1. 为什么不同的网络可以相互连接？
2. 路由器为何能够使数据包到达任意网络，而交换机则不行？
3. 路由器怎样为收到的数据包找到的行进路线？
4. 三层交换设备为什么没能取代路由器？

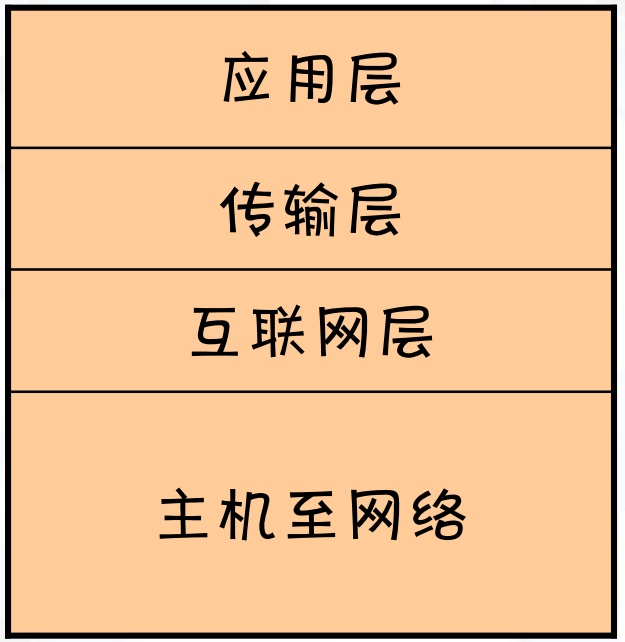
一、 网络互连的发展

Internet的发展

- 1970年，第一个分组交换网ARPA诞生，连接了四所大学
- 1972年，ARPA网有40个网点，应用有e-mail、rlogin、FTP，至此，网络的核心技术产生并开始研究网络的互联
- 1974年产生了两个基本的Internet协议：IP协议和TCP协议
- 80年代后期，产生了NSF网，它连接了美国所有的超级计算中心，并逐步取代了ARPA网

技术特点

- 使用TCP/IP协议
- 网络互联结构：各网之间一般采用路由器加专线连接
- 层次结构的域名及网络管理
- 分布式的管理模式，没有一个Internet管理中心
- 开发了通用的应用技术



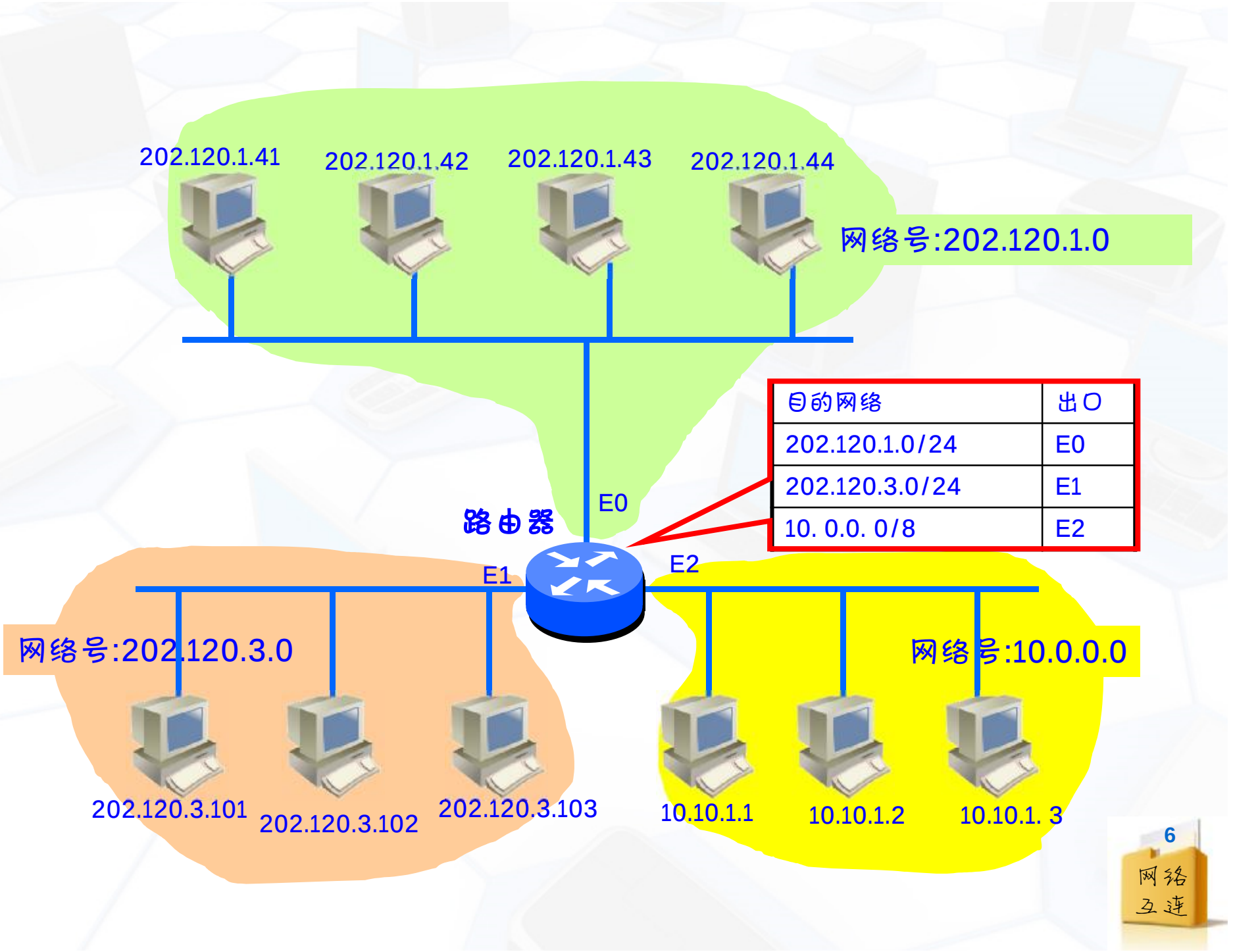
TCP/IP分层模型

二、 路由与路由算法

- 所谓路由是指为到达目的网络所进行的最佳路线选择
- 路由是网络层最重要的功能，路由选择是网络节点在收到一个分组后，基于通信子网在源与目的间提供的多条可能的传输路径，确定向下一节点传送路径的过程
- 在网络层完成路由功能的设备称为路由器，路由器是专门用于实现网络层功能的网络互连设备

1. 路由器与路由表

- 路由器是网络层的一个智能设备，承担了路由选择的任务，选择路由的依据是一张路由表，路由表指明了要到达某个地址该走哪一条路径
- 在路由表中，并非为每一个具体的目标IP地址指明路径，而是为目标IP地址所在的网络指明路径，这样路由表的大小才落在可操作的范围内，因此查找路由表的依据是目标主机的网络地址
- 路由器取出每一个接收到分组的目标IP地址，然后根据该地址中的网络地址查找路由表，确定下一步的传输路径，并从相应的路由器端口将分组送出，传送路径是由所经过的路由器一步一步确定的


202.120.1.41
202.120.1.42
202.120.1.43
202.120.1.44
网络号:202.120.1.0
E0
路由器
E1
E2
目的网络 出口
202.120.1.0/24 E0
202.120.3.0/24 E1
10. 0.0. 0/8 E2
网络号:202.120.3.0
网络号:10.0.0.0
202.120.3.101
202.120.3.102
202.120.3.103
10.10.1.1
10.10.1.2
10.10.1. 3

n 查看路由表

- ü Linux系统：ip route show
- ü Windows系统：route print

```
Active Routes:
Network Destination        Netmask          Gateway       Interface  Metric
          0.0.0.0          0.0.0.0      192.168.1.1    192.168.1.70      10
        127.0.0.0        255.0.0.0        127.0.0.1       127.0.0.1       1
      192.168.1.0    255.255.255.0    192.168.1.70    192.168.1.70      10
     192.168.1.70  255.255.255.255        127.0.0.1       127.0.0.1      10
    192.168.1.255  255.255.255.255    192.168.1.70    192.168.1.70      10
     192.168.50.0    255.255.255.0    192.168.50.1    192.168.50.1      20
     192.168.50.1  255.255.255.255        127.0.0.1       127.0.0.1      20
   192.168.50.255  255.255.255.255    192.168.50.1    192.168.50.1      20
     192.168.73.0    255.255.255.0    192.168.73.1    192.168.73.1      20
     192.168.73.1  255.255.255.255        127.0.0.1       127.0.0.1      20
   192.168.73.255  255.255.255.255    192.168.73.1    192.168.73.1      20
        224.0.0.0        240.0.0.0    192.168.1.70    192.168.1.70      10
        224.0.0.0        240.0.0.0    192.168.50.1    192.168.50.1      20
        224.0.0.0        240.0.0.0    192.168.73.1    192.168.73.1      20
  255.255.255.255  255.255.255.255    192.168.1.70    192.168.1.70       1
  255.255.255.255  255.255.255.255    192.168.50.1    192.168.50.1       1
  255.255.255.255  255.255.255.255    192.168.73.1    192.168.73.1       1
Default Gateway:       192.168.1.1
===========================================================================
Persistent Routes:
  None
```

2. 路由算法

- 路由算法设计必须考虑的问题

  正确性 简单性 健壮性 稳定性 公平性 最优性

- 路由算法的分类

  静态算法和动态算法
  单路径和多路径
  路由算法分区域内和区域间
  集中式和分布式

路由算法中的度量标准

- 路径长度：hop数
- 可靠性：线路出错率
- 延迟时间：路径延迟(带宽、路由器的排队长度)
- 带宽：最大带宽
- 路由器的负载
- 通信成本

相关的路由算法

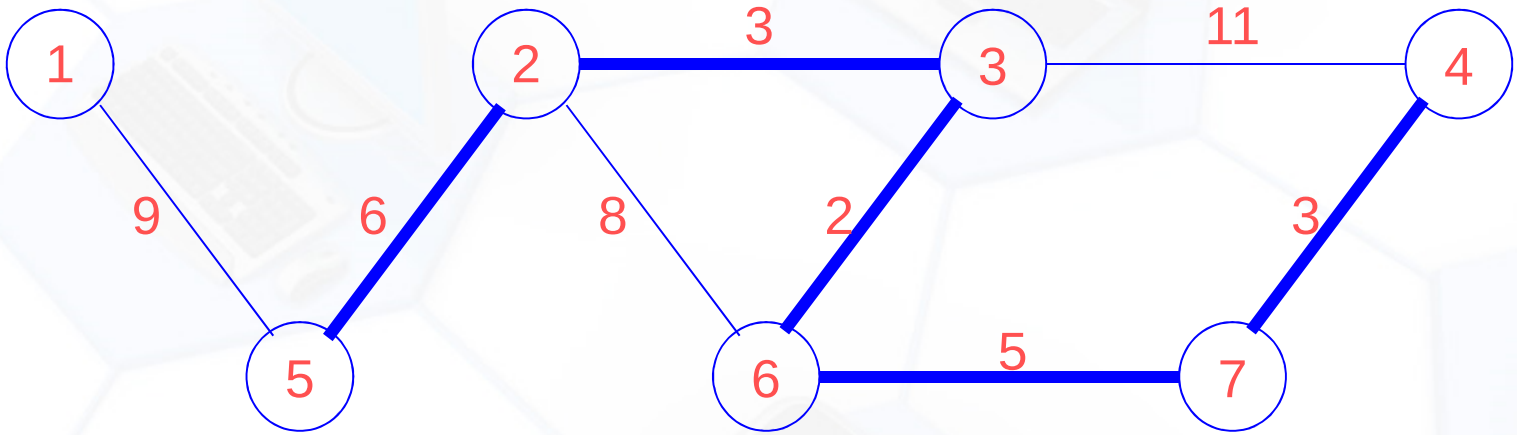
- 最短路径算法（Dijkstra）
- 扩散法(flooding)
- 基于流量的路由选择
- 距离矢量算法
- 链路状态算法
- 广播路由
- 多址传输路由选择(Multicast Routing)

1) 最短路由选择

§ Dijkstra算法（1959）：通过用边的权值作为距离的度量来计算最短路径，有最少边数的路径不一定是最短路径

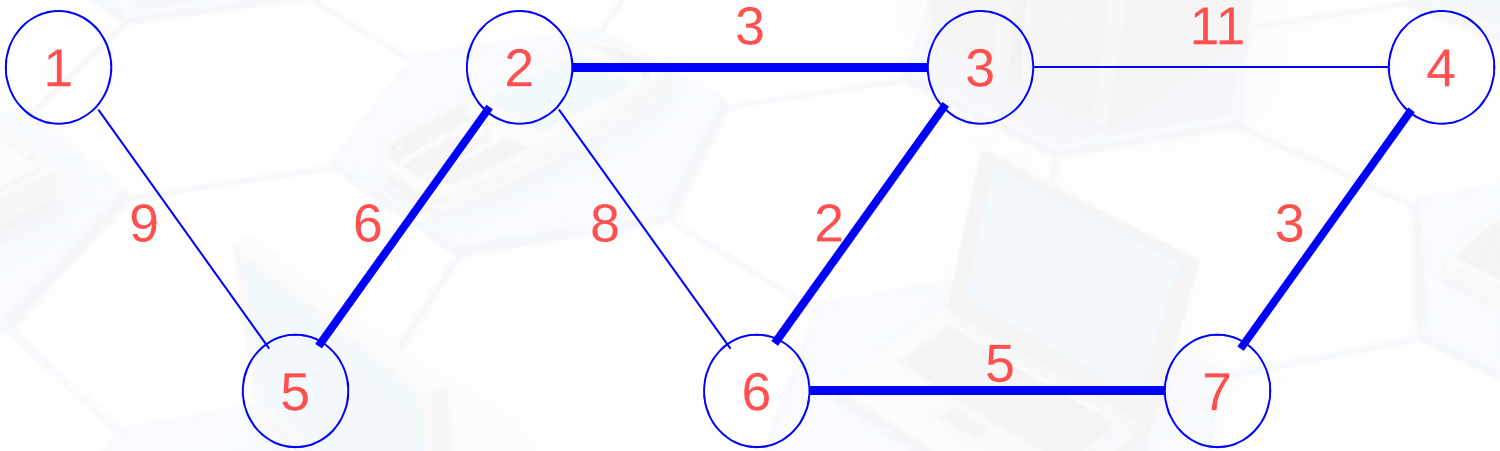
如下图：5和4之间边数最少的路径是5-2-3-4，

但最短路径是5-2-3-6-7-4



2) 扩散法(flooding)

§ 不计算路径，有路就走



如从5出发到4：

数据包从5®1,2；2®3,6；3®6,4；6®2,7；7®4

要解决的问题：数据包重复到达某一节点，如3，6

§ 数据包重复的解决方法

p 在数据包头设一计数器，每经过一个节点自动加1，达到规定值时，丢弃数据包

p 在每个节点上建立登记表，则数据包再次经过时丢弃

缺点：重复数据包多，浪费带宽

优点：可靠性高，路径最短，常用于军事

3) 距离矢量算法（Distance Vector Routing）

- 动态、分布式算法（RIP协议中使用本算法）
- 实现分布式算法的三要素：

The measurement process（测量）

The update protocol（更新邻接点距离矢量）

The calculation（计算）

pD-V算法的工作原理

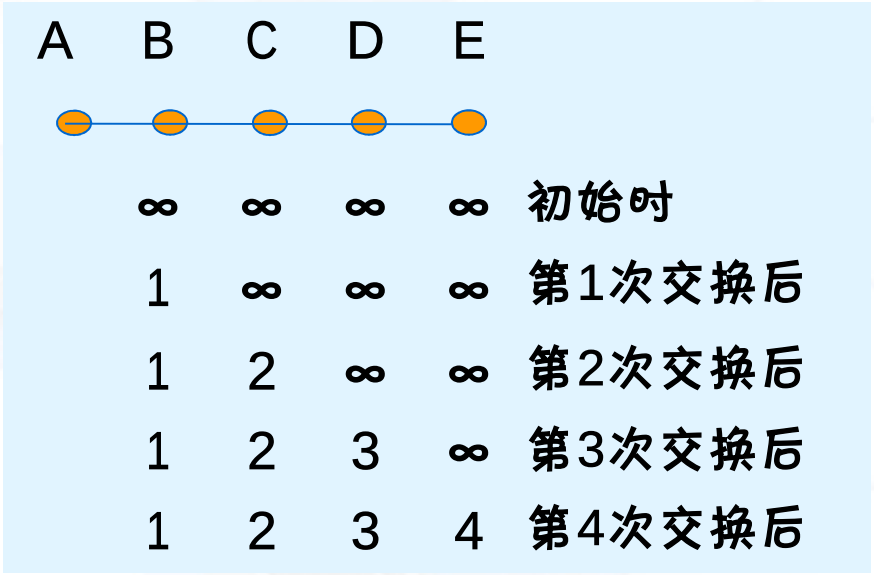
- ü 每个路由器用两个向量Di和Si来表示该点到网上所有节点的路径距离及其下一个节点
- ü 相邻路由器之间交换路径信息
- ü 各节点根据路径信息更新路由表

pD-V算法的缺点

- ü 交换的路径信息量大
- ü 路径信息不一致
- ü 收敛速度慢（坏消息）
- ü 距离矢量中不考虑带宽因子
- ü 不适合大型网络

p 无穷计算问题

——好消息传播得快，坏消息传播得慢



正常情况时

有突发情况时

网络互连

克服D-V算法收敛速度慢的方法

- 水平分裂
  - 同距离矢量法，只是到X的距离并不是真正的距离，对下方点通知真正的距离，对上方点给出无穷大
  - 如上图中的C点，它向D通知到A的真正距离，而向B通知到A的距离是无穷大

- Holddown
  - 当发现不通时，不重新选路径，而是把它设成无穷大

4) 链路状态算法(Link State Routing)

- 发现它的邻接节点，并得到其网络地址
- 测量它到各邻接节点的延迟或开销
- 组装一个分组以告知它刚知道的所有信息
- 将这个分组发给所有其他路由器
- 计算到每个其他路由器的最短路径

- 发现邻接节点

  当一个路由器启动后，向每个点到点线路发送HELLO分组，另一端的路由器发送回来一个应答来说明它是谁

- 测量线路开销

  发送一个ECHO分组要求对方立即响应，通过测量一个来回时间再除以2，发送方就可以得到一个延迟估计值，想要更精确些，可以重复这一过程，取其平均值

- 构造分组

  子网及其节点到其邻节点（路由器）的线路开销测量值（即延时，假设以毫秒计）

- 发布链路状态

  用扩散法（向邻接的节点）发布链路状态分组

- 存在的问题
  - 状态分组的重复到达
  - 如果序号循环使用，就会发生重复
  - 如果一个路由器被重启，序号将从0开始重新计数，但这些分组会被当成过时分组
  - 如果序号发生错误（如4被看成65540，第16位的0被误传成了1），则很多分组将被看成过时分组（此时5-65539均为过时分组，因为当前的分组序号是65540）

- 解决办法
  - 使用一个32位序号，即使每秒钟发送一个分组，137年才会循环一次
  - 在每个分组中加一年龄字段（如初值为60），每秒钟将年龄减1，为0后该分组将被丢弃，否则不会被认为是过时分组
- 计算新路由
  - 用Dijkstra算法计算到每个节点的路由
  - 得到该节点到每个节点的最短路径

p L-S路由算法的优缺点

ü各路由器的路由信息的一致性好

ü收敛性好，坏消息也一样传播得快

ü适用于大型网络，报文长度与网络规模关系不大

× 每个路由器需要有较大的存储空间

× 计算工作量大

三、 基于路由的互连

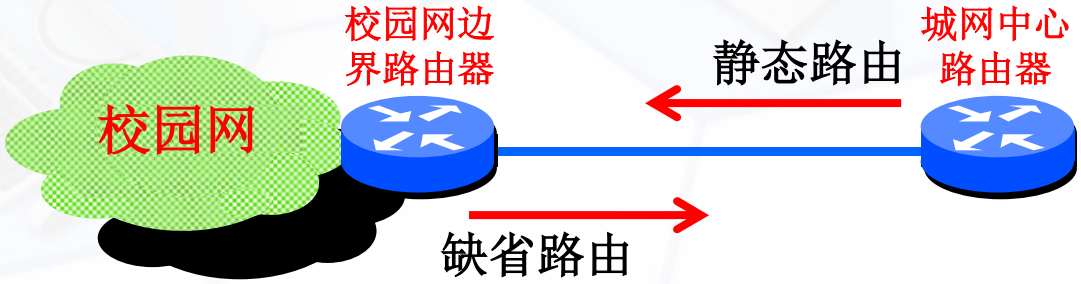
1. 静态路由与动态路由

- 静态路由(Static Route) — 人工在路由器上配置路由表
- 优点是路由器不必为路由表项的生成花费大量时间，有时可以抑制路由表的增长
- 缺点是人工配置开销大，网络拓扑结构变更时需重新配置路由表，一般只在小型网络或部分链路上使用

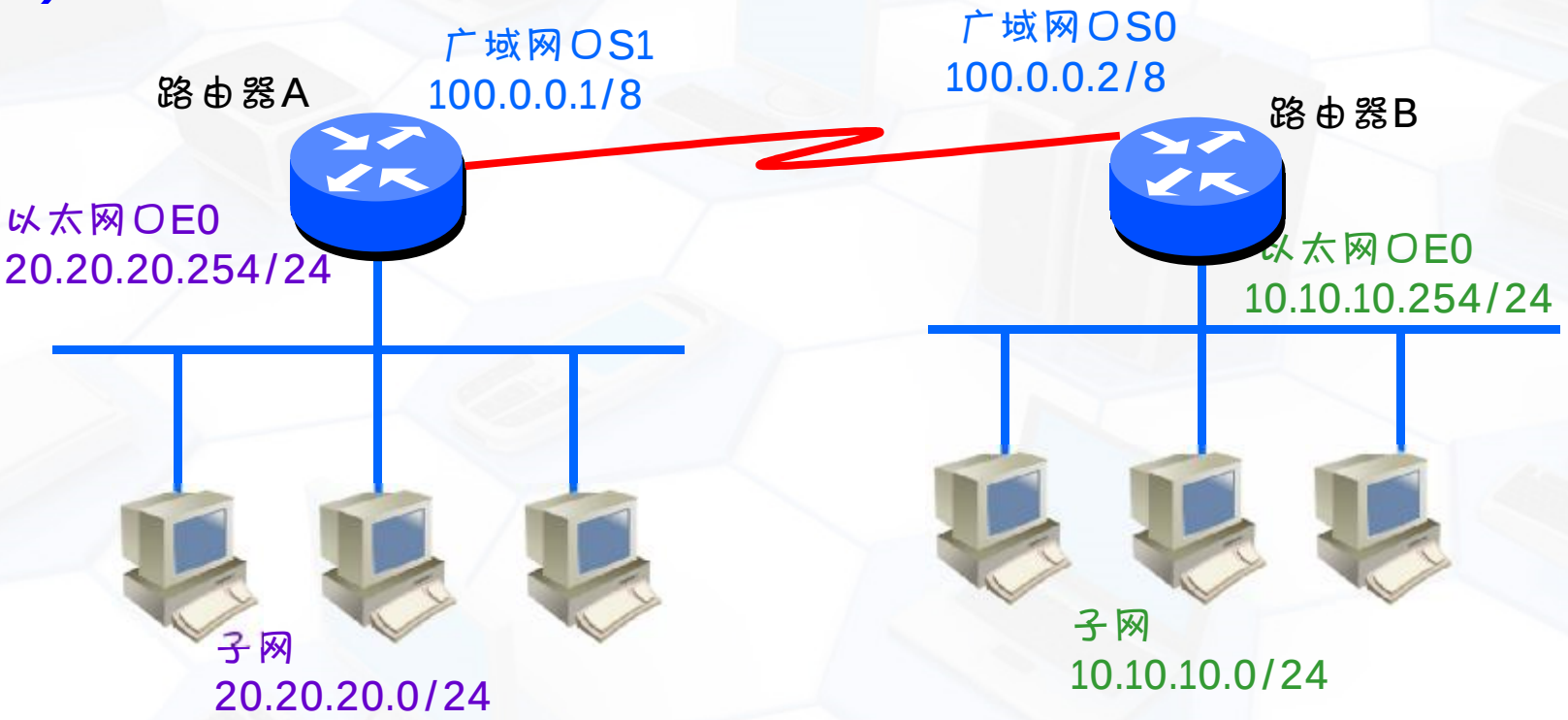
- 动态路由(Dynamic Route) — 由动态路由协议自动生成
- 优点是网络拓扑发生变化时，动态路由协议自动更新路由表
- 缺点是路由器路由计算开销大

1) 静态和缺省路由

- 缺省路由是静态路由的特例，也需要人工配置
- 互联网上有太多的网络和子网，受路由表大小的限制，路由器不可能也没有必要为互联网上所有网络和子网指明路径
- 凡是在路由表中无法查到的目标网络，在路由表中明确指定一个出口，这种路由方法称之为缺省路由



2) 静态路由设置

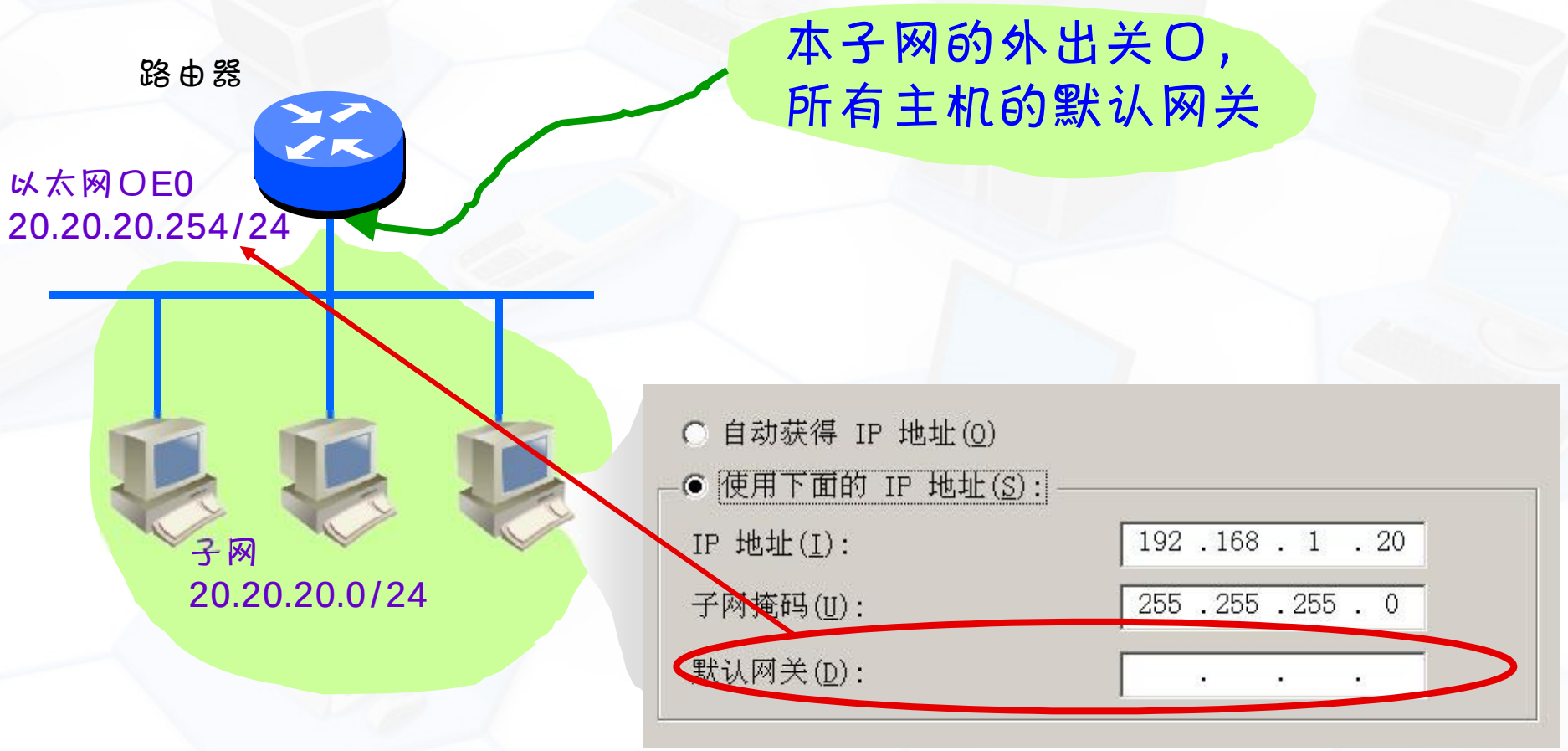


静态路由关键参数：目标网络、掩码、下一跳

路由A上的设置：ip route 10.10.10.0 255.255.255.0 100.0.0.2

路由B上的设置：ip route 20.20.20.0 255.255.255.0 100.0.0.1

3) 与路由设置相关的主机设定



2. 路由协议

- 网络层用于动态生成路由表信息的协议被称为路由协议
- 路由协议使网络中的路由设备能够相互交换网络状态信息，从而在内部生成关于网络连通性的数据并由此计算出到达不同目的网络的最佳路径或确定相应的转发端口

按路由选择算法分类

- 距离矢量路由协议
  - RIP
  - IGRP
- 链路状态路由协议
  - OSPF
- 混合型路由协议
  - IS-IS
  - EIGRP

按作用范围和目标分类

- 内部网关协议IGP
- 外部网关协议EGP

1) 有层次的路由

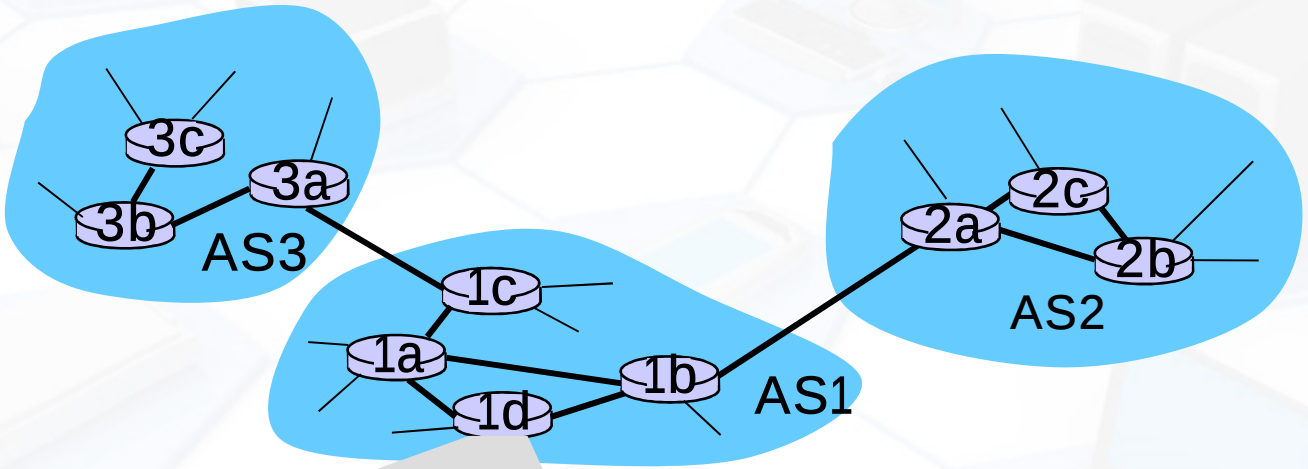
- 将路由划分到“区域”，称为“自治系统”(AS)
- 处于相同AS中的路由使用相同的路由协议
- 不同AS中的路由器可以使用不同的自治系统内路由协议

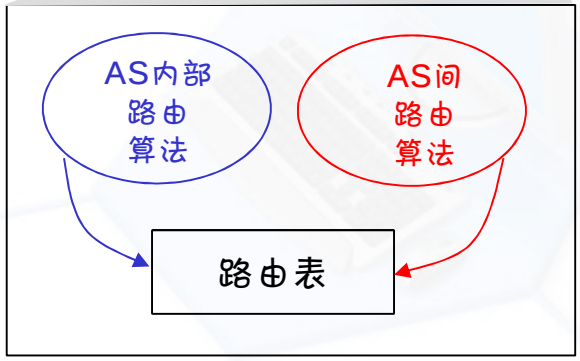
- ü 自治系统包含一个组织管理的一整套路由器和网络
- ü 除了发生故障的情况以外，一个自治系统是连通的(从图论的角度看)，即在任意一对节点之间都存在一条通路

网关路由器

- 直接通过链路连到其他AS中的路由器
- 属于AS的边界

2) AS间的互连





路由转发表由AS内和AS间路由算法共同更新

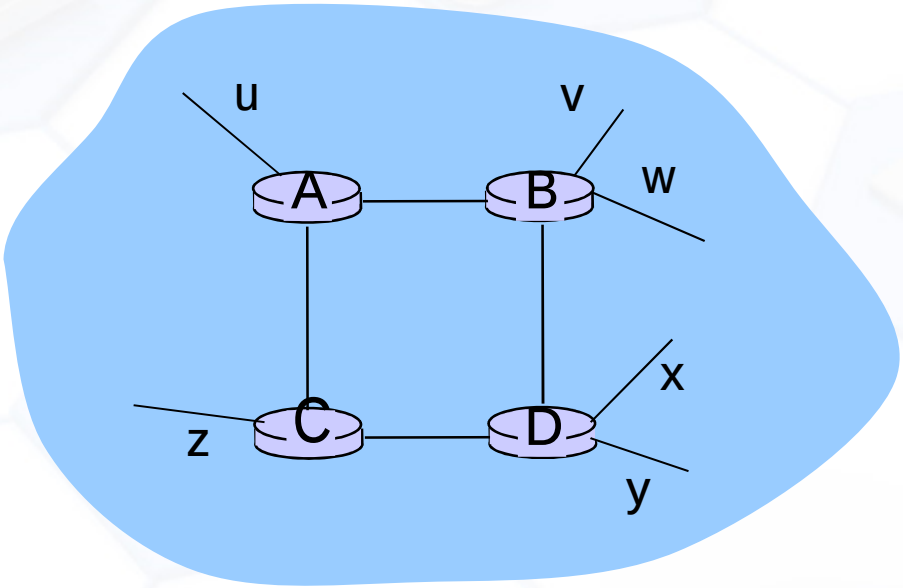
- AS内算法完成内部目的地的表项更新
- AS间与AS内算法 提供外部目的地的更新

3) AS内路由

- 也称为内部网关协议 (IGP)
- 常用AS内路由协议:
  - RIP: Routing Information Protocol
  - OSPF: Open Shortest Path First
  - IGRP: Interior Gateway Routing Protocol (Cisco专用)

(1) RIP协议（Routing Information Protocol）

- § RIP采用D-V路由算法，是Internet的一个主要路由协议
- § 版本号：有1和2两种，V2支持VLSM（可变长子网掩码）
- § 距离：一般为hop数，≤15



destination	hops
u	1
v	2
w	2
x	3
y	3
z	2

从路由A 出发的路由表情况

(a) RIP 的消息通告

- 使用称为“通告”的反馈消息，以30秒为间隔在相邻路由间交换信息
- 每个通告消息中最多可以列出一个自治系统中的25个目标子网

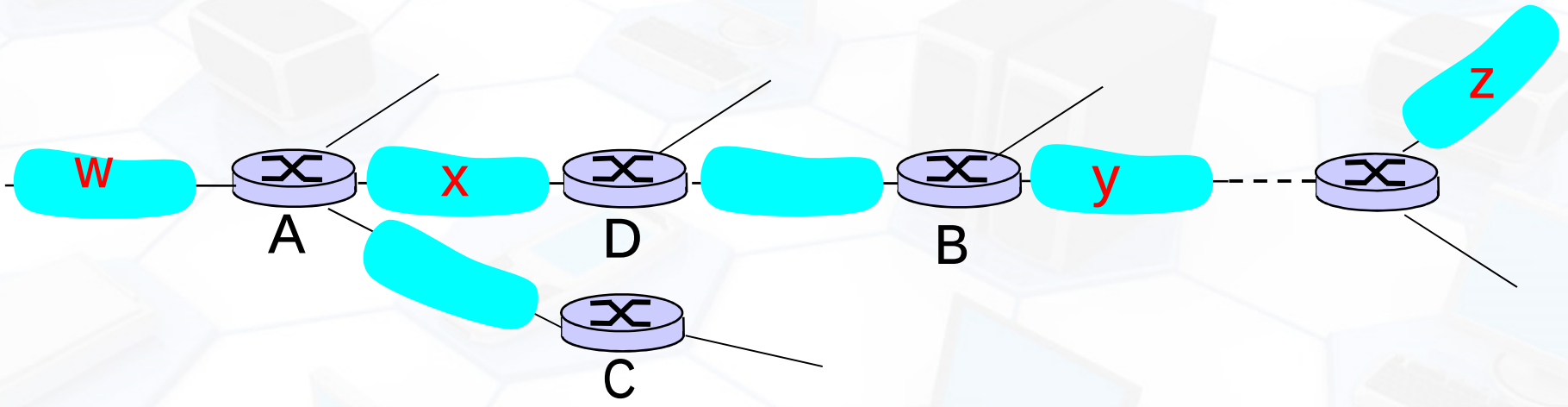
命令：

1—请求包

2—响应包

RIP报文格式

RIP举例——初始情况



目标网络	下一路由	至目标的跳数
w	A	2
y	B	2
z	B	7
x	—	1
….	….	….

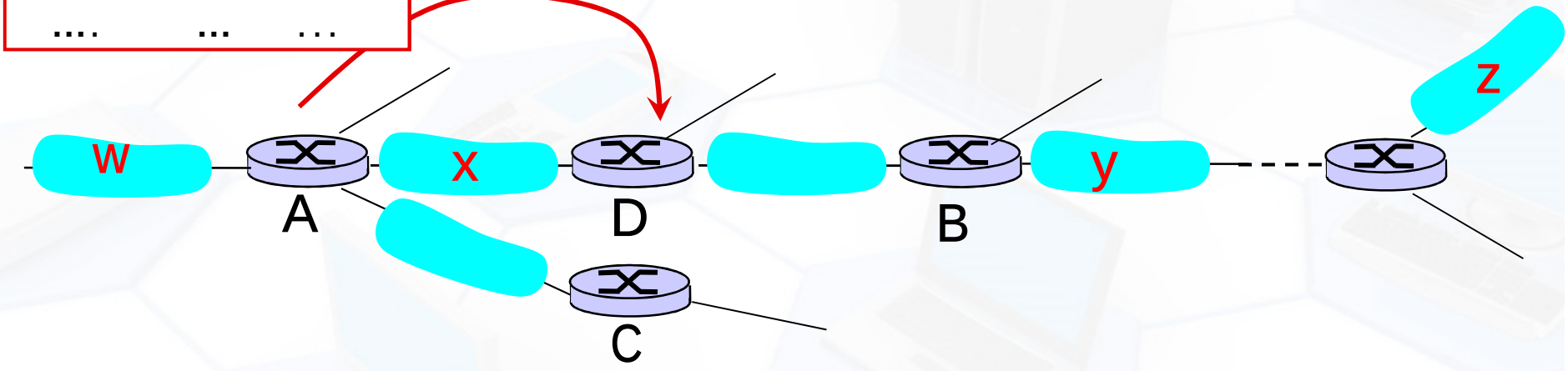
D的路由表

网络互连

RIP举例——A有新通告的情况

Dest	Next	hops
w	-	1
x	-	1
z	C	4
....	...	...

A发给D的通告消息



目标网络	下一路由	至目标的跳数
w	A	2
y	B	2
z	~~B~~ A	~~7~~ 5
x	--	1
....	....	....

D的路由表更新

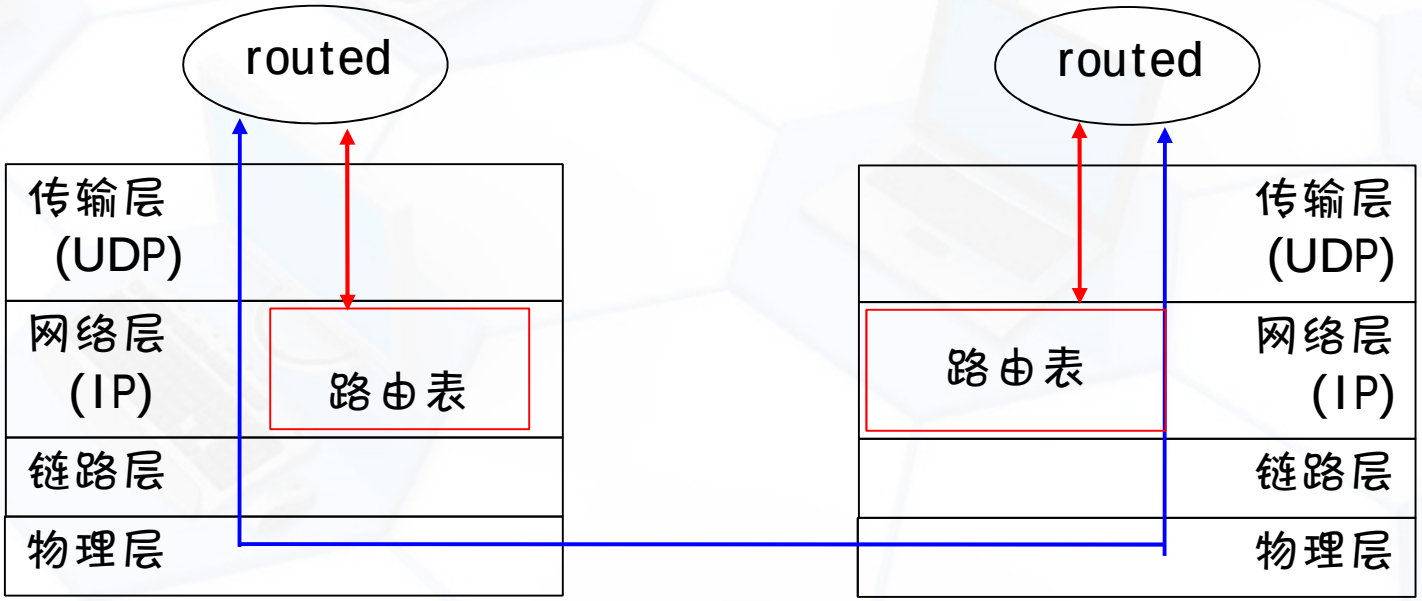
(b) RIP的链路错误与恢复

若180秒后未收到通告消息，相邻路由或链路将被视为失效

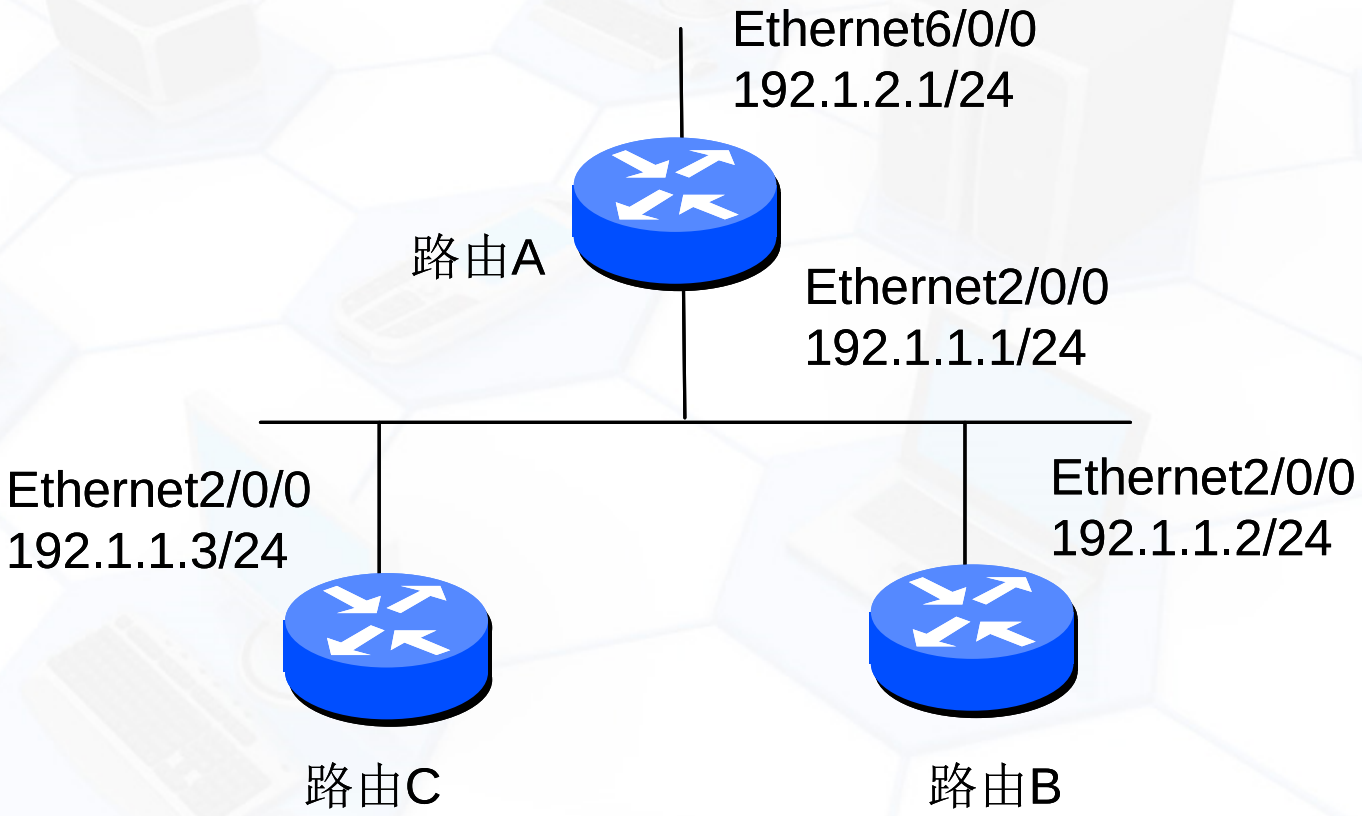
- 经由邻居节点的路线变为无效
- 新的通告被发往其他相邻路由
- 相邻路由轮流向外发出新通告(若路由表发生变化)
- 链路失效信息能够快速传播到整个网络
- 利用“毒害反向(poison reverse)”来避免往复循环(初始距离设定为16 hops)

(c) RIP 表的处理

- RIP路由表由属于应用层的进程 route-d (常驻进程)管理
- 通告以UDP包形式发送，并定期重复



(d) RIP典型配置


Ethernet6/0/0
192.1.2.1/24
路由A
Ethernet2/0/0
192.1.1.1/24
Ethernet2/0/0
192.1.1.3/24
Ethernet2/0/0
192.1.1.2/24
路由C
路由B

p 配置路由器A

配置接口Ethernet2/0/0和Ethernet6/0/0

```
[Router] interface ethernet 2/0/0
[Router-Ethernet2/0/0] ip address 192.1.1.1 255.255.255.0
[Router-Ethernet2/0/0] quit
[Router] interface ethernet 6/0/0
[Router-Ethernet6/0/0] ip address 192.1.2.1 255.255.255.0
```

启动RIP，并配置在接口Ethernet2/0/0和Ethernet6/0/0上运行RIP

```
[Router] rip
[Router-rip] network 192.1.1.0
[Router-rip] network 192.1.2.0
```

配置接口Ethernet 6/0/0只接收RIP报文

```
[Router] interface ethernet 6/0/0
[Router-Ethernet6/0/0] undo rip output
[Router-Ethernet6/0/0] rip input
```


p 配置路由器B

配置接口Ethernet2/0/0

```
[Router] interface Ethernet 2/0/0
[Router-Ethernet2/0/0] ip address 192.1.1.2 255.255.255.0
```

启动RIP，并配置在接口Ethernet2/0/0上运行RIP

```
[Router] rip
[Router-rip] network 192.1.1.0
[Router-rip] import direct
```


p 配置路由器C

配置接口Ethernet 2/0/0
[Router] interface Ethernet 2/0/0
[Router-Ethernet2/0/0] ip address 192.1.1.3 255.255.255.0

启动RIP，并配置在接口Ethernet2/0/0运行RIP
[Router] rip
[Router-rip] network 192.1.1.0
[Router-rip] import direct

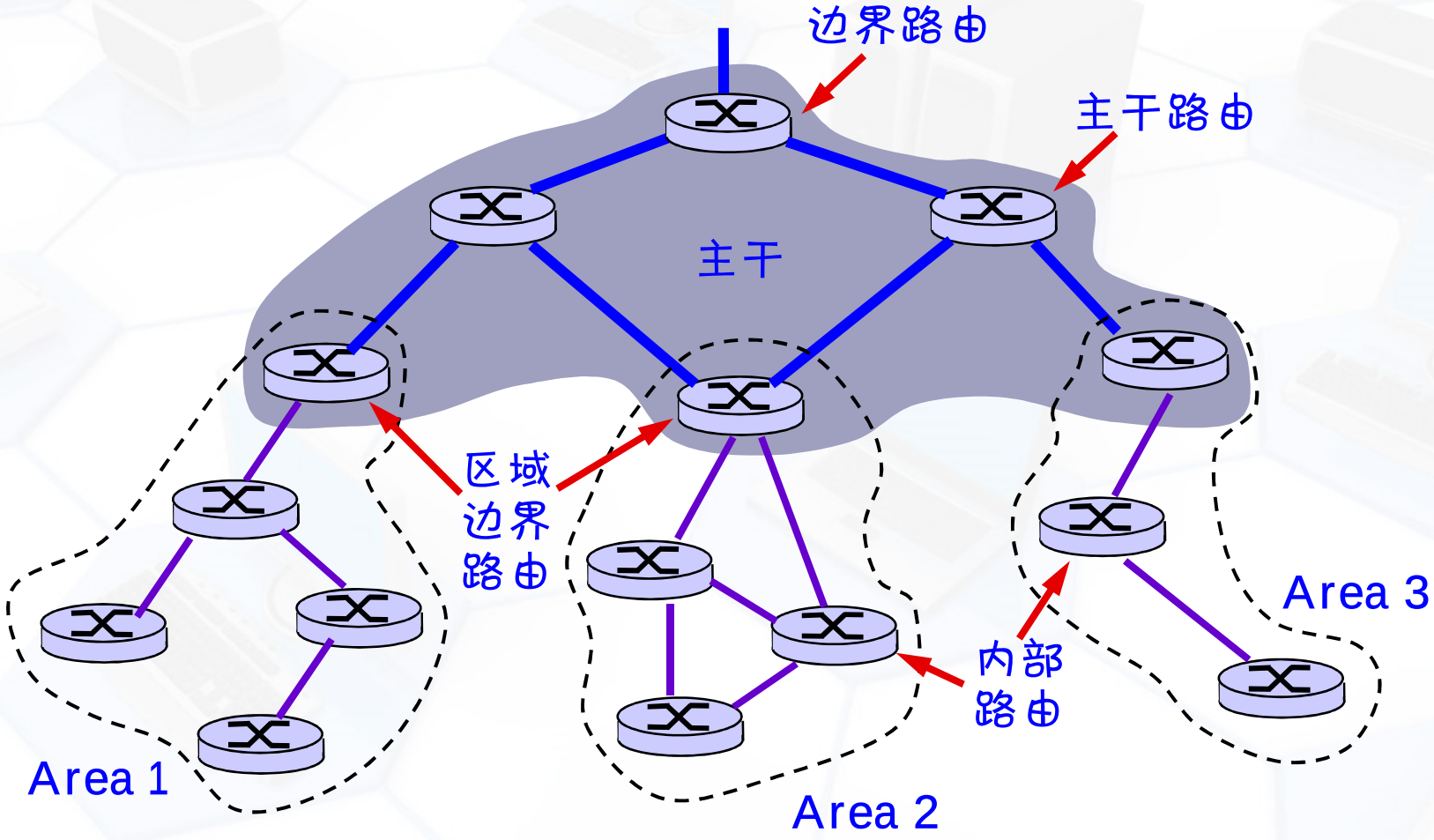
(2) OSPF协议(Open Shortest Path First)

- OSPF是Internet上主要的内部网关协议，负责AS内部路由
- 1988年开始制定，1990年成为标准，采用L-S路由算法
- 每个节点都有拓扑图
- 使用Dijkstra算法计算路由
- L-S通告信息在整个AS内以洪泛方式传递
- OSPF消息直接通过IP传递 (而非 TCP 或 UDP)

(a) OSPF 相对于RIP的优势

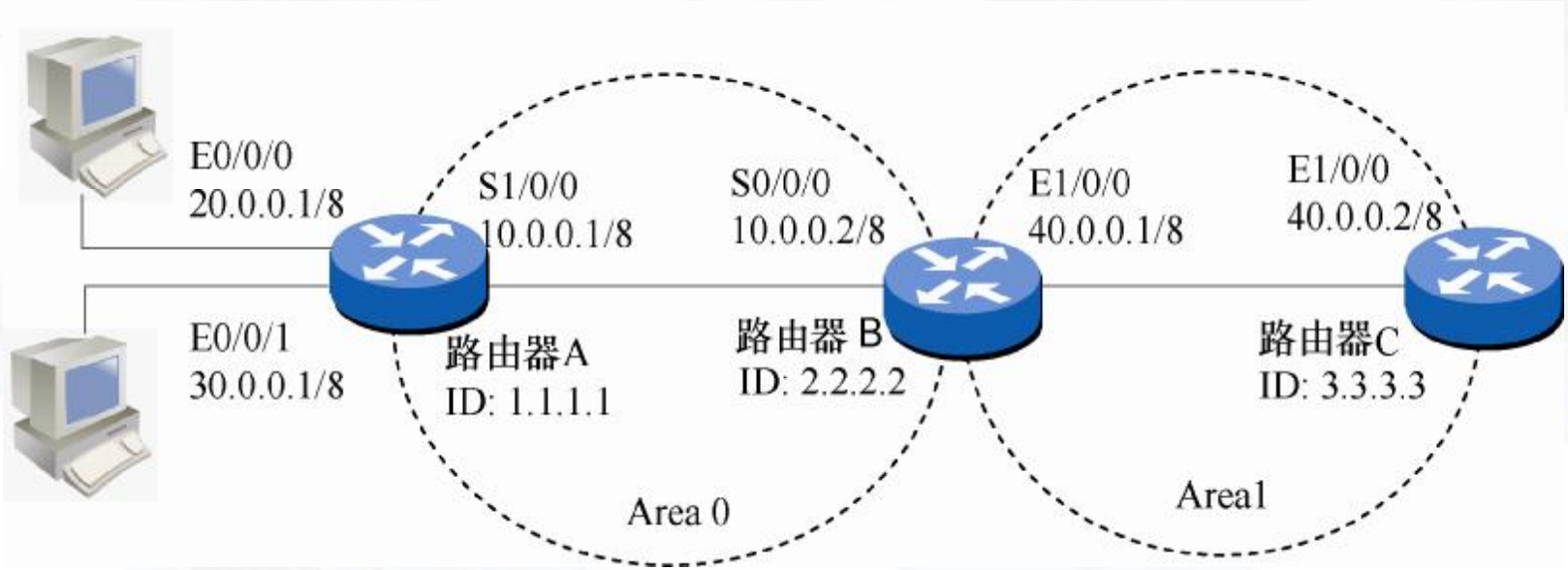
- **安全** 所有OSPF消息都经过鉴别以防止恶意代码
- **允许多条相同费用路径共存** (RIP中只能有一条)
- 每条链路允许针对不同TOS要求给出不同费用说明
- **集成了单播/组播支持**组播OSPF (MOSPF)使用与OSPF相同的拓扑结构数据库
- **在大型路由域内支持层次化的OSPF**

(b) OSPF的层次



- 两级层次：区域和主干
  - 链路状态通告仅在区域内
  - 每个节点都有详细的区域拓扑结构
  - 仅知道去其他区域子网的最短路径
- 区域边界路由器：汇总到本区域中其他子网的距离，向其他区域边界路由发送通告
- 主干路由器：执行OSPF但仅在主干范围路由选择
- 边界路由器：连接到其他的AS

(c) OSPF典型配置


E0/0/0
20.0.0.1/8
E0/0/1
30.0.0.1/8
S1/0/0
10.0.0.1/8
S0/0/0
10.0.0.2/8
E1/0/0
40.0.0.1/8
E1/0/0
40.0.0.2/8
路由器A
ID: 1.1.1.1
路由器 B
ID: 2.2.2.2
路由器C
ID: 3.3.3.3
Area 0
Area1

p 配置路由器A

```
<Router> system-view
[Router] router id 1.1.1.1
[Router] interface serial1/0/0
[Router-serial1/0/0] ip address 10.0.0.1 255.0.0.0
[Router-serial1/0/0] interface ethernet0/0/0
[Router-ethernet 0/0/0] ip address 20.0.0.1 255.0.0.0
[Router- ethernet 0/0/0] interface ethernet0/0/1
[Router- ethernet 0/0/1] ip address 30.0.0.1 255.0.0.0
[Router- ethernet 0/0/1] quit
[Router] ospf
[Router-ospf-1] area 0
[Router-ospf-1-area-0.0.0.0] network 10.0.0.1 0.255.255.255
[Router-ospf-1 -area-0.0.0.0] network 20.0.0.1 0.255.255.255
[Router-ospf-1 -area-0.0.0.0] network 30.0.0.1 0.255.255.255
```

p 配置路由器B

```
<Router> system-view
[Router] router id 2.2.2.2
[Router] internet serial0/0/0
[Router-serial0/0/0] ip address 10.0.0.2 255.0.0.0
[Router-serial0/0/0] interface ethernet 1/0/0
[Router-ethernet 1/0/0] ip address 40.0.0.1 255.0.0.0
[Router-ethernet 1/0/0] quit
[Router] ospf
[Router-ospf-1] area 0
[Router-ospf-1-area-0.0.0.0] network 10.0.0.2  0.255.255.255
[Router-ospf-1-area-0.0.0.0] area 1
[Router-ospf-1-area-0.0.0.1] network 40.0.0.1  0.255.255.255
```

p 配置路由器C

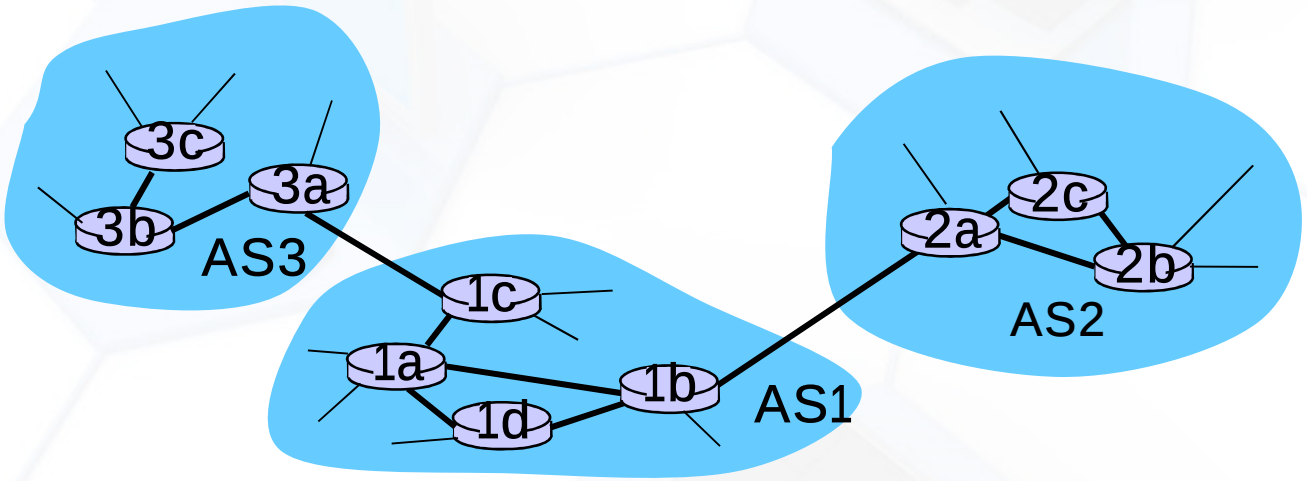
```
<Router> system-view
[Router] router id 3.3.3.3
[Router] interface ethernet 1/0/0
[Router-ethernet 1/0/0] ip address 40.0.0.2 255.0.0.0
[Router-ethernet 1/0/0] quit
[Router] ospf
[Router-ospf-1] area 1
[Router-ospf-1-area-0.0.0.1] network 40.0.0.2  0.255.255.255
```

4) AS之间的要做的工作

- 如果AS1中的路由 接收那些目的地在AS1之外的数据报，路由器需要将分组转发到某个作为网关的路由，但是哪个呢？

AS1 需完成：

- 学习通过AS2和AS3分别可以到达哪些目的地
- 将这些可去方向通知到AS1中的所有路由

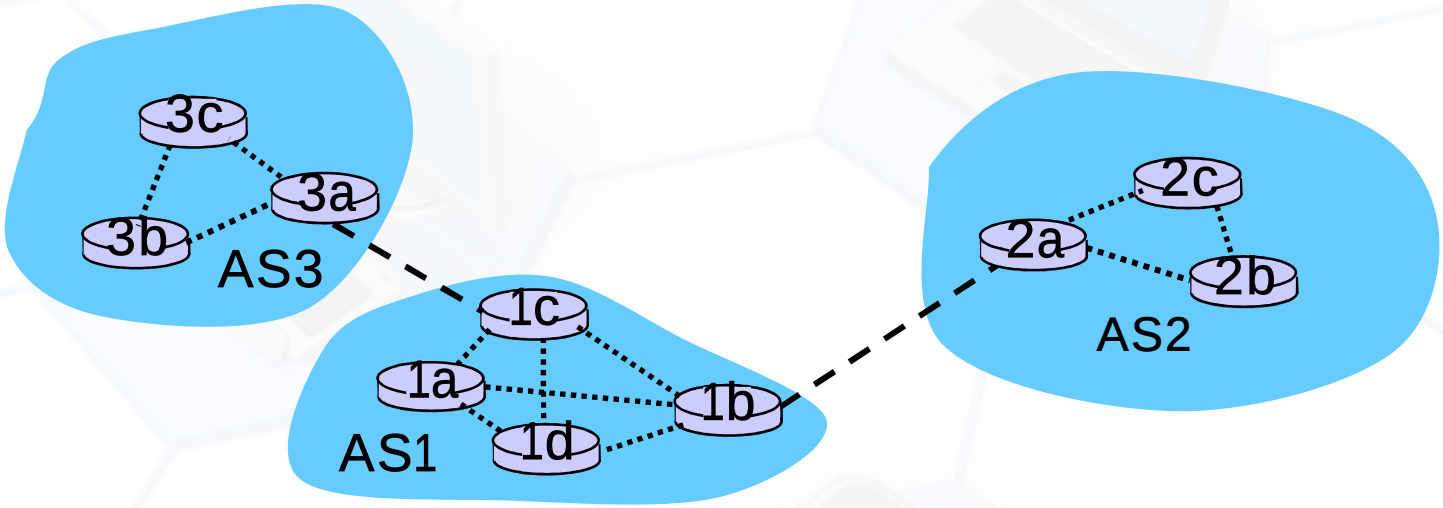


(1) AS间路由协议: Border Gateway Protocol

- BGP 为每个AS提供了一种手段:
  - 由邻居AS获取子网的可达性信息
  - 将可达性信息传递给AS中所有的内部路由
  - 依据可达性信息和路由策略确定到达子网的最佳路线
- 运行子网向互联网上的其他网络通告其存在

(2) BGP要点

- 成对的路由间(对等BGP) 通过半永久性的TCP连接交换路由信息(称为BGP 会话)
  - BGP会话不需要与实际物理链路有关
- 当AS2将一个网络号前缀通告给AS1, 就表明AS2确定会把所有带有该网络号的数据报发往该网段
  - AS2能够在其通告中汇集网络号



- 考虑AS之间的路由，使用的是距离矢量算法，但只在路由状态发生变化是才发送变化信息，使用TCP类连接传输层，进行信息交换
- BGP是用在自治系统之间实现网络可达信息的交换，整个交换过程要求建立在可靠的传输连接基础上
- BGP使用TCP作为传输协议，它可以将所有的差错控制功能交给传输协议来处理，这样其本身就变得简单多了
- BGP的当前版本称为BGP-4

四、 网络地址转换（Network Address Translation）

- 网络地址转换允许一个机构内的局域网以一个合法地址出现在互联网上。NAT将每个局域网节点上的地址转换成一个合法的IP地址，反之亦然
- NAT也可以应用到防火墙技术中，将个别IP地址隐藏起来，使外界无法直接访问内部网络设备
- 同时，NAT还可以帮助网络超越地址的限制，合理地安排网络中私有IP地址的使用

1. NAT基本原理

两种IP地址

- 全局IP地址：用于Internet上的分组转发，要求在Internet范围内唯一
- 私有IP地址：用于指定网络内的分组转发，只要求在指定网内部唯一

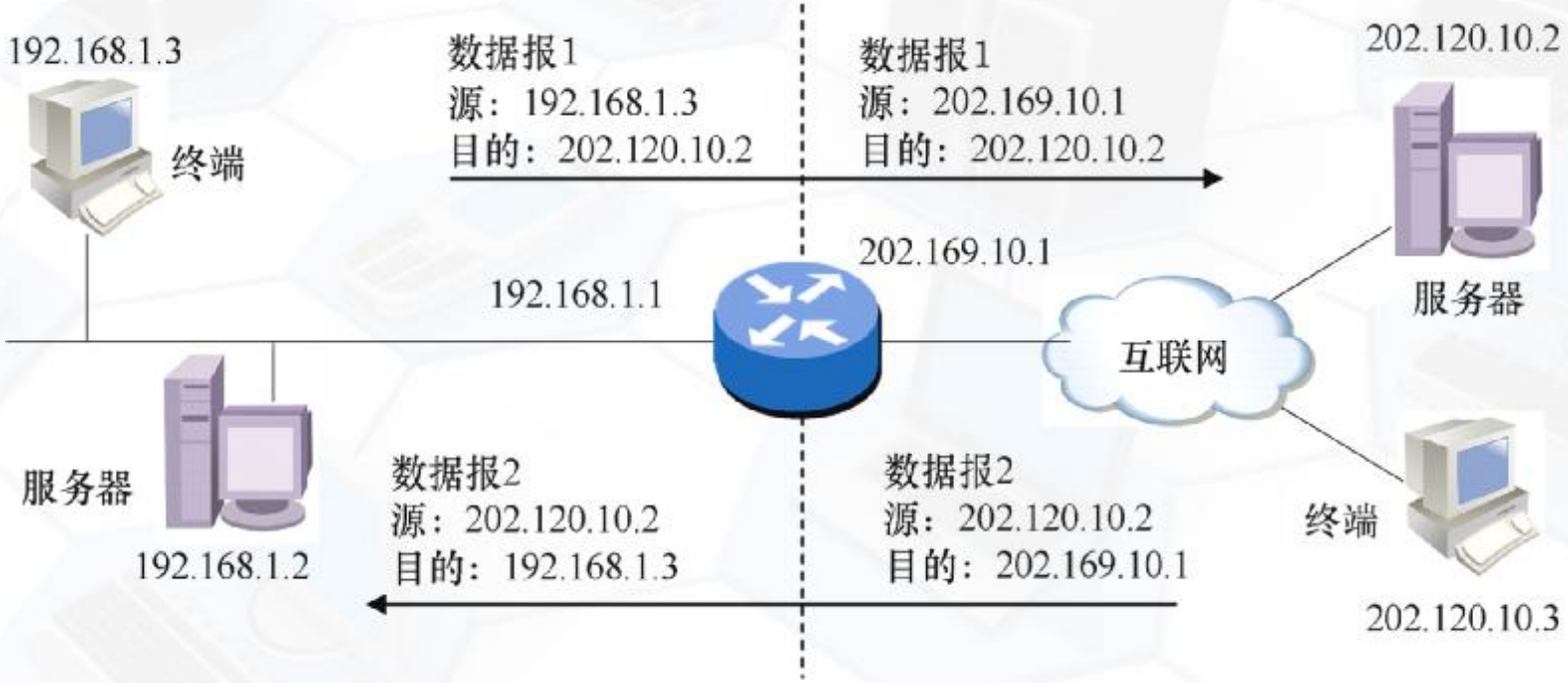
NAT：实现网络内的多台主机共享一个全局的IP地址

- NAT的具体的做法是将IP包内的地址域用合法的IP地址替换。NAT设备维护一个状态表，用来把非法的IP地址映射到合法的IP地址上去。每个IP包在NAT设备中都会被更换相关的IP地址

常用的私有IP地址

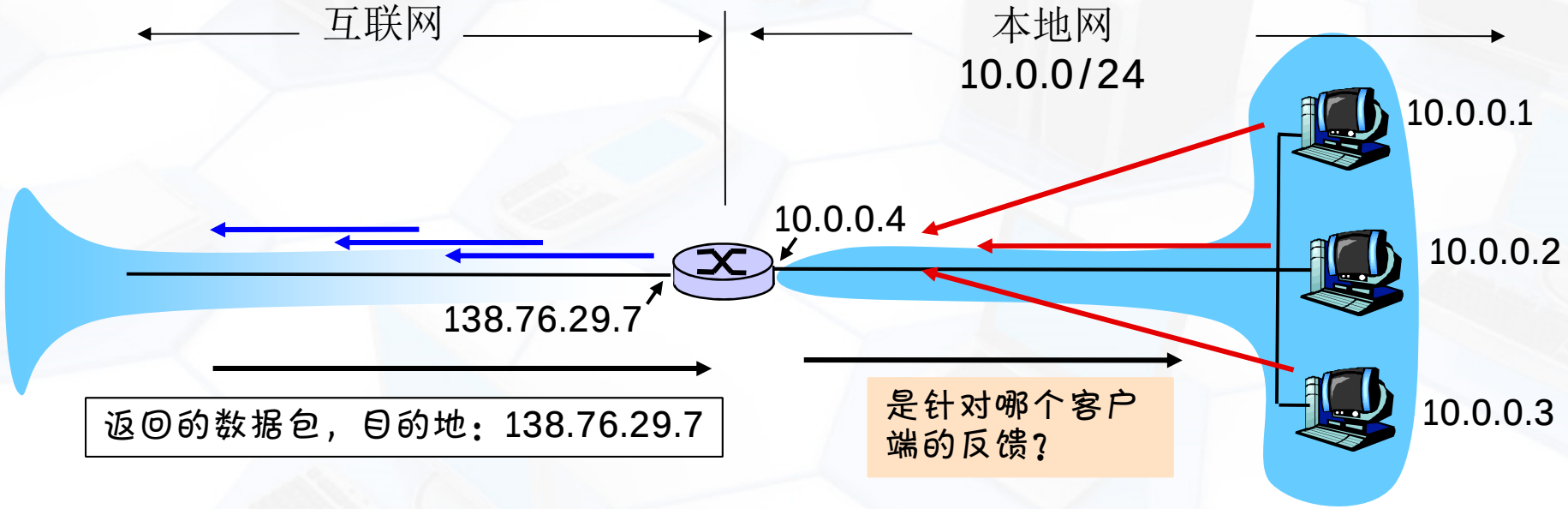
- 10.0.0.0 - 10.255.255.255
  - A single Class A network
- 172.16 .0.0- 172.31 .255.255
  - 16 contiguous Class B networks
- 192.168.0.0 - 192.168.255 .255
  - 256 contiguous Class C networks

2. NAT的具体实现类型

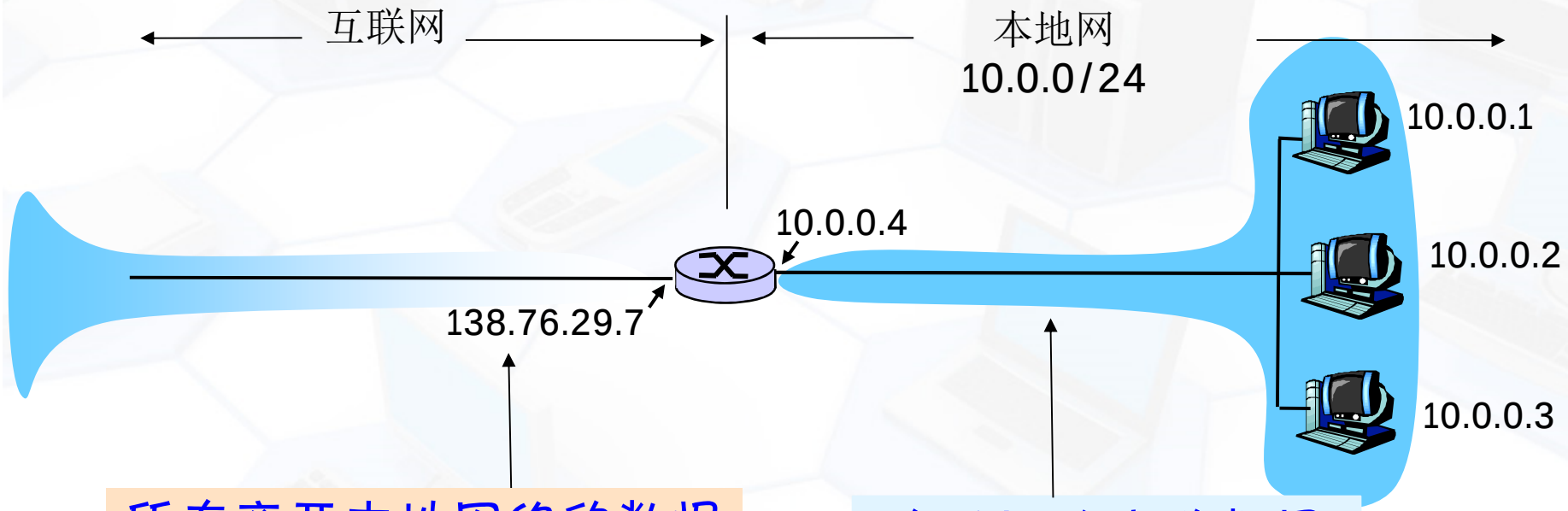


地址转换的基本过程

- 在只有一个全局IP地址的情况下，对于从Internet到内部网络上的分组，NAT设备无法知道分组所对应的内部主机，或者同一台内部主机上的应用



- 解决方法：使用TCP或者UDP端口号来区分，即NAT的依据不再仅仅是IP地址，还包括端口号

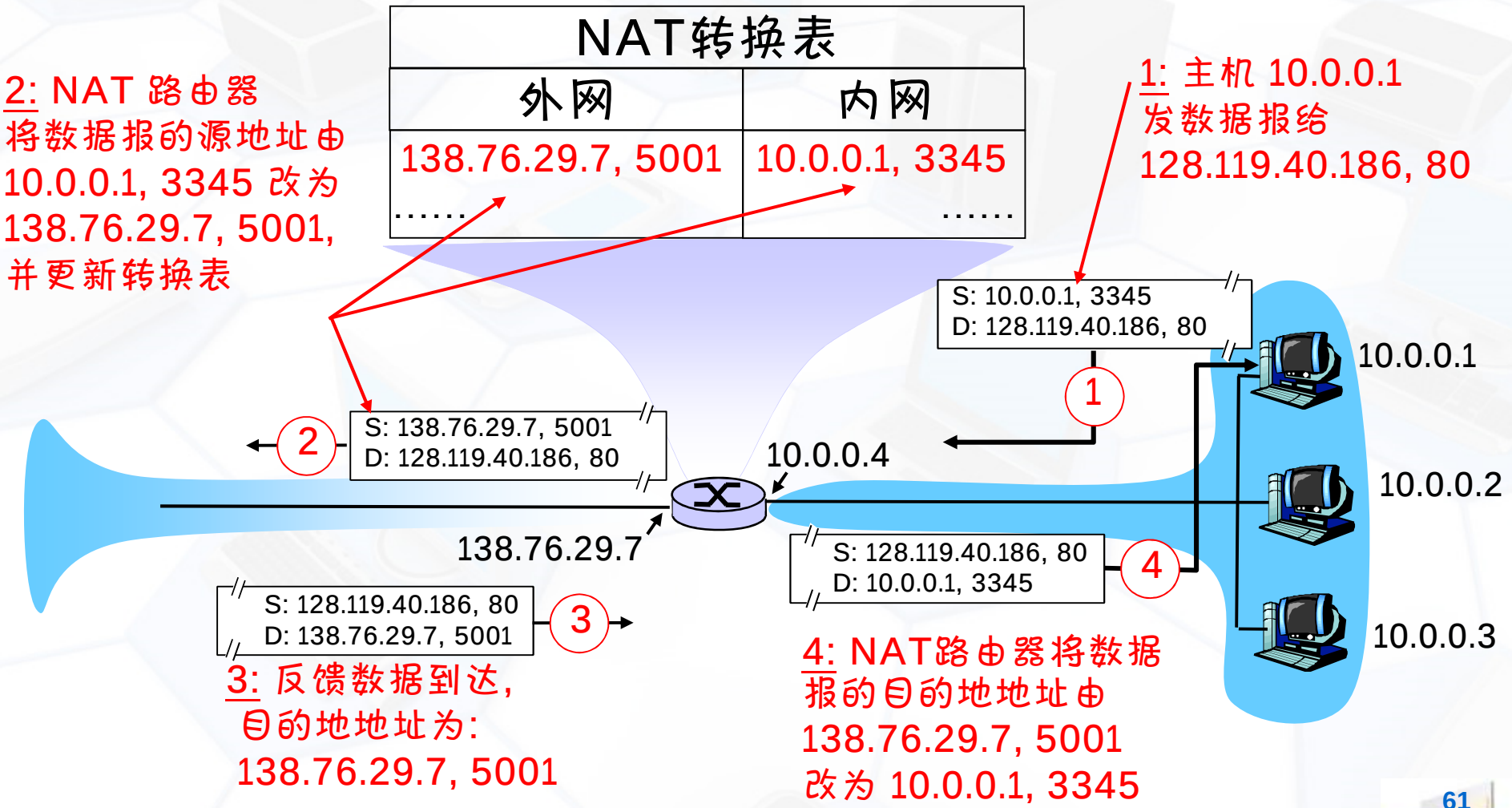


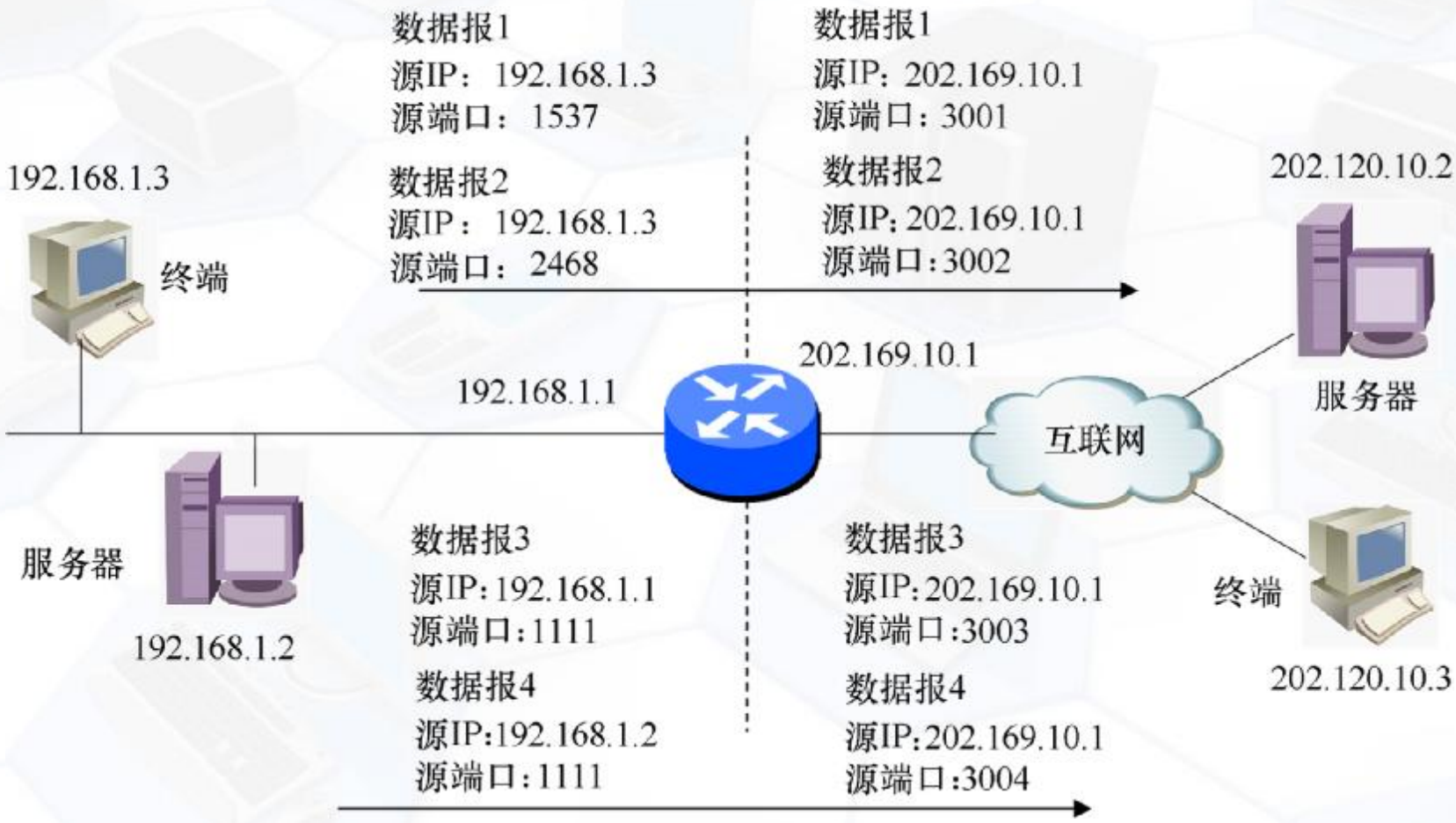
所有离开本地网络的数据报 都有相同的源NAT IP地址: 138.76.29.7, 不同的源端口号

在此网络中的数据报，其源地址都是10.0.0/24 网段，目的地则各异

3. 网络地址端口转换

- NAPT允许多个内部地址映射到同一个公有地址上，也可称为“多对一地址转换”或地址复用
- 多个带有内部地址的数据到达NAT服务器，其中有些数据来自同一个内部地址但有不同的源端口号，有些数据来自不同的内部地址但具有相同的源端口号
- 通过NAT映射，所有数据都被转换到外部地址，但每个数据都赋予了不同的源端口号，因而仍保留了报文之间的区别。当回应报文到达时，NAT进程仍能根据回应报文的目的地址和端口号来区别该报文应转发到的内部主机


NAT转换表
外网
内网
138.76.29.7, 5001
10.0.0.1, 3345
.....
.....
2: NAT 路由器
将数据报的源地址由
10.0.0.1, 3345 改为
138.76.29.7, 5001,
并更新转换表
1: 主机 10.0.0.1
发数据报给
128.119.40.186, 80
S: 10.0.0.1, 3345
D: 128.119.40.186, 80
1
10.0.0.1
10.0.0.2
10.0.0.3
10.0.0.4
S: 138.76.29.7, 5001
D: 128.119.40.186, 80
2
138.76.29.7
S: 128.119.40.186, 80
D: 10.0.0.1, 3345
4
S: 128.119.40.186, 80
D: 138.76.29.7, 5001
3
3: 反馈数据到达,
目的地地址为:
138.76.29.7, 5001
4: NAT路由器将数据
报的目的地地址由
138.76.29.7, 5001
改为 10.0.0.1, 3345


数据报1
源IP：192.168.1.3
源端口：1537
数据报1
源IP：202.169.10.1
源端口：3001
192.168.1.3
终端
数据报2
源IP：192.168.1.3
源端口：2468
数据报2
源IP：202.169.10.1
源端口：3002
202.120.10.2
服务器
202.169.10.1
192.168.1.1
互联网
服务器
192.168.1.2
数据报3
源IP：192.168.1.1
源端口：1111
数据报3
源IP：202.169.10.1
源端口：3003
终端
202.120.10.3
数据报4
源IP：192.168.1.2
源端口：1111
数据报4
源IP：202.169.10.1
源端口：3004

NAPT地址复用

3. NAT的优缺点

- 优点

为内部主机提供了“隐私”保护，实现了内部网络的主机通过该功能访问外部网络资源可能

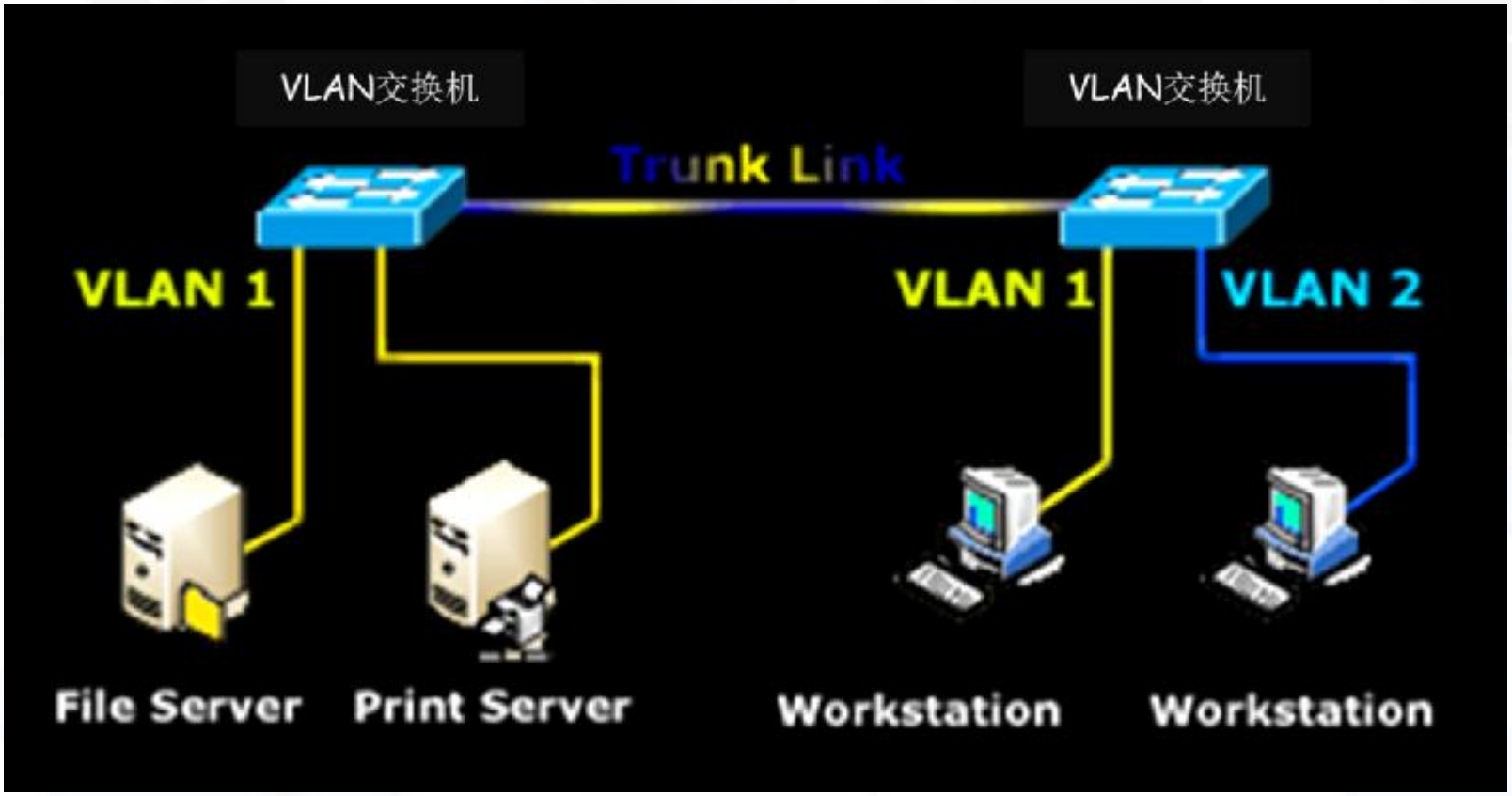
- 缺点

ü 需要对数据报文进行IP地址转换，所以所涉及的IP地址的数据报的报头不能被加密

ü 网络调试变得更加困难

ü 在链路的带宽低于10Mb/s速率时，地址转换对网络性能基本不构成影响；但当速率高于10Mb/s时，地址转换将对路由器性能产生一定影响

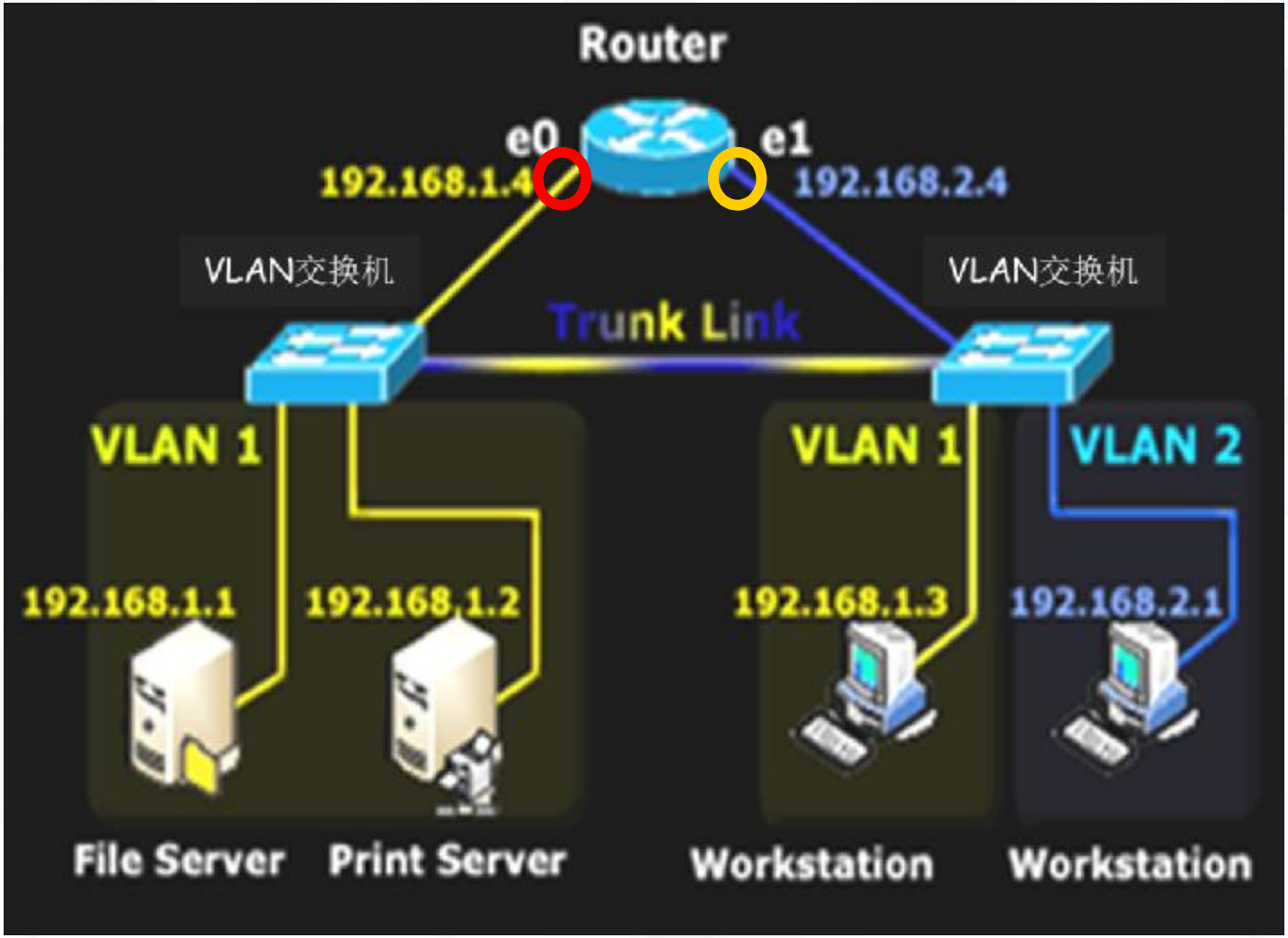
五、VLAN间的互通



需要互访的VLAN环境示意

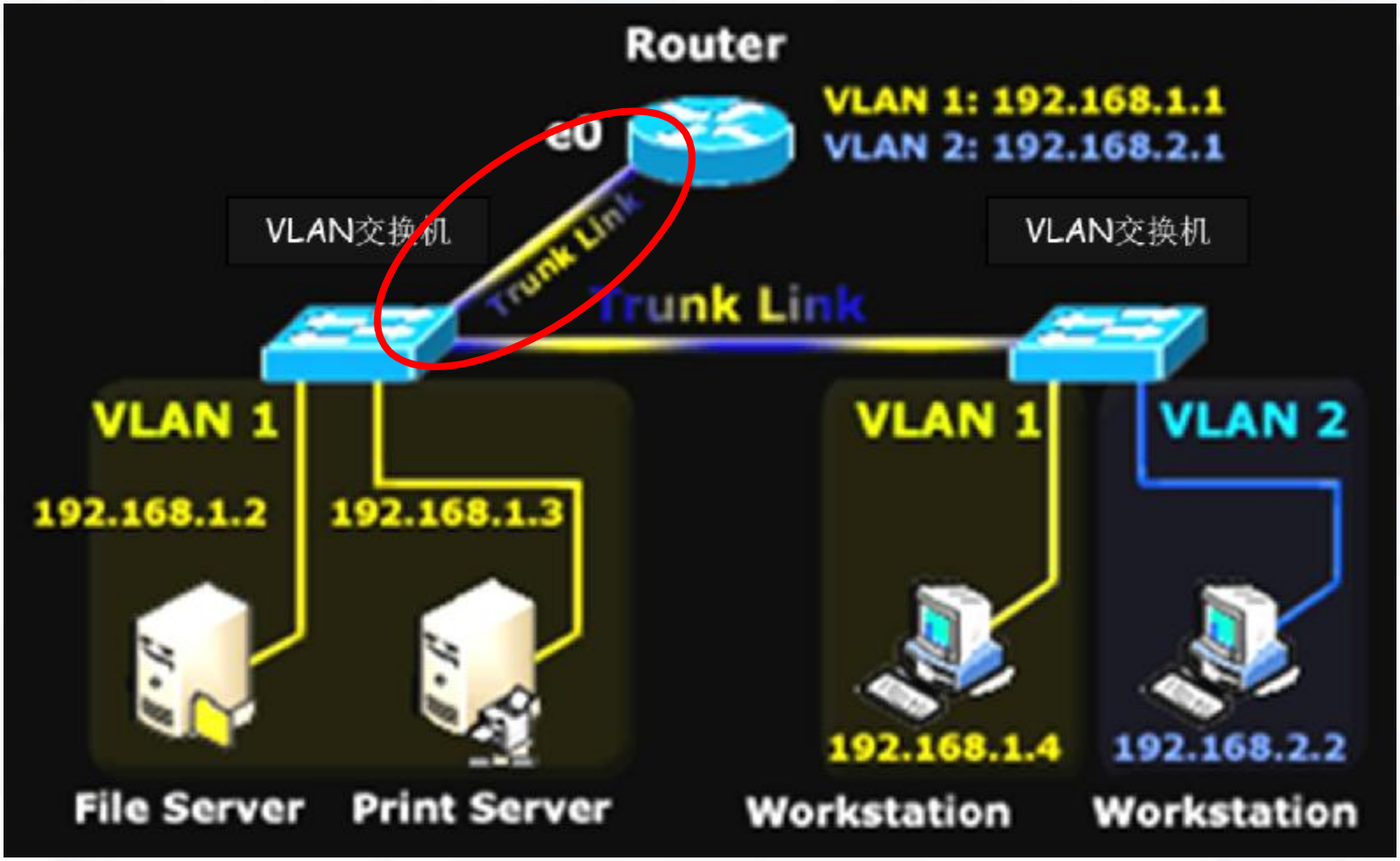
1. 互通方案

1）采用路由器



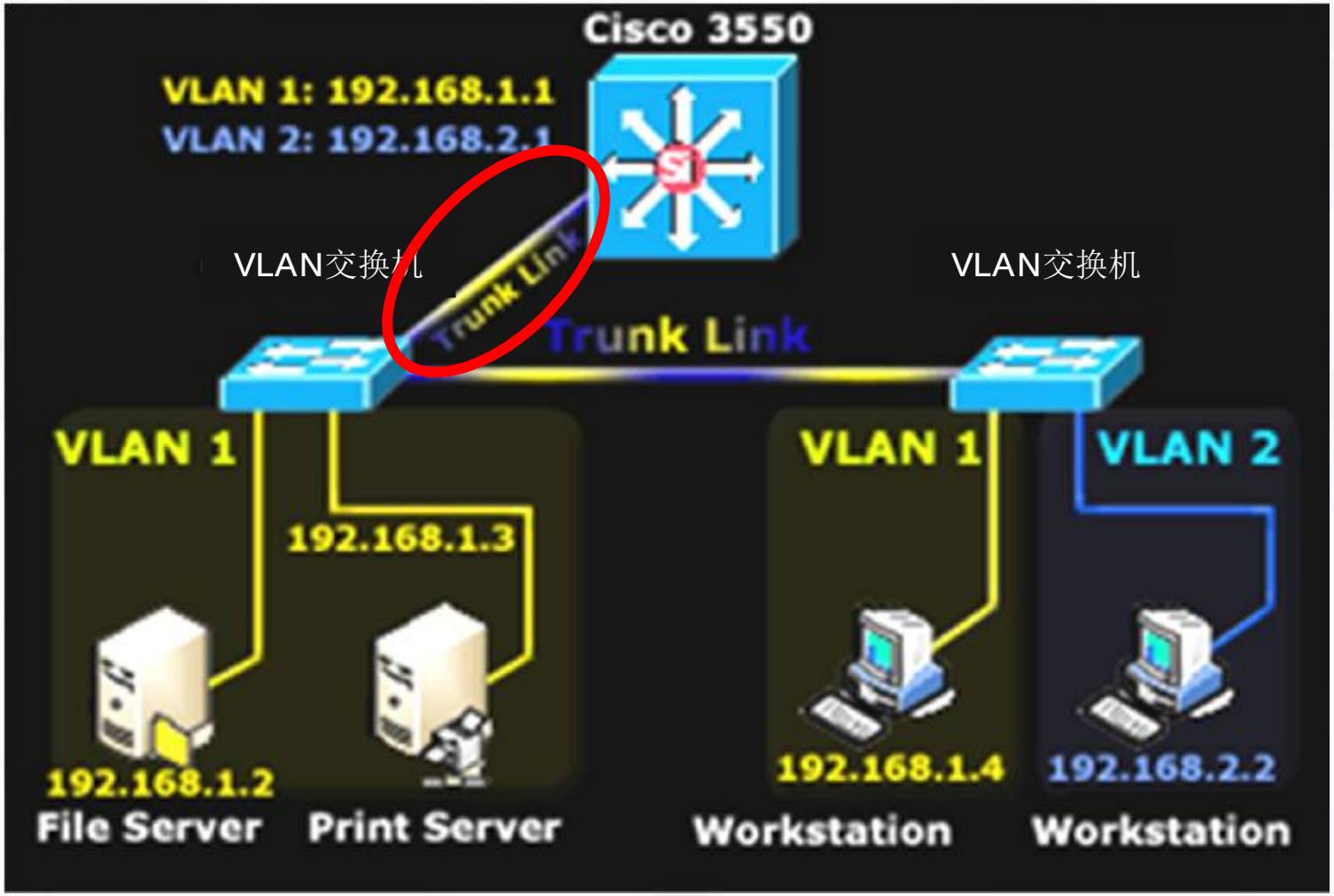
多端口路由解决方案示意

2） 采用支持主干以太链路的路由



使用以太主干链路的解决方案示意

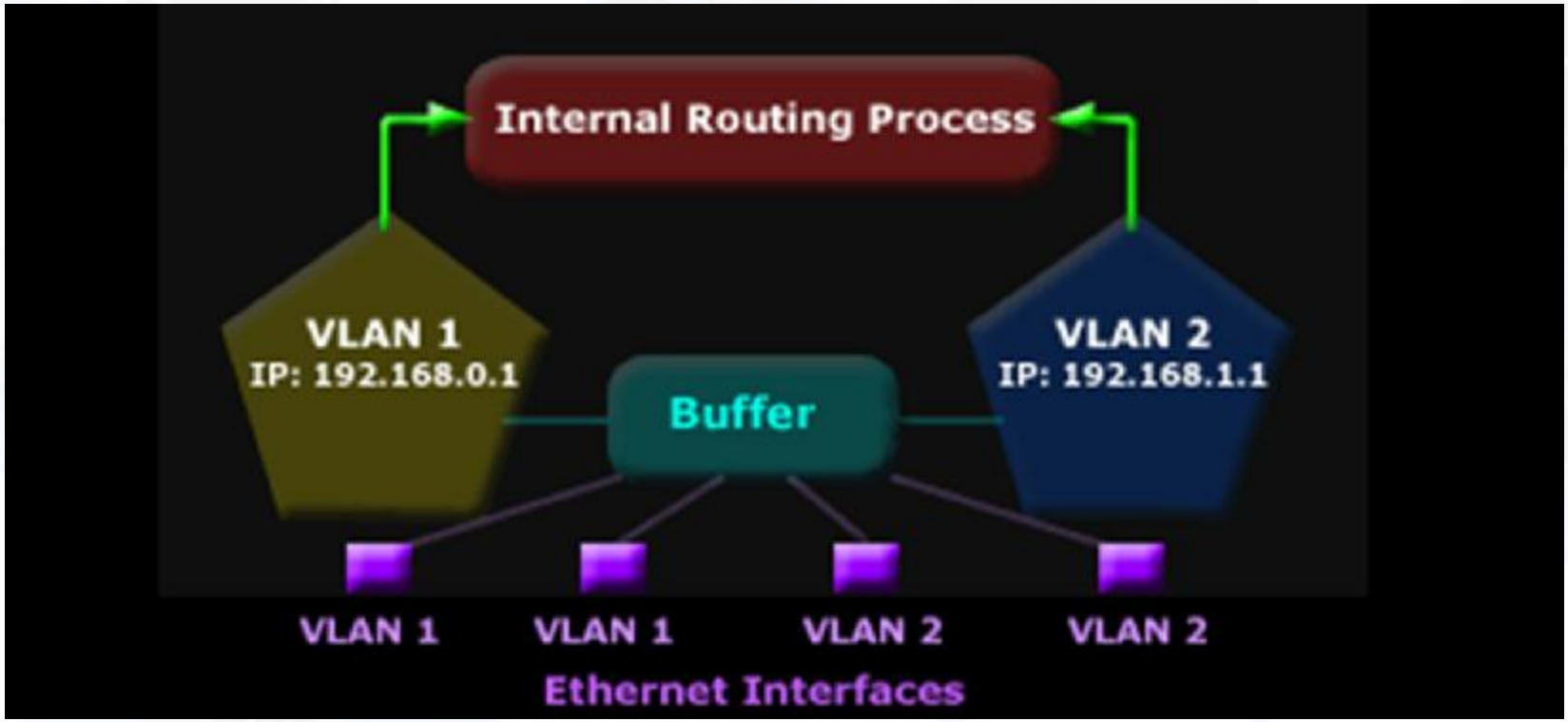
3） 采用三层交换



使用三层交换的解决方案示意

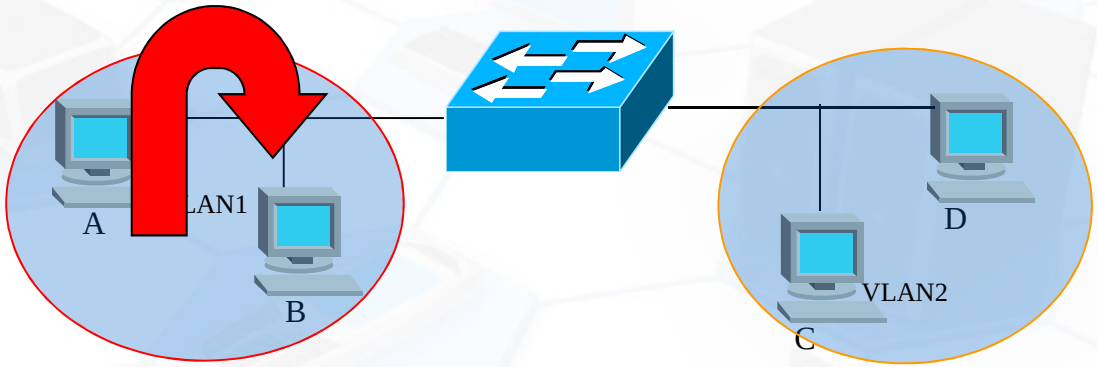
2. 三层交换原理

- 三层交换技术也称为IP交换
- 二层交换和三层路由的优势结合
- 利用三层协议中的信息来加强二层交换功能
- 不同于传统路由的软件驱动
- 三层交换削减了处理的协议数（仅对IP）
- 只完成交换和路由功能，限制特殊服务
- 使用专用集成电路（ASIC）构造更多功能，减少软件的运行


Internal Routing Process
VLAN 1
IP: 192.168.0.1
VLAN 2
IP: 192.168.1.1
Buffer
VLAN 1
VLAN 1
VLAN 2
VLAN 2
Ethernet Interfaces

三层交换机结构示意

3. 三层交换的处理过程



交换机上划分了两个VLAN，在VLAN1，VLAN 2 上配置了路由接口用来实现vlan1 和 vlan 2 之间的互通

#A 和B 之间的互通（以A 向B 发起ping 请求为例）：

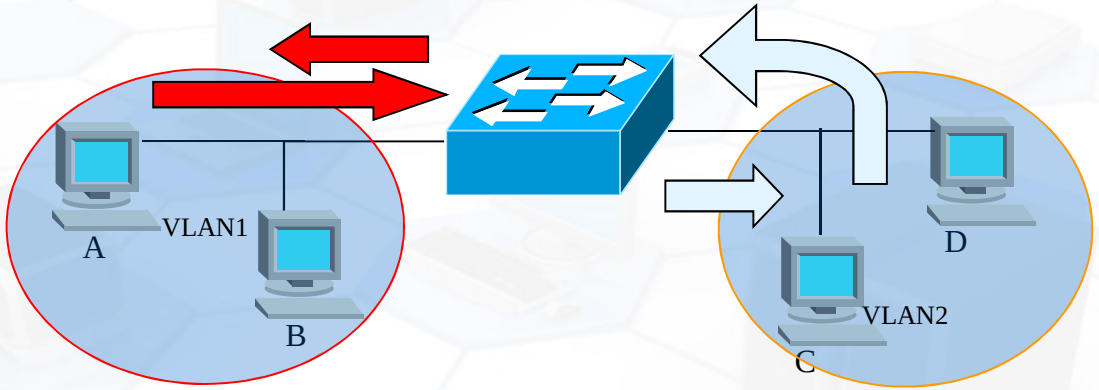
1）A 检查报文的目的IP 地址，发现B和自己在同一个网段

2）A向B发ARP请求报文，该报文在VLAN1 内广播

3）B对A发ARP回应报文

4）A向B发ICMP request

5）B向A发ICMP reply



#A 和C 之间的互通（以A 向C 发起ping 请求为例）：

1）A 检查报文的目的IP 地址，发现C和自己不在同一个网段

2）A向交换机（int vlan 1）发ARP 请求报文，该报文在VLAN1 内广播

3）网关向A发ARP 回应报文

4）A向交换机发ICMP request（目的MAC 是 int vlan 1 的MAC，源MAC 是A 的MAC，目的IP 是C，源IP 是A）

5）交换机收到报文后判断出是三层的报文，检查报文的目的IP 地址，发现是在自己的直连网段

6）交换机在向（int vlan 2）C发ARP 请求报文，该报文在VLAN2 内广播

7）C向交换机（int vlan 2）发ARP 回应报文

8）交换机向（int vlan 2）的C发ICMP request （目的MAC 是 C 的MAC，源MAC 是 int vlan 2 的MAC，目的IP 是C，源IP 是A）

9）后续步骤与4）相同

一个报文从端口进入后，交换设备如何来区分二层报文还是三层报文？

- 从A 到B 的报文由于在同一个VLAN 内部，报文的目的MAC 地址将是主机B 的MAC 地址，而从A 到C 的报文，要跨越VLAN，报文的目的MAC 地址是设备虚接口VLAN1 上的MAC 地址

- 因此交换机区分二三层报文的标准就是看报文的目的MAC地址是否等于交换机虚接口上的MAC 地址

- 三层交换的硬件处理流程主要依赖的表项是二层MAC地址表和三层的ip fdb 表
- 两个表中用于保存转发信息，在转发信息比较全的情况下，报文的转发和处理全部由硬件来完成处理，不需要软件的干预
- 两个表的功能是独立的，没有相互的关系，一个报文只要一进入交换机，硬件就会区分出这个包是二层还是三层

MAC地址表举例：show mac all

MAC ADDR	VLAN ID	STATE	PORT / INDEX	AGING TIME(s)
0000.21cf.73f4	1	Learned	Ethernet0/19	266
0002.557c.5a79	1	Learned	Ethernet0/12	225
0005.5df5.9f64	1	Learned	Ethernet0/16	300

MAC 地址表是精确匹配的方式， 其中关键的参数是：Vlan ID, Port/ index

IP转发表举例：show ipfdb all

Ip Address	RtIf	Vtag	VTValid	Port	Mac	Status
10.11.83.77	2	2	Invalid	GigabitEthernet2/1	00-e0-fc-00-55-18	1
10.63.32.2	2	2	Invalid	GigabitEthernet2/1	00-e0-fc-00-55-18	1
10.75.35.106	2	2	Invalid	GigabitEthernet2/1	00-e0-fc-00-55-18	2

- 路由接口索引（RtIf）：该索引用来确定该转发表项位于哪个路由接口下面
- Vlan tag: 该值用来表明所处的VLAN，该VLAN 和路由接口是对应的
- Vlan tag 有效位（VTValid）：用来标识转发出去的报文中是否需要插入Vlan tag 标记
- 端口索引（Port）：用来说明该转发表项的出端口；
- 下一跳MAC：三层设备每完成一跳的转发，会重新封装报文中的MAC头，硬件ASIC 芯片一般依据这个域里面的数值来封装报文头

- 每次收到报文，ASIC 都会从其中提取出源和目的地址在MAC地址表或者 IP Fdb表中进行查找，如果地址在转发表中可以找到，则认为该地址是解析的，如果找不到，则认为该地址是未解析的
- 根据地址是源、还是目的，还可以有源解析、目的未解析等等的组合
- 对于二层未解析，硬件本身可以将该报文在VLAN 内广播，而对于三层报文地址的未解析报文硬件本身则不对该报文进行任何的处理，而产生CPU 中断，靠软件来处理

硬件处理方法总结

- 收到报文后，判断该报文是二或是三层报文，然后判断其中的源，目的地址是否已经解析，如果已解析，则硬件完成该报文的转发，如果是未解析的，则产生CPU 中断，靠软件来学习该未解析的地址

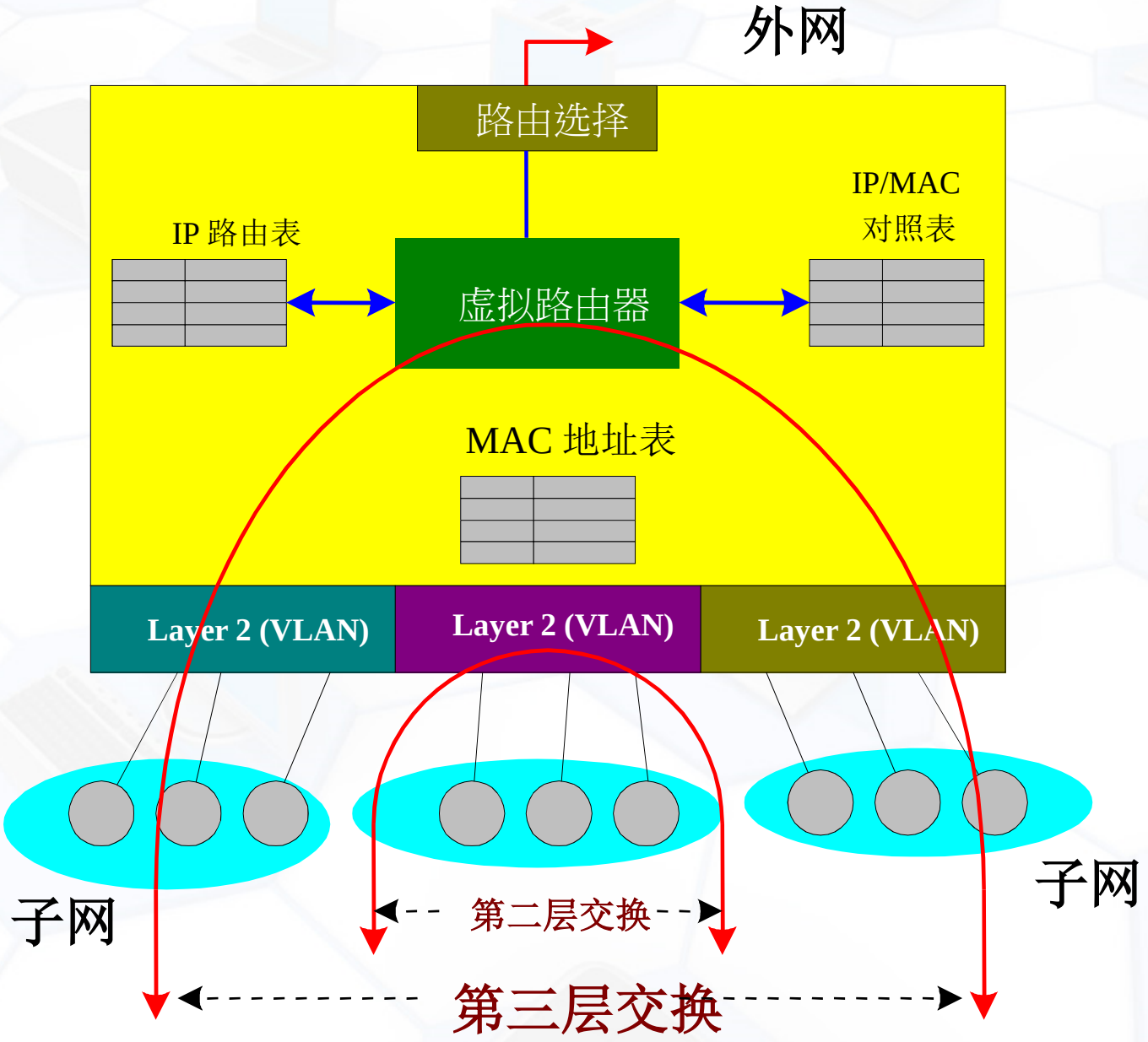
三层转发主要涉及的两个关键线程

1）报文转发线程主要根据地址学习线程生成的转发表（MAC table）信息来对报文进行转发，如果信息足够多，这个转发的过程全部由硬件来完成，如果信息不够，则会要求地址学习线程来进行学习，同时该报文硬件不能转发，会交给软件协议栈来进行转发

2）地址学习线程主要用来生成三层转发表（ipfdb table），它和二层的MAC 地址表类似，只不过里面的具体表项所代表的含义和所起的作用不同

- 在路由器等软件转发引擎中，每收一个报文都会去查路由表查下一跳，然后再查ARP 表找下一跳的MAC，而在三层交换机中，报文转发的时候不需要去查路由表和ARP 表

- 在三层转发流程中，第一个报文若硬件无法转发，要进行IP 地址的学习，同时为保证不丢包，该报文也由软件来进行转发。在学习完成后，第二，第三个报文以后就一直是由硬件来完成转发了。该过程可以套用“一次路由，多次交换”来形象的进行总结，在一次路由中，要利用路由表和ARP 表来学习IP 地址，和转发第一个报文，在以后的多次交换过程中，则只要利用ipfdb table 就可以了


外网
路由选择
IP/MAC
对照表
IP 路由表
虚拟路由器
MAC 地址表
Layer 2 (VLAN)
Layer 2 (VLAN)
Layer 2 (VLAN)
子网
子网
第二层交换
第三层交换

4. 三层交换机实例

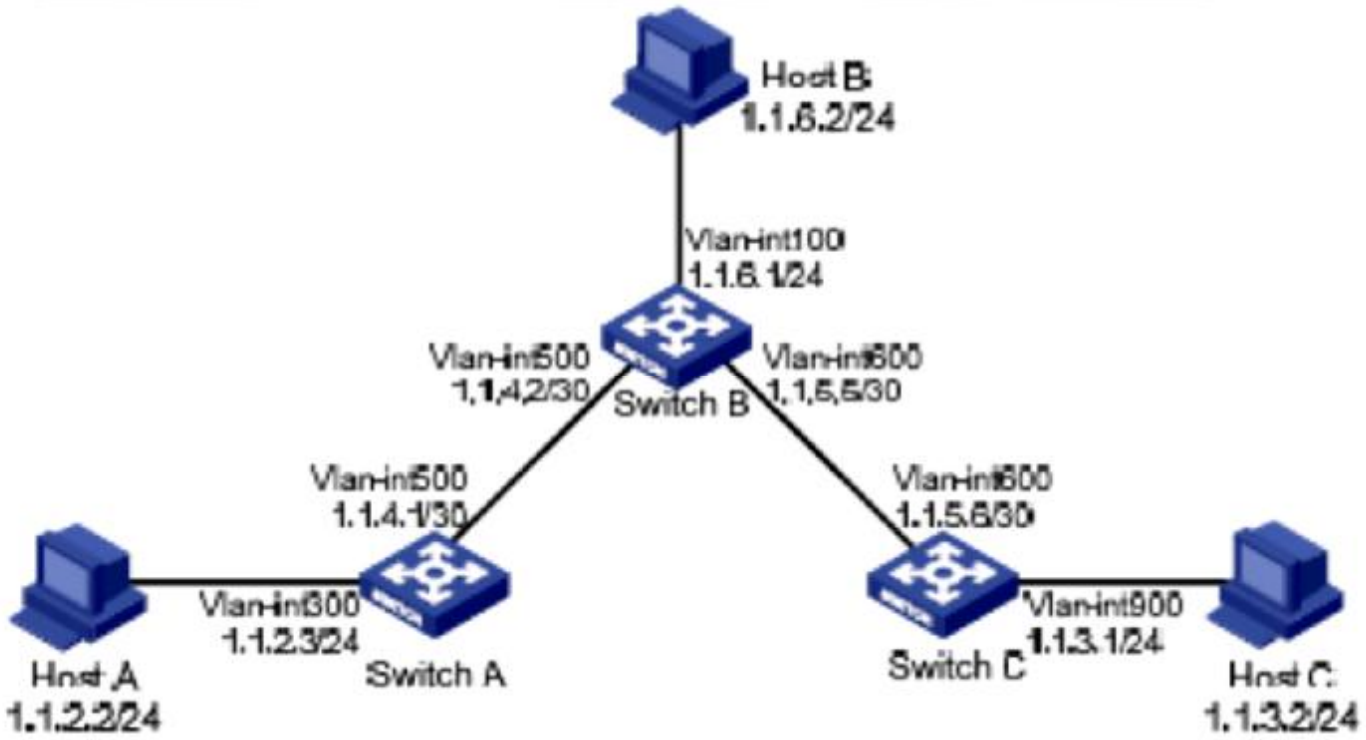
1) VLAN 接口

- 不同VLAN 间的主机不能直接通信，需要通过路由器或三层交换机等网络层设备进行转发
- 三层交换机提供VLAN 接口实现对报文进行三层转发的功能
- VLAN 接口是一种三层模式下的虚拟接口，主要用于实现VLAN 间的三层互通，它不作为物理实体存在于设备上
- 每个VLAN 对应一个VLAN 接口，该VLAN 接口可以为本VLAN 内端口收到的报文进行网络层转发操作
- 由于VLAN 能够隔离广播域，因此每个VLAN 也对应一个IP网段，VLAN 接口将作为该网段的网关对需要跨网段的报文进行基于IP 地址的三层转发

配置	命令	说明
进入系统视图	**system-view**	-
创建VLAN接口并进入VLAN接口视图	**interface Vlan-interface** *vlan-interface-id*	必选 如果该VLAN接口已经存在，则直接进入该VLAN接口视图
配置VLAN接口的IP地址	**ip address** *ip-address* { *mask* | *mask-length* } [**sub**]	可选 缺省情况下，没有配置VLAN接口的IP地址
为VLAN接口指定一个描述字符串	**description** *text*	可选 缺省情况下，VLAN接口的描述字符串为该VLAN接口的接口名，如“Vlan-interface1 Interface”

VLAN接口的相关命令

2） 三层交换的静态路由配置



交换机各接口及主机的IP 地址和掩码如图所示，采用静态路由，使任意两台主机之间都能互通

2） 三层交换的静态路由配置（续）

在Switch A 上配置缺省路由

[SwitchA> system-view

[SwitchA] ip route-static 0.0.0.0 0.0.0.0 1.1.4.2

在Switch B 上配置两条静态路由

<SwitchB> system-view

[SwitchB] ip route-static 1.1.2.0 255.255.255.0 1.1.4.1

[SwitchB] ip route-static 1.1.3.0 255.255.255.0 1.1.5.6

在Switch C 上配置缺省路由

<SwitchC> system-view

[SwitchC] ip route-static 0.0.0.0 0.0.0.0 1.1.5.5

配置主机

配置Host A 的缺省网关为1.1.2.3，Host B 的缺省网关为1.1.6.1，Host C 的缺省网关为1.1.3.1

2） 三层交换的静态路由配置（续）

查看配置结果，显示各交换机的IP 路由表

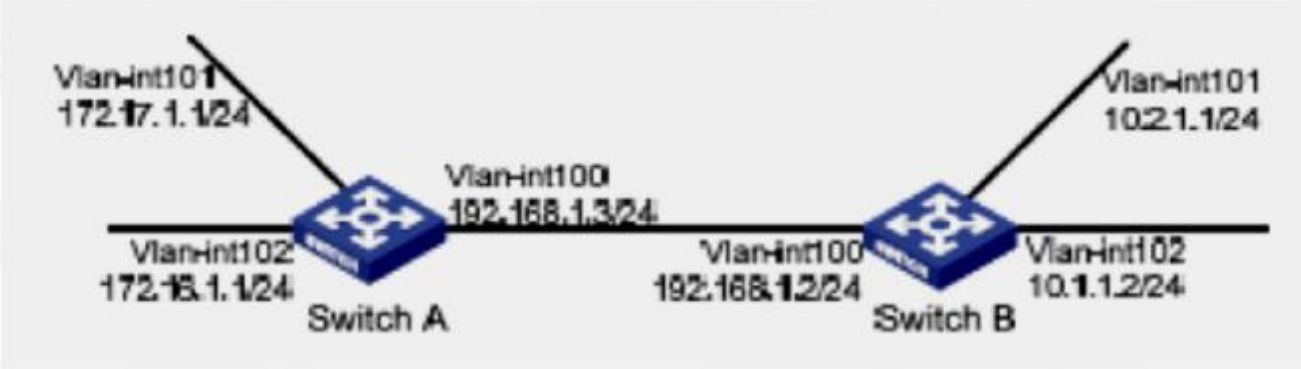
[SwitchA] display ip routing-table

```
Routing Tables: Public
         Destinations : 7        Routes : 7

Destination/Mask    Proto   Pre  Cost        NextHop         Interface

0.0.0.0/0           Static 60   0           1.1.4.2         Vlan500
1.1.2.0/24          Direct 0    0           1.1.2.3         Vlan300
1.1.2.3/32          Direct 0    0           127.0.0.1       InLoop0
1.1.4.0/30          Direct 0    0           1.1.4.1         Vlan500
1.1.4.1/32          Direct 0    0           127.0.0.1       InLoop0
127.0.0.0/8         Direct 0    0           127.0.0.1       InLoop0
127.0.0.1/32        Direct 0    0           127.0.0.1       InLoop0
```

3） 三层交换的RIP路由配置



要求在Switch A和Switch B的所有接口上使能RIP，并使用RIP-2 进行网络互连

3） 三层交换的RIP路由配置（续）

配置Switch A

<SwitchA> system-view

[SwitchA] rip

[SwitchA-rip-1] network 192.168.1.0

[SwitchA-rip-1] network 172.16.0.0

[SwitchA-rip-1] network 172.17.0.0

[SwitchA-rip-1] quit

配置Switch B

<SwitchB> system-view

[SwitchB] rip

[SwitchB-rip-1] network 192.168.1.0

[SwitchB-rip-1] network 10.0.0.0

[SwitchB-rip-1] quit

3） 三层交换的RIP路由配置（续）

查看Switch A 的RIP 路由表

```
Route Flags: R - RIP, T - TRIP
             P - Permanent, A - Aging, S - Suppressed, G - Garbage-collect
----------------------------------------------------------------------------

 Peer 192.168.1.2  on Vlan-interface100
      Destination/Mask        Nexthop      Cost     Tag    Flags    Sec
         10.0.0.0/8        192.168.1.2        1        0     RA       11
```

从路由表中可以看出，RIP-1 发布的路由信息使用的是自然掩码。

在Switch A 上配置RIP-2

- **[SwitchA] rip**
- **[SwitchA-rip-1] version 2**
- **[SwitchA-rip-1] undo summary**

```
[SwitchA] display rip 1 route
Route Flags: R - RIP, T - TRIP
             P - Permanent, A - Aging, S - Suppressed, G - Garbage-collect
------------------------------------------------------------------------------
 Peer 192.168.1.2  on Vlan-interface100
      Destination/Mask        Nexthop      Cost      Tag    Flags    Sec
         10.2.1.0/24        192.168.1.2       1        0     RA       16
         10.1.1.0/24        192.168.1.2       1        0     RA       16
```

从路由表中可以看出，RIP-2 发布的路由中带有更为精确的子网掩码信息。

使用RIP-2 以后路由表的情况

六、四层交换

- 第四层交换技术利用第三层和第四层包头中的信息来识别应用数据流会话
- 信息指TCP／UDP端口号、标记会话开始与结束的“SYN/FIN”位、以及源／目的IP地址
- 第四层交换机可以利用此信息做出向何处转发会话传输流的决定
- 第四层交换技术从头至尾跟踪和维持各个会话，根据会话和应用层信息做出转发决定

- 用户的请求可以根据不同的规则被转发到“最佳”的服务器上。第四层交换技术是用于传输数据和实现多台服务器间负载均衡的理想机制

- 具有第四层功能的交换机能够起到与服务器相连接的“虚拟IP”(VIP)前端的作用。每台服务器和支持单一或通用应用的服务器组都配置一个VIP地址，该VIP地址被发送出去并在域名系统上注册

- 在发出一个服务请求时，第四层交换机通过判定TCP开始，来识别一次会话的开始。然后它利用复杂的算法来确定处理这个请求的最佳服务器。确定之后，交换机就将会话与一个具体的IP地址联系在一起，并用该服务器真正的IP地址来代替服务器上的VIP地址

- 每台第四层交换机都保存一个与被选择的服务器相配的源IP地址以及源TCP 端口相关联的连接表。第四层交换机向这台服务器转发连接请求。
- 在使用第四层交换的情况下，接入可以与真正的服务器连接在一起来满足用户制定的规则，诸如使每台服务器上有相等数量的接入或根据不同服务器的容量来分配传输流
- 单功能负载均衡产品可以每秒连接400到800个接入。而基于专用负载均衡功能硬件、同时具有第二层和第四层功能的产品，接入速度则超过了每秒10万次
- 第四层交换中的关键问题是如何确定传输流转发给哪台最可用的服务器

- 目前做出负载均衡决定时，根据所需负载均衡的颗粒度，第四层交换将应用会话分配到服务器上的方法包括：求权数最小接入的简单加权循环、测量往返时延和服务器自身的闭合环路反馈等等
- 第四层交换机在形式和功能上不同于专用负载均衡器，它能够支持100Mbps或千兆的速率
- 第四层交换除了负载均衡功能外还支持如基于应用类型和用户ID的传输流控制功能
- 采用多级排队技术，第四层交换机可以根据应用来标记传输流以及为传输流分配优先级