Network Working Group
Internet-Draft
Expires: August 26, 2003

R. Stewart
Cisco Systems, Inc.
C. Bestler
Consultant
J. Jim
Microsoft
S. Ganguly
Iomega Corp, Inc.
H. Shah
Intel Corporation
V. Kashyap
IBM

February 25, 2003

Stream Control Transmission Protocol (SCTP) Remote Direct Memory Access (RDMA) Direct Data Placement (DDP) Adaptation

draft-stewart-rddp-sctp-02.txt

Status of this Memo

Copyright Notice

Abstract

This document describes a method to adapt Direct Data Placement (DDP)

and Remote Direct Memory Access (RDMA) to Stream Control Transmission Protocol (SCTP) RFC2960 [2] using a generic description found in [RDMA-Draft] [4] and [DDP-Draft] [3].

Table of Contents

1. Introduction

This document describes a method to adapt Direct Data Placement (DDP) and Remote Direct Memory Access (RDMA) to Stream Control Transmission Protocol (SCTP) RFC2960 [2] using a generic description found in [RDMA-Draft] [4] and [DDP-Draft] [3] This adaptation provides a method for two peers to know that each side is performing DDP or RDMA thus enabling hardware acceleration if available.

Some implementations may include this adaptation layer within their SCTP implementations to obtain maximum performance but the behavior of SCTP will be unaffected. In order to accomplish this we specify the use of the new adaptation layer indication as defined in [ADDIP-Draft] [6]

1.1 Definitions

DDP stream - A bi-directional pair of SCTP streams which have the same SCTP Stream identifier.

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

SCTP Stream - A uni-directional logical channel established from one to another associated SCTP endpoint. An SCTP Stream is used to form one direction of a DDP stream.

1.2 Conventions

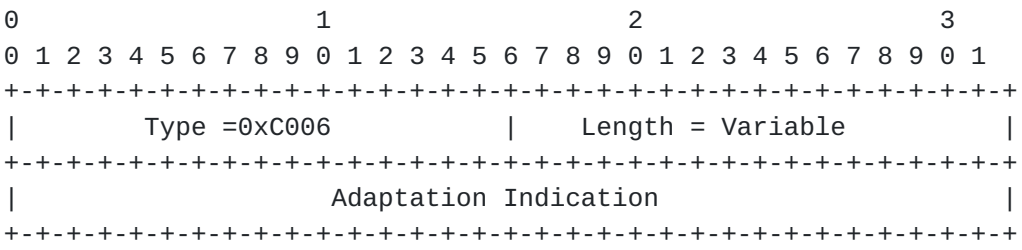
The keywords MUST, MUST NOT, REQUIRED, SHALL, SHALL NOT, SHOULD, SHOULD NOT, RECOMMENDED, NOT RECOMMENDED, MAY, and OPTIONAL, when they appear in this document, are to be interpreted as described in RFC2119 [1].

2. Data Formats

2.1 Adaptation Layer Indicator

This mapping places an entire SCTP association into a specific DDP mode: DDP or DDP+RDMA. 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 mix DDP and non-DDP streams in a single association. An application that needs to mix DDP and non-DDP traffic must use use more than a single association.

We define a adaptation indication which MUST appear in the INIT or INIT-ACK with the following format as defined in [ADDIP-Draft] [6]



Adaptation Indication:

The following values are defined for DDP in this document:

DDP	- 0x00000001
DDP+RDMA	- 0x00000002

The DDP 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.

2.2 Payload Protocol Identifier

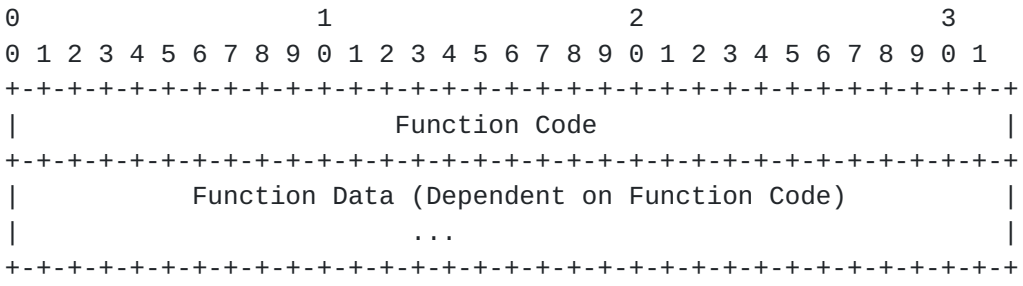
SCTP provides for delivery of user data messages. Each user message consists of a length, the data bytes and a Payload Protocol Identifier. The DDP SCTP adaptation uses two different Payload Protocol Identifiers.

Payload Protocol Identifier:

The following value are defined for DDP in this document:

DDP Message - 0x00000001
Adaptation Layer Control - 0x00000002

DDP Messages are as defined in [DDP-Draft]. Adaptation Layer Control messages are defined in this document.



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

DDP Stream Reset - 0x00000001

A DDP Stream Reset message MUST be presented to the SCTP layer with a Payload Protocol Identifer of one and a length of 4 bytes (i.e. the Function Code 0x00000001 with no Function dependant data).

3. SCTP Endpoints

3.1 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.

3.2 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 no additional bus traffic or duplication of memory registration resources. This will frequently result in a different default than for SCTP endpoints that are not DDP enabled.

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

4. 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.

DDP should request the maximum it will wish to use from SCTP. DDP Streams cannot be used without prior pre-posting of receive operations and/or enabling of STags. Therefore DDP will be able to initialize each stream on an "as needed" basis.

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

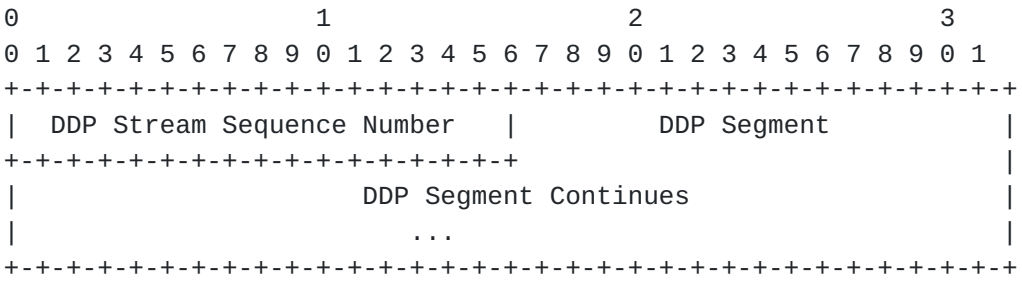
5. Fragmentation

A DDP/SCTP Receiver already must deal with fragementation at both the IP and DDP Layers. Therefore the sending DDP layer MUST disable SCTP layer segmenting of data chunks. If the DDP layer presents messages that are too large, the result will be IP fragmentation. While SCTP layer fragmentation is theoretically preferable, virtually all fragmentation will be done at the DDP layer. Because SCTP layer fragmentation would only be invoked under corner conditions, its benefits do not justify the complexity of its inclusion.

When disabling SCTP fragmentation, SCTP will reject messages that are known to be larger than the MTU size. This means that the DDP layer MUST be prepared to handle this error case.

6. Sequenced Unordered Operation

Each DDP Segment MUST be encoded within a single Data Chunk, along with a DDP Stream Sequence Number (DDP-SSN), as follows:



The DDP Stream Sequence Number represented the sequence of DDP segments sent for this DDP stream from one end. It is initialized to 1 when a DDP Stream is initialized or reset. It wraps to zero after 65535.

DDP MUST use the Unordered option on all Data Chunks (U Flag set to one). Each DDP segment within an SCTP Data Chunk may be placed immediately upon receipt from the SCTP layer.

A DDP segment is not deliverable until after it has been placed and all prior DDP Segments for the same DDP stream have been delivered.

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 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.

Note that no special logic is required on either end if the the maximum number of in-flight messages is less than 32768. No special DDP logic is required if the sending SCTP accepts no more than 32767 Data Chunks for a single stream without assigning SSNs.

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

7. Procedures

7.1 Association Initialization

At the startup of an association, an endpoint wishing to perform DDP, RDMA, or DDP+RDMA placement MUST include an adaptation layer indication in its INIT or INIT-ACK (as defined in Section 2.1. 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 DDP (but not RDMA). Procedures outlined in [DDP-Draft] MUST be followed.
DDP+RDMA	This indicates that the peer supports BOTH RDMA and DDP. If the receiving endpoint indicated the same, then the procedures in both [RDMA-Draft] and [DDP-Draft] MUST be followed. If the local endpoint only indicated DDP, then ONLY the procedures in [DDP-Draft] MUST be followed.
ANY-OTHER Indication	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.

7.2 Stream Reset

DDP [DDP-Draft] requires that a DDP Stream be aborted upon certain error conditions such as receiving an untagged message which the receiving side ULP had not enabled the reception of.

When a DDP stream is aborted, no further incoming packets will be

accepted for that stream until the stream is re-established. Many ULPs will maintain a session by re-establishing the DDP stream after such a termination.

Once a DDP Stream is declared to be aborted all DDP Messages on that stream MUST be discarded. Placement MUST NOT be performed. DDP Messages MUST NOT be delivered to the ULP. New DDP Messages from the ULP MUST NOT be accepted.

The aborted state MUST continue until a DDP Stream Reset Message is received. When this packet is received, the inbound SCTP stream will be re-enabled for normal handling of DDP Messages.

The DDP layer MAY send a DDP SCTP Stream Reset message in a Data Chunk to enable re-use of a Stream Identifier within an association for a new DDP Stream. However, it SHOULD select a previously unused stream first, if one is available.

The ability to re-use a Stream Identifier allows an SCTP association between two endpoints to remain open indefinitely. Isolated ULP faults will only impact the ULP components using the faulted stream, not those merely sharing the same association.

7.3 Chunk Bundling

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

7.4 STag Validation

STag validation is to be performed on a per stream basis. An integrated DDP/SCTP implementation MUST NOT enable an STag for an entire SCTP association merely because it is enabled for a single stream on that association. The ULP MUST be able to control STag enabling on a per stream basis, without regard to which SCTP association each stream is a part of.

8. IANA considerations

This document defines two new Adaptation Layer Indication codepoints:

```
    DDP                         - 0x00000001
    DDP+RDMA                    - 0x00000002
```

This document also defines two new Payload Protocol Identifier (PPIDs):

```
    DDP Message                 - 0x00000001
    Adaptation Layer Control    - 0x00000002
```

9. Security Considerations

Any direct placement of memory could pose a significant security risk if adequate local controls are not provided. These threats should be addressed in the appropriate DDP [DDP-Draft] [3] or RDMA [RDMA-Draft] [4] drafts. This document does not add any additional security risks over those found in RFC2960 [2].

10. Acknowledgments

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

References

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

[2] Stewart, R., Xie, Q., Morneault, K., Sharp, C., Schwarzbauer, H., Taylor, T., Rytina, I., Kalla, M., Zhang, L. and V. Paxson, "Stream Control Transmission Protocol", RFC 2960, October 2000.

[3] Culley, P., "Direct Data Placement over Reliable Transports", draft-shah-iwarp-ddp-00 (work in progress), September 2002.

[4] Recio, R., "An RDMA Protocol Specification", draft-recio-iwarp-rdma-01 (work in progress), November 2002.

[5] Stewart, R., "Sockets API Extensions for Stream Control Transmission Protocol", draft-ietf-tsvwg-sctpsocket-05 (work in progress), September 2002.

[6] Stewart, R., "Stream Control Transmission Protocol (SCTP) Dynamic Address Reconfiguration", draft-ietf-tsvwg-addip-sctp-06 (work in progress), September 2002.

Authors' Addresses

Randall R. Stewart
Cisco Systems, Inc.
8725 West Higgins Road
Suite 300
Chicago, IL 60631
USA

Phone: +1-815-477-2127
EMail: rrs@cisco.com

Caitlin Bestler
Consultant
1241 W. North Shore
2G
Chicago, IL 60626
USA

Phone: +1-773-743-1594
EMail: cait@asomi.com

Jim Pinkerton
Microsoft

Bellevue, Wa
USA

Phone: +1-xxx-xxx-xxx
EMail: jpink@microsoft.com

Sukanta Ganguly
Iomega Corp, Inc.
4435 Eastgate Mall
Suite 300
San Diego, CA 92121
USA

Phone: +1-858-795-7026
EMail: ganguly@iomega.com

Hemal V. Shah
Intel Corporation
Mailstop: PTL1
1501 S. Mopac Expressway, #400
Austin, TX 78746
USA

Phone: +1-512-732-3963
EMail: hemal.shah@intel.com

Vivek Kashyap
IBM
15450 SW Koll Parkway
Beaverton, OR 57006
USA

Phone: +1-503-578-3422
EMail: vivk@us.ibm.com

Intellectual Property Statement

Full Copyright Statement

Acknowledgement

Funding for the RFC Editor function is currently provided by the Internet Society.