SRP-2 Multichannel Architecture

Cris Simpson
cris.simpson@intel.com

17 March 2003

Blue text - additions Red text - deletions

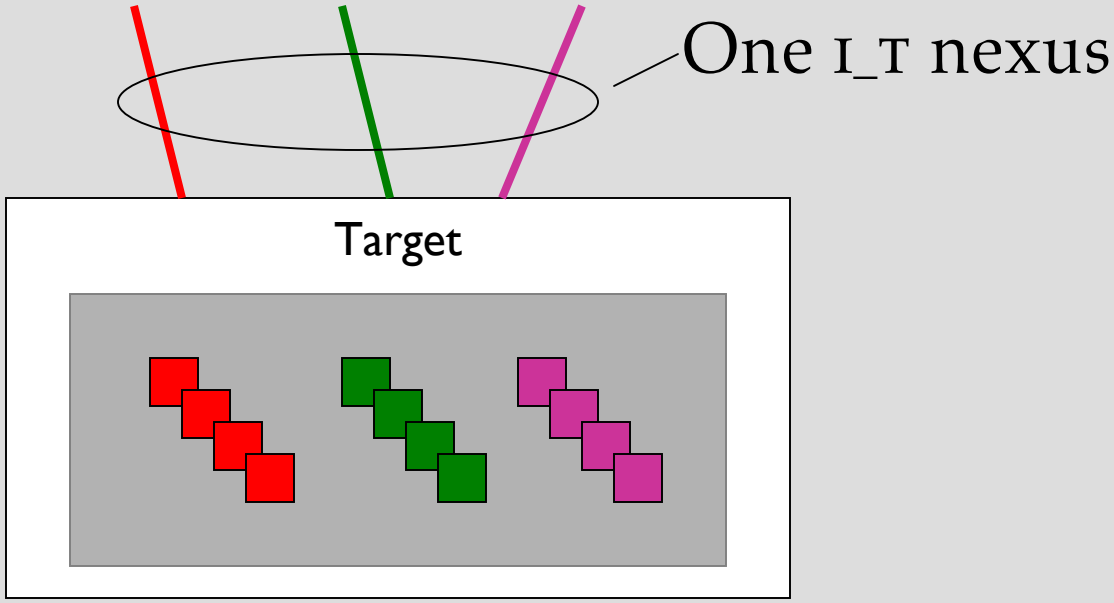
Goals

- Increase availability
  - Robust multiple paths
- Support distributed initiators
- Improve data transfer performance

This presentation intended to gather support for proposed directions.

SRP multichannel support

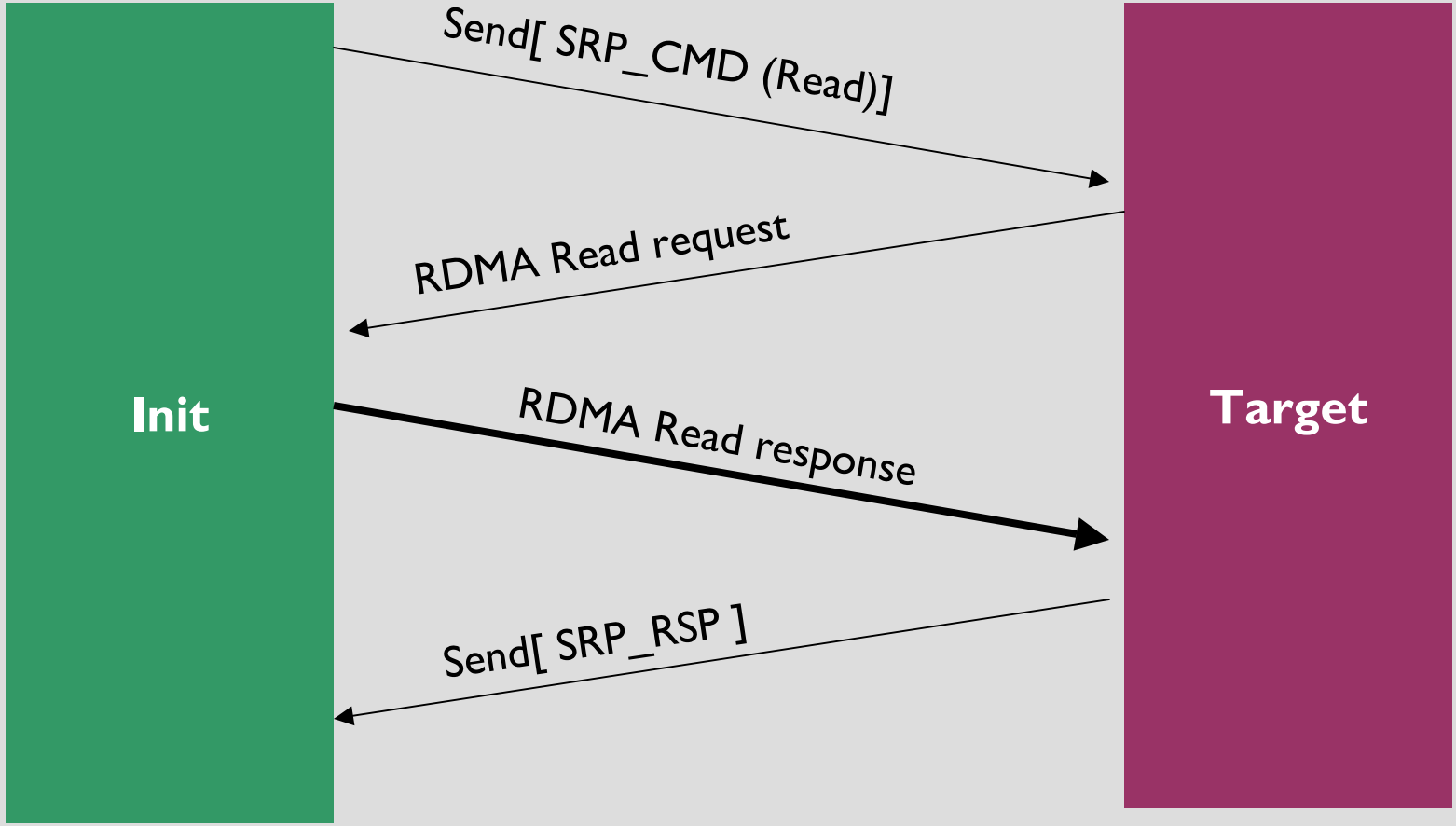
- Multiple channels, one I_T nexus
- Tasks are affiliated with channels
  - Data can be transferred only on sending channel
  - When channel dies, the target aborts affiliated tasks
    - This model reduces robustness



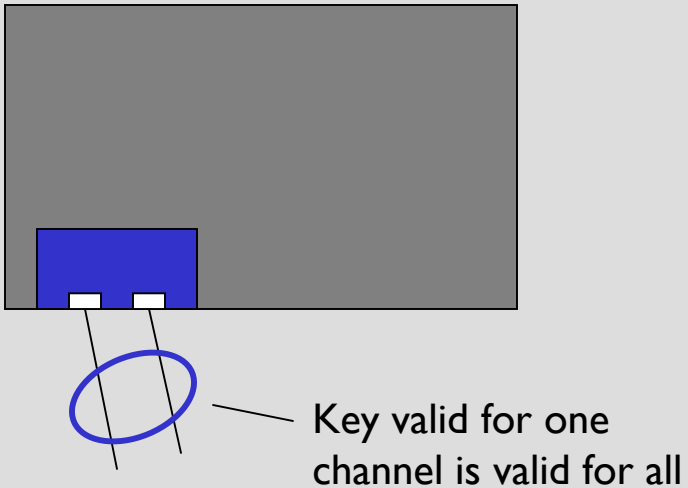
Two classes of RDMA operations

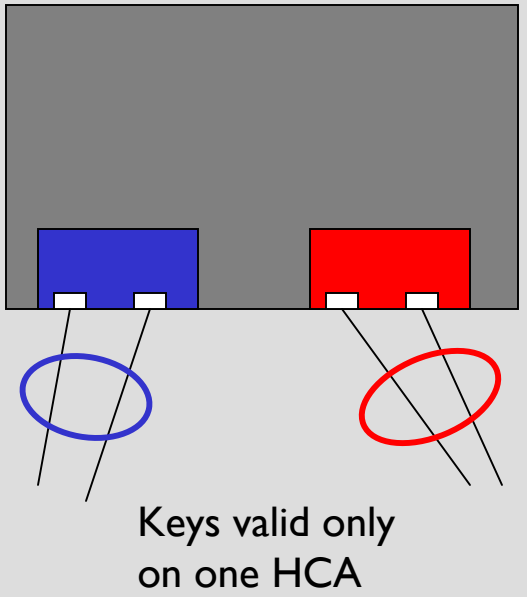
- Sends (command and status)
  - Data placed into receiver-specified buffer
- RDMA (data)
  - Data transferred to/from requestor-specified buffer
  - Transfers initiated by target
  - Buffer advertised (in this context) by initiator
  - Buffer described by [key, virtual address]
  - Key generated by operating system, value not predictable

SRP operations



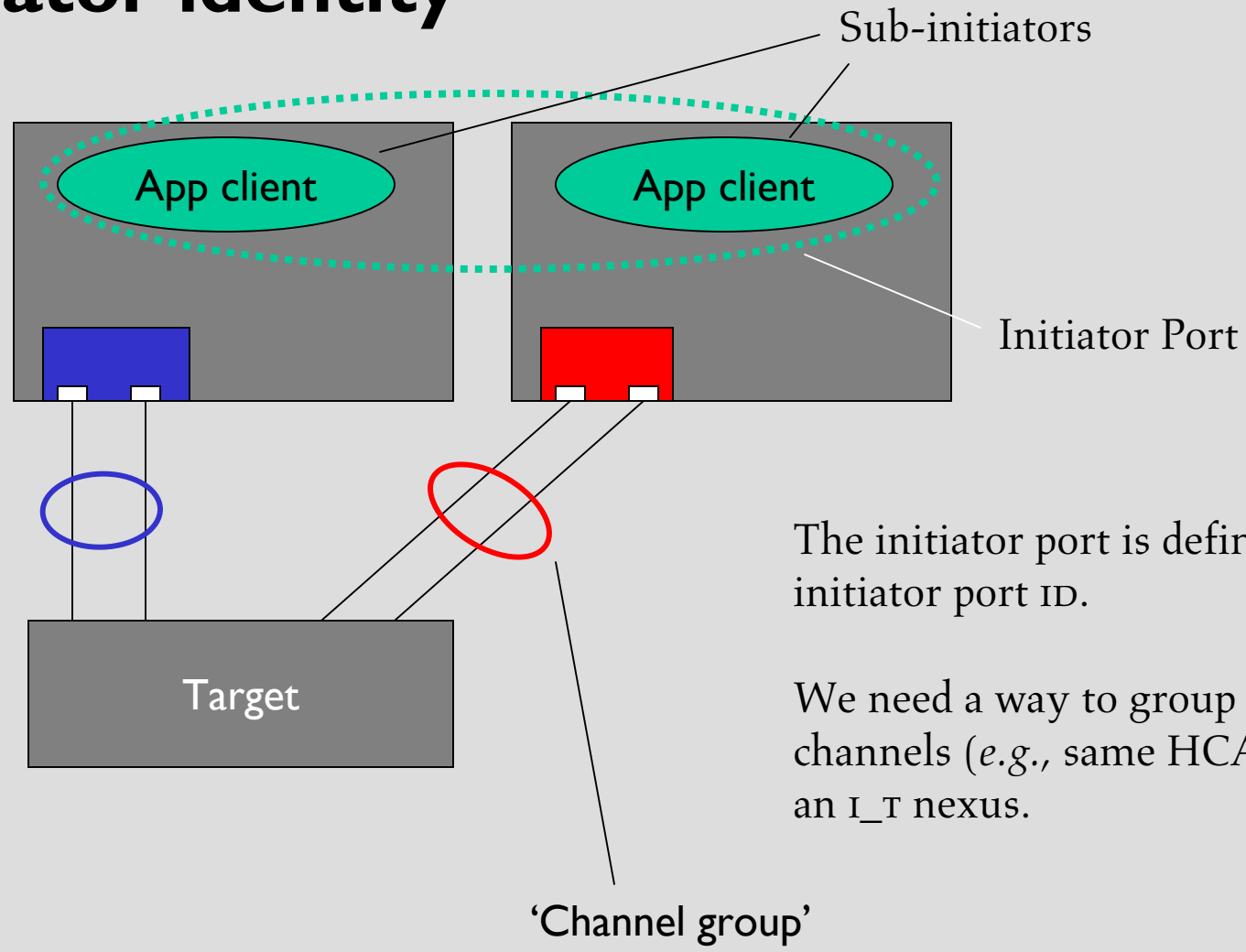
Key affinity





Data traffic is only possible on specified HCA

Initiator identity



The initiator port is defined by the initiator port ID.

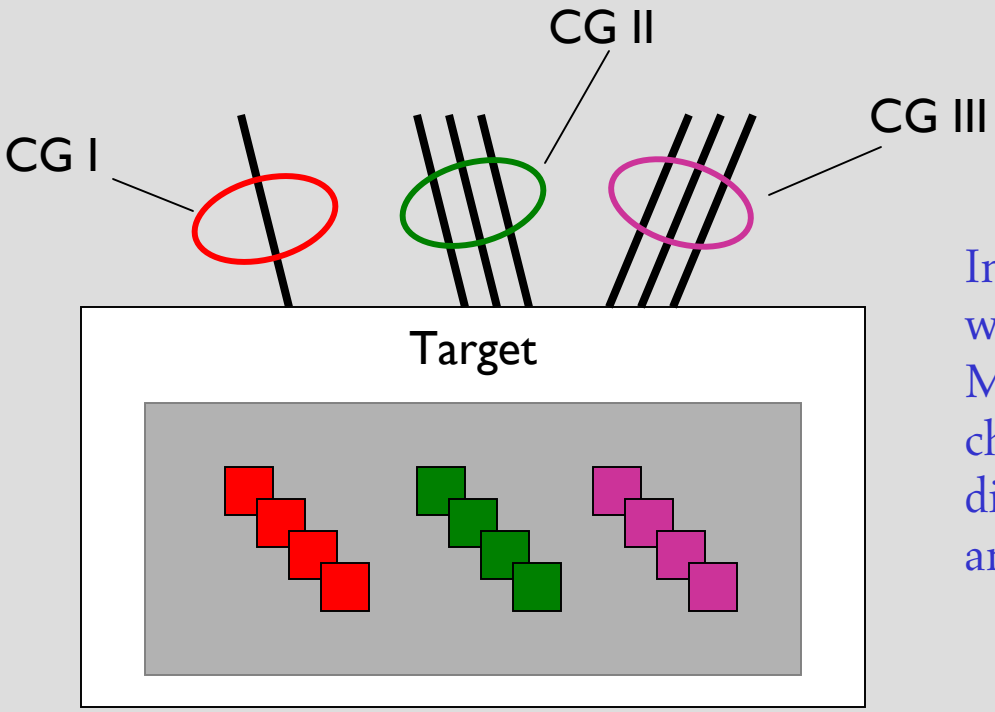
We need a way to group related channels (*e.g.*, same HCA) under an I_T nexus.

Channel groups

Figure shows one I_T nexus with three channel groups.
Tasks may be submitted, managed over any channel.

Data shall be transferred over one or more of the channels in the specified CG.
Status should be returned

1) On the specified submitting channel;
2) On a channel in the same channel group; or
3) On any other channel



Initiator specifies CG for data xfer - where does status go?
Must consider ordering issues across channels for data/status, including distinction between ACK to remote and availability of data to application.

Channel failures

- A channel failure does not cause a Task Abort
  - Only the failure of all channels in the channel group
  - *Presumption is that failed channel will be quickly re-instantiated in the channel group*
- A channel failure does not cause an I_T nexus loss
  - Unless it was the last channel of the nexus

Buffer credits

- RDMA endpoints must pre-post receive descriptors
- SRP
  - Explicit per-channel I ➨ T credits
  - One explicit credit for target-initiated IUs
    - For AER, credit management
  - Implicit T ➨ I credits, based on issued tasks
- SRP-2
  - How to enable multichannel flexibility? (*i.e.*, returning response on different channel)
    - Targets may be able to pool descriptors – study as option
    - Harder on initiator side – may need to add explicit T ➨ I credits

Ordering

- There is no ordering expectation across channels
- Since sub-initiators must work together (*i.e.*, tag assignments) there may not be a need for a command sequence number
  - On the the other hand, tag assignments need not be in the speed path

Channel establishment

- First channel of I_T nexus established
  - Target returns CHANNEL GROUP ID
- Subsequent channels
  - If adding to existing group
    - Initiator presents CHANNEL GROUP ID
  - Creating a new group
    - Initiator requests a new group
    - Target returns new CHANNEL GROUP ID