# INTEGRATING HIGH SPEED FABRICS INTO THE LINUX NETWORK STACK

*Christoph Lameter,*
2017 linux.conf.au in Hobart
*cl@linux.com*

# OVERVIEW

- High speed fabrics vs. regular networking
- The RDMA side car in the kernel
- An RDMA device is a network device after all
- An RDMA transport is a special protocol run on the device
- Infiniband/Omnipath/etc are lower layer protocols supporting basic messaging like any other protocol.'

- Full integration into the network stack and other mechanisms.
- Fast network I/O is a key problem of the Linux network stack today. Above 10GB/sec the existing Posix mechanisms have not enough performance anymore.
- Need to add more mechanisms to the kernel to do fast I/o. NOHZ already allows one to dedicate a cpu to a thread eliminating OS latencies.
- Posix mechanism is too slow and requires copy. We need to add new fast communication mechanisms to the kernel.

# INFINIBAND PROTOCOL SUPPORT

- Infiniband Addresses supported
- Messaging using native IB addressing via Posix Calls (sendmsg, rcvmsg)
- Using ethtool / netstat /ss whatever with these devices
- Potential support in netfilter etc.

# QP SUPPORT FOR ARBITRARY DEVICES

- Method to do offload with any netdevice.
- QPs for pipes
- QPs for Ethernet devices
- Flow specifications to direct traffic to QPs setup.

# RDMA ON ANY NETDEVICE

- Connect a QP to a netdevice and specify the end point. Register memory etc etc.
- Protocol specific way to establish a RDMA connection. If its ethernet using ROCE. Perform software emulation if the device does not support it. IWarp is possible if the other side does not support ROCE.
- Omnipath/Infiniband support native RDMA and would be able to do this without emulation.

# WITH THIS

- We are able to generically run RDMA connections between any endpoint.
- RDMA can go mainstream.
- It can be used in other contexts like communication with accelerators or FGPA and all sorts of special devices.
- No need for ROCE devices

# MULTICAST IMPLICATIONS

- Send/Receive both with Posix and QPs to devices.
- Important for third party libraries provided by Exchanges
- Native IB without IPoIB encapsulation in the fabric simplifying multicast handling and shortening packets.
- At the border to Ethernet we would need a translation/ conversion from IB multicast group to IPv4 MC group. Gateway without IGMP on the IB side. Querying the SM. One place where MC subscription information is kept.
- Ethernet mode can use regular offloads via networks stack while also doing QP based messaging I/O and true RDMA transfers.

# EXAMPLES FOR NEW "SOCKET" FUNCTIONS

- fd = queue_pair(domain, flags, *qpdesc)
- setqpopt, getqpopt
- qp_connect(fd, sockaddr, len)
- fd = qp_accept(fd, *qpdesc)