Internet Engineering Task Force (IETF)
Request for Comments: 5811
Category: Standards Track
ISSN: 2070-1721

J. Hadi Salim
Mojatatu Networks
K. Ogawa
NTT Corporation

March 2010

SCTP-Based Transport Mapping Layer (TML) for the Forwarding and Control Element Separation (ForCES) Protocol

Abstract

This document defines the SCTP-based TML (Transport Mapping Layer) for the ForCES (Forwarding and Control Element Separation) protocol. It explains the rationale for choosing the SCTP (Stream Control Transmission Protocol) and also describes how this TML addresses all the requirements required by and the ForCES protocol.

Status of This Memo

This is an Internet Standards Track document.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Further information on Internet Standards is available in Section 2 of RFC 5741.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at http://www.rfc-editor.org/info/rfc5811.

Table of Contents

1. Introduction

The ForCES (Forwarding and Control Element Separation) working group in the IETF defines the architecture and protocol for separation of control elements (CEs) and forwarding elements (FEs) in network elements (NEs) such as routers. [RFC3654] and [RFC3746], respectively, define architectural and protocol requirements for the communication between CEs and FEs. The ForCES protocol layer specification [RFC5810] describes the protocol semantics and workings. The ForCES protocol layer operates on top of an inter-connect hiding layer known as the TML. The relationship is illustrated in Figure 1.

This document defines the SCTP-based TML for the ForCES protocol layer. It also addresses all the requirements for the TML including security, reliability, and etc., as defined in [RFC5810].

2. Definitions

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 [RFC2119].

The following definitions are taken from [RFC3654] and [RFC3746]:

LFB: Logical Functional Block. A template that represents a fine-grained, logically separate aspect of FE processing.

ForCES Protocol: The protocol used at the Fp reference point in the ForCES Framework in [RFC3746].

ForCES PL: ForCES Protocol Layer. A layer in the ForCES architecture that embodies the ForCES protocol and the state transfer mechanisms as defined in [RFC5810].

ForCES TML: ForCES Protocol Transport Mapping Layer. A layer in the ForCES protocol architecture that specifically addresses the protocol message transportation issues, such as how the protocol messages are mapped to different transport media (like SCTP, IP, TCP, UDP, ATM, Ethernet, etc.), and how to achieve and implement reliability, security, etc.

3. Protocol Framework Overview

The reader is referred to the Framework document [RFC3746], and in particular Sections 3 and 4, for an architectural overview and explanation of where and how the ForCES protocol fits in.

There is some content overlap between the ForCES protocol specification [RFC5810] and this section (Section 3) in order to provide basic context to the reader of this document.

The ForCES protocol layering constitutes two pieces, the PL and TML. This is depicted in Figure 1.

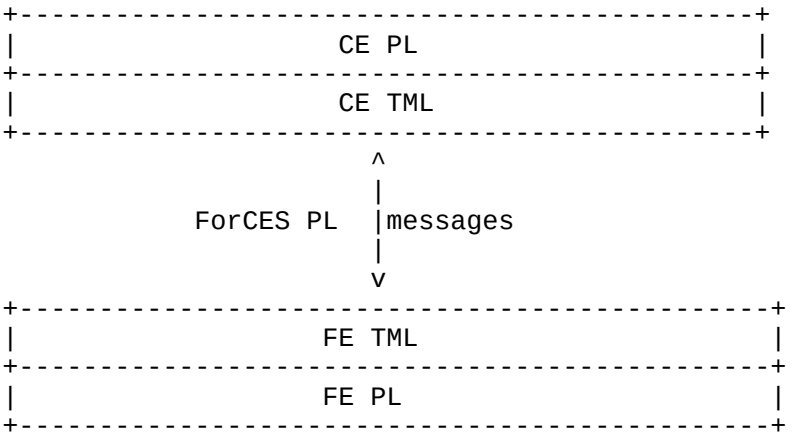

Figure 1: Message Exchange between CE and FE to Establish an NE Association

The PL is in charge of the ForCES protocol. Its semantics and message layout are defined in [RFC5810]. The TML is necessary to connect two ForCES endpoints as shown in Figure 1.

Both the PL and TML are standardized by the IETF. While only one PL is defined, different TMLs are expected to be standardized. The TML at each of the nodes (CE and FE) is expected to be of the same definition in order to inter-operate.

When transmitting from a ForCES endpoint, the PL delivers its messages to the TML. The TML then delivers the PL message to the destination TML(s).

On reception of a message, the TML delivers the message to its destination PL (as described in the ForCES header).

3.1. The PL

The PL is common to all implementations of ForCES and is standardized by the IETF [RFC5810]. The PL is responsible for associating an FE or CE to an NE. It is also responsible for tearing down such associations.

An FE may use the PL to asynchronously send packets to the CE. The FE may redirect various control protocol packets (e.g., OSPF, etc.) to the CE via the PL (from outside the NE). Additionally, the FE delivers various events that the CE has subscribed to via the PL [RFC5812].

The CE and FE may interact synchronously via the PL. The CE issues status requests to the FE and receives responses via the PL. The CE also configures the components of the associated FE's LFBs using the PL [RFC5812].

3.2. The TML

The TML is responsible for the transport of the PL messages. [RFC5810], Section 5 defines the requirements that need to be met by a TML specification. The SCTP TML specified in this document meets all the requirements specified in [RFC5810], Section 5. Section 4.2.2 of this document describes how the TML requirements are met.

3.2.1. TML and PL Interfaces

There are two interfaces to the PL and TML. The specification of these interfaces is out of scope for this document, but the interfaces are introduced to show how they fit into the architecture and summarize the function provided at the interfaces. The first interface is between the PL and TML and the other is the CE Manager (CEM)/FE Manager (FEM) [RFC3746] interface to both the PL and TML. Both interfaces are shown in Figure 2.

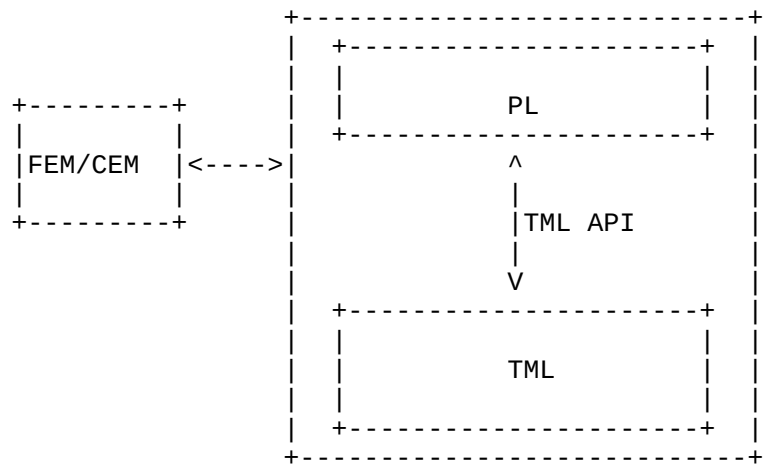

Figure 2: The TML-PL Interface

The CEM/FEM [RFC3746] interface is responsible for bootstrapping and parameterization of the TML. In its most basic form, the CEM/FEM interface takes the form of a simple static config file that is read on startup in the pre-association phase.

Appendix B discusses the service interfaces in more detail.

3.2.2. TML Parameterization

It is expected that it should be possible to use a configuration reference point, such as the FEM or the CEM, to configure the TML.

Some of the configured parameters may include:

- PL ID
- Connection Type and associated data. For example, if a TML uses IP/SCTP, then parameters such as SCTP ports and IP addresses need to be configured.
- Number of transport connections
- Connection Capability, such as bandwidth, etc.
- Allowed/Supported Connection Quality of Service (QoS) Policy (or Congestion Control Policy)

4. SCTP TML Overview

SCTP [RFC4960] is an end-to-end transport protocol that is equivalent to TCP, UDP, or DCCP in many aspects. With a few exceptions, SCTP can do most of what UDP, TCP, or DCCP can achieve. SCTP as can also do most of what a combination of the other transport protocols can achieve (e.g., TCP and DCCP or TCP and UDP).

Like TCP, it provides ordered, reliable, connection-oriented, flow-controlled, congestion-controlled data exchange. Unlike TCP, it does not provide byte streaming and instead provides message boundaries.

Like UDP, it can provide unreliable, unordered data exchange. Unlike UDP, it does not provide multicast support

Like DCCP, it can provide unreliable, ordered, congestion controlled, connection-oriented data exchange.

SCTP also provides other services that none of the three transport protocols mentioned above provide that we found attractive. These include:

- Multi-homing
- Runtime IP address binding
- A range of reliability shades with congestion control
- Built-in heartbeats
- Multi-streaming
- Message boundaries with reliability
- Improved SYN DOS protection
- Simpler transport events
- Simplified replicasting

4.1. Rationale for Using SCTP for TML

SCTP has all the features required to provide a robust TML. As a transport that is all-encompassing, it negates the need for having multiple transport protocols in order to satisfy the TML requirements ([RFC5810], Section 5). As a result, it allows for simpler coding and therefore reduces a lot of the interoperability concerns.

SCTP is also very mature and widely used, making it a good choice for ubiquitous deployment.

4.2. Meeting TML Requirements

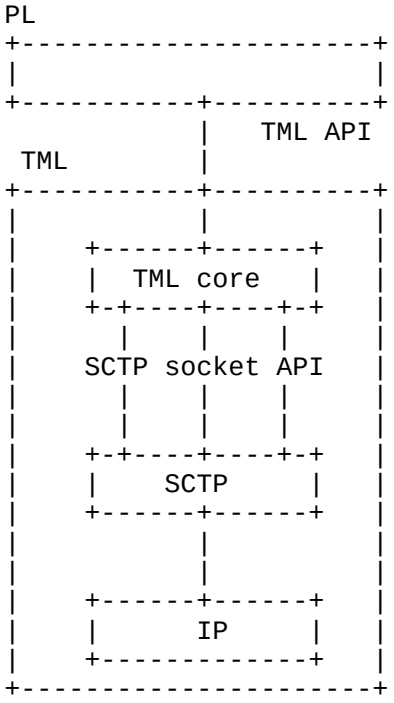

Figure 3: The TML-SCTP Interface

Figure 3 details the interfacing between the PL and SCTP TML and the internals of the SCTP TML. The core of the TML interacts on its northbound interface to the PL (utilizing the TML API). On the southbound interface, the TML core interfaces to the SCTP layer utilizing the standard socket interface [TSVWG-SCTPSOCKET]. There are three SCTP socket connections opened between any two PL endpoints (whether FE or CE).

4.2.1. SCTP TML Channels

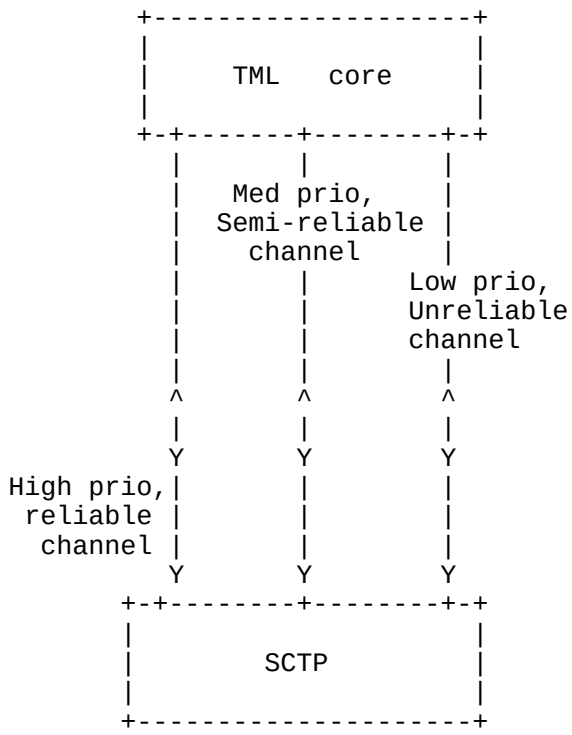

Figure 4: The TML-SCTP Channels

Figure 4 details further the interfacing between the TML core and SCTP layers. There are three channels used to group and prioritize the work for different types of ForCES traffic. Each channel constitutes an SCTP socket interface that has different properties. It should be noted that all SCTP channels are congestion aware (and for that reason that detail is left out of the description of the three channels). SCTP ports 6704, 6705, and 6706 are used for the higher-, medium-, and lower-priority channels, respectively. SCTP Payload Protocol ID (PPID) values of 21, 22, and 23 are used for the higher-, medium-, and lower-priority channels, respectively.

4.2.1.1. Justifying Choice of Three Sockets

SCTP allows up to 64 K streams to be sent over a single socket interface. The authors initially envisioned using a single socket for all three channels (mapping a channel to an SCTP stream). This simplifies programming of the TML as well as conserves use of SCTP ports.

Further analysis revealed head-of-line blocking issues with this initial approach. Lower-priority packets not needing reliable delivery could block higher-priority packets (needing reliable delivery) under congestion situations for an indeterminate period of time (depending on how many outstanding lower-priority packets are pending). For this reason, we elected to go with mapping each of the three channels to a different SCTP socket (instead of a different stream within a single socket).

4.2.1.2. Higher-Priority, Reliable Channel

The higher-priority (HP) channel uses a standard SCTP reliable socket on port 6704. SCTP PPID 21 is used for all messages on the HP channel. The HP channel is used for CE-solicited messages and their responses:

1. ForCES configuration messages flowing from CE to FE and responses from the FE to CE.

2. ForCES query messages flowing from CE to FE and responses from the FE to the CE.

PL priorities 4-7 MUST be used for all PL messages using this channel. The following PL messages MUST use the HP channel for transport:

- AssociationSetup (default priority: 7)

- AssociationSetupResponse (default priority: 7)

- AssociationTeardown (default priority: 7)

- Config (default priority: 4)

- ConfigResponse (default priority: 4)

- Query (default priority: 4)

- QueryResponse (default priority: 4)

If PL priorities outside of the specified range priority (4-7), PPID, or PL message types other than the above are received on the HP channel, then the PL message MUST be dropped.

Although an implementation may choose different values from the defined range (4-7), it is RECOMMENDED that default priorities be used. A response to a ForCES message MUST contain the same priority

as the request. For example, a config sent by the CE with priority 5 MUST have a config-response from the FE with priority 5.

4.2.1.3. Medium-Priority, Semi-Reliable Channel

The medium-priority (MP) channel uses SCTP-PR on port 6705. SCTP PPID 22 MUST be used for all messages on the MP channel. Time limits on how long a message is valid are set on each outgoing message. This channel is used for events from the FE to the CE that are obsoleted over time. Events that are accumulative in nature and are recoverable by the CE (by issuing a query to the FE) can tolerate lost events and therefore should use this channel. For example, a generated event that carries the value of a counter that is monotonically incrementing is fit to use this channel.

PL priority 3 MUST be used for PL messages on this channel. The following PL messages MUST use the MP channel for transport:

- Event Notification (default priority: 3)

If PL priorities outside of the specified priority, PPID, or PL message type other than the above are received on the MP channel, then the PL message MUST be dropped.

4.2.1.4. Lower-Priority, Unreliable Channel

The lower-priority (LP) channel uses SCTP port 6706. SCTP PPID 23 is used for all messages on the LP channel. The LP channel also MUST use SCTP-PR with lower timeout values than the MP channel. The reason an unreliable channel is used for redirect messages is to allow the control protocol at both the CE and its peer-endpoint to take charge of how the end-to-end semantics of the said control protocol's operations. For example:

1. Some control protocols are reliable in nature, therefore making this channel reliable introduces an extra layer of reliability that could be harmful. So any end-to-end retransmits will happen remotely.

2. Some control protocols may desire having obsolescence of messages over retransmissions; making this channel reliable contradicts that desire.

Given ForCES PL heartbeats are traffic sensitive, sending them over the LP channel also makes sense. If the other end is not processing other channels, it will eventually get heartbeats; and if it is busy processing other channels, heartbeats will be obsoleted locally over time (and it does not matter if they did not make it).

PL priorities 1-2 MUST be used for PL messages on this channel. PL messages that MUST use the MP channel for transport are:

- PacketRedirect (default priority: 2)
- Heartbeat (default priority: 1)

If PL priorities outside of the specified priority range, PPID, or PL message types other than the above are received on the LP channel, then the PL message MUST be dropped.

4.2.1.5. Scheduling of the Three Channels

In processing the sending and receiving of the PL messages, the TML core uses strict priority work-conserving scheduling, as shown in Figure 5.

This means that the HP messages are always processed first until there are no more left. The LP channel is processed only if channels that are a higher priority than itself have no messages left to process. This means that under a congestion situation, a higher-priority channel with sufficient messages that occupy the available bandwidth would starve lower-priority channel(s).

The design intent of the SCTP TML is to tie processing prioritization, as described in Section 4.2.1.1, and transport congestion control to provide implicit node congestion control. This is further detailed in Appendix A.2.

It should be emphasized that the work scheduling prioritization scheme prescribed in this document is receiver-based processing. Fully arrived packets on any of the channels are a source of work whose output may result in transmitted packets. However, we have no control on the order in which the SCTP/OS/network chooses to send transmitted packets across and make them available to the receiver. This is a limitation that we try to ameliorate by our choice of channel properties, ForCES message grouping, and the tying of CE and FE work scheduling. While that helps us ameliorate some of these issues, it does not fully resolve all.

From a ForCES perspective, we can tolerate some reordering. For example, if an FE transmits a config response (HP) followed by 10000 OSPF redirect packets (LP) and the CE gets 5 OSPF redirects (LP) first before the config response (HP), that is tolerable. What matters is the CE gets to processing the HP message soon (instead of sitting in long periods of time processing OSPF packets that would have happened if we use a single socket with three streams). This is

particularly important in order to deal with node overload well, as discussed in Section 4.2.2.6.

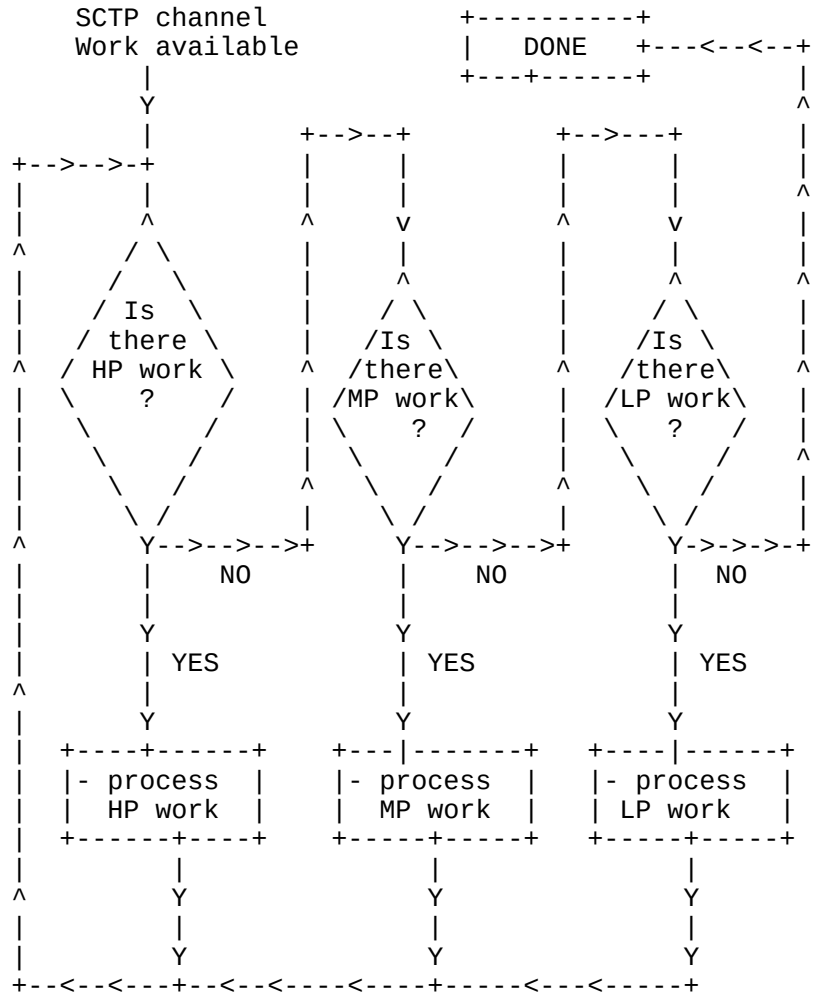

Figure 5: SCTP TML Strict Priority Scheduling

4.2.1.6. SCTP TML Parameterization

The following is a list of parameters needed for booting the TML. It is expected these parameters will be extracted via the FEM/CEM interface for each PL ID.

1. The IP address(es) or a resolvable DNS/hostname(s) of the CE/FE.

2. Whether or not to use IPsec. If IPsec is used, how to parameterize the different required ciphers, keys, etc., as described in Section 7.1

3. The HP SCTP port, as discussed in Section 4.2.1.2. The default HP port value is 6704 (Section 6).

4. The MP SCTP port, as discussed in Section 4.2.1.3. The default MP port value is 6705 (Section 6).

5. The LP SCTP port, as discussed in Section 4.2.1.4. The default LP port value is 6706 (Section 6).

4.2.2. Satisfying TML Requirements

[RFC5810], Section 5 lists requirements that a TML needs to meet. This section describes how the SCTP TML satisfies those requirements.

4.2.2.1. Satisfying Reliability Requirement

As mentioned earlier, a shade of reliability ranges is possible in SCTP. Therefore, this requirement is met.

4.2.2.2. Satisfying Congestion Control Requirement

Congestion control is built into SCTP. Therefore, this requirement is met.

4.2.2.3. Satisfying Timeliness and Prioritization Requirement

By using three sockets in conjunction with the partial-reliability feature [RFC3758], both timeliness and prioritization requirements are addressed.

4.2.2.4. Satisfying Addressing Requirement

There are no extra headers required for SCTP to fulfill this requirement. SCTP can be told to replicast packets to multiple destinations. The TML implementation will need to translate PL addresses to a variety of unicast IP addresses in order to emulate multicast and broadcast PL addresses.

4.2.2.5. Satisfying High-Availability Requirement

Transport link resiliency is one of SCTP’s strongest points. Failure detection and recovery is built in, as mentioned earlier.

- The SCTP multi-homing feature is used to provide path diversity. Should one of the peer IP addresses become unreachable, the others are used without needing lower-layer convergence (routing, for example) or even the TML becoming aware.

- SCTP heartbeats and data transmission thresholds are used on a per-peer IP address to detect reachability faults. The faults could be a result of an unreachable address or peer, which may be caused by a variety of reasons, like interface, network, or endpoint failures. The cause of the fault is noted.

- With the ADDIP feature, one can migrate IP addresses to other nodes at runtime. This is not unlike the Virtual Router Redundancy Protocol (VRRP) [RFC5798] use. This feature is used in addition to multi-homing in a planned migration of activity from one FE/CE to another. In such a case, part of the provisioning recipe at the CE for replacing an FE involves migrating activity of one FE to another.

4.2.2.6. Satisfying Node Overload Prevention Requirement

The architecture of this TML defines three separate channels, one per socket, to be used within any FE-CE setup. The work scheduling design for processing the TML channels (Section 4.2.1.5) is a strict priority. A fundamental desire of the strict prioritization is to ensure that more important processing work always gets node resources over less important work.

When a ForCES node CPU is overwhelmed because the incoming packet rate is higher than it can keep up with, the channel queues grow and transport congestion subsequently follows. By virtue of using SCTP, the congestion is propagated back to the source of the incoming packets and eventually alleviated.

The HP channel work gets prioritized at the expense of the MP, which gets prioritized over LP channels. The preferential scheduling only kicks in when there is node overload regardless of whether there is transport congestion. As a result of the preferential work treatment, the ForCES node achieves a robust steady processing capacity. Refer to Appendix A.2 for details on scheduling.

For an example of how the overload prevention works, consider a scenario where an overwhelming amount of redirected packets (from outside the NE) coming into the NE may overload the FE while it has outstanding config work from the CE. In such a case, the FE, while it is busy processing config requests from the CE, essentially ignores processing the redirect packets on the LP channel. If enough redirect packets accumulate, they are dropped either because the LP

channel threshold is exceeded or because they are obsoleted. If on the other hand, the FE has successfully processed the higher-priority channels and their related work, then it can proceed and process the LP channel. So as demonstrated in this case, the TML ties transport congestion and node overload implicitly together.

4.2.2.7. Satisfying Encapsulation Requirement

The SCTP TML sets SCTP PPIDs to identify channels used as described in Section 4.2.1.1.

5. SCTP TML Channel Work

There are two levels of TML channel work within an NE when a ForCES node (CE or FE) is connected to multiple other ForCES nodes:

1. NE-level I/O work where a ForCES node (CE or FE) needs to choose which of the peer nodes to process.

2. Node-level I/O work where a ForCES node, handles the three SCTP TML channels separately for each single ForCES endpoint.

NE-level scheduling definition is left up to the implementation and is considered out of scope for this document. Appendix A.4 briefly discusses some constraints about which an implementer needs to worry.

This document provides suggestions on SCTP channel work implementation in Appendix A.

The FE SHOULD do channel connections to the CE in the order of incrementing priorities, i.e., LP socket first, followed by MP, and ending with HP socket connection. The CE, however, MUST NOT assume that there is ordering of socket connections from any FE.

6. IANA Considerations

Following the policies outlined in "Guidelines for Writing an IANA Considerations Section in RFCs" [RFC5226], the following namespaces are defined in ForCES SCTP TML.

- SCTP port 6704 for the HP channel, 6705 for the MP channel, and 6706 for the LP channel.

- SCTP Payload Protocol ID (PPID) 21 for the HP channel (ForCES-HP), 22 for the MP channel (ForCES-MP), and 23 for the LP channel (ForCES-LP).

7. Security Considerations

The SCTP TML provides the following security services to the PL:

- o A mechanism to authenticate ForCES CEs and FEs at the transport level in order to prevent the participation of unauthorized CEs and unauthorized FEs in the control and data path processing of a ForCES NE.

- o A mechanism to ensure message authentication of PL data and headers transferred from the CE to FE (and vice versa) in order to prevent the injection of incorrect data into PL messages.

- o A mechanism to ensure the confidentiality of PL data and headers transferred from the CE to FE (and vice versa), in order to prevent disclosure of PL information transported via the TML.

Security choices provided by the TML are made by the operator and take effect during the pre-association phase of the ForCES protocol. An operator may choose to use all, some or none of the security services provided by the TML in a CE-FE connection.

When operating under a secured environment, or for other operational concerns (in some cases performance issues) the operator may turn off all the security functions between CE and FE.

IP Security Protocol (IPsec) [RFC4301] is used to provide needed security mechanisms.

IPsec is an IP-level security scheme transparent to the higher-layer applications and therefore can provide security for any transport layer protocol. This gives IPsec the advantage that it can be used to secure everything between the CE and FE without expecting the TML implementation to be aware of the details.

The IPsec architecture is designed to provide message integrity and message confidentiality outlined in the TML security requirements [RFC5810]. Mutual authentication and key exchange protocol are provided by Internet Key Exchange (IKE) [RFC2409].

7.1. IPsec Usage

A ForCES FE or CE MUST support the following:

- o Internet Key Exchange (IKE)[RFC2409] with certificates for endpoint authentication.

- o Transport Mode Encapsulating Security Payload (ESP) [RFC4303].

- o HMAC-SHA1-96 [RFC2404] for message integrity protection
- o AES-CBC with 128-bit keys [RFC3602] for message confidentiality.
- o Replay protection [RFC4301].

A compliant implementation SHOULD provide operational means for configuring the CE and FE to negotiate other cipher suites and even use manual keying.

7.1.1. SAD and SPD Setup

To minimize the operational configuration, it is RECOMMENDED that only the IANA-issued SCTP protocol number (132) be used as a selector in the Security Policy Database (SPD) for ForCES. In such a case, only a single SPD and SAD entry is needed.

Setup MAY alternatively extend the above policy so that it uses the three SCTP TML port numbers as SPD selectors. But as noted above, this choice will require an increased number of SPD entries.

In scenarios where multiple IP addresses are used within a single association, and there is desire to configure different policies on a per-IP address, then following [RFC3554] is RECOMMENDED.

8. Acknowledgements

The authors would like to thank Joel Halpern, Michael Tuxen, Randy Stewart, Evangelos Haleplidis, Chuanhuang Li, Lars Eggert, Avshalom Houri, Adrian Farrel, Juergen Quittek, Magnus Westerlund, and Pasi Eronen for engaging us in discussions that have made this document better.

Ross Callon was an excellent manager who persevered in providing us guidance and Joel Halpern was an excellent document shepherd without whom this document would have taken longer to publish.

9. References

9.1. Normative References

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

[RFC2404] Madson, C. and R. Glenn, "The Use of HMAC-SHA-1-96 within ESP and AH", RFC 2404, November 1998.

[RFC2409] Harkins, D. and D. Carrel, "The Internet Key Exchange (IKE)", RFC 2409, November 1998.

[RFC3554] Bellovin, S., Ioannidis, J., Keromytis, A., and R. Stewart, "On the Use of Stream Control Transmission Protocol (SCTP) with IPsec", RFC 3554, July 2003.

[RFC3602] Frankel, S., Glenn, R., and S. Kelly, "The AES-CBC Cipher Algorithm and Its Use with IPsec", RFC 3602, September 2003.

[RFC3758] Stewart, R., Ramalho, M., Xie, Q., Tuexen, M., and P. Conrad, "Stream Control Transmission Protocol (SCTP) Partial Reliability Extension", RFC 3758, May 2004.

[RFC4301] Kent, S. and K. Seo, "Security Architecture for the Internet Protocol", RFC 4301, December 2005.

[RFC4303] Kent, S., "IP Encapsulating Security Payload (ESP)", RFC 4303, December 2005.

[RFC4960] Stewart, R., "Stream Control Transmission Protocol", RFC 4960, September 2007.

[RFC5226] Narten, T. and H. Alvestrand, "Guidelines for Writing an IANA Considerations Section in RFCs", BCP 26, RFC 5226, May 2008.

[RFC5810] Doria, A., Ed., Hadi Salim, J., Ed., HAAS, R., Ed., Khosravi, H., Ed., Wang, W., Ed., Dong, L., Gopal, R., and J. Halpern, "Forwarding and Control Element Separation (ForCES) Protocol Specification", RFC 5810, March 2010.

9.2. Informative References

[RFC3654] Khosravi, H. and T. Anderson, "Requirements for Separation of IP Control and Forwarding", RFC 3654, November 2003.

[RFC3746] Yang, L., Dantu, R., Anderson, T., and R. Gopal, "Forwarding and Control Element Separation (ForCES) Framework", RFC 3746, April 2004.

[RFC5812] Halpern, J. and J. Hadi Salim, "Forwarding and Control Element Separation (ForCES) Forwarding Element Model", RFC 5812, March 2010.

[RFC5798] Nadas, S., Ed., "Virtual Router Redundancy Protocol (VRRP) Version 3 for IPv4 and IPv6", RFC 5798, March 2010.

[TSVWG-SCTPSOCKET] Stewart, R., Poon, K., Tuexen, M., Yasevich, V., and P. Lei, "Sockets API Extensions for Stream Control Transmission Protocol (SCTP)", Work in Progress, March 2010.

Appendix A. Suggested SCTP TML Channel Work Implementation

As mentioned in Section 5, there are two levels of TML channel work within an NE when a ForCES node (CE or FE) is connected to multiple other ForCES nodes:

1. NE-level I/O work where a ForCES node (CE or FE) needs to choose which of the peer nodes to process.

2. Node-level I/O work where a ForCES node, handles the three SCTP TML channels separately for each single ForCES endpoint.

NE-level scheduling definition is left up to the implementation and is considered out of scope for this document. Appendix A.4 briefly discusses some constraints about which an implementer needs to worry.

This document, and in particular Appendix A.1, Appendix A.2, and Appendix A.3 discuss details of node-level I/O work.

A.1. SCTP TML Channel Initialization

As discussed in Section 5, it is recommended that the FE SHOULD do socket connections to the CE in the order of incrementing priorities, i.e., LP socket first, followed by MP, and ending with HP socket connection. The CE, however, MUST NOT assume that there is ordering of socket connections from any FE. Appendix B.1 has more details on the expected initialization of SCTP channel work.

A.2. Channel Work Scheduling

This section provides high-level details of the scheduling view of the SCTP TML core (Section 4.2.1). A practical scheduler implementation takes care of many little details (such as timers, work quanta, etc.) not described in this document. It is left to the implementer to take care of those details.

The CE(s) and FE(s) are coupled together in the principles of the scheduling scheme described here to tie together node overload with transport congestion. The design intent is to provide the highest possible robust work throughput for the NE under any network or processing congestion.

A.2.1. FE Channel Work Scheduling

The FE scheduling, in priority order, needs to I/O process:

1. The HP channel I/O in the following priority order:

1. Transmitting back to the CE any outstanding result of executed work via the HP channel transmit path.

2. Taking new incoming work from the CE that creates ForCES work to be executed by the FE.

2. ForCES events that result in transmission of unsolicited ForCES packets to the CE via the MP channel.

3. Incoming Redirect work in the form of control packets that come from the CE via LP channel. After redirect processing, these packets get sent out on external (to the NE) interface.

4. Incoming Redirect work in the form of control packets that come from other NEs via external (to the NE) interfaces. After some processing, such packets are sent to the CE.

It is worth emphasizing, at this point again, that the SCTP TML processes the channel work in strict priority. For example, as long as there are messages to send to the CE on the HP channel, they will be processed first until there are no more left before processing the next priority work (which is to read new messages on the HP channel incoming from the CE).

A.2.2. CE Channel Work Scheduling

The CE scheduling, in priority order, needs to deal with:

1. The HP channel I/O in the following priority order:

    1. Process incoming responses to requests of work it made to the FE(s).

    2. Transmit any outstanding HP work it needs the FE(s) to complete.

2. Incoming ForCES events from the FE(s) via the MP channel.

3. Outgoing Redirect work in the form of control packets that get sent from the CE via LP channel destined to external (to the NE) interface on FE(s).

4. Incoming Redirect work in the form of control packets that come from other NEs via external interfaces (to the NE) on the FE(s).

It is worth repeating, for emphasis, that the SCTP TML processes the channel work in strict priority. For example, if there are messages incoming from an FE on the HP channel, they will be processed first

until there are no more left before processing the next priority work, which is to transmit any outstanding HP channel messages going to the FE.

A.3. SCTP TML Channel Termination

Appendix B.2 describes a controlled disassociation of the FE from the NE.

It is also possible for connectivity to be lost between the FE and CE on one or more sockets. In cases where SCTP multi-homing features are used for path availability, the disconnection of a socket will only occur if all paths are unreachable; otherwise, SCTP will ensure reachability. In the situation of a total connectivity loss of even one SCTP socket, it is recommended that the FE and CE SHOULD assume a state equivalent to ForCES Association Teardown being issued and follow the sequence described in Appendix B.2.

A CE could also disconnect sockets to an FE to indicate an "emergency teardown". The "emergency teardown" may be necessary in cases when a CE needs to disconnect an FE but knows that an FE is busy processing a lot of outstanding commands (some of which the FE hasn’t gotten around to processing, yet). By virtue of the CE closing the connections, the FE will immediately be asynchronously notified and will not have to process any outstanding commands from the CE.

A.4. SCTP TML NE-Level Channel Scheduling

In handling NE-level I/O work, an implementation needs to worry about being both fair and robust across peer ForCES nodes.

Fairness is desired so that each peer node makes progress across the NE. For the sake of illustration, consider two FEs connected to a CE; whereas one FE has a few HP messages that need to be processed by the CE, another may have infinite HP messages. The scheduling scheme may decide to use a quota scheduling system to ensure that the second FE does not hog the CE cycles.

Robustness is desired so that the NE does not succumb to a Denial-of-Service (DoS) attack from hostile entities and always achieves a maximum stable workload processing level. For the sake of illustration, consider again two FEs connected to a CE. Consider FE1 as having a large number of HP and MP messages and FE2 having a large number of MP and LP messages. The scheduling scheme needs to ensure that while FE1 always gets its messages processed, at some point we allow FE2 messages to be processed. A promotion and preemption-based scheduling could be used by the CE to resolve this issue.

Appendix B. Suggested Service Interface

This section outlines a high-level service interface between FEM/CEM and TML, the PL and TML, and between local and remote TMLs. The intent of this interface discussion is to provide general guidelines. The implementer is expected to care of details and even follow a different approach if needed.

The theory of operation for the PL-TML service is as follows:

1. The PL starts up and bootstraps the TML. The end result of a successful TML bootstrap is that the CE TML and the FE TML connect to each other at the transport level.

2. Transmission and reception of the PL messages commences after a successful TML bootstrap. The PL uses send and receive PL-TML interfaces to communicate to its peers. The TML is agnostic to the nature of the messages being sent or received. The first message exchanges that happen are to establish ForCES association. Subsequent messages may be either unsolicited events from the FE PL, control message redirects to/from the CE to/from FE, or configuration from the CE to the FE, and their responses flowing from the FE to the CE.

3. The PL does a shutdown of the TML after terminating ForCES association.

B.1. TML Bootstrapping

Figure 6 illustrates a flow for the TML bootstrapped by the PL.

When the PL starts up (possibly after some internal initialization), it boots up the TML. The TML first interacts with the FEM/CEM and acquires the necessary TML parameterization (Section 4.2.1.6). Next, the TML uses the information it retrieved from the FEM/CEM interface to initialize itself.

The TML on the FE proceeds to connect the three channels to the CE. The socket interface is used for each of the channels. The TML continues to re-try the connections to the CE until all three channels are connected. It is advisable that the number of connection retry attempts and the time between each retry is also configurable via the FEM. On failure to connect one or more channels, and after the configured number of retry thresholds is exceeded, the TML will return an appropriate failure indicator to the PL. On success (as shown in Figure 6), a success indication is presented to the PL.

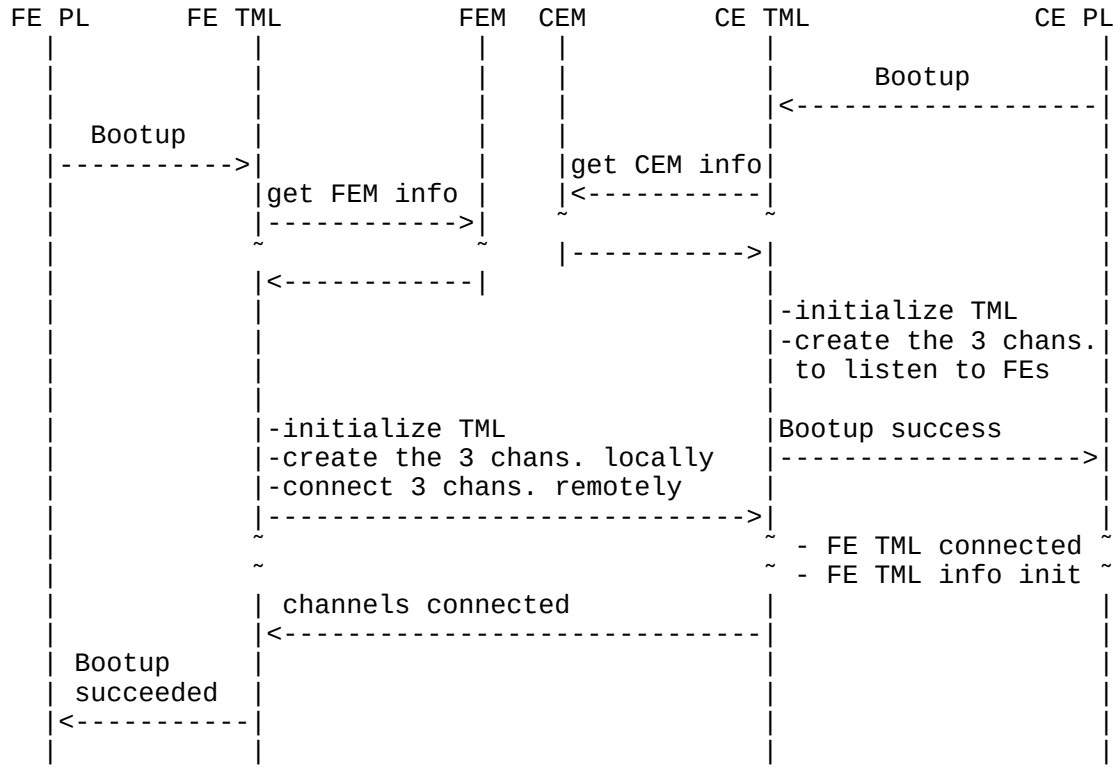

Figure 6: SCTP TML Bootstrapping

On the CE, things are slightly different. After initializing from the CEM, the TML on the CE side proceeds to initialize the three channels to listen to remote connections from the FEs. The success or failure indication is passed on to the CE PL (in the same manner as was done in the FE).

Post bootup, the CE TML waits for connections from the FEs. Upon a successful connection by an FE, the CE TML level keeps track of the transport-level details of the FE. Note, at this stage only transport-level connection has been established; ForCES-level association follows using send/receive PL-TML interfaces (refer to Appendix B.3 and Figure 8).

B.2. TML Shutdown

Figure 7 shows an example of an FE shutting down the TML. It is assumed at this point that the ForCES Association Teardown has been issued by the CE. It should also be noted that different implementations may have different procedures for cleaning up state, etc.

When the FE PL issues a shutdown to its TML for a specific PL ID, the TML releases all the channel connections to the CE. This is achieved by closing the sockets used to communicate to the CE. This results in the stack sending a SCTP shutdown, which is received on the CE.

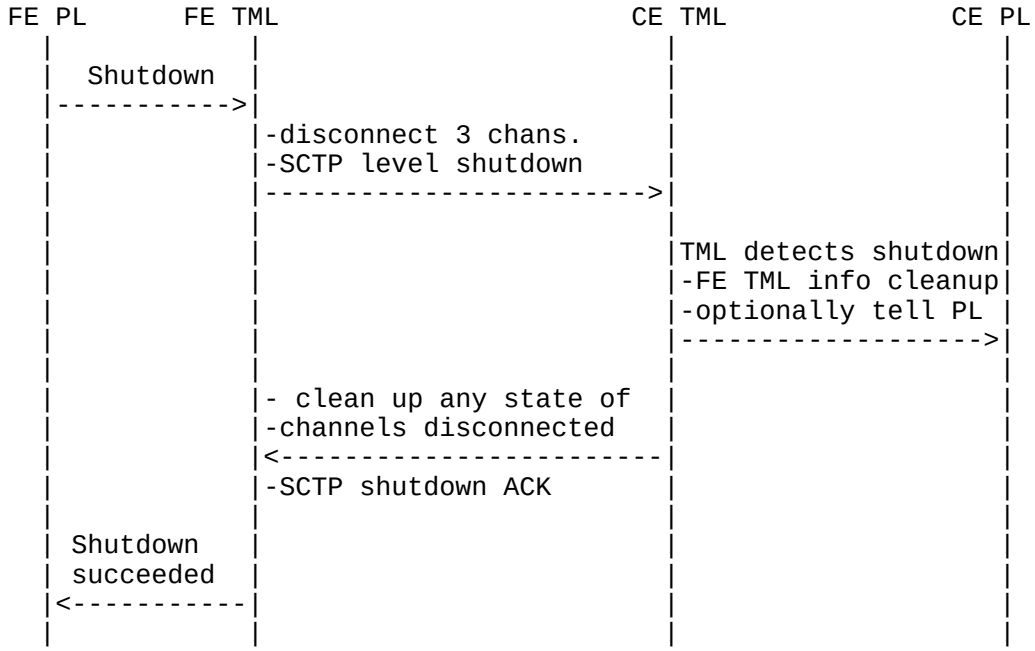

Figure 7: FE Shutting Down

On the CE side, a TML disconnection would result in possible cleanup of the FE state. Optionally, depending on the implementation, there may be need to inform the PL about the TML disconnection. The CE-stack-level SCTP sends an acknowledgement to the FE TML in response to the earlier SCTP shutdown.

B.3. TML Sending and Receiving

The TML should be agnostic to the content of the PL messages, or their operations. The PL should provide enough information to the TML for it to assign an appropriate priority and loss behavior to the message. Figure 8 shows an example of a message exchange originated at the FE and sent to the CE (such as a ForCES association message), which illustrates all the necessary service interfaces for sending and receiving.

When the FE PL sends a message to the TML, the TML is expected to pick one of HP/MP/LP channels and send out the ForCES message.

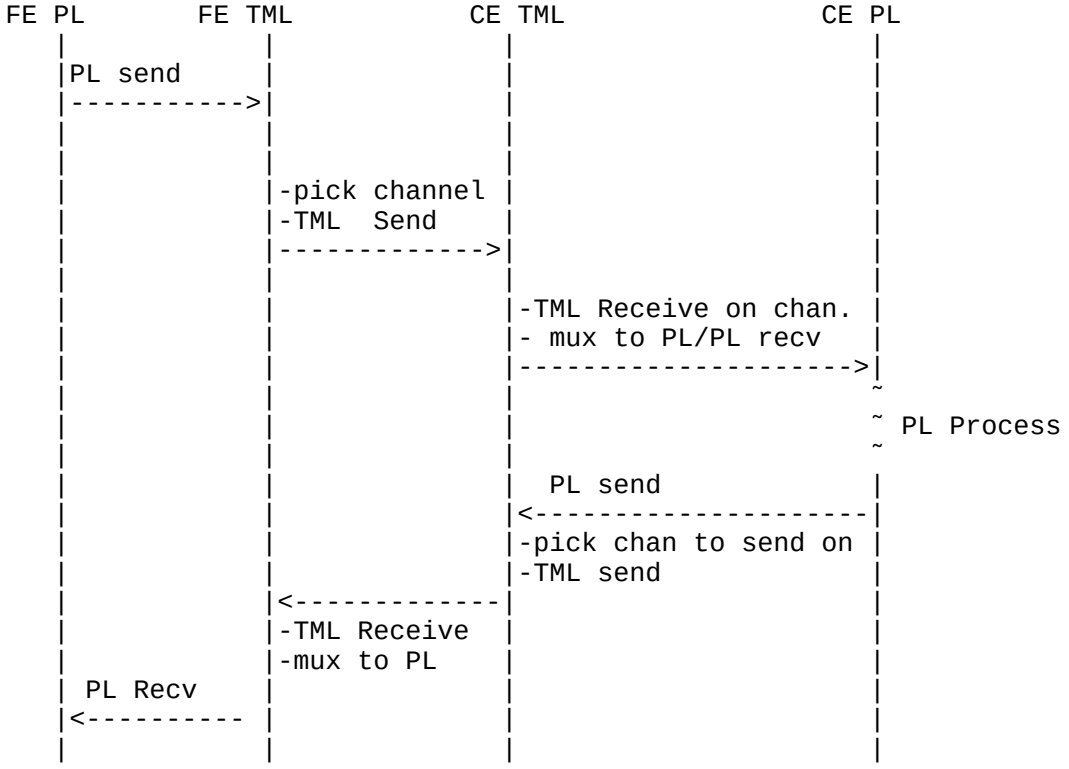

Figure 8: Send and Recv Flow

When the CE TML receives the ForCES message on the channel on which it was sent, it demultiplexes the message to the CE PL.

The CE PL, after some processing (in this example, dealing with the FE's association), sends the TML the response. As in the case of FE PL, the CE TML picks the channel to send on before sending.

The processing of the ForCES message upon arrival at the FE TML and delivery to the FE PL is similar to the CE side equivalent as shown above in Appendix B.3.

Authors' Addresses

Jamal Hadi Salim
Mojatatu Networks
Ottawa, Ontario
Canada

EMail: hadi@mojatatu.com

Kentaro Ogawa
NTT Corporation
3-9-11 Midori-cho
Musashino-shi, Tokyo 180-8585
Japan

EMail: ogawa.kentaro@lab.ntt.co.jp