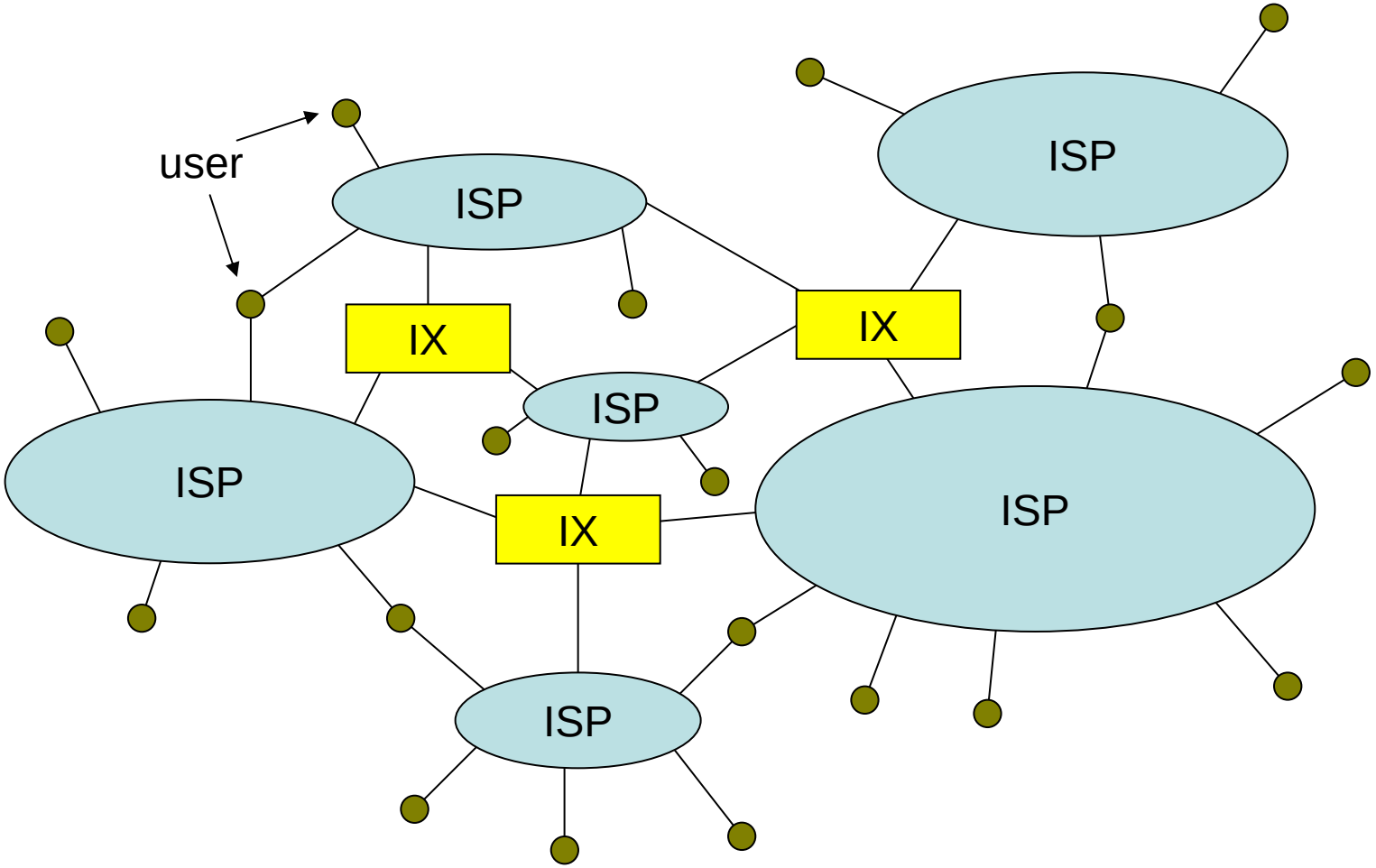
BGP4
overview and operation

Terutaka Komorizono <teru@ntt.co.th>

Introduction

- Presentation has many configuration examples
- Using Cisco IOS CLI
- Aimed at Service Providers
  - Techniques can be used by many enterprises too
- Feel free to ask questions

The Internet architecture



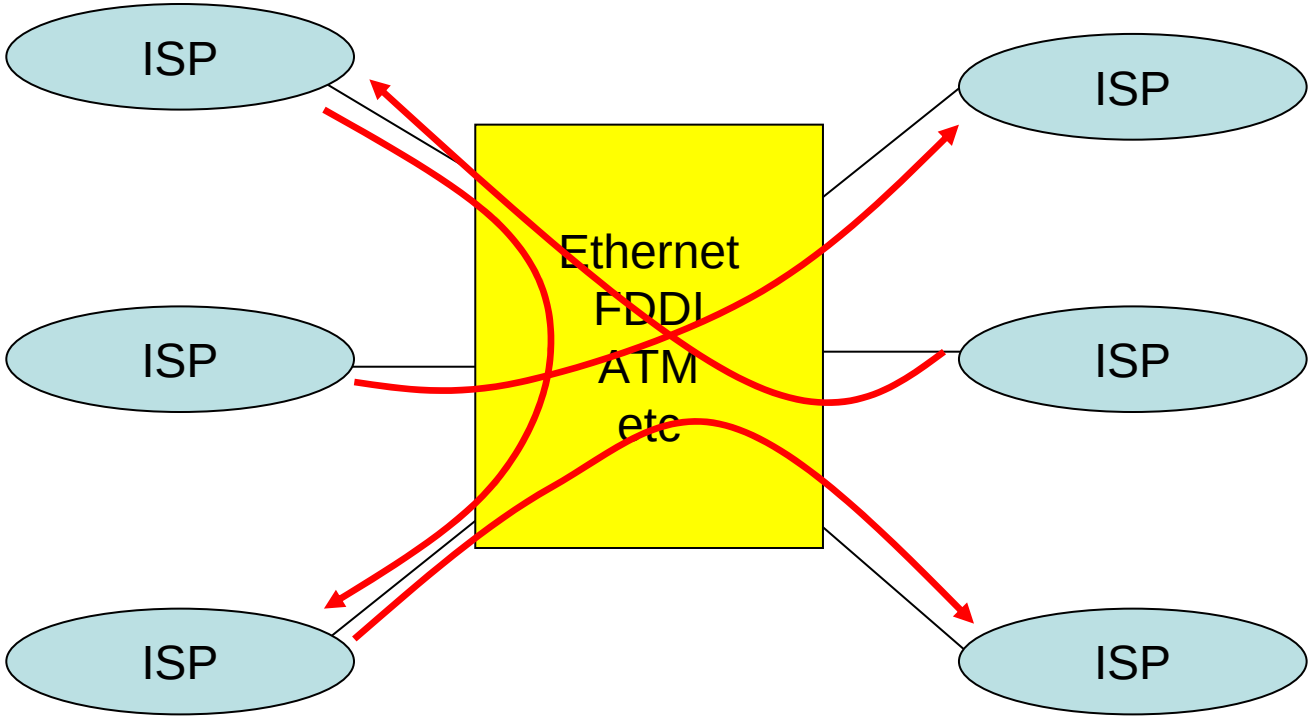
What is ISP?

- ISP provides connectivity to internet
- Mostly ISP has a interconnection with others, and the whole is composed
  - “internet”
  - Type of interconnection
    - via I.X.
    - Directly connection
- Users obtain the connectivity by several ISPs

I.X. (Internet Exchange Point)

- Service of connectivity between ISP and ISP.
- Exchange each ISPs traffic.
- Multilink type connection. Ethernet, FDDI, ATM.
- Two or more ISPs can be connected via same data-link media.
- i.e,
  - Network Access Point (NAP)
  - Metropolitan Area Exchange (MAE)
  - LINX, dix-ie (NSPIXP), JPIX, MEX, HKIX, etc.

Basic concept of I.X.



What is routing?

- Establishment of a connection in network layer between unique users who connect to internet
  - Addressing
  - Exchange routing information
- Traffic flow control on internet
  - Load balancing
  - Alternative route
  - Elimination of bottlenecks

Routing layer

- Type of routing
  - Inside ISP
  - Between ISP and ISP
- Interior Gateway (or Routing) Protocol (IGP)
  - Routing by cost
    - OSPF, RIP2
- External Gateway (or Routing) Protocol (EGP)
  - Routing by policy
    - BGP4

Scalability

- 2 problems
  - Depletion IP address block
  - Burst of routing table
- Temporary solution
  - CIDR (Class-less Inter-Domain Routing)
  - Private address (RFC1918)
- Permanent solution
  - IPv6 (RFC1883)
    - Extension IP address block (32bit to 128bit)
  - Hierarchical IP address assignment and aggregate routing entry

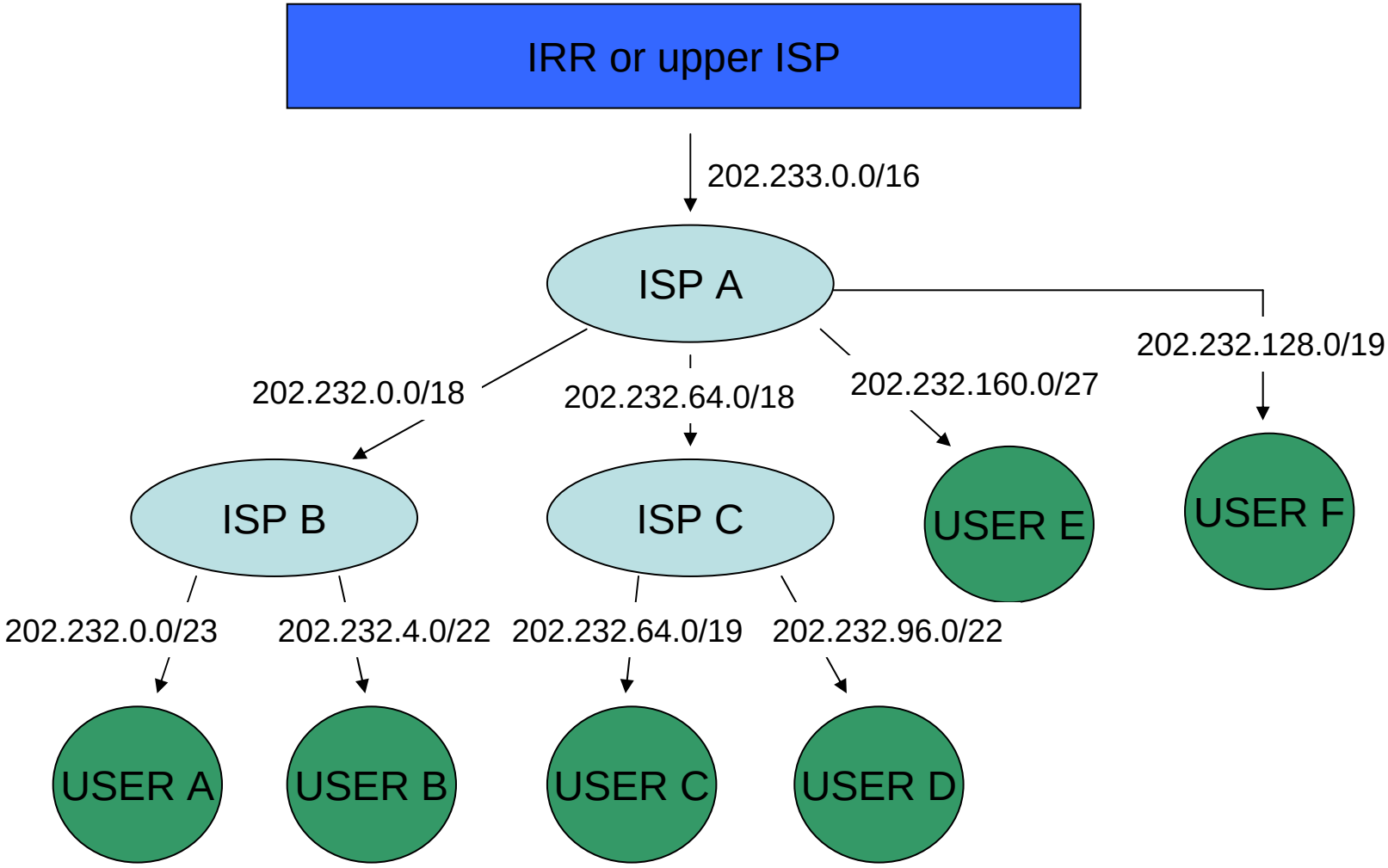
Classless Inter-Domain Routing

- Objective
  - Cast aside to bad effort of Classfull concept
  - Effective utilization of IPv4 address space
  - Reduction of routing entry in routing table
- Hierarchical IP address assignment
  - Assignment with bit (IP) boundaries
- Aggregate routing entry
- IP address prefix form
  - 202.232.68.0 – 202.232.68.63 = 202.232.68.0/26

Classless Routing

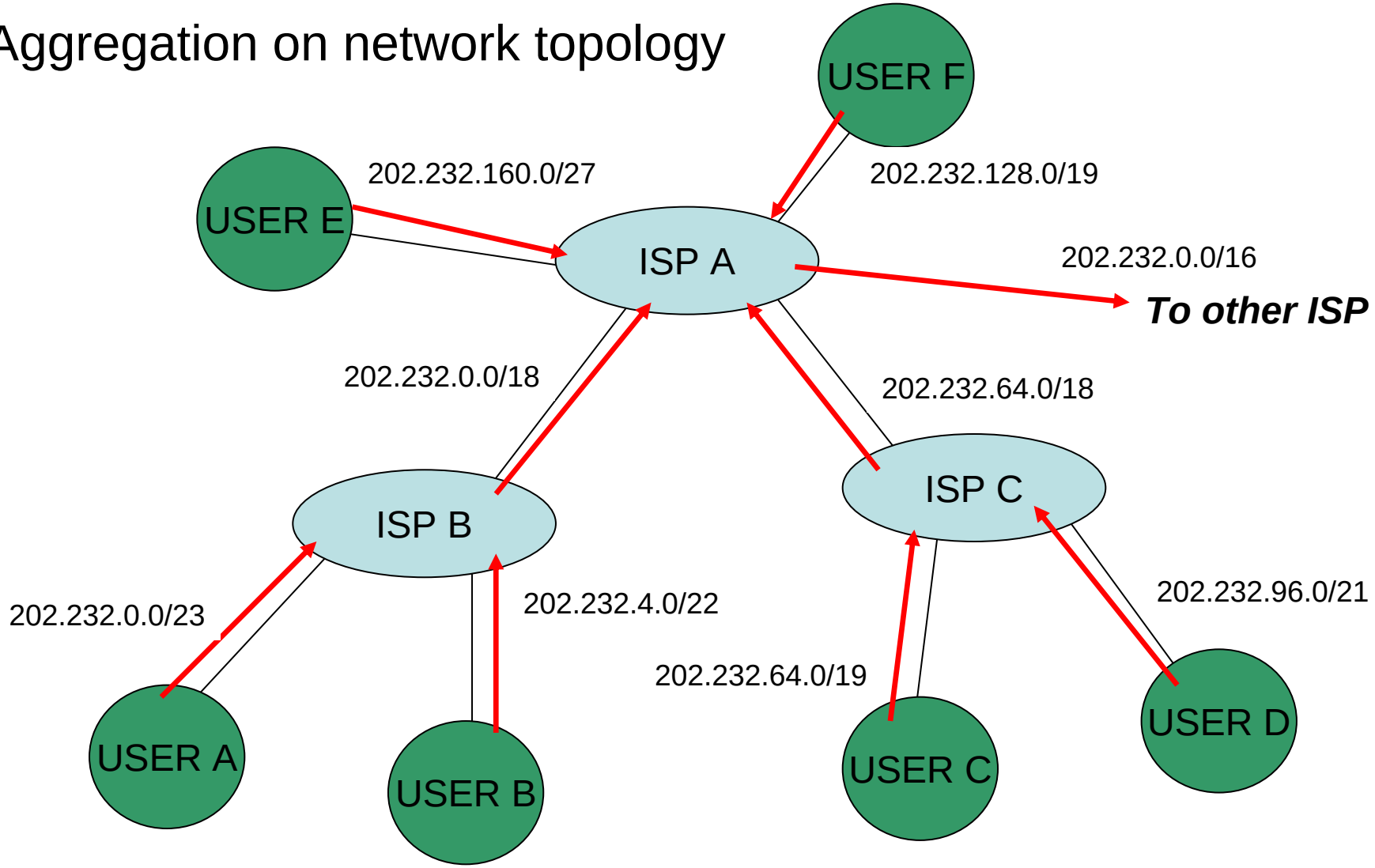
- Support VLSM
  - Interface / routing table / routing protocol
- Support supernet
  - Address / aggregate routing
- Exclusion of “Classfull” assignment and concept of “Classfull” routing
  - All-0 subnet, all-1 subnet, etc
- Classless routing information
  - Advertise netmask length

Hierarchical assignment

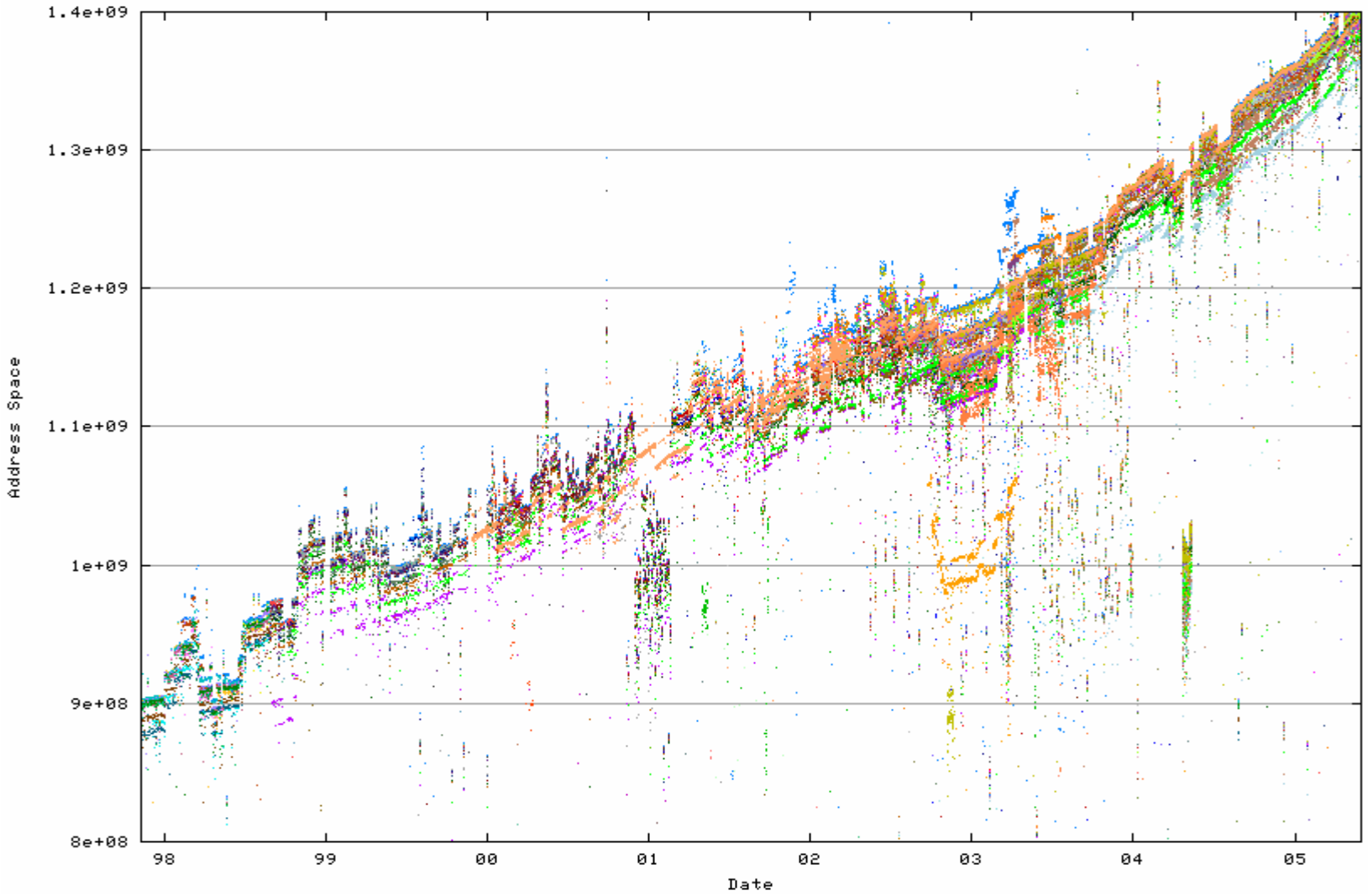


Aggregate routing information

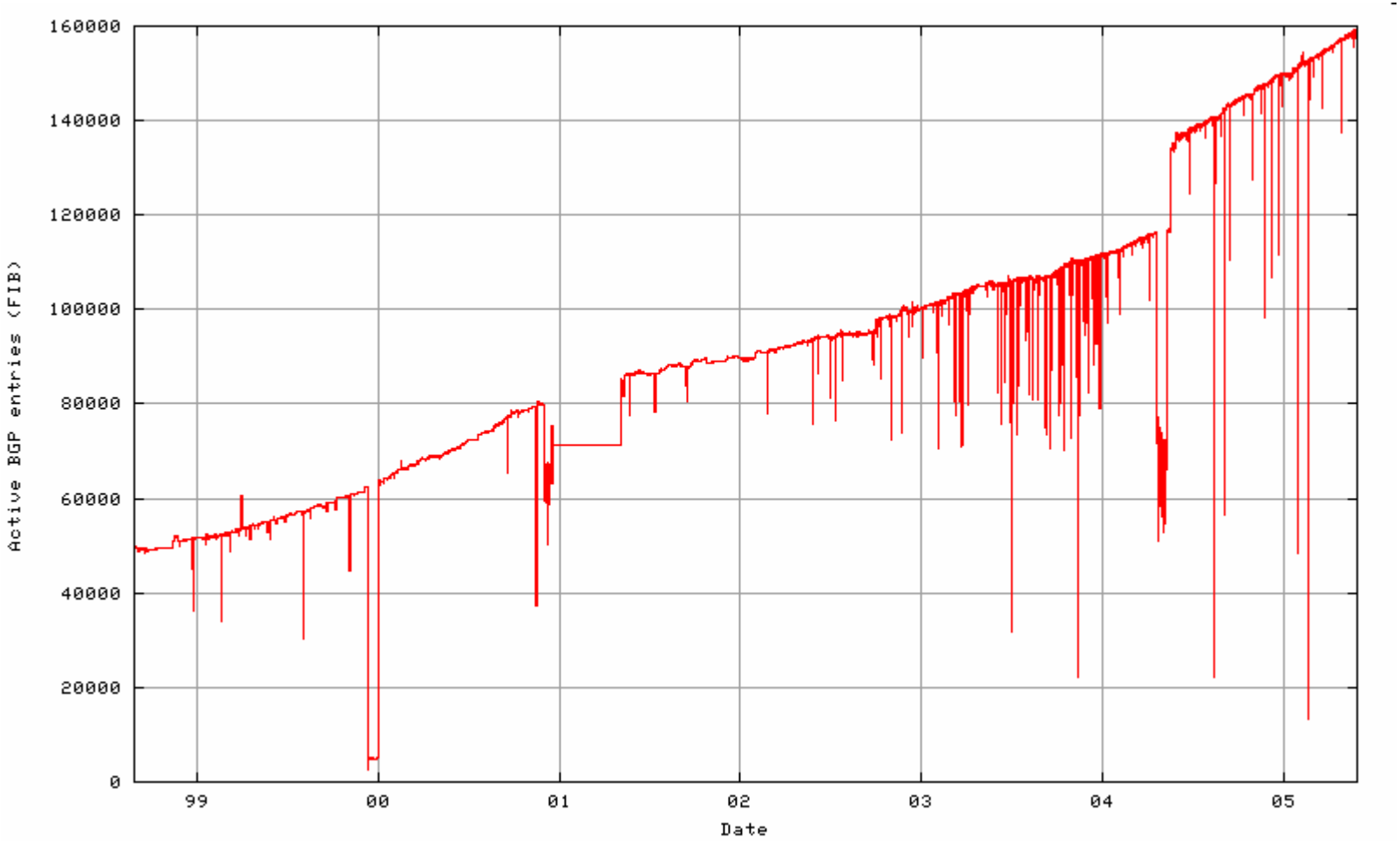
Aggregation on network topology



Status of IP address utilization



Status of Routing table



BGP4 overview

What is this BGP thing?

BGP4 (Border Gateway Protocol)

- RFC1771

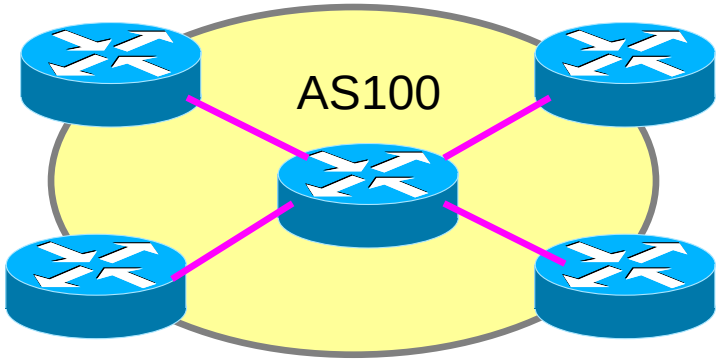
  Work in progress to update

  http://www.ietf.org/internet-drafts/draft-ietf-idr-bgp4-26.txt
- De-facto standard protocol of routing between AS and AS
  - Autonomous System (AS)
    - It is used to uniquely identify networks with common routing policy
    - ISP ≒ AS
    - Internet is aggregated AS
- Supported CIDR
  - Be vital to the realization of CIDR

Features

- Use TCP/179
  - Exchange routing information between peer (routers) and peer (routers)
  - Reliability is secured for the exchange of routing information
  - Incremental exchange information, different from RIP and etc
- Path vector routing protocol
  - Route selection by “Path attribute” that is added to routing information
  - AS Path, Origin, Next Hop, Multi-Exit-Discriminator (MED), Local Preference, etc.

Autonomous System (AS)



- Collection of networks with same routing policy
- Single routing protocol
- Usually under single ownership, trust and administration control
- Identified by unique number

Autonomous System Number (ASN)

- An ASN is 16 bit integer
  - 1-64511 are public network use
  - 64512-65534 are for private use and should never appear on the internet
  - 0 and 65535 are reserved
- 32 bit ASNs are coming soon

  http://www.ietf.org/internet-drafts/draft-ietf-idr-as4bytes-09.txt

  With ASN 23456 reserved for the transition

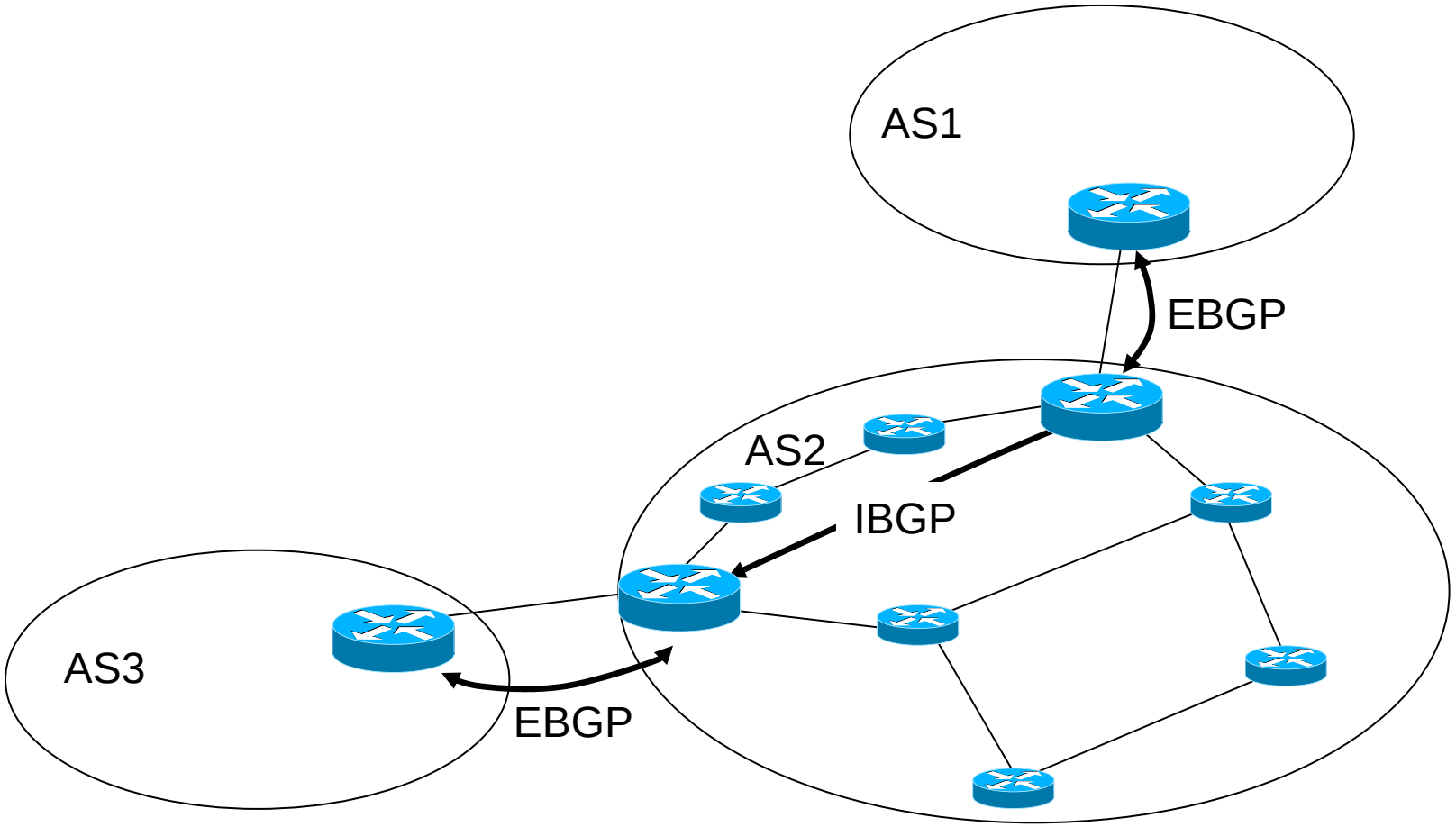
Autonomous System Number (ASN)

- ASNs are distributed by the Regional Internet Registries
  - Also available from upstream ISPs who are members of one of the RIRs
- Current ASN allocations up to 37887 have been made to the RIRs
  - Of these, around 19500 are visible on the internet
- Current estimates are that 4-byte ASNs will be required by July 2010

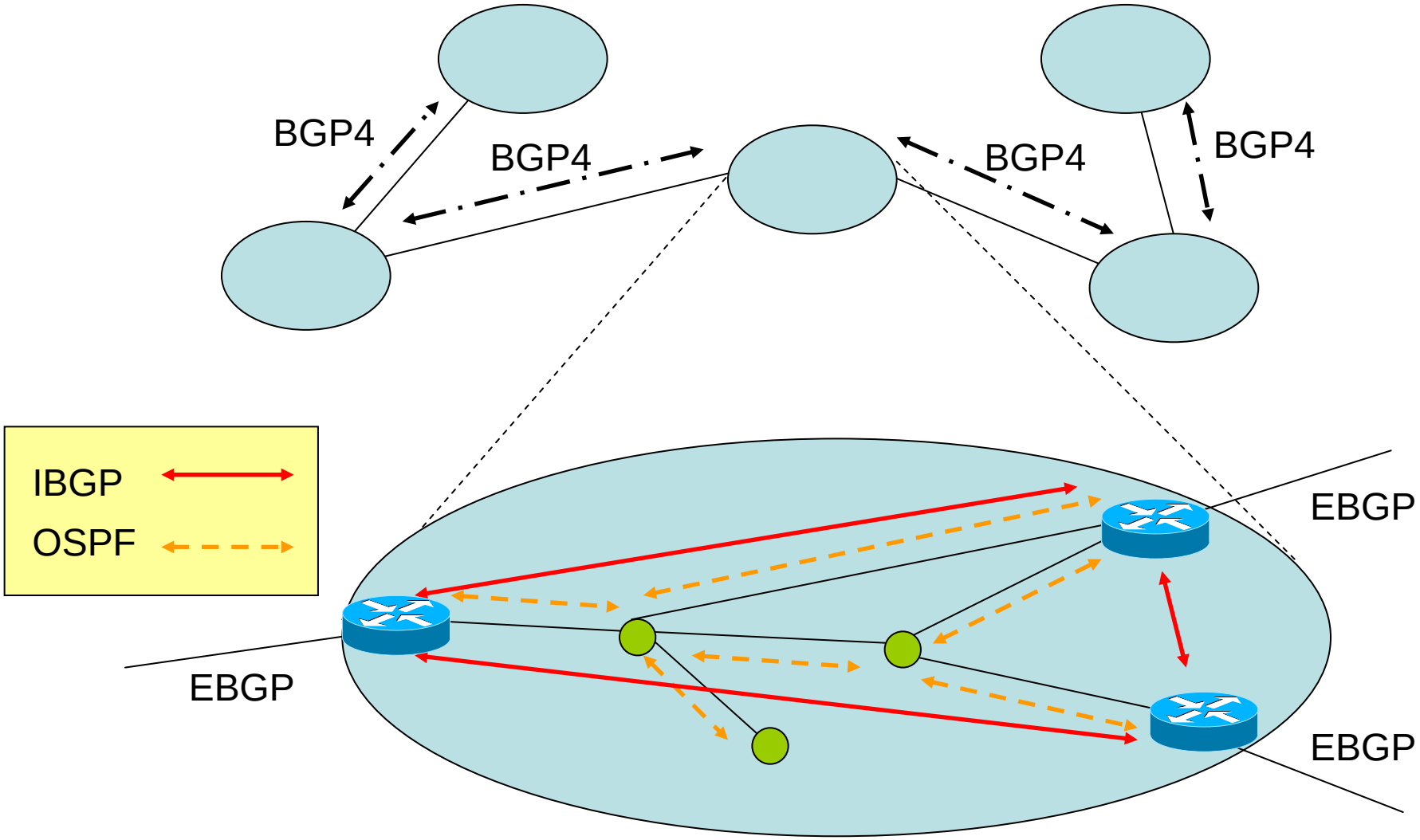
EBGP and IBGP

- BGP speaker (border router)
  - Router of running bgp
- EBGP (External BGP)
  - bgp peering session between bgp speaker and another AS's bgp speaker
- IBGP (Internal BGP)
  - bgp peering session between bgp speaker and bgp speaker in same AS
    - Full mesh
      - * RR and Confederation
    - Advertise routing information that learns from another AS bgp speaker
    - Do not advertise routing information that learns from another ibgp speaker

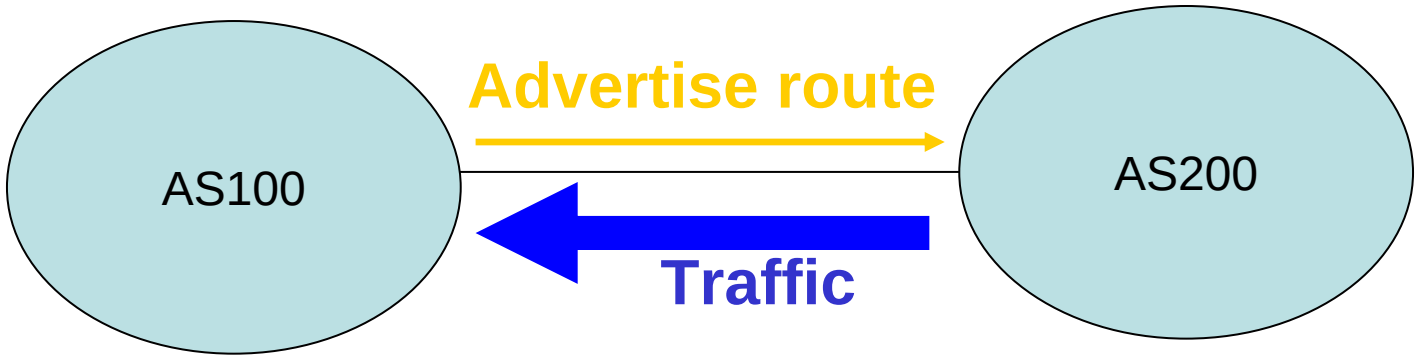
Routing information exchange between AS and AS by BGP



Routing control internal/external AS

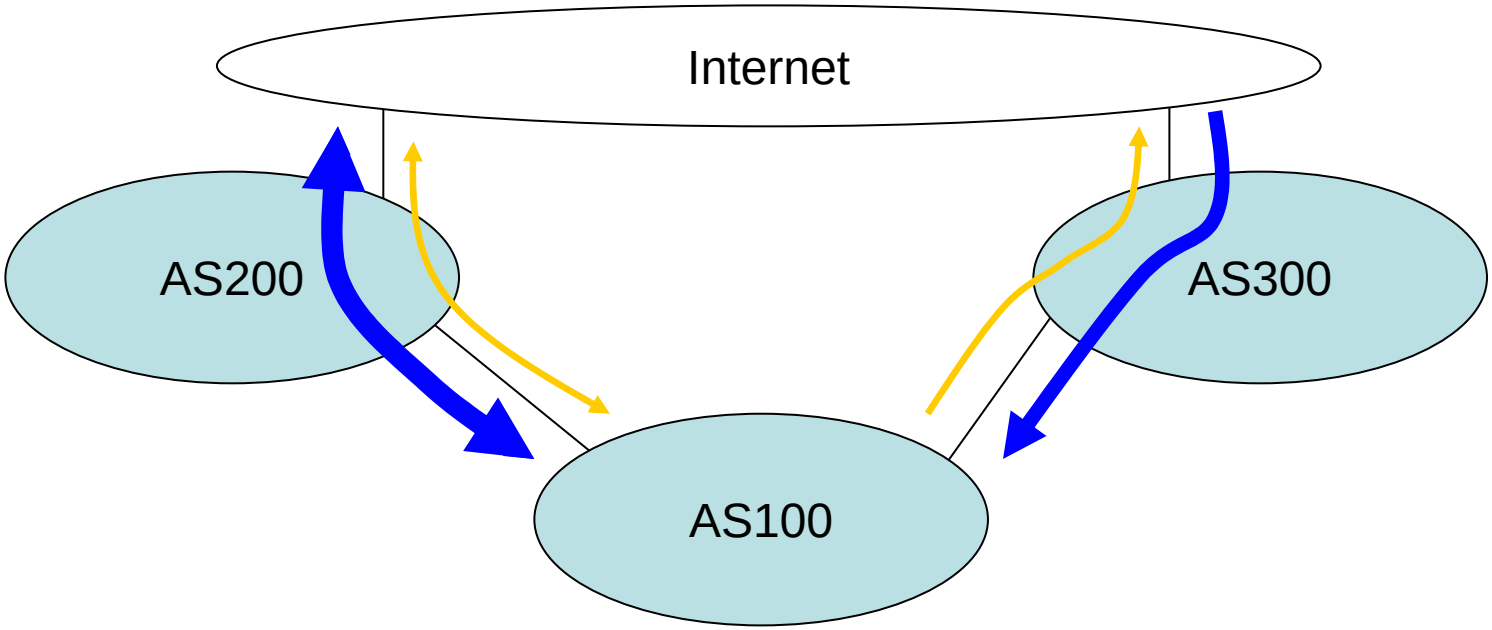


Basically BGP Traffic flow 1



1. When AS100 advertises own routes to AS200
2. Traffic to AS100 coming from AS200

Basically BGP Traffic flow 2



- AS200 transits AS100's traffic to internet

1. AS100 has a connectivity to the internet via AS300
2. AS100 starts to advertise own routes to AS300, and AS300 receives AS100 routes and advertises AS100's routes to the internet
3. The traffic to AS100 comes from internet via AS300
4. The traffic to AS100 via AS200 will be decreased due to some traffic from internet to AS100 is changed to via AS300

PATH Attributes

- Attributes of advertised (received) routes
  - For selecting route from multi route*****
  - Policy
- Transitive attribute and Non-transitive attribute
- Mandatory attribute and Optional attribute

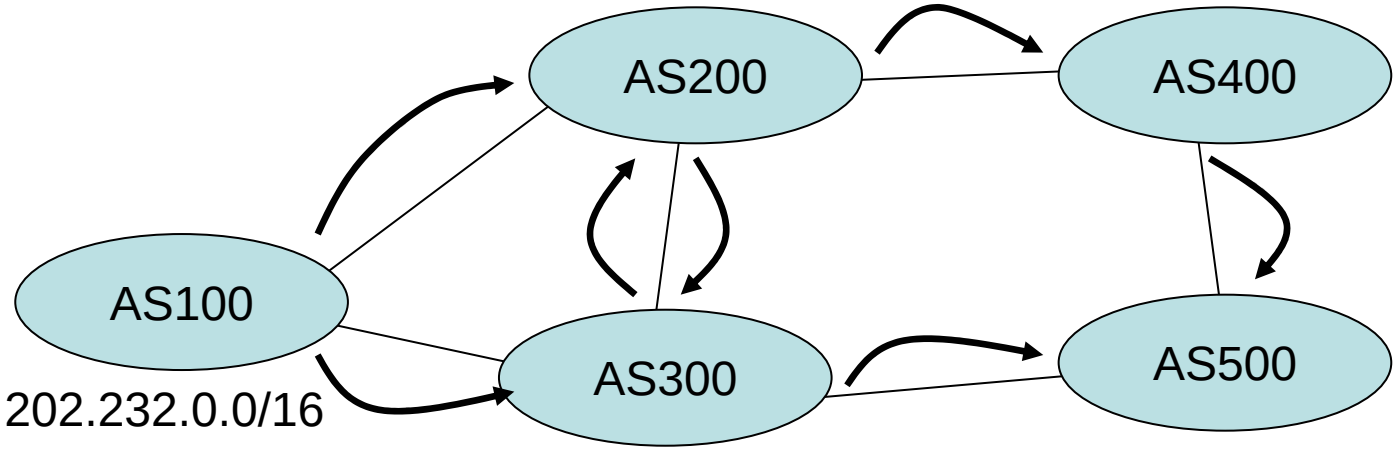
Origin Attribute

- Originate ASN
- It is set when firstly advertising the route
- We seldom use it in actual operation
- “Historical” attribute
- Mandatory attribute
- Available values
  - IGP
    - Almost of origin attribute is IGP
  - EGP
    - It’s hardly seen now
  - Incomplete
    - Can not be known where to had been distributed to bgp

AS-PATH attribute

- The AS-PATH attribute is actually the list of AS Ns that a route has traversed in order to reach a destination.
- Detection loop
- Shortest AS-PATH wins
  - Depend on the policy
  - prepend, stuffing, etc
- Mandatory attribute

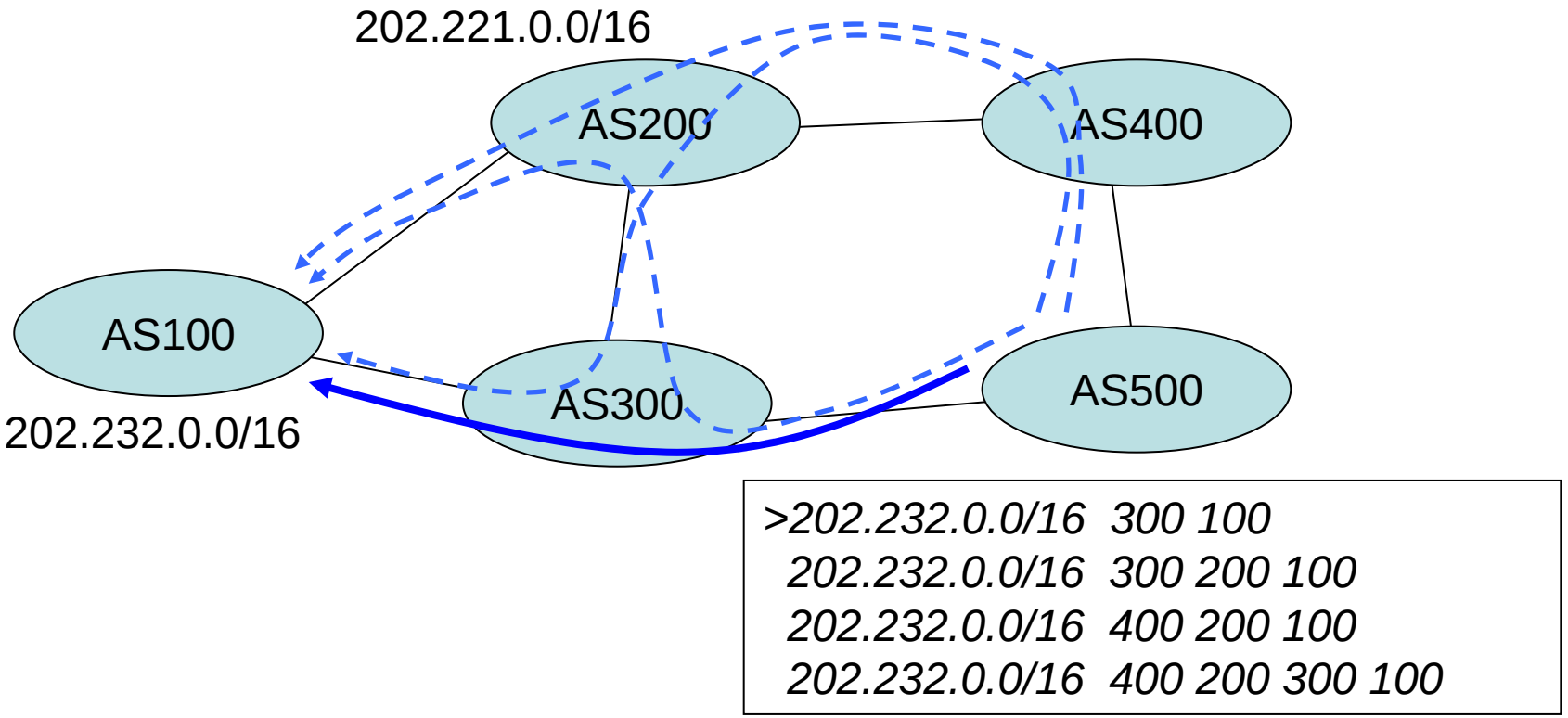
Sample AS-PATH attribute



202.232.0.0/16 300 100
202.232.0.0/16 300 200 100
202.232.0.0/16 400 200 100
202.232.0.0/16 400 200 300 100

- AS100 announce 202.232.0.0/16

Shortest AS-PATH route wins

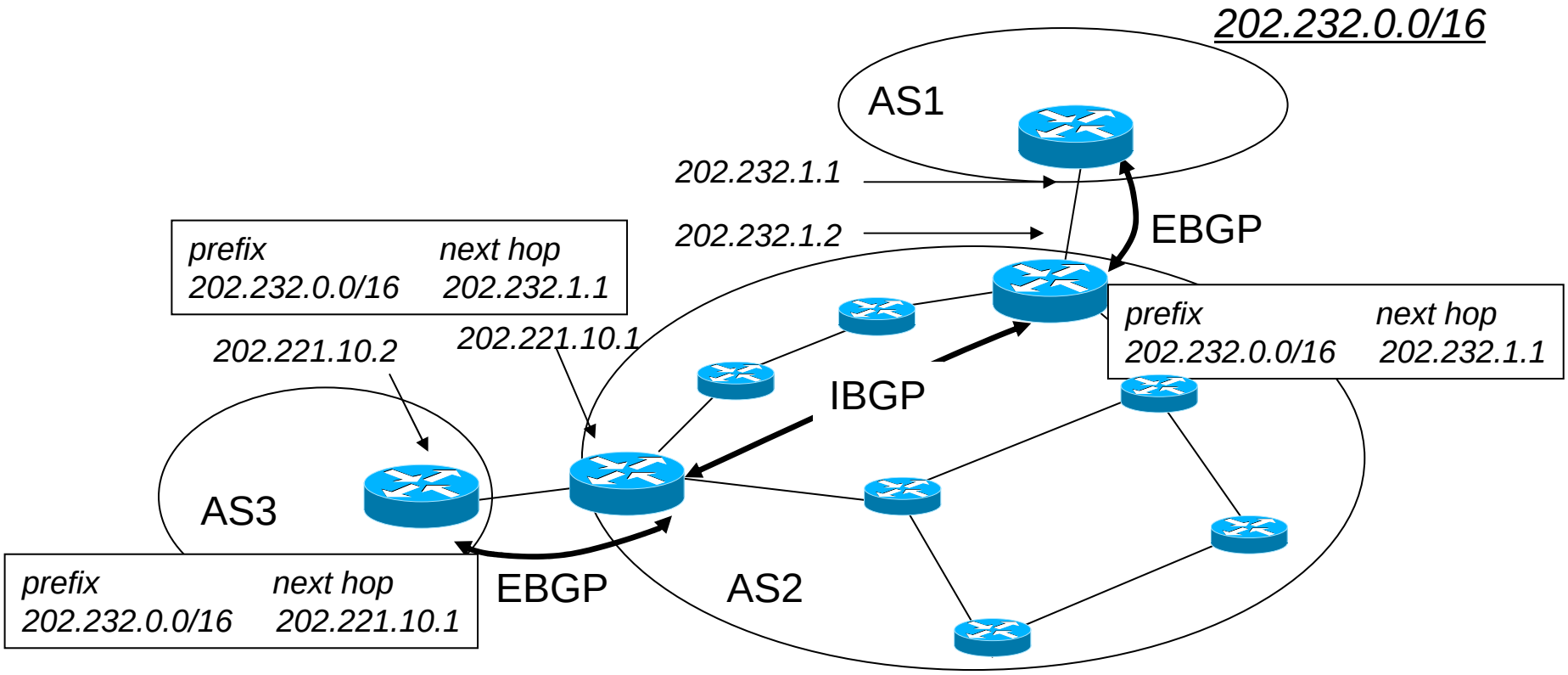


- Prefer via AS300 route from AS500 to 202.232.0.0/16

Next-Hop attribute

- In case of the next-hop address is unreachable, it will delete the prefix from own routing table
  - Delete the routes while receiving in eBGP
- There are two methods for resolution of next-hop of external routes
  - Set own loopback address to Next-Hop attribute while receiving the routes by eBGP
    - Set “next-hop-self” for iBGP (cisco)
    - Routing the loopback address by IGP
  - Redistribute to IGP such a PtP /30 (connected) address for eBGP peering address
    - redistribute connected
    - network command + passive

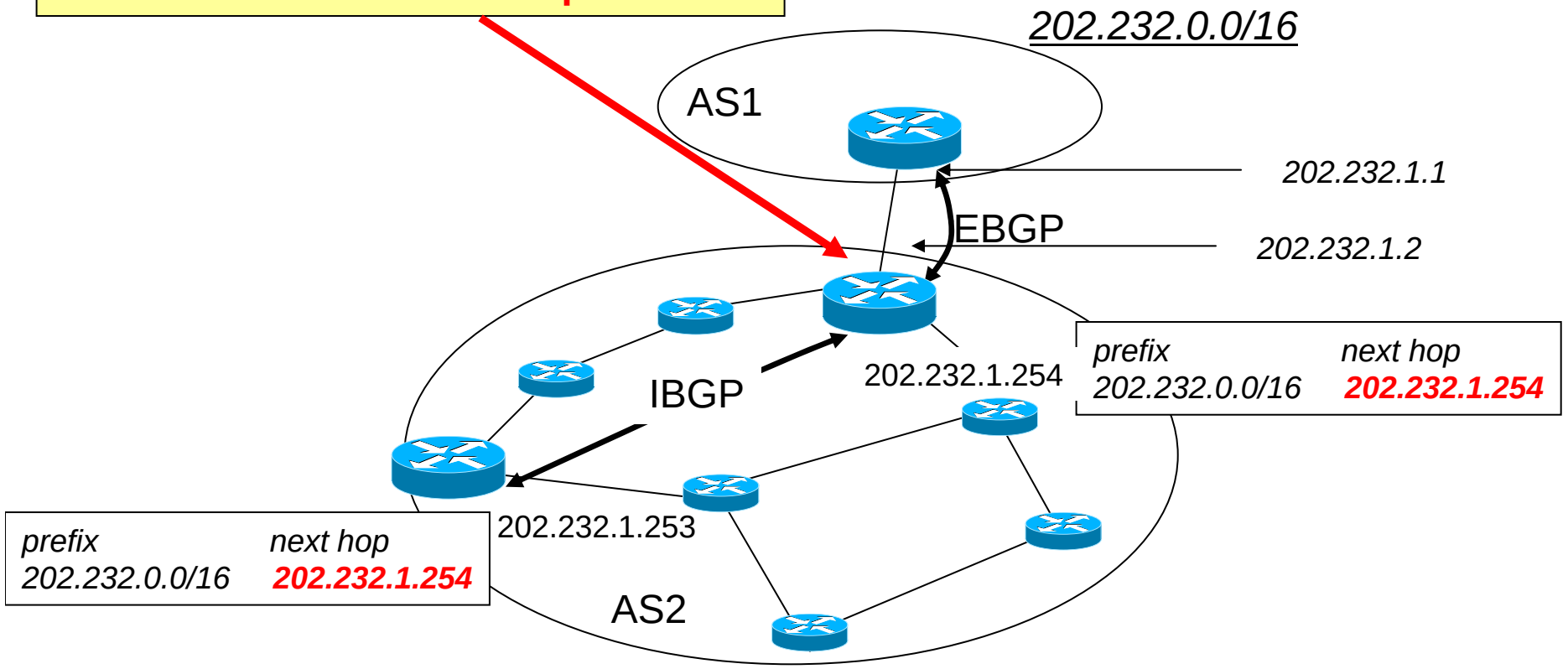
Next-Hop attribute



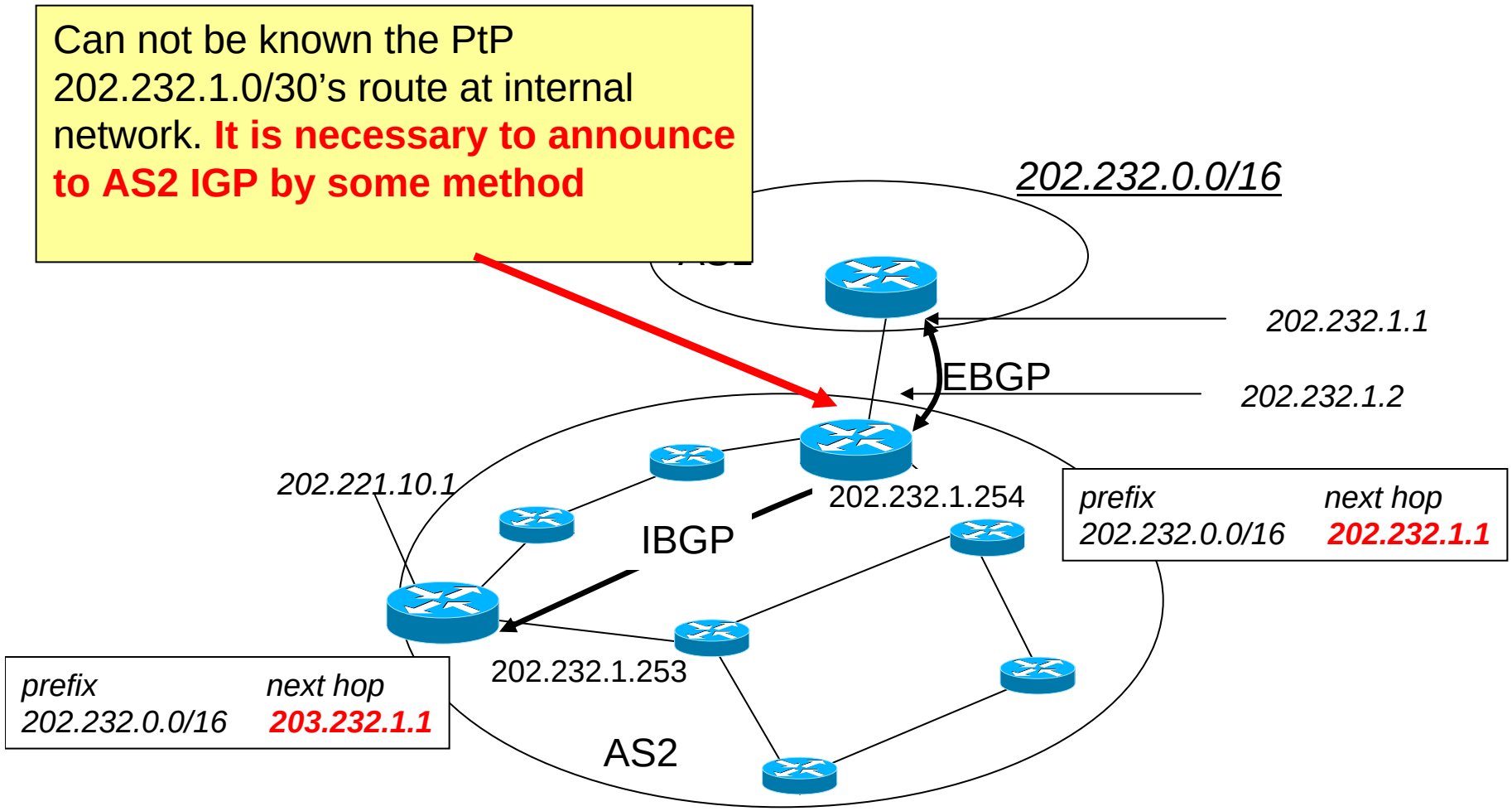
- Neighbor border router IP address on routing table
- No change next-hop attribute value in iBGP advertisement
- Routing from R3 to R1 is resolved by IGP

examples: set “next-hop-self”

When redistribute ebgp routes to ibgp,
R2 change next-hop attribute to as
“I’m the destination of the prefix”



Examples: redistribute ebgp routes to ibgp



Next-Hop (summary)

- IGP should carry route to next hops
- Recursive route look-up
- Unlink BGP from actual physical topology
- Allow IGP to make intelligent forwarding decision

Next-Hop attribute – hot potato/hold potato

- “Hot Potato” and “Cold Potato”
  - Hot Potato
    - Put out traffic from the closest point
    - “Closest-exit”. Put our traffic from the closest address Normally it routed by highest IGP cost
  - Cold Potato
    - It is Policy Routing. Routing is done by the policy even in case of the roundabout

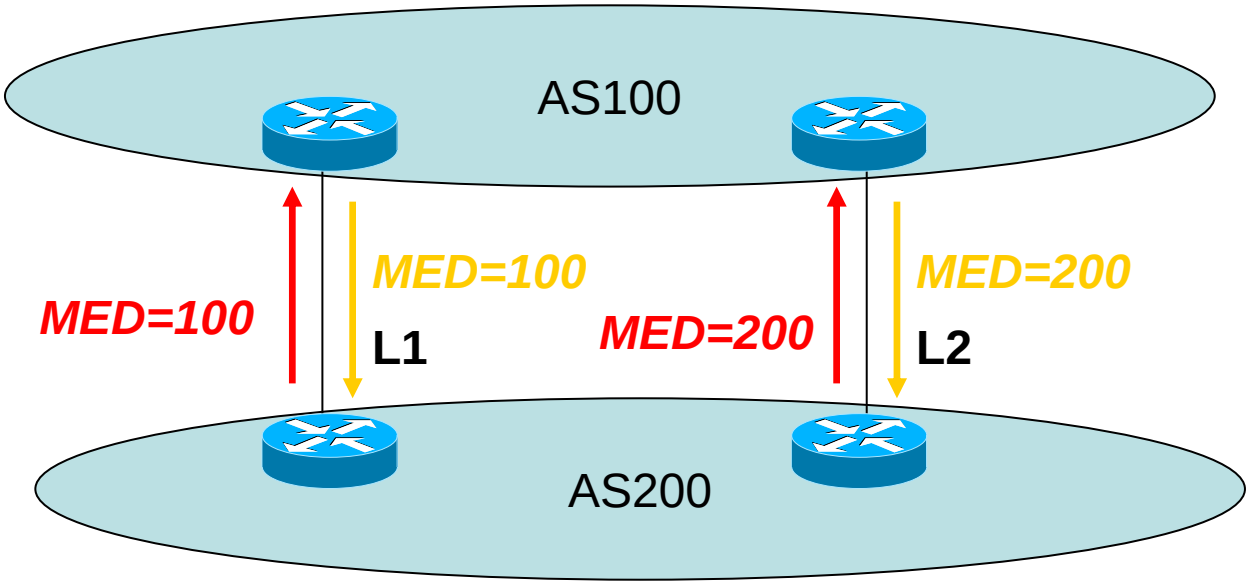
Multi-Exit Discriminator (MED)

- Inter-AS – non transitive
- Used to convey the relative preference of entry points

  Determines best path for inbound traffic
- Comparable if paths are from same AS
- IGP metric can be conveyed as MED

  set metric-type internal in route-map

Sample MED

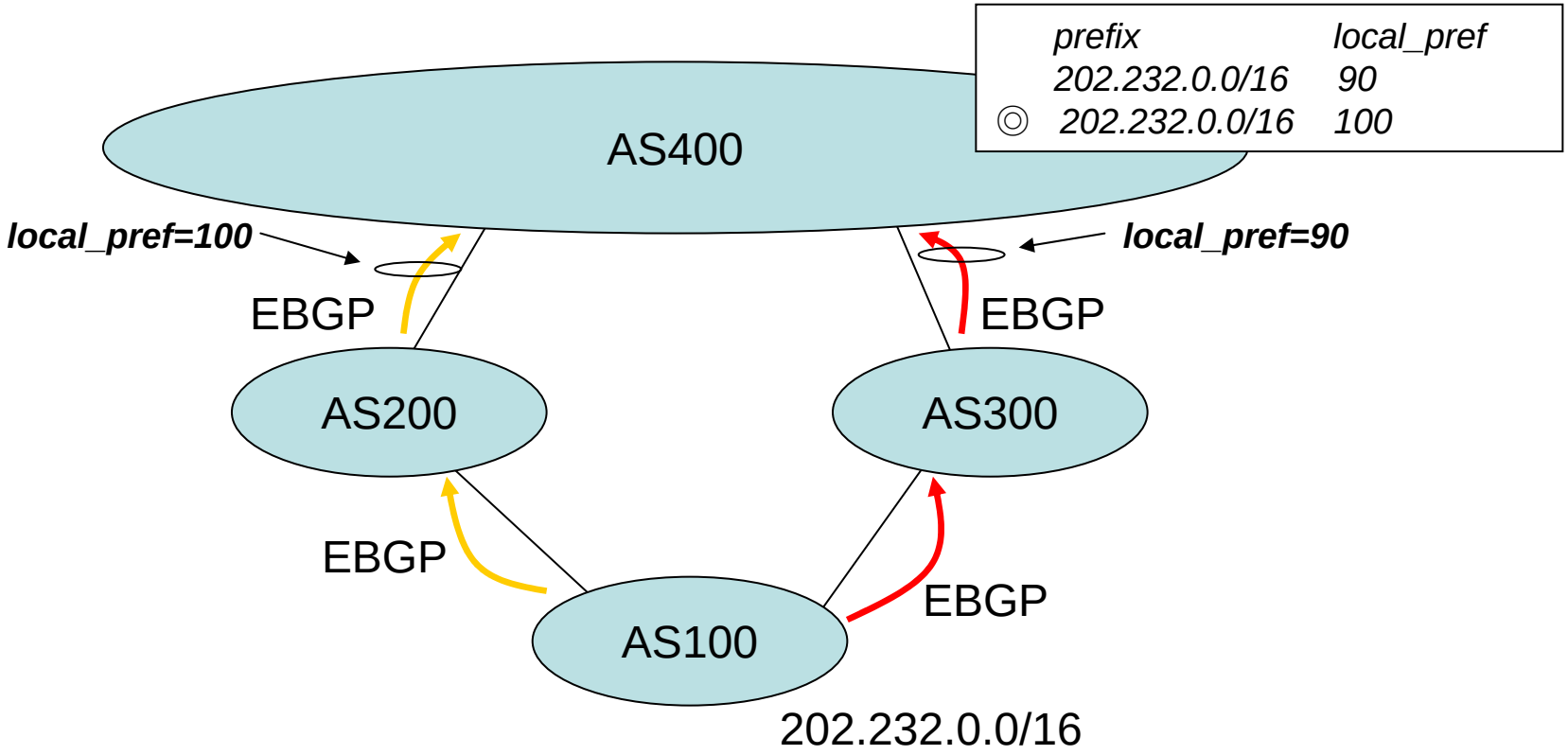


- Priority L1
- Independent MED value AS100 and AS200

local-preference attribute

- Local to an AS – non-transitive
  - Default local preference is 100 (cisco)
- Path with highest local preference wins
- Used to influence BGP path selection
  - Determines best path for outbound traffic
- It is general to use such as 90/100/110 to upstream/peer/customer respectively.
  - Do not use **expensive** upstream link as much as possible.
  - On the other hand, use **cheap** (ISP gotten money oppositely) link as much as possible. Maybe :-).
- Non-transitive attribute

Sample local-preference



- Preferred AS200_AS100 path in AS400

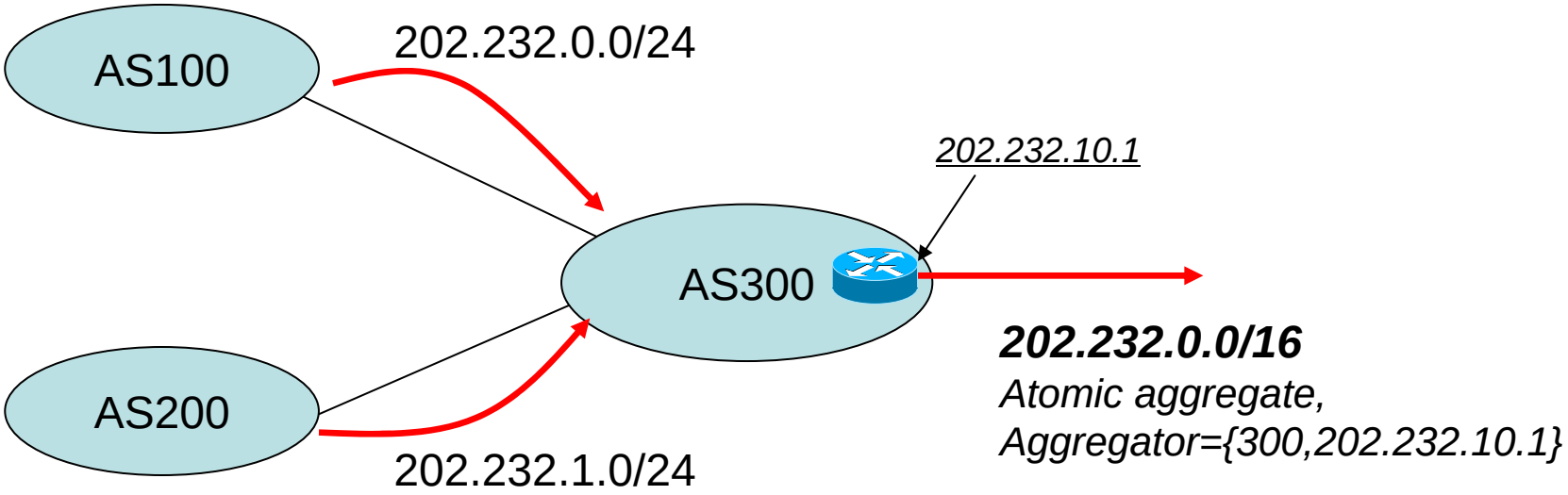
Atomic Aggregate attribute

- It is used by a BGP speaker to inform other BGP speakers that the local system selected a less specific route without selecting a more specific route which is included in it.
- It is unable to return to detailed original routes again
- We seldom use it in actual operation

Aggregator attribute

- The attribute contains the last AS number that formed the aggregate route (encoded as 2 octets), followed by the IP address of the BGP speaker that formed the aggregate route
- We seldom use it in actual operation

Sample Aggregation



- Atomic_Aggregate attribute and Aggregator attribute are set

Community Attribute

- RFC1997
- Route coloring
  - A community is a group of destinations which share some common property
  - Each autonomous system administrator may define which communities a destination belongs to. By default, all destinations belong to the general Internet community
- 32bit integer
  - Represented as two 16 bit integers (RFC1998)

Community attribute values (common)

- Well-known community

  no-export

  - Do not advertise to eBGP peers

  no-advertise

  - Do not advertise to any of peers

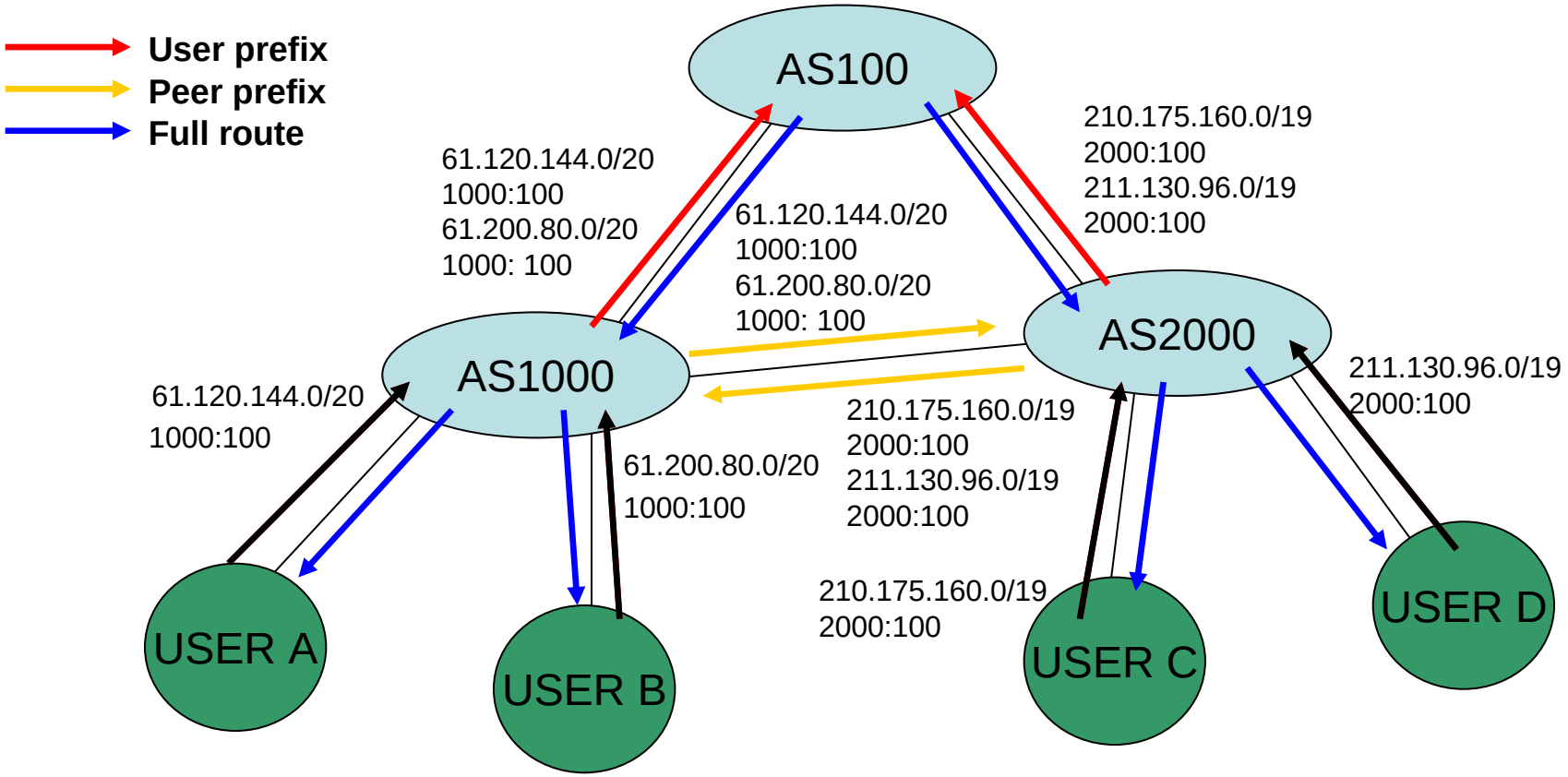
  Local-AS

  - Do not advertise outside local AS (only used with confederations)

Community attribute values (user area)

- Can set original community not reserved area per AS
  - Upper 16bit: AS number
  - Lower 16bit: original community value
  - Notation: ASnum:community value
    - e.g. 2914:100 customer prefix
      2914:200 peer prefix

Sample BGP community attribute



- Upstream ISP must advertise full route to down stream ISP
- User advertise own route to upstream ISP
- ISP B and ISP C must advertise their user route to upstream ISP and Peer ISP
- ISP B and ISP C should not advertise full route and peer's route to upstream.

BGP Best Path Selection Algorithm

1	Prefer highest weight attribute (CISCO original)
2	Prefer highest local_preference value
3	Prefer the path that was locally originated or via a network
4	Prefer the path with the shortest AS_PATH
5	Prefer the path with the lowest origin type (IGP < EGP < INCOMPLETE)
6	Prefer the lowest MED value
7	Prefer EBGP over internal IBGP path (if bestpath selected, go to STEP 9, multipath)
8	Prefer the path was the lowest IGP metric to the BGP next hop
9	Check if multiple paths need to be installed in the routing table for BGP multipath. (Continue, if bestpath is not selected yet)
10	When both paths are external, prefer the path that was received first (the oldest one)
11	Prefer router-ID
12	Prefer the path with the minimum cluster list length
13	Prefer the path coming from the lowest neighbor address

IRR and Route-server

IRR

- Background and needs
  - BGP routing information on internet does contain any proof.
  - Therefore, it is likely to interfere to the communication when it doesn't intend and the routing information is advertised by mistake. According to circumstances, there is a possibility that the routing information of the lie by malice flows.
  - The validity of the routing information can be confirmed by using IRR as one method of proving that routing information of BGP is accurate.
  - In a word, if information on IRR is referred, it can be judged whether the routing information of the correspondence is correct.
  - http://www.irr.net/
  - http://www.radb.net/

Who should use Route Registries? Why?

- For anyone who peers at public peering points, If you have customers who want to reach any of our sites (and these include most of the major US Government research laboratories), you need to register or use one of the big transit providers that do not filter your traffic.
- On more pragmatic level, it provides information of troubleshooting failures. If a customer reports that he can't reach 10.1.1.1 and it's not registered, it takes a LOT longer to figure out whom to call to resolve the problem.

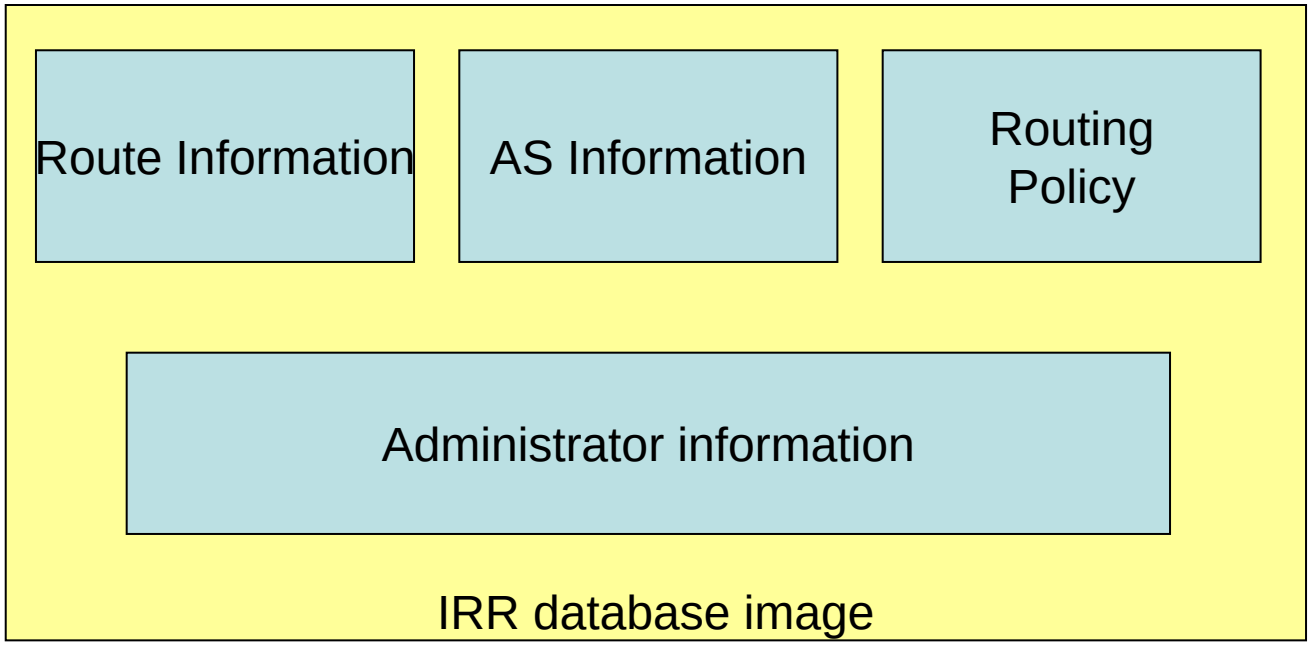
Ref: http://www.irr.net/docs/faq.html

IRRd

- The IRR (Internet Routing Registry) is the union of a growing number of world-wide routing policy databases that use the RPSL (Routing Policy Specification Language)
- IRRd is a stand-alone Internet Routing Registry database server. IRRd can store information and answer queries about local network, campus, and ISP backbone topology
- IRRd supports the RPSL registry syntax. As of version 2.2.0, IRRd also support the RPSLng IPv6 and multicast extensions to RPSL
- http://www.irrd.ner/

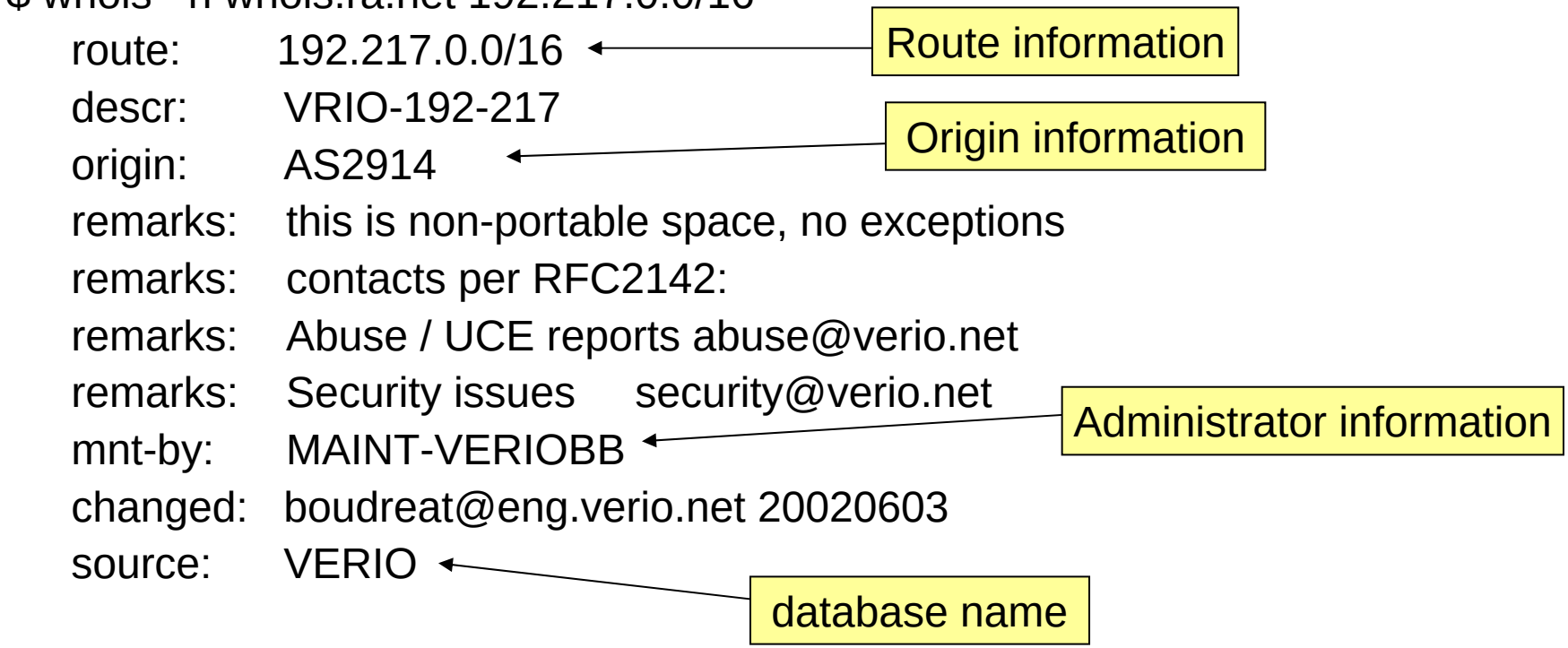
IRR database

- These various information is actually stored in shape "Object" in the IRR database
- Typical object information are “route object”, “AS object”, “routing policy”, and “Administrator object” existing



Sample: route object

```
$ whois –h whois.ra.net 192.217.0.0/16
    route:      192.217.0.0/16
    descr:      VRIO-192-217
    origin:     AS2914
    remarks:    this is non-portable space, no exceptions
    remarks:    contacts per RFC2142:
    remarks:    Abuse / UCE reports abuse@verio.net
    remarks:    Security issues     security@verio.net
    mnt-by:     MAINT-VERIOBB
    changed:    boudreat@eng.verio.net 20020603
    source:     VERIO
```



RPSL RFC

- RFC1786: RIPE-181

  Recommended reading
  - Representation of IP Routing Policies in a Routing Registry
- RFC2622

  Highly Recommended reading
  - Routing Policy Specification Language (RPSL)
- RFC2650

  Highly Recommended reading
  - Using RPSL in Practice
- RFC2725

  Recommended reading
  - Routing Policy System Security
- RFC2726

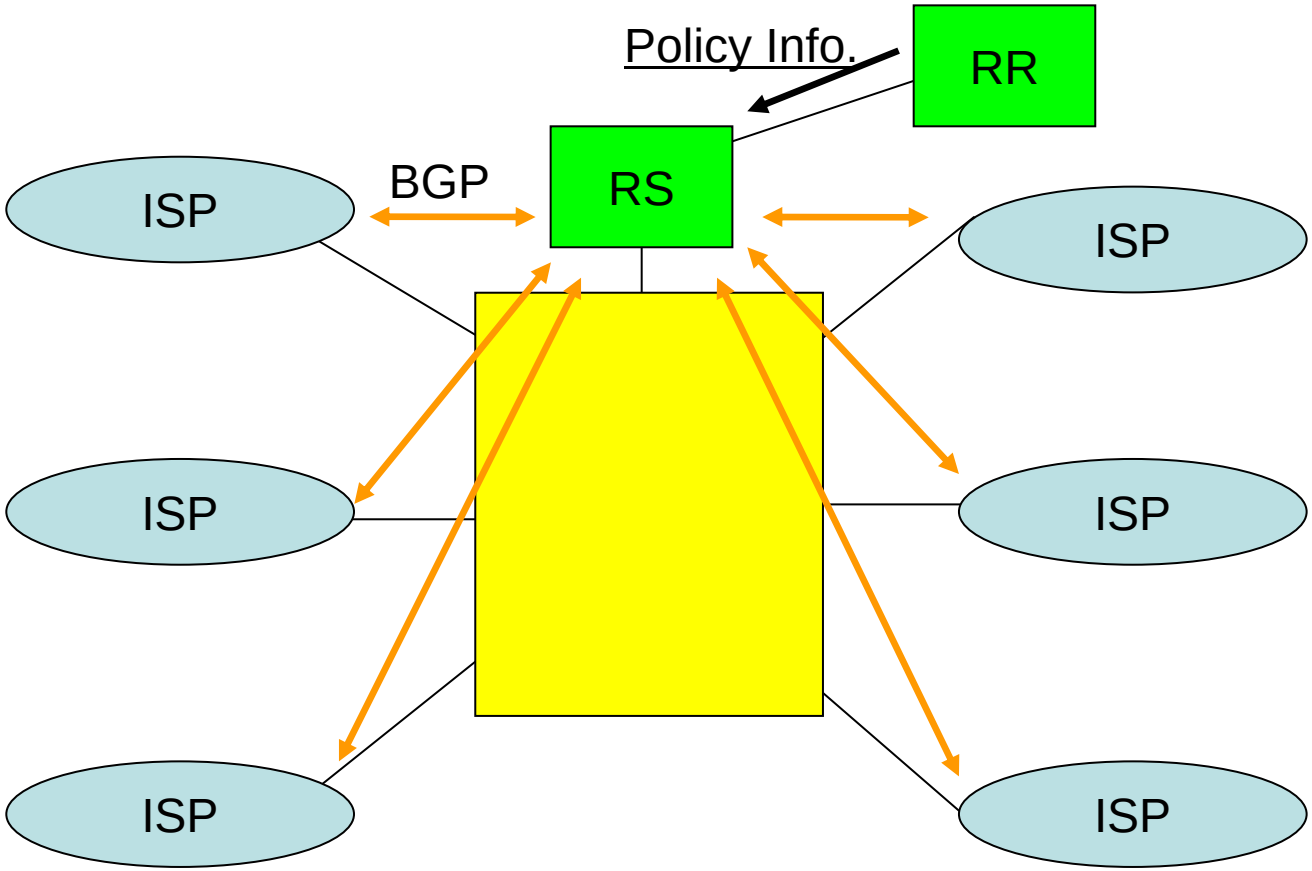
  Operational reading
  - PGP Authentication for RIPE Database Updates
- RFC2769

  Operational reading
  - Routing Policy System Replication

Routing Registry and Root Server

- Separate process packet forwarding and route selection
- Routing Registry
  - ASes routing policy database
- Root Server
  - it have BGP peering session with ISP who connected I.X. in a 2nd level.
  - Calculate each ISP routing table from policy on the Routing Registry

RR and RS



Major route server

- AT&T telnet://route-server.jp.att.net
- BBN Planet telnet://ner-routes.bbnplanet.net
- CERF telnet://route-server.cerf.net
- Exodus (US) telnet://route-server.exodus.net
- Exodus (ASIA) telnet://route-server-ap.exodus.net
- Exodus (EU) telnet://route-server-eu.exodus.net
- Oregon-IX telnet://route-views.oregon-ix.net

bogon routes

bogon routes - introduction

- A bogon prefix is a route that should never appear in the Internet routing table
- A packet routed over the public Internet (not including over VPN or other tunnels) should never have a source address in a bogon range
- These are commonly found as the source addresses of DDoS attacks.

bogon routes

- There are a variety of ways to track the bogons and updated IANA allocations

  http://www.iana.org/assignments/ipv4-address-space

- BGP Peering Bogon Tracking

  http://www.cymru.com/BGP/bogon-rs.html

- MAINT-BOGON-FILTERS in radb

  http://www.radb.net/cgi-bin/radb/whois.cgi?obj=MAINT-BOGON-FILTERS

  or do ‘whois –h whois.ra.net MAINT-BOGON-FILTERS’

BGP4
overview and operation

End of Tutorial

Many Thanks to:
Toshiya Asaba asaba@iij.ad.jp
Philip Smith pfs@cisco.com