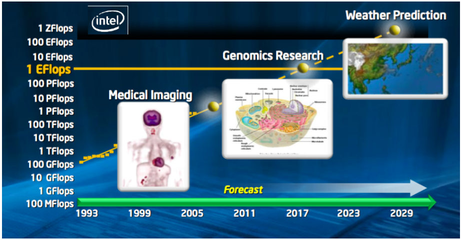
Large scale low power architectures computing system: status of ExaNeSt and EuroExa projects

Piero Vicini
for the ExaNeSt/EuroExa INFN team

ParCo2017 - Bologna - September 13, 2017

Horizon 2020

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.
- HPC for Big Data...

Let's talk of Basic Science...

- High Energy & Nuclear Physics
  - LQCD, Dark-energy and dark matter, Fission/Fusion reactions,...
- 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

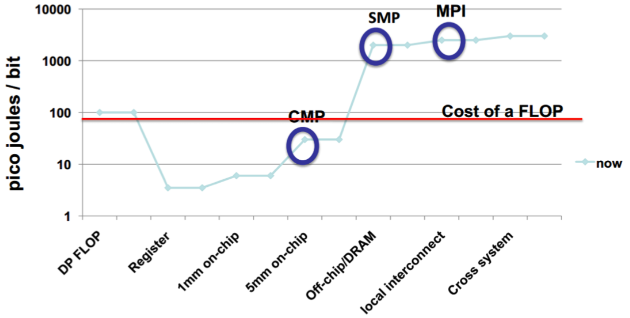
Big numbers, big problems: Power

- General agreement on the fact that data center power budget is less than 20 MW
  - half for cooling -> only 10MW for active electronics
- Current processors performances are
  - multi-core CPU: 1 TFlops/100W
  - GPGPU: 5-10 TFlops/300W but worst sustained/peak (and needs CPU) so only a factor 1.5 better
  - add few tens of watt for distributed storage and memory per node
- ExaScale sustained (where $\epsilon = 50\% - 70\%$)
  - 10^6 computing nodes
  - 100 MW of power –> *low power* approach is needed

Big numbers, big problems: system packing and cooling

- Current computing node assembly:
  - 8 processors into 1U box
  - ~30 1Uboxes per 42U rack (25% of volume dedicated to rack services)
- Summing up
  - 4000 racks per ExaFlops sustained
  - 6000 m^2 of floor space
  - service racks (storage, network infrastructure, power&controls, chillers,...) not included (!!)

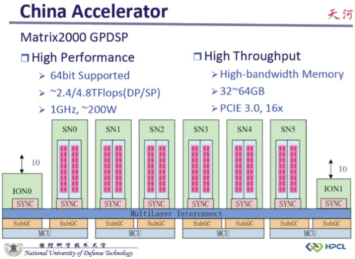
- It needs:
  - New mechanics for denser systems
  - New cooling technology (liquid/gas cooling) for reduce impact of cooling system on power consumption and data center space



- Needed improved hierarchical architectures for memory and storage
  - distributed hierachical memory
  - zero-copy through R(emote)DMA, P(artitioned)G(lobal)A(ddress)S(pace) leveraging on affinity to exploit data locality
- low latency, high bandwidth network

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



What next in Europe?

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?

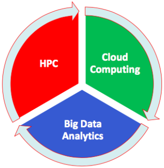
HPC Objectives (1)

- **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



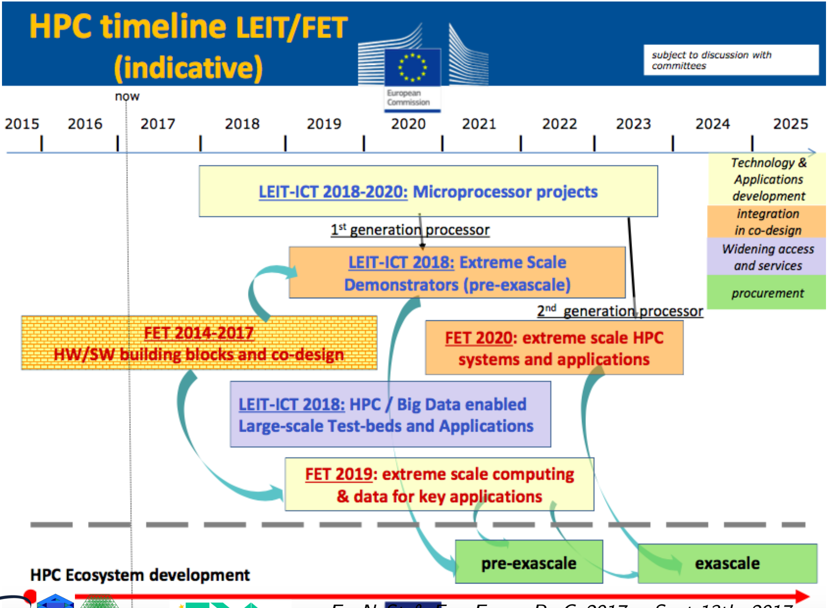
- Major investments so far both at MS and EU level [FP7, H2020]
- Numerous research players (academia and industry)
- HPC and Big Data PPPs, PRACE, GEANT, etc.

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?



An emerging new player in hybrid HPC: FPGA

- More and more different fields of applications thanks to combination of software-like flexibility and hardware perfromances...
- Xilinx Virtex UltraScale+ (Zynq optional), introduction 2017
  - TSMC FinFet 16nm -> 60% less than old generation power consumption
  - 128 transceivers @28Gbps (56Gbps?) for chip-to-chip and backplane interconnection
  - Many industrial standards: HBM (460 GB/s), PCIe gen3(4), DDR4 up to 2666 Mb/s, Ethernet,...
  - 21Tflops of DSP single precision FP
  - Multiple (4->8) ARM Cores (a53/57) @1.5GHz
- Similar in performance: Altera Stratix10

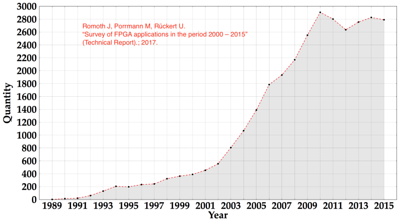


Fig. 1. IEEE listed FPGA related publications per year

Energy–Efficiency

Indicative numbers from various sources

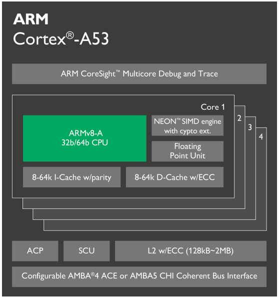
Type	Device	GFLOPS (SP)	Cost (€)	Power (W)	GFLOPS/€	GFLOPS/W
Multi-core	Intel E5-2630v3 8x2.4GHz	600	700	85	0.85	7.05
	Intel E5-2630v3 10x2.3GHz	740	1250	105	0.59	7.04
Many-core	Xeon Phi, knights corner, 16GB	2416	3500	270	0.69	8.94
	Xeon Phi, knights landing, 16GB	7000	3500	300	2.00	23.3
GPU	Nvidia GeForce Titan X	7000	1000	250	7.00	28
	Nvidia Tesla K80	8740	7000	300	1.24	29.13
	Nvidia Tegra X1	512	450	7	19.42	73
	Radeon firepro S9150	5070	3500	235	1.44	21.5
	ARM Mali T880 MP16	374	?	5?	?	74
FPGA	Altera Arria 10	1500	3000	30	1.00	50
	Altera Stratix 10	10000	2000?	125?	5.00?	80
	Xilinx Ultrascale+	4600	2000	40?	2.30	115

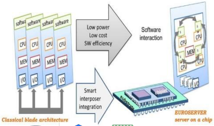
- *NVIDIA/ARM GPU's vs Altera/Xilinx FPGA's*
- Max Performance per Watt may not be the best metric

9

Low power CPU: ARM

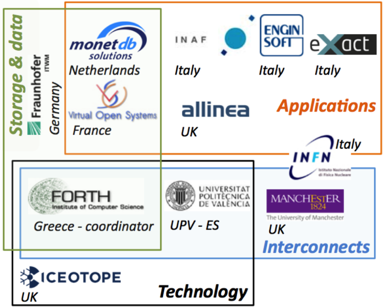
- ARM is (was?) the only "European" CPUs maker
- Innovative business model: ARM sell Intellectual Properties hw/sw instead of physical chip;
  - Pervasive technology: Android and Apple phones and tablets, RaspberryPI, Arduino, set-top box and multimedia, ARM-based uP in FPGA, ...
  - From 1990, *60 billion* of ARM-based chips delivered
- Architecture specialised for embedded/mobile processors
- Few generations of high end (64 bits) processors delivered





ARM low power processors in HPC

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



ExaNeSt: European Exascale System Interconnection Network & Storage

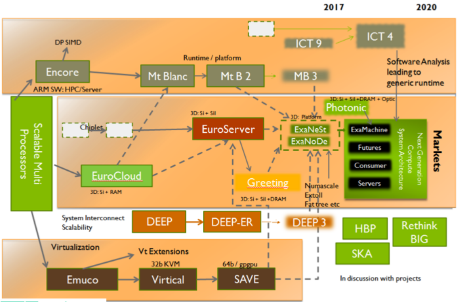
- EU Funded project H2020-FETHPC-1-2014
- Duration: 3 years (2016-2018). Overall budget about 7 MEuro.
- Coordination FORTH (Foundation for Research & Technology, GR)
- 12 Partners in Europe (6 industrial partners)

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

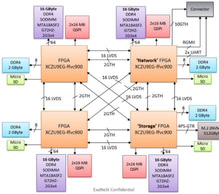
- System architecture for datacentric Exascale-class HPC
  - Fast, distributed in-node non-volatile-memory
  - Storage Low-latency unified Interconnect (compute & storage traffic)
    - RDMA + PGAS to reduce overhead
- Extreme compute-power density
  - Advanced totally-liquid cooling technology
  - Scalable 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

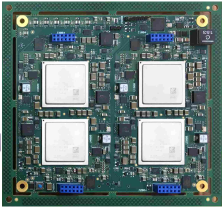
ExaNeSt ecosystem

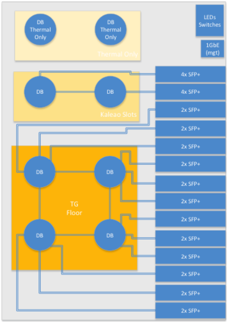
- EuroServer: Green Computing Node for European microservers
  - *UNIMEM* PGAS model among ARM computing nodes
- INFN EURETILE project: *brain inspired* systems and applications
  - APEnet+ network on FPGA + brain simulation (DPSNN) scalable application
- Kaleao: Energy-efficient uServers for Scalable 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







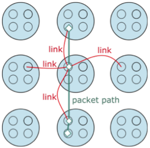
- Computing module based on Xilinx Zynq UltraScale+ FPGA...
  - Quad-core 64-bit ARM A53
  - ∼1 TFLOPS of DSP logic
- ... placed on small Daugther Board (QFDB) with
  - 4 FPGAs, 64 GB DDR4,
  - 0.5-1 TB SSD,
  - 10x 16Gb/s serial links-based I/O per QFDB
- mezzanine(blade) to host 8 (16 in second phase)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

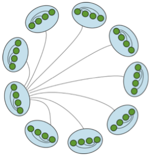
- ExaNeSt high density innovative mechanics...
  - 16 QFDBs per blade
  - 8 blades per chassis (only 6 computing)
  - 5 chassis per rack
  - 20 racks per "HPC container"
  - ExaScale in 30-50 self-consistent (computing, switches, PDUs, cooling) containers
- ...totally liquid cooling
  - track 1: immersed 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...)

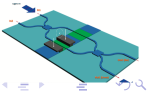
ExaNeSt Interconnect

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

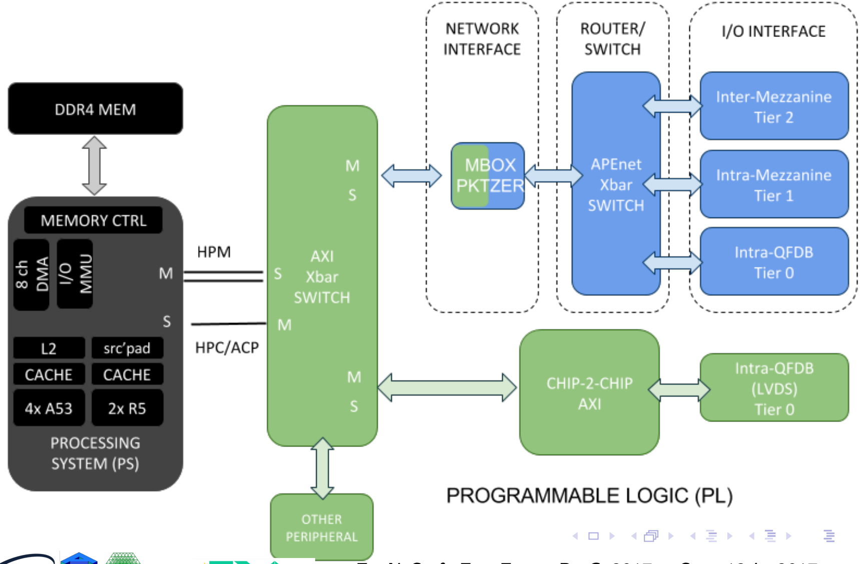
- Unified approach
  - interprocessor and storage traffic on same network medium
  - PGAS architecture and RDMA mechanisms to reduce communication overhead
- innovative routing functions and control flow (congestion managements)
- explore performances of different topologies
  - Direct blade-to-blade 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







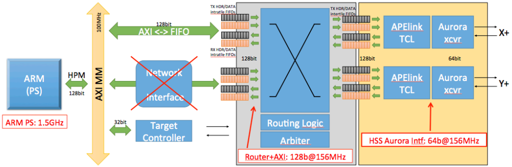
ExaNeSt network architecture at Unit level



ExaNeSt highlights: KARMA testbed

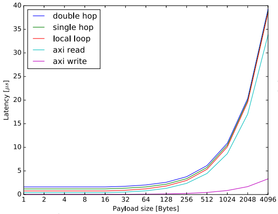
KARMA (*King ARM Architecture*): software testbed for INFN network router

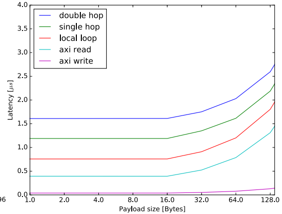
- Router FIFOs connected to ARM HPM AXI ports via an adapter IP
- *Target Controller*: a set of configuration/status registers AXI-readable

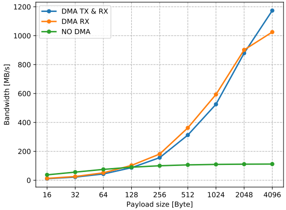


ExaNeSt Trenz cluster in Rome

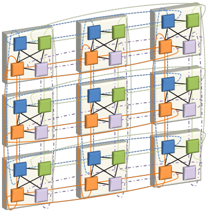
single/double hops; no interrupts, no virt-to-phys add. transl: sub-μS latency

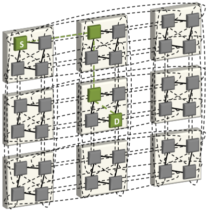




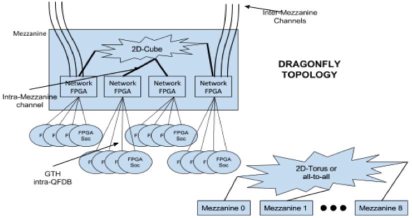


- Inter-blade direct topology
  - 4x 2D-Torus interconnects (3x3)
  - each QFDB of a mezzanine connected with its counterpart on neighboring mezzanine
  - 3 hops for longest path
  - INFN legacy: APEnet+ architecture





- Analysis of multi-level Dragonfly capabilities
  - QFDB –> blade –> system
  - Small diameter
  - few expensive global wires

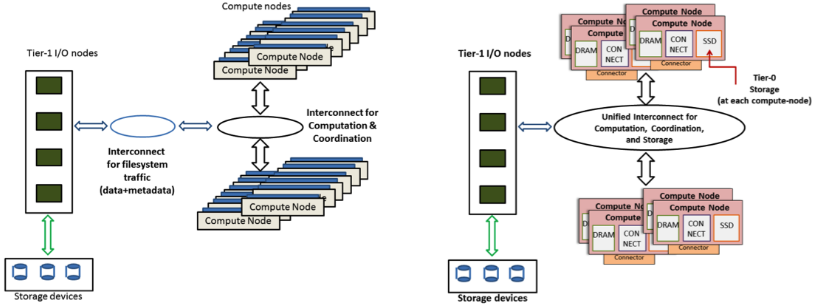


Role of applications in ExaNeSt

Co-design approach

- Applications define quantitative requirements for the system under design
- Applications evaluate the hw/sw system
- List of ExaNeSt applications:
  - Cosmological n-Body and hydrodynamical code(s) (INAF)
    - Large-scale, high-resolution numerical simulations of cosmic structures formation and evolution
  - Brain Simulation: DPSNN (INFN) <– see A. Biagioni talk...
    - 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 Bee**GFS** 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

- **EuroExa**: Co-designed Innovation and System for Resilient Exascale Computing in Europe: From Applications to Silicon
- Work Program Topic: FETHPC-01-2016, RIA
- Coordinator: G. Goumas ICCS (GR)

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

EU Invests Big In Supercomputer Developments for ExaScale

Published on September 6, 2017

Kick-off meeting at BSC (Barcellona) on September 4-5,2017

... 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; ARM 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...*

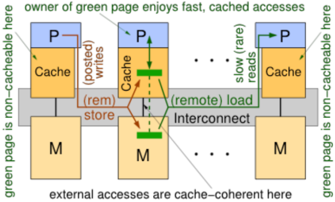
... *a unique hybrid, geographically-addressed, switching and topology interconnect within the rack offering low-latency and high-switching bandwidth...*

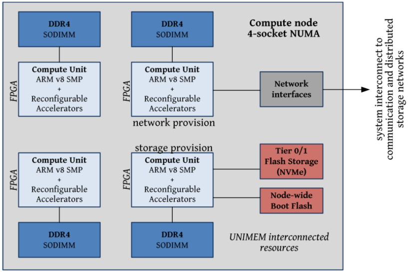
... *a rich mix of key HPC applications from across climate/weather, physics/energy and life-science/bioinformatics domains*

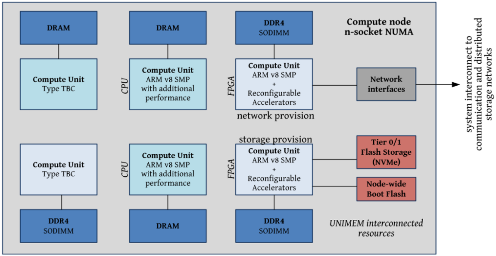
... *deployment of an integrated and operational peta-flop level prototype hosted at STFC, monitored and controlled by advanced runtime capabilities, equipped by platform-wide resilience mechanisms.*

EuroExa (few) details

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

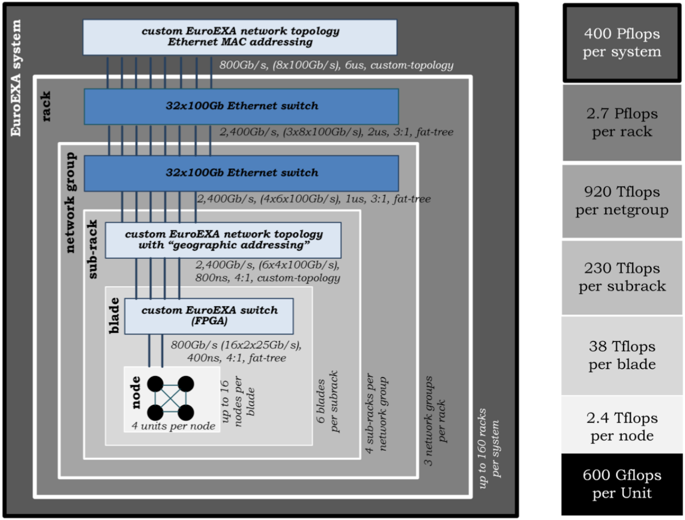






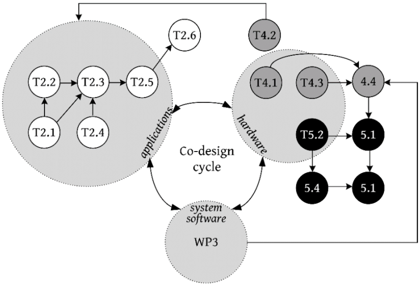
EuroExa (few) details

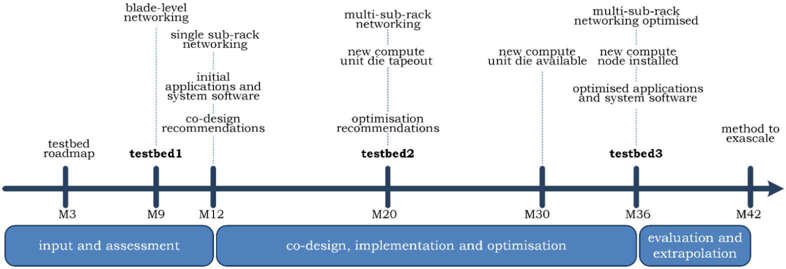
- Balanced, hierarchical network...



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 RM1 and FE mainly in :
  - benchmarking through applications: neural network simulator (RM1, link with HBP projects), LBM simulation (FE)
  - Network design at sub-rack level (RM1)
- INFN budget: 730 kEuro, 3 FTEs for the whole project duration

One slide summary

- HPC has a long and successfull history (mainly not-European...)
- Fundamental scientific and engineering computing problems needs ExaScale computing power
- The race toward ExaScale is started and Europe is trying to compete with established and emerging actors (USA, Japan, China,...) pushing for HPC technologies developments (EuroHPC, EXDCI, IPCEI,...)
- Many challenging issues require huge R&D efforts: power, interconnect, system packing and effective software frameworks
- ExaNeSt and EuroExa will contribute to the evaluation and selection of ExaScale enabling technologies, leveraging on Europe traditional expertise: embedded systems (ARM), excellence in scientific programming, design of non-mainstream network architecture