IDR Interim Meeting before IETF 103

agenda-interim-2018-idr-01-idr-01-01
Date: 10/26/2018, Time: 10:00-12:00

Web-ex link:
https://ietf.webex.com/ietf/ldr.php?RCID=6de208c314e7fcb229041a5ac290703d (streaming)
https://ietf.webex.com/ietf/lsr.php?RCID=9f63197e2f731e3a938472c49f0c8421 (download)

Attendees: Alvaro Retana, Alexander Azimov, Acee Lindem, Ajun Wang, Christoph Loibl, Erik Aman, Huaimo Chen, Jakob Heitz, Jeff Haas, Jim Uttaro, Jorge Rabadan, Kirill Kasavchenko, Ketan Talaulikar, Krishna Swamy, Linda Dunbar, Peter Psenak, Shyam, Sriram, Shunwan Zhuang, Zafar Ali, Erik Aman, Ruediger Volk, James Uttarro [23+]

Scribe: Susan Hares from recording

Agenda:

1.. Chair's Slides [10 min]
Includes discussion on BGP auto-discovery

2. BGP Auto-discovery: Chairs [10 min]

3. BGP Extension for SDWAN Overlay Networks Linda Dunbar [10 min]
https://tools.ietf.org/html/draft-dunbar-idr-bgp-sdwan-overlay-ext-00

4.. Application Specific Attributes Advertisement with BGP Link-State
Ketan Talaulikar/Peter Psenak [10 min]
https://tools.ietf.org/html/draft-ketant-idr-bgp-ls-app-specific-attr

5.. Discussion on growing BGP-LS attributes [20 min]

6.. Flowspec Indirection-id Redirect for SRv6 Huaimo Chen [5 min]
https://tools.ietf.org/html/draft-ietf-idr-srv6-flowspec-path-redirect/

7.. BGP FlowSpec Extensions for Routing Policy Distribution (RPD) Huaimo Chen [10 min]
https://tools.ietf.org/html/draft-li-idr-flowspec-rpd/

8.. BGP Extensions for IDs Allocation Huaimo Chen [10 min]
https://tools.ietf.org/html/draft-wu-idr-bgp-segment-allocation-ext/

9.. RFC 5575-bis: Dissemination of Flow Specification Rules Christoph Loibl [10]
https://tools.ietf.org/html/draft-ietf-idr-rfc5575bis-09

10. Detection and Mitigation of BGP Route Leaks Alexander / Sriram [10]
https://tools.ietf.org/html/draft-ietf-idr-route-leak-detection-mitigation-10

11. Aggregating BGP routes in Massive Scale Data Centers Jakob Heitz [15 min]
https://tools.ietf.org/html/draft-heitz-idr-msdc-bgp-aggregation-00

WG chairs Summary of requested actions:

1) WG Adoptions requests in interim meeting [10/26/2018]

- draft-ketant-idr-bgp-ls-app-specific-attr
- draft-ietf-idr-srv6-flowspec-path-redirect/
- draft-li-idr-flowspec-rpd -
- draft-wu-idr-bgp-segment-allocation-ext
- draft-heitz-idr-msdc-bgp-aggregation

2) WG adoptions prior to interim meeting

- draft-li-idr-bgp-ls-sbfd-extensions
- draft-ketant-idr-bgp-ls-flex-algo

3) WG LC pending

On hold due to RFC5575bis resolution, will restart after documents align with Rfc5575bis

- draft-ietf-idr-bgp-flowspec-oid –
- draft-ietf-idr-bgp-flowspec-interfaces-set –

Code Allocation drafts have high priority for WG LC

- draft-ietf-idr-wide-bgp-communities (7/29/2019 expires)
- draft-ietf-idr-bgp-open-policy (3/29/2019 expires)
- draft-ietf-idr-segment-routing-te-policy (10/12/2019 expires)
- draft-ietf-idr-bgp-ls-segment-routing-msd (IGP and BGP) (11/2/2019 expires)
- draft-ietf-idr-bgp-ls-segment-routing-rld (IGP, ? need BGP)
- draft-ietf-idr-bgp-eag-distribution (4/9/2019 expires)

4) WG LC with 2 implementations – Awaiting shepherd report

- draft-ietf-idr-rfc5575bis-10 (shepherd: John Scudder)
- draft-ietf-idr-bgp-optimal-route-reflection-17 (shepherd: John Scudder) -
- draft-ietf-idr-tunnel-encaps-10 (shepherd: John Scudder)

5) WG LC passed – awaiting implementations

- draft-ietf-idr-bgp-bestpath-selection-criteria
- driaft-ietf-idr-bgp-ls-node-admin-tag-extension-01
- draft-ietf-idr-ls-trill
- draft-ietf-idr-route-oscillations-stop (1 implementation)
- draft-ietf-idr-rs-bfd-06.txt -
- draft-ietf-idr-rtc-no-rt (1 implementation)

5) At IESG With AD

- draft-ietf-idr-bgp-gr-notification-15.txt – approved, revision needed
- draft-ietf-idr-te-pm-bgp-10 – publication-requested (10/9/18), in IETF LC, Telechat 12/12/2018
- draft-ietf-idr-bgpls-sgement-routing-epe-17 – pub-requested(10/19/18), in AD review
- draft-ietf-idr-bgp-ls-segment-routing-ext-11 – pub-requested (11/30/2018), in AD **review**

6) At RFC editor

- draft-ietf-iddr-bgp-prefix-sid-27 – at RFC Editor queue (MISREF)

Meeting Minutes:

Agenda:

1.. Chair's Slides [10 min]

- John indicated the status would be posted by IETF
- FCFS policy – means no substantive review of the document. As long as the request is formatted correctly, the IANA
- [comments on poor audio quality]

2.. BGP Autodiscovery: Chairs [10 min]

John: There is interest in the topic because we have 5 drafts on the topic.

- draft-ymbk-lsvr-lsoe-02
- draft-acee-idr-lldp-peer-discovery-03
- draft-heitz-idr-msdc-fabric-autoconf-00
- draft-xu-idr-neighbor-autodiscovery
- draft-raszuk-idr-bgp-auto-sesson-setup-00.

Alvaro requested that we try to have one proposal.

John reviewed the differences in the proposal. One solution (draft-heitz-idr-msdc-fabric-autoconf) is centralized. The rest of the documents are distributed with varying protocols mechanisms.

John (as WG chair) does not favor a requirements document. However as he looks at the proposals the one common proposal is to bootstrap single-hop E-BGP. The Union is large.

- Potential next step: There is enough interest so should work the group. The next step is an interim with proposal. A design team approach is also possible.
- Recommendation to the LSVR chairs: Please go ahead with LsoE adoption, but IDR may not use it.

Discussion:

- Ketan: There were some comments on our work that the state machine was not provided in our work with enough details. Version 10 does this
- Jakob: The mechanism I have is to discovery the whole fabric.
- John: I would encourage people to read Jakob's work.
- Jakob: The routing discovery is in the RTGWG.
- John: It is not clear that it is BGP.
- Acee: It is independent of BGP. You could use it with a different protocol.
- Jakob: That is why I put it in the routing working group.
- John: I think that we should have any interim that focuses the problem. We would like to have a small group that works on the topic. Does anyone strongly disagree with that approach? Please speak up if you have a difference mind.
  [silence]
- John: We will set up a design team that

3.. BGP Extension for SDWAN Overlay Networks
Linda Dunbar [10 min]
https://tools.ietf.org/html/draft-dunbar-idr-bgp-sdwan-overlay-ext-00

Summary: Proposing a new SAFI.
Why:

- Tunnel-Encap removed the SAFI=7 (RFC5512) for distributing encapsulation information. Current Tunnel-Encap requires Tunnels being associated with routes. Some people suggested adding a fake-route, but the configuration overhead is too high for (10K loads)
- Gap on draft-rosen-bess-secure-l3vpn – The solution is for only a few nodes where the secure keys are hand carried into the deployments. The authors did not want to expand this draft to the larger approach.

Discussion:
[After discussion on requirement of fake route]

- Acee: I do not want to dwell on this, but the IDR encapsulation draft has a color. Why do you need a separate for every endpoint? You will need an address for every site, but you will need that anyway. You will not need one for every tunnel. You can use the color for the tunnel. You need one address for CPE.
- Linda: The CPE can be a dhcp-generated address or a private address. Thank you for the input. I will change the chat.

[after discussion of BGP solution briefing]

- Acee: That's not true. There are lots solutions that do not use NHRP/DSVPN.
- Linda: There are some solutions that use the group key distribution. The end points need to distribute end point properties to the peers. NHRP/DSVPN does not scale. BGP can allow the CPE to advertise their attributes to Route Reflectors.

[after picture]

- Linda: Thank you John for your comments.

- Acee: The keys are used to bootstrap IKE-V2 and you use IKE-V2 on the sessions between the sites.
- Linda: This is correct.
- Jakob: You can make the port number 8 bytes and then you can use Ipv6 interface identifier or MAC.
- Linda: During a discussion yesterday with [x] for the port ID we need Ipv6 addresses. We do need to consider that the private address is Ipv6.
- Jakob: I meant the Ipv6 identifier which is 8 bytes. You could also use the whole address.
- John: Thank you Linda. There are several people who were interested in this proposal.

4.. Application Specific Attributes Advertisement with BGP Link-State
Ketan Talaulikar/Peter Psenak [10 min]
https://tools.ietf.org/html/draft-ketant-idr-bgp-ls-app-specific-attr

- [see presentation] – Request WG Adoption
  - Summary: Allows link attributes to be signal per application (RSVP TE, SRTE, LFA, Flexible Algorithm). A new BGP-LS top-level attribute TLV specified per application specific link attributes (ASLA TLV). Reuse the current TLVs and code points as sub-TLVS .
  - Request: WG adoption
- Comments:
  - John: This seems to be uncontroversial except for the next topic on our agenda. That aside, I'm going to assume the adoption is not going to be to difficult.

Comments

5.. Discussion on growing BGP-LS attributes [20 min]

- John: A heated discussion on the growth of BGP-LS attributes. When does the BGP-LS attribute additions seem to be too much. The key people concerned with this issues (Robert Raszuk and Adrian Farrell) are not on the call.
- Acee: If we did exceed the size at the expense of complexity, we could add a sequence number to the NLRI . Then it would be a different NLRI. It would be multiple NLRIs for the same prefix. It does not seem that it is a problem that is not solvable with some expense on computational complexities.
- Jeff: You are correct some form of sequence numbering for fragmentation or sequencing of the information can be done. One thing that has been ugly is that when we have done BGP or routing fragmentation before. One of the quasi-recent example is when we took PIM-RP messages and splitting these up for MVPN purposes. One challenging thing is that the standard NLRI is an mostly opaque key. There are some places where we ignore content (e.g. BGP labels). However, the problem with the fragmentation is knowing how many fragments are present. An NRLI cannot be considered whole until you have all 5 fragments. The general procedure for sending and receiving individual items. What happens if the fragments pass through the route-reflector that does not care about the content. Due to this fact, you may have ordering issues in order to get all the 5 fragments through. These

problems are not unsolvable, but they have been noted as problems before. Thus far we have tried to avoid doing fragmentation. It is very messy.

- Acee: I am aware the messiness. We have similar problems in OSPFv3 where we can do similar things with the router-LSA. It is a smaller scope problem in OSPFv3 because we have specific information. I agree there are problems regarding repacking. Using the information which has been repacking.
- John: Are we happy to accept a mathematician's solution that indicates we've found a problem and go on. Or do we believe that we need to bump this to a problem space.
- Jeff: Although I'm not caught up on the IDR list's discussion, I offer two comments
    - 1) Carrying this information on top of the standard BGP session and this gums up the work. One question is "How does one partition the work?" This is not specific to BGP-LS. BGP-LS is only one example case.
    - 2) General proliferation of when we add information to the IGPs these are bleeding over into BGP-LS. If we did not have the challenge of maximum size BGP PDU, then we could just state if you adopt a given feature in the IGP inside the LSR groups that you auto-adopt this in the BGP WG. However since BGP has this problem and the IGPs do not – then eventually we will grow too stuff in our existing problem.

I believe the length problem needs to be solved and soon-"ish".

- Ketan: I would concur. We should have a proposal or to look at one.
- John: Did you say you were going to propose a solution or that the WG needs a proposal?
- Ketan: I would suggest that we go what Acee and I mentioned in the thread earlier today. I suggest we have multiple instances of the NLRI (similar to OSPF) to allow the attributes to grow without going over the 4K boundary. I would not look at it as fragments, but multiple instances. This similar to the router information LSA rather than the router LSA of OSPF.
- Jeff: One final comment, we do have a proposal for large BGP messages. Does this address the problem and we can stop worrying about this for the short term?
- John: I was wondering if someone was going to mention that point? Has someone restarted this discussion on the mail list with that as a subject line?

6.. Flowspec Indirection-id Redirect for SRv6

Huaimo Chen [5 min]

https://tools.ietf.org/html/draft-ietf-idr-srv6-flowspec-path-redirect/

- Existing draft: draft-ietf-idr-flowspec-path-redirect (IPv4). This draft suggests: New sub-type for SRv6 and allows 3 new ID-Types for mapping.
    - ID-Type 6 (Node): indirection-id is an index into SRv6 SID global SID
    - ID-Type 7 (Binding): indirection-id is an global offset into SRV6 binding SID
    - ID-Type8 (Binding): indirection-id as Global SRv6 SID that maps SRv6 binding SID
- Draft: Request for adoption
- Discussion: none

7.. BGP FlowSpec Extensions for Routing Policy Distribution (RPD)

Huaimo Chen [10 min]
https://tools.ietf.org/html/draft-li-idr-flowspec-rpd/

- Draft proposes: two options: BGP attribute or BGP Wide Community
- Authors request Adoption
- Discussion:
    - Jeff: I have no immediate comments. I am an author of Wide Communities so I will take a look at your proposal and post it to the mailing list.
    - John: We will allow a bit of time for discussion before we do a WG adoption call.

8.. BGP Extensions for IDs Allocation

Huaimo Chen [10 min]
https://tools.ietf.org/html/draft-wu-idr-bgp-segment-allocation-ext/

- Draft proposes: BGP-LS
    - The Allocation of new BGP-LS Protocol-ID for allocating IDs associated with node.
    - A Node NRLI with "allocation" protocol ID will allow allocation of IDs associated with node: (SR-Capabilities for node, SR local Block, Indirection ID). The indirection ID will allow flow-specification to do a path redirect (see draft-ietf-idr-srv6-flowspec-path-redirect)
    - A Link NLRI with "allocation" protocol ID will allow allocating of ID associated with link.
    - A prefix NLRI with "allocation" protocol ID" will allow allocation of Prefix IDs (Prefix-SID TLV, Prefix RangeTLV)
- Authors: request adoption
- Comments:
    - Acee: I think you do need the extra protocol ID. This is the right approach.

[1:23:30 time in video]

9.. RFC 5575-bis: Dissemination of Flow Specification Rules

Christoph Loibl [10]
https://tools.ietf.org/html/draft-ietf-idr-rfc5575bis-09
Discussion: Two issues:

- Issue #1: BGP treats NLRI as opaque key to an entry in the database.
    - Con: garbage may be propagated.
    - Pro: NLRIs can be used to encode undefined work.
    - If remove opaque from the draft what happens?
        - Most implementation treat NLRI as "non-opaque"
        - Most consistent with the protocol (validation and BGP update filters)
- Fix to issue #1:
    - Draft changes are minimal , primarily to allow future FS-NLRI extensions
    - If there are errors , we should treat as withdraw + plus have "always true" extensions that allows for graceful extension of NLRIs
- Issue #2: RFC5575bis - section 7.6 rules on Traffic Action interference

    - Current draft: Resolve the conflicts with treat as withdraw
    - Problem: It may sensible for multiple redirects
        - Another option is to go to less restrictive behavior that is still predictable.
        - For example, if lowest redirect then it gets applied.
        - Another example, “lowest” rate-limit gets applied.
- Two questions for the WG
    - Shall we remove the opaque property from the draft and resolve the extensibility issues by “treating as withdraw++”?
    - Make interfering actions less restrictive and achieve predictability by sorting them?

  If we do remove the opaque property, most implementations will be in more compliance with the RFC5575bis than the original RFC5575.

- Call for comments on draft-10.txt 11/14/2018
    - Only
- Comments:
    - John Scudder: I have a question. You observe that by making the change #1 to remove the opaque, we would add more implementations to the compliance. With change #2, we will move more implementations to the non-compliant.
    - Christoph: On change #2, this is not true since RFC5575 does not specify this action. We did not do any checks to determine what is happening in the implementations. This is the predictable behavior.
    - Jeff Haas: At least for our implementation (Juniper) if you specify more than one community restricting the bandwidth, we will pick the first one. However, the first one is based on the fact we internally sort the communities by doing a “mem-comparison”. I suspect this behavior could be standardized at this point, but it is not consistent [across implementations?].
      We ran into a similar problem in the redirect community. In some cases you want to specify more than one redirect community, but the redirections should be done based on whether one can resolve the next hop you are redirecting to. For example, if you want to redirect to IP addresses-1, IP-Address-2, and IP-Address-3. If IPaddresses-1 and IPAddress-3 resolve, and IPaddress-1 is the closest one via the IGP you may want to choose IPaddress-1 rather than IPAddress-3.
      Before we change the specification, it may be helpful to query the implementation to determine if there is any level of consistency.
        - John: Christoph, you are indicating that no implementation really complies to an under specified requirement. Unless one is really luck that everyone just happens to comply with an approach, any change will move people into the non-compliant state.
        - Christoph: this is correct.
        - John Scudder: [As WG participant]: Your change #1 seems to be reasonable to document existing practice. Thank you for the future extensibility.
        - John [WG chair]: On #2, it seems that it is less clear. We started out this document to simply document the current state of implementations. If we change this portion we should turn re-do WG LC. Without a change #2, we can just send this document

forward to the IESG. #2 should be done in a future document. If we survey the implementations, and the restrictive sections (7.6) is common to all implementations then this is a big win.

- Christoph: If we remove this from the document, it we have unpredictable behavior.
- Jeff Haas: This is true.
- John Scudder: [WG chair]: I'm not stating this action is correct behavior, but that it could be fixed in a later document.
- Jeff Haas: While it is under specified, the communities are easy to modify on a hop-by-hop basis. If an implementation has a specific need, then these implementations could change the action upon these communities.
- Jeff: I am less personally concerned with resolving the conflict between traffic actions. We could simply provide guidance as a "should" after the survey [indicates there is no common] mechanism. I am far more concerned about the opaque behaviors.
- Christoph: Should we just document the current status in the document?
- Jeff: "Should" is the way out of making someone non-compliant. If you say "choose the lowest one" as a SHOULD be chosen, then the implementation that selects something else is still compliant. "SHOULD" provides long-term guidance to the implementations. It is a "net-good" to offer advice when we feel advice should be offered.
- Jakob: Did I hear first community? There is no such thing as a "FIRST" community is there?
- Jeff: As I mentioned, this is an implementation detail.
- John: Jeff is talking about lexical ordering.
- Jeff: This is correct. In our implementation we take and lexically sort the communities to provide a canonical form.
- Jakob: We do the same.
- Jeff: This is a common trick. This is my point if we are all doing the common trick [of sorting to a canonical form] and we all choose the first one after sorting, then we have a behavior that we can recommend.
- Christoph: If take this question to the mail list, will some people share their behavior on the communities.
- John: It is a fine idea. If you could propose some text, then see if the text applies to the implementation.
- Christoph: For #2, I will query the implementations with the text. For #1, I think we are writing down what is out there. All I need to change is to have the format for new implementations.
- Alvaro [x]: As the AD reviewing the document, I have a few comments. If we end up recommending something and no specifying something, SHOULD open the door. As the AD, I will ask the question why not "MUST". If SHOULD is in the text, then I will look for a reason why "SHOULD" is necessary. If the reason is, implementations do it this way – that is not an acceptable reason.
- John: I will work with Christoph to address this.

10. Detection and Mitigation of BGP Route Leaks
    Alexander / Sriram [10]
    https://tools.ietf.org/html/draft-ietf-idr-route-leak-detection-mitigation-10

- Proposal: Switch from BGP Attribute to BGP Community
    - Two indications: Down-only, Leak

Comments:

- John: I'm looking at the clock, could you jump to the request for input. You have 2 more minutes.

[feedback]:

- John: We do have time for questions. You are presenting this in the grow WG.
- Jakob: Communities are typically deleted as they go forward.
- Alexander Azimov: There is a presentation by Randy Bush that indicates that show that community (?) propagation is not what you think it is. It is more of a problem than you think. I think there is some information that should be reviewed.

11. Aggregating BGP routes in Massive Scale Data Centers
    Jakob Heitz [15 min]
    https://tools.ietf.org/html/draft-heitz-idr-msdc-bgp-aggregation-00

- Premise: Anything that uses SPF to aggregate the routes does not scale.
    - CLOS leaves do not provide any additional connectivity.
    - Server redundancy (1,2, or 4 links per server)
    - Massive – future: 1 million servers, 8 million links, 130K switches
    - BGP with route aggregation – 100s of routes
    - Massive routes cause many negative routes. IGP race conditions will have positive/negative routes.
- Definitions: hole-punch, punch-accept, do not aggregate, chad, negative routs
    - Hole-punch (punch a hole of aggregation),
    - Punch-accept (take a hole-punch),
    - do not aggregate
    - Chad route – punched out route (punched out of the aggregate with do not forward)
- Massive failures – Reduce costs – means reduce FIB space
    - Solution:
        - do not install all chad routes,
        - do not install aggregate with too many holes,
        - If no clean spines left, then install cleanest and then add in individual routes.
        - Fabric controller has the best view, and can install the most important route
- All of this means auto-configuration of BGP
    - Needs to allocate IP addresses and configure the routes as part of aggregate
    - Configuration – via another protocol
- Requests: WG adoption

Comments:

- John: All this fancy aggregation was something we did 20 years ago. We stopped. Why should we do it now? My answer was this work does not work in the Big I – Internet, but in the data center it is a different story.
- Jakob: I agree with John's point. I was trying to solve the problems of the negative routes. This is the place where the whole thing was sticking. There was a hole avoidance protocol which was doing something in the same vein.
- John: There have been drafts where there has virtual aggregation in grow which are in a related space.
- Jeff: This is similar to that [the virtual aggregation proposals in grow].
- Jakob: There is simple aggregation and virtual aggregation. The virtual aggregation spreads the routes around. To get to the real router you go through the aggregate route. The simple virtual aggregation is to reduce the RIB to put in the FIB. The difference here is the IP addresses are already aggregated. The IP addresses have to be allocated to the switches in a way that can be aggregated. If they cannot be aggregated, there is nothing you can save in the FIB. It needs to be a regular topology that you can take advantage of.
- John: I look forward to reading this again when I have more brain cells.
- Jakob: Acee is joining the draft.

John closes the meeting [2:18]