Internet Engineering Task Force
Internet-Draft
Intended status: Informational
Expires: August 10, 2019

R. Szarecki, Ed.
K. Vairavakkalai
N. Venkataraman
Juniper Networks Inc.

February 6, 2019

Use of Abstract NH in Scale-Out peering architecture draft-szarecki-grow-abstract-nh-scaleout-peering-00

Abstract

Many large-scale service provider networks use some form of scale-out architecture at peering sites. In such an architecture, each participating Autonomous System (AS) deploys multiple independent Autonomous System Border Routers (ASBRs) for peering, and Equal Cost Multi-Path (ECMP) load balancing is used between them. There are numerous benefits to this architecture, including but not limited to N+1 redundancy and the ability to flexibly increase capacity as needed. A cost of this architecture is an increase in the amount of state in both the control and data planes. This has negative consequences for network convergence time and scale.

In this document we describe how to mitigate these negative consequences through configuration of the routing protocols, both BGP and IGP, to utilize what we term the "Abstract Next-Hop" (ANH). Use of ANH allows us to both reduce the number of BGP paths in the control plane and enable rapid path invalidation (hence, network convergence and traffic restoration). We require no new protocol features to achieve these benefits.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on August 10, 2019.

Table of Contents

1. Introduction

Common to all large Internet networks are the requirements for large aggregate bandwidth and low latency. As network sizes and traffic volumes have increased, it has become common to use scale-out architectures to satisfy these requirements. Use of these techniques within individual networks is well-known. Here, we explore a scale-out architecture for interconnecting different Autonomous Systems (ASes).

Below, we show an example topology. Content is hosted within AS 2, consumers connect via the various ISP Metro ASes.

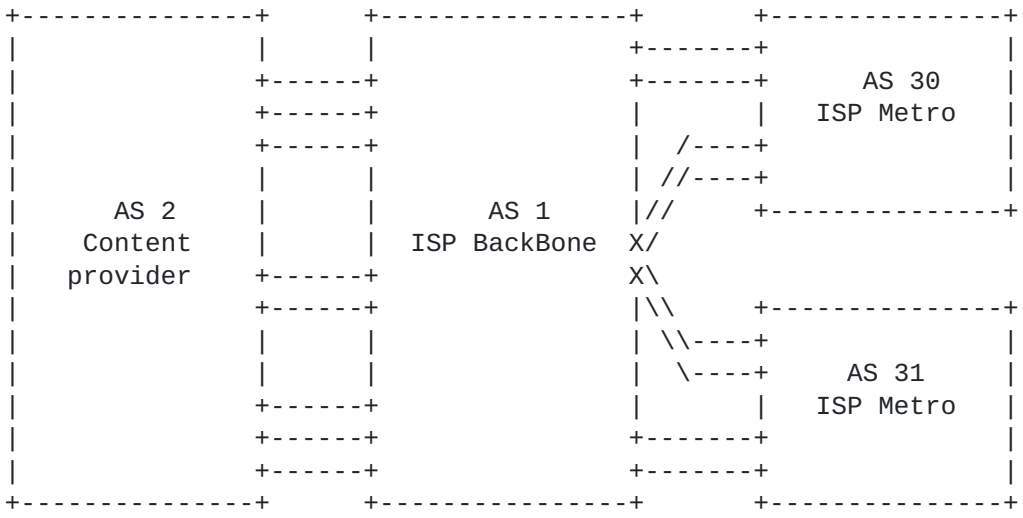

Figure 1

ASes 1 and 2 are connected at multiple, geographically diverse, sites. Geographic diversity is required for reasons including resiliency, minimization of latency, and minimization of cost associated with long-distance data transmission.

1.1. Scale-Out peering

The same trends that have driven the use of scale-out architectures within ASes drive interest in using them at peering sites. In such an architecture, each AS at the peering site deploys multiple independent Autonomous System Border Routers (ASBRs). Benefits that can be realized include N+1 redundancy and the ability to flexibly increase capacity as needed. The ASBRs are often connected to the rest of their AS in a leaf-spine topology through core routers, and augmented with a per-site pair of BGP route reflectors (RRs). See for example SITE1 in Figure 2, below.

The fundamental requirements in this architecture are:

a. Keep traffic on a path that has low latency.

b. Utilize all peering links that offer low latency.

c. In the event of failure, minimize the time needed to restore service.

1.1.1. Low latency

BGP, the Border Gateway Protocol, does not directly carry delay information. We make the general assumption in this document that paths selected by the BGP best path algorithm [RFC4271] will provide lower latency than those not selected. This assumption is not guaranteed to be true, but lacking special arrangements between peering ASes, it is what the protocol is able to provide.

1.1.2. All equal cost paths utilization

In order to use all links between peering ASes that provide the same BGP path costs to the destination prefix, at a minimum BGP speakers need to be enabled for multi-path operation. Additionally, all AS ingress BGP speakers need to know at least all equal and best paths to the destination via multiple ASBRs. If a full IBGP mesh is used, this happens naturally. However, IBGP full meshes are uncommon in large networks and are even more impractical in scale-out architectures due to the high total number of ASBRs.

The well-known techniques to deal with full-mesh scale challenges - Route Reflection [RFC4456] and Confederations [RFC5065] - hide redundant paths, as they advertise only a single selected path to their clients. While this helps keep path and session scale manageable, it makes BGP multipath unusable. We overcome this by using BGP ADD-PATH [RFC7911] between the RR and its clients (or among sub-ASes).

1.1.3. Summary

In summary, for a scale-out peering architecture:

- BGP multipath needs to be enabled on all IBGP sessions inside the AS.
- BGP multipath needs to be enabled on all EBGP sessions of each ASBR.
- BGP ADD-PATH needs to be enabled on all IBGP sessions.
  - RRs need to be able to send multiple paths per prefix. The upper limit depends on:
    - The maximum number of ASBRs per site (say N).
    - Possibly also on the maximum number of EBGP sessions held by a single ASBR with single peer AS (say M), depending on BGP next-hop attribute (BGP-NH) configuration.
  - RR clients/ASBRs may need to be able to send multiple paths per prefix if BGP-NH configuration is "next hop unchanged". The upper limit depends on the maximum number of EBGP sessions held by a single ASBR with single peer AS (say M).

For further consideration the following network diagram will be used for reference:

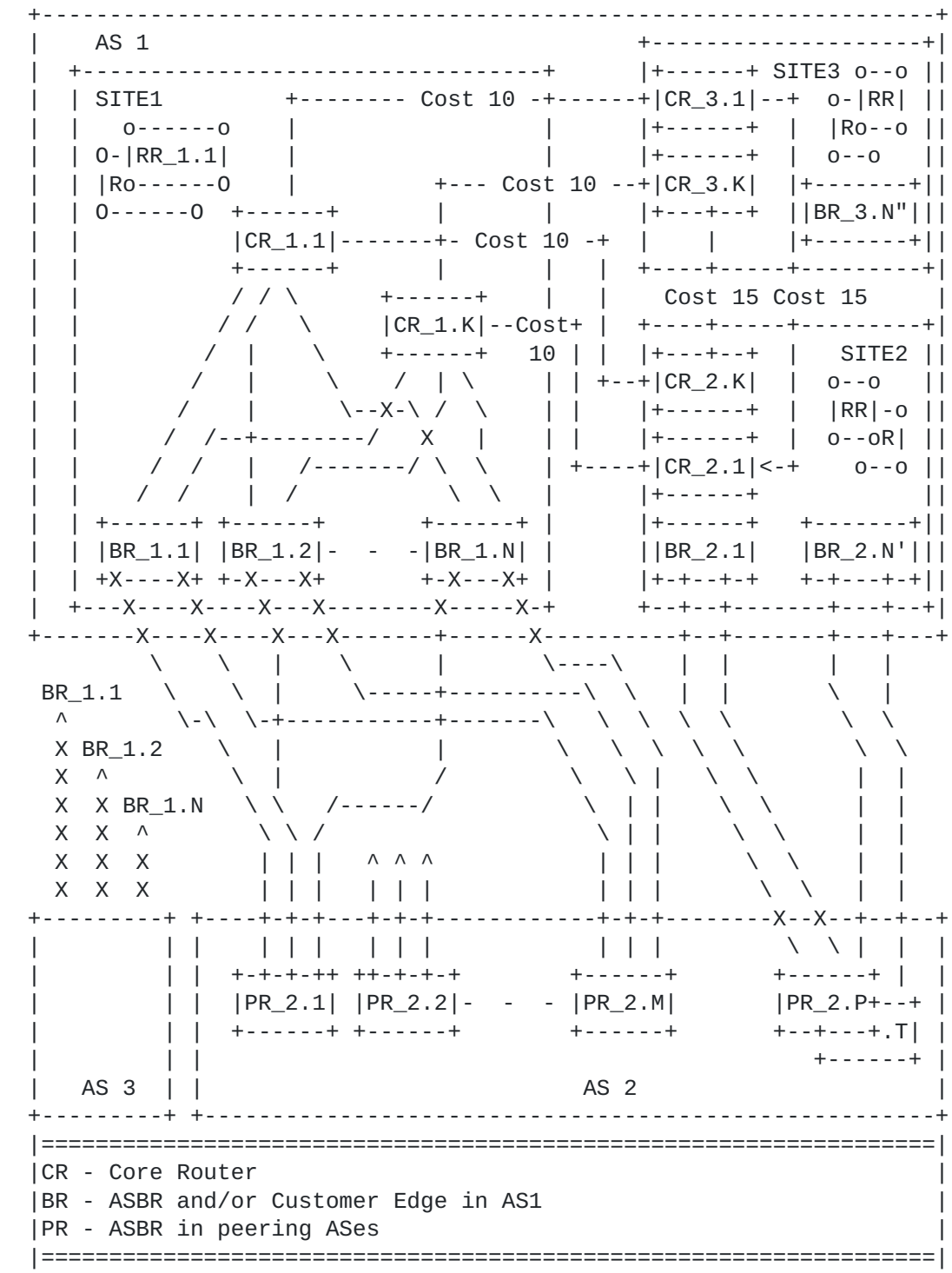
AS 1
SITE1
Cost 10
Cost 10
Cost 10
Cost 15 Cost 15
SITE3
SITE2
RR
BR_1.1
BR_1.2
BR_1.N
BR_3.N"
BR_2.N'
PR_2.1
PR_2.2
PR_2.M
PR_2.P
AS 3
AS 2
CR - Core Router
BR - ASBR and/or Customer Edge in AS1
PR - ASBR in peering ASes

Figure 2

1.2. Common BGP Deployment Configurations

1.2.1. IBGP with Next-Hop Unchanged

In one standard BGP configuration, an ASBR, when it advertises an externally learned prefix into IBGP, does not modify the BGP-NH. So, the BGP-NH is set to the IP address of an interface on the external peering router. The strength of this technique is the shorter time needed to restore connectivity with all equal cost multi-path (ECMP) in-use and on low latency paths. The drawback is extremely high BGP Routing Information Base (RIB) scale - proportional to the number of inter-AS links.

1.2.1.1. Example

Let's assume that in the network of Figure 2, all PR2.x of AS2 advertise the same set of prefixes on all sessions to AS1.

If BR1.1-BR1.N and BR2.1-BR2.N' each advertise only one path per prefix to their respective RRs, then as the result of ADD-PATH among RRs, BRs and CRs, at site 3 the BRs and CRs will learn N+N' paths per prefix learned from AS2. This is sufficient to equally distribute load among all N ASBRs on site 1 (note the IGP cost between site 2 and site 3).

However, when interfaces over which all BR1.1-BR_1.N learned their best path become unavailable (say interfaces to PR_2.1 in all cases, as a result of the failure of PR_2.1), the route to the BGP BGP-NH - that is, the IP address of the PR_2.1 interface - is removed from the IGP. BGP speakers at other sites (BR_3.x) will react by temporarily directing traffic to site 2 (BR_2.1-BR_2.N'). This switchover may happen in sub-second time, in a prefix-scale-independent manner, thanks to techniques commonly known as BGP PIC Edge [I-D.ietf-rtgwg-bgp-pic]. As a result, traffic is on a path other than the lowest cost path, as the connection from site 1 to AS2 is not entirely broken (links to PR_2.2-PR_2.M are operational).

Subsequently, all BR1.x will update their RRs with a new best path (say for PR_2.2) for each prefix (for example, 100,000 of them), triggering global convergence. Such a convergence, for a large number of prefixes, may take many minutes.

In the above example, BRs, RRs, and possibly CRs keep N+N' paths per prefix (N from site 1, and N' from site 2). Provided N=N'=4, this makes 8 path per prefix.

The solution for sub-optimal routing right after the failure would be to enable each BR to advertise multiple paths to its RRs, and for

them in turn to propagate it to all other RRs and hence BRs. So, each of BR1.x at site 1 will advertise M paths (from PR_2.1-PR_2.M), RR1.x will have N*M ECMP best paths and advertise them to other sites (site 3). As a result, BGP speakers at other sites (BR3.x at site 3) are provided with N*M paths per prefix from site 1 and N'*M' from site 2. Therefore to achieve optimal routing immediately after failure, a considerably higher scale of BGP paths needs to be handled. If M=N=N'=M'=4 then for each prefix we have 16 best paths and 16 non-best, a total of 32. If AS2 advertises 100,000 prefixes, this becomes 3.2M paths.

Although this solution provides a mean of fast, prefix-scale-independent traffic switchover, it does it only if an ASBR external interface goes down, which triggers an IGP event. In case an EBGP session fails but the underlying interface remains up (misconfiguration, software defect, etc), recovery still requires per-prefix withdrawal/update that could take many minutes at high scale.

1.2.2. IBGP with Next-Hop-Self

The other common technique is to modify BGP-NH to "self" (a local IP address, typically a loopback) when the BR advertises an externally learned path into IBGP. This technique allows the reduction of the number of paths per prefix, while keeping optimal forwarding - least cost and ECMP - in case of failure discussed above (e.g. PR_2.1 node failure). Actually, because IP addresses of BGP-NH as seen by other BGP speakers do not change in response to external failure events, and are resolvable by the IGP, there is no need to reprogram the Forwarding Information Base (FIB) at all. Unfortunately, other failures - loss of all connectivity between a single BR (say BR1.1) and a peer AS (all PRs in AS2) would not be handled quickly. As the BGP-NH advertised by BR_1.1 is not changed and is reachable by the IGP, BGP speakers in AS1 (BRs, CRs) will keep BR_1.1 as a feasible exit point until they receive BGP withdraws on a prefix-by-prefix basis. This is a global convergence process that at high scale can take minutes, during which time packets may be discarded or loop.

2. The BGP Abstract Next-Hop

The Abstract Next Hop (ANH) concept presented below does not require any changes to the BGP protocol itself. It is architectural solution to network configuration, that uses existing protocols' capabilities while achieving higher scale and faster routing convergence when scale-out peering sites exist.

When a BGP speaker advertises a path to its IBGP peer, it modifies the Protocol Next-Hop to be the ANH value. The ANH is just an IP

address that identifies the BGP session or a set of BGP sessions. The set of BGP sessions is defined by the operator in local configuration, according to network design needs. For example, an ANH might identify:

- a set of BGP sessions with the same peer AS and handled by a given single ASBR
- a set of BGP sessions with same the peer AS and handled by one or more ASBRs at a given site
- a set of BGP sessions with any upstream provider AS
- a set of BGP sessions with a given peer device and handled by one or more of ASBRs of the local AS

A host route to the ANH is installed in the relevant RIB and redistributed into the IGP. BGP maintains the ANH host route based on the state of the associated group of BGP sessions:

- As soon as all BGP sessions in the set go down, the ANH route is removed.
- When at least one BGP session in of the set comes up, the ANH route is created only after initial route convergence is complete for the peer (End-of-RIB (EoR) [RFC4724] is received).

Taken together, these procedures ensure that as soon as the final session in the set goes down, ingress routers will see the associated ANH withdrawn from the IGP. Since the ANH is used to resolve the associated BGP next hops, the ingress routers are triggered to converge to send traffic to their alternate (new best) route. They also ensure that as soon as one session in the set comes up and is synchronized (that is, the EoR is received), ingress routers will see the ANH advertised in the IGP and will be able to reconverge to use routes that are associated with that next hop.

The ANH can be any IP address that the router is eligible to advertise according to the local network's IP address management scheme. More details are given in Section 3.3.

3. Use of Abstract Next-Hop in scale-out peering design

In traditional configurations as described in Section 1.2 the meaning of the BGP-NH is either:

- An egress interface in the case of next-hop-unchanged configuration, or

o An egress ASBR in the case of next-hop-self configuration.

The meaning of Abstract Next Hop is more context-dependent. This document describes network configurations when the BGP-NH identifies:

a. An (egress ASBR, peer AS) pair. The ANH should be advertised into the IGP if, and only if, the given egress ASBR has at least one EBGP session in the ESTABLISHED state with the given peer AS, and the EoR marker has been received on that session. We call this the ASBR-Peer AS Abstract Next Hop (AP-ANH).

b. An (egress site in local AS, peer AS) pair, where a "site" may include multiple ASBRs. The ANH should be advertised into the IGP if, and only if, at least one ASBR of the given site has at least one EBGP session in the ESTABLISHED state with the given peer AS, and the EoR marker has been received on this session. We call this the Site-Peer AS Abstract Next Hop (SP-ANH).

Note that reachability of the ANH address in the IGP depends on EBGP session state and not inter-AS interface state, although of course, interface state may impact session state. How the IP route to the ANH address is instantiated on an ASBR and inserted into the IGP on particular device is a matter of local implementation.

3.1. Egress ASBR-Peer AS Abstract Next Hop (AP-ANH)

The AP-ANH is unique to an ASBR and its peer AS. For example, in the network of Figure 2, BR_1.1 would have two AP-ANH assigned - one for its peering with AS2 and the other for AS3. Similarly, BR_1.2 would have two AP-ANH, one per peer AS, with values different from the AP-ANH of BR_1.1, and so on. All AP-ANH are exported into the IGP by their ASBRs. Each ASBR advertises only one path per prefix to its RR, with the BGP-NH set to the appropriate AP-ANH. The RR will propagate it through the entire AS by means of IBGP ADD-PATH. In consequence, the number of paths learned per prefix is equal to number of ASBRs servicing a given peer AS. In the network as of Figure 2, for AS2 prefixes, this would be N+N' (from site_1 + from site_2) paths per prefix. This sets the scale requirements of this solution to be on par with Next-Hop-Self (Section 1.2.2). However, thanks to the properties of ANH, more failures are covered by prefix-independent techniques, as withdrawal of the ANH from the IGP makes the BGP-NH unresolvable.

Provided that all ASBRs in a given site (site1 in Figure 2) receive the same routing information from their peer AS (AS2), in non-faulty conditions, one could consider setting the ANH value on all ASBRs the same. However, failure(s) can create situations when multiple ASBRs will have a session in ESTABLISHED state with a given peer AS, but

some prefixes would be learned from EBGP only on a subset of these ASBRs. To prevent problems from arising in this situation, the per-ASBR AP-ANH needs to be advertised into the IGP and ASBRs need to set it as the BGP-NH when advertising routes to the site's Route Reflectors. However, for IBGP path advertisement being propagated beyond the site (into the RR mesh), the BGP-NH may be replaced by another ANH value, the Site-Peer AS ANH.

3.2. The Site-Peer AS Abstract Next Hop (SP-ANH)

The AP-ANH works on an ASBR level. From a given local AS perspective, the number of ANH is proportional to the number of pairs of ASBRs and ASes each of them peers with. With hundreds of peer ASes, tens of sites and ~10 ASBRs per site, the number of AP-ANH may scale into the thousands. At the same time, it may not be necessary or even desirable for every BGP speaker in the network to have visibility to every path down to individual egress ASBR granularity. With symmetrical multiplane backbone and/or leaf-spine designs, it is sufficient that BGP speakers on other sites have information that a given site (site1 in Figure 2) has at least one ASBR with an ESTABLISHED session to the peer AS (AS2). For example, in the network of Figure 2, even if BR3.1 has only one path with its BGP-NH equal to the ANH of BR1.1, BR3.1 resolves the BGP-NH in the IGP and spreads traffic among all CRs on site 3. Thus, traffic will be delivered to CR1.x at site 1. As long as CR1.x has visibility to all paths, traffic will be distributed equally to all site 1 ASBRs.

At the same time, when multiple paths are available on BGP speakers, every change is propagated, with consequent transmission and processing costs on all BGP speakers across the network. This will be true even if the route change doesn't impact the forwarding plane. For example, in the network of Figure 2, even if BR3.1 has N paths with BGP-NHs set to the ANHs of BR1.1 through BR1.N, BR3.1 will resolve those BGP-NHs in the IGP and spread traffic among all CRs of site 3. When one of the egress ASBRs (say BR1.2) loses its connectivity to the peer AS, the affected BGP routes (those with BGP-NH equal to AP-ANH of BR1.2) are withdrawn from all BGP speakers (e.g. BR3.1) of the network. All BGP speakers perform path selection and possibly update their forwarding data structures. Since the actual forwarding paths do not change, all this work represents unnecessary churn.

To avoid the above drawbacks, the RR of a given site (site1 in Figure 2), when re-advertising a BGP path learned from its ASBR client, modifies the BGP-NH to another abstract value - the Site-Peer AS Abstract NH (SP-ANH). This value is unique per (site, peer AS) pair, and is shared by all RRs of a given site. With this modification, it is sufficient that inter-site IBGP sessions carry

only one path per prefix (no ADD-PATH needed). Consequently, BGP RIB scale is reduced significantly. This frees up memory, reduces the amount of data RRs need to exchange, and mitigates churn. The BGP speakers in other sites of AS 1 need to resolve SP-ANH in order to build their local FIBs. Therefore SP-ANH have to be present in the IGP - some router(s) in the local site (RR, ASBR or CR) need to inject it into the IGP. While the selection of role that is responsible of SP-ANH injection is discussed below, in any case, the SP-ANH should be reachable in the IGP if, and only if, at least one of AP-ANH (for the same peer AS and ASBR belonging to given site) is reachable. Figure 3 illustrates routing information flow in a network such as that of Figure 2:

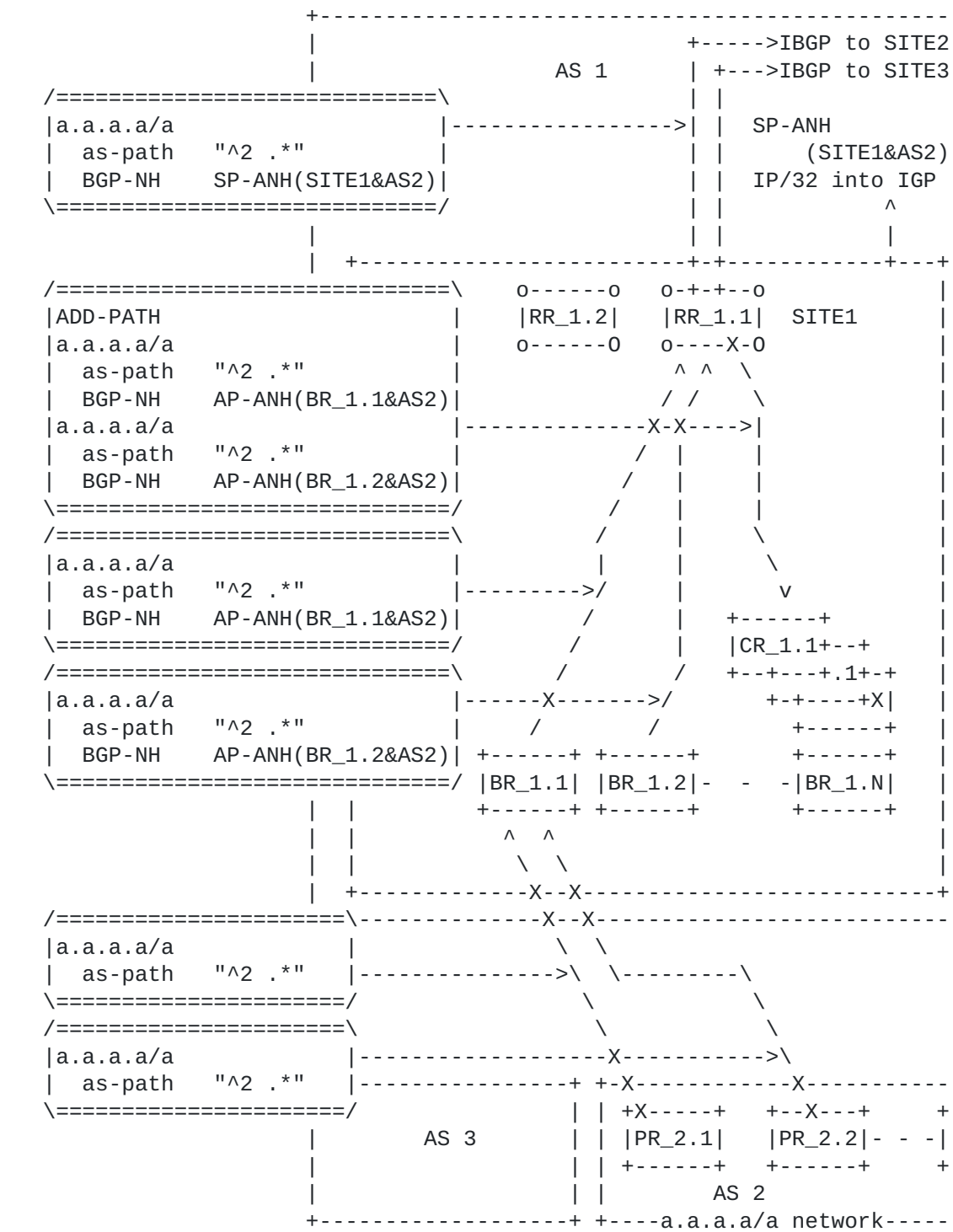
AS 1
SP-ANH
(SITE1&AS2)
IP/32 into IGP
IBGP to SITE2
IBGP to SITE3
a.a.a.a/a
as-path "^2 .*"
BGP-NH SP-ANH(SITE1&AS2)
ADD-PATH
AP-ANH(BR_1.1&AS2)
AP-ANH(BR_1.2&AS2)
RR_1.2
RR_1.1
SITE1
CR_1.1
BR_1.1
BR_1.2
BR_1.N
AS 3
PR_2.1
PR_2.2
AS 2
a.a.a.a/a network

Figure 3

3.3. Assignment of Abstract Next Hops

In the following subsections we provide more details of how abstract next hops can be injected in several different common network architectures.

3.3.1. Native IP Networks

In this network every router, including core routers, has full BGP routing information and forwards each packet based on destination IP lookup. Provided that all routers at an egress site receive multiple paths with BGP-NH set to AP-ANH (and not SP-ANH), it is a matter of the operator's decision which node - RR, ASBR or CR - will inject the SP-ANH route into the IGP. One may argue that injection of SP-ANH by ASBRs may be simpler, as it will be done by the same procedure and policy as injection of AP-ANH. Others may prefer injection at RR, as it limits the number of configuration touch-points.

3.3.2. MPLS

3.3.2.1. Identical BGP address space and paths received on all ASBRs

In the MPLS network, since traffic is carried over LSP tunnels, the SP-ANH needs to be injected into the IGP by a node that has the ability to perform an IP lookup. This eliminates the RR, and possibly CRs (in "BGP-free core" architectures). Instead, all ASBRs are used to insert SP-ANH addresses into the IGP. In case of LDP-based networks, this is sufficient. The CR will create an ECMP forwarding structure for labels of SP-ANH FEC coming from other sites. In RSVP-TE based networks, ECMP needs to happen on the ingress LSR and therefore, every BGP speaker needs to establish an LSP to every ASBR, and the SP-ANH address needs to be part of the FEC for its respective LSP. If SP-ANH is used as an RSVP (signaling) destination, some other means (such as affinity groups) needs to be used to ensure the desired 1:1 LSP to egress ASBR mapping.

3.3.2.2. Different address space sets or paths received on different ASBRs

In the case when the set of prefixes received from a given peer AS by one ASBR is different from the set received by another one, a combination of SP-ANH and MPLS-based load balancing on a CR may lead to a situation where an IP packet will be directed to an ASBR that lacks external routing information and hence can't forward traffic directly out of the AS. Similarly, if path attributes for a given prefix received by one ASBR are different from those received by another, again packets can be directed to the "wrong" ASBR. In this case the ASBR would use the IBGP route it learned from another ASBR

of the same site (via RR, with AP-ANH) and forward traffic over an LSP to the "correct" ASBR. This extra hop constitutes a sub-optimal traffic path through the network.

For example in the network of Figure 2, let's assume that prefix P2 is advertised to BR1.2-BR1.N by AS2 but not to BR1.1. BR3.1 has a BGP best route to P2 with its BGP-NH set to the SP-ANH of (site1, AS2). It resolves it by ECMP over N MPLS LSPs, terminating on BR1.1-BR1.N. So, some packets are forwarded by BR3.1 over an LSP via CR1.x and terminated on BR1.1. BR1.1 has no external route to P2, but it has (N-1) IBGP routes to P2 w/ BGP-NHs equal to the AP-ANHs of BR1.2-BR1.N. Therefore BR1.1 performs an IP lookup and forwards this packet over LSPs via CR1.x and terminated on BR1.2-BR1.N. Traffic is U-turned on BR1.1 and traverses CRs at site 1 twice.

Such asymmetry may be considered acceptable by the provider, as long as it's a transient condition. However, in the general case such a situation could be persistent, as the result of intentional configuration on the peer AS's ASBRs. Therefore the better solution would be to insert the SP-ANH into the IGP on CRs. In this case, CRs need to perform forwarding based on destination IP lookup. Therefore CRs would have to be able to learn and handle large IP routing and forwarding tables - at least all prefixes learned from peer ASes by the local ASBRs.

3.3.3. SPRING

3.3.3.1. Identical BGP address space and path received on all ASBRs

For SPRING based networks, we can take advantage of the unique capability of Anycast-SID [RFC8402]. The ASBRs of a single site allocate an Anycast-SID for each SP-ANH address. This SID can be used as the only SID by an ingress BGP speaker or, if a TE routed path is desired, depending on TE constraints, the TE controller can provision a SPRING path with the Anycast-SID at the end, instructing the CR to perform load balancing among connected ASBRs.

3.3.3.2. Different address space sets or paths received on different ASBRs

Similarly to a classic MPLS environment, such a situation may lead to suboptimal routing (redirecting from one ASBR to another), or may require the CR (instead of ASBR) to insert the SP-ANH into the IGP and generate a PREFIX-SID (or Anycast-SID if there is more then one CR) for it.

4. Worked Examples

Below we illustrate the operation of the proposal by working through its operation in the context of several different types of failures. Here, we assume that each ASBR in a given site of the local AS (site 1 of AS1 in Figure 2), that has an EBGP session with the given peer AS (AS2 in Figure 2), receives from its peer routers (PR2.x) routes to exactly same address space on each session.

4.1. Failure of a proper subset of EBGP sessions with a given peer AS on a single ASBR

- The impacted ASBR keeps advertising the AP-ANH into the IGP, as at least one session to the peer AS remains in the ESTABLISHED state.
- The impacted ASBR may send UPDATEs to RRs, however the BGP-NH remains the same and equal to the pre-failure AP-ANH.
- The RRs may send UPDATEs to their clients (CRs, BRs) and to RRs in other sites, however the BGP-NH remains the same as its pre-failure value: AP-ANH and SP-ANH respectively.
- As BGP-NH do not change, there are no changes in forwarding data structures (FIB) on any BGP speaker across the network, except possibly the ASBR that holds the impacted session.

4.2. Failure of a proper subset of EBGP sessions with a given peer AS on each ASBR of a given site

- The impacted ASBRs keep advertising the AP-ANH into the IGP, as at least one session to the peer AS remains in the ESTABLISHED state on each ASBR.
- The impacted ASBRs may send UPDATEs to RRs, however the BGP-NH remains the same and equal to the pre-failure AP-ANH.
- The RRs may send UPDATEs to their clients (CRs, BRs) and to RRs in other sites, however the BGP-NH remains the same and equal to its pre-failure value: AP-ANH and SP-ANH respectively.
- As BGP-NH do not change, there are no changes in forwarding data structures (FIB) on any BGP speaker across the network, except possibly the ASBRs that hold the impacted sessions.

4.3. Failure of all EBGP sessions with a given peer AS on single ASBR; Failure of a single ASBR

- The impacted ASBR stops advertising the AP-ANH into the IGP, as it has lost all sessions with given peer AS.
- The SP-ANH is kept reachable in the IGP.
- All other BGP speakers at the impacted site invalidate all paths with BGP-NH equal to the AP-ANH. This may trigger prefix-independent FIB data-structure patching/temporary fixing for sub-second traffic restoration.
- The impacted ASBR sends WITHDRAWs to its RRs.
- Each RR:
  * Sends WITHDRAWs to its clients at the local site (CRs, BRs) for paths from the impacted ASBR. As these sessions support ADD-PATH, paths from other ASBRs will remain. Other BGP speakers at this site have to modify their FIBs.
  * May send UPDATEs to RRs in other sites, however the BGP-NH remains the same, equal to the pre-failure SP-ANH. As the BGP-NH does not change, there are no changes in forwarding data structure (FIB) on any of BGP speakers across network, except those at the impacted site.
- Routing churn is mitigated in many cases to a single peering site, and does not propagate across the network. FIB changes are limited to a single peering site, and do not propagate across the network.

4.4. All EBGP sessions with a given peer AS on all ASBRs

- Each ASBR stops advertising its AP-ANH into the IGP, as it has lost all sessions with the given peer AS.
- The SP-ANH is no longer reachable in the IGP, as none of AP-ANH are reachable.
- All other BGP speakers across the network invalidate all paths with a BGP-NH equal to the removed AP-ANH or SP-ANH. This may trigger prefix-independent FIB data-structure patching/temporary fixing for sub-second traffic restoration.
- Each impacted ASBR sends WITHDRAWs to its RRs.

- The RRs send WITHDRAWs to their clients at the local site (CRs, BRs) and RRs in other sites for paths from the impacted ASBRs. As these sessions support ADD-PATH, paths from ASBRs at other sites will remain. The BGP speakers across the network may need to modify their FIBs.

5. Acknowledgements

Valuable comments and suggestions on solution covered by this document was provided by Mannan Venkatesan, John Scudder and Ron Bonica. Special thanks to John Scudder, who also helped with editorial changes.

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

Since this is a deployment architecture and not a protocol modification, it doesn't introduce any new issues to the BGP protocol itself. General BGP security considerations are discussed in [RFC4271] and [RFC4272], BGP deployment best practices are documented in [RFC7454], and nothing in this proposal impedes their use. Many of the practices recommended in that document are self-evidently still applicable, for example the use of cryptographic session protection methods such as TCP MD5 [RFC2385] or the TCP Authentication Option [RFC5925], and the Generalized TTL Security Mechanism [RFC5082]. Since we propose a novel use of IP addresses to assign ANHs, it's worth considering if anything new is required to protect them. We conclude there isn't, they fall into the existing category of "Prefixes Belonging to the Local AS" discussed in section 6.1.4 of [RFC7454].

8. Informative References

[I-D.ietf-rtgwg-bgp-pic]
Bashandy, A., Filsfils, C., and P. Mohapatra, "BGP Prefix Independent Convergence", draft-ietf-rtgwg-bgp-pic-08 (work in progress), September 2018.

[RFC2385] Heffernan, A., "Protection of BGP Sessions via the TCP MD5 Signature Option", RFC 2385, DOI 10.17487/RFC2385, August 1998, <https://www.rfc-editor.org/info/rfc2385>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4272] Murphy, S., "BGP Security Vulnerabilities Analysis", RFC 4272, DOI 10.17487/RFC4272, January 2006, <https://www.rfc-editor.org/info/rfc4272>.

[RFC4456] Bates, T., Chen, E., and R. Chandra, "BGP Route Reflection: An Alternative to Full Mesh Internal BGP (IBGP)", RFC 4456, DOI 10.17487/RFC4456, April 2006, <https://www.rfc-editor.org/info/rfc4456>.

[RFC4724] Sangli, S., Chen, E., Fernando, R., Scudder, J., and Y. Rekhter, "Graceful Restart Mechanism for BGP", RFC 4724, DOI 10.17487/RFC4724, January 2007, <https://www.rfc-editor.org/info/rfc4724>.

[RFC5065] Traina, P., McPherson, D., and J. Scudder, "Autonomous System Confederations for BGP", RFC 5065, DOI 10.17487/RFC5065, August 2007, <https://www.rfc-editor.org/info/rfc5065>.

[RFC5082] Gill, V., Heasley, J., Meyer, D., Savola, P., Ed., and C. Pignataro, "The Generalized TTL Security Mechanism (GTSM)", RFC 5082, DOI 10.17487/RFC5082, October 2007, <https://www.rfc-editor.org/info/rfc5082>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC7454] Durand, J., Pepelnjak, I., and G. Doering, "BGP Operations and Security", BCP 194, RFC 7454, DOI 10.17487/RFC7454, February 2015, <https://www.rfc-editor.org/info/rfc7454>.

[RFC7911] Walton, D., Retana, A., Chen, E., and J. Scudder, "Advertisement of Multiple Paths in BGP", RFC 7911, DOI 10.17487/RFC7911, July 2016, <https://www.rfc-editor.org/info/rfc7911>.

[RFC8402] Filsfils, C., Ed., Previdi, S., Ed., Ginsberg, L., Decraene, B., Litkowski, S., and R. Shakir, "Segment Routing Architecture", RFC 8402, DOI 10.17487/RFC8402, July 2018, <https://www.rfc-editor.org/info/rfc8402>.

Authors' Addresses

Rafal Jan Szarecki (editor)
Juniper Networks Inc.
1133 Innovation Way
Sunnyvale, CA 94089
US

Phone: +1(408)680-9604
Email: rafal@juniper.net

Kaliraj Vairavakkalai
Juniper Networks Inc.
1133 Innovation Way
Sunnyvale, CA 94089
US

Phone: +1(408)936-8872
Email: kaliraj@juniper.net

Natrajan Venkataraman
Juniper Networks Inc.
1133 Innovation Way
Sunnyvale, CA 94089
US

Phone: +1(408)936-6597
Email: natv@juniper.net