

ICSC ITWG Meeting Minutes
6/7/2005

6/7/05 ICSC ITWG Meeting Minutes

Taking minutes: HP (FW)

Y HP Jim Hamrick (JH)
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	+	+	-	-

ACTION summary:

PENDING AIs:

1. AI (JR): Review status of MM Detailed Requirements and propose sections ready for voting.

NEW AIs:

2. AI (Memory Management group): Add text to implementer's guide clarifying proper behavior if VA on input and output differ
3. AI (JH): Post errata with proposed text as resolution
4. AI (JR): Determine the text in ITAPI 2.0 that needs to be clarified that defines relationship between # of outstanding work requests and the size of the work queue
5. AI (JH): Create formal proposal for preferred method of DTO list requirements and present to RNICPI
6. AI (Memory Management subgroup): Replace wording on line 4251 that currently reads "refers to the same physical memory pages as the new LMR" with "the identical set of physical memory pages as the new LMR" and add to the end of the sentence "and has the same length".
7. AI (Memory management subgroup): Research current wording in `it_lmr_create`
8. AI (JR): Research previous discussions of physical to bus address translation responsibility and report to WG

- Agenda bashing, approve minutes
 - o Email from Fred Worley, subject "ICSC ITWG minutes for 5/24/05", sent 5/24/05 9:55AM PT
 - o [Minutes Approved](#)
- Action item review

ICSC ITWG Meeting Minutes

6/7/2005

- o AI (FN): Report back to ITWG on semantics of IOVA output in IB's Register Physical Memory verb.
 - § See email thread started by Fredy Neeser, subject "Semantics of IOVA in IB's Register Physical Memory Region verb", sent 5/31/05 8:00AM PT
 - § Address naming question: Why is "IO virtual address" used instead of "virtual address"
 - A: may not correspond to a valid virtual address in a process
 - May be a mapping that the consumer performs by itself
 - Does have significance on the wire
 - Has local significance and, if absolute addressing is used, remote significance
 - FN: Suggestion to call all such addresses "virtual addresses"
 - There is no semantic difference on input – may be a semantic difference on output
 - o iWARP doesn't include a parameter for output; input only
 - § What is significance of IOVA on output?
 - As long as returned IOVA matches requested IOVA then there is no difference in the paradigm for IB and iWARP
 - Need to be careful only if there is a difference
 - One source of difference – lack of byte granularity for a memory region
 - Example: If there are no virtual addresses (purely physically addressed system)
 - o Case may not be relevant
 - o Could use 0-based virtual addressing for this case
 - Ex: User space memory region, enter system call and access layer does pinning and is aware of the buffer list; VA must be VA that the users expects; user will use that VA to post requests on that address range;
 - o JR: 1.1 and earlier verbs give the latitude not to respect that...
 - o FN: then this usage of the verbs might fail
 - o JR: expect that register Phys Mem is probably for use by kernel clients
 - o FN: IB 1.2 sez returned address should match the requested address
 - o Suggestion: if returned address deviates from requested address ...
 - § Should check that lower address equals the lower bound that you can get from the query memory region verb (?)
 - o After creating the memory region, check to make sure that, if addresses are different, that the only change is to make the address properly aligned (e.g. page aligned)

ICSC ITWG Meeting Minutes

6/7/2005

- Otherwise, fail the LMR Create
- Suggest adding text to the implementer's guide
- AI (Memory Management group): Add text to implementer's guide clarifying proper behavior if VA on input and output differ

§ Closed

- AI (JH): Add text to the usage section of all posting calls to clarify appropriate usage to avoid WQ and CQ overflow

§ See email from Jim Hamrick, subject "AI: Proposed text for avoiding WQ and CQ overflow", sent 5/31/05 2:34PM PT

§ New text approved by WG

§ Closed

§ IA (JH): Post errata with proposed text as resolution

- AI (JR): Determine how DTO 6 high level requirement vote was resolved

§ See email from Jay Rosser, subject "AI - Determine how DTO 6 high level requirement vote was resolved", sent 6/1/05 5:20PM PT

§ Discussion of the vote on the semantic came up because there is a scenario where a strict definition would help...

- Strict definition appeared useful in posting lists of work requests
- If we get enhancement requested to RNICPI (coalesce flag passed through) then error checking is done immediately and the issue of a sloppy semantic is resolved (w.r.t. posting lists of work requests).
- Discussion was around how to guide consumer to determine the # of outstanding work requests

§ Closed

§ AI (JR): Determine the text in ITAPI 2.0 that needs to be clarified that defines relationship between # of outstanding work requests and the size of the work queue

- AI (JH): Provide additional description of what is meant in list of work requests proposal by # of work requests processed; include usage example

§ See email thread started by Jim Hamrick, subject "AI: Description of number of WRs processed for DTO list requirements", sent 5/31/05 3:49PM PT

§ Proposal on the table to ask the RNICPI WG for a "more friendly" method for posting a list of work requests that is in better alignment with previously published versions of the ITAPI

§ Proposal would simplify error semantics

§ Do not expect performance disadvantage to proposal

§ Putting work request into a struct and then extracting it again is an overhead

- Proposed solution is at least as efficient

§ Closed

§ AI (JH): Create formal proposal for preferred method of DTO list requirements and present to RNICPI

ICSC ITWG Meeting Minutes
6/7/2005

- o AI (FN): Propose resolution for ambiguity of shared state of LMRs for LMRs that are a) created in a shared state and b) transition into a shared state
 - § Proposing new flag to define non-sharable LMR
 - § For sharable LMR, return the actual state (whether it is truly shared or not)
 - § Implementation of LMR create, when it tries to share an new LMR, it will go and search for a matching LMR – but it will only search among those LMRs which do not have the non-sharable flag set
 - § Allows a consumer to have a non-sharable LMR which has the desired semantics – e.g. can be sure you can unlink the LMR later on (can not be inadvertently converted to a shared LMR later on)
 - Agreement that this is valuable
 - § What does it mean for 2 regions to match?
 - May not be possible to determine that 2 regions are “matching”
 - Text in ITAPI 2.0 is not precise – does not specifically state that the set of pages have to exactly match
 - o Difficult to imagine what else it could mean, however:
 - o Could meaningfully refer to a subset of the initial set of pages
 - o Not clear that there is a compelling reason to support a definition of “matching” that includes subsets
 - § Proposal: Add wording that it may not be possible (or efficient) for an implementation to identify two matching reasons
 - Implementation might react to a sharing request by not sharing if it is too difficult to determine that the regions requested to be shared match
 - § If consumer insists on sharing an LMR, they should be aware that shared LMR is not necessarily shared – may have to create another LMR
 - § Environment where you can not determine (can’t find) a match
 - Initially created LMR (without a match – non-shared state)
 - Then create a shared LMR with the same ID but the imp can’t find a match (can’t support finding a match)
 - What state are those LMRs in?
 - o Would be considered unshared from both an ITAPI perspective and a device perspective
 - § Could be unlinked
 - o Note that implementation could choose to track sufficient state such that the implementation would not need to query the device to determine the shared state of an LMR
 - § This may have unacceptable overhead
 - o Figuring out the match may be expensive in the absence of a shared ID

ICSC ITWG Meeting Minutes

6/7/2005

- If we expose the underlying state of the LMR and show that it is not shared even if the user has requested it to be shared) then consumer would at least have the ability to validate
- Agreed that exposing the non-sharable flag (as proposed in email by FN, subject “Erratum GN217=IBM204: Shared state of an LMR cannot be reliably inferred”, 5/31/05 7:27AM PT) is preferred over the original proposal, which could create overhead in the implementation

§ Suggested new wording:

- AI (Memory Management subgroup): Replace wording on line 4251 that currently reads “refers to the same physical memory pages as the new LMR” with “the identical set of physical memory pages as the new LMR” and add to the end of the sentence “and has the same length”.

§ Example:

- User process registers a memory region – passed to register mem verb; actual pinning of and at least mapping of the address range is performed by the verb;
- Second process registers the same region
- Even tho it is the same region it is not clear that it would have the same mapping of bus addresses
 - Same physical memory (shared memory between two processes)
 - RNIC may have two different IOVAs for this region
 - It is conceivable for the same physical memory to have two different IO address mappings
- What happens if the imp can’t determine if the regions are shared?
 - May be different reasons for the LMR not to be shared...
 - § Could be not-shared because there is no primary to share with
 - § Exists already a primary memory region that matches but the imp is unable to determine that they match
 - § Consumer needs to be aware that the existence of a matching memory region does not guarantee that it can be found
 - § Currently, ITAPI does guarantee that a match can be found (text in lmr_create)

§ AI (Memory management subgroup): Research current wording in `it_lmr_create`

§ Does iWARP guarantee that a match can be found?

- Don’t you have to provide the handle of an existing region in order to share?
 - In this case, there is no search process

ICSC ITWG Meeting Minutes

6/7/2005

- Only introduced at the verbs level because we don't use the handle
- Just have a global ID which allows sharing between processes that don't have the same PD
- There is merit in having the search process

§ Why can the search fail?

- Search could fail if the shared IDs do not match
- Text refers to shared ID case; should be fine
- "if unrelated callers supply the same value for Shared ID, matches will be found (although it may take longer on some implementations)"

§ If you are trying to create a snared LRM and you submit a bogus shared ID

§ Is there a use model for sharing without specifying the shared ID?

- Issue is unrelated applications
- They have no idea about another LMR that does not exist
- Does it still make sense for them to share a memory region?

§ Two scenarios

- Cooperating processes that have shared memory somewhere
 - If cooperating, likely they could use the same Shared ID
 - No reason for cooperating processes to use different Shared IDs
- User mode buffer that is shared by two different kernel consumers
 - The kernel consumers may not be cooperating

§ Does the current text cover this?

§ Summary:

- If you use a unique Shared ID, you are guaranteed to share the mapping
- If you use a common Shared ID for multiple LMRs, then the Implementation will make a best-effort to find the matching address range for a specific pair (set) of LMRs
 - There may be a performance impact, though.

○

§ Still undergoing internal review

§ Majority is complete and ready to vote

§ As kernel consumer, if I want to register phys mem, do I have to first map to convert physical addresses to bus addresses and create a list of bus addresses that is then passed to register phys memory verb or does the verb expect a list of physical addresses and it will do the mapping?

§ Does PBL contain physical addresses or bus addresses?

§ Currently the mapping service describes the case of going from a VA to PBL that contains bus addresses

ICSC ITWG Meeting Minutes

6/7/2005

- § Do we need different types of PBLs? In one case they contain already-mapped bus addresses and in the other they contain physical addresses
- § Pending
- § AI (JR): Research previous discussions of physical to bus address translation responsibility and report to WG
- o AI (JR): Forward FN's email, subject "Semantics of IOVA in IB's Register Physical Memory Region verb", to HP's IBTA folks for comment.
 - § Example that could justify this use model:
 - § IOVA being an output that is different provided in the interface to support potential, exotic architectures
 - May not be relevant today
 - § Proposal (as prev. discussed): After LMR create creates a memory region using register_physcail verb that it go and check the actual IOVA that was allocated; make sure that the only adjustment is that it be adjusted to a page boundary – otherwise reject the registration
 - § Closed
- o AI (JR): Generate an Erratum for changing the names of it_lmr_sync_rdma_read and it_lmr_sync_rdma_write to something more intuitive. Names proposed are it_lmr_flush_to_mem (performs writeback of dirty cache lines) and it_lmr_refresh_from_mem (performs invalidation of cache lines).
 - § See email from Jay Rosser, subject "Erratum GN218 - rename it_lmr_sync_rdma_read and it_lmr_sync_rdma_write for clarify", sent 6/1/05 4:54PM PT
 - § Closed
- o AI (All): See if there is any value in S-RQ support for UD (see also Erratum GN216 = IBM203 thread).
 - § See also response from Jay Rosser to IBM203 thread sent 6/2/05 5:10PM PT
 - § Some use cases identified for S-RQ for UD
 - § None of them currently considered sufficiently compelling
 - § Agreed that support for S-RQ for UD can remain a Phase 3 High Level Requirement
 - § However, there is a source code compatibility issue:
 - § SQ handle is not in the generic portion of the endpoint attribute; in the RC-only
 - If we wished to added S-RQ for UD later we would need to break source code compatibility in the future
 - § AI (All): Consider moving S-RQ handle from RC-only attributes to generic attributes
 - § Closed
- Callbacks
 - o Email from Jim Hamrick, subject "Call for vote on detailed requirements for callbacks", sent 5/31/05 9:48AM PT
 - o No discussion; no disagreement

ICSC ITWG Meeting Minutes
6/7/2005

- General IT-API 2.1 discussion
- Bugzilla
 - o Seems to be similar to what we have today (for content)
 - o No strong reason not to use Bugzilla
 - o Provides a well-defined process to submit an errata
 - o Need to determine how to capture current global IDs and company IDs for existing and new errata
 - o [Agreed to switch to Bugzilla as errata reporting and tracking mechanism](#)
- Any other business
 - o Comments on Shared RQ for UD service type
 - § Current summary: No compelling use model for Shared RQ for UD service type

Meeting adjourned, 9:05am PDT