iSER on InfiniBand Networks

Introduction

There has been great interest in defining how the iSCSI/iSER (iSCSI Extensions for RDMA) can be made to operate over an InfiniBand network instead of a previously defined protocol called SRP (SCSI RDMA Protocol). The SRP protocol lacked Storage Management and Discovery processes. Therefore, great interest in iSCSI/iSER (iSER for short) has developed since it has the management and discovery structure defined, built, or being built for iSCSI. This will also consolidate the efforts of both Ethernet and InfiniBand communities, and reduce the number of Storage protocols a user has to learn and maintain.

Several things are needed in order for an iSCSI/iSER over an InfiniBand (iSER/IB) network to operate. They are:

1. Being able to exploit the fact that all the IB Initiator and Target nodes can also have an IP Transport (IP-over-IB a.k.a. IPoIB) and associated IP Addresses for standard iSCSI discovery and for address resolution.

2. A way for an iSCSI/iSER on IB Initiator to Login directly to an InfiniBand attached iSER storage node (without having to connect and send Login PDUs via TCP/IP).

3. A way for an IB Initiator to choose, if it wishes, directly attached IB iSER Storage Nodes in preference to an iSER to iSCSI Gateway, or in preference to using TCP over IPoIB (Internet Protocol over InfiniBand) connection.

4. A way for iSER to operate on InfiniBand Channel Adapters supporting versions prior to 1.2 that does not support what is called a Zero Based Tagged Offset (ZBTO) or SendInvSE Messages[1].

Therefore this appendix is divided into four parts to identify the approaches needed.

[1] Though the IB specification has been updated to support the new RDMA features that were introduced by iWARP, there are likely to be a number of systems that are only using IB hardware that only meet the pre version 1.2 specifications. This would mean that they would not support the features called ZeroBasedTO (called in InfiniBand ZBVA) or SendInvSE Messages.

iSER on InfiniBand Networks

I. iSCSI/iSER's use of IP address on an InfiniBand Network

The iSCSI protocol has specified a discovery process which defines how the iSCSI storage controllers will advertise their existence.

Any storage portal within a storage enclosure (called a storage network entity) can advertise the existence of all other nodes to iSNS or SLP or respond to the iSCSI SendTargets Text Command.

The initiator can discover the targets to which it is authorized by

1. A TCP/IP Address:Port of an iSCSI portal is made known by the administrator to an initiator which then uses that information to obtain all the logical target names and their IP Address:Ports, and a port clustering ID called the Portal Group Tag (PGT). This information is extracted from this storage portal by the initiator logging into what is called a Discovery Session and issuing the "SendTargets" command. The Discovery Session is currently defined as a type of iSCSI TCP/IP session.

2. By contacting the SLP (Service Location Protocol) Data Base. SLP is a (perhaps remote database) into which the Storage Controller has advertised its portals (like that available via "SendTargets" in 1. above). The DB can be located within the storage enclosure itself and may only contain its own information but is address using the SLP protocol. This connection is defined as a TCP/IP based protocol.

3. By contacting the iSNS (internet Storage Name Server). iSNS is a more elaborate and more highly available DB server, to which the Storage Controller will advertise its portals (like that available via "SendTargets" in 1. above) This connection is defined as a TCP/IP based protocol.

Because the more advanced discovery processes (SLP and iSNS) of iSCSI and iSCSI/iSER creates, advertises and uses TCP/IP Address:Port as its method of addressing its portals, it is a goal to use that same approach for operating the discovery process for iSCSI/iSER on an IB network.

All IB nodes have a valid IP address through the standard IP over InfiniBand (IPoIB) mapping, and it is possible to advertise these portal addresses to the SLP or iSNS databases (or they can be kept as part of the "SendTargets" approach as mention in 1 above).

The resolution from IP address to InfiniBand Hardware Address (GID) will be conducted through the standard IP routing and ARP Mechanism (RFC 826) using the IP-over-IB interface. (Specific details on ARP over InfiniBand (IB ARP) can be found in the IP-over-IB specification [IBoIP].)

The iSCSI Portal TCP Port number is mapped into a IB "ServiceID" (the IB equivalent of a TCP Service, used by the IB Communication manager to make an IB Reliable Connection [RC]) by taking an "iSER Base SERVICEID" and adding to it the 16 bit iSCSI TCP Port number.

The “iSER Base SERVICEID” will be allocated by IBTA (InfiniBand Trade Association), and until then a value of "0x0000 0000 0002 0000" will be used.

(Note that the IB ServiceID is 64 bit long and the TCP port is only 16 bits long, which will result in the lower 16 bits of the ServiceID being equal to the TCP Port number.)

II. iSER over InfiniBand Login

The iSCSI and iSCSI/iSER over iWARP requires that TCP/IP Byte Stream protocol be used for Login. iSCSI uses TCP/IP Byte Stream protocol for all operations where as the TCP version of iSER/iWARP uses TCP/IP Byte Stream protocol only to Login and negotiate parameters. After the completion of Login iSER/iWARP will enter into Full Feature mode and then from that point on only issue RDMA operations.

iSER/iWARP for TCP requires the RNIC (RDMA Network Interface Controller) take over the TCP/IP connection used during Login and then, transparent to the iSCSI part of iSCSI/iSER, send the RDMA operations over the same TCP/IP connection.

If iSER is to operate over IB, it must establish the equivalent of a TCP connection over IB Reliable Connection (RC), on which the Login PDUs can flow from the Initiator to the Target. Then that same RC connection must be taken over for InfiniBand RC RDMA operations.

The series of Login and Login Response exchanges occur in a half duplex type operation. That is, first there is a Login request from the initiator to the target, and then after the target receives the Login request, the target sends a login response. Then after receiving the Login response, the initiator may continue with an additional Login request, etc. This process continues until all the login parameters are negotiated.
Unfortunately, it is not possible for iSER/IB to use TCP/IP on normal [IPoIB] since IPoIB is currently defined to operate on Unreliable Datagrams (UD), and iSER/IB requires Reliable Connections. Therefore, we need another way to exchange the Login parameters on a Reliable Connection, so that the Full Feature mode of iSER can issue the RDMA RC Operations on the same connection.

Since the iSCSI login PDUs are all defined to operate with a default MaxRecvDataSegmentLength = 8192, it is a relatively straight forward process to have the Initiator and Target, pre-post the right size buffers for the sending and receiving of the Login PDUs via the IB RDMA Send Message[2].

The above approach to Login, will also apply to the post-login Hello and HelloReply messages.

So as a first step, before sending the Hello, or Login PDU messages, the

[2] It is also possible that this approach would permit a version of iSCSI/iSER to operate on an SCTP Network. Therefore, it may be possible to leverage this potential in order to get both iSCSI/iSER support on IB as well as iSCSI/iSER support on SCTP Networks.

iSER Initiating Node will establish an InfiniBand reliable connection. [In InfiniBand, the connection establishment is done out-of-band through an entity called Connection Management (CM). The CM of the client sends a connection Request (REQ) to the server which sends a connection Response (REP) or connection Reject (REJ). The CM can also accept "Private Data" in its messages that can be used by the protocols.]

Before establishing a connection the Initiator will resolve the address from IP to the InfiniBand Hardware address (GID) using standard ARP (specifically IPoIB IB ARP). Then the Initiator CM will send a CM REQ specifying the appropriate "SERVICEID" that identifies, to the IB Endport node, which service is being requested at that node (this is analogous to the TCP Port Number). Also there will be some information in the "Private Data" area, specifying some additional information, which is useful in establishing the appropriate connection.

Since the CM is forming the equivalent of a TCP/IP connection establishment, the target or an intermediate gateway needs to have the Source and Destination IP addresses and Destination TCP Port, which are usually provided in the TCP Syn packet in the iWARP/TCP connection establishment.

The Destination TCP Port is present in the lower 16 bit portion of the ServiceID (Since the SERVICEID will be the "iSER Base SERVICEID" + TCP/IP port number). The IP addresses of the source and the destination will be embedded in the CM REQ Private Data field, additional IB version information and preference flags will also be added to the "Private Data".

The "Private Data" will give the peer node additional important information about the IB features that can be used in the connections and, for Gateway nodes, information about the Ultimate Target's IP Address, etc. The SERVICEID and "Private Data" (on both the REQ and REP messages) are defined in such a way that it will permit the iSER initiator to talk either directly to an iSER/IB Storage Endport node or to an iSER Gateway node (which converts iSER/IB to iSCSI, iSER/iWARP, or Fibre Channel) in a transparent manner.

See the Section 15.6 for details on the Private Data field structure and usage.

III. iSER Node Preferences

As part of iSCSI and iSER discovery is the fact that an iSCSI based protocol will advertise its connections into collections called Portal Groups. Each Portal Group will be of the same type of connection (iSCSI only, iSCSI-iSER/iWARP, iSER/IB, etc.). A storage enclosure (Network Entity) may have multiple Logical Targets. Each Logical Target may have multiple Portal Groups which can be used to connect to it. Each of these Portal Groups can be made up of the same connection type that makes up other Portal Groups connected to the Logical Target, or can be made up of a different type of connection.

Therefore, it is useful for an iSER/IB Initiator to be able to select a connection point from a Target Portal Group of IB connections instead of an indirect connection through an iSER Gateway. Further, if there is no

iSER/IB direct connection or an iSER/IB indirect connection (via a Gateway), the initiator should be able to use normal iSCSI and TCP/IP via [IPoIB]. The following describes how this type of selection can be made.

The key to this approach is the fact that each node has a type which can be individually requested via “Private Data” in the CM REQ Message. When you couple this connection process with the normal SLP/iSNS discovery process, it can be determined if the connection is a direct iSER/IB connection to storage, or a connection to an iSER Gateway or a normal iSCSI connection over IPoIB. In this way the most efficient connection type can be used to address the storage controller. The process for this determination is described below:

1. Pick a connection (portal) IP Address and locate its associated hardware address (GID) using ARP.
2. Send a connection request (CM REQ Message) to each of the GIDs (in turn), which represent the IP Addresses returned by the iSCSI Discovery Process.
3. Include the following Flags in the “Private Data” part of the CM REQ Message:
    - 00b ==> Make connection if either an iSER/IB End Target Node or an iSER/IB Gateway
    - 01b ==> Make connection only if node is an iSER/IB Target End Node (Do NOT make a Gateway connection)
    - 10b ==> Make connection only if node is an iSER/IB Gateway (Do NOT make an End Node connection)
    - 11b ==> Reserved & Invalid value (Do NOT make either an End Node nor a Gateway connection)

    Note: Reserved values are usually 0, however, with these flags, the use of negative logic is employed so the Initiator, which does not care to distinguish between an iSER/IB Storage Endport Node, or an iSER Gateway node can just leave the bit fields set to the default 00b.

    The target of this CM REQ Message, must either reject the connection, if it can not meet the flag requirements (using the standard CM REJ message), or accept the connection (via the standard CM REP message with appropriate “Private Data” specified later).

    If the Target or Gateway node receiving the CM REQ Message doesn’t support the iSER Service (specified by the ServiceID), that node will respond in the standard way for such a situation by returning a CM REJ (reject) message with reason = 8 (Invalid Service ID). In this case the initiator node can assume that it should use normal iSCSI over IPoIB (This is because, in spite of the rejection, the IP address did resolve to the rejecting node’s GID).

    Otherwise, the target node will respond either in a positive CM REQ Message, or a CM REJ message with appropriate reason information, some of which maybe in the response Message. If the node does support the iSER Service, but doesn’t accept the parameters in the private data field it will respond by returning the CM REJ message with reason = 28 (Consumer Reject), the ARI

text field (72 byte of additional Reject Info) can be used to describe the exact reason (due to the IP address, gateway/target, version, ..)

This means that the initiator will have the ability (if it wishes) to first connect to IB Storage Endports, if possible, and if not possible, then connect to, iSER/IB Gateways, and if none of those exist, then connect via iSCSI using IPoIB.

IV. Handling the older IB networks

There are two important default assumptions that have been used in the iSER/iWARP specification. These two assumptions are:

1. A target device can assume that the initiator supports the Zero Based Tagged Office (ZBTO).

2. An initiator supports the receiving of the Send with Invalidate type operation.

However, depending on the age/implementation specification that is supported by the IB Initiator HW, these assumptions are not always appropriate. Some IB nodes do not support ZBTO, and some do not support the Send with Invalidate type operation.

The second assumption is easily over come since the iSER protocol requires the initiator to invalidate any STag (R_Key) that is not automatically invalidated by the HW. Therefore, if a target knows that the initiator HW does not support Send with Invalidate type messages, it should just return the operational status via the “Send” type message (not the “Send with invalidate” message).

To address these issues, there will be flags designated in the iSER Hello/HelloReply Message. With these flags the initiator can ask to only use VA Based TO type Messages, and the other flag informs the initiator that it should not issue the Send with Invalidate type message.

There is also another reason that the “Initiator Invalidate” flag might be set in the Hello Message by the initiator. That reason is that some iSER implementations operating in Non Privilege State may want to reuse, or invalidate the STag itself for performance reasons. In that case it may decide to set the Initiator Invalidate Flag.

If the initiator sets the VABTO flag, and the target accepts it, then the Target MUST support Virtual Address Based Tagged Offsets in the iSER Headers.

The target side should be able to operate with or w/o ZBTO support in its HW, since it does not advertise its STags (R_Keys). However, it MAY contain software in its iSER/IB support to tolerate the initiator’s Virtual Address Based Tagged Offset (VABTO), and in that case it should respond to the initiator’s request by setting the VABTO flag in it’s HelloReply message.

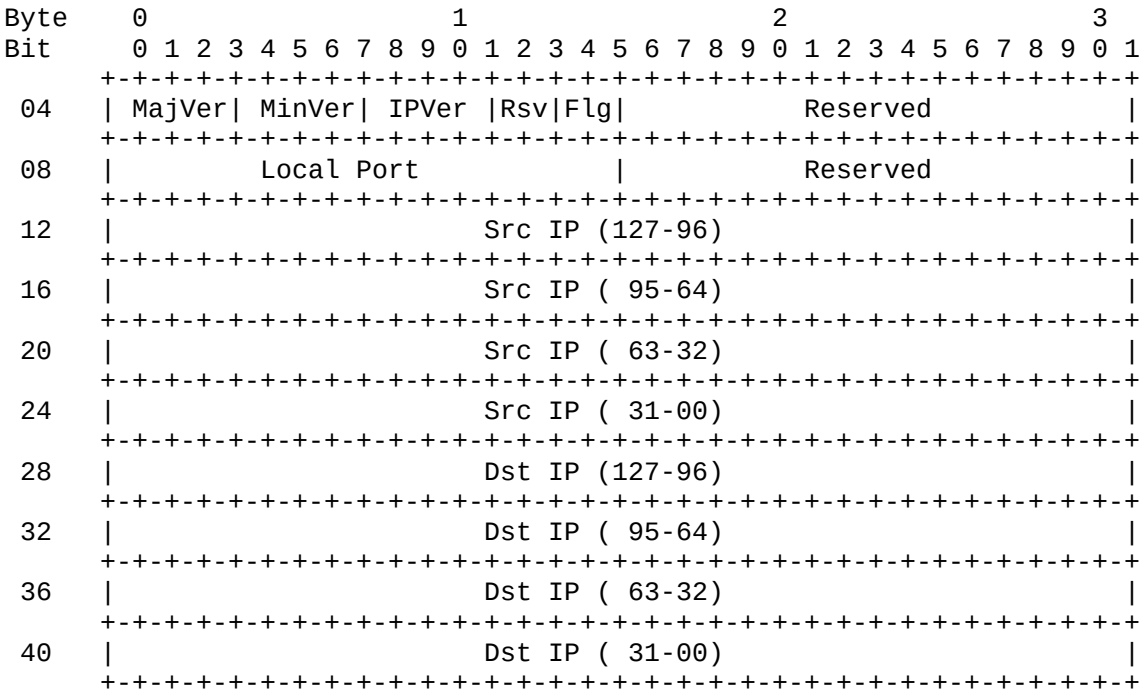
The ability to use the VABTO merely requires the initiator to send to the target, in the iSER headers, a 96 bit logical R_Key (the actual 32 bit R-Key and 64 bits of VA) for each SCSI direction (instead of the 32 Bits of R-Key (STag) used with a ZBTO). The Target will then need to remember the 96 bit Virtual R-Key (Virtual STag) instead of just the 32 bits of the actual R-Key (actual STag).

The Target must not set the VABTO Flag in its HelloReply Message unless requested by the Initiator.

APPENDIX

Format of the CM REQ/REP/REJ Messages & Private Data

(Note: this set of descriptions is currently a work in progress with the InfiniBand Trade Association.)



MajVer – TBD

MinVer are TBD

IPVer

- 0000b indicates IPV4 is being used
- 0001b indicates IPV6 is being used

all other values are reserved

Flg is the Connection Flag Field

- 00b ==> Make connection if either an iSER/IB End Target Node or an iSER/IB Gateway
- 01b ==> Make connection only if node is an iSER/IB Target End Node (Do NOT make a Gateway connection)
- 10b ==> Make connection only if node is an iSER/IB Gateway (Do NOT make an End Node connection)
- 11b ==> Reserved & Invalid value (Do NOT make either an End Node nor a Gateway connection)

Local Port

This field is for use by implementations that require the Source

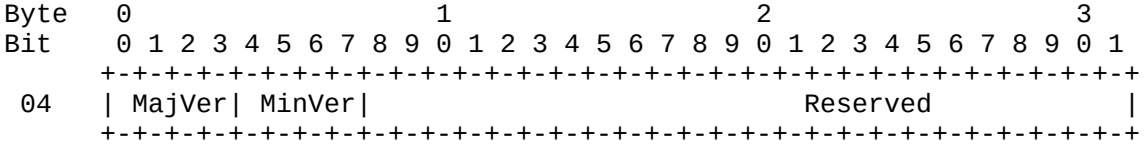
TCP Port number.

Src IP
The Source IP address (in the format determined by IPVer). This is the IP address of the IB Initiator.

Dst IP
The Source IP address (in the format determined by IPVer). This is the IP address of the IB Initiator.

Note: the Destination Port Number is located in the 16 Least Significant Bits (LSB) in the SERVICE ID field of the CM REQ Message. When the iSER SERVICE ID is constructed, the iSER Base SERVICEID should always be added to the port# no matter if it's talking to a Target or a Gateway (the iSER gateway consumer will listen on the SERVICEID bit range that signifies it as an iSER request, and then extract the Destination Port Number by subtracting the iSER base from the ServiceID to get the TCP Port to use when/if he needs it)

Format of the CM REP Message Private Data



Where
MaxVer
TBD

VMinVer
TBD

Valid CM REJ Messages in the ARI field:
TBD