INTERNET-DRAFT
Internet Engineering Task Force
Issued: June 2003
Expires: December 2003

L. Coene
Siemens

Multihoming issues in the Stream Control Transmission Protocol

<draft-coene-sctp-multihome-04.txt>

Status of this Memo

Abstract

This document describes issues of the Stream Control Transmission Protocol (SCTP)[RFC2960] in regard to multihoming on the Internet. It explores cases where through situations in the internet, single points-of-failure can occur even when using multihoming and what the impact is of multihoming on the routing tables of the host.

Table of Contents

1 Introduction

SCTP[RFC2960] is a transport protocol that uses multihoming. This draft is an attempt at identifying the issues that may arise in the layers below SCTP when multihoming is used by SCTP. Some issues are already being addresses in various other WGs and this document will try to highlight them. If the solutions would resolve the issues presented in this document then the problem presented is no longer an issue. This document will also try to gauge the effectiveness of the present multihoming architecture.

1.1 Terminology

The following terms are commonly identified in this work:

Association: SCTP connection between two endpoints.

Transport address: A combination of IP address and SCTP port number.

Multihoming: Assigning more than one IP network interface to a

single endpoint.

TLA: Top Level Aggregation

1.2 Contributors

The following people contributed to the document: L. Coene(Editor), M. Tuexen, G. Verwimp, J. Loughney, R.R. Stewart, Qiaobing Xie, M. Holdrege, M.C. Belinchon, A. Jungmaier, and L. Ong.

2 SCTP multihoming

2.1 Architecture of SCTP multihoming

A single message transmitted over an SCTP association from the originating host to the destination host will be sent using a single destination IP address chosen from the set of destination IP addresses available for that association. The paths used by the IP packets across the network might be different depending on the destination IP address. If a message fails to reach its destination, SCTP may retransmit the message using a different destination IP address.

SCTP does not have any way to determine whether two paths share links and routers when traversing the network.

The route of a path through a network can be static(example manual configuration) or dynamic(via routing protocols such as OSPF, BGP...). The route that a path takes through the network will change over time according to the routing protocols or routing decisions employed by the IP network layer.

If somewhere along the path a link or/and router fails then SCTP can detect the failure of the path via a heartbeat message(or in the worst case by the halting of the regular traffic). These actions done by SCTP are independant of the actions performed by the lower layers for failure detection and restoration and might de done in different timescales.

2.2 Interaction with routing

For fault resilient communication between two SCTP endpoints, the multihoming feature needs more than one IP network interface for each endpoint. The number of paths used is the minimum of network interfaces used by any of the endpoints. It is recommended to bind the association to all the IP source addresses of the endpoint.

Eeach network interface can have more than one IP address.

Under the assumption that every IP address will have a different, seperate path towards the remote endpoint, (this is the

responsibility of the routing protocols or of manual configuration) , if the transport to one of the IP addresses (= 1 particular path) fails then the traffic can migrate to the other remaining IP address (= other paths) within the SCTP association.

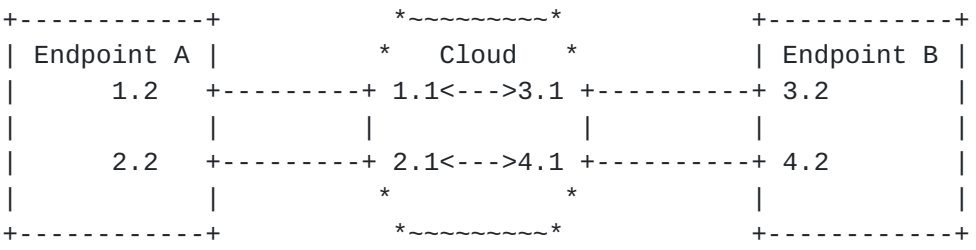

Figure 2.1.1: Two hosts with redundant networks.

Consider figure 2.1.1, if the host routing tables look as follows the endpoint will achieve maximum use of the multi-homing feature:

Endpoint A		Endpoint B	
Destination	Gateway	Destination	Gateway
3.0	1.1	1.0	3.1
4.0	2.1	2.0	4.1

Now if you consider figure 2.1.1, if the host routing table looks as follows, the association is subject to a single point of failure in that if any interface breaks, the whole association will break(See figure 2.1.2).

Host A		Host B	
Destination	Gateway	Destination	Gateway
3.0	1.1	1.0	4.1
4.0	2.1	2.0	3.1

Example: link 4.2-4.1 fails

Primary path: link 1.2-1.1 - link 3.1-3.2
Second Path : Link 2.2-2.1 - link 4.1-4.2

Endpoint A

```
+-------+--------+------+
|S= 1.2 | D= 3.2 | DATA |  ------->----- Arrives at Endpoint B
+-------+--------+------+
```

Endpoint B answers with SACK

```
    +-------+--------+------+
    |S= 4.2 | D= 1.2 | SACK | Gets lost, because send out on the failed
    +-------+--------+------+  4.1-4.2 link
```

After X time, retransmit on the other path by endpoint A

```
Endpoint A
+-------+--------+------+
|S= 2.2 | D= 4.2 | DATA | Is send out on link 2.2-2.1, but gets lost,
+-------+--------+------+ as msg has to pass via failed  4.1-4.2 link
```

The same scenario will play out for failures on the other links

Note : S = Source address
D = Destination address

Figure 2.1.2: Single point of failure case in redundant network due to routing table in host B

When an endpoint selects its source address, careful consideration must be taken. If the same source address is always used, then it is possible that the endpoint will be subject to the same single point of failure illustrated above. If possible the endpoint should always select the source address of the packet to correspond to the IP address of the Network interface where the packet will be emitted.

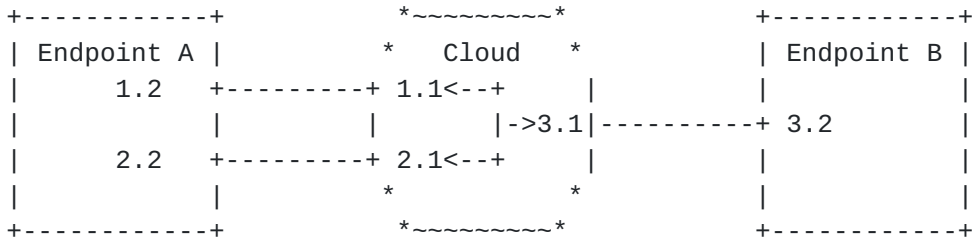

Figure 2.1.3: Two hosts with asymmetric networks.

In Figure 2.1.3 consider the following host routing table:

Endpoint A		Endpoint B	
Destination	Gateway	Destination	Gateway
3.0	1.1	1.0	3.1
		2.0	3.1

In this case the fault tolerance becomes limited by two seperate

issues. If the path between 3.1 and 3.2 breaks in both directions any association will break between endpoint A and endpoint B. The second failure will occur for the whole the association as well due to a breakage between 1.2 and 1.1 in both directions, since no

alternative route exists to 3.2 and all traffic is being routed through one interface.

Now one of these issues can be remedied by the following: In Figure 2.1.3 consider the following host routing table:

Endpoint A		Endpoint B	
Destination	Gateway	Destination	Gateway
3.0	1.1	1.0	3.1
3.0	2.1	2.0	3.1

The SCTP implementation on Endpoint A needs to consider both the source and destination address when identifying the transport. If it treats 1.2/3.2 as a separate transport address from 2.2/3.2, normal SCTP failover mechanisms work correctly. This mechanism is suggested in RFC 2960.

If the underlying OS does not support multiple routes to the same destination, or if the SCTP implementation can not control which route gets used (as is typical for user space implementations), we can still solve this issue by the following modification even when only one interface exists on endpoint B.

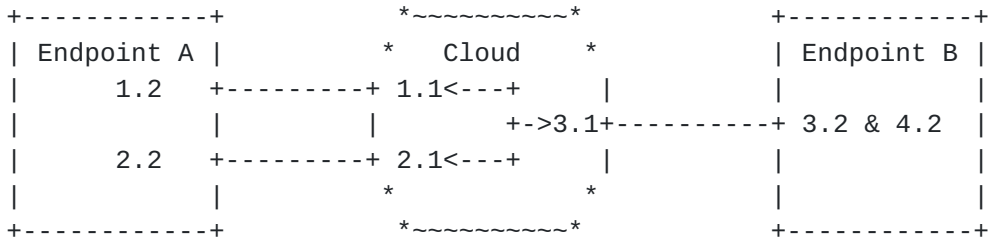

Figure 2.1.4: Two hosts with asymmetric networks, but symmetric addresses.

In Figure 2.1.4 consider the following host routing table:

Endpoint A		Endpoint B	
Destination	Gateway	Destination	Gateway
3.0	1.1	1.0	3.1
4.0	2.1	2.0	3.1

Now with the duplicate IP addresses assigned to the same interface

and the above routing tables, even if the interface between 1.1 and 1.2 breaks, an association will still survive this failure.

As a practical matter, it is recommended that IP addresses in a

multihomed endpoint be assigned IP endpoints from different TLA's to ensure against network failure.

In IP implementations the outgoing interface of multihomed hosts is often determined by the destination IP address. The mapping is done by a lookup in a routing table maintained by the operating system. Therefore the outgoing interface is not determined by SCTP. Using such implementations, it should be noted that a multihomed host cannot make use of the multiple local IP addresses if the peer is singlehomed. The multihomed host has only one path and will normally use only one of its interfaces to send the SCTP datagrams to the peer. If this physical path fails, the IP routing table in the multihome host has to be changed. This problem is out of scope for SCTP.

SCTP will always send its traffic to a certain transport address (= destination address + port number combination) for as long as the transmission is uninterrupted (= primary). The other transport addresses (secondary paths) will act as a backup in case the primary path goes out of service. The changeover between primary and backup will occur without packet loss and is completely transparent to the application. The secondary path can also be used for retransmissions(per section 6.4 of [RFC2960]).

The port number is the same for all transport addresses of that specific association.

Applications directly using SCTP may choose to control the multihoming service themselves. The applications have then to supply the specific IP address to SCTP for each outbound user message. This might be done for reasons of load-sharing and load-balancing across the different paths. This might not be advisable as the throughput of any of the paths is not known in advance and constantly changes due to the actions of other associations and transport protocols along that particular path, would require very tight feedback of each of the paths to the loadsharing functions of the user.

By sending a heartbeat chunk/message (=SCTP internal keep alive message) on all the multiple paths that are not used for active transmission of messages across the association, it is possible for SCTP to detect whether one or more paths have failed. SCTP will not use these failed paths when a changeover is required.

The transmission rate of sending heartbeat messages should be modifiable and the possible loss of the heartbeat message could be used for the monitoring and measurements of the concerned paths.

2.3 SCTP multihoming and the size of routing tables

As multihoming means that more than one destination address is used on the host, that would mean that a routing descision must be made

on the host in IP. The host does not know beforehand to which other host it is going to send something, so that would in theory require that all possible paths to all possible destinations should be known on that host. This amounts to the host being a part of the distribution of the routing information in the network.

Possible solutions would require to ask only for the paths to host that are actually in use(meaning a association is about to be setup with that particular host). This is a viable solution for hosts with a small number of associations to different hosts.

If the host has many associations with a lot of different host then then this becomes cumbersome(getting the specific paths from the routers and the updates and all) and leads in practice to same problem of distributing prefixes from the edge router(s) to the host.

It might be useful to explore ways where no distribution of routing information to the host for using multihoming is needed or where the interface/link selection is not based on the use of different prefixes. Not all hosts have facilities for containing possible large routing tables/databases.

2.4 SCTP multihoming and Network Adress Translators(NAT)

For multihoming the NAT must have a public IP address for each represented internal IP address. The host can preconfigure IP address that the NAT can substitute. Or the NAT can have internal Application Layer Gateway (ALG) which will intelligently translate the IP addresses in the INIT and INIT ACK chunks. See Figure 1.

If Network Address Port Translation is used with a multihomed SCTP endpoint, then any port translation must be applied on a per-association basis such that an SCTP endpoint continues to receive the same port number for all messages within a given association.

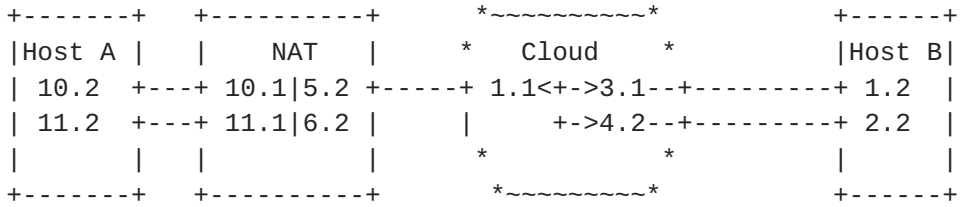

Fig 1: SCTP through NAT with multihoming

It should be noted that the NAT box becomes a single point-of-failure in this case, as ALL the paths of the SCTP

association have to go through that single NAT box.

2.5 SCTP multihoming and IPsec

IPsec was not designed to support multihomed connections and that imposes some difficulties when using IPsec with SCTP hosts that make use of several addresses in a single association.

The Security Policy Database (SPD) entries should use as selectors, among other fields of the IP header, the source and destination address. The easy and expensive way to scale the SPD to multihomed host, is simply creating a new entry for all the possible combinations of source and a destination addresses. A much better implementation approach is to simply use groups of addresses instead of single ones.

The same problem arises when identifying a Security Association (SA). An SA should be identified by the extended triplet ({set of destination addresses}, Security Parameter Index, Security Protocol).

Moreover, when exchanging keys using the Internet Key Exchange (IKE) protocol there are some extra difficulties. IKE only allows the use of a single source and destination address, and so the initial solution to the problem would be creating a number S*D of SAs, where S is the number of source addresses, and D the number of destination addresses. This solution unnecessarily consumes both time and resources. Other more complex and suitable approaches would need modifications in IKE itself.

A specific discussion about the problems that crop up when using IPsec with SCTP can be seen in [IPsec-SCTP]. All SCTP+IPSEC implementations would have to do the above to be compatible

3 Security considerations

SCTP only tries to increase the availability of a network. SCTP does not contain any protocol mechanisms which are directly related to user message authentication, integrity and confidentiality functions. For such features, it depends on the IPSEC protocols and architecture and/or on security features of its user protocols.

The solutions needed for allowing multihoming may provide security

risks.

4 References and related work

[RFC2960] Stewart, R. R., Xie, Q., Morneault, K., Sharp, C. , , Schwarzbauer, H. J., Taylor, T., Rytina, I., Kalla, M., Zhang, L. and Paxson, V."Stream Control Transmission Protocol", RFC2960, October 2000.

[RFC2663] Srisuresh, P. and Holdrege, M., "IP Network Address Translator (NAT) Terminology and Considerations", RFC2663, August 1999

[RFC2694] Srisuresh, P., Tsirtsis, G., Akkiraju, P. and Heffernan, A., "DNS extensions to Network Address Translators (DNS_ALG)", RFC2694, September 1999

[IPsec-SCTP] Bellovin, S. M., Ioannidis, J., Keromytis, A. D. and Stewart, R. R., "On the Use of SCTP with IPsec", draft-ietf-ipsec-sctp-06.txt, work in progress

5 Acknowledgments

This document was initially developed by a design team consisting of Lode Coene, John Loughney, Michel Tuexen, Randall R. Stewart, Qiaobing Xie, Matt Holdrege, Maria-Carmen Belinchon, Andreas Jungmaier, Gery Verwimp and Lyndon Ong.

The authors wish to thank Renee Revis, I. Rytina, H.J. Schwarzbauer, J.P. Martin-Flatin, T. Taylor, G. Sidebottom, K. Morneault, T. George, M. Stillman, N. Makinae, S. Bradner, A. Mankin, G. Camarillo, H. Schulzrinne, R. Kantola, J. Rosenberg, I. Arias-Rodriguez, D. Lehmann, La Monte Henry Yaroll, P. Savola, H. Alvestrand and many others for their invaluable comments.

6 Author's Address

The following authors have contributed to this document.

Lode Coene Phone: +32-14-252081
Siemens Atea EMail: lode.coene@siemens.com
Atealaan 34
B-2200 Herentals
Belgium

Expires: December 31, 2003

removing the copyright notice or references to the Internet Society or other Internet organizations, except as needed for the purpose of