To: T10 Technical Committee
From: Greg Pellegrino (Greg.Pellegrino@compaq.com)
and Rob Elliott, Compaq Computer Corporation (Robert.Elliott@compaq.com)
Date: 28 March 2001
Subject: SRP InfiniBand™ annex

Revision History

Revision 0: 5 January 2001 first revision
Revision 1: 11 January 2001 updates from Houston SRP meeting.
Revision 2: 18 February 2001 updates from Orlando SRP meeting and San Francisco IBTA AWG FTF meeting. Change bars removed due to amount of changes.
Revision 3: 3 March 2001 updates from Denver SRP and IBTA AWG joint meeting. Still no change bars. Removed most support material.
Revision 4: 28 March 2001 updates from Dallas SRP meeting.

Related Documents

Some of these references were used in development of the annex, although no relationship exists with the final result.

T10/srp-r03 – SCSI over RDMA protocol revision 3 (by Ed Gardner)
T10/fcp2r06 – SCSI over Fibre Channel protocol revision 6 (by Bob Snively)

Access controls:
T10/99-245r9 Access Controls (by Jim Hafner)
T10/00-261r0 Discussion of editorial changes to Access Controls (by Jim Hafner)
T10/00-287r1 TransportIDs for Access Controls (by Jim Hafner)
T10/00-381r0 Three minor modifications to Access Controls (by Jim Hafner)
T10/01-026r0 SPC-3 Access Control conflicts due to TransportIDs (by Rob Elliott)

T10/00-268r6 Defining Targets/Initiators as Ports (by George Penokie)
T10/00-425r1 Long Identifiers in SPC-3, SAM-2, SBC-2 and other XOR issues (by Jim Hafner)

T11/FC-GS-3 Revision 6.42 (Section 6.1.2.3 Platform Object)
T11/99-697v0 Management Server Platform Extension (by Duane Baldwin) (source of FC-GS-3)

InfiniBand Architecture Volume 1 – General Specifications, Release 1.0
InfiniBand Architecture Volume 2 – Physical Specifications, Release 1.0
InfiniBand Architecture Volume 3 – Application of InfiniBand, Release 0.9 (draft)

IETF/draft-ietf-ips-iscsi-disc-reqts-01.txt – iSCSI naming and discovery requirements (IPS working group)
IETF/draft-ietf-ips-iscsi-name-disc-00.txt - iSCSI naming and discovery (IPS working group)
IETF/draft-bakke-iscsi-wwui-urn-00 - A URN Namespace for iSCSI World-Wide Unique Identifiers (IPS working group)

Overview

This proposes topics and text for an InfiniBand annex for the SCSI over RDMA (SRP) standard.

The goal is to identify all optional InfiniBand features that must be implemented to ensure useful, interoperable SRP devices. An annex in InfiniBand Volume 1, taken from Volume 3, will describe how boot devices, a subset of SRP devices, are specifically identified and the minimum command sets that may be depended upon (the “Storage Boot Wire Protocol”).

All text outside [brackets] is part of the suggested text.

Suggested text.

Annex A (normative)
SRP for InfiniBand™ Architecture

[footnote] InfiniBand is a trademark and service mark of the InfiniBand Trade Association.

A.1 Related documents

InfiniBand™ Architecture Volume 1 – General Specifications. Release 1.0. InfiniBand℠ Trade Association.

IETF RFC 2373 - IP Version 6 Addressing Architecture. R. Hinden and S. Deering. Internet Engineering Task Force.

A.2 Glossary

A.2.1 Introduction

See the InfiniBand Architecture Volume 1 glossary for full definitions of InfiniBand terms.

A.2.2 Definitions

A.2.2.1 **Channel adapter (CA):** Device that terminates a link and executes transport-level functions. Also called a node. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.2 **Communication manager:** The software, hardware, or combination of the two that supports the communication management mechanisms and protocols used to establish and release InfiniBand connections. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.3 **Consumer:** The direct user of verbs. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.4 **Global ID (GID):** A port address used for directing packets between subnets. A GID is a valid 128-bit IPv6 address. Source and destination GIDs are carried in an optional global routing header. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.5 **Globally unique identifier (GUID):** A number that uniquely identifies a device or component. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.6 **General service interface:** An interface providing management services other than subnet management. Uses the well-known queue pair 1. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.7 **IO unit:** One or more IO controllers attached to the fabric through a single channel adapter. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.8 **IO controller:** The part of an IO unit that provides IO services. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.9 **IPv6 address:** A 128-bit address constructed in accordance with IETF RFC 2373 for Internet Protocol version 6. See IETF RFC 2373.

A.2.2.10 **Local ID (LID):** A port address used for directing packets within a subnet. Source and Destination LIDs are carried in every packet header. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.11 **Management datagram (MAD):** A packet used for communication to manage an InfiniBand network. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.12 **Packet:** The indivisible unit of InfiniBand Architecture data transfer and routing, consisting of one or more headers, a packet payload, and one or two CRCs. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.13 **Port:** Location on a channel adapter to which a link connects. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.14 **Processor unit:** One or more consumers attached to the fabric through one or more channel adapters.

A.2.2.15 **Queue pair (QP):** An interface used for communication. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.16 **Queue pair number (QPN):** A value that identifies a queue pair within a channel adapter. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.17 **Service ID:** A value that allows a communication manager to associate an incoming connection request with the entity providing the service. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.18 **Subnet:** A set of InfiniBand ports connected via switches that have a common subnet ID and are managed by a common subnet manager. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.19 **Subnet manager:** Entity that configures and controls a subnet. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.2.20 **Verbs:** An abstract description of the functionality of a channel adapter. An operating system may expose some or all of the verb functionality through its programming interface. See InfiniBand™ Architecture Volume 1 – General Specifications, release 1.0.

A.2.3 Acronyms

CA: Channel adapter
CRC: Cyclic redundancy check
GID: Global ID
GUID: Globally unique identifier
IBA: InfiniBand™ architecture
IPv6: Internet Protocol version 6
LID: Local ID
MAD: Management datagram
QP: Queue pair
QPN: Queue pair number
REP: Communication Management Reply MAD
REQ: Communication Management Request MAD
RTU: Communication Management Ready to Use MAD
ROM: Read only memory
SRP: SCSI over RDMA protocol

A.3 InfiniBand Architecture overview

This annex specifies requirements for mapping SCSI over RDMA protocol (SRP) onto the InfiniBand Architecture (IBA), a transport that provides the necessary RDMA semantics.

An IBA processor unit contains consumers and one or more channel adapters, each containing one or more ports and queue pairs (QPs) (see Figure 1).

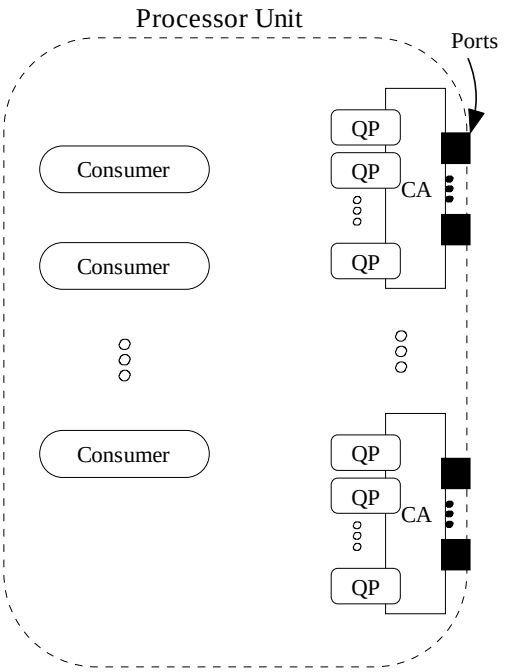

Figure 1. Processor unit (derived from InfiniBand Architecture specification).

An IBA IO unit contains a channel adapter with one or more ports, one or more queue pairs, and one or more IO controllers (see Figure 2).

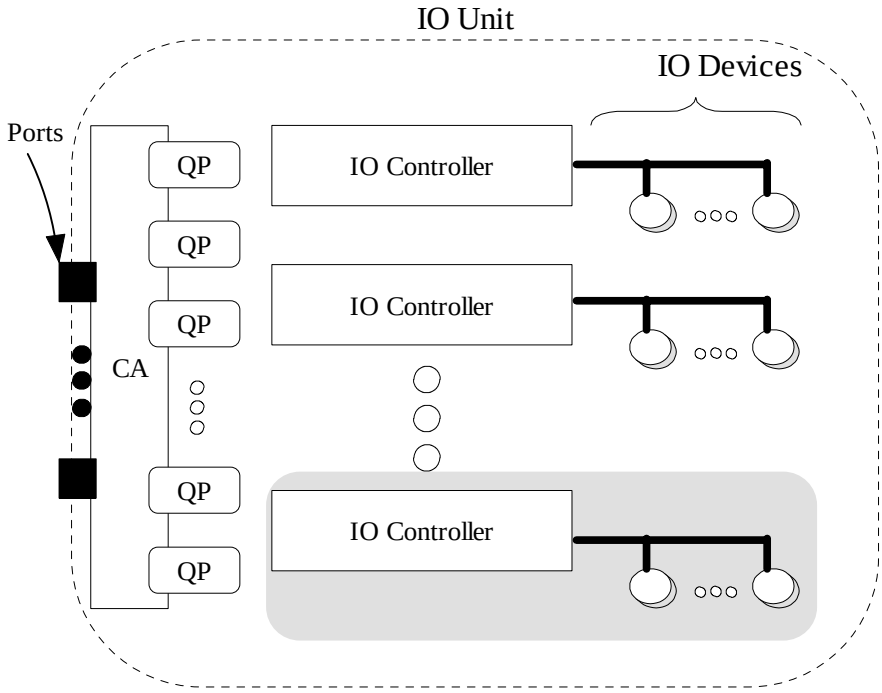

Figure 2. IO unit (derived from InfiniBand Architecture specification).

Each port has a 64-bit globally unique identifier (GUID) called a port GUID. Each channel adapter has a channel adapter GUID (which is shared by all ports on the channel adapter). Each IO controller has an IO controller GUID.

Each port is assigned a 16-bit local ID (LID) or a range of LIDs by the subnet manager. Each port has one or more 128-bit global IDs (GID). Each GID is globally unique, and may be formed in part from the port GUID. A GID is an IPv6 address. The subnet manager provides GUID to GID/LID resolution.

Table 1 summarizes the IBA names (GUIDs) and addresses (IDs) relevant to SRP.

Table 1. InfiniBand Architecture names and addresses

Name	Scope of uniqueness	Size	Description
Port GUID	worldwide	64 bits	Identifies a port within a channel adapter
Channel adapter GUID (Node GUID)	worldwide	64 bits	Identifies a channel adapter
IO controller GUID	worldwide	64 bits	Identifies an IO controller in an IO unit
LID	subnet	16 bits	Address assigned by the subnet manager to each port
GID (IPv6)	worldwide	128 bits	Address assigned by the subnet manager; (e.g. subnet prefix plus the port GUID)

A.4 SCSI architecture mapping

Figure 3 illustrates how SCSI initiator devices, initiator ports, target ports, and target devices map to InfiniBand Architecture objects. The figure also illustrates the mapping of the I-T-L nexus onto InfiniBand Architecture objects. The figure shows an initiator in a processor unit and a target in an IO unit.

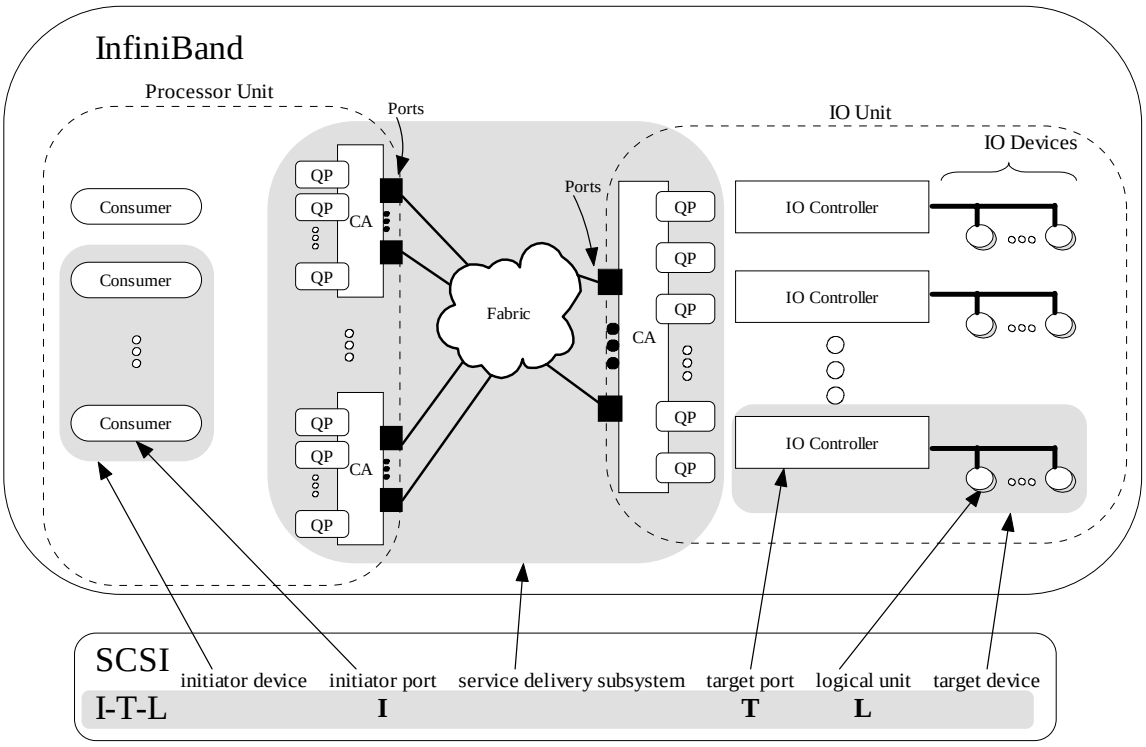

Figure 3. SCSI architecture to InfiniBand Architecture mapping example

SRP initiators may be implemented in processor nodes or IO units.

An SRP initiator device in a processor node is a set of consumers. An SRP initiator device in an IO unit is an IO controller.

An SRP initiator port in a processor node is a consumer. An SRP initiator port in an IO unit is an IO controller. The initiator port identifier shall be a worldwide unique identifier, and should be constructed by concatenating a GUID such as a channel adapter GUID with an identifier extension.

SRP targets may be implemented in processor nodes or IO units. In both cases, the SRP target shall include a device management agent to provide IOUnit, IOController, and ServiceEntries attributes and make available a worldwide unique IO controller GUID.

An SRP target device in a processor node is a set of consumers accessing a single channel adapter. An SRP target device in an IO unit is an IO controller plus one or more IO devices.

An SRP target port in a processor node is a consumer. An SRP target port in an IO unit is an IO controller. The target port identifier shall be equal to the IO controller GUID.

A service delivery subsystem contains queue pairs, channel adapters, and the InfiniBand fabric.

A.5 Communication management

A.5.1 Communication management overview

Communications managers on each InfiniBand device manage InfiniBand connections using MADs transported over the general service interface. SRP initiator and target ports shall use the active/passive (client/server) connection establishment protocol. The processor node or IO controller containing the SRP target port shall act as the server and the processor node or IO controller containing the SRP initiator port shall act as the client.

A.5.2 Discovering SRP target ports

To discover the service ID of an SRP target port in an IO unit, an SRP initiator port may use this sequence:
1) Retrieve the IOUnitInfo attribute from an IO unit using a DevMgtGet MAD to determine the presence and slot number of each IO controller attached to the IO unit;
2) retrieve the IOControllerProfile attributes from each IO controller, each of which includes a ServiceEntries table; and,
3) search the ServiceEntries table for entries with service names of “SRP.T10.NCITS”.

The service ID associated with each matching service name may be used in the communication management process to open InfiniBand connections to IO controllers acting as SRP target ports.

A.5.3 Establishing a connection

To establish an InfiniBand connection, the client places the service ID in a communication management Request (REQ) message. The server associates the request with the appropriate SRP target port. The PrivateData field of the REQ message shall include an SRP_LOGIN_REQ IU. The SRP target port may choose to refuse the connection based on the SRP_LOGIN_REQ IU content by using a communication manager reject value of “consumer reject.”

If the server accepts the connection request and SRP login, the server returns a queue pair number (QPN) in a Response (REP) message. The PrivateData field of the REP message shall include an SRP_LOGIN_RSP IU. The SRP initiator port may choose to refuse the connection based on the SRP_LOGIN_REQ IU content by using a communication manager reject value of “consumer reject.”

If the client accepts the connection reply and the SRP login response, it replies with a Ready To Use (RTU) message indicating both an InfiniBand and an SRP connection are open. It may start using the connection for communication.

A.5.4 Releasing a connection

Either the SRP initiator port or SRP target port may send an SRP_LOGOUT IU with a SEND operation. The sender shall disconnect upon receipt of an InfiniBand transport level acknowledgement to the SRP_LOGOUT IU. The sender may disconnect if its send queue has transitioned to an error state. The receiver of a LOGOUT IU shall respond with an InfiniBand transport acknowledgement and disconnect.

A.6 InfiniBand protocol requirements

SRP target ports and SRP initiator ports shall support the Reliable Connection transport service type.

SRP target ports shall implement the device management class of general management services.

SRP initiator ports and SRP target ports shall support the transport functions described in Table 2.

Table 2. Transport operation support requirements

Transport functions	SRP initiator port	SRP target port
Send to	Mandatory	Mandatory
Send from	Mandatory	Mandatory
RDMA write to	Mandatory	Not used
RDMA write from	Not used	Mandatory
RDMA read to	Mandatory for data-out commands	Not used
RDMA read from	Not used	Mandatory for data-out commands
RDMA Write with immediate data (to or from)	Not used	Not used
ATOMIC (to or from)	Not used	Not used

IO units containing an IO controller acting as an SRP target port shall report the device management IOUnit attributes as described in Table 3.

Table 3. IOUnit attributes for SRP target ports

Field	SRP requirements
Change ID	
Max Controllers	At least one
Option ROM	
Controller List	At least one IO controller must be present

IO controllers acting as SRP target ports shall report the device management IOControllerProfile attributes as described in Table 4.

Table 4. IOControllerProfile attributes for SRP target ports

Field	SRP requirements
[IO controller] GUID	
Device ID	
Vendor ID	
Device Version	
Subsystem Vendor ID	

Subsystem [Device] ID	
IO Class	[TBD see T10/01-104]
IO Subclass	[TBD see T10/01-104]
Protocol	[TBD see T10/01-104]
Protocol Version	[TBD see T10/01-104]
Service Connections	At least one
Initiators Supported	At least one
Send Message Depth	At least one
RDMA Read Depth [Editor's note: is this inbound or outbound? Cris Simpson thinks inbound. He will research and perhaps recommend IBTA delete these fields]	
Send Message Size [inbound or outbound?]	[Large enough to hold minimum SRP IU]
RDMA Transfer Size [inbound or outbound?]	At least one
Controller Operations Capability Mask	These bits shall be set to one: 0: ST (Send Messages to IO controllers) 1: SF (Send Messages from IO controllers) 5: WF (RDMA Write Requests from IO controllers) This bit shall be set to one by SRP target ports supporting data-out commands: 3: RF (RDMA Read Requests from IO controllers) These bits may be set to zero: 2: RT (RDMA Read Requests to IO controllers) 4: WT (RDMA Write Requests to IO controllers) 6: AT (Atomic Operations to IO controllers) 7: AF (Atomic Operations from IO controllers)
Controller Services Capability Mask	Bit 1 may be set for SRP ports with boot support: 1: SBWP Storage Boot Wire Protocol
Service Entries	At least one
ID String	

An IO controller acting as an SRP target port shall register with its Communications Manager a Service Name string of "SRP.T10.NCITS". This string is assigned an "IO SERVICE ID" type service ID by the Communications Manager.

IO controllers acting as SRP target ports shall include at least one ServiceName/ServiceID pair in the device management ServiceEntries attribute pair as described in Table 5.

Table 5. ServiceEntries attribute pair for SRP target ports

Field	Length	SRP requirements
ServiceName_n	320	"SRP.T10.NCITS"
ServiceID_n	64	Assigned by the IO controller

A.7 Extended Copy target descriptor

[Editor's note: this goes in SPC-3 rather than in the protocol standard.]

[existing section in SPC-3] 7.2.5 Descriptor type codes

Target descriptors and segment descriptors share a single set of code values that identify the type of descriptor (see table 16). Segment descriptors use codes in the range 00h to BFh. The definitions of codes between C0h and DFh are vendor specific. Target descriptors use codes in the range E0h to FFh.

Table 16 — EXTENDED COPY descriptor type codes (part 2 of 2)

Descriptor type code	Reference	Description	Shorthand
E0h	7.2.6.2	Fibre Channel World Wide Name target descriptor	
E1h	7.2.6.3	Fibre Channel N_Port target descriptor	
E2h	7.2.6.4	Fibre Channel N_Port with World Wide Name checking target descriptor	
E3h	7.2.6.5	Parallel Interface T_L target descriptor	
E4h	7.2.6.6	Identification descriptor target descriptor	
E5h	*7.2.6.x*	*SRP target descriptor* ***[this line is new]***	
E6h - FFh		Reserved for target descriptors	

[new section for SPC-3] 7.2.6.x SRP target descriptor format

The target descriptor format shown in table xx is used to identify a target using its SRP target port identifier.

The DESCRIPTOR TYPE CODE, PERIPHERAL DEVICE TYPE and NUL fields and the device type specific parameters are described in 7.2.6.1.

The LOGICAL UNIT NUMBER field specifies the logical unit within the SCSI device identified by the data in the TARGET PORT IDENTIFIER field that shall be the source or destination for EXTENDED COPY operations.

The TARGET PORT IDENTIFIER field contains the SRP target port identifier.

Table xx — SRP target descriptor format

Byte Bit	7	6	5	4	3	2	2	0
0	DESCRIPTOR TYPE CODE (E3H)							
1	RESERVED			NUL	PERIPHERAL DEVICE TYPE			
2	RESERVED							
3	RESERVED							
4 … 11	LOGICAL UNIT NUMBER							
12 … 19	TARGET PORT IDENTIFIER (8 BYTES)							
20 … 27	RESERVED							
28 … 31	DEVICE TYPE SPECIFIC PARAMETERS							

A.8 REPORT ALIASES and CHANGE ALIASES Identifier format

[Editor's note: This is based on Jim Hafner's T10/00-425r1, which is new and subject to change. This section may go into SPC-3 or SRP depending on the resolution of that proposal.]

The identifier formats in Table 6 are used by the REPORT ALIASES and CHANGE ALIASES commands to identify SRP target ports.

Table 6. Identifier formats for SRP

Protocol Identifier	Protocol Description	Type	Type Description	Format/Length
04h	SRP	00h	Target port identifier	8 bytes
04h	SRP	10h	InfiniBand GID with target port identifier checking	16 bytes + 8 bytes = 24 bytes

Table 7 describes the SRP target port identifier identifier format.

Table 7. SRP target port identifier identifier format.

Byte Bit	7	6	5	4	3	2	2	0
0	TARGET PORT IDENTIFIER (8 BYTES)							
..								
7								

The TARGET PORT IDENTIFIER field contains an SRP target port identifier.

Table 8 describes the SRP InfiniBand GID with target port identifier checking format.

Table 8. SRP InfiniBand GID with target port identifer checking identifier format.

Byte Bit	7	6	5	4	3	2	2	0
0	INFINIBAND GID (16 BYTES)							
..								
15								
16	TARGET PORT IDENTIFIER (8 BYTES)							
…								
23								

The INFINIBAND GID field contains an InfiniBand global identifier (GID) of an InfiniBand port connected to an SRP target port.

The TARGET PORT IDENTIFIER field contains the SRP target port identifier.

When a third party data manager first processes a segment descriptor that references this target descriptor, it shall confirm that the target port identifier is accessible via the InfiniBand GID. If the association cannot be confirmed, the third-party command shall be terminated because the target is unavailable. The third party manager shall track configuration changes that affect the InfiniBand GID value for the duration of the third party commands. An application client generating the third party commands is responsible for tracking configuration changes between commands.

A.9 Access Controls TransportID

[Editor's note: this goes in SPC-3 rather than in the protocol standard.]

The TransportID used by a target to identify initiators for access controls shall have the format described in Table 9.

[Editor's note: this is not currently InfiniBand specific – all SRP transports should be compatible]

Table 9. TransportID for SRP

Byte Bit	7	6	5	4	3	2	2	0
0	TYPE (02H)							
1	RESERVED							
..	RESERVED							
7	RESERVED							
8	INITIATOR PORT IDENTIFIER (16 BYTES)							
…								
23								