To: T10 Technical Committee
From: Greg Pellegrino (Greg.Pellegrino@compaq.com)
and Rob Elliott, Compaq Computer Corporation (Robert.Elliott@compaq.com)
Date: 3 January 2001
Subject: SRP InfiniBand™ annex

Revision History
Revision 0: first revision

Related Documents
srp-r01 – SCSI over RDMA revision 1
InfiniBand™ Architecture Release 1.0: Volume 1 – General Specifications
InfiniBand™ Architecture Release 0.9: Volume 3 – Application of InfiniBand

Overview
This proposes topics and text for an InfiniBand annex for the SCSI over RDMA standard. InfiniBand's volume 3 section 9 (Storage Boot Wire Protocol) should be largely removed, deferring to SRP and this annex.

The goal is to identify all optional InfiniBand features that must be implemented to ensure useful, interoperable SRP devices.

Suggested text.

Annex A (normative)
SRP for InfiniBand™ Architecture

A.1 Overview
This annex describes how SCSI over RDMA protocol is used over InfiniBand, a transport that provides the necessary RDMA semantics.

An IBA host contains one or more SRP initiators. The IBA host may use multiple GIDs to represent multiple SRP initiators. Within each GID, the IBA host may use multiple QPs to represent multiple initiators. The host uses different QPs to communicate with different targets.

An IBA target contains one or more SRP targets. Each SRP target is an IBA IO controller. The IBA target may contain multiple GIDs or just one GID to represent multiple SRP targets. Within a GID, the IBA target may implement multiple QPs to represent multiple targets. The target uses different QPs to communicate with different initiators.

A.2 Communication Management
Communications Managers manage connections using MADs over the GSI on each system. The Active/Passive (Client/Server) model is used. The SCSI target acts as the server and the SCSI initiator acts as the client.

The client places a ServiceID in a Request (REQ) message. The server associates the request with the appropriate SCSI target. The server returns the Queue Pair Number (QPN) in a Response (REP) MAD.

Alternate Path Migration (APM) may be supported. The initiator or target may include a second address in its REQ or REP packet that may be used in case of failure of the path to the original address.

Editor's note: if Reliable Datagrams were supported, then the EE Context would need to be retained along with the QPN.

Connections are created and destroyed through the Communication Manger (CM) residing at each node. The IO controller registers with its CM and specifies a protocol string of "SBWP.IBTA" which is assigned an IO SERVICE ID type service ID by the CM.

To establish a connection, the host sends a DevMgtGet MAD with AttributeID=ServiceEntries to the IO unit to get the IOC's protocol list which lists all the supported protocols and their associated Service IDs. The host filters the list to find the Service ID for "SBWP.IBTA". The host then initiates the connection with the IO controller by specifying that Service ID.

To release a connection, both Targets and Initiators may terminate a connection by sending a Communication Management Disconnection Request, CM DREQ, as described in IBA volume 1 release 1.0 section 12.6.10.

A.3 Initiator-specific requirements

Initiators shall support these inbound transport functions: Send, RDMA Read, and RDMA Write.

An SRP host shall have at least one QP dedicated to SRP conforming to the connection requirements of SRP.

Initiators shall follow all the HCA rules.

Should an initiator be required to check inbound RDMA transfer lengths? (Volume 1 Section 9.7.4.1.6) It can verify that the maximum transfer size of 2^31 was not exceeded, and that the sum of the packet payloads adds up to the requested transfer size.

A.4 Target-specific requirements

Targets shall support this inbound transport function: Send.

An IO controller indicates that it supports the SRP storage protocol by setting the SRP bit in the IO Controller Profile accessible via a DevMgtGet MAD. IOControllerProfile.ControllerServicesCapabilityMask.SBWP=1). Refer to IBA volume 1 release 1.0 Table 222 in section 16.3.3.4.

The target shall provide one Queue Pair for SRP for each host that it supports.

A target is not required to support more than one host at a time and may support multiple host connections as long as each connection and its operation are independent from the others.

If more hosts request connections than a SBWP IO controller has QPs, then the IO controller rejects additional connection requests with Rejection Reason=1 - No QP Available, as described in IBA volume 1 release 1.0 section 12.6.7.

SRP does not require the target make any memory available to the host.

Is End-to-end flow control required in the target? It is required in the host.

Targets shall follow all the TCA rules.

Targets implementing multiple GIDs shall follow all the Router rules.

Should the target be allowed to issue RDMA Write with Immediate Data when running data-in transfers? The AETH Credit Count (CCCCC) (Vol 1 Section 9.10.1) indicates how many receive work queue elements (WQE) are available for Sends (used for SRP management and AER) or RDMA Write with Immediates. If the host only wants to dedicate one WQE to keep the management stream going and the target uses it for RDMA Write data, is there a problem?

Should SRP targets be required to assume trivial subnet manager responsibilities?

A.5 Common initiator and target requirements

Targets and initiators shall support the Reliable Connection transport service type. Reliable Datagram support is not defined (this include Q_Keys and EE Contexts).

Atomic operations are not used by SRP.

Are Solicited Events needed?

Path MTU size support is device-specific.

Multicast operations are not used by SRP.

Should automatic failover support (path migration) be mandatory? We could require the target support responding to automatic path migration requests (Alternate Path Response APR) but not require it to generate them (Load Alternate Path LAP) if it loses contact with the initiator. This way the initiator can force a failover if it wants. This feature would be optional to the initiator.

Any Service Level may be negotiated and used. Any Virtual Lanes may be negotiated and used.

Should SRP devices be required to support more than two P_KEYs per port, the minimum IBA requirement?

Should SRP devices be required to support the Bad P_Key Trap and P_Key Violations Counter (Volume 1 section 10.9.3-4)?

Should SRP devices be required to check M_Keys for reads? (10.9.9) Should they be required to maintain M_Keys through power loss?

Any minimum RNR NAK Timer field values?

Do we need to require the Fence Indicator be used at any times? (Section 10.8.3) If the target does an RDMA READ then a SEND to indicate command completion, Table 67 seems to imply the SEND might be initiated (but not completed) before the RDMA READ completes. Would the initiator release the application client data-out buffer at this time?

DevMgt Class?

Performance management features are all optional (16.1.3.2)

The Device Management MADs shall be supported (Section 16.3) These indicate the number of IO Controllers supported by a TCA (maximum FFh), the EUI-64 of each IOC, the protocol supported (SRP!), etc. [All the fields should be reviewed and those required should be listed]

IBA volume 3 section 9.3 mentions something called a “Persistent Device Identifier”. This annex should describe how it is allocated and the format. It should have attributes similar to the WWNs used in FCP; nonvolatile persistence, worldwide uniqueness.

A.6 Extended Copy and Access Control Descriptors

Editor's note: update to match Jim Hafner's ALIAS IN/OUT terminology

The target descriptor contains:

Node GUID
Service ID

The copy manager should query the subnet manager to map the Node GUID into a GID/LID when needed. The final target maps the Service ID into a QPN when the copy manager opens communication.