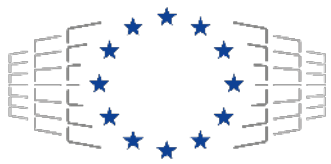
RED-SEA: network solution for ExaScale Architectures 25 May 2022

Andrea Biagioni, INFN, on the behalf of the APE LAB team

Workshop sul calcolo nell'INFN, Paestum 23 - 27 Maggio 2022

(25 minutes)



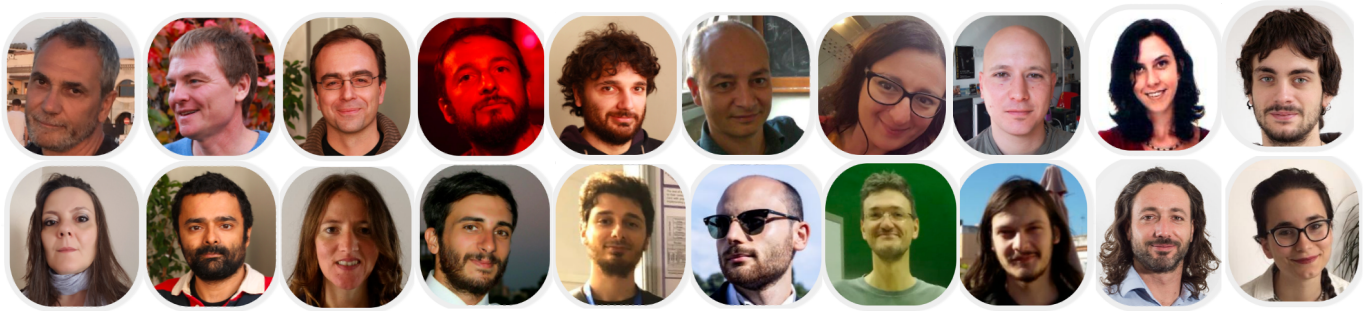
EuroHPC Joint Undertaking

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.

APE LAB

https://apegate.roma1.infn.it/

@APELab_INFN



- APE Parallel/Distributed Computing Lab
- 20 members (12 permanent staff + 8 fixed-term)
- 3 main research lines
  - HPC (system architecture, scalable network, application optimization)
  - NeuroScience (Brain Simulation, models, neuromorphic system)
  - HEP Computing (Fast read-out systems, online trigger, ML methods)
- Our Know-How
  - ASIC design, FPGA design, GPU programming and integration, Network design, dense system integration, parallel programming and application coding (LQCD, neural networks, brain simulation, complex systems), system software, compiler and languages,...
- International research network and industrial collaborations (a sample list):
  - ATOS, FORTH, UPC/BSC, Julich Forschungszentrum, Manchester Univ., Fraunhofer, CERN, NVidia, E4, Iceotope, ….

The RED-SEA consortium

Project start: 01/04/2021
Project duration: 36 months
Project budget: 8 M€

We are one of the “SEA” projects

3 complementary projects addressing Exascale challenges in a Modular Supercomputing Architecture (MSA) context

- In line with several HW/SW Exascale projects funded under previous European programmes
- Funded by the EuroHPC 2019-1 call focused on SW and applications
  - The EuroHPC Joint Undertaking targets Exascale computers in Europe in 2023-24
  - Should contain as many European components as possible
- Coordinated with other on-going European projects, particularly the European Processor Initiative

DEEP-SEA: DEEP Software for Exascale Architectures

- Better manage and program compute and memory heterogeneity
- Targets easier programming for Modular Supercomputers
- Continuation of the DEEP projects series

IO-SEA: Input/Output Software for Exascale Architectures

- Improve I/O and data management in large scale systems
- Builds upon results of SAGE1-2 projects and MAESTRO

RED-SEA: Network Solution for Exascale Architectures

- Develop European network solution
- Focus on BXI (Bull eXascale Interconnect)

RED-SEA objectives

Enable

Enable the design of a new generation of high performance network interconnect

- Leveraging existing European technology (BXI, Exanest …)
- Able to power the future EU Exascale systems

Explore

Explore new innovative solutions

- End-to-end network services – from programming models to reliability, security, low latency, and new processors

Develop

Develop the ecosystem and create a broader community of users and developers combining Research and Industrial teams

- Leveraging open standard and compatible API to develop innovative re-useable libraries and Fabrics management solutions

Objective

- O1 (scalability, reliability): >100k nodes; communication libraries (MPI) and AI to data-centric applications
- O2 (sustainability, HPC/datacenter convergence): Integrate Internet Protocol, Ethernet, RoCE
- O3 (Throughput & BW): x4 bw and message rate; x2 link freq. x2 NI for each process (multi-rail)
- O4 (congestion, QoS, isolation)
- O5 (programmability, latency): configure the net. offload engine, compute-in-network, improving lat. and energy efficiency
- O6 (new processor, EPI): ARM + RISC-V interoperability
- O7 (new indicators): new key features for apps, communication/computation overlap and offloading
- O8 (protection): partitioning into multiple clouds
- O9 (application and highlight): obtaining better benchmarks scores
- O10 (go to market)

The four pillars of RED-SEA research

Architecture, co-design and performance

Optimizing the fit with the other EuroHPC projects and with the EPI processors

High-performance Ethernet

Development of a high-performance, low-latency, seamless bridge with Ethernet

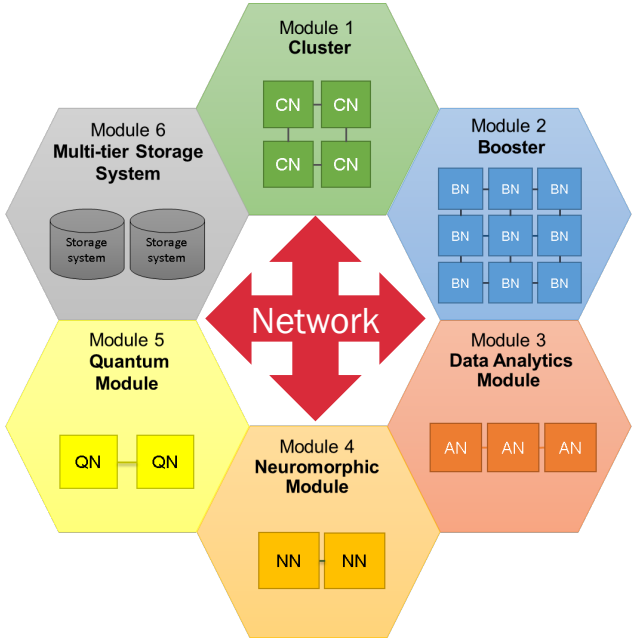
Efficient Network Resource management

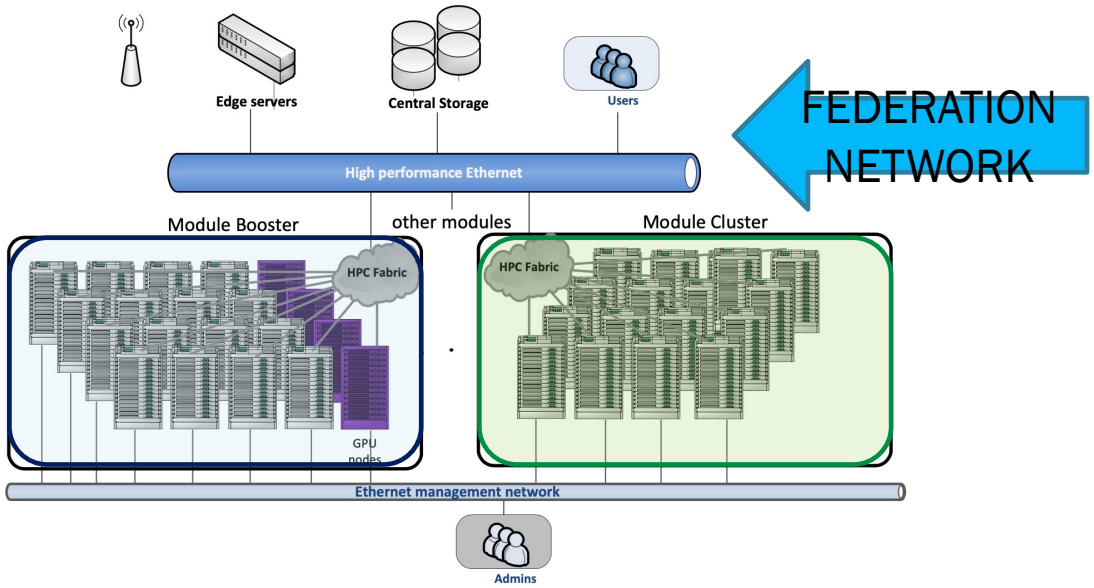
Including congestion management and Quality-of-Service targets while sharing the platform across application and users

Endpoint functions and reliability

End-to-end enhancements to network services - from programming models to reliability & security and to in-network compute

RED-SEA: MSA network architecture

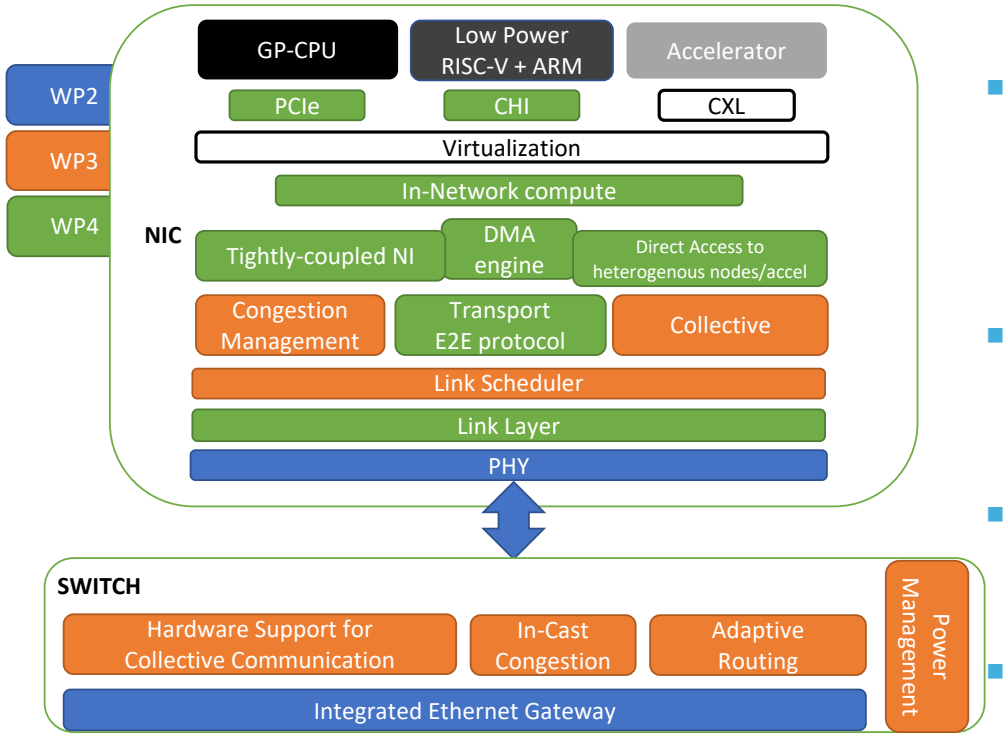




- HPC (High Performance Computing) ; HPDA (High-Performance Data Analytics); AI (Artificial Intelligence)
- Supercomputer: aggregation of resources that are organized to facilitate the mapping of applicative workflows
- HPC is part of the continuum of computing

- 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.

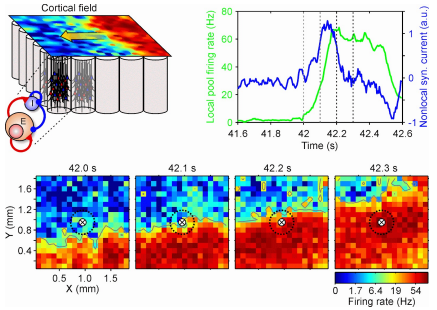
RED-SEA architecture

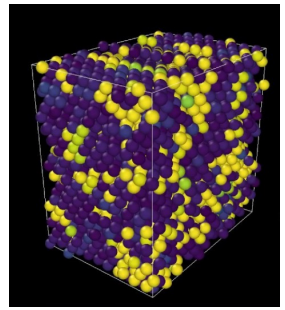


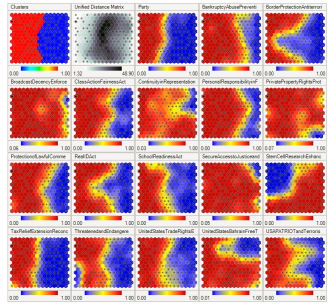
- Physical Layer (PHY)
  - MAC and PCS modular IPs (Ethernet/BXI), 200Gb/s per dir. 4*56gbps independent differential lanes
- Transport Layer
  - E2E reliability IP providing recovery mechanism for ensuring *message integrity, ordering* and *delivery via* a go- back-N protocol is used to retransmit lost or corrupted message
- Host Interface
  - direct access to the low-power processor (ARM/RISC-V) via a coherent interface to reduce lat. and simplify the SW interface.
  - PCIe low-latency interface equipped with multiple RDMA engines, to be compliant with off-the-shelf computer clusters
- Sofware environment
  - SW stack and libraries to exploit BXI offloading capabilities and HP collective
  - new benchmark for the efficiency of the communication using a smart NIC
- QoS and congestion management
  - adaptive routing, smart and responsive congestion management and highly flexible QoS (key decision made by NIC and the network switches)
- Ethernet Gateway
  - develop a HW bridging solution providing connectivity between a virtual Ethernet network (on top of the HPC fabric supporting Eth over BXI) and a physical Ethernet network (using the RoCE semantics)
  - bridging solution can be integrated inside each port of the switch ASIC

RED-SEA: methodology for Co-Design Activity

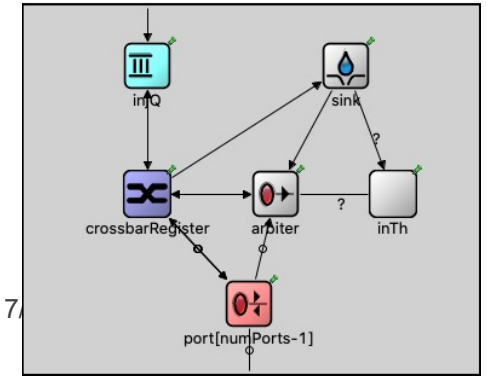
- **Application portfolio**
  - NEST: simulator for spiking neural network models
  - LAMMPS: molecular dynamic engine with focus on material modelling
  - SOM: artificial neural networks used in the context of unsupervised ML



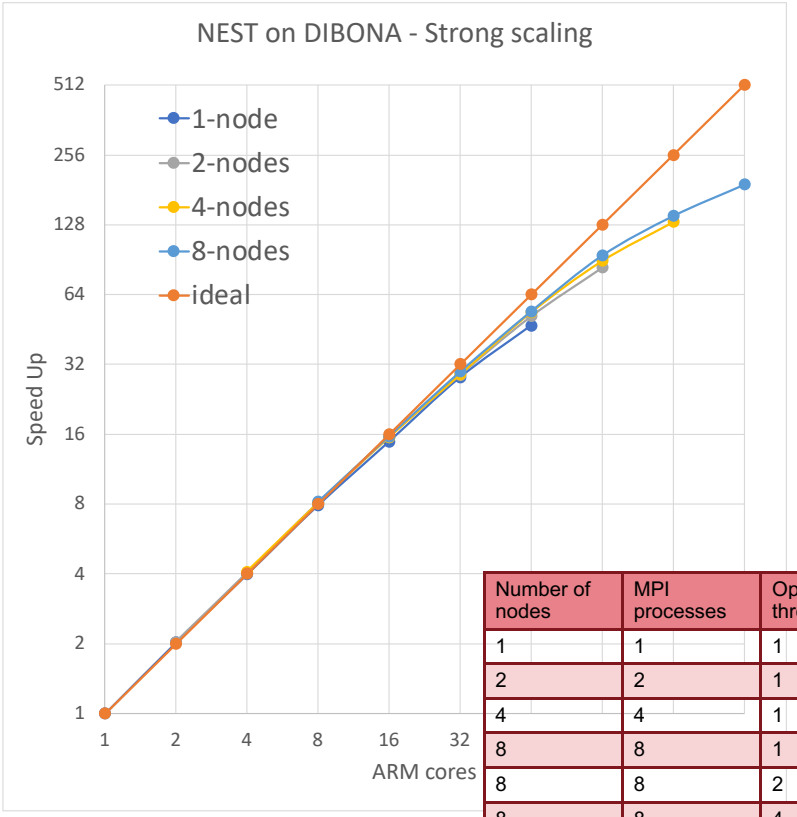




- **Benchmark portfolio**
  - GSAS: Global Shared Address Space environment provides a shared memory abstraction model to distributed applications
  - DAW: stress the NI capabilities at scale and the QoS capabilities of the interconnect
  - LinkTest: scalable benchmark for point-to-point communications
  - PCVS: validation engine designed to evaluate the offloading capabilities of high-speed network
- **Collection and Analysis of MPI Network Traces generated by applications**
  - VEF traces + DIBONA (12 nodes, 768 ARM cores, BXI interconnect)
  - Requirements for the applications and co-design recommendations
- **Simulator as reference to support the design and implementation of novel IPs proposed in the project**
  - Network traces feed the project simulators
  - Extrapolation of the behaviour at large scales (up to 10k nodes)

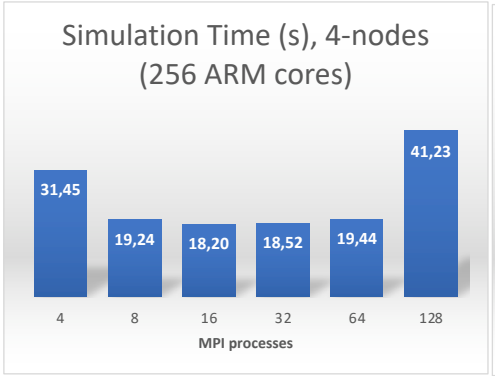


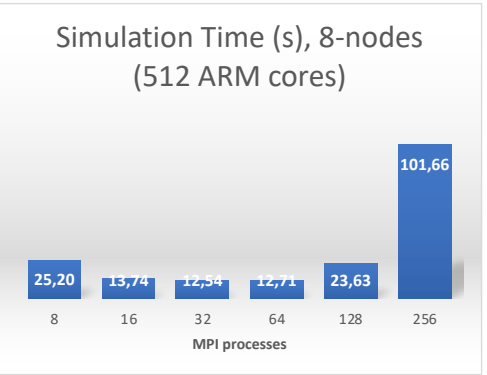
Co-Design 2022: Requirement for Neural communication



- Dibona (11-nodes)
  - Bi-socket Cavium ThunderX2
  - 704 ARM v8 cores@2GHZ
  - Memory: 48GB per node
  - Interconnect: BXI1.3

Number of nodes	MPI processes	OpenMP threads	ARM cores	Simulation time[s]	Ideal [s]	Deviation [%]	Speed up
1	1	1	1	2390.71	2390.71	0.00	1.00
2	2	1	2	1180.21	1195.35	-1.27	2.03
4	4	1	4	586.45	597.68	-1.88	4.08
8	8	1	8	292.39	298.84	-2.16	8.18
8	8	2	16	152.50	149.42	2,06	15.68
8	8	4	32	80.45	74.71	7.68	29.72
8	8	8	64	44.13	37.35	18.12	54.18
8	8	16	128	25.35	18.68	35.71	94.32
8	32	8	256	17.11	9.34	83.19	139.75
8	32	16	512	12.54	4.67	168.56	190.65
11	44	16	704	11.27	3.40	231.87	212.13

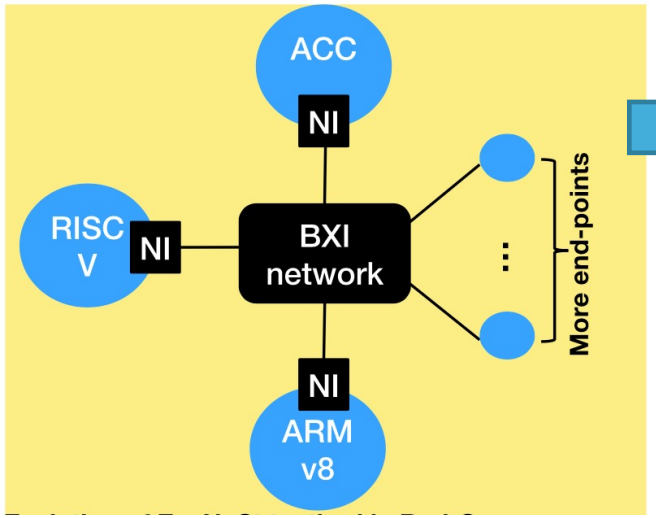




MPI process	OpenMP Threads	Number of messages	total bytes	average size [B]	Range of Interest	Messages in range [%]	Average Size [B]	Average Size Ratio
8	64	6,82E+04	8,02E+08	11760	512B-16kB	95,9	4854	
16	32	2,90E+05	1,24E+09	4259	256B-8kB	97,1	2558	0,53
32	16	1,19E+06	2,05E+09	1716	128B-4kB	97,8	1300	0,51
64	8	4,84E+06	3,97E+09	821	128B-2kB	96,6	718	0,55
128	4	1,95E+07	7,99E+09	410	64B-1kB	97,8	383	0,53
256	2	7,79E+07	2,06E+10	264	64B-512B	97,2	244	0,64

- Deviation from ideal scaling is already significant exploiting 256 ARM cores
- The NEST simulation could be distributed on a maximum of **64 MPI processes.**

BXI SYSTEM: INFN APEnetX



END-POINT: INFN APEnetX

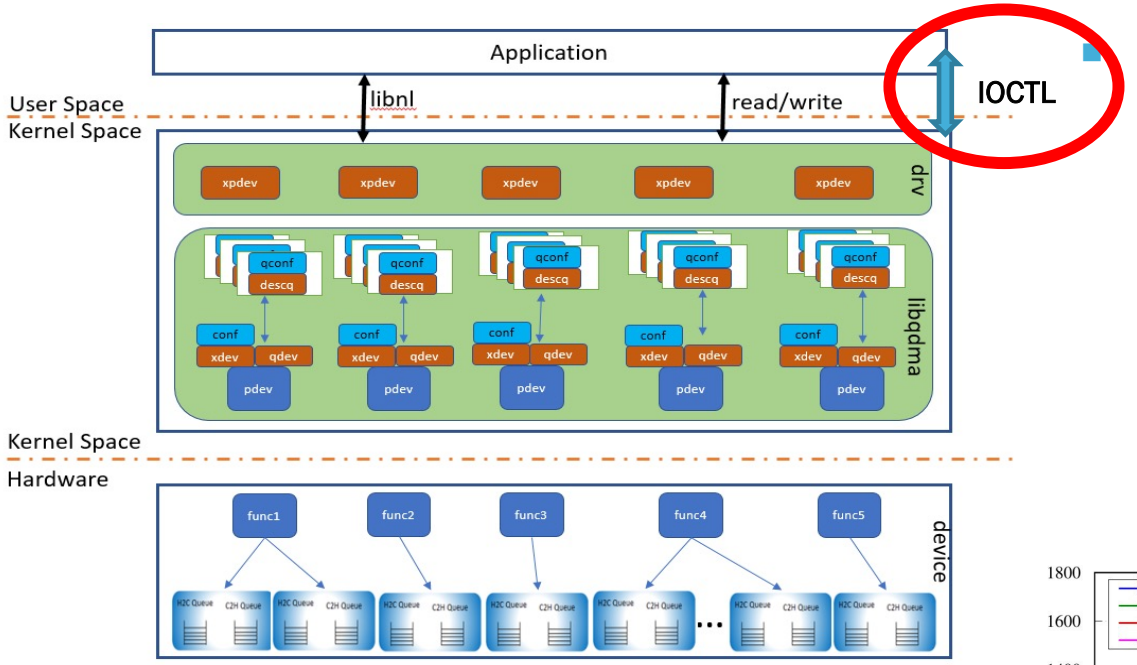
- to tightly **integrate** the network interfaces (NIs) **to RISC-V and ARMv8 cores and to FPGA-based accelerators and GPUs**
  - To prepare a number of EPI-related IPs
  - To create a highly heterogeneous programmable platform connected with state-of-the-art interconnect technologies.

- INFN duties
  - Network Interface Card (APEnetX)
    - PCIe gen4 (GPU+CPU) + BXI link (Xilinx Alveo FPGA)
  - Co-Design through applications (NEST)
- TARGET (Q4-2023)
  - Developing network IPs to optimize spiking neural network communication

APEnetX

- Re-design of the hardware blocks for the Network interface (APEni)
  - Lean, tightly-coupled network interfaces (M1-M36)
    - PCIe-based CPU interface (gen4)
    - Support for ExaNet protocol
  - Direct access to heterogeneous nodes/accelerators (M1-M36)
    - Tight integration of APEni to GPU accelerators and/or FPGA (synergies with other EU projects)
- Design of the APEnet architecture
  - Xilinx QDMA-based Network interface + enhanced network IP (Inherited fro previous EU project)
- Integration of APEnet architecture with BXI infrastructure
  - BXI link integrated in APEnetX (BXI end-point)

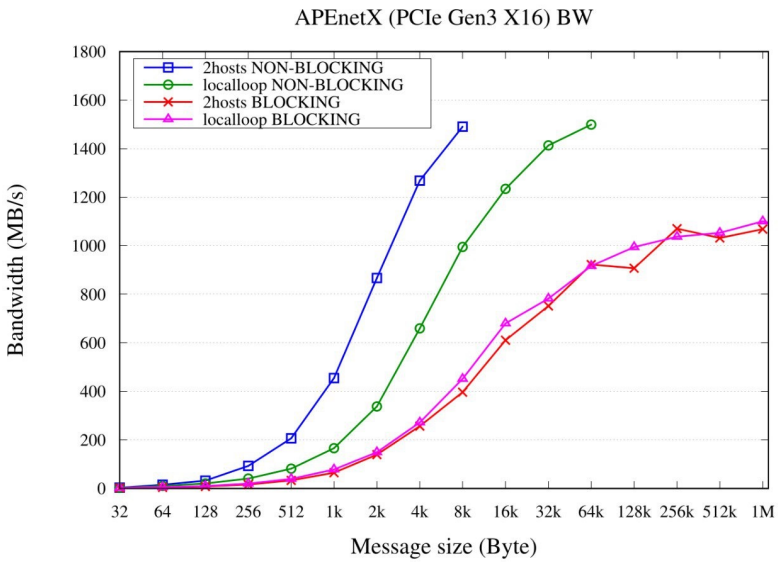
analysis of the QDMA-BASED NI performance

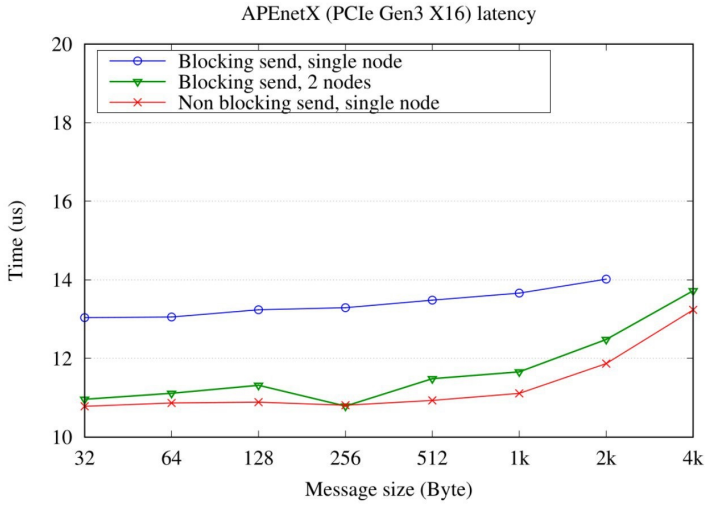


To not modify the existing Xilinx driver we used the IOCTL syscall as driver entry point (instead of already taken read/write) for:

- Register (pin & lock) RX buffer → **the IOMMU is used as translator**
- Starting the send phase → **custom TX descriptor** with bypass mode
- Get RX completion → **custom completion**
- Notify the completion to the user application
- Deregister all at the end of the test

- Preliminary BW test using Xilinx QDMA IP
  - Xilinx Alveo U200;
  - PCIe Gen3 x16
  - Network IP (128bit@100MHz)

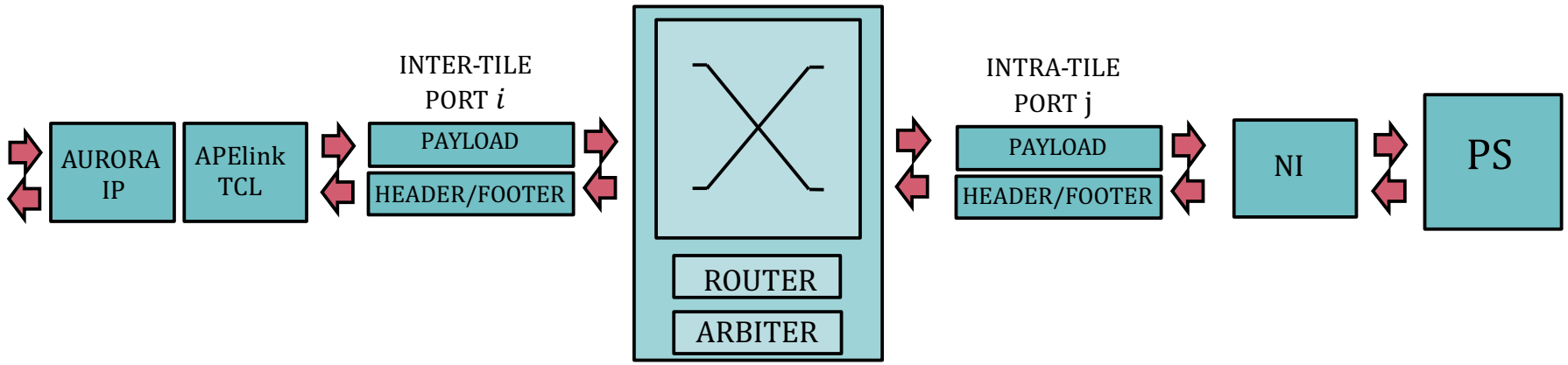




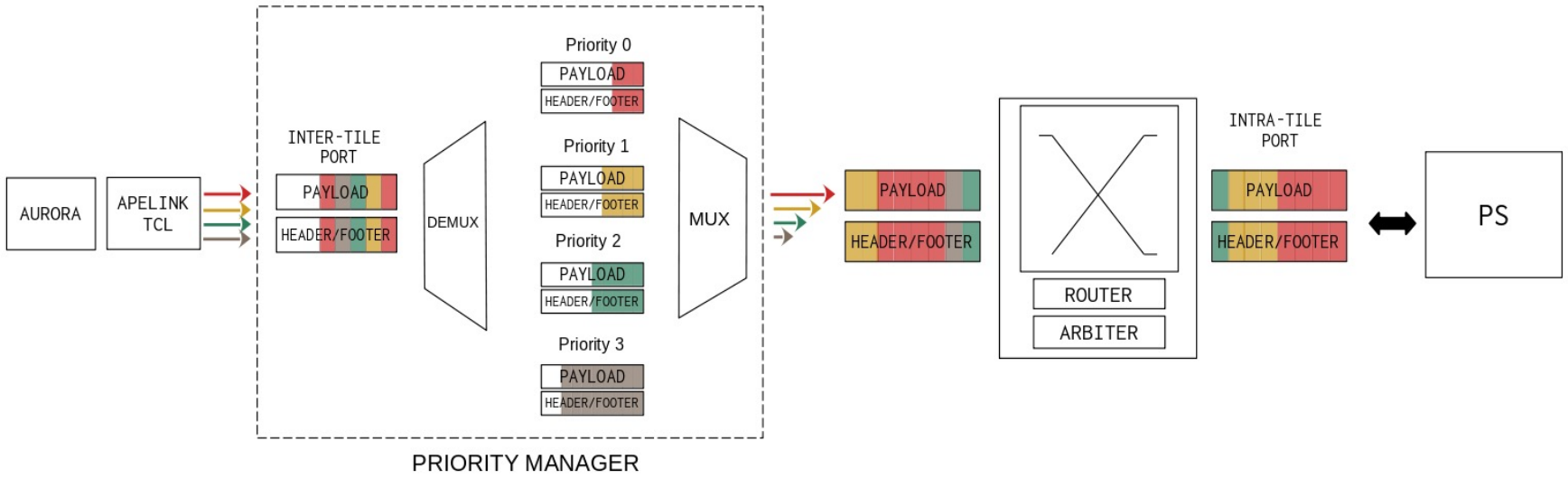
QoS: priority mechanism

Carlotta Chiarini Master Thesis
Investigation and design of a priority management mechanism for data transmission in scientific computing-oriented interconnection networks

Data flow

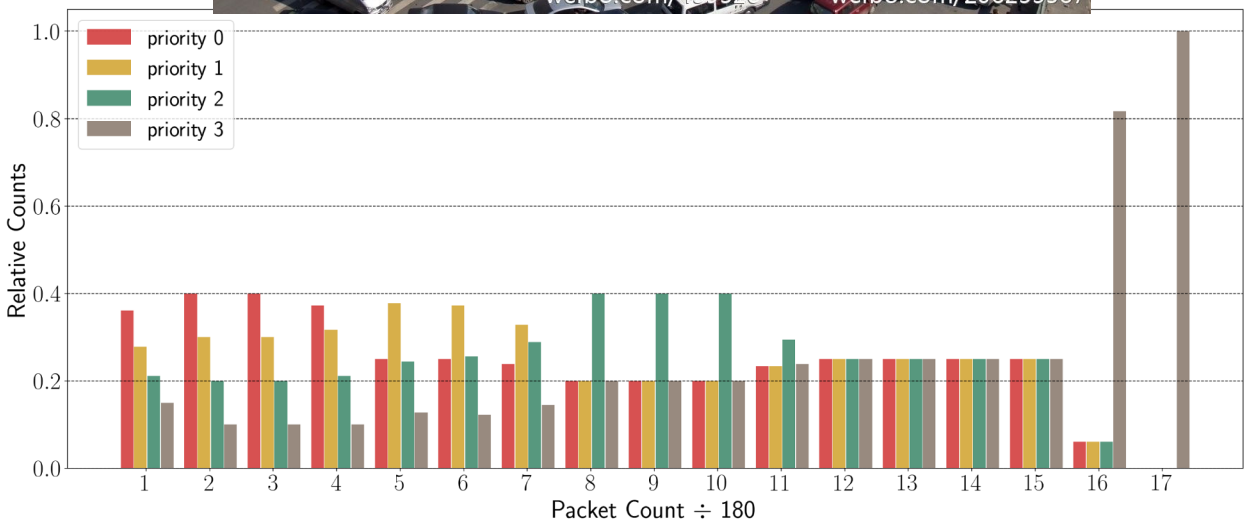


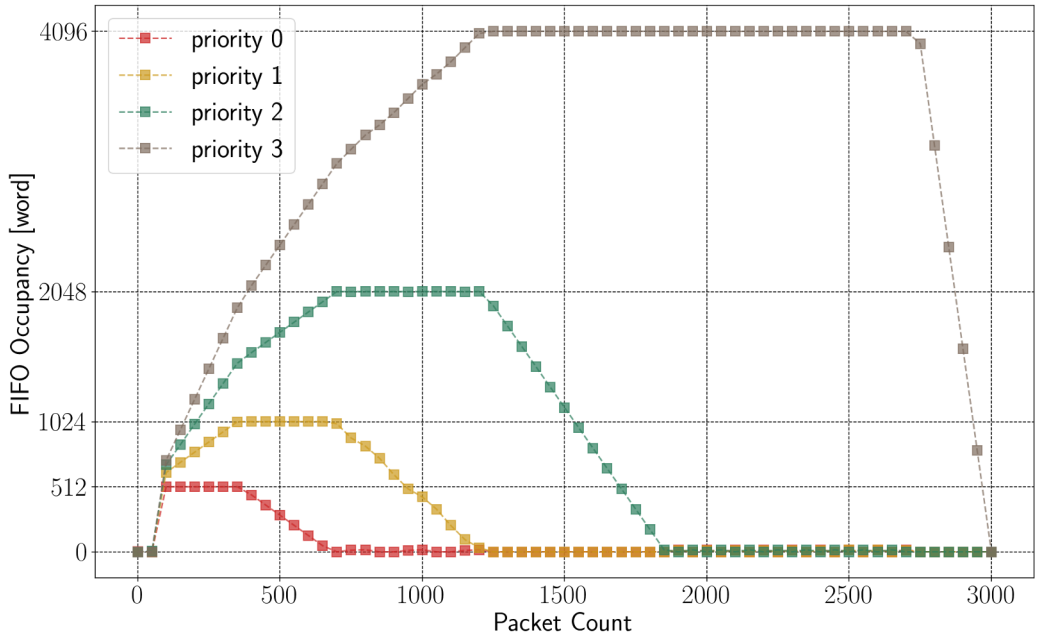
QoS enhanced Data flow



Priority mechanism: FIFO occupancy

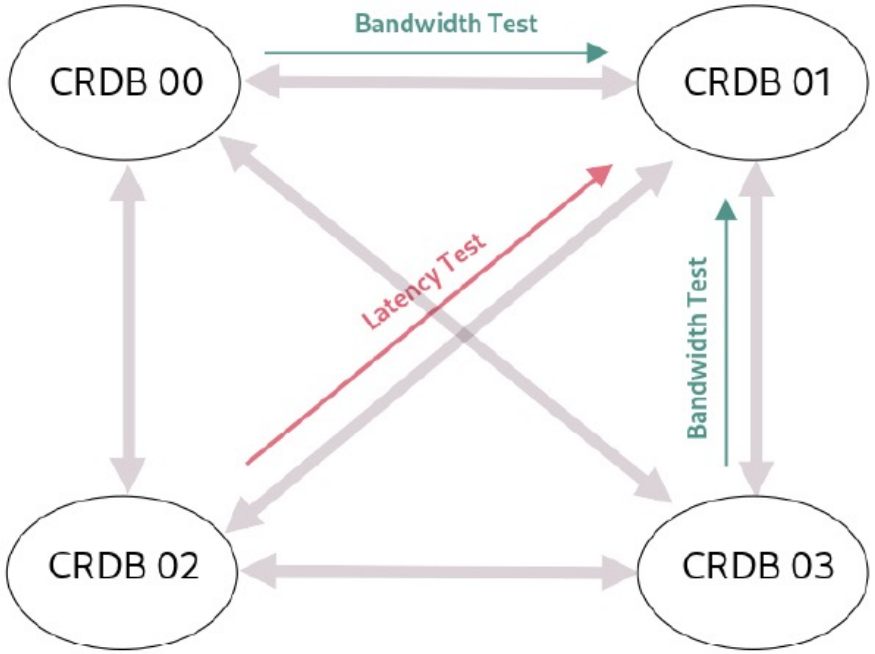
- Congestion
- QoS: data flow priority





Priority IP performance

- Bandwidth test: Low priority (congestion)
- Latency test: High-priority

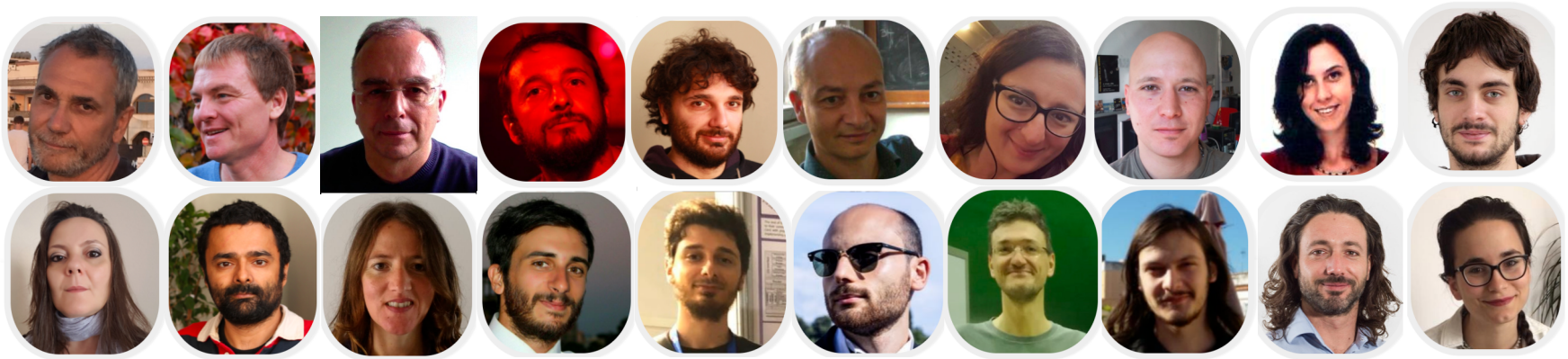


Size [byte]	No congestion [μs]	Priority OFF [μs]	Priority ON [μs]
0	2.40	12.03	7.71
1	2.37	11.98	7.76
2	2.37	11.97	7.77
4	2.35	11.93	7.72
8	2.34	11.89	7.76
16	2.34	11.92	7.74
32	2.35	11.91	7.74
64	10.78	58.58	37.95
128	10.98	59.34	38.56

Gain of 35%!

https://apegate.roma1.infn.it/

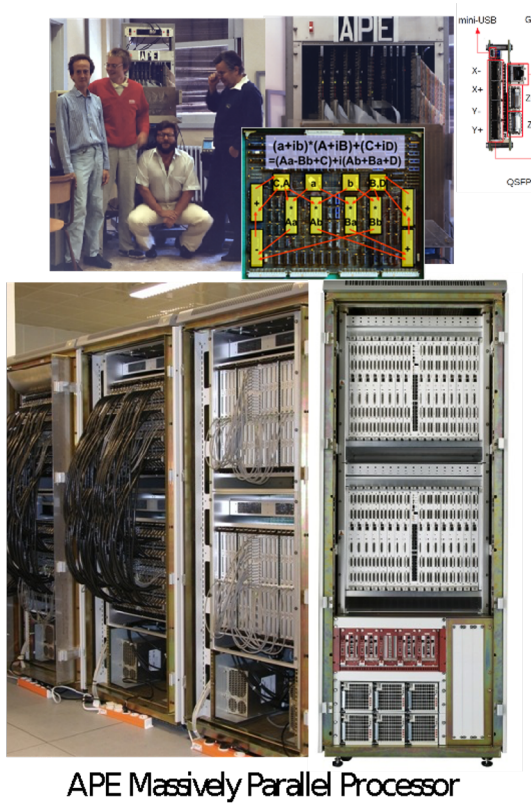
@APELab_INFN



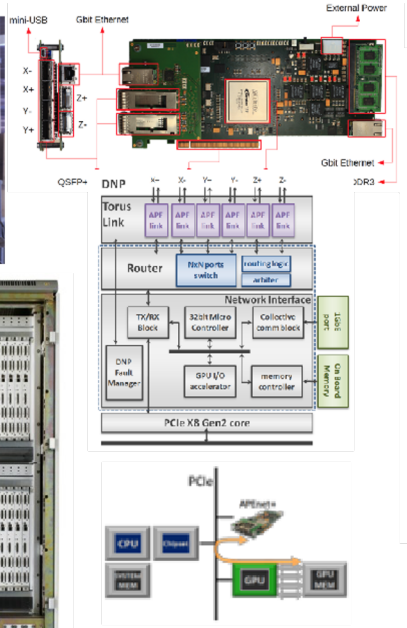
Thank you!!!

A bit of history & current R&D activities

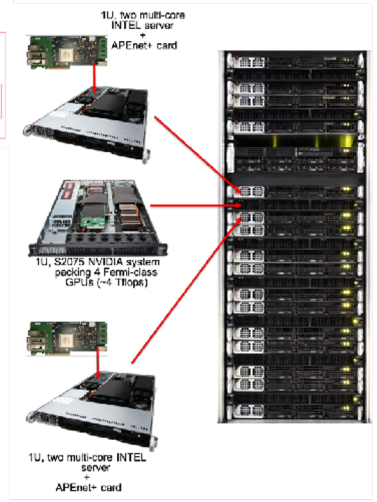
Our Legacy: MPP, Network design, Hybrid Systems



APE Massively Parallel Processor



APENet(+): torus network



QuOnG: hybrid cluster

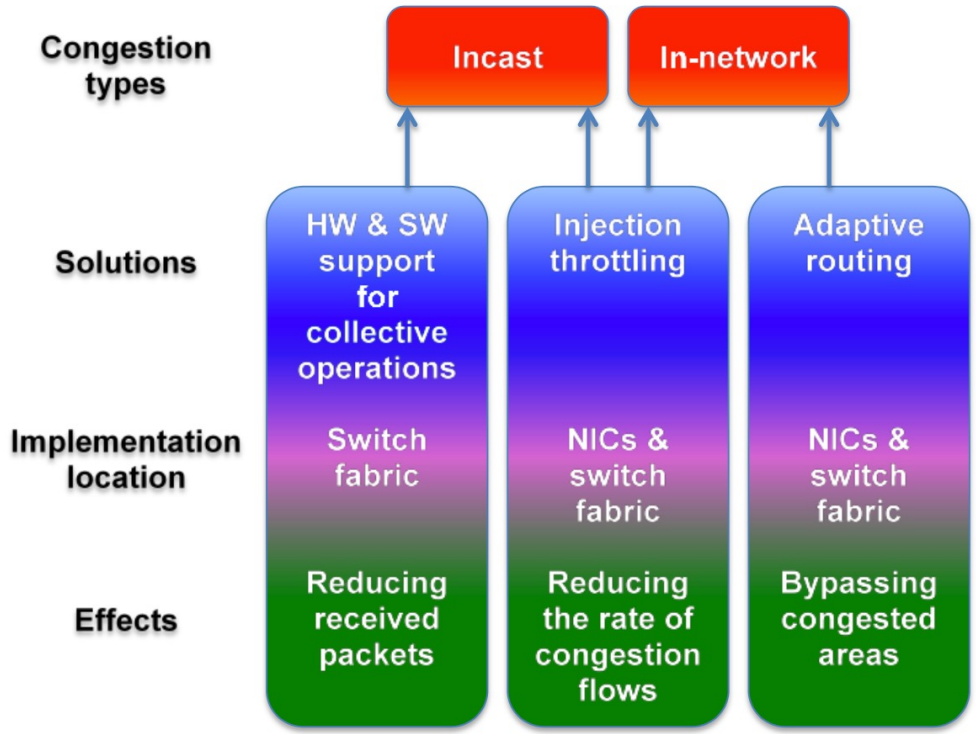
- RED-SEA & TextaRossa (EuroHPC 2021-2024)
  - TextaRossa: achieve extreme computing efficiency for heterogenous scalable HPC platforms
  - RED-SEA: scalable network for ExaScale systems
- ExaNeSt & EuroEXA (H2020 2015–2021)
  - Co-design and platform benchmarking activities: DPSNN, LBM
  - Design and prototyping of FPGA-based, direct network architecture: ExaNet, Custom Switch
- HBP & Wavescales (2016 – 2023)
  - Understand physical mechanism of cognition and brainstates
  - Development of a distributed, parallel and scalable spiking neural network simulator
- NaNet/APEIRON (2020 -)
  - FPGA-based stream computing and GPU-based online low-level trigger for HEP (NA62)
  - Programming Model based on Kahn Process Networks (KPNs), DNN and Spiking Network as reference approach for trigger, implementation via HLS language

RED-SEA novelties

- Physical Layer
  - The project focuses on the development of MAC and PCS modular IPs which can be reused for both Ethernet links and future BXI links, targetting 200Gb/s link per direction which are made of four independent differential lanes running at 56Gb/s.
- Transport Layer
  - 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.
- Host Interface
  - have a direct access to the low power processor cores via a coherent interface to reduce latency and simplify the software interface. (FPGA prototype of the direct NI to Arm and RISC-V)
  - optionally integrate a PCIExpress 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

RED-SEA novelties (2)

- Sofware environment
  - the software stack and the libraries to take advantage of the BXI offloading capabilities such as high-performance collective operations
  - new worldwide reference for benchmarking the efficiency of the communication using a smart network offering offloaded functions
- QoS and congestion management
  - fine grain and medium grain adaptive routing, smart and responsive congestion management and highly flexible QoS (key decision made by NIC and the network switches
    - First, the protocol definition and the specification of the hardware probes to monitor the status of the Fabric.
    - Second, the algorithms to make the best decisions for adaptive routing and injection throttling.
    - Third, the support for congestion management tailored for collective operations.



RED-SEA novelties (3)

- Ethernet Gateway
  - 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.
    - on Ethernet side using the RoCE semantics,
    - on HPC fabric side with better support for Ethernet over BXI
    - bridging solution can be integrated inside each port of the switch ASIC
- Low-latency Ethernet IPs
  - 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