TELEDYNE LECROY
Everywhereyoulook™

Emulating RoCEv2 flows with Z800 Freya and E100 Chimera

APPLICATION NOTE

What is RoCEv2?

Introduction

What is RoCEv2?

RoCEv2 (RDMA over Converged Ethernet version 2) is a network protocol that enables remote direct memory access (RDMA) over routable IP networks using UDP encapsulation. RDMA is a high-performance protocol that allows servers to directly read from or write to the memory of another server without involving the CPU, operating system, or intermediate data copying. This is commonly used in high-speed data center networks to reduce latency and CPU overhead.

RoCEv2 header format

RoCEv2 frames are essentially InfiniBand (IB) encapsulated in UDP, IP, and Ethernet, as shown in Figure 1. The UDP Destination Port 4791 indicates that the payload is RoCEv2.

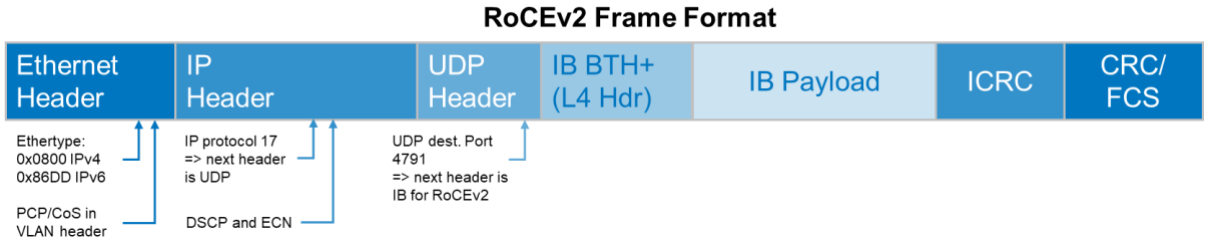

Figure 1: RoCEv2 frame format

RDMA protocol

The RDMA protocol involves the following high-level steps for the exchange of data from one memory location at one server to a memory location on another server:

1. **Connection Establishment**
   - Two devices establish a connection using RDMA-capable hardware (e.g., NICs supporting RDMA).
   - Both devices create queue pairs (send queue, receive queue and completion queue).

2. **Memory Registration**
    - The sending and receiving devices register their memory regions with the RDMA hardware.
    - Registered memory regions get pinned, meaning they cannot be moved or swapped by the OS during RDMA operations.

3. **Exchange of Memory Keys**
    - Devices exchange memory keys (e.g., Remote Memory Address and access permissions) that allow direct access to the registered memory regions.

4. **Data Transfer**
    - One device directly reads from or writes to the remote device's memory, bypassing the CPU on both ends.
    - Transfers occur with minimal latency since there is no involvement of intermediate buffers or kernel context switching.

5. **Completion Notification**
    - RDMA hardware informs the application when the operation is completed, ensuring reliability and consistency.

6. **Connection Teardown**
    - Once data exchange is complete, the RDMA connection is closed, and memory regions are deregistered.

The different types of packets exchanged between the two devices are encoded in the Base Transport Header (BTH). Since the data exchange is described in bullet 4 above, this is where performance testing becomes especially important.

Data can be exchanged as either IPv4 or IPv6 and as Reliable Connection (RC), Reliable Datagram (RD), Unreliable Connection (UC) and Unreliable Datagram (UD). Furthermore, every data exchange consists of a SEND FIRST frame, several SEND MIDDLE frames and a SEND LAST frame. This is exemplified on Figure 2 for the case of a RC SEND operation. The figure also illustrates how the BTH Opcode encodes the three types of RC SEND frames.

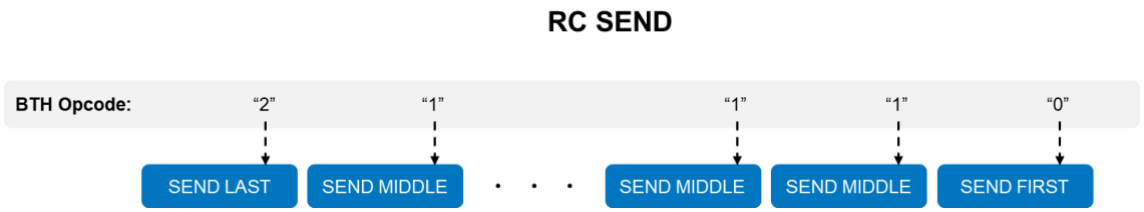

Figure 2: RC SEND data exchange.

Emulating RoCEv2 flows with Z800 Freya Traffic Generator

The Teledyne LeCroy Z800 Freya Traffic Generator (TG) can be used to generate RoCEv2 flows to enable performance testing of NICs, switches, and entire networks under various workloads. It is important to note that while the TG generates the RoCEv2 flows, it does not participate in the RDMA protocol itself. This helps to pinpoint potential issues by isolating the traffic generation from the protocol handling, allowing for more accurate identification of performance bottlenecks and network inefficiencies.

Z800 Freya can be configured to emulate RoCEv2 flows from the Xena Manager software by using one of our config files. For more advanced scripting we also offer a python code to be used with the Xena Open Automation (XOA) environment.

Configuring RoCEv2 with Xena Manager and config files

To ease the configuration of RoCEv2 flows we have created four configuration files that you can use:

- ipv6_rocev2_rc_send: IPv6, RoCEv2, Reliable Connection SEND
- ipv6_rocev2_ud_send: IPv6, RoCEv2, Unreliable Datagram SEND
- ipv4_rocev2_rc_send: IPv4, RoCEv2, Reliable Connection SEND
- ipv4_rocev2_ud_send: IPv4, RoCEv2, Unreliable Datagram SEND

The files can be loaded via the Xena Manager by right-clicking on the port and select *Load Port Configuration*, as shown on Figure 3.

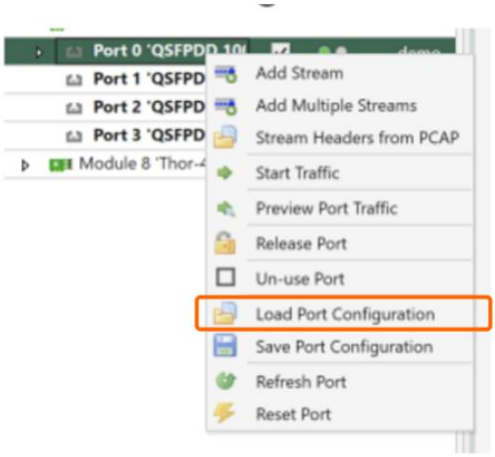

Figure 3: Loading a port configuration in Xena Manager.

Once the RoCEv2 configuration is loaded, you can see it in Port Properties as illustrated on Figure 4.

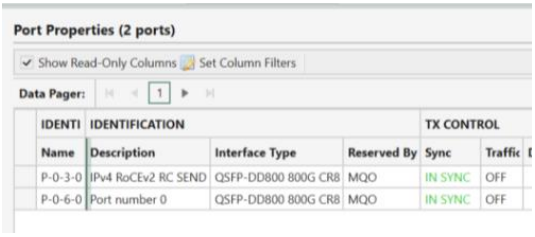

Figure 4: Port Properties showing the port is configure for IPv4 RoCEv2 RC SEND.

Configuring RoCEv2 with XOA

To set up more advanced RoCEv2 flows and potentially integrate with our own test environment, we have created an example Python script. This script provides flexible configuration options, such as adjusting the number of frames per flow, traffic rate, burstiness, repetition, and more.

The example can be found here: open-automation-script-library/rocev2 at main · xenanetworks/open-automation-script-library · GitHub

The file “headers.py” includes all the definitions of the various BTH Opcodes and other bits in the BTH. The file “roce2_emulate.py” defines the number of frames per flows, frame sizes etc.

Impairment emulation for RoCEv2 flows with E100 Chimera

The Teledyne LeCroy Xena E100 Chimera can emulate a broad range of network impairments, such as latency, jitter, packet and port impairments, flexible distributions, and bandwidth shaping. This is especially important for RoCEv2 traffic that is sensitive to latency, packet misordering and other such impairments.

We have e configuration file available for Chimera, that once uploaded will filter RoCEv2 frames based on the UDP Destination Port 4791. That makes it easy to create a test scenario where only RocEv2 frames are being impaired.

Pricing Information

For pricing details, please contact xena-sales@teledyne.com

Book a Discovery Call

Set up a quick call with a Xena tech expert to see if the Xena Ethernet Test Platform is the right solution for your needs: https://xenanetworks.com/discovery-call-booking/