减低解的复杂性

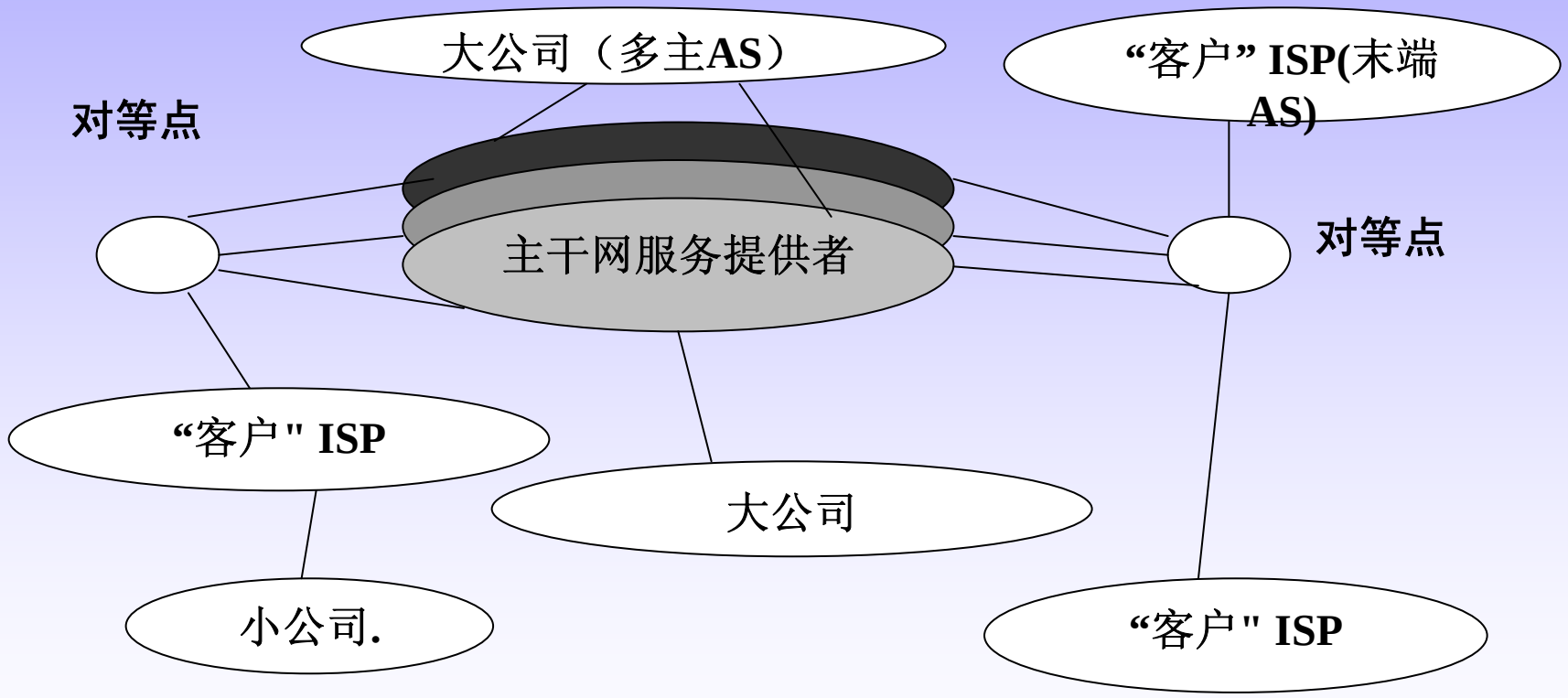
路由协议

- 因特网使用的路由协议分为两类：
  - 内部网关协议IGP：在一个AS内运行
  - 外部网关协议EGP：在AS间运行
- 路由算法分为两大类
  - 距离向量（D-V）算法：如RIP采用
  - 链路状态（L-S）算法：如OSPF采用
- AS：在一个权威管理机构下运行的网络，唯一编号以区别不同AS。整个因特网是由多个AS组成

全球互联网的物理互连


AS#
Backbone
主干络
AS#
AS#
Transit
地区/中转网络
AS#
R
R
AS#
Area 0
Area 1
Area 2
Area 5
AS#
R
NAT
R
AS#
组织性网络
组织性网络
Stub
大AS可内分为若干路由区
私网10./8...

非树复杂互连



划分AS域的基本思想

- AS:独立的**管理区**
  - **具有独立的选路策略；**
  - 运行独立的IGP；
  - AS间通过BGP交换路由信息。
  - 另一个名字--路由域
- AS基本思想:
  - 减轻域内路由协议的路由任务
  - 分级合并大网路由信息，提高了扩展性
- 大规模路由问题可分成两部分:
  - AS内路由--域内路由
  - AS间路由--域间路由

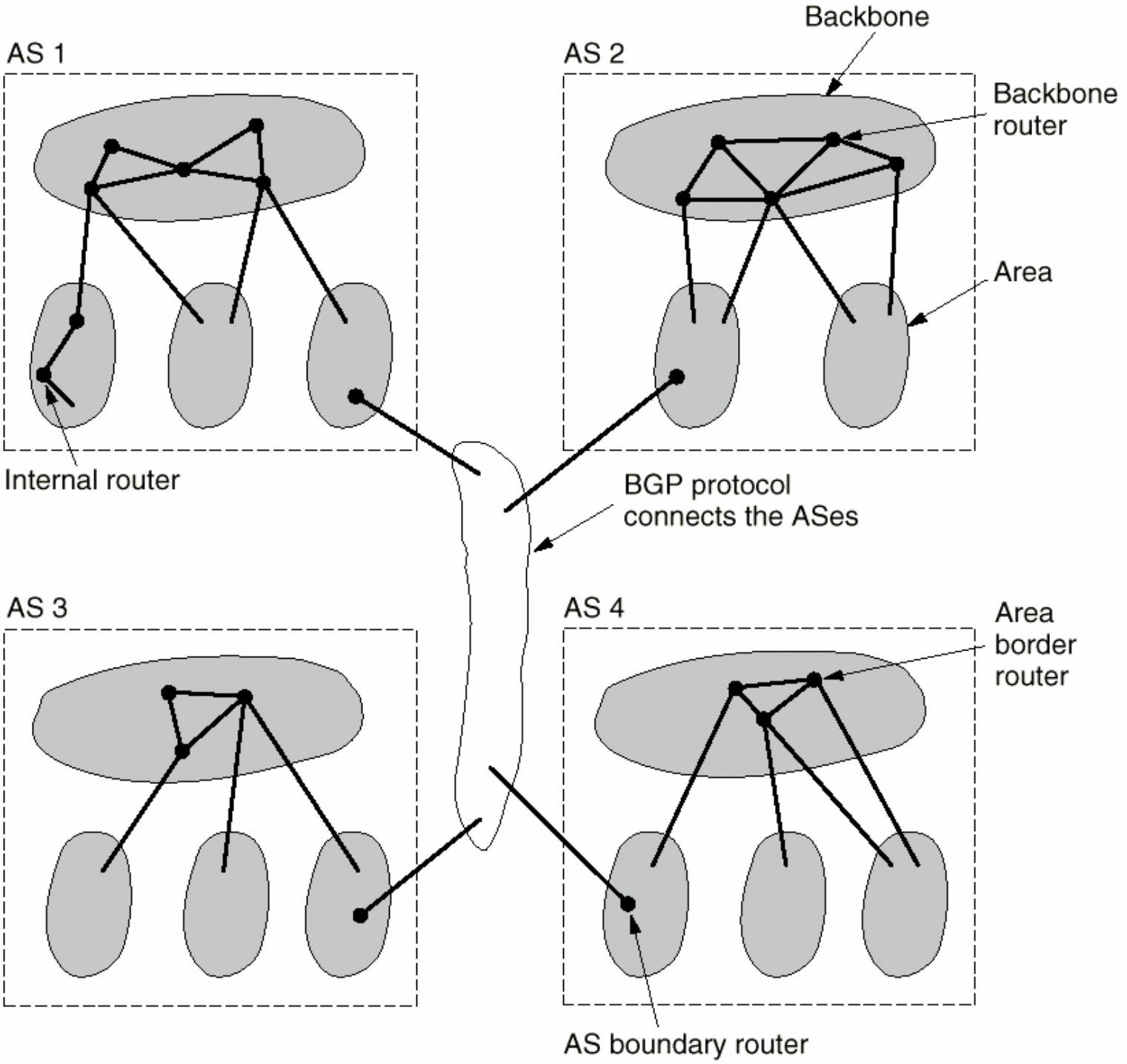
网络的分层路由

- 全球物理互连网 = AS1+AS2+AS3...;
- 一AS域内可划分若干个子域：区（areas）
  - AS = Area0(主干) + Area1 + Area2... ...
  - 区:由管理者配置为可彼此交换链路状态信息的路由器集合.主干网是一个特别区--Area 0
  - 通常存在三种路由域并使用相应路由协议
    - 同一AS内
      - 不分区
      - 分区
      - 使用协议 IGP = RIP,OSPF,IBGP... ...
    - 不同AS间
      - BGP4，任意结构
      - 早期EGP，只能针对树结构

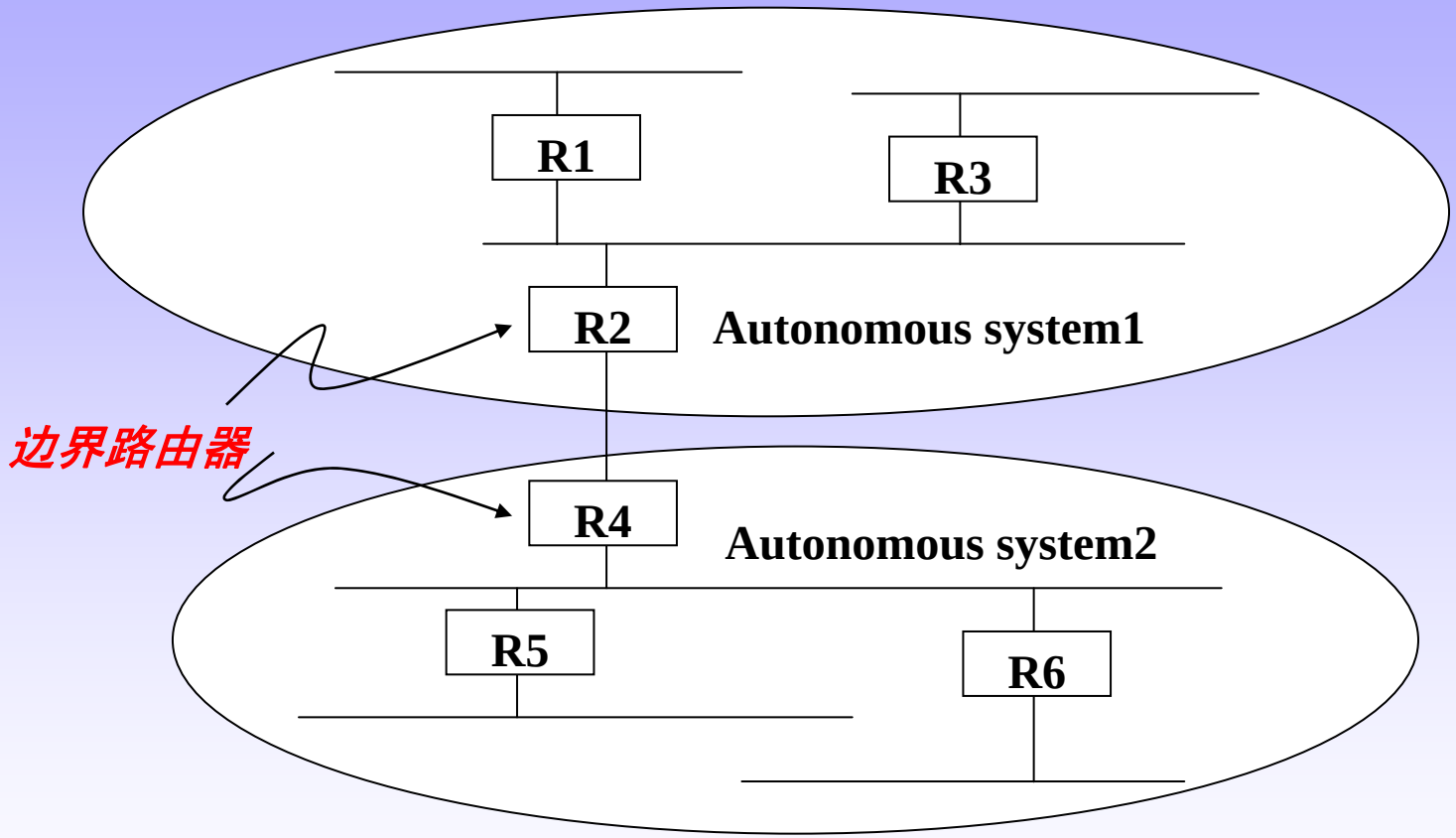
四类路由器，允许重叠

- 完全在一个区域内的内部路由器；
- 连接多个区域的区域边界路由器；
- 主干路由器；
- 自治系统边界路由器。
- AS网关路由器的任务（由BGP4完成）
  - 从相邻AS获得可达性信息
  - 传播对本AS中所有路由器的可达性信息

AS之间，OSPF域和主干的关系

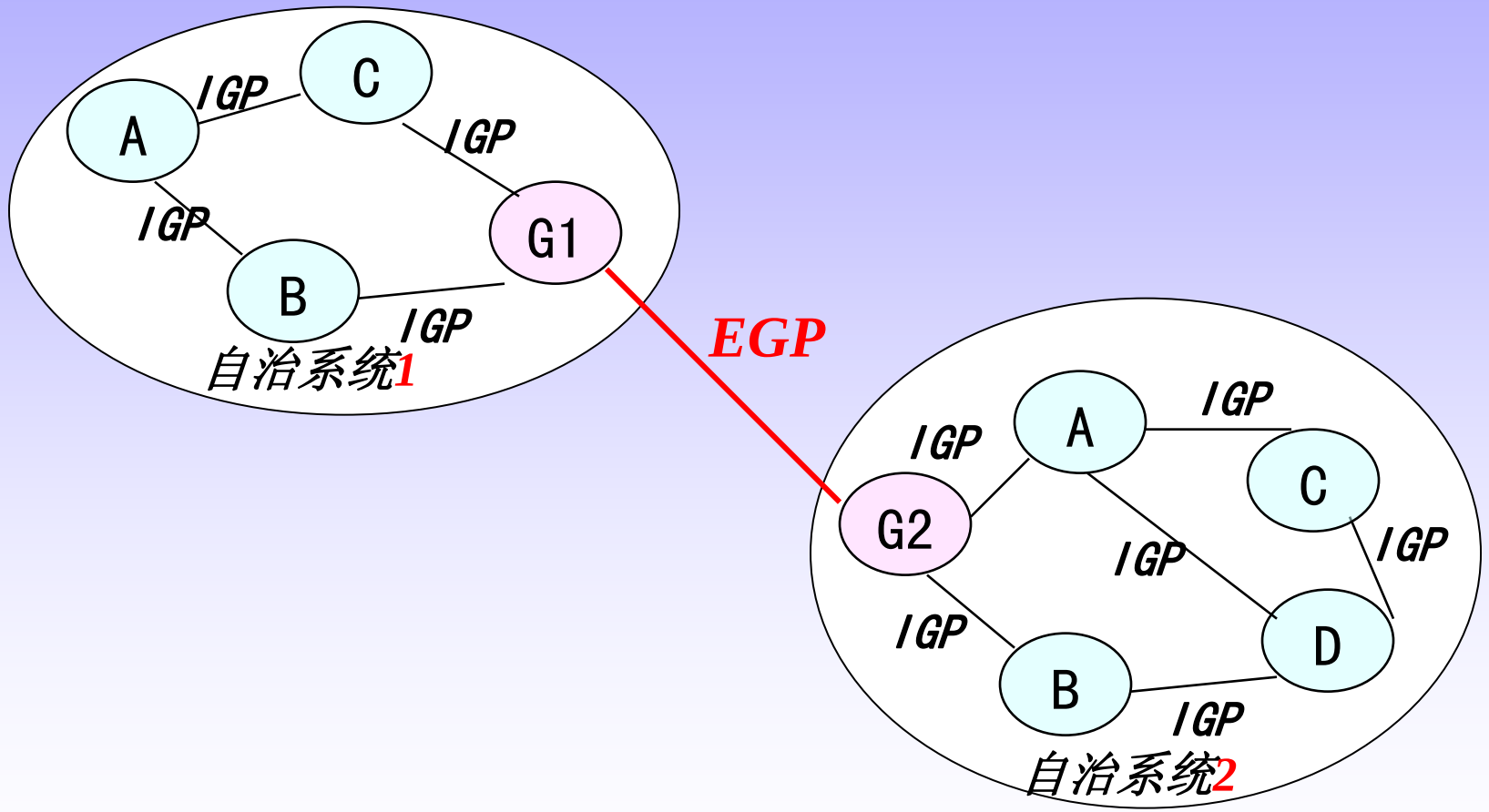


域间路由(BGP)

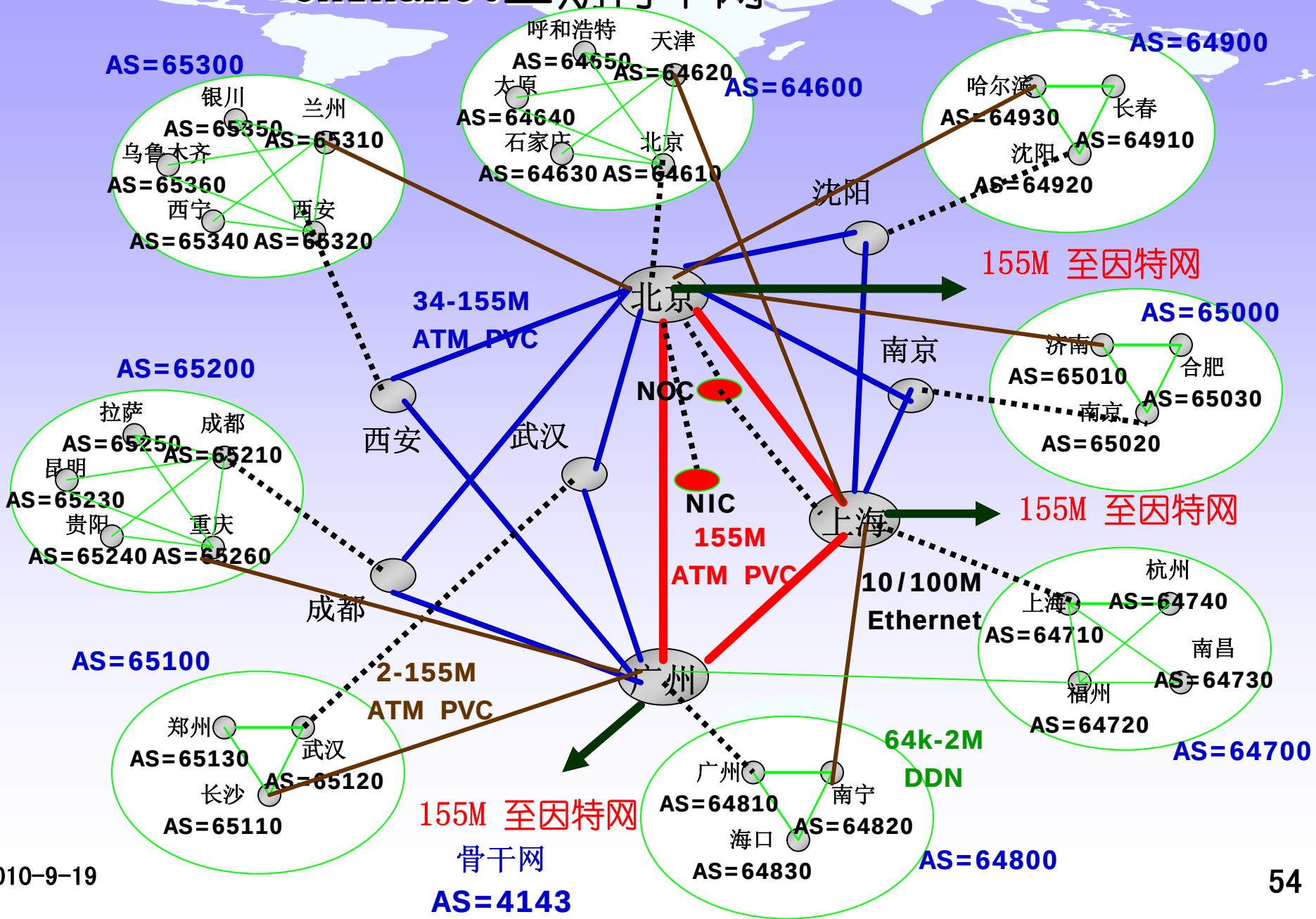


- 2个网络在一个自治域中
- AS:在单个管理实体控制下的网络系统

外部网关与内部网关

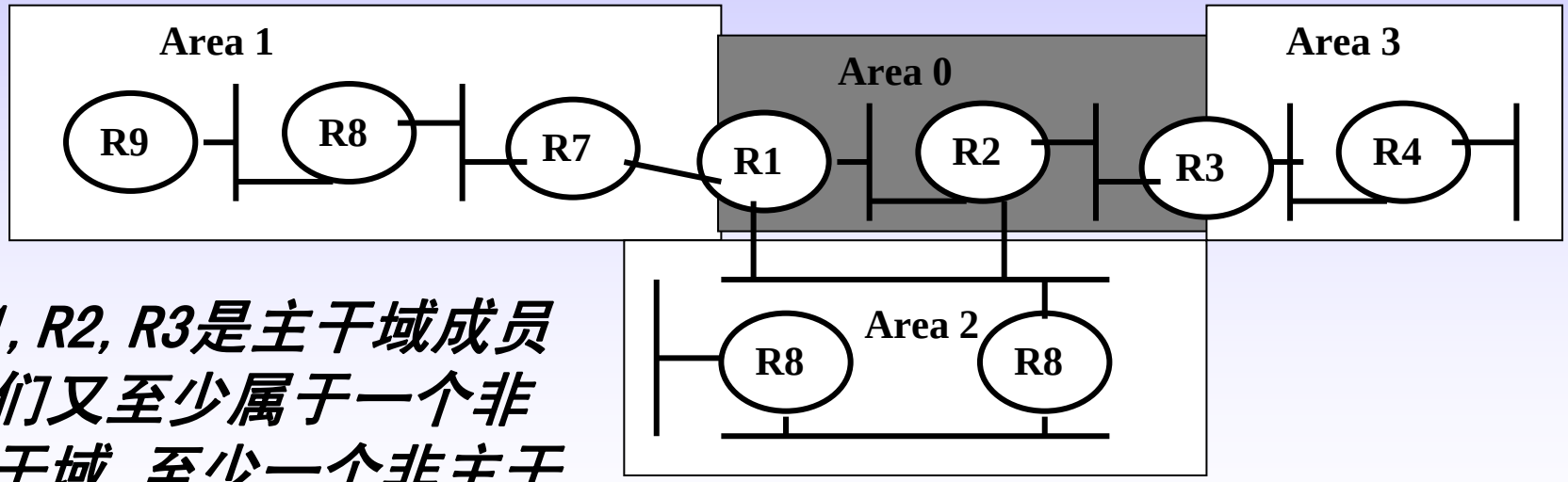

C
A
IGP
IGP
IGP
G1
B
IGP
自治系统1
EGP
IGP
A
IGP
C
G2
IGP
IGP
IGP
D
B
IGP
自治系统2

Chinanet二期骨干网



路由区

- 象OSPF一样,把域分成许多区,未过载域内路由协议就能分层,使域增大
- 区:管理上配置成和其它节点交换链路状态信息的路由器集. 主干网是一个特别区--Area 0



- *R1, R2, R3是主干域成员它们又至少属于一个非主干域, 至少一个非主干域的区边界路由器ABR*

1.3.5 数据运输

- 问题:怎样实现远程进程间通信
  - 主机-主机的包传输转化成进程-进程通信通道
  - 网络层结构，支持端应用程序--端到端协议
- **什么是连接?**
  - 一条连接就是不同系统内的两个实体之间的一个临时性的逻辑关联通路。
  - 在连接持续期间，每个实体都跟踪从对方到达和发送到对方的PDU，以便调节PDU的流量以及对丢失和损坏的PDU进行恢复。
- 互联网的全部功能，最基本、最小粒度的服务
  - **端到端数据传输**

对传输层协议的希望与IP层现实

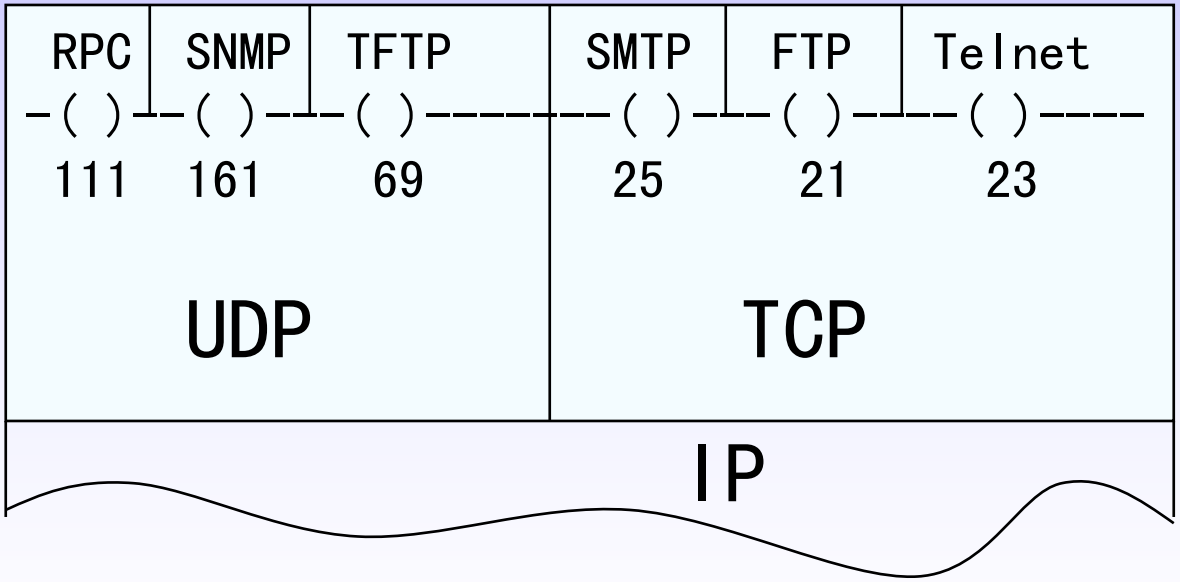
- 希望
  - 保障报文传输
  - 以发送相同的顺序传输报文
  - 每个报文最多传输一个拷贝
  - 支持任意长报文
  - 支持收、发之间的同步
  - 允许收方应用流控发方
  - 支持每个主机上的多个进程
- 现实（IP层提供的服务）
  - 丢包
  - 报文重排序
  - 对给定报文传输重复拷贝
  - 限制报文在某个有限大小
  - 在任意长延迟后传输报文
  - 以上是best-effort 层次上的饿服务，如IP

I）简单多路器(UDP)

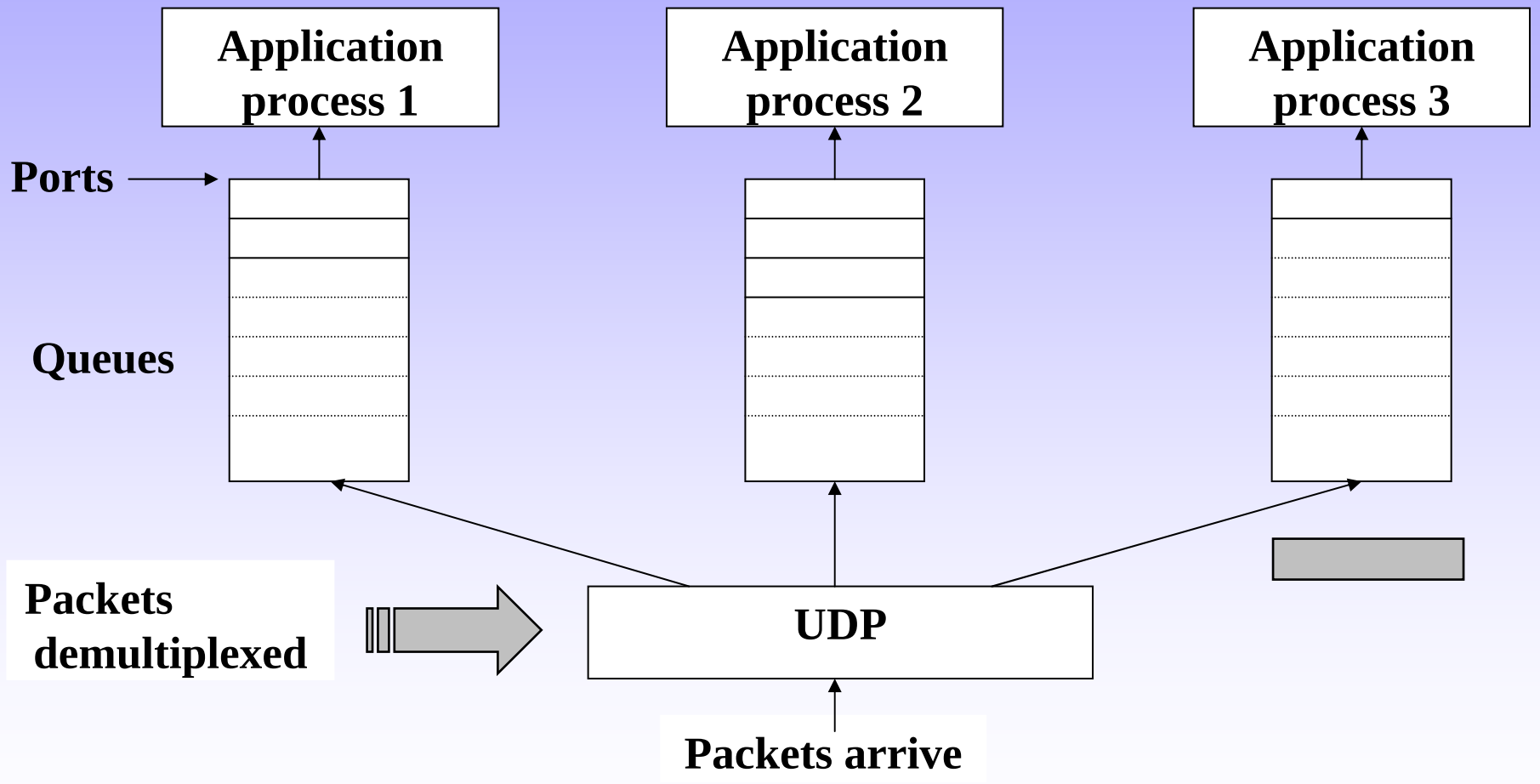
- 最简单的传输层协议是:把主机-主机的传输服务在IP协议基础上扩展成进程-进程的通信服务
- 在一台主机上可能运行多个进程,故协议需要加一多路开关层,使它们共享网络
- 传输层协议对best-effort未加任何功能.如User Datagram Protocol-UDP
- UDP: bare minimum
  - just port numbers, and an optional checksum
  - no flow control, no congestion control, no reliability or ordering

端口的概念

- 唯一区别全网上的每个进程或目的主机上的进程
  - 端口:间接区别每个进程的抽象定位器-数字
  - 唯一标识=主机IP地址 + 端口号



UDP消息队列


Application process 1
Application process 2
Application process 3
Ports
Queues
Packets demultiplexed
UDP
Packets arrive

UDP协议

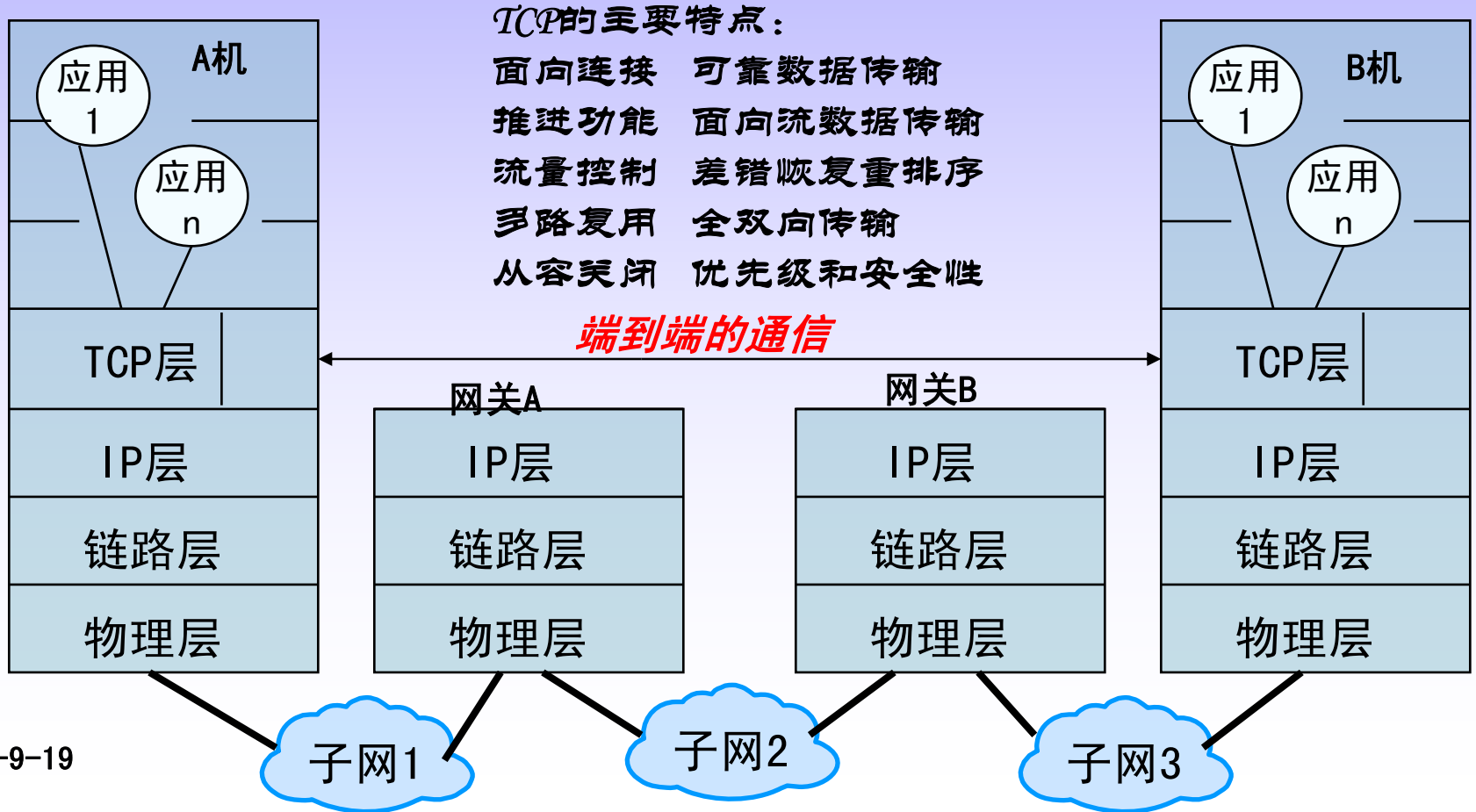
- 提供无连接服务，不保证数据完整到达目的地，减轻了网络的通信负担
- 适应C/S模式的简单请求/响应通信需要
- 应用程序要**实施超时重传机制**，并对数据包编号，但增加了应用程序的复杂性
- UDP保留各报文间的边界，不把应用多次发送的数据合并成一个包发出去，**且发包后不对该包缓存**，这对简单请求/响应很方便
- 需要**组播**的应用、多数音视频都建立在UDP之上。

II) 可靠字节流协议(TCP)

- UDP只是简单的分路协议
- TCP:更成熟的传输协议
  - 提供可靠,面向连接,按序字节流
  - 全双工,每个连接支持一对字节流,每个流一个方向
  - 流控机制:允许每个字节流的接收端在给定时间内限制其发送端的数据速率
  - 支持多路输出机制,允许一个主机上同时有多个会话对
  - 还提供拥塞控制机制
- 流控与拥控之差别:
  - 流控:防止发送者超过接收者的能力,流控是端到端的发射
  - 拥塞控制:防止太多的数据注入到网络中,从而引起交换机或链路超载,拥塞控制是关于主机和网络的关系

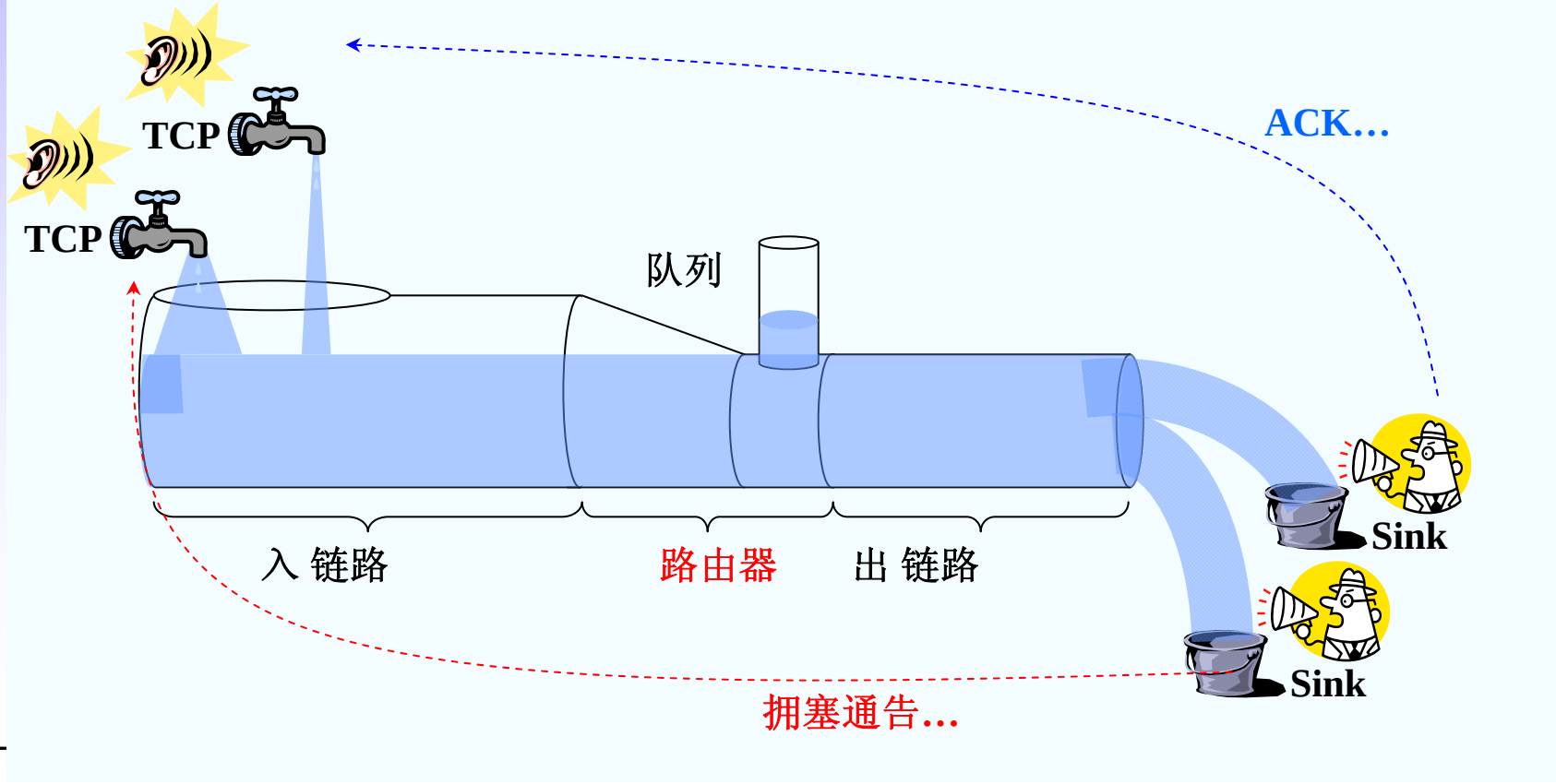
TCP层次与功能

- 处于应用层和网络层之间，实现端到端Peer to Peer的通信：在H上执行，屏蔽下层的服务质量差
- 对上层提供面向连接、端到端可靠通信服务：先连接后传输、任一方可断连、点点全双工，即两个方向同时传输数据，但不能组播



TCP friendly

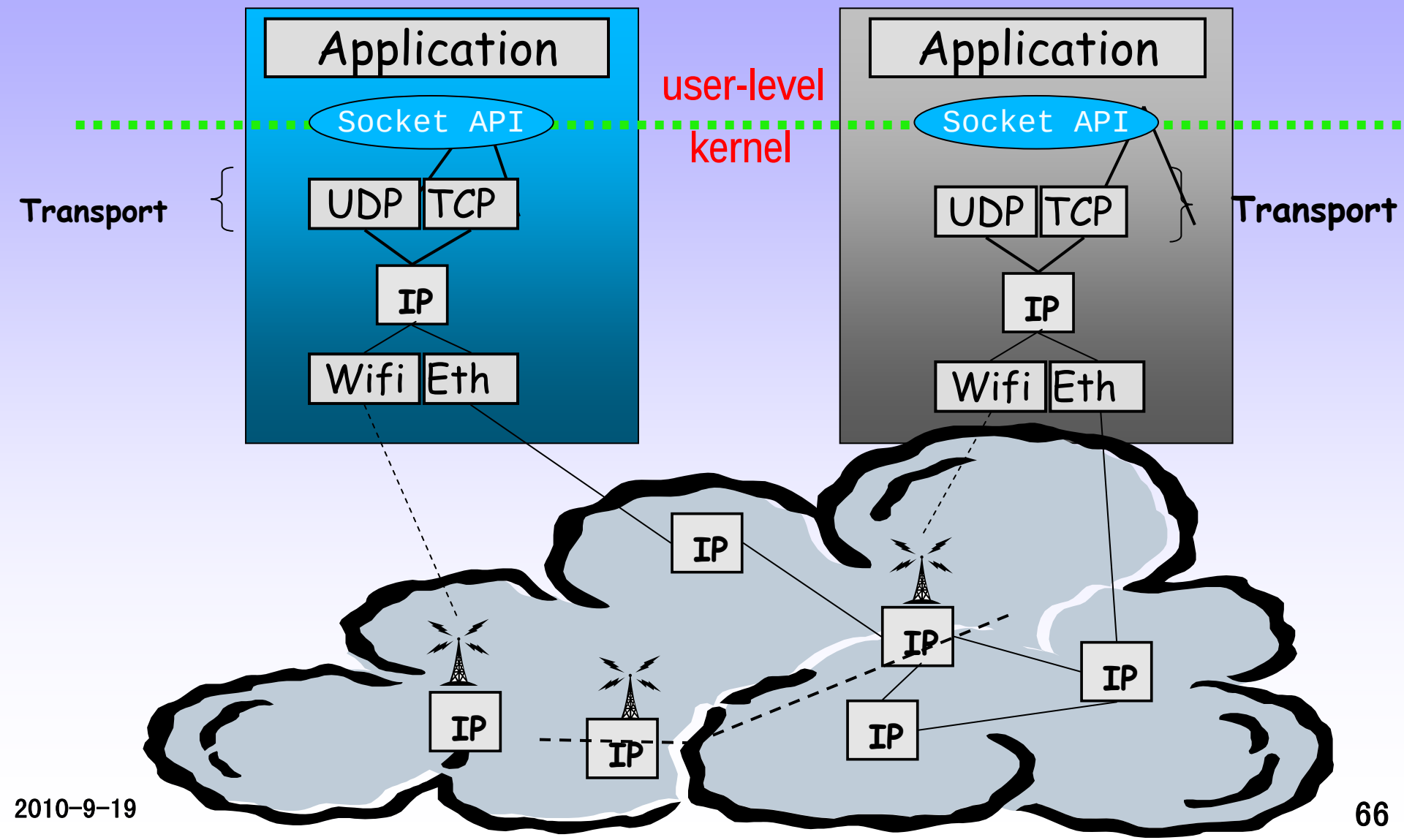
- TCP: a package deal
  - flow control, congestion control, *byte-stream orientation*
  - *total* ordering and *total* reliability



III）流控制传输协议：SCTP

- RFC 2960, Oct. 2000
  - SCTP is designed to transport PSTN signaling messages over IP networks, but is capable of broader applications.
- 为什么还需要新的传输协议：SCTP
- 考虑下列需求的特点，传统UDP/TCP很难满足
  - Multi-homing
  - Multi-streaming
  - Message boundaries (with reliability*)
  - Improved SYN-flood protection
  - Tunable parameters (Timeout, Retrans, etc.)
  - A range of reliability and order (full to partial to none)along with congestion control
  - and more...

传统UDP/TCP不能满足某些需求？



实际应用需要—信令传输

- 目前ＩＰ网中的信令消息交换通常是使用ＵＤＰ或ＴＣＰ来完成。但这两者都不能完全满足**电信网中信令承载的要求**。
- **ＵＤＰ**是基于消息的，提供快速的**无连接**业务。这使其**适合传输时延敏感**的信令消息。但是，ＵＤＰ本身仅提供**不可靠**的数据报业务。而差错控制，包括消息顺序、消息重复检测和丢失消息重传等，只能由**上层**应用来完成。
- **ＴＣＰ**虽然**提供差错和流控**，但对传输信令消息来说，却存在着诸多缺陷：
  - ＴＣＰ是面向字节流的。这意味着消息的描述必需由**应用来完成**，而且要在消息结束时**显式通知**ＴＣＰ以迫使其立即发送相应的字节数据；
  - **许多应用只需要信令消息的部分有序**，例如属于同一呼叫或同一会话的消息就是这样。而ＴＣＰ只提供严格的数据按序传输，这会导致不必要的队头阻塞并使消息的传输时延增大；
  - ＴＣＰ连接直接由一对传输地址（ＩＰ地址和端口号）识别，从而无法提供对多宿主机的透明支持；
  - 典型的ＴＣＰ实现不允许高层应用设定协议控制参数。但是一些应用可能会需要调**节传输协议的属性**以满足其特定要求，例如某些**信令协议有较高的时延**要求，而另一些信令协议则只要求较高的**可靠性**。
- 总之，UDP有边界但不可靠，TCP可靠而又无边界！**信令需要**

SCTP 更关键的特点之一

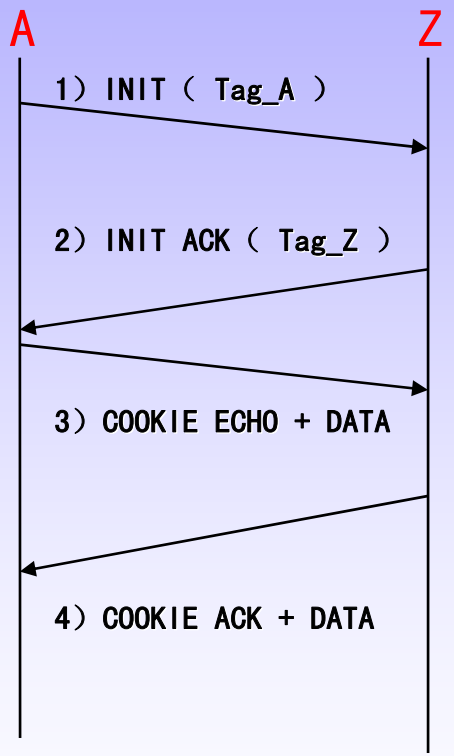
- Multi-homing *improved robustness to failures*
  - In TCP, 连接仅在 <IP addr,port> 与 <IP addr, port>之间进行
  - 对 multi-homed, 每端仍有一个IP地址可选
  - 如果接口down，**整个连接down**
  - In SCTP, 每端可列出许多 IP addresses
  - 如果某接口down ，仍可通过任何其它地址保持连接
- *Multi-streaming reduced delay*
  - 部分保序. 消减 Head of Line (HOL) 阻塞
  - In TCP, 所有数据保序; 队列头的丢失导致整个数据段延迟交付
  - In SCTP, 你可发送**多达 64K 的独立流, 每个保序流独立**
  - **某个流上的丢失并不导致其它流延迟交付**
- *Message boundaries preserved easier coding*
  - TCP 打包并不保留报文的边界
  - SCTP 保护报文边界
  - Application protocols easier to write, and application code simpler.

SCTP 更关键的特点之二

- Improved SYN-flood protection *more secure*
  - TCP 易受 SYN flooding攻击;
  - SCTP 采用四次握手，保护免受SYN flooding攻击
- Tunable parameters (Timeout, Retrans, etc.) *more flexibility*
  - TCP 参数调整只有系统管理员才能进行，实施内核的改变和锁定等
  - SCTP 参数可由socket basis调整
- Congestion controlled unreliable/unordered data *more flexibility*
  - TCP 虽有拥控，但不能做不可靠/失序的交付
  - UDP 虽能不能做不可靠/失序的交付，但没有拥控
  - SCTP 总有拥控，且能在没有—全有范围提供可靠性、保序的服务
  - SCTP, 可靠/不可靠数据都能在相同连接上多路

Normal Establishment of an Association -- four-way handshake

- “A” 发送 INIT 块到 “Z”. 块中含验证标志 Tag (Tag_A) （1 to 4294967295随机数）
- “Z” 响应 INIT ACK 块. 其中含自己的验证块 (Tag_Z)
- “A” 发送 COOKIE ECHO 块到 “Z”. 可与DATA 块绑定
- “Z” 回答 COOKIE ACK 到“A”，可与DATA 块绑定



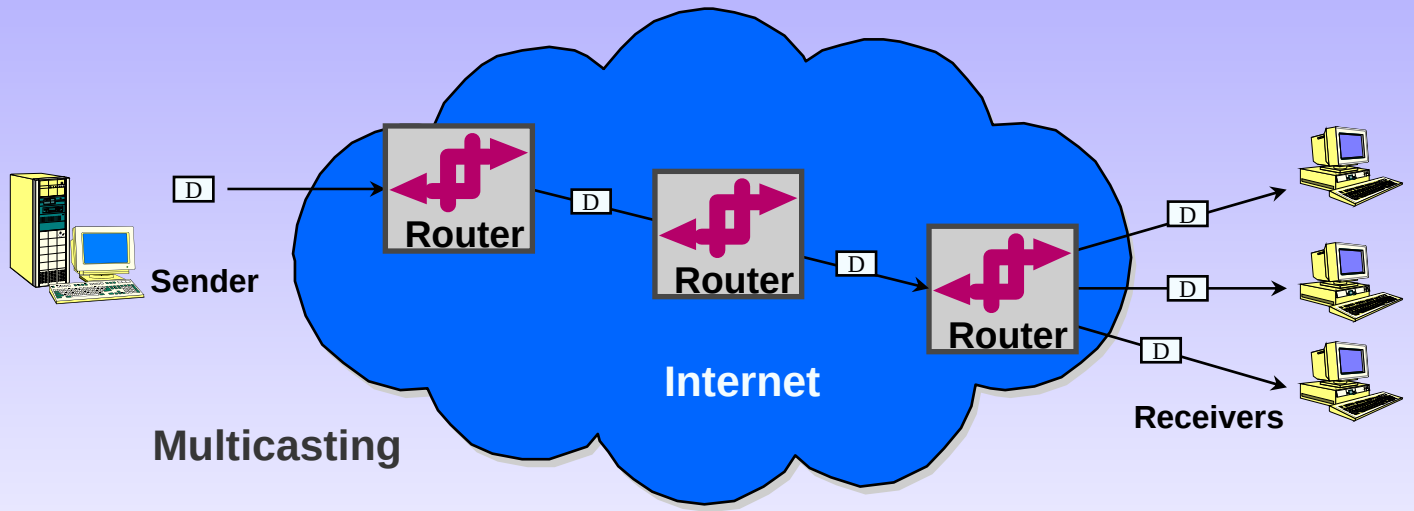
1.3.6 Multicasting-组播

- IP组播的基本定义：
  - 在LAN/WAN上能使IP数据报从一个源同时到多个目的传输过程
  - 接收组成员参加组播会议，应用只需发送一份拷贝到需要接收的组
  - 组播技术让包只寻址到组，而不是单个接收者
  - 开放组播的结点要运行一套能接收组播报文的TCP/IP协议
  - 由IETF推荐的1112RFC定义的对IP的扩展
- 相关缩略语
  - BSR:Boot Strap Router,自举路由器
  - IGMP:Internet Group Management Protocol,互联网组管协议
  - MBGP:Multi-protocol Border Gateway Protocol, 多协议边界网关协议
  - MSDP:Multicast Source Discovery Protocol,组播源发现协议
  - PIM-SM:Protocol Independent Multicast-Spars mode,协议独立组播－稀疏模式
  - PIM-DM:Protocol Independent Multicast-Dense mode,协议独立组播－密集模式
  - RP:Rendezvous Point,汇集点
  - RPT: Rendezvous Point Tree , PIM-SM协议共享树
  - SPT: Shortest Path Tree , 最短路径树
  - DR：Designated Router，指定路由器（运行IGMP协议）

IP组播路由的基本问题与方法

- 组播的基本矛盾？
  - 组播IP地址只是一个组成员集合的逻辑名字，不具有单播IP地址可直接全网寻址的功能
  - 组播地址用以区别网上播发的会话（session），而不是特别的物理目的地，源不需知道所有相关可直接寻址的地址
  - 组成员可能分散在Internet各个地方，不可能进行CIDR类似的聚类，而方便寻址
  - 基本问题：**知道目的组地址，但不知其包往何处发！**
- 组播的基本任务：
  - 建构以组播源为根的支撑树—找到接收对象
  - 在支撑树上传输IP组播流—发送组播流
- 组播的特点
  - 对不同的接收,许多会话数据流是相同的
  - 组成员可随时进出，导致组播树会变化
  - 组播支撑树的构造方法随组播协议的不同而变化
  - 单播转发路由关心包发往哪里，组播路由则只关心包来自哪里？故称—逆向路径？寻找源在何处？

先进的组播技术

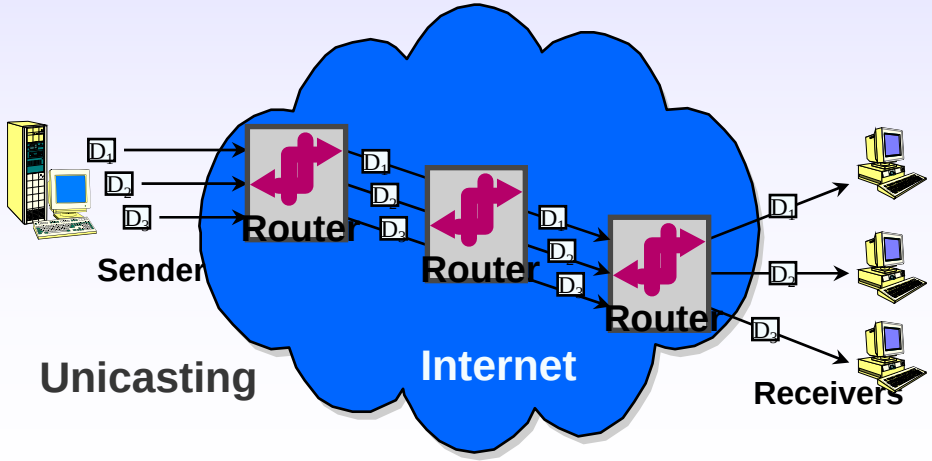


Multicast

-在 IP 网上一对多的传输

-支持视频会议,

-e-learning, 培训等

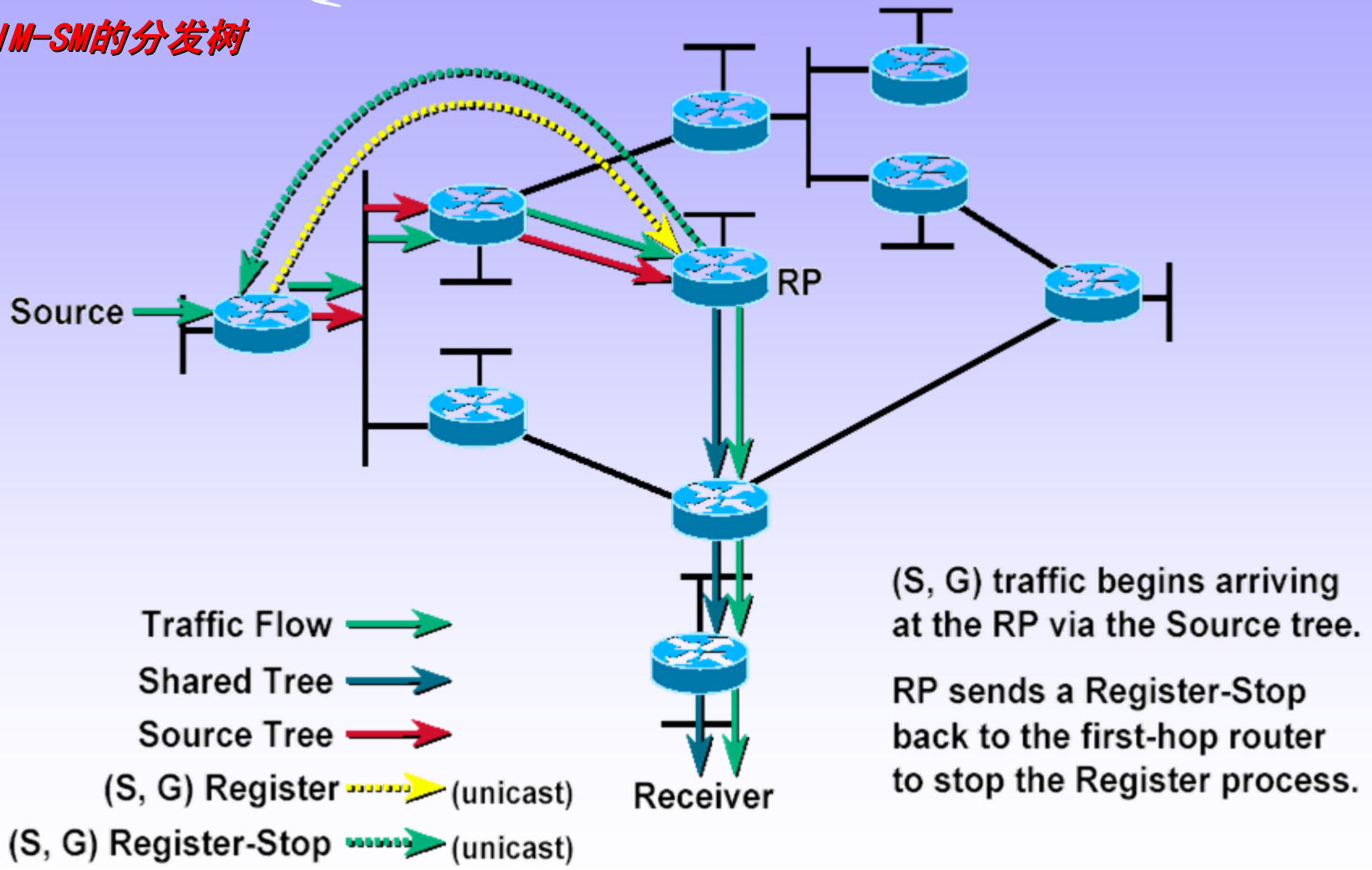


1）组播通信模型与协议

- 构成：1个核心 ＋ 3个发现
  - 分发树为核心
  - 源发现、接收者发现、拓扑分离
- 任务：组播路由协议根据组播**源**信息、**接收者**信息、网**络拓扑**（源和接收者间的连接关系）信息来构造组播**分发树**
- 协议：组播协议 ＝ 组播接收者发现协议 ＋ 组播路由协议 ＋ 组播源发现协议 ＋组播拓扑分离协议

一个核心（分发树）

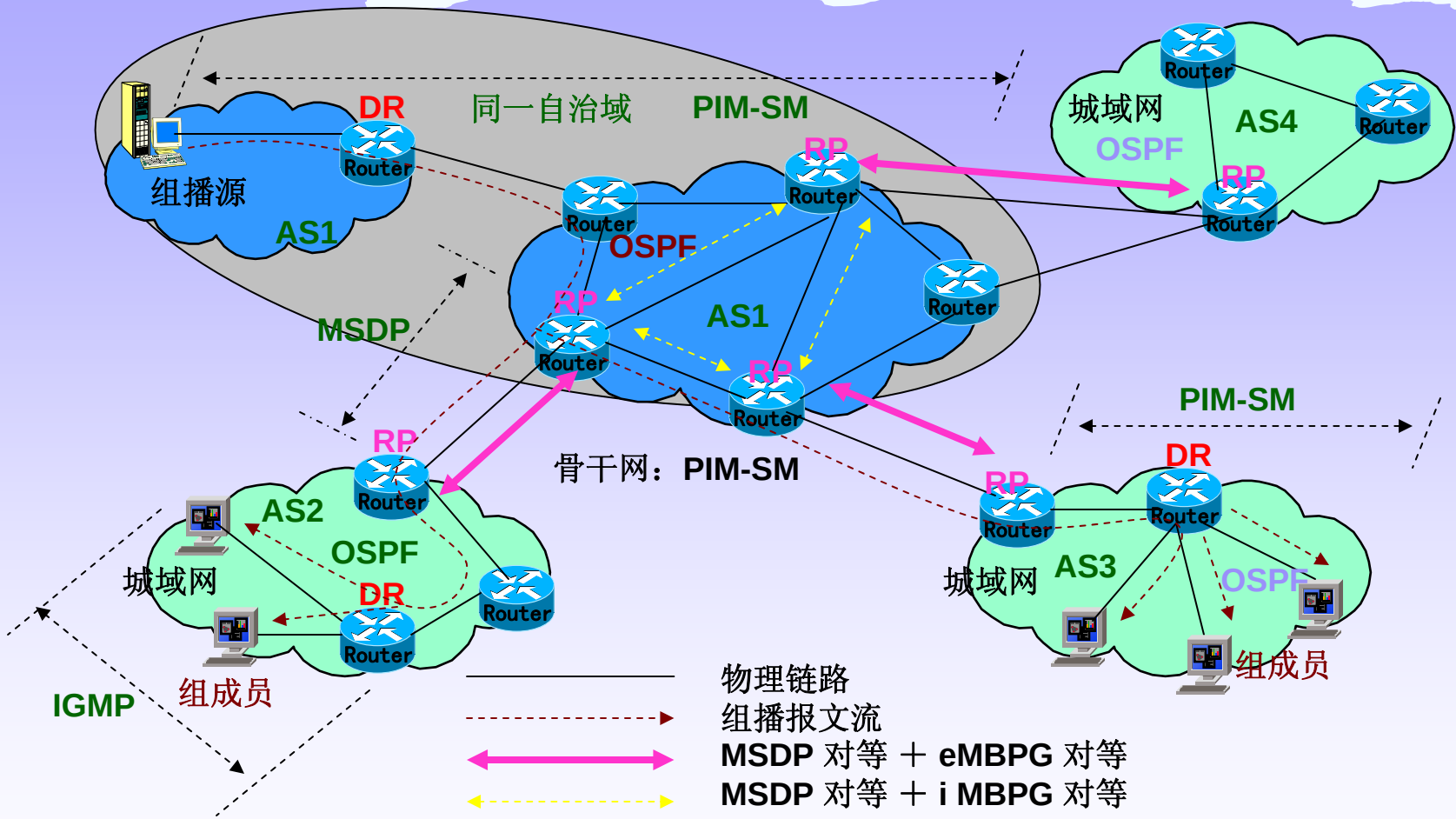
PIM-SM的分发树



三个发现

- 源发现协议
  - PIM-SM：完成AS域内的源发现
  - MSDP：完成AS域间的源信息发现和传播
- 接收者发现协议IGMP：
  - 位置：运行在主机和指定路由器DR之间
  - 任务：维护组播组是否有成员、获得该组成员信息；不关心到底是哪些成员，有多少成员
  - 优点：状态信息不因组播成员增加而增加；具有良好的可扩展性
- 拓扑分离协议
  - 目的：用于传播AS间的组播拓扑信息，指导跨AS组播分发树的建立
  - 原因：
    - PIM-SM在跨AS时，通过BGP协议掌握其它AS的网络拓扑信息
    - 但是有时需要单播、组播流量沿不同跨域路径转发，或应用不同路由策略
  - MBGP = eMBGP + iMBGP

跨AS的组播结构



IGMP：　完成路由器直连网段接收者发现　　OSPF：　完成AS内拓扑发现

MSDP：　交换AS之间的源信息　　MBGP：　交换AS之间的拓扑信息

PIM-SM：　完成AS内的源发现，根据上述信息构建组播分发树，达到域间组播转发

两大类组播路由协议

- 密集模式组播路由协议
  - 特点：
    - 假设组播组成员密集分布于网络上的许多子网，并至少包含一个组成员
    - 都需要大量的带宽
  - 典型路由协议
    - DVMRP
    - MOSPF
    - PIM-DM
  - 基本方法：
    - 依靠洪泛技术传播信息到所有组播路由器
- 稀疏模式组播路由协议
  - 特点：
    - 假设组播组成员稀疏分布于网络上，并至少包含一个组成员
    - 不需要大量的带宽，组成员可能由ISDN连通
  - 典型路由协议
    - CBT:Core-Based Tree
    - PIM-SM
  - 基本方法：
    - 不用洪泛方式，因组成员稀疏分布，否则会浪费带宽并引发严重性能下降

II）主要组播路由算法

- 组播路由协议中使用的主要算法有
  - Flooding
  - Spanning tree
  - Reverse-Path Broadcasting(RPB)
  - Truncated Reverse-Path Broadcasting(TRPB)
  - Reverse-Path Multicasting(MPB)
  - Core-based tree

PIM：组播路由协议

- 常用组播路由协议：PIM-SM/DM
- 任务:根据IGMP掌握的接收者信息，单播路由协议（如OSPF等）掌握的拓扑信息来
  - 完成源发现
  - 构建分发树
- PIM-DM：假设接收者在网上**密集**分布。首先将数据推到全网，后用协议信令剪枝不需要数据的网段，即为**扩散－剪枝**方式，其构建的分发树属于**源树**
- PIM-SM：假设接收者在网上**稀疏**分布。采用**按需发送**方式组播数据，即只向那些需要数据的网段转发。该方式首先构建**共享树**，当用户接收到组播数据后**切换到源树**
- RP:是PIM-SM的核心路由器， RP通常为一个或多个组播组服务。
  - 组播用户所直连的路由器采取“显式加入”机制主动加入以RP为根的共享树
  - 当用户接收到组播数据后还可切换到源树

III）组播的挑战、问题与发展

- 有些域间组播路由是基于UDP/IP的
- 尽力传送（Best effort）
  - 会产生丢包
  - 不可能有很可靠的数据传输，应有针对性设计，可靠组播有待进一步研究
- 不能避免拥塞
  - 缺乏TCP滑动窗口，且“慢启动”会导致拥塞，可尝试检测和避免机制
- 复制：某些协议会导致偶尔生成重复的包
- 无序发送：一些协议机制会导致无序发送

组播的挑战

- 没有更多的ISPs和OEM厂商开发有用的组播应用,缺乏组播工具和平台
- 与防火墙的交互作用:UDP能有效防止组成员ACKs的内爆,但Firewall对其失去控制作用.解决办--应用网关?
- QoS:探讨用ATM,RSVP等
- 基于Internet的组播实际上很少成功案例
- 应用层组播发展迅猛-P2P...IPTV...

组播的安全问题

- 组管理和访问控制
- 真实性（授权与认证）机密性和完整性
- 组密钥的分发
- 成员的加入
- 成员的脱离
- 组密钥的更新
- 允许外部审计

- 非法组播源侵入
- 非法组播接收者接收
- 组播核心路由器仿冒
  - BSR仿冒
  - RP仿冒
- 跨网络的非法组播加入
- 审计与计费

组播的发展

- 支持组播的主要高层应用协议:支持可靠数据传输
  - RTP:Real Time Transport Protocol
  - RTCP:Real Time Control Protocol
  - RTP:Real Time Streaming Protocol
  - RSVP:Resource Reservation Protocol
  - RMP: Reliable Multicast Protocol
  - RMF: Reliable Multicast Framework Protocol
  - RAMP: Reliable Adaptive Multicast Framework Protocol
  - Reliable Multicast Transport Protocol
    - Lucent 在其e-cast 用RMTP处理文件传输

IP组播应用软件/客户端

- EMULive Image Corp.’s Active Theater
- ICAST Corp.’s ICAST Viewer
- IVS(INRIA Videoconference System)
- Fantaswtic Corp.’s MeadiaSurfer
- Microsoft’s NetShow Theater Sever
- Live Networks Inc.’s Multikit
- Precept IP/TV
- Intel’s Proshare
- RealNetworks RealSyStems G2
- … … …

习题

- P156(英文书), P102（中文书）
  - 3；4；7；8；9；12；14；15；20；22；24；26；32；34；40；46；
- P235(英文书), P153（中文书）
  - 1；3；4；9；11
- P354(英文书), P233（中文书）
  - 14；16；38；46；47
- P433(英文书), P285（中文书）
  - 4；11；14；29；49

Questions