CASE STUDY

REDIGO Readout over High-Performance Networking at National Laboratory of Legnaro (LNL)

Performance test on ConnectX®-2 EN 10 Gigabit Ethernet Adapters with RoCE

Introduction

Fast data transmission technologies - in particular optical - are being deployed on an unprecedented scale in the Large Hadron Collider (LHC) physics experiments era [1990-2009]. LHC is a collider dedicated to physics research, and it has been working from 2010 at the European Organization for Nuclear Research (CERN) in Switzerland. Four experiments – A Large Ion Collider Experiment (ALICE), A Toroidal LHC ApparatuS (ATLAS), Compact Muon Solenoid (CMS), LHCb (bottom quark) - for particle and high energy physics are installed on the LHC ring and they are in a data collection phase.

The machine and the experiments use optical links:

- To connect the detector front ends, instead of doing both data readout and detector control;
- To transmit data between cards, crates and racks in the counting rooms.

Future experiments require a new generation of data acquisition (DAQ) systems. The use of standard networking protocols (TCP/IP, RDMA, RoCE, iWARP) and hardware (network adapter cards from commercial companies) ensures the compatibility between different components of the detectors and, in the meantime, it allows for seamless incremental upgrades to individual sub systems.

Some new experiments, which are in study and project phases, will take data without the use of a hardware trigger. Consequently, to minimize down time, all experimental data must be read from the detector in the 200 ms gaps between beam bunch trains. This semi-continuous data stream must be routed to DAQ computers, currently assumed to be PCs, processed and then sent to offline storage.

It is highly desirable to use commercial networking devices and protocols as much as possible. One option would be the use of 10 Gigabit Ethernet.

Without some form of traffic shaping between the concentrators/producers and the destination/consumers PCs, there would be the classic bottleneck problem on the egress of the Ethernet switch with data queuing for transmission to the processing node, and the possibility of packet loss. The growing convergence of storage protocols (RoCE, iWARP, RDMA over TCP) around 10 Gigabit Ethernet standard makes it attractive for deployment in new data acquisition systems for a number of reasons.

IEEE has been developing standards they collectively refer to as "Data Center Bridging" (DCB) and that are also sometimes referred to as "Converged Enhanced Ethernet" (CEE).

Istituto Nazionale di Fisica Nucleare
Laboratori Nazionali di Legnaro

OVERVIEW

The REDIGO (an Italian acronym for Readout at 10 Gbits/s) project at INFN National Laboratory of Legnaro (LNL) involved use of 10 Gigabit Ethernet commercial networking devices and protocols to assist with physics experiments and optimize performance of data transmission links for a Large Hadron Collider (LHC). Using Mellanox technology, REDIGO and LNL conducted four experiments for particle and high energy physics. They used ConnectX-2 EN with RoCE adapters to optimize performance and satisfy the electronic requirements of the REDIGO project proposal.

The main new features are:

- Priority-Based Flow Control (802.1Qbb), sometimes called "per-priority pause"
- Enhanced Transmission Selection (802.1Qaz);
- Congestion Notification (802.1Qau).

This paper shows the Mellanox ConnectX-2 EN cards solution with RDMA over Converged Ethernet (RoCE).

Key Results Summary

- Hardware installation, driver mlx4_0, firmware are user friendly.
- Latency is about 1.6 µs for buffer size < 48 bytes.
- Throughput is about 9.0 Gbits/s for buffer size > 4096 bytes

Test Goals

The goal is the performance study of Mellanox's ConnectX-2 EN adapter (OPN = MNPH29C-XTR | PSID = MT_0DB0120010)

The test objectives are:

- Latency measurement in back-to-back configuration;
- Latency measurement with switch between cards;
- Bandwidth measurement (both configurations).

System Apparatus

Machines:

- Vendor Model: IBM x3455 Type 7986 (dual core)
- Processors: Dual-Core AMD Opteron Processor 2214 2200 MHz
- Cache: 2 MB Level 2 cache
- Front Side Bus Speed: 1000 MHz
- Memory: 4 GB

10GigE Adapter:

- Mellanox ConnectX-2, PSID: MT_0DB0120010
- ConnectX-2 EN with RoCE, PCIe 2.0 5.0 GT/s
- Driver: mlx4_0 from OFED 1.5.2-rxe

Software:

- Operating System: Scientific Linux SL release 5.4
- Version: 2.6.18-164.2.1.el5 x86_64

Network Elements:

- Switch: Extreme Network Summit x650-24x

Network Accessories:

- Connectors: SFP+
- Cables: copper cables full duplex (3 m long)

Latency Test Description

The test setup consists of two machines (redigo-01 and redigo-02) connected back-to-back via cables (configuration A) or through a Summit x650-24x switch (configuration B). The procedure is a simple reflector test. Machine redigo-02 (the client) sends messages to machine redigo-01 (the server). The messages are immediately sent back to machine redigo-02. The programs use RDMA verbs with RoCE.

The round trip time (RTT) is the period of a complete cycle. The latency is calculated as half of the round trip time.

All latency tests run at 104 cycles. For each fixed buffer size of values between 1 bytes to 8192 bytes (x axis of graph in Figure 1), latency mean value is calculated and reported on y axis of graph in Figure 1. The designed upper value of switch latency is 2 µs, so this is used as true value for calculations.

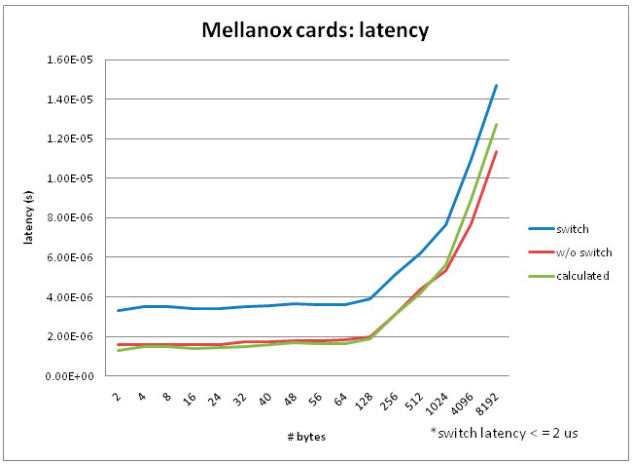

Figure 1: Latency test performance graph with Mellanox cards. Blue line "switch" shows results in configuration B and red line "w/o switch" shows data in configuration A. Green line "calculated" values are found by subtraction configuration B values minus switch latency (nominally 2 µs).

Key Point

The latency value remains consistently low for the tested solution. It is under 2 µs until buffer size is under 64 bytes; it grows up to 12 µs for 8192 bytes size.

Throughput Test Description

The test setup consists of two machines (redigo-01 and redigo-02) connected back-to-back via cables (configuration A) or through a Summit x650-24x switch (configuration B). The procedure is via simple reflector tests. Machine redigo-02 (client) sends messages to machine redigo-01 (server). The messages are immediately sent back to machine redigo-02. The programs use RDMA verbs with RoCE.

The throughput, transfer speed or bandwidth is the number of transferred bytes in one operation time.

Throughput tests run from 107 times for buffer size < 4096 bytes to 103 times for the biggest buffer (48000000 bytes). For each fixed buffer size values between 1 bytes to 48000000 bytes (x axis of graph in Figure 2), throughput value is calculated and reported on y axis of graph in Figure 2.

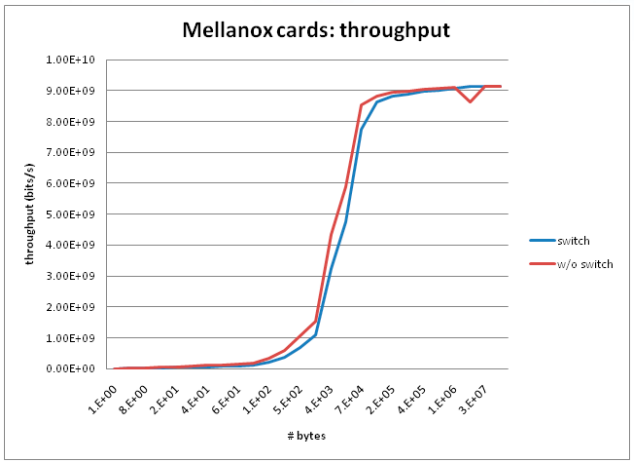

Figure 2: Throughput test performance graph with Mellanox cards. Blue points "switch" show results in configuration B and red squares "w/o switch" show data in configuration A.

Key Point

The tested solution can deliver close to a full 10GigE of network throughput across different message sizes (buffer size > 65536 bytes).

Test Configuration

Figure 3 shows test configurations:

- A where cards are connected back to back without switch;
- B where switch is present.

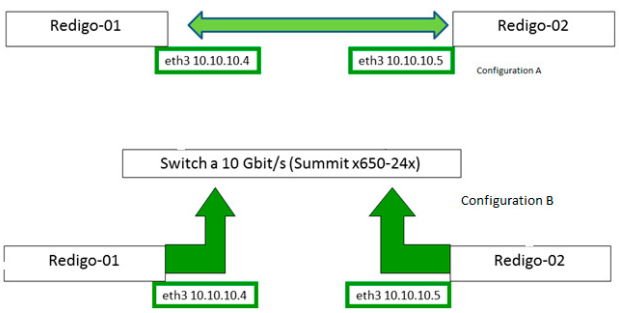

Figure 4 is a photo of REDIGO test apparatus at INFN LNL.

Summary

The results reported here clearly show the significant performance of a 10Gigabit Ethernet solution based on Mellanox ConnectX-2 EN with RoCE adapters.

For applications that require high bandwidth (for example, data storage) Mellanox connectivity solution delivers about 90% utilization of available network bandwidth.

The ability to forward traffic with consistently low latency (around 1.6 µs) increases is important in data acquisition setup of physical experiments, where multiple flows of data need to be delivered effectively to multiple subscribers to maximize application performance.

Mellanox ConnectX-2 EN with RoCE cards offers good performance to satisfy the electronics requirements of REDIGO project proposal.

Acknowledgments

The author wishes to express his gratitude:

- To his supervisor, Dr. Gaetano Maron;
- To REDIGO national reference, Dr. Marco Angelo Bellato;
- To REDIGO local reference in LNL, Dr. Michele Gulmini.

They offered me invaluable assistance, support and guidance.

Deepest gratitude are also due to the members of Technologies Informatics and Electronics Service (STIE) group, Dr. Simone Badoer , Dr. Massimo Biasotto, Dr. Damiano Bortolato, Dr. Sergio Fantinel, Dr. Pietro Molini, Dr. Nicola Toniolo, without whose knowledge and assistance this study would not have been successful.

More about REDIGO

REDIGO Readout at 10 Gbits/s "Development and study of 10 Gbits/s network for readout in physics experiments" is a proposal project approved in 2009 by Italian National Institute of Nuclear Physics (INFN) commission V, dedicated to electronics and software. It covers three years (2010, 2011, 2012) and it involves people from INFN Padua, INFN National Laboratory of Legnaro (LNL) and INFN Bologna.

For more information, please contact dott. Andrea Gozzelino

INFN - Laboratori Nazionali di Legnaro (LNL)

Viale dell'Università, 2 - I - 35020 - Legnaro (PD) - ITALIA

Office at LNL: E-101

Tel: +39 049 8068346

Fax: +39 049 641925

Mail: andrea.gozzelino@lnl.infn.it

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com