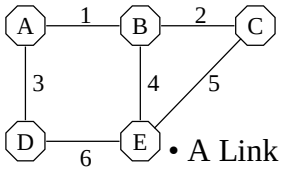
Link State Routing Principles

The Goal is to avoid the routing loops typical of DV routing and to scale to bigger networks and to varying topologies.

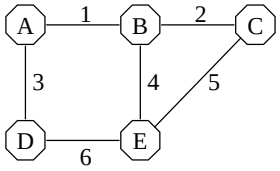
Open Shortest Path First(OSPF) is a recommended link state protocol for Interior routing in Internet



Example network

- A Link state protocol maintains the topology map (Link state DB) of the network.
  - When topology changes, maps are updated quickly.
  - The map is used to produce the Routing Table.
- 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

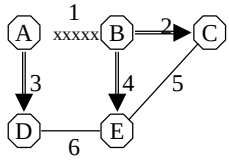


- One node is responsible for a particular entry
- Link directions are separate entries

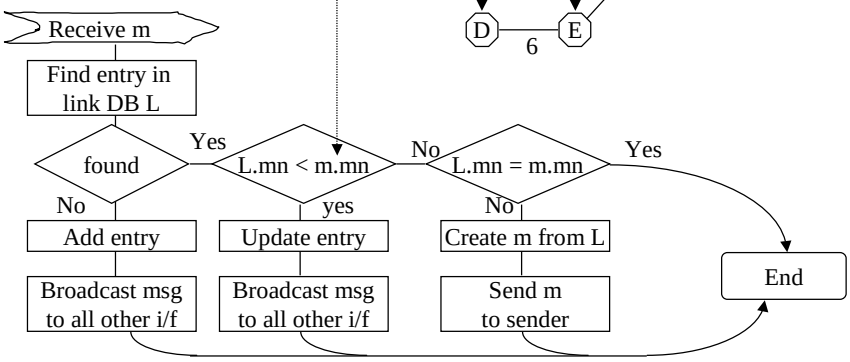
From	To	Link	Distance
A	B	1	1
A	D	3	1
B	A	1	1
B	C	2	1
B	E	4	1
C	B	2	1
C	E	5	1
D	A	3	1
D	E	6	1
E	B	4	1
E	C	5	1
E	D	6	1

Flooding protocol distributes information about topology changes

From A to B, link 1, dist = inf, message number

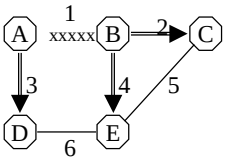


Flooding Algorithm is



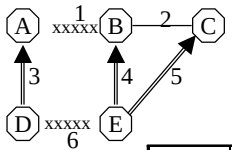
Link DB after distribution of failure of link AB

From	To	Link	Distance	M-Nr
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



- 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.

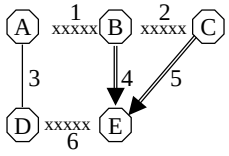
If network splits into islands, DBs in islands may diverge



From	To	Link	Distance	M-nr
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	Distance	M-nr
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

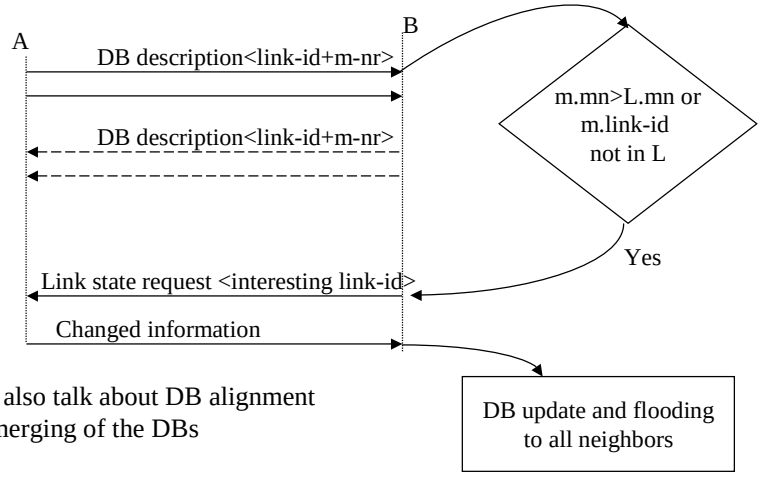


There is no immediate problem,
but if link 1 goes up again ...

DBs at B, C and E:

From	To	Link	Distance	M-nr
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

After reconnection of the islands “Bringing Up Adjacencies ” is required



We also talk about DB alignment
or merging of the DBs

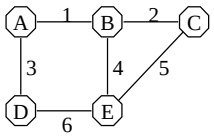
Integrity of the Link DB must be secured

- *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*

OSPF is based on Dijkstra's SPF algorithm

- SPF - shortest path first -algorithm computes the shortest path from source node *S* to all other 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.

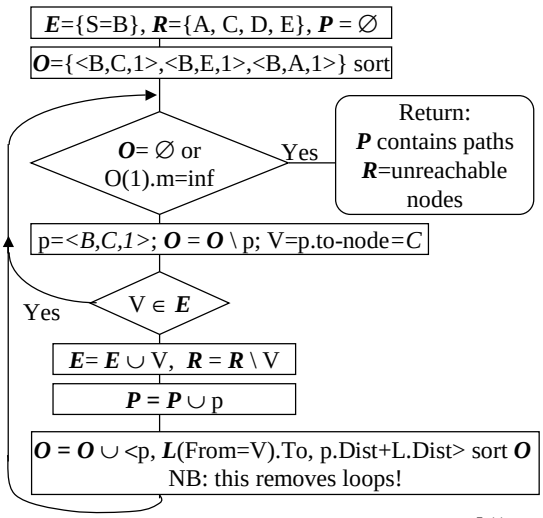
Dijkstra's shortest-path-first algorithm



L

From	To	Link	Dist.
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

Converges faster than Bellman-Ford
O(M.logM) < O(N.M)



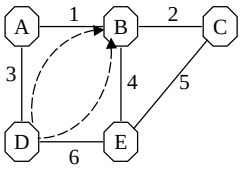
Advantages of Link State Protocols include

- Link State DBs converge quickly, no loops are formed
- Metrics can be quite accurate. 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 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.
- A Routing loop is possible if different nodes use different metrics for one user packet.

Spreading load to alternative equidistant paths improves network efficiency



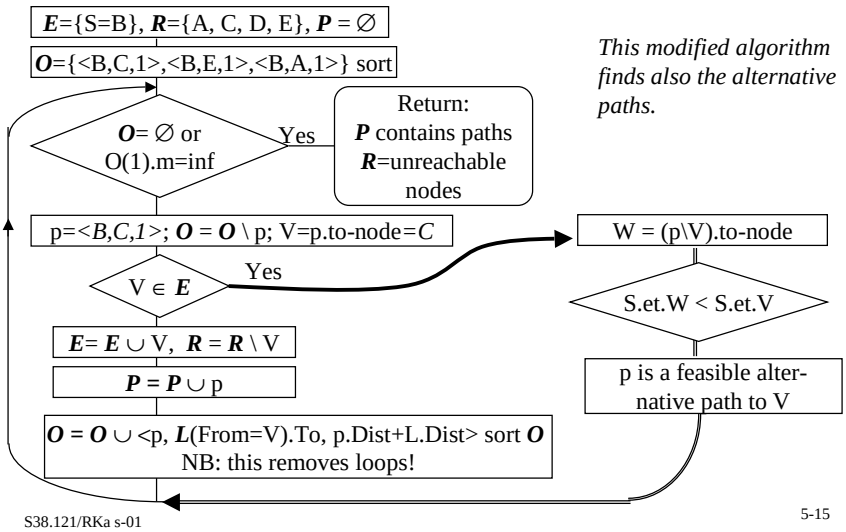
- Queues in nodes become shorter
- Average delay is decreased
- End-to-end jitter decreases
- Less traffic to reroute under failure conditions

– May change packet order because paths may have different delay (queue lengths in nodes)

– Difficulty: 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?

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



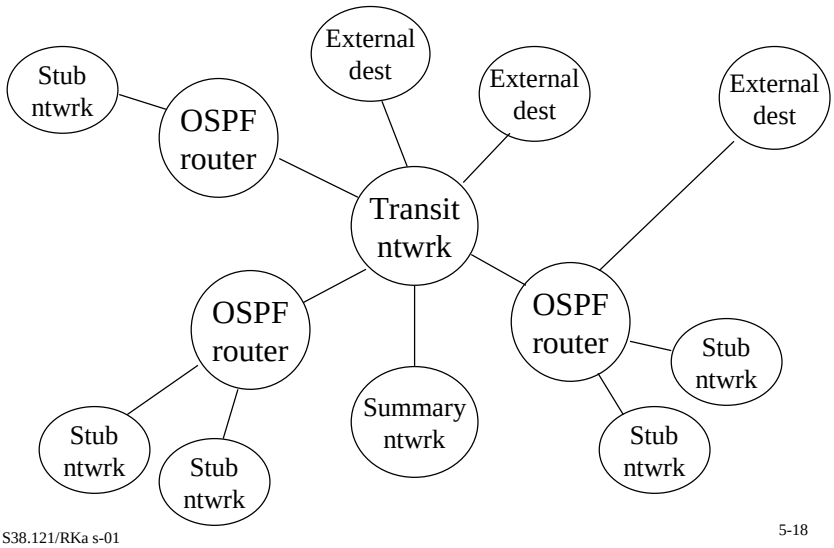
This modified algorithm finds also the alternative paths.

Link state protocol can describe several external routes with accurate metrics

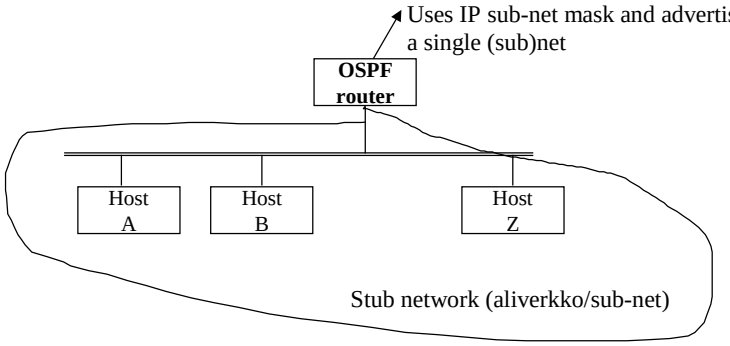
- DV-protocol capability to describe external routes is limited due to *counting to infinity* problem and due to *complexity of Bellman-Ford algorithm* (O(N**2)
- Link state protocol is free of those limitations. SPF route computation converges as O(N.logN) - where N = nrof external routes
- E.g. if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

OSPF Protocol Principles

OSPF sees the network as a graph



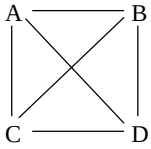
OSPF makes a difference between a router and a host



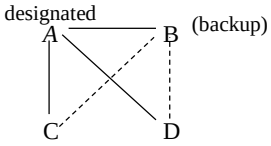
OSPF supports Broadcast networks

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*



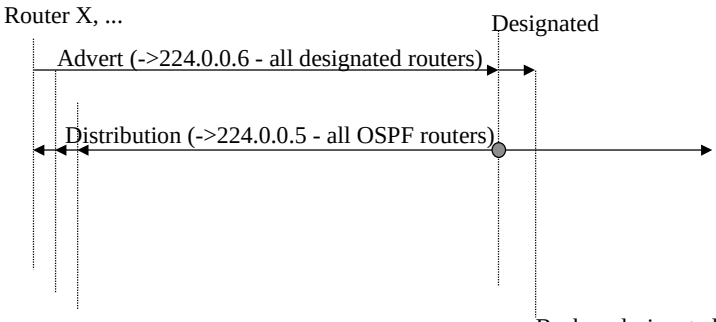
N*(N-1)/2 adjacencies (known neighbors)



Adjacencies are formed only with the Designated router(A) (edusreititin) ==>

- *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.

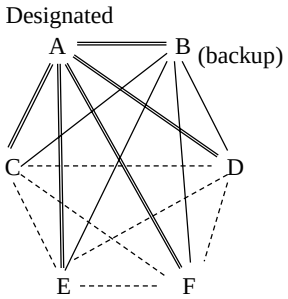
OSPF Flooding Protocol in a Broadcast network



Backup designated stays as silent as possible

==> no need to process acks from all other routers in the sub-net

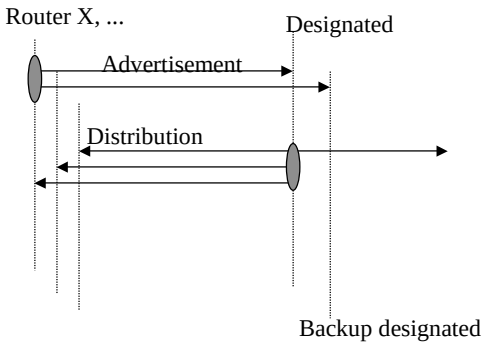
In non-BC nets OSPF works in the same way except that Bcasts are replaced by point-to-point messages



Permanent connection with designated

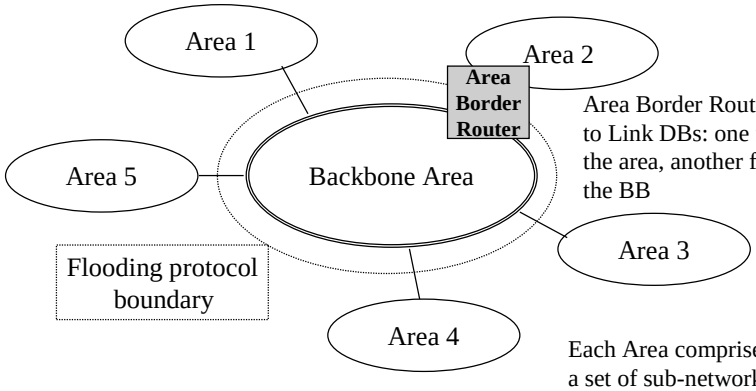
Permanent connection with backup designated

Dial-up connection with other routers



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



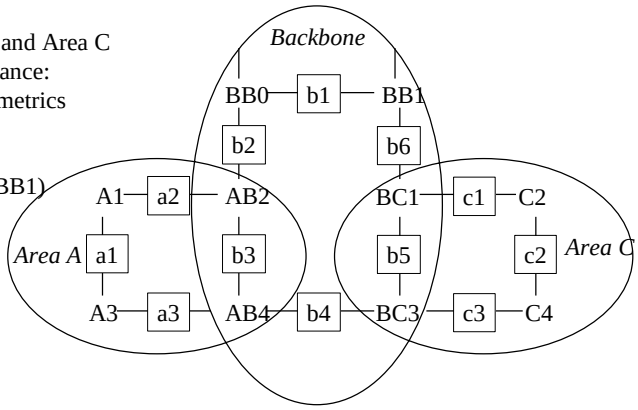
Area Border Router has to Link DBs: one for the area, another for the BB

Each Area comprises of a set of sub-networks

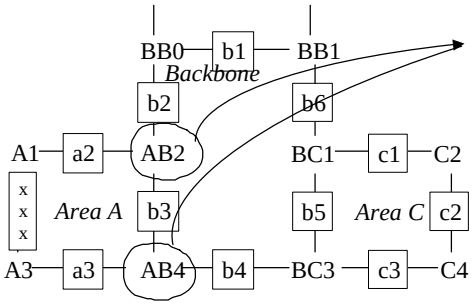
(Area = alue BB = runkoverkko, ABR - aluerajareititin)

(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) with distance: ABx--bz or ABx--cy (metrics are summed).
- external records (<-AB2,AB4<-- BB0,BB1)

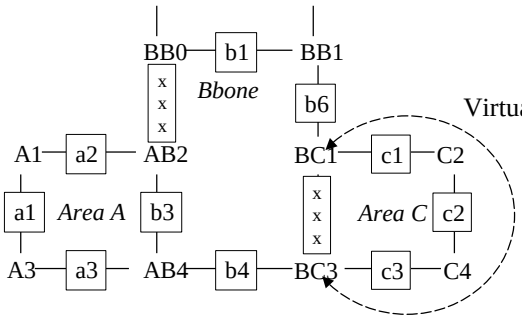


OSPF easily recovers from failures in Areas



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 Bbone failures a Virtual link can help if Bbone 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.

Stub Area = tynkäalue

LSA types in OSPF are

LS Type = 1 Router LSA -- describes set of active interfaces 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 -- descr ext routes

} Hierarchical Routing

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)

NBMA - non-broadcast multiple access e.g. ATM or FR

All OSPF routers on an Area have identical Link DBs

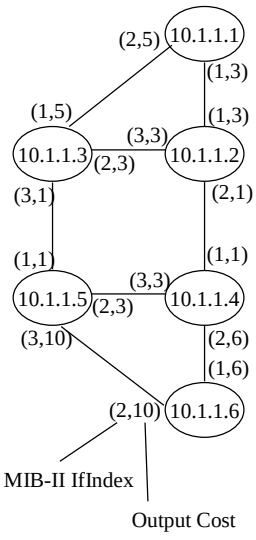
Common header of Link State Advertisement (LSA)

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 ...

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

LS age - seconds from advertisement
options: E - external links
T - type of service -- when many metrics are in use

Example of the Router LSA

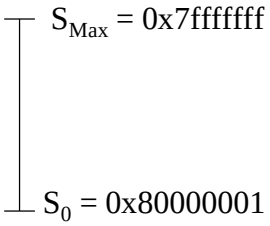


0 seconds	LS Age	Options	LS Type	E-bit, LS Type 1, (Router LSA)
	Link State ID			10.1.1.1
	Advertising Router			10.1.1.1
	LS Sequence Number			0x80000006
0x9b47	LS Checksum	Length		60 bytes
0 (ordinary)	Router Type	0	Nrof links	3
	Link ID			10.1.1.2 (Neighb)
	Link Data			IfIndex 1 (Unnum)
1 (pt-t-pt), 0	Lnk Type	#Tos met	Metric	3
	Link ID			10.1.1.3 (Neighb)
	Link Data			IfIndex 2 (Unnum)
1 (pt-t-pt), 0	Lnk Type	#Tos met	Metric	5
	Link ID			10.1.1.1
	Link Data			255.255.255.255
3 (stub netw), 0 Tos metrics	Lnk Type	#Tos met	Metric	0

Router 10.1.1.1’s router-LSA

Length = 24 + 3 * 12 = 60 bytes
Router with 100 interfaces:
length = 24 + 100 * 12 = 1224 bytes

LSA Sequence Numbers

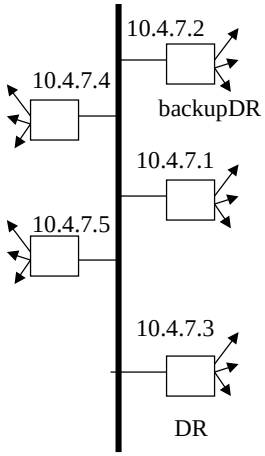


- To roll the space over, first delete record with S_{Max}
- A router may update a self originated record only once in 5 sec.
- In absence of errors rolling the space over takes at least 600 years.
- LS Age is updated during flooding at each step. Records with max Age are discarded. This breaks inf. loops.

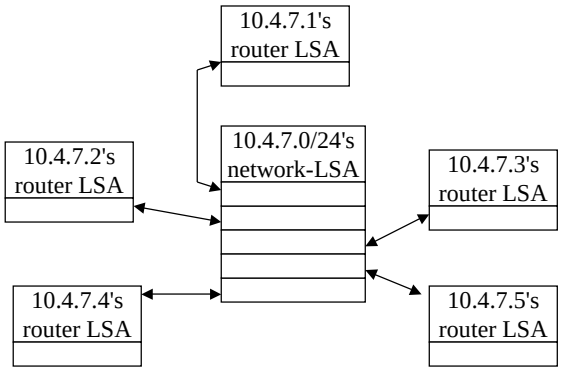
OSPF timeouts - LS Age field

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.

Network-LSA reduces Link DB for BC networks



DR - designated router

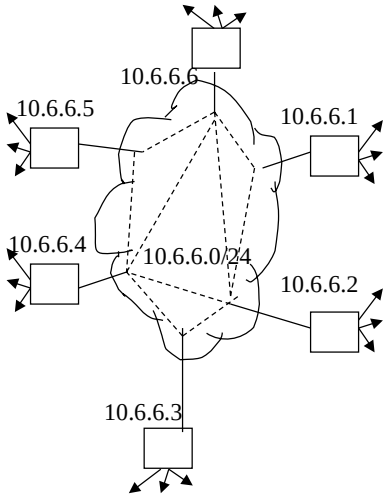


Network-LSA is generated by DR
Network-LSA reflects DB sync status with DRboth ways
This reduces nrof link records from O(n*(n-1)) to n*2.
Particularly important if Network is ATM or FR with a lot of routers attached!

Non-broadcast Multi-access(NBMA) sub-nets support many routers communicating directly but do not have BC capability

- Examples are ATM, Frame Relay, X25
- IP routing requires more manual configuration
- Designated router and backup DR concept reduce the nrof adjacencies
- The model is prone to failures that may be hard to track

Point-to-multi-point sub-net is more robust but less efficient



- There is no DR nor backup DR
- Every OSPF router maintains adjacencies with all neighbors with whom it has direct connectivity
- Alternative is a set of NBMA networks
- Next hop routing protocol improves scalability

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
- Dest IP address = Neighbors IP address or AllSPFRouters (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:

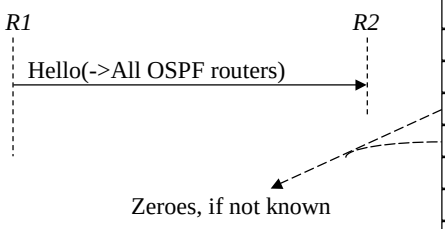
- Hello (huomio) protocol
- Exchange (tiedon vaihto) protocol
- Flooding (levitys) protocol

Common OSPF message header is:

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

OSPF current version is 2.
Type differentiates OSPF message types
Authentication may be based on
- passwords - poor security
- cryptographic methods (since 1995).

Hello protocol ensures that links are working and selects DR and Backup DR



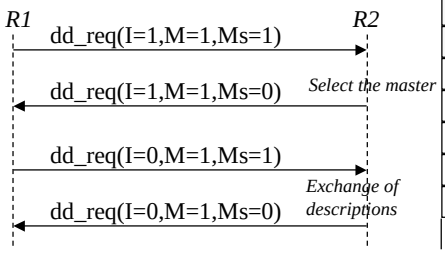
OSPF packet header type = 1		
Network mask		
Hello interval	options	priority
Dead interval		
Designated router		
Backup designated router		
Neighbor		
---------- --------		
Neighbor		

- 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.
- Options - E -external links, T - TOS routing capability.
- Priority tells about eligibility for the role of Designated Router.
- A Hello packet must be sent and received before a link becomes operational

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

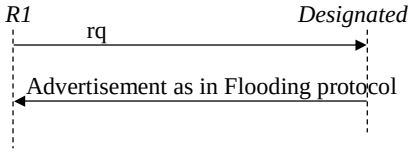


OSPF packet header type = 2 (dd)			
0	0	options	0 IMMs
dd sequence number			
Link state type			
Link state ID			
Advertising router			
Link state sequence number			
Link state checksum		Link state age	
----- -----			

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

I = initialize
M= more
Ms = Master/slave
(initial packet is colored)

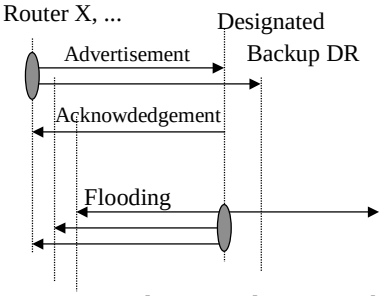
Request packets are used to get record contents. Rqs are acknowledged by Flooding protocol packets



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



OSPF packet header type = 4 (upd.)
Nrof advertisements
Link State Advertisements(see. LSA format)
----- -----

OSPF packet header tupe = 5 (ack.)
LSA headers
----- -----

- 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 Upd message.
- One ack packet can acknowledge may LSAs.

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

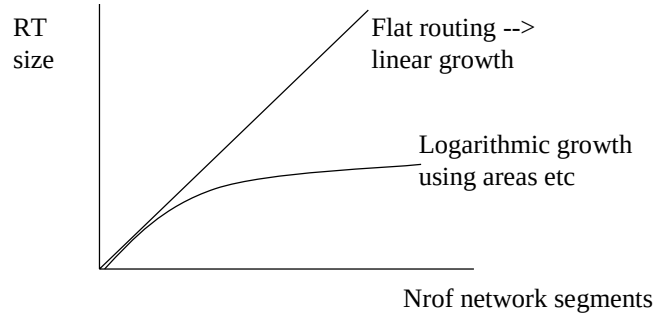
1. Age = nrof hops thru which Advertisement has travelled + seconds from reception
2. Maximum-age = 1 hour
3. Each record has to be advertised at least once in 30 min.
   A fresh advertisement resets the Age and increments record Sequence. nr.
4. When age reaches MaxAge=1h, an advertisement is sent.
5. MaxAge advertisement is accepted and flooded - removes obsolete info.
6. If age difference of Advertisement to DB is small, Advert is not flooded to avoid overloading the network with multiple copies of the same info.
7. If MaxAge record is not found, advertisement has not impact, (router most likely has already removed the dead LSA.)

Summary of OSPF subprotocols

	Hello msg (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

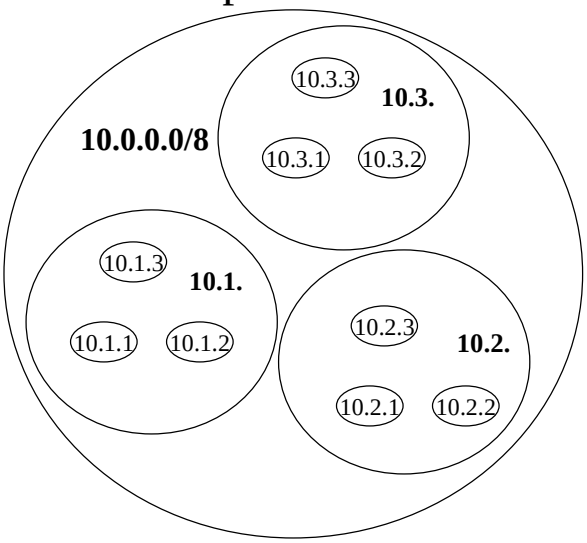
OSPF without Dijksta's algoritm and with more generic data objects is SCSP (Server Cache Synchronization Protocol) which is proposed as a basis for *Telephony Routing Information Protocol* - studied in our Lab. in IMELIO -project.

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



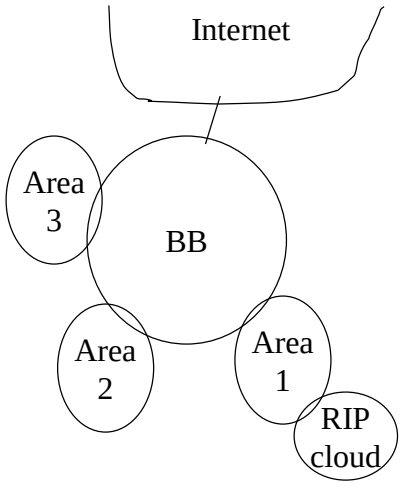
The cost is: sometimes sub-optimal routes.

Example of use 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

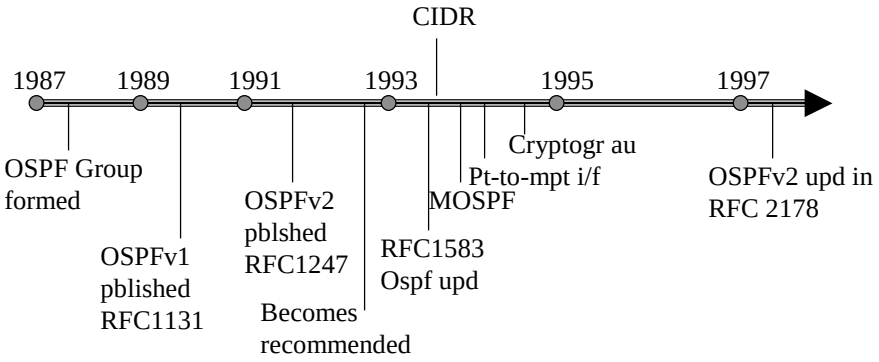
- Type 1 metrics are of the same order as OSPF metrics, e.g. hop count (for RIP and OSPF)
- Type 2 metrics are always more significant then OSPF internal metrics

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



CIDR - Classless Inter Domain Routing

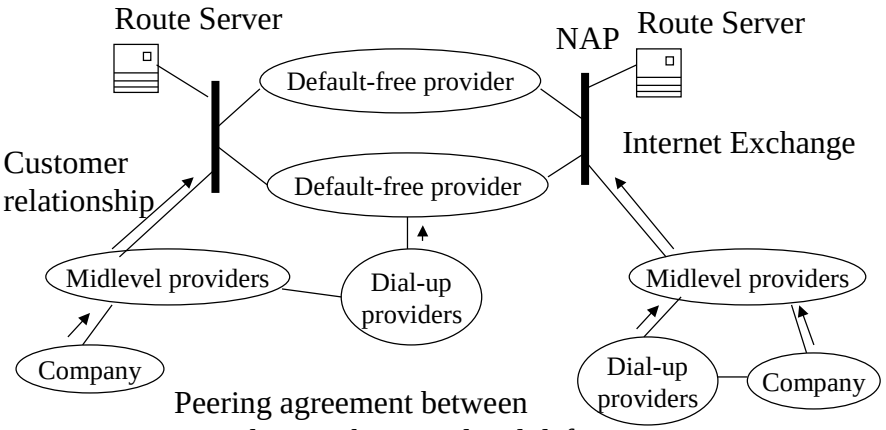
Internet growth has forced the adoption of CIDR address arithmetics to improve the efficiency of using IP address space.

Forwarding process
Router implementation

CIDR affected many routing protocols

- AS - Autonomous System is a part of the Internet owned by a single organization.
- In an AS usually one interior routing protocol is used e.g. OSPF or IS-IS.
- Between Ass exterior routing protocol - currently BGPv4 is used.

Organisation of the Internet as Autonomous Systems

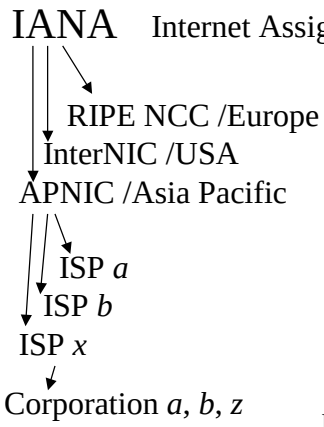


Peering agreement between providers on the same level define exchange of routing information

History of the Internet Core

…..1985 Arpanet
…..1987 NSFNET 56k lines
…..1992 NSFNET T1 lines (1.5M)
…. 1995 NSFNET T3 lines (24M)
1995 NSFNET decommissioned
1995… **Commercial** (UUNET,MCI, Sprint...

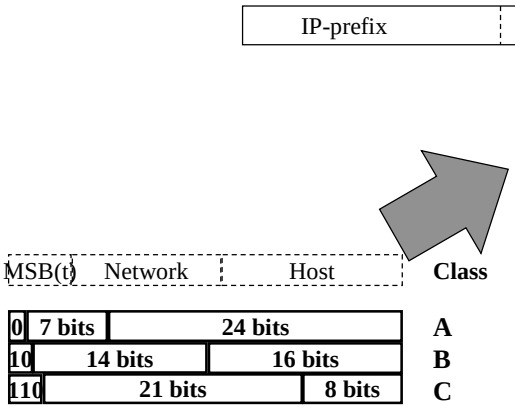
Internet Addresses are assigned by a hierarchy of registrars



- This model leads to provider addressing.
- Due to Provider addressing an ISP needs to advertise shorter prefixes leading to savings in routing table size in the Backbone

ISP= Internet Service Provider

CIDR allows splitting 32-bit IP-addresses freely into prefix and tail



A sequence of C-class addresses can be represented:

194.51.120.0 - 194.51.127.255 =

Start = 194.51.120.0
mask = 255.255.248.0

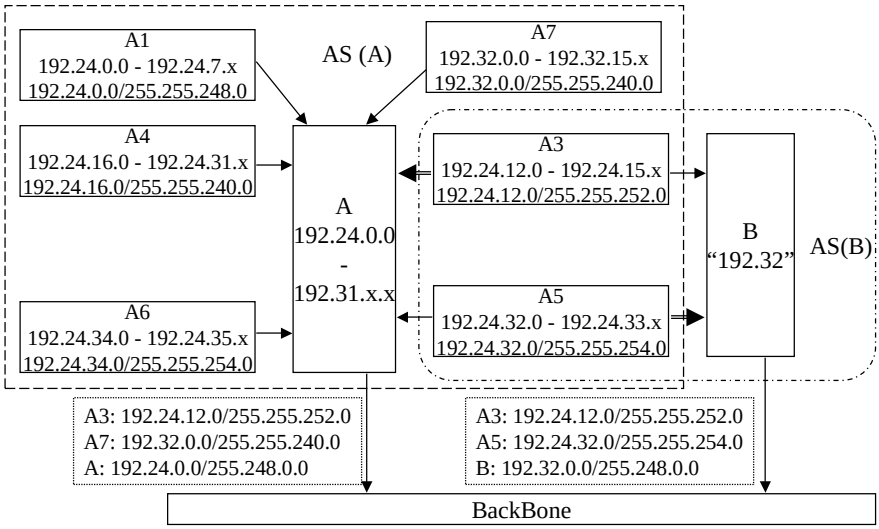
Short notation: 194.51.120.0/21

CIDR changes the way routes are advertised

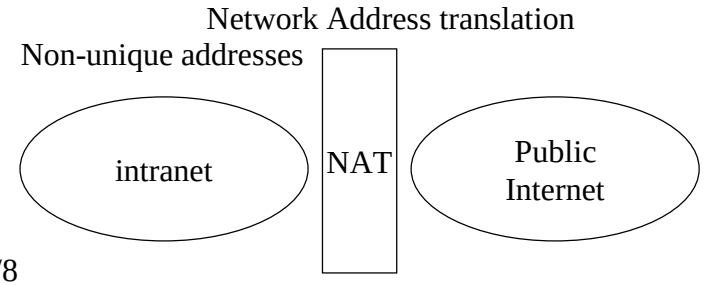
Rule1: Routing always looks for longest match address with the destination.
---> addresses of multi-homed networks can not be aggregated.
(multi-homed network connects to many ASs.)

Rule 2: Network that aggregates a set of routes must delete packets that match with the aggregated prefix but with none of the network addresses that went into the aggregate.
(this helps to avoid loops).

Example



Network Address Translation (NAT) preserves address space and improves security



10/8
172.16/12
192.168/16
--> not routable in public Internet

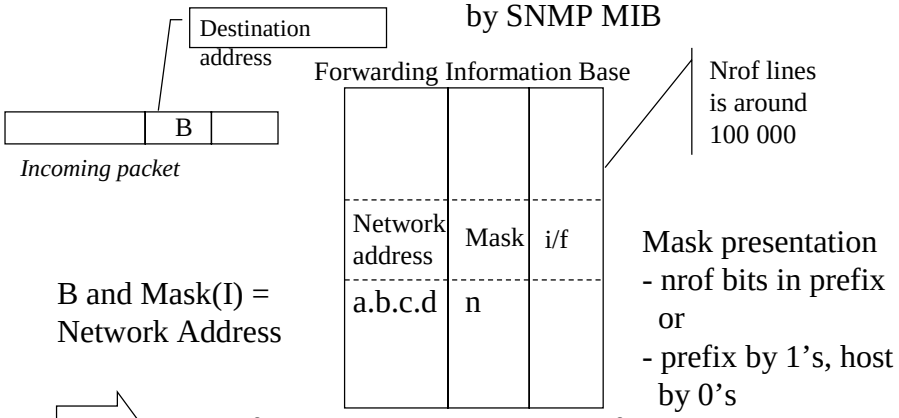
Internal IP-address -- <public IP address, port>

⇒ **Internet has become an A-subscriber network.**

Packet Forwarding and Router Architectures

Packet forwarding in a router

RFC 2097 - view routing table by SNMP MIB



B and Mask(I) = Network Address

Mask presentation
- nrof bits in prefix or
- prefix by 1's, host by 0's

List of Candidate destination interfaces

Take the longest match (= longest mask)

Modifications of forwarding process

- Multipath routing
  - e.g. hash(source IP address, destination IP address) produces one of the possible next hops.
- TOS routing
  - never widely used
  - has been removed from recommendations (RFC 2178 in 7/97)
- Source Routing (strict or loose)
  - max 9 hops can be specified in header options
  - has performance penalty
  - is considered a security hole (all packet may be dropped)

More modifications of the forwarding process

- When there are too many packets to forward, some need to be dropped. To maintain a fair service drop algorithms are used
  - e.g Random Early Detection (RED)
- Scheduling algorithms manage the share of connections in the available bandwidth
  - e.g give 15 kbit/s to an audio conference or half of the link bandwidth to interactive services
  - Weighted Fair Queuing (WFQ) and Class Based Queuing (CBQ) are examples of scheduling algorithms

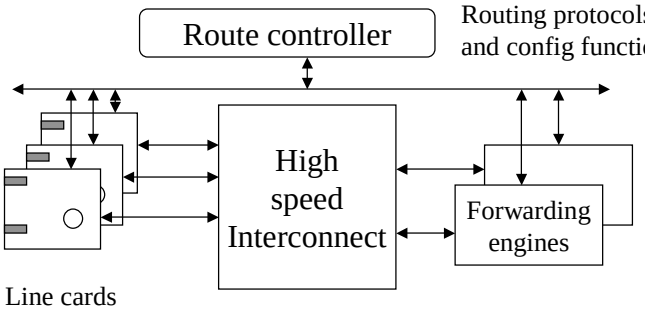
Routers support Security and problem resolution

- Security includes e.g. preventing unauthorised access to a company intranet
  - we talk about Firewalls
  - forwarding needs to check filtering rules on IP addresses and TCP port numbers
  - ISP routers may check all source IP addresses to trace security attacks
- A router may support RMON MIB
  - router allows traffic tracing for routing problem analysis

Routers can collect Statistics

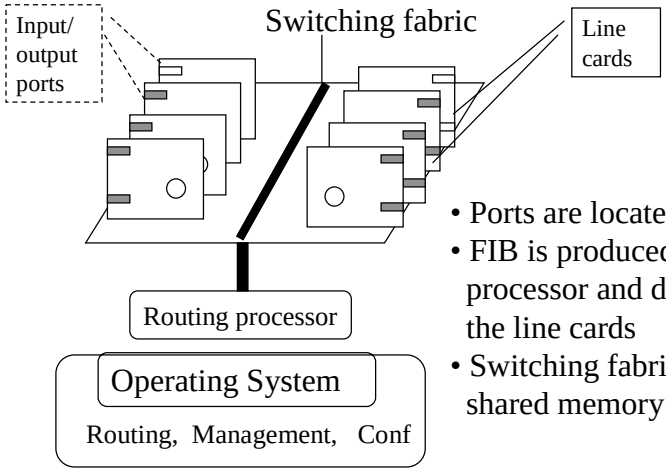
- Statistics are needed for Network Planning,
- Inter ISP accounting and
- Usage based charging

A non-trivial router architecture is



Forwarding capacity is increased by sharing the load between several forwarding engines

A faster router architecture is



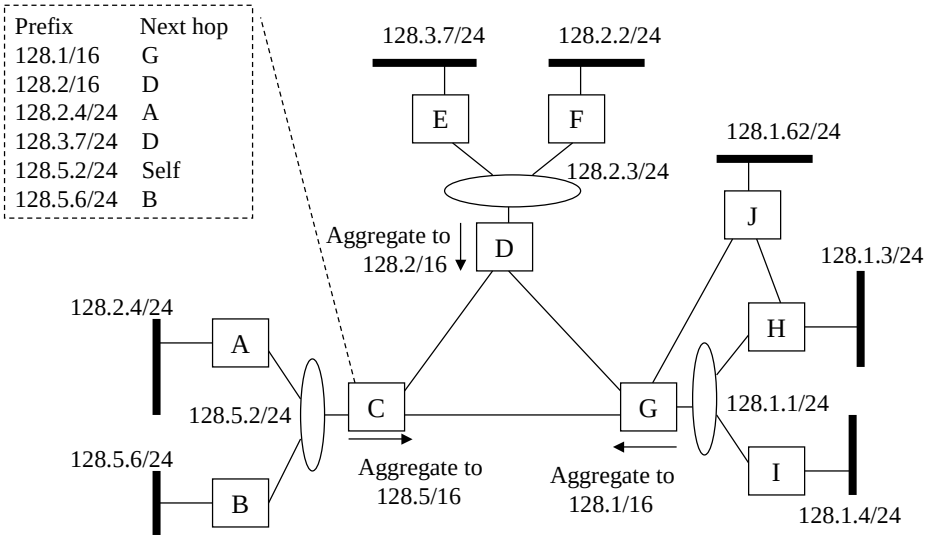
- Ports are located on line cards
- FIB is produced by Routing processor and downloaded to the line cards
- Switching fabric is a bus or shared memory

FIB Lookup speed is determined by nrof reads

- E.g. access time 30ns x 8 reads = 240 ns ⇒ >4 M lookups/s
- May need to backtrack --> poor worst case performance.
- Ways to improve performance:
  - Hardware oriented techniques
  - Table compaction techniques
    -- e.g. long trie branches with few leaves are packed into a node
  - Hashing techniques
    -- problem is unknown mask length -->
    -- (e.g. binary search on prefix length)

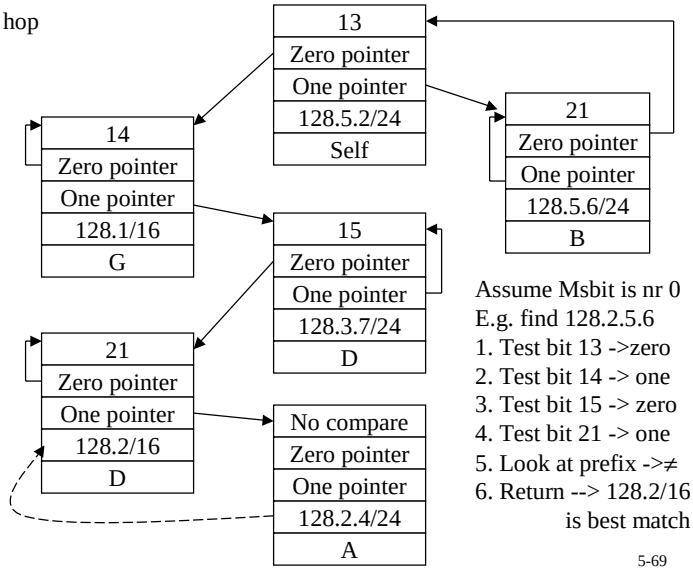
Route aggregation example

Prefix	Next hop
128.1/16	G
128.2/16	D
128.2.4/24	A
128.3.7/24	D
128.5.2/24	Self
128.5.6/24	B



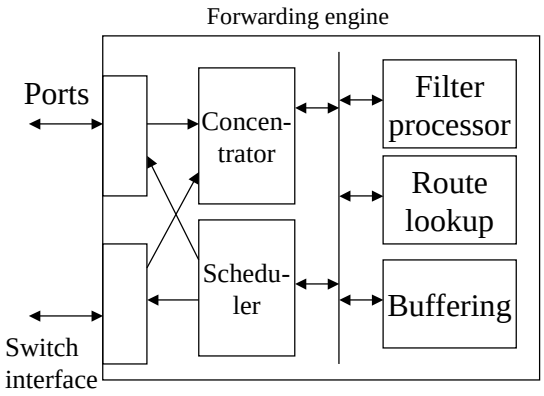
Route lookup may be based on Patricia tree

Prefix	Next hop
128.1/16	G
128.2/16	D
128.2.4/24	A
128.3.7/24	D
128.5.2/24	Self
128.5.6/24	B



Assume Msbit is nr 0
E.g. find 128.2.5.6
1. Test bit 13 ->zero
2. Test bit 14 -> one
3. Test bit 15 -> zero
4. Test bit 21 -> one
5. Look at prefix ->≠
6. Return --> 128.2/16 is best match

Forwarding speed can be increased by parallel processing



Example based on Bell Labs prototype.

Boxes are 33MHz… 66MHz FPGAs.

Can process all headers prior to buffering at 1Gbit/s line speeds!

--> can provide QoS.

Forw time = 40 x 8 b/1G/s = 320 ns.