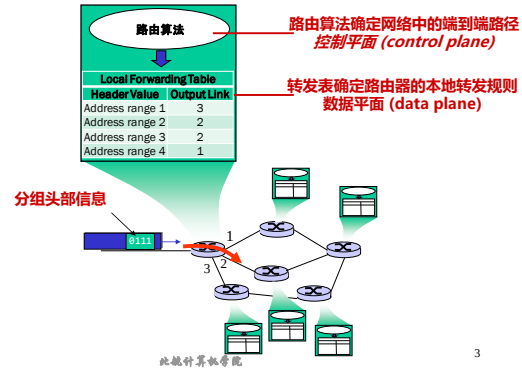
主要内容

- Internet的路由技术
  - 路由算法概述
    - 基本路由算法（基础知识回顾）
    - 广播，选播
    - 多播，移动主机路由……（可选）
  - BGP路由协议（预习）

路由Routing 和 转发Forwarding



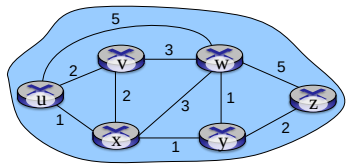
路由 Routing

- 网络层：选择路径
  - 路径：Choosing paths along which messages will travel from source to destination
  - 第三层（网络层）功能
- 其他层的功能 （后续课程介绍）
  - 第二层（数据链路层）：生成树算法（ Ethernet spanning tree protocol ）
  - 高层（传输层以上）：
    - CDN(Content delivery overlays), P2P（distributed hash tables）等
    - 网络虚拟化network virtualization

路由算法

- 数据报网络
  - 确定每个输入分组该被发送到哪条输出线路上
- 虚电路网络
  - 当建立一条新的虚电路时进行路由决策
  - 会话路由：在一次会话过程中（如VPN的一次登录），路径保持有效
- 满足以下特性：
  - 正确性，简单性，鲁棒性，稳定性，公平性，有效性
- 分类
  - 非自适应算法，也称为静态路由（static routing）
  - 自适应算法，也称为动态路由（dynamic routing）

图抽象：Graph abstraction



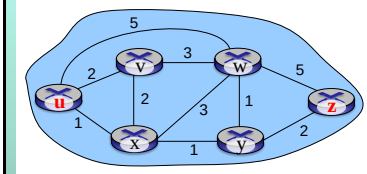
用图表示网络：

Graph: G = (N,E)

N = 节点，路由器集合set of routers = { u, v, w, x, y, z }

E = 边，链路集合 set of links ={ (u,v), (u,x), (v,x), (v,w), (x,w), (x,y), (w,y), (w,z), (y,z) }

Graph abstraction: 开销costs



c(x,x') = cost of link (x,x')

- e.g., c(w,z) = 5

开销cost 通常设为1，也可以设置与延迟、花费、带宽或拥塞参数相关的值（正比或反比）。

Cost of path $(x_1, x_2, x_3, \ldots, x_p) = c(x_1,x_2) + c(x_2,x_3) + \ldots + c(x_{p-1},x_p)$

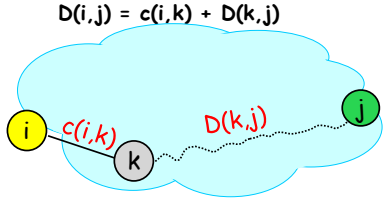
目标：找出 u 和 z 之间的最小开销路径

路由算法： algorithm that finds least-cost path

说明

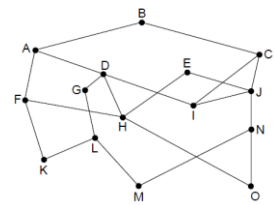
- 两个相邻节点之间的最短路径包含在最短路径子集中
- 说明:
  - 节点 i 到节点 j 的最短路径为D(i,j)，并经过其相邻节点 k, i到k的链路开销为 c(i,k), 那么，

$$D(i,j) = c(i,k) + D(k,j)$$

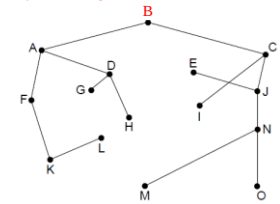


最优化原则

- 最优路径树
  - Each portion of a best path is also a best path; the union of them to a router is a tree called the sink tree（汇集树）
  - 有向无环图 DAG(Directed Acyclic Graph)



Network



Sink tree of best paths to router B

路由协议的考虑

- 谁来决定选择路径?
  - The network or the end host?
- 路径选择的目标的复杂度?
  - Shortest-path vs. policy-based routing
- 路由参与方是否协作?
  - Willing to share information?
  - Have a common goal in selecting paths?
- 考虑大规模行为的影响?
  - Stability of the network topology 拓扑结构稳定性
  - State and message overhead 通信开销、存储开销
  - Disruptions during routing convergence 收敛问题

路由算法分析

- 最短路径算法
- 洪泛算法
- 距离向量算法
- 链路状态路由
- 层次路由
- 广播路由
- 组播路由
- 选播路由
- 移动主机路由
- 自组织网络路由

参考书：Andrew S.Tanenbaum, Computer Networks , 清华大学出版社，（第五版，《计算机网络》）第5章

最短路径算法

- Shortest Path Algorithm
- 路径的度量指标
  - 跳数，物理距离，带宽，延迟，流量，成本开销等
- 计算两个节点之间的最短路径
  - 例：Dijkstra 算法（1959）

问题描述

- 问题描述：给定加权有向图G＝(V, E)和源点v∈V，求从v到G中其余各顶点的最短路径。
  - 应用：怎样找到一种最经济的方式，从一台计算机向网上所有其它计算机发送一条消息？
- 单源点的最短路径问题：给定加权有向图G和源点v，求从v到G中其余各顶点的最短路径。

Dijkstra 算法

◆使用广度优先搜索解决加权有向图或者无向图的单源最短路径问题，算法最终得到一个最短路径树

◆算法思路：贪心策略

符号表示

- 链路开销 c(x,y): link cost from node x to y; = ∞ if not direct neighbors
- 路径开销 D(v) : current value of cost of path from source to dest. v
- 邻居节点p(v): predecessor node along path from source to v
- 已知节点集合N': set of nodes whose least cost path definitively known

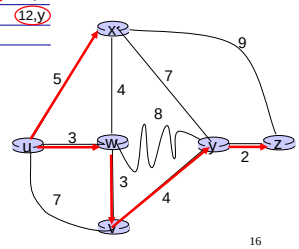
算法参考教材：James F.Kurose，Keith W. Ross，《计算机网络：自顶向下方法》，第六版，第七版，机械工业出版社

Dijkstra 算法说明

```
1  Initialization:  /*计算从源节点u到网络中每个节点的最短距离*/
2    N' = {u}
3    for all nodes v
4      if v adjacent to u      /*相邻节点距离计算*/
5         then D(v) = c(u,v)
6      else D(v) = ∞
7
8  Loop
9    find w not in N' such that D(w) is a minimum
10   add w to N'
11   update D(v) for all v adjacent to w and not in N' :
12      D(v) = min( D(v), D(w) + c(w,v) )
13   /* new cost to v is either old cost to v or known
14    shortest path cost to w plus cost from w to v */
15 until all nodes in N'
```


例 1

Step	N'	D(v) p(v)	D(w) p(w)	D(x) p(x)	D(y) p(y)	D(z) p(z)
0	u	7,u	3,u	5,u	∞	∞
1	uw	6,w		5,u	11,w	∞
2	uwx	6,w			11,w	14,x
3	uwxv				10,v	14,x
4	uwxvy					12,y
5	uwxvyz					

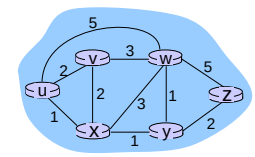


注意:

- **构建最短路径树：by tracing predecessor nodes**
- **存在相同开销路径：ties can exist (can be broken arbitrarily)**

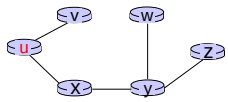
例 2

Step	N'	D(v),p(v)	D(w),p(w)	D(x),p(x)	D(y),p(y)	D(z),p(z)
0	u	2,u	5,u	1,u	∞	∞
1	ux	2,u	4,x		2,x	∞
2	uxy	2,u	3,y			4,y
3	uxyv		3,y			4,y
4	uxyvw					4,y
5	uxyvwz					



生成最短路径树

resulting shortest-path tree from u:



转发表：resulting forwarding table in u:

destination	link
v	(u,v)
x	(u,x)
y	(u,x)
w	(u,x)
z	(u,x)

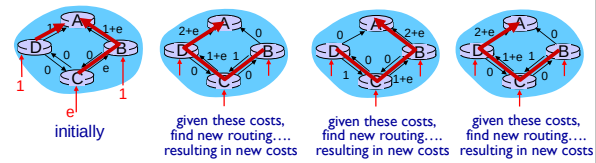
Dijkstra 算法讨论

算法复杂度 Algorithm complexity: *n* nodes

- Each iteration: need to check all nodes, *w*, not in *N*
- *搜索节点总数n*(n+1)/2，时间复杂度: O(n^2)*
- 高效实现：堆， O(*n* log *n*)

可能产生路由震荡Oscillations possible: 负载敏感算法的问题

- e.g., link cost = amount of carried traffic （负载）



initially

given these costs, find new routing…. resulting in new costs

given these costs, find new routing…. resulting in new costs

given these costs, find new routing…. resulting in new costs

路由算法分析

- 最短路径算法
- 洪泛算法
- 距离向量算法
- 链路状态路由
- 层次路由
- 广播路由
- 组播路由
- 选播路由
- 移动主机路由
- 自组织网络路由

洪泛算法

- 洪泛算法
  - Each node floods a new packet received on an incoming link by sending it out all of the other links
- 抑制重复包
  - 跳计数器
  - 路由器跟踪已经被扩散过的包
    - Nodes need to keep track of flooded packets to stop the flood; even using a hop limit can blow up exponentially
  - 例：无线路由算法
- 用途： 广播；鲁棒性；基准

路由算法分析

- 最短路径算法
- 洪泛算法
- 距离向量算法
- 链路状态路由
- 层次路由
- 广播路由
- 组播路由
- 选播路由
- 移动主机路由
- 自组织网络路由

距离向量算法

- 距离向量算法
  - 分布式路由算法
  - 例：Bellman-Ford路由算法（1957，1962）
  - ARPANET最早使用的路由算法，RIP协议
- 方法
  - 每个路由器维护一张路由表
    - 到目标路由器的相邻节点和距离度量值
  - 路由更新过程
    - 相邻节点交换距离向量
    - 在计算过程中不使用旧路由表

距离向量算法

- *Bellman-Ford equation (dynamic programming)*

let

$d_x(y)$:= cost of least-cost path from x to y

then

$$d_x(y) = \min_v \{c(x,v) + d_v(y)\}$$

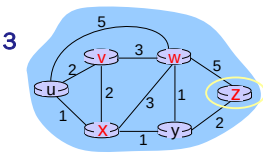
$c(x,v)$: cost to neighbor v

$d_v(y)$: cost from neighbor v to destination y

min taken over all neighbors v of x

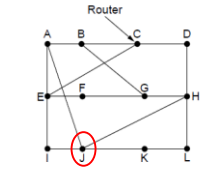
例子

- $d_v(z) = 5$, $d_x(z) = 3$, $d_w(z) = 3$



$$d_u(z) = \min \{ c(u,v) + d_v(z), \; c(u,x) + d_x(z), \; c(u,w) + d_w(z) \}$$
$$= \min \{2 + 5, \; 1 + 3, \; 5 + 3\} = 4$$

例子



问题：计算J→F的路由

To	A	I	H	K	New estimated delay from J	Line
A	0	24	20	21	8	A
B	12	36	31	28	20	A
C	25	18	19	36	28	I
D	40	27	8	24	20	H
E	14	7	30	22	17	I
F	23	20	19	40	30	I
G	18	31	6	31	18	H
H	17	20	0	19	12	H
I	21	0	14	22	10	I
J	9	11	7	10	0	–
K	24	22	22	0	6	K
L	29	33	9	9	15	K
	JA delay is 8	JI delay is 10	JH delay is 12	JK delay is 6	New routing table for J	

Vectors received from J's four neighbors

无穷计算问题

- 路由收敛（convergence）：查找最佳路径的过程
  - 好消息：例如节点A连通
  - 坏消息：例如节点A链路故障

A	B	C	D	E	
•	•	•	•	•	Initially
1	•	•	•	•	After 1 exchange
1	2	•	•	•	After 2 exchanges
1	2	3	•	•	After 3 exchanges
1	2	3	4	•	After 4 exchanges

好消息：Good news of a path to *A* spreads quickly

A	B	C	D	E	
•	1	2	3	4	Initially
X	3	2	3	4	After 1 exchange
	3	4	3	4	After 2 exchanges
	5	4	5	4	After 3 exchanges
	5	6	5	6	After 4 exchanges
	7	6	7	6	After 5 exchanges
	7	8	7	8	After 6 exchanges
	⋮				
	•	•	•	•	

坏消息：Bad news of no path to *A* is learned slowly

距离向量算法

迭代，异步

Iterative，asynchronous:

each local iteration caused by:

- local link cost change
- DV update message from neighbor

***distributed:* 分布式**

- each node notifies neighbors *only* when its DV changes
  - neighbors then notify their neighbors if necessary

每个节点计算：

wait for (change in local link cost or msg from neighbor)

↓

recompute estimates

↓

if DV to any dest has changed, *notify* neighbors

路由算法分析

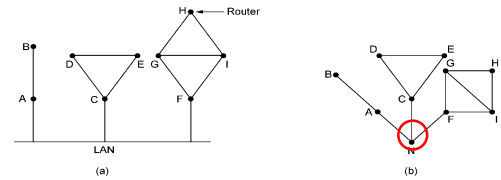
- 最短路径算法
- 洪泛算法
- 距离向量算法
- 链路状态路由
- 层次路由
- 广播路由
- 组播路由
- 选播路由
- 移动主机路由
- 自组织网络路由

链路状态路由算法

- 1979年前：距离向量路由算法，慢收敛问题
  - IS-IS, OSPF
  - 应用于大型网络或Internet
- 算法
  - Each node 洪泛floods information about its neighbors in LSPs (链路状态分组Link State Packets);
  - all nodes learn the full network graph（全局网络拓扑结构）
  - Each node runs Dijkstra's algorithm to compute the path to take for each destination

发现邻居节点

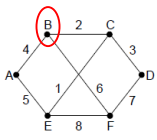
- 路由器启动时，向邻居节点发送HELLO数据包
- 返回应答ACK
- 局域网的情况（通过广播链路连接）
  - 指定路由器DR（designated router）



(a) Nine routers and a broadcast LAN. (b) A graph model of (a).

LSP链路状态分组

- LSP (Link State Packet)：邻居节点，链路权重，序号，年龄字段



网络

Link State Packets					
A	B	C	D	E	F
Seq.	Seq.	Seq.	Seq.	Seq.	Seq.
Age	Age	Age	Age	Age	Age
B 4	A 4	B 2	C 3	A 5	B 6
E 5	C 2	D 3	F 7	C 1	D 7
	F 6	E 1		F 8	E 8

每个节点的LSP

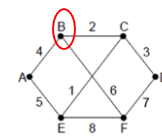
问题：如何分发链路状态分组，使得所有路由器能够快速可靠地获得全部链路状态数据包？

可靠洪泛Reliable Flooding

- 两个字段：序号seq. 和 年龄age
  - 新数据包，序号递增；每秒钟年龄减1
  - 新的LSPs首先放到缓冲区等待，丢弃重复/旧的分组；

  发送分组后，还要根据确认标志发送确认

例子：路由器B的LSP缓冲区



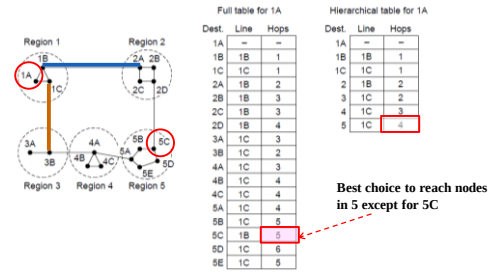
Source	Seq.	Age	Send flags A	Send flags C	Send flags F	ACK flags A	ACK flags C	ACK flags F	Data
A	21	60	0	1	1	1	0	0	
F	21	60	1	1	0	0	0	1	
E	21	59	0	1	0	1	0	1	
C	20	60	1	0	1	0	1	0	
D	21	59	1	0	0	0	1	1	

路由算法分析

- ◆最短路径算法
- ◆洪泛算法
- ◆距离向量算法
- ◆链路状态路由
- ◆层次路由
- ◆广播路由
- ◆组播路由
- ◆选播路由
- ◆移动主机路由
- ◆自组织网络路由

层次路由

- ◆层次路由Hierarchical routing
  - ❖ Hierarchical routing reduces the work of route computation but may result in slightly longer paths than flat routing



分层的数量

- ◆ 当单个网络非常大时，应该分多少层？
- ◆ 例如：720个路由器的子网
  - ❖ 不分层：每个路由器需要720个表项
  - ❖ 分24个区域，30个路由器/区域
    - ➢ 每个路由器的表项数：30+23=53项
  - ❖ 三层结构：8个簇，9个区域/簇，10个路由器/区域
    - ➢ 每个路由器的表项数：7+8+10=25项
- ◆【Kamoun，Kleinrock，1979】对于一个包含N个路由器的网络，最优层数 lnN；每个路由器所需的表项 $elnN$ 个

路由算法分析

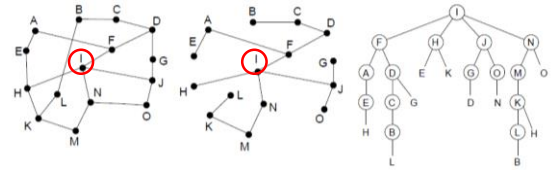
- ◆最短路径算法
- ◆洪泛算法
- ◆距离向量算法
- ◆链路状态路由
- ◆层次路由
- ◆广播路由
- ◆组播路由
- ◆选播路由
- ◆移动主机路由
- ◆自组织网络路由

广播路由算法

- 广播Broadcast的几种方法
  - 源节点给每个目标节点分别发送数据包
    - 需要知道全部目的地址；浪费带宽
  - 改进：多目标路由：每个数据包包含一组目标地址（位图），由路由器确定输出线路
    - 路由器为每个要用到的输出线路生成一份副本
    - 提高带宽利用率
    - 需要知道全部目的地址
  - 洪泛（flooding）
  - 逆向路径转发（计算出最短路径后的优化）

逆向路径转发

- 逆向路径转发 RPF (Reverse Path Forwarding)：如果该分组是从最佳路径被转发来的，就向除到来的那条线路以外的所有其他线路转发
- 建立一棵包括所有节点的汇集树（sink trees）
  - 汇集树（4hops，14packets）；RPF(5hops，24packets)
  - 优化：沿着汇集树进行逆向路径转发



Network　　Sink tree for *I* is efficient broadcast　　RPF from *I* is larger than sink tree

广播路由-3

- 逆向路径转发
  - 优点
    - 有效，易于实现
    - 在每个方向的链路上发送一次广播数据包
    - 路由器只需要知道如何到达全部目标（路由表）
- 改进：建立以发起广播的路由器为根的汇集树
  - 汇集树是生成树的一种
  - 每个路由器必须知道该生成树
    - 例：链路状态路由算法

路由算法分析

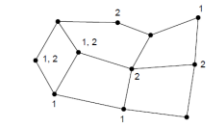
- 最短路径算法
- 洪泛算法
- 距离向量算法
- 链路状态路由
- 层次路由
- 广播路由
- 组播路由
- 选播路由
- 移动主机路由
- 自组织网络路由

组播路由-1

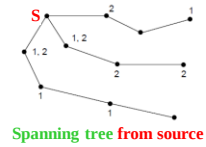
- 组播Multicasting：sends to a subset of the nodes called a group
  - 应用：视频会议，点播
- 组管理：创建，撤销，组成员维护
- 组播路由：与组的密度分布相关
  - 密集分布：接收者遍布在网络的大部分区域
    - 广播与剪枝
  - 稀疏分布：大部分网络都不属于组

组播路由-2 密集分布

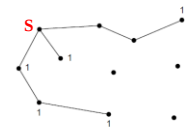
- 密集分布 Dense Case：不同的组播组有不同的生成树



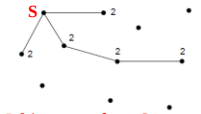
Network with groups 1 & 2



Spanning tree from source S



Multicast tree from S to group 1



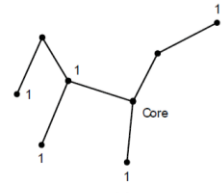
Multicast tree from S to group 2

生成树的剪枝方法

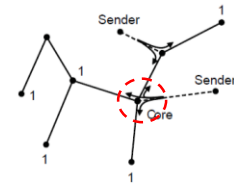
- MOSPF, Multicast OSPF，1994
  - 构造一棵以发送者为根的汇集树，删除所有不连到组成员的链路
- DVMRP, 距离向量组播路由协议，1988
  - 逆向路径转发
  - 处理剪枝 PRUNE消息，
  - 通过递归方法修剪一棵生成树

组播路由-3

- 稀疏分布 Sparse Case：CBT (Core-Based Tree) 计算某个组的单棵生成树（基于核心树）
  - Tree is the sink tree from core node to group members
  - Multicast heads to the core until it reaches the CBT



Sink tree from core to group 1



Multicast is send to the core then down when it reaches the sink tree

Internet的组播

- IPv4
  - D类地址
  - 实验床 Mbone，使用隧道技术(tunneling)
- IPv6
  - 可扩展性问题
- 组成员管理
  - IPV4：IGMP (Internet Group Management Protocol) 协议 RFC3376
  - IPv6: the protocol is Multicast Listener Discovery (MLD)
  - 查询和应答
- 组播路由协议

Internet的组播路由协议

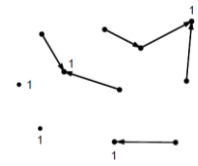
- AS内，协议独立组播协议（PIM）,2006
  - 密集模式（PIM-DM）：逆向路径转发树（RPF with pruning）
    - 应用：数据中心网络把文件分发给多个服务器
  - 稀疏模式(PIM-SM)：基于核心树（core-based trees）
    - 应用：内容提供商组播IP TV节目
- AS内其他路由协议
  - 距离向量组播路由协议DVMRP, 1988
  - 组播OSPF(MOSPF), 1994
- AS间
  - BGP或隧道的组播扩展

路由算法分析

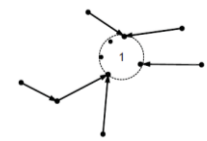
- 最短路径算法
- 洪泛算法
- 距离向量算法
- 链路状态路由
- 层次路由
- 广播路由
- 组播路由
- 选播路由
- 移动主机路由
- 自组织网络路由

选播路由Anycast-1

- Anycast：sends a packet to one (nearest) group member，RFC1546 (11/93) , RFC2101 (2/97), RFC2181 (7/97)
- IPV4：多个主机共享同一个单播地址（DNS支持）



Anycast routes to group 1



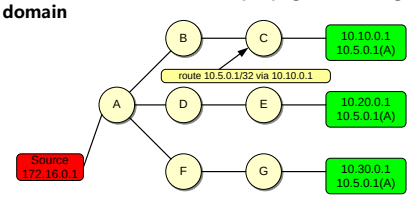
Apparent topology of sink tree to “node” 1

选播路由Anycast-2

- IPv6 anycast
  - Architecture (RFC1884, now RFC3513)
  - Reserved anycast addresses (RFC2526)
  - Anycast v4 prefix for 6to4 routers (RFC3068)
  - Source address selection (RFC3484)
  - DHCP (RFC3315)
- Anycast authoritative name service (RFC3258)
- Anycast for multicast RP (RFC3446)

选播路由Anycast-3

- 静态路由：配置简单
  - 主机路由：configure static routes on first-hop routers (host routes)
  - 路由传播：Ensure routes are propagated through domain

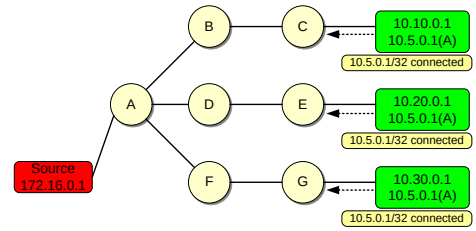


选播路由Anycast-4

- 动态路由
  - 服务器主机上运行路由软件（IGP路由协议：RIP, OSPF）
    - Run a host-based routing daemon on anycast servers
    - 例：GateD，Zebra/Quagga（开源路由器软件）
  - 主机发出路由消息
    - Host itself is route originator
    - When host is down, route is withdrawn
    - Leverages routing infrastructure

选播路由Anycast-5

- 动态路由：Dynamic IGP Routes
  - Each host announces route to IGP cloud



应用

- DNS（Domain Name System，域名系统）
  - 分布式域名服务器共享相同的IP地址，在IP层进行透明的服务定位
    - 例如，在IPv6网络中它可以共享一个熟知的IP地址，用户不需要特殊配置也不用关心访问的是哪一台DNS服务器；
  - 路由器选择“最近”的服务，缩短了服务响应的时间，同时减轻了网络负载
  - 路由器可以选择轻负载、高带宽路径转发报文

移动主机路由

（可选）

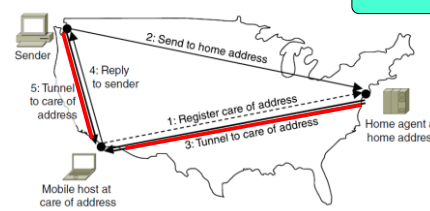
移动主机路由-1

- 如何路由一个数据包到移动主机？
- 模型
  - 家乡位置；家乡地址（home address）
  - 家乡代理（home agent）：Mobile hosts can be reached via a home agent
- 在网络层之上提供移动，例如：笔记本电脑的使用
  - 高层应用的位置服务（如Skype的再次登录）
  - 无法保持网络连接
- 网络层的移动性：移动路由
  - 移动主机把当前位置报告给家乡代理
  - 家乡代理进行数据转发

移动主机路由-2

- 例子：采用隧道技术tunneling
  - Fixed home agent tunnels packets to reach the mobile host; reply can optimize path for subsequent packets
  - No changes to routers or fixed hosts

三角路由问题



移动主机路由-3

- 如何向移动主机投递分组？
  - How does the home agent intercept拦截 a packet that is destined for the mobile node?
    - **Proxy ARP（ARP代理）**
  - How does the home agent then deliver the packet to the foreign agent?
    - **IP tunnel（隧道技术）**
    - 移交地址Care-of-address：当移动设备连接到非家乡网络时，为使之能够收发信息而分配给移动设备的临时IP地址。
  - How does the foreign agent deliver the packet to the mobile node?

自组织网络路由

（可选）

自组织网络的路由

- 自组织网络（Ad hoc）
  - 每个节点用无线通信，同时承担路由器和主机的双重角色，在网络中彼此靠近
  - 移动自组织网络（MANET, Mobile Ad hoc NETworks）
- 动态网络拓扑结构
- 路由协议
  - DSDV
  - DSR
  - AODV

DSDV

- Destination-Sequenced Distance Vector (DSDV) 节点序列距离矢量协议
- 逐跳hop-by-hop的距离向量路由协议DSDV，使用Bellman-Ford 算法
  - each node periodically broadcast routing updates
  - it guarantees loop-free (traditional DV doesn't)
- 每个节点维护一个到每个目的节点的“下一跳”路由表
- DSDV用序号标记每条路由 (the higher, the better)
- 每个节点通告路由的序号单调递增
- 每个节点周期性广播更新报文

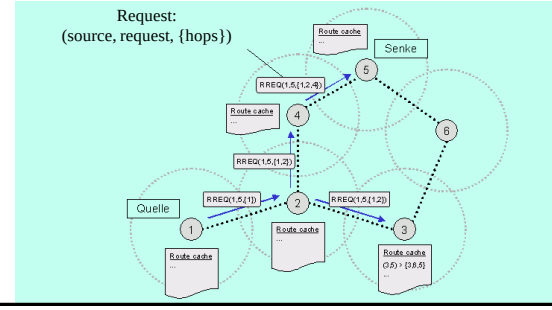
DSR

- 动态源路由协议 Dynamic Source Routing(DSR)
- 使用源路由source routing方法，每个分组带有一个路由节点序列
  - 中间节点不需要维护最新的路由信息
  - 不需要周期性路由通告和邻居发现分组
- DSR 协议包括：
  - 路由发现Route Discovery
  - 路由维护Route Maintenance.

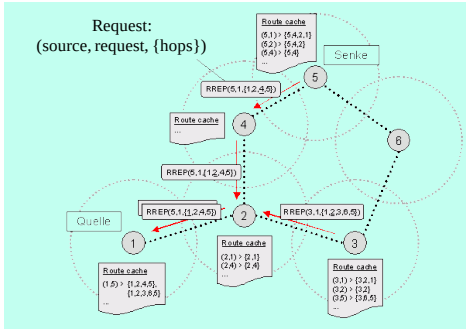
67

DSR路由发现

- Start: 源节点广播一个REQUEST 分组，洪泛到整个网络
- Propagate: 目的节点或其他知道到目的节点路由的节点进行响应REPLY.



DSR



69

DSR实现

- 检测：拓扑结构发生变化时进行检测
  - source node is notified with a ROUTE ERROR packet.
- 决策:
  - 是否使用另一个路由
  - 是否启动发现新的路由： Route Discovery protocol
- 路由发现
  - 收到ROUTE REQUEST 分组的节点返回 ROUTE REPLY 消息
- 节点发送 TTL=0的ROUTE REQUEST. 若超时，洪泛发送一个 ROUTE REQUEST.

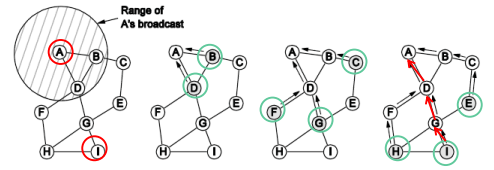
70

AODV

- Ad hoc按需距离向量算法AODV（Ad Hoc Ondemand Distance Vector）
  - 节点带宽有限
  - 电池寿命较短
- AODV 是DSR和DSDV的结合
  - 使用DSR中的按需机制（Route Discovery和 Route Maintenance ）
  - 使用DSDV中的 hop-by-hop 方法， 包括序号，周期性beacon信息等。

AODV

- 按需发现路由
  - 模型：每个节点可以与位于其覆盖范围内的其他节点通信；连接对称



A's starts to find route to *I*

A's broadcast reaches *B* & *D*

B's and *D*'s broadcast reach *C*, *F* & *G*

C's, *F*'s and *G*'s broadcast reach *H* & *I*

实验环境 ns-3

模拟工具ns-3

https://www.nsnam.org/

- Node mobility
- A realistic physical layer including:
  - a radio propagation model
  - supporting propagation delay
  - capture effects
  - carrier sense
- Radio network interfaces with properties such as:
  - transmission power
  - antenna gain
  - receiver sensitivity
- IEEE 802.11 MAC protocol using DCF
- Attenuates the power of a signal
- Reference distance
- .....

路由协议性能比较

- DSR was the best.
- DSDV performs well when load and mobility is low, poorly as mobility increases.
- AODV performs nearly as well as DSR, but has high overhead at high mobility levels.

BGP协议

（预习）

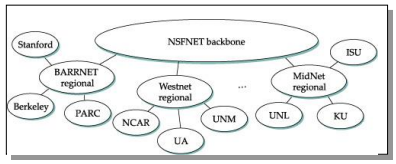
BGP协议

分层路由 Hierarchical routing

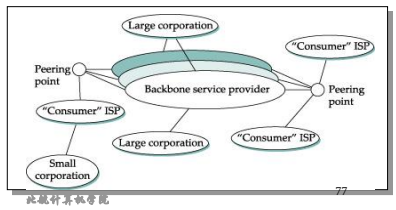
- 体系结构
  - ASes, Policies
  - BGP Attributes
- 路由
  - BGP Path Selection
  - iBGP
- 安全考虑
  - BGP的安全问题
  - 安全路由协议

全球化的Internet

- 1990年的树形结构

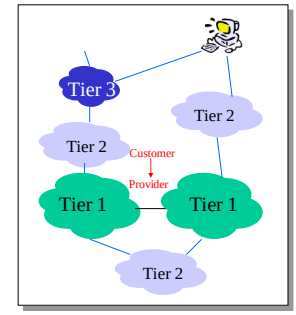


- 目前的结构



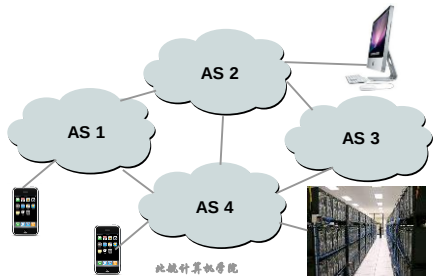
Internet的逻辑视图

- 国家级National (Tier 1 ISP)
  - "Default-free" with global reachability info（ global BGP table ）
  - Eg: AT & T, UUNET, Sprint
- 地区级Regional (Tier 2 ISP)
  - Regional or country-wide
  - Eg: Pacific Bell
- 本地Local (Tier 3 ISP)
  - Eg: Telerama DSL



将Internet连接起来

- 资源分配：Distributed cooperation for resource allocation
  - 路由BGP: what end-to-end *paths* to take (for ~50K ASes)
  - 传输TCP: what *rate* to send over each path (for ~3B hosts)



目标：路由可扩展性

Making routing scalable

- 可扩展性（scale）：数十亿级节点
  - 路由表限制：can't store all destinations in routing tables!
  - 链路带宽限制：routing table exchange would swamp links
- 自治管理：
  - internet = network of networks 网络的网络
  - each network admin may want to control routing in its own network

域间路由（Interdomain Routing）

- Internet分为不同自治系统AS（Autonomous Systems）
  - 不同管理域（domain）
  - Routers/links由单个机构进行管理
  - 包括：Service provider, company, university, ...
- 自治系统的层次结构
  - 顶层域：大型、tier-1的提供者具有国家范围内主干
  - 中等规模的地区级提供者具有较小主干
  - 由单个公司或大学管理的小型网络
- 自治系统间的交互
  - ASes之间不共享内部拓扑结构。
  - 而相邻ASes之间进行交互，协调路由。

自治系统AS的概念

- 自治系统 （Autonomous System，AS）
  - A set of routers under a single technical administration, using an interior gateway protocol (IGP) and common metrics to route packets within the AS
  - and using an exterior gateway protocol (EGP) to route packets to other AS's
- 自治系统号
  - Each AS assigned unique ID
- 信息交换
  - AS's peer at network exchanges message

自治系统号（AS Numbers）

AS Numbers are 16 bit values.

Currently over 50,000 in use.

- MIT: 3
- Harvard: 11
- Yale: 29
- Princeton: 88
- AT&T: 7018, 6341, 5074, …
- Verizon: 701, 702, 284, 12199, …
- Sprint: 1239, 1240, 6211, 6242, …
- …

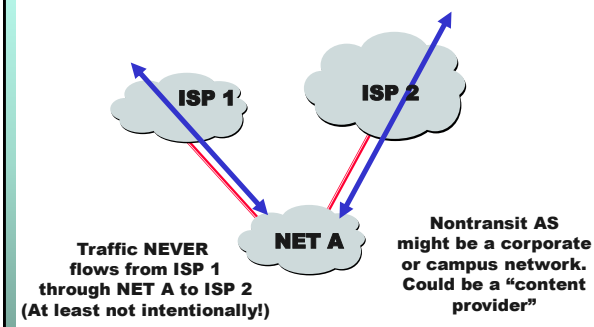
中国AS分配情况：

- CSTNET：AS7497
- HINANET:AS4808,AS4809，AS4810
- CERNET：AS4538
- 电子部信息化工程总体研究中心：AS7576
- 北京市信息中心：AS7638
- 首都公用信息平台:AS7639

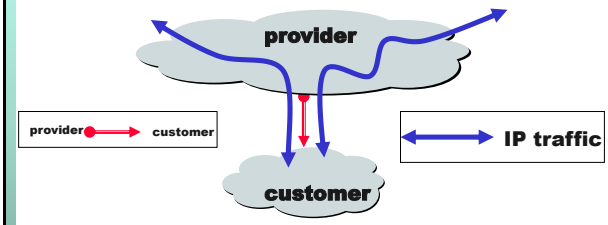
AS的类型

- ◆AS 流量类型
  - ❖ 本地流量Local traffic: starts or ends within an AS
  - ❖ 中转流量Transit traffic: passes through an AS
- ◆AS 类型
  - ❖桩stub AS: has a single connection to one other AS
    - ➢ carries local traffic only （85%）
  - ❖多宿multihomed AS: has connections to more than one AS
    - ➢ refuses to carry transit traffic
  - ❖中转transit AS: has connections to more than one AS
    - ➢ carries both transit and local traffic

非中转Nontransit vs. 中转Transit ASes



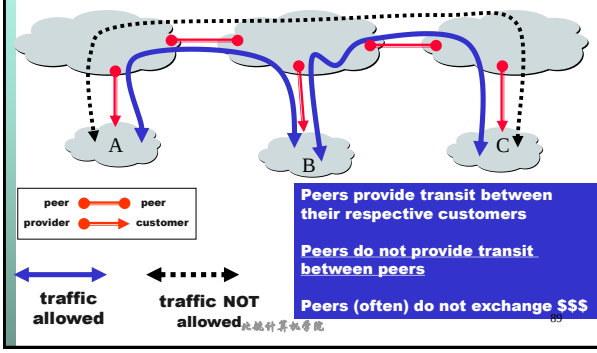
客户Customers vs.提供商Providers



付费：Customer pays provider for access to the Internet

对等关系

◆对等关系（The Peering Relationship）



路由问题

◆IGP协议是否可以适用

- ❖OSPF(Link state) or IRP（distance vector）?
- ❖没有统一的路径度量参数 – 策略policy决定

◆距离向量算法（distance-vector）的问题

- ❖慢收敛：Bellman-Ford algorithm may not converge

◆链路状态算法（link state）的问题:

- ❖度量参数：Metric used by routers not the same
- ❖LSP数据库：LS database too large – entire Internet
- ❖隐私：May expose policies to other AS's

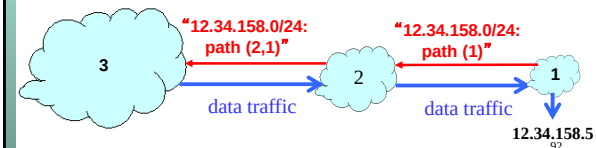
路径向量（Path-Vector）路由

◆扩展距离向量算法

- ❖交换完整路径信息：Each routing update carries the entire path
- ❖支持灵活的路由策略
- ❖环路检测，避免无穷计算问题（count-to-infinity problem）

◆主要思想：通告全部路径

- ❖RIP(距离向量): 对每个目的地址，发送距离参数（*metric*）
- ❖路径向量：对每个目的地址，发送整个路径（*entire path*）



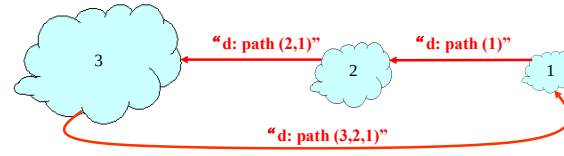
加速环路检测

◆环路检测

- ❖Look for its own node identifier in the path
- ❖E.g., node 1 sees itself in the path "3, 2, 1"

◆去除环路

- ❖E.g., node 1 simply discards the advertisement



EGP 和 BGP

- 域间路由协议（Inter-domain Routing Protocols）
  - Exterior Gateway Protocol (EGP)
    - Internet树形结构
      - 单一主干，自治系统分层连接，不支持对等端（peers）
  - Border Gateway Protocol (BGP)
    - 自治系统任意连接方式
- BGP的应用场景
  - 大型公司直接连接到一个或多个主干，其他的则连接到较小的，非主干服务提供商。
  - 很多服务提供商主要向“客户”提供服务（家庭用户），这些服务提供商必须连接到主干网提供商
  - 很多服务提供商通过“peering point”彼此互联。

Border Gateway Protocol（BGP）

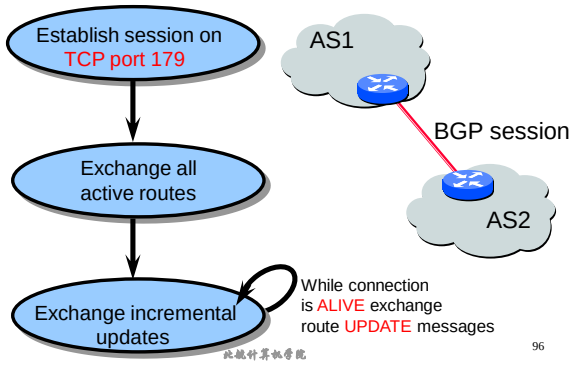
- Internet的域间路由协议
  - 基于前缀的路径向量路由协议

    Prefix-based path-vector protocol
  - 基于策略的路由协议

    Policy-based routing based on AS Paths

- 1989：BGP-1 [RFC 1105], replacement for EGP
- 1990：BGP-2 [RFC 1163]
- 1991：BGP-3 [RFC 1267]
- 1995：BGP-4 [RFC 1771], support for CIDR
- 2006：BGP-4 [RFC 4271], update

BGP 操作：会话



BGP Peers的互连

- BGP使用TCP协议，端口号179连接对等点（Peers）。
- 优点:
  - 简化BGP实现
  - 不需要周期性刷新： routes are valid until withdrawn, or the connection is lost
  - 增量更新 Incremental updates
- 缺点：
  - 拥塞控制对路由协议的影响?
  - TCP的脆弱性：Inherits TCP vulnerabilities!
  - 重负载性能下降：Poor interaction during high load

BGP 如何工作?

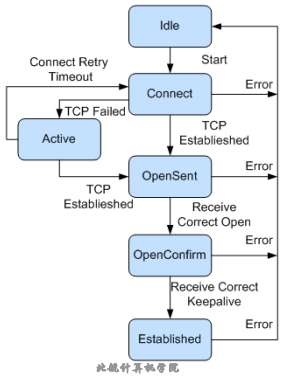
- 每个AS管理者选择BGP speaker
- 对等(Peer)路由器探测和鉴别
  - BGP peers/neighbors → BGP speaker
- 建立TCP连接
  - BGP session between speakers
  - Reliable session
- 交换BGP路由信息
  - prefix/AS path/etc.
  - CIDR

BGP 的四种消息

- Open：Peer之间建立TCP连接, 鉴别发送方身份，协商参数
- Notification：差错报告，也可用于关闭连接
- Keepalive：在没有update消息的情况下保持连接，可以作为特定帧的响应帧或链路维持帧(默认60s)
- Update：通告路由信息；初始时交换全部路由表，此后，采用增量更新的方式，即只声明新路由或撤销无效路由

announcement =
prefix + attributes values

BGP状态机



BGP状态机说明

1. 初始状态Idle：BGP收到Start事件后，和其它BGP对等体进行TCP连接，并转至Connect状态。
2. Connect状态：启动连接重传定时器（Connect Retry），等待TCP完成连接。
3. Active状态：建立TCP连接
4. OpenSent状态：BGP等待对等体的Open报文，并对收到的Open报文中的AS号、版本号、认证码等进行检查
5. OpenConfirm状态：BGP等待Keepalive或Notification报文。如果收到Keepalive报文，则转至Established状态，如果收到Notification报文，则转至Idle状态。
6. Established状态：BGP可以和对等体交换Update、Keepalive、Route-refresh报文和Notification报文

BGP的策略

- BGP支持策略（policies）配置功能
  - 策略不属于BGP协议
- 强化BGP策略：
  - 路径选择：choosing paths from multiple alternatives
  - 控制路径通告：controlling advertisement to other AS's
- 输入Import policy
  - 如何处理从邻居节点学到的路由？
  - 选择最优路径
- 输出Export policy
  - 向邻居节点通告哪些路由？
  - 取决于和邻居节点之间的关系

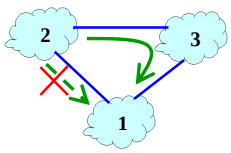
例：设置不同策略

- 拒绝转发：A multi-homed AS refuses to act as transit
  - Limit path advertisement
- 部分转发：A multi-homed AS can become transit for some AS's
  - Only advertise paths to some AS's
  - Eg: A Tier-2 provider multi-homed to Tier-1 providers
- AS选择：An AS can favor or disfavor certain AS's for traffic transit from itself

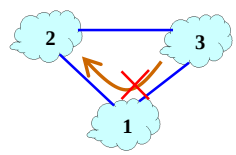
例：BGP的策略

- 每个节点使用本地策略
  - 路径选择: Which path to use?
  - 路径输出: Which paths to advertise?

路径选择：Node 2 prefers "2, 3, 1" over "2, 1"

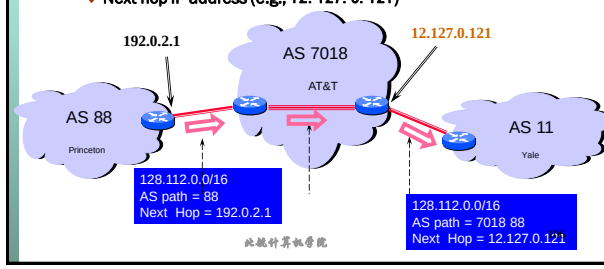


路径输出：Node 1 doesn't let 3 hear the path "1, 2"

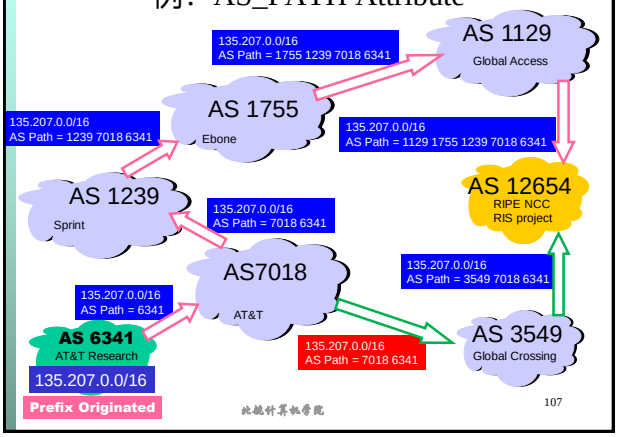


例：BGP 路由

- 目的地前缀表示 (e.g., 128. 112. 0. 0/16)
- 路由属性包括：
  - AS path (e.g., "7018 88")
  - Next-hop IP address (e.g., 12. 127. 0. 121)

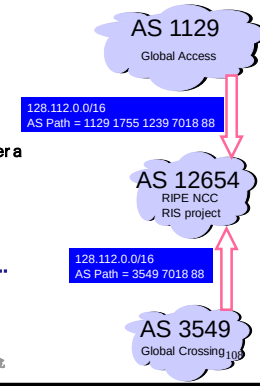


例：AS_PATH Attribute



BGP 路径选择

- 简单情况
  - Shortest AS path
  - Arbitrary tie break
- 例子
  - Three-hop AS path preferred over a five-hop AS path
  - AS 12654 prefers path through Global Crossing
- BGP采用最短路径路由,但是...
  - 基于策略的路由
    Policy-based routing



BGP是增量Update协议

- 节点学习到目的节点的多个路径
  - 在路由表中存储所有路由
  - 应用策略选择一个活动路由
  - 向邻居节点通告路由：List of reachable prefixes
- 增量更新
  - 通告Announcement
    - 选择一个新的活动路由后，在路径中增加节点ID
    - 向邻居节点通告路由（可选的）
  - 撤销Withdrawal
    - 如果活动路由不再有效
    - 向邻居节点发送撤销消息

路径选择方法

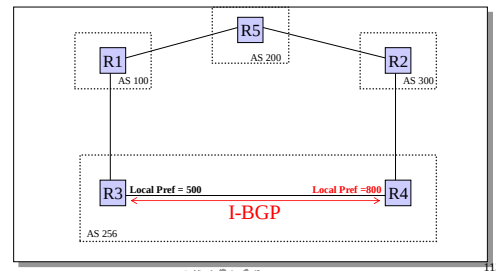
- 基于路径属性
  - 属性Attributes + 外部策略external (policy) information
- 例如:
  - 跳数 Hop count
  - 策略 Policy considerations
    - Preference for AS
    - Presence or absence of certain AS
  - 源路径 Path origin
  - 链路：Link dynamics

BGP重要路径属性

- Local Preference
- AS-Path
- MED
- Next hop

本地优先级LOCAL PREF

- 优先选择有最大LOCAL PREF值的路由
  - Local (within an AS) mechanism to provide relative priority among BGP routers （IBGP选择离开AS的首选路径）

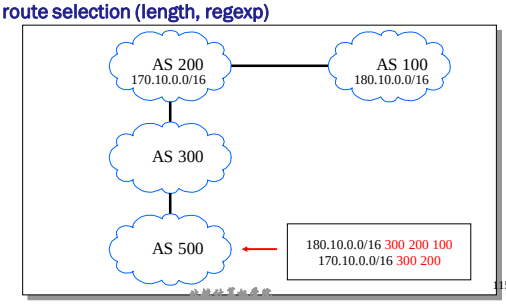


LOCAL PREF 使用方法

- 路由通告：Handle routes advertised to multi-homed transit customers
  - Should use direct connection (multihoming typically has a primary/backup arrangement)
- Peering vs. transit
  - Prefer to use peering connection
  - In general, customer > peer > provider
  - Use LOCAL PREF to ensure this

AS路径AS_PATH

- 最短AS_PATH路由优先：List of traversed AS's
- 环路检测：Useful for loop checking and for path-based route selection (length, regexp)



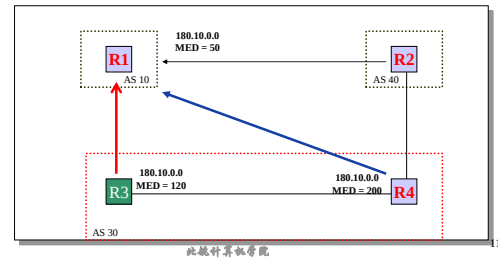
Multi-Exit Discriminator (MED)

- 多出口鉴别属性
- AS之间（EBGP）路径优先级：最小MED 值的路由优先
- Hint to external neighbors about the preferred path into an AS
  - 非传递性 Non-transitive attribute
  - 选择不同值：Different AS choose different scales
  - MED的取值范围：0~4294967295（32位）
- 两个AS之间有多个连接，选MED值小的为最优路由

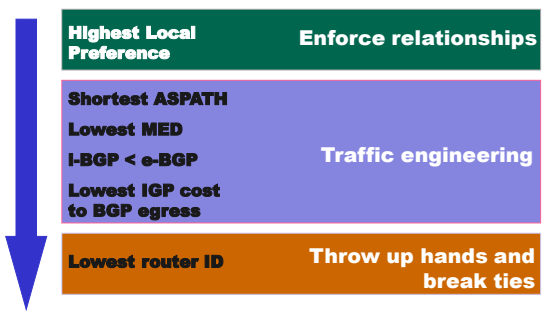
  Used when two AS's connect to each other in more than one place
  - 外部度量值。只有在AS序列号中第一个AS号码一致时，才进行MED比较

例：MED

- Typically used when two ASes peer at multiple locations
- 例：如图，优先使用R3的路由
- Cannot compare AS40's values to AS30's



路径选择过程



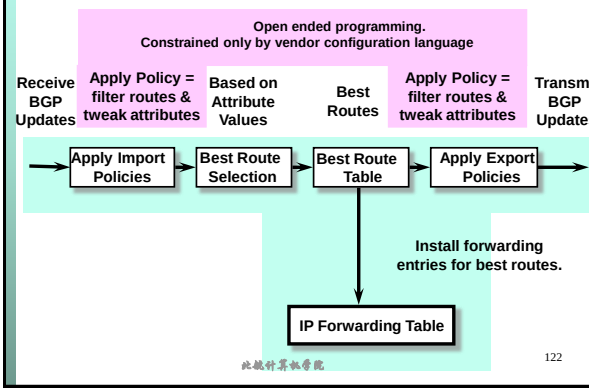
BGP协议路由选择过程

- Highest local preference
  - *Set by import policies* upon receiving advertisement
- Shortest AS path
  - Included in the route advertisement
- Lowest origin type
  - Included in advertisement or reset by import policy （IGP>EGP）
- Smallest multiple exit discriminator（MED）
  - Included in the advertisement or reset by import policy
- Smallest internal path cost to the next hop
  - Based on intradomain routing protocol (e.g., OSPF)
- Smallest next-hop router id
  - Final tie-break

BGP协议路由选择过程

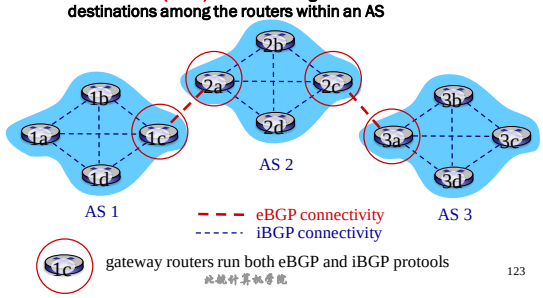
- **路由数据库： Routing Information Base**
  - Store all BGP routes for each destination prefix
  - Withdrawal message: remove the route entry
  - Advertisement message: update the route entry
- **路径选择：Selecting the best route**
  - Consider all BGP routes for the prefix
  - Apply rules for comparing the routes
  - Select the one best route
    - Use this route in the forwarding table
    - Send this route to neighbors

BGP Policy: 处理过程



AS 中的多个路由器

- **AS中有多个路由器**
  - External BGP (eBGP): exchanging routes between ASes
  - Internal BGP (iBGP): disseminating routes to external destinations among the routers within an AS



eBGP 和 iBGP

- **路由器的角色**
  - Speaker：发送BGP报文的设备称为BGP Speaker，它接收或产生新的报文信息，并发布（Advertise）给其它BGP Speaker。
  - Peer：相互交换报文的Speaker之间互称对等体（Peer）。若干相关的对等体可以构成对等体组（Peer Group）。
- **配置BGP时，每个AS中至少有一个节点作为BGP speaker**
- **BGP speaker的交互：与其他相邻AS、或在本AS内进行交互**
  - eBGP：相邻AS之间
  - iBGP：本AS内
    - 一个AS内有多个BGP speakers
    - 在BGP 路由器之间发布路由信息

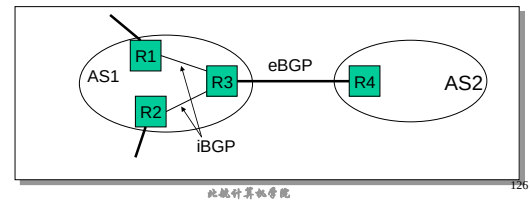
iBGP

- 在同一个AS内，多个BGP 路由器之间需要交换从相邻AS学到的路由信息
  - 问题：为什么不使用AS内的路由协议IGP?
  - iBGP must be full-mesh: each pair of BGP speaking routers has an iBGP session（逻辑连接）
- 特点
  - 消息格式与eBGP相同，前缀通告规则不同:
    - Prefix learned from eBGP can be advertised to IBGP neighbor and vice-versa, but
    - Prefix learned from one IBGP neighbor cannot be advertised to another IBGP neighbor
  - 说明：为了防止AS内产生环路，BGP设备不将从iBGP对等体学到的路由通告给其他iBGP对等体。为了解决iBGP对等体的连接数量太多的问题，BGP设计了路由反射器和BGP联盟。

例子：iBGP

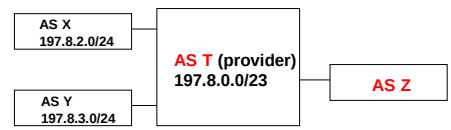
- R3 can tell R1 and R2 prefixes from R4
- R3 can tell R4 prefixes from R1 and R2
- R3 cannot tell R2 prefixes from R1

R2 can only find these prefixes through a *direct connection* to R1
Result: iBGP routers must be fully connected (via TCP)!
- contrast with eBGP sessions that map to physical links



IBGP	EBGP
AS内部会话必须全连接（TCP），因IBGP宣告者不允许把从IBGP邻居接收到的更新发给其它的IBGP邻居	无全连接要求
通告LOCAL_PREF属性，默认不修改下一跳和AS_PATH;	ebgp环境不通告LOCAL_PREF，通告MED；默认修改下一跳和AS_PATH
不需要直接连通，因AS内可达性由IGP提供；	缺省需要直接连通，默认EBGP对等体间是不需要共享相同的IGP的；例外就是EBGP多跳；
默认IBGP需要和IGP保持同步，如果IGP是OSPF要求：OSPF Router_ID和BGP Router_ID要一致；在全连接和不提供穿越服务时可关闭同步；同步的应用场景是转发链路上不是所有路由器都运行BGP。	无此要求
缺省条件下，即使配置了BGP到IGP的重分布，IBGP路由也不会被重分布到IGP中（防止环路），命令bgp redistribute-internal可强制IBGP重分布到IGP中；	无此限制

CIDR 和 BGP



问题：T 如何向 Z 通告路由？

例1：路由通告

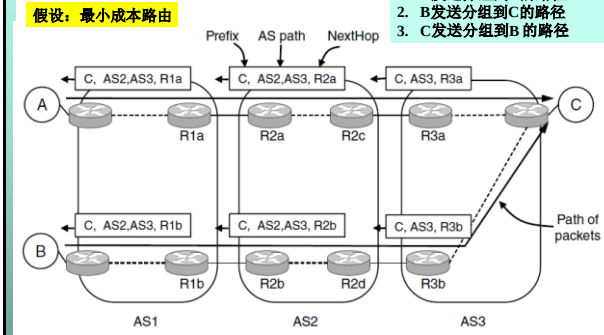
- **通告所有路径:**
  - Path 1: through T can reach 197.8.0.0/23
  - Path 2: through T can reach 197.8.2.0/24
  - Path 3: through T can reach 197.8.3.0/24
- **减小路由表：**
  - Path 1: through T can reach 197.8.0.0/22

要求：计算CIDR地址聚合

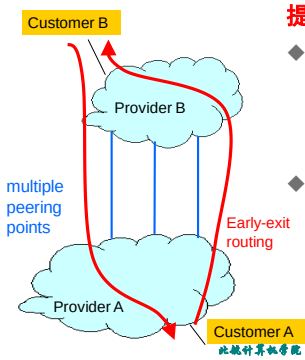
例2：BGP 路由通告传播

假设：最小成本路由

1. A发送分组到C的路径
2. B发送分组到C的路径
3. C发送分组到B的路径



Early-Exit or Hot-Potato Routing

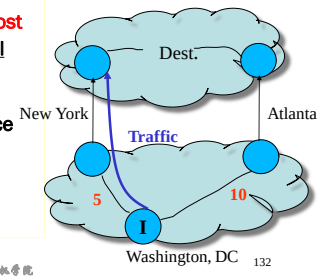


提前退出/热土豆路由

- **Selfish routing**
  - Each provider dumps traffic on the other
  - As early as possible
- **Asymmetric routing**
  - Traffic does not flow on same path in both directions

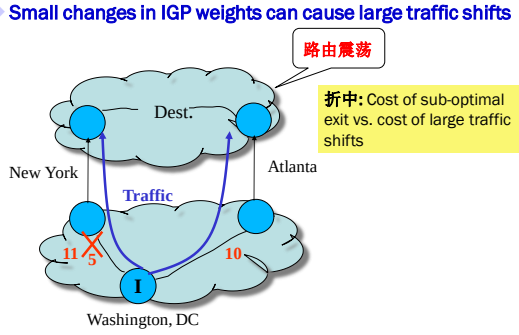
实现：Realizing Hot-Potato Routing

- **Hot-potato routing**
  - Each router selects the closest egress point
  - ... based on the path cost in intradomain protocol
- **BGP decision process**
  - Highest local preference
  - Shortest AS path
  - Closest egress point
  - Arbitrary tie break



问题：Hot-Potato Routing

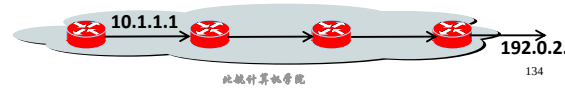
- Small changes in IGP weights can cause large traffic shifts



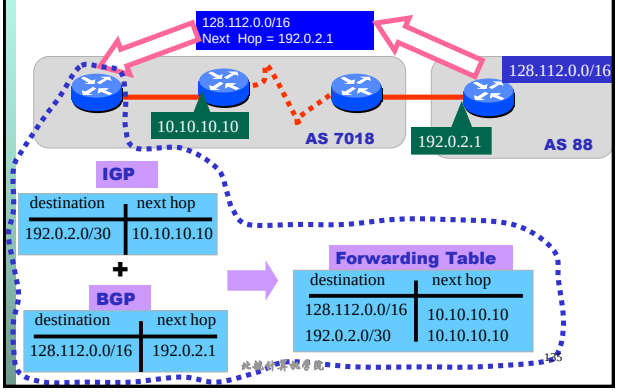
折中: Cost of sub-optimal exit vs. cost of large traffic shifts

BGP 和 IGP 路由信息的结合

- Border Gateway Protocol (BGP)
  - Announces reachability to external destinations
  - Maps a destination prefix to an egress point
    - 128.112.0.0/16 reached via 192.0.2.1
- Interior Gateway Protocol (IGP)
  - Used to compute paths within the AS
  - Maps an egress point to an outgoing link
    - 192.0.2.1 reached via 10.1.1.1



例子：结合BGP和 IGP 信息的路由表



BGP小结

- 路由变化：The only constant is change
  - Planned topology and configuration changes
  - Unplanned failure and recovery
- 路由协议收敛：Routing-protocol convergence
  - Transient period of disagreement
  - Blackholes, loops, and out-of-order packets
- 路由不稳定：Routing instability
  - Permanent conflicts in routing policy
  - Leading to bi-stability or oscillation

BGP协议的安全缺陷

- 由BGP底层协议TCP的安全漏洞引入的缺陷
- 缺乏认证机制
  - 没有对源AS是否能宣告前缀路由信息进行验证的机制
  - 没有对AS路径（AS PATH）进行验证的机制
  - 没有对路由撤销（Withdrawn）进行验证的机制
  - 没有对路径属性进行验证的机制

BGP安全问题：路由劫持

- 路由劫持（前缀劫持）
  - 攻击者通过误配置或恶意通告伪造路由信息，使得其它AS选择到达受害前缀的虚假路由，从而劫持互联网上到达该前缀的流量
- 按照攻击者对劫持流量处理方式的不同，可将前缀劫持分为三类
  - 黑洞：丢弃所吸附的网络流量,制造路由黑洞,阻断被劫持网络提供的服务
  - 伪装：使用属于被劫持网络的IP地址进行spam攻击
  - 窃听：将吸附的流量发回到被劫持网络,实现隐蔽的“中间人攻击”

典型案例

- 1997 年AS 7007事件。 AS 7007是一个小型ISP，策略配置上发生错误，导致互联网上路由表容量增加了一倍，使得很多路由器难以处理而崩溃。
- 2008年4月，著名的巴基斯坦电信劫持事件使得YouTube从互联网上消失了近两个小时
- 2010年4月，中国电信的一个下属AS将属于170多个国家的5万条前缀劫持了将近20分钟，受影响的前缀数量相当于当时全球路由表中前缀总数的15%
- 2014年向Google DNS美国服务器发送的部分请求被重路由经过委内瑞拉的一个网络

安全路由协议

- 基于公钥基础设施PKI（Public Key Infrastructure）对BGP路由更新消息进行签名和认证
  - RPKI：引入数字证书和签名机制，采用PKI验证路由通告签名者所持有的公钥，该签名者的IP 地址分配上游为其签发证书，一方面验证其公钥，一方面验证该实体对某个IP 地址前缀的管理权。
  - IETF: BGPSEC：BGPsec Protocol Specification,RFC 8205，2017年9月27日
  - S-BGP：BBN 公司Stephen Kent 提出,采用附加签名的BGP 消息格式， 用以验证路由通告中IP 地址前缀和传播路径上AS 号的绑定关系， 从而避免路由劫持

域间路由的挑战

- 规模scale
  - 前缀规模: 250,000, and growing
  - 自治系统规模: 50,000, and growing
  - 路由器规模: at least in the millions…
- 私有性Privacy
  - 保护内部网络拓扑结构
  - 保护商业关系
- 策略Policy
  - 链路度量参数的一致性：No Internet-wide notion of a link cost metric
  - 流量控制
    - Need control over where you send traffic
    - … and who can send traffic through you
- 安全security
  - 路由劫持

思考

- 不同路径上流量的类型可能不同
  - 实时应用程序：低延时，低抖动路径；大数据应用：低丢包率、高带宽路径
  - 如何根据应用的需要将流量进行分段管理？
- 让应用程序或策略管理器进行流量转发？
  - 分段路由SR（Segment Routing）
    - IETF SPRING（Source Packet Routing in Networking）工作组

完成小作业（2）

- 专题2 “SDN”

1.任意选择1篇论文进行阅读

2.每人独立完成论文评论（paper review），评论内容要求：

- 作者主要观点和要解决的问题
- 研究方法评论（关键技术，优点和局限性）
- 论文的主要贡献
- 其他
- 注意：不是翻译，篇幅不限

3.作业提交（两个文档）

- .docx文件
- .pptx文件（约 10 页左右，请勿超过15页，课堂讨论用）