THE UNIVERSITY of NEW MEXICO | SCHOOL of ENGINEERING

Department of Electrical & Computer Engineering

Final Technical Progress Report

DOE Grant #:	DE-SC0002113
Project Title:	STIMULUS: End-System Network Interface Controller for 100 Gb/s Wide Area Networks
Principal Investigator:	Prof. Payman Zarkesh-Ha
Institution:	University of New Mexico
Project Collaborator:	Dr. Jesse Wen, Acadia Optronics LLC
Reporting Period:	15 September 2009 to 14 September 2013

1. Synopsis of the Program

Because of the excessive network bandwidth requirements in recent years, there is an urgent need for the development of data exchange rate at 100 Gb/s and beyond. Internet exchanges, high-performance computing data for collaborating research efforts, as well as personalized content such as in YouTube, IPTV, and HDTV. Although blazing-fast data exchange is one of the primary objectives of the Department of Energy for collaborative scientific applications, many commercial companies such as Amazon, Google and Yahoo have also expressed their immediate requirement of 100 Gb/s network infrastructure. Applications such as high-energy physics, climate forecasts, genomics data and computation and others all demand increasingly high aggregate bandwidths. Large supercomputers such as Roadrunner (Los Alamos National Lab - LANL) and Jaguar (Oak Ridge National Laboratory - ORNL) are expected to have tens of thousands of transceivers, each providing aggregate point-to-point interconnect bandwidths of 100 Gb/s or more.

The main goal of this research grant is to develop a system-level solution leveraging novel technologies that enable network communications at 100 Gb/s or beyond. University of New Mexico in collaboration with Acadia Optronics LLC has been working on this project to develop the 100 Gb/s Network Interface Controller (NIC) under this Department of Energy (DOE) grant. This final report summarizes the research performed during September 2009 to September 2013, which includes the no-cost extension for year 4.

2. *Research Activities Performed*

The overall objective for this project in the four-year period consists of the following two major milestones; 1) Development of the 40 Gb/s NIC and 2) Development of the 100 Gb/s NIC. The key objectives for each milestone of this project are listed in the outline below, where the focus of our research activities in the final year was to complete the development of 100G Ethernet Media Access Control (MAC) and Physical Coding Sub-layer (PCS) IP cores for the 100 Gb/s NICs.

40 Gb/s NIC (1st Milestone)

- Board design
  - Schematic capture
  - Layout
  - Fabrication and assembly
- 40 G framework
  - DMA core
  - Memory controller
  - PCI-e/DMA integration
  - 40G MAC and PCS/PMA
  - Load balancing module
  - Performance monitoring module
- Offload engines
  - Checksum
  - Fragmentation
  - Aggregation
- 40 G Device driver
- Kernel Tweak
- Demonstration

100 Gb/s NIC (2nd Milestone)

- Upgrade to PCIe 3.0 framework
- Board design, layout and fabrication
- Optimize offload engines
- 100 G MAC and PCS/PMA interface
- 100 G Device driver
- Kernel tweak

3. *Research Accomplishments*

The project was divided into two phases (development of 40 Gb/s and 100 Gb/s NICs), which were allocated between Acadia Optronics LLC and University of New Mexico (UNM) based on our expertise. The objective of the development of a 40 Gb/s prototype NIC card was to further scale it to 100 Gb/s. The maturity of the system components/modules operating at 40 Gb/s made this an attractive milestone for the first year. A 40 Gb/s NIC card based on a Virtex-6 HX380T FPGA with a PCIe2.0 x16 host interface was developed. This was accompanied by the development of a device driver and Linux kernel tweaks to achieve high performance at 40 Gb/s. Following the 40 G NIC card a 100 G capable FPGA based NIC card with high-speed interconnect and Altera Stratix V FPGA 5SGXE7N2F45C2 was then also developed.

Figure 1 illustrates the architectural overview block of our proposed 40/100 Gb/s NIC, highlighting the major building blocks. It was determined that Acadia Optronics LLC should develop the frontend of the system, including PICe Endpoint, DMA Controller, Glue Logic, and Offloading Functions, and that UNM should develop the backend of the system including the Ethernet MAC and PCS cores.

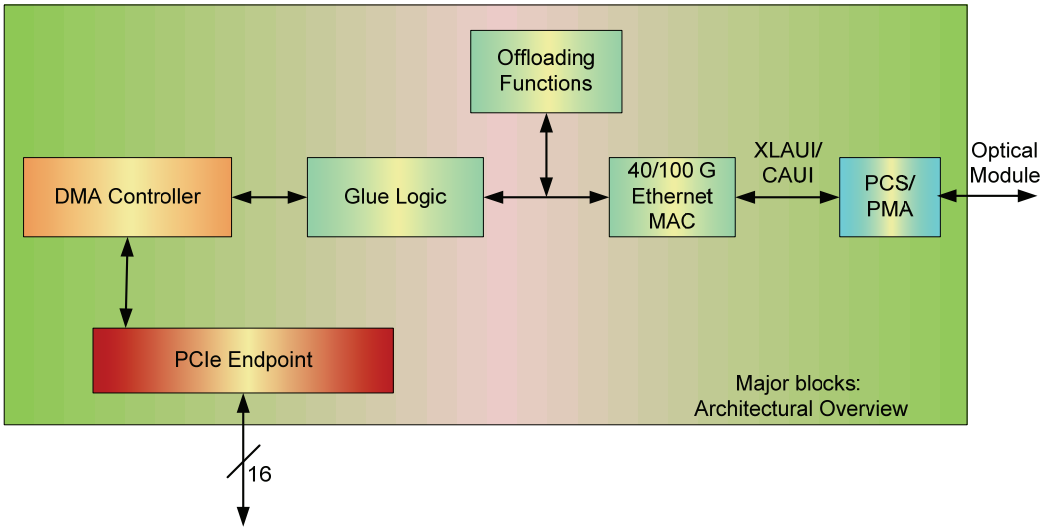

Fig. 1 – Basic building block of the proposed 40/100 Gb/s NIC

Due to its capability in achieving over 40 Gb/s data throughput, we chose a Xilinx Virtex-6 FPGA (HX380T-FF1923) for the first phase of the project. The FPGA-based NIC hardware was fully designed and developed by Acadia Optronics LLC. Based on the designed Virtex-6 FPGA development board, we implemented the initial MAC/PCS for the 40 G NIC at UNM and then extended the definitions to develop the 100 G NIC.

Phase I – 40G NIC Development

Once the initial definitions for the 40 Gb/s MAC and PCS were completed and the 40 G NIC hardware using the Xilinx Virtex-6 FPGA (HX380T-FF1923) was designed and fabricated by Acadia Optronics LLC in the first phase, we started implementing the 40 G MAC and PCS IP cores at the University of New Mexico in this phase.

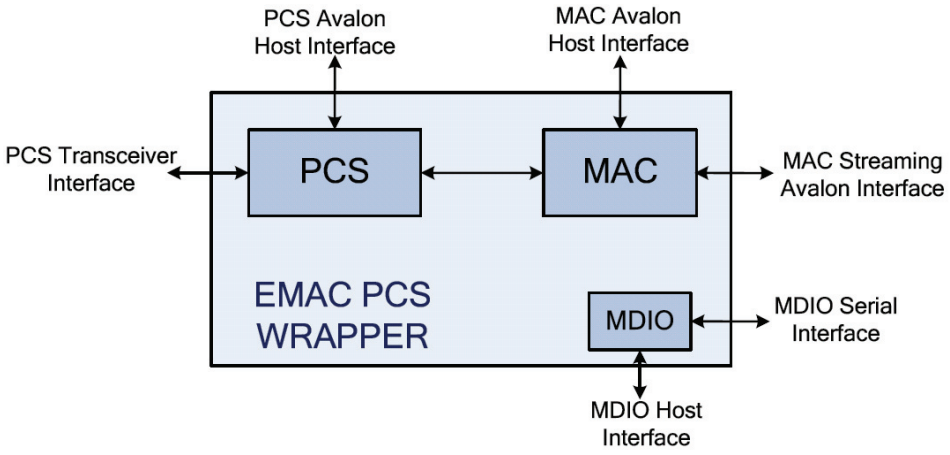

Fig. 2 – 40 G MAC PCS Wrapper Architecture

Figure 2 illustrates the 40G MAC PCS wrapper architecture, where Management Data Input / Output (MDIO) is used for optical module status and control. The MAC and PCS IP cores and the transceiver wrapper were developed for the Virtex-6 HX380T FPGA with a PCIe2.0 x16 host interface. The MAC and PCS cores are designed with a 128-bit data path operating at 312.5MHz. The implemented MAC and PCS cores are fully compliant with the IEEE802.3ba.

The summary of the tasks accomplished in the first phase of this project is listed below:

Task 1.a – The implementation of 40 G MAC was completed at UNM for the Virtex-6 FPGA in the second year. Figure 3 illustrates the simplified 40 G MAC block diagram, where it implements IEEE 802.3bd specifications with a 128 bit XLGMII interface operating at 312.5MHz for 40 Gb/s data rate. It also includes frame padding generation, CRC generation and checking, as well as preamble/SFD generation. In addition, it implements reconciliation sub-layer functionality with start- and end-control characters alignment, error control character and fault sequence insertion and detection, and many more functions for 40 G MAC.

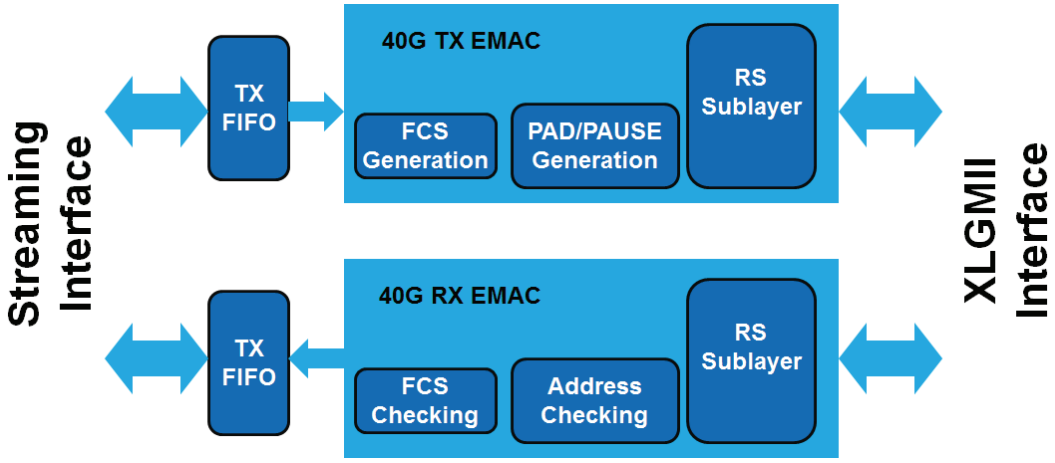

Fig. 3 – Simplified 40G MAC Block Diagram

Task 1.b – Complete development of 40 Gb/s PCS was also performed at UNM for the Virtex-6 FPGA in the second year. Figure 4 illustrates the simplified 40 G PCS block diagram, where it implements IEEE 802.3ba specifications with a 128 bit XLGMII interface operating at 312.5MHz for 40 Gb/s data rate. The developed PCS core implements a 128-bit XLGMII interface for 40G Ethernet compliant with IEEE 802.3ba specifications. It also includes scrambling/descrambling algorithms, Alignment Marker insertion and deletion, lane recording, and many more functions needed for 40 G PCS.

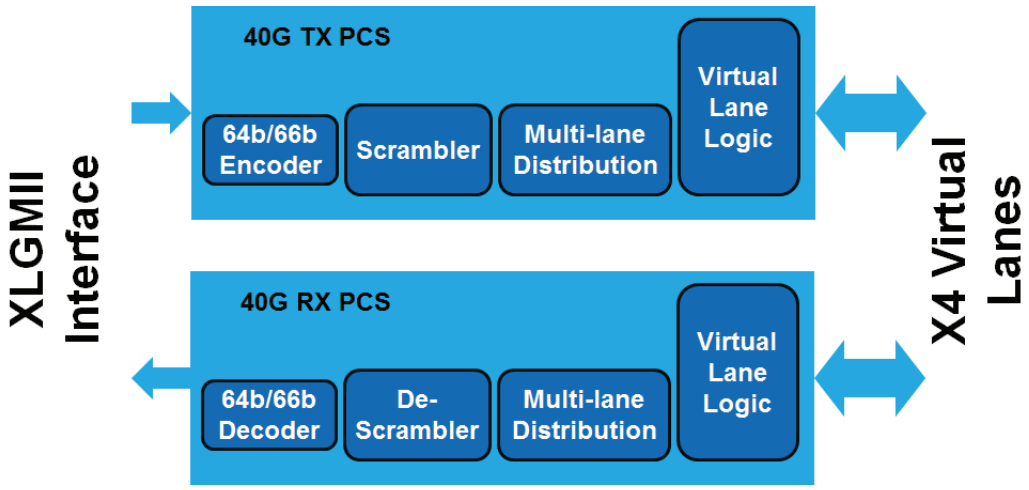

Fig. 4 – Simplified 40G PCS Block Diagram

Phase II – 100G NIC Development

The development of 40 G NIC card in the first phase was a stepping stone toward design and fabrication of the 100 G NIC in the second phase. The schematic and PCB layout design and fabrication of the 100G NIC circuit were completed by Acadia Optronics LLC. This milestone was achieved by extending the same framework used to design our 40 Gb/s NIC. The 100 Gb/s NIC contains an state-of-the-art Altera Stratix-V GX FPGA with support for 4 PCIe 3.0 8-lane endpoints to support 100 Gb/s Ethernet. The final PCB layout design and the fabricated 100 Gb/s designed are illustrated in Figure 5. To interface the developed NIC card with motherboard, we have chosen the standard PCIe 3.0 interface using a 16 lane host connection PCIe to achieve the desired 100 Gb/s data throughput.

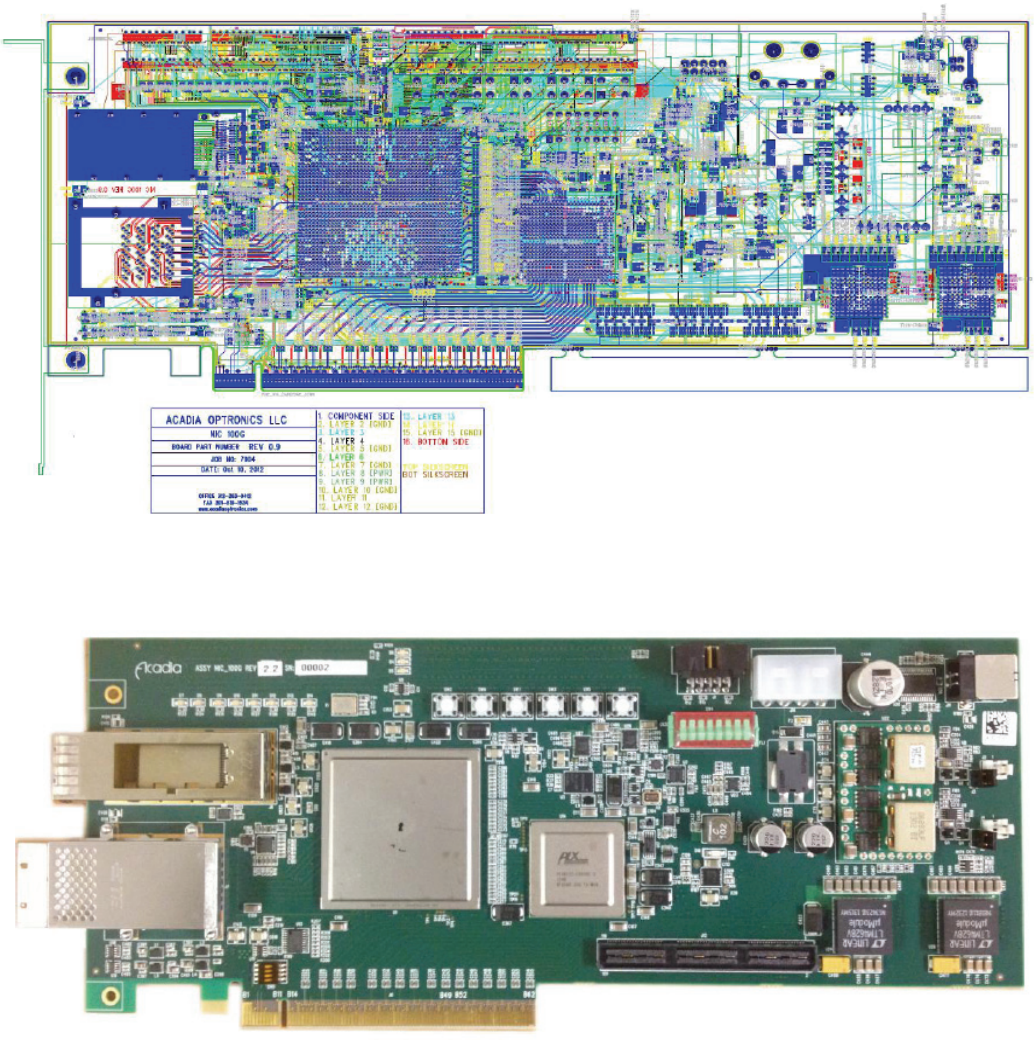

Fig. 5 – The final hardware PCB layout and fabricated 100 Gb/s NIC board

Similar to our initial 40G NIC design, the developed 100 G NIC also includes a 40-100 G Ethernet MAC core in accordance with the IEEE 804.3ba 2010 Higher Speed Ethernet Standard. This IP core handles frame encapsulation and flow of data between client and the network through a 40-100 G Ethernet PCS and PMA.

The block diagrams for the MAC and PCS are illustrated previously in Figures 3 and 4. On the transmit side, the MAC core accepts client frames, inserts the inter-packet gap, preamble, header, start of frame and checksum before passing the frame to the PHY. The PHY further encodes the MAC frame as required for reliable transmission over the media to the remote end. On the receive side, the MAC accepts packets from the PHY, performs the checksum, and strips out the CRC, SFP, and preamble before passing it to the client logic.

The implementation of MAC and PCS on the developed 100 G NIC was completed at the University of New Mexico. In addition, we helped Acadia Optronics LLC with resolving bugs, simulations, and performance measurements.

4. The 100G NIC Hardware Specifications

The complete specifications of the NIC card with board diagram are shown below.

FPGA

- 48 GTX Transceivers (up to 14.1 Gb/s each)
- 4 inbuilt PCIe 3.0 x8 endpoints
- 234,720 ALMs
- 938,880 registers

Network Interconnect

- 1 CXP (10x10) 100 Gb/s
- 1 QSFP (4x10) 40 Gb/s
- 1 JTAG connector for programming and debugging
- 1 USB connector
- 1 QSE connector for programming and debugging.

Memory

- 2 double data rate-small outline dual in-line memory (DDR3 SODIMM) modules at 1600 MHz providing up to 8 Gb/s storage.

5. Conclusions

Although there were many challenges along the way, this project was completed with success. The first milestone, which was 40 G NIC, was achieved during the first two years of the program originated by the department of energy (DoE). Our success in development of the 40 G NIC card was a steppingstone to reach our final product, i.e. 100 G NIC card. After successful design and fabrication of the 100G NIC, thorough testing and debugging were performed at each level. The final network-level testing was performed at NASA Goddard, where we had a chance to use their CXP-based 100 G switch to connect to the 100 Gb/s network. At the NASA Goddard lab, we tested the compatibility of our card with the 100 G Brocade MLXe switch and also performed performance benchmarking of our 100 G card. Using this setup we successfully passed data through the switch at rates approaching the theoretical maximum of 100 Gb/s.