Interdomain Routing (IDR)

IETF-87, Berlin

August 1, 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

- In interest of time
  - Both in this short session,
  - And the time I ran out of before it
- … the chairs will send a document status update to the list next week.
- Questions/comments welcome now, though.

North-Bound Distribution of Link-State and TE Information using BGP

Hannes Gredler (hannes@juniper.net)
Jan Medved (jmedved@cisco.com)
Stefano Previdi (sprevidi@cisco.com)
Adrian Farrel (adrian@olddog.co.uk)
Saikat Ray (sairay@cisco.com)

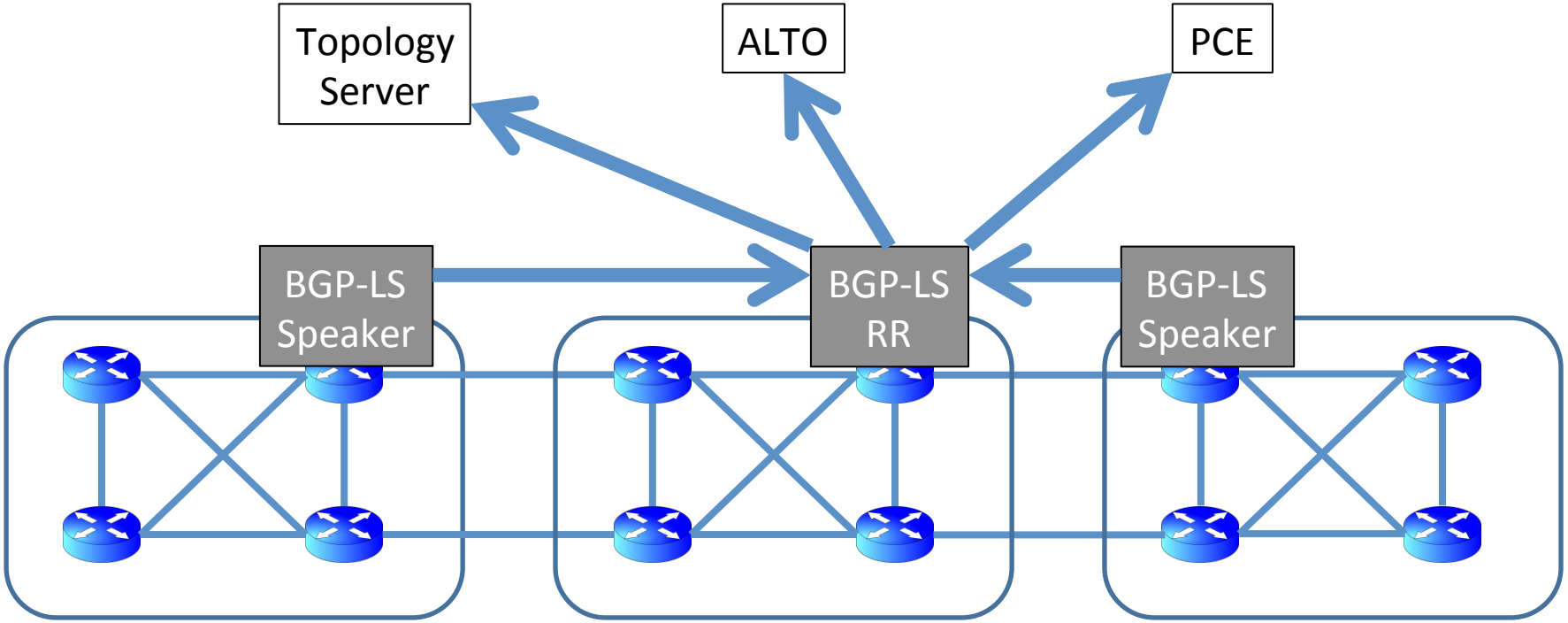
IETF 87 August 2013, Berlin

BGP-LS Overview

- BGP-LS is an address-family (afi=16388, safi=71) defined to carry IGP link-state database via BGP
  - Supports both IS-IS and OSPF(v2/v3)
  - Delivers topology information to outside agents
    - Topology servers, orchestration elements, ALTO servers
  - Allows a topology server to construct the full topology (even across ASes)
  - BGP allows policy-based control to aggregation, information-hiding, abstraction, etc.
  - Out of scope: Leak LS information back to routing

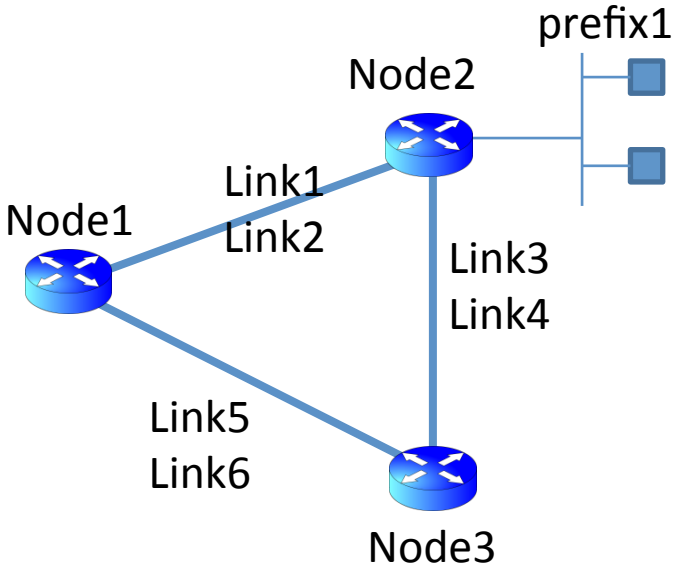
BGP-LS Overview

- Deployment model
  - IGP redistribution into BGP-LS
  - Advertisement of BGP-LS NLRIs to RR.
  - RR sends information to external agents



BGP-LS Overview

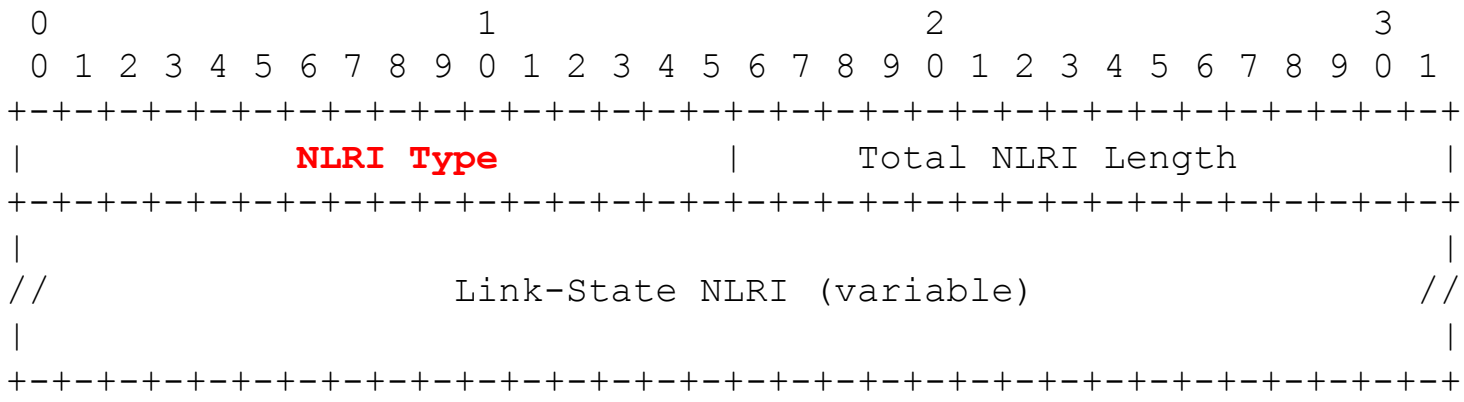
- A common topology abstraction model: An IGP network is modeled as three classes of objects
  - Nodes, Links (pair of nodes), prefixes



BGP-LS Objects:
- 3 nodes
- 6 links
- 1 prefix

BGP-LS Overview

- BGP-LS NLRI
  - NLRI Type defines the object class (node/link/prefix)
  - NLRI body is a set of TLV
  - NLRI contains the data that identifies an object
    - NLRI is the key for the object
      - Minimal data needed to remove ambiguity



BGP-LS Overview

- BGP-LS attribute
  - Optional non-transitive
  - Encode properties of the object
  - Data consists of TLVs
    - TLVs are specific to the object class
      - Node attribute TLVs (MT-ID, Flag bits, Node-name, etc.)
      - Link attribute TLVs (local/remote ipv4/ipv6 router-id, admin-group, link BW, SRLG, etc.)
      - Prefix attribute TLVs (IGP flags, (Exteneded) route tags, etc.)

Changes from -02

- One 64 bit *identifier*
  - Identifies the IGP instance
    - Needs to be "globally" unique
    - No semantics imposed.
- NLRI types are
  - Node = 1, Link = 2, IPv4 Prefix = 3, IPv6 Prefix = 4
- Node descriptor TLV uses a uniform "IGP Router-ID"
- OSPF route-type in prefix descriptor. Only one prefix in a prefix NLRI.

Changes from -02

- Node name and link name TLVs.
- OSPF area goes in the NLRI (key). New TLV for IS-IS area (goes in node attribute)
- Specifying auxiliary IPv4/IPv6 local/remote router-id in link attribute (for IS-IS links) is MUST
  - Helps with TE
- All TLV code-points from one space
- Some section reorganizations and clean-up

Status

- 3(.5) implementations
  - Cisco/Juniper
- Inter-op planned around September/October
- Still need BGP-LS Path Attribute code-point from IANA
- Comments?

Distribution of TE LSP State using BGP

draft-dong-idr-te-lsp-distribution-03

Jie Dong – jie.dong@huawei.com

Mach Chen – mach.chen@huawei.com

Hannes Gredler – hannes@juniper.net

Stefano Previdi – sprevidi@cisco.com

IETF87 IDR Aug. 2013 Berlin

Background

- The states of TE LSPs are required by some external components
  - Centralized Network Controller
  - Stateful PCE
  - NMS
  - …
- A general mechanism is needed to collect and distribute the states of TE LSPs
  - draft-ietf-idr-ls-distribution describes a mechanism to distribute link state and TE information using BGP
  - This document extends the scope of draft-idr-ls-distribution for TE LSP states

Proposed Solution

- Two new “NLRI Type” in the BGP Link State NLRI:
  - NLRI Type = 5: IPv4 TE LSP NLRI
  - NLRI Type = 6: IPv6 TE LSP NLRI

- A new TLV in BGP LINK_STATE Attribute
  - Describes the attributes & states of TE LSPs
    - path, metric, bandwidth, protection, admin status, etc.
  - TE LSP objects are regarded as Sub-TLVs

Updates after IETF85

- BGP extensions comply with draft-ietf-idr-ls-distribution
- New co-authors
  - Hannes Gredler
  - Stefano Previdi
- Editorial changes

Next Steps

- Appreciate comments from WG
- WG adoption？

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

draft-wu-idr-te-pm-bgp-01

Qin Wu

Danhua Wang

BGP attributes for NB Distribution of TE performance metrics

- Objective
  - Using BGP to share additional TE performance related information to external components beyond linkstate and TE information contained in [I-D.ietf-idr-ls-distribution]
  - External components can be ALTO server or PCE server.
- Motivation
  - As described in [I-D.ietf-idr-ls-distribution] links state and traffic engineering information (collected from IGP domain) can be distributed using BGP and share with external party (e.g., ALTO server, PCE server)
  - As described in [I-D.ietf-pce-pcep-service-aware], network performance info can be distributed via OSPF or ISIS
    - PCE uses network performance info for end to end path computation
    - However in some cases, PCE participant in the different IGP(e.g.,Inter-AS, Hierarchy PCE)

Why use BGP to distribute network performance info

- Inter-AS PCE computation
  - Cooperating PCEs to compute inter-domain path using BRPC
  - Fall short when PCE in each AS participant in different IGP
- Hierarchy of PCE
  - A child PCE must be configured with the address of its parent PCE[RFC6805]
  - Configuration system is challenged by handling changes in parent PCE identities and coping with failure events
  - parent PCEs to advertise their presence to child PCEs when they are not a part of the same routing domain is unspecified.
- Topology and Cost Info gathering for ALTO server
  - The ALTO Server can aggregate information from multiple systems to provide an abstract and unified view that can be more useful to applications.
    - Examples of other systems include routing protocol
    - ALTO server may be external component for BGP distribution
  - Gather network performance info using BGP and form Map service(i.e.,Cost Map service)

Why use BGP to distribute network performance info

- In the section 3 of [I-D.ietf-pce-pcep-service-aware], PCEP should satisfy 5 requirements regarding network performance constraints

1. PCE supporting this draft MUST have the capability to compute end-to-end path with latency, latency-variation and packet loss constraints. It MUST also support the combination of network performance constraint (latency, latency-variation, loss...) with existing constraints (cost, hop-limit...)

2. PCC MUST be able to request for network performance constraint(s) in PCReq message as the key constraint to be optimized or to suggest boundary condition that should not be crossed.

3. PCEs are not required to support service aware path computation. Therefore, it MUST be possible for a PCE to reject a PCReq message with a reason code that indicates no support for service-aware path computation.

4. PCEP SHOULD provide a means to return end to end network performance information of the computed path in a PCRep message.

5. PCEP SHOULD provide mechanism to compute multi-domain (e.g., Inter-AS, Inter-Area or Multi-Layer) service aware paths.

Brief Introduction of New BGP attribute

- [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 network performance information:
  - Link delay
  - Delay variation
  - Packet loss
  - Residual bandwidth
  - Available bandwidth
  - Link utilization
  - Channel throughput
- These BGP TLVs Applied to PCE server TED and ALTO Server, etc.

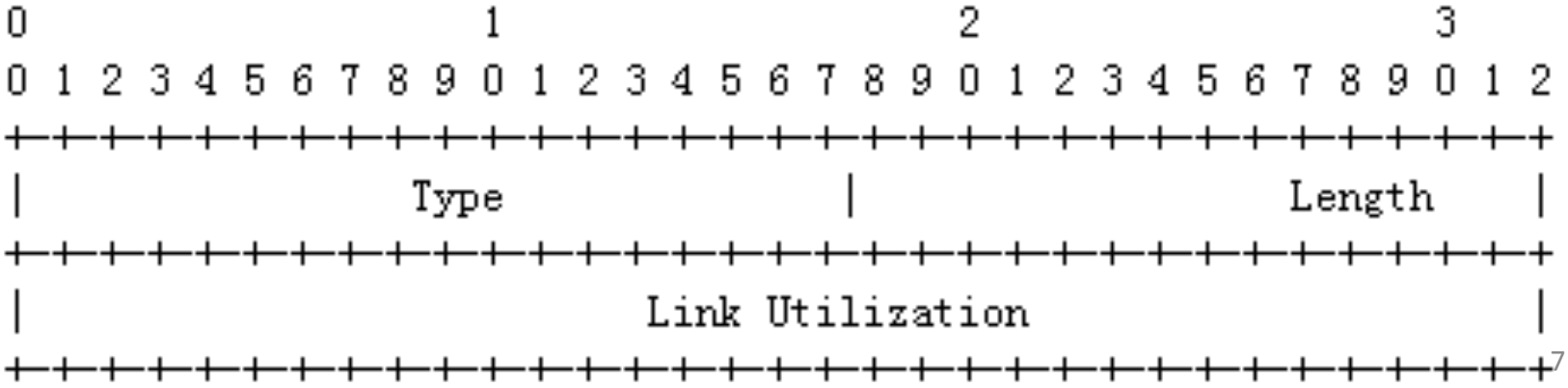
BGP Link Attribute TLVs

TLV Code Point	Description	IS-IS TLV/Sub-TLV	Defined in:
xxxx	Unidirectional Link Delay	22/xx	[ISIS-TE]/4.1
xxxx	Min/Max Unidirection Link Delay	22/xx	[ISIS-TE]/4.2
xxxx	Unidirectional Delay Variation	22/xx	[ISIS-TE]/4.3
xxxx	Unidirectional Link Loss	22/xx	[ISIS-TE]/4.4
xxxx	Unidirectional Residual Bandwidth	22/xx	[ISIS-TE]/4.5
xxxx	Unidirectional Available Bandwidth	22/xx	[ISIS-TE]/4.6
xxxx	Link Utilization	----	section 5.1
xxxx	Channel Throughput	----	section 5.2

1. [ISIS-TE] is referred to draft-ietf-isis-te-metric-extensions-00.
2.They are all Link attributes used with link NLRI defined in [I.D-ietf-idr-ls-distribution].
3. The first 5 TLVs are from IS-IS Extended IS Reachability sub-TLVs
4. The last 2 link asstribute TLVs are defined in this draft.

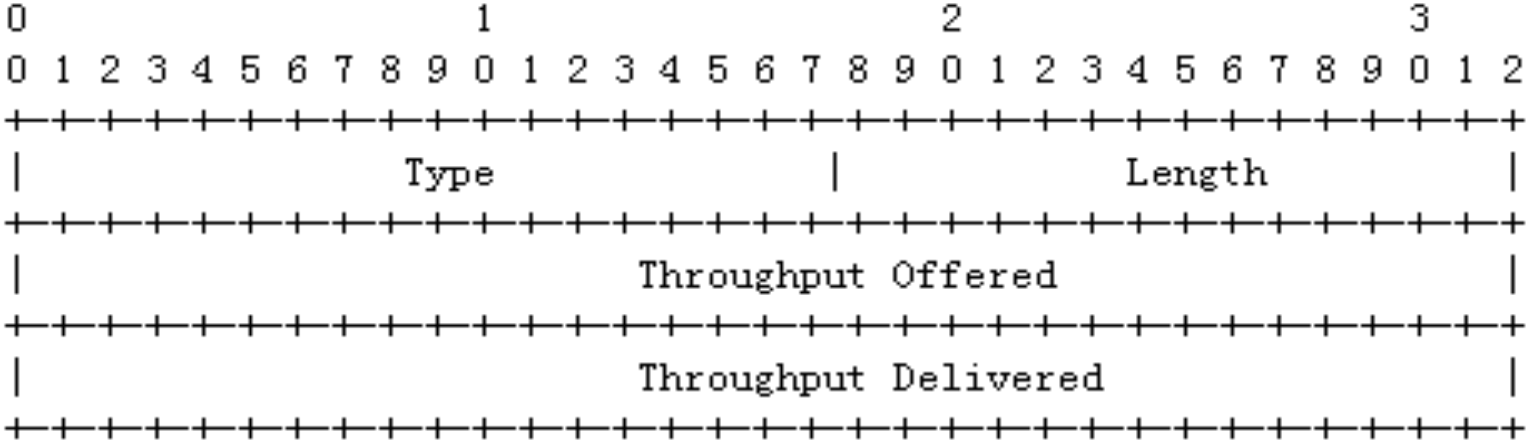
Link Utilization TLV

- Advertise the average link utilization between two directly connected IS-IS neighbors or BGP peers.
- Be the utilization percentage per interval (e.g., 5 minutes) from the local neighbor to the remote one.
- The measurement method is defined in section 6.4 of [RFC6703].
- This TLV carries aggregated link property and is more applicable to best effort network service.



Channel Throughput TLV

- Advertise the average Channel Throughput between two directly connected IS-IS neighbors or BGP peers.
- Be the throughput between the local neighbor and the remote ones over a configurable interval.
- The measurement method is defined in section 2.3 of [RFC6374].
- This TLV carries aggregated link property and ismore applicable to best effort network service.



Questions?

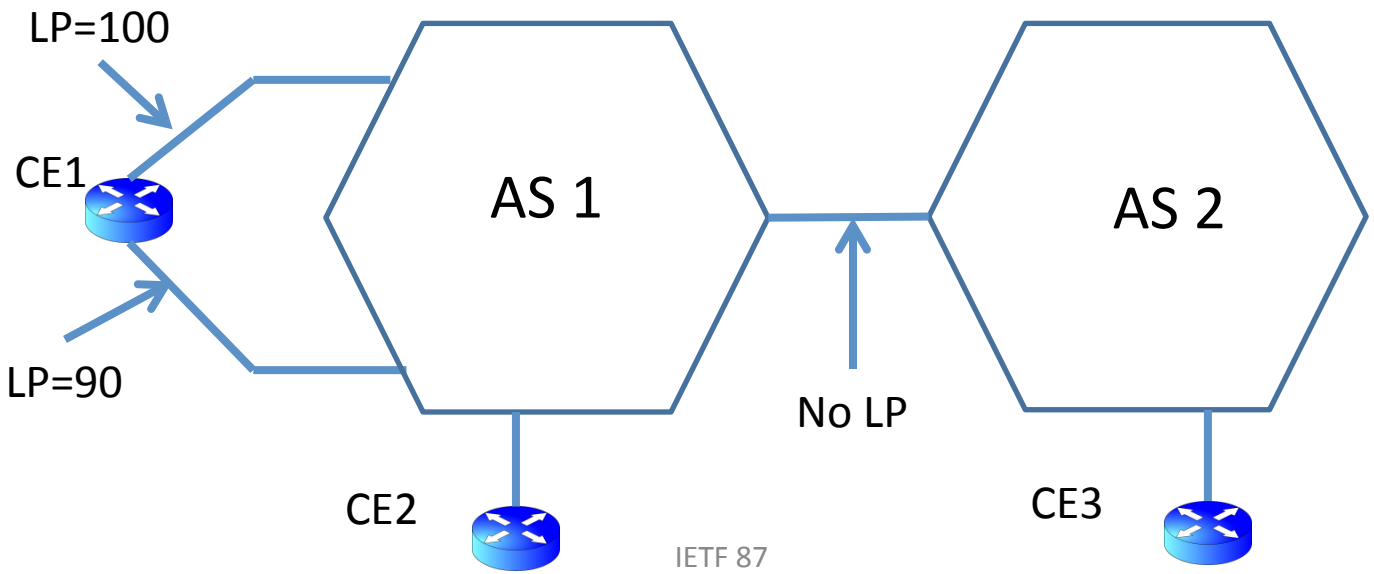
One Administrative Domain

James Uttaro (uttaro@att.com)
Saikat Ray (sairay@cisco.com)
Pradosh Mohapatra
(pmohapat@cumulusnetworks.com)

IETF 87 August 2013, Berlin

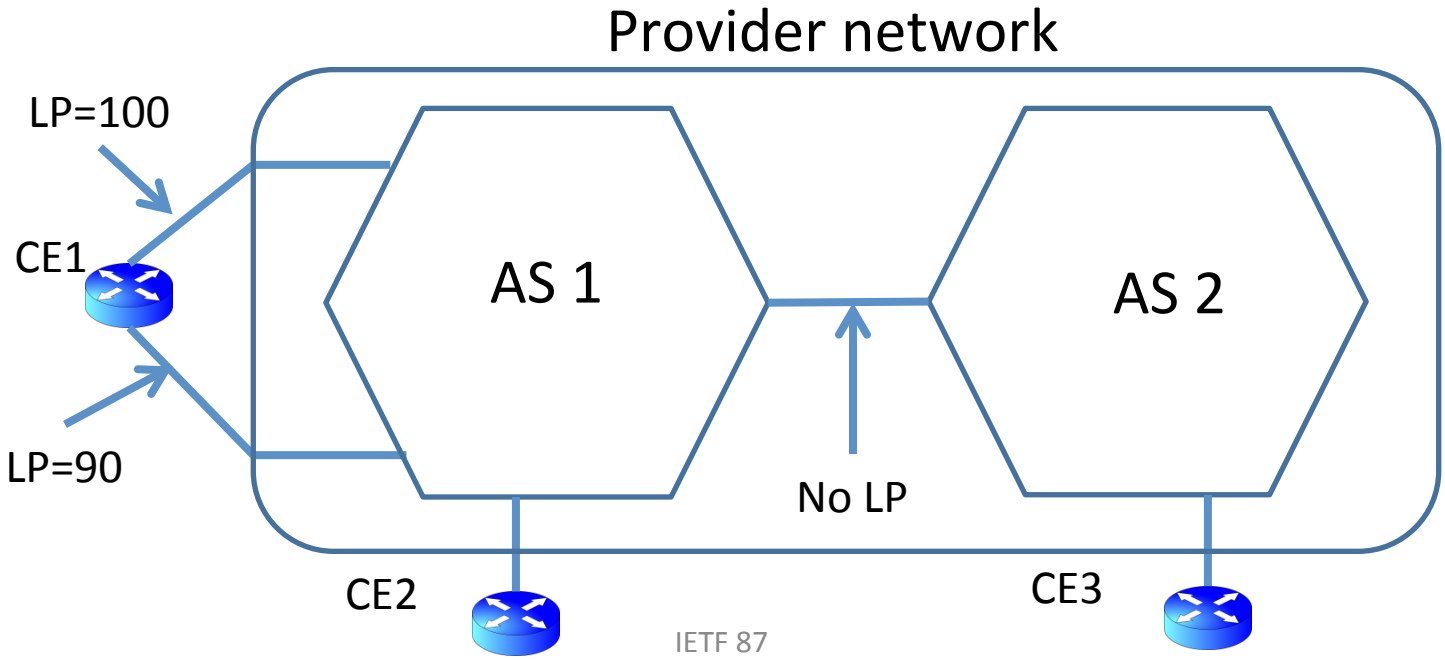
OAD Overview

- Currently autonomous systems are equated with administrative domains
  - Different ASes assumed to have independent policies
    - Attributes like LOCAL_PREF are not propagated across AS boundary
    - Traffic from CE2 honors LP; traffic from CE3 doesn't.



OAD Overview

- Reality is, a large Provider may own multiple Ases
  - Renumbering would be nice, but may not be viable
  - The customers of the Provider expect all sites to behave the same way
    - Traffic from CE3 should use the primary link to reach CE1

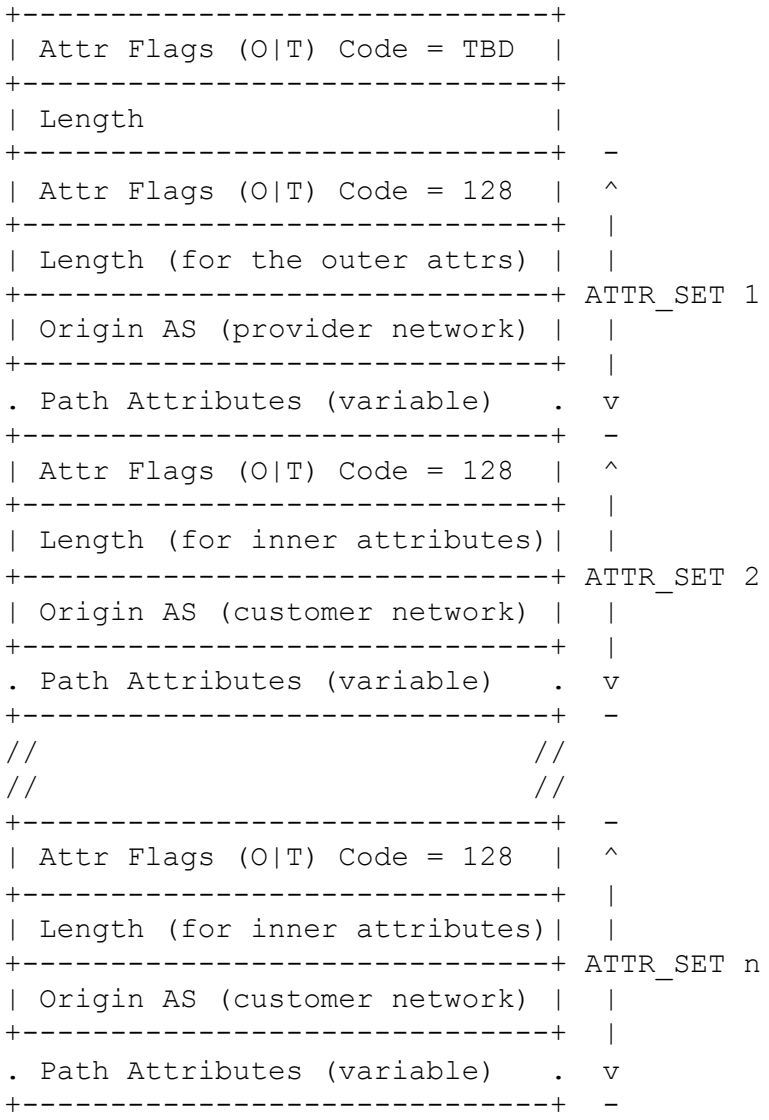


OAD Overview

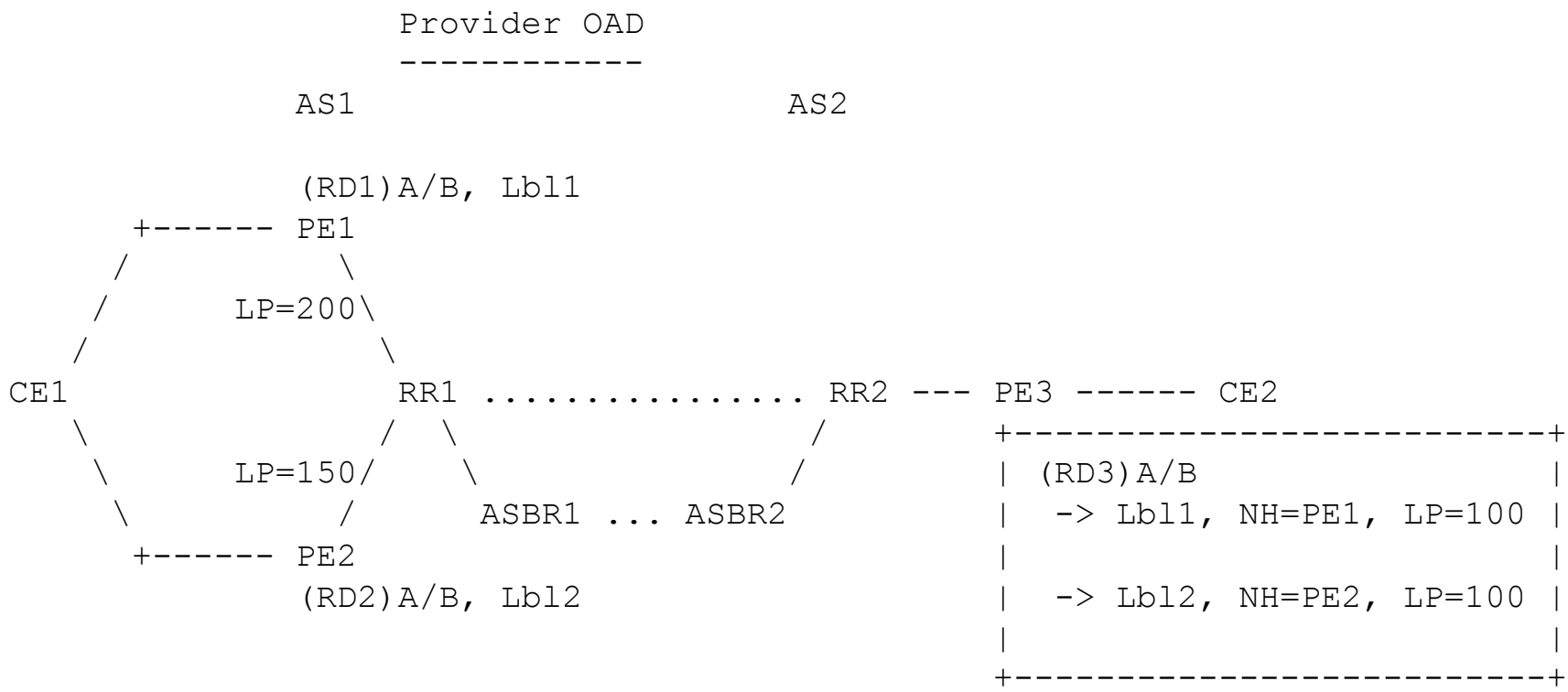
- This draft proposes tunneling attributes across ASes to create “admin-domains” where uniform policy is used and enforced
  - An admin domain need not be contiguous
- Extend ATTR_SET attribute for tunneling multiple sets of attributes
  - ATTR_SET by itself cannot be used since it has an existing semantics for CE attributes

ATTR_SET_STACK

- Optional, transitive attribute
- Creates a stack of ATTR_SET attributes
  - Encodes multiple set of attributes
  - Encodes the sequence
- Rules for encoding two ATTR_SET in this draft

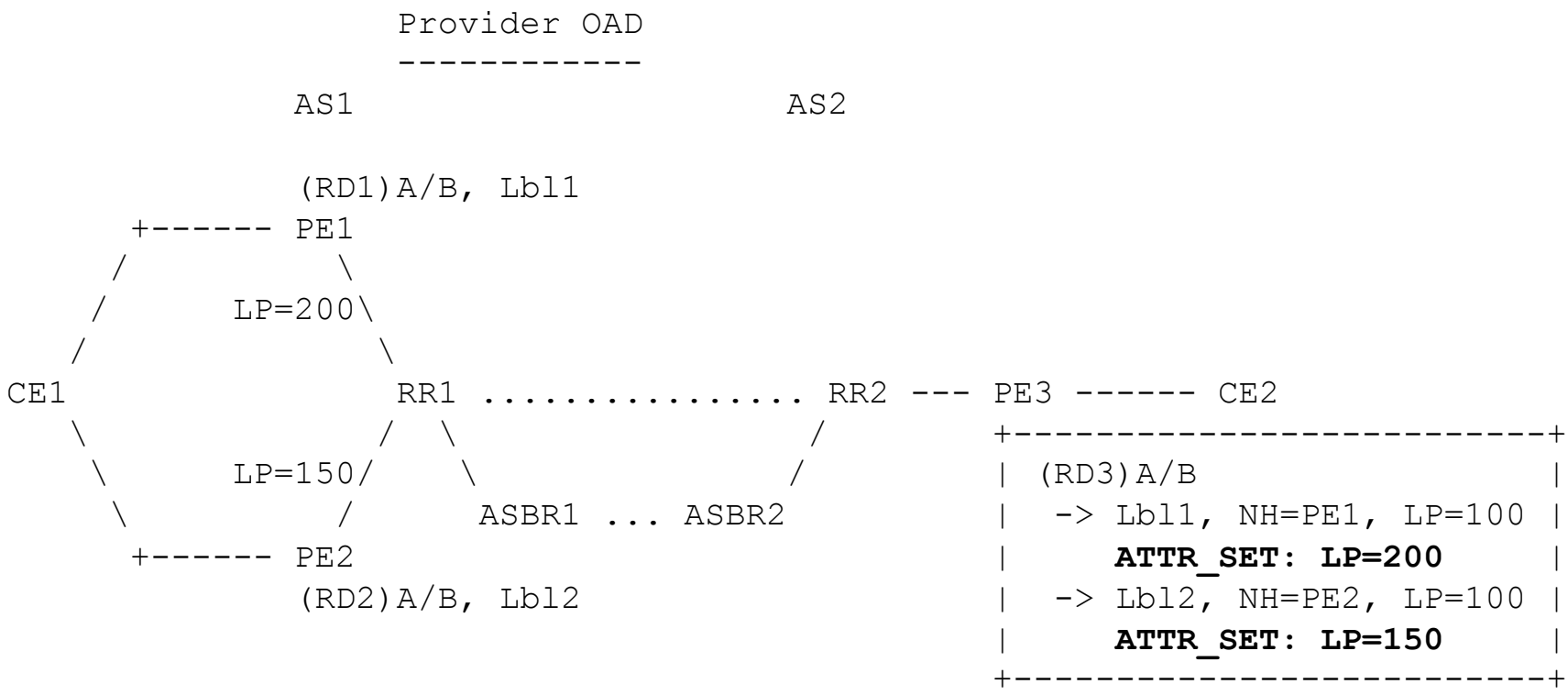


Example: Option C



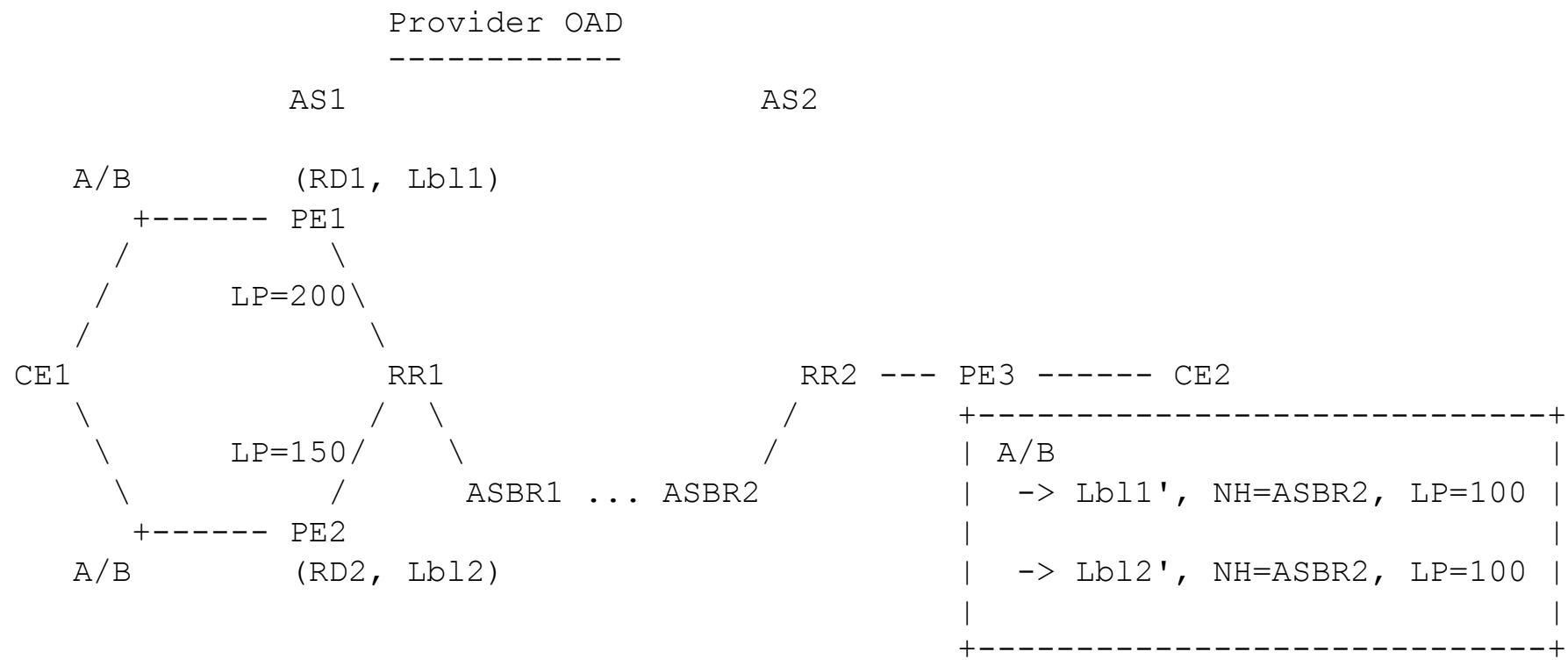
Existing behavior

Example: Option C



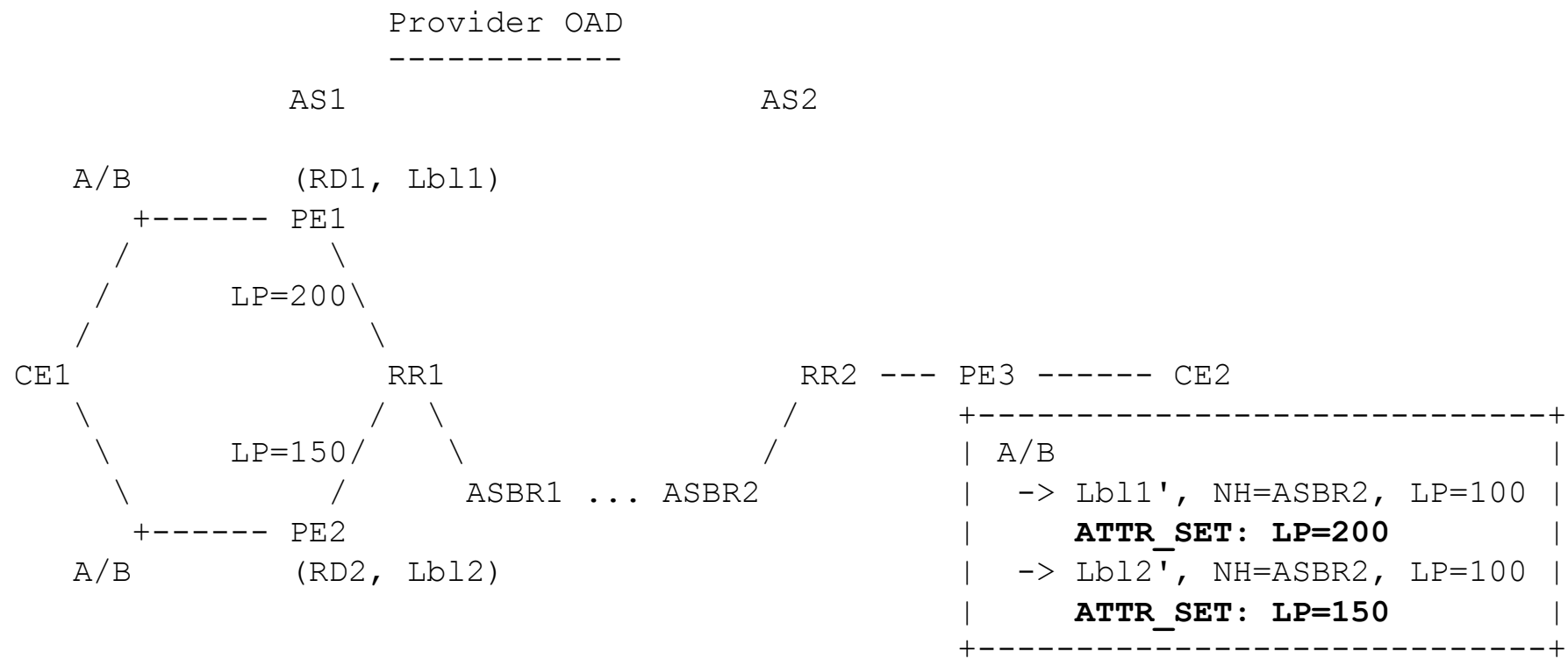
OAD behavior

Example: Option B



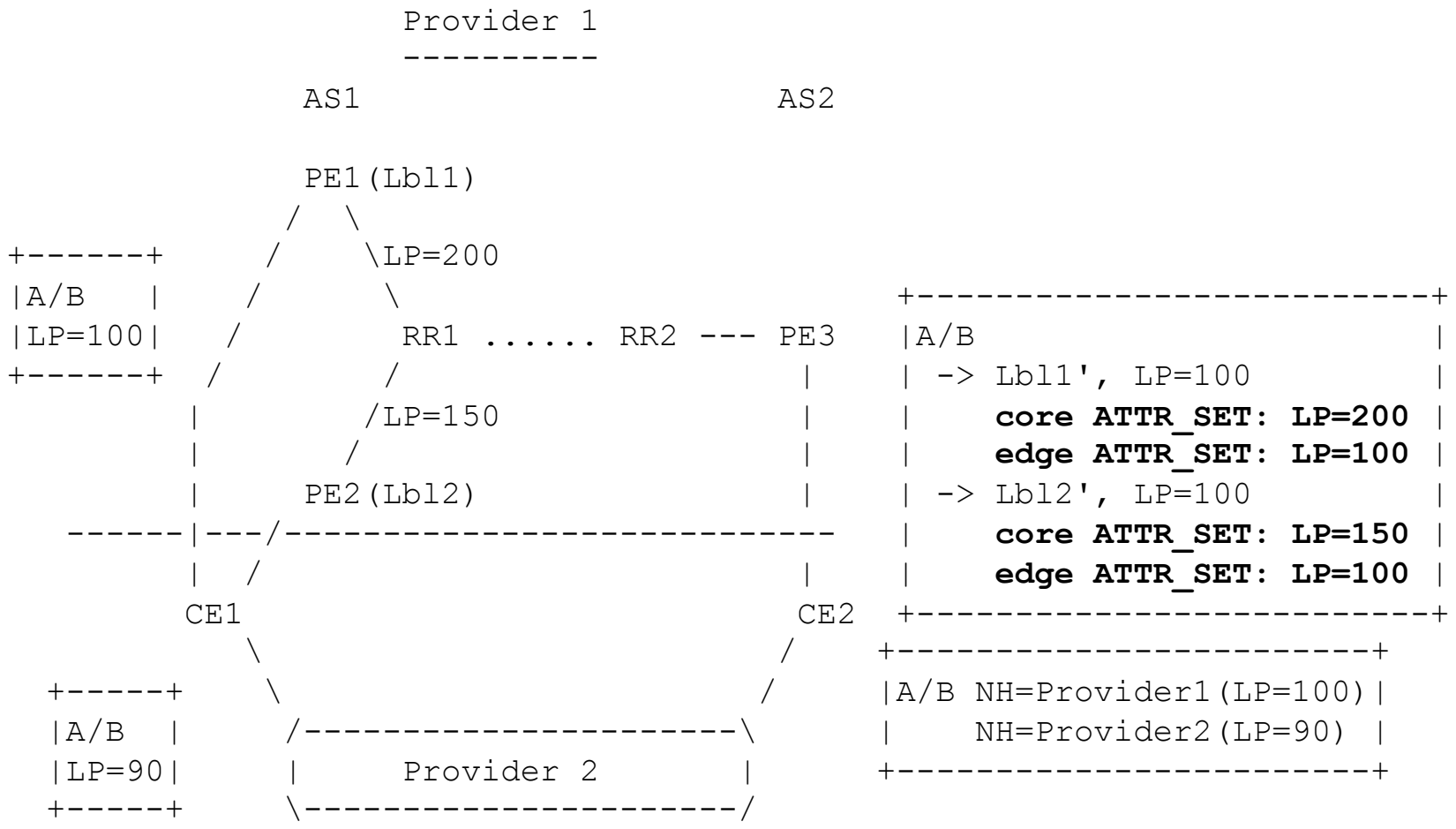
Existing behavior

Example: Option B



OAD behavior

Example: Dual Provider case



OAD behavior

OAD behavior

- ATTR_SET_STACK encodes both the edge attributes (iBGP PE-CE) and the core attributes (PE-PE/RR)
  - Needs the ATTR_SET_STACK wrapper even if only the core attributes are being sent.
- At the ingress PE, the core attributes (from the egress AS) as well as the local attributes are available
  - Local policy to make a choice
- Rules for encoding/use only at ingress and egress ASes; intermediate ASes do not use it
- Use policy to decide whether to send ATTR_SET_STACK attribute to an EBGP peer or not

Next steps

- Make the draft a WG document
- Comments?

BGP Path Marking

Camilo Cardona (juancamilo.cardona@imdea.org)
Pierre Francois (pierre.francois@imdea.org)
Saikat Ray (sairay@cisco.com)
Keyur Patel (keyupate@cisco.com)
Paolo Lucente (plucente@cisco.com)
Pradosh Mohapatra (pmohapat@cumulusnetworks.com)

IETF 87 August 2013, Berlin

Motivation

- With extensions like add-path or best-external, BGP may send non-bestpaths
- Useful to know whether a path advertised by a BGP speaker was a best path
  - Monitoring applications
  - Avoiding suboptimal routing in Inter-AS VPN

Path Type Extended Community

- Optional transitive Extended Community
  - Value field encodes path's role as a flag

Value	Path type
0x0000	Unknown
0x0001	Best-path
0x0002	Best-external path
0x0004	Multi-path
0x0008	Backup path
0x0010	Uninstalled path
0x0020	Unreachable path

  - Operational considerations
    - Attach the extended community if the path is to be advertised anyway

Comments?

BGP Persistence a.k.a. Long-Lived Graceful Restart

John Scudder

IDR, IETF-87, August 1, 2013

List of Authors

- Jim Uttaro
- Enke Chen
- Bruno Decraene
- John Scudder
- Clarence Filsfils
- Pradosh Mohapatra
- Yakov Rekhter
- Rob Shakir
- Adam Simpson

In a Nutshell

- When BGP session goes down,
  - allow relevant routes to “persist” (remain installed, but stale) for a long period of time.
  - Routes are “depreferenced” (only selected as a last resort)
- Intended use
  - “dinosaur killer” rare-but-severe control plane outages
  - Restricted/carefully considered AFI/SAFI and/or topologies

History

- 01 requested IDR adoption in mid-2012, to strong debate (love, hatred) but no clear consensus.
- Strongest objection was, if used for Internet AFI/SAFIs, possibility of leakage to the Internet At Large.
- 02 is a major revision intended to address this
  - Also analysis, clarity, terminology, {code, spec} reuse

Regular vs. Long-Lived GR

- Normal GR: don't react to session outage
  - Routes kept, no signaling to rest of network
  - Prioritizes network stability. Assumption is short duration with reversion to previous state.
- LLGR: do react
  - Routes kept but depreferenced: signaling required, network state may change
  - Stale routes are a last resort. Assumption is long duration, use up-to-date state whenever possible.

High-Level Description

- Many semantics of GR useful for Persistence
  - … so rather than reinvent, reference.
  - Implementation – minimize new/divergent code
- So what's new/different?
  - Routes can be stale for up to 2^24-1 seconds
  - Capability to signal support and constrain propagation
    - Stale routes may only be advertised to supporting peers, and are marked as "LLGR_STALE"
    - Hack for partial deployment, using NO_EXPORT
  - "NO_LLGR" community to suppress LLGR treatment

Operational

- Default off
  - Enable per AFI/SAFI after consideration
  - Generally: avoid if very dynamic, topological diversity. Consider if “semi-static”, topologically boring
- Probably usually scope to a single AS
  - But anyway, limit scope of LLGR routes to “consenting adults”

To Do

- Multicast VPN requires special consideration
  - Emerging strategy is to never use stale routes in making a new determination of Upstream PE or Upstream Multicast Hop
    - Effectively, a more draconian version of “depreference”
  - Placeholder in -02, detailed language for -03
- Note other option: don’t use LLGR for M-VPN
  - When in doubt, leave it off. Default is off.

Other issues from 01 debate

- Multi-fault scenario unlikely, poor network design
  - There are more things in heaven and earth, Horatio, than are dreamt of in your philosophy
- Depreferencing may be wrong strategy in face of supernets
  - In some cases yes, in some, no. In main use cases, no.
- Problem too marginal to justify using IDR time
  - Prefer to standardize properly rather than publishing as Informational or Individual Submission
- Solution isn't perfect
  - Perfect is the enemy of good

Next Steps

- Several implementations underway
- (Re-) Requesting WG adoption

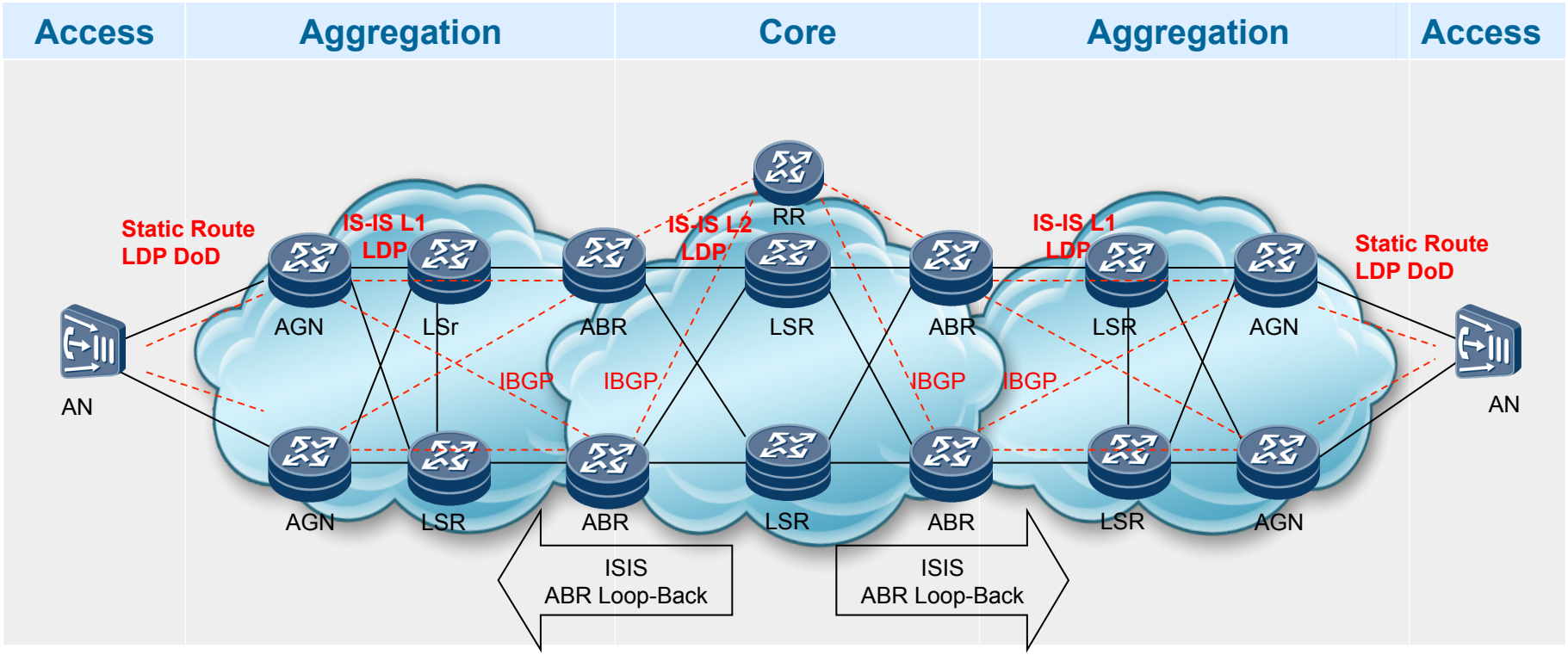
NEXTHOP_PATH ATTRIBUTE for BGP

draft-zhang-idr-nexthop-path-attr-00

Zhenbin Li, Li Zhang
Huawei Technologies

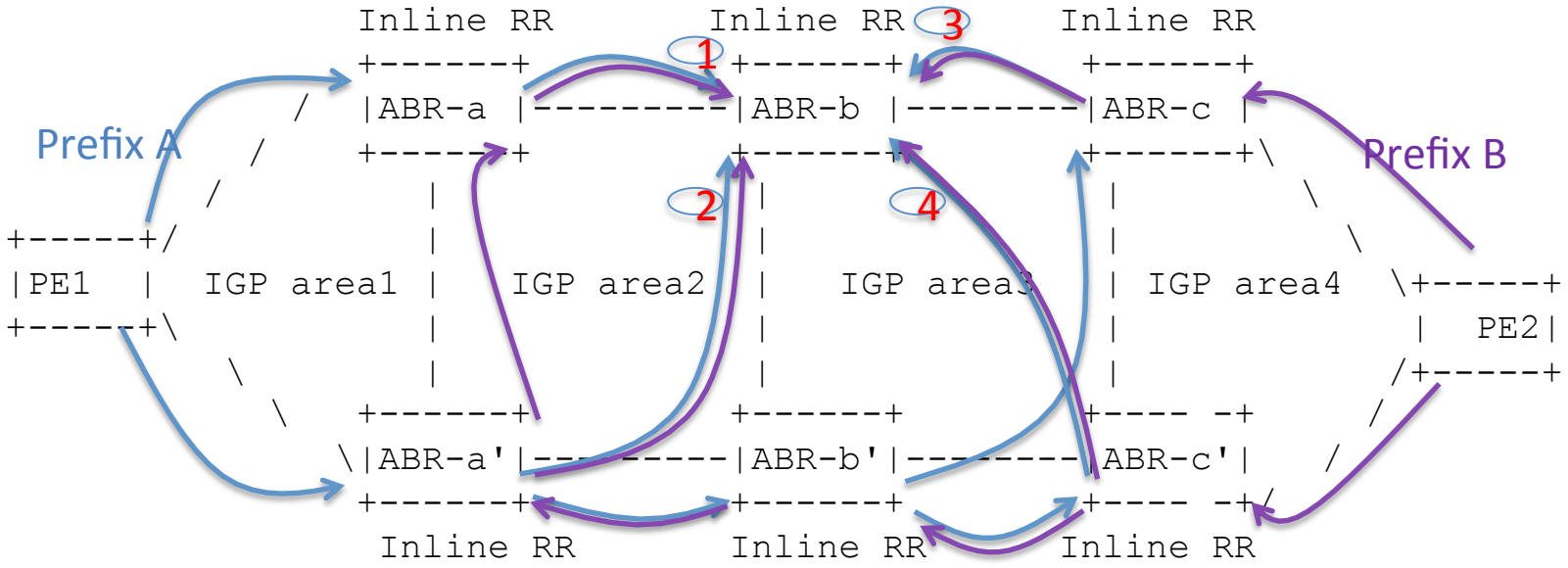
IETF 87, Berlin, Germany

Seamless MPLS Network Architecture



- The requirement of the integration of mobile backhaul networks and core/aggregation networks has been proposed
- The network will be divided into multiple IGP areas for access, aggregation and core network
- IBGP run among the Area Boarder Routers (ABRs)
- BGP ABRs should work as inline RR, which will reflect the route with next hop self (NHF)

BGP Route Selection Problem



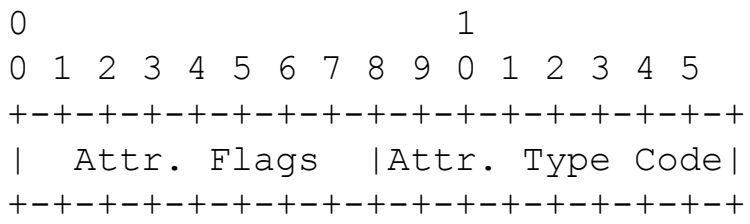
- **Inline RR will reflect the route with next hop self (NHF)**
- **For prefix A, ABR-b should select optimal route with next hop of ABR-a or ABR-a'; while for prefix B, ABR-b should select optimal route with next hop of ABR-c or ABR-c'**
- **To achieve this result, a complex route policy should be predesigned and configured for every peer every prefix**

BGP NEXTHOP_PATH ATTRIBUTE Description

- NEXTHOP_PATH ATTRIBUTE
  - is a optional transitive BGP Path Attribute
  - is composed of a sequence of next hop path segments
  - will record the distribution path of a route in Seamless MPLS network by the list of next hop path segment
- Effect of NEXTHOP_PATH ATTRIBUTE
  - for BGP route selection, which can reduce the route policy complexity
  - to get the service path in transport network, which will be used for network operation and maintenance

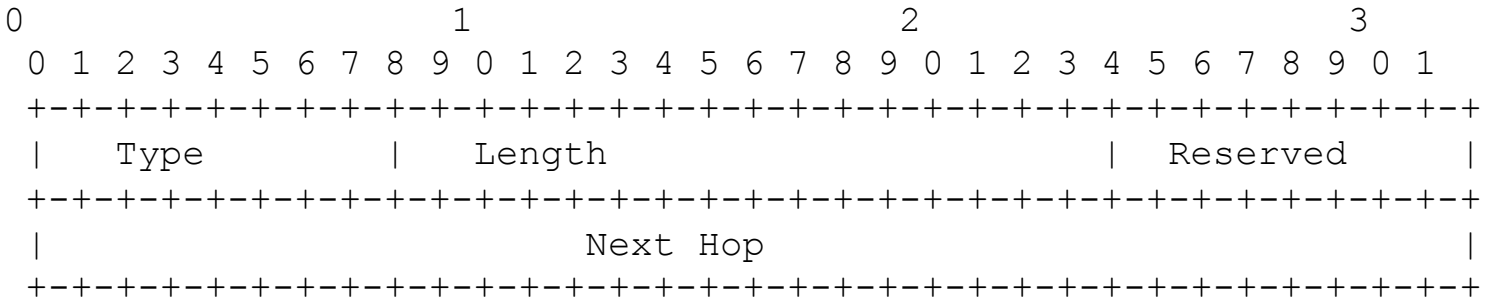
NEXTHOP_PATH ATTRIBUTE extension

- NEXTHOP_PATH ATTRIBUTE



Attr.Flags should be optional transitive
Attr.Type Code should be allocated by IANA

- Next hop path segment



Next Hop is the route next hop address

NEXTHOP_PATH ATTIBUTE Process in BGP(1)

- Creating and modification process

1. If the route is originated in this BGP speaker
   - If the attribute is supported, the TLV SHOULD be originated including the BGP speaker's own next hop address in a next hop path segment
   - If the attribute is not supported, the route will be distributed without NEXTHOP_PATH ATTRIBUTE
2. if the route is received from one BGP speaker's UPDATE message
   - If the NEXTHOP_PATH ATTRIBUTE is NULL and the local BGP speaker support this attribute, when the route is propagated with NHS, the TLV SHOULD be originated including the BGP speaker's own next hop address in a next hop path segment

NEXTHOP_PATH ATTIBUTE Process in BGP(2)

- If the attribute is non-NULL and the local BGP speaker support it, when the route is propagated with (NHS), the BGP speaker MUST appends its own next hop address as the last one of the next hop path segments.
- If the attribute is NULL and the local BGP speaker support it, when the route is propagated without NHS, the BGP speaker MUST NOT originate the NEXTHOP_PATH ATTRIBUTE.
- If the attribute is non-NULL and the local BGP speaker it, when the route is propagated without NHS, the BGP speaker MUST NOT change the next hop path sequence.
- If the BGP speaker does not support NEXTHOP_PATH ATTRIBUTE, it SHOULD keep the NEXTHOP_PATH ATTRIBUTE unchanged

NEXTHOP_PATH ATTIBUTE Process in BGP(2)

- Decision process
  - Next hop path loop detection should be done for scanning the full next hop path
  - When the NEXTHOP_PATH ATTRIBUTE is used for optimal route selection, the priority of this attribute is the same as AS_PATH ATTRIBUTE
  - When the NEXTHOP_PATH ATTRIBUTE is used for optimal route selection, the route with least next hops should be selected

Next Steps

- Get feedback on the NEXTHOP_PATH ATTRIBUTE extension and application
- The procedure for next hop path segment usage for IPv6 or other extensions will be discussed later

MULTICAST DISTRIBUTION AND REACHABILITY SIGNALING

DRAFT-REKHTER-GEO-DISTRIBUTION-CONTROL-03
DRAFT-REKHTER-MDRS-00
DRAFT-REKHTER-MDCS-00

Huajin Jeng – AT&T

Jeffrey Haas, Yakov Rekhter, Jeffrey Zhang – Juniper

IETF 87, July 2013

NOTE

This material was originally presented to IDR and MBONED in IETF 83. The full use case is presented in those sessions.

As was recommended in those sessions, the BGP specific changes have been extracted from the geo-distribution draft into the mdcs and mdrs drafts for IDR. The operational case for geo-distribution will be requested to be made a MBONED working group document.

http://www.ietf.org/proceedings/83/slides/slides-83-idr-6.pdf

PROBLEM 1: CAN THE CUSTOMER RECEIVE CONTENT VIA MULTICAST

- Ability of content-provider to determine content-receiver network destination areas where multicast-delivery option is available at a given current time period.

  This is especially critical for the successful introduction of multicast service since multicast enablement of global network infrastructure (which entails network equipment hardware/software/configuration updates) will not be flashed cut network-wide but rather will be phased in by areas over some extended period of time

PROBLEM 1: CAN THE CUSTOMER RECEIVE CONTENT VIA MULTICAST

Why not just annotate unicast routes for the customers?

- Those routes are not guaranteed to be in any specific protocol. For example, may be in an IGP or BGP.
- Unicast routes for customer networks usually represent aggregated networks. More specific prefixes that represent subsets of customers who could/could not receive multicast traffic would bias unicast forwarding.

PROBLEM 2: IMPLEMENTING BROADCAST BLACKOUTS

- Ability of content-provider to restrict multicast delivery of a given content on a designated multicast channel (S,G) to exclude a set of content-receiver network destination areas

  This is to support compliance with geo-restriction ("black-out") requirements that frequently exist for certain categories of live-event content distribution

"In broadcasting, the term blackout refers to the non-airing of television or radio programming in a certain media market. It is particularly prevalent in the broadcasting of sports events, although other television or radio programs may be blacked out as well."

http://en.wikipedia.org/wiki/Blackout_(broadcasting)

PROBLEM 2: IMPLEMENTING BROADCAST BLACKOUTS

Why shouldn't CPE provide this filtering?

- CPE devices may be tampered with. Such tampering may include interception of signaling information that may otherwise be useful for limiting content distribution.
- E.g. http://m.computerworld.com/s/article/9224838/Ore._man_convicted_for_helping_thousands_steal_Internet_service

MULTICAST DISTRIBUTION CONTROL SIGNALING (MDCS)

Document request to IDR:

- We need a new SAFI that will be associated with a flowspec encoding that is used for multicast control plane filtering.
- We're documenting a use case where Constrained Route-Target Filtering is being used for non-VPN reachability. (This is already permitted by the spec, we're not asking for a protocol change.)
- We'd like IDR to adopt this draft to document the usage of flowspec encoding with this SAFI for this application.
- That's it.

MULTICAST DISTRIBUTION REACHABILITY SIGNALING (MDRS)

Document request to IDR:

- We need a new SAFI.
- We'd like IDR to adopt this draft to document its use.
- That's it.

BGP FlowSpec IPv6
(draft-ietf-idr-flow-spec-v6-03)

Andy Karch

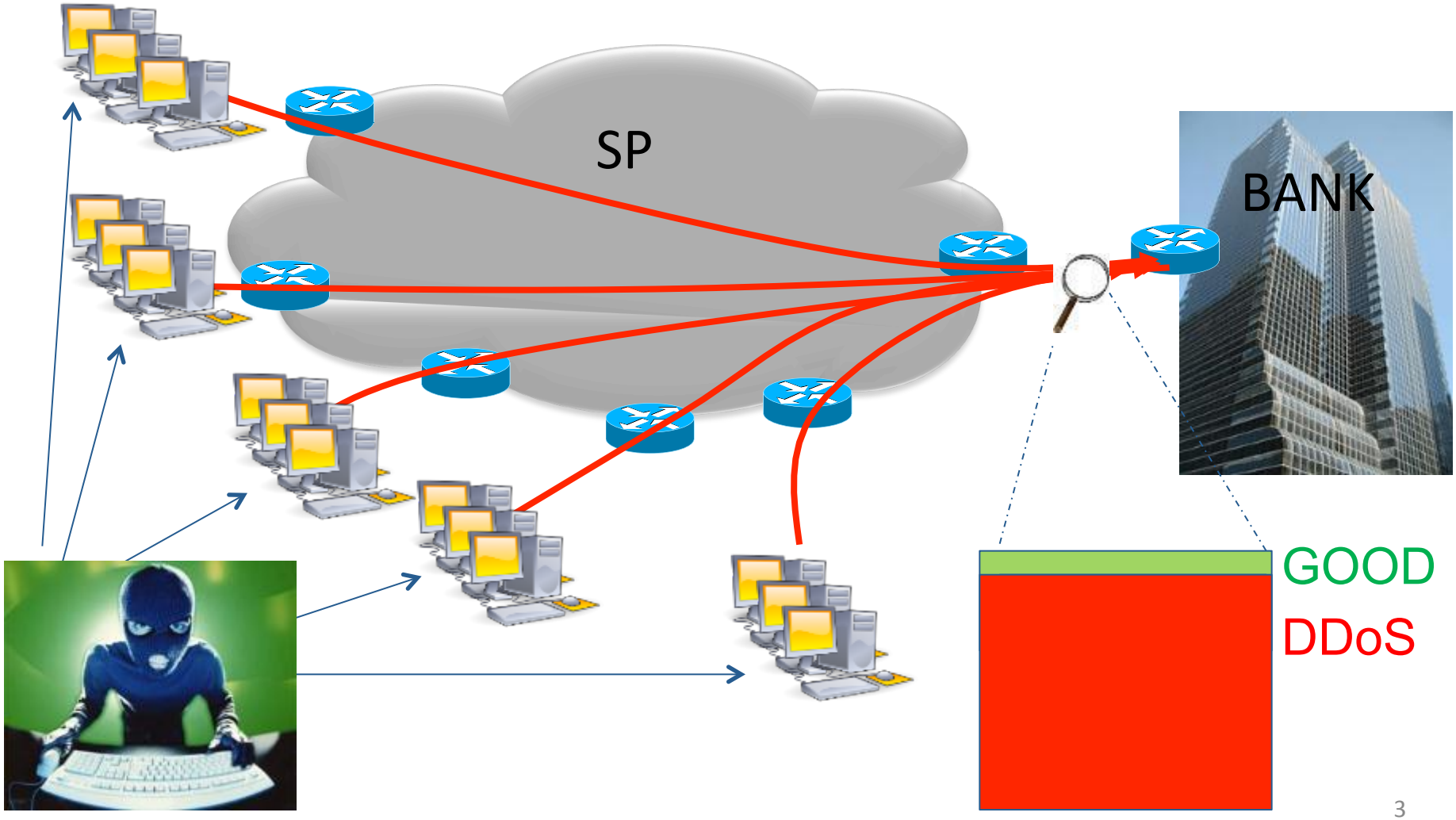
Robert Raszuk

Keyur Patel

What is FlowSpec?

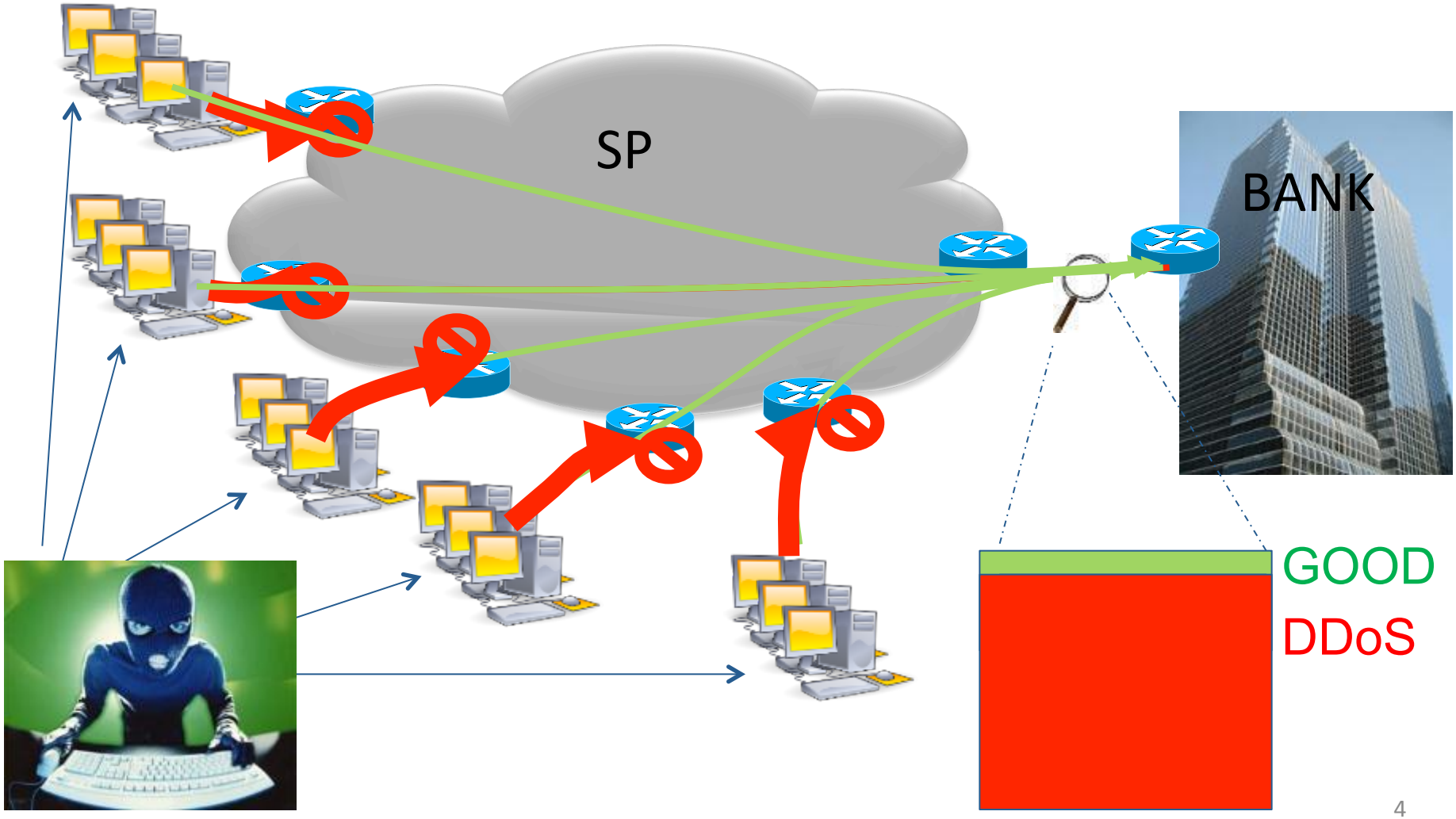
- RFC5575 - Dissemination of Flow Specification Rules
  - IPv4-Only
- BGP NLRI
  - SAFI 133 – IP
  - SAFI 134 – VPN
- Match and action
  - Match on all IP header fields, not just IP destination
  - Actions – rate-limit, drop, redirect, mark, sample
  - Similar to access-lists, policies, and filters
- Use-cases
  - DDoS Mitigation
  - Traffic Filtering

DDoS Impact



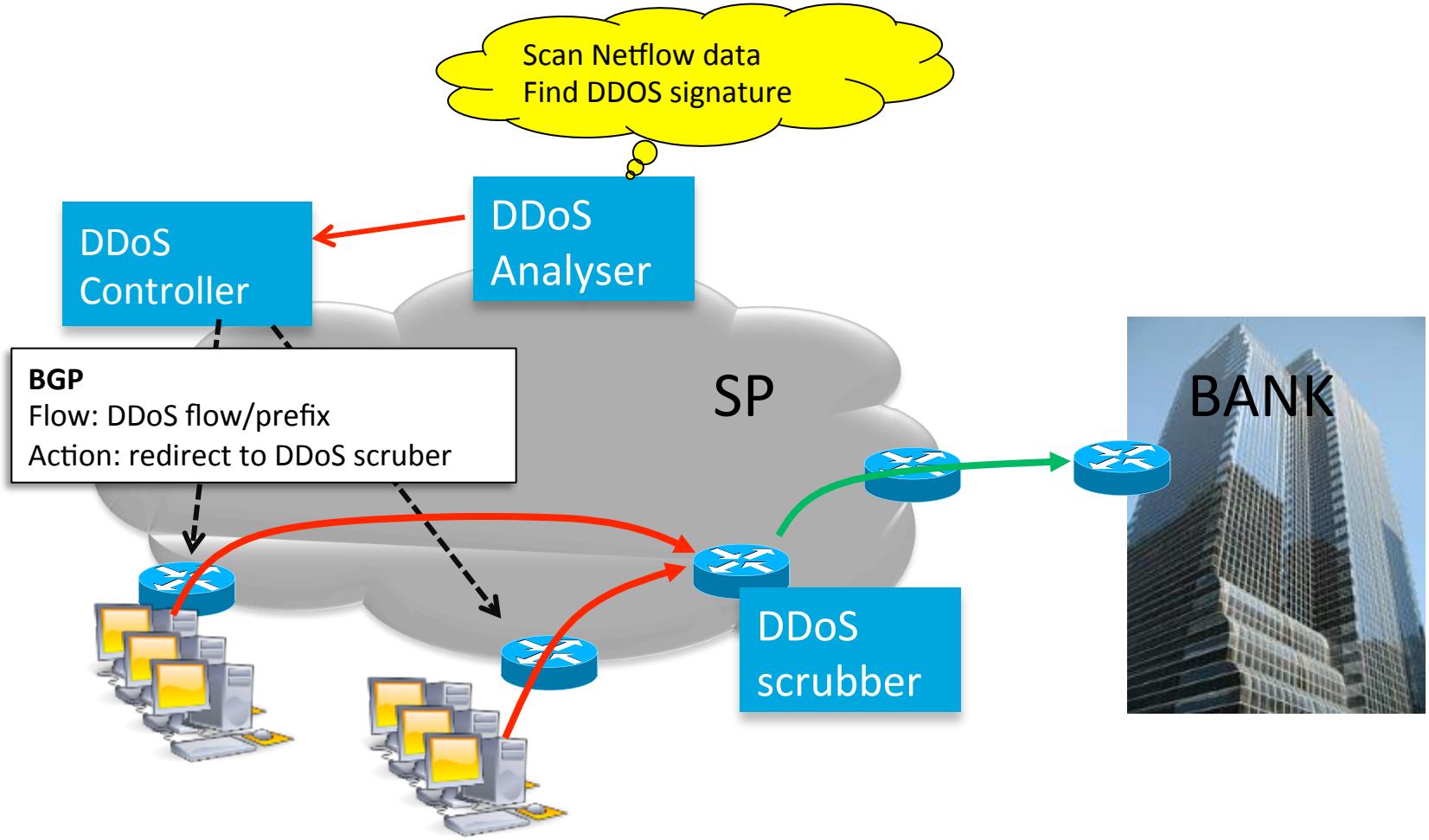
DDoS Mitigation

Drop at network ingress



DDoS Mitigation

Redirect traffic to DDoS scrubber



Changes for IPv6 Match Criteria

- Added
  - IPv6 Flow Label
- Modified
  - Source Address (prefix offset)
  - Destination Address (prefix offset)
  - IP Protocol -> Next Header
  - DSCP -> Traffic Class
- Removed
  - Fragment

Points for Discussion

1) Prefix Offset

2) Ordering of Traffic Filtering Rules

3) Fragmentation

Prefix Offset

- Don't care bits
- Field inside IPv6 prefix components
  - Destination IPv6 Prefix
  - Source IPv6 Prefix
- New for IPv6
- Allows flexible match on part of the IPv6 address
  - Match on end or interior of address.
- Prefix Encoding
  - Based on MP_REACH_NLRI in BGP UPDATE

Encoding: <type (1 octet), prefix length (1 octet), **prefix offset (1 octet)**, prefix>

Prefix Offset
Problems

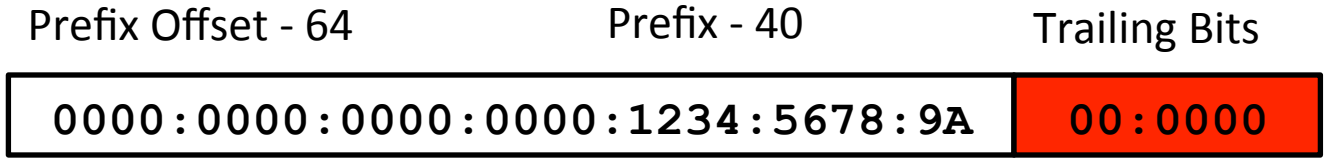
- Encoding
  - Do we include the offset bits?

- Order of Traffic Filtering Rules
  - How does the offset affect ordering?

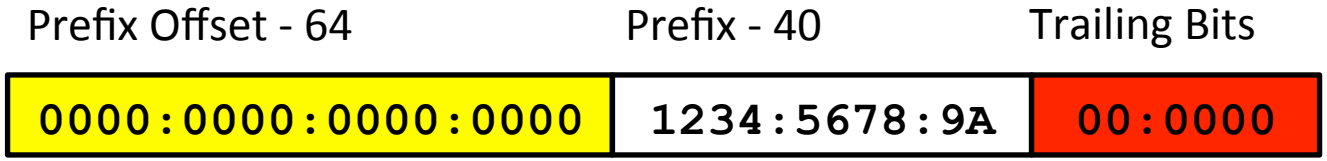
Prefix Offset Encoding

- **0000:0000:0000:0000:1234:5678:9A00:0000**
  - Length - 104
  - Don't care – 64
  - (40 match bits)

- Do we encode the entire <PrefixLength> bits?
  - The <PrefixOffset> bits are dead weight.



- Do we encode <PrefixLength> - <PrefixOffset> bits?
  - Most efficient, but perhaps <PrefixLength> is misleading.



Prefix Offset Encoding
Example - offset bits encoded

An example of a flow specification encoding for: "all packets to ::1234:5678:9A/80-104 from 192::/8 and port {range [137, 139] or 8080}".

destination	source	port
0x01 68 40 **00 00 00 00 00** **00 00 00 12 34 56 78 9A**	02 08 00 c0	04 03 89 45 8b 91 1f 90

- Destination prefix component length - 16
- NLRI Total Length - 28

Prefix Offset Encoding Example - offset bits omitted

An example of a flow specification encoding for: "all packets to ::1234:5678:9A/80-104 from 192::/8 and port {range [137, 139] or 8080}".

destination	source	port
0x01 68 40 **12 34 56 78 9A**	02 08 00 c0	04 03 89 45 8b 91 1f 90

- Destination prefix component length - 8
- NLRI Total Length – 20

Order of Traffic Filtering Rules

Problem:

- More than one rule may match a particular traffic flow.

Solution Requirements:

- Order must be constant in the network.
- Order must not depend on the arrival order of the flow specification's rules

Analogous to Longest-Prefix-Match

Order of Traffic Filtering Rules RFC5575

- IP prefix values (IP destination and source prefix):
  1. Lowest IP value of the common prefix length;
  2. if the common prefix is equal, then the most specific prefix has precedence.
  3. **NO PREFIX OFFSET**

```
if (component_type(comp1) == IP_DESTINATION || IP_SOURCE) {
    common = MIN(prefix_length(comp1), prefix_length(comp2));
    cmp = prefix_compare(comp1, comp2, common);
    // not equal, lowest value has precedence
    // equal, longest match has precedence
} else {
```

Order of Traffic Filtering Rules
Problem

- Without considering prefix offset, multiple flows that differ only by offset may appear equal in priority!

- Example
  - <Length – 32>, <Offset – 1>, <Prefix - 0x01020304)>
  - <Length – 32>, <Offset – 0>, <Prefix - 0x01020304)>

Order of Traffic Filtering Rules IPv6 FlowSpec

- IP prefix values (IP destination and source prefix):
  1. Lowest offset has precendence
     1. RATIONALE: Lowest offset matches more bits
  2. If the offset is equal, lowest IP value of the common prefix length;
  3. if the common prefix is equal, then the most specific prefix has precedence.

```
if (component_type(comp1) == IPV6_DESTINATION || IPV6_SOURCE) {
    // offset not equal, lowest offset has precedence
    // offset equal ...
    common_len = MIN(prefix_length(comp1), prefix_length(comp2));
    cmp = prefix_compare(comp1, comp2, offset, common_len);
    // not equal, lowest value has precedence
    // equal, longest match has precedence
} else {
```

Fragmentation

- Defined in RFC5575
  - Don’t Fragment
  - Is a Fragment
  - First Fragment
  - Last Fragment
- Removed in IPv6 draft

Fragmentation Undetermined Transport

- Unknown EH
- Upper-layer protocol field not in first fragment.
  - Last EH next-header field
- Upper-layer header not in first fragment
  - TCP, UDP, SCTP, ICMP…

Goals

1) Gather feedback and comments

2) Update draft
   - Converge on encoding, ordering, fragmentation
   - Clarify language
   - Provide encoding examples

3) Identify 2 implementations

4) Inter-op in the next few months.

5) Move towards RFC status

Backup Slides

Prefix Offset Component (offset bits omitted)

Type 1 - Destination IPv6 Prefix

Encoding: <type (1 octet), prefix length (1 octet), prefix offset (1 octet), prefix>

Defines the destination prefix to match. Prefix offset has been defined to allow for flexible matching on part of the IPv6 address where we want to skip (don't care) of N first bits of the address. This can be especially useful where part of the IPv6 address consists of an embedded IPv4 address and matching needs to happen only on the embedded IPv4 address. The encoded prefix contains enough octets for the bits used in matching (length minus offset bits).

Prefix Offset Component (offset bits encoded)

Type 1 - Destination IPv6 Prefix

Encoding: <type (1 octet), prefix length (1 octet), prefix offset (1 octet), prefix>

Defines the destination prefix to match. Prefix offset has been defined to allow for flexible matching on part of the IPv6 address where we want to skip (don't care) of N first bits of the address. This can be especially useful where part of the IPv6 address consists of an embedded IPv4 address and matching needs to happen only on the embedded IPv4 address. The default value for prefix offset bits SHOULD be 0, where matching uses subsequent bits up to prefix length. Otherwise prefixes are encoded as in BGP UPDATE messages, prefix length in bits is followed by enough octets to contain the prefix information.