Network Solution for Exascale Architectures

D2.9 BXI to Ethernet bridging demonstrator

Document Properties

Contract Number	955776
Contractual Deadline	M36 (31.03.2024)
Dissemination Level	Public
Nature	Other
Edited by	Damien Berton, ATOS
Authors	Damien Berton, ATOS
Reviewers	Jesús Escudero-Sahuquillo (UCLM) Vangelis Mageiropoulos (FORTH)
Date	28th March 2024
Keywords	High Performance Ethernet, 400 Gbit/s, FPGA
Status	Final
Release	1.0

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955776. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Greece, Germany, Spain, Italy, Switzerland.

History of Changes

Release	Date	Author, Organization	Description of Changes
0.1	13/03/2024	Damien Berton, Atos	Initial release
0.7	28/03/2024	Damien Berton, Atos	Review feedback added
1.0	28/03/2024	Damien Berton, Atos	Final Release

Table of Contents

List of Figures

List of Tables

Executive Summary

This document D2.9 “BXI to Ethernet bridging demonstrator” is produced by Task T2.6 “BXI to Ethernet Gateway prototyping and testing”, which is part of Work package 2 (WP2) “High-performance Ethernet” of the RED-SEA project.

This document contains the BXI to Ethernet bridging demonstrator description, the demonstrator validation report, and finally reports main test results obtained with the demonstrator.

1 Introduction

The purpose of the T2.6 task “BXI to Ethernet Gateway prototyping and testing” is to demonstrate the new generation (BXIv3) BXI to Ethernet gateway functionality solution using FPGA prototyping platforms and off-the-shelf Ethernet devices.

The testbed (also called demonstrator) integrates as much as possible the elements developed in the work package to enable functional validation and testing.

The demonstrator includes:
- One FPGA prototype of the BXIv3 NIC providing the Ethernet gateway functionality.
- One FPGA prototype implementing a test tool for Ethernet traffic observation.
- One “off the shelf” Ethernet switch.
- One Ethernet card inserted in a storage server.
- The software stack including the Ethernet driver for “IPoverBXI” communications.

Note: The low-latency Ethernet MAC and PCS soft IPs developed in the work package (task 2.5) could not be integrated into the demonstrator. Even if the MAC PCS IP ASIC version is not limited in frequency, the FPGA version developed for verification purposes is limited to 100 Gbps data rate, which is not compatible with BXIv3 NIC RTL code, which uses 400 Gbps Ethernet data rate. The demonstrator uses the Intel Agilex HIP called Ftile to implement MAC&PCS layer.

In the next pages, this document includes the following main chapters:

Chapter 2: Ethernet Gateway Demonstrator Description
Chapter 3: Ethernet Gateway Demonstrator Validation Report
Chapter 4 Ethernet gateway performance comparison
Chapter 5: BXIv3 integration test results.

2 BXI to Ethernet gateway testbed description

2.1 Testbed global description

2.1.1 Ethernet gateway testbed initial version

The four modules that constitute the Ethernet gateway testbed are presented in Figure 1:

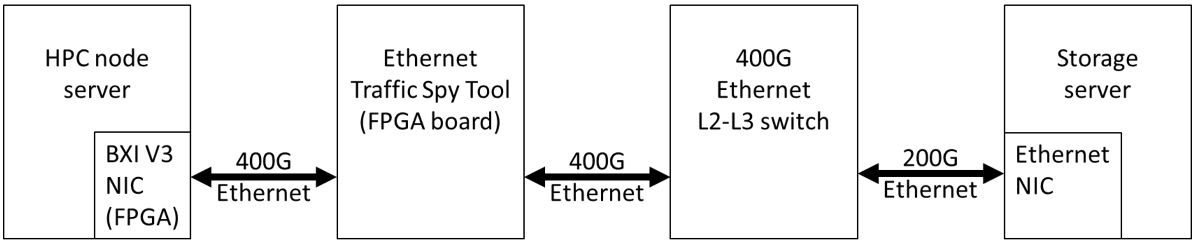


Figure 1 : Ethernet gateway testbed initial version overview.

The purpose of this testbed is to demonstrate communication between an HPC compute note, a member of a BXI fabric, with an off-the-shelf storage server with Ethernet connectivity. The objective is to demonstrate next-generation BXI capability to efficiently communicate with off-the-shelf data center infrastructure.

The role of each module is as follows:

- The HPC node is a high-performance computing server, in charge of communicating its input and output data through an integrated communication device called a Network Interface Card (NIC), which contains a high-performance BXI interface that uses Ethernet standards at the physical level.
- The Ethernet Traffic Spy (ETS) is an observation tool developed in the project, which allows visualizing Ethernet traffic between the HPC node and the BXI switch.
- The L2-L3 switch is in charge of packet routing in the BXI fabric domain(L2 routing configuration), and also outside(L3 routing configuration), to deliver computing results in the data center (or eventually to another HPC fabric in the case of a Modular Supercomputing Architecture (MSA).
- The storage server is in charge of exchanging computing data with the HPC node using a standard Ethernet NIC.

2.1.2 Demonstrator with added test features

Project progress led to the need to add a Test node server with an Ethernet NIC to the test bed initial version, with objective to start validation before availability of the included deliverable of the project: IP over Next-Gen BXI driver (deliverable RED-SEA D2.6) and Ethernet Gateway NIC RTL (deliverable RED-SEA D2.7).

The testbed with the additional test feature is represented in Figure 2.

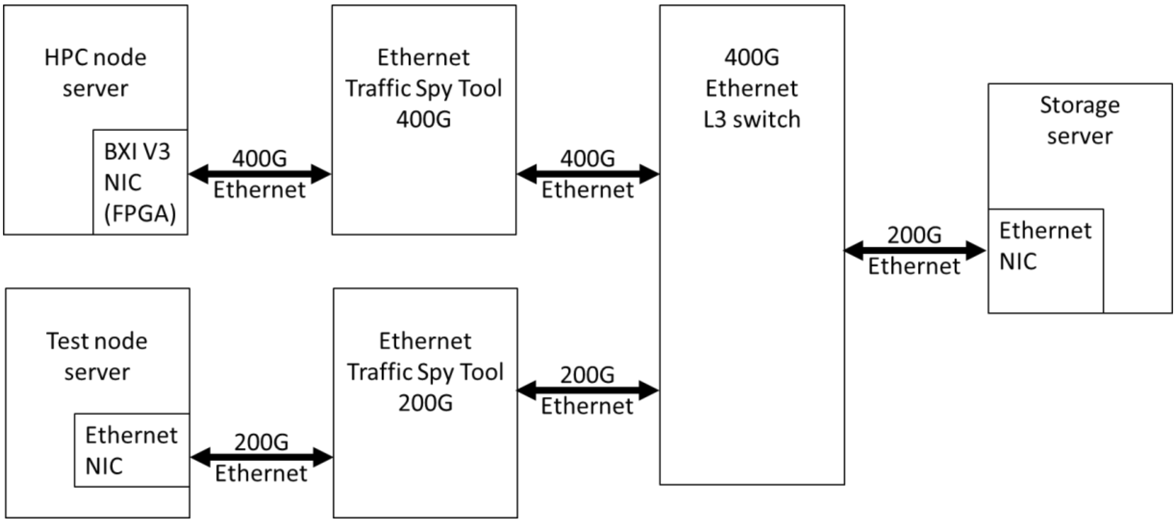


Figure 2 : Ethernet gateway testbed with added test modules

The added test node server provides ability to test:

- Linux-level communication (ping command) for the storage server; refer to Section 3.1.
- “IP over BXI driver” deliverable communication with the storage server independently of the NIC; refer to Section 3.4.

The added 200G Ethernet traffic spy tool provides ability to observe packet content between the test node server and the storage server.

2.2 Testbed module description

2.2.1 “HPC node server” description

The HPC node server is the module intended to host BXIv3 software driver and the NIC RTL code for the Ethernet gateway. It is mainly made of a processor board and a FPGA board, as shown in Figure 3.

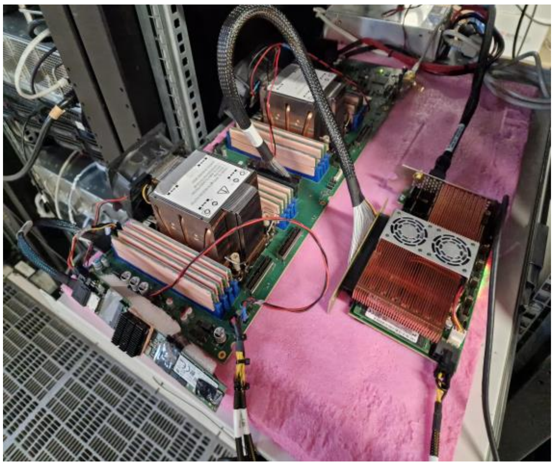


Figure 3 : HPC node server (on the left) with FPGA NIC board (on the right)

The processor board is based on the Central processing Unit (CPU) Intel® Xeon® Platinum 8480+ from 4th Generation Intel® Xeon® Scalable Processors family [2]. It contains 56 cores and has a Thermal Design Power of 350 Watts. This processor supports the latest Pcie express gen5 standard with 32 Gbps lane data-rate that is necessary to manage the full 400 Gbps Ethernet bandwidth targeted by the BXI NIC. The processor is hosted by a Sequana3 (name for Atos latest generation HPC hardware platform) prototype board called C4E.

The BXIv3 NIC RTL (called NICIA) design for Ethernet gateway is hosted by a Reflex CES FPGA board named “XpressSX AGI-FH400G” [3] based on Intel FPGA component AGI027. This board was the first board commercially available that can manage all following FPGA features needed for NICIA: 400G Ethernet HIP, Pcie_gen5_x16 interface, activated integrated ARM A53 processor called HPS, two banks of DDR4 memories, and 2.7 million of FPGA logic elements.

Concerning software aspect, the HPC node is based on Red Hat Enterprise Linux Operating system version 9. BXI over IP driver developed in RED-SEA task 2.3 is installed on the node to communicate with NICIA through the PCie gen5 x16 link.

2.2.2 400G Ethernet Traffic Spy (ETS) description

The 400G Ethernet Traffic Spy (ETS) test tool has been designed during the RED-SEA project with its main feature being Ethernet traffic observation at packet level. The tool is inserted between BXI NIC and BXI switch as described in Section 2.1.

Hardware card used for the ETS

This tool is hosted in an Intel FPGA board called “Agilex Transceiver Soc Development Kit” (Figure 4) in which we use mainly following elements:

- An Agilex FPGA component that contains a 2.7 million logic element matrix, and two Hardware Intellectual Property (HIP) Ethernet Blocks each capable of managing one 400G Ethernet port.
- two 400G Ethernet port using QSFPDD connectors.
- an external JTAG interface used for FPGA configuration and to monitor and parametrize the RTL design and Ethernet HIP blocks.

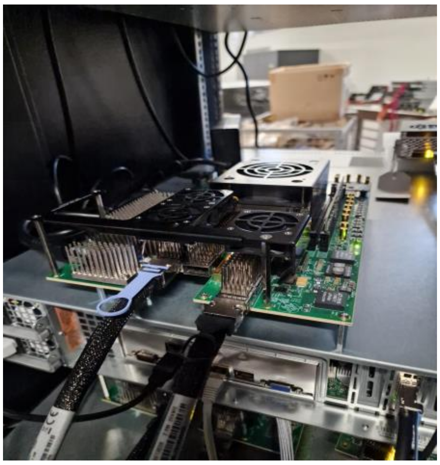


Figure 4 : Ethernet Traffic Spy FPGA board

Ethernet traffic Spy mode of operation

This FPGA design has 3 operating modes that are programmed using the JTAG interface:

- SPY_MODE is the operational mode: packets are transported from one Ethernet port to the other, and the JTAG interface allows accessing internal memories that store packet content for observation.
- MAC_LOOPBACK_MODE: each Ethernet port is loop-backed at the Mac interface; this is a test mode allowing checking Ethernet link connection with a remote Ethernet device.
- PACKET_GEN-ANA_MODE: in this test mode, a programmable packet generator activates the transmitter side, and the receiver side is used to count and analyze the received packets.

Ethernet traffic Spy architecture

The Ethernet Traffic Spy architecture is presented in the figure below:

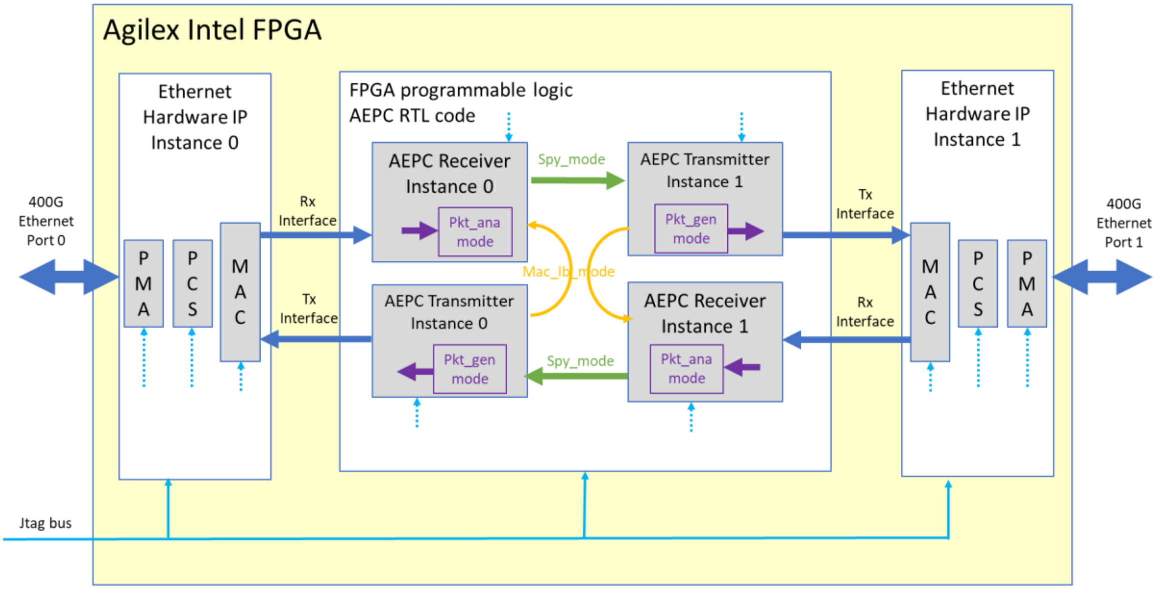


Figure 5 : Ethernet Traffic Spy architecture

The RTL design is made of four sub-blocks, each one is fed by a HIP interface.
The two Atos Ethernet Packet Client Receiver (one sub-block per HIP) are in charge of:

- buffering Ethernet received packets for delivery to Spy_mode interface and mac loopback mode interface.
- add timestamp information to each packet.
- provide an observation buffer to provide ability of packet observation through JTAG interface.

The two Atos Ethernet Packet Client transmitters (one sub-block per HIP) are in charge of:

- managing HIP interface back pressure mechanism,
- multiplexing the HIP Tx interface depending on the working mode programmed through JTAG: spy_mode, mac_loopback mode and packet generator/analysis mode.

The JTAG interface connects all blocks and allows accessing much information in each design block; most useful are:

- HIP initialization status: Tx_Pll_locked, Tx_lane_stable, Rx_clock_recovered, Rx_ready
- MAC level packet statistics for HIP Tx & Rx blocks: number of transmitted/received packets, packet sizes, damaged packets, number of transmitted and received bytes ….
- Received packet content and associated timestamp.

“Format_Eth” packet visualization tool description

“Format_Eth” is a program developed for this project with Tcl language that runs on a computer connected to an Ethernet traffic Spy through the JTAG interface. Its purpose is to visualize the received packets that have been captured by the ETS. The tool provides the following features:

- Display the packets in the order it has been received thanks to timestamp with microsecond granularity independently of the incoming port.
- Rename the IP address and the MAC address with the machine name (from a static table) for clarity.
- Rename some MAC header and IP header fields with their name to facilitate packet type recognition.
- Display packet content in hexadecimal format.

This tool supports both 200G and 400G Ethernet versions of the ETS.

Packet generation and analysis mode usage with NIC FPGA board

The ETS has also been adapted to be used with a NIC FPGA board (which means a board with only one Ethernet port and a PCIe port) and specifically with the Reflex-Ces board described in 2.2.1. This allows using the ETS with packet generation & analysis mode instead of the BXIv3 NIC design.

This ETS capability has been used for demonstrator validation before the BXIv3 NIC RTL design is finalized for hardware test.

2.2.3 BXI switch description

The demonstrator uses the Broadcom Tomahawk5 SVK [4] represented in Figure 6 for the BXI switch prototyping.

This switch provides 64 programmable Ethernet ports from 10 Gbps up to 800 Gbps. The Ethernet gateway demonstrator uses one 400G port for the NICIA interface and two 200G Ethernet ports to connect other servers. This is represented Figure 2.

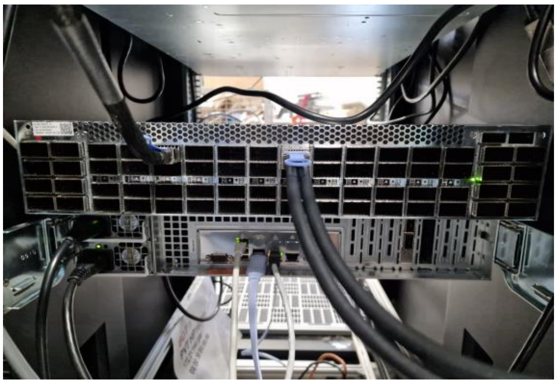


Figure 6 : Broadcom Tomahawk5 SVK prototyping switch

This switch has two main modes for delivering packets from one port to the other:

- Level 2 switching: packets are transferred based on their destination MAC address. This mode is used for the demonstrator validation with an “all to all” approach, in which all connected servers are placed in a Local Area Network (LAN) and can communicate together.
- Level 3 switching: packets are transferred based on their destination IP address. In this mode the switch is used to connect two (or more than 2) LANs. This mode is used for the BXI to Ethernet gateway function.

2.2.4 “Storage server” description

The “storage server” (represented Figure 7) is an Off-the-shelf Linux server equipped with a 200G Ethernet NIC board from Nvidia. The installed software is based on Red Hat Enterprise Linux version 8.

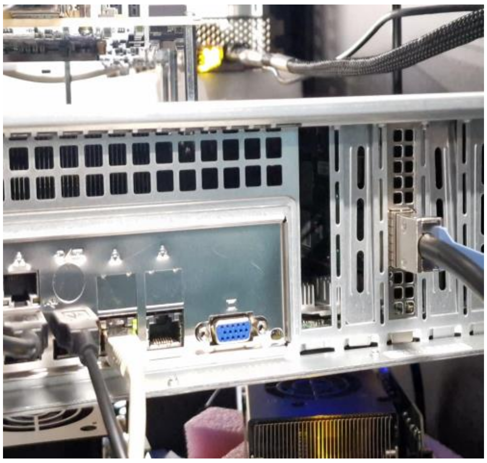


Figure 7 : Storage Server with 200G Ethernet NIC connection (on the right)

2.2.5 200G Ethernet Traffic Spy description

The 200G Ethernet Traffic Spy test tool has been designed as an initial version of the 400G Ethernet Traffic Spy described in Section 2.2.2. It has been added in the demonstrator to provide additional debug and observation capabilities to the demonstrator. The RTL source code is common with the 400G version. Data-rate-dependent source code is controlled with a system Verilog parameter.

2.2.6 “Test node server” description

The “test node server” has been added to the demonstrator to provide the capability of testing the BXI driver independently of the BXI NIC.

Hardware is the same as the “storage server” and operating system is also the same. On top of that, a Linux virtual machine is created with QEMU (Quick Emulator) and KVM (Kernel Virtual Machine) software. This virtual machine included an emulated BXI NIC whose interface is configured to be directly connected to the standard Ethernet interface of the host system. This allows running the BXI driver software on the emulated BXI NIC and performing Ethernet communications with the “storage server”.

3 BXI to Ethernet gateway testbed validation report

This chapter presents the test results showing the proper functioning of the Ethernet gateway testbed.

3.1 Storage server communication test

3.1.1 Test description

This test checks that "storage server" communicates with the "test server" through the switch prototype and the 200G version of ETS as represented Figure 8.

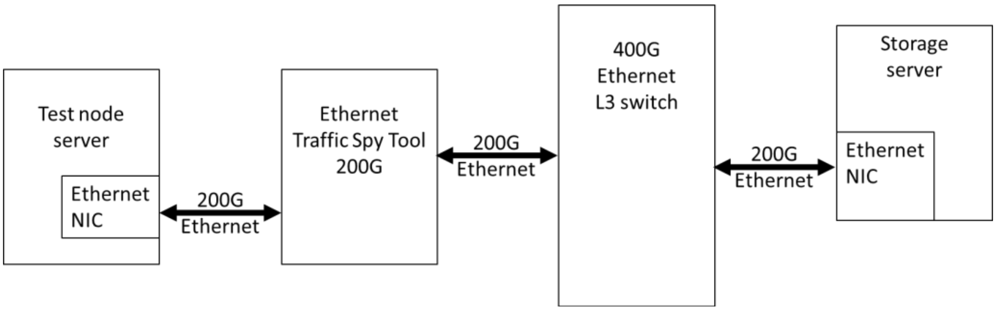
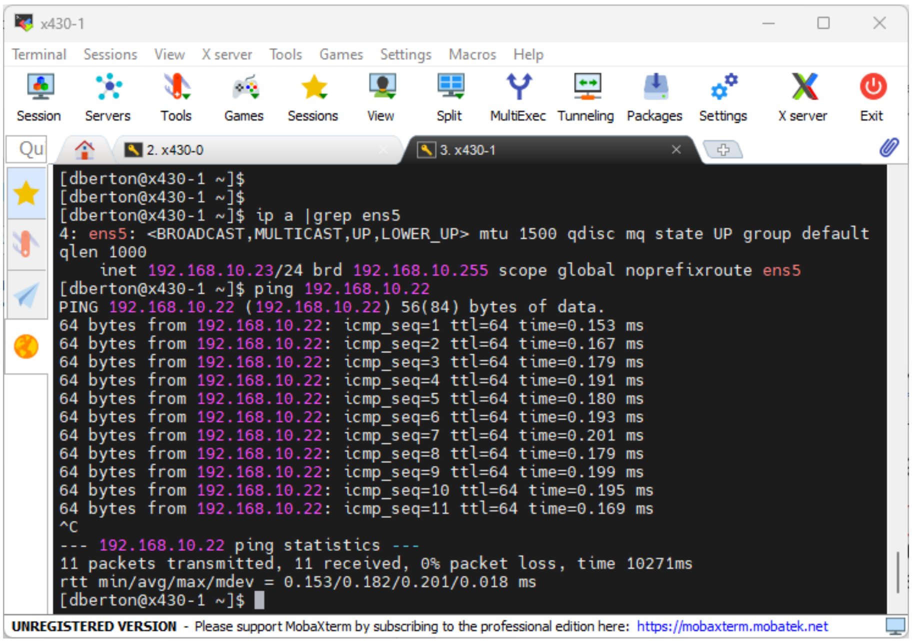


Figure 8 : Storage server communication test overview

The test is performed at OSI layer 3 with the Linux "ping" command.
The ETS is used in 200G spy mode with the "Format_eth" tool to provide a packet-level visualization of the communication between the two servers.
The switch is configured with an "all-to-all" level 2 routing configuration scheme.

3.1.2 Test results

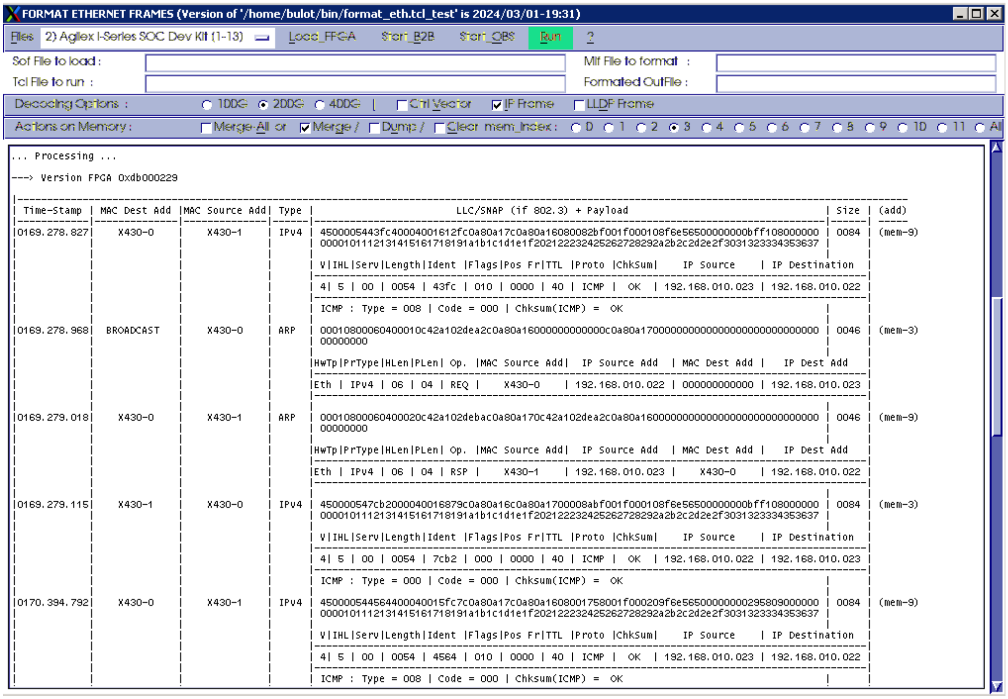
The ping command is successful between the two servers without packet loss and the average delay is approximately 180 µs, which is a standard result with Ethernet 200G NIC.


x430-1
Terminal Sessions View X server Tools Games Settings Macros Help
Session Servers Tools Games Sessions View Split MultiExec Tunneling Packages Settings X server Exit
2. x430-0
3. x430-1

```
[dberton@x430-1 ~]$
[dberton@x430-1 ~]$
[dberton@x430-1 ~]$ ip a |grep ens5
4: ens5: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc mq state UP group default
qlen 1000
    inet 192.168.10.23/24 brd 192.168.10.255 scope global noprefixroute ens5
[dberton@x430-1 ~]$ ping 192.168.10.22
PING 192.168.10.22 (192.168.10.22) 56(84) bytes of data.
64 bytes from 192.168.10.22: icmp_seq=1 ttl=64 time=0.153 ms
64 bytes from 192.168.10.22: icmp_seq=2 ttl=64 time=0.167 ms
64 bytes from 192.168.10.22: icmp_seq=3 ttl=64 time=0.179 ms
64 bytes from 192.168.10.22: icmp_seq=4 ttl=64 time=0.191 ms
64 bytes from 192.168.10.22: icmp_seq=5 ttl=64 time=0.180 ms
64 bytes from 192.168.10.22: icmp_seq=6 ttl=64 time=0.193 ms
64 bytes from 192.168.10.22: icmp_seq=7 ttl=64 time=0.201 ms
64 bytes from 192.168.10.22: icmp_seq=8 ttl=64 time=0.179 ms
64 bytes from 192.168.10.22: icmp_seq=9 ttl=64 time=0.199 ms
64 bytes from 192.168.10.22: icmp_seq=10 ttl=64 time=0.195 ms
64 bytes from 192.168.10.22: icmp_seq=11 ttl=64 time=0.169 ms
^C
--- 192.168.10.22 ping statistics ---
11 packets transmitted, 11 received, 0% packet loss, time 10271ms
rtt min/avg/max/mdev = 0.153/0.182/0.201/0.018 ms
[dberton@x430-1 ~]$
```

UNREGISTERED VERSION - Please support MobaXterm by subscribing to the professional edition here: https://mobaxterm.mobatek.net

Figure 9 : Storage server Ping test result



FORMAT ETHERNET FRAMES (Version of '/home/bulot/bin/format_eth.tcl_test' is 2024/03/01-19:31)

Sof File to load :
Tcl File to run :
Mif File to format :
Formated OutFile :

Decoding Options : ○ 100G ● 200G ○ 400G | □ Ctrl Vector ☑ IP Frame □ LLDP Frame

Actions on Memory : □ Merge-All or ☑ Merge / □ Dump / □ Clear mem_Index : ○ 0 ○ 1 ○ 2 ● 3 ○ 4 ○ 5 ○ 6 ○ 7 ○ 8 ○ 9 ○ 10 ○ 11 ○ All

```
... Processing ...

---> Version FPGA 0xdb000229

|-----------------------------------------------------------------------------------------------------------------------------------------------------
| Time-Stamp | MAC Dest Add |MAC Source Add| Type |                          LLC/SNAP (if 802.3) + Payload                               | Size | (add)
|------------|--------------|--------------|------|--------------------------------------------------------------------------------------|------|  -----
|0169.278.827|    X430-0    |    X430-1    | IPv4 | 4500005443fc40004001612fc0a80a17c0a80a16080082bf001f000108f6e56500000000bff108000000 | 0084 | (mem-9)
|            |              |              |      | 000010111213141516171819 1a1b1c1d1e1f202122232425262728292a2b2c2d2e2f3031323334353637 |      |
|            |              |              |      |                                                                                      |      |
|            |              |              |      | V|IHL|Serv|Length|Ident |Flags|Pos Fr|TTL |Proto |ChkSum|    IP Source    | IP Destination |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      | 4| 5 | 00 | 0054 | 43fc | 010 | 0000 | 40 | ICMP |  OK  | 192.168.010.023 | 192.168.010.022 |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      | ICMP : Type = 008 | Code = 000 | Chksum(ICMP) =  OK                                    |      |
|            |              |              |      |                                                                                      |      |
|0169.278.968|  BROADCAST   |    X430-0    | ARP  | 00010800060400010c42a102dea2c0a80a16000000000000c0a80a17000000000000000000000000000 | 0046 | (mem-3)
|            |              |              |      | 00000000                                                                             |      |
|            |              |              |      |                                                                                      |      |
|            |              |              |      |HwTp|PrType|HLen|PLen| Op. |MAC Source Add|  IP Source Add  | MAC Dest Add |  IP Dest Add  |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      |Eth | IPv4 | 06 | 04 | REQ |    X430-0    | 192.168.010.022 | 000000000000 | 192.168.010.023 |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      |                                                                                      |      |
|0169.279.018|    X430-0    |    X430-1    | ARP  | 00010800060400020c42a102debac0a80a170c42a102dea2c0a80a16000000000000000000000000000 | 0046 | (mem-9)
|            |              |              |      | 00000000                                                                             |      |
|            |              |              |      |                                                                                      |      |
|            |              |              |      |HwTp|PrType|HLen|PLen| Op. |MAC Source Add|  IP Source Add  | MAC Dest Add |  IP Dest Add  |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      |Eth | IPv4 | 06 | 04 | RSP |    X430-1    | 192.168.010.023 |    X430-0    | 192.168.010.022 |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      |                                                                                      |      |
|0169.279.115|    X430-1    |    X430-0    | IPv4 | 450000547cb2000040016879c0a80a16c0a80a1700008abf001f000108f6e56500000000bff108000000 | 0084 | (mem-3)
|            |              |              |      | 000010111213141516171819 1a1b1c1d1e1f202122232425262728292a2b2c2d2e2f3031323334353637 |      |
|            |              |              |      |                                                                                      |      |
|            |              |              |      | V|IHL|Serv|Length|Ident |Flags|Pos Fr|TTL |Proto |ChkSum|    IP Source    | IP Destination |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      | 4| 5 | 00 | 0054 | 7cb2 | 000 | 0000 | 40 | ICMP |  OK  | 192.168.010.022 | 192.168.010.023 |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      | ICMP : Type = 000 | Code = 000 | Chksum(ICMP) =  OK                                    |      |
|            |              |              |      |                                                                                      |      |
|0170.394.792|    X430-0    |    X430-1    | IPv4 | 4500005445644000400015fc7c0a80a17c0a80a1608001758001f000209f6e565000000000295809000000 | 0084 | (mem-9)
|            |              |              |      | 000010111213141516171819 1a1b1c1d1e1f202122232425262728292a2b2c2d2e2f3031323334353637 |      |
|            |              |              |      |                                                                                      |      |
|            |              |              |      | V|IHL|Serv|Length|Ident |Flags|Pos Fr|TTL |Proto |ChkSum|    IP Source    | IP Destination |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      | 4| 5 | 00 | 0054 | 4564 | 010 | 0000 | 40 | ICMP |  OK  | 192.168.010.023 | 192.168.010.022 |
|            |              |              |      |--------------------------------------------------------------------------------------|
|            |              |              |      | ICMP : Type = 008 | Code = 000 | Chksum(ICMP) =  OK                                    |      |
```

Figure 10 : Storage Ping test result Ethernet packet visualization with “Format_eth”

Packet observation shows the ARP and ICMP packets used by the ping command as expected.

The test node is named x430-0 and its IP address is 192.168.10.22.
The storage node is named x431-0 and its IP address is 192.168.10.23.

In the trace shown in Figure 10, we can see:

- a first trace from a previous execution (should be ignored).
- the ARP request from the test node that is broadcasted to everyone on the network.
- the ARP response from the storage node to the test node.
- the ICMP request from the test node to the storage node.
- the ICMP response from the storage node to the test node.

3.1.3 Conclusion

This test demonstrates:

- Successful configuration of the “storage server” and “test server” with their 200G NIC communication card.
- Successful configuration of the switch prototype for the two 200G ports.
- Successful behaviour of the ETS as an Ethernet spy and capability to observe Ethernet packet content.

3.2 Ethernet test from the HPC node server beyond the switch

3.2.1 Test description

Initial objective of this test was to test all Ethernet hardware links included in the demonstrator. Unfortunately, the Nvidia 200G NIC has an external loopback mode restricted to 100 Gbps data rate, that is not compliant with our 200 Gbps minimum objective. That is why we use the 200G version of the ETS tool for loopback.

In this test, the FPGA content dedicated to BXIv3 NIC is replaced also with the ETS FPGA design in packet generation & analysis mode, which is used as stimulus and result control for the test.

The 400G ETS is tested in nominal 400G spy mode and feeds the switch used as a 400G to 200G gateway. At last, the 200G version of ETS is used in loopback mode to send back packets to the same path.

The switch is configured with layer 2 routing to obtain packet transmission without any packet header modification.

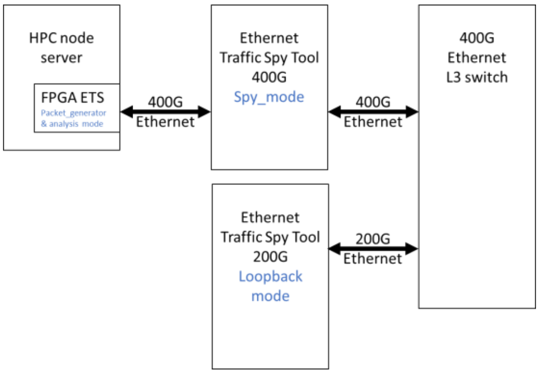
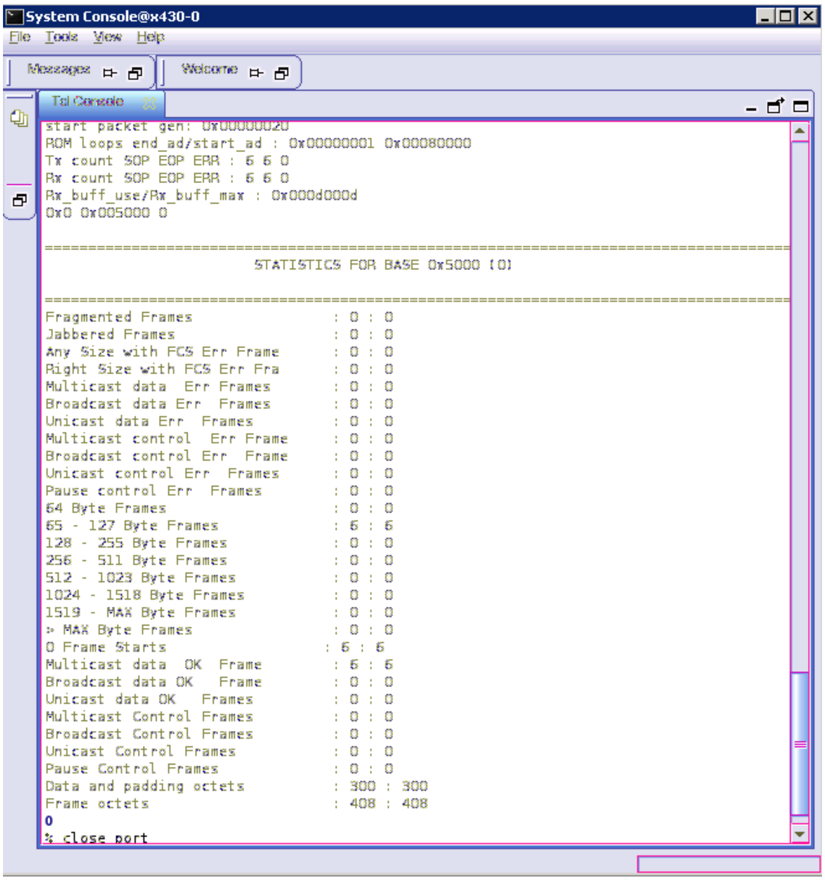


Figure 11 : HPC node beyond the switch test path representation

3.2.2 Test results

It is observed (Figure 12) that transmitted packets (left column statistics) are all received (right column) without error or modification.



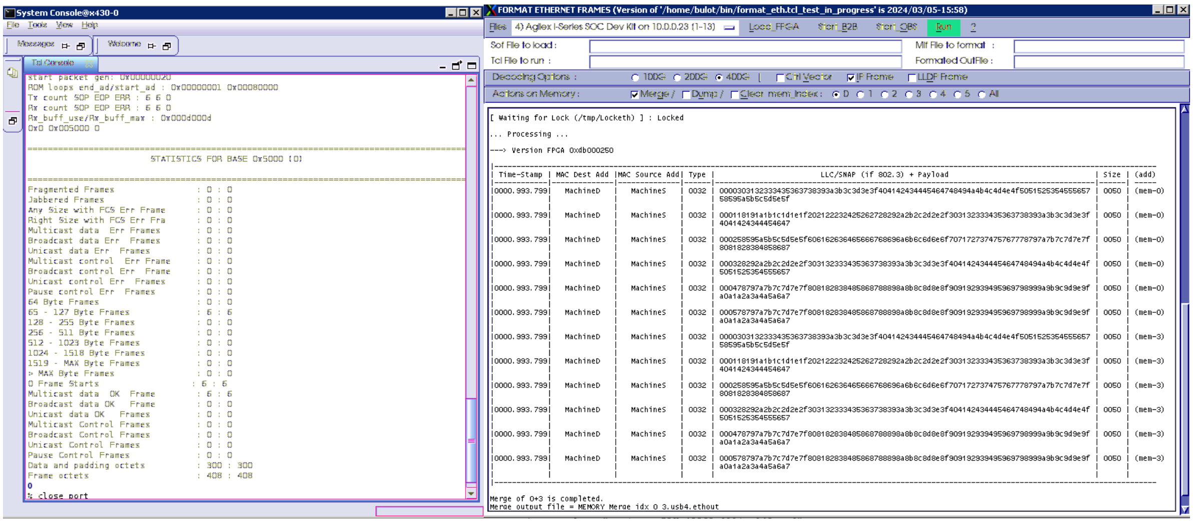
```
start packet gen: 0x00000020
ROM loops end_ad/start_ad : 0x00000001 0x00080000
Tx count SOP EOP ERR : 6 6 0
Rx count SOP EOP ERR : 6 6 0
Rx_buff_use/Rx_buff_max : 0x000d000d
0x0 0x005000 0

==========================================================================================
                          STATISTICS FOR BASE 0x5000 (0)

==========================================================================================
Fragmented Frames                   : 0 : 0
Jabbered Frames                     : 0 : 0
Any Size with FCS Err Frame         : 0 : 0
Right Size with FCS Err Fra         : 0 : 0
Multicast data  Err Frames          : 0 : 0
Broadcast data Err  Frames          : 0 : 0
Unicast data Err  Frames            : 0 : 0
Multicast control  Err Frame        : 0 : 0
Broadcast control Err  Frame        : 0 : 0
Unicast control Err  Frames         : 0 : 0
Pause control Err  Frames           : 0 : 0
64 Byte Frames                      : 0 : 0
65 - 127 Byte Frames                : 6 : 6
128 - 255 Byte Frames               : 0 : 0
256 - 511 Byte Frames               : 0 : 0
512 - 1023 Byte Frames              : 0 : 0
1024 - 1518 Byte Frames             : 0 : 0
1519 - MAX Byte Frames              : 0 : 0
> MAX Byte Frames                   : 0 : 0
0 Frame Starts                    : 6 : 6
Multicast data  OK  Frame           : 6 : 6
Broadcast data OK   Frame           : 0 : 0
Unicast data OK   Frames            : 0 : 0
Multicast Control Frames            : 0 : 0
Broadcast Control Frames            : 0 : 0
Unicast Control Frames              : 0 : 0
Pause Control Frames                : 0 : 0
Data and padding octets             : 300 : 300
Frame octets                        : 408 : 408
0
% close port
```

Figure 12 : Ethernet Packet generation and analysis statistics result

Packet content displayed with Format_eth tool (Figure 13) at the 400G ETS confirms that transmitted packets (the six first one stored in mem0) are observed identical to those received after loopback (the last six packets stored in mem3).



```
start packet gen: 0x00000020
ROM loops end_ad/start_ad : 0x00000001 0x00080000
Tx count SOP EOP ERR : 6 6 0
Rx count SOP EOP ERR : 6 6 0
Rx_buff_use/Rx_buff_max : 0x000d000d
0x0 0x005000 0

==========================================================================================
                          STATISTICS FOR BASE 0x5000 (0)

==========================================================================================
Fragmented Frames                   : 0 : 0
Jabbered Frames                     : 0 : 0
Any Size with FCS Err Frame         : 0 : 0
Right Size with FCS Err Fra         : 0 : 0
Multicast data  Err Frames          : 0 : 0
Broadcast data Err  Frames          : 0 : 0
Unicast data Err  Frames            : 0 : 0
Multicast control  Err Frame        : 0 : 0
Broadcast control Err  Frame        : 0 : 0
Unicast control Err  Frames         : 0 : 0
Pause control Err  Frames           : 0 : 0
64 Byte Frames                      : 0 : 0
65 - 127 Byte Frames                : 6 : 6
128 - 255 Byte Frames               : 0 : 0
256 - 511 Byte Frames               : 0 : 0
512 - 1023 Byte Frames              : 0 : 0
1024 - 1518 Byte Frames             : 0 : 0
1519 - MAX Byte Frames              : 0 : 0
> MAX Byte Frames                   : 0 : 0
0 Frame Starts                    : 6 : 6
Multicast data  OK  Frame           : 6 : 6
Broadcast data OK   Frame           : 0 : 0
Unicast data OK   Frames            : 0 : 0
Multicast Control Frames            : 0 : 0
Broadcast Control Frames            : 0 : 0
Unicast Control Frames              : 0 : 0
Pause Control Frames                : 0 : 0
Data and padding octets             : 300 : 300
Frame octets                        : 408 : 408
0
% close port
```

FORMAT ETHERNET FRAMES (Version of '/home/bulot/bin/format_eth.tcl_test_in_progress' is 2024/03/05-15:58)
Files
Load_FPGA
Start_B2B
Start_OBS
Run
?

4) Agilex I-Series SOC Dev Kit on 10.0.0.23 (1-13)

Sof File to load :
Tcl File to run :
Mif File to format :
Formated OutFile :

Decoding Options : ○ 100G ○ 200G ● 400G | □ Ctrl Vector ☑ IF Frame □ LLDP Frame

Actions on Memory : ☑ Merge / □ Dump / □ Clear mem_Index : ● 0 ○ 1 ○ 2 ○ 3 ○ 4 ○ 5 ○ All

```
[ Waiting for Lock (/tmp/Locketh) ] : Locked

... Processing ...

---> Version FPGA 0xdb000250

|-----------------------------------------------------------------------------------------------------------------------------------------------
| Time-Stamp | MAC Dest Add |MAC Source Add| Type |                         LLC/SNAP (if 802.3) + Payload                            | Size | (add)
|------------|--------------|--------------|------|----------------------------------------------------------------------------------|------|------
|0000.993.799|   MachineD   |   MachineS   | 0032 | 0000303132333435363738393a3b3c3d3e3f404142434445464748494a4b4c4d4e4f5051525354555657 | 0050 | (mem-0)
|            |              |              |      | 58595a5b5c5d5e5f                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000118191a1b1c1d1e1f202122232425262728292a2b2c2d2e2f303132333435363738393a3b3c3d3e3f | 0050 | (mem-0)
|            |              |              |      | 4041424344454647                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000258595a5b5c5d5e5f606162636465666768696a6b6c6d6e6f707172737475767778797a7b7c7d7e7f | 0050 | (mem-0)
|            |              |              |      | 8081828384858687                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000328292a2b2c2d2e2f303132333435363738393a3b3c3d3e3f404142434445464748494a4b4c4d4e4f | 0050 | (mem-0)
|            |              |              |      | 5051525354555657                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000478797a7b7c7d7e7f808182838485868788898a8b8c8d8e8f909192939495969798999a9b9c9d9e9f | 0050 | (mem-0)
|            |              |              |      | a0a1a2a3a4a5a6a7                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000578797a7b7c7d7e7f808182838485868788898a8b8c8d8e8f909192939495969798999a9b9c9d9e9f | 0050 | (mem-0)
|            |              |              |      | a0a1a2a3a4a5a6a7                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 0000303132333435363738393a3b3c3d3e3f404142434445464748494a4b4c4d4e4f5051525354555657 | 0050 | (mem-3)
|            |              |              |      | 58595a5b5c5d5e5f                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000118191a1b1c1d1e1f202122232425262728292a2b2c2d2e2f303132333435363738393a3b3c3d3e3f | 0050 | (mem-3)
|            |              |              |      | 4041424344454647                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000258595a5b5c5d5e5f606162636465666768696a6b6c6d6e6f707172737475767778797a7b7c7d7e7f | 0050 | (mem-3)
|            |              |              |      | 8081828384858687                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000328292a2b2c2d2e2f303132333435363738393a3b3c3d3e3f404142434445464748494a4b4c4d4e4f | 0050 | (mem-3)
|            |              |              |      | 5051525354555657                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000478797a7b7c7d7e7f808182838485868788898a8b8c8d8e8f909192939495969798999a9b9c9d9e9f | 0050 | (mem-3)
|            |              |              |      | a0a1a2a3a4a5a6a7                                                                 |      |
|0000.993.799|   MachineD   |   MachineS   | 0032 | 000578797a7b7c7d7e7f808182838485868788898a8b8c8d8e8f909192939495969798999a9b9c9d9e9f | 0050 | (mem-3)
|            |              |              |      | a0a1a2a3a4a5a6a7                                                                 |      |
|-----------------------------------------------------------------------------------------------------------------------------------------------

Merge of 0+3 is completed.
Merge output file = MEMORY Merge idx 0 3.usb4.ethout
```

Figure 13 : 400G Ethernet Traffic Spy “Format_eth” result

3.2.3 Conclusion

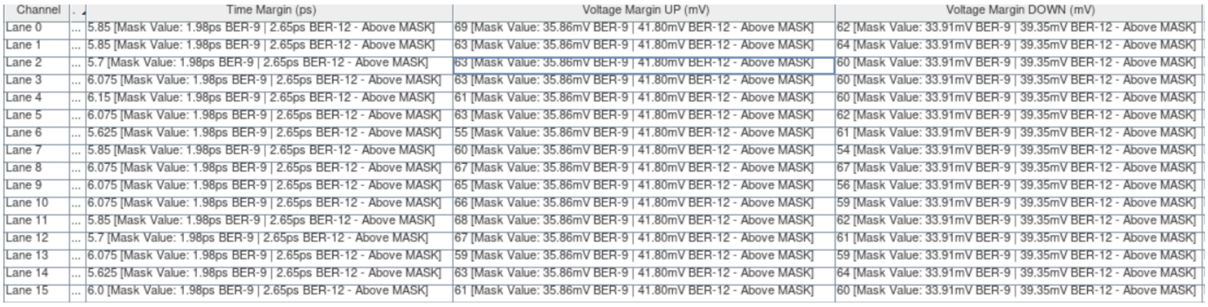
We deduce from this test:

- The 400G ETS can be inserted in a 400G Ethernet link without creating perturbation and provides the ability to observe all packets in both directions.
- The switch provides 400G Ethernet to 200G Ethernet transformation without altering packets.
- Consequently, 400 Gbps Ethernet communication from the HPC node NIC card beyond the switch is validated.
- The 200G ETS “mac_loopback” mode is functional.

3.3 HPC node PCIe gen5 test with FPGA board

One important requirement for the FPGA board for the 400G NIC RTL code is its PCIe bandwidth with the host processor. The objective of 400 Gbps Ethernet bandwidth for the NIC requires the same bandwidth on the PCIe interface. This can be achieved only with the new “gen5” generation of PCIe standard used with 16 lanes: Each pcie_gen5 lane has 32 Gbps bandwidth that generates a total bandwidth of 512 Gbps superior to the 400 Gbps we needed to use Ethernet at its maximum bandwidth.

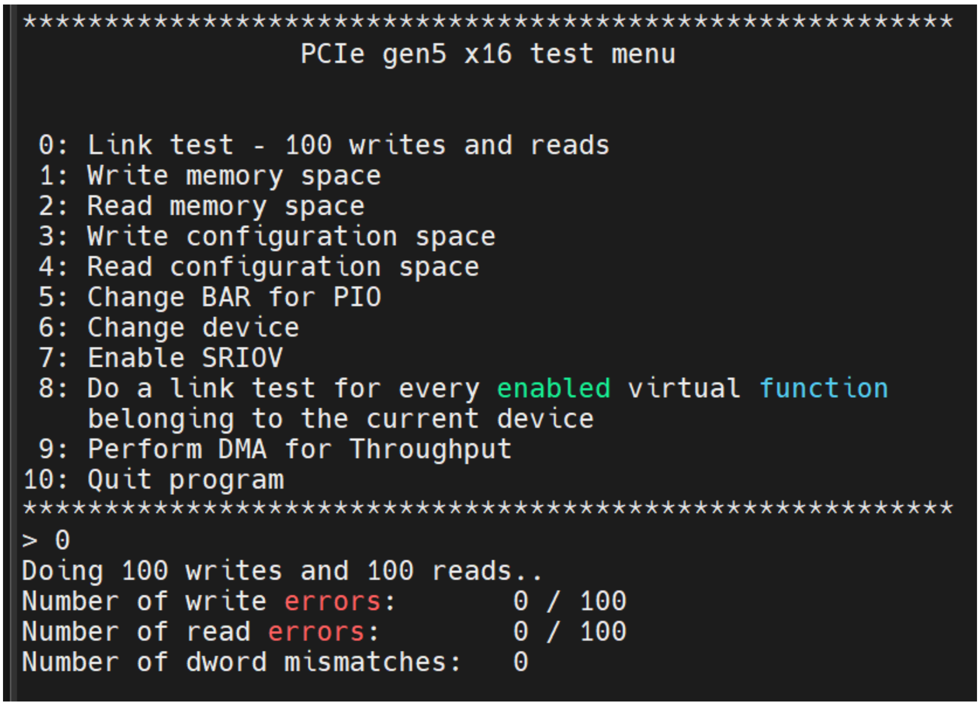
This test is performed on the “HPC node server” (represented in Figure 1) with its connected Reflex CES FPGA board. An FPGA design code specific to the PCIe gen5 test is generated with the Intel Quartus tool. This FPGA design provides a “PCIe toolkit” software that provides eye diagram analysis to check each lane signal integrity at the physical level. The eye diagram is not visualized but the main eye characteristics are displayed: horizontal length (in picoseconds) and vertical height (in millivolts). The analysis shows that the obtained values are above the mask-specified minimum values.



Channel	.	Time Margin (ps)	Voltage Margin UP (mV)	Voltage Margin DOWN (mV)
Lane 0	...	5.85 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	69 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	62 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 1	...	5.85 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	63 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	64 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 2	...	5.7 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	63 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	60 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 3	...	6.075 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	63 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	60 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 4	...	6.15 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	61 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	60 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 5	...	6.075 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	63 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	62 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 6	...	5.625 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	55 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	61 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 7	...	5.85 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	60 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	54 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 8	...	6.075 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	67 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	67 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 9	...	6.075 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	65 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	56 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 10	...	6.075 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	66 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	59 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 11	...	5.85 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	68 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	62 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 12	...	5.7 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	67 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	61 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 13	...	6.075 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	59 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	59 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 14	...	5.625 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	63 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	64 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]
Lane 15	...	6.0 [Mask Value: 1.98ps BER-9 | 2.65ps BER-12 - Above MASK]	61 [Mask Value: 35.86mV BER-9 | 41.80mV BER-12 - Above MASK]	60 [Mask Value: 33.91mV BER-9 | 39.35mV BER-12 - Above MASK]

Figure 14 : Pcie gen5 eye diagram result for the FPGA board

We also use Intel PCIe gen5 test software that runs on the host processor to check at upper level that the host can write and read in the FPGA memory without errors.



```
*********************************************************
                 PCIe gen5 x16 test menu


 0: Link test - 100 writes and reads
 1: Write memory space
 2: Read memory space
 3: Write configuration space
 4: Read configuration space
 5: Change BAR for PIO
 6: Change device
 7: Enable SRIOV
 8: Do a link test for every enabled virtual function
    belonging to the current device
 9: Perform DMA for Throughput
10: Quit program
*********************************************************
> 0
Doing 100 writes and 100 reads..
Number of write errors:       0 / 100
Number of read errors:        0 / 100
Number of dword mismatches:   0
```

Figure 15 : PCIe gen5 test software result on the host processor

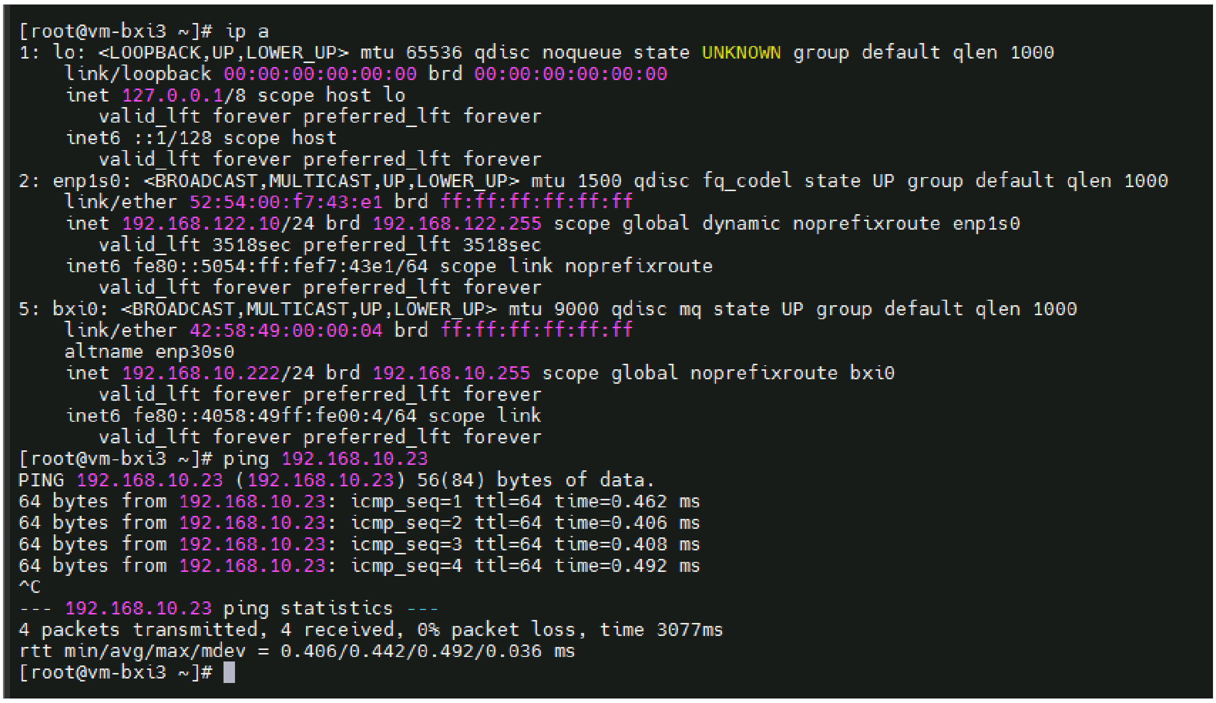
3.4 BXI-over-IP driver test

The main objective of this test is to check that the BXI-over-IP V3 software driver connects successfully with its two external software interfaces: the Linux operating system and the NIC card.

For this test, The BXI-over-IP V3 software is hosted in the “test node server” described in Section 2.1.2 & 2.2.6 and we check communication with the “storage server”. The QEMU software is used to place the driver in a virtual machine, which allows connecting the driver to the Ethernet network card of the server. An IP interface is created on the Test node server for the BXI driver with IP address 192.168.10.222.

The test consists of checking whether “ping” Linux command is successful with the distant “storage server” whose network interface is configured with IP address 192.168.10.23.

For this test, the test node and the storage server have self-assigned Ips in same subnet, and the switch is configured with a n “all-to all” level 2 switching configuration.



```
[root@vm-bxi3 ~]# ip a
1: lo: <LOOPBACK,UP,LOWER_UP> mtu 65536 qdisc noqueue state UNKNOWN group default qlen 1000
    link/loopback 00:00:00:00:00:00 brd 00:00:00:00:00:00
    inet 127.0.0.1/8 scope host lo
       valid_lft forever preferred_lft forever
    inet6 ::1/128 scope host
       valid_lft forever preferred_lft forever
2: enp1s0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 1500 qdisc fq_codel state UP group default qlen 1000
    link/ether 52:54:00:f7:43:e1 brd ff:ff:ff:ff:ff:ff
    inet 192.168.122.10/24 brd 192.168.122.255 scope global dynamic noprefixroute enp1s0
       valid_lft 3518sec preferred_lft 3518sec
    inet6 fe80::5054:ff:fef7:43e1/64 scope link noprefixroute
       valid_lft forever preferred_lft forever
5: bxi0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9000 qdisc mq state UP group default qlen 1000
    link/ether 42:58:49:00:00:04 brd ff:ff:ff:ff:ff:ff
    altname enp30s0
    inet 192.168.10.222/24 brd 192.168.10.255 scope global noprefixroute bxi0
       valid_lft forever preferred_lft forever
    inet6 fe80::4058:49ff:fe00:4/64 scope link
       valid_lft forever preferred_lft forever
[root@vm-bxi3 ~]# ping 192.168.10.23
PING 192.168.10.23 (192.168.10.23) 56(84) bytes of data.
64 bytes from 192.168.10.23: icmp_seq=1 ttl=64 time=0.462 ms
64 bytes from 192.168.10.23: icmp_seq=2 ttl=64 time=0.406 ms
64 bytes from 192.168.10.23: icmp_seq=3 ttl=64 time=0.408 ms
64 bytes from 192.168.10.23: icmp_seq=4 ttl=64 time=0.492 ms
^C
--- 192.168.10.23 ping statistics ---
4 packets transmitted, 4 received, 0% packet loss, time 3077ms
rtt min/avg/max/mdev = 0.406/0.442/0.492/0.036 ms
[root@vm-bxi3 ~]#
```

Figure 16 : Ping test between the BXI driver connected to Ethernet Nic and the distant “storage server”

We observe in Figure 16 that the test is successful. Due to the emulated connection with the Ethernet NIC, the ping latency measurement cannot be considered as a performance result.

3.5 Ethernet gateway testbed validation conclusion

The distribution of test coverage between the different sub-chapters is represented Figure 17.

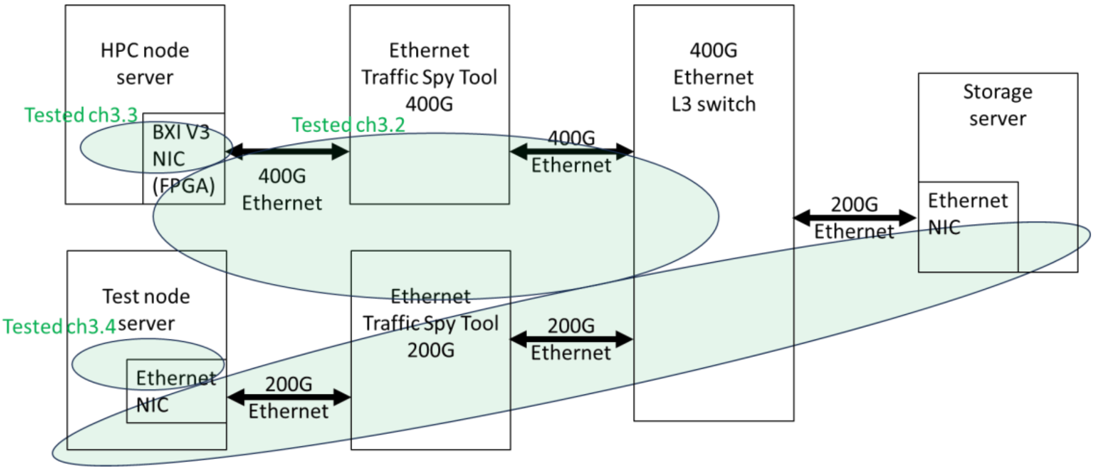


Figure 17 : Representation of demonstrator test coverage per sub-chapter

Tests presented in Chapter 3 check all the interfaces between the different modules of the demonstrator except the BXIv3 NIC RTL code for which only the UVM verification approach can be used for module validation. We deduce from these results that all unitary module behavior is correct. Consequently, the next chapter is dedicated to testing interactions between modules rather than individual module behavior.

4 BXI to Ethernet Gateway performance comparison between BXIv2 and BXIv3

In this chapter, we compare the Ethernet gateway latency between BXIv2 and BXIv3. Figure 18 shows the different setups used to perform the comparison:

- (A) shows the "DiBona" (Dibona is an Arm based HPC cluster prototype with BXIv2 interconnect from Mont-Blanc2020 project) BXIv2 elements that we used to perform the ping test, to determine the BXIv2 Ethernet gateway delay.
- (B) shows BXIv3 equivalent architecture. This configuration is not available for test but is represented for test model understanding.
- (C) shows the BXIv3 demonstrator we can use for the test. Note that the BXIv3 driver is used with The QEMU emulator to interface an off-the-shelf NIC Ethernet card.

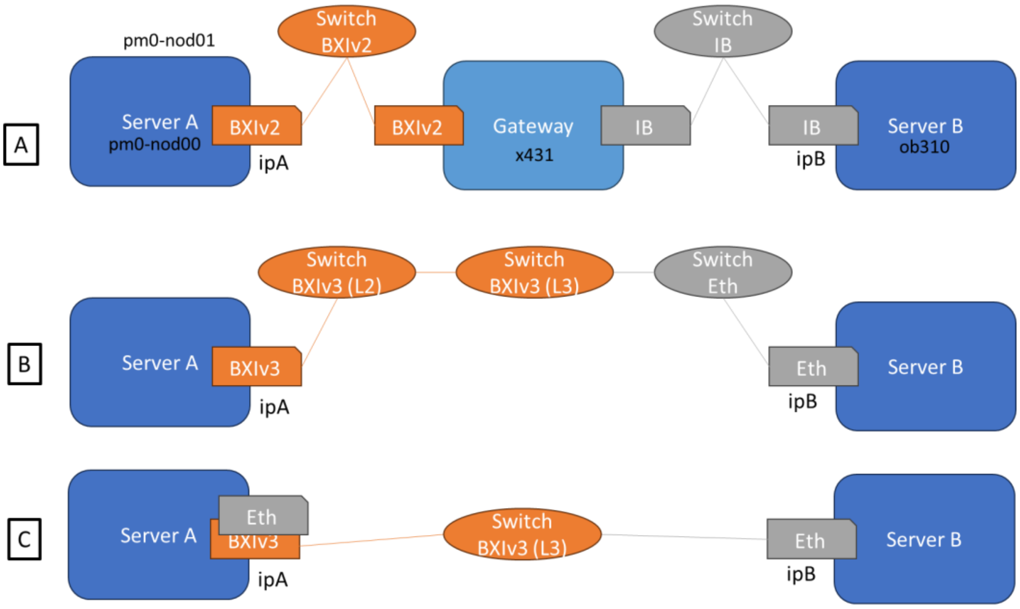
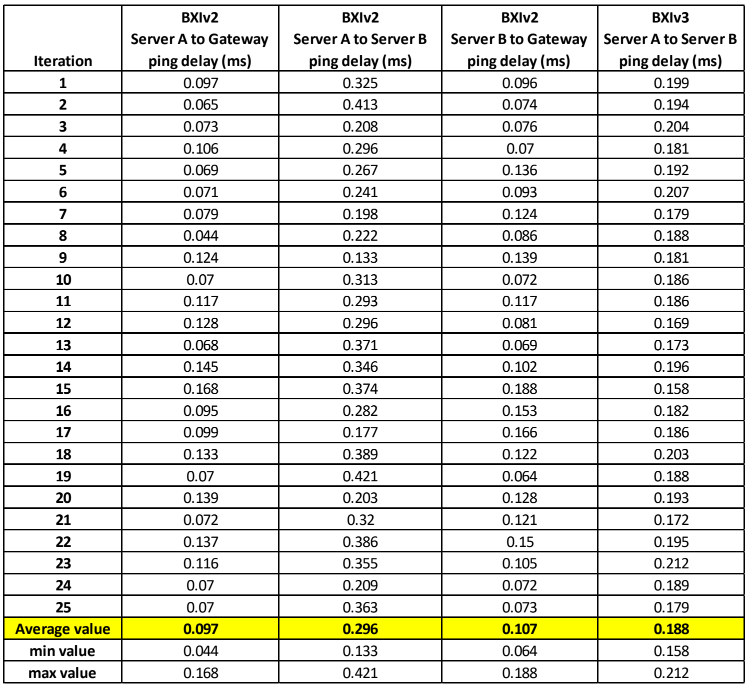


Figure 18 : BXI Ethernet gateway architecture comparison

4.1 BXIv2 gateway latency

Table 1 displays "ping" test results for the BXI Ethernet gateway. BXIv2 Ethernet gateway specific hardware allows ping test from server A (and respectively server B) to the Gateway with configuration (A).



Iteration	BXIv2 Server A to Gateway ping delay (ms)	BXIv2 Server A to Server B ping delay (ms)	BXIv2 Server B to Gateway ping delay (ms)	BXIv3 Server A to Server B ping delay (ms)
1	0.097	0.325	0.096	0.199
2	0.065	0.413	0.074	0.194
3	0.073	0.208	0.076	0.204
4	0.106	0.296	0.07	0.181
5	0.069	0.267	0.136	0.192
6	0.071	0.241	0.093	0.207
7	0.079	0.198	0.124	0.179
8	0.044	0.222	0.086	0.188
9	0.124	0.133	0.139	0.181
10	0.07	0.313	0.072	0.186
11	0.117	0.293	0.117	0.186
12	0.128	0.296	0.081	0.169
13	0.068	0.371	0.069	0.173
14	0.145	0.346	0.102	0.196
15	0.168	0.374	0.188	0.158
16	0.095	0.282	0.153	0.182
17	0.099	0.177	0.166	0.186
18	0.133	0.389	0.122	0.203
19	0.07	0.421	0.064	0.188
20	0.139	0.203	0.128	0.193
21	0.072	0.32	0.121	0.172
22	0.137	0.386	0.15	0.195
23	0.116	0.355	0.105	0.212
24	0.07	0.209	0.072	0.189
25	0.07	0.363	0.073	0.179
Average value	**0.097**	**0.296**	**0.107**	**0.188**
min value	0.044	0.133	0.064	0.158
max value	0.168	0.421	0.188	0.212

Table 1 : Ping test result with BXI Ethernet gateway

Combining these results with the end-to-end Server A to Server B ping delay measurement provides a way to calculate the Gateway added delay based on ping average values:
Gateway_delay = ServerA-ServerB_delay – ServerA-Gateway_delay – ServerB-gateway_delay
Gateway_delay = 0.296 – 0.097 – 0.107 = 0.092 ms.

From the BXIv2 switch specification, we know that the switch delay is in the range of 200 ns. It is consequently negligible compared to Server and gateway delays.

We deduce that the Gateway delay is approximately 100 µs.

4.2 BXIv3 Ethernet gateway latency

Concerning BXIv3, the (B) setup of Figure 18is not available because we have only one 400G Ethernet switch in the demonstrator. However, the BXIv3 prototype switch specification specifies a delay inferior to one microsecond, independently of L2 or L3 routing configuration, and out of congestion scenario. Consequently, the switch delay is negligible, and we can use scenario (C) to measure gateway delay.

BXIv3 uses an Ethernet switch (instead of a gateway server) which cannot answer ping requests. Consequently, we can only measure the end-to-end ping delay for BXIv3, which is measured to 188 µs in Table 1 and that we approximate to 200 µs.

4.3 BXI to Ethernet gateway delay conclusion

The above measurements conduct to following estimates:

- BXIv2 Ethernet gateway adds a measured 100 µs delay, which is eliminated with BXIv3 optimized architecture.
- “ping command” has a delay of approximately end-to-end delay of 200 µs independently of the use of a BXI (V2 or V3) or Ethernet driver.

We deduce that BXIv3 Ethernet gateway ping delay has a gain of one-third compared to BXIv2.

When we compare delay at the physical line only, BXIv3 delay estimate is inferior to 1us compared to the 100 µs BXIv2 delay.

5 BXI to Ethernet Gateway integration test

5.1 BXIv3 over IP RTL test results

This subchapter describes tests performed to check that the two NICIA external interfaces (PCIe and Ethernet) work successfully:

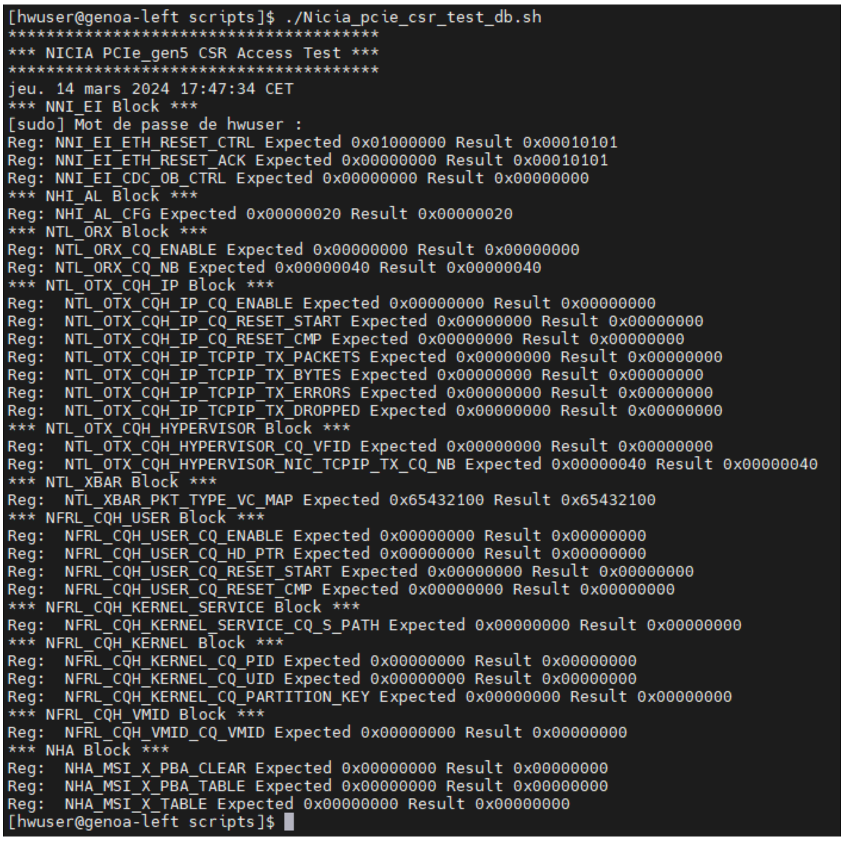
- The PCIe CSR test checks that the NIC configuration and status registers can be accessed successfully for write accesses and read accesses.
- The Ftile initialization test checks that the NIC can successfully start the Ethernet interface. “Ftile” is the name of Intel HIP block used for 400G Ethernet interface.

5.1.1 PCIe CSR access test

PCIe CSR access is the interface that the BXI software uses to drive the BXI NIC. Consequently, this test checks:

- that the Host processor of the HPC node accesses the FPGA prototyping board through PCIe,
- that the NICIA RTL design is loaded and started in the FPGA device,
- and that the NICIA RTL code successfully connects to the PCIe HIP block of the FPGA.

The successful result of this test is represented in Figure 19. Test has been shortened for figure visibility.



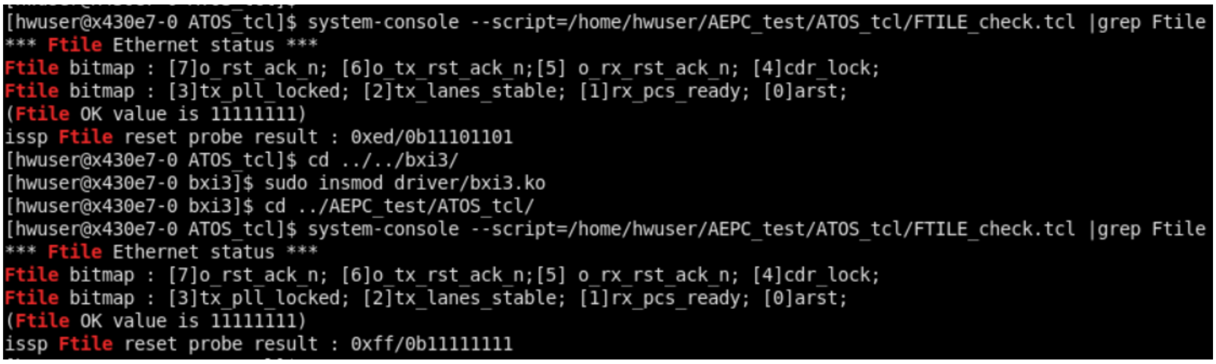
```
[hwuser@genoa-left scripts]$ ./Nicia_pcie_csr_test_db.sh
****************************************
*** NICIA PCIe_gen5 CSR Access Test ***
****************************************
jeu. 14 mars 2024 17:47:34 CET
*** NNI_EI Block ***
[sudo] Mot de passe de hwuser :
Reg: NNI_EI_ETH_RESET_CTRL Expected 0x01000000 Result 0x00010101
Reg: NNI_EI_ETH_RESET_ACK Expected 0x00000000 Result 0x00010101
Reg: NNI_EI_CDC_OB_CTRL Expected 0x00000000 Result 0x00000000
*** NHI_AL Block ***
Reg: NHI_AL_CFG Expected 0x00000020 Result 0x00000020
*** NTL_ORX Block ***
Reg: NTL_ORX_CQ_ENABLE Expected 0x00000000 Result 0x00000000
Reg: NTL_ORX_CQ_NB Expected 0x00000040 Result 0x00000040
*** NTL_OTX_CQH_IP Block ***
Reg:  NTL_OTX_CQH_IP_CQ_ENABLE Expected 0x00000000 Result 0x00000000
Reg:  NTL_OTX_CQH_IP_CQ_RESET_START Expected 0x00000000 Result 0x00000000
Reg:  NTL_OTX_CQH_IP_CQ_RESET_CMP Expected 0x00000000 Result 0x00000000
Reg:  NTL_OTX_CQH_IP_TCPIP_TX_PACKETS Expected 0x00000000 Result 0x00000000
Reg:  NTL_OTX_CQH_IP_TCPIP_TX_BYTES Expected 0x00000000 Result 0x00000000
Reg:  NTL_OTX_CQH_IP_TCPIP_TX_ERRORS Expected 0x00000000 Result 0x00000000
Reg:  NTL_OTX_CQH_IP_TCPIP_TX_DROPPED Expected 0x00000000 Result 0x00000000
*** NTL_OTX_CQH_HYPERVISOR Block ***
Reg:  NTL_OTX_CQH_HYPERVISOR_CQ_VFID Expected 0x00000000 Result 0x00000000
Reg:  NTL_OTX_CQH_HYPERVISOR_NIC_TCPIP_TX_CQ_NB Expected 0x00000040 Result 0x00000040
*** NTL_XBAR Block ***
Reg:  NTL_XBAR_PKT_TYPE_VC_MAP Expected 0x65432100 Result 0x65432100
*** NFRL_CQH_USER Block ***
Reg:  NFRL_CQH_USER_CQ_ENABLE Expected 0x00000000 Result 0x00000000
Reg:  NFRL_CQH_USER_CQ_HD_PTR Expected 0x00000000 Result 0x00000000
Reg:  NFRL_CQH_USER_CQ_RESET_START Expected 0x00000000 Result 0x00000000
Reg:  NFRL_CQH_USER_CQ_RESET_CMP Expected 0x00000000 Result 0x00000000
*** NFRL_CQH_KERNEL_SERVICE Block ***
Reg:  NFRL_CQH_KERNEL_SERVICE_CQ_S_PATH Expected 0x00000000 Result 0x00000000
*** NFRL_CQH_KERNEL Block ***
Reg:  NFRL_CQH_KERNEL_CQ_PID Expected 0x00000000 Result 0x00000000
Reg:  NFRL_CQH_KERNEL_CQ_UID Expected 0x00000000 Result 0x00000000
Reg:  NFRL_CQH_KERNEL_CQ_PARTITION_KEY Expected 0x00000000 Result 0x00000000
*** NFRL_CQH_VMID Block ***
Reg:  NFRL_CQH_VMID_CQ_VMID Expected 0x00000000 Result 0x00000000
*** NHA Block ***
Reg:  NHA_MSI_X_PBA_CLEAR Expected 0x00000000 Result 0x00000000
Reg:  NHA_MSI_X_PBA_TABLE Expected 0x00000000 Result 0x00000000
Reg:  NHA_MSI_X_TABLE Expected 0x00000000 Result 0x00000000
[hwuser@genoa-left scripts]$ █
```

Figure 19 : Pcie CSR access test result on the HPC node

5.1.2 BXI driver initialization with NICIA

This test consists of loading the BXI driver module in the HPC node operating system. During this operation, the BXI driver uses the PCIe interface to initialize NICIA concerning the Ethernet interface.

Figure 20 reports the test scenario: the ETS tool is first used to check that the Ethernet link is not ready at the start by checking the Ftile HIP status values: we see that the receiver information (cdr_lock and rx_pcs_ready) are not ok, then the BXI module is loaded in the Linux kernel. After this, we observe that the Ftile values have evolved to the expected values. This “Ftile_OK” status corresponds to the “link-up” information of a Linux system.



```
[hwuser@x430e7-0 ATOS_tcl]$ system-console --script=/home/hwuser/AEPC_test/ATOS_tcl/FTILE_check.tcl |grep Ftile
*** Ftile Ethernet status ***
Ftile bitmap : [7]o_rst_ack_n; [6]o_tx_rst_ack_n;[5] o_rx_rst_ack_n; [4]cdr_lock;
Ftile bitmap : [3]tx_pll_locked; [2]tx_lanes_stable; [1]rx_pcs_ready; [0]arst;
(Ftile OK value is 11111111)
issp Ftile reset probe result : 0xed/0b11101101
[hwuser@x430e7-0 ATOS_tcl]$ cd ../../bxi3/
[hwuser@x430e7-0 bxi3]$ sudo insmod driver/bxi3.ko
[hwuser@x430e7-0 bxi3]$ cd ../AEPC_test/ATOS_tcl/
[hwuser@x430e7-0 ATOS_tcl]$ system-console --script=/home/hwuser/AEPC_test/ATOS_tcl/FTILE_check.tcl |grep Ftile
*** Ftile Ethernet status ***
Ftile bitmap : [7]o_rst_ack_n; [6]o_tx_rst_ack_n;[5] o_rx_rst_ack_n; [4]cdr_lock;
Ftile bitmap : [3]tx_pll_locked; [2]tx_lanes_stable; [1]rx_pcs_ready; [0]arst;
(Ftile OK value is 11111111)
issp Ftile reset probe result : 0xff/0b11111111
```

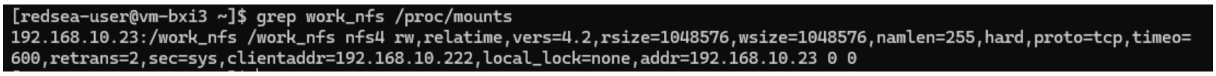
Figure 20 : BXI driver first test with NICIA

This test demonstrates the first successful communication between the BXIv3 driver and the BXIv3 NIC RTL.
The next step (currently in the debug phase) is to obtain a successful Linux “ping” command on the HPC node to join the storage server. This will demonstrate the first end-to-end BXIv3 communication involving the BXI driver and NIC RTL developed in the project.

5.2 End-to-end Ethernet gateway demonstrator test

This test demonstrates the usage of BXIv3 protocol for NFS service between a “HPC test node” and a “storage server” through a gateway, which creates communication between the BXI HPC interconnect network and a data center IP-over-Ethernet network. The overall test path is represented Figure 1 at the beginning of this document.
This test uses a virtualized NIC based on QEMU in the HPC node.

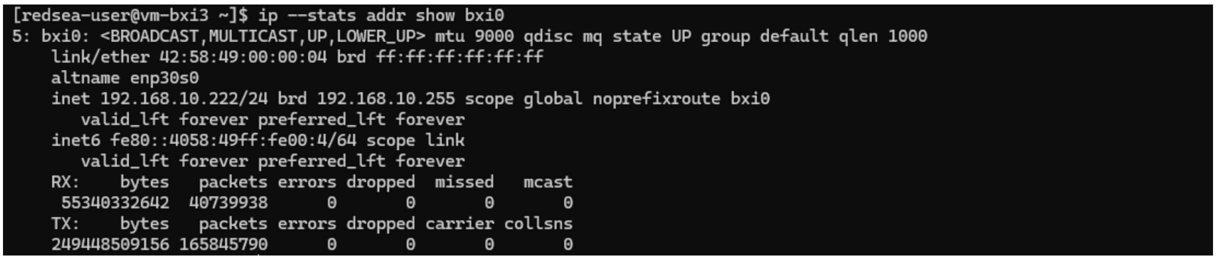
On the HPC test node (in the virtual machine with the emulated BXI NIC device), the “work_nfs” NFS filesystem has been mounted on the “/work_nfs” mount point. The NFS server is the storage server named x430-1 with a 192.168.10.23 IP address.



```
[redsea-user@vm-bxi3 ~]$ grep work_nfs /proc/mounts
192.168.10.23:/work_nfs /work_nfs nfs4 rw,relatime,vers=4.2,rsize=1048576,wsize=1048576,namlen=255,hard,proto=tcp,timeo=
600,retrans=2,sec=sys,clientaddr=192.168.10.222,local_lock=none,addr=192.168.10.23 0 0
```

Figure 21 : Mounted NFS filesystem on the “HPC node”

The NFS communications are using TCP/IP protocol and goes through the bxi0 interface. The “ip” linux command line tool can be used to display interface status and attributes.

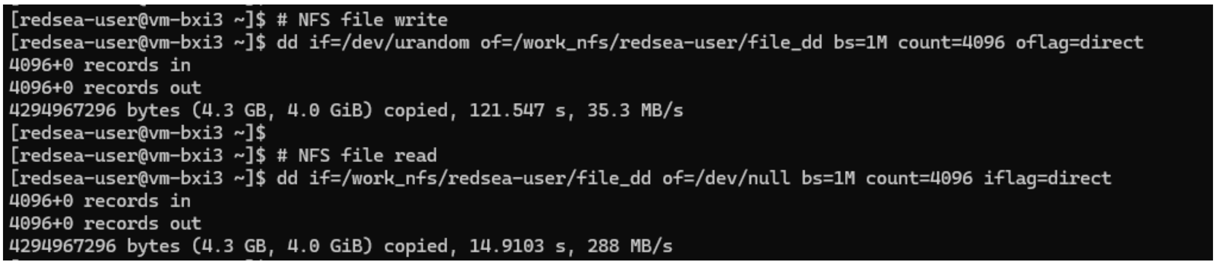


```
[redsea-user@vm-bxi3 ~]$ ip --stats addr show bxi0
5: bxi0: <BROADCAST,MULTICAST,UP,LOWER_UP> mtu 9000 qdisc mq state UP group default qlen 1000
    link/ether 42:58:49:00:00:04 brd ff:ff:ff:ff:ff:ff
    altname enp30s0
    inet 192.168.10.222/24 brd 192.168.10.255 scope global noprefixroute bxi0
       valid_lft forever preferred_lft forever
    inet6 fe80::4058:49ff:fe00:4/64 scope link
       valid_lft forever preferred_lft forever
    RX:    bytes   packets errors dropped  missed   mcast
     55340332642  40739938      0       0       0       0
    TX:    bytes   packets errors dropped carrier collsns
    249448509156 165845790      0       0       0       0
```

Figure 22 : BXI Communication interface on the HPC node

The test consists of NFS filesystem accesses from the test node:

- write a 4 GiB file into the NFS filesystem at /work_nfs/redsea-user/file_dd
- read a 4 GiB file from the NFS filesystem at /work_nfs/readsea-user/file_dd



```
[redsea-user@vm-bxi3 ~]$ # NFS file write
[redsea-user@vm-bxi3 ~]$ dd if=/dev/urandom of=/work_nfs/redsea-user/file_dd bs=1M count=4096 oflag=direct
4096+0 records in
4096+0 records out
4294967296 bytes (4.3 GB, 4.0 GiB) copied, 121.547 s, 35.3 MB/s
[redsea-user@vm-bxi3 ~]$
[redsea-user@vm-bxi3 ~]$ # NFS file read
[redsea-user@vm-bxi3 ~]$ dd if=/work_nfs/redsea-user/file_dd of=/dev/null bs=1M count=4096 iflag=direct
4096+0 records in
4096+0 records out
4294967296 bytes (4.3 GB, 4.0 GiB) copied, 14.9103 s, 288 MB/s
```

Figure 23 : NFS test command on the HPC node

In parallel with the test, the “Performance Co-pilot” monitoring tool (https://pcp.io/) is used to harvest system data, including network interface data. The recorded data is then graphically presented with the “Grafana” visualization tool (https://grafana.com/). The “RED-SEA WP2 Demo” dashboard displays four metrics:

- Network throughput (In): reception bandwidth.
- Network throughput (Out): transmission bandwidth.
- Network packets (In): reception packet rate.
- Network packets (Out): transmission packet rate.

We observe bandwidth and packet rate increase on the bxi0 interface during the file write (2 graphs on the right) and during the file read (2 graphs on the left). The transmission bandwidth reaches 40 MiB/s during file write. The reception bandwidth reaches 300 MiB/s during file reading. The performance is currently limited by the emulated BXI NIC device.

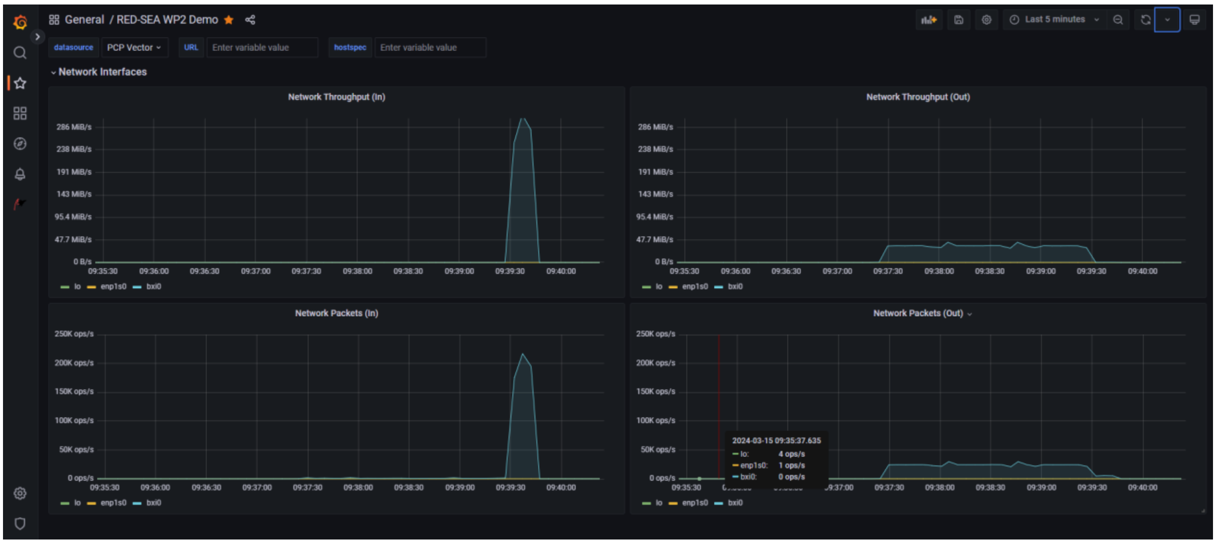


Figure 24 : NFS test network traffic visualization

5.3 BXI to Ethernet gateway integration test conclusion

BXIv3 integration test performed in this chapter demonstrates the validity of the Ethernet Gateway Hardware Architecture Specification (RED-SEA deliverable 2.1) concerning the enhancement of the BXI to Ethernet bridging function.

Some future tests with this same demonstrator will include the BXIv3 NIC RTL code in the NFS test application. In this context, the Ethernet gateway’s full 400Gbps bandwidth can be demonstrated.

6 Conclusion

This document describes the Ethernet gateway testbed as defined in the latest project amendment. However, some additional modules have been used to improve debug capabilities and facilitate tests. The Demonstrator has been fully tested and its usage shows a functional BXiV3 gateway.

Removal of BXI to IP/Ethernet server gateway with the method described in D2.1 has been demonstrated. Consequently, we obtain the 10k euro cost reduction (corresponds approx. to BXIv2 Ethernet gateway server price) as targeted by KPI7.

Moreover, most of the D2.1 [1] concepts have been demonstrated with tests performed in chapters 3 and 5 This illustrates the BXIv3 capability to implement an efficient Modular Supercomputing architecture with 400 Gbps bandwidth, which is four times more than BXIv2, thus reaching KPI2. Another important gain is the latency decrease by eliminating the BXIv2 gateway server delay (approximately 100 µs) which allows to have only a switch delay (inferior to 1 µs) when we connect two BXI Fabrics as described in the modular supercomputing architecture.

Eviden plans to continue using this demonstrator to validate BXIv3 maturity progress beyond the RED-SEA project. One important next step-item is to demonstrate we approach the 400 Gbps bandwidth for the BXIv3 Ethernet gateway.

7 Acronyms and Abbreviations

Term	Definition
ASIC	Application Specific Integrated Circuit
CSR	Control and Status Register
ETS	Ethernet Traffic Spy
FEC	Forward Error Correction
FPGA	Field Programmable Gate Array
FIFO	First-In-First-Out buffer
HAS	High level Architecture Specification
HPC	High Performance Computing
HIP	Hardware Intellectual Property (block)
IP	Internet Protocol
IP-Offload	Internet Protocol Offload
JTAG	Joint Test Action Group
MAC	Media Access Control
MSA	Modular Supercomputing Architecture
NIC	Network Interface Controller
PCS	Physical Coding Sublayer
PHY	Ethernet IP PHYsical layer which contains SerDes
RTL	Register Transfer Level
SW	Software: generally, it means the application or the driver executing on the host

Table 2: Acronyms and Abbreviations

8 Bibliography

[1] RED-SEA D2.1 High level Architecture Specification (HAS) of the Ethernet Gateway IP

[2] Intel® Xeon® Platinum 8480+ Processor specification

[3] REFLEX CES XpressSX AGI-FH400G Presentation of the Intel® Agilex™ board

[4] Broadcom Tomahawk 5 / BCM78900 Series Overview