

ICSC ITWG Meeting Minutes
3/22/3005

3/22/05 ICSC ITWG Meeting Minutes

Taking minutes: HP (FW)

Y HP Jim Hamrick (JH) (joined at 10:17 PST)
Y HP Jay Rosser (JR)
Y HP Fred Worley (FW)
Y IBM Fredy Neeser (FN)
N NetApp Arkady Kanevksy (AK)
N Sun Matt Pearson (MP)

cascading ascii art attendance diagram. (if you have more than 1 minus visible, you are not eligible to vote.)

	hp	ibm	netapp	sun
	----	-----	-----	----
m-3	+	+	-	-
m-2	+	+	-	-
m-1	+	+	-	-
m-0	+	+	-	-

Next Meeting:

Tuesday **3/29/05**, 10am-12pm PST

ACTION summary:

PENDING AIs:

1. AI (JR): Do some experimentation to fix how `it_post_rdma_read_to_rmr` appears in the "legal blackline" comparison document.
2. AI (JR): Propose errata on 2.0 that having successfully completed socket conversion calling close on the SD has no effect on the RDMA connection
3. AI (JR): Request status update from Martin with expected delivery date of the document back to the WG
4. AI (FN): Update MM Detailed Requirements document to reflect v2.0 (e.g. Absolute vs. Relative Addressing)

NEW AIs:

5. AI (FN): Generate formal proposal for alternate name for PBLs
6. AI (FN): Investigate how RNICPI will address the keep alive option and whether they will add a transport dependent mechanism to deal with this or simply expose through LLP handle (the latter being implementation dependent)
7. AI (JR): Add text in `it_evd_events`:
 - "Indeterminate behavior is constrained to the evd concerned; overflow on a CQ associated with a particular EVD must not effect any other EVD or endpoint associated with other EVDs"
8. AI (FN): Add mapping service requirements to memory management detailed requirements

ICSC ITWG Meeting Minutes

3/22/3005

9. AI (FN): Change the terminology and associated type names in the memory management requirements from “Physical Buffer List” to “I/O Buffer List”
10. AI (JH): Request more justification from the requestor of 0 as a valid watermark; specifically how this feature would be used
11. AI (FN): Make a best effort attempt to document all resolved defects in Appendix D

Minutes:

Agenda bashing, approve minutes:

- Minutes to approve:
 - Minutes for 8. March. 2005, email "ICSC ITWG draft minutes for 3/8/05" sent on Feb 8 by Fred Worley.

Minutes approved

Additional agenda items:

- Backward compatibility (JR)
 - Need to discuss current proposals that would break source code compatibility between v2.0 and v2.1 (potentially requiring a rev of the major number)
- Subgroup formation for 2.1 effort
- Issue with formatting of the header file
- Consumer use of PBL types
 - Should there be a message from the ITWG to consumers regarding PBL types?
 - Discussion of naming conventions:
 - § Is PBL an appropriate term to use for I/O addresses?
 - § Is IOBL (IO Buffer List) a preferable (more neutral) term?
 - Could use “it_ioaddr_t”
 - § Could be either physical or virtual address depending on system characteristics
 - § Would not distinguish physical address from bus addresses – just use the IOBL terminology
 - § Objection raised to the term Physical Buffer List as these addresses may or may not actually be physical addresses – an alternate name could prevent confusion
- AI (FN): Generate formal proposal for alternate name for PBLs

Action item review:

- AI (JR): Lead further discussion to complete: Produce requirements for RDDP WG informational draft to support IOH functionality.
 - [Is this still a goal given the last calls in RDDP WG?
 - Have defined IOH such that interop should be possible
 - However, RDDP WG has completed their last calls, so the window of opportunity for synchronization between ITWG and RDDP WG is closed

ICSC ITWG Meeting Minutes
3/22/3005

o Completed

- AI (JR): do some experimentation to fix how `it_post_rdma_read_to_rmr` appears in the “legal blackline” comparison document.
 - o Awaiting response
 - o Pending
- AI (JR): Add new requirement: It is mapper’s responsibility to ensure that the mapping is w.r.t. the appropriate IA
 - o JR revised mapping service proposal on 3/10/05
 - o Completed
- AI (JR): Propose errata on 2.0 that having successfully completed socket conversion calling close on the SD has no effect on the RDMA connection
See also related email thread "Erratum on 2.0: "close()" shall have no effect on a converted socket"
 - o Implementations exist where the fd is removed on conversion; no fd exists on which to change options such as keep alive
 - o IHVs may or may not present an interface that allows manipulation of the underlying TCP connection after conversion of the connection to RMDA

[JH joins]

- o Not clear how this will be addressed in the RNICPI
 - o Have to query the LLP handle after doing the `modify_qp_to_rts`
 - o Not clear how the LLP handle will be defined
- Discussion of use models:
 - o User of conversion: User mode SDP implementation built on ITAPI
 - o Counter: kernel mode SDP implementation without kernel socket support could not make use of this feature
 - o Thought process for using sockets interface for controlling TCP connection aspects after conversion to RDMA mode:
 - § Sockets API already exposes a means to set keep alive; already using the sockets API prior to socket convert
 - § Not usable by kernel consumers where kernel sockets are not available
- Will RNICPI proposal be compatible with ITWG?
 - o LP handle must be queried and can be changed as a result of `modify_qp`
 - o AI (FN): Investigate how RNICPI will address the keep alive option and whether they will add a transport dependent mechanism to deal with this or simply expose through LLP handle (the latter being implementation dependent)
- Pending

AI (FN, All): Provide feedback on JR’s proposal for compatibility header file text and backward compatibility section

- FN has generated a list of typos

ICSC ITWG Meeting Minutes
3/22/3005

- Completed

AI (JR): Fix DTO defect (Errata global numbers 200, 201)

- JR provided updated version of the spec to ITWG participants directly
- Still being updated based on email exchange
- What happens if there is no overflow protection?
 - o Default overflow behavior in DTO is notification enabled
 - o For an IA that supports overflow detection, ...
 - o Only supported when IT_EVD_OVERFLOW_DETECTION flag is true
 - o Overflow of an EVD containing an event stream can result in undefined behavior if the IA does not support overflow detection – the consumer is responsible for verifying the overflow detection capabilities of the IA and ensuring that the overflow condition does not occur if detection is not available
 - o Should it be called out in the man page that notifications could be lost?
 - o iWARP spec states that overflow of one CQ must not effect other CQs
 - o What do we mean by indeterminate?
 - § Indeterminate in the effect on the EVD and the QPs that depend on the EVD but no effect on other EVDs/CQs/QPs
 - o AI (JR): Add text in it_evd_events:
 - § Indeterminate behavior is constrained to the evd concerned; overflow on a CQ associated with a particular EVD must not effect any other EVD or endpoint associated with other EVDs
- Is this a catastrophic error?
 - o How do you recover from an error with indeterminate results?
 - o Would expect application programmer to treat this as a catastrophic failure
 - o Difficulty in that there is no clean way to detect the catastrophic condition
 - o Agreement to resolve only with the clarification of text in this and new (above) AIs

- Completed

AI (JR by Friday, 3/11/05): Investigate and respond to propose to use source incompatible solution to address type issue (Errata global number 202)

- Completed

AI (JR): Request status update from Martin with expected delivery date of the document back to the WG

- JR received commitment from Martin to return document to WG on Monday, 3/21/05
- Document not yet received
- Pending

AI(JR): Make ORD 3.0 a High Priority HLR for v2.1

- Completed

ICSC ITWG Meeting Minutes 3/22/3005

AI (JR): Mark MM-9.11 as complete and remove from the priority list

- Completed

AI (FN): Update MM Detailed Requirements document to reflect v2.0 (e.g. Absolute vs. Relative Addressing)

- [Review] Talked in last meeting on terminology for PBLs
 - o Observed that “Physical Buffer List” may be a confusing term for something that could be physical memory addresses or I/O bus addresses
 - o FN proposes using the term “I/O Buffer List” (IOBL) instead
 - o **AGREED to proposal to change the terminology in the memory management requirements**
 - o New type would be `it_ioaddr_t` (an OS dependent type)
 - o **AI (FN): Add mapping service requirements to memory management detailed requirements**
 - o **AI (FN): Change the terminology and associated type names in the memory management requirements from “Physical Buffer List” to “I/O Buffer List”**
- What is the type of an offset?
 - o Length is of type `it_length_t`
 - o First impression is that offset and length should be the same type
 - o However, if you think about what transport underneath offers, it can offer 64 bits for offset and (32 bits - 1) for length
 - o However, in the context, this is not a transport function but a host-card interface function
 - o But if the length type is used and an implementation defines the length to be 32 bits, this could artificially limit the size of an offset
 - o Should therefore use `uint64_t` for offset type
 - o Reason for `it_length_t` was to allow implementation to define length is a size that is natural for it to manipulate (to avoid unnatural type casts)
- The length on the wire and the length in the triplets are not synonymous and need not be the same type
 - o The length limit on the wire is a limit that applies to the sum of all triplets in a given transfer request
- Does anything prevent an implementation from posting an LMR that is >32 bits in length?
 - o No
 - o However, `it_length_t` is used for the argument to `it_lmr_create`
 - o Note that `it_length_t` is defined as 32 bits on 32-bit platforms and 64 bits on 64-bit platforms
 - o **AGREED that the offset should use the same type as the length of the LMR (which is `it_length_t`)**
- Pending

Fixing errata against v2.0 draft:

- Erratum GN 200 and 201
 - o Discussed above

ICSC ITWG Meeting Minutes

3/22/3005

- Erratum GN 202: `it_lmr_triplet` modification for IT-API 2.0.
Remaining questions:
 - Which type should be used for `addr.rel` and for the IN argument of `it_make_rdma_addr_relative`?
 - Which type should be used for LMR / LMR Triplet lengths?
- Other errata against v2.0 draft that should be fixed before publication?
 - o Discussion of proposal for 0 as a valid watermark
 - § Today, 0 is a special watermark value meaning “disabled”
 - § Example: Associated with endpoint; first incoming send operation would cause an affiliated async event to be generated
 - § Today, you would need at least two messages for an affiliated async event to be generated
 - § Purpose is to provide protection against denial of service attacks
 - § Could this be used?
 - If you believe that responding to these events could put off starvation then perhaps responding to them sooner would be good
 - Not clear that proposal is implementable
 - § AI (JH): Request more justification from the requestor of 0 as a valid watermark; specifically how this feature would be used
 - o Erratum GN 210
 - § Resolved by JR
 - o Request for motivation to use shared receive queues
 - o Erratum GN 43
 - § Discussed and agreed to reject proposed resolution
 - § Rejected
 - o Erratum GN 204 – style question of “`it_ia_handle_t`” vs “IA handle”
 - § Use of the type allows hot-link to the corresponding man page
 - § Rejected
 - o Erratum GN 211 – some defects have been resolved but not noted as such in Appendix D
 - § Agree that it would be nice to document these
 - § AI (FN): Make a best effort attempt to document all resolved defects in Appendix D

Other issues:

- Register phys mem in the IB verbs treats the address as both input and output whereas in iWARP the address is an input only
 - o In the implementers guide, the use of `register_mem` is currently discouraged in favor of register phys mem
 - o For IB, register phys mem may return a different address than passed in as input
 - o Advice to implementer needs modification.

o Publication Process for IT-API v2.0

ICSC ITWG Meeting Minutes
3/22/3005

- Discussed during action item review
- See action items section for discussion notes

- o Callback discussion for IT-API 2.1. See email thread started by Jim Hamrick on 03/14/2005
 - § [AGREEMENT with interface proposed in mail by JH 3/14/05, Subject: "Callback discussion for IT-API 2.1"](#)
 - § In future meeting, should discuss what is allowed to be done from within a callback handler

- o Continue discussion on mapping service for IT-API 2.1. See JR's revised proposal from 03/11/2005

- o Prioritization for IT-API v2.1
 - § see email thread started by Jay Rosser on 03/07/2005
 - § Additional items:
 - o Multicast support for MPI implementations on top of IT-API
 - § interest in this from HPC community
 - § Assume for UD
 - § better understanding of requirements needed to address the concern

- o Next steps
 - o Priorities for coming meetings
 - New memory management requirements [FN + JR + FW]
 - Callbacks [JH]
 - Support for DTOs [JH]
 - IB atomics [JR]
 - UD Multicast

- o Any other business
 - § tab alignment in header files
 - in 1.0, header files were generated with scripts based on man page documents
 - in 2.0, a different mechanism needed to be used (using different scripts and hand massage)
 - chose not to run through cbeautify or equiv as that would change the look and feel from v1.0
 - v2.1 / v2.0 source code compatibility
 - Currently, we define a minor version number update to be one that is source code compatible with the previous (same major number) version(s)
 - Several options may exist to allow source code compatibility
 - additional review and discussion will be required to resolve
 - § Need for additional modes
 - In MM requirements, some optional features for opening an RNIC in certain combined modes; not just page or block modes, including variable length page lists
 - RNICPI WG is not particularly eager to review this feature set

ICSC ITWG Meeting Minutes
3/22/3005

- They support just a variable length list
 - o Even if an RNIC vendor does not want to implement variable length lists, they would be forced to in order to support the RNICPI
 - o Solution lacks some flexibility
- Should there be a more formal statement to the RNICPI WG

Meeting adjourned, 12:10pm PST