Accelerated Wire-Speed Packet Capture at 200 Gbps

Lukáš Kekely, Martin Špinler, Štepán Friedl, Jiří Sikora, Jan Kořenek
CESNET a. l. e.
Zikova 4, 160 00 Prague, Czech Republic
Email: kekely,spinler,friedl,jiri.sikora,korenek@cesnet.cz

***Abstract*—We present our latest FPGA acceleration card NFB-200G2QL that is specifically designed to enable traffic processing at 200 Gbps. Unique high-speed DMA engines in the FPGA together with highly optimized Linux drivers enable data transfer through PCIe interfaces with minimal CPU overhead. Captured traffic can be independently distributed between individual cores of two physical CPUs (NUMA nodes) without utilization of QPI. As a result, wire-speed packet capture to the host memory from two fully saturated 100 Gbps Ethernet interfaces (QSFP28+ cages) is achieved and various network monitoring applications can utilize the power of the latest FPGAs and CPUs for data processing. This is especially useful when both directions of a single 100GbE link are monitored.**

The live demonstration shows how the packets are received from two 100 Gbps Ethernet links at wire-speed and captured to the host memory at 200 Gbps without a loss. The opposite direction of communication is also shown, i.e. how the packets are transmitted from the host memory and fully saturate the two 100GbE network interfaces. Achieved speeds are demonstrated by counters and gauges showing generated, received/transmitted and captured packets. We also show statistics of CPU load during the packet capture/transmission for different packet lengths.

Fig. 1. The FPGA based NFB-200G2QL acceleration card.

I. Technology Overview

100 Gigabit Ethernet (100GbE) was defined by the IEEE 802.3ba standard [1] in 2010 and currently is the fastest deployed Ethernet for computer networks. It enables for transmitting frames at a rate of 100 Gbps, which translates up to nearly 150 millions frames per second. Please note that packet rate this high means that a new frame is transferred every 6.7 ns. The 100GbE standard encompasses a number of different physical layer specifications, most notably 100GBASE-LR4 and 100GBASE-SR4 both working with four lanes (wavelengths of light) at 25 Gbps in a single-mode or multi-mode fiber.

The low-profile NFB-200G2QL card is shown in Figure 1 and commercially available from our partner [2]. It is world's first PCI Express adapter equipped with two 100GbE ports that is designed to enable wire-speed processing of traffic at full speed of both ports. This hardware-accelerated card with FPGA uses unique high-speed DMA modules that enable to achieve 200 Gbps throughput of data transfers over PCI Express between the card and memory of the host computer. This high throughput makes the card ideal for deployment in the fastest backbone networks and in high-throughput data centers. Main unique features of the NFB-200G2QL include:

- two 100GbE QSFP28+ transceiver interfaces (cages) that also supports 4×50G, 2×40G, 8×25G or 8×10G modes
- powerful Virtex UltraScale+ VU7P FPGA chip,
- three static QDR-IIIe memories (max. 288 Mb each),
- two PCI Express Gen3 interfaces with 16 lanes each,
- PCI Express half-length and low-profile form factor,
- power consumption of less than 65 W,
- passive cooling with NACA/NASA-shaped air scoop [3],
- external PPS input for precise timestamps.

The Xilinx Virtex UltraScale+ VU7P FPGA chip [4] is the heart of the card. Compared to other computing devices, such as fixed ASICs or programmable CPUs, FPGAs allow changing their internal structure by programming their firmware. A typical arrangement of the FPGA firmware for high-speed applications is a pipelined processing, which takes advantage of FPGA's inherent massive parallelism to achieve the required throughput and performance.

As a base of our FPGA firmware, we have developed a platform for rapid development of hardware-accelerated applications. The platform includes a set of firmware IP cores, especially blocks for network interfaces (from 1GbE up to 100GbE) and a unique high-performance programmable DMA bus-master connection to the software layer via PCIe bus. The software layer consists of Linux device drivers, tools for card management, and libraries for high-speed data transfers between the card and the host memory (DPDK or proprietary SZE2). The framework also specifies a generic interface to optional traffic processing pipeline in FPGA that

can be described using P4 language [5] to perform different operations on passing network data [6]. Our HaNIC solution is an example of such traffic processing in the FPGA firmware. It extends the functionality of a basic NIC by the support of packet parsing, filtering, and configurable hash-based distribution among multiple CPU cores.

Two PCI Express endpoints are needed as a workaround of the missing $\times 32$ PCIe endpoint support in current FPGAs, motherboards, and CPUs. This is required because the effective throughput of PCIe Gen3 $\times 16$ is only slightly above 100 Gbps. Therefore, our card utilizes two $\times 16$ slot in order to achieve required 200 Gbps throughput into the host memory. Furthermore, using two PCIe endpoints enable direct data transfers between the card and two physical CPUs (NUMA nodes) without the QPI bottleneck.

II. Demo Description

The goal of the proposed demo is to present the unique performance of the low-profile NFB-200G2QL card. A similar (simpler) demonstration has already been performed at [7]. We want to especially stress out the ability of the card to:

- operate both 100 GbE interfaces at wire-speed,
- transfer all received data via PCIe into the host memory at full 200 Gbps regardless of the frame length,
- transfer data from the host memory via PCIe at full 200 Gbps regardless of the frame length.

Illustration of the demo architecture can be seen in the Figure 2. The NFB-200G2QL card is connected into PCIe slots of standard server motherboard with two relatively fast multicore CPUs and fully filled memory banks (for maximal memory throughput). Inside the card's FPGA there is our HaNIC firmware configured to capture all of the incoming traffic and distribute it among available CPU cores. Both Ethernet ports of the card are connected to a tester device that can generate and receive (analyze) 100GbE network traffic. Since conventional hardware testers supporting 100 GbE ports (e.g. Spirent TestCenter) are too large and heavy to transport, we can instead implement the required traffic generation and capture capabilities inside our FPGA firmware and connect the optic cables in a loopback. This also shows the versatility of the FPGA firmware. Described demo architecture can operate in two basic modes: packet capture and packet replay.

In packet capture mode, packets of configurable length are generated at the maximum allowed rate and sent over the fiber into both 100 Gbps Ethernet ports. There, the packets are received by on-card PMA, PCS and MAC engines, distributed into multiple DMA channels and transferred via 2 PCIe endpoints at 200 Gbps into the ring buffers inside server's main memory. In the memory, the packets are accessed and counted. Processing has the form of only a simple accounting because we want to demonstrate that the card is capable of delivering the 200 Gbps of data into the software and not the performance of some specific advanced packet processing in the CPUs. Finally, live packet capture performance statistics are shown in the GUI on the screen. This mode corresponds to typical

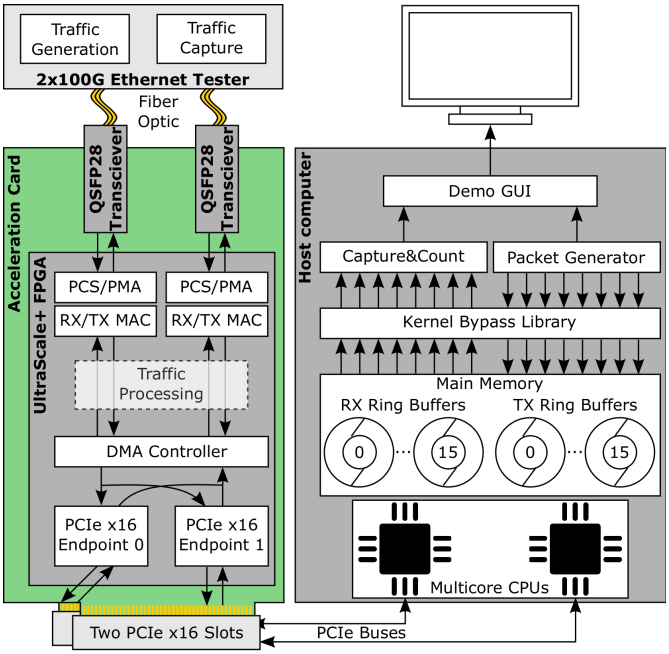

Fig. 2. Demo system architecture illustration.

network monitoring or security scenarios, where traffic of both directions of a tapped 100GbE link needs to be processed.

In packet replay mode, packets of configurable length are prepared by CPUs in the host memory and copied into multiple DMA ring buffers. From there, they are picked up by the DMA controllers in the card's FPGA and transferred via PCIe into its local memory. Then, they are transferred using standard Ethernet layers onto two optical 100 GbE lanes. Finally, live sending performance statistics are shown in the GUI on the screen. This mode corresponds to data center deployment, where large amounts of data need to be transferred.

Acknowledgment

The presented work is supported by the project Reg. No. CZ.02.1.01/0.0/0.0/16_013/0001797, and by the Technology Agency of the Czech Republic project TH02010214.

References

[1] IEEE Computer Society, "Amendment 4: Media access control parameters, physical layers, and management parameters for 40 Gb/s and 100 Gb/s operation," *IEEE Standard 802.3ba-2010*, pp. 1–457, June 2010.

[2] Netcope Technologies, "NFB-200G2QL FPGA-based hardware," *White Paper: Product Brief*, January 2018. [Online]. Available: https://www.netcope.com/en/resources/nfb-200g2ql-product-brief

[3] Charles W. Frick et al., "An experimental investigation of NACA submerged-duct entrances," *NACA ACR No. 5120*, November 1945.

[4] Xilinx, "UltraScale architecture and product data sheet: Overview," *Preliminary Product Specification DS890 (v3.2)*, pp. 1–44, January 2018.

[5] Pat Bosshart et al., "P4: Programming protocol-independent packet processors," *SIGCOMM Computer Communication Review*, vol. 44, no. 3, pp. 87–95, Jul. 2014.

[6] P. Benáček, V. Puš, H. Kubátová, and T. Čejka, "P4-To-VHDL: Automatic generation of high-speed input and output network blocks," *Microprocessors and Microsystems*, vol. 56, pp. 22–33, 2018.

[7] L. Kekely, M. Špinler, Š. Friedl, J. Sikora, and J. Kořenek, "Live demonstration of FPGA based networking accelerator for 200 Gbps data transfers," *The 30th IEEE/IFIP Network Operations and Management Symposium (NOMS)*, April 2018.