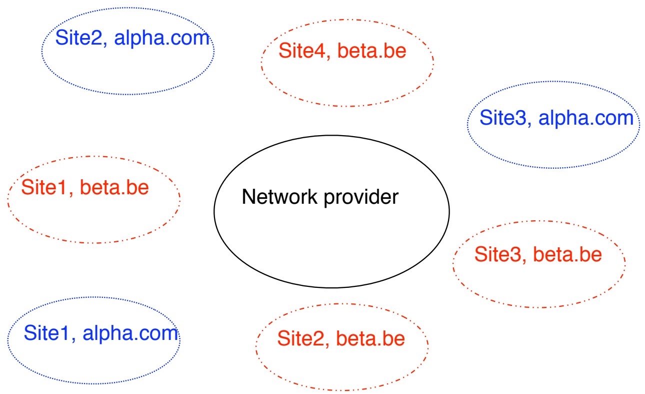
Computer Networks : Protocols and Practice

Part 11: Virtual Private Networks

Olivier Bonaventure
http://inl.info.ucl.ac.be/

The VPN problem

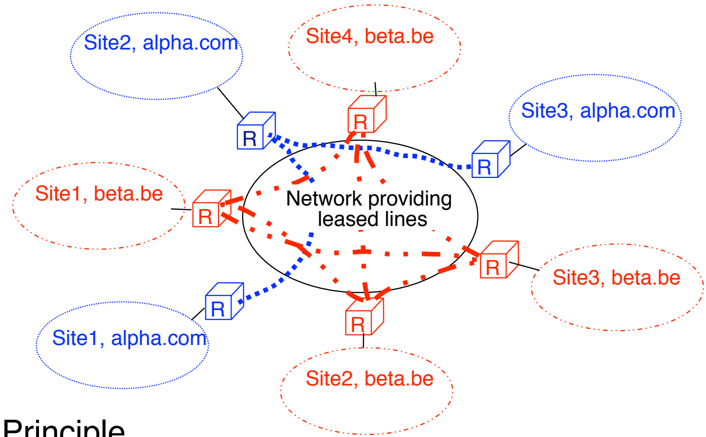


- How to efficiently create
  - one network containing the sites from alpha.com
  - one network containing the sites from beta.be

What should be the goal of a good VPN ?

- A good VPN service should
  - Support multiple corporate customers
    - in this case, the traffic from these customers should be isolated
    - some security features should be supported to ensure that packets from public Internet can be introduced inside VPN
  - provide QoS guarantees for corporate customers
    - typical solution is to reuse the classical mechanisms
  - be easy to use and manage
    - from the customer viewpoint
    - from the service provider viewpoint

The classical solution

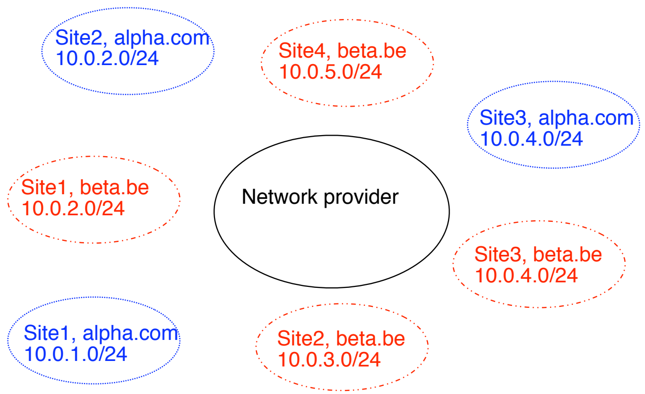


- Principle
  - Create leased lines between sites
    - full mesh (beta.be), hub and spoke (alpha.com) topologies

Evaluation of the classical solution

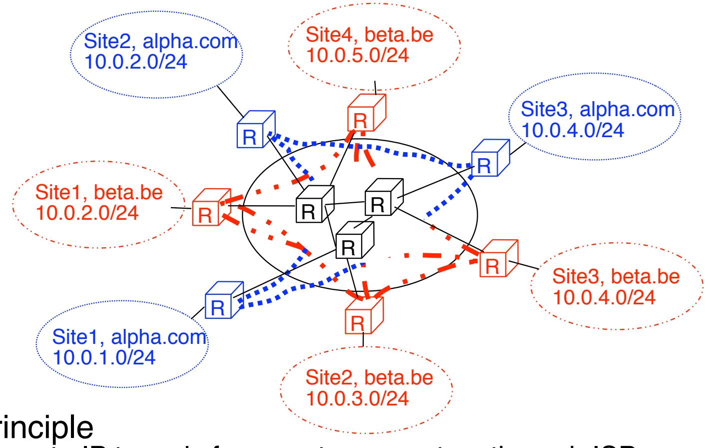
- Advantage
  - the quality of the service provided by the service provider is usually very good
- Drawbacks
  - the number of leased lines can be high
    - n*(n-1)/2 leased lines in total for full mesh
    - For a VPN with n sites, each router needs n-1 interfaces to obtain a full mesh
  - Flexibility
    - addition of a VPN may require several new lines
    - installation of leased line may require O(months)
  - cost can be high
    - no statistical multiplexing on provider's backbone
    - link costs even if no traffic is exchanged

The IP-VPN problem



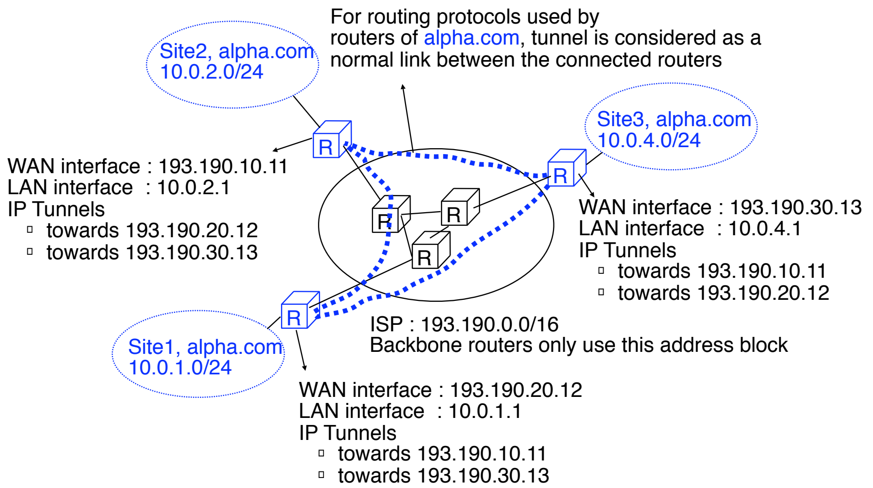
- How to efficiently create
  - one network containing the sites from alpha.com
  - one network containing the sites from beta.be

A customer-provisioned IP VPN



- Principle
  - create IP tunnels from customer routers through ISP
  - drawback : configuration burden on customer routers

A customer-provisioned IP-VPN (2)

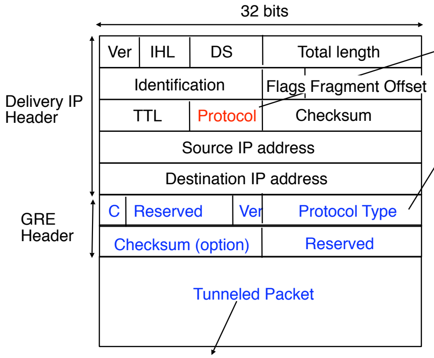


IP Tunnels

- Many IP tunnelling protocols exist
  - IP in IP tunnelling
    - can be used to carry IP packets inside IP packets
  - Generic Routing Encapsulation
    - can be used to carry network layer packets inside IP packets
  - Point-to-point tunnelling protocol
    - can be used to carry PPP frames through TCP/IP network
  - Layer 2 Tunnelling protocol
    - can be used to carry PPP frames through TCP/IP network
  - IPSec
    - security (authentication/confidentiality) extensions to IP also include tunnelling capabilities

GRE Tunnel

- Principle
  - Tunnel is used to carry network layer packets



IP Protocol
Value 47 is reserved to indicate that IP packet carries GRE-encapsulated packet

Protocol Type
Indicates the type of network layer packet carried by tunnel
same values as Ethernet type field (0x800 for IP packet)

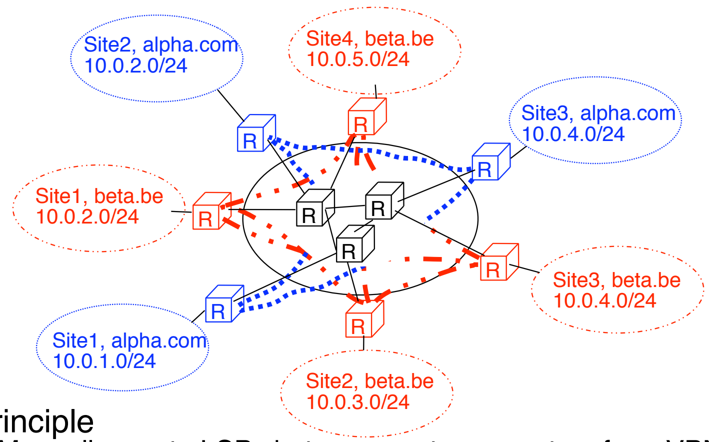
Tunneled packet may be optionally protected by Checksum in GRE header

Can contain any network layer packet understood by destination system that can be placed inside Ethernet frame

Evaluation of the simple IP solution

- Advantage
  - Flexibility
    - a single physical interface on each router
  - Cost
    - VPN site can multiplex traffic to different sites on this link
- Drawbacks
  - the number of tunnels can be high
    - n*(n-1)/2 tunnels in total for full mesh
    - For a VPN with n sites, each router needs n-1 tunnels to obtain a full mesh
  - Flexibility
    - addition of a VPN require adding new tunnels
  - Security
    - depends on tunnelling mechanism used
      - weak with GRE, better with IPsec

A simple MPLS-based solution



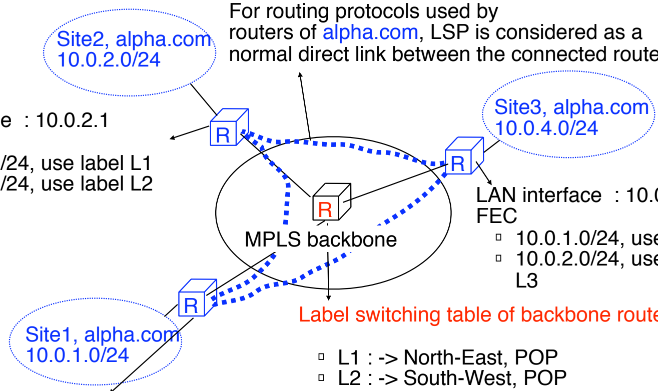
- Principle
  - Manually create LSPs between customer routers from VPN sites through MPLS backbone

This simple MPLS-based solution is similar in principle to the solution used to support VPN with technologies based on the label switching paradigm like

- ATM : Asynchronous Transfer Mode
- Frame Relay

A simple MPLS-based solution (2)

For routing protocols used by routers of alpha.com, LSP is considered as a normal direct link between the connected routers



LAN interface : 10.0.2.1
FEC
- 10.0.4.0/24, use label L1
- 10.0.1.0/24, use label L2

LAN interface : 10.0.4.1
FEC
- 10.0.1.0/24, use label L6
- 10.0.2.0/24, use label L3

LAN interface : 10.0.1.1
FEC
- 10.0.4.0/24, use label L4
- 10.0.2.0/24, use label L5

Label switching table of backbone router

- L1 : -> North-East, POP
- L2 : -> South-West, POP
- L3 : -> North-West, POP
- L4 : -> North-East, POP
- L5 : -> North-West, POP
- L6 : -> South-West, POP

Evaluation of the simple MPLS solution

- Advantages
  - a single physical line per VPN site
  - QoS can be provided on a per-LSP basis
  - Flexibility
    - bandwidth of each LSP can be easily updated
  - Cost
    - statistical multiplexing is possible on MPLS backbone
- Drawbacks
  - MPLS support
    - routers of the VPN sites must support MPLS
    - backbone routers must support MPLS
  - configuration burden
    - backbone routers must be configured for each new LSP
    - customer routers must be configured for each new site

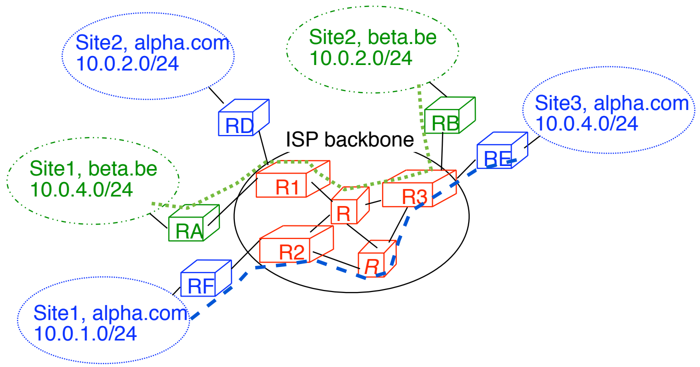
Provider-provisioned MPLS VPN

- Objective
  - Find a solution that is as automatic as possible
    - for the service provider
    - for the customers of the VPN service
  - Addition of a new site to an existing VPN
    - only the new customer router should need to be configured on the VPN
    - only a single router from the service provider should need to be configured on the provider's backbone

The provider-provisionned MPLS VPNs are defined in
RFC2547 BGP/MPLS VPNs. E. Rosen, Y. Rekhter. March 1999.

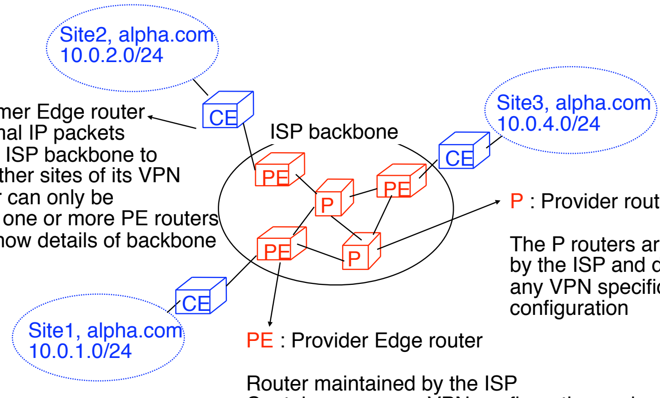
Provider-provisioned MPLS VPN (2)

- Principle of the solution



- transmission of one packet in beta.be, site1to site2
- transmission of one packet in alpha.com, site1to site3

Provider-based MPLS solution (3)



CE : Customer Edge router
Sends normal IP packets through the ISP backbone to reach the other sites of its VPN
A CE router can only be attached to one or more PE routers
Does not know details of backbone

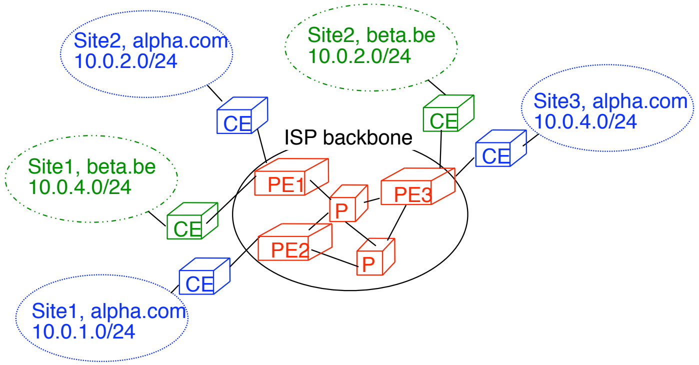
P : Provider router

The P routers are managed by the ISP and do not carry any VPN specific configuration

PE : Provider Edge router

Router maintained by the ISP
Contains some per-VPN configuration and ensures that the IP packets sent by a particular VPN site are delivered to the PE router attached to the destination VPN site

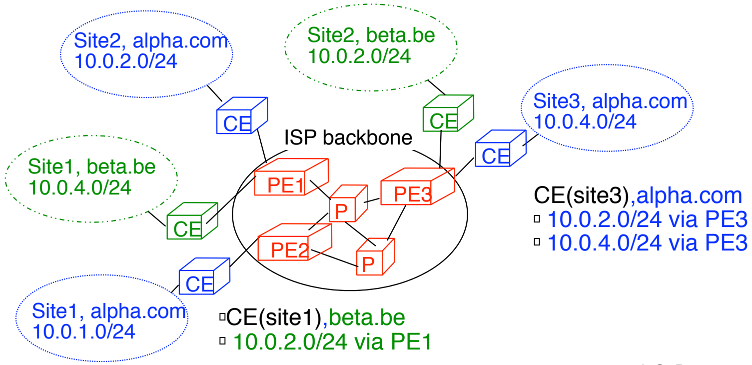
Problems to solve



- How to forward the packets from one CE router to the appropriate CE router of the same VPN ?
  - Need routing tables on CE, PE and P routers
  - How to efficiently distribute these routing tables ?

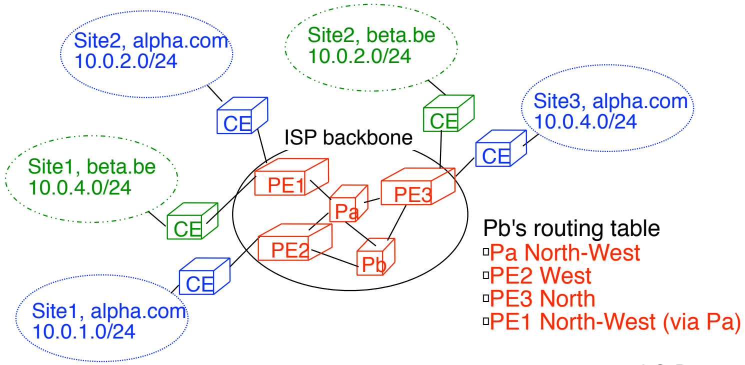
Routing tables on the CE routers

- Principle
  - Each CE router contains one routing table with the routes belonging to its VPN
    - CE does not know anything about ISP besides its attached PE



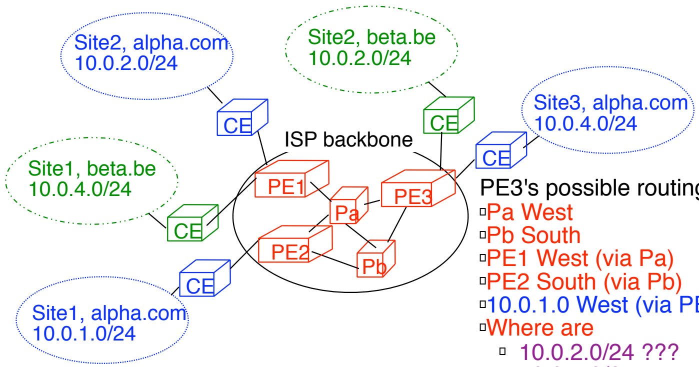
Routing tables on the P routers

- Principle
  - P routers only know how to reach the routers in their backbone
    - P routers do not know anything about VPNs



Routing tables on the PE routers

- Problem
  - Corporate networks often use RFC1918 addresses
  - Two different VPNs may use same IP subnets

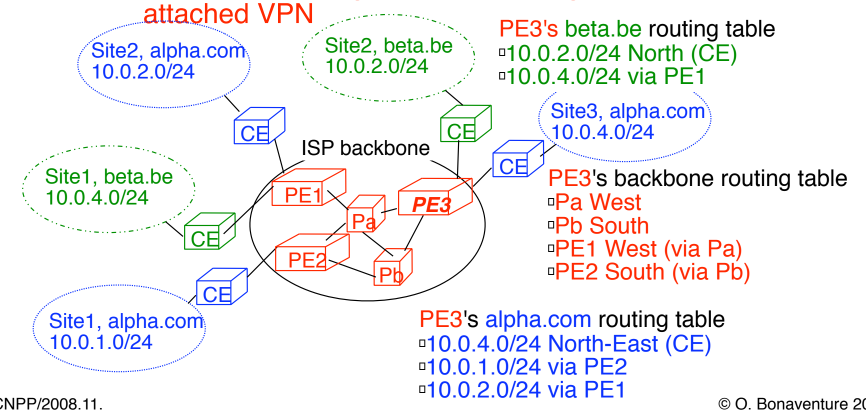


PE3's possible routing table

- Pa West
- Pb South
- PE1 West (via Pa)
- PE2 South (via Pb)
- 10.0.1.0 West (via PE2)
- Where are
  - 10.0.2.0/24 ???
  - 10.0.4.0/24

Routing tables on PE routers (2)

- Principle
  - Each PE router maintains several routing tables
    - standard routing table
    - one **VPN Routing and Forwarding table (VRF)** per attached VPN



The VRF contains all the routes belonging to a given VPN. This VRF is used to forward the packets that are received inside the corresponding VPN. For example, when considering PE3, it will use the beta.be VRF to forward a packet received on its North interface while it will use the alpha.com VRF to forward a packet received on its Nort-East interface.

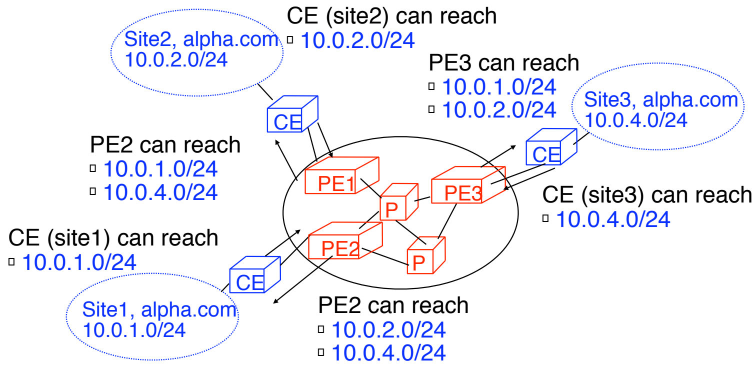
Distribution of the routing tables

- Routing problem
  - How can we correctly distribute the routing information to the CE, PE and P routers ?
    - A CE router must advertise its local routes to its attached PE and must receive the remote routes (or a default route) from this router
    - A PE router must receive two types of routing information
      - per VPN routing information for the routes reachable through attached CE routers and through remote PE routers
        - For scalability reasons, a PE router should only know the routing information about the VPNs that it directly supports
      - ISP routing information to be able to reach other PE routers
    - A P router must maintain routing information for the ISP
      - For scalability reasons, a P router should not know any VPN specific information

Distribution of routing information(2)

- Route distribution between CE and PE
  - static routes
    - both PE and CE are configured with static routes
    - suitable for small VPN sites with a single link
  - RIP
    - RIP is used by the CE to announce the routes reachable on its local network
    - RIP is used by the PE to announce the routes of the same VPN learned from the other PE routers
    - useful for medium VPN sites with multiple links
  - Other routing protocols
    - OSPF
      - This is a special OSPF instance between PE and CE, not the OSPF that is used inside the ISP backbone
    - eBGP
      - CE router uses eBGP session to advertise routes to PE

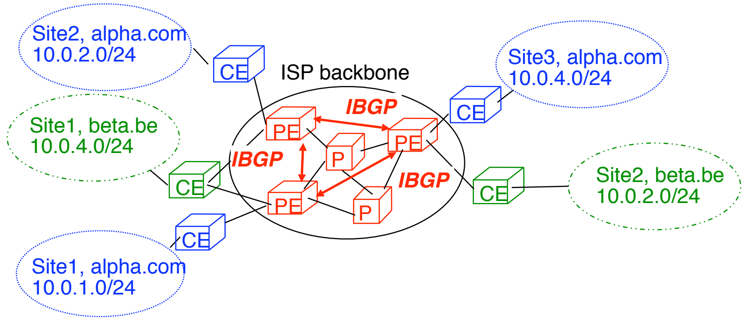
Distribution of routing information(3)



- In the backbone, all P and PE routers know the backbone topology by using OSPF

Distribution of routing information (4)

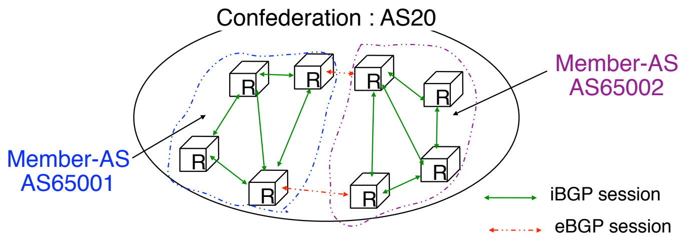
- Distribution of per VPN routes between PE



  - Principle
    - iBGP full mesh between PE routers
      - P routers do not need to run iBGP since they do not maintain per-VPN routes
    - iBGP sessions are used to redistribute the routes learned from CE routers to distant PE routers

How to scale iBGP in large domains ?

- Confederations
  - Divide the large domain in smaller sub-domains
    - Use iBGP full mesh inside each sub-domain
    - Use eBGP between sub-domains



  - Each router is configured with two AS numbers
    - Its confederation AS number
    - Its Member-AS AS number
  - Usually, a single IGP covers the entire domain

CNPP/2008.11.

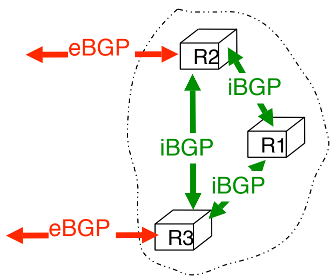
BGP confederations are discussed in :

P. Traina, D. McPherson, J. Scudder, "Autonomous System Confederations for BGP", RFC 3065, February 2001.

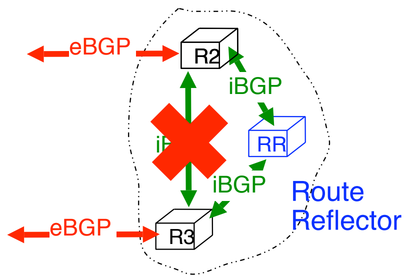
Route reflectors
An alternative to confederations

- Route reflectors
  - A route reflector is a special router that is allowed to propagate the routes learned over iBGP sessions on other iBGP sessions

Normal iBGP full mesh



iBGP with one route reflector

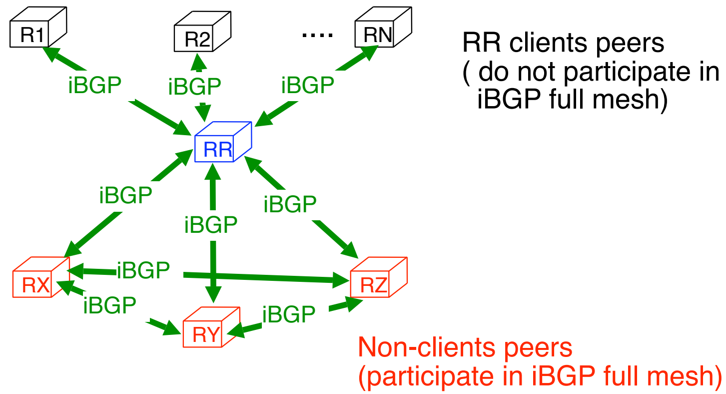


Route reflectors are defined in :

T. Bates, R. Chandra, E. Chen, "BGP Route Reflection - An Alternative to Full Mesh iBGP", RFC 2796, April 2000.

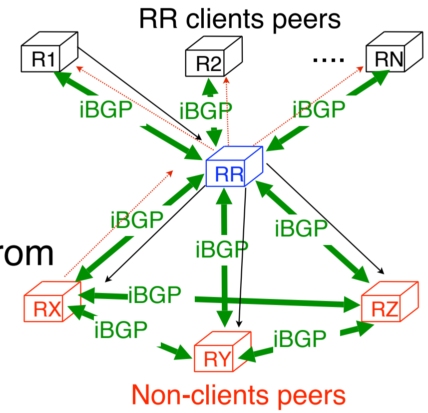
Behaviour of a Route Reflector

- Two types of iBGP peers of a route reflector



Behaviour of a Route Reflector

- Route received from an eBGP session or a client peer
  - Select best path
  - Advertise to
    - All client peers
    - All non-client peers

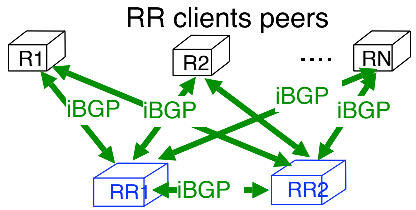


- Route received from non-client peer
  - Select best path
  - Advertise to :
    - All client peers

It should be noted that when a route reflector advertises its best path to client or non-client peers, it does not change the nexthop of the advertised route.

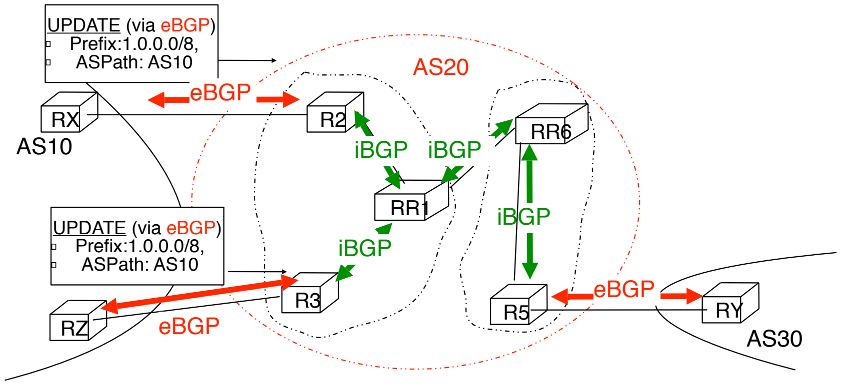
Fault tolerance of route reflectors

- How to avoid having the RR as a single point of failure ?
  - Solution
    - Allow each client peer to be connected at 2 RRs



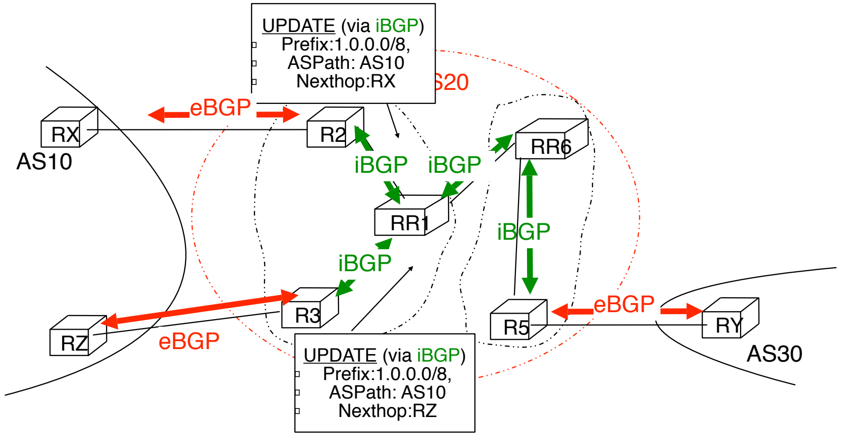
  - Issue
    - Configuration errors may cause redistribution loops
      - ORIGINATOR_ID used to carry router ID of originator of route
      - CLUSTER_LIST contains the list of RR that sent the UPDATE message inside the current AS

Route reflectors : an example



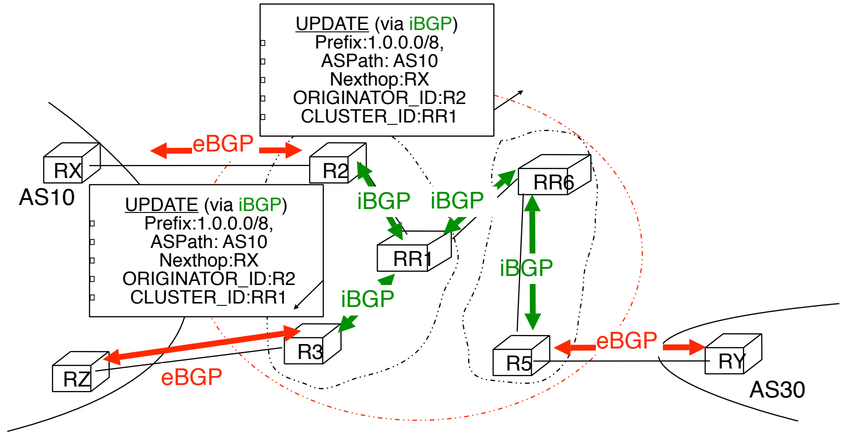
- R2 and R3 are clients of Route Reflector RR1
- RR1 and RR6 are in iBGP full mesh
- R5 is client of Route Reflector RR6

Route reflectors : an example (2)



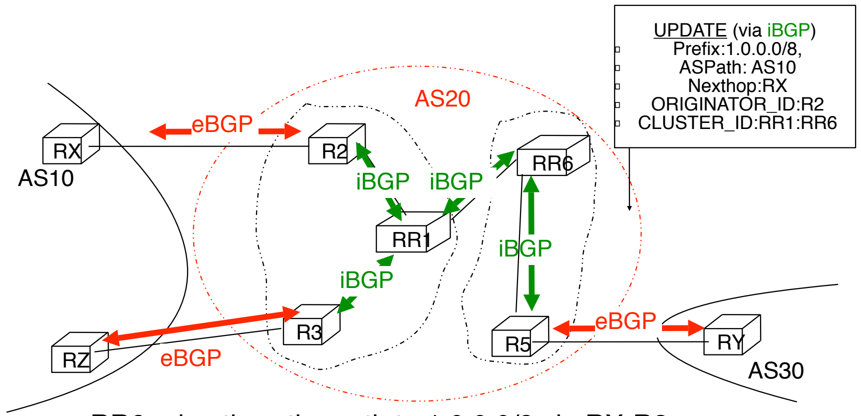
- RR1 will select its best path towards 1.0.0.0/8 and will re-advertise it by adding the ORIGINATOR_ID and the CLUSTERID

Route reflectors : an example (3)



- RR1 prefers the path to 1.0.0.0/8 via RX-R2
  - RR1 advertises this path to its client peer (R3)
    - the path is not advertised to R2 since R2 already received it
  - RR1 advertises this path to its non-client peer (RR6)

Route reflectors : an example (4)



- RR6 advertises the path to 1.0.0.0/8 via RX-R2
  - to its client peer R5
- R5 will remove ORIGINATOR_ID and CLUSTER_ID before advertising the path to RY via eBGP

Confederations versus Route reflectors

- Confederations
  - Solves iBGP scaling
  - Redundancy with iBGP full-mesh inside each MemberAS
  - Possible to run one IGP per Member AS
  - Requires manual router configuration
  - Can be used when merging domains
  - Can lead to some routing oscillations
- Route reflectors
  - Solves iBGP scaling
  - Redundancy by using Redundant RRs
  - Usually a single IGP for the whole AS
  - Requires manual router configuration
  - Can lead to some routing oscillations

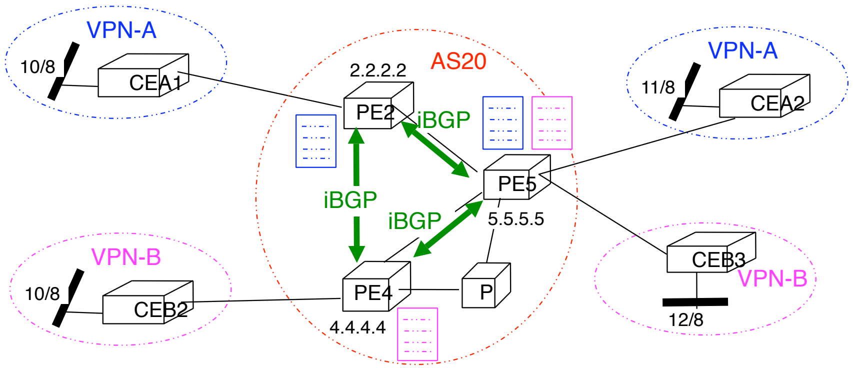
Note that besides route reflectors and confederations, some companies are developing proprietary solutions to solve the iBGP full mesh problem.

See e.g.

V. Jacobson, C. Alaettinoglu, and K. Poduri, BST - BGP Scalable Transport, NANOG26, October 2002, http://www.nanog.org/mtg-0302/bst.html

The distribution of the VPN routes by the PE routers

- Two issues
  - How to distribute the A and B routes for 10/8 ?
  - How to ensure that PE4 only receives B routes ?



MP-BGP and the VPN-IPv4 addresses

- MultiProtocol-BGP
  - an extension to BGP that allows a BGP router to advertise non-IPv4 routes
    - IPv6
    - IP multicast
    - VPN-IPv4
- The VPN-IPv4 address family
  - a method used by PE routers to encode IP v4 VPN addresses before advertising them with MP-BGP
    - a VPN-IPv4 address contains
      - an 8 bytes route distinguisher
      - an IPv4 prefix
    - BGP considers VPN-IPv4 addresses as *opaque bitstring*
    - two types of route distinguishers
      - `AS:value`
      - `IPaddress:value`

Controlling the distribution of VPN routes

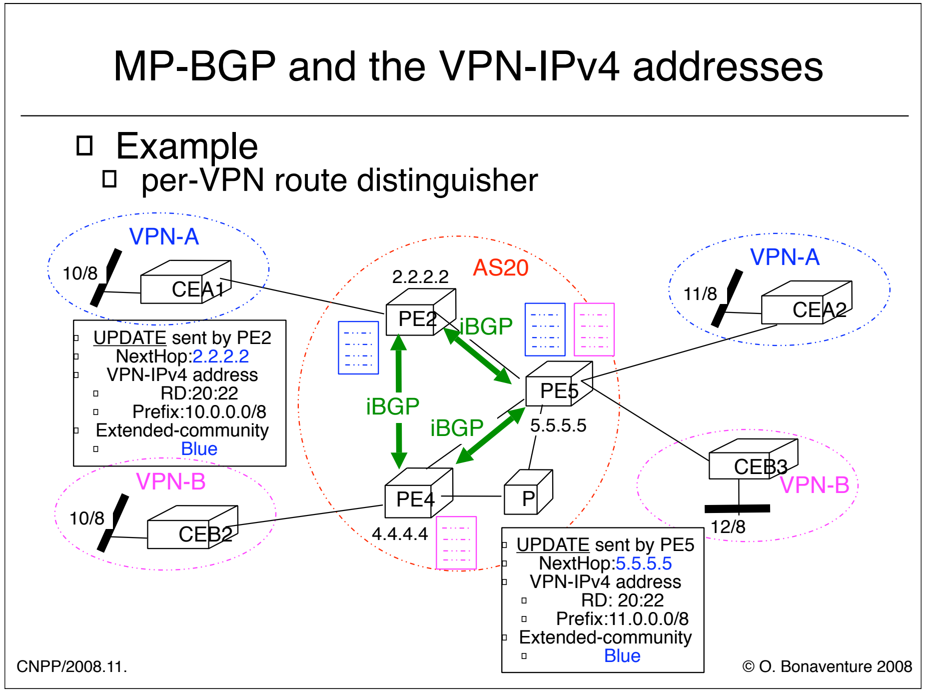
- How to ensure that VPN-IPv4 routes only reach the PE routers attached to those VPNs ?
  - associate one or more route targets to each VRF
  - a route associated with RT x must be distributed to all PE routers that have a VRF with RT=x
  - RT is encoded as an BGP extended community
    - `ASnumber:value`
    - `IPv4address:value`
  - Control of the distribution
    - PE router knows the RT supported by each of its peers and only advertises the appropriate VPN-IPv4 routes
    - or PE router advertises all its VPN-IPv4 routes and peers filter the received routes based on the attached RT

The BGP Extended Community attribute is defined in :

Sangli, Tappan and Rekhter, “BGP Extended Communities Attribute”, Internet draft, draft-ietf-idr-bgp-ext-communities-06.txt, work in progress, Aug. 2003

Compared to the classical communities, the main advantage of the extended communities is their size. The classical communities are 32-bits wide, and a block of 2^{16} values is allocated to each AS (ASX:1 to ASX:65535). If the communities were used to support VPNs, an AS could only define 2^{16} route target values. With extended communities, each AS can define 2^{32} different route target values.

The cooperative route filtering mechanism that can be used by PE router to advertise to their peers the routes that they wish to receive is defined in :
Chen, Rekhter, “Cooperative Route Filtering Capability for BGP-4”, Internet draft, draft-ietf-idr-route-filter-09.txt, work in progress, August 2003



An additional element of the RFC2457 architecture that does not appear in the slides is that each PE router defines, for each VPN attached to the router:

- an import policy to specifiy, which routes received via BGP or the PE-CE protocol can be installed in the VRF
- an export policy to specify which routes installed in the VRF need to be advertised by using the PE-CE protocol or BGP

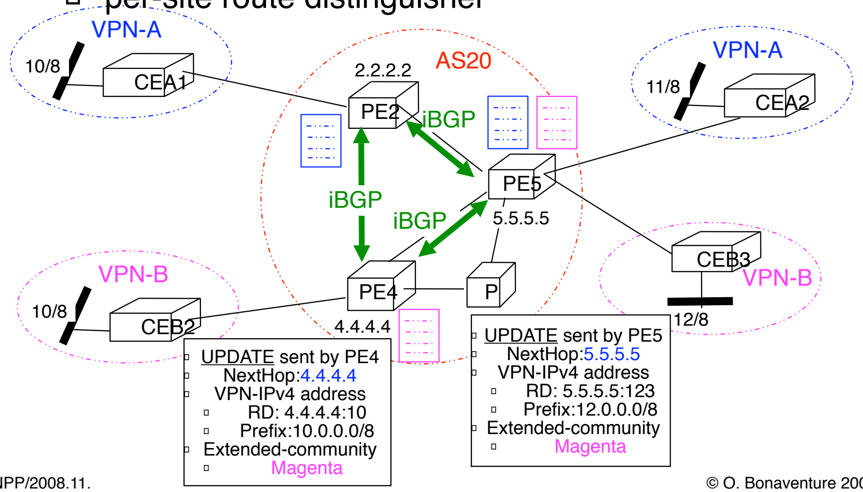
Of course, those policies will depend on the route distinguishers and the route targets being used.

In this example, the following import filters and import policies will be used

- PE5 imports the iBGP advertisements with extended communities blue and magenta since it has a CE route of VPNA and VPNB attached
  - The routes with RD 20:222 that are received by PE5 are placed in its VPN-A VRF
- PE4 does not import the BGP advertisements that carry the Blue extended community since no CE router of VPNA is attached to PE4

MP-BGP and the VPN-IPv4 addresses (2)

- Example
  - per-site route distinguisher

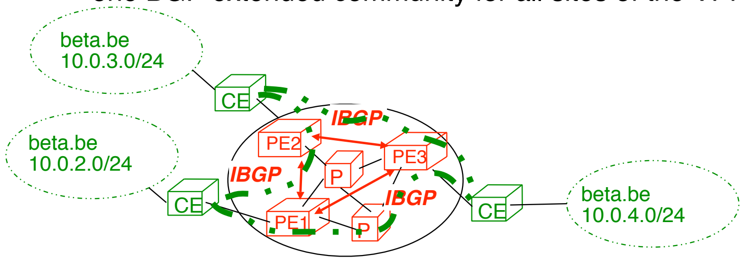


In this example, the following import filters and import policies will be used

- PE5 imports the iBGP advertisements with extended communities blue and magenta since it has a CE route of VPNA and VPNB attached
  - The routes with RD 4.4.4.4:10 that are received by PE5 are placed in its VPN-B VRF
- PE2 does not import the BGP advertisements that carry the Magenta extended community since no CE router of VPNB is attached to PE2

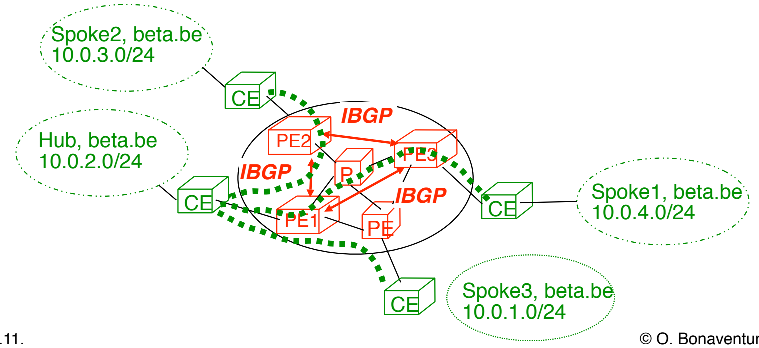
Types of VPN connectivity

- Utilisation of the BGP extended community attribute
  - depends on the type of inter-sites connectivity within each supported VPN
- Full mesh connectivity
  - all sites are equal
  - one BGP extended community for all sites of the VPN



Types of VPN connectivity (2)

- Hub & spoke connectivity
  - two types of sites
    - large (hub) site sends to all
    - small (spoke) sites use hub as relay site to reach others
  - one BGP extended community for Hub
  - one BGP extended community for all spoke sites

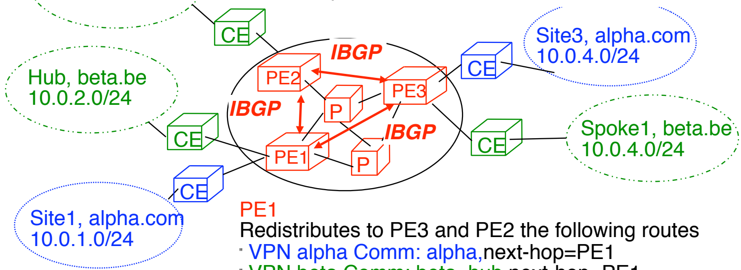


Types of VPN connectivity (3)

PE2
- Redistributes : VPN beta Comm: beta_spoke,next-hop=PE2
- Imports routes with community : beta_hub

PE3
- Redistributes
  - VPN betaComm: beta_spoke,next-hop=PE3
  - VPN alpha,Comm: alpha,next-hop=PE3

Imports the routes with communities : beta_hub, alpha



PE1

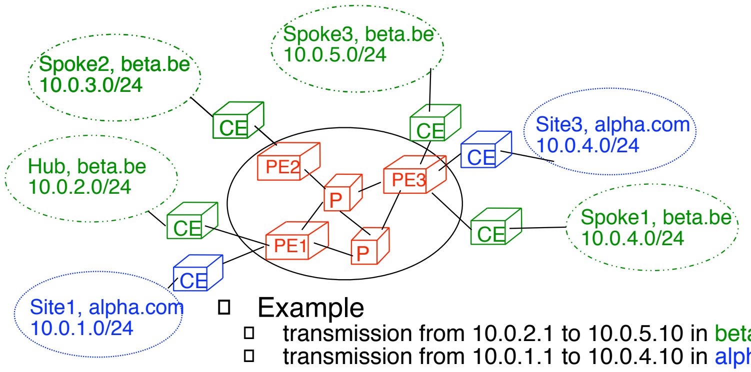
Redistributes to PE3 and PE2 the following routes
- VPN alpha Comm: alpha,next-hop=PE1
- VPN beta Comm: beta_hub,next-hop=PE1

Imports the routes with communities
- alpha, beta_spoke

Solving the forwarding problem

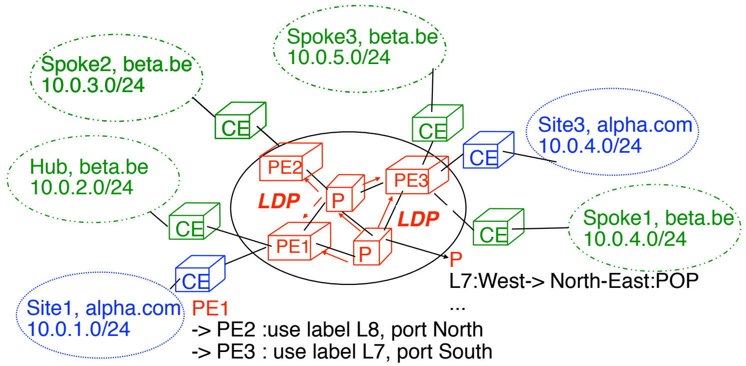
- How to forward the packets from each VPN through the provider's backbone ?
  - sending pure IP packets is not possible
    - P routers cannot know VPN-specific routes
    - different VPNs use the same RFC1918 address space
- Principle of the solution
  - CE routers send normal IP packets
    - CE routers remain as simple as possible
  - PE routers maintain several routing tables
    - one routing table per VPN attached to PE router
    - one routing table for the ISP backbone
  - PE encapsulate VPN packets
    - Common solution is to encapsulate with MPLS
    - Some ISPs are using GRE, L2TP or IPSec

Solving the forwarding problem with MPLS



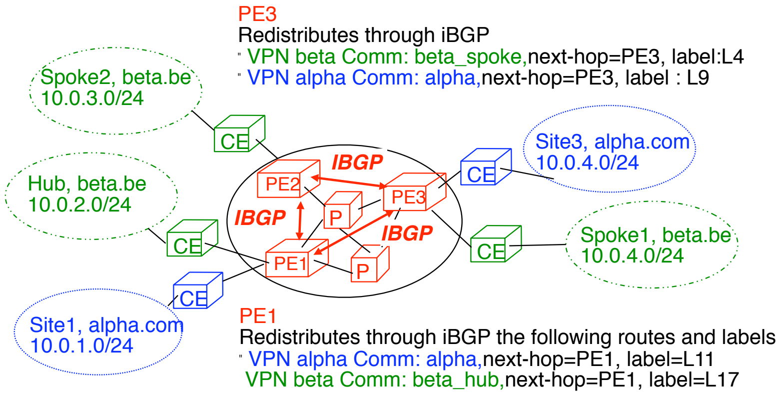
- Example
  - transmission from 10.0.2.1 to 10.0.5.10 in beta.be
  - transmission from 10.0.1.1 to 10.0.4.10 in alpha.com
- Principle of the solution : two levels of labels
  - one level of label is used to reach the next-hop PE
  - one level of label is used to indicate the VRF to be used (and thus the outgoing CE) in the egress PE

Distribution of labels



- Inside ISP backbone, use LDP to distribute labels between P and PE routers
  - each PE knows the label to use to reach any PE router
  - number of labels in P router depends on the number of PE, and not on the number of VPN sites

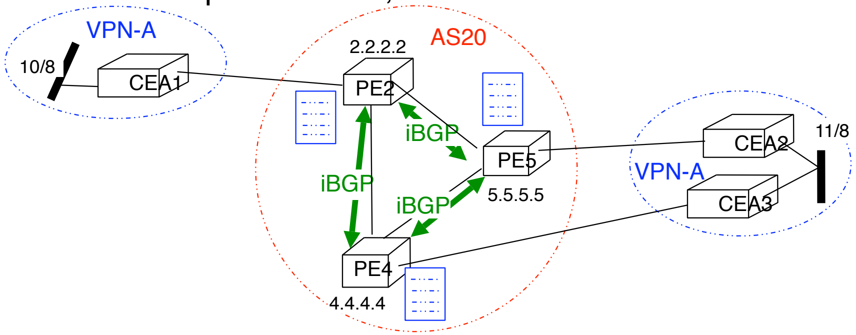
Distribution of labels (2)



- Principle
  - use iBGP to distribute VPN labels between PE routers

Packet flow in RFC2457 VPNs

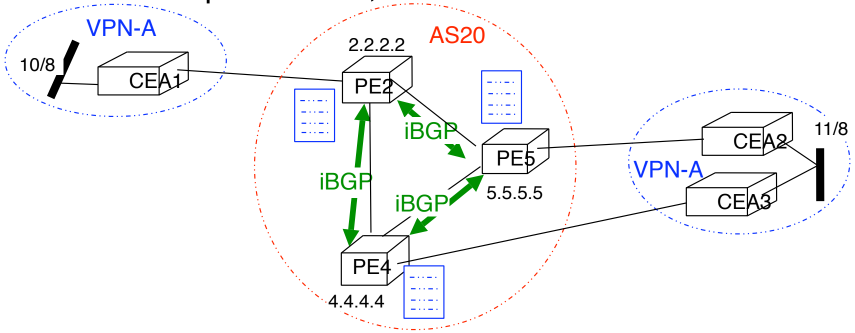
- with per-VPN RD, how does PE2 reach 11/8 ?



- PE2 receives two routes for 20:10:11/8
  - 20:10:11/8 from PE4 with nexthop = 4.4.4.4 (PE4)
  - 20:10:11/8 from PE5 with nexthop = 5.5.5.5 (PE5)
- PE2 selects the best route with its BGP decision process and installs it inside its VPN-A VRF
  - PE2 may use two LSPs to reach 11/8 via PE4 and PE5

Packet flow in RFC2457 VPNs (2)

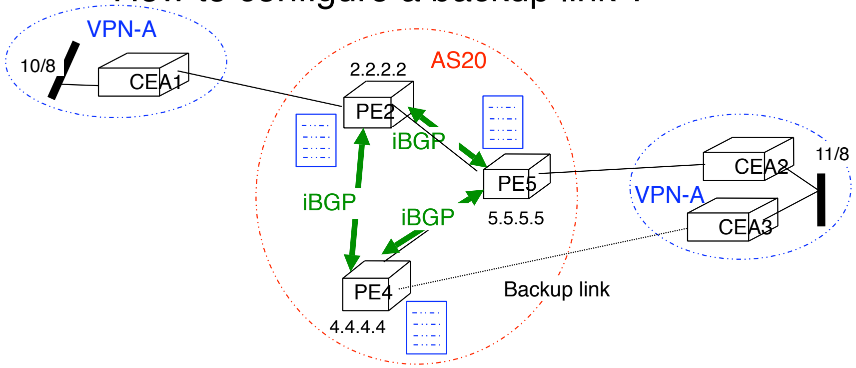
- with per-site RD, how does PE2 reach 11/8 ?



- PE2 receives two routes for 11/8
  - 4.4.4.4:123:11/8 from PE4 with nexthop = 4.4.4.4 (PE4)
  - 5.5.5.5:456:11/8 from PE5 with nexthop = 5.5.5.5 (PE5)
- BGP does not help PE2 to select which route is the best, the selection is done when installing in VPN-A VRF
  - PE2 may use two LSPs to reach 11/8 via PE4 and PE5

Backup links with RFC2457 VPNs

- How to configure a backup link ?



- PE4 adds localpref=50 to route learned from CEA3
- PE4 and all routers will prefer the route via PE5/CEA2
- Failure of link CEA2-PE5 will force PE5 to withdraw its VPN route towards 11/8 and the route via PE4 will be used

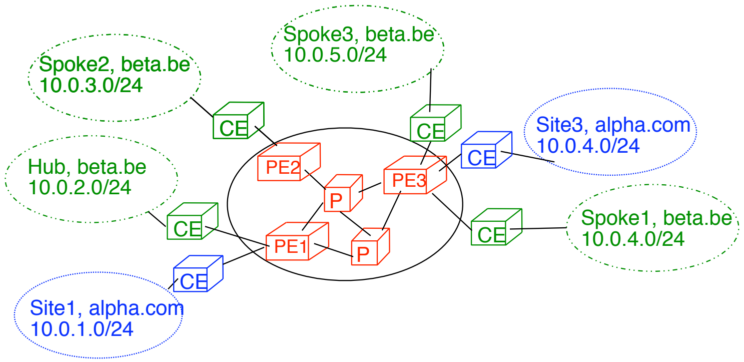
In this scenario, the convergence time in case of failure will depend on several factors :
- the time to detect the failure of the PE5-CEA2 link
  the best solution is clearly to detect the failure at layer1 or layer2. If the PE-CE protocol is used to detect the failure, then it may elapse several tens of seconds before the failure is actually detect and PE5 withdraws its VPN-IPv4 route

The type of route distinguishers used by PE4 and PE5 may influence the convergence time in large networks.

If PE4 and PE5 use the same route distinguishers for the routes learned from respectively CEA3 and CEA2, then when PE4 learns the RD:11/8 via iBGP, it will withdraw its own RD:11/8 route. When link PE5-CEA2 fails, PE4 will need to advertise its own route to all PE routers in the blue VPN. The propagation of this advertisement may take some time.

If PE4 and PE5 use different route distinguishers, e.g. 4.4.4.4:20 and 5.5.5.5:21, then both VPN-IPv4 routes will be received by all PE routers attached to CE routers in VPN-A. When installing the routes in their VRF, all PE routers will prefer the route with the 5.5.5.5:21 RD since it has the highest localpref value. However, all PE routers will always know both routes. Thus, if the route with RD=5.5.5.5:21 is withdrawn, then each PE router can quickly switch to the route with RD=4.4.4.4:20 provided, of course, that there is already a LSP between this PE router and PE4.

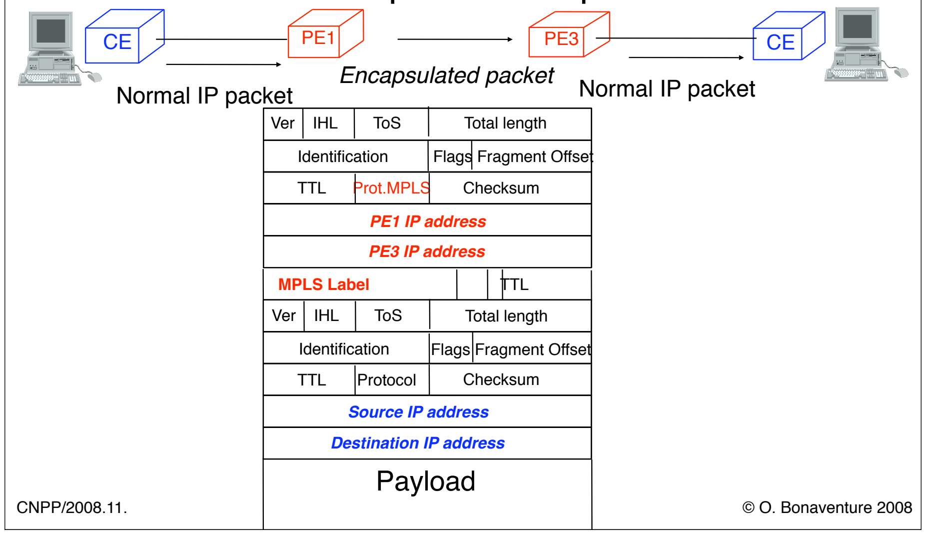
Solving the forwarding problem with tunnels



- Principle of the solution : Tunnel+MPLS
  - one tunnel is used to reach the next-hop PE
  - one MPLS label is used to indicate the VRF to be used (and thus the outgoing CE) in the egress PE

Solving the forwarding problem with tunnels (2)

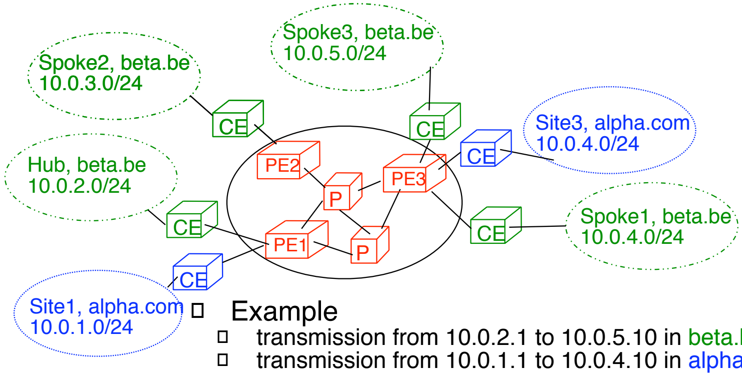
- How to the encapsulate the packets ?



It is also possible to use GRE tunnels to reach the egress PE instead of using MPLS-over-IP tunnel.

Solving the forwarding problem with tunnels (3)

PE3
Redistributes via iBGP
- VPN beta Comm: beta_spoke,next-hop=PE3, 10.0.5.0/24:label:L4
- VPN beta Comm: beta_spoke,next-hop=PE3, 10.0.4.0/24:label:L5
- VPN alpha,Comm: alpha,next-hop=PE3, label : 10.0.4.0/24L9



- Example
  - transmission from 10.0.2.1 to 10.0.5.10 in beta.be
  - transmission from 10.0.1.1 to 10.0.4.10 in alpha.com

Comparison of VPN solutions

- Provider-provisioned BGP/MPLS VPNs
  - Easy to configure for customer and provider
  - Provider can provider special QoS to VPN
  - But customer routes are distributed inside the provider's network by iBGP
    - provider may need to carry a large number of routes if clients use /32, /30 or /28 subnets
      - some ISPs report BGP/MPLS VPN tables larger than the BGP tables of backbone Internet routers
    - stability and convergence time of routing in the customer network depends on provider's iBGP
      - BGP has a rather slow convergence
      - Customer does not entirely controls routing in its VPN

Comparison of VPN solutions (2)

- Customer-provisioned VPNs
  - Providers are not involved in the provisioning of the VPN
    - no per-VPN routing tables to maintain and distribute
    - no revenue for value-added service
  - Customer builds VPN by establishing tunnels
    - it may be difficult to automate the tunnel establishment
    - a large number of tunnels may be required
  - Customer has full control over routing in the VPN
    - Routing protocol can be tuned for fast convergence, load balancing or whatever
      - no direct interactions between ISP's routing and VPN routing
    - Customer must be able to configure routers correctly

Layer 2 VPNs

- Service provided by RFC2457 VPNs is transport of IPv4 packets
  - CE devices are routers that send and receive IPv4 packets

- Some customers or operators prefer to offer layer 2 service
  - CE devices are capable of sending and receiving Ethernet frames, possibly with VLAN identifiers
  - Network managed by operator is similar to large Ethernet switch
    - PE devices need to learn MAC addresses reachable via each CE device
    - PE devices need to advertise to other PE devices the reachable MAC addresses
    - Broadcast and multicast Ethernet needs to be supported