To: T10 Technical Committee
From: Rob Elliott, Compaq Computer Corporation (Robert.Elliott@compaq.com)
Date: 29-30 November 2000
Subject: Minutes of the SRP WG - November 29-30, 2000 – Redmond, WA

Revision History
Revision 0: first revision

Attendance

Name	Company	EMail
T10 members:		
Dave Peterson	Cisco	dap@cisco.com
Rob Elliott	Compaq	robert.elliott@compaq.com
Dal Allan	ENDL	endlcom@acm.org
George Penokie	IBM/Tivoli	gpenokie@tivoli.com
Ed Gardner	Ophidian	eag@ophidian.com
Mark Evan	Quantum	mark.evans@quantum.com
Visitors:		
Litko Chan	Agilent	litko_chan@agilent.com
Bill Edwards	Compaq	bill.edwards@compaq.com
Greg Pellegrino	Compaq	greg.pellegrino@compaq.com
Bob Nixon	Emulex	bob.nixon@emulex.com
Cris Simpson	Intel	cris.simpson@intel.com
Russ Henry	LSI Logic	russ.henry@lsil.com
Keith Holt	LSI Logic	keith.holt@lsil.com
Rob Haydt	Microsoft	robhay@microsoft.com
Frank Campbell	QLogic	frank.campbell@qlogic.com
Seth Abrahams	Sun	seth.abrahams@sun.com
David Brean	Sun	david.brean@east.sun.com
Bill Terrell	Troika Networks	terrelll@troikanetworks.com

Results of meeting
Ed Gardner opened the SCSI over RDMA protocol meeting at 1pm Wednesday 29 November 2000 and thanked Rob Haydt of Microsoft for hosting. This protocol standard maps SCSI over InfiniBand™ Architecture, Virtual Interface (VI) Architecture, and similar transports supporting RDMA (remote direct memory access).

Agenda
The agenda was created at the meeting.

1. Scatter/Gather lists and bidirectional data transfers for SRP (00-410r0 by Ed Gardner)
2. Mode pages for SRP (00-426r0 by Rob Elliott)
3. SRP Asynchronous Event Reporting (01-010r0 by Rob Elliott and Greg Pellegrino)
4. RSP IU structure (discussion)
    a. Sense data location
    b. Residual data
    c. Bidirectional residual
5. Identifying devices in SRP (00-357r1 by Ed Gardner)
6. Flow control (discussion)
7. What to negotiation (discussion)
8. Document review (srp-r01 by Ed Gardner)

9. Meeting schedule

Scatter/Gather and Bi-directional Data Transfers for SRP (00-410r0 by Ed Gardner, Ophidian Designs)

[Wednesday]

The group discussed why scatter-gather is needed when the host adapter - the Host Channel Adapter (HCA) in InfiniBand – could do the mapping itself. Reasons given were that memory registration is a hot spot in kernel. With scatter-gather support, the host driver could assign a region of memory for I/O and let applications use SG lists to point to buffers scattered within that region.

Straw poll to work on scatter-gather proposal: unanimous

Two terms were chosen:

- A memory descriptor (MD) consists of a 64-bit address, 32-bit length, and 32-bit memory handle (R_KEY in InfiniBand).
- An indirect memory descriptor (IMD) consists of a 64-bit address, 32-bit length, and 32-bit memory handle (R_KEY in InfiniBand). The address and memory handle point to the SG table in host RDMA space. The length is the length of the SG table in bytes. This corresponds to the number of MDs in the SG table * 16 (16 is the size of each MD in bytes).

Dal suggests negotiating how many memory descriptors are in the CMND IU

The group discussed four models for what the CMND IU could contain (for unidirectional commands):

1. address of data - no scatter-gather list
2. address of data and an indirect pointer - one address to let the target start a data transfer before fetching the SG table
3. n addresses of data - scatter-gather list is entirely contained within the CMND IU; limited to length of n
4. indirect pointer only – all data transfers require a RDMA READ of the SG table

Two additional variations were created:

2b. n addresses of data and indirect pointer – several addresses to let the target start data transfers before fetching the SG table

4b. indirect pointer and n addresses copied from the table - like SBP-3 Fast Start proposal

The group preferred the last proposal (4b). This includes some addresses so the target can issue the RDMA READ for the SG table, then immediately start using RDMA WRITE to return SCSI read data. Full-duplex transports like InfiniBand can overlap the read and write data transfers. Since the addresses are copies from the table, targets are not required to use them – they can always fetch the SG table if they don't want to store the addresses.

Ed diagrammed a CMND IU based on 4b. The pertinent fields are:

type code
indirect bit
Number of MDs copied into the CMND IU
...
CDB length
CDB (may be variable length)
,,,
IMD (if the indirect bit is set)
n MDs

The group requested that existing opcodes be optimized. Three type codes would be used
1. one for <=16 byte CDBs (16 byte CDB field; zero fill unused bytes for 6, 10, and 12-byte CDBs)
2. one for <=32 byte CDBs (32 byte CDB field; zero fill unused bytes)
3. one for >32 byte CDBs (variable length CDB field; no zero fill)

Initiators must use the smallest function code possible. For example, 6-bit CDBs could not be sent in the >32 byte type code.

It was noted that no total data length is included in the CMND IU, unlike FCP which has a Data Length (DL) field in its CMND IU. An InfiniBand SRP to Fibre Channel FCP bridge will have to read the full SG table and add up the MD lengths to create the proper FC DL field.

The SG table is just a series of MDs. The length of the table is indicated by the length field in the IMD. The group agreed that no SG chaining is needed.

The group discussed bidirectional commands. The solution recommended was to add three additional type codes, representing the same CDB sizes described above. For bidirectional commands, two sets of IMDs, MDs, indirect bits, and Number of MDs would be present. The first is for the Data Out direction and the second is for the Data In direction.

The group agreed to keep one byte length field for unidirectional with the RDDATA and WRDATA bits indicating direction as in FCP.

[Thu]
Ed presented a new format, dropping the multiple type codes. The pertinent fields are:

Data out indirect bit
Data in indirect bit
Data out Number of MDs
Data in Number of MDs
...
CDB length
CDB (may be variable length)
,,,
Data Out IMD (if the indirect bit is set)
n MDs for Data Out
Data In IMD (if the indirect bit is set)
n MDs for Data In

The group reasserted that it wants type codes that optimize 16 byte and 32 byte CDBs, with a third for >16 variable length CDBs.

The group discussed the order of the IMDs and MDs. Possibilities include (each item of each is optional):

Option A	**Option B**
Data Out IMD	Data Out IMD
Data Out MDs	Data In IMD
Data In IMD	Data Out MDs
Data In MDs	Data In MDs

The group preferred option A.

Ed asked whether task management functions should use a different type code than commands. They are currently overloaded in protocols like FCP and SPI. The group agreed to use a separate type code for SRP. The task management format will still have both a tag field and a "tag of task to be managed" field.

Ed will prepare a revision 1 based on inputs in the WG for the next SRP meeting.

Mode pages for SRP (00-426r0 by Rob Elliott, Compaq)
[Wed]
This proposal adds the disconnect-reconnect mode page to SRP, including the EMDP (enable modify data pointers) and Maximum Burst Length fields. It also adds text for the Port Control mode page and Logical Unit Control mode page, although no bits are defined in those pages.

Two editorial changes were requested:
1. strike interconnect tenancy sentence
2. strike FCP LUN 0 sentence

The working group recommended unanimously that 00-426r0 as revised (to 00-426r1) be included in SRP.

SRP Asynchronous Event Reporting (01-010r0 by Rob Elliott and Greg Pellegrino, Compaq)
[Wed]
This proposal suggests adding special support for asynchronous event reporting in SRP.

The group requested several changes:
1. The group felt it appropriate to mandate this feature rather than make it optional. Use the same wording as CMND_IU to describe support for the AER IUs. Both initiators and targets must support the AER IUs, just like CMND and RSP IUs.
2. use these names for the IUs: SRP_AER and SRP_AER_RSP
3. use the SRP_RSP data format
4. keep the tag field
5. verify the lengths of the IUs are correct in the table

The working group recommended unanimously that 01-010r0 as revised (to 01-010r1) be included in SRP.

RSP IU structure
[Thu]
Someone requested that sense data be in a fixed location in the RSP IU. The response data field could be set to a fixed length to avoid the problem of two variable-length fields. The group noted that this was just copied from SPI and FCP. To be compatible and make bridges work, the group agreed leave it the same.

While reviewing the RSP IU structure, the group noted that the residual data field needs better definition when scatter-gather is added. Unlike FCP there is no total DATA LENGTH field in the CMD IU; however, the sum of the LENGTHs in all the MDs can be calculated. It was alsos noted that a bidirectional read overflow bit, bidirectional read underflow bit, and a bidirectional read residual field need to be added for bidirectional commands.

Identifying device in SRP (00-357r1 by Ed Gardner, Ophidian Designs)
[Thu]
The group discussed talk-through vs. talk-to at length.

00-357 term	Another term	Meaning
IB-SCSI	talk-to	A single SCSI target is presented on the SRP transport (e.g.

		InfiniBand) interface. The target contains multiple LUNs. For a bridge (gateway, router, ...) the LUNs may be compiled from multiple targets behind the bridge.
IB-FC	talk-through	Multiple SCSI targets are presented on the SRP transport interface. The target/LUN relationship from devices behind the behind the bridge (gateway, router, ...) is preserved.

Ed presented some text from an email proposing some commands to map devices to LUNs on a single target (the talk-to model).

The group decided that, as with other SCSI protocols, the transport sets up the I_T nexus and the SRP protocol sets up the L_Q. In other words, SRP does not need to concern itself with the target identifier for the transport carrying SRP traffic. In InfiniBand, the target may be addressed by a LID (local ID), GID (global ID), and queue pair number.

George Penokie took an action item to investigate the 8-byte LUN field in SAM-2 and SCC-2 and describe it in the next meeting.

Flow Control
[Thu]
Someone asked for the reason SRP include application level flow control (e.g. MaxCmdIU, MaxRspIU, etc.) when InfiniBand already includes flow control.

1. This protocol is not designed solely for InfiniBand. Generic VI transports do not include flow control – if a queue is overflowed, the entire connection is closed.
2. IB flow control may not be sufficient. IB flow control stalls the whole connection, preventing task management functions from being issued. SRP potentially allows prioritization.

The group debated whether the service type for IB connections should be mandated. For example, require RC (reliable connection; one-to-one connection from the initiator to the target) and prohibit RD (reliable datagram; the initiator connects one queue pair to many targets' queue pairs). Ed argued that there is nothing in SRP prohibiting it from operating on RD, and some sort of profile document is where requirements of that sort belong. Rob Haydt argued that SRP should include any requirements necessary to create an interoperable hardware market. It was noted that SBP-2 (SCSI over 1394) include a normative annex detailing "Minimum Serial Bus node capabilities."

What to negotiate
[Thu]
Given the changes the scatter-gather list adds to the CMD IU format, Ed asked what exactly should be negotiated during login. The group agreed that these two values should be negotiated:

- maximum IU length (as in the current specification)
- whether indirect MDs are supported

The latter value decides between two cases:
a) single MD per direction in the command IU
b) IMD and cache of MDs allowed per direction (cache size determined by the maximum IU length)
If indirect MDs are supported, both a) and b) are supported. If it is not supported, only a) is supported.

The group rejected negotiation of these values:

- maximum CDB size – this is implicit in the IU length
- maximum SG list size – for MDs within the IU, the maximum IU length covers this, for MDs in host memory, the only limit is the length field of the IMD

- which type codes supported – the group felt that all the type codes for different sized CDBs are mandatory.

Meeting schedule
[Thu]
Another SRP meeting will be held in Houston, TX at the Wyndham Greenspoint, hosted by Compaq.

4 January 2001 Thu 9am – 6pm
5 January 2001 Fri 9am – 1pm

The next T10 week in Orlando, FL at the Grosvenor Resort, hosted by Adaptec, will also have special SRP meetings in addition to the usual CAP meeting.

17 January 2001 Wed CAP (Commands Architecture and Protocol)
18 January 2001 Thu SRP (begins after T10 plenary, usually around 2pm)
19 January 2001 Fri SRP 9am – 1pm

Ed Gardner intents to request a letter ballot after the 19 Jan meeting.

The meeting adjourned around 3pm Thursday.