Developments regarding the integration of FPGA RDMA into the ATLAS Readout with FELIX in High Luminosity LHC

Matei-Eugen Vasile, Sorin Martoiu, Nayib Boukadida, Gabriel Stoicea, Petru Micu, Alexandru Dumitru, Andrei-Alexandru Ulmamei, Radu Hobincu, Cristina-Cerasela Iordache

***Abstract*—RDMA (Remote Direct Memory Access) is used by the ATLAS experiment at CERN in the new readout system based on FELIX (Front-End Link eXchange) for its networking layer. The FELIX system is used to interface the front-end electronics to commodity hardware in the server farm. In the current implementation of FELIX, RDMA communication is implemented using software on both ends of the RDMA links. FELIX is using RDMA through RoCE (RDMA over Converged Ethernet) to transmit data from its servers to the Software Readout Driver devices in the server farm using off-the-shelf networking equipment. As a consequence of the High Luminosity LHC upgrade, improvements in the data throughput will be needed. These improvements can be achieved by implementing RDMA support in the FELIX FPGA to simplify the path the data is taking through the readout system. This FPGA implementation of the RDMA protocol has been developed and tested. Now, a version of FELIX that uses this implementation is being proposed and demonstrated.**

I. Introduction

REmote Direct Memory Access (RDMA) is a network protocol designed to transfer data between remote devices with as little CPU involvement as possible, and can be implemented on top of various networking technologies such as InfiniBand [1] or TCP/IP [2]. In large-scale detector systems where relatively large quantities of data from multiple source nodes need to be distributed to multiple sink nodes, RDMA may add a significant performance advantage.

In the ATLAS experiment at CERN [3], the FELIX (Front-End Link eXchange) [4] system is used for interfacing the front-end detector electronics to the readout system and the high-level trigger farm. The system is based on a custom FPGA board which receives data from the front-end detector electronics via a large number of optical links and outputs data via a PCIe interface to a host computer which manages processing and relaying the data further to the readout system over a high bandwidth output link. The FELIX host computer uses the RDMA (Remote Direct Memory Access) support offered by network interface cards with RoCE (RDMA over Converged Ethernet) to transmit data further toward the readout systems over an Ethernet network. A possible improvement for enhancing the data throughput, as part of the High Luminosity LHC upgrade [5] [6], is the implementation of RDMA support in the FELIX FPGA itself, so that the data path from front-end detector electronics to the readout system bypasses the PCIe interface and the host computer, flowing directly to the Ethernet network.

An advantage of this approach is that the data-path is considerably shortened and potential bottlenecks related to the PCIe transfer or host operating system may be avoided. Data encapsulation and de-encapsulation required by the PCIe transfer to host memory is also avoided. These potential advantages come with the expense of increased complexity in the FPGA logic.

II. Background

The proposed FPGA implementation of the RDMA protocol was developed and tested [7], the results of the performance testing showing that it can fully utilize the bandwidth of an RDMA link over 100Gbps Ethernet network. The basis for the RDMA is an open source RDMA HLS core [8] which has been modified and expanded to make it work with existing off-the-shelf networking equipment and allow us to add our new features. The existing core had to be modified to allow for dynamic establishment of RDMA links and to improve the handling of packet retransmits so that high transfer bandwiths are reachable.

For this work, the platform used was a Xilinx VCU128 development board [9], with a Xilinx Virtex Ultrascale FPGA, which is one of the candidate FPGA platforms for the FELIX upgrade. The VCU128 board has both DDR external memory and HBM on-chip memory.

Manuscript submitted on: December 9, 2022

Matei-Eugen Vasile is with Institutul National de Cercetare-Dezvoltare pentru Fizică si Inginerie Nucleară Horia Hulubei (IFIN-HH), Măgurele, România (e-mail: matei.vasile@cern.ch)

Sorin Martoiu is with Institutul National de Cercetare-Dezvoltare pentru Fizică si Inginerie Nucleară Horia Hulubei (IFIN-HH), Măgurele, România (e-mail: sorin.martoiu@cern.ch)

Nayib Boukadida is with the Nikhef National institute for subatomic physics, Amsterdam, The Netherlands (e-mail: nayib.boukhadida@cern.ch)

Gabriel Stoicea is with Institutul National de Cercetare-Dezvoltare pentru Fizică si Inginerie Nucleară Horia Hulubei (IFIN-HH), Măgurele, România (e-mail: gabriel.stoicea@cern.ch)

Petru Micu is with Universitatea Politehnica București, București, România (e-mail: petru.micu@cern.ch)

Alexandru Dumitru is with Universitatea Politehnica București, București, România (e-mail: alexandru.dumitru@cern.ch)

Andrei-Alexandru Ulmamei is with Universitatea Politehnica București, București, România (e-mail: andrei-alexandru.ulmamei@cern.ch)

Radu Hobincu is with Universitatea Politehnica București, București, România (e-mail: radu.hobincu@cern.ch)

Cristina-Cerasela Iordache is with Universitatea Politehnica București, București, România (e-mail: cristina-cerasela.iordache@cern.ch)

978-1-6654-8872-3/22/$31.00 ©2022 IEEE

After this initial proof-of-concept demonstrator setup has been built, the next step was to begin working on integrating the FPGA RDMA implementation of the RDMA protocol in the FELIX system [10], thus offering a working alternative for the data path to the one currently implemented in FELIX.

During this second phase of developing the FPGA RDMA implementation, the Xilinx Alveo platform was also used. Xilinx Alveo boards are FPGA accelerator platforms with PCIe interfaces and multiple 100Gbps network interfaces. More precisely Alveo U250 boards [11] have been used, alongside the Xilinx Runtime Library (XRT) [12], in order to simplify the implementation of the communication between the host PC, the FPGA board and the hardware design running in the FPGA.

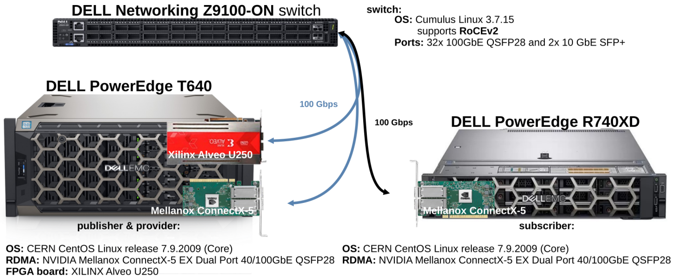
III. Implementation

The new features of the FPGA RDMA implementation are the support for multiple simultaneous RDMA connections. Without going into the specifics of the RDMA technology, what we refer to here as an RDMA connection, in more precise RDMA terminology, is known as an RDMA Queue Pair. An RDMA Queue Pair is similar to a TCP/IP socket.

The support for multiple simultaneous RDMA connections is desirable because we are investigating whether or not it is feasible to use a separate RDMA connection for each FELIX elink.

A. Design and Setup

The current overall design of the implementation, as well as the development setup, as shown in Fig. 1, is the same as described in [10].



DELL Networking Z9100-ON switch
switch:
OS: Cumulus Linux 3.7.15
supports RoCEv2
Ports: 32x 100GbE QSFP28 and 2x 10 GbE SFP+
DELL PowerEdge T640
100 Gbps
100 Gbps
DELL PowerEdge R740XD
Xilinx Alveo U250
Mellanox ConnectX-5
Mellanox ConnectX-5
publisher & provider:
subscriber:
OS: CERN CentOS Linux release 7.9.2009 (Core)
RDMA: NVIDIA Mellanox ConnectX-5 EX Dual Port 40/100GbE QSFP28
FPGA board: XILINX Alveo U250
OS: CERN CentOS Linux release 7.9.2009 (Core)
RDMA: NVIDIA Mellanox ConnectX-5 EX Dual Port 40/100GbE QSFP28

Fig. 1. The FPGA RDMA development setup.

B. Firmware and Software

In order for the multiple simultaneous RDMA connection support to be implemented, changes had to be implemented both in the FPGA core, and in the software framework. The foundation for this upgrade were already laid by the design choices made previously and described in [10].

In short, RDMA is used for the data layer. In the control layer, meaning the subscription management and the mapping of logical links between RDMA-capable endpoints, changes needed to be made so that the FELIX software no longer handles the flow of data between network enpoints but only instructs the RDMA firmware in the FPGA and the data receiver software on the receiver on how to establish the RDMA connections and to send the data over them, once they have been established. This mechanism has been described at more length in [10].

The novelty consists in the actual implementation of the support for creating multiple simultaneous RDMA connections between a provider device, which is the FPGA device on which the RDMA core is running, and one or more subscriber devices, which are the devices receiving the data from the provider.

The current implementation, on the provider side, needs to wait for all subscribers to connect and, only when all expected subscribers have connected, the data will begin to be sent to those subscribers. Each connection is a separate logical link which can carry different data, if needed. Implementing this has been simplified by the design choice of making the provider software that controls the provider firmware run in a separate process. Once started, this process communicates with the FELIX software through a TCP/IP socket connection, and making it work with multiple subscribers was a matter for having the provider software wait for several subscriber connection requests instead of a single one. The number of expected subscribers is entered as a parameter, when the entire setup is launched.

The current FPGA RDMA core implementation supports only four simultaneous connections.

IV. Results and Conclusions

A. Test Setup and Results

The test setup is similar to what was previously described in [7] and [10]. In short, the hardware and network setup is that shown in Fig. 1. Two PCs are used: one for the subscriber/receiver and one for the publisher. The publisher is also the host of the FPGA device on which the FPGA RDMA core is running. Both PCs and the FPGA device are connected to a switch with RDMA support. The publisher software waits for the tested number of connections to be requested by one or more subscribers. Once the requisite number of connections have been established between the provider FPGA device and the receiver(s) on the connected subscriber(s), the FPGA RDMA core begins sending data to the receiver(s). The bandwidth measurement is done on the FPGA device. The amount of data to be sent is configured as a parameter of the testing software. The time it takes for all the data to be sent is measured, for each connection, by saving two timestamps and simply subtracting the first from the second. The first timestamp is the time when the data starts to be sent on the respective connection. The second timestamp is the time when the acknowledgment for the last sent message is received on the FPGA RDMA core.

The presented results will be for message sizes ranging from 128B to 1MB, sent in bursts of 10 messages at a time.

For each test, two references have been used:

1) The bandwidth reported by the *ib_write_bw* test from the *Perftest* package (the red dotted line in Fig. 2 3 and 4)
2) The bandwidth measured running the test with a single subscriber and a single connection for that subscriber (the solid blue line in Fig. 2 3 and 4)

For multiple connections, the following setups have been tested:

- two connections distributed as:
  - both connections on the same subscriber
  - two subscribers, each with a single connections
- four connections distributed as:
  - all four connections on the same subscriber
  - two subscribers, one with three connections, one with one connection
  - two subscribers, each with two connections

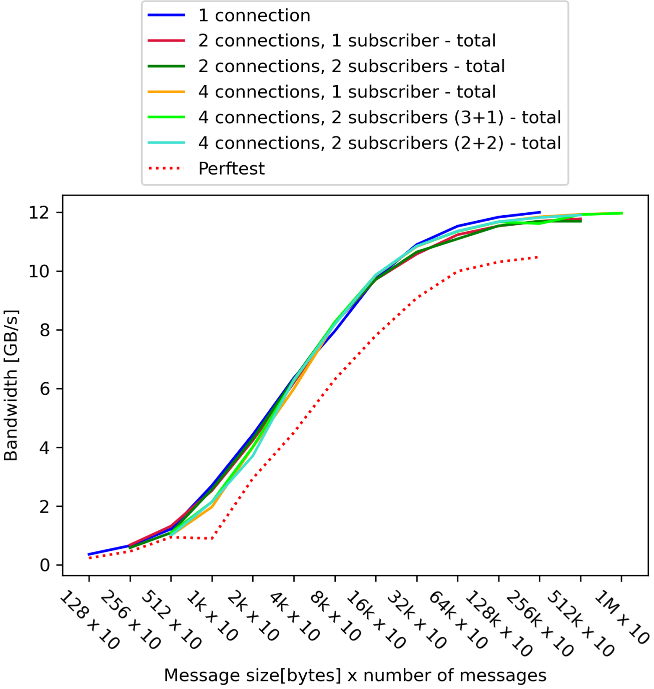


Fig. 2. Totals, from all the test configurations.

At link saturation, all multiple connection setups exhibit an overhead compared to the single connection setup. The largest overhead penalty measured is 3.19% compared to the single connection reference measurement.

B. Future Work

The most obvious future work is increasing the supported number of simultaneous connections. Using a connection for each FELIX elink is the goal here. In order for that to be feasible, support for hundreds of simultaneous connections will be needed. A less obvious, but not less relevant improvement feature, will be to implement support for subscribers connecting and disconnecting at any time instead of requring to have all subscribers connected first and only then starting the provider firmware to serve those subscribers.

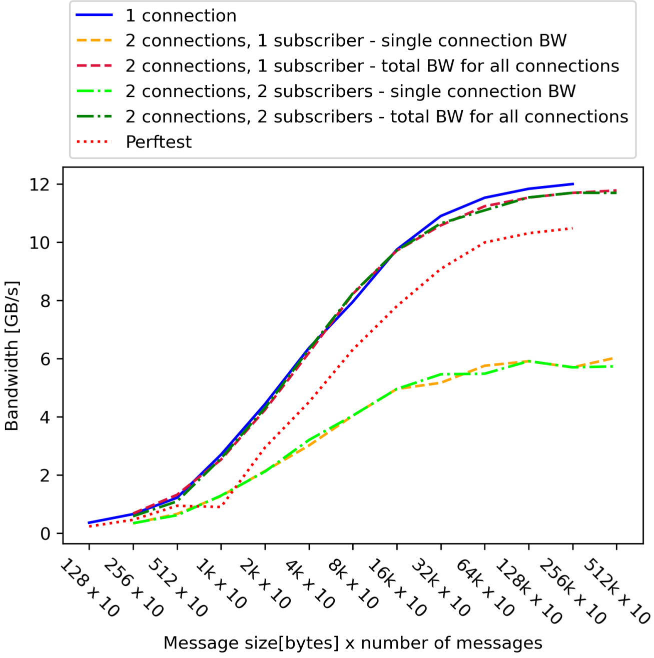


Fig. 3. 2 connection test configuration results.

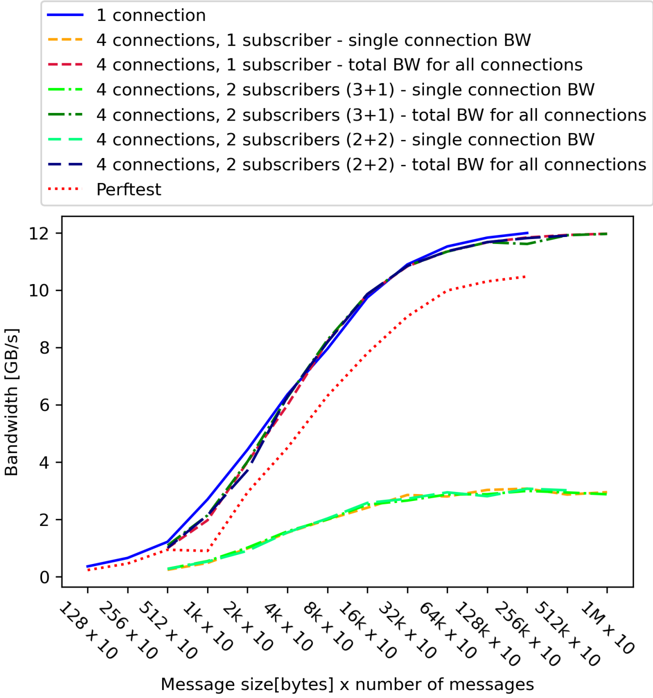


Fig. 4. 4 connections test configuration results.

References

[1] InfiniBand Trade Association, "InfiniBand™ Architecture Specification". [Online]. Available: https://www.infinibandta.org/ibta-specification/. Accessed on: November 25, 2022.

[2] RDMA Consortium, "Architectural Specifications for RDMA over TCP/IP". [Online]. Available: http://www.rdmaconsortium.org/. Accessed on: November 25, 2022.

[3] ATLAS Collaboration, "The ATLAS Experiment at the CERN Large Hadron Collider," *JINST*, vol. 3, pp. S08003, 2008, doi: 10.1088/1748-0221/3/08/S08003.

[4] C. A. Gottardo, "FELIX and SW ROD Commissioning of the New ATLAS Readout System," 2020 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), 2020, pp. 1-5, doi: 10.1109/NSS/MIC42677.2020.9507984.

[5] L. Evans and P. Bryant, “LHC Machine,” *JINST*, vol. 3, pp. S08001, 2008, doi: 10.1088/1748-0221/3/08/S08001.

[6] G. Apollinari, I. Béjar Alonso, O. Brüning, M. Lamont, and L. Rossi, “High-Luminosity Large Hadron Collider (HL-LHC): Preliminary Design Report,” CERN, Geneva, Switzerland, 2015, doi: 10.5170/CERN-2015-005.

[7] M. E. Vasile, N. Boukadida, S. Martoiu, M. Antonescu, A. A. Ulmamei, G. Stoicea, R. Hobincu, and C. C. Iordache , “FPGA implementation of RDMA for ATLAS Readout with FELIX in High Luminosity LHC,” *JINST*, vol. 17, pp. C05022, 2022, doi: 10.1088/1748-0221/17/05/C05022.

[8] D. Sidler, Z. István and G. Alonso, “Low-latency TCP/IP stack for data center applications,” 2016 26th International Conference on Field Programmable Logic and Applications (FPL), 2016, pp. 1-4, doi: 10.1109/FPL.2016.7577319.

[9] Xilinx Inc., “VCU128 FPGA Evaluation Kit”. [Online]. Available: https://www.xilinx.com/products/boards-and-kits/alveo/u250.html. Accessed on: November 25, 2022.

[10] M. E. Vasile, S. Martoiu, N. Boukadida, G. Stoicea, P. Micu, A. Dumitru, A. A. Ulmamei, R. Hobincu, and C. C. Iordache , ”Integration of FPGA RDMA into the ATLAS Readout with FELIX in High Luminosity LHC” presented at the TWEPP 2022 Topical Workshop on Electronics for Particle Physics, Bergen, Norway , Sep. 19-23, 2022.

[11] Xilinx Inc., “Alveo U250 Data Center Accelerator Card”. [Online]. Available: https://www.xilinx.com/products/boards-and-kits/alveo/u250.html. Accessed on: November 25, 2022.

[12] Xilinx Inc., “Xilinx Runtime Library (XRT)”. [Online]. Available: https://www.xilinx.com/products/design-tools/vitis/xrt.html. Accessed on: November 25, 2022.