Understanding Congestion and SCTP Multihoming with Cisco ITP and Mobile Vendor MSC

Troubleshooting Guide

Contents

1. Congestion Problem Analysis

1.1 Congestion Problem

The service provider customer has experienced several problems related to congestion happening when flapping in the IP network. This congestion has been linked to drops in the Mobile Network.

We have seen the congestion in the log analysis:

```
December 27 17:15:47.093: %CS7XUA-5-SCTPCONGESTONSET: ASP(MICAMC01-IPB) Level 0 ->
Level 1
December 27 17:15:48.333: %CS7XUA-5-SCTPCONGESTONSET: ASP(MICAMC01-IPB) Level 1 ->
Level 2
December 27 17:15:49.517: %CS7XUA-5-SCTPCONGESTONSET: ASP(MICAMC01-IPB) Level 2 ->
Level 3
…
December 27 17:15:50.353: %CS7XUA-5-SCTPCONGESTONSET: ASP(MICAMC01-IPA) Level 0 ->
Level 1
December 27 17:15:50.397: %CS7XUA-5-SCTPCONGESTABATE: ASP(MICAMC01-IPA) Level 1 ->
Level 0
```

We have discovered the congestion symptoms using the following command-line interface (CLI) commands:

1. Increasing retransmissions in the SCTP association indicate a problem:

   SI01ITP#SHOW IP SCTP ASSO PARA 0x1E0201FF

   …

   Num retrans: 6371 Max retrans: 2 Num times failed: 44

   10.176.112.197 retrans: 7 10.176.80.197 retrans: 0

2. Signaling Congestion (SCON) is being sent by Cisco IP Transfer Point (ITP):

   SI202ITP#SHOW CS7 ASP STA DETAIL

   …

   Cong 0 SCONs Sent: 0 **Cong 1 SCONs Sent: 236**

   Cong 2 SCONs Sent: 0 Cong 3 SCONs Sent: 0

   Inbound SSNM to SS7 Stats

3. There is an increase in Retx and Fast Retx together with a high water over the 500 threshold mark:

   SI01ITP#show ip sctp asso stat 0x1D0300A2

   …

   Transmit-queue-depth

   Current: 22 **High Water Mark: 633**

   DataGrams Sent: 453605 DataGrams Rcvd: 258333

   RexmitTO: 148 RexmitFAST: 41875

   SI01ITP#show ip sctp asso stat 0x1D0300A2

   ** SCTP Association Statistics AssocId:0x1D0300A2 **

   Transmit-queue-depth

   Current: 29 **High Water Mark: 633**

   DataGrams Sent: 466114 DataGrams Rcvd: 265824

```
RexmitTO: 148 RexmitFAST: 42534      <--- Increasing Fast Retx
```

By default, with tx-queue-depth = 1000, here are thresholds for each congestion level:

```
Thresholds for congestion on transmit queue
  Level 1 onset:        500   50% of Tx Q         Level 1 abate:             300
  Level 2 onset:        700    70% of Tx Q        Level 2 abate:             500
  Level 3 onset:        900     90% of Tx Q      Level 3 abate:             700
 Level 4 onset:        1000   100% of Tx Q        Level 4 abate:             900
```

If we are over the 500 mark, we will have SCON level 1, and over the 1000 mark, we will have SCON level 4. We can see below SCON 4 with High Water 1494:

```
SI202ITP#
SI202ITP#show ip sctp asso stat  0x1D0202AE

** SCTP Association Statistics AssocId:0x1D0202AE **
...
Transmit-queue-depth
  Current: 8  High Water Mark: 1494   <--- High Water over 1000, SCON level 4
DataGrams Sent: 2150147  DataGrams Rcvd: 1312332
RexmitTO: 0 RexmitFAST: 0

SI202ITP#show ip sctp asso parame  0x1D0202AE
…
Min RTO: 250 ms  Max RTO: 500 ms
LocalRwnd: 64000  Low: 62632   RemoteRwnd: 31356  Low: 116
Congest levels: 4  current level: 0  high mark: 1494  chkSum: crc32


SI202ITP#show cs7 asp na IMPA4-SLI01 STAT det
…
  ERR Inv Network App:     0            ERR Missing Parm:       44

  ERR Unsupported Type:    0            ERR Traffic Mode:       0
  ERR Unexpected Msg:      12           ERR Protocol Error:     0
  …

  ERR Inv Network App:     0            ERR Missing Parm:       1
  …
  NOTIFY-Alt ASP Active:   0            NOTIFY-ASP Failure:     12
…
  DRSTs Sent:              0            DUPUs Sent:             1
  Cong 0 SCONs Sent:       0            Cong 1 SCONs Sent:      212
  Cong 2 SCONs Sent:       0            Cong 3 SCONs Sent:      0
        Inbound SSNM to SS7 Stats
…
  ERR Pkts Dropped:        0            DUNA Pkts Dropped:      0
  DAVA Pkts Dropped:       0            SCON Pkts Dropped:      26
  DUPU Pkts Dropped:       0            DRST Pkts Dropped:      0
```

```
  Pkts Dropped by VIP:     0
  Level 1 Congestion Cnt:  20          Level 2 Congestion Cnt: 11
  Level 3 Congestion Cnt:  32          Level 4 Congestion Cnt: 1
```

1.2 Congestion Analysis

To avoid packet drops in the Mobile Network, the service provider customer has shut down the ASPs that showed problems in Site 1 and Site 2.

```
SI01ITP: SIGTRAN association in shutdown: IMNA5-SLI01 / IHNA2-SLI01 / ISNA1-SLI01

SI202ITP: SIGTRAN association in shutdown: IMPA4-SLI23
```

The congestion happened during the daytime hours of several consecutive days after IP flapping, but not during the hours around midnight.

1. We have verified the Stream Control Transmission Protocol (SCTP) settings for M3UA in ITP in associations with the problems:

SI01ITP

```
ASP name: IMNA5-SLI01                          Type: M3UA
…
Local receive window :   64000          Cumulative sack timeout:   100 ms
Assoc retrans:           5              Path retrans:              2
Max init retrans:        8              Max init RTO:              1000  ms
Minimum RTO:             250   ms       Maximum RTO:               500   ms
Bundle status:           on             Bundle timeout:            5 ms
Keep alive status:       true           Keep alive timeout:        30000 ms
SCTP congestion level:   0              SCON congestion level:     0
Unordered priority:      equal          Transmit queue depth:   1000
Initial cwnd:            3000           Idle cwnd rate:            50
Retrans cwnd rate:       50             Retrans cwnd mode:         RFC
FastRetrans cwnd rate:   50
Thresholds for congestion on transmit queue
  Level 1 onset:         500            Level 1 abate:             300
  Level 2 onset:         700            Level 2 abate:             500
  Level 3 onset:         900            Level 3 abate:             700
  Level 4 onset:         1000           Level 4 abate:             900
…
```

2. ITP and end nodes (MSC, HLR) are set with different SCTP parameters. A mismatch will cause unnecessary retransmission, which will cause congestion during high-volume traffic.

   In Table 1 we have the ITP SCTP settings for M3UA

Table 1. ITP SCTP Parameters

M3UA Parameter	ITP Default Value	ITP Service Provider Customer
path-retransmit	4	2
assoc-retransmit	10	5
RTO min (ms)	1000	250
RTO max (ms)	1000	500

M3UA Parameter	ITP Default Value	ITP Service Provider Customer
bundling (ms)	5	5
keepalive (ms)	30000	30000
cumulative-sack (ms)	200	100
tx-queue-depth (packets)	1000	1000
init-retransmit	8	8
init-timeout (ms)	1000	1000
receive-window (bytes)	64000	64000
max-inbound-streams	17	17
unordered-priority	equal	equal

Table 2 gives the SCTP MSC settings.

Table 2. MSC SCTP Parameters

Parameter	Value	Default	Description		
RTOMI	250	100	Minimum retransmission timeout		
RTOMA	500	250	Maximum retransmission timeout		
RTOI	300	150	Initial retransmission timeout		
BUF	256	128	Size of the sending buffer for the association		
THR	192	96	Sending buffer threshold		
ARW	65535	8192	Advertised receiver window		
	32768	8192	Advertised receiver window (RPB-E setting)		
TSACK	100	40	SACK delay		
IMR	8	8	Initial retransmission attempts		
MTU	1452	1416	Maximum transmission unit		
PMR	4	2	Path maximum retransmission		
AMR	8	2	Association maximum retransmission		
MIS	17	256	Maximum number of incoming streams for an association		
MOS	17	256	Maximum number of outgoing streams for an association		
AICL	30	30	Allowed increment for the cookie life span		
HBI	30	30	Heartbeat interval		
HBS	1	1	Heartbeat status		
KCP	4	4	Key change period		
LBS	1	1	Local bundling status		
RTOA	3	3	Retransmission timeout alpha index		
RTOB	2	2	Retransmission timeout beta index		
SMR	5	5	Maximum number of retransmission during shutdown of association		
VCL	60	60	Valid cookie life span		

Table 3 shows the main differences in the SCTP parameters.

Table 3. ITP MSC SCTP Parameters Mismatch

Cisco ITP Parameters	ITP Default Value	ITP Customer settings	MSC Customer settings	Mobile Vendor Parameters
path-retransmit	4	**2**	**4**	PMR
assoc-retransmit	10	**5**	**8**	AMR
bundling (ms)	5	**5**	**1**	LBS
tx-queue-depth (packets)	1000	**1000**	**256**	BUF
init-timeout (ms)	1000	**1000**	**300**	RTOI
receive-window (bytes)	64000	64000	65535/32768	ARW

2. Congestion Action Plan

After the analysis, we have identified two problems:

1. There are different settings in SCTP for M3UA in ITP and MSC, HLR.
2. There is a multiplicative effect once there is flapping, with congestion if there is high traffic in the network related to the lack of tuning in the SCTP queue.

The following action plan has been set in place:

1. Understanding why we have different SCTP settings with SCTP multihoming in ITP and MSC.
2. Matching the same parameters on both sides, mainly tx-queue-depth and values related to the queue that creates the congestion.

Step 1. Understanding the Congestion Reaction in MSC

The MSC with congestion management reacts to the SCON received from the ITP, limiting the traffic coming from the RNC to avoid the increase in congestion in the SIGTRAN network, and this implies dropping traffic from end users.

Step 2. Understanding Congestion in ITP

Congestion is triggered by either a packet timeout or fast retransmission that can be caused by IP network instability, flapping, bandwidth issues, and so on. As we can see in the previous sections, there is an increment of both in the specific associations with congestion:

```
    Current: 22  High Water Mark: 633
DataGrams Sent: 453605  DataGrams Rcvd: 258333
RexmitTO: 148 RexmitFAST: 41875
```

Once that happens, the transmit window is decreased. Depending on the rate of the traffic, this can start to cause congestion. The ITP will send SCON messages when the SCTP transmit queue is congested, as we have indicated previously:

Thresholds for congestion on transmit queue

```
    Level 1 onset:        500   50% of Tx Q        Level 1 abate:            300
    Level 2 onset:        700    70% of Tx Q       Level 2 abate:            500
    Level 3 onset:        900     90% of Tx Q     Level 3 abate:            700
   Level 4 onset:        1000   100% of Tx Q       Level 4 abate:            900
```

The MSC has a 31356 byte receive-window. If that window starts to fill, the MSC could have trouble keeping up and responding to packets. With a 250 ms minimum retransmission timeout (RTOmin), it is possible for the ITP to have to retransmit packets. Once a retransmission occurs, the transmit window decreases and, depending on the rate of the traffic, congestion could occur.

Following is a detailed explanation of the congestion control mechanism in ITP.

SCTP employs congestion control algorithms to adjust the amount of unacknowledged data that can be injected into the network and to retransmit segments dropped by the network. The SCTP congestion control algorithms respond to packet loss as an indication of network congestion. Packet loss detected by SCTP congestion control algorithms can put the sender in slow-start with a reduced congestion window, thereby limiting the amount of data that can be transmitted. The slow-start algorithm will force the sender to wait for an acknowledgment before transmitting new data. The slow-start and congestion control algorithms can force poor utilization of the available channel bandwidth when using long delay networks.

SCTP congestion control uses two state variables to accomplish congestion control. The first variable is the congestion window (cwnd). The congestion window is an upper bound on the amount of data the sender can inject into the network before receiving an acknowledgment. The second variable is the slow-start threshold (ssthresh). The slow-start threshold variable determines which algorithm is used to increase cwnd. If cwnd is less than or equal to ssthresh, the slow-start algorithm is used to increase cwnd.

If cwnd is greater than ssthresh the congestion avoidance algorithm is used to increase cwnd. There are two methods of packet-loss detection (interpreted as congestion notification by the SCTP congestion controls) defined in SCTP:

- Timeout of the retransmission timer. The congestion control algorithms resets the congestion control state variables cwnd and ssthresh. The setting of the congestion control state variables have the effect of putting the sender in slow start and assure that no more than one packet is outstanding until it receives an acknowledgment.

  ```
  ssthresh = max (cwnd/2, 2*MTU)
  cwnd = 1*MTU
  ```

- Detection of gaps in received Transmission Sequence Numbers (TSNs) through Gap Ack reports in a Selective Acknowledgment (SACK). Normally a sender will wait four consecutive Gap Ack reports before reacting to the indication of packet loss. The congestion control algorithms reset the congestion control state variables cwnd and ssthresh as a result of detecting the packet loss. The setting of the congestion control variables will put the sender in slow-start with a reduced cwnd effectively limiting the amount of data the sender can transmit.

  ```
  sshtresh = max(cwnd/2, 2*MTU)
  cwnd = ssthresh
  ```

Another aspect important to note is the relation between queue and SCTP association. In the ITP, each SCTP association has a unique code that is not shared per SCTP port, as we can see in the following parameters.

We can change the queue per ASP:

```
SI01ITP(config)#cs7 asp ICMI3-NODE1 2919 8101 m3ua
SI01ITP(config-cs7-asp)#tx-queue-depth?
  <100-20000>  queue depth in packets
  <cr>
```

If we take the example of local SCTP in Site 1 8101, with three different ASPs (each of them implies one SCTP association):

```
ICMI3-NODE1   ICMI3         active    M3UA  2919     10.176.72.129   0x1D0300BB
cs7 asp ICMI3-NODE1 2919 8101 m3ua

ICRM3-NODE1   ICRM3         active    M3UA  2907     10.176.98.129   0x1D0300BA
cs7 asp ICRM3-NODE1 2907 8101 m3ua

ICRM4-NODE1   ICRM4         active    M3UA  2907     10.176.98.137   0x1D0300B9
cs7 asp ICRM4-NODE1 2907 8101 m3ua
```

We can see each association with a different mark in the queue for each association:

```
SI01ITP#show ip sctp asso stat  0x1D0300BB
```

```
** SCTP Association Statistics AssocId:0x1D0300BB **
…
Transmit-queue-depth
  Current: 2  High Water Mark: 35
DataGrams Sent: 18347349  DataGrams Rcvd: 12028436
RexmitTO: 6 RexmitFAST: 1


SI01ITP#show ip sctp asso stat  0x1D0300BA
…
Transmit-queue-depth
  Current: 1  High Water Mark: 302
DataGrams Sent: 33363641  DataGrams Rcvd: 13441660
RexmitTO: 1 RexmitFAST: 6


SI01ITP#show ip sctp asso stat  0x1D0300B9

** SCTP Association Statistics AssocId:0x1D0300B9 **

…
Transmit-queue-depth
  Current: 0  High Water Mark: 23
DataGrams Sent: 9979604  DataGrams Rcvd: 1837680
RexmitTO: 252 RexmitFAST: 1
```

Step 3. Understanding SCTP Multihoming in ITP

SCTP provides several protocol parameters that can be customized by the upper layer protocol. These protocol parameters can be customized to control and influence SCTP performance behavior. Different network designs and implementations pose their own unique performance requirements. It is not possible to provide customized protocol parameters that are suitable for all implementations. The tuning information in this step is provided as a guide for understanding what the SCTP protocol parameters are and how they affect the various SCTP algorithms.

Connection Establishment

The protocol parameters assoc-retransmit, init-retransmit, and init-timeout can be customized to control connection establishment. During SCTP association initialization, sometimes packet retransmissions occur. The first initialization packet timeout occurs after 1 second. When initialization packet retransmissions occur, the timeout value is doubled for each retransmission. The maximum timeout value is bound by the init-timeout parameter. The init-timeout parameter is used to control the time between initialization packet retries. As a general rule, init-timeout should be configured to reflect the round-trip time for packets to traverse the network. An init-timeout value that is too small can cause excessive retries of initialization packets. Large init-timeout values can increase connection establishment times.

The number of retries allowed for connection establishment packets is controlled by the init-retransmit protocol parameter. When selecting the number of retries, the number of attempts should take into account varying network conditions that may prevent initialization packets from traversing the network.

The defaults used by M3UA/M2PA are recommendations from RFC 2960. The init-timeout default is 1 second. The init-retransmit default is set for 8. The init-retransmit and init-timeout defaults are suitable for most high-speed links. The defaults may require adjusting for slower links.

Here is an example of connection establishment:

Send an INIT:

```
src:A1:INIT(A1, A2)---->to:(B1)
```

The INIT ack comes back:

```
<-------to:A1-INIT-ACK(B1,B2) src:B1
```

At this stage, each side selects a primary destination. How they do this depends on the implementation. This means that unless the application does a “setprimary” socket option to assure that the primary is where they expect (on each side, by the way) then there is no assurance that this won’t occur.

Now the implementation in the ITP and also in the MSC and HLR (most of the implementation but not all) will make the first listed address in the INIT the primary; in the above example it would be A1.

But we can see implementations that are listed in another order:

```
A1:INIT(A1, A2)---->(to either address)   is valid
A1:INIT(A2, A1)---->(to either address)   is valid
```

SCTP Multihoming

A key feature of SCTP is multihoming. An SCTP endpoint is considered multihomed if more than one IP address can be used as a destination to reach that endpoint. Upon failure of the primary destination address SCTP switches to an alternate address.

In the configuration of a multihomed endpoint, the first remote IP address specified on the peer link is defined as the primary address. If the primary address is determined to be unreachable, SCTP multihoming switches to one of the alternate addresses specified on the peer link. SCTP will monitor the reachability of the failed destination address. Upon notification that reachability is reestablished to the primary address, M3UA/M2PA directs SCTP to switch back to the primary address.

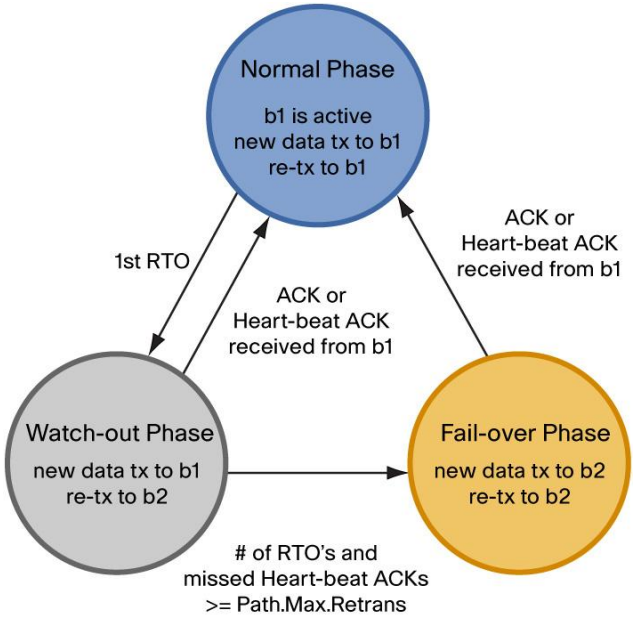
SCTP sends the data to the primary address. If timeout occurs, it resends the data to an alternate address. All new data will still be sent to the primary address. If a timeout occurs again, it resends the data to the alternate address. It continues this process until path-retransmit is reached on the primary address. Once path-retransmit is reached on the primary address, it marks the primary address unreachable and sends all data to the alternate address. It begins heartbeats on the primary address. When the heartbeat is successful, it marks the primary address reachable and starts sending data to the primary address again

The protocol parameters path-retransmit and retransmit-timeout can be customized to control how long SCTP waits before switching to an alternate address. The path-retransmit parameter controls the number of times that SCTP attempts to retransmit a packet before declaring the destination address unreachable.

The retransmit-timeout parameter is used to determine whether a packet must be retransmitted. If an acknowledgement is not received by the time the retransmission timer expires, all packets that have been transmitted, but not acknowledged, are retransmitted.

Figure 1 shows the state transition diagram governing the three-phase failover process between the primary and alternate path of the SCTP association. Under the “Normal” phase, new TPDUs are transmitted to the primary address of the destined association endpoint, that is, from A1 to B1, unless the SCTP user explicitly specifies the destination (and possibly) the source IP address to use. Upon a retransmission timeout (due to primary path loss/error), the association enters the “Watch-out” phase, in which the source attempts to select the most divergent source-destination pair, that is, A2-B2, from the original primary path to perform the retransmission. New incoming data are still being sent along the primary path during the “Watch-out” phase. In the meantime, the source endpoint will closely monitor the status of the primary path by (1) sending periodic heartbeat messages (with exponential backoff) to the primary destination address a1 while keeping track of the number of RTOs and missing heartbeat acknowledgements. If the sum of these two counts exceeds a so-called Path.Max.Retrans threshold, the primary path will be declared to be inactive, and the association enters the third “Failover,” which continues until the positive acknowledgements of data or heartbeat messages is received from the original primary destination address.

Figure 1. SCTP Association State Transition



Here is an example of SCTP multihoming to describe the algorithm from the local ITP side (round-robin scheme per destination):

```
local-peer 5000
  local-ip A1
  local-ip A2

link 0 sctp B1 B2 5000 5000
```

By default, the primary destination address is B1.

Assume there is no data traffic, so the heartbeat (HB) timer is running and expires for destination B1.

ITP will send a HB from A1-B1.

If ITP don’t get a HB-Ack then ITP keeps track of this HB failure.

When the HB timer expires for B2, ITP sends a HB from A2-B2; if ITP doesn’t get a HB-Ack, then ITP keeps track of this HB failure.

When the HB timer expires for B1 again, ITP sends a HB from A2-B1, also tracking the success or failure. When the HB timer expires for B2, ITP sends a HB from A1-B2.

ITP continues this round-robin of the source addresses with the destination addresses and monitors the success/failure.

When data traffic starts on the primary destination, ITP uses the information from monitoring the heartbeats and data acknowledgements to help determine which path, A1-B1 or A2-B1, is successful. The path that has been the most successful is the path that is selected. While the primary path is available, ITP continues to round-robin the source addresses in the HB for destination B2 on the alternate path and to monitor the success/failure.

In the service provider customer implementation, we have implemented in the IP network two paths instead of the four paths that are possible for simplicity purposes, so ITP with the HB mechanism keep tracks of the two possible paths (A1-B1 and A2-B2), marking the paths A1-B2 and B2-A1 as unreachable.

Following is an explanation for each SCTP parameter:

Path-retransmit

The path-retransmit parameter is the number of packet retries before the destination address is deemed unreachable (independently of the source). The number of path-retransmits multiplied by the retransmission timer ultimately controls how fast an alternate address becomes the primary path for multihomed nodes. This relationship suggests the RTO parameters and path-retransmit parameter should be considered together. Configuring the default RTO values and the default path retransmit value of 4 allows a multihomed node to switch to an alternate destination address within 4 seconds.

Retransmit-timeout

The RTO should be adjusted for round-trip delay between nodes. Preferably, the retransmission timeout should be greater than the round-trip delay between nodes. There will be a compromise between allowing a long delay and having responsive discovery of lost frames. We can calculate a simplistic estimate of round-trip times (RTT) for various packet sizes (ignoring propagation delay and latencies in transmission equipment) using the following estimated RTT equation:

```
estimated RTT =  ( ( packet size * bits per byte )  /  link speed) * 2
```

Assume a packet with a 20-byte IP header, 32-byte SCTP header, and 100 bytes of user data and a 1,544,000 bits/sec link between two nodes. Using the estimated RTT equation shown in the previous paragraph, we estimate an RTT of 1.5 ms.

SCTP computes RTO values based on RTT measurements. When packet retransmission occurs, the timeout value is doubled for each retransmission, with an upper limit of max RTO. Multihomed nodes will have to compromise between allowing a long delay and having responsive switching to an alternate IP address. Switching to an alternate path is of primary importance for multihomed nodes. The maximum RTO value for multihomed nodes should be set equal to or just slightly higher than the minimum RTO value. The number of outstanding bytes allowed decreases with each retransmission timeout. The trade-off of bounding the maximum RTO close to the minimum RTO is the frequency of retransmissions versus increasing transmission delays for packets on the transmit queue. During periods of retransmissions multihomed nodes send duplicate packets until the alternate address becomes the primary path. The alternate address becomes the primary when the number of retries exceeds the path-retransmit parameter. The default value for minimum and maximum RTO is 1 second. Propagation delays and latencies vary in networks, so care should be taken when selecting an RTO value.

Bundling

It is recommended that bundling be enabled for high packet rates (1000 pps or higher) with small packets (50 bytes and lower). Bundling can be less than optimal for lower data rates with small or large packets because of the transmission delay. Bundling is found to be effective for large packets at high data rates in networks with symmetrical traffic. The default bundling delay is 5 ms. Applications with low data rates should disable bundling if the increase in round-trip time is undesirable. It is recommended that bundling be enabled for applications sending small packets that may start with low data rates but are capable of increasing to higher sustained data rates.

Cumulative Selective Ack

The cumulative selective ack (cs-ack) is commonly known as “delayed ack.” The cs-ack parameter controls how long a receiver can delay before sending an acknowledgement. The ack is delayed hoping to have data going in the same direction as the ack, so the ack can “piggyback” with the data. The default of cs-ack is 200 ms. The cs-ack configured at the receiver must be less than the RTO minimum value configured at the sender. When the cs-ack of the receiver is greater than the RTO of the sender, unnecessary retransmissions may occur because the sender RTO expires before the receiver sends the delayed acknowledgment.

Receive-window

The size of the receive-window offered by the receiver generally can affect performance. SCTP adapts its transmission rate to suit the available network capacity by using a congestion-sensitive, sliding-window flow control mechanism described in RFC 2581. At any given instance only a certain number of bytes can be outstanding through the network. Keeping the path full of packets requires both congestion window (cwnd) and receive-window (rwnd) to reach the effective size of the “pipe” represented by the so-called bandwidth-delay product. We can calculate the capacity of the pipe using the following capacity equation:

```
capacity (bits) = bandwidth (bits/sec) x round-trip-time(sec)
```

The bandwidth-delay product can vary widely depending on the network speed and round-trip time between the two endpoints. Using the capacity equation shown in the previous paragraph, we can estimate the minimum buffer size given the bandwidth of the communication media and the round-trip time between the nodes. Assuming the nodes are connected by a 1,544,000 bits/sec T1 link with a round-trip time of 60 ms gives an estimated minimum buffer size of 11,580 bytes. The receive-window parameter default is set for 64,000 bytes. The congestion control and window ing algorithms adjust to network conditions by controlling the number of bytes that can be outstanding through the network.

Transmit Queue

The tx-queue-depth parameter is used to determine the onset and abate thresholds for congestion on the transmit queue. The tx-queue-depth parameter controls the number of packets allowed on the transmit queue. The tx-queue exists to absorb inevitable traffic bursts. When selecting the tx-queue-depth, there will be a compromise between hitting transmit congestion thresholds causing dropped packets and transmit delays due to queuing times. Applications that are sensitive to small delays should account for transmit delays due to queuing when selecting a tx-queue-depth. During periods of SCTP link congestion, the tx-queue-depth will control the number of packets that can be queued before packets are discarded, causing application retransmissions. The default tx-queue-depth is 1000 packets for M3UA, M2PA, and SUA. The default tx-queue-depth is 20,000 packets for SGMP.

Step 4. Understanding SCTP Multihoming in Mobile Vendor MSC

The information comes from the Mobile Vendor team regarding the general multihomed scenario where A1 and A2 are local IP addresses of MSC and B1 and B2 are remote IP addresses of ITP.

Initial Association Establishment.

The first phase is initial association establishment: it is always executed toward one remote IP address and can be repeated for a specified number of times (application parameter MAXofINITIALRetransmit). In case of a missing answer to the first attempt, more attempts are performed by changing the local IP address for each new attempt.

So, considering the above general scenario and assuming that B1 is the first remote IP address on ITP that is going to be attempted, the sequence is:

1. A1 -> B1
2. A2 -> B1
3. A1 -> B1
4. A2 -> B1 … and so on, until MAXofINITIALRetransmit expires.

At the first successful answer, the association is established.

In the case that B1 is not answering, B2 is attempted, with the same rotation of source addresses as follows:

1. A1 -> B2
2. A2 -> B2
3. A1 -> B2
4. A2 -> B2 … and so on, until MAXofINITIALRetransmit expires or B2 answers.

If the SCTP association is finally established, a primary SCTP path is established and the reconfiguration is such that the primary source IP address is the first IP address defined in the endpoint definition.

Normal Traffic Handling

An endpoint always transmits, by default, to the primary path. When the remote peer is multihomed, an endpoint tries to retransmit a timeout chunk to an active destination transport address that is different from the last destination address to which the data chunk was sent.

The scheme used is:

- For the next retransmission of this packet, a new remote peer address is selected (by rotation) out of the available remote peer addresses.
- For the old remote peer address (on which transmission failed) a new local source address is selected (by rotation) to improve the chance of successful transmission next time this remote address is selected.

Now, let’s look at the observed disturbance happened in customer network.

Suppose that B1 on ITP is unreachable from MSC but B2 is reachable: according to the actual settings of PMR (PathMaxRetrans = 4) and AMR (AssocMaxRetrans = 8), and considering that the cross-secondary path doesn’t work, the retransmission should be attempted on the secondary paths in the following order:

1. A1 -> B1 (transmission of DATA message 1 on primary path -> timeout 250 msec -> retransmission starts)
2. A1 -> B2 (retransmission of DATA message 1 on secondary path -> first timeout 250 msec for B2 -> first timeout for association)
3. A2 -> B1 (retransmission of DATA message 1 on secondary path -> first timeout 250 msec for B1 -> second timeout for association)
4. A2 -> B2 (retransmission of DATA message 1 on secondary path -> successful retransmission -> counter for association reset)
5. A1 -> B1 (transmission of DATA message 2 on primary path -> timeout -> second timeout 250 msec for B1 -> first timeout for association)

6. A1 -> B2 (retransmission of DATA message 2 on secondary path -> first timeout 250 msec for B2 -> second timeout for association)
7. A2 -> B1 (retransmission of DATA message 2 on secondary path -> third timeout 250 msec for B1 -> third timeout for association)
8. A2 -> B2 (retransmission of DATA message 2 on secondary path -> successful retransmission -> counter for association reset)
9. A1 -> B1 (transmission of DATA message 3 on primary path -> timeout -> fourth timeout 250 msec for B1 -> first timeout for association -> PATH A1-B1 DOWN)

Effectively, primary path A1-B1 is marked DOWN after 1750 msecs.

The availability of B1 is then verified with heartbeats.

When B1 is reachable again (by heartbeats or association restart by remote peer) the sequence starts again. SCTP always resets the local IP to the primary local IP if the valid HEARTBEAT or HEARTBEAT_ACK chunks are received on the primary local address. This way SCTP switches to the local primary as soon as it discovers the address is alive and A1-B1 path is restored.

Step 5. SCTP Parameters Calculation and Fine-Tuning

We do see different parameters in ITP and MSC, HLR; this mismatch will be related to unnecessary SCTP retransmissions that will create congestion during flapping and high volume traffic. These parameters have to be tuned to adapt to the service provider customer topology, focusing on the tx-queue-depth calculation.

From the ITP side, here is the calculation of the tx-queue-depth:

Following the heartbeat procedure, the ITP determines the primary path for A1-B1 and the secondary path A2-B2, with A1-B2 and A2-B1 being unreachable. We keep sending HB through the secondary paths, but receive acknowledgement only from A2-B2.

Let’s imagine we have a failure in the path A1-B1. We have path-rtx:2 and assoc-rtx:5, RTOmin = 250 msec, RTOmax = 500 msec.

- Data message 1 is sent from A1–B1; we have a timeout (don´t rx the ack), timeout 1. RTO=RTOmin:250 msec
- Data message 1 is retransmitted with the second path IP, A2–B2, rtx successful.
- Data message 2 is sent from A1-B1, we have a timeout (don´t rx the ack), timeout 2. As we reached path-retransmit:2 and timeout:2, we declare the path-rtx:2 so we mark this path unavailable. RTO=2*RTOmin=500 msec.
- Data message 2 is retransmitted with the second path IP, A2–B2, rtx successful.
- All subsequent messages will be sent A2–B2 until the HB procedure determines that the path A1–B1 is reachable again.

As you can see, with these settings, the ITP will declare the path A1-B1 down after 750 msec, but the MSC will declare it later, around 250 + 500 * 3 = 1750 msec.

The values are different in the ITP and MSC. Following research with the service provider customer and Mobile Vendor, it has been indicated that MSC can´t set the PMR = 2 and AMR = 5, as the MSC use a two-way HB mechanism instead of the four-way HB mechanism implemented by Cisco.

When Mobile Vendor MSC receives the HB from the ITP, respond with the HB-ACK and mark this path available without sending a HB to the ITP for this path, if at that moment, we have a failure in this path; Mobile Vendor MSC will take longer to discover the path unreachability than the ITP, as the ITP will always send the HB and wait for the HB_ACK to mark it as available.

Conversely, the ITP using the four-way heartbeat mechanism can mark as unvailable the paths that don´t have connectivity, the cross paths, A1-B2 and A2-B1, but the MSC needs an iteration of the algorithm to realize that these paths are not available.

Here is information from Mobile Vendor regarding MSC:

Considering the IP selection mechanism in MSC, the safest couple of PMR/AMR parameters is 4/8. The current implementation of SCTP in MSC foresees selection of path A2-B2 only after rotating first the local and then the remote IP addresses.

```
A1------------------B1

A2 ------------------B2
```

In case of failure on B1, MSC is going to use the secondary “wired” path A2-B2 with PMR(B1) = 2 and AMR = 3. Any kind of further disturbances can cause the first path to become inactive with the association to go down.

For the fine-tuning, we consider several items:

- By lowering the RTOmin parameter, the failover times as well the maximum message delays can be further reduced. However, with very low values of RTOmin, associations may become more susceptible to early, unwanted retransmission timer timeouts, and thus retransmissions. With very low PRL values this may even result in use of the secondary path before any actual failure has occurred, so it is generally not recommended to lower the RTOmin parameter below RTOmin;rec = 2 _ RTT. These spurious timeouts must also be avoided since they have a negative effect on the protocol throughput.
- RTOmax can be different in each customer, as it is tunable to adapt to each IP/Multiprotocol Label Switching (MPLS) network. The rule of thumb is: Verify the RTT in the worst-case scenario, in the suboptimal path. Make the RTO significantly bigger than the experienced RTT.
- The trade-off of bounding the maximum RTO close to the minimum RTO is the frequency of retransmissions versus increasing transmit delays for packets on the transmit queue. We have normally deployed values of RTOmax < 2 x RTOmin or with equal values for both, taking into account the compromise between allowing a long delay and having responsive switching to an alternate IP address.
- Another aspect is the HB-TIMEOUT and RTOmax. If the heartbeat interval is too close to RTO values (retransmission timeout), that is, if MAX_RTO = 2000, HB_TIMEOUT = 2000, this means the time we wait for an appropriate heartbeat response (RTO) is too close to the heartbeat interval itself hb_timeout+RTO ms.We don’t want to hit the max_path_retrans too quickly and mark paths inactive by sending heartbeats too frequently and marking them unsuccessful within the RTO timeout just because the stack is busy processing other high-priority messages. So the heartbeat interval has to be set to a value a lot greater than the max_rto. In SP CUSTOMER ITP, we have MAX_RTO = 500, HB_TIMEOUT = 30000.

Step 6. Tx-Queue-Depth Calculation

If we consider the test performed in the service provider customer network, we will have:

- As we have seen, the RTTmax = 60 msec in the worst case scenario and, considering the SACK = 100 MSC, we will be more aggressive in the RTO, so we propose a RTOmin = 200 and RTOmax = 250.
- We will consider the limitations in the MSC values, PMR = 4, AMR = 8 to provide proper functionality for SCTP multihoming.
- The general formula is:

```
Association down time (msec) = (RTOmin) + (AMR-1) * (RTOmax if RTOmax < 2 *
RTOmin)) msec
```

```
Queue (MSU) = Association down (msec) * (MSU/sec)
```

- With these values, we will have the association down with (RTOmin:200) + 7 * (RTOmax:250) = 1950 msec. To avoid congestion level 1, with 1000MSU/sec, the queue will be double 1950 * 2 = 3900MSU.

3. M2PA Configuration and SCTP Settings

Below we explain how the different SCTP parameters for M2PA configuration work, following this example:

```
path-retransmit 4
assoc-retransmit 10
retransmit-timeout 250 500
cumulative sack 100
tx-queue-depth 2000
```

There is a fundamental relationship between the following link configuration parameters: path-retransmit, retransmit-timeout, and tx-queue-depth. We indicate again the definition of the parameters: Path retransmit controls how many times SCTP will retransmit a packet on a given network path. Retransmit timeout controls how long SCTP waits for an acknowledgement of an outstanding packet. These two parameters together control how long SCTP waits before it declares that a destination is unreachable. The tx-queue-depth parameter controls how many packets are permitted to be queued on the link transmit queue. Queue thresholds determine when link-level congestion procedures are executed.

As an example to show the relationship between the path-retransmit, retransmit-timeout, tx-queue-depth MSU rates and the number of links, consider two multihomed M2PA links using the settings shown above and a rate of 800 MSU/sec per link. With a double failure (failing both multihomed paths concurrently), it will take approximately 3 seconds for an M2PA link to fail. During the time it takes for the link to fail, the ITP will buffer about 3 seconds of packets at 800 MSU/sec, which comes to about 2400 packets. To avoid triggering link congestion procedures during changeover, a tx-queue-depth of about 2800-3000 packets would be required. Notice that as the MSU rate increases, the tx-queue-depth must be adjusted to handle the anticipated buffering requirements.

A key design point is that each FW processor is designed to normally handle up to .4 erlang of the stated capacity of a single FlexWan for the protocol and feature set. For M2PA MTP3 traffic with no GTT, this translates to a maximum of 2700 MSU/sec. Based on our understanding of the service provider customer production deployment of seven links per FlexWan, that puts approximately 386 MSU/sec per link. As another example we will use the recommended settings and 386 MSU/sec per link. In this example we also assume a double failure (failing both multihomed paths concurrently). It will take approximately 3 seconds for the M2PA link to fail. During changeover, we may have to buffer about 3 seconds of packets at 386 MSU/sec, which comes to be about 1200 packets. Again to avoid triggering congestion procedures during changeover, tx-queue-depth should be set higher. So the configured tx-queue-depth of 2000 would be sufficient.

The point here is to establish how the tx-queue-depth is influenced by the link failure time (path-retransmit and retransmit-timeout), the number of links, and the expected MSU/sec rate. Changing any one of these parameters requires a reevaluation to understand how the link changeover is affected.

4. Conclusion

In order to deploy and troubleshoot SIGTRAN application and devices, we will need to understand how the SIGTRAN stack behaves in each device and fine-tune each parameter to accommodate to the real IP next-generation network (NGN).

Americas Headquarters
Cisco Systems, Inc.
San Jose, CA

Asia Pacific Headquarters
Cisco Systems (USA) Pte. Ltd.
Singapore

Europe Headquarters
Cisco Systems International BV Amsterdam,
The Netherlands

Cisco has more than 200 offices worldwide. Addresses, phone numbers, and fax numbers are listed on the Cisco Website at **www.cisco.com/go/offices.**

Printed in USA C07-651862-00 04/11