Advanced Transport

Table of contents

Atomic Operations

Atomic Operations in mlx5 Driver

To enable atomic operation with this endianness contradiction, use the ibv_create_qp to create the QP and set the IBV_QP_CREATE_ATOMIC_BE_REPLY flag on create_flags.

XRC - eXtended Reliable Connected Transport Service for InfiniBand

XRC allows significant savings in the number of QPs and the associated memory resources required to establish all to all process connectivity in large clusters.
It significantly improves the scalability of the solution for large clusters of multicore end-nodes by reducing the required resources.
For further details, please refer to the "Annex A14 Supplement to InfiniBand Architecture Specification Volume 1.2.1"
A new API can be used by user space applications to work with the XRC transport. The legacy API is currently supported in both binary and source modes, however it is deprecated. Thus we recommend using the new API.
The new verbs to be used are:

- ibv_open_xrcd/ibv_close_xrcd
- ibv_create_srq_ex
- ibv_get_srq_num
- ibv_create_qp_ex
- ibv_open_qp

Please use ibv_xsrq_pingpong for basic tests and code reference. For detailed information regarding the various options for these verbs, please refer to their appropriate man pages.

Dynamically Connected Transport (DCT)

Dynamically Connected transport (DCT) service is an extension to transport services to enable a higher degree of scalability while maintaining high performance for sparse traffic. Utilization of DCT reduces the total number of QPs required system wide by having Reliable type QPs dynamically connect and disconnect from any remote node. DCT connections only stay connected while they are active. This results in smaller memory footprint, less overhead to set connections and higher on-chip cache utilization and hence increased performance. DCT is supported only in mlx5 driver.

Note

Please note that ConnectX-4 supports DCT v0 and ConnectX-5 and above support DCT v1. DCTv0 and DCT v1 are not interoperable.

MPI Tag Matching and Rendezvous Offloads

Note

Supported in ConnectX®-5 and above adapter cards.

Tag Matching and Rendezvous Offloads is a technology employed by NVIDIA to offload the processing of MPI messages from the host machine onto the network card. Employing this technology enables a zero copy of MPI messages, i.e. messages are scattered directly to the user's buffer without intermediate buffering and copies. It also provides a complete rendezvous progress by NVIDIA devices. Such overlap capability enables the CPU to perform the application's computational tasks while the remote data is gathered by the adapter.
For more information Tag Matching Offload, please refer to the Understanding MPI Tag Matching and Rendezvous Offloads (ConnectX-5) Community post .