INFN H2020 projects for new hardware technologies: ExaNeSt & picoLO

P. Vicini – INFN Roma

Workshop CCR

27 Maggio 2015

Introduzione

- Perche' partecipare a call EU?
  - Cogliere opportunita' di finanziamento;
  - Creazione di networking
    - Trasversale, non necessariamente canonico,….;
  - Nuove aree di R&D tecnologico (i.e. non istituzionali per INFN);
  - Generare prospettive di Trasferimento Tecnologico (bidirezionale…);
  - ….

- La nostra (APE e dintorni) esperienza:
  - Presenti da almeno 10 anni in consorzi FP5-FP6-FP7
    - SHAPES, EURETILE,.....
    - Progetti integrati large e medium scale
  - Risultati: know-how tecnologico, papers, IPs, brevetti, formazione…
  - l'unica possibilita' per continuare a fare ricerca con un gruppo di giovani "fully committed" su tecnologie (ai tempi…) non propriamente INFN "core business" ma di prospettiva interessante
    - NaNet → NA62 e KM3Net
    - APEnet+ → Collaborazione industriali (Nvidia, Eurotech,….)
    - DPSNN → Corticonic , HumanBrainProject..

Outlook del talk

- In questo talk, report su due particolari iniziative a diverso grado di maturita':
  - ExaNeSt
    - H2020 FETHPC Call 1 2014
    - Approvato e finanziato
  - picoLO
    - LEIT Research and Innovation Action (RIA) H2020 ICT04 – 2015 (a)
    - Risultati della valutazione attesi per settembre/ottobre 2015

ExaNeSt

EuroExa → ExaNeSt

- Obiettivo iniziale: un GRANDE progetto HPC-oriented per la call FETHPC Nov. 2015
  - Tecnologia "Europea" (!) per la prossima generazione di sistemi HPC scalabili all'Exaflops
  - Grande collaborazione con vocazione prettamente industriale
    - Partner originari: Eurotech (Coordinatore), Nvidia, STM, INFN, UniBo, European top computing centers(BSC, Julich, Cineca,…), Xyratech….
    - Aggiunti in corso d'opera: ARM, Altera, Fraunhofer, Forth,….
- Partenza lanciata ma forse troppo anticipata (kickoff meeting dicembre 2014…)
  - Target iniziale 20 ME nonostante gli 8-10 ME indicati nel bando
- Luglio-Settembre 2014: Eurotech abbandona progetto, Nvidia out, Cineca out…
- Ottobre/Novembre 2014: due (tre) nuovi progetti "gemelli" con taglia ridotta e richieste di budget in linea con la call
  - ExaNode -> Follow up di Euroserver; focus on processore/memoria
  - **ExaNeSt** -> Focus on Storage e Interconnect **← dove siamo noi**
  - EcoScale -> Energy-efficient Heterogeneous COmputing at exaSCALE (essenzialmente software tools)

→ Marzo 2015: Evaluation results: progetti finanziati!!!!!

ExaNeSt

H2020-FETHPC-1-2014 ExaNeSt

Technical Annex

FET-PROACTIVE-TOWARDS EXASCALE HIGH PERFORMANCE COMPUTING

H2020-FETHPC-1-2014

Full title: **European Exascale System Interconnect and Storage**

Acronym: ExaNeSt

Date of preparation: 25.11.2014

Participant no.	Participant organisation names	short name	Country
P1 (Coordinator)	Foundation for Research and Technology – Hellas	FORTH	Greece
P2	Iceotope Research & Development Ltd.	ICE	UK
P3	Allinea Software Ltd.	ALNS	UK
P4	EnginSoft S.p.A.	ES	Italy
P5	eXact Lab Srl	XLAB	Italy
P6	MonetDB Solutions B.V.	MDBS	Netherlands
P7	Virtual Open Systems	VOSYS	France
P8	National Institute for Astrophysics	INAF	Italy
P9	National Institute for Nuclear Physics	INFN	Italy
P10	The University of Manchester	UOM	UK
P11	Univeritat Politecnica de Valencia	UPV	Spain
P12	Fraunhofer Gesellschaft	FHG	Germany

Type of funding scheme: Research and Innovation Actions (RIA), funding rate 100%

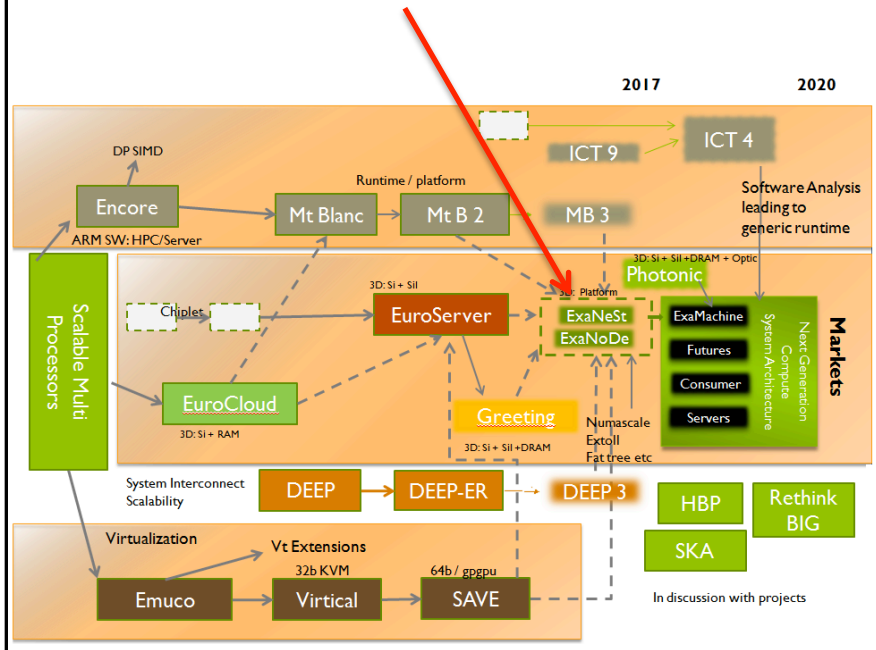
Topic: FETHPC-1-2014 (a): HPC Core Technologies and Architectures

Coordinator: Manolis Katevenis (FORTH)

e-mail: kateveni@ics.forth.gr

Tel.: +30 2811.39.16.64

fax: +30 2811.39.16.61



ExaNeSt abstract (1)

ExaNeSt will develop, evaluate, and prototype the physical platform and architectural solution for a *unified Communication and Storage Interconnect, plus the physical rack and environmental structures required to deliver European Exascale Systems.*

….The consortium brings technology, skills, and knowledge across the entire value chain from computing IP to packaging and system deployment; and from operating systems, storage, and communication to HPC with big data management, algorithms, applications, and frameworks….

→ Consorzio adeguato; know-how che copre l'intera catena tecnologica per progettare (e anche realizzare…) sistemi HPC

…. Building on a decade of advanced R&D, ExaNeSt will deliver the solution that can support exascale deployment in the follow-up industrial commercialization phases….

→ Il deliverable di progetto non e' un sistema HPC completo ma un dimostratore di tecnologie abilitanti di network e storage per sistemi di calcolo all'ExaScale

ExaNeSt abstract (2)

…. *Using direction from the ETP4HPC roadmap and soon-available high-density, high-efficiency compute, we will model, simulate, and validate through prototype, a system with*:

1. *A totally liquid cooled, high throughput, low latency connectivity*, suitable for exascale-level compute, their storage and I/O with congestion mitigation, QoS guarantees, and resilience.

2. Support for *distributed storage located with the compute elements* providing low latency that non-volatile memories require, while reducing energy, complexity, and costs.

3. *Support for task-to-data sw locality models* to ensure minimum data communication energy overheads and property maintenance in databases.

4. *Hyper-density system integration scheme* that will develop a modular, commercial, European-sourced advanced cooling system for exascale in ~200 racks while maintaining reliability and cost of ownership.

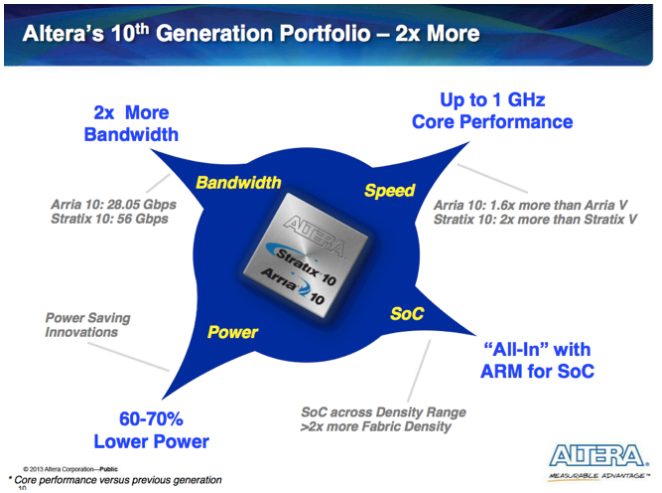
5. *The platform management scheme* for big-data I/O to this resilient, unified distributed storage compute architecture.

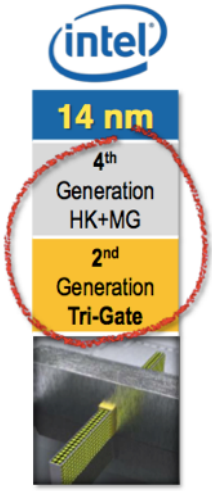
6. Demonstrate the *applicability of the platform* for the complete spectrum of Big Data applications, e.g. from HPC simulations to Business Intelligence support. All aspects will be steered and validated with the first-hand experience of HPC applications and experts, through kernel tuning and subsequent data management and application analysis.

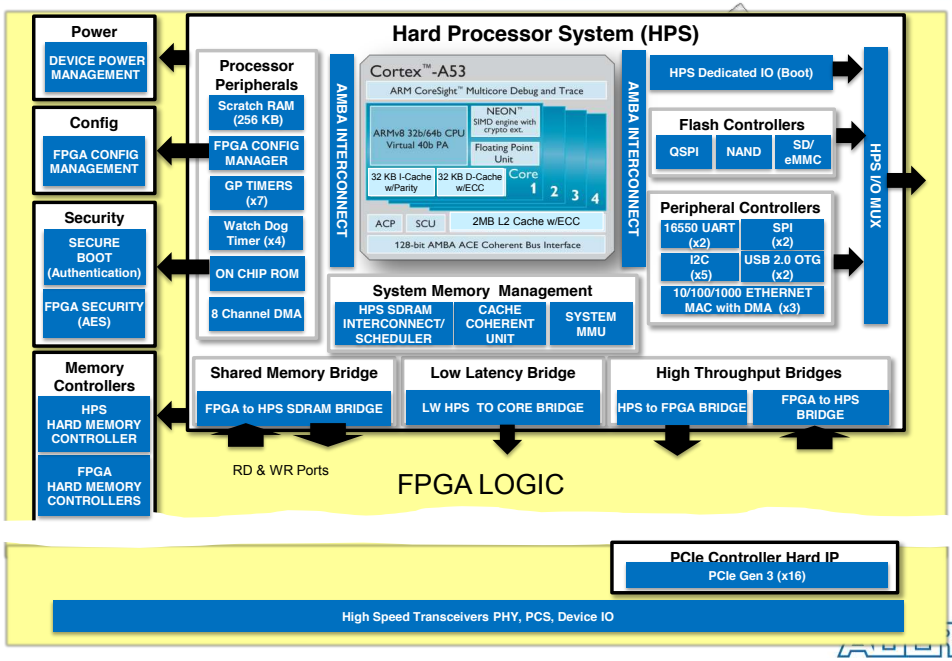
FPGA in ExaNeSt

In ExaNeSt il processore target sara' un SoC di ultima generazione: Altera Generation 10 FPGA

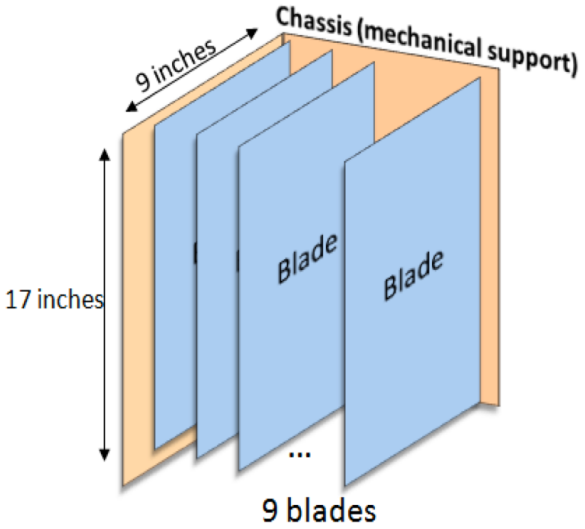
- Arria10 mid range (20nm) (from 2014)
- Stratix10 high-end
  - Introduction 2015
  - INTEL TriGate 14nm -> 70% of StratixV power consumption
  - 96 transceivers @32Gbps (56Gbps?) for chip-to-chip interconnection and @28Gbps for backplane/cable interconnection
  - Many industrial standards supported
    - included CAUI-x (Nvlink compatible)
    - PCIe Gen3(4)
    - 40/100GEth
  - tons of programmable logic @1GHz
  - ...and "for free"...
    - Support to HMC
    - 10 Tflops of DSP single precision FP performances
    - Multiple (4->8) ARM Cores (a53) @1.5GHz

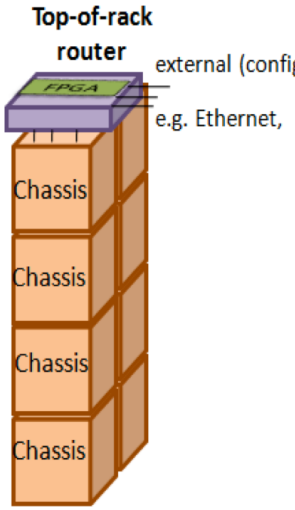


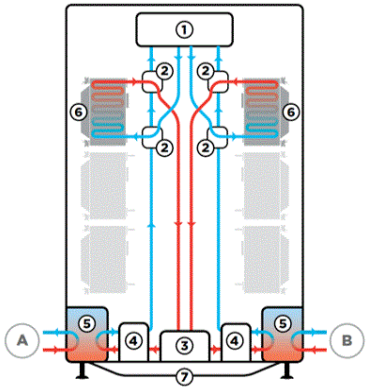




ExaNeSt: architettura, network e meccanica di sistema





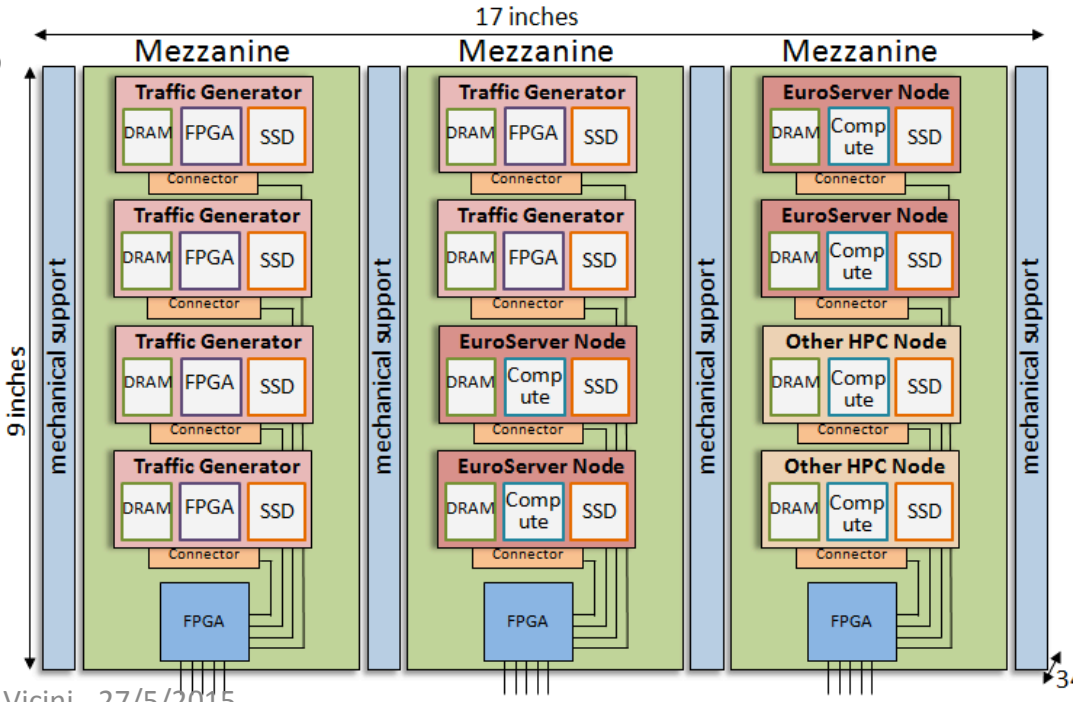


- Low Pressure System
- Tier 4 Ready Design
- 2N Coolant to the Cabinet
- Hot Swap Clean and Dry
- Water to the Cabinet: 45C In 50C Out

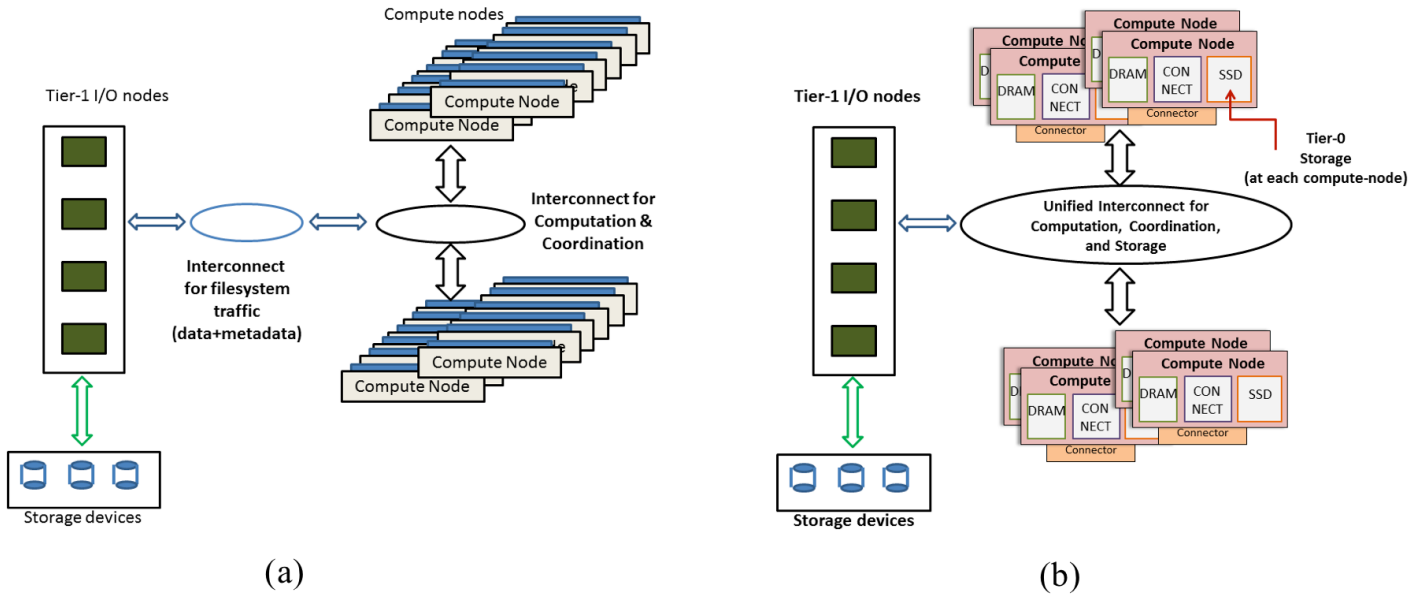
9 blades x 8 chassis = 72 blades

ExaNest assembla un rack completo interamente popolato e "fully functional"

- Blade: N(3) mezzanine card che ospitano
  - FPGA SoC: "Traffic Generator Module"
  - Nodi di calcolo (Euroserver, Exanode) con la propria memoria e lo storage locale appena disponibili…
  - High-end FPGA per network di comunicazione
- 9 blades/chassis 8 chassis/rack
- ToR configurabile per scalabilita' all'ExaFlops e high performance I/O
  - Testbed per link ottici
- Nuova meccanica di sistema totalmente liquid cooled



ExaNeSt Storage…



- ExaNest sara' Linux based ma con storage distribuito: NVM vicino al nodo computazionale (low latency and low power access to data)
- Infrastruttura di interconnessione unificata per storage e computing data network con architettura dello storage HPC realizzata come estensione dei parallel file systems esistenti open-source tipo FhGFS/BeeGFS.
- "*Highly optimized I/O path in the Linux kernel beyond the state of the art*"

ExaNeSt: benchmarking through real applications

- ExaNeSt realizzera' un sistema di esplorazione architetturale e valutazione di prestazioni sulla base di applicazioni challenging per i futuri sistemi HPCBlade,
- Co-design di sistema
  - Prima fase: benchmarks sintetici che emulano traffico di rete e di storage a partire dalle applicazioni reali
  - Seconda fase: re-engineering delle applicazioni attraverso ottimizzazione degli algoritmi di data distribution, data communication e storage.
- Deliverables finali: esecuzione delle applicazioni ottimizzate sulla piattaforma ExaNeSt (forse anche produzione?)

Application	Programming model	Storage/communication characteristics	Scale: current / target	Bottleneck
N-Body	MPI + OpenMP	100+ TBs/sim; Bandwidth intensive at checkpoint/restart; latency-sensitive compute commun; non nearest neighbor as particles move.	up to 100B particles / 100-1000x needed	Interconnect and storage /data.
Hydrodyn	MPI + OpenMP	100+ hundreds TB; non nearest neighbor communication	Multi-scale simulations / 10-100x better resolution needed	Memory size & system, (low-latency) interconnect
Brain simulations: Distributed Polychronous Spiking Neural Net with Synaptic Time Dependent Plasticity (DPSNN-STDP)	MPI + DOL (Distributed Operated Layer)	Currently mostly neighbor communication; runs on commodity clusters & APEnet+.	Several 100s cores / at least 1000x in next five years	Interconnect for higher-resolution and multi-scale/region simulations
Lattice QCD	MPI, SIMD	8 Flops per byte; nearest (16-32) neighbor; embarrassingly parallel. Used as benchmark for all classes of HPC systems	Scales as you add nodes.	Compute bottlenecks; number of nodes in system.
Material Science	MPI + OpenMP	Nearest neighbor communication	1K-10K atoms/ more than 100x needed	Placing tasks close to data
Weather/climate	MPI	Data parallel application; 15 TB per sim; several interacting sims run in parallel.	400x800x20 levels per sim; 2-3 concurrent sims / need 2x	Storage / data bottleneck
Data analytics (scientific & statistical DBs)	Column-store DB	Irregular communication, data-intensive, latency-critical	Current 100G DB / Ddesired 10TB; more interactive.	Compute, interconnect and storage

ExaNeSt: struttura del progetto

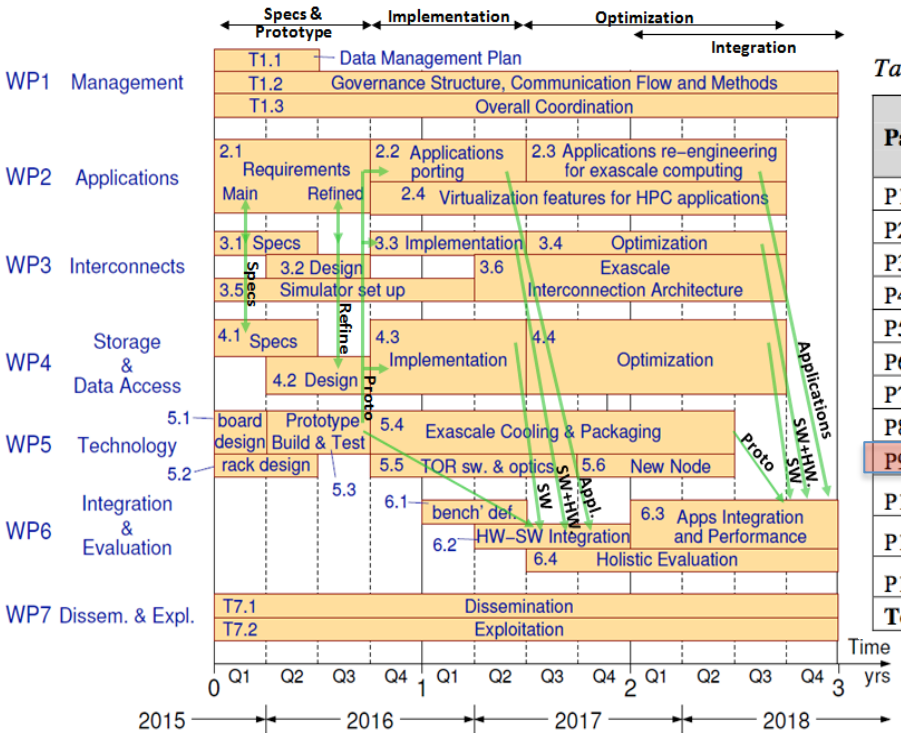


Table 7 - Summary of staff effort

Partner	WP1	WP2	WP3	WP4	WP5	WP6	WP7	Tot. Person/ Months per Partner
P1 FORTH	12	3	62	40	46	9	3	175
P2 ICE	2	0	0	0	108	24	6	140
P3 ALNS	2	24	0	0	0	4	2	32
P4 ES	8	20	2	0	0	33	2	65
P5 XLAB	1	20	6	0	0	0	2	29
P6 MDBS	3	36	3	22	3	6	3	76
P7 VOSYS	0	27	0	29	0	4	6	66
P8 INAF	4	72	0	0	5	12	6	99
P9 INFN	0	19	71	23	0	23	4	140
P10 UOM	4.2	3	41.3	0	16	15.5	2.2	82.2
P11 UPV	2	0	54	0	10	4	4	74
P12 FHG	1	0	0	32	0	3	1	37
Tot. Person/Months	**39.2**	**224**	**239.3**	**146**	**188**	**137.5**	**41.2**	**1015.2**

- Starting date: 1 Dicembre 2015
- Durata: 36 mesi
- Total budget: 8.442.548 Euro (di cui ~2ME per procurement hardware)
- INFN (Roma1 + CNAF) principalmente in
  - WP2 (applicazioni), WP3 (network design), WP4 (storage) e WP6 (integrazione di sistema)
  - Previsti 140 MM ~ 4FTE per l'intera durata del progetto
  - INFN Budget: 770 KEuro

picoLO

LEIT Call ICT04-2015

- LEIT (Leadership in enabling and industrial technologies) Research and Innovation Action (RIA) H2020 ICT04 – 2015 (a)
  - **Next generation servers, micro-server and highly parallel embedded computing systems based on ultra-low power architectures**:
    - *The target is highly performing low-power low-cost micro-servers, using cutting-edge technologies like, for example, optical interconnects, 3D integrated system on chip, innovative power management, which can be deployed across the full spectrum of home, embedded, and business applications. Focus is on integration of hardware and software components into fully working prototypes and including validation under real-life workloads from various application areas. Specific emphasis is given on low-power, low-cost, high-density, secure, reliable, scalable small form-factor datacentres ("datacentre-in-a-box").*
  - **New cross-layer programming approaches**
    - *empowering developers to effectively master and exploit the full potential of the next generations of computing systems based on heterogeneous parallel architectures and constituting the computing continuum. Beyond performance, optimisation should include energy efficiency, time-criticality, dependability, data movement, security and cost-effectiveness. Research should also aim at radically increasing the productivity in programming and maintaining intrinsically parallel code by marginalising the need for dual expertise - application engineering and computer system engineering. Focus is on holistic approaches hiding the complexity between the computing HW component level and the level of application families.*
- Chiusa il 14-04-2015.
- Risultati della valutazione attesi per fine settembre 2015.

picoLO: obiettivi

- In linea con quanto richiesto nella Call…

  - An hardware and software integration effort that will realize a *working, customized, heterogeneous, low -power computing prototype* based on embedded European technology

  - Targets *highly performing low-power low-cost micro-servers* which can be deployed across the full spectrum of home, embedded, and business applications.

  - Aims to target the growing diffusion of *real-time applications for cyber physical systems, IoT, automotive, medical and diagnostic instruments*, etc. requiring high and flexible computing in a compact, portable, low--power and scalable fashion

Research and Innovation Action
ICT-04-2015 - Customised and low power computing

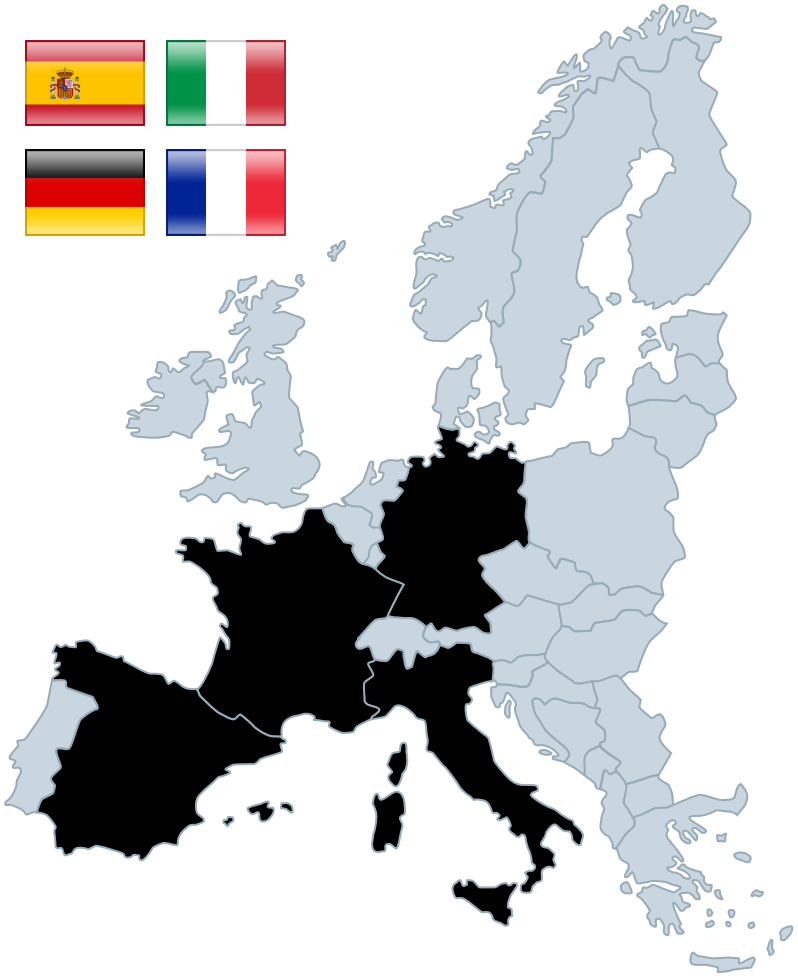
Energy efficient, cost effective, scalable, heterogeneous computing platform and productive software environment for advanced embedded applications

picoLO

List of participants

Participant N.	Participant organisation name	Participant short name	Country
1	Barcelona Supercomputing Center	BSC	ES
2	Siemens AG	SIE	DE
3	Istituto Nazionale di Fisica Nucleare	INFN	IT
4	UJF-TIMA	UJF-TIMA	FR
5	SECO	SECO	IT
6	E4	E4	IT
7	Jülich Supercomputing Centre	JUELICH	DE
8	Leibniz Universität Hannover	LUH	DE
9	FEV GmbH	FEV	DE

picoLO team



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

SIEMENS

INFN

iima

FEV

SECO

E4 COMPUTER ENGINEERING

JÜLICH

picoLO: la piattaforma di calcolo

Architettura di calcolo **embedded scalabile** caratterizzata da valori ottimali dei rapporti FLOPS/W and FLOPS/€

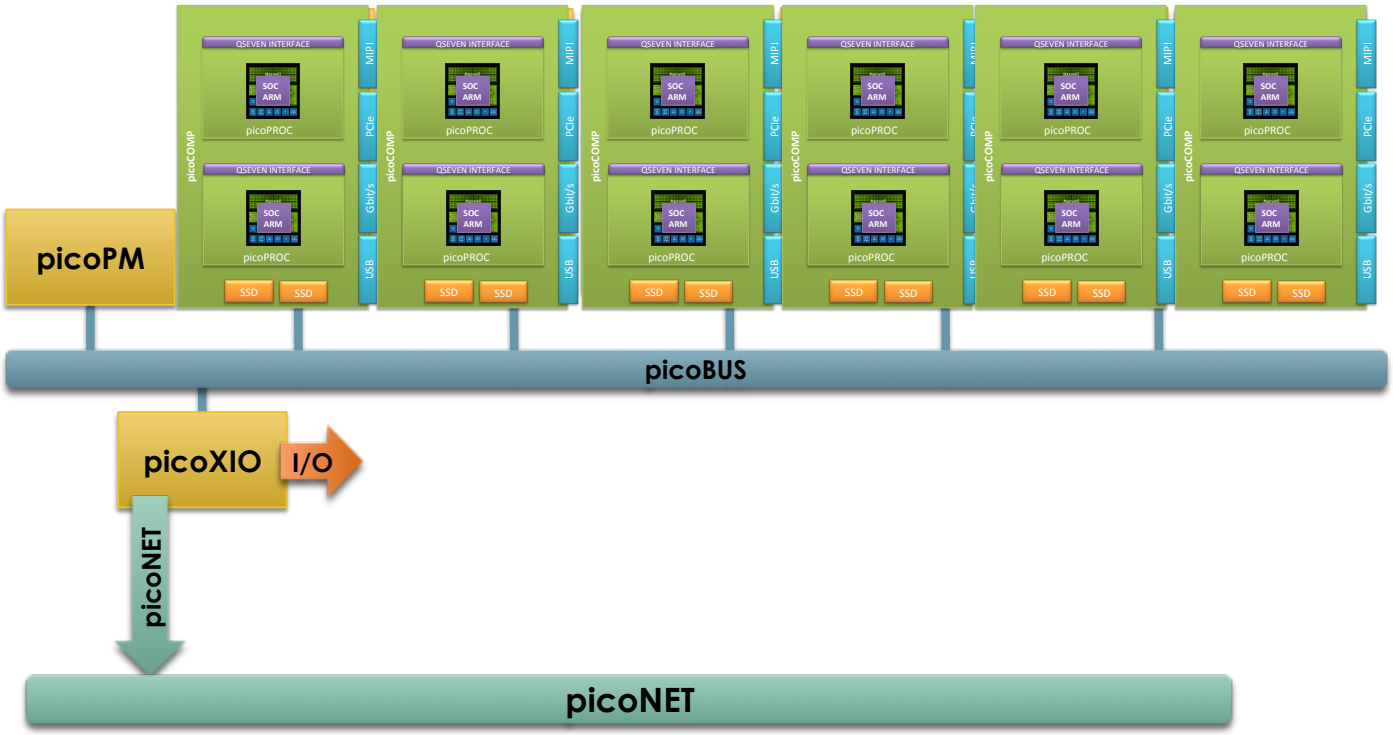
- La picoCARD rappresenta il modulo elementare attorno a cui è costruita l'architettura. Due tipi:
  - picoCOMP, implementano le capacità di calcolo, storage e I/O locali;
  - picoXIO, implementano la connettività a bassa latenza e l'I/O di sistema (vedi oltre).

- picoCOMP è una carrier board che alloggia un certo numero (ad es. 2) di Computer on Module (picoPROC) (standard da definire, ad es. Qseven). Due tipi:
  - picoPROC-1: SoC multi-core ARM con acceleratore many-core (ad es. NVidia Tegra X1).
  - picoPROC-2: FPGA SoC device (ad es. ALTERA Arria-10).

picoPROCs in picoCOMP

picoBOX

Un certo numero (max 6) di picoCOMP possono essere assemblate assieme ad una board picoXIO in modo da costruire un sistema picoBOX ottimale per l'applicazione di interesse.

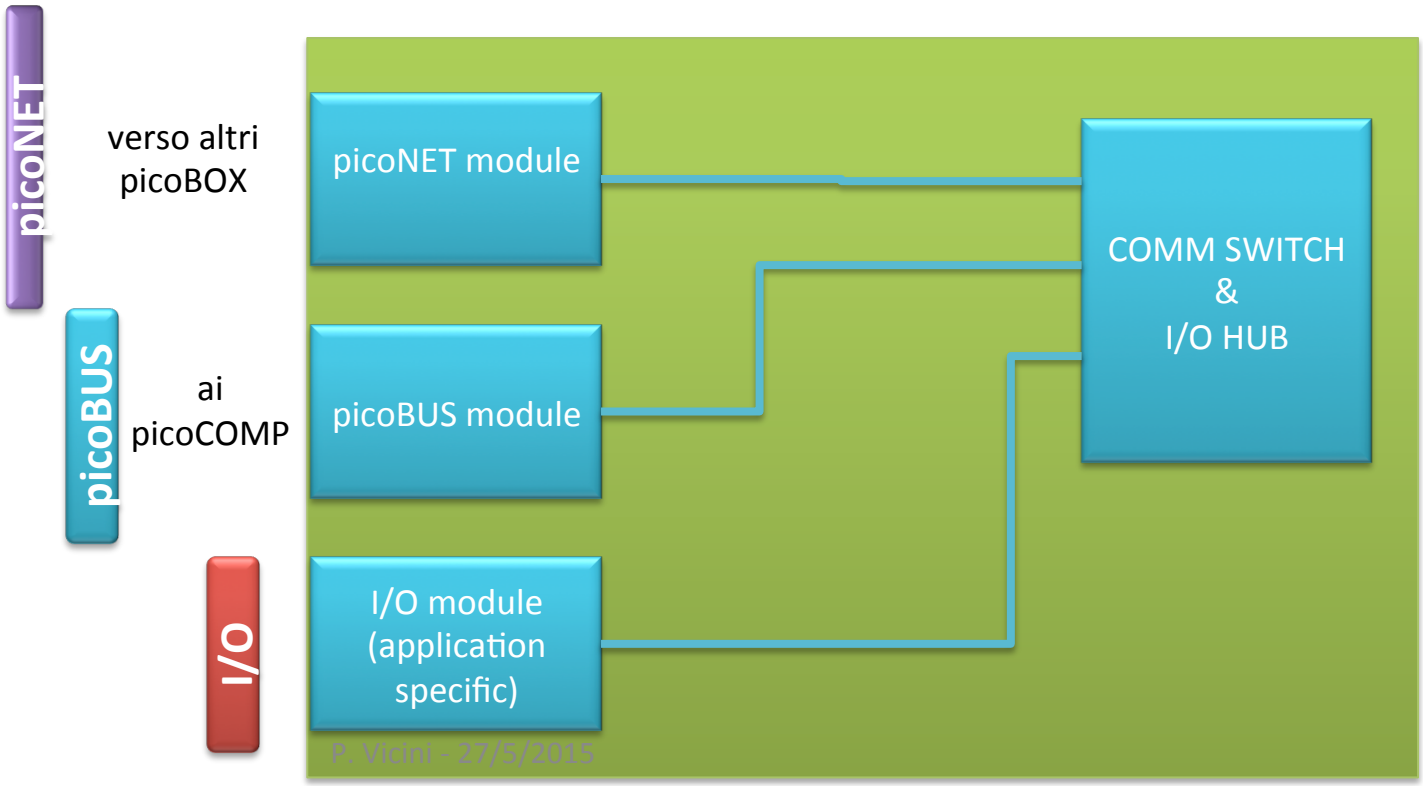


Il modulo **picoPM** implementa il power supply ed i servizi di monitoring e management remoti per i componenti del picoBOX.

picoXIO

La picoXIO card implementa

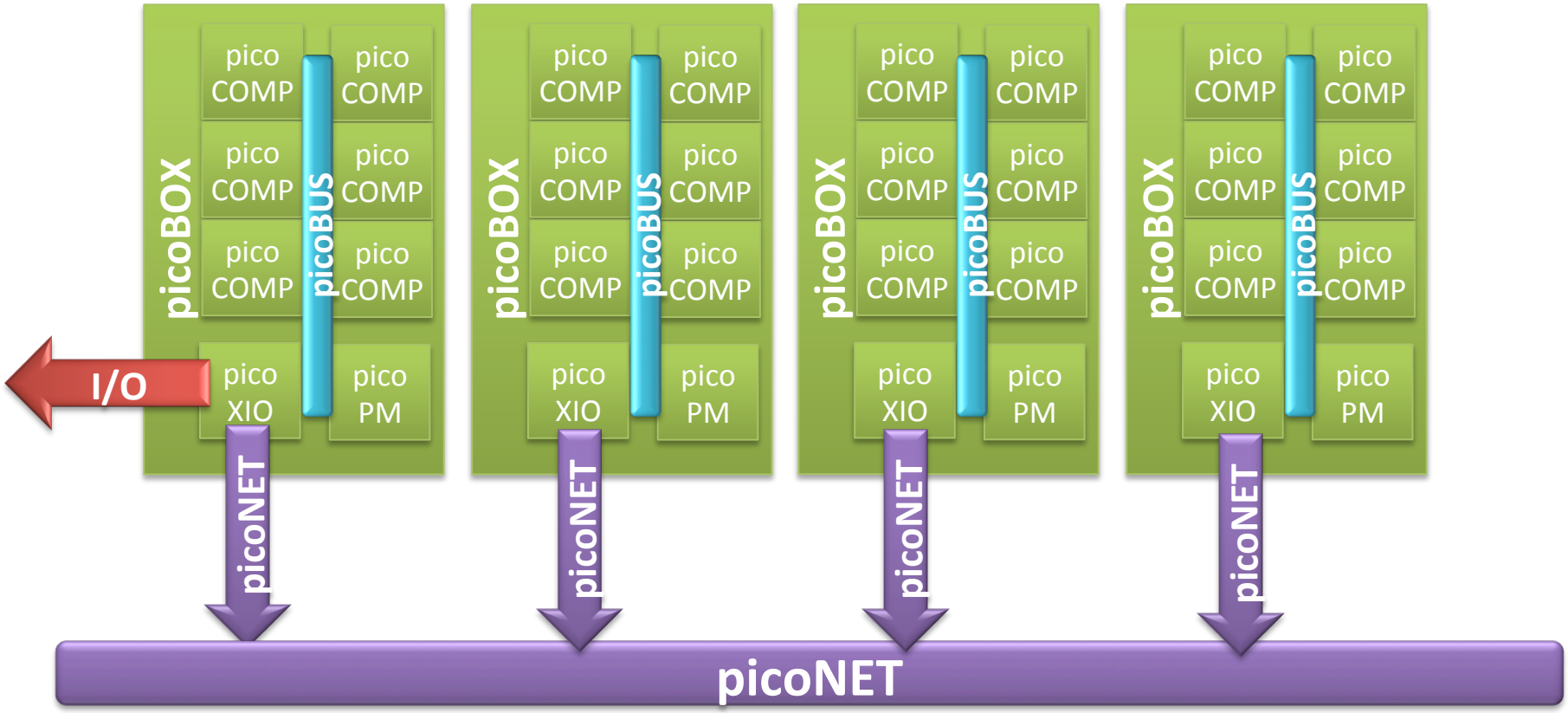
1. le comunicazioni a bassa latenza tra picoPROC che appartengono alla stessa picoBOX e con le omologhe picoXIO che appartengono a picoBOX diverse.
2. Un certo numero di canali di I/O standard (ad es. 10GbE) condivisi dai picoPROC della picoBOX.
3. Uno o più socket per aggiungere funzionalità di I/O specifiche per la applicazione, tramite una daughter card di espansione.



picoLO cluster

picoLO e' un sistema scalabile

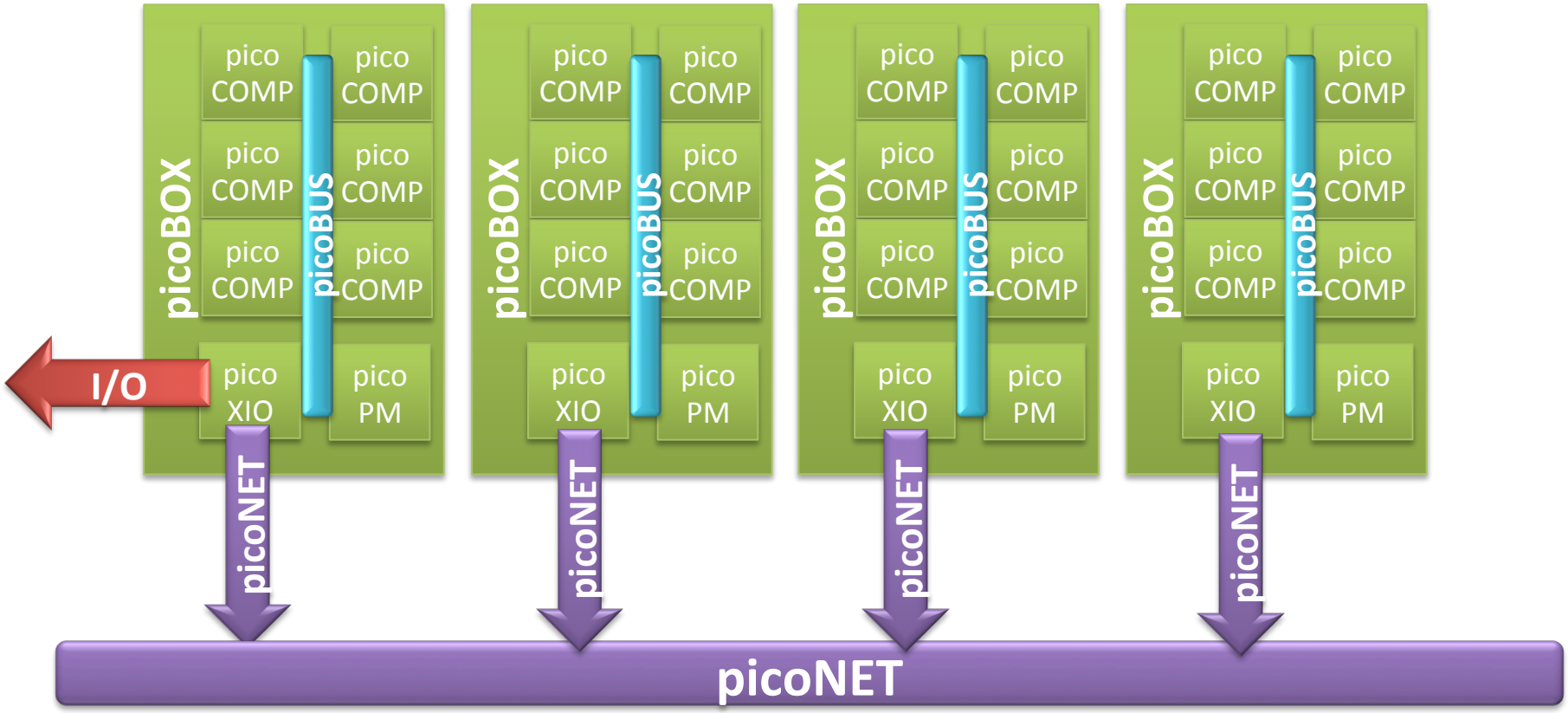
- diversi picoBOX possono essere collegati dalla rete picoNET per costruire un cluster (**picoSTACK**).



picoLO cluster

picoLO e' un sistema scalabile

- diversi picoBOX possono essere collegati dalla rete picoNET per costruire un cluster (**picoSTACK**).

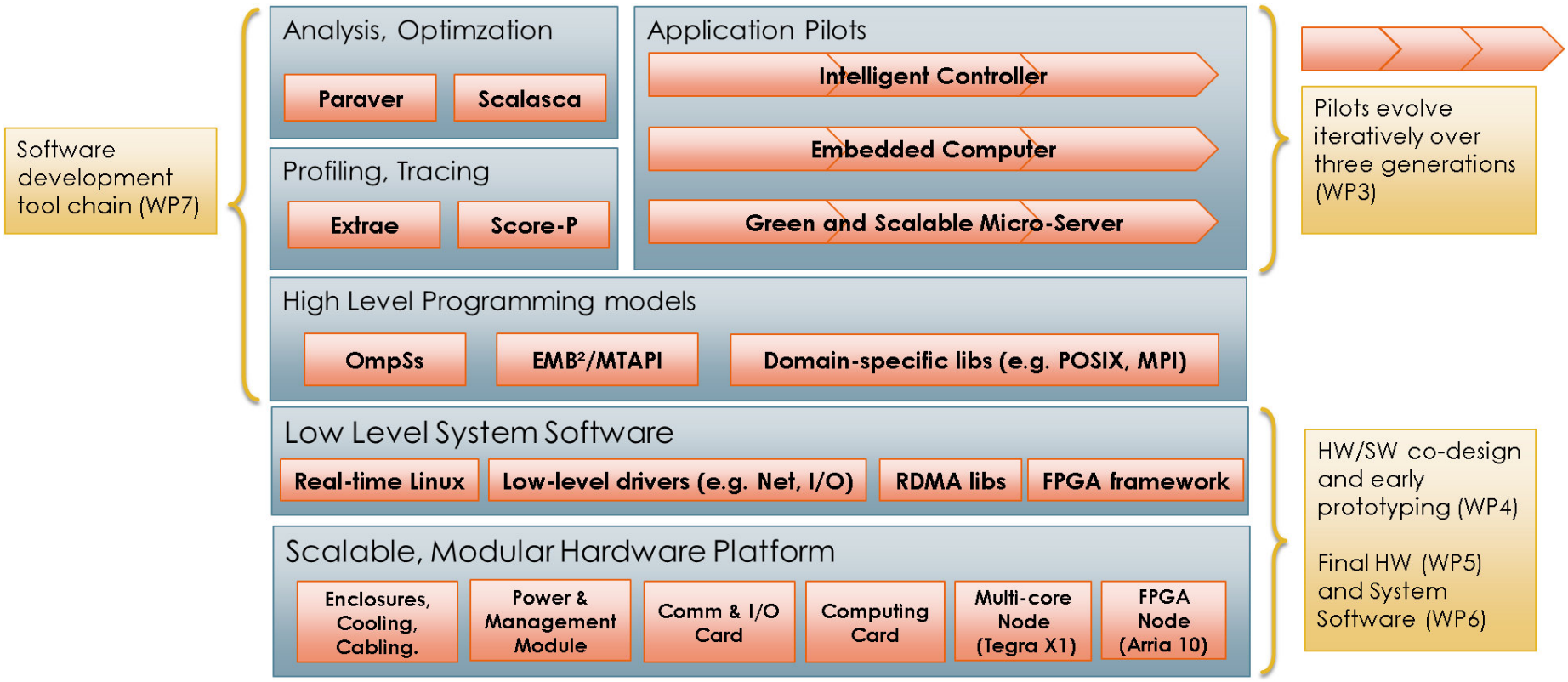


picoLO benchmarks

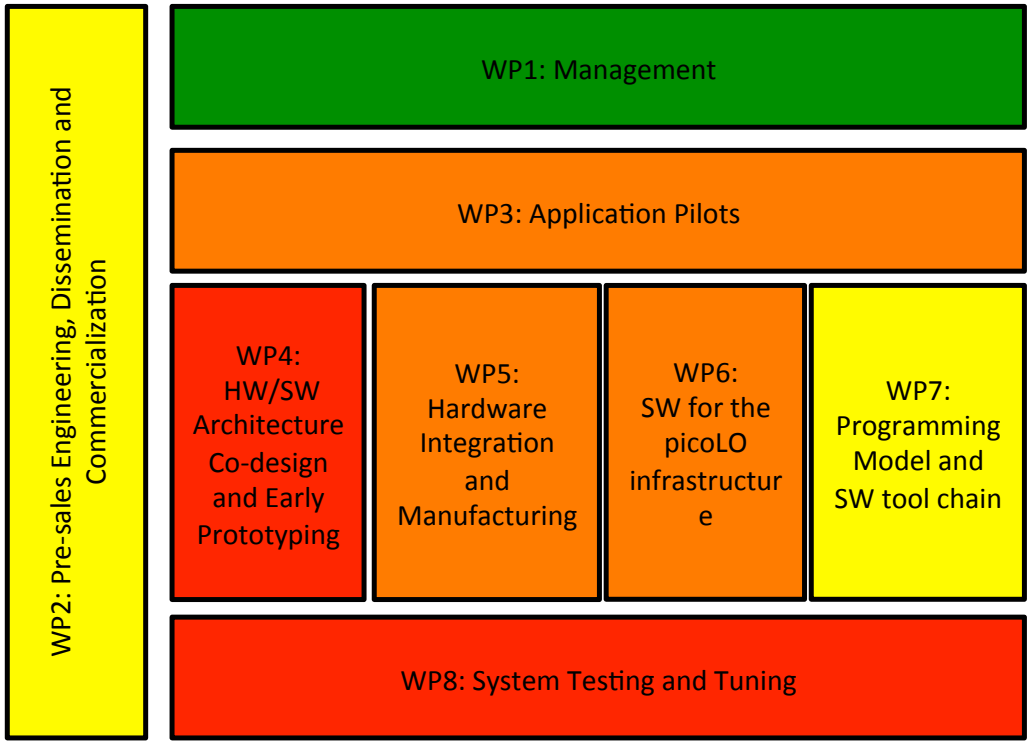
Application pilot che individuano tre diverse tipologie di use case della piattaforma:

- Intelligent CONTROLLER
  - Vision-based automatic train operation
  - Autonomous longitudinal driving cars
  - High Energy Physics low-level trigger.
- Embedded COMPUTER
  - Medical CT image reconstruction
  - Automated somatic mutation detection
  - Low power computing for X-Ray Tomography to Cultural Heritage diagnostics
- Green Scalable MICRO-SERVER
  - Large scale unsupervised neural network learning (DPSNN-STDP)

picoLO at a glance…



picoLO: organizzazione del progetto



Durata: 36 Mesi.
Budget totale: 7.8 M€ di cui 1.2 M€ per INFN
Personale INFN: 5-7 unità.
Sezioni coinvolte: Roma, CNAF.
people@Roma:
Alessandro Lonardo (coordinatore INFN), Pier Stanislao Paolucci, PV,…
people@CNAF:
Daniele Cesini, Andrea Ferraro,…

WP1: Management
Leader: BSC, Contributors: All partners

WP2: Pre-sales Engineering, Dissemination and Commercialization
Leader: E4, Contributors: BSC, SIE, INFN, SECO

WP3: Applications
Leader: SIE, Contributors: BSC, INFN, LUH, FEV

WP4: Hardware/Software Architecture Co-design and Early Prototyping
Leader: INFN, Contributors: BSC, UJF-TIMA, SECO, E4

WP5: Hardware Integration and Manufacturing
Leader: SECO, Contributors: INFN, E4

WP6: Software for the picoLO infrastructure
Leader: UJF-TIMA, Contributors: INFN, E4

WP7: Programming model and Software toolchain
Leader: BSC, Contributors: SIE, INFN, UJF-TIMA, JUELICH, LUH

WP8: System Testing and Tuning
Leader: INFN, Contributors: BSC, SIE, UJF-TIMA, SECO, FEV

Conclusioni

- L'expertise ed il know-how accumulato con la partecipazione decennale a progetti FET ci ha permesso di
  - consolidare il nostro patrimonio di conoscenze in ambito tecnologico (comprese IP's e brevetti)
  - contribuire in maniera sostanziale alla redazione di progetti innovativi in ambito tecnologico e assumere ruoli di leadership in questi ultimi
  - finanziare attivita' di ricerca (essenzialmente man-power) rilevanti per le attivita' correnti e future dell'INFN

- Il "circolo e' virtuoso" se esistono sinergie reali tra obiettivi ed attivita' dei progetti FET e interessi/attivita' di ricerca (ad esempio in ambito calcolo INFN)
  - Utilizzo di FPGA e computing node low-power in ExaNest

    → uso di ARM embedded in FPGA per realizzazione di network custom ottimizzate per sistemi di calcolo dedicati alle applicazioni scientifiche (progetto COSA)
  - Sviluppo ed ottimizzazione di canali I/O standard in picoLO

    → interconnessione a bassa latenza con detector HEP e sistemi di low level trigger (progetto NaNet)
  - Sviluppo/ottimizzazione di applicazioni scientifiche challenging

    → LQCD, simulazione di reti neurali,…

- Ad oggi:
  - ExaNeSt in FETHPC approvato. 3 anni a partire da Dicembre 2015. RM1+CNAF team
  - picoLO (crossing fingers…)

- Le attivita' contribuiranno (tra l'altro) allo sviluppo di nuove idee di progetto da sottomettere nelle prossime call 2016-2017…