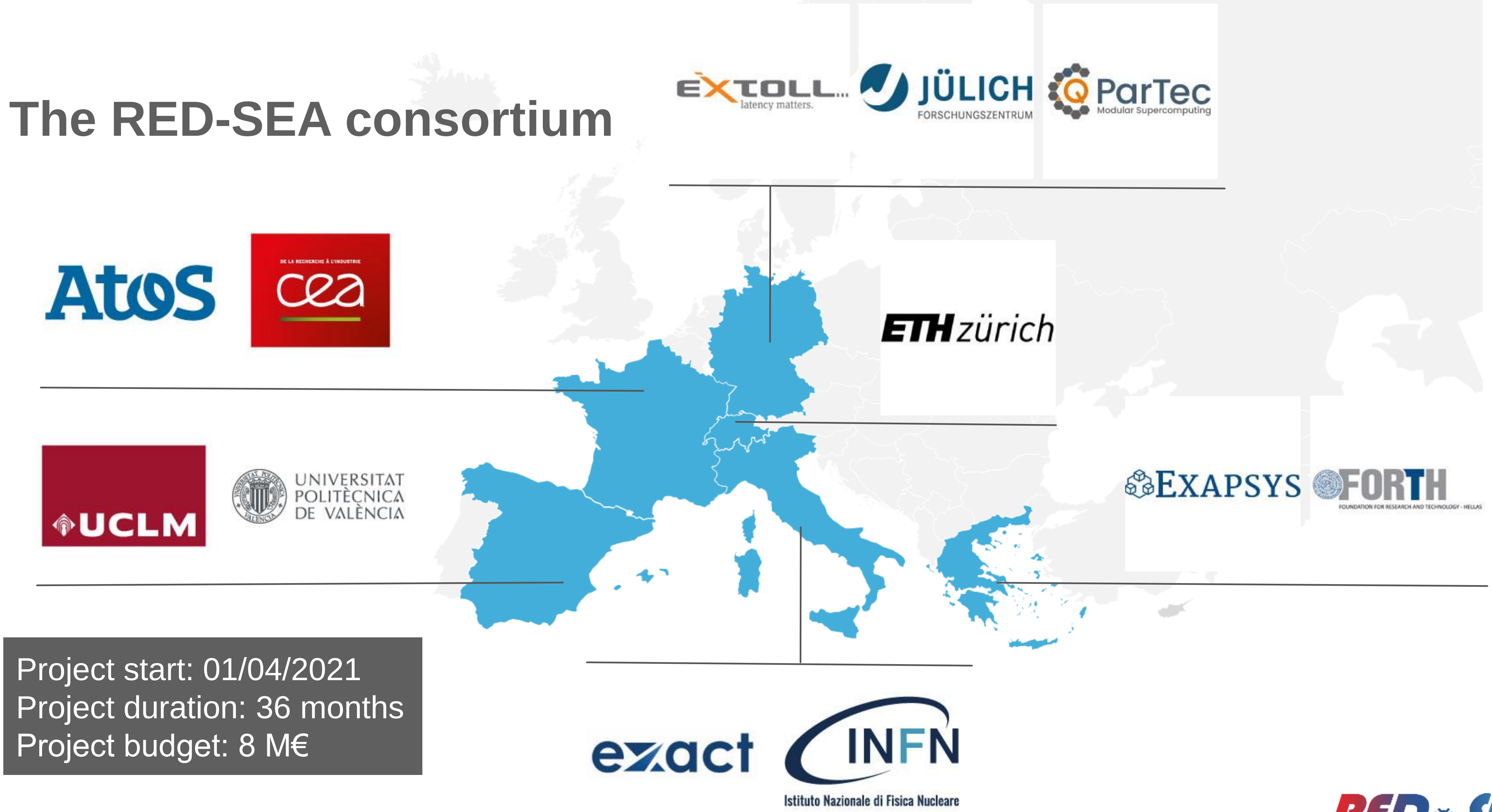
redsea-project.eu | @redsea_EU | @redsea-project

RED-SEA overview

Pedro J. García & Jesús Escudero-Sahuquillo (UCLM)

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.

The RED-SEA consortium


EXTOLL latency matters.
JÜLICH FORSCHUNGSZENTRUM
ParTec Modular Supercomputing
Atos
DE LA RECHERCHE À L'INDUSTRIE cea
ETH zürich
UCLM
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
EXAPSYS
FORTH FOUNDATION FOR RESEARCH AND TECHNOLOGY - HELLAS
exact
INFN Istituto Nazionale di Fisica Nucleare

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 are 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 motivation

- At Exascale, the **interconnect can become the bottleneck**
  - Number of components and their heterogeneity is increasing, requirements are diverse
- Crucial aspects for the network:
  - **Scalability, reliability**: beyond 100K nodes keeping key performance and reliability
  - **Sustainability, HPC/datacenter convergence**
    - integrate Internet Protocol (IP) and Ethernet and RoCE (RDMA over Converged Ethernet) traffic over the HPC interconnect, at low latency and high message rates
  - **Throughput & bandwidth**: ×4 BW and message rate for each endpoint of the network
    - ×2 link frequency (up to 200Gb/s) and ×2 network interfaces per process (multi-rail)
  - **Congestion control, quality of service, isolation, protection, sharing:** partition existing HPC system into multiple (private) clouds
  - **Programmability, latency**: configure the network offload engine, enable compute-in-network, better latency and energy efficiency.
- **Overall goal**: **extend and optimize BXI interconnect for Exascale**

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

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

INFN Istituto Nazionale di Fisica Nucleare

High-performance Ethernet

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

Atos

Efficient Network Resource management

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

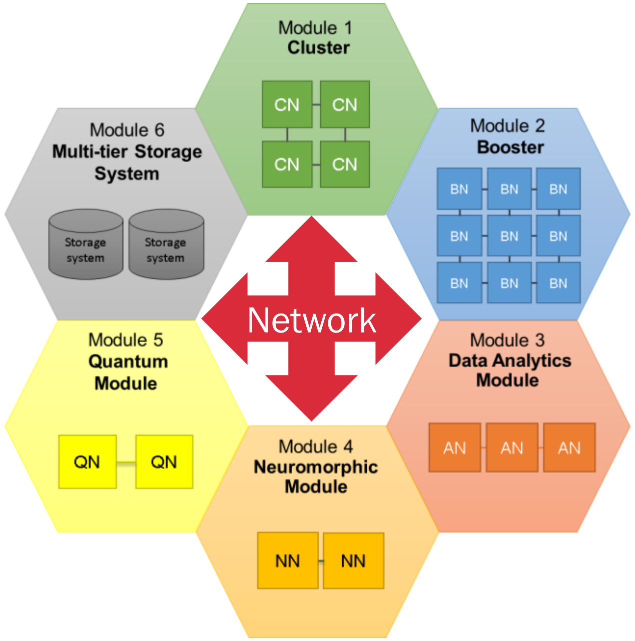
Universitat Politècnica de València

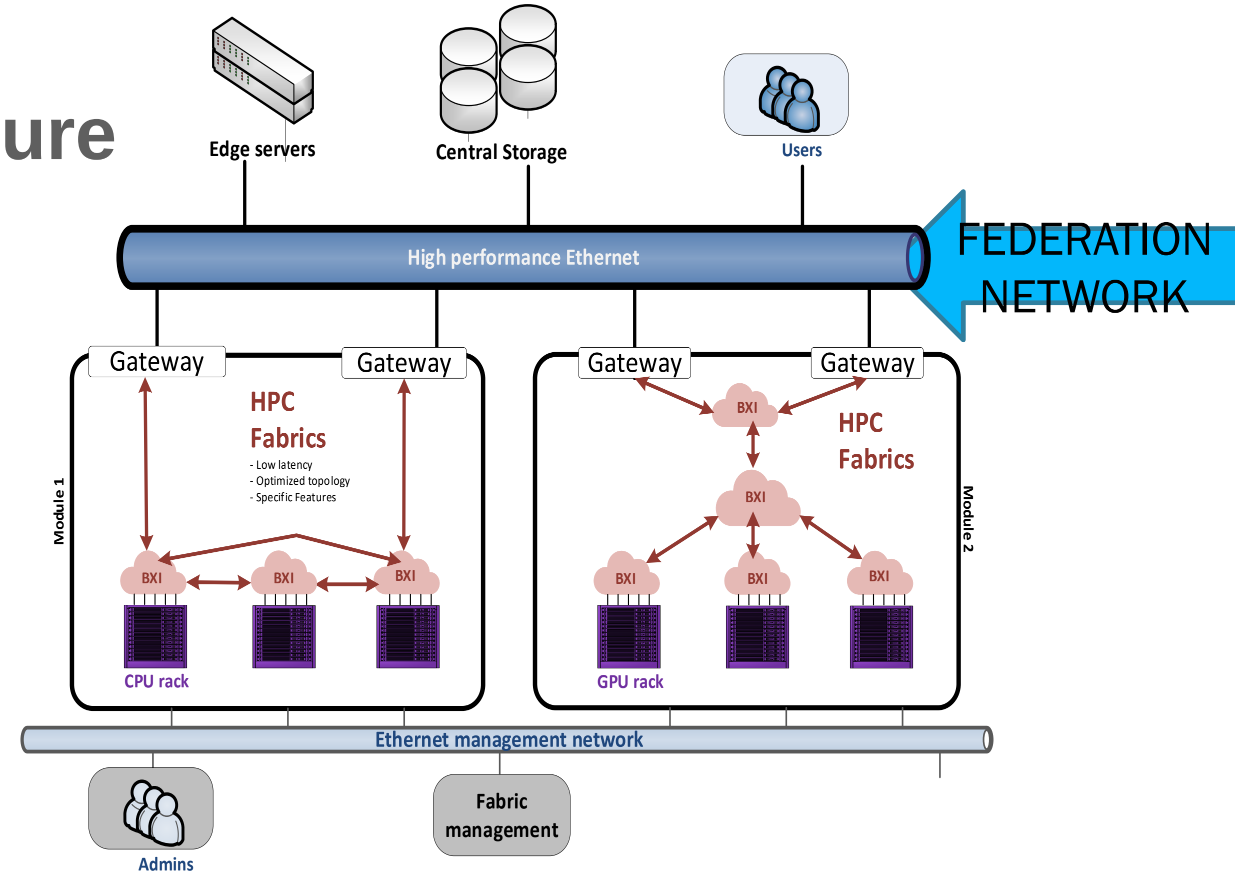
Endpoint functions and reliability

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

FORTH

RED-SEA: MSA network architecture



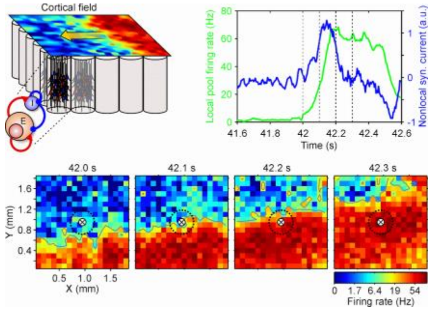


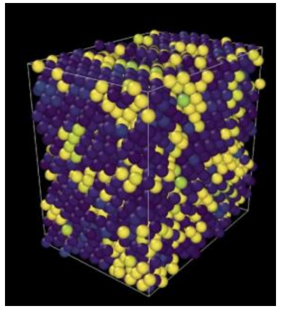
- HPC (High Performance Computing) ; HPDA (High-Performance Data Analytics); AI (Artificial Intelligence)
- Supercomputer: aggregation of resources that are organized to facilitate the mapping of applicative workflows
- HPC is part of the continuum of computing

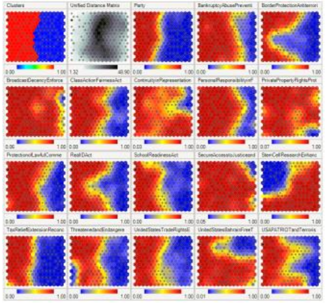
- High performance Ethernet as federation network featuring state-of-the-art low latency RDMA communication semantics;
- BXI as the HPC fabric consisting of two discrete components, a BXI NIC plus a BXI switch, and the BXI fabric manager.

RED-SEA: methodology for Co-Design Activity

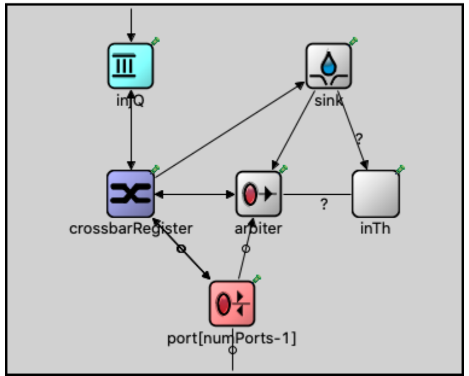
- **Application portfolio**
  - NEST: simulator for spiking neural network models
  - LAMMPS: molecular dynamic engine with focus on material modelling
  - SOM: artificial neural networks used in the context of unsupervised ML







- **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 100K nodes)



RED-SEA: Hardware Testbeds

TESTBED	Features	Outcome	Availability Date	Remote Access
DIBONA	4 blades; 768 Arm v8 cores (12 nodes) OS: RHEL 8.4 Memory: 256GB per Node: 16x16GB DDR4@2666MT/s	Analysis of BXI 1.3 • net. Traces of apps • benchmarks	16 November 2021	YES
DEEPcluster	2CN + BXI switch	T1.2: partec	Q4 2021	YES
ExaNeSt	64 arm cores; 16 QFDB; 4 mezzanines	Prototype of FORTH RDMA + cong. mgmt	Q4 2021	NO
INFN-dev	Alveo board (u50; u200; U280) PCIe gen3/gen4 I/O 100gbps (APElink; BXI-link) ExaNet protocol compliant	• Prototype of APEnetX • Debug & development INFN WP3 and WP4 IPs	Q3 2021 APEnet v6 (0.1): Q4 2022	NO
TGCC KNL	828 nodes (276 blades) Intel(R) Xeon Phi(TM) CPU 7250 96 Go of memory (6x16) + 16 Go mcdram OS: RHEL-7.9; interconnect: BXI v1.2	VEF traces / BXI traces	now (only to CEA partner and subject to quota availability)	YES (up to 14/11/22)
INTI-BXI	nodes (AMD rome); 2*64 cores/node Mem: 240Go /node 4 BXI NICs /node	WP4 – T4.5 multirail	Q1 2022	No Only to CEA

RED-SEA: Simulators (I)

Simulator (partner)	Features	Tasks involved
COSSIM (EXAPSYS)	**Current** • Processing in ARM, RISC-V (work-in-progress/eProcessor), Intel (deprecated) • Network topologies, routing algos, switches, etc are those supported by OMNET++ • Main change of OMNET++ has to do with INET packages that have been adapted so as to support full IP, Linux-compatible packets (e.g. including payload) **RED-SEA:** • NIC Architectural model with several implementation details needed • Interconnection scheme of CPU with NIC	**T1.4 :** MPI packets generated in COSSIM can be integrated in SAURON (VEF Traces) Identify if COSSIM can be connected to SAURON instead of plain OMNET++ From **WP2** get NIC design compatible with GEM5
SAURON (UCLM)	**Current:** • Network topologies: Fat-trees, Dragonflies, Slim-flies, KNS, etc. • Routing algorithms: deterministic (D-mod-K, DESTRO), Oblivious (VLB), and adaptive (PAR, UGAL, Fully, ARNs, etc.) • Switch buffer organizations (input-queued, virtual output queues, etc.) • Congestion management and QoS models • Compatible with **VEF Traces Framework** **RED-SEA:** • BXI3 Architecture (NIC and switch) • Protocols designed in WP3 and WP4	**T1.4 :** - Migration to OMNET++ 6.0 - Exploring connection with COSSIM All the tasks in **WP3:** modeling new network management proposals **T4.1:** modeling e2e protocols

RED-SEA: Simulators (II)

Simulator (partner)	Features	Tasks involved
DQN_SIM (INFN)	**Current** • Simulation models developed from scratch using the OMNeT++ 5.4 framework • N-dim Torus Topology • Modelled after the APEnet RDMA network architecture: data-link layer (buffers, virtual channels), network layer (VCT switching, deterministic routing (DOR), Oblivious (random) and Adaptive Routing (*ch, DQN-Routing), transport layer (packet definition, network interface). • Interface between OMNeT++ and the Ray distributed execution framework to exploit its services in order to get routing actions from the Deep Q-Network reinforcement learning agent. **RED-SEA:** • Port the models to the SAURON framework in order to assess DQN scalability and performance under realistic traffic conditions • Study the application of the DQN adaptive routing algo to other topologies and/or network architectures.	

Questions ?