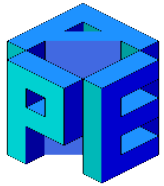


EuroHPC JU Projects for High-Performance Networks for HPC

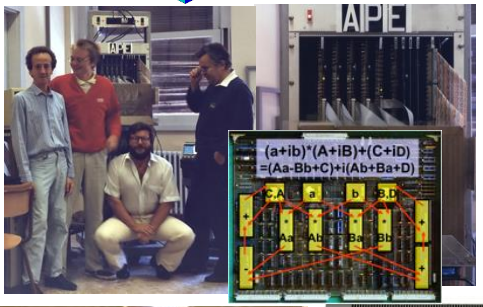
Andrea Biagioni
INFN – Sezione di Roma
APE LAB team

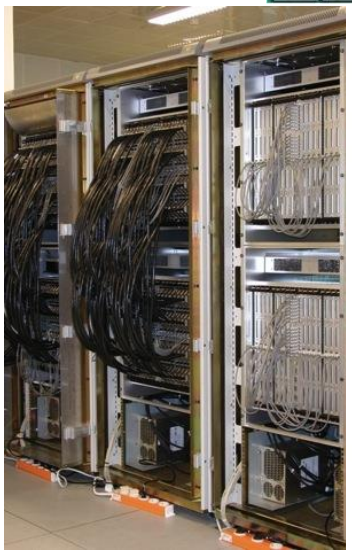
Workshop Computing@CSN5 applications and innovations at INFN
Bari, Italy, 14-16 October 2024

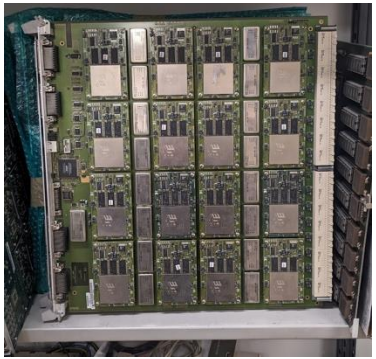
APE LAB TEAM

- 18 members (11 staff + 2 fixed-term + 5 PHD)
- 3 main research lines
  - HPC (system architectures, scalable networks, apps optimization)
  - Neuroscience (brain simulations, models, neuromorphic systems)
  - HEP Computing (Read-out systems, online trigger)
- Know-how
  - ASIC design, FPGA design, GPU programming and integration, network design, dense system integration, parallel programming and application coding (LQCD, neural networks, complex systems), system software, compilers and languages, data analysis, data processing, mathematical physics, theoretical models, statistics…
- National and International research network and industrial collaborations:
  - Grenoble Univ., Athena, FORTH, UPC, CINECA, CNR, Julich, LENS, Manchester Univ, UniMi, CERN, NVidia, EuroTech, E4, IceoTope, IDIBAPS, MonetDB, ATOS, EVIDEN, BULL, UCLM, UPV, ISS

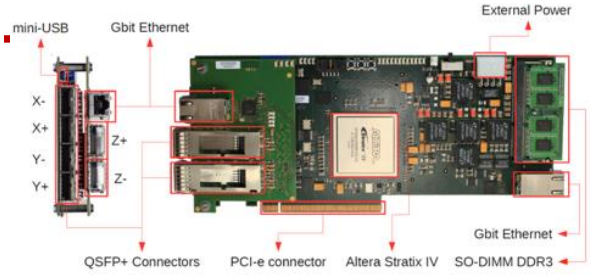
A bit of history…

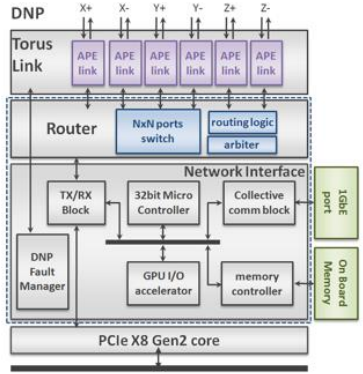




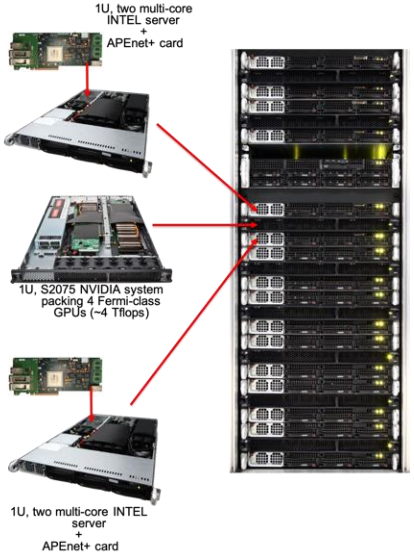


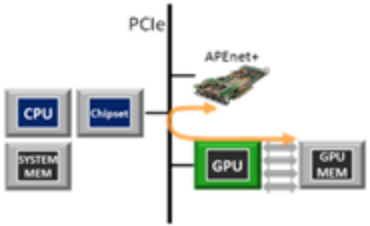
APE Massively Parallel Processor





QUonG: hybrid cluster

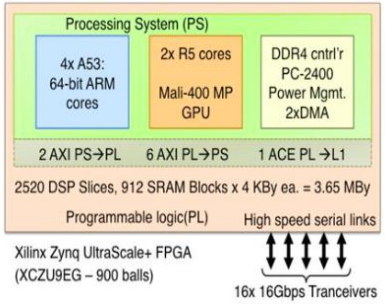




APEnet(+): torus network

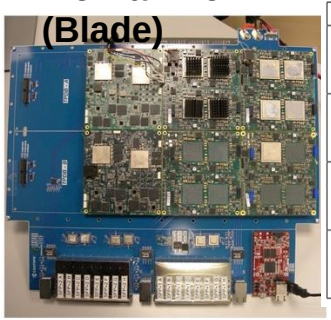
ExaNeSt Overview

Un



QFDB: Node

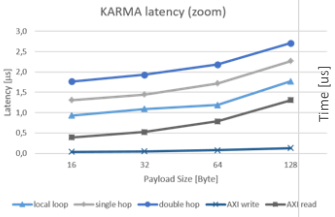
Mezzanine (Blade)

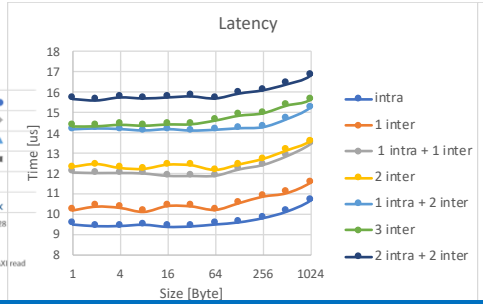


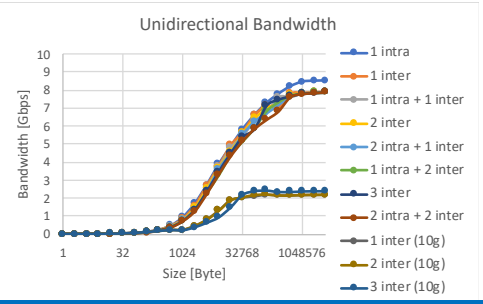
	ExaNeSt
cores per blade	64
memory per blade [GB]	256
FPGAs per blade	16
cores per chassis	576
memory per chassis [GB]	2304
FPGAs per blade	144
core per rack	1728
memory per rack [GB]	6912
FPGAs per blade	432
core per equivalent 1u	~43
memory per equivalent 1u [GB]	~173
FPGAs per equivalent 1u	~11

Rac

	Hierarchy	Fanout	Switching	Topology	Bandwidth	Latency
Tier 4	System	500 Racks	Optical			
Tier 3	Rack	3 chassis	10GbE (ExaNet)	Fat-Tree (Torus)	10 Gbps	
Tier 2	Chassis	9 mezzanines	ExaNet	3D-Torus	4x10 Gbps	400 ns per hop
Tier 1	Mezzanine	4 nodes	ExaNet	Ring	2x10 Gbps	400 ns per hop
Tier 0	Node	4 FPGAs	ExaNet	All-to-All	16 Gbps	400 ns
FPGA	Unit	ZU9				
CORE		A53				





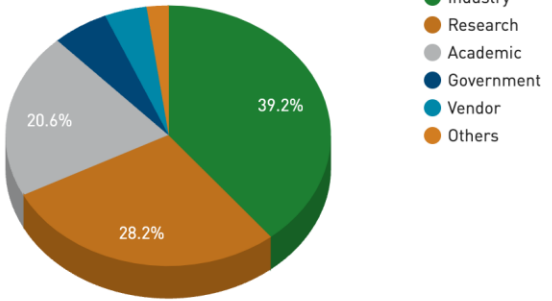


Skills

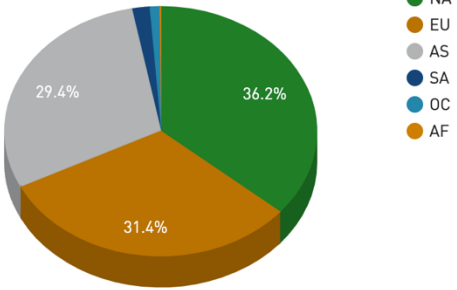
- Hardware (Andrea Biagioni, Carlotta Chiarini, Ottorino Frezza, Francesca Lo Cicero, Piero Vicini)
  - Architecture Design
  - FPGA (Altera & Xilinx) firmware coding (VHDL, Vivado, Quartus, Modelsim)
  - HPC (Routing and Switching), ICT (Data Transmission), GPU (GPUDirect RDMA, NVP2P)
  - HEP (Trigger)
- Software (Alessandro Lonardo, Michele Martinelli, Pierpaolo Perticaroli, Luca Pontisso, Cristian Rossi)
  - Driver, API (RDMA), MPI library
  - GPU in HEP trigger system
  - High Level Synthesis (C, C++, OpenCL)
  - Architectural Simulation and Architecture Design
- Brain studies: physics models, simulations, data analysis (Fabrizio Capuani, Alessandra Cardinale, Giulia De Bonis, Cosimo Lupo,, Pier Stanislao Paolucci, Elena Pastorelli, Francesco Simula, Leonardo Tonielli)
  - Parallel computing
  - Brain Simulation and Modelling (DPSNN, Nest), Brain Inspired Fast Learning
  - Data Analysis of Brain Data and Cortex Dynamics (Matlab, Python)

Why HPC? Why Custom Interconnects? TOP 500

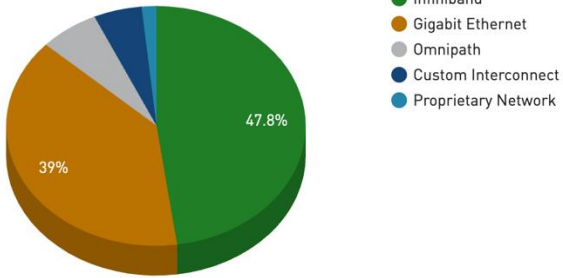
Segments System Share



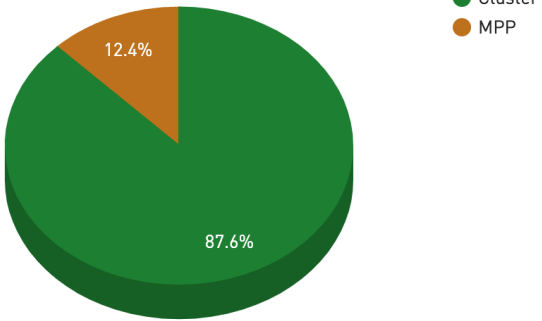
Continents System Share



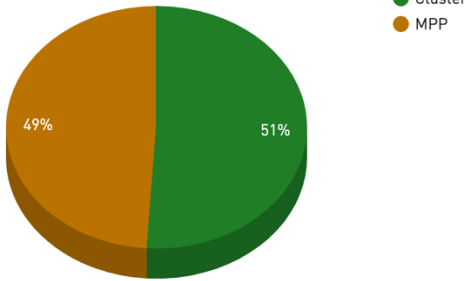
Interconnect Family System Share



Architecture System Share



Architecture Performance Share



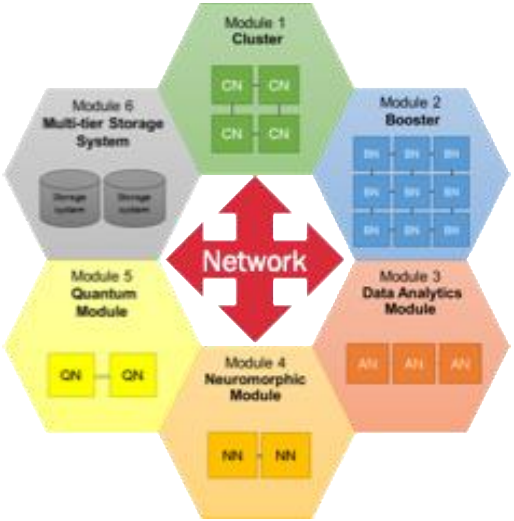
- **Custom or proprietary networks are still necessary.**
- **Application specific solutions and optimizations!**

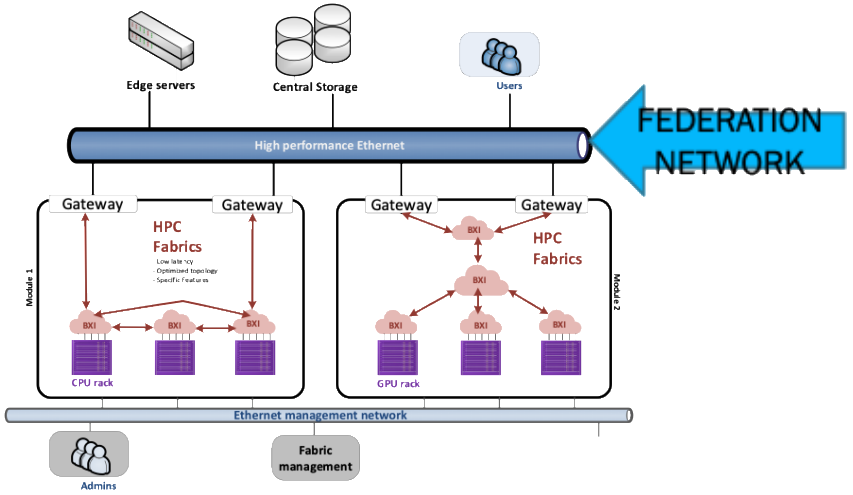
RED-SEA project

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

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



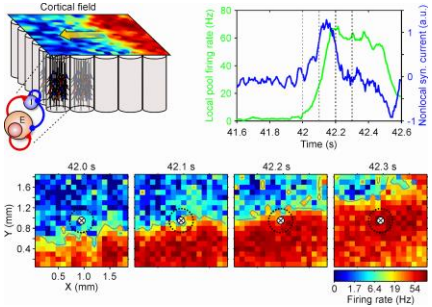


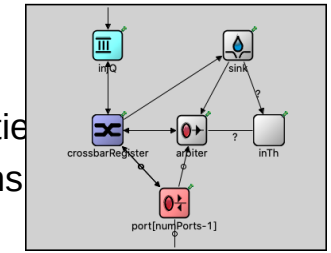
- HPC (High Performance Computing) ; HPDA (High-Performance Data Analytics); AI (Artificial Intelligence)
- Supercomputer: aggregation of resources that are organized to facilitate the mapping of applicative workflows
- HPC is part of the continuum of computing

- 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: methodology for Co-Design Activity

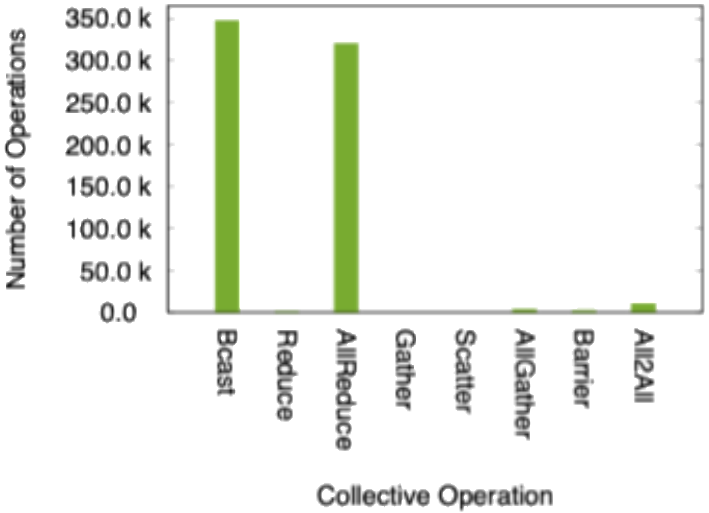
- Application portfolio
  - **NEST: simulator for spiking neural network models**
  - LAMMPS: molecular dynamic engine with focus on material modelling
  - SOM: artificial neural networks used in the context of unsupervised ML
- Benchmark portfolio
  - GSAS: provides a shared memory abstraction model to distributed applications
  - DAW: stress the NI capabilities at scale and the QoS capabilities
  - LinkTest: scalable benchmark for point-to-point communications
  - PCVS: validation engine designed to evaluate the offloading capabilitie
- 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)

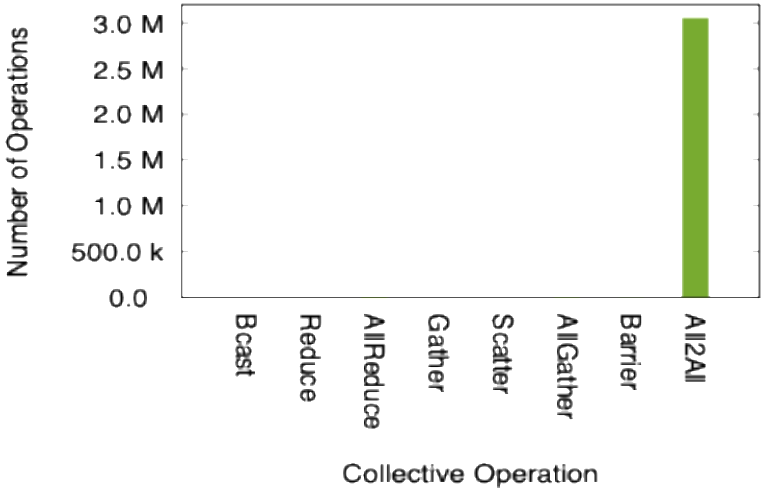




Static analysis

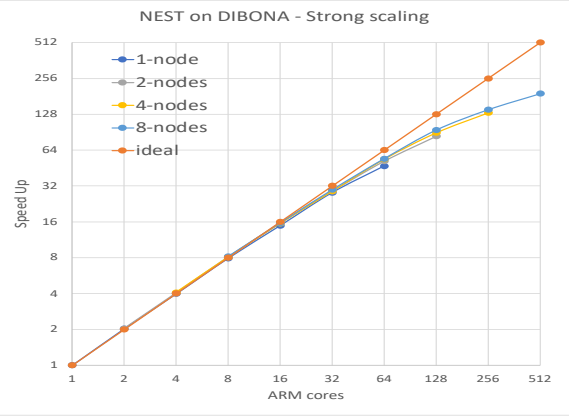
- Congestion characterization
  - Static analysis: different behaviour, LAMMPS wider varity of collectives, NEST dominated by All2All





NEST on DIBONA

- the dynamics undergoing Slow Waves Activity of the cortex of one brain hemisphere of a mouse in a deep sleep state
- The parallelization scheme assigns neurons to different virtual processes (core) in a round-robin fashion
  - MPI processes may spread across the nodes of the system
  - OpenMP threads depends on the CPU architecture of the node
  - Total Core = MPI process X OpenMP threads
- Virtual processes belonging to different MPI processes keep their status in sync exchanging data with each other by calling MPI_Alltoall at the end of every step of integration



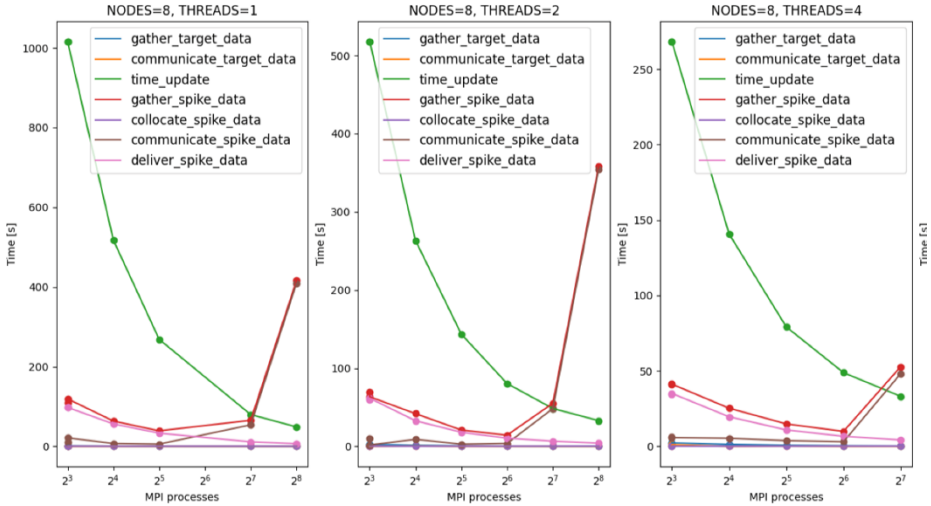
MPI process	OpenMP Threads	Number of messages	total bytes	average size [B]	Range of Interest	Messages in range [%]	Average Size [B]	Average Size Ratio
8	64	6,82E+04	8,02E+08	11760	512B-16kB	95,9	4854	
16	32	2,90E+05	1,24E+09	4259	256B-8kB	97,1	2558	0,53
32	16	1,19E+06	2,05E+09	1716	128B-4kB	97,8	1300	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.**

Co-Design: NEST requirements

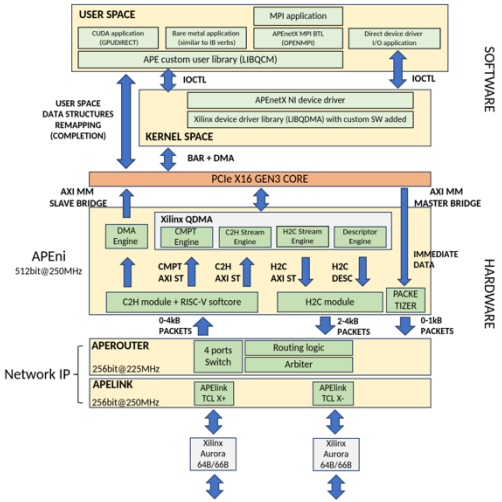
Name of timer	Exaplanation	Part of
time_gather_target_data	Cumulative time for communicating connection information from postsynaptic to presynaptic side	time_communicate_prepare
time_communicate_target_data	Cumulative time for core MPI communication when gathering target data	time_gather_target_data
time_update	Time for neuron update	time_simulate
time_gather_spike_data	Time for complete spike exchange after update phase	
time_collocate_spike_data	Time to collocate MPI send buffer from spike register	time_gather_spike_data
time_communicate_spike_data	Time for communicating spikes between compute nodes	
time_deliver_spike_data	Time to deliver events from the MPI receive buffers to their local synaptic targets (including synaptic update, e.g., STDP synapses) and to the spike ring buffers of the corresponding postsynaptic neurons	

time gather spike data (time for complete spike exchange) decays up to 64 MPI processes and increases with the number of processes, disregarding the number of threads and nodes, becoming a bottleneck in the speedup of the system as supposed during the analysis based on the VEF-traces results.



APEnetX Network Interface Card

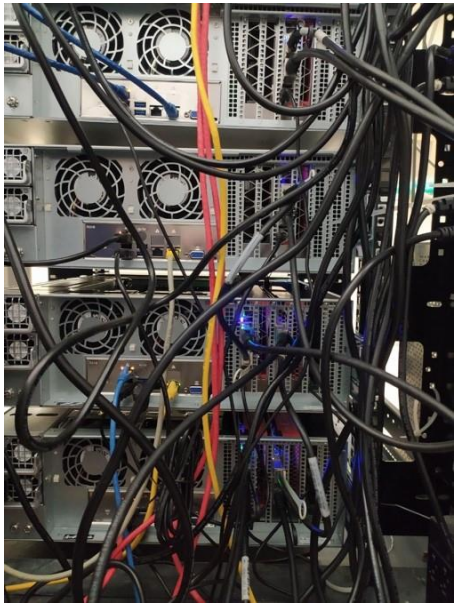
- INFN duties
  - Network Interface Card (APEnetX)
    - PCIe gen4 (GPU+CPU) + BXI link (Xilinx Alveo FPGA)
  - Co-Design through applications (NEST)
- TARGET
  - Developing network IPs to optimize spiking neural network communication



- Xilinx Alveo Board DMA engine
- Matching requirement for the communication generated by NEST
- Providing proprietary software driver and low-level communication library
- NVIDIA GPUDirect RDMA
- Custom OpenMPI BTL
- **Bandwidth per channel 57.6 Gbps**
- **Latency 1.9us**
- Validated through HPC-benchmark
- Large-scale simulation environment (NEST traces)
- Interoperability with the BXI interconnect
- Proprietary priority management mechanism to improve QoS of the data transmission system

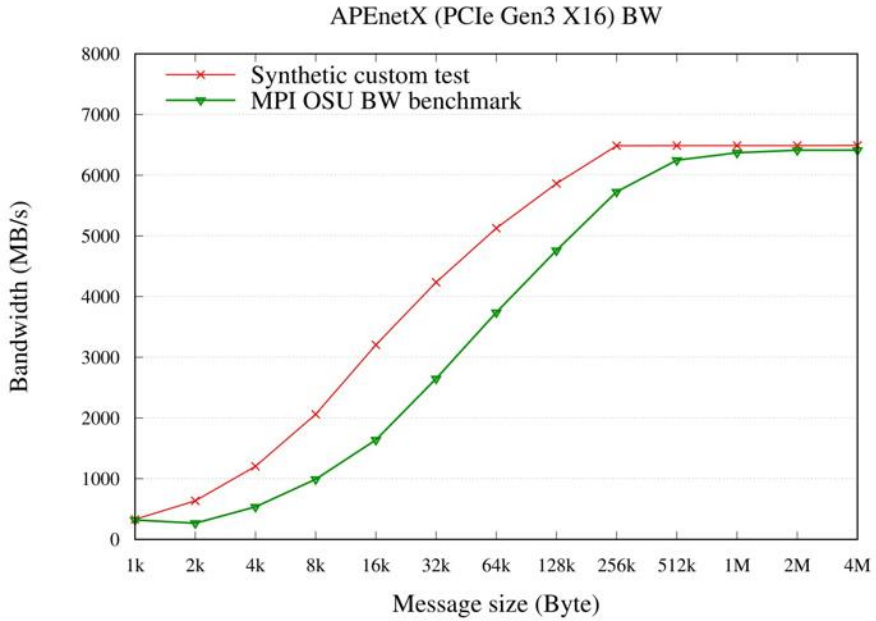
Hardware testbed

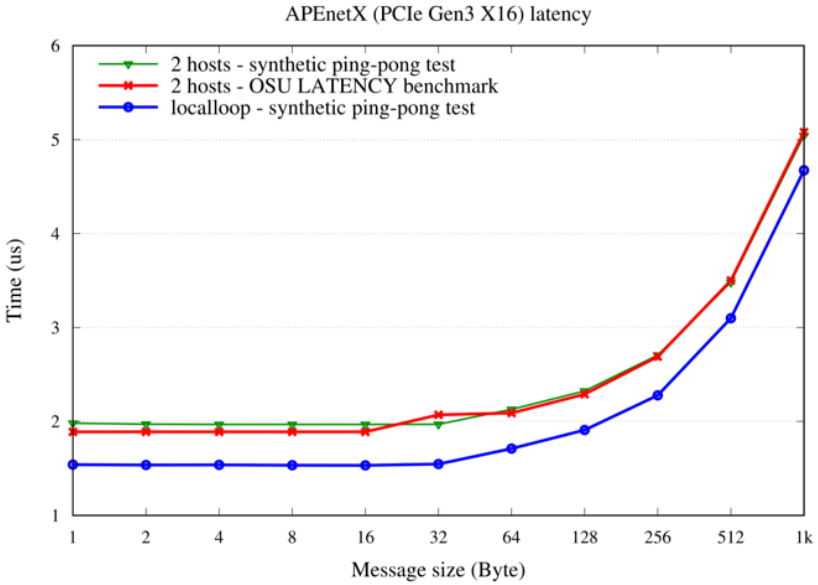
- 2x Supermicro SuperWorkstation 7049GP-TRT server (MidTerm)
  - 2 x 8-cores 4200-series 14nm Intel Xeon Scalable Silver Processors (Cascade Lake) running @ 2.10GHz (PCIe gen3 support)
  - Memory: 192GB DDR4 @3.2GHz
- 2x SuperMicro A+ Server 3014TS-I
  - 1x 16-cores AMD EPYC 7313P @3÷3.7GHz (PCIe gen4 support)
  - Memory: 128GB DDR4 @3.2GHz
- 2x SuperMicro X13SEI-F, single socket
  - 1 x 10-cores Intel Xeon Silver 4410T running @ 4.00GHz (PCIe gen5)
  - Memory: 128GB DDR5 @2.4GHz
- OS: GNU/Linux Centos 8.8 with Linux 4.18.0 kernel.
- **APEnetX prototype**: 4x Xilinx Alveo U200 built on the Xilinx 16nm UltraScale architecture,



Latency & bandwidth test

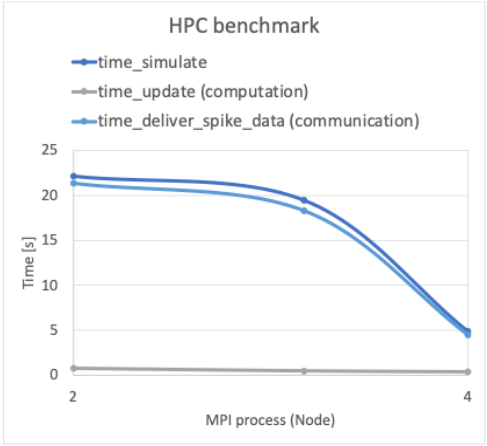
- Latency and bandwidth results for both our synthetic tests and for the standard latency/bandwidth tests in the OSU Micro Benchmarks set.

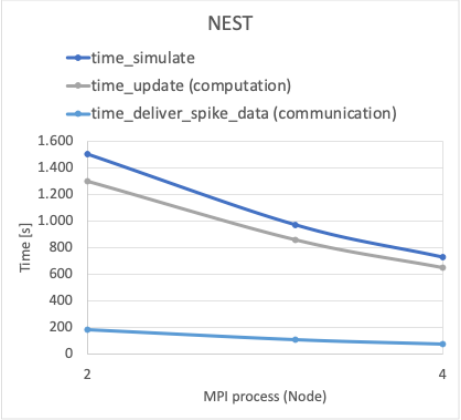




NEST on APEnetX development platform

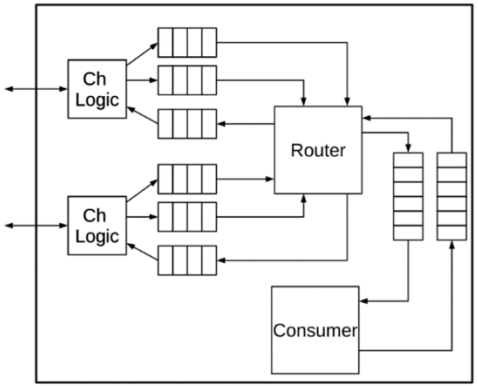
- **Slow Waves Activity of the cortex of one brain hemisphere of a mouse (NEST)**
  - Network size: 61596 neurons (30798 excitatory + 30798 inhibitory)
  - Simulated time: 4000ms
- **HPC benchmark (NEST)**
  - Network size: 11250 neurons (9000 excitatory + 2250 inhibitory)
  - Simulated time: 250ms
- **Limitations:**
  - Fast Send Issue (no hardware optimization for small messages <1kB)
  - 1 MPI process per node (multi-process under debug)





The APEnetX simulator

- **Porting to OMNet++5 and then to OMNeT++ 6.0**
- Network Node
  - **Buffer**: FIFO functionality
  - **Channel**: APElink functionality (link control)
  - **Producer**: packet injection to the TX FIFOs
  - **Consumer:** message from the RX queue
  - **Router** → APErouter (VCT + DOR)



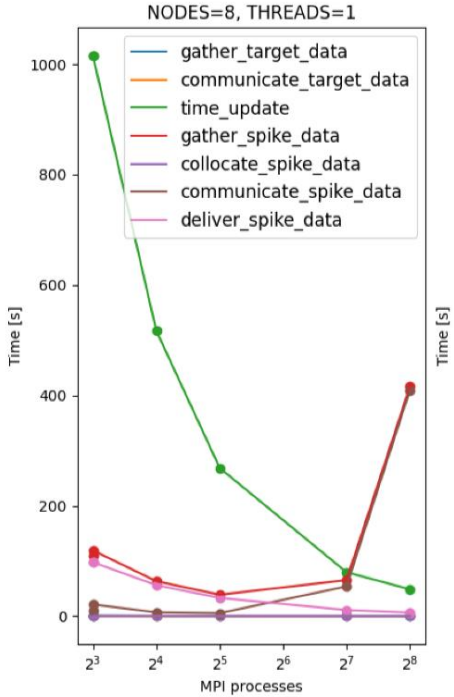
Simulator platform improvement

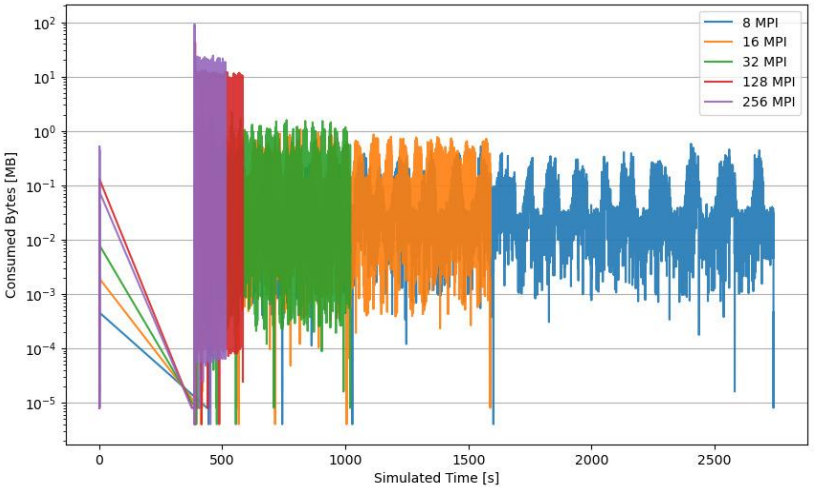
- VEF-traceLIB integration to inject and the network traces and analyze the proposed model
- Tuning the simulator with real APEnetX hw timing
  - PCIe Gen3 X16 → Producer & Consumer
    - FAST SEND: Transmission of small packets (<2048B)
    - DMA transaction for big packets
  - Transceivers and Link Layer (250MHz) → Channel
    - 62 Clock cycle latency
  - Switch (225 MHz clock) → Router
    - 10 clock cycles latency; Completely pipelined
- Virtual priority queue model as an extension of the Buffer Module (to support T3.5 activity)
- Multiple std::queues — one for each priority rank — are instantiated within the same Buffer

NEST on APEnetX simulator

- simulated time of the NEST traces injected in the APEnetX simulator (DQN_SIM).
- Communication seems not an issue in the simulator

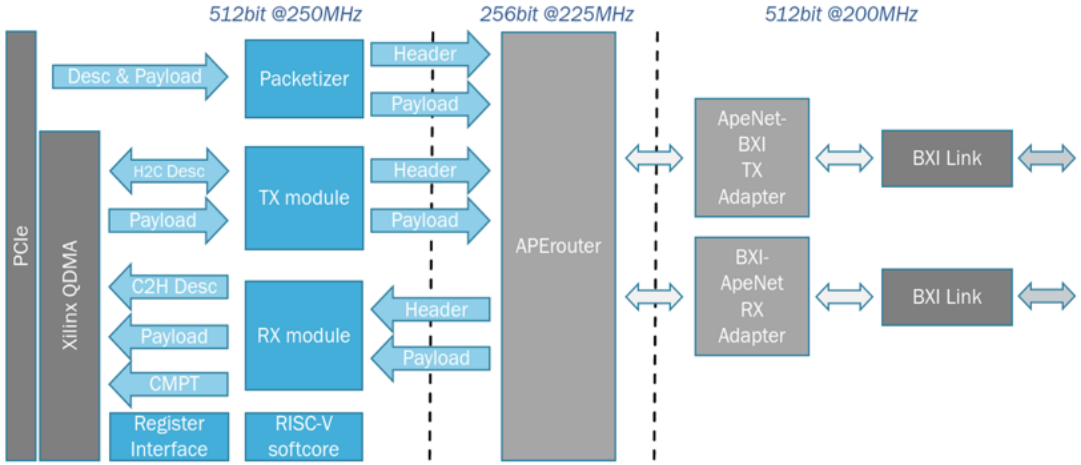
Configuration difference:
SIM: 256 nodes 1P/N
HW: 8 nodes 64P/N

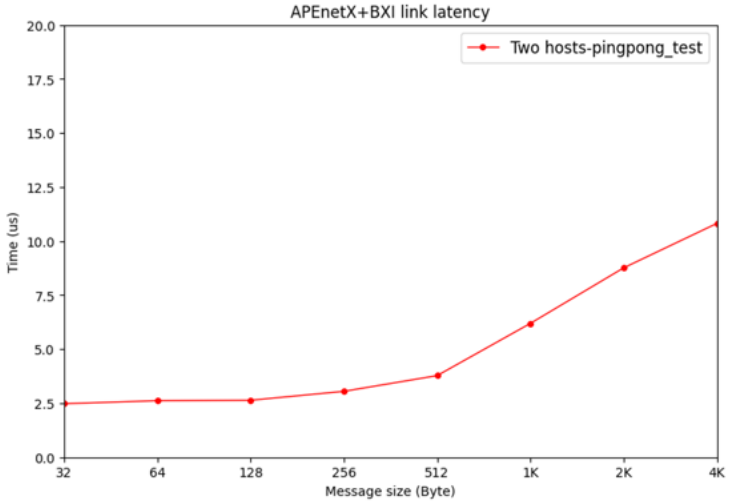




Integration of APEnetX with BXIv2

- porting of the BXIv2 link layer firmware onto APEnetX prototype (Alveo U200)
- an adapter between the BXIv2 and APEnetX communication protocol was designed and implemented.





NET4EXA

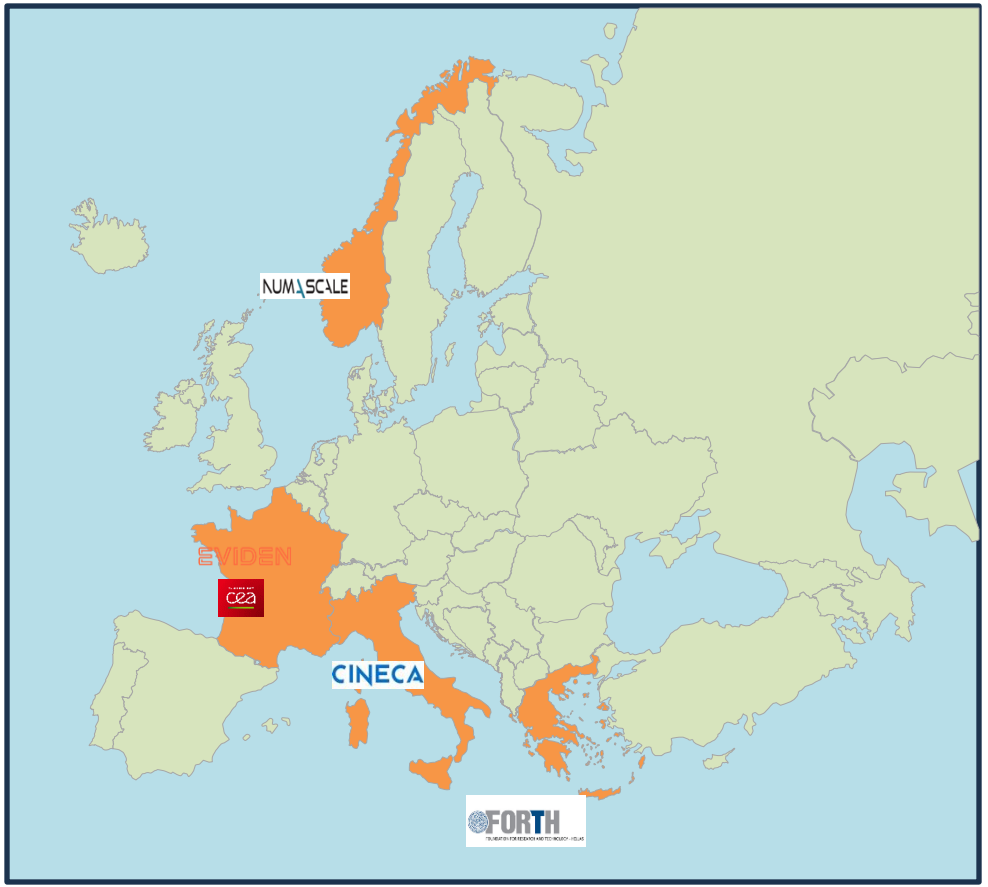
- Acronym: NET4EXA
- Title: NeTwork for EXAscale systems
- EuroHPC Call: HORIZON-EUROHPC-JU-2023-INTER-02
  - Type of action: HORIZON-JU-IA HORIZON JU Innovation Actions (w/ TRL 8)
  - Proposal number: 101175702
- Total costs : 71 126 351 €;
  - EU funding: 26 916 520,70 €;
  - + countries' funding
  - + in-kind contribution for industrial beneficiaries.
- Project Start date: Sep. 1st, 2024; Duration: 30 months
- 5 beneficiaries (see detail in the next slide)

- NET4EXA aims at
  - Developing & demonstrating BXIv3, a new generation of fast interconnect for HPC and AI systems.

NET4EXA Consortium

- 5 Beneficiaries;
- 3 CINECA Affiliate Entities;
- 3 subcontractors
- 4 European countries

TYPE	NAME	Country
Large company	1 - BULL	FR
SMEs	2 - NUMASCALE AS	NO
	4 – Subco SCINTIL	FR
	4 – Subco Spearl	FR
	5 – Subco	IT
Large Datacenters & Research centers	4 - CEA	FR
	5 – CINECA	IT
Academic partners	3 – FORTH	GR
	5.1 CINECA – UNITRENTO	IT
	5.2 CINECA – UNIROMA1	IT
	5.3 CINECA - INFN	IT



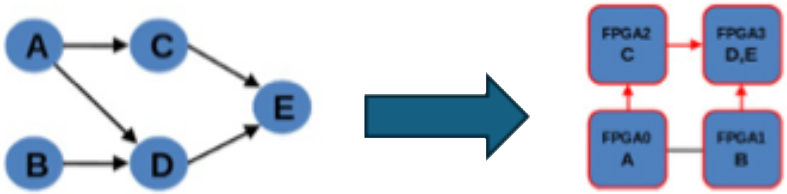
TEXTAROSSA project

Project start: 01/04/2021
Project duration: 36 months
Project budget: 6 M€

- The project has developed **core technologies** for the emerging generations of high-end heterogeneous computing architectures towards exascale-class systems, extensively applying the **top-down co-design process** from the applications to the prototypes of HW architectures and SW systems.
- **The ambition was leveraging and further extending** the expertise of the core partners involved in the European Processor Initiative (EPI) to realize the EuroHPC roadmap for energy-efficiency, high-performance and secure services by enabling new computation paradigms for HPC, AI and HPDA applications.

APEIRON

- Enables the mapping the dataflow graph of the application on the distributed FPGA system and offering runtime support for the execution.
- Allows users, with no (or little) experience in hardware design tools, to develop their applications on such distributed FPGA-based platforms:
  - Tasks are implemented in C++ using High Level Synthesis tools (Xilinx Vitis®).
  - Lightweight C++ communication API: Non-blocking send() / Blocking receive().



Host Interface IP: interface the FPGA logic with the host through the system bus

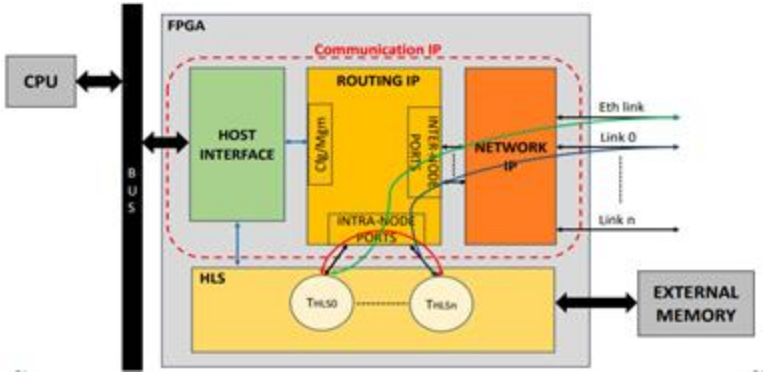
- Xilinx® XDMA PCIe Gen3

Routing IP: routing of intra-node and inter-node messages between the processing tasks on FPGA

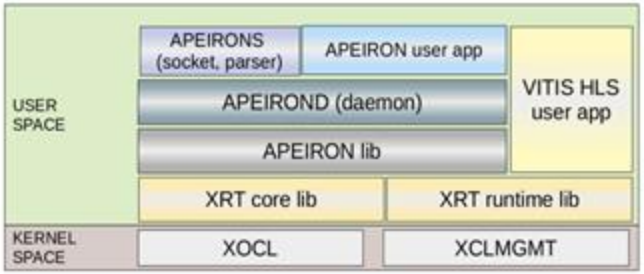
Network IP: network channels and Application-dependent I/O

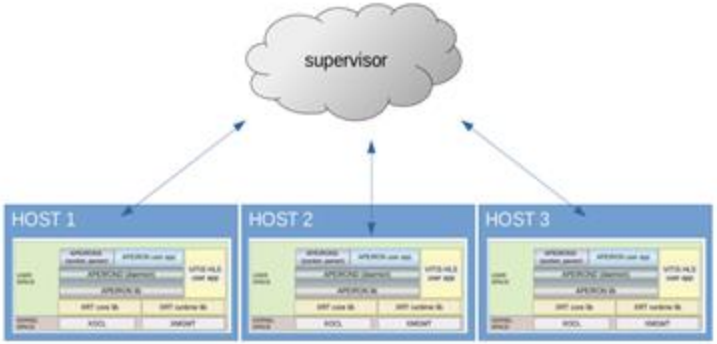
- Custom APElink 20/40 Gbps
- UDP/IP over 1/10/25/40 GbE

HLS Kernels: user defined processing tasks



APEIRON SOFTWARE STACK

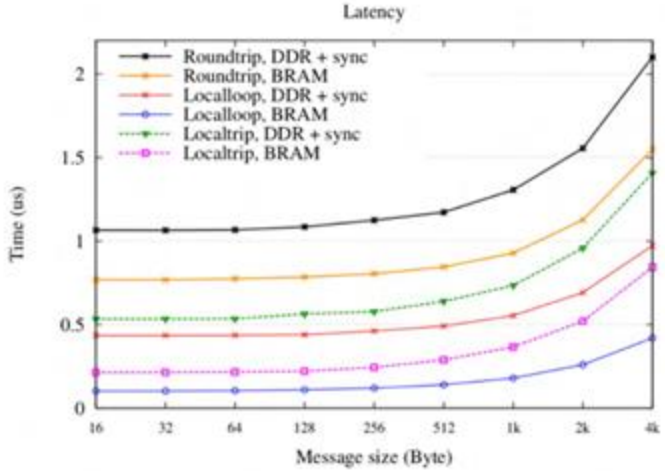


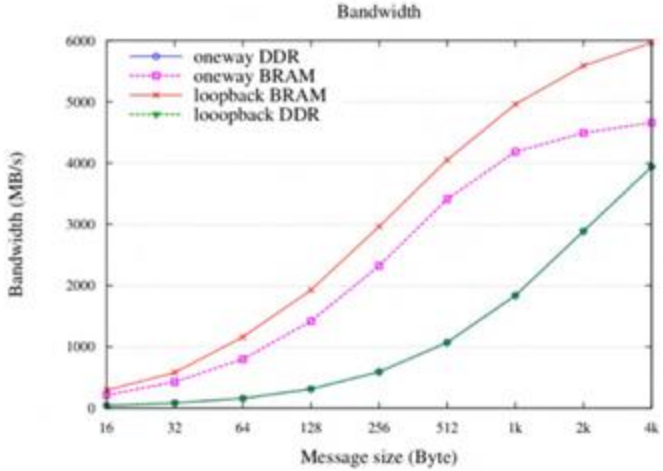


- The APEIRON runtime software stack is built on top of the Xilinx® XRT one adding three layers to:
  - add the functionalities required to manage multiple FPGA execution platforms (e.g., program the devices, configure the IPs, start/stop execution, monitor the status of IPs, ...);
  - reduce the impact of changes in XRT API introduced with any new version of Vitis on the APEIRON host-side applications;
  - decouple the APEIRON SW stack from the specific platform, easing the future porting of the framework to different platforms/vendors.
- **APEIRONlib** is a wrapper of XRT functions, allowing host user code to acquire a handler over the device, output redir.
- **APEIROND** is a persistent daemon managing multiple access request from user apps to the board using the APEIRON lib exposed functions
- **APEIRONS** module receiving and interpreting commands from the supervisor application through the network and forwarding results and applications output to it.
- **SUPERVISOR** is the software component used to manage

APEIRON PERFORMANCE

Communication IP: 256 bit datapath @200MHz



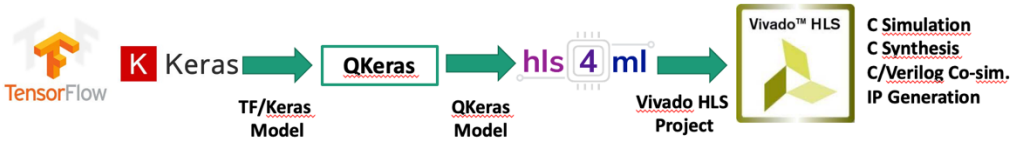


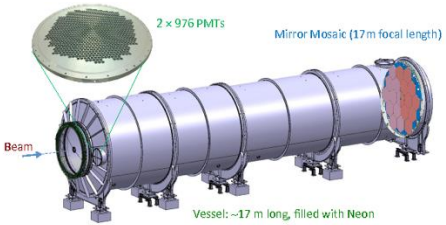
Latency	DDR (ns)	BRAM (ns)
Intranode (localtrip)	533	213
Internode (roundtrip)	1065	768

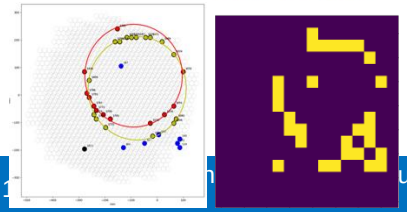
Bandwidth	DDR (MB/s)	BRAM (MB/s)
Intranode (localtrip)	3938	5967
Internode (roundtrip)	3938	4658

RAIDER: Real-time AI-based Data analytics on hetERogeneneous distributed svstem

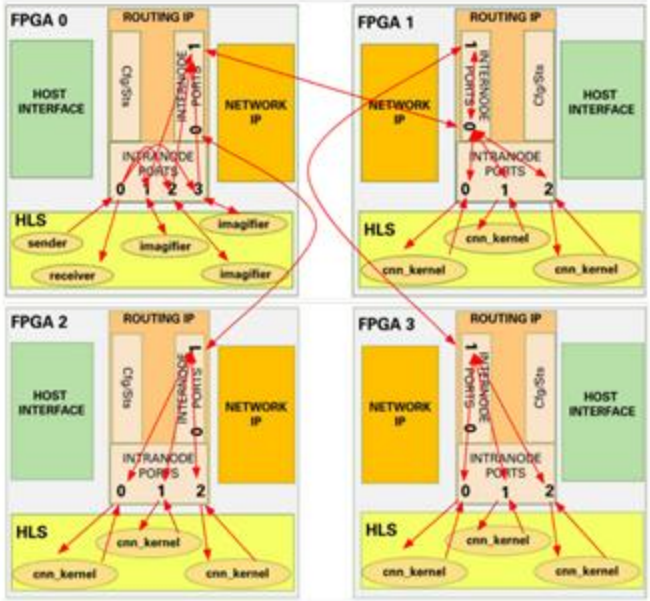
- CERN NA62 Experiment: study of the very rare decay K->pinunu.
- Need to produce physics events at high rate (10 MHz) to collect statistics.
- Problem: count the number of charged particles in each physics events using the RICH detector to improve the online selection (trigger) performance.
-





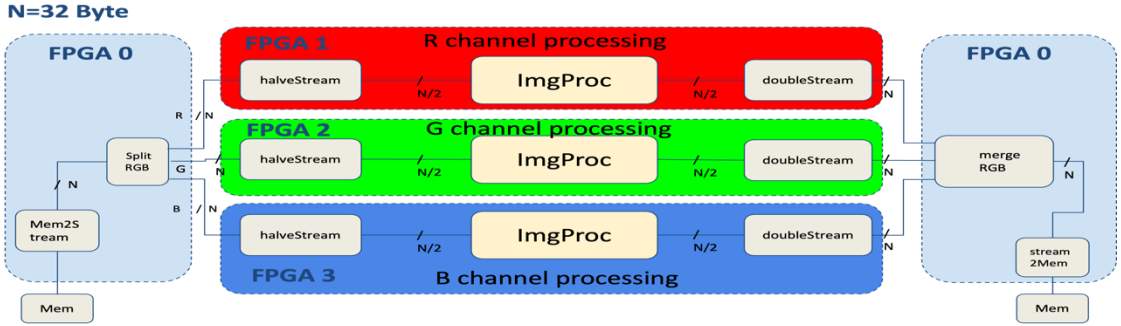


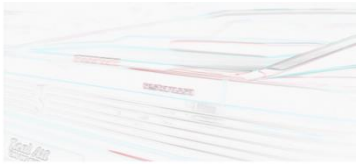
KPI	CNN CPU tensorflow	CNN CPU+GPU tensorflow
time to solution [s]	158.521	125.963
throughput [events/s]	**189250**	**238165**
energy to solution [J]	11091.919	17497.783 (8724.648 GPU)
energy efficiency [events/J]	**270.467**	**154.305**



KPI	RAIDER @200 MHZ [4 FPGA, 9CNNs]
time to solution [s]	0.554
throughput [events/s]	**4873646.209** x20
energy to solution [J]	165.277 (101.055 FPGA)
energy efficiency [events/J]	**16336.183 (26718.126 FPGA)** x100

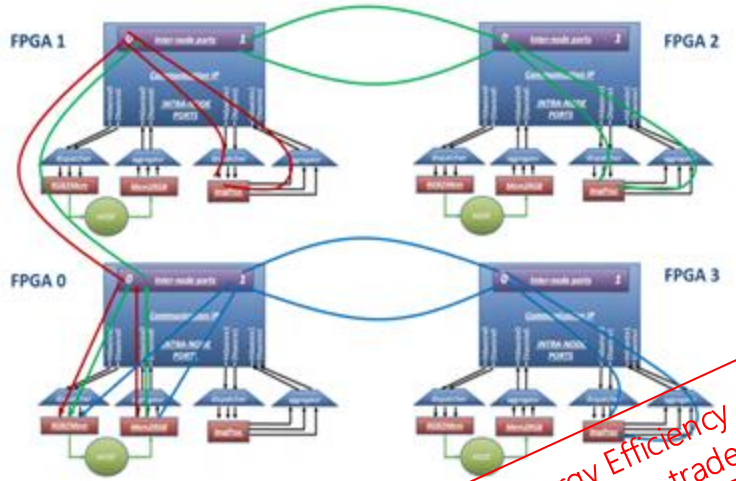
FIPLib-MultiFPGA

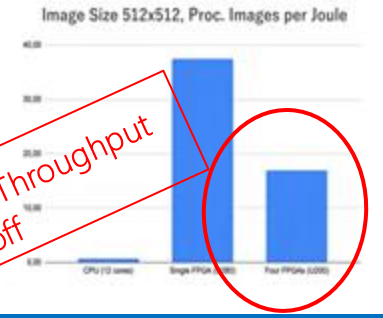




FPGA Image Processing Library ➔ multi-FPGA implementation with APEIRON

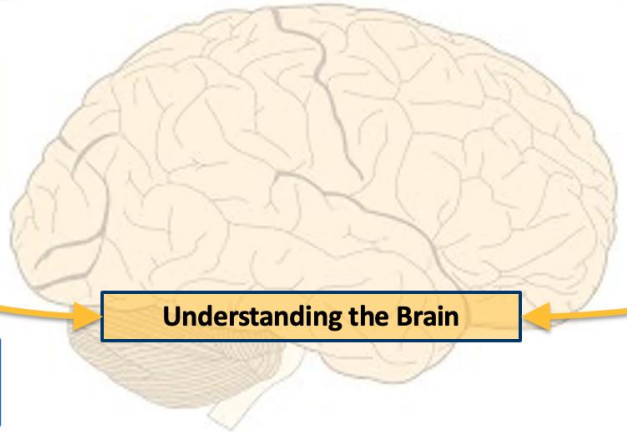
- Developed by ENEA in C++, it employs the Vitis HLS flow to construct the library's kernels for the execution of image processing algorithms.
- On a multi-FPGA setup, we were able to split the overall image processing by implementing a single RGB kernel on each node: one I/O node (FPGA 0) and three (R, G, B channels) processing FPGAs (1, 2, 3).





Energy Efficiency - Throughput trade-off

BRAINSTAIN



WP1
Analysis pipelines
data processing, **data analysis**

WP2
Brain models
&
Simulations

Task 1
Algorithms & Methods

Task 2
Data Analysis

Task 3
Clinical Applications

Understanding the Brain

Hardware/Software co-design
of Computing Architectures

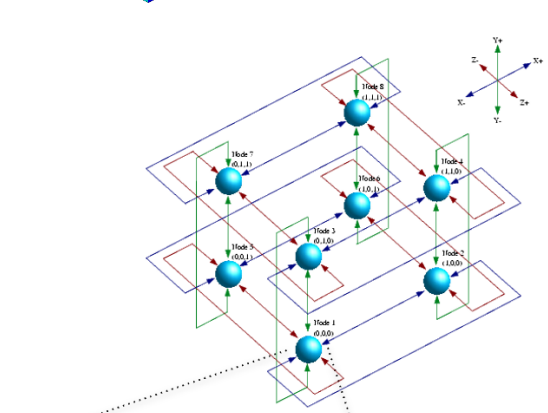
WP3
Design of Brain-inspired
Computing Architectures for AI

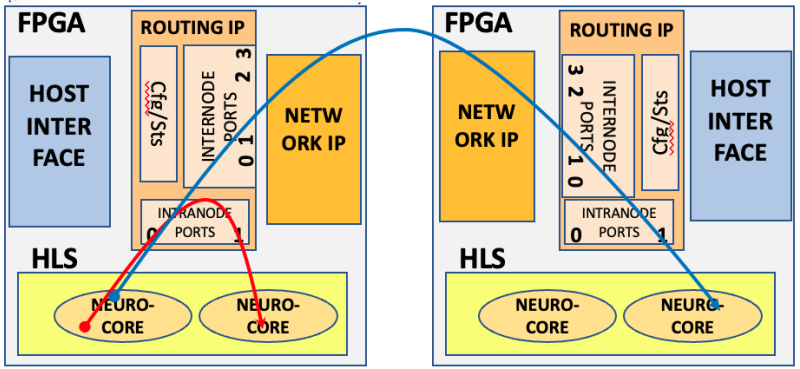
Task 1
From brain models to architecture

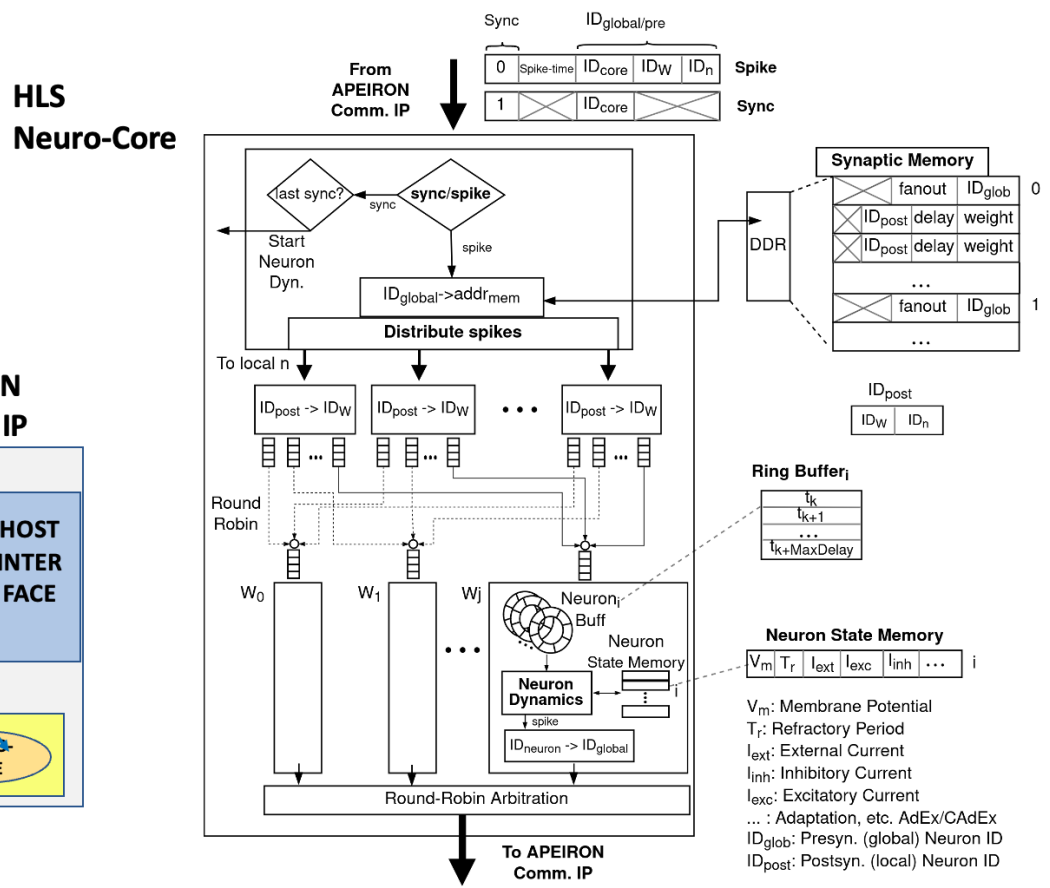
Task 2
Architectures prototyping

Task 3
Validation and Perf. Estimation

Design of Brain-inspired Computing Architecture for AI Tasks







Conclusion

- Custom interconnects are still necessary
- INFN involved in several EuroHPC projects to design high performance networks
  - Net4exa
- CSN5 activities
  - BRAINSTAIN

APE LAB Team

THANK YOU

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

https://twitter.com/APELab_INFN

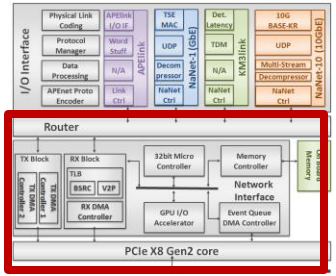
From HPC to HEP: NaNet Project

NaNet: Design and implementation of a family of FPGA-based PCIe Network Interface Cards :

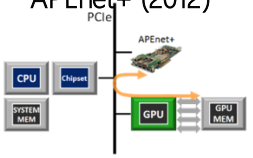
- Bridging the front-end electronics and the software trigger computing nodes.
- Supporting multiple link technologies and network protocols.
- Enabling a low and stable communication latency.
- Having a high bandwidth.
- Processing data streams from detectors on the fly (data compression/decompression and re-formatting, coalescing of event fragments, …).
- Optimi NA62 KM3NeT NA62 with GPU accelerators.

	NaNet-1	NaNet3	NaNet-10	NaNet-40
Year	Q3 - 2013	Q1 - 2015	Q2 - 2016	Q3 - 2019
Device Family	Altera Stratix IV	Altera Stratix V	Altera Stratix V	Altera Stratix V
Channel Technology	1 GbE	KM3link	10 GbE	40 GbE
Transmission Protocol	UDP	TDM	UDP	UDP
Number of Channel	1	4	4*	2
PCIe	Gen2 x8	Gen2 x8	Gen3 x8**	Gen3 x8
SoC	NO	NO	NO	NO
High Level Synthesis	NO	NO	NO	YES
nVIDIA GPUDirect RDMA	YES	YES	YES	YES
Real-time Processing	Decomp.	Decomp.	Decomp. Merger	?

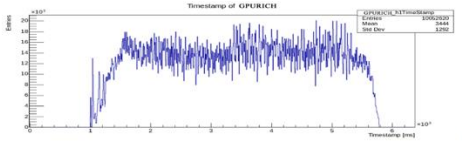
NaNet architecture



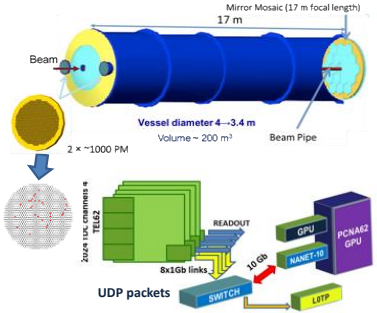
FP7-ICT
EURETILE
APEnet+ (2012)



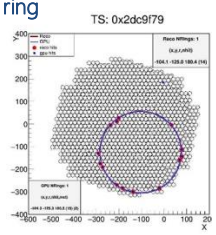
GPU-RICH generated primitives (late Oct 2018)



GPU-RICH overview



GPU 1 ring == Reco 1 ring



GPU 1 rings
Reco 2 rings