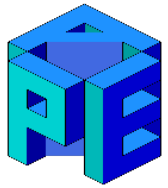
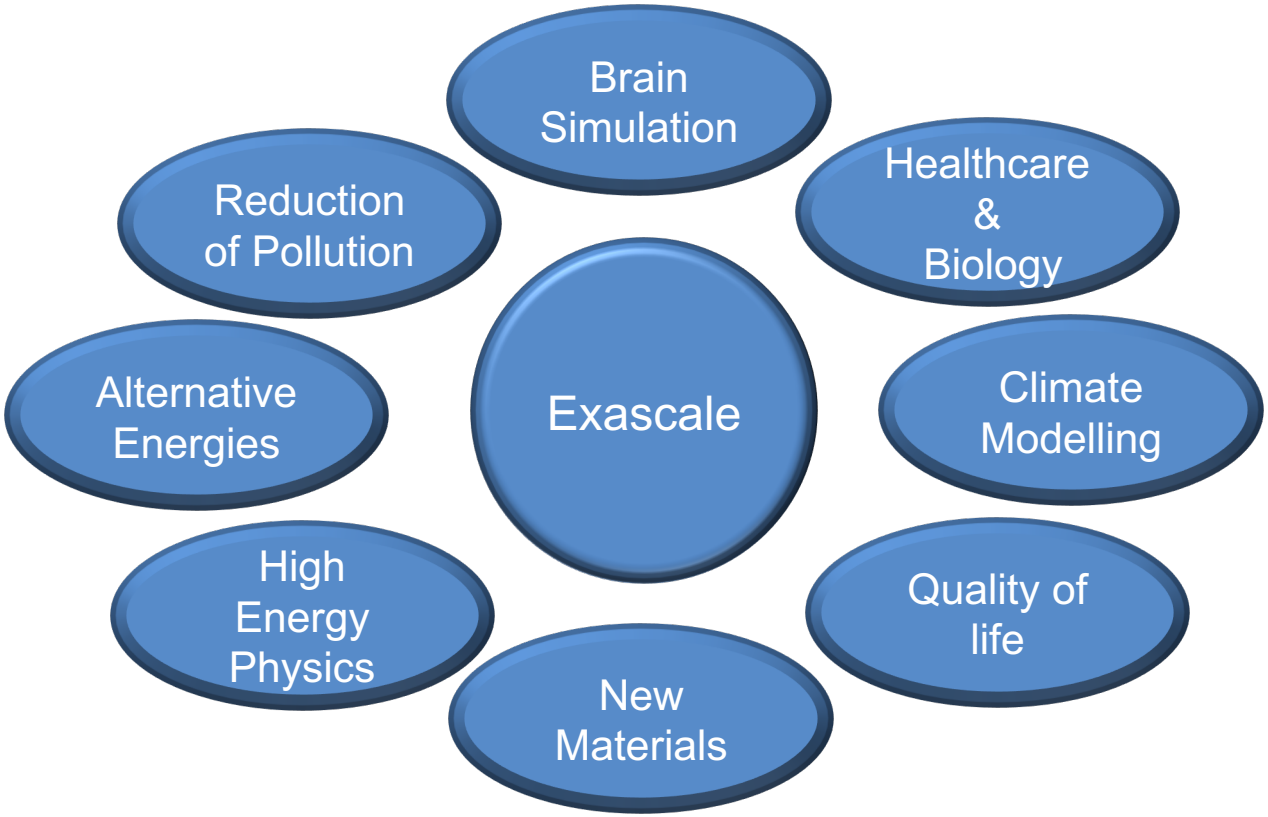


Supercalcolo

Andrea Biagioni
INFN – Sezione di Roma
APE LAB team

Retreat Fisica Particelle Elementari
Assisi, Italy, 16-18 June 2019

The Exascale value



P. Messina, "The exascale computing project," *Computing in Science Engineering*, vol. 19, no. 3, pp. 63–67, May 2017.

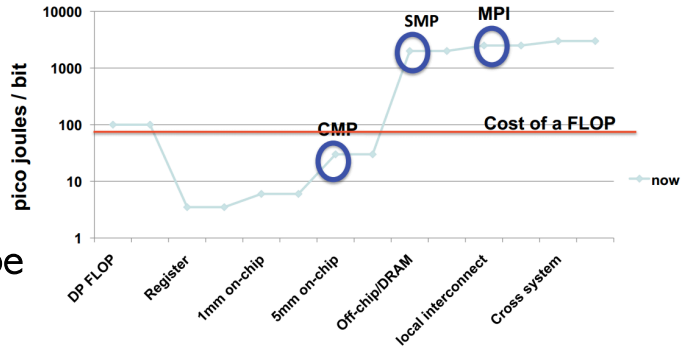
Challenges for Exascale System

Advanced scientific computing research," accessed: 29/Sep/2017. [Online]. Available: https://science.energy.gov/ascr/

- System power
  - Scaling from today's requirement for a petaflop computer, the exaflop computer in 2020 would require 200MW. The target is 20-40 MW.
- Clock frequencies
  - They decrease to conserve power; the number of processing unit per chip increases.
- Memory Bandwidth and capacity
  - Not enough memory per processor. Scaling approaches useless.
- Cost of data movement
  - Both in energy consumed and performance is not expected to improve as much as that of floating point (data movement minimization).
- The I/O system (chip-to mem, mem to I/O, I/O to disk) will be much harder to manage
- Reliability and resiliency
- Hardware/Software CoDesign

J. Shalf, S. Dosanjh, and J. Morrison, *Exascale Computing Technology Challenges*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 1–25. [Online].

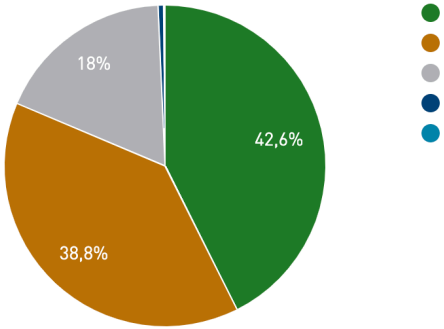
The Cost of Data Movement



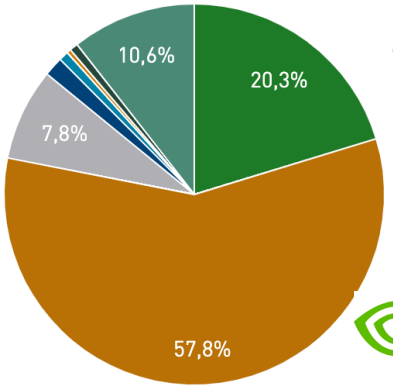
TOP500 list overview (Nov18)

- Europe runs after Asia and America
- 10-best HPC system
  - USA (5): Summit (200 PFlops); Sierra (125 PFlops); Trinity (41 Pflops); Titan (27 Pflops); Sequoia (20 Pflops)
  - CHINA (2): Sunway (125 Pflops); Tianhe-2A (100 PFlops)
  - JAPAN (1): AI Bridging Cloud Infrastructure (32 Pflops)
  - EUROPE (2): Piz Daint (27 Pflops); SuperMUC-NG (26 Pflops)

Continents Performance Share



Accelerator/CP Family Performance Share



86%

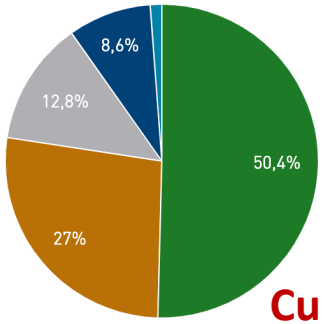
NVIDIA

Processor Generation	Count	System Share (%)
Intel Xeon E5 (Broadwell)	233	46,6
Xeon Gold	99	19,8
Intel Xeon E5 (Haswell)	71	14,2
Xeon Platinum	22	4,4
Intel Xeon Phi	18	3,6

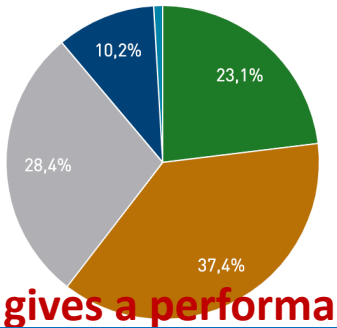
88%

intel

Interconnect Family System Share



Interconnect Family Performance Share



25G Ethernet 50G Consortium

INFINIBAND TRADE ASSOCIATION

Custom Interconnect gives a performance boost!!

What next in Exascale HPC?

- FPGA
  - Different fields of applications thanks to combination of software-like flexibility and hardware performance
  - Xilinx Virtex Ultrascale+
    - TSMC FinFet 16nm → 60% less than old generation power consumption
    - Industrial standard: PCIe gen3, DDR4, Ethernet
    - 21 TFLOPS of DSP single precision
    - Multiple ARM cores (4/8)
- ARM processors
  - Is (was?) the only European CPUs maker
  - Innovative business model: sell Intellectual Properties instead of chip
  - Architecture specialised for embedded/mobile processors delivered
  - ARM in HPC: AMCC X-gene 3, CAVIUM ThunderX
  - EU funded projects: Mont-blanc (BSC); UniServer

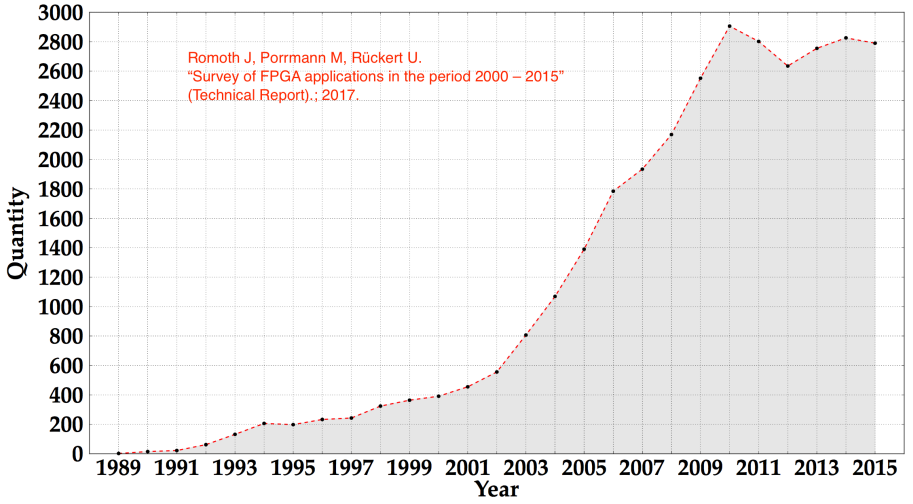
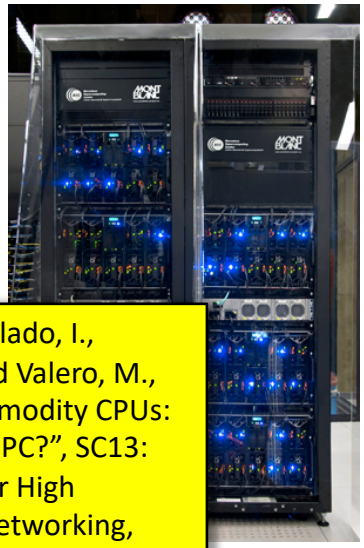


Fig. 1. IEEE listed FPGA related publications per year



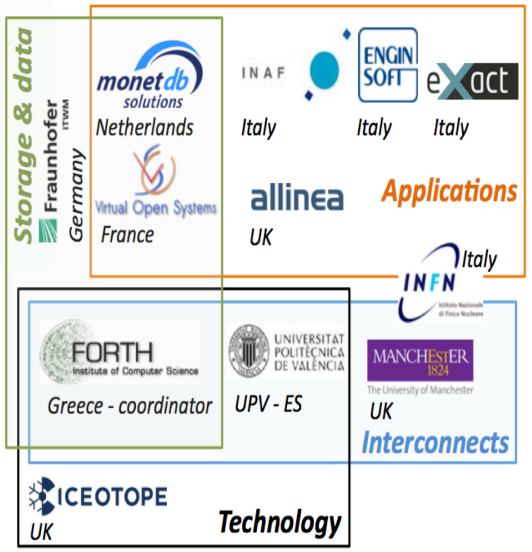
N. Rajovic, Carpenter, P., Gelado, I., Puzovic, N., Ramirez, A., and Valero, M., "Supercomputing with commodity CPUs: are mobile SoCs ready for HPC?", SC13: International Conference for High Performance Computing, Networking, Storage and Analysis. Denver, United States, pp. 40?40, 2013.

EU-funded project

- H2020 project ExaNeSt (December 2015 – May 2019)
  - System architecture definition: Network, Storage, Packaging & Cooling
  - FPGA (network and accelerator) + low power ARM processor
  - Co-Design: applications identify system requirements and evaluate the prototype
- H2020 project EuroEXA (September 2017 – February 2021)
  - ExaNeSt, ECOSCALE and ExaNoDe Follow up
    - ExaNeSt: architecture, hierarchical interconnect (ExaNet), Packaging & Cooling
    - ECOSCALE: programming environment for FPGA accelerator
    - ExaNoDe: computing node development
- HBP Wavescales2 (2016 – 2023) (Pier Paolucci talk)
  - Understand physical mechanism of cognition and brainstates
  - Development of a distributed, parallel and scalable spiking neural network simulator

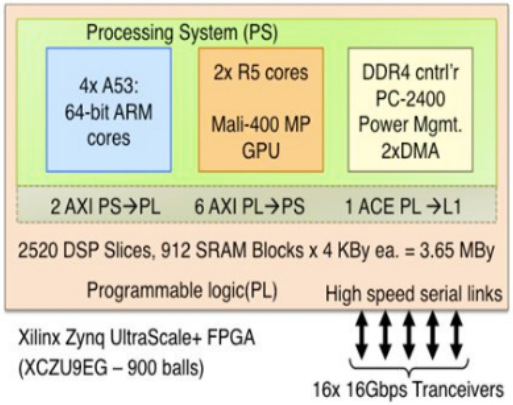
H2020 – FETHPC-2014 ExaNeSt (2015-2019)

- Evaluation of Exascale-enabling technologies (Total Budget 8M€)
  - Low-latency unified Interconnect (compute & storage traffic)
  - Fast, distributed in-node non-volatile-memory
- Extreme compute-power density
  - ARM-based (v8, 64-bit) microserver + FPGA accelerator
  - Advanced totally-liquid cooling technology
- Real scientific and data-center applications: HW/SW CO-Design
- July19-Prototype: 768 ARM cores; 192 FPGAs; 3TB of DDR3 memory
- INFN Budget (contracts and equipment): 770k€ (590k€ + 30k€ @Roma)
- INFN Roma duties:
  - Architecture Definition and Integration
    - A. Biagioni, P. Cretaro, O. Frezza, F. Lo Cicero, A. Lonardo, M. Martinelli, P. S. Paolucci, E. Pastorelli, F. Simula, P. Vicini et al., "Next generation of Exascale-class systems: ExaNeSt project and the status of its interconnect and storage development", Microprocessors and Microsystems, Volume 61, pp. 58-71, 2018 [link]
    - A. Biagioni, P. Cretaro, O. Frezza, F. L. Cicero, A. Lonardo, M. Martinelli, P. S. Paolucci, E. Pastorelli, F. Simula, P. Vicini, et al. "The next generation of Exascale-class systems: The ExaNeSt project," in 2017 Euromicro Conference on Digital System Design (DSD), pp. 510–515, Aug 2017 [link]
  - Design and implementation of the interconnect (ExaNet)
    - A. Biagioni, F. Capuani, P. Cretaro, G. De Bonis, F. Lo Cicero, A. Lonardo, M. Martinelli, P. Paolucci, E. Pastorelli, L. Pontisso, F. Simula, P. Vicini et al., "Large scale low power computing system: Status of network design in ExaNeSt and EuroEXA projects," Advances in Parallel Computing, vol. 32, pp. 750–759, 2018 [link]
    - A. Biagioni, P. Cretaro, O. Frezza, F. L. Cicero, A. Lonardo, M. Martinelli, P. S. Paolucci, E. Pastorelli, F. Simula, P. Vicini et al., "Low latency network and distributed storage for next generation HPC systems: the ExaNeSt project," Journal of Physics: Conference Series, vol. 898, no. 8, p. 082045, 2017 [link]
  - Identification of the requirements and prototype benchmarking through the proprietary spiking neural network (DPSNN)
    - A. Biagioni, F. Capuani, P. Cretaro, G. De Bonis, F. Lo Cicero, A. Lonardo, M. Martinelli, P. Paolucci, E. Pastorelli, L. Pontisso, F. Simula, and P. Vicini et al., "The brain on low power architectures: Efficient simulation of cortical slow waves and asynchronous states," Advances in Parallel Computing, vol. 32, pp. 760–769,2018 [link]

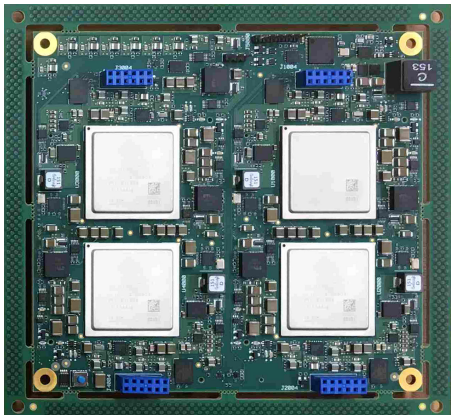


ExaNeSt Overview

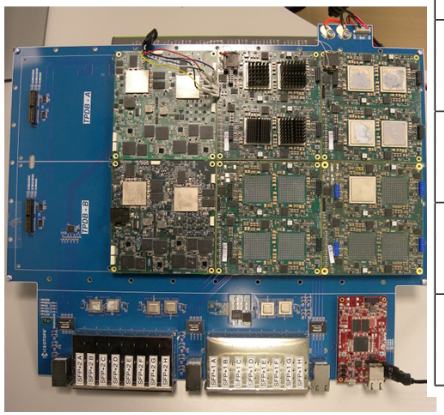
Unit



QFDB: Node



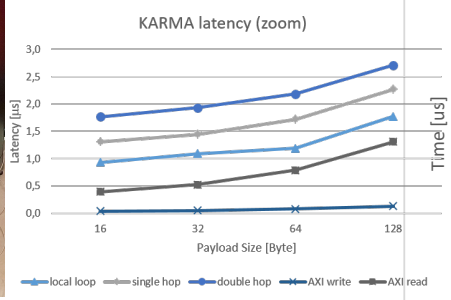
Mezzanine (Blade)

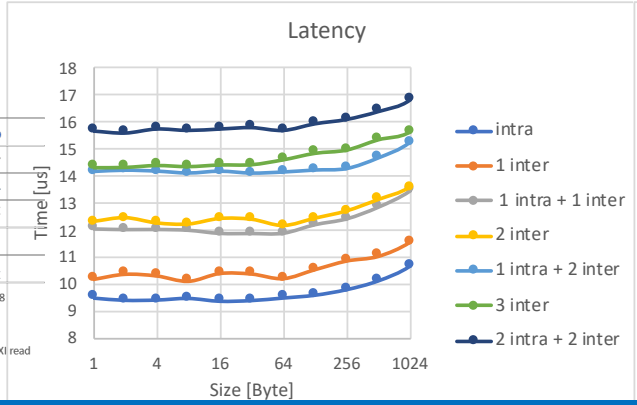


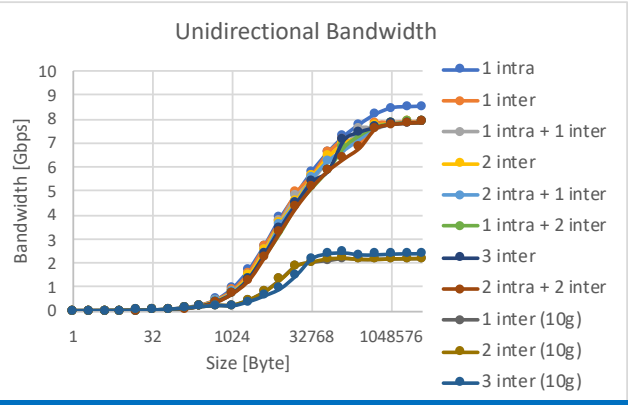
	ExaNeSt
cores per blade	64
memory per blade [GB]	256
FPGAs per blade	16
cores per chassis	576
memory per chassis [GB]	2304
FPGAs per blade	144
core per rack	1728
memory per rack [GB]	6912
FPGAs per blade	432
core per equivalent 1u	~43
memory per equivalent 1u [GB]	~173
FPGAs per equivalent 1u	~11

Rack

	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				







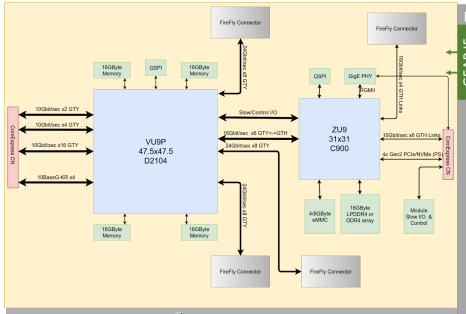
H2020 FET-HPC 2016 EuroEXA (2017-2021)

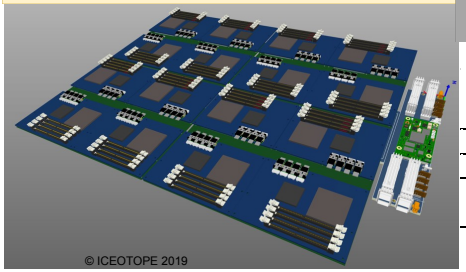
EuroEXA leverages on ExaNeSt, ExaNoDe and ECOSCALE results to deliver a world-class HPC pre-Exascale demonstrator (Total Budget 20M€)

- Energy Efficiency
  - Tighter integration, customization and hardware acceleration
  - Advanced cooling
  - Mitigation of data transfer cost (memory compression, hyperconverged storage)
- Scalability
  - UNIMEM architecture
  - Unified (for data and storage traffic) low latency, high throughput, RDMA-based interconnect
  - Hierarchical network topology
- Q1-2020 prototype: 768 ARM cores, 384 FPGAs (192 VU9+192 ZU9), 12 TB of DDR4 memory
- INFN Budget 730k€ (490k€ + 50 k€ @Roma)
- INFN Roma duties
  - Network design at sub-rack level
  - Benchmarking through application: spiking neural network simulator (DPSNN)

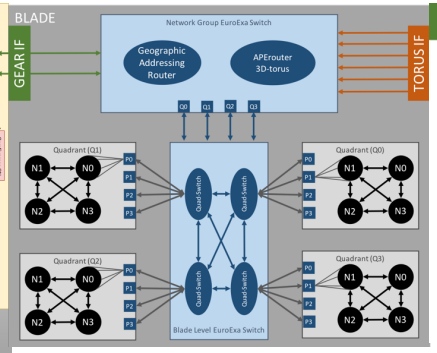
EuroEXA

CRDB: Node

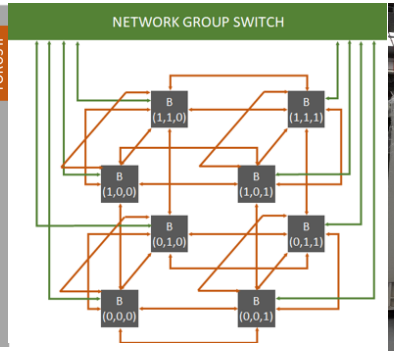




Blade

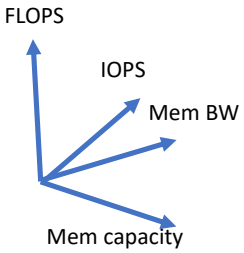


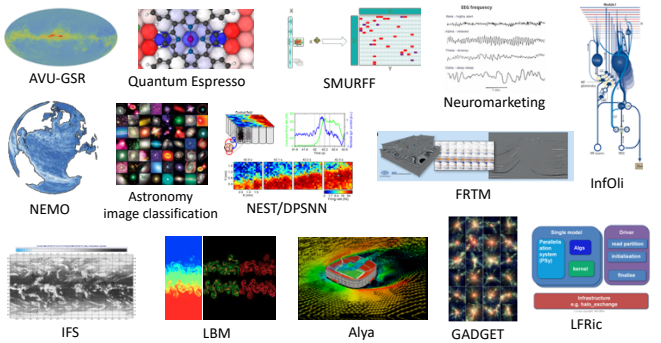
Network Group



Container

	ExaNeSt						EuroEXA					
Hierarchy	**Name**	**Fanout**	**Switching**	**Topology**	**BW**	**Latency**	**Name**	**Fanout**	**Switching**	**Topology**	**BW**	**Latency**
TIER 4	System	1000 Racks					System	250 Racks				
TIER 3	Rack	3 Chassis	10GbE ExaNet	Fat-Tree 3D-Torus	10Gbps 6x10Gbps		Cabinet	4NGs	100GbE GEAR	Fat-Tree	2x100Gbps	
TIER 2	Chassis	9 Mezzanines	ExaNet	3D-Torus	4x10Gbps	400ns per hop	Network-Group	8 Blades	ExaNet	3D-Torus	6x100Gbps	1400ns
TIER 1	Mezzanine	4 Nodes	ExaNet	Ring	2x10Gbps	1st neighbour: 400ns 2nd neighbour 800ns	Blade	16 Nodes	ExaNet	All-to-All Full Crossbar	2x16Gbps 4x16Gbps	Quadrant: 400ns Blade: 800ns
TIER 0	QFDB	4 FPGAs	ExaNet	All-to-All	16Gbps	400ns	CRDB	2 FPGAs	Chip2Chip		6x16Gbps	
FPGA	Unit	ZU9 (4 core)						VU9+ZU9				
CORE		A53						A53				





From HPC to HEP: NaNet Project

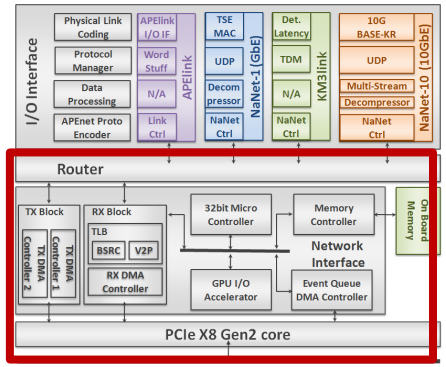
NaNet: Design and implementation of a family of FPGA-based PCIe Network Interface Cards :

- 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, …).
- Optimizing data transfers with GPU accelerators.

NA62 KM3NeT NA62

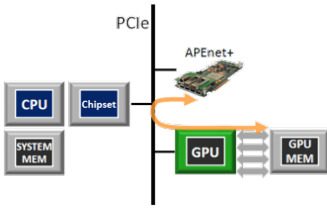
	NaNet-1	NaNet³	NaNet-10	NaNet-40
Year	Q3 - 2013	Q1 - 2015	Q2 - 2016	Q3 - 2019
Device Family	Altera Stratix IV	Altera Stratix V	Altera Stratix V	Altera Stratix V
Channel Technology	1 GbE	KM3link	10 GbE	40 GbE
Transmission Protocol	UDP	TDM	UDP	UDP
Number of Channel	1	4	4*	2
PCIe	Gen2 x8	Gen2 x8	Gen3 x8**	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

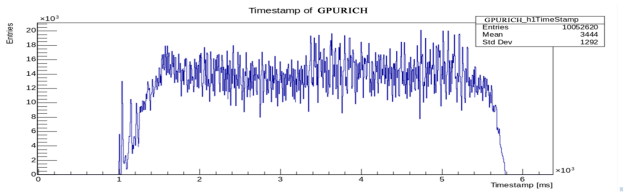


CPU+GPU

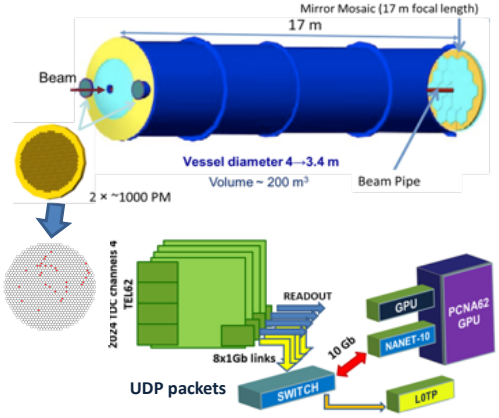
FP7-ICT EURETILE APEnet+ (2012)



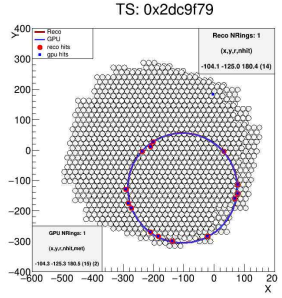
GPU-RICH generated primitives (late Oct 2018)



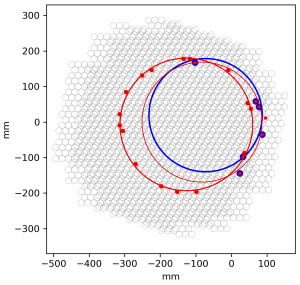
GPU-RICH overview



GPU 1 ring == Reco 1 ring



GPU 1 rings
Reco 2 rings



NaNet Publications

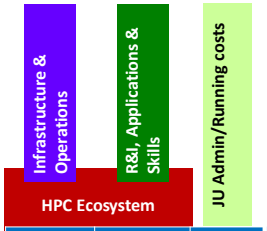
- R. Ammendola, M. Barbanera, A. Biagioni, P. Cretaro, O. Frezza, G. Lamanna, F. L. Cicero, A. Lonardo, M. Martinelli, E. Pastorelli, P. S. Paolucci, R. Piandani, D. Rossetti, F. Simula, M. Sozzi, P. Valente and P. Vicini "**Real-time heterogeneous stream processing with NaNet in the NA62 experiment,**" Journal of Physics: Conference Series, vol. 1085, p. 032022, 2018 [link]
- R. Ammendola, A. Biagioni, P. Cretaro, S. Di Lorenzo, M. Fiorini, O. Frezza, G. Lamanna, F. L. Cicero, A. Lonardo, M. Martinelli, et al., "**Development of network interface cards for TRIDAQ systems with the NaNet framework,**" Journal of Instrumentation, vol. 12, no. 03, p. C03037, 2017 [link]
- R. Ammendola, A. Biagioni, P. Cretaro, O. Frezza, G. Lamanna, F. Lo Cicero, A. Lonardo, M. Martinelli, P. Paolucci, E. Pastorelli, et al., "**Reconfigurable PCI-express cards for low-latency data transport in HEP experiments**," in Nuovo Cimento C-Colloquia and Communication in Physics vol.40 Soc Italiana Fisica, via Saragozza, 12, I-40213 Bologna, Italy, 2017 [link]
- Ammendola, R., Biagioni, A., Frezza, O., Lo Cicero, F., Martinelli, M., Paolucci, P.S., Pontisso, L., Simula, F., Vicini, P., Ameli, F., Nicolau, C.A., Pastorelli, E., Simeone, F., Tosoratto, L., and Lonardo, A., "**Nanet3: The on-shore readout and slow-control board for the KM3NeT-Italia underwater neutrino telescope**," EPJ Web of Conferences, vol. 116, p. 05008, 2016 [link]
- R. Ammendola, A. Biagioni, M. Fiorini, O. Frezza, A. Lonardo, G. Lamanna, F. Lo Cicero, M. Martinelli, I. Neri, P. Paolucci, E. Pastorelli, L. Pontisso, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "**Nanet-10: a 10GbE network interface card for the GPU-based low-level trigger of the NA62 RICH detector**," Journal of Instrumentation, vol. 11, no. 03, p. C03030, 2016 [link]
- R. Ammendola, A. Biagioni, O. Frezza, G. Lamanna, F. L. Cicero, A. Lonardo, M. Martinelli, P. S. Paolucci, E. Pastorelli, L. Pontisso, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "**NaNet: Design of FPGA-based network interface cards for real-time trigger and data acquisition systems in HEP experiments**," in 2015 IEEE Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC), pp. 1–3, Oct 2015 [link]
- R. Ammendola, A. Biagioni, O. Frezza, G. Lamanna, F. L. Cicero, A. Lonardo, M. Martinelli, P. S. Paolucci, E. Pastorelli, L. Pontisso, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "**A multi-port 10GbE PCIe NIC featuring UDP offload and GPUDirect capabilities.**," Journal of Physics: Conference Series, vol. 664, no. 9, p. 092002, 2015 [link]
- A. Lonardo, F. Ameli, R. Ammendola, A. Biagioni, A. C. Ramusino, M. Fiorini, O. Frezza, G. Lamanna, F. Lo Cicero, M. Martinelli, I. Neri, P. Paolucci, E. Pastorelli, L. Pontisso, D. Rossetti, F. Simeone, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "**NaNet: a Configurable NIC Bridging the Gap Between HPC and Real-time HEP GPU Computing**," Journal of Instrumentation, vol. 10, no. 04, p. C04011, 2015 [link]
- A. Lonardo, F. Ameli, R. Ammendola, A. Biagioni, A. Cotta Ramusino, M. Fiorini, O. Frezza, G. Lamanna, F. Lo Cicero, M. Martinelli, I. Neri, P. S. Paolucci, E. Pastorelli, L. Pontisso, D. Rossetti, F. Simeone, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "**A FPGA-based Network Interface Card with GPUDirect enabling realtime GPU computing in HEP experiments**," in Proceedings, GPU Computing in High-Energy Physics (GPUHEP2014), pp. 86–91, 2015 [link]
- R. Ammendola, A. Biagioni, R. Fantechi, O. Frezza, G. Lamanna, F. Lo Cicero, A. Lonardo, P. S. Paolucci, F. Pantaleo, R. Piandani, L. Pontisso, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "**NaNet: a low–latency NIC enabling GPU–based, real–time low level trigger systems**," Journal of Physics: Conference Series, vol. 513, no. 1, p. 012018, 2014 [link]
- R. Ammendola, A. Biagioni, O. Frezza, G. Lamanna, A. Lonardo, F. Lo Cicero, P. S. Paolucci, F. Pantaleo, D. Rossetti, F. Simula, M. Sozzi, L. Tosoratto, and P. Vicini, "**NaNet: a flexible and configurable low-latency NIC for real-time trigger systems based on GPUs**," Journal of Instrumentation, vol. 9, no. 02, p. C02023, 2014 [link]

Skill and Needs

- Hardware (Andrea Biagioni, 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)
  - Master Thesis Students, INFN External Funds support
- Software (Paolo Cretaro, Alessandro Lonardo, Luca Pontisso)
  - Driver, API (RDMA), MPI library
  - GPU in HEP trigger system
  - High Level Synthesis (C, C++, OpenCL)
  - Architectural Simulation and Architecture Design
- Physics Models and simulations and data analysis (Cristiano Capone, Fabrizio Capuani, Giulia De Bonis, *Chiara De Luca* , *Paolo Muratore*, Pier Stanislao Paolucci, Elena Pastorelli, Francesco Simula)
  - Parallel computing
  - Brain Simulation and Modelling (DPSNN, Nest), Brain Inspired Fast Learning
  - Data Analysis of Brain Experimental Data (Matlab, Python)
- Other Activities
  - Membro di commissione (PV), Referente Formazione (GDB) e Trasferimento Tecnologico (AL)
  - Teaching (PV, AL, AB), Lab2Go (GDB, PV), Lab2 (AL, LP, GDB, PSP, EP, CC)
  - Collaboration with NA62 and KM3NeT

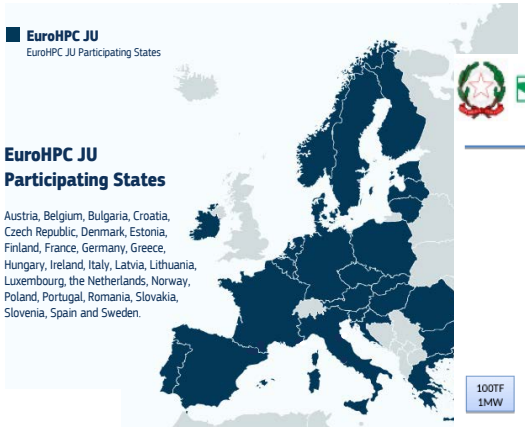
Future Activities: EuroHPC

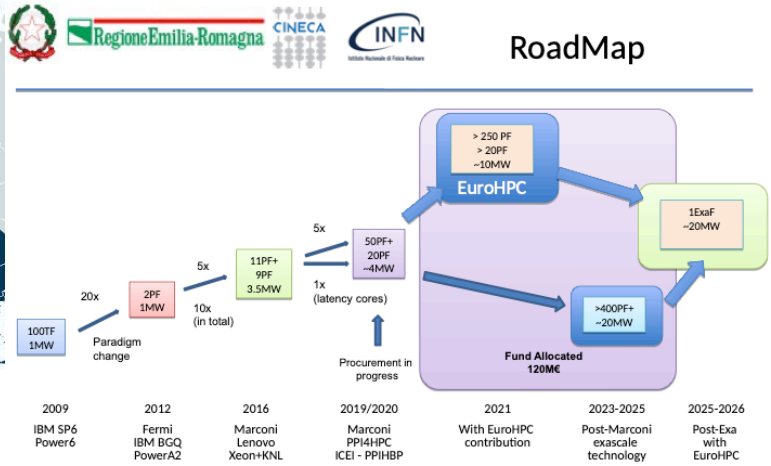
EuroHPC – Activities



- **Infrastructure + Operations**
  Acquisition of 2 pre-exascale machines and several (tbd) mid-range machines
- **Applications & Skills + R&I**
  R&I, exascale technologies and systems (incl. low-power processor); applications
- **JU Admin/running costs**
  - **JU Operation: 2019 to 2026**

~270	min 180	10	486 m€	EC
~290	~186	10	486 m€	Participating States
560	392	20	972 m€	Total
0	~420 (in kind)	2	422 m€	Private Members





~ 280 M€

Precursors to exascale

At least 2 Precursors to exascale

EU contribution: ≤50% of CAPEX and ≤50% of OPEX
MAX EU budget: 250 M€

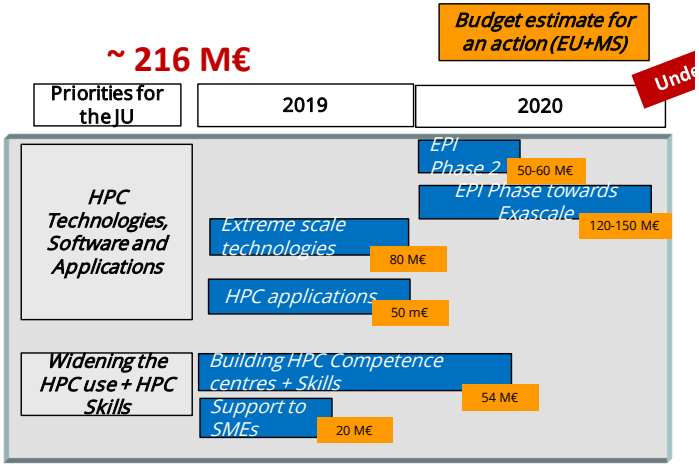
Petascale

At least 2 Petascale

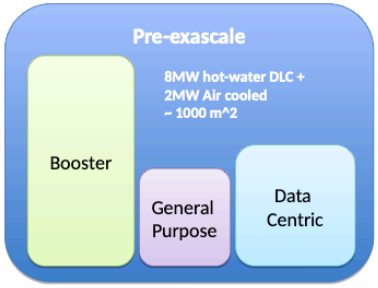
EU contribution: ≤35% of CAPEX
MAX EU budget: 30 M€

~ 216 M€

Budget estimate for an action (EU+MS)



Proposed Systems



Use cases:
- 10x computing capability in a large set of key applications for science, industry and society (CoEs, HEP, Pharma, Oil&GAS), and keep the European leadership.
- gain sovereignty on strategic technologies for the European economic wealth, like Artificial Intelligence, Cybersecurity and Internet of Thing,
- tackle relevant and urgent societal challenges.

Advanced experimental platform towards exascale

BASED on European IPs (e.g. PRACE PCP)

Use cases:
- Validate Inference engines on FPGA (in collaboration with ST micro), with training performed on the Booster.
- Acceleration of Quantum inspirited Algorithm for basic science.
- Image/Video processing & Cybersecurity
- Large scale Spiking neural networks
- Data processing for HEP experiments

Sanzio Bassini, March 2019 — Cineca - SuperComputing Applications & Innovations — 6

R&I activity

- ❑ Budget: 3% (3-4M€)
- ❑ FPGA+ARM server
- ❑ INFN apps
  - Quantum
  - Neural network
  - HEP data processing
- ❑ Funded (R&I under discussion)

APE LAB Team

THANK YOU

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