

ICSC ITWG Meeting Minutes, Dec. 14, 2004  
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	dec 14	+	+	-	-

Present:

jim hamrick (jh) [HP]  
fredy neeser (fn) [IBM]  
jay rosner (jr) [HP]

Minutes to approve: Email from Jay Rosner, subject "Re: ICSC ITWG draft minutes for 11/30/04 (second draft)", sent 12/6/04 9:40PM PT - [Approved](#).

Action item review

- AI (FN): Adjust current text describing how to unbind a narrow RMR after QP disconnected.
  - FN - Modified pages
  - Closed
- AI (FN): Add a sentence per RDMA DTO man page describing access privileges for sinks and sources.
  - FN - Modified all the man pages
  - JR - Merge still needed for some pages (editorial issue)
  - Closed
- AI (FN): Add advice to Consumers that creating write-only local access is a transport-dependent programming practice; providing for both local read and local write is transport-independent. Make sure `IT_PRIV_DEFAULT` is the portable value.
  - FN - Wouldn't `IT_PRIV_LOCAL_DEFAULT` be a better name?
    - § JH - inclined to keep `IT_PRIV_DEFAULT` since it does not require the application writer to think about access privileges and they still get useful LMRs for sends/receives without exposing themselves to risk
    - § FN - okay, we will keep `IT_PRIV_DEFAULT` as is.
  - FN - Information added to access privileges to represent this in `it_lmr_create` man page

⊖ Closed

- AI (FN): Publish updated MM man pages by 12/10/04.
  - ⊖ FN - released pages yesterday addressing all calls except for `it_post_rdma_read_to_rmr()` (cannot release this page since it depends on a vote that is still pending).
  - ⊖ Closed
  
- AI (FN): Point out cases where marker control for IT-API Consumer still has value even if the IETF proposal for RNIC compatibility is adopted.
  - ⊖ Pending
  
- AI (JR): Add discussion of how IRD/ORD negotiation is done into app usage section of `it_ep_connect` man page, and reference to it from app usage section of `it_ep_accept` man page.
  - ⊖ Pending
  
- AI (JR): Determine if there are any remote operations in iWARP (such as remote invalidation of an MR or MW) that can be performed on things that only have local access permission.
  - ⊖ FN - the answer for iWARP is NO - in section 7.8 of RDMAC iWARP verbs (pg 115 lines 1-4).
    - § FN - remote access rights must be enabled to allow an incoming remote operation.
    - § FN - MM-21.2 and MM-21.3 always set the remote access rights.
    - § FN - second condition is Stag must have remote read/write privileges.
  - ⊖ JH - motivation behind AI was to determine if there is a scenario where a Consumer would end up confused about the context of an "RMR context" (i.e. is it a local or remote entity and can the Consumer be mistaken about use of it).
    - § FN - States that for an RMR context based on an LMR, cannot have no remote access.
    - § JH - Thinks that the "RMR context" is reasonable given the above.
    - § FN - Agrees with the legacy API, but notes that remote invalidation can fail for one class of "RMR context" (where based on an LMR) vs the other class of "RMR context" (window, fast registered, etc).
      - FN - thinks the iWARP consumer will be confused by this
      - FN - thinks that expressing error semantics is very difficult with existing terminology
  
- AI (JR): Generate a proposal for a mapping service to generate a PBL (with bus addresses) given a virtual address range and a virtual address space ID(virtual addresses)
  - ⊖ May discuss under "If time permits" below
  - ⊖ Pending

- AI (JR): Produce requirements for RDDP WG informational draft to support IOH functionality.  
      $\bar{O}$  Pending

#### Bitwise vs logical OR

- FN - did a search on "logical" in IT API 1.0 and found the bit sets shown in the pages really were intended to be bitwise used rather than logically.
  - JR - then action is simply to change "logical" to "bitwise" where it occurs.
  - FN - yes
- JR - notes that FN had sent private email listing the other pages where this shows up.
  - AI - JR - revise the man pages FN suggests.

#### MM issues (<= 40 min)

- Vote on email proposal sent out by FN on Dec. 7 with subject "AI - Revised proposal for cleaning up access privilege bits"
  - JH - we were going to break source code compatibility for Consumers where they were using "0" as a privilege
    - § FN - besides this, all other constants are kept and should have compatibility.
    - § JH - since we are breaking compatibility in other areas, this is acceptable.
  - FN - notes that this proposal fixes a number of errata.
    - § JH - agrees
  - FN - for `it_lmr_link`, must do immediate checking on the fast path since for InfiniBand, must return an immediate error when someone tries to create a "write-only" region.
    - § This requires that `it_lmr_link()` have a check for this in the fast path
    - § FN - JH was okay with this since it is a trivial check.
    - § FN - no check is needed for iWARP.
  - FN - calls a formal vote on this proposal
    - § HP votes yes, IBM votes yes.
    - § **Approved.**
- Continue discussion based on MM Detailed Requirements v0.98
  - FN - MM Detailed Requirements: Remote invalidation (MM-15.0)
    - § FN - MM-15.0.D2 - had wondered whether we could reuse the `it_post_send()` call to accomplish remote invalidation and felt could not since an additional parameter is needed (the RMR context).
    - § FN - MM-15.1.D1 - notes that unlinking of both an RMR and LMR supported since "RMR context" can represent either.
      - FN - notes his dismay with "RMR context" terminology
    - § FN - MM-15.1.D1.2.2 - text in square brackets following requirement justifies use of completion

event indicating to remote consumer that unlink has occurred.

§ MM-15.1.D1.2.2.1.1 - JH - wonders why you need to distinguish between LMR and RMR being unlinked?

- JH - Could just indicate that an RMR context has been unlinked
- FN - if no distinction, then it is a burden on Consumer to track what the RMR context refers to.
  - o FN - not sure if it is easier for Implementation or Consumer to figure out the correspondence.

§ JH - thinks Implementation can do this slightly more easily since RMR context is opaque to Consumer

§ MM-15.1.D1.2.2.1.2 - JH - recommends replacing "IT\_NO\_ADDR" with "IT\_NULL\_HANDLE" since it already appears in API and means the same thing.

- FN - what is difference between IT\_NULL\_HANDLE and NULL?
- JH - probably no difference for most implementations but it is intended to be used for handles

- o FN - continues with MM-16.X
- o FN - MM-L.X moved into MM-16.2
- o JH - back to MM-15.1

§ JH - recommends DTO completion event just supplying the `it_handle_t` corresponding to the RMR or LMR or `IT_NULL_HANDLE`

§ FN - how would consumer know what kind of handle this is?

- JH - could just call `it_get_handle_type()` on the handle.
- FN - is it always that case that casting a handle to `it_handle_t` and then passing it into `it_get_handle_type()` will retrieve the correct underlying type?

- o JH - thinks this is possible.
- o FN - if this is the case, then likes JH's proposal.

§ JH - consults "it\_get\_handle\_type" man page and thinks that it does support the required behavior.

§ FN - thinks that the handle is implied to be a pointer in all cases because of the implications

- o FN - agrees and accepts JH's recommendation to get rid of MM-15.1.D1.2.2.1.1 and only keep MM-15.1.D1.2.2.1.2 with the `IT_NULL_HANDLE` change above

§ AI - FN - make the above change to the requirements.

- o FN - continues with MM-16.2
  - § FN - has collected all remotely detected error conditions and how they manifest is described in MM-16.4
  - § FN - MM-16.2.3
    - InfiniBand does not allow remote invalidation of Wide RMRs - however, rationale is unclear as to why. Not a security issue, may just be a backwards compatibility issue.
  - § JH - is MM-16.2.5 different from MM-16.2.2?
    - FN - RMR context in 16.2.2 could be an underlying LMR
    - JH - are we going to distinguish these two errors to the Consumer of the IT-API?
      - o FN - intent of requirements is a comprehensive list of all the possible errors in the requirements and not sure if every one of these errors should manifest to the Consumer separately.
  - § FN - asks for opinions on whether the semantics should be exposed to Consumers?
    - JH - asks if for InfiniBand, do these errors get manifested synchronously or asynchronously?
    - FN - says he needs to cover 16.3 to help explain this.
      - o (FN - 16.3 describes how the errors manifest remotely and 16.4 describes how the errors manifest locally)
- o FN - continues with MM-16.3
  - § FN - buffer contents of the send may or may not be placed (depending on transport)
  - § JR - however, you still get a completion error, so the Consumer cannot assume anything about the buffer contents.
- o FN - continues with MM-16.4
  - § JH - least common denominator is the connection breaks
    - JH - some transports may give a little extra reporting, but that is transport-dependent.
    - JH - should we attempt anything beyond least common denominator?
  - § FN - notes that for IB, if the send with unlink fails remotely, you will get a successful completion for this operation, but the QP will then transition to the error state. Thus the next operation will fail.
  - § FN - notes that for iWARP, you will likely get an affiliated async error in this case
  - § JH - summarizes: At remote, recv WR will complete with an error. The QP will be torn down. At sender, on iWARP, get Aff Async Error, on IB, get successful completion (on both IB and iWARP, next send will then fail).

- JH - proposes only exposing receive failing on the remote side.
- JH - proposes NOT manifesting the Async Aff error on sender for iWARP
  - This way, IT API will be consistent for both transports - sender can only find out via the next operation.
  - FN - wonders if this buys us much? We are just tossing some additional debugging information.
  - FN - thinks maybe we could just define the Aff Async error and could use it for debugging on some transport.
    - § JR - recommends we not define the transport-dependent programming practice of observing this error on iWARP (i.e. do not document it for Consumer)
  - JH - thinks the Consumer will find Async Event is not particularly useful and thinks it should be dropped.
  - JR - it could still be manifested in an OS-dependent manner for debugging purposes, so okay with dropping the support for the Async Event.
  - FN - okay.
  - AI - FN - get rid of MM-16.4.D2.1 and MM-16.4.D2.2
- § AI - FN - document that a unlink operation that fails remotely will not manifest any behavior locally. Also add text to Implementor's guide that they need to absorb the Async Error and not manifest it via the IT API.

CM issues (<= 40 min)

- Discussion of third draft of CM man pages
- FN - had generic question on page called "Connection Management for iWARP". This page contains text on the TII and FN felt it not clear whether the TII should be described here. Recommends that this text should be moved to a chapter that applies to both IB and iWARP.
  - ◌ JR - agrees with the sentiment - wonders where to put it. Concerned about putting too much detail into overview sections.
  - ◌ JR - how about a new chapter, "Connection Management".
    - § FN - yes and it could have a section dealing with TII for all transports and a subsection specifically on iWARP (containing TDI).
    - § JH/FN - okay.
- FN - suggests postponing the discussion to next time.

If time permits:

- From physical addresses to bus addresses
  - See first discussion in Minutes for 23. Nov. 2004
    - § FN - discusses the address space issue.
    - § AI - JR - look more at 11/23/04 minutes to derive questions needing answering in mapping proposal.
  - Why a mapping/unmapping service?
    - § FN - in RNIC-PI, they have a map call that is used in connection with the STag of zero.
      - FN - wonders what the benefit of STag of zero is if you have to map stuff?
      - JH - some of the additional things required to use STag of zero are asking the O/S to pin memory. However, do not have to register on card.
      - <minute-taker - missed the discussion here>

Ending discussion

- FN - thinks we are closing in on IT API 2.0 - will send a few more CM comments to JR.

Meeting adjourns at 7-minutes over.