BGP Advanced Topics

Internal versus External BGP

Route Reflectors, Confederations, Multiprotocol-BGP

Agenda

- <u>IBGP internals</u>
- Route reflectors
- Confederations
- Route servers
- IGP as BGP Transit
  - Introduction
  - OSPF Interaction
  - RIPv2 Interaction
- MP-BGP

Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

IBGP internals

- **main IBGP aspects inside an AS**
  - continuity
    - all packets entering the AS that were not blocked by some policies should reach the proper exit BGP router
    - all transit routers inside the AS should have a consistent view about the routing topology
  - synchronization
    - we should not cheat external partners by declaring that we know how to reach a destination, when we cannot really deliver packets to this destination
    - synchronization with the IGP can solve the packet drop problem

BGP Continuity 1

- **routes learned by internal peers are not advertised to other internal BGP peers**
  - prevents routing loops inside the AS
- **therefore a full IBGP mesh is necessary to keep the continuity of BGP updates**
  - every BGP router builds BGP sessions with all other internal BGP routers within an AS
  - this might cause manageability and resource problems if the number of BGP sessions in one router exceeds 100

BGP Continuity 2

- **however, the full mesh of IBGP sessions assures only that the BGP routers know the next hop IP address for all transit routes**
  - next hop IP address should be in the IP routing table to forward packets into the right direction
    - e.g., learned by IGP, configured by static route
- **if transit routers do not run BGP**
  - then they should know how to forward all transit packets to the proper border routers
- **so BGP routes should be injected into IGP**
  - to inform all transit routers about proper forwarding directions

BGP Continuity 3

- **remark:**
  - injecting (redistributing) BGP routes into IGP might cause resource problems
    - so it is not recommended in general
  - but then all transit routers within an AS should run a fully meshed IBGP
    - so all transit routers should know the destinations without using an IGP
- **if BGP routes are injected into IGP**
  - BGP synchronization is necessary

BGP Synchronization 1

- **BGP synchronization**
  - when a router receives a BGP update for a certain destination from an IBGP peer, the router verifies that destination is reachable before this route is advertised to an other EBGP peer
  - means that an IGP Update must be received for a route first before that IBGP learned route is passed on
- **packet drop problem without synchronization**
  - we get incoming traffic for the advertised destination but we cannot deliver, so packets are dropped

BGP synchronization 2

- **BGP synchronization is not necessary if**
  - the AS is not a transit AS
  - or this is the only transit AS for the transit destinations
  - or all transit routers participate in IBGP
- **synchronization causes IGP routing instabilities in the transit path to appear in the global internet**
  - possible solutions:
    - use topologies where synchronization is not necessary

IBGP Scalability

- **IBGP full meshed**
  - becomes a scalability issue with many border routers and many BGP routes
  - resource intensive
    - CPU, memory, bandwidth, configuration
- **several ways to solve full-mesh scalability problem**
  - BGP core with private AS´s
  - confederations (RFC 1965, experimental, Cisco)
  - route reflectors (RFC 1966, experimental, Cisco and RFC 2796, Proposed Standard)
  - route server (RFC 1863, experimental, Bay Networks)
  - new approach using MPLS in Transit-AS

Agenda

- **IBGP internals**
- **Route reflectors**
- **Confederations**
- **Route servers**
- **IGP as BGP Transit**
  - Introduction
  - OSPF Interaction
  - RIPv2 Interaction
- **MP-BGP**

BGP Route Reflectors 1

- **basic concept**
  - concentration router acts as focal point for internal BGP sessions
    - the other endpoint of an IBGP sessions is called client
  - multiple BGP routers can peer with a central point (route reflector, RR)
  - multiple route reflectors can peer together
- **naming conventions and operation rules**
  - clients together with their RR are called a cluster
  - all peers of RR which are not part of the cluster are non-clients
  - non-clients must be fully meshed with RR

BGP Route Reflectors 2

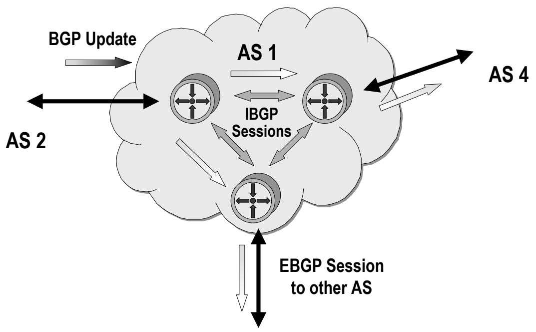
- RR function is implemented only on the route reflector
- clients and non-clients are normal BGP peers that have no notion of the route reflector
- any RR that receives multiple routes for the same destination will select the best path following usual BGP decision process
- the best path will be propagated inside the AS on the following rules:
  - if a route is received from a non-client, reflect it to clients only
  - if a route is received from a client, reflect it to all non-clients and also to other clients
  - if a route is received from an EBGP peer, reflect it to and clients and non-clients

© 2005, D.I. Manfred Lindner

BGP Route Reflectors 3

- **RR does not change IBGP behavior**
  - preserves attributes received from clients and non-clients
  - for loop avoidance inside an AS
    - additional Originator_ID and Cluster_List attributes were defined
- **Client multi-homing to redundant RR´s inside a cluster is possible**
  - however, all clients must be multi-homed in the same way
- **Hierarchical RR design is possible**
  - RR itself is a client for a higher RR level
  - a cluster cannot span over multiple levels

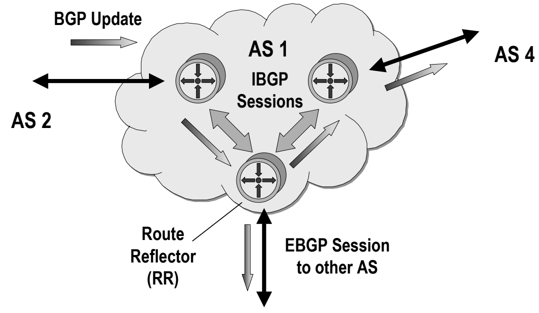
Example 1: Without Route Reflectors



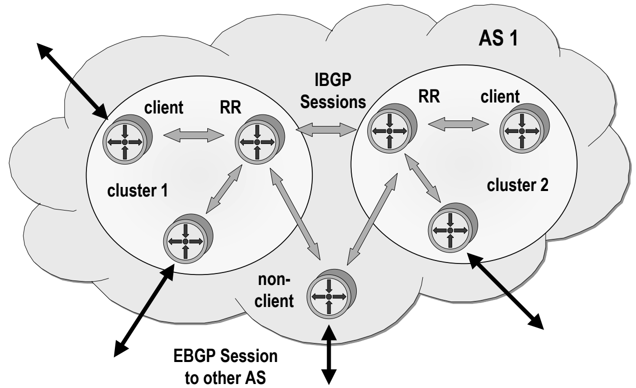
Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

Example 1: With Route Reflectors



Example 2: With Route Reflectors



© 2005, D.I. Manfred Lindner

Agenda

- **IBGP internals**
- **Route reflectors**
- **<u>Confederations</u>**
- **Route servers**
- **IGP as BGP Transit**
  - Introduction
  - OSPF Interaction
  - RIPv2 Interaction
- **MP-BGP**

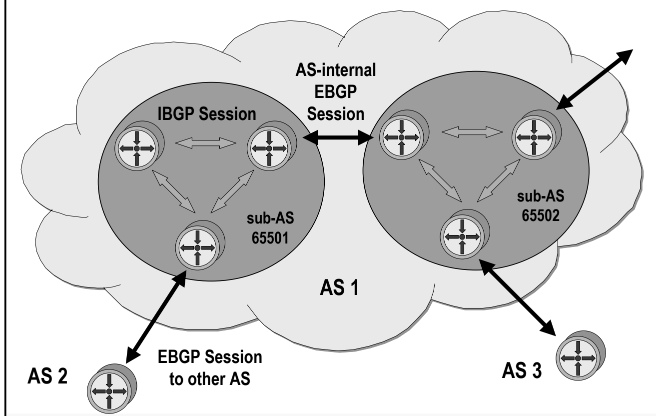
BGP Confederations

- **basic concept**
  - an AS is split into multiple sub-AS´s
  - every sub-AS is given a different private AS number
    - range 65412 - 65535
  - inside each sub-AS all the rule of IBGP apply
  - between sub-AS´s EBGP is used
  - although EBGP is used between sub-AS´s routing inside the confederation behaves like IBGP routing in a single AS
    - Next_Hop-, Local_Preference-, and MED-information is preserved when crossing sub-Ass boundaries
  - to the outside world a confederation looks like a single AS
    - private AS numbers will be removed from AS_Path when a route is advertised to the outside world

Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

Confederations Example



Agenda

- **IBGP internals**
- **Route reflectors**
- **Confederations**
- **<u>Route servers</u>**
- **IGP as BGP Transit**
  - Introduction
  - OSPF Interaction
  - RIPv2 Interaction
- **MP-BGP**

BGP Route Server

- **an approach to solve the IBGP full-mesh scalability problem developed by Bay Networks**
  - can be used additionally as exchange point for EBGP sessions (EBGP Route Server)
- **slightly different to Cisco´s Route Reflectors**
- **both are specified in RFCs classified experimental**
- **IETF is working to move one Router Server standard forward**

Agenda

Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

Concept of IGP as BGP Transit

- **in some cases you might be forced to use non-BGP routers in your BGP transit network**
  - multi-vendor environment, old models
  - IBGP scaling solutions are not available on intermediate routers
- **if the IGP support route tagging, then you might have a chance to convert BGP information into IGP route tags and back**
  - however, exporting and importing BGP information must be done in a consistent way in the whole routing domain
  - some limitations might exist on this conversion, but at least mandatory attributes should be converted and transited

Candidate IGP´s for BGP Transit

- **Route tagging is available in**
  - OSPF
  - RIPv2
- **Route tagging is not available in**
  - RIP
  - IGRP
  - EIGRP

Agenda

- **IBGP internals**
- **Route reflectors**
- **Confederations**
- **Route servers**
- **IGP as BGP Transit**
  - Introduction
  - OSPF Interaction
  - RIPv2 Interaction
- **MP-BGP**

BGP and OSPF Interaction

- **defined by RFC 1745**
- **provides a standard for translating BGP attributes to fields in OSPF and vice versa**
- **gives interaction rules and implementation guidelines**
  - when OSPF functionality and BGP functionality are used at the same time
    - OSPF as IGP
    - BGP-4 as EGP
- **natural point for interaction**
  - OSPF autonomous system boundary router (ASBR)

Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

BGP and OSPF Interaction

- **OSPF aspects**
  - internal, external type1, external type 2 routes
  - route aggregation
  - external routes can be tagged (tag field in external LSA)
    - tag is set when importing routes form external domains
    - tag will be used when exporting this routes to another domain using BGP
  - forwarding address to support redirection to other router than ASBR (forwarding address field in external LSA)
  - exporting rules: what and when route information should be given to BGP process
    - should be activated by explicit manual configuration
    - MED not taken from OSPF metric
    - Local_Preference not used

BGP and OSPF Interaction

- **BGP aspects**
  - exporting rules: what and when route information should be given to OSPF process
    - should be activated by explicit manual configuration
    - OSPF metric defaults for external type 2
    - never mirror IBGP information back into OSPF
    - handling of default network
    - MED not used
  - BGP identifier and OSPF router ID must be the same
  - OSPF tags used
    - to set Origin and Path Attributes in BGP

BGP and OSPF Interaction

- **Route tag in OSPF has 32 bits**
  - manual generation
    - use of the bits is not specified in the standard
    - export into BGP as origin EGP and only local AS in AS_PATH
  - automatic generation
    - use is specified in the standard
    - path information will be truncated to a maximum of 2 hops at re-export into BGP, so be careful about potential routing loops
      - typically in such cases the route is advertised with origin EGP to reflect the fact that the AS path was truncated
      - origin IGP is generated only if a full path information is really available

Agenda

- **IBGP internals**
- **Route reflectors**
- **Confederations**
- **Route servers**
- **IGP as BGP Transit**
  - Introduction
  - OSPF Interaction
  - RIPv2 Interaction
- **MP-BGP**

Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

BGP and RIPv2 Interaction

- **RIPv2 has some special information fields:**
  - routing domain - 16 bits
  - Address Family Identifier (AFI) - 16 bits
  - route tag - 16 bits
  - next hop - 32 bits
  - metric - 32 bits
- **these fields might be enough to map BGP information**
  - of course, with similar limitations as with OSPF

BGP and RIPv2 Interaction

- **from RFC2453 (4.2 Route Tag):**
  - "This allows for the possibility of a BGP-RIP protocol interactions document, which would describe methods for synchronizing routing in a transit network"
- **but this document did not born yet**
- **you might use your own scheme**
  - maybe something similar to the OSPF interaction
  - e.g. route tag = local or neighbor AS number
  - e.g. routing domain = flags for different types

Agenda

- **IBGP internals**
- **Route reflectors**
- **Confederations**
- **Route servers**
- **IGP as BGP Transit**
  - Introduction
  - OSPF Interaction
  - RIPv2 Interaction
- **MP-BGP**

Multiprotocol BGP 1

- **BGP-4 (RFC 1771) is capable of carrying routing information only for IPv4**
- **The only three pieces of information carried by BGP-4 that are IPv4 specific are**
  - the NEXT_HOP attribute (expressed as an IPv4 address),
  - the AGGREGATOR (contains an IPv4 address)
  - the NLRI (expressed as IPv4 address prefixes)
- **Multiprotocol Extensions to BGP-4**
  - RFC 2858
  - enable it to carry routing information for multiple network layer protocols (e.g., IPv6, IPX, etc...).

Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

Multiprotocol BGP 2

- **To enable BGP-4 to support routing for multiple network layer protocols two things have to be added**
  - the ability to associate a particular network layer protocol with the next hop information
  - the ability to associate a particular network layer protocol with a NLRI
- **To identify individual network layer protocols**
  - Address Family Identifiers (AFI) are used
  - values defined in RFC 1700
    - RFC 1700 is historic, obsoleted by RFC 3232
    - RFC 3232 specifies a Online Database for ASSIGNED NUMBERS
      - www.iana.org

Address Family Numbers (RFC 1700)

```
Number    Description
------    ------------------------------------------------------------
     0    Reserved
     1    IP (IP version 4)
     2    IP6 (IP version 6)
     3    NSAP
     4    HDLC (8-bit multidrop)
     5    BBN 1822
     6    802 (includes all 802 media plus Ethernet "canonical format")
     7    E.163
     8    E.164 (SMDS, Frame Relay, ATM)
     9    F.69 (Telex)
    10    X.121 (X.25, Frame Relay)
    11    IPX
    12    AppleTalk
    13    Decnet IV
    14    Banyan Vines
 65535    Reserved
```

Multiprotocol BGP 4

- **Address Family Identifier (AFI) in MP-BGP**
  - this parameter is used to differentiate routing updates of different protocols carried across the same BGP session
  - it is a 16-bit value
- **MP-BGP uses an additional Sub-Address Family Identifier (SAFI)**
  - it is a 8-bit value
    - 1 NLRI used for unicast forwarding
    - 2 NLRI used for multicast forwarding
    - 3 NLRI used for both unicast and multicast forwarding
- **Usual notation AFI/SAFI (i.e. x/y)**
    - 1/1 IP version 4 unicast
    - 1/2 IP version 4 multicast
    - 1/128 VPN-IPv4 unicast (used for MPLS-VPN)

Multiprotocol BGP 3

- **Capability Advertisement Procedures are used**
  - by a BGP speaker that to determine whether the speaker could use multiprotocol extensions with a particular peer or not -> RFC 3392
  - done during BGP Open with Capabilities Optional Parameter (Parameter Type 2)

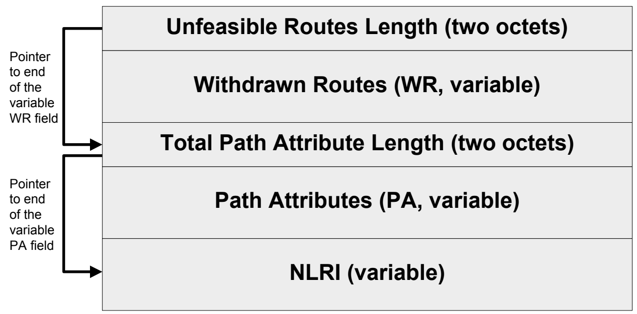
```
+------------------------------+
| Capability Code (1 octet)    |
+------------------------------+
| Capability Length (1 octet)  |
+------------------------------+
| Capability Value (variable)  |
+------------------------------+
```

    - Capability Code is unambiguously identifies individual capabilities. Capability Value is interpreted according to the value of the Capability Code field.

Multiprotocol BGP 4

- **Two new attributes**
  - Multiprotocol Reachable NLRI (MP_REACH_NLRI)
  - Multiprotocol Unreachable NLRI (MP_UNREACH_NLRI)
- **MP_REACH_NLRI is used**
  - to carry the set of reachable destinations together with the next hop information to be used for forwarding to these destinations
- **MP_UNREACH_NLRI is used**
  - to carry the set of unreachable destinations
- **Both of these attributes**
  - are optional and non-transitive

BGP Update Message Format for IPv4



BGP Update Message Details for IPv4

- **NLRI**
  - 2-tuples of (length, prefix)
    - length = number of masking bits (1 octet)
    - prefix = IP address prefix (1 - 4 octets)
    - note: prefix field contains only necessary bits to completely specify the IP address followed by enough trailing bits to make the end of the field fall on an octet boundary
- **path attributes are composed of**
  - triples of (type, length, value) -> TLV notation
    - attribute type (two octets)
      - 8 bit attribute flags, 8 bit attribute type code
    - attribute length (one or two octets)
      - signaled by attribute flag-bit nr.4
    - attribute value (variable length)
      - content depends on meaning signaled by attribute type code

IPv4 Path Attribute Format / NLRI Format

Attribute Type

1 octet	1 octet	1-2 octet
Attr. Flags	Attr. Type Code	Attribute Length
Attribute Value (variable octets)		

Path Attribute Format

1 octet	
Length	Prefix (1 - 4 octets)

NLRI

Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

VPN-IPv4 BGP Update with MP_Reach_NLRI

1 octet	1 octet	1 octet
Attr. Flags	Type Code = 14	Attribute Length
AFI= 1		SAFI = 128
Length of NHA		
Next Hop Address (NHA, 1- 4 octets)		

Path Attribute MP_Reach_NLRI

1 octet
Length = 120
Label (3 octets = 24 bits)
Route Distinguisher (8 octets = 64 bits)
IPv4 address (4 octets = 32 bits)

NLRI for VPN-IPv4

BGP Draft Attributes 1

- **BGP Extended Communities Attribute**
  - consists of a set of "extended communities”
    - optional transitive; type 16
    - defined in draft-ietf-idr-bgp-ext-communities-07.txt
  - two important enhancements over the existing BGP Community Attribute:
    - it provides an extended range, ensuring that communities can be assigned for a plethora of uses, without fear of overlap.
    - the addition of a type field provides structure for the community space.
  - Important for MPLS_VPN
    - **Route Target Community**
    - **Route Origin Community**

BGP Draft Attributes 2

- **Route Target:**
  - The Route Target Community identifies one or more routers that may receive a set of routes (that carry this Community) carried by BGP. This is transitive across the Autonomous system boundary.
  - It really identifies only a set of sites which will be able to use the route, without prejudice to whether those sites constitute what might intuitively be called a VPN.
- **Route Origin:**
  - The Route Origin Community identifies one or more routers that inject a set of routes (that carry this Community) into BGP. This is transitive across the Autonomous system boundary.

BGP Draft Attributes 3

- **Route Target and Router Origin**
  - type: 2 octets (extended form of this attribute)
    - high octet -> 00, 01, 02 -> defines the structure of the value field
    - low octet -> defines the actual type
  - value: 6 octets
- **Route Target:**
  - high octet type: 0x00 or 0x01 or 0x02
  - low octet type: 0x02
- **Route Origin:**
  - high octet type: 0x00 or 0x01 or 0x02
  - low octet type: 0x03

Institute of Computer Technology - Vienna University of Technology

L49 - BGP Advanced Topics

BGP Draft Attributes 4

- **Structure of value field based on high octet part of type**
  - 0x00:
    - 2 octets Global Administrator Field (IANA assigned AS #)
    - 4 octets Local Administrator Field (actual value of given type contained in low octet part of type)
  - 0x01:
    - 4 octets Global Administrator Field (IP address assigned by IANA)
    - 2 octets Local Administrator Field
  - 0x02:
    - 4 octets Global Administrator Field (IANA assigned 4 octet AS #)
    - 2 octets Local Administrator Field

© 2005, D.I. Manfred Lindner