Istituto Nazionale di Fisica Nucleare

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

Outlines in hardware and software for new generations of exascale interconnects

Michele Martinelli, INFN

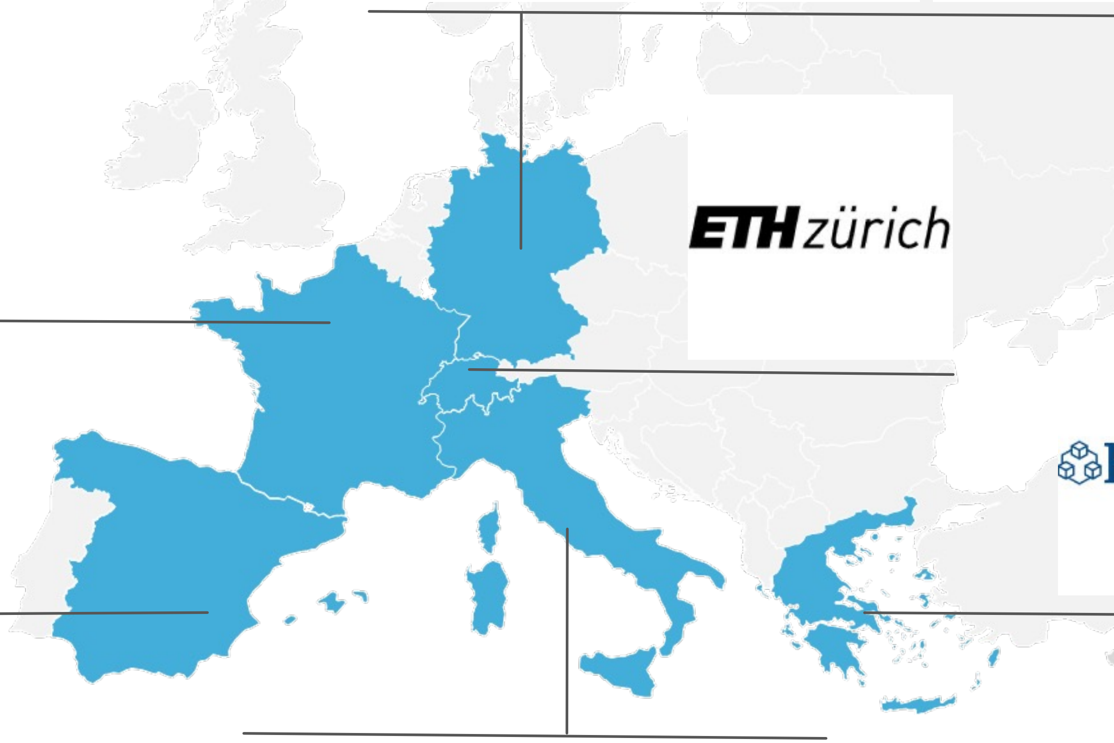
26TH INTERNATIONAL CONFERENCE ON COMPUTING IN HIGH ENERGY & NUCLEAR PHYSICS (CHEP2023)

May, 8 2023

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.

The RED-SEA consortium


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

1. Network and interconnects
2. Performance Evaluation
3. Simulation frameworks
4. Hardware design and tools
5. HPC System and integration

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

RED-SEA objectives

Enable

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

- Exploiting existing European technology both in the academic and industrial field (BXI, ExaNeSt, ExaNet)
- Able to power the future EU Exascale systems

Explore

Explore new innovative solutions

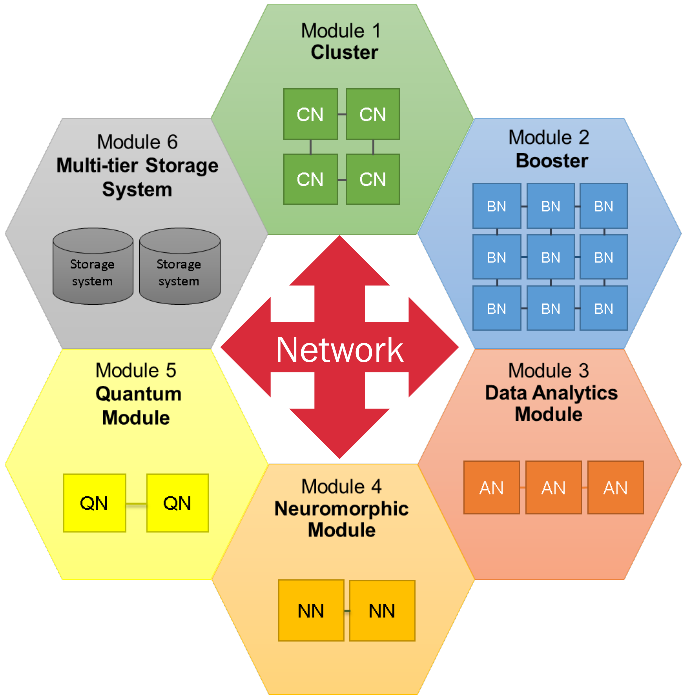
- End-to-end network services – from programming models to reliability, security, low latency, and new processors
- Management of data traffic (congestion generated by collective communication), QoS delivery mechanism

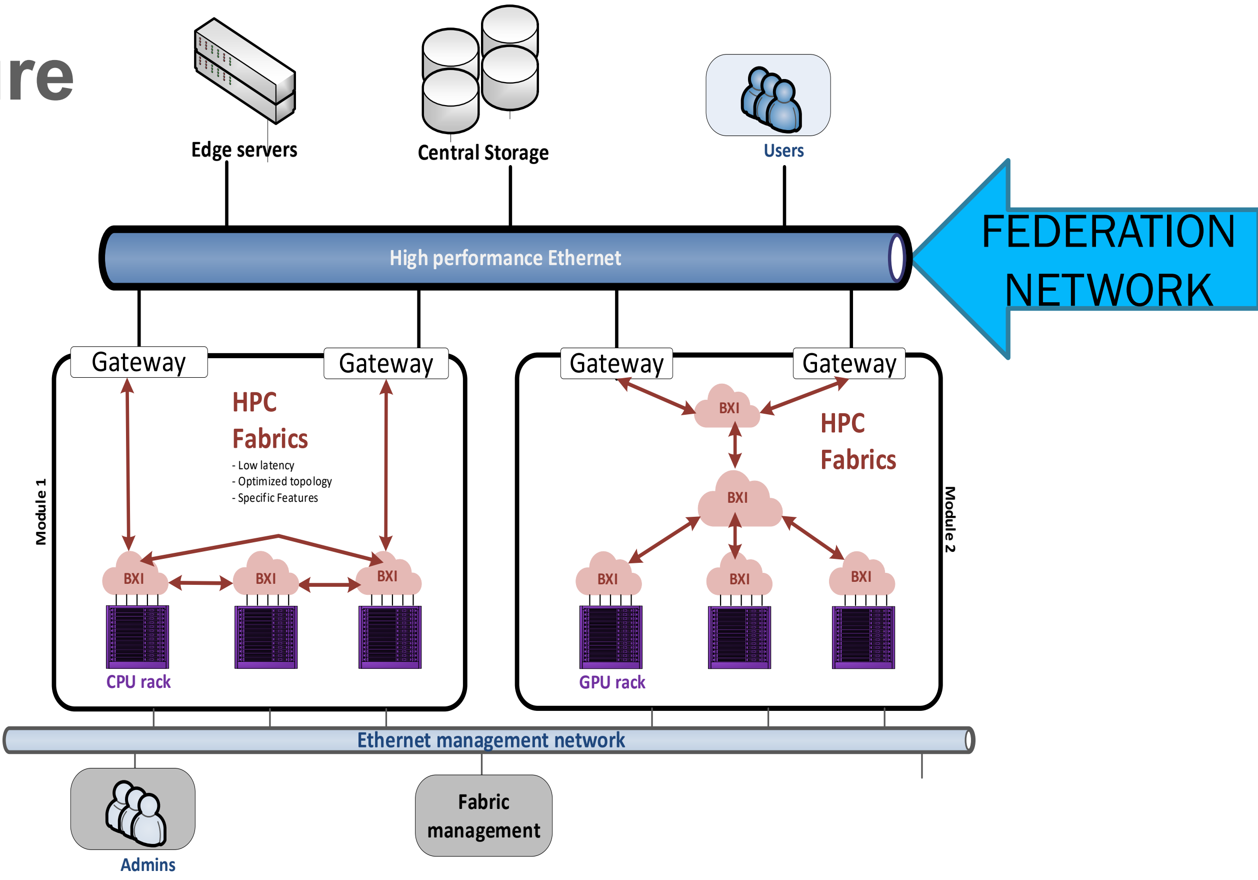
Develop

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

RED-SEA: MSA network architecture



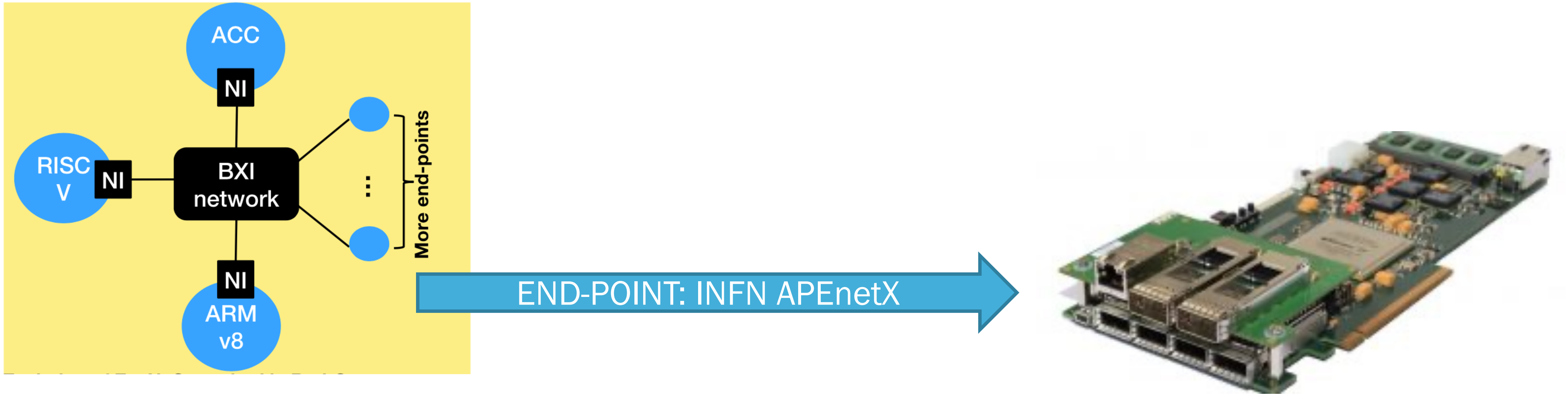


Modular Supercomputer Architecture (MSA)

- aggregation of resources that are organized to facilitate the mapping of applicative workflows
  - HPC (High-Performance Computing)
  - HPDA (High-Performance Data Analytics)
  - AI (Artificial Intelligence)

- 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
  - the BXI fabric manager

BXI ECOSYSTEM: INFN APEnetX



Goals:

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

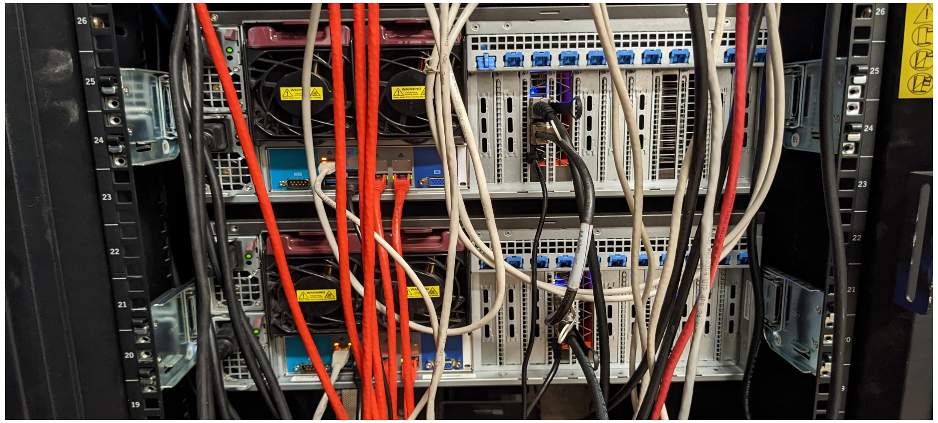
- INFN duties
  - Network Interface Card (APEnetX)

    PCIe gen4 (**RDMA semantics** GPU+CPU) + BXI link on Xilinx Alveo FPGA
  - Co-Design through applications (NEST)

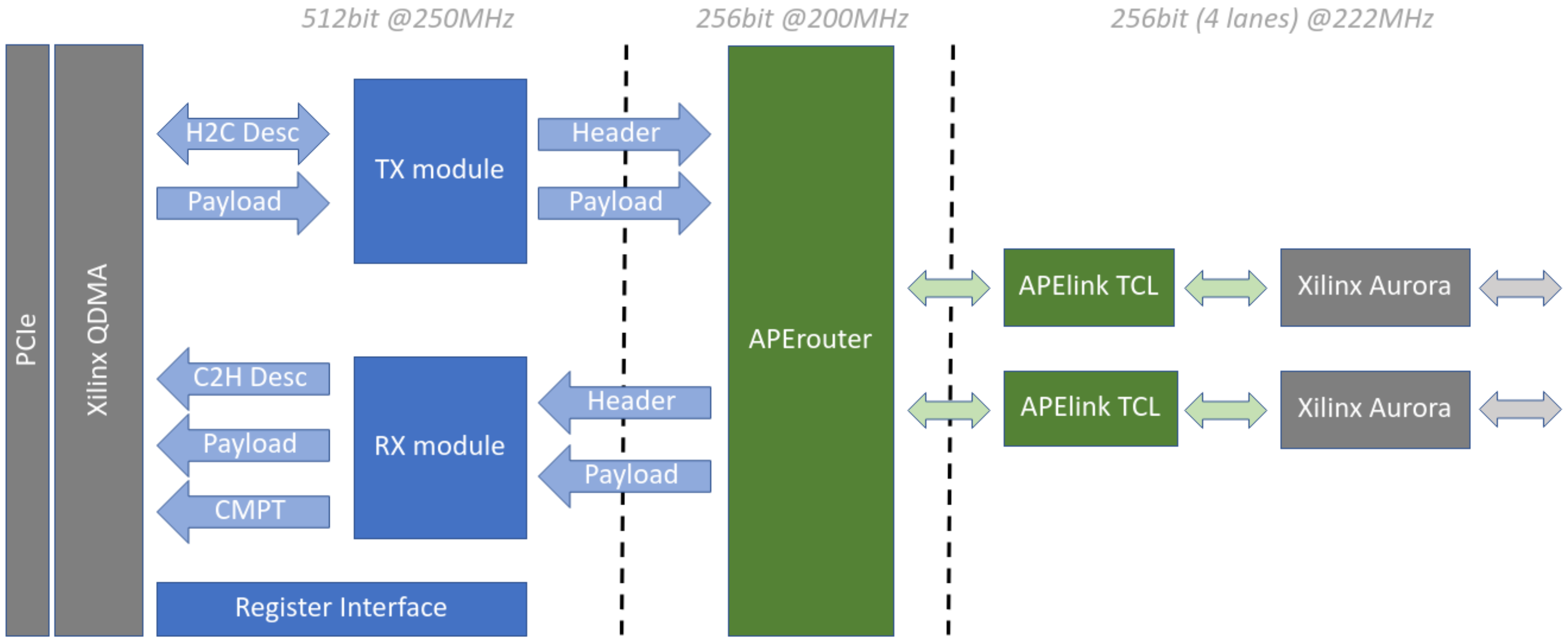
INFN APEnetX TESTBED

- 2x Supermicro SuperWorkstation 7049GP-TRT server
  - 2 x 8-cores 4200-series 14nm Intel Xeon Scalable Silver Processors (Cascade Lake) running @ 2.10GHz.
    - PCI gen3 support (not PCIe gen4)
  - Memory: 192GB DDR4 @3.2GHz
  - **APEnetX prototype**: Xilinx Alveo U200 built on the Xilinx 16nm UltraScale architecture, which natively supports the QDMA IP.
  - OS: GNU/Linux Centos 8 with Linux 4.18.0 kernel.

- connected point-to-point through custom "apelink" protocol with QSFP+ cables



APEnetX HW architecture



PCIe interface + QDMA engine

network interface

routing and switching IP + trasmission control logic

embedded transceivers

QDMA software stack
("Xilinx open source driver" with custom software added)

IOCTL syscall:

- entry-point for our custom sw addition

Register/deregister (pin and lock) RX buffer AND completion buffer

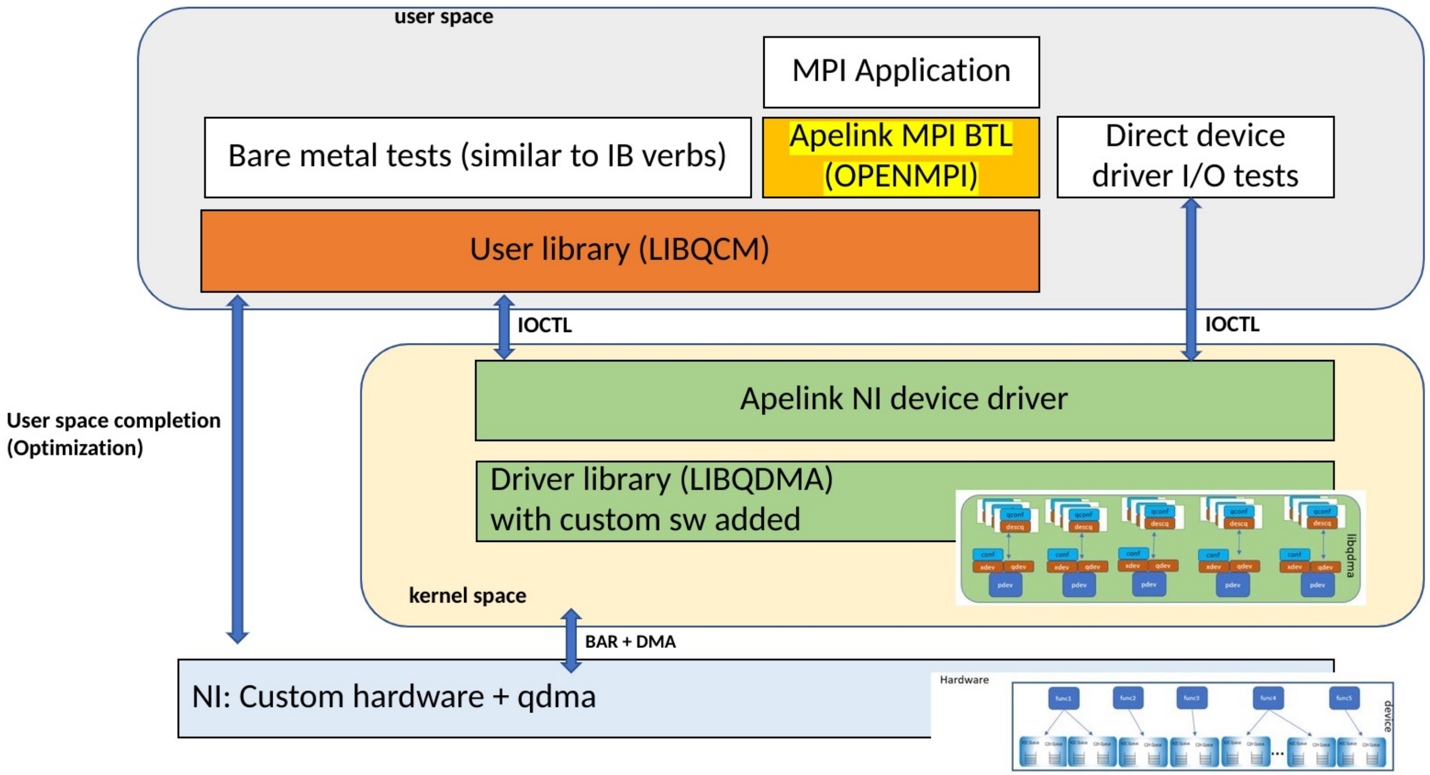
- System IOMMU is used as address translator (p2v v2p)

Sending phase:

- custom TX descriptor

Receiving phase:

- custom completion, notify the user application (polling)



user space

MPI Application

Bare metal tests (similar to IB verbs)

Apelink MPI BTL (OPENMPI)

Direct device driver I/O tests

User library (LIBQCM)

IOCTL

IOCTL

User space completion (Optimization)

Apelink NI device driver

Driver library (LIBQDMA) with custom sw added

kernel space

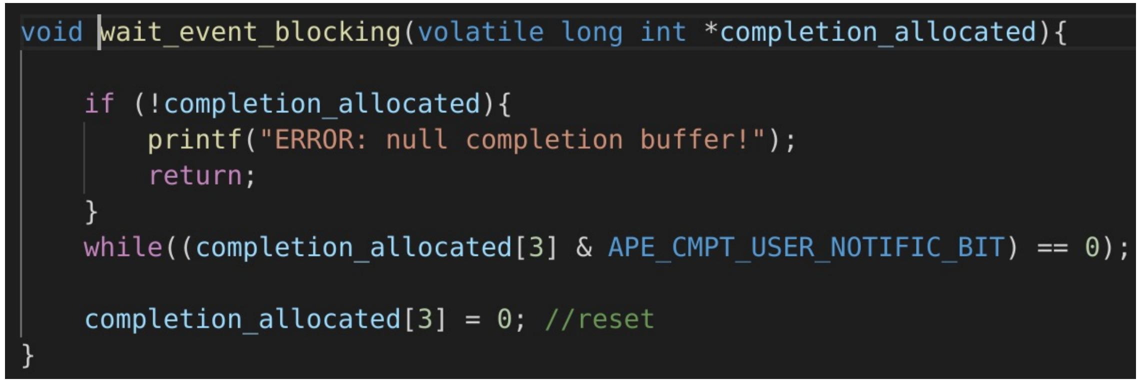
BAR + DMA

NI: Custom hardware + qdma

Completions to userspace

The naïf idea is to have a completion for every RX entry (descriptor) so the userspace sw needs to check a single (kmap-ed) memory area where the kernel copy the new completion

1. User allocates a completion buffer (1 page)
2. Driver remaps this buffer in kernel space
3. When a new completion arrives:
   - the driver copies the completion words to the (remapped) user address
   - Last bit of the completion is used as a signal from kernel to user
   - User data is (asyncronously) DMA-copied in memory

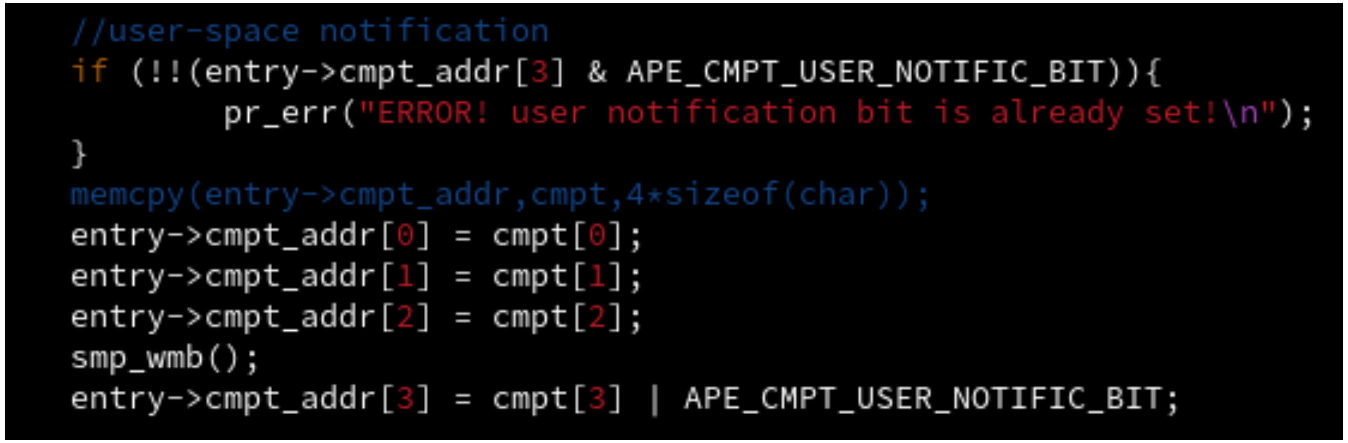


```
void wait_event_blocking(volatile long int *completion_allocated){

    if (!completion_allocated){
        printf("ERROR: null completion buffer!");
        return;
    }
    while((completion_allocated[3] & APE_CMPT_USER_NOTIFIC_BIT) == 0);

    completion_allocated[3] = 0; //reset
}
```

User space (polling)



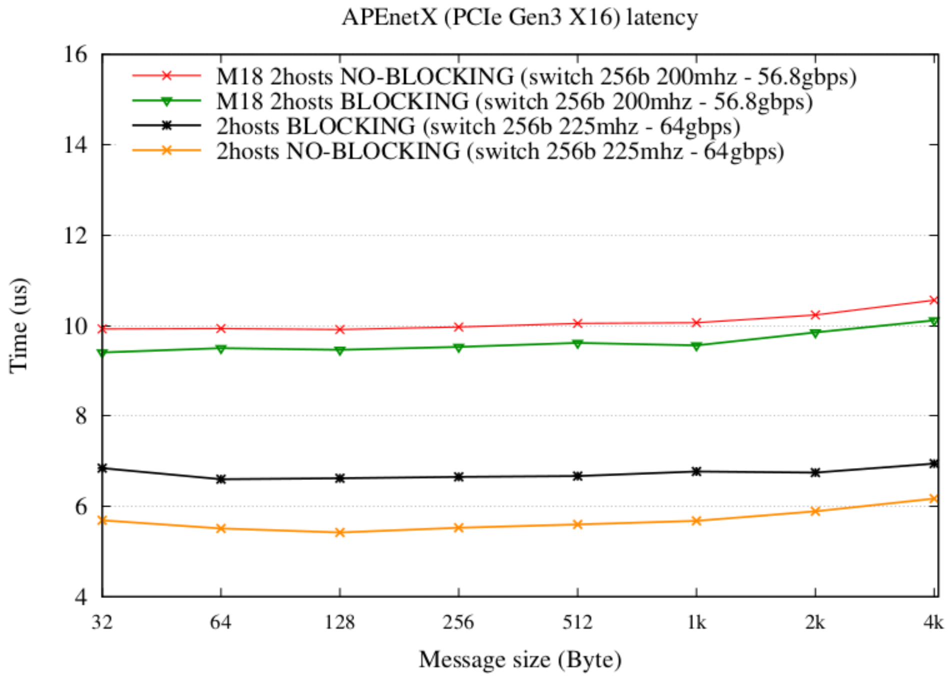
```
//user-space notification
if (!!(entry->cmpt_addr[3] & APE_CMPT_USER_NOTIFIC_BIT)){
        pr_err("ERROR! user notification bit is already set!\n");
}
memcpy(entry->cmpt_addr,cmpt,4*sizeof(char));
entry->cmpt_addr[0] = cmpt[0];
entry->cmpt_addr[1] = cmpt[1];
entry->cmpt_addr[2] = cmpt[2];
smp_wmb();
entry->cmpt_addr[3] = cmpt[3] | APE_CMPT_USER_NOTIFIC_BIT;
```

Kernel space (interrupt context)

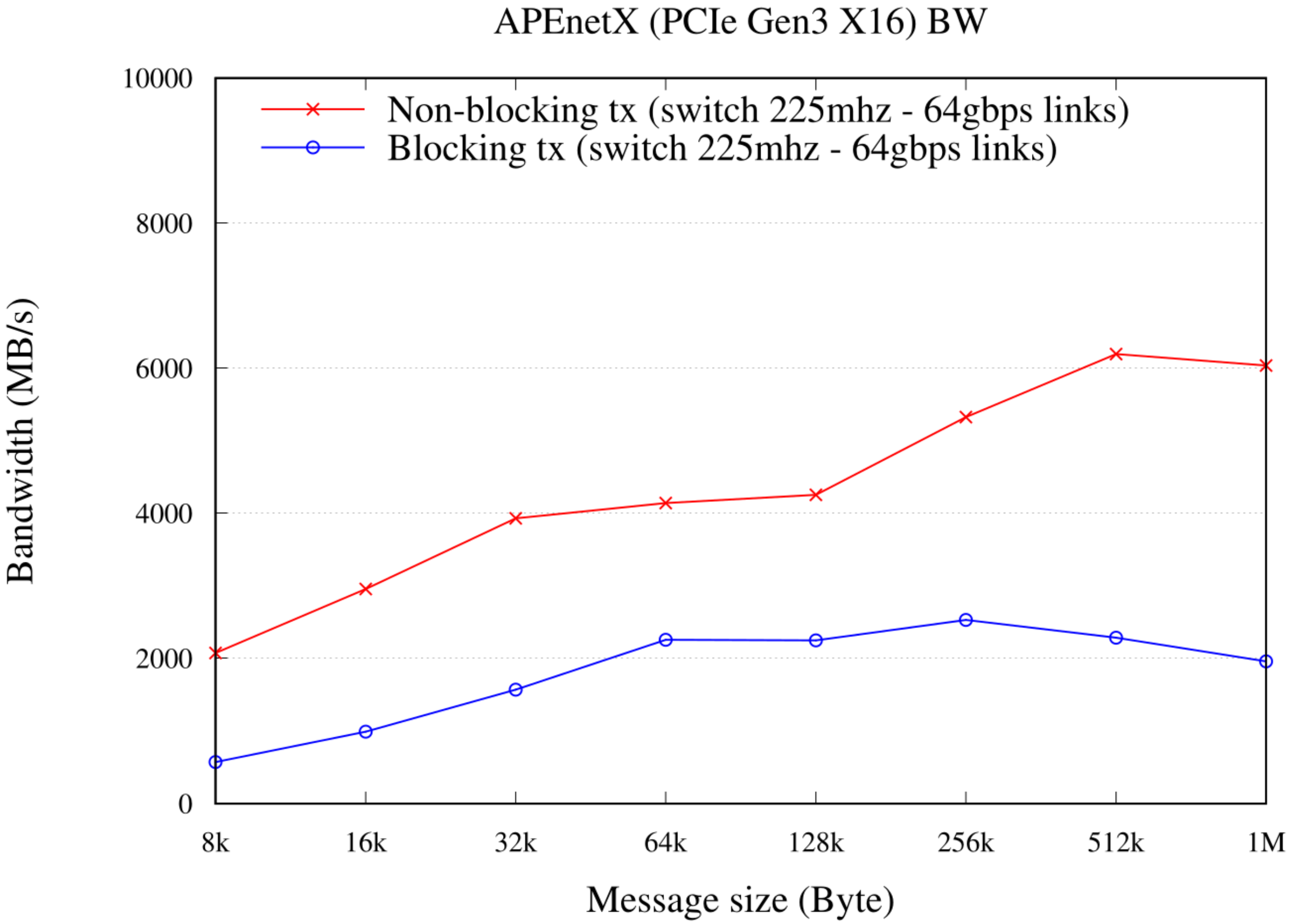
Completions to userspace

Optimization: nodes exchange information on the completion address during the initial handshake

- TX descriptors carries the destination completion vaddr
  - completion vaddr travel with the pkt
- RX HW uses the vaddr to DMA-write the completion
  - no xilinx driver in RX, receiver process polls on its completion addr

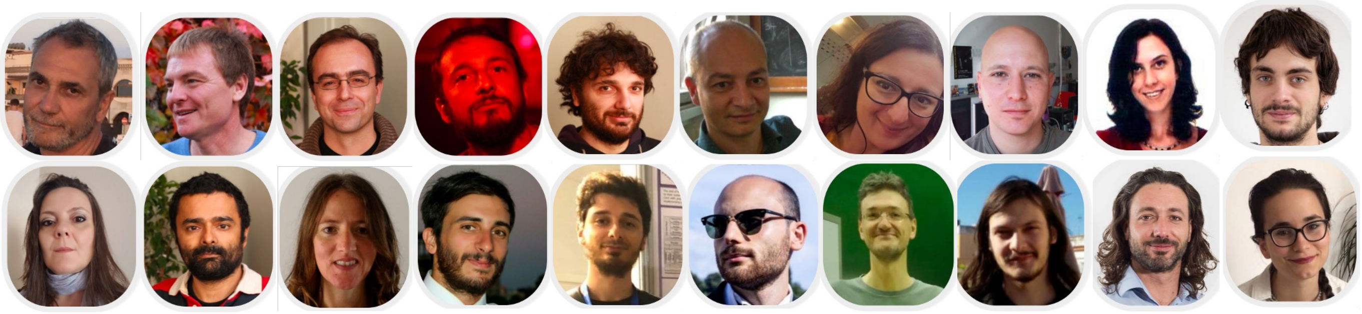


Preliminary results – bandwidth



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

@APELab_INFN

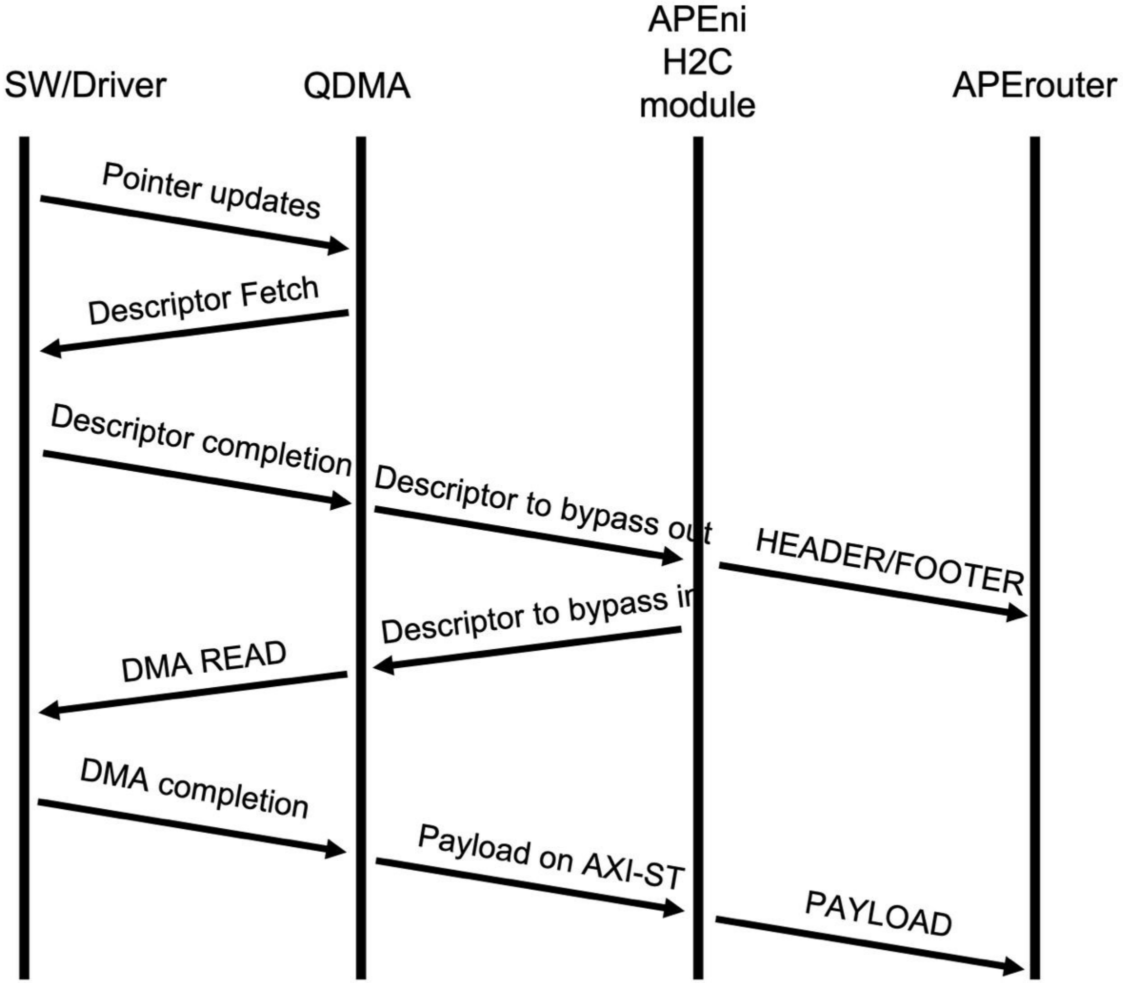


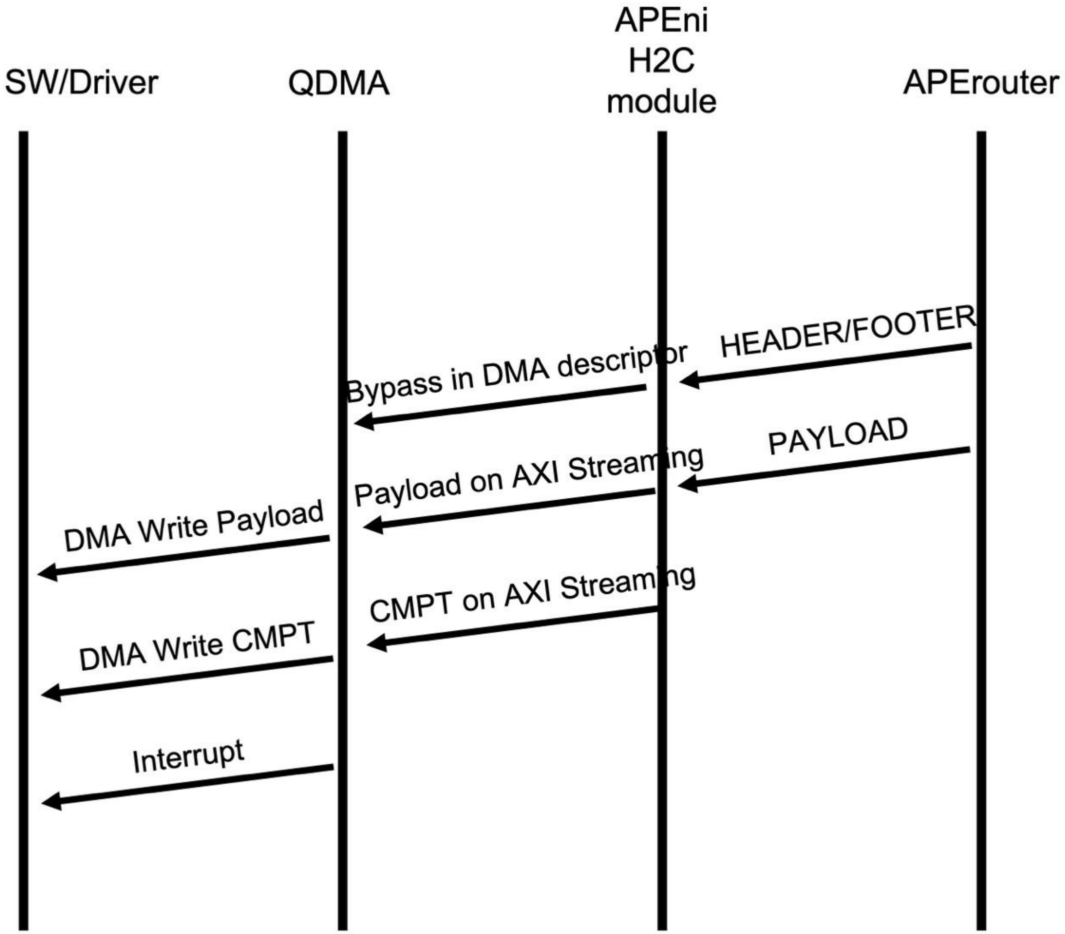
Thank you!

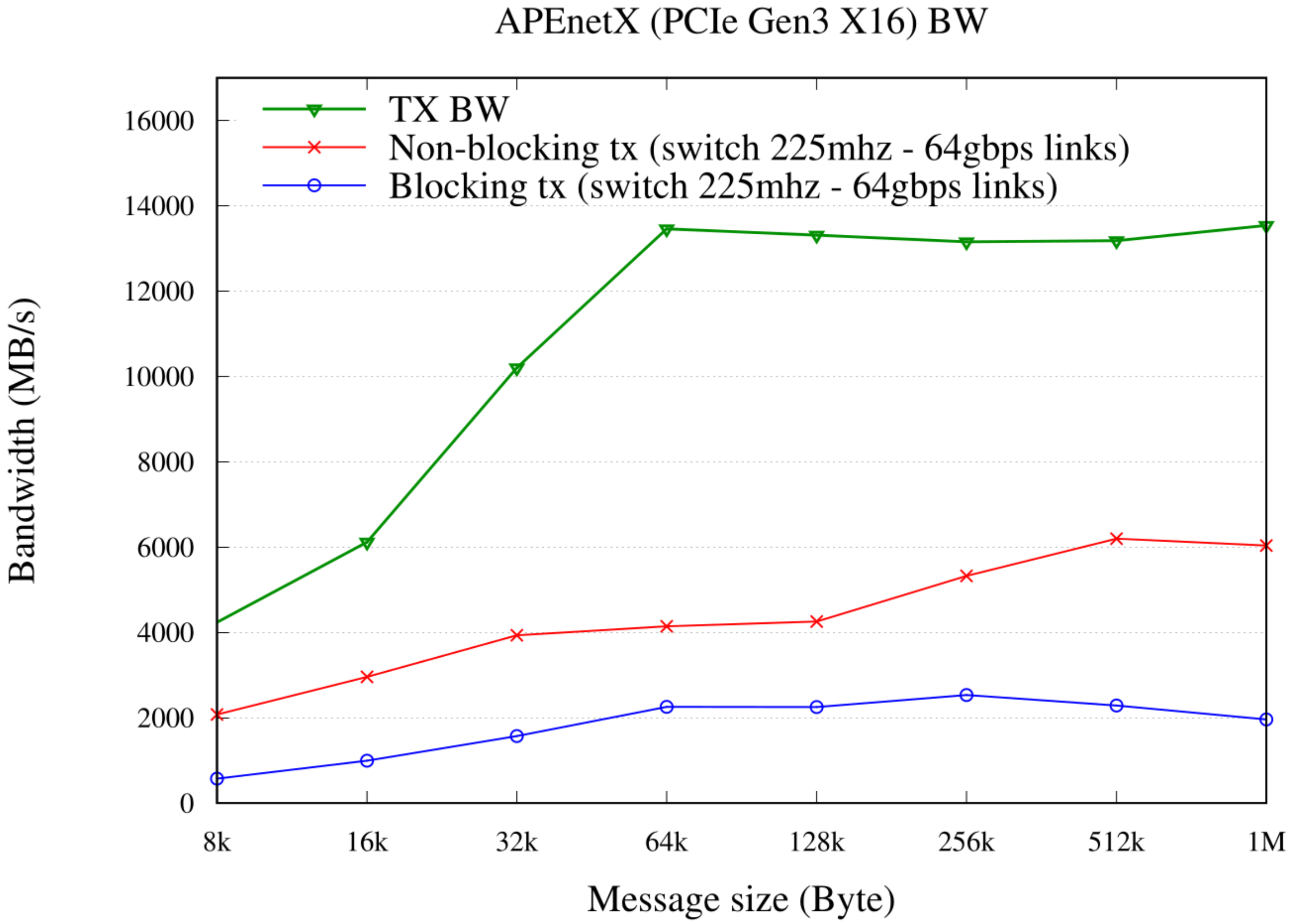
BACKUP SLIDES

Work in progress

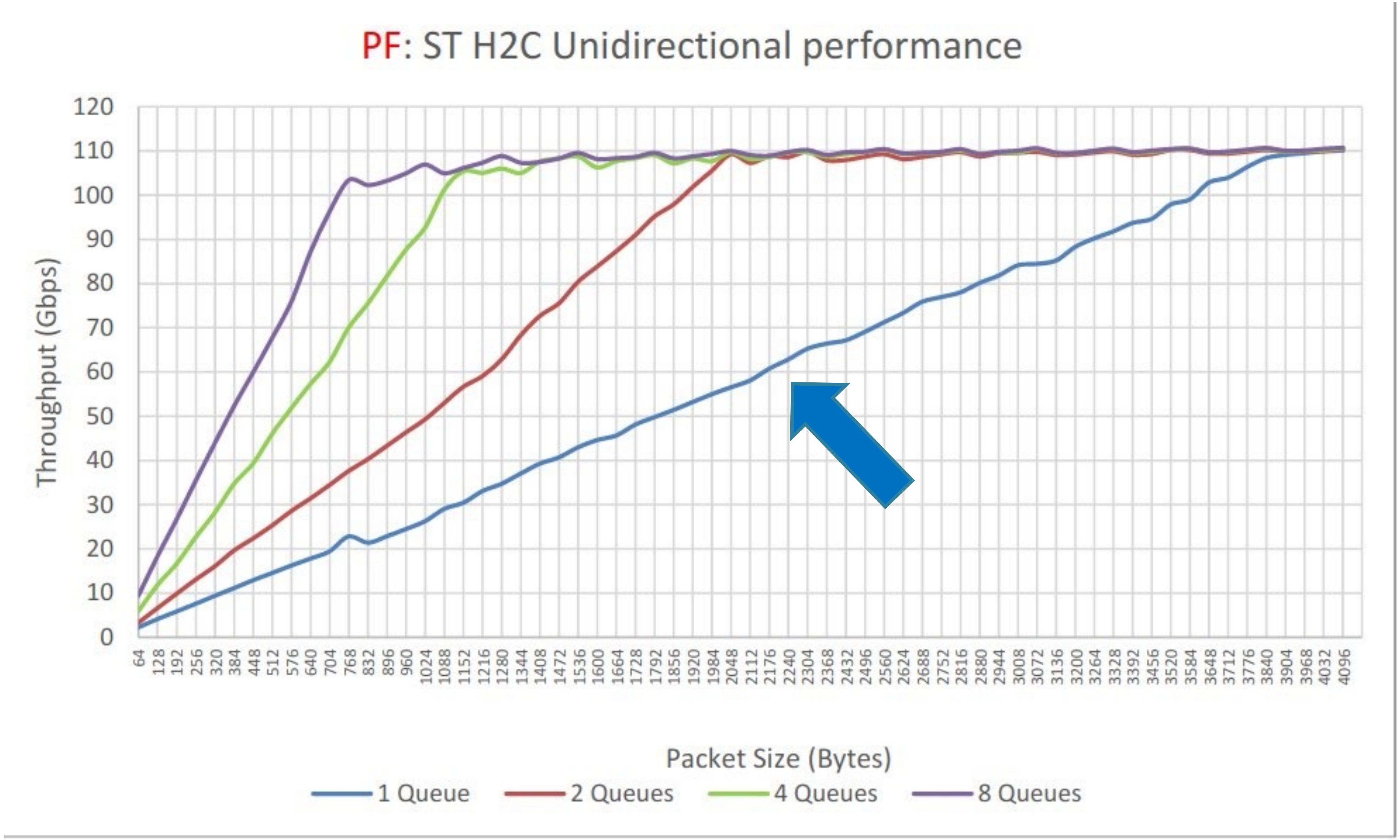
- optimize the performance (especially bw)
- GPU support
- BXI integration
- PCI GEN4
- MPI


SW/Driver
QDMA
APEni H2C module
APErouter
Pointer updates
Descriptor Fetch
Descriptor completion
Descriptor to bypass out
HEADER/FOOTER
Descriptor to bypass in
DMA READ
DMA completion
Payload on AXI-ST
PAYLOAD


SW/Driver
QDMA
APEni H2C module
APErouter
HEADER/FOOTER
Bypass in DMA descriptor
PAYLOAD
Payload on AXI Streaming
DMA Write Payload
CMPT on AXI Streaming
DMA Write CMPT
Interrupt


APEnetX (PCIe Gen3 X16) BW
TX BW
Non-blocking tx (switch 225mhz - 64gbps links)
Blocking tx (switch 225mhz - 64gbps links)
Bandwidth (MB/s)
16000
14000
12000
10000
8000
6000
4000
2000
0
8k
16k
32k
64k
128k
256k
512k
1M
Message size (Byte)

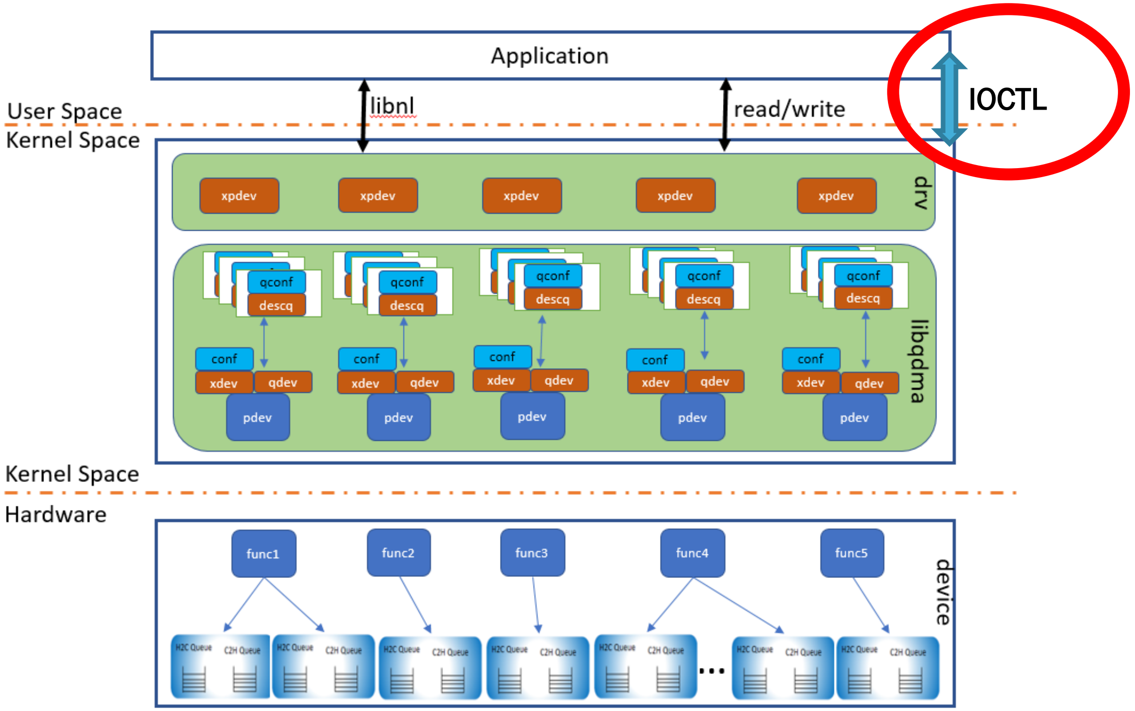
MAX bandwidth

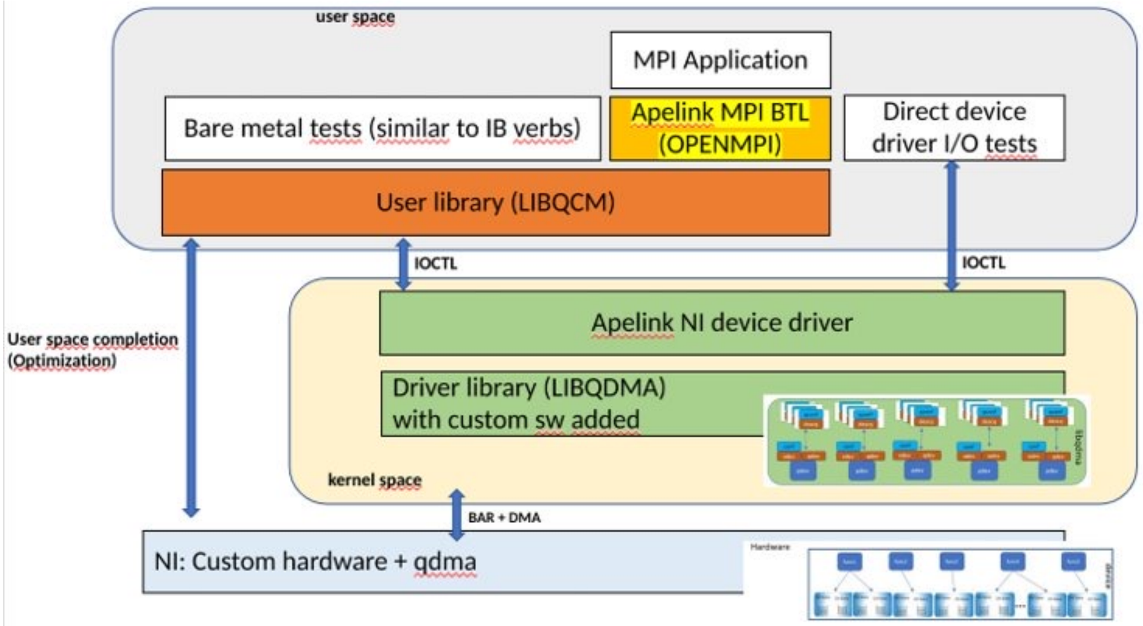


RED-SEA consortium

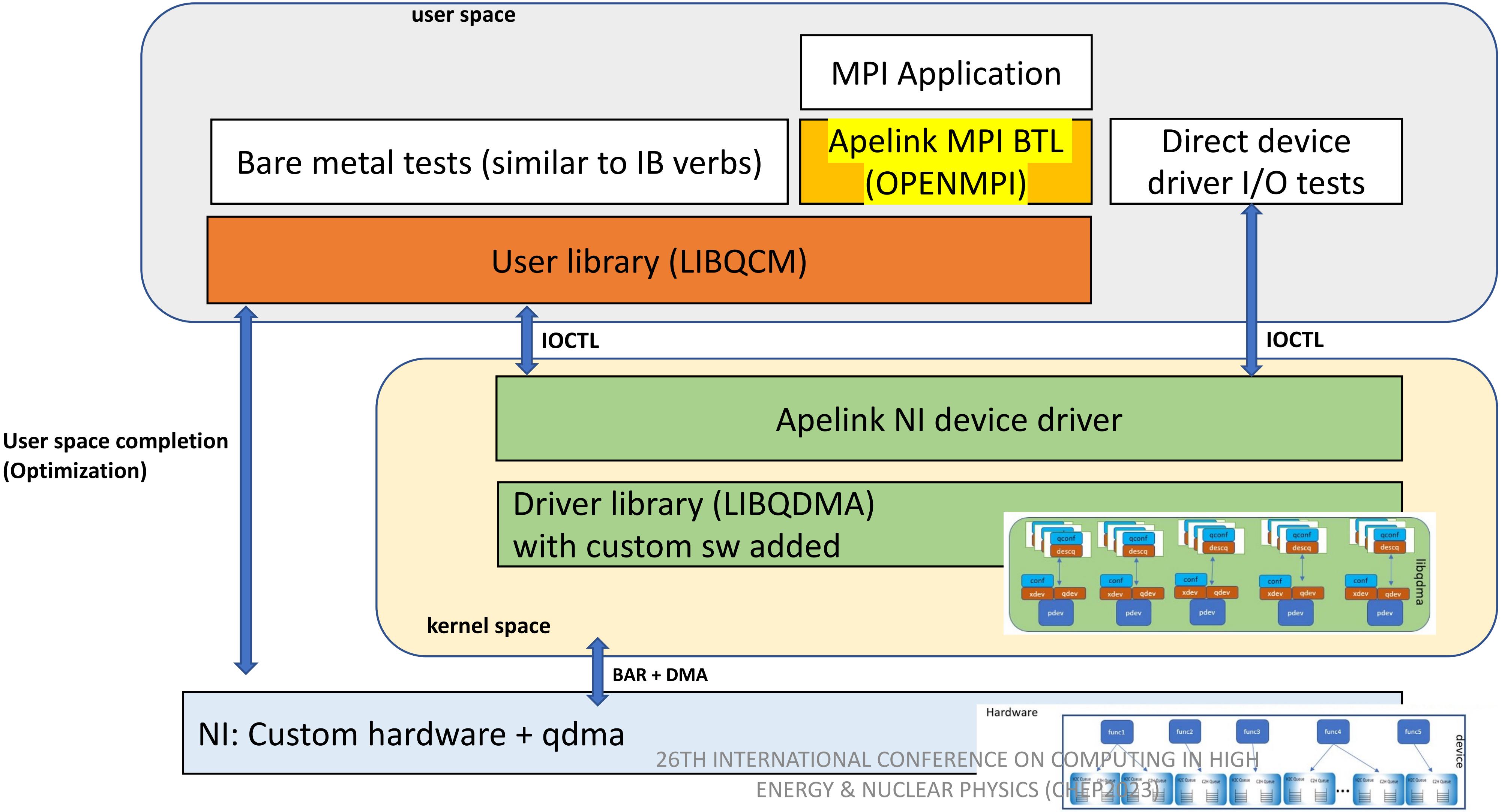
- RED-SEA is coordinated by ATOS / BULL the French industrial partner with a long experience in designing on-chip and off-chip interconnection systems.
- CEA focuses on the optimization and porting of MPI communication libraries, and explores mechanism to allow efficient use of multiple NICs in the same compute node
- UCLM and UPV work on techniques on all aspects of interconnects, defining efficient topologies, adaptive routing strategies, quality of service mechanism
- FORTH works on congestion management mechanism, provides end-to-end flow control and has experience in the integration of HW and SW environment
- ExaPSYS is a young startup with knowledge on simulations and heterogeneous systems.
- Julich, is one of the most important computing centers in Europe and in the project is the responsible for the development of a benchmarking and performance evaluation tool
- Partec develops MPI communication libraries
- EXTOLL, a spin-off from University of Heidelberg, offers low-latency, state- of-the-art, high-performance IPs
- ETH research is focused on smart NICs for programmable in-network computing
- ExactLab optimizes HPC codes

QDMA software stack
("Xilinx open source driver" with custom software added)




user space
MPI Application
Bare metal tests (similar to IB verbs)
Apelink MPI BTL (OPENMPI)
Direct device driver I/O tests
User library (LIBQCM)
IOCTL
IOCTL
User space completion (Optimization)
Apelink NI device driver
Driver library (LIBQDMA) with custom sw added
kernel space
BAR + DMA
NI: Custom hardware + qdma

Current sw stack overview



user space

MPI Application

Bare metal tests (similar to IB verbs)

Apelink MPI BTL (OPENMPI)

Direct device driver I/O tests

User library (LIBQCM)

IOCTL

IOCTL

User space completion (Optimization)

Apelink NI device driver

Driver library (LIBQDMA) with custom sw added

kernel space

BAR + DMA

NI: Custom hardware + qdma

Memory	
Off-chip Memory Capacity	64 GB
Off-chip Total Bandwidth	77 GB/s
Internal SRAM Capacity	35 MB
Internal SRAM Total Bandwidth	31 TB/s
Interfaces	
PCI Express	Gen3x16
Network Interfaces	2x QSFP28 (100GbE)
Logic Resources	
Look-up Tables (LUTs)	892,000
Power and Thermal	
Maximum Total Power	225W
Thermal Cooling	Passive

Passive Option

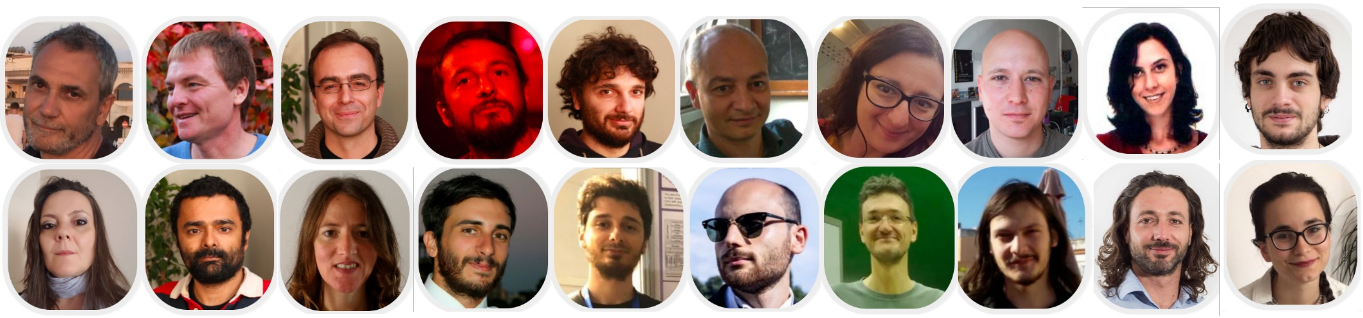
Objective

- O1 (scalability, reliability): >100k nodes; communication libraries (MPI), AI and data-centric applications
- O2 (sustainability, HPC/datacenter convergence): seamless Integration of Internet Protocol, Ethernet, RoCE
- O3 (Throughput & BW): x4 bandwidth and message rate per endpoint; doubling link frequency and Network Interface for each process (multi-rail)
- O4 (congestion, QoS, isolation): new congestion mechanism algorithms, adaptive routing, link scheduler algorithms
- O5 (programmability, latency): configure the network offload engine, compute-in-network, improving data transfer latency and energy efficiency ratio
- O6 (new processor, EPI): ARM + RISC-V interoperability
- O7 (new indicators): new key features for apps, communication/computation overlap and offloading
- O8 (protection): partitioning HPC system into multiple private clouds maintaining protection, security, isolation
- O9 (application and highlight): obtaining better benchmarks scores to demonstrate data-centric performance
- O10 (go to market)

APE LAB

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

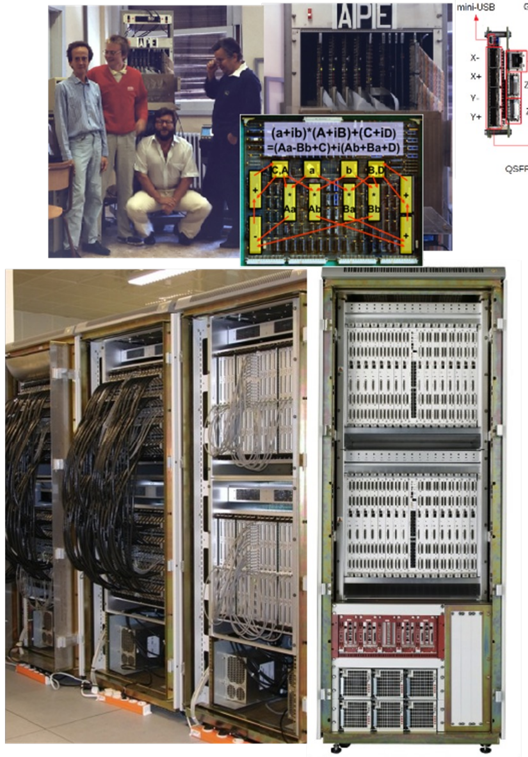
@APELab_INFN



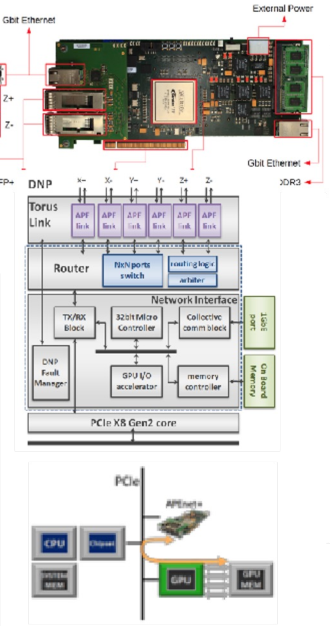
- APE Parallel/Distributed Computing Lab
- 20 members (12 permanent staff + 8 fixed-term)
- 3 main research lines
  - HPC (system architecture, scalable network, application optimization)
  - NeuroScience (Brain Simulation, models, neuromorphic system)
  - HEP Computing (Fast read-out systems, online 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, ….

A bit of history & current R&D activities

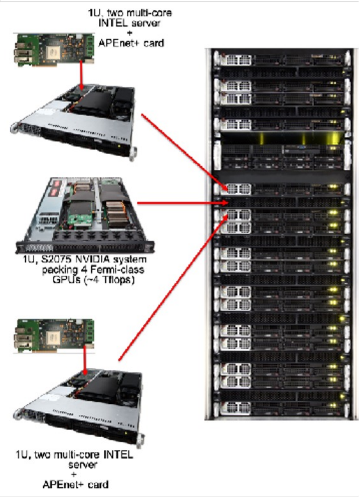
Our Legacy: MPP, Network design, Hybrid Systems



APE Massively Parallel Processor



APENet(+): torus network


1U, two multi-core INTEL server + APEnet+ card
1U, S2075 NVIDIA system packing 4 Fermi-class GPUs (~4 Tflops)
1U, two multi-core INTEL server + APEnet+ card

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

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)

HW testbed: Dibona

Compute Blades:
X4 Up & Running
x1 Under test

Login Server:
mb3-host

Compute Rack

BullSequana X
Live innovation to the fullest
Atos
BullSequana
To meet the new challenges of the 21st century, you need the outstanding computing power of Bull Sequana supercomputers coupled with the high performance data analytics and artificial intelligence delivered by Atos Codex.
Contact an Atos expert at atos.net/bullsequana-x
Atos

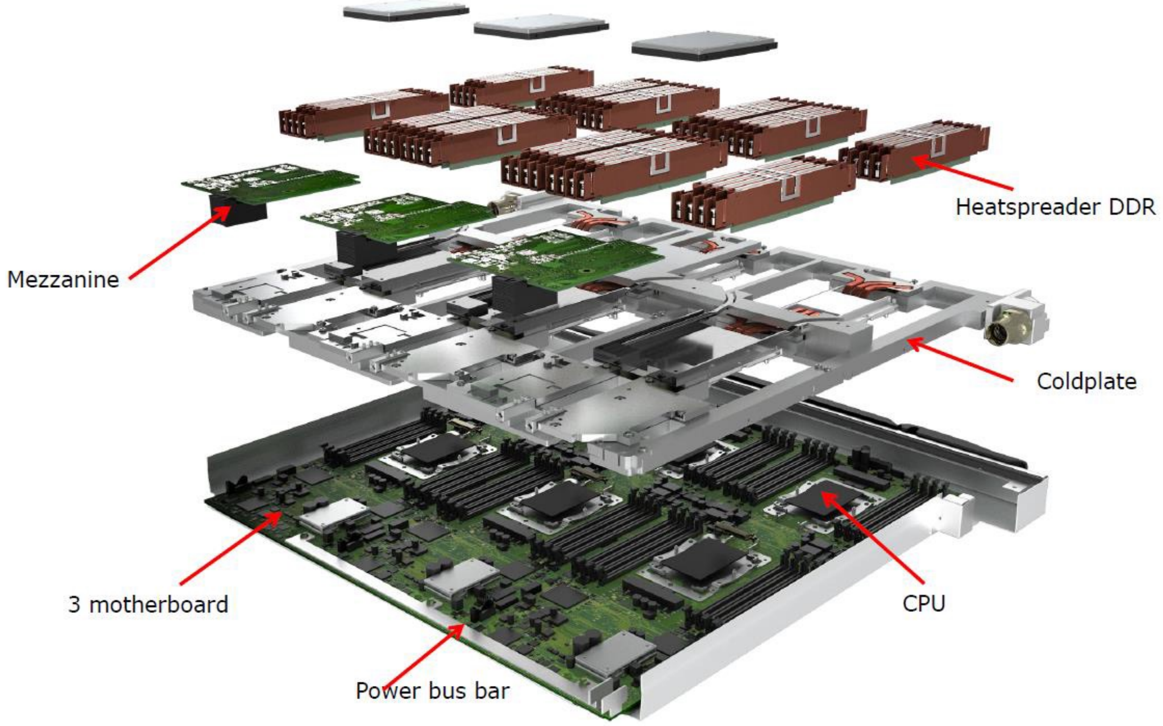
Switch Rack

Management Node

1 BXI Switch:
Connects the 12 up & running nodes

HW testbed Dibona (ATOS)

	Dibona Blade (x4)
Design	1U blade comprising 3 compute nodes side-by-side
Processors	3 Bi-socket Cavium® ThunderX2™ Armv8 processors with 32 cores @2GHz
Architecture	3 motherboard compatible with Cavium reference platform
Memory	3 x16 DDR4 memory slots (max 1024 GB with 64 GB DIMMs)
I/O Slots	BXI1.3 Port mezzanine board
Power Supply	In cabinet
Cooling	Cooling by direct contact with DLC cold plate or through heat spreaders for DIMMS
OS	Red Hat Entreprise Linux 8.4 && Smart Management Center (SMC) & Smart Software Suit (SLURM, OpenMPI)


Heatspreader DDR
Mezzanine
Coldplate
3 motherboard
CPU
Power bus bar