



# RDMA Software Support for Broadcom Ethernet NICs

**Dr. Hemal V. Shah**

Data Center Solutions Group (DCSG), Broadcom Inc.

## Agenda

- **Broadcom Ethernet NICs for HPC/ML**
- **RDMA Software Components**
- **Linux RDMA Software Infrastructure Support**
- **RDMA Software Infrastructure Support for HPC/AI/ML**
- **Peer Mem Direct Infrastructure Support**
- **Summary**

# Broadcom Ethernet NICs for HPC and AI/ML

## Ethernet NIC

- 200 Gbps port speed
- Single/Dual port configurations
- Low latency data path
- High packet per second pipeline

## Offloads

- RDMA (RoCE)
- QoS
- Flow processing
- Partitioning and SR-IOV

## RoCE

- Verbs
- UCX
- MPI (OpenMPI, MVAPICH2...)
- Peer Mem Direct & Collectives

## Software Support

- Congestion state management
- QP rate control
- ECN marking
- CNP generation

## Congestion Control

## Summary of Broadcom Ethernet NIC RDMA SW Components

Environment	Software Components
Linux	<ul style="list-style-type: none"><li>• User mode verbs library (libbnxt_re) – open source, upstream</li><li>• Kernel mode driver (bnxt_re) – open source, upstream, in-box</li><li>• Peer Memory Module – out-of-box</li><li>• NVM configuration tool (bnxtnvm), QoS tool (bnxtqos)</li><li>• Support for standard Linux tools including ethtool, devlink, and lldptool</li><li>• Installation and configuration scripts to simplify RoCE configuration and deployment</li></ul>
VMware	<ul style="list-style-type: none"><li>• Kernel mode driver – in-box, PVRDMA support</li><li>• NVM configuration tool (bnxtnvm)</li><li>• Support for standard VMware tools including esxcli for QoS and DCB</li></ul>
Windows	<ul style="list-style-type: none"><li>• User mode library (supports NDSPI)</li><li>• Kernel mode driver (supports NDKPI) – in-box</li><li>• All three RDMA modes supported: RDMA over PF, RDMA over vNIC, RDMA over VF</li><li>• NVM configuration tool (bnxtnvm)</li><li>• Support for standard Windows tools for QoS and DCB configuration</li></ul>
Firmware	Control plane firmware for RDMA resource management (OS-independent)

<https://techdocs.broadcom.com/us/en/storage-and-ethernet-connectivity/ethernet-nic-controllers/bcm957xxx/1-0/RDMA-over-Converged-Ethernet.html>

# Linux RDMA Core Software

- **Native RDMA Support**

- Most Linux distros and kernels

- **rdma-core userspace**

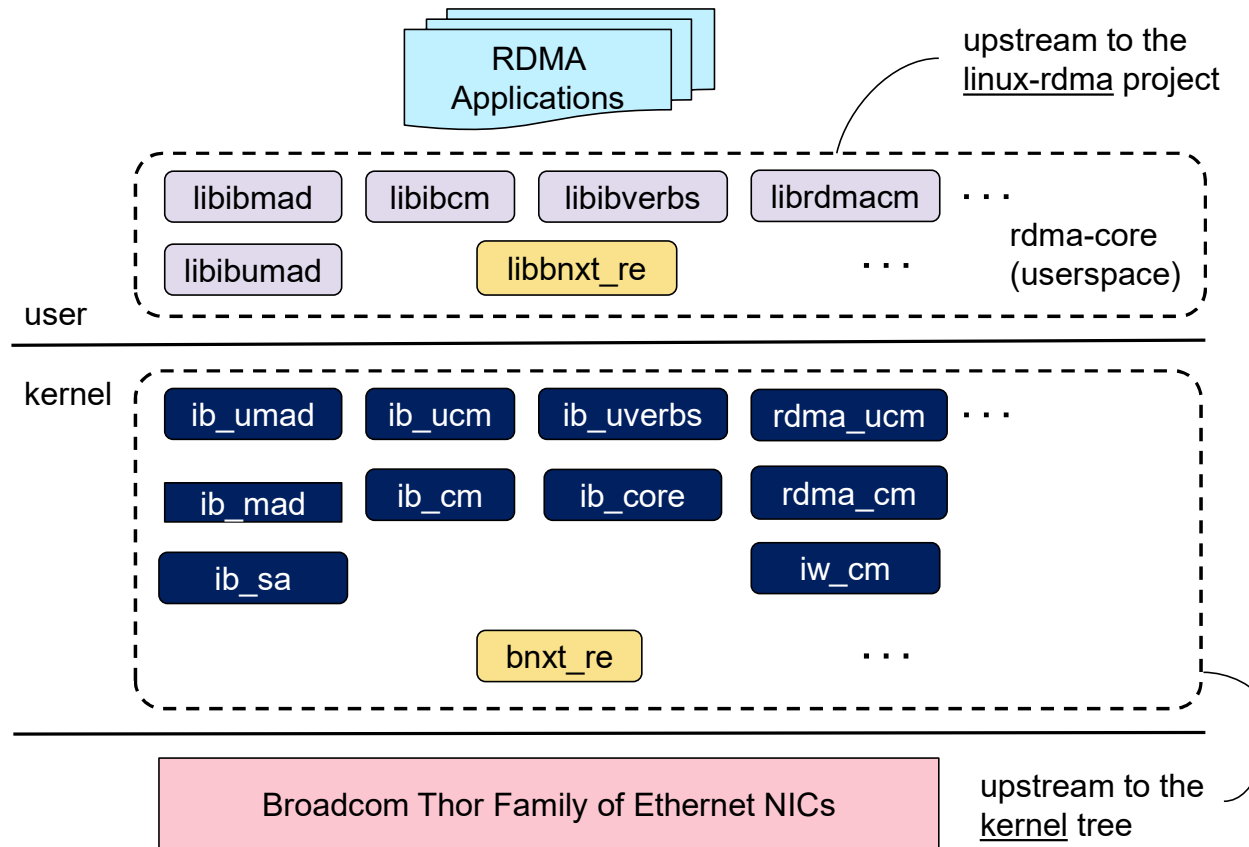
- **libbnxt\_re** user-space library for RoCE
- Upstream in linux-rdma project

- **Kernel driver: bnxt\_re**

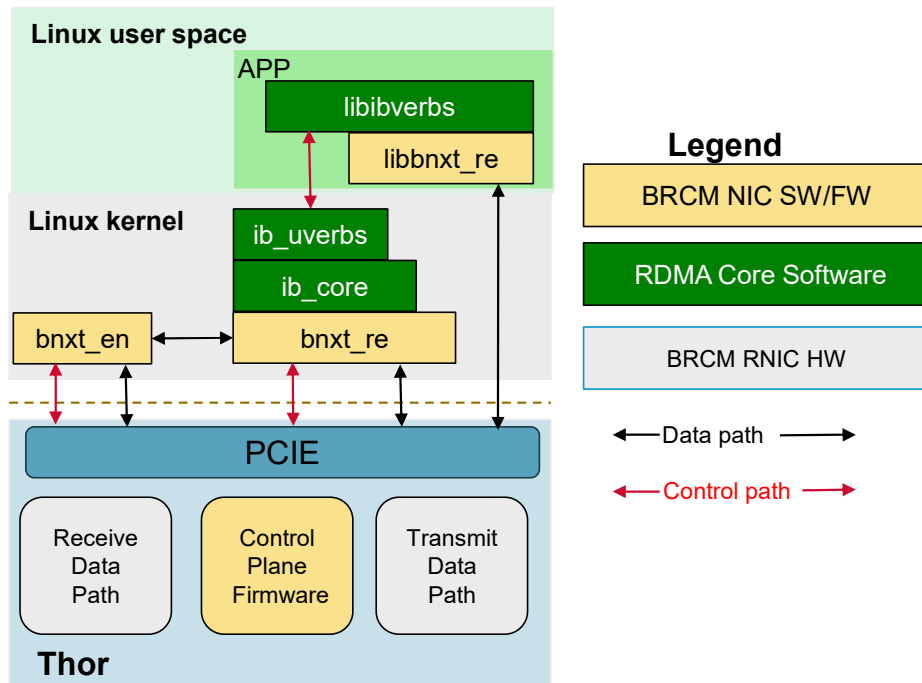
- Broadcom provides **bnxt\_re** RoCE driver
- Upstream in Linux kernel tree

- **Out of Box versions**

- **libbnxt\_re** and **bnxt\_re** out of box versions available from Broadcom for latest features



# Thor RoCE Software for Linux



## • Broadcom Linux RoCE Components

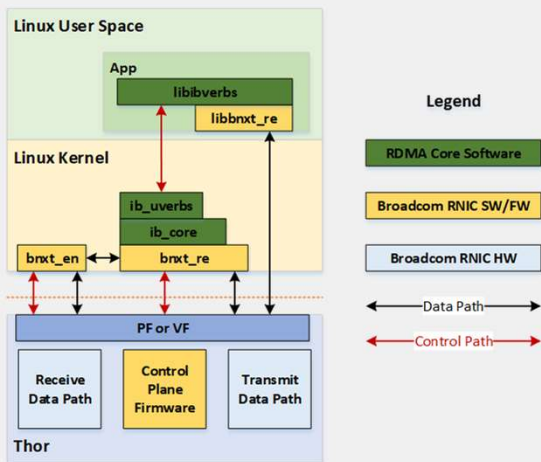
- RoCE User Library (libbnxt\_re)
- RoCE driver (bnxt\_re)
- NIC driver (bnxt\_en)
- RoCE Control Plane Firmware

## • Advanced Software Features

- Performance profiles
- Doorbell pacing and recovery
- Error recovery
- ...

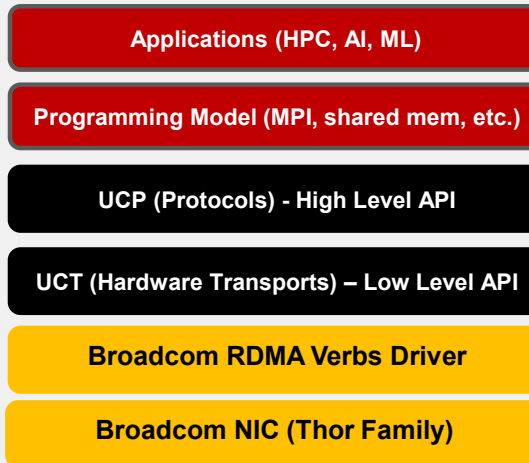
# Linux Software Infrastructure Support for HPC/AI/ML

## RDMA Verbs



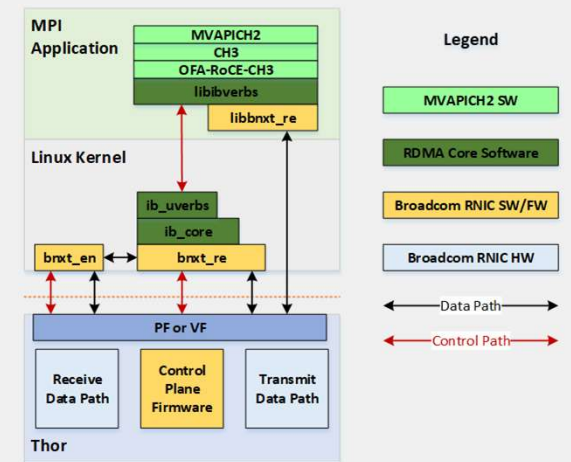
- Supports Host OS/VM/Container
- Same Driver for PF or VF
- Enables open source RDMA stack
  - Aligned with Linux kernel
  - Kernel modules are upstream
  - User libs in OFED/linux-rdma

## UCX



- Enabled by RDMA Verbs Driver
- Unmodified UCX applications

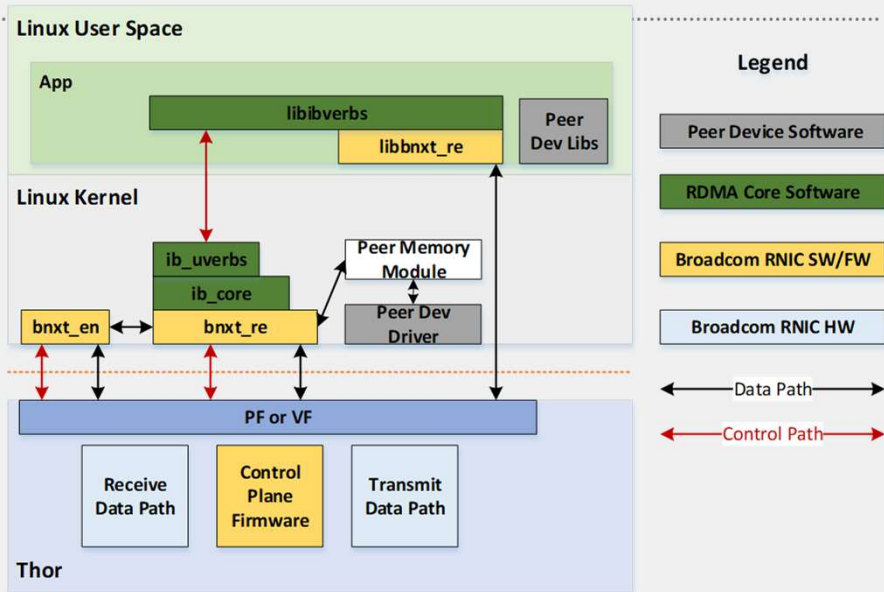
## MPI



- Enabled by Verbs Provider
- MPI/Verbs, MPI/UCX/Verbs
- Unmodified MPI applications
- Multiple MPI Implementations
  - OpenMPI, MVAPICH2, Intel MPI....

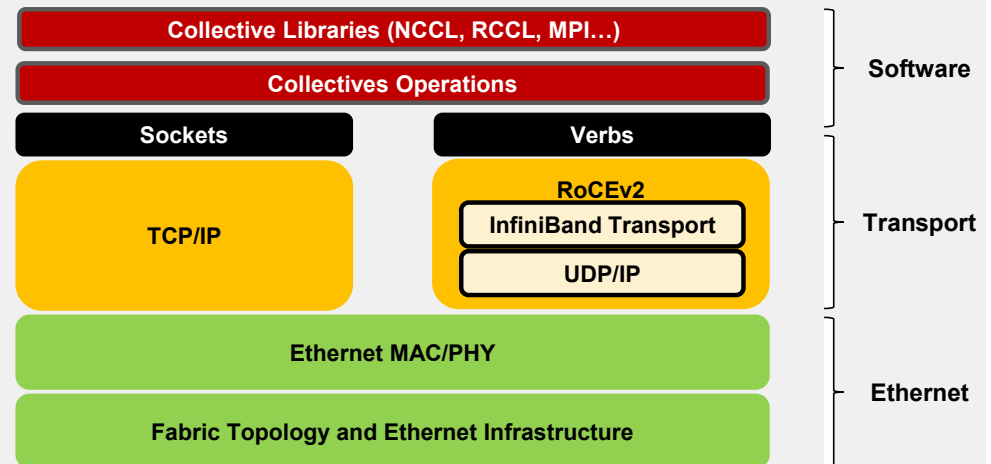
# Peer Mem Direct Infrastructure Support for HPC/AI/ML

## Baseline Peer Mem Direct



- **Peer memory model support**
- **Linux support**
  - `ib_peer_mem` (Broadcom) - supported
  - `nv_peer_mem` (NVIDIA) – supported
  - `dma-buf` (upstream) – plan to be supported

## Collectives



- **Enabled by Verbs Provider**
- **Unmodified applications**
- **Multiple Collective Libs supported**
  - NCCL, RCCL, MPI....



---

## Summary

- **Broadcom Ethernet NICs are widely deployed in HPC/AI/ML markets**
- **Ease of use, deployment, and configuration are focus areas for RDMA SW**
- **Broadcom RDMA SW is mature & supports standard infrastructures**
- **Broadcom continues to enhance RDMA software support**

Thank You





**BROADCOM**®

connecting everything®