中山大學
SUN YAT-SEN UNIVERSITY

第五单元 网络层
-OSPF协议

2017.5.18

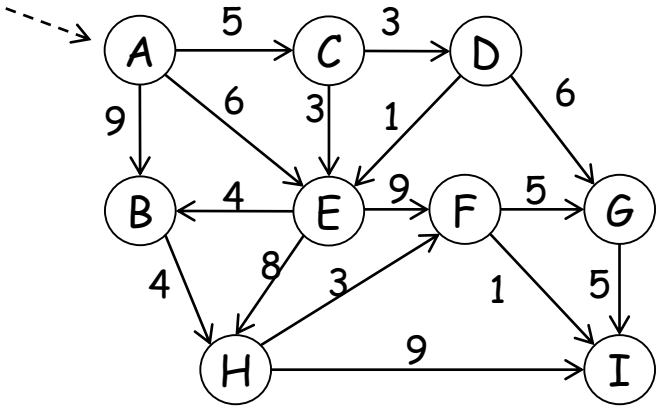
本节内容

- 概述
- 把网络转变为图
- OSPF协议的详细描述
- OSPF开销
- 数据库同步
- 路由器ID
- 指定路由器
- OSPF分组格式
- LS Update分组
- Dijkstra最短路经算法

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>

http://tools.ietf.org/html/rfc2328

- **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;

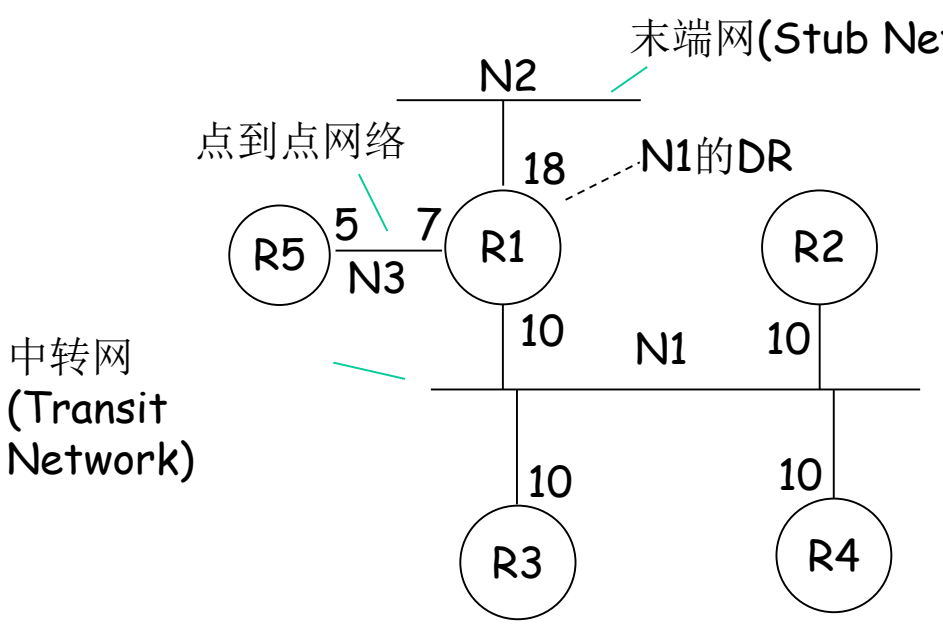
  (4) 利用这些路径上的下一跳建立路由表。

如何利用把整个网络(AS)转化为AS的拓扑结构图？

末端网：只跟一个路由器有联系

把网络转变为图

R1的路由表内有三项，就是到N1、N2、N3

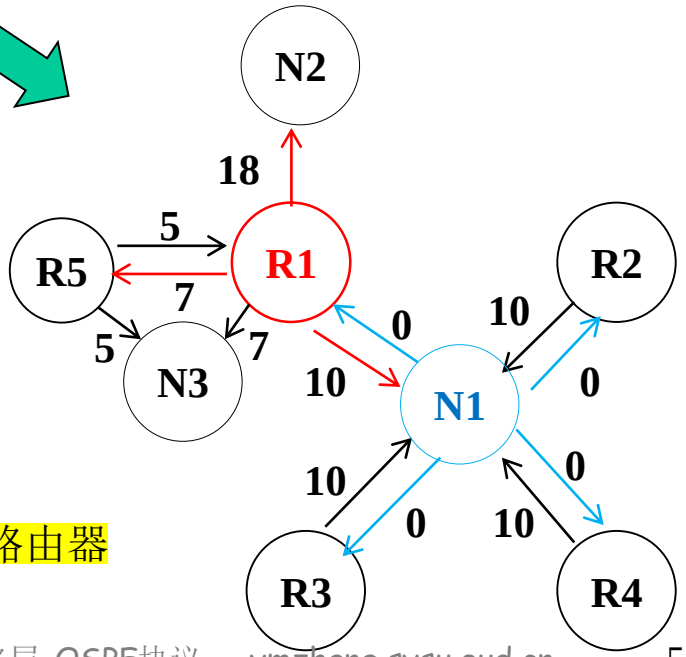


链路状态通告 路由器的发出边

(Link State Advertisement)

- N1和N2是多路访问网络，例如：以太网
- N3为点到点网络，例如：ppp

所有的路由器的LSA集不齐图中所有的边，因为中转网的发出边没有。



R1's Router LSA:

	R1 (From)
N1	10
N2	18
R5	7
N3	7

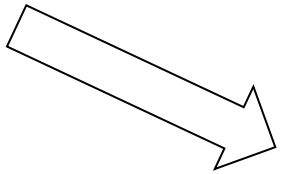
N1's Network LSA:

	N1(From)
R1	0
R2	0
R3	0
R4	0

- 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) 由它来收集和发散N1的LSA。
- 如果图中点到点网络没有配置IP地址，则不要节点N3

N1's Network LSA:

	N1(From)
R1	0
R2	0
R3	0
R4	0

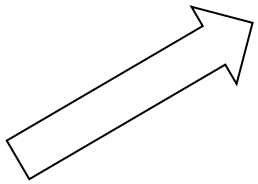


链路状态数据库

	R1	R2	R3	R4	R5	R6	N1	N2	(From)
R1							0		
R2							0		
R3							0		
R4							0		
R5	7								
R6									
N1	10								
N2	18								
N3	7								

R1's Router LSA:

	R1 (From)
N1	10
N2	18
R5	7
N3	7



- 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。

OSPF协议的详细描述

- 发现邻居

  OSPF路由器每10秒 (Hello Inverval)向邻居发送Hello分组。如果40秒(Dead Interval, 4 times hello interval)都收不到邻居发来的Hello分组，则把到邻居的链路标记为失效。多路访问网络采用多播 (224.0.0.5, all OSPF routers) 发送Hello分组。一个Hello分组包含优先权、已知的邻居（收到过Hello ）、DR和BDR。

BDR：备份指定路由器

- 完全相邻

  在发现邻居之后，OSPF路由器将与邻居交换链路状态数据库中的LSA，请求得到更新的或者没有的LSA。在与邻居的链路状态数据库变得完全一样时，它们就处于完全相邻状态（ fully adjacency）。

- 生成LSA

  每30分钟或链路变化时，每个OSPF路由器会生成 router LSA，中转网的DR会生成 network LSA。

- 扩散LSA

  产生的 LSA立即封装为Update分组，被可靠地扩散出去 (需要确认)。每次产生的LSA的序号会加1。序列号越大表示越新。若通过收到多个 LSA，由发出此LSA的路由器ID(发通告路由器), 链路状态和序列号唯一确定。通过序号，也可以防止扩散形成回路，第二次收到来自相同的发通告路由器、相同LSA类型和相同序号的LSA将丢弃它。

- 收集LSA

  路由器收集到LSA之后，用新LSA替换链路状态数据库中旧LSA。如果一个LSA在60分钟(max age)没有被更新，它将从链路状态数据库移除。

- 计算最短路径

  当链路状态数据库被改变时，OSPF路由器将利用Dijkstra算法计算到所有网络的最短路径。

- 建立路由表

  利用得到的最短路径产生路由表。

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

数据库同步

接口处于shutdown状态

Down

向对方不停地发送Hello

Attempt

Hello(DR=0，Seen=0) ---------->

<----------Hello(DR=0，Seen=0)

收到对方的Hello

Init

Hello(DR=0，Seen=0)

<----------

对方hello指出已收到自己的Hello

Two-way

Hello(DR=0，Seen=R1,…)

<----------

neighbor

广播型网络要选举DR

DR Elected if needed

Hello(DR=0，Seen=R1，DR=R1)

<---------->

ExStart

发送不带LSA的DD确定主从关系

RID大的成为Master

DD(Seq=x,I,M,Master) ---------->

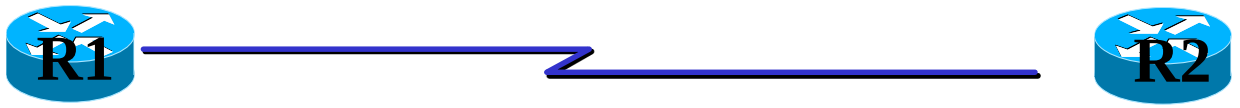
M-More

I- initialize

DD(Seq=y,I,M,Master)

<----------

DD指 Database Description Packet



由Slave发起，Master响应，都把自己的LSA头部发给对方，以便知道自己的LS数据库中缺少哪些LSA或哪些LSA过时了

Exchange

DD(Seq=y,M,Slave) →

← DD(Seq=y+1,M,Master)

DD(Seq=y+1,M,Slave) →

…

← DD(Seq=y+n,Master)

DD(Seq=y+n,Slave) →

发送完LSA头部的一方令M=0，并等待对方发送完毕

双方都可以发出请求，要求对方发送完整的LSA

Loading

LS Request →

← LS Update

LS Request →

← LS Update

可以在Exchange没有结束时就开始Loading过程，即 让Exchange和Loading交替进行。

收到了所有过时或缺少的LSA

Full

Fully adjacent

DROther把所有DD多播(224.0.0.6)给DR和BDR，DR单播确认，DR把新的DD多播(224.0.0.5)给该网络的所有OSPF路由器。对于无DR网络，OSPF路由器直接多播(224.0.0.5)DD和LSU给所有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*

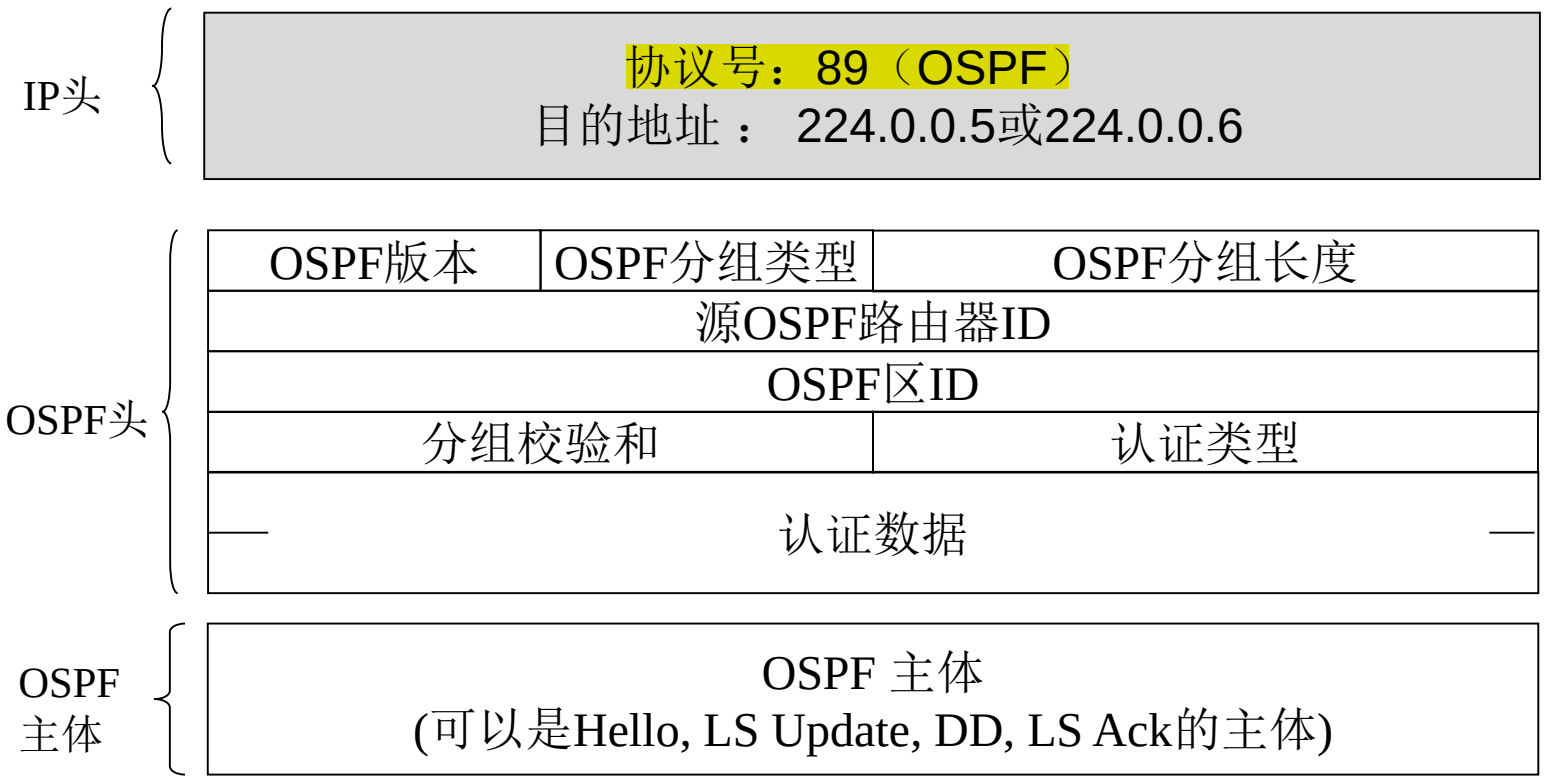
- 当多路访问网络重启时，选择DR的过程就开始了。在等待时间结束(Wait Time, Dead Interval，40秒)时，带有最高和次高优先权的路由器分别成为DR和 BDR(Backup DR)。如果优先权相同，RID更大的成为DR，次大的成为BDR。
- 如果路由器不希望参与选举，则应该把优先权设置为0。如果优先权相同，具有更高RID的路由器成为DR。如果收到的Hello列出了DR(RID不是0.0.0.0)，路由器成为DR。
- 如果一个新的路由器在选举之后到达或者有路由器修改为更高的优先权，它也不可能抢占现存的 DR (或BDR)和变为DR(或BDR)。
- 当DR失效时，BDR成为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。防止接触不良
- 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。二次保护
- 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

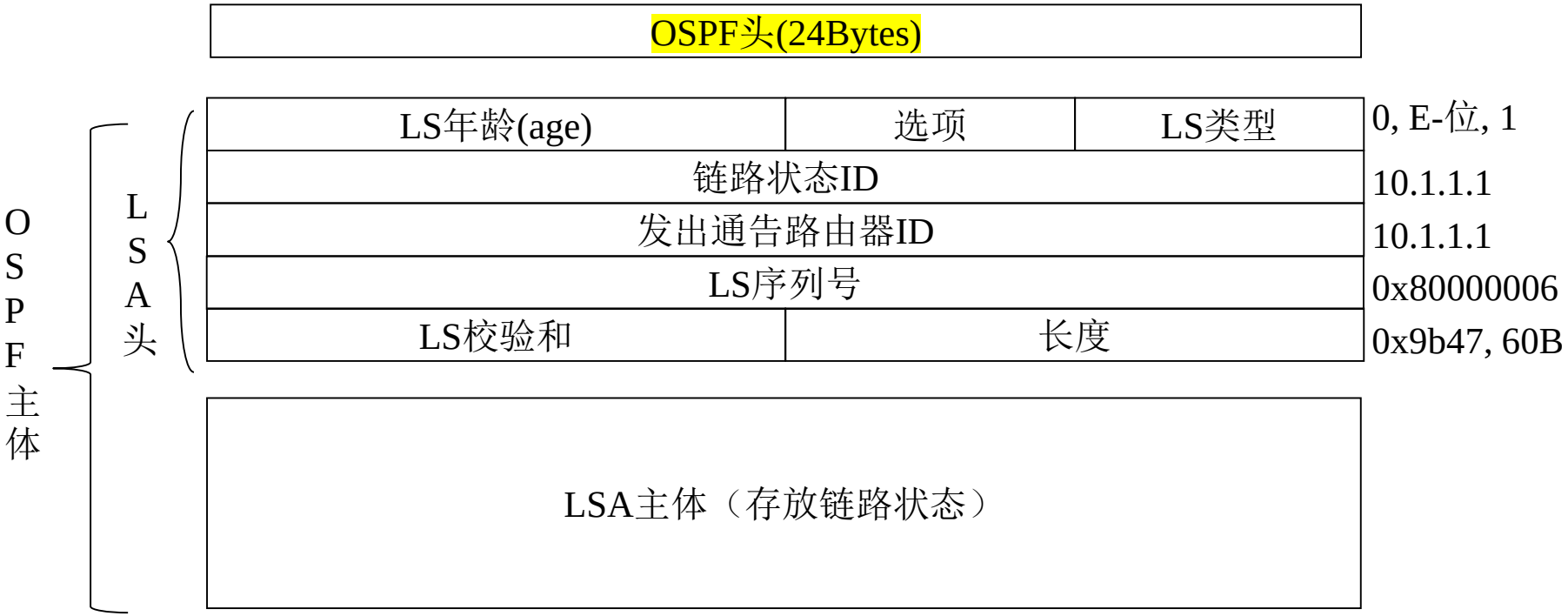


OSPF分组类型：

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）

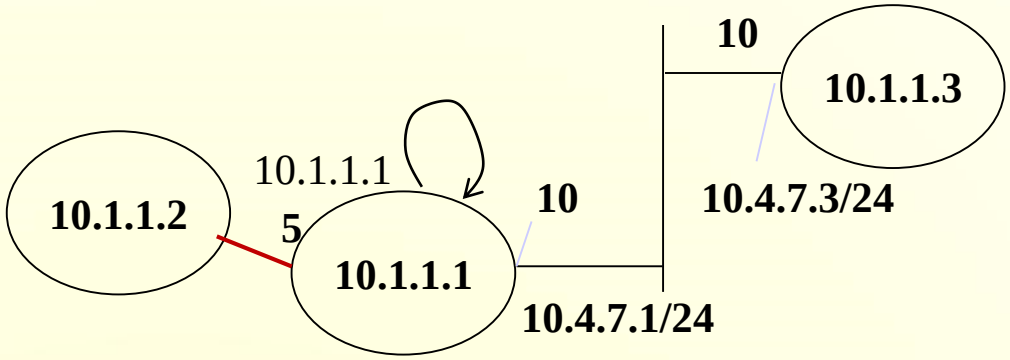
LS Update分组



- 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。

Router LSA



10.1.1.1的Router LSA

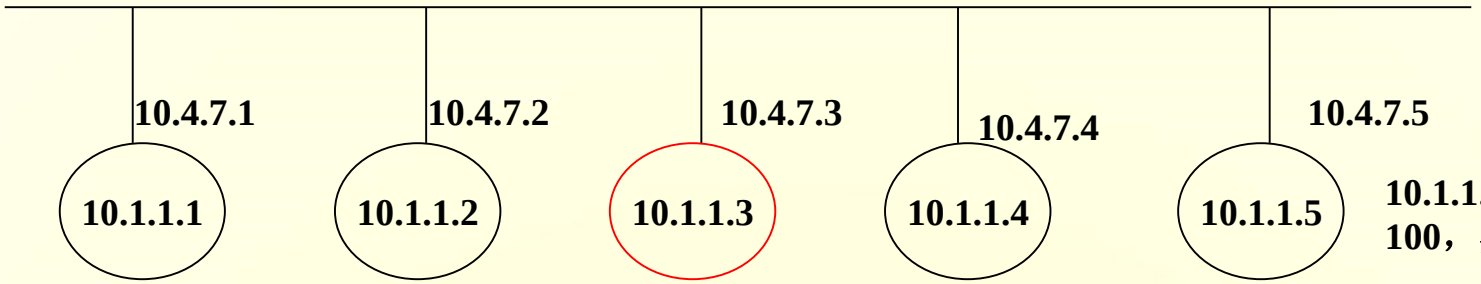
LS A 头

LS Type =1　**链路状态ID**:10.1.1.1，**长度**:60 **发出通告路由器ID(AR)**:10.1.1.1　**LS序列号**:0x80000006					20B
0	**V**	**E**	**B**	**0(8b)**	**链路总数(16b)**	0,000,0,3(链路数)
链路1	**链路ID**					10.1.1.2(邻居RID)
	链路数据					10.1.1.1(AR)
	链路类型			**TOS个数** 服务类型	**度量值**	1(点到点), 0, 5(开销)
	…					
	TOS			**0**	**TOS度量值**	这里只是举例，实际没有使用TOS度量
链路2	**链路ID**					10.4.7.1(接口IP地址)
	链路数据					10.1.1.1(AR)
	链路类型			**TOS个数**	**度量值**	2(广播网络), 0, 10(开销)
链路3	**链路ID**					10.1.1.1(接口IP地址)
	链路数据					255.255.255.255(掩码)
	链路类型			**TOS个数**	**度量值**	3(末端网), 0, 0 (开销)

- ✓ 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

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

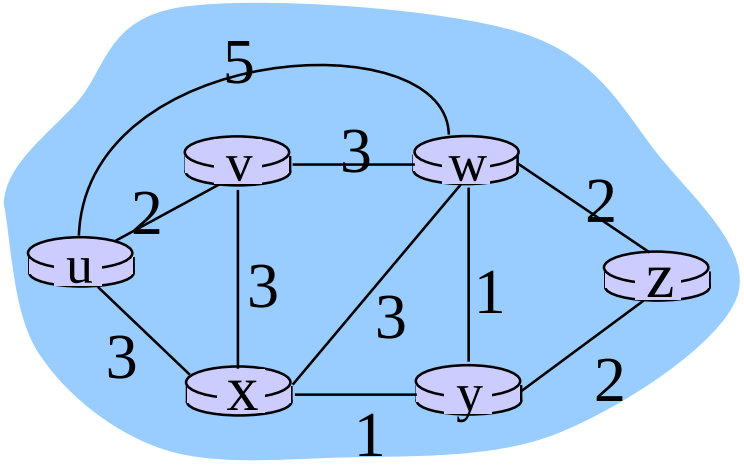
Notation:

- 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					



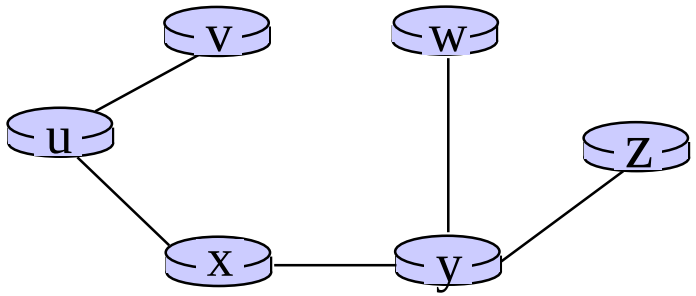
算法复杂性: n个节点

- 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:



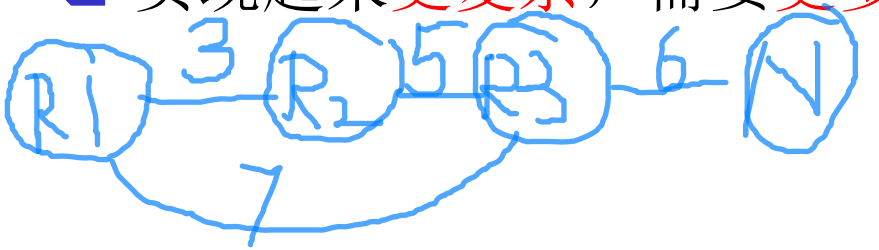
u的转发表:

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

OSPF的特点

路由器存的图不一样可能会导致有回路。

- 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。



R2中图没了R2与R3的边，那么R2要到N将会在R1与R2之间来回发送。

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

为什么LS数据库不一致可能导致回路问题？举例说明。

总结

- 概述
- 把网络转变为图
- OSPF协议的详细描述
- OSPF开销
- 数据库同步
- 路由器ID
- 指定路由器
- 分区OSPF
- OSPF分组格式
- LS Update分组
- Dijkstra最短路经算法