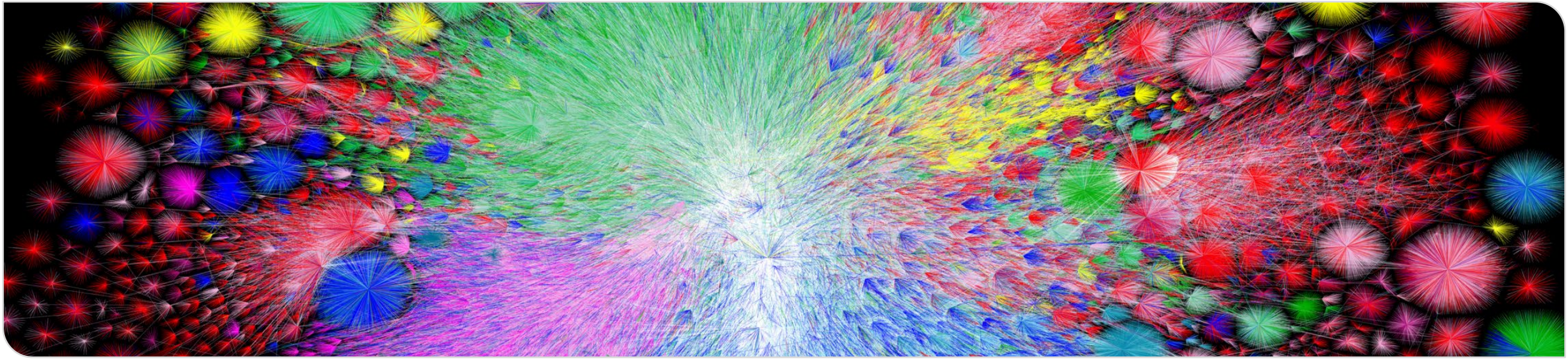
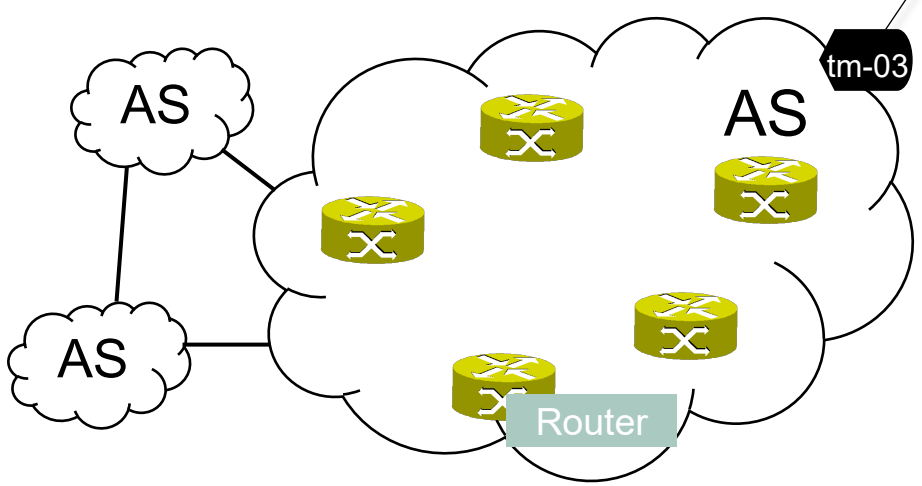
Karlsruher Institut für Technologie

Telematik

03. Internet Routing



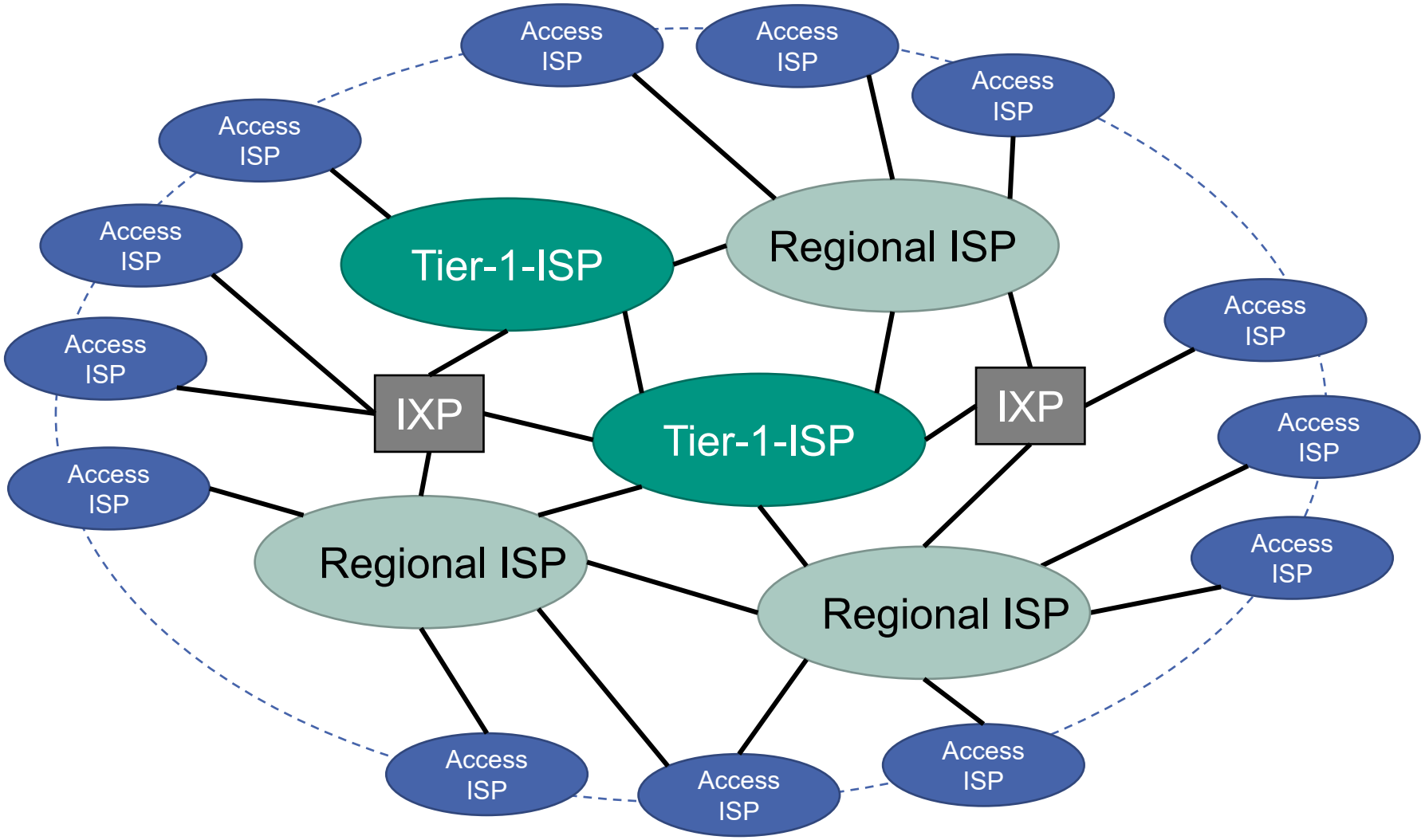


Structure of the Internet
Autonomous systems (AS)
…

3 Internet Routing

3.1 Basics

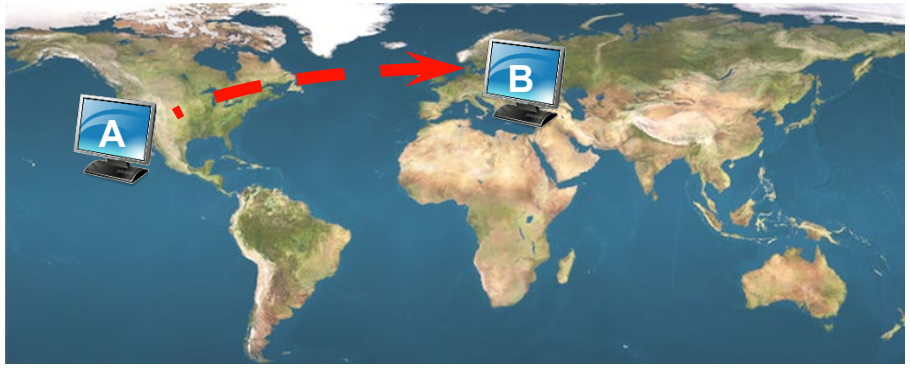
Internet: Network of Networks



ISP: Internet Service Provider
IXP: Internet Exchange Point

Routing

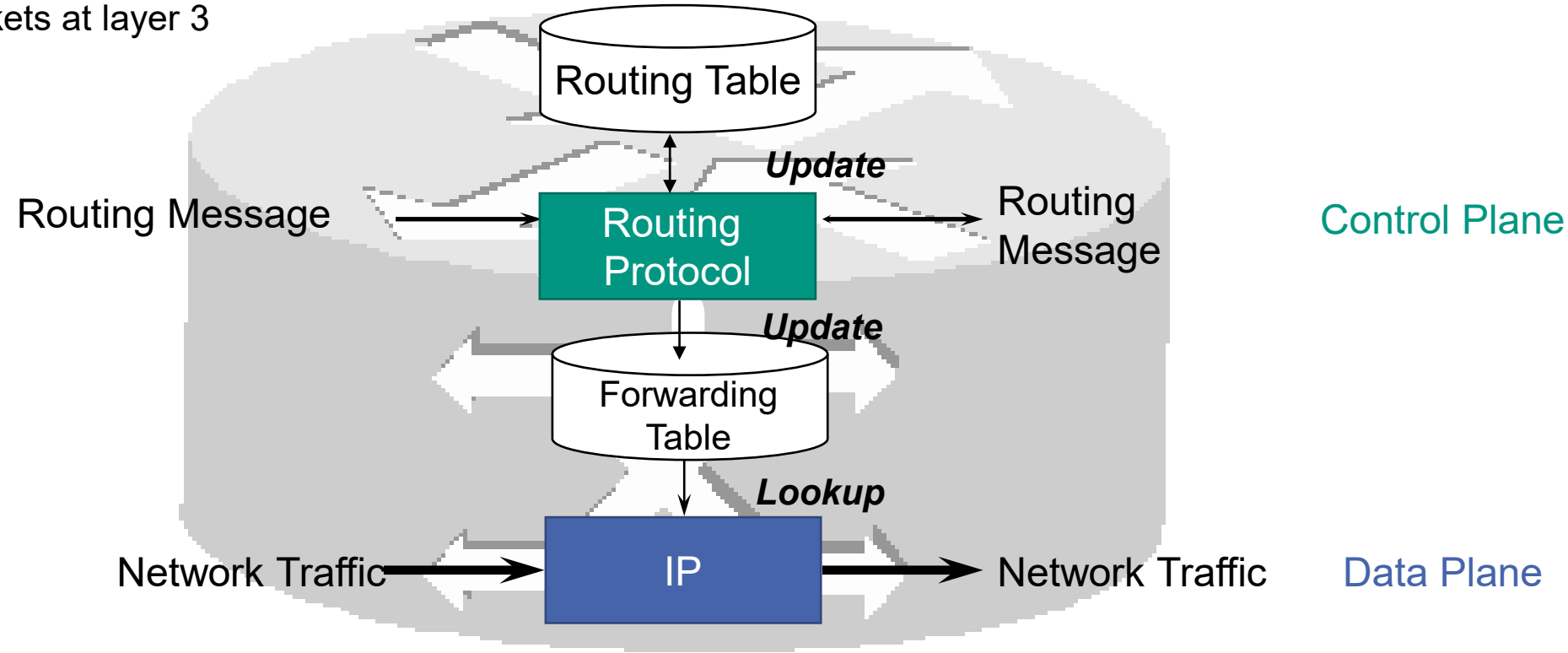
- Determines the path that the packets follow



- Routing is part of the control path
  - Requires routing algorithms and routing protocols

High-level View on an IP Router

- Control Plane
  - Routing protocols
  - Exchange of routing messages for calculation of routes
- Data Plane
  - Lookup
  - Forwarding of packets at layer 3



Definitions

- Routing table
  - Generated by routing protocol (control plane)
  - Entries
    - Mapping of destination IP prefixes to next hop (IP address)
  - Optimized for the particular routing algorithm
    - Changes in the topology
  - Performance is not critical
    - Implemented in software

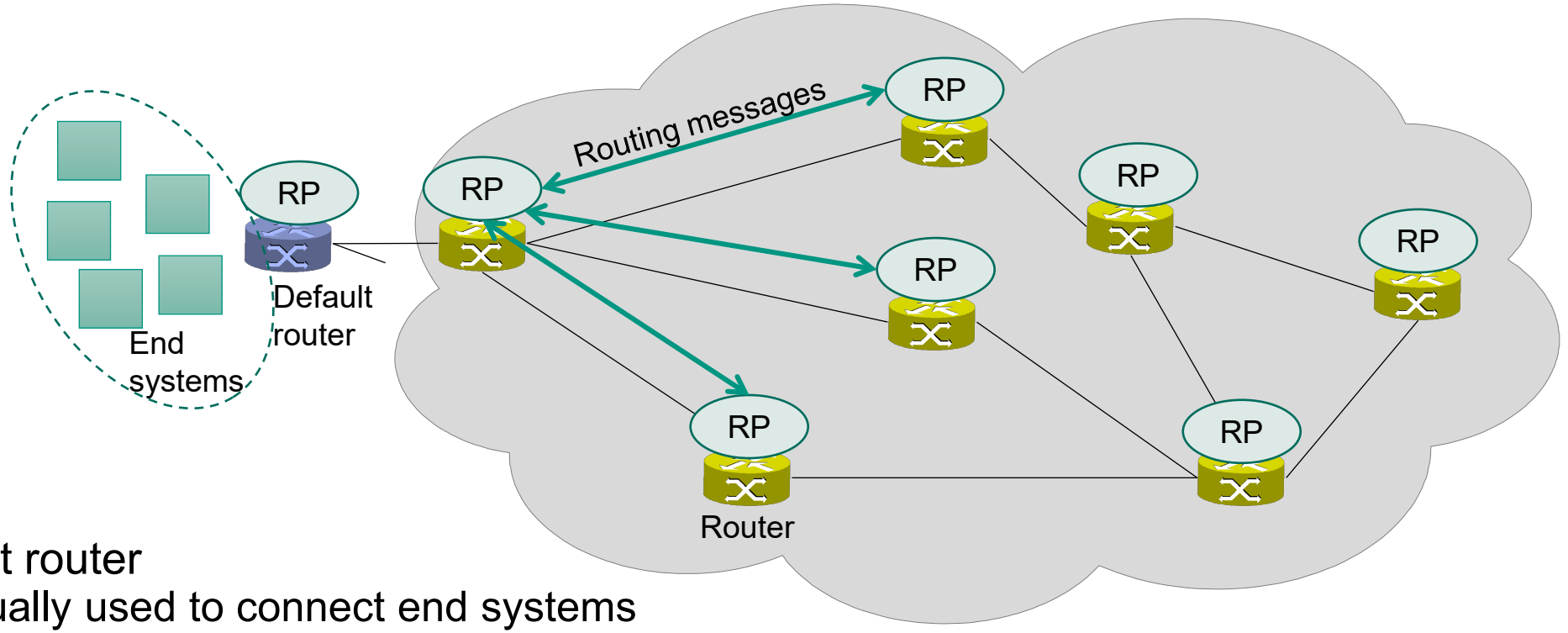
- Forwarding table
  - Used for packet forwarding (data plane)
  - Entries
    - Mapping of IP prefixes to outgoing ports (interface ID and MAC address)
  - Optimized for longest prefix matching
  - Performance is critical (lookup in line speed)
    - Partially uses dedicated hardware

Definitions

- Routing *metric* (also named cost, weight)
  - Metric used by a router to make routing decision
  - Can be applied to an individual link or to the overall path
  - Examples
    - Utilization, latency, data rate
    - Number of hops
- Routing *policy*
  - Policy-based routing decisions
  - Policies are defined by network operator / owner
  - Example
    - Prefer routes over specific neighbor-networks

Distributed Adaptive Routing

- Currently commonly used in the Internet
- Each router hosts an instance of the routing protocol
  - Exchange of routing information via routing messages
- Adaptation of paths to current situation in the network



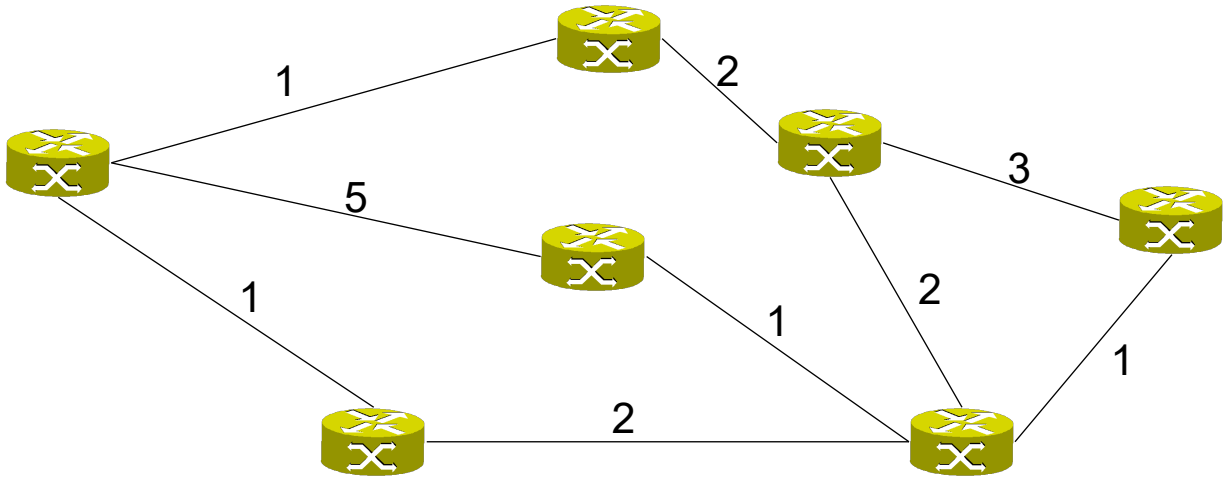
- Default router
  - Usually used to connect end systems

Instance of the Routing Protocol RP

Path Computation

- Network is modeled as graph
  - Graph $G = (N, E)$
    - N - nodes
      - Routers are nodes 
    - E - edges
      - Links between routers are edges
      - Edges are associated with metric

- Example

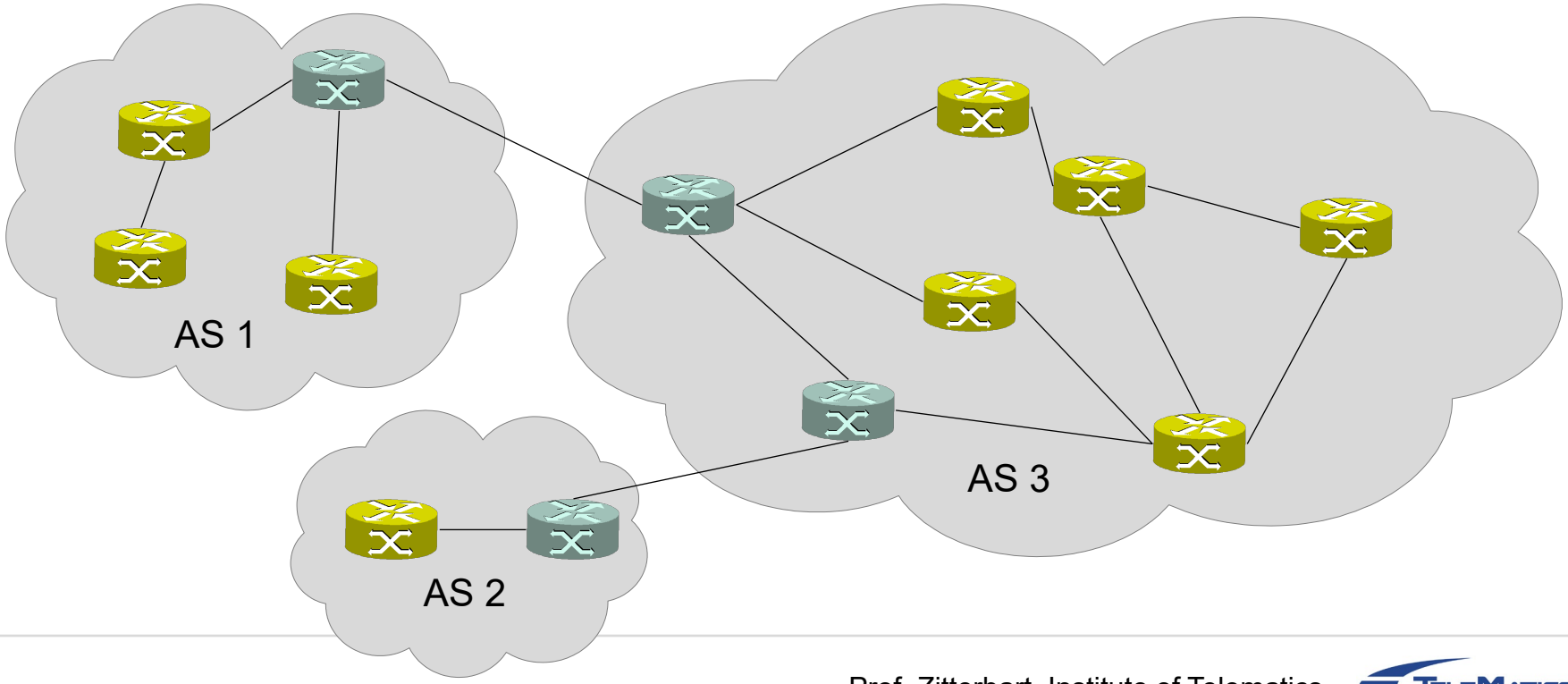


3.2 Autonomous Systems

3.2.1 Structuring into autonomous systems

Internet: Network of Networks

- Divided into Autonomous Systems (AS)
  - Routing inside an autonomous system
    - Interior Gateway Protocol (IGP)
  - Routing between autonomous systems
    - Exterior Gateway Protocol (EGP)



Autonomous System

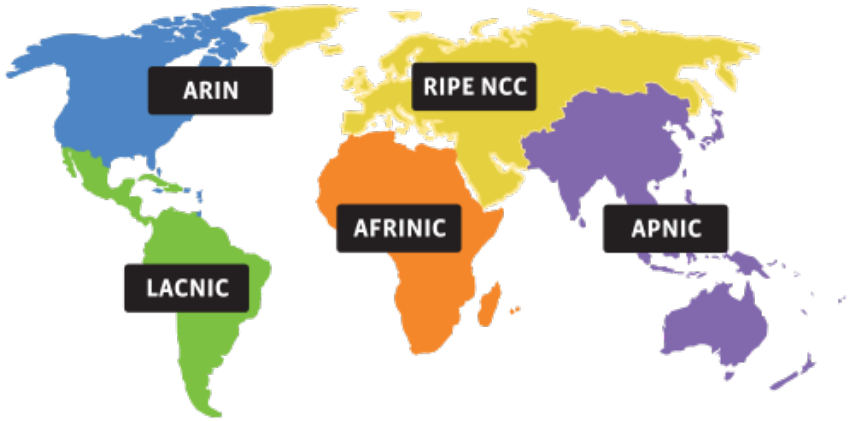
- Identification of an AS
  - Unique number - Autonomous Systems Number (ASN)
    - earlier 16 bit; now 32 bit
- Properties
  - Appears as a single entity to the outside
    - The same technical administration applies
  - Uniform routing policy
  - Typically uniform interior routing protocol
    - Different ASes can use different interior routing protocols
- Advantages
  - Separated administrative domains
  - Scalability by using two logical levels
    - Routing protocol inside an AS (not global)
    - Routing protocol between ASes

Important Properties

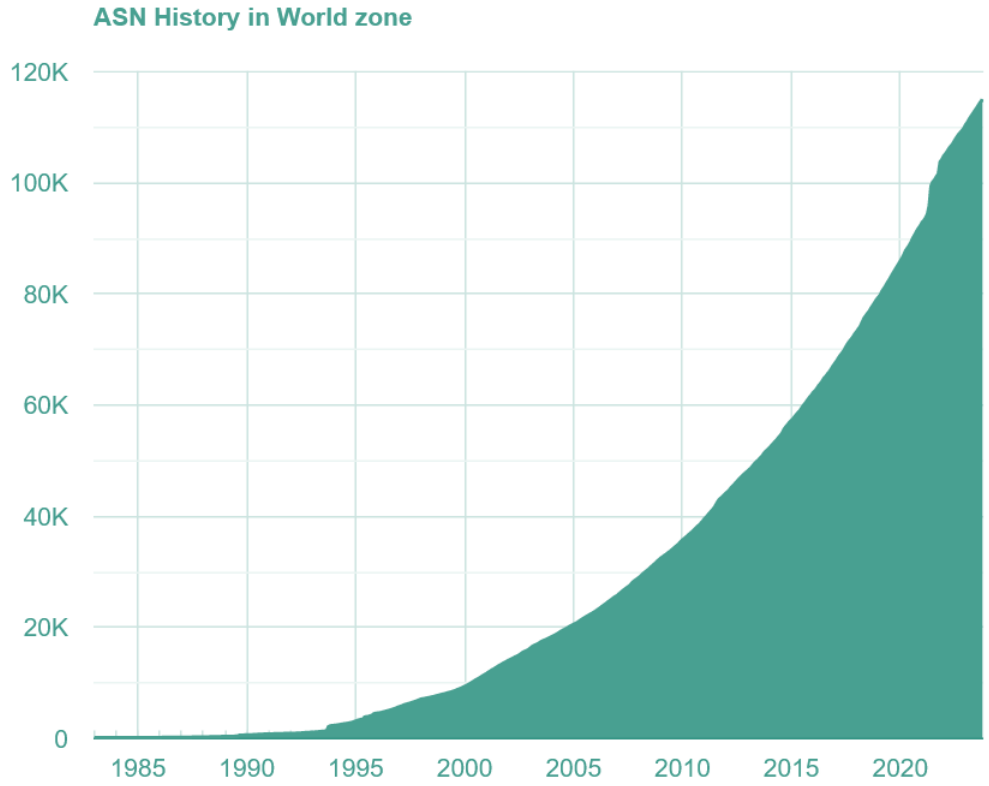
- Scalability of routing protocols
  - Overhead increases with size of the network
    - Space for storing routing information
    - Number of routing messages to exchange
    - Computation overhead
- Operator autonomy
  - Choice of interior routing protocol
  - Hiding of internal network structure

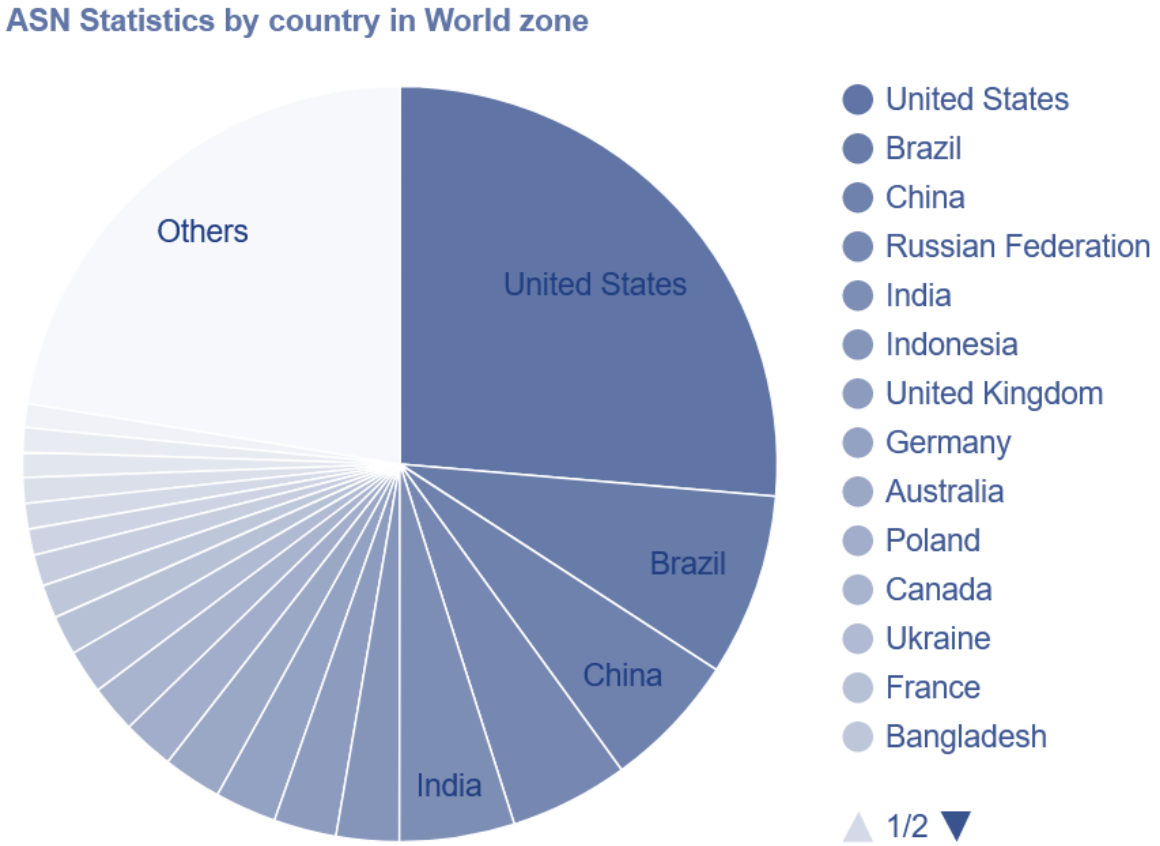
Allocation of AS-Numbers

- IANA (Internet Assigned Numbers Authority) delegates allocation to Regional Internet Registries (RIR), e.g.,
  - ARIN (North America)
  - RIPE NCC (Europe, Middle East and Central Asia)
  - APNIC (Asia-Pacific)
  - LACNIC (Latin America, Caribbean)
  - AfriNIC (Africa)



Allocation of AS-Numbers





[https://www-public.imtbstsp.eu/~maigron/RIR_Stats/RIR_Delegations/World/ASN-ByNb.html]

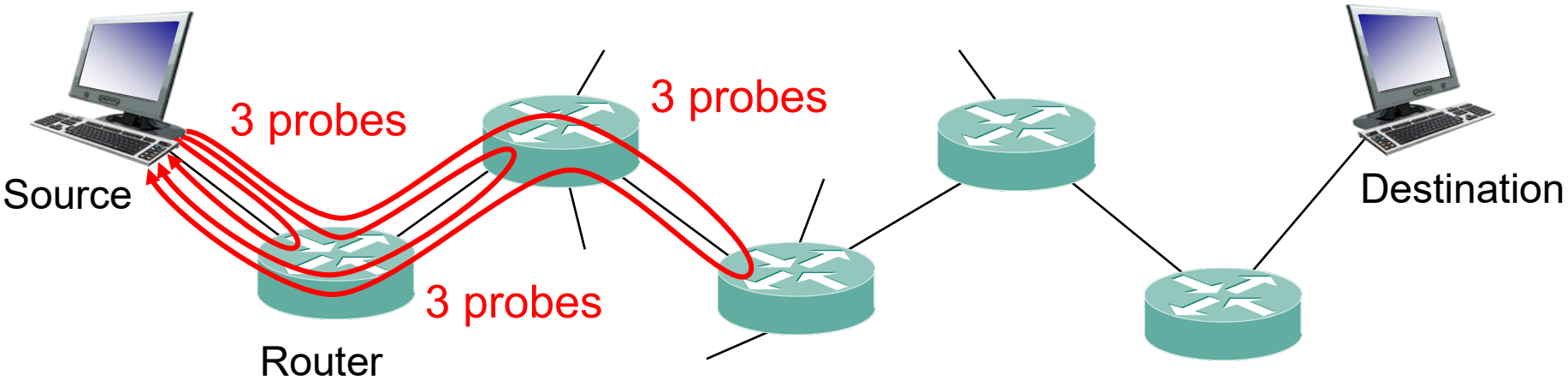
Traceroute

- Determines path towards destination

Windows: `>tracert kit.edu`
Linux: `$ traceroute kit.edu`

- Recap
  - Utilizes ICMP Echo Request packets with increasing time to live (1, 2, 3 …)
    - Routers R_1, R_2, R_3 … on the way to the destination
  - Routers reply with response packet

Allows to determine which ASes are traversed



Routing Example (traceroute)

- Traceroute from KIT network to a host in the USA (Sep. 2012):

```
$ traceroute -A www.stanford.edu
traceroute to www.stanford.edu (171.67.215.200), 30 hops max, 60 byte packets
 1  i72marbgate.tm.uni-karlsruhe.de (141.3.71.126) [AS34878]  0.152 ms  0.146 ms  0.137 ms
 2  172.16.4.1 (172.16.4.1) [*]  0.298 ms  0.292 ms  0.286 ms
 3  nbi-bd-rtr.link.uni-karlsruhe.de (141.3.127.100) [AS34878]  2.110 ms
 4  info-gw-int.link.uni-karlsruhe.de (141.3.127.190) [AS34878]  0.570 ms
 5  tr-v901-r-ir-cs-1.scc.kit.edu (129.13.70.201) [AS34878]  1.109 ms  1.181 ms  1.270 ms
 6  tr-v1288-r-ir-cn-1.scc.kit.edu (141.52.249.169) [AS34878]  1.560 ms  8.884 ms  8.973 ms
 7  xr-fzk1-te1-3-906.x-win.dfn.de (188.1.38.221) [AS680]  6.439 ms  2.3
 8  zr-fra1-te0-7-0-7.x-win.dfn.de (188.1.145.49) [AS680]  5.704 ms  5.8
 9  dfn.rt1.fra.de.geant.net (62.40.124.33) [AS20965]  4.695 ms  4.914 ms
10  abilene-wash-gw.rt1.fra.de.geant.net (62.40.125.18) [AS20965]  96.77
11  ae-8.10.rtr.atla.net.internet2.edu (64.57.28.6) [*]  111.541 ms  111.341 ms  126.955 ms
12  xe-1-0-0.0.rtr.hous.net.internet2.edu (64.57.28.112) [*]  133.702 ms
13  * ge-6-1-0.0.rtr.losa.net.internet2.edu (64.57.28.96) [*]  616.438 ms
14  hpr-lax-hpr--i2-newnet.cenic.net (137.164.26.133) [AS2152]  164.168 ms  164.484 ms  164.272 ms
15  svl-hpr2--lax-hpr2-10g.cenic.net (137.164.25.38) [AS2152]  172.160 ms
16  hpr-stanford--svl-hpr2-10ge.cenic.net (137.164.27.62) [AS2152]  172.8
17  boundarya-rtr.Stanford.EDU (171.66.0.34) [AS32]  185.682 ms  185.478 ms  185.664 ms
18  * * *
19  * * *
20  www-v6.Stanford.EDU (171.67.215.200) [AS32]  173.472 ms  173.541 ms  173.697 ms
```

KIT Karlsruhe Institute of Technology (KIT)

DFN Verein zur Förderung eines Deutschen Forschungsnetzes e.V.

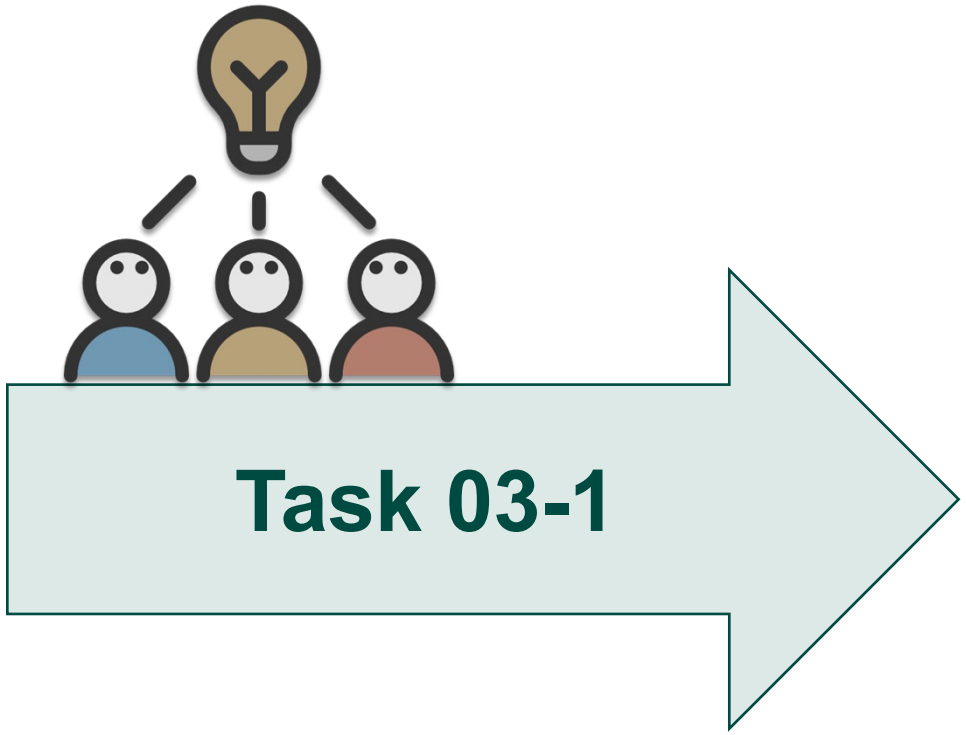
GEANT The GEANT IP Service

ABILENE - Internet2

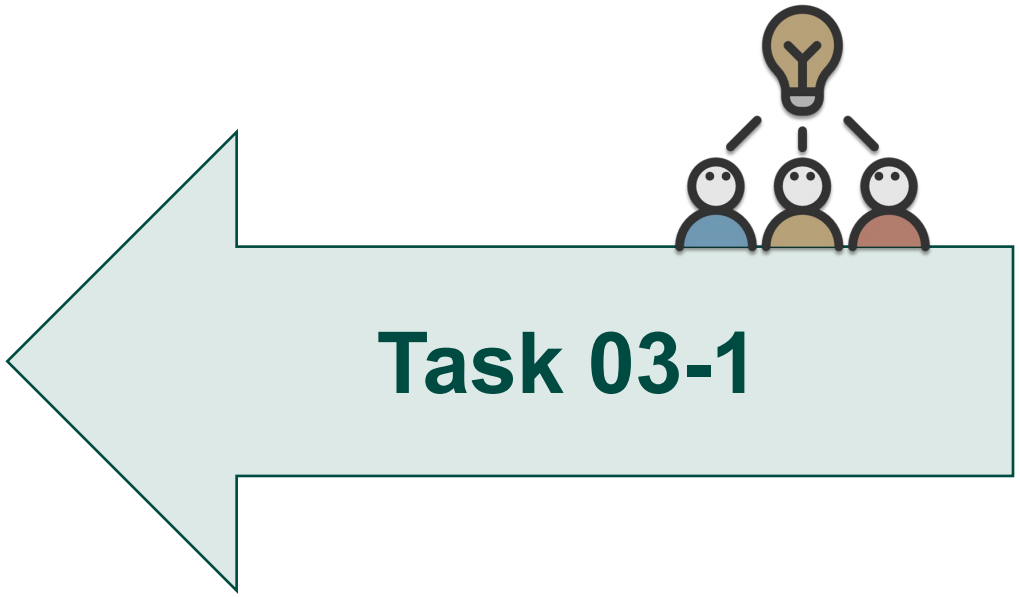
CSUNET-NW - California State University Network

STANFORD - Stanford University

- Average AS path length in Internet is approximatelly 3,83

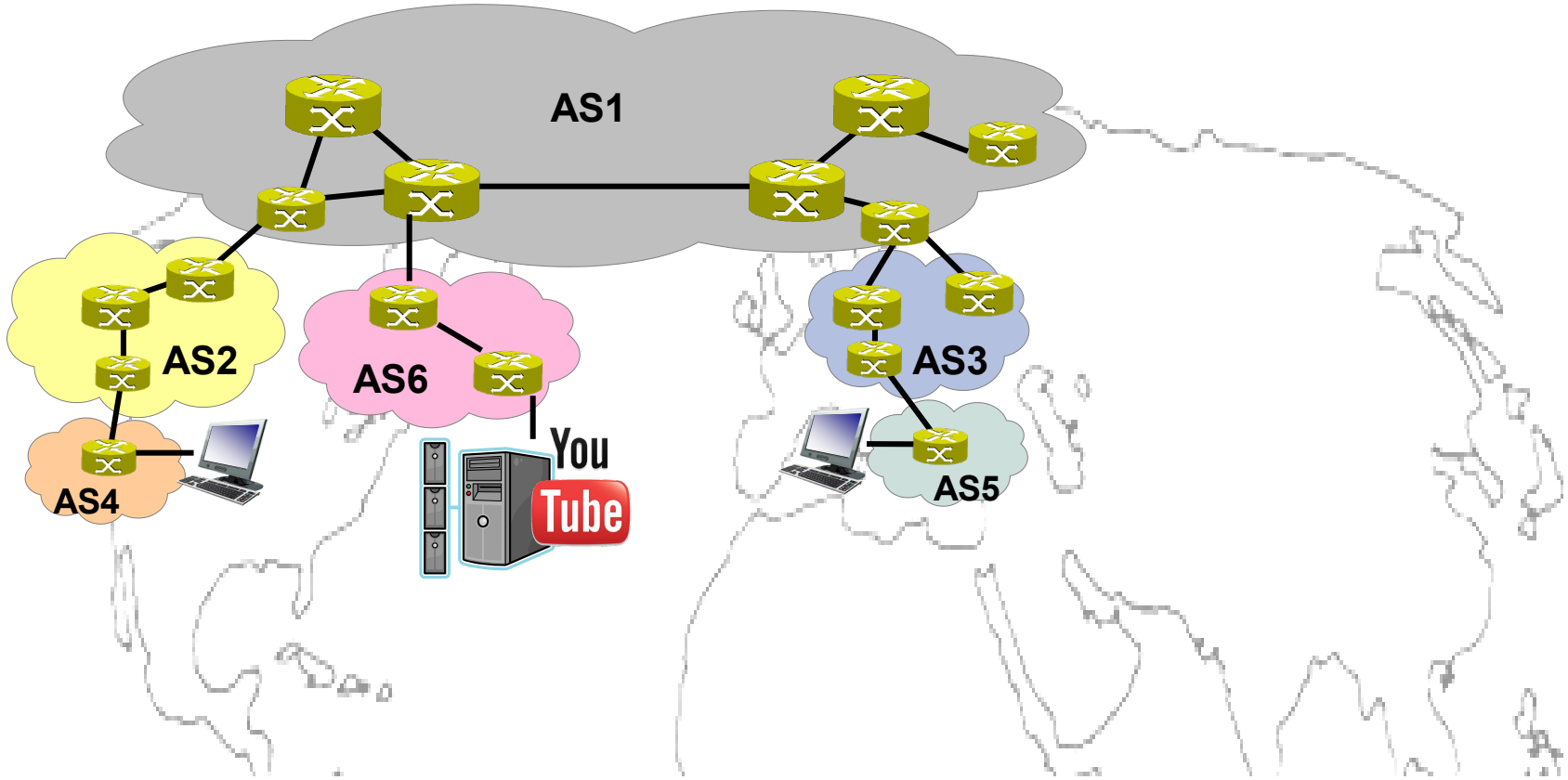


Traceroute,
Autonomous
Systems (ASes)


Task 03-1

Subdivision into Autonomous Systems

- Illustration



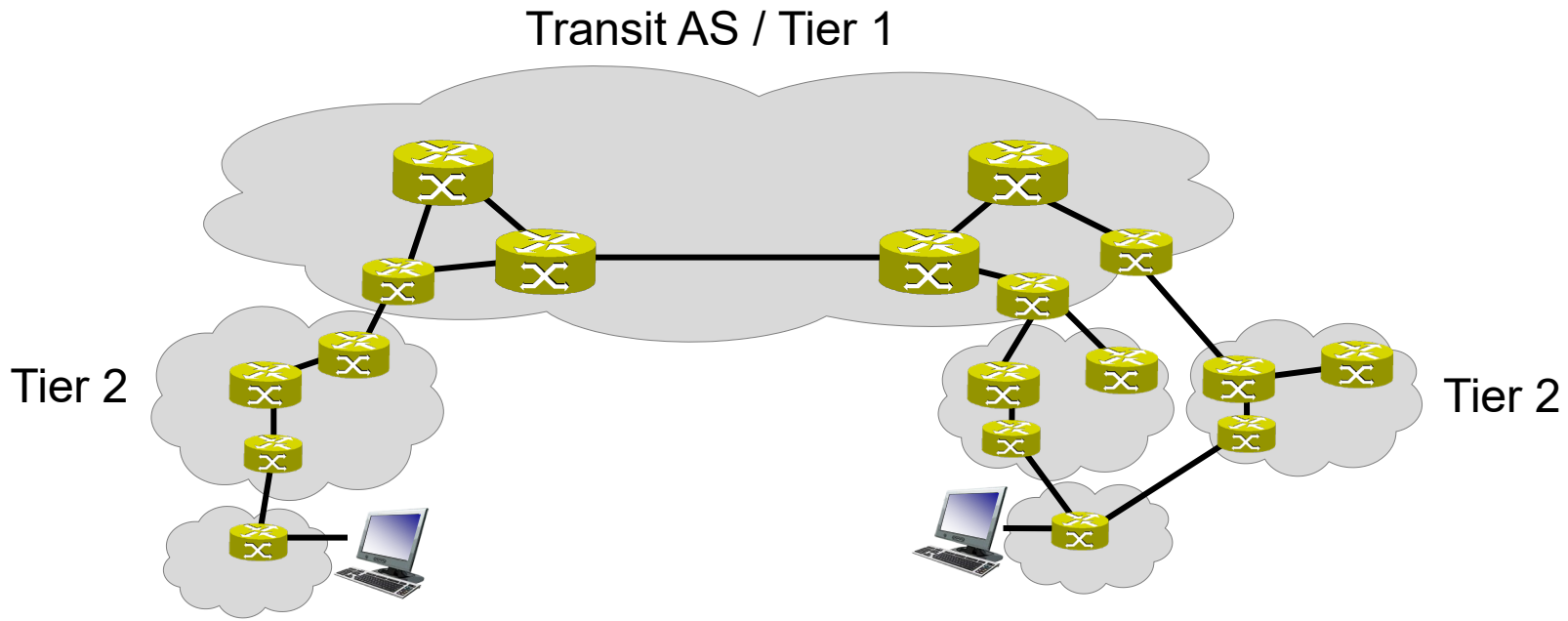
Classification of Autonomous Systems

- Classification based on role
  - Stub AS
    - Small organizations and enterprises
      - Mostly operate only regionally
    - Connected to exactly one provider
    - No transit traffic
  - Multihomed AS
    - Large enterprises
    - Connected to several providers (reliability)
    - No transit traffic
  - Transit AS
    - Provider
      - Often global scope

Classification of Autonomous Systems

- Classification based on “economic position/influence”
  - Tier 1, tier 2, tier 3 …

- Illustration



Different Roles

- End customer
  - Uses Internet application
  - Examples
    - Universities
    - Enterprises
    - Customers of Internet Service Providers (ISP), e.g.,
      - 1&1, KabelBW, AOL, Telekom, O2, Vodafone

- Content delivery provider
  - Requested by end customers / Internet application
    - Provide content
  - Examples
    - Google, Akamai, Yahoo, YouTube, Facebook

… how to ensure reachability *(and performance) ?*

3.2.2 Reachability across autonomous systems

Reachability

- Problem
  - How to ensure mutual reachability?
  - Cooperation among autonomous systems?

- Basic concepts
  - Transit
    - Purchased connectivity
  - Peering
    - Direct connection, typically between ASes of the same tier

Connectivity and Transit

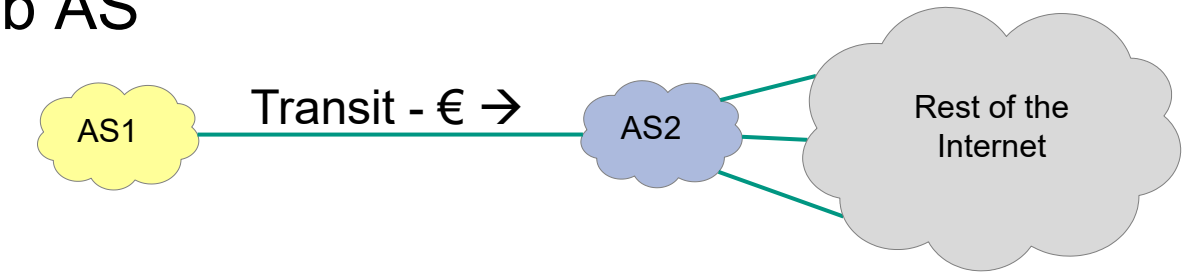
- Establish connectivity
  - Establish paths to all other ASes in the Internet
  - AS operator purchases connectivity from one or more ASes

- Transit
  - Purchased connectivity
    - Upstream: provider (seller) of transit
    - Downstream: customer (buyer)

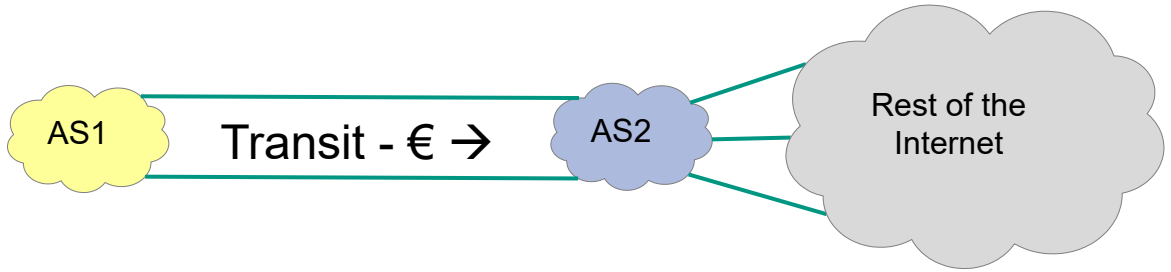
      … hierarchical customer/provider structure
  - Traffic exchange
    - In both directions
    - Only downstream AS must pay; usually volume rate
  - Transit AS
    - Provider AS, that offers transit

Connectivity and Transit

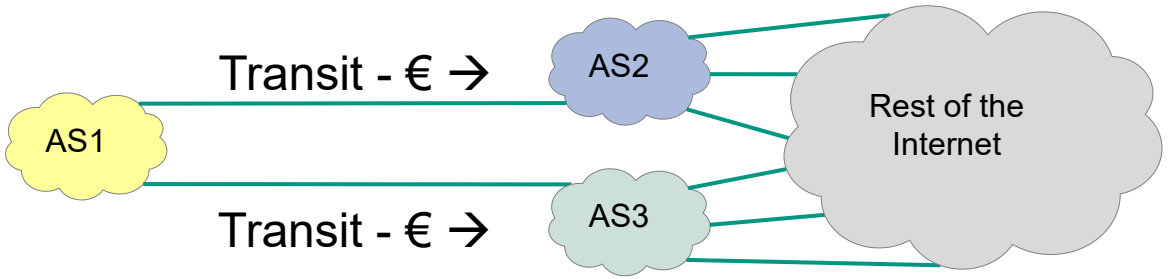
- Options for connecting a stub AS
  - Stub AS



  - Dualhomed stub AS



  - Multihomed stub AS



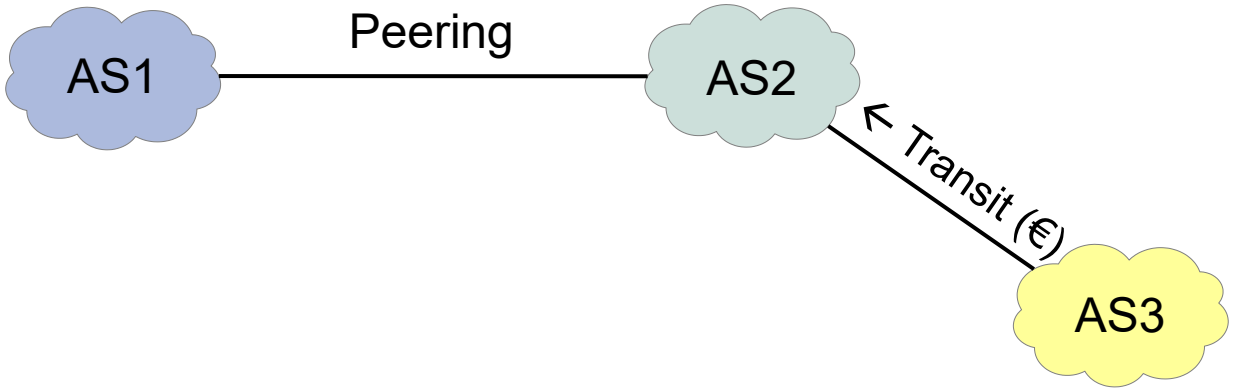
Peering

- Private peering
  - Direct connection between two ASes, usually of same tier
  - No cost for traffic exchange; costs for network infrastructure apply
  - However
    - Mostly only data traffic between privately peered ASes
    - No transit traffic of other ASes

  … connectivity is not achieved by private peering

https://youtu.be/T2jb1tzXzMw

- Example: peering and transit

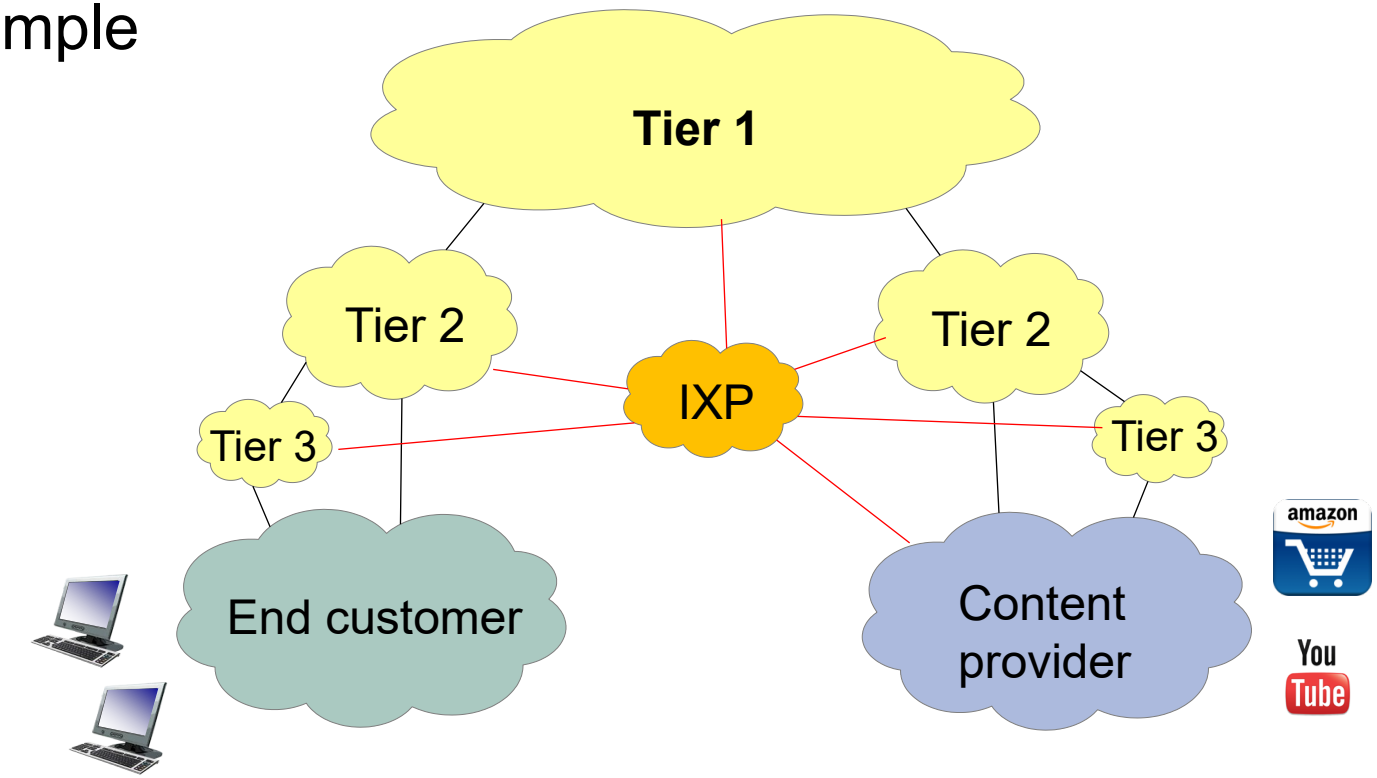


Peering

- Private peering
  - Advantages
    - Benefits both ASes: save transit costs, that otherwise would apply
    - Shorter data paths: fewer AS hops between source and destination
  - Problems
    - Direct connection of ASes complicated
      - Different geographical locations
    - Full mesh of n ASes
      - $\frac{(n-1) * n}{2}$ separate connections
      - *… currently more than 1 billion*

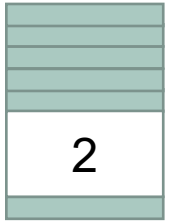
Peering

- Public peering
  - Through Internet exchange points (IXPs)
    - Central public authority for interconnection
- Example



Peering

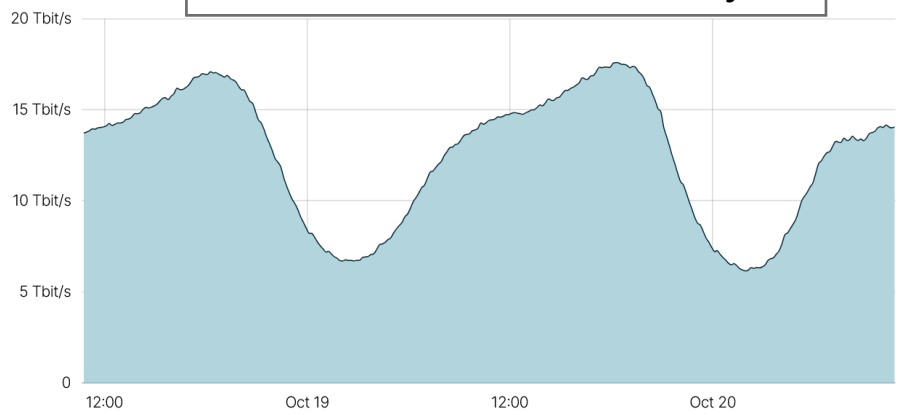
- Public peering
  - Internet exchange point
    - Neutral traffic forwarding on layer 2
    - No differentiation regardless of customer, content, or type of service
    - Examples
      - DE-CIX, AMS-IX, LINX, Equinix
  - Members / customers
    - Monthly fixed charges per network port
      - Necessary for operation and maintenance of IXP‘s switching platform
    - Usage
      - E.g., for peering, transit, PVLANs



Peering

- Public peering
  - Different peering policies
    - Open: AS is open for peering with all other ASes
    - Selective: Peering only under given terms and conditions
    - Restrictive: AS does not engage in new peering relationships
    - No Peering: AS does not do any peering

Traffic Frankfurt – 2 days



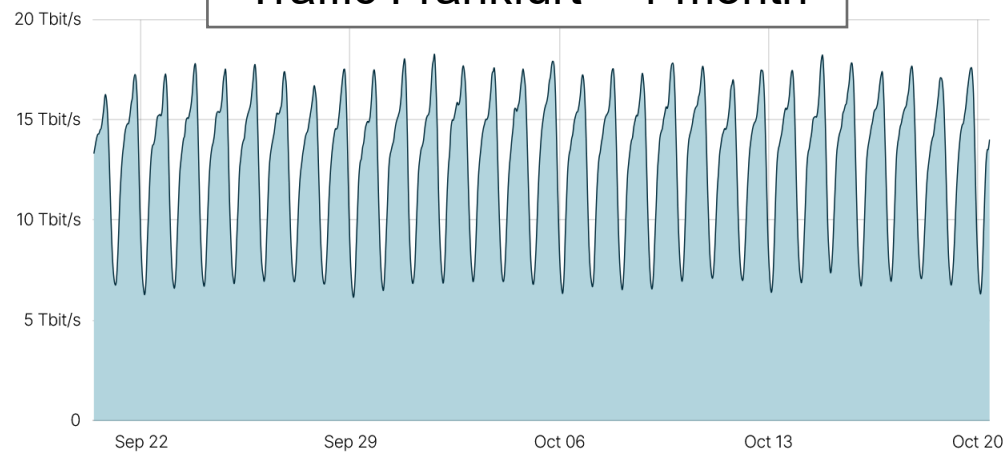
ALL-TIME PEAK **18.22** Tbit/s

GRAPH PEAK **17.63** Tbit/s

GRAPH AVERAGE **13.70** Tbit/s

CURRENT **13.95** Tbit/s

Traffic Frankfurt – 1 month



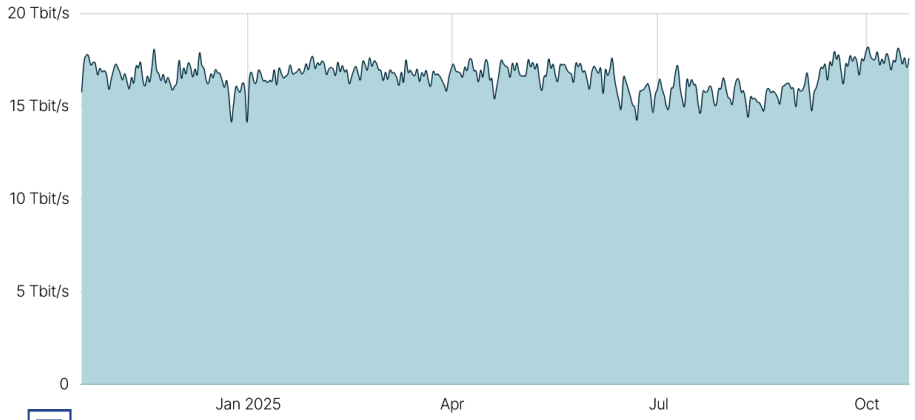
ALL-TIME PEAK **18.22** Tbit/s

GRAPH PEAK **18.22** Tbit/s

GRAPH AVERAGE **13.72** Tbit/s

CURRENT **13.95** Tbit/s

Traffic Frankfurt – 1 year



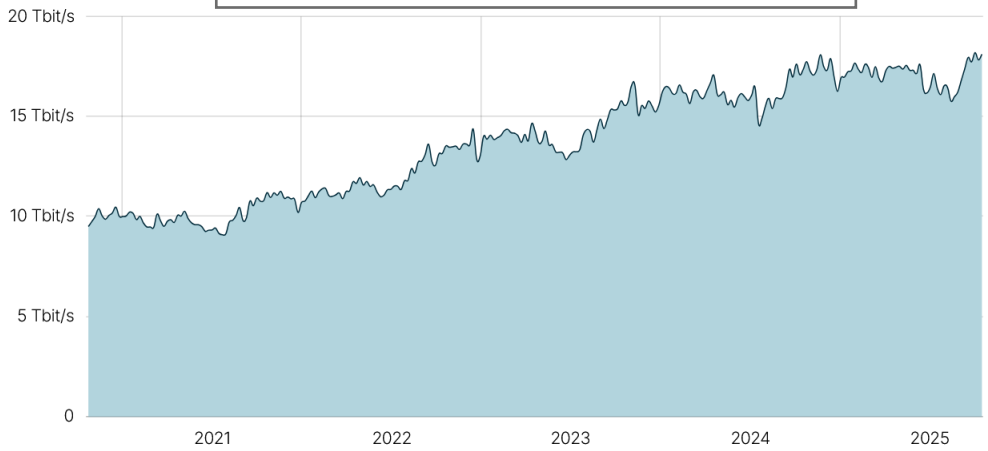
ALL-TIME PEAK **18.22** Tbit/s

GRAPH PEAK **18.22** Tbit/s

GRAPH AVERAGE **13.39** Tbit/s

CURRENT **14.10** Tbit/s

Traffic Frankfurt – 5 years



ALL-TIME PEAK **18.22** Tbit/s

GRAPH PEAK **18.22** Tbit/s

GRAPH AVERAGE **9.82** Tbit/s

CURRENT **14.10** Tbit/s

[https://www.de-cix.net/en/locations/frankfurt/statistics], October 20th, 2025

Autonomous Systems and Transit/Peering

- Tier 1
  - Large global ASes with access to (all) other ASes
    - Do *not* buy any transit
    - Sell transit
    - Peering with other tier 1 ASes
  - Examples
    - Deutsche Telekom, AT&T, NTT,Tata Communications, TeliaSonera
- Tier 2
  - Big national and inter-regional ASes
    - Connection to providers of Internet applications
  - Downstream of tier 1 ASes
    - Sell transit to other ASes
    - Usually employ peering
  - Examples
    - Vodafone, Comcast, Tele2

Autonomous Systems and Transit/Peering

- Tier 3
  - Small mostly regional ASes
    - Connections with small providers of Internet applications
  - Downstream of tier 2 providers
    - Usually do not sell transit to other ASes
    - Sell transit mostly to end customers/users
    - Usually employ peering
  - Examples
    - KabelBW, NETHINKS, Alice, Congstar, Versatel

Content Provider

- Goal
  - Fast delivery of content

  … low latencies are important!

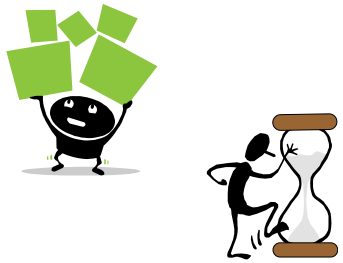
- For this purpose
  - Locations close to tier 1 peering points are preferred

  → peering with eyeball networks (i.e., access networks)
    - ISPs that sell Internet access to end users

- Two basic alternatives
  - Web servers are hosted directly in tier 1 ASes
    - Does not require an own AS number
  - Web servers are connected over own routers
    - Content delivery network (CDN)
    - Own AS number required
    - Peering with essential content providers at important peering points
    - Examples
      - Google, Yahoo, Akamai

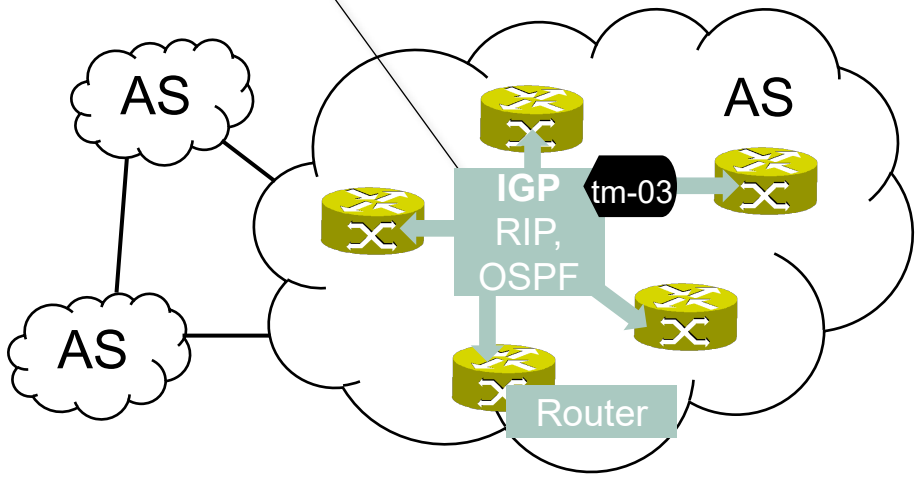
Content Delivery Network

- World wide network with own AS number
  - Thousands of Points of Presence (PoP) spread across the world
- Point of Presence
  - Consists of access routers und core routers
    - Access router at the edge of a CDN
    - Core router inside a CDN
  - Customers are connecting through access routers
- Objectives
  - Load balancing at access routers
    - Selection of the best suitable access routers
  - Be close to customers … low latencies



Karlsruher Institut für Technologie

Control path
Routing protocols inside ASes

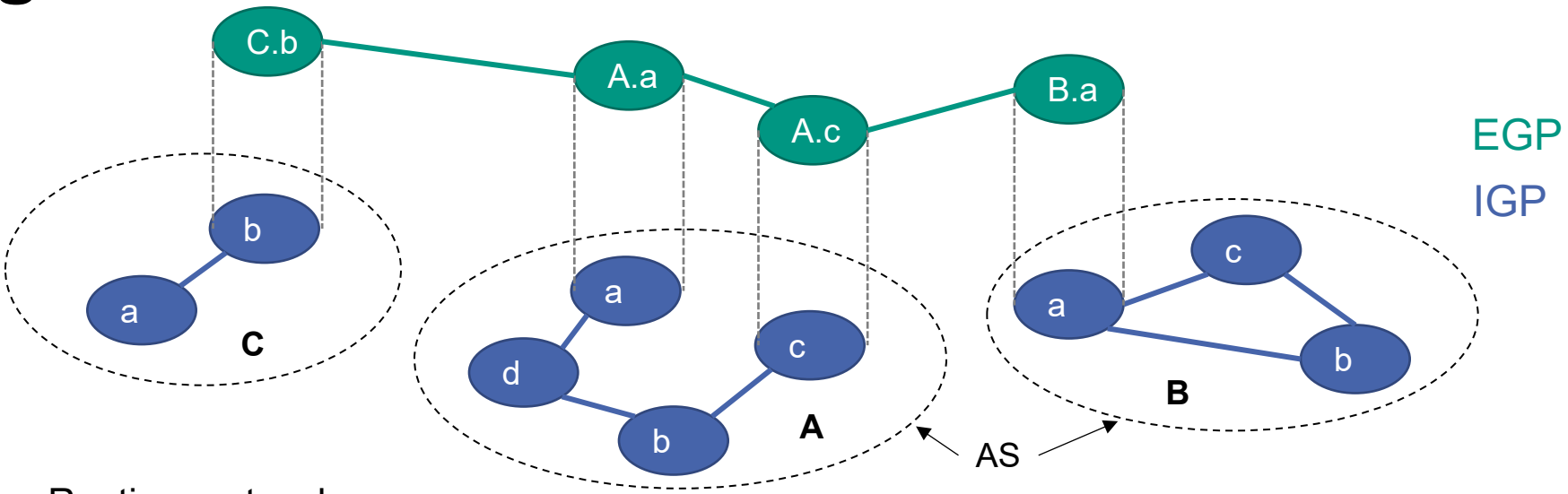


3.2.3 Routing in and between Autonomous Systems

Routing Protocols

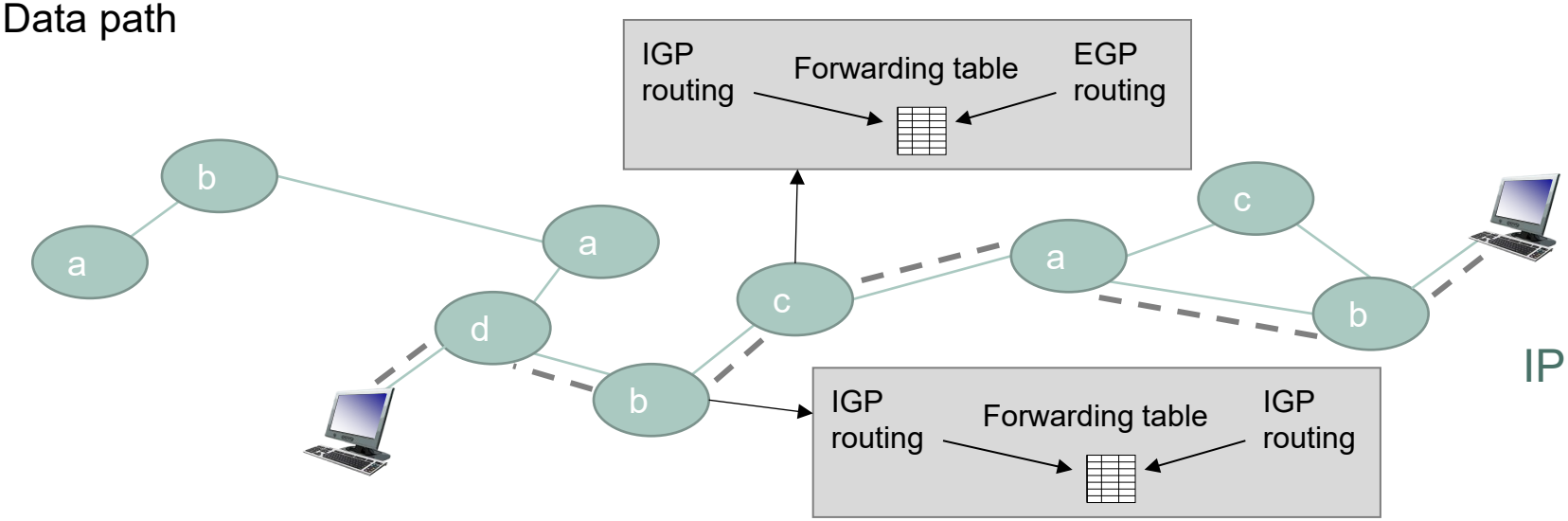
- Classification in
  - Interior gateway protocols (IGPs) inside one AS
    - Also named intra-domain routing protocols
    - Are encapsulated inside an AS, i.e., not visible to the outside
    - Different IGPs in different ASes possible
    - Metric-based
  - Exterior gateway protocols (EGPs) between ASes
    - Also named inter-domain routing protocols
    - Single protocol between *all* ASes
    - Policy-based

Routing with IGPs snd EGPs



Routing protocols

Data path



Interior Gateway Protocols

- RIP (*Routing Information Protocol*)
  - Based on distance vector algorithm
  - Very simple, does not scale to larger networks
  - Not used in ASes


[Huit00]

- OSPF (*Open Shortest Path First*)
  - Based on link state algorithm
  - For some time “the” IGP protocol

- IS-IS (*Intra-Domain Intermediate System to Intermediate System Routing Protocol*)
  - Standardized as ISO-Standard 10589
  - Based on link state algorithm
  - Increasingly used


[HoWa06]

- EIGRP (*Enhanced Interior Gateway Routing Protocol*)
  - CISCO proprietary, based on distance vector algorithm
  - “extension” of RIP to solve count-to-infinity


[RFC7868]

3.3 RIP: Routing Information Protocol

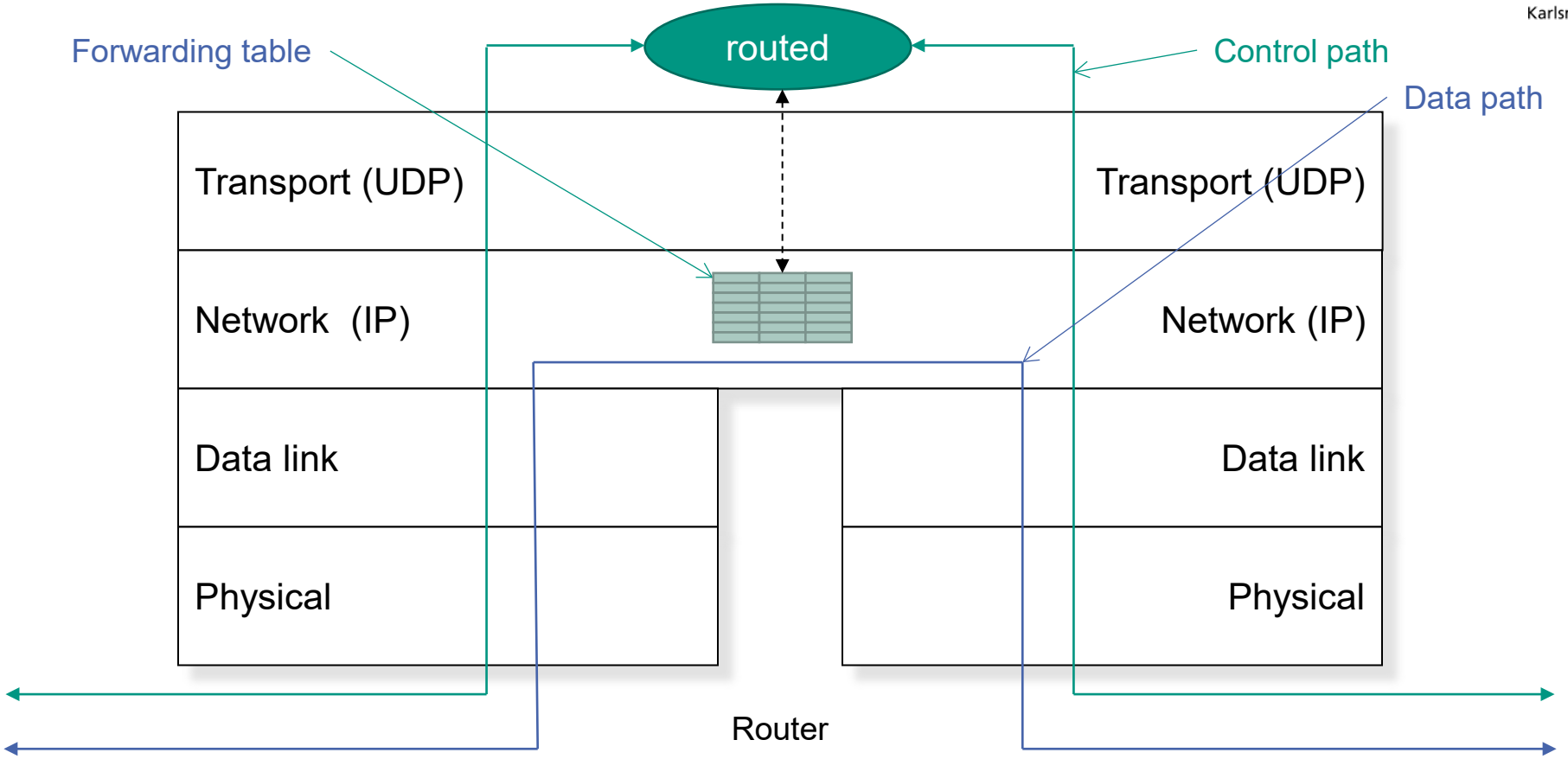
Routing Information Protocol (RIP)

- Interior gateway protocol
- One of the first routing protocols in the Internet
  - In BSD unix implemented and installed as „routed“ (*route management daemon*)
- Very simple protocol that requires very little configuration
- Documented in
  - RFC 1058 in June 1988 (RIPv1 – Version 1), historic
  - Suggestions for improvement were already included
    - Address common problems of distance-vector algorithms: loops, count-to-infinity
    - Suggestions: split horizon etc.
  - RFC 2453 (RIPv2 – Version 2), actual version

[RFC1058]

[RFC2453]

RIP in the Protocol Stack



- Application process routed implements RIP and manages forwarding table
- RIP routing messages are sent over UDP
  - Not reliable

Routing Metric

- Distance between source and destination corresponds to number of hops on the path (hop count)

- Hop count
  - Limited range of values: 1 - 15
  - Value of 16 corresponds to „infinity“
  - Limiting the maximum value helps to cope with, e.g., count-to-infinity

RIP Routing Messages

- RIP protocol entities exchange routing messages
  - UDP is used as transport protocol

- Types of routing messages
  - Request message
    - Requests complete routing table or part of it
  - Response message for different reasons
    - Response to specific query
    - Regular update
      - Broadcasted/multicasted every 30 seconds
    - Triggered update
      - Metric for a route changed

Outgoing Routing Updates

- Regular routing update
  - Periodically, every 30 seconds
  - Sends entire routing table to all its neighbors
  - Entries in the routing table are periodically refreshed
  - No refresh for at least 180 seconds
    - → Hop-Count is set to 16 („infinite“), corresponding route is invalidated

- Metric for route changes (triggered update)
  - Only changes since last update are communicated, not complete routing table
  - Rate limitation in order to reduce load on the network
    - Minimal time interval between consecutive triggered updates
      - Randomized value between 1 and 5 seconds
      - All changes during this period are accumulated and sent in a single routing update

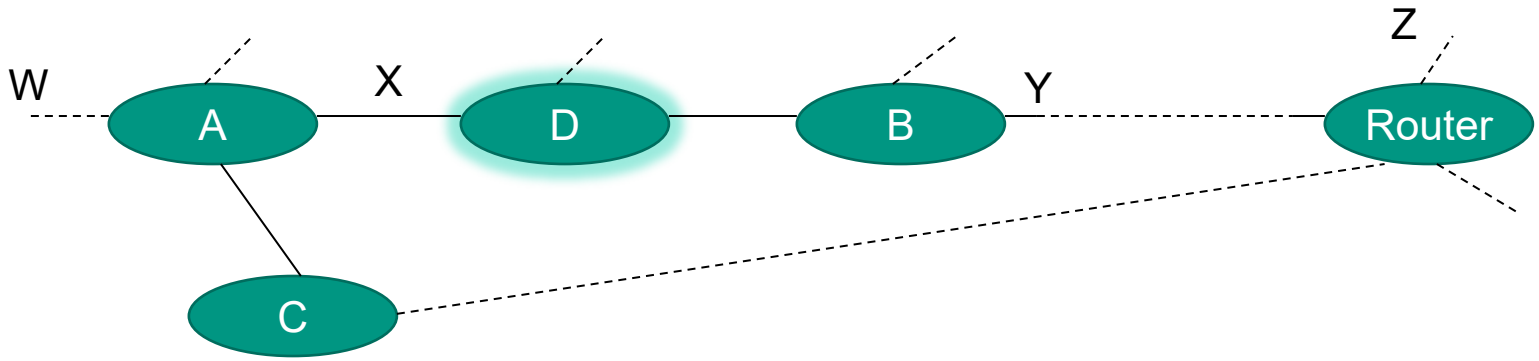
[RFC2453]

Incoming Routing Updates

- The following cases can be distinguished
  - Entry for a destination address does not exist in routing table and received metric is not „infinite“
    - Insert new entry in routing table
  - Current entry for a destination address in routing table has larger metric or routing update was sent by the “next router” for this destination
    - Modify entry
  - Otherwise
    - Ignore routing update

Example (I)

- Scenario
  - Connecting lines represent either direct links or LANs between routers
  - Ovals represent routers

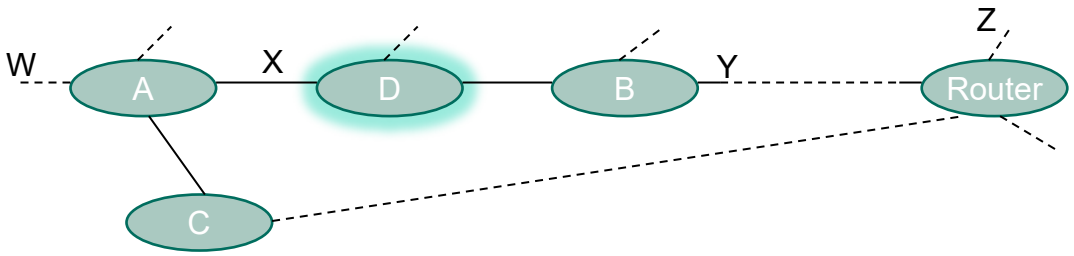


- Routing table of router D

Target Prefix	Next Router	Hop Count
W	A	2
Y	B	2
Z	B	7
X	–	1
…	…	…

Example (II)

- 30 seconds later
  - D receives new routing update from A



Target Prefix	Hop Count
Z	4
W	1
X	1
…	…

  - New routing table of D

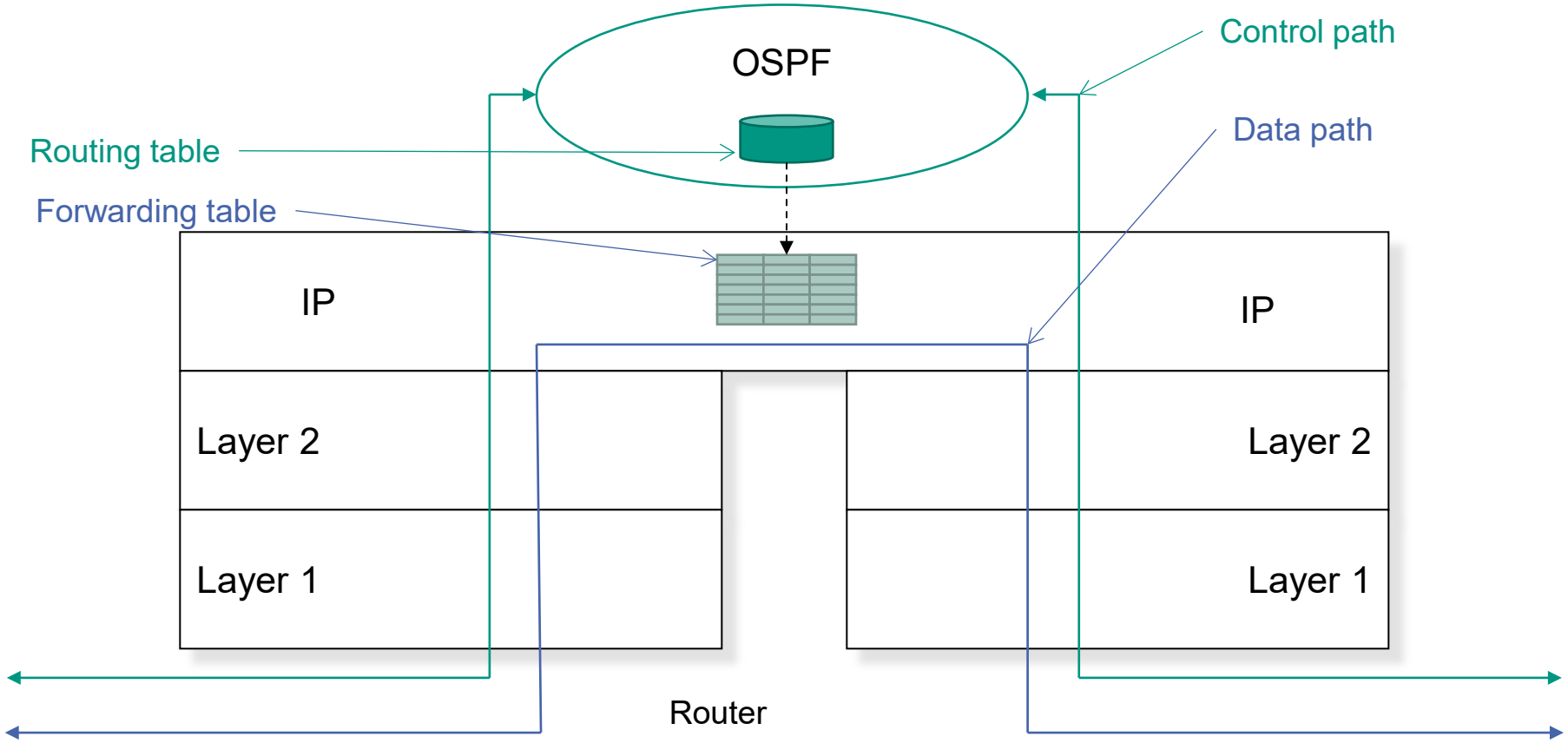
Target Prefix	Next Router	Hop Count
W	A	2
Y	B	2
Z	B	7
X	–	1
…	…	…

Target Prefix	Next Router	Hop Count
W	A	2
Y	B	2
Z	A	5
…	…	…

3.4 OSPF: Open Shortest Path First

3.4.1 Brief Overview of OSPF

OSPF in the Protocol Stack



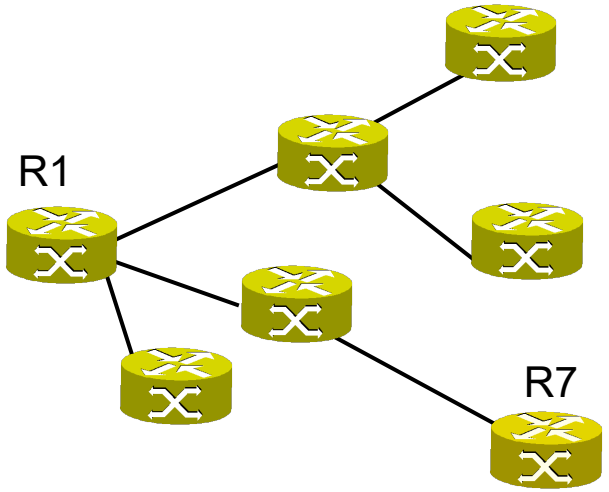
- OSPF is located on top of IP

→ OSPF uses an unreliable communication service

OSPF

- Interior gateway protocol
- Link state protocol
  - Every router in the network needs to learn complete network topology
    - Nodes and links with their costs (weights)
      - Nodes are OSPF routers
  - Every router separately computes shortest paths based on network topology
    - Constructs a tree of shortest paths with itself as a root
      - Example for router R1
    - Uses dijkstra shortest path algorithm
- OSPF is defined in RFC 2328

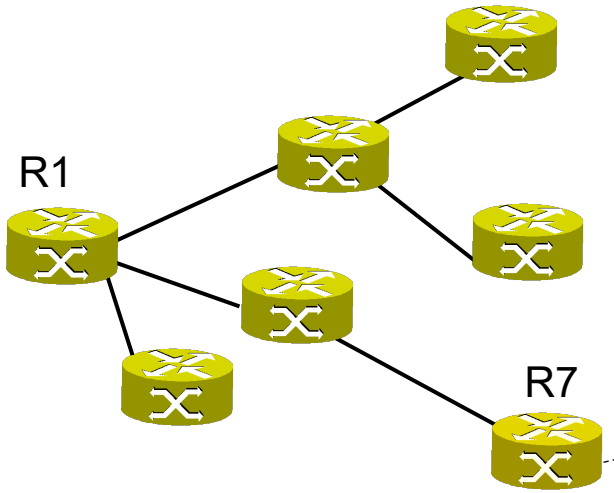
[RFC2328]

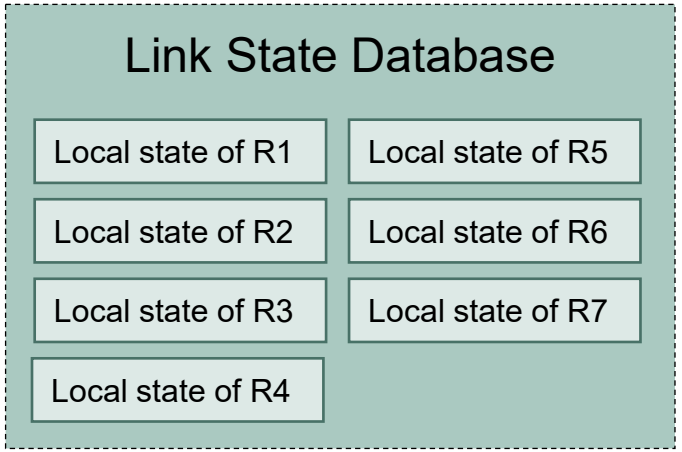


Every router needs to have identical knowledge of the network topology. Otherwise, calculated paths are inconsistent.

Link State Database

- Each router has a link state database
  - Identical on every router (after convergence)
  - Stores local state of all routers in the network
    - Usable interfaces and reachable neighbors



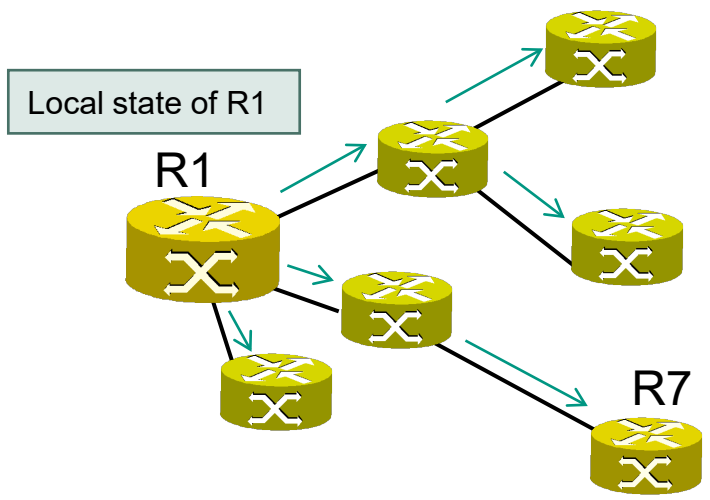


- Link state database is used to
  - Construct topology graph of the network and
  - Calculate routing table

Routers have identical knowledge of network topology iff their link state databases are synchronized, i.e., they have identical content at all routers.

Flooding of Router´s Local State

- Every router floods its local state throughout the routing domain
  - All routers in the network must receive an identical copy of this information
    - It is stored in the link state database of every router



- All OSPF protocol exchanges are authenticated
  - Only trusted routers can participate in routing

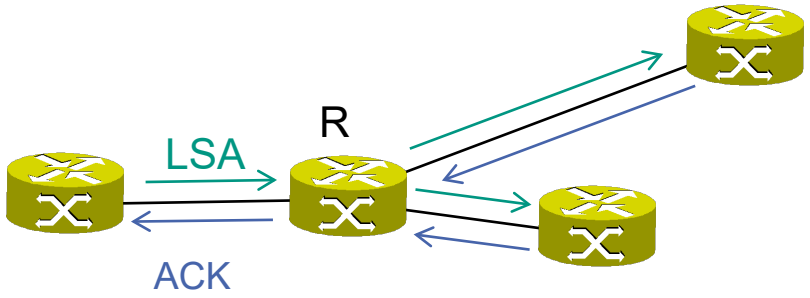
Synchronized Link State Databases (I)

- Goal: link state databases of all routers need to have identical content
  - Need to be synchronized

- The following actions are needed
  - Ensure that every LSA is received by every router in the network
    → reliable flooding

Reliable Flooding

- Reception of an LSA is acknowledged by neighboring router
  - These are hop-by-hop acknowledgements



Synchronized Link State Databases (II)

- Goal: link state databases of all routers need to have identical content
  - Need to be synchronized

- The following actions are needed
  - Ensure that every LSA is received by every router in the network
    → reliable flooding

  - Ensure that all routers consistently either store or discard each LSA
    → fully deterministic comparison rules

  - Ensure that expired LSAs are deleted from link state databases of every router
    - LSA lifetime associated with each LSA
    - LSAs age out while being in the database, but
      - Clocks of the routers may not be synchronized
        → LSAs can expire faster on some routers
    - Synchronize deletion of LSAs among routers
      → flooding of expired LSAs and special handling of MaxAge LSAs

Routing Metric

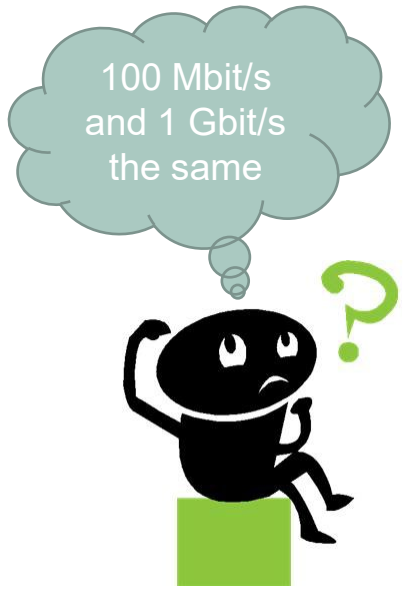
- Each link is associated with link costs
  - Example: prefer links with higher data rate

$$Cost = \frac{ReferenceDataRate}{InterfaceDataRate}$$

  - $ReferenceDataRate = 100\ Mbit/s$ (default value on Cisco routers)

LinkDataRate	Link Cost
64 kbit/s	1562
100 Mbit/s	1
1 Gbit/s	1

  - $ReferenceDataRate$ can be configured
    - E.g., to $1\ Gbit/s$ or $10\ Gbit/s$
    - Should be consistent across all routers in network



3.4.2 Link State Advertisement

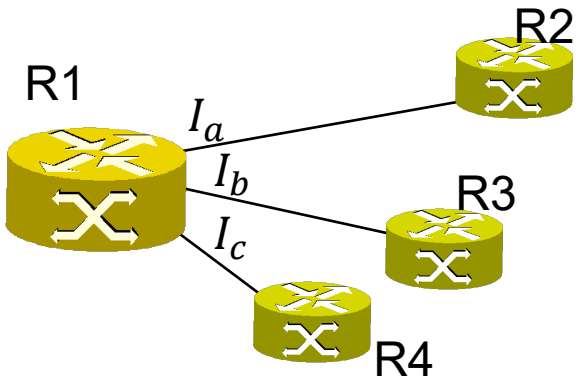
Link State Advertisement

- Each router periodically advertises its state
  - State is called link state
- In case the routers state changes
  - This is also advertised
- Each router constructs router link state advertisements (LSAs)
  - Router LSAs consist of information about its neighbors and links
  - Each LSA is flooded throughout the routing domain
    - The collected LSAs from all routers form the link state database

Link State Advertisement: Example

- Generic example

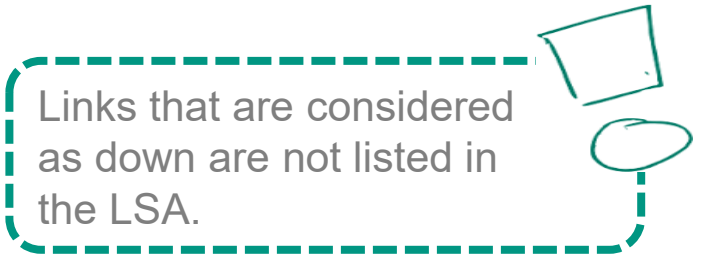
Advertising router R1
I_a Link state:
Neighbor R2
Link cost *x*
I_b Link state:
Neighbor R3
Link cost *y*
I_c Link state:
Neighbor R4
Link cost *z*



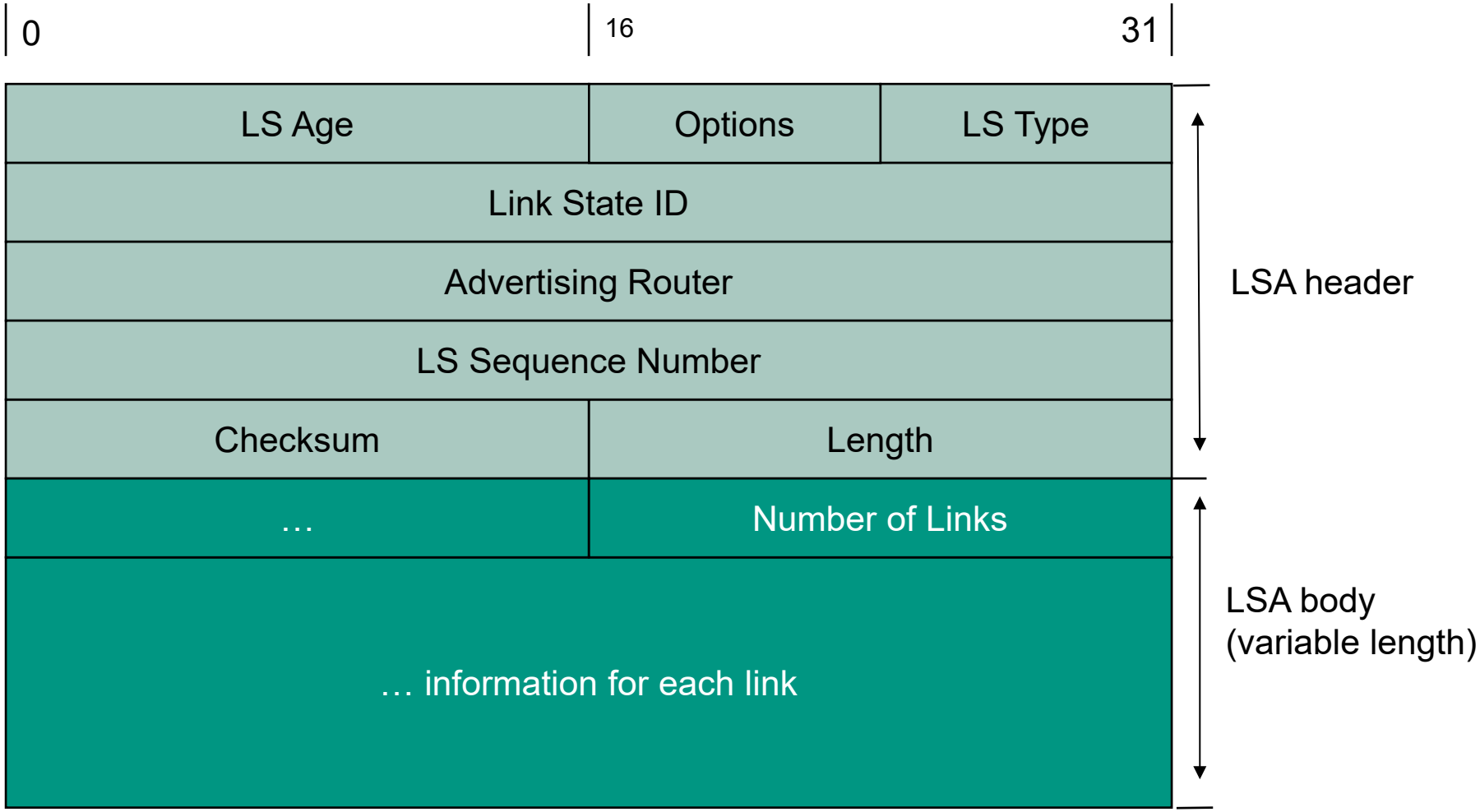
Each link state needs to be advertised separately

Structure of a Link State Advertisement

- LSA consists of a header and a body
- LSA header contains information used to uniquely identify the LSA
  - Advertising router
  - Sequence number of LSA at advertising router
  - …
- LSA body contains information of all operational links of the router
  - Associated cost
  - Type of link
    - To another router
    - To an end system
    - …
  - Reachability information
    - Router ID of another router
    - IP address of end system
  - …



Link State Update Message

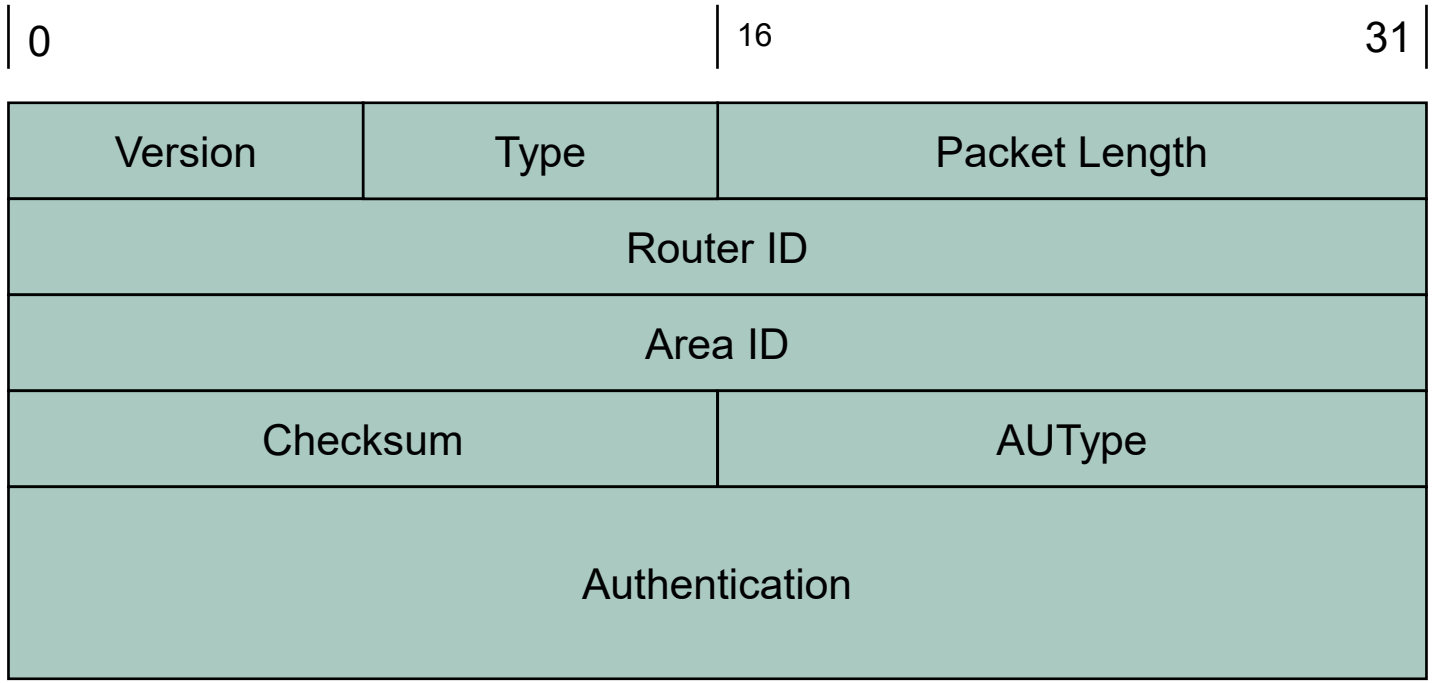


LSA Header

- LS Age
  - Time in seconds since LSA was originated
- Options
  - Optional capabilities supported by OSPF domain
- LS Type
  - Router LSA, network LSA …
- Link State ID
  - Identifies piece of routing domain that is described by the LSA
- Advertising Router
  - OSPF router ID of originating router
- LS Sequence Number
  - Incremented each time a new LSA is generated
- Checksum
  - Over entire message except age field
- Length
  - # bytes for entire LSA including header

Header of OSPF Messages

- All OSPF messages begin with a standard header



- Link state update is carried in the body of the OSPF message

[RFC2328]

Header of OSPF Messages

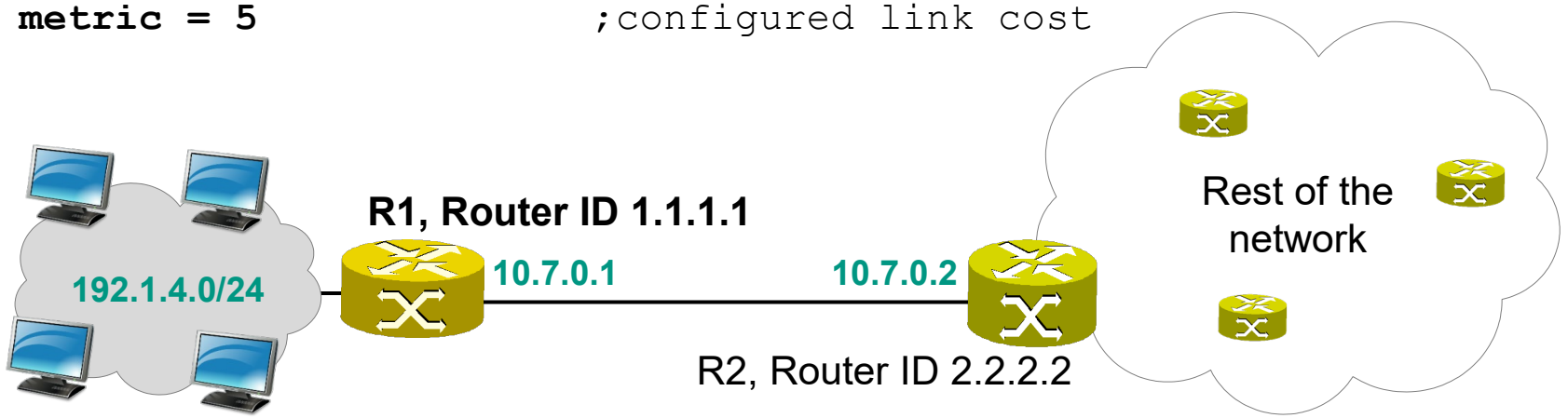
- Version
  - OSPF Version, currently 2 for IPv4 and 3 for IPv6
- Type
  - Hello
  - Database description
  - Link state request
  - Link state update
  - Link state acknowledgement
- Router ID
  - ID of originating router
    - IP address of one of its interfaces (e.g., smallest IP address)
- Area ID
  - OSPF area – explained later
- Checksum
  - Internet checksum over OSPF message
    - Excluding authentication field
- AUType and Authentication
  - Authentication of originating router

Example of a Link State Advertisement

- LSA of router R1 (excerpt)

```
Advertising router = 1.1.1.1

#links = 2
 Type = 1                  ;Point-to-point link to a router
 Link ID = 2.2.2.2         ;Neighbor's router ID
 Link data = 10.7.0.1      ;IP address of (this) router's interface
 metric = 3                ;configured link cost
 Type = 3                  ;Stub network (no other router is attached)
 Link ID = 192.1.4.0       ;IP network address
 Link data = 0xffffff00    ;Subnet mask
 metric = 5                ;configured link cost
```



Lifetime of Link State Advertisements

- Each LSA is associated with a lifetime (LS Age)
- Set to “0” by advertising router
  - When flooded, incremented by transmission delay (estimated value)
  - As LSA is stored in database, Age is incremented over time
- When LSA’s age reaches MaxAge, LSA is considered out-of-date
  - MaxAge is set to 1 hour
- Consequence: routers must refresh their LSAs every LSRefreshTime
  - LSRefreshTime is set to 30 minutes
  - Minimum value between generation of any particular LSA: 5 seconds

Issuing LSAs

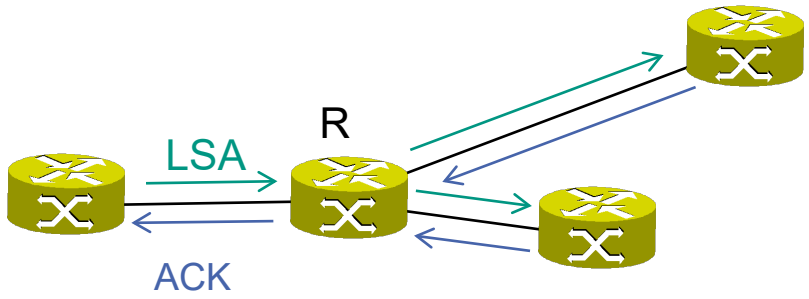
- If nothing changes (link, router), nothing needs to be reported with respect to routing
  - → keep quiet
  - LSAs are refreshed every 30 minutes

- Besides periodic refreshes, communication is only needed in case of changes, for example
  - Interface changed to up or down
  - Neighboring router on link is unreachable
  - Configuration changes

- Minimum time between two consecutive LSAs of a router is set to 5 seconds
  - Due to stability reasons

Handling Received LSAs

- Router R receives an LSA
  - Does a router store and forward the received LSA?



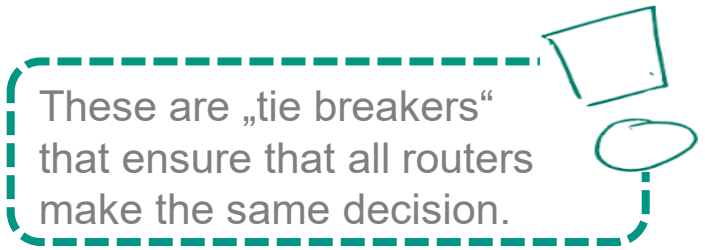
- Router R stores received LSA
  - If R does not have an LSA from the advertising router
  - If the received LSA is newer than the one in the link state database

Re-compute shortest paths if content of link state database changed

- If router R stores the LSA, it forwards it to its neighbors
  - Uses multicast address 224.0.0.5 with hop limit of 1

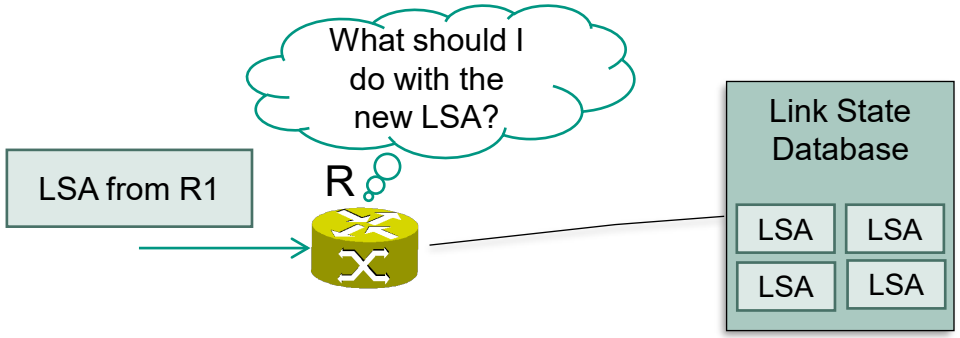
Newer LSA?

- An LSA from an advertising router is considered as newer and, thus, stored in link state database
  - LSA having higher sequence number is more recent
  - If both instances have the same sequence number
    - If both instances have different checksums
      - LSA with “larger” checksum is considered more recent
    - Else …



Reliable Flooding

- Example: router R receives LSA from advertising router R1



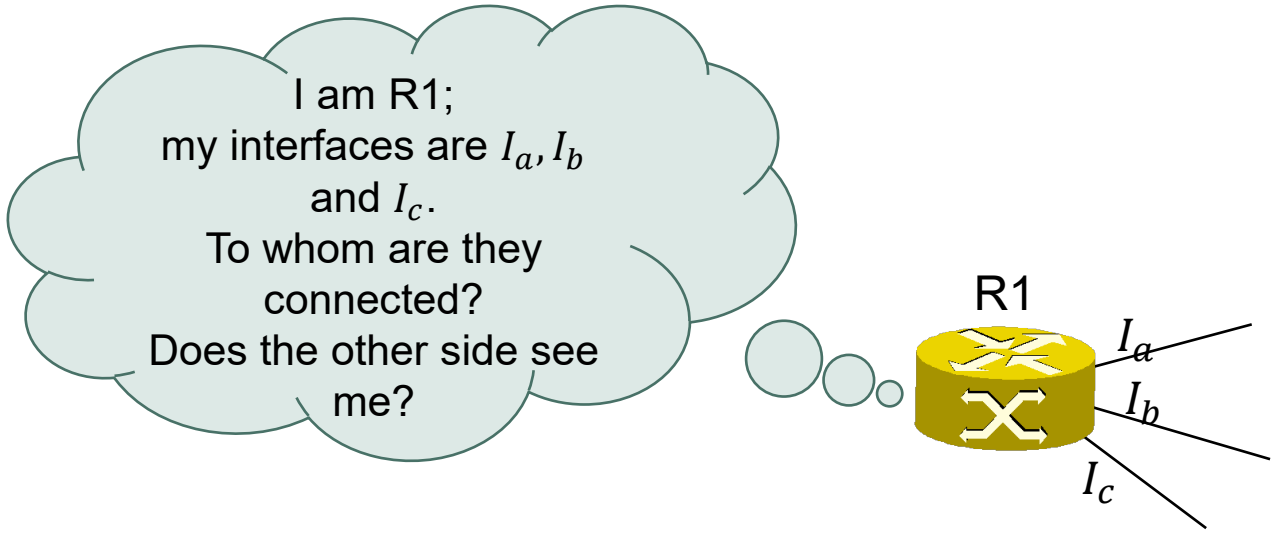
  - If $Age == MaxAge$ (of received LSA) and no LSA from R1 is known
    - Send ACK to sending neighbor and discard LSA
  - If there is no LSA from R1 in database or received LSA is newer
    - Store/replace LSA
    - Send ACK
    - Update Age and flood LSA to neighbors
  - If already stored copy is newer
    - Send stored copy back to advertising router R1
  - If LSA and stored copy are identical
    - Discard LSA

We abstracted here from some details

3.4.3 Bringing Up Adjacencies

Hello Protocol

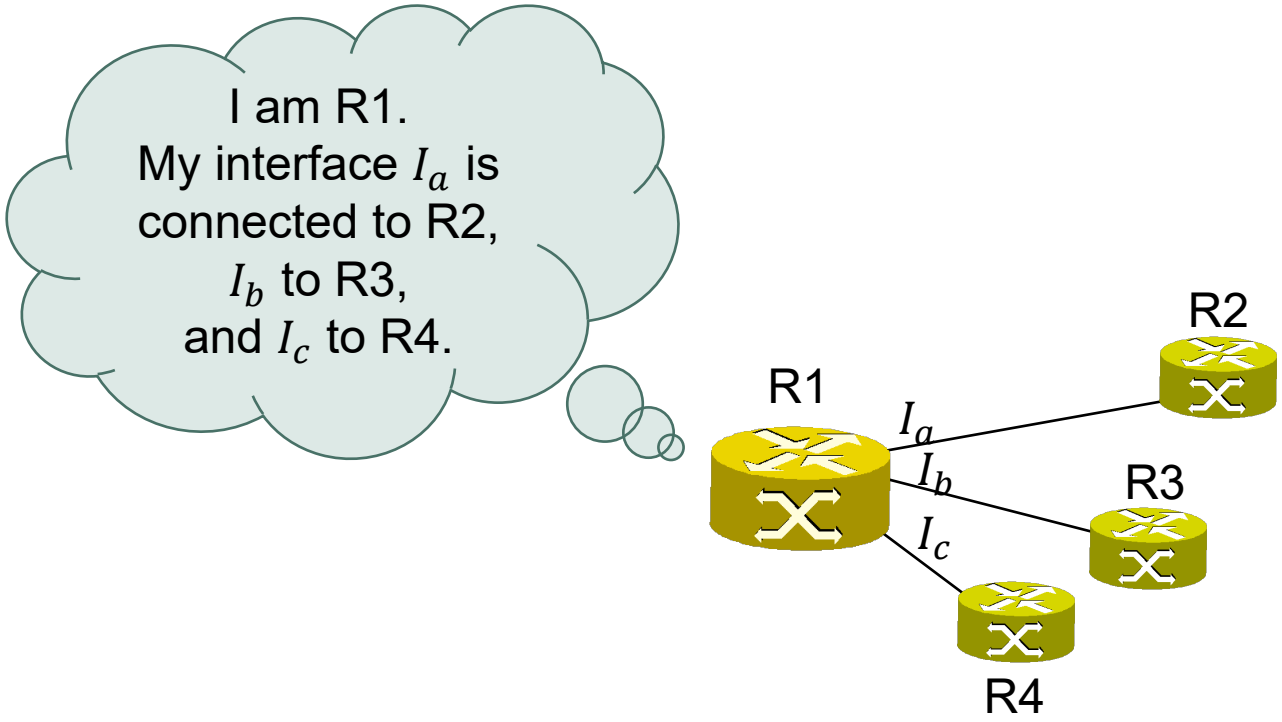
- Goals
  - Establish and maintain neighbor relationships
  - Ensure bi-directional communication between neighboring OSPF routers



- Hello protocol determines identity and liveliness of neighboring routers

Hello Protocol

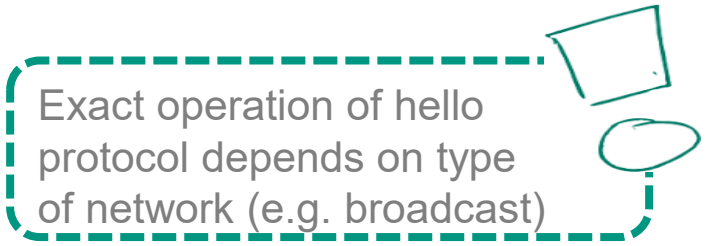
- Each router learns its neighbors and monitors the state of the links to them



Hello Message

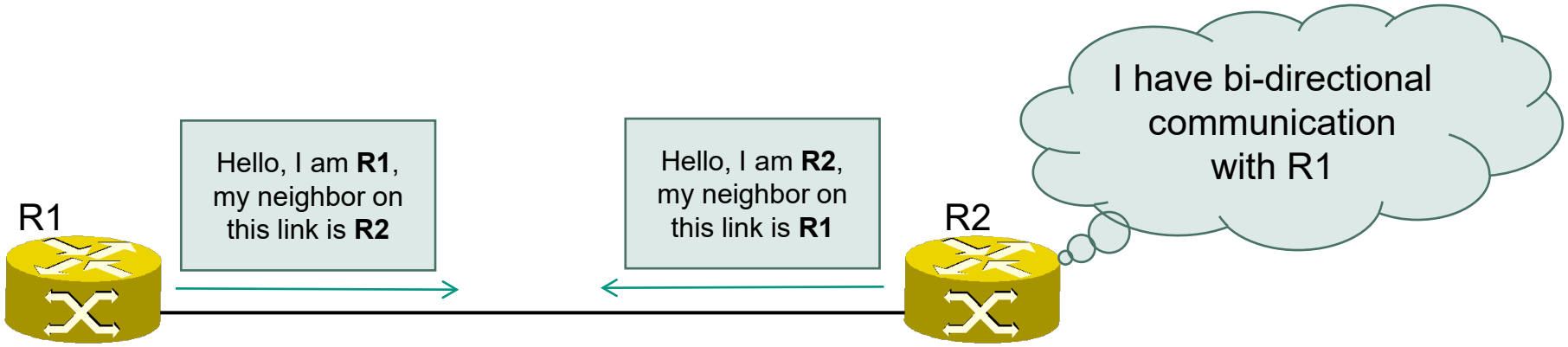
- Contains own $router\ ID$
- Contains $router\ ID$ of neighboring router, if known
  - If not yet known, $router\ ID$ is set to 0.0.0.0
  - If own $router\ ID$ is contained in neighbor's hello message, communication is considered to be bi-directional
- Destination IP address of hello message
  - 224.0.0.5 (multicast address, "AllSPFRouters")

  → hello message is received and processed only by OSPF routers



Simplified Workflow

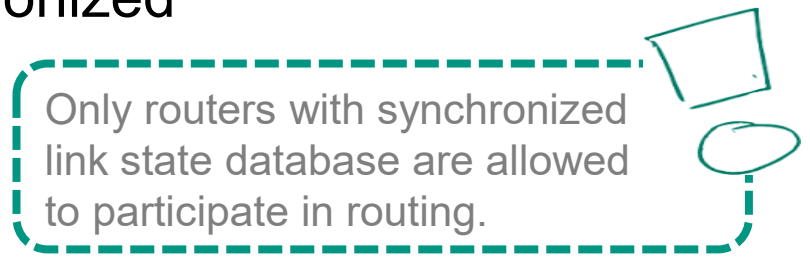
- A router periodically sends a hello message on all its links
  - “Hello, I am R1, I am still here”
  - If known, hello message contains $router\ ID$ of neighboring router
    - “my neighbor on this link is R2”



- If no hello message is received for a pre-defined period of time the link is considered to be down
  - Standard value for periodic hello messages: every 10-30 seconds
  - Fast hello extension: < 1 second

Synchronization of Link State Databases

- Link state databases must be synchronized
  - OSPF (only) requires adjancent routers to remain synchronized



- Initial database synchronization
  - Router asks neighboring router to share its database
  - Performed immediately after a “handshake” of the hello protocol
  - Routers exchange LSA headers with each other
  - If an LSA is missing it is requested from the neighbor router

→ the routers are now considered as adjacent to each other

Designated Router

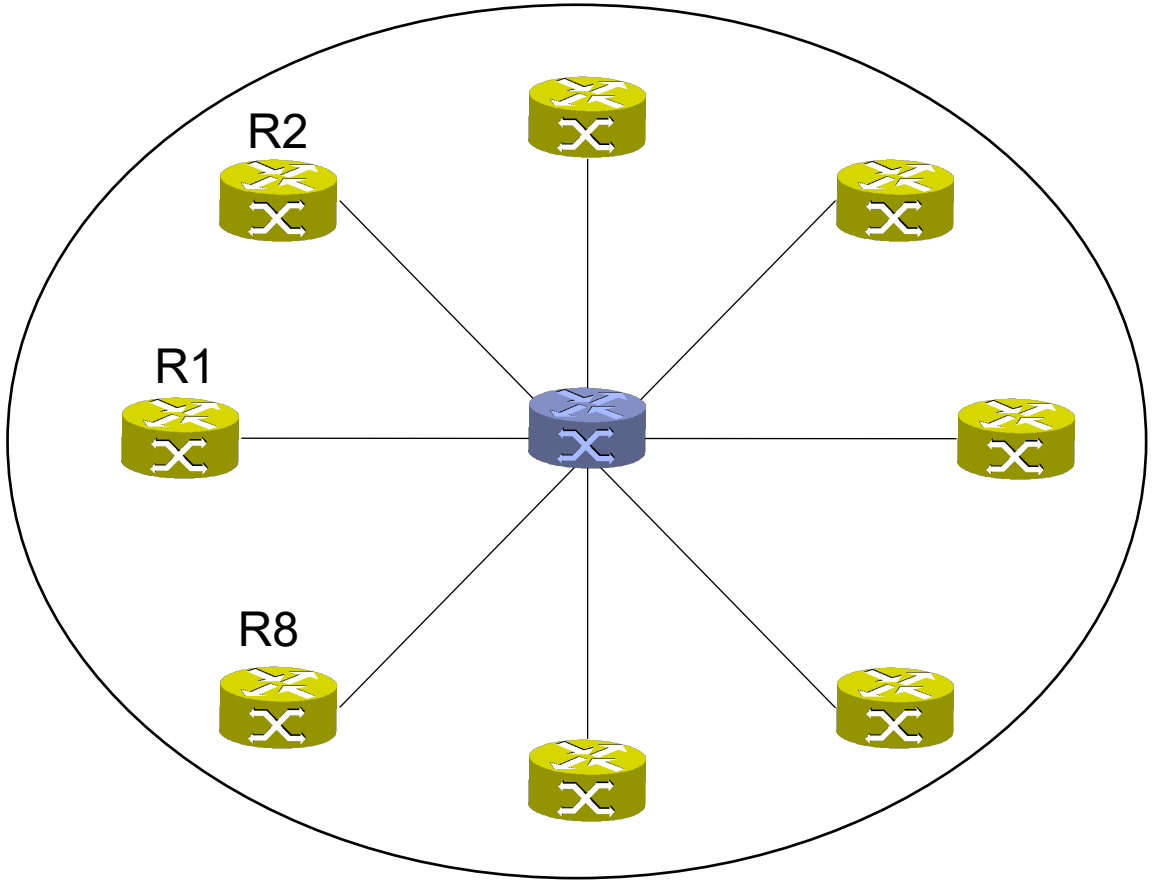
- Every broadcast network or NBMA has a Designated Router

- Designated Router
  - Originates network LSA on behalf of network
    - Lists set of routers currently attached to network
  - Becomes adjacent to all other routers in the network
    - Is endpoint of many adjacencies
    - Multicasts link state update to reduce number of packets
  - Backup Designated Router takes over if Designated Router crashes

NBMA: Non-broadcast-multi-access

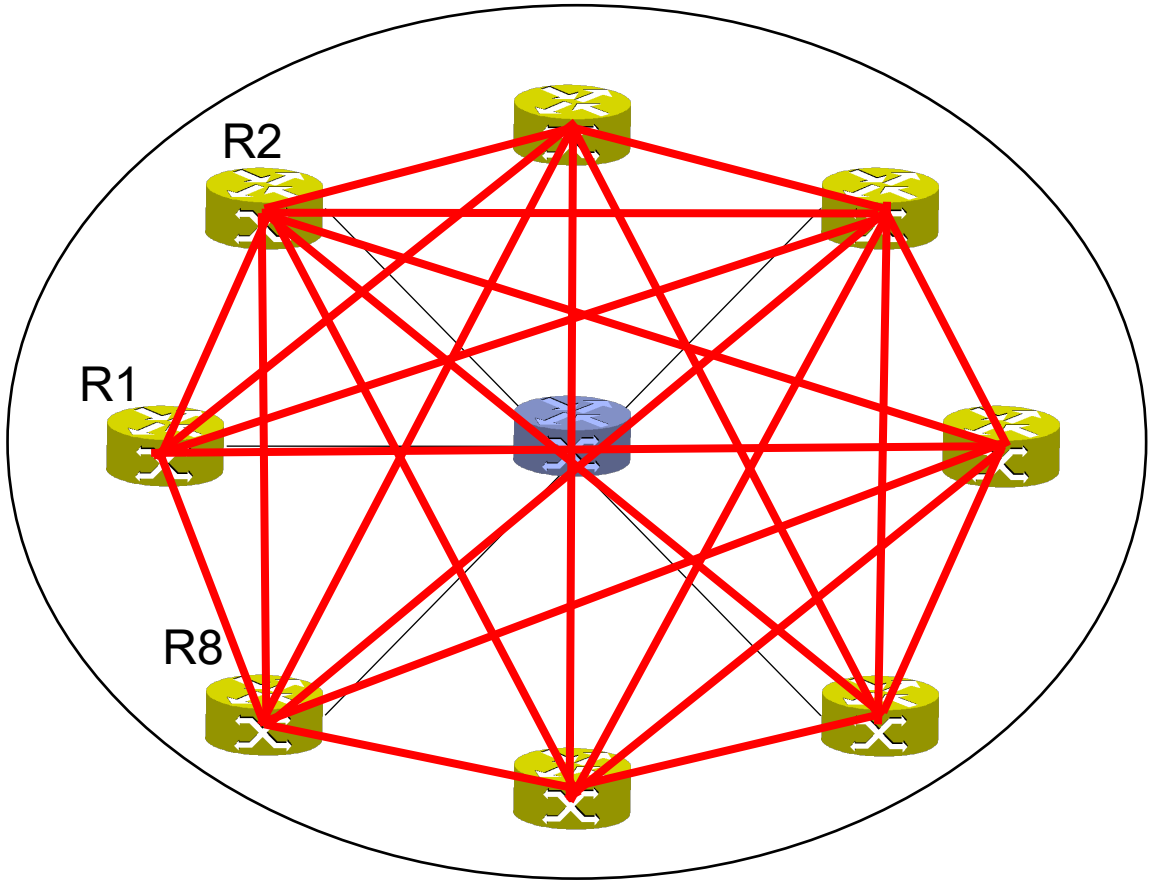
Example Network

- Non-broadcast multi-access network
- Network consists of 8 routers
  - R1 … R8
  - These routers are all connected to the same switch (star topology)



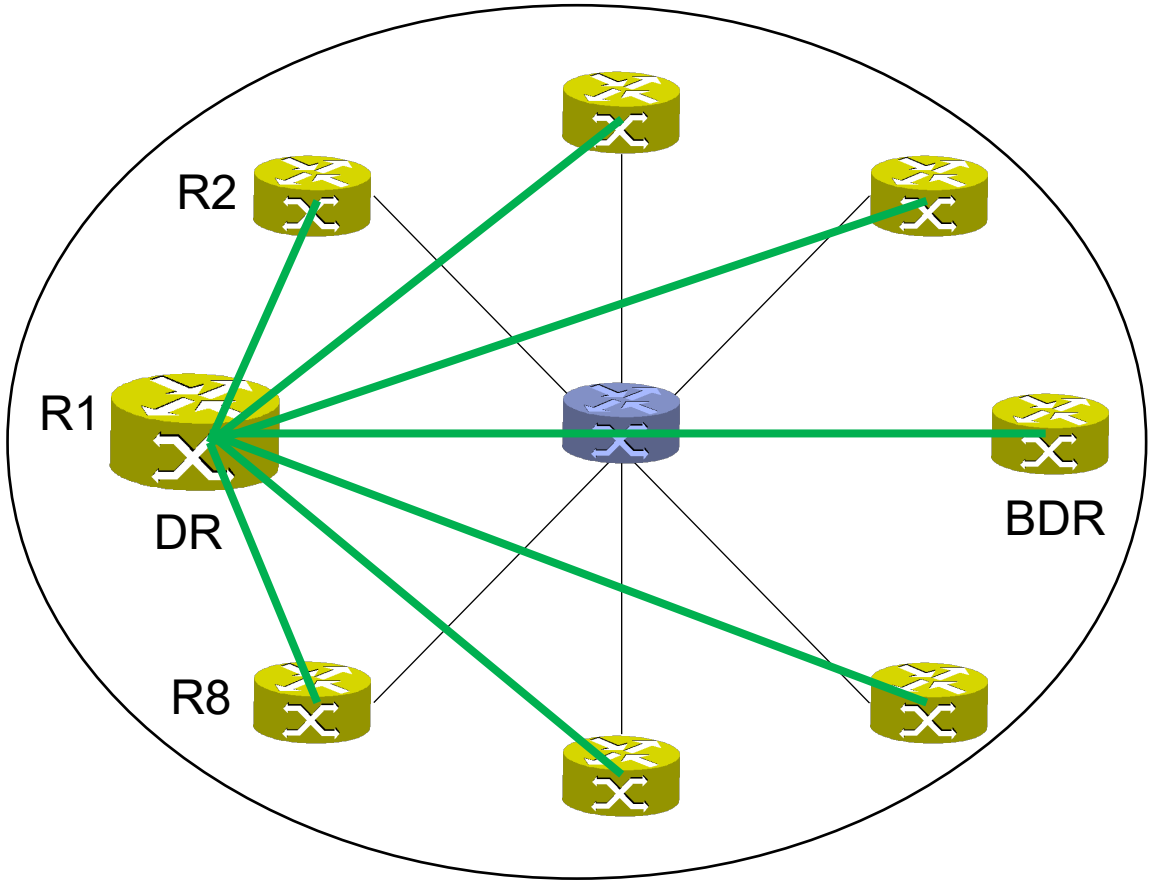
Example Network

- Neighbor discovery
  - Each router has 7 neighbors
  - → Full mesh of neighbors
- LSA Flooding
  - Each router floods LSAs to all other routers
  - → High OSPF traffic overhead



Example Network

- Establish a Designated Router (DR)
- Established adjacency
  - All routers only adjacent to DR
  - DR distributes LSAs
- Backup DR (BDR)
  takes over if DR crashes

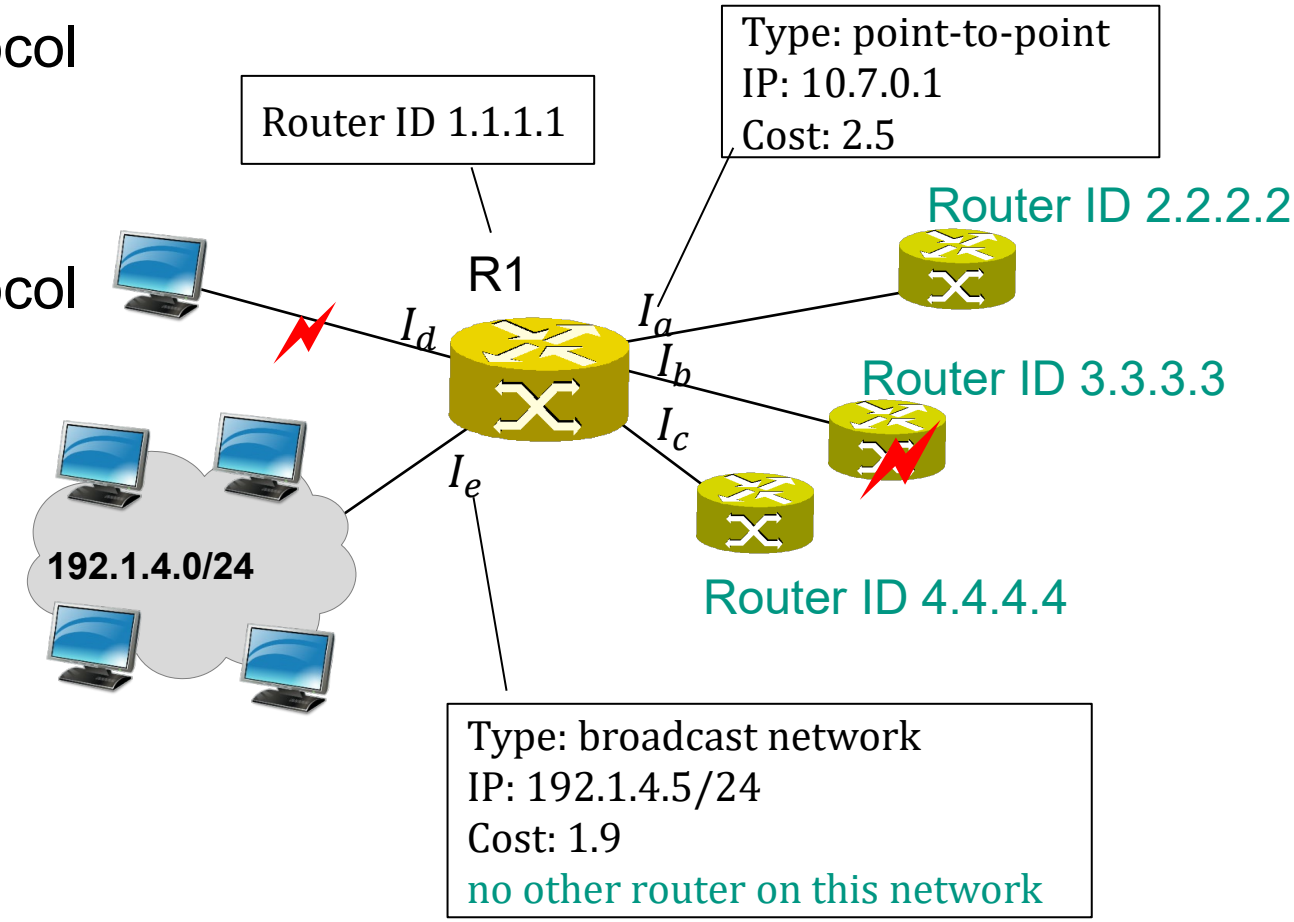


Pre-Configuration of OSPF Router

- Each router is pre-configured with the following parameters
  - Router ID – unique ID of a router in the network
    - E.g., smallest IP address of its interfaces
  - Per-interface parameters
    - Interface IP address (and mask)
    - Interface output cost – metric
      - Typically, inversely proportional to link data rate

Link States of a Router

- Router ID of neighbors
  - dynamically discovered by hello protocol
- Availability (↯)
  - dynamically discovered by hello protocol or physical layer
- Everything else is configured



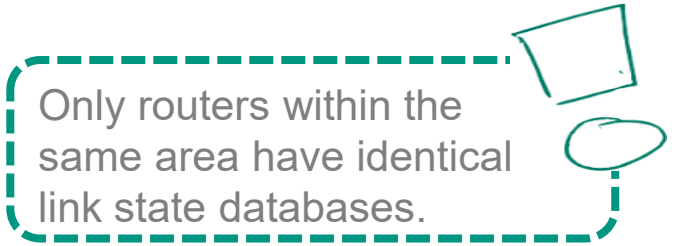
3.4.4 OSPF Areas

OSPF Areas

- Autonomous systems can grow rather large
  - Scalability problem
    - LSA flooding and
    - Route computation overhead

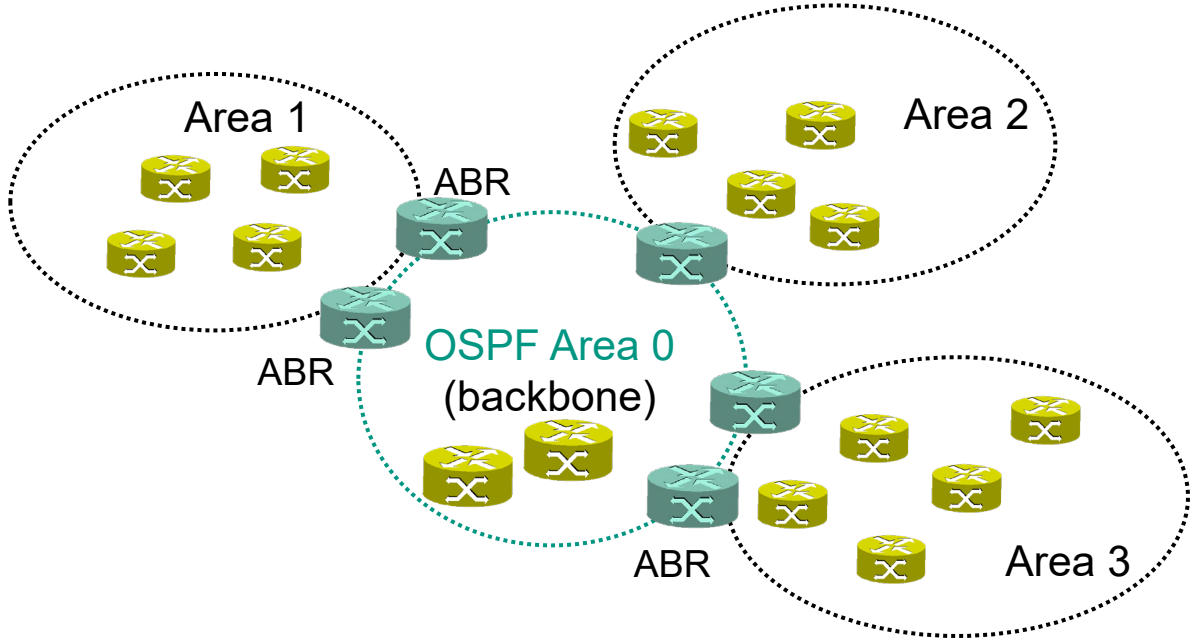
  do not scale

- Basic concept
  - Divide an AS into areas
    - Apply routing as discussed only within an area (intra-area-routing)
      - LSA flooding and route computation limited to an area
      - Topology of an area is not visible outside the area
    - Areas exchange summary information with each other
      - Addresses reachability from these areas
    - Typical size of an area: less than about 100 routers



Backbone of the Autonomous System

- OSPF Area 0 – backbone of the autonomous system
  - Area 0.0.0.0
  - Must be always connected
    - Virtual links may be used
  - Contains all ABRs of the AS
- All other areas …
  - are directly connected to backbone
- Area border routers (ABRs) interconnect areas
  - They belong to both: their area and the backbone
  - They run an instance of OSPF for each area they are connected to



Area border

Backbone border

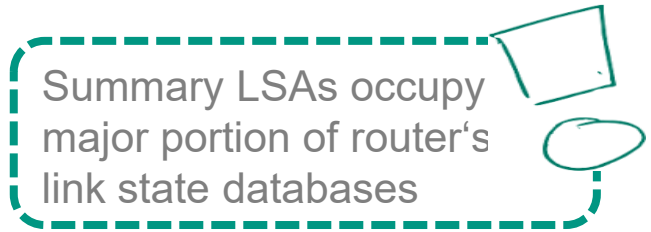
ABR

Propagation of Route Information across Areas

- Area border routers generate summary LSAs
  - Contain ABR's routing table for corresponding area
    - List of destinations reachable within the area
    - Associated with path cost from the ABR to destination
  - ABR´s routing table is constructed after intra-area path computation

- Handling of summary LSAs
  - Same way as "regular" LSAs

  - ABR forward summary LSAs of an area into backbone
  - ABR forward summary LSAs from backbone into an area



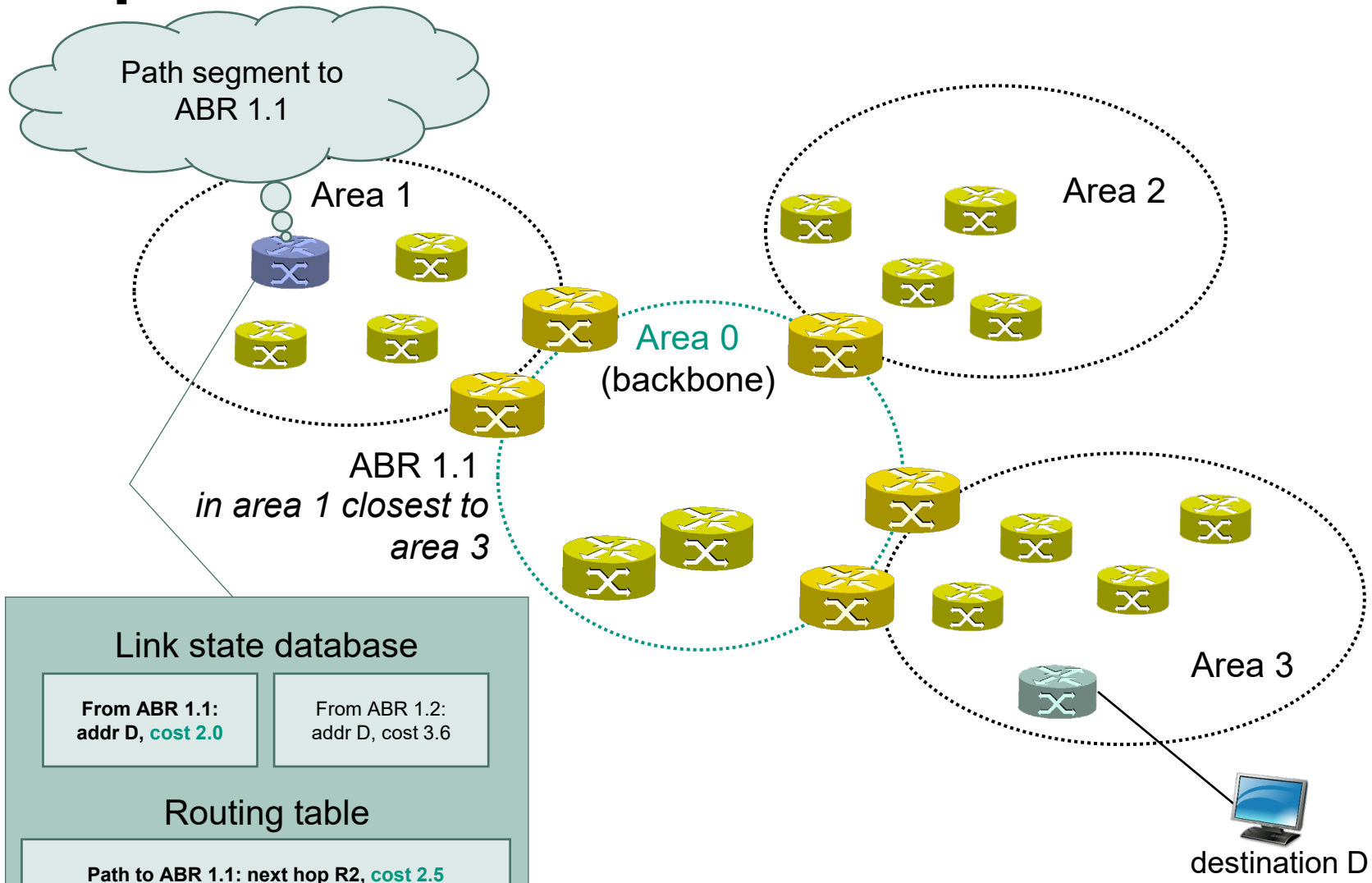
Inter-Area Forwarding

- Data between areas are forwarded through backbone (Area 0)

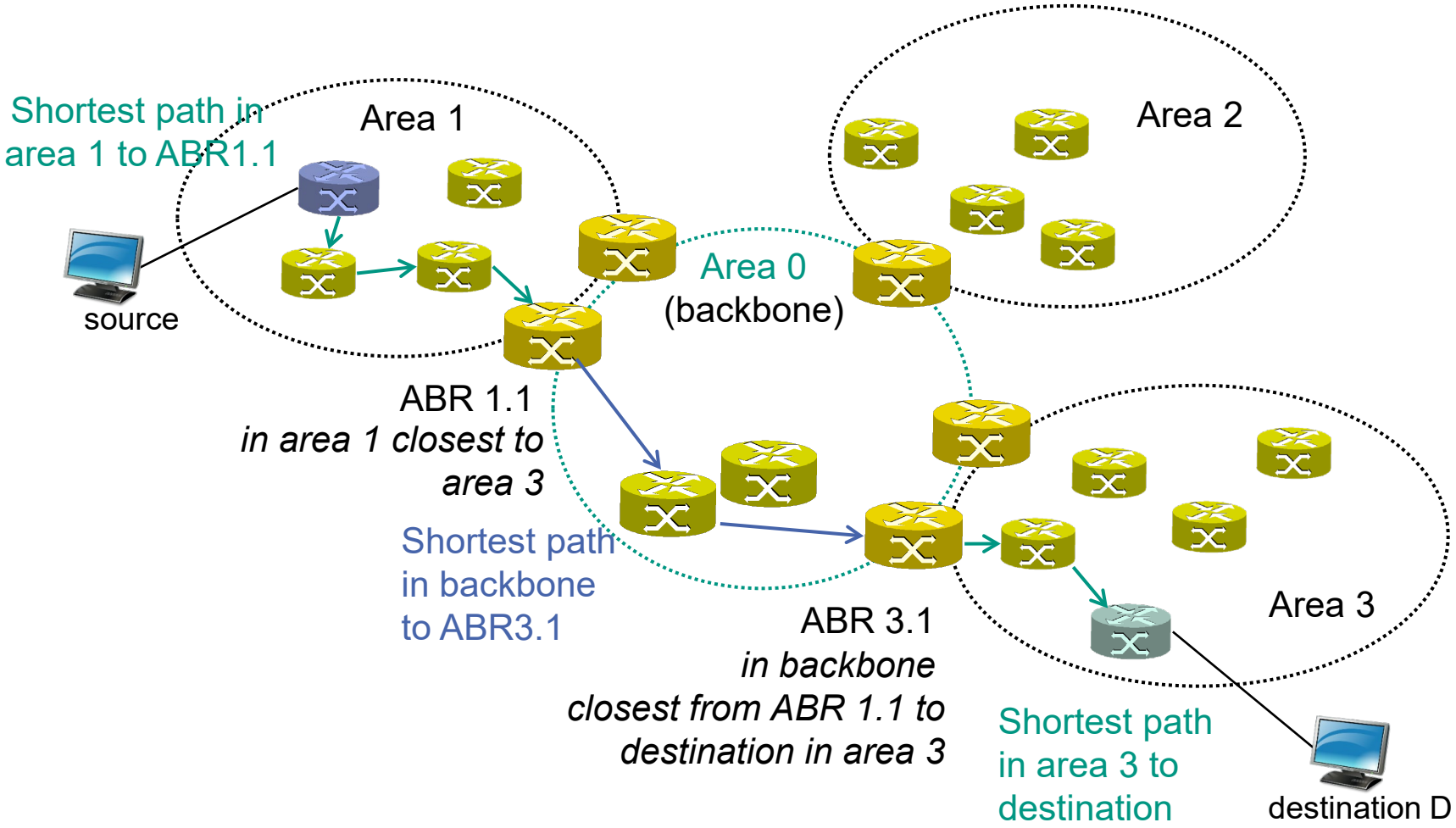
- End-to-end path consists of path segments
  - Segment between source and ABR of originating area
  - Segment between two ABRs in area 0, and
  - Segment between ABR of target area and destination

- Routers within an area select ABRs so that resulting end-to-end path is a shortest path
  - Based on path costs of ABRs

Example: Inter-Area Route Calculation

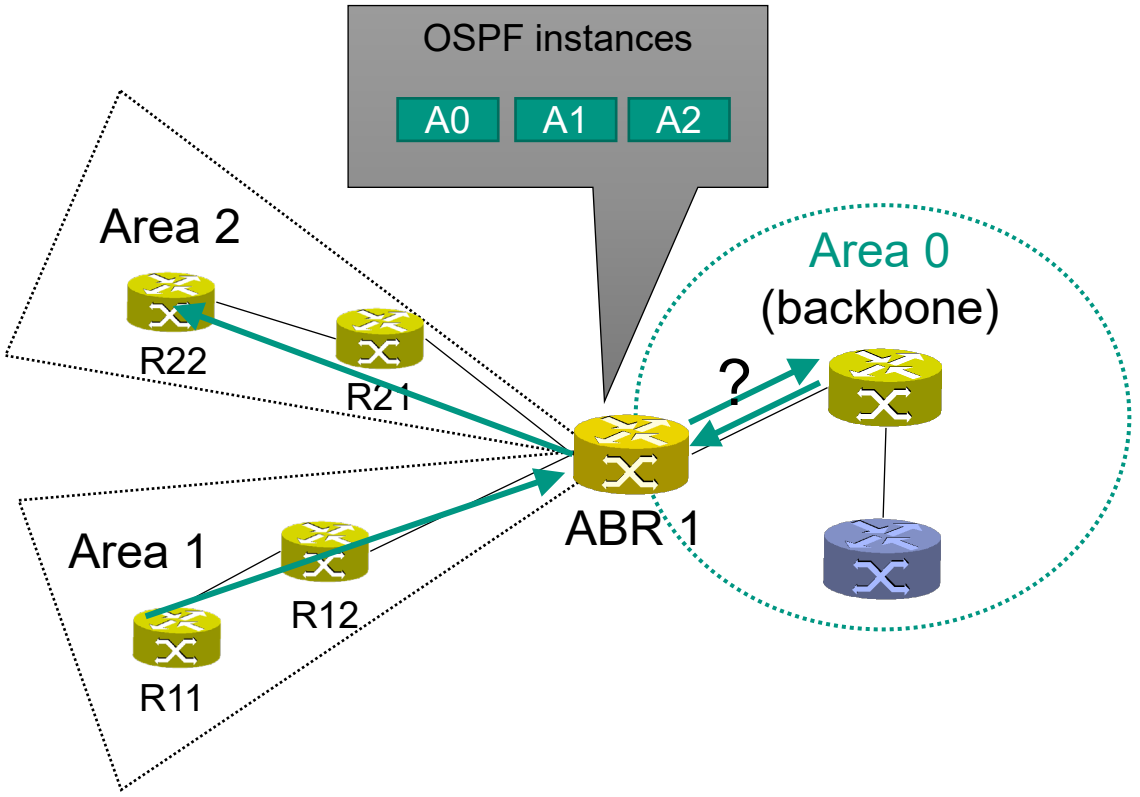


Example: Inter-Area Forwarding



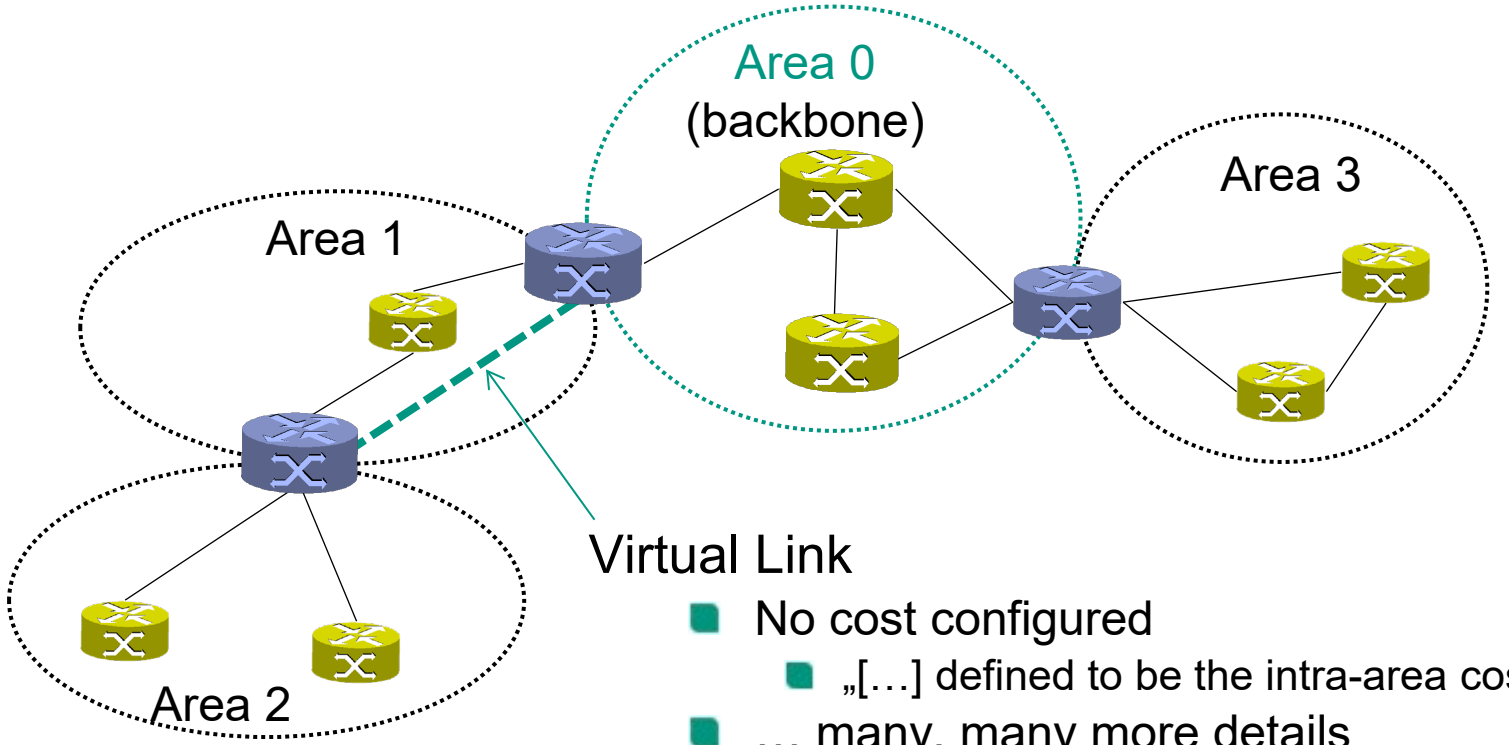
Multiple Areas Attached to Single ABR

- Which route from *R11* to *R22*?
- What happens at ABR 1?
  - Into the backbone?
- No. ABR1 is part of Area 2
  - It has an intra-area route to R22
  - Sends packet directly to Area 2



Virtual Links

- RFC 2328, section 15
  - „[…] virtual links can be configured through non-backbone areas.“
  - „The virtual link is […] belonging to the backbone […] joining two ABRs.“



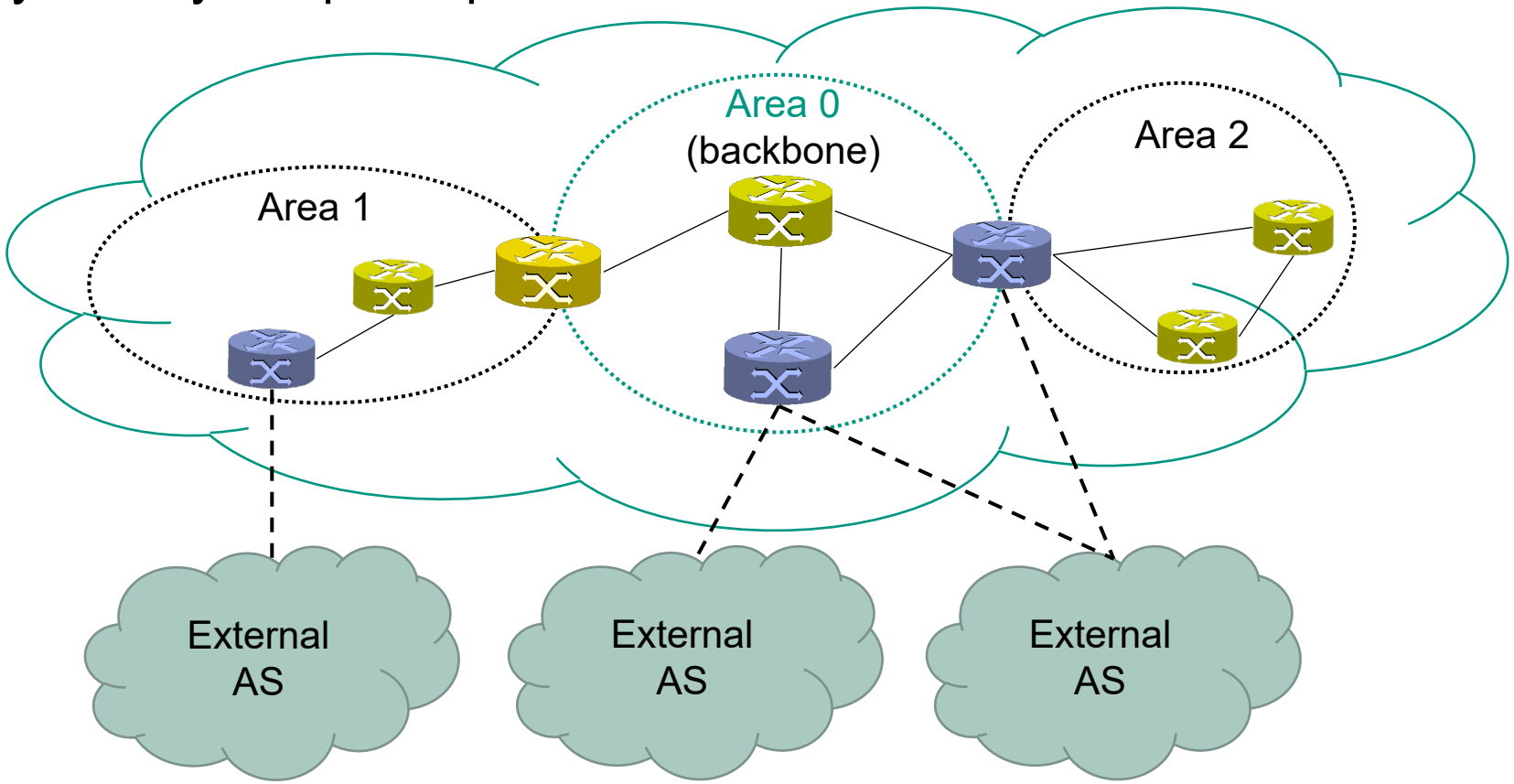
- Virtual Link
  - No cost configured
    - „[…] defined to be the intra-area cost […]“
  - … many, many more details

Classification of Routers in OSPF

- Internal routers
  - Router with all interfaces belong to same network
  - Single routing instance
- Area border routers (ABRs)
  - Attached to multiple areas
  - Multiple routing instances
    - One for each attached area
- Backbone routers
  - Router that has an interface to backbone area
  - Includes all ABRs
- AS boundary routers (ASBRs)
  - Exchange routing information with routers belonging to other autonomous systems
  - Path to AS boundary routers are known by every router in the AS

Location of AS Boundary Routers

- AS boundary routers
  - may be internal or area border routers
  - may or may not participate in the backbone



3.4.5 Types of LSAs

Different Types of LSAs

- Router LSA (Type 1)
  - All links a router knows and the associated cost for using it
  - Stays within the local area
  - Used to calculate intra-area part of the routing table
- Network LSA (Type 2)
  - List of all routers in a broadcast-capable network
  - Only sent by designated router
  - Stays within the local area
  - Also used to calculate intra-area part of the routing table
- Routing table calculation: See Section 16 in RFC 2328

LS Type	LSA description
1	These are the router-LSAs. They describe the collected states of the router's interfaces. For more information, consult Section 12.4.1.
2	These are the network-LSAs. They describe the set of routers attached to the network. For more information, consult Section 12.4.2.
3 or 4	These are the summary-LSAs. They describe inter-area routes, and enable the condensation of routing information at area borders. Originated by area border routers, the Type 3 summary-LSAs describe routes to networks while the Type 4 summary-LSAs describe routes to AS boundary routers.
5	These are the AS-external-LSAs. Originated by AS boundary routers, they describe routes to destinations external to the Autonomous System. A default route for the Autonomous System can also be described by an AS-external-LSA.

Different Types of LSAs

- Summary LSA (Types 3 and 4)
  - Only used in case of multiple OSPF areas
  - Required to calculate inter-area part of routing table
  - Type 3 = routes to other areas
  - Type 4 = preferred route to AS boundary router(s)
    - Not required for stub areas
    - Can use default routes

LS Type	LSA description
1	These are the router-LSAs. They describe the collected states of the router's interfaces. For more information, consult Section 12.4.1.
2	These are the network-LSAs. They describe the set of routers attached to the network. For more information, consult Section 12.4.2.
3 or 4	These are the summary-LSAs. They describe inter-area routes, and enable the condensation of routing information at area borders. Originated by area border routers, the Type 3 summary-LSAs describe routes to networks while the Type 4 summary-LSAs describe routes to AS boundary routers.
5	These are the AS-external-LSAs. Originated by AS boundary routers, they describe routes to destinations external to the Autonomous System. A default route for the Autonomous System can also be described by an AS-external-LSA.

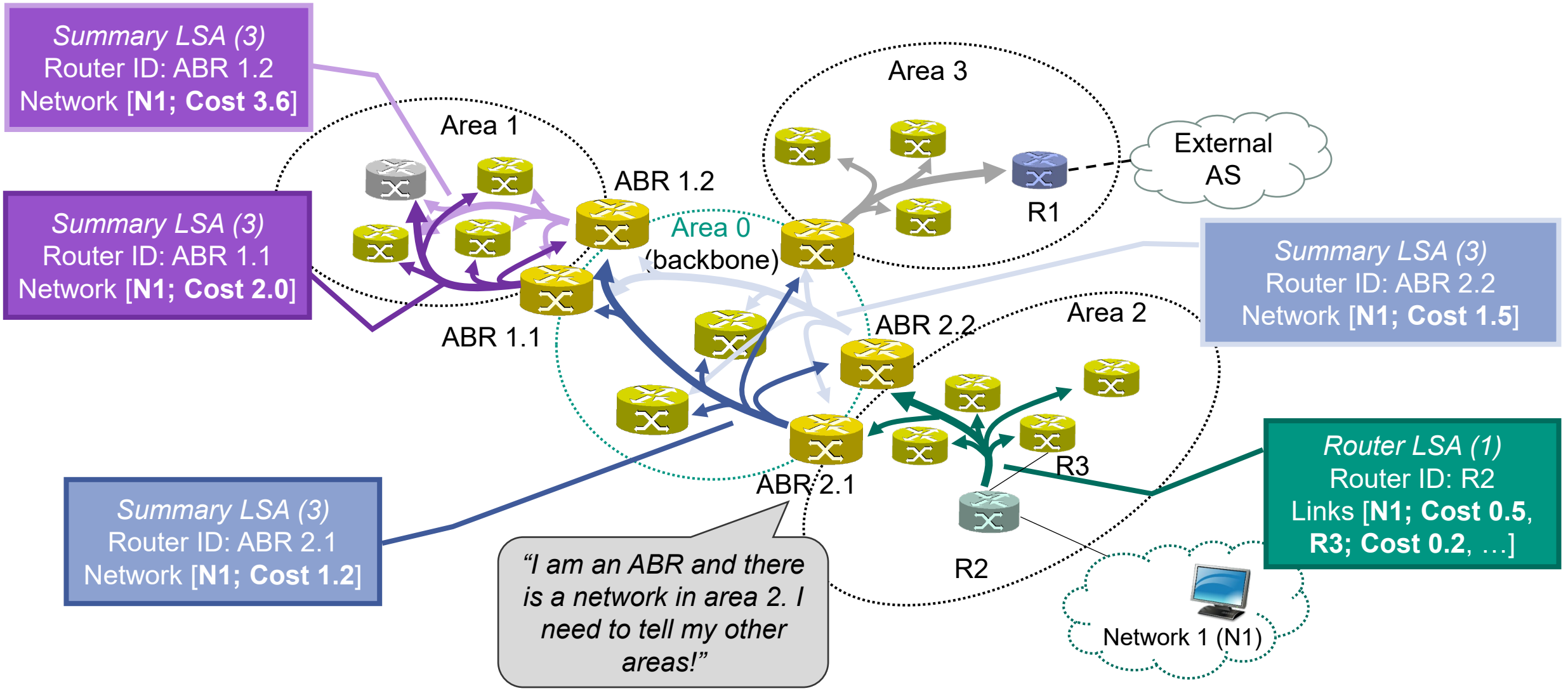
Different Types of LSAs

- AS-external LSA (Type 5)
  - Announce AS external routes
  - Sent by AS boundary routers
  - Flooded to all routers in the AS not just within the area
  - Note: explicit AS external routes are never advertised in summary-LSAs

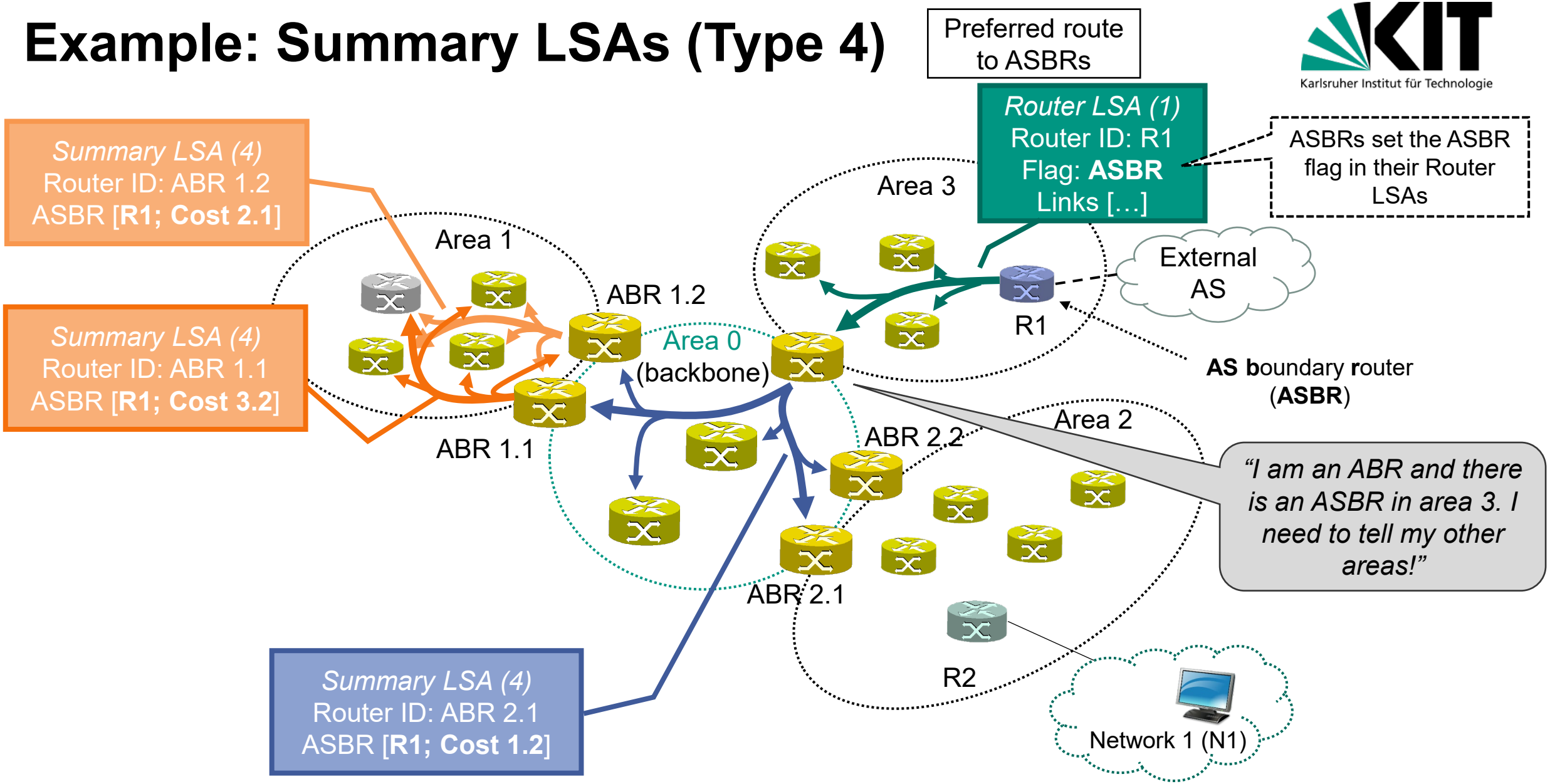
LS Type	LSA description
1	These are the router-LSAs. They describe the collected states of the router's interfaces. For more information, consult Section 12.4.1.
2	These are the network-LSAs. They describe the set of routers attached to the network. For more information, consult Section 12.4.2.
3 or 4	These are the summary-LSAs. They describe inter-area routes, and enable the condensation of routing information at area borders. Originated by area border routers, the Type 3 summary-LSAs describe routes to networks while the Type 4 summary-LSAs describe routes to AS boundary routers.
5	These are the AS-external-LSAs. Originated by AS boundary routers, they describe routes to destinations external to the Autonomous System. A default route for the Autonomous System can also be described by an AS-external-LSA.

Example: Summary LSAs (Type 3)

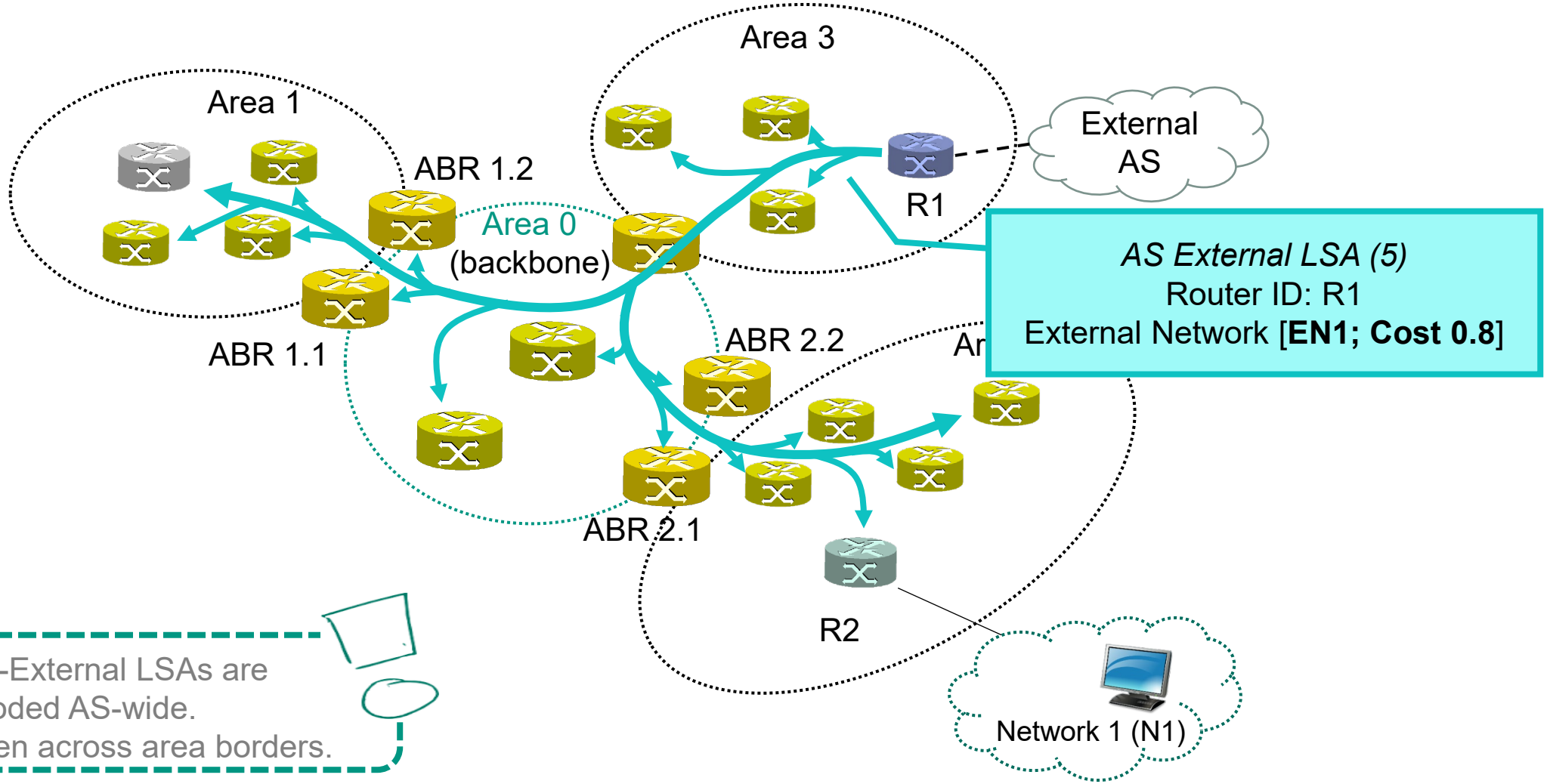
Routes to networks in other areas



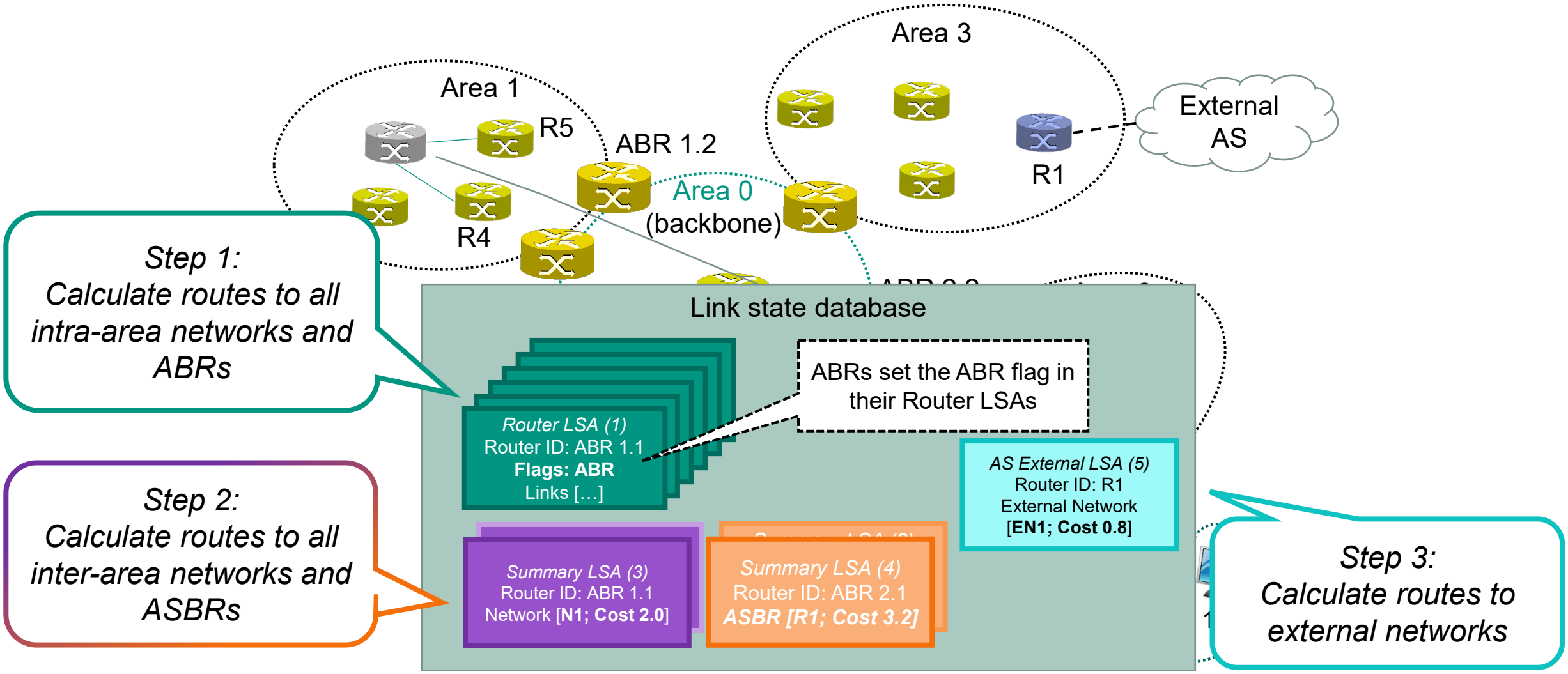
Example: Summary LSAs (Type 4)



Example: AS-External LSA (Type 5)



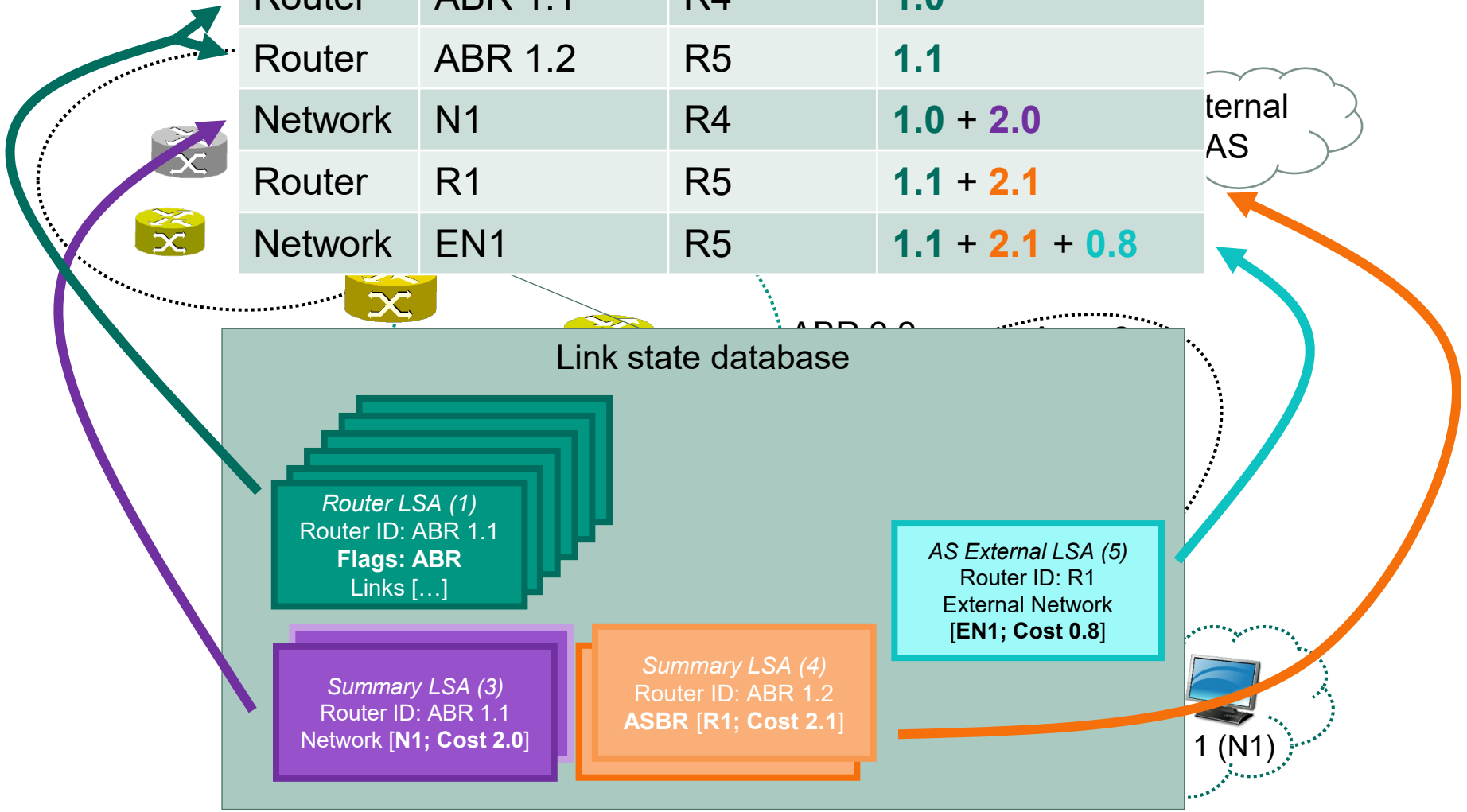
Example: Calculating Routing Tables



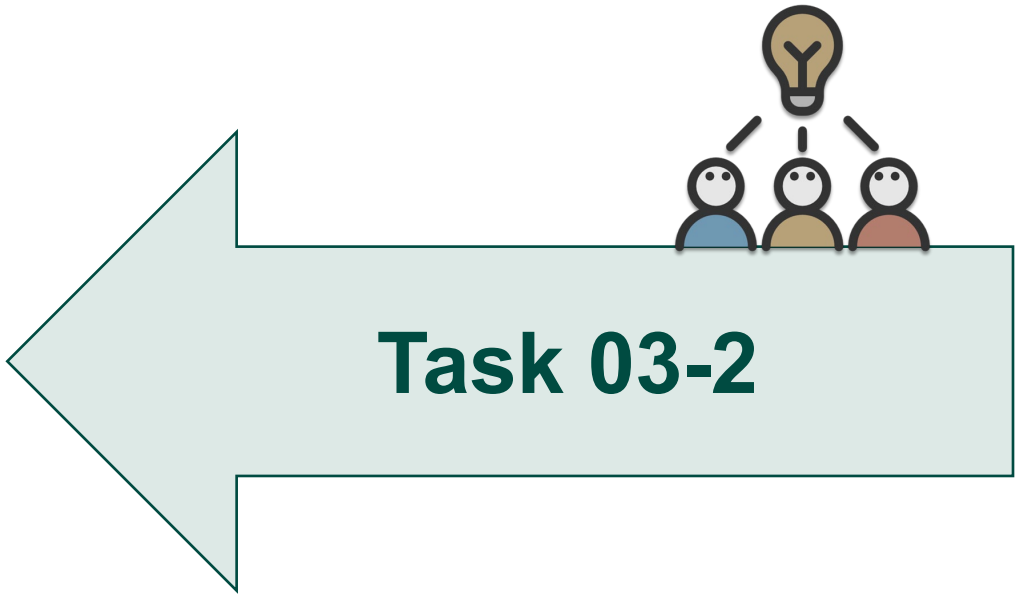
Example: Calculating Routing Tables

1. build graph
2. calculate shortest paths (Dijkstra)

Type	Destination	Next Hop	Cost
Router	ABR 1.1	R4	1.0
Router	ABR 1.2	R5	1.1
Network	N1	R4	1.0 + 2.0
Router	R1	R5	1.1 + 2.1
Network	EN1	R5	1.1 + 2.1 + 0.8




Task 03-2
OSPF
trace


Task 03-2

3.4.8 RIP vs. OSPF

RIP vs. OSPF

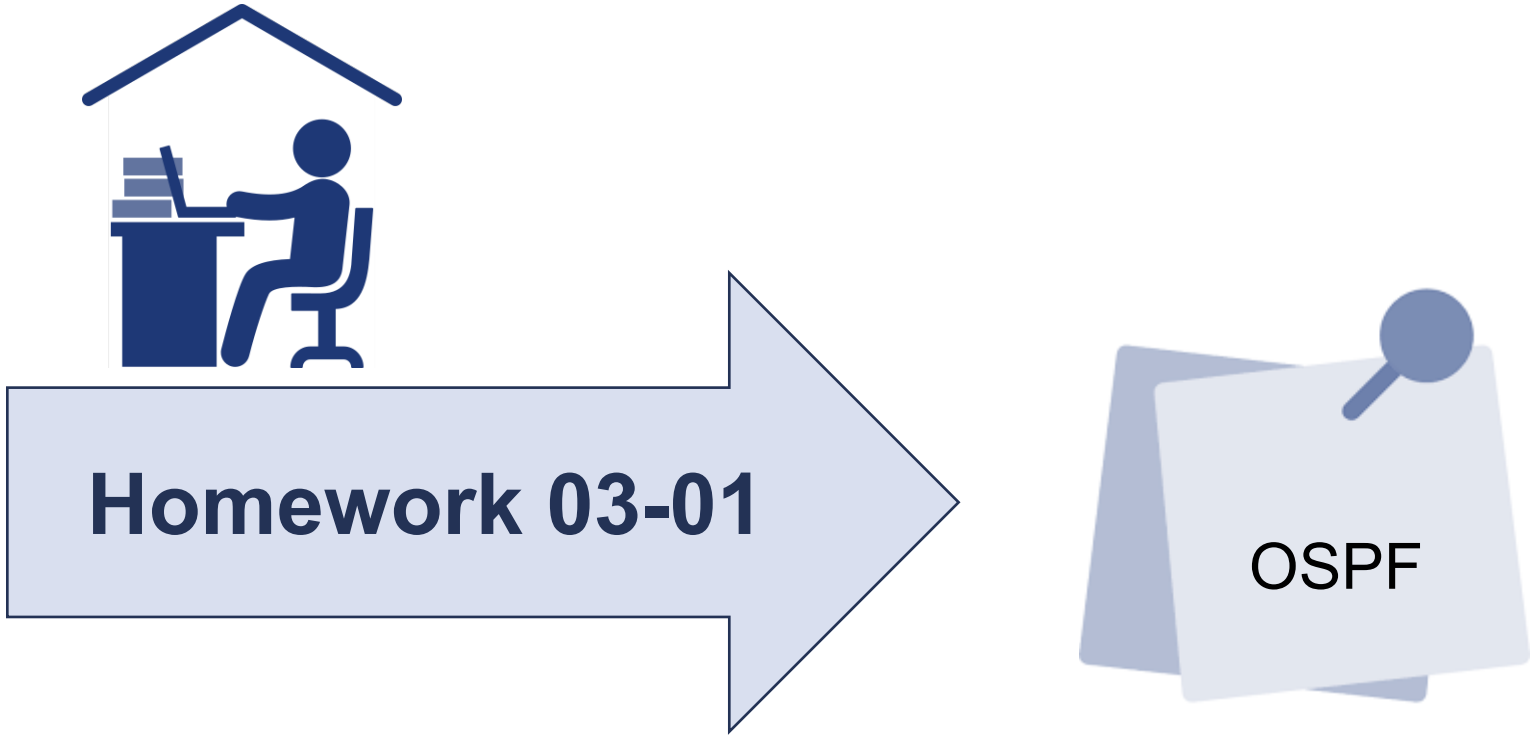
- RIP
  - Limited in metric selection and size
    - Only one metric (hop count)
    - Maximum path length of 15 hops
  - Periodic updates every 30 seconds, even without changes
  - Slow convergence, count-to-infinity
    - → Not suitable for large networks
  - But: easier and requires less resources than OSPF
  - Still sometimes used in small networks

distance vector

- OSPF
  - Addresses shortcomings of RIP
    - Faster convergence, no count-to-infinity, lower signaling overhead …
  - Large networks can be divided into areas
  - Standard in large ASes (together with IS-IS)

link-state

Homework



Homework

Homework 03-02

3.5 Additional Issues

Observation

- Shortest path algorithms select shortest path !

  - Selection depends on routing metric chosen
    - Finding suited metric is non-trivial

  - All data towards a destination (i.e., IP address) follow this shortest path
    - Traffic engineering (e.g., load balancing) not directly possible

    - Several possibilities in order to support traffic engineering exist, e.g.,
      - ECMP: Equal cost multipath … utilize multiple shortest paths
      - Opaque LSA: carries information related to traffic engineering

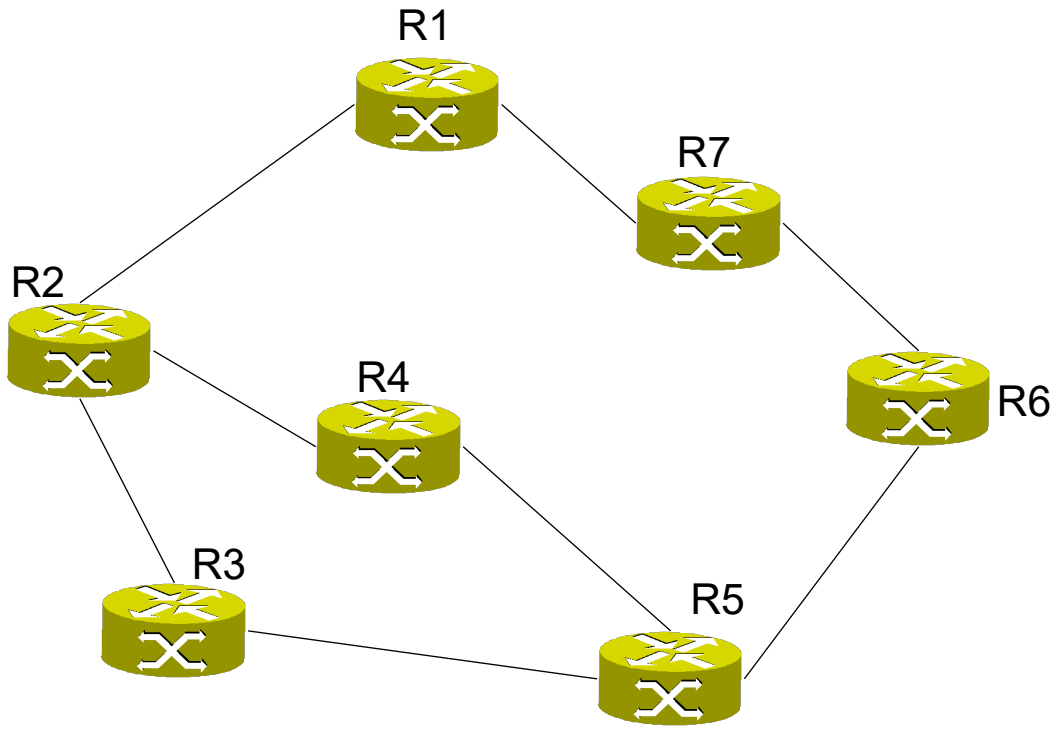
3.5.1 ECMP: Equal Cost Multipath

Multiple Paths with Lowest Cost

- Dijkstra´s algorithm generates shortest paths
  - Multiple paths with lowest cost may exist
  - Example
    - Link weights on all links: 1
    - Same cost for paths from R1 to R5
      - R1 – R2 – R3 – R5
      - R1 – R2 – R4 – R5
      - R1 – R7 – R6 – R5

- Equal-cost multi-path routing (ECMP)
  - Traffic from R1 to R5 can be equally split towards R2 and R7
  - R2 can, again, equally split on R3 and R4

  → Allows for load balancing



OSPF supports ECMP

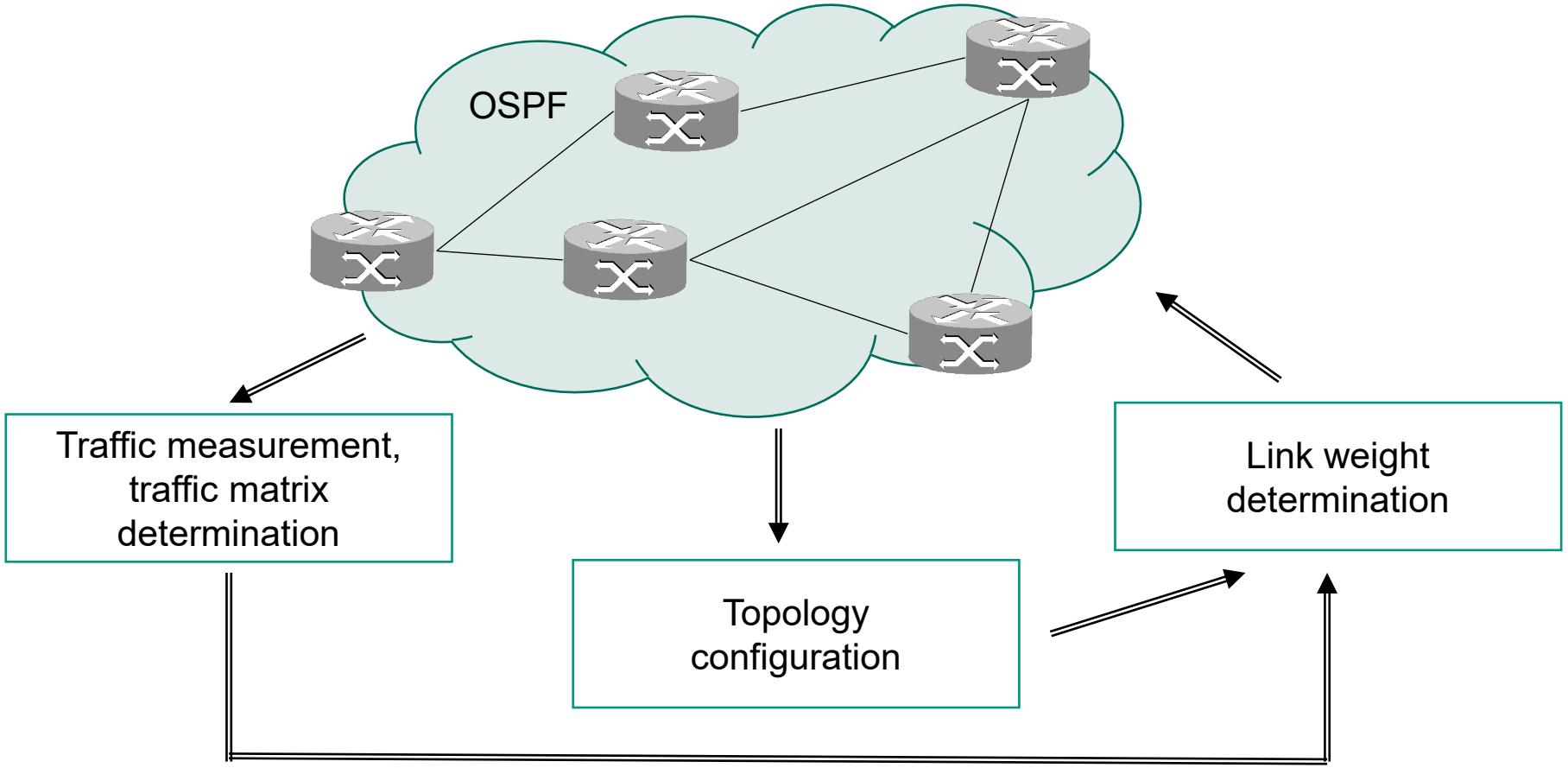
3.5.2 Traffic Engineering

Traffic Engineering

- Goal
  - Performance optimization of operational networks
    - Performance requirements are met
    - Network resources are well utilized
- Main task
  - Determine proper link weights
- Traffic engineering
  - Addresses medium term goals and overall behavior of operational networks

[Medh17]

Architectural Framework

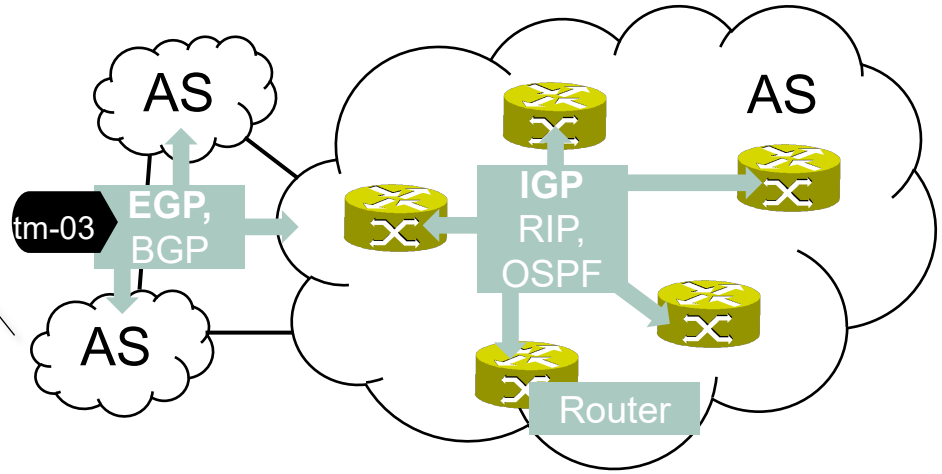


- Traffic engineering actions occur outside of the actual network

[Medh17]

Karlsruher Institut für Technologie

Control path
Routing between ASes



3.6 BGP: Border Gateway Protocol

3.6.1 Basics

Exterior Gateway Protocols

- Background
  - Division of large networks into autonomous systems (AS)
  - In each autonomous system, there is at least one special intermediate system that serves as an interface to other ASs

- Advantages
  - Scalability
    - Size of routing tables depends on size of AS
    - Changes in routing tables are only propagated within an AS
  - Autonomy
    - Internet = network of networks
    - Routing can be controlled in the own network
      - Uniform interior routing protocol within the AS
      - Interior routing protocols of different ASes do not have to be identical

[Huit00]

Border Gateway Protokoll (BGP)

- BGP is the most important exterior gateway protocol
  - Basis of today's internet-wide routing
  - Worldwide usage among all autonomous systems

- BPG is a path vector protocol
  - Extension of distance vector approach
  - BGP distributes paths, not metrics like costs etc.
    - With paths it is easy to guarantee that no loops exist
    - Paths are associated with path attributes

  - Based on policies of network operator
    - E.g., choosing economically viable paths, observation of contractual agreements, etc.

[Huit00, KuRo20]

Path Attributes

- Different categories of path attributes exist
  - We focus on well-known mandatory attributes only

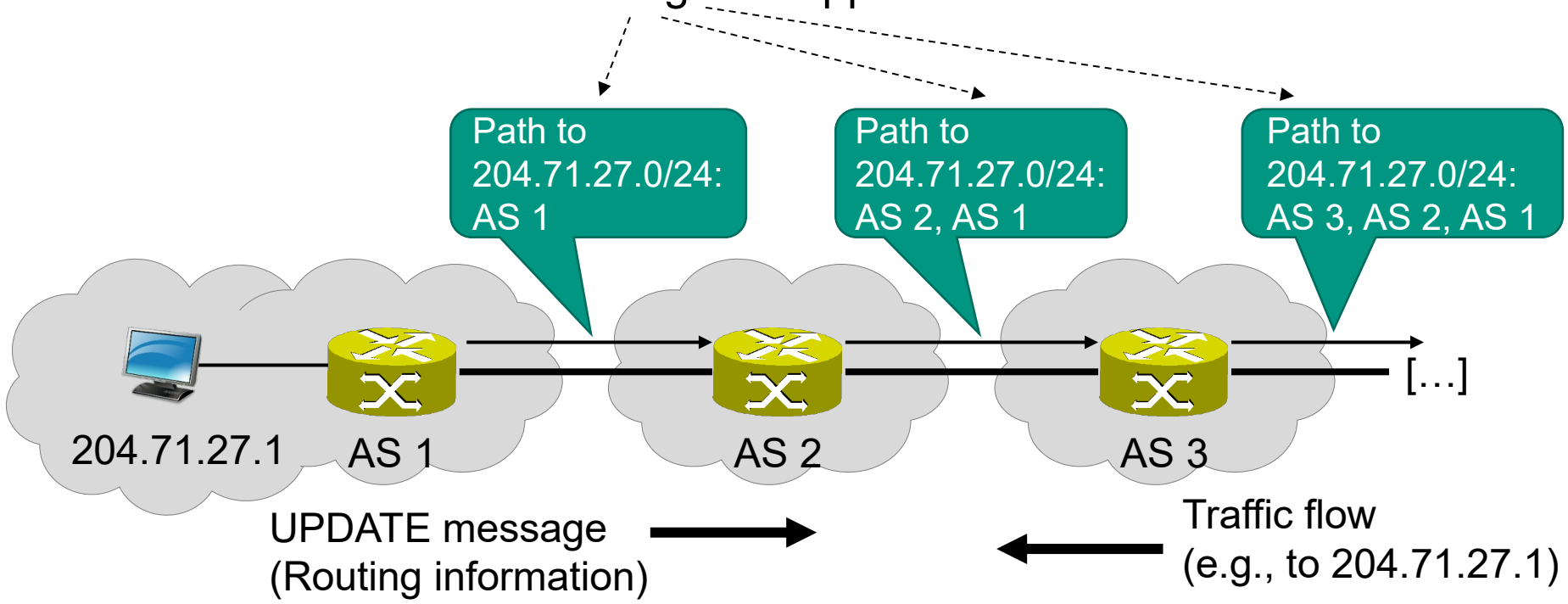
- ORIGIN
  - Mechanism how an IP prefix is first announced
    - IGP: obtained from interior gateway protocol, e.g., OSPF
    - EGP: obtained from exterior gateway protocol
    - Incomplete: IP prefix is unknown (e.g., in case of static routes)

- AS-PATH
  - Sequence of AS numbers that identifies the ASes which this UPDATE message has passed

- NEXT-HOP
  - IP address of next hop router on the way to the destination IP prefix

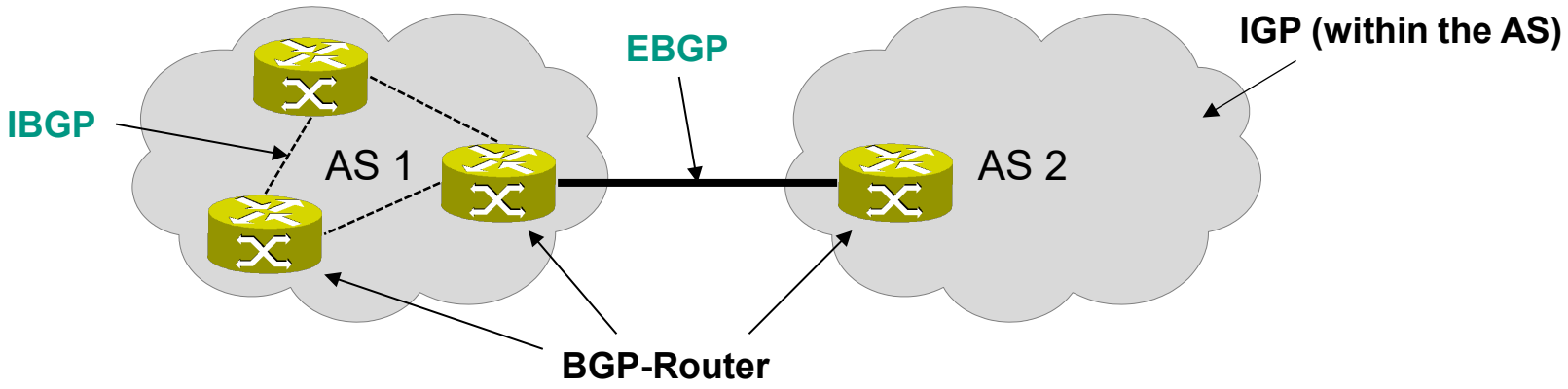
BGP in a Nutshell

- What exactly is being distributed?
  - Paths (also called routes) that consist of
    - Target: prefixes (also called: network, network prefixes, IP address ranges)
    - Attributes: path, next hop, …
      - Each traversed AS adds its own AS number to the path
- Traffic "follows" UPDATE messages in opposite direction

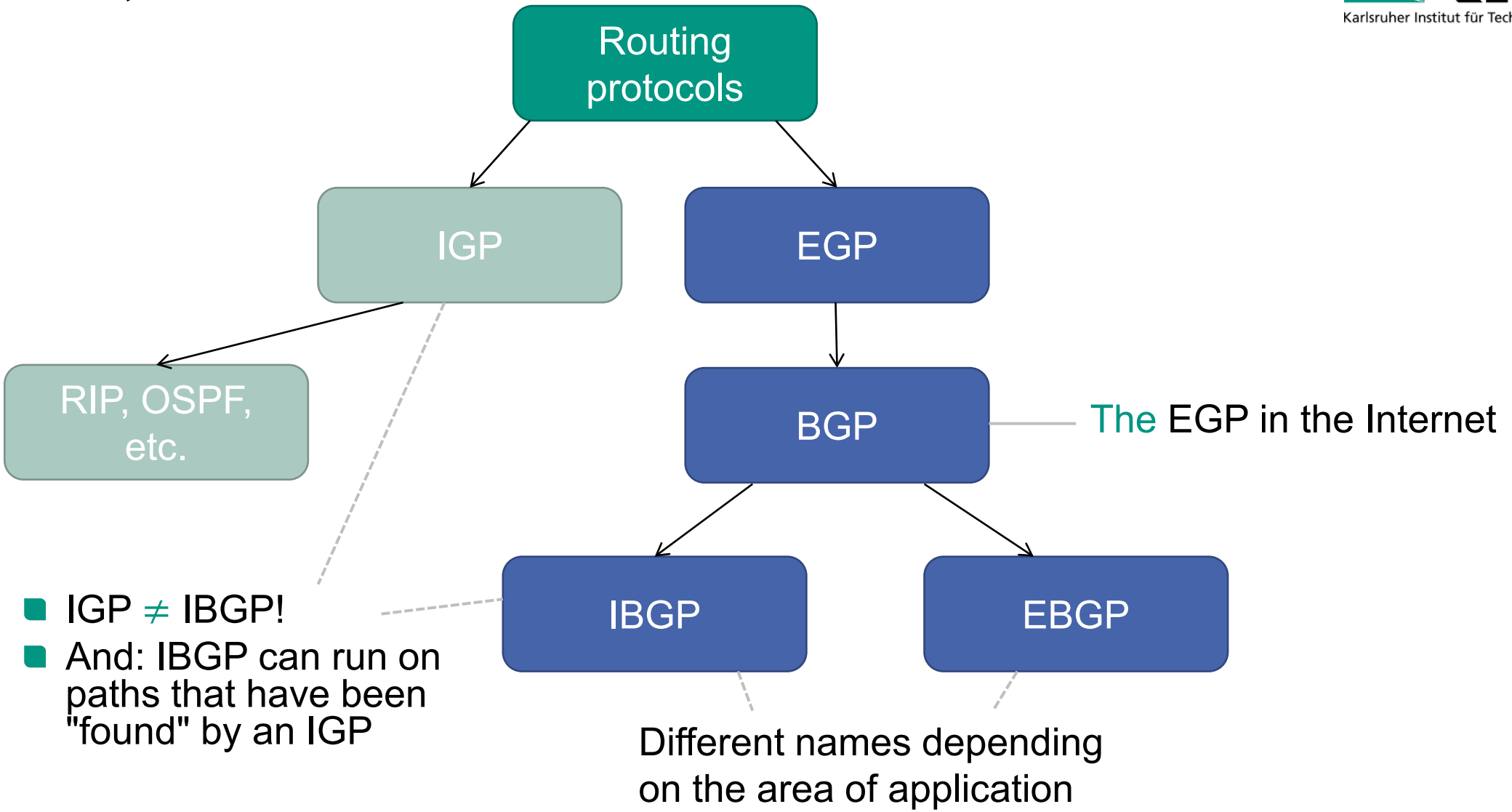


BGP - Structure

- External BGP (EBGP)
  - Spoken between BGP routers of neighboring ASs
  - Announcement and forwarding of path information
  - Internal details of AS are not exchanged
- Internal BGP (IBGP)
  - Spoken between BGP routers within an AS
  - Synchronization of BGP routers of an AS
  - Establishment of transit routes
    - For transit ASs

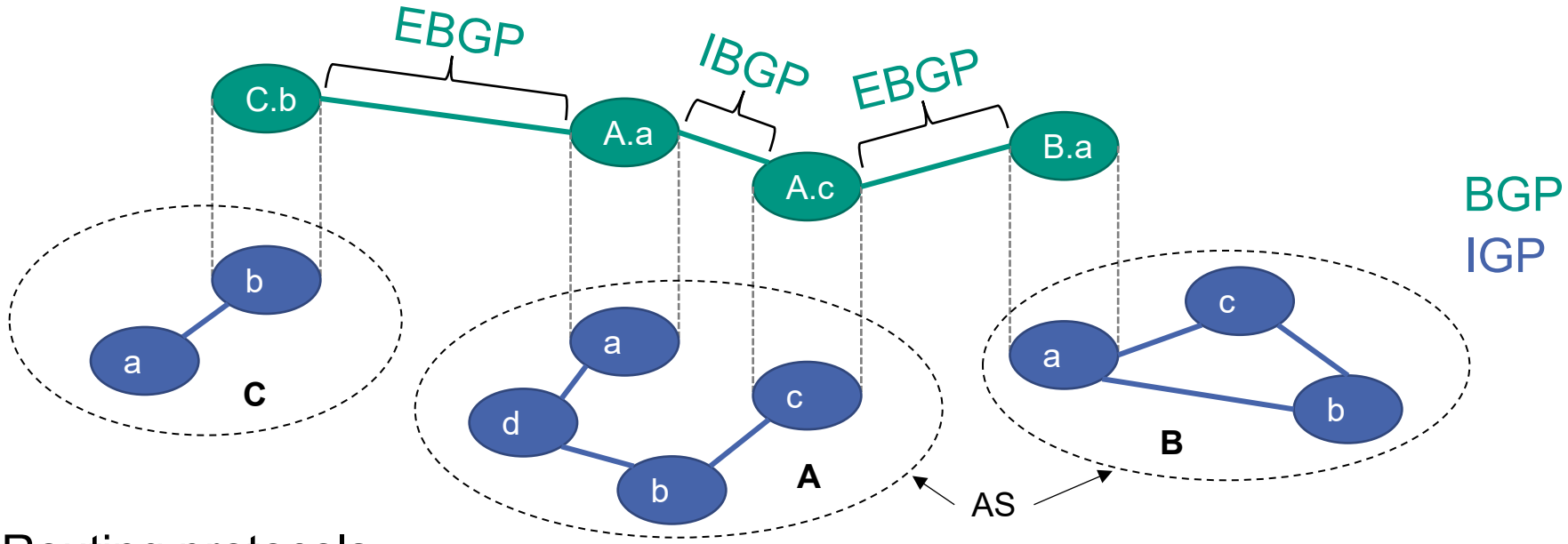


BGP, IGP, IBGP…



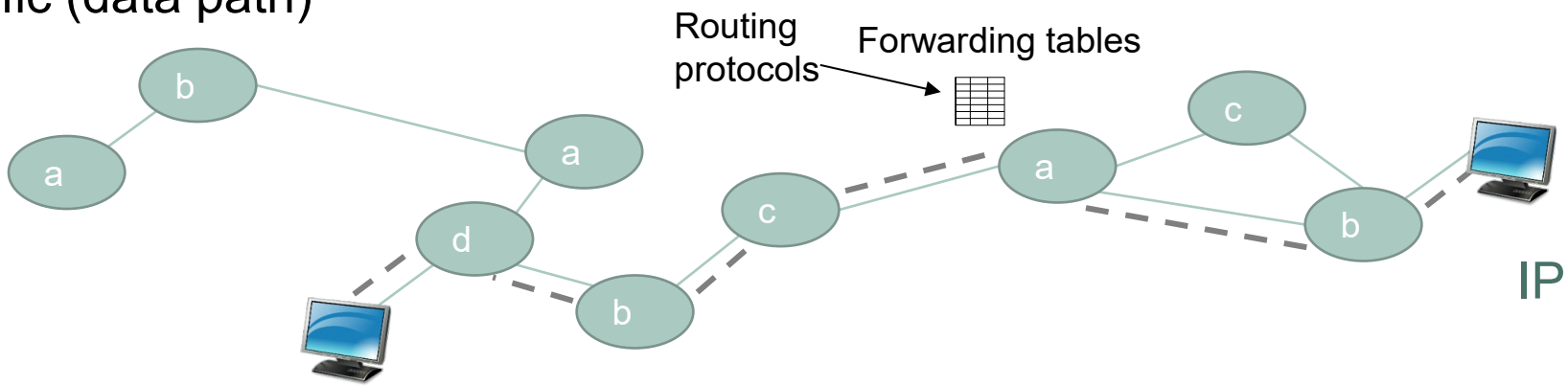
- IGP ≠ IBGP!
- And: IBGP can run on paths that have been "found" by an IGP

Interplay of the Routing Approaches



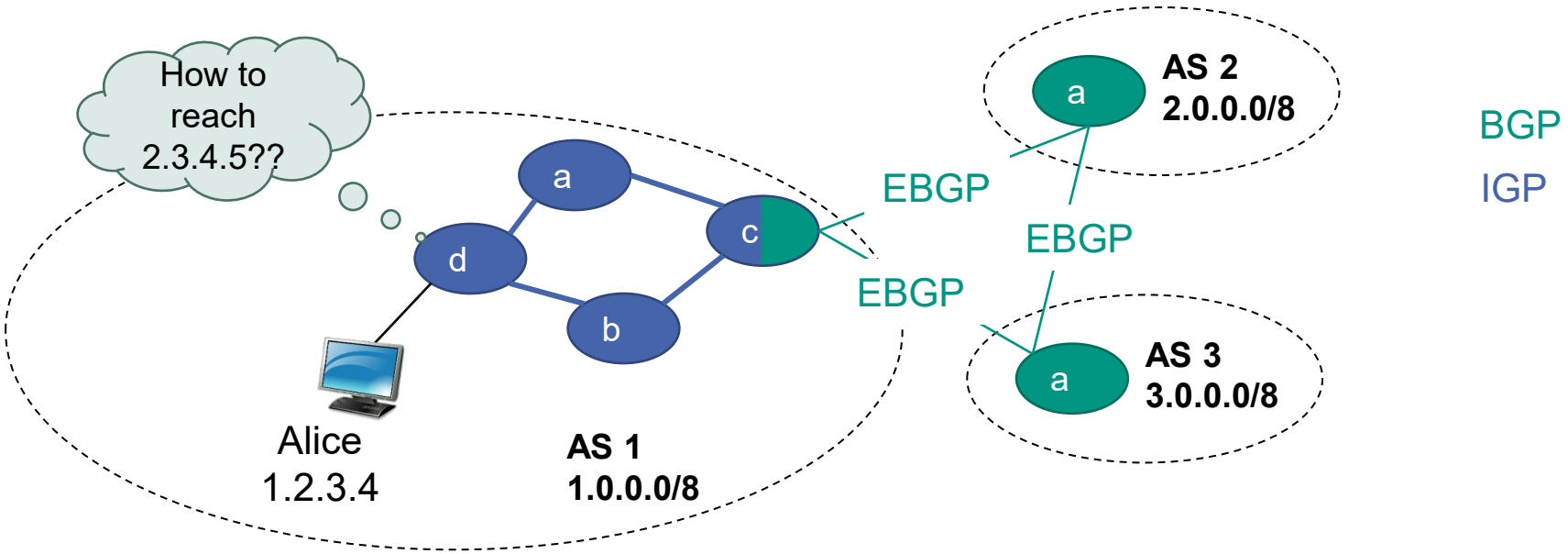
Routing protocols

Traffic (data path)



Routing with BGP and IGPs

- Assume Alice wants to send a packet to an external target
  - External target = not part of the local IGP domain, e.g., 2.3.4.5
- How does IGP router know what to do with this packet?
  - Is not strictly prescribed by BGP
  - Network operators can configure this freely
  - Different approaches possible

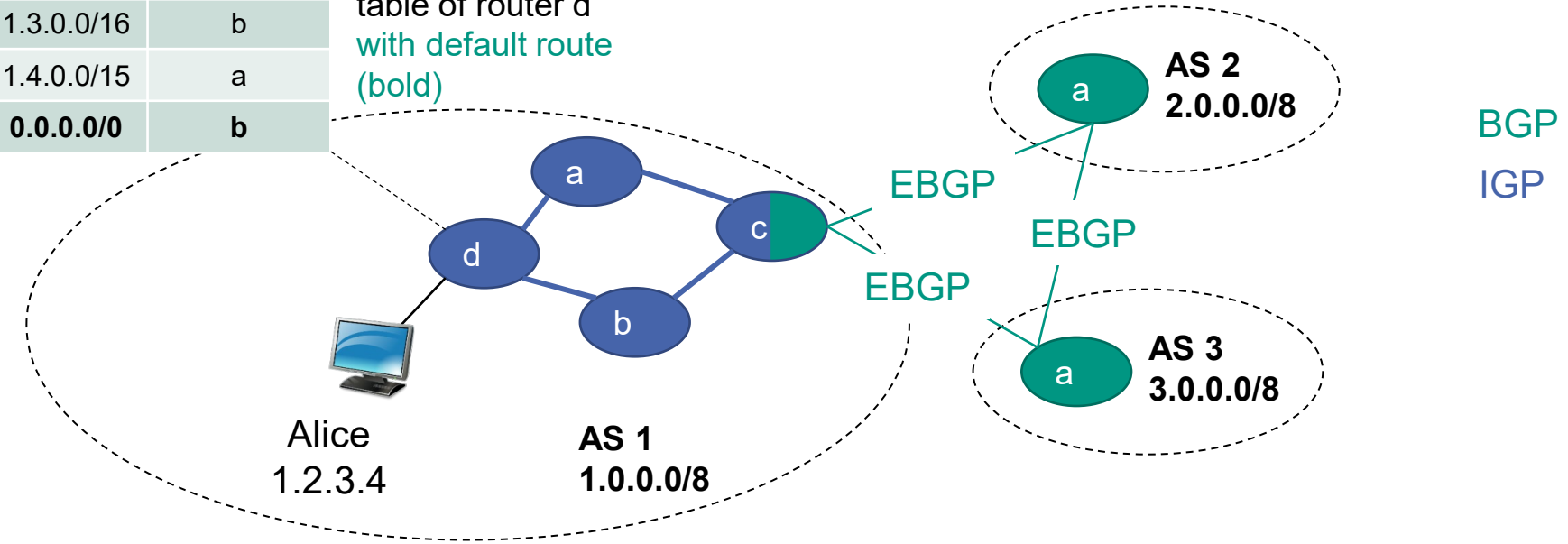


Routing with BGP and IGPs

- Approach 1: IGP distributes "default" routes
  - Unknown address/prefix packets are routed to default BGP router via shortest path
  - Good option for stub ASes
  - Not practicable for transit ASes

Target	Next Hop
1.3.0.0/16	b
1.4.0.0/15	a
0.0.0.0/0	**b**

Simplified routing table of router d with default route (bold)

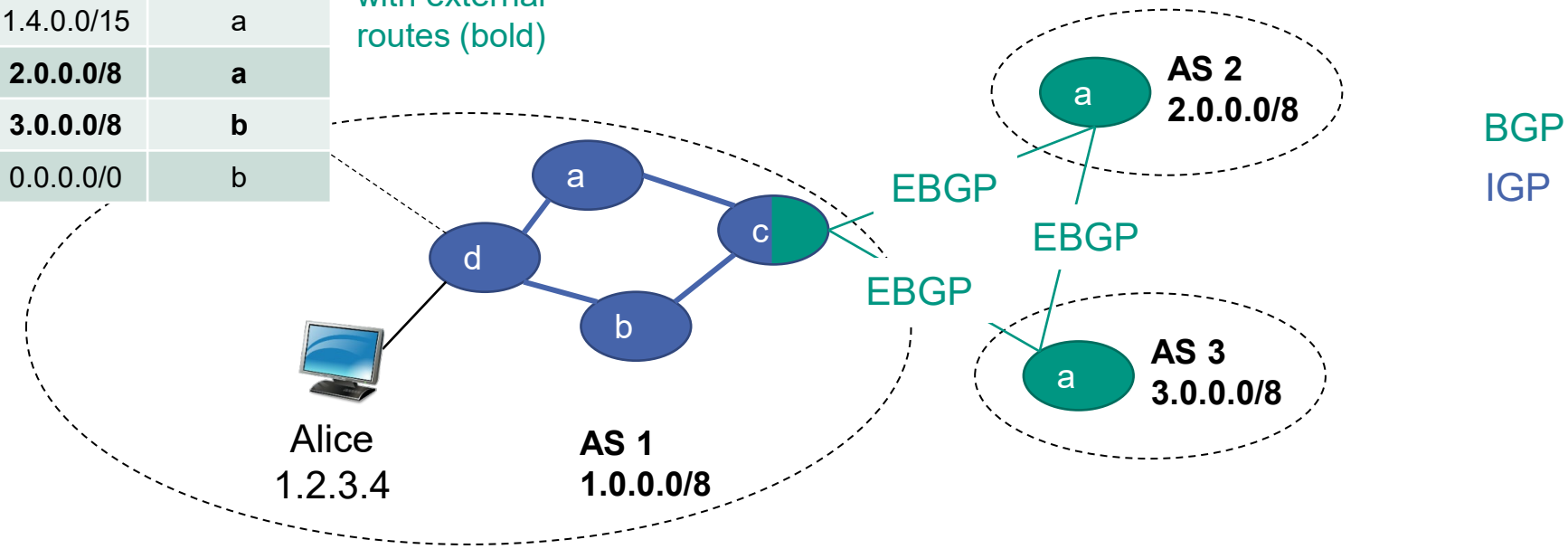


Routing with BGP and IGPs

- Approach 2: Publication of external routes via IGP
  - Allows more fine-grained control such as „all Google traffic goes this way“
  - Cannot be done with all external routes (scalability!)
  - Usually combined with default route

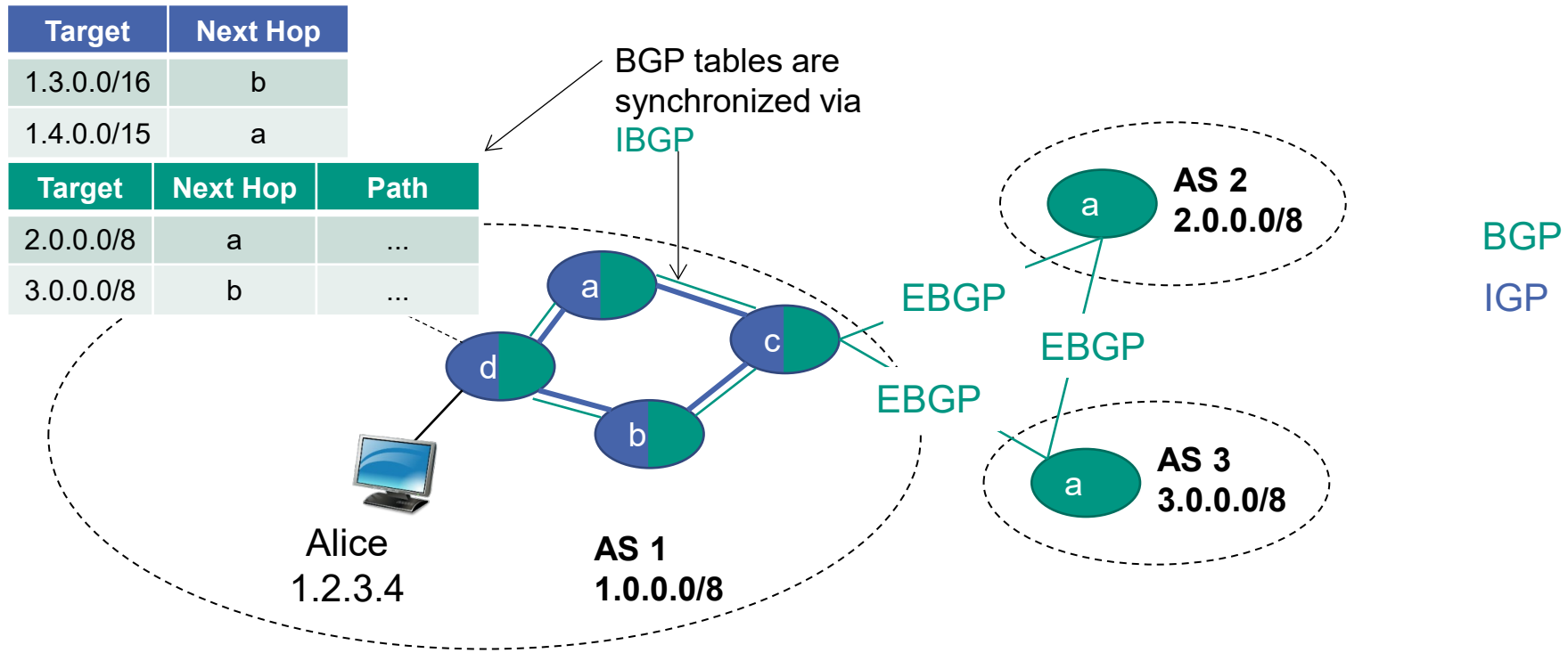
Target	Next Hop
1.3.0.0/16	b
1.4.0.0/15	a
2.0.0.0/8	**a**
3.0.0.0/8	**b**
0.0.0.0/0	b

Simplified routing table of router d with external routes (bold)



Routing with BGP and IGPs

- Approach 3: IGP router also speaks BGP
  - Forwarding table is build from two routing tables (BGP + IGP)
  - Often the case with big backbone providers



BGP-Sessions

- Point-to-point
  - Usually only between directly connected routers
    - Neighbors are called "peers"
  - BGP uses TCP connections between these routers
- „Chicken and egg“ problem?
  - How to establish TCP connection without working IP routing?
  - IBGP
    - IGP of AS can be used
  - EBGP
    - Usually direct physical connection → no routing required
    - Manual configuration at both ends
      - Neighboring network operators may need to coordinate with each other

IBGP Connections

- Simplest case: all BGP routers are fully meshed and connected directly to each other
  - BGP sessions must be kept active all the time
  - Bad scalability

- Alternatives
  - Concentrate IBGP traffic in a single router
    - Called route reflector
    - Only route reflector has to maintain sessions with everyone else
      - Forwards messages
    - More than one reflector used in practice for reliability reasons

[RFC4456]

  - Form hierarchies of sub ASes
    - Called AS confederations
    - Can also be used to implement more complex policies
    - Confederation appears to outside as single AS
      - Sub ASes receive private AS numbers

[RFC5065]

BGP - Messages

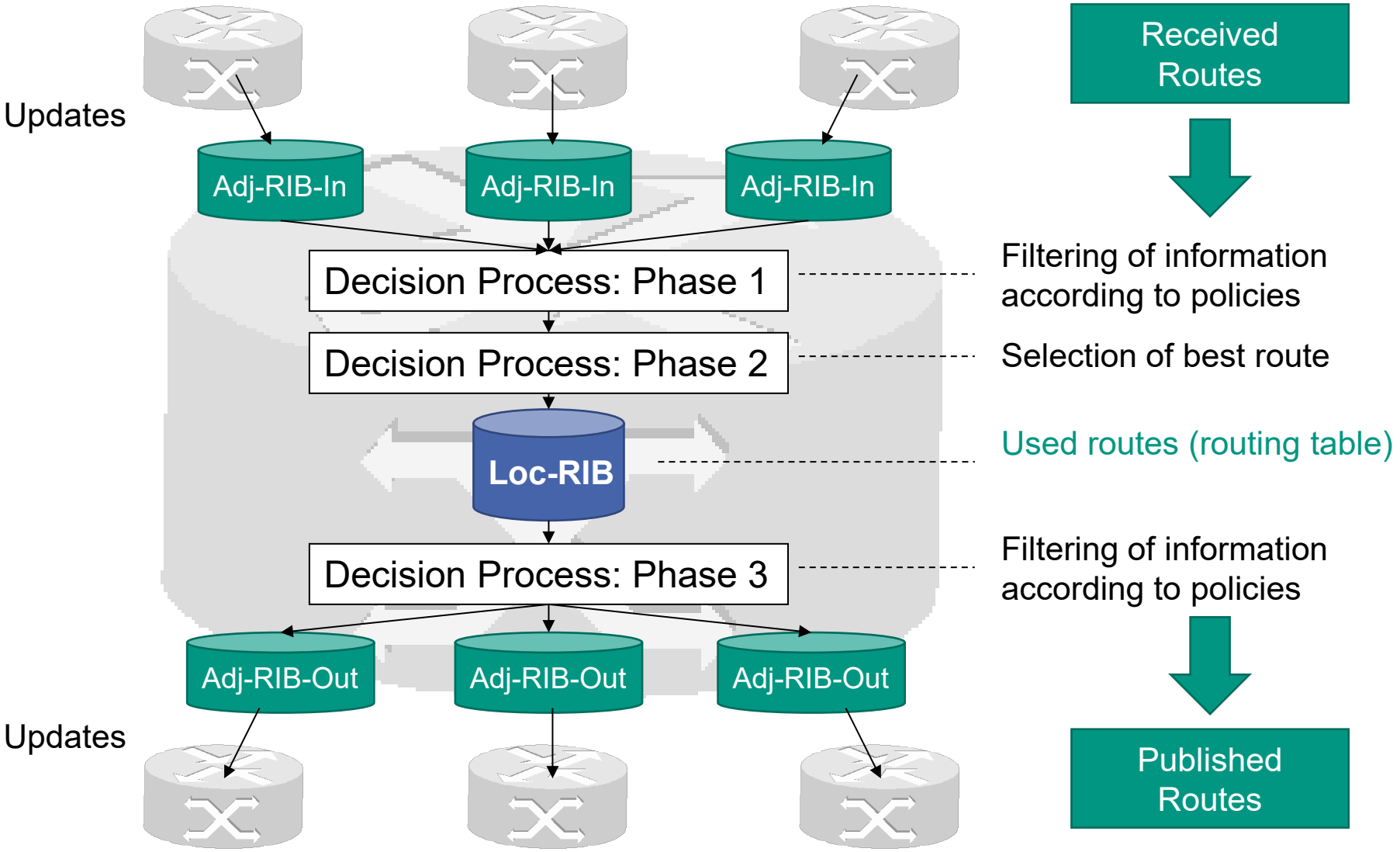
- OPEN
  - Establishment of BGP connection to peer BGP router
    - Important: TCP connection must already exist
  - Authentication
- UPDATE
  - Announcement of new or withdrawal of outdated path
  - Attention: Only sent if new, better paths available
    - For better scalability (without news, BGP is relatively silent)
- KEEPALIVE
  - Keeps connection alive in absence of UDPATE messages
  - Acknowledgment for an OPEN request
  - Recommended KeepAliveTimer: 30 s
- NOTIFICATION
  - Error message and tear down of BGP connection

* Note: If for a duration of 90 seconds neither an update nor a keepalive message is received from a peering BGP speaker, the peer is considered as not reachable

Routing Information Base

- BGP provides mechanisms for distributing path information
  - Does not dictate how routes should be chosen
    → No predefined routing metric
- BGP uses policies
  - Which routes should be considered?
  - Which is the “best” route?
  - Which routes should be published to which neighbors?
- BGP instance of a router collects received and dispatched routing information in various internal tables
  - Routing Information Base (RIB)
  - Mainly for logical structuring

Routing Information Base



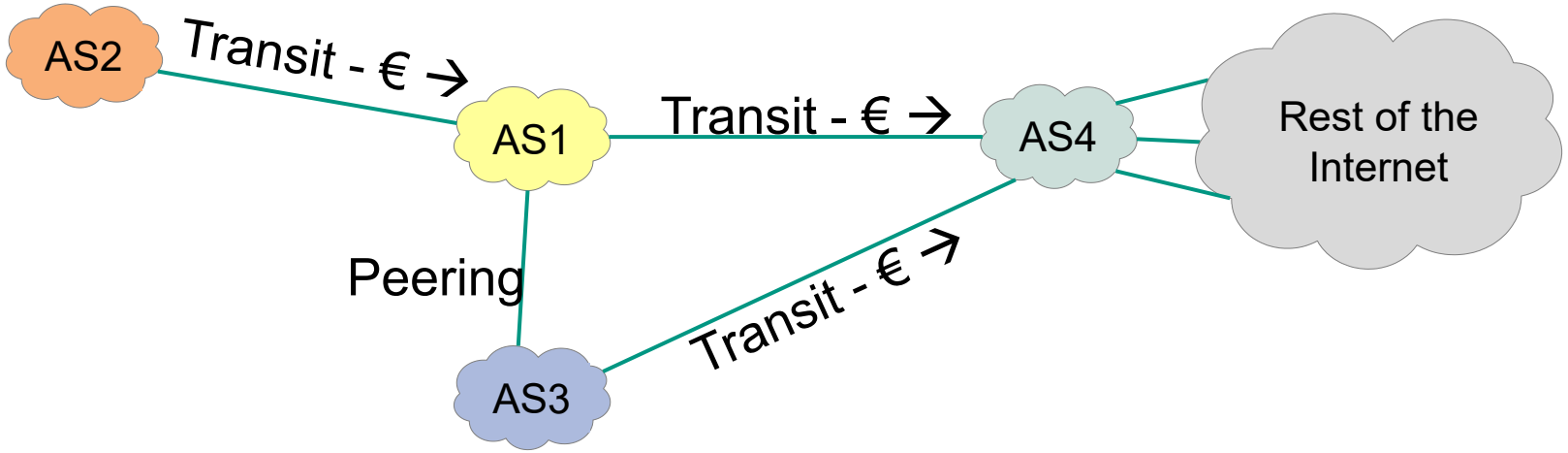
Routing Information Base

- Adj-RIB-In (*Adjacent RIB Incoming*)
  - Exists per peer
  - Stores information received from this peer in received UPDATE messages
- Loc-RIB (*Local RIB*, *Routing Information Base*)
  - „Actual routing table“
    - Only preferred (= best) routes to destination networks are included here
      - Selected by the BGP speaker´s Decision Process
    - Forwarding Information Base (FIB) is build based on Loc-RIB
- Adj-RIB-Out (*Adjacent RIB Outgoing*)
  - Exists per peer
  - Contains routes published to this peer through UPDATE messages

Routing Decision Process

- Phase 1
  - Calculate degree of preference for each received route
    - May be based on preconfigured policy information
- Phase 2
  - Choose best route for each destination
    - Loops must be avoided (check own AS number in AS path)
    - Select route with highest degree of preference to destination
      - If multiple exist … apply tie breaking rules
        - Select route originated locally at BGP speaker
        - Select route with smallest number of ASes in AS path
        - …
  - Install route in Loc-RIB
- Phase 3
  - Routes in Loc-RIB are processed according to configured policy
    - A route in Loc-RIB may be excluded from a particular Adj-RIB-Out

Example: Policies



- AS1

```
Import: from AS2 all paths whose first hop is AS2
Import: from AS3 all paths whose first hop is AS3
Import: from AS4 all paths
Export: to AS2 all paths
Export: to AS3 all paths whose first hops are AS1, AS2
Export: to AS4 all paths whose first hops are AS1, AS2
```

Example: Excerpt from BGP Routing Table

```
myrouter>show ip bgp
BGP table version is 6543445, local router ID is .....
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network  Next Hop            Metric LocPrf Weight Path
*  3.0.0.0  202.249.2.86                           0 7500 2516 701 80 i
*           167.142.3.6                            0 5056 701 80 i
*           195.66.224.82         23302            0 4513 701 80 i
*           195.219.96.239                         0 8297 6453 701 80 i
*           192.121.154.25                         0 1755 701 80 i
*           205.215.45.50                          0 4006 701 80 i
*           195.211.29.22                          0 5409 6667 8543 […] 80 i
*           207.172.6.173            22            0 6079 701 80 i
*           206.220.240.222                        0 10764 1 701 80 i
*>          157.130.185.17                         0 701 80 i
*           157.22.9.7                             0 715 701 80 i
*  4.40.228.0/23   206.220.240.222                  0 10764 11537 5661 13755 i
*                  203.181.248.242                 0 7660 11537 5661 13755 i
*                  193.0.0.56                      0 3333 1103 […] 13755 i
*                  134.55.20.229                   0 293 11537 5661 13755 i
*>                 198.32.8.252                    0 11537 5661 13755 i
```

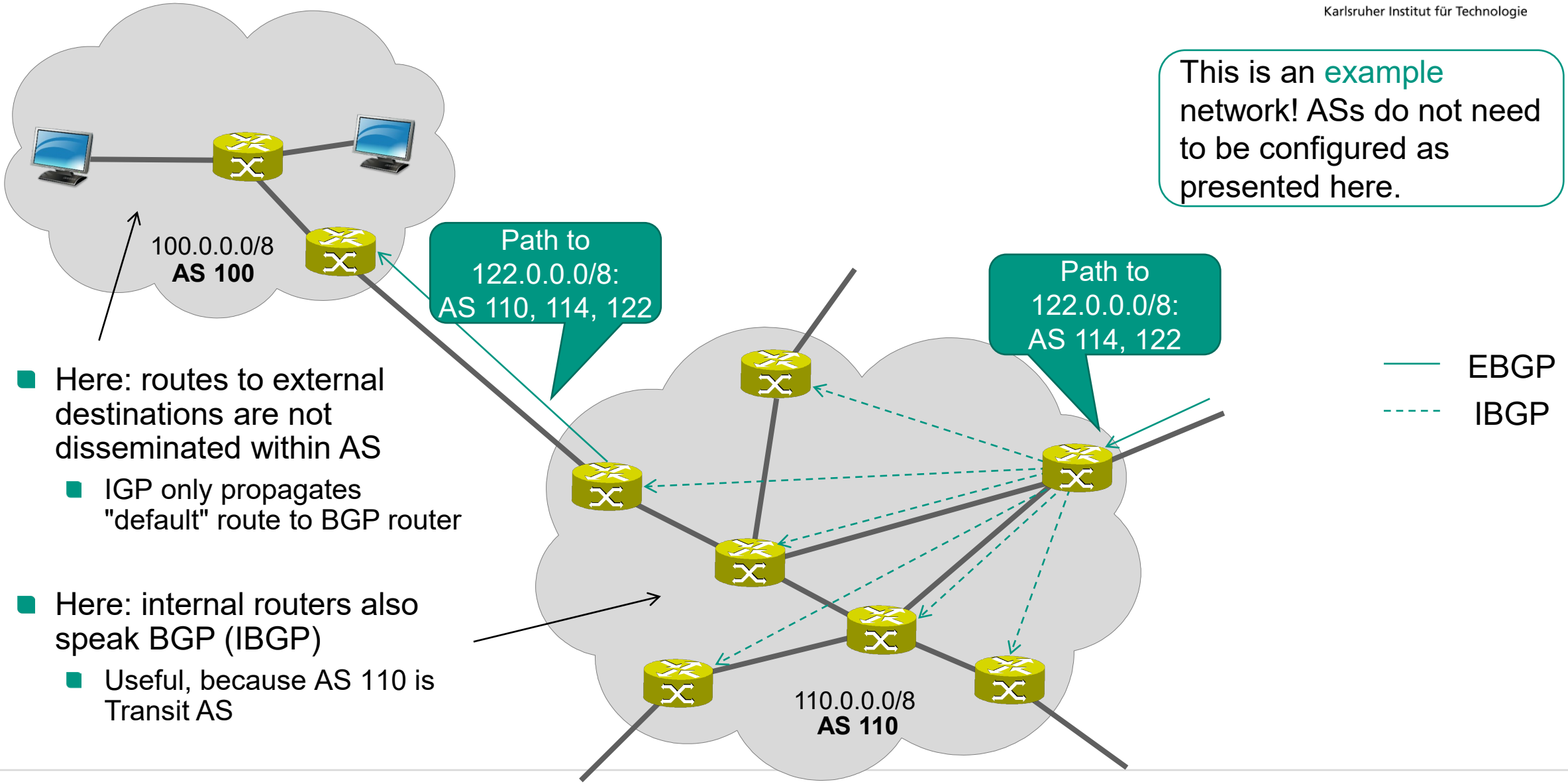
AS Pfad: AS 7500, AS 2561

Target AS

Best route

Target Network

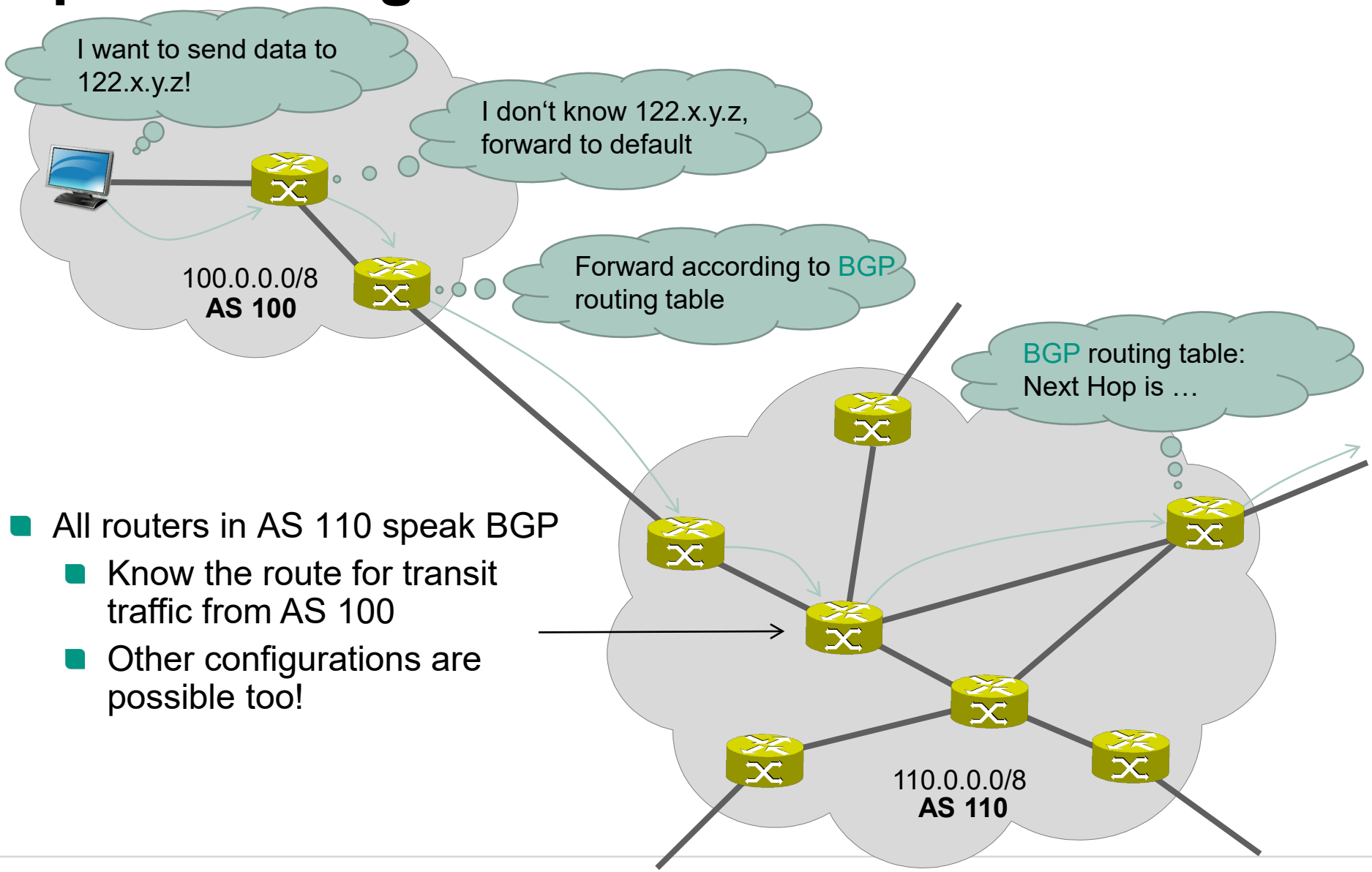
Example: Distribution of Path Information



This is an example network! ASs do not need to be configured as presented here.

- Here: routes to external destinations are not disseminated within AS
  - IGP only propagates "default" route to BGP router
- Here: internal routers also speak BGP (IBGP)
  - Useful, because AS 110 is Transit AS

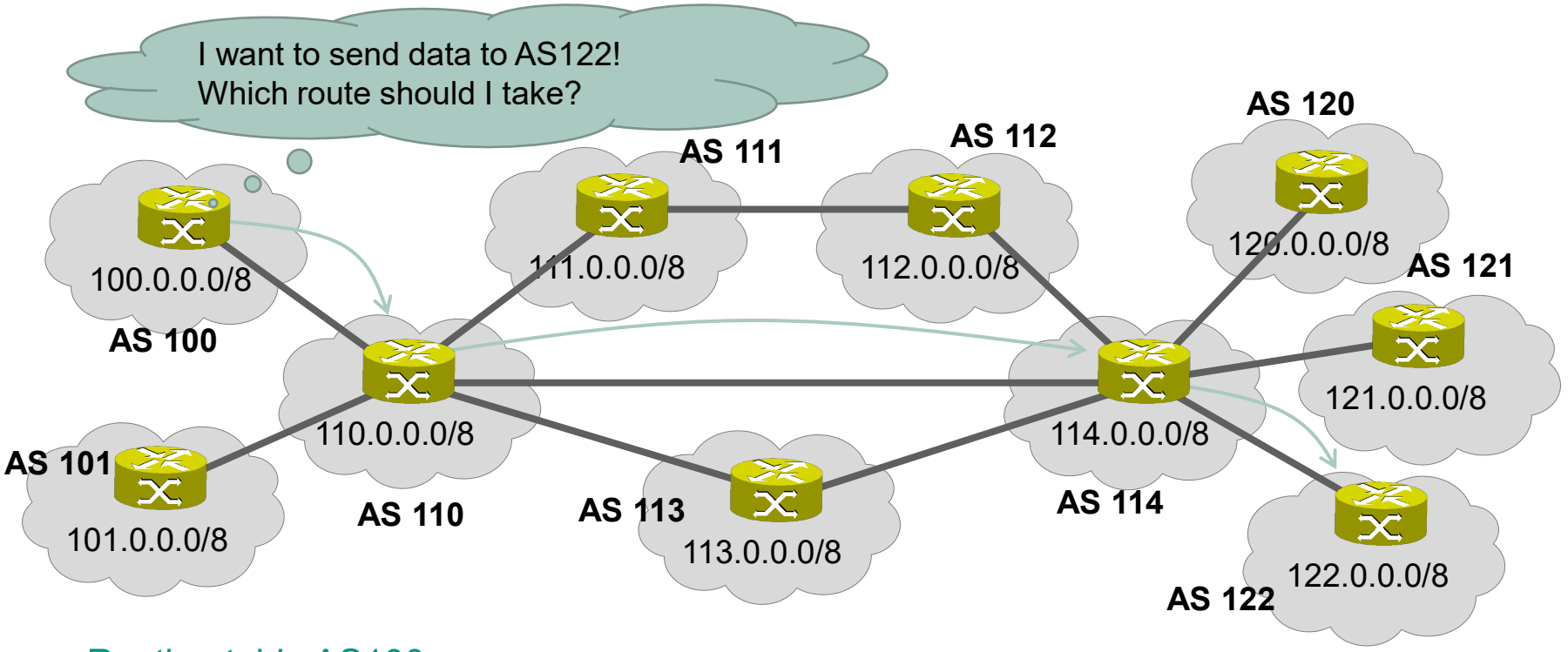
Example: Routing



Data

- All routers in AS 110 speak BGP
  - Know the route for transit traffic from AS 100
  - Other configurations are possible too!

Example: Routing between multiple ASs



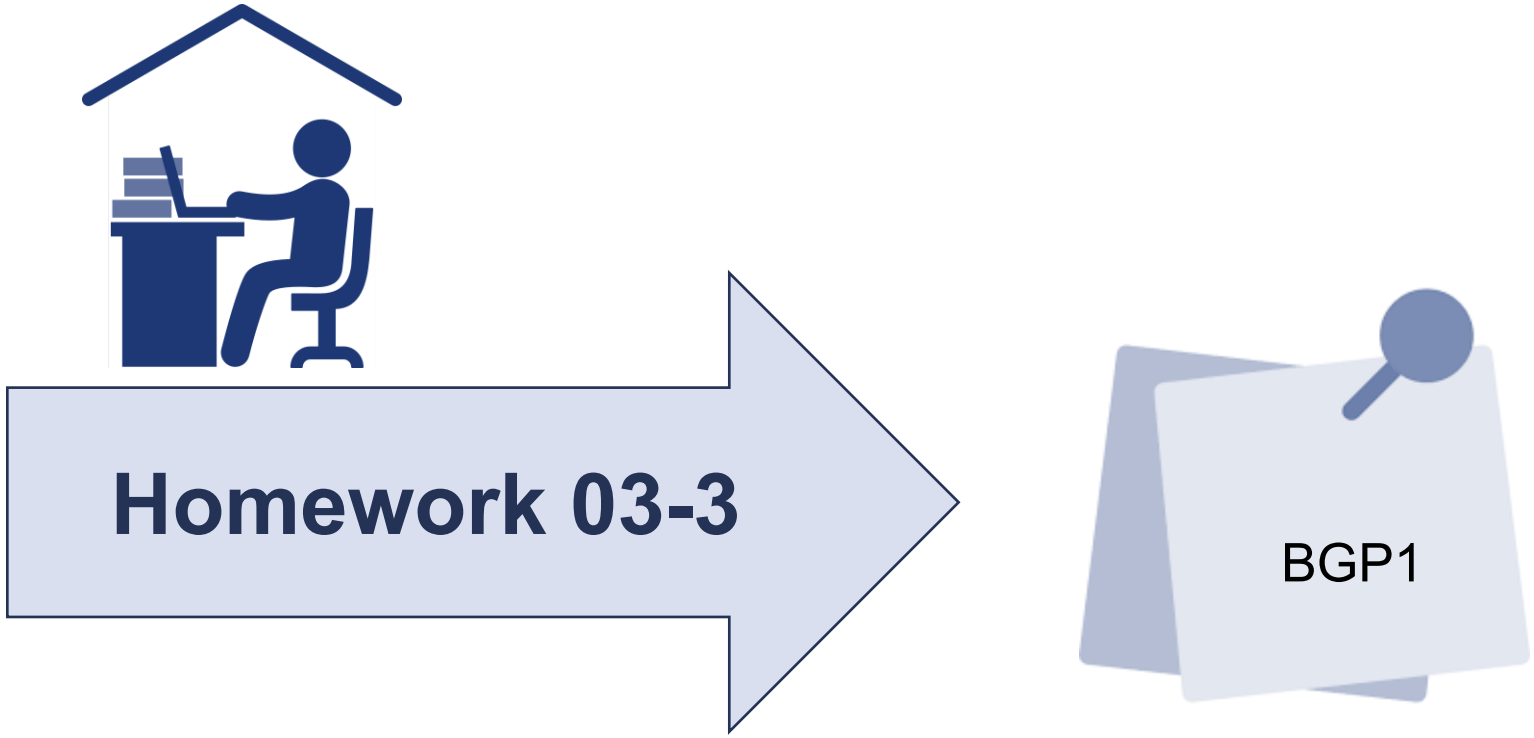
Routing table AS100

Routing table AS110

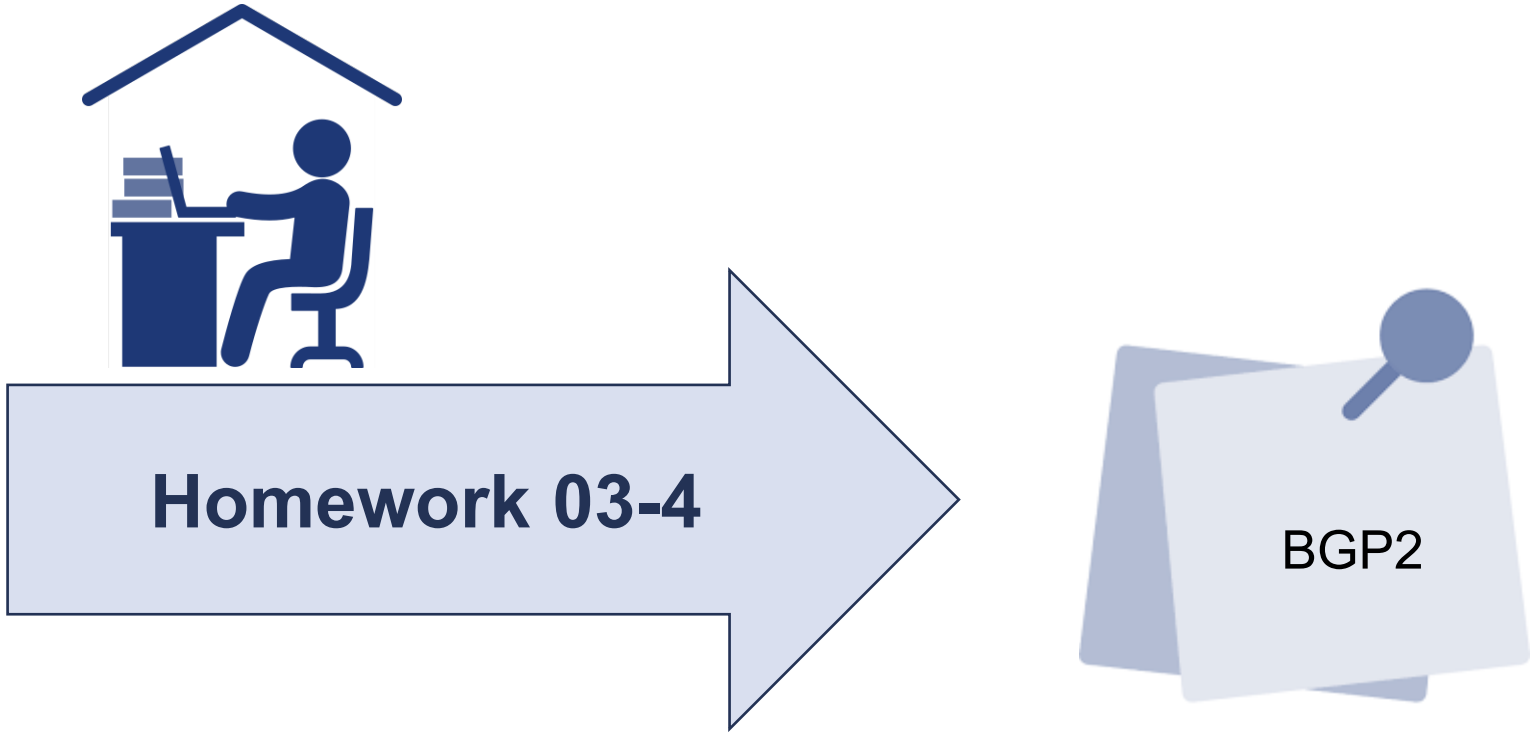
Routing table AS114

```
  Network          Next Hop            Metric LocPrf Weight Path
*> 112.0.0.0       10.1.1.112               0             0 112 i
*                  10.1.1.110                             0 110 111 112 i
…
*> 122.0.0.0       10.1.1.122               0             0 122 i
```

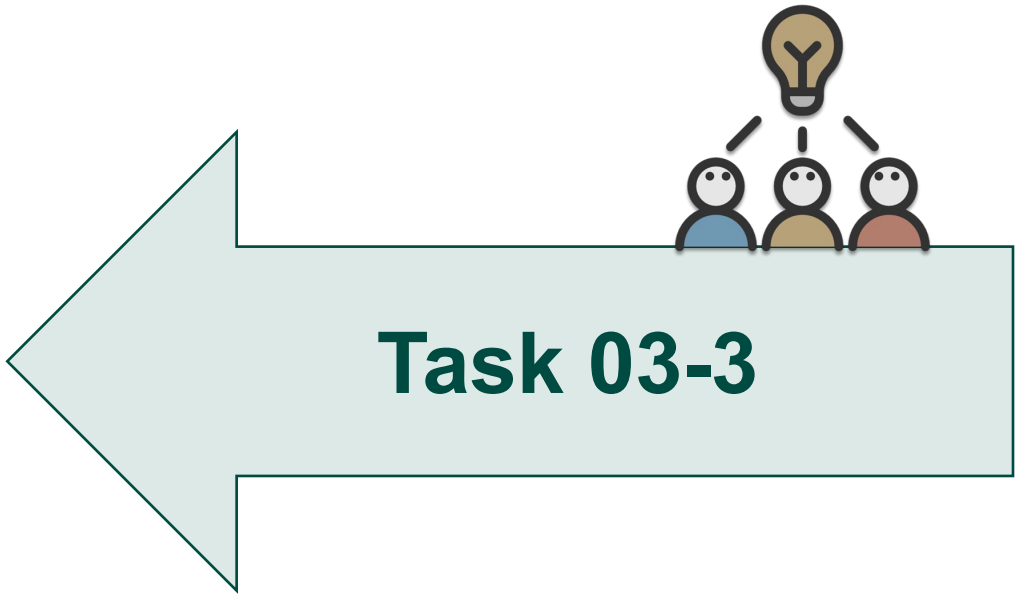
Homework



Homework




Task 03-3
BGP


Task 03-3

3.6.2 Challenges of BGP

Challenges

- BGP is a complex protocol
  - … we just touch the surface in this lecture
- Critical aspects
  - BGP is very sensitive to configuration errors
  - BGP has security issues
    - Unauthorized BGP originations (Prefix Hijacks)
    - Unauthorized BGP update modifications (Path Hijacks)
    - BGP policy violations (Route Leaks)
- Other challenges
  - Vulnerabilities related to routing
    - Spoofed source addresses
    - Reflection amplification attacks
  - Maintaining scalability
    - Increasing deaggregation of routing information
      - For example, due to the multi-homing of ASs
    - Growth of routing tables
    - Increasing dynamics of routing changes
  - Increased demands on the Internet

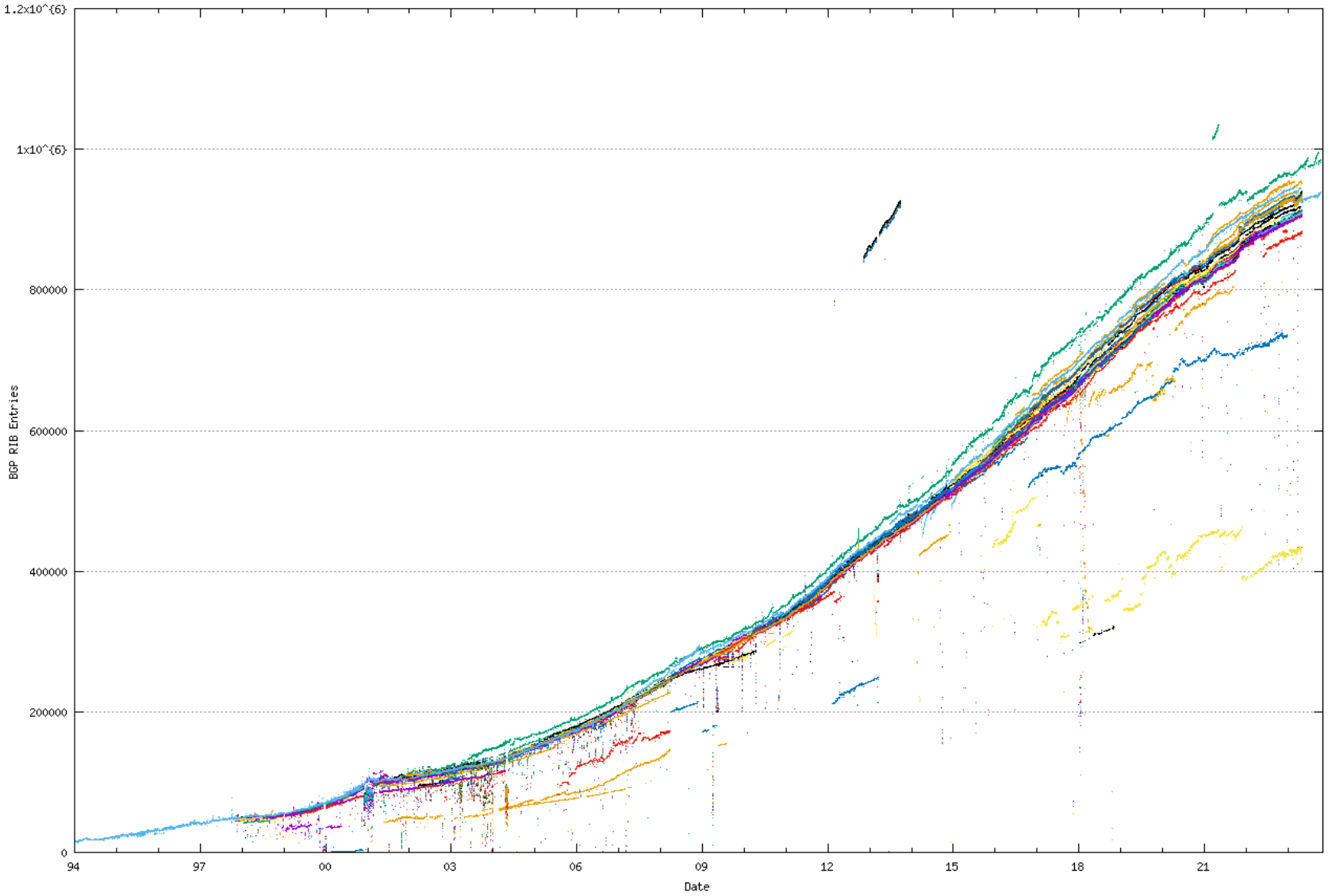
3.6.2.1 Scalability

Multi-Homing

- What is multi-homing?
  - AS at the edge of the Internet is connected to the Internet via several ASs
- Why have "multi-homed" autonomous systems?
  - Resilience of the Internet connection
  - Costs (important traffic uses fast but expensive uplink, other traffic uses cheap uplink)
- Where is the problem?
  - Aggregation of prefixes is broken up
  - Changes to preferred route must be propagated throughout the whole Internet in the worst case
- Possible remedy
  - NOPEER attribute: Restricts the propagation of changes to the edge of the Internet

[RFC3765]

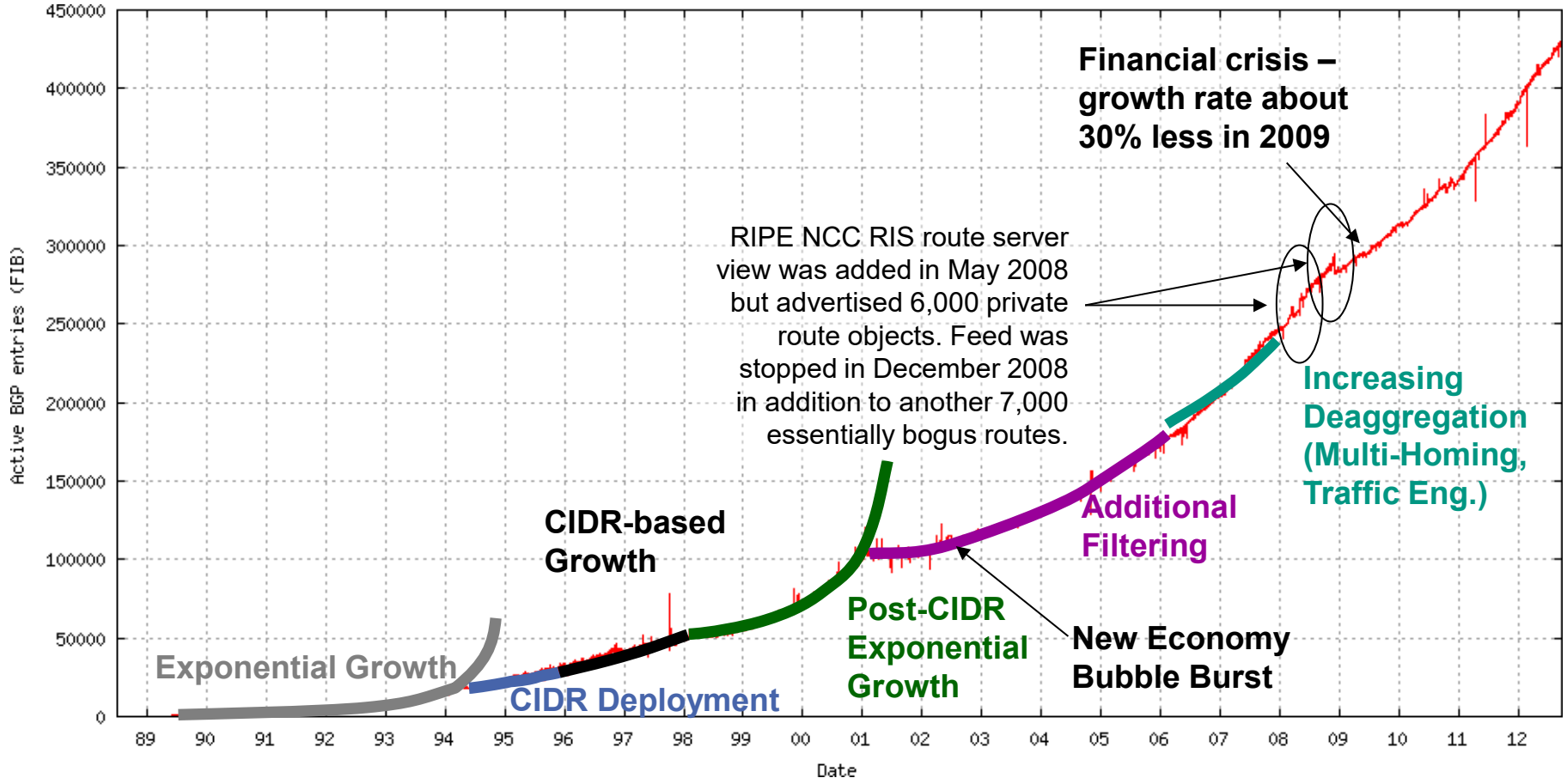
Size of Routing Tables (IPv4)



FIB: Forwarding Information Base
BGP: Border Gateway Protocol

Source: http://bgp.potaroo.net/

Growth of Routing Tables



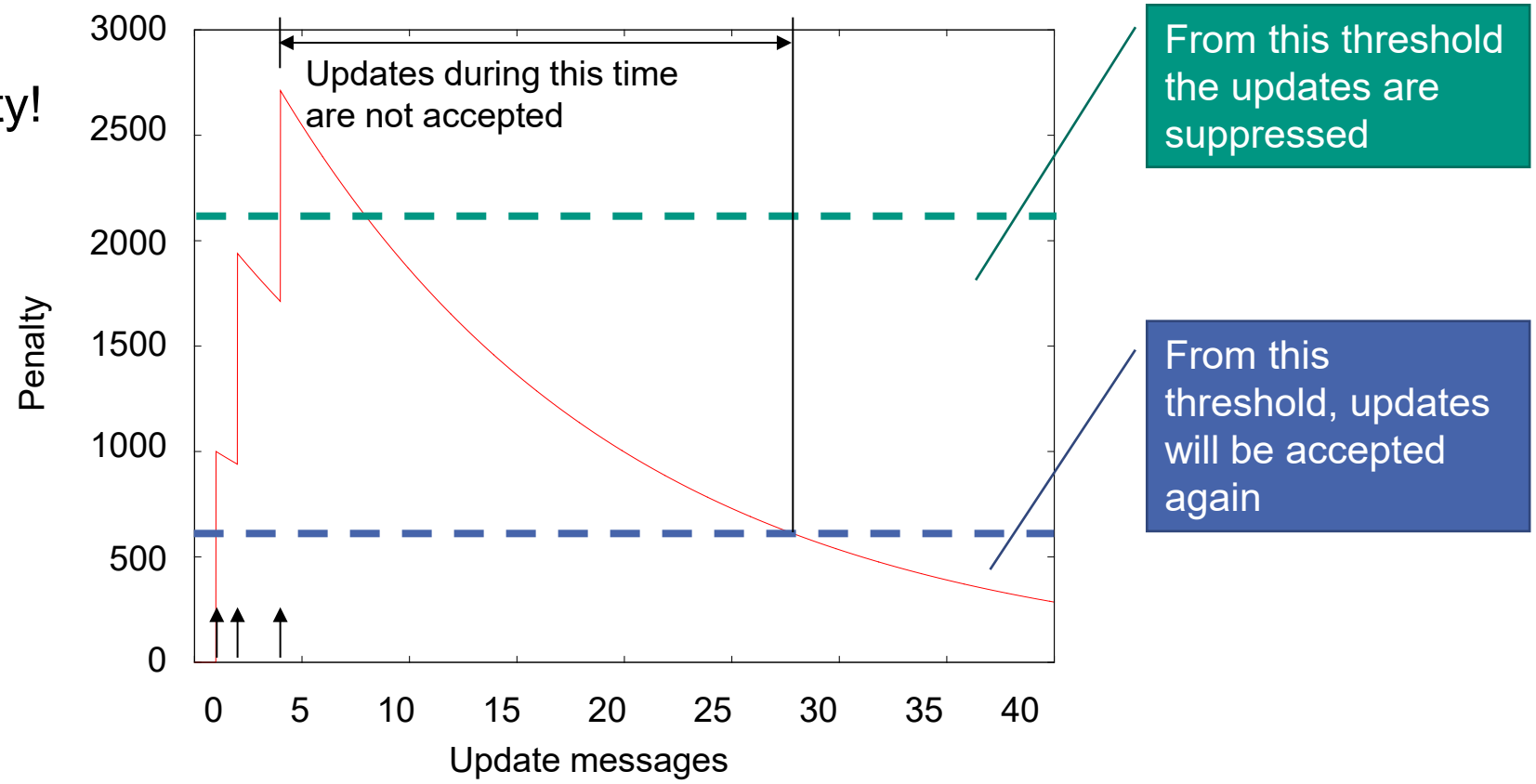
Increasing Dynamics of Routing Changes

- Problem
  - Numbering of ASes is flat, not subdivided
  - Autonomous systems differ in their functionality
    - Stub-AS or multi-homed AS
    - Transit-AS
  - BGP knows no hierarchy
    - How can the propagation of topology changes be "reasonably" restricted?
- Effects
  - Aggregation of prefixes continues to decline due to multi-homing and traffic engineering
  - Meshing of the autonomous systems with each other becomes denser
  - Routes in routing table change more frequently
- Possible remedy
  - Route flap damping
    - Temporarily suppress unstable routes

 [RFC2439]  [RFC4989]  [Hust06]

Route Flap Damping

- Temporarily suppress changes of unstable routes
- A penalty will be increased per update
  - here 1000 points per update
- Penalty value drops exponentially over time
- Can lead to loss of connectivity!



Increasing Quality Requirements

- Increasing demands of applications on the transmission quality
  - VoIP, Skype etc. need low jitter and small delays
    - Delay and jitter depend on routes / route changes
    - Effects: What happens if a route is not stable?
      - Voice connection sounds tinny or has dropouts
      - Cancellation of the VoIP session at approx. 2% packet loss (depending on the codec)

- BGP can not choose high quality routes
  - BGP only provides length of AS path as a metric

- Possible remedy
  - No other BGP metrics in sight
  - Reduction of route change frequency
    - Route flap damping – can cause connection loss
    - NOPEER Attribute – only helps with autonomous systems at the edge
  - All approaches only help with partial problems

3.6.2.2 Configuration Problems

Routing Incident: Example OVH

- October 13th, 2021
  - Worldwide outage, duration ~ 1 hour
- What happened?
  - OVH intended to increase its DDoS processing capacity by adding new infrastructures
  - Entire BGP routing table was announced in the OVHcloud IGP
    - More than 800000 IPv4 routes
    - … neither OSPF nor routers scale for such high load
    - OSPF routing table was full
    - RAM and CPU became overloaded
    - Only IPv4 traffic was affected
  - → IPv4 routing inoperable

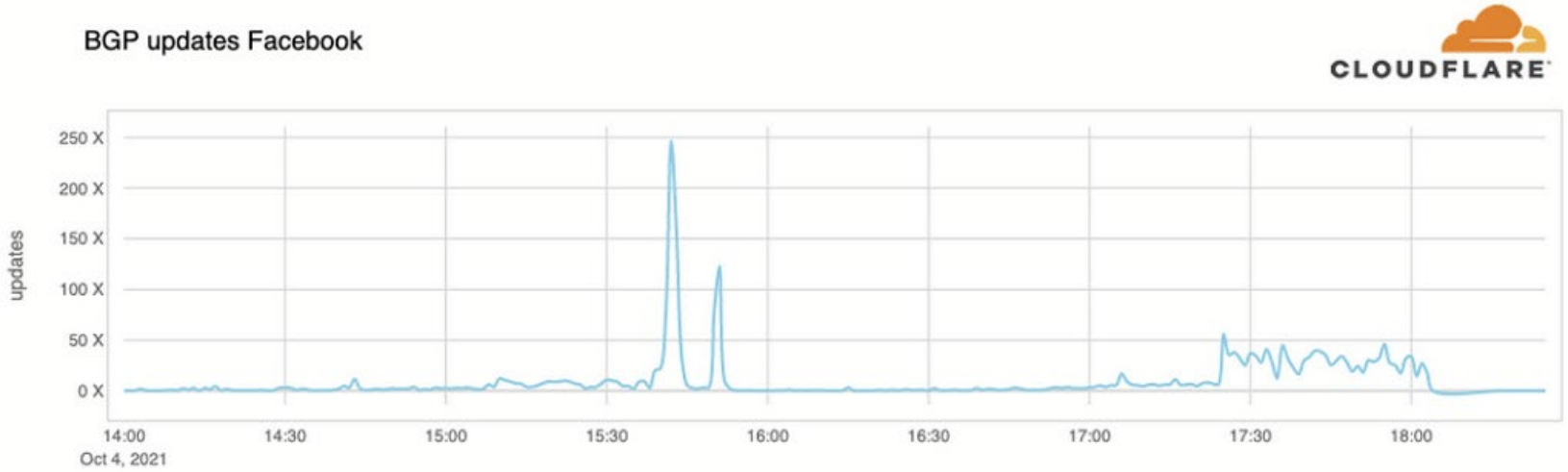
[https://www.theregister.com/2021/10/13/ovh_outage/
https://corporate.ovhcloud.com/en/newsroom/news/network-incident/]

Example: Facebook not available

- Facebook, WhatsApp, Instagram were down
  - October 4th 2021, ~ 15:50 UCT until ~21:20 UCT
- Attack, configuration fault … ???

Facebook not available

- Cloudflare: „Facebook DNS lookup returning SERVFAIL“
  - ~15:40 UCT: peak of routing changes from Facebook
  - 15:58 UCT: Facebook stopped announcing routes to their DNS prefixes

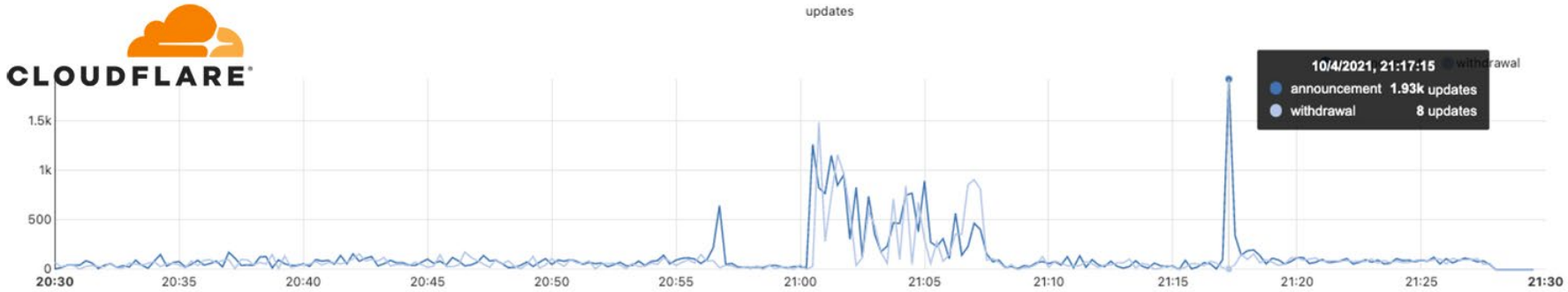


[https://blog.cloudflare.com/october-2021-facebook-outage]

- Announcements and withdrawals (light blue)

Facebook not available

- Update
  - ~21:00 UCT: renewed BGP activity from Facebook



[https://blog.cloudflare.com/october-2021-facebook-outage]

Facebook not available

- Centralized infrastructure
  - All peering routers operated in own infrastructure
  - All DNS servers operated in own infrastructure

- No routes available to Facebooks systems
  - DNS systems not reachable although they were still operational

  - Internal systems not reachable from outside
    - Internal tools could not operate as usual
  - Networked locking system caused problems

  - Engineers were sent to facility
    - High levels of physical and system security … access difficult

DNS

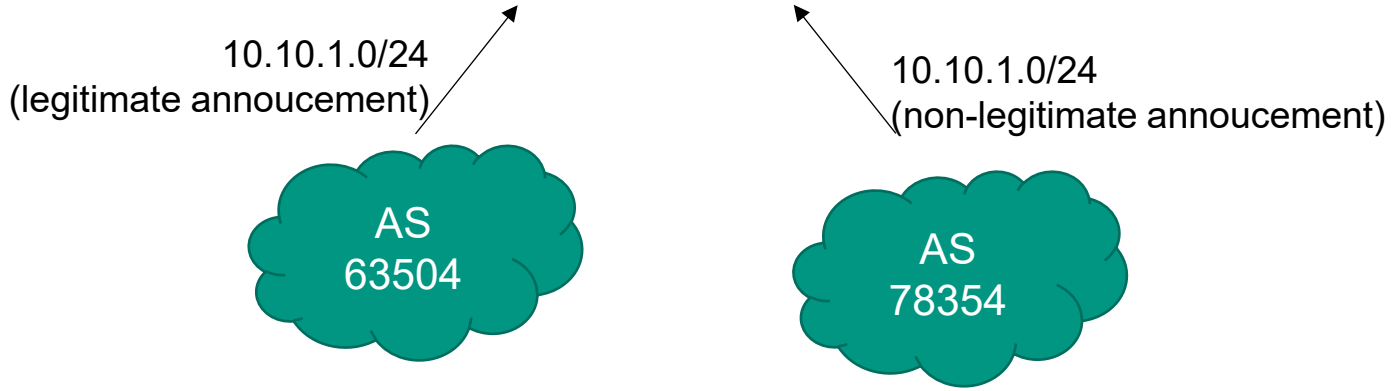
BGP

[https://www.heise.de/-6209377]

3.6.2.3 Prefix Hijacking

Announcement of IP Prefixes

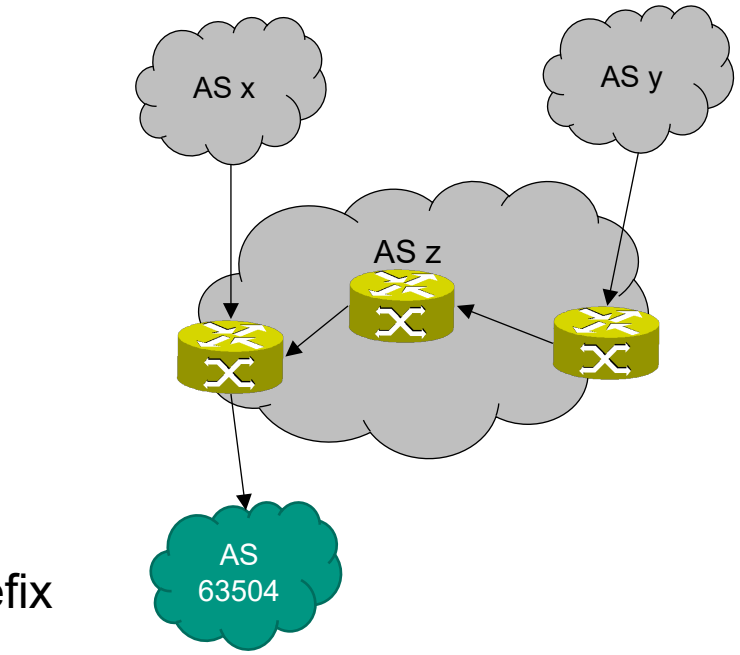
- The following steps are needed
  - Organization obtains an IP prefix
  - IP prefix must be tied to an AS
  - Announcement to the rest of the Internet through BGP
- Annoucement based on trust
  - AS that announces IP prefix owns this IP prefix
- Example



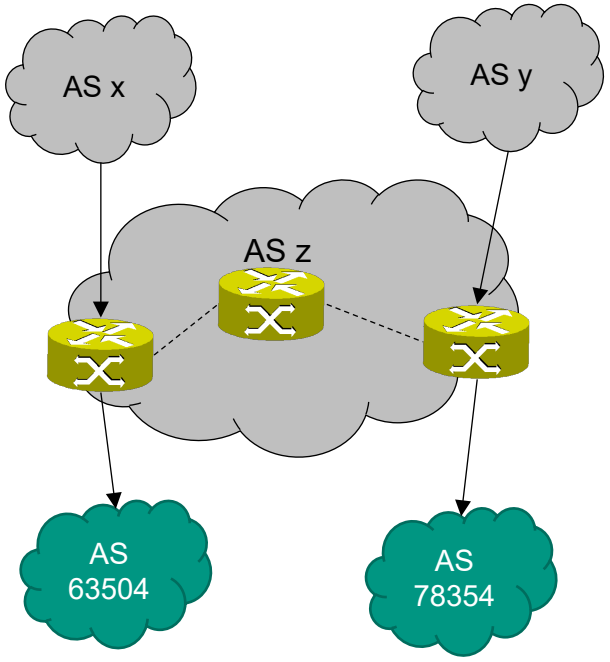
  - Since BGP works on trust it accepts both annoucements

IP Prefix Hijacking

- Situation
  - AS announces IP prefix without any authorization
    - i.e., AS does not own this IP prefix
- Consequence
  - Traffic for respective IP prefix forwarded to non-legitimate AS
- Problem
  - Misconfigurations may lead to prefix hijacking



10.10.1.0/24
(legitimate annoucement)



10.10.1.0/24
(non-legitimate annoucement)

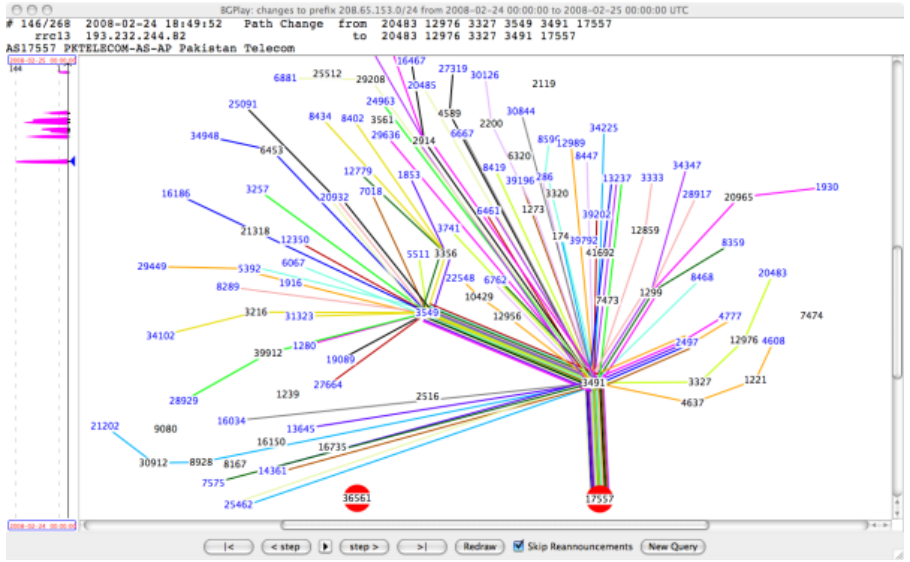
Example: YouTube Failure

- On February 24, 2008, the Pakistani government decides to ban YouTube nationwide
  - Pakistan Telecom (AS 17557) announces the prefix 208.65.153.0/24 via BGP, which actually belongs to YouTube (AS 36561)
  - This spreads worldwide within a few minutes
    - Result: YouTube requests are routed directly to Pakistan
      → YouTube is no longer available
  - YouTube announces two prefixes 208.65.153.0/25 and 208.65.153.128/25 after about 90 minutes
    - Due to the longest matching prefix, the correct routes are used again
  - The faulty routes are also artificially degraded until they are completely withdrawn after about 140 minutes

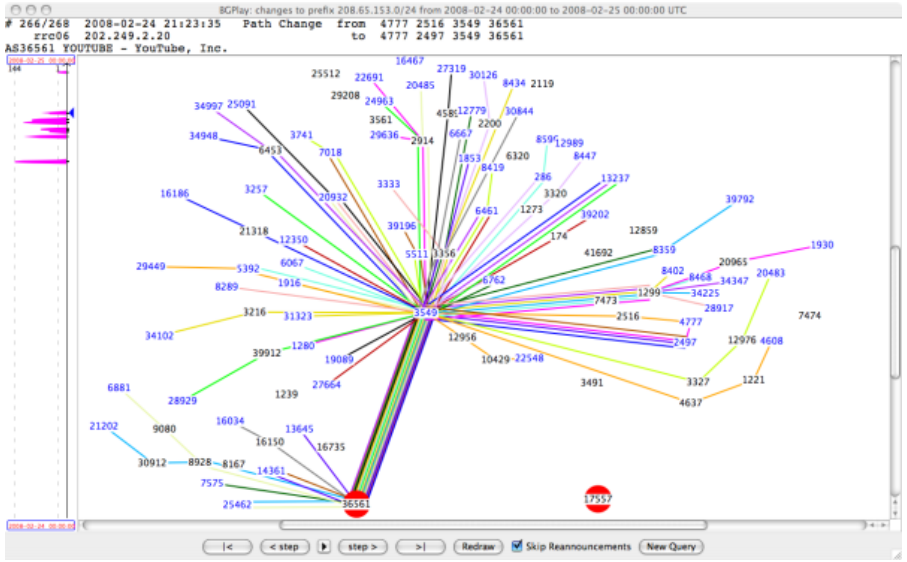
Example: YouTube Failure

- All this could be monitored via the Routing Information Service (RIS) of the European registrar RIPE
  - http://www.ripe.net/news/study-youtube-hijacking.html

https://youtu.be/IzLPKuAOe50



The routes to YouTube point to Pakistan Telecom (AS 17557)



The routes to YouTube point back to the correct AS 36561

Problem: Parts of the Internet “hijacked”

- Initial situation
  - Configuration error at Chinese provider IDC China (AS 23724)
    - Announces routing responsibility for approximately 37,000 prefixes at the beginning of April 2010
      - Dell, CNN, Apple, Amazon.com, …
    - Temporarily also for German Telekom
      - However, in most cases, shorter path lengths already existed for these prefixes
- Result
  - Views of affected websites "landed" at IDC China
    - Not further resolvable there
  - Duration limited to only about 15 minutes
- Reason

  → Missing cryptographic security of routing information

Example: Hijacking Apples Prefixes

- Russian Provider Rostelecom announced IP prefixes of Apple
  - July 22nd, 2022, duration roughly 12 hours
- Apple typically announces IP prefixes through AS714
  - Typically 17.0.0.0/9
- Rostelecom announced through AS12389
  - Prefix 17.70.96.0/19 which is part of Apple´s 17.0.0.0/8 block
  - Route announced was propagated globally
  - Consequently, Apple´s data are flowing through Russian routers
  - Apple´s services were not disrupted
- Apple later announced even more specific prefix 17.70.96.0/21

[https://www.heise.de/news/Russischer-Provider-kapert-Apple-Adressraum-7193705.html]

Example: Redirected Traffic

- Traffic from an AS in Denver to an AS in Denver is routed through an AS in Iceland

Man in the Middle Hijacking

- Attacking AS redirects traffic to its "victim" over itself
  - Simply by announcing prefixes of victim
  - Made possible by
    - Lack of control within BGP
  - First described by Kapela
    - “Stealing the Internet” in 2008

- In 2013, Renesys was watching over 1,500 Man in the Middle attacks on different prefixes

- Problem
  - Such attacks are difficult to detect
    - How to decide if hops on the transmission path are legitimate nodes or belong to a man in the middle attack?

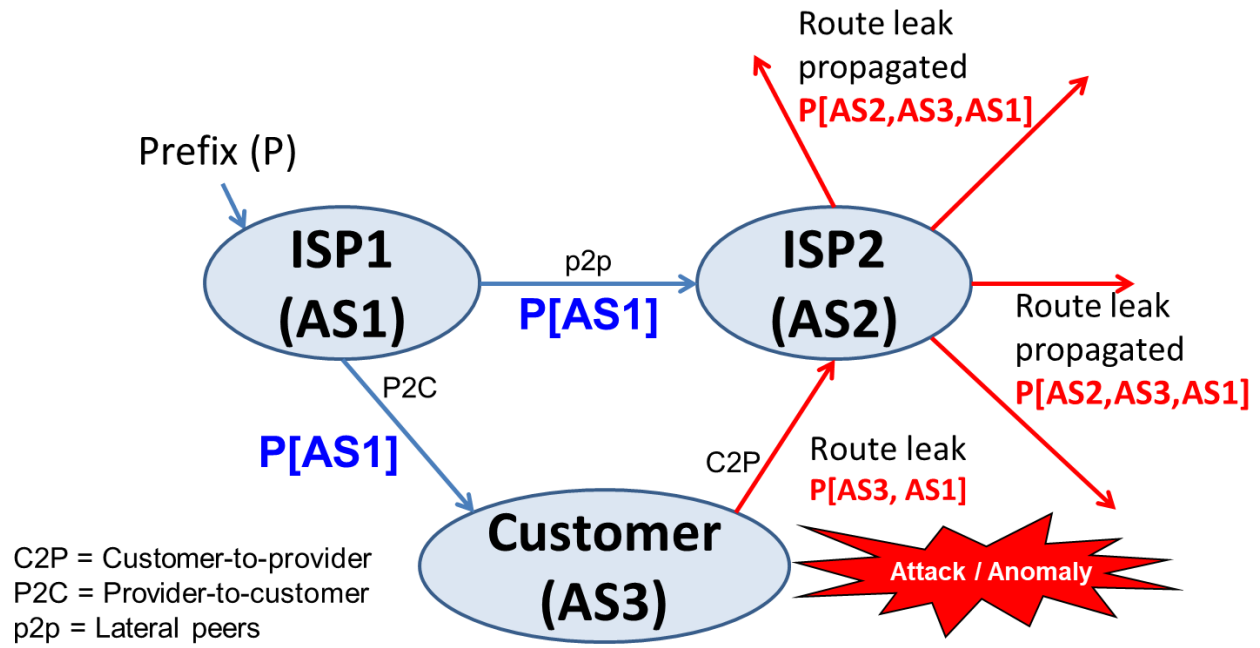
Man in the Middle Hijacking

- Problem
  - AS prefixes are easy to hijack
    - Lack of trust in assigning prefixes
    - There are no verifiable mappings between ASN and IP prefixes
- Counter measures
  - Current standard procedure, to protect against prefix hijacking
    - Alarm systems that monitor global routing information and report erroneous announcements of their own prefixes
    - Filter incoming routes from other ASs (time-consuming manual configuration)
- But
  - As long as not all ASs filter routes of their customers, problem persists!
  - Weakest link is enough for an attacker
- Possible solution
  - Establishment of a cryptographically secure chain of trust for routing information

3.6.2.4 Route Leak

Route Leak

- Description
  - Propagation of routing announcements beyond their intended scope
    - This is a policy violation
- Example
  - Multi-homed customer (AS3)
    - Learns update from one provider (ISP1)
    - Leaks update to another transit provider (ISP2)



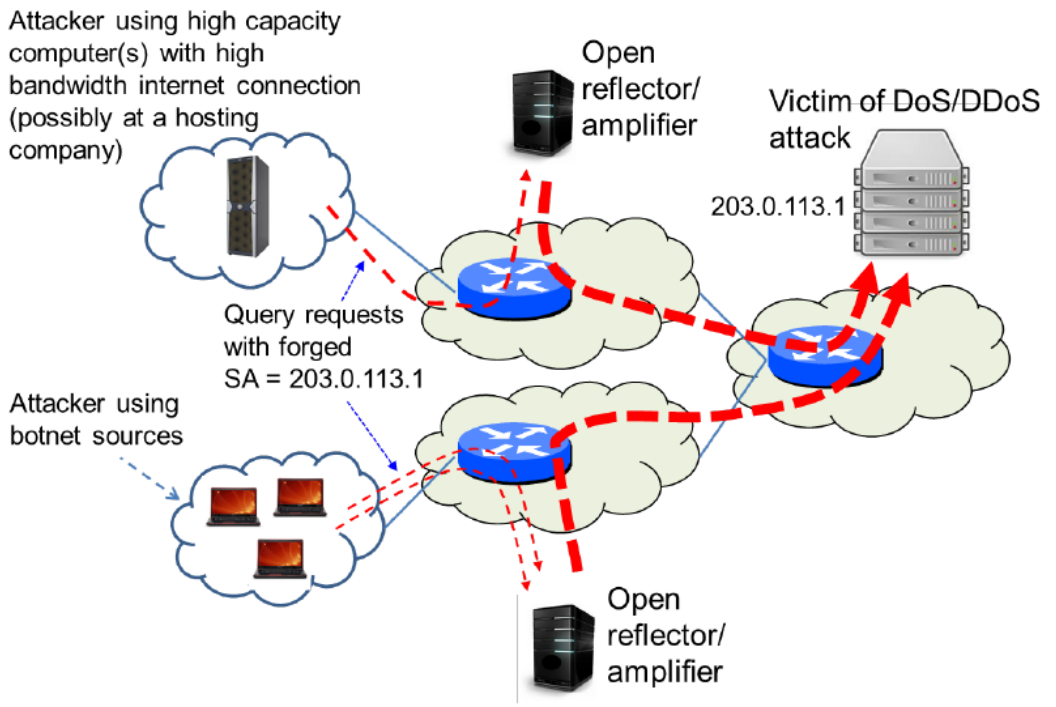
In general, ISPs prefer customer route announcements over those from others.

[NIST25]

3.6.2.5 IP address spoofing and reflection and amplification

IP Spoofing with Reflector/Amplifier

- Often
  - IP source address is spoofed to avoid traceability
  - Combined with reflection and amplification
- Possible result: DDoS
  - Distributed denial of service
  - Responses from Internet server are directed towards victim
    - Response is much larger than request



[NIST25]

3.6.2.6 Improving BGP Security and Resilience

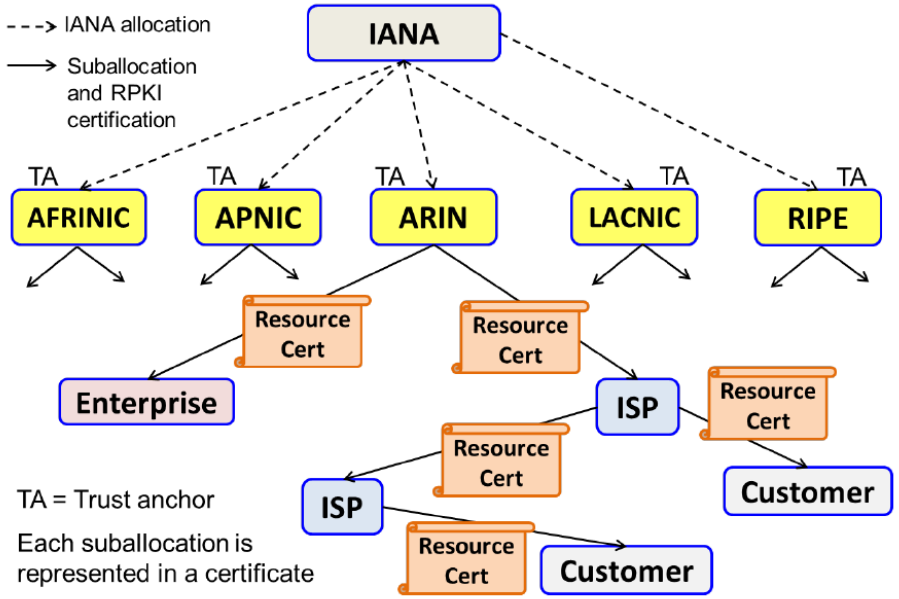
More details in NS

Solutions and Recommendations

- Registration of Route Objects
  - Internet routing registries (IRRs)
  - Maintained by RIR (regional Internet registries, e.g., RIPE)
  - Completeness, correctness, freshness, and consistency of data vary widely

Transport layer security is key to integrity of messages communicated in BGP sessions

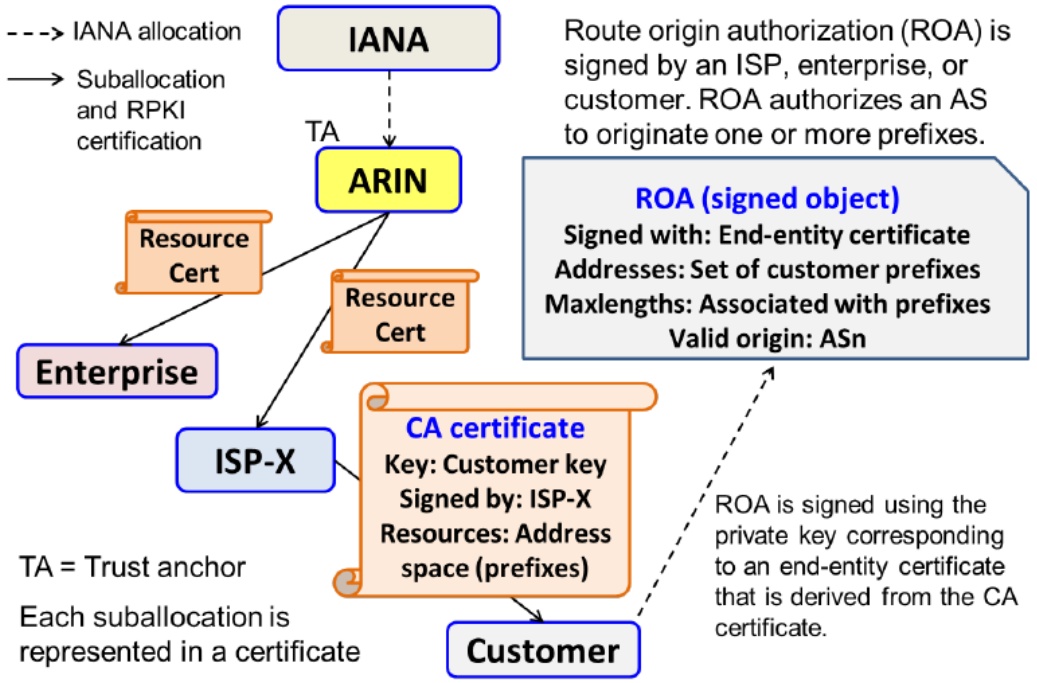
- Certification of resources in Resource Public Key Infrastructure (RPKI)
  - Cryptographically secured registries



[NIST25]

Solutions and Recommendations

- Route origin authorization (ROA)
  - Declares specific AS as an authorized originator of BGP announcements for the prefix
- ROA-based route origin validation (ROA-ROV)
  - Router checks whether advertised routes are valid



[NIST25]

Solutions and Recommendations

- Prefix filters
  - Only prefixes expected in a peering relationship are accepted
  - Should be implemented with respect to incoming and outgoing prefixes
- Source address validation
  - Typically in network edge devices, e.g., border routers
- …

[NIST25]

Secure BGP

- Extensions of BGP to secure the route information
  - soBGP (secure origin BGP) from Cisco
  - S-BGP (secure BGP) from BBN-Technologies

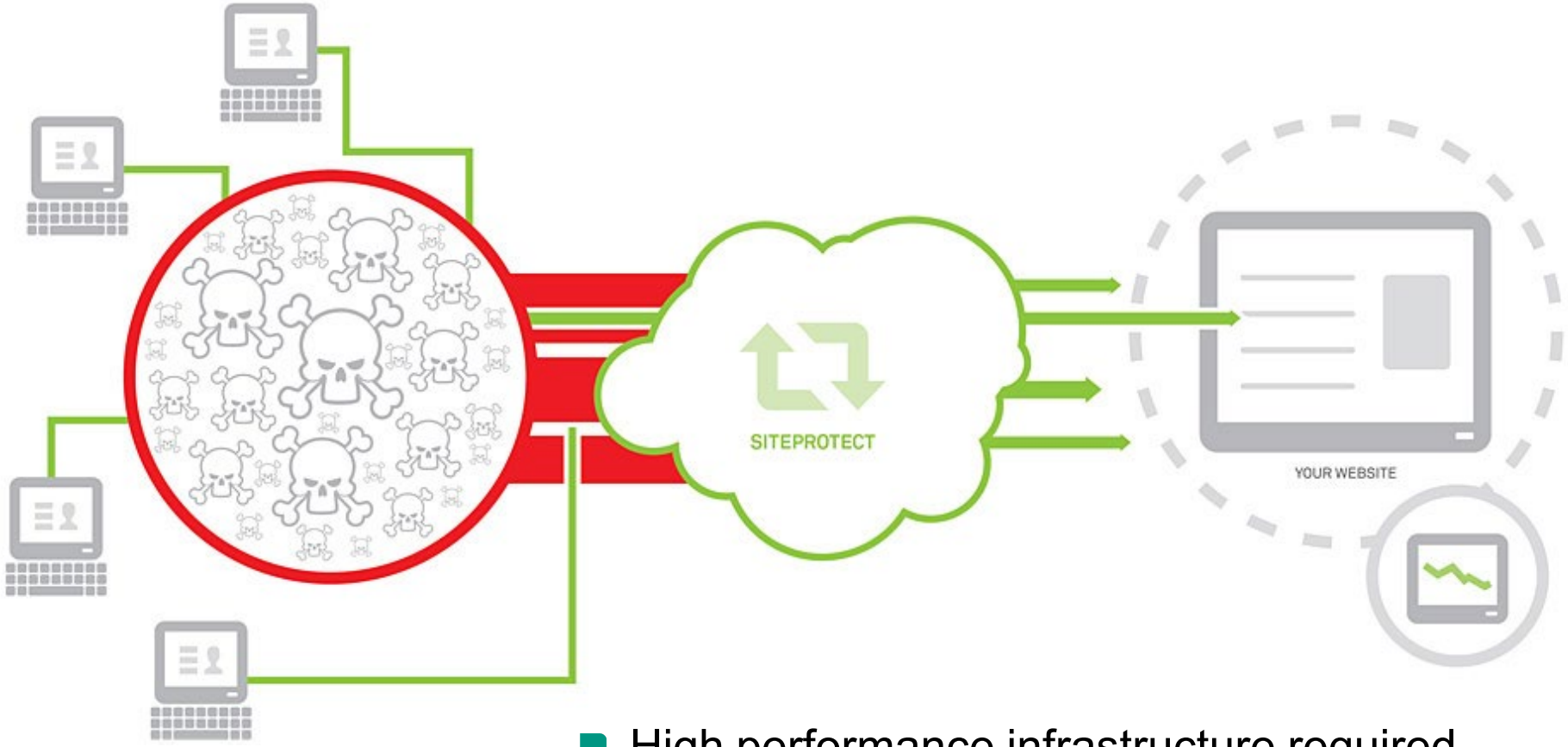
  - In principle, similar approaches
    - Authentication and authorization of senders
    - Integrity protection and encryption of the content of routing information
  - Nevertheless, many differences
    - For example, transfer of certificates in-band or out-of-band

[Hust05,Hust05a]

Dealing with DDoS Attacks

- Possible DDoS attack on an AS
  - Upstream with numerous (and large) packets
- The attacked AS ...
  - Has no way to protect against the heavy traffic
  - Can not filter attack from the incoming data stream and handle the legitimate requests
  - External help is necessary!
- Approach: "Cleaning Center"
  - Redirect traffic towards the attacked ASs to a "cleaning center" (Cloud)
    - "Cleaning Center" must announce the prefixes of the affected AS
    - In the "Cleaning Center" DDoS attack traffic and legitimate traffic are separated
    - Legitimate traffic is then routed back to the AS via a "clean pipe"
- Providers of such services are e.g. Tata, Verisign, AT & T, Abor

„Cleaning Center“

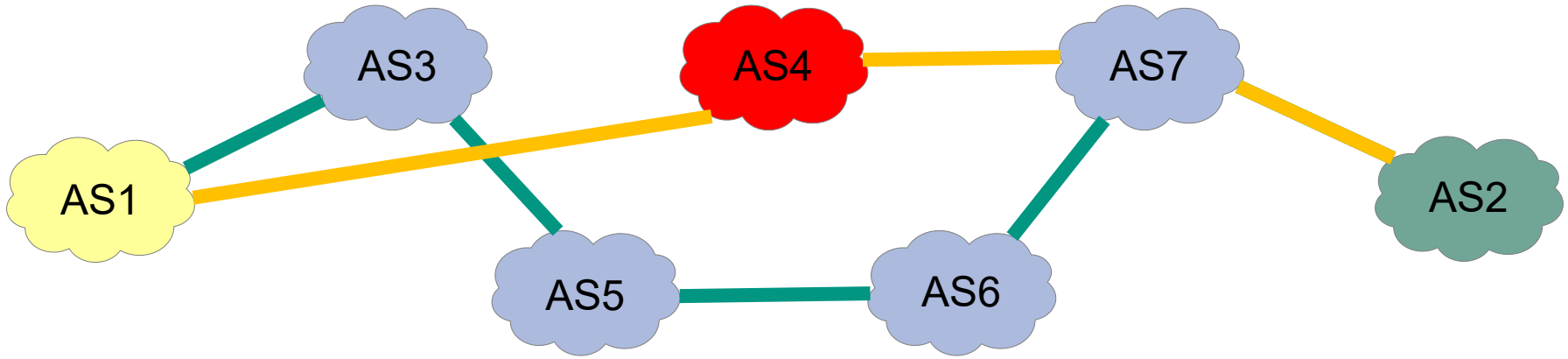


- High performance infrastructure required (cloud)
  - Traffic monitoring
  - Attack detection
  - Redirecting the traffic

DDoS Mitigation vs. Route Hijacking

- Path back to affected AS exists: the “clean pipe"
  - All traffic to an AS is routed via another AS
  - Does not disappear in the sink like with route hijacking

- Well…
  - ... this AS sees all packets, can delete, change or add new packets
    → strong position for a "man in the middle“ attacker



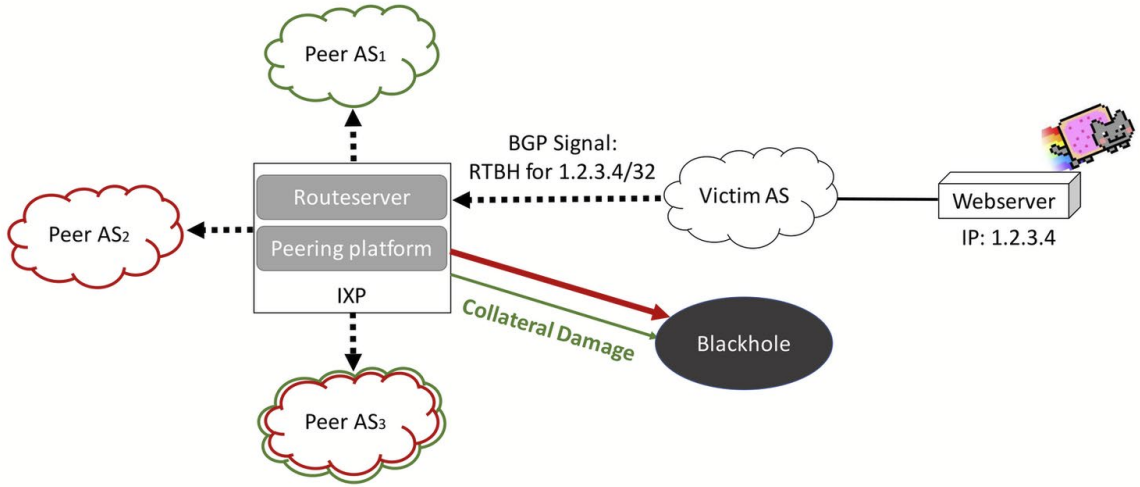
Remote Triggered Black Hole filtering (RTBH)

- Aka BGP Blackholing
- Goal: Early traffic mitigation in case of (D)DoS attack

- Operation
  - Announce an owned IP prefix to neighboring AS
    - Route server at IXP forwards RTBH message to neighboring ASes
    - ASes forward traffic to victim prefix to blackhole
  - Discard traffic towards this owned IP prefix
    - Reduces traffic towards own AS

- Problem
  - Also drops benign traffic
    - High false positive rates possible
    - Success depends on prefix length

Careful usage required



[https://labs.ripe.net/author/marcin_nawrocki/down-the-black-hole-dismantling-operational-practices-of-bgp-blackholing-at-ixps/]

PROBLEMS

1) What is the difference between control path and data path?
2) What information is stored in the routing table, what in the forwarding table?
3) Compare routing policy and routing metric. What are the similarities, what are the differences?
4) Which problems can occur if the routing is (solely) based on the current delay metric? How can these problems be solved?
5) Why does the Internet needs autonomous systems?
6) Can you get an AS number for your home network?
7) In which way can an autonomous system be classified (there might be several categories!)?
8) Is private peering free of charge? What about public peering?
9) What are the main benefits of public peering compared to transit?
10) What is an IXP?
11) Where does an ISP use interior gateway protocols? Where exterior gateway protocols?
12) What transport protocol does RIP use? What about OSPF?
13) Would it be possible to use RIP as the only routing protocol for the whole Internet? What about OSPF?

14) What is stored in the link state database?
15) Which information is gathered by the hello protocol?
16) What will happen if all control plane messages in an OSPF network start being dropped after the network was operational for some time? When will the network stop forwarding packets? When will the link-state databases be emptied?
17) Why does a new OSPF router synchronize its database before participating in OSPF LSA exchange? What would it have to do if the mechanism was not available?
18) How can BGP save routing messages via aggregation?
19) For what purpose do we need IBGP?
20) Do the SYN and SYN-ACK packets of a TCP handshake with a public server on the Internet visit (always) the same routers?
21) Why does the BGP routing process requires multiple (logical) tables such as Adj-RIB-In and Adj-RIB-out?
22) Is it possible to intercept all traffic towards and from an ISP without alerting the operators working at the ISP?
23) Why is the routing table of a BGP router in the default free zone so big?

Die von uns zur Erstellung der Folien genutzte

LITERATUR

Literatur

[HoWa06] Ch. Hopps, D. Ward; IS-IS for IP Internets; IETF, Working Group isis; http://www.ietf.org/html.charters/isis-charter.html

[Huit00] C. Huitema; Routing in the Internet; Prentice-Hall, 2nd Edition, 2000

[Hust05] G. Huston; An Operational Perspective on BGP Security; IETF 63; Aug 2005; http://www.ietf.org/old/2009/proceedings/05aug/slides/grow-2.pdf

[Hust05a] G. Huston; Securing Inter-Domain Routing; The ISP Column; March 2005; http://www.potaroo.net/papers/isoc/2005-03/route-sec-2-ispcol.pdf

[Hust06] G. Huston; Aggregation Reports; http://bgp.potaroo.net/as1221/bgp-active.html

[KhZi89] A. Khanna, J. Zinky; The Revised ARPANet Routing Metric; ACM Computer Communication Review, Vol 19, Issue 4, Sep 1989, pp. 45-56

[KuRo20] J.F. Kurose, K.W. Ross; Computer Networking – A Top-Down Approach; Addison Wesley, 6th Edition, 2020

[Medh17] Deepankar Medhi, Karthikeyan Ramasamy; Network Routing: Algorithms, Protocols, and Architectures; Morgan Kaufmann, 2nd Edition 2017

[NIST25] NIST Special Publication 800; Border Gateway Protocol Security and Resilience; initial public draft, January 2025

Literatur

[RFC1058] C. Hedrick; Routing Information Protocol; IETF, RFC 1058, Jun 1988

[RFC2328] J. Moy; OSPF Version 2; IETF, RFC 2328, Apr 1998

[RFC2439] C. Villamizar; Route Flap Damping; IETF, RFC 2439, Nov 1998

[RFC2453] G. Malkin; RIP Version 2; IETF, RFC 2453, Nov 1998

[RFC3765] G. Huston; NOPEER Community for Border Gateway Protocol (BGP) Route Scope Control; IETF, RFC 3765, Apr 2004

[RFC4456] T. Bates, E. Chen, R. Chandra; BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP); IETF, RFC 4456, Apr 2006

[RFC4989] D. Meyer, L. Zhang, K. Fall (Eds.); Report from the IAB Workshop on Routing and Adressing; IETF, RFC 4989, Sep 2007

[RFC5065] P. Traina, D. McPherson, J. Scudder; Autonomous System Confederations for BGP; IETF, RFC 5065, Aug 2007

[RFC7868] D. Savage at all; Cisco's Enhanced Interior Gateway Routing Protocol (EIGRP); IETF, RFC 7868, Mai 2016