ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

External Routing
BGP

Jean-Yves Le Boudec

Contents

1. Inter-Domain Routing
2. Policy Routing
3. Route Aggregation
4. BGP protocol and implementation
5. Interaction BGP—IGP—Packet Forwarding
6. A CISCO view
7. Other Bells and Whistles
8. Examples
9. Illustrations and Statistics

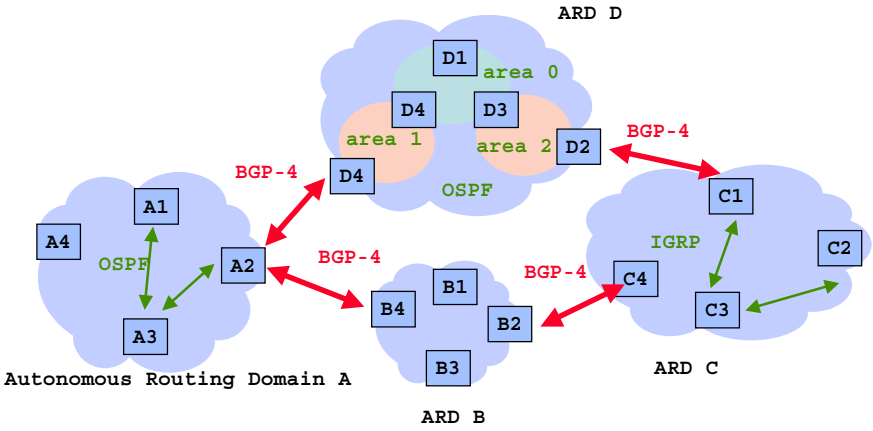
1. Routing in the Internet

- The Internet is too large to be run by one routing protocol
- Hierarchical routing is used
  - the Internet is split into Domains, or Autonomous Systems
  - with OSPF: large domains are split into Areas
- Routing protocols are said
  - **interior**: (Internal Gateway Protocols, IGPs): inside ASs: RIP, OSPF (standard), IGRP (Cisco)
  - **exterior**: between ASs: EGP (old) and BGP-1 to BGP-4 (today), IDRP (tomorrow)

Autonomous Routing Domains Autonomous Systems (ASs)

- ARD = routing domain under one single administration
  - one or more border routers
  - all subnetworks inside an ARD should be connected
  - should learn about other subnetworks - the routing tables of internal routers should contain entries of all destination of the Internet
- AS are ARD with a number ("AS number")
  - 16 bits
  - public: 1 - 64511
  - private: 64512 – 65535
- ARDs with default route to the rest of the world do not need a number
- Examples
  - AS1942 - CICG-GRENOBLE, AS1717, AS2200 - Renater
  - AS559 - SWITCH Teleinformatics Services
  - AS5511 – OPENTRANSIT
  - EPFL: one ARD, no number

- the figure shows three domains, or ARDs.
- ARDs can be transit (B and D), stub (A) or multihomed (C). Only non stub domains need an AS number, as we can see on the BGP slides later on.



Hierarchical Routing

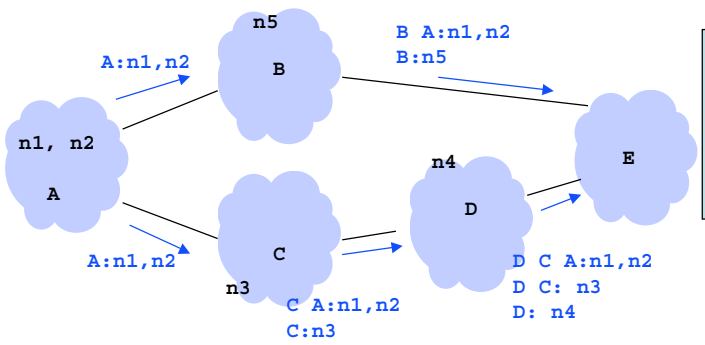
- Hierarchical routing is different case by case, however, we can distinguish three elements
  - 1. **routing method** used in the higher level
  - 2. **mapping** higher level nodes to lower level nodes
  - 3. **inter-level** routing information
- We know two examples
  - hierarchical routing with OSPF (inside a large domain)
    - Centrally Organized
  - inter-domain routing with BGP-4
    - Self-Organized

Inter-Domain Routing

- Inter domain routing hierarchies
  - BGP-4: one level of hierarchy (one ARD is a virtual node in BGP)
    The ARD interconnection layer is **self-organized**
  - IDRP: several levels of hierarchy (ARDs can be aggregated)
- The principles of BGP-4 :
  - 1. routing method used in the higher level:
    - *path vector*
    - with *policy* routing
  - 2. mapping higher level nodes to lower level nodes
    - border gateways (= BGP speakers)
  - 3. inter-level routing information
    - summary link state records are injected into the interior routing protocol (OSPF, RIP, etc)

Path Vector Routing

- a message between neighbours is a set of: (path, dest) (called "routes")
- every node (here: one AS) maintains a table of best paths known so far
- paths are announced to neighbours using the same principles as distance vector, ie. AS announces the best paths it knows
- applies to inter-domain routing
  - no global meaning for costs can be assumed (heterogeneous environment)
  - ASs want control over which paths they use (see policy routing, later)
- Q. Explain how E chooses the paths to n1 and n2
- Q. How can loops be avoided ?

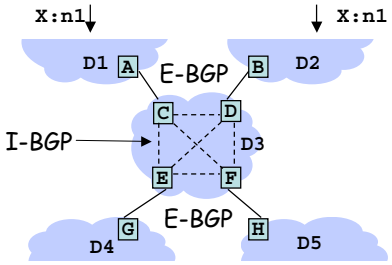


BGP table in E

dest	AS path
n1	B A
n2	B A
n3	D C
n4	D
n5	B

Border Gateways, E-BGP and I-BGP

- BGP runs on routers called *border gateways = "BGP speakers"*-- belong to one AS only
  - two border gateways per boundary
  - Q: compare to OSPF
- In addition, BGP speakers talks to each other inside the AS using "Internal-BGP" (I-BGP) over TCP connections
  - full mesh called the "BGP mesh"
  - I-BGP is the same as E-BGP except for one rule: routes learned from a neighbour in the mesh are not repeated inside the mesh (Q. why ?)
  - Q: Is there a need for all BGP speakers in one network to be adjacent ?

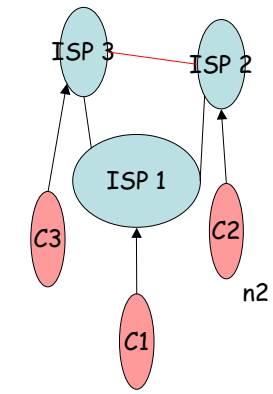



```
A->C:   D1,X:n1
C->E:   D1,X: n1
E->G:   D3,D1,X:n1
```

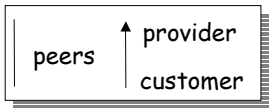

2. Policy Routing

- Interconnection of ASs is self-organized
  - point to point links between networks: ex: EPFL to Switch, Switch to Telianet
  - interconnection points: NAP (Network Access Point), MAE (Metropolitan Area Ethernet), CIX (Commercial Internet eXchange), GIX (Global Internet eXchange), IXP, SFINX, LINX
- Mainly 3 types of relations, depending on money flows
  - customer: EPFL is customer of Switch. EPFL pays Switch
  - provider. Switch is provider for EPFL; Switch is paid by EPFL
  - peer: EPFL and CERN are peers: costs of interconnection is shared

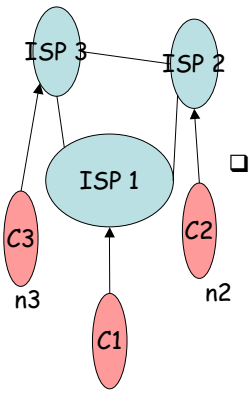
What is the Goal of Policy Routing ?



- Example:
  - ISP3-ISP2 is transatlantic link, cost shared between ISP2 and ISP 3
  - ISP 3- ISP 1 is a local, inexpensive link
  - Ci is customer of ISPi, ISPs are peers
- It is advantageous for ISP3 to send traffic to n2 via ISP1
- ISP1 does not agree to carry traffic from C3 to C2
  - ISP1 offers a "transit service" to C1 and a "non-transit" service to ISP 2 and ISP3
- The goal of "policy routing" is to support this and other similar requirements

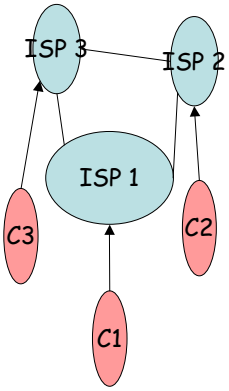


How does Policy Routing Work ?



- Policy routing is implemented by rules imposed to BGP speakers inside an AS, who may
  - refuse to import or announce some paths
  - modify the attributes that control which path is preferred (see later)
- Example
  - ISP 1 announces to ISP 3 all networks of C1 – so that C1 can be reached by all sources in the world
  - ISP 1 announces to C1 all routes it has learnt from ISP3 and ISP2 – so that C1 can send traffic to all destinations in the world
  - ISP2 announces "ISP2 n2" to ISP3 and ISP1 ; assume that ISP1 annouces "ISP1 ISP2 n2" to ISP3.
  - ISP 3 has two routes to n2: "ISP2 n2" and "ISP1 ISP2 n2"; assume that ISP3 gives preference to the latter
  - packets from n3 to n2 are routed via ISP1 – undesired
  - solution: ISP 1 announces to ISP3 only routes to ISP3's customers

Typical Policy Routing Rules

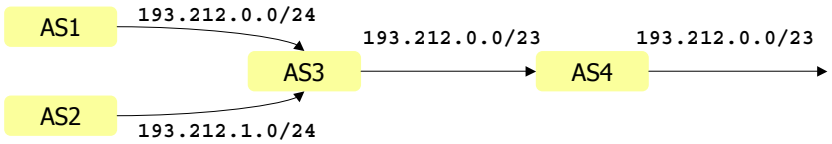


- Provider (ISP1) to customer (C1)
  - announce all routes learnt from other ISs
  - import only routes that belong to domain C1 example: import from EPFL only one route 128.178/15
- Customer (C1) to Provider (ISP1)
  - announce all routes that belong to domain C1
  - import all routes
- Peers (ISP1 to ISP3)
  - announce only routes to all customers of ISP1
  - import only routes to ISP3's customer
  - these routes are defined as part of peering agreement
- The rules are defined by every AS (self-organization) and implemented in all BGP speakers in one AS

3. Aggregation

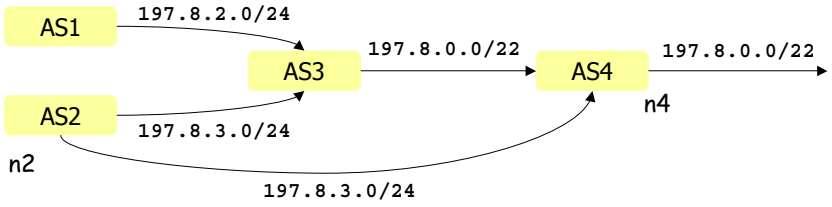
- Domains that do not have a default route (i.e. all transit ISPs) must know all routes in the world (> 120 000)
  - in IP routing tables unless default routes are used
  - in BGP announcements
- Aggregation is a way to reduce the number of routes

Aggregation Example 1



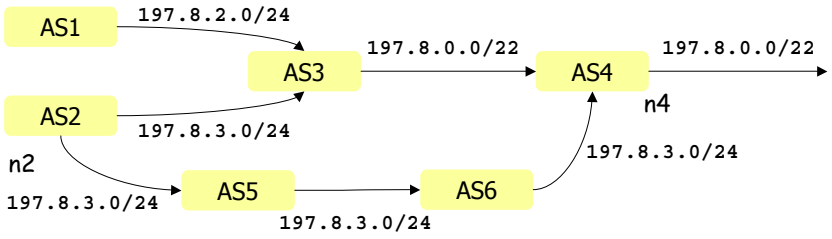
- AS1: 193.212.0.0/24 AS_PATH: 1
- AS2: 193.212.1.0/24 AS_PATH: 2
- AS3: 193.212.0.0/23 AS_PATH: 3 {1 2}
- AS4: 193.212.0.0/23 AS_PATH: 4 3 {1 2}

Aggregation Example 2



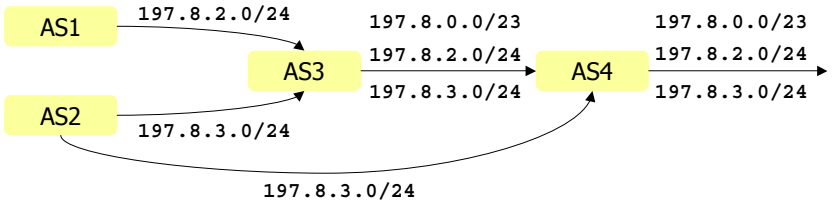
- AS4 receives
  - 197.8.0.0/22 AS_PATH: 3 {1 2}
  - 197.8.3.0/24 AS_PATH: 2
- Both routes are injected into AS4's routing tables
- Q: what happens to packets from n4 to n2 ?

Aggregation Example 3



- AS4 receives
  - 197.8.0.0/22 AS_PATH: 3 {1 2}
  - 197.8.3.0/24 AS_PATH: 6 5 2
- Both routes are received by AS4; only shortest AS paths routes are injected into routing tables
  - Q: what happens to packets from n4 to n2 ?

Example Without Aggregation



- Q: If AS3 does not aggregate, what are the routes announced by AS 4 ? Is there any benefit ?

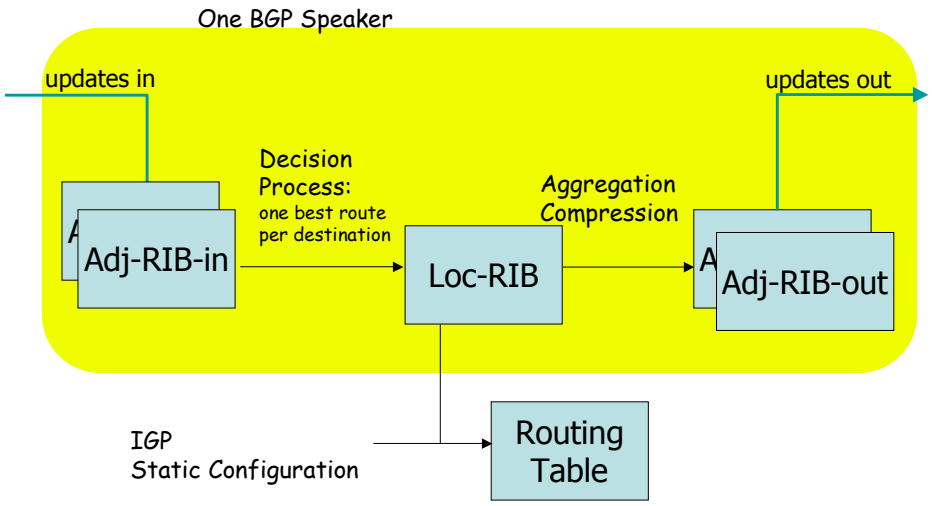
4. BGP (Border Gateway Protocol)

- BGP-4, RFC 1771
- AS border router - BGP speaker
  - peer-to peer relation with another AS border router
  - connected communication
    - on top of a TCP connection, port 179 (vs. datagram (RIP, OSPF))
  - external connections (E-BGP)
    - with border routers of different AS
  - internal connections (I-BGP)
    - with border routers of the same AS
  - BGP only transmits modifications

Routes

- **Route** - unit of information; contains:
  - destination (subnetwork prefix) A
  - path to the destination (AS-PATH)
  - attributes
    - degree of preference (LOCAL-PREF)
    - origin of announcement (ORIGIN)
    - others, see late
- Advertised between a pair of BGP speakers
- Stored locally in RIBs (Routing Information Base)
- Every BGP speaker can add or modify the path attributes, using its *decision process*

Routing Information Bases



Operation of BGP Speaker

BGP speaker :

- stores received routes in **Adj-RIB-in**
  - one per BGP peer (internal or external)
- applies decision process and stores results in **Loc-RIB** (global to BGP speaker)
  - decide which routes to accept
  - decide how to rank them (set LOCAL-PREF)
  - decide which routes to export and with which attributes
- dispatches results per outgoing interface into **Adj-RIB-out** (one per BGP peer), after aggregation and information reduction
- maintains adjacency to peers (over TCP connection): open, keep-alive
- sends updates when Adj-RIB-out changes
- Write forwarding entries in its routing table; redistributes routes learnt from E-BGP from Loc-RIB into IGP and vice-versa, unless other mechanisms are used (See Examples)

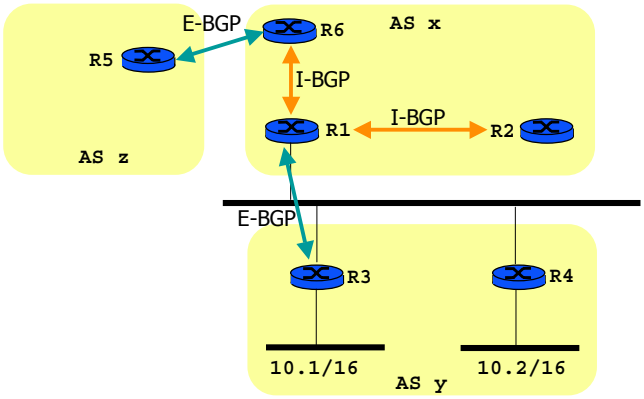
BGP messages

- 4 types
  - OPEN
  - KEEPALIVE
  - NOTIFICATION
  - UPDATE
- Size: from 19 to 4096 bytes
- Security by MD5

Route Attributes

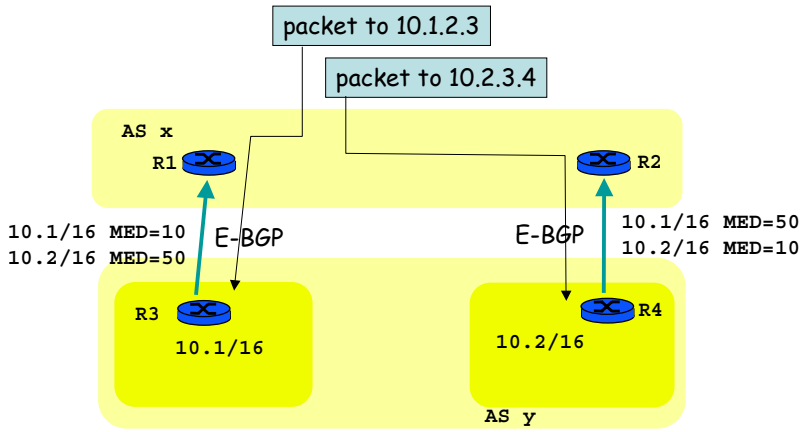
- Well-known Mandatory
  - ORIGIN (route learnt from IGP, BGP or static)
  - AS-PATH
  - NEXT-HOP (see later)
- Well-known Discretionary
  - LOCAL-PREF (see later)
  - ATOMIC-AGGREGATE (= route cannot be dis-aggregated)
- Optional Transitive
  - MULTI-EXIT-DISC (MED)(see later)
  - AGGREGATOR (who aggregated this route)
- Optional Nontransitive
  - WEIGHT (see later)

NEXT-HOP



- R3 advertises 10.2/16 to R1, NEXT-HOP = R4 IP address
- R6 advertises 10.2/16 to R5, NEXT-HOP = R6 IP address
- Q. where is such a scenario likely to happen ?

MULTI-EXIT-DISC (MED)



- One AS connected to another over several links
  - ex: multinational company connected to worldwide ISP
  - AS y advertises its prefixes with different MEDs (low = preferred)
  - If AS x accepts to use MEDs put by ASy: traffic goes on preferred link

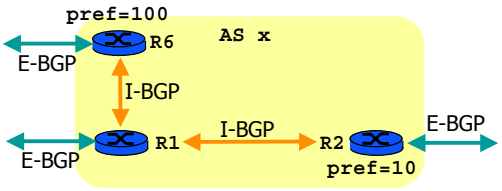
MED Example

- Q1: by which mechanisms will R1 and R2 make sure that packets to ASy use the preferred links ?
- Q2: router R3 crashes; can 10.1/16 still be reached ? explain the sequence of actions.

MED Question

- Q1: Assume now ASx and ASy are peers (ex: both are ISPs). Explain why ASx is not interested in taking MED into account.
- Q2: By which mechanisms can ASx pick the nearest route to ASy ?

LOCAL-PREF

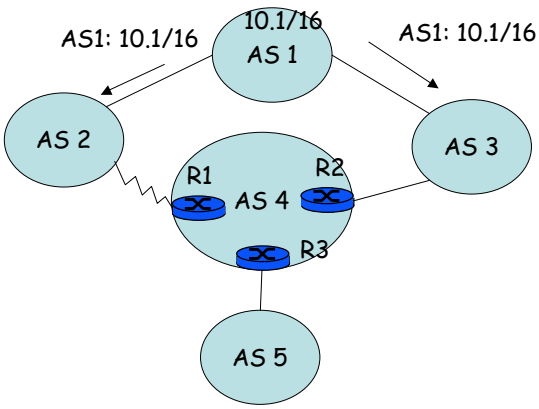


- Used inside an AS to select a best *AS path*
- Assigned by border router when receiving route over E-BGP
  - Propagated without change over I-BGP
- Example
  - R6 associates pref=100, R2 pref=10
  - R1 chooses the largest preference

`bgp default local-preference` *pref-value*

LOCAL-PREF Example

- Q1: The link AS2-AS4 is expensive. How should AS 4 set local-prefs on routes received from AS 3 and AS 2 in order to route traffic preferably through AS 3 ?
- Q2:Explain the sequence of events for R1, R2 and R3.



LOCAL-PREF Question

- Q: Compare MED to LOCAL-PREF

Choice of the best route

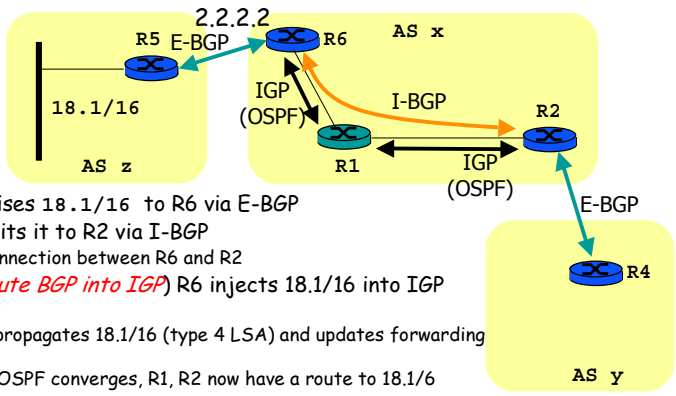
- Done by decision process ; result is: route installed in Loc-RIB
- At most one best route to exactly the same prefix is chosen
  - Only one route to 2.2/16 can be chosen
  - But there can be different routes to 2.2.2/24 and 2.2/16
- Decision Process uses the following priorities (for example)
  1. Highest LOCAL-PREF
  2. Shortest AS-PATH
  3. Lowest MED, if taken seriously by this network
  4. E-BGP > I-BGP
  5. Shortest path to NEXT-HOP, according to IGP
  6. Lowest BGP identifier

5. Interaction BGP—IGP—Packet Forwarding

There are three interactions between BGP and internal routing that you have to know

- *Redistribution*: routes learnt by BGP are passed to IGP (ex: OSPF)
  - Called "redistribution of BGP into OSPF"
  - OSPF propagates the routes using type 4 LSAs to all routers in OSPF cloud
- *Injection*: routes learnt by BGP are written into the forwarding table of this router
  - Routes do not propagate; this helps only this router
- *Synchronization*: see later

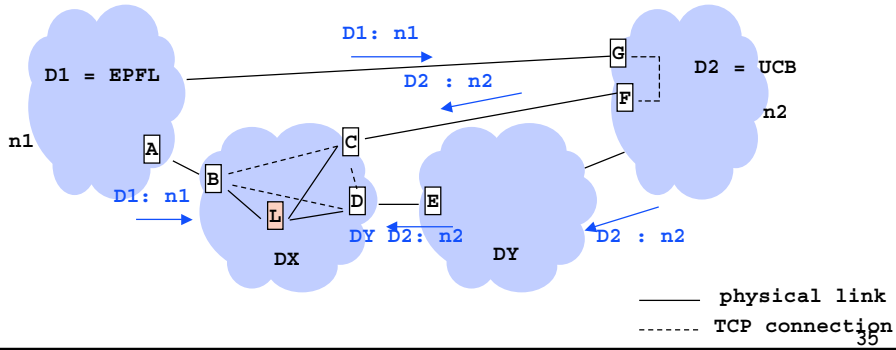
Redistribution Example



- R5 advertises `18.1/16` to R6 via E-BGP
- R6 transmits it to R2 via I-BGP
  - TCP connection between R6 and R2
- (*redistribute BGP into IGP*) R6 injects 18.1/16 into IGP (OSPF)
  - OSPF propagates 18.1/16 (type 4 LSA) and updates forwarding tables
  - After OSPF converges, R1, R2 now have a route to 18.1/6
- R2 advertises route to R4 via E-BGP
  - (*synchronize with IGP*) R2 must wait for the OSPF entry to 18.1/6 before advertising via E-BGP
- Packet to `18.1/16` from AS y finds forwarding table entries in R2, R1 and R6

Example with Re-Distribution

- by ___, F learns from G the route D2-D1-n1
- C redistributes the external route D2:n2 into OSPF
- by _____, D learns the route D2:n2; by _____ D learns the route DYD2:n2; D selects D2:n2 and does not redistribute it to OSPF
- by ___, B learns the route D2:n2
- by ___, A learns the route DX:D2:n2
- by ___, L learns the route to n2 via C



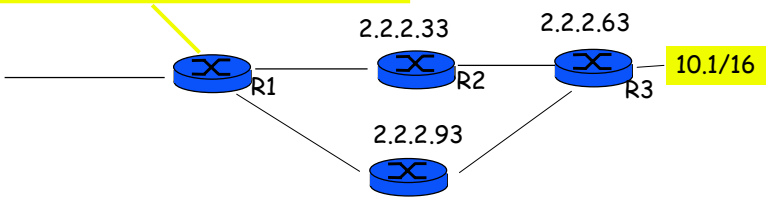
Re-Distribution Considered Harmful

- In practice, operators avoid re-distribution of BGP into IGP
  - Large number of routing entries in IGP
  - Reconvergence time after failures is large if IGP has many routing table entries
- A classical solution is based on *recursive table lookup*
  - When IP packet is submitted to router, the forwarding table may indicate a "NEXT-HOP" which is not on-link with router
  - A second table lookup needs to be done to resolve the next-hop into an on-link neighbour

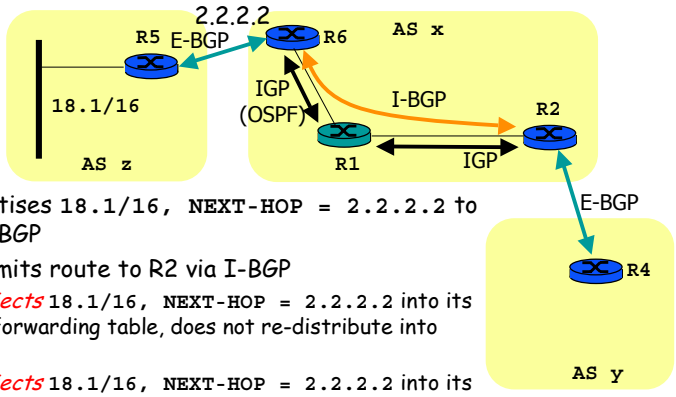
Example: Recursive Table Lookup

- At R1, data packet to 10.1.x.y is received
- The forwarding table at R1 is looked up
  - Q: what are the next events ?

To	*NEXT-HOP*	*layer-2 addr*
10.1/16	2.2.2.63	N/A
2.2.2.63	2.2.2.33	x09:F1:6A:33:76:21

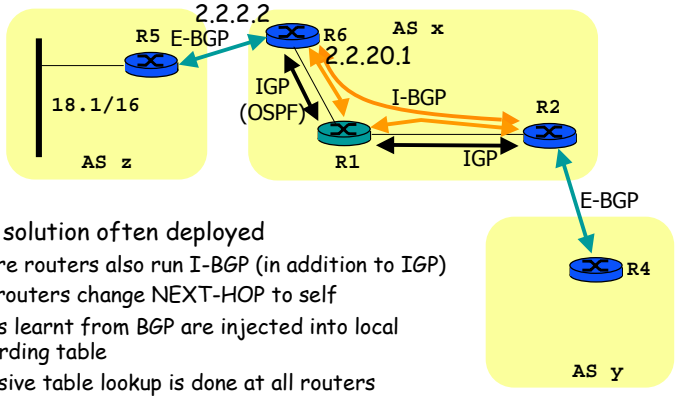


Avoid Redistribution: Combine Recursive Lookup and NEXT-HOP



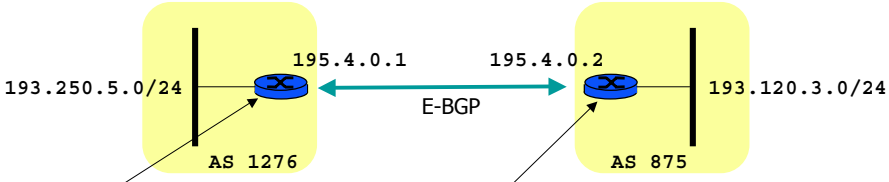
- R5 advertises `18.1/16, NEXT-HOP = 2.2.2.2` to R6 via E-BGP
- R6 transmits route to R2 via I-BGP
  - R6 *injects* `18.1/16, NEXT-HOP = 2.2.2.2` into its local forwarding table, does not re-distribute into OSPF
  - R2 *injects* `18.1/16, NEXT-HOP = 2.2.2.2` into its local forwarding table
- Data packet to 18.1.2.3 is received by R2
  - Recursive table lookup at R2 can be used
  - Q: there is a problem at R1: how can we solve it ?

Avoid Redistribution: Practical Solution



- Practical solution often deployed
  - All core routers also run I-BGP (in addition to IGP)
  - Edge routers change NEXT-HOP to self
  - Routes learnt from BGP are injected into local forwarding table
  - Recursive table lookup is done at all routers
  - Q: repeat the sequence of previous slide with this new assumption
- Potential problem: I-BGP mesh -> use reflectors
- IGP handles only internal networks – very few

6. Configuration on CISCO




```
router bgp 1276
neighbor 195.4.0.2 remote-as 875
network 193.250.5.0 255.255.255.0
```

```
router bgp 875
neighbor 195.4.0.1 remote-as 1276
network 193.120.3.0 255.255.255.0
```

Route filtering

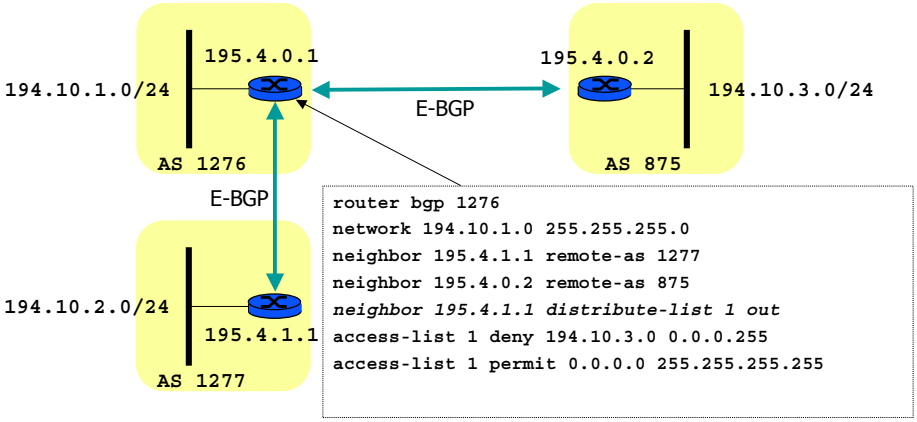
- Associate an access list with a neighbor

```
neighbor ID distribute-list no-of-the-list [in/out]
```

- Define an access list
  - non-significant-bits (inverse of the netmask)
  - if no action specified at the end of the list, apply "deny everything else"

```
access-list No-of-the-list [deny/permit]
                    IP-address non-significant-bits
```


Route filtering



- AS 1276 does not want to forward traffic to `194.10.3.0/24` of AS 875 - it does not re-advertise this prefix

Path filtering

- Associate a filter list with a neighbor

neighbor *ID* **filter-list** *no-of-the-list* **[in/out]**

- Define a filter list

ip as-path access-list *no-of-the-list* **[deny/permit]**
regular-expression

- Regular expressions

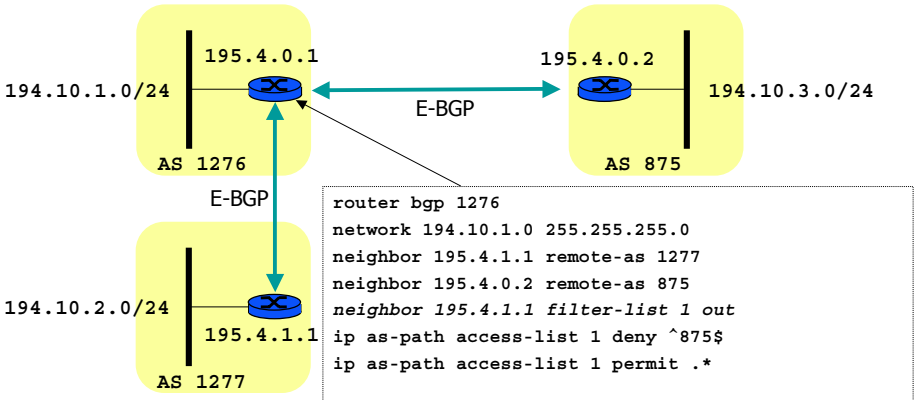
```
^  beginning of the path
$  end of the path
.  any character
?  one character
_  matches ^ $ ( ) 'space'
*  any number of characters (zero included)
+  any number of characters (at least one)
```


Path filtering

- Examples

`^$` - local routes only (empty AS_PATH)
`.*` - all routes (all paths AS_PATH)
`^300$` - AS_PATH = 300
`^300_` - all routes coming from 300 (e.g. AS_PATH = 300 200 100)
`_300$` - all routes originated at 300 (e.g. AS_PATH = 100 200 300)
`_300_` - all routes passing via 300 (e.g. AS_PATH = 200 300 100)

Path filtering



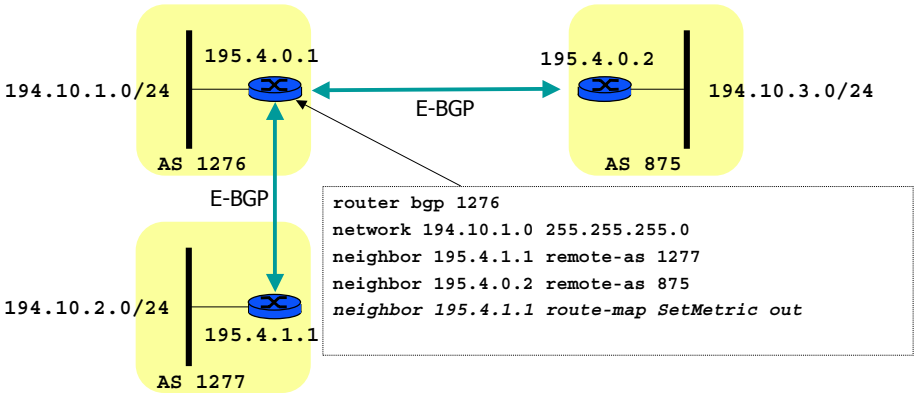
- AS 1276 does not want to forward traffic for all internal routes of AS 875

Route maps

```
route-map map-tag [permit|deny] instance-no
first-instance-conditions: set match
next-instance-conditions: set match
...
route-map SetMetric permit 10
match ip address 1
set metric 200

route-map SetMetric permit 20
set metric 300

access-list 1 permit 194.10.3.0 0.0.0.255
```

- Set metric 200 (MED) on route `194.10.3/24`

Route maps



- Set metric 200 on route `194.10.3/24`, 300 otherwise

Route maps

```
neighbor 192.68.5.2 route-map SetLocal in

route-map SetLocal permit 10
set local-preference 300
```

- Set LOCAL_PREF to 300

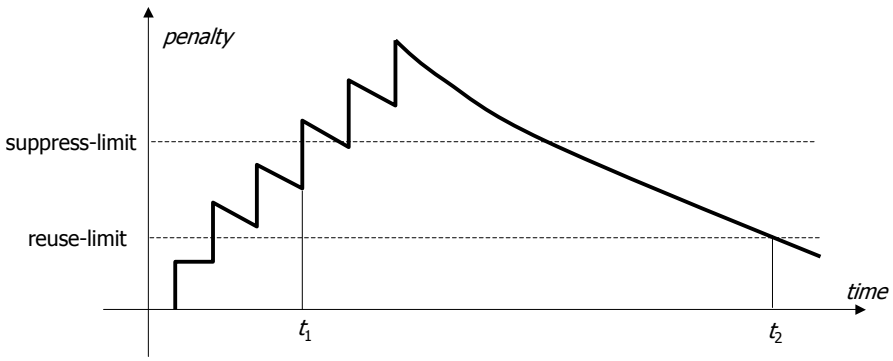
```
neighbor 172.16.2.2 route-map AddASnum out

route-map AddASnum permit 10
set as-path prepend 801 801
```

- Prepend AS 801 801 to AS_PATH (makes it longer)

7. Other Mechanisms in BGP
Route Flap Dampening

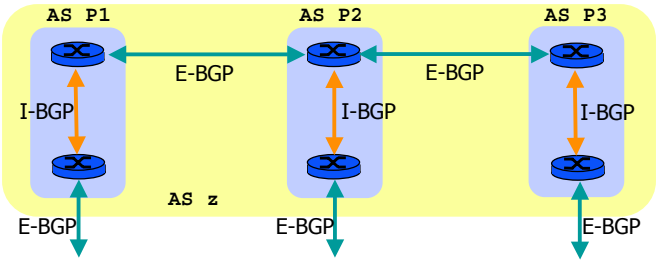
- Route modification propagates everywhere
- Sometimes routes are *flapping*
  - successive UPDATE and WITHDRAW
  - caused for example by BGP speaker that often crashes and reboots
- Solution:
  - decision process eliminates flapping routes
- How
  - withdrawn routes are kept in Adj-RIN-in
  - if comes up again soon (ie : flap), route receives a penalty
  - penalty fades out exponentially
  - used to suppress or restore routes

Route Flap Dampening



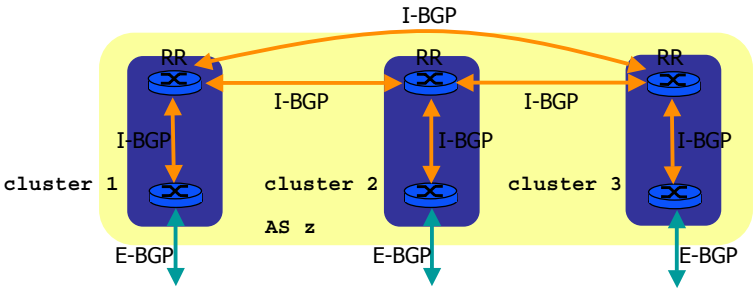
- Route suppressed at t_1, restored at t_2

Avoid I-BGP Mesh: Confederations



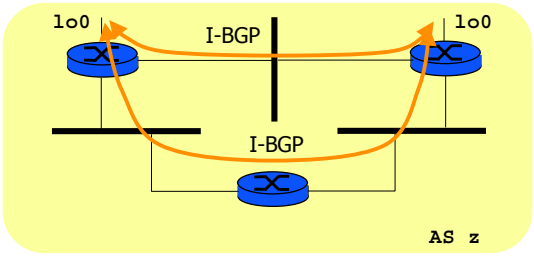
- AS decomposed into sub-AS
  - private AS number
  - similar to OSPF areas
    - I-BGP inside sub-AS (full interconnection)
    - E-BGP between sub-AS

Avoid I-BGP Mesh : Route reflectors



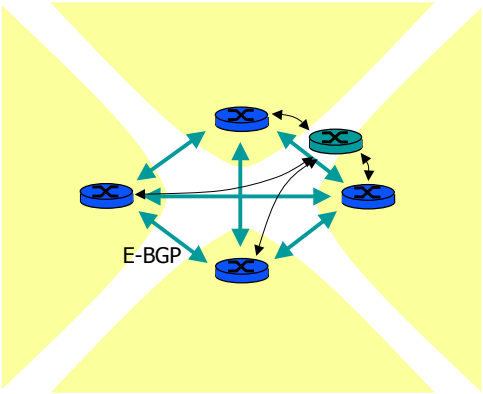
- Cluster of routers
  - one I-BGP session between one client and RR
  - CLUSTER_ID
- Route reflector
  - re-advertises a route learnt via I-BGP
  - to avoid loops
    - ORIGINATOR_ID attribute associated with the advertisement

I-BGP configuration



- I-BGP configured on loopback interface (lo0)
  - interface always up
  - IP address associated with the interface
  - IGP routing guarantees packet forwarding to the interface

Avoid E-BGP mesh: Route server



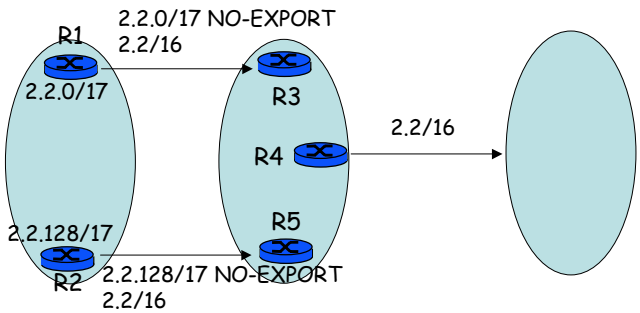
- At interconnection point
- Instead of $n(n-1)/2$ peer to peer E-BGP connections
- n connections to Route Server
- To avoid loops ADVERTISER attribute indicates which router in the AS generated the route

Communities

- Other attributes can be associated with routes in order to *simplify* rules. They are called « communities »
  - Pre-defined: Example: NO-EXPORT (a well known, pre-defined attribute) – see later for an example
  - Defined by one AS (a label of the form ASN:x where AS= AS number, x = a 2 byte—number)

NO—EXPORT

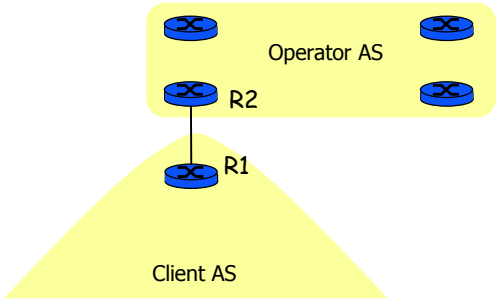
- Written on E-BGP by one AS, transmitted on I-BGP by accepting AS, not forwarded
- Example: AS2 has different routes to AS1 but AS2 sends only one aggregate route to AS3
  - simplifies the aggregation rules at AS2
  - What is the route followed by a packet sent to 2.2.48 received by R4 ?



8. Examples

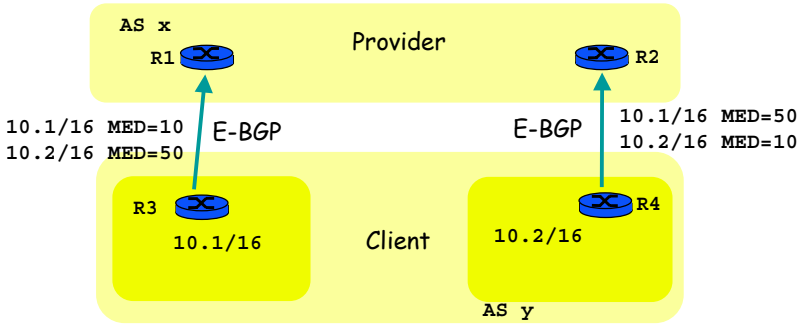
- Dual Homing
- Hot potato routing

Ex1: Stub Area



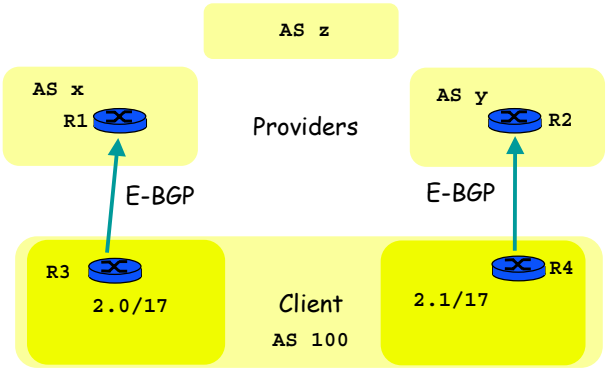
- BGP not needed between Client and Operator
- No AS number for client
- R2 learns all prefixes in Client by static configuration or RIP on link R1—R2
- Example: EPFL and Switch
- Q: what if R1 fails ?

Ex2: Stub Area, Dual Homing to Single Provider



- With numbered Client AS
  - Use MED to share traffic from ISP to Client on two links
  - Use Client IGP configuration to share traffic from Client to two links
  - Q1: is it possible to avoid distributing BGP routes into Client IGP ?
  - Q2: is it possible to avoid assigning an AS number to Client ?
  - Q3: is it possible to avoid BGP between Client and Provider ?

Ex3: Stub Area, Dual Homing to Several Providers

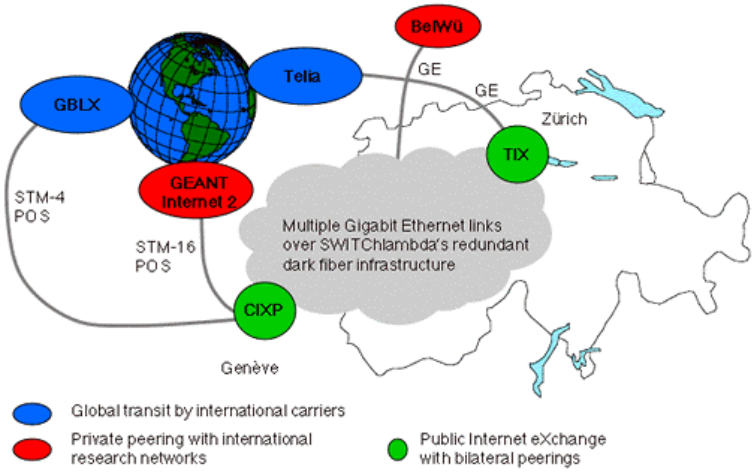


- Client has own address space and AS number
- Q: how can routes be announced between AS 100 and AS x ? AS x and AS z ?
- Q: assume Client wants most traffic to favour AS x. How can that be done ?

Ex4: Hot Potato Routing

- Packets from Customer 2 to Customer 1
  - Both R21 and R22 have a route to Customer 1
  - Shortest path routing favours R21
  - Q1: by which mechanism is that done ?
- Q2: what is the path followed in the reverse direction ?

9. Illustrations: Switch



An Interconnection Point

swissix SWISS INTERNET EXCHANGE

E-Mail | Credits

Expand all | Collapse all

- General Information
- Services
- Costs
  - Membership fees
  - Connection fees
- Legal
  - Articles of association
  - Peering Policy
  - Connection agreement
- Members
  - Member list
  - Board members
  - Membership application
- Member Login
- Tech Corner
- Links

Welcome to swissix

The Swissix (Swiss Internet Exchange) in Zurich, Switzerland, is now open. We are pleased to welcome ISPs and hosting companies as members and peering partners.

With continued growth of Internet traffic, we want to make sure that there is sufficient reliability built into the Swiss Internet. By exchanging traffic at multiple exchanges points, you can help ensure that consumers have fast Internet access and network operators have multiple routes for their traffic flows.

The Swiss Internet Exchange (swissix) is a neutral and independent exchange and a place for Internet Service Providers (ISPs) to interconnect and exchange IP traffic with each other at a national or international level.

63

swissix SWISS INTERNET EXCHANGE

E-Mail | Credits

Expand all | Collapse all

- General Information
- Services
- Costs
  - Membership fees
  - Connection fees
- Legal
  - Articles of association
  - Peering Policy
  - Connection agreement
- Members
  - Member list
  - Board members
  - Membership application
- Member Login
- Tech Corner
- Links

Membership fees

The yearly membership fee is CHF 100.- per company. The membership fee is not refundable.

Connection fees

Action: Till end of 2003 all port and connectionfees are free.

The connection fees consitst of a monthly and a one-time installation fee and depend on the connection port type.

Port type	Monthly CHF	One-time CHF
100BaseTX	275.00	1150.00
1000BaseSX/LX	1520.00	3000.00

Prices are excluding VAT (7.6%)

Deployment of 100BaseTX: immediately
Deployment of 1000BaseSX/LX: 2-4 weeks

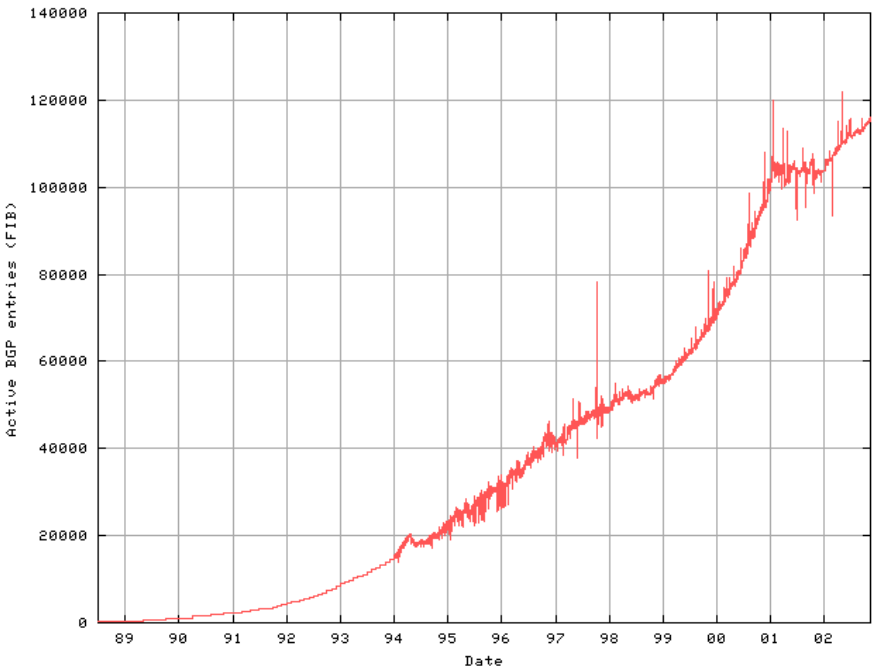
64

from www.ris.ripe.net: all routes to 128.178.0.0/15 on RIPE Route Collectors

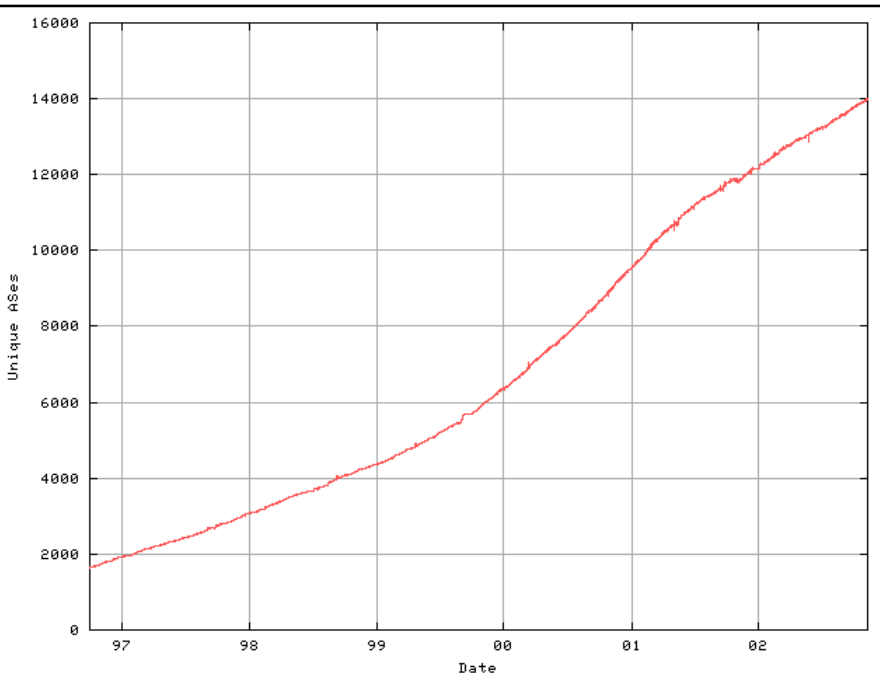
Type				Next HOP	MED	Origin		Community	RRC ID
A	128.178.0.0 /15	2003-10-02 05:05:49Z	129.250.0.2 32	129.250.0.2 32	9	Not defined	2914 1299 559	2914:420 2914:2000 2914:3000	RIPE NCC
A	128.178.0.0 /15	2003-10-02 06:16:00Z	193.10.252. 5	193.10.252. 5	0	IGP	2603 3356 1299 559	2603:666 3356:2 3356:86 3356:507 3356:666 3356:2076	Netnod
A	128.178.0.0 /15	2003-10-02 06:16:17Z	194.68.48.1	194.68.48.1	0	IGP	12381 1653 2603 20965 559	12381:1653	Netnod
A	128.178.0.0 /15	2003-10-02 06:16:37Z	194.68.48.1	194.68.48.1	0	IGP	12381 1653 2603 3356 1299 559	12381:1653	Netnod
A	128.178.0.0 /15	2003-10-02 06:21:08Z	193.10.252. 5	193.10.252. 5	0	IGP	2603 20965 559	2603:222 2603:666 20965:155	Netnod
A	128.178.0.0 /15	2003-10-02 06:21:17Z	194.68.48.1	194.68.48.1	0	IGP	12381 1653 2603 20965 559	12381:1653	Netnod
A	128.178.0.0 /15	2003-10-02 07:24:06Z	129.250.0.2 32	129.250.0.2 32	9	Not defined	2914 3549 559	2914:420	

Some statistics

- Number of routes
  - 1988-1994: exponential increase
  - 1994-1995: CIDR
  - 1995-1998: linear increase (10000/year)
  - 1999-2000: return to exponential increase (42% per year)
  - since 2001: return to linear increase, ~120,000
- Number of ASs
  - 51% per year for 4 last years
  - 14000 AS effectively used
- Number of IP addresses
  - 162,128,493 (Jul 2002)
  - 7% per year

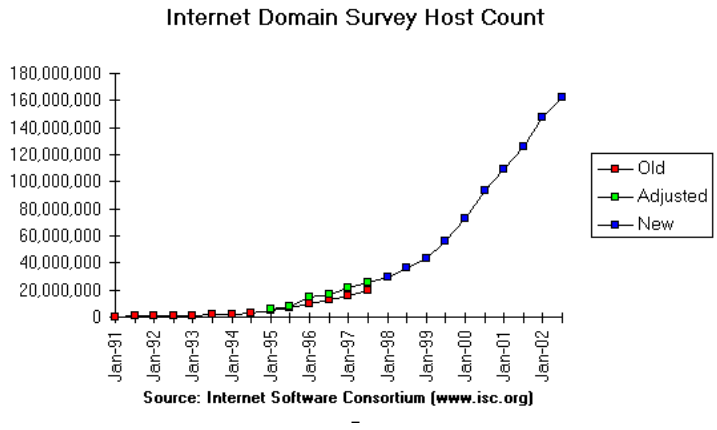

Active BGP entries (FIB)
140000
120000
100000
80000
60000
40000
20000
0
89
90
91
92
93
94
95
96
97
98
99
00
01
02
Date

67


Unique ASes
16000
14000
12000
10000
8000
6000
4000
2000
0
97
98
99
00
01
02
Date

68

Number of hosts

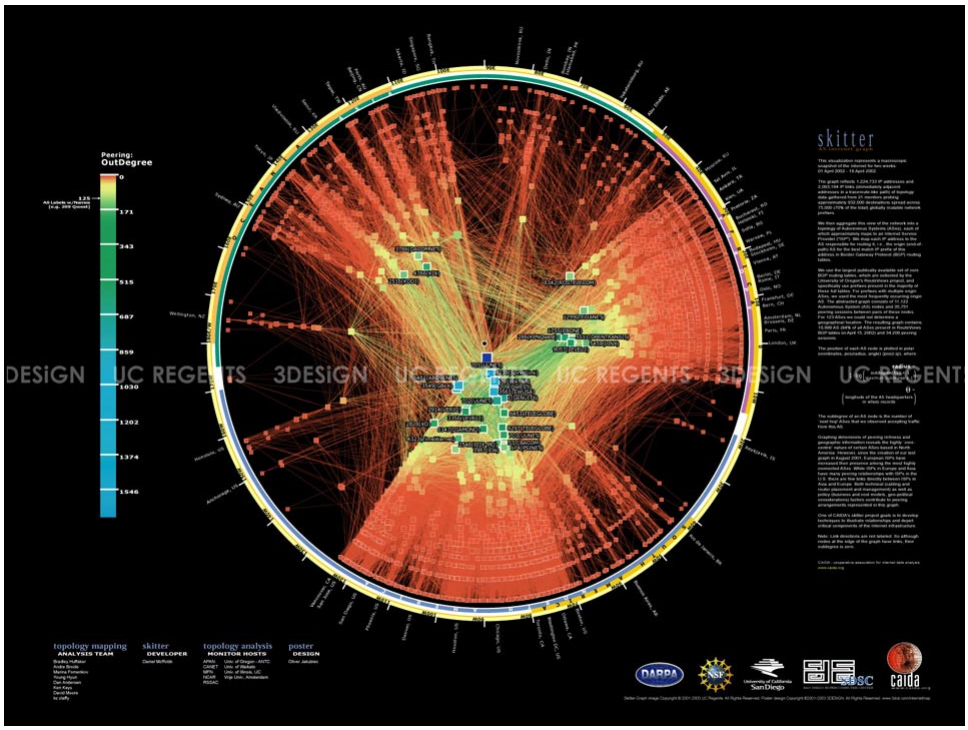


BGP statistics

BGP routing table entries examined: 1	17013
Total ASes present in the Internet Routing Table: 1	4042
Origin-only ASes present in the Internet Routing Table:	12159
Transit ASes present in the Internet Routing Table:	1883
Transit-only ASes present in the Internet Routing Table:	63
Average AS path length visible in the Internet Routing Table:	5.3
Max AS path length visible:	23
Number of addresses announced to Internet:	1182831464
Equivalent to 70 /8s, 128 /16s and 147 /24s	
Percentage of available address space announced:	31.9
Percentage of allocated address space announced:	58.5

Prefix length distribution

/1:0 /2:0 /3:0 /4:0 /5:0 /6:0
/7:0 /8:17 /9:5 /10:8 /11:12 /12:46
/13:90 /14:239 /15:430 /16:7308 /17:1529 /18:2726
/19:7895 /20:7524 /21:5361 /22:8216 /23:9925 /24:64838
/25:185 /26:221 /27:126 /28:105 /29:85 /30:93
/31:0 /32:29



Exercise

- What ASs does EPFL receive service from ?
- What ASs does Switch receive service from ?
- Find the names of the networks that have these AS numbers

Exercise

- Lookup http://rpsl.info.ucl.ac.be. to find out the relationships between Switch and other providers
- How does the software on this site decide whether a relationship is client, provider or peer ?

References

- Timothy Griffin's home page at Intel
- **The stable paths problem and interdomain routing**
  Griffin, T.G.; Shepherd, F.B.; Wilfong, G.
  ACM/IEEE ToN April 2002, Page(s): 232-243

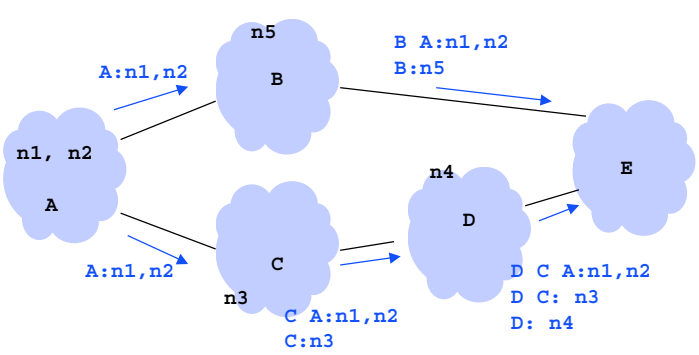
  (fundamental issues in BGP – which policies are implementable)

- RFC 1771 (BGP-4)
- C. Huitema, "Le Routage dans l'Internet"
- John W. Stewart III " BGP 4"
- www.ris.ripe.net : AS paths
- www.cidr-report.org aggregation statistics
- www.caida.org map of Internet
- rpsl.info.ucl.ac.be relations between ASs

Solutions

Path Vector Routing

- Q. Explain how E chooses the paths to n1 and n2
  - A. E receives the routes "B A n1" and "D C A n1". E selects as best routes the ones with shorter AS path.
- Q. How can loops be avoided ?
  - A: BGP routers recognize looping announcements by the repetition of the same AS in the path. Such announcements are discarded

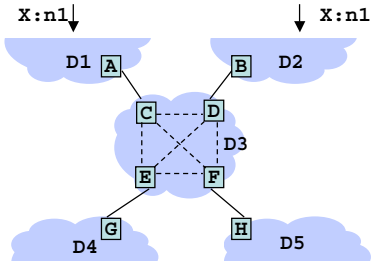


BGP table in E

dest	AS path
n1	B A
n2	B A
n3	D C
n4	D
n5	B

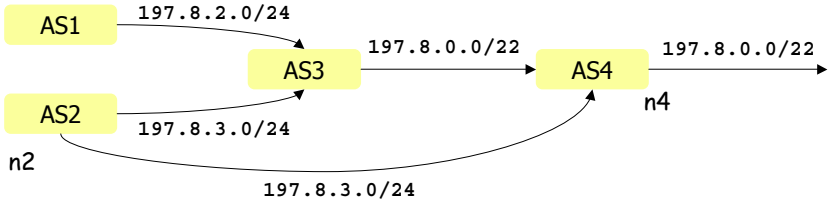
Border Gateways, e-BGP and I-BGP

- BGP runs on routers called *border gateways = "BGP speakers"*-- belong to one AS only
  - Q: compare to OSPF
  - A: there is one single inter-area router per area boundary: it belongs to both areas
- In addition, BGP speakers talks to each other inside the AS using "Internal-BGP" (I-BGP) over TCP connections
  - I-BGP is the same as E-BGP except for one rule: routes learned from a neighbour in the mesh are not repeated inside the mesh (Q. why ?)
    A: otherwise loops cannot be avoided (same AS number !)
  - Q: Is there a need for all BGP speakers in one network to be adjacent ?
    A: no, they are generally not. The mesh is over TCP connections.



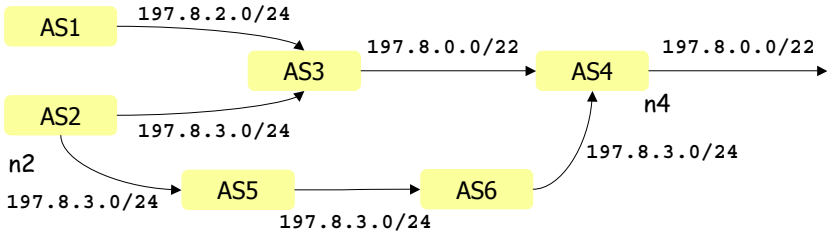

```
A->C:   D1,X:n1
C->E:   D1,X: n1
E->G:   D3,D1,X:n1
```

Aggregation Example 2



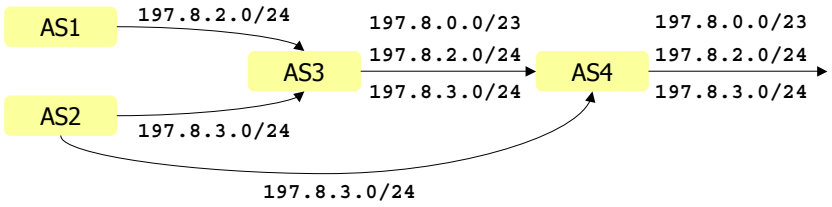
- AS4 receives
  - `197.8.0.0/22` `AS_PATH: 3 {1 2}`
  - `197.8.3.0/24` `AS_PATH: 2`
- Both routes are injected into AS4's routing tables
  - Q: what happens to packets from n4 to n2 ?
  - A: it depends on the attributes set by the rules in AS4; by default, the direct route to n2 is preferred (fewer Ass in path). There are two routing entries in AS4 routers: one for `197.8.0.0/22` and one for `197.8.3.0/24`. Longest prefix match in the packet forwarding algorithm ensures that packets to n2 go on the direct route.

Aggregation Example 3



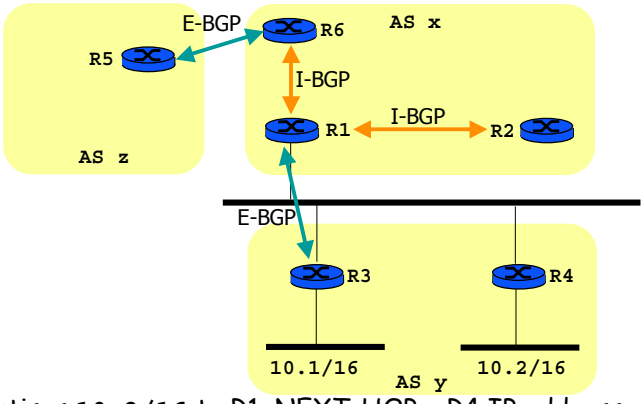
- AS4 receives
  - `197.8.0.0/22` `AS_PATH: 3 {1 2}`
  - `197.8.3.0/24` `AS_PATH: 6 5 2`
- Both routes are received by AS4; only shortest AS paths routes are injected into routing tables Q: what happens to packets from n4 to n2 ?
  - A: now packets to n2 go via AS3

Example Without Aggregation



- Q: If AS3 does not aggregate, what are the routes announced by AS 4 ? Is there any benefit ?
- A:
  - `197.8.0.0/23` `AS_PATH: 4 3`
  - `197.8.2.0/24` `AS_PATH: 4 3 1`
  - `197.8.3.0/24` `AS_PATH: 4 2`
- A: there is no benefit since all routes go via AS 4 anyhow. AS4 should aggregate.

NEXT-HOP



- R3 advertises `10.2/16` to R1, NEXT-HOP = R4 IP address
- R6 advertises `10.2/16` to R5, NEXT-HOP = R6 IP address
- Q. where is such a scenario likely to happen ?
- A: in interconnection points with many providers interconnected on one LAN

MED Example

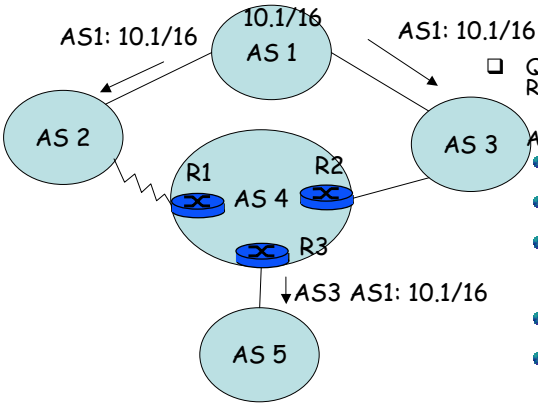
- Q1: by which mechanisms will R1 and R2 make sure that packets to ASy use the preferred links ?
  A:
  - R1 and R2 exchange their routes to AX y via I-BGP
  - R1 has 2 routes to 10.1/16, one of them learnt over E-BGP; prefers route via R1; injects it into IGP
  - R1 has 2 routes to 10.2/16, one of them learnt over E-BGP; prefers route via R2; does not inject a route to 10.2/16 into IGP
- Q2: router R3 crashes; can 10.1/16 still be reached ? explain the sequence of actions.
  A:
  - R1 clears routes to ASy learnt from R1 (keep-alive mechanism)
  - R2 is informed of the route suppression by I-BGP
  - R2 has now only 1 route to 10.1/16 and 1 route to 10.2/16;. keeps both routes in its local RIB and injects them into IGP since both were learnt via E-BGP
  - traffic to 10.1/16 now goes to R2

MED Question

- Q1: Assume now ASx and ASy are peers (ex: both are ISPs). Explain why ASx is not interested in taking MED into account.
  A: ASx is interested in sending traffic to ASy to the nearest exit, avoiding transit inside ASx as much as possible. Thus ASx will choose the nearest route to ASy, and will ignore MEDs
- Q2: By which mechanisms can ASx pick the nearest route to ASy ?
  A: it depends on the IGP. With OSPF: all routes to ASy are injected into OSPF by means type 5 LSAs. These LSAs say: send to router R3 or R4. Every OSPF router inside ASx knows the cost (determined by OSPF weights) of the path from self to R3 and R4. Packets to 10.1/16 and 10.2/16 are routed to the nearest among R3 and R4 (nearest = lowest OSPF cost).

LOCAL-PREF Example

- Q1: The link AS2-AS4 is expensive. How should AS 4 set local-prefs on routes received from AS 3 and AS 2 in order to route traffic preferably through AS 3 ?
  A: for example: set LOCAL-PREF to 100 to all routes received from AS 3 and to 50 to all routes received from AS 2



- Q2: Explain the sequence of events for R1, R2 and R3

A:

- R1 receives the route AS2 AS1 10.1/16 over E-BGP; sets LOCAL-PREF to 50
- R2 receives the route AS3 AS1 10.1/16 over E-BGP; sets LOCAL-PREF to 100
- R3 receives AS2 AS1 10.1/16, LOCAL-PREF=50 from R1 over I-BGP and AS3 AS1 10.1/16, LOCAL-PREF=100 from R1 over I-BGP
- R3 selects AS3 AS1 10.1/16, LOCAL-PREF=100 and installs it into local-RIB
- R3 announces only AS3 AS1 10.1/16 to AS 5

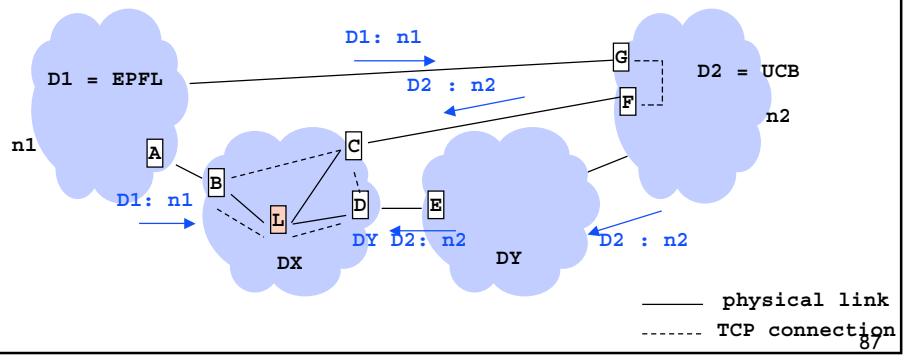
LOCAL-PREF Question

- Q: Compare MED to LOCAL-PREF
  A:
  - MED is used between ASs (i.e. over E-BGP); LOCAL-PREF is used inside one AS (over I-BGP)
  - MED is used to tell one provider AS which *entry link* to prefer; LOCAL-PREF is used to tell the rest of the world which *AS path* we want to use, by not announcing the other ones.

Example with Re-Distribution

- by I-BGP, F learns from G the route to D2-D1-n1
- C redistributes the external route D2:n2 into OSPF;
- by I-BGP, D learns the route D2:n2; by E-BGP D learns the route DYD2:n2; D selects D2:n2 and does not redistribute it to OSPF
- by I-BGP, B learns the route D2:n2 from C
- by E-BGP, A learns the route DX:D2:n2
- by OSPF, L learns the route to n2 via C

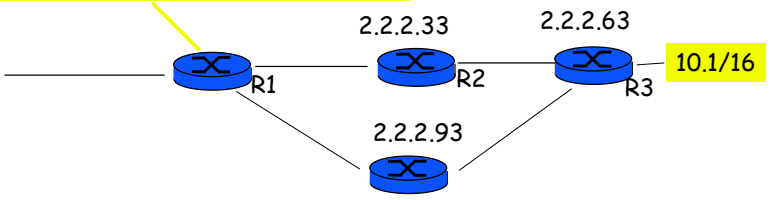
I-BGP - internal BGP
E-BGP - external BGP



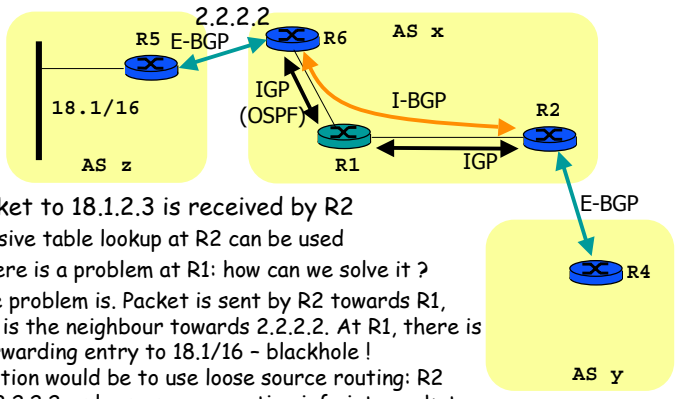
Example: Recursive Table Lookup

- At R1, data packet to 10.1.x.y is received
- The forwarding table at R1 is looked up
  - Q: what are the next events ?
  - A: first, the nex-hop 2.2.2.63 is found; a second lookup for 2.2.2.63 is done; the packet is sent to MAC address `x09:F1:6A:33:76:21`

To	*NEXT-HOP*	*layer-2 addr*
10.1/16	2.2.2.63	N/A
2.2.2.63	2.2.2.33	x09:F1:6A:33:76:21

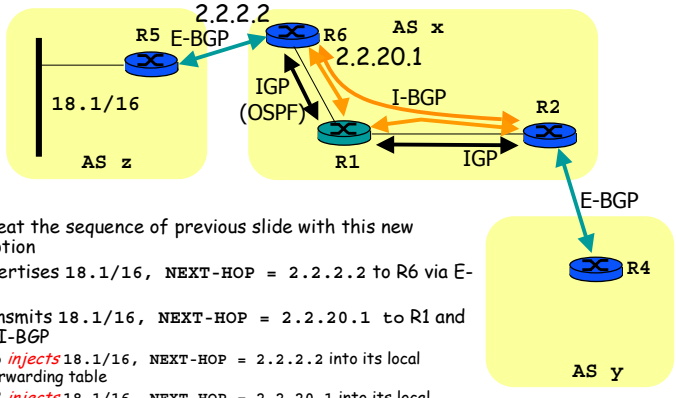


Avoid Redistribution: Combine Recursive Lookup and NEXT-HOP



- Data packet to 18.1.2.3 is received by R2
  - Recursive table lookup at R2 can be used
  - Q: there is a problem at R1: how can we solve it ?
  - A: the problem is. Packet is sent by R2 towards R1, which is the neighbour towards 2.2.2.2. At R1, there is no forwarding entry to 18.1/16 – blackhole !
    A solution would be to use loose source routing: R2 adds 2.2.2.2 as loose source routing info into packet. In practice however, source routing is not used with IPv4. See later in the section for another solution.

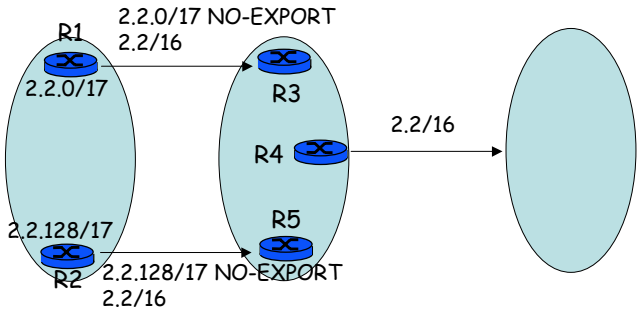
Avoid Redistribution: Practical Solution



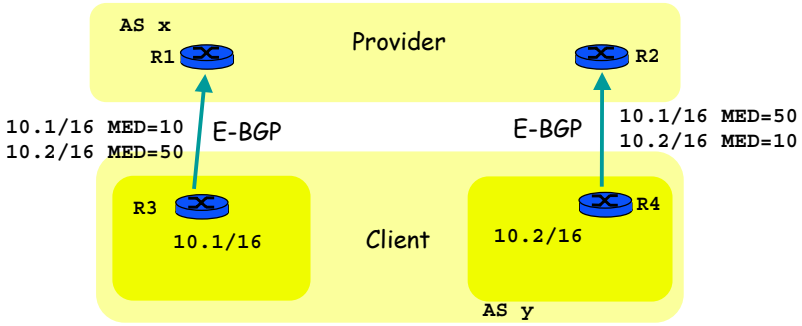
- Q: repeat the sequence of previous slide with this new assumption
- R5 advertises `18.1/16, NEXT-HOP = 2.2.2.2` to R6 via E-BGP
- R6 transmits `18.1/16, NEXT-HOP = 2.2.20.1` to R1 and R2 via I-BGP
  - R6 *injects* `18.1/16, NEXT-HOP = 2.2.2.2` into its local forwarding table
  - R2 *injects* `18.1/16, NEXT-HOP = 2.2.20.1` into its local forwarding table
- Independently, IGP finds that, at R2, packets to 2.2.10.1 should be sent to R1
- Data packet to 18.1.2.3 is received by R2
  - At R2, recursive table lookup determines that packet should be forwarded to R1
  - At R1, recursive table lookup determines that packet should be forwarded to R6
  - At R6, recursive table lookup determines that packet should be forwarded to 2.2.2.2

NO—EXPORT

- Q: What is the route followed by a packet sent to 2.2.48 received by R4 ?
- A: the packet is sent via R3 and R1

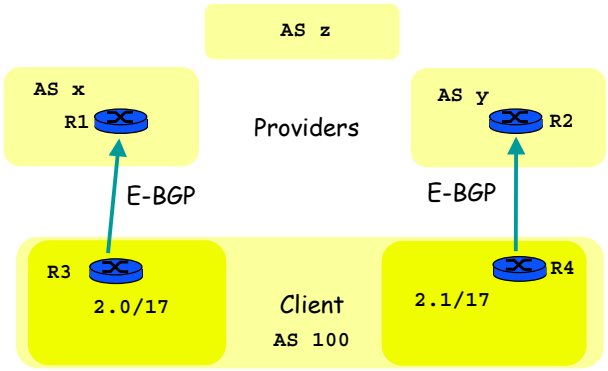


Ex2: Stub Area, Dual Homing to Single Provider



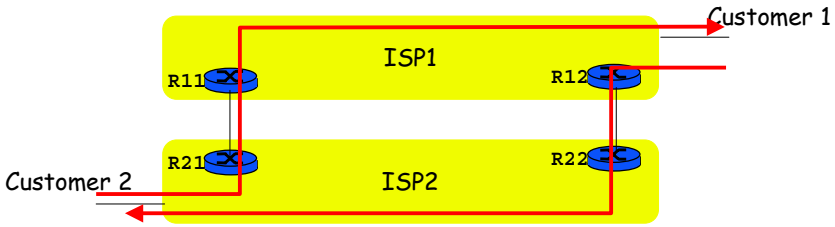
- Q1: is it possible to avoid distributing BGP routes into Client IGP ?
- A: yes, for example: configure R3 and R4 as default routers in Client AS; traffic from Client AS is forwarded to nearest of R3 and R4. If R3 or R4 fails, to the remaining one
- Q2: is it possible to avoid assigning an AS number to Client ?
- A: Yes, it is sufficient to assign to Client a private AS number: Provider translates this number to its own.
- Q3: is it possible to avoid BGP between Client and Provider ?
- A: Yes, by running a protocol like RIP between Client and Provider and redistributing Client routes into Provider IGP. Thus Provider pretends to the rest of the world that the prefixes of Client are its own.

Ex3: Stub Area, Dual Homing to Several Providers



- Client has own address space and AS number
- Q: how can routes be announced between AS 100 and AS x ? AS x and AS z ?
  A: R3 announces 2.0/17 and 2.0/16; traffic from AS x to 2.0/17 will flow via AS x; if R3 fails, it will use the longer prefix and flow via Asy.
  ASx announces 2.0/17 and 2.0/16 to AS z
- Q: assume Client wants most traffic to favour AS x. How can that be done ?
  A: R3 announces an artificially inflated path: 100 100 100 100 : 2.0/17. AS z will favour the path via Asy which has a shorter AS path length

Ex4: Hot Potato Routing



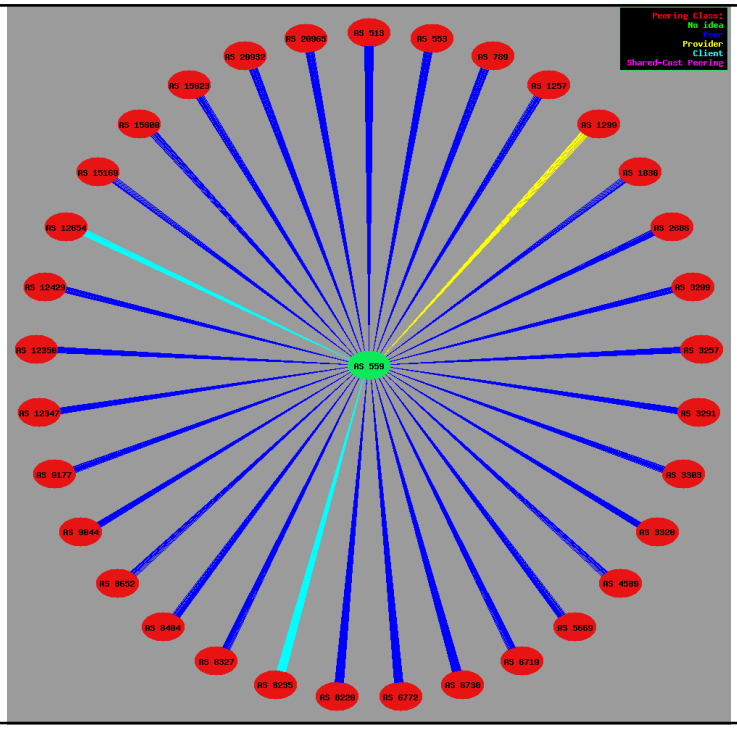
- Packets from Customer 2 to Customer 1
  - Both R21 and R22 have a route to Customer 1
  - Shortest path routing favours R21
  - Q1: by which mechanism is that done ?
  - A: « Choice of the best route » (criteria 5), assuming alll routers in ISP2 run BGP
- Q2: what is the path followed in the reverse direction ?
  - A: see picture. Note the asymmetric routing

Exercise

- What ASs does EPFL receive service from ?
  - from the previous routes, we find AS 559 (Switch)
- What ASs does Switch receive service from ?
  - from the previous routes we see that there are at least:
    - AS 1299
    - AS 20965
    - AS 3549
- Find the names of the networks that have these AS numbers
  - from whois on www.ripe.net:
    - AS 1299: Telianet
    - AS 20965: Geant
    - AS 3549: Global Crossing

Exercise

- Lookup http://rpsl.info.ucl.ac.be. to find out the relationships between Switch and other providers
- How does the software on this site decide whether a relationship is client, provider or peer ?
  - AS X is client of Switch if AS X accepts ANY path and announces only self (AS X)
  - AS X is provider of Switch if AS X announces ANY path and accepts only AS Switch
  - AS X is a peer if AS X accepts and announces only a small set of routes


Peering Class:
No idea
Peer
Provider
Client
Shared-Cost Peering
AS 559
AS 513
AS 553
AS 789
AS 1257
AS 1299
AS 1838
AS 2686
AS 3209
AS 3257
AS 3291
AS 3303
AS 3320
AS 4589
AS 5669
AS 6719
AS 6730
AS 6772
AS 8220
AS 8235
AS 8327
AS 8404
AS 8652
AS 9044
AS 9177
AS 12347
AS 12350
AS 12429
AS 12654
AS 15169
AS 15600
AS 15623
AS 20932
AS 20965