Internet Engineering Task Force (IETF)
Request for Comments: 7306
Category: Standards Track
ISSN: 2070-1721

H. Shah
Broadcom Corporation
F. Marti
W. Noureddine
A. Eiriksson
Chelsio Communications, Inc.
R. Sharp
Intel Corporation

June 2014

Remote Direct Memory Access (RDMA) Protocol Extensions

Abstract

This document specifies extensions to the IETF Remote Direct Memory Access Protocol (RDMAP) as specified in RFC 5040. RDMAP provides read and write services directly to applications and enables data to be transferred directly into Upper-Layer Protocol (ULP) Buffers without intermediate data copies. The extensions specified in this document provide the following capabilities and/or improvements: Atomic Operations and Immediate Data.

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/rfc7306.

Table of Contents

1. Introduction

The RDMA Protocol [RFC5040] provides capabilities for zero-copy data communications that preserve memory protection semantics, enabling more efficient network protocol implementations. The RDMA Protocol is part of the iWARP family of specifications which also include RFC 5041 [RFC5041], RFC 5044 [RFC5044], and RFC 6581 [RFC6581]. This document specifies the following extensions to the RDMA Protocol (RDMAP):

- Atomic Operations can be performed on remote memory locations. Support for Atomic Operations enhances the usability of RDMAP in distributed shared-memory environments.

- Immediate Data messages allow the ULP at the sender to provide a small amount of data. When an Immediate Data message is sent following an RDMA Write Message, the combination of the two messages is an implementation of RDMA Write with Immediate message that is found in other RDMA transport protocols.

Other RDMA transport protocols define the functionality added by these extensions leading to differences in RDMA applications and/or Upper-Layer Protocols. Removing these differences in the transport protocols simplifies these applications and ULPs, and that is the main motivation for the extensions specified in this document.

RSockets [RSOCKETS] is an example of RDMA-enabled middleware that provides a socket interface as the upper-edge interface and utilizes RDMA to provide more efficient networking for socket-based applications. RSockets is aware of Immediate Data support in InfiniBand [IB]. RSockets cannot utilize the RDMA Write with Immediate Data operation from InfiniBand. The addition of the Immediate Data operation specified in this document will alleviate this difference in RSockets when running on InfiniBand and iWARP.

Structured high-performance computing applications based on the Message-Passing Interface [MPI] may use Atomic Operations defined in this specification. DAT Atomics [DAT_ATOMICS] is an example of RDMA-enabled middleware that provides a portable RDMA programming interface for various RDMA transport protocols. DAT Atomics includes a primitive for InfiniBand that is not supported by iWARP RDMA-enabled Network Interface Controllers or RNICs. The addition of Atomic Operations as specified in this document will allow Atomic Operations in DAT Atomics to work for both InfiniBand and RNICs interchangeably.

For more background on RDMA Protocol applicability, see "Applicability of Remote Direct Memory Access Protocol (RDMA) and Direct Data Placement Protocol (DDP)" [RFC5045].

1.1. Discovery of RDMAP Extensions

Today there are RDMA applications and/or ULPs that are aware of the existence of Atomic and Immediate Data operations for RDMA transports such as InfiniBand and application programming interfaces such as Open Fabrics Verbs [OFAVERBS]. Today, these applications need to be aware that RDMAP does not support certain of these operations. Typically, the availability of these capabilities is exposed to the applications through adapter query interfaces in software. Applications then have to decide to use or not use Immediate Data or Atomic Operations based on the results of the query interfaces. Such query interfaces typically return the scope of atomicity guarantees, not the individual Atomic Operations supported. Therefore, this specification requires all Atomic Operations defined within to be supported if an RNIC supports any Atomic Operations.

In cases where heterogeneous hardware, with differing support for Atomic Operations and Immediate Data Operations, is deployed for use by RDMA applications and/or ULPs, applications are either statically configured to use or not use optional features or use application-specific negotiation mechanisms. For the extensions covered by this document, it is RECOMMENDED that RDMA applications and/or ULPs negotiate at the application or ULP level the usage of these extensions. The definition of such application-specific mechanisms is outside the scope of this specification. For backward compatibility, existing applications and/or ULPs should not assume that these extensions are supported.

In the absence of application-specific negotiation of the features defined within this specification, the new operations can be attempted, and reported errors can be used to determine a remote peer’s capabilities. In the case of Atomics, a FetchAdd operation with "Add Data" set to 0 can safely be used to determine the existence of Atomic Operations without modifying the content of a remote peer’s memory. A Remote Operation Error or Unexpected OpCode error will be reported by the remote peer if there is an Immediate Data or Atomic Operation that is not supported by the remote peer.

2. Requirements Language

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].

3. Glossary

This document is an extension of RFC 5040, and key words are defined in the glossary of that document.

Atomic Operation - an operation that results in an execution of a memory operation at a specific ULP Buffer address on a remote node using the Tagged Buffer data transfer model. The consumer can use Atomic Operations to read, modify, and write memory at the destination ULP Buffer address, while at the same time guaranteeing that no other Atomic Operation read or write accesses to the ULP Buffer address targeted by the Atomic Operation will occur across any other RDMAP Streams on an RNIC at the Responder.

Atomic Operation Request - an RDMA Message used by the Data Source to perform an Atomic Operation at the Responder.

Atomic Operation Response - an RDMA Message used by the Responder to describe the completion of an Atomic Operation at the Responder.

CmpSwap - an Atomic Operation that is used to compare and swap a value at a specific address on a remote node.

FetchAdd - an Atomic Operation that is used to atomically increment a value at a specific ULP Buffer address on a remote node.

Immediate Data - a small fixed-size portion of data sent from the Data Source to a Data Sink.

Immediate Data Message - an RDMA Message used by the Data Source to send Immediate Data to the Data Sink.

Immediate Data with Solicited Event (SE) Message - an RDMA Message used by the Data Source to send Immediate Data with Solicited Event to the Data Sink.

iWARP - a suite of wire protocols comprised of RFC 5040, RFC 5041, RFC 5044, and RFC 6581.

Requester - the sender of an RDMA Atomic Operation request.

Responder - the receiver of an RDMA Atomic Operation request.

RNIC - RDMA-enabled Network Interface Controller. In this context, this would be a network I/O adapter or embedded controller with iWARP functionality.

ULP - Upper-Layer Protocol. The protocol layer above the one currently being referenced. The ULP for RFC 5040 / RFC 5041 is expected to be an OS, Application, adaptation layer, or proprietary device. The RFC 5040 / RFC 5041 documents do not specify a ULP -- they provide a set of semantics that allow a ULP to be designed to utilize RFC 5040 / RFC 5041.

4. Header Format Extensions

The control information of RDMA Messages is included in header fields defined in RFC 5041, the Direct Data Placement (DDP) protocol. RFC 5040 defines the RDMAP header formats layered on the DDP header definition. This specification extends RFC 5040 with the following new formats:

- Four new RDMA Messages carry additional RDMAP headers. The Immediate Data operation and Immediate Data with Solicited Event operation each include 8 bytes of data following the RDMAP header. Atomic Operations include Atomic Request or Atomic Response headers following the RDMAP header. The RDMAP header for Atomic Request messages is 52 bytes long as specified in Figure 4. The RDMAP header for Atomic Response Messages is 32 bytes long as specified in Figure 5.

- Introduction of a new queue for untagged Buffers (QN=3) used for Atomic Response tracking.

4.1. RDMAP Control and Invalidate STag Fields

For reference, Figure 1 depicts the format of the DDP Control and RDMAP Control Fields, in the style and convention of RFC 5040:

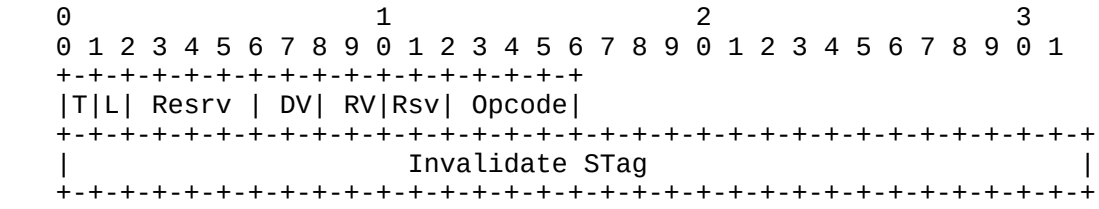


Figure 1: DDP Control and RDMAP Control Fields

The DDP Control Field consists of the T (Tagged), L (Last), Resrv, and DV (DDP protocol Version) fields [RFC5041]. The RDMAP Control Field consists of the RV (RDMA Version), Rsv, and Opcode fields [RFC5040].

This specification adds values for the RDMA Opcode field to those specified in RFC 5040. Figure 2 defines the new values of the RDMA Opcode field that are used for the RDMA Messages defined in this specification.

As shown in Figure 2, STag (Steering Tag) and Tagged Offset are not applicable for the RDMA Messages defined in this specification. Figure 2 also shows the appropriate Queue Number for each Opcode.

All RDMA Messages defined in this specification MUST have:

The RDMA Version (RV) field: 01b.

Opcode field: Set to one of the values in Figure 2.

Invalidate STag: Set to zero by the sender, ignored by the receiver.

RDMA Opcode	Message Type	Tagged Flag	STag and TO	Queue Number	In-validate STag	Message Length Communicated between DDP and RDMAP
1000b	Immediate Data	0	N/A	0	N/A	Yes
1001b	Immediate Data with SE	0	N/A	0	N/A	Yes
1010b	Atomic Request	0	N/A	1	N/A	Yes
1011b	Atomic Response	0	N/A	3	N/A	Yes

Figure 2: Additional RDMA Usage of DDP Fields

Note: N/A means Not Applicable.

This extension defines RDMAP use of Queue Number 3 for Untagged Buffers for Atomic Responses. This queue is used for tracking outstanding Atomic Requests.

All other DDP and RDMAP Control Fields are set as described in RFC 5040.

4.2. RDMA Message Definitions

The following figure defines which RDMA Headers are used on each new RDMA Message and which new RDMA Messages are allowed to carry ULP payload.

RDMA Message OpCode	Message Type	RDMA Header Used	ULP Message allowed in the RDMA Message
1000b	Immediate Data	Immediate Data Header	No
1001b	Immediate Data with SE	Immediate Data Header	No
1010b	Atomic Request	Atomic Request Header	No
1011b	Atomic Response	Atomic Response Header	No

Figure 3: RDMA Message Definitions

5. Atomic Operations

The RDMA Protocol Specification in RFC 5040 does not include support for Atomic Operations, which are an important building block for implementing distributed shared memory.

This document extends the RDMA Protocol specification with a set of basic Atomic Operations and specifies their resource and ordering rules. The Atomic Operations specified in this document provide equivalent functionality to the InfiniBand RDMA transport as well as extended Atomic Operations defined in Open Fabrics Verbs, to allow applications that use these primitives to work interchangeably over iWARP. Other operations are left for future consideration.

Atomic Operations as specified in this document execute a 64-bit memory operation at a specified destination ULP Buffer address on a Responder node using the Tagged Buffer data transfer model. The operations atomically read, modify, and write back the contents of the destination ULP Buffer address and guarantee that Atomic Operations on this ULP Buffer address by other RDMAP Streams on the

same RNIC do not occur between the read and the write caused by the Atomic Operation. Therefore, the Responder RNIC MUST implement mechanisms to prevent Atomic Operations to a memory registered for Atomic Operations while an Atomic Operation targeting the memory is in progress. The Requester of an Atomic Operation cannot rely on Atomic Operation behavior at the Responder across multiple RNICs or with respect to other applications/ULPs running at the Responder that can access the ULP Buffer. It is OPTIONAL for an RNIC to provide such behavior when implementing the Atomic Operations specified in this document. An RNIC that supports Atomic Operations as specified in this document MUST implement both the FetchAdd operation as specified in Section 5.1.1 and the CmpSwap operation as specified in Section 5.1.2. The advertisement of Tagged Buffer information for Atomic Operations is outside the scope of this specification and is handled by the ULPs.

Implementation note: It is RECOMMENDED that the applications do not use the ULP Buffer addresses used for Atomic Operations for other RDMA operations due to the lack of atomicity guarantees between operations other than Atomic Operations.

Implementation note: Errors related to the alignment in the following sections cover Atomic Operations targeted at a ULP Buffer address that is not aligned to a 64-bit boundary.

Atomic Operation Request Messages use the same remote addressing mechanism as RDMA Reads and Writes. The ULP Buffer address specified in the request is in the address space of the Remote Peer to which the Atomic Operation is targeted.

Atomic Operation Response Messages MUST use the Untagged Buffer model with QN=3. Queue number 3 will be used to track outstanding Atomic Operation Request messages at the Requester. When the Atomic Operation Response message is received, the Message Sequence Number (MSN) will be used to locate the corresponding Atomic Operation request in order to complete the Atomic Operation request.

5.1. Atomic Operation Details

The following subsections describe the Atomic Operations in more detail.

5.1.1. FetchAdd

The FetchAdd Atomic Operation requests the Responder to read a 64-bit Original Remote Data Value at a 64-bit aligned ULP Buffer address in the Responder’s memory, perform the FetchAdd operation on multiple fields of selectable length specified by 64-bit "Add Mask", and write

the result back to the same ULP Buffer address. The Atomic addition is performed independently on each one of these fields. A bit set in the Add Mask field specifies the field boundary; for each field, a bit is set at the most significant bit position for each field, causing any carry out of that bit position to be discarded when the addition is performed.

FetchAdd Atomic Operations MUST target ULP Buffer addresses that are 64-bit aligned. FetchAdd Atomic Operations that target ULP Buffer addresses that are not 64-bit aligned MUST be surfaced as errors, and the Responder's memory MUST NOT be modified in such cases. Additionally, an error MUST be surfaced and a terminate message MUST be generated. The setting of the Add Mask field to 0x0000000000000000 results in Atomic Add of 64-bit Original Remote Data Value and 64-bit "Add Data".

The pseudocode below describes a masked FetchAdd Atomic Operation.

```
bit_location = 1

carry = 0

Remote Data Value = 0

for bit = 0 to 63

{

   if (bit != 0 ) bit_location = bit_location << 1

   val1 = (Original Remote Data Value & bit_location) >> bit

   val2 = (Add Data & bit_location) >> bit
   sum = carry + val1 + val2

   carry = (sum & 2) >> 1

   sum = sum & 1

   if (sum)

      Remote Data Value |= bit_location

   carry = ((carry) && (!(Add Mask & bit_location)))

}
```

The FetchAdd operation is performed in the endian format of the target memory. The "Original Remote Data Value" is converted from the endian format of the target memory for return and returned to the Requester. The fields are in big-endian format on the wire.

The Requester specifies:

- Remote STag
- Remote Tagged Offset
- Add Data
- Add Mask

The Responder returns:

- Original Remote Data

5.1.2. CmpSwap

The CmpSwap Atomic Operation requires the Responder to read a 64-bit value at a ULP Buffer address that is 64-bit aligned in the Responder's memory, to perform an AND logical operation using the 64-bit Compare Mask field in the Atomic Operation Request header, then to compare it with the result of a logical AND operation of the Compare Mask and the Compare Data fields in the header. If the two values are equal, the Responder is required to swap masked bits in the same ULP Buffer address with the masked Swap Data. If the two masked compare values are not equal, the contents of the Responder's memory are not changed. In either case, the original value read from the ULP Buffer address is converted from the endian format of the target memory for return and returned to the Requester. The fields are in big-endian format on the wire.

The Requester specifies:

- Remote STag
- Remote Tagged Offset
- Swap Data
- Swap Mask
- Compare Data
- Compare Mask

The Responder returns:

- Original Remote Data Value

The following pseudocode describes the masked CmpSwap operation result.

```
if (!((Compare Data ^ Original Remote Data Value) &

      Compare Mask))

then

   Remote Data Value =

     (Original Remote Data Value & ~(Swap Mask))

                        | (Swap Data & Swap Mask)

else

   Remote Data Value = Original Remote Data Value
```

After the operation, the remote data Buffer MUST contain the "Original Remote Data Value" (if comparison did not match) or the masked "Swap Data" (if the comparison did match). CmpSwap Atomic Operations MUST target ULP Buffer addresses that are 64-bit aligned.

If a CmpSwap Atomic Operation is attempted on a target ULP Buffer address that is not 64-bit aligned:

- The operation MUST NOT be performed,
- The Responder's memory MUST NOT be modified,
- The result MUST be surfaced as an error, and
- A terminate message MUST be generated. (See Section 8.2 for the contents of the terminate message.)

5.2. Atomic Operations

The Atomic Operation Request and Response are RDMA Messages. An Atomic Operation makes use of the DDP Untagged Buffer Model. Atomic Operation Request messages MUST use the same Queue Number as RDMA Read Requests (QN=1). Reusing the same Queue Number for Atomic Request messages allows the Atomic Operations to reuse the same infrastructure (e.g., Outbound and Inbound RDMA Read Queue Depth

(ORD/IRD) flow control) as defined for RDMA Read Requests. Atomic Operation Response messages MUST set Queue Number (QN) to 3 in the DDP header.

The RDMA Message OpCode for an Atomic Request Message is 1010b. The RDMA Message OpCode for an Atomic Response Message is 1011b.

5.2.1. Atomic Operation Request Message

The Atomic Operation Request Message carries an Atomic Operation Header that describes the ULP Buffer address in the Responder's memory. The Atomic Operation Request header immediately follows the DDP header. The RDMAP layer passes to the DDP layer a RDMAP Control Field. The following figure depicts the Atomic Operation Request Header that is used for all Atomic Operation Request Messages:

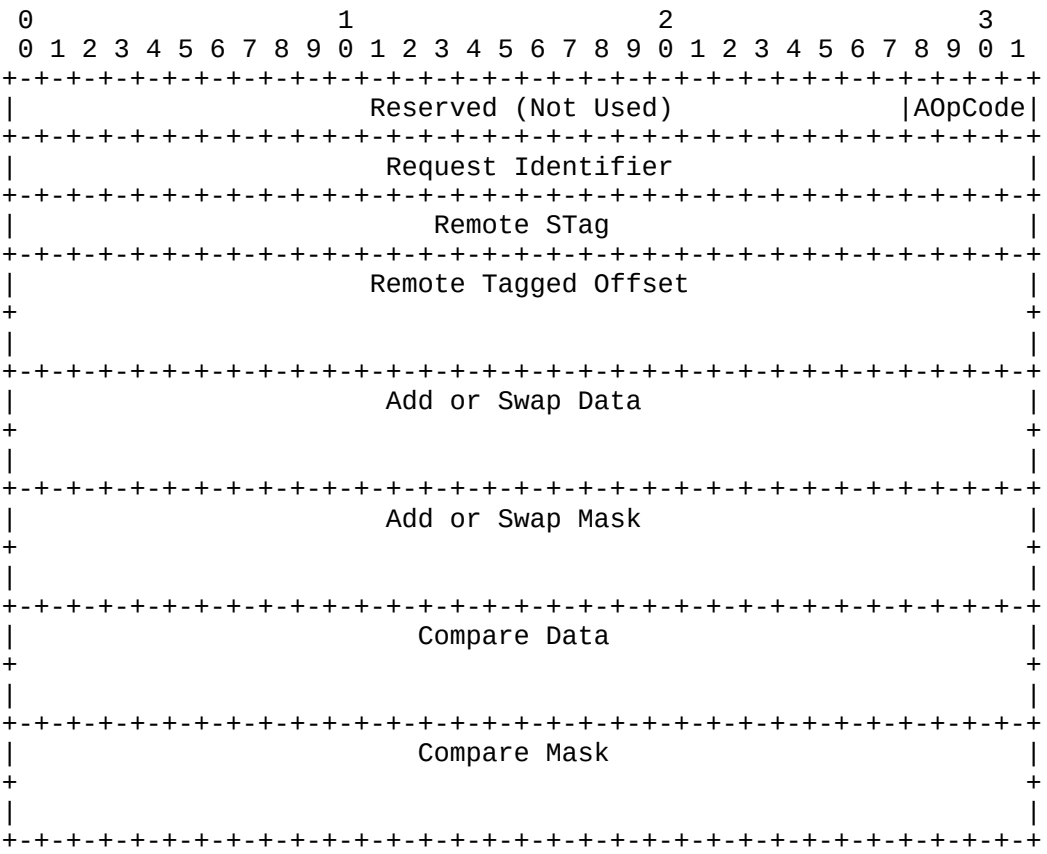


Figure 4: Atomic Operation Request Header

Reserved (Not Used): 28 bits

This field is set to zero on transmit, ignored on receive.

Atomic Operation Code (AOpCode): 4 bits.

See Figure 5. All Atomic Operation Codes from Figure 5 MUST be implemented by an RNIC that supports Atomic Operations.

Request Identifier: 32 bits.

The Request Identifier specifies a number that is used to identify the Atomic Operation Request Message. The value used in this field is selected by the RNIC that sends the message, and it is reflected back to the Local Peer in the Atomic Operation Response message.

Remote STag: 32 bits.

The Remote STag identifies the Remote Peer's Tagged Buffer targeted by the Atomic Operation. The Remote STag is associated with the RDMAP Stream through a mechanism that is outside the scope of the RDMAP specification.

Remote Tagged Offset: 64 bits.

The Remote Tagged Offset specifies the starting offset, in octets, from the base of the Remote Peer's Tagged Buffer targeted by the Atomic Operation. The Remote Tagged Offset MAY start at an arbitrary offset but MUST represent a ULP Buffer address that is 64-bit aligned.

Add or Swap Data: 64 bits.

The Add or Swap Data field specifies the 64-bit "Add Data" value in an Atomic FetchAdd Operation or the 64-bit "Swap Data" value in an Atomic Swap or CmpSwap Operation.

Add or Swap Mask: 64 bits

This field is used in masked Atomic Operations (FetchAdd and CmpSwap) to perform a bitwise logical AND operation as specified in the definition of these operations. For non-masked Atomic Operations (Swap), this field is set to ffffffffffffffffh on transmit and ignored by the receiver.

Compare Data: 64 bits.

The Compare Data field specifies the 64-bit "Compare Data" value in an Atomic CmpSwap Operation. For Atomic Operations FetchAdd and Atomic Swap, the Compare Data field is set to zero on transmit and ignored by the receiver.

Compare Mask: 64 bits

This field is used in masked Atomic Operation CmpSwap to perform a bitwise logical AND operation as specified in the definition of these operations. For Atomic Operations FetchAdd and Swap, this field is set to ffffffffffffffffh on transmit and ignored by the receiver.

Atomic Operation Code	Atomic Operation	Add or Swap Data	Add or Swap Mask	Compare Data	Compare Mask
0000b	FetchAdd	Add Data	Add Mask	N/A	N/A
0010b	CmpSwap	Swap Data	Swap Mask	Valid	Valid

Figure 5: Atomic Operation Message Definitions

The Atomic Operation Request Message has the following semantics:

1. An Atomic Operation Request Message MUST reference an Untagged Buffer. That is, the Local Peer's RDMAP layer MUST request that the DDP mark the Message as Untagged.

2. One Atomic Operation Request Message MUST consume one Untagged Buffer.

3. The Responder's RDMAP layer MUST process an Atomic Operation Request Message. A valid Atomic Operation Request Message MUST NOT be delivered to the Responder's ULP (i.e., it is processed by the RDMAP layer).

4. At the Responder, an error MUST be surfaced in response to delivery to the Remote Peer's RDMAP layer of an Atomic Operation Request Message with an Atomic Operation Code that the RNIC does not support.

5. An Atomic Operation Request Message MUST reference the RDMA Read Request Queue. That is, the Requester's RDMAP layer MUST request that the DDP layer set the Queue Number field to one.

6. The Requester MUST pass to the DDP layer Atomic Operation Request Messages in the order they were submitted by the ULP.

7. The Responder MUST process the Atomic Operation Request Messages in the order they were sent.

8. If the Responder receives a valid Atomic Operation Request Message, it MUST respond with a valid Atomic Operation Response Message.

5.2.2. Atomic Operation Response Message

The Atomic Operation Response Message carries an Atomic Operation Response Header that contains the "Original Request Identifier" and "Original Remote Data Value". The Atomic Operation Response Header immediately follows the DDP header. The RDMAP layer passes to the DDP layer a RDMAP Control Field. The following figure depicts the Atomic Operation Response header that is used for all Atomic Operation Response Messages:

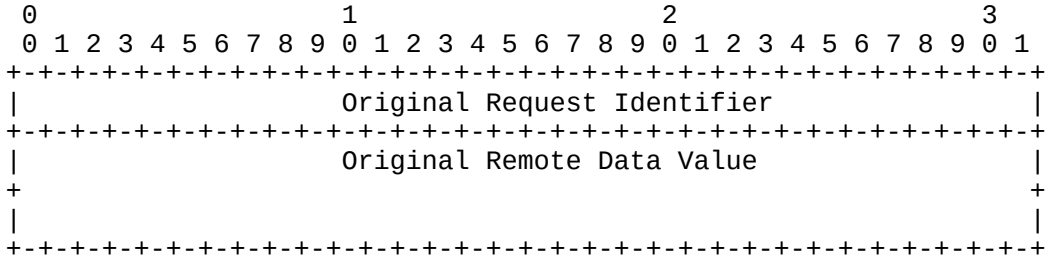


Figure 6: Atomic Operation Response Header

Original Request Identifier: 32 bits.

The Original Request Identifier is set to the value specified in the Request Identifier field that was originally provided in the corresponding Atomic Operation Request Message.

Original Remote Data Value: 64 bits.

The Original Remote Value specifies the original 64-bit value stored at the ULP Buffer address targeted by the Atomic Operation.

The Atomic Operation Response Message has the following semantics:

1. The Atomic Operation Response Message for the associated Atomic Operation Request Message travels in the opposite direction.

2. An Atomic Operation Response Message MUST consume an Untagged Buffer. That is, the Responder RDMAP layer MUST request that the DDP mark the Message as Untagged.

3. An Atomic Operation Response Message MUST reference the Queue Number 3. That is, the Responder's RDMAP layer MUST request that the DDP layer set the Queue Number field to 3.

4. The Responder MUST ensure that a sufficient number of Untagged Buffers are available on the RDMA Read Request Queue (Queue with DDP Queue Number 1) to support the maximum number of Atomic Operation Requests negotiated by the ULP in addition to the maximum number of RDMA Read Requests negotiated by the ULP.

5. The Requester MUST ensure that a sufficient number of Untagged Buffers are available on the RDMA Atomic Response Queue (Queue with DDP Queue Number 3) to support the maximum number of Atomic Operation Requests negotiated by the ULP.

6. The RDMAP layer MUST Deliver the Atomic Operation Response Message to the ULP.

7. At the Requester, when an invalid Atomic Operation Response Message is delivered to the Remote Peer's RDMAP layer, an error is surfaced.

8. When the Responder receives Atomic Operation Request messages, the Responder RDMAP layer MUST pass Atomic Operation Response Messages to the DDP layer, in the order that the Atomic Operation Request Messages were received by the RDMAP layer, at the Responder.

5.3. Atomicity Guarantees

Atomicity of the Read-Modify-Write (RMW) on the Responder's node by the Atomic Operation MUST be assured in the context of concurrent atomic accesses by other RDMAP Streams on the same RNIC.

5.4. Atomic Operations Ordering and Completion Rules

In addition to the ordering and completion rules described in RFC 5040, the following rules apply to implementations of the Atomic Operations.

1. For an Atomic Operation, the Requester MUST NOT consider the contents of the Tagged Buffer at the Responder to be modified by that specific Atomic Operation until the Atomic Operation Response Message has been Delivered to RDMAP at the Requester.

2. Atomicity guarantees MUST be provided within the scope of a single RNIC.

   Implementation Note: This requirement for atomicity among operations is limited to the scope of a single RNIC. Atomicity guarantees are OPTIONAL with respect to access to the Tagged Buffer by any other method than an Atomic Operation via the same RNIC. Examples of such accesses that may not be atomic with respect to an Atomic Operation include accesses via other RNICs and local processor memory access to the Tagged Buffer.

3. Atomic Operation Request Messages MUST NOT start processing at the Responder until they have been Delivered to RDMAP by DDP.

4. Atomic Operation Response Messages MAY be generated at the Responder after subsequent RDMA Write Messages or Send Messages have been Placed or Delivered.

5. Atomic Operation Response Message processing at the Responder MUST be started only after the Atomic Operation Request Message has been Delivered by the DDP layer (thus, all previous RDMA Messages on that DDP Stream have been Delivered).

6. Send Messages MAY be Completed at the Responder before prior incoming Atomic Operation Request Messages have completed their response processing.

7. An Atomic Operation MUST NOT be Completed at the Requester until the DDP layer Delivers the associated incoming Atomic Operation Response Message.

8. If more than one outstanding Atomic Request Message is supported by both peers, the Atomic Operation Request Messages MUST be processed in the order they were delivered by the DDP layer on the Responder. Atomic Operation Response Messages MUST be submitted to the DDP layer on the Responder in the order the Atomic Operation Request Messages were Delivered by DDP.

6. Immediate Data

The Immediate Data operation is typically used in conjunction with an RDMA Write Operation to improve ULP processing efficiency. The efficiency is gained by causing an RDMA Completion to be generated immediately following the RDMA Write operation. This RDMA Completion delivers 8 bytes of Immediate Data at the Remote Peer. The combination of an RDMA Write Message followed by an Immediate Data Operation has the same behavior as the RDMA Write with Immediate Data operation found in InfiniBand. An Immediate Data operation that is not preceded by an RDMA Write operation causes an RDMA Completion.

6.1. RDMAP Interactions with ULP for Immediate Data

For Immediate Data operations, the following are the interactions between the RDMAP Layer and the ULP:

- At the Data Source:
  - The ULP passes to the RDMAP Layer the following:
    - 8 bytes of ULP Immediate Data
  - When the Immediate Data operation Completes, an indication of the Completion results.
- At the Data Sink:
  - If the Immediate Data operation is Completed successfully, the RDMAP Layer passes the following information to the ULP Layer:
    - 8 bytes of Immediate Data
    - An Event, if the Data Sink is configured to generate an Event.
  - If the Immediate Data operation is Completed in error, the Data Sink RDMAP Layer will pass up the corresponding error information to the Data Sink ULP and send a Terminate Message to the Data Source RDMAP Layer. The Data Source RDMAP Layer will then pass up the Terminate Message to the ULP.

6.2. Immediate Data Header Format

The Immediate Data and Immediate Data with SE Messages carry Immediate Data as shown in Figure 7. The RDMAP layer passes to the DDP layer an RDMAP Control Field and 8 bytes of Immediate Data. The first 8 bytes of the data following the DDP header contains the Immediate Data. See Appendix A.3 for the DDP segment format of an Immediate Data or Immediate Data with SE Message.

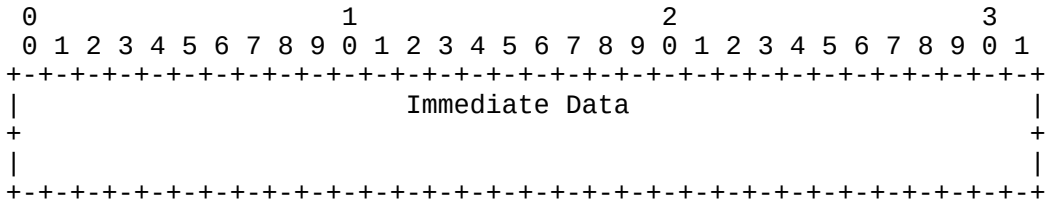


Figure 7: Immediate Data or Immediate Data with SE Message Header

Immediate Data: 64 bits.
8 bytes of data transferred from the Data Source to an untagged Buffer at the Data Sink.

6.3. Immediate Data or Immediate Data with SE Message

The Immediate Data or Immediate Data with SE Message uses the DDP Untagged Buffer Model to transfer Immediate Data from the Data Source to the Data Sink.

- An Immediate Data or Immediate Data with SE Message MUST reference an Untagged Buffer. That is, the Local Peer's RDMAP Layer MUST request that the DDP layer mark the Message as Untagged.

- One Immediate Data or Immediate Data with SE Message MUST consume one Untagged Buffer.

- At the Remote Peer, the Immediate Data and Immediate Data with SE Messages MUST be Delivered to the Remote Peer's ULP in the order they were sent.

- For an Immediate Data or Immediate Data with SE Message, the Local Peer's RDMAP Layer MUST request that the DDP layer set the Queue Number field to zero.

- For an Immediate Data or Immediate Data with SE Message, the Local Peer's RDMAP Layer MUST request that the DDP layer transmit 8 bytes of data.

- The Local Peer MUST issue Immediate Data and Immediate Data with SE Messages in the order they were submitted by the ULP.

- The Remote Peer MUST check that Immediate Data and Immediate Data with SE Messages include exactly 8 bytes of data from the DDP layer. The DDP header carries the length field that is reported by the DDP layer.

6.4. Ordering and Completions

Ordering and completion rules for Immediate Data are the same as those for a Send operation as described in Section 5.5 of RFC 5040.

7. Ordering and Completions Table

The following table summarizes the ordering relationships for Atomic and Immediate Data operations from the standpoint of the Local Peer issuing the Operations. Note that in the table that follows, Send includes Send, Send with Invalidate, Send with Solicited Event, and Send with Solicited Event and Invalidate. Also note that in the table below, Immediate Data includes Immediate Data and Immediate Data with Solicited Event.

First Operation	Second Operation	Placement Guarantee at Remote Peer	Placement Guarantee at Local Peer	Ordering Guarantee at Remote Peer
Immediate Data	Send	No Placement Guarantee between Send Payload and Immediate Data	Not Applicable	Completed in Order
Immediate Data	RDMA Write	No Placement Guarantee between RDMA Write Payload and Immediate Data	Not Applicable	Not Applicable

Immediate Data	RDMA Read	No Placement Guarantee between Immediate Data and RDMA Read Request	RDMA Read Response will not be Placed until Immediate Data is Placed at Remote Peer	RDMA Read Response Message will not be generated until Immediate Data has been Completed
Immediate Data	Atomic	No Placement Guarantee between Immediate Data and Atomic Request	Atomic Response will not be Placed until Immediate Data is Placed at Remote Peer	Atomic Response Message will not be generated until Immediate Data has been Completed
Immediate Data or Send	Immediate Data	No Placement Guarantee	Not Applicable	Completed in Order
RDMA Write	Immediate Data	No Placement Guarantee	Not Applicable	Immediate Data is Completed after RDMA Write is Placed and Delivered
RDMA Read	Immediate Data	No Placement Guarantee between Immediate Data and RDMA Read Request	Immediate Data MAY be Placed before RDMA Read Response is generated	Not Applicable
Atomic	Immediate Data	No Placement Guarantee between Immediate Data and Atomic Request	Immediate Data MAY be Placed before Atomic Response is generated	Not Applicable

Atomic	Send	No Placement Guarantee between Send Payload and Atomic Request	Send Payload MAY be Placed before Atomic Response is generated	Not Applicable
Atomic	RDMA Write	No Placement Guarantee between RDMA Write Payload and Atomic Request	RDMA Write Payload MAY be Placed before Atomic Response is generated	Not Applicable
Atomic	RDMA Read	No Placement Guarantee between Atomic Request and RDMA Read Request	No Placement Guarantee between Atomic Response and RDMA Read Response	RDMA Read Response Message will not be generated until Atomic Response Message has been generated
Atomic	Atomic	Placed in order	No Placement Guarantee between two Atomic Responses	Second Atomic Request Message will not be processed until first Atomic Response has been generated
Send	Atomic	No Placement Guarantee between Send Payload and Atomic Request	Atomic Response will not be Placed at the Local Peer until Send Payload is Placed at the Remote Peer	Atomic Response Message will not be generated until Send has been Completed

RDMA Write	Atomic	No Placement Guarantee between RDMA Write Payload and Atomic Request	Atomic Response will not be Placed at the Local Peer until RDMA Write Payload is Placed at the Remote Peer	Not Applicable
RDMA Read	Atomic	No Placement Guarantee between Atomic Request and RDMA Read Request	No Placement Guarantee between Atomic Response and RDMA Read Response	Atomic Response Message will not be generated until RDMA Read Response has been generated

8. Error Processing

In addition to the error processing described in Section 7 of RFC 5040, the following rules apply for the new RDMA Messages defined in this specification.

8.1. Errors Detected at the Local Peer

The Local Peer MUST send a Terminate Message for each of the following cases:

1. For errors detected while creating an Atomic Request, Atomic Response, Immediate Data, or Immediate Data with SE Message, or other reasons not directly associated with an incoming Message, the Terminate Message and Error code are sent instead of the Message. In this case, the Error Type and Error Code fields are included in the Terminate Message, but the Terminated DDP Header and Terminated RDMA Header fields are set to zero.

2. For errors detected on an incoming Atomic Request, Atomic Response, Immediate Data, or Immediate Data with SE (after the Message has been Delivered by DDP), the Terminate Message is sent at the earliest possible opportunity, preferably in the next

outgoing RDMA Message. In this case, the Error Type, Error Code, and Terminated DDP Header fields are included in the Terminate Message, but the Terminated RDMA Header field is set to zero.

8.2. Errors Detected at the Remote Peer

On incoming Atomic Requests, Atomic Responses, Immediate Data, and Immediate Data with Solicited Event, the following MUST be validated:

- The DDP layer MUST validate all DDP Segment fields.
- The RDMA OpCode MUST be valid.
- The RDMA Version MUST be valid.

On incoming Atomic requests the following additional validation MUST be performed:

- The RDMAP layer MUST validate that the Remote Peer's Tagged ULP Buffer address references a ULP Buffer address that is 64-bit aligned. In the case of an error, the RDMAP layer MUST generate a Terminate Message indicating RDMA Layer Remote Operation Error with Error Code Name "Catastrophic error, localized to RDMAP Stream" as described in Section 4.8 of RFC 5040. Implementation Note: A ULP implementation can avoid this error by having the target ULP Buffer of an Atomic Operation 64-bit aligned.

9. Security Considerations

This document specifies extensions to the RDMA Protocol specification in RFC 5040, and as such the Security Considerations discussed in Section 8 of RFC 5040 apply. In particular, Atomic Operations use ULP Buffer addresses for the Remote Peer Buffer addressing used in RFC 5040 as required by the security model described in RFC 5042 [RFC5042].

RDMAP and related protocols may be used by applications that exhibit distinctive traffic characteristics such as message timing, source, destination, and size patterns. Examples include structured high-performance computing applications based on the MPI interface. For such applications, analysis of encrypted traffic could reveal sensitive information, e.g., the nature of the application, size of data set being used, and information about the application's rate of progress. Such information can be hidden from passive observation via use of Encapsulating Security Payload version 3 (ESPv3) Traffic Flow Confidentiality [RFC4303] to obfuscate the encrypted traffic's characteristics. ESPv3 implementation requirements for RDMAP are specified in [RFC7146].

10. IANA Considerations

IANA has added the following entries to the "RDMAP Message Operation Codes" registry of "Remote Direct Data Placement (RDDP)" registry:

0x8, Immediate Data, this specification

0x9, Immediate Data with Solicited Event, this specification

0xA, Atomic Request, this specification

0xB, Atomic Response, this specification

In addition, the following registry has been added to the "Remote Direct Data Placement (RDDP)" registry. The following section specifies the registry, its initial contents, and the administration policy in more detail.

10.1. RDMAP Message Atomic Operation Subcodes

Name of the registry: "RDMAP Message Atomic Operation Subcodes"

Namespace details: RDMAP Message Atomic Operation Subcodes are 4-bit values.

Information that must be provided to assign a new value: An IESG-approved Standards Track specification defining the semantics and interoperability requirements of the proposed new value and the fields to be recorded in the registry.

Fields to record in the registry: RDMAP Message Atomic Operation Subcode, Atomic Operation, RFC Reference.

Initial registry contents:

0x0, FetchAdd, this specification

0x1, Reserved, this specification

0x2, CmpSwap, this specification

Note: An experimental RDMAP Message Operation Code has already been allocated; hence, there is no need for an experimental RDMAP Message Atomic Operation Subcode.

All other values are Unassigned and available to IANA for assignment. New RDMAP Message Atomic Operation Subcodes should be assigned sequentially in order to better support implementations that process RDMAP Message Atomic Operations in hardware.

Allocation Policy: Standards Action [RFC5226]

10.2. RDMAP Queue Numbers

Name of the registry: "RDMAP DDP Untagged Queue Numbers"

Namespace details: RDMAP DDP Untagged Queue numbers are 32-bit values.

Information that must be provided to assign a new value: An IESG-approved Standards Track specification defining the semantics and interoperability requirements of the proposed new value and the fields to be recorded in the registry.

Fields to record in the registry: RDMAP DDP Untagged Queue Numbers, Queue Usage Description, RFC Reference.

Initial registry contents:

0x00000000, Queue 0 (Send operation Variants), [RFC5040]

0x00000001, Queue 1 (RDMA Read Request operations), [RFC5040]

0x00000002, Queue 2 (Terminate operations), [RFC5040]

0x00000003, Queue 3 (Atomic Response operations), this specification

Note: An experimental RDMAP Message Operation Code has already been allocated; hence, there is no need for an experimental RDMAP DDP Untagged Queue Number.

All other values are Unassigned and available to IANA for assignment. New RDMAP queue numbers should be assigned sequentially in order to better support implementations that perform RDMAP queue selection in hardware.

Allocation Policy: Standards Action [RFC5226]

11. References

11.1. Normative References

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

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

[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.

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

[RFC7146] Black, D. and P. Koning, "Securing Block Storage Protocols over IP: RFC 3723 Requirements Update for IPsec v3", RFC 7146, April 2014.

11.2. Informative References

[DAT_ATOMICS] DAT Collaborative, "IB Transport Specific Extensions for DAT 2.0", User Direct Access Programming Library, <http://www.datcollaborative.org/DAT_IB_Extensions.pdf>.

[IB] InfiniBand Trade Association, "InfiniBand Architecture Specification Volumes 1 and 2", Release 1.1, November 2002, <http://www.infinibandta.org/specs>.

[MPI] Message Passing Interface Forum, "MPI: A Message-Passing Interface Standard, Version 3.0", September 2012, <http://www.mpi-forum.org/docs/mpi-3.0/mpi30-report.pdf>.

[OFAVERBS] Rosenstock, H., "Subject: Re: [PATCH 0/2] Add support for enhanced atomic operations", message to the linux-rdma mailing list, <http://www.spinics.net/lists/linux-rdma/msg02405.html>.

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

[RFC5045] Bestler, C., Ed., and L. Coene, "Applicability of Remote Direct Memory Access Protocol (RDMA) and Direct Data Placement (DDP)", RFC 5045, October 2007.

[RFC6581] Kanevsky, A., Ed., Bestler, C., Ed., Sharp, R., and S. Wise, "Enhanced Remote Direct Memory Access (RDMA) Connection Establishment", RFC 6581, April 2012.

[RSOCKETS] Hefty, S., "RDMA CM - RDMA enabled Sockets library for Open Fabrics", <http://git.openfabrics.org/?p=~shefty/librdmacm.git;a=summary>.

12. Acknowledgments

The authors would like to acknowledge the following individuals who provided valuable comments and suggestions.

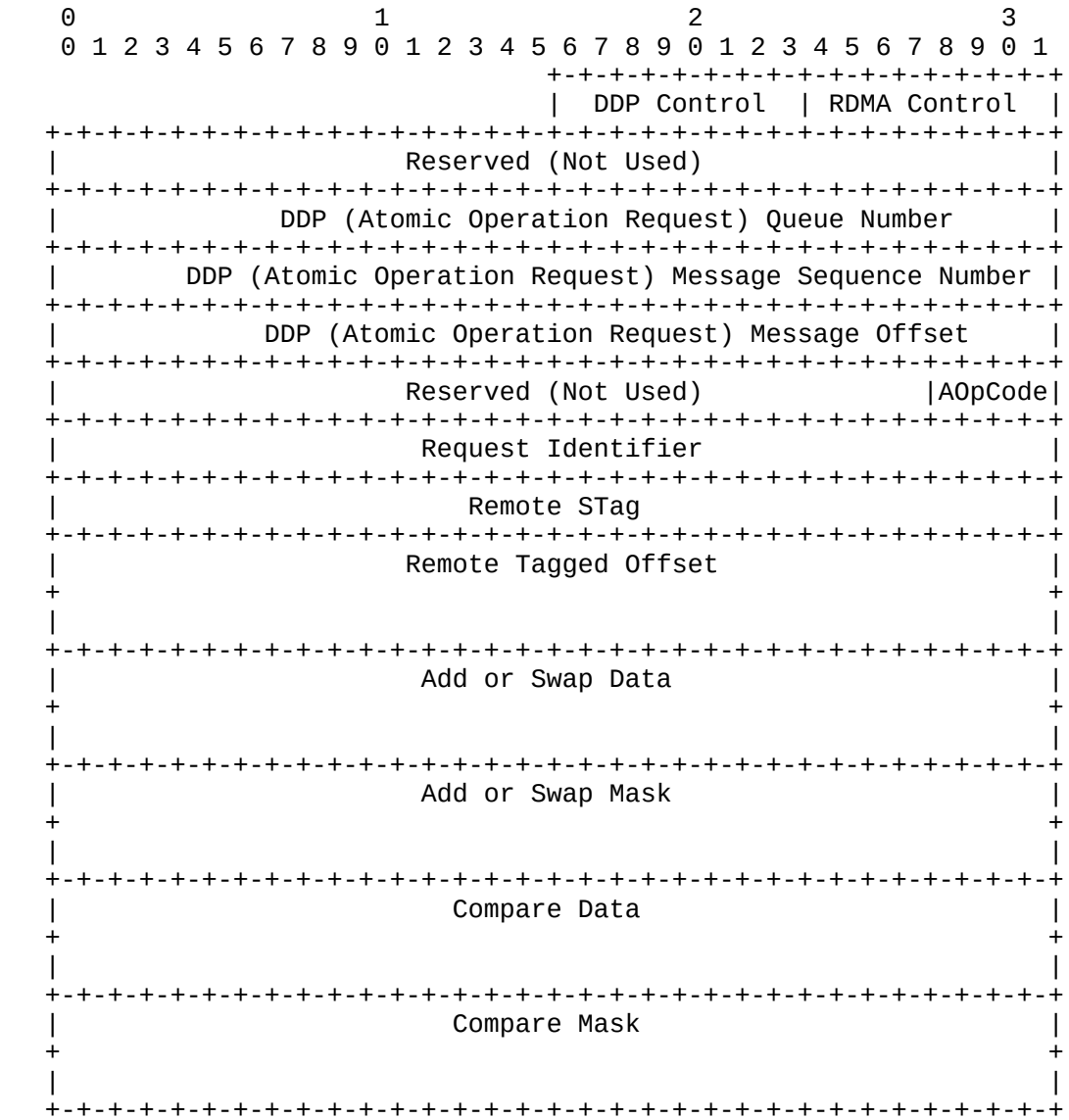
- David Black
- Arkady Kanevsky
- Bernard Metzler
- Jim Pinkerton
- Tom Talpey
- Steve Wise
- Don Wood

Appendix A. DDP Segment Formats for RDMA Messages

This appendix is for information only and is NOT part of the standard. It simply depicts the DDP Segment format for the various RDMA Messages.

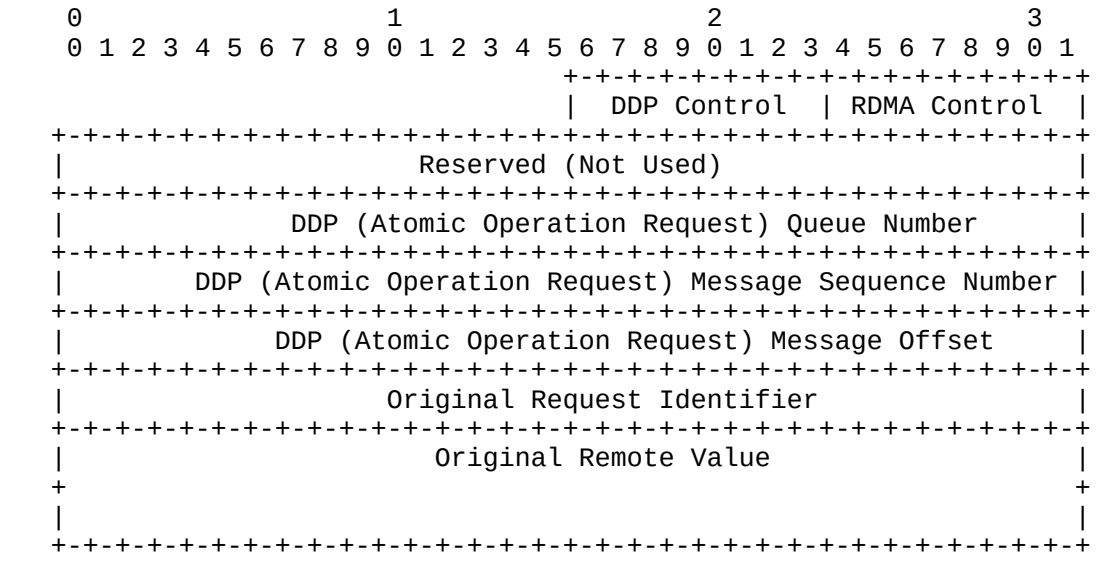
A.1. DDP Segment for Atomic Operation Request

The following figure depicts an Atomic Operation Request, DDP Segment:



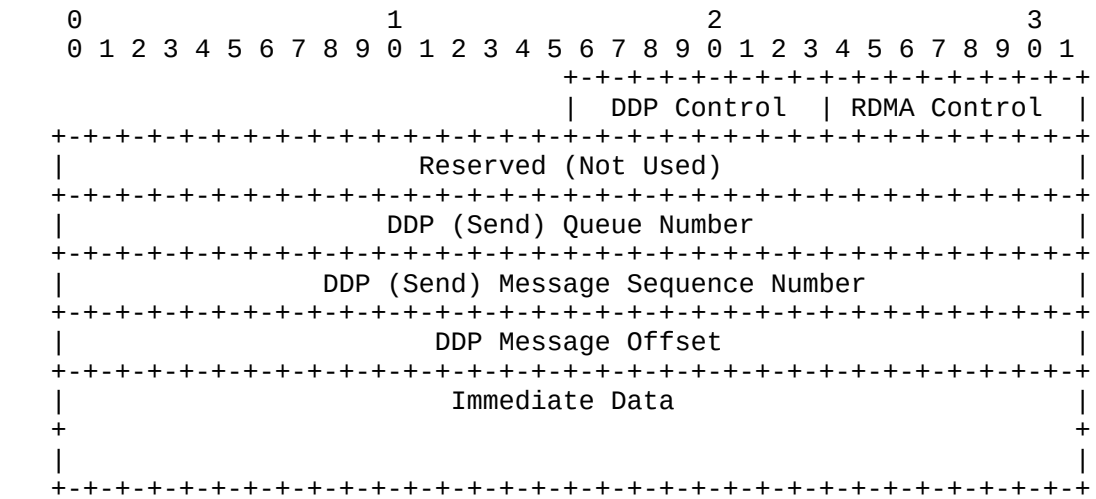
A.2. DDP Segment for Atomic Response

The following figure depicts an Atomic Operation Response, DDP Segment:



A.3. DDP Segment for Immediate Data and Immediate Data with SE

The following figure depicts an Immediate Data or Immediate Data with SE, DDP Segment:



Authors' Addresses

Hemal Shah
Broadcom Corporation
5300 California Avenue
Irvine, CA 92617
US
Phone: 1-949-926-6941
EMail: hemal@broadcom.com

Felix Marti
Chelsio Communications, Inc.
370 San Aleso Ave.
Sunnyvale, CA 94085
US
Phone: 1-408-962-3600
EMail: felix@chelsio.com

Asgeir Eiriksson
Chelsio Communications, Inc.
370 San Aleso Ave.
Sunnyvale, CA 94085
US
Phone: 1-408-962-3600
EMail: asgeir@chelsio.com

Wael Noureddine
Chelsio Communications, Inc.
370 San Aleso Ave.
Sunnyvale, CA 94085
US
Phone: 1-408-962-3600
EMail: wael@chelsio.com

Robert Sharp
Intel Corporation
1300 South Mopac Expy, Mailstop: AN4-4B
Austin, TX 78746
US
Phone: 1-512-362-1407
EMail: robert.o.sharp@intel.com