Virtual Aggregation (VA)

Paul Francis, MPI-SWS
Xiaohu Xu, Huawei,
Hitesh Ballani, Cornell
Dan Jen, UCLA
Robert Raszuk, Cisco
Lixia Zhang, UCLA
Draft activity

• New version of main draft (01)
  – draft-ietf-grow-va-01
    • Minor changes

• Two new drafts:
  – draft-ietf-grow-va-mpls-innerlabel-00
  – draft-ietf-grow-va-auto-00
Changes to draft-ietf-grow-va-01

• Fix interaction problem with PIM multicast
  – Pointed out by John Scudder
-00 draft: FIB Suppression can happen here
-00 draft: FIB Suppression can happen here
But this breaks PIM!!
-01 draft: FIB Suppression must happen here
Post Routing Table FIB suppression

• Already implemented by Huawei
  – Tag entries in routing table as being suppressable
  – Suppress just before loading into FIB

• Comments?
• In VA, tunnels are “targeted” to remote ASBR (external peers)
• If MPLS is tunnel type, this can amount to a lot of LSPs
• This draft proposes “inner label”
  – Only require one LSP per local ASBR
  – More in line with MPLS TE
• Essentially same as used for MPLS VPNs
Three encapsulations

Stacked labels (RFC3032):
Payload | IP | Inner label | Outer label | link | ==> 

MPLS-in-IP (RFC4023):
Payload | IP | Inner label | Outer IP header | link | ==> 

MPLS-in-GRE (RFC4023):
Payload | IP | Inner label | GRE | Outer IP header | link | ==>
Three encapsulations

Payload | IP | Inner label | Outer label | link | ==>  
Payload | IP | Inner label | Outer IP header | link | ==>  
Payload | IP | Inner label | GRE | Outer IP header | link | ==>  

Outer header gets packets to the local ASBR
Three encapsulations

Payload | IP | Inner label | Outer label | link | ==>  

Payload | IP | Inner label | Outer IP header | link | ==>  

Payload | IP | Inner label | GRE | Outer IP header | link | ==>  

Inner label identifies the remote ASBR to the local ASBR
Mechanism

• When local ASBR advertises a route in iBGP
  – Set NEXT_HOP to itself
  – Assign a label
    • Inner label, used to identify remote ASBR
  – Convey label with RFC3107
    • “Carrying Label Information in BGP-4”

• Use RFC5512 to indicate outer header of IP or GRE-IP
  – "BGP Encapsulation SAFI and BGP Tunnel Encapsulation Attribute"
## Range of options

<table>
<thead>
<tr>
<th>Inner label?</th>
<th>5512 label?</th>
<th>LSP to Next Hop?</th>
<th>Tunnel Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Don't tunnel packet (normal behavior without VA)</td>
</tr>
<tr>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Use LSP</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Use 5512 tunnel to next hop</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Use 5512 tunnel to Next Hop if possible, else use LSP</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Use IP tunnel to Next Hop with inner label</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Use LSP (stacked labels)</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Use 5512 tunnel to Next Hop with inner label</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Use 5512 tunnel to Next Hop with inner label if possible, else use LSP</td>
</tr>
</tbody>
</table>
Comments?

• Q: Do we need to specify a “required” tunnel type?
• Four configs in VA
  – APR: It’s own VPs
  – Every router: VP-list
  – Every router: Popular prefixes (optional)
    • Some are trivially auto-configured (customer routes, routes for which router is egress)
    • High-volume popular prefixes require config
  – Every router: Tunnel type
draft-ietf-grow-va-auto-00

• Four configs in VA
  – APR: It’s own VPs
  – Every router: VP-list
  – Every router: Popular prefixes (optional)
    • Some are trivially auto configured (customer routes, routes for which router is egress)
    • High-volume popular prefixes require config
  – Every router: Tunnel type
A simple and useful deployment model (Robert Raszuk)

- One VP (0/0)
- All RRs are APRs for 0/0
  - (all RRs have full FIB)
- Edge routers have “default” plus simple popular prefixes
  - Routes for which edge router is egress
  - Customer routes
  - If room, routes with shortest iBGP metrics
- All paths are shortest path---no need for volume-based popular prefixes
A simple and useful deployment model (Robert Raszuk)

• This model can require very little configuration
  – If vendor provides it as a “special case”
  – “enable raszuk mode”

• More complex config only required if even RRs cannot hold entire FIB
  – Must deal with VP-list and volume-based popular prefixes
Automating config of high volume popular prefixes

• This feature is optional

• Model:
  – Management device receives netflow records from router
  – When netflow records indicate high-volume for some sub-prefix, management device tells router to FIB-install

• Router can be ASBR or RR
  – Must transmit iBGP updates
Automating config of high-volume popular-prefixes

- Note that it is the ingress router that needs to FIB-install to obtain shortest-path benefit

Two cases:

1. Router sees high volume incoming
   - Independently FIB-install high-volume sub-prefixes

2. Router sees high volume outgoing
   - Can be from many ingress routers, few of which see high-volume
   - Must somehow inform the ingress routers
For identified high-volume sub-prefixes:

- ASBR/RR attaches a “should FIB-install” tag (non-transitive extended attribute) to BGP updates for the sub-prefix
  - Send immediately or later
- Other routers use this as a hint in their FIB-installing decision process
  - i.e. don’t need to FIB-install if there isn’t room
- For RR, some corner cases whereby not all routers receive the tag
  - At worst, causes inefficiencies, not errors
  - See draft
• Comments?
How to know what to FIB-install?

• Routers must install VP routes
  – Routers must also tunnel packets to APRs
  – Therefore, routers must either know which routes are VP routes, or tunnel all packets
• APR must install VP sub-prefixes
• Installation of all other routes is optional
How to know what to FIB-install?

• Routers must install VP routes
  – Routers must also tunnel packets to APRs
  – Therefore, routers must either know which routes are VP routes, or tunnel all packets

• APR must install VP sub-prefixes

• Installation of all other routes is optional

Current approach:
Configure “VP-list” in all routers
How to know what to FIB-install?

• Keep VP-list approach as default mandatory approach
  – Note that VP-list doesn’t need to change very often

• Allow optional auto-config of VP-list or equivalent info
  – Draft defines two approaches:
    • “VP-route” tag
    • “Can-suppress” tag
VP-route tag

- APRs tag VP routes with non-transitive extended attribute
  - (Note these also tagged with NO_EXPORT)
- Receivers of tag know:
  - They must install VP route
  - They must tunnel packets to NEXT_HOP (the APR)
  - They may suppress sub-prefixes within the VP
VP-route tag: during BGP session startup

• During session startup (before End-of-RIB marker) router “assumes” that sub-prefixes are suppressible
  – After End-of-RIB marker, router knows all VPs, therefore knows what must be installed

• For many packets, delivery delayed until after end-of-RIB
  – Though alleviated by Graceful Restart
VP-route tag: VP route churn

- What if the only VP route for a given VP has churn?
- Two possible policies:
  - Allow this to lead to FIB churn
  - Dampen VP routes to avoid FIB churn (with penalty of non-delivery of packets)
Second approach: “can suppress” tag

- Configure ASBRs with “VP-range”
  - Ranges of addresses covered by all VPs
  - Eventually a single 0/0 entry
  - Non-ASBR routers need no such configuration

- ASBR tags routes within range with “can suppress” tag
  - Non-transitive Extended Attribute
  - Exception: VP routes are never tagged
  - May also not tag other routes according to policy, for instance customer routes
“Can suppress” tag

• Routers receiving the tag determine if they really can suppress
  – APR must FIB-install sub-prefixes within VP
• If all VP-routes go down, sub-prefix routes are never-the-less still tagged “can suppress”
• Packet could have both “can suppress” and “should install” tags
• Comments?