RED-SEA: Network Solution for Exascale Architectures

Andrea Biagioni, Paolo Cretaro, Ottorino Frezza, Francesca Lo Cicero, Alessandro Lonardo,
Michele Martinelli, Pier Stanislao Paolucci, Elena Pastorelli, Francesco Simula, Matteo Turisini, Piero Vicini
INFN, Sezione di Roma, Italy, name.surname@roma1.infn.it

Roberto Ammendola
INFN, Sezione di Roma Tor Vergata, Italy, name.surname@roma2.infn.it

Pascale Bernier-Bruna, Claire Chen, Said Derradji,
Stephane Guez, Pierre-Axel Lagadec, Gregoire Pichon, Etienne Walter
Atos, France, name.surname@atos.net

Gaetan De Gassowski, Matthieu Hautreaux, Stephane Mathieu, Gilles Moreau, Marc Perache, Hugo Taboada
CEA, France, name.surname@cea.fr

Torsten Hoefler, Timo Schneider
ETH Zürich, Switzerland, timos@inf.ethz.ch

Matteo Barnaba, Giuseppe Piero Brandino, Francesco De Giorgi, Matteo Poggi
Exact Lab, Italy, brandino@exact-lab.it

Iakovos Mavroidis, Yannis Papaefstathiou, Nikolaos Tampouratzis
Exapsys, Greece, ygp@exapsys.eu

Benjamin Kalisch, Ulrich Krackhardt, Mondrian Nuessle
Extoll, Germany, name.surname@extoll.de

Pantelis Xirouchakis, Vangelis Mageiropoulos, Michalis Gianioudis, Harisis Loukas, Aggelos Ioannou,
Nikos Kallimanis, Nikos Chrysos, Manolis Katevenis
Forth, Greece, nchrysos@ics.forth.gr

Wolfang Frings, Dominik Gottwald, Felime Guimaraes, Max Holicki, Volker Marx, Yannik Muller
Julich Research Centre, Germany, n.surname@fz-juelich.de

Carsten Clauss, Hugo Falter, Xu Huang, Jennifer Lopez Barillao, Thomas Moschny, Simon Pickartz
ParTec, Germany, surname@par-tec.com

Francisco J. Alfaro, Jesus Escudero-Sahuquillo, Pedro Javier Garcia, Francisco J. Quiles, Jose L. Sanchez
University of Castilla-La Mancha, Spain, name.surname@uclm.es

Adrián Castelló, Jose Duro, Maria Engracia Gomez, Enrique Quintana, Julio Sahuquillo, Eugenio Stabile
Universidad Politecnica de Valencia, Spain, megomez@disca.upv.es

2022 25th Euromicro Conference on Digital System Design (DSD) | 978-1-6654-7404-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/DSD57027.2022.00100

***Abstract*—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. RED-SEA will pave the way to the next generation of European Exascale interconnects, including the next generation of BXI, as follows: (i) specify the new architecture using hardware-software co-design and a set of applications representative of the new terrain of converging HPC, HPDA, and AI; (ii) test, evaluate, and/or implement the new architectural features at multiple levels, according to the nature of each of them, ranging from mathematical analysis and modeling, to simulation, or to emulation or implementation on FPGA testbeds; (iii) enable seamless communication within and between resource clusters, and therefore development of a high-performance low latency gateway, bridging seamlessly with Ethernet; (iv) add efficient network resource management, thus improving congestion resiliency, virtualization, adaptive routing, collective operations; (v) open the interconnect to new kinds of applications and hardware, with enhancements for end-to-end network services – from programming models to reliability, security, low- latency, and new processors; (vi) leverage open standards and compatible APIs to develop innovative reusable libraries and Fabrics management solutions.**

***Index Terms*—Interconnect, HPC, congestion mechanism, datacenter, collective communication, Low-Latency Ethernet, QoS**

2771-2508/22/$31.00 ©2022 IEEE
DOI 10.1109/DSD57027.2022.00100

I. Introduction

To meet the scientific and industrial challenges of the coming decade, Exascale systems are required to simulate and understand more complex phenomena — e.g. multi-physics, multiple phases heterogeneous workflows — with refined accuracy, more frequently and faster (shorter time to solution), while handling an exponential growth of data. To this purpose, new usage models have been created improving the global efficiency of HPC systems. At the same time, the convergence of High Performance Computing (HPC), High-Performance Data Analytics (HPDA) and Artificial Intelligence (AI) workloads results in a modular supercomputer architecture (see Figure 1).

The optimization of application workflows has become a major source of improvement of production efficiency. The network backbone federating specialized clusters must present the supercomputer as an aggregation of resources that are organized to facilitate the mapping of applicative workflows. Modular supercomputer performance mainly relies on network quality. Moreover, the HPC systems will be part of the continuum of computing and will interact securely with the outside world including public clouds, edge servers, or third party HPC systems. These new challenges motivate the definition of a new generation of Exascale class Interconnect solutions.

Meanwhile, the Interconnect market landscape is changing: there are no more pure players following the acquisition of Mellanox by Nvidia in 2019. HPE/Cray future systems will be based on a proprietary interconnect named Slingshot [1], while Intel removed Omni-Path [2] from its roadmap. US, Japanese, and Chinese planned Exascale systems will embed "home-grown" technology for both the processor and the interconnect. Atos BXI [3] remains the only HPC interconnect independent from the compute solution available to European HPC solution providers. A reinforcement of the BXI roadmap is required, jointly with the EPI project, to guarantee technological sovereignty for upcoming European Exascale systems.

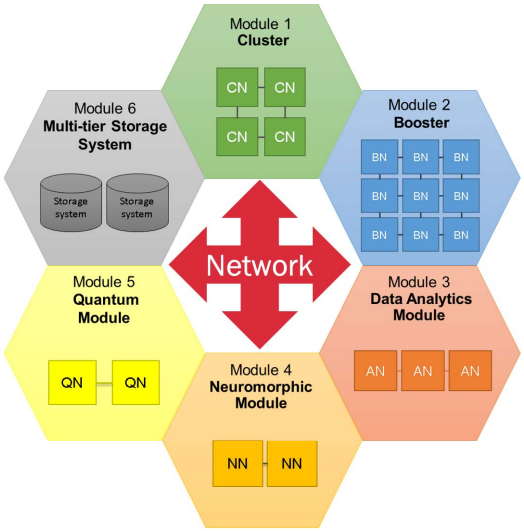

Fig. 1. Schematic representation of Modular Supercomputing Architecture.

The next generation of Exascale systems will critically require an efficient network. This network will need to support massively parallel processing systems (hundreds of thousands of nodes, millions of cores), provide a set of features allowing applications to scale efficiently at Exascale level and beyond, be prepared for power-efficient accelerators and compute units, and support wide-spread and emerging datacentric and AI-related applications. The RED-SEA consortium pursues this target, leveraging key European competences and background, including BXI, the key production-proven European Interconnect, as well as results from a number of EU-funded projects on interconnects and HPC systems. RED-SEA will support and drive extreme scale computing and data driven technologies within Europe by extending and optimizing the European BXI Exascale-class interconnect to anticipate the requirements of systems in the 2022-2025 timeframe.

A. Related work and previous projects achievements

During the last years many attempts were done to implement high-speed networks, targeting HPC computing, in FPGA-based system. INFN APEnet+ project [4] integrated a 3D Torus low latency interconnect equipped with a RDMA engine and a specialised HW/SW interface to a NVIDIA GPU implementing a "Direct-GPU" protocol to avoid multiple hops inside the host. APEnet+ was integrated in a heterogeneous PC cluster used mainly for LQCD computing and spiking neural network simulation. More recently Axiom project [5] built a FPGA SoC-based module integrating a multi-channel, high-speed, interconnects managed by a custom efficient Network Interface (NI) with multi (4) channels routed through USB-C mechanics connectors. The interconnect is balanced with the average computing throughput of the multi-core ARM embedded in the ZYNQ devices and provides system scalability exploiting the bi-directional links and different network topologies. In the framework of ExaNeSt [6] project an high performance low latency FPGA-based interconnect, ExaNet, has been designed, verified and deployed on a small/medium testbed of 128 Zynq Ultrascale+ SoC. ExaNet is a hierarchical network interconnect characterised by all-to-all topology at node level (made of 4 interconnected FPGAs) and a scalable 3D Torus network for inter-node interconnect at rack level. ExaNeSt integrates an optimized embedded ARM-oriented NI and a high performance switch/router block with 6 3D Torus bidirectional links up to 32 gbps, in a single FPGA. EuroEXA [7] leverage on ExaNet to push the concept of "hybrid topology scalable interconnect" (*TriFeCta*) at extreme scale. EuroEXA designed an enhanced version of ExaNet architecture providing different topologies and features at the various level of the network hierarchy. EuroEXA designed an innovative "Custom Switch" based on a single FPGA Virtex Ultrascale+ implementing 2-hops all-to-all topology at board level, a 3D Torus network at rack level and a ExaNet-Ethernet 100/200G bridge for inter-racks connectivity.

B. Consortium

The RED-SEA project sets out to deliver high-performance European interconnect technologies for extreme-scale, heterogeneous systems, by bringing together a number of well-established research teams across Europe, with a long experience in interconnects, including network design, deployment and evaluation. The consortium has hands-on experience on the following areas: 1) networks and interconnects, 2) performance evaluation, 3) simulation frameworks, 4) hardware design and tools and 5) HPC systems and integration. UPV and UCLM have developed state-of-the-art techniques along nearly all aspects of interconnects, including efficient topologies, deadlock-free adaptive routing strategies, flow control, QoS provision, congestion management techniques and fault-tolerant routing strategies. FORTH pioneered per-flow back-pressure and congestion management, weighted round-robin scheduling, as well as end-to-end scheduled flow control, with demonstrated experience on hardware and software prototyping. ETH brings cutting-edge research on scalable high-performance networking, parallel programming (excelling in MPI), and programmable in-network computing. INFN, the developer of APEnet [8], has a long experience in system prototyping and in efficient HPC codes. EXTOLL, an EU-based company, spin-off from University of Heidelberg, offering low-latency, state- of-the-art, high-performance IPs (especially high-speed SERDES) for HPC interconnects; CEA, with activities that span from the operation of large HPC deployments to new hardware and software technologies for systems and interconnects. FZJ hosts one of the biggest Supercomputing Centers in Europe, and conducts scientific research on High-Performance-Computing codes. ParTec develops MPI runtimes for interconnects, ExactLab optimizes HPC codes, and ExaP-SYS is a young startup with good knowledge on simulations and heterogeneous systems. CEA hosts one of the biggest Supercomputing Centers in Europe, and conducts scientific research as part of its mission. Atos and CEA are strongly cooperating in French national exascale program. Last but not least, the coordinator of the project, Atos/Bull, Europe's only computer manufacturer, has experienced research and development teams that have produced a number of commercial on-chip and off-chip interconnects and systems.

II. Objectives

The RED-SEA overall objective is to prepare a new-generation European Interconnect, capable of powering the EU Exascale systems to come, through an economically viable and technologically efficient interconnect, leveraging European interconnect technology (BXI) associated with standard and mature technology (Ethernet), previous EU-funded initiatives, such as ExaNeSt [6], EuroEXA, ECOSCALE, Mont-Blanc [9], DEEP projects [10], and the European processor (EPI) project, as well as open standards and compatible APIs.

The RED-SEA project will trigger the third generation of the BXI2 interconnect, contributing to its roadmap by: (i) defining the architecture blueprint and the corresponding simulation models; (ii) designing the new building blocks (IPs) necessary to address the new challenges of modular supercomputers; (iii) delivering initial proof-of-concept demonstration of its critical components on real life applications; and (iv) developing the ecosystem and creating a broader community of users and developers combining Research and Industrial teams.

The key challenges to achieve the desired performance and openness are listed below: 1) *scalability, reliability*: Demonstrate ways to scale industrial interconnects beyond 100 K nodes, while meeting key performance and reliability targets, and meeting diverse requirements, from communication libraries (MPI) and AI to data-centric applications; 2) *HPC/datacenter convergence*: develop and demonstrate product-level methods that optimally integrate Internet Protocol (IP) and Ethernet and RoCE (RDMA over Converged Ethernet) traffic over an HPC interconnect, achieving low latency and high message rates; 3) *throughput, bandwidth*: Multiply by 4 the bandwidth and the message rate available for each endpoint of the network by doubling the frequency of the link (up to 200 Gb/s) and by doubling the number of network interface for each process (multi-rail); 4) *quality of service*: Develop new congestion control algorithms and QoS-provision mechanisms suitable for agile data-centric HPC environments, evaluate them in platforms and scalable simulation models, and outline their path to the interconnect product; 5) *programmability, latency*: Develop ways to programmatically configure the network offload engine, while enabling also compute-in-network, achieving better latency / energy efficiency. 6) *new processors, relationship with EPI*: Demonstrate interoperability of the designed interconnect with components from the European Processor Initiative, such as Arm and RISC-V processors and accelerators and define alternative network architectures for European Exascale systems; 7) *new indicators*: Develop a set of benchmarks (including new ones and extensions of existing ones) highlighting new key features for applications at large scale such as computation/communication overlap and offloading. These benchmarks will help to design, validate, and compare high speed networks; 8) *protection, sharing*: Demonstrate methods for partitioning an existing HPC system, cluster into multiple (private) clouds, while maintaining protection, security, and isolation; 9) *applications and highlight*: Demonstrate the improvement of Data-centric applications (key-value store, AI, other) performance by obtaining better Benchmarks scores§; 10) *Go to market / impact*: Define a Go to Market path and optimize our chances to have a major part of these European IPs and outcomes used in main European systems by horizon 2022-23, while strengthening our current positions on the Interconnect market segment.

III. Concept

Next-generation HPC and data-driven systems will be heterogeneous in the computing devices that they will use, including low-power Arm and RISC-V processors, high-end CPUs, vector acceleration units and GPUs suitable for massive single-instruction multiple-data (SIMD) workloads, as well as

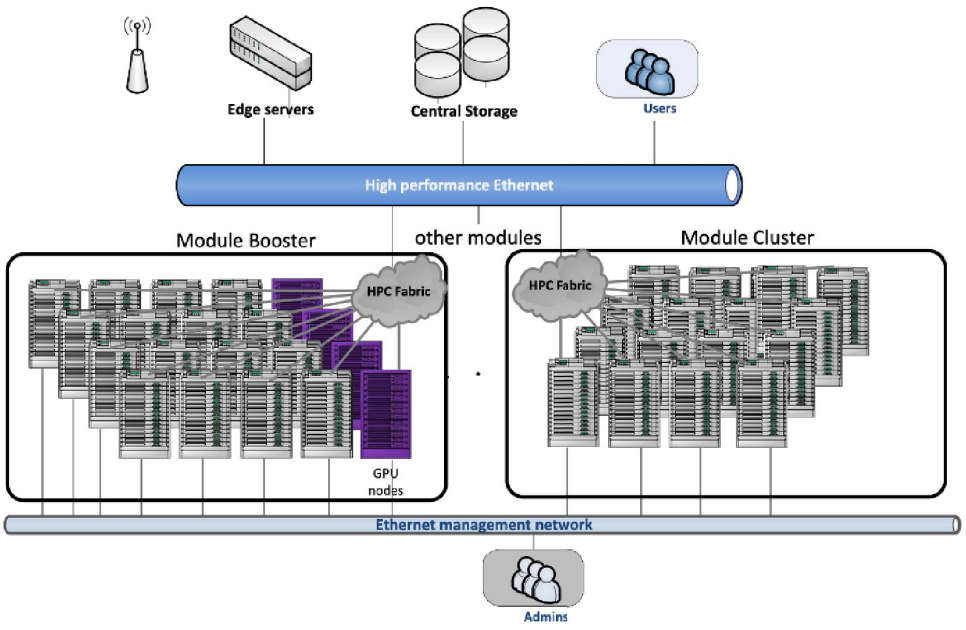

Fig. 2. Network architecture of a Modular Supercomputing Architecture.

FPGA and ASIC designs tailored for extremely power-efficient custom codes [11]. These compute units will be surrounded by distributed, heterogeneous (often deep) memory hierarchies, including high-bandwidth HBM2e memories (410 GByte/s raw bandwidth per stack) and distributed SRAMs, as well as by fast NVM devices offering microsecond-level access time. At the same time, modern data-parallel processing units such as GPUs and vector accelerators can crunch data at amazing rates (tens of TFLOPS). In this landscape, the network is likely to become the next big bottleneck, similar to memory in single node systems. the RED-SEA consortium proposes to address these challenges using (see Figure 2):

- High performance Ethernet as federation network featuring state-of-the-art low latency RDMA communication semantics;
- BXI as the HPC fabric consisting of two discrete components, a BXI NIC plus a BXI switch, and the BXI fabric manager. BXI third generation will add new features and boost its performance to match the listed objectives.

A. RED-SEA Physical Layer

The continuing increase of the bandwidth leads to more and more sophisticated serial links. The project targets 200 Gb/s link per direction which are made of four independent differential lanes running at 56 Gb/s. The quality of the link is measured by its Bit Error Rate (number of errors per unit time) for long distance and its interoperability with copper or optical cables from several providers. That is why the physical layer of the Ethernet and BXI port is critical to achieve the scalability, performance and reliability goals at system level within an acceptable cost. The project focuses on the development of MAC and PCS modular IPs which can be reused for both Ethernet links and future BXI links. Their design is optimized for low latency because in modern ASIC components the chip traversal is improving with the silicon thickness, while the part of the physical layer control and encoding is becoming prominent. The consortium IPs portfolio for designing network components will be enriched, thus providing an advantage over the existing IPs in the market.

B. RED-SEA Transport Layer

The scalability is a major concern. Installing production systems counting more than 100 K nodes is already a challenge but scaling the performance is the real requirement. The global performance depends directly on the network behaviour. The reliability requirements are following the explosion of the number of endpoints and the End-to-End reliability mechanism must be designed to simultaneously support up to 100 K nodes and to sustain the performance of 200 Gb/s for each link. The project will design an E2E reliability IP providing recovery mechanism for transient and permanent failures ensuring message integrity message ordering and message delivery via a go-back-N protocol is used to retransmit lost or corrupted message in the transport layer.

C. RED-SEA Host Interface

The project is targeting a very aggressive reduction of the latency between host processors and the network. This goal will be achieved in two different ways. The most disruptive change is to remove the standard PCIe interface and to have a direct access to the low power processor cores via a coherent

interface to reduce latency and simplify the software interface. The consortium will develop an FPGA prototype of the direct network interface to Arm and RISC-V cores. We will reuse the network interface from the ExaNeSt project and will adapt it to make it compatible with BXI switches. On the other hand, in the same RED-SEA prototype, we plan to optionally integrate a PCIe low-latency interface equipped with multiple RDMA engines, to allow for comparison of different CPU interface solutions and evaluation of innovative RED-SEA network architecture for off-the-shelf computer clusters.

D. RED-SEA Software Environment

The project aims to develop the software stack and the libraries to take advantage of the BXI offloading capabilities such as high-performance collective operations. In addition, the project wants to establish a new worldwide reference for benchmarking the efficiency of the communication using a smart network offering offloaded functions. The BXI network is based on the API Portals 4. Portals 4 [12] is a standard which was developed in collaboration by Sandia National Labs and by the University of New Mexico. It was chosen because it is the only interface available that supports both MPI and PGAS, while also providing appropriate building blocks for system software communications (parallel I/O, job launch). Portals 4.0 provides an abstraction that offloads the MPI matching semantics to provide a clean interface for offload and independent progress. Smart interconnects, based on the ability to offload MPI semantics from the host CPU to hardware, can be translated directly to greater application performance. Indeed, communications and computations progress completely independently. As a consequence, performance is not impacted by a heavy communication load on the host CPU. The BXI interconnect is the unique native hardware implementation of Portals 4.

E. RED-SEA QoS and Congestion Mechanism

RED-SEA will propose and evaluate new QoS and congestion management solutions for the BXI3 fabric (see Figure 3. The main techniques envisioned are: fine grain and medium grain adaptive routing, smart and responsive congestion management and highly flexible QoS. Key decisions should be taken as soon as possible in the hardware in order to increase the efficiency of the proposed techniques, designing and implementing mechanisms immediately at Network Interface Card or in the network switches Furthermore, the switches must provide a global view of the fabric. The global congestion management solution includes innovations developed by the partners at several levels:

- the protocol definition and the specification of the hardware probes to monitor the status of the Fabric.
- the algorithms to make the best decisions for adaptive routing and injection throttling.
- the support for congestion management tailored for collective operations.

The developed IPs span from the hardware modules in the BXI ports to the firmware modules running in the components and in the global fabric management software.

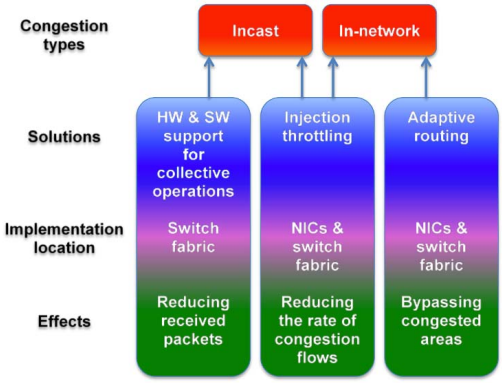

Fig. 3. RED-SEA Congestion management strategy.

F. RED-SEA Ethernet gateway

Exascale systems are not monolithic and closed systems anymore. They are hybrid, composed of specialized partitions and they must be open to the external world. They communicate with other supercomputers, Hybrid clouds, and Edge servers to participate to a global workflow. This constitutes a continuum computing approach. This is a disruptive change in the nature of supercomputers. As a consequence, no more compromise is possible with security matters, and a second consequence is that the gateways to the external networks are under pressure. One of the demonstrators of the project is an Ethernet gateway prototype connecting the HPC fabric to the Ethernet storage network. The Gateway is implemented as a high-end FPGA tightened to the ports of the BXI switches interfacing with the Ethernet switches. The FPGA prototype will be evaluated at speed, under real data workloads. The FPGA Gateway could be productized and then added in the future Atos systems to replace usual Gateways servers. The gateway nodes are usually servers embedding network adapters from both protocols and the translation is done in software, copying data in memory. That is why the achieved bandwidth is limited and these servers must be duplicated to sustain the required performance. The bridging can be optimized in cost and in performance on both sides: on Ethernet side using the RoCE semantics, and on HPC fabric side with better support for Ethernet over BXI. First, the HPC network must feature the transport layer attributes suitable for IP protocol: for example, a native broadcast capability and an unreliable communication option are desired. The key objective here is to develop a hardware bridging solution which can provide a connectivity between a virtual Ethernet network (on top of the HPC fabric) and a physical Ethernet network. This bridging solution can be integrated inside each port of the switch ASIC drastically reducing cost and matching the performance of the HPC fabric. Gateway servers will be replaced by few square millimetres

of silicon in the switch ASIC. The following figure illustrates the layers to traverse in each component of the path from the compute node to the storage node. The first table describes the IP router case which is a gateway server embedding one BXI card plus one Ethernet card. The second table describes the case of a BXI switch featuring the Ethernet Gateway. The number of components and the number of layers to cross are both reduced when using a switch with embedded Gateway.

G. Low-Latency Ethernet IPs

Any Ethernet implementation in the scope of HPC needs high-bandwidth but also low-latency operations. Crucial components in any Ethernet-port implementation be it in a switch or a NIC include the Media Access Control (MAC) component, the physical physical-coding sub-layer (PCS) as well as the PHY. Modern, very high-bandwidth Ethernet standards like 25G, 50G, 100G, and beyond, generally employ a technique called forward error correction (FEC) to efficiently handle channels with relatively high bit error rate (BER), which can be attributed to the high signalling rate on the physical medium for these standards. For the FPGA prototype built in the project, FPGA integrated hard-macro MACs are generally the best solution, because they do not consume costly logic resources of the FPGA and also avoid the potentially challenging (if at all possible) implementation of a high-performance MAC in LUTs. On the other hand, these general hard- macros do not feature the performance required for true HPC implementations and may miss some features that could be very valuable for the FPGA prototype as well as later FPGA-based exploitation of the project results. Low-latency, high-bandwidth MAC and PCS layer suited for HPC are a true challenge to design. FEC further complicates this task and special design choices must be made to efficiently support these needed features while keeping latency minimal (for example extensive use of speculative forwarding of frames).

IV. Methodologies

This project foresees several development and evaluation platforms, including analytical models and calculations, existing or new computer simulation models, as well as hardware emulation platforms combining FPGAs, servers, and commercial equipment (including BXI). In order to maximize productivity, the most suitable development and evaluation method will be selected on a per-problem and per-partner basis.

Additionally, depending on the maturity of the solution and of the evaluation platform, the new technologies may be tested under various workloads, including running applications and frameworks, synthetic workloads, stress-test and adversarial traffic patterns.

A. Connection between research and BXI

As shown in Figure 4 , the RED-SEA project is crafting a positive feedback loop-back between research and industry. The project will have the opportunity to leverage the BXI2 technology, which is used in operational HPC systems, offering high (100-200 Gb/s) link speeds, high-radix switches (48 ports), per-VC flow control, end- to-end and hop-by-hop retransmissions, and advanced network offload. Many of the activities in RED-SEA will use BXI2 and its existing high-performance capabilities in order to test and validate new IPs. This gives the research teams of RED-SEA a competitive advantage in furthering the state-of-the-art of industry-class interconnects. The consortium aims at a bidirectional and positive relationship between RED-SEA research and BXI. Its new technologies will be based on the specifications of BXI2 and will be tested on platforms incorporating BXI2 equipment, but the designs are really targeting BXI3 (the next generation of BXI). To serve this ambition, RED-SEA will define and evaluate ideas and alternatives on paper and in simulations, and will also realize many of these ideas in hardware IPs within emulation platforms, which can serve as a good starting point for BXI3 ASICs development.

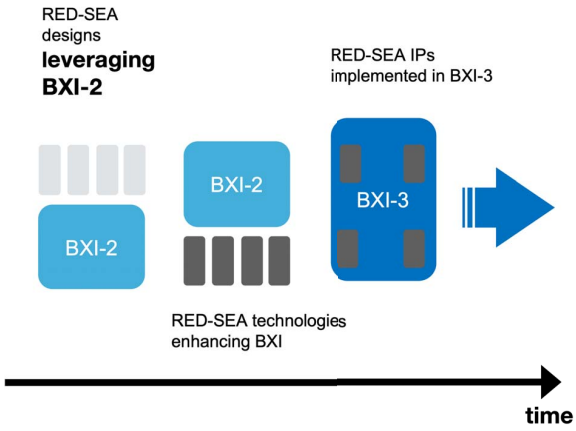

Fig. 4. In the RED-SEA project, we will leverage the existing BXI interconnect to develop, test and evaluate new technologies, planning to integrate many of them to advance BXI..

B. Co-Design

In RED-SEA, co-design is understood as a collective iterative process where (i) relevant applications, workloads, and ecosystems (requirements) (ii) interconnect and processor designers (new solutions/IPs) and (iii) Atos company (market considerations) meet to shape, evaluate and promote the interconnect IP developed in RED-SEA. Central to the overall success of this effort is the BXI interconnect and its evolution.

Application owners, HPC platform providers, and experienced interconnect designers will work on an extended list of in-house or widely-known applications and workloads. As an output of this process, the consortium will distill a number of envisioned workloads, associated with related performance targets. All solutions proposed by network designers within RED-SEA will be evaluated against the workloads and performance targets selected by the RED-SEA co-design.

In RED-SEA, we will use a mix of platforms to develop, test and verify the proposed solutions. The evaluation platforms

that are foreseen include a set of simulation models and hardware testbeds to match the particular requirements of each task and to boost productivity.

C. Simulation platforms and models

In RED-SEA, we will use network simulation models of varying fidelity in order to evaluate the envisioned solutions. Widely-adopted frameworks used to study HPC and datacenter interconnects include OMNeT++ [13]. However, in many cases, custom simulation platforms are preferred for their simulation speed. The accuracy of a simulation model in describing the inner workings of a network (such as a hardware or software component) varies widely across simulation platforms, and in principle affects the achievable scale and the time to solution – the simulation time of sizable configurations may range form minutes to days or even weeks. Certain simplifications are generally accepted in order to improve the time to solution, with regard to simulation evaluations. When studying congestion management for instance, the simulations help to evaluate different solutions under different congestive scenarios inside the network, and therefore the model does not have to capture the details of the processor. On the other hand, when examining the network interface, detailed models of the path connecting the processor with the network interface may be needed in order to accurately capture latency and message rate. SystemC and RTL models, tools that are used by hardware design teams, are also frequently used to examine in detail the behaviour and correctness of IPs under various workloads. The simulation tools that will be used by the consortium include: OMNeT++, NS3 [14], SystemC, Verilog/SystemVerilog, Gem5 [15], Custom frameworks [16].

D. Hardware emulation strategy

One of the targets of the RED-SEA project is to develop new hardware IPs that can be used toenhance the BXI interconnect. Functional level simulations will be used to evaluate the performance impact of these IPs under synthetic traffic patterns or mini-application traces. RED- SEA will additionally test the performance of these IPs in real deployments using hardware emulation platform, that mix commercial/verified boards (e.g. BXI switch) programmable FPGAs with modifiable models of BXI components (e.g. BXI switches and NICs), computer servers, and FPGA boards.

1) Small scale testbeds: Many of the tasks envisioned in this project concern the development of hardware and software IPs. The consortium will use small-scale testbeds for emulation purposes that include FPGA boards coupled with existing components, such as BXI switches, BXI network interface cards (NICs), Ethernet switches and adaptors, in order to validate the functionality of IPs, and measure their performance. These testbeds accelerate the simulation speed versus RTL simulation, and also make it possible to run real applications.

2) Large scale testbeds: The consortium will leverage the multi-FPGA platform from the ExaNeSt project to develop and test protocols and heterogeneous end-points, such as RISC-V processors and FPGA accelerators, against workload generators and real HPC / Data-centric / Deep-learning applications. The current ExaNeSt platform consists of twelve (12) liquid-cooled blades, each one housing four (4) Quad-FPGA-DaughterBoards (QFDBs), each with four (4) Xilinx Ultrascale+ MPSoCs containing four (4) ARMv8 (A53) cores, 64 GB DRAM, and a 256 GB SSD. The QFDBs are interconnected using both a 10G Ethernet network and a custom HPC interconnect. The prototype can be modified to support a number of network topologies.The currently planned configuration is a multipath-capable hybrid topology, with four spine switches and multi-groups of QFDBs. This prototype will support our efforts on simple network interfaces connected to ARM or RISC-V processors and to FPGA accelerators, and will also be used to demonstrate selected results from congestion management and multi-path routing.

3) Dibona: Arm-based HPC Cluster: The Dibona cluster is an Arm-based HPC class machine designed in the Mont-Blanc 3 project. Dibona will be reused for BXI performance optimizations and analysis. Dibona features ARMv8 ThunderX2 processors connected using InfiniBand or BXI fabric in a fat tree topology. Its software stack provides optimized operating system, HPC programming models and libraries. Dibona's compute blade is composed of 3 boards (compute nodes), each of them is optimized to embed 2 sockets (CPUs) and 16 memory channels. In the context of this project, the motherboards will be updated with BXI support on the interconnect mezzanine. This leads to major update of the Dibona cluster at the management node for the required runtimes (OS, drivers, firmware) and hardware components.

E. Relevant HPC and Datacentre Workloads

RED-SEA aims to improve the performance and energy-efficiency of applications in the following categories

- HPC simulations and codes
- Brain simulations
- Datacentre analytics frameworks, such as key-value stores.
- Storage frameworks using Ethernet or IP over BXI
- Deep learning applications

The list of applications and specific network benchmarks selected in the project is:

1) DPSNN/NEST: Distributed and Plastic Spiking Neural Network is an application developed to model brain cortex behavior and more recently to study sleep-related learning activity. It is a scalable neural network simulation C++/MPI code for HPC platforms at extreme scales. It simulates the spiking dynamics of the brain cortex by slicing it into a grid of cortical columns populated with neurons and their interconnecting synapses. Work on this subject is currently being moved over to the NEST simulator by the NEST Initiative; it is a well-established application encompassing a vastly broader range of neuron models and synaptic connectivity – supporting parallel execution via an MPI and OpenMP hybrid – which aptly provides a Python interface for easier setup and interoperability with codes for further algebraic manipulation

and statistical investigation over the simulated network and its dynamics.

2) LAMMPS: Large-scale Atomic/Molecular Massively Parallel Simulator is a classic molecular dynamic engine with a focus on material modelling. It is used widely in several branches of science: solid state physics, computational chemistry, biophysics and many others. It is also well known for its ease of compiling and running on several different computer architectures (from laptop to large cluster). The fact that it is used to simulate dynamics for atomic (atomic gases), meso (large molecules such as proteins) and continuum scale (metals), makes it a perfect codesign reference tool.

3) SOM: Self-organizing maps (SOMs) are artificial neural networks that are used in the context of unsupervised machine learning. As it happens when one tries to parallelize some learning algorithm, some modifications to the algorithm itself are in order (in the specific case from the so-called on-line version to the off-line and sparse off-line). The goal is to develop a massively-parallel implementation of this algorithm based on MPI and optimized for the RED-SEA architecture. We will also develop a PGAS implementation of the SOM algorithm, using the SAS environment, and compare it to the MPI implementation.

4) DAW: DAW (Datacentre-inspired adversarial workloads) is a testbed scenario-generator suite in order to replay interesting workloads of the ExaNeSt large-scale platform that stress the network interface capabilities at scale and the QoS capabilities of the interconnect, by mixing synthetic traffic-patterns (on platform) with running analytics (cluster-level) applications.

5) LinkTest: LinkTest is a tool for the scalable benchmarking of communication APIs. The APIs and associated communication hardware are benchmarked by sending messages between tasks hosted either on the same or different CPUs/GPUs. Messages can be sent between two tasks in parallel with one task sending its message to the other, as the other is sending its message back. Alternatively, the messages can be sent one after the other. Furthermore, the location where these messages are stored can be controlled. They can reside either in CPU RAM or GPU RAM.

6) PCVS: PCVS (Parallel Computing Validation Suite) is a validation engine designed to evaluate the offloading capabilities of high-speed network by running large test bases in a scalable manner, taking advantage of highly parallel environments to reduce their time to result, improving subsequently the project efficiency thanks to a more regular validation process. This set of benchmarks will be released as an open-source product included in the PCVS framework in order to be easily deployed on various architectures. In the end, this set of benchmarks will allow to characterize the capacity of exascale machines to deal with highly scalable applications such as task-based ones.

ACKNOWLEDGMENT

This work is supported by RED-SEA, a project that receives 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.

REFERENCES

[1] D. De Sensi, S. Di Girolamo, K. H. McMahon, D. Roweth, and T. Hoefler, "An in-depth analysis of the slingshot interconnect," in *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–14, 2020.

[2] M. S. Birrittella, M. Debbage, R. Huggahalli, J. Kunz, T. Lovett, T. Rimmer, K. D. Underwood, and R. C. Zak, "Intel® omni-path architecture: Enabling scalable, high performance fabrics," in *2015 IEEE 23rd Annual Symposium on High-Performance Interconnects*, pp. 1–9, 2015.

[3] S. Derradji, T. Palfer-Sollier, J.-P. Panziera, A. Poudes, and F. W. Atos, "The bxi interconnect architecture," in *2015 IEEE 23rd Annual Symposium on High-Performance Interconnects*, pp. 18–25, IEEE, 2015.

[4] R. Ammendola, M. Bernaschi, A. Biagioni, M. Bisson, M. Fatica, O. Frezza, F. Lo Cicero, A. Lonardo, E. Mastrostefano, P. S. Paolucci, D. Rossetti, F. Simula, L. Tosoratto, and P. Vicini, "Gpu peer-to-peer techniques applied to a cluster interconnect," in *2013 IEEE International Symposium on Parallel Distributed Processing, Workshops and Phd Forum*, pp. 806–815, 2013.

[5] Theodoropoulos *et al.*, "The axiom project (agile, extensible, fast i/o module)," in *2015 International Conference on Embedded Computer Systems: Architectures, Modeling, and Simulation (SAMOS)*, pp. 262–269, 2015.

[6] M. Katevenis *et al.*, "Next generation of exascale-class systems: Exanest project and the status of its interconnect and storage development," *Microprocessors and Microsystems*, vol. 61, pp. 58–71, 2018.

[7] Biagioni, Andrea *et al.*, "Euroexa custom switch: an innovative fpga-based system for extreme scale computing in europe," *EPJ Web Conf.*, vol. 245, p. 09004, 2020.

[8] R. Ammendola, A. Biagioni, O. Frezza, F. L. Cicero, A. Lonardo, P. S. Paolucci, D. Rossetti, F. Simula, L. Tosoratto, and P. Vicini, "APEnet+: a 3D torus network optimized for GPU-based HPC systems," *Journal of Physics: Conference Series*, vol. 396, p. 042059, dec 2012.

[9] A. Armejach, B. Brank, J. Cortina, F. Dolique, T. Hayes, N. Ho, P.-A. Lagadec, R. Lemaire, G. López-Paradís, L. Marliac, M. Moretó, P. Marcuello, D. Pleiter, X. Tan, and S. Derradji, "Mont-blanc 2020: Towards scalable and power efficient european hpc processors," in *2021 Design, Automation Test in Europe Conference Exhibition (DATE)*, pp. 136–141, 2021.

[10] N. Eicker, T. Lippert, T. Moschny, E. Suarez, and for the DEEP project, "The deep project an alternative approach to heterogeneous cluster-computing in the many-core era," *Concurrency and Computation: Practice and Experience*, vol. 28, no. 8, pp. 2394–2411, 2016.

[11] J. S. Vetter, R. Brightwell, M. Gokhale, P. McCormick, R. Ross, J. Shalf, K. Antypas, D. Donofrio, T. Humble, C. Schuman, *et al.*, "Extreme heterogeneity 2018-productive computational science in the era of extreme heterogeneity: Report for doe ascr workshop on extreme heterogeneity," 2022.

[12] K. Raffenetti, A. J. Pena, and P. Balaji, "Toward implementing robust support for portals 4 networks in mpich," in *2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*, pp. 1173–1176, IEEE, 2015.

[13] A. Varga, *OMNeT++*, pp. 35–59. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[14] G. F. Riley and T. R. Henderson, *The ns-3 Network Simulator*, pp. 15–34. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010.

[15] N. Binkert *et al.*, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, p. 1–7, aug 2011.

[16] N. Tampouratzis, I. Papaefstathiou, A. Nikitakis, A. Brokalakis, S. Andrianakis, A. Dollas, M. Marcon, and E. Plebani, "A novel, highly integrated simulator for parallel and distributed systems," *ACM Trans. Archit. Code Optim.*, vol. 17, mar 2020.