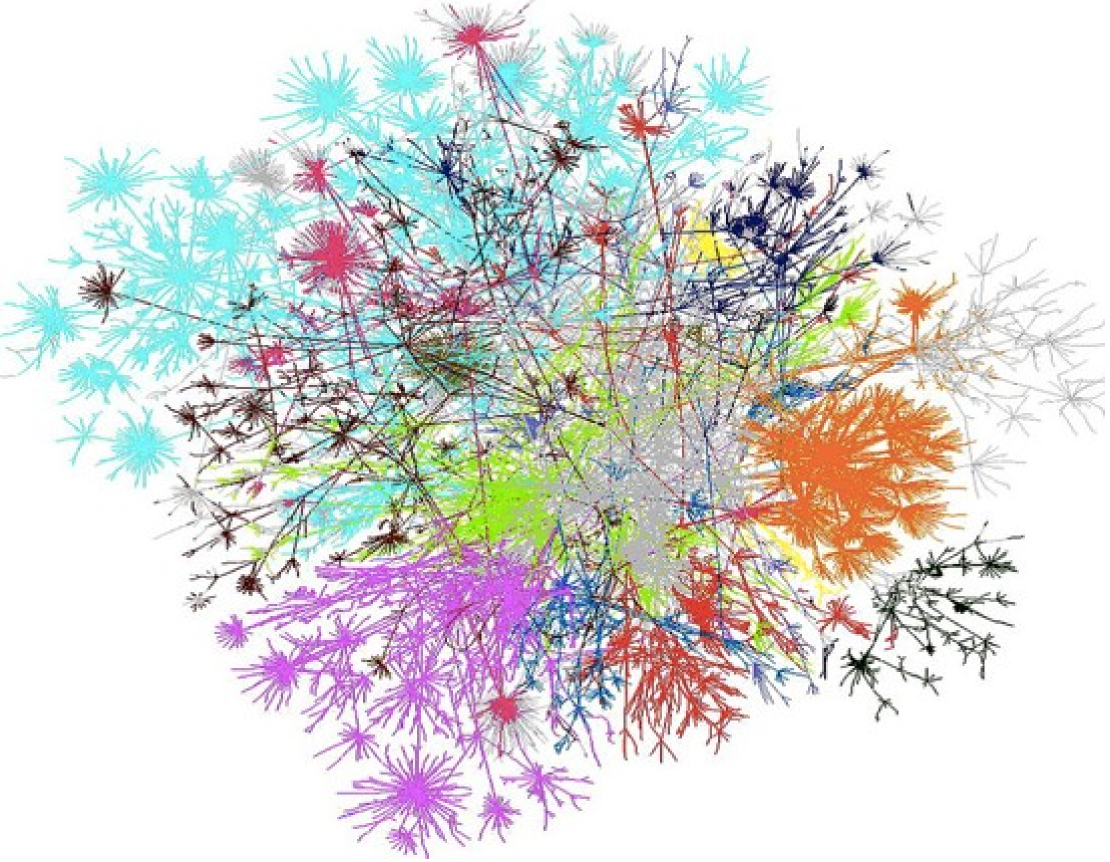
CS 31006: Computer Networks – Internet Routing



Department of Computer Science and Engineering

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Rajat Subhra Chakraborty
rschakraborty@cse.iitkgp.ac.in

Sandip Chakraborty
sandipc@cse.iitkgp.ac.in

Limitations of Distance Vector Routing / RIP

- The resolution to the counting to infinity problem enforces a maximum cost for a network path (generally 15 in RIP). This limits the diameter of a AS to a maximum of 15 hops.

- High signaling overhead - Periodic broadcasting of the distance vector table can result in increased utilization of the network resources for signaling.

- The algorithm is relatively slow to converge; you require information from all the nodes in the AS.

Link State Routing

- 1979: The ARPANET routing protocol was replaced by link state routing, as an impact to count-to-infinity problem (convergence become slow)

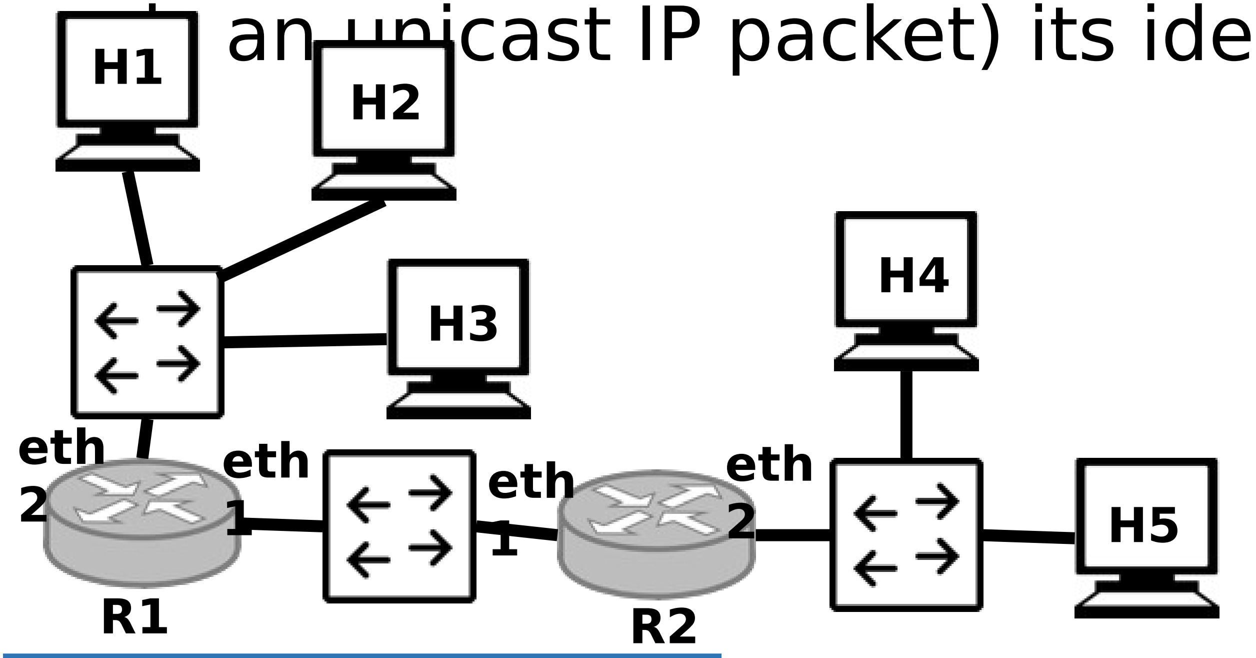
- The routing protocol – **Open Shortest Path First (OSPF)**

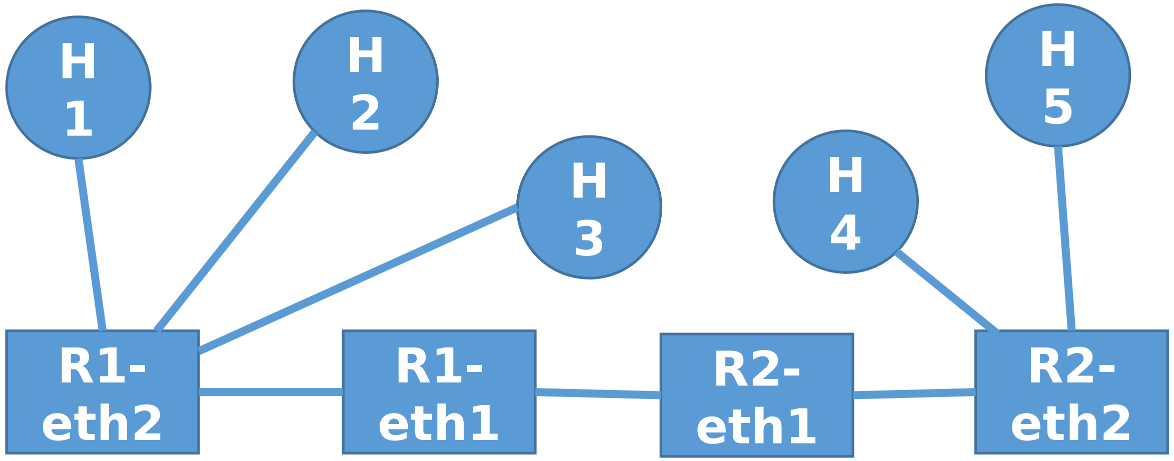
- The protocol is fairly simple
  - Discover neighbors and learn their network addresses
  - Set the distance or cost metric to each of the neighbors
  - Construct a packet telling all it has learned
  - Broadcast this packet – every router periodically learns the **link state** of the network graph
  - Compute the shortest path to every other routers

LSR - Learning about Routers

- When a router is booted, it first learn the neighbors – broadcast a HELLO packet on each point to point line **– note the use of broadcast IP address**

- Once a router receives a HELLO message, it sends back (this is an unicast IP packet) its identity (IP address)





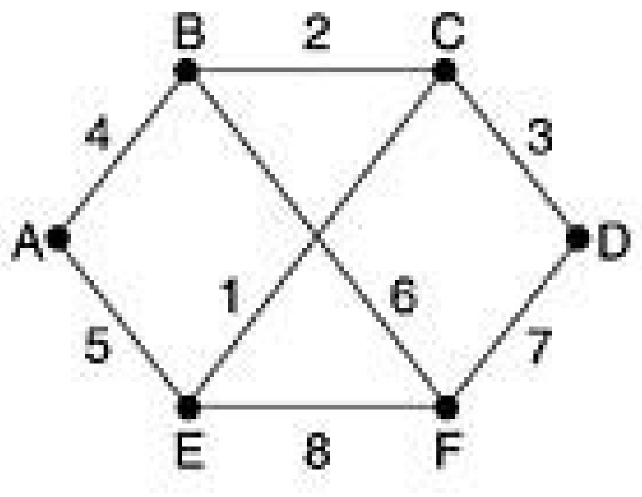
LSR– Setting Link Costs

- Each link is assigned with a link cost or distance (hop count, delay) which is used as the routing metric to find out the shortest path

- A standard approach – inverse of the link bandwidth - higher capacity paths are better choices (minimize the routing cost)

- Some networks use link delay - computed through a ICMP ECHO packet from the IP layer
  - ICMP – **Internet Control Message Protocol –** a set of message suites for IP layer management functionalities
  - ICMP Echo Request and ICMP Echo Reply

LSR – Building Link State Packets



(a)

Link State Packets

A		B		C		D		E		F	
Seq.		Seq.		Seq.		Seq.		Seq.		Seq.	
Age		Age		Age		Age		Age		Age	
B	4	A	4	B	2	C	3	A	5	B	6
E	5	C	2	D	3	F	7	C	1	D	7
		F	6	E	1			F	8	E	8

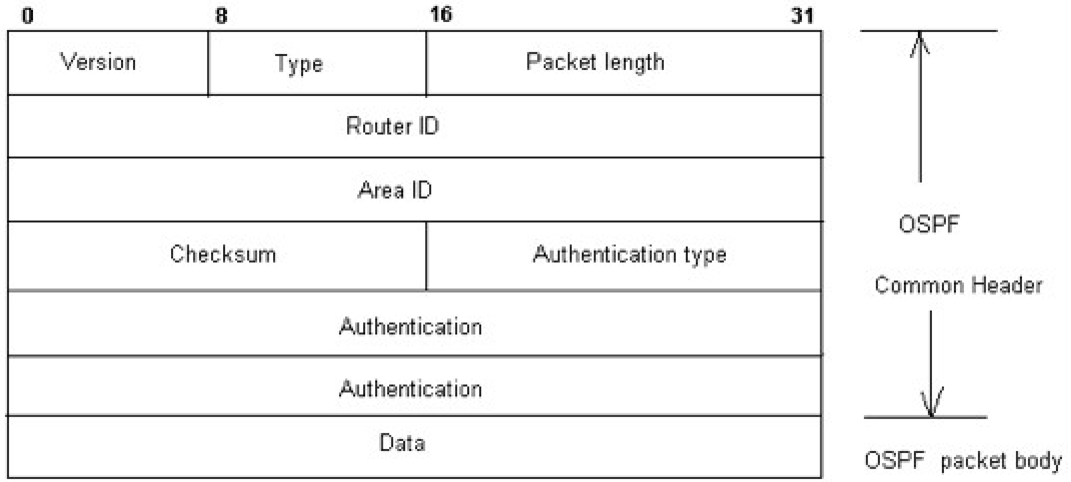
(b)

Image: Computer Networks, Andrew S. Tanenbaum, David J. Wetherall

LSR – Distributing the Link State Packets

- All of the routers must get all of the link state packets quickly and reliably
  - If different routers have different information, the routing inconsistency may occur

- Use flooding to distribute the link state packets to all the



s a sequence number that is

ket sent – **used to identify stale**

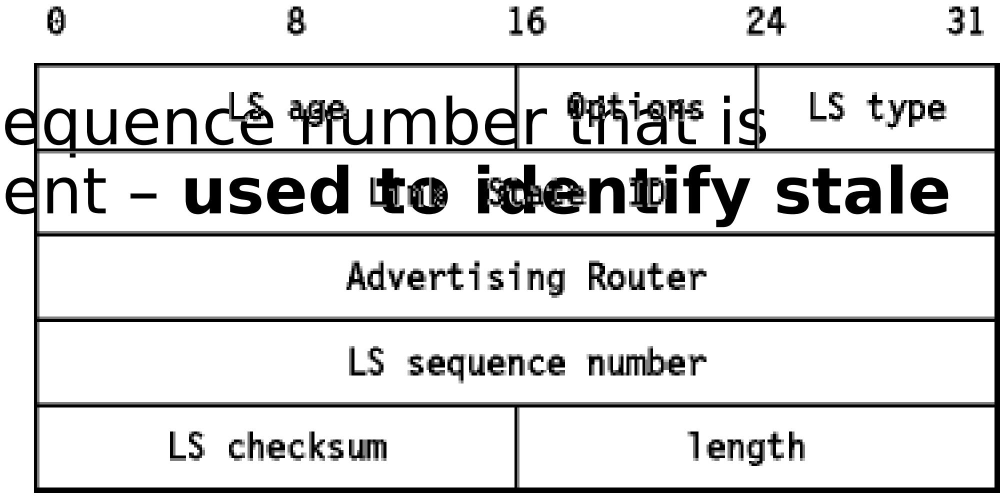


Image: http://www.danzig.jct.ac.il/tcp-ip-lab/ibm-tutorial/3376c33.html

LSR – Distributing the Link State Packets

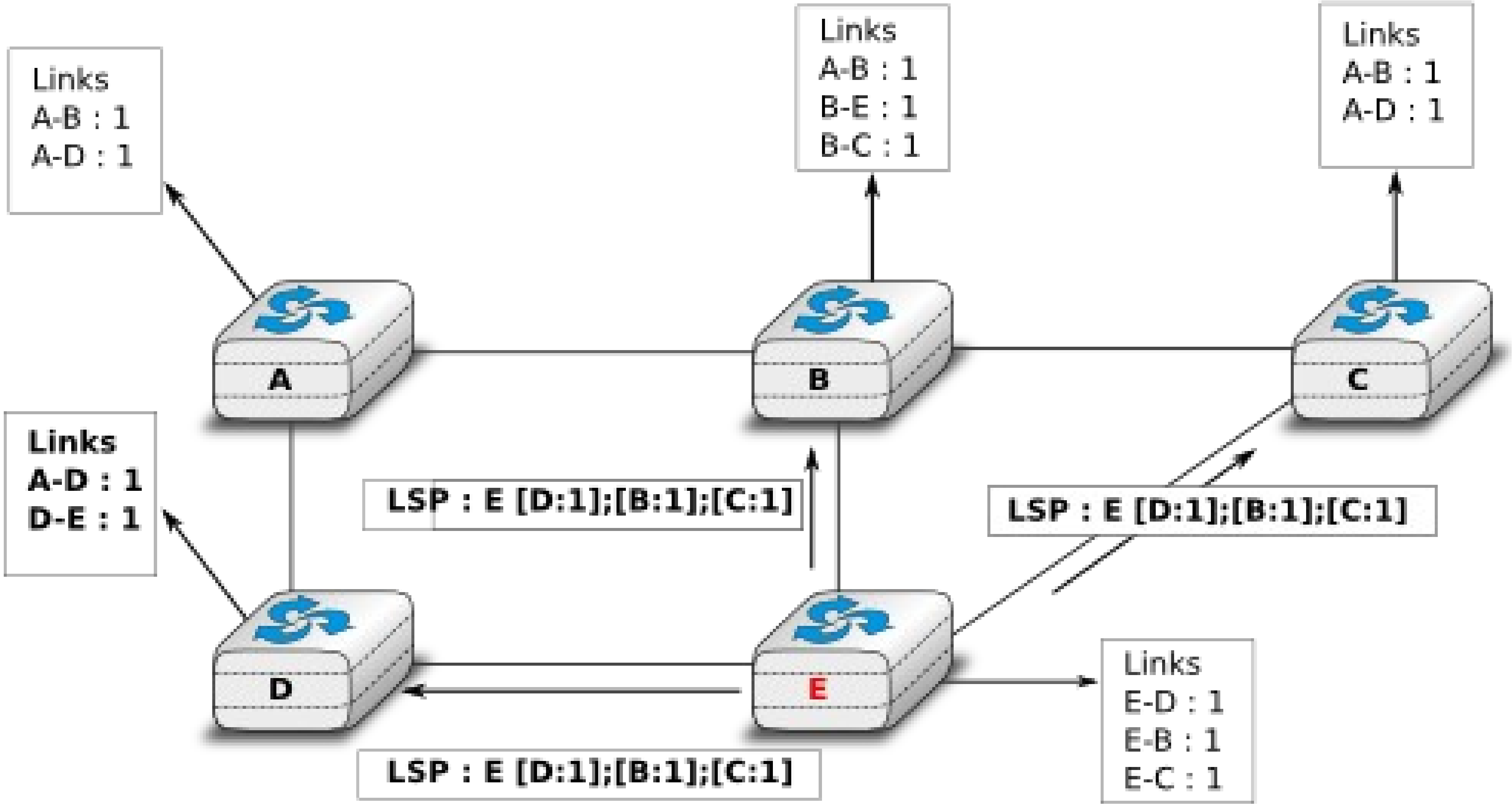


Image: http://cnp3book.info.ucl.ac.be/1st/html/network/network.html

LSR – Distributing the Link State Packets

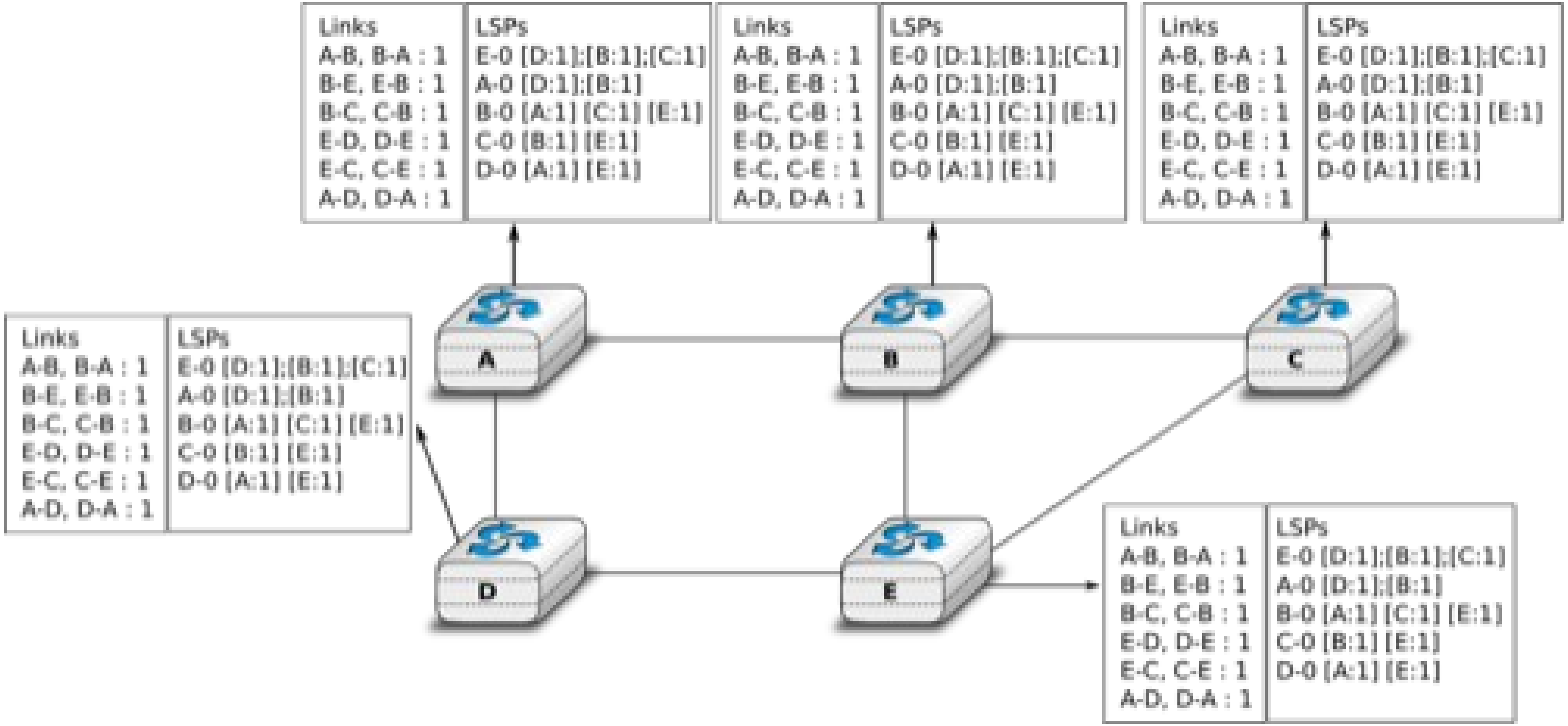
- Once a Link State (LS) packet is received, the sequence number is checked
  - If the sequence number is higher than the last observed LS packet, then it is accepted; otherwise it is discarded

- **What if the sequence number wraps around?**
  - Use a 32 bit sequence number, and 1 LS packet per second – it would take 137 years to wrap around

- Every entry in the router is associated with an **age** – denotes the lifetime of an entry
  - Deletes the old entries from the routing table

LSR – Computing the New Routes

- Construct the network graph from the link state packets



Links	LSPs
A-B, B-A : 1	E-0 [D:1];[B:1];[C:1]
B-E, E-B : 1	A-0 [D:1];[B:1]
B-C, C-B : 1	B-0 [A:1] [C:1] [E:1]
E-D, D-E : 1	C-0 [B:1] [E:1]
E-C, C-E : 1	D-0 [A:1] [E:1]
A-D, D-A : 1	

Links	LSPs
A-B, B-A : 1	E-0 [D:1];[B:1];[C:1]
B-E, E-B : 1	A-0 [D:1];[B:1]
B-C, C-B : 1	B-0 [A:1] [C:1] [E:1]
E-D, D-E : 1	C-0 [B:1] [E:1]
E-C, C-E : 1	D-0 [A:1] [E:1]
A-D, D-A : 1	

Links	LSPs
A-B, B-A : 1	E-0 [D:1];[B:1];[C:1]
B-E, E-B : 1	A-0 [D:1];[B:1]
B-C, C-B : 1	B-0 [A:1] [C:1] [E:1]
E-D, D-E : 1	C-0 [B:1] [E:1]
E-C, C-E : 1	D-0 [A:1] [E:1]
A-D, D-A : 1	

Links	LSPs
A-B, B-A : 1	E-0 [D:1];[B:1];[C:1]
B-E, E-B : 1	A-0 [D:1];[B:1]
B-C, C-B : 1	B-0 [A:1] [C:1] [E:1]
E-D, D-E : 1	C-0 [B:1] [E:1]
E-C, C-E : 1	D-0 [A:1] [E:1]
A-D, D-A : 1	

Links	LSPs
A-B, B-A : 1	E-0 [D:1];[B:1];[C:1]
B-E, E-B : 1	A-0 [D:1];[B:1]
B-C, C-B : 1	B-0 [A:1] [C:1] [E:1]
E-D, D-E : 1	C-0 [B:1] [E:1]
E-C, C-E : 1	D-0 [A:1] [E:1]
A-D, D-A : 1	

Image: http://cnp3book.info.ucl.ac.be/1st/html/network/network.html

Link State Routing - OSPF

- Link State Messages – **Overhead for the LSR**

- **OSPF Messages**
  - HELLO: Used to establish neighborhood
  - Database Descriptor (DD or DBD): Broadcast the local routing database among neighbors – check consistency of database information among routers
  - Link State Requests (LSR): Explicitly requests for link state information based on the database comparison
  - Link State Updates (LSU): Forward link state information
  - Link State Acknowledgements (LSAck): Acknowledges the receipt of link state information

Inter-domain Routing Protocols

- Routing between two autonomous systems (AS) – we call them as **routing domains**
  - **Exterior Gateway Protocol (EGP)**

- **Multi-homed AS:** has connection to more than one AS

- **Stub AS:** Connected to only one other AS

- **Transit AS:** Provides connection to other AS

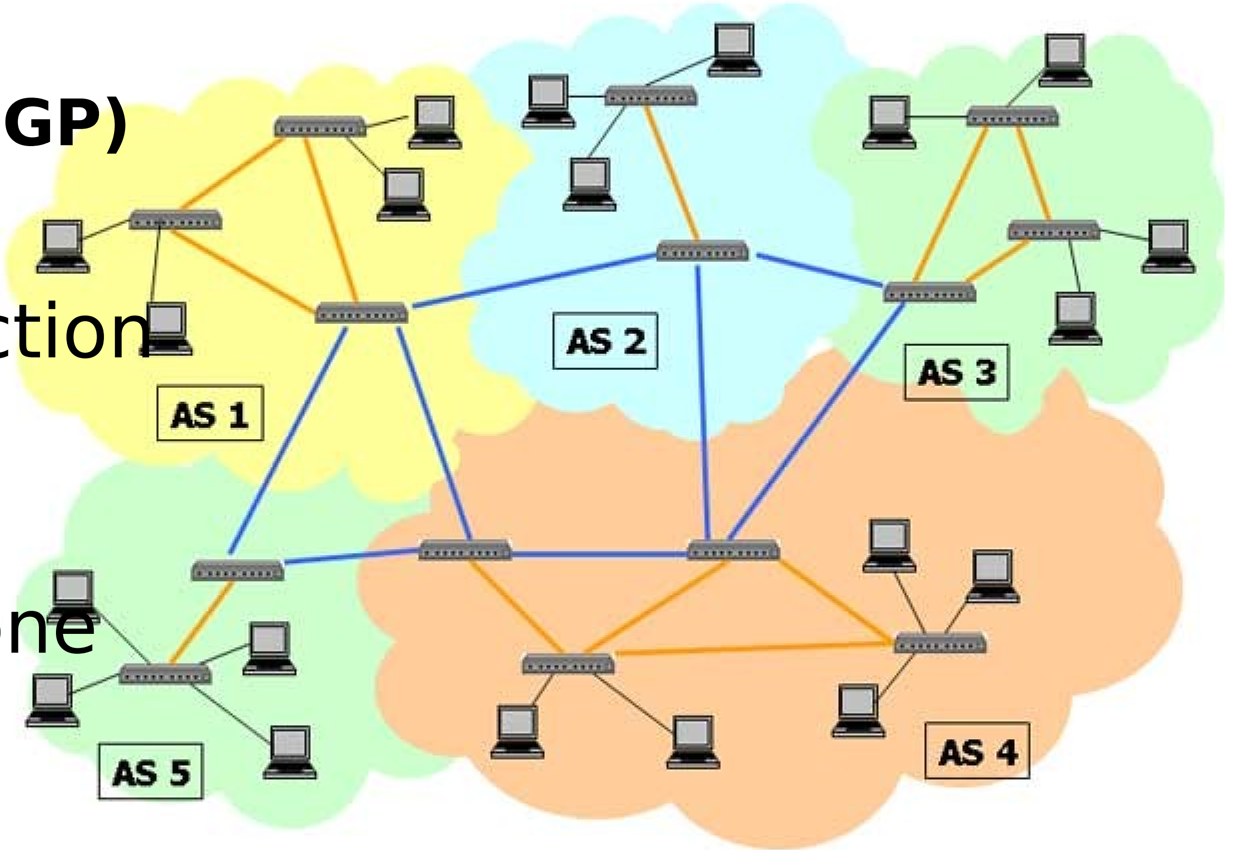


Photo courtesy: http://www.web3.lu/

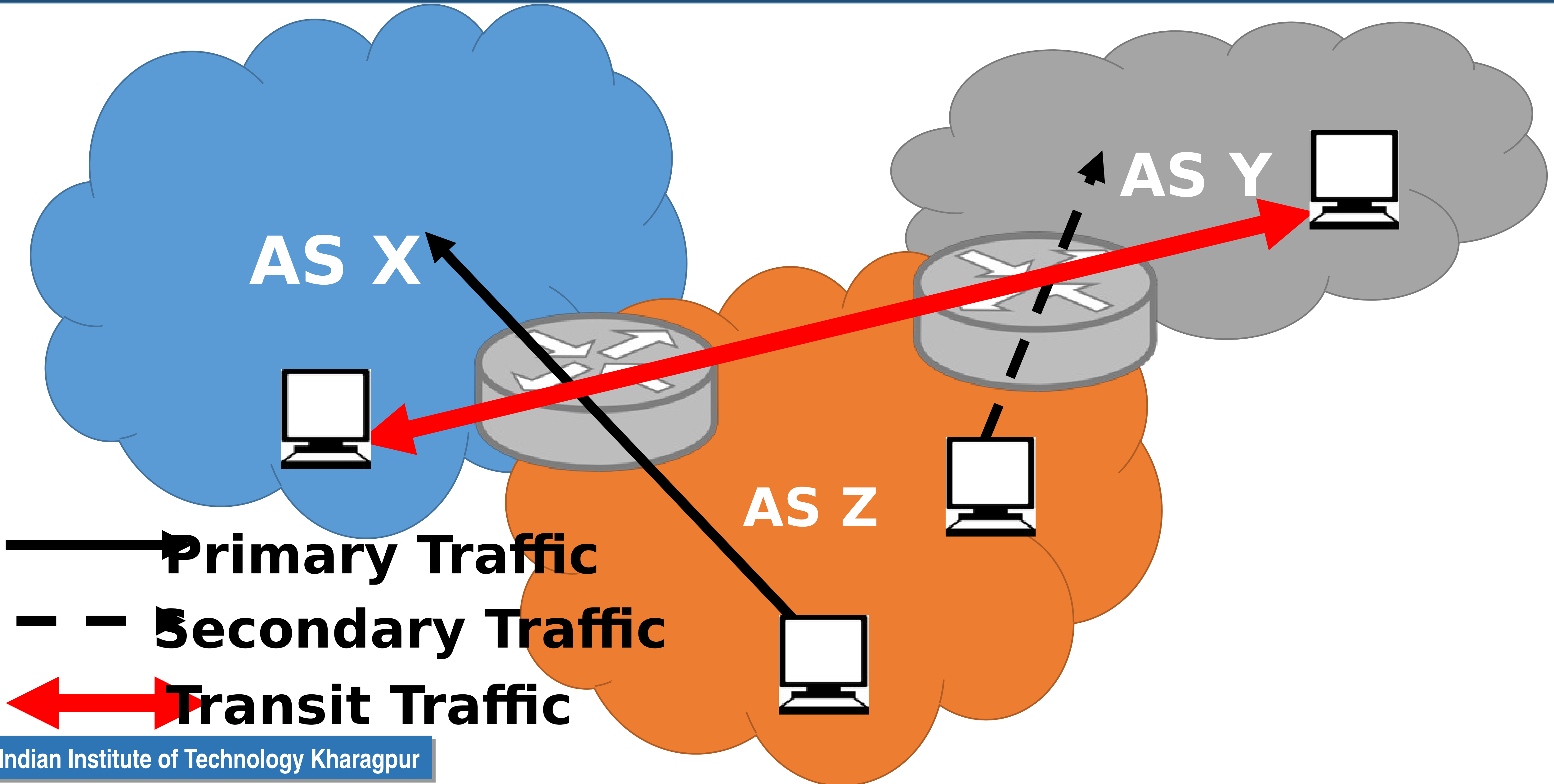
Routing Protocols inside and across AS

- Each AS can run its own intra-domain routing protocols, we call them as **interior gateway protocols (IGP)**
  - Open Shortest Path First (OSPF)
  - Routing Information Protocol (RIP)
  - Can even use static routing or a mixed of IGPs at different subnets

- Inter-domain routing problem – **the AS shares *reachability information* –** description of the set of IP addresses that can be reached via a given AS

- **Challenge –** Each AS has to determine its own ***routing policies*** (can be complex)
  - *Whenever possible, I prefer to send traffic via AS X than via AS Y, but I'll use AS Y if it is the only path; and I never want to carry traffic*

Routing Policies in the Internet



Border Gateway Protocol (BGP)

- The initial protocol (called EGP) was designed for specialized topology, such as a tree topology.

- BGP replaces EGP – generalizes the topology structure of the Internet.

- BGP assumes that the Internet is an arbitrary interconnected set of ASs.

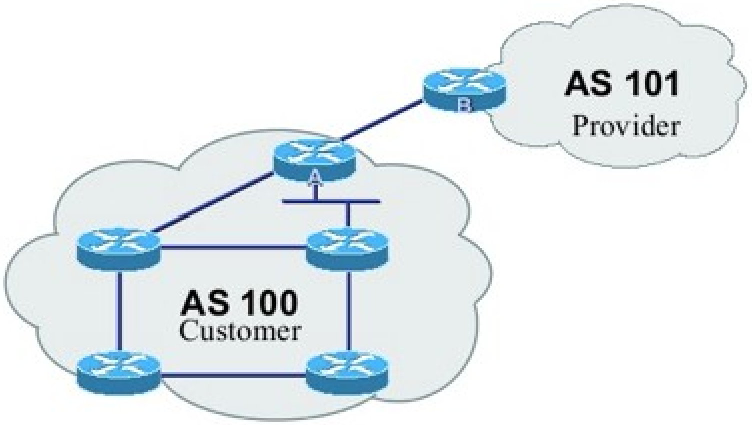
- **Local Traffic:** Originates at or terminates on nodes within an AS

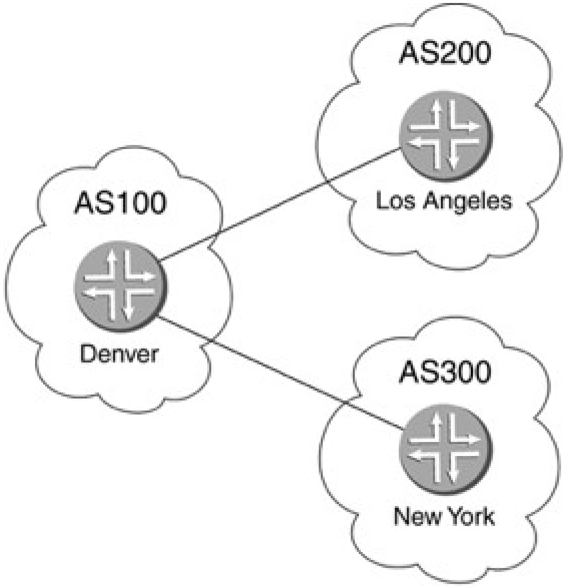
Traffic through an AS

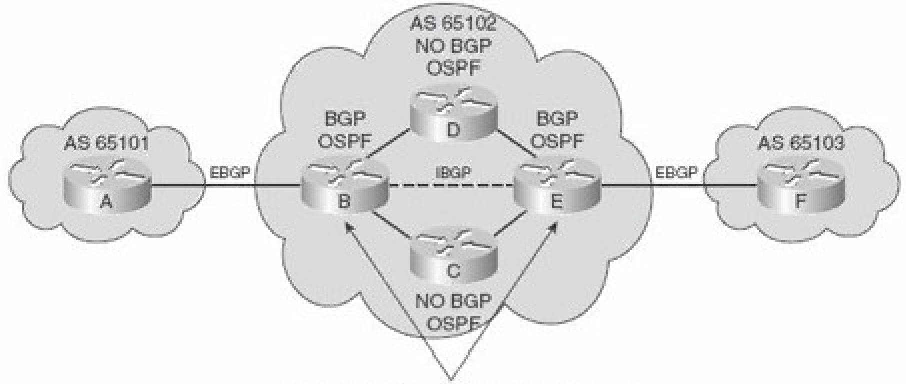
- **Stub AS:** Only carry local traffic

- **Multi-homed AS:** Only carry local traffic, refuses to carry transit traffic

- **Transit AS:** Carry both transit and local traffic







Redistributing BGP into OSPF is not recommended; instead run IBGP on all routers within the AS.

Objectives of BGP

- **Best non-looping policy-complaint path**
  - Loop free path through the ASs
  - Complaint with the policies of the various ASs along the path

- **Scaling** – CIDR at the Internet scale may not be scalable – You need to store IP/netmask information for thousands of subnets
  - Define paths by AS numbers, not the IP; Example AS12-AS14-AS76-AS132-AS45-AS61

- **Path Cost** – the autonomous systems are "*autonomous*" – every AS has their own interior routing protocol and own routing metrics – how to define a path metric for a BGP path?
  - Advertise only the ***reachability*** information, no path costs are

Network Layer Reachability Information (NLRI)

- AS Reachability information shared by the BGP supported routers (called the BGP peers)

- Shared between BGP peers through BGP UPDATE messages

- An IPv4 prefix (with IPv4 protocol) with the corresponding network IP
  - Example: 110.12, /16

- Used to find out the AS reachability as well as route aggregation through CIDR
  - Combine 110.12/16 and 110.12.8/24 together to form 110.12/16 – You do not need to maintain duplicate paths

BGP Speakers and BGP Peers

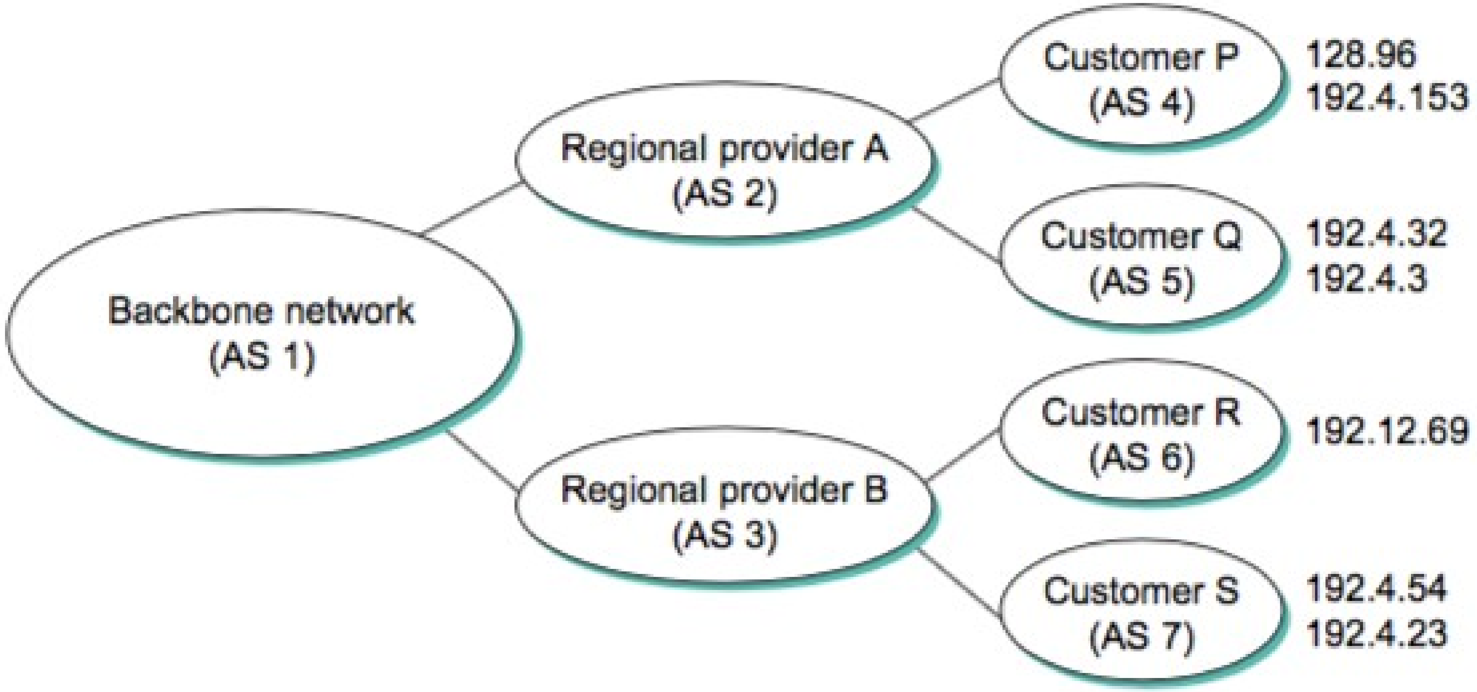
- **BGP Speakers** – A router configured with BGP – a spoke-person for the entire AS
  - Advertises the reachability information for this AS

- Once initialized, uses the well known BGP port (TCP port 179) to connect to other configured BGP peers in the Internet, and share the AS reachability information

- BGP speakers advertised the path information with the BGP speakers in the peer ASs.

BGP Path Information

- BGP advertises ***complete path*** information as enumerated list of ASs to reach a particular network
  - Necessary to enable the sorts of policy decisions
  - Enables routing loops to be readily detected



- AS 2 can advertise NLRI for the subnets given to Customer P and Customer Q

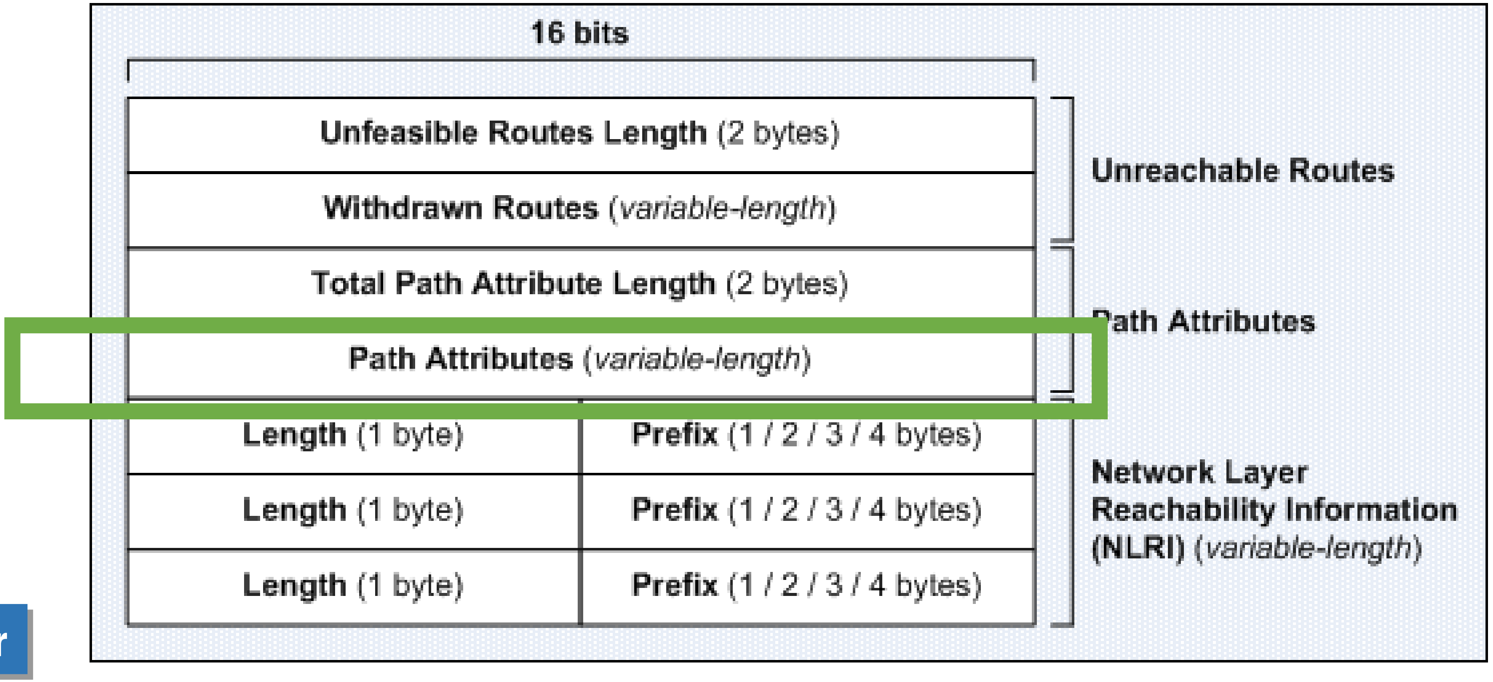
- AS 3 can reach to the subnet 128.96 via the AS Path AS3-AS1-AS2-AS4

Image courtesy: Computer Networks, Larry L Peterson and Bruce S Davie

BGP AS_PATH Attribute in UPDATE Message

- Stores all the paths across various ASs through which a BGP UPDATE message has passed

- Every BGP speaker, when receives an UPDATE message, appends its own AS number and advertise that to the BGP peers



16 bits

Unfeasible Routes Length (2 bytes)

Withdrawn Routes (*variable-length*)

Unreachable Routes

Total Path Attribute Length (2 bytes)

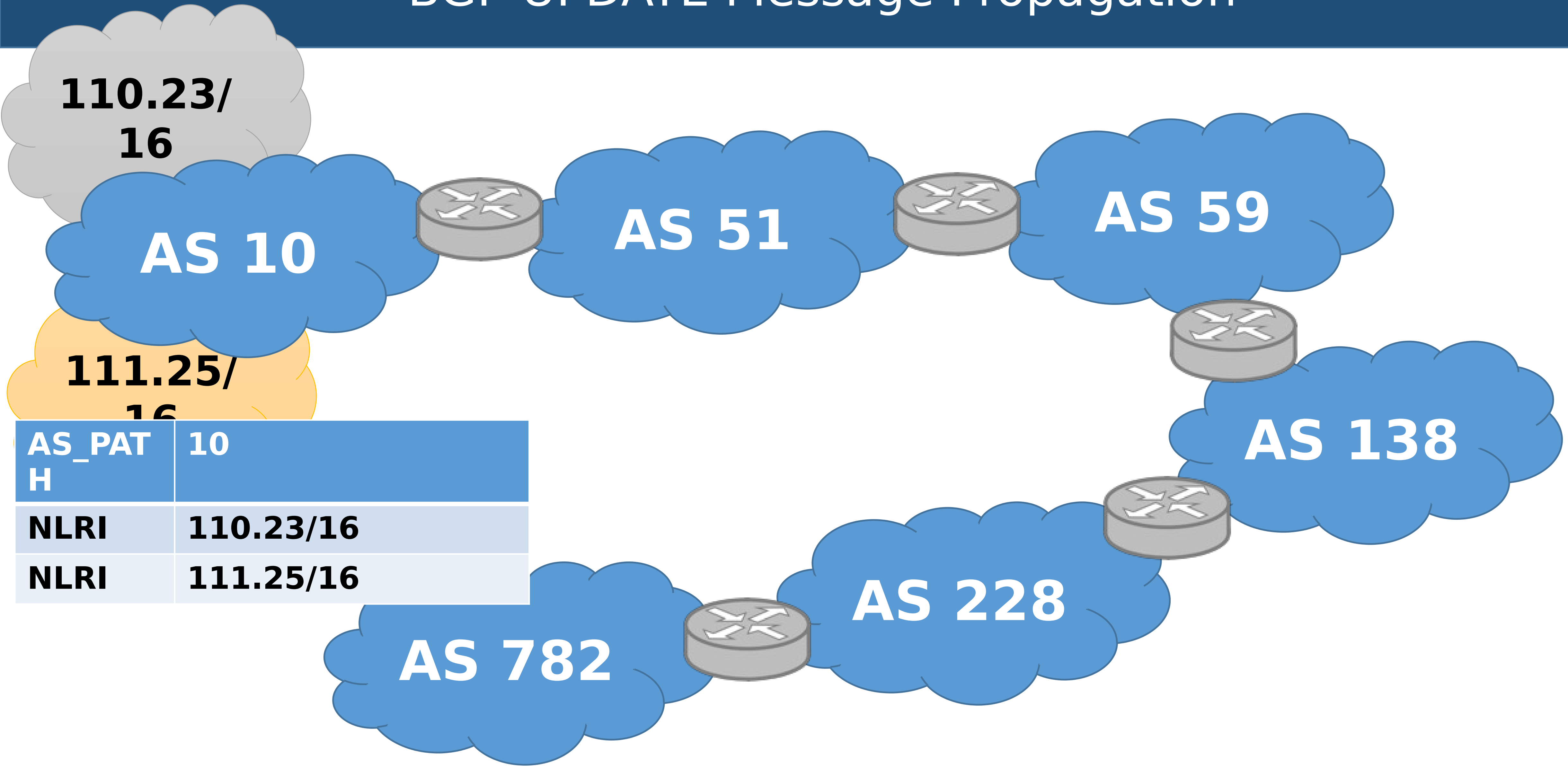
Path Attributes (*variable-length*)

Path Attributes

Length (1 byte)	**Prefix** (1 / 2 / 3 / 4 bytes)
Length (1 byte)	**Prefix** (1 / 2 / 3 / 4 bytes)
Length (1 byte)	**Prefix** (1 / 2 / 3 / 4 bytes)

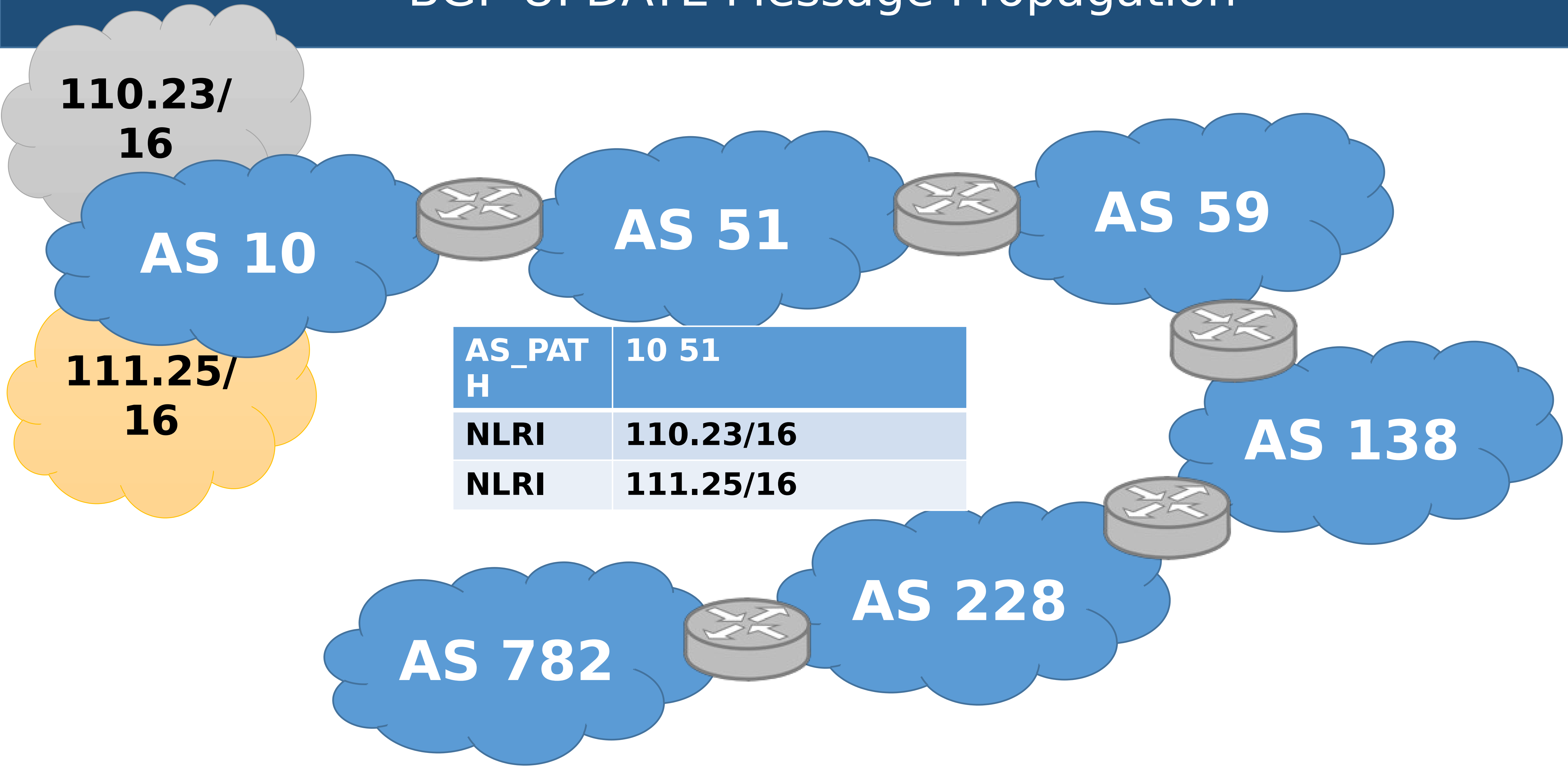
Network Layer Reachability Information (NLRI) (*variable-length*)

BGP UPDATE Message Propagation



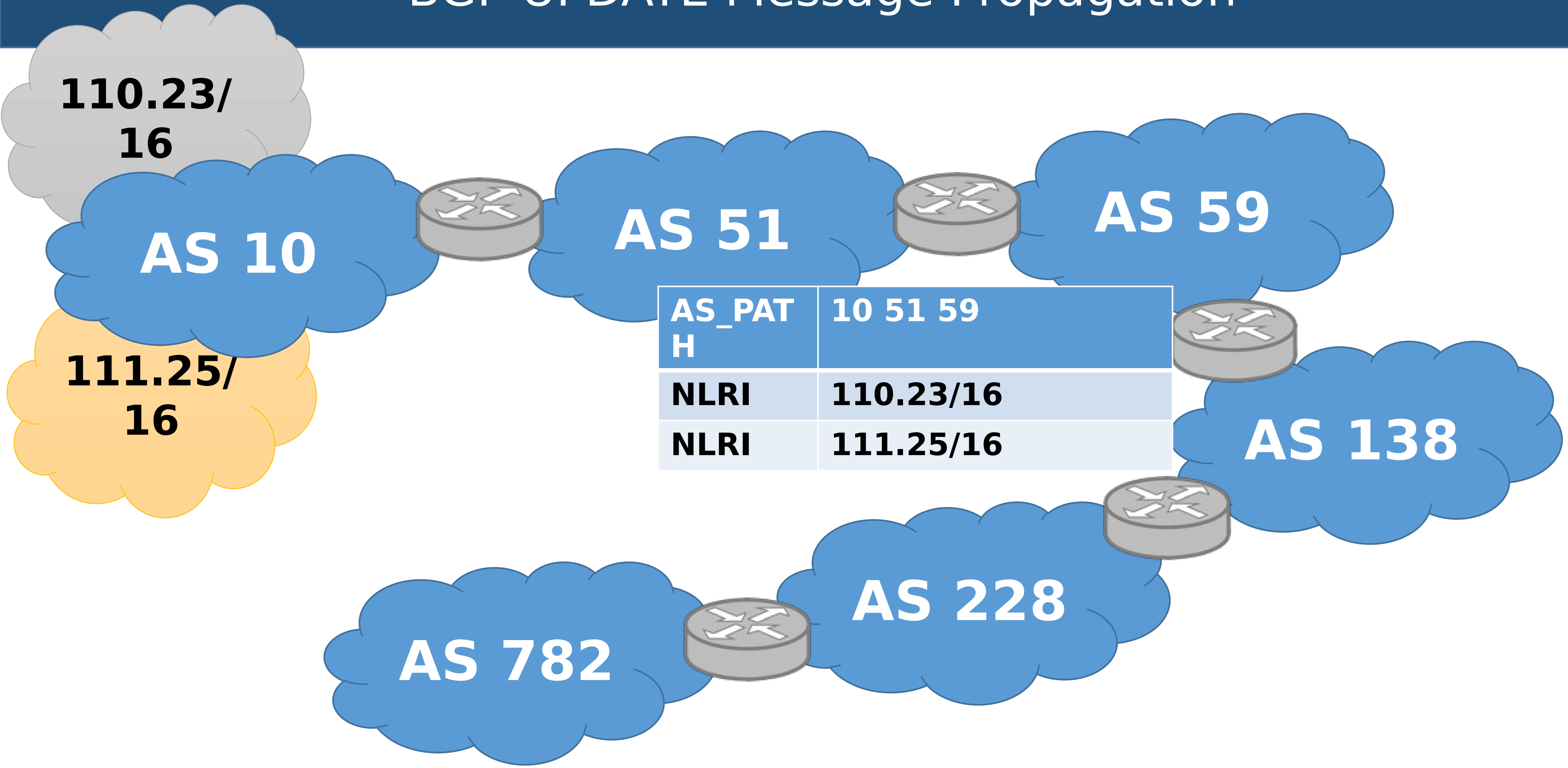
AS_PATH	10
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation



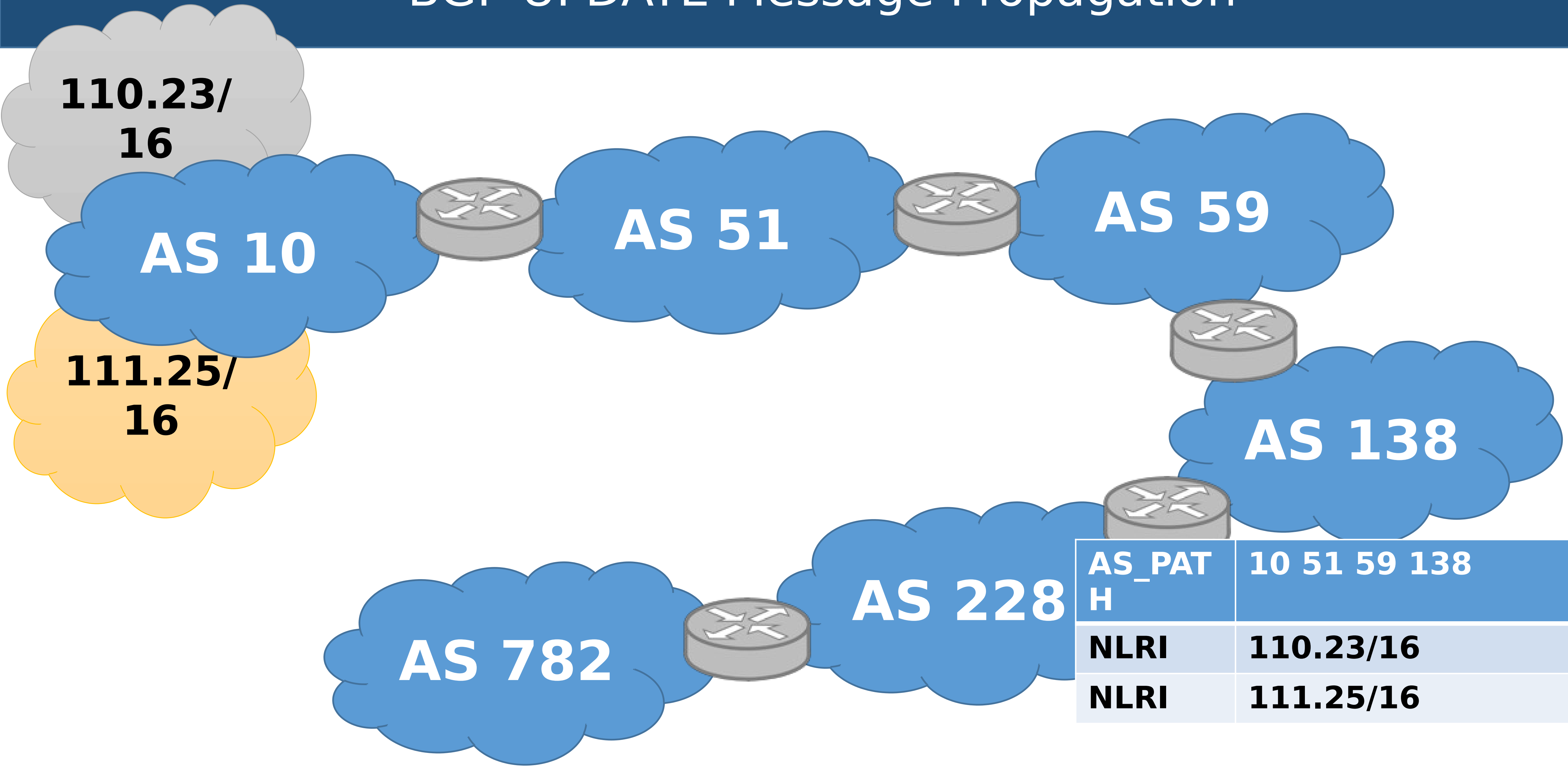
AS_PATH	10 51
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation



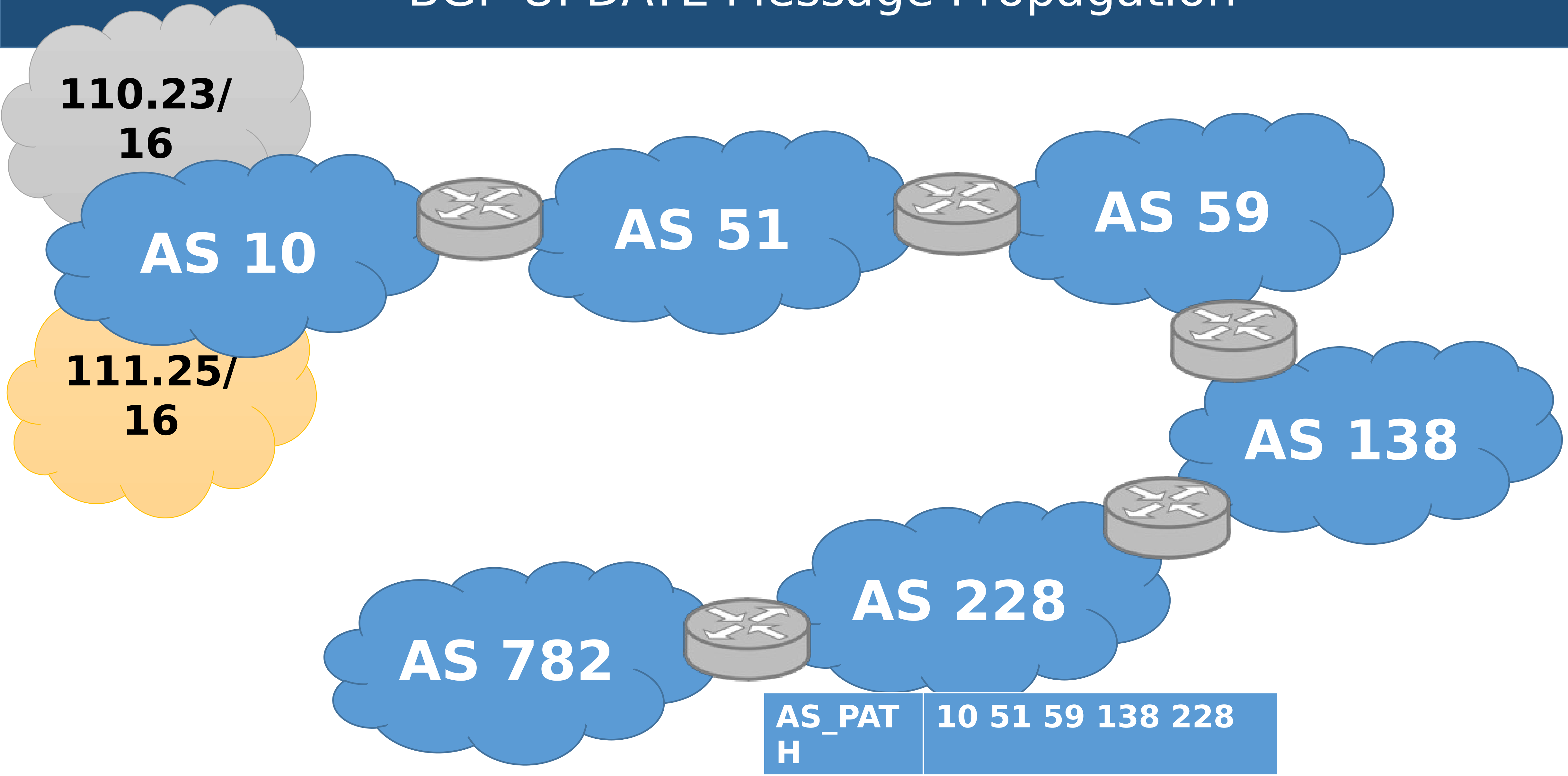
AS_PATH	10 51 59
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation



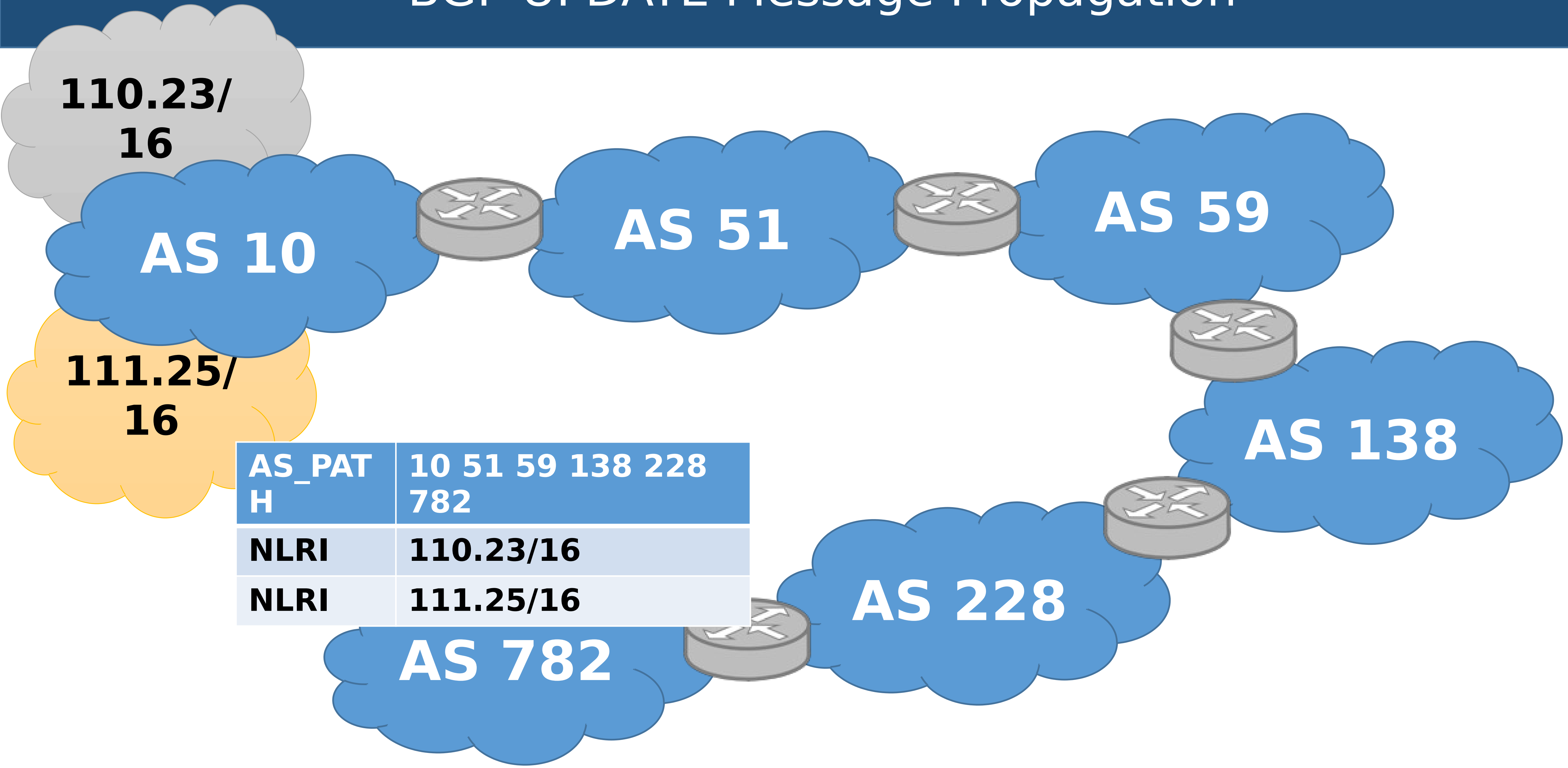
AS_PATH	10 51 59 138
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation



AS_PATH	10 51 59 138 228
NLRI	110.23/16
NLRI	111.25/16

BGP UPDATE Message Propagation

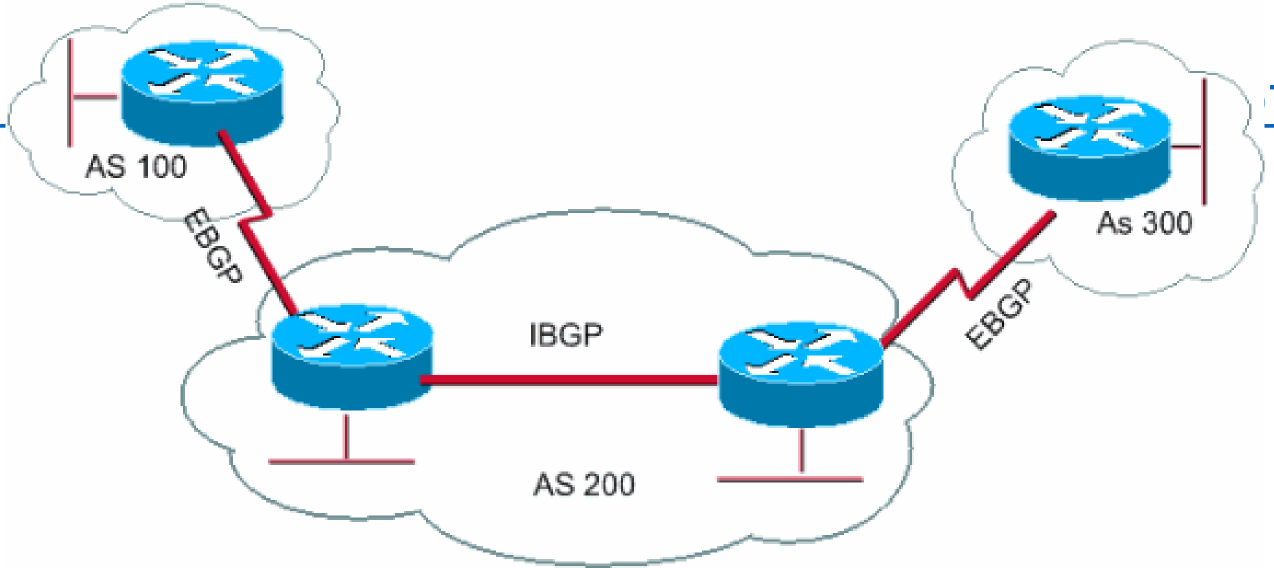


AS_PATH	10 51 59 138 228 782
NLRI	110.23/16
NLRI	111.25/16

BGP Route Establishment

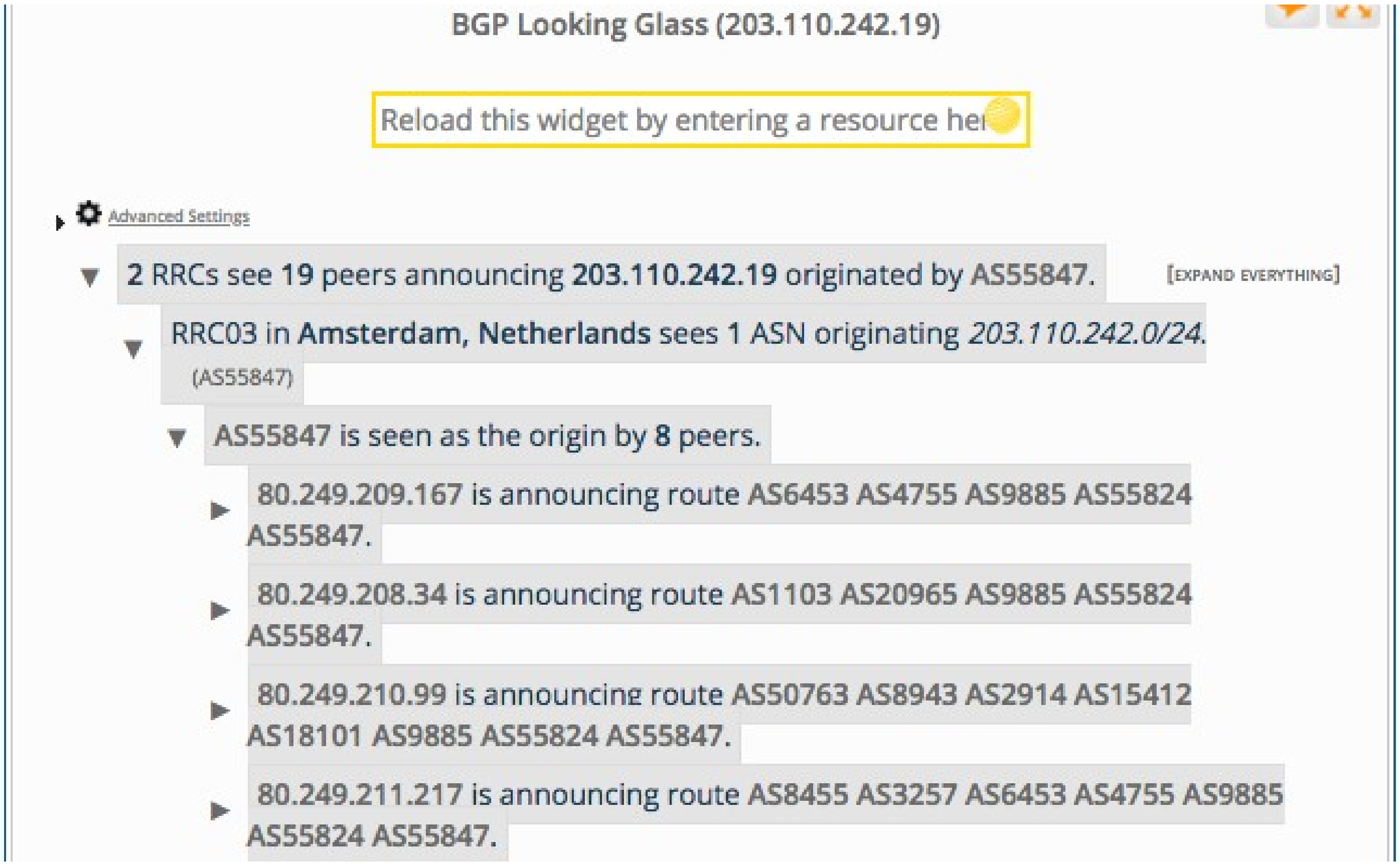
- Based on the BGP UPDATE message, a BGP speaker may have multiple paths to a subnet

- The BGP speaker chooses the best one according to its own local policies -> It advertises this route in the next BGP UPDATE message
  - Check the set of rules that are followed for BGP path establishment algorithm: https://www.cisco.com/c/en toc ol-bgp/13753-25.html



age courtesy: **https://www.cisco.com**

BGP Looking Glass



BGP Looking Glass (203.110.242.19)

▼ 2 RRCs see **19** peers announcing **203.110.242.19** originated by **AS55847.**

▼ RRC03 in **Amsterdam, Netherlands** sees 1 ASN originating *203.110.242.0/24.* (AS55847)

▼ AS55847 is seen as the origin by **8** peers.

▶ 80.249.209.167 is announcing route AS6453 AS4755 AS9885 AS55824 AS55847.

▶ 80.249.208.34 is announcing route AS1103 AS20965 AS9885 AS55824 AS55847.

▶ 80.249.210.99 is announcing route AS50763 AS8943 AS2914 AS15412 AS18101 AS9885 AS55824 AS55847.

▶ 80.249.211.217 is announcing route AS8455 AS3257 AS6453 AS4755 AS9885 AS55824 AS55847.

https://stat.ripe.net/widget/looking-glass