ExaNeSt status and INFN perspective in H2020 HPC EU framework

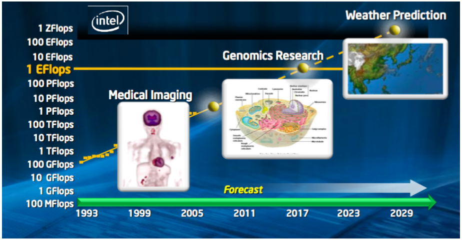
Piero Vicini (INFN Rome)

Workshop CCR
May 22-26, 2017,
LNGS

Contents

- Brief technology survey
- What is happening in HPC arena
- EU goals
- INFN activity in EU H2020 FET FP
  - ExaNeSt project: recap and status
  - Introduction to EuroExa project
- New initiative to support use of many-core architectures for INFN computing

The needs for ExaScale systems in science



- HPC is mandatory to compare observations with theoretical models
- HPC infrastructure is the theoretical laboratory to test the physical processes.

Let's talk of Basic Science...

- High Energy & Nuclear Physics
  - LQCD (again...), Dark-energy and dark matter, Fission/Fusion reactions (ITER)
- Facility and experiments design
  - Effective design of accelerators (also for medical Physics, GEANT...)
  - Astrophysics: SKA, CTA
  - ...
- Life science
  - Personal medicine: individual or genomic medicine
  - Brain Simulation <– HBP (Human Brain Project) flagship project

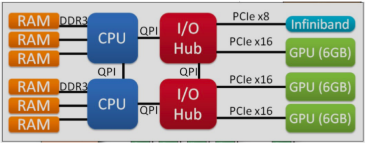
Technological challenges for ExaScale systems

Just to name a few....

- Power efficiency and compute density
  - huge number of nodes but limited data center power and space
- Memory and Network technology
  - memory hierarchies: move data faster and closer...
  - increase memory size per node with high bandwidth and ultra-low latency
  - distribute data across the whole system node set but access them with minimal latency...
- Reliability and resiliency
  - solutions for decreased reliability (extreme number of state -of-the-art components) and a new model for resiliency
- Software and programming model
  - New programming model (and tools) needed for hierarchical approach to parallelism (intra-node, inter-node, intra-rack....)
  - system management, OS not yetready for ExaScale...
- Effective system design methods
  - CO-DESIGN: a set of a hierarchical performance models and simulators as well as commitment from apps, software and architecture communities

Hybrid Supercomputer: CPU + Accelerators

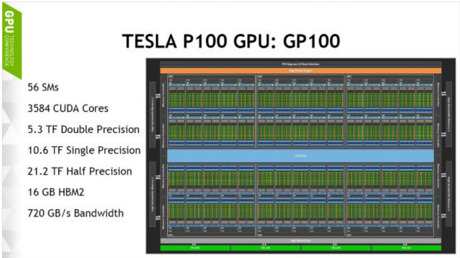
Most high-end HPC systems are characterized by *hybrid architecture*



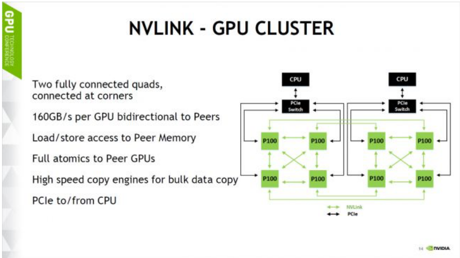
- ASIP, FPGA or commodity components (GPGPU...)
- Better $/*PeakFlops*: offload cpu task to accelerator able to perform faster
- May consume less energy and may be better at streaming data.
- —> warning!!!:
  - computing efficency ϵ (Sustained/Peak) not impressive
  - it's a function of accelerator and network...

	Nazione	Score	Numero Nodi	Tipologia Acc.	Peak Perf(Pflops)	Linpack Perf (Pflops)	Efficiency	Power (MW)	Interconnect
Tianhe-2	China	1	16000 (2CPU+3PHI)	Xeon + PHI	54,9	33,8	62%	17,8	Proprietary
Titan	USA(Oak R.)	2	18000(1CPU+1K20x)	Opteron + K20x	27,1	17,6	65%	8,2	Cray Gemini
Piz Daint	Switzerland	6	5272(1CPU+1K20x)	Xeon + K20x	7,8	6,2	79%	2,3	Cray Aries
Stampede	USA (TACC)	7	6400	Xeon + PHI	8,5	5,1	60%	4,5	Infiniband

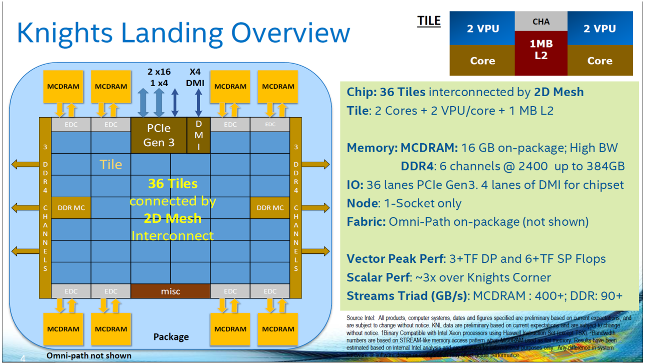
NVidia Pascal P100 and the last generation Volta V100 (1.5x) recently announced...

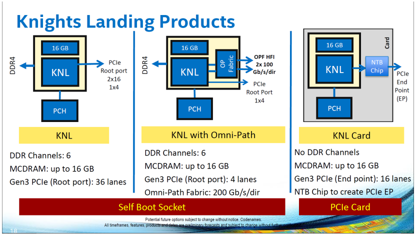


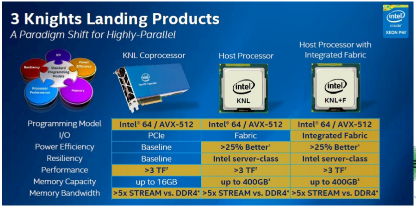
Tesla Products	Tesla P100	Tesla K80	Tesla K40	Tesla M40
GPU	GP100 (Pascal)	2 x GK210 (Kepler)	GK110 (Kepler)	GM200 (Maxwell)
SMs	56	26 (13 per GPU)	15	24
CUDA cores	3840	4992 (2 x 2496)	2880	3072
Base Clock	1328 MHz	560 MHz	745 MHz	948 MHz
GPU Boost Clock	1480 MHz	875 MHz	810/875 MHz	1114 MHz
Peak Double Precision	5.3 TFLOPS	2.91 TFLOPS	1.68 TFLOPS	.2 TFLOPS
Peak Single Precision	10.6 TFLOPS	8.73 TFLOPS	5.04 TFLOPS	7 TFLOPS
Memory Interface	4096-bit HBM2	2 x 384-bit GDDR5	384-bit GDDR5	384-bit GDDR5
Memory Size	16 GB	24GB (12GB per GPU)	12 GB	24 GB
Peak Bandwidth	720 GB/s	480 GB/s (240 GB/s per GPU)	288 GB/s	288 GB/sec
TDP	300 Watts	300 Watts	235 Watts	250 Watts
Transistors	15.3 billion	2 x 7.1 billion	7.1 billion	8 billion
GPU Die Size	610 mm²	2 x 561mm²	551 mm²	601 mm²
Manufacturing Process	16-nm	28-nm	28-nm	28-nm



Accelerators: INTEL PHY (ex-MIC)

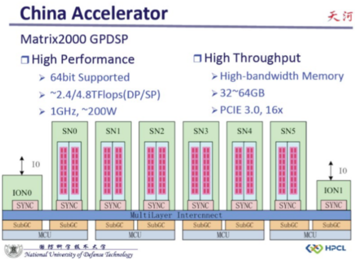






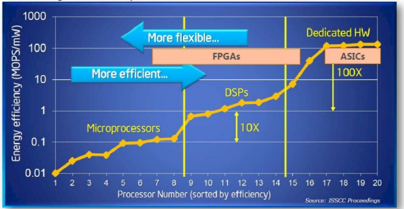
Next (almost) ExaScale systems around the World

- US CORAL (Collaboration of Oak Ridge, Argonne, and Livermore) project, 525+M$ from DOE, for 3 100-200 PetaFlops systems in 2018-19 (Pre-Exascale system), ExaScale in 2023
  - *Summit/Sierra* OpenPower-based (IBM P9 + NVidia GPU + Mellanox) 150(300) PFLops/10MW
  - *Aurora* Intel-based (CRAY/INTEL, Xeon PHI Knights Hill, Omnipath) 180(400) PFlops/13MW
- JAPAN FLAGSHIP2020 RIKEN + Fujitsu
  - derived from Fujitsu K-computer, SPARC64-based + Tofu interconnect, delivered in 2020
- CHINA ??? , NUDT + Government
  - ShenWei and FeiTang CPUs plus proprietary GPU and network... delivered in 2020

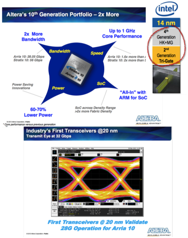


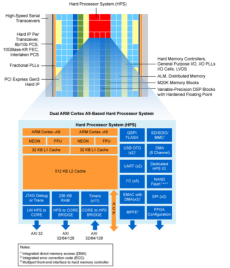
An emerging new player in hybrid HPC: FPGA

- Stratix10 high-end, introduction 2016
- INTEL TriGate 14nm -> 30% less than old generation power consumption
- 96 transceivers @32Gbps (56Gbps?) for chip-to-chip interconnection and @28Gbps for backplane/cable interconnection
- Many industrial standards supported included CAUI-x (Nvlink)
- tons of programmable logic @1GHz
- ...and "for free"
  - 10 Tflops of DSP single precision FP
  - HMC (3D mem, high bandwidth) support
  - Multiple (4->8) ARM Cores (a53/57) @1.5GHz
- Similar in performance: XILINX Zynq UltraScale+ MPSoC Devices



Source: Bob Broderson, Berkeley Wireless group

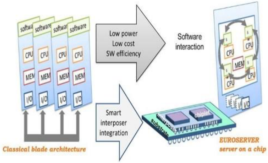




ARM & HPC

Several attempts to use ARM low power processors in high end computing

- Server and micro-server ARM-based
  - AMCC X-gene 3, 32 v8-A cores@3GHz,
  - CAVIUM ThunderX SoCs up to 48 v8-A cores@2.4GHz
  - Broadcom/Qualcomm multi-core, Samsung SoC Exynos
- EU-funded projects
  - Mont-blanc project (BSC)
  - UniServer
  - ....



- INFN COSA project measured energy efficiency of low power architecture ARM based for scientific computing (Astrophysics, Brain simulation, Lattice-Boltzmann fluid-dynamics,..). On average:
  - $\sim$3x ratio x86 core / ARM core performances
  - but $\sim$10x ratio x86 core / ARM power consumption
  - –> ARM architectures *3x less* energy to solution for scientific applications

European Commission President Jean-Claude Juncker

"Our ambition is for Europe to become one of the top 3 world leaders in high-performance computing by 2020"

French-German Conference on Digital;
Paris, 27 October 2015

—> EuroHPC: 7 countries agreement on pushing HPC development in Europe (Digital Day, March 2017)

What next in Europe?

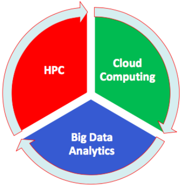
HPC Objectives (1)

- **Acquisition** (in 2020-2021) of 2 operational **pre-exascale** and (in 2022-2023) two full **exascale** machines (of which one based on European technology)
- **Interconnection and federation** of national and European HPC resources and creation of an HPC and Big Data service infrastructure facility
- **Demonstrating and testing** technology performance towards exascale through scientific & industrial compute-intensive applications

HPC Objectives (2)

Build a world-class European High Performance Computing (HPC), Big Data and Cloud Ecosystem

Enabled by the Convergence of 3 big technologies



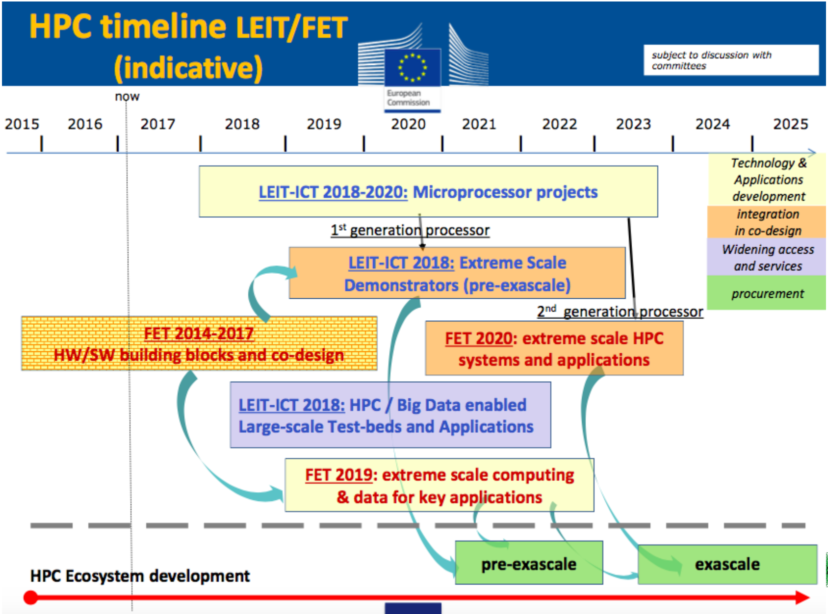
- Major investments so far both at MS and EU level [FP7, H2020]
- Numerous research players (academia and industry)
- HPC and Big Data PPPs, PRACE, GEANT, etc.

HPC/EDI – Funding needs [COM(2016) 178 of 19/4/2016]

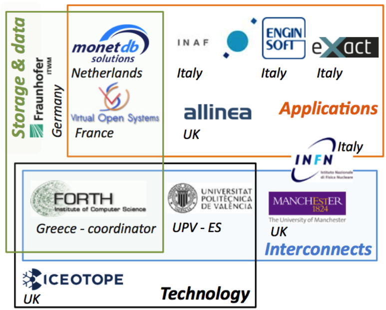
- **1.5 B€** for 2 pre-exascale and 2 exascale machines
- **1.7 B€** for the interconnection and federation of supercomputing infrastructures
- **0.5 B€** for processor and for wider access to HPC facilities for SMEs
- **1.0-1.5 B€** for demo and testing of industrial applications

- Total: 4.7 - 5.2 BEuro needed....
- mainly from National and Regional funds...
- 1.5 BEuro for sytems procurement
- 0.15 BEuro for European Processor NRE

What next in Europe?



ExaNeSt: FETHPC 2014



ExaNeSt: European Exascale System Interconnection Network & Storage

- EU Funded project H2020-FETHPC-1-2014
- Duration: 3 years (2016-2018). Overall budget about 7 MEuro.
- Coordination FORTH (Foundation for Research & Technology, GR)
- 12 Partners in Europe (6 industrial partners)

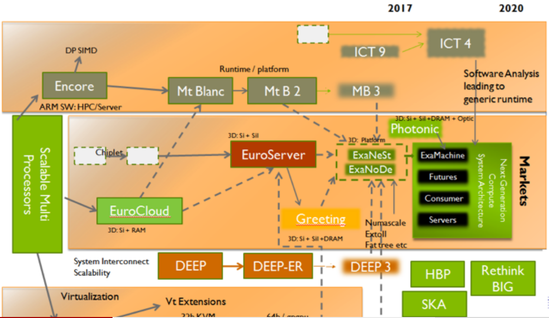
ExaNeSt objectives

"...Overall long-term strategy is to develop a European low-power high-performance Exascale infrastructure based on ARM-based micro servers..."

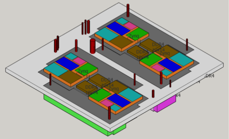
- System architecture for datacentric Exascale-class HPC
  - Fast, distributed in-node non-volatile-memory
  - Storage Low-latency unified Interconnect (compute & storage traffic)
    - RDMA + PGAS to reduce overhead
- Extreme compute-power density
  - Advanced totally-liquid cooling technology
  - Scalable packaging for ARM-based (v8, 64-bit) microserver
- Real scientific and data-center applications
  - Applications used to identify system requirements
  - Tuned versions will evaluate our solutions

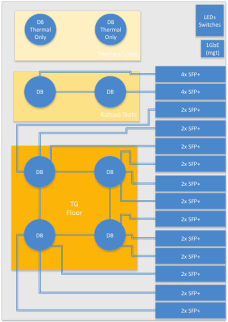
ExaNeSt ecosystem

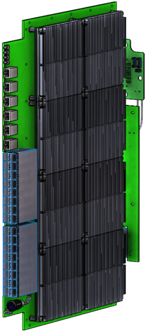
- EuroServer: Green Computing Node for European microservers
  - *UNIMEM* PGAS model among ARM computing nodes
- INFN EURETILE project: *brain inspired* systems and applications
  - APEnet+ network on FPGA + brain simulation (DPSNN) scalable application
- Kaleao: Energy-efficient uServers for Scalable Cloud Datacenters
  - startup company interested in commercialisation of results
- *Twin* projects: ExaNode and EcoScale
  - ExaNode: ARM-based Chiplets on silicon Interposer design
  - EcoScale: efficient programming of heterogenous infrastructure (ARM + FPGA accelerators)



ExaNeSt prototypes







- Computing module based on Xilinx Zynq UltrScale+ FPGA...
  - Quad-core 64-bit ARM A53
  - ∼1 TFLOPS of DSP logic
- ... placed on small Daugther Board (QFDB) with
  - 4 FPGAs, 64 GB DDR4,
  - 0.5-1 TB SSD,
  - 10x 16Gb/s serial links-based I/O per QFDB
- mezzanine(blade) to host 8 (16 in second phase) QFDBs
  - intra-mezzanine QFDB-QFDB direct network
  - lots of connectors to explore topologies for inter-mezzanine network

ExaNeSt prototypes

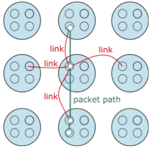
- ExaNeSt high density innovative mechanics...
  - 8(16) QFDBs per mezzanine
  - 9 blades per chassis
  - 8-12 chassis per rack
- ...totally liquid cooling
  - track 1: immersed liquid cooled systems based on convection flow
  - track 2: phase-change (boiling liquid) and convection flow cooling (up to 350 kW of power dissipation capability...)

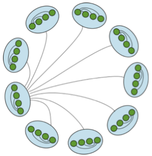
- $\sim 7PFlops$ per racks and $20GFlops/W$
- Extrapolating from current technology, ExaNeSt-based Exascale system with 140 racks, 21M ARM cores and 50MW

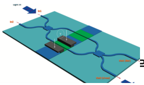
ExaNeSt Interconnect

ExaNeSt is working testbed FPGA-based to explore and evaluate innovative network architectures, network topologies and related high performance technologies.

- Unified approach
  - merge interprocessor and storage traffic on same network medium
  - PGAS architecture and RDMA mechanisms to reduce communication overhead
- innovative routing functions and control flow (congestion managements)
- explore performances of different topologies
  - Direct blade-to-blade networks (Torus, Dragonfly,...)
  - Indirect blade-switch-blade networks
- All-optical switch for rack-to-rack interconnect (ToR switch)
- Support for resiliency: error/detect correct, multipath routing,...
- Scalable network simulator to test large scale effects in topologies





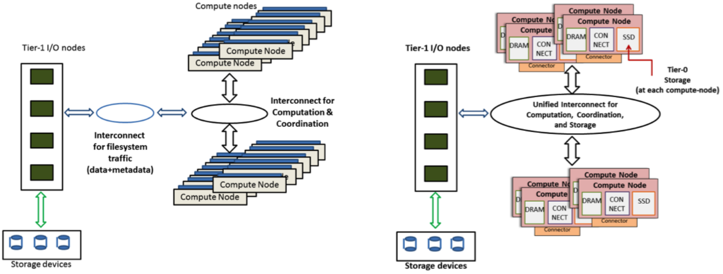


Role of applications in ExaNeSt

Co-design approach

- Applications define quantitative requirements for the system under design
- Applications evaluate the hw/sw system
- Applications list:
  - Cosmological n-Body and hydrodynamical code(s) (INAF)
    - Large-scale, high-resolution numerical simulations of cosmic structures formation and evolution
  - Brain Simulation (DPSNN) (INFN)
    - Large scale spiking behaviours and synaptic connectivity exhibiting optimal scaling with the number of hardware processing nodes (INFN).
    - Mainly multicast communications (all-to-all, all-to-many).
  - Weather and climate simulation (ExactLab)
  - Material science simulations (ExactLab and EngineSoft)
  - Workloads for database management on the platform and initial assessment against competing approaches in the market (MonetDB)
  - Virtualization Systems (Virtual Open systems)

ExaNeSt storage



- Distributed storage: NVM close to the computing node to get low access latency and low power access to data
- based on BeeGFS open source parallel filesystem with caching and replication extensions
- Unified interconnect infrastructure per storage and inter-node data communication
- Highly optimized I/O path in the Linux kernel

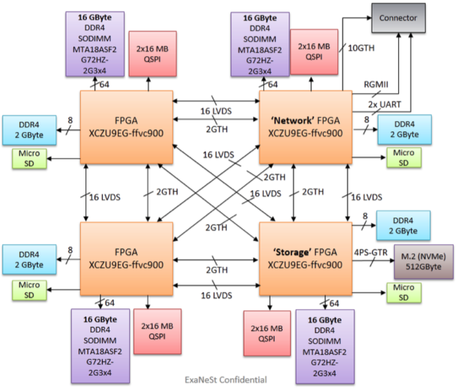
ExaNeSt status

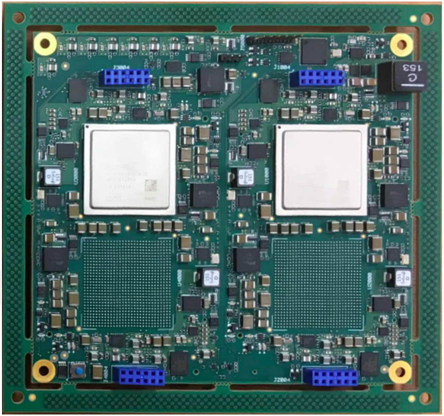
- First review at end of July
- We decided to go for *rising of expectations*...
  - a more compact and powerful system: QFDB hosting 4 FPGAs
  - a fully working prototype (not only a demonstrator...)
  - a tighter and fruitful collaboration with our twins projects (signed agreements with EcoScale and ExaNode) to get a complete software stack included MPI libraries and OpenCL framework for acceleration

ExaNeSt status

- First review at end of July
- We decided to go for *rising of expectations*...
  - a more compact and powerful system: QFDB hosting 4 FPGAs
  - a fully working prototype (not only a demonstrator...)
  - a tighter and fruitful collaboration with our twins projects (signed agreements with EcoScale and ExaNode) to get a complete software stack included MPI libraries and OpenCL framework for acceleration
- but obviously got some delays...
  - QFDB module expected for the end of summer
  - new ICEOTOPE mechanics and cooling will be installed in Crete at end of July
  - large QFDB-based testbed expected end of the year...
  - first ExaNeSt MPI release not before Oct 2017

QFDB under test and validation...

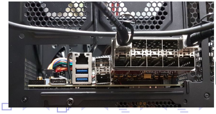




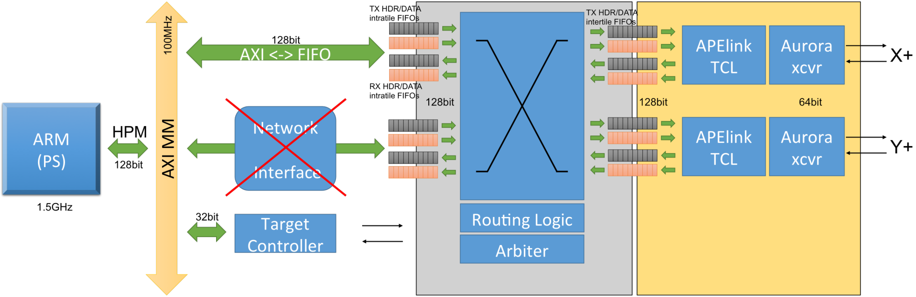
ExaNeSt highlights: QFDB and testbed

Currently assembled in Rome (and Heraklion) and used for FPGA firmware development and hardware test

- Trenz TE0808-03ES2-S: XIlinx Zynq UltraScale+ devkit (same FPGA of QFDB)
- 4(+2) nodes interconnected via 1GbE commodity network
- 2x2 (X,Y) Direct topology using on-board couple of 10gbps links on SFP+ connectors
  - succesfully tested FMC splitter module 10port SFP+
  - succesfully tested copper cables (2m) and AOC (active optical cable)
  - preliminary SI analysis of Xilinx links using Tektronix real time scope @50GSamples
  - execution of 48 hours of ping-pong test



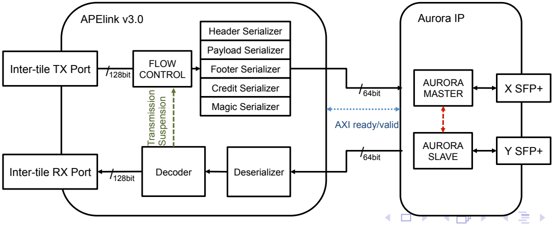
ExaNeSt highlights: INFN ExaNet Router architecture



	CLB LUTs (274080)	CLB Registers (548160)	CLB (34260)	Block RAM Tile (912)	GTHE4 CHANNEL (16)
Top Exanet	20505 (7%)	30608 (6%)	5410 (16%)	113.5 (13%)	2 (13%)
Top Core	8547	14104	2508	14.5	0
Switch	5917	6950	1687	96	0
Links	5594	8949	1782	3	2

ExaNet router it's NOT a simple APEnet+ porting...

- compliant with ExaNet protocol
- totally parametric (header/footer width, virtual channel number, credit width)
  - currently implemented 2 VCHs/direction to avoid deadlocks
- a brand new Data Link Controller (APElink TCL)
  - low latency, AXI compliant, valid/ready interface with Aurora IP
  - new, low latency credit management: 8 bit per VCH , programmable threshold values
- byte enable management developed and currently under test
- Routing&Arbiter infrastructure allows to implement an enhanced DOr routing function and VCH select based on *priority* ...



INFN ExaNet Router: timing details

TX SIDE: 3 clock cycles (from header FIFO empty to axi IF Aurora)

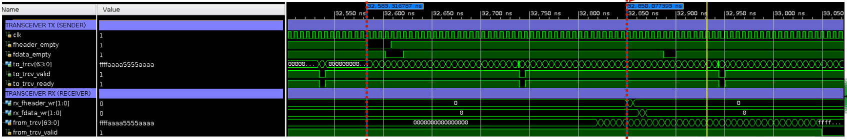
RX SIDE: 5 clock cycles (from AXI IF Aurora to header in the FIFO)

Write Enable

APElink – single hop latency
267ns (~40 clock cycles)

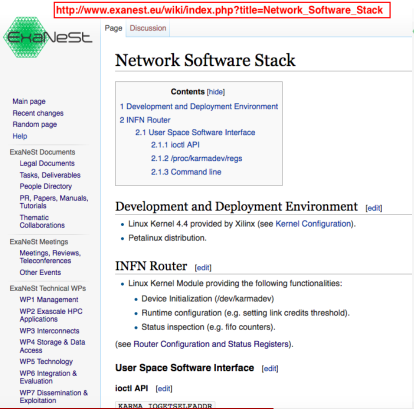
FIFO Header not empty

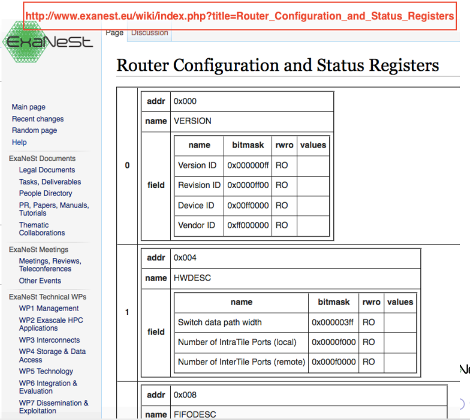
FIFO Header write enable



KARMA (*King ARM Architecture*) is a software-oriented testbed for our network subsystem

- Router FIFOs directly connected to the ARM HPM AXI port through an adapter IP (streaming –> mem map protocol)
- A set of configuration/status registers accessed through AXI and embedded in a custom IP (*Target Controller*

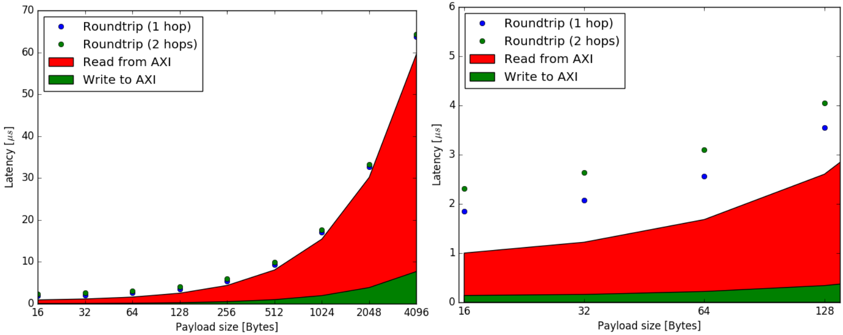




0	addr	0x000				
	name	VERSION				
	field	name	bitmask	rwro	values	
		Version ID	0x000000ff	RO		
		Revision ID	0x0000ff00	RO		
		Device ID	0x00ff0000	RO		
		Vendor ID	0xff000000	RO		
1	addr	0x004				
	name	HWDESC				
	field	name	bitmask	rwro	values	
		Switch data path width	0x000003ff	RO		
		Number of IntraTile Ports (local)	0x0000f000	RO		
		Number of InterTile Ports (remote)	0x000f0000	RO		
	addr	0x008				
	name	FIFODESC				

First sketch of test (user space) writes commands/data to the hardware

- single and dual hops test; no DMA, no interrupts, no system-wide locking and no fast virtual-to-physical address translation



- Bus width(frequency): Router+AXI 128b(100MHz), Aurora Intf 64b(156.25MHz)
- AXI Write $\sim$ 4$cycles$@100MHz; AXI Read $\sim$ 21$cycles$@100MHz
- Router Hardware Latency $\sim$ 840ns per hop
  - Router $\sim$ 20cycles@100MHz, APElink $\sim$ 10cycles@156MHz, + Aurora...
  - AURORA latency for TX-ready-RX-ready in loopback configuration $\sim$ 30$cycles$@156MHz ($\sim$ 200nS)

ExaNeSt highlights: DPSNN on TRENZ testbed

- Execution time of 5 seconds of simulated cortical activity on TRENZ and INTEL platforms for a configuration of an 8 by 8 bi-dimensional grid of neural columns, mapped on a growing number of MPI processes.

Nodes	MPI procs[1]	TRENZ[2] (s)	INTEL[3] (s)
1	1	3656.5	632.9
1	2	1964.6	336.0
1	4	1151.8	181.6
2	8	600.5	83.2
4	16	317.1	40.7

1. Results here are for OpenMPI; MPICH numbers are the same order of magnitude and not reported.
2. TRENZ cluster: 4 nodes, 4 ARM cores (A-53 1.5GHz, 64 bit arch).
3. INTEL cluster: 4 dual-socket nodes, 2 processes per socket, (corresponding to only 2 used cores out of a six-core Intel(R) Xeon(R) E5-2630 v2 CPU (clocked at 2.60GHz). Nodes are interconnected through a Mellanox InfiniBand network.

–> Activity co-funded by EU projects HBP and EXANEST

ExaNeSt highlights: DPSNN Code Optimization

- Some optimizations have been already implemented on the DPSNN code (in view of *million cores* scaling):
  - Optimization of message sizes: payloads were split to a fixed length part plus a remainder (empty most of the times) to reduce the total number of messages
  - Memory layout was changed to increase buffer contiguity: better cache locality and streamlining of messages
  - Memory optimization in axonal-spikes management
- Planned code re-engineering
  - A two-level hierarchy of communications can be implemented - different communicators are made for *local* and *distal* messages and can use different, simultaneous channels (e.g. shared memory for local + available network device for distal)
  - Coherency island support
  - Differentiation among collectives and point-to-point communications at different levels of the hierarchy
- Further possible optimizations
  - Communication in critical areas can be changed from standard (MPI) send/receive semantics to RDMA semantics...

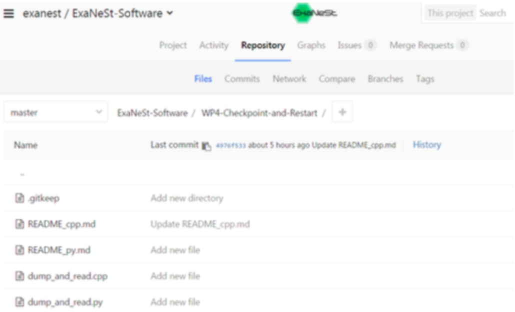
–> Activity co-funded by EU projects HBP and EXANEST

CNAF contribution...

Checkpoint-and-restart simulator

3

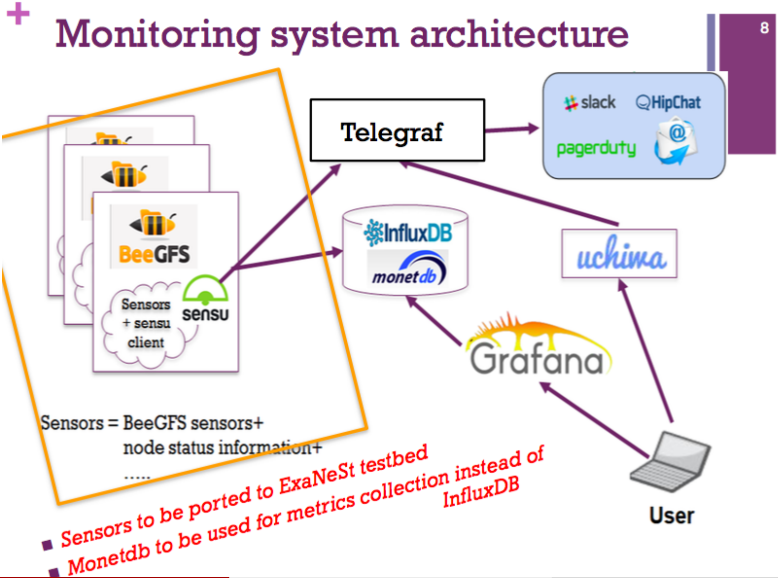
- **Synthetic test to stress the storage system simulating I/O behaviours of data-intensive HPC applications**
  - Developed a C++ version and a Python version
  - Tested on BeeGFS testbed @ INFN-CNAF
  - Available in the GitLab of the project
  - *To be tested on the ExaNeSt prototype*



Parameters that can be configured:

- Number of MPI processes
- Size of the array handled by each MPI process
- Frequency of store/load operations

CNAF contribution...



EuroExa: FETHPC 2016

PROJECT FULL TITLE: Co-designed Innovation and System for Resilient Exascale Computing in Europe: From Applications to Silicon

ACRONYM: EuroEXA

WORK PROGRAM TOPIC: FETHPC-01-2016

TYPE OF ACTION: Research and Innovation Action (RIA)

NAME OF COORDINATING PERSON: Dr. Georgios Goumas

European Commission

EuroEXA

Resilient Exascale Computing in Europe: From Applications to Silicon

LIST OF PARTICIPANTS

Part. No	Participant Organisation name	Short Name	Country
1	Institute of Communications and Computer Systems	ICCS	GR
2	University of Manchester	UNIMAN	UK
3	Barcelona Supercomputing Center	BSC	ES
4	Foundation for Research and Technology – Hellas	FORTH	GR
5	Science and Technology Facilities Council	STFC	UK
6	Interuniversitair Micro-Electronica centrum IMEC VZW	IMEC	BE
7	ZeroPoint Technologies AB	ZPT	SE
8	Iceotope Research & Development Ltd.	ICE	UK
9	Allinea Software Ltd	ALLIN	UK
10	Synelixis Lyseis Plirof. Automatismou & Tilepikoinonion Monoprosopi EPE	SYN	GR
11	Maxeler Technologies Limited	MAX	UK
12	Neurasmus BV	NEUR	NL
13	Istituto Nazionale di Fisica Nucleare	INFN	IT
14	Istituto Nazionale di Astrofisica	INAF	IT
15	European Centre for Medium-range Weather Forecasts	ECMWF	INT
16	Fraunhofer Gesellschaft zur Foerderung der Angewandten Forschung E.V.	FRAUN	DE

ExaNeSt

EuroExa: abstract(1)

... *EuroEXA brings a holistic foundation from multiple European HPC projects and partners together with the industrial SME (MAXeler for FPGA data-flow; ICEotope for infrastructure; ALLINea for HPC tooling and ZPT to collapse the memory bottleneck)...*

–> Computing platform as a whole thanks to consortium based on SME and key European academic partners

... *co- design a ground-breaking platform capable of scaling peak performance to 400 PFlops in a peak system power envelope of 30MW ... we target a PUE parity rating of 1.0 through use of renewables and immersion-based cooling... modular-integration approach, novel inter-die links and the tape-out of a resulting EuroEXA processing unit with integration of FPGA for prototyping and data-flow acceleration.*

–> challenging targets achievable through adoption of beyond-state-of-the-art tech.

EuroExa: abstract(2)

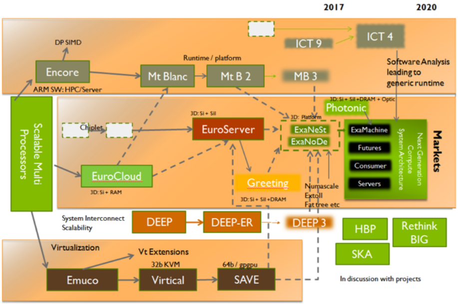
... *a homogenised software platform offering heterogeneous acceleration with scalable shared memory access...*

... *a unique hybrid, geographically-addressed, switching and topology interconnect within the rack offering low-latency and high-switching bandwidth...*

... *a rich mix of key HPC applications from across climate/weather, physics/energy and life-science/bioinformatics domains*

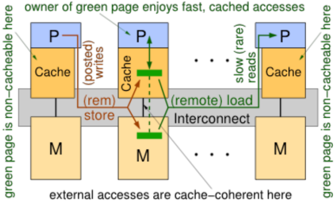
... *deployment of an integrated and operational peta-flop level prototype hosted at STFC, monitored and controlled by advanced runtime capabilities, equipped by platform-wide resilience mechanisms.*

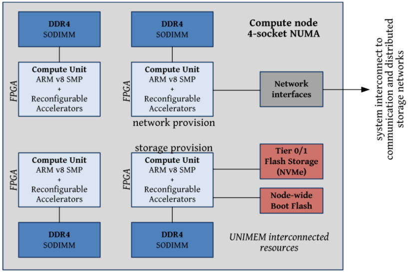
- EuroExa will leverage on results of previous projects
  - UniServer: general approach to low-power based HPC computing and UniMem architecture;
  - ExaNode: low power, high performance, multi core ARM-based CPU
  - ExaNeSt: high-enf FPGA in HPC system, network architecture, advanced system mechanics and cooling;
  - EcoScale: FPGA programming and use as application customized computing accellerators

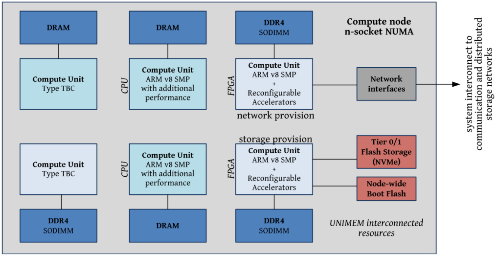


EuroExa (few) details

- high efficiency computing node with low latency (local and remote) memory access...

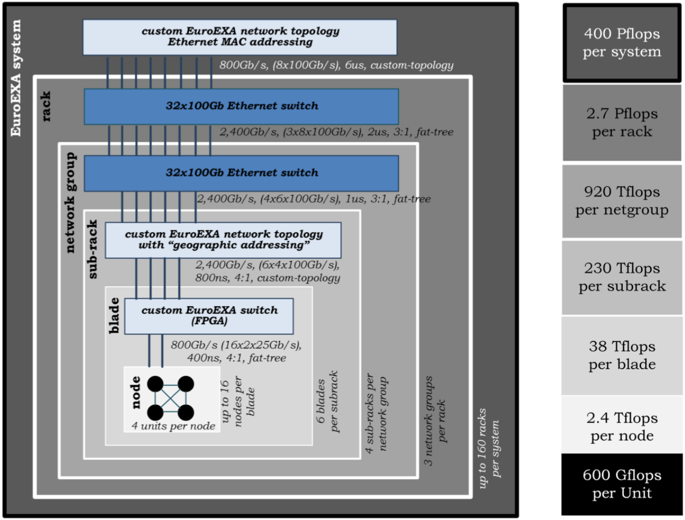






EuroExa (few) details

- Balanced, hierarchical network...



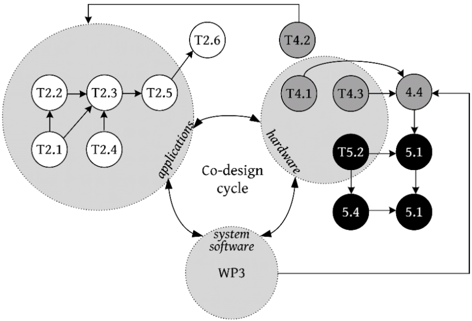
EuroExa (few) details

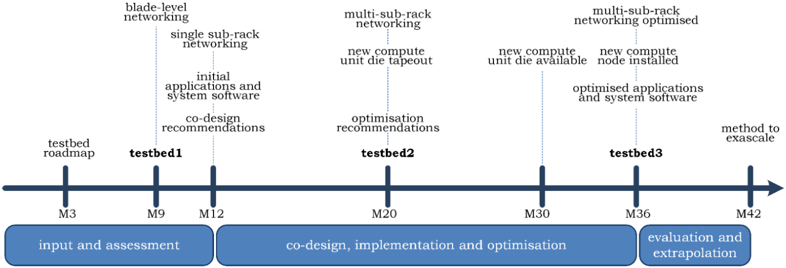
- Low power...

Hierarchy	Proposed number of hierarchical elements	Power (W)	Description
Compute Unit	1 device ARMv8 + FPGA	25	Power of compute centric code on UltraScale+ (Xilinx power estimator)
Node	4 units	125	Units plus PCB, SSD, NIC
Blade	16 nodes	2,050	Nodes plus embedded switch
Sub-Rack	6 blades	12,900	Add 5% PSU inefficiency
Net-Group	4 sub-racks	53,000	Subracks plus level-2 Mellanox switch
Rack	3 net-groups	175,000	Top-of-rack switch and infrastructure
System	160 racks	30 MW	Estimated peak 400 PFLOP system

EuroExa methods...

- EuroExa will use a strong co-design approach and incremental system design and integration





EuroExa structure

	WP1	WP2	WP3	WP4	WP5	WP6	Total PMs
ICCS	**18**	68	22	0	0	10	**118**
UNIMAN	10	24	62	**163**	40	5	**304**
BSC	10	**92**	94	4	0	5	**205**
FORTH	1	29	**88**	70	16	6	**210**
STFC	1	36	18	6	**36**	3	**100**
IMEC	1	36	0	0	0	5	**42**
ZPT	1	3	4	52	0	3	**63**
ICE	3	4	0	14	50	**32**	**103**
ALLIN	1	12	14	2	0	3	**32**
SYN	1	35	28	0	6	5	**75**
MAX	1	6	94	4	0	3	**108**
NEUR	1	40	11	0	0	3	**55**
INFN	1	38	24	10	40	2	**115**
INAF	1	48	13	2	0	2	**66**
ECMWF	1	39	0	0	0	2	**42**
FRAUN	1	31	37	0	0	2	**71**
						Total PMs	**1709**

- Start date and duration: September 1st, 2017, 42 months
- Total budget: 20MEuro (>7MEuro for hardware procurement and NRE for silicon);
- INFN and UniFE mainly in :
    - benchmarking through applications: neural network simulator (RM1, link with HBP projects), LBE simulation (UniFE)
    - Network design at sub-rack level (RM1)
- INFN budget: 730 kEuro, 3 FTEs for the whole project duration

- Europe (finally...) started to fund and push for HPC technologies developments: EuroHPC, EXDCI, IPCEI,...
- A couple of FET HPC calls to explore methods and technologies useful to deploy "European" HPC ExaScale systems
- We are hardly working on it: ExaNeSt (to study distributed storage and network for ExaScale system); EuroExa (to build the next generation pre-ExaScale prototype)
- Up to now, some delays and few achievements...
- ... but preliminary results and activities ramp-up are really encouraging and synergic with INFN HPC core business
- A new initiative has started to support the introduction of many-core architectures for HPC/HTC computing in INFN (see next slides...)

Progetto CIPE e iniziativa di acquisizione di sistemi many core "next gen"
per il computing HPC e HTC in ambito INFN

Progetto CIPE

- Nel 2016 finanziamento per il calcolo INFN: il progetto "CIPE".
- Finanziamento per
  - HPC di produzione per i gruppi teorici (sulla base del documento di settembre 2014 attualmente in corso di aggiornamento)
  - update infrastrutture di calcolo degli sperimentali (Tier-xx)
  - sperimentazione di nuove architetture e convergenza delle piattaforme di calcolo (ad esempio come e cosa serve per strutturare il calcolo "opportunistico")
  - overheads vari...
- Ad oggi:
  - finanziate consistente numero di ass.ric. per teorici ($\sim 1ME$)
  - finanziate borse per sperimentali ($\sim 1ME$)
  - in progress il rifinanziamento dell'accordo attivo INFN-CINECA per core-hours aggiuntive su sistema Marconi
  - in progress accordo con CINECA per acquisto e hosting sistemi per HTC basati su partizione A1
- budget CIPE non ancora esaurito....
  - spazio per limitato finanziamento di attivita' di esplorazione "tecnologica" legata al computing HPC e HTC

Attuali sistemi HPC di produzione: uno sguardo d'insieme

- *Multi million cores* supercomputers
  - $10^5 \div 10^6$ processori, $10^2 \div 10^3$ cores per processore
  - efficaci dal punto di vista dei ratio $/Flops e $/Power
  - Alta granularita' e architettura ibrida –> CPU + acceleratori computazionali *many-core*
- Un esempio, a noi "vicino", di approccio ibrido: MARCONI
  - La roadmap d'installazione (Aprile 2016-Luglio2017) prevede:
    - Aprile '16: Cluster basato su PC server; CPU Broadwell E5-2600 v4 (18 cores, 2 proc per node), 2PFlops integrati
    - Fine 2016 (Inizio 2017): Cluster addizionale basato su INTEL PHY (KNL, 70 cores) –> 11 PFlops
    - 2017 (estate): Sky Lakes (architettura server "standard" ma con 20/30 cores) per 5 PFlops addizionali

quindi...

- Ad oggi NON esistono alternative commerciali a scala larga che implementino modelli alternativi a CPU + acceleratori many-cores
- I nostri codici scientifici non sono particolarmente ottimizzati per sfruttare il parallelismo estremo dei sistemi ibridi many-core.
- Simili discorsi valgono anche per il calcolo HTC degli sperimentali per ovvii motivi di opportunità e contenimento dei costi di procurement e operativi.
  - necessita' di scala larga per il computing offline per (ad esempio) gli esperimenti di HL-LHC.
  - "computing esotico": L0-1 trigger on-line,...
- La frazione maggiore di PFlops ottenibili dai sistemi HPC correnti e futuri viene, e verra', dal computing sugli acceleratori many-cores

–> dobbiamo creare le condizioni per un loro **utilizzo efficente**

Questa iniziativa...

- ...vuole lanciare un'attivita' di alfabetizzazione/discovery di queste architetture di calcolo orientatata ai giovani fisici dei gruppi computazionali dell'INFN (teorico e sperimentali) e ai (giovani..) tecnologi per imparare a
  - valutare le necessita' computazionale e la complessita' dei problemi di calcolo,
  - effettuare il loro porting efficiente sui sistemi many core
  - gestire l'hosting ed il supporto sistemisitico di piattaforme a scala larga
- basare questa attivita' sul procurement di uno o piu' sistemi di calcolo di taglia piccola/media, NON di produzione, da
  - installare in casa;
  - composto da componenti che (possiamo aspettarci) diventino il mainstream dei sistemi HPC del prossimo futuro.
  - equipaggiato da tutto il software e dalle librerie necessarie (e opzionali);
- Il tutto gestito (almeno inizialmente...) da un comitato ristretto
  - teorici: Biferale, Cosmai, Pepe;
  - sperimentali: Boccali, De Salvo;
  - esperti di tecnologia ed infrastrutture: Maron, Schifano, Vicini

Attivita' iniziale

- Il comitato di gestione dovra' produrre, in tempi molto brevi,
  - una survey sulla tecnologia corrente e futura
  - una selezione dei codici di nostro interesse da usare come benchmark per questa attivita' (non e' detto che sia efficiente investire nel porting di TUTTI i nostri codici)
  - una collezione dei requirements algoritmici e computazionali di tali codici ed una survey dei tools software necessari (compilatori, librerie, eventuali framework di supporto alla programmazione parallela)
- Sulla base di questa analisi preliminare
  - valutazione dei costi di procurement di sistema (attraverso interazione con fornitori)
  - realizzazione del capitolato e supporto alla procedura di gara
  - individuazione del sito d'installazione

- Individuati alcuni codici di interesse su cui fondare le attivita' di benchmarking di sistema
  - LQCD (MILC, OpenQCD), Lattice Boltzmann (LBM, Codice pseudo-spettrale), 3DFFT fast, Astro, Complex systems, Bio (???)
  - HEP: in progress, frameworks piuttosto che singoli codici computazionali monolitici
- Iniziata attivita' di technology survey con obiettivo di acquisire il(i) sistema(i) per la fine del 2017
  - Caveat: procurement compatibile con la disponibilita' dei sistemi scelti, non in overlap con quanto gia' nelle nostre disponibilita' (es. CINECA)
  - Incontrati per adesso Intel, NVidia, IBM:
    - a 6+ mesi da oggi nessuna particolare novita'
    - Xeon+FPGA, Power9 con NVLink2.0 integrato, sistemi JBOD basati su Volta
  - esplorazione di framework di collaborazione (stile OpenLAB) con i system providers per avere "early access" alle nuove architetture ed ai tools di programmazione correlati, accesso alle roadmaps e/o scontistica sostanziale
  - disseminazione dell'iniziativa per trovare convergenze di obiettivi e sinergie di attivita' con altri gruppi interessati a tale esplorazione

A lungo termine...

Il procurement di macchina e' solo l'inizio dell'attivita'

- serve supporto (i.e. manpower) dai gruppi teorici e sperimentali.
  - sinergie di fatto con i borsisti "CIPE" sperimentali;
  - quadro di collaborazione da costruire con i prossimi assegnisti "CIPE" teorici e (soprattutto) i loro tutor scientifici...
- sul budget di progetto prevediamo il reclutamento di alcuni "giovani" tecnologi
- nei prossimi mesi dovremo completare lo schema delle attivita' post-installazione di macchina (scuole, workshop, seminari specifici e attivita' hands-on con professionisti,...).

–> il successo dell'iniziativa contribuira' a **mantenere ed incrementare il know-how necessario all'utilizzo efficente delle macchine HPC many-core di prossima generazione.**