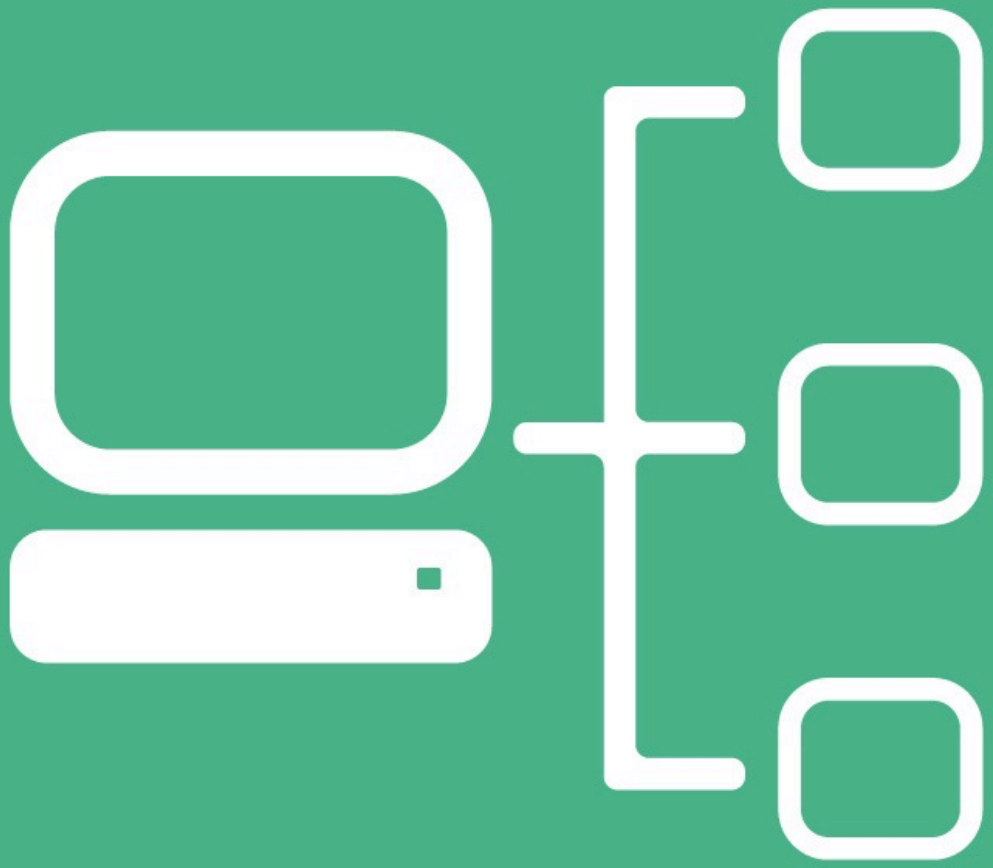
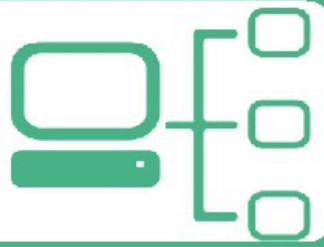


COSA:
Computing on SoC Architecture

Obiettivi



- Acquisizione know-how
  - Porting e benchmarking su System on Chip low power
  - Gestione di sistemi SoC in ambiente Linux
  - Benchmarking di architetture ibride
- Studio di interconnessioni dedicate toroidali a bassa latenza tramite sistemi ARM+FPGA ← *ROMA*
  - Realizzazione di un cluster basato su FPGA che integri le CPU ARM embedded, low power e network basato su protocollo APEnet+
- Aree di interesse applicativo (…benchmark)
  - Area teorica: astrofisica, LQCD, fluido dinamica, dinamica molecolare
  - Area sperimentale: High Level Trigger, Montecarlo, Imaging medico (tomografia assiale)
  - Reti Neurali (DPSNN) ← *ROMA*

I Cluster di COSA

CNAF

~25 **board SoC Based**
(GFLOPS nominali di 2 server tradizionali con GPU) — Anno I

+ ~10 nuove board — Anno II

ex **cluster COKA** (2 server)
+ nuove acquisizioni (1 server) — Anno I

+ nuove acquisizione (1 server) — Anno II

(**server = cpu + acceleratori**)

ROMA1

4 **board ARM+FPGA** based
+ 1 server — Anno I

16 board ARM+FPGA based
+ 4 server — Anno II

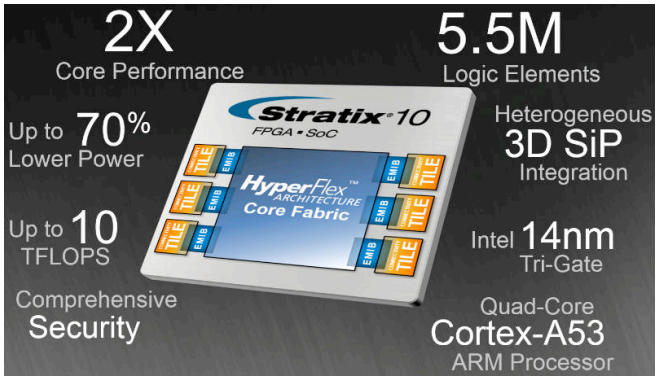
PD

Ex cluster HEPMARK
+ nuove acquisizioni (~ 3 server) — Anno I

+ nuove acquisizione (~ 3 server) — Anno II

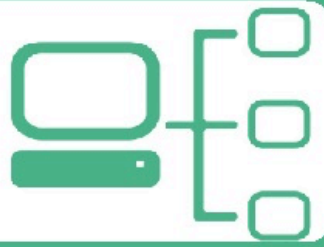
FPGA: scenario attuale

A bit of politics: INTEL acquisisce ALTERA (end 2015)



- In principio, le FPGA ALTERA dovrebbero ricevere un boost
  - INTEL e' fra le principali fonderie di ALTERA (14nm, FINFET)
  - Chiara strategia di lungo termine con integrazione on-die di XEON (o low power CPU) e FPGA (applicazioni data center, big data…)
- Ad oggi: Stratix 10 SoC (primi samples 10/2016)
  - CPU low power (Quad core ARM A53)
  - memoria allo stato dell'arte: in package HBM2 (High Bandwidth Memory), 1Tb/s
  - Nuova architettura di interconnessione interna (HyperFlex)
  - Transceivers 28g -> 56g, N*Tflops, low power
- FPGA SoC nel futuro immediato (???)
  - Omnipath interfacciato direttamente al core FPGA?
  - CPU-FPGA interface su QPI? Risolto il bottleneck PCIe?

FPGA: scenario attuale(2)

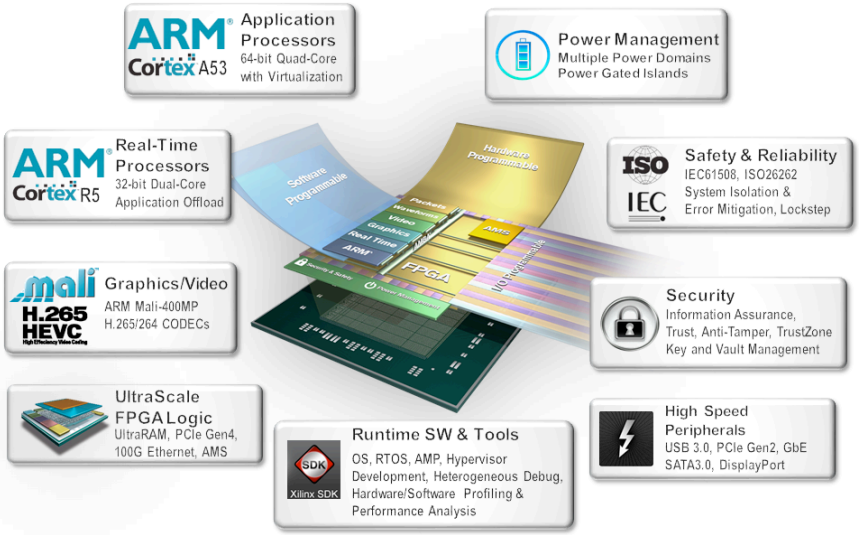


In realta', per utenti indipendenti che ne fanno un uso avanzato si misurano piu' problemi che vantaggi

- Situazione a Roma oggi:
  - FPGA dev kit ARRIA10 SoC (24nm):
    - prevista estate 2015, ordine effettuato settembre 2015, consegna prevista 09/2016 (!!!)
    - Stiamo ancora aspettando…
  - FPGA dev kit SoC STRATIX10 (14nm):
    - sperabilmente 06/2017 ma la storia passata…
- Esiste un'alternativa?

FPGA: scenario attuale(3)

- Xilinx ZYNQ Ultrascale+
- 16nm FINFET, GTH+GTY (100G)
- Quad Core ARM A53 + MALI
  - Coherent Interconnect sub-sys
- Disponibili ora (anche se in ES2) ed ordinabili.



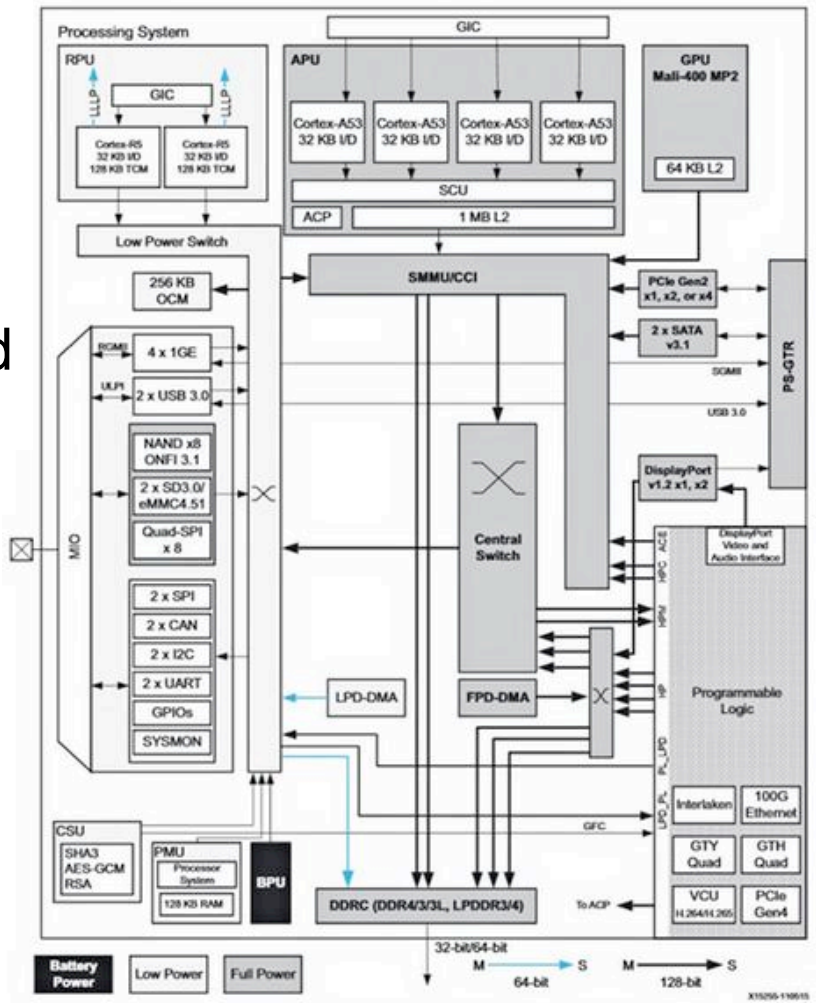
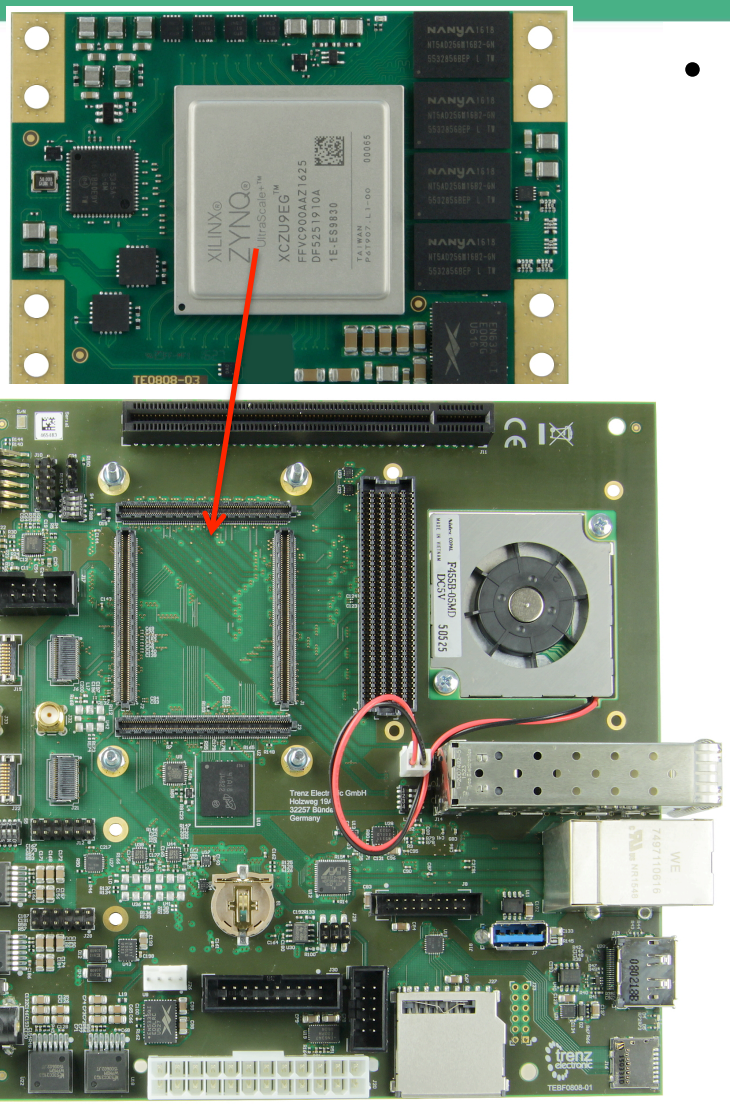


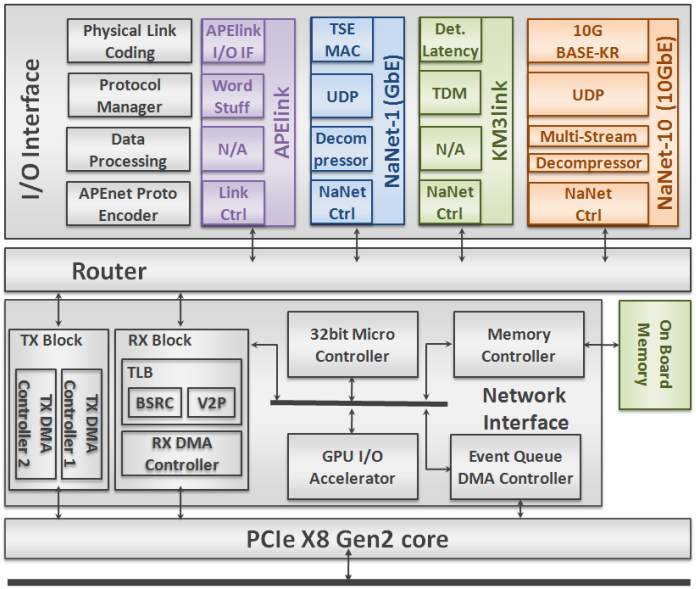
Figure 1-1: **Zynq UltraScale+ MPSoC Top-Level Block Diagram**

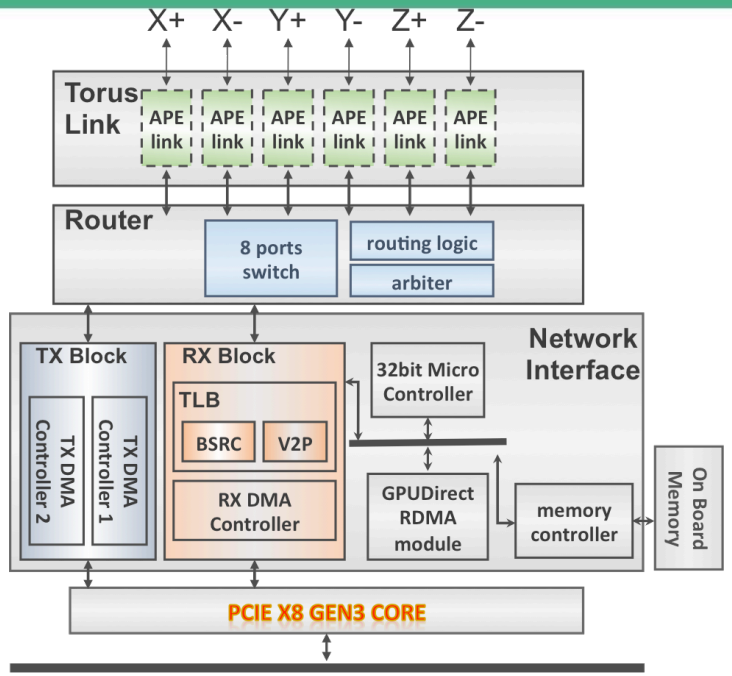
FPGA: scenario attuale(4)



- Cosa abbiamo oggi in lab?
  - 2x Dev kit SoM ZYNQ UltraScale+ (XCZU09) della Trenz (DE)
    - PCI Gen3 x8, 20 GTH@16gbps (2xSFP+, QSFP+ on FMC)
  - Plug and Play (PetaLinux)
  - OpenCL support…
  - 3.8 kE incluso box e cavi

Highlights: APEnet V5(Vx) SW/Driver





APEnet+ V5

Based on Altera Stratix V 28nm FPGA

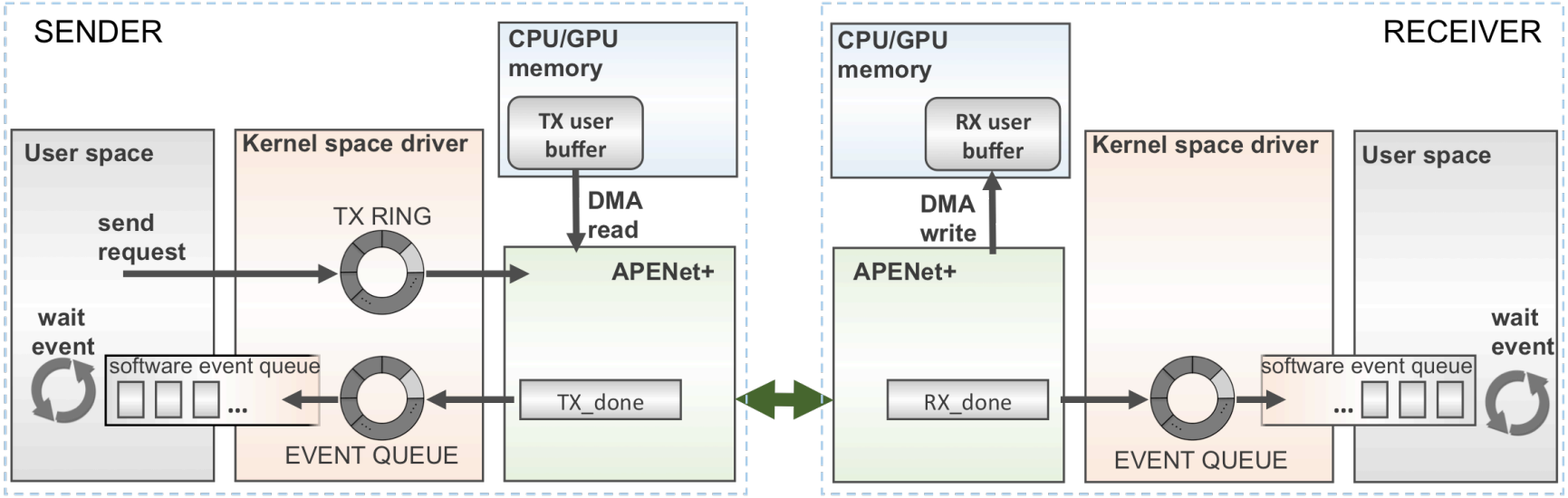
PCIe Gen3 x8 (proprietary PLDA / custom IP core)

Used in:

QUonG: HPC platform built up as a cluster of hybrid elementary computing nodes

M. Martinelli

Highlights: APEnet V5(Vx) SW/Driver



Buffers allocation (pin and lock memory)

On tx side: buffer with data to transfer

On rx side: buffer where data must be stored

Send request -> descriptor with TX physical address and RX virtual address

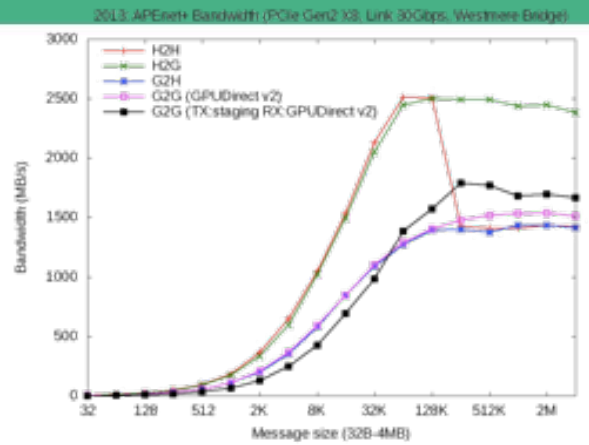
Wait event:

On tx side: wait for the TX_DONE event

On rx side: wait for RX_DONE event

M. Martinelli

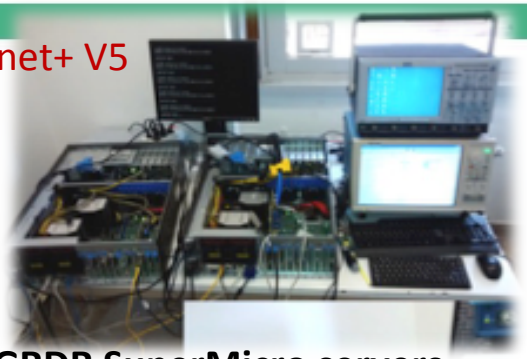
Highlights: APEnet V5(Vx) SW/Driver test



OLD V4!

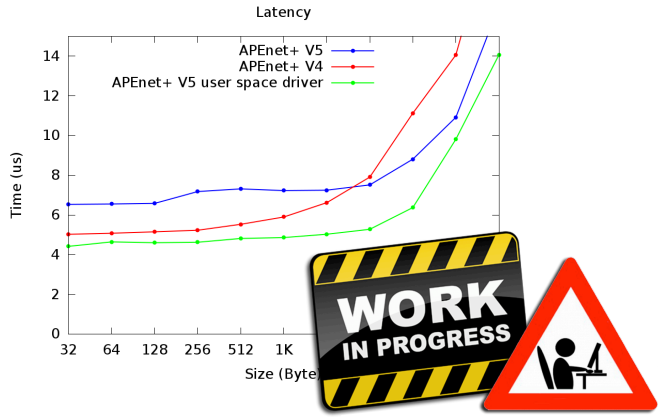
NEW V5

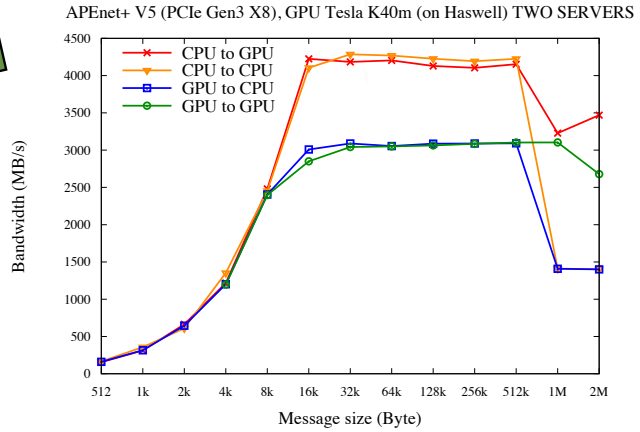
APEnet+ V5



7048GRDR SuperMicro servers

- Dual socket E5-2620 V3





Latency: sending one packet of fixed size and measuring the time required for the "sent" and "receive" events to arrive

Bandwidth: all the packets are sent in a once. Then wait for all the events.

M. Martinelli

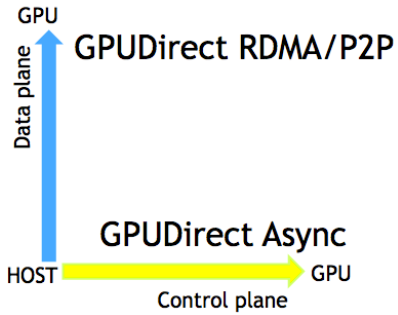
APEnet V5(Vx) SW/Driver: next...

There is a lot of space for the improvement of high performances FPGA-based NIC devices by carefully codesigning the hw and the sw stack.

New idea: providing many-core accelerators with the capabilities of autonomously initiate network data transfer is a relevant shift of paradigm that also boost perfomances, lowering significantly the latency.

GPUDIRECT
scopes

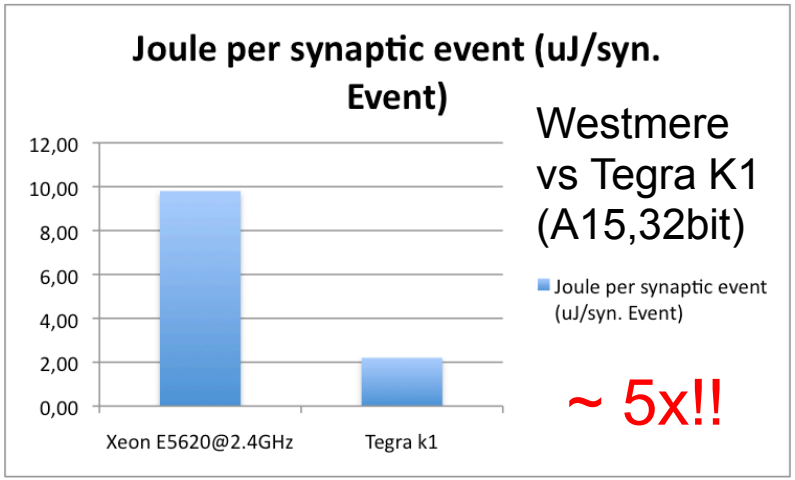
- GPUDirect P2P → data
  - GPUs both master and slave
- GPUDirect RDMA → data
  - GPU slave, 3rd party device master
- GPUDirect Async → control
  - GPU & 3rd party device master & slave

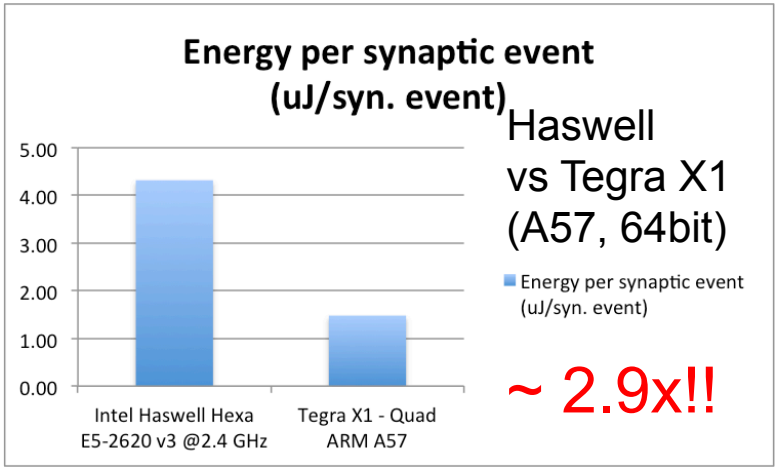


M. Martinelli

DPSNN and Low Power CPUs

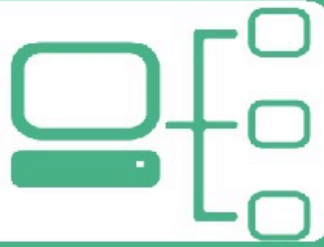
- Attivita' di benchmarking delle architetture low power (Nvidia Tegra)
- Risultati preliminari mostrano che per il TK1, il rapporto FLOPS/W e' migliore di quello per il TX1





- Stiamo già vedendo il limite delle architetture low power (almeno per applicazioni "strane")?
- E comunque:
  - SpiNNaker (specialized multi-core ARM): 20 nJ/syn. Evt.,
  - TrueNorth (ASIC): 26 pJ/syn. Evt.,
  - Human brain: 1–10 fJ/syn. evt. Range.

COSA @Roma oggi…



Issues:

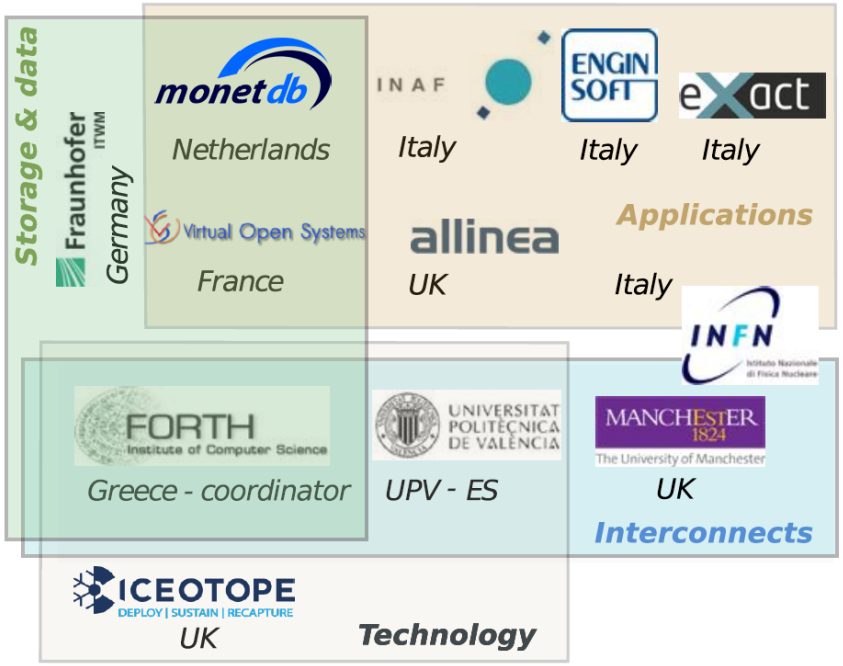
- Disponibilita' dei materiali di ricerca in forte ritardo
  - → Budget 2016 ridotto, bulk del finanziamento per acquisto dei materiali spostato in avanti
  - → Richiesta prolungamento al 2017 (+ 1 anno)
  - → Necessita' di ridurre gli FTE in COSA per coprire inefficienze amministrative alla partenza dei nuovi progetti EU in cui siamo coinvolti

Opportunita':

- Sinergia con progetto H2020 FET-HPC ExaNeSt e HBP Wavescales
  - RM+CNAF nel progetto per R&D su reti toroidali HPC per CPU ARM (su piattaforma basata su FPGA) e storage distribuito
  - Applicazione DPSNN tra i principali benchmark applicativi di ExaNeSt e simulatore di riferimento in WaveScales.
  - Co-design di un sistema ottimizzato per simulazioni di reti neurali

COSA + ExaNeSt

ExaNeSt Consortium



- European Exascale System Interconnection Network & Storage
- EU Funded project H2020-FETHPC-1-2014
- Duration: 3 year (2016-2018)
- Coordination FORTH (Foundation for Reaserch Technology, GR)
- 12 partners in Europe (6 industrial partners)
- www.exanest.eu

COSA + ExaNeSt

Exanest Objectives

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

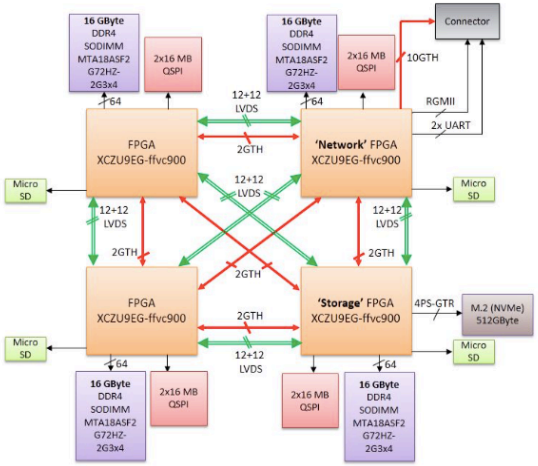
INFN

INFN activities are strongly synergic with project objectives:

- APE supercomputer: VLSI, system design, high density packing
- APEnet: FPGA-based NIC for clusters (low-latency, high-throughput)

COSA + ExaNeSt

ExaNeSt HW



- QFDB: 4 Ultrascale+ FPGAs (16 cores)
  - all-to-all connectivity (2 x HSS + 16 x LVDS)
- 64 GBytes DDR4 (16 GB/FPGA @ 160 Gb/s)
- 512 GBytes SSD/NVMe
  - 4x PCIe v2 (8 GBytes/s)
- 10 HSS links to remote
- 120mm x 130mm (in fabrication)

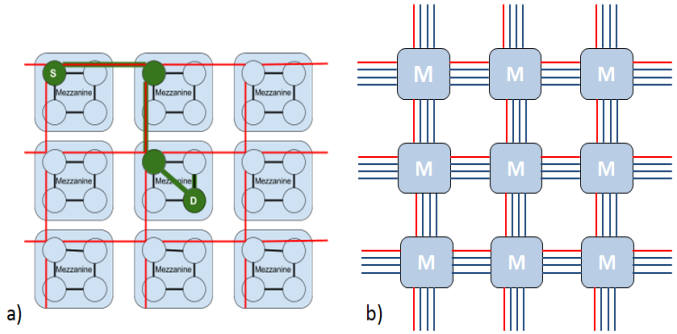
COSA + ExaNeSt

ExaNeSt: Interconnection Network

- Multi-tiered network: hierarchical infrastructure of separate networks interacting through a suitable set of communication protocols.
- Evaluate network architecture, topologies and related high performance technologies
- Unified approach:
  - Low latency RDMA
  - PGAS architecture
  - Merge heavy storage traffic and interprocessor data (Flow Prioritization)
- All-optical switch for rack-to-rack interconnect using 2×2/4×4 building blocks
- Support for resiliency
  - error detection, system and link diagnostic, multipath routing
- Topologies
  - Direct blade-to-blade networks (Torus, Dragonfly,...)
  - Indirect blade-switch-blade networks

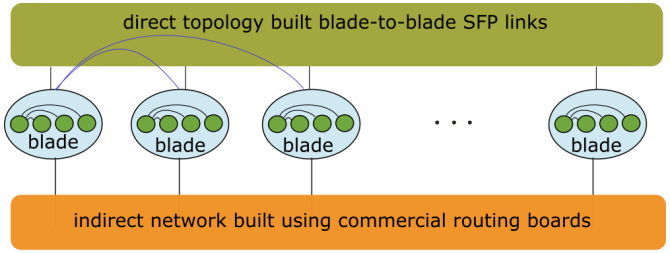
COSA + ExaNeSt

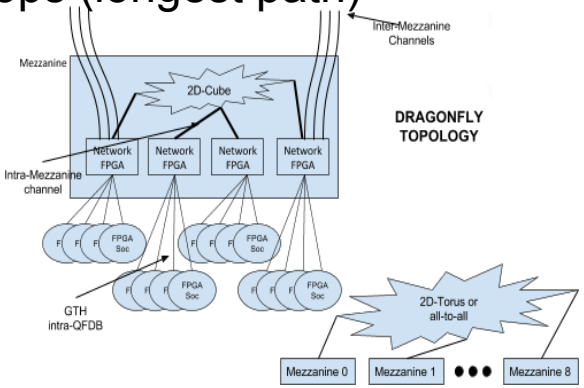
Configurable Interconnect Topologies



- 4x 2D-Torus interconnects (3x3)
- each QFDB of a mezzanine is connected with their counterparts on neighbouring mezzanines
- 3 hops (longest path)

- Exploration of Multi-level Dragonfly
  - QFDB → blade → system
  - Small diameter
  - Few expensive global wires





- Hibrid direct + indirect networks
  - Segregate throughput- from latency-sensitive traffic

DPSNN in HBP: WaveScales

SP3-WP2: WaveScalES

Towards a multi-scale perturbational atlas of the cerebral cortex. Linking the macro (TMS/EEG) to the microscale (cortical slices) through simultaneous recording of hd-EEG and Stereo-EEG responses to intracortical stimulation

Photostimulation and photoinhibition of slow-wave activity by light-regulated ligands of neuronal receptors

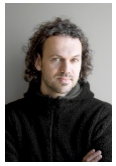
SP4, SP3, SP6, SP7

Multiscale theory/model of cortical dynamics of slow-waves. Matching theory and simulations with experiments

Parallel simulations ported from DPSNN prototyping engine to nest ::


Maria Victoria Sanchez-Vives


Marcello Massimini

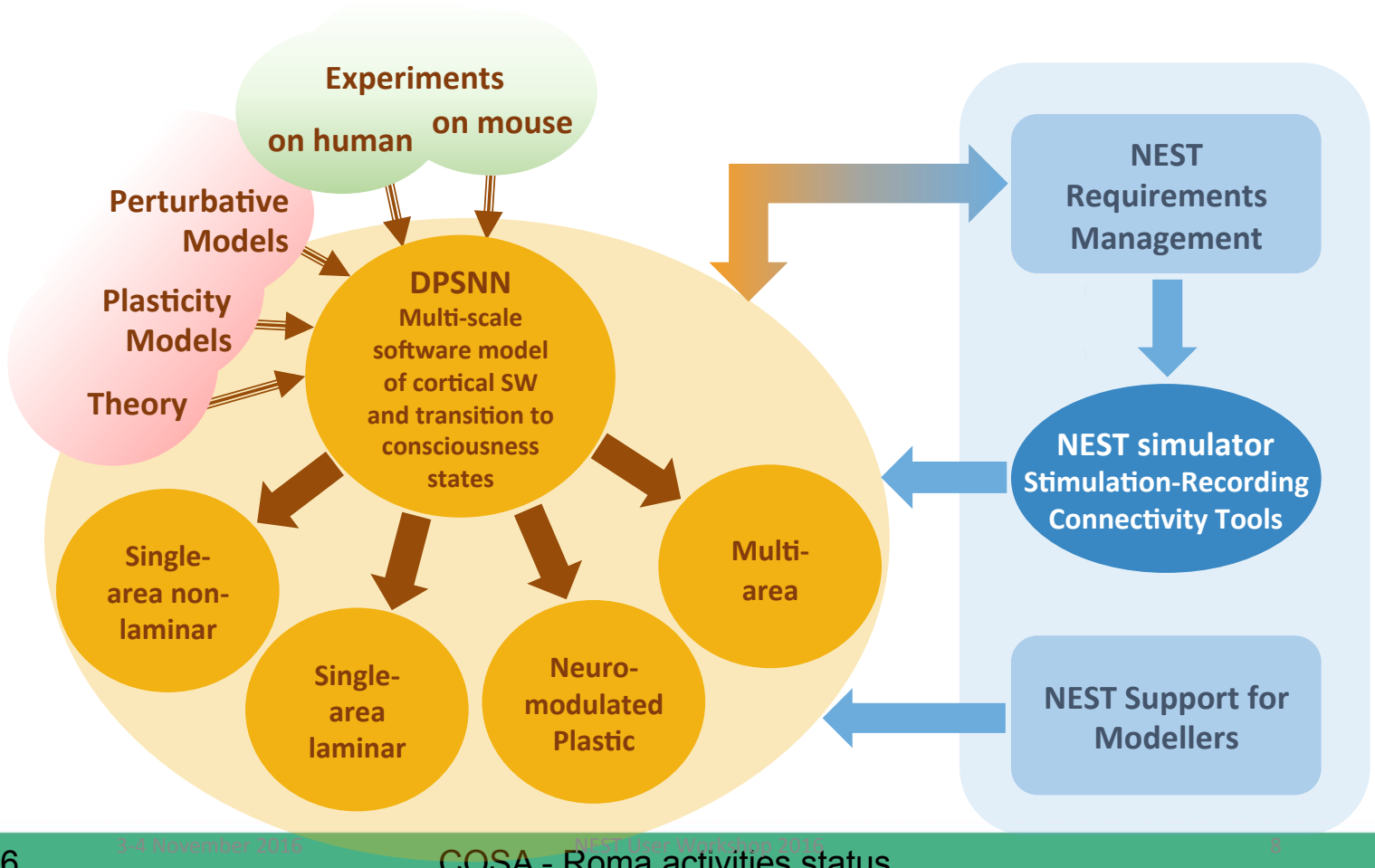

Pau Gorostiza


Maurizio Mattia


Pier Stanislao Paolucci

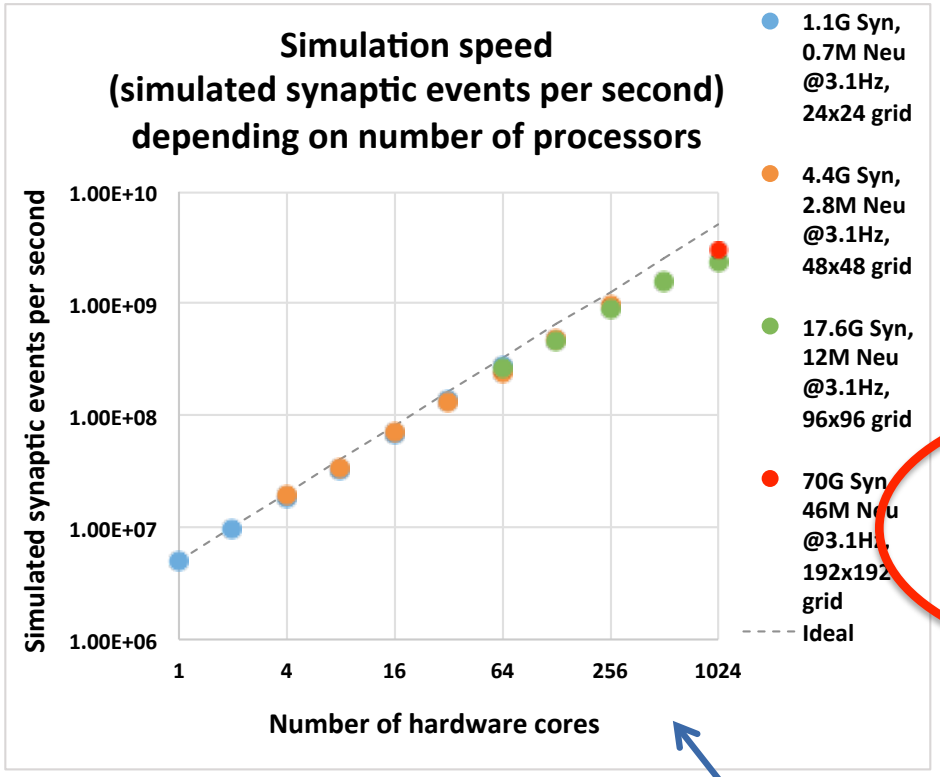
DPSNN in HBP: WaveScales

Collaboration with NEST



DPSNN in HBP: WaveScales

Large Scale Parallel Simulation



Efficient simulation of tens of billions of synapses, projected by grids of columns of point spiking neurons, distributed on thousands of hardware cores and software processes.

Study of hardware and software technologies for neural simulations

INFN

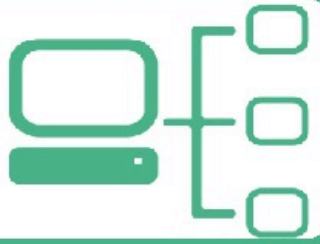
The APE parallel computing lab. of INFN

On proprietary DPSNN (Distributed Plastic Spiking Net Simulator) to be ported to

nest ::

HBP research infrastructure

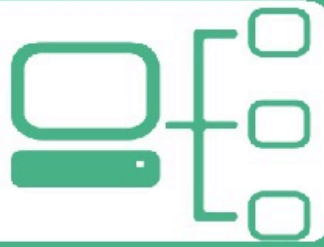
COSA, ExaNest and HBP...



COSA, ExaNeSt e WaveScales sono progetti fortemente sinergici e garantiscono il co-finanziamento del personale coinvolto.

- ExaNeSt -> nuova release APEnet+ integrata con ARM a 64bit per il computing e per il supporto ai network tasks
  - Principali nuove features:
    - PGAS per accesso coerente a bassa latenza alla memoria dei nodi remoti,
    - nuova Network Interface con CPU ARM a piu' bassa latenza,
    - nuovo link "resilient", routing adattivo e collettive ottimizzate
- WaveScales + ExaNeSt
  - Ottimizzazione del codice di simulazione DPSNN e sua integrazione in framework state-of–the-art (NEST)
  - Sperimentazione architetture HW/SW ottimizzate per reti neurali
- Partecipazione a progetto EuroExa, follow-up delle attivita' di sviluppo network e DPSNN benchmark in ExaNeSt
  - proposal sottomesso nella call FET-HPC 2016-17)
- Idee per nuova proposta di architettura di calcolo ottimizzata per simulazioni di reti neurali in ambito HBP (2018->)

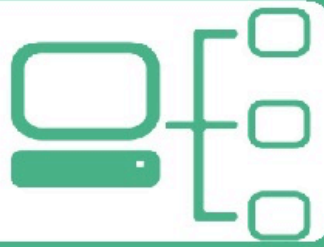
Attività RM 2016/17



Attivita' in corso nel 2016:

- Acquisizione 4 schede FPGA Dev Kit ZynQ Ultrascale+ (TRENZ) e schede d'interfaccia FMC multiple SFP+ e QSFP28/40 per integrazione FPGA cluster di Roma (gara MEPA conclusa, in "pietosa" attesa di emissione dell'ordine…)
- Gia acquisite ed in lab 2 di questi sistemi per prove preliminari
- In attesa della risoluzione della "querelle ALTERA-XILINX", porting preliminare di APEnet+ su Arria10 ed in parallelo su piattaforma XILINX ZYNQ Ultrascale+ in ambito ExaNeSt
  - Sezione custom di APEnet (switch, router block, etc…) done
  - In progress interfacciamento di APEnet ai links GTH XILINX
  - Benchmark iniziale su interfaccia AXI verso il sub-sys ARM
- Benchmark della piattaforma ZynQ UltraScale+ - based attraverso DPSNN

Attività RM 2016/17



Nel 2017 (forte ridimensionamento dei finanziamenti ma per fortuna piena sinergia con i progetti EU):

- completamento porting V5 su piattaforma Arria10 e Xilinx Zynq U+ (TRENZ), e QFDB-ExaNeSt
- Rilascio di APEnet+ V10 firmware (ARM supported)
  - Integrazione delle nuove features sviluppate in ambito ExaNeSt (principalmente su piattaforma Xilinx)
  - test di V10 su nuove generazioni NVidia GPU (in collaborazione con i progetti NaNet e NaNet-T)
- Completamento cluster FPGA-based
- DPSNN: completamento delle analisi di scaling e power efficiency su cluster low-power
  - Individuazione di criticita' architetturali di tali sistemi
  - Esplorazione di architetture ottimizzate per simulazioni di reti neurali

EuroExa (FETHPC 2016/17)

EuroExa fa leva sulle attivita' sinergiche di 3 iniziative attive nella call FETHPC-2014 a cui l'INFN partecipa (ExaNeSt).

- Obiettivi
  - prototipo funzionante di maturita' tecnologica alta, a scala medio/grande (alcuni Pflops) e basato su processori a bassa potenza, componenti riconfigurabili (FPGA) per accelerazione ottimizzata del calcolo e implementazione di una innovativa infrastruttura di network gerarchica e ibrida con topologie "dirette" (Torus e DragonFly) ed "indirette" (fat-tree).
  - nuovo ambiente di programmazione parallela basato su open-source esistente, ma ottimizzato ed integrato con framework di programmazione per i componenti riconfigurabili basato su OpenCL.

- Benchmark di sistema attraverso l'esecuzione di applicazioni scientifiche "grand challenge" a larga scala in una prospettiva di co-design applicazioni/HW/SW
  - codice DPSNN in NEST (simulazioni di reti neurali a larga scala, in sinergia con le attivita' del progetto HBP WaveScales) e
  - kernel di simulazioni di fluido-dinamica alla Lattice Boltzmann multidimensionale (contributo della sezione di Ferrara).

- INFN partecipa inoltre al design, debug, test e integrazione della nuova architettura di rete includendo tutti gli sviluppi low-level software necessari per la utilizzazione efficente della network stessa.
- Durata 42 mesi, budget INFN ~900kE (di cui 600kE per TD)

Richieste Roma per COSA Anno III (a luglio 2016…)

	MI	INV	INV SJ	CONS	SW	TOTALE
CNAF	2	24	3	2	6	37
FERRARA	2	10				12
PADOVA	2					2
PARMA	2					2
PISA	2					2
ROMA	2	10	24	2		38
TOTALE	**12**	**44**	**23**	**4**	**6**	**93**

Sede	Personale	FTE	WP
ROMA1	Alessandro Lonardo 0.2 Pier Stanislao Paolucci 0.2 Piero Vicini 0.2 TD (ExaNest) 0.4	1.0	1,2,4,5

- Inventariabile: 34kE
  - 4 schede FPGA Arria10/Stratix10 per lo sviluppo del prototipo di cluster basato su FPGA (6 kE)
  - 10 kE per sistema Nvidia PASCAL
- Consumo: 2 kE
  - Jumper cables e materiale per assemblare cluster FPGA
- Missioni: 2 kE
  - Riunioni di collaborazione, presentazione a conferenza, contatti con fornitori

COME E' ANDATA A FINIRE???