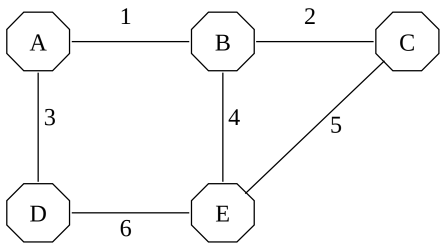
Link State Routing Principles

Link state routing

- The goal is to avoid the routing loops typical of DV routing and to scale to bigger networks and to varying topologies.
- A link state protocol maintains the topology map (link state DB) of the network.
  - Same map in every node
  - When the topology changes, the maps are updated quickly
- OSPF is IETF specified link state protocol for Internet.
  - OSPF is recommended as the follower of RIP.

The map is the full list of all links

- Example network

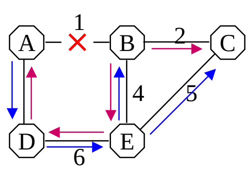


- One node is responsible for a particular entry
- Link directions are separate entries
- Same map in every node
  - No loops

From	To	Link	Cost
A	B	1	1
A	D	3	1
B	A	1	1
B	C	2	1
B	E	4	1
C	B	2	1
C	E	5	1
D	A	3	1
D	E	6	1
E	B	4	1
E	C	5	1
E	D	6	1

Flooding protocol distributes information about topology changes

- The updates are distributed to the whole network



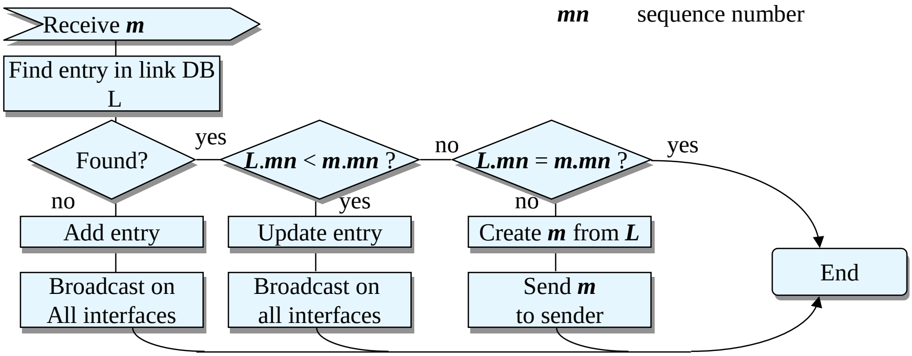
From	To	Link	Cost	Seq.num
A	B	1	inf	2

From	To	Link	Cost	Seq.num
B	A	1	inf	2

Flooding protocol distributes information about topology changes

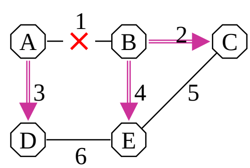
Flooding Algorithm:

m	received entry
L	corresponding row in link DB
mn	sequence number



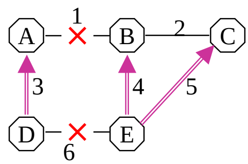
Link DB after distribution of failure of link AB

- Message numbering starts from 1 on node restart.
- Modulo arithmetic is used to determine what is "a little bigger than"
  - ⇒ message numbering can overflow without problems.
  - ⇒ 4294967295 + 1 = 0



From	To	Link	Cost	Seq.num
A	B	1	inf	2
A	D	3	1	1
B	A	1	inf	2
B	C	2	1	1
B	E	4	1	1
C	B	2	1	1
C	E	5	1	1
D	A	3	1	1
D	E	6	1	1
E	B	4	1	1
E	C	5	1	1
E	D	6	1	1

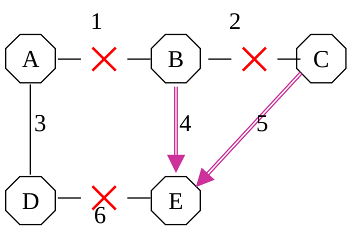
If network splits into islands, DBs in islands may diverge



From	To	Link	Cost	Seq.num
A	B	1	inf	2
A	D	3	1	1
B	A	1	inf	2
B	C	2	1	1
B	E	4	1	1
C	B	2	1	1
C	E	5	1	1
D	A	3	1	1
D	E	6	inf	2
E	B	4	1	1
E	C	5	1	1
E	D	6	1	1

From	To	Link	Cost	Seq.num
A	B	1	inf	2
A	D	3	1	1
B	A	1	inf	2
B	C	2	1	1
B	E	4	1	1
C	B	2	1	1
C	E	5	1	1
D	A	3	1	1
D	E	6	1	1
E	B	4	1	1
E	C	5	1	1
E	D	6	inf	2

Link 2 fails ⇒ DBs diverge even more

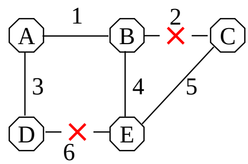


There is no immediate problem, but if link 1 goes up again ...

DBs in B, C and E:

From	To	Link	Cost	Seq.num.
A	B	1	inf	2
A	D	3	1	1
B	A	1	inf	2
B	C	2	inf	2
B	E	4	1	1
C	B	2	inf	2
C	E	5	1	1
D	A	3	1	1
D	E	6	1	1
E	B	4	1	1
E	C	5	1	1
E	D	6	inf	2

Link 1 goes up

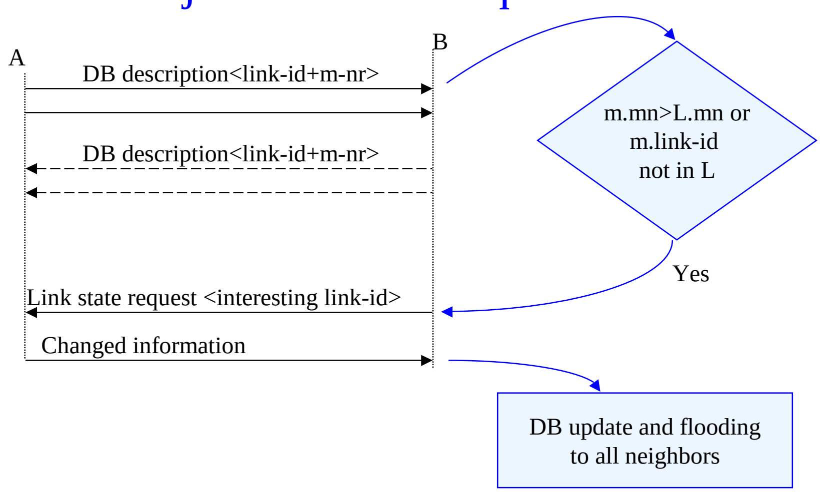


From	To	Link	Cost	Seq.num
A	B	1	1	3
A	D	3	1	1
B	A	1	1	3
B	C	2	1	1
B	E	4	1	1
C	B	2	1	1
C	E	5	1	1
D	A	3	1	1
D	E	6	inf	2
E	B	4	1	1
E	C	5	1	1
E	D	6	1	1

From	To	Link	Cost	Seq.num
A	B	1	1	3
A	D	3	1	1
B	A	1	1	3
B	C	2	inf	2
B	E	4	1	1
C	B	2	inf	2
C	E	5	1	1
D	A	3	1	1
D	E	6	1	1
E	B	4	1	1
E	C	5	1	1
E	D	6	inf	2

From	To	Link	Cost	Seq.num
A	B	1	1	3
A	D	3	1	1
B	A	1	1	3
B	C	2	inf	2
B	E	4	1	1
C	B	2	inf	2
C	E	5	1	1
D	A	3	1	1
D	E	6	inf	2
E	B	4	1	1
E	C	5	1	1
E	D	6	inf	2

After reconnection of the islands “bringing up adjacencies ” is required



What happens if router C is restarted?

- There may be LSAs in the network that were distributed by C before the restart
  - After the restart, the router numbers the LSA:s with the number “InitialSequenceNumber”
  - The neighbor replies that is has newer information.
  - If C wants to keep it own LSAs alive, it increases the sequence number by 1 and redistributes.
  - If the information of the neighbor is no longer valid, C removes it by distributing the same entry with the age = MaxAge.

Integrity of the Link DB must be secured

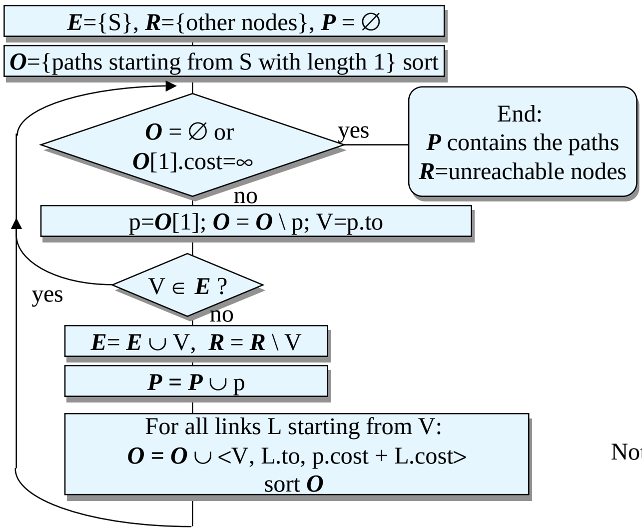
- *Flooding messages* are acknowledged link by link
- DB *description messages* are acknowledged
- Each DB entry is protected by obsolescence timer, if an update does not arrive in time, entry is removed.
- Each Entry is protected by a checksum
- Messages carry also authentication info
- *But: while update is in progress, some nodes receive info earlier than others --> routing mistakes happen*

Dijkstra's shortest-path-first algorithm

Dijkstra's shortest-path-first algorithm

- OSPF is based on Dijkstra's shortest-path-first algorithm.
- The purpose is to create a routing table.
- SPF - shortest path first - algorithm computes the shortest path from source node S to all other nodes
- Dijkstra's algorithm converges faster than Bellman-Ford
  - $O(M \log M) < O(N \cdot M)$
  - M is number of links, N is number of nodes
- Initially nodes are divided to Evaluated E, the paths from which are known and to other nodes R.
- In addition an ordered list of paths O is needed

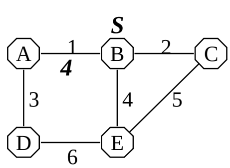
Dijkstra’s shortest-path-first algorithm



S	start node
E	evaluated nodes
R	other nodes
O	ordered list of paths <from, to, cost>
P	paths (result)

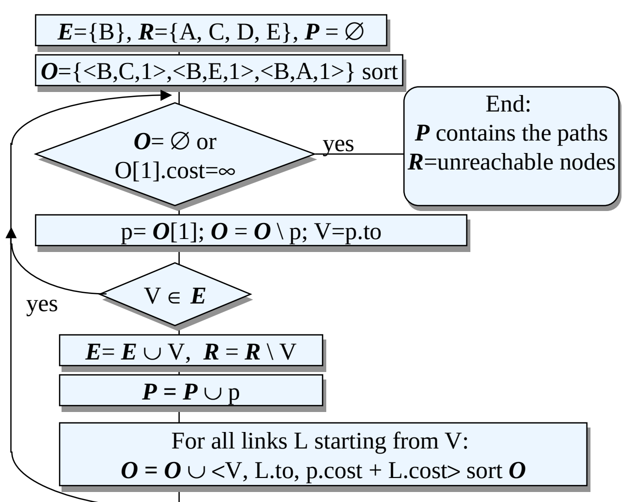
Note: the loops are removed here

Dijkstra’s shortest-path-first algorithm – example

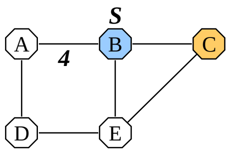


L

From	To	Link	Cost
A	B	1	4
A	D	3	1
B	A	1	4
B	C	2	1
B	E	4	1
C	B	2	1
C	E	5	1
D	A	3	1
D	E	6	1
E	B	4	1
E	C	5	1
E	D	6	1



Dijkstra’s shortest-path-first algorithm – example



E

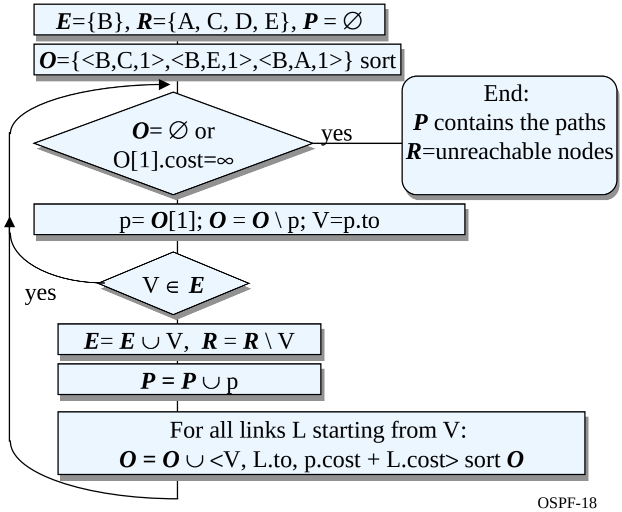
B

R

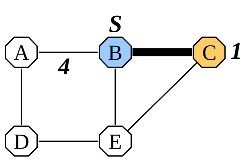
A, C, D, E

O

<B,C,1>
<B,E,1>
<B,A,4>



Dijkstra’s shortest-path-first algorithm – example



E

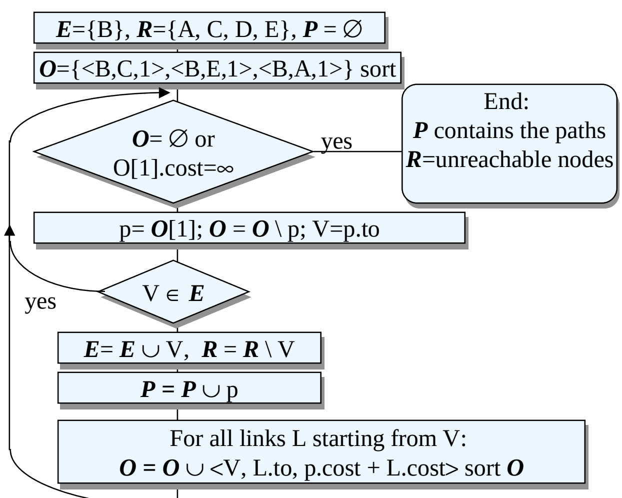
B, C

R

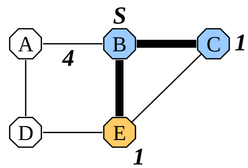
A, D, E

O

<B,E,1>
<C,E,2>
<B,A,4>



Dijkstra's shortest-path-first algorithm – example



E

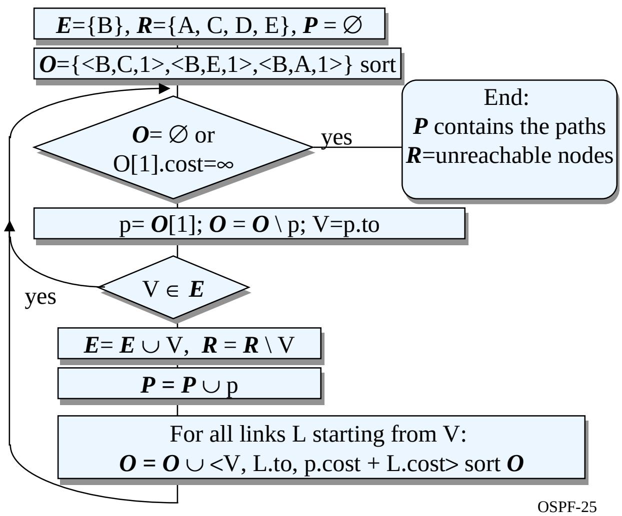
B, C, E

R

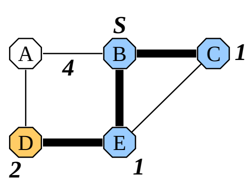
A, D

O

<C,E,2>
<E,D,2>
<B,A,4>



Dijkstra's shortest-path-first algorithm – example



E

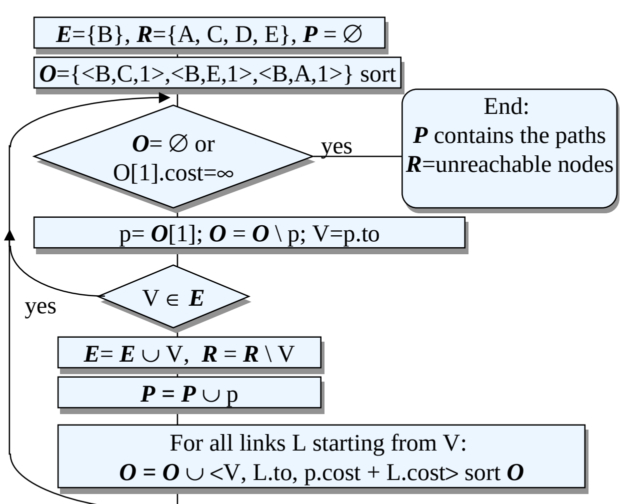
B, C, E, D

R

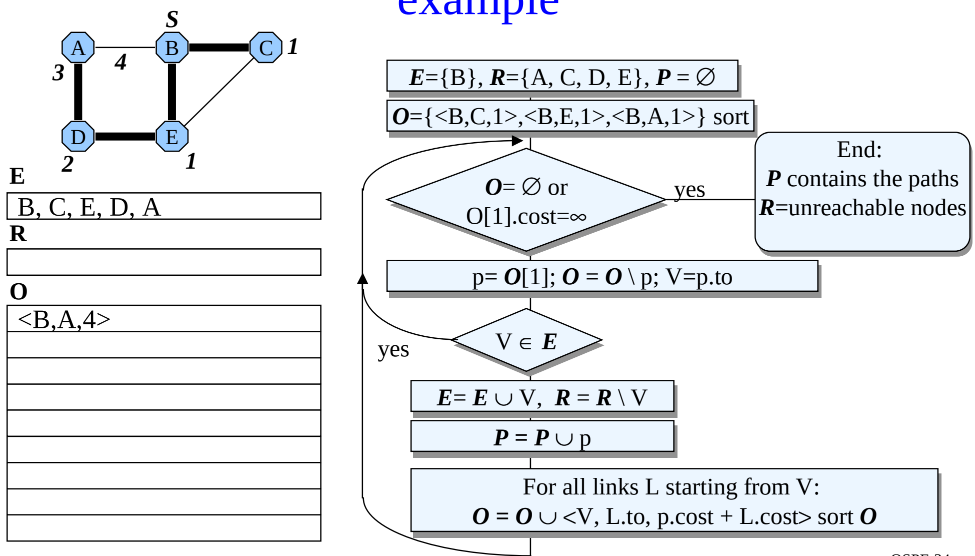
A

O

<B,A,4>



Dijkstra’s shortest-path-first algorithm – example



Advantages of Link State Protocols

- Link State DBs converge quickly, no loops are formed
  - O(M log M) *M = numer of links*
- Metrics can be quite accurate.
  - In DV-protocols counting-to-infinity limits (inf=16)
- One protocol can easily support several metrics:
  - Capacity, delay, cost, reliability.
- Can maintain several routes to a destination.
- Exterior routes can have their own representation.

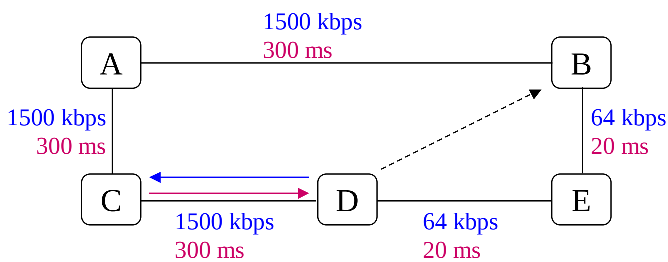
Using several metrics (1)

Using several metrics requires:

- Metrics must be stored for each link (L.et1, L.et2 ...)
- Computing separate Routing Tables for each metric (P(et1), P(et2) ...)
- Link protocol must carry all metrics
- User packets must be marked with the required metric.

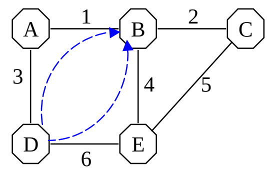
Using several metrics (2)

A Routing loop is possible if different nodes use different metrics for one user packet



⇒ User packets must be marked with the required metric

Spreading load to alternative equidistant paths improves network efficiency

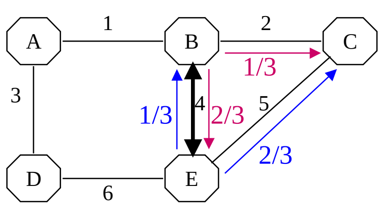


- + Queues in nodes become shorter
- + Average delay is decreased
- + End-to-end jitter decreases
- + Less traffic to reroute under failure conditions

- - May change packet order because paths may have different delay (queue lengths in nodes)
- - Existing traffic can not be pinned down to primary path so that only overload would take the alternative path ⇒ stability is a problem
- ? When are paths equidistant enough?

When are paths equidistant enough?

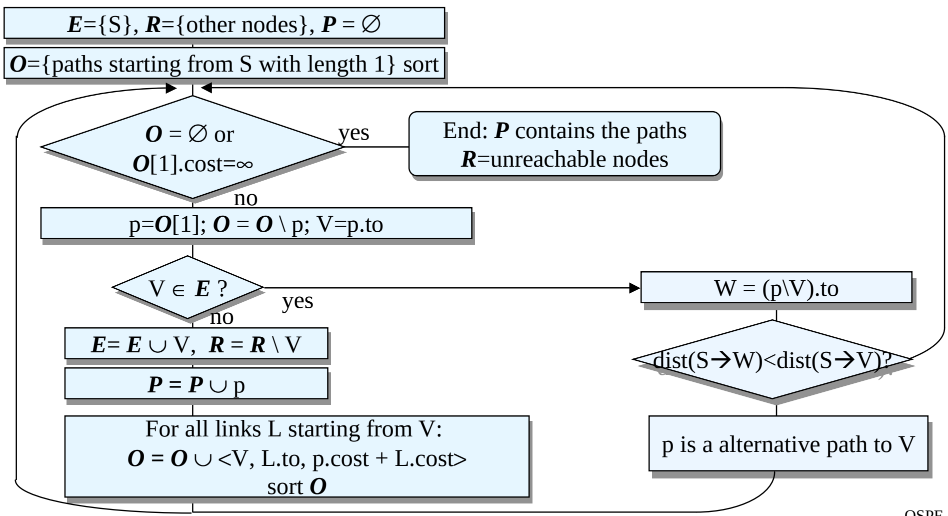
- What happens if the traffic to C is divided between two alternative paths?



⇒ The packet to X can be sent through Y is Y is closer to the destination than the local node

- Rule A->Y...->X, if Y.et.X < A.et.X accepts only monotonic alternative routes

Dijkstra’s shortest-path-first algorithm that finds alternative paths

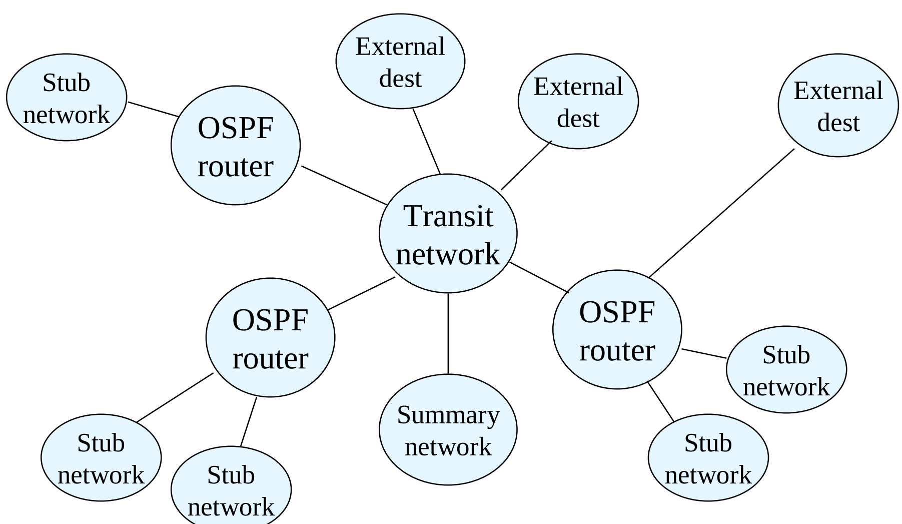


Link state protocol can describe several external routes with accurate metrics

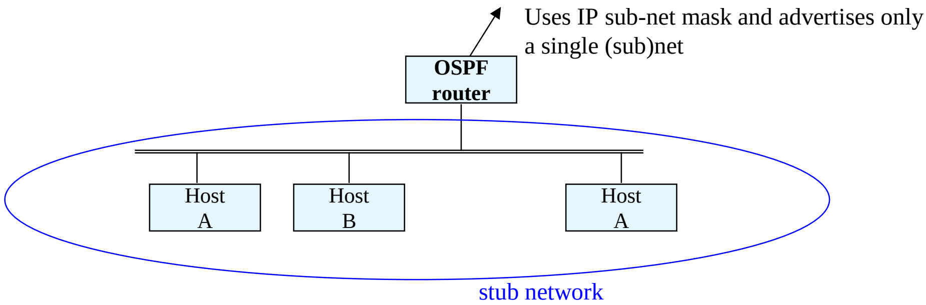
- DV-protocol capability to describe external routes is limited due to counting to infinity problem and due to complexity of Bellman-Ford algorithm
  - Inf=16 ⇒ maximum distance limited
  - Bellman-Ford complexity is O(N**2)
- Link state protocol is free of those limitations.
  - SPF route computation converges as O(N.logN) - where N = nrof external routes
- E.g. if there are 30 000 external routes => 10exp9 vs. 450 000

The OSPF protocol

OSPF sees the network as a graph



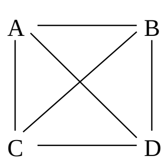
OSPF makes a difference between a router and a host



OSPF supports broadcast networks (1)

In a broadcast network

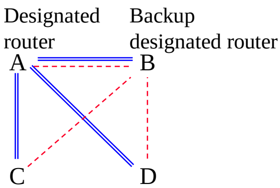
- Each device can send to each other
- One can send to all or to a sub-set of connected devices
- If it has N routers, they have N*(N-1)/2 adjacencies and
- Each router would advertise N-1 routes to other routers + one stub network ⇒ N^2

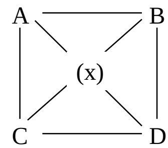


N*(N-1)/2 adjacencies (known neighbors)

E.g. Ethernet, Token ring, FDDI

OSPF supports broadcast networks (2)

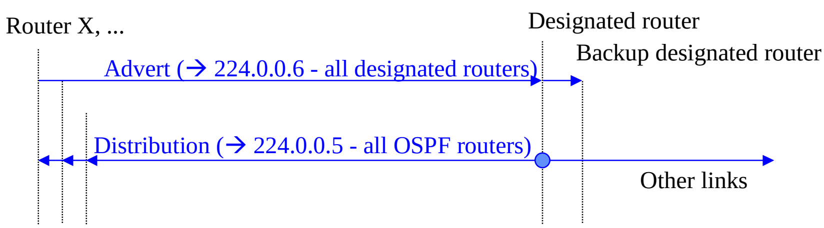




- Adjacencies are formed only with the designated router (A)
  - ⇒ Must be selected using the Hello protocol
  - ⇒ Synchronization of link DBs becomes simpler
- Backup designated router (B) should be selected together with the designated.

- The broadcast network is modeled using a ”virtual router”
- The links from the virtual router are network links
  - Advertised by the designated router
  - Cost = 0
- The links from the routers to the virtual router
  - Advertised by the routers

OSPF flooding protocol in a broadcast network

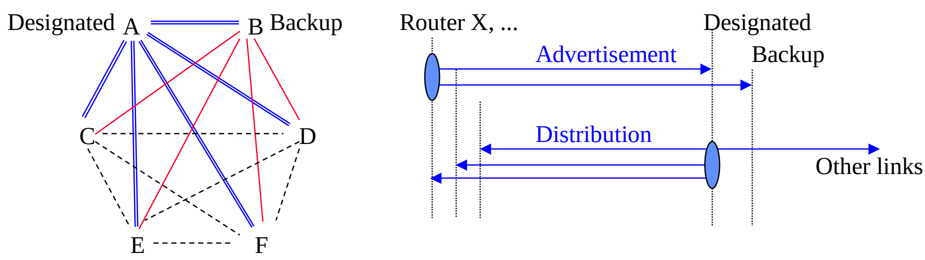


⇒ No need to process acks from all other routers in the sub-net

Backup designated stays as silent as possible

OSPF flooding protocol in a non-broadcast network

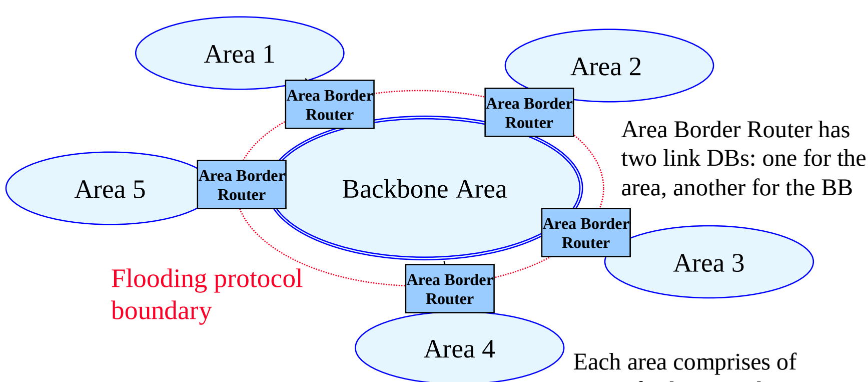
- In non-broadcast nets OSPF works in the same way except that broadcasts are replaced by point-to-point messages



Permanent connection with designated
Permanent connection with backup designated
Dial-up connection with other routers (other traffic)

NB: it makes sense to minimize permanent connections due to their cost

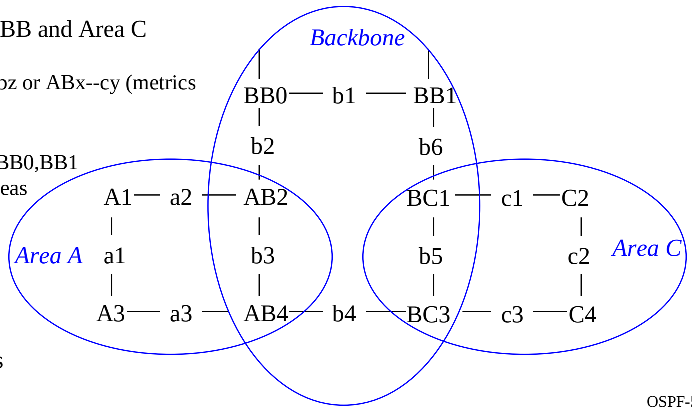
By breaking down a large network into areas OSPF eases flooding and reduces the size of link DBs



(Sub)networks of other areas are described in summary records – the metric is computed in “RIP-style”

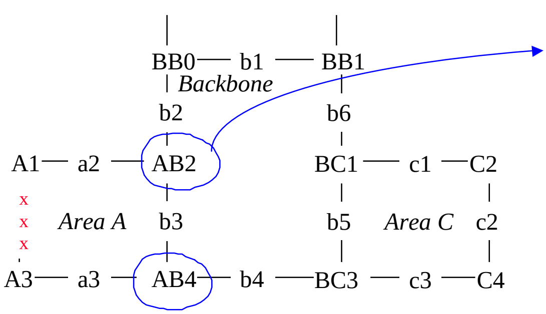
Link DB for Area A:

- a1
- a2
- a3
- sub-net records of BB and Area C
  - ← AB2,AB4
  - Distance ABx--bz or ABx--cy (metrics are summed).
- external records
  - ←AB2,AB4 ←BB0,BB1
  - All info in all areas



Hierarchy ⇒ No loops

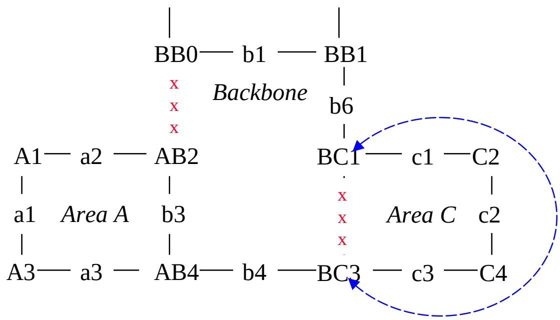
OSPF easily recovers from failures in areas



AB2 and AB4 advertise only those sub-nets which they can reach: AB2: a2 and AB4: a3.

BBone does not know exact structure of Area A - but knows identities of all reachable sub-nets.

In backbone failures a virtual link can help if backbone splits into isolated segments



Virtual link through Area C:
distance=c1+c2+c3

On a stub area all external routes are summed to the default route

- If an OSPF Area has only one Area Border Router, all traffic to and from the Internet goes thru this ABR. It is no use to advertise all Internet Routes separately towards such an Area.
- There can even be several ABRs but the best of them can not be selected based on destination prefix (leading bits (<32) of IP address)
- NSSA - “Not So Stubby Area” is an Area, on which all external routes have been summed into the default route except for some.

OSPF link state records

LSA types in OSPF

- LS Type = 1 Router LSA
  - describes set of active If and neighbors
- LS Type = 2 Network LSA
  - describes a network segment (BC or NBMA) along with the IDs of currently attached routers
- LS Type = 3 Summary LSA
- LS Type = 4 AS Border Router Summary LSA
- LS Type = 5 AS- external LSA
  - describes external routes
- LS Type = 6 Group Membership LSA
  - MOSPF – Multicast
- LS Type = 7 NSSA LSA
  - to import limited external info
- LS Type = 8 (proposed) external attributes LSA
  - in lieu of Internal BGP

Hierarchical Routing (LS Types 3–5)

NBMA = Non-Broadcast Multiple Access, e.g ATM

All OSPF routers on an area have identical link databases

Common header of Link State Advertisement (LSA)

LS age	**options**	**LS type**
Link state ID		
Advertising router		
LS sequence number		
LS checksum	**length**	

Key

← 32 bits →

LS age:

- Seconds from advertisement

Options:

- E - external links
- T - type of service
  when many metrics are in use

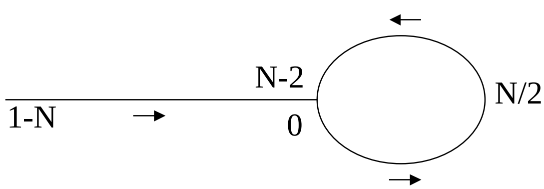
Link DB has "LS" record/entry types

1. router LSA
2. network LSA
3. Summary link (IP network)
4. Summary link (to a border router)
5. External link
   .. Summary records have the same format
6. Multicast LSA
7. NSSA record

etc ...

LSA Sequence Numbers

- "Lollipop sequence space"



$N = 2^{31}$

- If one of the number is < 0
  - The higher number is newer
- If both numbers are ≥ 0
  - If (b-a) < (N-1)/2 then b is newer

Router LSA (type 1)

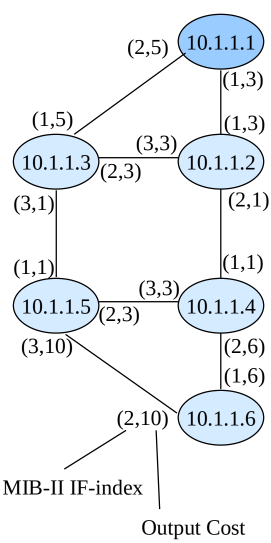
RouterType	0	Number of links
Link ID		
Link data		
Type	# TOS	TOS 0 metric
TOS=x	0	TOS x metric
TOS=y	0	TOS y metric
. . .		
TOS=z	0	TOS z metric

Router type
- E-bit (External)
  - This router is an area-border router
- B-bit (Border)
  - This router is a border router

TOS 0 metric when no TOS is used

Type
1. Link is a point-to-point link to another router
   - Link ID = router's OSPF ID, link data = router's interface address
2. Link connects to a transit network
   - Link ID = IP address of designated router's interface, link data = router's interface address
3. Link connects to a stub network
   - Link ID = Network/subnet number, link data = network/subnet mask

Example of the Router LSA



Router 10.1.1.1's router-LSA:

	LS Age = 0 seconds	Options	LS type=1	E-bit; 1=Router LSA
	Link State ID = 10.1.1.1			
	Advertising Router = 10.1.1.1			
	LS Sequence Number = 0x80000006			
	Checksum= 0x9b47	Length = 60 bytes		
0=ordinary	RouterType=0	0	Nrof links = 3	
	Link ID = 10.1.1.2 (neighb)			
	Link Data = IF-index 1 (unnum)			
1=pt-t-pt	Type=1	#TOS=0	Metric=3	
	Link ID = 10.1.1.3 (neighb)			
	Link Data = IF-index 2 (unnum)			
1=pt-t-pt	Type=1	#TOS=0	Metric=5	
	Link ID = 10.1.1.1			
	Link Data = 255.255.255.255			
3=stub network	Type=3	#TOS=0	Metric=0	

Length = 24 + 3 * 12 = 60 bytes

Router with 100 interfaces:
- Length = 24 + 100 * 12 = 1224 bytes

Network Link LSA (type 2)

Network mask
Attached router
Attached router
. . .
Atttached router

- Advertised by designated routers for transit networks
- Link state ID = IP interface ID
- Attached router = OSPF identifier of the attached router

Summary Link LSA (type 3,4)

Network mask		
0	0	TOS 0 metric
TOS=x	0	TOS x metric
TOS=y	0	TOS y metric
. . .		
TOS=z	0	TOS z metric

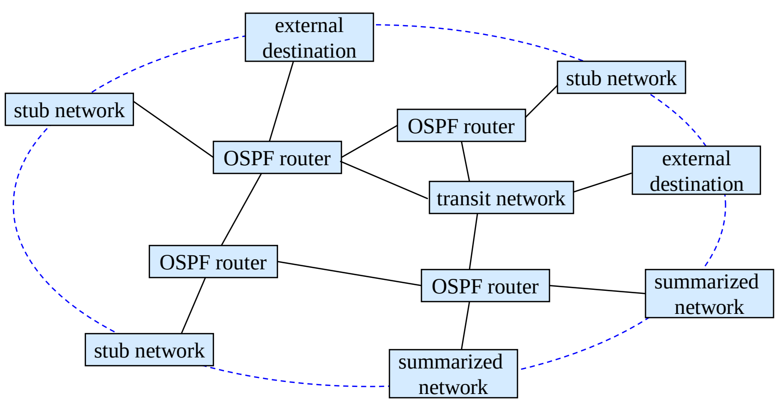
- For IP networks
  - LS type = 3
  - Network mask of network/subnet
  - Link state ID = IP network/subnet number
- For border routers
  - LS type = 4
  - Network mask = FFFFFFFF
  - Link state ID = IP address of border router
- One separate advertisement for each destination

External Link LSA (type 5)

Network mask		
E,TOS=0	0	TOS 0 metric
External route tag (0)		
E,TOS=x	0	TOS x metric
External route tag (x)		
. . .		
E,TOS=z	0	TOS z metric
External route tag (z)		

- Advertised by border routers
  - Information from external gateway protocols (BGP-4)
- One destination per record
- Link state ID = IP network/subnet of destination
- Network mask = network/subnet mask
- E-bit indicates that distance is not comparable to internal metrics
  - Larger than any internal metric
- Route tag is only used by border routers (not used by OSPF)

Computation of routes



- Separate routes for each TOS and for TOS 0
  - Possibly unreachable destinations for some TOS if not all routers support TOS ⇒ routed with TOS 0

The OSPF protocol

OSPF packets – the protocol itself

- OSPF works directly on top of IP. OSPF protocol number is 89.
- For most packets TTL = 1, except for hierarchical routing
- Destination IP address = Neighbors IP address or AllOSPFRouters (224.0.0.5) or AllDRouters (224.0.0.6)
- Packet types are
  - Type 1: Hello
  - Type 2: Database Description packet
  - Type 3: Link State Request packet
  - Type 4: Link State Update packet
  - Type 5: Link State Acknowledgement packet

OSPF protocol runs directly on IP

OSPF has 3 sub-protocols:

1. Hello protocol
2. Exchange protocol
3. Flooding protocol

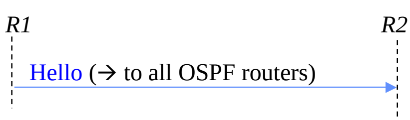
The common OSPF message header:

Version	Type	Packet length
Router ID		
Area ID		
Checksum		Authentic. type
Authentication		
Authentication		

- Current version of OSPF is 2
- Type differentiates OSPF messages
- Authentication type
  - 0 = No authentication
  - 1 = Password
    - – Limited protection
  - 2 = Cryptographic authentication
    - – MD5

0	Key ID	Length
Cryptographic sequence number		

Hello protocol ensures that links are working and selects designated router and backup DR



- Neighbor – list of neighbors that have sent a hello packet during last dead interval seconds.
- Hello interval tells in seconds how often hello packets are sent.
- Priority tells about eligibility for the role of Designated Router.
- A hello packet must be sent and received before a link becomes operational

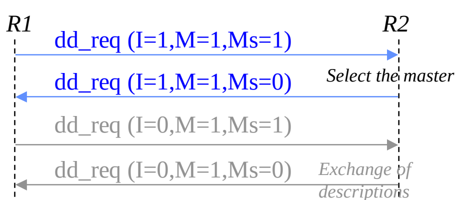
OSPF packet header type = 1		
Network mask		
Hello interval	Options	Priority
Dead interval		
Designated router		
Backup designated router		
Neighbor		
- - -		
Neighbor		

- Options
  - – E = external links.
  - – T = TOS routing capability.
- DR and Backup DR = 0 if not known

Hello protocol selects the DR and the Backup DR

1. Eligibility is achieved after one dead interval provided two-way reachability is OK.
2. From the routers that announced eligibility, the one with highest priority is elected to Backup Designated. Tie is broken by electing the one with highest ID.
3. If no neighbor proposed itself to backup DR, the neigbor with the highest priority is selected. Tie is broken by selecting the one with highest ID.
4. Designated in selected among those that proposed with rules 2 and 3.
5. If none proposed itself to DR, the backup DR is promoted. Actions 2 and 3 are repeated to re-select the backup DR.
6. A high priority former DR postpones its proposal to retake the position of DR after recovery to minimize changes. Actions 2....5 are continuous.

Exchange protocol initially synchronizes link DB with the designated router (1)

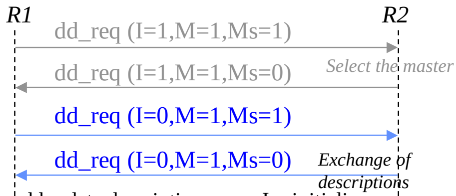


dd = data description I = initialize
M = more Ms = Master/slave

OSPF packet header type = 2 (dd)			
0	0	Options	0 IMMs
dd sequence number			

- Exchange protocol uses database description packets
- First the master and slave are selected
- If both want to be masters, the highest address wins
- Retransmission if the packet is lost
- The same sequence number in the replies

Exchange protocol initially synchronizes link DB with the designated router (2)

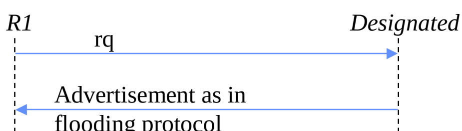


dd = data description
M = more
I = initialize
Ms = Master/slave

OSPF packet header type = 2 (dd)			
0	0	Options	0 IMMs
dd sequence number			
Link state type			
Link state ID			
Advertising router			
Link state sequence number			
Link state checksum		Link state age	
- - -			

- Master sends its Link DB description in sequence numbered packets
- Slave acks by sending its corresponding description packets.
- Exchange continues until all descriptions are sent and acknowledged. (M=0)
- Differences are recorded on the list of “records-to-request”.

Request packets are used to get record contents. Requests are acknowledged by flooding protocol packets

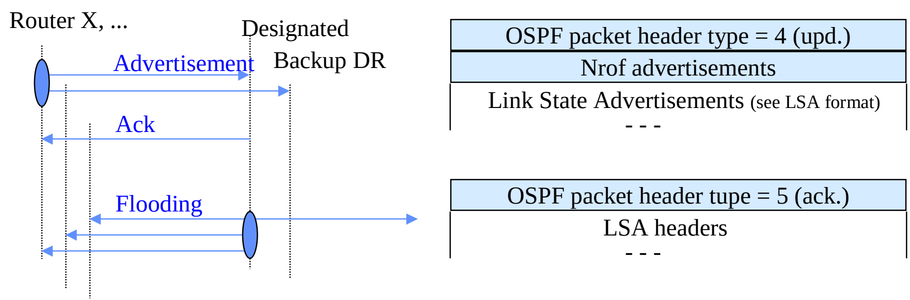


rq = request

OSPF packet header type = 3 (rq)
Link state type
Link state ID
Advertising router
- - -

- Router waits for ack for resend interval. If no response, Rq is repeated.
- “Records-to-request” may be split into may Requests, there are too many.
- If something goes wrong, backup to role negotiation is the typical remedy.
- First Request can be sent immediately when first interesting record has been detected. Then dd-packet exchange and Rq packet exchange take place in parallel.

Flooding protocol continuously maintains Link DB integrity



- Original LSA is always sent by the router responsible for that link.
- Advertisement is distributed acc to flooding rules to the Area (age=age+1).
- Ack of a new record by DR can be replaced in BC network by update message.
- One ack packet can acknowledge may LSAs.

Summary of OSPF subprotocols

	Hello (1)	DD (2)	LS rq (3)	LS upd (4)	LS ack (5)
Hello protocol	X				
Database exchange		X	X	X	X
Flooding protocol				X	X

Server Cache Synchronization Protocol (SCSP) is OSPF without Dijkstra's algorithm and with more generic data objects.

Link records have an age, old/dead ones are removed from Link DB (1)

- Old information must be removed from DB
- Every node must use the same information
  - ⇒ The removals must be synchronized
- The LSAs of OSPF have an age
  - Age = 0 when the advertisement is created
  - Age = number of hops thru which the advertisement has traveled + seconds from reception
- Max age is 1 hour
  - Not used in the calculation of routes
  - Must be removed
- Every entry must be advertised at a 30 min interval.
  - The new advertisement zeroes the age and increments the sequence number.

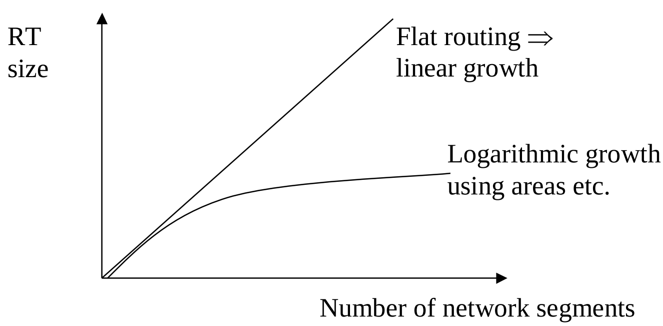
Link records have an age, old/dead ones are removed from Link DB (2)

- When the age reaches MaxAge (= 1 h) the entry is removed
  - The router must send an advertisement to the neighbors when the aged entry is removed
- The flooding algorithm examines the age of the received advertisement
  1. MaxAge advertisement is accepted and flooded - removes obsolete info.
  2. If the age difference of the advertisement to the DB is small, the advertisement is not flooded to avoid overloading the network with multiple copies of the same info. This is due to normal routing when the entry is received on different paths.
  3. If the age difference is large (>MaxAgeDiff), the newest advertisement is accepted and distributed. In this case, the router has probably been restarted.
  4. If MaxAge record is not found, advertisement has not impact. The router most likely has already removed the dead LSA.

OSPF timeouts - LS Age field

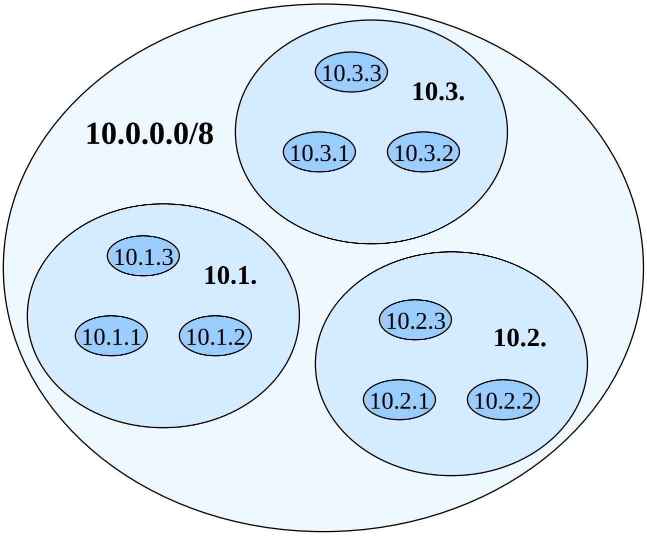
Constant	Value	Action of OSPF router
MinLSArrival	1 second	Max rate at which a router will accept updates of any LSA via flooding
MinLSInterval	5 seconds	Max rate at which a router can update an LSA
CheckAge	5 min	Rate to verify an LSA Checksum in DB
MaxAgeDiff	15 min	When Ages differ more than 15 min, they are considered separate. Smaller LS age - newer!
LSRefreshTime	30 min	A Router must refresh any self-originated LSA whose age has reached 30 min.
MaxAge	1 hour	LSA is removed from DB.

The purpose of hierarchical routing in OSPF is to reduce routing table growth



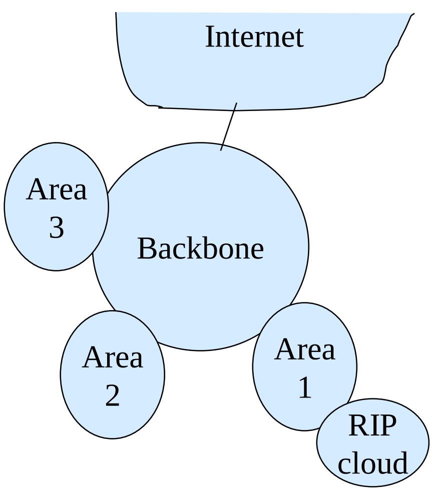
The cost is: sometimes suboptimal routes.

Example of routing hierarchy



Example:
- 16 segments in each lowest level network
- flat routing: RTsize= 16*9=144
- areas 10.1.1: 16 local routes + 10.1.2/24 10.1.3/24 10.2/16 10.3/16 == 20 RT entries!

OSPF supports 4 level routing hierarchy



Level	Description
1	Intra-area routing
2	Inter-area routing
3	External Type 1 metrics
4	External Type 2 metrics

- Type 1 metrics are of the same order as OSPF metrics, e.g. hop count (for RIP and OSPF)
- Type 2 metrics are always more significant than OSPF internal metrics

Why is it difficult to route packets around network congestion?

- BBN ARPANET link state metric varied with the length of the output queue of the link ⇒ lead to route trashing.
- The problem is there is no route pin-down for existing traffic.
- By limiting the range of the metric changes, an equilibrium could be reached. Nevertheless routing instability is the problem.

When QoS or Class of Service a'la DiffServ is introduced this problem again becomes important.

OSPF development history