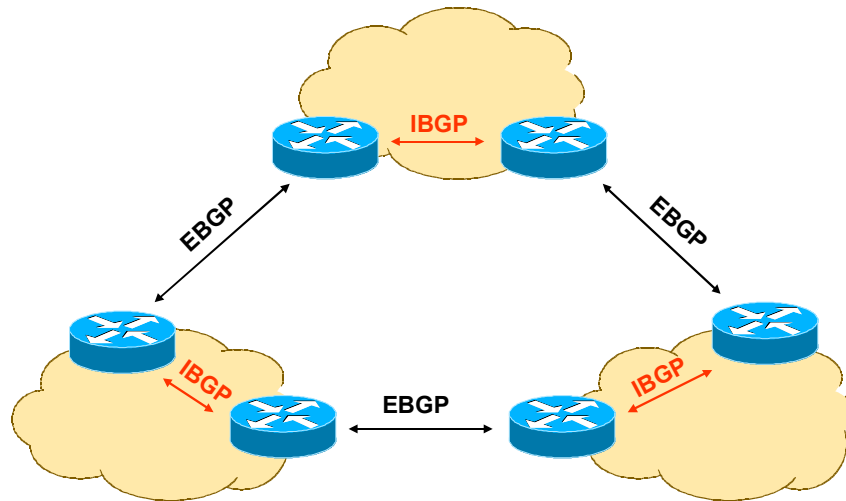


# **BGP**

## Internal and External BGP

# EBGP and IBGP



(C) Herbert Haas 2005/03/11

2

Interior BGP or "IBGP" allows edge routers to share NLRI and associated attributes, in order to enforce an AS-wide routing policy.

IBGP is responsible to assure connectivity to the "outside world" i. e. to other autonomous systems. That is, all packets entering this AS and were not blocked by policies should reach the proper exit BGP router. All transit routers inside the autonomous system should have a consistent view about the routing topology. Furthermore, IBGP routers must assure "synchronization" with the IGP, because packets cannot be continuously forwarded if the IGP routers have no idea about the route. Thus, IBGP routers must await the IGP convergence time inside the AS. Obviously this aspect assumes that BGP routes are injected to transit IGP routers by redistribution. The story with synchronization is explained a few slides later...

# Internal and External BGP



- **EBGP** messages are exchanged between peers of different ASs
  - ◆ EBGP peers **should** be **directly connected**
- Inside an AS this information is forwarded via **IBGP** to the next BGP router
  - ◆ IBGP messages have same structure like EBGP messages
- **Administrative Distance**
  - ◆ IBGP: **200**
  - ◆ EBGP: **20** (preferred over all IGP)

Some vendors including Cisco also allow EBGP peers to be logically linked over other hops inbetween. This "Multi-Hop" feature might introduce BGP-inconsistency and weakens the reliability as the BGP-TCP sessions cross other routers, so in practice a direct peering should be achieved.

Routing information learned by IBGP messages has much higher administrative distance than information learned by EBGP. Because of this, routes are preferred that do not cross the own autonomous system.

## Loop Detection



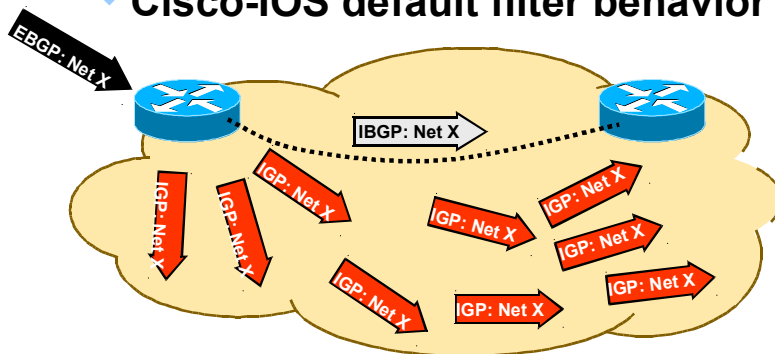
- **Update is only forwarded if own AS number is not already contained in AS\_Path**
- **Thus, routing loops are avoided easily**
- **But this principle doesn't work with IBGP updates (!)**
- **Therefore IBGP speaking routers must be fully meshed !!!**

For EBGP sessions loop-free topology is guaranteed by checking AS-Path, but it is not the case for IBGP sessions.

# BGP → IGP Redistribution



- Only routes learned via EBGP are redistributed into IGP
  - ◆ To assure optimal load distribution
  - ◆ Cisco-IOS default filter behavior



(C) Herbert Haas 2005/03/11

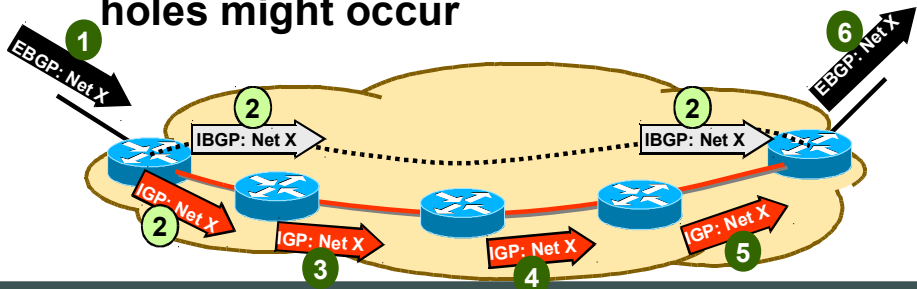
5

Routes learned via IBGP are never redistributed into IGP. This is the Cisco IOS "default filter" behavior. Obviously, if a router learned a route via IBGP, it is not a external (direct) peer for this route.

# Synchronization With IGP



- **Routes learned via IBGP may only be propagated via EBGP if same information has been also learned via IGP**
  - ◆ That is, same routes also found in routing table (= are really reachable)
- **Without this "IGP-Synchronization" black holes might occur**



(C) Herbert Haas 2005/03/11

6

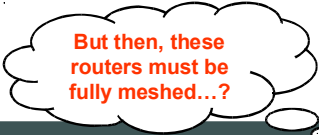
When a BGP router learns about an exterior network via an IBGP session, this router does not enter this route into its routing table nor propagates this route via EBGP because the IGP-transit routers might not be aware about this route and therefore convergence has not been occurred yet. The BGP router should propagate the learned route until this route has been entered into its routing table by IGP.

To understand this issue remember that BGP routing information is transported almost instantaneous between two BGP peers, while IGP updates might need quite a long time until reaching the other side of the AS. As illustrated in step 2 in the picture above, the IBGP message has been received by the BGP peer on the right border already, while the first IGP update (advertising the same network X) was injected by the left BGP peer and only reached the next IGP router at this time.

# Avoid Synchronization



- **Synchronization with IGP means injecting thousands of routes into IGP**
  - ◆ IGP might get overloaded
  - ◆ Synchronization dramatically affects BGP's convergence time
- **Alternatives**
  - ◆ Set default routes leading to BGP routers (might lead to suboptimal routing)
  - ◆ **Use only BGP-routers inside the AS !**



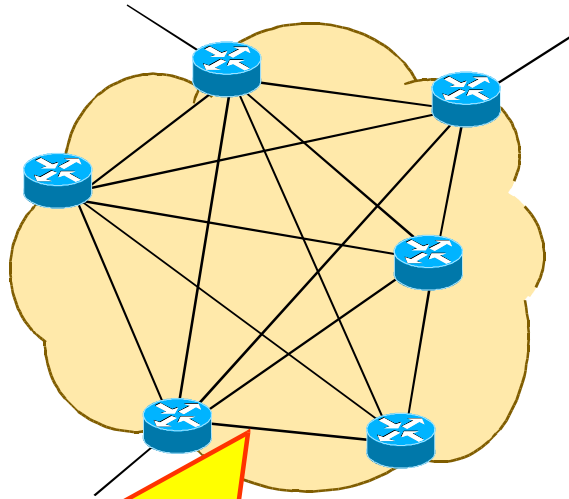
But then, these routers must be fully meshed...?

Synchronization is an old idea and leads to unwanted effects. First of all, most IGPs are not designed to carry a huge number of routes as needed in the Internet. Thus IGPs might get overloaded when ten thousands of external routes should be propagated in addition to the interior routes.

Furthermore, external routes are not needed inside an AS and typically a default route pointing to an BGP border router is sufficient (however this might lead to suboptimal routes as the default route might not be the best route). And finally, the consistency of the global BGP routing map would depend on the convergence of several (lots of) IGP routers – a situation that should be avoided!

Note that BGP injection into IGP and required BGP synchronization is not necessary if the AS is a transit AS only, such as many ISP networks. ISP networks have typically BGP routers only and thus need no synchronization. Fortunately many routers today (including Cisco routers) support the option to turn off synchronization.

# Fully Meshed IBGP Routers



- Does not scale
  - ◆  $n(n-1)/2$  links
- Resource and configuration challenge
- Solutions:
  - ◆ Route Reflectors
  - ◆ Confederations

**Note: These are logical IBGP connections!  
The physical topology might look different!**

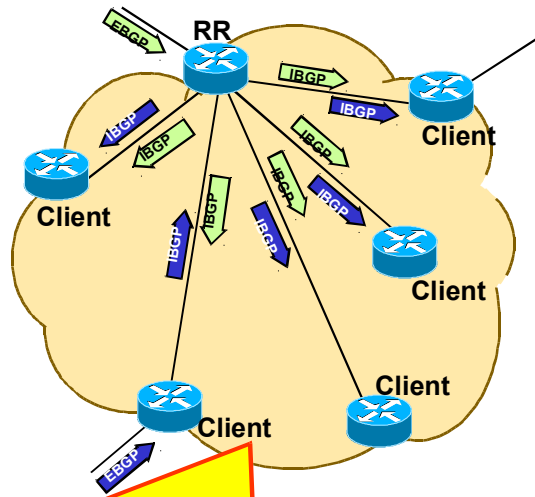
Every BGP router maintains IBGP sessions with all other internal BGP routers of an AS. Obviously, this fully meshed approach does not scale, especially it becomes a resource and manageability problem if the number of BGP sessions in one router exceeds 100.

Remember that each BGP session corresponds to a TCP connection, which requires a lot of system resources. Additionally BGP sessions must be manually established, so a fully meshed environment is also a configuration problem. This is also the reason, why BGP cannot replace traditional IGP in "normal" autonomous systems. ISPs demand for fast BGP convergence and do not need IGP in general.

Generally, there are two solutions to circumvent this problem: Route Reflectors and Confederations. Both techniques are discussed in the next slides.



# Route Reflector



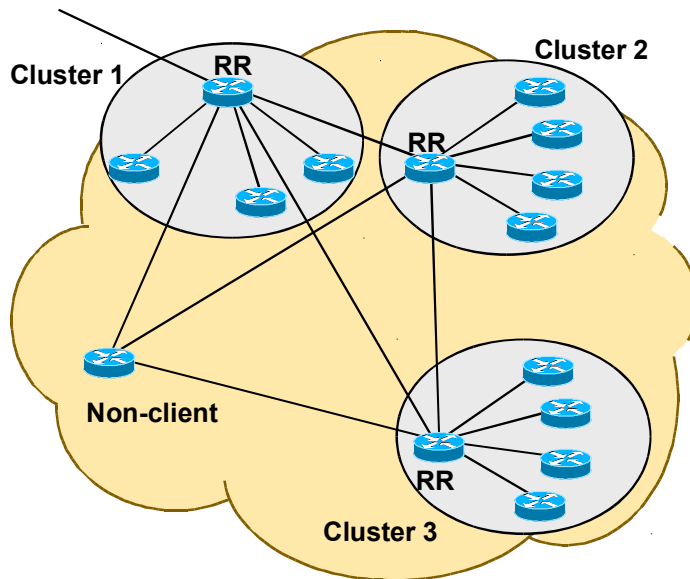
**Note:** Although these are logical IBGP connections, the physical topology should be the **main indicator** for an efficient cluster design (which router becomes RR)

- RR mirrors BGP messages for "clients"
- RR and clients belong to a "cluster"
- Only RR must be configured
  - ◆ Clients are not aware of the RR

Route reflectors are dedicated BGP routers that act like a mirror for IBGP messages. All BGP routers that peer with a RR are called "clients" and belong to a "cluster". Clients are normal BGP routers and have no special configuration – they have no awareness of a RR.

Using RRs there are only n-1 links.

# RR Clusters



- Only RRs are fully meshed
- Special Attributes care for loop-avoidance
- "Non-clients" must be fully meshed with RRs
  - ◆ And with other non-clients

Clients are considered as such because the RR lists them as clients.

# RR Issues



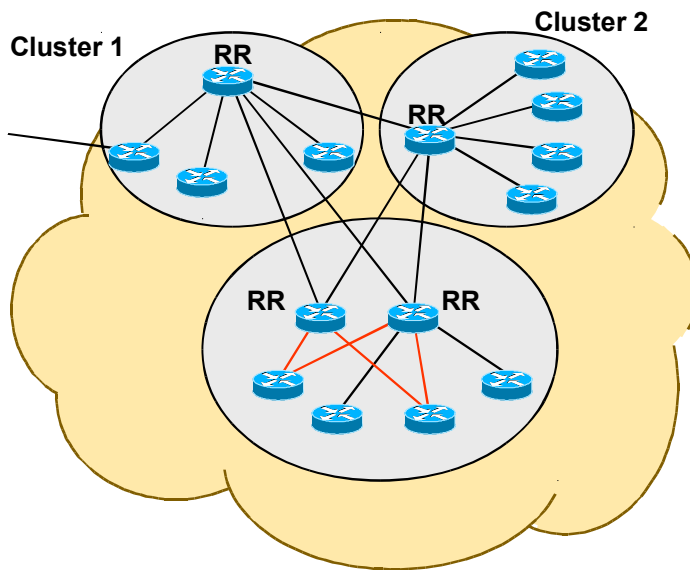
- RRs do **not change** IBGP behavior or attributes
- RRs only propagate **best routes**
- Special attributes to avoid routing updates **reentering** the cluster (routing loops)
  - ♦ **ORIGINATOR\_ID**  
Contains router-id of the route's originator in the local AS;  
attached by RR (Optional, Non-Trans.)
  - ♦ **CLUSTER\_LIST**  
Sequence of cluster-ids; RR appends own cluster-id when  
route is sent to non-clients outside the cluster  
(Optional, Non-Transitive)

It is important to know that RRs preserve IBGP attributes. Even the NEXT\_HOP remains the same, otherwise routing loops might occur. Imagine two clusters whose RRs are logically interconnected via IBGP but physically via clients. If one of these RRs learns about a NLRI from the other RR, this RR would reflect that information to its clients – also to that client who forwarded this NLRI information to this RR.

Obviously the NEXT\_HOP attribute must remain the same, that is pointing to the RR of the other cluster and not to the local RR, because there is no physical connection between the RRs.

If a RR learns the same NLRI from multiple client peers, only one path will be propagated to other peers. Therefore, when RRs are used, the number of path available to reach a given destination might be lower than that of a fully-meshed approach. Thus, suboptimal routing can only be avoided if the logical topology maps the physical topology as close as possible.

# Redundant RRs



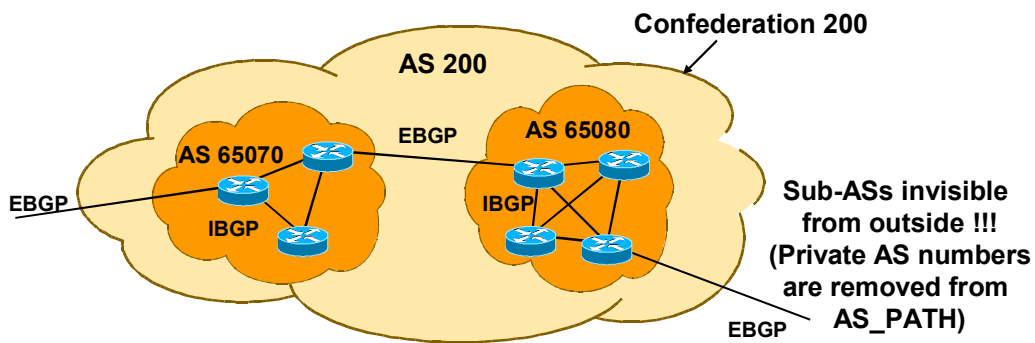
- RR is single point of failure
  - ◆ Other than fully meshed approach
- Redundant RRs can be configured
  - ◆ Clients attached to several RRs

Clients are considered as such because the RR lists them as clients.

# Confederations



- Alternative to route reflectors
- Idea: AS can be broken into multiple sub-ASs
- Loop-avoidance based on AS\_Path
- All BGP routers inside a sub-AS must be fully meshed
- EBGP is used between sub-ASs



(C) Herbert Haas 2005/03/11

13

Sub-ASs should utilize the private range of AS numbers (64512-65534).

# RRs versus Confederations



- **RRs are more popular**
  - ◆ **Simple migration** (only RRs needs to be configured accordingly)
  - ◆ Best scalability
- **Confederations drawbacks**
  - ◆ Introducing confederations require complete AS-renumbering inside an AS
  - ◆ Major change in logical topology
  - ◆ Suboptimal routing (Sub-ASs do not influence external AS\_PATH length)
- **Confederations benefits**
  - ◆ Can be used with RRs
  - ◆ Policies could be applied to route traffic between sub-ASs