OSPF协议

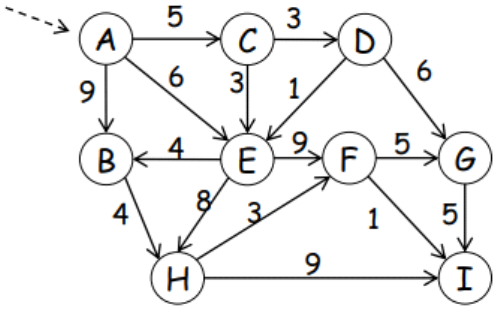
2019年5月15日 1:30

概述

- 链路状态(link state)路由算法:

  (1) 利用最短路径算法(例如：Dijkstra最短路经算法)求出一个节点(源节点)到所有其它节点的最短路经。

  (2) 利用这些最短路经上的下一个节点作为下一跳得到源节点的转发表(路由表)。



节点A的转发表

目的	下一跳	距离
B	B	9
C	C	5
D	C	8
E	E	6
F	B	14
G	C	14
H	B	13
I	E	15 16

节点A的链路状态：<AB,9> <AC,5> <AE,6>

- OSPF 协议(Open Shortest Path First)采用链路状态路由算法，它可能是在大型企业中使用最广泛的内部网关协议 (IGP)
- OSPF协议的简单描述：

  (1) 周期性地收集链路状态，并扩散给AS中的所有路由器;

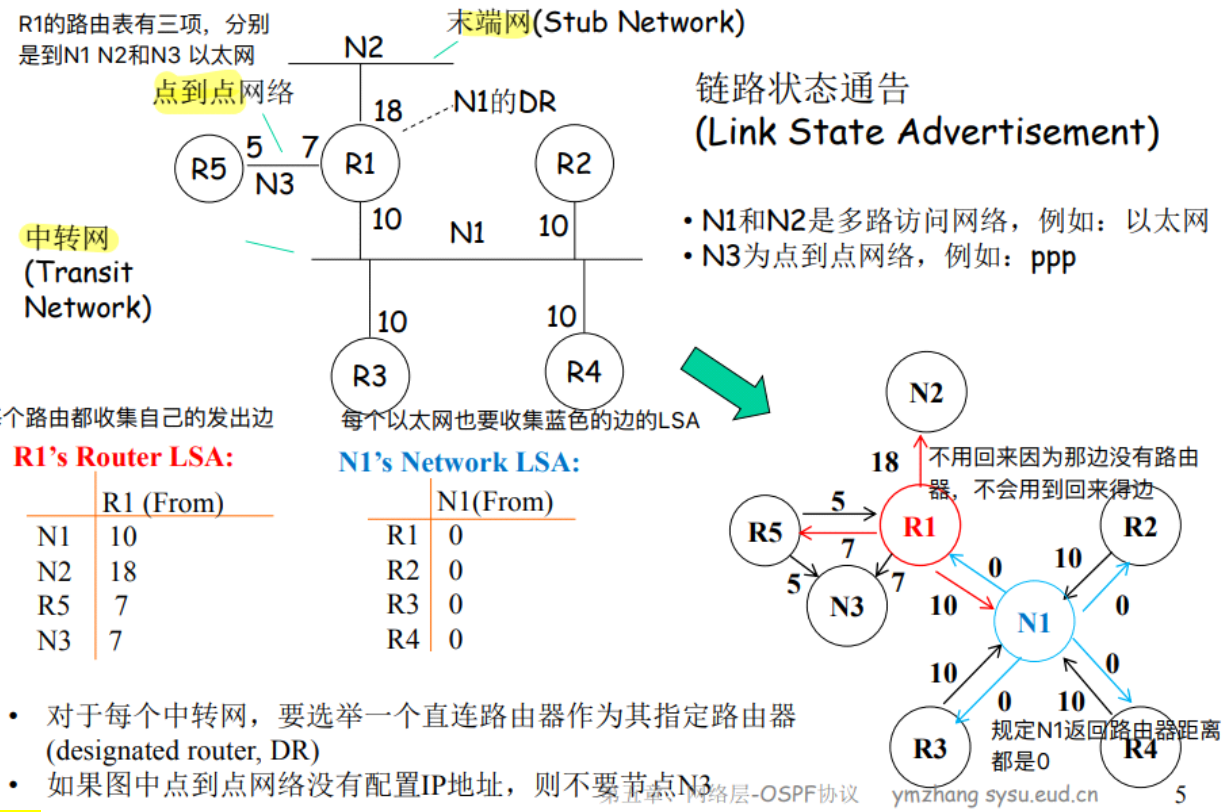
  (2) 用收到的链路状态建立整个AS的拓扑结构图;

  (3) 利用Dijkstra算法计算到AS中所有网络的最短路径;

  (4) 利用这些路径上的下一跳建立路由表。

如何利用把整个网络(AS)转化为AS的拓扑结构图？

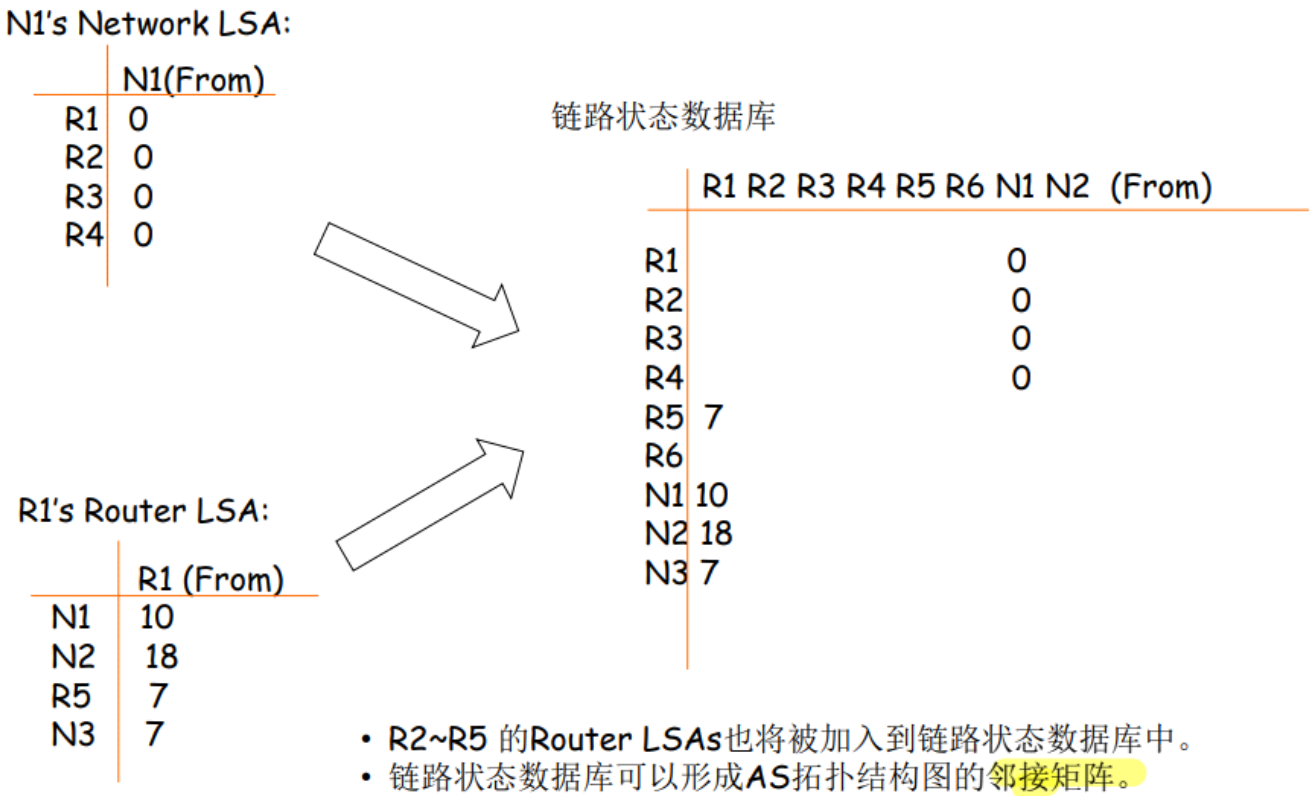
把网络转变为图



- 对于每个中转网，要选举一个直连路由器作为其指定路由器(designated router, DR)
- 如果图中点到点网络没有配置IP地址，则不要节点N3

中转网：多路访问网，而且有多个OSPF路由器和这个直连网直接相连

环回网络：环回网路就是一个末端网



- R2~R5 的Router LSAs也将被加入到链路状态数据库中。
- 链路状态数据库可以形成AS拓扑结构图的邻接矩阵。

RID：每个路由器的唯一标识，用来区分不同的路由器。路由器会选取它所有loopback接口上数值最高的IP地址。若没配置loopback接口的IP地址，那么路由器就在所有活动物理端口中选取一个数值最高的IP地址作为路由器的Router ID

DR：运行OSPF的路由器通过与邻居路由器建立邻接关系，互相传递链路状态信息。如果每两个路由器之间都要建立邻接关系，那么就会构成n(n-1)/2个邻接关系，这时就有些混乱了，而且浪费了很多不必要的网络资源。为了可以避免这些问题的发生，可以在该网段上选举一个指定路由器

（Designated Router，DR）。由DR和网络中的其他路由器建立邻接关系，并负责将网段上的变化告知它们。

BDR：为了实现冗余，当DR失效时，需要有一个新的DR来接替DR的工作，这便是BDR（Backup Designated Router，BDR）。网络上所有的路由器将和DR和BDR同时形成邻接关系。ＤＲ和ＢＤＲ之间也将形成邻接关系。

自动选举：网段上router ID最大的路由器将被选举为DR，第二大的将被选举为BDR，这样的选举可能不是最佳的
手工选举DR和BDR：需要设置路由器的优先级，每台路由器的接口都有一个路由器优先级，优先级的大小范围是0~255,数值越大，优先级越高

OSPF开销

接口类型	带宽（bps）	OSPF开销
Ethernet	10M	10
Fast Ethernet	100M	1
Gigabit Ethernet	1G	1
T1	1.544M	64
E1	2.048M	48
56000 point-to-point network	56K	1785
19.2K point-to-point network	19.6K	5208

- 开销的实际计算方法： reference-bandwidth/bandwidth (Mbps)
  开销必须为大于0的整数， reference-bandwidth的默认值为100 。
- 修改reference-bandwidth的方法：
  #ospf auto-cost reference-bandwidth 1000

在OSPF的配置过程中（多在汇聚和核心交换机上出现），往往会看到下面的配置语句：auto-cost reference-bandwidth 10000 。那这句话起到什么作用呢?

对于路由协议而言，常用的有基于距离矢量的路由协议（如RIP）和链路状态路由协议（如OSPF）。相对而言，距离矢量路由协议使用的是跳数值来选择最优路径的，这种算法比较容易理解，每经过一个路由器加一跳；对于相同目的地的，取跳数最少的做为最优路由加入到路由表。

而对于OSPF协议而言，它是一种链路状态路由协议，使用的度量值是一种COST值，其由带宽、时延、可靠性等共同决定。一般来说是用cost=10^8/bandwidth来计算的（其中bandwidth需要划算成以byte为单位）。

例如对100M链路而言，其cost=10^8/(100M*10^6byte)=1。此时便出现一个问题，对于1000M链路而言，就会出现cost=0.1的情况，而对于10000M链路会出现cost=0.01的情况。于是auto-cost reference-bandwidth就被引入了进来。它的出现通过人为的改变参照值实现可以在1000M或者更高的链路上实现OSPF 的cost值自动计算。

OSPF数据包

承载在IP数据包内，使用协议号89

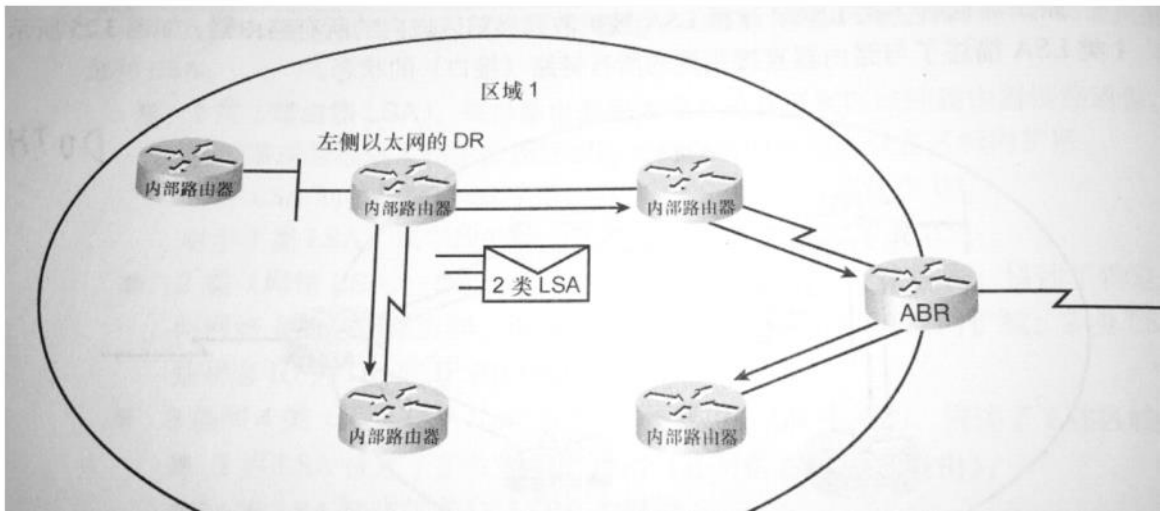
OSPF的包类型

OSPF的包类型	描　述
Hello包	用于发现和维持邻居关系，选举DR和BDR
数据库描述包（DBD）	用于向邻居发送摘要信息以同步链路状态数据库
链路状态请求包（LSR）	在路由器收到包含新信息的DBD后发送，用于请求更详细的信息
链路状态更新包（LSU）	收到LSR后发送链路状态通告（LSA），一个LSU数据包可能包含几个LSA
链路状态确认包（LSAck）	确认已经收到LSU，每个LSA需要被分别确认

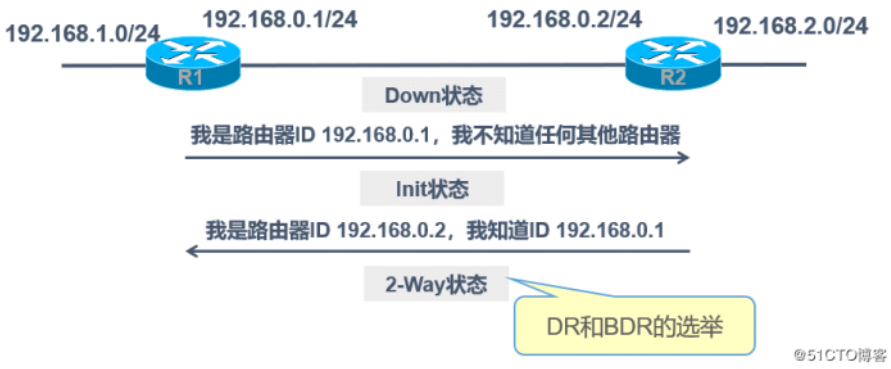
（2）网络 LSA（Network LSA）：2 类 LSA 是 DR 为区域中每个中转的广播网络或 NBMA 网络生成的。中转网络至少与两台 OSPF 路由器直接相连，诸如以太网等多路访问网络就属于中转网络。2 类 LSA 列出了构成中转网络的所有路由器（包括 DR 本身）和链路的子网掩码。中转链路的 DR 负责通告 2 类 LSA，2 类 LSA 随后被扩散到区域内所有的路由器，2 类 LSA 不会跨越区域边界进行传播（如下图所示）。其链路状态 ID 为通告它的 DR 的 IP 接口地址。使用命令 show ip ospf database network 可以查看网络 LSA 通告的信息。请注意，和路由器 LSA 不同，网络 LSA 中没有度量字段。

关于网络 LSA 的其他解释：

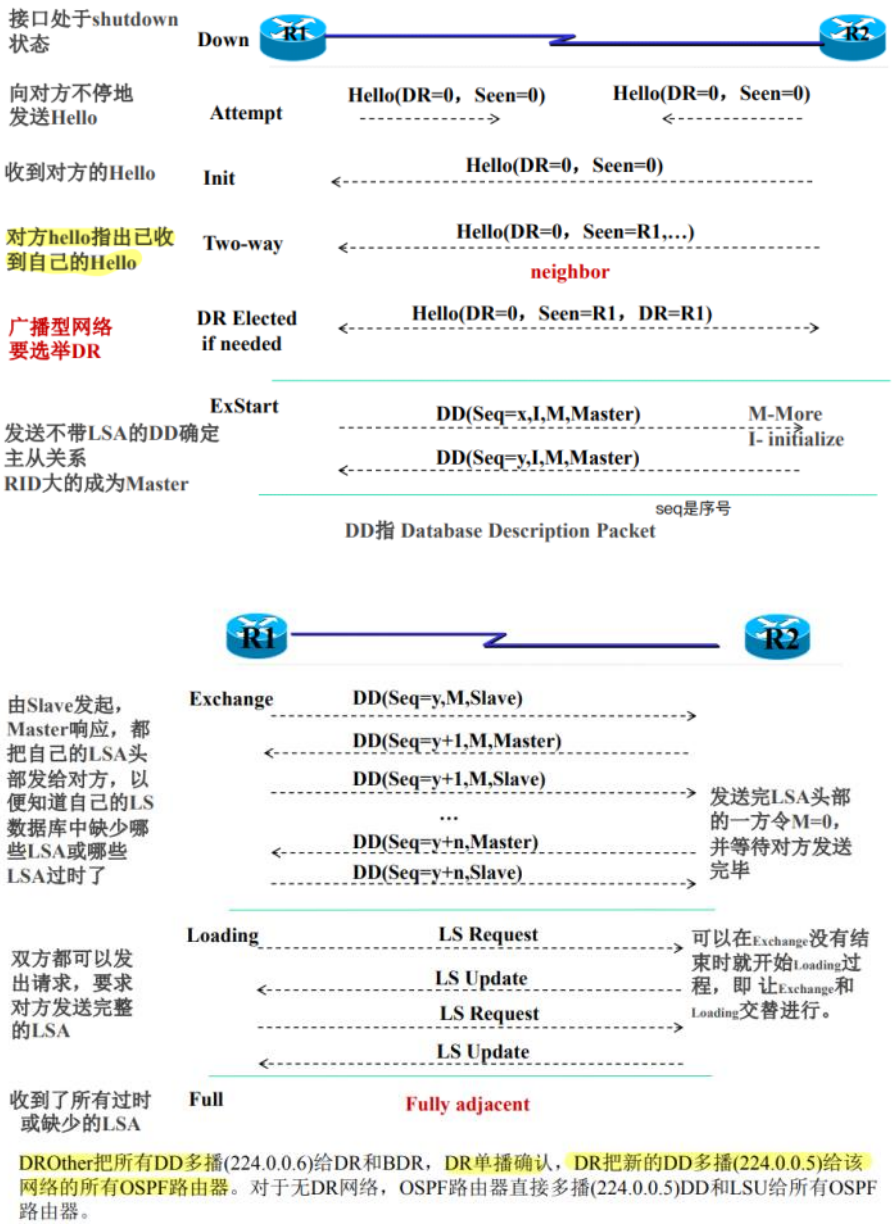
DR 路由器可以看作一个“伪”节点，或是一个虚拟路由器，用来描绘一个多路访问网络和与之相连的所有路由器。从这个角度来看，一条网络 LSA 通告也可以描绘一个逻辑上的“伪”节点，就像一条路由器 LSA 通告描绘一个物理上的单台路由器一样。网络 LSA 通告列出了所有与之相连的路由器，包括 DR 路由器本身。就像路由器 LSA 一样，网络 LSA 也仅仅在产生这条网络 LSA 的区域内部进行泛洪扩散。



OSPF启动的第一个阶段是使用Hello报文建立双向通信的过程



数据库同步



有DR的先发给DR再统一多播该网络，因为多路访问网中的路由器只和DR和BDR建立邻居关系，如果没有就直接多播出去给建立相邻关系的其他OSPF路由器

路由器ID

- OSPF协议采用路由器ID(Router ID，RID)标识每一个路由器。
- 路由器ID由以下方法得到：
  ① 使用直接配置的RID（#router-id id）。
  ② 所有活动环回接口中最大的IP地址。
  ③ 所有活动物理接口中最大的IP地址。
- 除非路由器重启、所选接口故障或关闭或IP地址改变、重新执行了router-id命令，RID将保持不变。

指定路由器*

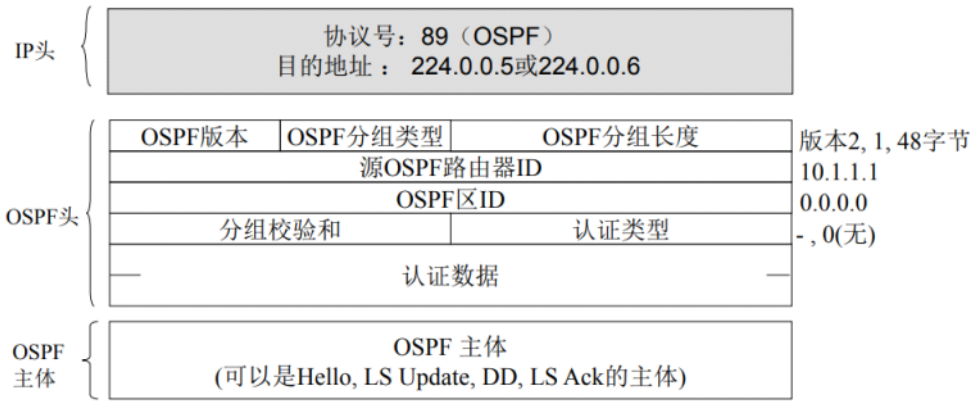
- 当多路访问网络重启时，选择DR的过程就开始了。在等待时间结束(Wait Time, Dead Interval，40秒)时，带有最高和次高优先权的路由器分别成为DR和 BDR(Backup DR)。如果优先权相同，RID更大的成为DR，次大的成为BDR。
- 如果路由器不希望参与选举，则应该把优先权设置为0。如果优先权相同，具有更高RID的路由器成为DR。如果收到的Hello列出了DR(RID不是0.0.0.0)，路由器成为DR。
- 如果一个新的路由器在选举之后到达或者有路由器修改为更高的优先权，它也不可能抢占现存的 DR (或BDR)和变为DR(或BDR)。
- 当DR失效时，BDR成为DR，将开始一个新的选举过程来选出BDR。
- 一个多路访问网络中的OSPF路由器只与DR和BDR建立相邻关系。
- 收到一个LSA后，一个多路访问网络中的OSPF路由器将把它首先多播(224.0.0.6)给DR和BDR，然后 DR再把它多播 (224.0.0.5)给所有OSPF路由器。

LSA定时器*

- 每10秒(Hello Interval)向邻居发送一次Hello，4倍的hello interval (Dead Interval，40秒)没有收到邻居的Hello就认为邻居失效。
- 每30分钟会产生新的LSA，最小间隔时间为5秒。
- 每个LSA都有年龄字段(age)，发给邻居时被设置为0，在链路状态数据库中age会不断增长，增长到Max Age(默认为60分钟)时LSA被标记为失效。失效的LSA会被扩散到整个AS，令AS的所有路由器把该LSA从链路状态数据库中移除。
- 存储在链路状态数据库中的LSA每10分钟会被计算校验和，如果有错将被删除。
- 接收来自邻居的LSA的最小间隔时间为1秒。
- 计算最短路径的最小间隔时间为10秒。

OSPF分组格式

224.0.0.5 -- All OSPF Router
224.0.0.6 -- OSPF DR or BDR



OSPF分组类型：
1--Hello Packet
2--Database Description Packet
3--Link State Request Packet
4--Link State Update Packet
5--Link-State Acknowledge Packet

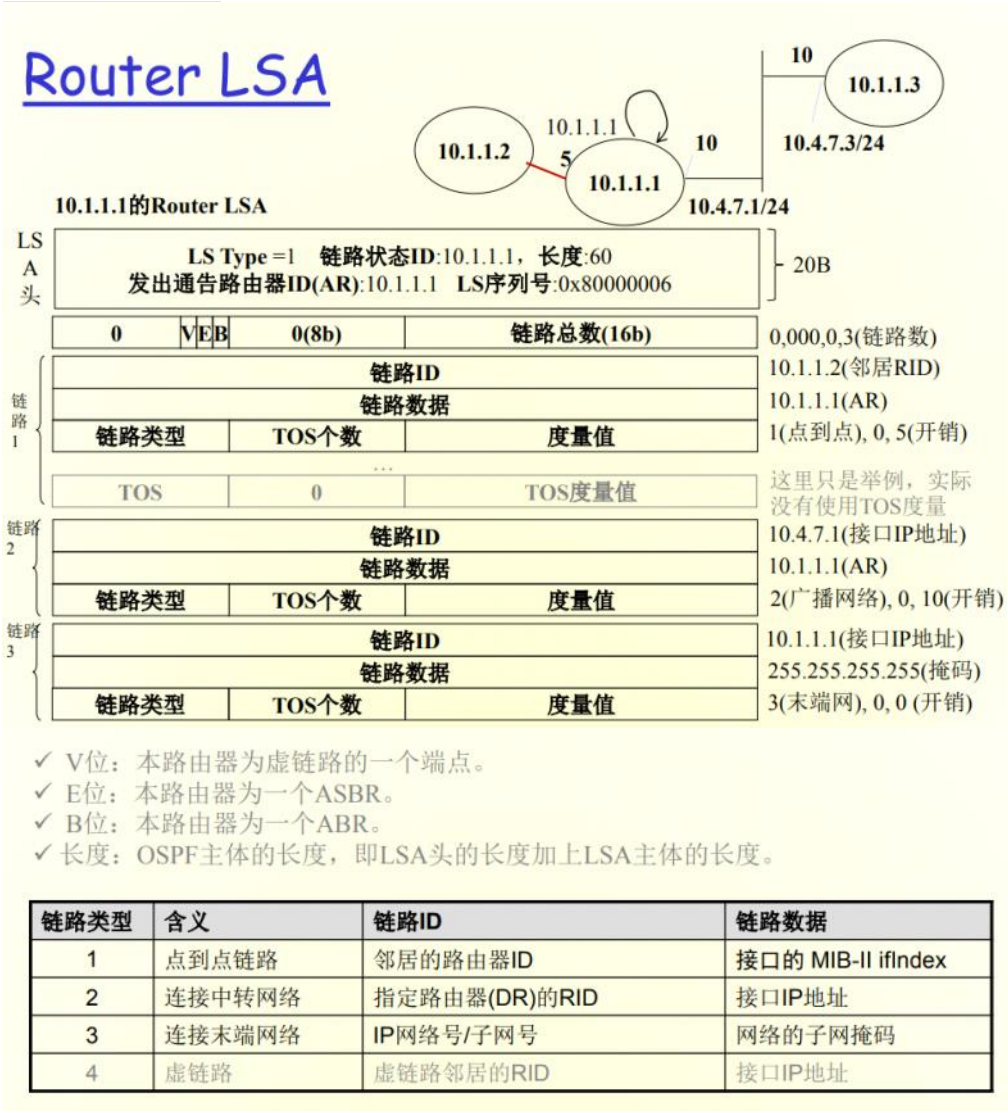
OSPF可以分区，至少要有一个区，即主干区（0.0.0.0）

LS Update分组

- 发出通告路由器ID为产生本分组的路由器ID。
- 链路状态ID用于区分不同的。Router LSA的链路状态ID与发出通告路由器ID相同，Network LSA的链路状态ID为网络号。
- LS类型：1--Router LSA　　2--Network LSA
  3--Network Summary LSA　　4--ASBR Summary LSA(E1和E2)
  5--AS-External LSA　　6--Group Membership LSA
  7--NSSA External LSA(N1和N2)

* type1~4都被限制在本区扩散。

OSPF包头有24个字节



✓ V位：本路由器为虚链路的一个端点。
✓ E位：本路由器为一个ASBR。
✓ B位：本路由器为一个ABR。
✓ 长度：OSPF主体的长度，即LSA头的长度加上LSA主体的长度。

链路类型	含义	链路ID	链路数据
1	点到点链路	邻居的路由器ID	接口的 MIB-II ifIndex
2	连接中转网络	指定路由器(DR)的RID	接口IP地址
3	连接末端网络	IP网络号/子网号	网络的子网掩码
4	虚链路	虚链路邻居的RID	接口IP地址

区域边界路由器**ABR**（Area Border Routers）：

该类路由器可以同时属于两个以上的区域，但其中一个必须是骨干区域（area 0）。

ABR用来连接骨干区域和非骨干区域，可以是实际连接，也可以是虚连接。

自治系统边界路由器**ASBR**（AS Boundary Routers）

与其他AS交换路由信息的路由器称为ASBR。 使用了多种路由协议！

只要一台OSPF路由器引入了外部路由的信息，他就称为了ASBR，

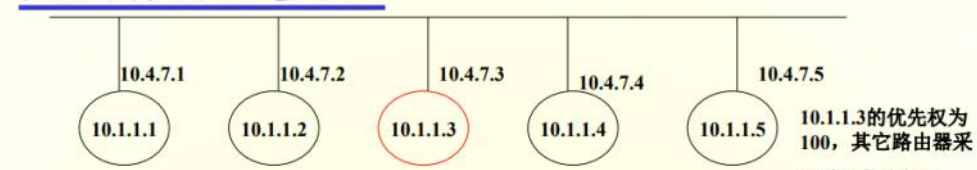
它有可能是ABR，区域路由器，不一定位于AS边界。

路由聚合

路由聚合是指ABR或ASBR将具有相同前缀的路由信息聚合，只发布一条路由到其它区域。

AS被划分成不同的区域后，区域间可以通过路由聚合来减少路由信息，减小路由表的规模，提高路由器的运算速度。

Network LSA



10.1.1.3的优先权为100，其它路由器采用默认优先权1。

网络10.4.7.0/24的Network LSA

LSA头	LS Type：2　链路状态ID：10.4.7.3 (接口IP地址),长度:44 发出通告路由器ID：10.1.1.3　LS序列号：0x80000010	20B
	子网掩码	255.255.255.0
	相连的路由器1的RID	10.1.1.1
	相连的路由器2的RID	10.1.1.2
	相连的路由器3的RID	10.1.1.3
	相连的路由器4的RID	10.1.1.4
	相连的路由器5的RID	10.1.1.5

由指定路由器发出，包括该网络的与指定路由器所有完全相邻的所有路由器。

Dijkstra最短路经算法

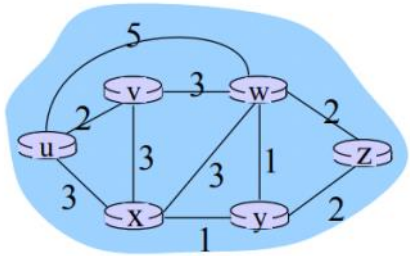
Notation:

- c(x,y): 从x到y的链路开销; 如果不是直接邻居，则为∞
- D(v): 从源节点到目的节点v的当前路经开销。
- p(v): 从源节点到目的节点v的路经上最靠近v的上一个节点。
- N': 已经知道最短路径的节点集合

```
1  Initialization:
2   N' = {u}
3   for all nodes v
4     if v adjacent to u
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'
```

举例：

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					



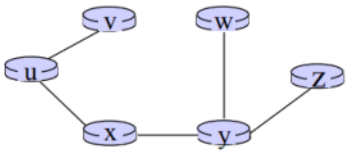
算法复杂性: n个节点

- 每次循环需要检查所有不在N中的节点 $n(n+1)/2$次比较: $O(n^2)$
- 更有效地算法: $O(n\log n)$

震荡的可能性:

- 例如，链路开销=实际流量大小。

从u出发的最短路径树:



u的转发表:

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

OSPF的特点

- 所有的OSPF消息都要认证 (防止恶意入侵)。
- 路由表中允许多个相同开销的路经存在(RIP只允许一条路经)，可以实现负载均衡。
- 对于每条链路，允许同时有多个(TOS)开销。
- 多播OSPF (MOSPF)使用与OSPF相同的链路状态数据库(思科路由器不支持)
- 在大型路由选择域中OSPF可以分区。
- 比RIP收敛快而且更安静。
- 实现起来更复杂，需要更多的计算开销。

LS算法和DV算法比较

消息复杂性

- LS: n个节点, E条链路, 要发送O(nE)条消息
- DV: 只在邻居之间交换消息

收敛速度

- LS: $O(n^2)$算法需要 O(nE) 条消息
  - 可能会震荡
- DV: 收敛时间变化
  - 可能出现路由循环
  - 计数到无穷问题

健壮性: 路由器失效时会出现什么情况?

LS:

- LS节点可能通告不正确的链路开销
- 每个节点只计算自己的路由表

DV:

- DV节点可能通告不正确的路径开销
- 每个节点的路由表被其它节点所用
  - 错误会通过网络传播开