

ICSC ITWG Meeting Minutes, 10 May 2005

Taking minutes: HP (jr)

Meeting attendance diagram (if you have more than 1 minus within the last four ITWG meetings, you are not eligible to vote):

mtg	date	hp	ibm	netapp	sun
---	-----	--	---	-----	---
m-3		+	+	-	-
m-2		+	+	-	-
m-1		+	+	-	-
m-0	may 10	+	+	-	-

Present:

jim hamrick (jh) [HP] (left after 1 hour)
fredy neeser (fn) [IBM]
fred worley (fw) [HP]
jay rosner (jr) [HP]

Minutes to approve:

- Email from Jay Rosner, subject "ICSC ITWG draft minutes for 19 April 2005", sent 04/20/2005 - **Approved**.
- Email from Fredy Neeser, subject "ICSC ITWG draft minutes for 26 April 2005", sent 04/27/2005 - **Approved**.

Pending Action items:

- **AI (FN)**: Revisit `it_lmr_create` to determine if an address space qualifier is necessary
- **AI (FW)** - pursue discussion of RNIC failover and host memory movement w.r.t. memory registration.
- **AI (FN)**: Report back to ITWG on semantics of IOVA output in IB's Register Physical Memory verb.
- **AI (JR)**: Put IB Atomics requirements to vote.

New Action items:

- **AI (ITWG)** - deprecate "out of resources" error code from the Bind calls in IT-API 2.1 and add Bind to list of routines valid to call from a callback.
- **AI (All)** - research whether Terminate message can be correlated to pending RDMA Read.
- **AI (JR,FW)** - determine if option Two (supporting non-aligned relative addressing for Atomics) is truly necessary and return a proposal.
- **AI (JR)** - add requirement for Atomics stating that if the Consumer wants to use non-coherent memory at the Atomic target and wishes to access the Atomic target operand via CPU rather than HCA, then they must use the sync operations.
- **AI (All)** - consider whether we wish to rename `it_lmr_sync_rdma_read` and `it_lmr_sync_rdma_write` to more intuitive names and then deprecate the existing calls.
- **AI (JR)** - find the original note on the subject of posting API to ITWG page and follow up with Martin.

- **AI (FN)** - Ask Martin/Cathy whether possible to make hyperlinks blue in PDF documents.

Action item review:

- AI (FN, JR): Revisit `it_lmr_create` to determine if an address space qualifier is necessary
 - FN - still pending
 - JR - does not see this as necessary
 - FN - asks for another week to think about this
 - **Pending**
- AI (JR, FN) - determine if bind does do allocation for some IB implementations (i.e. is there a likely issue)
 - FN - did not get any feedback on this issue, but given that HP derived an answer that this is not an issue, FN okay with allowing bind in ISR
 - FN - asks if the error code should be removed from the `it_rmr_link` call
 - JH - agrees
 - **AI (ITWG)** - deprecate "out of resources" error code from the Bind calls in IT-API 2.1 and add Bind to list of routines valid to call from a callback.
 - **Closed**
- AI (FW) - pursue discussion of RNIC failover and host memory movement w.r.t. memory registration.
 - FW - investigating use of space ID in `lmr_create` under failure scenarios
 - FW - however found implementation options to be able to work around lack of space ID in `lmr_create` call
 - FN - asks for failover case refresher discussion
 - FW - reminds us that the original question was whether having a space ID parameter in `it_lmr_create` would be beneficial to a failover implementation
 - FW - describes the idea of moving memory registrations from one card to another - will the regeneration of IOVA (bus address) be problematic without the space ID?
 - FW - thinks that if the failover is performed internally to IT-API implementations, then no space ID needed. If the failover is done by applications outside of IT-API, then space ID would be probably required but FW thinks there is no way to support reprogramming the LMR in a secure way
 - FN - wonders if FW should explore this further?
 - FW - wonders if the IT-API already includes a capability to perform some feature, should the additional features be added?
 - JH - general IT-API philosophy has been to avoid functional redundancy except where there is a profound performance benefit.
 - FW - not fully yet convinced of performance implications - thinks that space ID allows a different use model - one where an IT-API Consumer can be a

service provider to other applications outside of their context

- FW gives an example of a storage subsystem provider that performs all memory registration on the behalf of the upper level consumers
- o FN - asks for a summary of the two options for handling failover
 - FN - thinks one option is IT-API consumer handles failover
 - FN - other option is IT-API implementation handles failover
 - FW - agreed
 - FW - notes that his preference is failover underneath IT-API and above RNIC-PI since IT-API is an OSV component.
 - FN - notes that DAT is looking at high availability
 - FW - will release a summary of his thoughts for our review
- o Pending
- AI (FN) - to research basis of following text in ASYNCHRONOUS ERRORS section of `it_post_rdma_read`: "Despite using the RC service, an RDMA Read DTO may fail to successfully deliver the contents of the section of the remote buffer into the local buffer. This may result in a broken Connection or lead to data corruption in the local buffer, which is detected locally with high probability. In case of local data corruption, either an `IT_ASYNC_AFF_EP_L_LLP_ERROR` Affiliated Asynchronous Error will surface after the DTO has already completed with `IT.DTO_SUCCESS`, or an `IT.DTO_ERR_TRANSPORT` Completion Error will occur."
 - o See also Minutes of the 26. April meeting and Email by Fredy Neeser, subject "Re: RDMA Read Failures on iWARP", sent 04/27/2005.
 - FN - gives a summary of his reply to Caitlin
 - FN - Say an RDMA Read causes an access violation at the remote peer. This causes a Terminate message to be generated by the remote. FN claims that it is easy to correlate the Terminate message to the offending request.
 - FN - notes that CB had made a comment that converting a Terminate message to an error would not always be possible.
 - o Specific case was where there were subsequent identical RDMA Reads and one fails
 - FN - thinks that processing of an incoming RDMA Write to a corresponding outstanding RDMA Read request (where incoming RDMA Write is response) is of the same order of difficulty as matching an incoming Terminate message to the outstanding RDMA Read.
 - o JR - was of the impression that Terminate processing was likely to be done in S/W as opposed to RDMA Read response processing

likely done in H/W (at the Verbs provider level).

- FN - would like to mandate that on receipt of a Terminate message that CAN be mapped to an outstanding RDMA Read, that a Completion error be manifested rather than an Async error.
- FN - has some proposed text for the error handling in RDMA read - will send out after meeting.
- JR - notes that we all need to explore this issue.
- AI (All) - research whether Terminate message can be correlated to pending RDMA Read.
- o Closed
- AI (FN): Report back to ITWG on semantics of IOVA output in IB's Register Physical Memory verb.
 - o FN - asked internal folks about this, but has received no response.
- o Pending
- AI (JR): Put IB Atomics requirements to vote.
 - o JR - will be put Atomics to vote for May 17th if discussion below is not controversial

IB Atomics

- Alignment restrictions for Absolute vs. Relative Addressing
 - o JR - describes the requirement to ensure RDMA address of the Atomic target is 8-byte aligned.
 - o FN - describes that the additional requirements added into the specification give the Consumer an algorithm to be able to determine that they are using an 8-byte aligned buffer.
 - o FN - describes the need to constrain further the relative addressing case - we need to ensure that the Atomic-initiating Consumer be able to determine the alignment.
 - JR - thinks that the alignment can be determined at the remote and theoretically could be conveyed to the local (via some ULP).
 - FN - inclined to further constrain the use of relative addressing so that Atomic-initiating Consumer can use 8-byte aligned offsets and guarantee 8-byte alignment.
 - FN - to achieve this, could constrain use of relative addressing for Atomics to require that the relative region MUST be 8-byte aligned.
 - o JR - states that there are the two options:
 - One - align the Relative region to 8-byte alignment and enable initiator to guarantee alignment by simply keeping FBO 8-byte aligned
 - Two - not constrain Relative region alignment and require remote and initiator to communicate FBO modulus needed to attain 8-byte alignment
 - o AI (JR,FW) - determine if option Two (supporting non-aligned relative addressing for Atomics) is truly necessary and return a proposal.
- Given that atomicity is guaranteed only on a single IA, how is the remote IA handling atomics identified/selected?

- o FN - feedback internally made him ask if a remote node has more than one HCA, how does the local Consumer determine which HCA to target to get atomic operations.
 - o JR - thinks this is strictly a ULP issue - the HCA arbitrating the use of Atomics is uniquely named (by address) in the fabric - the ULPs need to be aware of which device to target.
- Do we need an `it_lmr_sync_atomic` to make incoming atomics visible in case of non-coherent memory?
 - o FN - notes that we have synchronization calls for RDMA Reads and RDMA Writes.
 - o FN - describes a scenario where a barrier is performed by reading a counter via the local CPU where that counter is being Fetched-and-Added by remote nodes via Atomics. Two options exist:
 - One - at the barrier controller, read the counter using HCA Atomics
 - Two - simply read the counter with the CPU
 - o For scenario One, no coherency issues. For scenario Two, if the counter is in non-coherent memory, then synch operations need to be performed
 - o AI (JR) - add requirement for Atomics stating that if the Consumer wants to use non-coherent memory at the Atomic target and wishes to access the Atomic target operand via CPU rather than HCA, then they must use the sync operations.
 - o AI (All) - consider whether we wish to rename `it_lmr_sync_rdma_read` and `it_lmr_sync_rdma_write` to more intuitive names and then deprecate the existing calls.
- Email from Jay Rosser, subject "Third draft - IB Atomics detailed requirements", sent 05/09/2005
 - o Described above (in AIs).
 - o Draft needs to be updated with non-coherent AI and investigation of the relative addressing issue needs to be completed.
 - o Ballot should be postponed until above resolved.

Posting a list of Work Requests

- Continue discussion of detailed requirements and feedback
- Postponed till next meeting.

Various email threads

- Email by Caitlin Bestler, subject "Access Layer locking implications of multiple-element SGLs", sent 05/09/2005.
 - o FN - CB had looked at the possible implementations of SGLs in DAPL and IT-API
 - In DAPL, STag, Address, Offset
 - In IT-API, LMR Handle, Address, Offset
 - o FN - CB concluded that parallel array on the stack is desirable
 - o JR - is there any action we need to take here?
 - o FN - no
- Email thread started by Caitlin Bestler, subject "Implications of the missing RTR State", sent 05/02/2005.

- o FN - describes this as similar to our discussion of the subject.

Any other business

- ICSC ITWG "native" home page
(<http://www.opengroup.org/icsc/native/protected/>): Can IT-API v2.0 and header files be made visible on this page?
 - o AI (JR) - find the original note on the subject of posting API to ITWG page and follow up with Martin.
- Header file
 - o Snafu fixed and header differences file posted.
- Non-coherent memory w.r.t. RMRs.
 - o FN - If Consumer has linked an RMR to an LMR in non-coherent memory, how should Consumer sync the memory?
 - o JR - ideally only using the subset of memory represented by the RMR (since synch is an expensive operation)
 - o JR/FN/FW - we think the existing calls are sufficient for RMRs. No action needed.
- Hyperlinks are not colored in document and PDF.
 - o AI (FN) - Ask Martin/Cathy whether possible to make hyperlinks blue in PDF documents.

Ending discussion:

- None

Meeting adjourned at the 115-minute mark