



COMP9332

Network Routing & Switching

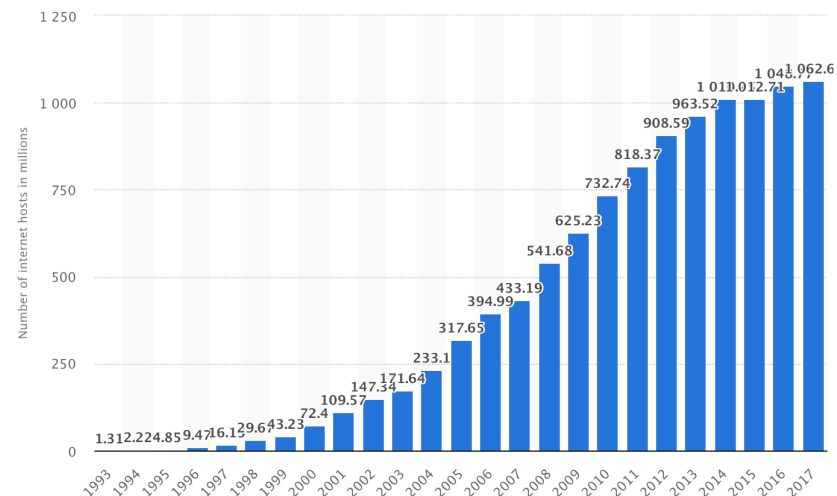
Inter-domain routing and BGP

■ www.cse.unsw.edu.au/~cs9332

The global Internet challenge (1)

- Must enable communication between any 2 hosts anywhere
- Given any IP address, any router must know how to route to that address
- The scale of the Internet means that some methods just **WILL NOT** work
 - E.g. With more than 1 billion hosts in today's Internet, host-based routing won't work. Problems:
 - Size of the routing table: memory requirement, look-up time
 - Route update overhead
 - Fault isolation becomes impossible

Number of worldwide internet hosts in the domain name system (DNS) from 1993 to 2017 (in millions)




The global Internet challenge (2)



- Key issue: Scalability
- Solutions you've seen
 - Hierarchical IP addressing
 - CIDR address aggregation
 - Network based routing rather than host based routing
- Hierarchy is a powerful way to provide scalable solutions
- Internet hierarchy: hosts, networks, autonomous systems

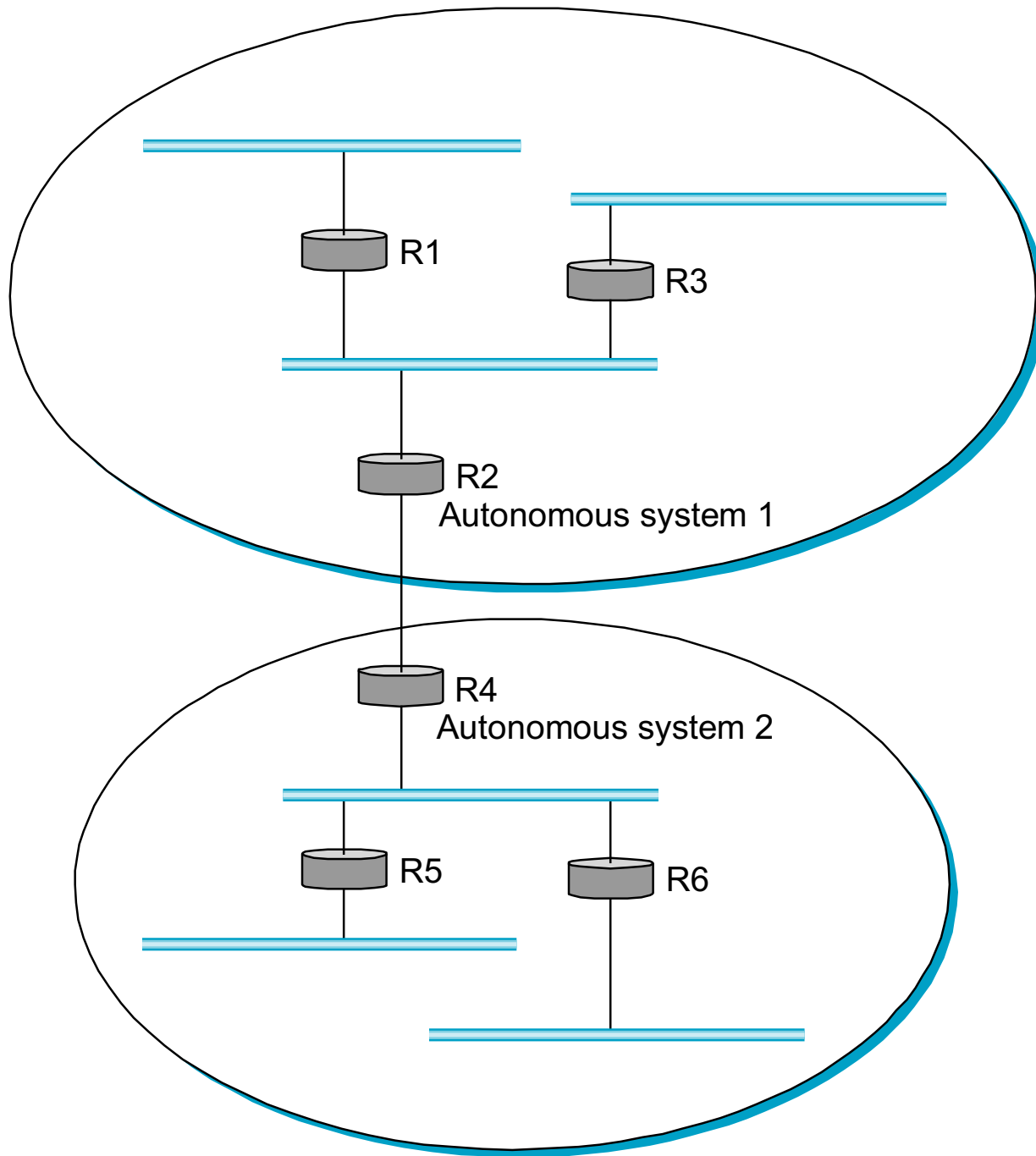
Outline

- 
- Organization of the global Internet
 - Inter-domain routing and BGP
 - Inter-domain traffic engineering
 - How multihomed and transit autonomous systems control their traffic

Internet organization



- Organized into hierarchy of Autonomous Systems (ASs) and networks
 - The number of unique autonomous systems of the Internet exceeded 5,000 in 1999, 30,000 in late 2008, 54,000 in mid-2016 and 60,000 in early 2018
 - consists of one or more networks [picture next page]
 - is administered by a single **administrative** authority
 - Examples: ISPs, company networks, university networks



AS traffic types



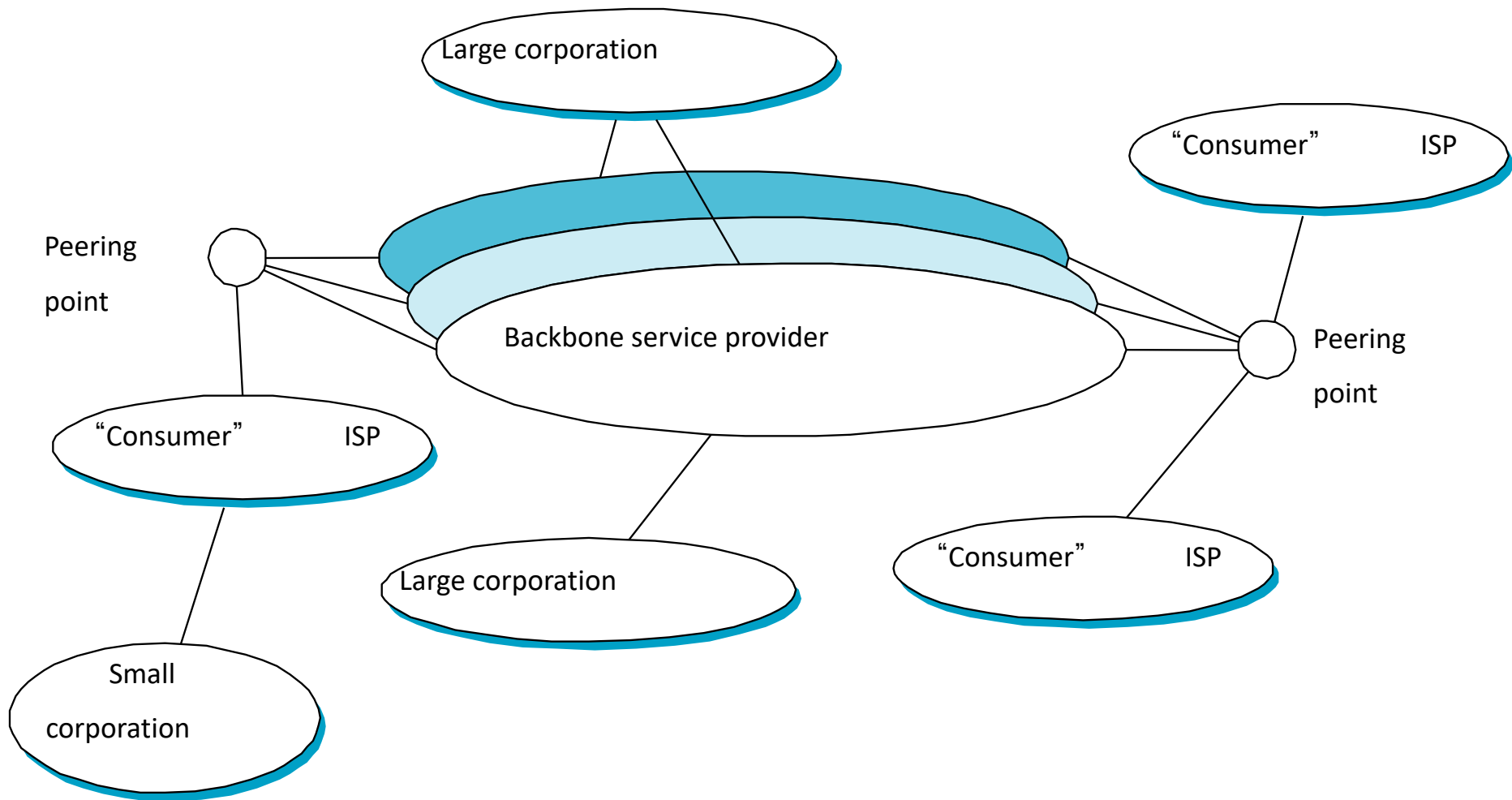
■ Local traffic

- Traffic local to an AS, either
 - » Originates from the AS, or
 - » Destined for the AS

■ Transit traffic

- Not local traffic

Interconnection of ASs



Classification of ASs



■ Stub AS

- One connection to another AS
- Carries only local traffic

■ Multihomed AS

- Connects to more than one AS but **does not** carry transit traffic

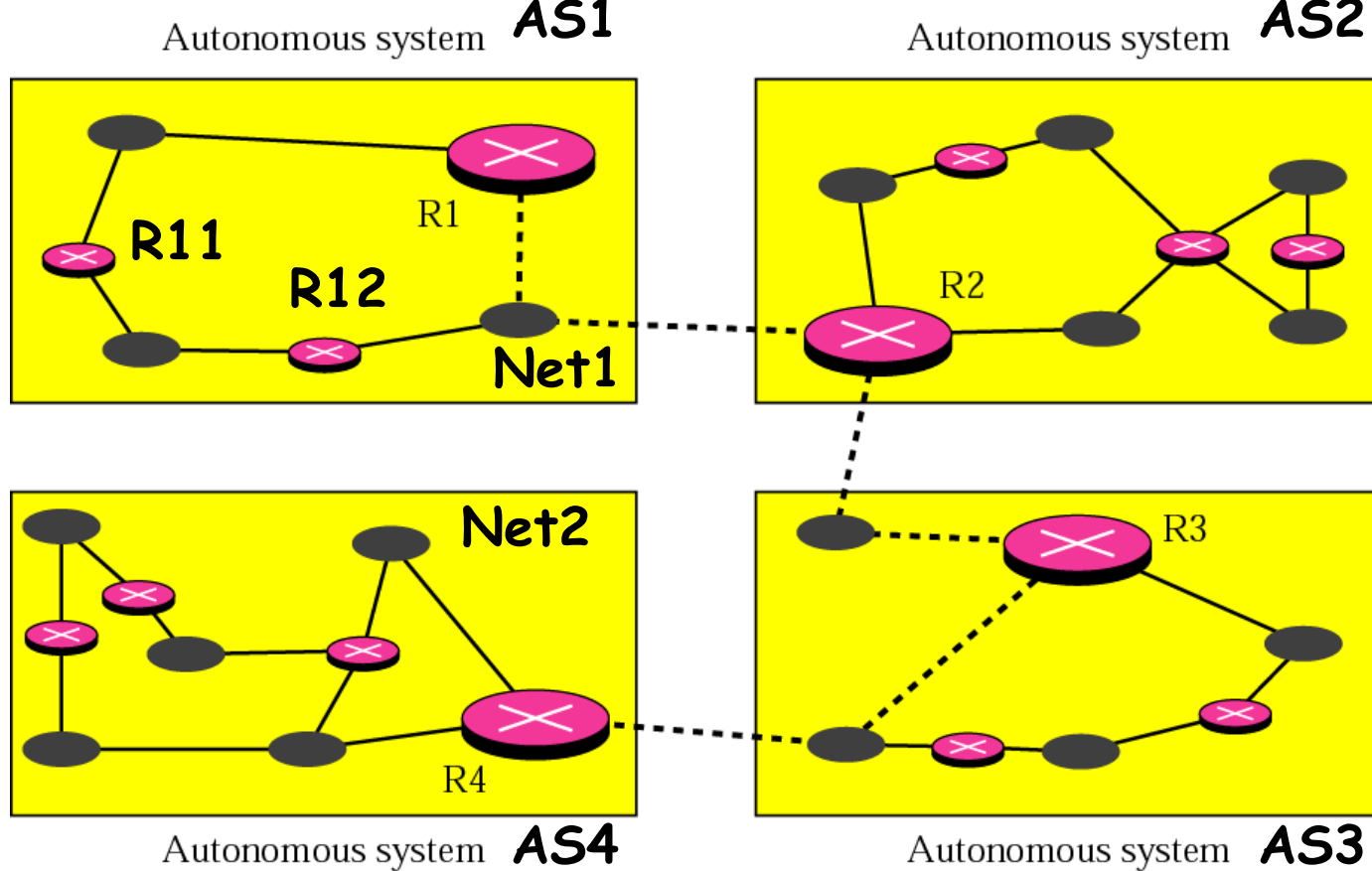
■ Transit AS

- Connects to two or more AS
- Carries both local and transit traffic

ASs and routing (1)



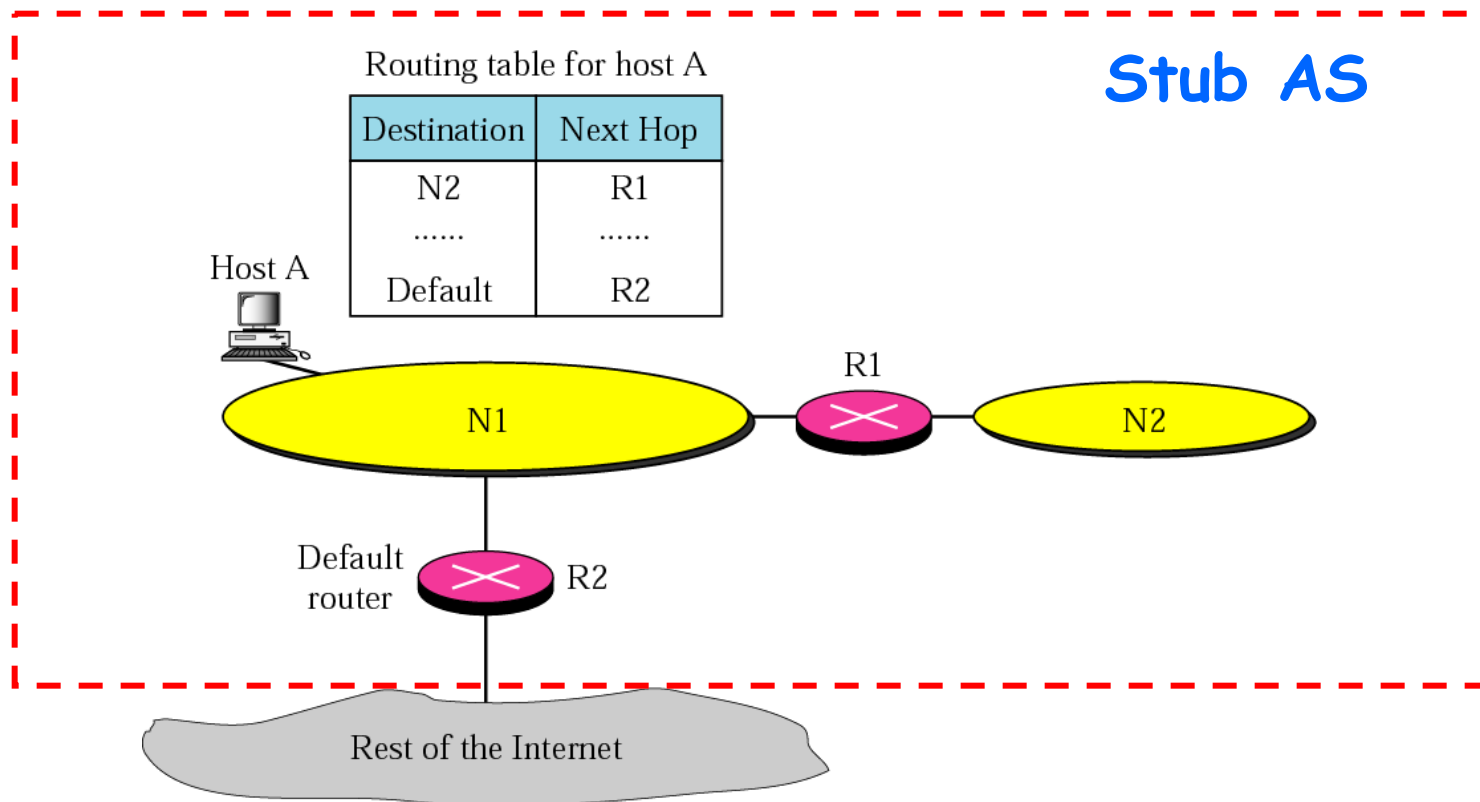
- ASs allow us to divide Internet routing into 2 sub-problems
 - Intra-domain routing: within an AS
 - Inter-domain routing: between ASs
- The AS model allows these two routing problems to be decoupled
 - An AS can choose any intra-domain routing protocol it prefers



- R11 learns from **intra**-domain routing protocol that the best route to Net1 is via R12
- AS1 learns from **inter**-domain routing protocol that the best route to Net2 is via AS2
 - Information about Net2 is propagated by inter-domain routing protocol to reach the other ASs

Routing in stub ASs

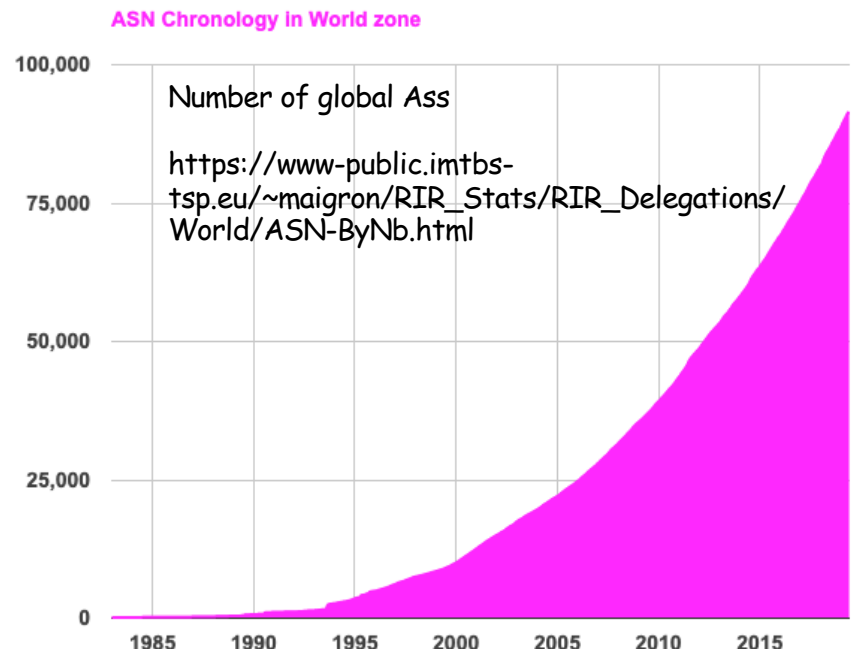
- For traffic to the rest of the Internet
 - A stub AS can use default route
 - The AS model allows route summarization
- The stub AS has to advertise itself via an inter-domain routing protocol to the rest of the Internet



Inter-domain routing

- Internet currently uses the Border Gateway Protocol version 4 (BGP-4) as its inter-domain routing protocol

- Scale of the problem:
 - More than 91000 ASs in June 2019
 - More than 784678 network prefixes (size of global routing table on 20 June 2019)




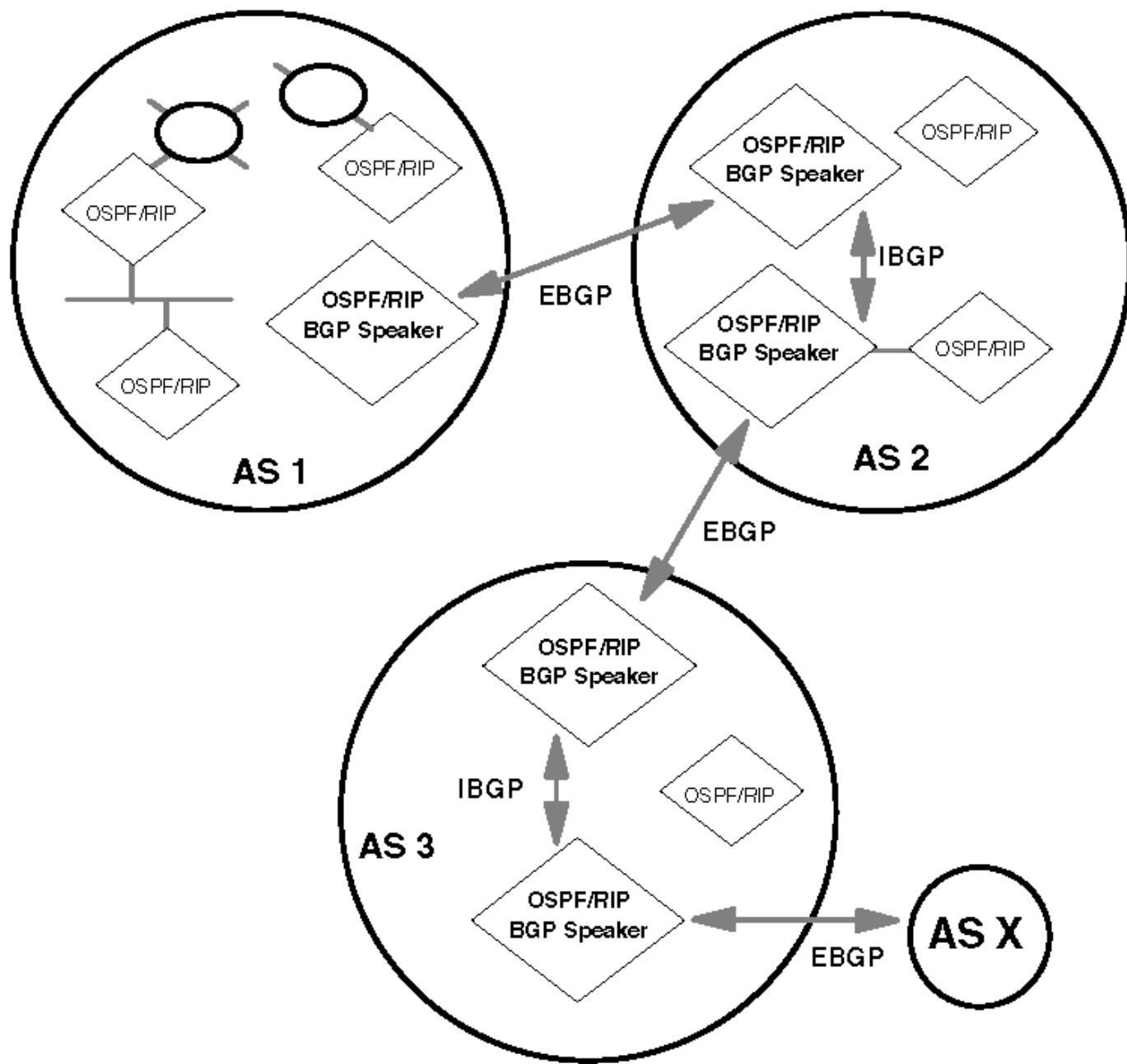
Challenges and features of BGP



- Only attempts to find a good enough loop-free path. Doesn't try to find the optimal.
- Concerns only with reachability: “You can reach network X via AS Y”
- Need to take routing policy into account
- Problem of trust and router misconfiguration

BGP basic

- 
- Each AS is identified by an AS number
 - Each AS has one or more routers running BGP
 - These routers are known as BGP speakers
 - BGP neighbours: A pair of BGP speakers that exchange routing information [illustration next page]
 - Internal (IBGP) neighbours
 - » A pair of BGP speakers within the same AS
 - External (EBGP) neighbours
 - » Two BGP speakers from two different AS

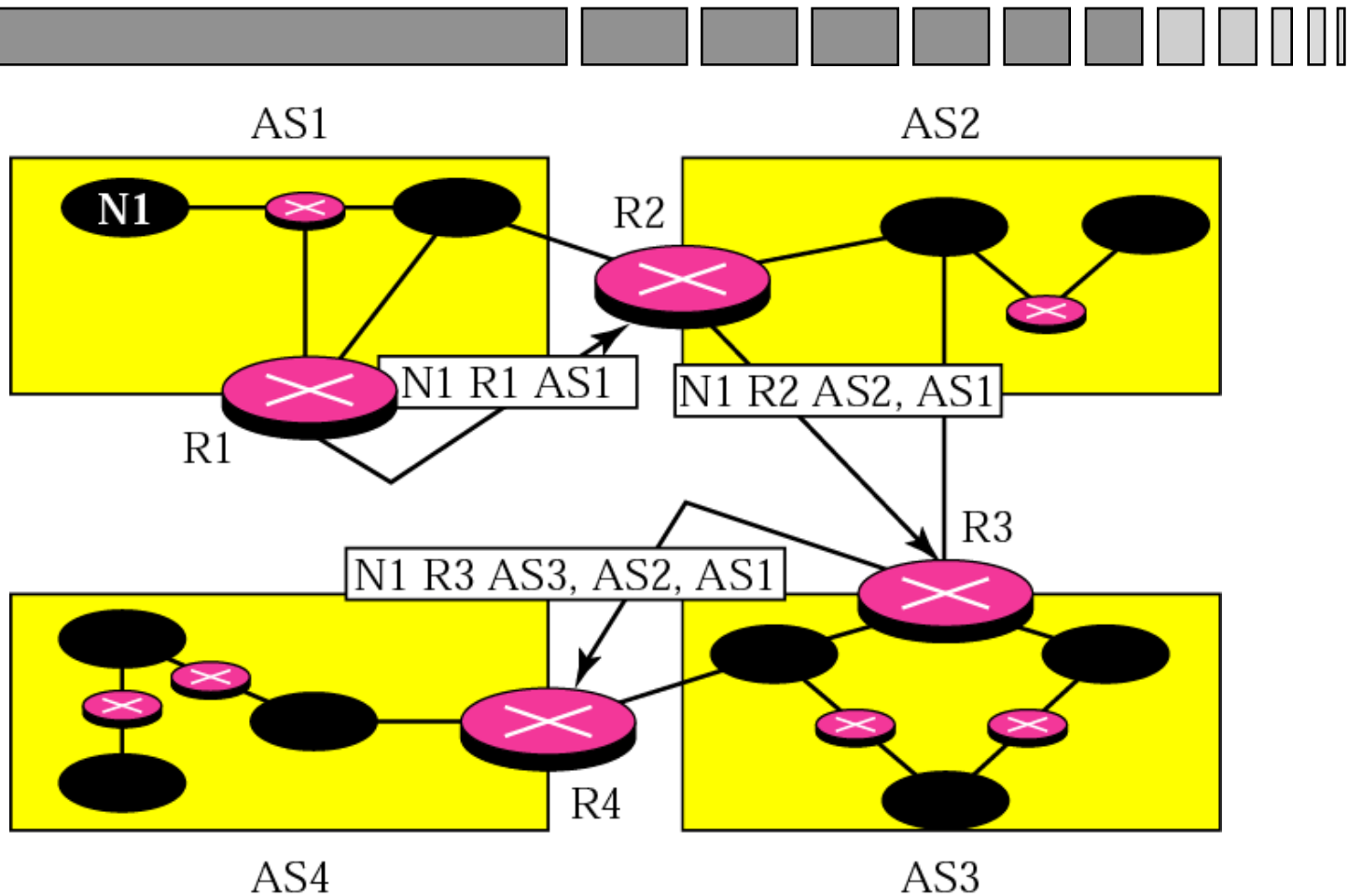


Path vector routing



- BGP uses a path vector to describe a route
- A path vector consists of
 - Destination network prefix
 - Next hop
 - A series of ASs which specifies a path leading to the destination
 - And some other attributes

Path vector (example)



Path attributes



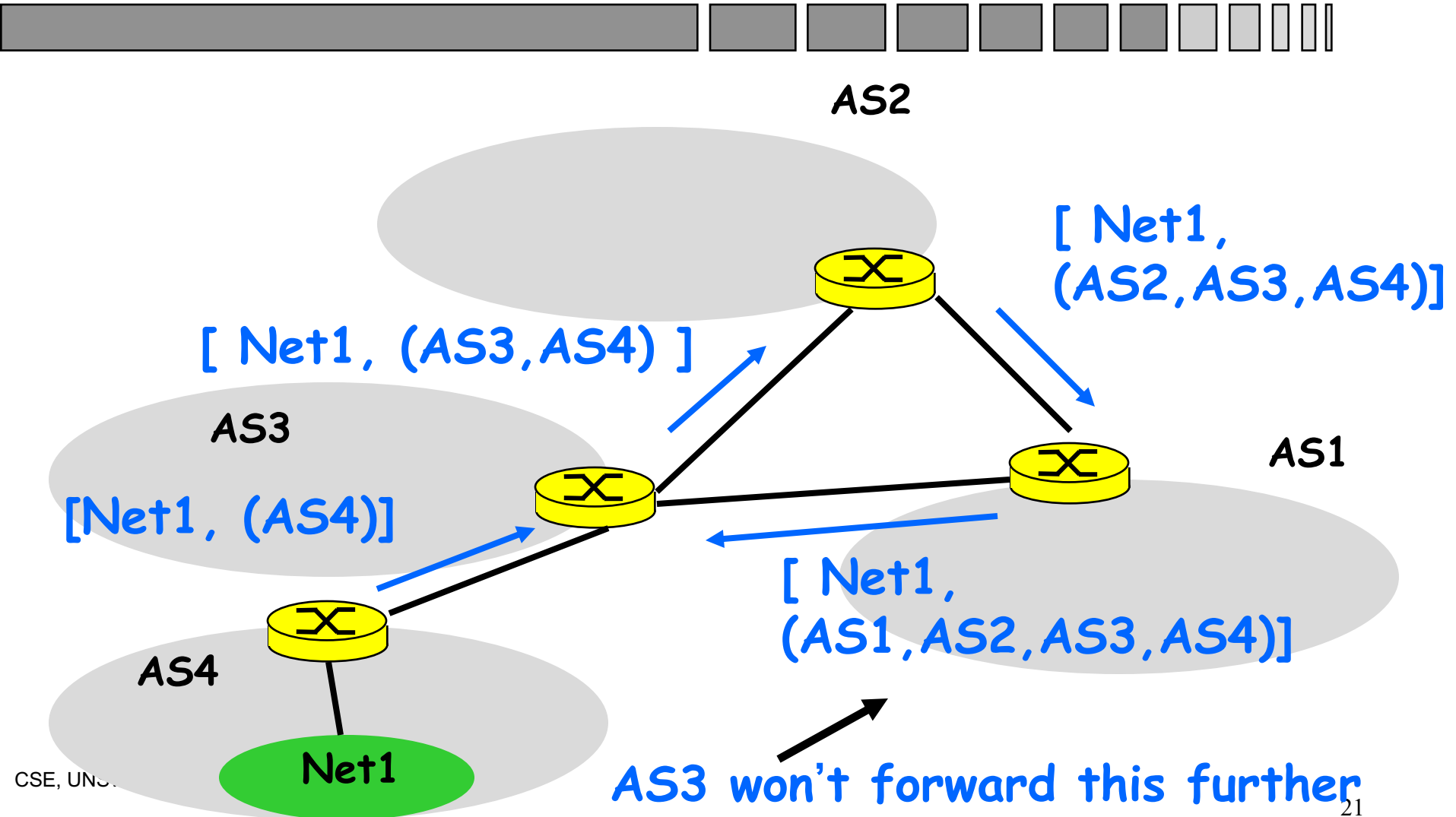
- In BGP, a path vector is defined by path attributes
- A path to a given destination has two main attributes
 - Next_HOP
 - AS_PATH
- Other attributes: local_pref, community
- Next_HOP is the IP address of the EBGP router in the remote AS
- AS_PATH is a list of AS numbers

The use of path vector



- Routing loop can be easily detected
- Enables policy routing

Routing loop prevention




BGP policy routing



- BGP speakers receive a lot of path-vector advertisements from its EBGP neighbours
- Advertisement processing involves 3 steps
 - Import policies
 - » Is the route useful to me? If not, eliminate it.
 - Path selection
 - » Which route should I choose?
 - Export policies
 - » Which route should I export to my EBGP neighbours?

Policy routing

- 
- Policies can be dictated by
 - Presence and absence of agreement to exchange traffic
 - Trust
 - Network preferences
 - Examples of policy
 - An AS may consider AS1 not secure
 - An AS may prefer to send its packets via AS1 rather than AS2 because the former charges a lower tariff

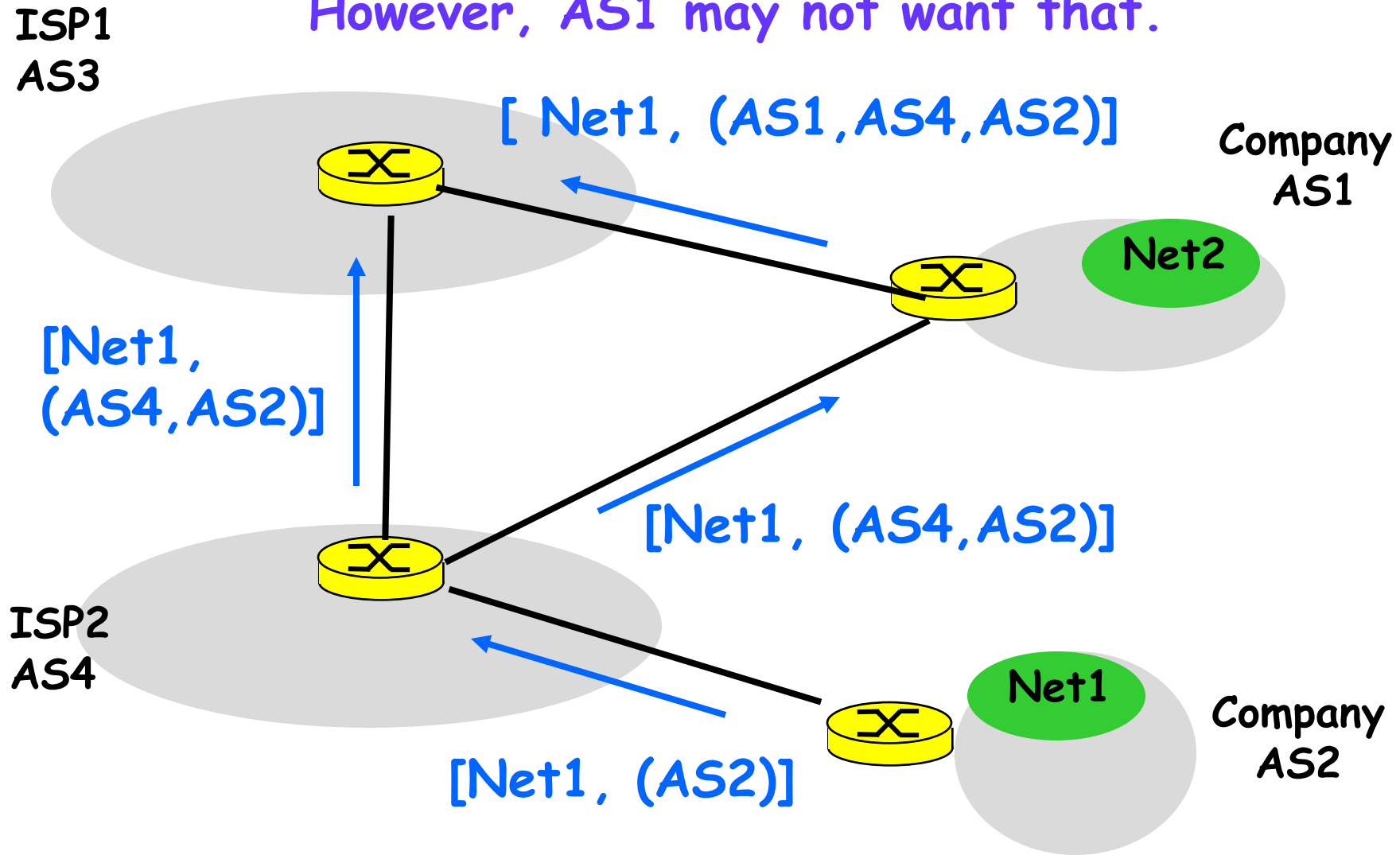
Path selection



- Policy routing and path vector
 - When a BGP speaker receives a path vector from its neighbour, it examines the list of ASs
 - If the path vector has one or more undesirable AS, it is dropped (import policy)
 - After that, other factors such as hop count (in terms of the number of ASs) come into consideration
 - The chosen path is entered into the BGP routing table
- We will examine path selection in detail later

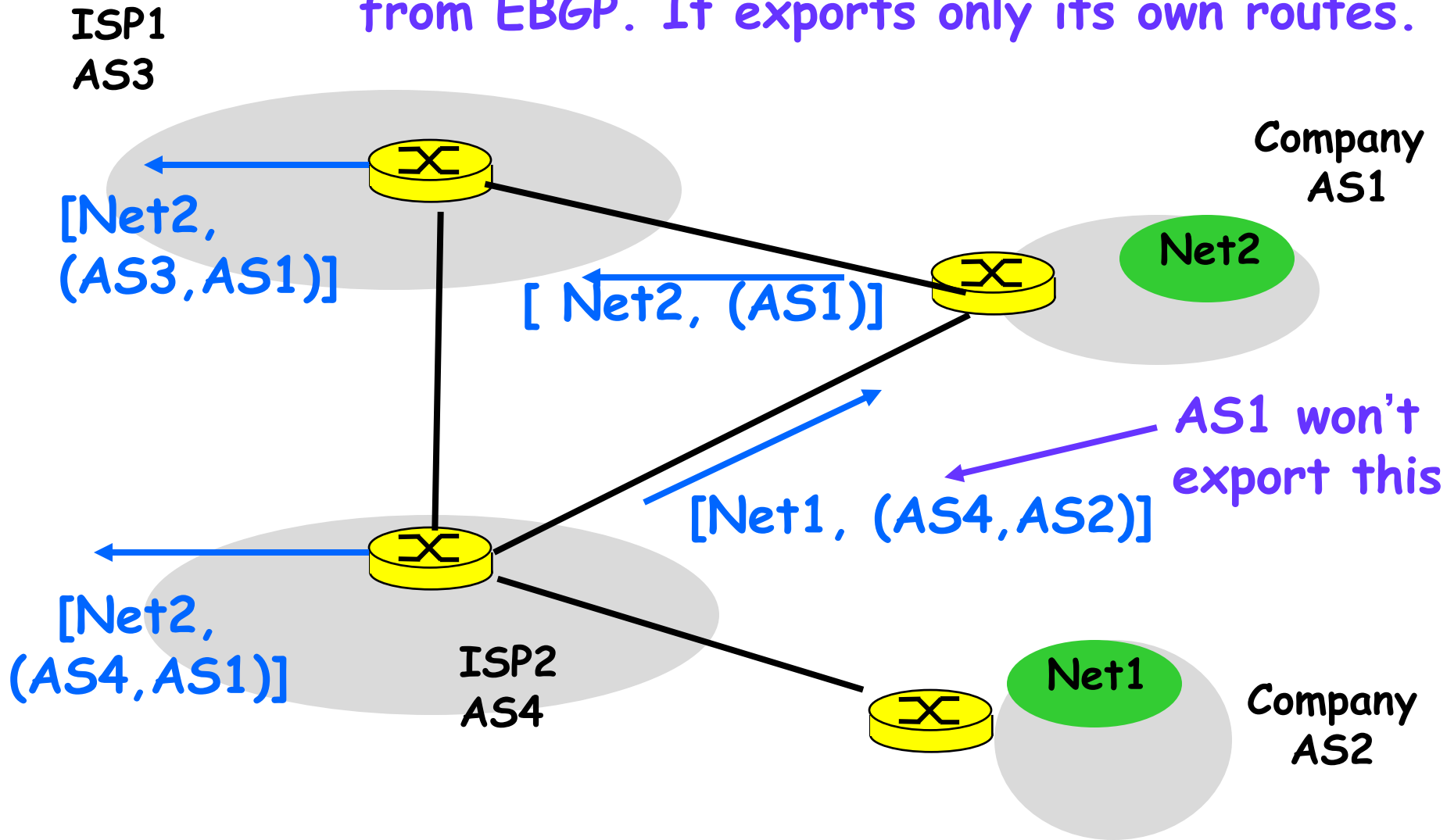
Using import and export policy (1)

If AS1 forwards this to AS3, then ISP1 may use AS1 to reach Net1 if it chooses. However, AS1 may not want that.



Using import and export policy (2)

If AS1 doesn't want to become a transit network, it does not export routes learnt from EBGP. It exports only its own routes.



Using import and export policy (3)



- The previous 2 pages show an example of how a multi-homed network uses policy routing to prevent itself becoming a transit network
- Export policy
 - Determines what other ASs know about your AS
 - This is the key tuning knob for traffic control or inter-domain traffic engineering

Drawback of path vector



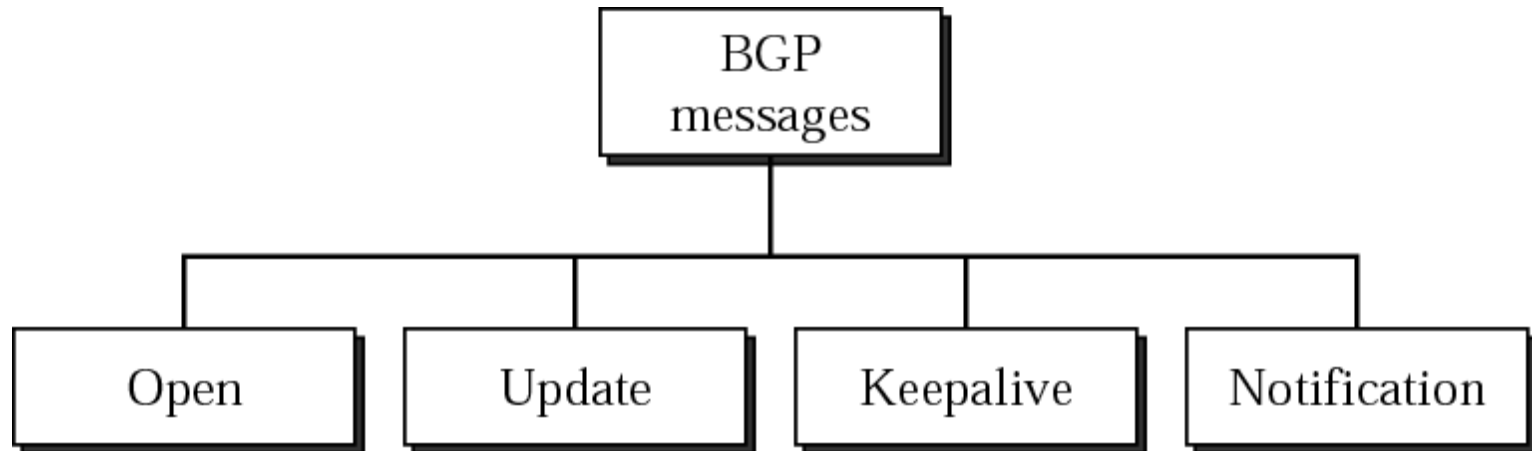
- Overheads in storing and transmitting path vectors. The overheads depend on
 - #network prefixes, length of AS_PATH, topology
- BGP has a number of features to reduce these overheads
 - E.g., **incremental updates** and **network prefix compression**
- Good news: The graph of Internet ASs has the small world property (small network span)
 - In 2000, nearly 6200 ASs but average distance between two ASs is only 3.76.

BGP protocol operations



- BGP establishes a reliable TCP connection between peers
- When two BGP speakers initially form a BGP session, they exchange their entire routing table
- After that, changes to the table are communicated as **incremental updates**
 - A BGP routing table can contain more than 500,000 entries!
 - » See <http://bgp.potaroo.net/> for up-to-date figures
 - The choice of TCP means the initial transfer is reliable, so only **incremental updates** are required
 - » This reduces transmission overhead

Types of BGP messages



BGP packet types



- BGP uses four types of packets
- Three of them are used to maintain the BGP connection, they are
 - OPEN
 - KEEP ALIVE
 - NOTIFICATION
- The last one, UPDATE, is used to exchange path vectors

BGP connection maintenance (1)



- BGP runs over TCP
- After the TCP connection has been established, BGP uses an OPEN packet to initiate a BGP session
- Once a connection is established, BGP keeps it

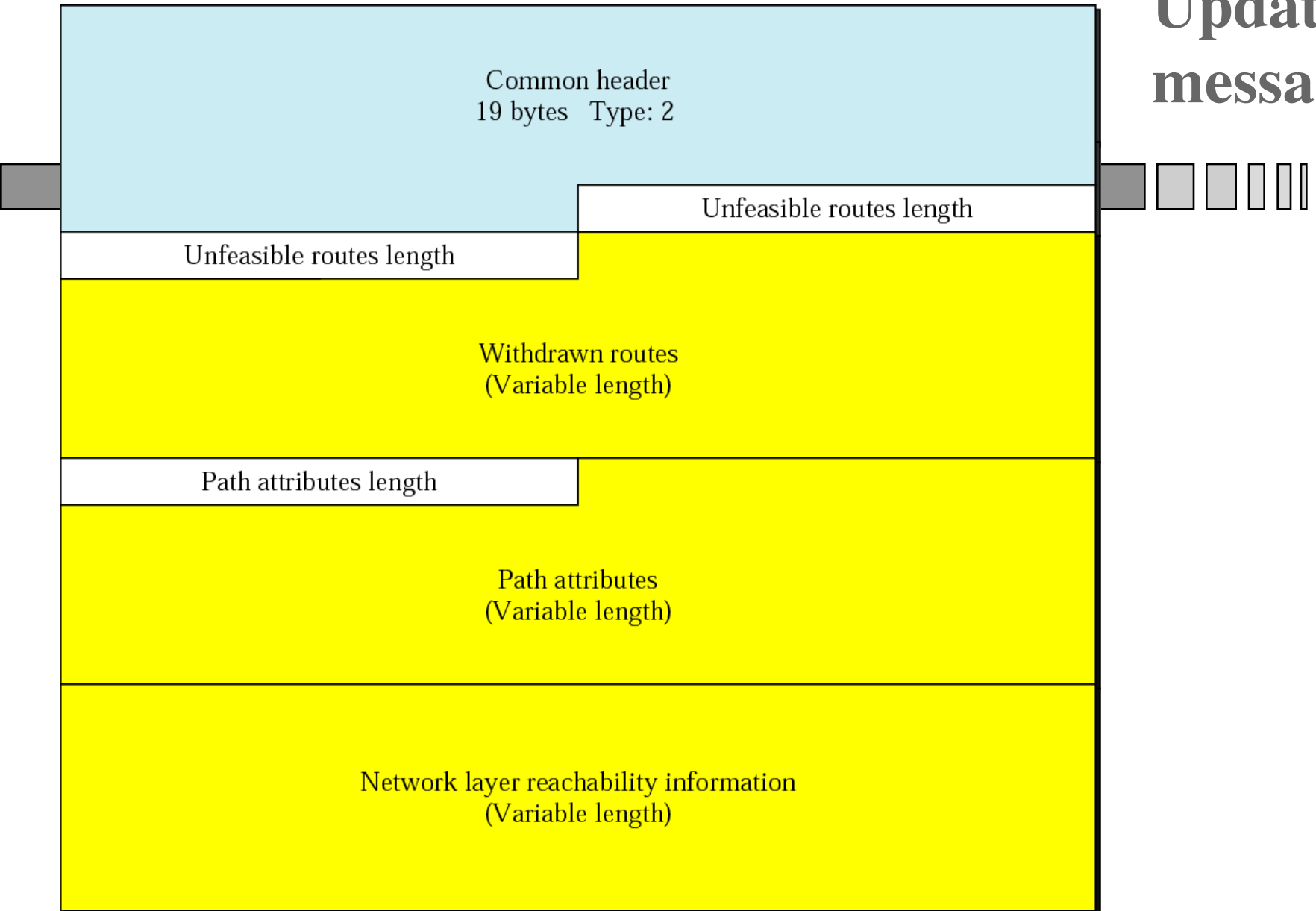
BGP connection maintenance (2)



- **KEEPALIVE** packets are regularly exchanged to keep the connection alive
- If no messages are exchanged for a period specified by **HOLDTIMER**, an error is assumed to have occurred
 - **NOTIFICATION** packet is sent & TCP connection closed
 - **HOLDTIMER** is specified in the **OPEN** packet

Figure 13-54

Update message



BGP UPDATE Message



- Used to advertise new routes or withdraw old routes
- To withdraw, list the IP network prefixes of destinations networks
- To add new routes, give path attributes of destination networks

BGP Update Message

How to code destination address



- To withdraw or to add a route, destination network address needs to be entered in update messages
- To support CIDR, a mask-address pair is needed
- A mask-address pair would occupy $4+4=8$ bytes uncompressed for IPv4
- BGP uses a **compression** technique to reduce the size of mask-address pairs

Mask-Address Compression (1)



- Two fields
 - Mask field
 - Network address field
- Mask field
 - Instead of specifying the mask, the prefix length is specified
 - Requires always 1 byte

Mask-Address Compression (2)



- Network address field
 - Contains only the non-zero octets of the network address
 - Length is variable
- Exercise: If a network address has a prefix length of 21 bits, how many bytes are required to code this mask-address pair
 1. Without using compression?
 2. With compression?

Mask-Address Compression (3)



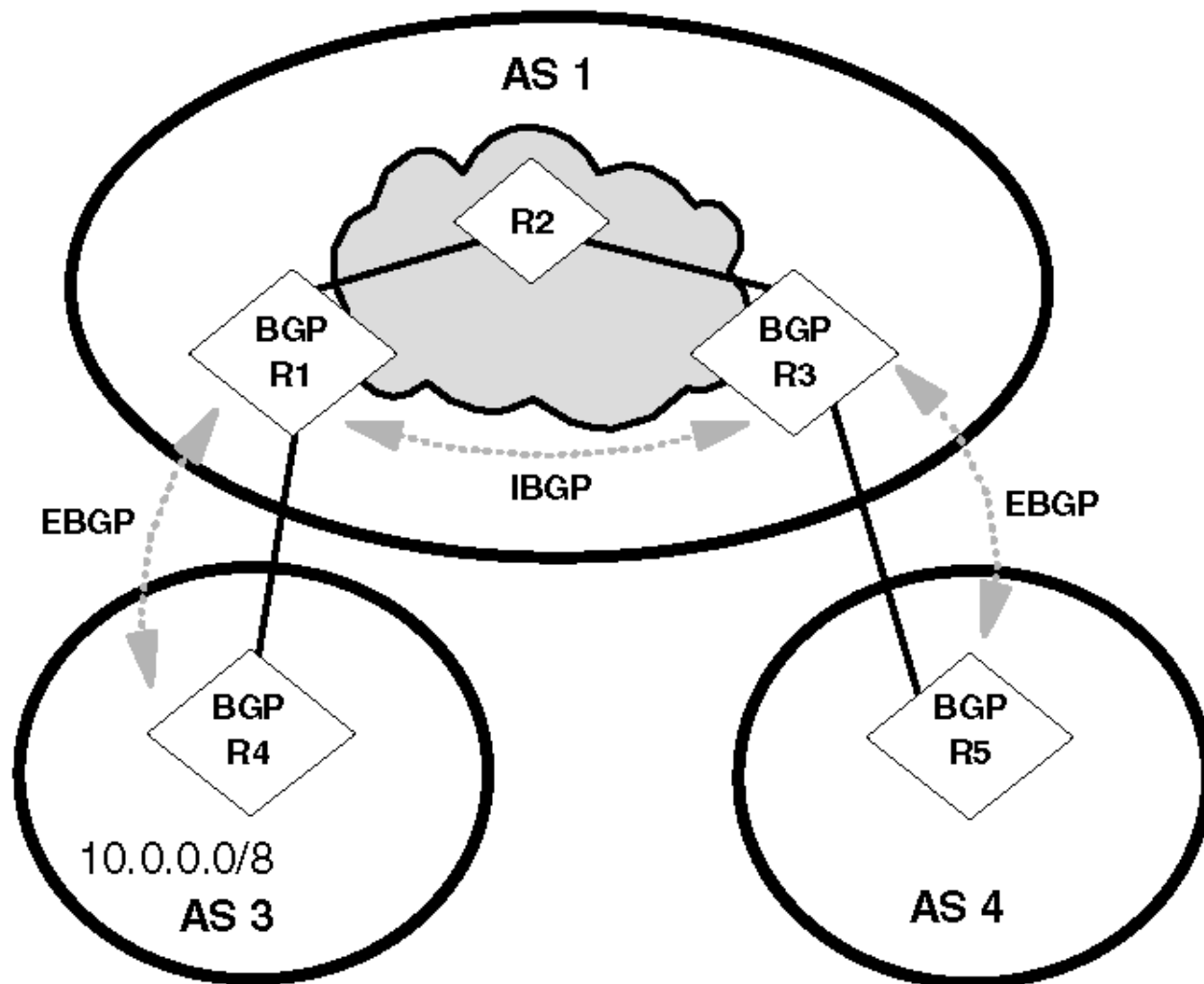
■ Solution:

1. 8 bytes without using compression
2. 4 bytes with compression
 - 1 byte for the mask field
 - 3 bytes for the address field
 - Since prefix length is 21, the first 3 bytes are non-zero, last byte is zero
- Compression reduces bandwidth overhead!

BGP synchronisation



- The problem (Refer to the figure on the next slide)
 - AS1 provides transit service between AS3 and AS4
 - Packets between R1 and R3 go through R2
 - R1 and R3 run BGP but not R2
 - R1 learns the route to 10/8 (in AS3), it exports it to R3 via IBGP
 - R3 exports the route to R5 using EBGP
 - Since R2 doesn't run BGP, it doesn't know about the networks in AS3
 - If R5 sends packets to 10/8 (in AS3), it must go through R2, but R2 won't know how to handle them, the packets will be dropped



BGP synchronisation (cont'd)



- BGP synchronisation requires a transit AS not to advertise a route to its neighbouring AS before its own internal routers have learnt about the route
- How can internal routers in an AS learn the external networks?

How routers learn external routes? (1)



- **Method 1:** Require all routers to run BGP
 - BGP requires there is an IBGP connection between every pair of BGP speakers in an AS
 - » Aka fully meshed set of IBGP connections
 - Questions:
 - » If an AS has N routers and all of them run BGP, how many IBGP connections are there?
 - » What is the implication for a large AS with many routers?

All routers running BGP



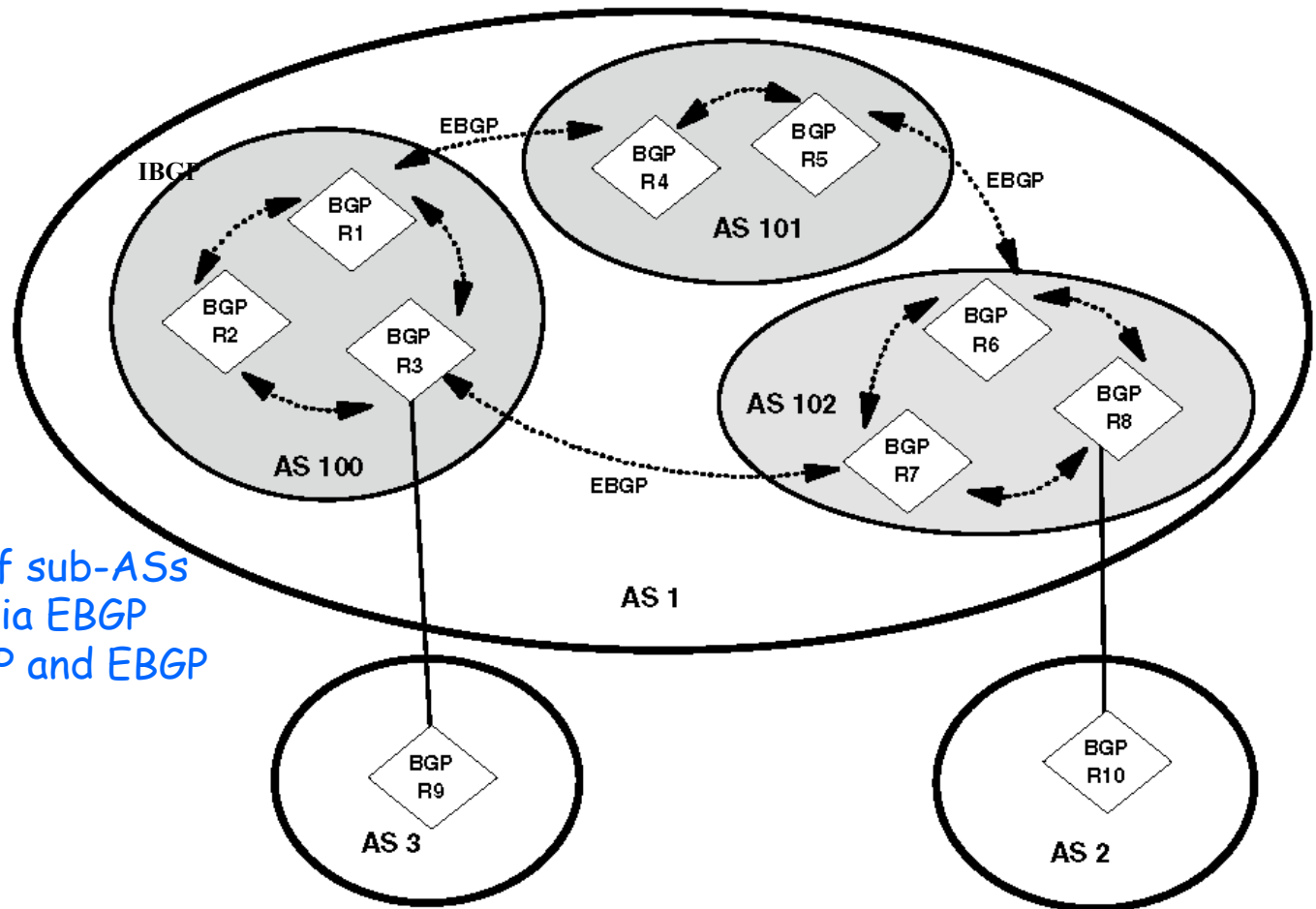
■ Answers

- $N(N-1)/2$
- High bandwidth overhead

■ If all routers are to run BGP, a large AS has two alternative configurations to reduce the number of IBGP connections

- BGP confederations
- BGP route reflector

BGP confederations



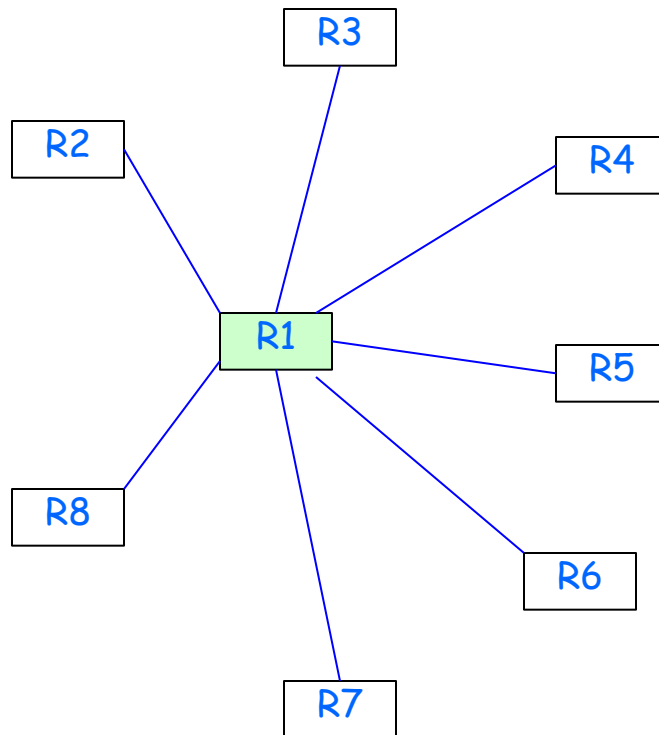
- Create a number of sub-ASs
- Connect sub-ASs via EBGP
- Contains both IBGP and EBGP

Route Reflector

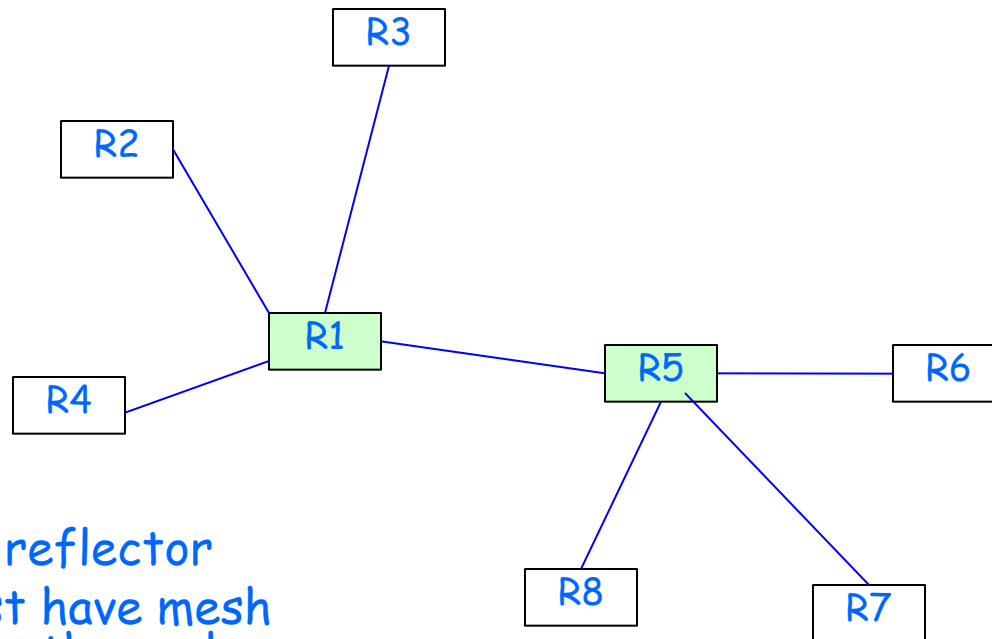


- IBGP speakers are not allowed to re-advertise routes learned from one IBGP speaker to other IBGP speakers
 - Hence the full-mesh connectivity
- This rule is relaxed for reflectors
- Reflectors are special IBGP speakers who are allowed to re-advertise
- By designating one or more IBGP speakers as reflectors, it is possible to avoid the burden of full-mesh connectivity

Single reflector



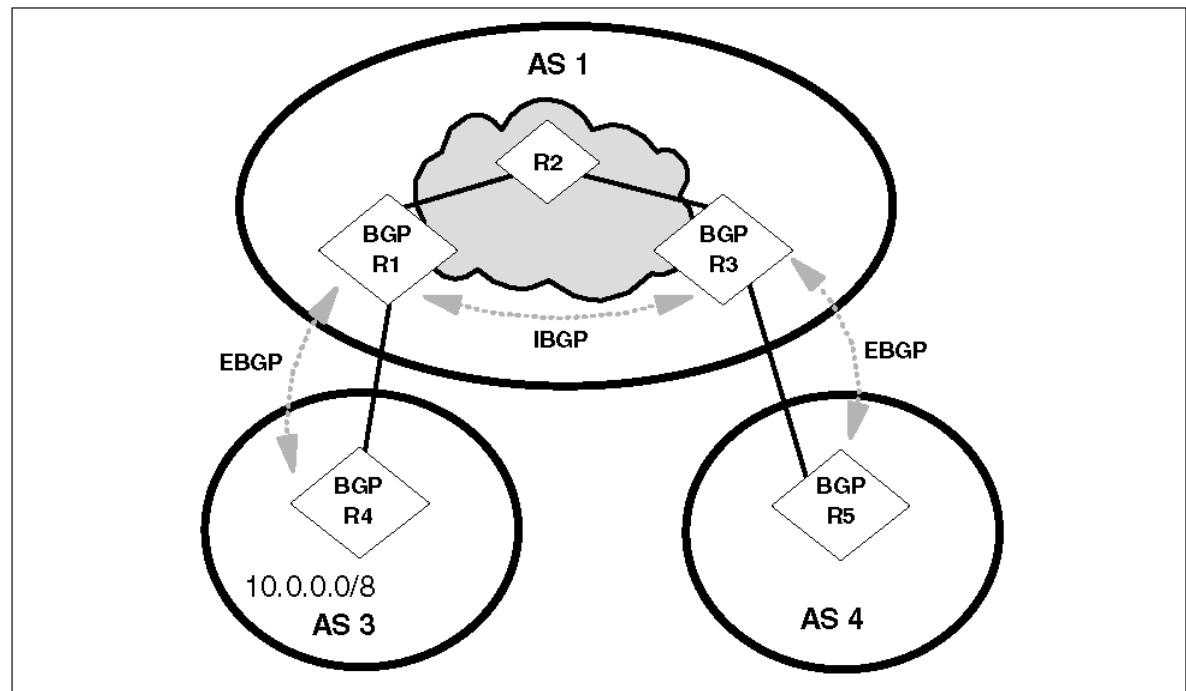
Multiple reflector



- Reduces load on a reflector
- All reflectors must have mesh connectivity between themselves

How routers learn external routes? (2)

- **Method 2:** Redistribute the external routes learnt from BGP into the AS using intra-domain routing protocol
 - Example: If AS1 uses RIP
 - » R1 adds an entry to the RIP table for 10.0.0.0/8
 - » This information will be propagated to its neighbours at the next RIP update (RIP tables will become very large)



Some advanced topics



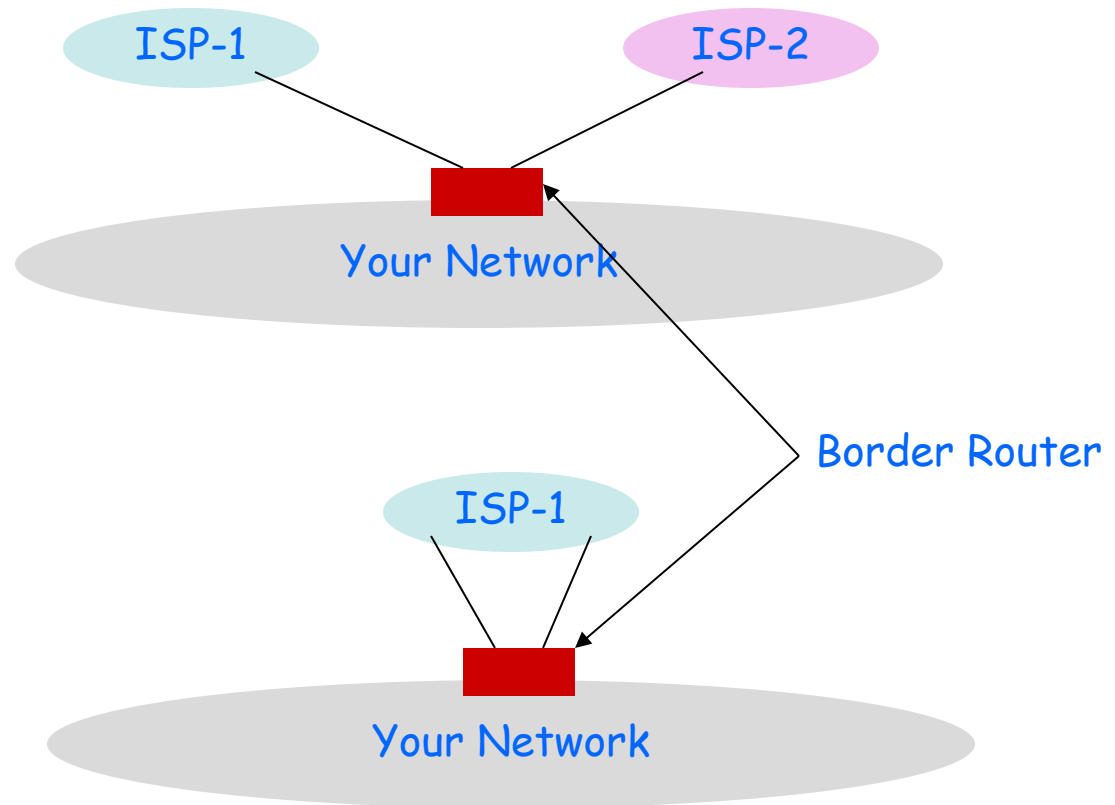
- Multihoming and Inter-domain traffic engineering
- Problems of BGP
 - Increasing use of multi-homing
 - Trust and router misconfiguration
- Other problems (not covered in COMP9332)
 - Route oscillation etc.



Multihoming

What is Multihoming

- Your network may be connected to the Internet via more than one physical connection
 - Different ISPs
 - Same ISP, but different kinds of links (e.g. ADSL, T1)



Why Multihoming



- Primarily for reliability
 - Usually two connections, one in use, other work as backup if the main connection fails
 - Very expensive to idle a backup connection
- Use both connections simultaneously
 - Offers load balancing
 - More cost effective solution

Routing over Multihoming



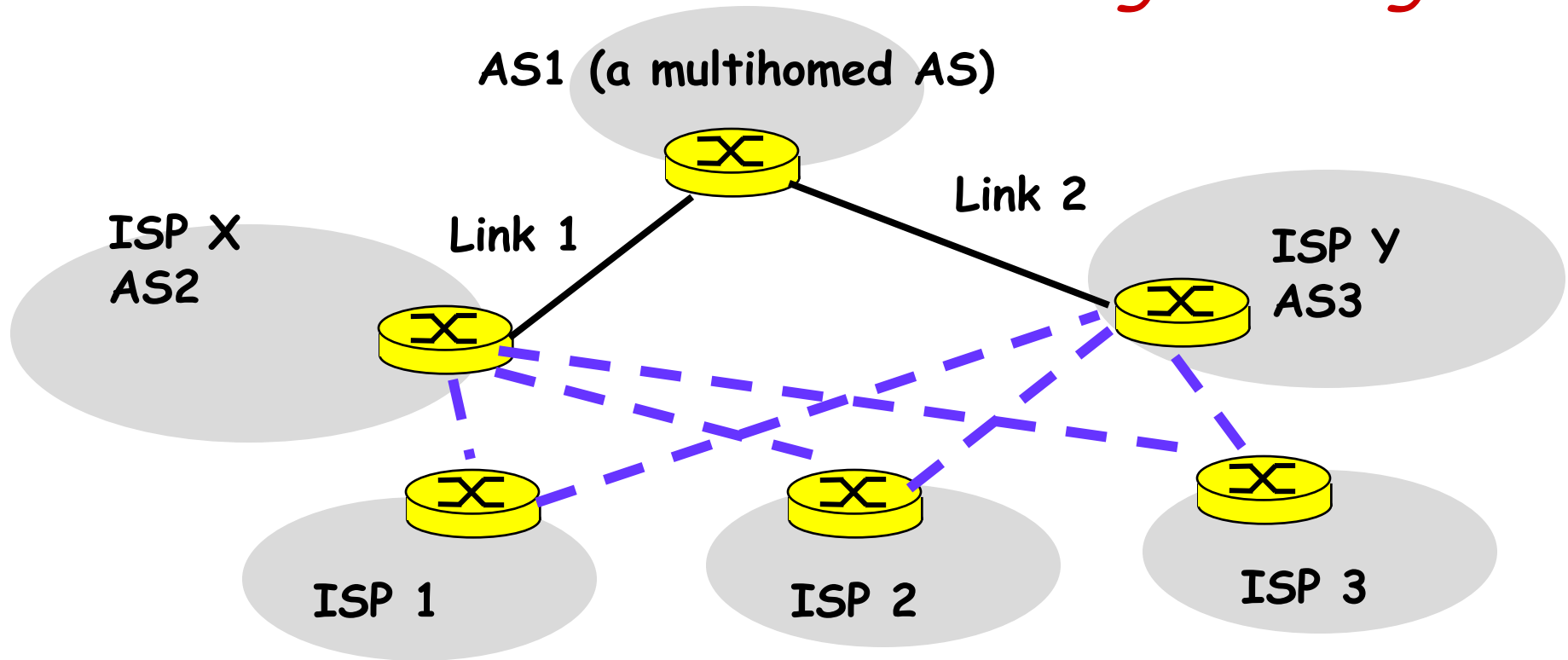
- Inbound traffic - traffic from ISP(s) to your network
- Outbound traffic - traffic from your network to ISP(s)
- Inbound traffic can arrive via either connection
- Outbound traffic can leave your network via either connection

Routing Control over Multihoming



- Can I control what traffic, inbound and outbound, should use which connection (or ISP)?
- Yes, you can
- How?
- Topic of interdomain traffic engineering (next)

What is inter-domain traffic engineering?



How can we control the outgoing traffic and incoming traffic of AS1?

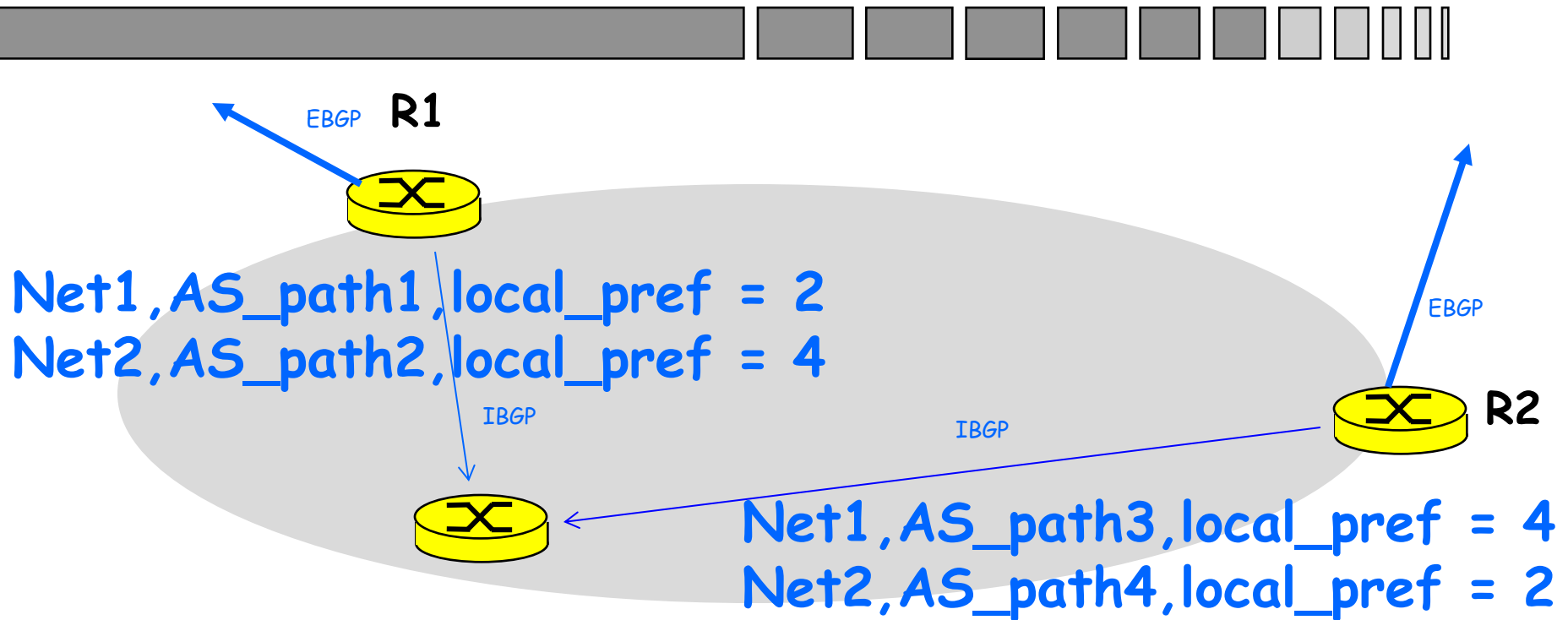
E.g. Is it possible to have (1) traffic coming from ISP1 and ISP2 always uses Link 1; and (2) traffic from ISP3 always uses Link 2.

Controlling traffic



- Controlling outgoing traffic
 - Using the **local-pref** attribute
- Controlling incoming traffic
 - Selective advertisement
 - Inflating AS_PATH length
 - Provider-supported BGP customer communities

Using local_pref to control outgoing traffic

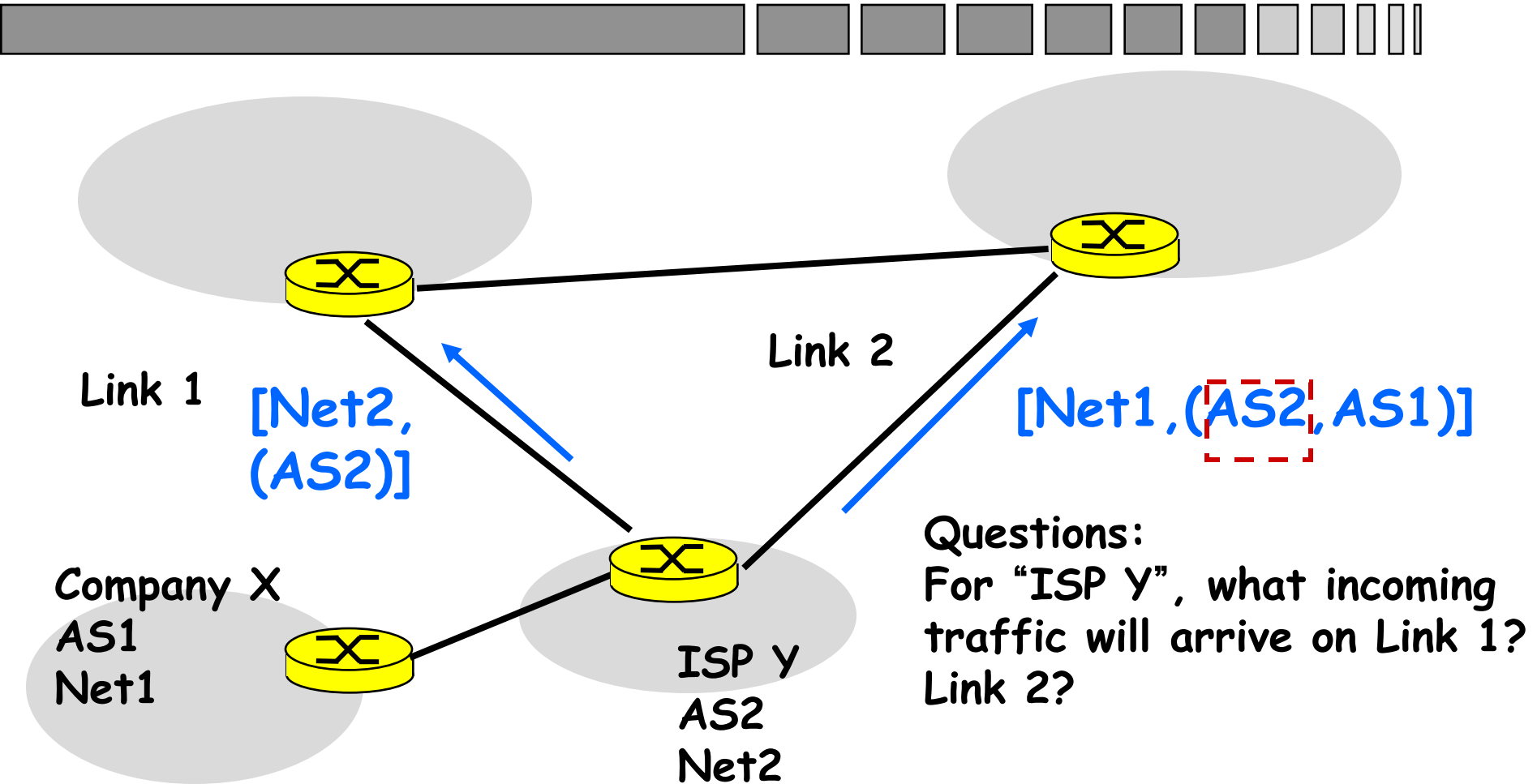


Local_pref is an IBGP attribute (never used on EBGP unless in confederation)
Both R1/R2 learn routes to Net1/Net2 from different external sources using EBGP
R1 sets LP=2 for Net1 and LP=4 for Net2 and advertises it to R3 using IBGP
R2 sets LP=4 for Net1 and LP=2 for Net2 and advertises it to R3 using IBGP
R3 selects R1 as next hop for Net2 and R2 for Net1

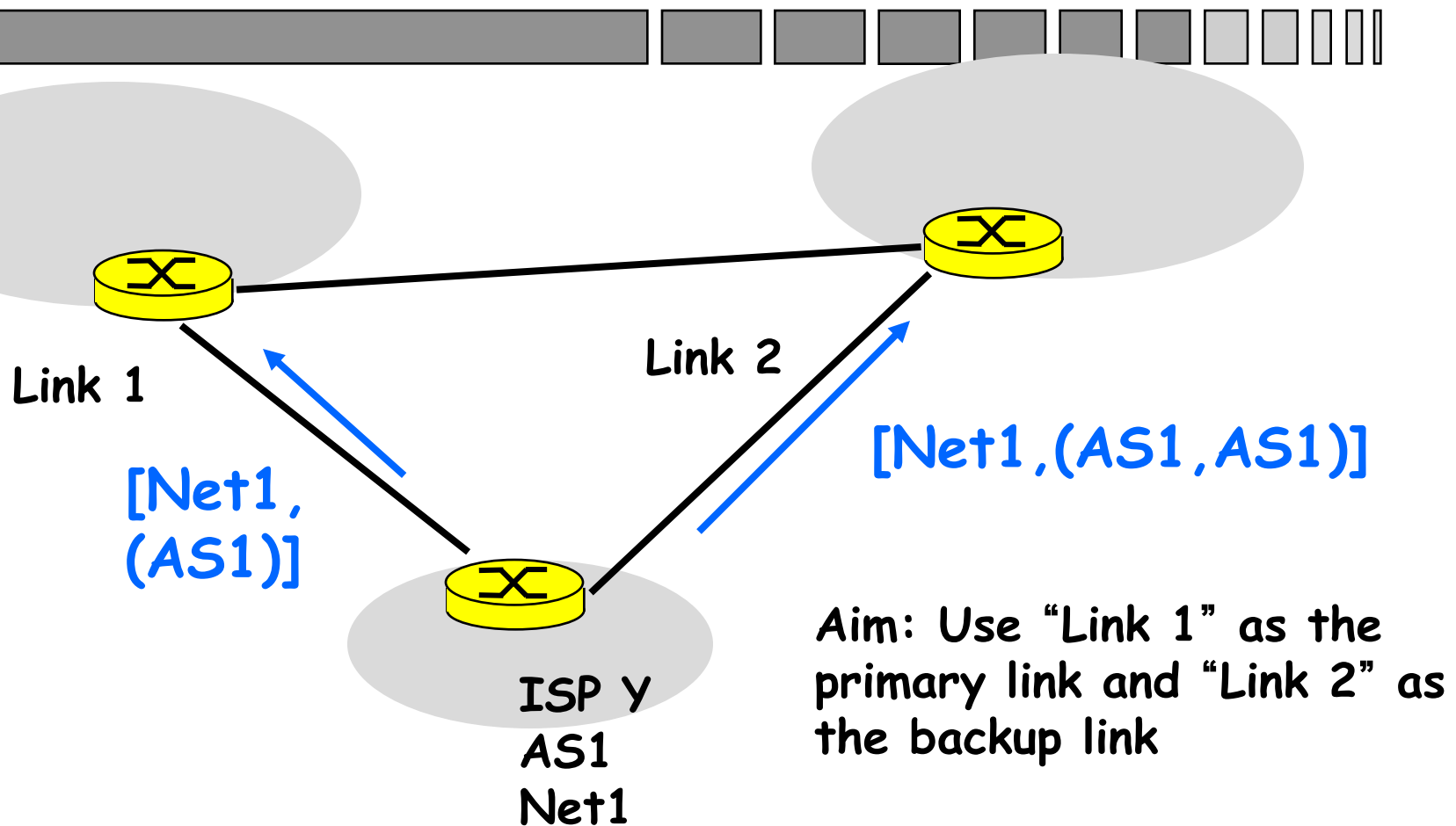


Controlling Incoming Traffic

Selective advertisement

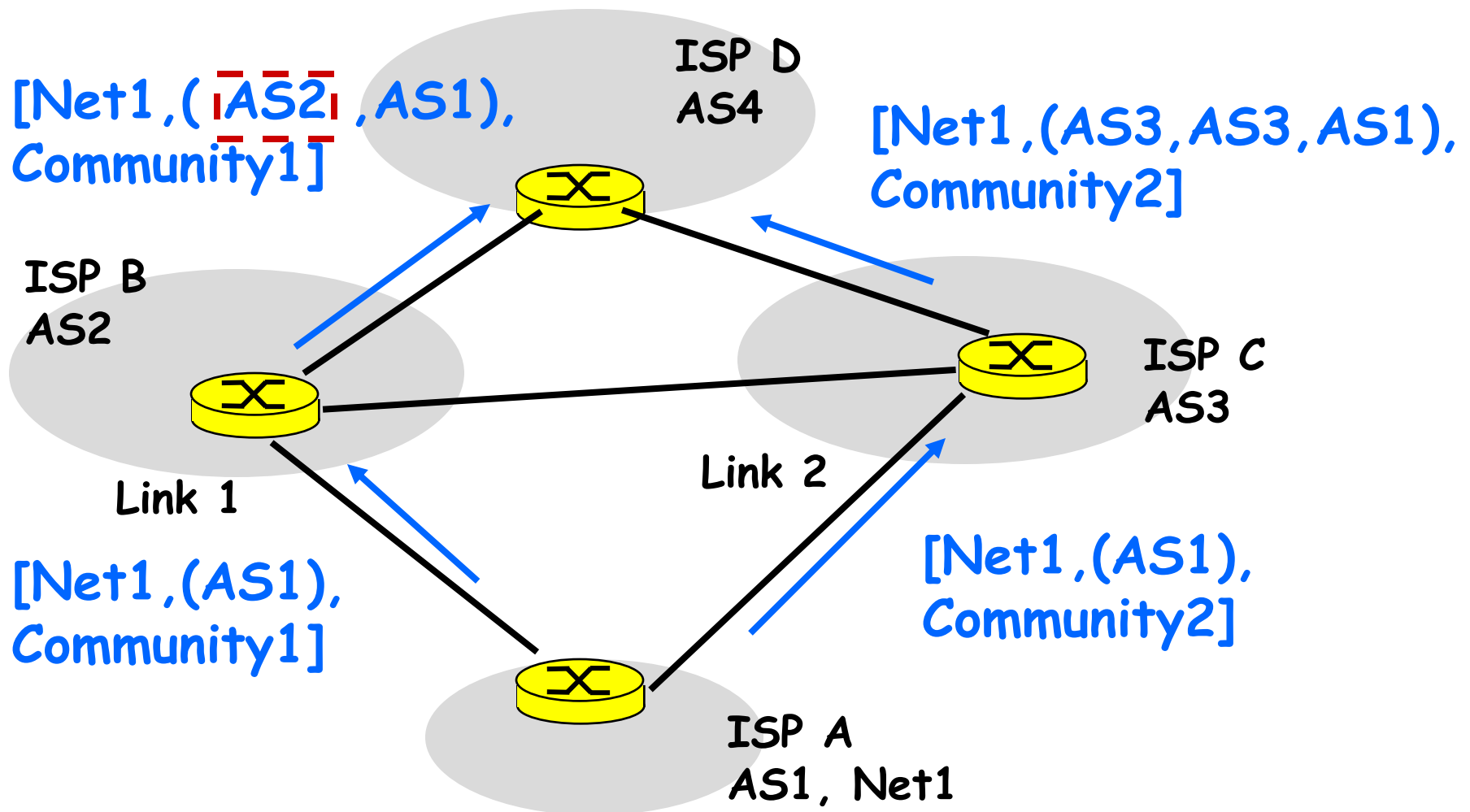


Inflating AS_PATH length



Provider supported BGP customer attributes

There is an agreement between neighbouring ISPs, when ISP C receives a route with community attribute "Community2" from ISP A, it will inflate the AS_PATH length (control beyond one-hop)



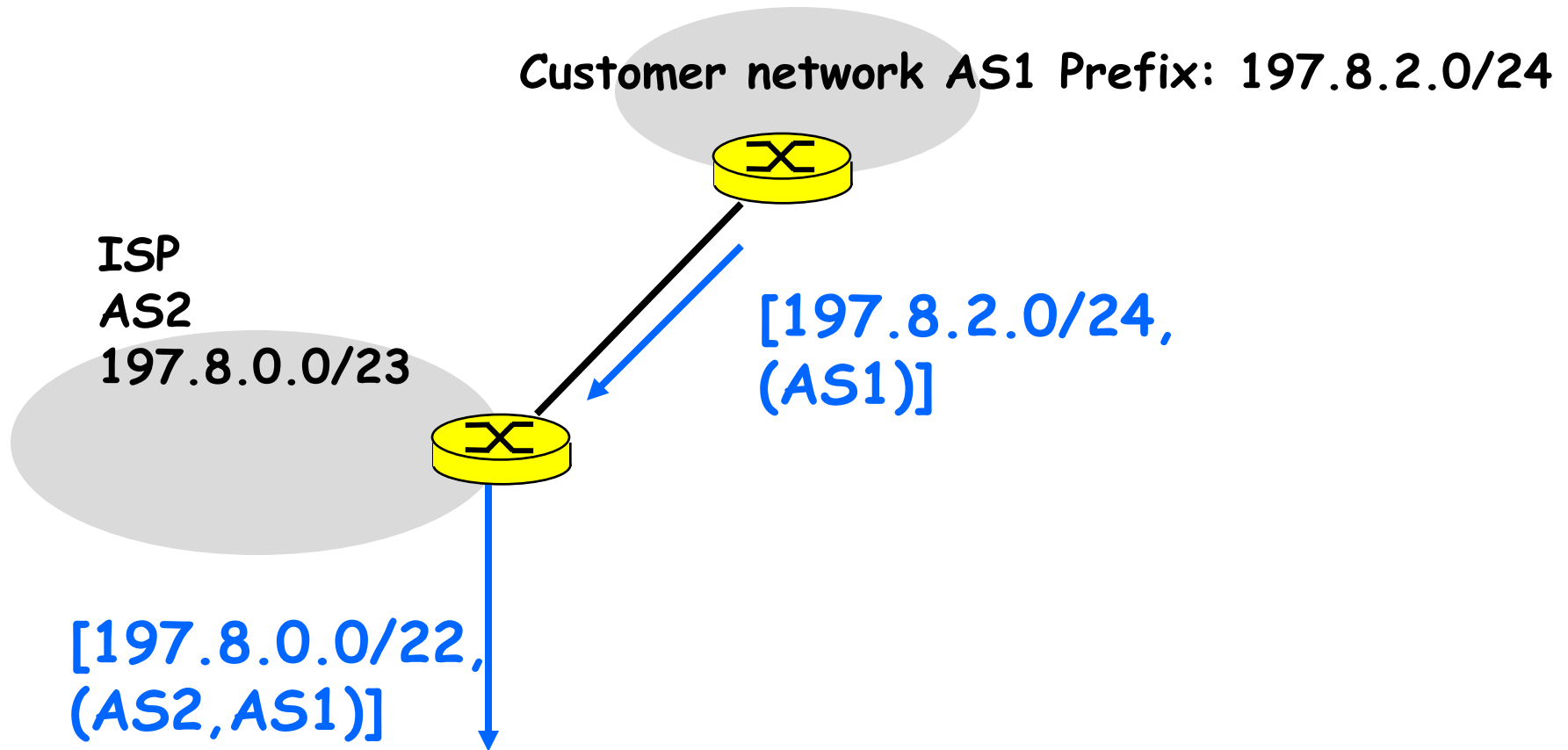
Some current BGP problems



- Multi-homing increasing the size of BGP tables
- Router misconfiguration

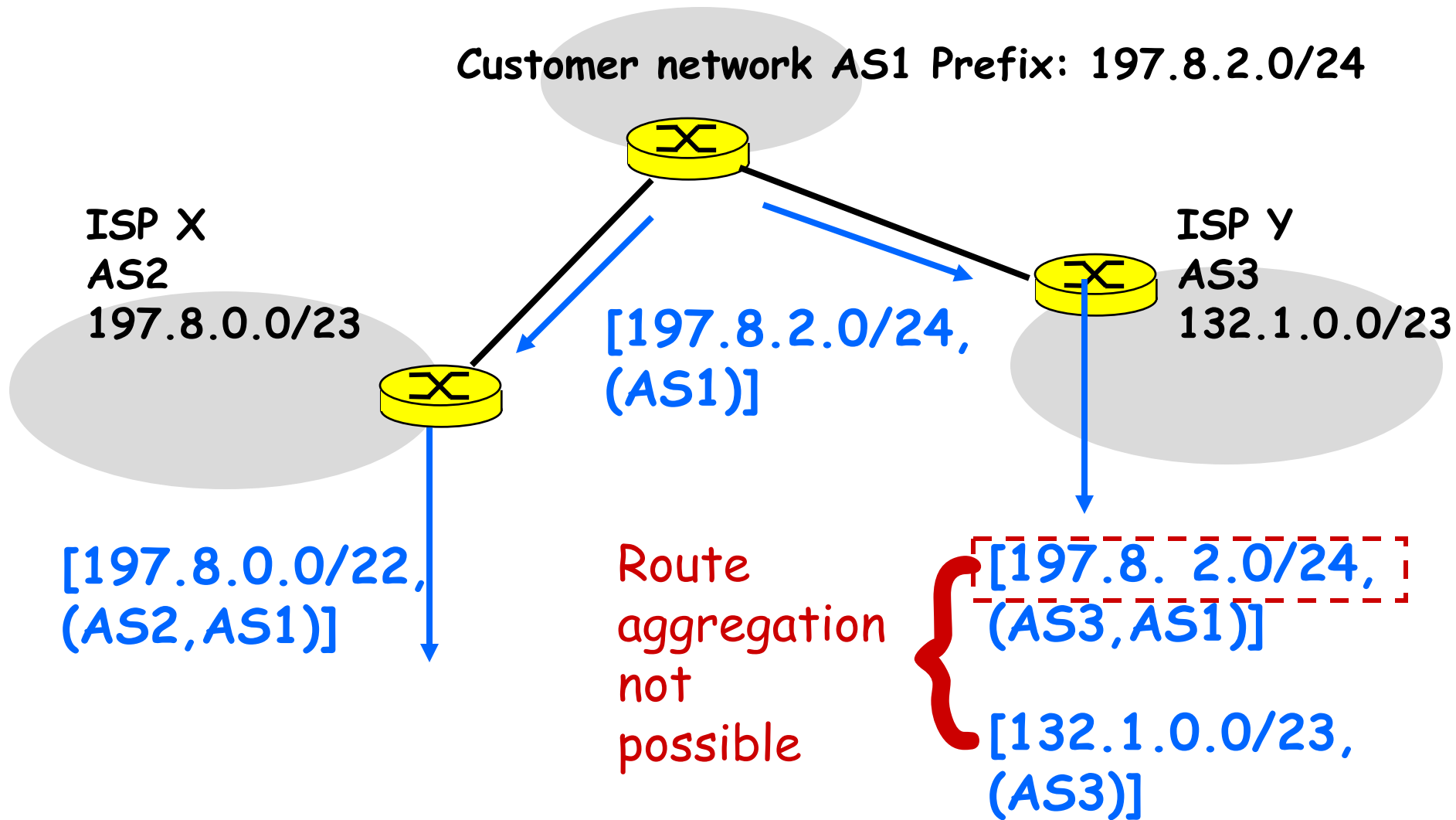
Problem of multi-homing (1)

When multi-homing is not used:



Problem of multi-homing (2)

When multi-homing is used:



Problem of multi-homing (3)

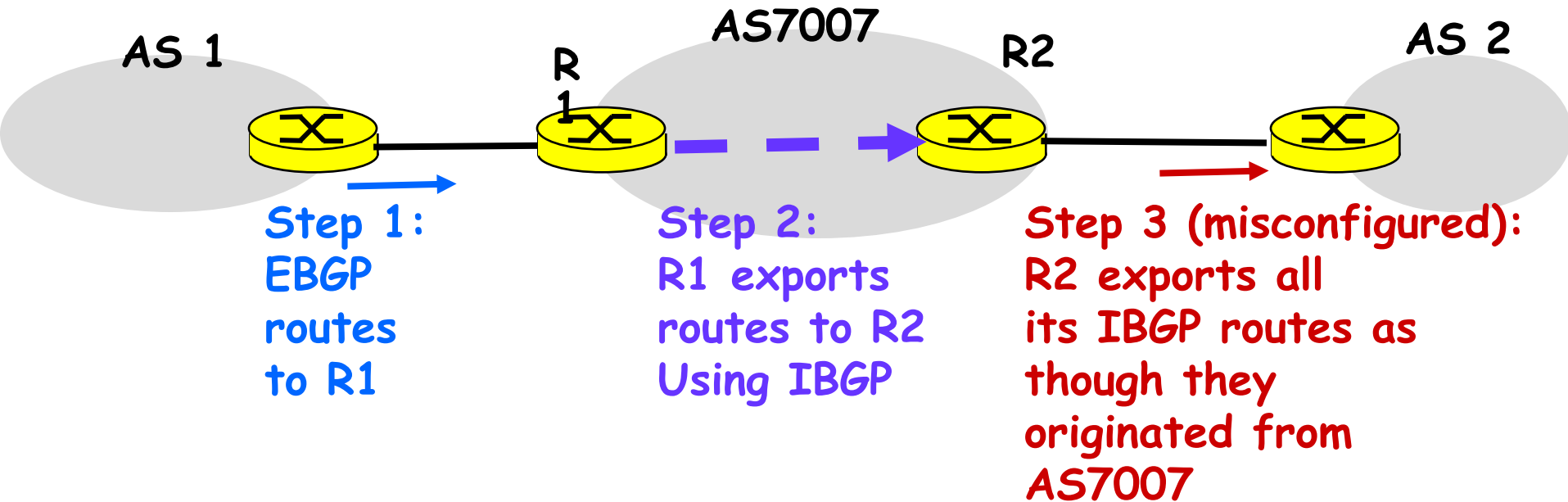


- Growth in multi-homing
 - Less than 1000 multi-homed stub ASs in 1998
 - More than 6000 multi-homed stub ASs in 2003
- Results:
 - Growth in BGP routing table size
- More memory and transmission overheads
- Can existing routers deal with this growth?
 - Scalability problem, again!

Trust and router misconfiguration (1)

The AS7007 incident (April 25, 1997):

accidentally leaking internal routing table to the Internet creating a blackhole



Trust and router misconfiguration (2)



- Most of the Internet sent traffic to AS7007
 - This is all caused by a router misconfiguration
 - Two hours of interruption for a large part of the Internet

References

- 
- IBM Redbook (Section 4.9)
 - Rorouzan (Chap 4 , 3rd Ed.)