Network Working Group
Internet Draft
Document: draft-mirtorabi-ospf-tunnel-adjacency-00.txt
Expiration Date: November 2003

Sina Mirtorabi
Peter Psenak
Cisco Systems
May 2003

OSPF Tunnel Adjacency
draft-mirtorabi-ospf-tunnel-adjacency-00.txt

Status of this Memo

Abstract

OSPF specification requires that intra-area paths are always preferred over Inter-area paths regardless of the path's cost. This document describes a solution that will remedy this limitation without introducing any significant change to the current specification. Further, this solution provides some other benefits such as automatic partition repair described in application section.

1. Motivation

There could be a requirement to prefer an Inter-area path over an intra-area path, for example in order to utilize the high bandwidth backbone path to transit the intra-area traffic from a non-backbone area. The current OSPF specification does not provide any generic

mechanism to be able to achieve this. In some situations Virtual Link (VL) can help, however there are some restrictions applied to VL:

a) Transit area needs to be a non-backbone regular area

b) VL prevents summarization of backbone prefixes into transit area

c) VL cannot be configured through Stub or NSSA [2] areas

2. Proposed Solution

The Tunnel Adjacency (TA) proposal uses a similar concept as virtual link, e.g. forming an (possibly multihop) adjacency between two ABRs through the transit area, however TA can be configured for any area and the transit area can also be any area.

Tunnel adjacency's concept is close to VL in the way of establishing an adjacency, sending unicast OSPF packets and synchronizing the database over it. Data packet forwarding between the two ABRs is different from a VL in that the packets are tunneled if the TA path spans multiple hops. This removes the requirement for routers internal to the transit area to have the TA area's unsummarised intra-area routes. The rest of this document describes TA specification.

3. Bringing up the tunnel adjacency

TA is configured between two ABRs attached to the same OSPF area. This area is called Tunnel-adjacency Transit Area (TTA). Similar to Virtual Link, TA is identified by the Router ID of the other endpoint. Once a tunnel adjacency for a given area is configured and an intra-area path exists between the two ABRs through the TTA, the router can start forming adjacency with the remote neighbor by sending unicast Hellos and synchronizing the database over TA as specified in OSPF [1].

The interface MTU should be set to 0 in Database Description Packets sent over the TA as it is done with virtual links. TA could be configured as a Demand Circuit (DC) in order to reduce Hello exchange and periodic LSA flooding.

4. Tunnel adjacency encapsulation

User traffic routed based on the presence of the TA will be encapsulated on the TA endpoints in the following way:

a) If both ends of the TA are directly connected to the same network and the best intra-area path to the TA endpoint in the

TTA is over this direct network connection, NO special encapsulation is needed.

b) Otherwise the traffic is further encapsulated (tunneled) and sent directly to the TA endpoint. The encapsulation type is left to the implementation and different encapsulation types could be specified through configuration. However, in order to have interoperability between vendors all implementation should support GRE encapsulation.

5. Advertising tunnel adjacency

TA is announced as an unnumbered point-to-point link. Once the router's TA reaches the FULL state it will be added as a link type 1 to the Router LSA with:

```
Link ID   = remote's Router ID
Link Data = router's own IP address associated with TA
Cost      = intra area cost to the TA endpoint in the TTA or the
            configured cost
```

The IP address specified in the link data is computed during the routing table build process for the TTA.

6. Tunnel adjacency interface data structure

An OSPF interface data structure is created for each configured tunnel adjacency. The cost of the TA is configurable allowing a traffic path to be selected independent of the intra-area path cost. The default cost is equal to the intra-area cost to reach the remote TA's neighbor in the TTA.

TA is considered as unnumbered point-to-point interface.

7. Tunnel adjacency interface FSM

TA Interface FSM is the same as specified in OSPF [1].
The InterfaceUp event for TA interfaces is generated once the intra-area path to the remote end of the TA becomes reachable through the TTA.

InterfaceDown event is generated for TA when the intra-area path to the remote end of the TA is lost in TTA.

8. Tunnel adjacency neighbor data structure

TA neighbor data structure is identical to the neighbor data structure for standard OSPF adjacencies as specified in OSPF [1].

9. Tunnel adjacency neighbor FSM

TA neighbor FSM is identical to the neighbor FSM for standard OSPF point-to-point adjacencies.

10. Tunnel adjacency OSPF control packet processing

OSPF control packet processing is specified in OSPF [1] section 8.
This section is modified as follow :

[...]

The IP source address should be set to the IP address of the sending interface. Interfaces to unnumbered point-to-point networks have no associated IP address. On these interfaces, the IP source should be set to any of the other IP addresses belonging to the router. For this reason, there must be at least one IP address assigned to the router. Note that, for most purposes, virtual links and tunnel adjacency act precisely the same as unnumbered point-to-point networks.

However, each virtual link or tunnel adjacency does have an IP interface address belonging to transit area or TTA (discovered during the routing table build process) which is used as the IP source when sending packets over the virtual link or tunnel adjacency. If there is not at least one IP address belonging to Transit area or TTA and the virtual link or TA is configured, a router could advertise any of its attached IP address as a stub link (Link ID set to the router's own IP interface address, Link Data set

to the mask 0xffffffff) to the transit area.

[...]

Receiving protocol packets as described in 8.2 is changed as follow:

Next, the OSPF packet header is verified. The fields specified in the header must match those configured for the receiving interface. If they do not, the packet should be discarded:

- o The version number field must specify protocol version 2.

- o The Area ID found in the OSPF header must be verified. If all of the following cases fail, the packet should be discarded. The Area ID specified in the header must either:

  (1) Match the Area ID of the receiving interface. In this case, the packet has been sent over a single hop. Therefore, the packet's IP source address is required to be on the same network as the receiving interface. This can be verified by comparing the packet's IP source address to the interface's IP address, after masking both addresses with the interface mask. This comparison should not be performed on point-to-point networks. On point-to-point networks, the

interface addresses of each end of the link are assigned independently, if they are assigned at all.

  (2) Indicate a non-backbone area. In this case, the packet has been sent over a tunnel adjacency. The receiving router must be an area border router, and the Router ID specified in the packet (the source router) must be the other end of a configured tunnel adjacency. The receiving interface must also attach to the TTA. If all of these checks succeed, the packet is accepted and is from now on associated with the tunnel adjacency for that area.

  (3) Indicate the backbone. In this case, the packet has been sent over a virtual link or tunnel adjacency. The receiving router must be an area border router, and the Router ID specified in the packet (the source router) must be the other end of a configured virtual link or tunnel adjacency. The receiving interface must also attach to the virtual link's configured

transit area or tunnel adjacency's configured TTA. If all of these checks succeed, the packet is accepted and is from now on associated with the virtual link or tunnel adjacency.

[Note if there is a match for both a VL and TA then this is a configuration error that should be handled at the configuration level.]

- o Packets whose IP destination is AllDRouters should only be accepted if the state of the receiving interface is DR or Backup (see Section 9.1).

[...]

11. Tunnel adjacency next hop calculation

The next-hop to reach the TA endpoint is equal to the next-hop associated with the TA endpoint inside the TTA.

Data packet forwarding between the two ABRs is different from a VL in that the packets are tunneled if the TA path spans multiple hops. This removes the requirement for routers internal to the transit area to have the TA area's unsummarised intra-area routes.

12. Virtual link - tunnel adjacency comparison

Virtual link has the following limitations:

1) The link should belong to a non-backbone area

2) Backbone area route cannot be summarized into the Transit area

3) VL can not be configured through Stub or NSSA area

Tunnel adjacency remedies all the above limitations. Further it will allow:

a) The cost of TA is configurable allowing a traffic path to be selected independent of the intra-area path cost, making it ideal to force a traffic path.

b) It can be used as an on demand partition repair. In this application, the TA will be established only if the two end of

TA are not reachable over a given area (see application section).

c) Multiple TAs could be configured over a TTA, each (TA) belonging to a different area in order to provide an intra area path for each area therefore saving cost of adding additional links (see application section).

Tunnel Adjacency can be considered as a generalization of Virtual Link.

13. Applications

In this section we give a few examples in which TA can be used.

13.1 Prefer Inter-area Path over intra-area Path

It is a common example that users would like to prefer the high bandwidth part of the backbone for traffic that can be strictly routed inside the non-backbone area.

Consider the following topology:

```
              R1-------backbone------R2
               |                      |
             area 1                 area 1
               |                      |
              R3--------area 1--------R4
```

Fig.1

The backbone link between R1 and R2 is a high speed link and could be used to forward part of the traffic of area 1 between R1 and R2. In the current OSPF specification, intra-area path are preferred over inter-area path. As a result R1 will always route traffic to R4 through area 1 involving lower speed links. Even to reach networks connected to R2 that belong to area 1, R1 will use the intra-area path over area 1.

By configuring a TA between R1 and R2 a p2p link will be advertised

into area 1 making the TA visible as a topological part of area 1 and by associating a low cost with TA, R1 will now compare two intra-area path and choose the one with lower cost.

Note that the above scenario can not be solved by VL since the link between R1 and R2 belongs to the backbone area and it is not desirable to move this backbone link into a non-backbone area.

It should also be noted that the connection between R1 and R2 in the backbone area could be multi-hop away, therefore there is no one hop limitation for TA.

13.2 On demand partition avoidance for the backbone

It can be desirable to not have a virtual link unless the backbone is partitioned, because the backbone's configured ranges are ignored when originating summary-LSA into a transit area. On demand partition repair requires checking to see if the two ends of TA are reachable through the backbone area before starting to form the adjacency.

When a TA is configured between the two ABRs a configuration option (automatic) will be used to not start sending Hello unless the other ABR is not reachable over the backbone area. Further, once the on demand adjacency is configured the check for ABR status is ignored during formation of the TA adjacency, because ABR may lose its backbone link and lose its ABR status, but the TA still needs to be established.

The cost of on-demand TA should automatically be set to maximum cost LSInfinity (16-bit value 0xFFFF). The reason to set the cost of TA to 0xFFFF in this case is to make it easier to detect that the partitioned area healed. During the SPF only the shortest path to the remote end of the TA is discovered and if the shortest path is via the TA itself, there is no simple way to find out that an alternative intra-area path to the remote end of the TA, other than over TA itself, exist. Setting the metric of TA to 0xFFFF makes this task easier.

```
  R1-------area 1--------R2
   |                      |
backbone              backbone
   |                      |
 R3--------backbone-----R4
```

Fig.2

In the above topology in order to have an on demand VL for the

backbone, an on demand TA can be configured between R1 and R2 for backbone through area 1. Should the backbone be partitioned, R1/R2 are not reachable over the backbone and they start forming adjacency through area 1 for the backbone.

13.3 On demand partition avoidance for summarized non-backbone area

In general when a non-backbone area is partitioned there is no need for partition repair as an intra-area route will be replaced by an Inter-area route for a segmented area. However this is not true any more if the area is summarized into the backbone. Consider the following topology:

```
  R1-------backbone------R2
   |                      |
 area 1                 area 1
   |                      |
 R3--------area 1--------R4
```

Fig.3

R1 and R2 are summarizing area 1 into the backbone area. when area becomes partitioned, for example when R3-R4 link is broken, R1 and R2 still continue to summarize area 1 into the backbone area. This can lead to blackholing of the traffic. The reason is that after the area partitioning, R1 or R2 will only have knowledge of their attached partitioned area. When R1 or R2 receives a packet that does not belong to it's attached partitioned area (as a result of advertising a summary) the packet will be discarded.

Note that R1 and R2 will install a discard route for the configured summary range. If the destination is not found in the attached area the packet is discarded following the discard route entry in the routing table.

By configuring an on demand TA for area 1 through the backbone, R1/R2 will establish an adjacency should area 1 becomes partitioned, that is when R1/R2 is not reachable over area 1.

Note that the cost of on-demand TA should be set to maximum cost

LSInfinity (16-bit value 0xFFFF).

13.4 Saving additional link between ABRs in a Hub and Spoke environment

Consider the typical Hub and Spoke topology in figure 4.


```
         R1---BB--R2
          | \    / |
          |  \  /  |
          |   \/   |
          |   /\   |
          |  /  \  |
        Spoke1  Spoke2
```

Fig.4

Only two Spokes are represented in figure 4, but in general we may have N spokes similar to Spoke1.

R1 and R2 are ABRs and can be multiple hops away over the backbone area (BB).Further, the ABRs are summarizing IP prefixes from all the attached areas into the backbone.

Case 1: Spoke1 and Spoke2 are in different area

Since both R1 and R2 are summarizing, there is a need for a link between R1 and R2 in each connected area. This is to guarantee an alternative path when the link between a spoke and Hub becomes unavailable.

For example imagine a network X advertised by Spoke1 and summarized by both R1 and R2. Later the link between R1 and Spoke1 goes down. When a packet arrives at R1 to be forwarded to Spoke1, R1 cannot send the packet to Spoke1 since the link is not available and R1 by summarizing may have installed a discard route for summarized range (here we assume the range is still 'active', as there may be other spokes in the same area as Spoke1, single attached to R1 and advertising prefixes that falls in the same range as X), so R1 will not use an inter-area path over R2. A link between R1 and R2, inside the same area as the link between R1 and Spoke1 is, would

prevent this problem.

Case 2: Spoke1 and Spoke2 are in the same area

Link between R1 and Spoke1 is broken. The path from R1 to Spoke1 is R1-Spoke2-R2-Spoke1 instead of R1-R2-Spoke1.

In general, for N areas being attached to the Hub routers, there is a need for N links between Hub routers. Multiple TA could be used through the backbone between the Hub routers to avoid using multiple physical links (each belonging to a different non-backbone area) between ABRs.

14. Tunnel adjacency parameters

Tunnel adjacency can be configured between area border routers having interfaces to a common area and it can belong to any area. The tunnel adjacency appears as an unnumbered point-to-point link in the graph for the configured area. Tunnel adjacency must be configured on both ends.

A tunnel adjacency is defined by the following configurable parameters:

- o The Router ID of the Tunnel adjacency's other endpoint.
- o The TTA area through which the tunnel adjacency runs.
- o The area to which the tunnel adjacency belong.

Optionally the following configurable parameters can be set:

- o cost of the tunnel adjacency which will overwrite the intra-area cost between the two endpoint of the TA.
- o Encapsulation type used between the two endpoint of the TA.
- o 'Automatic' option used for on demand partition repair.

15. Compatibility issues

All mechanisms described in this document are backward-compatible

with standard OSPF implementations.

16. Security

Tunnel adjacency specified in this document does not raise any security issues that are not already covered in [1].

17. Acknowledgments

Authors would like to thank Abhay Roy, Liem Nguyen, Acee Lindem and Pat Murphy for their comments on the document.

18. Reference

[1] Moy, J., "OSPF Version 2", STD 54, RFC 2328, April 1998.
[2] Murphy, P., "The OSPF Not-So-Stubby Area (NSSA) Option", RFC 3101, January 2003.

19. Authors' address

Sina Mirtorabi
Cisco Systems
225 West Tasman drive
San Jose, CA 95134
E-mail: sina@cisco.com

Peter Psenak
Cisco Systems
Parc Pegasus,
De Kleetlaan 6A
1831 Diegem
Belgium
E-mail: ppsenak@cisco.com