A Study of iSCSI Extensions for RDMA (iSER)

Mallikarjun Chadalapaka (HP)
Michael Ko (IBM)
Patricia Thaler (Agilent).

Uri Elzur (Broadcom)
Hemal Shah (Intel)

Outline

- Background
  - The Who, Where
- Motivation and case for iSER
  - The Why
- Layering of iSCSI, iSER & iWARP
  - Stack and functionality distribution
- iSER design features
  - Connection setup, Transformation, Data integriy management
- Changes/extensions to iSCSI
  - What is changed and why
- Enhancements in iWARP protocols
  - Automatic invalidation
- Enhancements to iWARP Verbs
  - Efficient registration of STags
- Next steps
  - Standardization
- Questions

Background

- The iSER paper
  - is based on a (just concluded) protocol design,
  - explores work done by contributors from several companies in the RDMA Consortium,
  - belongs to the “Experience” category – the “E” in NICELI.

- iSER is “iSCSI Extensions for RDMA (iSER)”, iSER maps the iSCSI protocol over the iWARP protocol suite (RDMA over TCP/IP). The focus of this paper is:
  - how iSER enables efficient data movement for iSCSI using generic RDMA hardware
  - how/why certain iWARP architectural features were conceived during the iSER design.

iSCSI, TCP and the challenges therein

- iSCSI is an "application protocol" designed to run on TCP/IP.
  - transports the SCSI protocol exchanges so SCSI I/Os can be done over TCP/IP.

- TCP copy overhead and reassembly buffer requirements were identified as serious acceptance/deployment barriers during iSCSI's early design.

- The iSCSI protocol thus includes an optional feature called "markers".
  - Markers delineate iSCSI PDU boundaries via recurring pointers showing up at fixed intervals within the TCP data stream.
  - The iSCSI markers however aided iSCSI-specific direct data placement (can also be done without employing markers, albeit needing more reassembly memory) that directly places each iSCSI PDU into its final memory location.
  - With or without markers, iSCSI-specific data placement needed an iSCSI-specific NIC to efficiently run iSCSI protocol avoiding TCP data copies.

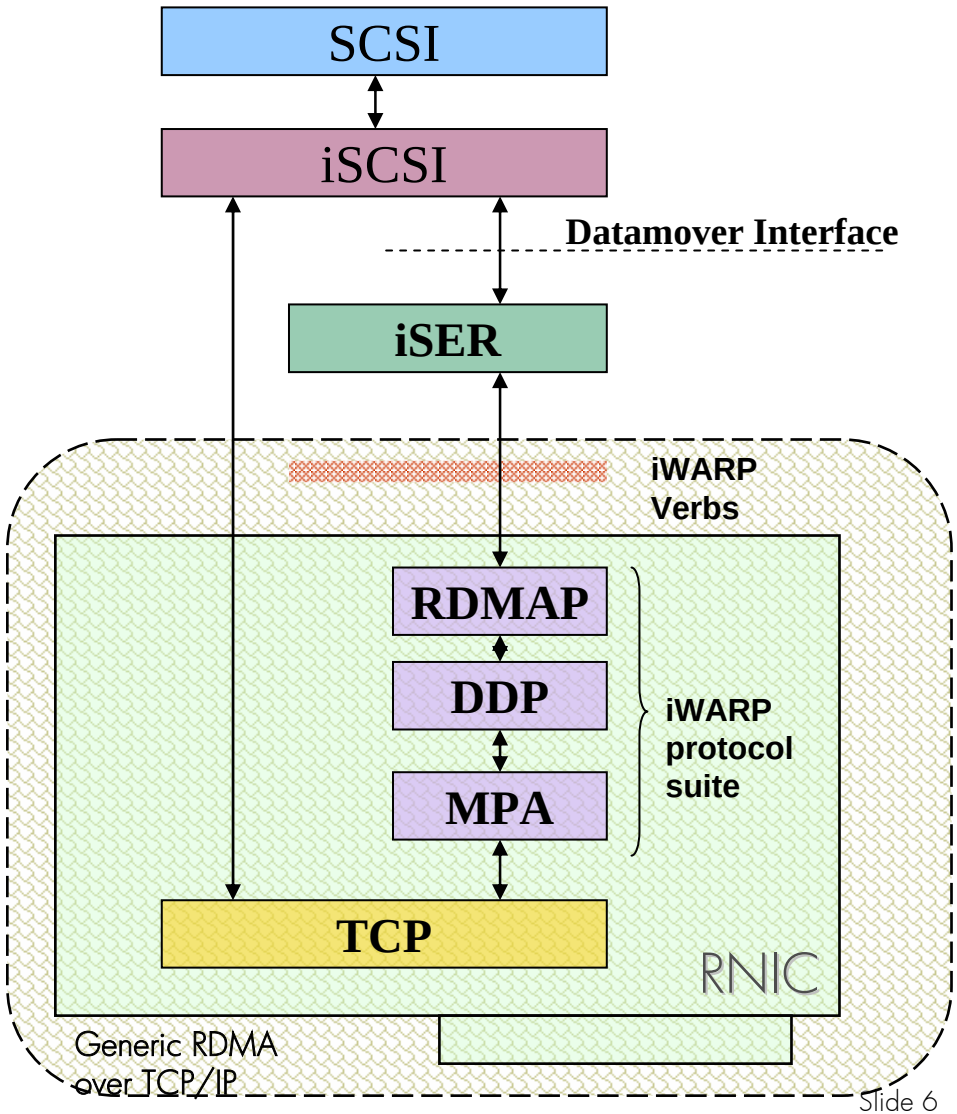
The case for iSER

- Considerations the designers pondered over were -
  - Does RDMA over TCP/IP technology satisfy the data movement needs of iSCSI? If so, when the RDMA technology advances, so does iSCSI.
  - Why tackle fundamental issues such as copy elimination via iSCSI-specific protocol?.
  - Did iWARP say it offers CRC-level reliability on TCP/IP? Let iSCSI take the opportunity to stop playing transport!
  - If nothing else, iSCSI needs iSER to run most efficiently on those (presumed to become) pervasive RNICs (RDMA-enabled NICs) in future.

- The iSCSI designers were thus ultimately convinced of the need for iSER, an "extension" to iSCSI to enable it to run on RDMA over TCP/IP (aka iWARP).

- iSER has the explicit design goal to let iSCSI run on RNICs requiring no greater number of interrupts than an iSCSI NIC does – i.e. run most efficiently on generic RNICs.

iSCSI, iSER and iWARP



- The iSER protocol is designed to run on RDMAP protocol of the iWARP suite.
  - ➔ The paper contains a discussion of why RDMAP was preferred over DDP.
- The iSER wire protocol is dependent only on RDMAP. However, the "iWARP Verbs" are a crucial part of the solution puzzle.
  - During the iSER design, certain Innovations in iWARP Verbs were also made to best meet the needs of iSER.
- The first step was to define an architecture model, "Datamover Architecture", that distilled the needs of iSCSI to generic data movement primitives.
  - iSER design was then mapping the primitives to RDMAP exchanges.

iSER design

- iSER protocol uses the well-known TCP port used for iSCSI connection establishment, rather than using a new iSER well-known port.
  - The iSCSI/iSER connection thus always starts in iSCSI "streaming" mode.
  - A new iSCSI login key used for turning the RDMA (iSER) mode on after login.
  - The existing discovery and boot mechanisms work with no changes.

- Transformation or Encapsulation?
  - A question not traditionally encountered in layered protocols.
  - The iSER protocol simply encapsulates certain iSCSI PDUs (called "control-type" PDUs) in iSER RDMA Send Messages, while it transforms certain other iSCSI PDUs (called "data-type" PDUs) into RDMA Writes or RDMA Reads.

- The iSER protocol relieves iSCSI of having to play transport role
  - iSER mandates that iSCSI-level PDU digests must not be used because iWARP guarantees CRC-level data integrity.
  - iSCSI CRC generation, checking, retransmission requests, retransmissions, timeout-based retransmissions - a lot of complexity in iSCSI is thus gone!

Changes to iSCSI

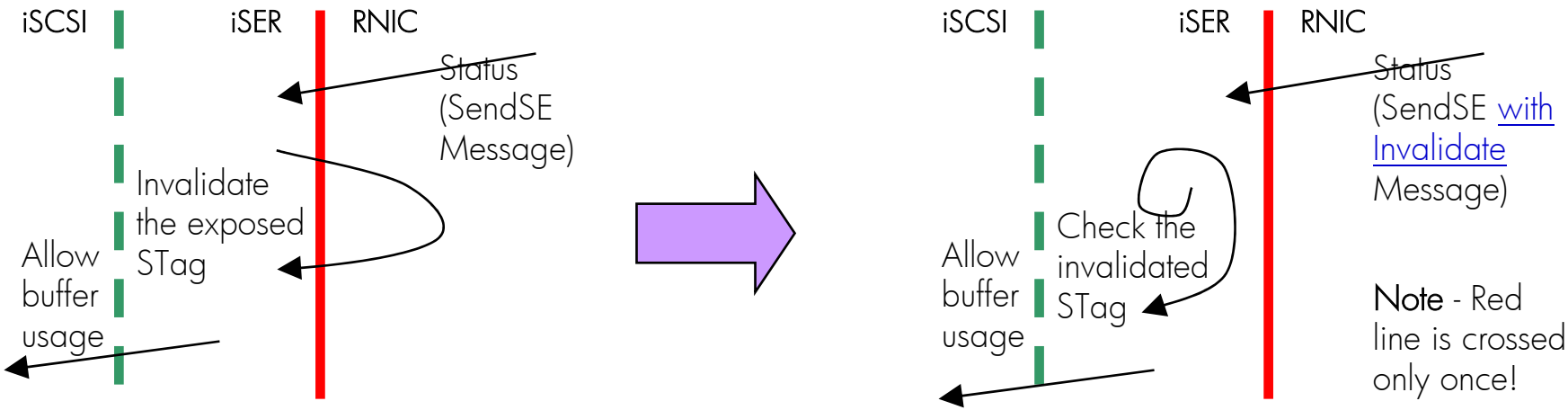
- The biggest set of changes to iSCSI in order to support iSER will be in the area of how iSCSI interfaces to its LLP (lower level protocol).
  - Traditional iSCSI interfaces directly with TCP.
  - Traditional iSCSI is involved in a lot of data movement activity.
  - In the new model, iSCSI simply yields the administration of data movement to iSER, and iSER and iWARP will work together to move the data.

- Wire protocol
  - iSCSI-level PDU digests (header & data) must not be used (so, don't bother to use the PDU level recovery features of iSCSI).
  - No piggybacking of status on the last read data PDU (the receiving RNIC doesn't demux during placement!)

- Other areas
  - Obviously, iSCSI should know to negotiate the new login key – to turn the RDMA (iSER) mode on after login.
  - iSCSI must "chunk" long unsolicited data sequences into PDUs so that each "mid-PDU" is exactly of negotiated max size.

Enhancement to RDMAP (automatic invalidation)

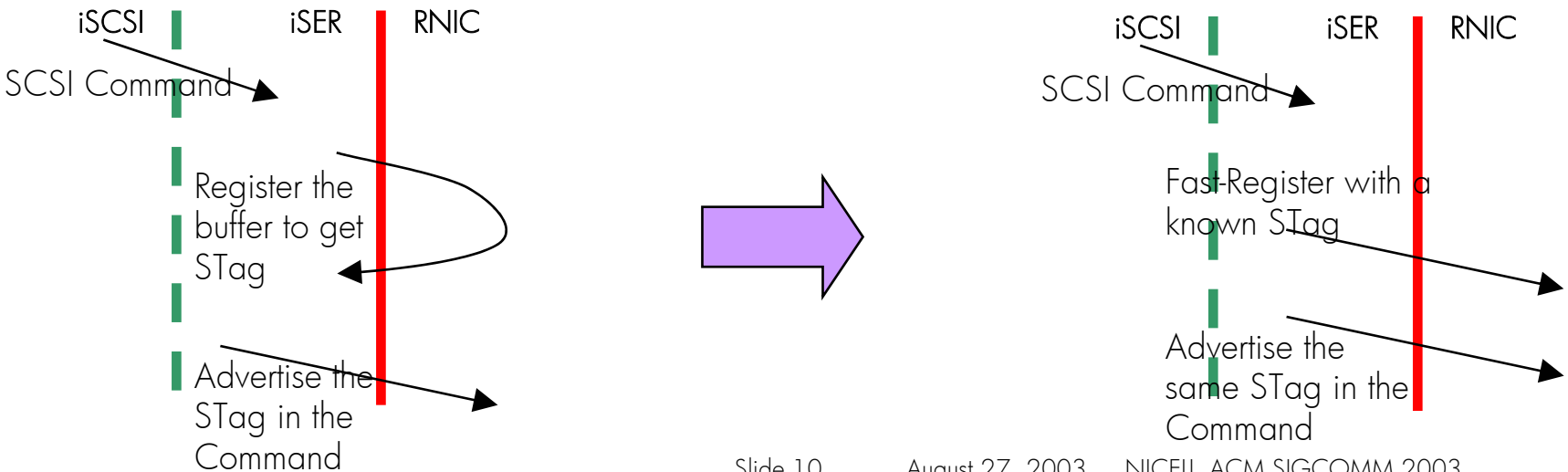
- SCSI has a clearly defined transactional model
  - Command (Initiator -> Target)
  - data (either way)
  - status (Target -> Initiator)
- The initiator iSER layer (client) exposes its STags to the target (server).
  - After receiving the status, initiator iSER layer will invalidate the STag mapping before using those buffers.
  - How about doing this invalidation automatically on receiving the status? That takes one hardware access out from the performance path.



Enhancements to iWARP Verbs (fast register)

- The initiator iSER layer (client) exposes its STags to the target (server).
  - The initiator iSER layer must register the Command buffer locally with the RNIC.
  - Registration process yields the STag, so must precede the advertisement.
  - This is a synchronous wait for a hardware response in the performance path.
- In the fast-register model, the STag is allocated to iSER apriori. It is merely associated with the Command buffer during runtime.
  - The "fast-registration" is now guaranteed to succeed.
  - The initiator iSER layer can post the fast-register and command requests to the hardware back-to-back, no more waiting.

➔ The paper also discusses automatic deregistration and Shared Receive Queues.



Next Steps

- The Datamover Architecture for iSCSI (DA) and iSCSI Extensions for RDMA (iSER) specifications were publicly released by the RDMA Consortium on July 21, 2003 (all specs available on **www.rdmaconsortium.org**).

- Several Consortium member companies are working on productization of the iWARP protocol suite and iSER.

- Both DA and iSER specs are submitted to IETF as Internet Drafts for pursuing standardization.

Thank you!

- Questions?