A Combinatorial Optimization Framework for the Design of Resilient iBGP Overlays

Cristina Mayr
Universidad de la República
Montevideo, Uruguay
mayr@fing.edu.uy

Claudio Risso
Universidad de la República
Montevideo, Uruguay
crisso@fing.edu.uy

Eduardo Grampín
Universidad de la República
Montevideo, Uruguay
grampin@fing.edu.uy
Internet: network of networks

- Domain Names
- Number Resources: global pool of IP and AS numbers
- Protocol Assignments
BGP

a path vector protocol

A routing protocol used to exchange routing information among different networks

Learns multiple paths from internal and external BGP speakers

Pervasive BGP model for large ASes: all routers in an AS are iBGP speakers
BGP

1) prefixes learned from an eBGP neighbour can be re-advertised to an iBGP neighbour, and vice versa

2) prefixes learned from an iBGP neighbour cannot be re-advertised to another iBGP (split-horizon rule)
BGP

RFC4456 - BGP Route Reflection: An Alternative to Full Mesh Internal BGP
Known issues with Route Reflection

- **Less robustness** - If a RR fails, its client routers also become disconnected, which affects network availability and routing stability.
- **Slow convergence** - An update message may take several hops before reaching the iBGP destination router.
- **Sub-optimal routes** - As each RR has a partial view of the network topology, and only propagates its better option, it might select a best path different from the choice in full mesh.
- **Increased probability of loops** - In general there is more than one RR, to avoid single points of failure. So clients connect to several RRs, which could potentially introduce data plane loops.
- **Non deterministic behavior** - This happens when routing decisions depend on the arrival order of announcements.
Dealing with Route Reflection

- Multi-path
  - RFC7911 defines a BGP extension that allows the advertisement of multiple paths for the same address prefix
- BGP Prefix-Independent Convergence (PIC)
  - Internet Draft: based on a shared hierarchical forwarding chain.
- BGP Advertise-Best-External
  - Internet Draft: provides the network with a backup external route to avoid loss of connectivity with the primary external route. Main router providers do implement this feature.
- Diverse-Path (Shadow Router)
  - RFC6774: BGP can distribute an alternative path other than the best path between BGP speakers
- BGP Optimal Route Reflection (BGP-ORR)
  - Internet Draft: suggests modifications to the algorithm to choose best path so that all RRs learn all the paths. It needs Diverse-Path among RRs is in place.
Optimal Route Reflector Topology Design (ORRTD)

- Motivation
  - Active BGP entries:
    - IPv4: 760040 prefixes, 411460 CIDR aggregated, IPv6: 62722 prefixes, 37123 CIDR aggregated (*)
  - Impact in every router of the AS: memory, CPU
    - Each router needs to store a local database of all prefixes announced by each routing peer
    - Hold complete set of best paths, and perform a lookup into this forwarding data structure for each packet.

(*) Source: https://www.cidr-report.org, 01/14/2019
Optimal Route Reflector Topology Design (ORRTD)

- Find the minimum quantity of RRs:
  - Such that the routes selected with the designated RRs are the same that would have been selected if instead of having RRs, the iBGP speakers were fully meshed.
Optimal Route Reflector Topology Design (ORRTD)

- Find the minimum quantity of RRs:
  - Such that the routes selected with the designated RRs are the same that would have been selected if instead of having RRs, the iBGP speakers were fully meshed.

  ➢ how to choose the RRs?
  ➢ how to assign clients to RRs?
BGP revisited

- Attributes:
  - Local pref: Statically configured ranking of routes within AS
  - AS_PATH
  - NEXT_HOP
  - MED (Multi Exit Discriminator)
- Finite State Machine
- Messages
  - Open, Update, Keepalive, Notification
- Default decision for route selection:
  - Highest local pref, shortest AS path, lowest MED, prefer eBGP over iBGP, lowest IGP cost, router id
- BGP+IGP
  - Hot potato routing: a router prefers the route with the shortest IGP path (the closest exit point)
ORRTD - Assumptions

- AS with BGP speaking routers connected by a pure IP network (hop-by-hop routing)
- Complete AS topology (costs included) is known.
- Fixed set of eBGP adjacencies, set of ASBRs that send and receive routing information to/from other ASes.
- All external prefixes are learned through BGP, filtered according to the path selection algorithm to get to a set of classes of prefixes;
- The sets of prefixes (attributes included) learned by each ASBR are known in advance
Objects and premises of the technique

- Which of the internal routers will be designated as RR?
  - Routers are internal routers (IR) or ASBRs;
  - Only an IR can be RR (an ASBR cannot be a RR);
  - Every Client-IR (i.e. not RR) must be peer of a unique RR per class of prefixes;
  - Whenever optimal for some IR, every ASBR must be connected to at least one RR per class of prefixes;
  - An ASBR cannot peer with a Client-IR.
From the raw problem to its Integer Programming Formulation

Graph with 2 classes of prefixes (A and B)

ASBR-IR for prefix class A  ASBR-IR for prefix class B

IR-IR for prefix class A  IR-IR for prefix class B

IR to Class A
<table>
<thead>
<tr>
<th>A</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
</tr>
</tbody>
</table>

IR to Class B
<table>
<thead>
<tr>
<th>B</th>
<th>AB</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
</tr>
</tbody>
</table>

Mathematical model
Mathematical model

Formalization

- An IGP weighted graph \( G=(V,E,W) \)
- A node partition \( V = ASBR \cup IR \)
  \((ASBR \cap IR = \emptyset), E \rightarrow V \times V, W:E \rightarrow N^+\)
- A set of pre filtered finite set of prefixes classes \( C \)
  represented by \( ASBR2CLASS: ASBR \rightarrow 2^C \), where \( 2^C \) is
  the set of all the subsets of \( C \).
- \( ASBR2IR4C: C \times ASBR \rightarrow 2^IR \)
- \( IR2IR4C: C \times IR \rightarrow 2^IR \)
Introducing resilience

- **Link failure**: an edge $e = (x_i, x_{i+1})$ fails. Suppose the new closest ASBR from $u$ is $w$. Then create a fictitious prefix class $C^f$ published from $w \in ASBR$.

- **Internal router failure**: if $x_i \in IR$ fails and it belongs to some best path $p$ from another router $u \in IR$ to $v \in ASBR$, then a new best path to an ASBR must be computed. If this new best path ends in a different router $w \in ASBR$, proceed as in the previous case.

- **Border router failure**: if $v \in ASBR$ fails and it is the best exit for some $u \in IR$, then a new best path to some $w \in ASBR$ must be computed. Then create a fictitious prefix class $C^f$ published from $w \in ASBR$. 

$p$: best path from $u \in IR$ to $v \in ASBR$ for prefix class $A$. 
$p = u, x_{i_1}, ..., x_{i_h}, v$. 

\[ \text{Diagram:} \]**
Improving the model

Formalization

• $\text{FCres} : C \times \text{ASBR} \rightarrow 2^{\text{IR}}$ : for each IR and preferred ASBR, and certain prefix class $C$, obtain the new fictitious prefix class

• $\text{PFCres} : C \times \text{ASBR} \times \text{IR} \rightarrow 2^{\text{IR}}$ : best alternative path for the new fictitious prefix class

• $\text{PIR2IR4Cres} : C \times \text{IR} \rightarrow 2^{\text{IR}}$ : affinity among routers belonging to the best alternative path for the new fictitious prefix class
Improving the model

Input

\[ C \] : set of prefixes classes
\{S^k\} : set of border-to-internal BGP affinity matrices
\quad S^k_{ij} = 1 \text{ if and only if } j \text{ in ASBR-to-IR for prefix class } k,\text{ with } k \in C, \quad i \in BR, \quad j \in IR
\{T^k\} : set of internal-to-internal BGP affinity matrices
\quad FC^i : set of fictitious prefix classes
\{P^l\} : set of new BGP best path nodes from internal-to-BR affinity matrices
\{Q^l\} : set of new BGP best path IR-to-IR affinity matrices

parameters

\quad BR : set of all Autonomous System Border Routers
\quad IR : set of all Internal Routers
\quad S' : \{S^k\} \cup \{FC^i\}
\quad C' : C \cup FC
\quad T' : \{T^k\} \cup \{Q^l\}

variables

\quad x_i : 1 \text{ if router } i \text{ is to be a RR and } 0 \text{ otherwise;}
\quad y_i^k : 1 \text{ if ASBR } i \text{ is to be iBGP adjacent to IR } j \text{ for prefixes class } k \text{ and } 0 \text{ otherwise;}
\quad z_i^k : 1 \text{ if IR } i \text{ is to be iBGP adjacent to IR } j \text{ for prefix class } k \text{ and } 0 \text{ otherwise;}
\quad w_{gh}^l : 1 \text{ if nodes } g, h \in P^l, \text{ i.e., the alternative best path}
Mathematical model

\[
\begin{align*}
\min & \sum_{i \in IR} x_i \\
\text{Subject to:} & \\
\sum_{(ij) \in S'^k} y^k_{ij} & \geq 1, \\
x_j - y^k_{ij} & \geq 0, \\
x_j + \sum_{(ij) \in T'^k} z^k_{ij} & \geq 1, \\
x_i + x_j - z^k_{ij} & \geq 0, \\
x_i + x_j + z^k_{ij} & \leq 2, \\
\sum_{(jh) \in S'^k} y^k_{jh} - z^h_{ih} & \geq 0, \\
\sum_{i \in IR} x_i & \geq 2, \\
u^l_{gh} & \geq y^k_{ij}, \\
\sum_{(ij) \in p^l} y^k_{ij} & \geq 1, \\
x_i, y^k_{ij}, z^k_{ij}, w^l_{gh} & \in \{0, 1\}, \quad \forall i, j \in V, k \in C', l \in FC
\end{align*}
\]
Experimental results

- RRLOC++ emulation environment
  - based on Quagga, MiniNExT and ExaBGP
- Two prefix classes
- We represent each prefix class by just one prefix
- Topologies in which all vertices have degree greater than one
- Which border routers learn each prefix class
Number of RRs

# Table

<table>
<thead>
<tr>
<th>Topology</th>
<th># IRs</th>
<th># ASBRs</th>
<th>RRs ORRTD</th>
<th>RRs BGPSep</th>
<th>BGPSepS</th>
<th>RRs Zhang</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abilene</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>AB5</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>5</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>AB10</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>10</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Airtel</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Cernet</td>
<td>37</td>
<td>4</td>
<td>1</td>
<td>9</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Garr</td>
<td>47</td>
<td>7</td>
<td>4</td>
<td>12</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>UniC</td>
<td>24</td>
<td>3</td>
<td>2</td>
<td>8</td>
<td>10</td>
<td>4</td>
</tr>
<tr>
<td>Uninett</td>
<td>67</td>
<td>3</td>
<td>2</td>
<td>27</td>
<td>26</td>
<td>4</td>
</tr>
</tbody>
</table>
Comparisson of loc_rib table

**Loc_rib table**: stores routes that have been selected by this BGP router and have been validated, by applying local policies for route selection.

<table>
<thead>
<tr>
<th>ORRTD Model loc_rib_ipv4 table</th>
<th>FM model loc_rib_ipv4 table</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>prefix</strong></td>
<td><strong>next_hop</strong></td>
</tr>
<tr>
<td>177.10.158.0/24</td>
<td>192.168.0.1</td>
</tr>
<tr>
<td>202.70.88.0/21</td>
<td>172.16.2.2</td>
</tr>
<tr>
<td>177.10.158.0/24</td>
<td>192.168.0.1</td>
</tr>
<tr>
<td>202.70.88.0/21</td>
<td>192.168.0.7</td>
</tr>
<tr>
<td>177.10.158.0/24</td>
<td>192.168.0.8</td>
</tr>
<tr>
<td>202.70.88.0/21</td>
<td>192.168.0.7</td>
</tr>
<tr>
<td>177.10.158.0/24</td>
<td>192.168.0.8</td>
</tr>
<tr>
<td>202.70.88.0/21</td>
<td>192.168.0.8</td>
</tr>
<tr>
<td>177.10.158.0/24</td>
<td>172.16.1.2</td>
</tr>
<tr>
<td>202.70.88.0/21</td>
<td>192.168.0.7</td>
</tr>
<tr>
<td>177.10.158.0/24</td>
<td>172.16.3.2</td>
</tr>
<tr>
<td>202.70.88.0/21</td>
<td>192.168.0.1</td>
</tr>
<tr>
<td>177.10.158.0/24</td>
<td>192.168.0.1</td>
</tr>
<tr>
<td>202.70.88.0/21</td>
<td>192.168.0.7</td>
</tr>
<tr>
<td>177.10.158.0/24</td>
<td>192.168.0.8</td>
</tr>
<tr>
<td>202.70.88.0/21</td>
<td>192.168.0.8</td>
</tr>
</tbody>
</table>
Number of RRs and equations

NOMINAL CASE vs. RESILIENT CASE

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Abilene</td>
<td>8</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>114</td>
<td>419</td>
</tr>
<tr>
<td>AB5</td>
<td>5</td>
<td>3</td>
<td>2</td>
<td>3</td>
<td>55</td>
<td>131</td>
</tr>
<tr>
<td>AB10</td>
<td>10</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>250</td>
<td>478</td>
</tr>
<tr>
<td>Airtel</td>
<td>3</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>32</td>
<td>131</td>
</tr>
<tr>
<td>Garr</td>
<td>47</td>
<td>7</td>
<td>4</td>
<td>4</td>
<td>3080</td>
<td>3852</td>
</tr>
<tr>
<td>UniC</td>
<td>24</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>902</td>
<td>944</td>
</tr>
<tr>
<td>Uran</td>
<td>18</td>
<td>5</td>
<td>3</td>
<td>5</td>
<td>599</td>
<td>687</td>
</tr>
<tr>
<td>Jgn2Plus</td>
<td>11</td>
<td>6</td>
<td>2</td>
<td>6</td>
<td>399</td>
<td>453</td>
</tr>
</tbody>
</table>
## BGP sessions

<table>
<thead>
<tr>
<th>Topology</th>
<th>resilient</th>
<th>ORRTD</th>
<th>Full Mesh</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abilene</td>
<td>20</td>
<td></td>
<td>55</td>
</tr>
<tr>
<td>AB5</td>
<td>18</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>AB10</td>
<td>23</td>
<td></td>
<td>78</td>
</tr>
<tr>
<td>Airtel</td>
<td>26</td>
<td></td>
<td>36</td>
</tr>
<tr>
<td>Garr</td>
<td>331</td>
<td></td>
<td>1431</td>
</tr>
<tr>
<td>UniC</td>
<td>24</td>
<td></td>
<td>351</td>
</tr>
<tr>
<td>Uran</td>
<td>57</td>
<td></td>
<td>253</td>
</tr>
<tr>
<td>Jgn2Plus</td>
<td>64</td>
<td></td>
<td>136</td>
</tr>
</tbody>
</table>
ASBRs can be RRs

<table>
<thead>
<tr>
<th>Topology</th>
<th>#RRs resilient ORRTD</th>
<th>#RRs resilient ORRTD - BRs</th>
<th>#Sessions resilient ORRTD</th>
<th>#Sessions resilient ORRTD - BRs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Garr</td>
<td>4</td>
<td>6</td>
<td>331</td>
<td>149</td>
</tr>
<tr>
<td>Abilene</td>
<td>2</td>
<td>2</td>
<td>20</td>
<td>17</td>
</tr>
<tr>
<td>AB5</td>
<td>3</td>
<td>3</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>AB10</td>
<td>2</td>
<td>2</td>
<td>23</td>
<td>19</td>
</tr>
<tr>
<td>Airtel</td>
<td>2</td>
<td>2</td>
<td>26</td>
<td>8</td>
</tr>
<tr>
<td>UniC</td>
<td>2</td>
<td>2</td>
<td>24</td>
<td>25</td>
</tr>
<tr>
<td>Uran</td>
<td>5</td>
<td>4</td>
<td>57</td>
<td>35</td>
</tr>
<tr>
<td>Jgn2Plus</td>
<td>6</td>
<td>6</td>
<td>64</td>
<td>44</td>
</tr>
</tbody>
</table>
Conclusions and next steps

• We proposed ORRTD, a mathematical model to tackle the problem of crafting a consistent, reliable and still optimal iBGP overlay, while minimizing the number of Route Reflectors (RRs) and sessions, even in the case of single failure.

• Experimental results corroborate not only the theoretical consistency of the ORRTD technique, but its outperformance over other alike heuristic approaches.
Conclusions and next steps

- Solving the problem for large instances and many prefix classes is difficult *(we proved it is a NP-hard problem).*

- What’s next?
  - Heuristic approaches
  - Prefix classes categorization
  - MPLS scenario
Questions?