Network Working Group
Request for Comments: 5043
Category: Standards Track

C. Bestler, Ed.
Neterion
R. Stewart, Ed.
Cisco Systems, Inc.

October 2007

Stream Control Transmission Protocol (SCTP) Direct Data Placement (DDP) Adaptation

Status of This Memo

This document specifies an Internet standards track protocol for the Internet community, and requests discussion and suggestions for improvements. Please refer to the current edition of the "Internet Official Protocol Standards" (STD 1) for the standardization state and status of this protocol. Distribution of this memo is unlimited.

Abstract

This document specifies an adaptation layer to provide a Lower Layer Protocol (LLP) service for Direct Data Placement (DDP) using the Stream Control Transmission Protocol (SCTP).

Table of Contents

1. Introduction

This document describes a method to adapt Direct Data Placement [RFC5041] to Stream Control Transmission Protocol (SCTP) [RFC4960].

Some implementations may include this adaptation layer within their SCTP implementations to obtain maximum performance, but the behavior of SCTP will be unaffected. An SCTP layer used solely by this adaptation layer is able to take certain optimizations based on the limited subset of SCTP capabilities used. In order to allow optimization for these implementations, we specify the use of the new adaptation layer indication as defined in [RFC5061]

1.1. Conventions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in RFC 2119 [RFC2119].

2. Definitions

DDP - See Direct Data Placement Protocol.

DDP Endpoint - The logical sender/receiver of DDP Segments. An SCTP stream pair is not assumed to have a DDP Endpoint. DDP Segments may only be sent once a DDP Endpoint has been assigned to an SCTP stream pair by a local interface.

DDP Source Stream Sequence Number (DDP-SSN) - A stream-specific sequence number assigned by the adaptation layer for each SCTP Data Chunk sent. This is the order that chunks were submitted to SCTP, no matter in what order they are actually sent or received.

DDP Segment - The smallest unit of data transfer for the DDP protocol. It includes a DDP Header and ULP Payload (if present). A DDP Segment should be sized to fit within the Lower Layer Protocol MULPDU (Marker PDU Aligned (MPA) Upper Layer PDU).

DDP Segment Chunk - An SCTP Payload Data Chunk that encapsulates the DDP-SSN and a DDP Segment.

DDP Stream - A sequence of DDP Segments whose ordering is defined by the LLP. For SCTP, a DDP stream maps directly to a bidirectional pair of SCTP streams with the same Stream IDs. Note that DDP has no ordering guarantees between DDP streams.

DDP Stream Session - A single pairing of DDP Endpoints over a DDP stream that lasts from an Initiation message through the Termination message(s).

DDP Stream Session Control Message - A message that is used to control the association of the DDP Endpoint with the DDP stream.

Direct Data Placement Protocol (DDP) - A wire protocol that supports Direct Data Placement by associating explicit memory buffer placement information with the LLP payload units.

Lower Layer Protocol (LLP) - In the context of DDP, the protocol layer beneath RDMA that provides a reliable transport service. The SCTP DDP adaption is one of the initially defined LLPs for DDP.

Protection Domain - A common local interface convention to control which Steering Tags (STags) are valid with which DDP Endpoints. Under this convention, both the Steering Tag and DDP Endpoint are created within the context of a Protection Domain, and the Steering Tag may only be enabled for DDP Endpoints created under the same Protection Domain.

RDMA - Remote Direct Memory Access.

RNIC - RDMA Network Interface Card.

SCTP association - A protocol relationship between two SCTP endpoints. An SCTP association supports multiple SCTP streams.

SCTP Data Chunk - An SCTP Chunk used to convey Payload Data. There can be multiple Chunks within each SCTP packet. Other Chunks are used to control the SCTP Association.

SCTP endpoint - The logical sender/receiver of SCTP packets. On a multihomed host, an SCTP endpoint is represented to its peers as a combination of an SCTP port number and a set of eligible destination transport addresses to which SCTP packets can be sent.

SCTP Stream - A unidirectional logical channel established from one to another associated SCTP endpoint. There can be multiple SCTP streams within each SCTP association. An SCTP stream is used to form one direction of a DDP stream.

Transmission Sequence Number (TSN) - A 32-bit sequence number used internally by SCTP. One TSN is attached to each chunk containing user data to permit the receiving SCTP endpoint to acknowledge its receipt and detect duplicate deliveries.

Upper Layer Protocol (ULP) - In the context of RDMA protocol specifications, this is the layer using RDMA services. Typically, this is an application or middleware. A primary goal of RDMA protocols is to enable direct transfer of payload to/from ULP Buffers.

3. Motivation

This document specifies an adaptation layer which fulfills the requirements of a Lower Layer Protocol (LLP) for DDP using a specific subset of SCTP capabilities.

The defined protocol is intended to be implementable over existing SCTP stacks, while clearly defining what portions of SCTP are required to enable an implementation to be optimized specifically to support DDP.

4. Overview

The adaptation layer uses a pair of like-numbered SCTP streams within an SCTP Association to provide a reliable DDP stream between two DDP Endpoints. Except as specifically noted, each DDP Segment submitted by the DDP layer is encoded as a single unordered SCTP Data Chunk. In addition to the DDP Segment, the Data Chunk also contains a sequence number (DDP-SSN) that reflects the order in which DDP submitted the segments for that stream.

A DDP Stream Session is defined by DDP Stream Session Control Chunks that manage the state of the DDP Stream Session. These Chunks dynamically bind DDP Endpoints to the DDP Stream Session, and DDP Segment Chunks are used to reliably deliver DDP Segments with the session.

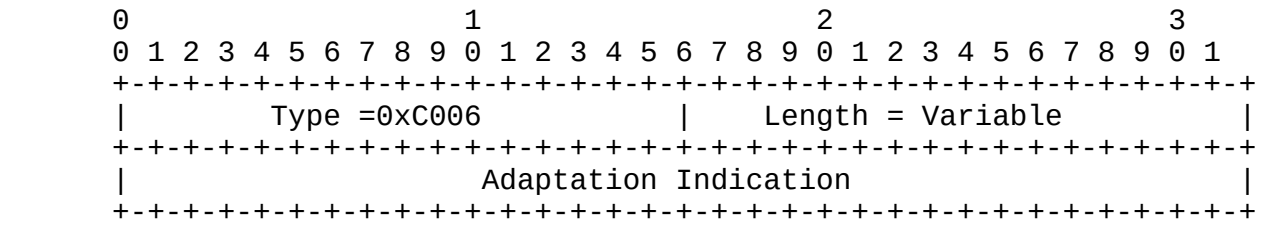
5. Data Formats

5.1. Adaptation Layer Indicator

The DDP/SCTP adaptation layer uses all streams within an SCTP association. An SCTP Association that has had the DDP Adaptation Indication negotiated will carry only SCTP Data Chunks as defined in this document.

It is presumed that the handling of incoming data chunks for DDP-enabled associations is sufficiently different than for routine SCTP associations that it is undesirable to require support for mixing DDP and non-DDP streams in a single association. More than a single association is required if an application desires to utilize both DDP and non-DDP traffic with the same remote host.

We define an Adaptation Indication that MUST appear in the INIT or INIT-ACK with the following format as defined in [RFC5061].



Adaptation Indication:

The following value has been assigned for DDP.

DDP - 0x00000001

5.2. Payload Data Chunks

The DDP SCTP adaptation uses two types of SCTP Payload Data Chunks, differentiated by the Payload Protocol Identifier:

DDP Segment Chunks are used to reliably deliver DDP Segments sent between DDP Endpoints.

DDP Stream Session Control Messages are used to establish and tear down DDP Stream Sessions, specifically by controlling the binding of DDP Endpoints with SCTP streams.

Payload Protocol Identifier:

The following value are defined for DDP in this document and have been assigned by IANA:

DDP Segment Chunk - 16
DDP Stream Session Control - 17

5.2.1. DDP Source Sequence Number (DDP-SSN)

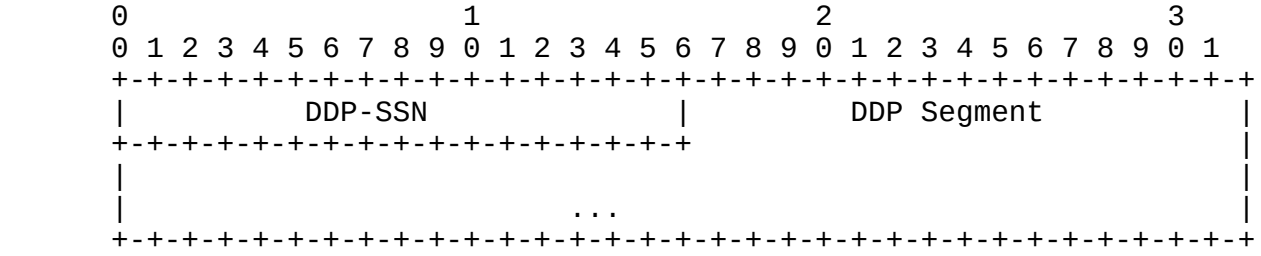
All SCTP Payload Data Chunks used by this adaptation layer include a DDP Source Sequence Number (DDP-SSN). The DDP-SSN tracks the sequence in which the messages were submitted to the SCTP layer for the SCTP stream in use. The DDP-SSN MUST have the same value that the SCTP Stream Sequence Number (SSN) would have been assigned had ordered SCTP Payload Data Chunks been used rather than unordered.

The rationale for specifying the DDP-SSN is as follows:

- The SCTP Stream Sequence Number (SSN) is not suitable for this purpose because all messages defined by this document use unordered Payload Data Chunks to ensure prompt delivery from the receiving SCTP layer.

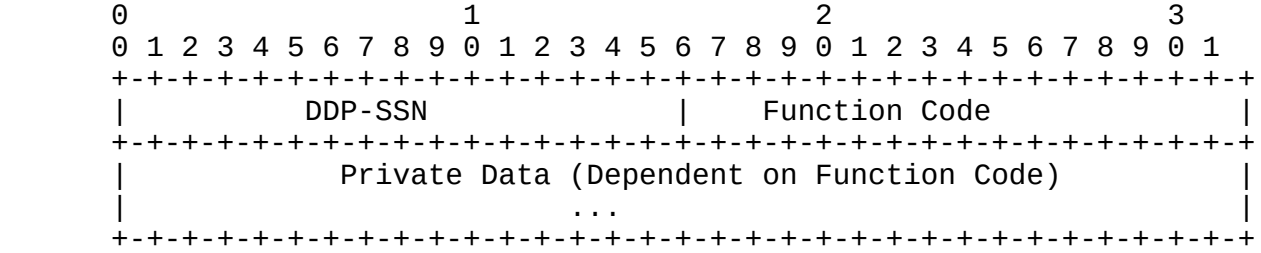
- The SCTP Transmission Sequence Number (TSN) is not suitable for determining the original order of Data Chunks within a stream. The sending SCTP layer is allowed to optimize the transmission sequence of unordered Data Chunks to encourage Chunk Bundling, or for other purposes.

5.2.2. DDP Segment Chunk



DDP Segments are as defined in [RFC5041]. The DDP Segment Chunk serves the same purpose as the MPA [RFC5044] Upper Layer PDU (MULPDU) in that it carries DDP Segments over a reliable protocol with added sequencing information.

5.2.3. DDP Stream Session Control



The following function code values are defined for DDP in this document:

```
DDP Stream Session Initiate         - 0x001
DDP Stream Session Accept           - 0x002
DDP Stream Session Reject           - 0x003
DDP Stream Session Terminate        - 0x004
```

ULP-supplied Private Data MUST be included for DDP Stream Session Initiate, DDP Stream Session Accept, and DDP Stream Session Reject messages. However, the ULP supplied Private DATA MAY be of zero length.

Private Data length MUST NOT exceed 512 bytes in any message.

Private Data MUST NOT be included in the DDP Stream Session Terminate message.

Received DDP Stream Session Control messages SHOULD be reported to the ULP. If reported, any supplied Private Data MUST be available for the ULP to examine.

The DDP/SCTP adaptation layer MAY limit the number of Session Initiate requests that it has submitted to the ULP. When a DDP Stream Session Initiate cannot be forwarded to the ULP due to such a limit, the adaptation layer MUST respond with a DDP Stream Session Terminate message.

6. DDP Stream Sessions

A DDP Endpoint is the logical sender/receiver of DDP Segments. A DDP stream connects two DDP Endpoints using a matched pair of SCTP streams having the same SCTP Stream Identifiers.

A DDP Stream Session defines the sequence of Data Chunks exchanged between two DDP Endpoints over a DDP stream that has a distinct beginning and end as defined in the following section. Data Chunks from one DDP Stream Session are never carried over to the next session. Each Data Chunk unambiguously belongs to exactly one session. The DDP-SSNs assigned to the Data Chunks for a session MUST NOT have any gaps.

The local interface MAY dynamically associate a DDP Endpoint with the DDP stream based upon the initial exchanges of a DDP Session, and dynamically terminate that association at the session’s end. Alternately, a specialized local interface could simply statically map DDP Endpoints to DDP streams.

Conventionally, local interfaces for RDMA have deferred the selection of the DDP Endpoint until after the ULP decides to accept an RDMA connection request. But that is a local interface choice and not a wire protocol requirement.

A DDP stream is associated with at most one Protection Domain during a single DDP Stream Session. On the passive side, the association is typically deferred until the DDP Stream Session Accept message.

6.1. Sequencing

The DDP-SSN is reset to zero at the beginning of each DDP Stream Session.

The normative sequence for considering Payload Data Chunks within a given session is based upon each Data Chunk's DDP-SSN. When considered in this normative sequence, all sessions MUST conform to one of the patterns defined in this section.

If the adaptation layer receives a Payload Data Chunk that conforms to none of the enumerated legal patterns, the DDP Stream Session MUST be terminated.

6.2. Legal Sequence: Active/Passive Session Accepted

In this DDP Stream Session sequence, one DDP Endpoint assumes the active role in requesting a DDP Stream Session, which the other side accepts.

Active side sends a DDP Stream Session Initiate message.

Passive side sends a DDP Stream Session Accept message.

Each side may then send zero or more DDP Segments with increasing DDP-SSNs, subject to any flow control imposed by other protocol layers.

The final User Data Chunk for each side MAY be a DDP Stream Terminate. At least one side MUST send a DDP Stream Terminate. Note that this would follow any RDMAP Terminate message, which to the adaptation layer is simply another DDP Segment.

6.3. Legal Sequence: Active/Passive Session Rejected

DDP Stream Sessions allow each party to send a single non-payload message before the other end commits specific resources to the session. This allows each end to determine which resources are to be used, and how they are to be configured, or even if the session should be granted.

These decisions MAY be influenced by the need to assign a specific Protection Domain, to determine how many RDMA Read Credits are required, or to determine how many receive operations the ULP should enable.

Because of these or other factors, the passive side MAY choose to reject a DDP Stream Session Request. This results in the following legal sequence:

Active side sends a DDP Stream Session Initiate message.

Passive side sends a DDP Stream Session Reject message.

A DDP Stream Session Reject message MUST NOT be sent unless the rejection is at the direction of the ULP.

6.4. Legal Sequence: Active/Passive Session Non-ULP Rejected

Acceptance or rejection of DDP Stream Session Initiate messages SHOULD be under the control of the ULP. This MAY require passing an event to the ULP. There MUST be a finite limit on the number of such requests that are pending a ULP decision. When more session requests are received, the passive side MUST respond to the Initiate message with a DDP Stream Terminate Message.

6.5. ULP-Specific Sequencing

An implementation MAY choose to support additional ULP-specific sequences, but MUST NOT do so unless requested to do so by the ULP.

A defined ULP MUST be able to operate using only the defined mandatory session sequences. Any additional sequences must be used only for optional optimizations.

6.6. Other Sequencing Rules

A DDP Stream Session Control message MUST NOT be sent if it may be received before a prior DDP Stream Session Control message within the same DDP Stream Session.

An active side of a DDP Stream Session MUST NOT send a DDP Segment that might be received before the DDP Stream Session Initiate message.

This MAY be determined by SCTP acking of the Data Chunk used to carry the DDP Stream Session Initiate message, or by receipt of a responsive DDP Stream Session Control message.

A DDP Stream Identifier MUST NOT be reused for another DDP Stream Session while any Data Chunk from a prior session might be outstanding.

7. SCTP Endpoints

7.1. Adaptation Layer Indication Restriction

The local interface MUST allow the ULP to specify an SCTP endpoint to use a specific Adaptation Indication. It MAY require the ULP to do so.

Once an endpoint decides on its acceptable Adaptation Indication(s), it SHOULD terminate all requests to establish an association with any different Adaptation Indication.

An SCTP implementation MAY choose to accept association requests for a given SCTP endpoint only until one association for the endpoint has been established. At that point, it MAY choose to restrict all further associations for the same endpoint to use the same Adaptation Indication.

7.2. Multihoming Implications

SCTP allows an SCTP endpoint to be associated with multiple IP addresses, potentially representing different interface devices. Distribution of the logic for a single DDP stream across multiple input devices can be very undesirable, resulting in complex cache coherency challenges. Therefore, the local interface MAY restrict DDP-enabled SCTP endpoints to a single IP address, or to a set of IP addresses that are all assigned to the same input device ("RNIC").

The default binding of a DDP-enabled SCTP endpoint SHOULD NOT cover more than a single IP address unless doing so results in neither additional bus traffic nor duplication of memory registration resources. This will frequently result in a different default than for SCTP endpoints that are not DDP enabled.

Applications MAY choose to avoid using out-of-band methods for communicating the set of IP addresses used by an SCTP endpoint when there is potential confusion as to the intended scope of the SCTP endpoint. For example, assuming the SCTP endpoint consists of all IP addresses Advertised by DNS may work for a general purpose SCTP endpoint but not a DDP-enabled one.

Even when multihoming is supported, ULPs are cautioned that they SHOULD NOT use ULP control of the source address in an attempt to load-balance a stream across multiple paths. A receiving DDP/SCTP implementation that chooses to support multihoming SHOULD optimize its design on the assumption that multihoming will be used for network fault tolerance, and not to load-balance between paths. This is consistent with recommended SCTP practices.

8. Number of Streams

DDP streams are bidirectional. They are always composed by pairing the inbound and outbound SCTP streams with the same SCTP Stream Identifier.

The adaptation layer should request the maximum number of SCTP streams it will wish to use over the lifetime of the association. SCTP streams must still be bound to DDP Endpoints, and a DDP-enabled SCTP association does not support ordered Data Chunks. Therefore, the mere existence of an SCTP stream is unlikely to require significant supporting resources.

This mapping uses an SCTP association to carry one or more DDP streams. Each DDP stream will be mapped to a pair of SCTP streams with the same SCTP stream number. The adaptation MUST initialize all of its SCTP associations with the same number of inbound and outbound streams.

9. Fragmentation

A DDP/SCTP Receiver already deals with fragmentation at both the IP and DDP layers. Therefore, use of SCTP layer segmenting will be avoided for most cases.

As a Lower Layer Protocol (LLP) for DDP, the SCTP adaptation layer MUST inform the DDP layer of the maximum DDP Segment size that will be supported. This should be the largest value that can be supported without use of IP or SCTP fragmentation, or 516 bytes, whichever is larger.

A minimum of 516 bytes is required to allow a DDP Stream Session Control Message with 512 bytes of Private Data.

SCTP data chunk fragmentation MUST NOT be used unless the alternative is IP fragmentation.

The SCTP adaptation layer SHOULD set the maximum DDP Segment size below the theoretical maximum in order to allow bundling of Control Chunks in the same SCTP packet.

The SCTP adaptation layer MUST reject DDP Segments that are larger than the maximum size specified.

10. Sequenced Unordered Operation

The adaptation layer MUST use the Unordered option on all Data Chunks (U Flag set to one). The SCTP layer is expected to deliver unordered Data Chunks without delay.

Because DDP employs unordered SCTP delivery, the receiver MUST NOT rely upon the SCTP Transmission Sequence Number (TSN) to imply ordering of DDP Segments. The fact that the SCTP Data Chunk for a DDP Segment is prior to the cumulative ack point does not guarantee that all prior DDP segments have been placed. The SCTP sender is not obligated to transmit unordered Data Chunks in the order presented.

The DDP-SSN can be used without special logic to determine the submission sequence when the maximum number of in-flight messages is less than 32768. This also applies if the sending SCTP accepts no more than 32767 Data Chunks for a single stream without assigning TSNs.

If SCTP does accept more than 32768 Data chunks for a single stream without assigning TSNs, the sending DDP must simply refrain from sending more than 32767 Data Chunks for a single stream without acknowledgment. Note that it MUST NOT rely upon ULP flow control for this purpose. Typical ULP flow control will deal exclusively with untagged messages, not with DDP segments.

The receiver MAY perform a validity check on received DDP-SSNs to ensure that any gap could be accounted for by unreceived Data Chunks. Implementations SHOULD NOT allocate resources on the assumption that DDP-SSNs are valid without first performing such a validity check. An invalid DDP-SSN MAY result in termination of the DDP stream.

11. Procedures

11.1. Association Initialization

At the startup of an association, a DDP/SCTP adaptation layer MUST include an adaptation layer indication in its INIT or INIT-ACK (as defined in Section 5.1). After the exchange of the initial first two SCTP chunks (INIT and INIT-ACK), an endpoint MUST verify and inspect the Adaptation Indication and compare it to the following table to determine proper action.

Indication type	Action
NONE	This indicates that the peer DOES NOT support ANY DDP or RDMA adaptation, and thus RDMA and DDP procedures MUST NOT be performed upon this association.
DDP	This indicates that the peer DOES support the DDP/SCTP adaptation layer defined here.
ANY-OTHER Indication	This indicates that the peer DOES NOT support the DDP adaptation, and thus DDP procedures MUST NOT be performed upon this association.

An implementation MAY require that all associations for a given SCTP endpoint be placed in the same mode.

The local interface MAY allow the ULP to accept only requests to establish an association in a specified mode.

11.2. Chunk Bundling

SCTP allows multiple Data Chunks to be bundled in a single SCTP packet. Data chunks containing DDP Segments with untagged messages SHOULD NOT be delayed to facilitate bundling. Data chunks containing DDP Segments with tagged messages will generally be full sized, and hence not subject to bundling. However, partial-size tagged messages MAY be delayed, as they are frequently followed by a short untagged message.

11.3. Association Termination

Termination of an SCTP Association due to errors should be handled at the SCTP layer. The RDMAP-defined RDMAP Terminate Message SHOULD NOT be sent on each DDP stream when a determination has been made to terminate an SCTP association. Sending that message on each SCTP stream could severely delay the termination of the association.

The local interface SHOULD notify all consumers of DDP streams when the underlying SCTP stream has been terminated.

Other RDMAP-defined Terminate Messages MUST be generated as specified when a DDP stream is terminated. Note that with the SCTP mapping, termination of a DDP Stream does not mandate termination of the Association.

12. IANA Considerations

This document defines a new SCTP Adaptation Layer Indication codepoint for DDP (0x00000001). [RFC5061] creates the registry from which this codepoint has been assigned.

This document also defines two new SCTP Payload Protocol Identifiers (PPIDs). RFC 4960 [RFC4960] creates the registry from which these identifiers have been assigned. The following values have been assigned:

```
DDP Segment Chunk           - 16
DDP Stream Session Control  - 17
```

13. Security Considerations

Any direct placement of memory could pose a significant security risk if adequate local controls are not provided. These threats are addressed in the appropriate DDP [RFC5041], RDMA [RFC5040], or Security [RFC5042] documents. This document does not add any additional security risks over those found in RFC 4960 [RFC4960].

The IPsec requirements for Remote Direct Data Placement (RDDP) are based on the version of IPsec specified in RFC 2401 [RFC2401] and related RFCs, as profiled by RFC 3723 [RFC3723], despite the existence of a newer version of IPsec specified in RFC 4301 [RFC4301] and related RFCs. One of the important early applications of the RDDP protocols is their use with iSCSI iSER [RFC5046]; RDDP's IPsec requirements follow those of IPsec in order to facilitate that usage by allowing a common profile of IPsec to be used with iSCSI and the RDDP protocols. In the future, RFC 3723 may be updated to the newer version of IPsec; the IPsec security requirements of any such update should apply uniformly to iSCSI and the RDDP protocols.

Additional requirements apply to security for RDDP over SCTP, due to the use of SCTP as the transport protocol. An implementation of IPsec for RDDP over SCTP:

1) MUST support IPsec functionality for SCTP equivalent to the IPsec functionality for TCP that is required by RFC 3723,

2) SHOULD support the same level of IPsec functionality for SCTP and TCP unless there is no support for TCP, and

3) MUST support at least the level of protocol and port selector functionality for SCTP that is supported for TCP.

14. Contributors

Many thanks to our contributors who have spent many hours reading and reviewing keeping us straight: Sukanta Ganguly an independent consultant, Vivek Kashyap of IBM, Jim Pinkerton of Microsoft, and Hemal Shah of Broadcom. Thanks for all your hard work.

15. Acknowledgments

The authors would like to thank the following people that have provided comments and input: Stephen Bailey, David Black, Douglas Otis, Allyn Romanow, and Jim Williams.

16. References

16.1. Normative References

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

[RFC3723] Aboba, B., Tseng, J., Walker, J., Rangan, V., and F. Travostino, "Securing Block Storage Protocols over IP", RFC 3723, April 2004.

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

[RFC5040] Recio, R., Metzler, B., Culley, P., Hilland, J., and D. Garcia, "A Remote Direct Memory Access Protocol Specification", RFC 5040, October 2007.

[RFC5041] Shah, H., Pinkerton, J., Recio, R., and P. Culley, "Direct Data Placement over Reliable Transports", RFC 5041, October 2007.

[RFC5042] Pinkerton, J. and E. Deleganes, "Direct Data Placement Protocol (DDP) / Remote Direct Memory Access Protocol (RDMAP) Security", RFC 5042, October 2007.

[RFC5061] Stewart, R., Xie, Q., Tuexen, M., Maruyama, S., and M. Kozuka, "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration", RFC 5061, September 2007.

16.2. Informative References

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

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

[RFC5044] Culley, P., Elzur, U., Recio, R., Bailey, S., and J. Carrier, "Marker PDU Aligned Framing for TCP Specification", RFC 5044, October 2007.

[RFC5046] Ko, M., Chadalapaka, M., Elzur, U., Shah, H., and P. Thaler, "Internet Small Computer System Interface (iSCSI) Extensions for Remote Direct Memory Access (RDMA)", RFC 5046, October 2007.

Authors' Addresses

Caitlin Bestler (editor)
Neterion
20230 Stevens Creek Blvd.
Suite C
Cupertino, CA 95014
USA

Phone: 408-366-4639
EMail: caitlin.bestler@neterion.com

Randall R. Stewart (editor)
Cisco Systems, Inc.
Forest Drive
Columbia, SC 29036
USA

Phone: +1-815-342-5222
EMail: rrs@cisco.com

Full Copyright Statement

Intellectual Property