INTERNET-DRAFT Arunkumar Arumuga Nainar
Intended Status: Informational draft Tata Communications Ltd
Expires: April 4, 2014 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

Table of Contents

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 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) n 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) n 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.

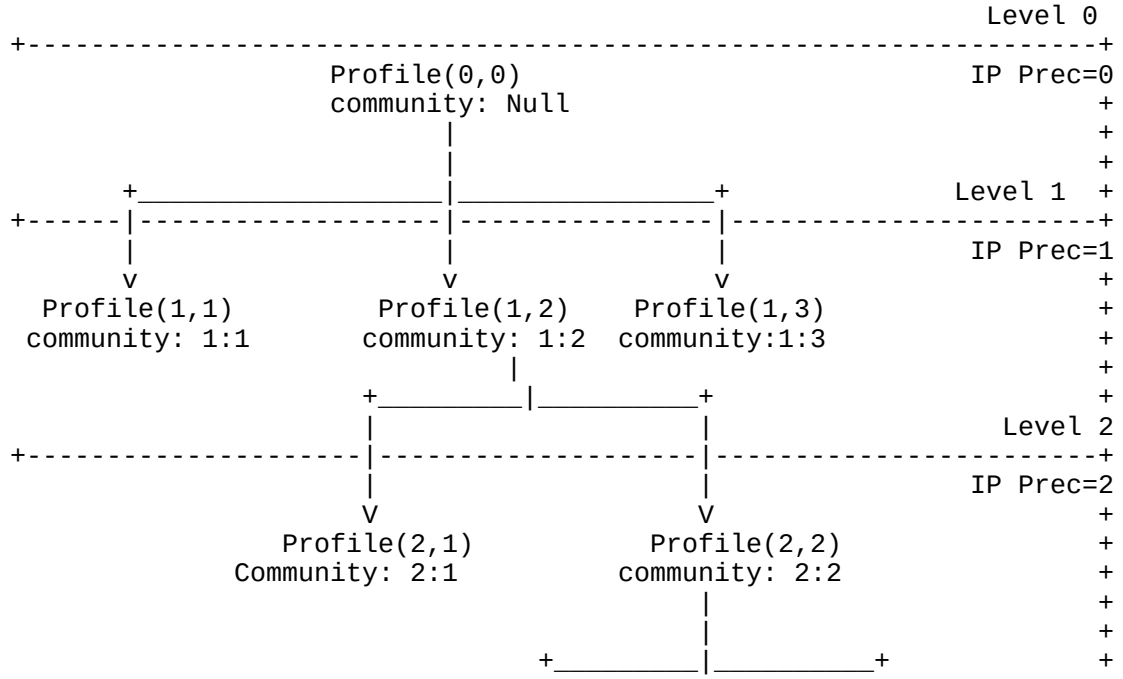
2. Helps translate the above concept/rules in deployment practice using available tools and technologies.

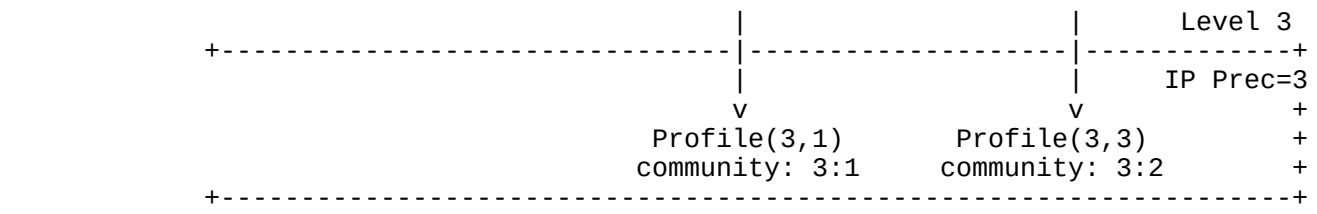
M-Tree based Association Model

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 Type1 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 be 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 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 in 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 through put 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

London EC4M 7AN
United Kingdom

EMail: arun.arumuganainar@tatacommunications.com