An Introduction to Interdomain Routing and the Border Gateway Protocol (BGP)

Eduardo Grampín Castro
grampin@fing.edu.uy
Universidad de la República
Uruguay

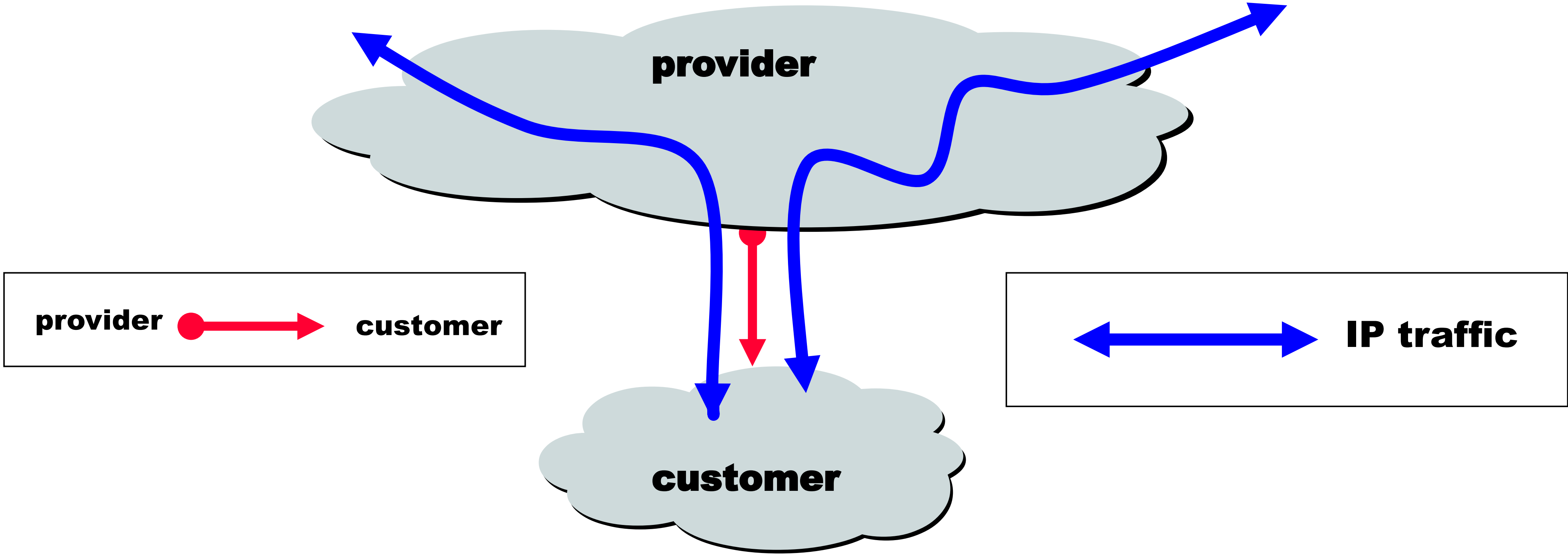
based on slides by Timothy Griffin, Alberto García & Geoff Huston

Outline

- Relationships Among Networks and Interdomain Routing
- Implementing Inter-Network Relationships with BGP
- BGP: a bit of theory
- BGP Scalability
- (Inter + Intra) domain Routing

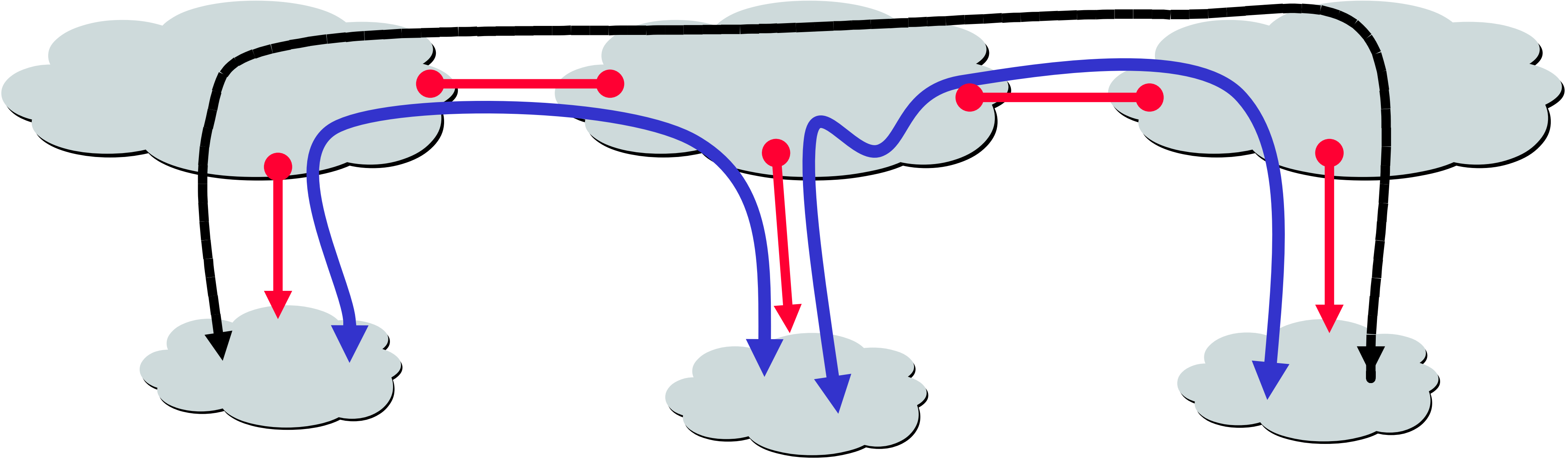
Relationships Among Networks and Interdomain Routing

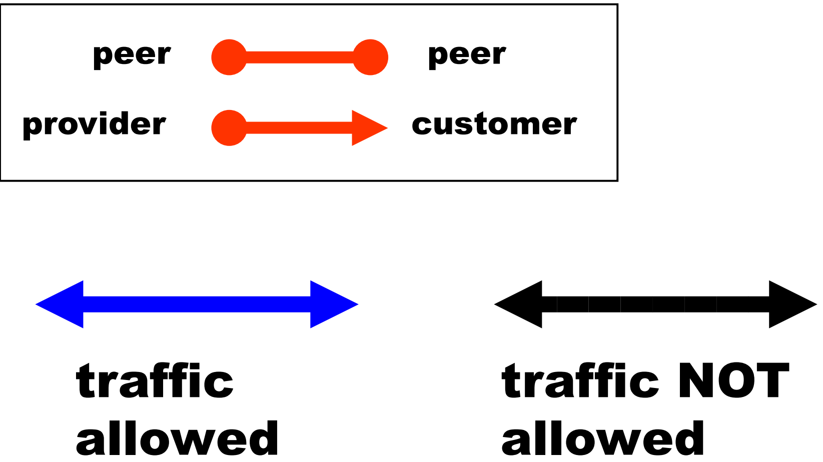
Customers and Providers



Customer pays provider for access to the Internet

The “peering” relationship



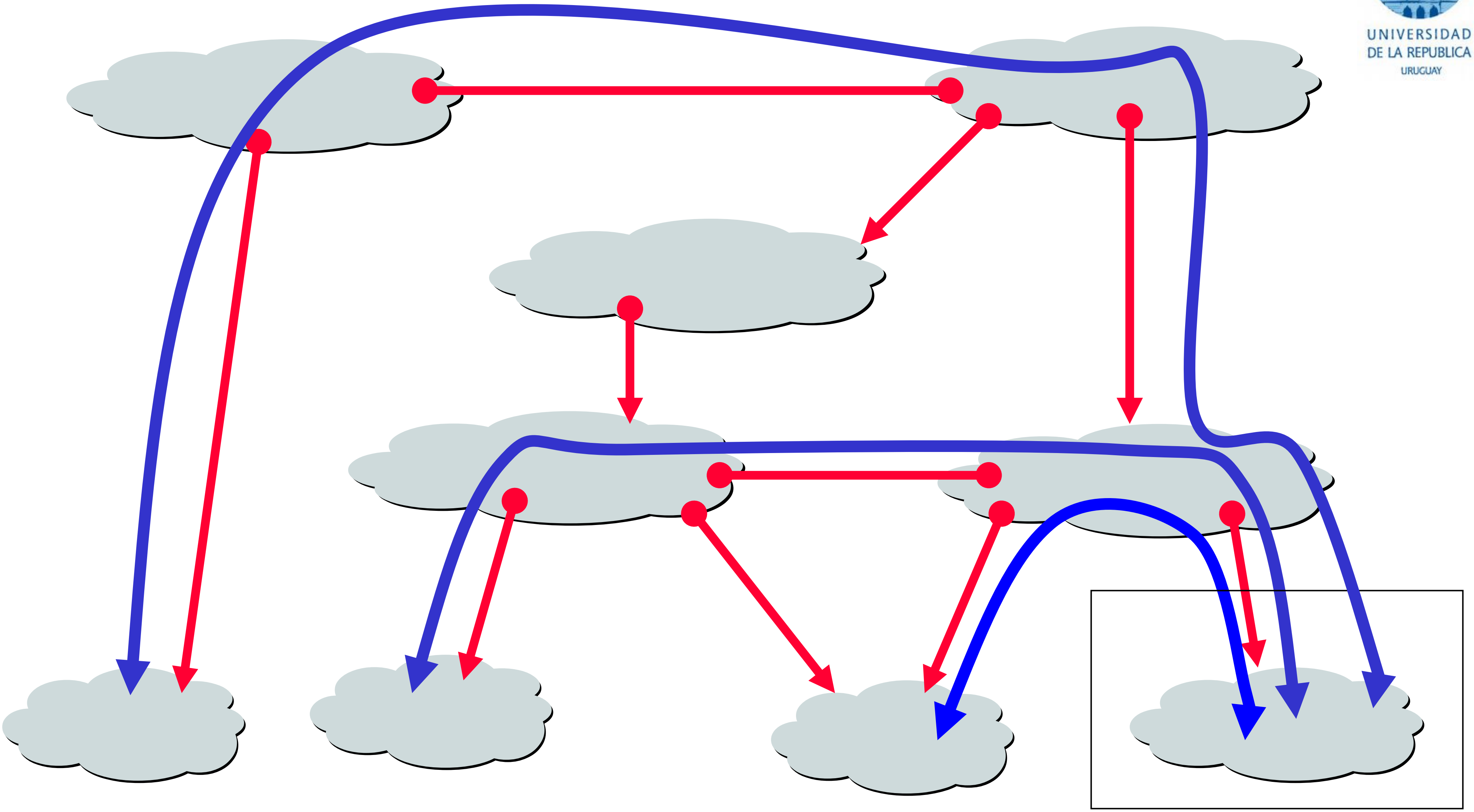


Peers provide transit between their respective customers

<u>Peers do not provide transit between peers</u>

Peers (often) do not exchange $$$

Peering: routing shortcuts



Peering also allows connectivity between the customers of “Tier 1” providers.

peer	●—●	**peer**
provider	●→	**customer**

To peer or not to peer

Peer

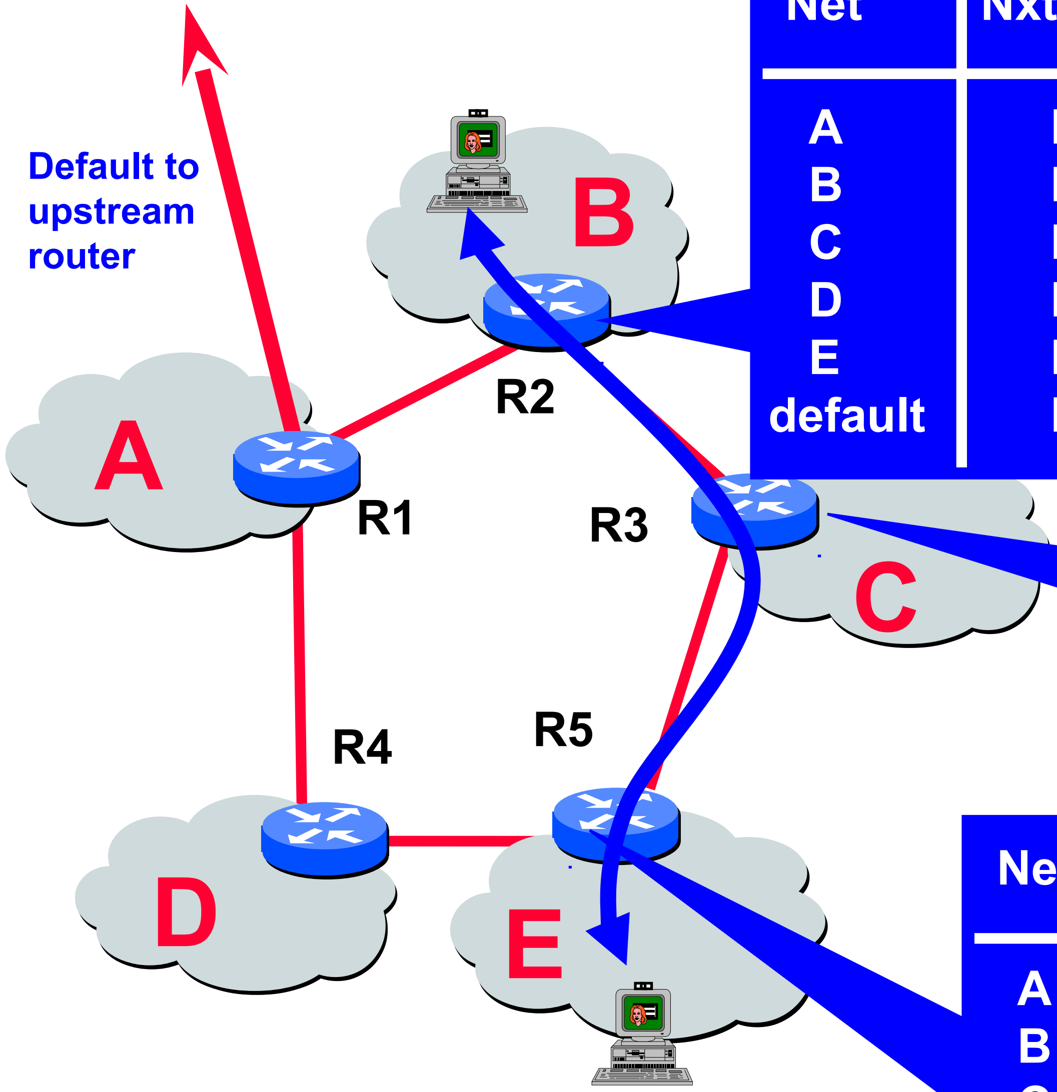
- **Reduces upstream transit costs**
- **Can increase end-to-end performance**
- **May be the only way to connect your customers to some part of the Internet (“Tier 1”)**

Don’t Peer

- **You would rather have customers**
- **Peers are usually your competition**
- **Peering relationships may require periodic renegotiation**

Peering agreements are often confidential

Routing vs. Forwarding



Net	Nxt Hop
A	R1
B	Direct
C	R3
D	R1
E	R3
default	R1

Forwarding always works

Routing can be badly broken

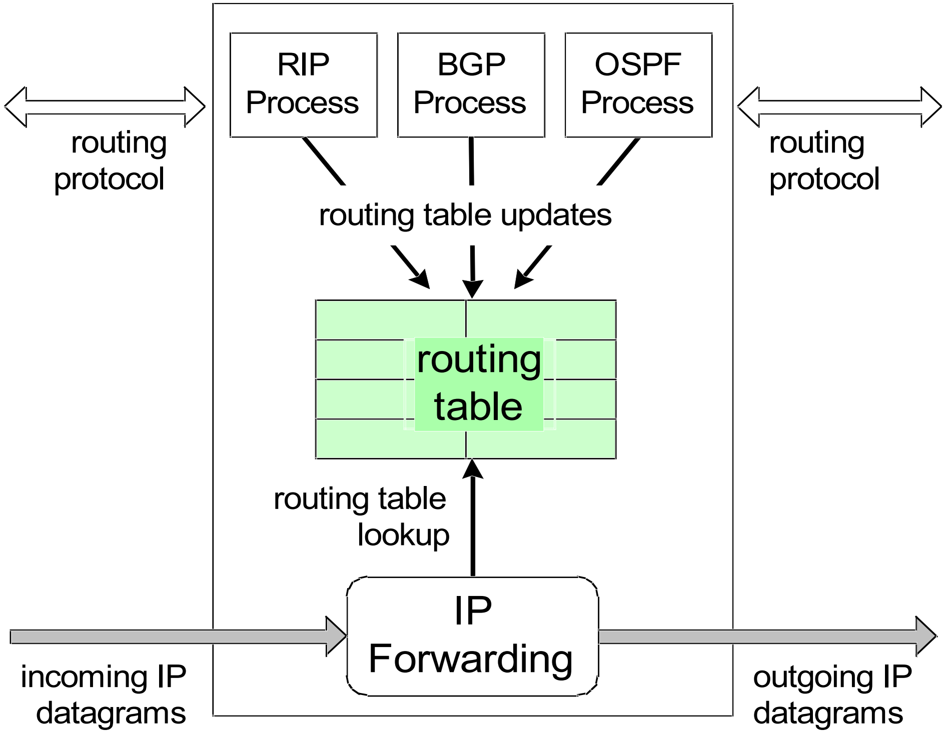
Net	Nxt Hop
A	R2
B	R2
C	Direct
D	R5
E	R5
default	R2

Net	Nxt Hop
A	R4
B	R3
C	R3
D	R4
E	Direct
default	R4

Forwarding: determine next hop

Routing: establish end-to-end paths

Routing vs. Forwarding



How Are Forwarding Tables Populated to implement Routing?

Statically

Administrator manually configures forwarding table entries

+ More control
+ Not restricted to destination-based forwarding
- Doesn’t scale
- Slow to adapt to network failures

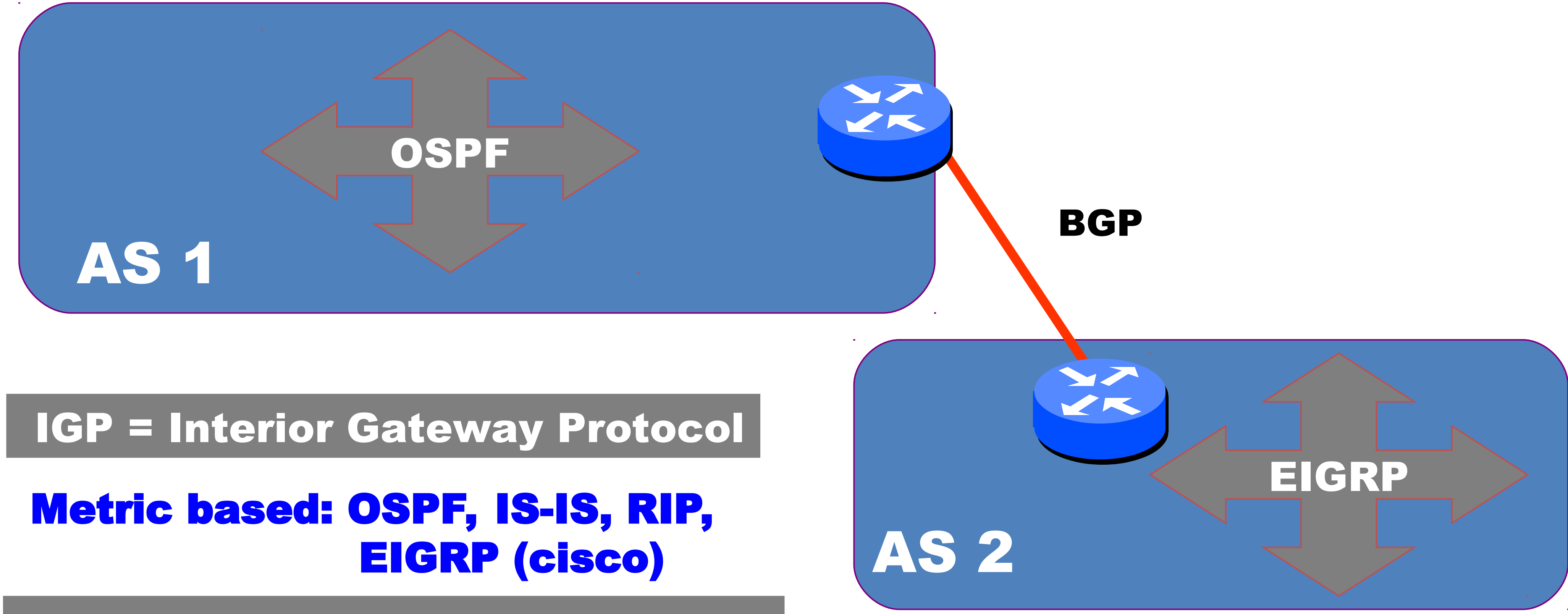
Dynamically

Routers exchange network reachability information using ROUTING PROTOCOLS. Routers use this to compute best routes

+ Can rapidly adapt to changes in network topology
+ Can be made to scale well
- Complex distributed algorithms
- Consume CPU, Bandwidth, Memory
- Debugging can be difficult
- Current protocols are destination-based

In practice : a mix of these.
Static routing mostly at the “edge”

Architecture of Dynamic Routing



IGP = Interior Gateway Protocol

Metric based: OSPF, IS-IS, RIP, EIGRP (cisco)

EGP = Exterior Gateway Protocol

Policy based: BGP

The Routing Domain of BGP is the entire Internet

Technology of Distributed Routing

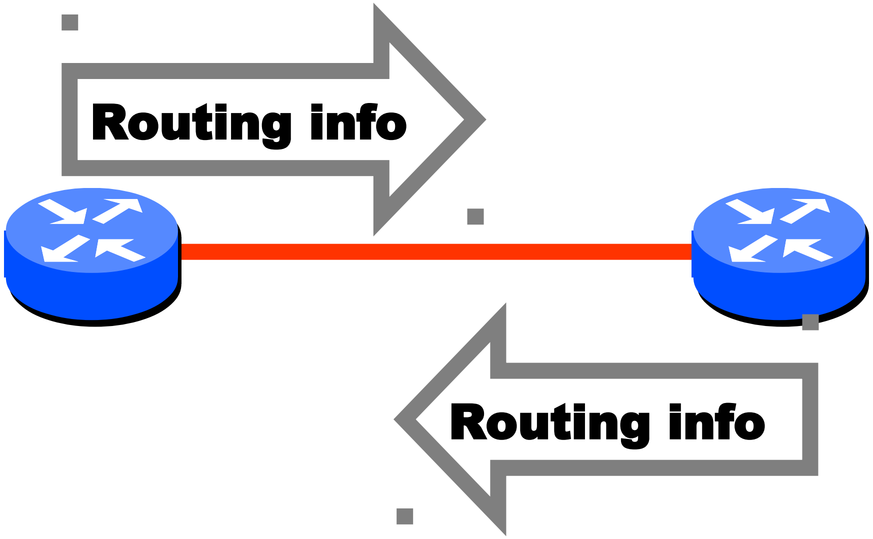
Link State

- Topology information is flooded within the routing domain
- Best end-to-end paths are computed locally at each router.
- Best end-to-end paths determine next-hops.
- Based on minimizing some notion of distance
- Works only if policy is shared and uniform
- Examples: OSPF, IS-IS

Vectoring

- Each router knows little about network topology
- Only best next-hops are chosen by each router for each destination network.
- Best end-to-end paths result from composition of all next-hop choices
- Does not require any notion of distance
- Does not require uniform policies at all routers
- Examples: RIP, BGP

Routers Talking to Routers



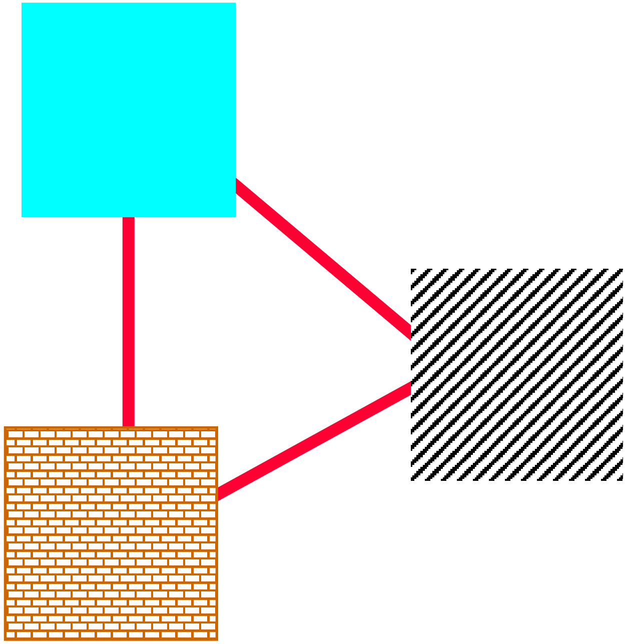
- Routing computation is distributed among routers within a routing domain
- Computation of best next hop based on routing information is the most CPU/memory intensive task on a router
- Routing messages are usually not routed, but exchanged via layer 2 between physically adjacent routers (internal BGP and multi-hop external BGP are exceptions)

Autonomous Systems and Routing Domains

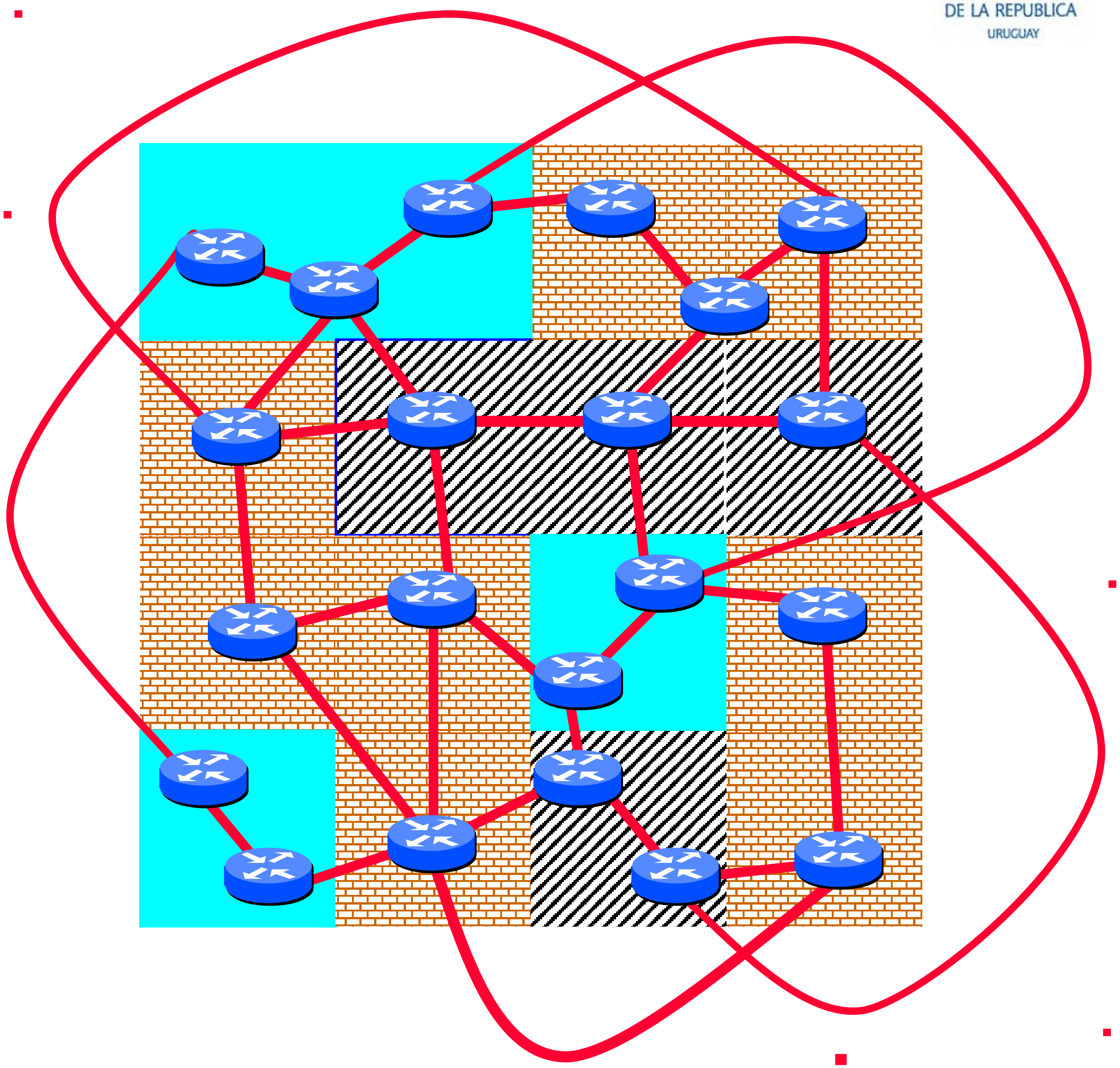
- Routing domain:
  - A collection of physical networks glued together using IP, that have a unified administrative routing policy (campus networks, corporate networks, …)
- Autonomous Systems (ASes):
  - An autonomous system is a routing domain that has been assigned an Autonomous System Number (ASN).
  - "… the administration of an AS appears to other ASes to have a single coherent interior routing plan and presents a consistent picture of what networks are reachable through it." (RFC 1930: Guidelines for creation, selection, and registration of an Autonomous System)

AS Graph != Internet Topology

BGP was designed to throw away information!

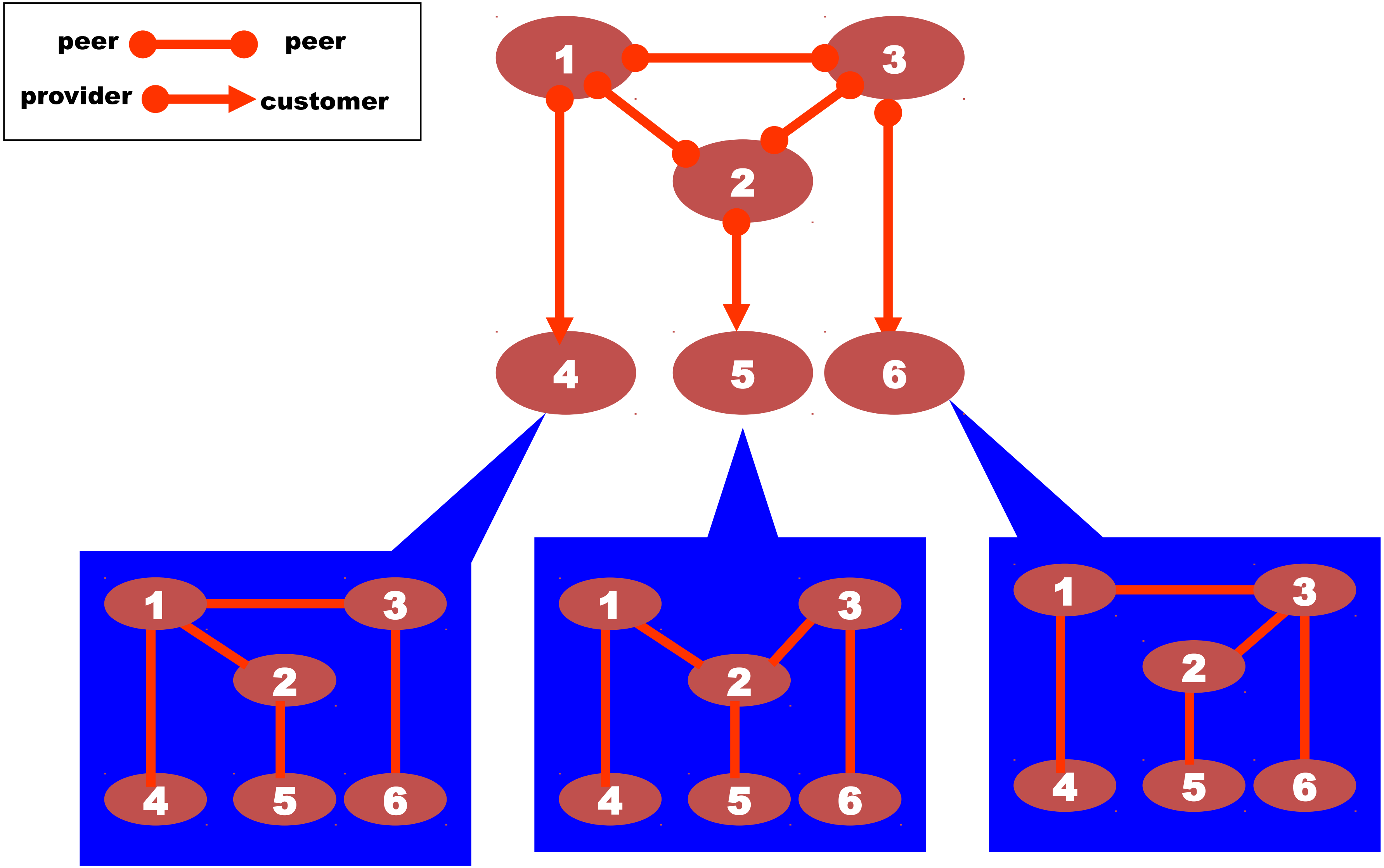


The AS graph may look like this.



Reality may be closer to this…

AS Graphs depend on the point of view



References

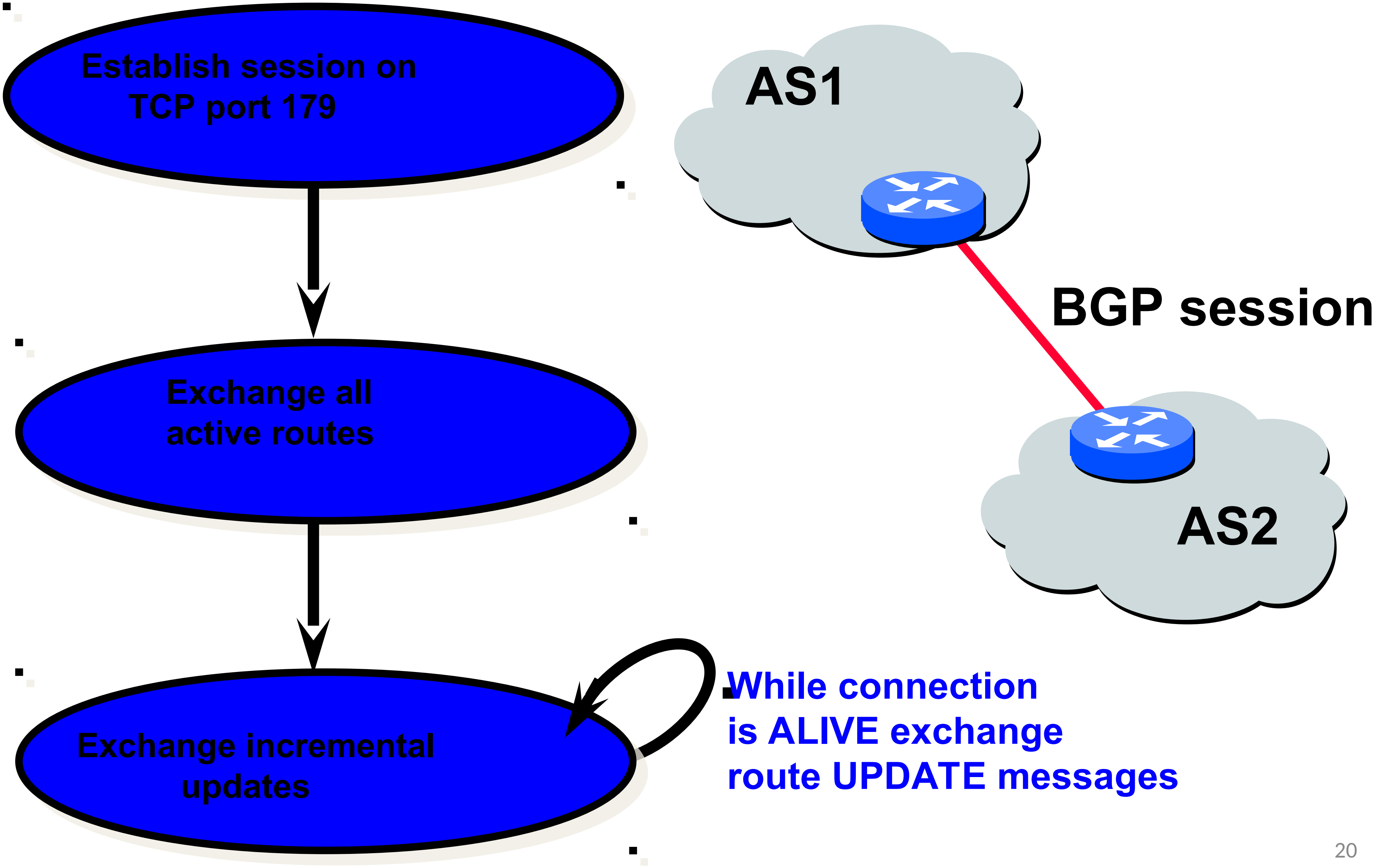
1. Lixin Gao, *“On Inferring Autonomous System Relationships in the Internet”*. IEEE/ACM TRANSACTIONS ON NETWORKING, VOL. 9, NO. 6, DECEMBER 2001.
2. Lixin Gao, Jennifer Rexford, *“Stable internet routing without global coordination”*. IEEE/ACM Transactions on Networking (TON) Volume 9 Issue 6, December 2001.

Implementing Inter-Network Relationships with BGP

BGP-4

- BGP = Border Gateway Protocol
- Is a Policy-Based routing protocol
- Is the de facto EGP of today's global Internet
- Relatively simple protocol, but configuration is complex and the entire world can see, and be impacted by, your mistakes.
  - 1989 : BGP-1 [RFC 1105]
    - Replacement for EGP (1984, RFC 904)
  - 1990 : BGP-2 [RFC 1163]
  - 1991 : BGP-3 [RFC 1267]
  - 1995 : BGP-4 [RFC 1771]
    - Support for Classless Interdomain Routing (CIDR)
  - 2006: BGP-4 [RFC 4271, obsoletes 1771]

BGP Operations (Simplified)



Establish session on TCP port 179

Exchange all active routes

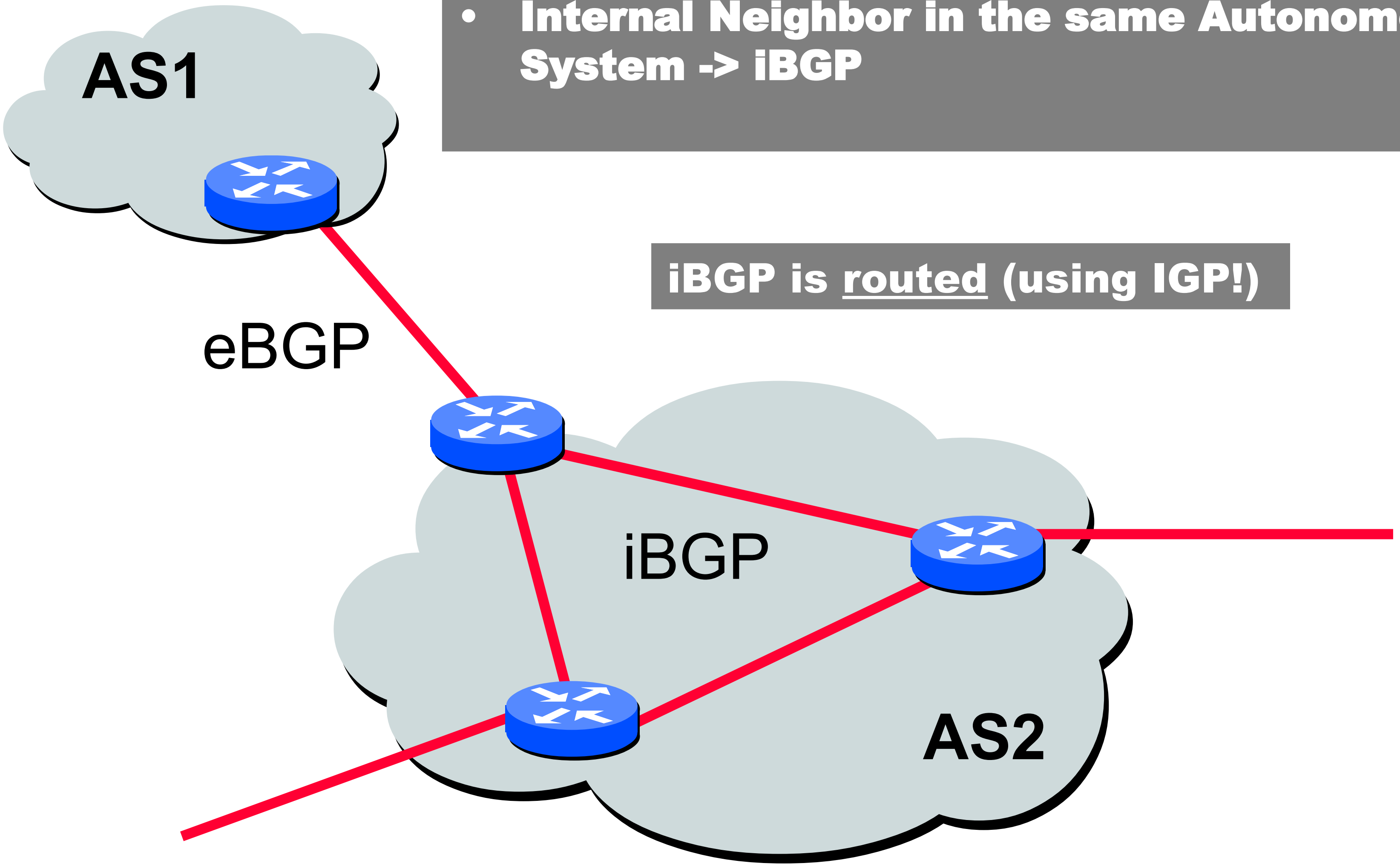
Exchange incremental updates

While connection is ALIVE exchange route UPDATE messages

Two Types of BGP Neighbor Relationships

- **External Neighbor in a different Autonomous Systems -> eBGP**
- **Internal Neighbor in the same Autonomous System -> iBGP**

iBGP is <u>routed</u> (using IGP!)



Four Types of BGP Messages

- Open : Establish a peering session.
- Keep Alive : Handshake at regular intervals.
- Notification : Shuts down a peering session.
- Update : Announcing new routes or withdrawing previously announced routes.

announcement

=

prefix + attributes values

BGP Attributes

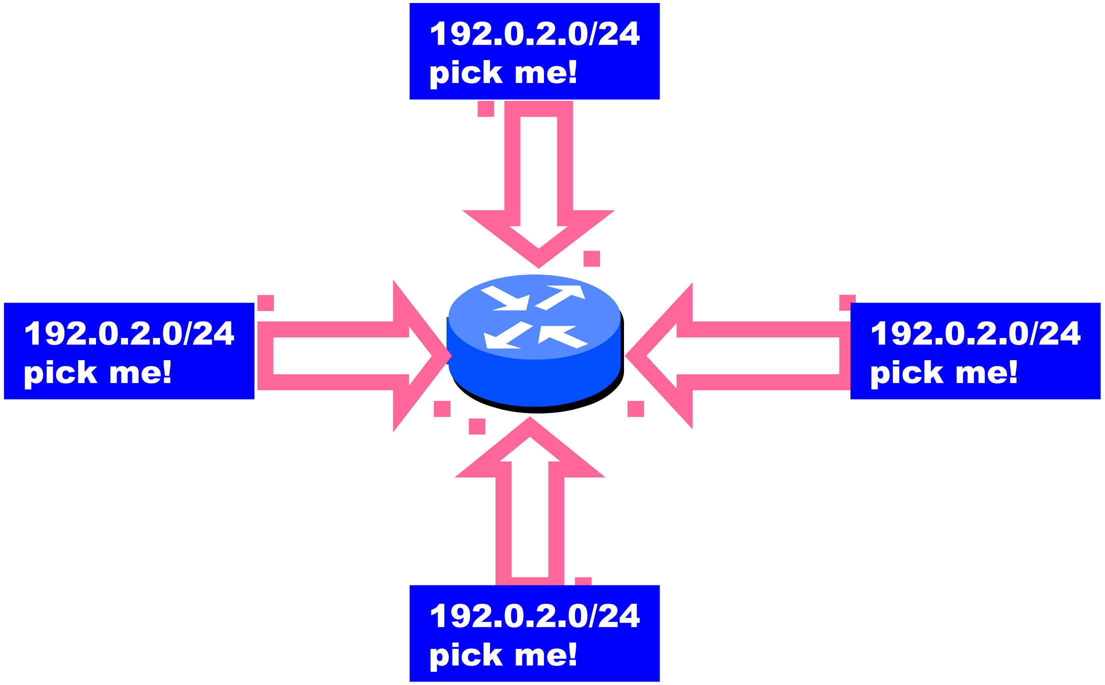
Value	Code	Reference
1	ORIGIN	[RFC1771]
2	AS_PATH	[RFC1771]
3	NEXT_HOP	[RFC1771]
4	MULTI_EXIT_DISC	[RFC1771]
5	LOCAL_PREF	[RFC1771]
6	ATOMIC_AGGREGATE	[RFC1771]
7	AGGREGATOR	[RFC1771]
8	COMMUNITY	[RFC1997]
9	ORIGINATOR_ID	[RFC2796]
10	CLUSTER_LIST	[RFC2796]
11	DPA	[Chen]
12	ADVERTISER	[RFC1863]
13	RCID_PATH / CLUSTER_ID	[RFC1863]
14	MP_REACH_NLRI	[RFC2283]
15	MP_UNREACH_NLRI	[RFC2283]
16	EXTENDED COMMUNITIES	[Rosen]
...		
255	reserved for development	

Most important attributes

From IANA: http://www.iana.org/assignments/bgp-parameters

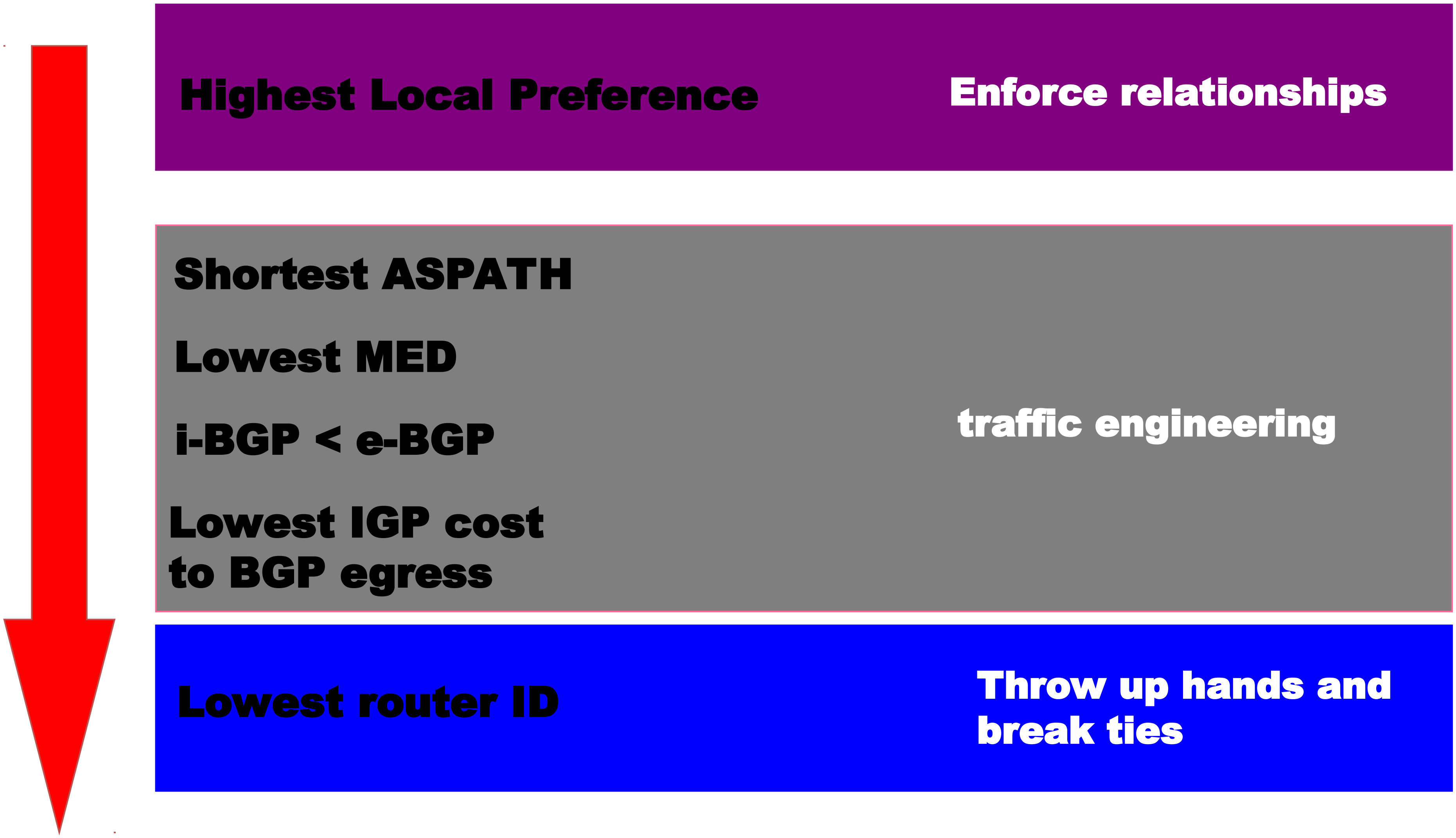
Not all attributes need to be present in every announcement

Attributes are Used to Select Best Routes



- BGP chooses only one path to reach the destination
- BGP propagates the best path to its neighbours
- BGP stores the non-selected routes to be able to recover them if needed

Route Selection Summary



Highest Local Preference

Enforce relationships

Shortest ASPATH

Lowest MED

i-BGP < e-BGP

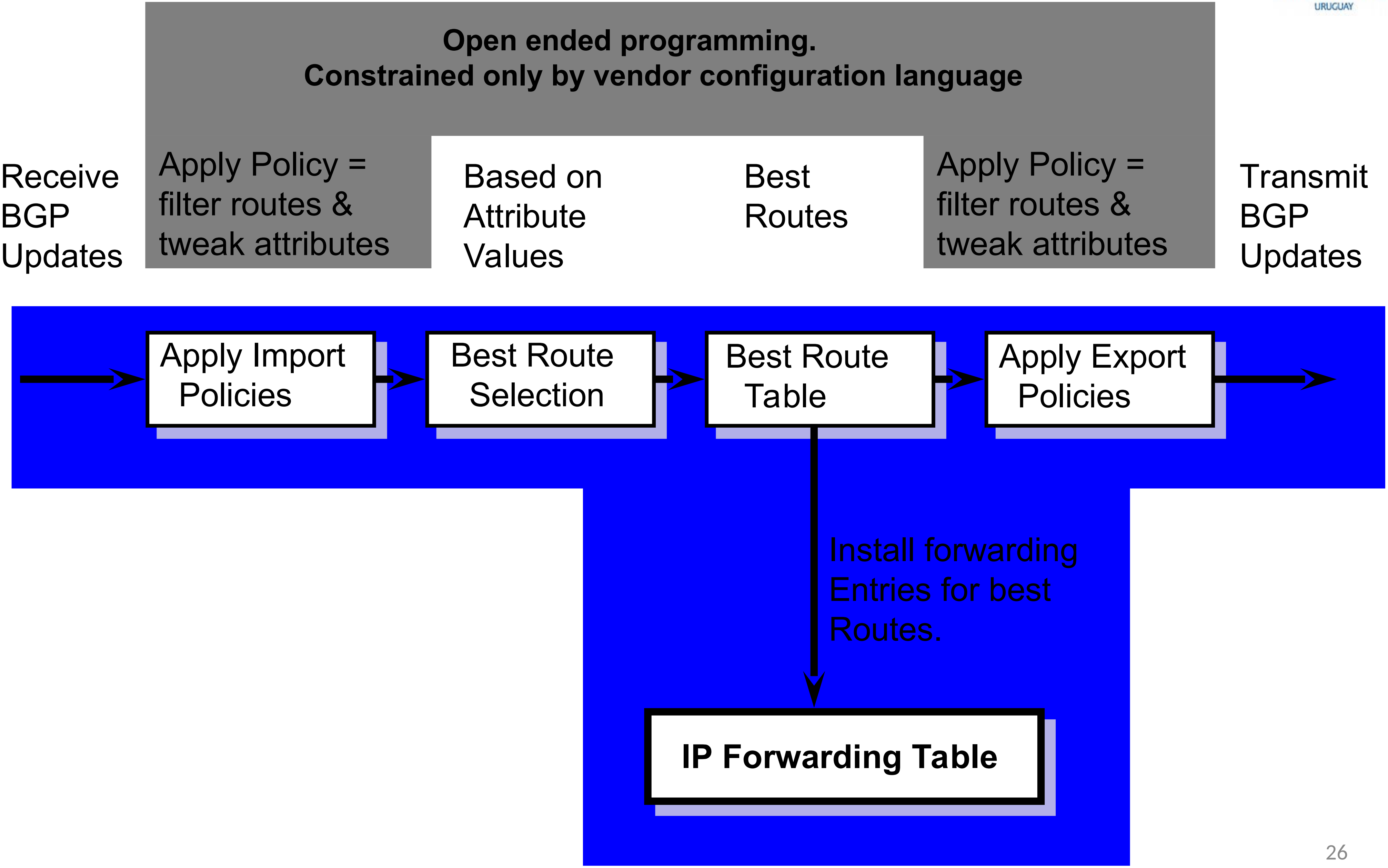
Lowest IGP cost to BGP egress

traffic engineering

Lowest router ID

Throw up hands and break ties

BGP Route Processing



Open ended programming.
Constrained only by vendor configuration language

Receive BGP Updates

Apply Policy = filter routes & tweak attributes

Based on Attribute Values

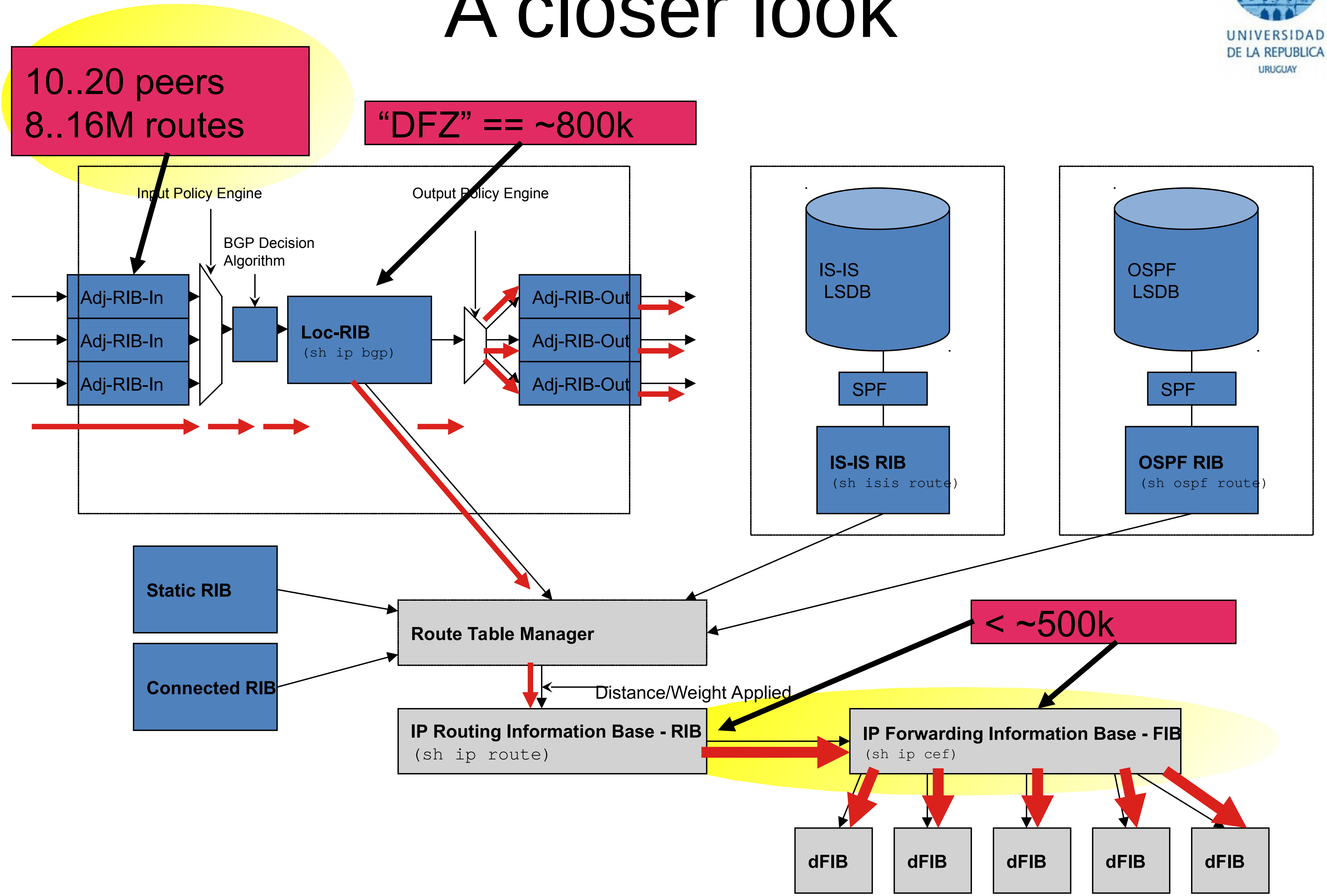
Best Routes

Apply Policy = filter routes & tweak attributes

Transmit BGP Updates

Install forwarding Entries for best Routes.

A closer look



show ip bgp

```
route-views>sh ip bgp
BGP table version is 4228790247, local router ID is 128.223.51.103
Status codes: s suppressed, d damped, h history, * valid, > best, i - internal,
              r RIB-failure, S Stale
Origin codes: i - IGP, e - EGP, ? - incomplete

   Network          Next Hop            Metric LocPrf Weight Path
*  1.9.0.0/16       144.228.241.130                        0 1239 3320 4788 i
*                   194.85.40.15                           0 3267 9002 4788 i
*                   194.85.102.33                          0 3277 3216 1273 4788 i
*                   64.71.255.61                           0 812 1273 4788 i
*                   154.11.98.225            0             0 852 3320 4788 i
*                   154.11.11.113            0             0 852 3320 4788 i
*                   209.124.176.223                        0 101 101 11164 4788 i
*                   69.31.111.244          183             0 4436 1273 4788 i
*                   66.185.128.48            7             0 1668 7018 4788 i
*                   129.250.0.11           302             0 2914 4788 i
*                   4.69.184.193             0             0 3356 1273 4788 i
*                   207.172.6.1              0             0 6079 4788 4788 i
*                   193.0.0.56                             0 3333 8218 4788 i
*                   65.106.7.139             3             0 2828 3549 4788 i
*                   208.74.64.40                           0 19214 26769 1273 4788 i
*                   207.46.32.34                           0 8075 4788 i
*>                  12.0.1.63                              0 7018 4788 i
*                   216.218.252.164                        0 6939 4788 i
*                   207.172.6.20             0             0 6079 2914 4788 i
*
*
```

Source: telnet route-views.routeviews.org

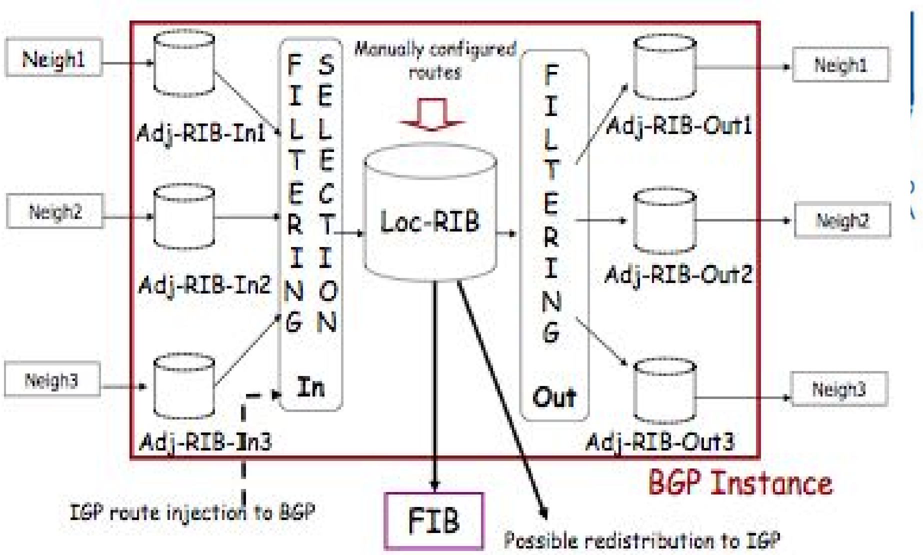
Full table: http://bgp.potaroo.net/as2.0/bgptable.txt

show ip route

```
route-views>show ip route 1.9.0.0 255.255.0.0
Routing entry for 1.9.0.0/16
  Known via "bgp 6447", distance 20, metric 0
  Tag 7018, type external
  Last update from 12.0.1.63 2w4d ago
  Routing Descriptor Blocks:
  * 12.0.1.63, from 12.0.1.63, 2w4d ago
      Route metric is 0, traffic share count is 1
      AS Hops 2
      Route tag 7018
```

Source: telnet route-views.routeviews.org

BGP operation model explained



- Consider a prefix advertised by peer1
  - The prefix is stored in the “database” Adj-RIB-In1 (adjacent Routing Information Base input from 1)
  - Then, the prefix is filtered (suppose not)
  - When a new prefix arrives, selection process is started again. For that, info for the same prefix in other Adj-RIB-In’s is considered
  - Suppose this prefix is preferred. Then, information is stored in the Local Routing Information Base, and then installed in the Forwarding Information Base (=IP forwarding table)
    - Note that maybe some translation is required – consider NEXT_HOP example
  - Then, the outgoing filter decides to which neighbors it must be advertised
    - The fact that it has been advertised is stored, to know to which neighbors send future withdraws or changes in the route

Basic Processing of BGP Routes

1. Input selection:
   - Filter received routes, delete non-acceptable routes
     - Routes with loops (loop detection)
     - Unacceptable routes (private addresses, non-allocated addresses)
     - Route filtered due to a policy (policy-based filtering)
       - Prefix
       - AS_PATH
       - COMMUNITY
     - Unstable routes
2. Route-selection algorithm:
   - Select the best route from the set of routes
     - Applying policy
3. Output selection:
   - Decide which routes to propagate to the peers
     - Applying policy

Outgoing route filtering and business model

- The business model suggests:
  - Never carry traffic between two of your providers
    - To do this, don’t advertise (filter out) to providers routes received from providers
  - Never carry traffic between a peer and a provider (and vice versa)
    - To do this, don’t advertise (filter out) to providers routes received from peers, and filter to peers route received from providers.
  - Send as much traffic as you can to your clients
    - Do not configure any filter (out) involving clients

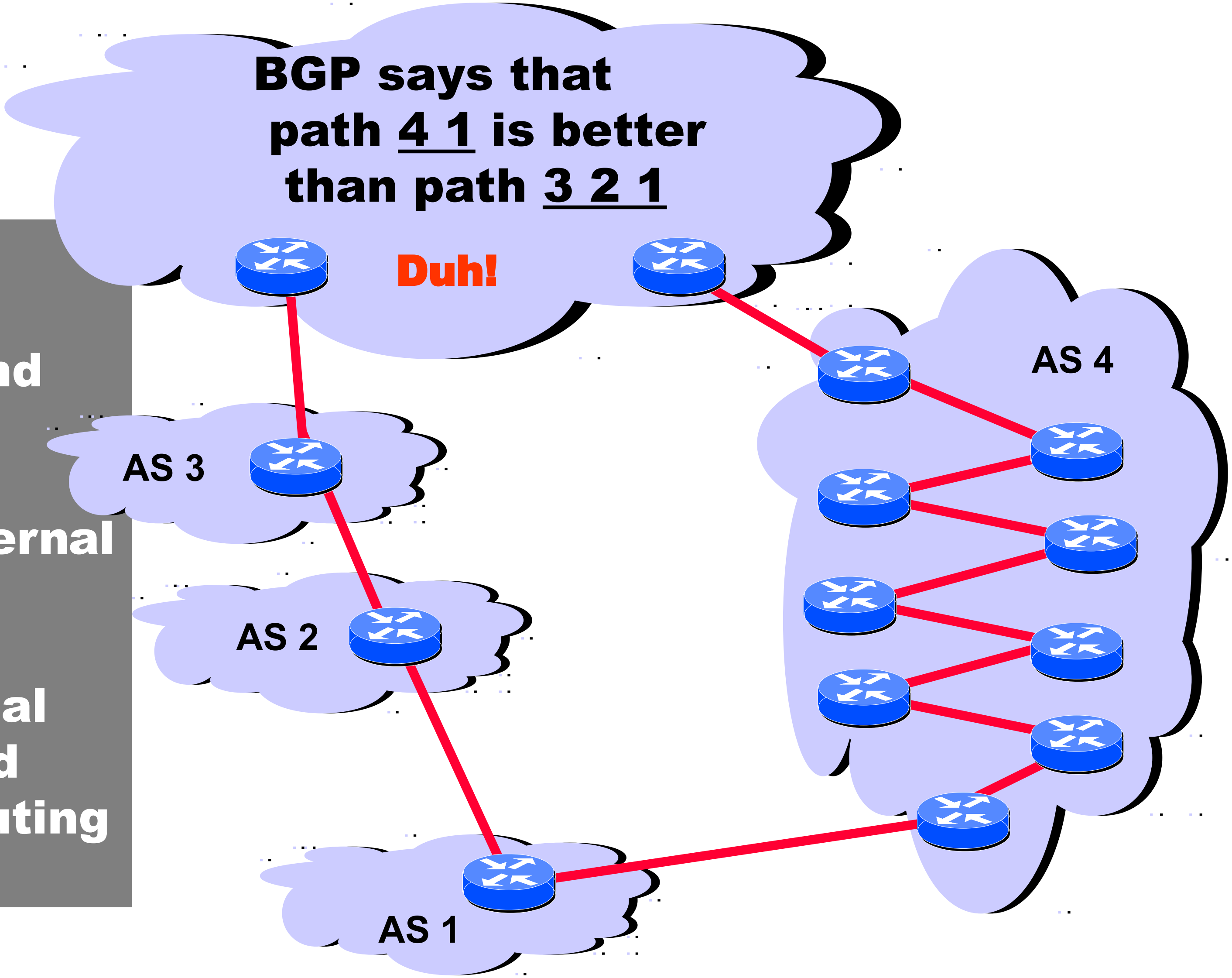
BGP Path Attributes: AS_PATH

- Contains the AS numbers traversed by the announced route
  - In so-called path segments (one per AS)
- For each UPDATE message passed along to another AS (EBGP):
  - The AS prepends (=inserts at the beginning) its AS number to the list of path segments
    - List must remain unchanged if UPDATE passed to a router within the AS (iBGP)
- Sequence of path segments
  - A path segment is described with:
    - Type (AS_SET or AS_SEQUENCE)
    - Length of the path segment (# of AS in the path segment)
    - Values (one or more AS numbers)
- This allows you to:
  - Apply routing policies based on the transit AS
  - Detect loops: if the receiving AS is already contained in the path

AS_PATH: beware

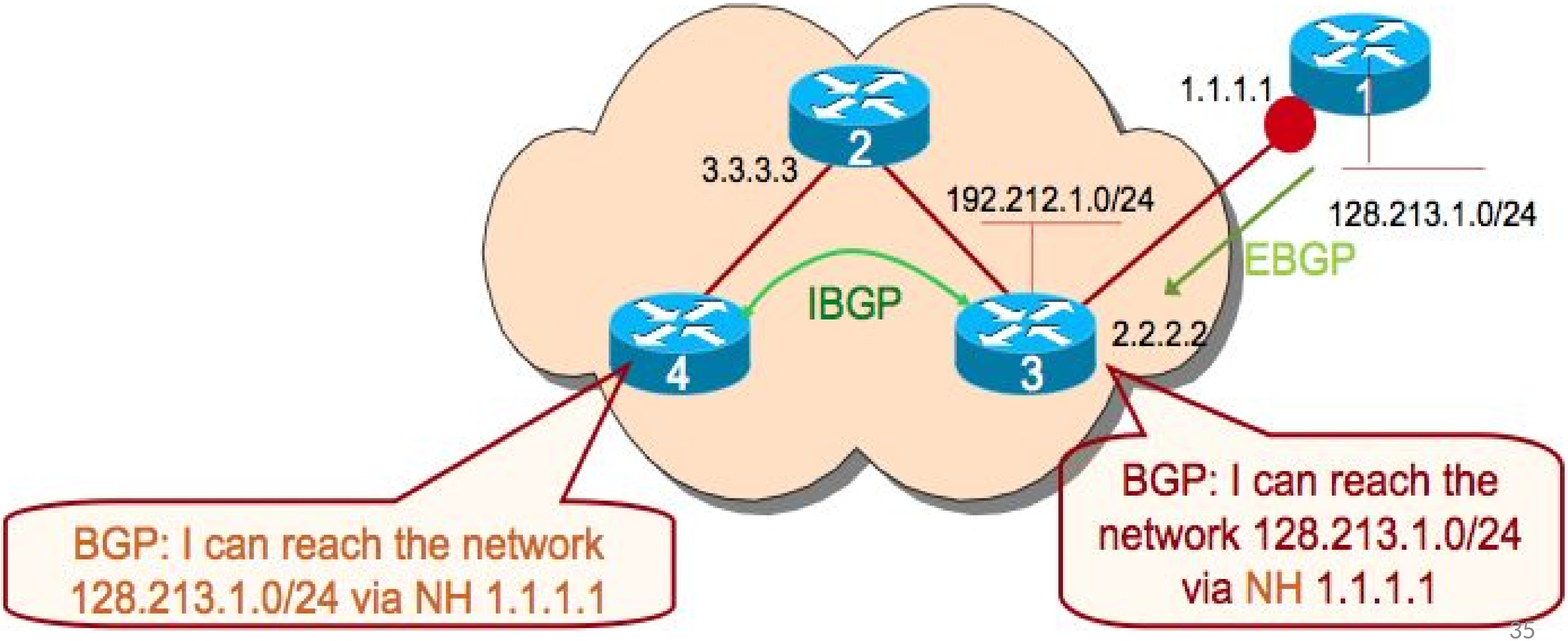
In fairness: could you do this “right” and still scale?

Exporting internal state would dramatically increase global instability and amount of routing state



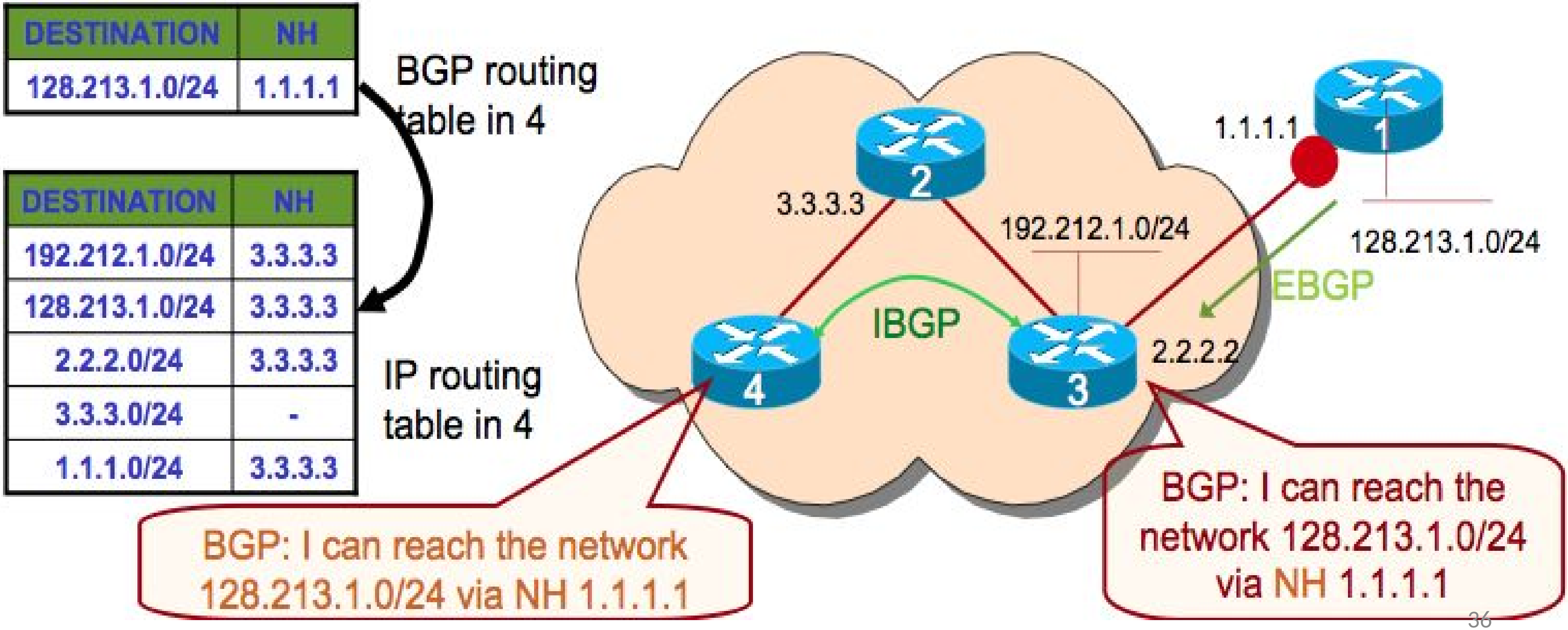
BGP Path Attributes: NEXT_HOP

- NEXT_HOP shows the IP address of the border router that provides access to the announced routes
- In the example, a route generated in router 1 is propagated to 3 (eBGP) and 4 (iBGP)
  - NH for both routers is 1.1.1.1
- If 4 propagates the route outside, it should insert its outgoing interface IP address as NEXT_HOP


1
1.1.1.1
2
3.3.3.3
192.212.1.0/24
128.213.1.0/24
EBGP
IBGP
4
3
2.2.2.2
BGP: I can reach the network 128.213.1.0/24 via NH 1.1.1.1
BGP: I can reach the network 128.213.1.0/24 via NH 1.1.1.1

BGP Path Attributes: NEXT_HOP

- The NEXT_HOP info along with the IP routing table is processed to generate a new entry in the IP routing table
- An entry for the NEXT_HOP must exist in the IP routing table (either through IGP or statically)
  - For example: R4 must know (through IGP or static route) how to route to 1.1.1.1



BGP routing table in 4

DESTINATION	NH
128.213.1.0/24	1.1.1.1

IP routing table in 4

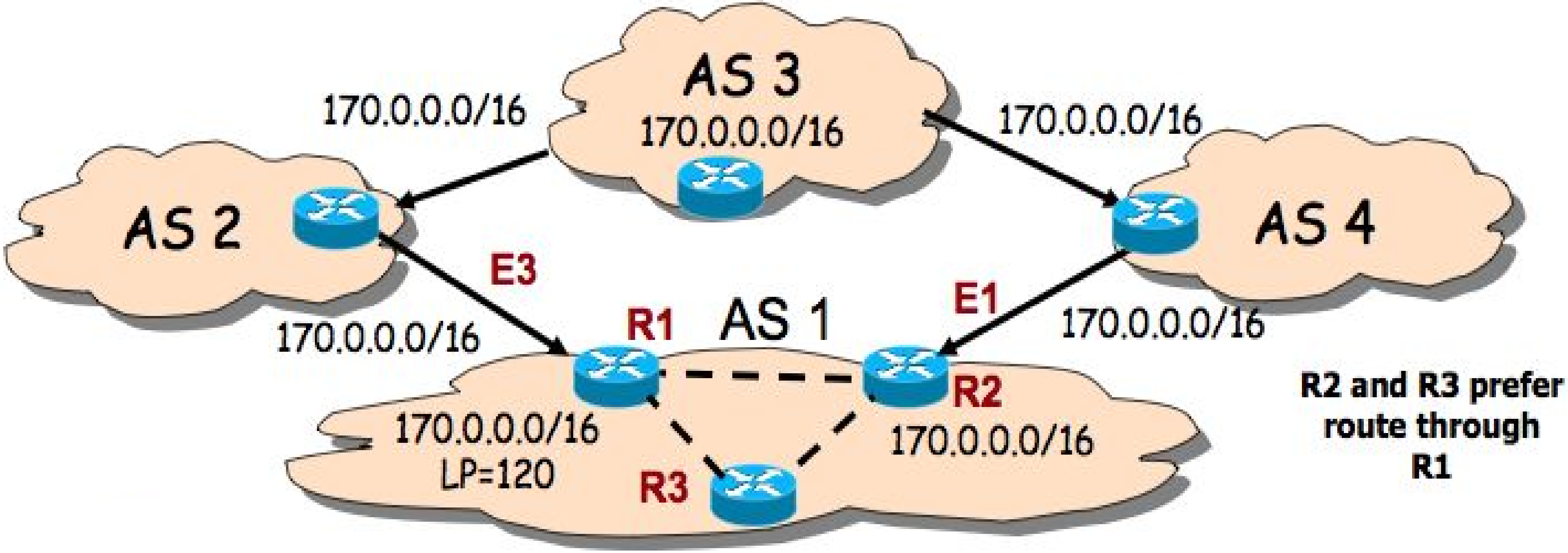
DESTINATION	NH
192.212.1.0/24	3.3.3.3
128.213.1.0/24	3.3.3.3
2.2.2.0/24	3.3.3.3
3.3.3.0/24	-
1.1.1.0/24	3.3.3.3

BGP: I can reach the network 128.213.1.0/24 via NH 1.1.1.1

BGP: I can reach the network 128.213.1.0/24 via NH 1.1.1.1

BGP Path Attributes: LOCAL_PREF

- Aim: allow the propagation of link preference for some external prefix inside an AS
  - It is configured in a single router, and it is propagated through iBGP to all internal peers
  - Prefer routes with the highest local preference
  - Default value of 100 (i.e. if it is not explicitly set, it is equivalent to 100)
  - Note: it is only used inside a given AS (it is only transmitted by iBGP)

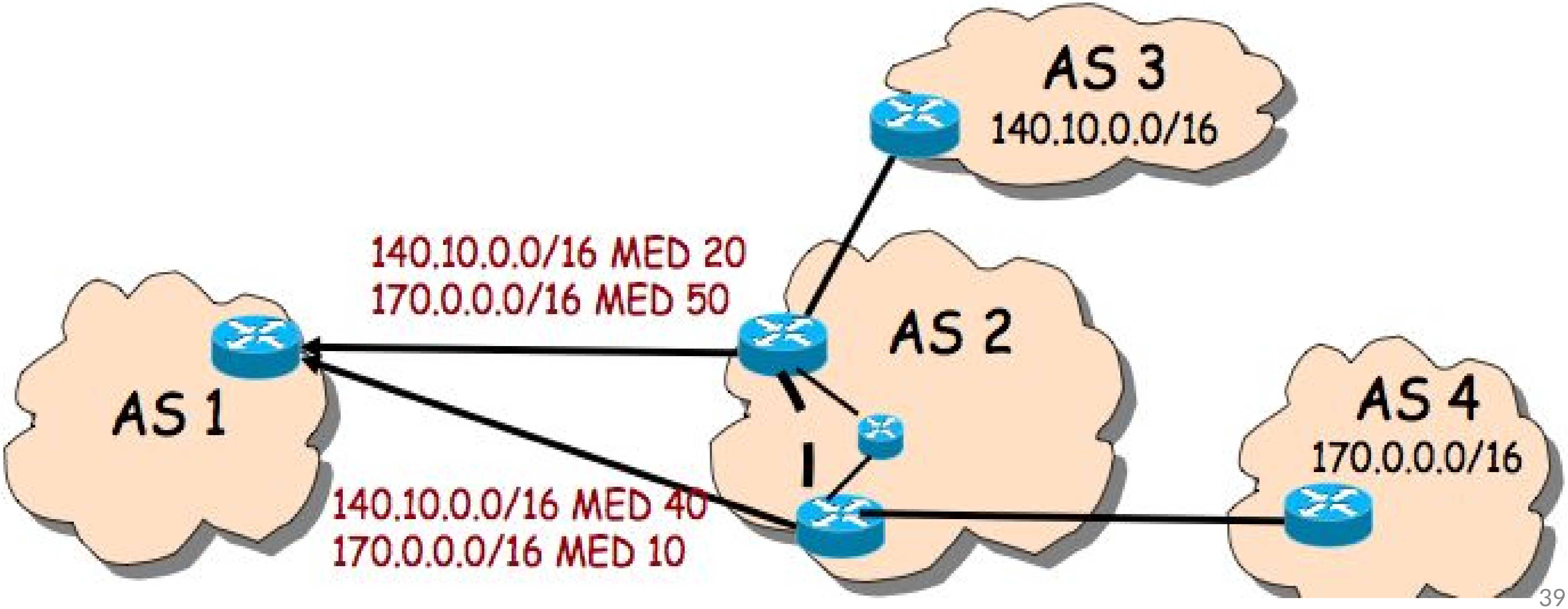


LOCAL_PREF to enforce business model

- The usual business model suggest the following route preference:
  - Prefer always routes to clients
    - When you send traffic through that link, you obtain PROFIT
  - If not, prefer routes to peers
    - When you send traffic through that link, you do at LOW COST, through short path to destination...
  - Else, transit
    - HIGH COST
- This behavior is enforced by proper configuration of LOCAL_PREF

MULTI-EXIT DISCRIMINATOR (MED)

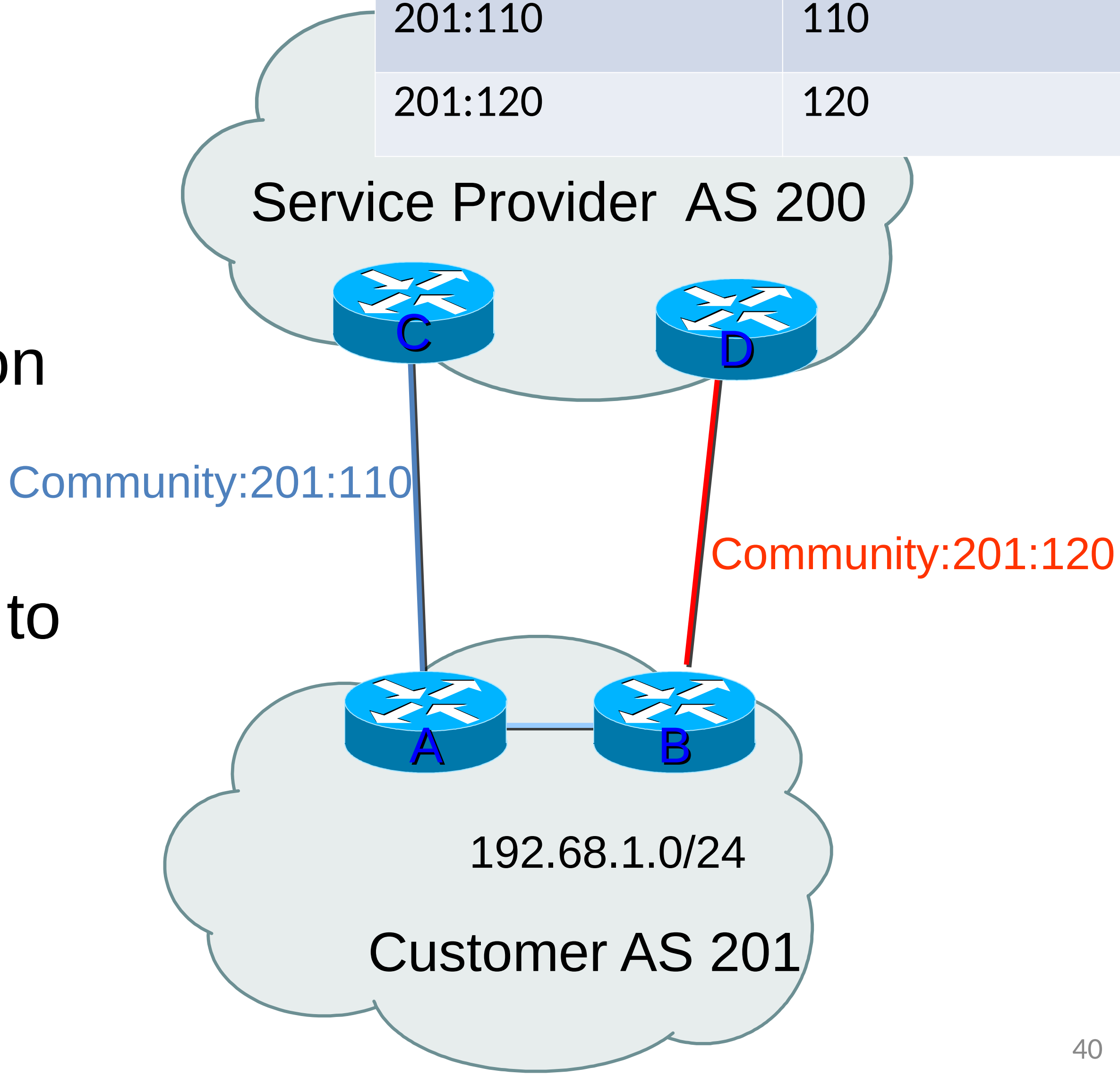
- Allows an AS to suggest to its neighbours a preferred connection (when multiple exist) for a given route
  - Distance metric: Always prefer the lower value
  - In principle it discriminates between routes with equal AS_PATH values
  - The metric is local between the two ASs, it is not propagated further



BGP Path Attributes:COMMUNITY

- COMMUNITY value:
  - Group of destinations sharing common properties
  - 32 bit number acting as a tag to qualify a route

Community	Local Preference
201:110	110
201:120	120

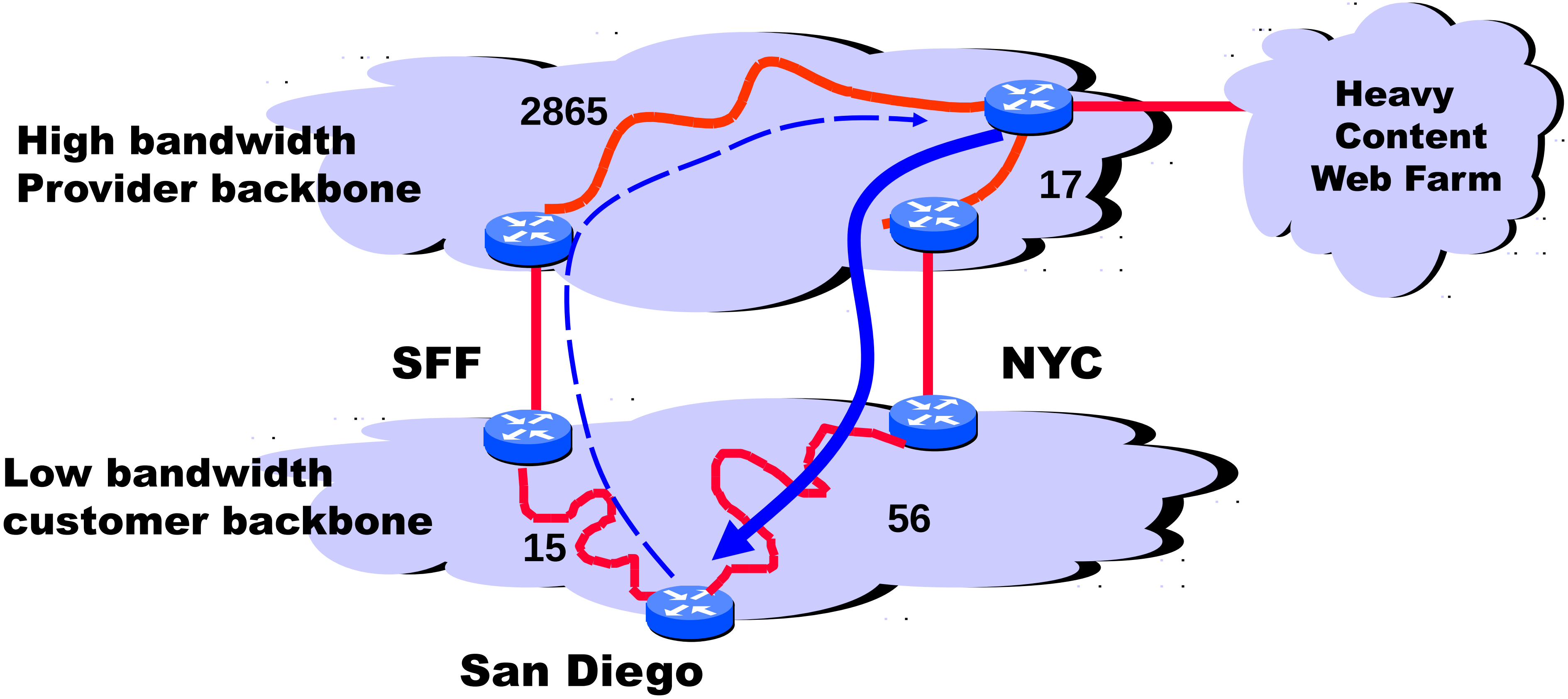


Review: BGP Route Selection Rules

- Remember: selection applies for the SAME PREFIX

1. If NEXT_HOP is not available (there is no route in the IP forwarding table), ignore the route.
2. Delete routes with lower LOCAL_PREF
   - Implementation of local policy -> trustworthy.
3. Delete routes with longest AS_PATH (larger amount of AS to transit).
   - Very much applied.
4. Delete routes with higher ORIGIN.
   - IGP<EGP<Incomplete
5. Delete routes with higher MED (coming from the same AS)
   - If two routes come from the same AS, it is probable that they will have the same AS_PATH, on the contrary rule 3 would not have been applied.
6. Delete routes that were learnt by IBGP, if there are routes learnt by EBGP.
   - Send traffic to the exterior if it is possible.
7. Delete routes to NEXT_HOP with higher costs.
   - Note that only considers AS own metric.
   - Hot potato: send traffic to the faster way to exterior.
8. Tie break: prefer routes announced by router with lower BGP identifier.

Hot potato

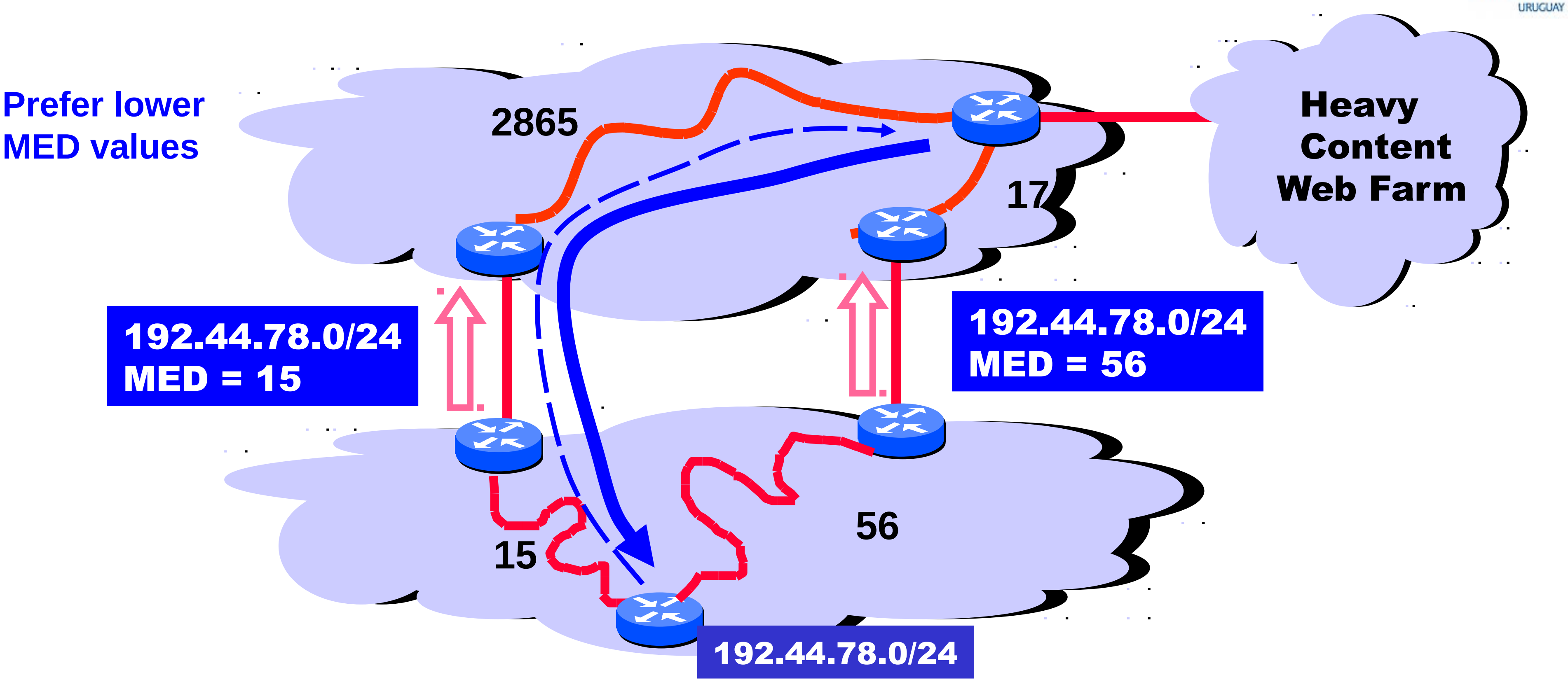


Remember: only for outbound traffic

- - - → **tiny http request**

━━━▶ **huge http reply**

Cold Potato: MED



Remember: MEDs (are considered) BEFORE IGP distance!

Note1 : some providers will not listen to MEDs

Note2 : MEDs need not be tied to IGP distance

References

- Books
    1. *Internet Routing Architectures*. Sam Halal, Danny McPherson. Cisco Press. 2000.
    2. *Routing in the Internet*. Christian Huitema. Prentice Hall. 2000.
- RFCs
    1. RFC 4271. *A Border Gateway Protocol 4 (BGP-4)*. Y. Rekhter, T. Li. S. Hares. Jan 06.
    2. RFC 4274. *BGP-4 Protocol Analysis*. D. Meyer, K. Patel. Jan 2006
    3. RFC 4276. *BGP-4 Implementation Report*. S. Hares, A. Retana. Jan 2006.
    4. RFC 4277. *Experience with the BGP-4 Protocol*. D. McPherson, K. Patel. Jan 06.
    5. RFC 1930. *Guidelines for creation, selection and registration of an Autonomous System (AS).*

BGP: a bit of theory

Awake!

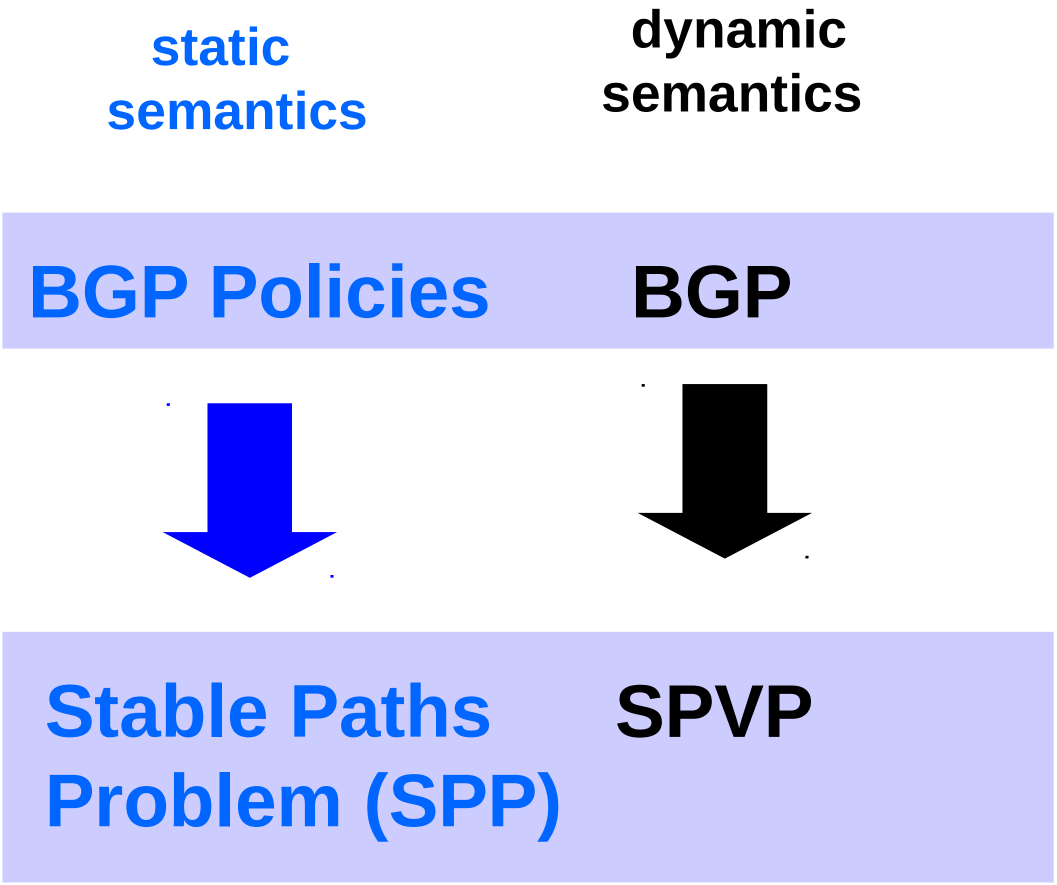
- BGP is not guaranteed to converge on a stable routing.
  - Policy interactions could lead to “livelock” protocol oscillations. See “Persistent Route Oscillations in Inter-domain Routing” by K. Varadhan, R. Govindan, and D. Estrin.

- Corollary: *BGP is not guaranteed to recover from network failures*

What Problem is BGP Solving?

Underlying problem	Distributed means of computing a solution.
Shortest Paths	RIP, OSPF, IS-IS
X?	BGP

Separate dynamic and static semantics



static semantics

dynamic semantics

BGP Policies

BGP

Stable Paths Problem (SPP)

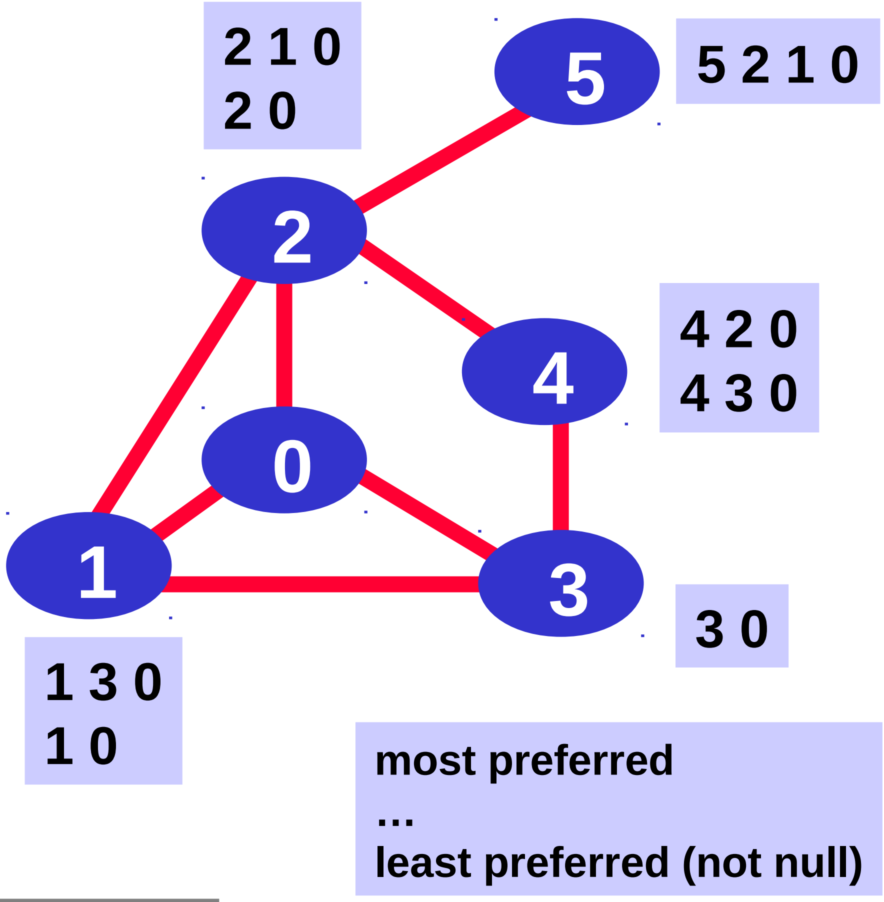
SPVP

SPVP = Simple Path Vector Protocol = a distributed algorithm for solving SPP

See [Griffin, Shepherd, Wilfong]

An instance of the Stable Paths Problem (SPP)

- **A graph of nodes and edges,**
- **Node 0, called *the origin*,**
- **For each non-zero node, a set or permitted paths to the origin. This set always contains the “null path”.**
- **A ranking of permitted paths at each node. Null path is always least preferred. (Not shown in diagram)**



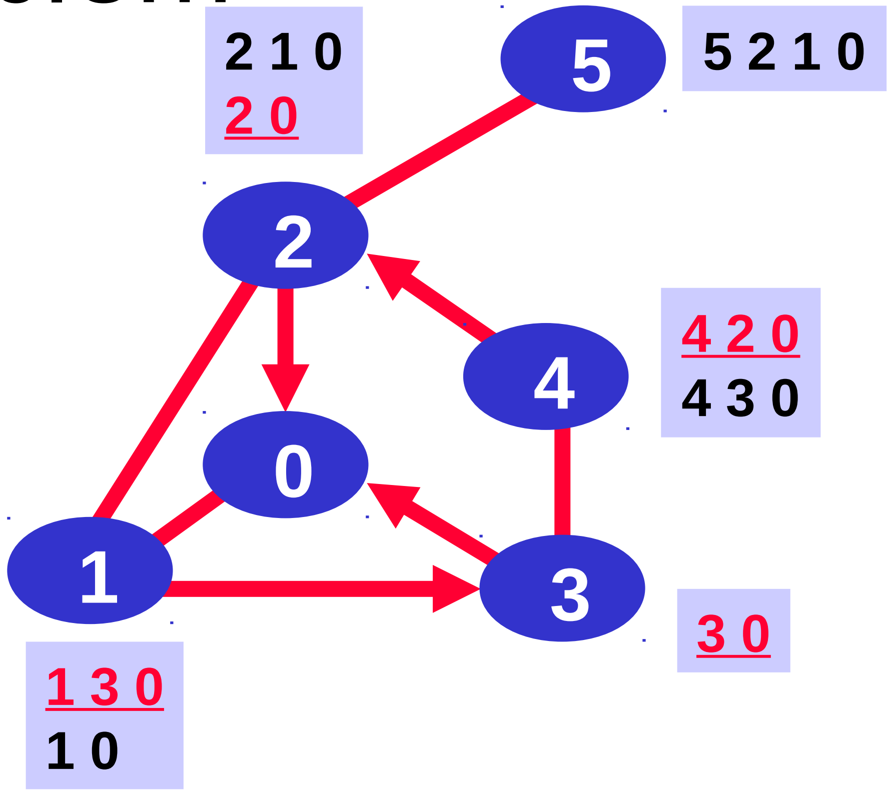
most preferred
…
least preferred (not null)

When modeling BGP : nodes represent BGP speaking routers, and 0 represents a node originating some address block

A Solution to a Stable Paths Problem

A *solution* is an assignment of permitted paths to each node such that

- **node u's assigned path is either the null path or is a path uwP, where wP is assigned to node w and {u,w} is an edge in the graph,**
- **each node is assigned the highest ranked path among those consistent with the paths assigned to its neighbors.**

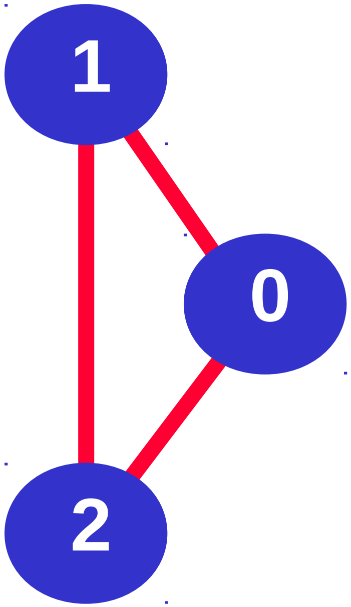


A Solution need not represent a shortest path tree, or a spanning tree.

An SPP may have multiple solutions

1 2 0

1 0



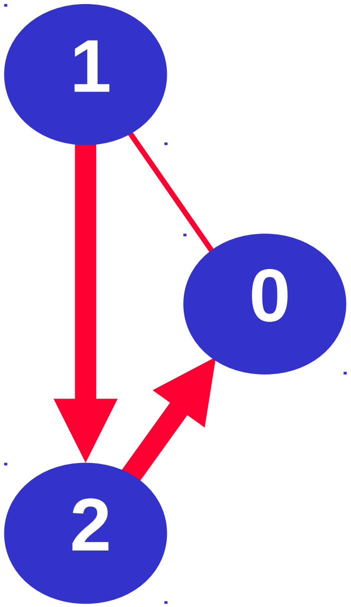
2 1 0

2 0

DISAGREE

1 2 0

1 0



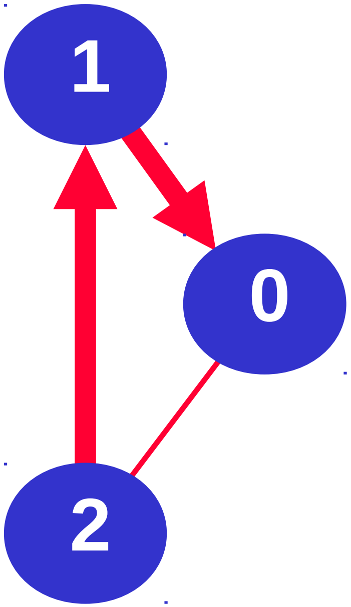
2 1 0

2 0

First solution

1 2 0

1 0

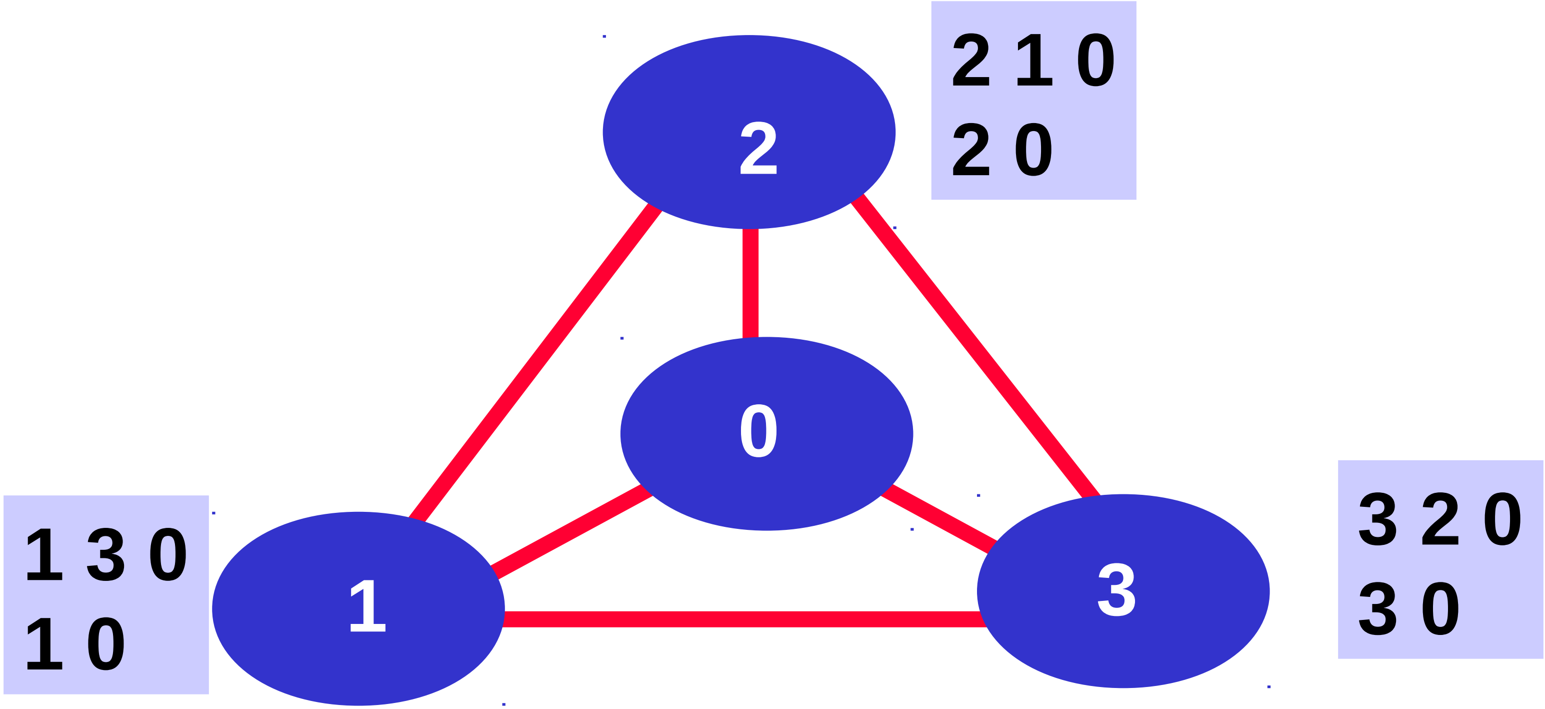


2 1 0

2 0

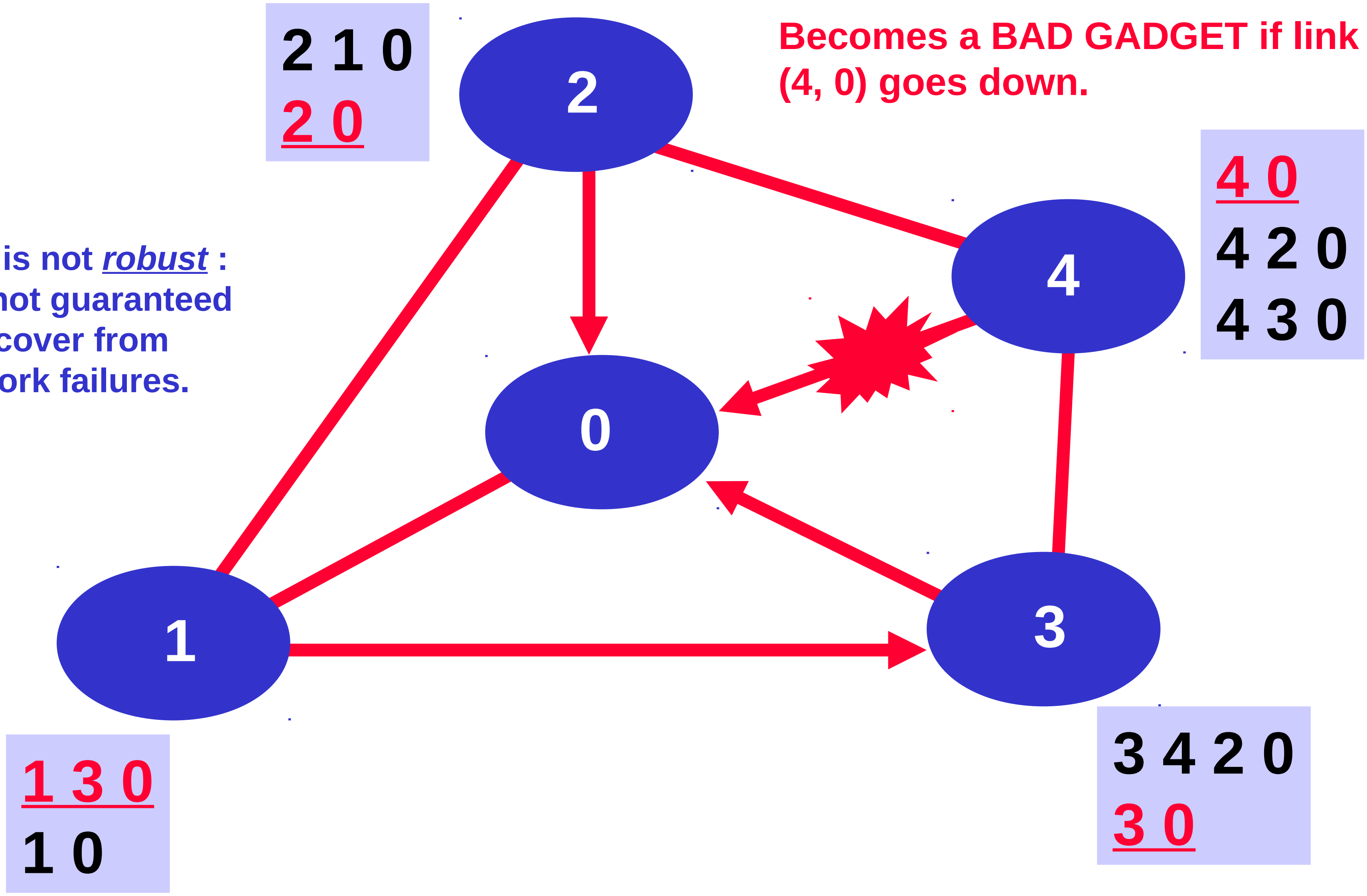
Second solution

Bad Gadget: no solution



This is an SPP version of the example first presented in
Persistent Route Oscillations in Inter-Domain Routing. Kannan Varadhan, Ramesh Govindan, and Deborah Estrin. Computer Networks, Jan. 2000

Beware of Backup Policies



2 1 0
2 0

4 0
4 2 0
4 3 0

1 3 0
1 0

3 4 2 0
3 0

Becomes a BAD GADGET if link (4, 0) goes down.

BGP is not *robust* : it is not guaranteed to recover from network failures.

References

1. Kannan Varadhan, Ramesh Govindan, and Deborah Estrin, “Persistent route oscillations in inter-domain routing”. Computer Networks, Volume 32, Issue 1, January 2000, Pages 1-16.

2. Timothy G. Griffin, F. Bruce Shepherd, and Gordon Wilfong, “The Stable Paths Problem and Interdomain Routing”. IEEE Transactions on Networking. Volume 10, Issue 2 (April 2002). Pages 232-243.

BGP Scalability

IAB Workshop on Inter-Domain routing in October 2006 – RFC 4984:

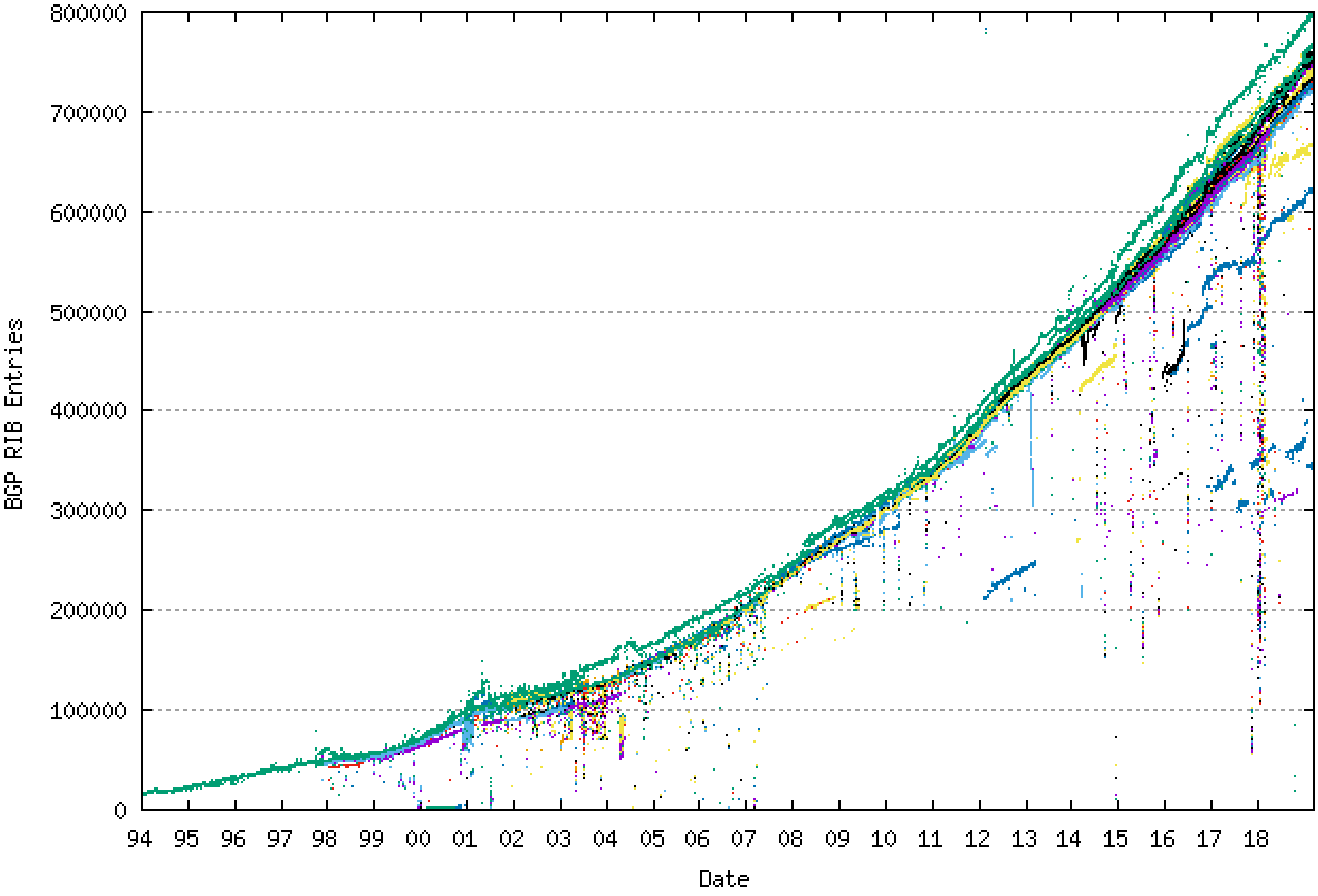
“routing scalability is the most important problem facing the Internet today and must be solved”

BGP measurements

There are a number of ways to "measure" BGP:

1. Assemble a large set of BGP peering sessions and record everything
   - RIPE NCC's RIS service
   - Route Views
2. Perform carefully controlled injections of route information and observe the propagation of information
   - Beacons
   - AS Set manipulation
   - Bogon Detection and Triangulation
3. Take a single BGP perspective and perform continuous recording of a number of BGP metrics over a long baseline -> potaroo.net

BGP Routing Table Size



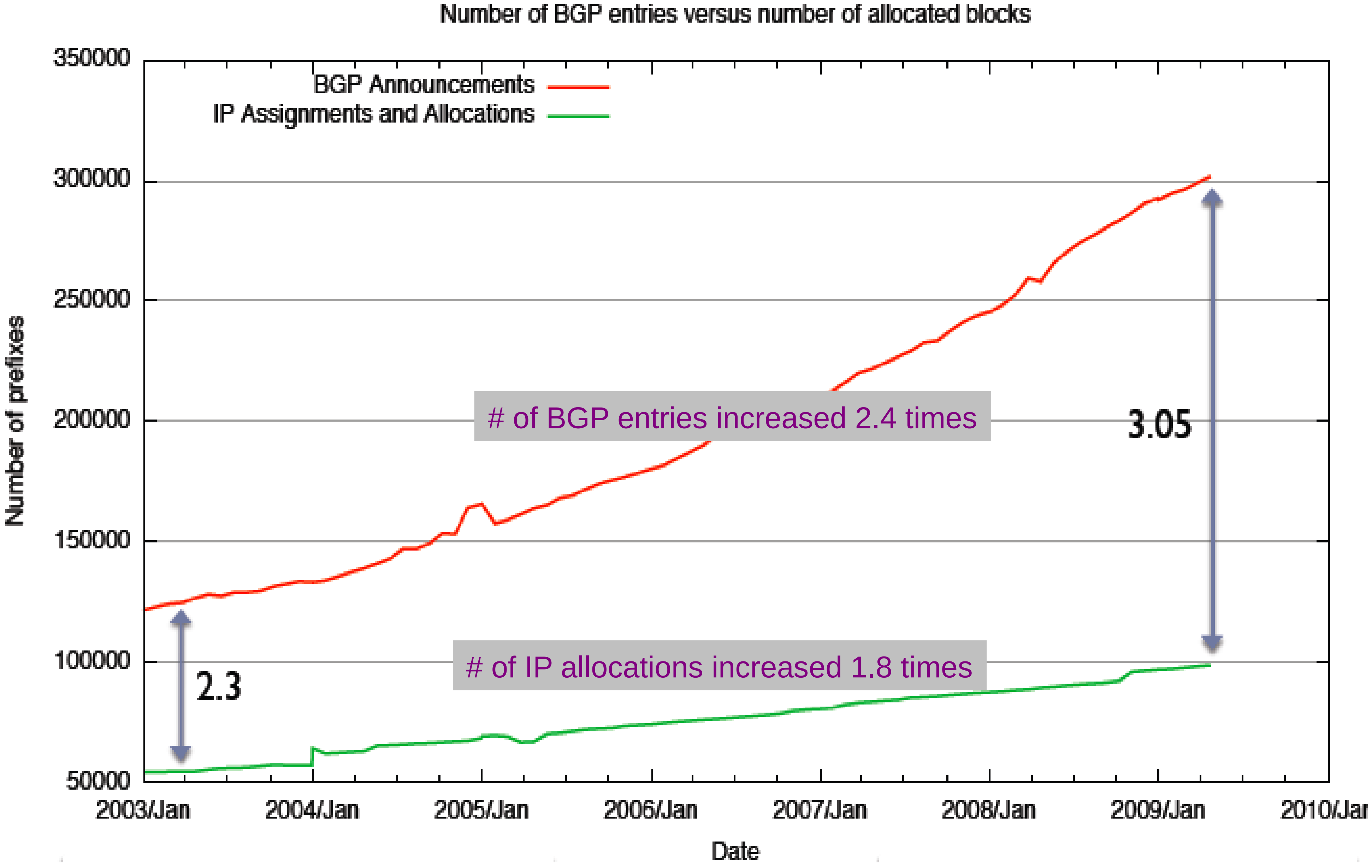
Source: potaroo.net

BGP Scaling and Stability

Is it the size of the RIB or the level of dynamic update and routing stability that is the concern here?

- What is the anticipated end of service life of the core routers?
- What's the price/performance curve for forwarding engine ASICS?
- What's a sustainable growth factor in FIB size that will allow for continued improvement in unit costs of routing?
- What is a reasonable margin of uncertainty in these projections?

BGP announced prefixes versus IP allocated prefixes (BGP vs RIRs)



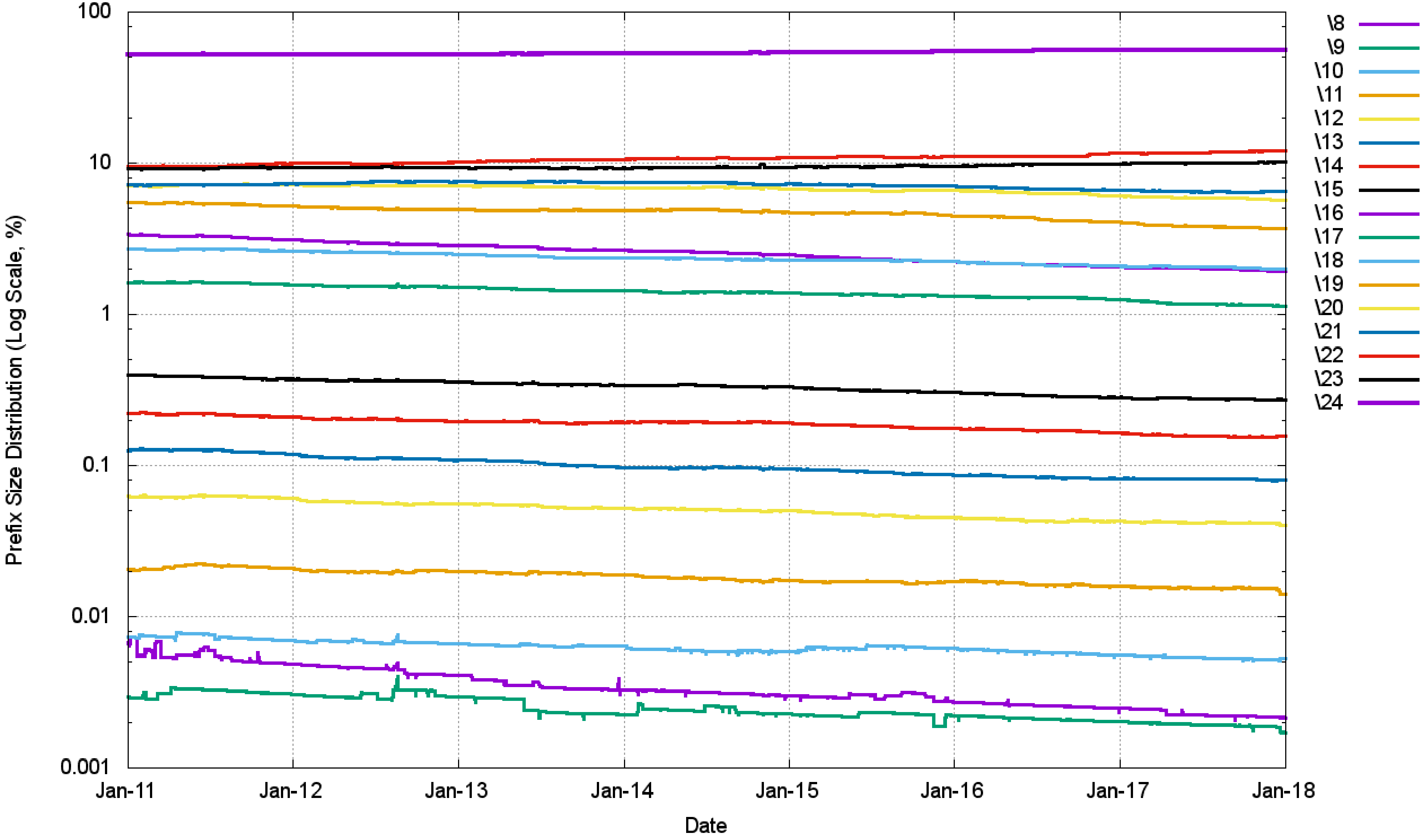
# of BGP entries increased 2.4 times

# of IP allocations increased 1.8 times

BGP growth and IP allocation

- BGP table has more than doubled in 6 years
  - Even though the growth rate is not exponential
- The BGP table growth outstrips IP allocation rate
- Multihoming and traffic engineering techniques introduce redundancy in BGP table (58% in 2009)

More detailed vision: size of BGP announcements



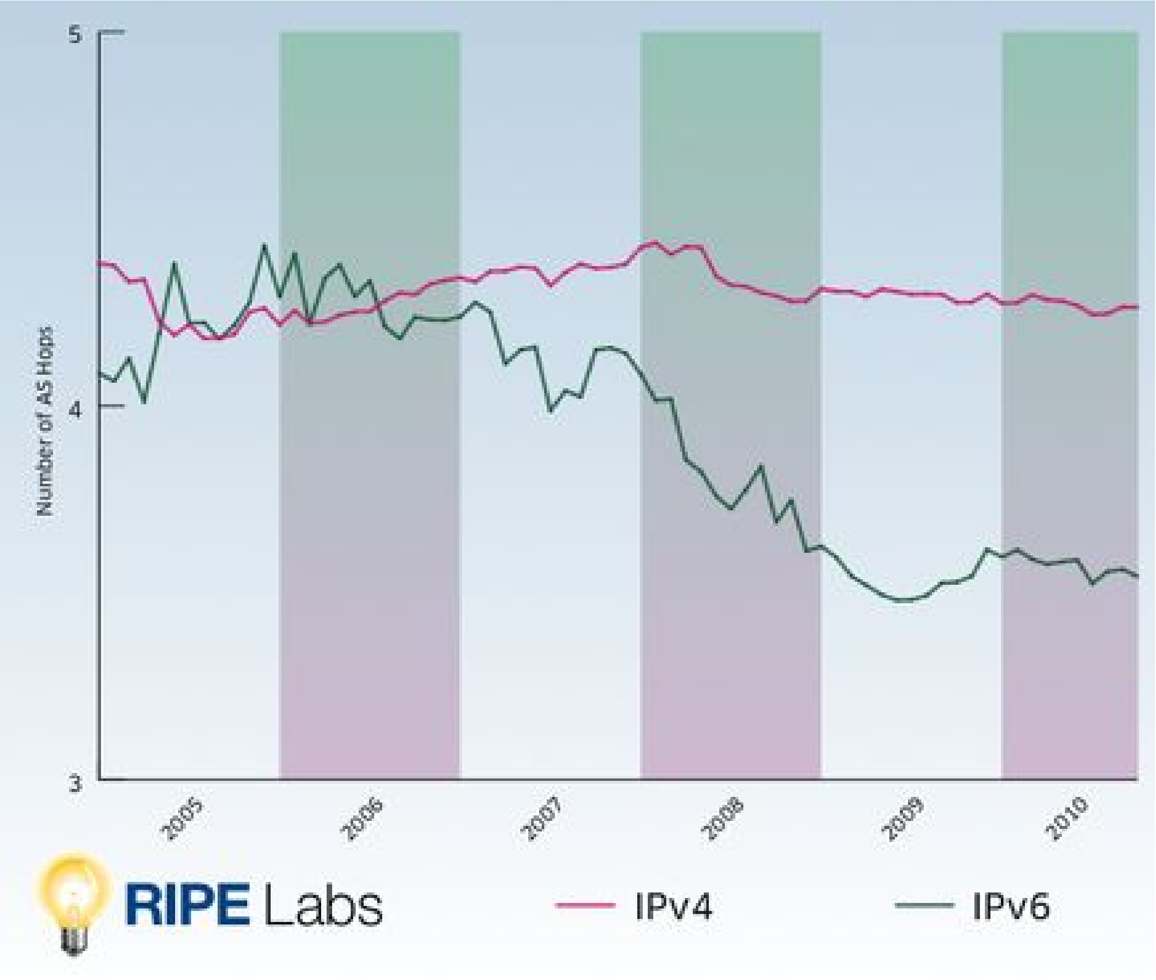
Source: https://blog.apnic.net/2019/01/16/bgp-in-2018-the-bgp-table/

What about BGP churn?

- Growth rates of BGP update activity appear to be far smaller than the growth rate of the routing space itself
  - IF we do not consider duplicates, which account for 40% of BGP updates

- Why are the levels of growth in BGP updates not proportional to the size of the routing table?

Average AS Path Length is long term stable



What is going on?

- The convergence instability factor for a path vector protocol like BGP is related to the AS path length, and average AS Path length has remained steady in the Internet for some years → the growth of the network appears to have been achieved by increasing the density of connectivity, rather than increasing the network's diameter

- Today's Internet of 30,000 ASes is more densely interconnected, but not more "stringier" than the internet of 5,000 ASes of 2,000 (2009)

- This is consistent with the observation that the number of protocol path exploration transitions leading to convergence to a new stable state is relatively stable over time

- Beware!
  - *maybe densification limits the visibility of routing changes (?)*

References

1. *Analyzing the Internet BGP Routing Table*, Geoff Huston, The Internet Protocol Journal - Volume 4, Number 1, March 2001.

2. *BGP in 2009*, Geoff Huston, RIPE 60, Prague, May 2010.

3. *BGP Routing Table: Trends and Challenges*, Alexander Afanasyev, Neil Tilley, Brent Longsta, and Lixia Zhang, 2010.

4. *BGP Churn Evolution: A perspective from the core*, Ahmed Elmokashfi, Amund Kvalbein, Constantine Dovrolis, INFOCOM 2010.

Recap

- Relationships Among Networks and Interdomain Routing
  - Gao-Rexford
- Implementing Inter-Network Relationships with BGP
  - Policy-based routing (which metric is minimized?)
    - Not Link-State nor Distance-Vector
    - Sessions vs. Flooding
  - Many attributes
    - AS_PATH: loop free
- BGP: a bit of theory
  - BGP is not safe: convergence is not guaranteed after a failure
- BGP Scalability
  - RIB size -> TE practices, deaggregation
  - Churn: mostly “quiet”, if we do not count duplicates (why duplicates, anyway?)

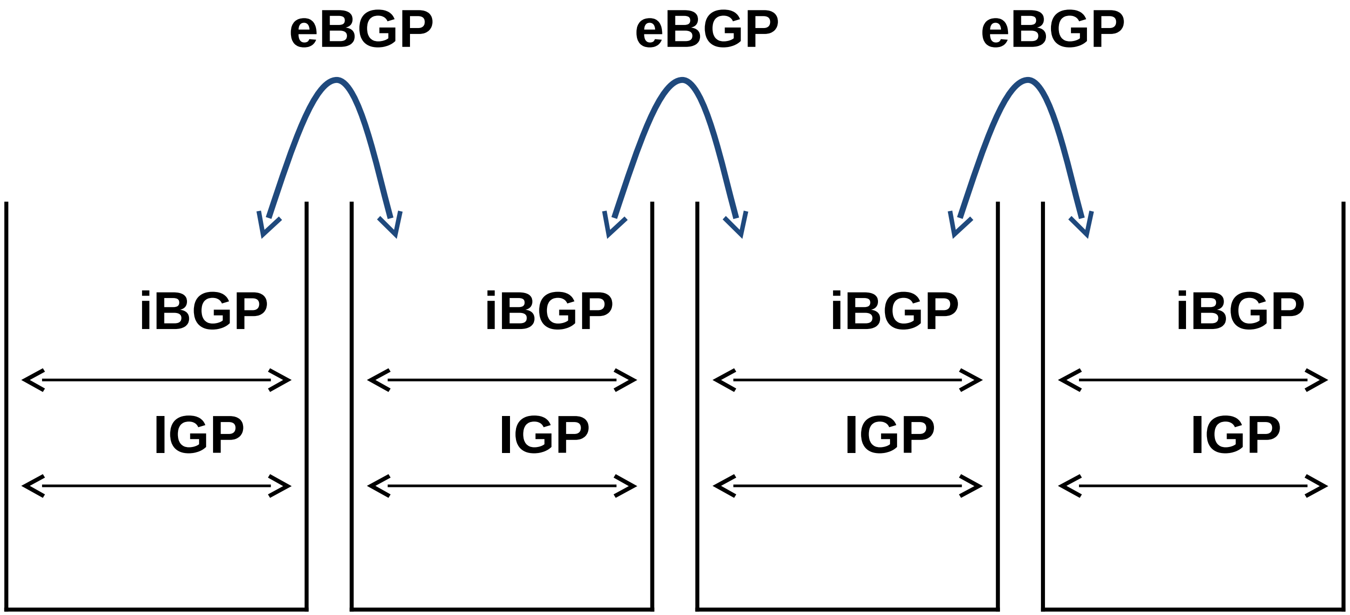
iBGP Scalability:
Route Reflectors topologies

Outline

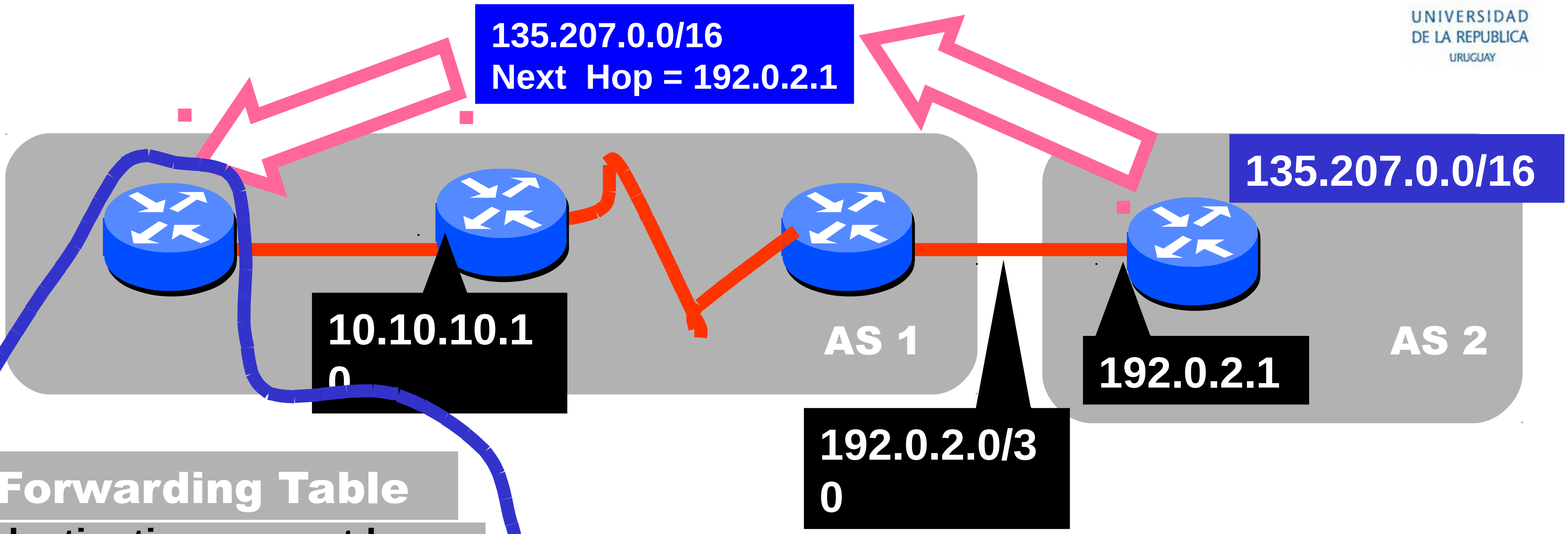
- (Inter + Intra)domain Routing
- Route Reflectors
  - Multiple, hierarchical RRs
  - Known issues
- iBGP Route Reflectors topologies
  - Practical design guidelines
  - Correct and scalable proposals
  - Others
    - IETF
    - Centralized solutions

BGP-IGP Interaction

- ASes exchange reachability information using (external) BGP
- Intradomain routing: IGP
- Propagation of BGP information intradomain: (internal) BGP



Remember: NEXT_HOP



Forwarding Table

destination	next hop
192.0.2.0/30	10.10.10.10

+

EGP

destination	next hop
135.207.0.0/16	192.0.2.1

Forwarding Table

destination	next hop
135.207.0.0/16	10.10.10.10
192.0.2.0/30	10.10.10.10

BGP-IGP interaction alternatives

- Propagation of BGP Information via the IGP, multicast or other efficient flooding mechanism
  - Mentioned in rfcs 1772 & 1773, implementation?
    - BGP Scalable Transport and other alternatives
- Redistribution/tagged IGP
  - Specified in rfcs 1403 & 1745 (moved to historical status)
- Encapsulation
  - MPLS tunnels among eBGP speakers
- Pervasive BGP

Pervasive Internal BGP (iBGP)

- All routers in an AS are iBGP speakers
- IGP is only used for routing within the AS
  - No BGP routes are imported into the IGP
- Routing table recursive lookup
  - First lookup determine BGP next hop (exit router)
  - Second lookup determine the IGP path to the exit router
- Need to make sure that
  - Internal transport of BGP info is loop-free (just BGP info!)
  - Internal routing is coherent (now, loop-freeness for data plane forwarding)

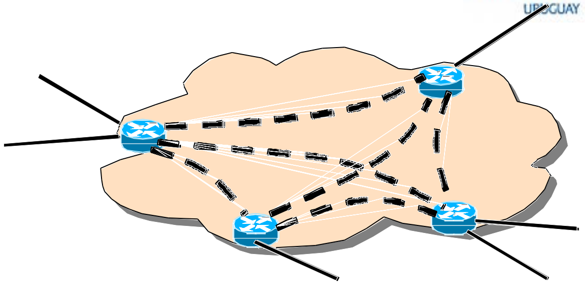
Internal BGP

- iBGP and eBGP are same protocol in that
  - same message types used
  - same attributes used
  - same state machine
  - BUT use different rules for readvertising prefixes

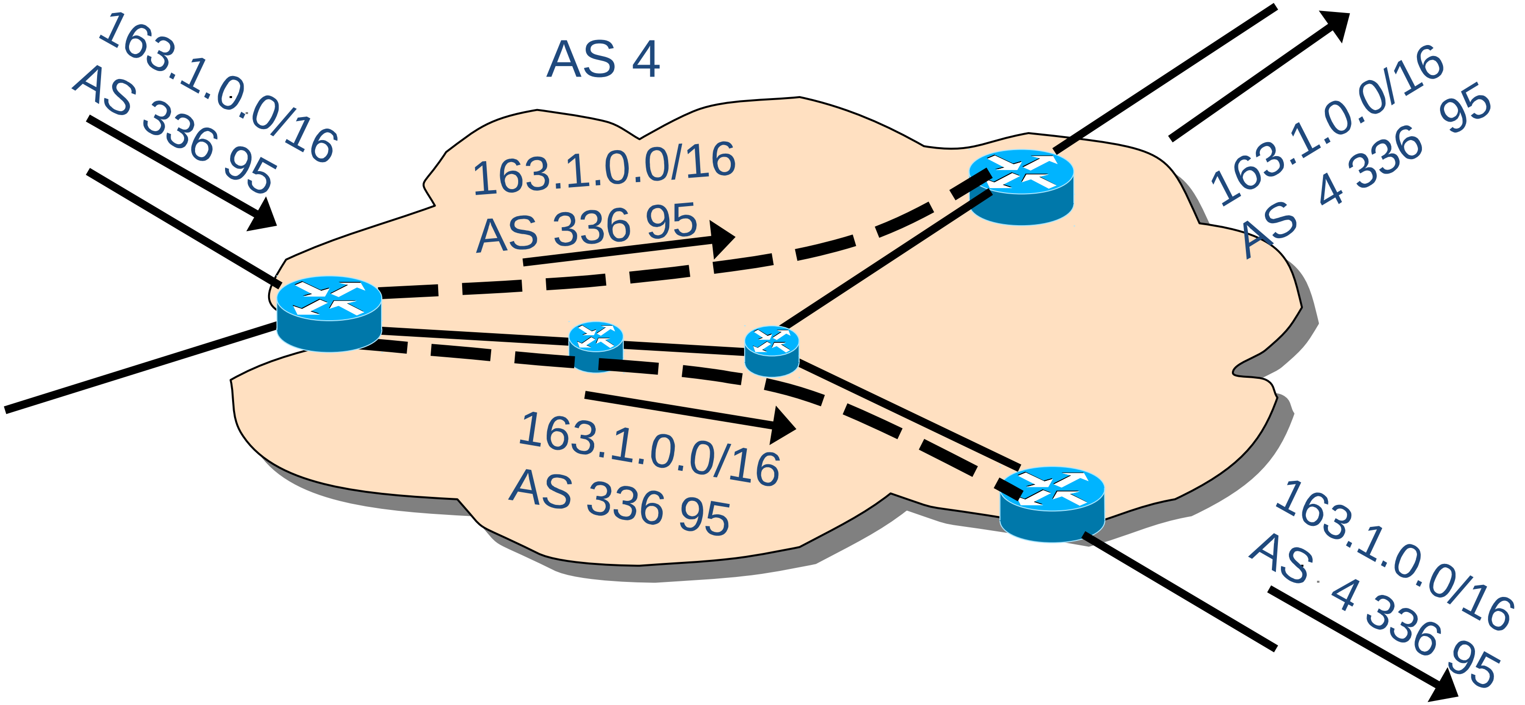
- Rules for iBGP
  - #1: prefixes learned from an eBGP neighbor *can* be readvertised to an iBGP neighbor, and vice versa
  - #2: prefixes learned from an iBGP neighbor *cannot* be readvertised to another iBGP neighbor

Loop-freeness of BGP info in iBGP

- Why rule #2? To prevent *BGP announcements* from looping
  - eBGP detect loops via AS-PATH
  - AS-PATH not changed in iBGP
- Implication of rule: a full mesh of iBGP sessions between each pair of routers in an AS is required



- Example of rule #1:



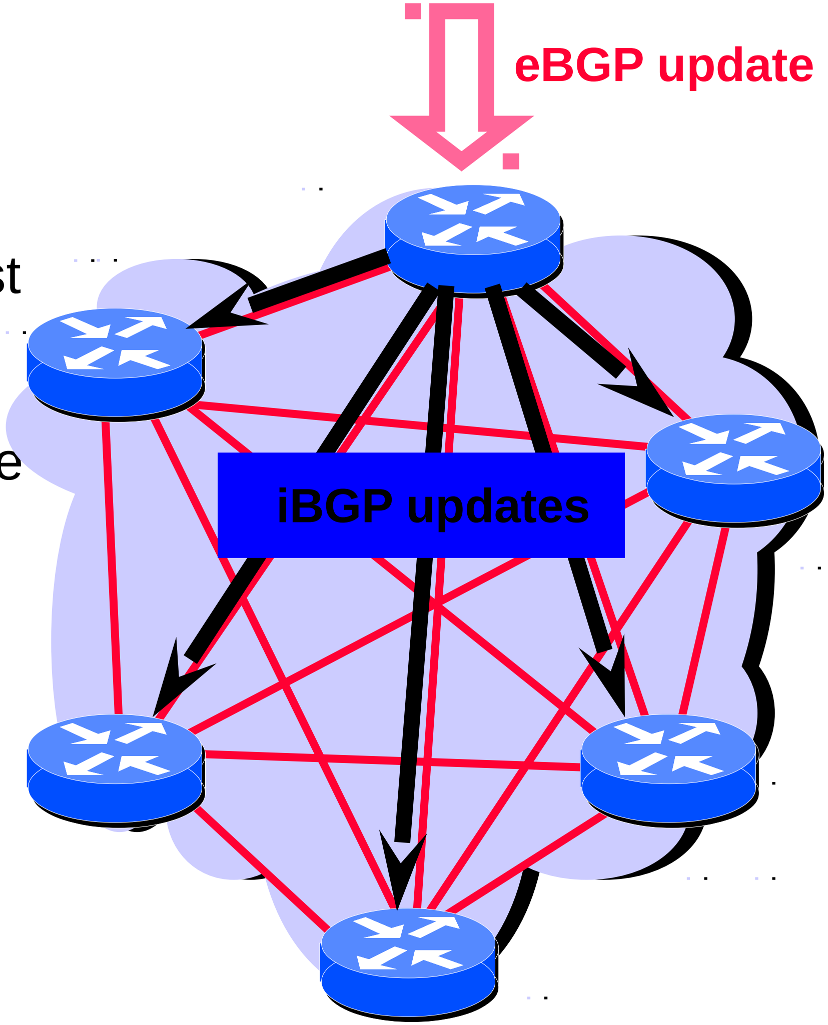
iBGP full-mesh scalability

- n*(n - 1)/2 iBGP sessions

- Configuration management
  - Each router must have n-1 iBGP sessions configured
  - The addition of a single iBGP speaker requires configuration changes to all other iBGP speakers
  - E.g. if we have 200 routers in our network that would give us 19900 BGP sessions!

iBGP full-mesh scalability

- Routing state
  - Many Adj-RIBs : most routes are not used
  - Size of iBGP routing table can be order n larger than number of best routes (remember alternate routes!)
  - Each router has to listen to update noise from each neighbor
  - CPU and memory resources -> large routers needed!
- Solutions
  - Route Reflectors
  - Confederations

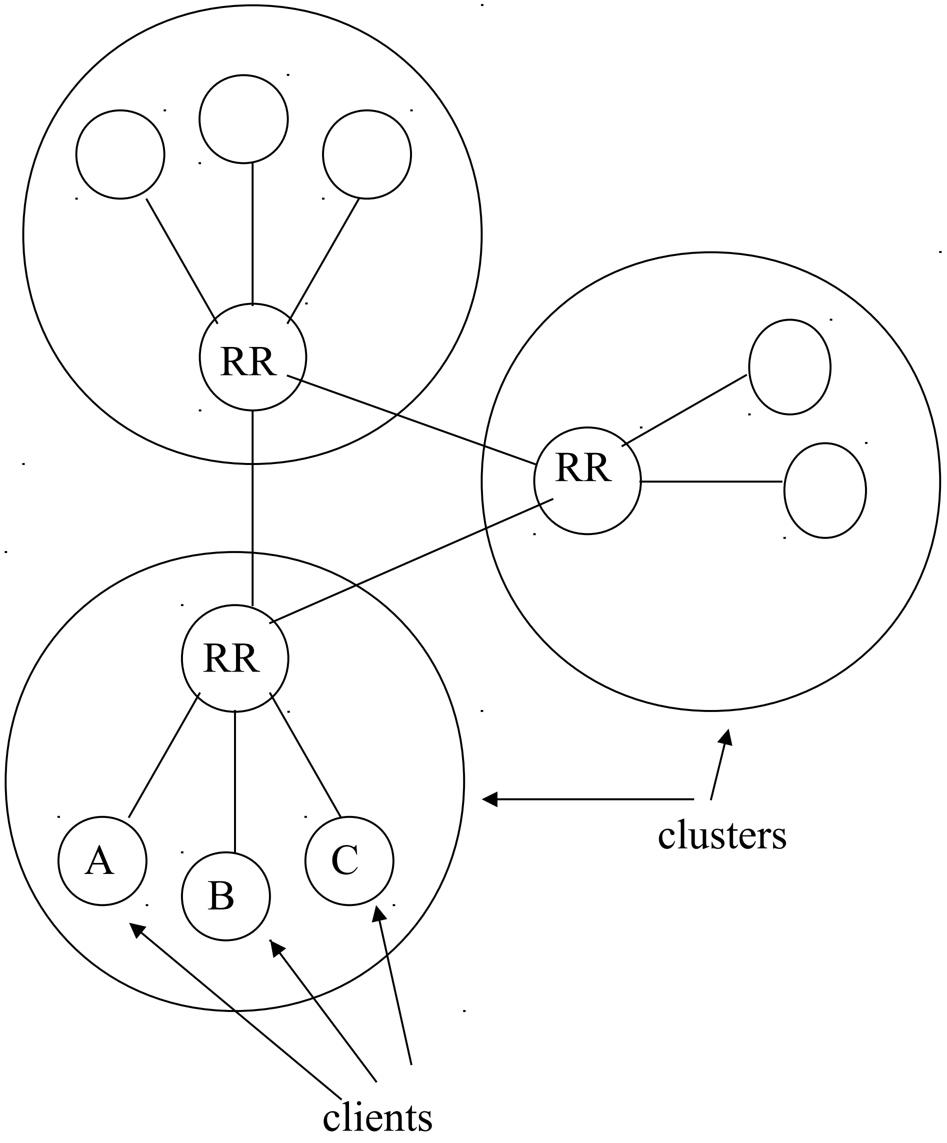


Route Reflectors

Route Reflectors

- Avoiding the virtual full mesh of iBGP sessions:
- group routers into clusters
- Assign a leader to each cluster, called a route reflector (RR)
- Members of a cluster are called clients of the RR

- The clients do not know they are clients and are configured as normal iBGP peers
- Only the best route to a destination is sent from a RR to a client



Route Reflectors: announcements

- If received from RR, reflect to clients
- If received from a client, reflect to RRs and clients
- If received from eBGP, reflect to all: RRs and clients
- RRs reflect only the best route to a given prefix, not all announcements they receive
  - helps size of routing table
  - sometimes clients don’t need to carry full table
- RR should not change the attributes
  - NEXT_HOP
  - AS_PATH
  - LOCAL_PREF
  - MED

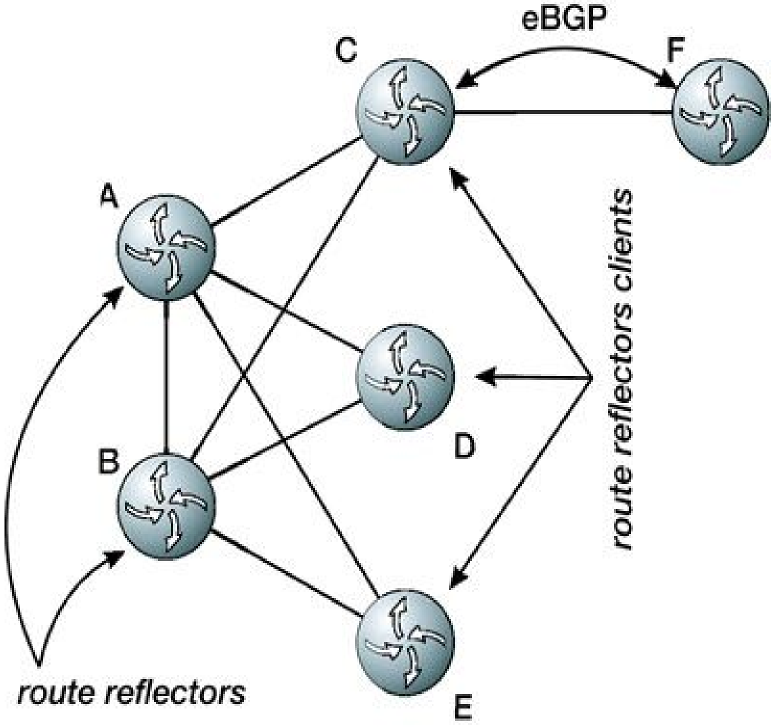
Avoiding Loops with Route Reflectors

- Loops cannot be detected by traditional approach using AS_PATH because AS_PATH not modified within an AS
- Announcements could leave a cluster and re-enter it
- Two new attributes added by RR *if a route is reflected*
  - ORIGINATOR_ID: Router ID of route’s originator in AS
    rule: announcement discarded if returns to originator
  - CLUSTER_LIST: a sequence of Cluster Ids, set by RRs
    rule: if an RR receives an update and the cluster list contains its Cluster ID, then update is discarded
- Both are optional, nontransitive (dont propagate to eBGP)

Multiple route reflectors

- For redundancy, is possible to have more than one route reflector in a cluster
  - Otherwise, the RR is a 'single-point-of-failure'

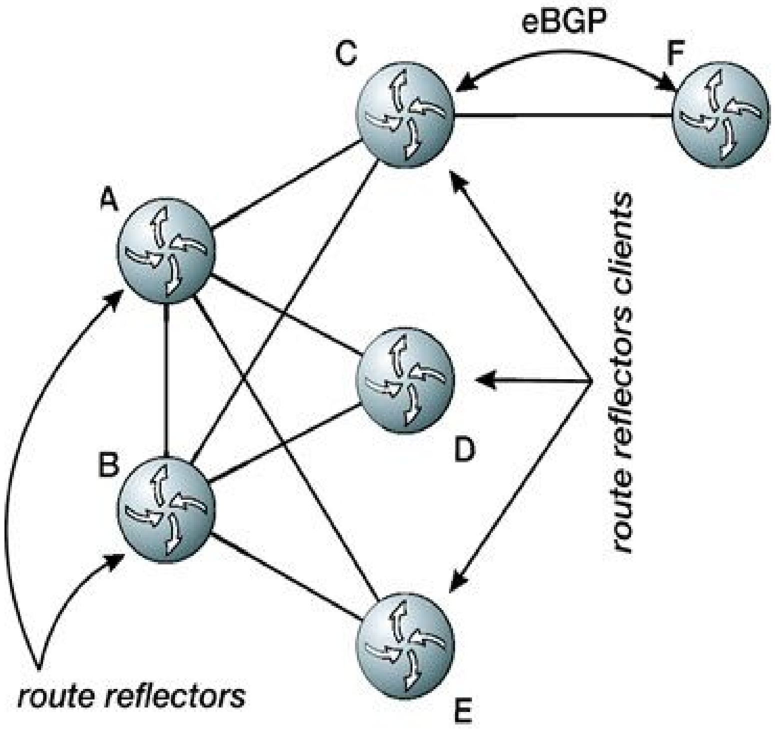
- RRs in a cluster may have the same cluster ID
  - …or different cluster IDs
  - *Let's discuss alternatives over this sample topology*



Source: Practical BGP

Multiple route reflectors

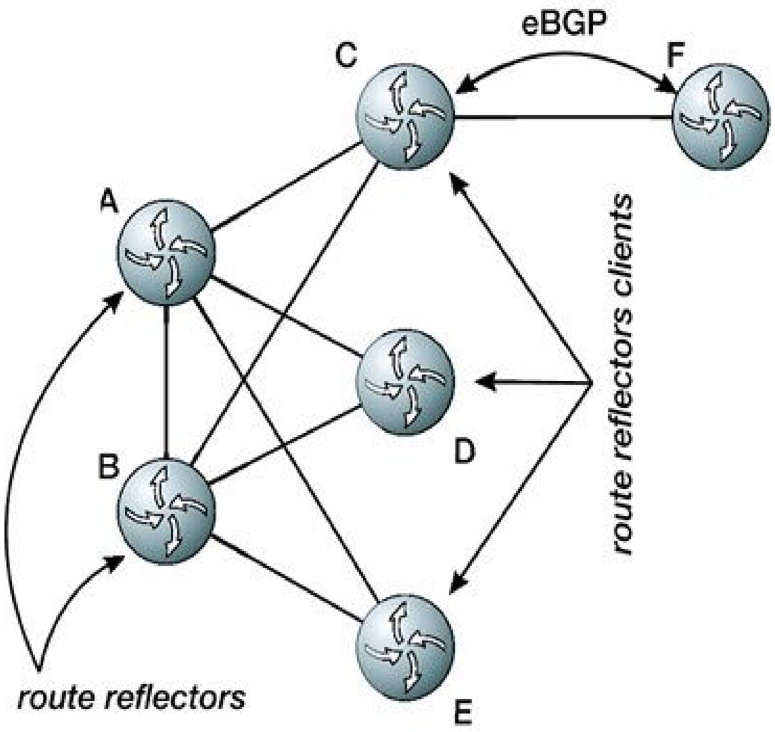
- Different Cluster IDs:
  - C receives some route *p* from F. It advertises *p* to A and B
  - A creates a new Cluster List attribute and inserts its router ID, and sets the originator ID to C. A advertises *p* to B, D, and E
  - B receives this update, and discovers that its cluster ID is not in the Cluster List. It accepts the advertisement and prepends its cluster ID to the Cluster List
  - B also receives the update from C, creates a new Cluster List attribute, and inserts its cluster ID. Also sets the originator ID to C. B then runs the best path algorithm and selects the direct path via C and advertises this update to D, E and A
  - Since A and B are using different Cluster IDs, D and E each receive two copies of the update



Source: Practical BGP

Multiple route reflectors

- Same cluster IDs:
  - C receives some route *p* from F. It advertises *p* to A and B
  - A creates a new Cluster List attribute, inserts its cluster ID (the same as B), and sets the originator ID to C. Router A then advertises *p* to B, D, and E.
  - B receives this update and discards it, because its locally defined cluster ID is already in the Cluster List
  - B also receives the update from C, adds a Cluster List attribute, inserts its cluster ID, and sets the originator ID to C. B advertises this update to D, E and A
  - When A receives the update from B, it discards it because the cluster list contains the locally configured cluster ID
  - Now routers A and B are only required to store one copy of the route

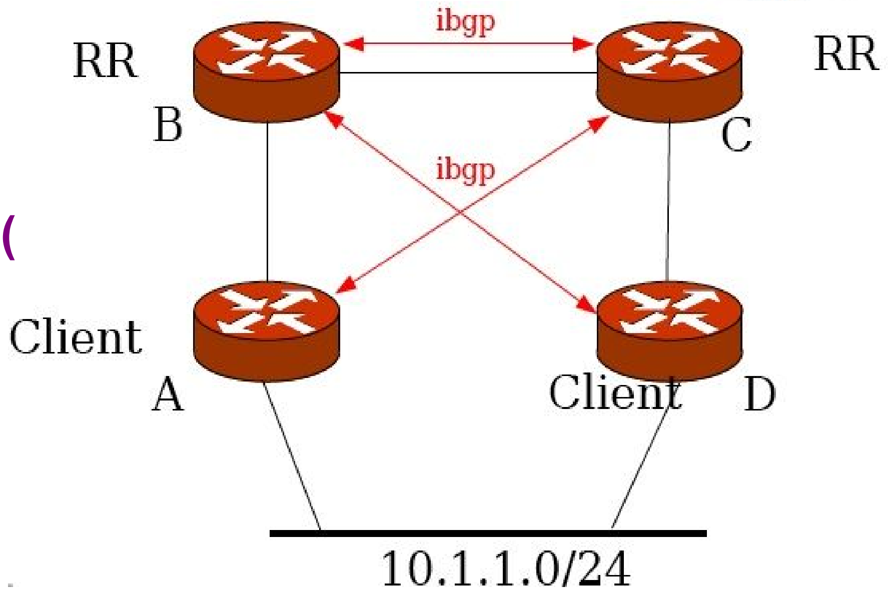


Source: Practical BGP

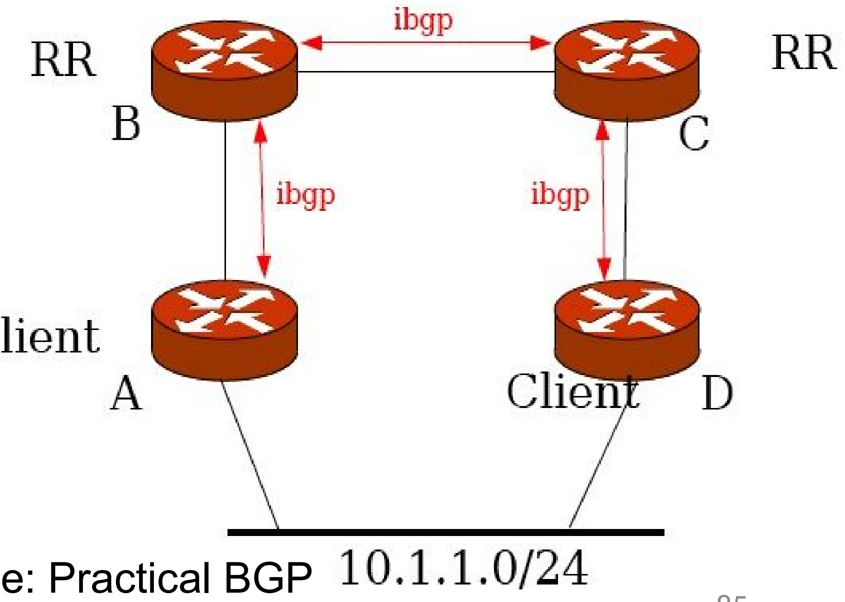
Multiple route reflectors

- A and D receive an advertisement for 10.1.1.0/24 from an external BGP peer
- B will prefer D and C will prefer A => routing loop!

:-(



:-)

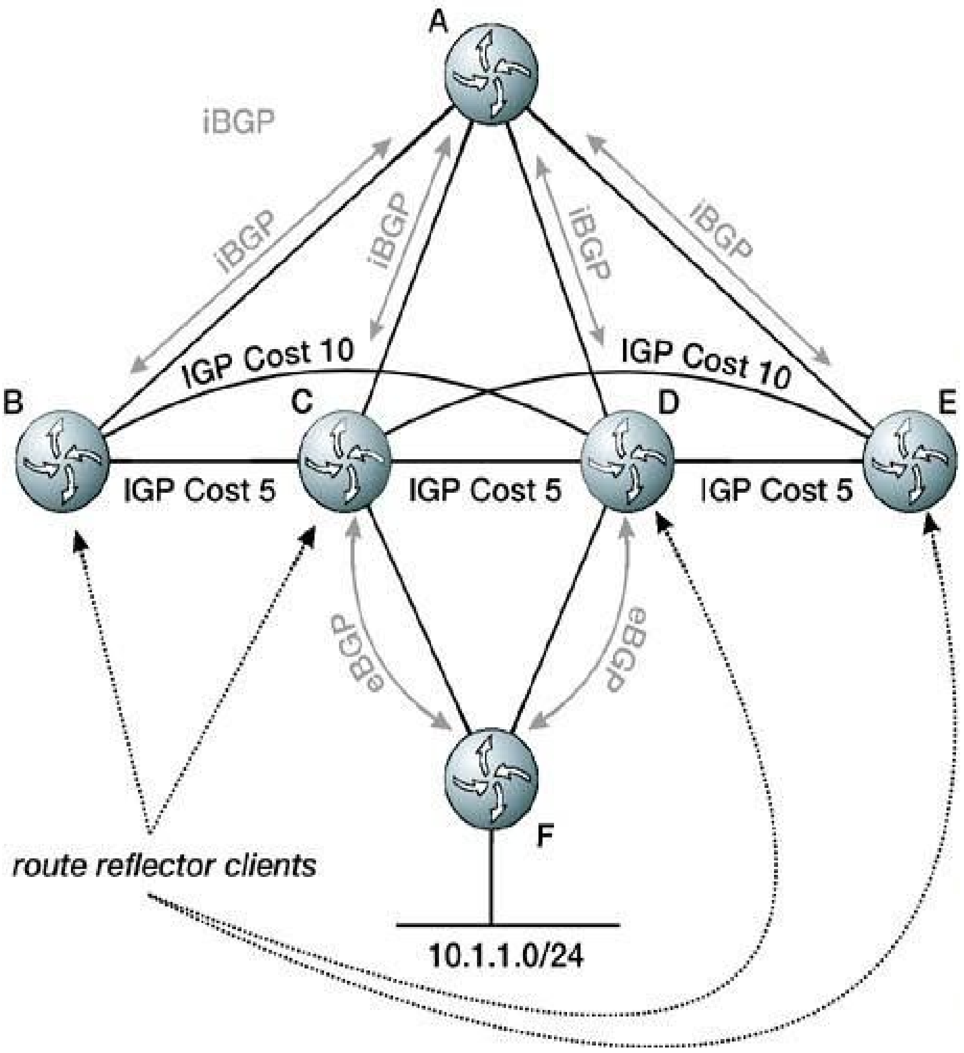


Route reflectors:
follow the physical topology

Source: Practical BGP

Route reflectors: suboptimal route selection

- First design: full mesh iBGP
  - Next hop selection follow IGP
- Using A as RR
  - Supose A chooses C
  - What happens?

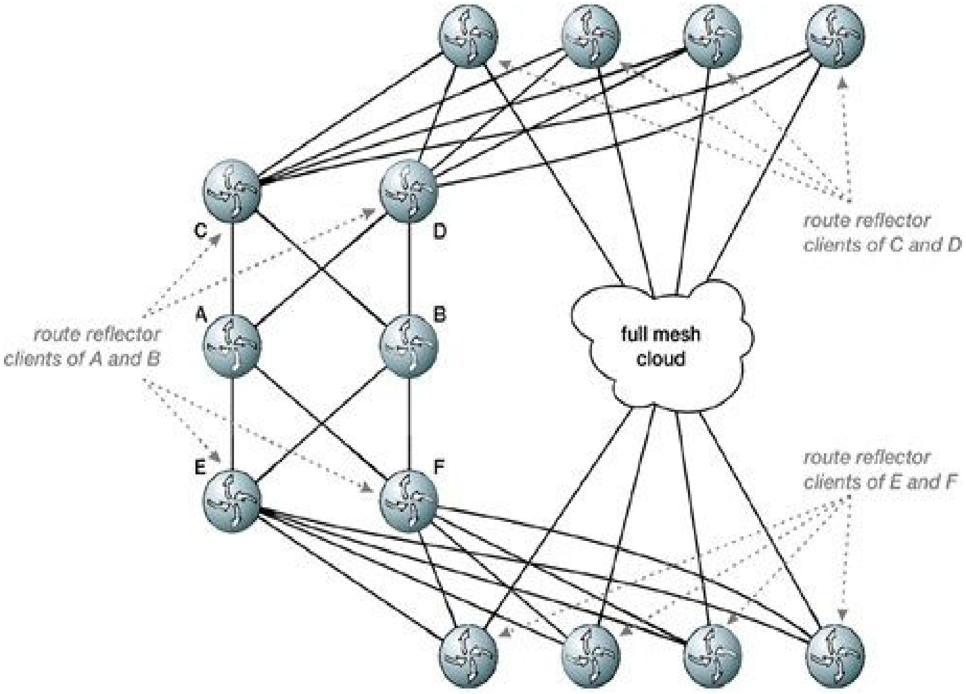


Obs: ignore LOCAL_PREF and MED

Source: Practical BGP

Hierarchical Route Reflectors

- Example with three route reflection clusters
  - Root cluster: A and B are route reflectors, while D, C, E, and F are clients.
  - Two lower-level clusters: C and D, E and F are route reflectors, respectively.
- Remember: RR topology should follow the physical topology of the interconnected iBGP speakers.
  - Avoiding forwarding loops


route reflector clients of C and D
C
D
A
B
route reflector clients of A and B
full mesh cloud
E
F
route reflector clients of E and F

Source: Practical BGP

MED oscillation

- RFC 3345
  - Bad news:
    - “In certain topologies involving either route reflectors or confederations, the partial visibility of the available exit points into a neighboring AS may result in an inconsistent best path selection decision as the routers don't have all the relevant information. If the inconsistencies span more than one peering router, they may result in a persistent route oscillation”
  - (Relative) good news
    - “The persistent route oscillation behavior is deterministic and can be avoided by employing some rudimentary BGP network design principles until protocol enhancements resolve the problem”

MED oscillation

- RR/confederation hides some information
  - RR/confederation sends best path only
  - not all routers know all best paths
- MED (Multi Exit Discriminator) vs IGP cost to the neighbor…
- Seen on a RR/confederation border:

```
 #show ip bgp 10.0.0.0 | include best #
Paths: (3 available, best #3)
 #show ip bgp 10.0.0.0 | include best #
Paths: (3 available, best #2)
 #show ip bgp 10.0.0.0 | include best #
Paths: (3 available, best #3)
 ...
```

MED oscillation: workarounds

- Use full mesh iBGP
  - Scalability…
- Do not listen to the MED (or only with stub-AS)
  - `set metric 0` on all prefixes
- `bgp always-compare-med`
  - Inconsistent! -> remember MEDs is per-AS attribute
- Use local-pref to force decision
  - exit no longer chosen by peer = more work :(
- Allow peer to set local-pref using community

- *Protocol improvement?*

iBGP Scalability: summary

- iBGP is necessary for core routers in a transit network
- BGP loop detection mechanism is based on AS_PATH-> IBGP peering must be fully meshed
- This leads to scaling problems
- Solutions:
  - Route reflectors
  - AS confederations

iBGP Scalability: summary

- Known issues
  - Routing loops
  - Divergence and/or non-deterministic update process
    - MED oscilation
  - Suboptimal intra-domain routing
  - Route deflection
    - Re-route at exit point!

References

1. *Practical BGP*, by Russ White, Danny McPherson, Sangli Srihari
2. *IBGP scaling: Route reflectors and confederations*, Olof Hagsand, Lecture at KTH /CSC
3. *BGP Oscillation*, Fabien Berger, Presentation at Swiss Network Operators Group - SwiNOG 3, 19 September 2001
4. RFC 1772, “Application of the Border Gateway Protocol in the Internet”, March 1995.
5. T. Bates, E. Chen, R. Chandra, *BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)*. IETF RFC 4456 (Obsoletes: 2796, 1966), April 2006.
6. P. Traina, D. McPherson, J. Scudder, *Autonomous System Confederations for BGP*. IETF RRFC 5065 (Obsoletes: 3065), August 2007.
7. D. McPherson, V. Gill, D. Walton, A. Retana, *Border Gateway Protocol (BGP) Persistent Route Oscillation Condition*. IETF RFC 3345, August 2002.

iBGP Route Reflectors topologies

Practical design guidelines

- Heuristics
  - Bates
  - Zhang
- Some IETF proposals
  - Best-external
  - Add-Path
  - N-plane Route Reflectors
  - BGP Optimal Route Reflection [BGP-ORR]

iBGP Correctness

- Signalling and forwarding correctnes
  - Path symmetry: in eBGP signalling and forwarding traffic flow along the same path (usually peering over directly connected link)
  - iBGP is routed, therefore path symmetry is not guaranteed
  - iBGP configuration correctness: stable, anomaly free routing (in particular, loop-free)
  - Checking the correctness of an iBGP graph is NP-complete
  - Two conditions ensure a correct (loop-free) iBGP graph:
    - 1) route-reflectors should prefer client routes to non-client routes
    - 2) every shortest path should be a valid signaling path

- Dissemination correctness
  - Every router should learn at least one route to each destination

Correct and scalable proposals

- BGPSep
  - And variants BGPSep_S, BGPSep_D
- Optimal iBGP topologies

  Fm-optimality
- Centralized solutions
  - Route Control Platfrom, among others
- Non-BGP solutions and/or with modifications to BGP behaviour/implementation
- An Integer Programming formulation: Optimal Route Reflector Topology Design (ORRTD) [MGR2018]

References

1. [RFC 4456] T. Bates, E. Chen, R. Chandra, *BGP route reflection - an alternative to full mesh internal BGP (IBGP)*, RFC 4456 (April 2006).
2. [PelsserCN2010] C. Pelsser, B. Quoitin, S. Uhlig, T. Takeda, and K. Shiomoto. *Providing scalable NH-diverse iBGP route redistribution to achieve sub-second switch-over time*. To appear in Computer Networks journal, 2010.
3. [BGPDesign] R. Zhang, M. Bartell, *BGP Design and Implementation*, 1st Edition, Cisco Press, 2003.
4. [GW02] Griffin, T.G.,Wilfong, G.: *On the correctness of iBGP configuration*. In: Proc. of ACM SIGCOMM (August 2002).
5. [Vutukuru2006how] *How to Construct a Correct and Scalable iBGP Configuration*, Mythili Vutukuru, Paul Valiant, Swastik Kopparty, and Hari Balakrishnan. IEEE INFOCOM 2006.
6. [BuobUM2008] M.-O. Buob, S. Uhlig, M. Meulle, *Designing optimal iBGP Route-Reflection topologies*. IFIP Networking 2008.

References

7. [BGP-ORR] R. Raszuk, C. Cassar, E. Aman, B. Decraene, K. Wang, BGP Optimal Route Reflection (BGP-ORR). Internet Draft <draft-ietf-idr-bgp-optimal-route-reflection-17>. Expires: April 13, 2019
8. [BST03] Kedar Poduri, Cengiz Alaettinoglu, Van Jacobson, *BST - BGP Scalable Transport*. NANOG 27, 2003
9. [LOUP] Nikola Gvozdiev, Brad Karp, Mark Handley, LOUP: the principles and practice of intra-domain route dissemination, In Proceedings, Proceedings of the 10th USENIX conference on Networked Systems Design and Implementation, nsdi'13, Pages 413-426.
10. [RCP] Nick Feamster, Hari Balakrishnan, Jennifer Rexford, Aman Shaikh, Jacobus van der Merwe, *The case for separating routing from routers*. In Proceedings FDNA '04, ACM SIGCOMM workshop on Future directions in network architecture
11. [MGR2018] Cristina Mayr, Eduardo Grampín, Claudio Risso, Optimal Route Reflection Topology Design. In Proceedings, 10th Latin America Networking Conference, LANC '18. Pages 65-72.