INTERNET-DRAFT
Internet Engineering Task Force
Issued: December 2000
Expires: June 2001

L. Coene
M. Tuexen
G. Verwimp
Siemens
J. Loughney
Nokia
R.R. Stewart
Cisco
Qiaobing Xie
Motorola
M. Holdrege
Ipverse
M.C. Belinchon
Ericsson
A. Jungmayer
University of Essen

Stream Control Transmission Protocol Applicability Statement
<draft-ietf-sigtran-sctp-applicability-03.txt>

Status of this Memo

Abstract

This document describes the applicability of the Stream Control Transmission Protocol (SCTP)[RFC2960] for general usage in the Internet. This document describes the key features of SCTP and how they are used for general purpose data transport.

TABLE OF CONTENTS

1 Introduction

1.1 Terminology

The following terms are commonly identified in related work:

Association: SCTP connection between two endpoints.

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

Upper layer: The user of the SCTP protocol, which may be an adaptation layer, a session layer protocol, or the user application directly.

1.2 Protocol Overview

The Stream Control Transmission Protocol (SCTP) provides a reliable transport between two endpoints.

The following functions are provided by SCTP:

- Reliable Data Transfer
- Multiple streams to help avoid head-of-line blocking
- Ordered and unordered data delivery on a per-stream basis
- Bundling and fragmentation of user data
- Congestion and flow control
- Support continuous monitoring of reachability
- Graceful termination of association
- Support of multi-homing for added reliability
- Protection against blind denial-of-service attacks
- Protection against blind masquerade attacks

2 Applicability of the Stream Control Transmission Protocol -- SCTP

This section describes where you could use SCTP for transporting application data.

- SCTP can be used as a general-purpose transport protocol for message-oriented applications. Message boundaries are preserved during data transport and so no message delineation is needed. The user data can be delivered by the order of transmission within a stream or the order of arrival.

- For data streams that lack boundaries or markers, the application must force an arbitrary boundary for the data it sends to SCTP. The choice of the boundary should result in a reasonable frame size based on the network MTU.

- SCTP can be used to provide redundancy and fault tolerance at the transport layer and below. Applications needing this level of fault tolerance can make use of SCTP's multi-homing support.

- SCTP can be used as the transport protocol for applications where head-of-line blocking is a concern. Such an application should use multiple streams to provide independent ordering of user messages.

- SCTP can be used as the transport protocol for applications where the average size of the user messages is small. For such applications, the bundling feature of SCTP will combine multiple small messages for efficient bandwidth utilization.

- SCTP can be used as the transport protocol for applications where the average size of the user messages is large. For such applications, SCTP will transparently fragment large user messages to conform to the appropriate packet size.

- SCTP can be used as the transport protocol for applications where multiple message streams need to be multiplexed over a single association. For such applications, SCTP will transparently combine messages from multiple streams at the transmission and demultiplex upon reception.

3 Issues affecting deployment of SCTP

3.1 SCTP multihoming and interaction with routing

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

Under the assumption that every IP address will have a different, seperate paths towards the remote endpoint, (this is the responsibility of the routing protocols or of manual configuration) , if the transport to one of the IP address (= 1 particular path) fails then the traffic can migrate to the other remaining IP address (= other paths) within the SCTP association.

```
+------------+           *~~~~~~~~~~*           +------------+
| Endpoint A |          *   Cloud   *           | Endpoint B |
|     1.2    +---------+ 1.1<--->3.1 +----------+ 3.2        |
|            |         |             |          |            |
|     2.2    +---------+ 2.1<--->4.1 +----------+ 4.2        |
|            |          *           *           |            |
+------------+           *~~~~~~~~~~*           +------------+
```

Figure 3.1.1: Two hosts with redundant networks.

Consider figure 3.1.1, 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 3.1.1, 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 3.1.2).

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 3.1.2: Single point of failure case in redundant network.

```
  +------------+           *~~~~~~~~~~*           +------------+
  | Endpoint A |          *   Cloud    *          | Endpoint B |
  |     1.2    +---------+ 1.1<--+      |          |            |
  |            |         |       |->3.1|----------+ 3.2        |
  |     2.2    +---------+ 2.1<--+      |          |            |
  |            |          *            *          |            |
  +------------+           *~~~~~~~~~~*           +------------+
```

Figure 3.1.3: Two hosts with asymmetric networks.

In Figure 3.1.3 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 modification even when only one interface exists on endpoint B.

```
  +------------+           *~~~~~~~~~~*            +------------+
  | Endpoint A |          *   Cloud    *           | Endpoint B |
  |      1.2   +---------+ 1.1<---+     |          |            |
  |            |         |        +->3.1+----------+ 3.2 & 4.2  |
  |      2.2   +---------+ 2.1<---+     |          |            |
  |            |          *            *           |            |
  +------------+           *~~~~~~~~~~*            +------------+
```

Figure 3.1.4: Two hosts with asymmetric networks, but symmetric addresses.

In Figure 3.1.4 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 TLV's to ensure against network failure.

In IP implementations the outgoing interface of multihomed hosts is ofter 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. Something which is out of scope of 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 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 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 keep alive message should be modifiable and the possible loss of keep alive message could be used for the monitoring and measurements of the concerned paths.

3.2 Use of SCTP in Network Address Translators (NAT) Networks [RFC2663]

When a NAT is present between two endpoints, the endpoint that is behind the NAT, i.e., one that does not have a publicly available network address, shall take one of the following options:

(1) When single homed sessions are to be used, no transport addresses should be sent in the INIT or INIT ACK chunk(Refer to section 3.3 of RFC2960 for chunk definitions). This will force the endpoint that receives this initiation message to consider the sender as only having that one address. This method can be used for a NAT, but any multi-homing configuration at the endpoint that is behind the NAT will not be visible to its peer, and thus not be taken advantage of. See figure 3.2.1.

```
       +-------+  +---------+      *~~~~~~~~~~*           +------+
       |Host A |  |   NAT   |     *   Cloud    *          |Host B|
       | 10.2  +--|10.1|2.1 |----|--------------|---------+ 1.2  |
       |       |  |    |    |     *            *          |      |
       +-------+  +---------+      *~~~~~~~~~~*           +------+
```

Fig 3.2.1: SCTP through NAT without multihoming

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 3.2.2.

```
  +--------+   +-----------+      *~~~~~~~~~~~*          +-------+
  |Host A |    |    NAT    |     *    Cloud    *          |Host B|
  | 10.2  +---+ 10.1|5.2 +------+ 1.1<+->3.1--+----------+ 1.2  |
  | 11.2  +---+ 11.1|6.2 |     |     +->4.2--+----------+ 2.2  |
  |       |   |          |      *           *          |      |
  +--------+   +-----------+       *~~~~~~~~~~*          +-------+
```

Fig 3.2.2: SCTP through NAT with multihoming

(2) Another alternative is to use the hostname feature and DNS to resolve the addresses. The hostname is included in the INIT of the association or in the INIT ACK. The hostname must be resolved by DNS before the association is completely set up. There are special issues regarding NAT and DNS, refer to RFC2694 for details.

4 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.

Mechanisms for reducing the risk of blind denial-of-service attacks and masquerade attacks are built into SCTP protocol. See RFC2960, section 11 for detailed information.

Currently the IPSEC working group is investigating the support of multihoming by IPSEC protocols. At the present time to use IPSEC, one must use 2 * N * M security associations if one endpoint uses N addresses and the other M addresses.

5 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.

[RFC2401] Kent, S., and Atkinson, R., "Security Architecture for the Internet Protocol", RFC 2401, November 1998.

[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

6 Acknowledgments

The authors wish to thank Renee Revis, I. Rytina, L. Ong, H.J. Schwarzbauer, J.P. Martin-Flatin, T. Taylor, G. Sidebottom, K. Morneault, T. George, M. Stillman and many others for their invaluable comments.

7 Author's Address

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

John Loughney   Phone: +358-9-43761
Nokia Research Center   EMail: john.loughney@nokia.com
Itamerenkatu 11-13
FIN-00180   Helsinki
Finland

Michel Tuexen   Phone: +49-89-722-47210
Siemens AG   EMail: Michael.Tuexen@icn.siemens.de
Hofmannstr. 51
81359 Munich

Germany

Randall R. Stewart   Phone: +1-815-477-2127
24 Burning Bush Trail.   EMail: rrs@cisco.com
Crystal Lake, IL 60012

USA

Qiaobing Xie   Phone: +1-847-632-3028
Motorola, Inc.   EMail: qxie1@email.mot.com
1501 W. Shure Drive
Arlington Heights, IL 60004
USA

Matt Holdrege   Phone: +1-408-830-3239
Ipverse   Email: matt@ipverse.com
223 Ximeno Avenue
Long Beach, CA 90803-1616
USA

Maria-Carmen Belinchon   Phone: +34-91-339-3535
Ericsson Espana S. A.   EMail: Maria.C.Belinchon@ericsson.com
Network Communication Services
Retama 7, 5th floor
Madrid, 28045
Spain

Andreas Jungmayer   Phone: +49-201-1837636
University of Essen   EMail: ajung@exp-math.uni-essen.de
Institute for experimental Mathematics
Ellernstrasse 29
D-45326 Essen
Germany

Gery Verwimp   Phone: +32-14-253424
Siemens Atea   EMail: gery.verwimp@siemens.atea.be
Atealaan 34
B-2200   Herentals
Belgium

Expires: June 30, 2001

Full Copyright Statement