iBGP Route Reflectors topologies

Eduardo Grampín
Universidad Carlos III de Madrid

Outline

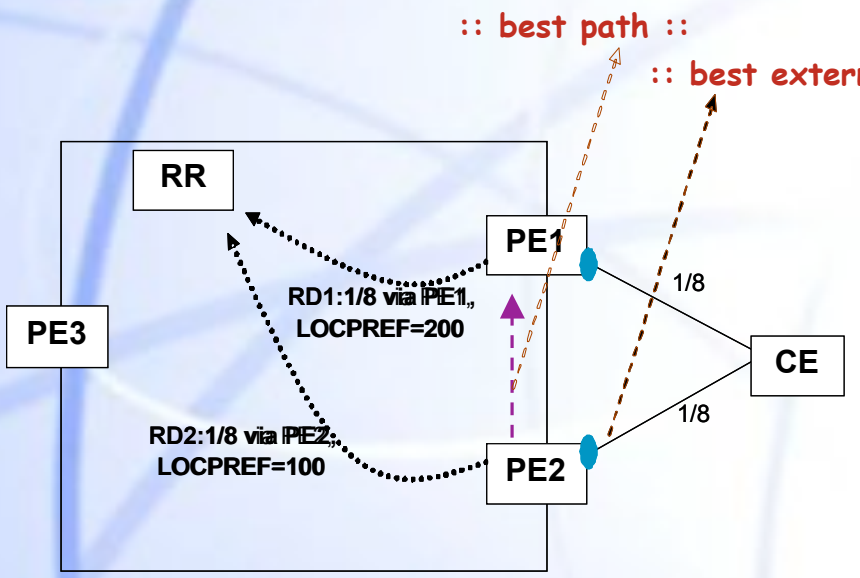
- Practical design guidelines
- Correct and scalable proposals
- **Recent IETF proposals**
- **Others**

Recent IETF proposals

- Best-external [BE-02]
- Add-Path [AP-04]
- N-plane Route Reflectors [NRR-02]

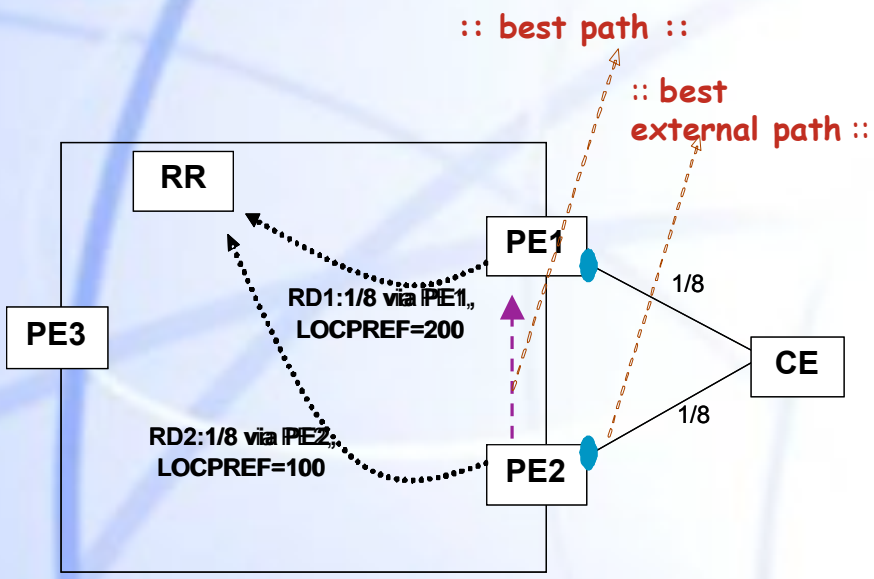
Advertisement of the best external route in BGP <draft-ietf-idr-best-external-02>

Basic idea



- PE2 selects iBGP-learned path as best
- Normally, PE2 withdraws (or fails to advertise) its eBGP-learned path
- With best-external, PE2's EBGP-learned path gets advertised to iBGP
- PE2 still installs the best path for forwarding

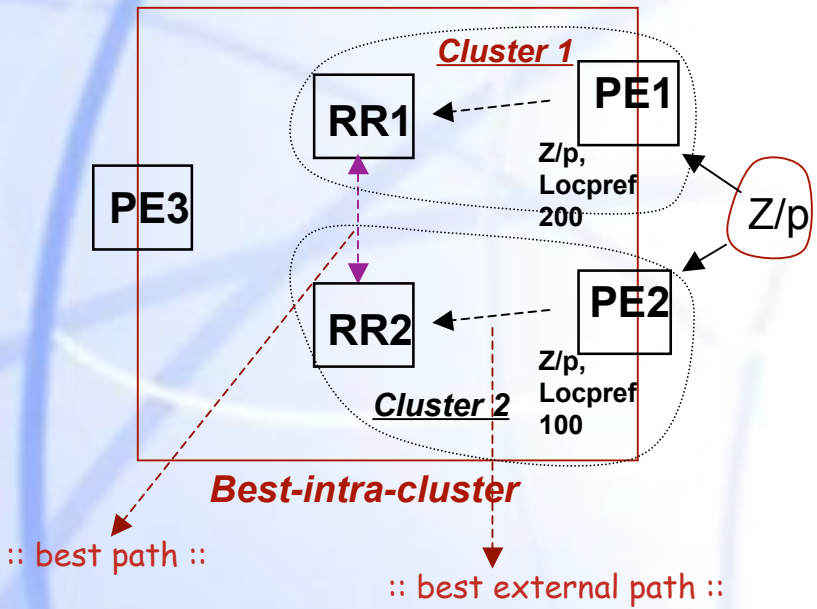
Motivation

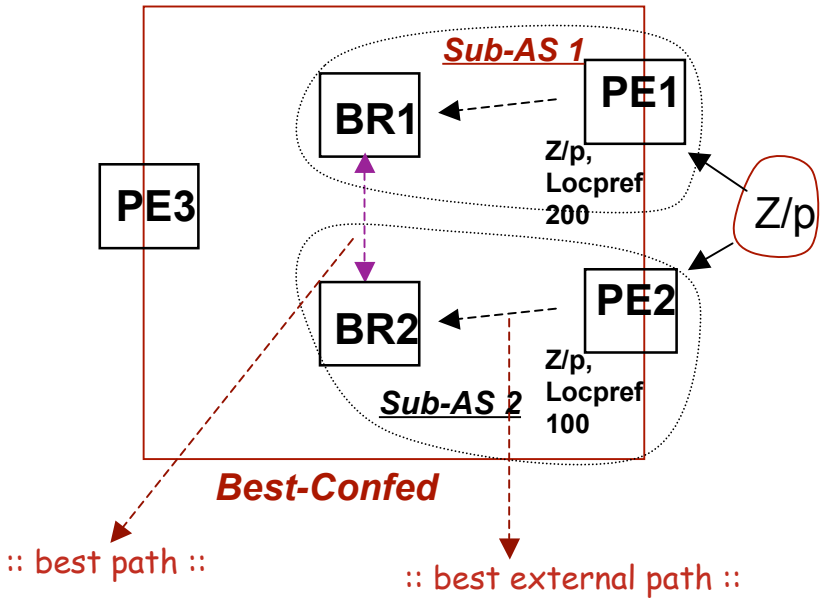


Example Scenario

- Create "primary-backup" topology by configuring LOCAL_PREF
  - primary = PE1-CE link
  - backup = PE2-CE link
- PE2 uses primary path for forwarding, but advertises backup path in iBGP
- If next hop of primary path becomes unresolvable, switch to backup can be made without waiting for PE1 and PE2 to send BGP messages

Other scenarios





Best External Path Selection

- Create a total ordering of all the paths based on the BGP decision process
- Best External path = First path in the total order that is external to the “domain”
- “Domain” can be:
  - AS
  - Cluster
  - Confederation

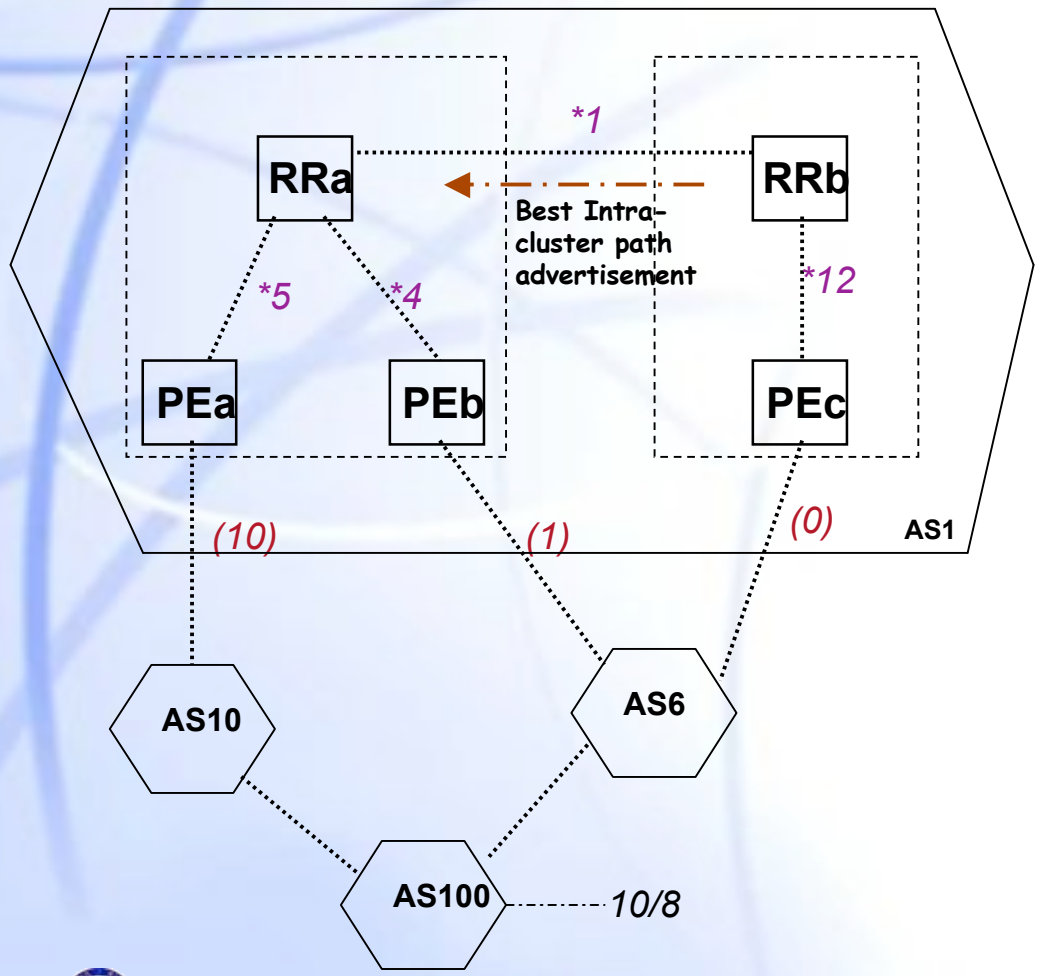
Applications

- **Fast Connectivity Restoration**
  - If next hop of primary route becomes unresolvable, switch to backup route can be made without waiting for BGP updates or withdraws from PEs
- **Inter-domain Churn Reduction**
  - Switching to secondary route might not affect what the upstream AS sees
- **Reducing persistent iBGP oscillation**

- **Note:**
  - Best external is implemented in both cisco and juniper routers

Reducing Persistent route oscillation

RFC3345 Example



- Reduce chances of persistent route oscillation by introducing additional information
- Global view of paths (ordered):

  [1] : '10 100, 10, <igp cost> (RRa:5, RRb:6)' *BEST

  [2] : '6 100, 0, <igp cost> (RRa:13, RRb:12)'

  [3] : '6 100, 1, <igp cost> (RRa:4, RRb:5)'
- Oscillation occurs because of circular dependency between paths
- Broken if RRb advertises its best intra-cluster path ([2]) when it chooses best ([1]).

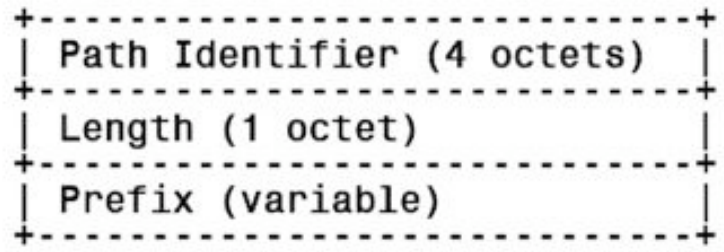
Advertisement of Multiple Paths in BGP <draft-ietf-idr-add-paths-04.txt>

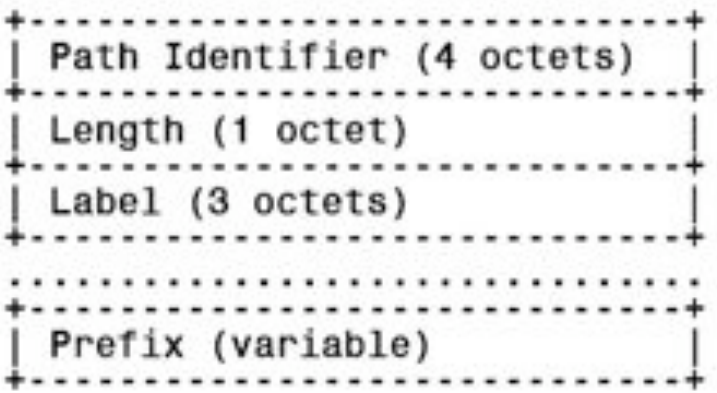
Basic Idea

- Mechanism that allows the advertisement of multiple paths for the same prefix without the new paths implicitly replacing any previous ones
- Summary: add a *path identifier* to the encoding
- The intent of this extension is to be used in a controlled fashion for applications that require only partial propagation of the routing information, or specific individual recipients

Modifications

- **NLRI Encoding**
  - The Path Identifier field is used to distinguish between different prefixes
  - Extension to
    - ✓ Multiprotocol Extensions for BGP-4 (RFC 2858) and the base spec (RFC 1771)
    - ✓ Carrying Label Information in BGP-4 (RFC 3107)
- **New Capability: ADD-PATH**





Selection modes

- How to advertise multiple paths over a single iBGP session?
- Different applications lead to different selection modes
  - All paths
  - N paths (max)
  - AS-Wide best paths
  - Neighbor-AS group best paths
  - Best Loc Pref / Second best Loc Pref paths
  - (other modes?)
- Interop issues when different path selection modes are applied by speakers in an AS?

Applications

- Preventing MED Oscillation
- Several multipath applications
- Route Server
- Others?

Distribution of diverse BGP paths <draft-ietf-grow-diverse-bgp-path-dist-02>

Multi plane route reflection

- Standard BGP4 specification allows for the selection and propagation of only one best path for each prefix
- Path diversity is desirable
  - Preventing MED Oscillation, multipath applications, etc
- Proposal: *multi plane route reflection*
  - The best path (main) reflector plane distributes the best path for each route as it does today
  - The second plane distributes the second best path for each route and so on
  - Distribution of N paths for each route can be achieved by using N reflector planes

Deployment

- Each plane of route reflectors is a logical entity
  - May or may not be co-located with the existing best path RRs
- Configuring an additional iBGP session from the current clients if required
  - No code changes required on the route reflector clients
- Claim:
  - The installation of one or more additional route reflector control planes is much cheaper and an easier than the need of upgrading 100s of route reflector clients in the entire network to support different protocol encoding -> ADD-PATH
- But…
  - What about routing state?
  - New sessions required (remember full-mesh configuration scalability issues)

Other proposals

- BGP Scalable Transport [BST03]
- Centralized solutions

BST - BGP Scalable Transport

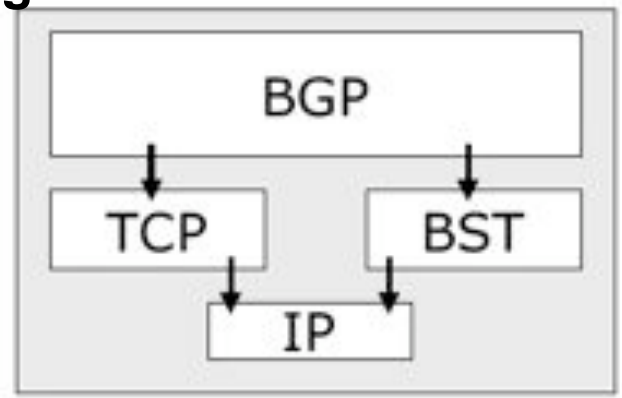
- **NANOG 27, 2003**
  - **Kedar Poduri, Cengiz Alaettinoglu, Van Jacobson**

BGP and TCP

- BGP is built from two separable pieces:
  - the BGP protocol and
  - the TCP transport used to carry protocol messages between peers
- It's relatively easy to add additional transport(s) to BGP
  - To quantify: adding BST to GateD required changing ~200 lines of GateD code
- If designed carefully, the new transport should have no effect on the protocol or its behavior: the same bits get delivered to the same peers in the same order, maybe a little faster and more reliably

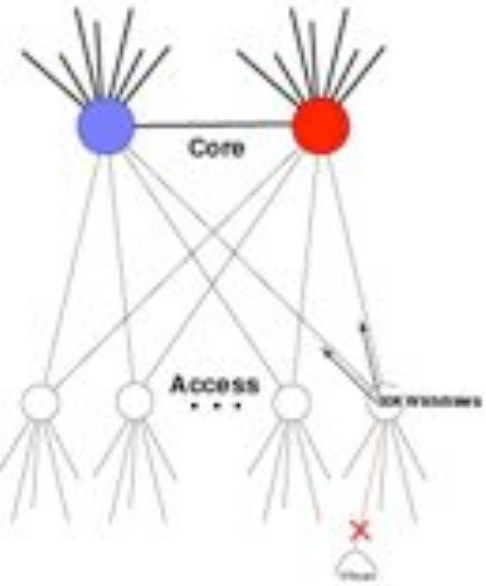
BGP and TCP

- The same BGP messages are sent (copied) to every BGP peer
- Rather than sending many copies, a multipoint transport could take care of delivering one copy to all interested parties
- Multipoint transport can be done two different ways:
  - Use multicast
  - Use application level replication & flooding like IS-IS or OSPF
- BST uses the second approach



Where can transport help?

- Take a typical PoP with core routers in a full iBGP mesh and acting as router reflectors for the PoP's access routers
  - Say one of the access routers loses a peering with UUnet. That router sends withdraws for ~50K prefixes to its local core routers (~500KB of data)
  - Since other UUnet peerings are in different PoPs, the core routers have to send those withdraws to all the other core routers (100–200 of them)
  - So each core router in the PoP sends 500KB on each of 200 different TCP connections – 100MB total
- Transport can't fix the 500KB of withdraws but a better transport can prevent inflating that 500KB into 100MB



Comments

- Replacing n parallel TCP connections with a single multipoint connection cuts down on traffic while improving reliability and convergence

- But, by itself a multipoint transport doesn't solve BGP's configuration issues:
  - every peer still has to be configured with the address of every other peer in the mesh and every configuration has to be updated when a router is added or removed from the mesh
- Neither it reduces the routing state for iBGP full mesh routers

Centralized solutions

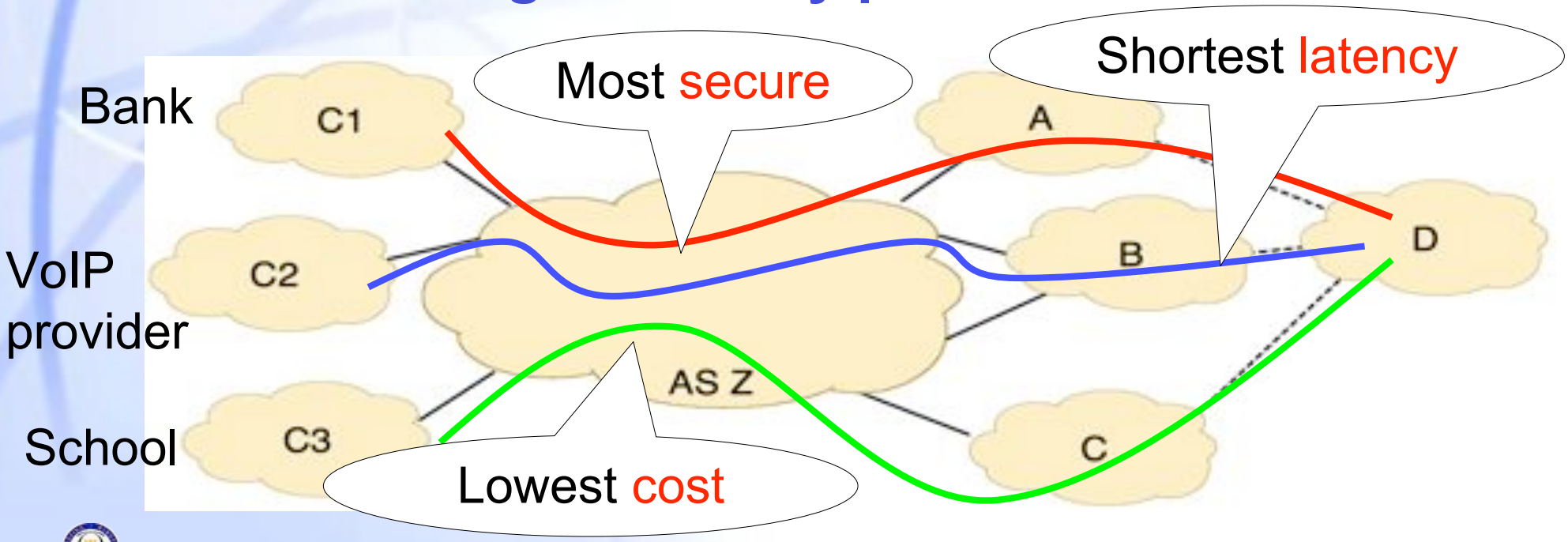
- Morpheus [MJSAC09]
- Others:
  - Route Control Platfrom [RCP]
  - iBGP Route Server Architecture [iBGPRS09]

Design for Configurability: Rethinking Interdomain Routing Policies from the Ground Up

- IEEE Journal on Selected Areas in Communications, April 2009
  - Yi Wang, Ioannis Avramopoulos, and Jennifer Rexford

A Case For Customized Route Selection

- Large ISPs usually have multiple paths to reach the same destination
- Different paths have different properties
- Different neighbors may prefer different routes

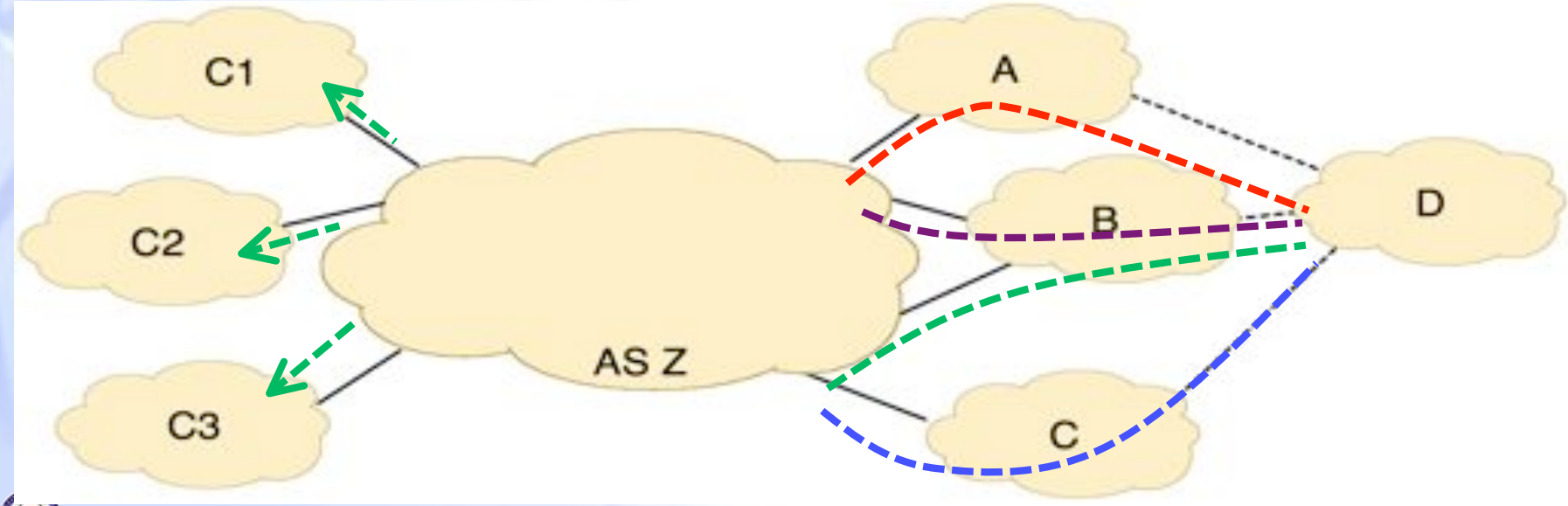


Exploit Path Diversity

- **Large ISPs Have Rich Path Diversity**
  - Top 2% ASes have 10 or more AS paths for certain destinations
- **Paths May Differ Significantly**
  - Security
    - ✓ Prefix / sub-prefix hijacking is a real threat
    - ✓ Avoiding an undesirable AS along the path
    - ✓ Large ASes are likely to have at least one valid/desirable route for most prefixes
  - Performance
    - ✓ Alternative BGP paths often have better performance than the default path
- **Path diversity gives large ISPs plenty of choices**

Exploit Path Diversity

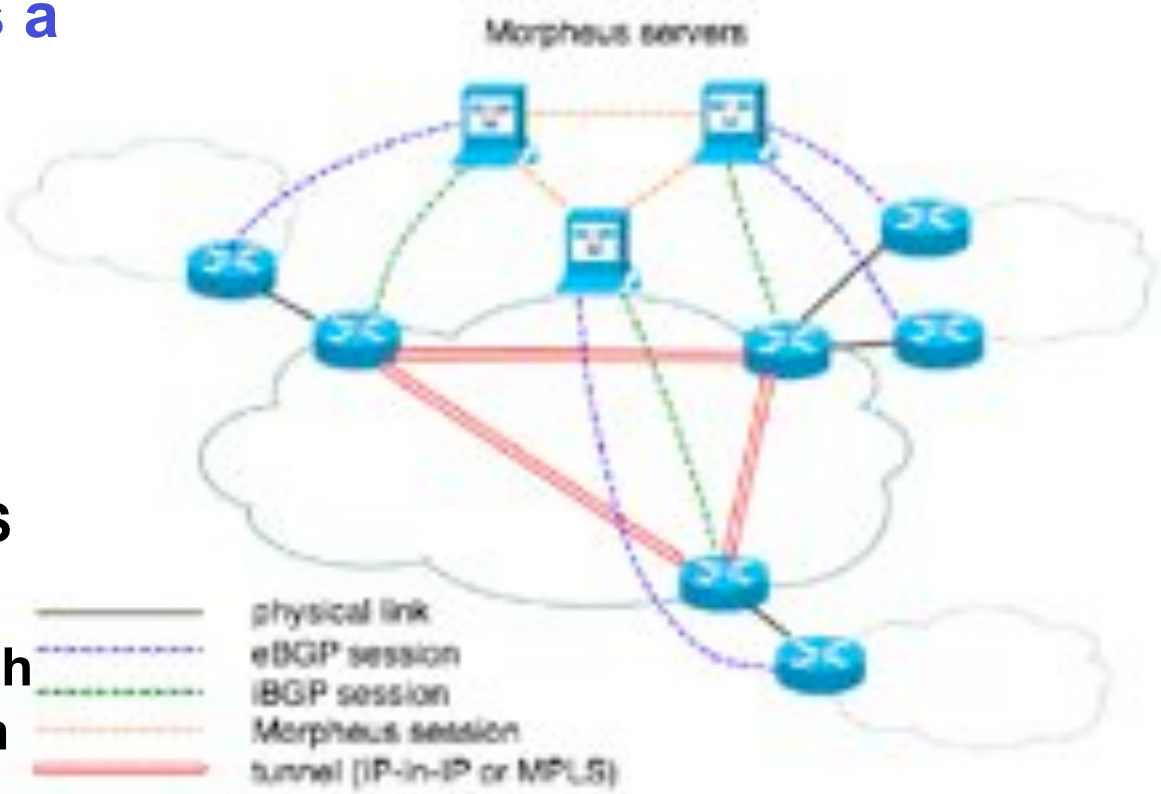
- **Flexibility Is Infeasible Today**
  - **BGP: The routing protocol ("glue") of the Internet**
    - ✓ **An ISP configures BGP to realize its routing policies**
  - **BGP uses a restrictive, "one-route-fits-all" model**
    - ✓ **Every router selects one best route (per destination) for all neighbors**



Morpheus: Enable Flexible Path Selection

- A *routing control platform* that enables a single ISP to flexibly pick paths for customers

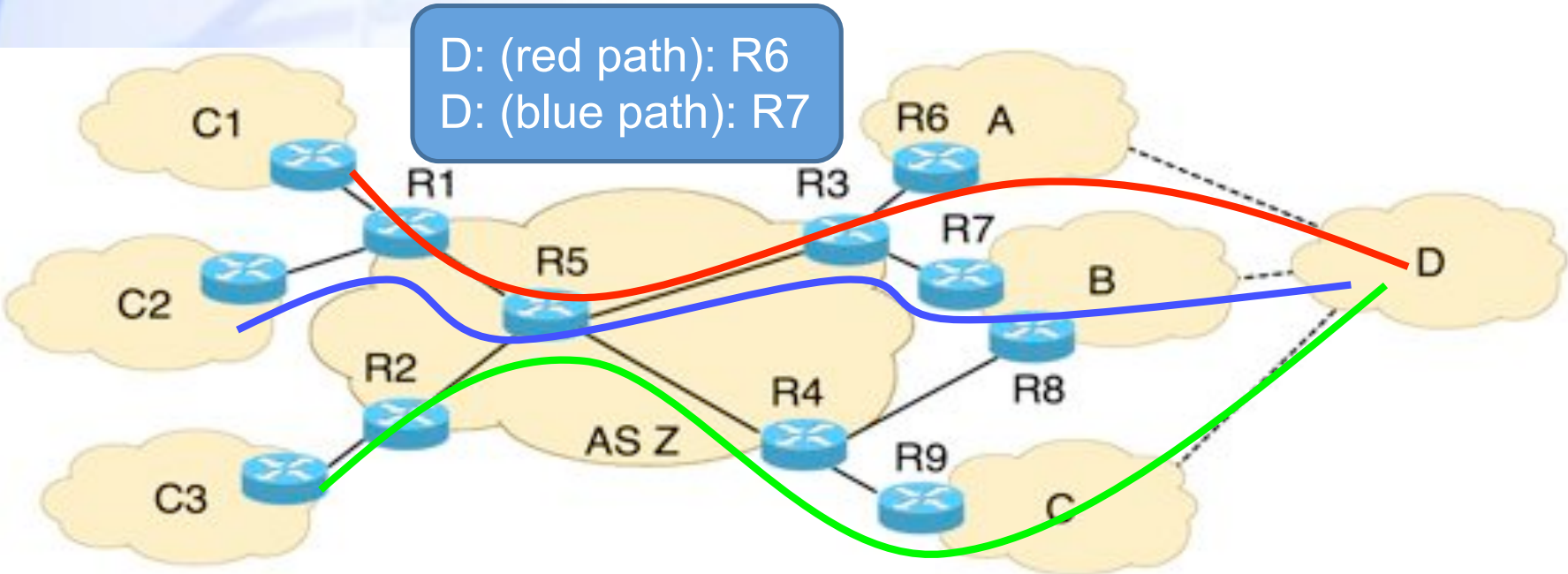
- Two components
  - Support from intra-AS routing architecture
  - Morpheus servers with flexible path selection processes



Flexible Route Assignment

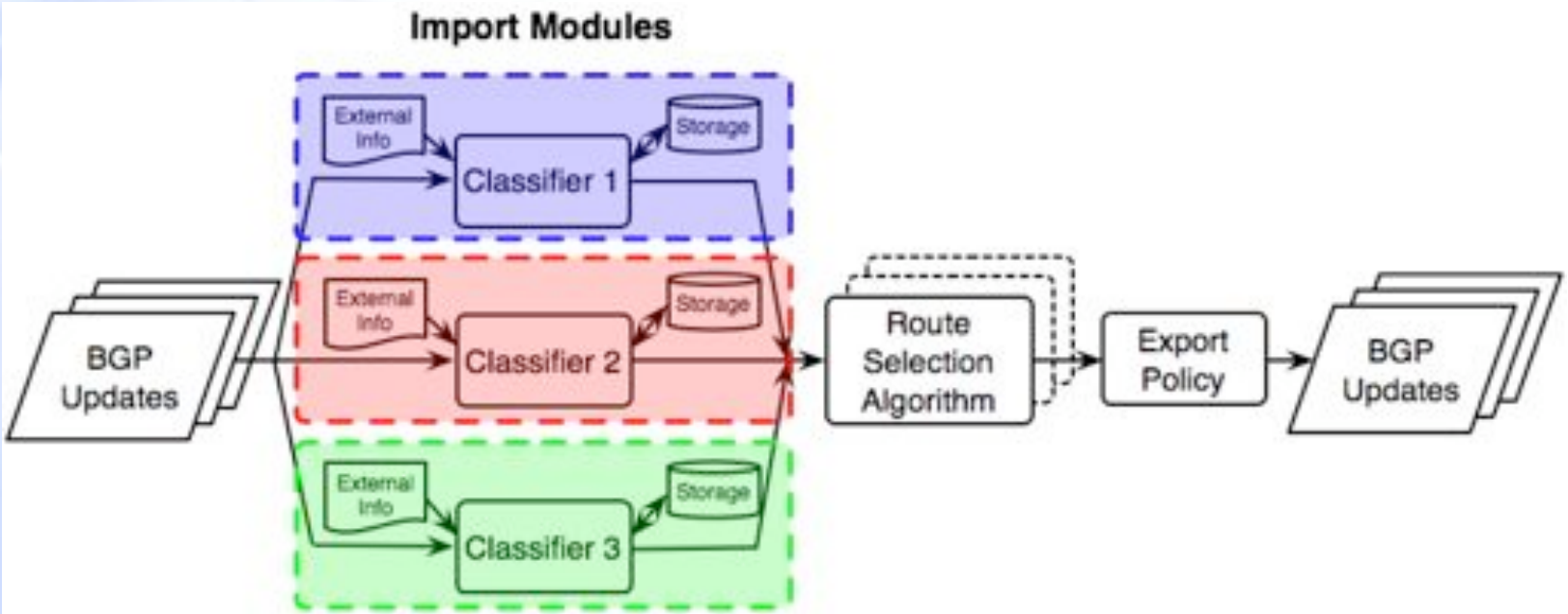
- **Support for multiple paths already available**
  - **"Virtual routing and forwarding (VRF)" (Cisco)**
  - **"Virtual router" (Juniper)**

    R3's forwarding table (FIB) entries



Inside Morpheus Server: Policy Objectives As Independent Modules

- Each module tags routes in separate spaces
- Easy to add side information
- Different modules can be implemented independently (e.g., by third-parties) – evolvability

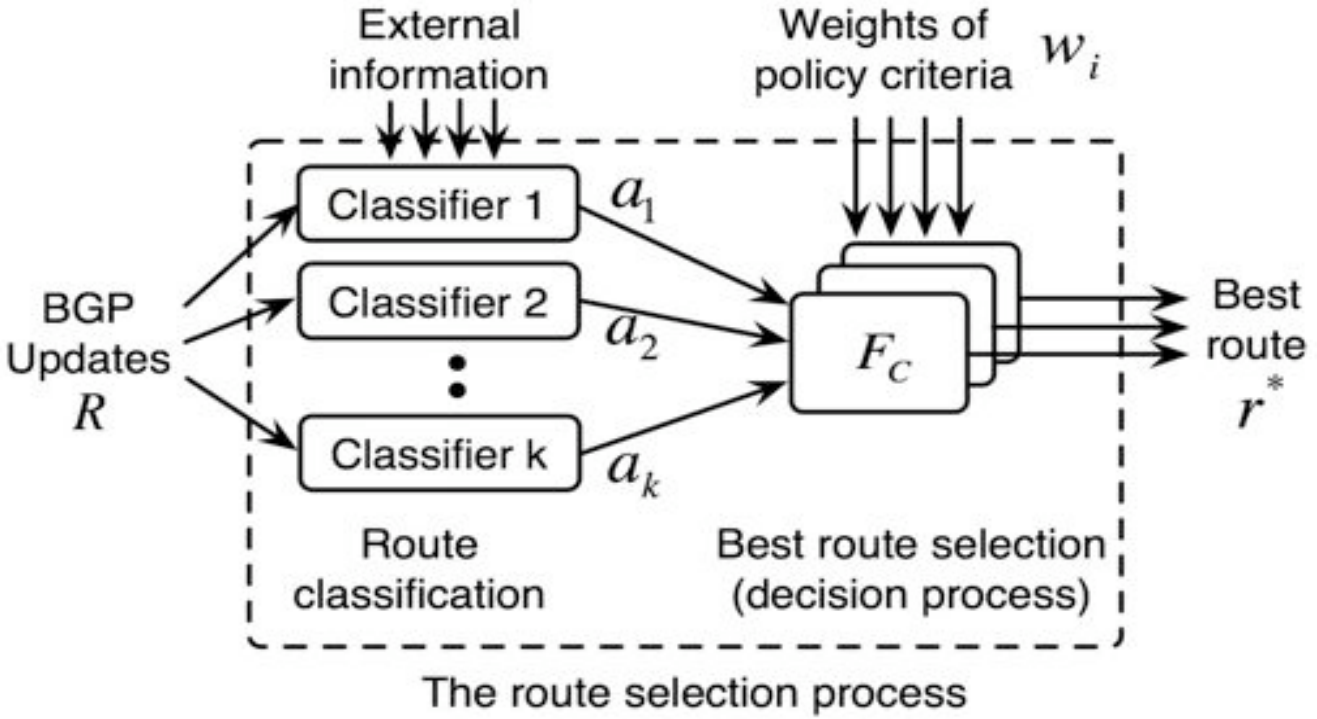


Policies

- Current BGP Implementations strictly rank one attribute over another (not possible to make trade-offs between policy objectives)
- E.g., a policy with trade-off between business relationships and stability

  "If all paths are somewhat unstable,
  pick the most stable path (of any length)
  Otherwise,
  pick the shortest path through a customer"
- Infeasible today

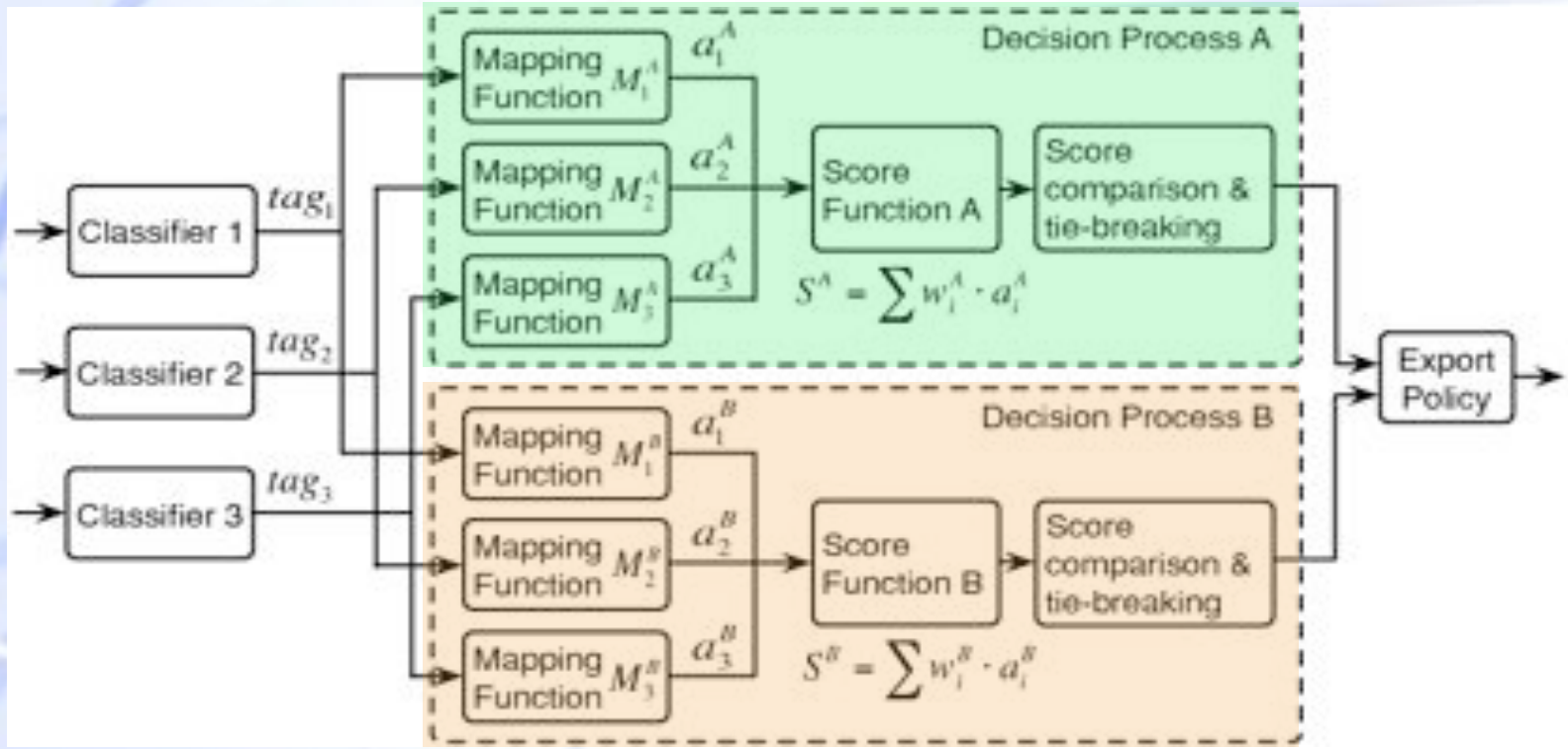
Use Weighted Sum Instead of Strict Ranking



The route selection process

- Every route r gets a value $a_i(r)$ of each criterion (policy objective) c_i (assigned by classifiers)
- Each criterion c_i is assigned a weight w_i
- Every route r has a final score $S(r)$: $$S(r) = \sum_{c_i \in C} w_i \cdot a_i(r)$$
- The route with highest $S(r)$ is selected as best: $$r^* = \arg\max_{r \in R} \left(\sum_{c_i \in C} w_{c_i} \cdot a_{c_i} \right)$$

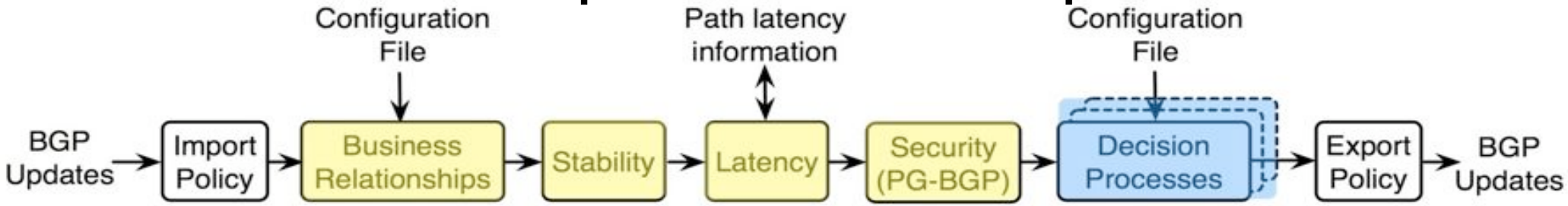
Multiple Decision Processes



- **Multiple decision processes running in parallel**
- **Each realizes a different policy with a different set of weights of policy objectives, selecting potentially different best routes**

Prototype Implementation

- **Implemented as an extension to XORP**
  - Four new classifier modules (as a pipeline)
  - New decision processes that run in parallel



- **Evaluation**
  - Classifiers work very efficiently
  - Morpheus is faster than the standard BGP decision process (w/ multiple alternative routes for a prefix)
  - Throughput: unoptimized prototype can support a large number of decision processes

iBGP Scalability and Topology Design: Summary

- iBGP full mesh: scaling problems
- Solutions:
  - Route reflectors
  - AS confederations
- Problem: correctness
  - Griffin: Checking the correctness of an iBGP graph is NP-complete, but two conditions ensure a correct (loop-free) iBGP graph:
    - ✓ 1) route-reflectors should prefer client routes to non-client routes
    - ✓ 2) every shortest path should be a valid signaling path
- Solutions:
  - iBGP Topology Design Problem: correctness and scalabilty
    - ✓ BGPSep
    - ✓ Fm-optimal
    - ✓ Skeleton

iBGP Scalability and Topology Design: Summary

- IETF: ongoing proposals
  - Improve route diversity
    - ✓ Add-Paths
    - ✓ Best-External
    - ✓ N-Plane RRs
- Centralized solutions
  - Improve route diversity and choice capabilities using an "omniscient" platform: *separating routing from routers*
  - Need help from routing architecture, e.g. MPLS tunnels from ingress to egress

References

1. [BE-02] P. Marques, R. Fernando, E. Chen, P. Mohapatra, *Advertisement of the best external route in BGP*. Internet Draft <draft-ietf-idr-best-external-02.txt>. Expires: February 8, 2011
2. [AP-04] D. Walton, A. Retana, E. Chen, J. Scudder, *Advertisement of Multiple Paths in BGP*. Internet Draft <draft-ietf-idr-add-paths-04.txt>. Expiration Date: February 2011
3. [NRR-02] R. Raszuk (Ed), R. Fernando, K. Patel, D. McPherson, K. Kumaki, *Distribution of diverse BGP paths*. Internet Draft <draft-ietf-grow-diverse-bgp-path-dist-02>. Expires: January 9, 2011
4. [BST03] Kedar Poduri, Cengiz Alaettinoglu, Van Jacobson, *BST - BGP Scalable Transport*. NANOG 27, 2003

References

5. [MJSAC09] Yi Wang, Ioannis Avramopoulos, and Jennifer Rexford, *Design for Configurability: Rethinking Interdomain Routing Policies from the Ground Up*. IEEE Journal on Selected Areas in Communications, April 2009
6. [RCP] Nick Feamster, Hari Balakrishnan, Jennifer Rexford, Aman Shaikh, Jacobus van der Merwe, *The case for separating routing from routers*. In Proceedings FDNA '04, ACM SIGCOMM workshop on Future directions in network architecture
7. [iBGPRS09] Uli Bornhauser, Peter Martini, Martin Horneffer, *An Inherently Anomaly-free iBGP Architecture*. IEEE 34th Conference on Local Computer Networks (LCN 2009). Zürich, Switzerland; 20-23 October 2009