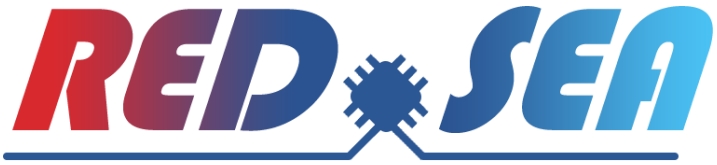
Network Solution for Exascale Architectures



D6.5: Final report

Document Properties

Contract Number	955776
Contractual Deadline	M36 (31/03/2024)
Dissemination Level	Public
Nature	Report
Edited by:	Claire Chen (BULL/Eviden); Pascale Bernier-Bruna (BULL/Eviden)
Authors	All partners
Reviewers	
Date	31/03/2024
Keywords	Interconnect networks; NIC; Switch; BXI
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	15/02/2024	C. Chen, BULL	ToC definition
0.2	05/03/2024	C. Chen, BULL	Ready for internal review
0.9	29/03/2024	C.Chen, BULL	Corrections after feedback of internal review
1.0	29/03/2024	C.Chen, BULL	Ready for submission

Table of Contents

List of Figures

List of Tables

1 Introduction

The upcoming generation of Exascale systems is heavily reliant on a streamlined network infrastructure. This network must be capable of accommodating massively parallel processing systems, consisting of hundreds of thousands of nodes and millions of cores. It should offer a range of functionalities to enable applications to scale effectively at Exascale and beyond, while also being adaptable for power-efficient accelerators and computing units. Furthermore, it should support a wide array of prevalent and emerging data-centric and AI-driven applications

In order to enable Exascale computing, next generation interconnection networks must scale to hundreds of thousands of nodes, and must provide features to also allow the HPC, HPDA, and AI applications to reach Exascale, while benefiting from new hardware and software trends.

The RED-SEA consortium has been dedicated to address the goal by leveraging key European expertise and background, including BullSequana eXascale Interconnect (BXI), the key production-proven European Interconnect, as well as results from a number of EU-funded projects on interconnects and HPC systems.

The RED-SEA project has been actively engaged in various aspects of European Exascale interconnect technologies to facilitate the development of the next generation of European Exascale interconnects, which includes preparations in the BXI technology. Specifically, the project has addressed the following aspects:

- Specification of the new architecture through hardware-software co-design, focusing on a set of representative applications from the realms of HPC, HPDA, and AI.
- Testing, evaluation, and implementation of new architectural features across multiple levels, including mathematical analysis, modelling, simulation, and FPGA-based implementations.
- Development of a high-performance, low-latency Ethernet gateway to facilitate seamless communication within and between resource clusters.
- Implementation of efficient network resource management to enhance congestion control, virtualisation, adaptive routing, and collective operations.
- Exploration of the BXI ecosystem to accommodate a variety of applications and hardware, with improvements in end-to-end network services such as programming models, reliability, security, low-latency, and support for new processors.
- Utilisation of open standards and compatible APIs to create innovative reusable libraries and solutions for Fabrics management.

After 36 months project lifespan, RED-SEA has successfully attained all objectives outlined in the Description of Actions (DoA). This report endeavours to outline the main scientific and technological results of the project. One significant outcome lies in the advancement of the European Interconnect network BXI, particularly in improving the current version (BXIv2) and preparing for its next generation (BXIv3). Another achievement of the project is its contribution to new, efficient network resource management schemes. These advancements enhance congestion control, virtualisation, adaptive routing, and collective operations. Additionally, the project has extended the BXI ecosystem, aiming to expand the applicability of the interconnect to various applications and hardware. Moreover, enhancing the tools and simulators chosen by the RED-SEA project is crucial for these achievements. Finally, several additional developments have been undertaken to enhance European technology Intellectual Properties (IPs) within the field of interconnect networks. These results will be detailed in the paragraph 3 of this document.

The current document represents one of the latest documents from the RED-SEA project. Its aim is to present the project's main outcomes in a manner easily understandable to the public, with a focus on exploitation and impact aspects. For a comprehensive understanding of the project's overall results, readers are encouraged to review deliverable D1.4 titled "Report on holistic evaluation of RED-SEA network technologies".

Beginning with a summary of the project objectives in paragraph 2, it proceeds to describe the main results in paragraph 3. Following this, paragraph 4 delves into the impact generated by the project, along with general findings and lessons learned.

2 Objectives

The RED-SEA's overall objective was to prepare a next-generation European Interconnect capable of supporting forthcoming EU Exascale systems. This involved initiating an economically feasible and technologically efficient interconnect, leveraging European interconnect technology (BXI) alongside standard and mature technologies like Ethernet. Furthermore, the project drew on previous EU-funded initiatives such as ExaNeSt, EuroEXA, and the European Processor Initiative (EPI), as well as open standards and compatible APIs.

In particular, the RED-SEA project triggered the third generation of the BXI interconnect, BXIv3, contributing to its roadmap by:

- defining the architecture blueprint and the corresponding simulation models;
- designing the new building blocks (Intellectual Properties) necessary to address the new challenges of modular supercomputers;
- delivering initial proof-of-concept demonstration of its critical components; and
- developing the ecosystem and creating a broader community of users and developers combining Research and Industrial teams.

Despite RED-SEA being a Research and Innovation Action (RIA) project, over 50% of the technologies developed within the project have been integrated into either the BXI roadmap or other commercial product roadmaps. Moreover, the RED-SEA has laid the groundwork for the initial releases of BXIv3. We have ambition to develop and industrialise BXIv3 and its subsequent version through follow-up funding projects.

3 Key results for exploitation

The RED-SEA project has yielded numerous advancements in both scientific and technological domains. In this section, we aim to spotlight a selection of the results for future exploitation.

The selected results have been categorised as follows:

- 3.1: **BXI: European interconnect technologies**: Highlights advancements in both the next-generation BXI (BXIv3) and enhancements for the current BXI version (BXIv2).
- 3.2: **Optimised network resource management**: Details advancements in network resource management.
- 3.3: **Extension of the BXI ecosystem**: Presents key outcomes related to expanding the BXI ecosystem.
- 3.4: **Tools and simulators**: Describes improvements and enhancements made to tools and simulators.
- 3.5: **Collateral developments**: Outlines additional developments undertaken within the RED-SEA project.

3.1 BXI: European interconnect technologies

BXI (BullSequana eXascale Interconnect) is Bull's initiative aimed at developing an interconnect for High-Performance Computing (HPC). Commencing at the end of 2010, the first NIC and switch ASICs (BXIv1) introduced in 2017, followed by BXIv2 in 2021.

BXI interconnect is made of two components: the NIC that allows the servers to access to the network, and the switch that establishes communication between the servers, as illustrated in *Figure 1* and *Figure 2*.

Figure 1: BXIv2 switch

Figure 2: BXIv2 NIC standard PCIe card

The RED-SEA project's most significant outcomes lie in its contribution to BXIv3 research, specification, partial design, and prototyping of some key components (e.g. Ethernet gateway). As a result, these achievements will be exploited in BXIv3. Moreover, two patents with the results relating to BXIv3 have been filed in the second half of the project lifespan. At the point of submitting the deliverable, the patents are still in the instructional stage.

Additionally, some optimisations have been made to enhance the performance of the current version of BXI, BXIv2.

The RED-SEA project has laid the groundwork for the development, implementation, and industrialisation of BXIv3. This foundation will be further advanced through the potential subsequent project, NET4EXA, which presents a proposal to the EuroHPC JU interconnect call (HORIZON-EUROHPC-JU-2023-INTER-02-01), if NET4EXA is accepted by the JU.

3.1.1 BXIv3 Fabric

A BXI fabric, a BXI network, aims to interconnect nodes with CPUs and/or GPUs and allow connectivity with an Ethernet Network. The adoption of Ethernet as the underlying link technology for BXIv3 marks a significant departure from the proprietary link protocol used in BXIv2. This shift allows for native

support of both traditional IP/Ethernet traffic and high-performance BXI/Portals traffic, as well as facilitating the federation of clusters through L3 Internet Protocol routers.

One of the main achievements carried out in the RED-SEA project is the accomplishment of the High-level Architecture Specification (HAS) of the BXIv3 fabric. The main target characteristics of BXIv3 fabric described in the HAS are summarised below.

A BXIv3 fabric is composed of a set of BXIv3 Network Interface Controllers (NIC) and BXIv3 Switches:
- A BXIv3 NIC is the entry points of the fabric, it connects a CPU or GPU to the BXI network. The NIC is connected to a CPU/GPU through PCIe links.
- A BXIv3 Switch has 64 identical ports that can be connected to NICs or Ethernet Layer 3 Switches.

One important point to note is that BXIv3 is not intended to replace Ethernet. Instead, it serves as a complementary technology, seamlessly interfacing with Ethernet. Ethernet offers an open ecosystem, complete flexibility in configuration, the ability to configure chaotic or unpredictable topologies, a high degree of heterogeneity in servers or network components, and the capability to address classic enterprise system requirements. By complementing Ethernet, BXIv3 can leverage the advantages of Ethernet mentioned above.

The main target characteristics of BXIv3 Switch include 64 ports (vs 48 ports in BXIv2); 400 Gb/s per port (x2 vs RED-SEA proposal). As the objective of BXIv3 is to integrate an Ethernet gateway into each switch port to provide a cost-effective solution with improved performance, the implementation of a standard Ethernet physical layer and MAC is mandatory. Additionally, other specifications have also been elaborated, such as Virtual Channels, VLAN, ports aggregation and congestion reporting necessary for BXIv3 Switch. Furthermore, the expected communication models for BXIv3 have been described in the HAS. Lastly, a brief overview of the frame format planned for use in BXIv3 has been provided.

This HAS is the groundwork for BXIv3 Switch which will be further exploited in other EU funded projects, such as EUPEX.

Final, future endeavours to advance this initiative will be pursued in subsequent projects, potentially supported by funding from EuroHPC JU, such as the NET4EXA project.

3.1.2 BXIv3 NIC

NICIA is the code name for the NIC of BXIv3. The goal of NICIA is to support the creation of high-performance computing (HPC) interconnects with up to hundreds of thousands nodes, running MPI on top of Portals [1] interface. In addition, much of network interface processing is offloaded to the NIC to achieve the features, such as: i) zero-copy and application-bypass: a data transfer process where information is directly read and written in the user space, bypassing any intermediate buffering in either user or kernel space; ii) OS-bypass: commands are sent to NIC by application without the need to go through the kernel. In addition to offloading the Portals programming interface to support MPI, OFI and PGAS, the NIC will also support offloading of IP (Internet Protocol) for Ethernet communications with dedicated path and resources to achieve high performance for UDP or TCP traffic.

Figure 3 shows the system view where NICIA is connected to the host through PCIe gen5 and connected to the switch using Ethernet protocol. On the host the NIC is accessed by the driver or directly the Portals library. NIC also reads and writes data in memory through DMA over PCIe.

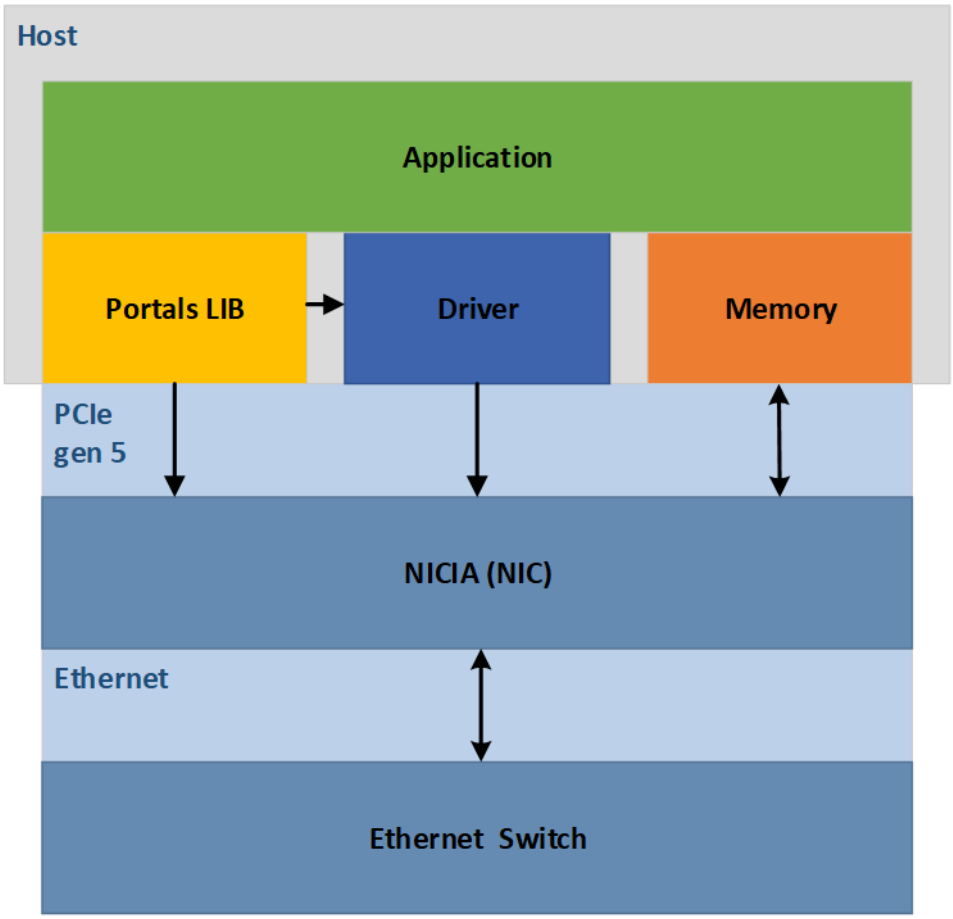


Figure 3: System view of NICIA

In order to achieve flexibility of the design, the NIC is implemented on an Agilex FPGA card, which contains FPGA fabric and a hard processing system (HPS).

The key results of the BXIv3 NIC achieved in the RED-SEA are depicted in the following subparagraphs. They will be exploited in European funded projects, such as the EUPEX project.

BXI3 NIC IP offload transport layer design

Following the high-level specification of the BXIv3 transport layer, the transport layer has also been designed to be integrated in BXIv3 NIC (Network Interface Card). Here, the emphasis is placed on offloading TCP/IP to the BXIv3 NIC.

Offloading TCP/IP to the Network Interface Controller (NIC) involves delegating the processing of TCP/IP protocol stack functions from the system's CPU to the NIC itself. By offloading TCP/IP processing to the NIC, the system's CPU is freed from handling these network-related tasks, enabling it to focus on other critical processing tasks. This can result in improved system responsiveness, reduced latency, and increased overall throughput, particularly in network-intensive applications.

NICIA facilitates TCP/IP offloading through a dedicated pathway and allocated resources. By utilising dedicated resources and pathways, NICIA enhances the throughput and responsiveness of UDP and TCP traffic, contributing to overall system efficiency and network performance.

The implementation of this feature was carried out in the RED-SEA project, especially in WP2 where the development of both the hardware dedicated to TCP/IP offload and its associated software, Ethernet Linux driver, has been delivered.

BXI3 Ethernet driver

The Ethernet driver is designed to achieve high levels of parallelism through independent resources and a zero-copy mechanism, aiming to meet targeted performance benchmarks. It uses state of the art Linux APIs to avoid locking in transmit and reception paths and avoid any unnecessary copying of data. Compatible with both x86 and ARM architectures, it has been successfully designed and developed. Additionally, it has been seamlessly integrated into the Linux network stack and administrative tools of BXIv3.

Ethernet gateway

One of the objectives of the RED-SEA project is to focus on the Ethernet gateway of BXIv3. The Ethernet gateway solution of BXIv3 involves utilising L3 Ethernet switches as gateways enabling connectivity and interoperability with existing Ethernet devices, representing a significant enhancement over previous versions of BXI (v1 and v2), which relied on a gateway server. The Ethernet gateway designed and developed for BXIv3 offers enhanced bandwidth and reduced latency for communications between the two networks. Moreover, employing switches instead of servers facilitates superior performance scalability and lowers investment and exploitation costs.

A series of tasks related to the Ethernet gateway solution have been accomplished. It commenced with the development of the high-level specification for the Ethernet gateway. Subsequently, an initial version of the Register Transfer Level (RTL) code was generated, encompassing all interfaces and capable of transferring basic messages.

Following the design phase and complete verification of the Ethernet gateway, the final RTL for an FPGA was developed. Lastly, the preparation for testing and validating the connectivity between the compute node in the HPC network fabric and the service node in the Ethernet backbone network involved development of an Ethernet gateway prototype.

The components of the Ethernet gateway prototype are illustrated in Figure 4. This prototype comprises:

- a compute node server with the RED-SEA Ethernet device and IP / Ethernet driver,
- a L3 Ethernet switch used as the gateway,
- a service node server with an off-the-shelf Ethernet device,
- two traffic spy FPGAs have been added to capture Ethernet frames that come in and out of the switch.

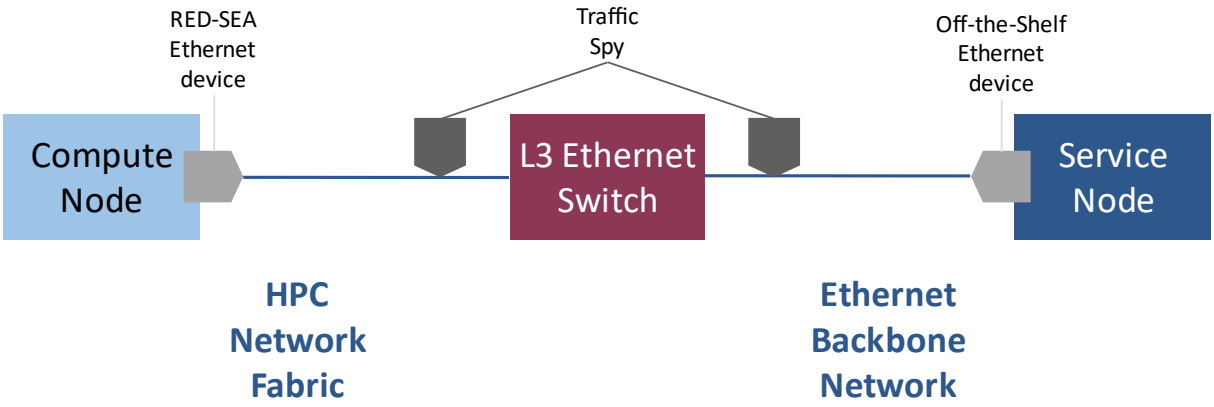


Figure 4: Ethernet gateway prototype

As of writing this report, the prototype employs an emulated Ethernet device within a virtualisation environment on the compute node. The basic connectivity tests, such as Internet Control Message Protocol (ICMP) echo request and echo response, have been successfully conducted. Once the bring-up activity is finalised, this emulation will be replaced by the RED-SEA Ethernet hardware device.

This prototype contributes to the test plan of BXIv3.

3.1.3 BXIv3 Patents

Two patents concerning BXIv3 have been developed and filed by BULL in the context of the RED-SEA project. However, since they are still in the instructional stage and have not been published, the specific details of their content cannot be disclosed at this time. The overview of these two patents is summarised below.

Method and system for intra- and inter-cluster communication

In packet transfers between two nodes, various issues can cause delays in the packets, such as a link may be down, the destination node may be down, or congestion may occur at the destination node. These issues must be distinguished to handle them appropriately. It is essential to specifically detect each issue to implement the appropriate reaction. This patent proposes methods to address the issues in both intra- and inter-cluster communication.

This patent has been filed in Europe (application EP24305074.7).
The method described in this patent will be implemented in BXIv3.

System and method for managing packet transmission issues in High-Performance Computers

Several package transmission challenges exist in HPC, including identifying endpoints during doing Portals communications across multiple clusters connected via an Ethernet backbone network, routing messages across multiple clusters during Portals communications, and avoiding the need for router nodes to cross cluster-backbone and backbone-cluster networks, which can introduce cost and performance bottlenecks. This patent proposes methods to address these package transmission issues.

This patent has been filed under number 23307012.7.
The method described in this patent will be implemented in BXIv3.

3.1.4 BXIv2 improvements

BXIv2 comes with a software stack that enables system services and user applications to exploit BXI hardware components (Figure 5).

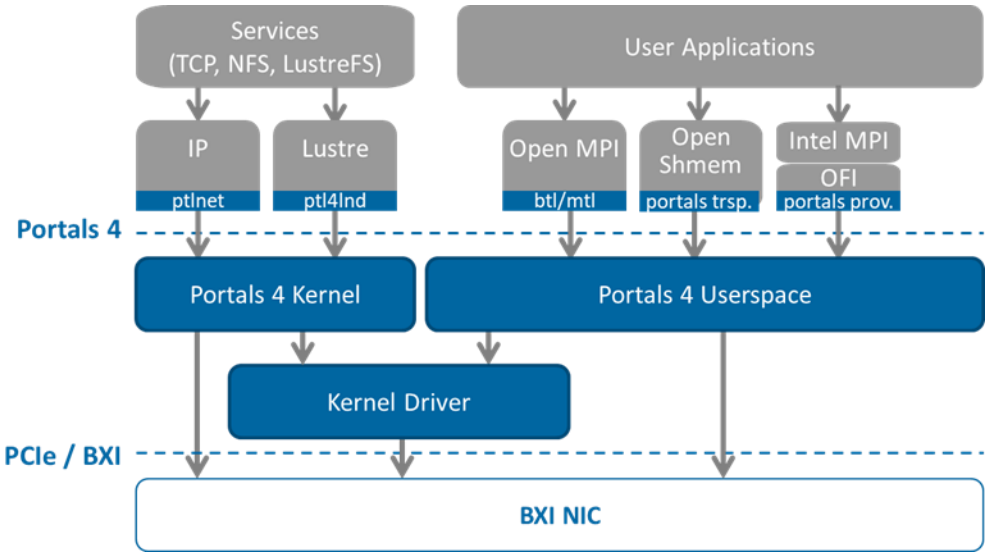


Figure 5: BXI software (compute) stack

Applications may offload most of their communication operations onto the BXI NIC with minimal processing of the host CPU, by using the Portals4 user-space API directly (bxi-portals library). Alternately, applications may seamlessly leverage existing socket-based communication model using Ptlnet software component, which exposes a network interface (e.g ptl0) similar to usual ethernet

devices. Any applicative IP traffic routed to/from ptl0 will be processed by Ptlnet component (IP over BXI).

Ptlnet is a Data Link Layer (L2) Linux network device (netdevice) over Portals4 and as such allows any Network Layer (L3) protocol such as IP (Internet Protocol) to be transmitted over BXI network by the Linux kernel. In order to leverage existing tools such as ifup, ifdown, iptools, tcpdump and ethtool, the L2 frame format has been chosen to be the same as the Ethernet one.

We have worked on the performance of the Linux kernel module (ptlnet) that implements the Internet Protocol (IP) over BXIv2. Several performance optimisations have been identified and prioritised. Three performance optimisations of the IP over BXI kernel module have been developed: simultaneous use of multiple send/receive communication queues, transmission and reception scatter-gather, segmentation and checksum offloading. The optimised module has been evaluated with a bandwidth performance campaign that covered a large range of message sizes, TCP and UDP upper protocol, monodirectional or bidirectional traffic, and IP gateway additional component. The results show significant bandwidth improvements both for TCP and UDP traffic, allowing to reach 70 to 80 Gbits/s throughput.

These optimisations have been included in the BXI Compute Stack as of BXIv2. Therefore, BXIv2 has already got benefits from these improvements.

Information regarding certain benchmark results for BXIv2 is available in a blog post accessible via the RED-SEA website [2]: https://redsea-project.eu/peek-bxi/.

3.2 Optimised network resource management

The challenge extends beyond merely offering a high-bandwidth connection and achieving theoretical minimal latency; it lies in attaining real-world performance when numerous diverse applications are concurrently utilising the system.
To be efficient, the interconnect network must be able to maintain balanced traffic flows, avoid congestions when multiple workloads are competing for the same resources, and route traffic through the optimal path, while overcoming hardware failures.

RED-SEA has been tackling these challenges and made results in the following topics that will be detailed in the following sub-paragraphs:

- Optimised realisations of MPI-based collective communication primitives,
- Provision of differentiated services,
- Optimized Congestion Management.

3.2.1 Optimised realisations of MPI-based collective communication primitives

In the context of the RED-SEA project with extreme computing and data components, parallel applications will generate large congestion due to collective operations. This means that improvements in the performance of these operations will have a very favourable impact on the performance of the applications that use them.

In the RED-SEA project, the impact on the network of collective communication primitives (CCPs) was firstly characterised, followed by the implementation and evaluation of optimisation mechanisms to enhance their performance. As a conclusion, it was found that the network performance is heavily influenced by the collective primitives (both type and frequency) involved in each workload. More precisely, the communication patterns defined by the collective primitives of the active workloads largely determine the network performance.

The software and hardware optimisations have been realised for the MPI-based collective communication primitives improving the performance of the data exchanges over the network.

On the software side, the following techniques in combination with three well-known realisations of the MPI standard, namely OpenMPI, MPICH and IntelMPI have been analysed:

- A performance evaluation of the blocking global reduction (blocking) MPI_Allreduce versus its non-blocking counterpart MPI_Iallreduce immediately followed by the corresponding blocking synchronization MPI_Wait.
- An analysis of two realizations of MPI_Iallreduce that offer significantly higher performance when applied to perform a blocking global reduction. These implementations operate by dividing the message (transparently to the user) into either a collection of messages of a specific smaller size or a fixed number of smaller messages, in both cases pipelining the transfers.
- A demonstration that splitting the messages to pipeline the transfers benefits not only the global reduction primitive but also other collective operations, such as broadcast or reduce-scatter.
- A study of the benefits of an optimal MPI processes to cores mapping. The mapping policies may offer a significant performance loss for some algorithms because they are not aware of where the processes are mapped.

On the hardware side, the primitives that we are interested in optimizing have been identified: the multicast primitives are the ones that most contribute to create congestion in the network. The proposed hardware optimization has been based on using multicast routing in the network switches. This means that packets that would share the same path or a part of the path to their respective destinations are replaced by a single packet and when these paths separate them the corresponding packet is duplicated and sent through both sub paths.

As results, taking into consideration the network topology in the implementation of CCPs can bring important benefits on the CCP performance and therefore in the application performance.

A blog post explaining optimisations for collective communications primitives (CCPs) has been available on the RED-SEA website [2]: https://redsea-project.eu/optimizations-for-collective-communications-primitives-ccp/ .

3.2.2 Provision of differentiated services

The virtualisation and QoS provision mechanisms for BXIv3 have been dealt with in the RED-SEA project. Moreover, these mechanisms have been integrated in the SAURON simulator that has been optimised as well.

The virtualisation is about interference measured with both application and synthetic traffics and with several applications running concurrently. The mechanism to reduce the interference was evaluated.

QoS (Quality of Services) consists of taking into account the mechanisms that BXIv3 incorporates for provision applications with differentiated services (i.e., three classes of traffic are defined to be generated at the hosts using BXIv3 NICs). New mechanisms for provision of QoS have been developed. These mechanisms include an arbitration table at every output port of BXIv3 NIC and Switch. Each table defines a set of entries, containing a weight that fixes the maximum amount of information that each service level or virtual channel can send.

The results of the RED-SEA project have been proposed to three projects Spanish national or regional projects to be re-used in these projects.

3.2.3 Optimised Congestion Management

Modern high-speed networks shift away from traditional TCP-based communication to adopting RDMA transfers in order to achieve lower latency. These networks need hardware-based congestion management in order to deal with saturation trees.

A new congestion control scheme named Accurate has been advanced and evaluated. The Accurate congestion control computes and assigns exact max-min fair rates to network flows, without relying on costly per-flow states inside the network, the mechanism relies on a simple hardware block in front of network links, with minimal latency and hardware cost overheads. Accurate was the result of the EU previously funded projects, such as the ExaNeSt EU project and the EuroEXA project.

In the RED-SEA project, we implemented and evaluated in real hardware Accurate, an efficient and fair congestion control scheme. Importantly, Accurate relies on simple hardware and comes with only a few

knobs that need very little tuning. Using simulations in ExaNeSt-based testbed of the RED-SEA project, we compared Accurate with TCP and PAUSE-only RDMA networks, showing that Accurate provides up to 10x faster flow completion times for latency-sensitive flows (Giannopoulos, et al., 2018). In this work, we collect significant performance improvements from an ARM-based cluster running HPC application, mixed with background, datacentre-inspired, workloads, Accurate, using the optimizations developed in RED-SEA, achieves up to 12x better tail latency of victim flows, up to 3x better LAMMPS application runtime, and finds the correct rate for victim flows in less than 60 µs.

The next generation of BXI, BXIv3 or its subsequent versions, will explore the results to enhance BXI network resource management.

3.3 Extension of the BXI ecosystem

One of the RED-SEA research pillars was to explore new innovative solutions regarding end-to-end network services from programming models to reliability, security, low latency, and new processors.

Following explorations have been made for enriching BXI ecosystem:

- expansion of the usage of BXI networks through the Portals API
- new Intellectual Properties blocs added in the Network Interfaces
- DIAPASOM: a new application enabling openness of the interconnect to new kinds of applications.

3.3.1 Expansion of BXI networks usage

Portals is the API specifications for end-to-end communication on modern interconnect. Portals has been implemented in BXI through BXI NIC Portals Software Stack as shown in Figure 5 in 3.1.4.

Some extensions through the Portals API have been realised in the RED-SEA project which allows developers to access and integrate BXI into their applications easily and provides the necessary API for efficiently taking advantage of BXI's smart Network Interface Cards. The extensions are as follows: 1) LinkTest extension to support BXI; 2) Parastation MPI extension to support BXI and 3) MPC dual-rail support.

LinkTest extension to support BXI

LinkTest[1] is a robust communications benchmarking tool designed to test point-to-point connections between processes in serial or parallel mode and is capable of handling very large numbers of processes (tested up to 1 800 000 MPI tasks). Before the RED-SEA project, LinkTest has to rely on MPI layer and then Portals4 layer to benchmark BXI communication, as it is shown in Figure 6:

[1] Available at https://gitlab.jsc.fz-juelich.de/cstao-public/linktest

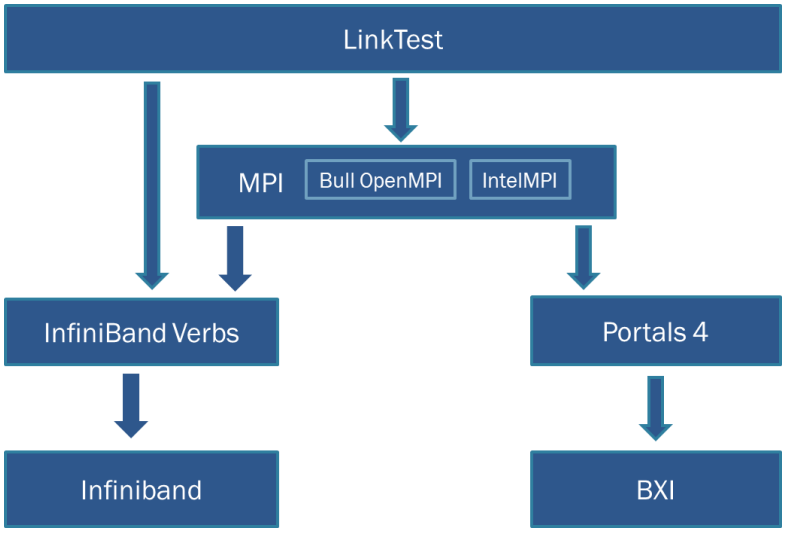


Figure 6: LinkTest before RED-SEA

LinkTest has been extended to support direct access to portals4 to perform BXI benchmarks directly thru Portals4 (Figure 7), as it is shown in Figure 8.

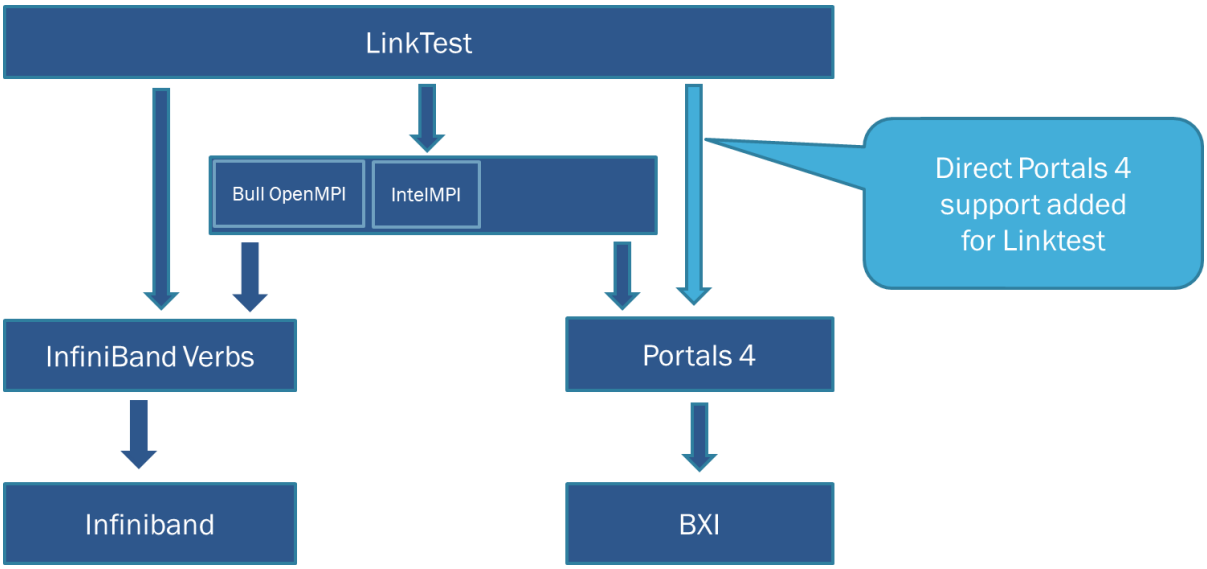


Figure 7: LinkTest: direct Portails4 support

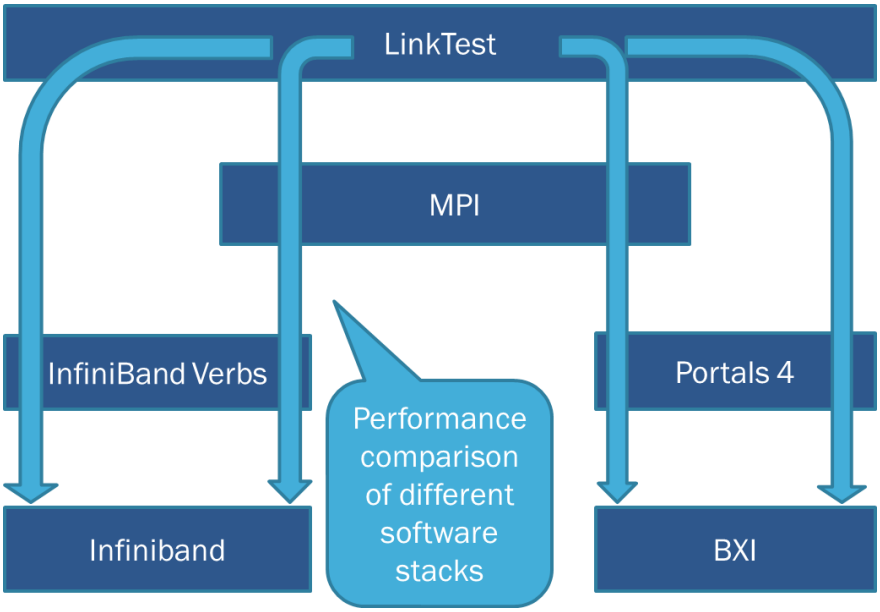


Figure 8: LinkTest: performance comparison

A blog post related to the LinkTest extension for BXI interconnects is accessible via the RED-SEA website: https://redsea-project.eu/extending-linktest-for-bxi-interconnects/.

Public release of a LinkTest version that can benchmark BXI communications is available on https://gitlab.jsc.fz-juelich.de/cstao-public/linktest.

ParaStation MPI extension to support BXI

ParaStation Modulo is a comprehensive software suite especially designed for MSA systems. The ParaStation MPI communication stack is a central pillar of ParaStation Modulo and enables efficient communication for MPI applications. ParaStation MPI is an MPICH derivate integrating its low-level communication layer pscom at the ADI3 layer (see Figure 9). The pscom library enables point-to-point communication among the MPI processes and abstracts the hardware with a variety of plugins supporting different interconnects and interfaces relevant to the HPC domain, e. g., InfiniBand, UCX, Extoll, and OmniPath.

In the RED-SEA project, the low-level communication layer of ParaStation MPI has been extended for BXI support. This extension enables both eager and rendezvous communication semantics in pscom4portals plugin; it also enables gateway communication with applications running on heterogeneous network landscapes.

Additionally, one-sided communication semantics in ParaStation MPI has been optimised on top of BXI to enable efficient RMA operations and one-sided communication semantics offered by MPI and PGAS. The support for one-sided communication via native RMA operations to ParaStation MPI on top of BXI allows ParaStation MPI get benefit from the RMA capabilities provided by BXI. Finally, the transparent bridging capabilities offered by pscom are further extended by the support for BXI interconnects. This allows MPI applications running on top of a heterogeneous network landscape using BXI among other high-speed interconnects.

These extensions are shown in Figure 9:

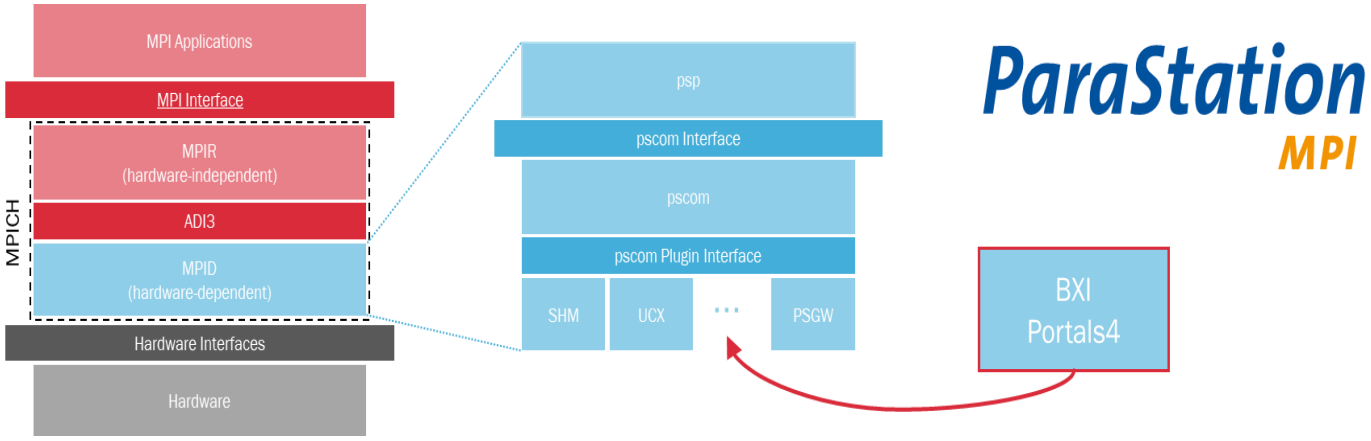


Figure 9: ParaStation MPI extensions

These extensions are part of commercial version of the ParaStation MPI communication stack. Furthermore, the ParaStation MPI will be used on the EUPEX Pilot.

MPC multirail feature added to the BXI network

MPC is an open source MPI implementation (http://mpc.hpcframework.com/). This implementation is part of the standard MPI implementation referenced by the MPI Forum.
MPC has been extended with the MPC multirail feature to the BXI network to take advantage of the underlying routing algorithms and balance traffic among NICs, thus increasing bandwidth. It allows the reduction of the cost of persistent collectives' communications through the improvement of the setup/registering step in high-speed network card and MPI runtime. More information can be found in this blog post available on the RED-SEA website: https://redsea-project.eu/development-of-multirail-feature-in-mpc-for-bxi-interconnect/.

The MPC multirail feature is available in github on main branch with tag MPC4.3.0 (official release to come T2 2024): https://github.com/cea-hpc/mpc/tree/MPC_4.3.0.

3.3.2 New Intellectual Properties blocs added in the Network Interfaces

One of the goals of RED-SEA is to tightly integrate the network interfaces (NIs) to RISC-V and ARMv8 cores and to FPGA-based accelerators.

Two network interfaces have been enriched in the RED-SEA project. They are:

- The CaRVnet network interface: a low-cost network interface end-points for BXI-enabled ports connecting with low-power processors (e.g. RISC-V from EPI) and accelerators based on evolution of the NI from the ExaNeSt / EuroExa projects. Please refer to the subsection below for further details.
- The APEnetX network interface: 1) this PCIe based NI designed and enriched to provide low-latency access to accelerators, as well as to x86 and ARM architectures, with the perspective of EPI CPU integration; 2) various optimisations for sending small messages made in APEnetX, significantly reducing descriptor and data latency over PCIe; 3) the APEnetX network interface has been adapted and successfully directly connected to BXIv2 links. Please refer to the subsection below for further details.

The caRVnet Network Interface

The CaRVnet network interface was firstly designed for low-power ARM-v8 processors.

In the RED-SEA project, the CaRVnet network interface has been optimised, especially the improvements made on the RDMA latency from 4µs to nearly 0.5µs, the link throughput from 20Gb/s to 100Gb/s, and RDMA message rate in FPGA tests from 1/3MOP/s to 50MOP/s for 8-byte transfers.

Moreover, the CaRVnet network interface has also been tightly coupled with low-power RISC-V processors, optimising latency while targeting a lean network interface design that can be integrated in the same chip with the RISC-V processor. This is a significant advancement towards integrating the network interface with the Arm GPP and the RISC-V on the same chip, ensuring the alignment with the EPI roadmap for this technology. A blog post relating to Low-latency Communication in RISC-V Clusters is available on the RED-SEA website: https://redsea-project.eu/low-latency-communication-in-risc-v-clusters/.

Finally, the CaRVnet network interface has been adapted and successfully directly connected to BXIv2 links. The integration is shown in Figure 10:

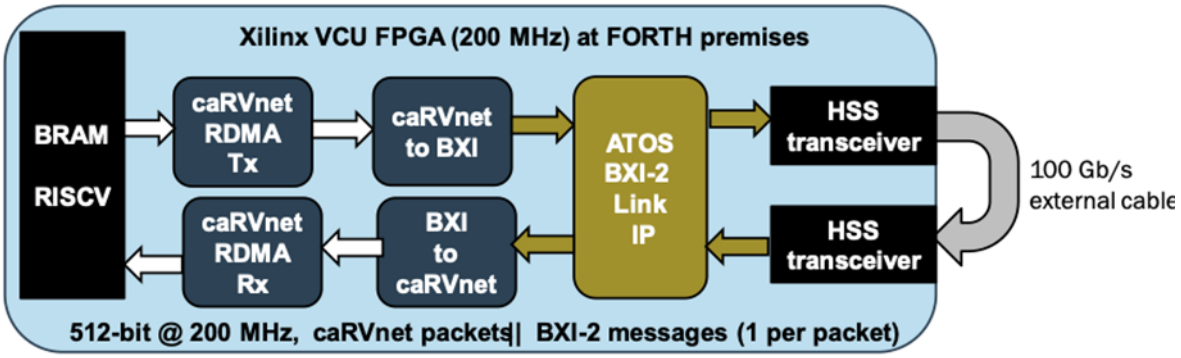


Figure 10: Block level diagram of BXIv2 + caRVnet integration

Two additional IPs: caRVnet-to-BXI and BXI-to-caRVnet, have been developed, they are connected before and after each BXIv2 Link to do protocol conversion.

This outcome is added to FORTH IP portfolio and is reuse in RISER EU project to connect RISC-V ASICs.

The APEnetX Network Interface

APEnetX, is a low-latency and high-throughput NIC based on a PCIe Gen3/Gen4 interface designed to increase the capability of the network and compliant with off-the-shelf clusters.

In the RED-SEA project, a proprietary Network Interface, known as the APEnetX network interface, has been designed and implemented to be compatible with the embedded DMA engine of Xilinx FPGAs. It is based on the PCIe interface and has been developed to provide low-latency access to accelerators, as well as to x86 and ARM architectures, with the perspective of EPI CPU integration. This development has been in the context of EPI-related intellectual proprieties' preparation. Moreover, APEnetX proposes various optimizations for sending small messages, significantly reducing descriptor and data latency over PCIe. Related network intellectual properties, mainly targeting the communication generated by spiking neural network applications, have been developed as well.The APEnetX network interface has been adapted and successfully directly connected to BXIv2 links. The integration of the APEnetX network interface with BXIv2 is shown in Figure 11:

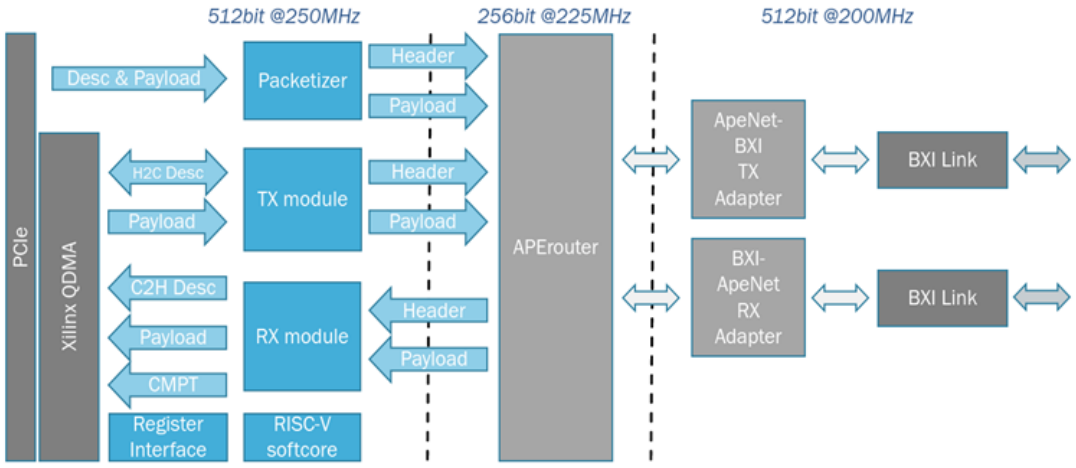


Figure 11: APEnetX BXIv2 integration Platform

This integration implies the design of APEnetX Transmission/Reception (TX/RX) adapter between the two domains – i.e., APEnetX-to-BXI and the BXI-to-APEnetX adapter – which translates the packets between the APEnet and BXI protocol and manages the credits.

A blog post relating to the APEnet interconnect architecture is available on the RED-SEA website: https://redsea-project.eu/the-apenet-interconnect-architecture-in-red-sea/.

This outcome is planned to be reused in further national (e.g. BRAINSTAIN) and potential European projects.

3.3.3 Opened the interconnect to new kinds of applications - DIAPASOM

One of the RED-SEA objectives was to enhance the performance and energy efficiency of the applications selected within the portfolios of RED-SEA partners. DIAPSOM was one of them.

DIAPSOM, stands for DIstributed And PArallel Self Organising Maps, is an in-house implementation of the parallel Self-Organising map algorithm (SOM) developed within the RED-SEA project. It has been successfully ported to two testbeds of the RED-SEA project (Dibona and TGCC KNL) and optimised for BXIv2. Self-organising maps (SOMs) are artificial neural networks that are used in the context of unsupervised machine learning with application in many research fields (genomics and computational biology, for instance). The "self-organisation" is particularly useful in clustering analysis since it provides additional insight into relationships between the identified clusters.

The DIAPSOM application was developed from scratch during RED-SEA. Both MPI-based and OPENSHMEM-based implementations were implemented. This is the first publicly available implementation of Self Organizing in Map using a Partitioned Global Address Space (PGAS) paradigm. This should pave the way to the implementation of other machine learning algorithms using the PGAS approach.

DIAPSOM has been released as an open-source package under the BSD 4-clauses license and is available at https://github.com/exactlab/diapasom.

3.4 Tools and simulators for high-performance interconnection network

RED-SEA also aimed at developing simulation platforms that complement the set of tools to perform co-design and evaluation activities in all phases of the project: to carry out the network requirements analysis, to support the new intellectual properties hardware implementation and test in design phase, and to accomplish the RED-SEA network characterisation and evaluation at large scale – not achievable

by the limited-size of the hardware testbeds. The following simulators and tools have been used and extended, as explained below, during the RED-SEA project life cycle:

- SAURON
- VEF
- COSSIM

Simulation is a popular method to evaluate the behaviour and performance of IT systems, such as HPC clusters or data centres, and has been extensively used to model and evaluate new designs for high-performance interconnection networks, such as those used in data centres. Many simulation tools have been proposed to model interconnection networks, such as OMNet++-based simulators (e.g., INET, ib_model, or SAURON), NS-3, SST, or CODES. These simulators can generally model the network components and their architecture, with different levels of detail or abstraction. Apart from the interconnection network architecture, another important aspect that network simulators need to offer is the ability to reproduce realistic workloads in the interconnection network.

SAURON simulator[2] is one of the simulators selected by the RED-SEA project. It is a packet-level event-driven simulation tool, based on the OMNeT++ framework that models interconnection networks for tens of thousands of nodes. Within the RED-SEA project, SAURON has been extended to model the BXI features, including the enhancements such as considering the latest version of OMNeT++. The extension and use of the SAURON simulator allowed the project partners dealing with network resource management (WP3) to have a wide and stable simulation infrastructure.

The VEF traces framework which was selected as a target tool by the RED-SEA project, is a set of open-source tools developed to facilitate the modelling and characterisation of the communication generated by MPI-based applications and to reproduce this modelling in network simulators. Essentially, the VEF traces framework offers a set of tools to capture the network traffic and generate self-related traces, called VEF traces. These traffic traces store the communication operations, both point-to-point and collective, and can be used to feed any third-party network simulator, provided that this simulator uses the TraceLib library included in the VEF traces framework. In the RED-SEA project, the VEF traces family of tools (open-source), has been extended to include the hardware and software support for collectives. Open-source VEF Traces are accessible via URLs:

- VEF Prospector: https://gitraap.i3a.info/fandujar/VEF-Prospector
- VEF TraceLib: https://gitraap.i3a.info/fandujar/VEF-TraceLIB
- Public repository: https://gitraap.i3a.info/jesus.escudero/vef-traces-repository

A blog post relating to the VEF traces framework is published on the RED-SEA website: https://redsea-project.eu/the-vef-traces-framework/.

Both SAURON simulation infrastructure and VEF trace framework will be used in further development of additional research within new projects and in future doctoral theses. Three Spanish national or regional projects are expected to utilise the version of SAURON and VEF extended by RED-SEA:

- TETRA2 (Efficient Techniques for Advanced Interconnect Technologies 2 - Técnicas Eficientes para Tecnologías de Red Avanzadas 2). Funded by: Regional Government of Castilla-La Mancha. Since: 01-09-2022. To: 31-08-2025.
- HEEDA (Highly Energy-Efficient Datacentre Architectures - Arquitecturas con alta eficiencia energética para centros de proceso de datos). Funded by: Spanish Ministry of Science and Innovation. Since: 01-12-2022. To: 30-11-2024
- DIDASI (Development and Improvement of HPC and Datacenter Applications, Services and Infrastructures - Desarrollo y Mejora de Aplicaciones, Servicios e Infraestructuras en HPC y Centros de Datos). Funded by: Spanish Ministry of Science and Innovation. Since: 01-09-2022. To: 31-08-2025

[2] P. Yébenes et al, "Towards Modeling Interconnection Networks of Exascale Systems with OMNet++", 21st Euromicro International Conference on Parallel, Distributed, and Network-Based Processing (PDP) 2013

COSSIM is another simulator dealt with in the RED-SEA project. COSSIM is the fully distributed simulator that can simulate several tens of thousands of nodes interconnected with any network topology/technology at a cycle accurate manner. COSSIM supports the efficient simulation of the processing units, based on GEM5, as well as of the on-chip and off-chip interconnections based on OMNET. In the RED-SEA project, COSSIM was extended so as to support full simulations of parallel systems incorporating RISC-V and ARM-based CPUs and different network protocols, technologies, topologies. The COSSIM parallel simulator framework fills a gap in the research community.

In terms of the cross-partners collaboration inside the RED-SEA consortium, VEF-Prospector, which captures the application MPI calls and gathers them in trace files using a special format (i.e., VEF), was successfully integrated in COSSIM simulator. Hence MPI traffic generated in COSSIM can be integrated within the SAURON simulator using VEF traces to simulate the SAURON network capabilities (e.g., BXI model).

It is expected that after the end of the project, the updated COSSIM simulator will be leveraged in the following ways:

- It will be marketed as a standalone software as well as a service to potential customers developing highly parallel systems,
- It will be utilised in-house for the development of novel EXAPSYS commercial services.

A blog post describing the COSSIM framework is available on the RED-SEA website: https://redsea-project.eu/cossim-framework/.

3.5 Collateral developments

Some additional developments have carried out with the RED-SEA project in order to enrich the European technologies in the realm of interconnect networks. They are: pspin synthesizable prototype and Low latency 100G Ethernet MAC & PCS.

3.5.1 pspin synthesizable prototype

SmartNICs are a recent movement towards offloaded packet processing to free the CPU from packet processing and thus spend more time handling the typical computation tasks. They come in different programming models and dataflow models. Among different paradigms, sPIN [3] developed at ETH Zurich proposes network accelerators with a micro-architecture optimised for packet processing and fine-grain memory hierarchies and data movement acceleration. It offers precise control to the programmers to build high-performance networked applications that are offloaded completely to the SmartNIC.

The sPIN paradigm has been evaluated extensively with diverse networked applications [4] [5] [6], showcasing its capability of offloading complicated applications to a sPIN-based network accelerator. Up to now, however, all evaluations of sPIN took place in simulation and there lacked a real-world end-to-end demo on hardware.

In the RED-SEA project, we built the first full-system demo of sPIN in hardware based on the PsPIN [7] implementation of sPIN and the Corundum [8] open-source Ethernet Network Interface Card (NIC). PsPIN is a RISC-V-based packet processing cluster implementing the sPIN in-network computing paradigm. This allows fast testing of packet handlers (code that runs on the sPIN cluster) in comparison to the slow cycle-accurate simulator. This facilitates the development of the sPIN ecosystem, including the platform itself and applications designed for it. Moreover, we also developed a synthesizable prototype of PsPIN in HDL on snitch RISC-V cores, including the necessary software ecosystem to compile and run sPIN handlers.

One of the central benefits of the sPIN architectures could be demonstrated with this prototype: it is possible to implement and run packet handlers undertaking tasks usually conducted by the host CPU. Therefore, the MPICH dataloop datatype engine has been ported to sPIN by implementing the corresponding handlers. This way, real overlap of communication and computation becomes possible

as the datatype (de-)serialisation can now be executed concurrently to the application's main computation phase.

The open-source software has been accessible since April 2021 and has been consistently updated throughout the project duration: https://github.com/spcl/pspin.
The implementation has been described and benchmarked in publications accessible via the RED-SEA website: Building Blocks for Network-Accelerated Distributed File Systems https://redsea-project.eu/wp-content/uploads/2022/11/3571885.3571898_Building-blocks.pdf.

A blog post providing information regarding Application-defined, high-performance packet processing with sPIN is accessible via the RED-SEA website: https://redsea-project.eu/application-defined-high-performance-packet-processing-with-spin/.

3.5.2 Low latency 100G Ethernet MAC & PCS

One of the interconnect technologies for high-performance computing (HPC) and Data Centre applications developed in the RED-SEA project is Low-latency 100G Ethernet MAC. This technology serves as a crucial building block for future Ethernet integration into HPC platforms and can target both ASIC and FPGA platforms.

Within RED-SEA, we developed a portable low-latency 100G Ethernet MAC block, called EX_EMAC_100G. This block can be used in an ASIC implementation (specifically for the GlobalFoundries GF22X process, already used in other EU projects) with PHY IP block from EXTOLL, or an FPGA implementation combined with a PHY IP block from the FPGA vendor. The 100 Gbit/s MAC aids in expanding the European IP eco system for interconnect technologies.

The low latency MAC & PCS is part of the IP (Intellectual Property) portfolio of EXTOLL GmbH and will become a commercial product. Usage in upcoming projects is also envisioned.

4 Impact, general findings and lessons learned

4.1 Impact

The RED-SEA project has made it possible to put the spotlight on the interconnection network, a topic that has hitherto received limited attention in European funded projects and, in general, little emphasis compared to, for example, processors or system software.

The most significant impact of RED-SEA is that, as the project is largely based on an existing product (BXI), a large proportion of the project results will be directly integrated into a new version of this product, BXIv3. The impact will be greatest if BXIv3 is actually deployed on the second European supercomputer to be installed in France. This is well under way, with BXIv3 already having secured a place in the EUPEX pilot - something not foreseen in the original EUPEX plans.

In addition to commercial products, the RED-SEA project partners have also made a major effort to share everything that is not proprietary IP in open source: 38% of the exploitable results are available in open source. This is the case, for example, of CEA's MPC and of ExactLab's DIAPASOM (DIstributed And PArallel Self Organising Maps). We can also mention the VEF traces, which have been the subject of numerous collaborations with other projects and have been widely disseminated by the UCLM - they have even been the subject of a recorded tutorial.

Finally, the impact of RED-SEA will also be perpetuated by numerous research projects that plan to take up and extend the results of RED-SEA.

4.2 Products

The RED-SEA project has played a significant role in facilitating the transformation of a substantial portion of our work into products, whether commercial or open-source. This is of utmost importance to us, as it will enable users to access the technology we have developed and provide feedback on its benefits or shortcomings. An example of this is the next-generation interconnect network, BXIv3, which will be integrated into the portfolio of BullSequana servers commercialised by BULL. Another example is the COSSIM simulator that has been extended to support full simulations of parallel systems incorporating RISC-V and ARM-based CPUs, along with various network protocols, technologies, and topologies. The beta version has already been rolled out, with the final version planned for commercial availability in 2025.

4.3 Industrial collaborations / exploitation

One of the notable strengths of the RED-SEA project lies in its fruitful collaboration between partners from both academia and industry within the project itself, as well as its collaboration with other EuroHPC Joint Undertaking funded projects.

The collaboration among partners from academia and industry within the project involved various stakeholders. One major contributor was the large enterprise (BULL), bringing extensive field experience in BXI (state-of-the-art Exascalable interconnect technology). Additionally, four active SMEs (Exact-lab, EXAPSYS, Extoll, PARTEC) added valuable knowledge and innovation capacity. Furthermore, prominent large data centres and research partners (FORTH, CEA, Jülich Supercomputing Centre, INFN, UPV, UCLM) provided the consortium with crucial challenges that required attention in the realm of interconnect technology.

Furthermore, the RED-SEA project has derived benefits from collaborating with other Joint Undertaking (JU) funded projects, both in terms of technical advancements and dissemination efforts. One noteworthy example on the technical front is the broadening of MPI-based traces from other projects (such as SEA projects and the MAELSROM project), which enabled us to conduct network analysis based on substantial trace sources. On the dissemination side, we successfully coordinated our efforts

by presenting common Birds of Feathers and workshops at various events. Additionally, we collaborated with other projects to share booths, thereby enhancing our visibility across multiple events and promoting the project effectively.

We consider that exploitation is relevant for all partners, whether industrial or academic. Thanks to the diversity of partners, RED-SEA achieved a great diversity of exploitations: products, patents, IP portfolio, open-source tools, further research. Finally, over 50% of the technologies developed, specifically 11 out of the 21 Exploitable Results, are either already integrated into a commercial product or are scheduled for implementation in the near future.

4.4 Dissemination

For dissemination, the collaboration with the other SEA projects and with other EuroHPC projects was essential to give visibility to RED-SEA. This collaboration initiated very early in the project made it possible for RED-SEA to have a significant and visible presence in many events (more than we initially anticipated), which would not have been possible on our own. This is not a new finding, several of the project partners were already used to collaborating between projects long before RED-SEA, for example by sharing a booth at major events. But it's a lesson worth perpetuating: when it comes to communication, you have to join forces to be visible!

Concerning dissemination, the very limited number of public deliverables hampered the dissemination of the project's findings - due to the sheer lack of material, especially as scientific publications are often derived from deliverables. This high level of confidentiality obviously has to do with the fact that the project is largely based on a commercial product with proprietary IP – an advantage for exploitation, but a problem for dissemination. However, RED-SEA managed to publish 19 academic papers, all available in Open Access. Whenever the publisher did not give the option to publish in open access, we made sure a pre-print was available in open access. Publications are listed in the About section of the RED-SEA website: https://redsea-project.eu/about/#publications.

4.5 Towards European Interconnect Network for Exascale and beyond (through NET4EXA project)

Several key considerations would be involved in creating an interconnect network specifically geared towards exascale and beyond, they are:

1. **High-speed Interconnects**: The network infrastructure would need to support high-speed interconnects capable of transferring vast amounts of data between nodes within exascale computing systems.
2. **Scalability**: The network should be designed to scale efficiently to accommodate the massive parallelism inherent in exascale computing systems. This involves considerations such as minimising latency, optimising bandwidth, and ensuring reliable communication between thousands or even millions of processing elements.
3. **Energy Efficiency**: Given the enormous power requirements of exascale computing, energy efficiency would be a critical consideration for the interconnect network. Designing low-power

communication protocols and optimising network topology to minimise energy consumption would be essential.

4. **Security**: With the increasing importance of data security and privacy, the interconnect network would need robust security features to protect sensitive data and prevent unauthorised access or tampering.
5. **Standards and Compatibility**: Establishing common standards and protocols for the interconnect network would be essential to ensure interoperability and compatibility across different hardware and software platforms. This would facilitate collaboration and the sharing of resources among European research institutions, universities, and industry partners.

For the sake of European technological sovereignty, it is imperative for Europe to develop a scalable and energy-efficient inter-node interconnect that integrates European technologies for exascale and post-exascale supercomputers. This European interconnect will serve as a cornerstone for advancing Europe's capabilities in high-performance computing.

The RED-SEA project has explored above-mentioned aspects by delivering outcomes for high-speed, scalable, energy-efficiency and secure interconnect networks. Moreover, BULL has played a role as funder in constructing the UltraEthernet Consortium (UEC) [9] and is the sole European company among the steering members of the UEC. The UEC announced in July 2023, has the vision to deliver an Ethernet-based, open, interoperable, high-performance, full-communications stack architecture to meet the growing network demands of AI/ML and HPC at scale. This represents a significant step towards openness for BXI. These outcomes have laid groundwork for the development, implementation, and industrialisation of BXIv3.

This foundation should be further advanced through full implementation of BXIv3, encompassing both switch and network interface components, as well as their corresponding hardware and software, at a high level of technical readiness. This effort is crucial to ensure that BXIv3 will be prepared for exascale and post-exascale supercomputers. We have submitted a project named NET4EXA which presents a proposal to the EuroHPC JU interconnect call (HORIZON-EUROHPC-JU-2023-INTER-02-01). This new project, serving as a follow-up to the RED-SEA project, aims to design, develop, and produce inter-node interconnects utilising European technologies for the integration in exascale and post-exascale supercomputers. Within this new project, we have planned to contribute to standards and compatibility, particularly within the UEC to pave the road for the compatibility, in parallel with the ongoing standardisation process led by the UEC.

The outcomes of RED-SEA in terms of the preparation for BXIv3 and advancements in network interconnect technologies will be used as input elements for NET4EXA.

4.6 General findings, Lessons learned / Vision

Overall, the RED-SEA project has contributed to advancing European capabilities in high-performance computing and interconnect solutions, laying the foundation for future innovations and collaborations in the field.

One of the primary findings of the RED-SEA project is the enhancement of interconnect technologies. This includes advancements in high-speed, scalable, energy-efficient, and secure interconnect networks, which have paved the way for the development of next-generation interconnect solutions. Notable achievements in this area include optimised network resource management and the exploration of innovative network solutions throughout the project's lifespan.

Another significant outcome of the RED-SEA project is the establishment of foundational work for BXIv3, a next-generation European interconnect technology. This includes the design of BXIv3 NIC transport layer, development of a BXIv3 Ethernet driver aimed at achieving high levels of parallelism through independent resources and a zero-copy mechanism, as well as the prototyping of BXIv3 Ethernet gateway solutions. These efforts represent crucial steps towards the realisation and industrialisation of BXIv3.

Furthermore, the collaboration within the consortium partners: the project has demonstrated the importance and effectiveness of cross-partner collaboration between academia and industry in driving innovation and achieving project goals. At the same time, the RED-SEA project has benefited from collaboration with other EuroHPC Joint Undertaking funded projects, leveraging synergies and maximising impact through knowledge exchange and joint dissemination activities.

Finally, the project has contributed to the development of European technologies to maintain technological sovereignty and competitiveness in the global landscape, particularly in the field of high-performance computing.

Moving forward, the vision is to continue advancing European technological capabilities in high-performance computing and interconnect solutions. This involves 1) further development and commercialisation of BXIv3, incorporating the latest advancements in hardware and software in future funded projects such as NET4EXA, 2) exploring new network technologies such as new protocols, photonics, 3) contributing to standards bodies and consortiums to ensure compatibility and interoperability across systems, fostering a more open and collaborative ecosystem, especially within the UltraEthernet Consortium.

We remain committed to pushing the boundaries of what is achievable in high-performance computing, especially in the field of interconnect networks. By doing so, we contribute to shaping the future of European technology and ensuring its competitiveness on the global stage.

5 Acronyms and Abbreviations

Abbreviation	Description
ASIC	Application Specific Circuit
BXI	BullSequana eXascale Interconnect
EPI	European Processor Initiative
FCS	Frame Check Sequence
FEC	Forward Error Correction
FIFO	First In First Out
FPGA	Field Programmable Gate Array
GF22X	GlobalFoundries GF22FDX process
Gbps	Giga bit per second = Gbit/s
HAS	High-level architecture specification
ICMP	Internet Control Message Protocol
I/O	Input / Output
IP	Intellectual Property
MAC	Medium Access Control
MPICH	a standard for message-passing for distributed-memory applications used in parallel computing
NIC	Network Interface Card
OFI	Open Fabric Interconnect.
Portals	Portals is a message passing interface intended to allow scalable, high-performance network communication between nodes of a parallel computing system
PCS	Physical Coding Sublayer
PGAS	Global Shared Address Space
PHY	Physical Layer = Serializer/Deserializer
RAM	Random Access Memory
RX	Receive
TX	Transmit

Table 1: Acronyms and Abbreviations

6 References

[1] Portals website: https://www.sandia.gov/portals/

[2] RED-SEA public website: https://redsea-project.eu/

[3] Torsten Hoefler et al. “sPIN: High-performance streaming Processing in the Network”. en. In: arXiv:1709.05483 [cs] (Oct. 2017). arXiv: 1709.05483. URL: http://arxiv.org/abs/1709.05483 (visited on 12/25/2020).

[4] Shiyi Cao, Salvatore Di Girolamo, and Torsten Hoefler. “Accelerating Data Serialization/Deserialization Protocols with In-Network Compute”. en. In: 2022 IEEE/ACM International Workshop on Exascale MPI (ExaMPI). Dallas, TX, USA: IEEE, Nov. 2022, pp. 22–30. ISBN: 978-1-66546-341-6. DOI: 10 .1109 / ExaMPI56604. 2022. 00008. URL: https: / / ieeexplore . ieee org / document / 10027020/ (visited on 08/09/2023).

[5] Salvatore Di Girolamo et al. “Network-accelerated non-contiguous memory transfers”. en. In: Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis. Denver Colorado: ACM, Nov. 2019, pp. 1–14. ISBN: 978-1-4503-6229-0. DOI: 10.1145/3295500.3356189. URL: https://dl.acm.org/doi/10.1145/3295500.3356189 (visited on 01/16/2022).

[6] Salvatore Di Girolamo et al. “Building Blocks for Network-Accelerated Distributed File Systems”. en. In: SC22: International Conference for High Performance Computing, Networking, Storage and Analysis. Dallas, TX, USA: IEEE, Nov. 2022, pp. 1–14. ISBN: 978-1-66545-444-5. DOI: 10.1109/SC41404.2022.00015. URL: https://ieeexplore.ieee.org/document/10046100/ (visited on 08/09/2023).

[7] Salvatore Di Girolamo et al. “PsPIN: A high-performance low-power architecture for flexible in-network compute”. In: arXiv:2010.03536 [cs] (June 2021). arXiv: 2010.03536. URL: http://arxiv.org/abs/2010.03536 (visited on 10/01/2021).

[8] Alex Forencich et al. “Corundum: An Open-Source 100-Gbps Nic”. en. In: 2020 IEEE 28th Annual International Symposium on Field-Programmable Custom Computing Machines (FCCM). Fayetteville, AR, USA: IEEE, May 2020, pp. 38–46. ISBN: 978-1-72815-803-7. DOI: 10.1109/FCCM48280.2020.00015. URL: https: //ieeexplore.ieee.org/document/9114811/ (visited on 01/18/2022).

[9] UltraEthernet Consortium website: https://ultraethernet.org/