BGP and Failure Recovery

SCC. 203 – Computer Networks

Week 20 – Monday 18th March 2024

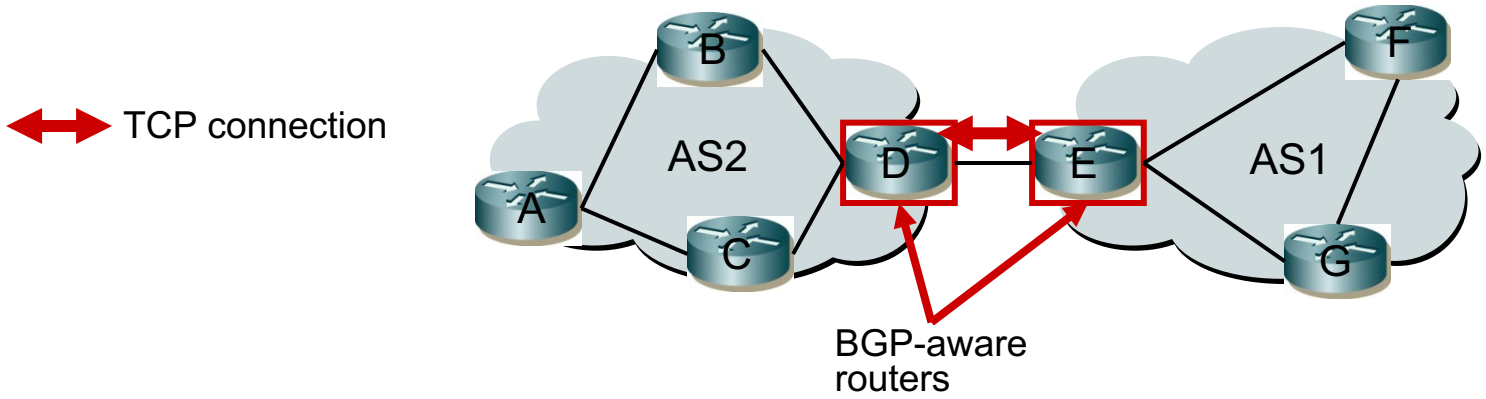
Onur Ascigil

BGP Recap

- *Intra-AS routing protocols such as OSPF (link-state) and RIP (distance vector) run between a set of routers all managed by the same Autonomous System (AS)*
- *On the other hand, BGP runs between different ASes.*
- *BGP extends the basic distance vector approach, to allow policy-driven routing*
  - *Policies are largely driven by economics, (i.e. business relationships between ASes)*
- *Consequently:*
  - *One AS may not wish to advertise reachability of certain routes to some of its neighbours, because the neighbour might then send traffic, increasing the AS's network load*
  - *An AS may wish to preferentially route its traffic via one neighbour than via another*
- *BGP allows routers to exchange information about the reachability of **IP destination address prefixes***
- *Border Gateway Protocol (BGP) is a **Path Vector protocol***

Basic BGP Message Exchange

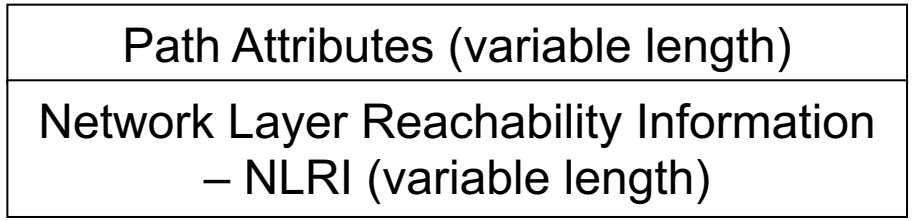
- *Two BGP routers establish a TCP connection between themselves*
  - *A BGP session is therefore between two (and only two) peers*



- *4 BGP message types*
  - ***OPEN**: initial message sent at the start of a BGP session. It allows each BGP peer to identify itself and agree optional parameters*
  - ***UPDATE: advertises paths to destinations, and associates attributes with these paths. This is the principal BGP message we are concerned with, see following slides***
  - ***NOTIFICATION**: error reporting*
  - ***KEEPALIVE**: allows BGP peers to confirm they are still running; based on a hold time interval (~some tens of seconds)*

BGP UPDATE Message Structure

- *BGP UPDATE message advertises routes to destinations and associates attributes with these routes. Its structure is:*
  - ***Path attributes:*** *a set of values associated with each IP destination prefix in the NLRI field*
  - ***Network Layer Reachability Information (NLRI):*** *a set of IP destination prefixes*



- *IP destination prefix format: network prefix, and the number of bits in the prefix (i.e. the address mask).*
  - *Example: 128.234.208.0/20 (CIDR) see Week 16 Lecture slides*

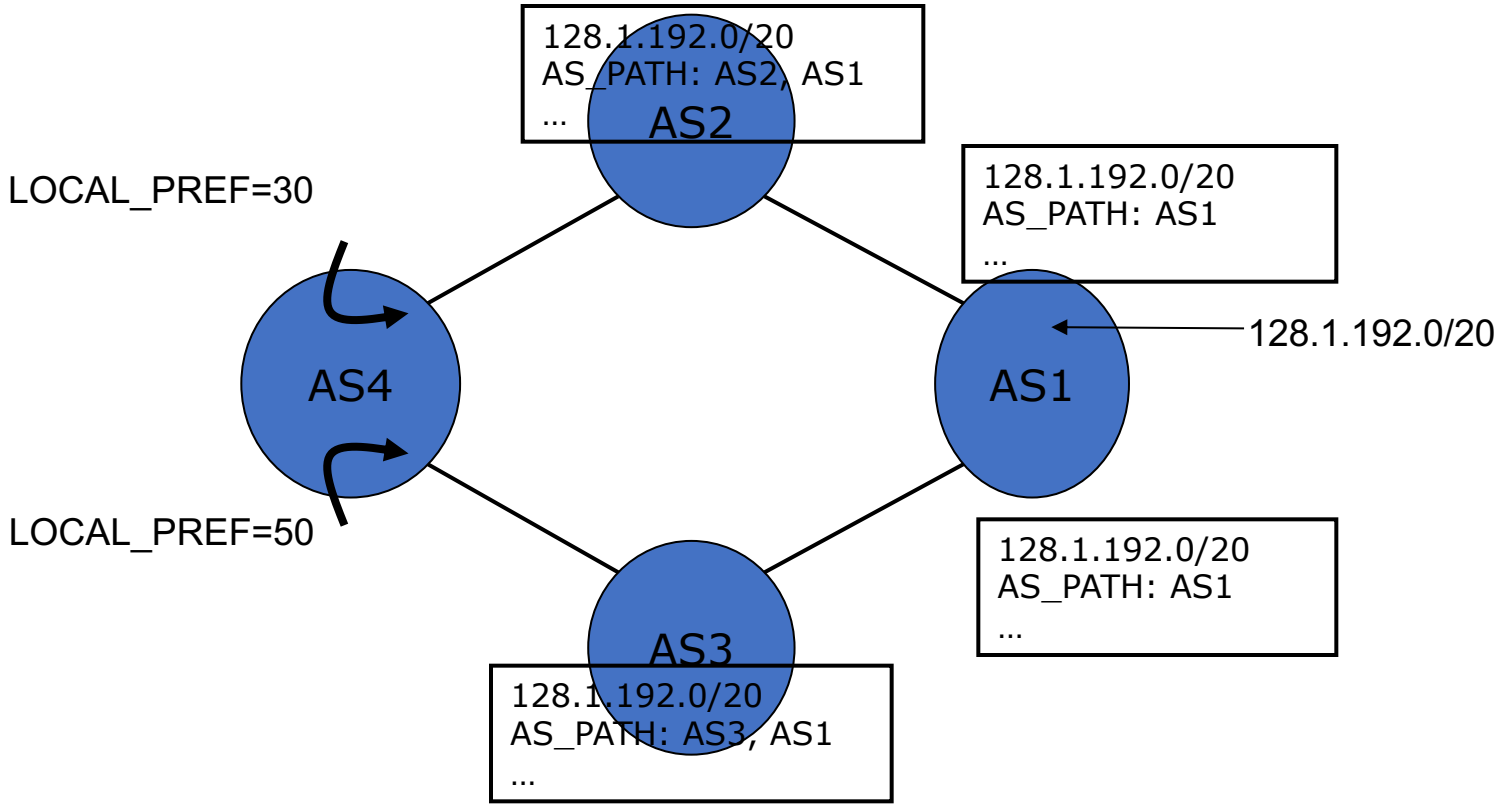
BGP Attributes

- *Attributes define information about the path to each destination prefix. They help BGP to select the best paths to IP destination prefixes while remaining scalable for the large number of ASs that constitute the Internet*

- *The principal attributes that we consider are the following:*
  - *Local preference, LOCAL_PREF*
  - *Autonomous System Path, AS_PATH*
  - *Multi-exit discriminator, MED*
  - *Next hop, NEXT_HOP*

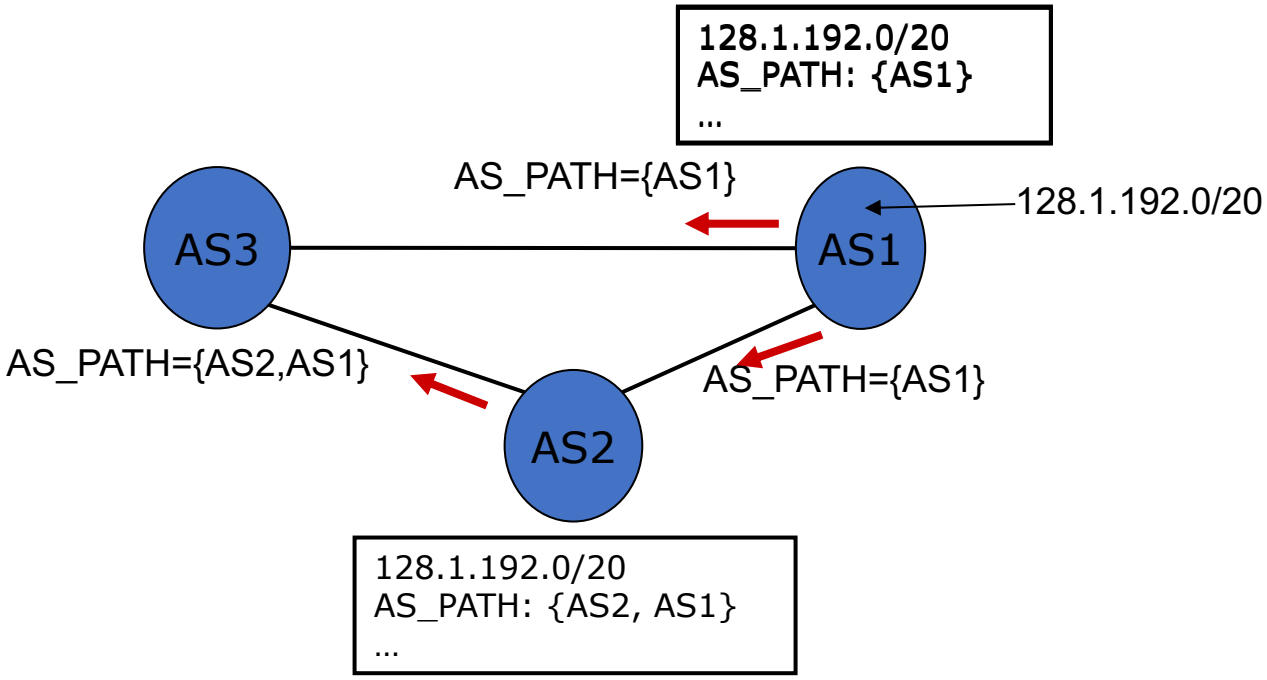
BGP Attributes: LOCAL_PREF

- *If there are multiple paths to a given destination prefix, the LOCAL_PREF attribute allows BGP to specify a preference within an AS for one route over the other(s)*
- *Only used locally within an AS*
- *The higher the LOCAL_PREF value, the more preferred a route is*



BGP Attributes: AS_PATH

- *Each AS in the Internet is assigned its own unique AS number*
- *The BGP AS_PATH attribute contains a list of all the AS numbers of the ASes through which the prefix announcement has passed*
- *The number of entries in the AS_PATH attribute is therefore effectively a measure of the "hop count" to reach the prefix(es) (where one hop = one domain / AS)*

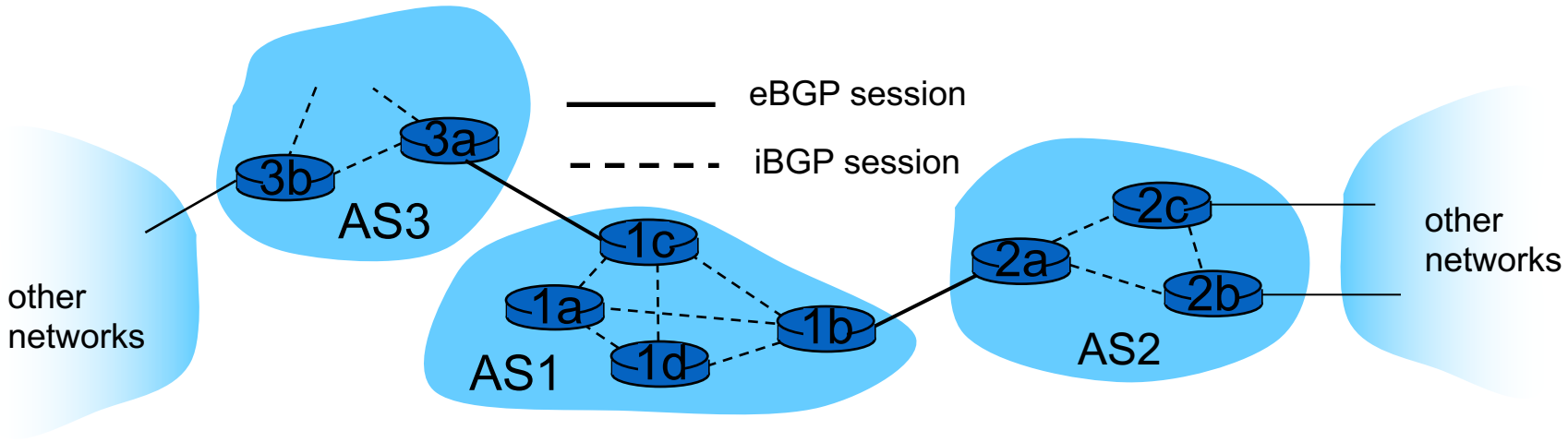


BGP NEXT_HOP: iBGP and eBGP

Using **eBGP** session between 3a and 1c, AS3 sends prefix reachability info to AS1.

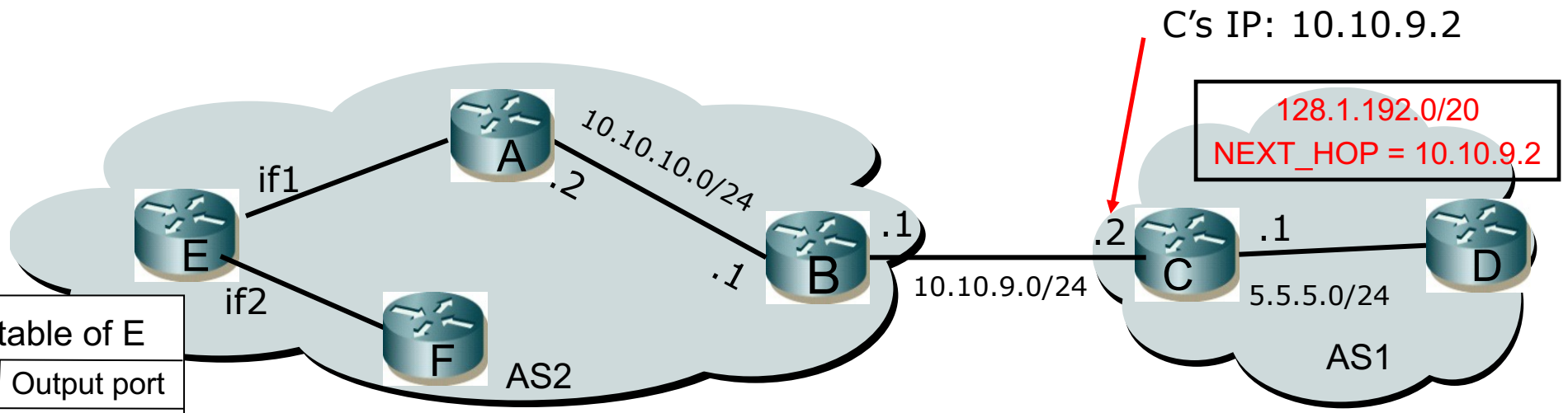
- 1c can then use **iBGP** do distribute new prefix information to all routers in AS1
- 1b can then re-advertise new reachability info to AS2 over 1b-to-2a **eBGP** session

When a router learns of a new prefix, it creates entry for prefix in its forwarding table.



BGP Attributes: NEXT_HOP

- *The NEXT_HOP attribute specifies the IP address of the router in the adjacent domain that is the next hop to the destination prefixes listed in the UPDATE message NLRI field*
- *The NEXT_HOP address is usually the same as the IP address of the of the BGP router that is sending the message that contains the NEXT_HOP attribute*
  - *e.g. C advertises reachability of 128.1.192.0/20 to B, specifying as NEXT_HOP the IP address of C*

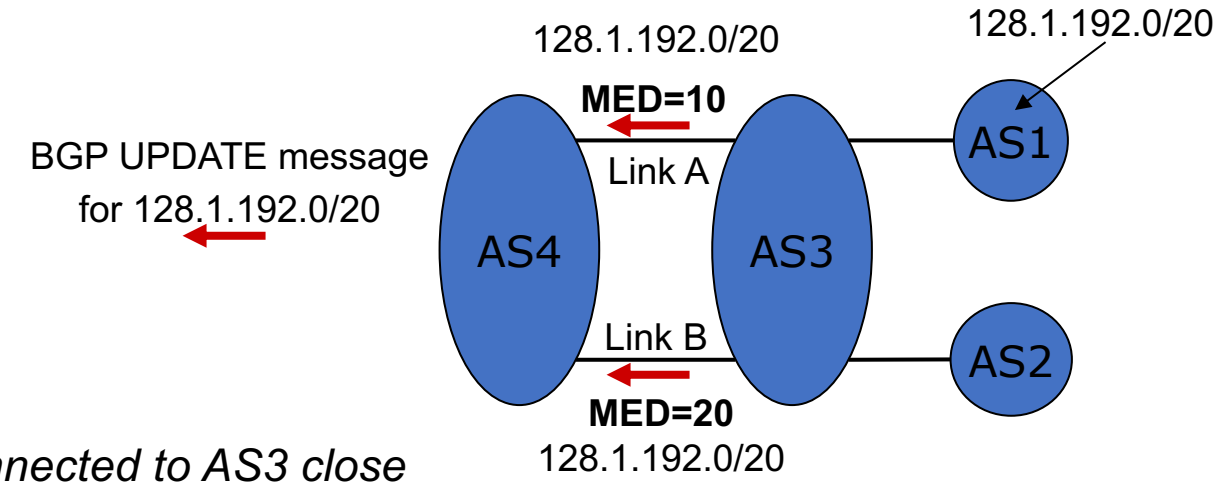


Forwarding table of E	
IP prefix	Output port
128.1.192.0/20	if1

BGP Attributes: MULTI_EXIT_DISCRIMINATOR (MED)

- *MED is used when two ASs have two or more direct connections. It allows one AS to express to the other AS a relative preference for each of the links: one AS sets the MED value, the other AS uses this information*
- *The lower the MED value, the more preferred a path is*

Note MED is a preference, not a rule, as we shall see in a later slide



- *Example:*
  - *AS3 and AS4 are connected by two links, A and B*
  - *BGP sessions run over each link*
  - *AS1 is connected to AS3 close to link A; AS2 is connected to AS3 close to link B*
  - *AS3 sets MED values so as to prefer link A for prefixes in AS1, and to prefer link B for prefixes in AS2*
- AS3 here could have only sent one advertisement over its preferred link, why does it advertise the destination over multiple links?

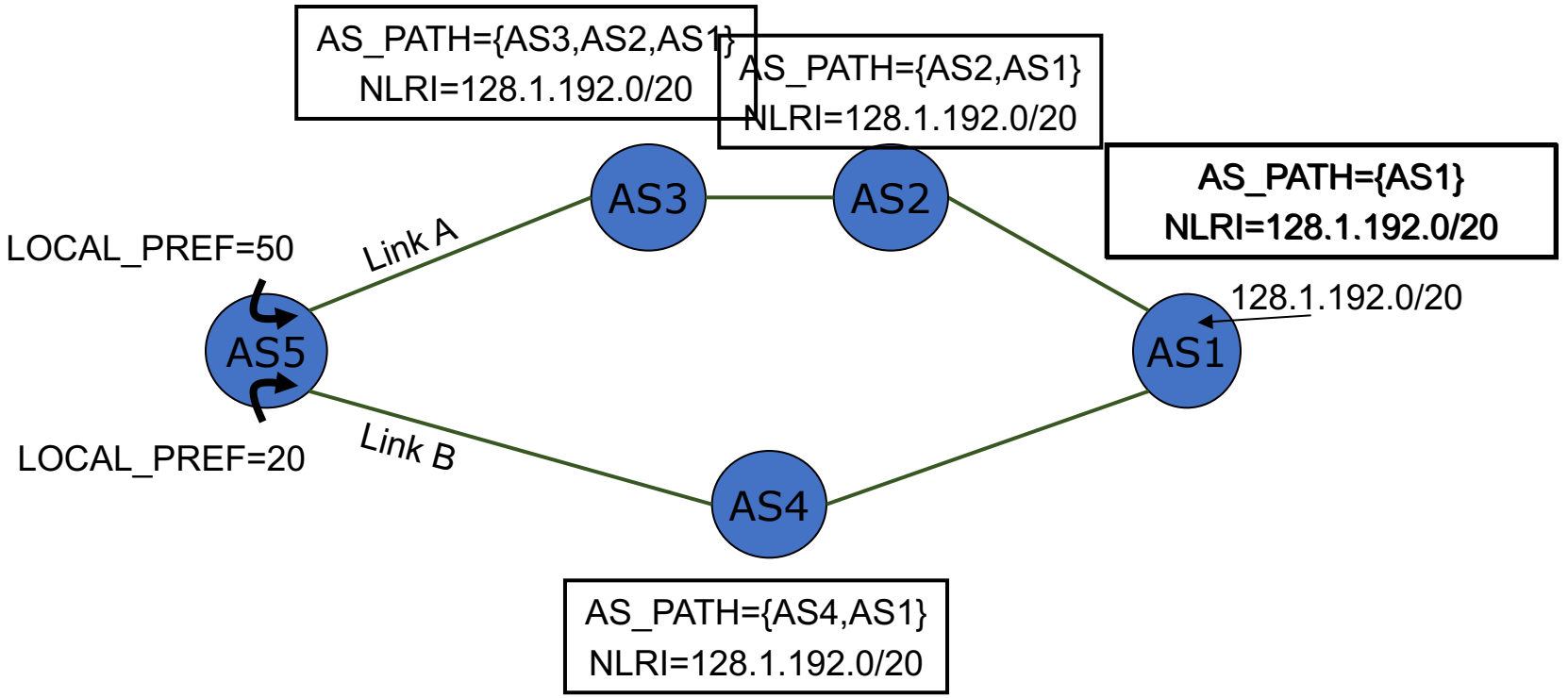
BGP Route Selection Process

- *If BGP learns about more than one route for a given address prefix, the following rules are applied:*
  - *Select the route with the highest LOCAL_PREF value. If still a tie, then*
  - *Select the route with the shortest AS_PATH (=hop count, in distance vector terms). If still a tie, then*
  - *Select the route with the lowest MULTI_EXIT_DISCRIM, if multiple routes were learned from the same AS. If still a tie, then*
  - *Select the route with the minimum cost to the NEXT_HOP. If still a tie, then*
  - *Select the route learned via eBGP (if only one), or the route learned via eBGP with the lowest BGP identifier. If still a tie, then*
  - *Select the route learned from the iBGP neighbour with the lowest BGP identifier (this is usually one of the router's IP addresses)*

Rules vary slightly depending on source. This set is based on "BGP4", J.W. Stewart

Example: LOCAL_PREF and AS_PATH

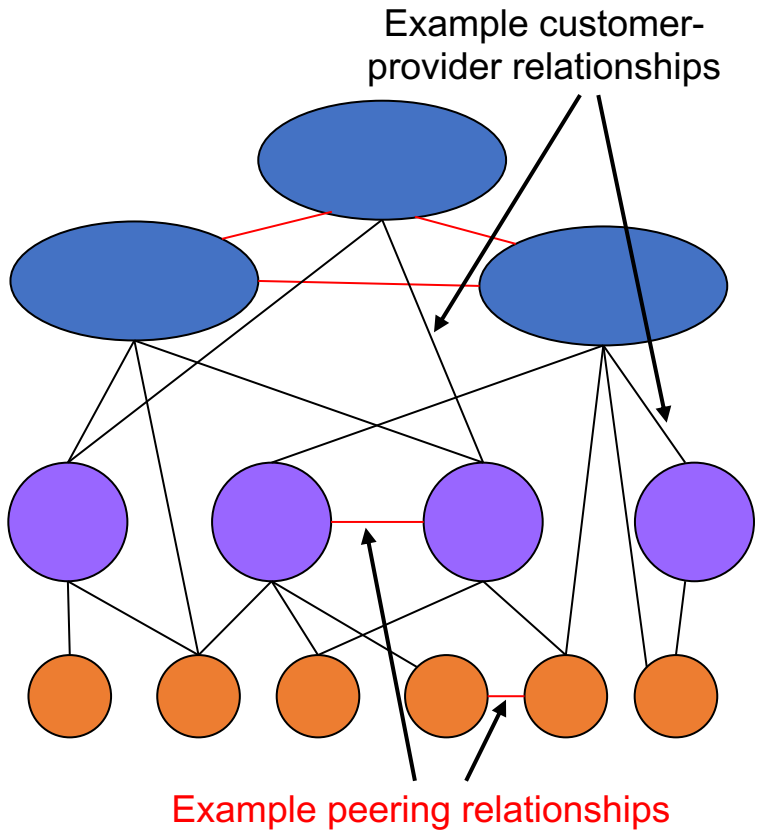
- *AS5 selects link A as the path to forward packets destined for 128.1.192.0/20 because of the LOCAL_PREF settings, even though this is not the shortest path (measured by the size of AS_PATH)*



Inter-domain Structure: Business Relationships

The various functions of the domains allow us to define a set of relationships:

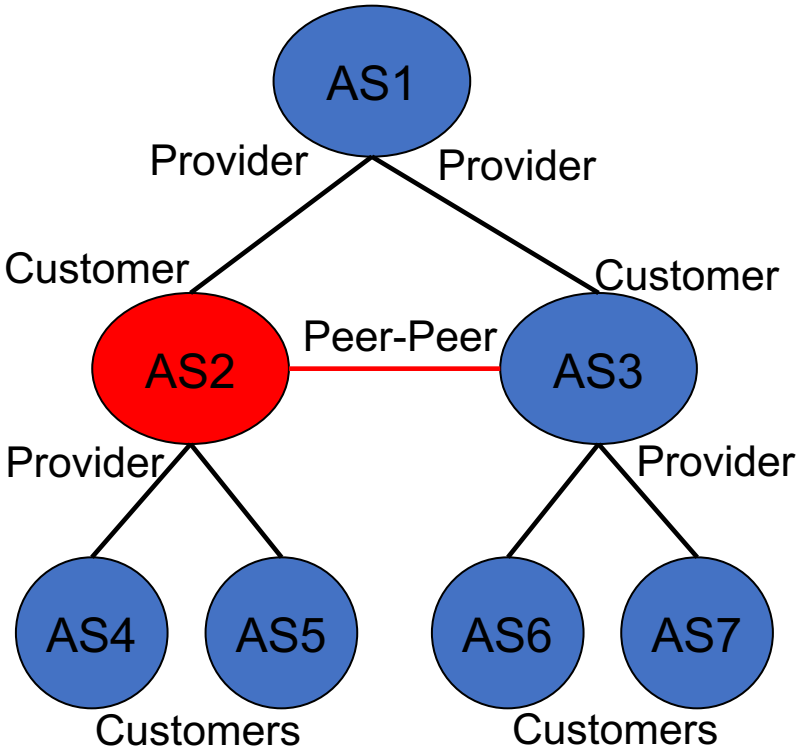
- ***Customer-provider relationship:***
  - *The customer pays the provider for access to the rest of the Internet. In the same way that a residential customer pays a monthly fee to their ISP, so a (smaller) domain pays a (larger) domain for access*
- ***Peer-Peer relationship:***
  - *Two domains (typically of similar size) agree to exchange traffic between their respective customers. Traffic flow volumes in either direction of a peer-to-peer relationship are usually similar*
- *These business relationships between domains impact the routing information exchanged between the domains*



BGP Export Policies: Achieving Policies via Advertisements

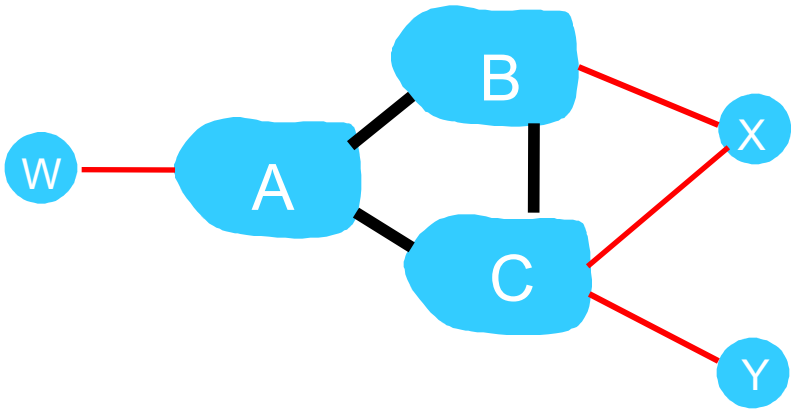
Example: AS2's BGP policies

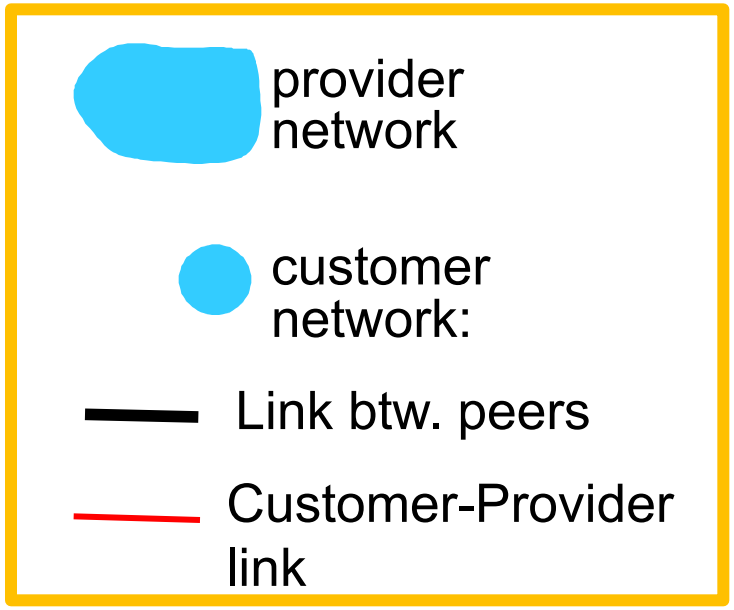
- *AS2 receives prefix advertisements of direct neighbours:*
  - *Prefixes from AS1 (AS_Path: AS1): advertise to 4 and 5*
  - *Prefixes from AS3 (AS_Path: AS3): advertise to 4 and 5*
  - *Prefixes from AS4 (AS_Path: AS4): advertise to 1, 3, and 5*
  - *Prefixes from AS5 (AS_Path: AS5): advertise to 1,3, and 4*
  - *Prefixes from AS2: advertise to everyone*
- *AS2 receives prefix advertisements via neighbours:*
  - *Prefixes from AS6 (AS_Path: AS3-AS6): advertise to 4 and 5*
  - *Prefixes from AS7 (AS_Path: AS3-AS7): advertise to 4 and 5*
  - *Prefixes from AS6 (AS_Path: AS1-AS3-AS6): advertise to 4 and 5*
  - *Prefixes from AS7 (AS_Path: AS1-AS3-AS7): advertise to 4 and 5*
  - *Prefixes from AS3 (AS_Path: AS1-AS3): advertise to 4 and 5*
  - *Prefixes from AS6 (AS_Path: AS1-AS3-AS6): advertise to 4 and 5*
  - *Prefixes from AS7 (AS_Path: AS1-AS3-AS7): advertise to 4 and 5*



- *AS2 does not advertise to AS3 the fact that it can reach prefixes in AS1, since it does not wish to carry (and pay for!) transit traffic for AS3*
- *Similarly, AS2 does not advertise to AS1 the fact that it can reach the subnet(s) in AS3*

BGP Export Policies: Achieving Policies via Advertisements





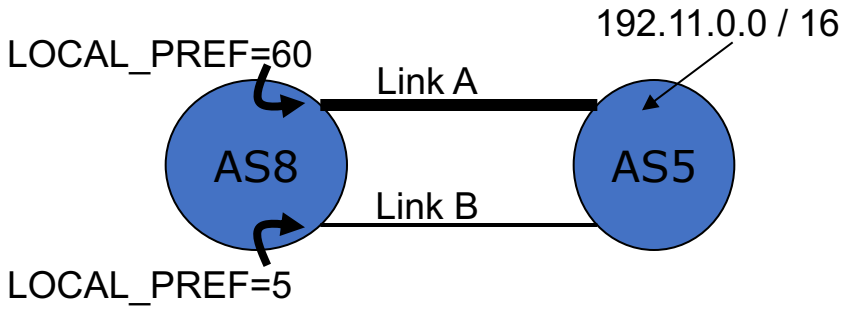
- C advertises path Cy to x.
- 1. Should x advertise xCy path to B?

- A advertises path Aw to B
- B advertises path BAw to x
- 2. Should B advertise path BAw to C?

BGP: Controlling Outbound Traffic

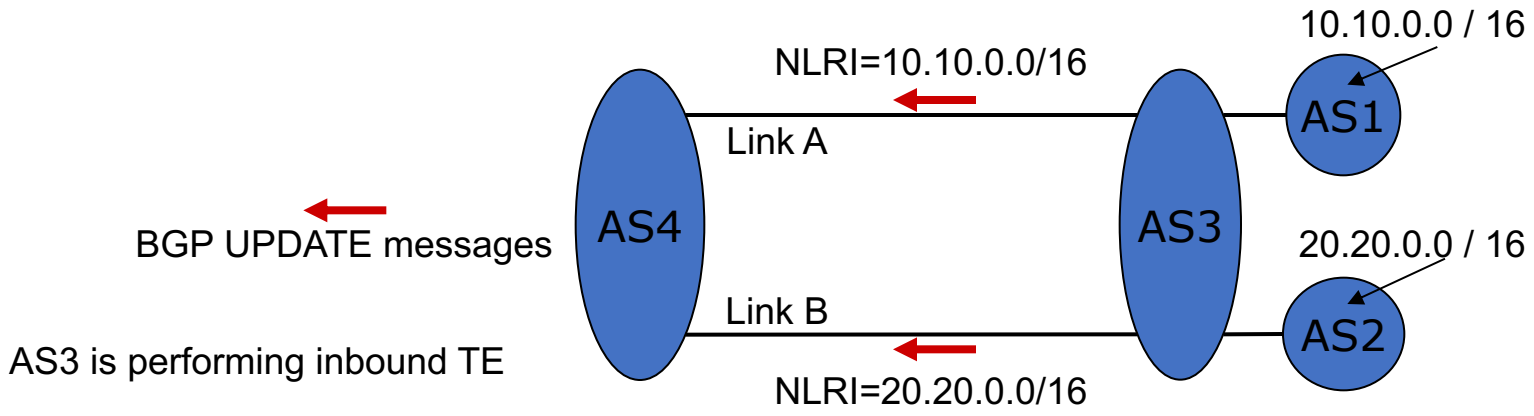
- *Definition: TE that optimises the flow of traffic leaving a domain*
- *Approach: configure the LOCAL_PREF attribute*
- *Example: 2 links between AS5 and AS8; link A is high bandwidth, link B is low bandwidth*
  - *Assign a higher LOCAL_PREF (=60) to the high bandwidth link and a lower LOCAL_PREF (=5) to the low bandwidth link; all traffic will then be routed over the high bandwidth link (Link A)*
  - *The low bandwidth link still exists as a backup path in the event of failure of the high bandwidth link*

AS 8 is using LOCAL_PREF to control outbound traffic



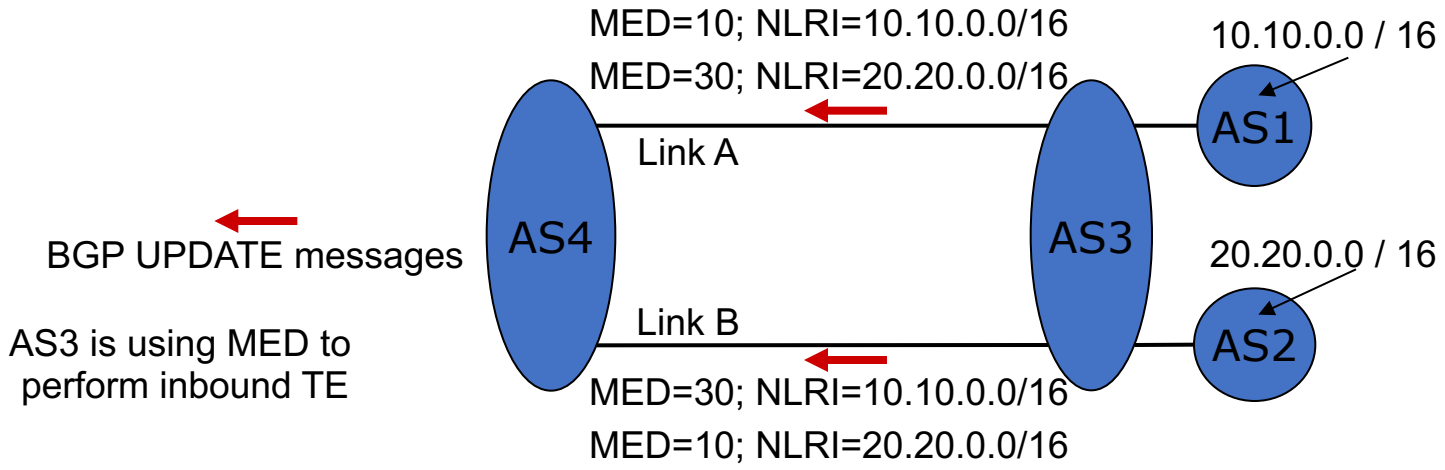
BGP: Controlling Inbound Traffic(1)

- *Approach 1: announce different advertisements on different links*
  - *Example: AS3 announces reachability of all AS1 addresses on Link A only, and reachability of all AS2 addresses on Link B only*
  - *AS4 therefore sends 10.10.0.0/16 traffic on Link A and 20.20.0.0/16 traffic on Link B*
  - *Disadvantage: if one link fails the corresponding destination becomes unreachable*



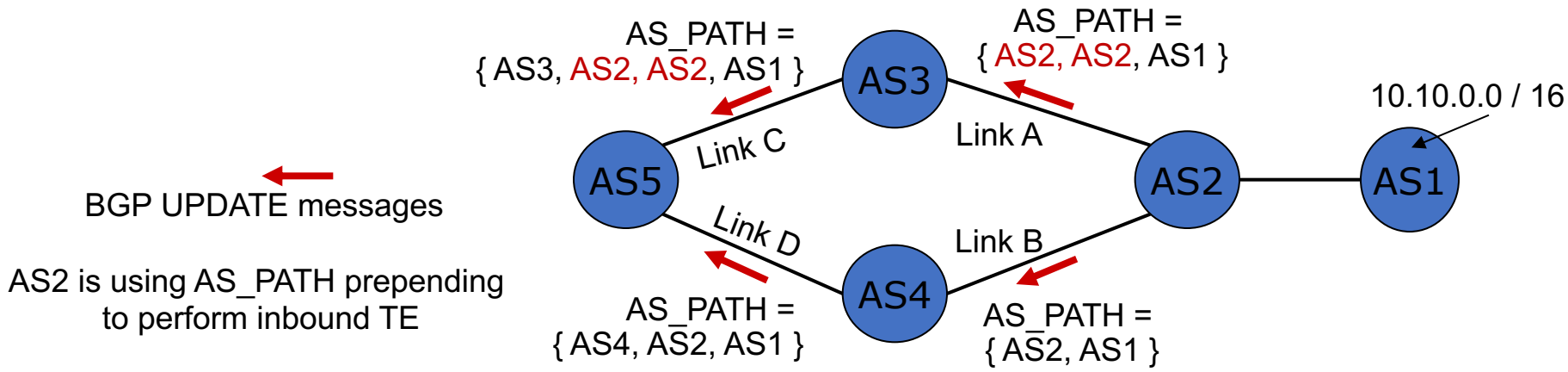
BGP: Controlling Inbound Traffic(2)

- *Approach 2: configure MULTI_EXIT_DISCRIM (MED) attribute*
  - *MED only works when two ASs have 2 or more direct links and they agree to implement MED*
  - *The lower the MED value the more preferred a path is; so AS3 will receive from AS 4 traffic destined to 10.10.0.0/16 on Link A and traffic destined to 20.20.0.0/16 on Link B*
  - *Advantage: backup path availability: Link A is a backup for 20.20.0.0/16 and Link B is a backup for 10.10.0.0/16*



BGP: Controlling Inbound Traffic(3)

- *Approach 3: artificially extend the AS Path length (AS_PATH prepending)*
  - *Example: AS2 announces reachability of AS1 on its inter-domain links to AS3 and AS4, but uses AS_PATH prepending on the link to AS3 so as to discourage traffic destined to AS1 from using this link (it uses twice its own AS to make the path appear longer, which is common)*
  - *Based on the AS_PATH attributes, AS5 will therefore select AS4 when forwarding traffic to 10.10.0.0 / 16*



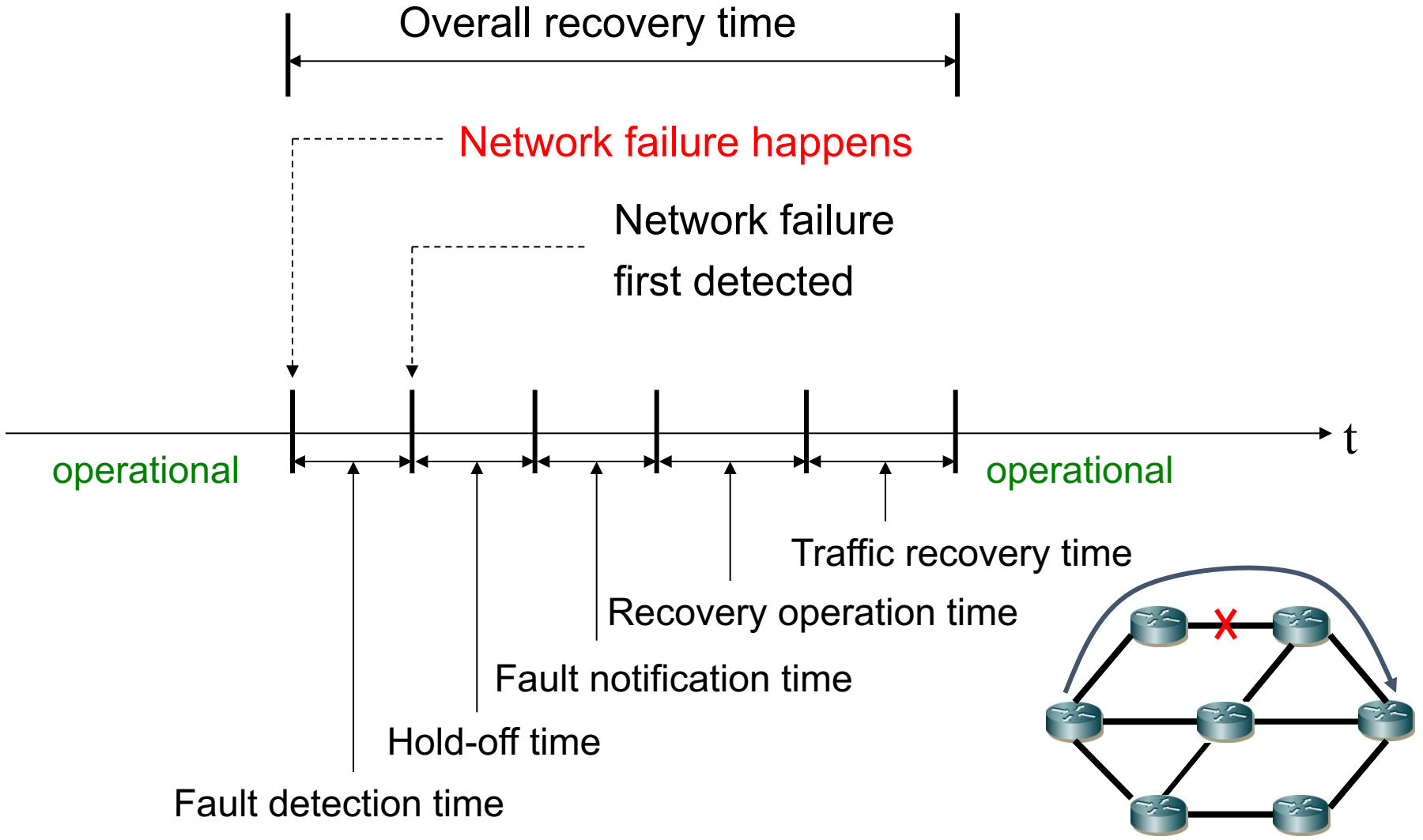
Part 2: Failure Recovery

Statistics of Network Failures

- *Around 20% of network failures are attributed to planned maintenance that can be completely anticipated, in which case "make-before-break" can be fully applicable*
- *Around 70% of unexpected failures are single link failures, i.e. the breakdown of a single interface*
- *Around 80% of link failures are transient ones – for most of the cases the link can be recovered within 10 minutes, and about 50% of them last less than a minute (e.g. due to router rebooting)*
- *More information available at:*

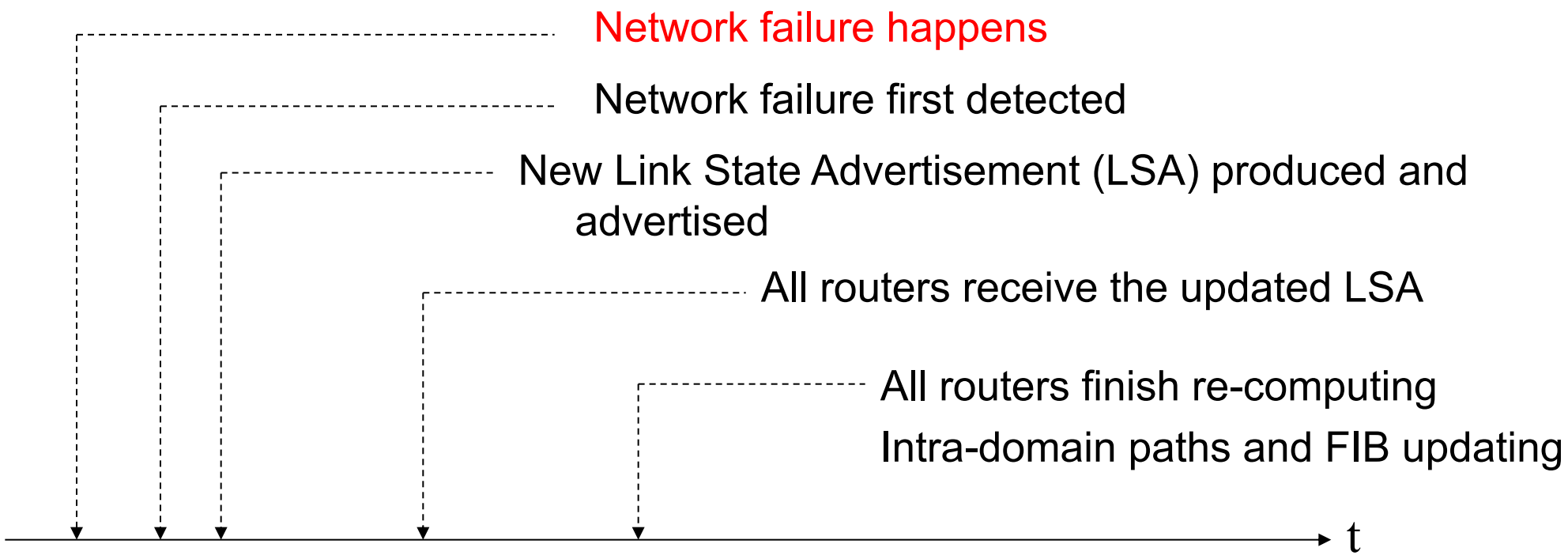
  - G. Iannaccone et al, "Analysis of Link Failures in a Large IP Backbone", Proc. ACM Internet Measurement Workshop (IMW), 2012

Network Recovery Cycle



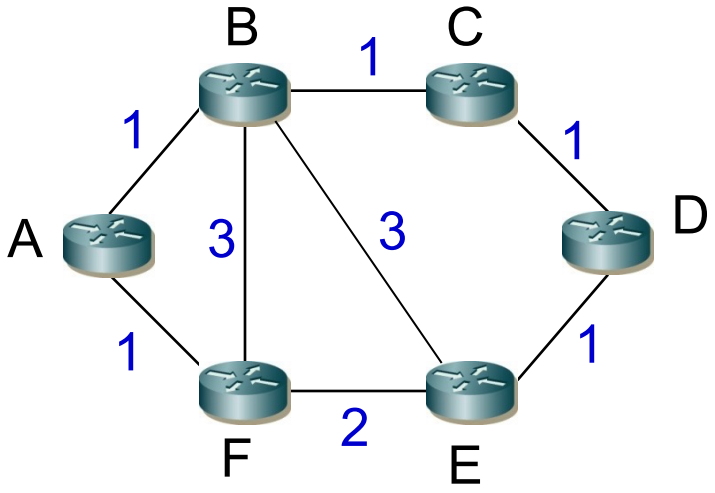
Post-failure Re-convergence

- *The time procedure between the link fails and all the routers within an AS reach consistent views on the new network topology and accordingly finish updating their routing/forwarding tables is known as re-convergence*



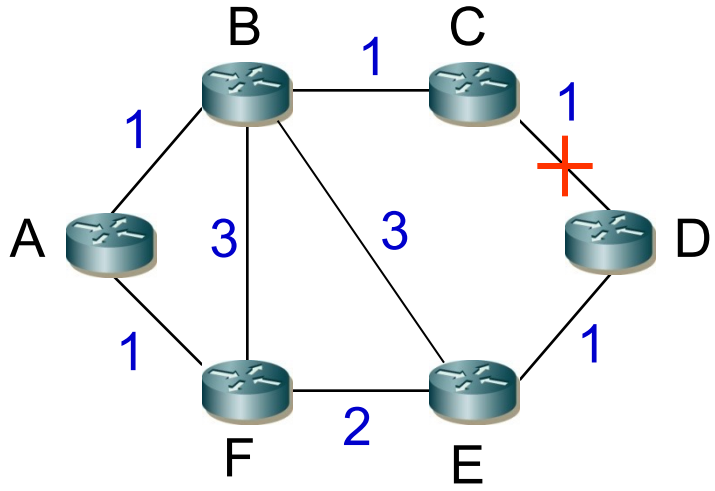
- Routers near the point of failure may have finished re-computing of paths and FIB update even before others receive the new LSA, resulting in inconsistency

Re-convergence Example (1)



(a) Normal state

	Destination	*Next Hop*
A's FIB	*D*	*B*
B's FIB	*D*	*C*
C's FIB	*D*	*D*

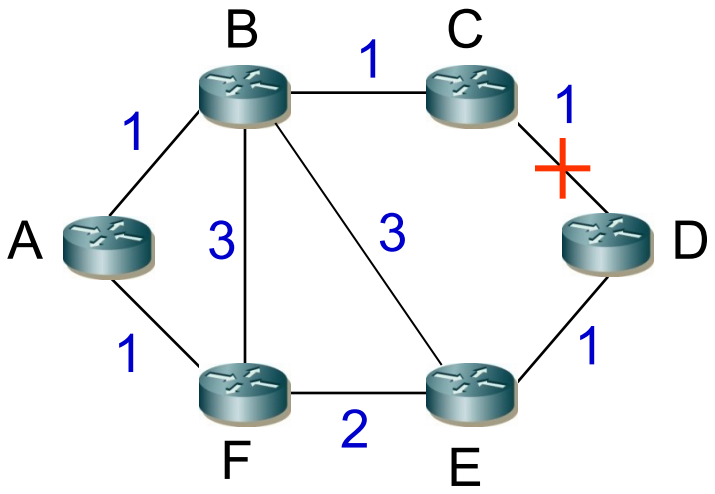


(b) C detects the failure but not yet finishes updating its FIB

	Destination	*Next Hop*
A's FIB	*D*	*B*
B's FIB	*D*	*C*
C's FIB	*D*	*D*

Packets from A/B/C to D start to be dropped at router C!

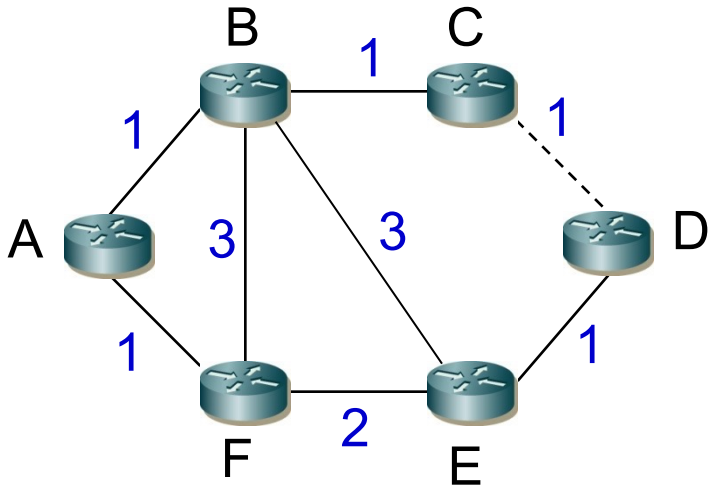
Re-convergence Example (2)



(c) C finishes updating its FIB but others have not done yet

	Destination	Next Hop
A's FIB	*D*	*B*
B's FIB	*D*	*C*
C's FIB	*D*	*B*

Packets from A to D encounter a transient loop between C and B!



(d) Re-convergence finished

	Destination	Next Hop
A's FIB	*D*	*F*
B's FIB	*D*	*E*
C's FIB	*D*	*B*

Fast Reroute (FRR)

- *Basic Idea*

  – The router directly attached to the failing link is responsible for local rerouting. This router is called the repairing router of the failing link

  – Backup solutions need to be pre-configured at the repairing router

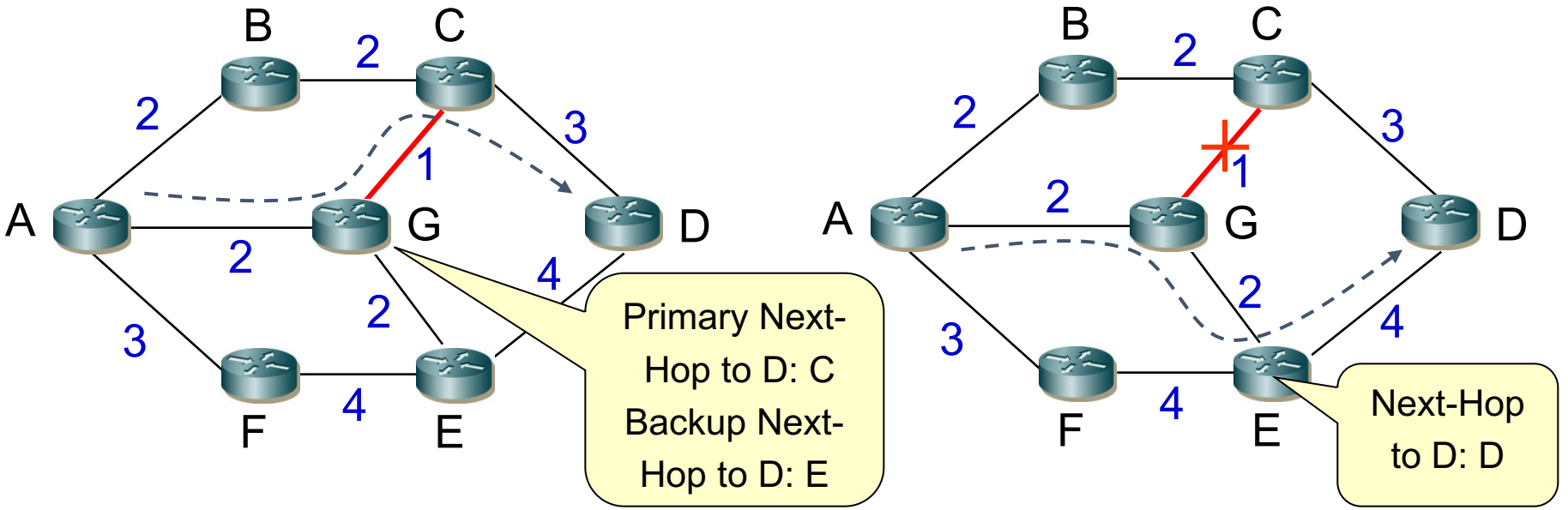
  – Link State Advertisement (LSA) about the failed link is suppressed – none of the remote routers are aware of the failure, and hence they do not need to update their routing/forwarding tables accordingly

*The routers pre-installs a backup **alternative next-hop** for each destination. This alternative next-hop must be a directly attached neighbour to the head node, but is NOT necessarily on the shortest path towards the destination*

– In case the protected link fails, the head node immediately forwards the packets towards the destination on the pre-installed backup alternative next-hop from where it can be natively delivered to the destination

An Example for Success

Consider traffic from A to D…



- Normal State

  - Link G→ C is to be protected. In addition to its primary next hop to D (which is C), the repairing router G also maintains the backup next-hop E.

  - G always uses C as the unique next hop to D in the normal state

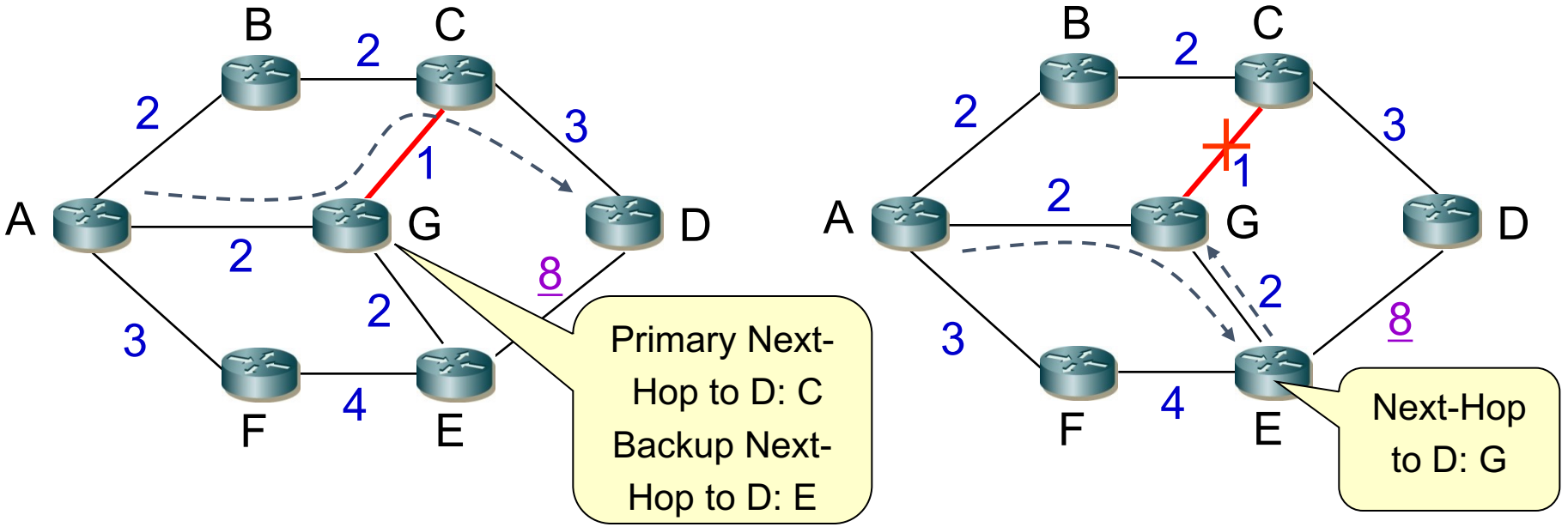
- Link G→ C Fails

  - G immediately forwards the packets to the backup next-hop E. E directly forwards packets to the destination D according to its forwarding table

  - A is NOT aware of the failure so it does not need to change its own next hop from G to B

An Example for Failed Protection

Sometimes you do not get that lucky…

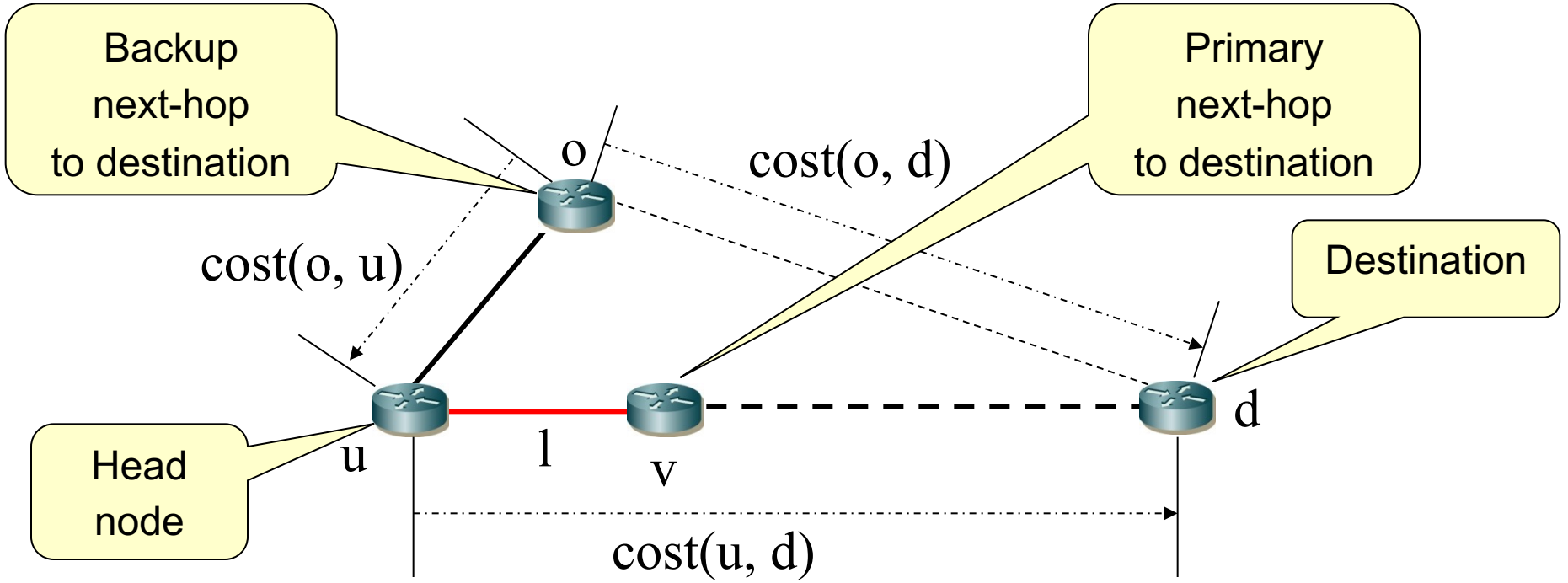


- Link G→ C Fails (Note the link cost between E and D is increased to 8!)

  - When G detects the failure and starts forwarding packets to the backup next-hop E, E will return them back to G, thus forming a loop

  - This is because E's next-hop in the normal state is G, as the shortest path from E to D is E→G→C→D and E is not notified about the failure between G and C

A Necessary Condition

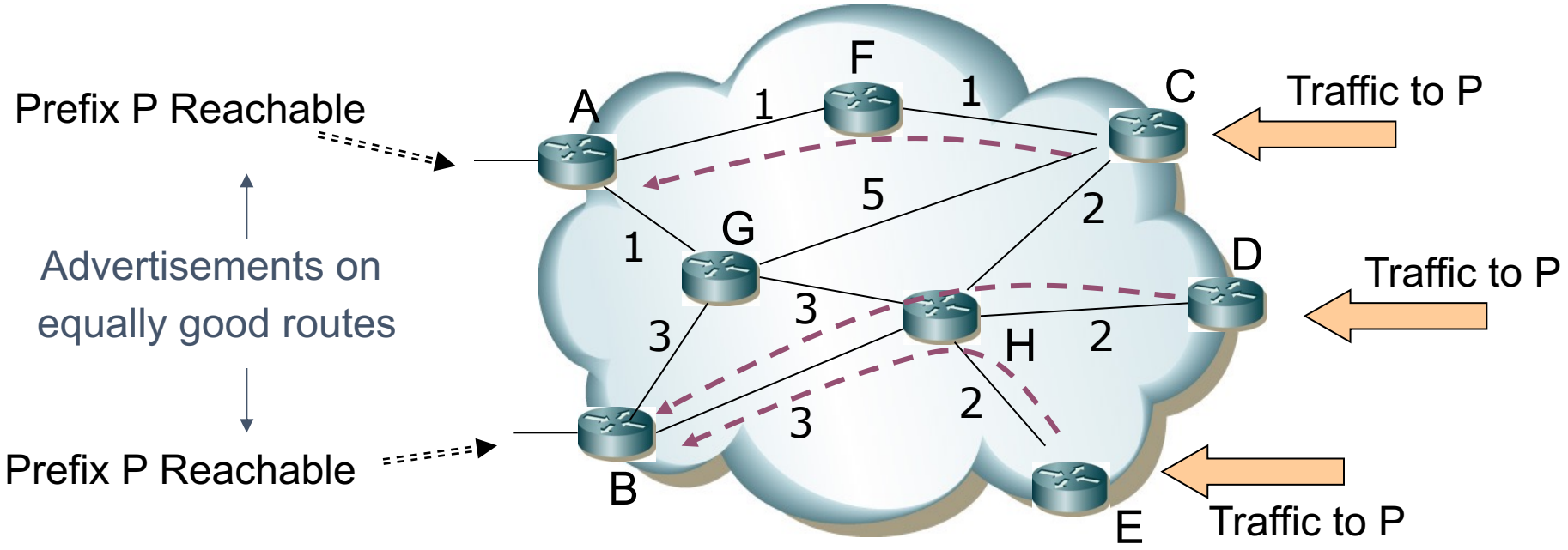


- The ingress node of the protected link l must NOT be on the shortest path from the backup alternative next-hop to the destination, i.e.

$$cost(o, d) < cost(o, u) + cost(u, d)$$

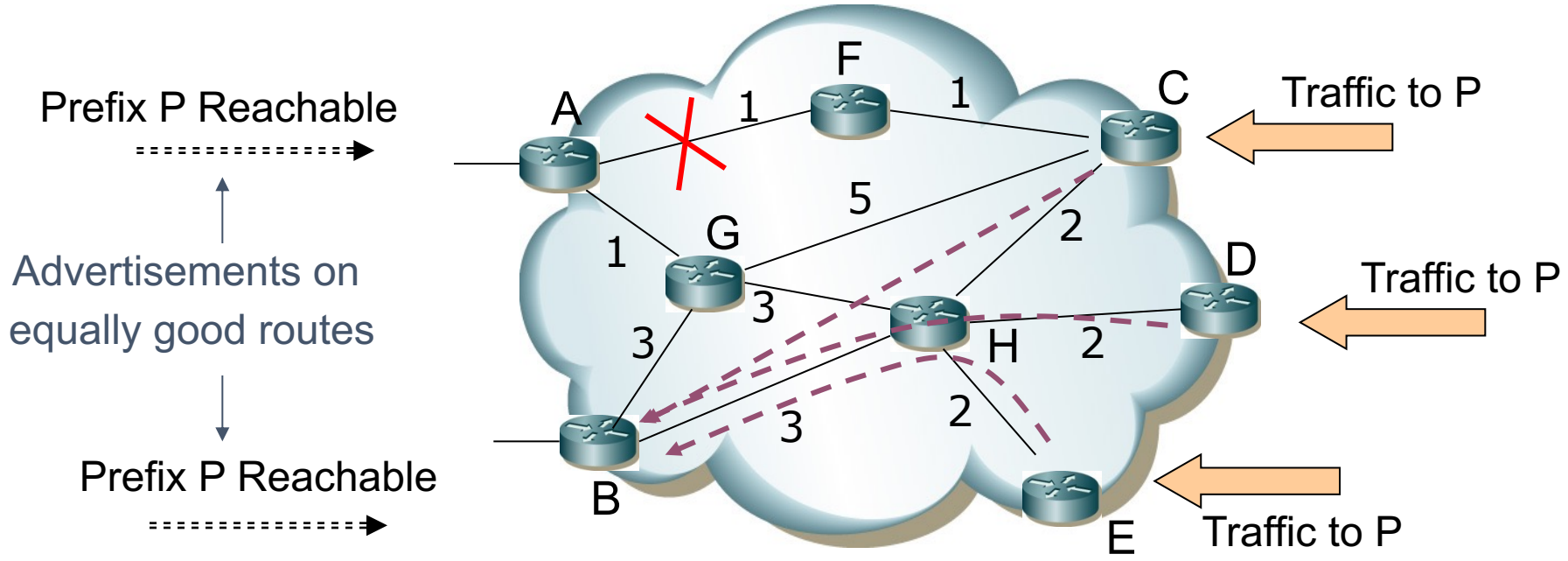
where u is the ingress node of the protected link l, d is the destination and o is a feasible backup next hop for l towards the destination

Hot Potato Routing – A Failure Example



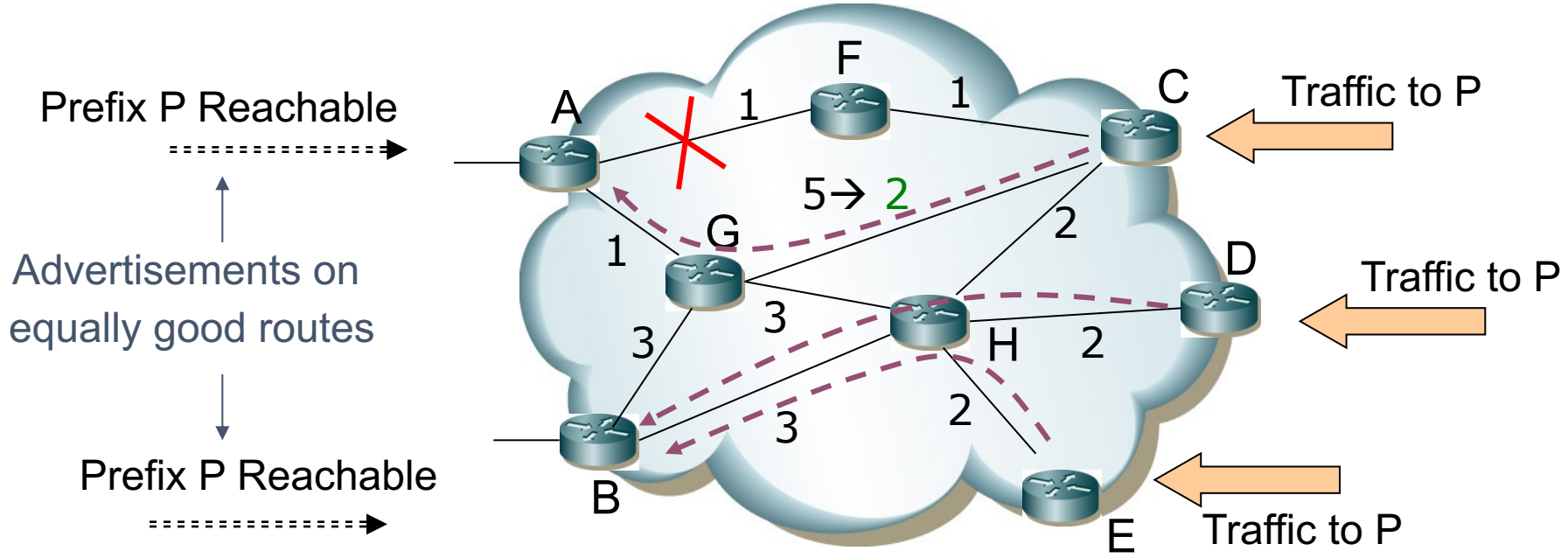
- A and B receive BGP advertisements on equally good routes towards P
- According to hot potato routing, C, D and E may select their own egress points according to the intra-domain path cost to A and B:
  - C selects A as the egress towards P, cost = 2
  - D and E select B as the egress towards P, cost = 5 respectively

Hot Potato Routing – A Failure Example



- In case the intra-domain link between F and A fails, router C will select B as the new egress point, as after re-convergence B is closer to C than A.
- As a result:
  - The original balanced load between inter-domain links attached to A and B will be impacted by the change of egress points at C
  - The downstream domain connected through B will have to carry some unexpected additional traffic due to the change of egress points at C

Hot Potato Routing – A Failure Example



- Strategy: to intelligently adjust link costs in order to avoid egress point switching even after intra-domain link failure occurs
- By changing the link cost between C and G from 5 to 2, router A is still closer to C than B even if the link between A and F fails. So C will not switch its egress point towards P from A to B, thus avoiding all the previous problems caused by the intra-domain link failure

Thanks for listening!

Any questions?