Inter-Domain Routing Working Group Kevin Fang, Cisco Systems
Internet Draft Feng Cai, Cisco Systems
Document: draft-zhiyfang-fecai-bgp-over-sctp-00.txt May.10 2009
Expires: November 2009

BGP-4 message transport over SCTP

Status of this Memo

This Internet-Draft will expire on November 10, 2009.

Copyright Notice

Abstract

This memo defines using SCTP for BGP-4 transport routing message. SCTP has many benefit for Signaling/Message transportation , BGP-4 transport over SCTP will enhance the link stability and efficiency.

Conventions used in this document

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].

Table of Contents

1. Introduction

This section explains the reasoning for using Stream Control Transmission Protocol(SCTP)transport Border Gateway Protocol 4(BGP-4) message.

1.1. Motivation

SCTP is a transport protocols defined in [RFC4960]. SCTP is designed to transport Public Switched Telephone Network (PSTN) signaling messages over IP networks, but is capable of broader applications.

We have observed that many of the NGN protocols(Sigtran,SIP,H.248,..) designed to support transport of such signaling are also useful for the transport of BGP.

BGP support for Four-octet AS Number Space [RFC4893], That means more and more Service Provider and Enterprise will get the AS Number, so it will becomes a large-scale network which will exchange a large amount of messages. As BGP-4 is transport independent, support SCTP is a relatively straightforward process, nearly identical to support for TCP.

1.2. Potential Benefits

Coene et. al. present some of the key benefits of SCTP[1]. We summarize some of these benefits to enhance BGP-4 transportation.

1.2.1. Fast Retransmission

SCTP can quickly determine the loss of a packet, as a result of its usage of SACK and a mechanism which sends SACK messages faster than normal when losses are detected.

When the Router working in HUB-SPKE environment(BGP Route-Reflector) if BGP-4 transport over TCP, the RR will receive a lot of TCP ACK, that may cause input-queue overflow. That may cause many TCP retransmission and Peering node lost, SCTP use SACK will be much better than TCP that may reduce the input-queue length.

When message lost, SACK mechanism will detect it faster than TCP.

1.2.2. SCTP Multi-Streaming

SCTP supports the delivery of multiple independent user message streams within a single SCTP association. This capability, when properly used, can alleviate the so-called head-of-line-blocking problem caused by the strict sequence delivery constraint imposed to the user data by TCP.

This can be particularly useful for applications that need to exchange multiple, logically separate message streams between two endpoints.

MPLS VPN is widely used in future network , It will require BGP-4 transport more and more routing informations, which means it will transport a large-number of messages. In BGP over TCP environment, Any peer failed to receive the message will cause TCP retransmit, that will cause Head of Line Blocking (HOL-Blocking). It will cause the Router can not send out message to other peering nodes. Multi-Streaming is a good mechanism to avoid such HOL-Blocking.

1.2.3 SCTP Multi-Homing

SCTP provides transparent support for communications between two endpoints of which one or both is multi-homed.

SCTP provides monitoring of the reachability of the addresses on the remote endpoint and in the case of failure can transparently failover from the primary address to an alternate address, without upper layer intervention.

BGP-4 over TCP will use a loopback interface to avoid the link failure. but in some particular scenario, BGP-4 message still transport over broken link. Although BGP-4 can support Bidirectional Forwarding Detection [BFD], but still can not provide multi-link solution.

If BGP-4 transport over SCTP , Routers can use Multi-homing to avoid single link failure.

1.3. Key Terms

Using SCTP transport BGP-4 message will offer the following services:

-- SCTP Multihoming gives a better redundancy solutions.
-- SCTP Multistreaming will avoid the HOL blocking.

See the BGP-4 specification [RFC4271] and Multiprotocol Extensions for BGP-4 [RFC4760] for an introduction to the concepts these textual conventions cover.

2. Using SCTP multistreming to avoid HOL blocking

BGP-4 now can support 4-Bytes ASN, also MultiProtocol BGP[RFC4760] extends BGP to allow information for multiple NLRI families and sub-families to transported in BGP. Current implementation just transport all the Routes in a single BGP session.

In fact, one malformed messages may cause the session HOL-blocking , and then terminate. Thus, it would be desirable to allow the session

related to that family to be terminated while leaving other AFI/SAFI unaffected. As BGP is commonly deployed, this is not possible.

Multisession BGP[3] was try to transport the AFI/SAFI over multiple session, but this is not a efficiency way. If BGP-4 message transport over SCTP, we can easily use SCTP-Multi-Streaming feature to avoid the HOL-Blocking.

Multi-streaming is used in transport layer, that means on application layer, BGP-4 will only see one SCTP-association to the peer node, but actually the message transport is over many streaming tunnel.

BGP-4 multi-streaming transport over SCTP as follows:

```
      _____________                                     _____________
     |    BGP-4    |                                   |    BGP-4    |
     | Application |                                   | Application |
     |-------------|                                   |-------------|
     |    SCTP     |<-------------Stream 1------------->|    SCTP     |
     |  Transport  |<-------------Stream 2------------->|  Transport  |
     |   Service   |                ....               |   Service   |
     |             |<-------------Stream N------------->|             |
     |-------------|                                   |-------------|
     |             |One or more    ----      One or more|             |
     | IP Network  |IP address      \/        IP address| IP Network  |
     |   Service   |appearances     /\       appearances|   Service   |
     |_____________|               ----                |_____________|

       SCTP Node A |<-------- Network transport ------->| SCTP Node B
```

2.1. Classify Route information

When BGP-4 support SCTP-multi-streaming, we need a way to distinguish the information/message to different streams. it can be classify by the following method:

```
   --  Classify by AFI/SAFI
   --  Classify by AS_PATH
   --  Classify by Route Distinguisher(RD)
```

The following format MUST be used for the SCTP DATA chunk:

```
  0                   1                   2                   3
  0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1 2 3 4 5 6 7 8 9 0 1
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |   Type = 0    | Reserved|U|B|E|    Length                     |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
 |                              TSN                              |
 +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

```
       |      Stream Identifier S      |   Stream Sequence Number n    |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       |                  Payload Protocol Identifier                  |
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
       \                                                               \
       /                 User Data (seq n of Stream S)                 /
       \                                                               \
       +-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+-+
```

In SCTP DATA chunk format , "Stream Identifier S" field is 16 bits unsigned integer, Thus it will support 65535 streams over a single SCTP Association. A hash algorithm is needed to classify the message as follows:

```
                                   _____________
                _____________     |             |---Stream 1---->
               | Hash based  |    |    SCTP     |---Stream 2---->
               | Classifier  |--->|  Transport  |    ....
               |_____________|    |   Service   |
                                  |_____________|---Stream N---->
```

The receiver will simply ignore the stream id.

2.2. Classification Analysis

2.2.1. Classify by AFI/SAFI

Classifier will use the AFI/SAFI as a Hash source data. But if one Router mark all AFI/SAFI with malformed community or other attribute, that will cause all the Streaming Queue blocked.

2.2.2. Classify by AS_PATH

Classifier will use the FIRST and LAST AS Number in AS_PATH Sequence as a Hash source data. this mechanism will avoid the HOL-blocking scenario describe in 2.2.1. but it may require 2-Level hash classifier as follows:

```
                                             _____________
      _____________      ______________     |             |-Stream 1--->
     | Src.AS Hash |    | Dest.AS Hash |    |    SCTP     |-Stream 2--->
     | Classifier  |--->| Classifier   |--->|  Transport  |    ....
     |_____________|    |______________|    |   Service   |
                                            |_____________|-Stream N--->
```

2.2.3. Classify by Route Distinguisher(RD)

Classifier will use the Route Distinguisher(RD) as a Hash source data. This mechanism can avoid large UPDATE message in some VPN. all the malformed messages from VPN will send in a single Streaming follows, that will not leaving other VPN unaffected.

3. Using SCTP multihoming for BGP connection

There's an article[4] to support BGP multihoming via TCP. Multihoming is a desired feature to enhance BGP redundancy and Reliability. Using SCTP multihoming feature is much more reasonable than multihoming over TCP.

3.1. BGP link via TCP limitation

Using multihoming over TCP will has some limitations, In this scenario, We often use a loopback interface as update source to avoid single link failure. But in eBGP multihops scenario as shown below:

```
            _____________        _________       ______________
           |             |--a1--|IP Cloud1|--b1--|              |
           | RtrA(lo 0)  |      |_________|      | RtrB(lo 0)   |
           | eBGP to RtrB|       _________       | eBGP to RtrA |
           |_____________|--a2--|IP Cloud2|--b2--|______________|
                                |_________|
```

RtrA use interface loopback 0 to establish a TCP sessions to RtrB's interface loopback 0 across a IP cloud, If link b1 down, RtrA will detect the link failure after the IP Cloud1 IGP convergence.
If RtrA run BFD can detect the link failure faster , then RtrA will advertise peer RtrC lost. but RtrA still can use link a2 communicate with RtrB.

This is caused by only one TCP sessions between two Routers.
Neighbor recover-time is depends on IGP convergence speed. When the link recover, the neighbor will be established again. The update message will be transmitted to all networks again. Which will cause the route flapping and networks instability.

3.2. Which link need BGP multihoming

SCTP provides transparent support for communications between two endpoints of which one or both is multi-homed.

iBGP link often has only 1 hop to the peering node, Thus will detect the link failure much faster. It will not require to establish multihoming, only use SCTP link via two Router's loopback interface is enough. But using SCTP transport is required to enhance the transport reliability, In iBGP to RR connections, SACK will increase the RR's performance. and multistreaming will avoid HOL-Blocking.

eBGP link connect to another AS, Inter-AS is not very stable and also will congestion in some time period. Establish a backup link to the peering node is necessary.

3.3. Init multihomging link for BGP connection

SCTP association need determine Primary Address , We can use link load, reliability, bandwidth as preference value , also we can use a pre-configured value as preference value.

eBGP multihoming link betweeen 2 Routers shown as below:

```
     _________________________             _______________________
    |                         |           | AS Y                  |
    |     __________          |           |          __________   |
    |    |          |--ip.a1--+-----------+-ip.b1---|          |  |
    |    |   RtrA   |         |           |         |   RtrB   |  |
    |    |__________|--ip.a2--+-----------+-ip.b2---|__________|  |
    | AS X                    |           |                       |
    |_________________________|           |_______________________|
```

SCTP multihoming can also init to different Router, but it will require RtrA config a route transmit packet to ip.b2 via link *ip.a2--ip.c2* as the follows:

```
     _________________________             _______________________
    |                         |           | AS Y                  |
    |     __________          |           |          __________   |
    |    |          |         |           |         |  RtrB    |  |
    |    |          |--ip.a1--+-----------+-ip.b1---|   (ip.b2)|  |
    |    |          |         |           |         |_____|____|  |
    |    |   RtrA   |         |           |               |       |
    |    |          |         |           |          _____|____   |
    |    |          |         |           |         |     |    |  |
    |    |          |--ip.a2--+-----------+-ip.c2---|   RtrC   |  |
    |    |__________|         |           |         |__________|  |
    | AS X                    |           |                       |
    |_________________________|           |_______________________|
```

3.4. Link failure detection and switchover procedure

SCTP provides monitoring of the reachability of the addresses on the remote endpoint and in the case of failure can transparently failover from the primary address to an alternate address, without upper layer intervention.

But in BGP-4 Multihoming implementation, when primary link failed We MUST notify the RIB/FIB to forwarding other packets to the alternate link. A withdraw a message need to send out.

4. BGP-4 Stack modification to support SCTP

BGP-4 transport over SCTP need to modify the BGP-4 Stack, the key terms as below:

-- modify neighbor FSM to init the SCTP link and also gives a backward capability to fallback TCP connections.

-- modify BGP Capability Advertisement to support SCTP Multistreaming transportations method.

-- modify NOTIFICATION Subcodes to notify the neighbor that failed to init SCTP connections or Primary/Alternate link failure.

4.1. Neighbor connection FSM modification

There are 2 Status added by support BGP-4 over SCTP:

o CONNECT-SCTP
o CONNECT-TCP

When BGP-4 process start, Neighbor status change from IDLE to CONNECT-SCTP. In this step, BGP speaker try to init SCTP connection to the peering node. Add SCTP-ConnectRetry Timer to monitor SCTP connections. If this timer expire, BGP will retry to init SCTP connecections.

If the SCTP-ConnectRetry Timer expire again, BGP-4 will fallback to init a TCP connection , and FSM change from CONNECT-SCTP to CONNECT-TCP. and a NOTIFICATION message will send out later to notice the remote peer that an error occur when init SCTP connection and fallback to TCP.

If still timeout, neighbor status will change to ACTIVE status. Then BGP Speaker listen on the configured interface.

If SCTP/TCP link successful established , OPEN message will send out and the neighbor status will change to OPENSENT.

4.2. New BGP Capability Advertisement

This specification defines SCTP transport capability:

```
Capability code (1 octet): TBD (Wants to reserve 69)
Capability length (1 octet): fixed 2bits
Capability value (2 bits):
  0 -- Do not use Multistreaming
  1 -- Use MultiStreaming and classify by AFI/SAFI
  2 -- Use MultiStreaming and classify by AS_PATH
  3 -- Use MultiStreaming and classify by Route Distinguisher(RD)
```

4.3. New NOTIFICATION Subcodes

This specification introduces three new subcodes:

- o TBD -- Init SCTP association failed, fallback to TCP connection.
- o TBD -- Primary SCTP link failure.
- o TBD -- Alternate SCTP link failure.

5. Security Considerations

from RFC3257:

"SCTP has been designed with the experiences made with TCP in mind. To make it hard for blind attackers (i.e., attackers that are not man-in-the-middle) to inject forged SCTP datagrams into existing associations, each side of an SCTP association uses a 32 bit value called "Verification Tag" to ensure that a datagram really belongs to the existing association. So in addition to a combination of source and destination transport addresses that belong to an established association, a valid SCTP datagram must also have the correct tag to be accepted by the recipient.

Unlike in TCP, usage of cookie in association establishment is made mandatory in SCTP. For the server, a new association is fully established after three messages (containing INIT, INIT-ACK, COOKIE-ECHO chunks) have been exchanged. The cookie is a variable length parameter that contains all relevant data to initialize the TCB on the server side, plus a HMAC used to secure it. This HMAC (MD5 as per [RFC1321] or SHA-1 [SHA1]) is computed over the cookie and a secret, server-owned key."

6. IANA Considerations

This document defines a new BGP capability - BGP transport over SCTP Capability. The Capability Code for BGP transport over SCTP Capability is TBD(Wants to reserve 69). currently used capability-codes as below:

http://www.iana.org/assignments/capability-codes/

7. References

7.1. Normative References

[1] Coene, L., "Stream Control Transmission Protocol Applicability Statement", RFC 3257, April 2002.

[2] M. Tim Jones , Better networking with SCTP: the Stream Control Transmission Protocol combines advantages from both TCP and UDP

[3] John, G Scudder., Chandra, Appanna. "Multisession BGP" draft-ietf-idr-bgp-multisession-03.txt, January 2007

[4] Philip, S. and Gaurab, U, "BGP Multihoming and Internet Exchange Points", SANOG 7. http://www.sanog.org/resources /sanog7/pfs-bgp-multihoming.pdf

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, March 1997.

[RFC4271] Rekhter, Y., Li, T., and S. Hares, "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, January 2006.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, January 2007.

7.2. Informative References

[BFD] Katz, D. and D. Ward, "Bidirectional Forwarding Detection", Work in Progress.

Authors' Addresses

Kevin Fang
Cisco Systems, Inc.
Edge Routing Business Unit

EMail: zhiyfang&cisco.com

Feng Cai
Cisco Systems, Inc.
Edge Routing Business Unit

EMail: fecai&cisco.com