IPoIB Stateless Offloads and More

Dror Goldenberg
Mellanox Technologies

Agenda

- Stateless offloads outstanding patches submitted by Eli Cohen
  - Checksum
  - TSO
  - LRO
  - Interrupt Moderation
- Future work
  - RCA, IPv6, LRO
- IPoIB topics
  - IPoIB-CM hw_csum patch
  - IPoIB-CM UC connection keep alive

Checksum Offload

- Supports
  - IPv4, TCP and UDP
    - No IP options
- Includes
  - Device capability
  - Send WQE
    - IP checksum bit
    - TCP/UDP checksum bit
  - Poll CQ
    - Checksum OK bit (refers to TCP/UDP)
  - IPoIB modification
    - Pass on checksum indications from/to verbs
    - HIGH_DMA, S/G
    - Checksum offload turned on (no disable)
- Requires
  - <none>

Support Matrix

	IPoIB UD	IPoIB CM
InfiniHost		
InfiniHost III-Ex	✓	
InfiniHost III-Lx		
ConnectX	✓	

TCP Segmentation Offload (TSO)

- Supports
  - TCP over IPv4
- Includes
  - Device capability
  - Send WQE
    - TSO opcode
    - Pointer to IPoIB header & TCP MSS
  - IPoIB modification
    - Pass on TSO indications to verbs
    - TSO turned on (no disable)
- Requires
  - Checksum offload

Support Matrix

	IPoIB UD	IPoIB CM
InfiniHost		
InfiniHost III-Ex		
InfiniHost III-Lx		
ConnectX	✓	

Large Receive Offloads (LRO)

- Supports
  - TCP over IPv4
- Includes
  - Identify LRO candidates
    - Mainstream TCP/IP segments
      - Non fragmented, aligned timestamp, no special flags (syn/rst/etc)
  - Accumulate
    - Linked list of skbs, up to 64KB per session
    - Sessions are accumulated till CQ is drained
  - lro_enabled – module parameter
    - Enabled by default
- Requires
  - Checksum offload
- TODO
  - Will be replaced by inet_lro.c (generic LRO in kernel)

Support Matrix

	IPoIB UD	IPoIB CM
InfiniHost		
InfiniHost III-Ex	✓	
InfiniHost III-Lx		
ConnectX	✓	

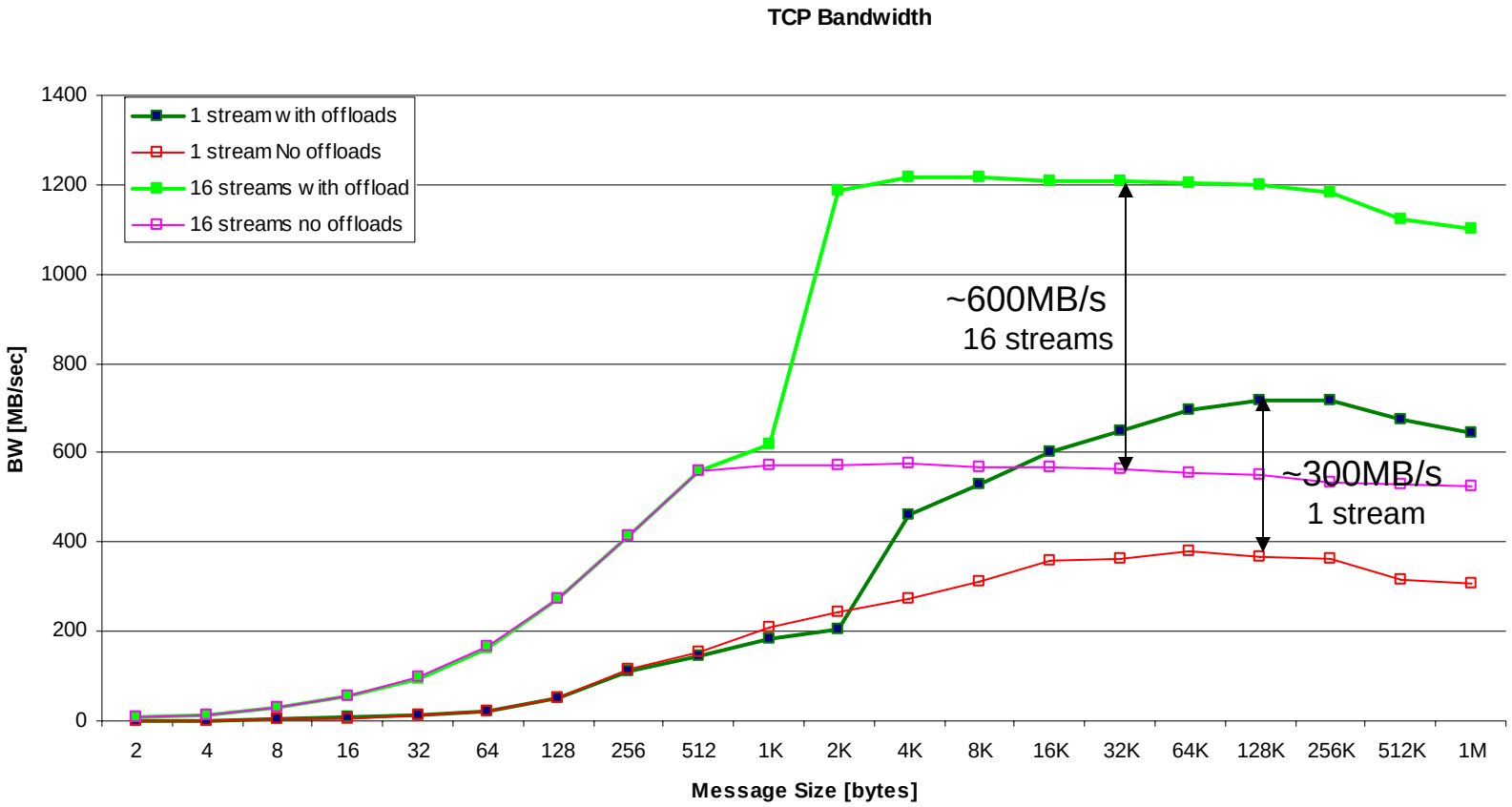
Interrupt Moderation

Support Matrix

	IPoIB UD	IPoIB CM
InfiniHost		
InfiniHost III-Ex		
InfiniHost III-Lx		
ConnectX	✓	✓

- Supports
  - IPv4, IPv6, TCP and UDP
- Includes
  - Modify CQ - Per CQ moderation parameters
    - Max CQEs to trigger an event
    - Max time from 1st CQE (usec) to trigger an event
  - IPoIB modification
    - Control moderation on Send/Receive CQs
    - Moderation settable by ethtool (0 – disable moderation)
      - ethtool -C ib<num> rx-frames <number>
      - ethtool -C ib<num> rx-usecs <number>
    - Enabled by default (through openibd script)
- Requires
  - <none>

Stateless Offloads Performance Improvement



IPoIB-UD
Iperf 1 and 16 streams
ConnectX DDR FW 2.2.0
Dell Power Edge 1950
2.6.18 Red Hat EL 5

Includes:
- Checksum offload
- Interrupt moderation
- TSO
- LRO

Status

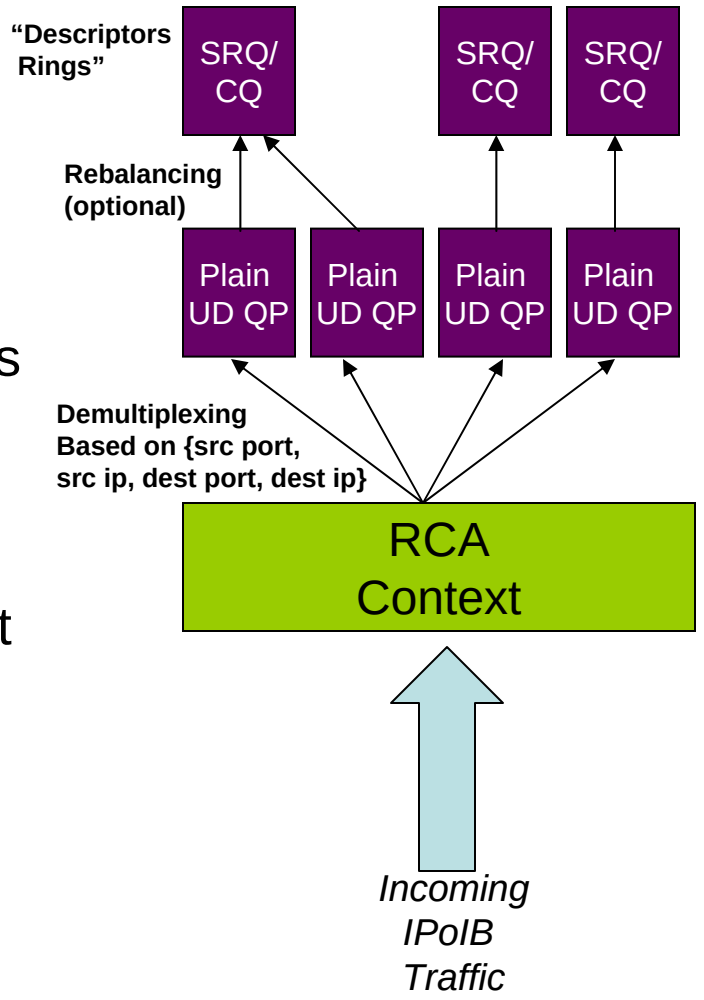
- Integrated into OFED 1.3
- Mainline kernel integration
  - Patches updated to 2.6.24, will be resent to review
- LRO
  - Will move to the new inet_lro.c patch

Future Work

- Add support for IPv6
  - Checksum offload
  - TSO
  - LRO (part of lro_inet.c)
- Receive Core Affinity (RCA)…

Receive Core Affinity (RCA) Architecture

- RCA Context
  - Replaces the IPoIB UD QP
  - Demultiplexes incoming IPoIB packets
    - Hashing {src/dst IP, src/dest port} - TCP
    - Hashing {src/dst IP} – UDP/Fragments
  - Packet reach a set of consecutive UD QPs
- UD QPs handle incoming packets
  - Can be the final destination; or
  - Can dequeue WQEs from SRQ and report CQEs to shared CQs
    - Rebalancing can be applied on run time



Receive Core Affinity (RCA) Implementation Suggestion

- Device Capability
- Create an RCA context
  - Includes UD QP (Send Queue) for IPoIB
  - Includes the RCA context with N child UD RQs
    - Have to be consecutive (mlx4 can reserve…)
  - Need to figure out the appropriate APIs
- Add support for multiple EQs
  - And multiple MSI-X
  - Enable CQ->EQ remapping for rebalancing
  - This is beneficial for other applications as well
- Enable Modify QP to alter CQ/SRQ affiliation

Support Matrix

	IPoIB	
	UD	CM
InfiniHost		
InfiniHost III-Ex		
InfiniHost III-Lx		
ConnectX	✓	

And More…

OPENFABRICS
ALLIANCE

Hw_csum patch

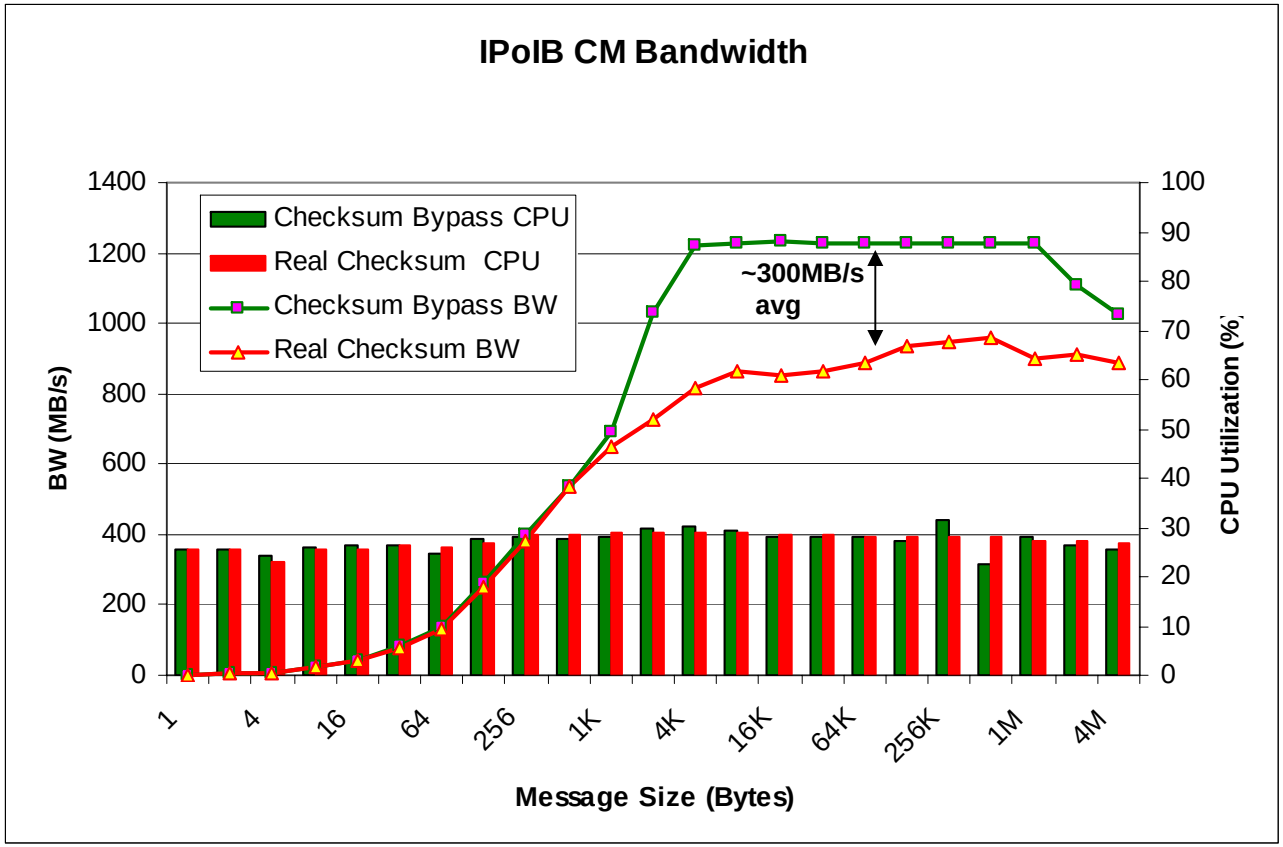
- IPoIB CM only
  - Enables checksum calculation bypass
  - End to end integrity ensured by IB ICRC
    - As long as we stay in the local subnet…
- HW Address
  - Added checksum bypass bit
  - If set, interface accepts packets without checksum
- IPoIB header
  - Added checksum bypass bit
  - If set, packet checksum is not valid
- Module parameter (hw_csum)
  - Turned off by default
  - Administrator decision whether to use or
  - Should not be turned on along with IP forwarding

Support Matrix

	IPoIB UD	IPoIB CM
InfiniHost		✓
InfiniHost III-Ex		✓
InfiniHost III-Lx		✓
ConnectX		✓

- Flow
  - Send
    - If peer supports checksum bypass
      - send packet without checksum and set IPoIB header bit
    - Else
      - Calculate checksum
  - Receive
    - If IPoIB header indicates checksum bypass
      - Indicate packet with CHECKSUM_UNNECESSARY
    - Else
      - OS will do the checksum

Hw_csum Performance



IPoIB-CM 64KB MTU
netperf
ConnectX DDR FW 2.2.0
Dell Power Edge 1950
2.6.22.1

Includes:
- Interrupt moderation
- Hw_csum patch

IPoIB-CM UC Connection Keep Alive

- RC indicates connection loss
  - Retransmission timeout ⇨ completion with error
- UC does not indicate connection loss
  - Current implementation QPs are unidirectional
  - CM can detect stale connections, but it can take forever
  - Remote side trying to establish an IPoIB-CM connection is not an indication
    - Remote side may choose to establish >1 connection
- Remote node reboot issue
  - It is likely to obtain the same IPoIB-UD QP number after a reboot
  - HW address does not change
  - IPoIB-CM keeps sending on the UC QP, remote side doesn't have it opened
    - Packets are discarded…
- What can we do about it?
  - Tear down and establish connections periodically
    - Not clean, can cause packet drop/reorder
  - Implement protocol specific (IPoIB-CM) keep alive
    - Not clean, why do we need something specific for IPoIB-CM ? It's clearly a UC issue.
  - Use CM mechanism to check if the connection is alive
    - One way is to use LAP/APR (try loading alternate path = primary path and get a different error if connection is alive or not)
    - Architect something in IBTA to solve this