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 Control**
- Congestion state management
- QP rate control
- ECN marking
- CNP generation
## Summary of Broadcom Ethernet NIC RDMA SW Components

<table>
<thead>
<tr>
<th>Environment</th>
<th>Software Components</th>
</tr>
</thead>
</table>
| Linux       | • User mode verbs library (libbnxt_re) – open source, upstream  
• Kernel mode driver (bnxt_re) – open source, upstream, in-box  
• Peer Memory Module – out-of-box  
• NVM configuration tool (bnxtnvm), QoS tool (bnxtqos)  
• Support for standard Linux tools including ethtool, devlink, and Ildptool  
• Installation and configuration scripts to simplify RoCE configuration and deployment |
| VMware      | • Kernel mode driver – in-box, PVRDMA support  
• NVM configuration tool (bnxtnvm)  
• Support for standard VMware tools including esxcli for QoS and DCB |
| Windows     | • User mode library (supports NDSPI)  
• Kernel mode driver (supports NDKPI) – in-box  
• All three RDMA modes supported: RDMA over PF, RDMA over vNIC, RDMA over VF  
• NVM configuration tool (bnxtnvm)  
• Support for standard Windows tools for QoS and DCB configuration |
| Firmware    | Control plane firmware for RDMA resource management (OS-independent) |

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