

INTERNET-DRAFT
Intended Status: Informational draft
Expires: April 4, 2014

Arunkumar Arumuga Nainar
Tata Communications Ltd
October 1, 2013

Dynamic Path Selection (DPS) Based on Application
draft-aumuganainar-rtgwg-dps-00

Abstract

The document describes a network design architecture for routing packets via different paths available in the network based on application port number. Primarily, this is targeted for Enterprise customers who have built up redundancy at their WAN edge but are suffering from a congested primary link whilst the secondary is idle.

The objective of this architecture is as follows

- 1) Offload bulky application on to the secondary link
- 2) Achieve the above with out introducing asymmetric routing in the network

Status of this Memo

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

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF), its areas, and its working groups. Note that other groups may also distribute working documents as Internet-Drafts.

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."

The list of current Internet-Drafts can be accessed at
<http://www.ietf.org/lid-abstracts.html>

The list of Internet-Draft Shadow Directories can be accessed at
<http://www.ietf.org/shadow.html>

Copyright and License Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1	Introduction	3
1.1	Terminology	4
2.	DPS Architecture Overview.	4
3.	DPS Signaling:-	4
4.	DPS Profile Based Packet Filter	10
5.	DPS Routing Frame Work:-	12
6.	DPS Fault-detection mechanism	14
8.	Implementation Details.	14
7.	Summary	16
8	Security Considerations	17
9	IANA Considerations	17
10	References	17
10.1	Normative References	17
10.2	Informative References	17
	Authors' Addresses	17

1 Introduction

The high availability puzzle can be resolved by building in resiliency to network designs. Whilst active/backup routing schemes are sufficient to create redundancy with low convergence times the following deficiencies and customer demands are not addressed comprehensively.

1. IP routing is essentially best path based. This will lead to underutilized or over utilized links.
2. WAN application performance could be adversely impacted due to congestion whilst the backup link remains idle. Techniques such as DiffServ QoS do address the problem effectively, but those approaches address only the symptoms and not the root cause.
3. Half of the network resources that the end customer has paid for, always remains unused .This is a matter of huge concern for small and mid-size customers as WAN circuit costs are very high and recurring.

Existing Solutions

One way to address the above problems is to load balance the traffic across the available links. To enable load balancing, there are several methods that are available today such as the following.

1. Equal Cost Load balancing
2. GLBP (Global Load Balancing Protocol) based load balancing
3. Optimized Edge Routing (OER) - Cisco proprietary feature
4. Policy based routing

However all these techniques can only be implemented at per-hop level. This would mean load balancing techniques need to be applied on each and every device that the traffic passes through. Failure to do so, might result in asymmetric routing and out of order packets. This invariably results in serious application performance issues.

Proposed solution:-

To address this problem, a new architecture called Dynamic Path Selection or DPS is being proposed. DPS provides the frame work for separating applications that have different QoS requirements and sends them along two different paths in the network. By sending different applications on different links, DPS will be able to

successfully address all the issues reported above with out compromising network availability.

1.1 Terminology

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. DPS Architecture Overview.

The objective of DPS is to achieve end-to-end application separation with out introducing asymmetric routing within the network. In order to ensure the above objectives, we should have a comprehensive mechanism to achieve the following tasks.

Task 1: Any two sites participating in DPS will have to agree on a common set of applications that it will send using either the primary routing path or the secondary routing path (also called a DPS path). This happens in the control plane and will be implemented at the time routing information is exchanged. Please refer to DPS Signaling section for more details.

Task 2: At the time of forwarding the packets, packet should be filtered based on application and the capabilities of remote sites. Packets should than be pushed in to appropriate paths. Please refer to DPS Profile Based Packet filter section for more details.

Task 3: If the packet is pushed in to a DPS path, it should always use the secondary link end to end. This is achieved by building an overlay VPN network (called DPS Routing Domain) over the normal IP/MPLS network using commonly available technologies such as DMVPN (Dynamic Multipoint VPN) tunnels and VRF (Virtual Routing and Forwarding) instances. Please refer to DPS Routing Frame Work section for more details.

Task 4: A comprehensive fault detection mechanism should be put in place to detect the faults in the DPS domain. In such a case, the DPS traffic should be re-routed via the normal routing domain. Please refer to the DPS Fault-Detection & Recovery mechanism section for more details.

3.DPS Signaling:-

DPS Signaling will enable sites to actively exchange their DPS

capabilities dynamically and agree on which set of applications that it will treat as critical and non-critical. DPS architecture assumes existence of dual links on sites that are participating in DPS. For the sake of discussion, the applications to be transported across the first link (also called a primary link) are termed a critical applications and the set of applications that need to be transported across the second link (also called a secondary link) are termed non-critical applications.

In order to achieve the above objective, the Network Manager will be required to define the application profile. Information defined in the application profile will be communicated to all participating sites and a decision will be taken locally based on the profile information received for forwarding the packet.

Definition of DPS Profile:-

A DPS profile is defined as a non-overlapping applications that is treated as critical. The Network Manager will be free to define multiple DPS profiles as long as the application defined in them does not overlap with any of the previously defined DPS profiles.

For example:-

```
Profile 1:  { Citrix, SAP, RTP, H.325 }
Profile 2:  { FTP , HTTP }
Profile 3:  { SMTP, POP3 }
```

.

.

So on and so forth...

Examples quoted above are purely arbitrary and in practice, the definition will be left to the discretion of Network Managers. Any application that is not a member of the critical application set will be treated as non-critical.

Note: Alternatively customers/Network managers can also define non-critical application. In such a application that is not a member of non-critical application set will be treated as Critical.

The definition is valid as long as no application is a member of more than 1 profile. A site on the network can be defined to conform to one or more profiles. In such a case, the list of applications that the given site can potentially treat as critical is the union of all the profiles that it conforms to.

Critical application set for site X = Union of all the conforming profiles.

DPS path selection is unidirectional. In order to avoid asymmetric routing, we must ensure any two participating sites should define a common set of applications as critical. In such a case, if X and Y are two participating sites, then:

Critical Application Set for (X, Y) Pair = Critical Application Set for Site (X) \cap Critical Application Set for Site (Y)

Note: Any application that is not a member of the Critical Application set will be treated as non-critical and will go over the DPS path.

Special Case:-

It is very much possible that there could be a site within the network that does not have DPS capability. For example:

1. Site might be a small site and might not have dual links and hence DPS will not be applicable to them.
2. When a network is being migrated, the sites that have not been migrated to the new network may not understand DPS and hence should not be treated as a DPS capable site.

In such cases routing to and from the sites will have to follow normal IP routing path. To handle this special case, a default profile will be defined called Profile 0:

Profile 0: { } is a null set.

When a DPS capable site X communicates with a non-DPS Capable site Z then:

Critical Application Set for (X,Z) pair =
Critical Application Set for Site (X) \cap Critical Application Set for Site (Z)
= { } or a Null set.

The behavior for Null set is that all traffic will be treated as critical and will be routed via normal routing domain.

Hierarchical model for associating profiles to the site.

In order to aid the following objectives, a hierarchical model based

on M-Tree is proposed for DPS. The M-Tree based approach is a design guideline that provides the network manager with the following benefits:

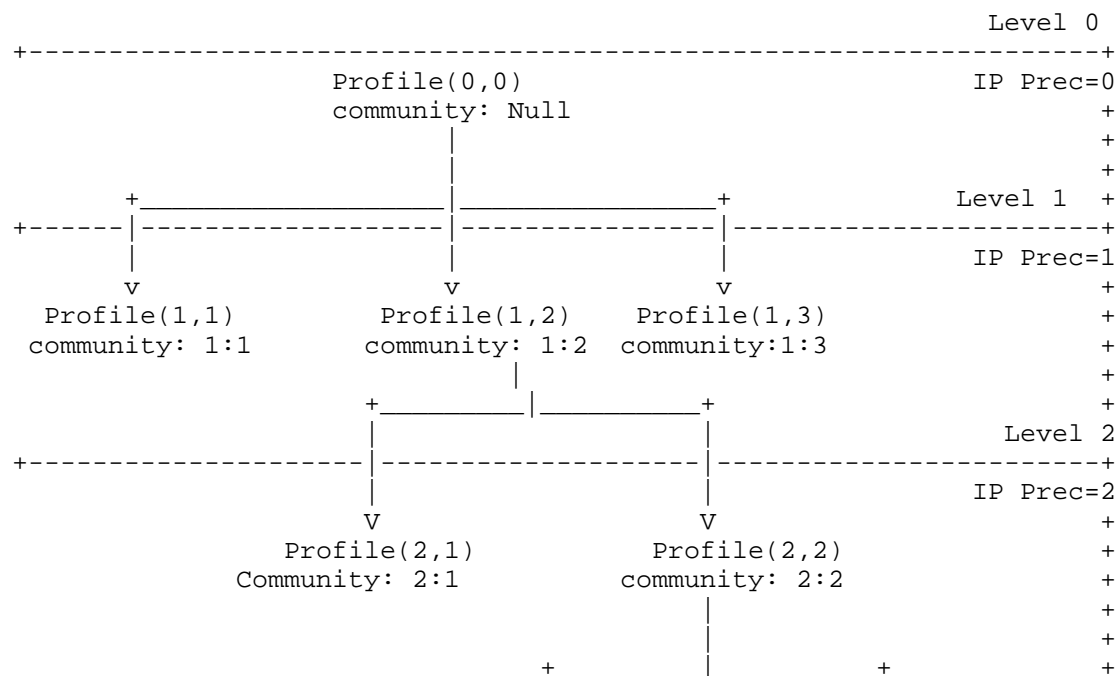
1. Provides guidelines for association rules between sites and application profiles.
2. Helps translate the above concept/rules in deployment practice using available tools and technologies.

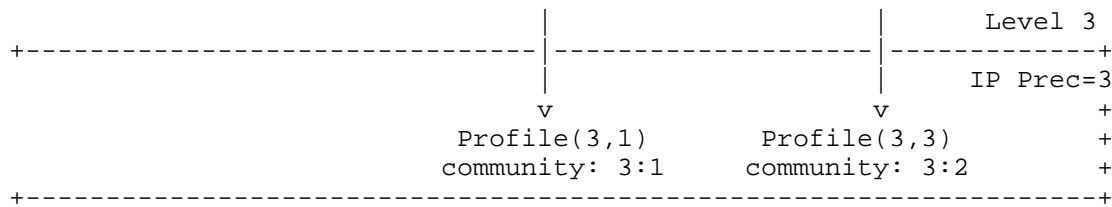
M-Tree based Association Model

As per this model, application profiles will be arranged in the form of the M-Tree as per the following rule:

Default profile or Profile 0 will form the root the tree. Other profile will be assigned as a child. Each parent can have any number of child.

Design Note: Technically the depth of tree could be infinite. However implementation schemes could impose its own restrictions. At present we rely on IP precedence to mark the depth of the tree. This restricts the depth of tree to 8 (8 levels including Level 0).





Usage of non-IP Precedence based marking could possibly extend the depth of the tree. Couple of mechanism are suggested as possible alternatives and listed below.

1. IP DSCP based marking scheme (up to 64 levels possible).
2. QOS Group based marking scheme (up to 100 levels possible).

However marking tree depth or DPS level using IP DSCP or QOS group is not possible using tools currently available in operating systems of networking devices such as Cisco's IOS. It will require minimum amount of code-development effort to take advantage of the above schemes. Till that time, IP Precedence will be used for implementing the framework on a production network and all implementations until that time will be subjected to the known restriction associated with IP Precedence.

In the above tree structure, a site can be associated with any of the profiles located in any of the levels. Under such a scenario, the critical application set is defined by following equation:

Critical Application Set for give Node i,j = Profile(i,j) U
 Profile(Parent of Profile(i,j)) for all values of i,j

In order to translate the tree structure in to actual deployment practice, each node or profile will be associated with a standard BGP community and each level will be associated with an IP precedence value. The choice of BGP community is arbitrary and is determined by the administrator. The IP precedence value chosen will be equal to the level at which the profile is located. Because DPS signaling relies on BGP community, when the network is deployed, it is mandatory that the primary link of the DPS capable site should run BGP and all the underlying providers support transport of BGP communities.

When a site advertises its routing information, it advertises the community associated with its own profile and all its parents' as well. It should be noted that at any given level, a profile will send only one community (along with the community list of its parent).

Once the communities are sent, the receiving site will interpret the communities. The interpretation of communities is limited to the communities that the given site advertises. Other communities are silently ignored. A site will receive a BGP prefix and associate an IP precedence to the prefix based on the highest level of the matching communities.

For example if a site is in Level N, then it will use following algorithm to associate an IP Precedence for the receiving profile.

```
If Level N community is present , then Set IP Precedence to N
If Level N-1 community is present then Set IP Precedence to N-1
.
.
.
If Level 2 community is a present then Set IP Precedence to 2
If Level 1 community is a present then Set IP Precedence to 1
If there is no matching community at all Set IP Precedence to 0
```

The deployment of above DPS Signaling Mechanism leverages an existing feature called QoS Policy Propagation via BGP (QPPB). This is commonly used feature on networking devices and it is used for propagating QOS marking information in the BGP advertisements. Even though it is not designed to carry DPS signaling, the QPPB functionality is leveraged to achieve DPS signaling. This would mean no additional code changes are required to be done on network devices to achieve this.

Note:- All of the above happen in the control plane (before the packet gets forwarded). However the actual marking happens when the packet hits the site's primary LAN interface. A packet will be remarked as the rules set above using QPPB. Once the packet is marked, then the packet will taken through profile based filtering where the decision will be taken about which routing domain will be referred to while forwarding the packet. Practical Illustration of DPS Profiles

Consider a small network consisting of 20 sites. The sites' profiles are categorized in to 3 types with the below configuration:

- * Type 1: Primary: 10 Mbps; Secondary: 2 Mbps
- * Type 2: Primary: 2 Mbps; Secondary: 8 Mbps/800 Kbps DSL
- * Type 3: Primary: 8 Mbps/800 Kbps DSL; Secondary: None

Common applications used on the network are Citrix, SAP, SMTP, FTP & HTTP. Among which Citrix and SAP are very critical to the business and needs to be protected.

The Network Manager wants to restrict Citrix and SAP to the primary link and the rest to the secondary link. This works well on Type 2 sites. These are small sites predominantly consisting of thin client. However on Type 1 sites are large sites with thick client. Users utilise applications such as SMTP and Lotus notes more than SAP and Citrix. Here a problem is noticed. There is high congestion on the 2 Mbps secondary link. SMTP and FTP are business traffic but by nature they are bulky. Because Type 1 sites have a large number of thick clients, the portion of this traffic is also high. Hence there is the desire to offload SMTP and FTP on to the large 10Mbps link.

Based on the above scenario Profile tree can be built as follows.

Profile 0: { } - This is null set ; BGP Community: None and Precedence = 0.

Profile 1: {Citrix, SAP } with BGP Community : 100:1 and Precedence = 1.

Profile 2: {SMTP, FTP} with BGP Community : 100:2 and Precedence = 2.

This configuration will result in following:

Case 1: When Type 1 talks to Type 1 Site:
Critical Application = {Citrix, SAP, SMTP, FTP}

Case 2: When Type 1 talks to Type 2 Site:
Critical Application = {Citrix, SAP}

Case 3: When Type 2 talks to Type 2 Site:
Critical Application = {Citrix, SAP}

Case 4: When Type 1 talks to Type 3 Site:
Critical Application = { }

Case 5: When Type 2 talks to Type 3 Site:
Critical Application = { }

Case 6: When Type 3 talks to Type 3 Site:
Critical Application = { }

4. DPS Profile Based Packet Filter

DPS Profile Based Packet Filter attempts to filter packets based on DPS profiles and pushes them in to the relevant DPS routing domain or the normal routing domain. It happens in two steps:

> STEP 1:- Colour or mark the packet based on DPS capabilities of the destination site as per the rules set by DPS Signaling.

> STEP 2:- Filter the packets based on application and the DPS capabilities of the source-destination pair.

STEP 1: Colouring or Marking of Packets.

The actual marking happens when the packet hits the routers LAN interface. The packet will be remarked as per the rules set during the DPS signaling by QPPB. Once the packet is marked, the packet will be taken through profile based filtering where the decision to forward it to the relevant routing domain will be taken.

Design Note: Because QPPB remarks the traffic, Trust based QoS model will not be supported when DPS is turned on in a given site. However, QoS can still be applied on DPS capable sites; this is achieved by performing explicit classification and marking at the router before applying QoS policies on the out bound interface.

Note: Current DPS implementation supports only IP Precedence based markings. However with a little bit of development effort other mechanisms such as QoS group can also be adopted. When this is done, restrictions on trust based QoS model will cease to exist. Here the packet is appropriately coloured so that we can pass this through a profile based filter.

1. Application of the incoming packet is an element of Critical Application Set for (X,Z) then it will be push to normal routing domain.

2. Otherwise it will be pushed to DPS routing domain.

- 3.Special condition rule also applies here, i.e. if Critical Application Set for (X,Z) is a null set then packet will be pushed to normal routing domain.

This Profile based filter will be applied on the LAN interface of the router. Once the traffic hits the primary router, the traffic gets separated as DPS traffic or as normal traffic and gets pushed to appropriate routing domain. Implementation models for Profile based filter is done through two common features/technologies:

1. Packet filters (Access Control List) based on TCP and UDP application port numbers and IP Precedence.

2. Policy based Routing (PBR).

PBR will use simple next hop feature to push the traffic in to the DPS domain (please refer to DPS Routing Framework section for more details). However in case of single router, dual circuit scenario, a modified version of PBR will be used. Here, PBR will be used to select the VRF domain based on which packet will have to be routed. This feature is called VRF selection based on PBR and it is common feature used on most of networking devices including Cisco.

It should be noted that there are several restrictions on PBR match criteria in most implementations such as matching IP Precedence using extend ACLs is not supported. However this mechanism has been tested and implemented in Cisco's software based routing platforms such as ISRs.

Also during our implementation, we have found that PBR had huge impact on routers performance. Hence future implementations based on sleek model using Layer 4 port numbers and IP Precedence could be done to make these processes more efficient.

5. DPS Routing Frame Work:-

DPS Routing frame work provides overlay routing domain for routing packets that belong to non-critical applications. DPS frame work assumes the following:

1. Customer sites consist of redundant routers and redundant links. The first link (also called a primary link) will connect to Router 1 (also called a primary router) and will be used to carry traffic belonging to critical applications. Primary link will also carry all the traffic destined for sites that do not support DPS. The second link (also called a secondary link) will connect to Router 2 (also called a secondary router) and will be used to carry traffic belonging to non-critical applications.

2. DPS routing framework also assumes that BGP is enabled across the primary link and the network provider supports transport of BGP communities end to end.

In order to create a DPS routing framework two new interfaces/sub interfaces will be configured and their details are listed below.

1. Dynamic multipoint tunnel interface (DMVPN tunnel interface). This will be created on the secondary router. The DMVPN tunnel is a point to multipoint tunnel interface commonly used in IP Networks for creating any-to-any overlay VPNs.

Source Address of the DMVPN tunnel will only be advertised via secondary link. At the primary router these source address will be

filtered out. This ensures that any traffic coming out of tunnel interface will leave the local site via the secondary link and enter the destination site via its secondary link

2. In addition to the tunnel interface, one more sub-interface will be created across the back to back link between the primary and secondary router.

In order to secure the normal and DPS routing domain, new virtual routing and forwarding instances (VRF) will be created on the secondary router. Both the DMVPN tunnel interface and the DPS back to back sub-interface on the secondary router will be assigned to the VRF.

Routing protocols will be enabled on the newly created interface and separate routing protocol instances will be run across the DPS domain. Following peers will be established across these interfaces:

1. 1st peering will be established across DPS back to back interface between primary and secondary router.

2. 2nd peering across DMVPN hub. It should be noted that though routing information is exchanged only with DMVPN hub device, traffic flow will be always happen directly between the spokes. This capability is defined by Next Hop Resolution Protocol (NHRP # RFC 2332) and it is built in to DMVPN tunnel technology. This capability is leveraged to provide any to any communication on the DPS Frame work.

Design Note:- In order to increase the availability of the DPS routing domains it is suggested to host additional DMVPN hubs. In such a case each DPS site will have two peering points via DMVPN tunnel interfaces.

All the LAN routes are pushed in to the DPS domain via peering established across back to back sub interface. This is then propagated across the entire network via a DMVPN tunnel interface. VRF configured on the secondary router ensures that DPS and normal routing information do not get mixed up with each other. If the DPS routing domain is built around the above guidelines, we can ensure that the packet will leave the local site via its secondary link and enter the remote site again via the secondary link.

The above design assumes two routers being used. However the design could be a single router, two circuits scenario as well. In such a case, there is no need for the DPS back to back sub-interface. The rest of details remain the same for the single router scenario.

6. DPS Fault-detection mechanism

As with any networks, faults can happen in a DPS routing domain. DPS by design has got several single points of failure. However DPS has been equipped with sound fault detection and recovery mechanisms. Fault detection and recovery mechanisms will dynamically allow a given router to detect faults that might have happened anywhere (local and remote faults) on the DPS domain. Once the fault is detected the packet is ejected out of the DPS domain and pushed on to the normal routing domain.

Fault detection is enabled through dynamic routing information exchanged via a routing protocol. A fault can happen any where within the site such as:

1. Secondary link could have failed.
2. Back to back link connecting primary and secondary router could have failed.
3. LAN interface on the primary router could have failed.

All of the above failures will result in routing information being withdrawn from the routing table. If a route for a given DPS capable site is not present in the DPS routing table then it is considered a fault.

To enable fault recovery, DPS uses a default static route to push the traffic out of the DPS domain and in to the normal routing domain. During the event default route is used inside the routing domain, we will have to use one or more summary route that encompasses all the LAN routes used with in the network instead of default static routes. This will enable DPS to push the traffic in to the DMVPN tunnel if a more specific route is available. In case a more specific route is not available (this might happen due to local or remote fault) it will use default static route to pop out of DPS domain and back out to the primary router and route via the normal routing domain.

8. Implementation Details.

This architecture has been developed using exiting features available in Cisco IOS. Details are given below.

- 1) DPS Signaling :- QPPB
- 2) Profile based Filter :- PBR and Extended ACL
- 3) Routing Framework :- OSPF, DMPVN and VRF

4) Fault Recovery :- Static Routing

All the components are put to gather as described in previous sections and has been thoroughly tested in labs and also implemented in the field. Current implementations are done using Cisco routers and IOS version 15.0M. OSPF has been used as routing protocol inside the DPS domain and it has been tweaked so that it scales well in large deployments. During lab testing, we were able to scale well using this architecture where it was tested up to 500 sites with 5000 prefixes. In the production environment, several implementations were done with largest one consisting of 300 sites & 2000 prefixes. Following are the challenges that we faced during this implementation. Some of them will require additional development effort:

1. Lack of trust based QoS model. This restriction is particularly important in converged environment where voice and data shares the same infrastructure space. Here customers wanted their providers to support trust based markings. Due to reliance of IP precedence based coloring for identifying DPS capabilities trust model could not be supported.

2. Matching using Extended ACLs based on IP Precedence inside the PBR was also a challenge. All hardware switching based platforms such as Cisco's Catalyst platforms failed during lab testing. However software switching based platforms such as Cisco's ISRs performed really well both in lab and also in the production environment.

3. PBR based filters had severe restriction on throughput of software based routing platform. Additional development work is required to accomplish light weight profile based filters.

To a greater extent, large scale implementation is possible in the present form with out any modifications on any networking hardware that supports the above mentioned features (eg: Cisco IOS). However, with little bit of development effort, we will be able to overcome some of the shortcomings as well. These are listed below

- 1) Lack of support for trust model has been a major drawback in the current architecture. Though QPPB can mark, QOS-GROUP field, it can not be matched inside a PBR. IOS in its current form only allows classification based on QoS-Group only on output policy. If support can be added for matching QOS-Group inside a PBR then we can do the coloring based on QoS-Group instead of IP Precedence. Hence trust model can be easily supported.

- 2) PBR is currently used for Profile based filtering. however throughput of the device is very much limited when this feature is turned

on. Since filtering is only done on IP Precedence and Application port-number, special filters could be developed to speed up this operations. This could improve the performance of the application even better.

7. Summary

By summarizing all the four components, true end to end application based routing scheme could be achieved. Such DPS frame work has the following advantages:

1. Give lots of room for Network Manager to determine which path should be used for which application.
2. This is very scalable framework.
3. Trouble shooting the setup is easy and simple since it is based on simple routing.
4. DPS capable sites can co-exists with non DPS sites and this capability provides enough room for phased migration. Hence DPS technology adoption is easy and simple.
5. It should be noted that DPS frame work and signaling, needs to be understood only by edge devices and all the devices in middle such as provider routers need not be aware of DPS.

```
Definitions and code {  
    line 1  
    line 2  
}
```

Special characters examples:

The characters , , ,

However, the characters \0, \&, \%, \" are displayed.

.ti 0 is displayed in text instead of used as a directive.

.\" is displayed in document instead of being treated as a comment

C:\dir\subdir\file.ext Shows inclusion of backslash \".

8 Security Considerations

TBD

9 IANA Considerations

TBD

10 References

10.1 Normative References

- [KEYWORDS] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC1776] Crocker, S., "The Address is the Message", RFC 1776, April 1 1995.
- [TRUTHS] Callon, R., "The Twelve Networking Truths", RFC 1925, April 1 1996.

10.2 Informative References

- [EVILBIT] Bellovin, S., "The Security Flag in the IPv4 Header", RFC 3514, April 1 2003.
- [RFC5513] Farrel, A., "IANA Considerations for Three Letter Acronyms", RFC 5513, April 1 2009.
- [RFC5514] Vyncke, E., "IPv6 over Social Networks", RFC 5514, April 1 2009.

11 Acknowledgements

The authors would like to thank Hesham Moussa for his review and comments.

Authors' Addresses

Arunkumar Arumuga Nainar
Tata Communications (UK)
1st Floor
20 Old Bailey

INTERNET DRAFTDynamic Path Selection Based on ApplicationOctober 01, 2013

London EC4M 7AN
United Kingdom

EMail: arun.arumuganainar@tatacommunications.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: October 29, 2016

F. Baker
Cisco Systems
M. Xu
S. Yang
J. Wu
Tsinghua University
April 27, 2016

Requirements and Use Cases for Source/Destination Routing
draft-baker-rtgwg-src-dst-routing-use-cases-02

Abstract

This note attempts to capture important use cases for source/destination routing.

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 <http://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 October 29, 2016.

Copyright Notice

Copyright (c) 2016 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
1.1. Requirements Language	3
2. Use Cases	3
2.1. Simple Egress Routing	3
2.2. General Egress Routing	5
2.3. Specialized Egress Routing	6
2.4. Intra-domain access control	7
2.5. Traffic Engineering	8
3. Derived Requirements	9
4. IANA Considerations	9
5. Security Considerations	9
6. Privacy Considerations	10
7. Acknowledgements	10
8. References	10
8.1. Normative References	10
8.2. Informative References	10
Appendix A. Change Log	11
Authors' Addresses	11

1. Introduction

Source/Destination routing has been proposed in the IPv6 community and specifically in homenet as a means of dealing with multihomed networks whose upstream networks give them provider-allocated addresses. An initial approach was suggested in [RFC3704], which assumed that a packet following a default route to an egress CPE Router might arrive at the wrong one, and need to be redirected to the right CPE Router. Subsequent approaches, including those listed in the bibliography, have focused on using routing protocols or routing procedures with extensions that make decisions based on both the source and the destination address.

"Source/Destination Routing" is defined as routing in which both the source and the destination address must be considered in selecting the next hop. It might be thought of as routing "to a destination with a constraint" - a router might have multiple routes to a given destination, and follow the one that also obeys the constraint, or it might have only one route to a destination but correctly fail to forward a packet that doesn't meet the constraint. From that perspective, the logic here extends to other cases in which a constraint might be placed on the route. As with all routing, a primary requirement is to follow the longest-match-first rule to the destination; following a less specific route may well take traffic to the wrong place.

As a side note, source address spoofing in this case will be limited to addresses from the indicated source prefixes, obviating the need for upstream ingress filtering. Ingress filtering within the domain in LAN switches can prevent spoofing of addresses within those prefixes.

This note attempts to capture common use cases. These will be in terms of a general statement of intent coupled with a specific example of the intent for clarity. The use cases are obviously not limited to these, but these should be a reasonably complete set.

1.1. Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC2119].

2. Use Cases

The use cases proposed here are not an exhaustive set, but are representative of a set of possibilities. At least three are presently-deployed use cases; the fourth is a possible use case within an edge network.

2.1. Simple Egress Routing

One use case is as shown in Figure 1. A customer network has two or more upstream networks, and a single CPE Router. Each upstream network allocates a prefix for use in the customer network, and the customer network configures a subnet from each of those ISP prefixes on each of its LANs. The CPE Router advertises default routes into the network that are "from" each PA prefix. Apart from prefix itself, the services of the upstream ISPs are indistinguishable; they each get the customer to the Internet.

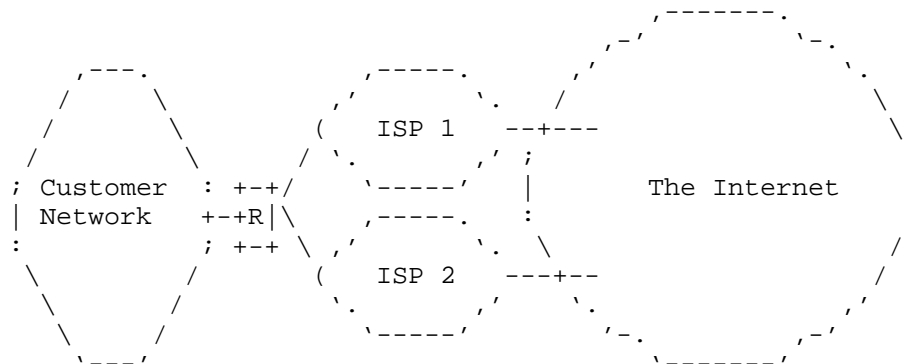


Figure 1: Egress Routing in a Multihomed Environment with One CPE Router

The big issue in this network is, of course, ingress filtering [RFC2827] by the upstream ISP. If packets intended for a remote destination pass through the wrong ISP, they will be blocked. In the ideal case, traffic following default route gets to the upstream network indicated by its source address.

The CPE Router could, at least in concept, advertise a single default route into the network, as all traffic to an upstream ISP must pass through that CPE Router. However, should another CPE Router be added later, it would have to change its behavior to accommodate that CPE Router (as in Section 2.2). Hence, the single CPE Router must advertise two default routes into the network, one "from" each PA prefix.

In this case, the destination prefix in routing is a default route, `::/0`. The source prefix is the prefix allocated by the ISP. In this case, routing within the network is largely unchanged, as all traffic to another network goes to the CPE Router, but the CPE Router must send it to the correct ISP.

Note that in this use case, if there are other routers or internal routes in the network, there is no need for them to specify source prefixes on their routes, and if they do, the prefix specified is likely to be `::/0`. The reason is that traffic arriving from the ISPs must be delivered to destinations within the network, so routing cannot preclude them.

2.2. General Egress Routing

A more general use case is as shown in Figure 2. A customer network has two or more upstream networks, with a separate CPE Router for each one. Each upstream network allocates a prefix for use in the customer network, and the customer network configures a subnet from each of those ISP prefixes on each of its LANs. Each CPE Router advertises a default route into the customer network. Apart from prefix itself, the services of the upstream ISPs are indistinguishable; they each get the customer to the Internet.

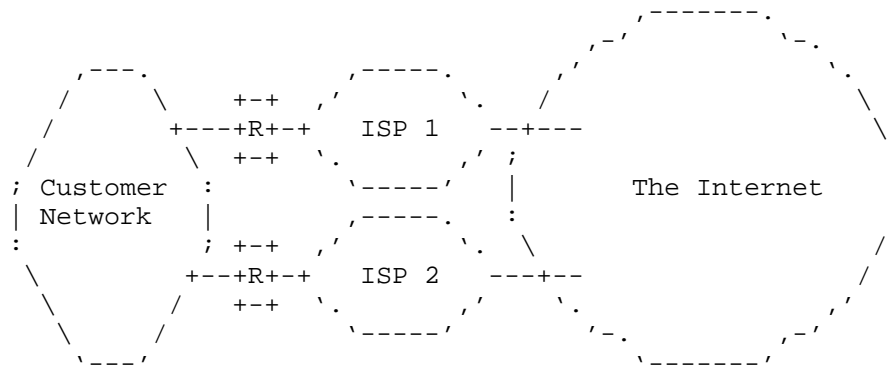


Figure 2: Egress Routing in a Multihomed Environment

The big issue in this network is again ingress filtering [RFC2827] by the upstream ISP. If packets intended for a remote destination pass through the wrong ISP, they will be blocked. Traffic following default route gets to the upstream network indicated by its source address.

In this case, the destination prefix in routing is a default route, `::/0`. The source prefix is the prefix allocated by the ISP. We want a routing algorithm that sends packets matching such a specification to the CPE Router advertising that default route.

Note that in this use case, if there are other routers or internal routes in the network, there is no need for them to specify source prefixes on their routes, and if they do, the prefix specified is likely to be `::/0`. The reason is that traffic arriving from the ISPs must be delivered to destinations within the network, so routing cannot preclude them.

2.3. Specialized Egress Routing

A more specialized use case is as shown in Figure 3. A customer network has two or more upstream networks, with one or more CPE Routers; the example shows a separate CPE Router for each one. Each upstream network allocates a prefix for use in the customer network, and the customer network configures a subnet from each of those ISP prefixes on each of its LANs. Some CPE Routers might advertise a default route into the customer network; one or more of the other CPE Routers, perhaps all of them, advertise a more-specific route. The services offered by the upstream networks differ in some important way.

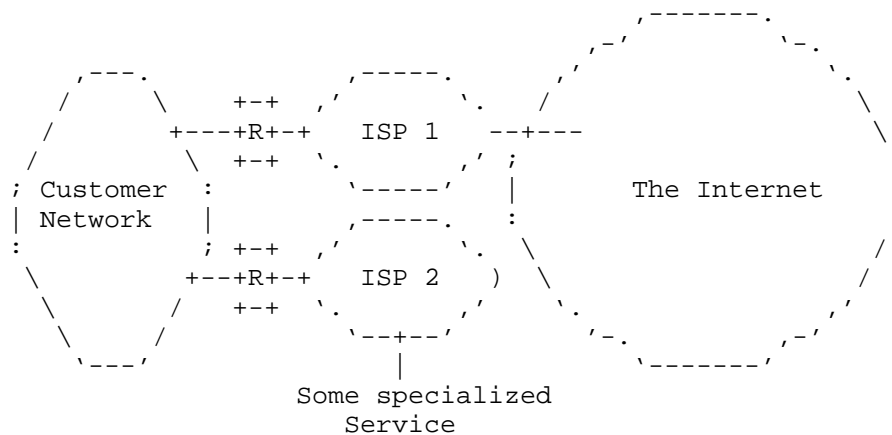


Figure 3: Egress Routing with a specialized upstream network

A specific example of such a service is the NTT B-FLETS video service in Japan; however, the use case describes any use with one or more walled gardens. In the B-FLETS case, a customer may purchase services from a number of ISPs, providing general Internet access. However, the video service requires customers accessing it to use its allocated prefix, and other ISPs (following [RFC2827]) will not accept that prefix as a source address. This is similar to the previous use cases, but

- o the only application at that "ISP" is the video service,
- o packets using the video service MUST use the video service's source and destination addresses, and
- o no other service will accept a video service address as a source address.

The big issue in this network is, once again, ingress filtering [RFC2827] by the upstream ISP, with the additional caveat that the upstream services are far from identical. If packets intended for a remote destination pass through the wrong ISP, they will be blocked. Additionally, while other ISPs advertise access to the general Internet, they may not provide service to the specialized service in question. Hence, egress routing in this case also ensures delivery to the intended destination using the bandwidth it provides. In the ideal case, traffic following default route gets to the upstream network indicated by its source address.

In this case, one or more ISPs might offer a default route as a destination prefix in routing, `::/0`. The source prefix is the prefix allocated by the ISP. In addition, the ISP offering the specialized service advertises one or more specific prefixes for those services, with appropriate source prefixes for their use. We want a routing algorithm that sends packets matching such a specification to the CPE Router advertising that indicated route, and dropping, perhaps with an ICMPv6 response, packets for which it effectively has no route.

Note that in this use case, if there are other routers or internal routes in the network, there is no need for them to specify source prefixes on their routes, and if they do, the prefix specified is likely to be `::/0`. The reason is that traffic arriving from the ISPs must be delivered to destinations within the network, so routing cannot preclude them.

2.4. Intra-domain access control

A use case within the confines of a single network is as shown in Figure 4. A network has one or more internal networks with differing access permission sets; the financial servers might only be accessible from a set of other prefixes that financial people are located in, or university grade records is only reachable from the offices of professors. This could be implemented using firewalls between the domains, or using application layer filters; in this case, the routing architecture replaces an exclusive firewall rule.

In this case, each domain advertises reachability to its prefix, listing acceptable source prefixes. Domains that are willing to be generally reached might advertise `::/0` as a source prefix, or the prefix in use in the general domain.

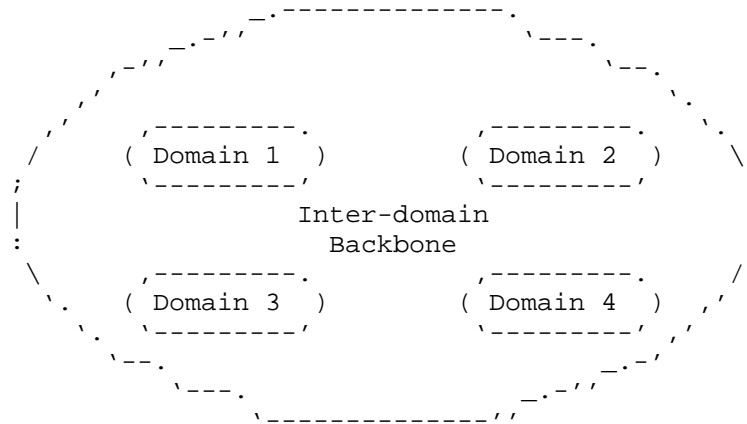


Figure 4: Intradomain Access Control

The big issue in this network is a difference in policy.

2.5. Traffic Engineering

This use case derives from real requirements of CERNET2, an IPv6 network with 59 PoPs and sites from 22 cities. The network shown in Figure 5 has multiple internal networks with different priorities when accessing the target network. For example, domain 1 and domain 2 need higher speed. At the same time, the egress router R1 is much more congested than R2, because traffic from almost all domains (including 1, 2, 3, 4) travel through R1. It is anticipated that network can divert traffic (from some domain to target network) to another egress router for reducing the total latency.

For a mid-size network, CERNET2 wants to make the operations more dynamic and does not want to use static routing or PBR. Also, CERNET2 does not want to use MPLS and MTR, because it does not have MPLS/MTR operators and the learning curve is quite high. So, CERNET2 desires to deploy src/dst routing.

In this case, the egress router advertises reachability from specific source prefixes to the target network, with different metric representing the priority. For example, by adjusting the advertised metrics, the path from domain 1 and 2 towards the target network will have much smaller metrics when going through R2 than through R1. Thus, the routers across the intra-domain will divert the traffic from domain 1 and 2 to R2 when forwarding to the target network.

This implementation uses Source/Destination Routing Using BGP-4 [I-D.xu-src-dst-bgp].

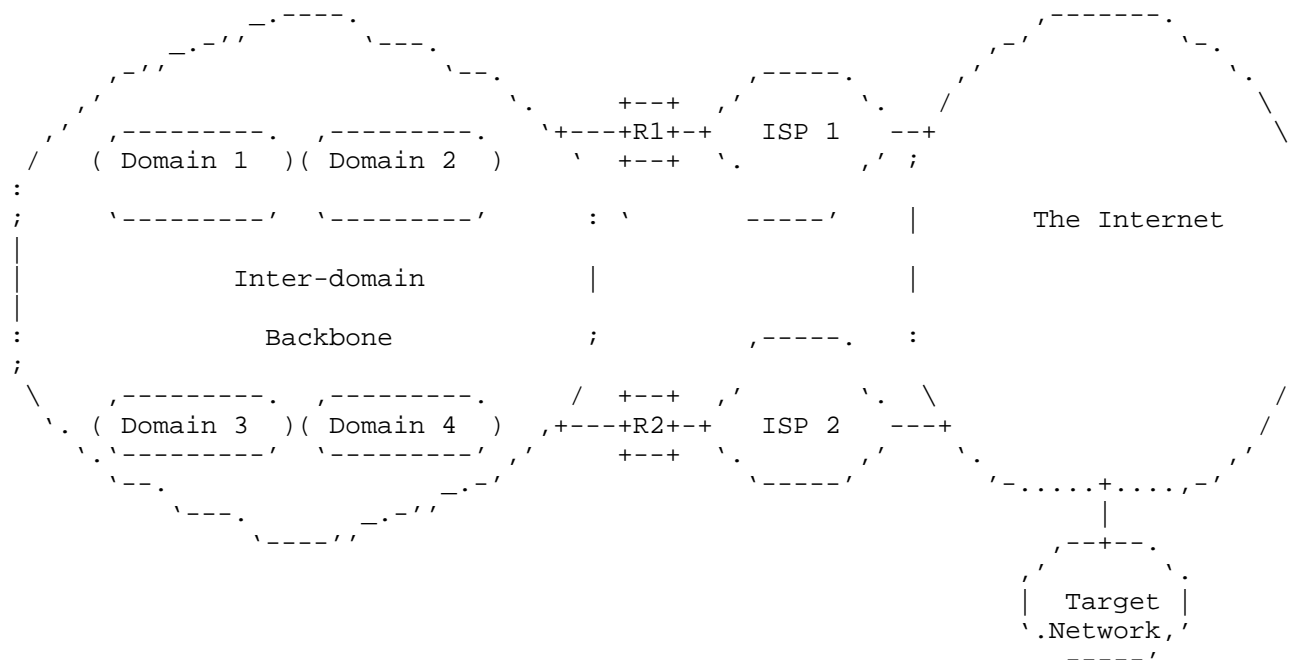


Figure 5: Traffic Engineering

3. Derived Requirements

The use cases in can each be met if:

- o The routing protocol or mechanism includes a source prefix. It is acceptable that a default source prefix of `::/0` (all addresses) applies to routes that don't specify a prefix.
- o The routing protocol or mechanism includes a destination prefix, which may be a default route (`::/0`) or any more specific prefix up to and including a host route (`/128`).
- o The FIB lookup yields the route with the most specific (e.g. longest-match) destination prefix that also matches the source prefix constraint, or no match.

4. IANA Considerations

This memo asks the IANA for no new parameters.

5. Security Considerations

As a descriptive document, this note adds no new security risks to the network.

6. Privacy Considerations

As a descriptive document, this note adds no new privacy risks to the network.

7. Acknowledgements

This note was discussed with Acee Lindem, Jianping Wu, Juliusz Chroboczek, Les Ginsberg, Lorenzo Colitti, Mark Townsley, Markus Stenberg, Matthieu Boutier, Ole Troan, Ray Bellis, Shu Yang, and Xia Yin.

8. References

8.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <<http://www.rfc-editor.org/info/rfc2119>>.

8.2. Informative References

- [I-D.baker-fun-routing-class]
Baker, F., "Routing a Traffic Class", draft-baker-fun-routing-class-00 (work in progress), July 2011.
- [I-D.baker-ipv6-isis-dst-src-routing]
Baker, F. and D. Lamparter, "IPv6 Source/Destination Routing using IS-IS", draft-baker-ipv6-isis-dst-src-routing-05 (work in progress), April 2016.
- [I-D.baker-ipv6-ospf-dst-src-routing]
Baker, F., "IPv6 Source/Destination Routing using OSPFv3", draft-baker-ipv6-ospf-dst-src-routing-03 (work in progress), August 2013.
- [I-D.boutier-homenet-source-specific-routing]
Boutier, M. and J. Chroboczek, "Source-specific Routing", draft-boutier-homenet-source-specific-routing-00 (work in progress), July 2013.
- [I-D.troan-homenet-sadr]
Troan, O. and L. Colitti, "IPv6 Multihoming with Source Address Dependent Routing (SADR)", draft-troan-homenet-sadr-01 (work in progress), September 2013.

[I-D.xu-homenet-traffic-class]

Xu, M., Yang, S., Wu, J., and F. Baker, "Traffic Class Routing Protocol in Home Networks", draft-xu-homenet-traffic-class-02 (work in progress), April 2014.

[I-D.xu-src-dst-bgp]

Xu, M., Yang, S., and J. Wu, "Source/Destination Routing Using BGP-4", draft-xu-src-dst-bgp-00 (work in progress), March 2016.

[RFC2827] Ferguson, P. and D. Senie, "Network Ingress Filtering: Defeating Denial of Service Attacks which employ IP Source Address Spoofing", BCP 38, RFC 2827, DOI 10.17487/RFC2827, May 2000, <<http://www.rfc-editor.org/info/rfc2827>>.

[RFC3704] Baker, F. and P. Savola, "Ingress Filtering for Multihomed Networks", BCP 84, RFC 3704, DOI 10.17487/RFC3704, March 2004, <<http://www.rfc-editor.org/info/rfc3704>>.

Appendix A. Change Log

Initial Version: August 2013

Repost: October 2014, initial draft reposted on request.

CERNET2: April 2016, CERNET2 use cases added.

Authors' Addresses

Fred Baker
Cisco Systems
Santa Barbara, California 93117
USA

Email: fred@cisco.com

Mingwei Xu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-1572
Email: xumw@tsinghua.edu.cn

Shu Yang
Graduate School at Shenzhen, Tsinghua University
Division of Information Science and Technology
Shenzhen 518055
P.R. China

Phone: +86-755-2603-6059
Email: yang.shu@sz.tsinghua.edu.cn

Jianping Wu
Tsinghua University
Department of Computer Science, Tsinghua University
Beijing 100084
P.R. China

Phone: +86-10-6278-5983
Email: jianping@cernet.edu.cn

RTGWG
Internet-Draft
Intended status: Informational
Expires: November 15, 2014

S. Ning
Tata Communications
A. Malis
Consultant
D. McDysan
Verizon
L. Yong
Huawei USA
C. Villamizar
Outer Cape Cod Network Consulting
May 14, 2014

Advanced Multipath Use Cases and Design Considerations
draft-ietf-rtgwg-cl-use-cases-06

Abstract

Advanced Multipath is a formalization of multipath techniques currently in use in IP and MPLS networks and a set of extensions to existing multipath techniques.

This document provides a set of use cases and design considerations for Advanced Multipath. Existing practices are described. Use cases made possible through Advanced Multipath extensions are described.

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 <http://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 November 15, 2014.

Copyright Notice

Copyright (c) 2014 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Assumptions	3
3. Terminology	3
4. Multipath Foundation Use Cases	5
5. Advanced Multipath Use Cases	8
5.1. Delay Sensitive Applications	8
5.2. Large Volume of IP and LDP Traffic	9
5.3. Multipath and Packet Ordering	9
5.3.1. MPLS-TP in network edges only	11
5.3.2. Multipath at core LSP ingress/egress	12
5.3.3. MPLS-TP as a MPLS client	13
6. IANA Considerations	13
7. Security Considerations	14
8. Acknowledgments	14
9. Informative References	14
Appendix A. Network Operator Practices and Protocol Usage	17
Appendix B. Existing Multipath Standards and Techniques	19
B.1. Common Multipath Load Splitting Techniques	19
B.2. Static and Dynamic Load Balancing Multipath	20
B.3. Traffic Split over Parallel Links	21
B.4. Traffic Split over Multiple Paths	21
Appendix C. Characteristics of Transport in Core Networks	22
Authors' Addresses	24

1. Introduction

Advanced Multipath requirements are specified in [RFC7226]. An Advanced Multipath framework is defined in [I-D.ietf-rtgwg-cl-framework].

Multipath techniques have been widely used in IP networks for over two decades. The use of MPLS began more than a decade ago. Multipath has been widely used in IP/MPLS networks for over a decade with very little protocol support dedicated to effective use of multipath.

The state of the art in multipath prior to Advanced Multipath is documented in Appendix B.

Both Ethernet Link Aggregation [IEEE-802.1AX] and MPLS link bundling [RFC4201] have been widely used in today's MPLS networks. Advanced Multipath differs in the following characteristics.

1. Advanced Multipath allows bundling of non-homogenous links together as a single logical link.
2. Advanced Multipath provides more information in the TE-LSDB and supports more explicit control over placement of LSP.

2. Assumptions

The supported services are, but not limited to, pseudowire (PW) based services ([RFC3985]), including Virtual Private Network (VPN) services, Internet traffic encapsulated by at least one MPLS label ([RFC3032]), and dynamically signaled MPLS ([RFC3209] or [RFC5036]) or MPLS-TP Label Switched Paths (LSPs) ([RFC5921]).

The MPLS LSPs supporting these services may be point-to-point, point-to-multipoint, or multipoint-to-multipoint. The MPLS LSPs may be signaled using RSVP-TE [RFC3209] or LDP [RFC5036]. With RSVP-TE, extensions to Interior Gateway Protocols (IGPs) may be used, specifically to OSPF-TE [RFC3630] or ISIS-TE [RFC5305].

The locations in a network where these requirements apply are a Label Edge Router (LER) or a Label Switch Router (LSR) as defined in [RFC3031].

The IP DSCP field [RFC2474] [RFC2475] cannot be used for flow identification since L3VPN requires Diffserv transparency (see RFC 4031 5.5.2 [RFC4031]), and in general network operators do not rely on the DSCP of Internet packets.

3. Terminology

Terminology defined in [RFC7226] and [RFC7190] is used in this document.

In addition, the following terms are used:

classic multipath:

Classic multipath refers to the most common current practice in implementation and deployment of multipath (see Appendix B). The most common current practice when applied to MPLS traffic makes use of a hash on the MPLS label stack, and if IPv4 or IPv6 are

indicated under the label stack, makes use of the IP source and destination addresses [RFC4385] [RFC4928].

classic link bundling:

Classic link bundling refers to the use of [RFC4201] where the "all ones" component is not used. Where the "all ones" component is used, link bundling behaves as classic multipath does. Classic link bundling selects a single component link to carry all of the traffic for a given LSP.

Among the important distinctions between classic multipath or classic link bundling and Advanced Multipath are:

1. Classic multipath has no provision to retain packet order within any specific LSP. Classic link bundling retains packet order among any given LSP but as a result does a poor job of splitting load among components and therefore is rarely (if ever) deployed. Advanced Multipath allows per LSP control of load split characteristics.
2. Classic multipath and classic link bundling do not provide a means to put some LSP on component links with lower delay. Advanced Multipath does.
3. Classic multipath will provide a load balance for IP and LDP traffic. Classic link bundling will not. Neither classic multipath or classic link bundling will measure IP and LDP traffic and reduce the RSVP-TE advertised "Available Bandwidth" as a result of that measurement. Advanced Multipath better supports RSVP-TE used with significant traffic levels of native IP and native LDP.
4. Classic link bundling cannot support an LSP that is greater in capacity than any single component link. Classic multipath supports this capability but may reorder traffic on such an LSP. Advanced Multipath can retain order of an LSP that is carried within an LSP that is greater in capacity than any single component link if the contained LSP has such a requirement.

None of these techniques, classic multipath, classic link bundling, or Advanced Multipath, will reorder traffic among IP microflows. None of these techniques will reorder traffic among PW, if a PWE3 Control Word is used [RFC4385].

4. Multipath Foundation Use Cases

A simple multipath composed entirely of physical links is illustrated in Figure 1, where an multipath is configured between LSR1 and LSR2. This multipath has three component links. Individual component links in a multipath may be supported by different transport technologies such as SONET, OTN, Ethernet, etc. Even if the transport technology implementing the component links is identical, the characteristics (e.g., bandwidth, latency) of the component links may differ.

The multipath in Figure 1 may carry LSP traffic flows and control plane packets. Control plane packets may appear as IP packets or may be carried within a generic associated channel (G-Ach) [RFC5586]. A LSP may be established over the link by either RSVP-TE [RFC3209] or LDP [RFC5036] signaling protocols. All component links in a multipath are summarized in the same forwarding adjacency LSP (FA-LSP) routing advertisement [RFC3945]. The multipath is summarized as one TE-Link advertised into the IGP by the multipath end points (the LER if the multipath is MPLS based). This information is used in path computation when a full MPLS control plane is in use.

If Advanced Multipath techniques are used, then the individual component links or groups of component links may optionally be advertised into the IGP as sub-TLV of the multipath FA advertisement to indicate capacity available with various characteristics, such as a delay range.

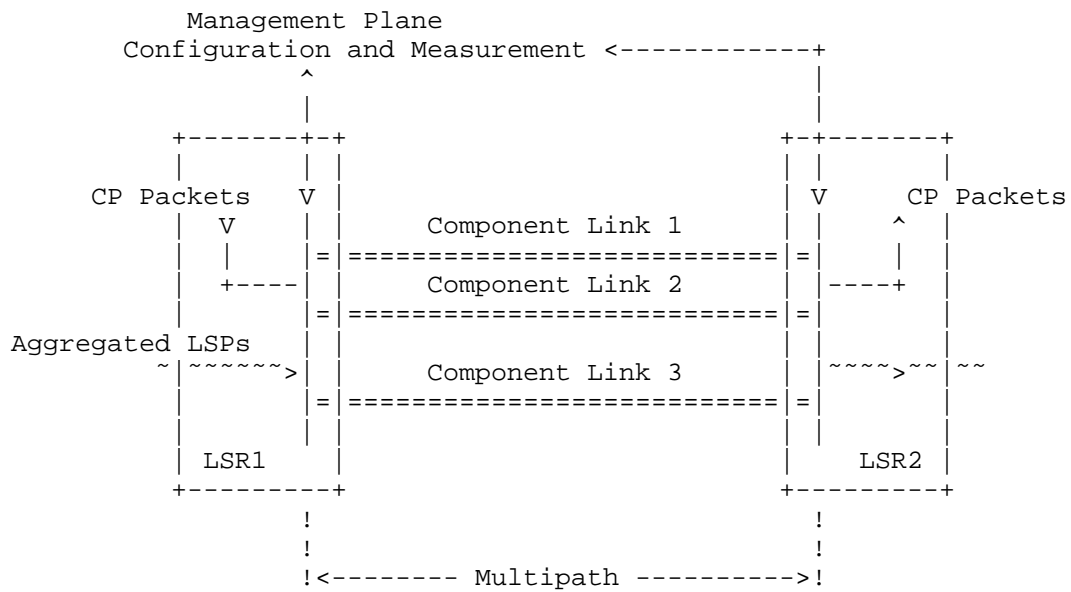


Figure 1: a multipath constructed with multiple physical links between two LSR

[RFC7226] specifies that component links may themselves be multipath. This is true for most implementations even prior to the Advanced Multipath work in [RFC7226]. For example, a component of a pre-Advanced Multipath MPLS Link Bundle or ISIS or OSPF ECMP could be an Ethernet LAG. In some implementations many other combinations or even arbitrary combinations could be supported. Figure 2 shows three forms of component links which may be deployed in a network.

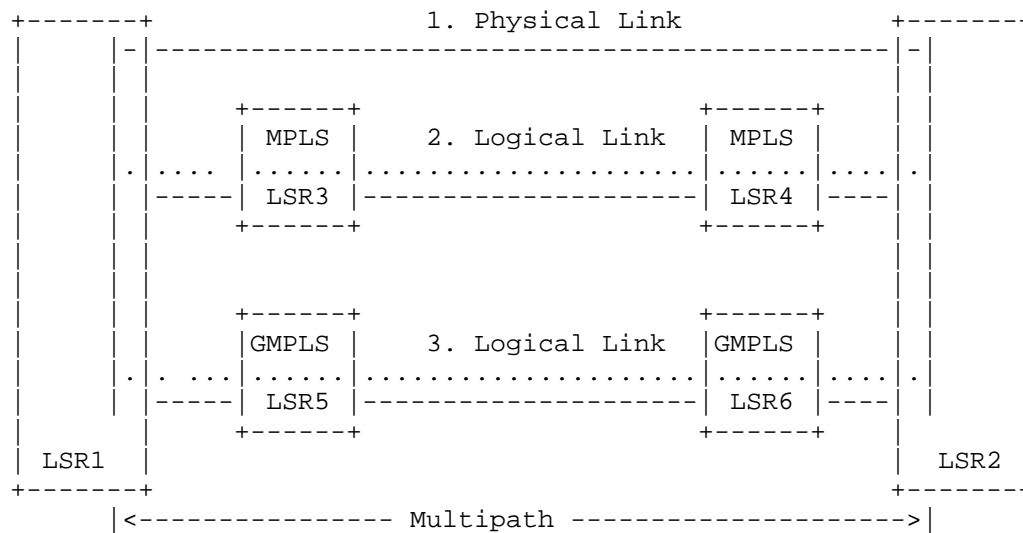


Figure 2: Illustration of Various Component Link Types

The three forms of component link shown in Figure 2 are:

1. The first component link is configured with direct physical media plus a link layer protocol. This case also includes emulated physical links, for example using pseudowire emulation.
2. The second component link is a TE tunnel that traverses LSR3 and LSR4, where LSR3 and LSR4 are the nodes supporting MPLS, but supporting few or no GMPLS extensions.
3. The third component link is formed by lower layer network that has GMPLS enabled. In this case, LSR5 and LSR6 are not the nodes controlled by the MPLS but provide the connectivity for the component link.

A multipath forms one logical link between connected LSR (LSR1 and LSR2 in Figure 1 and Figure 2) and is used to carry aggregated traffic. Multipath relies on its component links to carry the traffic but must distribute or load balance the traffic. The endpoints of the multipath maps incoming traffic into the set of component links.

For example, LSR1 in Figure 1 distributes the set of traffic flows including control plane packets among the set of component links. LSR2 in Figure 1 receives the packets from its component links and sends them to MPLS forwarding engine with no attempt to reorder packets arriving on different component links. The traffic in the

opposite direction, from LSR2 to LSR1, is distributed across the set of component links by the LSR2.

These three forms of component link are a limited set of very simple examples. Many other examples are possible. A component link may itself be a multipath. A segment of an LSP (single hop for that LSP) may be a multipath.

5. Advanced Multipath Use Cases

The following subsections provide some uses of the Advanced Multipath extensions. These are not the only uses, simply a set of examples.

5.1. Delay Sensitive Applications

Most applications benefit from lower delay. Some types of applications are far more sensitive than others. For example, real time bidirectional applications such as voice communication or two way video conferencing are far more sensitive to delay than unidirectional streaming audio or video. Non-interactive bulk transfer is almost insensitive to delay if a large enough TCP window is used.

Some applications are sensitive to delay but users of those applications are unwilling to pay extra to insure lower delay. For example, many SIP end users are willing to accept the delay offered to best effort services as long as call quality is good most of the time.

Other applications are sensitive to delay and willing to pay extra to insure lower delay. For example, financial trading applications are extremely sensitive to delay and with a lot at stake are willing to go to great lengths to reduce delay.

Among the requirements of Advanced Multipath are requirements to support non-homogeneous links. One solution in support of lower delay links is to advertise capacity available within configured ranges of delay within a given multipath and then support the ability to place an LSP only on component links that meeting that LSP's delay requirements.

The Advanced Multipath requirements to accommodate delay sensitive applications are analogous to Diffserv requirements to accommodate applications requiring higher quality of service on the same infrastructure as applications with less demanding requirements. The ability to share capacity with less demanding applications, with best effort applications generally being the least demanding, can greatly

reduce the cost of delivering service to the more demanding applications.

5.2. Large Volume of IP and LDP Traffic

IP and LDP do not support traffic engineering. Both make use of a shortest (lowest routing metric) path, with an option to use equal cost multipath (ECMP). Note that though ECMP is prohibited in LDP specifications, it is widely implemented. Where implemented for LDP, ECMP is generally disabled by default for standards compliance, but often enabled in LDP deployments.

Without traffic engineering capability, there must be sufficient capacity to accommodate the IP and LDP traffic. If not, persistent queuing delay and loss will occur. Unlike RSVP-TE, a subset of traffic cannot be routed using constraint based routing to avoid a congested portion of an infrastructure.

In existing networks which accommodate IP and/or LDP with RSVP-TE, either the IP and LDP can be carried over RSVP-TE, or where the traffic contribution of IP and LDP is small, IP and LDP can be carried native and the effect on RSVP-TE can be ignored. Ignoring the traffic contribution of IP is valid on high capacity networks where a very low volume of native IP is used primarily for control and network management and customer IP is carried within RSVP-TE.

Where it is desirable to carry native IP and/or LDP and IP and/or LDP traffic volumes are not negligible, RSVP-TE needs improvement. An enhancement offered by Advanced Multipath is an ability to measure the IP and LDP, filter the measurements, and reduce the capacity available to RSVP-TE to avoid congestion. The treatment given to the IP or LDP traffic is similar to the treatment when using the "auto-bandwidth" feature in some RSVP-TE implementations on that same traffic, and giving a higher priority (numerically lower setup priority and holding priority value) to the "auto-bandwidth" LSP. The difference is that the measurement is made at each hop and the reduction in advertised bandwidth is made more directly.

5.3. Multipath and Packet Ordering

A strong motivation for multipath is the need to provide LSP capacity in IP backbones that exceeds the capacity of single wavelengths provided by transport equipment and exceeds the practical capacity limits achievable through inverse multiplexing. Appendix C describes characteristics and limitations of transport systems today. Section 3 defines the terms "classic multipath" and "classic link bundling" used in this section.

For purpose of discussion, consider two very large cities, city A and city Z. For example, in the US high traffic cities might be New York and Los Angeles and in Europe high traffic cities might be London and Amsterdam. Two other high volume cities, city B and city Y may share common provider core network infrastructure. Using the same examples, the city B and Y may Washington DC and San Francisco or Paris and Stockholm. In the US, the common infrastructure may span Denver, Chicago, Detroit, and Cleveland. Other major traffic contributors on either US coast include Boston, northern Virginia on the east coast, and Seattle, and San Diego on the west coast. The capacity of IP/MPLS links within the shared infrastructure, for example city to city links in the Denver, Chicago, Detroit, and Cleveland path in the US example, have capacities for most of the 2000s decade that greatly exceeded single circuits available in transport networks.

For a case with four large traffic sources on either side of the shared infrastructure, up to sixteen core city to core city traffic flows in excess of transport circuit capacity may be accommodated on the shared infrastructure.

Today the most common IP/MPLS core network design makes use of very large links which consist of many smaller component links, but use classic multipath techniques. A component link typically corresponds to the largest circuit that the transport system is capable of providing (or the largest cost effective circuit). IP source and destination address hashing is used to distribute flows across the set of component links as described in Appendix B.3.

Classic multipath can handle large LSP up to the total capacity of the multipath (within limits, see Appendix B.2). A disadvantage of classic multipath is the reordering among traffic within a given core city to core city LSP. While there is no reordering within any microflow and therefore no customer visible issue, MPLS-TP cannot be used across an infrastructure where classic multipath is in use, except within pseudowires.

Capacity issues force the use of classic multipath today. Classic multipath excludes a direct use of MPLS-TP. The desire for OAM, offered by MPLS-TP, is in conflict with the use of classic multipath. There are a number of alternatives that satisfy both requirements. Some alternatives are described below.

MPLS-TP in network edges only

A simple approach which requires no change to the core is to disallow MPLS-TP across the core unless carried within a pseudowire (PW). MPLS-TP may be used within edge domains where

classic multipath is not used. PW may be signaled end to end using single segment PW (SS-PW), or stitched across domains using multisegment PW (MS-PW). The PW and anything carried within the PW may use OAM as long as fat-PW [RFC6391] load splitting is not used by the PW.

Advanced Multipath at core LSP ingress/egress

The interior of the core network may use classic link bundling, with the limitation that no LSP can exceed the capacity of a single circuit. Larger non-MPLS-TP LSP can be configured using multiple ingress to egress component MPLS-TP LSP. This can be accomplished using existing IP source and destination address hashing configured at LSP ingress and egress. Each component LSP, if constrained to be no larger than the capacity of a single circuit, can make use of MPLS-TP and offer OAM for all top level LSP across the core.

MPLS-TP as a MPLS client

A third approach involves making use of Entropy Labels [RFC6790] on all MPLS-TP LSP such that the entire MPLS-TP LSP is treated as a microflow by midpoint LSR, even if further encapsulated in very large server layer MPLS LSP.

The above list of alternatives allow packet ordering within an LSP to be maintained in some circumstances and allow very large LSP capacities. Each of these alternatives are discussed further in the following subsections.

5.3.1. MPLS-TP in network edges only

Classic MPLS link bundling is defined in [RFC4201] and has existed since early in the 2000s decade. Classic MPLS link bundling place any given LSP entirely on a single component link. Classic MPLS link bundling is not in widespread use as the means to accommodate large link capacities in core networks due to the simplicity and better multiplexing gain, and therefore lower network cost of classic multipath.

If MPLS-TP OAM capability in the IP/MPLS network core LSP is not required, then there is no need to change existing network designs which use classic multipath and both label stack and IP source and destination address based hashing as a basis for load splitting.

If MPLS-TP is needed for a subset of LSP, then those LSP can be carried within pseudowires. The pseudowires adds a thin layer of encapsulation and therefore a small overhead. If only a subset of

LSP need MPLS-TP OAM, then some LSP must make use of the pseudowires and other LSP avoid them. A straightforward way to accomplish this is with administrative attributes [RFC3209].

5.3.2. Multipath at core LSP ingress/egress

Multipath can be configured for large LSP that are made of smaller MPLS-TP component LSP. Some implementations already support this capability, though until Advanced Multipath no IETF document required it. This approach is capable of supporting MPLS-TP OAM over the entire set of component link LSP and therefore the entire set of top level LSP traversing the core.

There are two primary disadvantage of this approach. One is the number of top level LSP traversing the core can be dramatically increased. The other disadvantage is the loss of multiplexing gain that results from use of classic link bundling within the interior of the core network.

If component LSP use MPLS-TP, then no component LSP can exceed the capacity of a single circuit. For a given multipath LSP there can either be a number of equal capacity component LSP or some number of full capacity component links plus one LSP carrying the excess. For example, a 350 Gb/s multipath LSP over a 100 Gb/s infrastructure may use five 70 Gb/s component LSP or three 100 Gb/s LSP plus one 50 Gb/s LSP. Classic MPLS link bundling is needed to support MPLS-TP and suffers from a bin packing problem even if LSP traffic is completely predictable, which it never is in practice.

The common means of setting very large LSP link bandwidth parameters uses long term statistical measures. For example, at one time many providers based their LSP bandwidth parameters on the 95th percentile of carried traffic as measured over the prior one week period. It is common to add 10-30% to the 95th percentile value measured over the prior week and adjust bandwidth parameters of LSP weekly. It is also possible to measure traffic flow at the LSR and adjust bandwidth parameters somewhat more dynamically. This is less common in deployments and where deployed, makes use of filtering to track very long term trends in traffic levels. In either case, short term variation of traffic levels relative to signaled LSP capacity are common. Allowing a large over allocation of LSP bandwidth parameters (ie: adding 30% or more) avoids over utilization of any given LSP, but increases unused network capacity and increases network cost. Allowing a small over allocation of LSP bandwidth parameters (ie: 10-20% or less) results in both underutilization and over utilization but statistically results in a total utilization within the core that is under capacity most or all of the time.

The classic multipath solution accommodates the situation in which some very large LSP are under utilizing their signaled capacity and others are over utilizing their capacity with the need for far less unused network capacity to accommodate variation in actual traffic levels. If the actual traffic levels of LSP can be described by a probability distribution, the variation of the sum of LSP is less than the variation of any given LSP for all but a constant traffic level (where the variation of the sum and the variation of the components are both zero).

Splitting very large LSP at the ingress and carrying those large LSP within smaller MPLS-TP component LSP and then using classic link bundling to carry the MPLS-TP LSP is a viable approach. However this approach loses the statistical gain discussed in the prior paragraphs. Losing this statistical gain drives up network costs necessary to achieve the same very low probability of only mild congestion that is expected of provider networks.

There are two situations which can motivate the use of this approach. This design is favored if the provider values MPLS-TP OAM across the core more than efficiency (or is unaware of the efficiency issue). This design can also make sense if transport equipment or very low cost core LSR are available which support only classic link bundling and regardless of loss of multiplexing gain, are more cost effective at carrying transit traffic than using equipment which supports IP source and destination address hashing.

5.3.3. MPLS-TP as a MPLS client

Accommodating MPLS-TP as a MPLS client requires the small change to forwarding behavior necessary to support [RFC6790] and is therefore most applicable to major network overbuilds or new deployments. This approach is described in [RFC7190] and makes use of Entropy Labels [RFC6790] to prevent reordering of MPLS-TP LSP or any other LSP which requires that its traffic not be reordered for OAM or other reasons.

The advantage of this approach is an ability to accommodate MPLS-TP as a client LSP but retain the high multiplexing gain and therefore efficiency and low network cost of a pure MPLS deployment. The disadvantage is the need for a small change in forwarding to support [RFC6790].

6. IANA Considerations

This memo includes no request to IANA.

7. Security Considerations

This document is a use cases document. Existing protocols are referenced such as MPLS. Existing techniques such as MPLS link bundling and multipath techniques are referenced. These protocols and techniques are documented elsewhere and contain security considerations which are unchanged by this document.

This document also describes use cases for multipath and Advanced Multipath. Advanced Multipath requirements are defined in [RFC7226]. [I-D.ietf-rtgwg-cl-framework] defines a framework for Advanced Multipath. Advanced Multipath bears many similarities to MPLS link bundling and multipath techniques used with MPLS. Additional security considerations, if any, beyond those already identified for MPLS, MPLS link bundling and multipath techniques, will be documented in the framework document if specific to the overall framework of Advanced Multipath, or in protocol extensions if specific to a given protocol extension defined later to support Advanced Multipath.

8. Acknowledgments

In the interest of full disclosure of affiliation and in the interest of acknowledging sponsorship, past affiliations of authors are noted. Much of the work done by Ning So occurred while Ning was at Verizon. Much of the work done by Curtis Villamizar occurred while at Infinera. Much of the work done by Andy Malis occurred while Andy was at Verizon.

9. Informative References

[I-D.ietf-rtgwg-cl-framework]

Ning, S., McDysan, D., Osborne, E., Yong, L., and C. Villamizar, "Advanced Multipath Framework in MPLS", draft-ietf-rtgwg-cl-framework-04 (work in progress), July 2013.

[IEEE-802.1AX]

IEEE Standards Association, "IEEE Std 802.1AX-2008 IEEE Standard for Local and Metropolitan Area Networks - Link Aggregation", 2006, <<http://standards.ieee.org/getieee802/download/802.1AX-2008.pdf>>.

[ITU-T.G.694.2]

ITU-T, "Spectral grids for WDM applications: CWDM wavelength grid", 2003, <<http://www.itu.int/rec/T-REC-G.694.2-200312-I>>.

[RFC1717]

Sklower, K., Lloyd, B., McGregor, G., and D. Carr, "The PPP Multilink Protocol (MP)", RFC 1717, November 1994.

- [RFC2474] Nichols, K., Blake, S., Baker, F., and D. Black, "Definition of the Differentiated Services Field (DS Field) in the IPv4 and IPv6 Headers", RFC 2474, December 1998.
- [RFC2475] Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., and W. Weiss, "An Architecture for Differentiated Services", RFC 2475, December 1998.
- [RFC2597] Heinanen, J., Baker, F., Weiss, W., and J. Wroclawski, "Assured Forwarding PHB Group", RFC 2597, June 1999.
- [RFC2615] Malis, A. and W. Simpson, "PPP over SONET/SDH", RFC 2615, June 1999.
- [RFC2991] Thaler, D. and C. Hopps, "Multipath Issues in Unicast and Multicast Next-Hop Selection", RFC 2991, November 2000.
- [RFC2992] Hopps, C., "Analysis of an Equal-Cost Multi-Path Algorithm", RFC 2992, November 2000.
- [RFC3031] Rosen, E., Viswanathan, A., and R. Callon, "Multiprotocol Label Switching Architecture", RFC 3031, January 2001.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC3209] Awduche, D., Berger, L., Gan, D., Li, T., Srinivasan, V., and G. Swallow, "RSVP-TE: Extensions to RSVP for LSP Tunnels", RFC 3209, December 2001.
- [RFC3260] Grossman, D., "New Terminology and Clarifications for Diffserv", RFC 3260, April 2002.
- [RFC3270] Le Faucheur, F., Wu, L., Davie, B., Davari, S., Vaananen, P., Krishnan, R., Cheval, P., and J. Heinanen, "Multi-Protocol Label Switching (MPLS) Support of Differentiated Services", RFC 3270, May 2002.
- [RFC3630] Katz, D., Kompella, K., and D. Yeung, "Traffic Engineering (TE) Extensions to OSPF Version 2", RFC 3630, September 2003.
- [RFC3809] Nagarajan, A., "Generic Requirements for Provider Provisioned Virtual Private Networks (PPVPN)", RFC 3809, June 2004.

- [RFC3945] Mannie, E., "Generalized Multi-Protocol Label Switching (GMPLS) Architecture", RFC 3945, October 2004.
- [RFC3985] Bryant, S. and P. Pate, "Pseudo Wire Emulation Edge-to-Edge (PWE3) Architecture", RFC 3985, March 2005.
- [RFC4031] Carugi, M. and D. McDysan, "Service Requirements for Layer 3 Provider Provisioned Virtual Private Networks (PPVPNs)", RFC 4031, April 2005.
- [RFC4124] Le Faucheur, F., "Protocol Extensions for Support of Diffserv-aware MPLS Traffic Engineering", RFC 4124, June 2005.
- [RFC4201] Kompella, K., Rekhter, Y., and L. Berger, "Link Bundling in MPLS Traffic Engineering (TE)", RFC 4201, October 2005.
- [RFC4385] Bryant, S., Swallow, G., Martini, L., and D. McPherson, "Pseudowire Emulation Edge-to-Edge (PWE3) Control Word for Use over an MPLS PSN", RFC 4385, February 2006.
- [RFC4928] Swallow, G., Bryant, S., and L. Andersson, "Avoiding Equal Cost Multipath Treatment in MPLS Networks", BCP 128, RFC 4928, June 2007.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5586] Bocci, M., Vigoureux, M., and S. Bryant, "MPLS Generic Associated Channel", RFC 5586, June 2009.
- [RFC5921] Bocci, M., Bryant, S., Frost, D., Levrau, L., and L. Berger, "A Framework for MPLS in Transport Networks", RFC 5921, July 2010.
- [RFC6391] Bryant, S., Filsfils, C., Drafz, U., Kompella, V., Regan, J., and S. Amante, "Flow-Aware Transport of Pseudowires over an MPLS Packet Switched Network", RFC 6391, November 2011.
- [RFC6790] Kompella, K., Drake, J., Amante, S., Henderickx, W., and L. Yong, "The Use of Entropy Labels in MPLS Forwarding", RFC 6790, November 2012.

- [RFC7190] Villamizar, C., "Use of Multipath with MPLS and MPLS Transport Profile (MPLS-TP)", RFC 7190, March 2014.
- [RFC7226] Villamizar, C., McDysan, D., Ning, S., Malis, A., and L. Yong, "Requirements for Advanced Multipath in MPLS Networks", RFC 7226, May 2014.

Appendix A. Network Operator Practices and Protocol Usage

Often, network operators have a contractual Service Level Agreement (SLA) with customers for services that are comprised of numerical values for performance measures, principally availability, latency, delay variation. Additionally, network operators may have performance objectives for internal use by the operator. See RFC3809, Section 4.9 [RFC3809] for examples of the form of such SLA and performance objective specifications. In this document we use the term Performance Objective as defined in [RFC7226]. Applications and acceptable user experience have an important relationship to these performance parameters.

Consider latency as an example. In some cases, minimizing latency relates directly to the best customer experience (for example, in interactive applications closer is faster). In other cases, user experience is relatively insensitive to latency, up to a specific limit at which point user perception of quality degrades significantly (e.g., interactive human voice and multimedia conferencing). A number of Performance Objectives have a bound on point-to-point latency and as long as this bound is met the Performance Objective is met; decreasing the latency is not necessary. In some Performance Objectives, if the specified latency is not met, the user considers the service as unavailable. An unprotected LSP can be manually provisioned on a set of links to meet this type of Performance Objective, but this lowers availability since an alternate route that meets the latency Performance Objective cannot be determined.

Historically, when an IP/MPLS network was operated over a lower layer circuit switched network (e.g., SONET rings), a change in latency caused by the lower layer network (e.g., due to a maintenance action or failure) was not known to the MPLS network. This resulted in latency affecting end user experience, sometimes violating Performance Objectives or resulting in user complaints.

A response to this problem was to provision IP/MPLS networks over unprotected circuits and set the metric and/or TE-metric proportional to latency. This resulted in traffic being directed over the least latency path, even if this was not needed to meet a Performance Objective or meet user experience objectives. This results in

reduced flexibility and increased cost for network operators. Some providers prefer to use lower layer networks to provide restoration and grooming, but the inability to communicate performance parameters, in particular latency, from the lower layer network to the higher layer network is an important problem to be solved before this can be done.

Latency Performance Objectives for point-to-point services are often tied closely to geographic locations, while latency for multipoint services may be based upon a worst case within a region.

The time frames for restoration (i.e., as implemented by predetermined protection, convergence of routing protocols and/or signaling) for services range from on the order of 100 ms or less (e.g., for VPWS to emulate classical SDH/SONET protection switching), to several minutes (e.g., to allow BGP to reconverge for L3VPN) and may differ among the set of customers within a single service.

The presence of only three Traffic Class (TC) bits (previously known as EXP bits) in the MPLS shim header is limiting when a network operator needs to support QoS classes for multiple services (e.g., L2VPN VPWS, VPLS, L3VPN and Internet), each of which has a set of QoS classes that need to be supported and where the operator prefers to use only E-LSP [RFC3270]. In some cases one bit is used to indicate conformance to some ingress traffic classification, leaving only two bits for indicating the service QoS classes. One approach that has been taken is to aggregate these QoS classes into similar sets on LER-LSR and LSR-LSR links and continue to use only E-LSP. Another approach is to use L-LSP as defined in [RFC3270] or use the Class-Type as defined in [RFC4124] to support up to eight mappings of TC into Per-Hop Behavior (PHB).

The IP DSCP cannot be used for flow identification. The use of IP DSCP for flow identification is incompatible with Assured Forwarding services [RFC2597] or any other service which may use more than one DSCP code point to carry traffic for a given microflow. In general network operators do not rely on the DSCP of Internet packets in core networks but must preserve DSCP values for use closer to network edges.

A label is pushed onto Internet packets when they are carried along with L2VPN or L3VPN packets on the same link or lower layer network provides a mean to distinguish between the QoS class for these packets.

Operating an MPLS-TE network involves a different paradigm from operating an IGP metric-based LDP signaled MPLS network. The multipoint-to-point LDP signaled MPLS LSPs occur automatically, and

balancing across parallel links occurs if the IGP metrics are set "equally" (with equality a locally definable relation) and if ECMP is enabled for LDP, which network operators generally do in large networks.

Traffic is typically comprised of large (some very large) flows and a much larger number of small flows. In some cases, separate LSPs are established for very large flows. Very large microflows can occur even if the IP header information is inspected by a LSR. For example an IPsec tunnel that carries a large amount of traffic must be carried as a single large flow. An important example of large flows is that of a L2VPN or L3VPN customer who has an access line bandwidth comparable to a client-client component link bandwidth -- there could be flows that are on the order of the access line bandwidth.

Appendix B. Existing Multipath Standards and Techniques

Today the requirement to handle large aggregations of traffic, much larger than a single component link, can be handled by a number of techniques which we will collectively call multipath. Multipath applied to parallel links between the same set of nodes includes Ethernet Link Aggregation [IEEE-802.1AX], link bundling [RFC4201], or other aggregation techniques some of which may be vendor specific. Multipath applied to diverse paths rather than parallel links includes Equal Cost MultiPath (ECMP) as applied to OSPF, ISIS, LDP, or even BGP, and equal cost LSP, as described in Appendix B.4. Various multipath techniques have strengths and weaknesses.

Existing multipath techniques solve the problem of large aggregations of traffic, without addressing the other requirements outlined in this document, particularly those described in Section 5.

B.1. Common Multipath Load Splitting Techniques

Identical load balancing techniques are used for multipath both over parallel links and over diverse paths.

Large aggregates of IP traffic do not provide explicit signaling to indicate the expected traffic loads. Large aggregates of MPLS traffic are carried in MPLS tunnels supported by MPLS LSP. LSP which are signaled using RSVP-TE extensions do provide explicit signaling which includes the expected traffic load for the aggregate. LSP which are signaled using LDP do not provide an expected traffic load.

MPLS LSP may contain other MPLS LSP arranged hierarchically. When an MPLS LSR serves as a midpoint LSR in an LSP carrying client LSP as payload, there is no signaling associated with these client LSP. Therefore even when using RSVP-TE signaling there may be insufficient

information provided by signaling to adequately distribute load based solely on signaling.

Generally a set of label stack entries that is unique across the ordered set of label numbers in the label stack can safely be assumed to contain a group of flows. The reordering of traffic can therefore be considered to be acceptable unless reordering occurs within traffic containing a common unique set of label stack entries. Existing load splitting techniques take advantage of this property in addition to looking beyond the bottom of the label stack and determining if the payload is IPv4 or IPv6 to load balance traffic accordingly.

MPLS-TP OAM violates the assumption that it is safe to reorder traffic within an LSP. If MPLS-TP OAM is to be accommodated, then existing multipath techniques must be modified. [RFC6790] and [RFC7190] provide a solution but require a small forwarding change.

For example, a large aggregate of IP traffic may be subdivided into a large number of groups of flows using a hash on the IP source and destination addresses. This is as described in [RFC2475] and clarified in [RFC3260]. For MPLS traffic carrying IP, a similar hash can be performed on the set of labels in the label stack. These techniques are both examples of means to subdivide traffic into groups of flows for the purpose of load balancing traffic across aggregated link capacity. The means of identifying a group of flows should not be confused with the definition of a flow.

Discussion of whether a hash based approach provides a sufficiently even load balance using any particular hashing algorithm or method of distributing traffic across a set of component links is outside of the scope of this document.

The current load balancing techniques are referenced in [RFC4385] and [RFC4928]. The use of three hash based approaches are described in [RFC2991] and [RFC2992]. A mechanism to identify flows within PW is described in [RFC6391]. The use of hash based approaches is mentioned as an example of an existing set of techniques to distribute traffic over a set of component links. Other techniques are not precluded.

B.2. Static and Dynamic Load Balancing Multipath

Static multipath generally relies on the mathematical probability that given a very large number of small microflows, these microflows will tend to be distributed evenly across a hash space. Early very static multipath implementations assumed that all component links are of equal capacity and perform a modulo operation across the hashed

value. An alternate static multipath technique uses a table generally with a power of two size, and distributes the table entries proportionally among component links according to the capacity of each component link.

Static load balancing works well if there are a very large number of small microflows (i.e., microflow rate is much less than component link capacity). However, the case where there are even a few large microflows is not handled well by static load balancing.

A dynamic load balancing multipath technique is one where the traffic bound to each component link is measured and the load split is adjusted accordingly. As long as the adjustment is done within a single network element, then no protocol extensions are required and there are no interoperability issues.

Note that if the load balancing algorithm and/or its parameters is adjusted, then packets in some flows may be briefly delivered out of sequence, however in practice such adjustments can be made very infrequent.

B.3. Traffic Split over Parallel Links

The load splitting techniques defined in Appendix B.1 and Appendix B.2 are both used in splitting traffic over parallel links between the same pair of nodes. The best known technique, though far from being the first, is Ethernet Link Aggregation [IEEE-802.1AX]. This same technique had been applied much earlier using OSPF or ISIS Equal Cost MultiPath (ECMP) over parallel links between the same nodes. Multilink PPP [RFC1717] uses a technique that provides inverse multiplexing, however a number of vendors had provided proprietary extensions to PPP over SONET/SDH [RFC2615] that predated Ethernet Link Aggregation but are no longer used.

Link bundling [RFC4201] provides yet another means of handling parallel LSP. RFC4201 explicitly allow a special value of all ones to indicate a split across all members of the bundle. This "all ones" component link is signaled in the MPLS RESV to indicate that the link bundle is making use of classic multipath techniques.

B.4. Traffic Split over Multiple Paths

OSPF or ISIS Equal Cost MultiPath (ECMP) is a well known form of traffic split over multiple paths that may traverse intermediate nodes. ECMP is often incorrectly equated to only this case, and multipath over multiple diverse paths is often incorrectly equated to ECMP.

Many implementations are able to create more than one LSP between a pair of nodes, where these LSP are routed diversely to better make use of available capacity. The load on these LSP can be distributed proportionally to the reserved bandwidth of the LSP. These multiple LSP may be advertised as a single PSC FA and any LSP making use of the FA may be split over these multiple LSP.

Link bundling [RFC4201] component links may themselves be LSP. When this technique is used, any LSP which specifies the link bundle may be split across the multiple paths of the component LSP that comprise the bundle.

Appendix C. Characteristics of Transport in Core Networks

The characteristics of primary interest are the capacity of a single circuit and the use of wave division multiplexing (WDM) to provide a large number of parallel circuits.

Wave division multiplexing (WDM) supports multiple independent channels (independent ignoring crosstalk noise) at slightly different wavelengths of light, multiplexed onto a single fiber. Typical in the early 2000s was 40 wavelengths of 10 Gb/s capacity per wavelength. These wavelengths are in the C-band range, which is about 1530-1565 nm, though some work has been done using the L-band 1565-1625 nm.

The C-band has been carved up using a 100 GHz spacing from 191.7 THz to 196.1 THz by [ITU-T.G.694.2]. This yields 44 channels. If the outermost channels are not used, due to poorer transmission characteristics, then typically 40 are used. For practical reasons, a 50 GHz or 25 GHz spacing is used by more recent equipment, yielding 80 or 160 channels in practice.

The early optical modulation techniques used within a single channel yielded 2.5Gb/s and 10 Gb/s capacity per channel. As modulation techniques have improved 40 Gb/s and 100 Gb/s per channel have been achieved.

The 40 channels of 10 Gb/s common in the mid 2000s yields a total of 400 Gb/s. Tighter spacing and better modulations are yielding up to 8 Tb/s or more in more recent systems.

Over the optical modulation is an electrical encoding. In the 1990s this was typically Synchronous Optical Networking (SONET) or Synchronous Digital Hierarchy (SDH), with a maximum defined circuit capacity of 40 Gb/s (OC-768), though the 10 Gb/s OC-192 is more common. More recently the low level electrical encoding has been Optical Transport Network (OTN) defined by ITU-T. OTN currently

defines circuit capacities up to a nominal 100 Gb/s (ODU4). Both SONET/SDH and OTN make use of time division multiplexing (TDM) where the a higher capacity circuit such as a 100 Gb/s ODU4 in OTN may be subdivided into lower fixed capacity circuits such as ten 10 Gb/s ODU2.

In the 1990s, all IP and later IP/MPLS networks either used a fraction of maximum circuit capacity, or at most the full circuit capacity toward the end of the decade, when full circuit capacity was 2.5 Gb/s or 10 Gb/s. Beyond 2000, the TDM circuit multiplexing capability of SONET/SDH or OTN was rarely used.

Early in the 2000s both transport equipment and core LSR offered 40 Gb/s SONET OC-768. However 10 Gb/s transport equipment was predominantly deployed throughout the decade, partially because LSR 10GbE ports were far more cost effective than either OC-192 or OC-768 and 10GbE became practical in the second half of the decade.

Entering the 2010 decade, LSR 40GbE and 100GbE are expected to become widely available and cost effective. Slightly preceding this transport equipment making use of 40 Gb/s and 100 Gb/s modulations are becoming available. This transport equipment is capable or carrying 40 Gb/s ODU3 and 100 Gb/s ODU4 circuits.

Early in the 2000s decade IP/MPLS core networks were making use of single 10 Gb/s circuits. Capacity grew quickly in the first half of the decade but more IP/MPLS core networks had only a small number of IP/MPLS links requiring 4-8 parallel 10 Gb/s circuits. However, the use of multipath was necessary, was deemed the simplest and most cost effective alternative, and became thoroughly entrenched. By the end of the 2000s decade nearly all major IP/MPLS core service provider networks and a few content provider networks had IP/MPLS links which exceeded 100 Gb/s, long before 40GbE was available and 40 Gb/s transport in widespread use.

It is less clear when IP/MPLS LSP exceeded 10 Gb/s, 40 Gb/s, and 100 Gb/s. By 2010, many service providers have LSP in excess of 100 Gb/s, but few are willing to disclose how many LSP have reached this capacity.

By 2012 40GbE and 100GbE LSR products had become available, but were mostly still being evaluated or in trial use by service providers and content providers. The cost of components required to deliver 100GbE products remained high making these products less cost effective. This is expected to change within years.

The important point is that IP/MPLS core network links have long ago exceeded 100 Gb/s and some may have already exceeded a Tb/s and a

small number of IP/MPLS LSP exceed 100 Gb/s. By the time 100 Gb/s circuits are widely deployed, many IP/MPLS core network links are likely to exceed 1 Tb/s and many IP/MPLS LSP capacities are likely to exceed 100 Gb/s. The growth in service provider traffic has consistently outpaced growth in DWDM channel capacities and the growth in capacity of single interfaces and is expected to continue to do so. Therefore multipath techniques are likely here to stay.

Authors' Addresses

So Ning
Tata Communications

Email: ning.so@tatacommunications.com

Andrew Malis
Consultant

Email: agmalis@gmail.com

Dave McDysan
Verizon
22001 Loudoun County PKWY
Ashburn, VA 20147
USA

Email: dave.mcdysan@verizon.com

Lucy Yong
Huawei USA
5340 Legacy Dr.
Plano, TX 75025
USA

Phone: +1 469-277-5837
Email: lucy.yong@huawei.com

Curtis Villamizar
Outer Cape Cod Network Consulting

Email: curtis@occnc.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: August 3, 2015

S. Bryant
C. Filsfils
S. Previdi
Cisco Systems
M. Shand
Independent Contributor
N. So
Vinci Systems
January 30, 2015

Remote Loop-Free Alternate (LFA) Fast Re-Route (FRR)
draft-ietf-rtgwg-remote-lfa-11

Abstract

This document describes an extension to the basic IP fast re-route mechanism described in RFC5286, that provides additional backup connectivity for point to point link failures when none can be provided by the basic mechanisms.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC2119 [RFC2119].

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 <http://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 3, 2015.

Copyright Notice

Copyright (c) 2015 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	3
2. Terminology	3
3. Overview of Solution	4
4. Repair Paths	6
4.1. Tunnels as Repair Paths	6
4.2. Tunnel Requirements	7
5. Construction of Repair Paths	8
5.1. Identifying Required Tunneled Repair Paths	8
5.2. Determining Tunnel End Points	8
5.2.1. Computing Repair Paths	9
5.2.2. Selecting Repair Paths	11
5.3. A Cost Based RLFA Algorithm	12
5.4. Interactions with IS-IS Overload, RFC6987, and Costed Out Links	17
6. Example Application of Remote LFAs	18
7. Node Failures	18
8. Operation in an LDP environment	20
9. Analysis of Real World Topologies	21
9.1. Topology Details	21
9.2. LFA only	22
9.3. RLFA	23
9.4. Comparison of LFA an RLFA results	24
10. Management and Operational Considerations	25
11. Historical Note	26
12. IANA Considerations	26
13. Security Considerations	26
14. Acknowledgments	27
15. References	27
15.1. Normative References	27
15.2. Informative References	27
Authors' Addresses	29

1. Introduction

RFC 5714 [RFC5714] describes a framework for IP Fast Re-route (IPFRR) and provides a summary of various proposed IPFRR solutions. A basic mechanism using loop-free alternates (LFAs) is described in [RFC5286] that provides good repair coverage in many topologies [RFC6571], especially those that are highly meshed. However, some topologies, notably ring based topologies are not well protected by LFAs alone because there is no neighbor of the point of local repair (PLR) that has a cost to the destination without traversing the failure that is cheaper than the cost to the destination via the failure.

The method described in this document extends LFA approach described in [RFC5286] to cover many of these cases by tunneling the packets that require IPFRR to a node that is both reachable from the PLR and can reach the destination.

2. Terminology

This document uses the terms defined in [RFC5714]. This section defines additional terms that are used in this document.

Repair tunnel A tunnel established for the purpose of providing a virtual neighbor which is a Loop Free Alternate.

P-space The P-space of a router with respect to a protected link is the set of routers reachable from that specific router using the pre-convergence shortest paths, without any of those paths (including equal cost path splits) transiting that protected link.

For example, the P-space of S with respect to link S-E, is the set of routers that S can reach without using the protected link S-E.

Extended P-space

Consider the set of neighbours of a router protecting a link. Exclude from that set of routers the router reachable over the protected link. The extended P-space of the protecting router with respect to the protected link is the union of the P-spaces of the neighbours in that set of neighbours with respect to the protected link (see Section 5.2.1.2).

Q-space Q-space of a router with respect to a protected link is the set of routers from which that specific router

can be reached without any path (including equal cost path splits) transiting that protected link.

PQ node A PQ node of a node S with respect to a protected link S-E is a node which is a member of both the P-space (or the extended P-space) of S with respect to that protected link S-E and the Q-space of E with respect to that protected link S-E. A repair tunnel endpoint is chosen from the set of PQ-nodes.

Remote LFA (RLFA) The use of a PQ node rather than a neighbour of the repairing node as the next hop in an LFA repair [RFC5286].

In this document the notation X-Y is used to mean the path from X to Y over the link directly connecting X and Y, whilst the notation X->Y refers to the shortest path from X to Y via some set of unspecified nodes including the null set (i.e. Including over a link directly connecting X and Y).

3. Overview of Solution

The problem of LFA IPFRR reachability in some networks is illustrated by the network fragment shown in Figure 1 below.

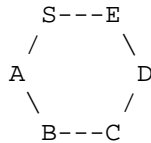


Figure 1: A simple ring topology

If all link costs are equal, traffic transiting link S-E cannot be fully protected by LFAs. The destination C is an ECMP from S, and so traffic to C can be protected when S-E fails, but traffic to D and E are not protectable using LFAs.

This document describes extensions to the basic repair mechanism in which tunnels are used to provide additional logical links which can then be used as loop free alternates where none exist in the original topology. In Figure 1 S can reach A, B, and C without going via S-E; these form S's extended P-space with respect to S-E. The routers that can reach E without going through S-E will be in E's Q-space with respect to link S-E; these are D and C. B has equal-cost paths to E via B-A-S-E and B-C-D-E and so the forwarder at S might choose to send a packet to E via link S-E. Hence B is not in the Q-space of

E with respect to link S-E. The single node in both S's extended P-space and E's Q-space is C; thus node C is selected as the repair tunnel's end-point. Thus, if a tunnel is provided between S and C as shown in Figure 2 then C, now being a direct neighbor of S would become an LFA for D and E. The definition of (extended-)P space and Q space are provided in Section 2 and details of the calculation of the tunnel end points is provided in Section 5.2.

The non-failure traffic distribution is not disrupted by the provision of such a tunnel since it is only used for repair traffic and MUST NOT be used for normal traffic. Note that Operations and Maintenance (OAM) traffic specifically to verify the viability of the repair MAY traverse the tunnel prior to a failure.

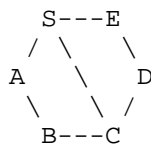


Figure 2: The addition of a tunnel

The use of this technique is not restricted to ring based topologies, but is a general mechanism which can be used to enhance the protection provided by LFAs. A study of the protection achieved using remote LFA in typical service provider core networks is provided in Section 9, and a side by side comparison between LFA and remote LFA is provided in Section 9.4.

Remote LFA is suitable for incremental deployment within a network, including a network that is already deploying LFA. Computation of the repair path requires acceptable CPU resources, and takes place exclusively on the repairing node. In MPLS networks the targeted LDP protocol needed to learn the label binding at the repair tunnel endpoint Section 8 is a well understood and widely deployed technology.

The technique described in this document is directed at providing repairs in the case of link failures. Considerations regarding node failures are discussed in Section 7. This memo describes a solution to the case where the failure occurs on a point to point link. It covers the case where the repair first hop is reached via a broadcast or non-broadcast multi-access (NBMA) link such as a LAN, and the case where the P or Q node is attached via such a link. It does not however cover the more complicated case where the failed interface is a broadcast or non-broadcast multi-access (NBMA) link.

This document considers the case when the repair path is confined to either a single area or to the level two routing domain. In all other cases, the chosen PQ node should be regarded as a tunnel adjacency of the repairing node, and the considerations described in Section 6 of [RFC5286] taken into account.

4. Repair Paths

As with LFA FRR, when a router detects an adjacent link failure, it uses one or more repair paths in place of the failed link. Repair paths are pre-computed in anticipation of later failures so they can be promptly activated when a failure is detected.

A tunneled repair path tunnels traffic to some staging point in the network from which it is known that, in the absence of a worse than anticipated failure, the traffic will travel to its destination using normal forwarding without looping back. This is equivalent to providing a virtual loop-free alternate to supplement the physical loop-free alternates. Hence the name "Remote LFA FRR". In its simplest form, when a link cannot be entirely protected with local LFA neighbors, the protecting router seeks the help of a remote LFA staging point. Network manageability considerations may lead to a repair strategy that uses a remote LFA more frequently [I-D.ietf-rtgwg-lfa-manageability].

Examples of worse failures are node failures (see Section 7), the failure of a shared risk link group (SRLG), the independent concurrent failures of multiple links, broadcast or non-broadcast multi-access (NBMA) links Section 3; protecting against such worse failures is out of scope for this specification.

4.1. Tunnels as Repair Paths

Consider an arbitrary protected link S-E. In LFA FRR, if a path to the destination from a neighbor N of S does not cause a packet to loop back over the link S-E (i.e. N is a loop-free alternate), then S can send the packet to N and the packet will be delivered to the destination using the pre-failure forwarding information. If there is no such LFA neighbor, then S may be able to create a virtual LFA by using a tunnel to carry the packet to a point in the network which is not a direct neighbor of S from which the packet will be delivered to the destination without looping back to S. In this document such a tunnel is termed a repair tunnel. The tail-end of this tunnel (the repair tunnel endpoint) is a "PQ node" and the repair mechanism is a "remote LFA". This tunnel MUST NOT traverse the link S-E.

Note that the repair tunnel terminates at some intermediate router between S and E, and not E itself. This is clearly the case, since

if it were possible to construct a tunnel from S to E then a conventional LFA would have been sufficient to effect the repair.

4.2. Tunnel Requirements

There are a number of IP in IP tunnel mechanisms that may be used to fulfil the requirements of this design, such as IP-in-IP [RFC1853] and GRE[RFC1701] .

In an MPLS enabled network using LDP[RFC5036], a simple label stack[RFC3032] may be used to provide the required repair tunnel. In this case the outer label is S's neighbor's label for the repair tunnel end point, and the inner label is the repair tunnel end point's label for the packet destination. In order for S to obtain the correct inner label it is necessary to establish a targeted LDP session[RFC5036] to the tunnel end point.

The selection of the specific tunnelling mechanism (and any necessary enhancements) used to provide a repair path is outside the scope of this document. The deployment in an MPLS/LDP environment is relatively simple in the data plane as an LDP LSP from S to the repair tunnel endpoint (the selected PQ node) is readily available, and hence does not require any new protocol extension or design change. This LSP is automatically established as a basic property of LDP behavior. The performance of the encapsulation and decapsulation is efficient as encapsulation is just a push of one label (like conventional MPLS TE FRR) and the decapsulation is normally configured to occur at the penultimate hop before the repair tunnel endpoint. In the control plane, a targeted LDP (TLDP) session is needed between the repairing node and the repair tunnel endpoint, which will need to be established and the labels processed before the tunnel can be used. The time to establish the TLDP session and acquire labels will limit the speed at which a new tunnel can be put into service. This is not anticipated to be a problem in normal operation since the managed introduction and removal of links is relatively rare as is the incidence of failure in a well managed network.

When a failure is detected, it is necessary to immediately redirect traffic to the repair path. Consequently, the repair tunnel used MUST be provisioned beforehand in anticipation of the failure. Since the location of the repair tunnels is dynamically determined it is necessary to automatically establish the repair tunnels. Multiple repair tunnels may share a tunnel end point.

5. Construction of Repair Paths

5.1. Identifying Required Tunneled Repair Paths

Not all links will require protection using a tunneled repair path. Referring to Figure 1, if E can already be protected via an LFA, S-E does not need to be protected using a repair tunnel, since all destinations normally reachable through E must therefore also be protectable by an LFA. Such an LFA is frequently termed a "link LFA". Tunneled repair paths (which may be calculated per-prefix) are only required for links which do not have a link or per-prefix LFA.

It should be noted that using the Q-space of E as a proxy for the Q-space of each destination can result in failing to identify valid remote LFAs. The extent to which this reduces the effective protection coverage is topology dependent.

5.2. Determining Tunnel End Points

The repair tunnel endpoint needs to be a node in the network reachable from S without traversing S-E. In addition, the repair tunnel end point needs to be a node from which packets will normally flow towards their destination without being attracted back to the failed link S-E.

Note that once released from the tunnel, the packet will be forwarded, as normal, on the shortest path from the release point to its destination. This may result in the packet traversing the router E at the far end of the protected link S-E, but this is obviously not required.

The properties that are required of repair tunnel end points are therefore:

- o The repair tunneled point MUST be reachable from the tunnel source without traversing the failed link; and
- o When released from the tunnel, packets MUST proceed towards their destination without being attracted back over the failed link.

Provided both these requirements are met, packets forwarded over the repair tunnel will reach their destination, and will not loop after a single link failure.

In some topologies it will not be possible to find a repair tunnel endpoint that exhibits both the required properties. For example if the ring topology illustrated in Figure 1 had a cost of 4 for the link B-C, while the remaining links were cost 1, then it would not be

possible to establish a tunnel from S to C (without resorting to some form of source routing).

5.2.1. Computing Repair Paths

To compute the repair path for link S-E it is necessary to determine the set of routers which can be reached from S without traversing S-E, and match this with the set of routers from which the node E can be reached, by normal forwarding, without traversing the link S-E.

The approach used in this memo is as follows:

- o The method of computing the set of routers which can be reached from S on the shortest path tree without traversing S-E is described. This is called the S's P-space with respect to the failure of link S-E.
- o The distance of the tunnel endpoint from the point of local repair (PLR) is increased by noting that S is able to use the P-Space of its neighbours with respect to the failure of link S-E, since S can determine which neighbour it will use as the next hop for the repair. This is called the S's Extended P-space with respect to the failure of link S-E. The use of extended P-space allows greater repair coverage and is the preferred approach.
- o Finally two methods of computing the set of routers from which the node E can be reached, by normal forwarding, without traversing the link S-E. This is called the Q-space of E with respect to the link S-E.

The selection of the preferred node from the set of nodes that are in both Extended P-Space and Q-Space with respect to the S-E is described in Section 5.2.2.

A suitable cost based algorithm to compute the set of nodes common to both extended P-space and Q-space with respect to the S-E is provided in Section 5.3.

5.2.1.1. P-space

The set of routers which can be reached from S on the shortest path tree without traversing S-E is termed the P-space of S with respect to the link S-E. This P-space can be obtained by computing a shortest path tree (SPT) rooted at S and excising the sub-tree reached via the link S-E (including those routers which are members of an ECMP that includes link S-E). The exclusion of routers reachable via an ECMP that includes S-E prevents the forwarding subsystem from attempting to execute a repair via the failed link

S-E. Thus for example, if the SPF computation stores at each node the next-hops to be used to reach that node from S, then the node can be added to P-space if none of its next-hops are link S-E. In the case of Figure 1 this P-space comprises nodes A and B only. Expressed in cost terms the set of routers {P} are those for which the shortest path cost S->P is strictly less than the shortest path cost S->E->P.

5.2.1.2. Extended P-space

The description in Section 5.2.1.1 calculated router S's P-space rooted at S itself. However, since router S will only use a repair path when it has detected the failure of the link S-E, the initial hop of the repair path need not be subject to S's normal forwarding decision process. Thus the concept of extended P-space is introduced. Router S's extended P-space is the union of the P-spaces of each of S's neighbours (N). This may be calculated by computing a shortest path tree (SPT) at each of S's neighbors (excluding E) and excising the subtree reached via the path N->S->E. Note this will excise those routers which are reachable through all ECMPs that includes link S-E. The use of extended P-space may allow router S to reach potential repair tunnel end points that were otherwise unreachable. In cost terms a router (P) is in extended P-space if the shortest path cost N->P is strictly less than the shortest path cost N->S->E->P. In other words, once the packet is forced to N by S, it is a lower cost for it to continue on to P by any path except one that takes it back to S and then across the S->E link.

Since in the case of Figure 1 node A is a per-prefix LFA for the destination node C, the set of extended P-space nodes with respect to link S-E comprises nodes A, B and C. Since node C is also in E's Q-space with respect to link S-E, there is now a node common to both extended P-space and Q-space which can be used as a repair tunnel end-point to protect the link S-E.

5.2.1.3. Q-space

The set of routers from which the node E can be reached, by normal forwarding, without traversing the link S-E is termed the Q-space of E with respect to the link S-E. The Q-space can be obtained by computing a reverse shortest path tree (rSPT) rooted at E, with the sub-tree which might traverse the protected link S-E excised (i.e. those nodes that would send the packet via S-E plus those nodes which have an ECMP set to E with one or more members of that ECMP set traversing the protected link S-E). The rSPT uses the cost towards the root rather than from it and yields the best paths towards the root from other nodes in the network. In the case of Figure 1 the Q-space of E with respect to S-E comprises nodes C and D only.

Expressed in cost terms the set of routers $\{Q\}$ are those for which the shortest path cost $Q \leftarrow E$ is strictly less than the shortest path cost $Q \leftarrow S \leftarrow E$. In Figure 1 the intersection of the E's Q-space with respect to S-E with S's P-space with respect to S-E defines the set of viable repair tunnel end-points, known as "PQ nodes". As can be seen, for the case of Figure 1 there is no common node and hence no viable repair tunnel end-point. However when the extended the extended P-space Section 5.2.1.2 at S with respect to S-E is considered, a suitable intersection is found at C.

Note that the Q-space calculation could be conducted for each individual destination and a per-destination repair tunnel end point determined. However this would, in the worst case, require an SPF computation per destination which is not currently considered to be scalable. Therefore the Q-space of E with respect to link S-E is used as a proxy for the Q-space of each destination. This approximation is obviously correct since the repair is only used for the set of destinations which were, prior to the failure, routed through node E. This is analogous to the use of link-LFAs rather than per-prefix LFAs.

5.2.2. Selecting Repair Paths

The mechanisms described above will identify all the possible repair tunnel end points that can be used to protect a particular link. In a well-connected network there are likely to be multiple possible release points for each protected link. All will deliver the packets correctly so, arguably, it does not matter which is chosen. However, one repair tunnel end point may be preferred over the others on the basis of path cost or some other selection criteria.

There is no technical requirement for the selection criteria to be consistent across all routers, but such consistency may be desirable from an operational point of view. In general there are advantages in choosing the repair tunnel end point closest (shortest metric) to S. Choosing the closest maximises the opportunity for the traffic to be load balanced once it has been released from the tunnel. For consistency in behavior, it is RECOMMENDED that the member of the set of routers $\{PQ\}$ with the lowest cost $S \rightarrow P$ be the default choice for P. In the event of a tie the router with the lowest node identifier SHOULD be selected.

It is a local matter whether the repair path selection policy used by the router favours LFA repairs over RLFA repairs. An LFA repair has the advantage of not requiring the use of tunnel, however network manageability considerations may lead to a repair strategy that uses a remote LFA more frequently [I-D.ietf-rtgwg-lfa-manageability].

As described in [RFC5286], always selecting a PQ node that is downstream to the destination with respect to the repairing node, prevents the formation of loops when the failure is worse than expected. The use of downstream nodes reduces the repair coverage, and operators are advised to determine whether adequate coverage is achieved before enabling this selection feature.

5.3. A Cost Based RLFA Algorithm

The preceding text has described the computation of the remote LFA repair target (PQ) in terms of the intersection of two reachability graphs computed using a shortest path first (SPF) algorithm. This section describes a method of computing the remote LFA repair target for a specific failed link using a cost based algorithm. The pseudo-code provided in this section avoids unnecessary SPF computations, but for the sake of readability, it does not otherwise try to optimize the code. The algorithm covers the case where the repair first hop is reached via a broadcast or non-broadcast multi-access (NBMA) link such as a LAN. It also covers the case where the P or Q node is attached via such a link. It does not cover the case where the failed interface is a broadcast or non-broadcast multi-access (NBMA) link. To address that case it is necessary to compute the Q space of each neighbor of the repairing router reachable through the LAN, i.e. to treat the pseudonode [RFC1195] as a node failure. This is because the Q spaces of the neighbors of the pseudonode may be disjoint requiring use of a neighbor specific PQ node. The reader is referred to [I-D.ietf-rtgwg-rlfa-node-protection] for further information on the use of RLFA for node repairs.

The following notation is used:

- o $D_{opt}(a,b)$ is the shortest distance from node a to node b as computed by the SPF.
- o `dest` is the packet destination
- o `fail_intf` is the failed interface (S-E in the example)
- o `fail_intf.remote_node` is the node reachable over interface `fail_intf` (node E in the example)
- o `intf.remote_node` is the set of nodes reachable over interface `intf`
- o `root` is the root of the SPF calculation
- o `self` is the node carrying out the computation
- o `y` is the node in the network under consideration

- o `y.pseudonode` is true if `y` is a pseudonode

```

////////////////////////////////////
//
//   Main Function

////////////////////////////////////
//
// We have already computed the forward SPF from self to all nodes
// y in network and thus we know D_opt (self, y). This is needed
// for normal forwarding.
// However for completeness.

Compute_and_Store_Forward_SPF(self)

// To extend P-space we compute the SPF at each neighbour except
// the neighbour that is reached via the link being protected.
// We will also need D_opt(fail_intf.remote_node,y) so compute
// that at the same time.

Compute_Neighbor_SPFs()

// Compute the set of nodes {P} reachable other than via the
// failed link

Compute_Extended_P_Space(fail_intf)

// Compute the set of nodes that can reach the node on the far
// side of the failed link without traversing the failed link.

Compute_Q_Space(fail_intf)

// Compute the set of candidate RLFA tunnel endpoints

Intersect_Extended_P_and_Q_Space()

// Make sure that we cannot get looping repairs when the
// failure is worse than expected.

if (guarantee_no_looping_on_worse_than_protected_failure)
    Apply_Downstream_Constraint()

//
//   End of Main Function
//
////////////////////////////////////

```

```
////////////////////////////////////
//
//  Procedures
//

////////////////////////////////////
//
//  This computes the SPF from root, and stores the optimum
//  distance from root to each node y

Compute_and_Store_Forward_SPF(root)
    Compute_Forward_SPF(root)
    foreach node y in network
        store D_opt(root,y)

////////////////////////////////////
//
//  This computes the optimum distance from each neighbour (other
//  than the neighbour reachable through the failed link) and
//  every other node in the network
//
//  Note that we compute this for all neighbours including the
//  neighbour on the far side the failure. This is done on the
//  expectation that more than one link will be protected, and
//  that the results are stored for later use.
//

Compute_Neighbor_SPFs()
    foreach interface intf in self
        Compute_and_Store_Forward_SPF(intf.remote_node)
```

```
////////////////////////////////////
//
// The reverse SPF computes the cost from each remote node to
// root. This is achieved by running the normal SPF algorithm,
// but using the link cost in the direction from the next hop
// back towards root in place of the link cost in the direction
// away from root towards the next hop.

Compute_and_Store_Reverse_SPF(root)
  Compute_Reverse_SPF(root)
  foreach node y in network
    store D_opt(y,root)

////////////////////////////////////
//
// Calculate extended P-space
//
// Note the strictly less than operator is needed to
// avoid ECMP issues.

Compute_Extended_P_Space(fail_intf)
  foreach node y in network
    y.in_extended_P_space = false
    // Extend P-space to the P-spaces of all reachable
    // neighbours
    foreach interface intf in self
      // Exclude failed interface, noting that
      // the node reachable via that interface may be
      // reachable via another interface (parallel path)
      if (intf != fail_intf)
        foreach neighbor n in intf.remote_node
          // Apply RFC5286 Inequality 1
          if ( D_opt(n, y) <
              D_opt(n,self) + D_opt(self, y))
            y.in_extended_P_space = true
```

```
////////////////////////////////////
//
// Compute the nodes in Q-space
//

Compute_Q_Space(fail_intf)
    // Compute the cost from every node the network to the
    // node normally reachable across the failed link
    Compute_and_Store_Reverse_SPF(fail_intf.remote_node)

    // Compute the cost from every node the network to self
    Compute_and_Store_Reverse_SPF(self)

    foreach node y in network
        if ( D_opt(y,fail_intf.remote_node) < D_opt(y,self) +
            D_opt(self,fail_intf.remote_node) )
            y.in_Q_space = true
        else
            y.in_Q_space = false

////////////////////////////////////
//
// Compute set of nodes in both extended P-space and in Q-space

Intersect_Extended_P_and_Q_Space()
    foreach node y in network
        if ( y.in_extended_P_space && y.in_Q_space &&
            y.pseudonode == False)
            y.valid_tunnel_endpoint = true
        else
            y.valid_tunnel_endpoint = false
```

```

////////////////////////////////////
//
// A downstream route is one where the next hop is strictly
// closer to the destination. By sending the packet to a
// PQ node that is downstream, we know that if the PQ node
// detects a failure, it will not loop the packet back to self.
// This is useful when there are two failures, or a node has
// failed rather than a link.

Apply_Downstream_Constraint()
    foreach node y in network
        if (y.valid_tunnel_endpoint)
            Compute_and_Store_Forward_SPF(y)
            if ((D_opt(y,dest) < D_opt(self,dest))
                y.valid_tunnel_endpoint = true
            else
                y.valid_tunnel_endpoint = false

//
////////////////////////////////////

```

5.4. Interactions with IS-IS Overload, RFC6987, and Costed Out Links

Since normal link state routing takes into account the IS-IS overload bit, [RFC6987], and costing out of links as described in Section 3.5 of [RFC5286], the forward SPF's performed by the PLR rooted at the neighbors of the PLR also need to take this into account. A repair tunnel path from a neighbor of the PLR to a repair tunnel endpoint will generally avoid the nodes and links excluded by the IGP overload/costing out rules. However, there are two situations where this behavior may result in a repair path traversing a link or router that should be excluded:

1. When the first hop on the repair tunnel path (from the PLR to a direct neighbor) does not follow the IGP shortest path. In this case, the PLR MUST NOT use a repair tunnel path whose first hop is along a link whose cost or reverse cost is MaxLinkMetric (for OSPF) or the maximum cost (for IS-IS) or, has the overload bit set (for IS-IS).
2. The IS-IS overload bit and the mechanism of [RFC6987] only prevent transit traffic from traversing a node. They do not prevent traffic destined to a node. The per-neighbor forward SPF's using the standard IGP overload rules will not prevent a PLR from choosing a repair tunnel endpoint that is advertising a

desire to not carry transit traffic. Therefore, the PLR MUST NOT use a repair tunnel endpoint with the IS-IS overload bit set, or where all outgoing interfaces have the cost set to MaxLinkMetric for OSPF.

6. Example Application of Remote LFAs

An example of a commonly deployed topology which is not fully protected by LFAs alone is shown in Figure 3. PE1 and PE2 are connected in the same site. P1 and P2 may be geographically separated (inter-site). In order to guarantee the lowest latency path from/to all other remote PEs, normally the shortest path follows the geographical distance of the site locations. Therefore, to ensure this, a lower IGP metric (5) is assigned between PE1 and PE2. A high metric (1000) is set on the P-PE links to prevent the PEs being used for transit traffic. The PEs are not individually dual-homed in order to reduce costs.

This is a common topology in SP networks.

When a failure occurs on the link between PE1 and P1, PE1 does not have an LFA for traffic reachable via P1. Similarly, by symmetry, if the link between PE2 and P2 fails, PE2 does not have an LFA for traffic reachable via P2.

Increasing the metric between PE1 and PE2 to allow the LFA would impact the normal traffic performance by potentially increasing the latency.

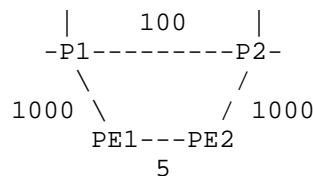


Figure 3: Example SP topology

Clearly, full protection can be provided, using the techniques described in this document, by PE1 choosing P2 as the remote LFA repair target node, and PE2 choosing P1 as the remote LFA repair target.

7. Node Failures

When the failure is a node failure rather than a point-to-point link failure there is a danger that the RLFA repair will loop. This is discussed in detail in [I-D.bryant-ipfrr-tunnels]. In summary the

problem is that two of more of E's neighbors each with E as the next hop to some destination D may attempt to repair a packet addressed to destination D via the other neighbor and then E, thus causing a loop to form. A similar problem exists in the case of a shared risk link group failure where the PLR for each failure attempts to repair via the other failure. As will be noted from [I-D.bryant-ipfrr-tunnels], this can rapidly become a complex problem to address.

There are a number of ways to minimize the probability of a loop forming when a node failure occurs and there exists the possibility that two of E's neighbors may form a mutual repair.

1. Detect when a packet has arrived on some interface I that is also the interface used to reach the first hop on the RLFA path to the remote LFA repair target, and drop the packet. This is useful in the case of a ring topology.
2. Require that the path from the remote LFA repair target to destination D never passes through E (including in the ECMP case), i.e. only use node protecting paths in which the cost from the remote LFA repair target to D is strictly less than the cost from the remote LFA repair target to E plus the cost E to D.
3. Require that where the packet may pass through another neighbor of E, that node is down stream (i.e. strictly closer to D than the repairing node). This means that some neighbor of E (X) can repair via some other neighbor of E (Y), but Y cannot repair via X.

Case 1 accepts that loops may form and suppresses them by dropping packets. Dropping packets may be considered less detrimental than looping packets. This approach may also lead to dropping some legitimate packets. Cases 2 and 3 above prevent the formation of a loop, but at the expense of a reduced repair coverage and at the cost of additional complexity in the algorithm to compute the repair path. Alternatively one might choose to assume that the probability of a node failure is sufficiently rare that the issue of looping RLFA repairs can be ignored.

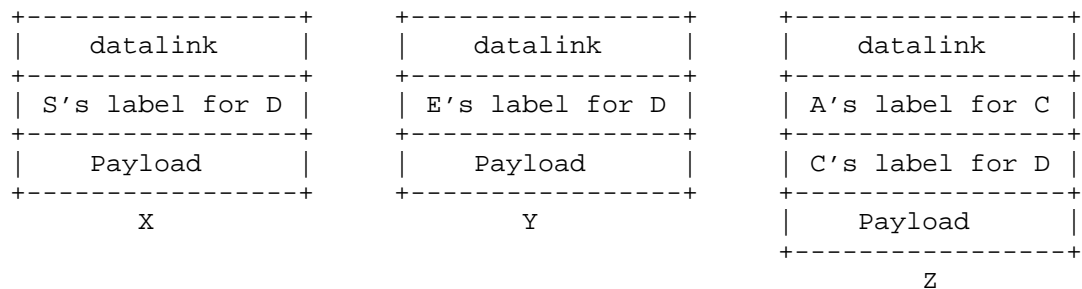
The probability of a node failure and the consequences of node failure in any particular topology will depend on the node design, the particular topology in use, and the strategy adopted under node failure. It is recommended that a network operator perform an analysis of the consequences and probability of node failure in their network, and determine whether the incidence and consequence of occurrence are acceptable.

This topic is further discussed in [I-D.ietf-rtgwg-rlfa-node-protection].

8. Operation in an LDP environment

Where this technique is used in an MPLS network using LDP [RFC5036], and S is a transit node, S will need to swap the top label in the stack for the remote LFA repair target's (PQ's) label to the destination, and to then push its own label for the remote LFA repair target.

In the example Figure 2 S already has the first hop (A) label for the remote LFA repair target (C) as a result of the ordinary operation of LDP. To get the remote LFA repair target's label (C's label) for the destination (D), S needs to establish a targeted LDP session with C. The label stack for normal operation and RLFA operation is shown below in Figure 4.



X = Normal label stack packet arriving at S
 Y = Normal label stack packet leaving S
 Z = RLFA label stack to D via C as the remote LFA repair target.

Figure 4

To establish an targeted LDP session with a candidate remote LFA repair target node the repairing node (S) needs to know what IP address that the remote LFA repair target is willing to use for targeted LDP sessions. Ideally this is provided by the remote LFA repair target advertising this address in the IGP in use. Which address is used, how this is advertised in the IGP, and whether this is a special IP address or an IP address also used for some other purpose is out of scope for this document and must be specified in an IGP specific RFC.

In the absence of a protocol to learn the preferred IP address for targeted LDP, an LSR should attempt a targeted LDP session with the Router ID [RFC2328] [RFC5305] [RFC5340] [RFC6119] [I-D.ietf-ospf-routable-ip-address], unless it is configured otherwise.

No protection is available until the TLDP session has been established and a label for the destination has been learned from the remote LFA repair target. If for any reason the TLDP session cannot be established, an implementation SHOULD advise the operator about the protection setup issue through the network management system.

9. Analysis of Real World Topologies

This section gives the results of analysing a number of real world service provider topologies collected between the end of 2012 and early 2013

9.1. Topology Details

The figure below characterises each topology (topo) studied in terms of :

- o The number of nodes (# nodes) excluding pseudonodes.
- o The number of bidirectional links (# links) including parallel links and links to and from pseudonodes.
- o The number of node pairs that are connected by one or more links (# pairs).
- o The number of node pairs that are connected by more than one (i.e. parallel) link (# para).
- o The number of links (excluding pseudonode links, which are by definition asymmetric) that have asymmetric metrics (#asym).

topo	# nodes	# links	# pairs	# para	# asym
1	315	570	560	10	3
2	158	373	312	33	0
3	655	1768	1314	275	1195
4	1281	2326	2248	70	10
5	364	811	659	80	86
6	114	318	197	101	4
7	55	237	159	67	2
8	779	1848	1441	199	437
9	263	482	413	41	12
10	86	375	145	64	22
11	162	1083	351	201	49
12	380	1174	763	231	0
13	1051	2087	2037	48	64
14	92	291	204	64	2

9.2. LFA only

The figure below shows the percentage of protected destinations (% prot) and percentage of guaranteed node protected destinations (% gtd N) for the set of topologies characterized in Section 9.1 achieved using only LFA repairs.

These statistics were generated by considering each node and then considering each link to each next hop to each destination. The percentage of such links across the entire network that are protected against link failure was determined. This is the percentage of protected destinations. If a link is protected against the failure of the next hop node, this is considered guaranteed node protecting (GNP) and percentage of guaranteed node protected destinations is calculated using the same method used for calculating the link protection coverage.

GNP is identical to Node-protecting as defined in [RFC5714] and does not include the additional node protection coverage obtained by the de facto node-protecting condition described in [RFC6571].

topo	% prot	% gtd N
1	78.5	36.9
2	97.3	52.4
3	99.3	58
4	83.1	63.1
5	99	59.1
6	86.4	21.4
7	93.9	35.4
8	95.3	48.1
9	82.2	49.5
10	98.5	14.9
11	99.6	24.8
12	99.5	62.4
13	92.4	51.6
14	99.3	48.6

9.3. RLFA

The figure below shows the percentage of protected destinations (% prot) and % guaranteed node protected destinations (% gtd N) for RLFA protection in the topologies studies. In addition, it show the percentage of destinations using an RLFA repair (% PQ) together with the total number of unidirectional RLFA targeted LDP session established (# PQ), the number of PQ sessions which would be required for complete protection, but which could not be established because there was no PQ node, i.e. the number of cases whether neither LFA or RLFA protection was possible (no PQ). It also shows the 50 (p50), 90 (p90) and 100 (p100) percentiles for the number of individual LDP sessions terminating at an individual node (whether used for TX, RX or both).

For example, if there were LDP sessions required A->B, A->C, C->A, C->D, these would be counted as 2, 1, 2, 1 at nodes A,B,C and D respectively because:-

A has two sessions (to nodes B and C)

B has one session (to node A)

C has two sessions (to nodes A and D)

D has one session (to node D)

In this study, remote LFA is only used when necessary. i.e. when there is at least one destination which is not reparable by a per

destination LFA, and a single remote LFA tunnel is used (if available) to repair traffic to all such destinations. The remote LFA repair target points are computed using extended P space and choosing the PQ node which has the lowest metric cost from the repairing node.

topo	% prot	% gtd N	% PQ	# PQ	no PQ	p50	p90	p100
1	99.7	53.3	21.2	295	3	1	5	14
2	97.5	52.4	0.2	7	40	0	0	2
3	99.999	58.4	0.7	63	5	0	1	5
4	99	74.8	16	1424	54	1	3	23
5	99.5	59.5	0.5	151	7	0	2	7
6	100	34.9	13.6	63	0	1	2	6
7	99.999	40.6	6.1	16	2	0	2	4
8	99.5	50.2	4.3	350	39	0	2	15
9	99.5	55	17.3	428	5	1	2	67
10	99.6	14.1	1	49	7	1	2	5
11	99.9	24.9	0.3	85	1	0	2	8
12	99.999	62.8	0.5	512	4	0	0	3
13	97.5	54.6	5.1	1188	95	0	2	27
14	100	48.6	0.7	79	0	0	2	4

Another study[ISOCORE2010] confirms the significant coverage increase provided by Remote LFAs.

9.4. Comparison of LFA and RLFA results

The table below provides a side by side comparison the LFA and the remote LFA results. This shows a significant improvement in the percentage of protected destinations and normally a modest improvement in the percentage of guaranteed node protected destinations.

topo	LFA % prot	RLFA %prot	LFA % gtd N	RLFA % gtd N
1	78.5	99.7	36.9	53.3
2	97.3	97.5	52.4	52.4
3	99.3	99.999	58	58.4
4	83.1	99	63.1	74.8
5	99	99.5	59.1	59.5
6	86.4	100	21.4	34.9
7	93.9	99.999	35.4	40.6
8	95.3	99.5	48.1	50.2
9	82.2	99.5	49.5	55
10	98.5	99.6	14.9	14.1
11	99.6	99.9	24.8	24.9
12	99.5	99.999	62.4	62.8
13	92.4	97.5	51.6	54.6
14	99.3	100	48.6	48.6

As shown in the table, remote LFA provides close to 100% prefix protection against link failure in 11 of the 14 topologies studied, and provides a significant improvement in two of the remaining three cases. Note that in an MPLS network the tunnels to the PQ nodes are always present as a property of an LDP-based deployment.

In the small number of cases where there is no intersection between the (extended)P-space and the Q-space, a number of solutions to providing a suitable path between such disjoint regions in the network have been discussed in the working group. For example an explicitly routed LSP between P and Q might be set up using RSVP-TE or using Segment Routing [I-D.filsfils-spring-segment-routing]. Such extended repair methods are outside the scope of this document.

10. Management and Operational Considerations

The management of LFA and remote LFA is the subject of ongoing work within the IETF [I-D.ietf-rtgwg-lfa-manageability] to which the reader is referred. Management considerations may lead to a preference for the use of a remote LFA over an available LFA. This preference is a matter for the network operator, and not a matter of protocol correctness.

When the network re-converges, microloops [RFC5715] can form due to transient inconsistencies in the forwarding tables of different routers. If it is determined that microloops are a significant issue in the deployment, then a suitable loop free convergence methods such

as one of those described in [RFC5715], [RFC6976], or [I-D.litkowski-rtgwg-uloop-delay] should be implemented.

11. Historical Note

The basic concepts behind Remote LFA were invented in 2002 and were later included in [I-D.bryant-ipfrr-tunnels], submitted in 2004.

[I-D.bryant-ipfrr-tunnels], targeted a 100% protection coverage and hence included additional mechanisms on top of the Remote LFA concept. The addition of these mechanisms made the proposal very complex and computationally intensive and it was therefore not pursued as a working group item.

As explained in [RFC6571], the purpose of the LFA FRR technology is not to provide coverage at any cost. A solution for this already exists with MPLS TE FRR. MPLS TE FRR is a mature technology which is able to provide protection in any topology thanks to the explicit routing capability of MPLS TE.

The purpose of LFA FRR technology is to provide for a simple FRR solution when such a solution is possible. The first step along this simplicity approach was "local" LFA [RFC5286]. This specification of "Remote LFA" is a natural second step.

12. IANA Considerations

There are no IANA considerations that arise from this architectural description of IPFRR. The RFC Editor may remove this section on publication.

13. Security Considerations

The security considerations of [RFC5286] also apply.

Targeted LDP sessions and MPLS tunnels are normal features of an MPLS network and their use in this application raises no additional security concerns.

IP repair tunnel endpoints (where used) SHOULD be assigned from a set of addresses that are not reachable from outside the routing domain. This would prevent their use as an attack vector.

Other than OAM traffic, used to verify the correct operation of a repair tunnel, only traffic that is being protected as a result of a link failure is placed a repair tunnel. The repair tunnel MUST NOT be advertised by the routing protocol as a link that may be used to carry normal user traffic, or routing protocol traffic.

14. Acknowledgments

The authors wish to thank Levente Csikor and Chris Bowers for their contribution to the cost based algorithm text. The authors thank Alia Atlas, Ross Callon, Stephane Litkowski, Bharath R, Pushpasis Sarkar and Adrian Farrel for their review of this document.

15. References

15.1. Normative References

- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC5286] Atlas, A. and A. Zinin, "Basic Specification for IP Fast Reroute: Loop-Free Alternates", RFC 5286, September 2008.
- [RFC5714] Shand, M. and S. Bryant, "IP Fast Reroute Framework", RFC 5714, January 2010.

15.2. Informative References

- [I-D.bryant-ipfrr-tunnels]
Bryant, S., Filsfils, C., Previdi, S., and M. Shand, "IP Fast Reroute using tunnels", draft-bryant-ipfrr-tunnels-03 (work in progress), November 2007.
- [I-D.filsfils-spring-segment-routing]
Filsfils, C., Previdi, S., Bashandy, A., Decraene, B., Litkowski, S., Horneffer, M., Milojevic, I., Shakir, R., Ytti, S., Henderickx, W., Tantsura, J., and E. Crabbe, "Segment Routing Architecture", draft-filsfils-spring-segment-routing-04 (work in progress), July 2014.
- [I-D.ietf-ospf-routable-ip-address]
Xu, X., Chunduri, U., and M. Bhatia, "Carrying Routable IP Addresses in OSPF RI LSA", draft-ietf-ospf-routable-ip-address-01 (work in progress), October 2014.
- [I-D.ietf-rtgwg-lfa-manageability]
Litkowski, S., Decraene, B., Filsfils, C., Raza, K., Horneffer, M., and P. Sarkar, "Operational management of Loop Free Alternates", draft-ietf-rtgwg-lfa-manageability-07 (work in progress), January 2015.

- [I-D.ietf-rtgwg-rlfa-node-protection]
Sarkar, P., Gredler, H., Hegde, S., Bowers, C., Litkowski, S., and H. Raghuvver, "Remote-LFA Node Protection and Manageability", draft-ietf-rtgwg-rlfa-node-protection-01 (work in progress), December 2014.
- [I-D.litkowski-rtgwg-uloop-delay]
Litkowski, S., Decraene, B., Filsfils, C., and P. Francois, "Microloop prevention by introducing a local convergence delay", draft-litkowski-rtgwg-uloop-delay-03 (work in progress), February 2014.
- [ISOCORE2010]
So, N., Lin, T., and C. Chen, "LFA (Loop Free Alternates) Case Studies in Verizon's LDP Network", 2010.
- [RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.
- [RFC1701] Hanks, S., Li, T., Farinacci, D., and P. Traina, "Generic Routing Encapsulation (GRE)", RFC 1701, October 1994.
- [RFC1853] Simpson, W., "IP in IP Tunneling", RFC 1853, October 1995.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC3032] Rosen, E., Tappan, D., Fedorkow, G., Rekhter, Y., Farinacci, D., Li, T., and A. Conta, "MPLS Label Stack Encoding", RFC 3032, January 2001.
- [RFC5036] Andersson, L., Minei, I., and B. Thomas, "LDP Specification", RFC 5036, October 2007.
- [RFC5305] Li, T. and H. Smit, "IS-IS Extensions for Traffic Engineering", RFC 5305, October 2008.
- [RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, July 2008.
- [RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.
- [RFC6119] Harrison, J., Berger, J., and M. Bartlett, "IPv6 Traffic Engineering in IS-IS", RFC 6119, February 2011.

- [RFC6571] Filsfils, C., Francois, P., Shand, M., Decraene, B., Uttaro, J., Leymann, N., and M. Horneffer, "Loop-Free Alternate (LFA) Applicability in Service Provider (SP) Networks", RFC 6571, June 2012.
- [RFC6976] Shand, M., Bryant, S., Previdi, S., Filsfils, C., Francois, P., and O. Bonaventure, "Framework for Loop-Free Convergence Using the Ordered Forwarding Information Base (oFIB) Approach", RFC 6976, July 2013.
- [RFC6987] Retana, A., Nguyen, L., Zinin, A., White, R., and D. McPherson, "OSPF Stub Router Advertisement", RFC 6987, September 2013.

Authors' Addresses

Stewart Bryant
Cisco Systems
250, Longwater, Green Park,
Reading RG2 6GB, UK
UK

Email: stbryant@cisco.com

Clarence Filsfils
Cisco Systems
De Kleetlaan 6a
1831 Diegem
Belgium

Email: cfilsfil@cisco.com

Stefano Previdi
Cisco Systems

Email: sprevidi@cisco.com

Mike Shand
Independent Contributor

Email: imc.shand@gmail.com

Ning So
Vinci Systems

Email: ning.so@vinci-systems.com

Network Working Group
Internet-Draft
Intended status: Informational
Expires: April 24, 2014

Z. Li
H. Chen
G. Yan
Huawei Technologies
October 21, 2013

An Architecture of Central Controlled Interior Gateway Protocol (IGP)
draft-li-rtgwg-cc-igp-arch-00

Abstract

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

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 <http://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 April 24, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents

(<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Terminology	3
3. Architecture	3
3.1. Reference Model	3
3.2. Deployment Mode	4
3.3. Requirement of IGP Extensions	4
3.3.1. Building Connectivity	4
3.3.2. Roles Auto-Discovery	5
3.3.3. Choosing Controller	5
3.3.4. High Availability	5
3.3.5. Security	6
4. Usecases	6
4.1. Network Topology Acquirement	6
4.2. Automated Dividing Multiple Domains	6
4.3. Centralized MPLS TE	7
4.4. MPLS Global Label Allocation	8
4.5. Virtual Link	8
5. IANA Considerations	9
6. Security Considerations	9
7. References	10
7.1. Normative References	10
7.2. Informative References	10
Authors' Addresses	11

1. Introduction

Interior Gateway Protocol (IGP) is a protocol for exchanging routing information between gateways (hosts with routers) within an autonomous network (for example, a system of corporate local area networks). The routing information can then be used by the Internet Protocol (IP) or other network protocols to specify how to route transmissions.

The internet is the most popular network, it is a distributed system. Depending on its configuration, each network device communicates with its neighbor, generates the FIB, and forwards the packet hop by hop. As the rise of SDN, central controlled IGP is becoming more important and new requirements for IGP are proposed as follows:

1. Build the central control architecture between the controller and the client. It includes building connectivity, collecting the topology, and dividing multiple areas automatically, etc.
2. Many new applications are emerging under the central controlled framework, such as network virtualization, centralized MPLS TE calculation, segment routing, etc. These new applications bring extension requirement to IGP.

This document defines an IGP-Based Central Control architecture and then use cases and corresponding IGP extensions under this architecture are described.

2. Terminology

BGP: Border Gateway Protocol

IGP: Interior Gateway Protocol

IS-IS: Intermediate System-Intermediate System

OSPF: Open Shortest Path First

SDN: Software Defined Network

3. Architecture

3.1. Reference Model

The following figure depicts a typical architecture of central controlled IGP. It consists of two essential network elements: IGP Controller and IGP Client. IGP Controller controls all the IGP Clients within its administrative domain by communicating with them. And the controller will also exchange the information each other through some protocol extensions which is out of scope of this document.

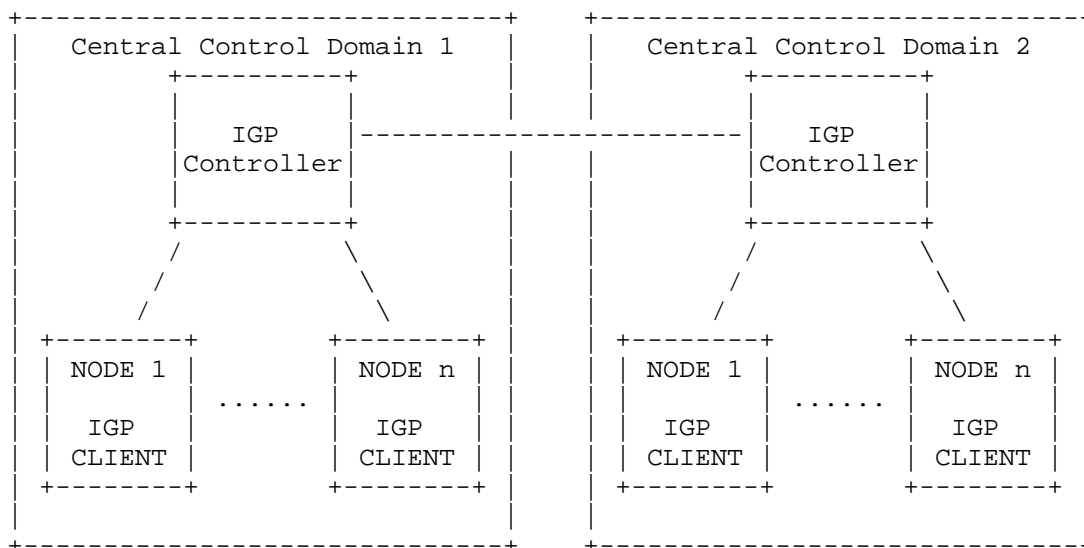


Figure 1: An Architecture of Central Controlled IGP

3.2. Deployment Mode

IGP Controller can run on a general-purpose server or a network device. If IGP Controller runs on a network device, it supports both central-controlled functionality and forwarding functionality. In this scenario, besides the control central point, the IGP controller can also work as a forwarding central point to receive traffic from one node and forward to other nodes. The forwarding model in this scenario is just like hub-spoke forwarding model. If IGP Controller runs on a server, it will not involve in the actual forwarding. It only works as the control central point to control the forwarding behaviors of the nodes. In this scenario the traffic will be distributed in the controlled nodes.

More than one controller can be deployed in a central control domain. These controllers can work on master-slave mode or load-sharing mode.

3.3. Requirement of IGP Extensions

Building a IGP-based Central Controlled Framework needs extensions to IGP, I2RS etc.

3.3.1. Building Connectivity

IGP protocol is very important to establish connective in the central control domain. When a new device connects to the this domain, the connectivity with the other node and the controller should be built at first. The procedures should be automated since the number of devices in this domain can be huge. Base on this initialization process, the controller can download the necessary configuration to this new node to drive it to set up adjacency with its neighbors and the controller. Then the topology information can be synchronized in the central control domain and the connectivity can be built.

3.3.2. Roles Auto-Discovery

In the central control domain, there are two basic roles: IGP controller and IGP client. The controller can centrally configure the client role through I2RS interface. The role information should be flooded through IGP extensions to support the auto discovery functionality.

3.3.3. Choosing Controller

After the roles of the elements are discovered, if there are multiple controllers in the domain, the client can determine which controller to join by its own, or the controllers can determine which controller the clients should join and set the configuration on the nodes through I2RS interface. When determine the controller to be joined, the work mode (master-slave, load-sharing, etc.) of multiple controllers, service type and some other constraints needs to be taken into account.

3.3.4. High Availability

In the IGP-based Central Controlled framework, IGP Controller plays a key role. To avoid one-point-failure of IGP Controller, it is possible to run redundant IGP Controllers for high availability.

Information should be synchronized between the controllers through necessary mechanisms or protocol extensions other than IGP. When the Primary IGP Controller failed, the Backup IGP Controllers will take over the work of the Primary IGP Controller.

To ensure IGP route persistence in case of occurrence of IGP Controller failure, the new Primary IGP Controller SHOULD perform resynchronization with IGP Clients.

When IGP Client loses connection with Primary IGP Controller, it SHOULD following IGP Graceful Restart routine.

3.3.5. Security

In IGP-based Central Controlled framework, it SHOULD be ensured that communications between IGP Controllers and IGP Clients conform to network security policy. The communication key used on IGP Client can be configured through I2RS or other way.

4. Usecases

In IGP-based Central Controlled framework, new use cases which are difficult to be supported in traditional networks are emerging. In some specific use cases, extension and enhancement of IGP protocol are necessary.

4.1. Network Topology Acquirement

In traditional network, it is very difficult for the application to get and use the topology. The application has to depend multiple protocols such as OSPF, ISIS, LLDP, etc. In some scenarios, the application has to communicate with these protocols directly. In the IGP-based central controlled framework, the topology acquirement procedures SHOULD be simplified. All topology related information SHOULD be able to be collected by IGP. Thus the complexity of network operation and management can be reduced. In the IGP-base central controlled framework, the controller can get the whole topology information of the central control domain which can be easily provided for applications through public interface.

4.2. Automated Dividing Multiple Domains

When there are mass devices in the network, not only LSDB synchronization, but also route convergence, will be big pressure for any device, so the network has to be divided into multiple domains. In the IGP-based central controlled the framework, the division can be done automatically by the controller which can calculate reasonable scale for IGP domains based on the whole network information and the possible constraints. IGP adjacency is only set up between nodes in the same IGP domain. The adjacency SHOULD not set up between nodes in different IGP domains. Thus the pressure on the nodes for LSDB synchronization can be reduced and route convergence performance can be improved. The configuration about domain division can be set through I2RS interfaces from the controller to the clients. The architecute for dividing multiple domains with the central controller is shown in the figure 2.

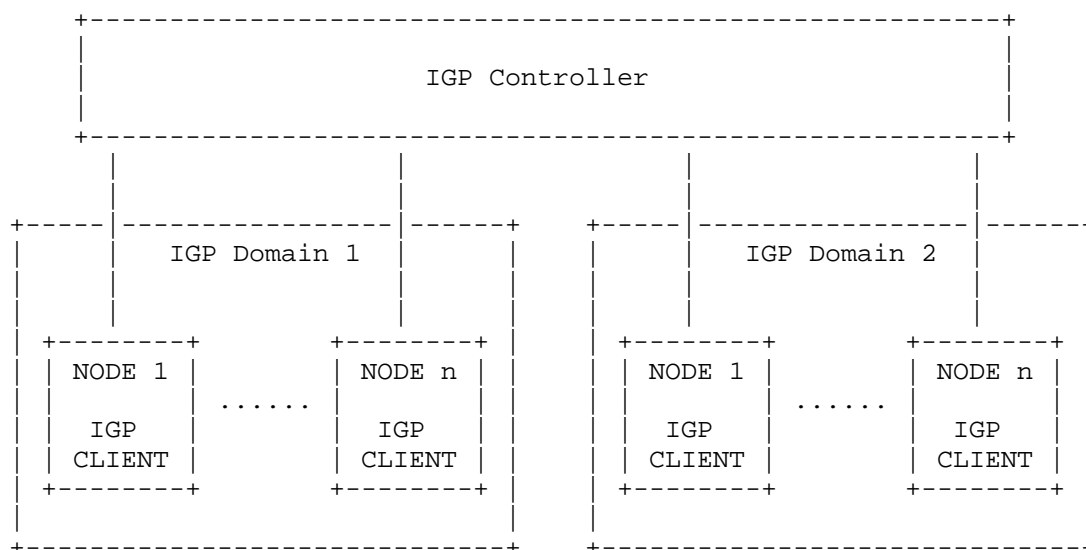


Figure 2: Automatic Division of Multiple IGP Domains

4.3. Centralized MPLS TE

In the IGP-based Central Controlled framework, the controller can implement better traffic engineering functionality because it can calculate more reasonable path based on complete topology information and state information of the whole network. Centralized MPLS TE calculation can avoid the flaw of non-best path proposed by the existing distributed MPLS TE calculation.

In order to support centralized MPLS TE path calculation, IGP SHOULD be able to collect more information from the network. There are two types of information for IGP to collect:

1. Static configuration: In traditional network, MPLS TE attributes should be configured on the link such as maximum reservable bandwidth, color, TE metric, etc. These information will be flooded in the work for MPLS TE path calculation. In the IGP-based Central Controlled framework, these configuration can be set by the controller. This means it is not necessary for the controller to get the TE link information through IGP flooding process. For the reason of compatibility, the IGP flooding process of MPLS TE link information can be kept in the central controlled framework. On the other hand, it provides a possible way for the inconsistency check on the configuration.

2. Running information: Some dynamic state information such as real traffic bandwidth, packet loss rate, delay, power consumption, etc. can be flooded through IGP extensions from the nodes to the controller. The running information can help the controller to calculate more reasonable path and calculate path for more constraints defined by applications.

4.4. MPLS Global Label Allocation

MPLS Global Label should be allocated centrally to guarantee all distributed network nodes can understand meaning of a specific global label in same. The IGP-based Central Controlled framework is particularly suitable to allocate MPLS Global Label through some necessary IGP extensions rather than traditional MPLS protocols(e.g. LDP, RSVP-TE etc.).

MPLS Global Label is defined in [I-D.li-mpls-global-label-framework] and related use cases are defined in [I-D.li-mpls-global-label-usecases].

The MPLS global label should be assigned centrally; each node in network should have same understanding about these labels. In the central control network, the global label will be handled by controller, and IGP protocol will flood these labels.

The extensions of IGP for MPLS global label include:

1. Collect the label capability of each node. The label capability is the global label space.
2. IGP Controller determines the COMMON label space for all its IGP Clients.
3. The controller will assign the global label for different services, and these label bindings will be flooded through IGP protocol to IGP clients.
4. IGP Client receives the MPLS Global Labels, and generates corresponding MPLS forwarding entries.

IGP is suitable for the use cases of MPLS global label in the intra-domain scenario. These use cases include MPLS virtual network and segment routing as defined in [I-D.li-mpls-global-label-usecases].

4.5. Virtual Link

When the IGP-based Central Controlled framework is applied, one possible scenario is partial deployment. That is, part of the

existing network will be converted to be controlled in the central control mode. The application scenario is shown in the following figure:

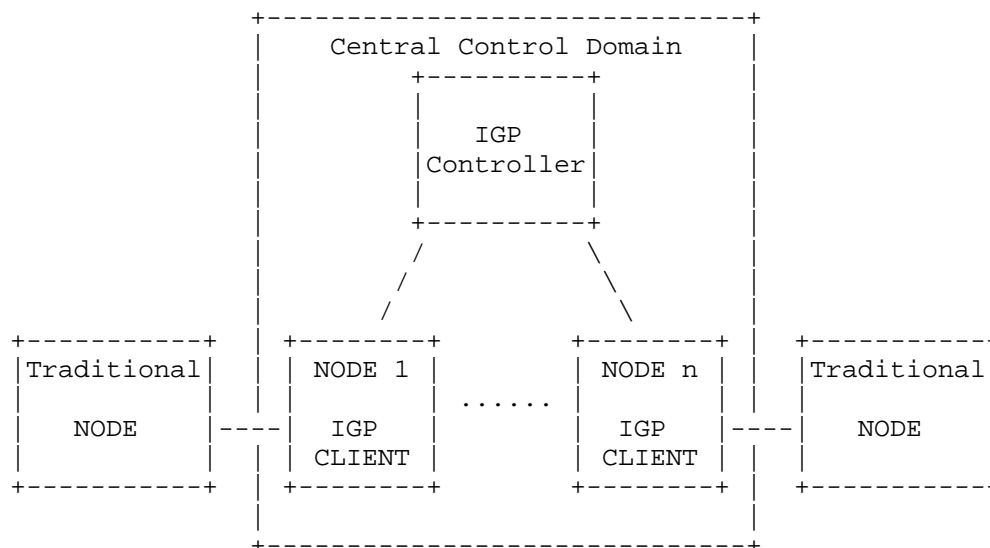


Figure 3: Partial Deployment of Central Controlled IGP

In this scenario, it is not necessary for the traditional nodes to learn the detailed topology information of the central control domain. The information flooded between the central control domain and the traditional nodes can be reduced. The central control domain can only advertise virtual links which connect the edge nodes in the domain that the traditional node can be aware of. The process can reduce the pressure of the traditional node for flooding and improve convergence performance.

In the central control domain, the controller can apply the policy defined by the applications to control whether the virtual link will be advertised to the outside and what metric is advertised to affect the route calculation of the outside network.

5. IANA Considerations

TBD.

6. Security Considerations

TBD.

7. References

7.1. Normative References

- [ISO/IEC 10589]
ISO, "Intermediate system to Intermediate system routing information exchange protocol for use in conjunction with the Protocol for providing the Connectionless-mode Network Service (ISO 8473)," ISO/IEC 10589:1992.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.

7.2. Informative References

- [I-D.chen-ospf-ttz]
Chen, H., Li, R., Cauchie, G., Retana, A., Ning, S., Toy, M., and L. Liu, "OSPF Topology-Transparent Zone", draft-chen-ospf-ttz-06 (work in progress), July 2013.
- [I-D.ietf-ospf-te-metric-extensions]
Giacalone, S., Ward, D., Drake, J., Atlas, A., and S. Previdi, "OSPF Traffic Engineering (TE) Metric Extensions", draft-ietf-ospf-te-metric-extensions-04 (work in progress), June 2013.
- [I-D.li-mpls-global-label-framework]
Li, Z., Zhao, Q., and T. Yang, "A Framework of MPLS Global Label", draft-li-mpls-global-label-framework-00 (work in progress), July 2013.
- [I-D.li-mpls-global-label-usecases]
Li, Z., Zhao, Q., and T. Yang, "Useases of MPLS Global Label", draft-li-mpls-global-label-usecases-00 (work in progress), July 2013.
- [I-D.li-ospf-ext-green-te]
Yan, G., Yang, J., and Z. Li, "OSPF Extensions for MPLS Green Traffic Engineering", draft-li-ospf-ext-green-te-01 (work in progress), October 2013.
- [I-D.ylz-ospf-lsdb-sync-group]
Yan, G., Liu, Y., and X. Zhang, "OSPF Extensions for Link State Database Synchronization Group", draft-ylz-ospf-lsdb-sync-group-01 (work in progress), October 2013.

Authors' Addresses

Zhenbin Li
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: lizhenbin@huawei.com

Huaimo Chen
Huawei Technologies
Boston, MA
USA

Email: huaimo.chen@huawei.com

Gang Yan
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: yangang@huawei.com

Network Working Group
Internet-Draft
Intended status: Standards Track
Expires: April 21, 2014

X. Zhang
G. Yan
Huawei Technologies
October 18, 2013

Algorithm for Ordered Metric Adjustment
draft-zxd-rtgwg-ordered-metric-adjustment-00

Abstract

Upon link down event or link up event, each device in network individually schedules route calculation. Because of different hardware capabilities and internal/external environments, the time to update forwarding entries on these devices are disordered which can cause a transient forwarding loop. This document introduces a method to prevent forwarding loop by adjusting link metric gradually for several times.

Requirements Language

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

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 <http://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 April 21, 2014.

Copyright Notice

Copyright (c) 2013 IETF Trust and the persons identified as the document authors. All rights reserved.

This document is subject to BCP 78 and the IETF Trust's Legal Provisions Relating to IETF Documents (<http://trustee.ietf.org/license-info>) in effect on the date of publication of this document. Please review these documents carefully, as they describe your rights and restrictions with respect to this document. Code Components extracted from this document must include Simplified BSD License text as described in Section 4.e of the Trust Legal Provisions and are provided without warranty as described in the Simplified BSD License.

Table of Contents

1. Introduction	2
2. Overview of Algorithm	3
2.1. Link up event	4
2.2. Link down event	7
3. Algorithm Sections	8
3.1. Calculating the adjustment range of link metric for each node	8
3.2. Determine existing forwarding loop or not between two direct nodes	8
3.3. Algorithm of multiple nodes simultaneously switch optimal path without forwarding loop	9
4. IANA Considerations	9
5. Security Considerations	10
6. Normative References	10
Authors' Addresses	10

1. Introduction

The internet is the most popular network, it is a distributed system, Depend on its configuration, each network device communicates with its neighbor, calculate routes and generate the FIB individually, finally, the packet will be forwarded hop by hop. But due to the difference of each device hardware capabilities and internal/external environments, the route calculation cannot be scheduled at same time, the micro-loop occur, and some mechanisms are already provided in IETF to resolve this issue, like ordered FIB.

This document tries to provide a different method to resolve this issue.

In figure 1, there are some forwarding loop scenarios:

- o Upon link BA down event, for the destination A, if B updates its forwarding entry before G, a transient forwarding loop occurs between B and G. Node failure MAY be treated as multiple links' failure, such as B fails, the links GB, BA, BC go to down. For

the destination A, if G updates its forwarding entry before I, a transient forwarding loop occurs between I and G.

- o Upon link BA up event, for the destination A, if G updates its forwarding entry before B, a transient forwarding loop occurs between B and G. Node failure recovery MAY be treated as multiple links' failure recovery, such as B recovers, the links GB, BA, BC go to up. For the destination A, if I updates its forwarding entry before G, a transient forwarding loop occurs between I and G.

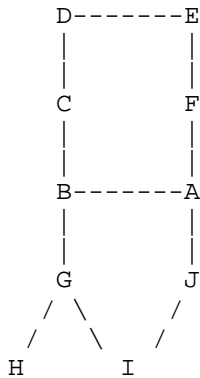


Figure 1 Topology (all links with metric 10 except links AF and AJ with metric 100)

2. Overview of Algorithm

This document introduces a method to prevent forwarding loop by adjusting link metric gradually. There are two cases to be considered here:

- o Link up event: The link metric will be decreased from maximum to configuration value. Node failure recovery MAY be treated as multiple links' failure recovery.
- o Link down event: The link metric will be increased from configuration value to maximum. Node failure MAY be treated as multiple links' failures.

2.1. Link up event

As we know, the optimal paths from other nodes to node R can be represented as the RSPF tree with root R. We assume that the link XR between node X and R goes to up, some nodes MAY switch their optimal paths to R and the new optimal paths include the link XR. If the metric of link XR is small enough, X will be the children of R in RSPF tree and the nodes of the sub-tree under X on the RSPF tree will switch their optimal paths to R, other nodes' optimal paths are not changed. The nodes whose optimal paths to the R are changed are denoted by set of S. If the nodes in S switch their optimal paths when link XR goes to up, some forwarding loop MAY exist as described in section 1. In order to prevent the forwarding loop, we can control the nodes in S to switch optimal paths gradually instead of switching all the nodes at the same time. For the case of link up event, the metric of link XR is decreased gradually by several times. Once the metric of link XR is adjusted, one node or several nodes any two of which do not have forwarding loop will switch optimal path to R. Until the metric of link XR is decreased to configuration value, all the nodes in S switch their optimal paths to R. We give an example to describe the procedure of adjusting link metric as follows:

In Figure 1, suppose the link BA is down. All the paths of the other nodes to node A can be represented as RSPF(Reverse Shortest Path First) tree with root A(as figure 2 below).

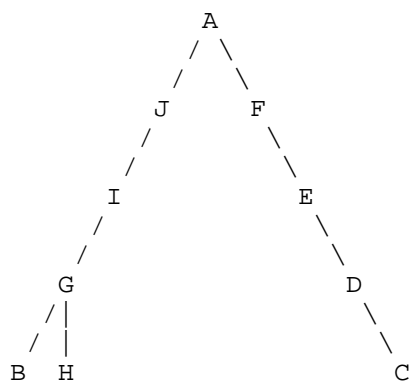


Figure 2 Reverse Shortest Path First Tree

After link BA going to up, node B will start to adjust the metric of link BA and repeat adjusting several times as below:

- o The metric of link BA is set to 121. For the destination A, only B's optimal path has changed. The following figure 3 describes the RSPF tree with root A after the first adjustment.

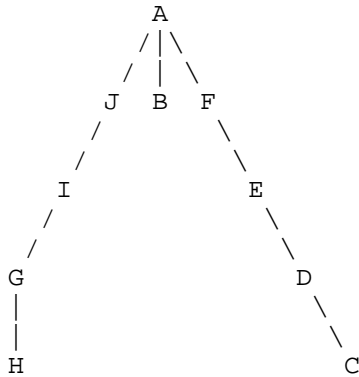


Figure 3 Reverse Shortest Path First Tree (The metric of link BA is 121)

- o The metric of link BA is set to 101. For the destination A, the node C, G and H's optimal paths have changed. The following figure 4 describes the RSPF tree with root A after this adjustment.

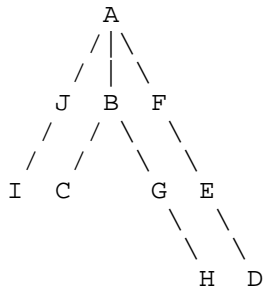


Figure 4 Reverse Shortest Path First Tree (The metric of link BA is 101)

- o The metric of link BA is set to 81. For the destination A, the node D and I's optimal paths have changed. The figure 5 below describes the RSPF tree with root A after this adjustment.

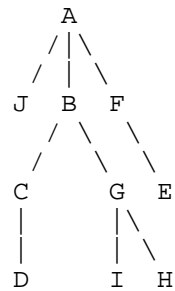


Figure 5 Reverse Shortest Path First Tree(The metric of link BA is 81)

- o The metric of link BA is set to 61. For the destination A, the node E and J's optimal paths have changed. The figure 6 below describes the RSPF tree with root A after this adjustment.

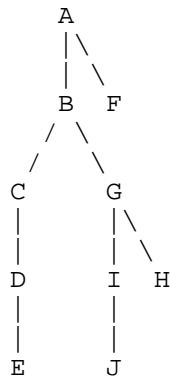


Figure 6 Reverse Shortest Path First Tree(The metric of link BA is 61)

- o The metric of link BA is set to 10 which is configuration value. For the destination A, node F's optimal path has changed. The figure 7 below describes the RSPF tree with root A after this adjustment.

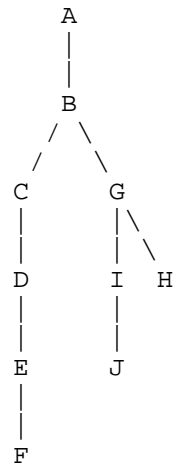


Figure 7 Reverse Shortest Path First Tree(The metric of link BA is 10)

As shown in the example above, one or more nodes' optimal paths to node A will be affected when the metric of link is adjusted every time. The nodes not affected keep the outgoing interfaces and next hops unchanged. There is no forwarding loop during this adjustment.(as in Figure 3 H, G, etc.).

Once the link metric is adjusted, the new LSP/LSA will be generated with the new metric information and will be flooded to the same area/level, and all the nodes in the same area/level will calculate routes. So all the nodes need enough time to flood new LSP/LSA and calculate routes before the next adjustment of link metric. Usually, it needs a few seconds to tens of seconds delay to start a new adjustment. The delay between different adjustments will avoid to affect each other.

Similarly, for the destination B, we can use the same method to adjust the metric of link AB to avoid forwarding loop.

2.2. Link down event

In Figure 1, supposing the link BA goes to down, for the destination A, it can avoid forwarding loop by increasing metric of link BA from configured value to maximum. The process of adjustment is reverse to the process of link up event. And the RSPF tree with root A is changed from figure 7 to figure 2 gradually.

The adjustment of metric just involves the nodes connected to the failure link, other nodes just do normal route calculation. So the method is very convenient for incremental deployment.

3. Algorithm Sections

In figure3, the metric of the link BA is set to 121, node B switches an optimal path to A, while the other nodes do not switch their optimal paths. In fact the metrics ranged from 121 to 129 of the link BA is also valid to ensure that only the node B has to switch to an optimal path to A. The following algorithm 3.1 is used to calculate a reasonable adjustment metric range for each node to switch its optimal path without forwarding loop.

We first assume the failure link is L(for example, link BA in figure 1), and its configuration metric value is K. The destination node of link L is root node(for example, node A) which is referenced in the following sections.

3.1. Calculating the adjustment range of link metric for each node

1. Supposing the link L(for example, link BA in figure 1) is down, the metric of link L can be considered as maximum. It is easy to use RSPF algorithm to calculate the distance of each node i to root node. The distance from each node i to the root is recorded as $D(i, \max)$.
2. Supposing the link L is up, the metric of link L is set to configured value. It is easy to use RSPF algorithm to calculate the distance of each node i to root which is the destination node of link L. The distance from each node i to the root is recorded as $D(i, \min)$.
3. If node i switches its optimal path to root node, the reasonable upper metric of link L can be adjusted is $\text{COST}(i, \max)$, $\text{COST}(i, \max) = D(i, \max) - D(i, \min) + K$.
4. If node i switches its optimal path to root node, the reasonable lower metric of link L can be adjusted is $\text{COST}(i, \min)$, $\text{COST}(i, \min) = \text{MAX}\{\text{COST}(j, \max)\}$, where j is the son of node i in case of link L being up.
5. When the link L's metric is set in the range of $(\text{COST}(i, \min), \text{COST}(i, \max))$, node i can be switched to its optimal path to the root node without forwarding loop.

3.2. Determine existing forwarding loop or not between two direct nodes

F is the parent node, S is its son node. if $COST(F, \max)$ equals $COST(S, \max)$, when F switches to the new optimal path to root because of link L's metric's adjustment, S will switches simultaneously with F without forwarding loop.

3.3. Algorithm of multiple nodes simultaneously switch optimal path without forwarding loop

1. Initialize three queues: TentList, CandList, OutPutList.
2. The destination node of link L is recorded as root.
3. Push the root node to TentList.
4. Get the node N from TentList, where N is the node whose $COST(i, \min)$ is maximum in TentList. if TentList is empty, we cannot get any node, then this algorithm terminates.
5. Move node N to the tail of OutPutList.
6. Push every son node S_i of N to CandList, where $COST(S_i, \max)$ does not equal $COST(N, \max)$.
7. For each node m_i in TentList, if $COST(m_i, \max) > COST(N, \min)$, remove the node m_i from TentList.
8. When m_i is deleted from TentList, push every son node S_j of m_i to CandList, where $COST(S_j, \max)$ does not equal $COST(m_i, \max)$.
9. Move all the nodes from CandList to Tentlist, then CandList is empty.
10. goto 4.
11. When the algorithm finishes, OutPutList stores node $N_1, N_2, \dots N_s$,
 - * In case of link L going to up, the adjustment process of metric of link L is $COST(N_1, \min)+1, COST(N_2, \min)+1, \dots COST(N_s, \min)+1$, and configuration value K.
 - * In case of link L going to down, the adjustment process of metric of link L is configuration value K, $COST(N_s, \min)+1, \dots COST(N_2, \min)+1$ and $COST(N_1, \min)+1$.

4. IANA Considerations

This document includes no request to IANA.

5. Security Considerations

This document is not currently believed to introduce new security concerns.

6. Normative References

- [RFC1195] Callon, R., "Use of OSI IS-IS for routing in TCP/IP and dual environments", RFC 1195, December 1990.
- [RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.
- [RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
- [RFC5715] Shand, M. and S. Bryant, "A Framework for Loop-Free Convergence", RFC 5715, January 2010.
- [RFC6976] Shand, M., Bryant, S., Previdi, S., Filsfils, C., Francois, P., and O. Bonaventure, "Framework for Loop-Free Convergence Using the Ordered Forwarding Information Base (oFIB) Approach", RFC 6976, July 2013.

Authors' Addresses

Xudong Zhang
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: zhangxudong@huawei.com

Gang Yan
Huawei Technologies
Huawei Bld., No.156 Beiqing Rd.
Beijing 100095
China

Email: yangang@huawei.com