计 算 机 网 络

高级网络互连

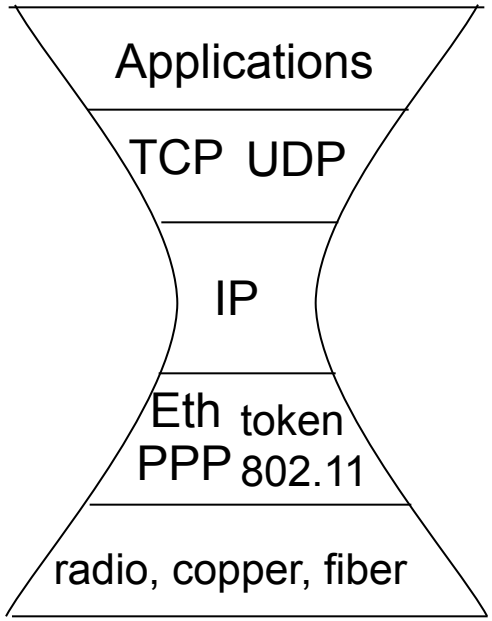
华中科技大学电信学院 2016

学习目标

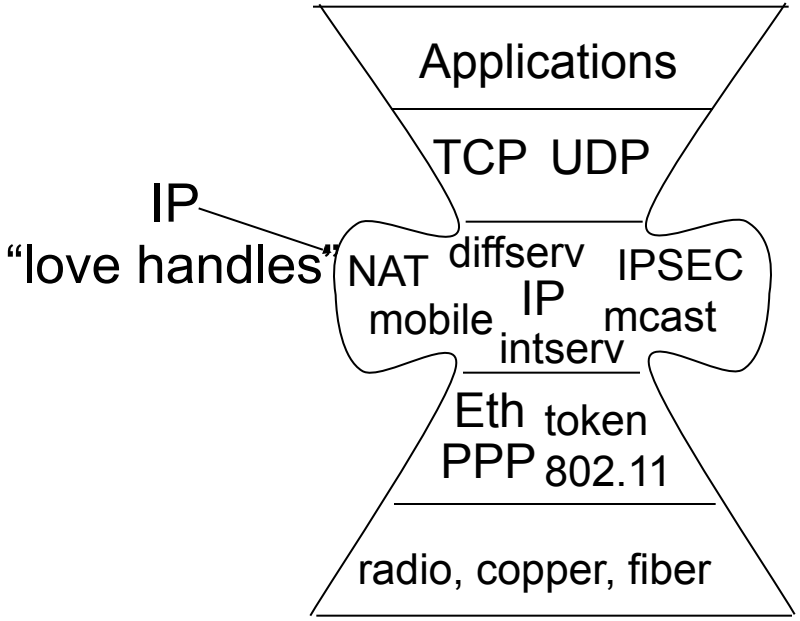
- 理解无类别域间路由的路由转发实现机制
- 掌握互联网域内路由和域间路由的概念，理解BGP协议的实现机制
- 理解IPv6的设计目标，了解IPv6的主要技术特征
- 理解多播的概念
- 了解移动主机的路由实现机制

赢者通吃：如何支持新的应用？

互联网中年危机: 思想狭隘, 心宽体胖？



IP “hourglass”



Middle-age IP “hourglass” ?

提纲

- 引言
  - 核心问题: 扩展到数十亿节点
- 全球互联网
- IPv6
- 多播
- 移动设备之间的路由
- 总结

全球互联网

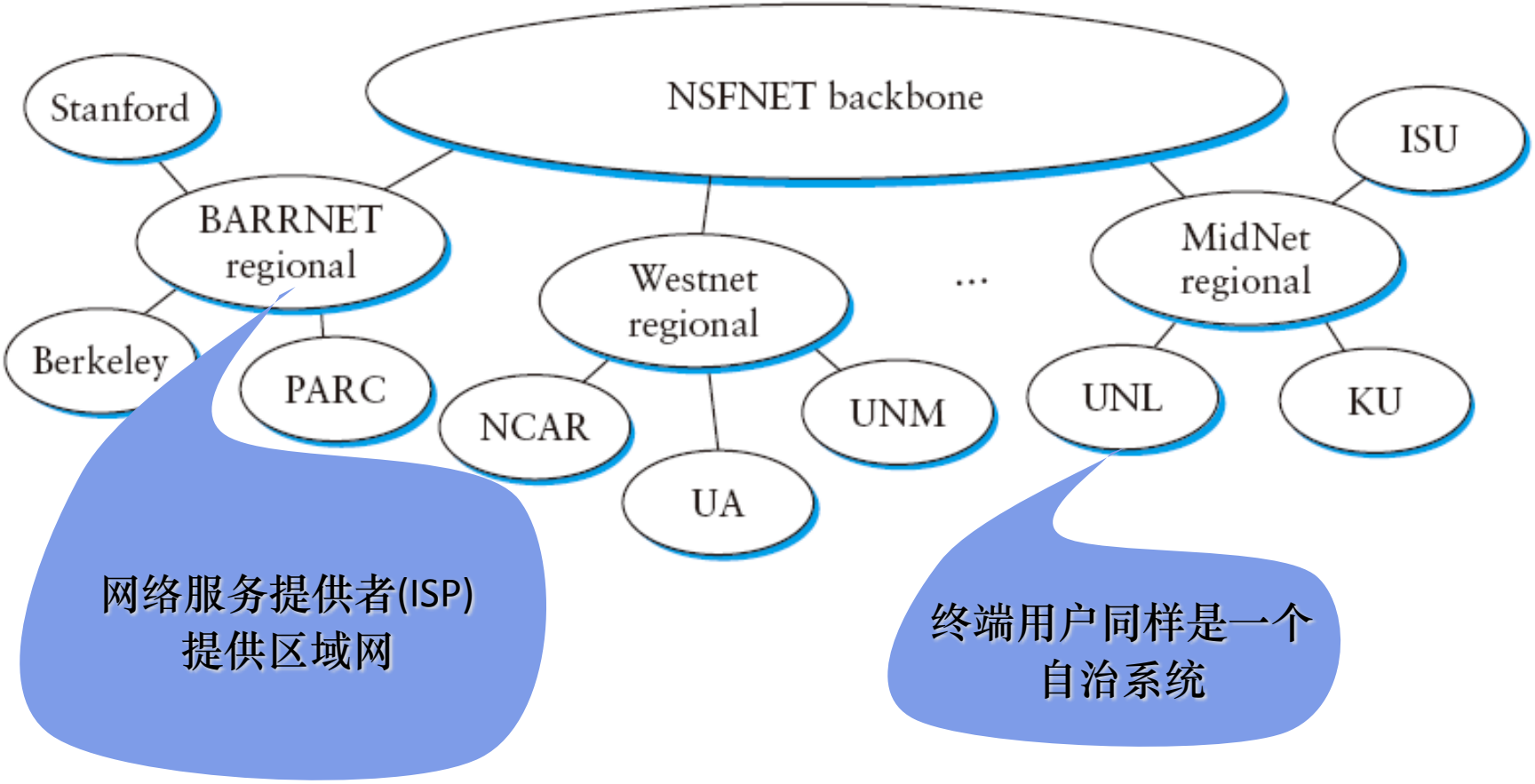
- 然而，简单的网络互联并不足以满足扩展性的要求

- IP地址的简单分层仅能实现“一定程度”的扩展
  - 路由器有必要知道连在互联网上的所有网络，这在现实中完全不可能达到

- 存在一些大幅度提高可扩展性的技术使得Internet发展到了当前的程度

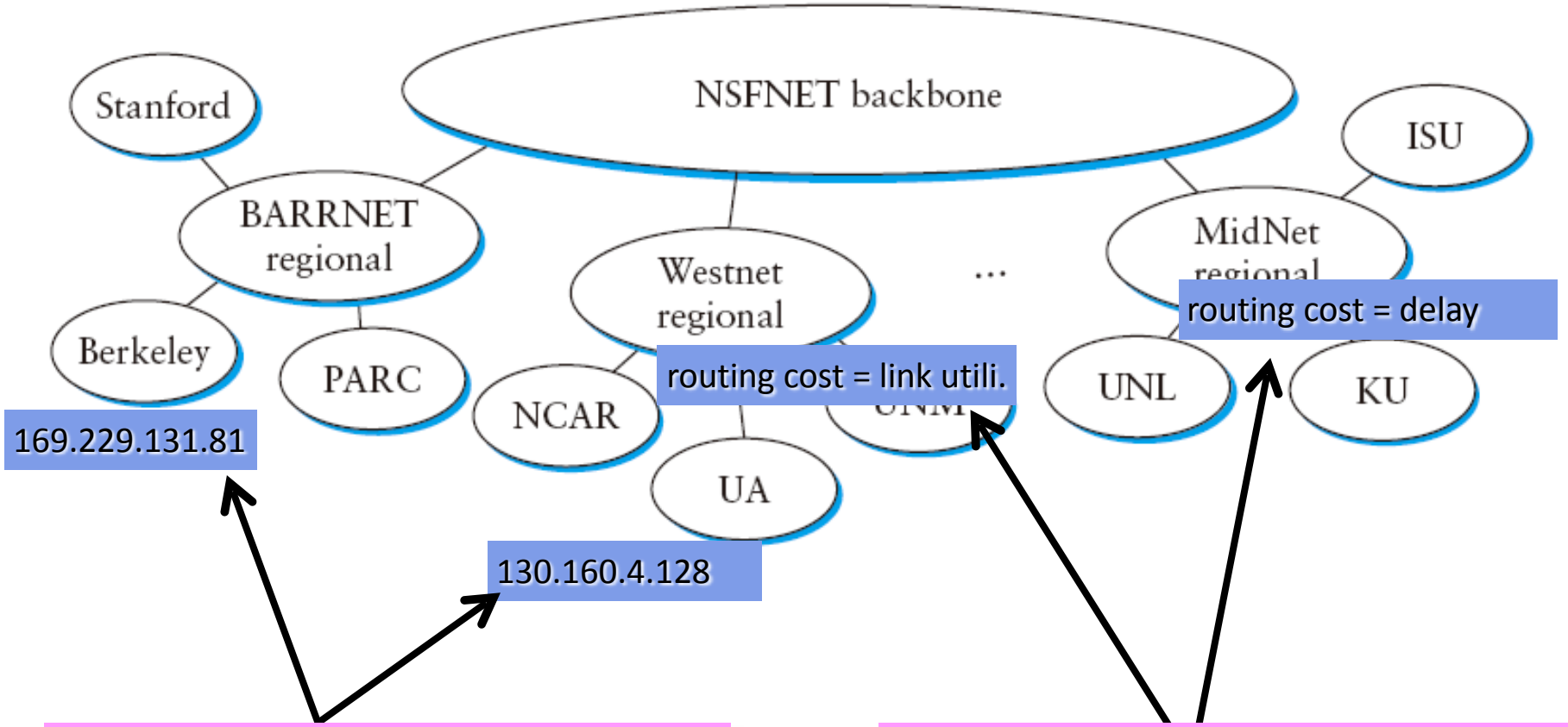
全球互联网的实际情况

1990年时Internet的树形结构



全球互联网的实际情况

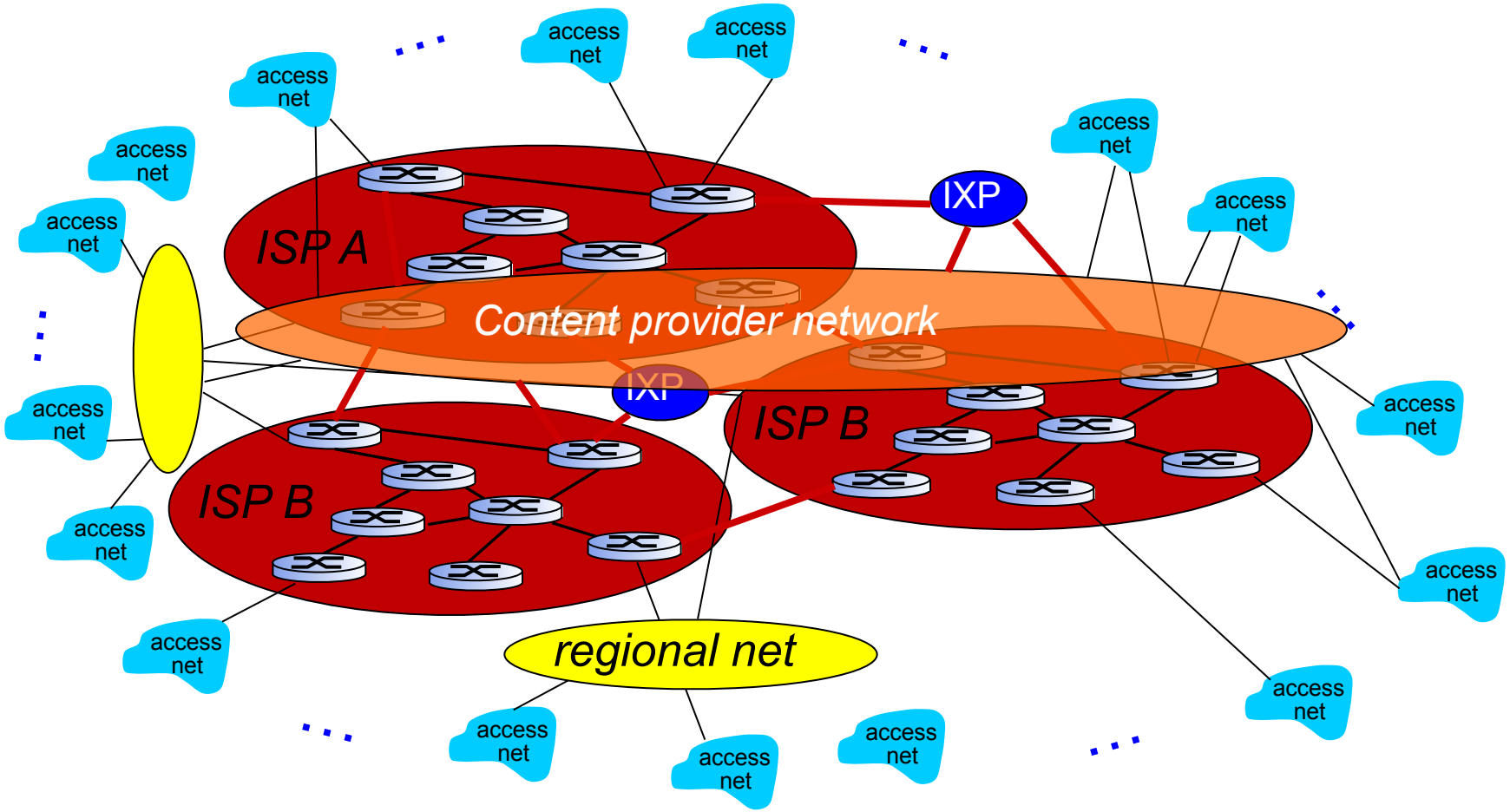
1990年时Internet的树形结构



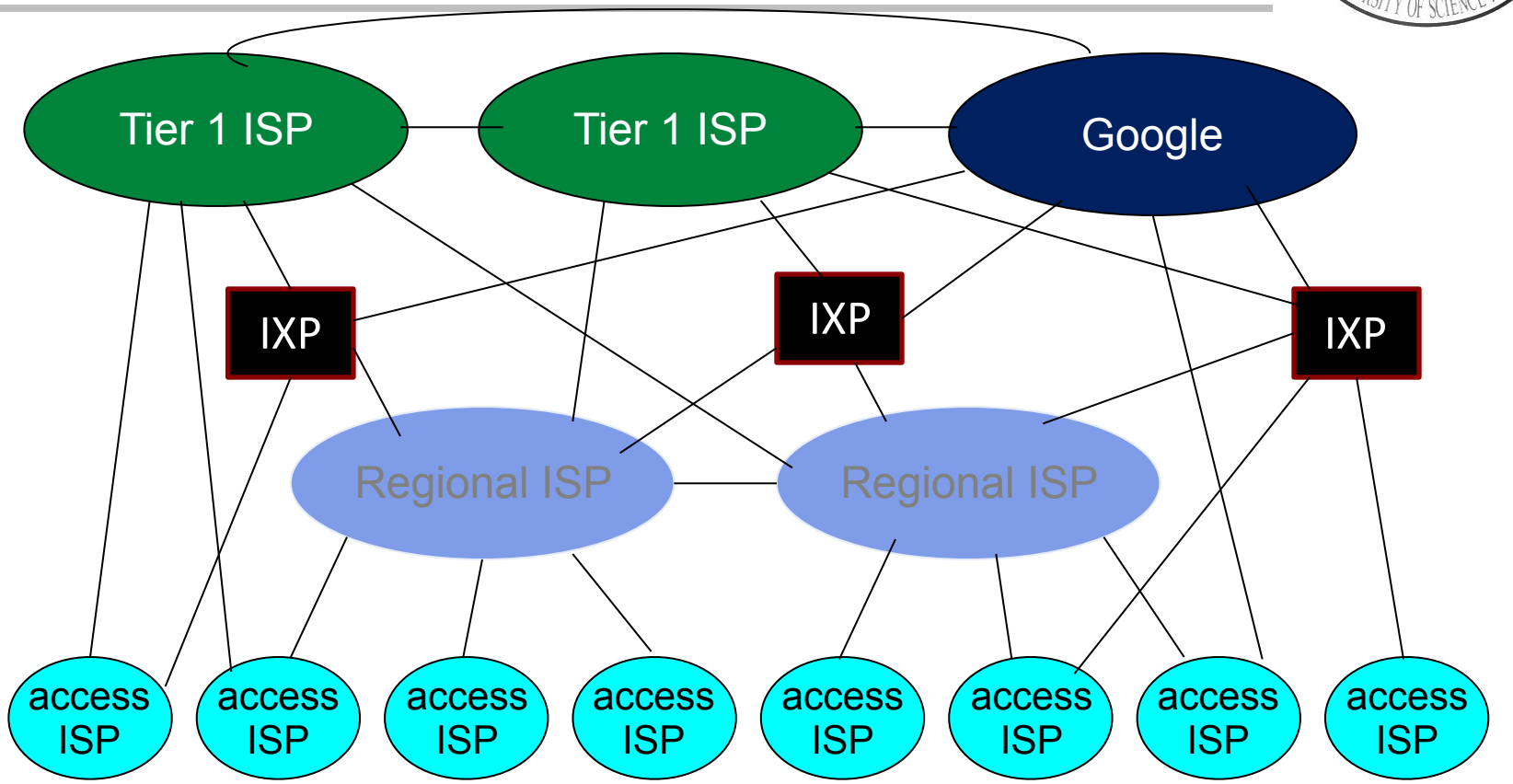
大学可以为内部用户免费分配 IP 地址
=> 地址空间的利用率

ISPs 对网络中所使用的最佳路由协议以及如何定义链路评价指标都有不同的看法
=> 自治系统

互联网架构：网络互联的网络



互联网架构：网络互联的网络



- 网络中心: 少量充分互联的大型网络
  - 顶级商业ISPs (e.g., Level 3, Sprint, AT&T, NTT, China Telcom, Unicom), 全国 & 全球覆盖
  - 内容服务网络 (e.g, Google): 建设私有网络将自己的数据中心直接连入互联网，而不通过顶级或者区域网络服务提供商

域间路由选择(BGP)

- Internet和自治系统
- 域间路由选择
- 路径向量路由选择
- BGP

Internet 和自治系统

- Internet 按照自治系统(也称为路由选择域)进行组织

- 每一个自治系统(AS)在一个独立的管理实体的控制之下
  - 示例: 校园网络, 公司网络

- 为什么提出自治系统?
  - 从管理和安全的角度考虑
  - 扩展性: 将大型互联网中路由选择信息进行层次汇聚的一种补充

自治系统

- 每一个自治系统(AS) 由一个单一的管理实体控制
- 每一个AS拥有一个 AS 号 (ASN)
- AS 号
  - 16 bits 的整数
  - 公共 AS 号: 1 – 64511
  - 私有 AS 号: 64512 – 65535
  - 示例
    - AT&T: 7018, 6431, ...
    - Sprint: 1239, 1240, ...
    - MIT: 3

AS号

- AS 号占有16-bit
  - 共存在65,536 个唯一的AS号
- 部分AS号被保留(例如, 私有AS号)
  - 只有64,510个AS号可以公共使用
- 由Internet Assigned Numbers Authority管理
  - 以1024为分配单元向区域性Internet注册中心(RIRs)进行分配
  - IANA 已向RIRs分配39,934个 AS号 (Jan'06)
- RIRs 向各机构分配AS号
  - RIRs 已分配AS号34,827个 (Jan'06)
  - 只有21,191号在域间路由选择过程中可见 (Jan'06)
- 近期已开始分配32-bit AS #s (2007)

域名管理权移交：美国想干啥？中国会受益么？
http://www.edu.cn/xxh/focus/li_lun_yj/201608/t20160822_1441513.shtml

AS的类型

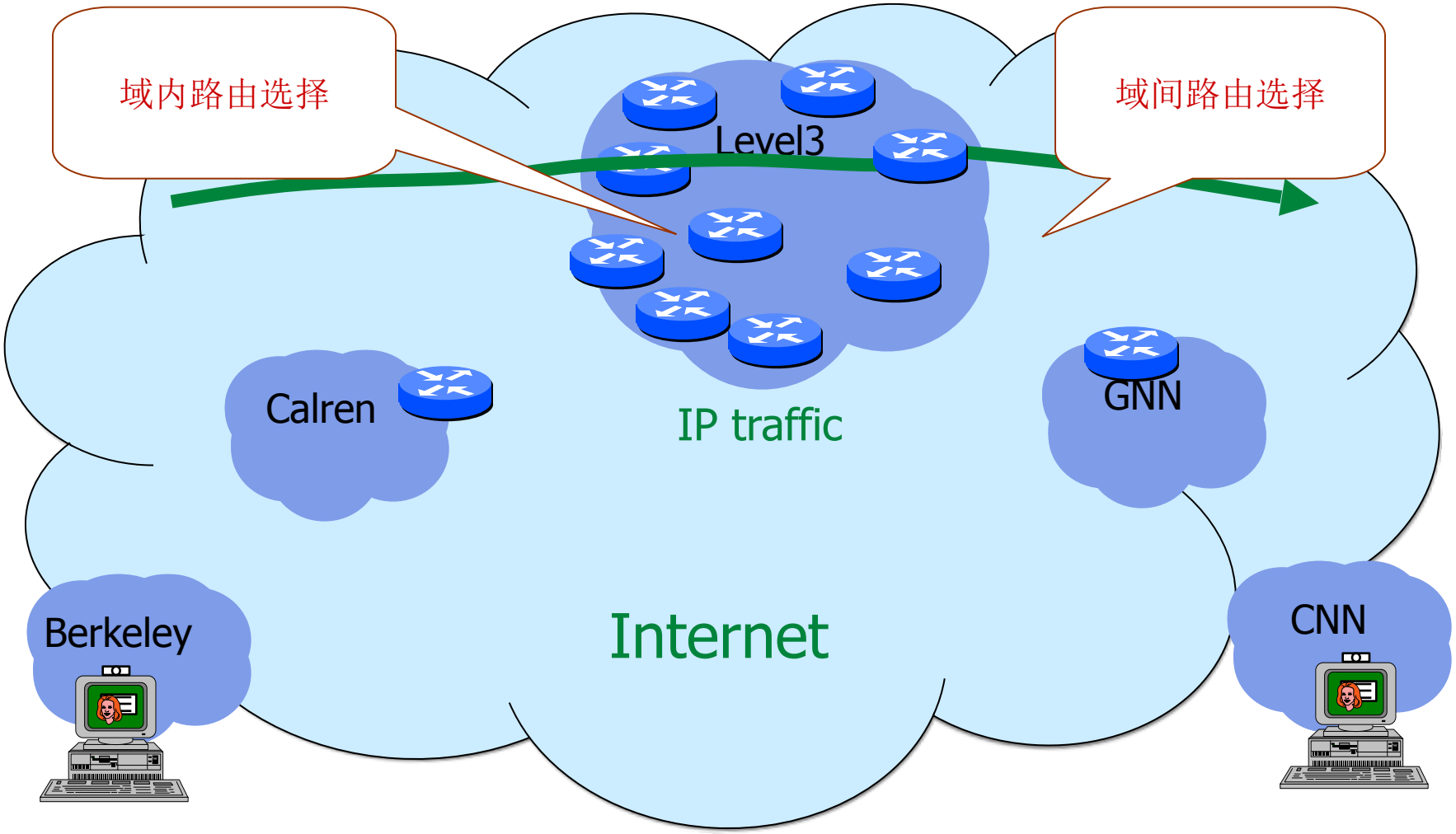
- 通信量类型
  - 本地流量: 在一个AS内部
  - 中转流量: 经过一个AS

- AS分类
  - 桩AS(Stub AS): 仅与一个其他AS相连, 因此仅包含本地流量
  - 多出口AS(Multi-home AS):与其它的自治系统具有多个连接, 但拒绝承载中转流量
  - 中转AS(Transit AS):与其它的自治系统具有多个连接, 允许承载中转流量

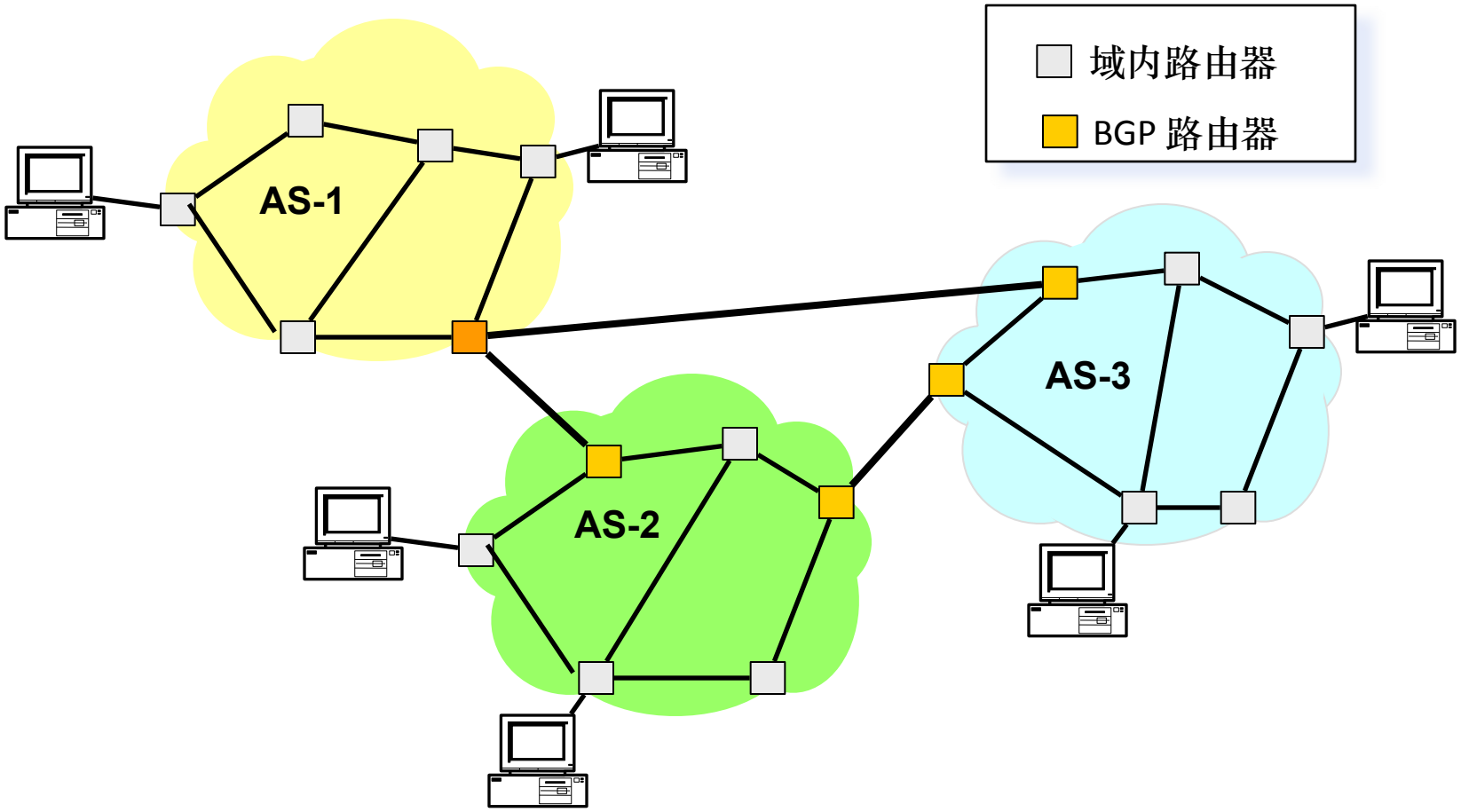
域间路由选择(BGP)

- Internet和自制系统
- 域间路由选择
- 路径向量路由选择
- BGP

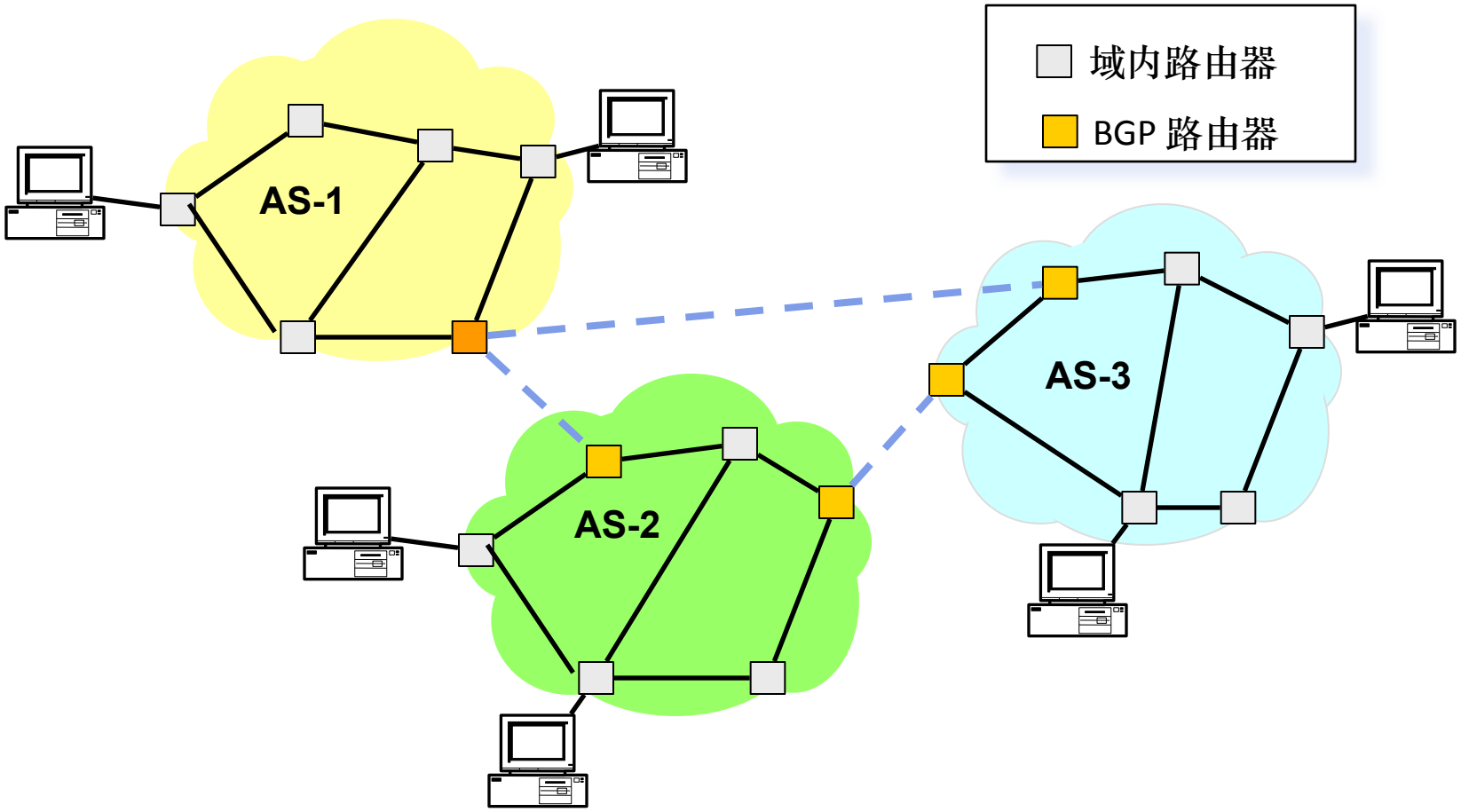
Internet 路由选择架构



示例

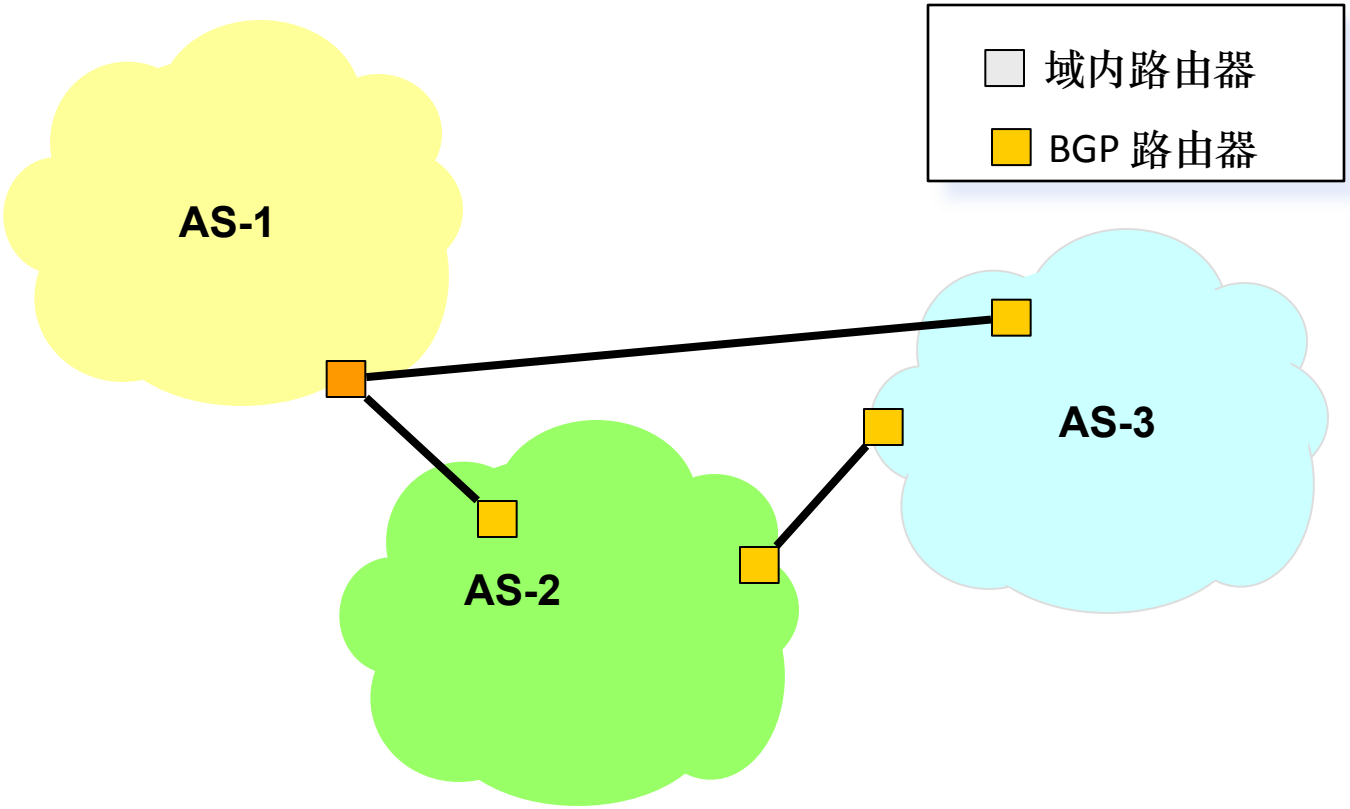


域内



域内路由选择协议又称为内部网关协议 (IGP), 例如 OSPF, RIP

域间



域间路由协议又称为外部网关协议 (EGP), 例如 BGP

两级路由

- 域内路由选择
  - 运行于一个特定的网络, 即一个自治系统内
  - 网络内两个节点之间的最优路由
  - 内部网关协议(IGP)
    - 基于评价指标
    - 示例: OSPF, RIP, IS-IS

- 域间路由选择
  - 运行于多个网络之间, 即自治系统之间 (ASes)
  - 提供整个Internet的全连接
  - 外部网关协议(EGP)
    - 基于策略
    - 示例: EGP(外部网关协议), BGP (边界网关协议)

域间路由选择面临的挑战

- 目标
  - 寻求一条通往预定目的地的无环路径
  - 更关注可达性而非最优性
- 挑战
  - 规模
    - 前缀: 200,000, 仍在不断增长
    - ASes: 已分配40K, 其中20,000+在使用中
    - 路由器: 数量至少上百万…
  - 隐私
    - ASes 不希望泄露其拓扑信息
    - … 以及与邻节点之间的商业关系
  - 策略
    - 不存在全Internet通用的链路代价评价指标
    - 需要控制从哪里传送流量
    - … 谁能通过你中转流量

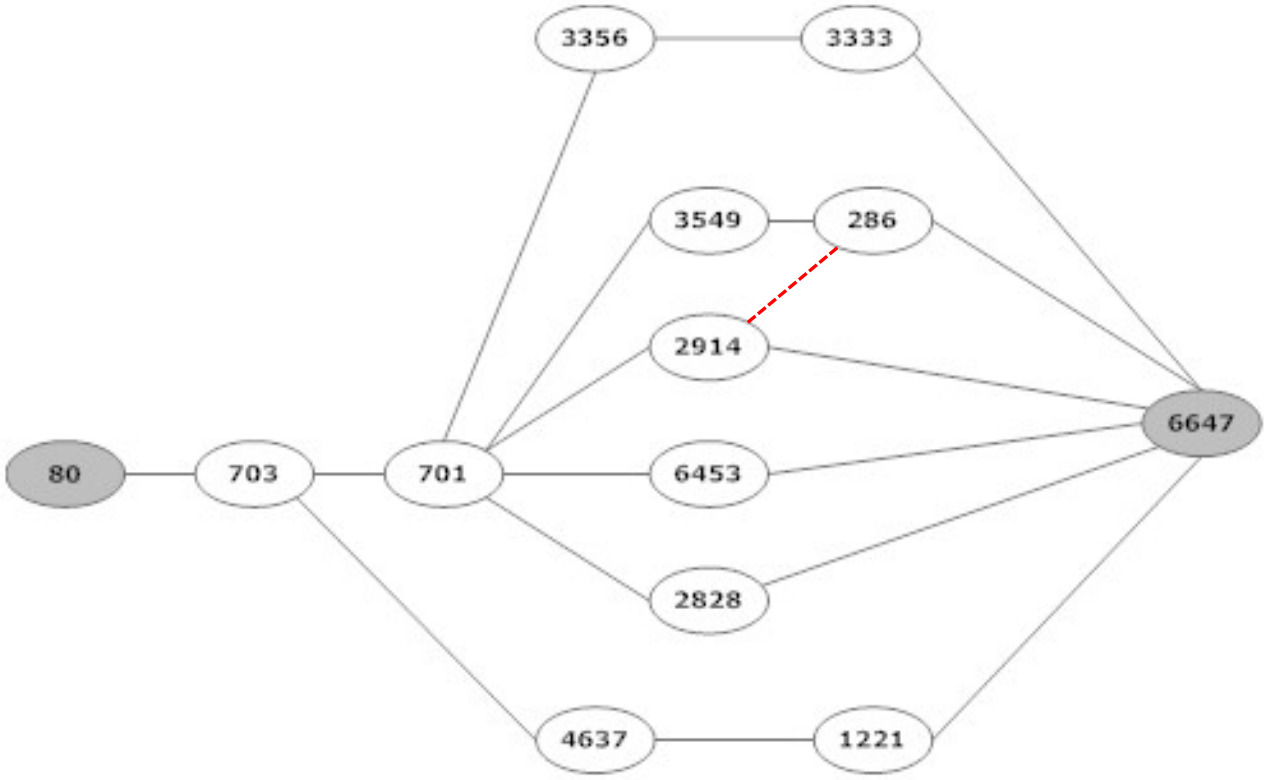
BGP路由表样例

	Network	Next Hop	Metric	LocPrf	Weight	Path
*	3.0.0.0	193.0.0.56			0	3333 3356 701 703 80 i
*		203.62.252.186			0	1221 4637 703 80 i
*		134.222.87.1			0	286 3549 701 703 80 i
*		195.219.96.239			0	6453 701 703 80 i
*		65.106.7.139	3		0	2828 701 703 80 i
*>		129.250.0.11	6		0	2914 701 703 80 i
*		157.130.10.233			0	701 703 80 i
*>	4.4.4.0/30	4.68.1.166	0		0	3356 701 703 80 i
*		203.62.252.186			0	1221 4637 4766 9318 18305?

from route-views.routervies.org(AS6647)

BGP表构造AS拓扑图

Path
3333 3356 701 703 80 i
1221 4637 703 80 i
286 3549 701 703 80 i
6453 701 703 80 i
2828 701 703 80 i
2914 701 703 80 i
701 703 80 i
3356 701 703 80 i
1221 4637 4766 9318 18305?



根据路由表构造的AS级拓扑图

BGP路由监控系统扩展

- BGP路由服务器
  - 解决了拓扑扩展问题，使BGP连接数从O(n^2)降到O(n)
  - 路由服务器为每个BGP会话维护各自的路由策略和路由表，可以通过show ip bgp等命令访问
- Looking Glasses服务器
  - 运行Looking Glasses软件，对BGP路由信息进行有限查询进行故障排查
  - 只具有Ping、Traceroute、show bgp summary等简单命令
- IRR路由信息库
  - IRR机构受理IP和AS号申请
  - 通过路由策略规范语言（RPSL）记录ISP的BGP路由信息(不强制执行) ——不完整、不具有实时性

IRR路由信息库查询样例

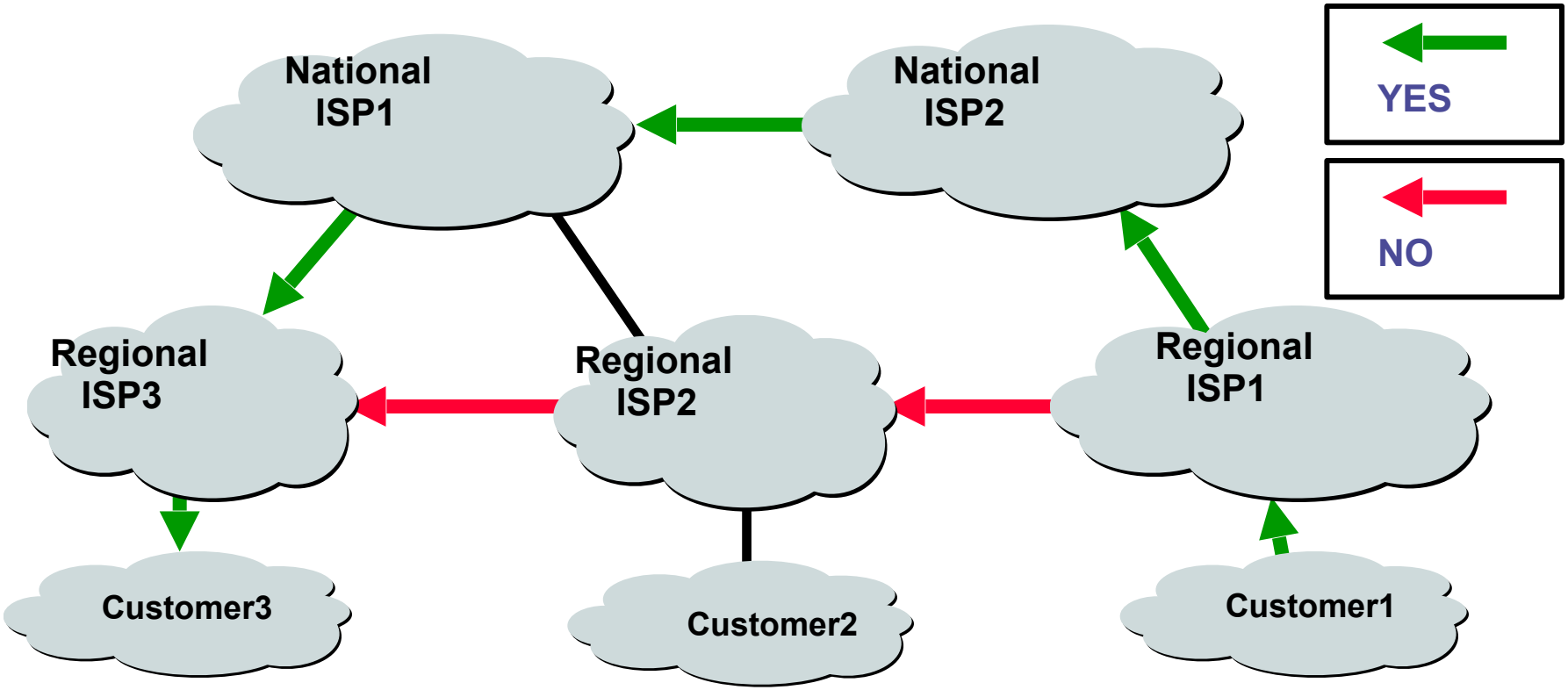
```
aut-num:        AS23910
as-name:        CNGI-CERNET2-AS-AP
descr:          China Next Generation Internet CERNET2
descr:          CNGI-CERNET
descr:          Beijing 100084, China
country:        CN
import:         from AS4538
                action pref=10;
                accept ANY
export:         to AS9406
                announce AS23910 AS4538 AS9407 AS4839 AS4840
default:        to AS4538
                action pref=10;
                networks ANY
admin-c:        CER-AP
tech-c:         CER-AP
remarks:        Multihome portion of CERNET
mnt-by:         MAINT-CERNET-AP
mnt-routes:     MAINT-CERNET-AP
changed:        hm-changed@apnic.net 20031014
source:         APNIC
role:           APNIC Hostmaster
address:        6 Cordelia Street
address:        South Brisbane
address:        QLD 4101
country:        AU
phone:          +61 7 3858 3100
fax-no:         +61 7 3858 3199
e-mail:         helpdesk@apnic.net
admin-c:        AMS11-AP
tech-c:         AH256-AP
nic-hdl:        HM20-AP
remarks:        Administrator for APNIC
```

域间路由选择(BGP)

- Internet和自制系统
- 域间路由选择
- 路径向量路由选择
- BGP

最短路由选择的约束

- 所有的流量必须通过最短路由传送
- 所有节点需要拥有统一的链路代价标识
- 无法体现商业关系



链路状态路由选择的问题

- 洪泛拓扑信息
  - 较高的带宽和存储开销
  - 强迫节点通告敏感信息

- 每一个节点本地计算所有路由
  - 在大型网络中会产生较大的处理开销

- 最小化某种意义上的距离
  - 要求策略共享且统一

- 主要应用于AS内部路由选择
  - 例如, OSPF and IS-IS

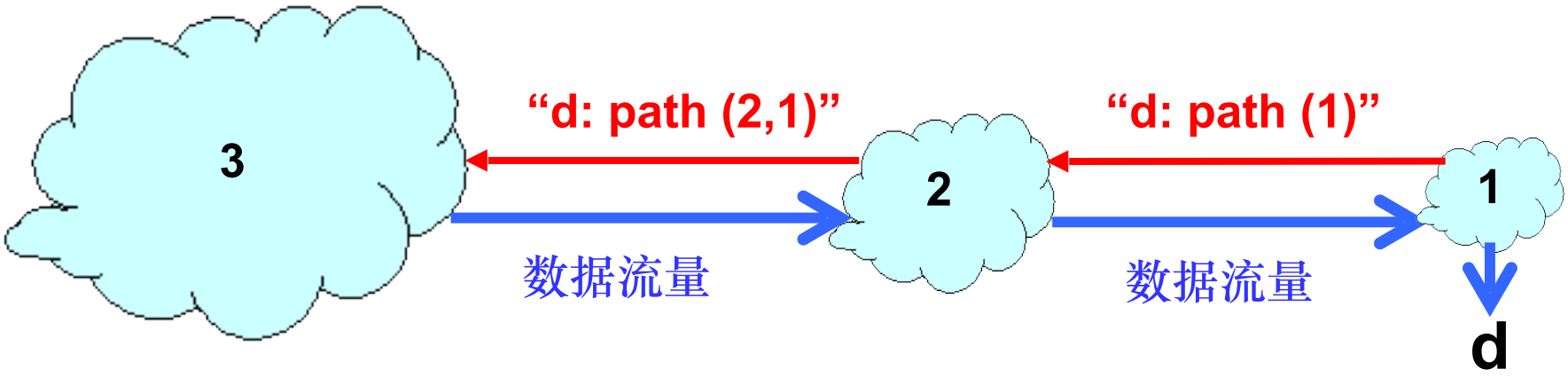
距离向量: 讨论

- 优点
  - 隐藏了网络拓扑的细节
  - 节点仅确定通往目的地的 “下一跳”
- 缺点
  - 最小化某种意义上的距离, 这在域间设置上非常困难
  - 无穷计算问题导致的收敛慢 (“坏消息传递慢”)
- 想法: 对距离向量进行扩展
  - 使其能够快速检测环路

路径向量路由选择

- 距离向量路由选择的扩展
  - 支持灵活的路由策略
  - 避免无穷计算问题

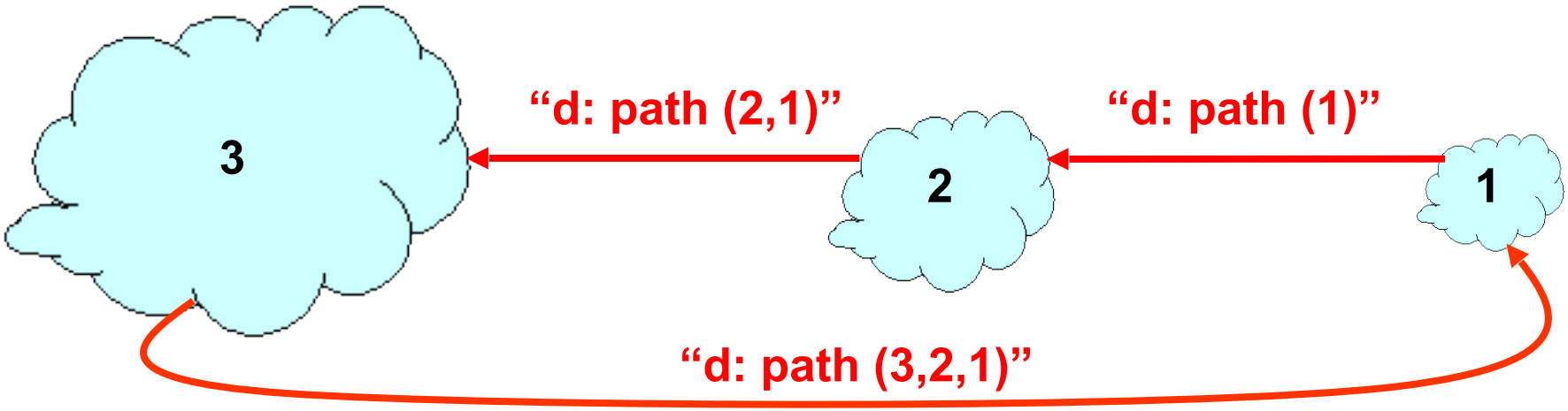
- 核心思想: 通告整个路径
  - 距离向量: 发送到每一个目的d的距离向量
  - 路径向量: 发送到每一个目的d的路径向量



快速环路检测

- 节点可以很容的检测环路路径
  - 在路径中查询自己的节点标识
  - 例如, 节点1发现自己的标识存在于路径“3, 2, 1”中

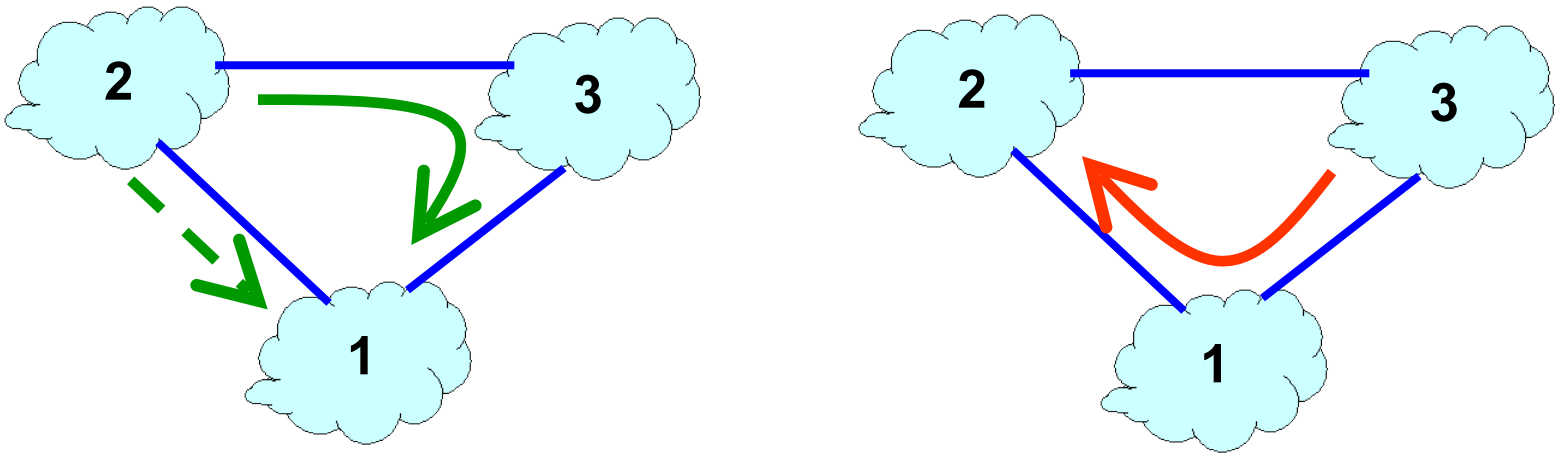
- 节点丢弃环路路径
  - 例如, 节点1丢弃该通告



灵活的策略

- 每一个节点可以采取本地策略
  - 路径选择: 采用哪一条路径?
  - 路径通告: 通告哪一条路径?

- 例如
  - 节点2更倾向于选择“2, 3, 1”而非“2, 1”
  - 节点 1不允许节点知道路径“1, 2”的存在



域间路由选择(BGP)

- Internet和自制系统
- 域间路由选择
- 路径向量路由选择
- BGP

边界网关协议

- Internet的域间路由选择协议
  - 基于前缀的路径向量路由选择协议
  - 基于策略进行路由选择构建AS路径
  - 过去的18年不断改进

- 1989 : BGP-1 [RFC 1105], 替代了最早的 EGP
- 1990 : BGP-2 [RFC 1163]
- 1991 : BGP-3 [RFC 1267]
- 1995 : BGP-4 [RFC 1771], 支持 CIDR
- 2006 : BGP-4 [RFC 4271], 修正

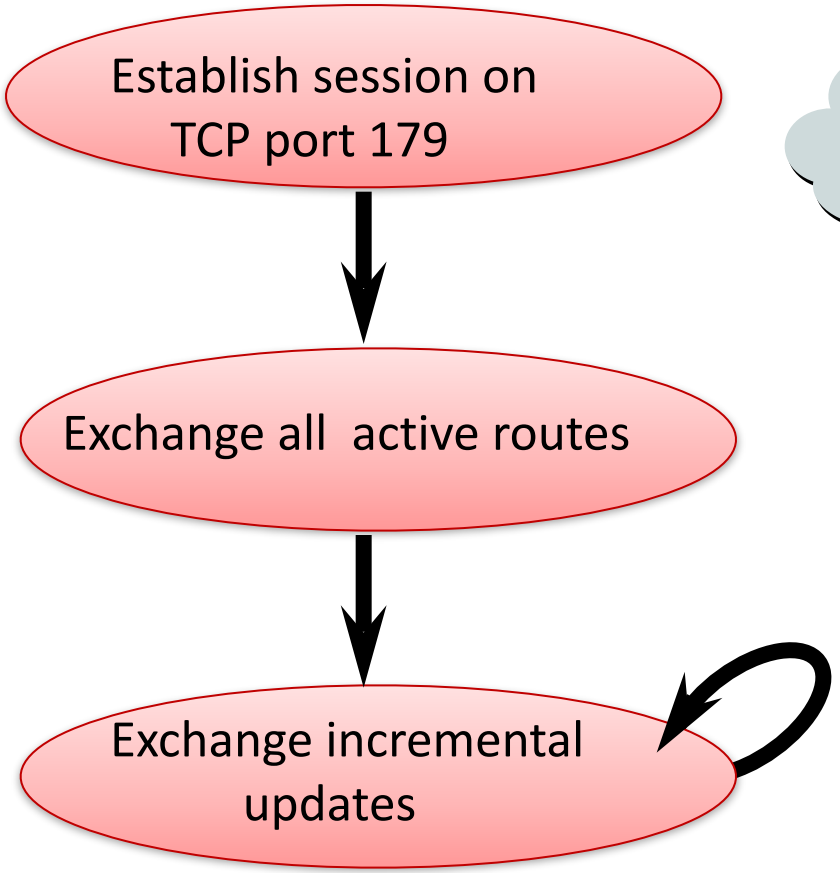
BGP的特点

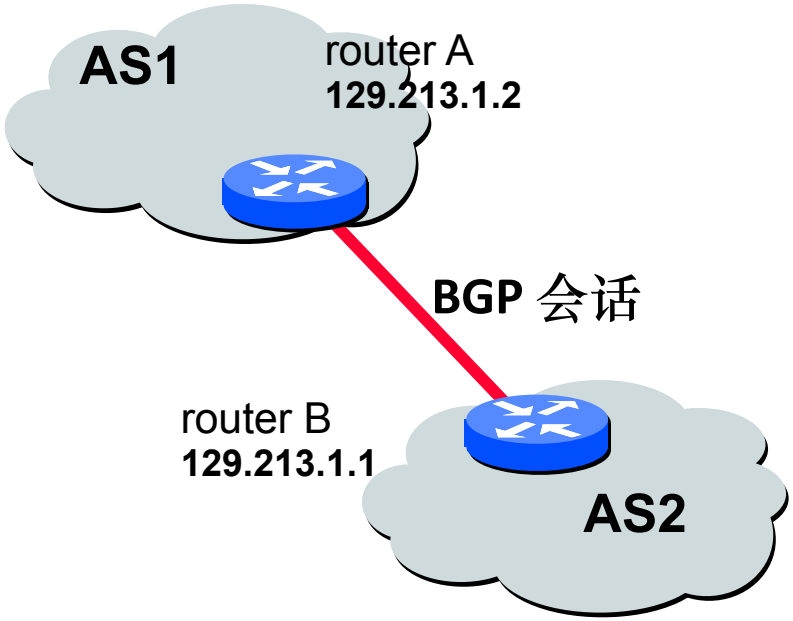
- 允许Ases向其他 ASes"路由" that they are "responsible" for and how to reach them
  - BGP-代言人 之间进行通信
  - 采用"路由通告",或"promises" – 也称为"NLRI"或"网络层可达信息"
  - 路径向量路由选择协议

- 基于策略: 允许ISPs表达其路由策略, 包括both in selecting outbound paths and in announcing internal routes

- 非常"简单"的协议, 但是配置相当复杂

BGP Operations





While connection is ALIVE, exchange route UPDATE messages

增量协议

- 节点知道多条到达目的地的路径
  - 在路由表中存储所有的路由
  - 采用策略选择一条最好的路由

- 增量更新
  - 通告
    - 一旦选择一条新的路由, 则将节点id加入路径向量
    - … 并(有选择性的)通告其他邻居
  - 撤销
    - 如果路由不再有效
    - … 发送撤销路由消息

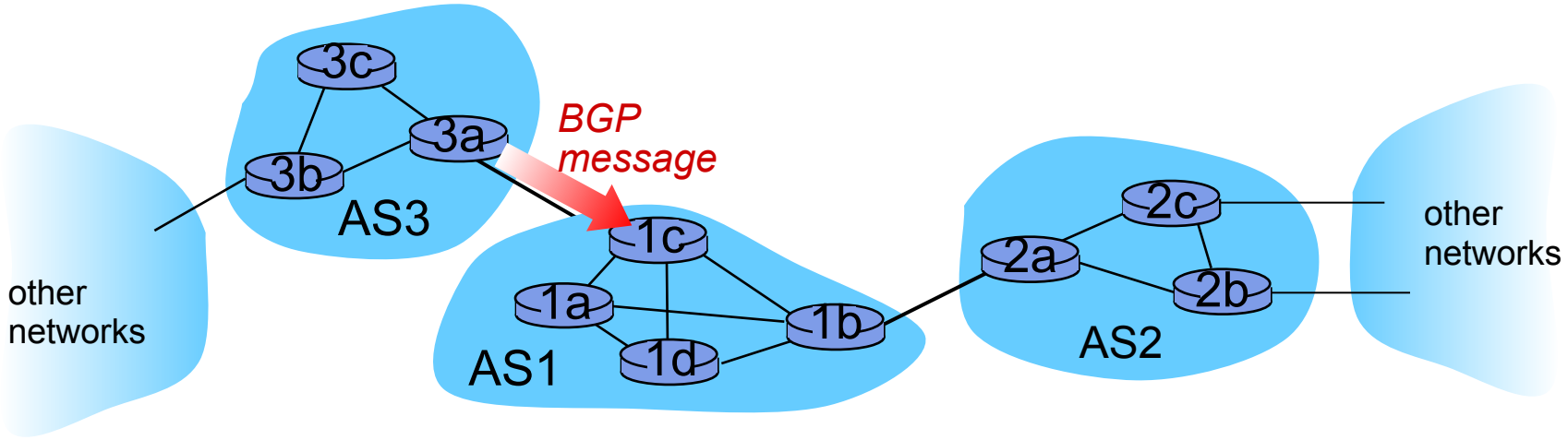
Internet inter-AS routing: BGP

- BGP (Border Gateway Protocol): *the* de facto inter-domain routing protocol
  - "glue that holds the Internet together"
- BGP provides each AS a means to:
  - eBGP: obtain subnet reachability information from neighboring ASs.
  - iBGP: propagate reachability information to all AS-internal routers.
  - determine "good" routes to other networks based on reachability information and policy.
- allows subnet to advertise its existence to rest of Internet: *"I am here"*

BGP basics

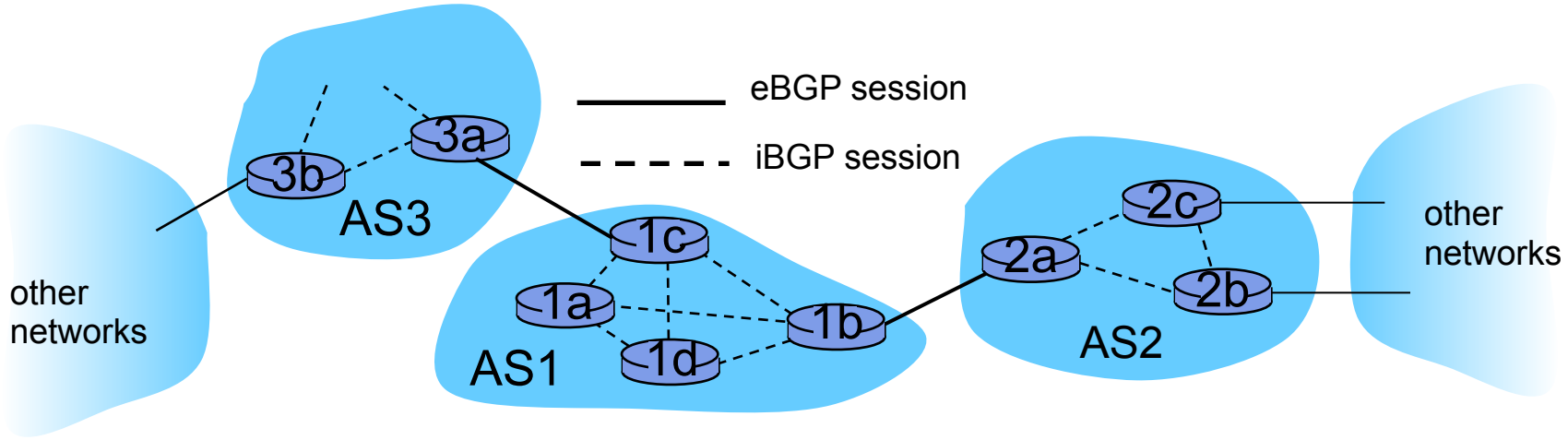
- BGP session: two BGP routers ("peers") exchange BGP messages:
  - advertising *paths* to different destination network prefixes ("path vector" protocol)
  - exchanged over semi-permanent TCP connections

- when AS3 advertises a prefix to AS1:
  - AS3 *promises* it will forward datagrams towards that prefix
  - AS3 can aggregate prefixes in its advertisement



BGP basics: distributing path information

- using eBGP session between 3a and 1c, AS3 sends prefix reachability info to AS1.
  - 1c can then use iBGP do distribute new prefix info to all routers in AS1
  - 1b can then re-advertise new reachability info to AS2 over 1b-to-2a eBGP session
- when router learns of new prefix, it creates entry for prefix in its forwarding table.



Path attributes and BGP routes

- advertised prefix includes BGP attributes
  - prefix + attributes = “route”
- two important attributes:
  - AS-PATH: contains ASs through which prefix advertisement has passed: e.g., AS 67, AS 17
  - NEXT-HOP: indicates specific internal-AS router to next-hop AS. (may be multiple links from current AS to next-hop-AS)
- gateway router receiving route advertisement uses import policy to accept/decline
  - e.g., never route through AS x
  - *policy-based* routing

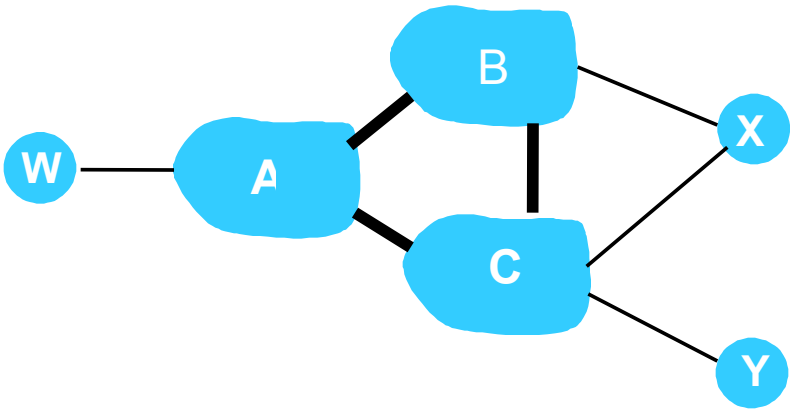
BGP route selection

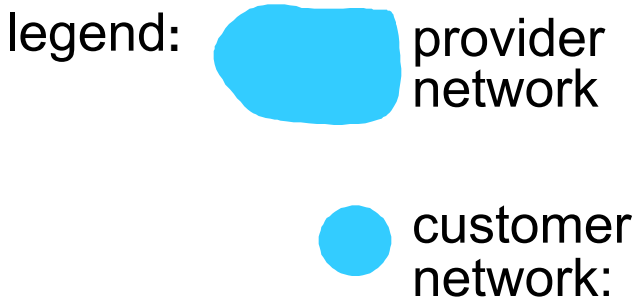
- router may learn about more than 1 route to destination AS, selects route based on:
  1. local preference value attribute: policy decision
  2. shortest AS-PATH
  3. closest NEXT-HOP router: hot potato routing
  4. additional criteria

BGP messages

- BGP messages exchanged between peers over TCP connection
- BGP messages:
  - OPEN: opens TCP connection to peer and authenticates sender
  - UPDATE: advertises new path (or withdraws old)
  - KEEPALIVE: keeps connection alive in absence of UPDATES; also ACKs OPEN request
  - NOTIFICATION: reports errors in previous msg; also used to close connection

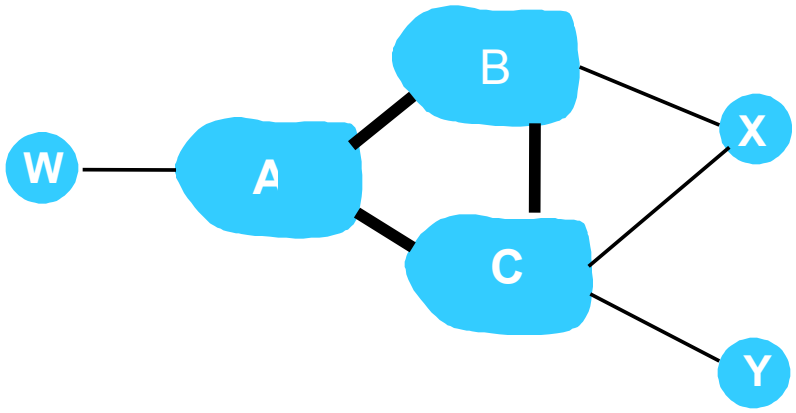
BGP routing policy

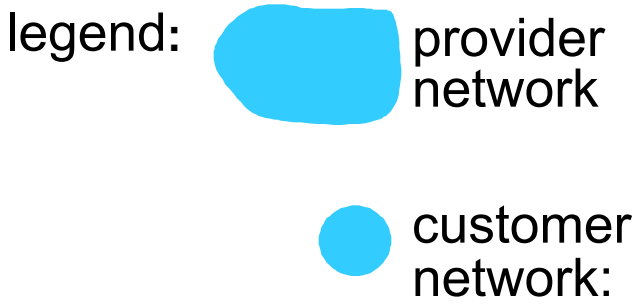




- A,B,C are *provider networks*
- X,W,Y are customer (of provider networks)
- X is *dual-homed:* attached to two networks
  - X does not want to route from B via X to C
  - .. so X will not advertise to B a route to C

BGP routing policy (2)





- A advertises path AW to B
- B advertises path BAW to X
- Should B advertise path BAW to C?
  - No way! B gets no "revenue" for routing CBAW since neither W nor C are B's customers
  - B wants to force C to route to w via A
  - B wants to route *only* to/from its customers!

Why different Intra-, Inter-AS routing ?

policy:

- inter-AS: admin wants control over how its traffic routed, who routes through its net.
- intra-AS: single admin, so no policy decisions needed

scale:

- hierarchical routing saves table size, reduced update traffic

performance:

- intra-AS: can focus on performance
- inter-AS: policy may dominate over performance

小结

- 路径向量路由选择协议
  - 快速收敛性(与距离向量路由选择协议)
  - 信息隐藏
  - 支持灵活的策略

- 域间路由
  - 自治系统(ASes)
  - 基于策略的路径向量路由选择协议

提纲

- 引言
  - 核心问题: 扩展到数十亿节点
- 全球互联网
- IPv6
- 多播
- 移动设备之间的路由
- 总结

IPv6: motivation

- *initial motivation:* 32-bit address space soon to be completely allocated.
- additional motivation:
  - header format helps speed processing/forwarding
  - header changes to facilitate QoS

IPv6 datagram format:

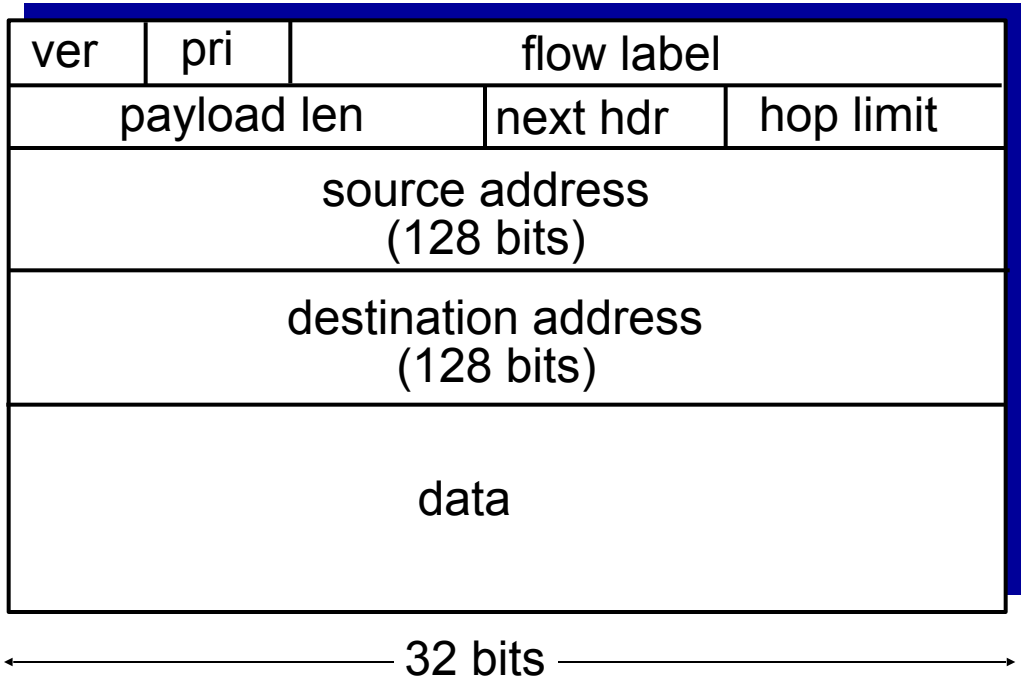
- fixed-length 40 byte header
- no fragmentation allowed

IPv6 datagram format

priority: identify priority among datagrams in flow

flow Label: identify datagrams in same "flow."
(concept of"flow" not well defined).

next header: identify upper layer protocol for data

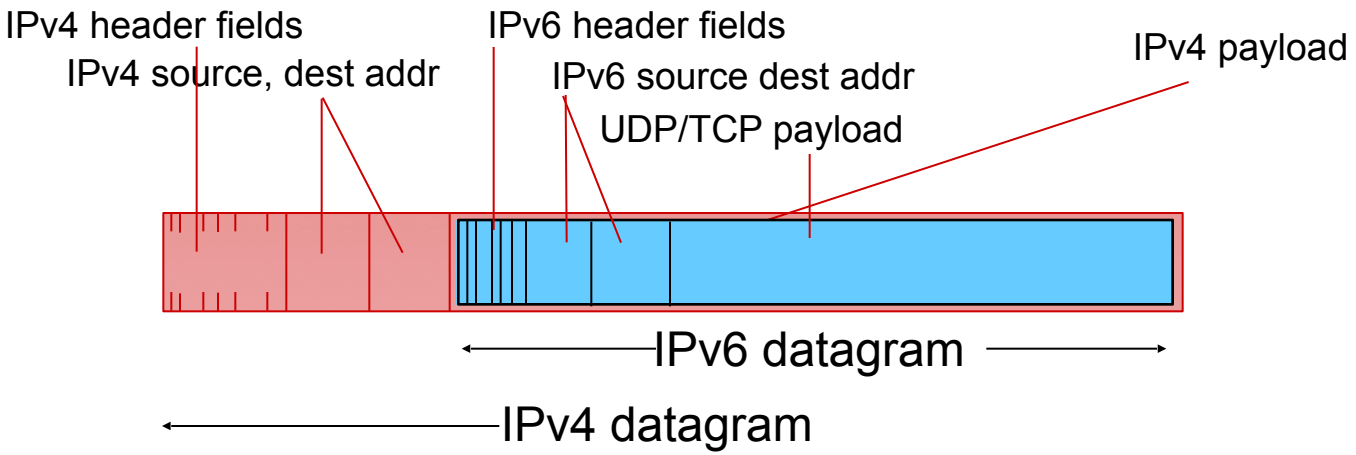


Other changes from IPv4

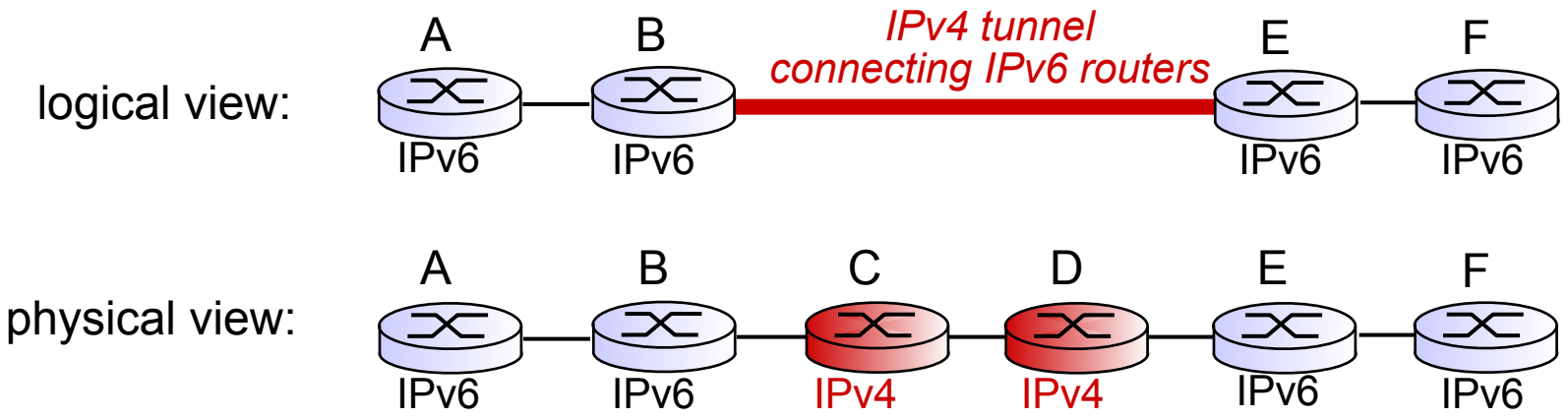
- *checksum*: removed entirely to reduce processing time at each hop
- *options:* allowed, but outside of header, indicated by “Next Header” field
- *ICMPv6:* new version of ICMP
  - additional message types, e.g. “Packet Too Big”
  - multicast group management functions

Transition from IPv4 to IPv6

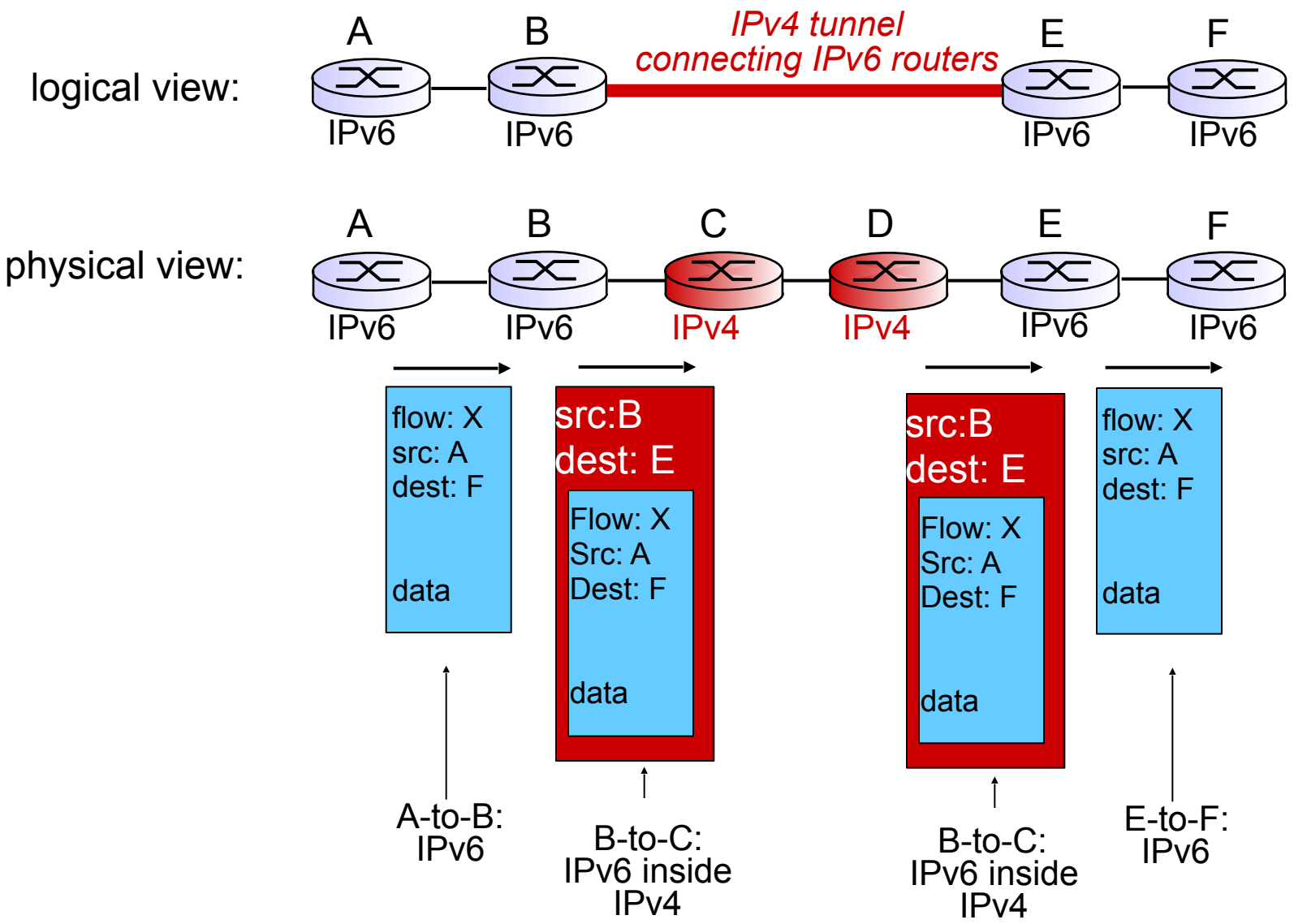
- not all routers can be upgraded simultaneously
  - no “flag days”
  - how will network operate with mixed IPv4 and IPv6 routers?
- *tunneling:* IPv6 datagram carried as *payload* in IPv4 datagram among IPv4 routers



Tunneling


logical view:
A
B
IPv4 tunnel
connecting IPv6 routers
E
F
IPv6
IPv6
IPv6
IPv6
physical view:
A
B
C
D
E
F
IPv6
IPv6
IPv4
IPv4
IPv6
IPv6

Tunneling

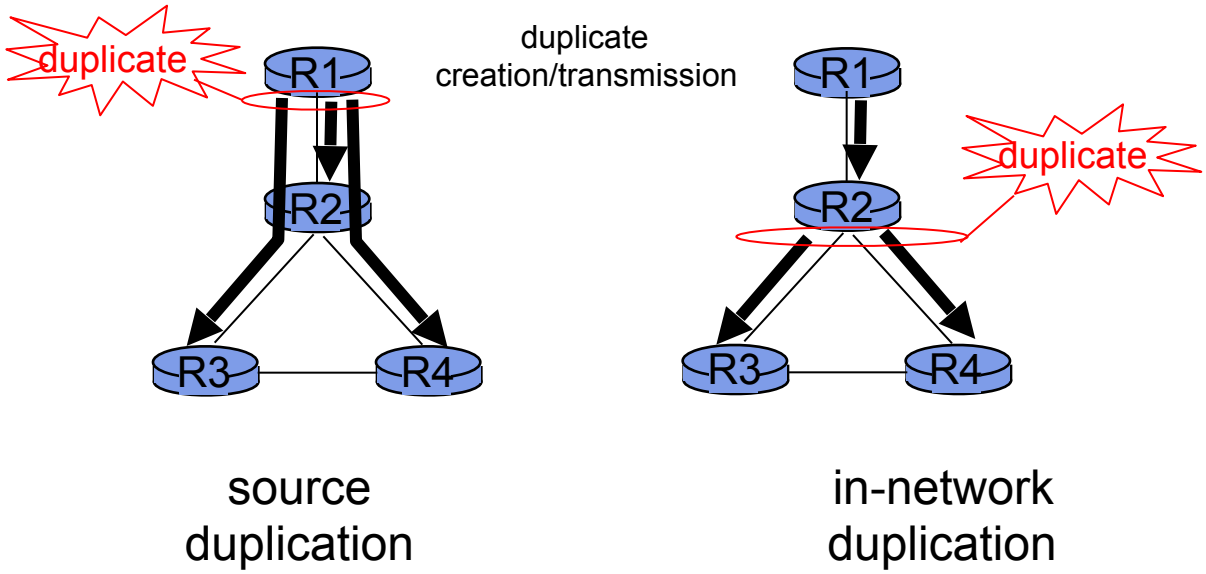

logical view:
A
B
IPv4 tunnel
connecting IPv6 routers
E
F
IPv6
IPv6
IPv6
IPv6
physical view:
A
B
C
D
E
F
IPv6
IPv6
IPv4
IPv4
IPv6
IPv6
flow: X
src: A
dest: F
data
src:B
dest: E
Flow: X
Src: A
Dest: F
data
src:B
dest: E
Flow: X
Src: A
Dest: F
data
flow: X
src: A
dest: F
data
A-to-B:
IPv6
B-to-C:
IPv6 inside
IPv4
B-to-C:
IPv6 inside
IPv4
E-to-F:
IPv6

提纲

- 引言
  - 核心问题: 扩展到数十亿节点
- 全球互联网
- IPv6
- 多播
- 移动设备之间的路由
- 总结

Broadcast routing

- deliver packets from source to all other nodes
- source duplication is inefficient:



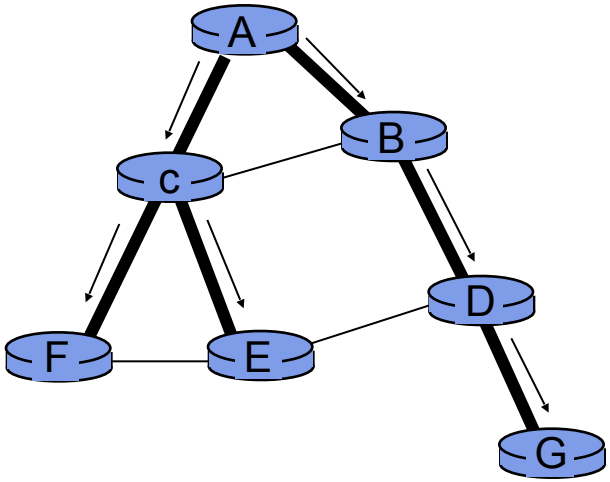
- source duplication: how does source determine recipient addresses?

In-network duplication

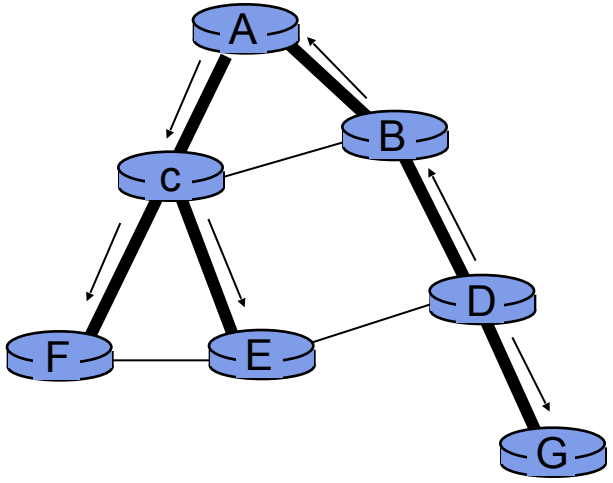
- *flooding:* when node receives broadcast packet, sends copy to all neighbors
  - problems: cycles & broadcast storm
- *controlled flooding:* node only broadcasts pkt if it hasn't broadcast same packet before
  - node keeps track of packet ids already broadacsted
  - or reverse path forwarding (RPF): only forward packet if it arrived on shortest path between node and source
- *spanning tree:*
  - no redundant packets received by any node

Spanning tree

- first construct a spanning tree
- nodes then forward/make copies only along spanning tree



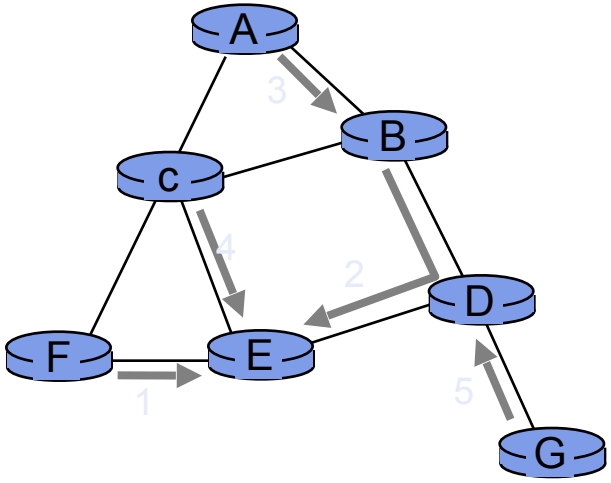
(a) broadcast initiated at A



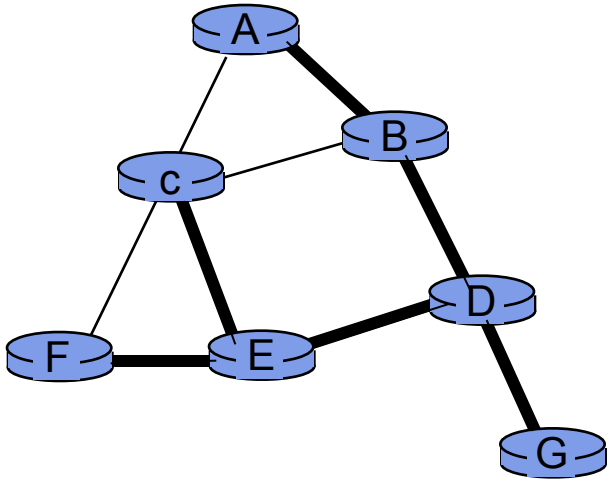
(b) broadcast initiated at D

Spanning tree: creation

- center node
- each node sends unicast join message to center node
  - message forwarded until it arrives at a node already belonging to spanning tree



(a) stepwise construction of spanning tree (center: E)

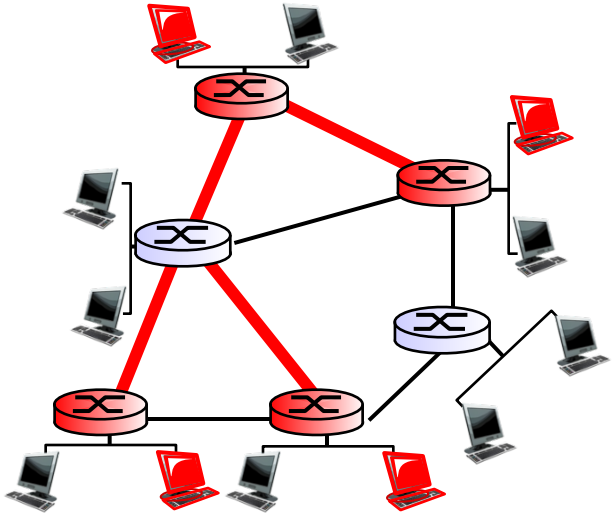


(b) constructed spanning tree

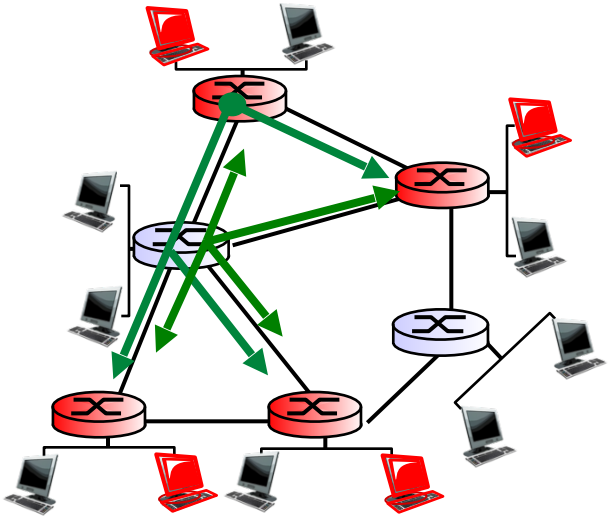
Multicast routing: problem statement

goal: find a tree (or trees) connecting routers having local mcast group members

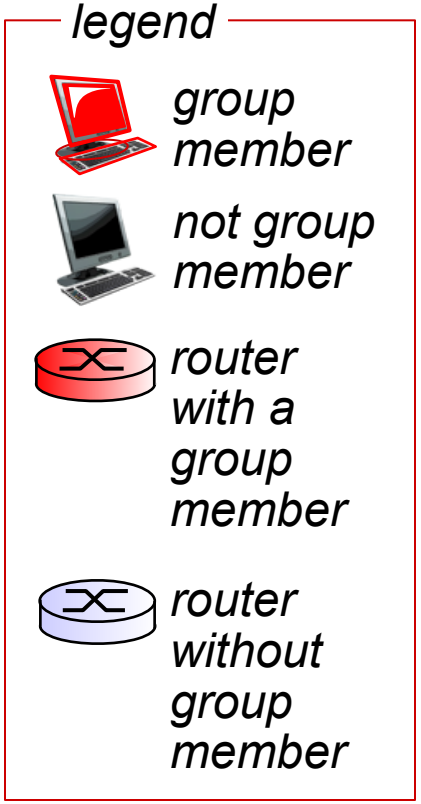
- *tree:* not all paths between routers used
- *shared-tree:* same tree used by all group members
- *source-based:* different tree from each sender to rcvrs



shared tree



source-based trees



Approaches for building mcast trees

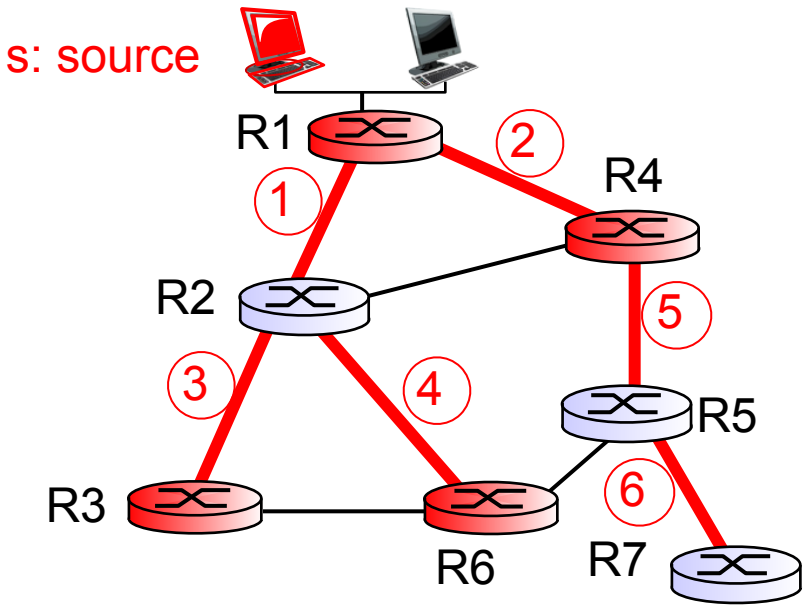
approaches:

- *source-based tree:* one tree per source
  - shortest path trees
  - reverse path forwarding
- *group-shared tree:* group uses one tree
  - minimal spanning (Steiner)
  - center-based trees

…we first look at basic approaches, then specific protocols adopting these approaches

Shortest path tree

- mcast forwarding tree: tree of shortest path routes from source to all receivers
  - Dijkstra's algorithm



LEGEND

router with attached group member

router with no attached group member

i — link used for forwarding, i indicates order link added by algorithm

Reverse path forwarding

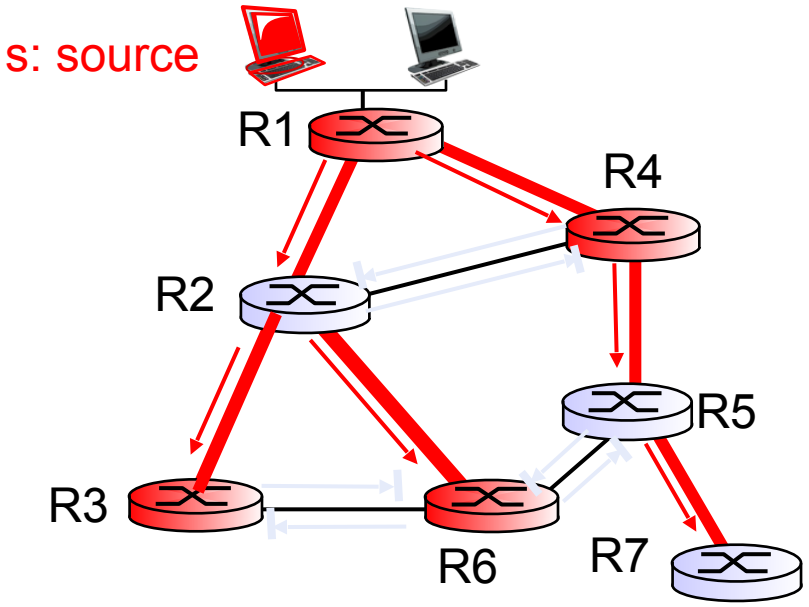
- rely on router's knowledge of unicast shortest path from it to sender
- each router has simple forwarding behavior:

if (mcast datagram received on incoming link on shortest path back to center)

then flood datagram onto all outgoing links

else ignore datagram

Reverse path forwarding: example



LEGEND

router with attached group member

router with no attached group member

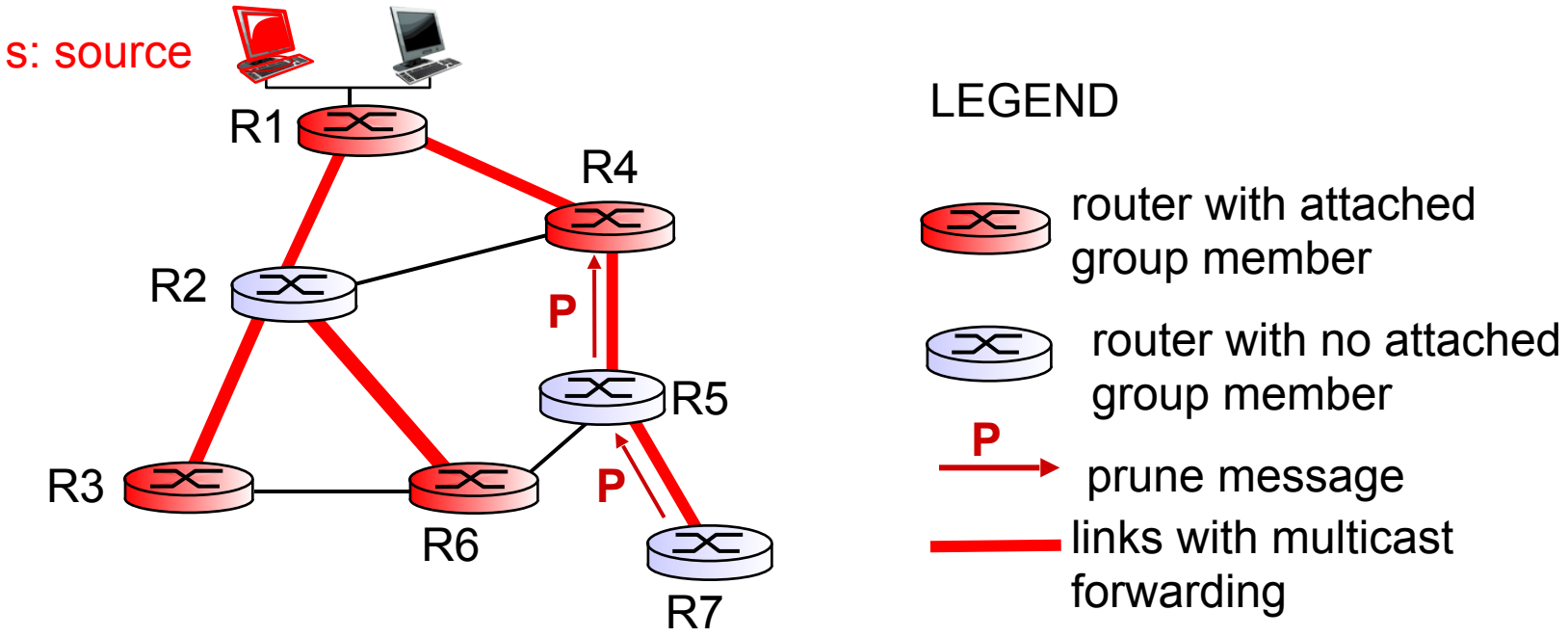
datagram will be forwarded

datagram will not be forwarded

- result is a source-specific *reverse* SPT
  - may be a bad choice with asymmetric links

Reverse path forwarding: pruning

- forwarding tree contains subtrees with no mcast group members
  - no need to forward datagrams down subtree
  - “prune” msgs sent upstream by router with no downstream group members



Shared-tree: steiner tree

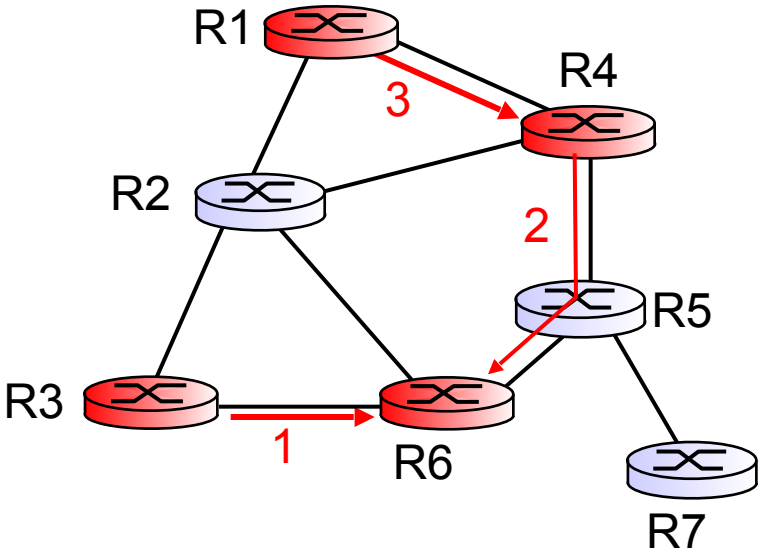
- *steiner tree:* minimum cost tree connecting all routers with attached group members
- problem is NP-complete
- excellent heuristics exists
- not used in practice:
  - computational complexity
  - information about entire network needed
  - monolithic: rerun whenever a router needs to join/ leave

Center-based trees

- single delivery tree shared by all
- one router identified as *"center"* of tree
- to join:
  - edge router sends unicast *join-msg* addressed to center router
  - *join-msg* "processed" by intermediate routers and forwarded towards center
  - *join-msg* either hits existing tree branch for this center, or arrives at center
  - path taken by *join-msg* becomes new branch of tree for this router

Center-based trees: example

suppose R6 chosen as center:



LEGEND

router with attached group member

router with no attached group member

1 path order in which join messages generated

Internet Multicasting Routing: DVMRP

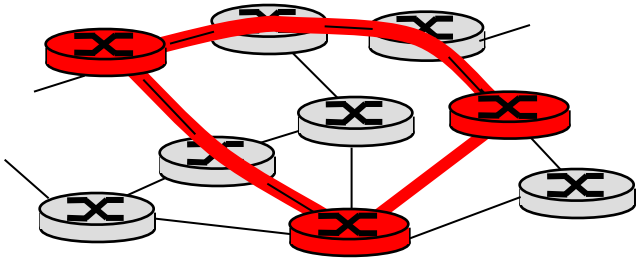
- DVMRP: distance vector multicast routing protocol, RFC1075
- *flood and prune:* reverse path forwarding, source-based tree
  - RPF tree based on DVMRP's own routing tables constructed by communicating DVMRP routers
  - no assumptions about underlying unicast
  - initial datagram to mcast group flooded everywhere via RPF
  - routers not wanting group: send upstream prune msgs

DVMRP: continued…

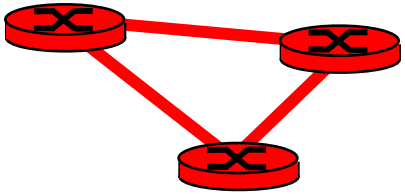
- *soft state:* DVMRP router periodically (1 min.) "forgets" branches are pruned:
  - mcast data again flows down unpruned branch
  - downstream router: reprune or else continue to receive data
- routers can quickly regraft to tree
  - following IGMP join at leaf
- odds and ends
  - commonly implemented in commercial router

Tunneling

Q: how to connect "islands" of multicast routers in a "sea" of unicast routers?



physical topology



logical topology

- mcast datagram encapsulated inside "normal" (non-multicast-addressed) datagram
- normal IP datagram sent thru "tunnel" via regular IP unicast to receiving mcast router (recall IPv6 inside IPv4 tunneling)
- receiving mcast router unencapsulates to get mcast datagram

PIM: Protocol Independent Multicast

- not dependent on any specific underlying unicast routing algorithm (works with all)
- two different multicast distribution scenarios :

dense:

- group members densely packed, in "close" proximity.
- bandwidth more plentiful

sparse:

- # networks with group members small wrt # interconnected networks
- group members "widely dispersed"
- bandwidth not plentiful

Consequences of sparse-dense dichotomy:

dense

- group membership by routers *assumed* until routers explicitly prune
- *data-driven* construction on mcast tree (e.g., RPF)
- bandwidth and non-group-router processing *profligate*

sparse:

- no membership until routers explicitly join
- *receiver- driven* construction of mcast tree (e.g., center-based)
- bandwidth and non-group-router processing *conservative*

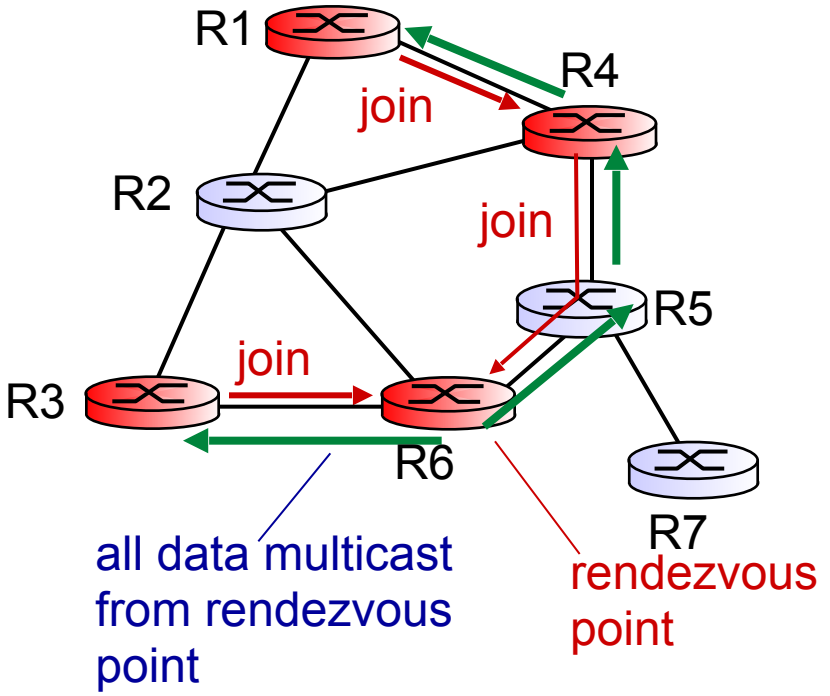
PIM- dense mode

flood-and-prune RPF: similar to DVMRP but…

- underlying unicast protocol provides RPF info for incoming datagram
- less complicated (less efficient) downstream flood than DVMRP reduces reliance on underlying routing algorithm
- has protocol mechanism for router to detect it is a leaf-node router

PIM - sparse mode

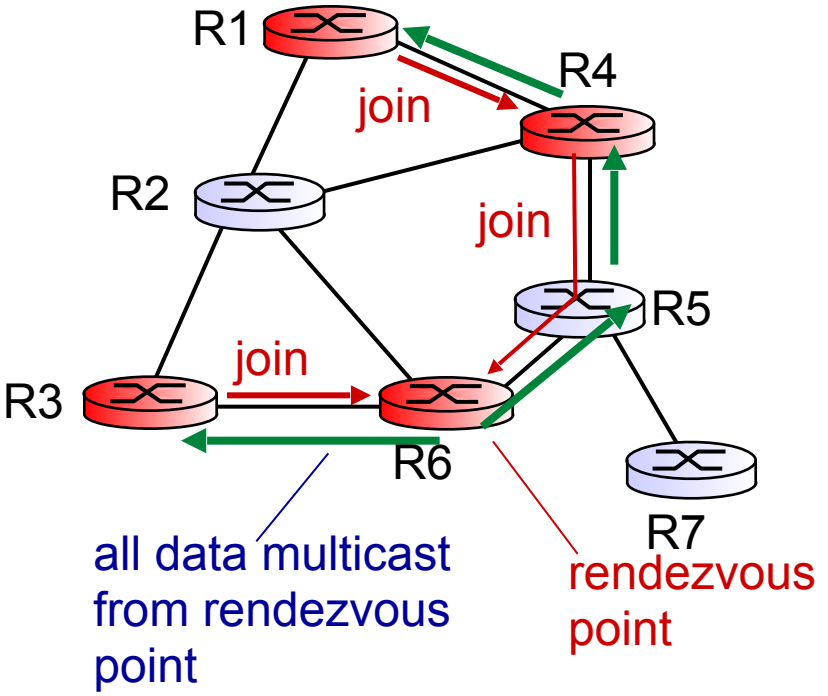
- center-based approach
- router sends *join* msg to rendezvous point (RP)
  - intermediate routers update state and forward *join*
- after joining via RP, router can switch to source-specific tree
  - increased performance: less concentration, shorter paths



PIM - sparse mode

sender(s):

- unicast data to RP, which distributes down RP-rooted tree
- RP can extend mcast tree upstream to source
- RP can send *stop* msg if no attached receivers
  - "no one is listening!"



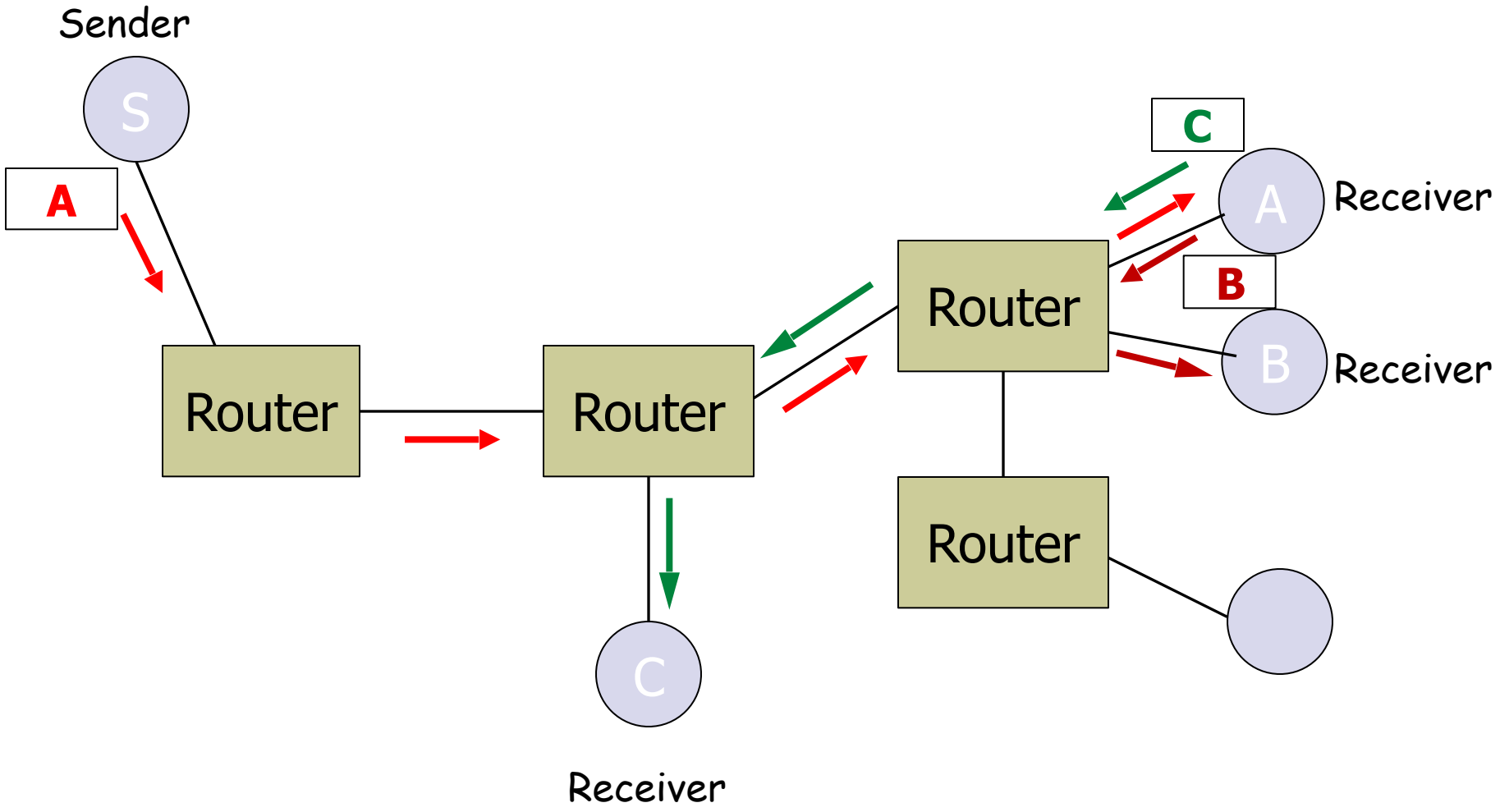
Application Level Multicast

- Provide IP multicast functionality above the IP layer
- Challenge: do this efficiently
- ALM is to have applications self-organized into a logical overlay network, and transfer data along the edges of the overlay network using unicast transport services.
  - Each application communicates only with its neighbors in the overlay network.
  - Multicasting is implemented by forwarding messages along trees that are embedded in the virtual overlay network.

Pros and Cons of ALM

- Pros:
  - No requirement for multicast support in the network layer
  - No need to allocate a global group identifier, such as an IP multicast address
  - Unicast traffic engineering techniques can be applied, such as flow control, congestion control, and reliable delivery services
- Cons:
  - End-to-end latencies can be high
  - Possible inefficient use of bandwidth

An Illustration of ALM

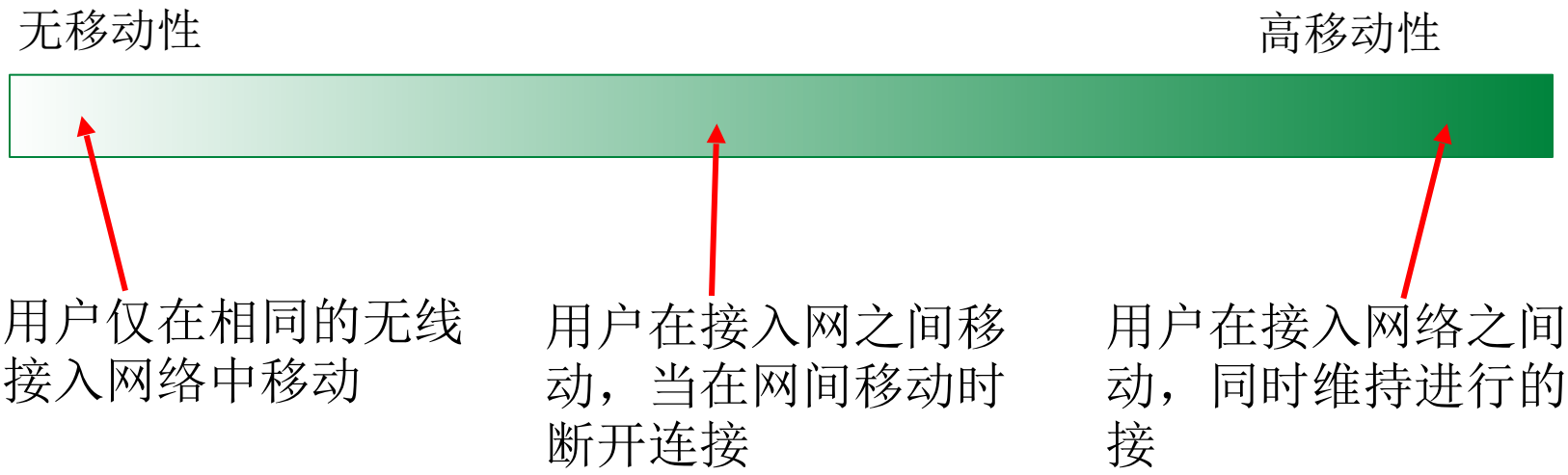


提纲

- 引言
  - 核心问题: 扩展到数十亿节点
- 全球互联网
- IPv6
- 多播
- 移动设备之间的路由
- 总结

什么是移动性?

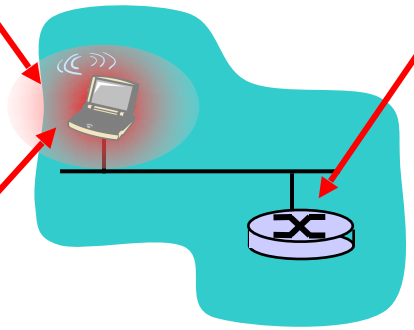
- 从网络层观点说明用户移动性程度谱:

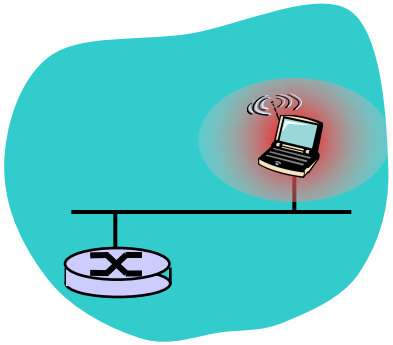


移动性：术语

归属网络：移动节点固定的"居所"
(如：128.119.40/24)

归属代理：执行移动管理功能的实体





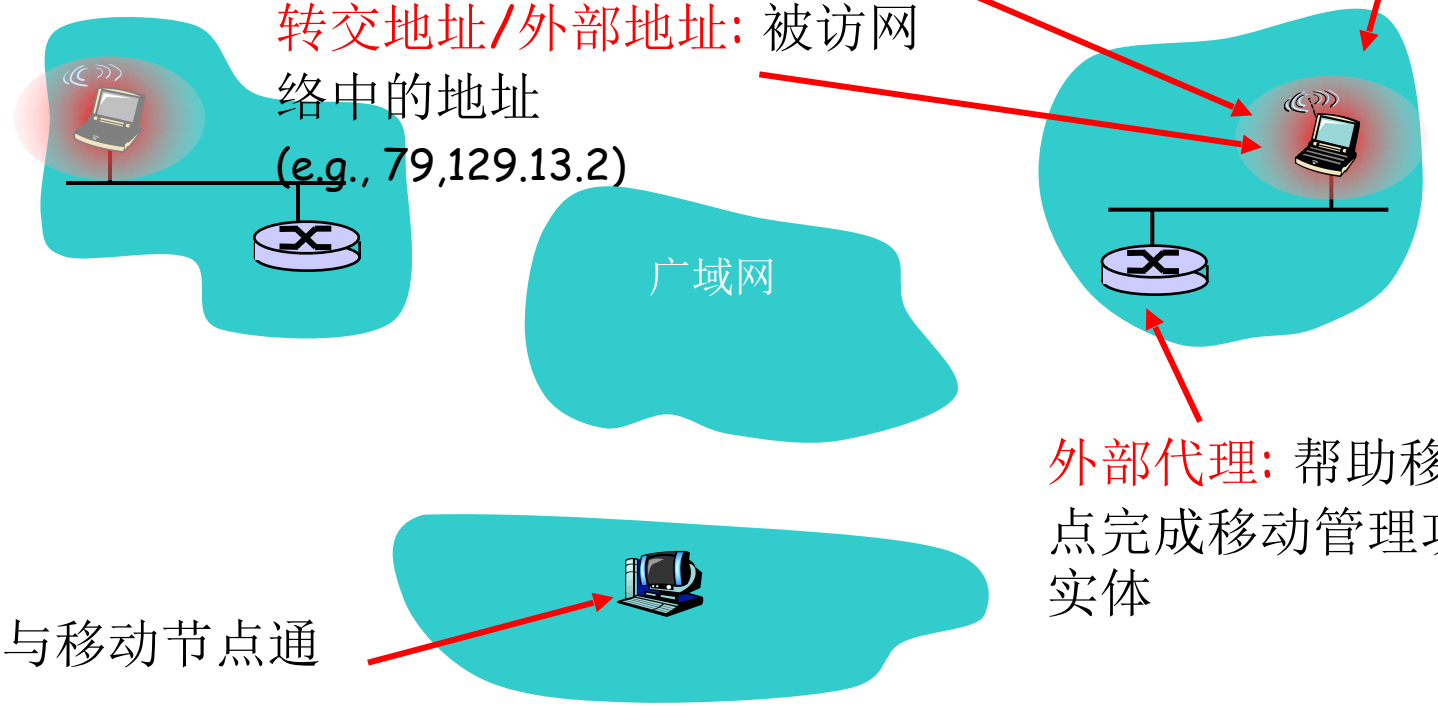
永久地址：归属网络中的地址，经常用来联系移动节点
e.g., 128.119.40.186

移动性：术语

永久地址：保持不变（e.g., 128.119.40.186）

被访网络：移动节点当前所在网络（e.g., 79.129.13/24）

转交地址/外部地址：被访网络中的地址（e.g., 79,129.13.2）



外部代理：帮助移动节点完成移动管理功能的实体

通信者：与移动节点通信

你如何与一个移动朋友联系？

考虑朋友频繁的改变地址，你如何找到她？

- 查找所有的电话册？
- 打电话给她的父母？
- 希望她让你知道她在哪里？

移动性: 方法

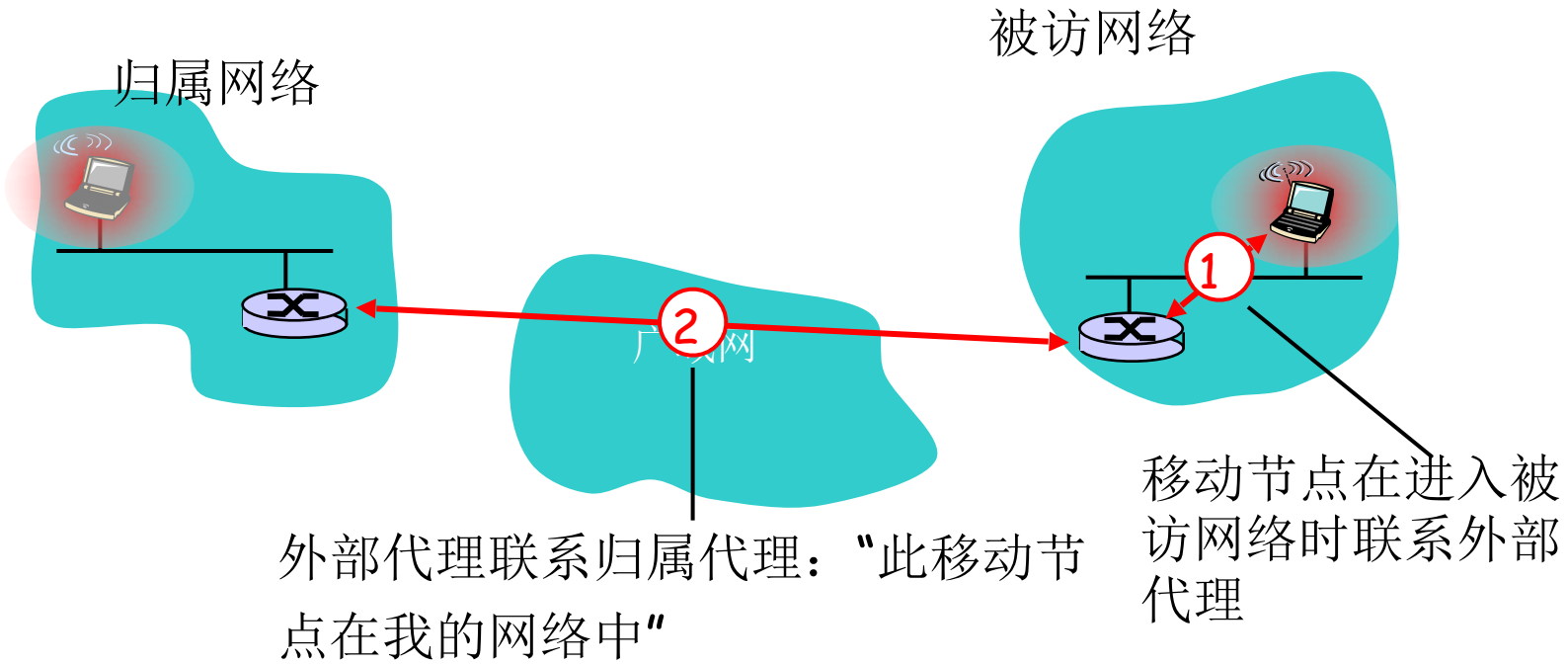
- 让路由处理: 路由器通过路由表交换，通知在它网络中的移动节点的永久地址
  - 路由表指出了每个移动节点的位置
  - 不用对终端系统作改动
- 让终端系统处理:
  - 间接选路: 从通信者到移动节点的通信，通过归属代理，然后被转发到远端
  - 直接选路: 通信者获得移动节点的外部地址/转交地址, 直接发送给移动节点

移动性: 方法

- 让路由处理: 路由[illegible]由表交换，通知在它网络中的移动节点的[illegible]
  - 路由表指出了每[illegible]位置
  - 不用对终端系统作[illegible]

- 让终端系统处理:
  - 间接选路: 从通信者到移动节点的通信，通过归属代理，然后被转发到远端
  - 直接选路: 通信者获得移动节点的外部地址/转交地址, 直接发送给移动节点

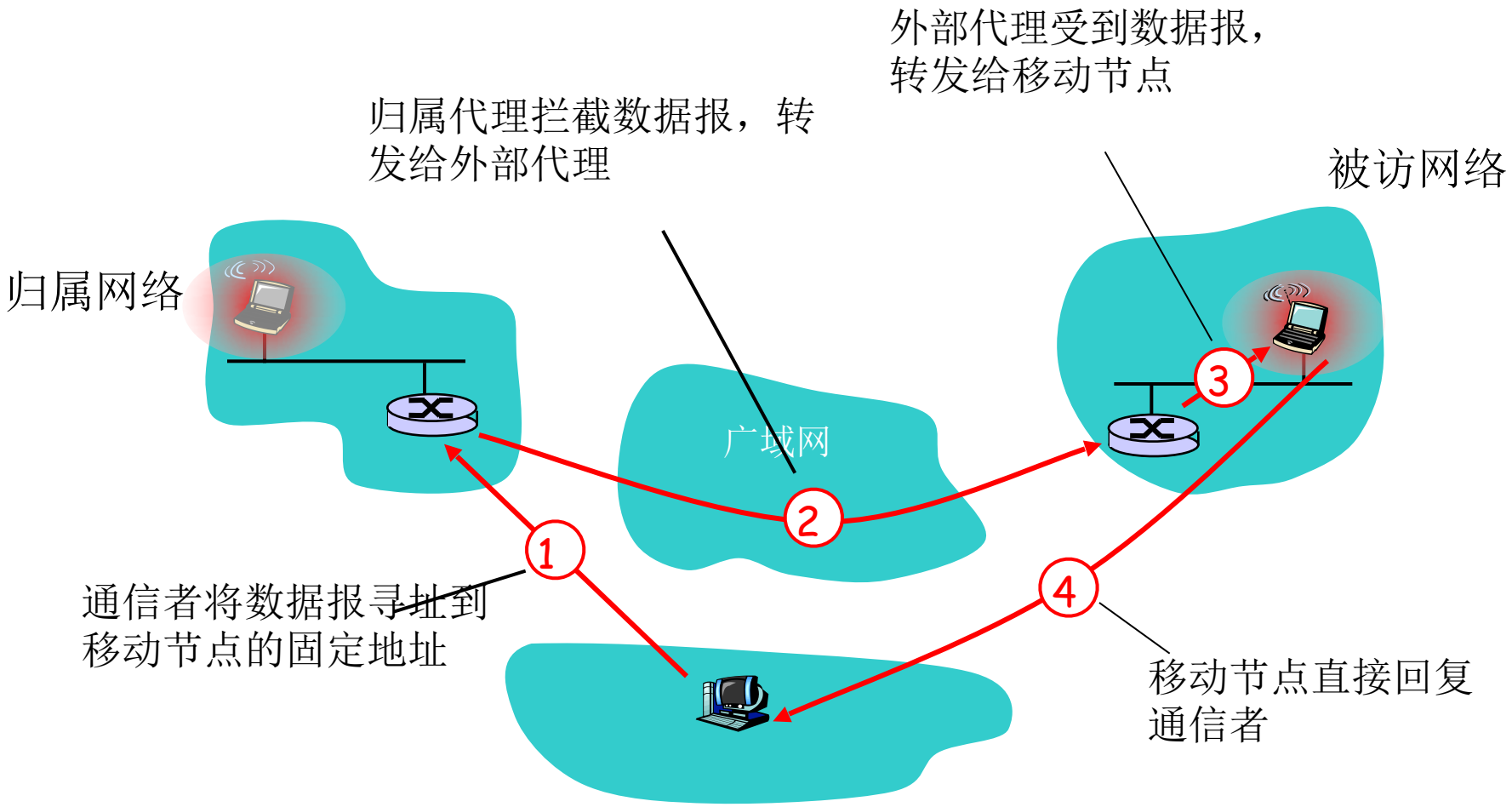
移动性：注册



最终结果：

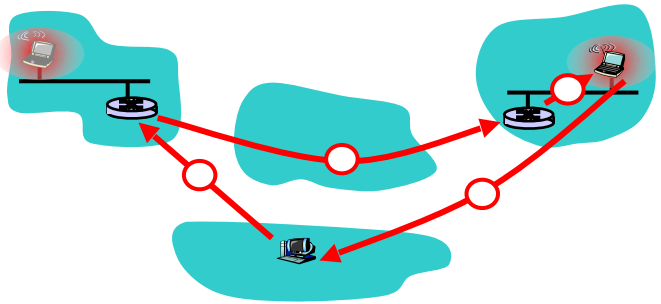
- 外部代理知道了移动节点
- 归属代理知道了移动节点的位置

间接选路


外部代理受到数据报，
转发给移动节点
归属代理拦截数据报，转
发给外部代理
被访网络
归属网络
广域网
1
2
3
4
通信者将数据报寻址到
移动节点的固定地址
移动节点直接回复
通信者

间接选路：评论

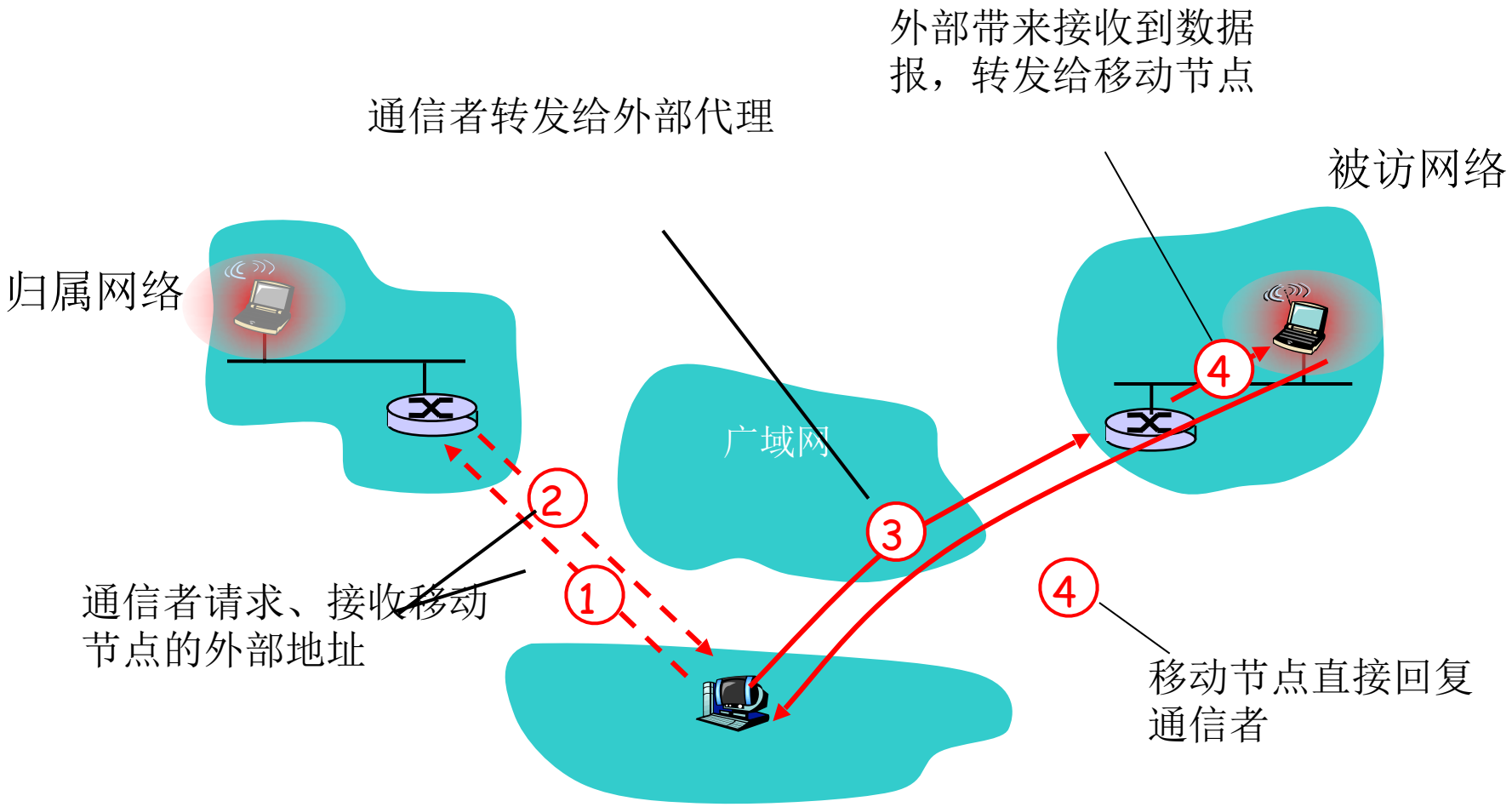
- 移动节点使用两个地址：
  - 永久地址：被通信者使用（因此移动节点的位置对通信者是透明的）
  - 转交地址：被归属代理用来转发数据报给移动节点
- 外部代理的功能可能由移动节点自己完成
- 三角路由：通信者-归属网络-移动节点
  - 当通信者、移动节点在同一网络时，效率很低



间接选路: 在网络间移动

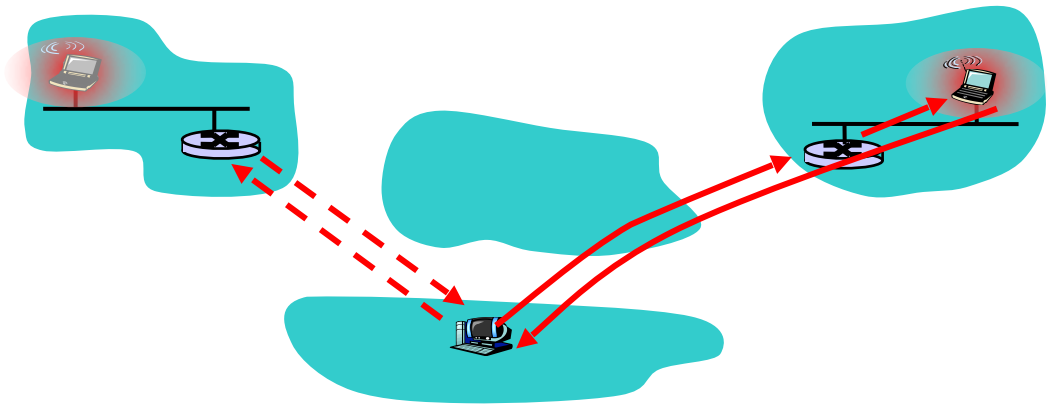
- 假设移动用户移动到另一个网络
  - 向新的外部代理注册
  - 新的外部代理向归属代理注册
  - 归属代理更新移动节点的转交地址
  - 数据报继续被转发到移动节点 (但是通过新的转交地址)
- 移动性透明的改变外部: 可维护持续的连接!

直接选路


外部带来接收到数据报，转发给移动节点
通信者转发给外部代理
被访网络
归属网络
广域网
2
3
4
1
4
通信者请求、接收移动节点的外部地址
移动节点直接回复通信者

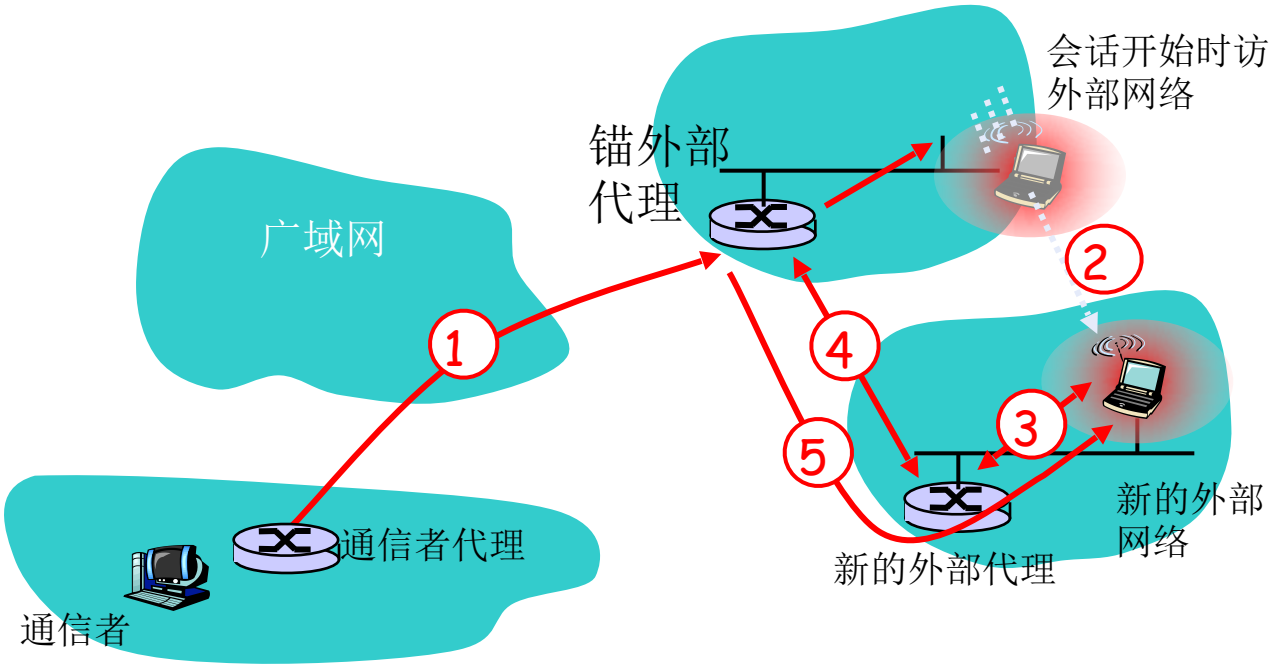
直接选路: 评价

- 克服了三角路由的问题
- 对通信者非透明: 通信者必须从归属代理获得转交地址
  - 当移动节点改变被访网络时会出现什么?



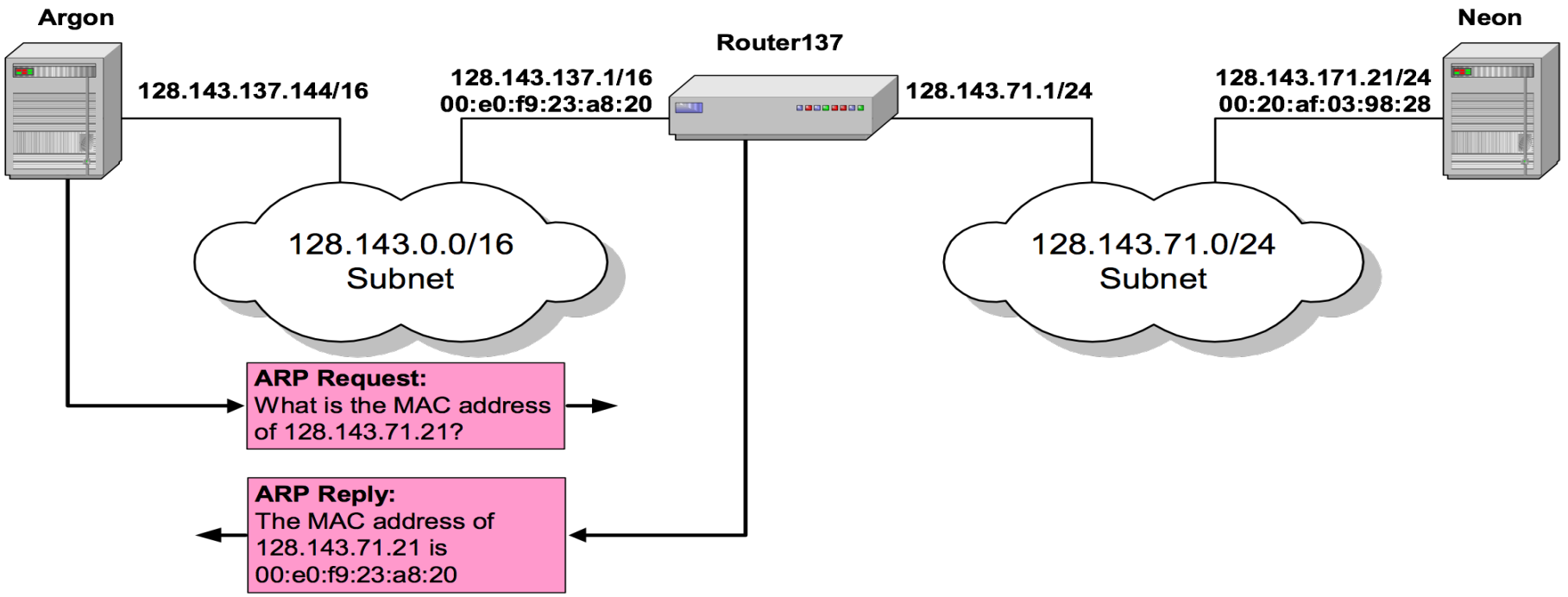
在直接选路下的移动性

- 锚外部代理: 第一个被访网络中的外部代理
- 数据总是首先被发送到锚外部代理
- 当移动节点移动时: 新的外部代理接收从旧的外部代理转发的数据



Implementation

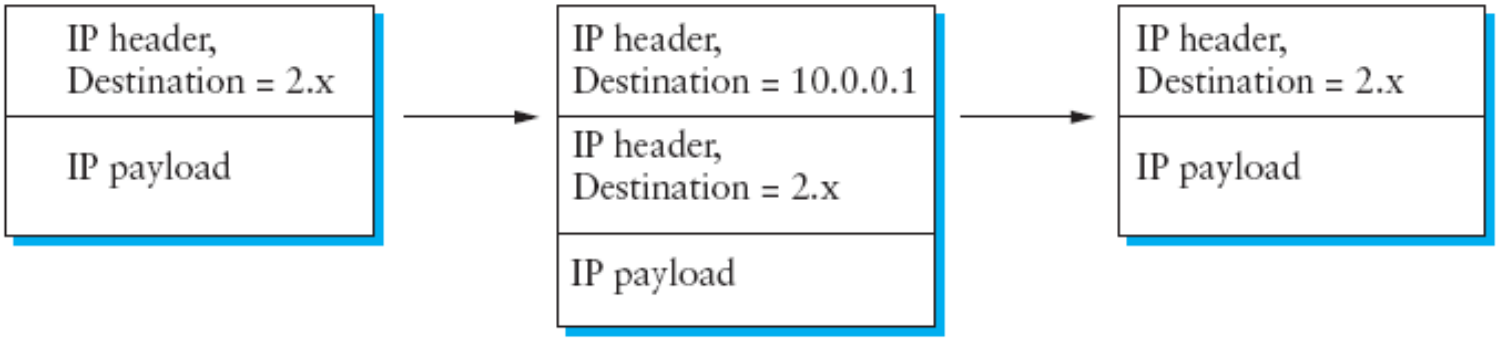
- **How can home agent get the packets for correspondent?**
  - Proxy ARP: Host or router responds to ARP Request that arrives from one of its connected networks for a host that is on another of its connected networks.



Implementation

- **How can home agent transmit packets to foreign agent?**
  - IP tunnel: IP tunnel is a virtual point-to-point link between a pair of nodes that are actually separated by an arbitrary number of networks.

NetworkNum	NextHop
1	Interface 0
2	Virtual interface 0
Default	Interface 1



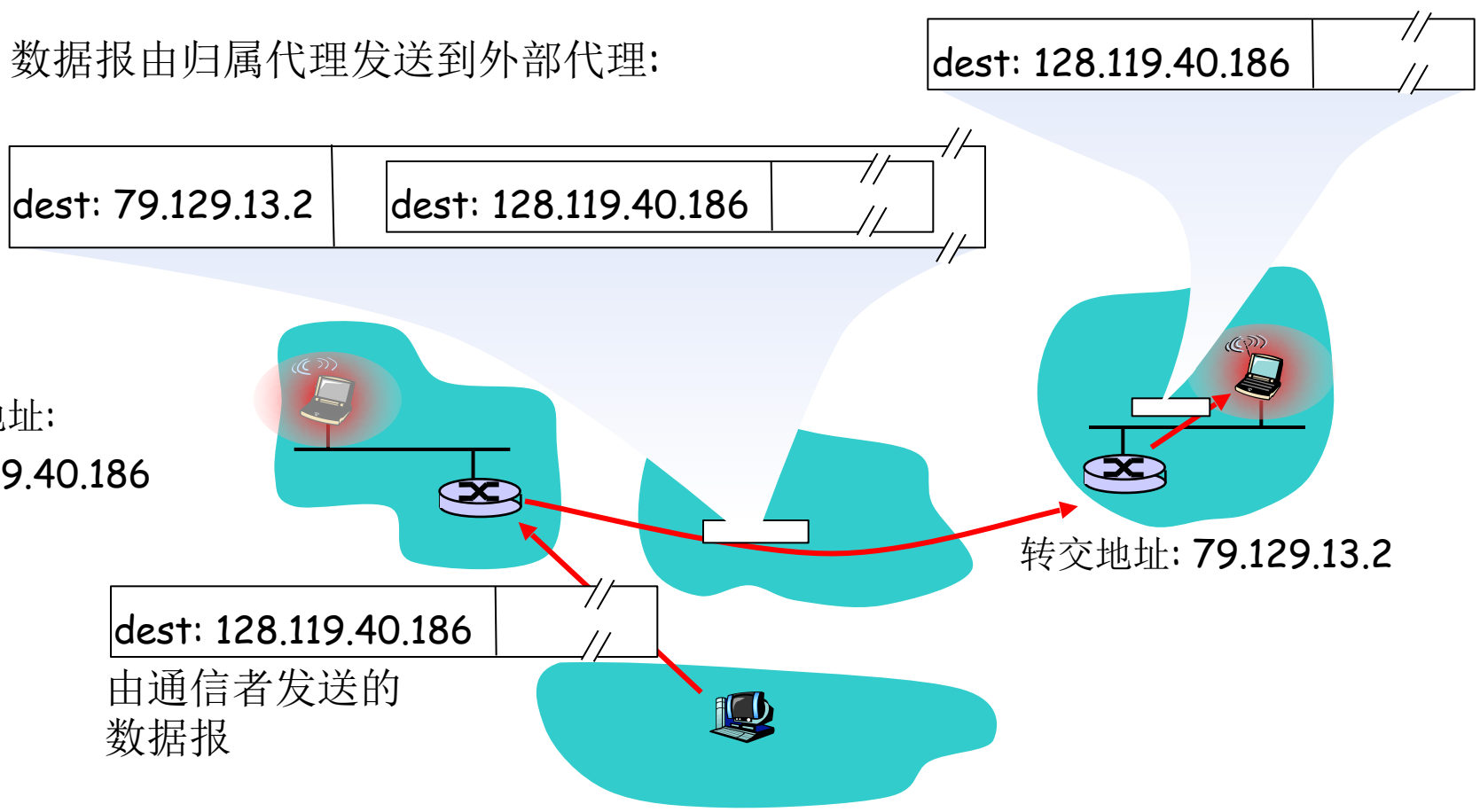
移动IP

- RFC 3344
- 包含我们考虑过的许多元素:
  - 归属代理，外部代理，外部代理注册，转交地址和封装
- 标准由三部分组成:
  - 数据报间接选路
  - 代理发现
  - 向归属代理注册

移动IP: 间接选路

外部代理到移动节点的数据报

数据报由归属代理发送到外部代理:

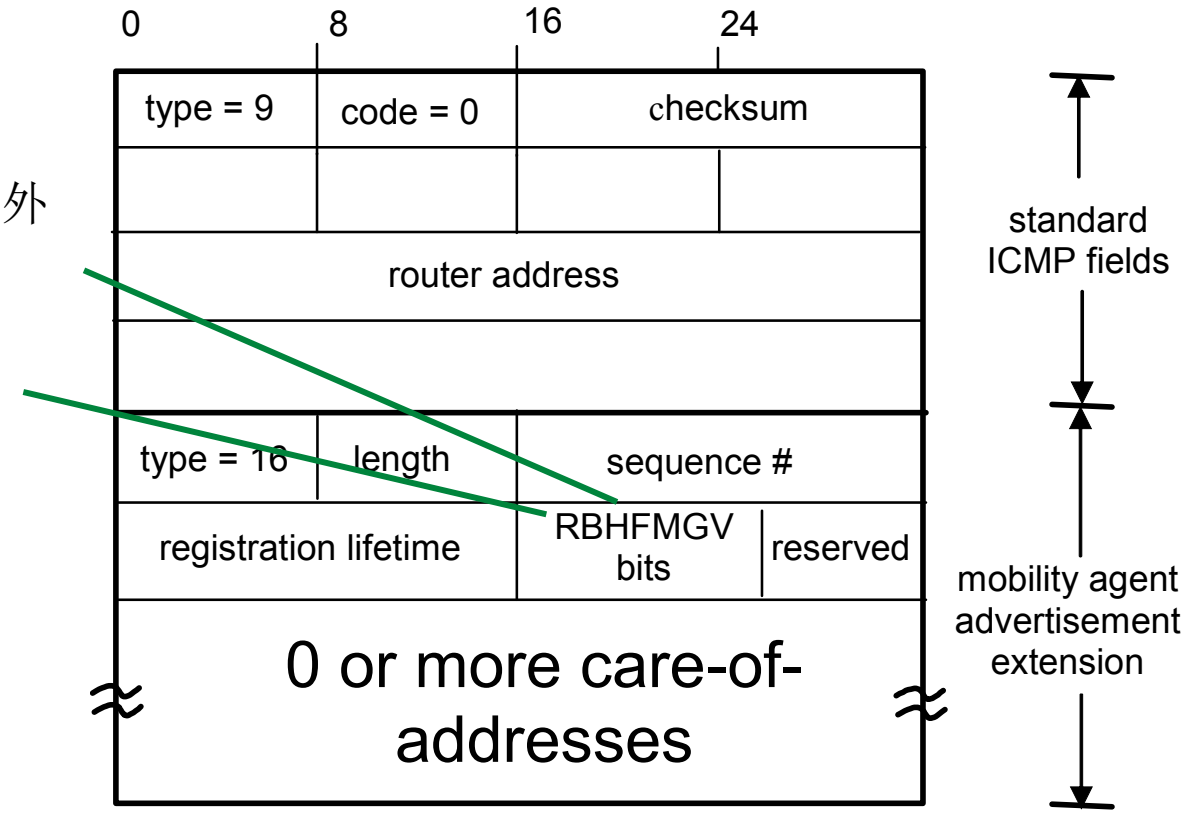


移动IP: 代理发现

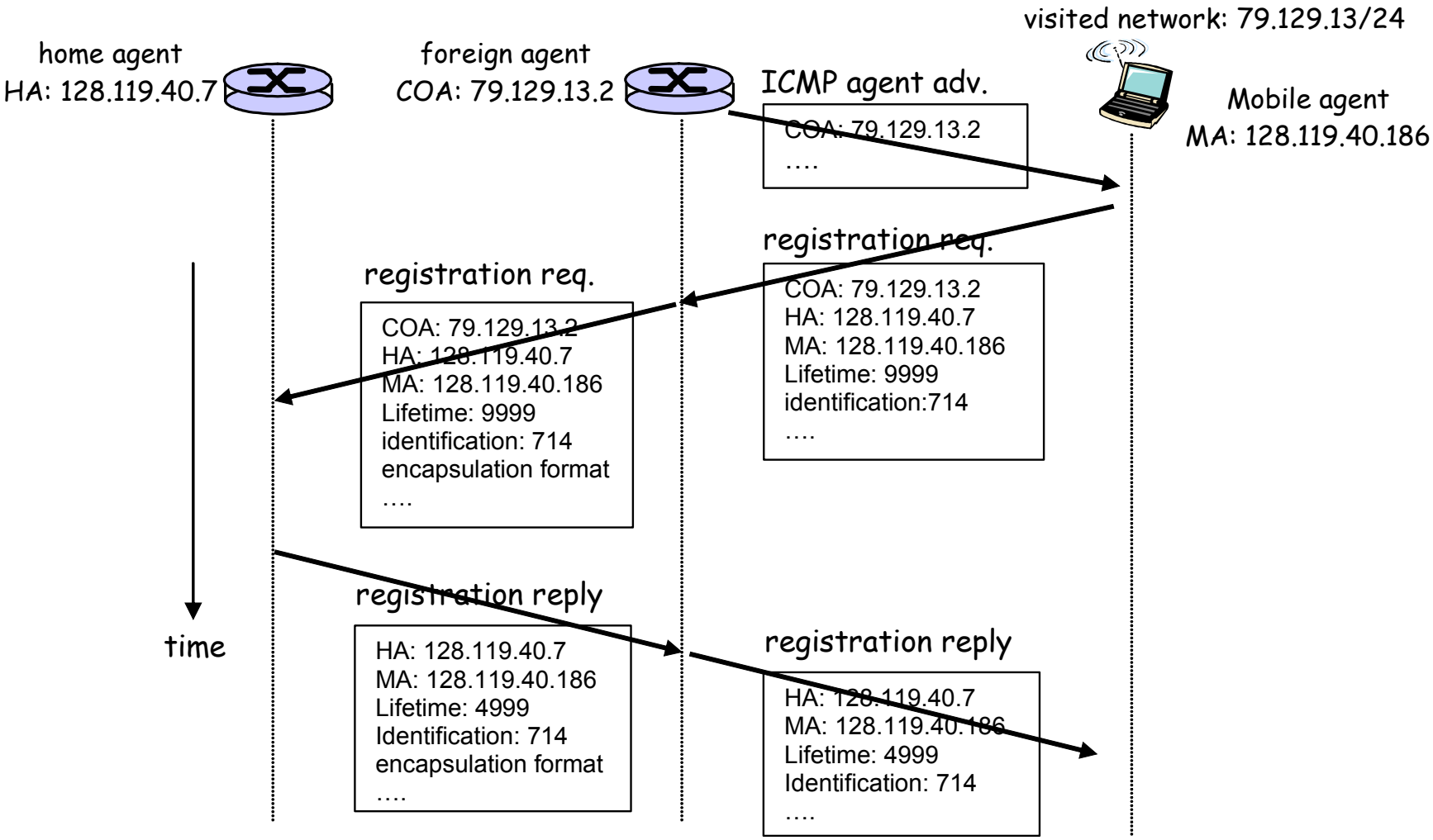
- 代理通告: 外部代理或归属代理广播一个类型字段为9的ICMP报文，来进行通告服务

H,F比特: 归属和/或外部代理

R比特: 注册请求

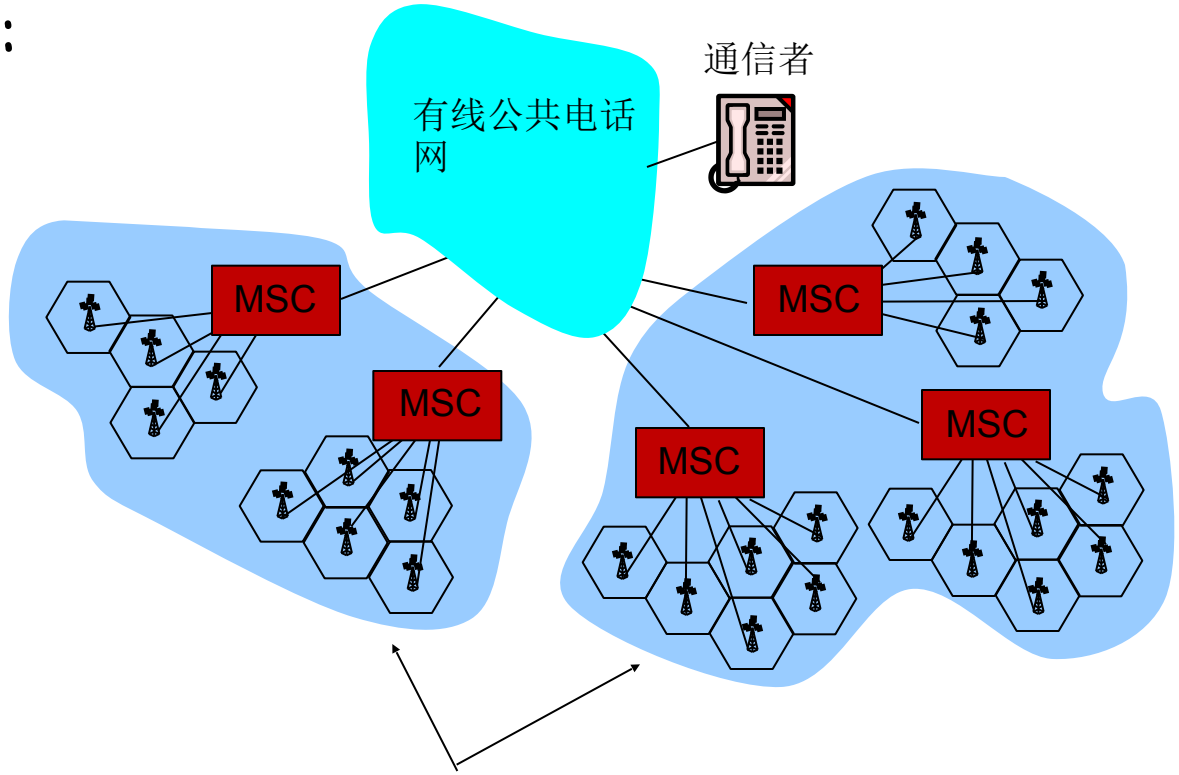


移动IP: 注册示例


visited network: 79.129.13/24
home agent
HA: 128.119.40.7
foreign agent
COA: 79.129.13.2
ICMP agent adv.
COA: 79.129.13.2
….
Mobile agent
MA: 128.119.40.186
registration req.
COA: 79.129.13.2
HA: 128.119.40.7
MA: 128.119.40.186
Lifetime: 9999
identification:714
….
registration req.
COA: 79.129.13.2
HA: 128.119.40.7
MA: 128.119.40.186
Lifetime: 9999
identification: 714
encapsulation format
….
time
registration reply
HA: 128.119.40.7
MA: 128.119.40.186
Lifetime: 4999
Identification: 714
encapsulation format
….
registration reply
HA: 128.119.40.7
MA: 128.119.40.186
Lifetime: 4999
Identification: 714
….

蜂窝网络体系结构

回想:

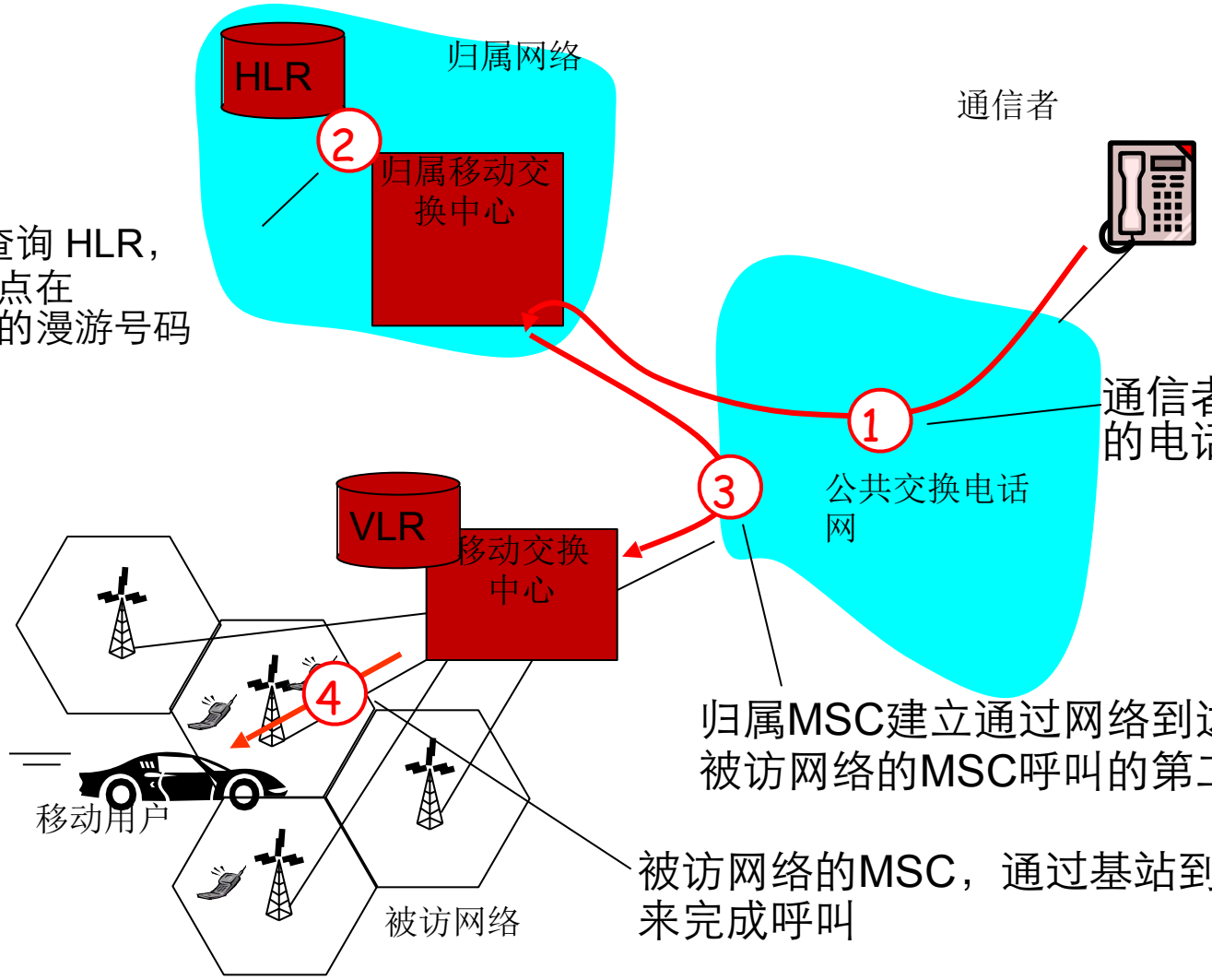


不同的蜂窝网，由不同的提供者运营

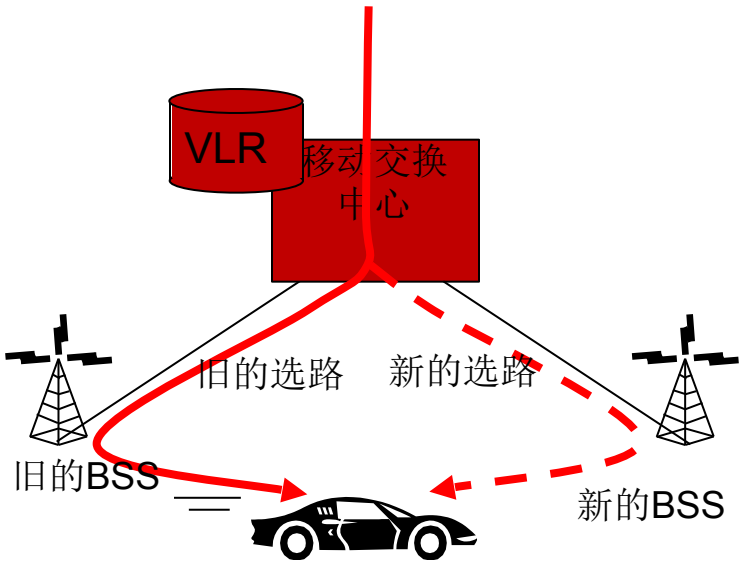
蜂窝网中的移动性处理

- 归属网络: 你提交请求的蜂窝网络（如Sprint PCS, Verizon)
  - 归属位置注册器（HLR): 归属网络中的数据库包含永久蜂窝电话号码，用户个人概要信息（服务，参数选择，账单），当前位置信息（可能是在另外的网络）
- 被访网络: 移动节点当前处于的网络
  - 访问者位置注册器（VLR): 包含网络中的每个当前用户的访问入口
  - 可能是归属网络

GSM: 到移动节点的间接选路

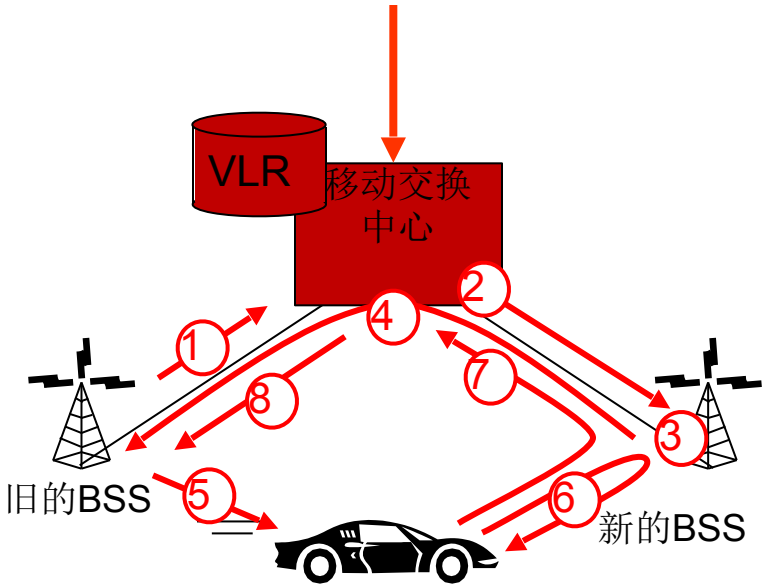


GSM：一般MSC的切换



- 切换的目标: 通过新的基站选路到移动节点 (不发生中断)
- 导致切换的原因:
  - 来自新BSS的较强信号
  - 负载均衡: 释放当前BSS中的通道
  - GSM没有说明为什么执行切换，仅仅说明了如何进行切换
- 切换由旧的BSS初始化（启动）

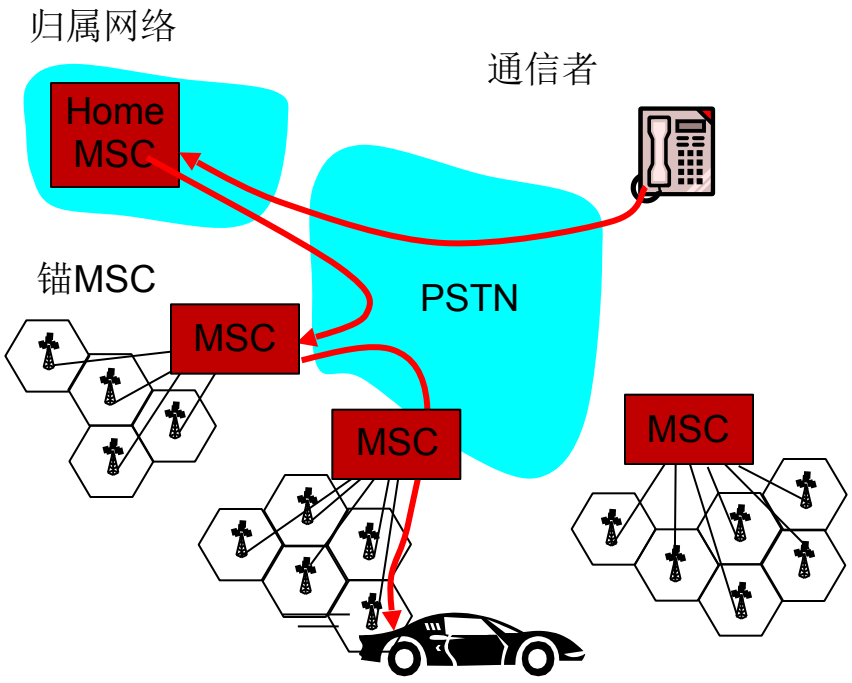
GSM：一般MSC的切换



1. 旧的BSS通知被访MSC即将要进行一个切换，通知移动用户切换时所涉及的新的BSS
2. MSC 建立到新BSS的路径 (分配资源)
3. 新的BSS 分配一个无线信道供移动用户使用
4. 新的BSS发出信令到MSC和旧的BSS: ready
5. 旧的BSS 告诉移动用户: 执行到新的BSS的切换
6. 移动用户和新的BSS交换一个或多个报文，一激活新的BSS中的新信道
7. 移动用户通过新的BSS向MSC发送一个切换完成报文，MSC重新选路正在进行的到移动用户的呼叫

8 沿着到旧的BSS的资源被释放

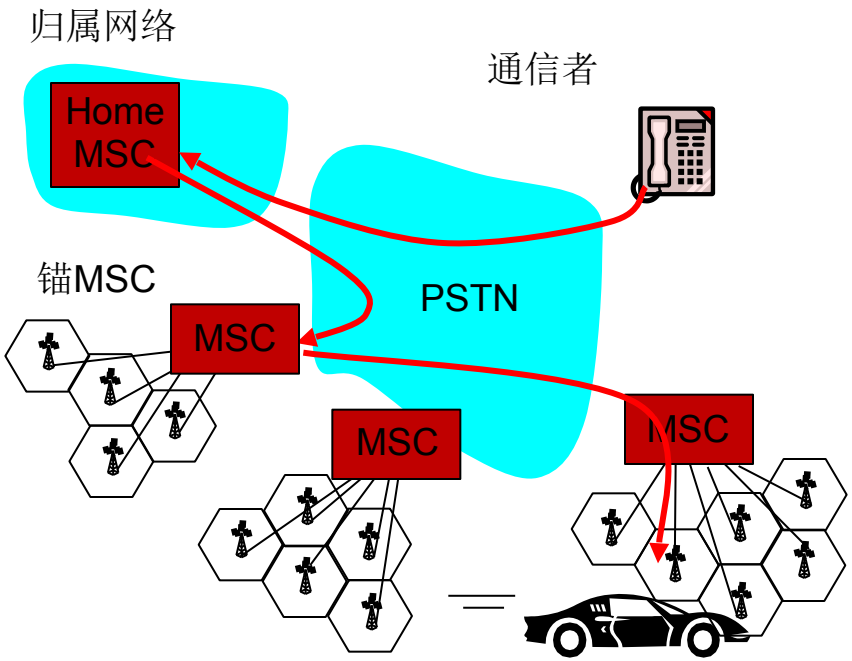
GSM: MSC间的切换



(a) 切换前

- 锚MSC: 呼叫首次开始时移动用户所访问的MSC
  - 呼叫总是选路到锚MSC
- 当移动用户移动到新的MSC时，新的MSC添加到MSC链的末尾
- IS-41 使用最少的可选路径来去除在锚MSC和当前被访问MSC之间的MSC

GSM: MSC间的切换



(b) 切换后

- 锚MSC: 呼叫首次开始时移动用户所访问的MSC
  - 呼叫总是选路到锚MSC
- 当移动用户移动到新的MSC时，新的MSC添加到MSC链的末尾
- IS-41 使用最少的可选路径来去除在锚MSC和当前被访问MSC之间的MSC

移动性: GSM vs 移动IP

GSM 元素	对GSM元素的评论	移动IP元素
归属系统	移动用户永久电话号码所归属的网络	归属网络
网关移动交换中心（或简称归属MSC），归属位置注册器（HLR）	归属MSC：获取移动用户路由地址的联系点。HLR：归属系统中包含移动用户永久电话号码、个人信息、当前位置和定制信息的数据库	归属代理
被访网络	移动用户当前所在的非归属网络	被访网络
被访移动服务交换中心，访问者位置注册器（VLR）	被访MSC：负责建立于MSC相关联的发射区中到/从移动节点的呼叫。VLR：被访网络中的临时数据库项，包含每个访问移动用户的订购信息。	外部代理
移动站点漫游号码（MSRN），或简称漫游号码	用于归属MSC和被访MSC之间电话呼叫的路由地址，对移动用户和通信者均不可见	转交地址

无线，移动性：对高层协议的影响

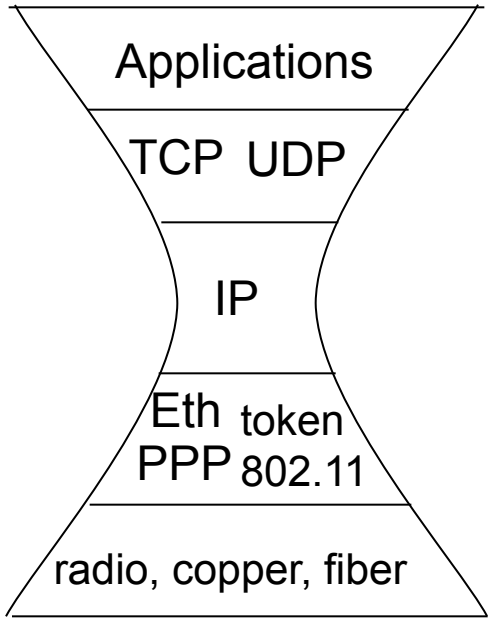
- 目标是最小化可能带来的影响
  - 尽力而为服务模型应该保持不变
  - TCP和UDP可以运行在具有无线链路的网络中
- 但是，带来性能方面的影响：
  - 比特错误或者移动切换带来数据包丢失/时延
  - TCP将丢失解释为拥塞，将不必要的减小拥塞窗口
  - 延迟对实时流量有较大的不利影响
  - 无线链路带宽有限

提纲

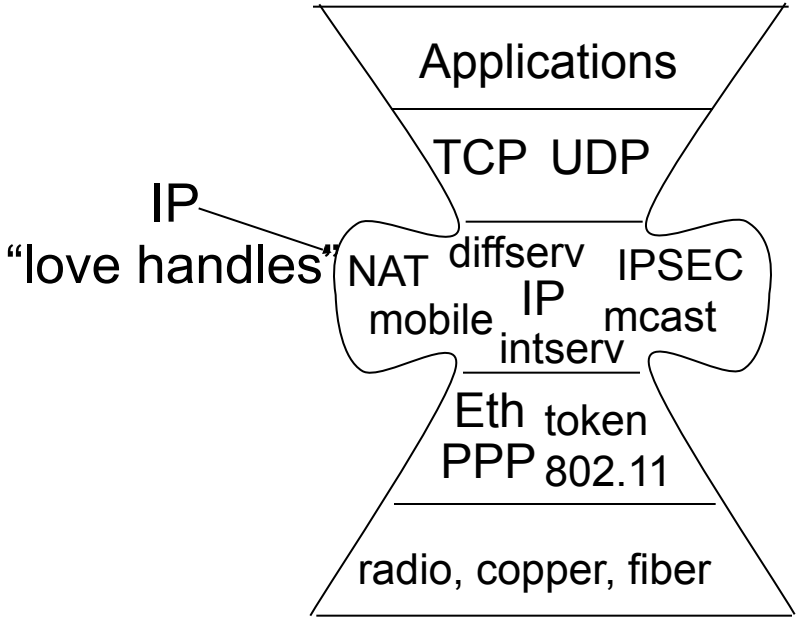
- 引言
  - 核心问题: 扩展到数十亿节点
- 全球互联网
- 多播
- 移动设备之间的路由
- 总结

赢者通吃：如何支持新的应用？

互联网中年危机: 思想狭隘, 心宽体胖？



IP “hourglass”



Middle-age IP “hourglass” ?

谢谢！

华中科技大学
电子信息与通信学院
Email: itec@hust.edu.cn
网址：http://itec.hust.edu.cn

参考资料

- *Chapter 4 in L. L. Peterson and B. S. Davie, Computer Networking: A System Approach (5th edition), Morgan Kaufmann, 2012*
- *Chapter 4 in James F. Kurose and Keith W. Ross, Computer Networking: A Top-Down Approach (6th edition), Pearson Education Inc., 2012*
- 吴功宜，计算机网络（第3版）,清华大学出版社，*2011*

附录