Interdomain Routing (IDR)

IETF-88, Vancouver

November 8, 2013

Note Well

This summary is only meant to point you in the right direction, and doesn't have all the nuances. The IETF's IPR Policy is set forth in BCP 79; please read it carefully.

The brief summary:

- **By participating with the IETF, you agree to follow IETF processes.**
- **If you are aware that a contribution of yours (something you write, say, or discuss in any IETF context) is covered by patents or patent applications, you need to disclose that fact.**
- **You understand that meetings might be recorded, broadcast, and publicly archived.**

For further information, talk to a chair, ask an Area Director, or review the following:

BCP 9 (on the Internet Standards Process)

BCP 25 (on the Working Group processes)

BCP 78 (on the IETF Trust)

BCP 79 (on Intellectual Property Rights in the IETF)

Document Status

- The dog ate my homework
  - Well OK, my laptop has been broken all week
- Thus the chairs will send a document status update to the list next week.

Agenda

- Administrivia (Chair) 10 minutes
- draft-patel-raszuk-bgp-vector-routing-00 (Keyur Patel) 10 minutes
- draft-ietf-idr-aigp-10 last call issues (Eric Rosen) 15 minutes
- draft-ietf-idr-add-paths-09 (Jeff Haas) 10 minutes
- draft-haas-idr-flowspec-redirect-rt-bis-00 (Jeff Haas) 5 minutes
- draft-ietf-idr-sla-exchange (Shitanshu Shah) 5 minutes
- draft-wu-idr-te-pm-bgp-03 (Qin Wu) 5 minutes
- draft-li-idr-cc-bgp-arch-00 (Lizhenbin (Robin)) 10 minutes

BGP Vector Routing

draft-patel-raszuk-bgp-vector-routing-01

Keyur Patel, Robert Raszuk, Burjiz Pithawala, Ali Sajassi, Eric Osborne, Jim Uttaro, Luay Jalil

IETF 88, November 2013, Vancouver, Canada

Motivation

- Network Architectures require additional control over the traffic paths (Inter as well as Intra domain)
  - Need to force the traffic to go through one or more Transit Nodes
  - Transit Nodes could be a TE Node
  - Other examples include Service Nodes like: Firewall, NAT, Load Balancers, etc
- Need a **scalable control plane solution** to advertise "information" so that the traffic gets routed through an ordered set of Transit points before it is forwarded to its destination
  - In context of Transit points as Service Nodes it is known as "Service Chaining". Otherwise it is known as "Traffic Engineering" (TE)

BGP Vector Routing

- BGP based mechanism to create arbitrary forwarding topologies as well as facilitate Service Chaining
  - Does not require changes to the forwarding plane
  - Assumes use of an existing encapsulation/tunneling techniques to forward data
- New BGP attribute known as a Vector Node attribute
- Vector Node attribute consist of one or more TLVs
  - TLVs carry ordered lists of IP Transit Hops that needs to be traversed before the packet is forwarded to its destination
  - TLV information is used to replace the NEXTHOP information when installing the route in RIB/FIB
- Two new TLVs defined as part of this draft
  - Type 1 and Type 2 TLV
- Rules to process and use TLV information of Vector Node Attribute

BGP Vector Routing (Cont'd)

- BGP Vector Node attribute can be applied to any BGP Address Family
- Creation of a BGP Vector Node attribute is outside the scope of the document
  - Assumed to be created using CLI on a router or using an Orchestrated system, or by some automated SDN policy computing engines
- Vector Node attribute is usually inserted at a single point in the network and advertised by BGP to all BGP speakers

BGP Vector Node Attribute TLVs

- TYPE1 TLV consists of a Vector Node address
  - Vector Node address is an address of a transit (services) router and is typically announced in an IGP protocol
- TYPE2 TLV consists of a Vector Node and a Service Node address
  - Vector Node address is an address of a Transit Services router and is typically announced in an IGP protocol
  - Service Node address is an address of a Service Appliance and is directly connected to Vector Node address and is not announced in an IGP. Alternatively Service Node Address could be a Local ID of a Transit Service Router pointing to an Appliance
  - Vector Nodes and Service Nodes may belong to a different Address Families
- Both the TLVs carry AS Number to facilitate Inter-AS announcements

BGP Vector Node Attribute Rules

- 4 Rules defined to process the BGP Vector Node Attribute
- 1st Rule describes Vector Node attribute and AS Number Validation
  - Missing Attribute or a failing AS Number Validation results in use of a BGP address from BGP MP_REACH attribute or from a NEXT_HOP attribute (if BGP MP_REACH Attribute is NOT present) as a NEXTHOP address when adding a route to RIB/FIB
- 2nd Rule describes a case where an AS Number Validation succeeds but a BGP Speaker Address (loopback or connected) is missing in the Vector Node Attribute
  - In such a case BGP Speaker should use the First TLV Vector Node address as a NEXTHOP address when adding a route to RIB/FIB
- 3rd Rule describes a case where an AS Number Validation succeeds but a BGP Speaker Address (loopback or connected) is present in the Vector Node Attribute TLV
  - In such a case BGP Speaker should use the next eligible Vector Node address as a NEXTHOP address when adding a route to RIB/FIB

BGP Vector Node Attribute Rules (Con't)

- 4th Rule describes a case where an AS Number Validation succeeds but a BGP Speaker Address (loopback or connected) is present as the Last Vector Node Attribute TLV address
  - In such a case BGP Speaker should use the BGP address from BGP MP_REACH attribute or from a NEXT_HOP attribute (if BGP MP_REACH Attribute is NOT present) as a NEXTHOP address when adding a route to RIB/FIB

Questions?

Request WG to adopt the draft as a WG document.

AIGP Last Call Issues

- After almost 5 years, 5 implementations, and significant deployment, draft finally reaches WG last call
- So folks not directly involved read the draft for the first time
- Some interesting issues raised during LC, some controversy about how to address those issues
- Some F2F discussion seems worthwhile before finalizing
- Note: no objections raised during LC to “meat” of draft, i.e. to rules for computing and using the value of the AIGP attribute (semantics)
- Objections raised to error handling, encoding, “leakage protection” at admin boundaries, i.e., stuff that might impact “somebody else”
- Want to focus discussion on LC issues …

AIGP

- BGP Path Attribute: Accumulated IGP Metric of path to prefix
- Allows IGP metric to be major determinant of bestpath selection for BGP-distributed internal routes
  - Provisioning determines the set of prefixes to which AIGP gets attached
  - BGP becomes a sort of IGP for those prefixes
- **Must not leak** out past administrative boundary
  - **Not** an inter-provider metric
  - AIGP is **non-transitive** attribute, discarded when not recognized
  - By default, even if recognized, AIGP treated as unrecognized (discarded) on EBGP sessions
    - All admin boundaries are EBGP sessions (converse not true)
- For possible future expansion, attribute coded as list of TLVs, but only type 1 *(IGP distance)* defined

Error Handling for Malformed AIGP Attribute

- Not clearly specified in draft
- What's best: *treat as withdraw,* or *discard attribute?*
- *Treat as withdraw* is default for attributes affecting bestpath selection
  - But AIGP is only to be used in scenarios where there is tunneling to the next hop; complete consistency not needed
- *Discard attribute* is therefore less disruptive way to handle malformed attribute
- *Discard attribute* is also very like what is done with an unrecognized transitive attribute
- Proposed resolution: use *discard attribute* as error handling method

Can the Non-Transitivity Break?

- R1---(ibgp)---ASBR1----(ebgp)----ASBR2
- AS containing ASBR2 uses AIGP
  - ASBR2 mistakenly sets the transitive bit on the AIGP attribute
  - ASBR2 mistakenly sends AIGP attribute to ASBR1
- ASBR1 does not understand attribute, sees transitive bit, forwards to R1 when really the attribute ought to be discarded
- R1 understands AIGP attribute and is provisioned to use it.
  - But now it mistakenly has received the attribute from across an admin boundary
  - Should R1:
    - Clear the transitive bit and forward the attribute (local repair)? Or
    - *Discard attribute* as malformed
  - Proposed resolution: *discard attribute* as malformed
    - Attribute isn't supposed to be processed by R1 or forwarded any further
    - Restores the proper non-transitive behavior

TLV Encoding Issues

- Length field not specified “correctly”, shouldn’t include length of type and length fields
  - Too late
  - Sorry ☹
- What if attribute contains multiple type 1 TLVs?
  - Is this malformed, or should one of the type 1 TLVs be used and the others ignored?
  - Proposed resolution: do not treat as malformed, use the first one.
  - Other TLV types to be ignored if not recognized, of course.

Disabled By Default

- Default per-session settings:
  - Do not originate routes with AIGP
  - On EBGP sessions, discard attribute if received
  - So:
    - On EBGP sessions, attribute shouldn't pass unless enabled on both sides
    - On IBGP sessions, attribute will pass if enabled on one side
  - Enough protection against leakage?
    - Think so; but controversial on mailing list.
  - Enough protection against errors?
    - Can't protect against all errors

Capability Needed?

- Capability needed?
  - No, shouldn't need a capability for every new (optional) attribute

Advancing add-path

Jeffrey Haas, et seq.
jhaas@juniper.net

add-path current status

- The base BGP add-path feature is well deployed and interoperable at this point:
  - Alcatel-Lucent
  - Cisco
  - Juniper
  - (and probably others…)

add-path concerns

- During the development of the add-path feature, there were a number of concerns about how the feature would behave from a route-selection standpoint.
- Those issues are much better understood these days. Many are documented in draft-ietf-idr-add-paths-guidelines.

eBGP and add-path

- draft-pmohapat-idr-fast-conn-restore is currently a NORMATIVE reference in the base add-path document.
- The Edge_Discriminator Path Attribute documented in that I-D is required for BGP to perform consistent path selection for eBGP routes distributed in Add-Path.
- There are no implementations of this feature?

Advancing add-path

- Operators are clearly seeing benefit from the add-path feature, even without the Edge_Discriminator feature.
- Introducing that feature has the usual incremental BGP deployment pain points.
- Should the feature be removed as a normative reference so the add-path feature can advance and get published as an RFC?

Discussion

draft-haas-idr-flowspec-redirect-rt-bis

Jeffrey Haas, Ed.
jhaas@juniper.net

RFC 5575 Redirect Extended Community

“Redirect: The redirect extended community allows the traffic to be redirected to a VRF routing instance that lists the specified route-target in its import policy. If several local instances match this criteria, the choice between them is a local matter (for example, the instance with the lowest Route Distinguisher value can be elected). **This extended community uses the same encoding as the Route Target extended community** [RFC4360].”

The Issue

- A Route Target is not only the 6 bytes of Value field but uses the Type-high octet as a "format specifier".
- The Flowspec RFC only shows a single type/sub-type allocated: 0x8008
- This has lead several implementers to the conclusion that you simply try out all RT types using the Value field.
- *This is not how the feature is deployed.*

The fix

- A small draft updating RFC 5575 noting that the type field for the Redirect Extended Community is used the same as the Route Target extended Community, just ORed with 0x80.
- IANA is requested to update its registry to make the appropriate allocations.

Inter-domain SLA Exchange

http://www.ietf.org/id/draft-ietf-idr-sla-exchange-02.txt

IETF 88, Nov 2013, Vancouver, Canada

Topics

- Take-away from IETF 86 (including feed-back from tsvwg)
- Changes since IETF 86
- Implementation Report
- Next Steps

Evaluate re-use of existing IANA types (This slide was presented at the IETF 86)

- RFC 5102 - IPFIX Information Element ids to represent Traffic Class (IANA Type = IPFIX Information Element Identifiers)
  - Re-use only Element Id + Abstract data-type
- RFC5575 – BGP Flow Specification (IANA Type = Flow Spec Component Types)
  - Limited set of traffic class
- RFC5975 – QSPEC Template (ref. QSPEC parameters)
  - Parameter ID IANA type
  - Limited set of traffic class
  - Some of the parameters are irrelevant to SLA

  Feed-back from tsvwg: look at RFC2212 as a reference (RFC5975 inherits from)

Changes since IETF 86

- Re-use of IPFIX Element identifiers for Traffic Classifier Element [RFC5102]

- Rate profile using exactly same format as Tspec [RFC2212]

- Modification for proper and consistent use of Terminology

  Eg.,

  SLA parameter exchange is not same as establishing SLA

  Generalize terminology to support more use-case applicability

Implementation Report

- Implementation on multiple Cisco OS
  - Supports use-cases (section “Deployment Considerations”) described in the draft

- Details of implementation report and inter-operability at

  http://www.ietf.org/internet-drafts/draft-svshah-idr-sla-exchange-impl-00.txt

- Looking for more implementations

Next Steps?

BGP attribute for North-Bound Distribution of Traffic Engineering (TE) performance Metric

draft-wu-idr-te-pm-bgp-03

Qin Wu (sunseawq@huawei.com)
Danhua Wang (wangdanhua@huawei.com)
Stefano Previdi (sprevidi@cisco.com)
Hannes Gredler (hannes@juniper.net)
Saikat Ray (sairay@cisco.com)

Recap.

- TE performance related information is required by some external components(e.g.,ALTO server,PCE server)
  - TE Performance information includes network delay, jitter, packet loss, bandwidths.
  - PCE Server can use network performance info as constraint for end to end path computation
  - ALTO server can gather and aggregate these dynamic network performance information and use these info to decide which endpoint to connect.

- TE performance can be hard to gather via ISIS or OSPF or need to gather using other means in some cases
  - Inter-AS PCE computation
  - Hierarchy of PCE
  - BGP
  - NMS/OSS

  ……

- A new general mechanism is needed to collect and distribute TE performance information
  - draft-ietf-idr-ls-distribution describes a mechanism to distribute link state and TE information using BGP
  - This draft uses BGP to share additional TE performance related information to external components beyond linkstate and TE information contained in [I-D.ietf-idr-ls-distribution]

New BGP TLV attribute for TE performance info

- [I-D.ietf-idr-ls-distribution] defines new BGP path attribute (BGP-LS attribute) to carry link, node, prefix properties.
- This draft reuses existing BGP-LS attribute and defines 7 new TLVs that can be announced as BGP-LS attribute used with link NLRI.
- These BGP TLVs populate the following network performance information:
  - Unidirectional Link Delay
  - Min/Max Unidirectional Link Delay
  - Unidirectional Delay Variation
  - Unidirectional Packet Loss
  - Available bandwidth
  - Unidirectional Residual Bandwidth
  - Unidirectional Available Bandwidth
  - Unidirectional Utilized Bandwidth
- These network performance information carried in BGP TLV is same as one In IS-IS Extended Reachability TLV [**I.D-ietf-isis-te-metric-extensions-00**]
- The format and semantics of the 'value' fields in these BGP TLVs is same as one defined as sub TLV of IS-IS Extended Reachability TLV.

Update after IETF 87

- Complimentary to [I-D.ietf-idr-ls-distribution]
- Changes compared to (v-01)
  - Remove new metric ‘channel throughput’ from this draft based on discussion with ISIS-TE-extension draft authors
  - Move new metric ‘link utilization’ to [**I.D-ietf-isis-te-metric-extensions-01**] and define it as ‘unidirectional utilized bandwidth’ Sub TLV of IS-IS Extended Reachability TLV
  - Change metric name and add “Min/Max Unidirectional Link Delay ” as a new metric to get inline with [I.D-ietf-isis-te-metric-extensions-00] .
  - Add ‘unidirectional utilized bandwidth’ as seventh metric carried in new BGP TLV.
  - Add ' Anomalous ' bit in the BGP TE performance TLV to indicate whether performance is in steady state.

- Thanks Hannes for arranging a offline discussion after Berlin meeting with ISIS-TE-extension authors on why two additional attributes should be added into IGP draft.

- New coauthors
  - Stefano Previdi
  - Hannes Gredler
  - Saikat Ray

Next Step

- Any comments?
- Request WG adoption

An Architecture of Central Controlled Border Gateway Protocol (BGP)

draft-li-idr-cc-bgp-arch-00

Zhenbin Li, Mach Chen, Shunwan Zhuang
Huawei Technologies

IETF 88, Vancouver, BC, Canada

Introduction

- As the Software Defined Networks (SDN) solution develops, BGP is extended to support central control.
- This document introduces an architecture of using BGP for central control.
- Some use cases under this new framework are also discussed. For specific use cases, making necessary extensions in BGP are required.

Architecture -- Reference Model

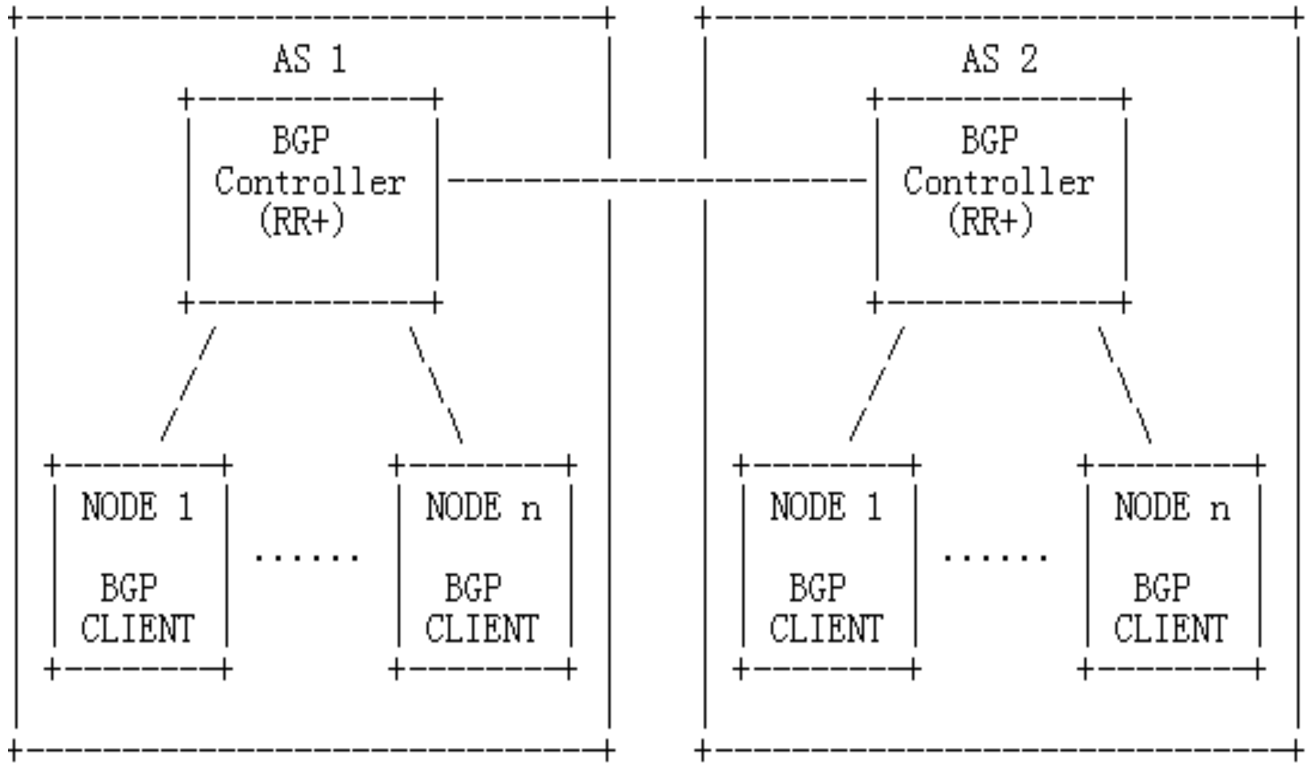

Figure 1: An Architecture of Central Controlled BGP

- BGP Controller controls all the BGP Clients within its administrative domain by communicating with them.
- BGP sessions are also set up between multiple BGP controllers.

Architecture -- Deployment Mode

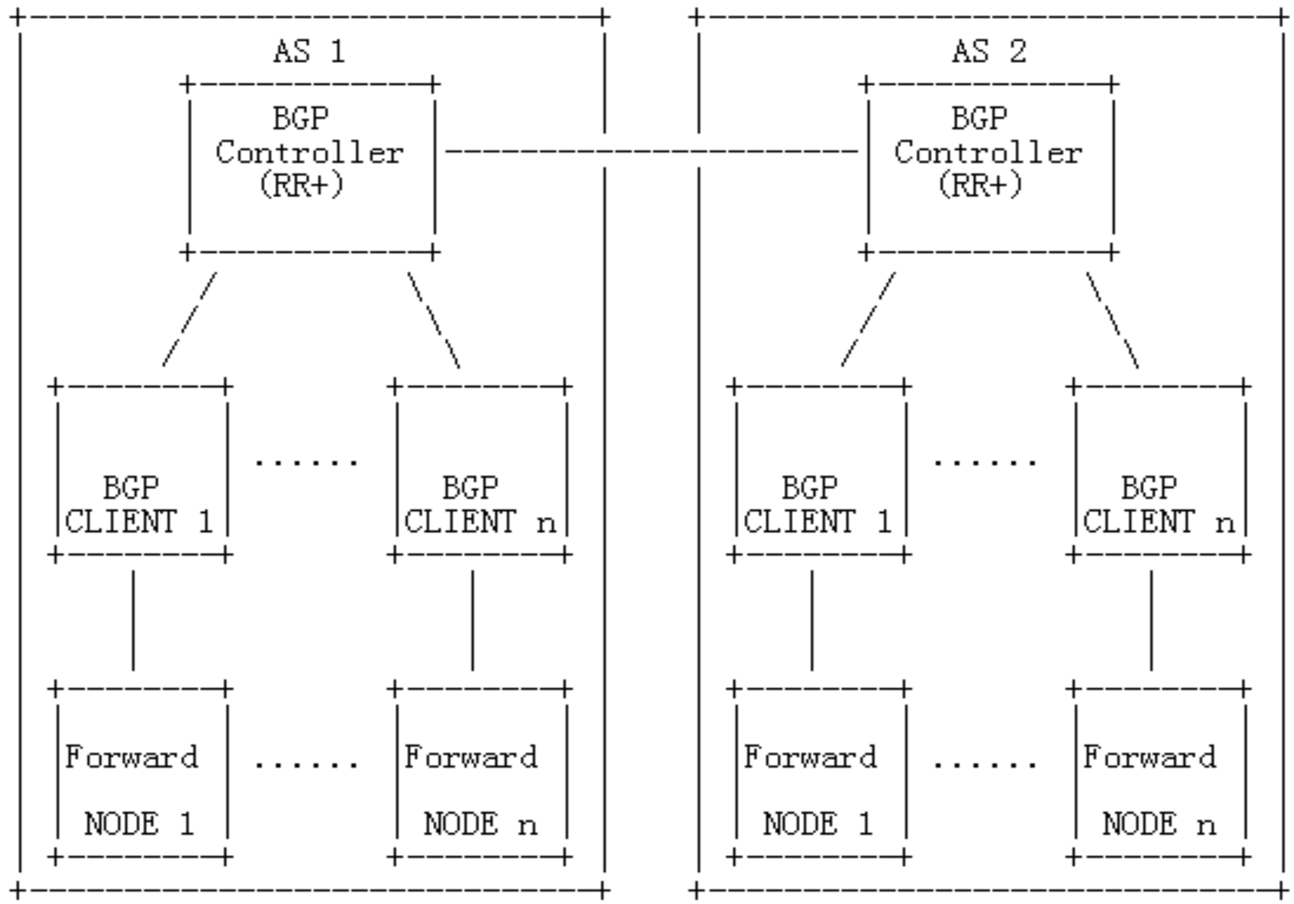

Figure 2: Decoupling BGP Client and Forwarding

- BGP Controller and BGP Client can run on a general-purpose server or a network device.
- It is more meaningful to decouple control plane and forwarding functionality on BGP Client because this manner enables network devices focusing on forwarding functionality.

Architecture -- Protocol Extensions

- Building Connectivity:
  - Connectivity between BGP Controller and BGP Clients in an AS can be built by extending IGP protocol.
  - In order to simplify network operations, such connectivity SHOULD be automatically established.
- Roles Auto-Discovery:
  - BGP Controller and BGP Client roles can be auto-discovered by extending IGP protocol to flooding the role information within an AS.
  - When IGP has finished the flooding process of role information, BGP Controller and BGP Client can establish a BGP session on demand.
- Capability Negotiation:
  - In order for BGP Controller and BGP Client to support BGP-based Central Controlled framework in a friendly way, this document suggests to defines a new BGP Central Control Capability.
- High Availability:
  - To void one-point-failure of BGP Controller, it is possible to run redundant BGP Controllers for high availability.
- Security

Use Cases

In BGP-based Central Controlled framework, new use cases are emerging：

- Network Topology Acquirement
  - BGP has been extended to distribute link-state and traffic engineering information.
- Simplifying Network Operation and Maintenance
  - By using I2RS APIs, it would allow network operator to setup BGP policy configuration and apply route policy easily from an central point.
  - In the new Central Controlled framework, VPN Service can be deployed rapidly according to business requirements. More detailed description could be found in [draft-li-l3vpn-instant-vpn-arch-00].

Use Cases(Cont.)

- MPLS Global Label Allocation
  - MPLS Global Label should be allocated in a central point to guarantee all distributed network nodes can understand meaning of a specific global label in same.
  - The new BGP-based Central Controlled framework is particularly suitable to allocate MPLS Global Label for services deployed on the network edge nodes.
  - [draft-li-mpls-global-label-usecases-00] proposes the use cases: 1) Identification of MVPN/VPLS, 2) Local Protection of PE Node, 3) Segment-Based EVPN, etc.
- RR-Based Traffic Steering
  - RR-based Traffic Steering (RRTS) defined in [draft-chen-idr-rr-based-traffic-steering-usecase-00], is an idea that leverages the BGP route reflection mechanism to realize traffic steering in the network.
  - Therefore the operators can conduct specific traffic to traverse specific path, domains and/or planes as demand.

Use Cases(Cont.)

- Inter-Controller Applications
  - The service set up between the nodes is proxied by the BGP Controllers.
  - More detailed description could be found in [draft-li-l2vpn-ccvpn-arch-00]

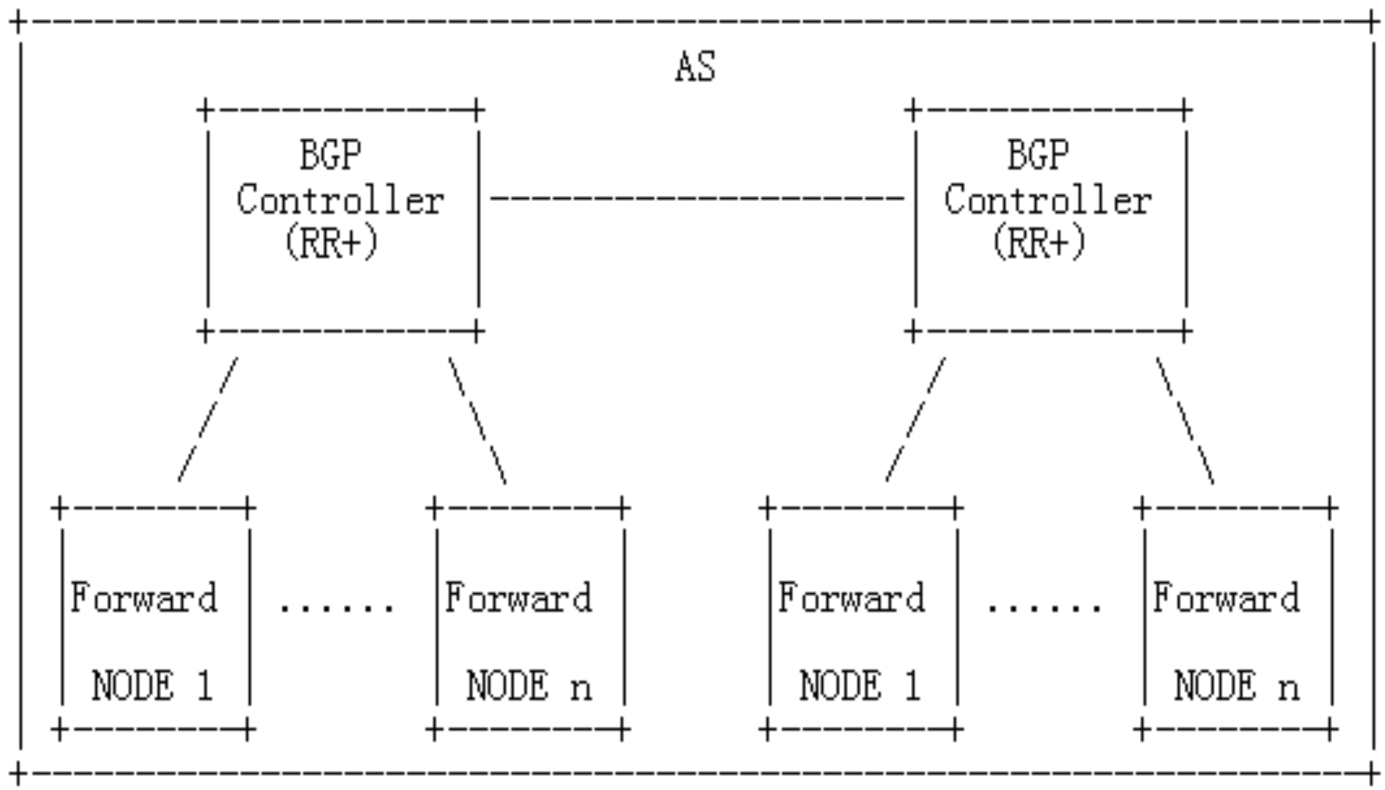

Figure 3: Removing BGP Session between Controller and NODE

Next Steps

- Solicit more comments & feedbacks
- Revise the draft