INTERNET-DRAFT
Intended Status: Informational draft
Expires: April 5, 2014

Arunkumar Arumuga Nainar
Tata Communications Ltd
October 2, 2014

Dynamic Path Selection Requirements
draft-arumuganainar-rtgwg-dps-requirements-01

Abstract

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.

Hence there is genuine requirement to offload some of the traffic to an alternate path while the primary path is still active and running. This document defines this requirement in detail.

http://www.ietf.org/1id-abstracts.html

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

Table of Contents

1 Introduction

In a large enterprise environment, network locations are often distributed over large geographical areas (often it spans different countries and continents). In such a case enterprise network managers often buy virtual private networks from service providers. Such a network is built over either MPLS or Internet/IPSEC extension to MPLS networks.

Generally the enterprise end user location connects in to the service provider using last mile connections. These last mile circuits comes in various technology options and bandwidth capacities. While the service provider core is very resilient and virtually has infinite capacity in terms of bandwidth, the weak link in the enterprise network is indeed the last mile.

Last miles are the costliest component for the enterprise and most of the times it contributes to 50-60% of total network costs. Yet these are the weakest link in the network. They constitute single point of failures and do fail frequently. Hence in order to increase the network availability the network managers buy redundant links.

With redundant links in place, the availability puzzle is fully resolved, however, it does pose a significant question. Half of the costliest resource that they have purchased will remain idle through out the life of this network. This happens while network managers work over time to resolve the issues that arise out of congestion on the primary circuit.

There is a genuine need to identify few select application and off load them on to a path that is considered to be the best path by the routing protocols. This draft defines and documents this requirement. This draft also describes the high level framework based on which such application off loading could be accomplished. The scope of draft also includes the evaluation of the existing techniques and their gap analysis.

1.1 Terminology

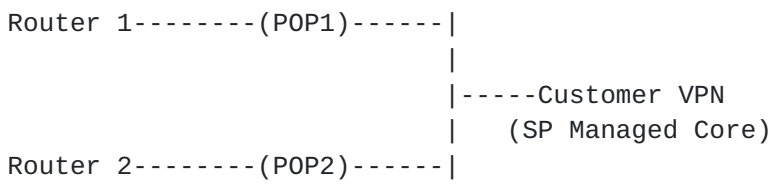
The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Basic assumption

This draft addresses a problem faced in an enterprise network. Typical enterprises comprise of collection of sites/Local Area

Networks (LANs) that are spread across a large geographical area (across cities, countries and continents). These sites are interconnected via a virtual private network of some form. Such enterprises typically go to single or even multiple service providers to establish this connectivity.

In such a case, the service provider will interconnect the site at their POP location via a last mile circuit. The choice of last mile technology could be diverse such as Ethernet, Frame-Relay, ATM, DSL, T1/E1/T3/E3, etc. Bandwidth capacity also varies from location to location (as low as 512 Kbps DSL to as high as 10 Gbps Ethernet). In cases where VPN connectivity is not possible, the VPN network can be extended in to a remote location via Internet/IPSEC technologies as well. Due to the diverse nature of the connectivity option, any frame work should be agonistic to technologies and choice of providers.



This draft assumes availability is very important for the target enterprises, hence there will be at least two last mile circuits that link them to the service provider core. Again, technology and the last mile provider can be chosen independently. Often enterprises selects a good or premium primary circuit and cheaper option for the secondary connectivity. For example

1) Primary: 10 Mbps Ethernet to MPLS; Secondary:- 10 Mbps Ethernet to MPLS

2) Primary: 10 Mbps Ethernet to MPLS; secondary:- 2 MBPS E1 circuit to MPLS

3) Primary: 1.5 Mbps T1; Secondary:- ADSL 2+ Internet/IPSEC

The above samples are for illustration taken from typical examples found in the field.

Routing schemes usually designate the premium circuit option as primary. Routing protocol metrics will hence be better for the link and all the traffic goes over it as long as they are active. Traffic shift to the secondary if the primary circuit goes down. This kind of routing scheme is called classical routing or Active-passive routing scheme.

3.Problem statement :-

In the classical routing scheme, all traffic goes over the primary circuit. Sizing of the circuit will depend on the bandwidth requirements of all the applications running on the LAN side which need remote connectivity. Last mile circuits generally constitute a large bulk of the wan network cost. In some networks that are geographically well spread this could constitute 50-60% of total WAN network costs, hence over sizing the network is not an option.

As an alternative, the Network Administrator could resort to Quality of Service. This will ensure applications that are critical to the business get enough bandwidth, but as collateral damage, applications that are not so critical could be starved of the bandwidth. This starvation could happen even while secondary circuit is active and contains lots of unused bandwidth.

Key challenge here is to find a way to select certain applications that are considered not-so critical for the business and off load them on to secondary circuit. This would ensure applications that would otherwise be starved of bandwidth in the primary circuit will have lots of room in the secondary circuit. This kind of active-active routing scheme will hence forth be called as DPS (Dynamic Path Selection) routing. This when achieved, will improve the overall user experience for the not-so-critical applications without compromising the critical applications. In majority of the situations this will result in a slightly better experience for the critical applications.

It should be noted that during the event of a primary circuit failure, both critical and non-critical applications will go over the secondary circuit. Hence the QoS mechanism will have to be applied even on the secondary circuit so as to punish the not-so-critical applications. Though this is a real problem, the occurrence of such an event is very rare. SLA for a premium circuit such as Carrier or Metro Ethernet could be as high as 99.5% . With this SLA in place network administrators can guarantee acceptable level of performance for all applications 99.5% of the time whilst for critical applications the experience can be guaranteed 99.999% of times.

If this DPS routing can be achieved, then it will directly translate in to greater user experience for users with out compromising the cost that the Network managers will have to spend on the WAN networks.

4. DPS Routing Requirements

The DPS requirements can be summarized in 5 simple steps.

1) Any two sites participating in the DPS routing, will have to agree on off loading patterns. This pattern is referred to as DPS Profiles. Example of a profile is illustrated below.

Profile 1: Critical Application: SAP, Citrix, VOIP ,Telnet; Non-Critical: remaining applications

Profile 2: Critical Application: SAP, Citrix, VOIP, Telnet, SMTP, HTTP; Non-Critical: remaining applications

Profile 3: Critical Application: All non-Internet traffic; Non-Critical: Internet-traffic

If three sites wanted to participate in DPS routing, they need to exchange the profile information and agree on a common profile to off load. It should be possible that the first site will negotiate to use profile 1 for the second site, and profile 2 with the third. The underling requirement is if two sites communicate they should be able to agree on a common profile.

2) Traffic hitting the input interface of router will have to be first classified based on application. Here the application can be defined as follows:

a) A specific known TCP port number

b) Combination of TCP Port and Server's IP Address

c) A known Layer 7 Signature

For example:

a) Traffic sent to TCP Port (80) could be classified as HTTP application.

b) Traffic sent to TCP Port 80 and Server IP (ex: proxy server) can be classified as Internet

c) Traffic sent to TCP Port 80 and Server IP (ex. proxy server)sent to the URL www.example.com can be classified as "EXAMPLE.COM APP". Here the classification was based on the Layer 7 Signature.

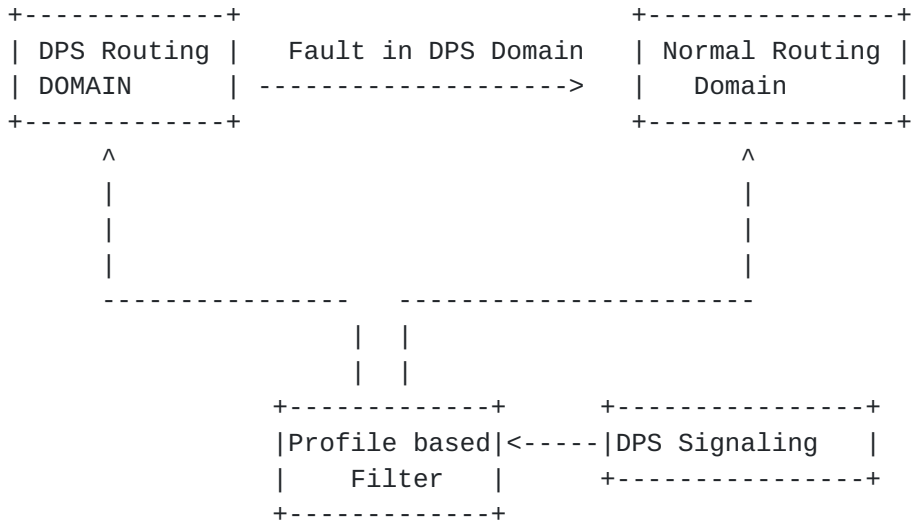
3) If the profile is agreed and classification is done, then the participating router should be able to filter the traffic and send it

across available paths.

4) During step 2, if the traffic is pushed to the secondary circuit (not the best path as per the routing protocol)it should stick with the circuit end to end. For example an application that has been off loaded will have to enter the remote site via the secondary circuit. Return path for that flow will have to take the secondary path at both locations so as to prevent asymmetric routing.

As the quality of the link does vary largely at any given site asymmetric routing will invariably create a application performance issues and hence should be avoided at all times.Hence, path symmetry should be honored even during positive and negative events on the network. For example if the secondary circuit fails, either locally or at the remote site, it is treated as a negative event and during this time path symmetry should not be compromised.

5. Dynamic Path selection Framework



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 will be off loaded to secondary path (also referred to as DPS path on this draft).

Task 2: At the time of forwarding the packets, packet should be filtered based on agreed application set. Packets should then be pushed in to appropriate paths.

Task 3: If the packet is pushed on to a DPS path, it should always use the secondary link end to end. .

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.

To achieve the above four tasks, this draft recommends four functional blocks as described in the above diagram. These functional blocks are individually designed and fine tuned to achieve various level of sophistication. Details of the building blocks are described in the below section.

Normal Routing Domain:

This is the Active-Passive or classical routing as described in section 2. Any site that is not running DPS (DPS unaware), will have only this module. Here, the routing protocol will configured in a traditional fashion. The primary circuit will have a best metric and will be preferred and the secondary will be acting as backup.

DPS Routing domain:

This will be overlay routing domain that will be built over the existing Normal Routing domain. However some of the paths will not be considered while building the domain. For example Primary Circuits can be ignored while building the overlay domain. Alternatively overlay domain can be built in such a way that it prefers secondary circuit as a path with better metric and hence primarily route over secondary path.

It is mandatory that all the LAN routes that are available on the normal routing domain is fully injected in to the overlay DPS Routing domain. However, in order to separate the traffic, the overlay DPS Routing domain will be put into a separate VRF. Once a packet is pushed in to the overlay DSP Routing domain, the Normal Routing domain is never consulted, until the time the fault tolerance mechanism pushes traffic out of the overlay DPS Routing domain.

DPS Signaling

One of the core objective of DPS is to avoid asymmetry while routing. Hence, if a particular application is off loaded, then the remote network should also off load the same application. The signaling

module will provide a mechanism for participating sites to communicate with each other and agree about the applications that need to be off loaded.

During the signaling, sites will exchange the Application Profiles as described in section 4. Sites could be able to support more than one profile, however when they communicate a common profile must be chosen. For example, profiles can be defined as follows:

Profile 1: Critical Application: SAP, Citrix, VoIP, Telnet; Non-Critical: remaining applications

Profile 2: Critical Application: SAP, Citrix, VoIP, Telnet, SMTP, HTTP; Non-Critical: remaining applications

Site A & B can support both profile 1 and profile 2 Site C can support profile 1 Site D can support Profile 2

When Site A talks to Site C, applications will be off loaded as per definition of Profile 1

When Site A talks to Site C, applications will be off loaded as per definition of Profile 2

When Site A talks with Site B, off loading can be done either by choosing one profile or by choosing a combination of both profiles. One of the profiles can be dynamically chosen based on the network conditions prevailing at the time.

For example Site A could signal to Site B its preference for profile 1 when the link utilization level is very high. At all other times it can prefer profile 2 when dealing with Site B. Such a constraint based profile selection should be explicitly acknowledged and agreed by site B. If agreement could not be reached then it will default to normal routing.

It should be noted that the profile definition as such is static. All of the application sets will have to be individually defined by the network administrator. But selection of profiles as such can be dynamic and it could depend on a) local or remote network conditions and b) capability of the site of which the target IP address belongs to.

Profile Based Filtering Module

Once off loaded applications are defined, the framework should support a classification and filtering engine. Classification could be done via variety of application signatures. For example an

application signature can be defined as:

1) specific TCP port or an Server IP address

2) A combination of TCP port and IP Address

3) A Layer 7 Signature (eg: specific URL like http://www.example.com)

The profile based filter should be able to classify the packets and then push the off loaded application traffic in to the DPS Routing domain. Other traffic should be routed normally.

6. Packet flow in DPS Environment

If a site is deployed conforming to the above specifications, the following things will happen when the packet hits router.

1) When the packet hits the input interface, it gets classified based on the Application Signature.

2)If signaling happens in the forwarding plane, then the Signaling Module will communicate with the remote site and agree on the profile that will be used. If the signaling is implemented in the control plane, then the profile negotiation would have already taken place. In both the scenarios, the profile information are fed in to the profile filter.

3) Once the profile and application signature is determined, the profile based filter will push the packet in to the appropriate routing domain.

4) In the case of the packet being pushed in to the DPS Routing domain, the forwarding happens based on information available in the DPS Routing domain.

5) Failures might happen in the network and this might result in loss of routing information in the DPS routing domain. In such a case the DPS routing domain will push the packet out of the domain and in to the Normal Routing domain. Under no circumstances should the packet be dropped due to non-availability of routing information in the DPS routing domain.

8. Implementation Details and Gap Analysis.

8.1 DPS Routing domain:-

The main constraint is that provider devices should be transparent to DPS Domain routing. This adds more complexity to the design. The following techniques were considered for implementation:

a) Multi-Topology routing:

Multi-Topology Routing was considered but found unsuitable for DPS routing. This is because it requires dependency on providers supporting the DPS Routing scheme. Also, it forces packets to be marked only with two DSCP values (1 for normal routing and another other for DPS routing). Some implementations mandate usage of the same or similar DSCP/IP Precedence values for traffic traveling through both routing domains.

And hence it was found that MTR as not suitable for this purpose.

b) A separate VPN for DPS:

This does work fine for the DPS Routing domain. Here we extend a second sub-interface to the same POP via the secondary last mile circuit. The second sub-interface could be put on a separate VPN.

Technically this is feasible. There few implementations in the field of this type. However this is not a preferred option. Any new VPN on a MPLS network will come with extra cost and most of the time, the costs are prohibitively high.

c) Overlay VPN Using Dynamic Mesh VPN tunnels:

Here the overlay network is built using DMVPN (Dynamic Mesh VPN). Details of DMVPN is described in detail in the draft-detienne-dmvpn. There is a vendor implementation that exists today that supports a multi-point overlay tunnel as described in the draft. This could be used to create an overlay DPS Routing domain. A standard routing protocol can be run over this overlay network.

This technique is the widely deployed technique used in field deployments today. In the production environment, several implementations were done with largest one consisting of 300 sites & 2000 prefixes.

It should be noted that the proposed draft draft-detienne-dmvpn suggests that binding IPSEC encryption on to the DMVPN tunnel is mandatory. DPS Routing does not require IPSEC. Current implementations were all done without encryption. This draft recommends the DMVPN draft to be amended so that IPSEC binding requirement is de-linked from the DMVPN standard.

8.2 DPS Signaling:-

DPS signaling can be implemented in both forwarding plane and also in the control plane.

Control plane implementations are done via BGP. Here, BGP will be used as a routing protocol for the Normal Routing domain. Profile information can be exchanged via standard community. For example:

1) Profile 1 can be represented by community 100:1

2) Profile 2 can be represented by community 100:2

If a given site supports only profile 1, it will advertise site prefixes with a community 100:1. If the same site wanted to support both the profiles then the prefix would be sent with both communities, 100:1 and 100:2. At the remote end, the site will understand the capabilities based on the communities associated with the prefix.

For example Site 1 supports both profile 1 and profile 2 and hence it advertise the prefix with community 100:1 and 100:2. But site 2 supports only profile 1 (100:1). The common profile between both the sites is 100:1 and hence profile 1 will be used. It should noted that both site 1 and site 2 with full certainty will use profile 1. Under such a scenario asymmetric routing will not happen.

Currently this scheme has been implemented using one vendor specific implementation. In the current implementation, the router will colour the packet with IP Precedence values. There will be one to one correspondence between the profile chosen and the IP Precedence values used for colouring. Once the packet is coloured, the profile based filter can applied. Filtering will involve inspecting both the colouring done by the Signaling Module and the application to which the packet belongs to.

While this works perfectly well on large scale network, a couple of short comings remains.

1) The trust model could not be supported when DPS is enabled. This is because the DPS Signaling Module will remark the TOS bits during the colouring operation. Hence an alternate mechanism needs to developed to color the packet with out modifying the TOS bits.

2)Because profiles are exchanged via BGP, the router negotiates the off loaded applications list at the time of exchanging the BGP routing information. This would mean routers will not be able to react to local or remote network condition. For example, the choice

of profiles can not be changes if the link utilization is very high. The draft draft-arumuganainar-rtgwg-DPS-L4-path-preference-negotiation-00 suggests an alternate mechanism that will allow profile negotiation to happen when the TCP connection gets established. This will enable link preference to be negotiated for each TCP connection and hence such a mechanism will be more dynamic then the BGP one

8.3 Profile based Filter

Each profile is associated with a pre-determined list of applications. These applications are determined by the network administrator as an application that can be offloaded.

The administrator can define multiple profiles and associate this with one or more sites. When two sites communicate, they signal each other and agree on a single common profile. The Signaling Module will then colour the packet indicating which profile to be used while deciding the application to be off loaded.

Once the steps are complete, the profile based filter can be applied. The filter will look for the colour and then matches the protocol with corresponding off loaded application. If a match is found the packet will be pushed in to the DPS Routing domain, otherwise the packet will be normally routed.

Current implementations are done using Policy Based Routing (PBR). Here, filtering is done using a special PBR policy. This will check the colour of the packet and the pre-determined list of applications. It could then push the packet in to the appropriate routing domain.

While in theory PBR works perfectly fine, PBR is a processor intensive mechanism and hence when it is applied, the throughput of the router is adversely impacted. Hence, router scalability should be kept in mind while sizing the router for any particular site. In the future a light weight mechanism optimised for profile based filtering could be developed as well.

7. Summary

To summarise, by adopting the DPS frame work, true end to end application based routing scheme could be achieved. The DPS framework has the following advantages:

1. Give lots of room for Network Manager to determine which path should be used for which application.

2. DPS also provides a scalable frame work for achieving the above

objective.

3. By modularizing the DPS components, advanced development of individual components becomes possible. Troubleshooting will also get greatly simplified.

4. DPS capable sites can co-exist with non-DPS sites and this capability provides enough room for a phased migration. Thus 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 any other devices in the path 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

10.2 Informative References

11 Acknowledgements

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

Authors' Addresses

Arunkumar Arumuga Nainar
Tata Communications (UK) Limited
1st Floor
20 Old Bailey
London EC4M 7AN
United Kingdom

EMail: arun.arumuganainar@tatacommunications.com