
哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY
立足航天，服务国防，面向国民经济主战场
《计算机网络》
第4章 网络层
主讲人：刘亚维

拥塞控制

拥塞:

- 非正式地：“太多的源发送太多的数据，速度太快，网络无法处理”
- 事件：
  - 长时间延迟（在路由器缓冲区中排队）
  - 数据包丢失（路由器上的缓冲区溢出）
- 与流量控制不同！
- top-10问题!

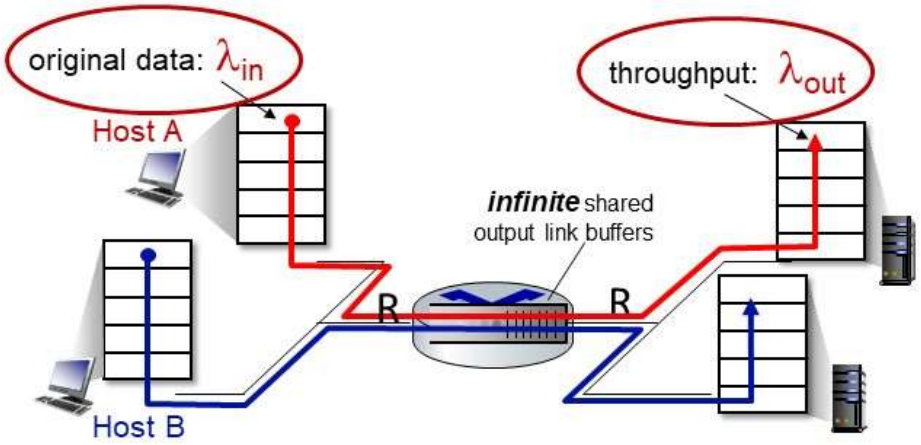
congestion control:
too many senders,
sending too fast

flow control: one sender
too fast for one receiver

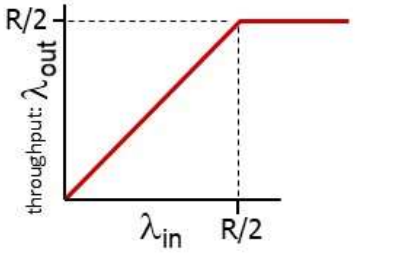
拥堵的原因/代价:场景1

最简单场景:

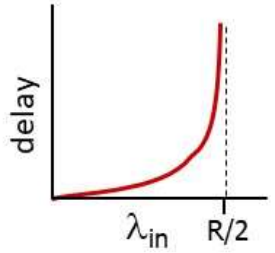
- **一台路由器，无限缓冲区**
- **输入、输出链路容量：R**
- **两个数据流**
- **无需重新传输**



Q: 问：当到达率 λ_{in} 接近 R/2 时会发生什么?



maximum per-connection throughput: R/2

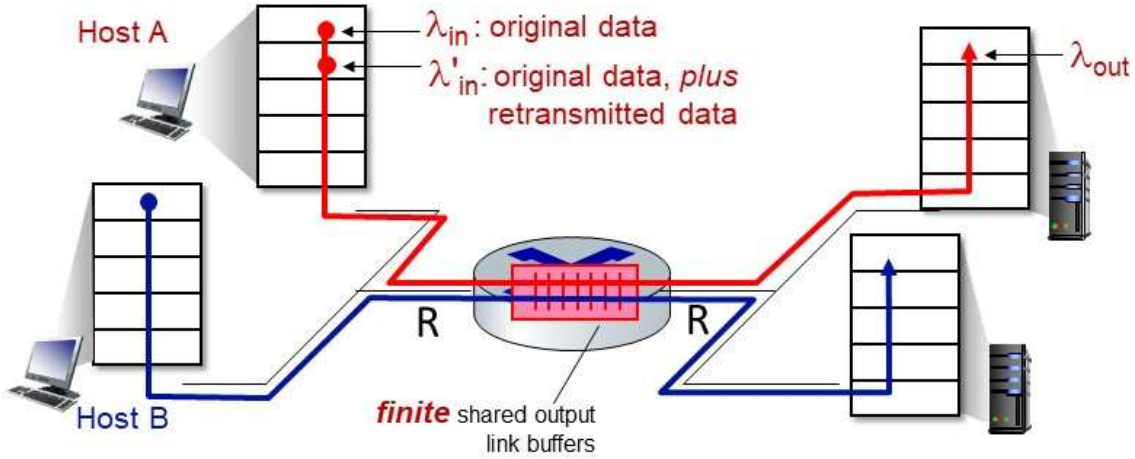


large delays as arrival rate λ_{in} approaches capacity

Transport Layer: 3-4

拥堵的原因/代价:场景2

- 一个路由器，有限缓冲区
- 发送方重新传输丢失、超时的分组
  - 应用层输入 = 应用层输出: $\lambda_{in} = \lambda_{out}$
  - 传输层输入包括重传分组: $\lambda'_{in} \geq \lambda_{in}$

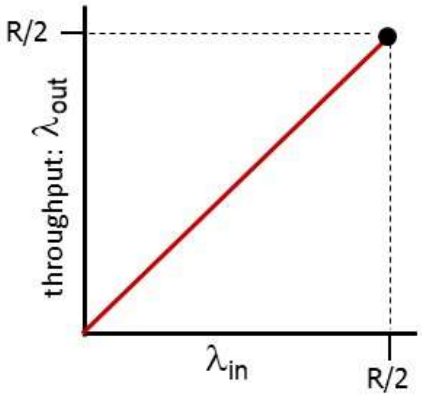


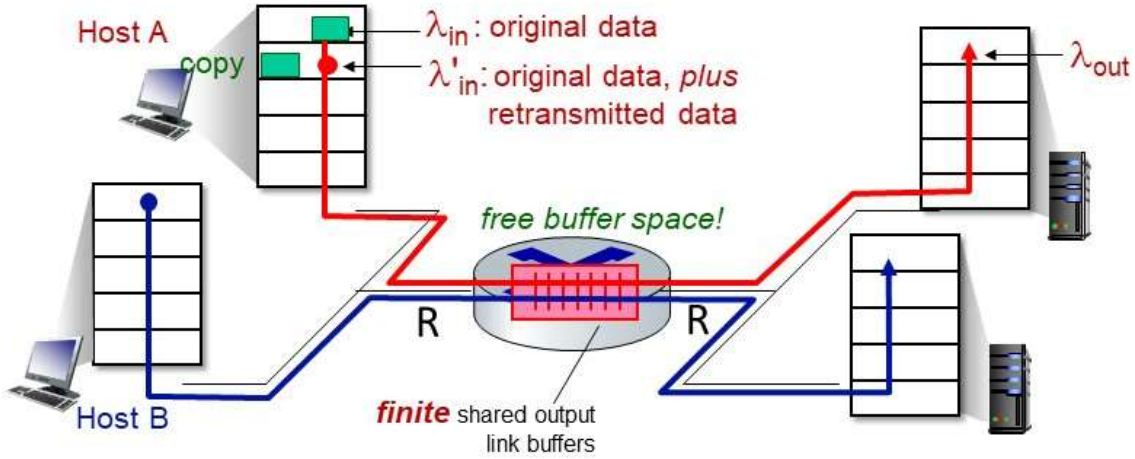
Transport Layer: 3-6

拥堵的原因/代价:场景2

理想化:完美的知识

- 仅在路由器缓冲区可用时，发送方发送数据



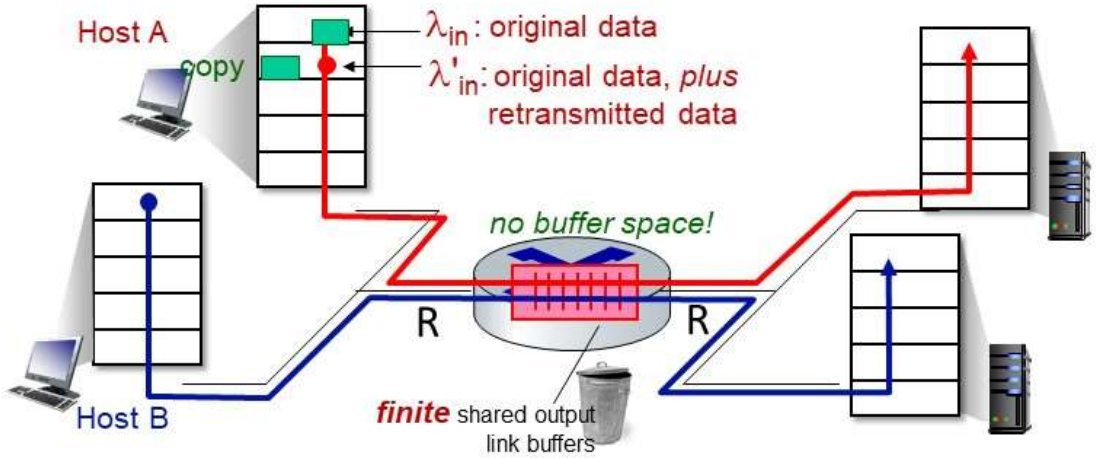


Transport Layer: 3-8

拥堵的原因/代价:场景2

理想化:部分完美的知识

- 由于缓冲区已满，数据包可能会丢失（在路由器上丢弃）
- 发送方知道数据包何时被丢弃：仅在已知数据包丢失时重新发送

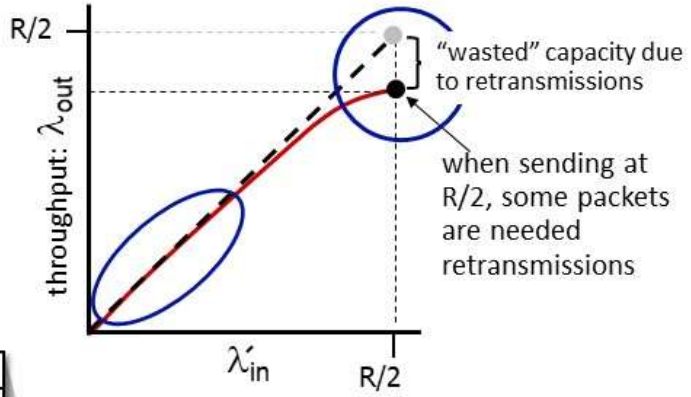


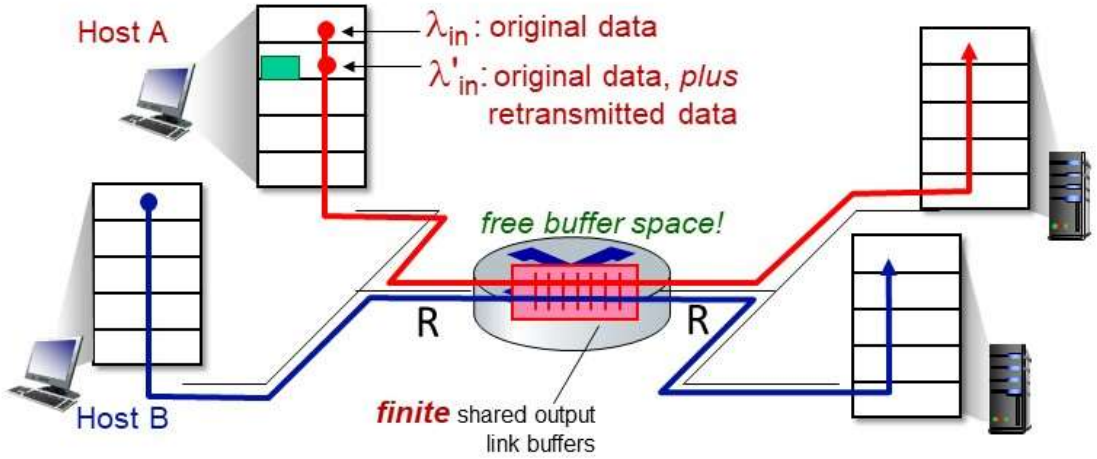
Transport Layer: 3-10

拥堵的原因/代价:场景2

理想化:部分完美的知识

- 由于缓冲区已满，数据包可能会丢失（在路由器上丢弃）
- 发送方知道数据包何时被丢弃：仅在已知数据包丢失时重新发送



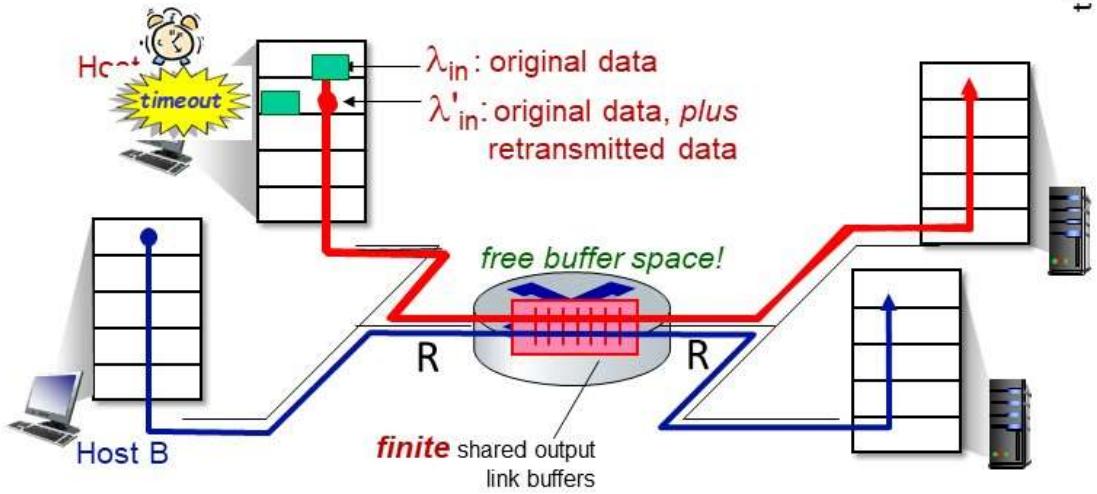


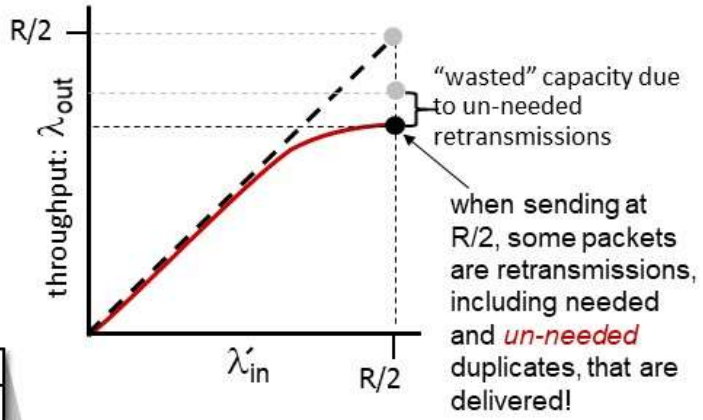
Transport Layer: 3-12

拥堵的原因/代价:场景2

真实场景:不必要的重传

- 由于缓冲区已满，数据包可能会丢失，在路由器上丢弃 - 需要重新传输
- 但是发件方定时器可能会过早超时，发送两份副本，而且两份副本都被送达



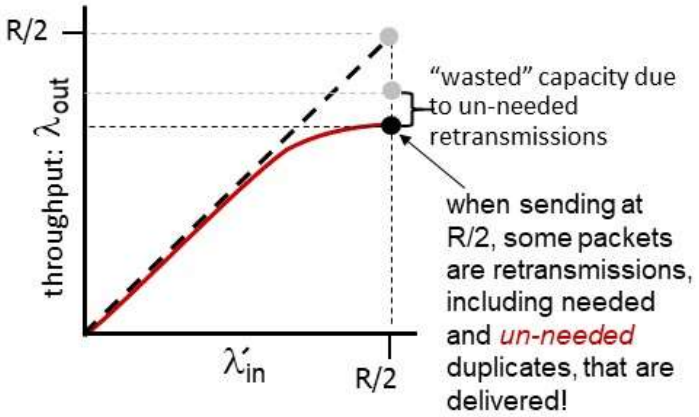


Transport Layer: 3-14

拥堵的原因/代价:场景2

真实场景:不必要的重传

- 由于缓冲区已满，数据包可能会丢失，在路由器上丢弃 - 需要重新传输
- 但是发件方定时器可能会过早超时，发送两份副本，而且两份副本都被送达



拥塞的“成本”：

- 要达到接收方希望的吞吐量，要做更多的工作（重传）
- 不必要的重传:链路携带数据包的多个副本
  - 降低可达到的最大吞吐量

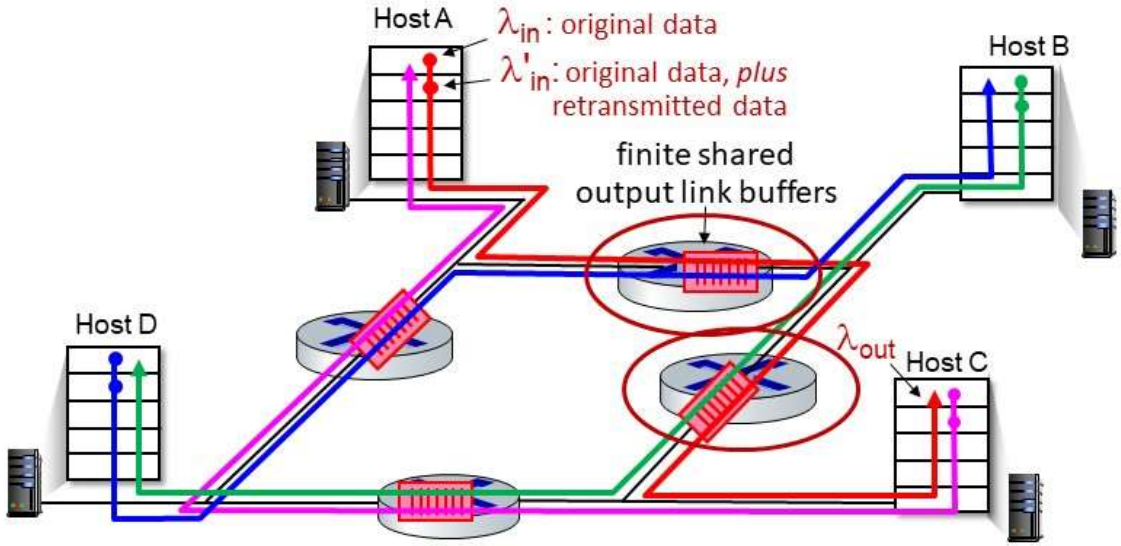
Transport Layer: 3-16

拥堵的原因/成本:场景3

- 四个发送方
- 多跳路径
- 超时/重新传输

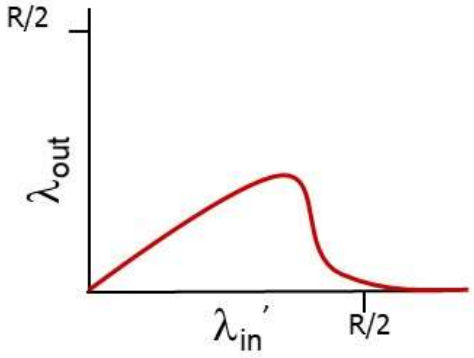
Q: 随着 λ_{in} 和 λ_{in}' 增加，会发生什么?

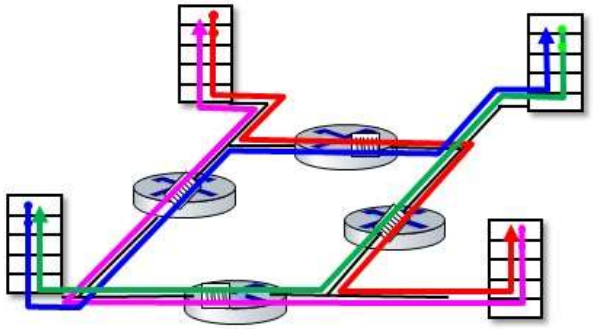
A:随着红色λ_{in}' 增加，所有到达上游队列的蓝色分组都会被丢弃，blue throughput → 0



Transport Layer: 3-18

拥堵的原因/代价:场景3





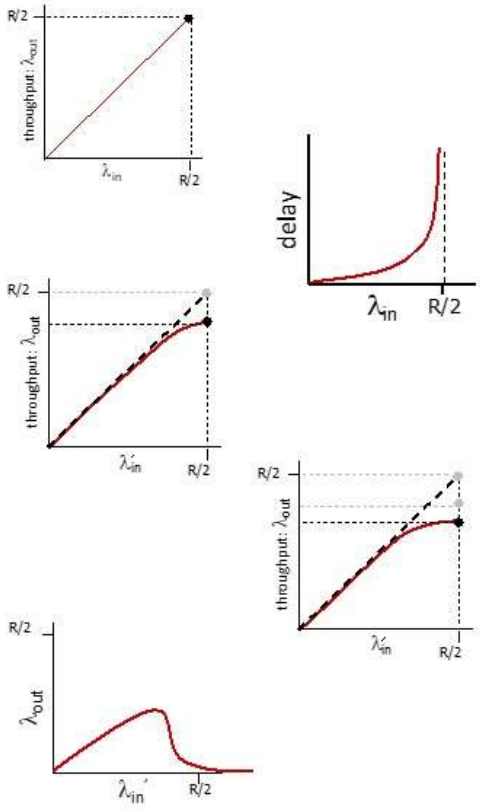
拥堵的另一个“代价”：

- **当分组丢弃时，用于该分组的任何上游传输容量和缓冲都被浪费了！**

Transport Layer: 3-20

拥堵的原因/代价: 总结

- 吞吐量永远不能超过容量
- 当接近容量，延迟也会增加
- 丢失/重传会降低有效吞吐量
- 不必要的重传会进一步降低有效吞吐量
- 上游的传输容量/缓冲因下游分组丢失而浪费

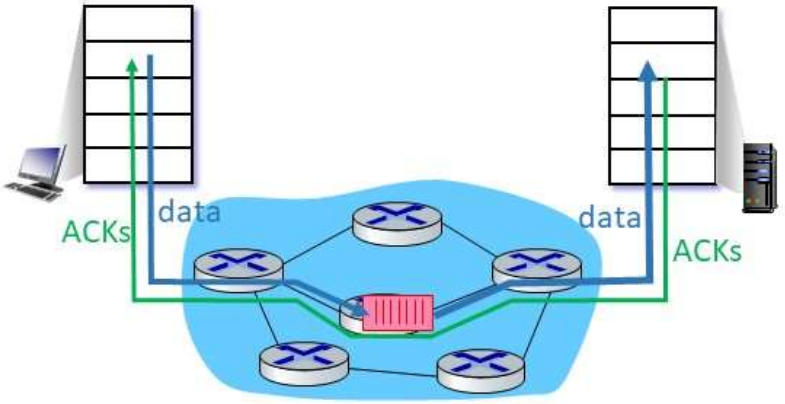


Transport Layer: 3-22

拥塞控制的方法

端到端拥塞控制：

- 没有来自网络的明确反馈
- 从观察到的损失、延迟**推断**出拥塞
- TCP采取的办法

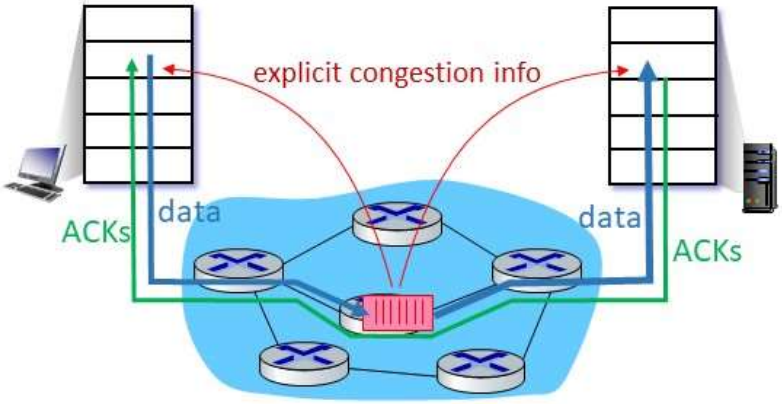


Transport Layer: 3-23

拥塞控制的方法

网络辅助拥塞控制：

- 当主机的流量通过拥塞的路由器时，路由器向发送/接收主机提供**直接**反馈
- 可以指示**拥塞级别**或明确**设置发送速率**



- TCP ECN, ATM, DECbit protocols

Transport Layer: 3-24

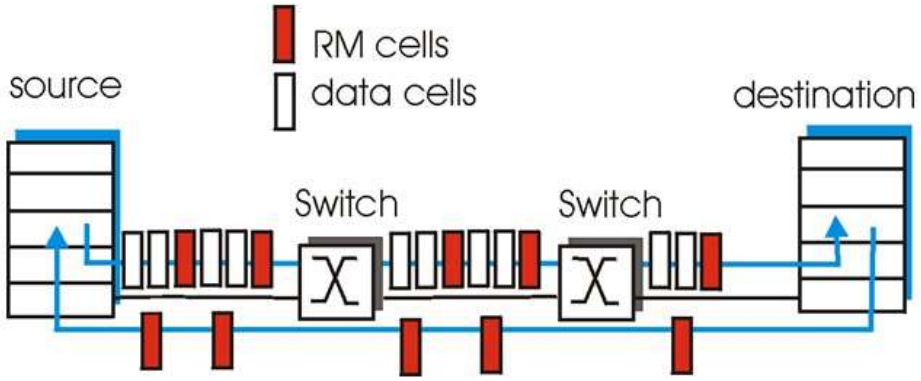
网络层拥塞控制策略

- 流量感知路由
- 准入控制
- 流量调节
  - 抑制分组
  - 背压
- 负载脱落

案例：ATM ABR拥塞控制

- ABR：available bit rate
  - “弹性服务”
  - 如果发送方路径“underloaded”
    - 使用可用带宽
  - 如果发送方路径拥塞
    - 将发送速率降到最低保障速率
- RM(resource management) cells
  - 发送方发送
  - 交换机设置RM cell位(网络辅助)
    - NI bit: 速率不许增长
    - CI bit: 拥塞指示
  - RM cell由接收方返回给发送方

案例： ATM ABR拥塞控制



- 在RM cell中有显式的速率(ER)字段：两个字节
  - 拥塞的交换机可以将ER置为更低的值
  - 发送方获知路径所能支持的最小速率
- 数据cell中的EFCI位: 拥塞的交换机将其设为1
  - 如果RM cell前面的data cell的EFCI位被设为1，那么发送方在返回的RM cell中置CI位

主要内容

本章学习目标

- 理解网络层服务
- 理解虚电路网络与数据报网络
- 掌握路由器体系结构
- 掌握IP协议
  - IP数据报
  - IP地址与子网划分
  - CIDR与路由聚合
- 掌握DHCP、NAT、ICMP、ARP等协议
- 掌握典型路由算法和路由协议

主要内容

- 4.1 网络层服务
- 4.2 虚电路网络与数据报网络
- 4.3 路由器体系结构
- 4.4 IP协议
- 4.5 IP相关协议
- 4.6 路由算法
- 4.7路由协议

35

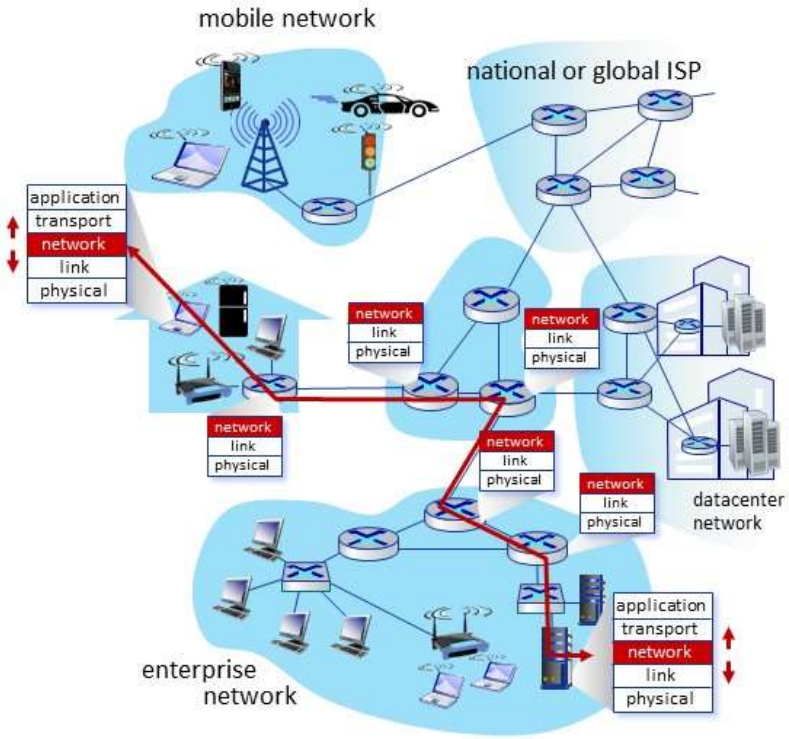

哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY
立足航天，服务国防，面向国民经济主战场

4.1 网络层服务

刘亚维

网络层

- 从发送主机向接收主机传送数据段（segment）
- **发送主机**：
  - 将数据段封装到数据报（datagram）中
- **接收主机**：
  - 向传输层交付数据段（segment）
- **每个主机和路由器都运行网络层协议**
- **路由器**
  - 检验所有通过它的IP数据报的头部域
  - 将数据报从输入端口移动到输出端口，以沿着端到端的路径传输数据报



网络层核心功能-转发与路由

- **转发(forwarding)**：将分组从路由器的输入端口转移到合适的输出端口
- **路由(routing)**：确定分组从源到目的经过的路径
  - 路由算法 (routing algorithms)

类比:旅行

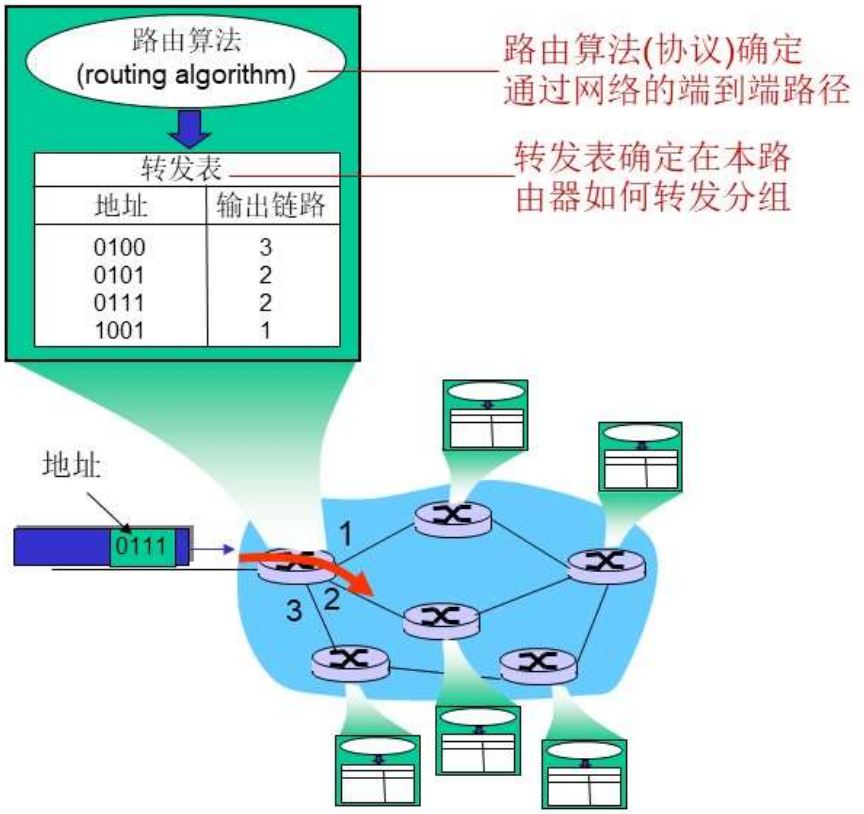
- **转发**：通过单一立交桥的过程
- **路由**：从出发地到目的地的行程规划过程

forwarding

routing

网络层核心功能-转发与路由

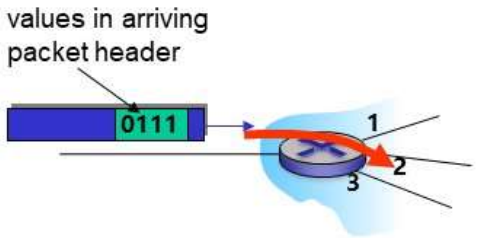
- **转发(forwarding)**：将分组从路由器的输入端口转移到合适的输出端口
- **路由(routing)**：确定分组从源到目的经过的路径
  - **路由算法**
    (routing algorithms)



网络层：数据平面、控制平面

数据平面:

- *本地*, per-router功能
- 确定如何将到达路由器输入端口的数据报**转发**到路由器输出端口



控制平面

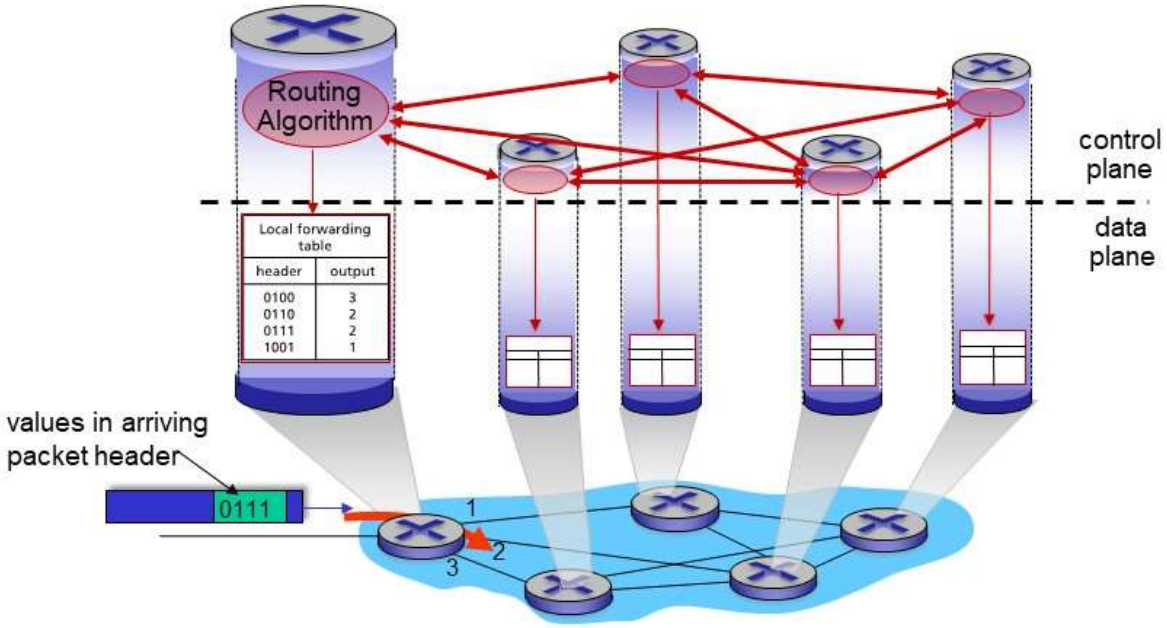
- *network-wide*逻辑
- 确定如何将数据报，沿着从源主机到目标主机间的**端到端路径**，在路由器之间**路由**

- 两种控制平面方法:
  - **传统路由算法**:在路由器中实现
  - **software-defined networking (SDN)**:在（远程）服务器中实现

Network Layer: 4-43

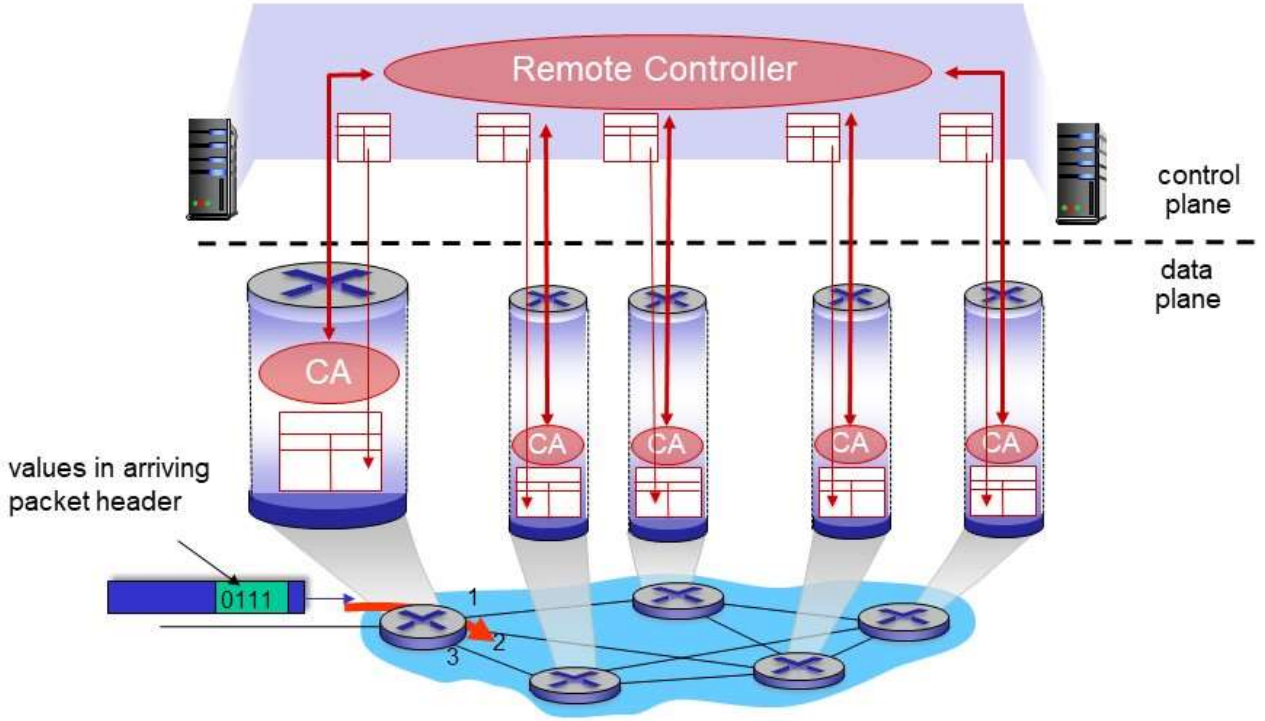
Per-router控制平面

位于每一个路由器中的独立路由算法组件，在控制平面中进行交互



Software-Defined Networking (SDN)控制平面

转发表由远程控制器进行计算，并在路由器中安装



Network Layer: 4-47

网络层服务模型

Q: 网络层为发送端（主机）到接收端（主机）的数据报传送“通道(channel)”提供什么样的服务模型(service model)?

网络架构	服务模型	是否保证？ 带宽	丢失	顺序	延迟	拥塞反馈
Internet	best effort	none	no	no	no	no (inferred via loss)
ATM	CBR（恒定速率）	constant rate	yes	yes	yes	no congestion
ATM	VBR（可变速率）	guaranteed rate	yes	yes	yes	no congestion
ATM	ABR（可用比特率）	guaranteed minimum	no	yes	no	yes
ATM	UBR（不指明比特率）	none	no	yes	no	no

52

关于best-effort service的思考

- **机制简单**，使得因特网得以广泛部署
- **配以充足的带宽**，在“绝大部分时间”内，都可以“满足” 实时应用程序（例如交互式语音、视频）的性能需求。
- **基于复制的、应用层的分布式服务**（数据中心、内容分发网络CDN），通过部署到客户网络附近，可以从多个位置提供服务
- 有助于实现“弹性”的拥塞控制服务

很难反驳best-effort服务模式的成功

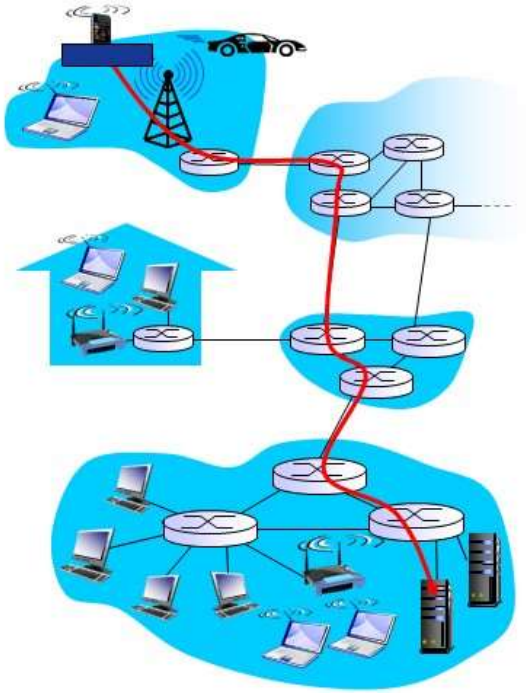
Network Layer: 4-53

网络层服务模型

- **无连接服务**(connection-less service):
  - 不事先为系列分组的传输确定传输路径
  - 每个分组独立确定传输路径
  - 不同分组可能传输路径不同
  - **数据报网络**(datagram network)
- **连接服务**(connection service):
  - 首先为系列分组的传输确定从源到目的经过的路径(建立连接)
  - 然后沿该路径（连接）传输系列分组
  - 系列分组传输路径相同
  - 传输结束后拆除连接
  - **虚电路网络**(virtual-circuit network)

网络层核心功能-连接建立

- 某些网络的重要功能:
  - ATM, 帧中继, X.25
- 数据分组传输之前两端主机需要首先建立虚拟/逻辑连接
  - 网络设备（如路由器）参与连接的建立
- 网络层连接与传输层连接的对比:
  - 网络层连接: 两个主机之间 (路径上的路由器等网络设备参与其中)
  - 传输层连接: 两个应用进程之间（对中间网络设备透明）



4.2 虚电路网络与数据报网络

刘亚维

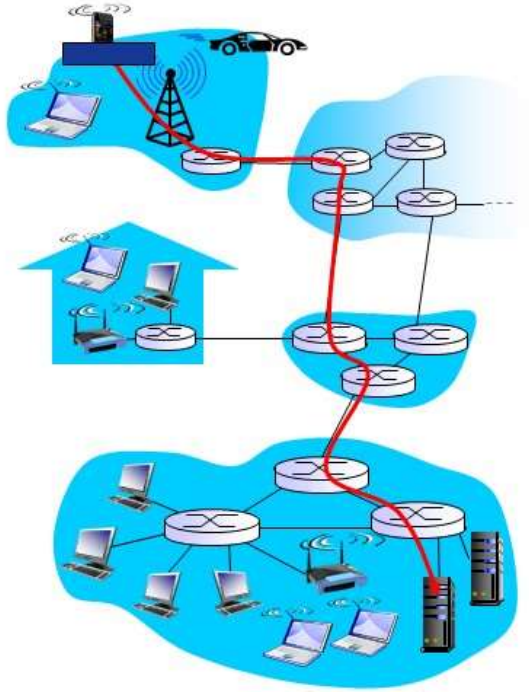
连接服务与无连接服务

- 数据报(datagram)网络与虚电路(virtual-circuit)网络是典型两类分组交换网络
- 数据报网络提供网络层无连接服务
- 虚电路网络提供网络层连接服务
- 类似于传输层的无连接服务（UDP）和面向连接服务（TCP），但是网络层服务：
  - 主机到主机服务
  - 网络核心实现

虚电路(Virtual circuits)

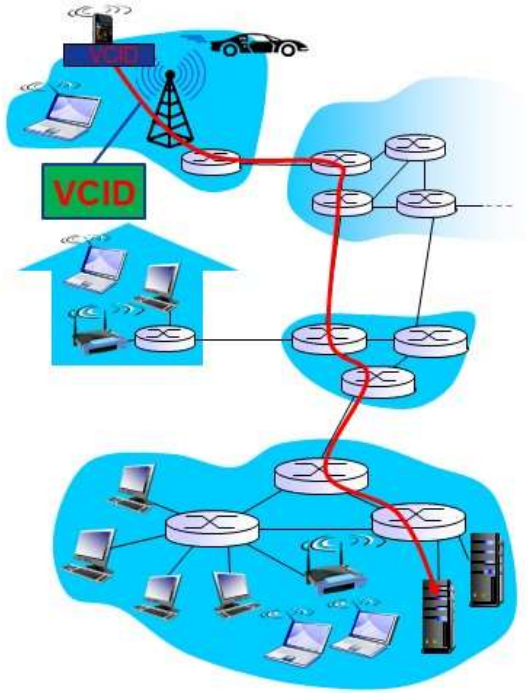
虚电路：一条从源主机到目的主机，类似于电路的路径(逻辑连接)

- 分组交换
- 每个分组的传输利用链路的全部带宽
- 源到目的路径经过的网络层设备共同完成虚电路功能



虚电路(Virtual circuits)

- 通信过程：
  - 呼叫建立(call setup)→数据传输→拆除呼叫
- 每个分组携带虚电路标识(VC ID)，而不是目的主机地址
- 虚电路经过的每个网络设备(如路由器)，维护每条经过它的虚电路连接状态
- 链路、网络设备资源(如带宽、缓存等)可以面向VC进行预分配
  - 预分配资源=可预期服务性能
  - 如ATM的电路仿真（CBR）

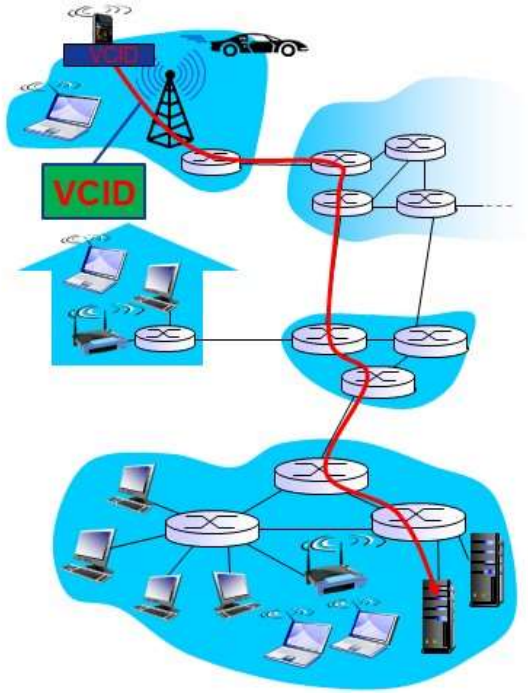


VC的具体实现

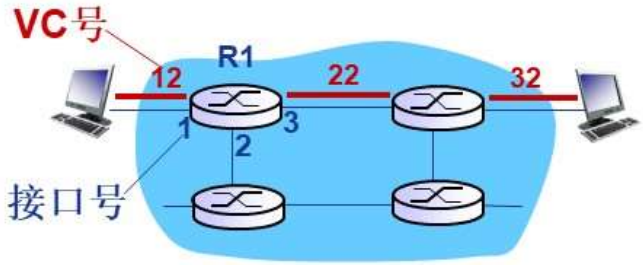
每条虚电路包括:

1. 从源主机到目的主机的一条路径
2. 虚电路号（VCID），沿路每段链路一个编号
3. 沿路每个网络层设备（如路由器），利用转发表记录经过的每条虚电路

- 沿某条虚电路传输的分组，携带对应虚电路的VCID，而不是目的地址
- 同一条VC，在每段链路上的VCID通常不同
  - 路由器转发分组时依据转发表改写/替换虚电路号



VC转发表



路由器R1的VC转发表:

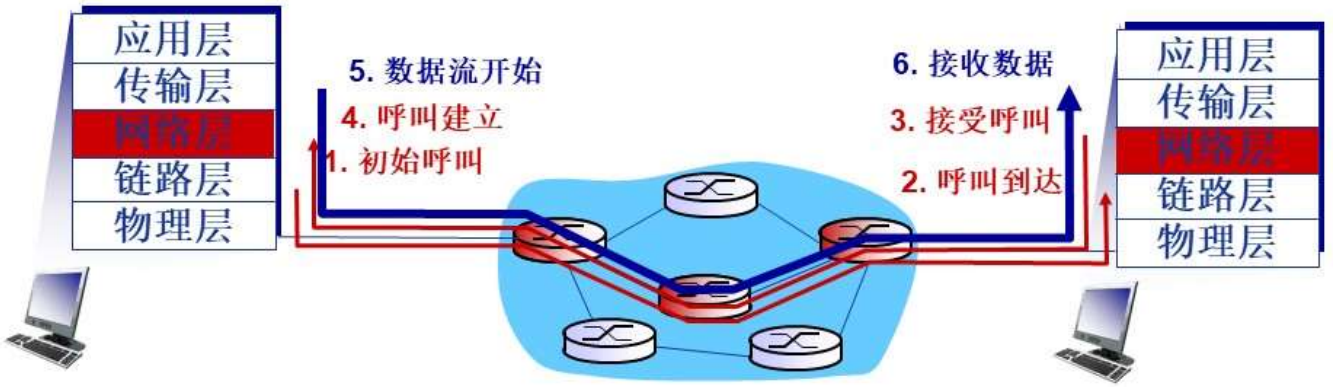
输入接口	输入VC #	输出接口	输出VC #
1	12	3	22
2	63	1	18
3	7	2	17
1	97	3	87
…	…	…	…

VC路径上每个路由器都需要维护VC连接的状态信息！

61

虚电路信令协议(signaling protocols)

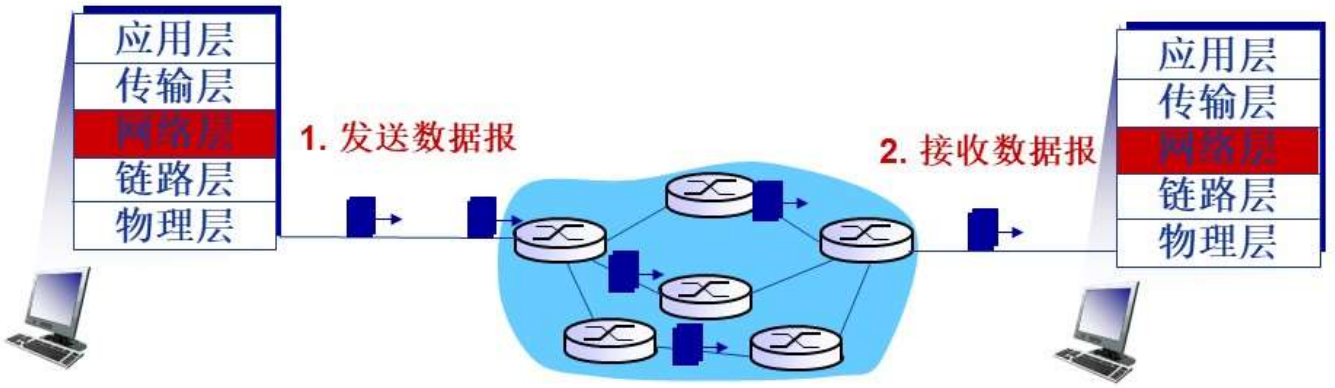
- 用于VC的建立、维护与拆除
  - 路径选择
- 应用于虚电路网络
  - 如ATM、帧中继(frame-relay)网络等
- 目前的Internet不采用



62

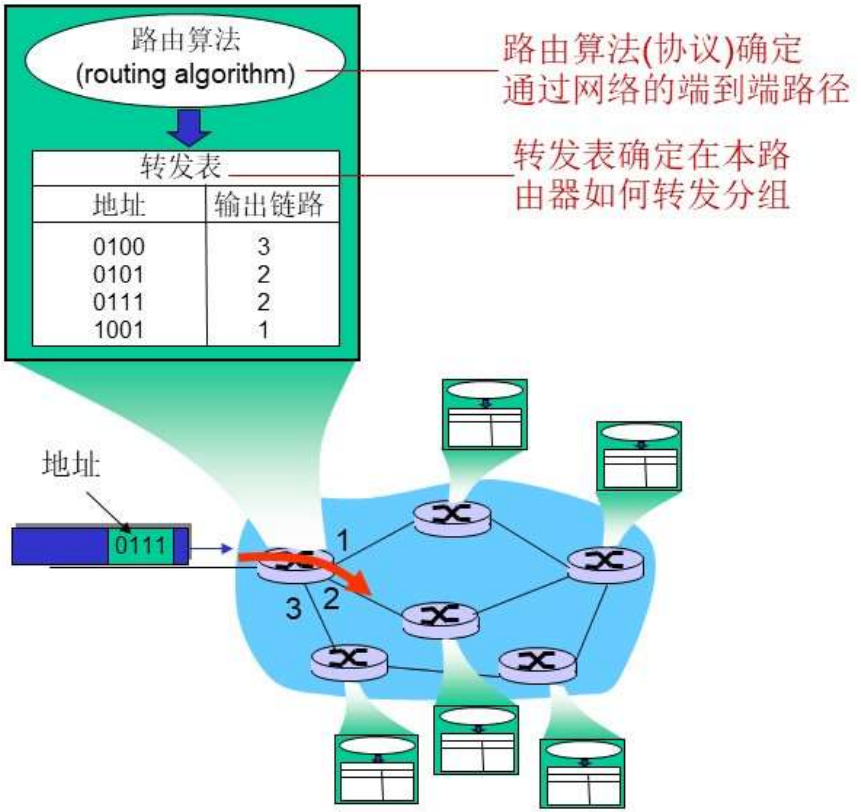
数据报网络

- 网络层无连接
- 每个分组携带目的地址
- 路由器根据分组的目的地址转发分组
  - 基于路由协议/算法构建转发表
  - 检索转发表
  - 每个分组独立选路

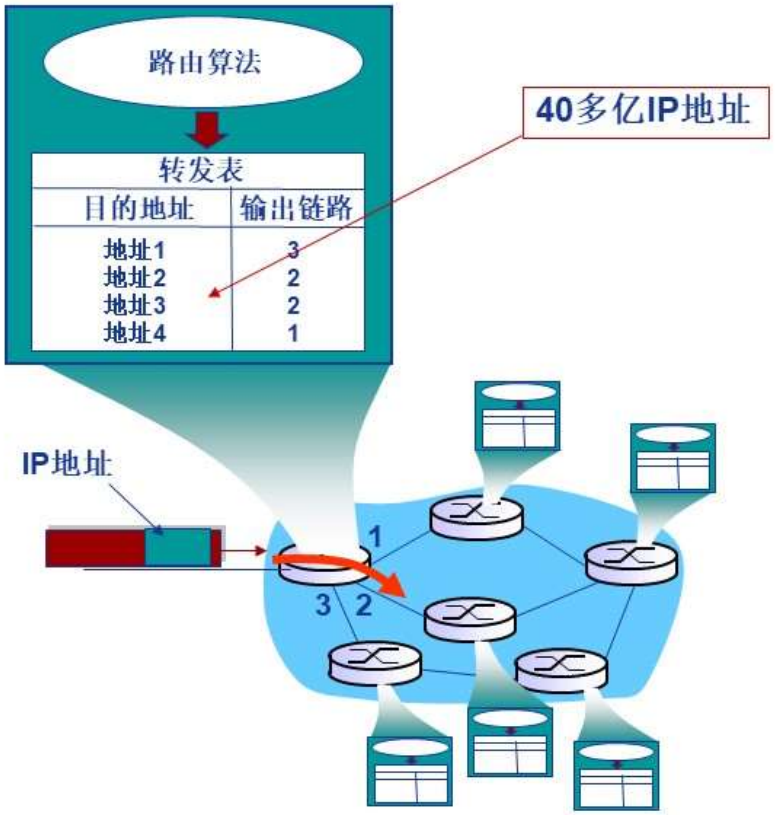


63

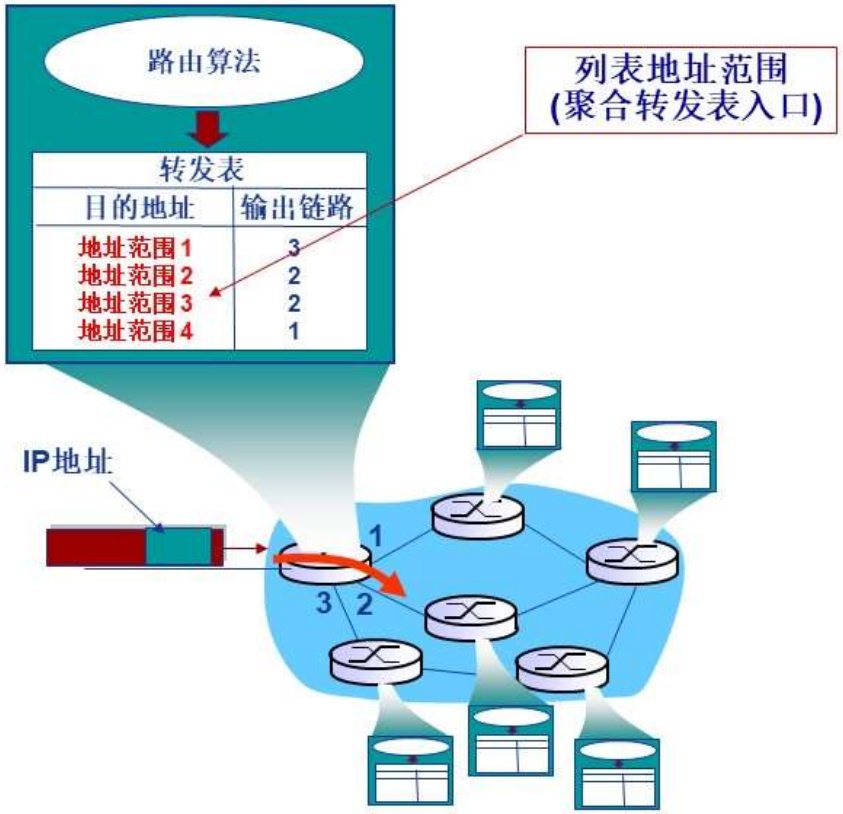
数据报转发表


路由算法
(routing algorithm)
转发表
地址
输出链路
0100 3
0101 2
0111 2
1001 1
路由算法(协议)确定
通过网络的端到端路径
转发表确定在本路
由器如何转发分组
地址
0111
1
2
3

数据报转发表



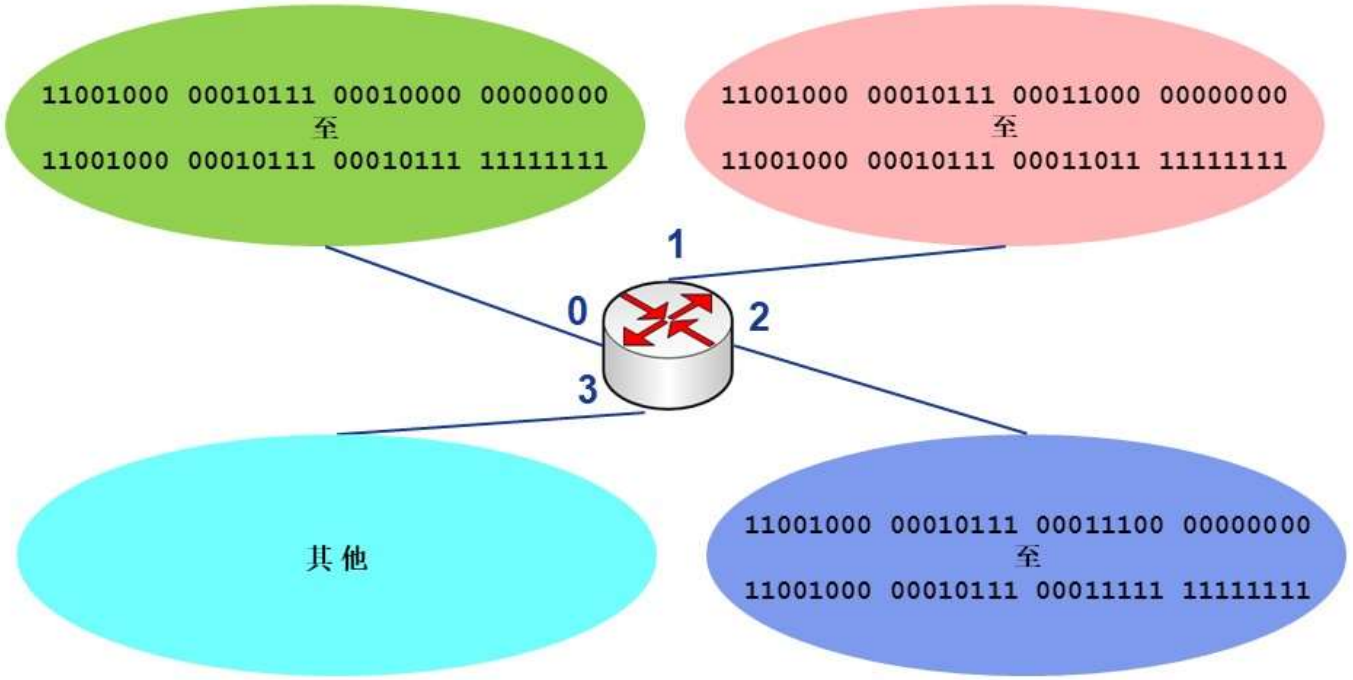
数据报转发表


路由算法
转发表
目的地址
输出链路
地址范围1
地址范围2
地址范围3
地址范围4
3
2
2
1
列表地址范围
(聚合转发表入口)
IP地址
1
3
2

数据报转发表

目的地址范围	链路接口
11001000 00010111 00010000 00000000 至 11001000 00010111 00010111 11111111	0
11001000 00010111 00011000 00000000 至 11001000 00010111 00011011 11111111	1
11001000 00010111 00011100 00000000 至 11001000 00010111 00011111 11111111	2
其他	3

数据报转发表



Q: 如果地址范围划分的不是这么“完美”会怎么样？

最长前缀匹配优先

例如：

目的地址范围	链路接口
11001000 00010111 00010*** *********	0
11001000 00010111 00011000 *********	1
11001000 00010111 00011*** *********	2
其他	3

DA: 11001000 00010111 00010110 10100001　从哪个接口转发？ A:0

DA: 11001000 00010111 00011000 10101010　从哪个接口转发？ A:1

最长前缀匹配优先

在检索转发表时，优先选择与分组目的地址匹配前缀最长的入口（entry）。

数据报网络 or VC网络?

Internet (数据报网络)

- 计算机之间的数据交换
  - “弹性”服务，没有严格时间需求
- 链路类型众多
  - 特点、性能各异
  - 统一服务困难
- “智能”端系统 (计算机)
  - 可以自适应、性能控制、差错恢复
- ***简化网络，复杂“边缘”***

ATM (VC网络)

- 电话网络演化而来
- 核心业务是实时对话：
  - 严格的时间、可靠性需求
  - 需要有保障的服务
- “哑(dumb)”端系统(非智能)
  - 电话机
  - 传真机
- ***简化“边缘”，复杂网络***