第六章 网络层进阶

崔勇

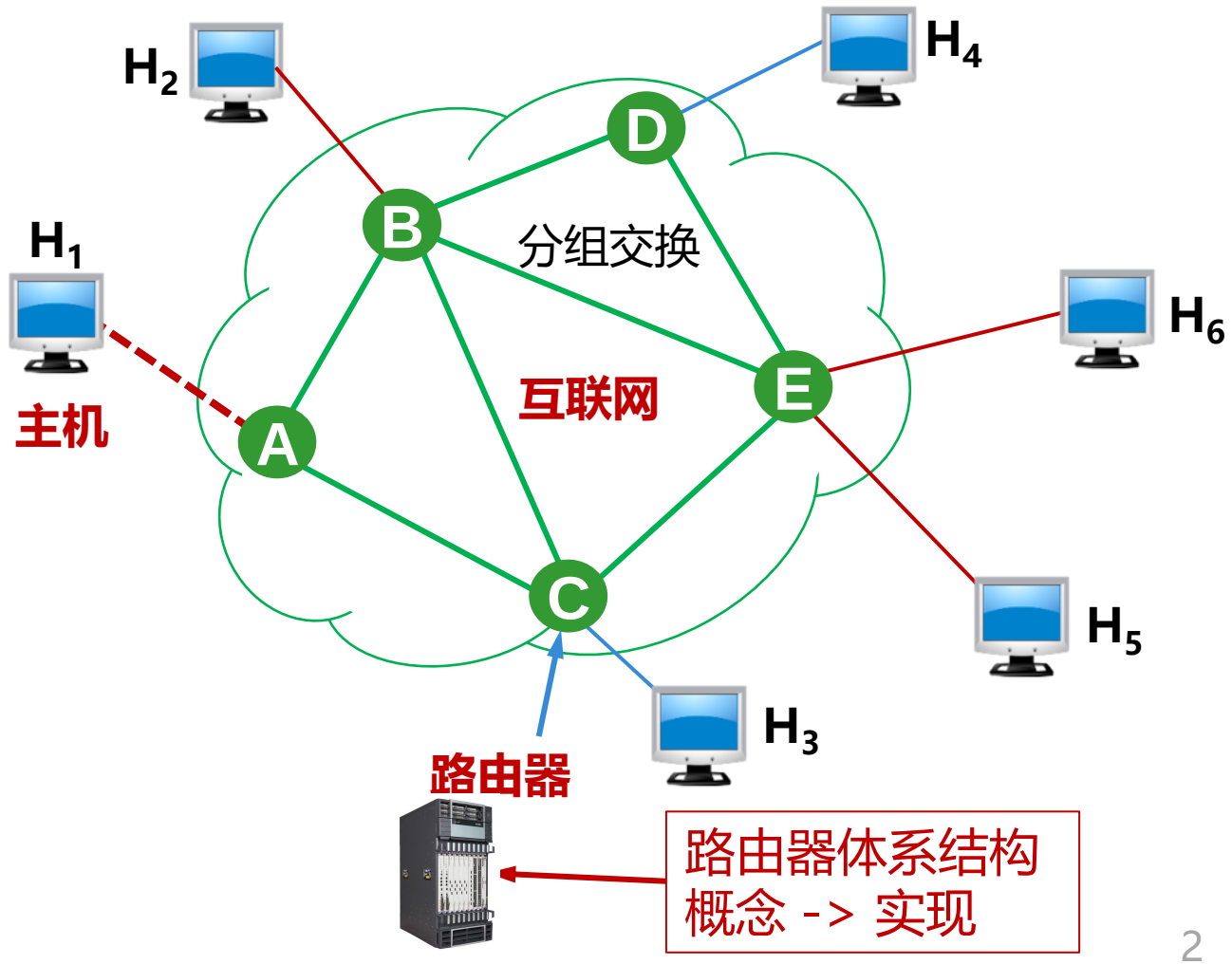
清华大学

计算机网络
教案社区

致谢社区成员	
首都师范大学 陈文龙	武汉大学 吴黎兵
西安邮电大学 谢晓燕	仲恺农业工程学院 邹莹
枣庄学院 李旭宏	辽宁大学 曲大鹏
西南民族大学 方诗虹	浙江理工大学 舒挺
内蒙古农业大学 白云莉	荆楚理工学院 余琨
华为技术有限公司 李振斌	

思考与展望

- 核心功能已完成
  - 两大功能：编址 & 寻址
  - 编址：IP协议族
  - 寻址：RIP、OSPF、BGP协议
- 更多问题……
  - 网络层概念->实现
    - 路由器的软硬件协同设计
  - IPv4地址耗尽问题：32位够吗?
    - IPv4优化还是更长地址
  - 更好的网络层服务：提升质量?
    - 什么是更好？从哪里着手?



本节目标

1. 了解标签交换概念，了解路由协议MPLS
2. 掌握路由器工作原理：控制层和数据层，报文转发
3. 掌握网络地址转换技术NAT
4. 掌握IPv6技术，了解IPv4/IPv6过渡机制
5. 了解服务质量控制技术和网络层拥塞控制技术
6. 了解软件定义网络SDN概念和应用

本节内容

6.6 标签交换和MPLS

6.7 路由器体系结构

6.8 NAT技术

6.9 IPv6技术

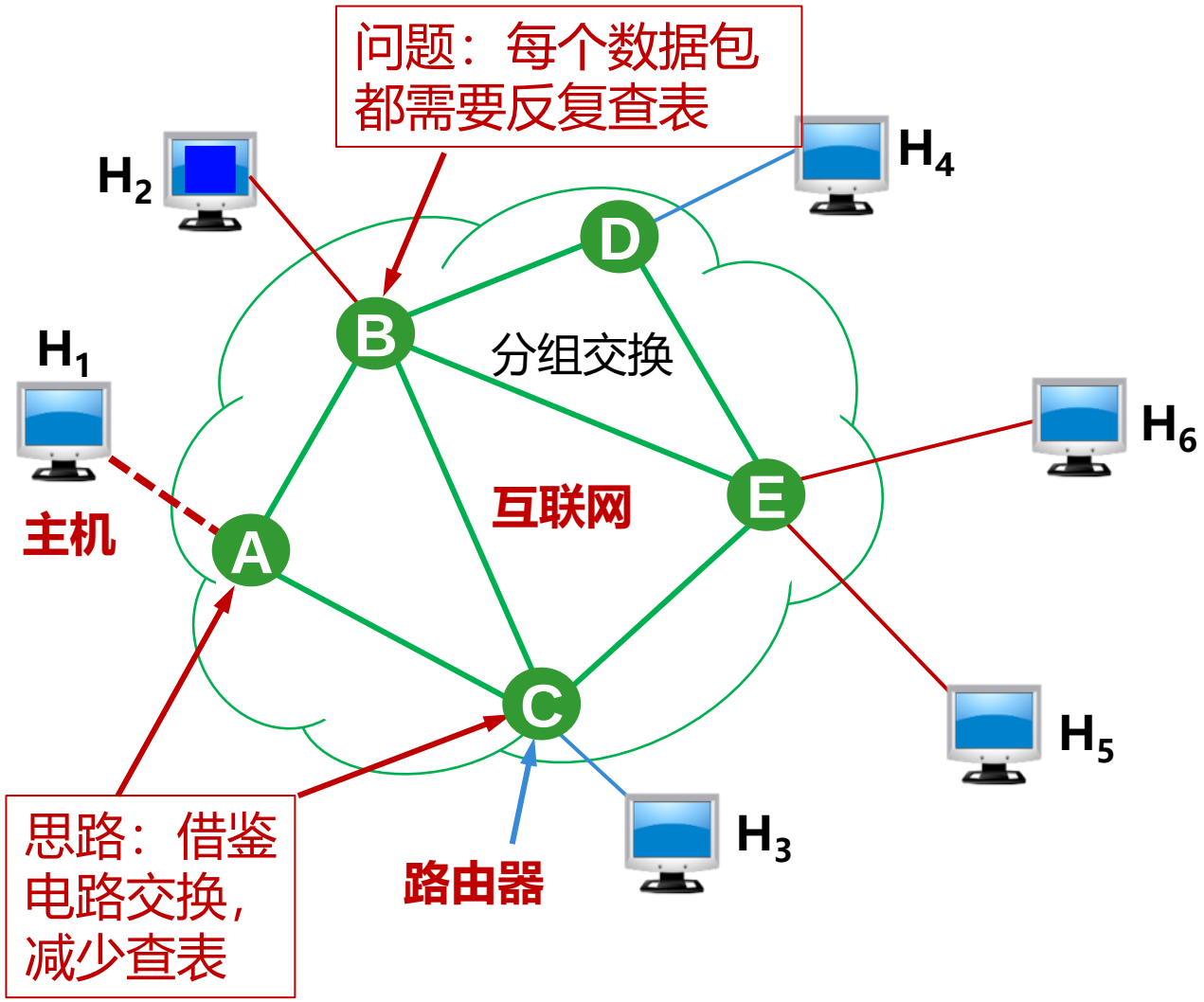
6.10 服务质量和拥塞控制算法

6.11 软件定义网络SDN

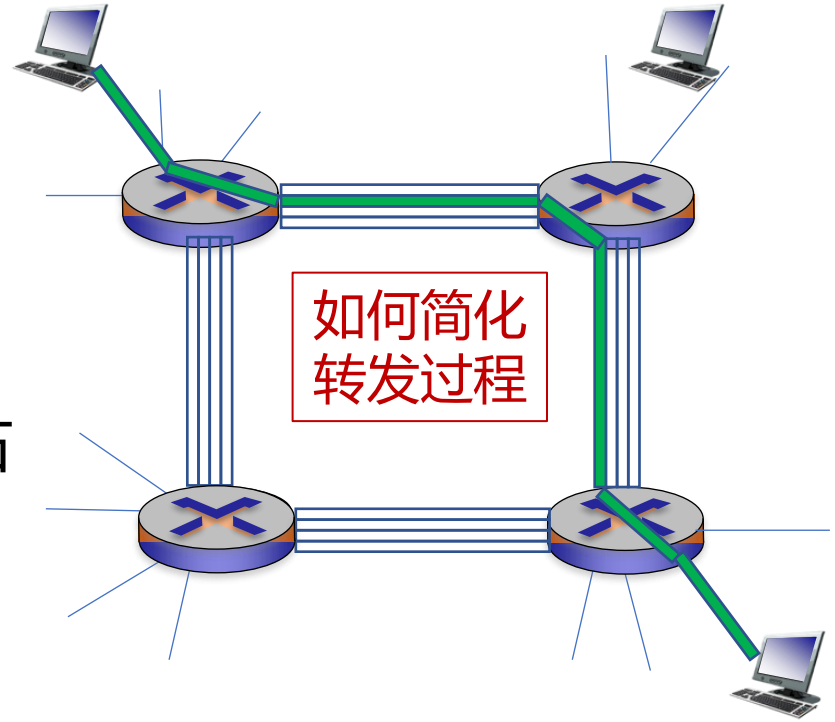
1. 虚电路交换方式
2. 标签交换和MPLS

思考与发明

- BGP结合RIP和OSPF优势
  - 支撑大规模网络，收敛快
  - BGP路由是否已经完美?
- 百万级路由，逐包查表?
  - 分组交换的必然结果
  - 电路交换不需要查表，借鉴?
- 思路：借鉴电路交换
  - 在分组交换网络上模拟电路交换
  - 如何维护“电路”状态?
  - 转发节点如何根据状态转发?



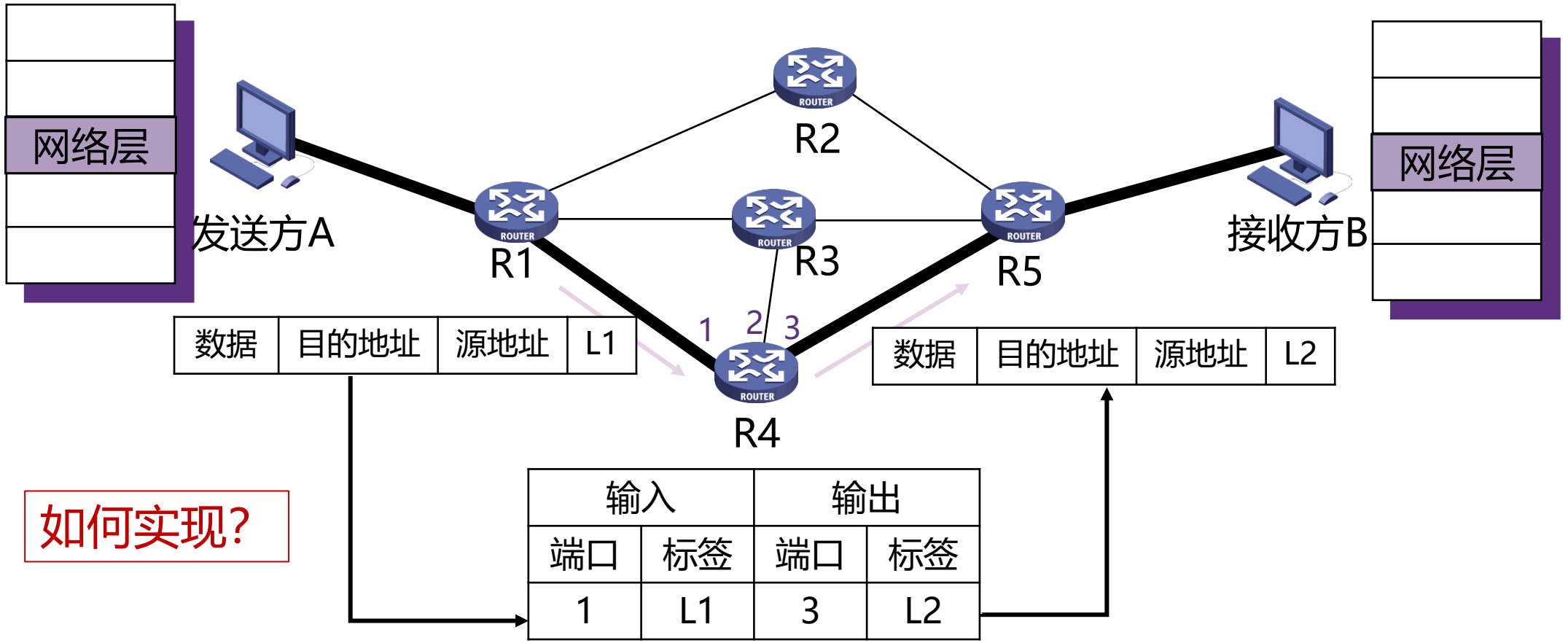
虚电路交换方式



- 回顾电路交换
  - 电路交换通常采用面向连接方式
  - 先呼叫建立连接，实现端到端的资源预留
  - 电路交换连接建立后，物理通路被通信双方独占
  - 中间设备按照物理通路状态直接转发
- 设计分组交换网络中的“电路交换”
  - 核心：相同路径的流量统一管理，聚合转发表（目的IP vs 聚合流）
  - 基于标签的虚电路网络：
    - 边缘路由器负责连接建立和拆除，维护“电路”状态（标签）
    - 内部路由器据“电路”状态转发，避免高频率查表带来的开销

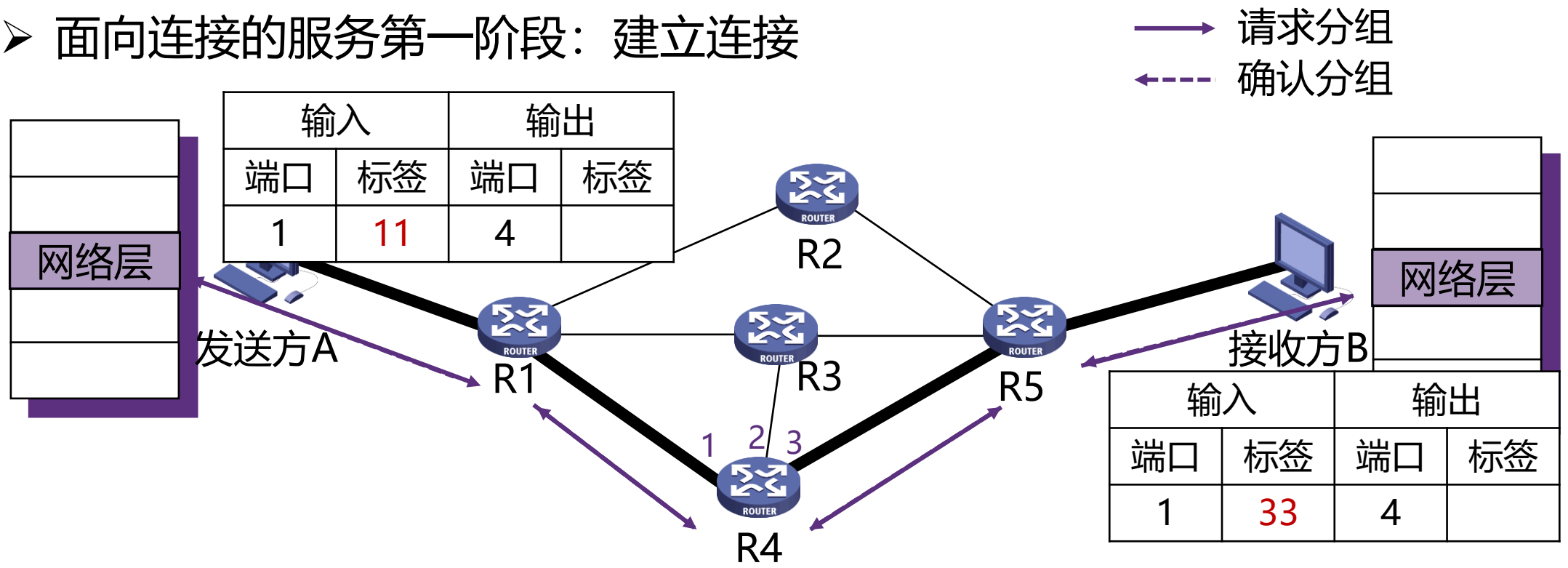
虚电路转发策略

- 虚电路的转发策略：虚电路转发决策基于分组标签，即虚电路号



面向连接的虚电路

- 面向连接的服务第一阶段：建立连接



输入		输出	
端口	标签	端口	标签
1	11	4	

输入		输出	
端口	标签	端口	标签
1	33	4	

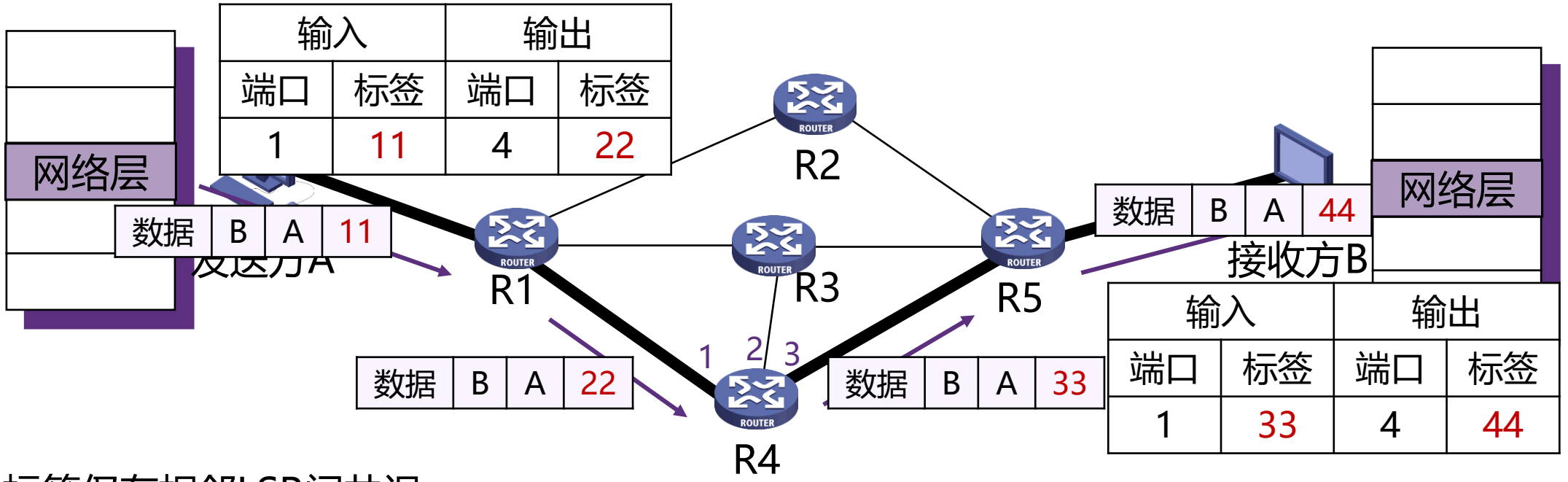
输入		输出	
端口	标签	端口	标签
1	22	3	

面向连接的虚电路

- 面向连接的服务第二阶段：发送数据

数据	目的地址	源地址	标签

分组



输入		输出	
端口	标签	端口	标签
1	11	4	22

输入		输出	
端口	标签	端口	标签
1	22	3	33

输入		输出	
端口	标签	端口	标签
1	33	4	44

标签仅在相邻LSR间共识

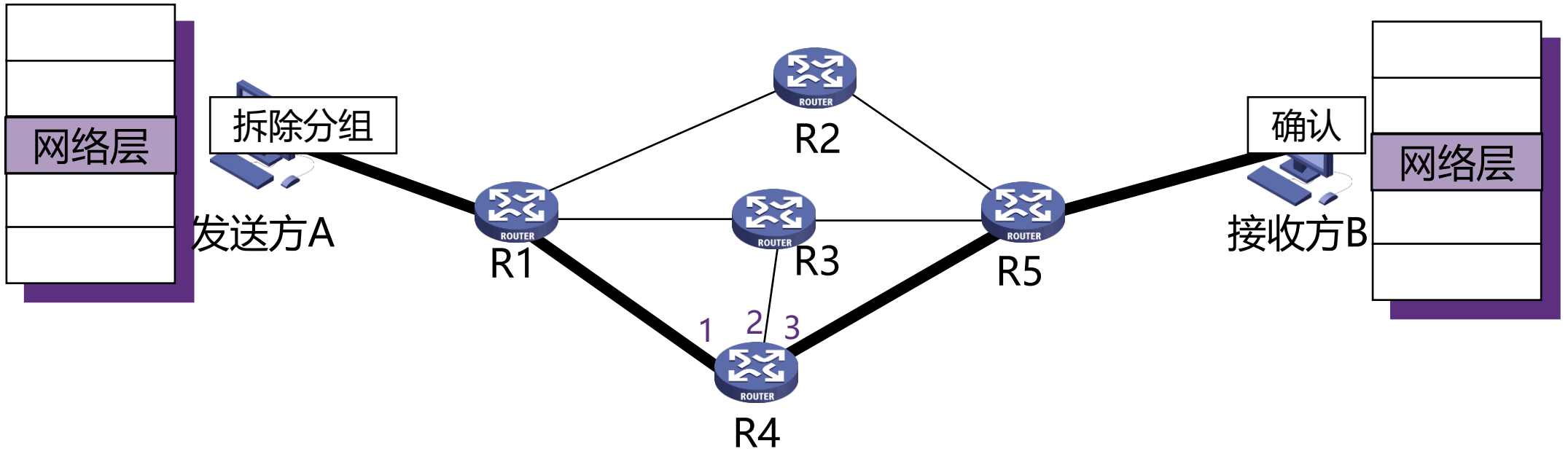
LSR会维护一张转发表

建立连接时进行资源预留

节点维护每流状态！？
清华->腾讯会议/云

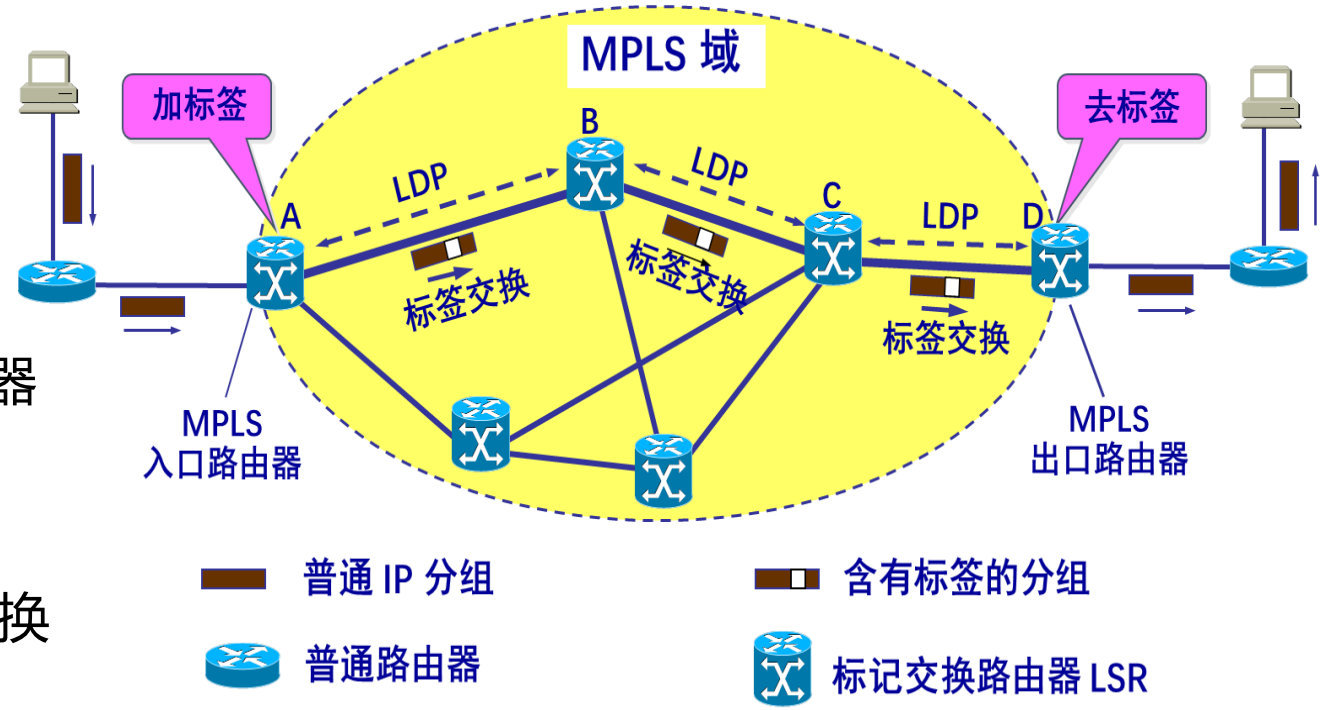
面向连接的虚电路

- 面向连接的服务第三阶段：释放连接



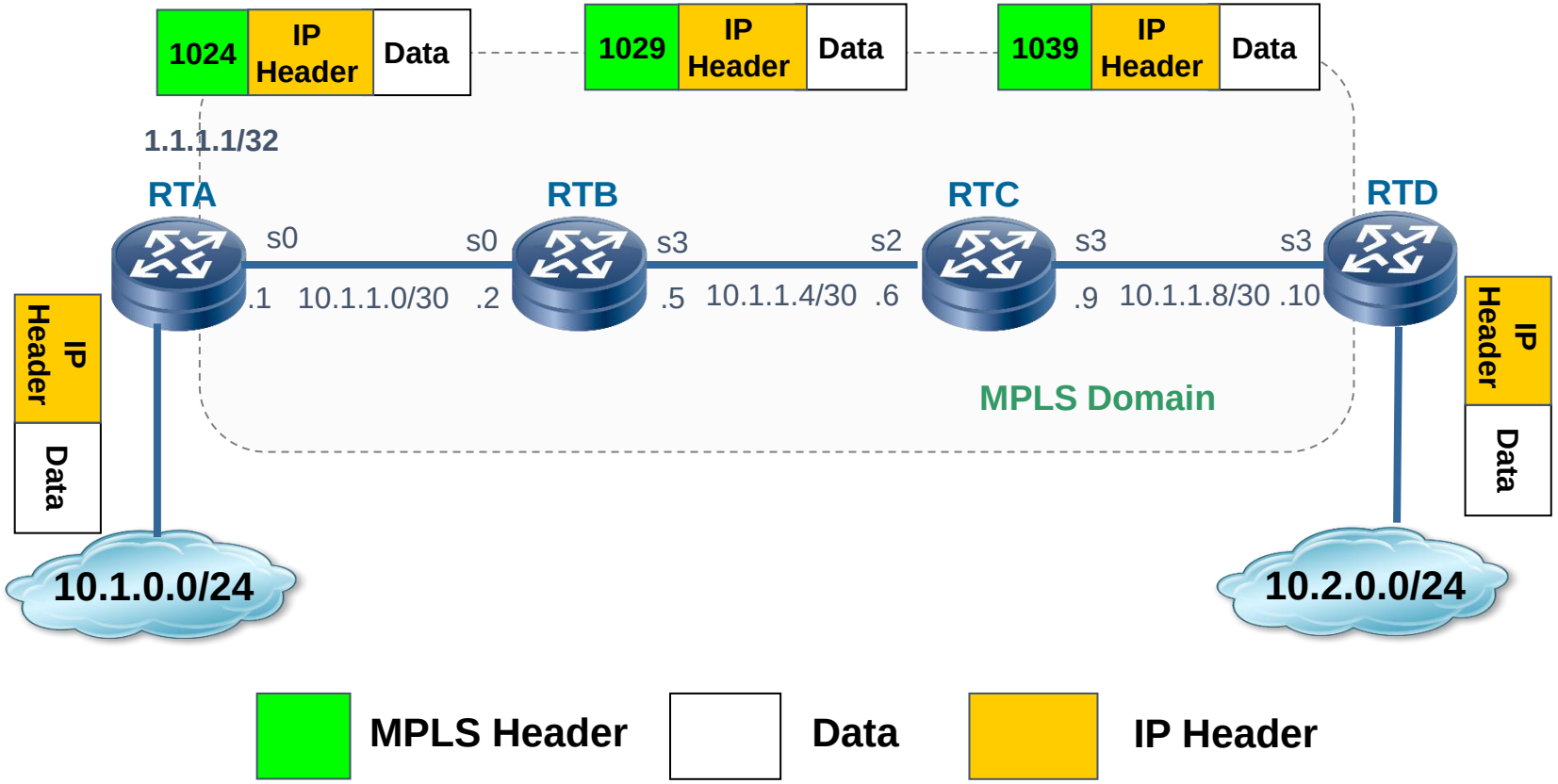
多协议标签交换MPLS

- 多协议标签交换MPLS (MultiProtocol Label Switching)
  - 多协议表示在 MPLS 的上层可以采用多种协议，例如：IP，IPv6、IPX
  - 标签是指每个分组被分配一个标签，路由器根据该标签对分组进行转发
  - 交换是指标签的交换，MPLS 报文交换和转发是基于标签的
- 标签交换路由器LSR
  - 支持MPLS的路由器
  - 具备标签交换、路由选择两种功能
- MPLS域
  - 所有相邻的支持MPLS技术的路由器构成的区域
- 标签分配协议LDP
  - 用来在LSR之间建立LDP 会话并交换Label/FEC映射信息



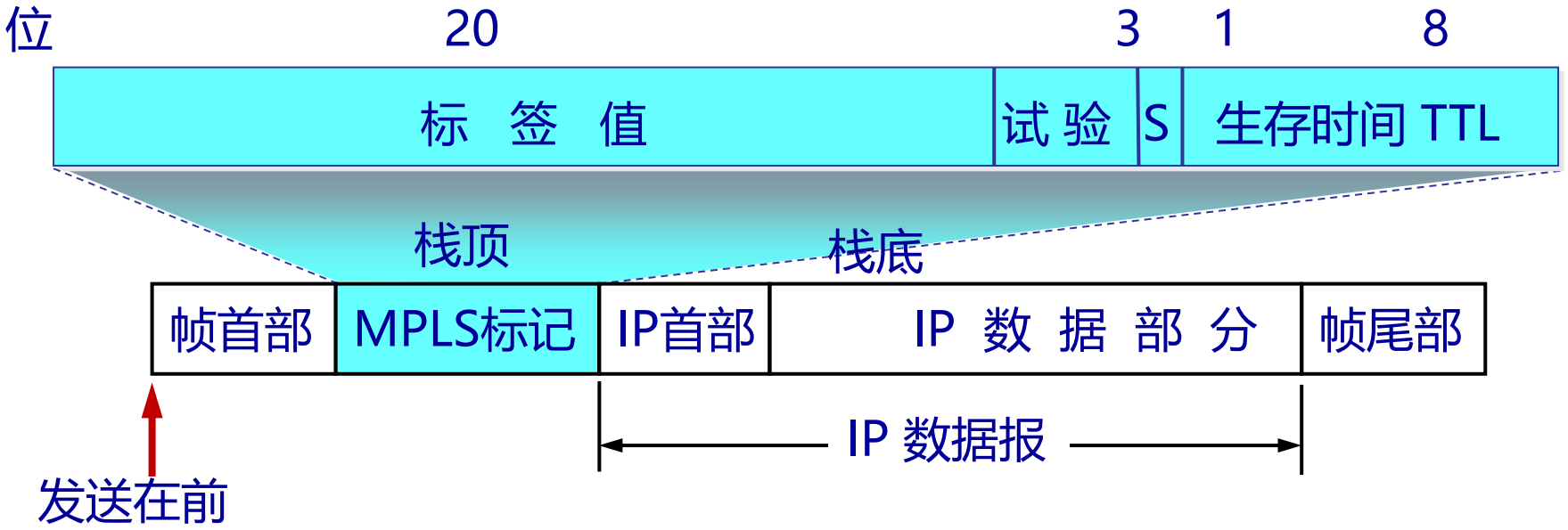
MPLS转发过程

- 加标签
  - 在 MPLS 域的入口处，给每一个 IP 数据报加上 标签，然后对加上标记的 IP 数据报用硬件进行转发
- 标签交换
  - 采用硬件技术对加上标记的 IP 数据报进行转发称为标签交换
- 去标签
  - 当分组离开 MPLS 域时，MPLS 出口路由器把分组的标签去除。后续按照一般IP分组的转发方法进行转发



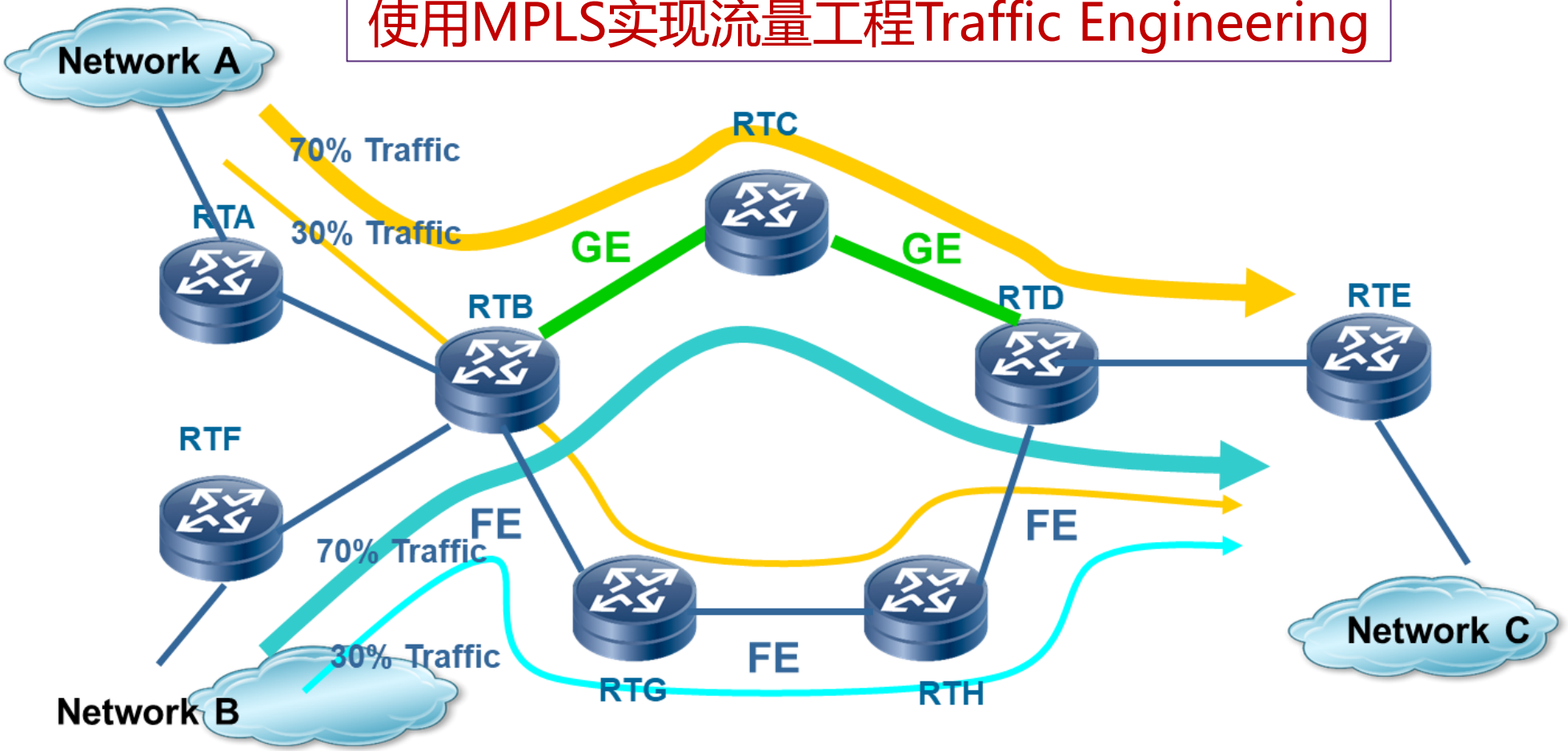
MPLS报文结构

- MPLS报文结构
  - “给 IP 数据报加标签”其实就是在以太网的帧首部和IP数据报的首部之间插入一个 4 字节的 MPLS 首部
  - MPLS又称为2.5层协议



MPLS典型应用

使用MPLS实现流量工程Traffic Engineering



路由协议小结（6.3-6.6）

- 思考与发明：RIP->OSPF->BGP
  - RIP：距离向量，邻居交互最优下一跳，分布式计算；收敛速度较慢
  - OSPF：链路状态，洪泛原始信息，集中计算最短路；带宽&计算开销大
  - BGP：融合RIP和OSPF设计经验，邻居交互路径向量；支撑大网路由
  - MPLS：借鉴电路交换，面向连接；更强控制，实现VPN和TE

多种路由协议生成的路由表
如何指导实际的转发？

本节内容

6.6 标签交换和MPLS

6.7 路由器体系结构

1. 路由器概述
2. 路由器控制平面
3. 路由器数据平面
4. 交换结构
5. 路由器拓展知识

6.8 NAT技术

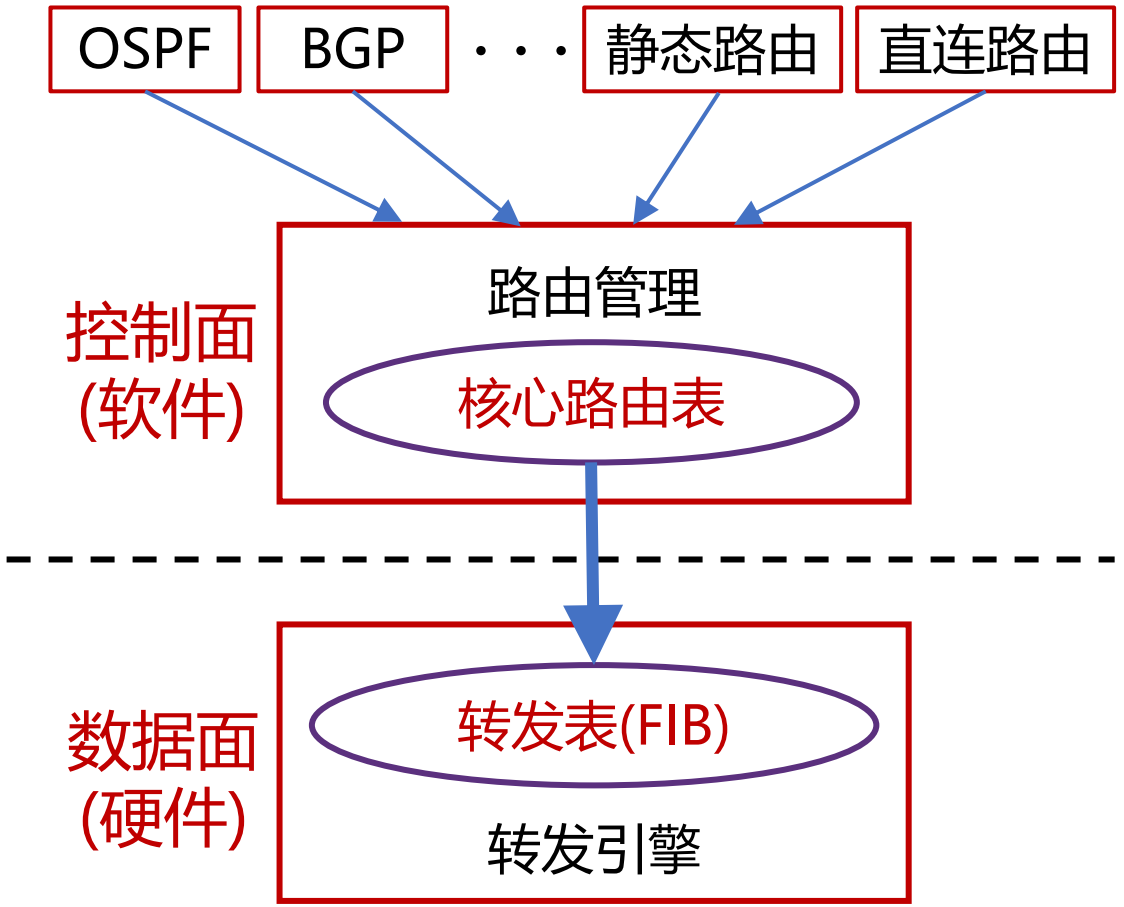
6.9 IPv6技术

6.10 服务质量和拥塞控制算法

6.11 软件定义网络SDN

路由器控制平面

- 路由器可同时运行多个路由协议
- 路由器也可不运行任何路由协议，只使用静态路由和直连路由
- 路由管理根据路由优先级，选择最佳路由，形成核心路由表
- 路由管理将核心路由表的信息，再提供给各个路由协议，实现控制面闭环
- 控制面将核心路由表下发到数据面，形成转发表（FIB）



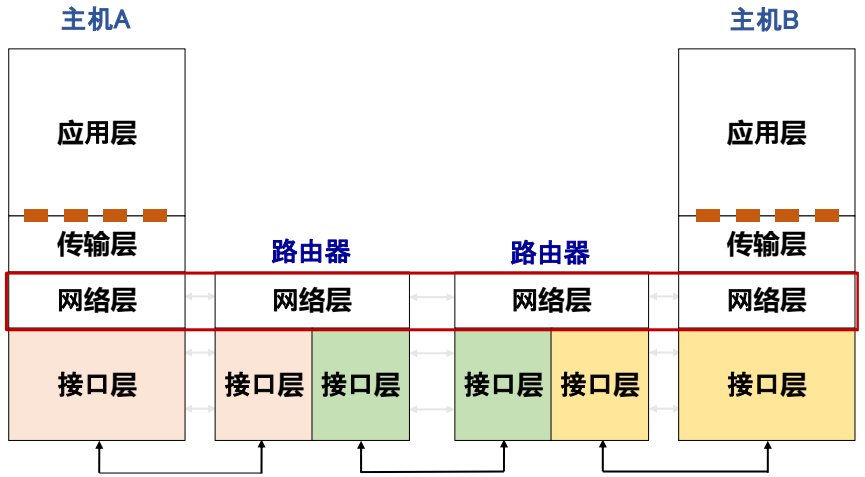
路由器控制平面

- 若存在多个“去往同一目的IP前缀”的不同类型路由，路由器根据优先级选择最佳路由（形成转发表）
- 优先级数值越小，优先级越高

路由种类	路由优先级
直连路由	0
静态路由	1
eBGP路由	20
OSPF路由	110
RIP路由	120
iBGP路由	200

路由器数据平面

- 路由器中IP报文转发过程
  - 链路层解封装（核对MAC地址、MAC校验）
  - 若上层为IP，进行IP头部校验，获取目的IP地址
  - 用目的IP地址和最长前缀匹配规则，查询转发表
  - 路由匹配查询失败，丢弃报文
  - 查询成功
    - 获取转发出接口和下一跳IP地址
    - IP头部“TTL”字段值减1，重新计算IP头部“校验和”
    - 重新进行链路层封装，发送报文

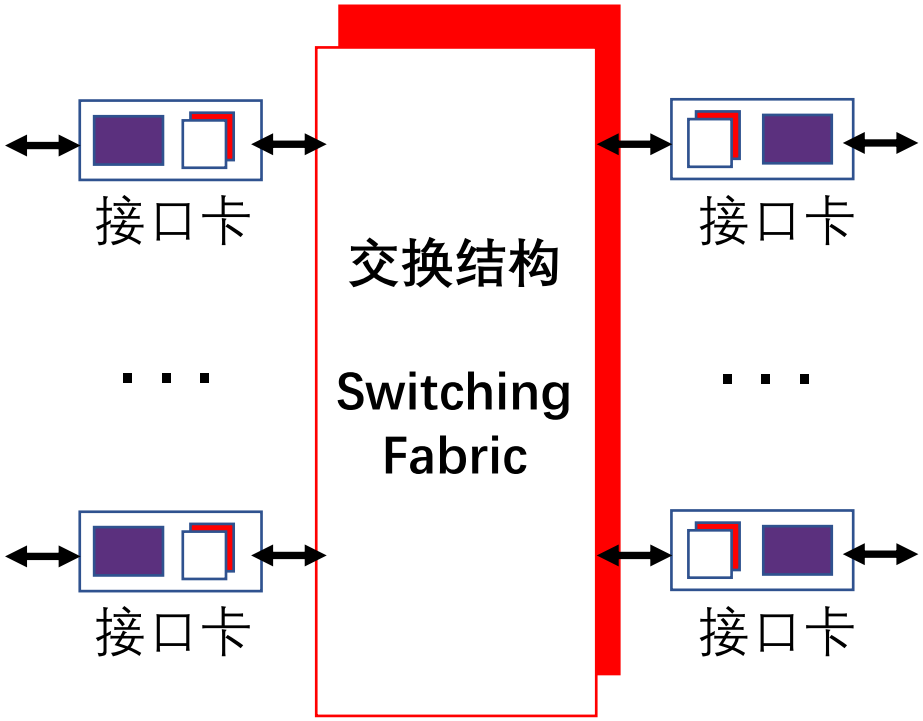


普通IP报文转发过程中，路由器不查看传输层及以上层协议的内容

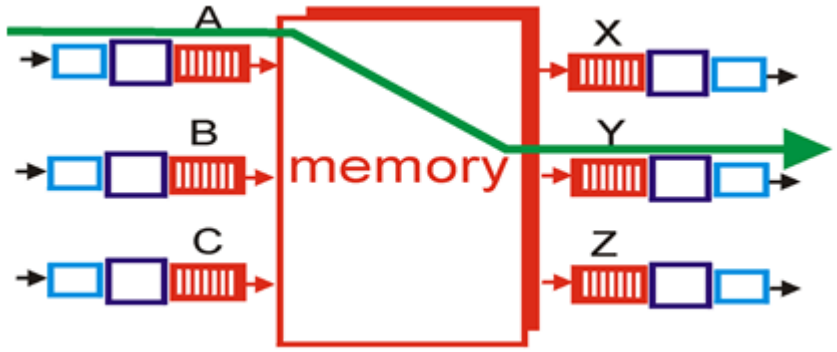
- IP报文在路由器转发前后的变化
  - 链路层封装更新（新头）；IP头部“TTL”减1，IP头部“校验和”更新

路由器数据层

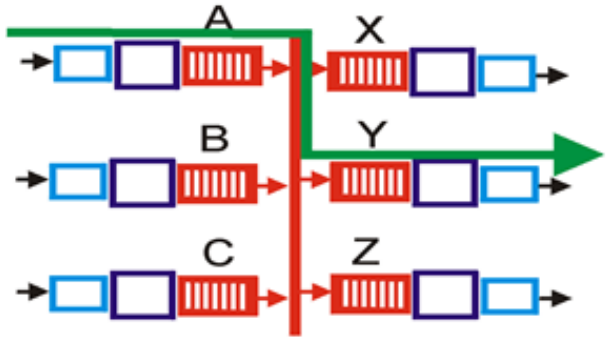
- 数据报在不同硬件单元的处理
  - 报文输入的接口卡
    - 链路层解封装
    - 转发表查询
    - 通过交换结构将报文排队并发往目的接口卡
  - 不同类型的交换结构
    - 从输入接口卡发往输出接口卡
  - 报文输出的接口卡
    - 从交换结构接收报文，排队进行后续处理
    - 链路层封装
    - 从输出接口发送报文



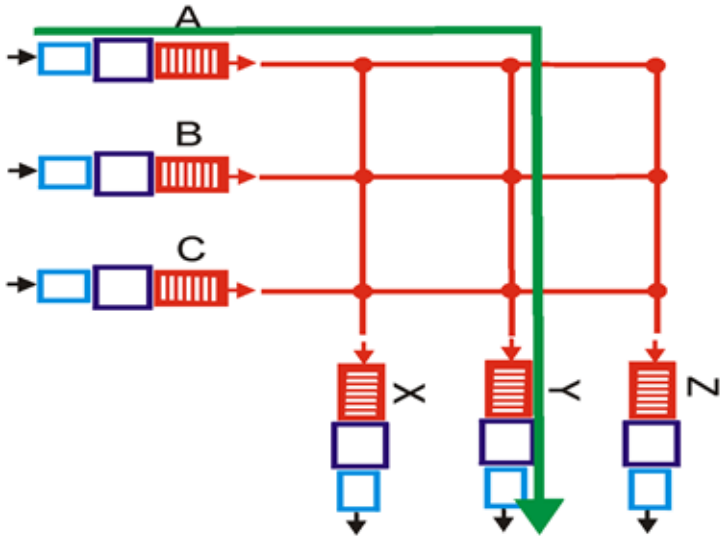
3种典型的交换结构



共享内存



共享总线



纵横式 Crossbar

应用于高性能分布式路由器

输入排队 v.s. 输出排队

队头丢弃 v.s. 队尾丢弃

路由器扩展知识

- 路由器的端系统角色
  - 也作为网络端系统进行协议交互
    - 远程网络管理，SNMP
    - 远程网络配置，SSH
    - 文件传输，FTP，TFTP
    - 各种路由协议交互
    - ...

  - 路由器系统包含完整TCP/IP协议栈
    - 传输层协议
    - 应用层协议

路由器有4/5两层协议吗?

- 家用路由器
  - 不运行动态路由协议（出口唯一）
  - 运行DHCP协议，分配私有IP
  - NAT地址转换
  - 本地DNS服务
  - 用户管理及认证
  - 防火墙功能
  - 无线AP
  - ...

与典型网络路由器差异较大

本节内容

6.6 标签交换和MPLS

6.7 路由器体系结构

6.8 NAT技术

网络地址转换NAT

6.9 IPv6技术

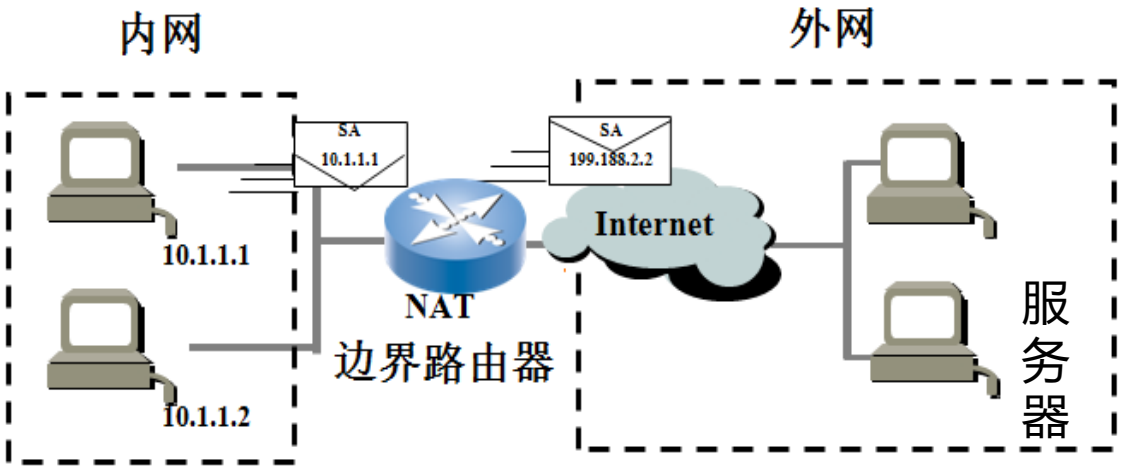
6.10 服务质量和拥塞控制算法

6.11 软件定义网络SDN

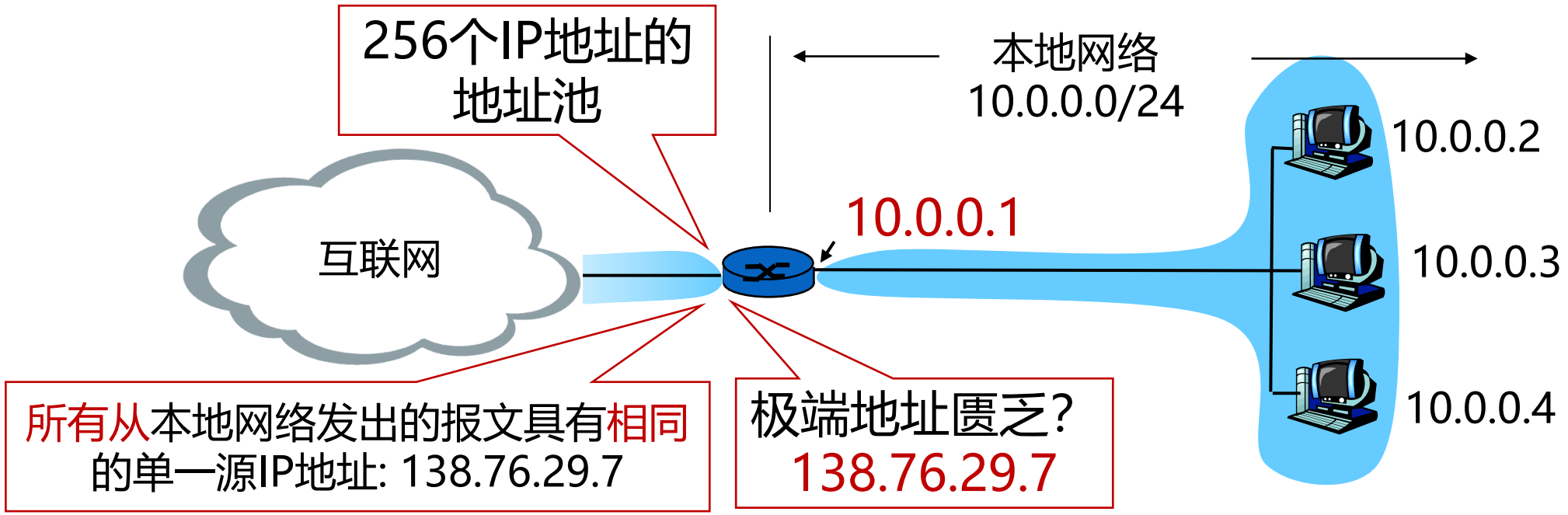
网络地址转换NAT

- 自己建网络自己用：私有IP地址
  - A类地址：10.0.0.0--10.255.255.255
  - B类地址：172.16.0.0--172.31.255.555
  - C类地址：192.168.0.0--192.168.255.255
- IP地址不够用怎么办？
  - 能否设计地址共享/复用机制？
- 网络地址转换(NAT)
  - 用于解决IPv4地址不足的问题，是一种将私有（保留）地址转化为公有IP地址的转换技术
  - 能否复用IP地址？
  - 如何映射：谁出去，回来给谁？

如何上互联网？



NAT工作机制

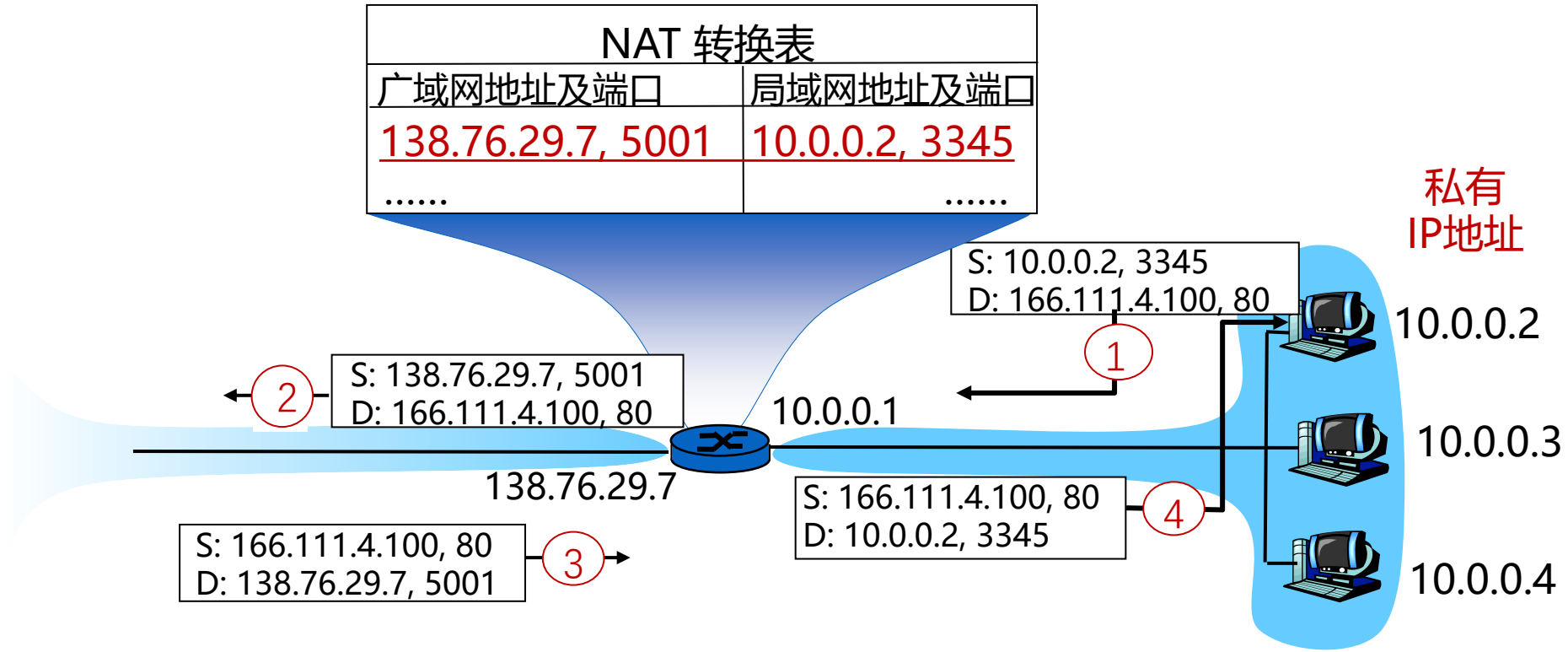


不同的源端口号

传输层TCP/UDP拥有16-bit 端口号字段
所以WAN侧一个地址可支持>60,000个并发连接

思考：同一主机不同应用，或者不同主机的同一端口，NAT转换如何处理?

NAT工作机制



什么时候建立NAT表?
自动建立还是手动建立?

网络地址转换

- 有效提升地址复用率，解决IPv4地址不足的问题
  - 一个WAN侧地址可支持>60,000个并行连接
  - NAT根据不同的IP上层协议进行NAT表项管理
  - TCP，UDP等
- NAT的优势
  - 节省合法地址，减少地址冲突
  - 灵活连接Internet
  - 保护局域网的私密性

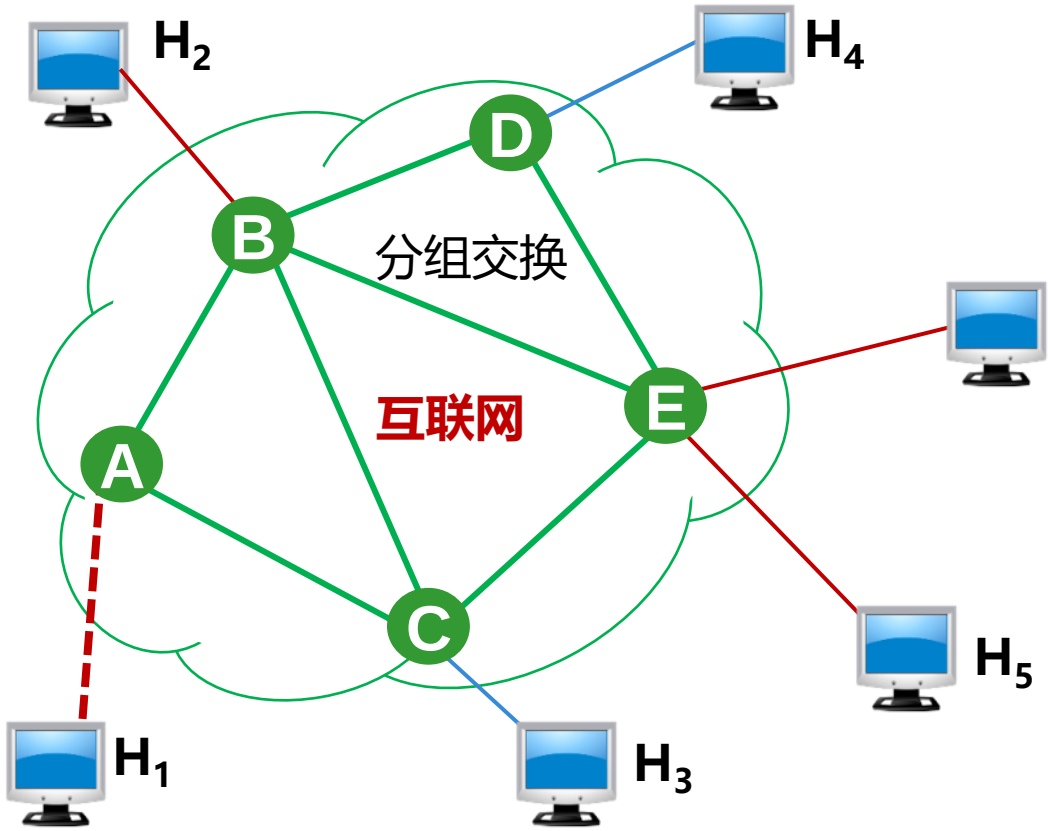
NAT是协议吗?
大家喜欢吗?

更根本的解决方案?
IPv6协议

- 问题或缺点
  - 违反了IP的结构模型，路由器处理传输层协议
  - 违反了端到端的原则
  - 违反了最基本的协议分层规则
  - 不能处理IP报头加密
  - 新型网络应用的设计者必须要考虑 NAT场景，如 P2P应用程序

思考与发明

- IP地址不够用怎么办?
  - 地址买卖?
  - NAT可以缓解地址不足
  - 根本方案：扩展地址空间
- IPv6协议设计
  - 更长的地址位数：128位
  - 还有什么需要改的?
    - 辅助协议：ARP，DHCP，ICMP
    - 路由协议：RIP，OSPF，BGP
  - 新旧互通?



IPv6地址：2402:f000:3::7d87:caa6:46de

本节内容

6.6 标签交换和MPLS

6.7 路由器体系结构

6.8 NAT技术

6.9 IPv6技术

6.10 服务质量和拥塞控制算法

6.11 软件定义网络SDN

1. IPv6简介
2. IPv6头部结构
3. IPv6扩展头
4. IPv6地址配置
5. 邻居发现
6. IPv6路由协议
7. IPv4/IPv6过渡技术

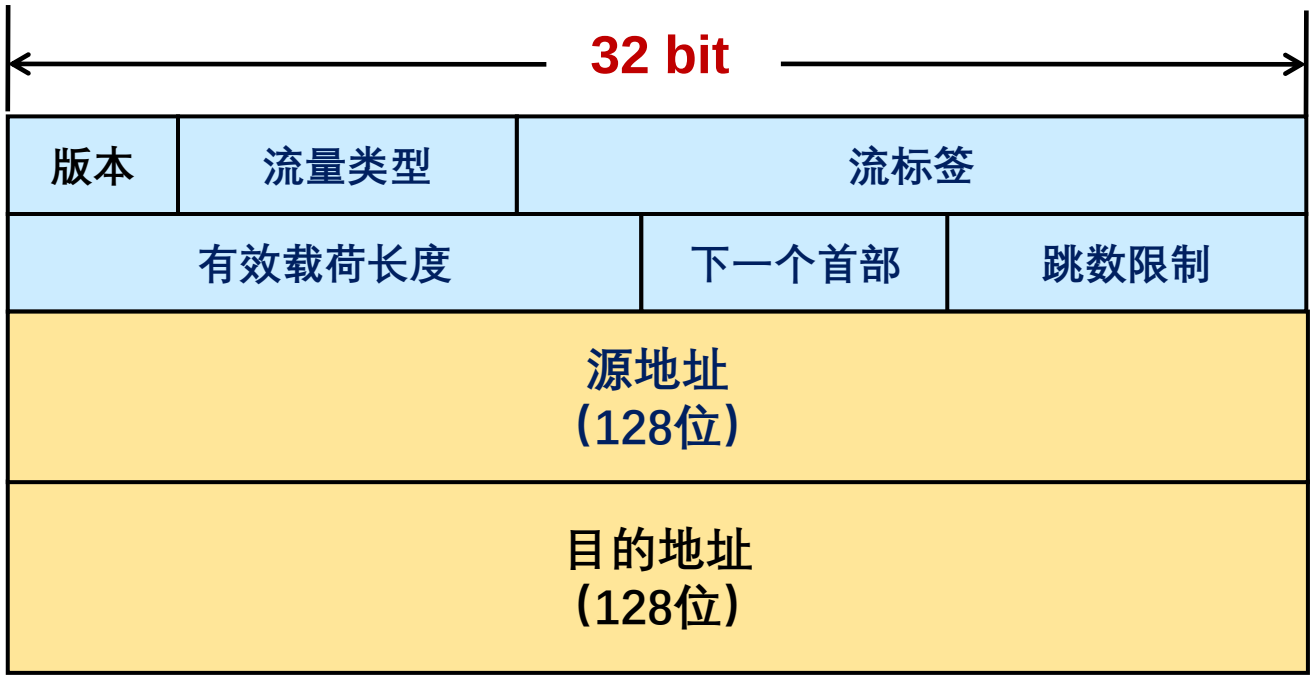
IPv6协议

- IPv6（Internet Protocol version 6）
  - 千年大计：32-bit地址空间耗尽
  - CIDR和NAT都无法从根本上解决地址短缺问题
  - 互联网工程任务组（IETF）设计的用于替代IPv4的下一代协议
- IPv6 地址
  - 地址长度为128bit，是IPv4地址长度的4倍
  - IPv6地址空间数量约为$3*10^{38}$
  - IPv6地址表示法，冒分十六进制，x:x:x:x:x:x:x:x
    - 简化方法：每个x前面的0可省略，可把连续的值为0的x表示为“::”，且“::”只能出现1次
    - 简化前地址，2001:0DA8:0000:0000:200C:0000:0000:00A5
    - 简化后地址，2001:DA8:0000:0000:200C::A5

IPv6头部

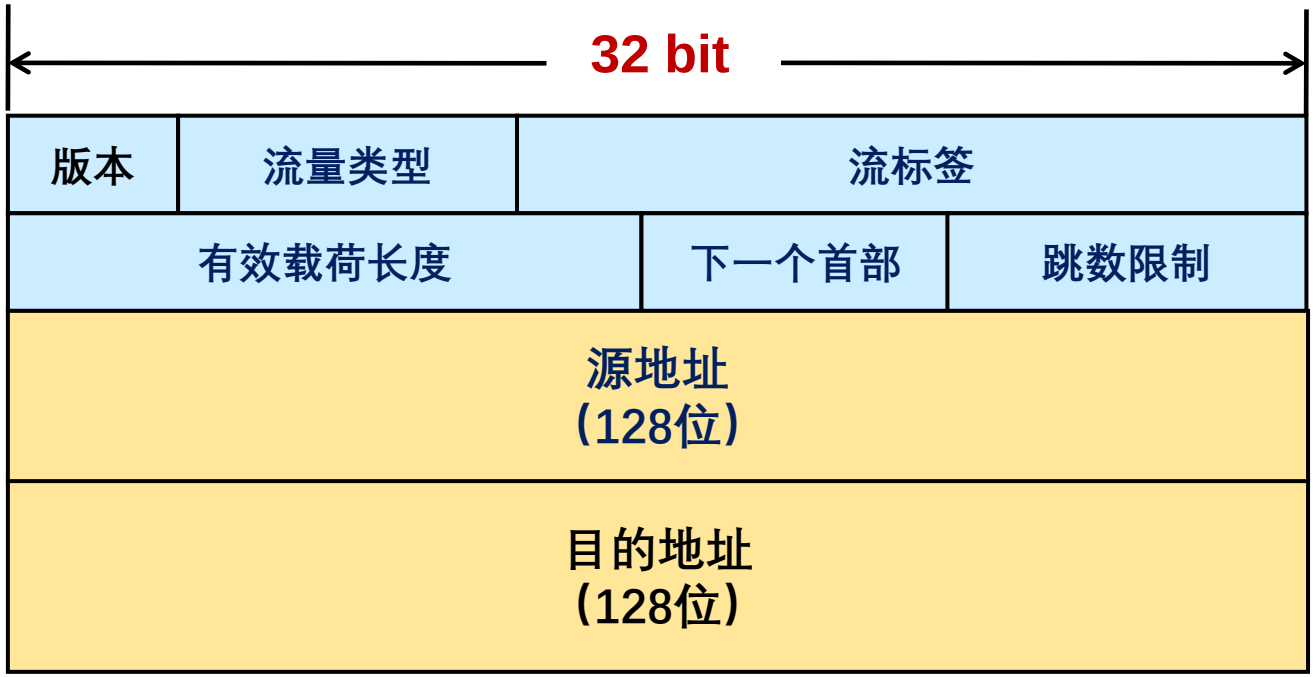
- 版本：4bit，协议版本号，值为6
- 流量类型：8bit，区分数据包的服务类别或优先级
- 流标签：20bit，标识同一个数据流
- 有效载荷长度：16bit，IPv6报头之后载荷的字节数（含扩展头），最大值64K

IPv4分组头



IPv6头部

- 下一个首部：8bit，IPv6报头后的协议类型，可能是TCP/UDP/ICMP等，也可能是扩展头
- 跳数限制：8bit，类似IPv4的TTL，每次转发跳数减1，值为0时包将会被丢弃
- 源地址： 128bit，标识该报文的源地址
- 目的地址： 128bit，标识该报文的目的地址



IPv6头部字段分析

如何在IPv4基础上提升分组转发处理速度?

- IPv6头部长度固定40字节，所有“选项”字段都在IPv6扩展头部分
- 与IPv4头部的比较
  - 去除“首部长度”（首部长度固定为40字节）
  - 去除“首部校验和”，提升转发速度
- IPv6分片机制
  - IPv6分组不能在传输途中分片，只在源端进行分片
  - IPv6支持Path MTU发现机制
  - 去除分片字段：“标识”“标志”“片偏移”，移至扩展头（分段头）

IPv6扩展头

- IPv6报文可承载多个扩展头
- 每个扩展头都包含“下一个首部”字段（IPv6首部固定字段也有）
  - 可指向下一个扩展头类型
  - 或指明传统上层协议类型（最后一个扩展头）： TCP/UDP/ICMP . . .
- 如有多个扩展头，需按规定顺序出现
  - 逐跳选项头，转发路径上每个节点都需检查该扩展头的信息
  - 路由头，指明转发途中需经过哪些节点，类似于IPv4的源路由机制
  - 分段头，包含类似IPv4分片处理信息：片偏移、“更多段”标志、标识符
  - 目的地选项头，目的端系统需要确认的信息
  - . . .

IPv6地址及配置

- IPv6地址分类
  - 未指定地址（::/128），不能分配给任何节点
  - 回环地址（::1/128），表示节点自己，不分配，类似IPv4中的127.0.0.1
  - 组播地址（FF00::/8）
  - 链路本地地址（FE80::/10），也称为Link-local地址，仅在本地链路上使用，网络设备根据接口MAC地址自动生成
  - 全局单播地址，其它地址
- IPv6地址配置方式
  - 手动配置
  - DHCPv6（IPv6动态主机配置协议）
  - 无状态地址自动配置，基于ND协议的RS报文的IPv6前缀信息，结合自己的链路层地址生成IPv6地址

邻居发现ND协议

- 邻居发现协议（Neighbor Discovery Protocol，ND）
  - 邻居发现基于ICMPv6实现，不同的Type值和Code值表示不同的ND消息
- 消息类型1：邻居请求(Neighbor Solicitation，NS)
  - 类似于IPv4中的ARP请求报文，获取邻居的链路层地址，验证邻居可达，重复地址检测
- 消息类型2 ： 邻居通告(Neighbor Adivertisment，NA)
  - 类似于IPv4中的ARP应答报文，对NS消息进行响应
- 消息类型3 ： 路由器请求(Router Solicitation，RS)
  - 端系统通过RS消息向路由器发出请求，请求地址前缀和其他信息，用于节点的自动配置
- 消息类型4 ： 路由器通告(Router Advertisement，RA)
  - 路由器通过RA消息向端系统发布地址前缀（IPv6地址自动配置）和其他配置信息
- 消息类型5 ： 重定向(Redirect)
  - 通知主机重新选择正确的下一跳地址（针对某个目的IPv6地址）

IPv6路由协议

- RIPng for IPv6，RFC 2080，对RIP修改以适应IPv6环境
  - 使用路由器的链路本地IPv6地址作为源地址，发送路由更新信息
- OSPFv3 for IPv6，RFC 5340，适应IPv6网络
  - 使用路由器的链路本地IPv6地址作为源地址，并作为下一跳地址
  - OSPFv3有7种类型的LSA，新增Link LSA和Intra Area Prefix LSA
- MP-BGP(Multi-Protocol BGP)，RFC 4760，支持多种网络层协议（IPv6和IPX）
  - MP-BGP向前兼容，并支持组播
  - BGP连接可以是IPv4或IPv6，报文内可传递其它网络协议的路由信息
  - 多协议可达NLRI描述了到达目的地的信息：地址属于哪个网络层协议，下一跳地址

IPv4 -> IPv6

网络层基础：IPv4
动态地址分配：DHCP
单跳网络处理：ARP
基础管理和控制：ICMP
距离向量路由：RIP
链路状态路由：OSPF
外部网关路由：BGP

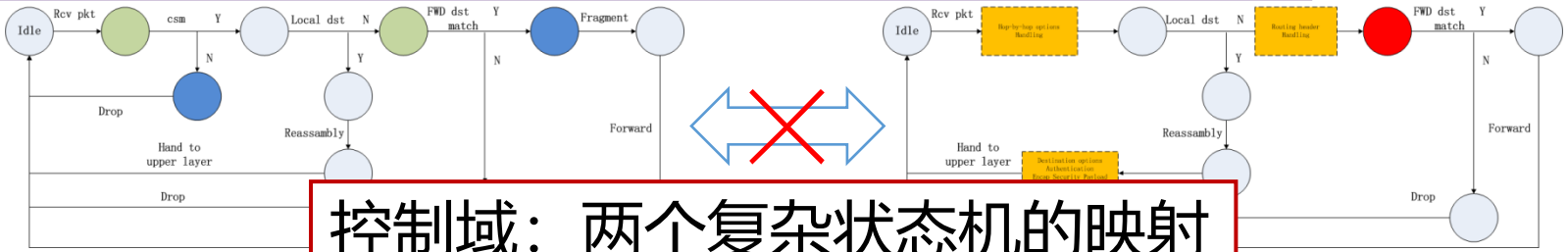
网络层基础：IPv6
动态地址分配：DHCPv6
单跳网络处理：ND
基础管理和控制：ICMPv6
距离向量路由：RIPng
链路状态路由：OSPFv3
外部网关路由：MP-BGP

- 双协议栈技术
  - 每个设备同时运行IPv4和IPv6两个协议栈
  - 无法解决IPv4地址不足的问题，不能促进IPv6部署

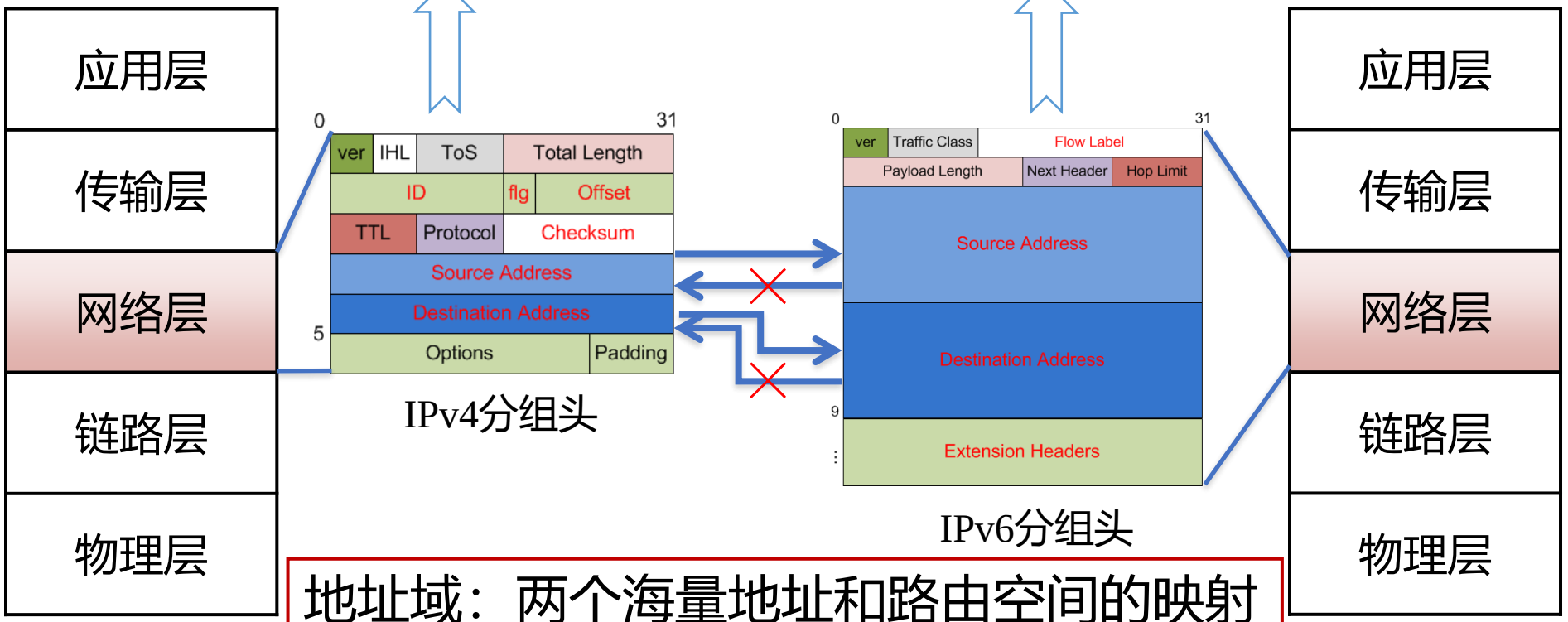
尽量保证端到端透明原则
（用户无感知）

- IPv4和IPv6技术并不能兼容，怎么办？
  - 完全抛弃IPv4，建立新的IPv6网络？
  - 设计过渡技术，同步发展，逐步切换到IPv6网络？更合理的选择

IPv4到IPv6过渡技术难点



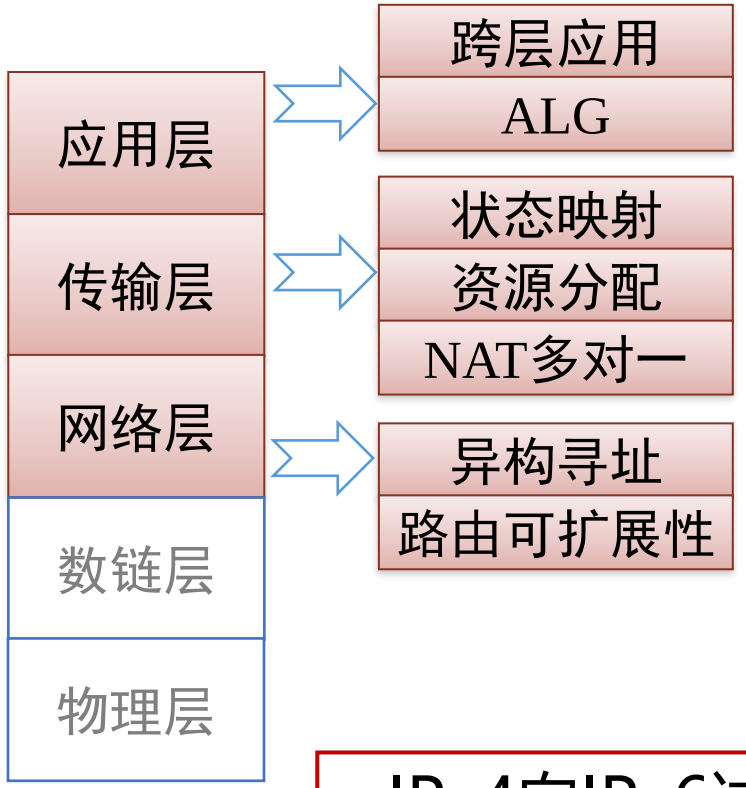
控制域：两个复杂状态机的映射



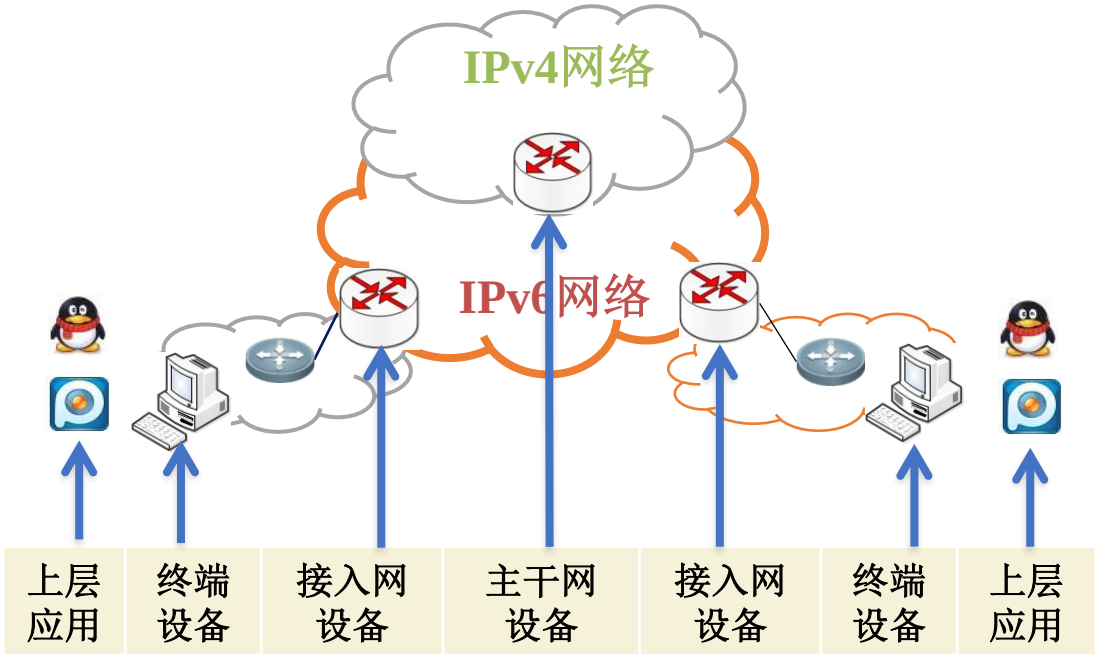
地址域：两个海量地址和路由空间的映射

IPv4到IPv6迁移及过渡技术难点

纵向贯穿体系结构各层次



横向涉及网络结构各网元

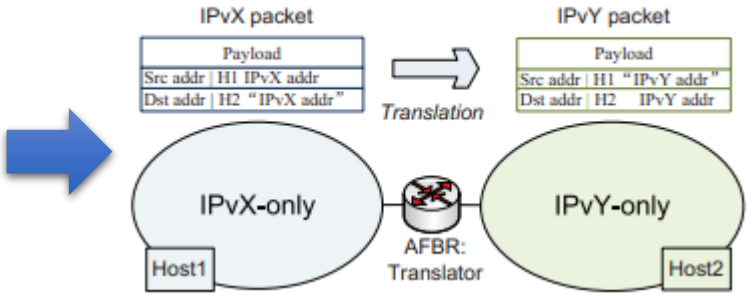


IPv4向IPv6过渡是发展下一代互联网的重大技术难题

两种过渡技术路线

翻译技术

形式化系统
相互转换

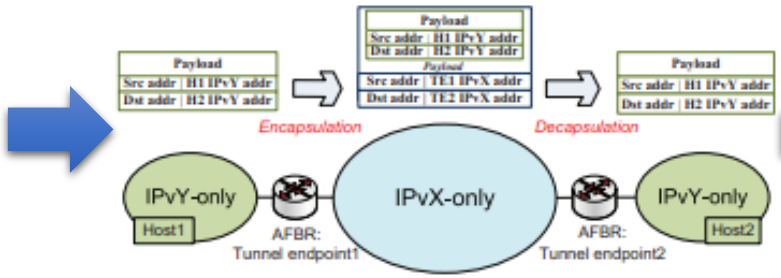


难实现完全转换

适用于特定
过渡场景
(RFC6144)

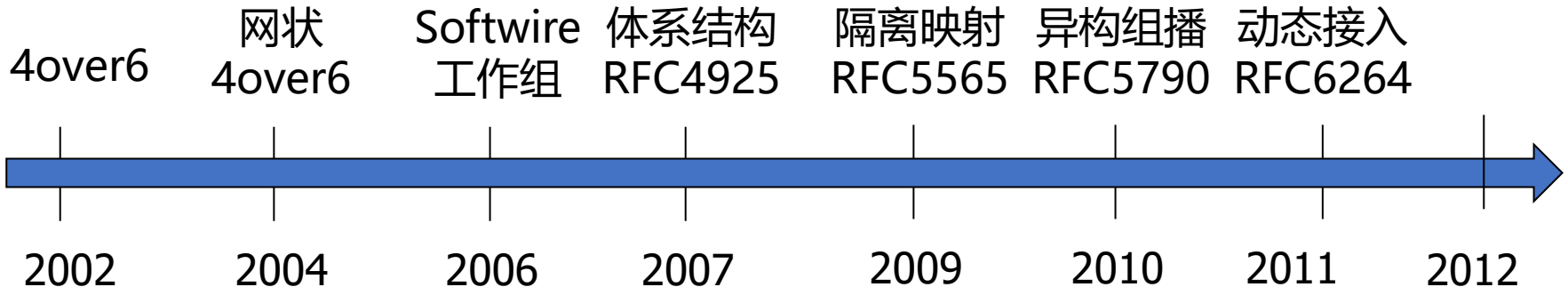
隧道技术

形式化系统
相互承载



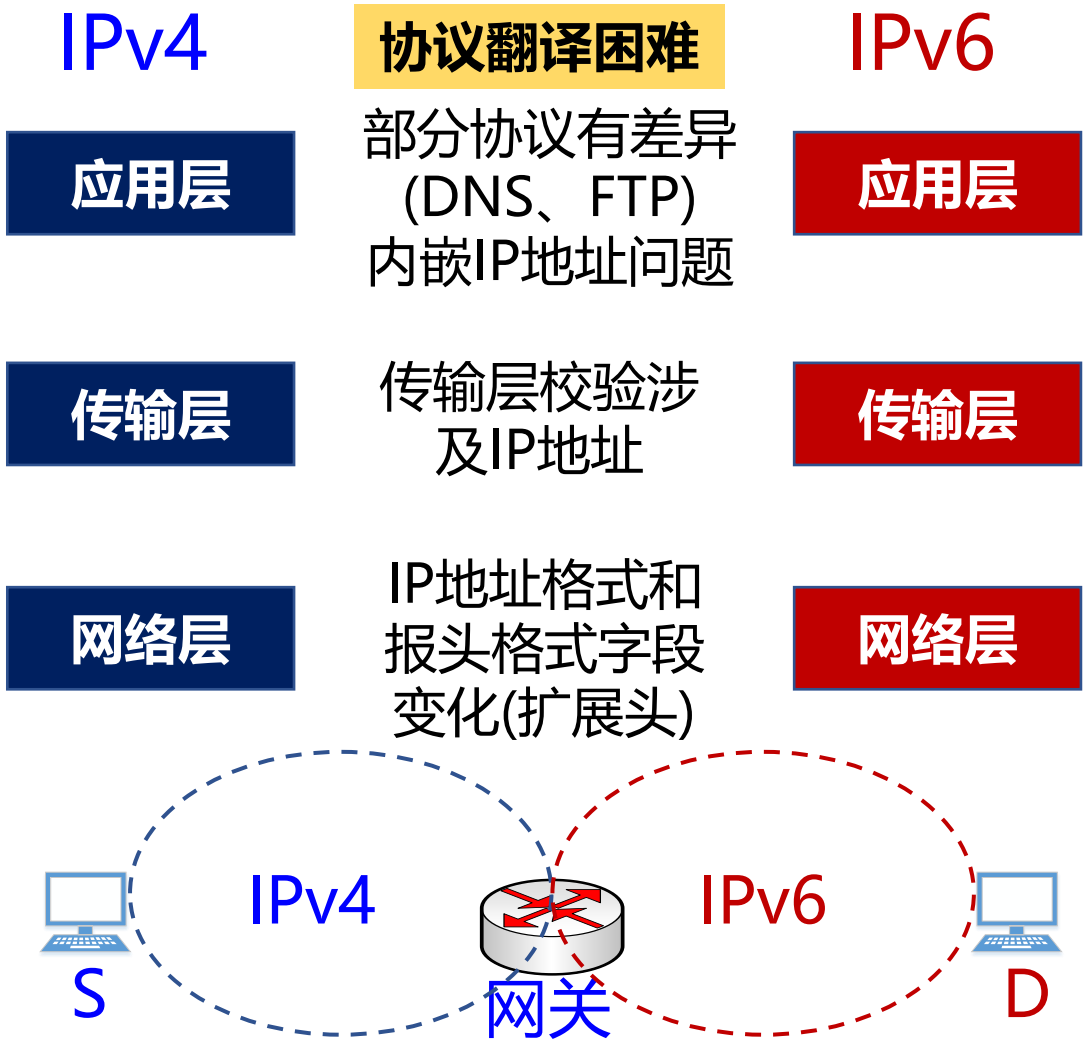
可实现完全承载

广泛适用于
各种场景

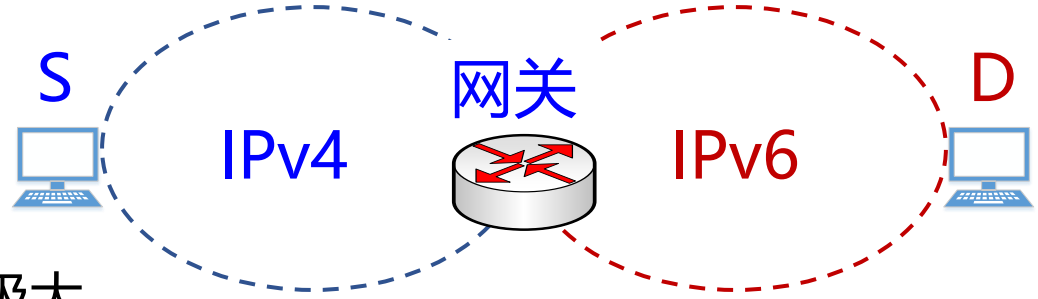


IPv4/IPv6翻译

- IPv4-IPv6地址往返转换
- IPv4报头和IPv6报头翻译
  - 各字段对应，包括IP选项翻译
- 传输层校验和转换，涉及IPv6地址
- ICMP翻译
  - ICMPv6重新设计了类型值和代码值
- 其它应用层协议翻译
  - FTP（命令名称变化，内嵌IP地址）
  - 内嵌IP地址的其它应用层协议
- DNS处理
  - 实现域名查询中A记录和4A记录的双向翻译



IPv4/IPv6翻译的问题



- 路由可扩展性问题
  - 96位前缀？给路由聚合带来巨大困难
- 应用层内嵌IP地址
  - 部分应用层协议内嵌IP地址，应用层翻译难度极大
- 破坏互联网端到端原则
  - 通信双方只看到通信对端在本网络对应的地址
  - 翻译网关处理传输层、应用层数据（上层数据无法透明传输）
- 异构地址寻址问题
  - 通信双方至少有一方需感知本次通信涉及翻译技术
  - 域名访问需DNS服务支撑
- 翻译导致报文变长可能导致分片问题，影响转发性能

端到端加密呢？

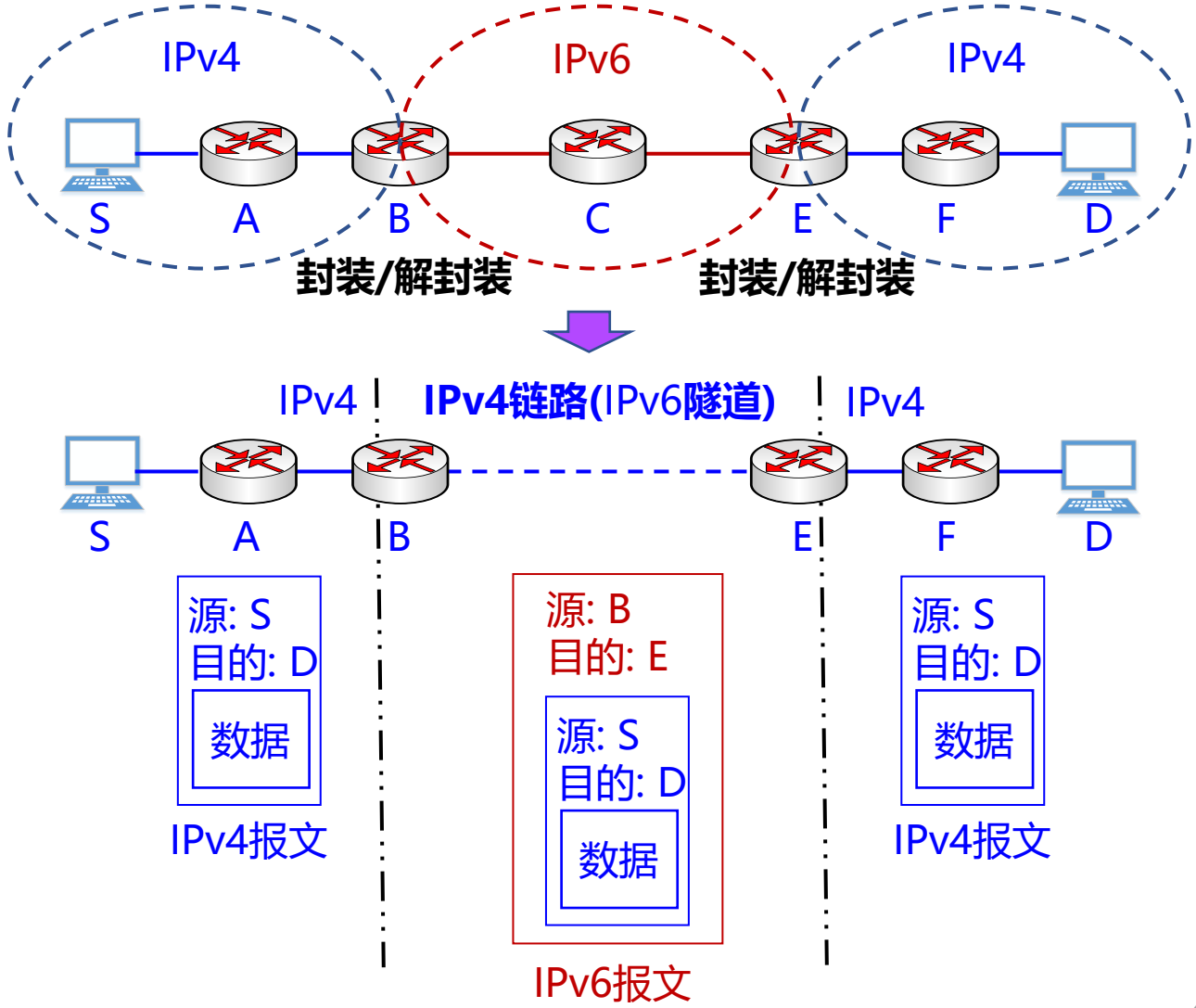
隧道技术

隧道技术

- 同构网络跨越异构网络进行通信
- 将A协议数据包封装在B协议中传输

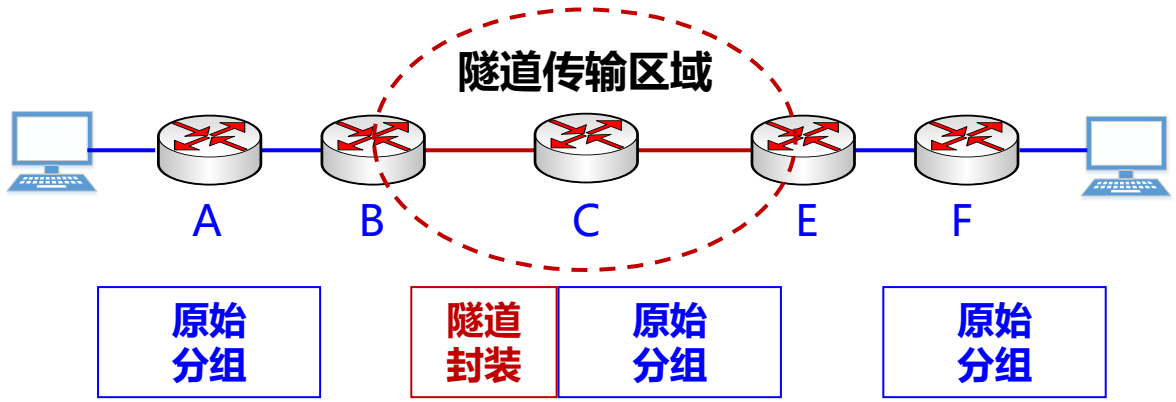
隧道类型

- 应用层隧道
  - SSH隧道，HTTPS隧道
- 传输层隧道
  - TCP隧道，UDP隧道
- 网络层隧道
  - 4 in 4，4 in 6，6 in 4
  - GRE，通用路由封装隧道
- 链路层隧道
  - L2TP协议，链路层隧道
  - PPTP协议，点对点隧道

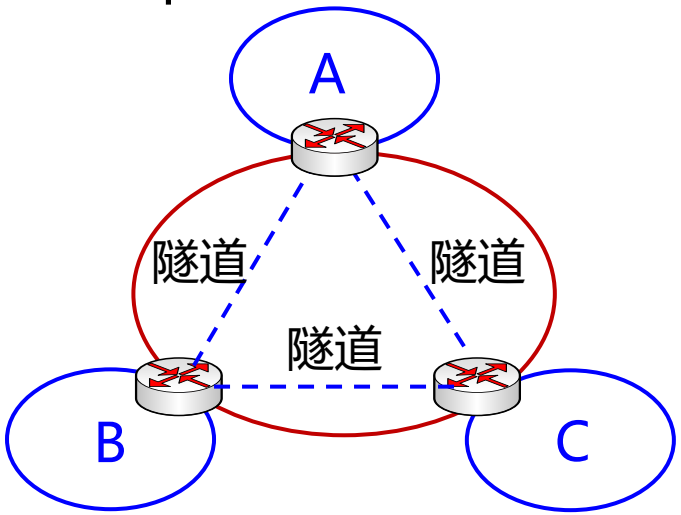


隧道技术

- 报文长度增大导致的分片问题
  - 途中分片与重组对传输性能影响较大
- 解决方法
  - 隧道网关提前分片
  - Path MTU发现机制
- 隧道链路的选择问题
  - A发往F的报文，一路上如何路由？



- 静态配置
  - ip route 30.0.0.0/8 tunnel 1
- 4over6：与路由结合的动态学习
  - 主干网：扩展核心路由协议BGP
  - 接入网：DHCP协同NAT，实现轻量级
  - 从嘲笑、怀疑，到争吵追随
  - OpenWRT，中国电信、德电、法电



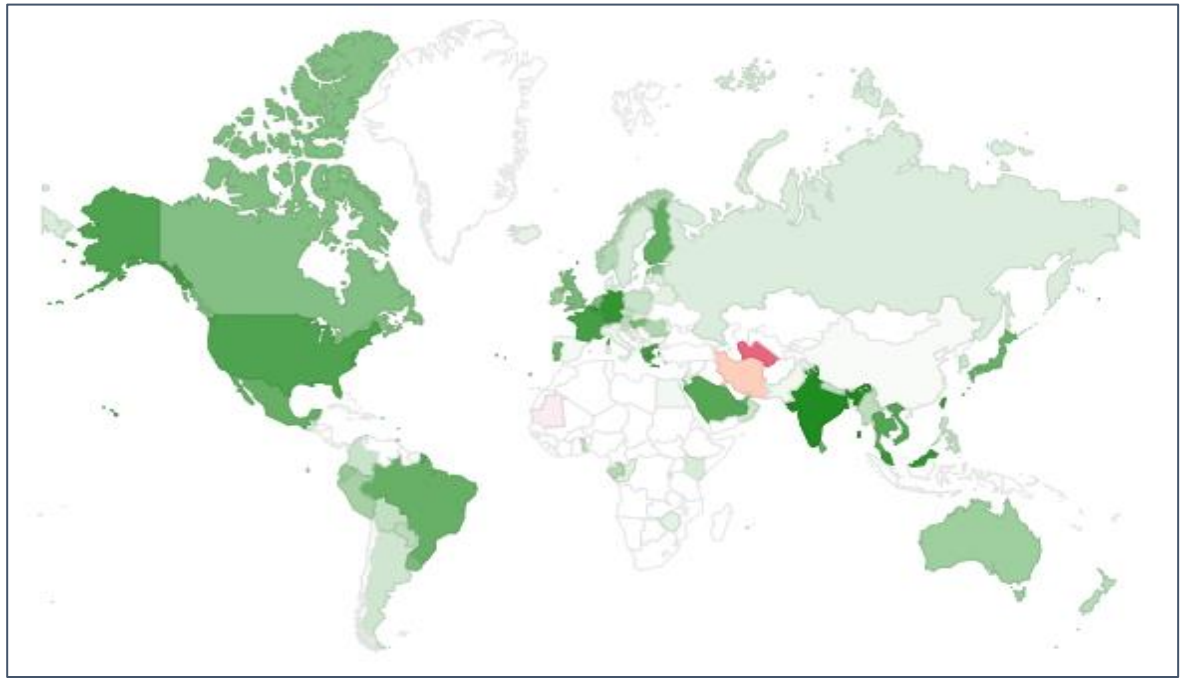
全球IPv6下一代互联网快速发展

全球IPv6互联网进入快速发展期

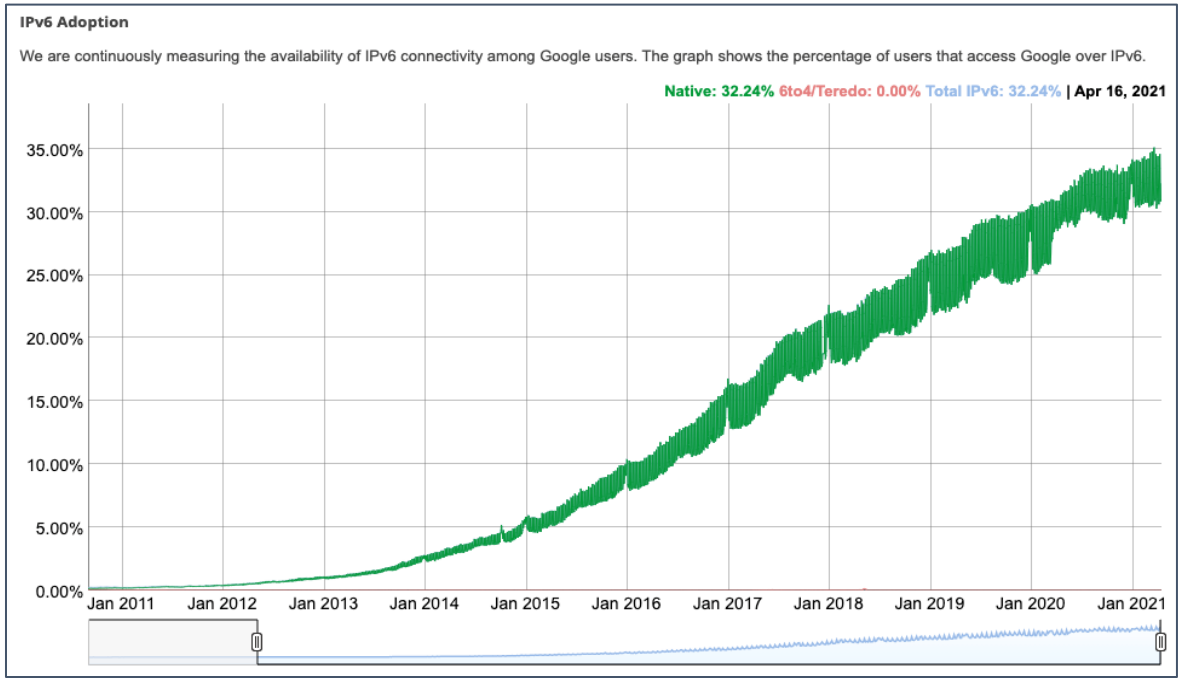
IPv6部署率：印度56.7%，德国51.6%，法国46.9%，美国44.7%，巴西38.7%，日本37.8英国34.1%，加拿大31.5%

双循环
全球化

截至2021年4月，IPv6部署率在30%以上占一半以上



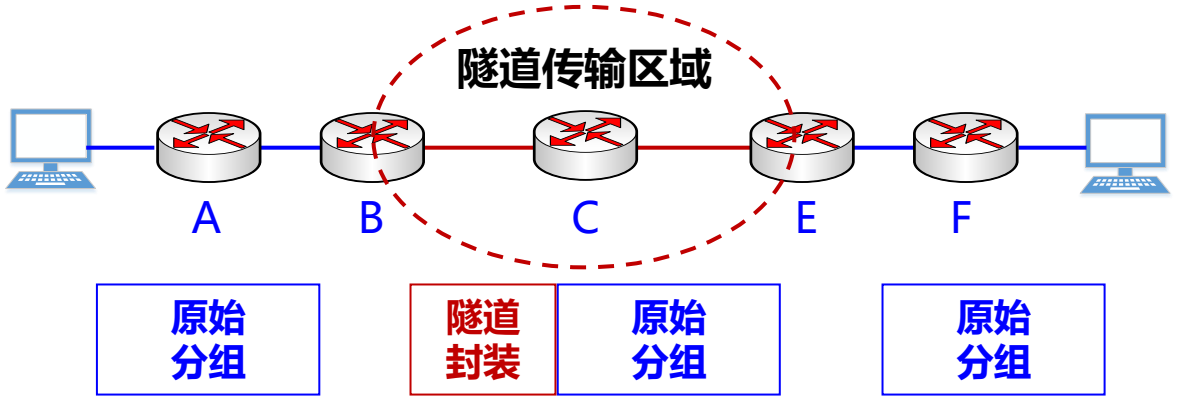
使用IPv6访问Google网站的用户占比最高达34.4%



NAT、IPv6技术小结

- 相似的出发点：IPv4地址不够用？
  - 在IPv4的框架中优化：NAT技术
  - 拓展地址位数，新设计：IPv6技术
- NAT技术
  - NAT转换表：内网(IP，端口)<->外网(IP，端口)
- IPv6技术
  - 地址空间：32 -> 128位
  - IP协议族和路由协议的IPv6设计
  - IPv4/IPv6共存：翻译、隧道

网络层基础：IPv6
动态地址分配：DHCPv6
单跳网络处理：邻居发现
基础管理和控制：ICMPv6
距离向量路由：RIPng
链路状态路由：OSPFv3
外部网关路由：MP-BGP



本节内容

6.6 标签交换和MPLS

6.7 路由器体系结构

6.8 NAT技术

6.9 IPv6技术

6.10 服务质量和拥塞控制算法

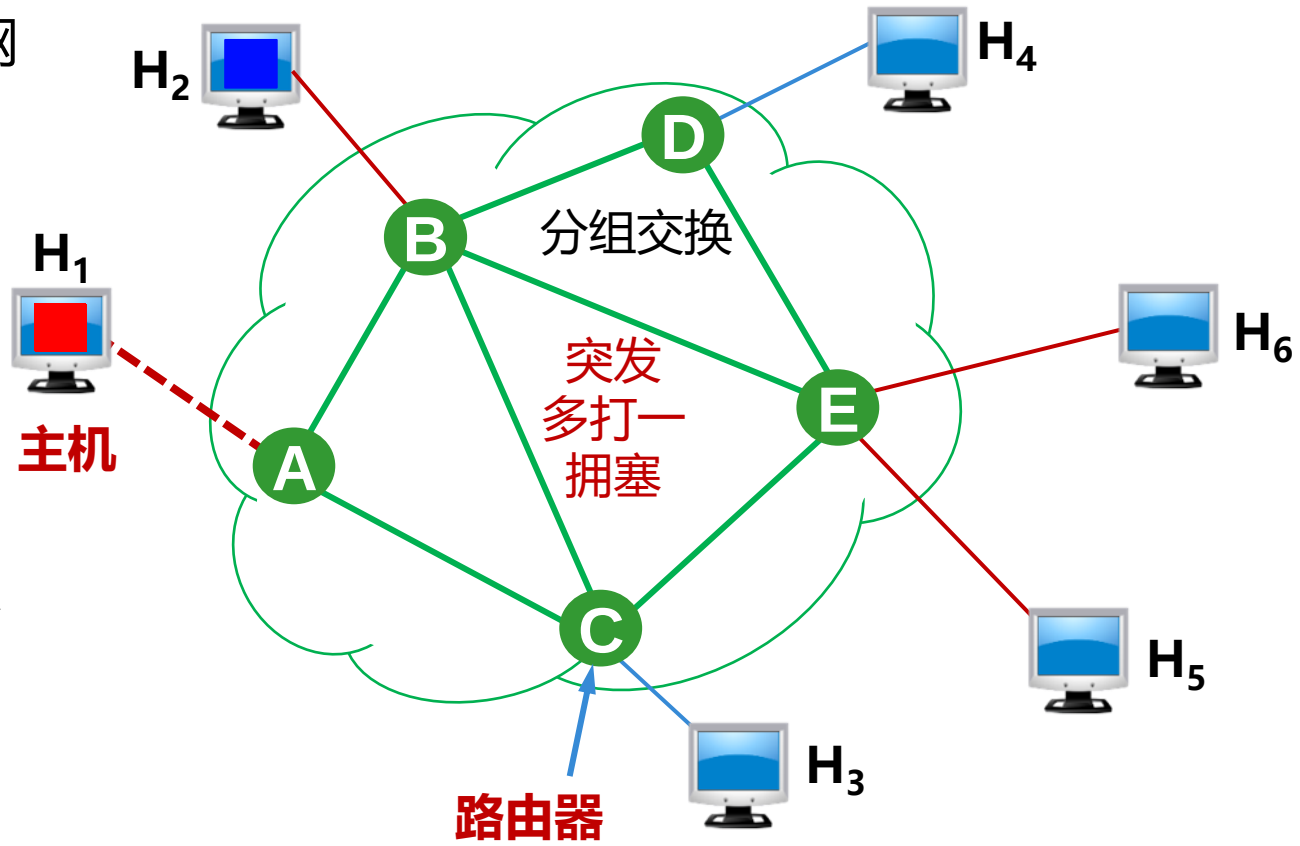
6.11 软件定义网络SDN

1. 网络服务质量概述
2. 流量整形
3. 综合服务
4. 区分服务
5. 拥塞控制概述
6. 流量感知路由
7. 流量调节
8. 随机早期检测

思考与发明

- 尽力而为的互联网☹
  - 视频会议、VR、工业互联网、车联网
- 服务质量QoS
  - 带宽、时延、抖动、丢包率等指标
  - 互联网v.s.电信网：如何保障QoS?
- 网络性能问题分析
  - 短时发送过快，占满节点缓存
  - 没有差异化服务，流量只能平等排队
  - 流量规模超出带宽上限，网络满载

如何避免令人讨厌的
多打一、突发和拥塞?



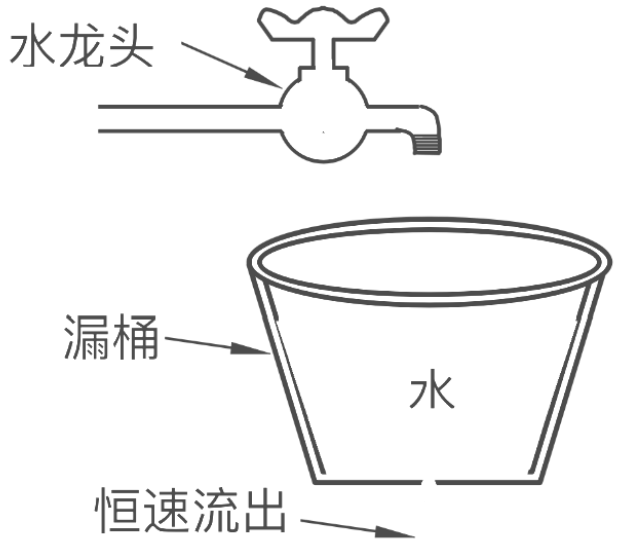
流量整形

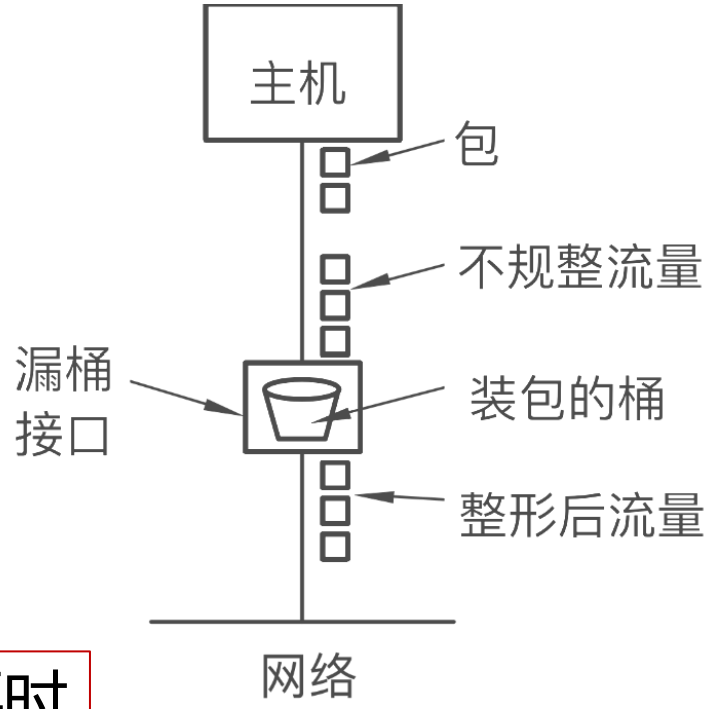
- 流量整形(traffic shaping)
  - 其作用是限制流出某一网络的某一连接的流量与突发，使这类报文以比较均匀的速度向外发送
- 流量整形算法包括漏桶算法和令牌桶算法
  - 漏桶算法（Leaky Bucket Algorithm）：主要目的是控制数据注入到网络的速率，平滑网络上的突发流量，突发流量可以被整形以便为网络提供一个稳定的流量
  - 令牌桶算法（Token Bucket Algorithm）：用来控制发送到网络上的数据的数目，并允许突发数据的发送

流量整形

漏桶算法原理

- 到达的数据包（网络层的PDU）被放置在底部具有漏孔的桶中（数据包缓存）
- 漏桶最多可以**排队b个字节**，漏桶的这个尺寸受限于内存。如果数据包到达的时候漏桶已经满了，那么数据包应被丢弃
- 数据包从漏桶中漏出，以常量速率**（r字节/秒）**注入网络，因此平滑了突发流量



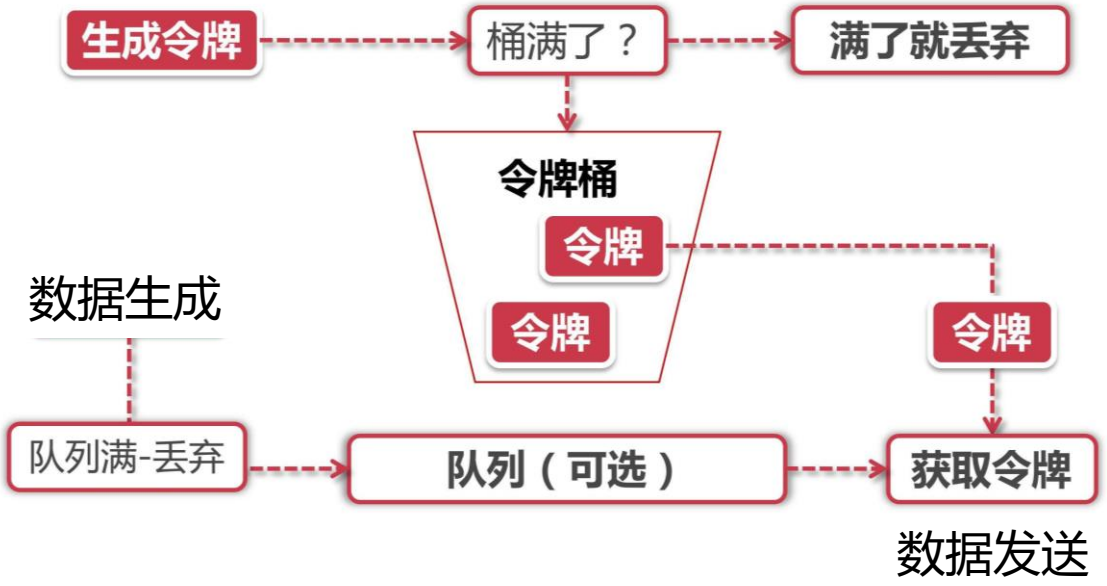


先多停一会儿，需要时能用我的配额吗？

流量整形

令牌桶算法工作原理

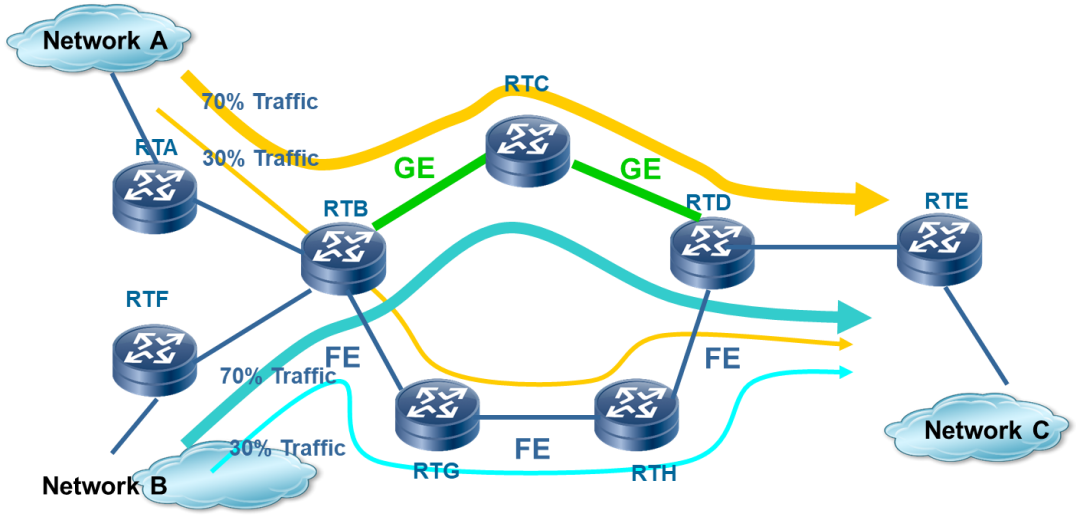
- 产生令牌：周期性的以速率r向令牌桶中增加令牌，桶中的令牌不断增多。如果桶中令牌数已到达上限，则丢弃多余令牌
- 消耗令牌：输入数据包会消耗桶中的令牌。在网络传输中，数据包的大小通常不一致。大的数据包相较于小的数据包消耗的令牌要多
- 判断是否通过：输入数据包经过令牌桶时存在两种可能：输出该数据包或者被丢弃。当桶中的令牌数量可以满足数据包对令牌的需求，则将数据包输出，否则将其丢弃



综合服务IntServ

- 综合服务/集成服务IntServ（Integrated Services）
  - 所有路由器处理每个流的消息，维护每个流的路径状态和资源预留状态，在路径上执行基于流的分类、调度、管理
  - 基于资源预留协议RSVP，面向连接，逐节点建立或拆除流的状态和资源预留状态，根据流的状态进行QoS路由
  - 在每个路由器上通过**每流**状态维护，实现QoS精细化管理
- 综合服务的特征：资源预分配、全局流状态、传输控制

Lixia Zhang
UCLA



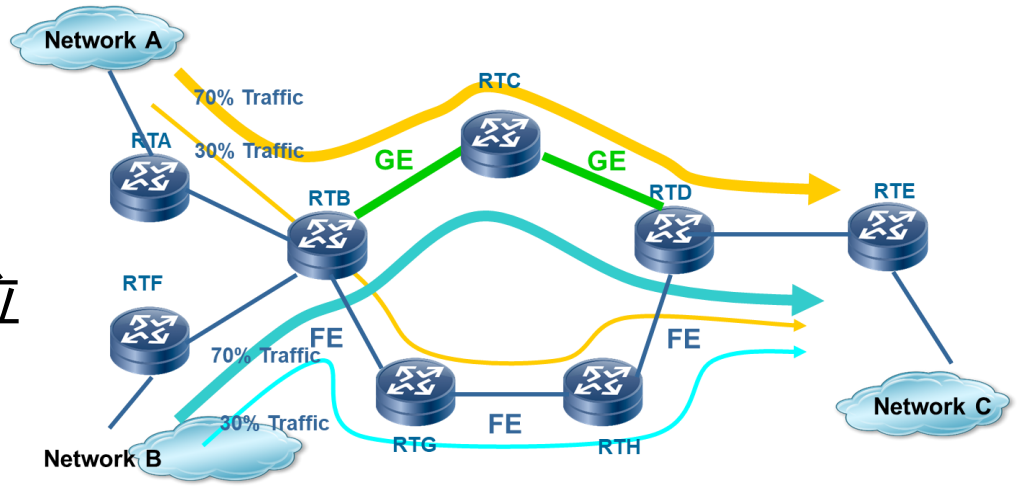
核心难点与改进思路?

互联网设计原则

区分服务DiffServ

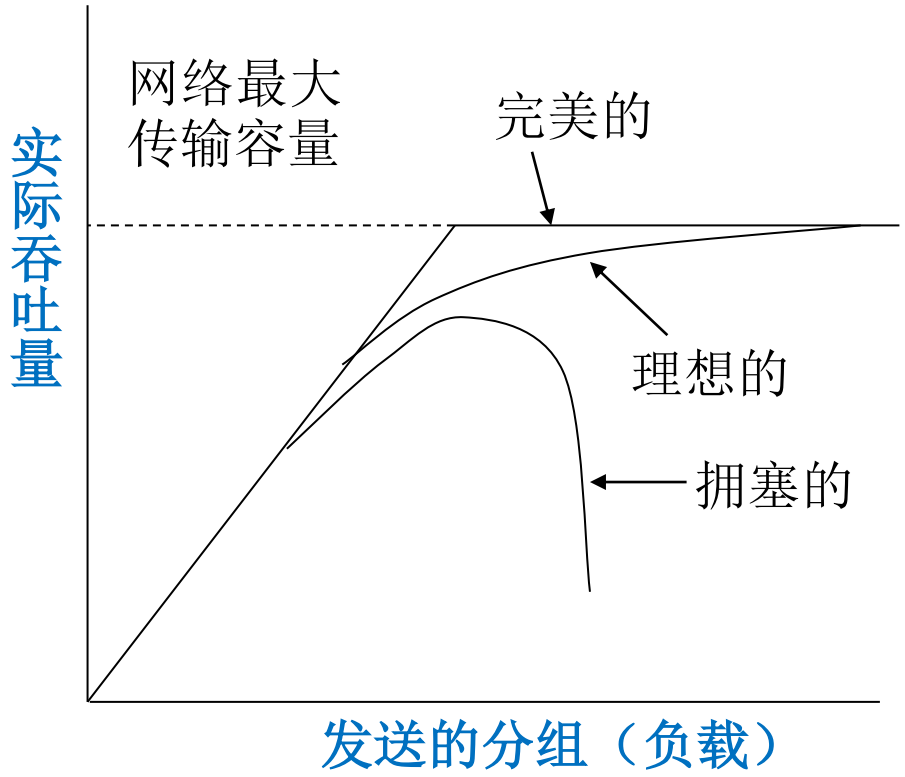
- 区分服务（DiffServ：Differentiated services）
  - 拒绝每流而采用每类，针对每个类别实现不同服务质量（领导人专车、救护车、出租、公交、自行车……）
  - DiffServ：一种简单且可扩展的机制，在IP网络上分类和管理网络流量，提供服务质量QoS控制
  - 边界节点根据约定好的QoS规定，把将要进入网络的流量分类成不同的流
  - IPv6报头的8位区分服务DS字段中，使用6位区分服务码点（DSCP）进行分组分类

核心难点与挑战？
边界如何映射



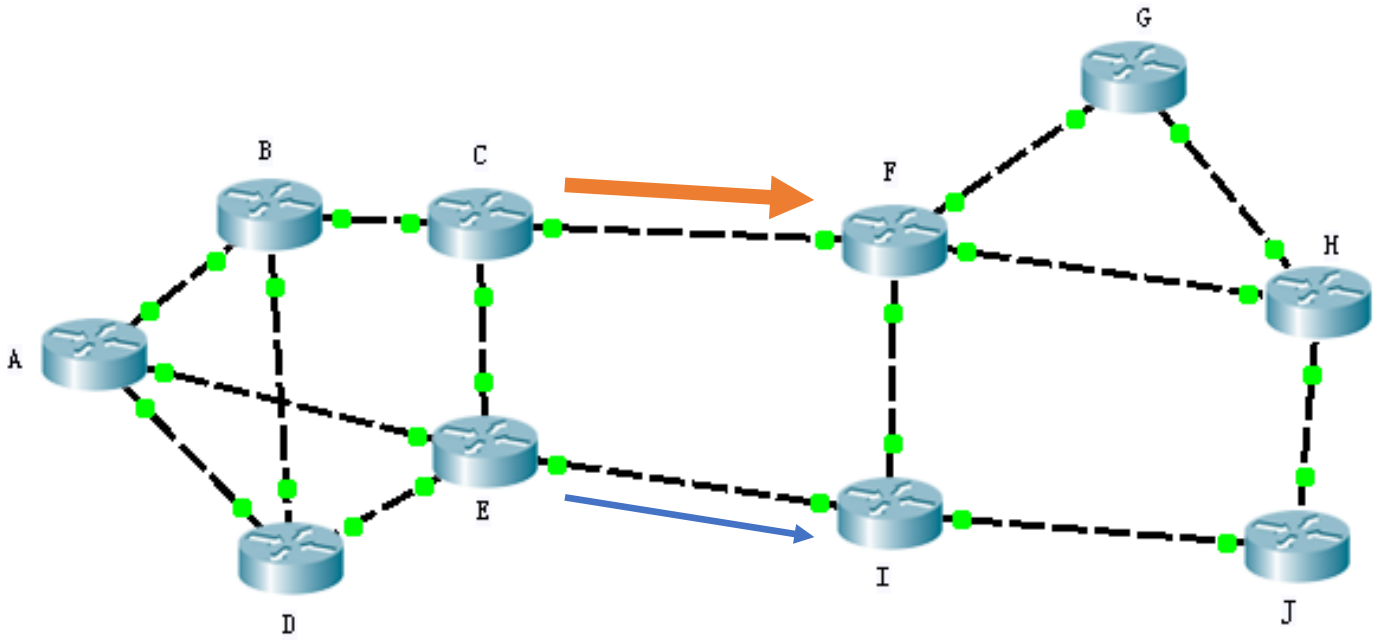
拥塞控制概述

- 拥塞
  - 网络中存在太多的数据包导致数据包传输延迟或丢失，从而导致网络吞吐量下降
- 拥塞控制（congestion control）
  - 需要确保通信子网能够承载用户提交的通信量，是全局性问题，涉及主机、路由器等多种因素
- 产生拥塞的原因
  - 主机发送到网络的数据包数量过多，超过了网络的承载能力
  - 突发的流量填满了路由器的缓冲区，造成某些数据包会被丢弃



服务质量路由QoSR

- 思路：绕开热门的区域，疏散流量
- 方法：计算路径权重时包含跳数、带宽、传输延迟、负载、排队延迟等
- 问题：路由表可能会出现反复变化，从而导致不稳定的路由



发的数据实在太多怎么办？

海量终端如何互相协调？

流量调节

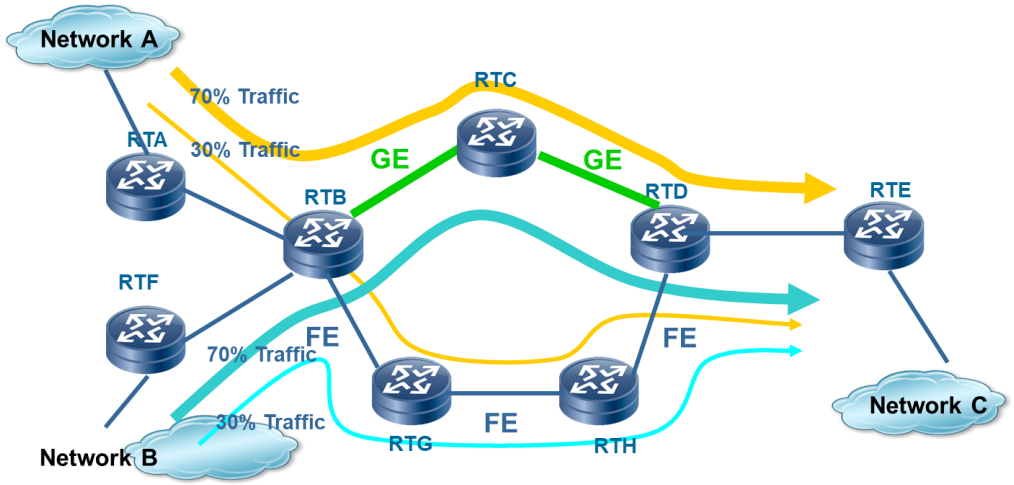
➢ 抑制包(Choke Packets)

- 用于通知发送方减小发送量，路由器选择一个被拥塞的数据包，给该数据包的源主机返回一个抑制包，抑制包中的目的地址取自该拥塞数据包
- 源主机收到抑制包后，减少发向特定目的地址的流量

如何提前预警，
让源端少发一些?

➢ 逐跳的抑制包(Hop-by-Hop Choke Packets)

- 在高速或长距离网络中，由于源主机响应太慢，抑制包算法对拥塞控制的效果并不好，可采用逐跳抑制方法
- 其核心思想是抑制包对它经过的每个路由器都起作用，能够迅速缓解发生拥塞处的拥塞，但要求上游路由器有更大的缓冲区



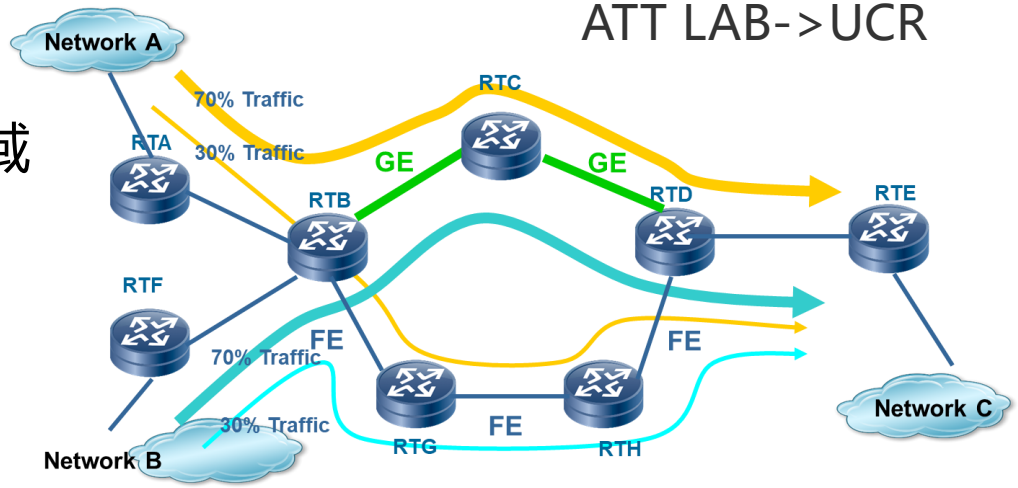
流量调节

- 显式拥塞通告ECN
  - ECN：Explicit Congestion Notification
  - 在IP包头中记录数据包是否经历了拥塞
  - 在数据包转发过程中，路由器可以在包头中标记为经历拥塞
  - 接收方在下一个应答数据包里回显该标记作为显式拥塞信号
  - 发送端降速
- 写入IETF国际标准
  - RFC3168定义IP头的TOS域未使用的两位为ECN域
  - 00：发送主机不支持ECN
  - 01或者10：发送主机支持ECN
  - 11：路由器正在经历拥塞

可能的问题?

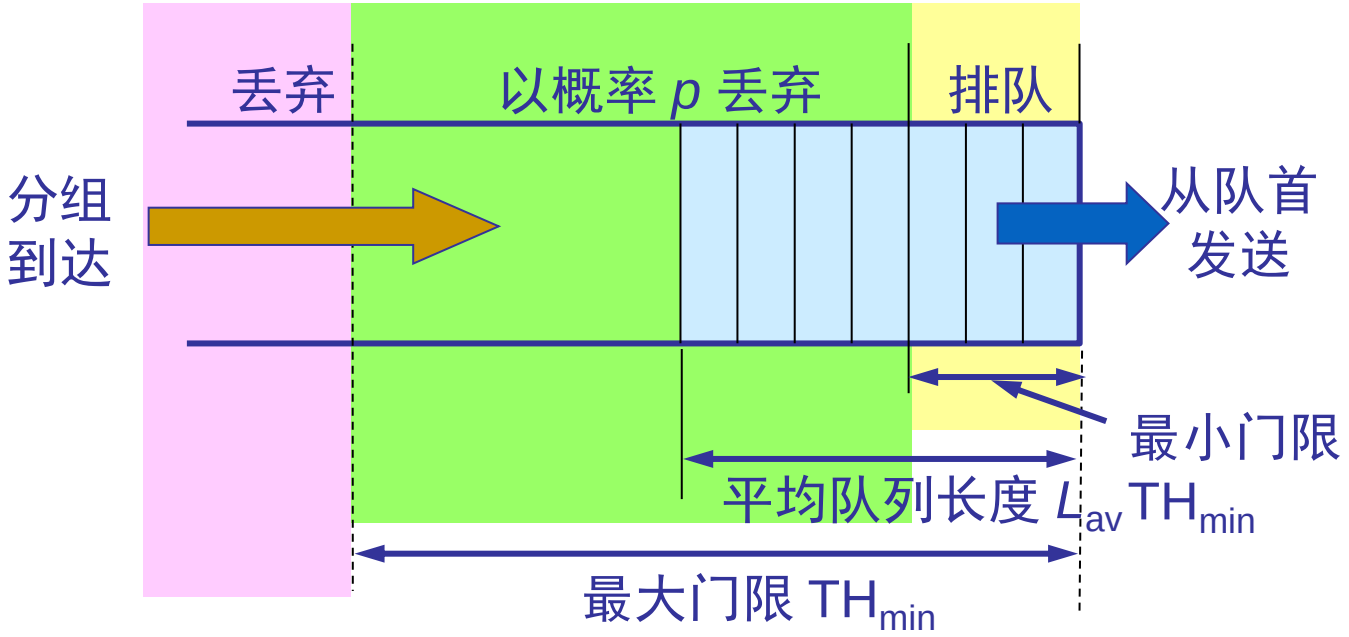
避免端网协同?

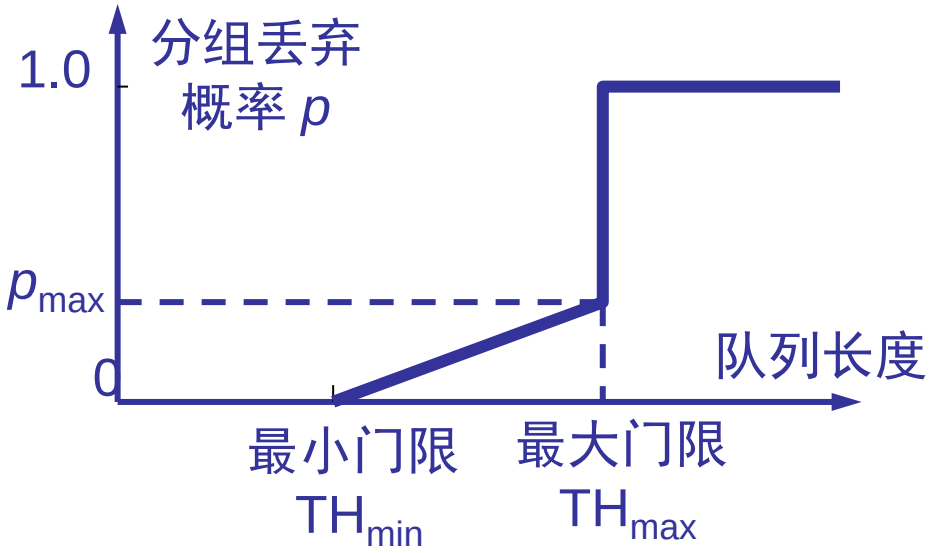
K. K. Ramakrishnan
ATT LAB->UCR



随机早期检测

- 传统方式的问题
  - Buffer满了只能全部丢掉，所有流量损失惨重，能否不要太生硬?
- 随机早期检测 RED (Random Early Detection)
  - RED 将路由器的到达队列划分成为三个区域





服务质量和拥塞控制算法小结

- 如何用上更好的网络服务?
  - 更好? 网络服务质量QoS（带宽、时延、抖动、丢包率）
- 网络层QoS
  - 突发流量 <-> 流量整形
    - 限制突发流量产生，避免节点缓存波动导致的丢包
  - 缺少差异化服务<->综合服务、区分服务
    - IntServ逐流维护状态，DiffServ按照每类处理优先级
  - 流量超出带宽<->拥塞控制
    - 流量感知路由：通过调度流量绕开热门区域，疏解流量
    - 端网协同的流量调节：抑制包（显示拥塞通告ECN），逐跳反压机制
    - 网络节点的随机早期检测：未雨绸缪，缓解即将到来的拥塞

本节内容

6.6 标签交换和MPLS

6.7 路由器体系结构

6.8 NAT技术

6.9 IPv6技术

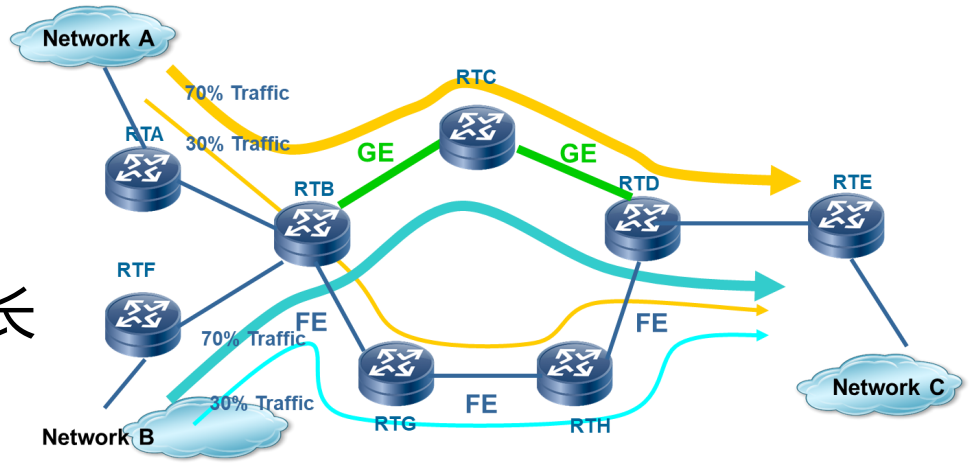
6.10 服务质量和拥塞控制算法

6.11 软件定义网络SDN

1. 新型网络需求分析
2. 软件定义网络SDN

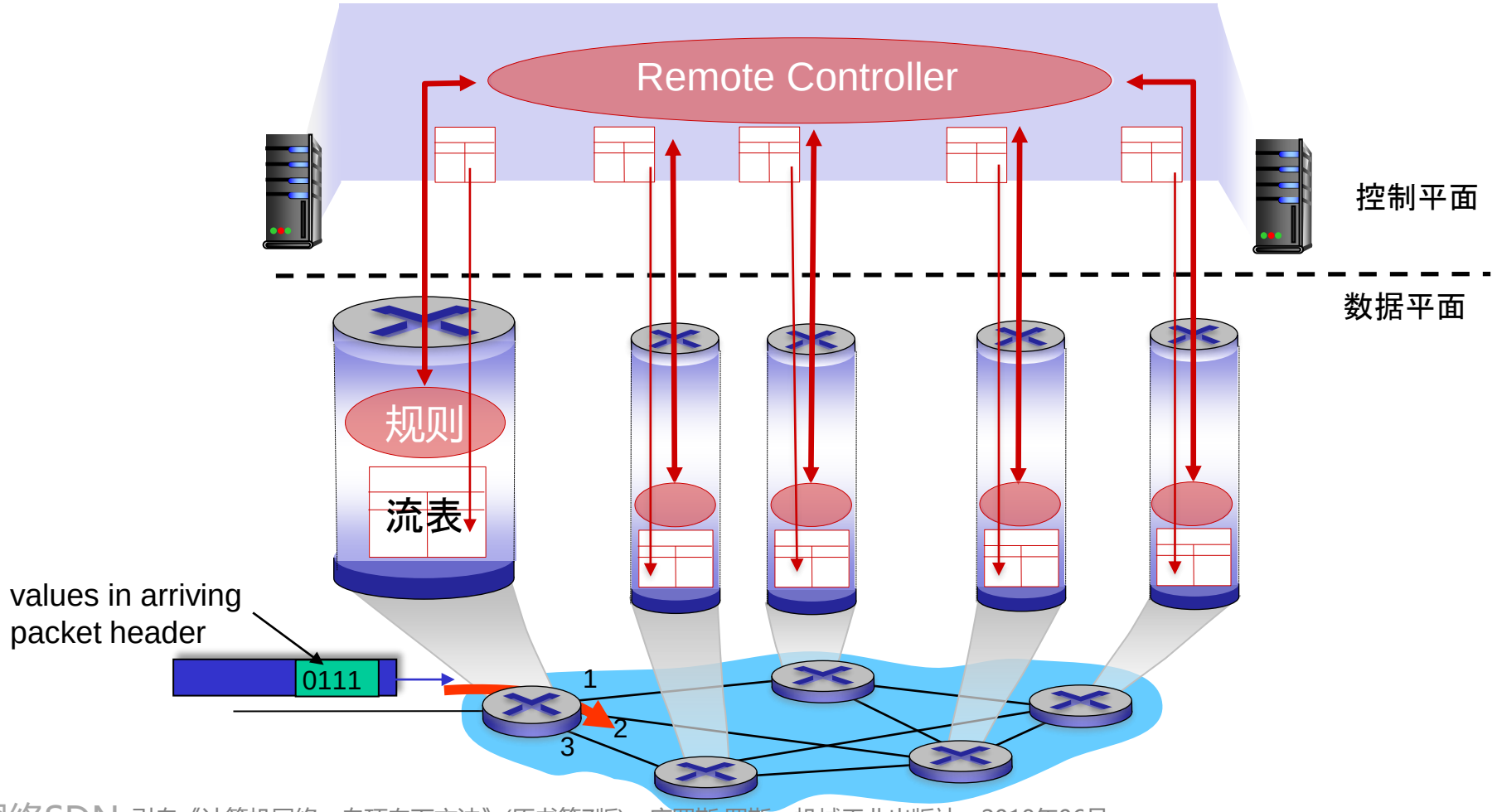
传统网络面临的问题

- 分布式路由的缺点
  - 网络设备以接力棒形式告诉下一跳邻居设备
  - 缺少全局控制，难以优化
  - 底层网络设备数量不断增加，路由收敛时间长
- 传统网络不可编程，运营商难以掌控
  - 互联网核心节点采用硬件实现，价格高，门槛高
  - 运营商的网络新需求难以设备难以支持，运营商难以配置
  - 很难实现高效、按需的数据传输
- 运营商和互联网厂商的呼唤
  - 是否能有简单的硬件，实现灵活的组网和配置？



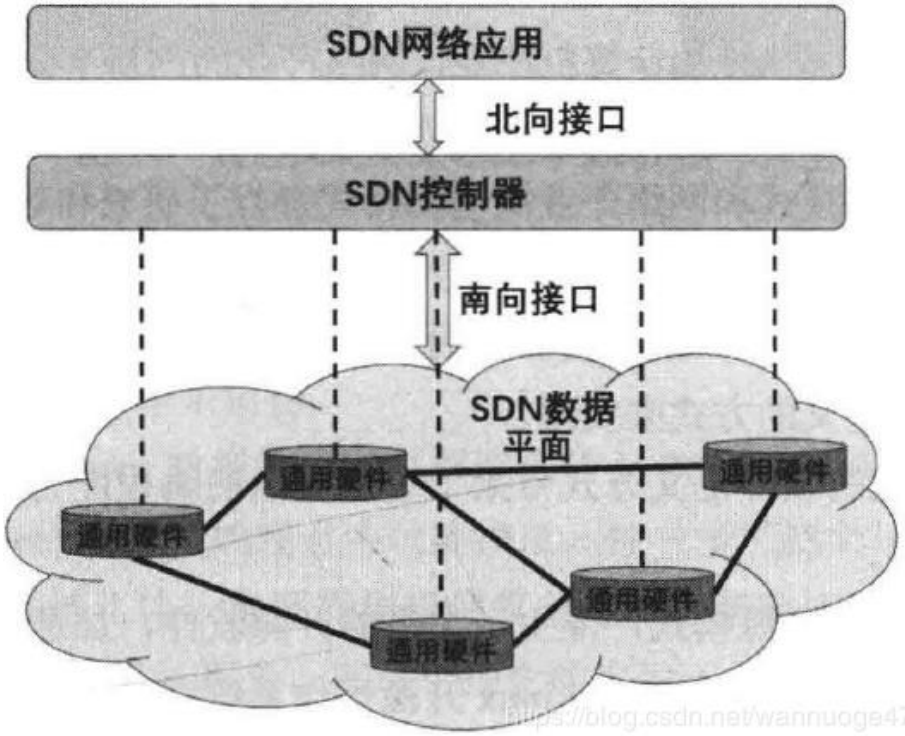
SDN：控制平面和数据平面分离

远程控制、计算、配置SDN交换机



软件定义网络SDN

- 软件定义网络（SDN）
  - 新的网络体系结构：分离转发面与控制面
- SDN控制器集中控制
  - 具有全局视野，可以动态控制网络
  - 根据应用数据对于网络的需求，为其计算出最优路径，并通过下发流表控制交换机转发此类数据，进而实现高效、按需的数据传输
- 主要接口
  - 北向接口：向上提供，进行适配应用和管理网络
  - 南向接口：向下提供，管理转发面设备



网络的成本去哪里了？

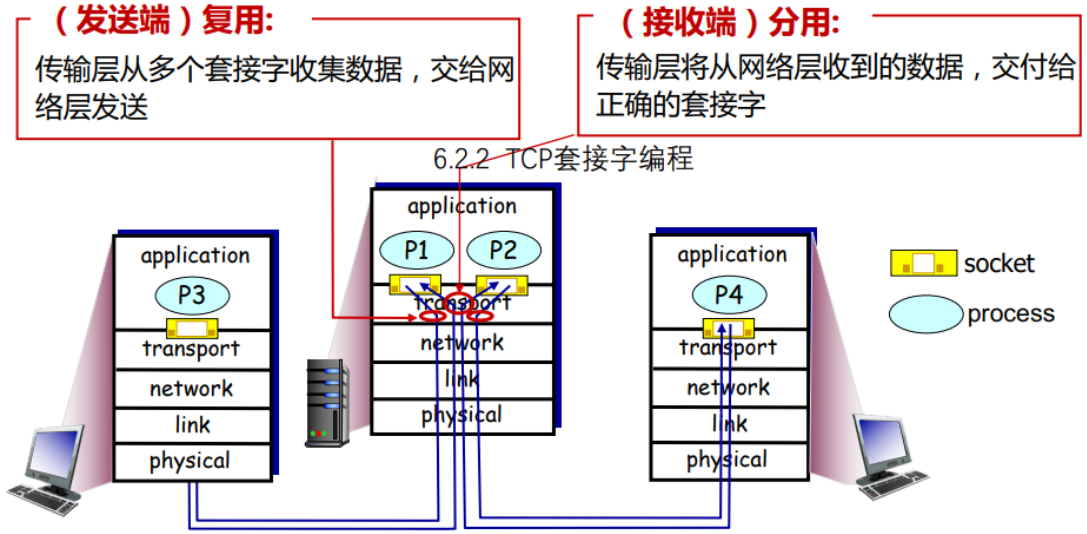
集中式网络与路由服务器

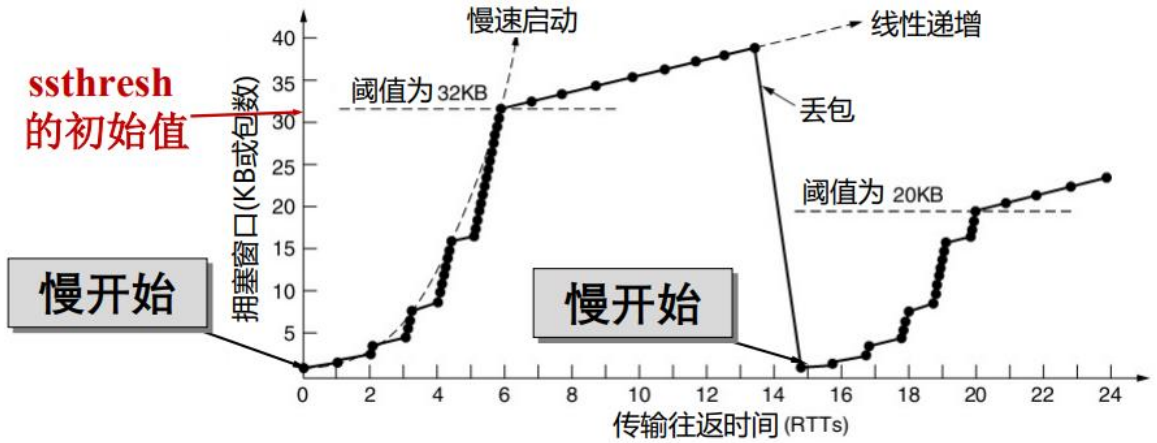
总结

- 标签交换和MPLS
  - 借鉴电路交换，面向连接；更强控制，实现VPN和TE
- 路由器体系结构
  - 解决从概念到实现的问题；控制层和数据层，报文转发，交换结构
- 下一代互联网IPv6
  - IPv4框架内思路：网络地址转换NAT，用端口号复用IP地址
  - IPv6技术：更长的地址位数，新协议设计，IPv4/IPv6过渡技术
- 服务质量和拥塞控制算法：网络层性能提升技术
- 软件定义网络SDN：网络层功能灵活管理

下周预告

- 网络层解决多跳网络传输，实现主机间的交付
  - 单个设备的多个应用进程间如何共享网络？
  - 你希望网络可靠吗？可惜……
  - 网络拥塞怎么办？
- 期待传输层的思考与发明之旅





作业

- 《Computer Networks-5th Edition》章节末习题
  - CHAPTER 5：2(虚电路交换)、22（区分服务）、40（IPv6）、42（IPv6邻居发现）
  - 小实验4（ICMPv6和ND数据包观察，见网络学堂附件）

- 截止时间：下下周三晚11:59，提交网络学堂

致谢社区本章贡献者

清華大學 Tsinghua University
计算机网络教案社区

贡献者姓名	单 位	贡献内容
陈文龙	首都师范大学	本章统稿 5.5 5.9(IPv6协议)
吴黎兵	武汉大学	5.6 5.7
谢晓燕	西安邮电大学	5.8
邹莹	仲恺农业工程学院	5.4.1 5.4.2 5.4.4
李旭宏	枣庄学院	5.1.2 5.1.3 5.2.3 5.2.4 5.2.6 5.4.3
曲大鹏	辽宁大学	5.3.1 5.3.2
方诗虹	西南民族大学	5.1.4 5.3.5 5.3.6 5.3.7 5.3.8
舒挺	浙江理工大学	5.1.1
白云莉	内蒙古农业大学	5.3.3 5.3.4
余琨	荆楚理工学院	5.2.1 5.2.2 5.2.5
李振斌	华为技术有限公司	5.9(SRv6)

致谢社区本章贡献者

陈文龙

首都师范大学

5.路由器工作原理
9.IPv6技术

吴黎兵

武汉大学

6.拥塞控制算法
7.服务质量

谢晓燕

西安邮电大学

8.三层交换和VPN

邹莹

仲恺农业工程学院

4.Internet路由协议

李旭宏

枣庄学院

1.网络层服务
2.Internet网际协议
4.Internet路由协议

《计算机网络：自顶向下方法》(原书第7版)，库罗斯 罗斯，机械工业出版社，2018年06月
《计算机网络（第5版）》，Tanenbaum & Wetherall，清华大学出版社，2012年3月
《计算机网络（第7版）》，谢希仁，电子工业出版社，2017年01月
《计算机网络教程（第6版）》，吴功宜，电子工业出版社，2018年03月
《计算机网络（第3版）》，徐敬东、张建忠，清华大学出版社，2013年6月1日

特别致谢：
部分内容取材于此

致谢社区本章贡献者

清华大学 Tsinghua University
计算机网络教案社区

曲大鹏

辽宁大学

3.路由算法

方诗虹

西南民族大学

1.网络层服务
3.路由算法

舒挺

浙江理工大学

1.网络层服务

白云莉

内蒙古农业大学

3.路由算法

余琨

荆楚理工学院

2.Internet网际协议

李振斌

华为技术公司

9.IPv6技术

《计算机网络：自顶向下方法》(原书第7版)，库罗斯 罗斯，机械工业出版社，2018年06月
《计算机网络（第5版）》，Tanenbaum & Wetherall，清华大学出版社，2012年3月
《计算机网络（第7版）》，谢希仁，电子工业出版社，2017年01月
《计算机网络教程（第6版）》，吴功宜，电子工业出版社，2018年03月
《计算机网络（第3版）》，徐敬东、张建忠，清华大学出版社，2013年6月1日

特别致谢：
部分内容取材于此