SoftiWARP – Project Update

A Software iWARP Driver for OpenFabrics

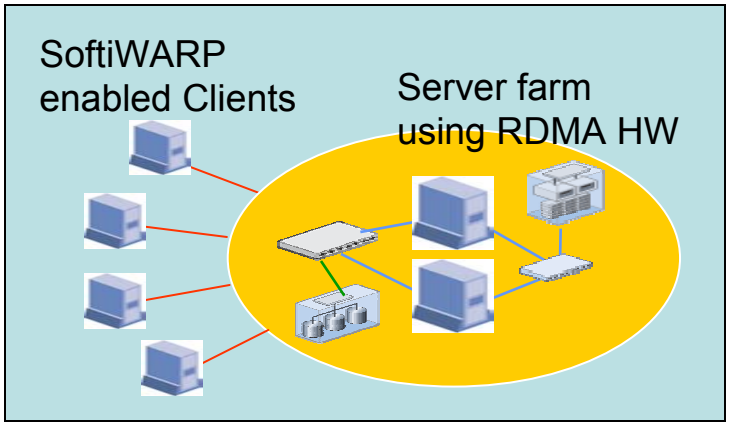
Bernard Metzler, Philip Frey, Animesh Trivedi
IBM Zurich Research Lab

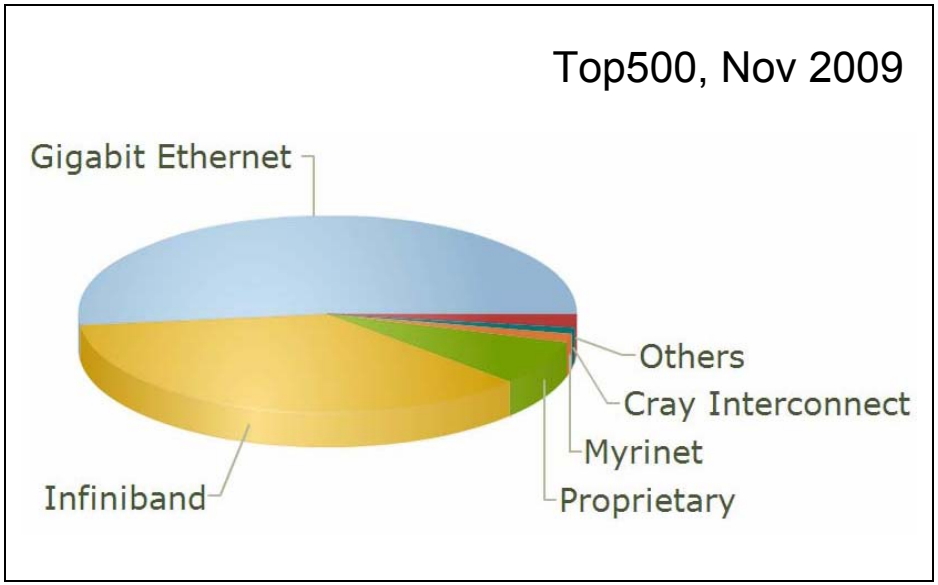
Agenda

- Background
- SoftiWARP and OFED
- Project Status
- Benefits
  - Data Touching Operations
  - Async. API
- Outlook

Background

- RDMA first guess: High perf./low latency
  - HW supported transport stack
- Also RDMA:
  - Async. communication semantics
  - One-sided operations
  - Explicit buffer ownership management
- There are benefits from running stack in SW:
  - *Enable* RDMA: Peer 'real' RDMA HW
  - RDMA semantics available to application

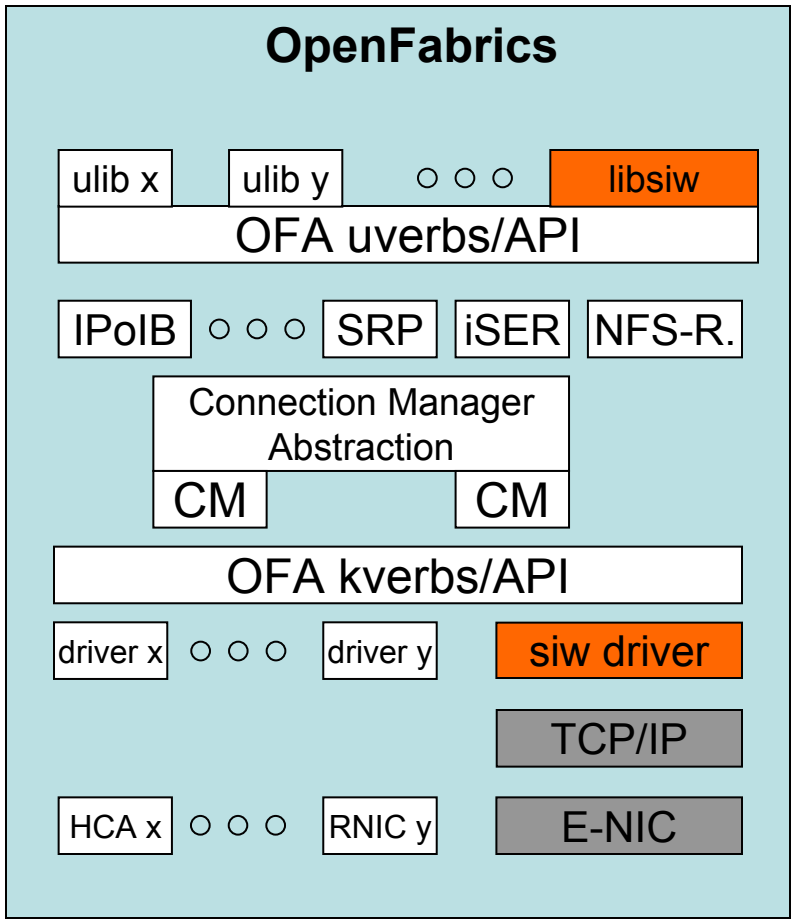




- OpenFabrics
  - Framework for RDMA Host Enablement
  - InfiniBand + iWARP + RDMAoE
- IBM Zurich Research
  - RDMA API standardization
  - IETF work on iWARP
  - Software iWARP stack

OpenFabrics Integration

- ../driver/infiniband/hw/siw
  - The ‘hw’ is the CPU
- ‘siw’ Kernel Module
  - exporting OFA interfaces
  - Uses non blocking TCP kernel sockets
- ‘libsiw’ User Library
  - integrates with libibverbs, librdmacm
- Current Build
  - Linux 2.6.34-rc1
  - Tested in OFA1.4.1



Project Status

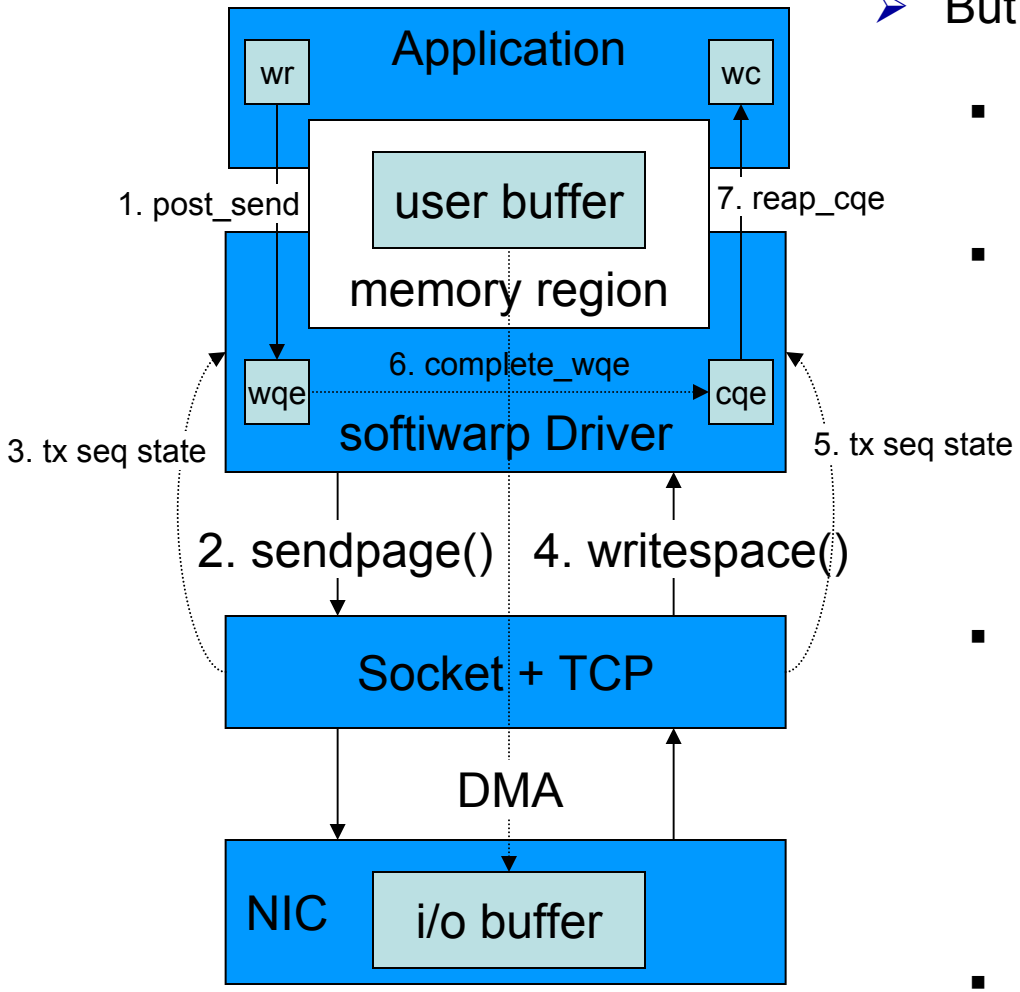
- Open Sourced at http://gitorious.org/softiwarp
  - Kernel Module at http://gitorious.org/softiwarp/kernel
  - User Lib at http://gitorious.org/softiwarp/userlib

- Work in progress:
  - Main functionality done. Peers with Chelsio T3.
  - Whats next?
    - Termination Messages
    - CM redesign (under way)
    - Kernel client support
    - Fast Registration
    - MW's
    - Sending MPA markers (?)
    - …its open source, so its up to you!

Benefits: Less Data Touching

- 0copy TX, if:
  - TCP can safely retransmit from source buffer,
  - API and Transport semantic allows for source buffer content altered.
- With RDMA we have:
  - Pinned source buffer (Memory Region),
  - Explicit send buffer ownership (Work Request/Completion).
- SoftiWARP experimentally does sendpage() on payload if:
  - Non-signalled work requests
    - All Read.Resonse, non-signaled WRITE/SEND,
    - Some byte length threshold exceeded.
  - No CRC (CRC would break, IPsum not)
  - Delayed memory dereferencing to be done
    - Use pipe cleaner like signaled zero length READ, or --->

Less Data Touching (2)



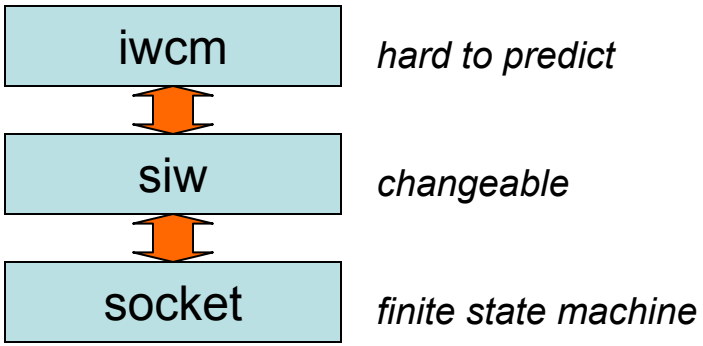
- But, what if TCP/socket tx state is known?
  - Per spec., iWARP completes signaled WR after handover to TCP (here is the copy)
  - What if WC gets delayed appropriately?
    - Needs some TCP tx state information
    - Compute TCP send sequence # or check tx queue length on write() to socket
    - Compare with TCP ack sequence # or queue length on writespace_available() callback
    - Generate WC only if data belonging to WR is ack'd by peer TCP entity.
  - Low hanging fruit, but:
    - User (siw) must trigger writespace() upcall and read some provider state (socket/tcp),
    - Violates layering,
    - Discuss/seek for advise from Linux TCP stack maintainers,
    - Currently not part of siw.
  - Useable for all end-to-end work requests

Benefits: Receive Path

- Data placement during socket callback (in softirq)
  - Technique known from iscsi_tcp driver.
  - All target buffers always pinned.
  - No extra thread for data placement needed.
  - No application scheduling (only for reaping completions)
- Read.Response: softirq schedules work on Inbound Read Queue
  - No remote application scheduling for pulling data
  - No data cache pollution
  - Efficient data serving (web, multimedia, …)

Current Activities

- SoftiWARP@BlueGene/P experiment
  - Scalability: 1000's of nodes to run siw
  - GPFS and MPI on iWARP/SoftiWARP
  - Linux-kernel + siw + libsiw
  - OFA standard verbs
  - GPFS on siw may benefit
    - Think about file write (it's a RDMA.READ)
- Code Development
  - Connection manager redesign (latest mvapich unveiled need)
  - Start looking into performance
  - Code the missing pieces



Future Activities

- Code stabilization/maintenance:
  - Get together in an Open Source project!
- Performance Considerations
- Interoperability testing
- Kernel client interface & kernel clients (iSER, NFS, what else?)
- Work towards a well accepted code base for mainline Linux integration

Summary

- SoftiWARP
  - Respects and tries to efficiently use current Linux Network Stack
  - Contributes to the RDMA ecosystem
  - Helps to spread OFED
  - Is RDMA semantics for the people
  - It’s green ;-)

Thank You.