OPENFABRICS ALLIANCE

iWarp 2.0

First Stab at a Potential Features List

felix marti (felix@chelsio.com)
March 17th, 2010

Agenda

- Introduction/History
- Lessons Learned
- Proposal
- Features
- Conclusion

Introduction/History

- iWarp support was added to OFED to support IETF RDDP (and RDMAC) compliant RNICs
- iWarp Verbs are *similar* to other RDMA transport Verbs and semantics… but there are differences
  - Same ‘baseset’ (RDMA WRITE, RDMA READ, SEND, RECEIVE, …) but each transport defines its set of additional Verbs
  - Differences in semantics

Lessons Learned

- Applications/ULPs deal with the differences; examples include MPIs of the world, DAPL, Middlewares, NFSoverRDMA, iSER, Custom Implementations
- Different Verbs may lead to transport aware code in the Applications/ULPs which we would like to avoid
- Differences in semantics may go unnoticed until something breaks somewhere

Proposal

- Close the gap between different RDMA transport Verbs and semantics to simplify application/ULP development
- Slightly ‘enhance’ iWarp protocol to ‘simplify’ RNIC implementations
- Disclaimers
  - This presentation is a starting point
  - This presentation is not a Chelsio product roadmap
  - Any standardization needs to be done by the IETF

Graceful Out of RQE Handling

- Motivation
  - Remove stringent Receive Buffer Flow-Control requirement
  - Available on other RDMA Transports – ease application/ULP portability
- E.g. use TCP receive buffering capabilities to hold on to incoming SENDs until RQEs become available
- Does not require a specification update

RDMA WRITE w/Immediate

- Motivation
  - Fuses RDMA WRITE + SEND into a single operation
  - Available on other RDMA Transports – eases application portability
- RDMA WRITE accepts a small amount of immediate data (8-bytes?) that gets placed in a CQE on the remote peer *after* the RDMA WRITE payload has been placed
- Requires a specification update

Atomic Memory Operations

- Motivation
  - Many, e.g. distributed locks, mathematical computation, databases, …
  - Available on other RDMA Transports – eases application portability
- Implement Atomic Memory Operations
  - Store
  - Fetch&Op
  - Compare&Swap
  - …

RDMA READ w/Local Invalidate

- Motivation
  - Reduce RDMA READ* scaling cost
- RDMA READ Response FPDUs are self-describing e.g. contain placement information
- BUT, per RDMA READ work request state is required to distinguish between
  - RDMA READ
  - RDMA READ w/Local Invalidate
- INSTEAD, distinguish based on wire opcodes which requires a specification update

Unreliable Datagram & Multicast

- Hmmm, what does it have to do with iWarp?
- On Ethernet, the unreliable datagram & multicast wire protocol of choice is UDP and not iWarp

We want to provide a solution for applications/ULPs that prefer a QP/CQ interface (with OS bypass) over vanilla sockets

Conclusion

- OFED supports different RDMA transports and we can improve application/ULP portability by minimizing differences in Verbs and semantics
- iWarp 2.0 can be geared towards it
- Requires feedback from application/ULP architects and implementers