

ICSC ITWG Meeting Minutes
4/28/2005

4/5/05 ICSC ITWG Meeting Minutes

Taking minutes: HP (FW)

Y HP Jim Hamrick (JH) (joined at 10:11 PST)
Y HP Jay Rosser (JR)
Y HP Fred Worley (FW)
Y IBM Fredy Neeser (FN)
N NetApp
N Sun

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 **4/19/05**, 10am-12pm PDT (**No Meeting 4/12/05**)

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

NEW AIs:

3. AI (JH): Propose initial list of actions that a callback function should not take (assuming an interrupt or signal handler context).
4. AI (JR): Provide a use case for a kernel mode consumer registering and using user-mode buffers when not in process context
5. AI (FN, JR): Revisit `it_lmr_create` to determine if an address space qualifier is necessary
6. AI (FN, JR): Registration and IO buffer lists – determine if IOBL and underlying RNICPI PBLs can be efficiently translated

- Agenda bashing, approve minutes
 - o Email from Fred Worley, subject "ICSC ITWG draft minutes for 3/22/05", sent 3/25/05 1:58PM PT
 - o **Approved**

ICSC ITWG Meeting Minutes
4/28/2005

- Action item review
 - o AI (JR): Do some experimentation to fix how `it_post_rdma_read_to_rmr` appears in the “legal blackline” comparison document.
 - § Pending
 - o AI (JR): Request status update from Martin with expected delivery date of the document back to the WG
 - § Martin states that document will be complete by 4/8/2005
 - § Pending
 - o AI (JH): Request more justification from the requestor of 0 as a valid watermark; specifically how this feature would be used
 - § JH communicated with requestor of the change
 - § Requestor is looking for a way to generate an event for efficient notification of pending work requests (for receipt of first message)
 - Want event on empty receive queue when a send arrives
 - § If value is 0, it is not possible to have a pending work request without having an event – would immediately get an async event
 - § If you are trying to catch all such events without checking in the data path then this feature would be nice
 - § JH believes this is an RNICPI issue, not an ITAPI issue
 - § No longer requesting for ITAPI
 - § However:
 - § In iWARP verbs, if QP RQ limit is set to 0, when the QP gets any send message, even if pending and not completely finished, then an event will be generated (async event associated with that QP)
 - § Currently, verbs say this only applies to SRQs
 - § Would not be difficult (in theory) to use this mech for other queues (not just SRQs)
 - There is also a low water mark associated with SRQs
 - High water mark keeps track of pending work requests – could same mech be used even if QP is not associated with shared RQ?
 - o iWARP verbs spec does not require that
 - o However, high prob. that the mech would work even if queues are not shared; may not be a lot of work for HW vendors to extend this mech to non-shared RQs
 - § Errata rejected
 - § Completed
 - o AI (FN): Update MM Detailed Requirements document to reflect v2.0 (e.g. Absolute vs. Relative Addressing)
 - § See email from FN, 4/5/05 [Add reference]
 - § Completed
 - o AI (FN): Draft improved text for Application Usage section of `it_socket_convert` regarding use of `sd` after Socket Conversion.

ICSC ITWG Meeting Minutes
4/28/2005

- § See Email thread started by Fredy Neeser, subject "TCP keepalive option; calling close during/after socket conversion", sent 3/29/05 8:51AM PT
- § See Email from Fredy Neeser, subject "AI: Improved guidance to consumers on keepalive issue", sent 3/30/05 5:14AM PT
- § No support committed from RNICPI
- § FN drafted additional text for it_socket_convert man page to address
- § New text provided to OpenGroup (Cathy) for inclusion in doc
- § **Completed**

- IT-API 2.0 Errata:
 - o Email from Fredy Neeser, subject "iWARP QP Bind Enable attribute- missing note in v2.0 Implementer's Guide", sent 4/5/05 7:55AM PT
 - o iWARP has an additional QP attribute that IB did not have (bind enable)
 - o No compelling reason found to expose this attribute at ITAPI level
 - o Seems appropriate to keep behavior consistent with IB by not exposing the attribute
 - o Propose to resolve by adding additional text to implementer's guide at the end of the section on it_ep_rc_create:
 - § "For the iWARP transport, the QP attribute for MW bind operations shall be set to "enabled"".
 - o **Completed**

- Publication process for IT-API 2.0
 - o Version vs Issue
 - § Not discussed
 - § Request from OpenGroup to use terminology "Issue 2, Version 0" as opposed to "Version 2.0 with a major number of 2 and a minor number of 0"
 - o Delivery update
 - § Martin states that document will be complete by 4/8/2005

- Work towards IT-API v2.1
 - o Callback discussion for IT-API 2.1. See email thread started by Jim Hamrick on 03/15/2005
 - § What is allowed to be done from within a callback function?
 - § Depending on scope, callback could be invoked directly from the ISR
 - § Function may be limited when executing on the ISR (e.g. may not be able to allocate memory)
 - § What guidance is appropriate to give to consumers/implementers?

ICSC ITWG Meeting Minutes

4/28/2005

- Idea is to give rules to the consumers as to what functions they can invoke from a callback routine
- § Is memory allocated by the consumer accessible during the callback?
 - For some architectures, memory context may not be accessible from the ISR
- § Should the ITAPI provide a definition of what a callback can do?
Some examples (not final):
 - Callback function may not sleep
 - Callback function may not invoke blocking calls
 - Callback function may not invoke creation interfaces
- § Looking for superset of reasonable restrictions across OSs
- § May also want to look at they types of things a consumer may want to do from a callback function
- § Could create two different classes of callback – e.g. classes of thread safety; could also divide callback implementations into interrupt context safe and not interrupt context safe
 - Richer callback environment but potentially lower performance
 - Would put bounds on what implementers would be required to implement
 - State that there are 3 classes of thread safety – implementation has to claim one
 - Not discovered programmatically; consumer must know that the desired thread-safe model exists within the implementation
 - E.g. thread safe ITAPI library and non-thread-safe ITAPI library would have different names – application would link to the one that it requires
- § Actions:
 - **AI (JH): Propose initial list of actions that a callback function should not take (assuming an interrupt or signal handler context).**
- § Alternatives:
 - Document types of constraints for callbacks:
 - Interrupt context (most restricted)
 - Signal handlers (somewhat restricted)
 - Threads (unrestricted)
- § Proposals:
 - Restrict consumers to one model – assume most restrictive
 - Superset of ISR restrictions and signal handler consumers
 - If consumers find callback restrictions too restrictive, consumer can register a (restricted) callback that invokes an (unrestricted) thread in order to defeat the restrictions
- o Email exchange with Caitlin Bestler (CB) RE providing events to callbacks
 - § CB and JH agree that there should be only one invocation of a callback at a time (callbacks do not have to guard against being invoked on two different processors simultaneously)

ICSC ITWG Meeting Minutes
4/28/2005

- § In the DAT collaborative callback model, the callback is furnished with an event
 - Debated previously in ITWG
 - Determined that callbacks should be furnished with notification but not an event
 - Can use ITWG model to implement DAT model
 - ITWG model doesn't break the other use model but may make an implementation of it somewhat less efficient
- o New MM Requirements:
 - § Mapping service for IT-API v2.1. See JR's revised proposal from 03/11/2005
 - § See mail from FN, Subject AI: Update MM Detailed Requirements, 04/05/2005
 - § Changed table at beginning of the document for optional features
 - Based on review of IB 1.2 specification, consumer is not guaranteed support of either Base Memory Management (BMM) support (verbs) or the individual features of BMM (e.g. direct LRM handle / reserve LKEY).
 - Having BMM just means that there is verbs support for querying – e.g. is the reserved LKEY available?
 - o Query may answer there is no support for reserve LKEY
 - o The verbs will look different if you have base memory management
 - o For an OS integrator, this may create some challenges
 - Is there a query interface to determine if you have BMM?
 - o No; just documentation
 - Flags like `direct_lmr_handle_support`; `lrm_link_support` – will be false if there is no base memory management support or if verbs are present and flag is false
 - § Fixed documentation to be consistent with version 2
 - Def of type `it_iobl_t` (see p6 of MM Detailed Requirements, April 5, 2005)
 - Provided clear text description of data structures, IO Buffer list
 - o Union of constant element length list and variable element length list
 - Cleaned up names a bit
 - o `it_map_mem` changed to `it_mem_map`
 - use model in 22.1.1
 - o non-priv consumer would use a system call to pin a buffer
 - o on behalf of consumer, kernel consumer calls `it_mem_map`
 - § may need address space identifier – may not be in user's context
 - o consumer constructs IO buffer list from the mapping

ICSC ITWG Meeting Minutes

4/28/2005

- kernel consumer uses IO buffer list to:
 - § link an LMR
 - or
 - § create a new LMR
 - non-priv consumer references the LMR and DTOs
 - § note that PZ must be shared between user process and kernel consumer
 - AI (JR): Provide a use case for a kernel mode consumer registering and using user-mode buffers when not in process context
 - use model in 22.1.2
 - discussion of 22.1.2.1
 - § S/G lists are short in iWARP (minimum is 4 elements, although implementation dependent)
 - § For IOBL, limit could be very large
 - § Could have implications on how devices are constructed
 - When you deal with user mode addresses, direct LMR handle is not of any help
 - § User mode process can not use direct LMR handle for security reasons
 - § Only usable to manipulate kernel buffers
 - § Kernel buffers may be more likely to be physically contiguous than user memory (although this is OS and system architecture dependent)
- § Additional discussion:
- Now have address space qualifier for mem map call
 - Could this also be useful for LMR create?
 - § Are there cases where multiple address spaces are valid simultaneously?
 - E.g. physical addresses, user process context, kernel context
 - § How do you tell LMR create which address space this address range belongs to?
 - § Potentially useful to allow a kernel process to register user memory on behalf of a user-space consumer
 - AI (FN, JR): Revisit `it_lmr_create` to determine if an address space qualifier is necessary
 - Registration and IO buffer lists
 - AI (FN, JR): Registration and IO buffer lists – determine if IOBL and underlying RNICPI PBLs can be efficiently translated
- § What MM requirements are ready for vote?
- Direct LMR

ICSC ITWG Meeting Minutes
4/28/2005

- Local registration of resources (close to ready for vote)
- Local regions
- Local fence semantics
- Endpoint attributes
- § What MM requirements still needs review?
 - Fast register LRM handle needs review
 - Addressing within LRMs and LMWs
 - How underlying memory in an LMR is defined
 - Depends on action item for RNICPI definition
 - Remote invalidation
 - Query local LMR (19.1)
 - Mapping service
- § Completed MM requirements
 - Narrow window binding
 - LMR creation and RDMA reads
- § Next steps:
 - Complete detailed requirements before completing the man pages
- Support for lists of Work Requests (WR seems to be a preferred term now over DTO, as it includes MM operations)
 - § No update
- IB Atomics
 - § No update
- UD Multicast support (e.g. for MPI implementation on top of IT-API)
 - § No update
- Any other business
 - Discussion in RNICPI WG of adding UDP support
 - § Data structures are done such that STAG and LKEY/RKEY concepts could be used
 - § Could do IB's UD service and IP and reliable datagram somehow

Meeting adjourned 11:50am PDT