

ITWG Meeting Minutes – 11/09/2004

List of those present at the meeting

| Company | Name |
|-------------------|--------------------------|
| Broadcomm | Uri E. (UE) |
| HP | Jim Hamrick (JH) |
| | Jay Rosser (JR) |
| | Fred Worley (FW) |
| IBM | Fredy Neeser (FN) |
| Network Appliance | Arkady Kanevsky (AK) |
| Sun | Matthew Pearson (MP) |

(People whose names are in **bold** letters were present)

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

```
      hp      ibm      netapp      sun      broadcomm
-----
m-3    +        +        -        -        - <- enforcement starts
m-2    +        +        -        -        -
m-1    +        +        -        -        -
m-0    +        +        -        -        -
```

Next Meeting

Tuesday 11/16/04

Action items appear in **red** embedded in the document below.

Votable items appear in **blue** embedded in the document below.

Discussion

Minutes to approve

FN: Minutes to approve. Fred hasn't sent draft minutes yet for 10/26 meeting. Jay, have you seen any minutes?

JR: No.

FN: Will wait for minutes to be posted to reflector.

AI Review

FN: Jay had AI to emit proposal for how to turn off IOH header for compatibility.

JR: Proposal sent to the reflector on 11/3/04 covering this topic. Proposal covers another AI assigned to me as well. Proposal is in two parts. One part is to change semantics of how the Transport Independent Interface will work by adding new fields in IOH, the other part is to achieve DAT compatibility by allowing transmission/reception of the IOH to be suppressed. (Jay proceeds to describe the proposal.)

JR: Got a comment on proposal from within HP regarding the description of how the downgrade bit conflicts with MPA spec. The conflict (as perceived by some within HP) is that the MPA spec won't allow the downgrade capability of IOH to override the Rx marker setting of an IETF RNIC that supports downgrade capability but prefers to not receive markers. The comment was that there is no need to include a description of the conflict if we avoid the conflict and make it a requirement that an IETF RNIC must advertise that it wants Rx markers if it has the capability to handle them.

FN: If one side desires no Rx markers (indicated in MPA REQ), side that sends reply would then specify that it wants to downgrade, results in ignoring request to turn off markers?

JR: Sort of. Model in this case is that active side in MPA REQ would say it doesn't want markers, passive side is RDMAC that can't do that, possibilities are reject or accept with understanding that active side would downgrade.

FN: If active side advertises it can downgrade, but says it prefers not to receive markers, then passive side accept with downgrade can compel active side to receive markers.

JR: Only way to fix this on active side is to enable markers.

JR: If we mandate that a compatible IETF RNIC must ask for markers, wouldn't have a problem.

FN: Another comment on this proposal. It might be better to call the new reject code `IT_IWARP_VERSION_MISMATCH`, so that readers don't confuse this with an IT-API version mismatch.

JR: Ok. **AI: Jay to change reject code name.**

FN: Another comment: T field corresponds to type of RNIC. (Protocol version of DDP and RDMAP.) Think the version fields for these protocol actually have two bits. Better to define T as a 2-bit field? Would allow us to reuse reject mechanism in future version if we get protocol version number beyond 1.

JR: We have a number of different protocol layers that may or may not remain consistent. Wonder if even two bits are enough.

JH: Why not throw bits at the problem? Make T field have as many bits as required to cover all possible combinations of protocol versions.

JR: Ok. **AI: Jay to figure out how many bits will be needed, and size the T field appropriately.**

FN: Part 1 of proposal seems otherwise fine to me.

JR: (Proceeds to discuss part 2 of proposal, which is allowing IOH header to be turned off.) Major side effect of proposal is that a conn_qual on passive side will be constrained to either support or not support the IOH, and so same conn_qual can't be used to listen for incoming traffic from DAT and IT-API active sides.

FN: it_listen_create has flags argument, but it is not used in the manner that you propose. Think it might be better to change it_listen_create to take a new transport-dependent parameter instead of adding a flag bit.

JR: Don't see problem with just using flag bit instead. What do you find objectionable?

FN: The asymmetry between parameter use on active side versus flag bit use on passive side.

JH: We could use a flag bit passed to it_ep_connect to remove the asymmetry.

FN: Why do we have flag on it_listen_create, anyway?

JR: To specify whether implementation should choose an ephemeral port or not.

(Tangential discussion about which conn_qual are used where ensues. Tangent eventually ends...)

JR: So, would using a bit rather than a parameter on active side be okay?

FN: Yes. We do, however, need to make sure that the fact that not all of the flag bits to it_ep_connect are orthogonal is documented. (e.g. You can't turn on both two-way and three-way connection establishment.)

FN: What happens with it_ep_accept? If IOH header is present, it_ep_accept will look at EP attributes and use them to fill in IOH in MPA REP, right? If IOH header isn't present, then EP attributes won't be communicated to peer, right?

JR: Right. EP attributes for IRD/ORD will be ignored if IOH is suppressed. **AI: Jay to specify that IRD/ORD from the EP attributes is only used if IOH Header is used.**

FN: Same problem happens for connect. If no IOH, can't provide IRD/ORD. Need to specify this.

FN: Technically, turning off IOH and changing content of IOH violates our voted-upon requirements. Need to vote to allow ourselves to make these changes. Call for Vote in one week on the following proposition: [The IT-API phase 2 implementation shall support suppression of the IOH on both the active and passive sides. The IOH shall support additional "type" and "downgrade capability" fields. There shall also be a defined IT_IWARP_VERSION_MISMATCH reject code for use in the IOH.](#)

FN: Next issue. [My AI to get an update from the RNIC-PI workgroup is still pending.](#)

FN: Next issue. Jay to get info on sensible use of Remote Access Flag.

JR: Did some digging, RAF allows two classes of Stags to be created, analogous to IB L_Key and R_Key. Always setting RAF functionally equivalent to creating an IB Memory Region requesting both an L_Key and R_Key. Access rights are orthogonal to whether keys exist, however. Only reason RAF seems to exist is to allow some implementations to optimize resource usage. Happy to accept your proposal to always set the RAF. Action item is closed.

FN: Next issue. Follow-up on AI regarding RTR state, and using RDMA Writes and Reads to cause a "First FPDU Received" event to be generated. Zero-length RDMA Write/Read will not result in checking of Stag. Means that we can use zero-length RDMA Write to send first FPDU from active side to get passive side into connected state.

JR: But we still don't have async event on passive side to indicate transition from RTR to RTS, right?

FN: Right, but RNIC-PI WG (at least John Carrier, anyway) seems to support adding this functionality.

JR: Need to know nature of all devices in cluster and adjust your ULP to meet their capabilities if RTR becomes an optional feature, though.

FN: Only reach RTS state after you've sent MPA REP, had that ACK'd, and gotten an FPDU.

JR: Major concern with all of this is that we don't orphan existing RNICs.

JH: Corner case you need to worry about if you use zero-length RDMA Write as a mechanism to cause FPDU received event to be generated. If you use a zero-length RDMA write to cause transition, need to make sure implementation has an extra entry available in Consumer's Work Queue to avoid potential Work Queue overflow.

FN: Need RNIC-PI WG to decide whether or not it will support new event.

JR: We can't get stuck waiting for decision from the RNIC-PI WG for too long.

CM man page review

FN: Next topic. Draft CM man pages. Still haven't read your latest draft.

JR: That's probably okay... looks like they are going to be changing again anyway.

FN: Uri had some valid comments regarding role of TCP ACK vis a vis your diagrams in the CM man pages. Said that it doesn't need to be a separate message from first RDMA Send; could be piggybacked. Should mention this in text next to figures showing TCP ACK.

FN: **AI: Jay to add text to diagram to say that TCP ACK could be piggybacked on the first FPDU.**

FN: Uri said IRD/ORD negotiation isn't covered. I explained in email how it was done. Think we should discuss this somewhere in the man pages.

FN: **AI: Jay to 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.**

Memory management issues

FN: Next topic, memory management issues. Thanks Jay for your comments, have incorporated them into man pages.

FN: I posted an email to the reflector about current IT-API semantic that LMR/RMR can only be linked/unlinked in connected state. Seems like a bad model. Don't want to impose this on all transports just because this is the way it works with IB.

(Discussion ensues about whether RDMAC verbs would allow bind/unbind/fast register/invalidate to be done in anything other than the RTS state. Jim expresses skepticism that there is anything in the RDMAC verbs that explicitly requires processing of bind/unbind/fast register/invalidate in the Init state. Fredy isn't sure. Group feels the verbs spec may be ambiguous on this point. Discussion then shifts to whether or not the programming model for dealing with unbind/invalidate when the QP used to create the binding moves out of RTS is broken. Jim thinks a mechanism that might work is for the QP to be connected to another of the Consumer's QPs, and then the unbind/invalidate WRs could then be posted/processed.)

FN: **AI: Fredy to ask RNIC-PI WG how to unbind a narrow RMR after a QP has been disconnected.**

FN: AI: Fredy to find out if Bind, Faster Register, Unbind, and Invalidate are supported in anything other than the RTS state.

FN: Regarding the issue I posted to the reflector about opening an interface adapter in priv. mode. Agree with feedback on reflector that we don't need this.

FN: Newx issue, handling of remotely-detected errors. Caitlin doesn't think on-the-fly conversion of terminate header to work completion would be worth it. But iWARP has a Work Completion that allows this to be done. Faced with possibility that some vendors will manifest errors in the Work Completion, and some will manifest those same errors only in the terminate header.

JR: Our current high level requirements say we'll interpret received terminate as either completion or async error.

FN: Means we can end up manifesting the same error in one of two different places. Not desirable.

JH: We could coerce all errors to appear as async errors, even if they were reported by the underlying RNIC as completion errors. Pro is that we only manifest these errors in one place. Cons are that it's not the same place as for the IB transport, and that reporting via async error is probably an inferior mechanism to reporting via Work Completion.

(Discussion ensues. Three different options are identified for how to manifest remotely-detected errors: manifest the error where the RNIC manifests it (either as an async event, or as a Work Completion), manifest the error in a single place in an IA-specific manner (e.g. for an IA running the IB transport, manifest as a Work Completion, but for an IA running the iWARP transport, manifest as an async event), or allow the Consumer to say that they want all remote errors regardless of transport type manifest as async events. No conclusion was reached as to what would be best.)

Meeting ends at ~12:17pm PST.