Network Working Group
Request for Comments: 5045
Category: Informational

C. Bestler, Ed.
Neterion
L. Coene
Nokia Siemens Networks

October 2007

Applicability of Remote Direct Memory Access Protocol (RDMA) and Direct Data Placement Protocol (DDP)

Status of This Memo

Abstract

This document describes the applicability of Remote Direct Memory Access Protocol (RDMAP) and the Direct Data Placement Protocol (DDP). It compares and contrasts the different transport options over IP that DDP can use, provides guidance to ULP developers on choosing between available transports and/or how to be indifferent to the specific transport layer used, compares use of DDP with direct use of the supporting transports, and compares DDP over IP transports with non-IP transports that support RDMA functionality.

Table of Contents

1. Introduction

Remote Direct Memory Access Protocol (RDMAP) [RFC5040] and Direct Data Placement (DDP) [RFC5041] work together to provide application-independent efficient placement of application payload directly into buffers specified by the Upper Layer Protocol (ULP).

The DDP protocol is responsible for direct placement of received payload into ULP-specified buffers. The RDMAP protocol provides completion notifications to the ULP and support for Data-Sink-initiated fetch of Advertised Buffers (RDMA Reads).

DDP and RDMAP are both application-independent protocols that allow the ULP to perform remote direct data placement. DDP can use multiple standard IP transports including SCTP and TCP.

By clarifying the situations where the functionality of these protocols is applicable, this document can guide implementers and application and protocol designers in selecting which protocols to use.

The applicability of RDMAP/DDP is driven by their unique capabilities:

- This document will discuss when common data placement procedures are of more benefit to applications than application-specific solutions built on top of direct use of the underlying transport.

- DDP supports both Untagged and Tagged Buffers. Tagged Buffers allow the Data Sink ULP to be indifferent to what order (or in what messages) the Data Source sent the data, or in what order packets are received. Typically, tagged data can be used for payload transfer, while untagged is best used for control messages. However each upper-layer protocol can determine the optimal use of Tagged and Untagged Messages for itself. This document will discuss when Data Source flexibility is of benefit to applications.

- RDMAP consolidates ULP notifications, thereby minimizing the number of required ULP interactions.

- RDMAP defines RDMA Reads, which allow remote access to Advertised Buffers. This document will review the advantages of using RDMA Reads as contrasted to alternate solutions.

  A more comprehensive introduction to the RDMAP and DDP protocols and discussion of their security considerations can be found in [RFC5042].

Some non-IP transports, such as InfiniBand, directly integrate RDMA features. This document will review the applicability of providing RDMA services over ubiquitous IP transports instead of over customized transport protocols. Due to the fact that DDP is defined cleanly as a layer over existing IP transports, DDP has simpler ordering rules than some prior RDMA protocols. This may have some implications for application designers.

The full capabilities of DDP and RDMAP can only be fully realized by applications that are designed to exploit them. The coexistence of RDMAP/DDP-aware local interfaces with traditional socket interfaces will also be explored.

Finally, DDP support is defined for at least two IP transports: SCTP [RFC5043] and TCP [RFC5044]. The rationale for supporting both transports is reviewed, as well as when each would be the appropriate selection.

2. Definitions

Advertisement - the act of informing a Remote Peer that a local RDMA Buffer is available to it. A Node makes available an RDMA Buffer for incoming RDMA Read or RDMA Write access by informing its RDMA/DDP peer of the Tagged Buffer identifiers (STag, base address, and buffer length). This Advertisement of Tagged Buffer information is not defined by RDMA/DDP and is left to the ULP. A typical method would be for the Local Peer to embed the Tagged Buffer's Steering Tag, base address, and length in a Send Message destined for the Remote Peer.

Data Sink - The peer receiving a data payload. Note that the Data Sink can be required to both send and receive RDMA/DDP Messages to transfer a data payload.

Data Source - The peer sending a data payload. Note that the Data Source can be required to both send and receive RDMA/DDP Messages to transfer a data payload.

Lower Layer Protocol (LLP) - The transport protocol that provides services to DDP. This is an IP transport with any required adaptation layer. Adaptation layers are defined for SCTP and TCP.

Steering Tag (STag) - An identifier of a Tagged Buffer on a Node, valid as defined within a protocol specification.

Tagged Message - A DDP message that is directed to a ULP-specified buffer based upon imbedded addressing information. In the immediate sense, the destination buffer is specified by the message sender. The message receiver is given no independent indication that a Tagged Message has been received.

Untagged Message - A DDP message that is directed to a ULP-specified buffer based upon a Message Sequence Number being matched with a receiver-supplied buffer. The destination buffer is specified by the message receiver. The message receiver is notified by some mechanism that an Untagged Message has been received.

Upper Layer Protocol (ULP) - The direct user of RDMAP/DDP services. In addition to protocols such as iSER [RFC5046] and NFSv4 over RDMA [NFSDIRECT], the ULP may be embedded in an application or a middleware layer, as is often the case for the Sockets Direct Protocol (SDP) and Remote Procedure Call (RPC) protocols.

3. Direct Placement

Direct Data Placement optimizes the placement of ULP Payload into the correct destination buffers, typically eliminating intermediate copying. Placement is enabled without regard to order of arrival, order of transmission, or requirement of per-placement interaction with the ULP.

RDMAP minimizes the required ULP interactions. This capability is most valuable for applications that require multiple transport layer packets for each required ULP interaction.

3.1. Direct Placement Using Only the LLP

Direct data placement can be achieved without RDMA. Pre-posting of receive buffers could allow a non-RDMA network stack to place data directly to user buffers.

The degree to which DDP optimizes depends on which transport it is being compared with, and on the nature of the local interface. Without RDMAP/DDP, pre-posting buffers require the receiving side to accurately predict the required buffers and their sizes. This is not feasible for all ULPs. By contrast, DDP only requires the ULP to predict the sequence and size of incoming Untagged Messages.

An application that could predict incoming messages and required nothing more than direct placement into buffers might be able to do so with a properly designed local interface to native SCTP or TCP (without RDMA). This is easier using native SCTP because the

application would only have to predict the sequence of messages and the maximum size of each message, not the exact size.

The main benefit of DDP for such an application would be that pre-posting of receive buffers is a mandated local interface capability, and that predictions can always be made on a per-message basis (not per byte).

The Lower Layer Protocol, LLP, can also be used directly if ULP-specific knowledge is built into the protocol stack to allow "parse and place" handling of received packets. Such a solution either requires interaction with the ULP or the protocol stack's knowledge of ULP-specific syntax rules.

DDP achieves the benefits of directly placing incoming payload without requiring tight coupling between the ULP and the protocol stack. However, "parse and place" capabilities can certainly provide equivalent services to a limited number of ULPs.

3.2. Fewer Required ULP Interactions

While reducing the number of required ULP interactions is in itself desirable, it is critical for high-speed connections. The burst packet rate for a high-speed interface could easily exceed the host system's ability to switch ULP contexts.

Content access applications are important examples of applications that require high bandwidth and can transfer a significant amount of content between required ULP interactions. These applications include file access protocols (NAS), storage access (SAN), database access, and other application-specific forms of content access such as HTTP, XML, and email.

4. Tagged Messages

This section covers the major benefits from the use of Tagged Messages.

A more critical advantage of DDP is the ability of the Data Source to use Tagged Buffers. Tagging messages allows the Data Source to choose the ordering and packetization of its payload deliveries. With direct data placement based solely upon pre-posted receives, the packetization and delivery of payload must be agreed by the ULP peers in advance.

The Upper Layer Protocol can allocate content between Untagged and/or Tagged Messages to maximize the potential optimizations. Placing content within an Untagged Message can deliver the content in the

same packet that signals completion to the receiver. This can improve latency. It can even eliminate round trips. But it requires making larger anonymous buffers to be available.

Some examples of data that typically belongs in the Untagged Message would include:

- short fixed-size control data that is inherently part of the control message. This is especially true when the data is a required part of the control message.

- relatively short payload that is almost always needed, especially when its inclusion would eliminate a round-trip to fetch the data. Examples would include the initial data on a write request and Advertisements of Tagged Buffers.

Tagged Messages standardize direct placement of data without per-packet interaction with the upper layers. Even if there is an upper-layer protocol encoding of what is being transferred, as is common with middleware solutions, this information is not understood at the application-independent layers. The directions on where to place the incoming data cannot be accessed without switching to the ULP first. DDP provides a standardized 'packing list', which can be interpreted without requiring ULP interaction. Indeed, it is designed to be implementable in hardware.

4.1. Order-Independent Reception

Tagged Messages are directed to a buffer based on an included Steering Tag. Additionally, no notice is provided to the ULP for each individual Tagged Message's arrival. Together these allow Tagged Messages received out of order to be processed without intermediate buffering or additional notifications to the ULP.

4.2. Reduced ULP Notifications

RDMAP offers both Tagged and Untagged Messages. No receiving-side ULP interactions are required for Tagged Messages. By optimally dividing traffic between Tagged and Untagged Messages, the ULP can limit the number of events that must be dealt with at the ULP layer. This typically reduces the number of context switches required and improves performance.

RDMAP further reduces required ULP interactions, consolidating completion notifications of Tagged Messages with the completion notification of a trailing Untagged Message. For most ULPs, this radically reduces the number of ULP required interactions even further.

While RDMAP consolidation of notices is beneficial to most applications, it may be detrimental to some applications that benefit from streamed delivery to enable ULP processing of received data as promptly as possible. A ULP that uses RDMAP cannot begin processing any portion of an exchange until it receives notification that the entire exchange has been placed. An "exchange" here is a set of zero or more Tagged Messages and a single terminating Untagged Message. An application that would prefer to begin work on the received payload as soon as possible, no matter what order it arrived in, might prefer to work directly with the LLP. RDMAP is optimized for applications that are more concerned when the entire exchange is complete.

An application that benefits from being able to begin processing of each received packet as quickly as possible may find RDMAP interferes with that goal.

Such an application might be able to retain most of the benefits of RDMAP by using the DDP layer directly. However, in addition to taking on the responsibilities of the RDMAP layer, the application would likely have more difficulty finding support for a DDP-only API. Many hardware implementations may choose to tightly couple RDMAP and DDP, and might not provide an API directly to DDP services.

These features minimize the required interactions with the ULP. This can be extremely beneficial for applications that use multiple transport layer packets to accomplish what is a single ULP interaction.

4.3. Simplified ULP Exchanges

The notification rules for Tagged Messages allows ULPs to create multi-message "exchanges" consisting of zero or more Tagged Messages that represent a single step in the ULP interaction. The receiving ULP is notified that the Untagged Message has arrived, and implicitly notified of any associated Tagged Messages.

If a ULP cannot effectively use Tagged Messages, it would derive little benefit from use of RDMAP/DDP by comparison to direct use of SCTP. But, while Tagged Buffers are the justification for RDMAP/DDP, Untagged Buffers are still necessary. Without Untagged Buffers, the only method to exchange buffer Advertisements would require out-of-band communications. Most RDMA-aware ULPs use Untagged Buffers for requests and responses. Buffer Advertisements are typically done within these Untagged Messages.

More importantly, there would be no reliable method for the upper-layer peers to synchronize. The absence of any guarantees about

ordering within or between Tagged Messages is fundamental to allowing the DDP layer to optimize transfer of tagged payload.

Therefore, no ULP can be defined entirely in terms of Tagged Messages. Eventually, a notification that confirms delivery must be generated from the RDMAP/DDP layer.

Limiting use of Untagged Buffers to requests and responses by moving all bulk data using tagged transfers can greatly simplify the amount of prediction that the Data Sink must perform in pre-posting receive buffers. For example, a typical RDMA-enabled interaction would consist of the following:

1. Client sends transaction request to server as an Untagged Message.

2. This message includes buffer Advertisements for the buffers where the results are to be placed.

3. The server sends multiple Tagged Messages to the Advertised buffers.

4. The server sends transaction reply as an Untagged Message to the client.

5. Client receives single notification, indicating completion of the interaction.

With this type of exchange, the pacing and required size of Untagged Buffers are highly predictable. The variability of response sizes is absorbed by tagged transfers.

4.4. Order-Independent Sending

Use of Tagged Messages is especially applicable when the Data Sink does not know the actual size, structure, or location of the content it is requesting (or updating).

For example, suppose the Data Sink ULP needs to fetch four related pieces of data into four separate buffers. With SCTP, the Data Sink ULP could receive four messages into four separate buffers, only having to predict the maximum size of each. However, it would have to dictate the order in which the Data Source supplied the separate pieces. If the Data Source found it advantageous to fetch them in a different order, it would have to use intermediate buffering to re-order the pieces into the expected order even though the application only required that all four be delivered and did not truly have an ordering requirement.

Techniques, such as RAID striping and mirroring, represent this same problem, but one step further. What appears to be a single resource to the Data Sink is actually stored in separate locations by the Data Source. Non RDMA protocols would either require the Data Source to fetch the material in the desired order or force the Data Source to use its own holding buffers to assemble an image of the destination buffer.

While sometimes referred to as a "buffer-to-buffer" solution, RDMA more fundamentally enables remote buffer access. The ULP is free to work with larger remote buffers than it has locally. This reduces buffering requirements and the number of times the data must be copied in an end-to-end transfer.

There are numerous reasons why the Data Sink would not know the true order or location of the requested data. It could be different for each client, different records selected and/or different sort orders, as well as RAID striping, file fragmentation, volume fragmentation, volume mirroring, and server-side dynamic compositing of content (such as server-side includes for HTTP).

In all of these cases, the Data Source is free to assemble the desired data in the Data Sink's buffer in whatever order the component data becomes available to it. It is not constrained on ordering. It does not have to assemble an image in its own memory before creating it in the Data Sink's buffers.

Note that while DDP enables use of Tagged Messages for bulk transfer, there are some application scenarios where Untagged Messages would still be used for bulk transfer. For example, a file server may not expose its own memory to its clients. A client wishing to write may Advertise a buffer upon which the server will issue RDMA Reads. However, when performing a small write, it may be preferable to include the data in the Untagged Message rather than incurring an additional round trip with the RDMA Read and its response.

Generally, the best use of an Untagged Message is to synchronize and to deliver data that is naturally tied to the same message as the synchronization. For initial data transfers, this has the additional benefit of avoiding the need to Advertise specific Tagged Buffers for indefinite time periods. Instead, anonymous buffers can be used for initial data reception. Because anonymous buffers do not need to be tied to specific messages in advance, this can be a major benefit.

4.5. Untagged Messages and Tagged Buffers as ULP Credits

The handling of end-to-end buffer credits differs considerably with DDP than when the ULP directly uses either TCP or SCTP.

With both TCP and SCTP, buffer credits are based upon the receiver granting transmit permission based on the total number of bytes. These credits reflect system buffering resources and/or simple flow control. They do not represent ULP resources.

DDP defines no standard flow control, but presumes the existence of a ULP mechanism. The presumed mechanism is that the Data Sink ULP has issued credits to the Data Source, allowing the Data Source to send a specific number of Untagged Messages.

The ULP peers must ensure that the sender is aware of the maximum size that can be sent to any specific target buffer. One method of doing so is to use a standard size for all Untagged Buffers within a given connection. For example, a ULP may specify an initial Untagged Buffer size to be used immediately after session establishment, and then optionally specify mechanisms for negotiating changes.

Tagged Buffers are ULP resources Advertised directly from ULP to ULP. A DDP put to a known Tagged Buffer is constrained only by transport level flow control, not by available system buffering.

Either Tagged or Untagged Buffers allows bypassing of system buffer resources. Use of Tagged Buffers additionally allows the Data Source to choose in what order to exercise the credits.

To the extent allowed by the ULP, Tagged Buffers are also divisible resources. The Data Sink can Advertise a single 100 KB buffer, and then receive notifications from its peer that it had written 50 KB, 20 KB, and 30 KB to that buffer in three successive transactions.

ULP management of Tagged Buffer resources, independent of transport and DDP layer credits, is an additional benefit of RDMA protocols. Large bulk transfers cannot be blocked by limited general-purpose buffering capacity. Applications can flow control based upon higher level abstractions, such as number of outstanding requests, independent of the amount of data that must be transferred.

However, use of system buffering, as offered by direct use of the underlying transports, can be preferable under certain circumstances.

One example would be when the number of target ULP Buffers is sufficiently large, and the rate at which any writes arrive is sufficiently low, that pinning all the target ULP Buffers in memory would be undesirable. The maximum transfer rate, and hence the maximum amount of system buffering required, may be more stable and predictable than the total ULP Buffer exposure.

Another example would be when the Data Sink wishes to receive a stream of data at a predictable rate, but does not know in advance what the size of each data packet will be. This is common from streaming media that has been encoded with a variable bit rate. With DDP, the Data Sink would either have to use Untagged Buffers large enough for the largest packet, or Advertise a circular buffer. If, for security or other reasons, the Data Sink did not want the size of its buffer to be publicly known, using the underlying SCTP transport directly may be preferable because of its byte-oriented credits.

5. RDMA Read

RDMA Reads are a further service provided by RDMAP. RDMA Reads allow the Data Sink to fetch exactly the portion of the peer ULP Buffer required on a "just in time" basis. This can be done without requiring per-fetch support from the Data Source ULP.

Storage servers may wish to limit the maximum write buffer allocated to any single session. The storage server may be a very minimal layer between the client and the disk storage media, or the server may merely wish to limit the total resources that would be required if all clients could push the entire payload they wished written at their own convenience.

In either case, there is little benefit in transferring data from the Data Source far in advance of when it will be written to the persistent storage media. RDMA Reads allow the Storage Server to fetch the payload on a "just in time" basis. In this fashion, a relatively small number of block-sized buffers can be used to execute a single transaction that specified writing a large file, or a Storage Server with numerous clients can fetch buffers from the individual clients in the order that is most convenient to the server.

This same capability can be used when the desired portion of the Advertised Buffer is not known in advance. For example, the Advertised Buffer could contain performance statistics. The Data Sink could request the portions of the data it required, without requiring an interaction with the Data Source ULP.

This is applicable for many applications that publish semi-volatile data that does not require transactional validity checking (i.e., authorized users have read access to the entire set of data). It is less applicable when there are ULP consistency checks that must be performed upon the data. Such applications would be better served by having the client send a request, and having the server use RDMA Writes to publish the requested data. Neither RDMAP nor DDP provide mechanisms for bundling multiple disjoint updates into an atomic

operation. Therefore, use of an Advertised Buffer as a data resource is subject to the same caveats as any randomly updated data resource, such as flat files, that do not enforce their own consistency.

6. LLP Comparisons

Normally, the choice of underlying IP transport is irrelevant to the ULP. RDMAP and DDP provides the same services over either. There may be performance impacts of the choice, however. It is the responsibility of the ULP to determine which IP transport is best suited to its needs.

SCTP provides for preservation of message boundaries. Each DDP Segment will be delivered within a single SCTP packet. The equivalent services are only available with TCP through the use of the MPA (Marker PDU Alignment) adaptation layer.

6.1. Multistreaming Implications

SCTP also provides multi-streaming. When the same pair of hosts have need for multiple DDP streams, this can be a major advantage. A single SCTP association carries multiple DDP streams, consolidating connection setup, congestion control, and acknowledgements.

Completions are controlled by the DDP Source Sequence Number (DDP-SSN) on a per-stream basis. Therefore, combining multiple DDP Streams into a single SCTP association cannot result in a dropped packet carrying data for one stream delaying completions on others.

6.2. Out-of-Order Reception Implications

The use of unordered Data Chunks with SCTP guarantees that the DDP layer will be able to perform placements when IP datagrams are received out of order.

Placement of out-of-order DDP Segments carried over MPA/TCP is not guaranteed, but certainly allowed. The ability of the MPA receiver to process out-of-order DDP Segments may be impaired when alignment of TCP segments and MPA FPDUs is lost. Using SCTP, each DDP Segment is encoded in a single Data Chunk and never spread over multiple IP datagrams.

6.3. Header and Marker Overhead

MPA and TCP headers together are smaller than the headers used by SCTP and its adaptation layer. However, this advantage can be reduced by the insertion of MPA markers. The difference in ULP Payload per IP Datagram is not likely to be a significant factor.

6.4. Middlebox Support

Even with the MPA adaptation layer, DDP traffic carried over MPA/TCP will appear to all network middleboxes as a normal TCP connection. In many environments, there may be a requirement to use only TCP connections to satisfy existing network elements and/or to facilitate monitoring and control of connections. While SCTP is certainly just as monitorable and controllable as TCP, there is no guarantee that the network management infrastructure has the required support for both.

6.5. Processing Overhead

A DDP stream delivered via MPA/TCP will require more processing effort than one delivered over SCTP. However, this extra work may be justified for many deployments where full SCTP support is unavailable in the endpoints of the network, or where middleboxes impair the usability of SCTP.

6.6. Data Integrity Implications

Both the SCTP [RFC4960] and MPA/TCP [RFC5044] adaptation provide end-to-end CRC32c protection against data accidental corruption, or its equivalent.

A ULP that requires a greater degree of protection may add its own. However, DDP and RDMAP headers will only be guaranteed to have the equivalent of end-to-end CRC32c protection. A ULP that requires data integrity checking more thorough than an end-to-end CRC32c should first invalidate all STags that reference a buffer before applying its own integrity check.

CRC32c only provides protection against random corruption. To protect against unauthorized alteration or forging of data packets, security methods must be applied. The RDMA security document [RFC5042] specifies usage of RFC 2406 [RFC2406] for both adaptation layers. As stated in [RFC5042], note that the IPsec requirements for 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.

6.6.1. MPA/TCP Specifics

It is mandatory for MPA/TCP implementations to implement CRC32c, but it is not mandatory to use the CRC32c during an RDMA connection. The activating or deactivating of the CRC in MPA/TCP is an administrative configuration operation at the local and remote end. The administration of the CRC (ON/OFF) is invisible to the ULP.

Applications should assume that disabling CRC32c will only be used when the end-to-end protection is at least as effective as a transport layer CRC32c. Applications should not use additional integrity checks based solely on the possibility that CRC32c could be disabled without equivalent integrity checks at a lower level.

CRC32c must not be disabled unless equivalent or better end-to-end integrity protection is provided.

If the CRC is active/used for one direction/end, then the use of the CRC is mandatory in both directions/ends.

If both ends have been configured not to use the CRC, then this is allowed as long as an equivalent protection (comparable to or better than CRC) from undetected errors on the connection is provided.

6.6.2. SCTP Specifics

SCTP provides CRC32c protection automatically. The adaptation to SCTP provides for no option to suppress SCTP CRC32c protection.

6.7. Non-IP Transports

DDP is defined to operate over ubiquitous IP transports such as SCTP and TCP. This enables a new DDP-enabled node to be added anywhere to an IP network. No DDP-specific support from middleboxes is required.

There are non-IP transport fabric offering RDMA capabilities. Because these capabilities are integrated with the transport protocol they have some technical advantages when compared to RDMA over IP. For example, fencing of RDMA Operations can be based upon transport level acks. Because DDP is cleanly layered over an IP transport, any explicit RDMA layer ack must be separate from the transport layer ack.

There may be deployments where the benefits of RDMA/transport integration outweigh the benefits of being on an IP network.

6.7.1. No RDMA-Layer Ack

DDP does not provide for its own acknowledgements. The only form of ack provided at the RDMAP layer is an RDMA Read Response. DDP and RDMAP rely almost entirely upon other layers for flow control and pacing. The LLP is relied upon to guarantee delivery and avoid network congestion, and ULP-level acking is relied upon for ULP pacing and to avoid ULP Buffer overruns.

Previous RDMA protocols, such as InfiniBand, have been able to use their integration with the transport layer to provide stronger ordering guarantees. It is important that application designers that require such guarantees provide them through ULP interaction.

Specifically:

There is no ability for a local interface to "fence" outbound messages to guarantee that prior Tagged Messages have been placed prior to sending a Tagged Message. The only guarantees available from the other side would be an RDMA Read Response (coming from the RDMAP layer) or a response from the ULP layer. Remember that the normal ordering rules only guarantee when the Data Sink ULP will be notified of Untagged Messages; it does not control when data is placed into receive buffers.

Re-use of Tagged Buffers must be done with extreme care. The fact that an Untagged Message indicates that all prior Tagged Messages have been placed does not guarantee that no later Tagged Message has. The best strategy is to change only the state of any given Advertised Buffers with Untagged Messages.

As covered elsewhere in this document, flow control of Untagged Messages is the responsibility of the ULP.

6.8. Other IP Transports

Both TCP and SCTP provide DDP with reliable transport with TCP-friendly rate control. Currently, DDP is defined to work over reliable transports and implicitly relies upon some form of rate control.

DDP is fully compatible with a non-reliable protocol. Out-of-order placement is obviously not dependent on whether the other DDP Segments ever actually arrive.

However, RDMAP requires the LLP to provide reliable service. An alternate completion handling protocol would be required if DDP were to be deployed over an unreliable IP transport.

As noted in the prior section on Tagged Buffers as ULP credits, neither RDMAP nor DDP provides any flow control for Tagged Messages. If no transport layer flow control is provided, an RDMAP/DDP application would be limited only by the link layer rate, almost inevitably resulting in severe network congestion.

RDMAP encourages applications to be ignorant of the underlying transport path MTU. The ULP is only notified when all messages ending in a single Untagged Message have completed. The ULP is not aware of the granularity or ordering of the underlying message. This approach assumes that the ULP is only interested in the complete set of messages, and has no use for a subset of them.

6.9. LLP-Independent Session Establishment

For an RDMAP/DDP application, the transport services provided by a pair of SCTP streams and by a TCP connection both provide the same service (reliable delivery of DDP Segments between two connected RDMAP/DDP endpoints).

6.9.1. RDMA-Only Session Establishment

It is also possible to allow for transport-neutral establishment of RDMAP/DDP sessions between endpoints. Combined, these two features would allow most applications to be unconcerned as to which LLP was actually in use.

Specifically, the procedures for DDP Stream Session establishment discussed in section 3 of the SCTP mapping, and section 13.3 of the MPA/TCP mapping, both allow for the exchange of ULP-specific data ("Private Data") before enabling the exchange of DDP Segments. This delay can allow for proper selection and/or configuration of the endpoints based upon the exchanged data. For example, each DDP Stream Session associated with a single client session might be assigned to the same DDP Protection Domain.

To be transport neutral, the applications should exchange Private Data as part of session establishment messages to determine how the RDMA endpoints are to be configured. One side must be the Initiator, and the other, the Responder.

With SCTP, a pair of SCTP streams can be used for successive sessions while the SCTP association remains open. With MPA/TCP, each connection can be used for, at most, one session. However, the same source/destination pair of ports can be re-used for a subsequent TCP connection, as allowed by TCP.

Both SCTP and MPA limit the private data size to a maximum of 512 bytes.

MPA/TCP requires the end of the TCP connection that initiated the conversion to MPA mode to send the first DDP Segment. SCTP does not have this requirement. ULPs that wish to be transport neutral should require the initiating end to send the first message. A zero-length RDMA Write can be used for this purpose if the ULP logic itself does naturally support this restriction.

6.9.2. RDMA-Conditional Session Establishment

It is sometimes desirable for the active side of a session to connect with the passive side before knowing whether the passive side supports RDMA.

This style of session establishment can be supported with either TCP or SCTP, but not as transparently as for RDMA-only sessions. Pre-existing non-RDMA servers are also far more likely to be using TCP than SCTP.

With TCP, a normal TCP connection is established. It is then used by the ULP to determine whether or not to convert to MPA mode and use RDMA. This will typically be integral with other session-establishment negotiations.

With SCTP, the establishment of an association tests whether RDMA is supported. If not supported, the application simply requests the association without the RDMA adaptation indication.

One key difference is that with SCTP the determination as to whether the peer can support RDMA is made before the transport layer association/connection is established, while with TCP the established connection itself is used to determine whether RDMA is supported.

7. Local Interface Implications

Full utilization of DDP and RDMAP capabilities requires a local interface that explicitly requests these services. Protocols such as Sockets Direct Protocol (SDP) can allow applications to keep their traditional byte-stream or message-stream interface and still enjoy many of the benefits of the optimized wire level protocols.

8. Security Considerations

RDMA security considerations are discussed in the RDMA security document [RFC5042]. This document will only deal with the more usage-oriented aspects, and where there are implications in the choice of underlying transport.

8.1. Connection/Association Setup

Both the SCTP and TCP adaptations allow for existing procedures to be followed for the establishment of the SCTP association or TCP connection. Use of DDP does not impair the use of any security measures to filter, validate, and/or log the remote end of an association/connection.

8.2. Tagged Buffer Exposure

DDP only exposes ULP memory to the extent explicitly allowed by ULP actions. These include posting of receive operations and enabling of Steering Tags.

Neither RDMAP nor DDP places requirements on how ULPs Advertise Buffers. A ULP may use a single Steering Tag for multiple buffer Advertisements. However, the ULP should be aware that enforcement on STag usage is likely limited to the overall range that is enabled. If the Remote Peer writes into the 'wrong' Advertised Buffer, neither the DDP nor the RDMAP layer will be aware of this. Nor is there any report to the ULP on how the Remote Peer specifically used Tagged Buffers.

Unless the ULP peers have an adequate basis for mutual trust, the receiving ULP might be well advised to use a distinct STag for each interaction, and to invalidate it after each use, or to require its peer to use the RDMAP option to invalidate the STag with its responding Untagged Message.

8.3. Impact of Encrypted Transports

While DDP is cleanly layered over the LLP, its maximum benefit may be limited when the LLP Stream is secured with a streaming cypher, such as Transport Layer Security (TLS) [RFC4346]. If the LLP must decrypt in order, it cannot provide out-of-order DDP Segments to the DDP layer for placement purposes. IPsec [RFC2401] tunnel mode encrypts entire IP Datagrams. IPsec transport mode encrypts TCP Segments or SCTP packets, as does use of Datagram TLS (DTLS) [RFC4347] over UDP beneath TCP or SCTP. Neither IPsec nor this use of DTLS precludes providing out-of-order DDP Segments to the DDP layer for placement.

Note that end-to-end use of cryptographic integrity protection may allow suppression of MPA CRC generation and checking under certain circumstances. This is one example where the LLP may be judged to have "or equivalent" protection to an end-to-end CRC32c.

9. References

9.1. Normative References

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

[RFC2406] Kent, S. and R. Atkinson, "IP Encapsulating Security Payload (ESP)", RFC 2406, November 1998.

[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, "DDP/RDMAP Security", RFC 5042, October 2007.

[RFC5043] Bestler, C. and R. Stewart, "Stream Control Transmission Protocol (SCTP) Direct Data Placement (DDP) Adaptation", RFC 5043, October 2007.

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

9.2. Informative References

[NFSDIRECT] Talpey, T., Callaghan, B., and I. Property, "NFS Direct Data Placement", Work in Progress, June 2007.

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

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

[RFC4346] Dierks, T. and E. Rescorla, "The Transport Layer Security (TLS) Protocol Version 1.1", RFC 4346, April 2006.

[RFC4347] Rescorla, E. and N. Modadugu, "Datagram Transport Layer Security", RFC 4347, April 2006.

[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

Lode Coene
Nokia Siemens Networks
Atealaan 26
Herentals 2200
Belgium

Phone: +32-14-252081
EMail: lode.coene@nsn.com

Full Copyright Statement

Intellectual Property