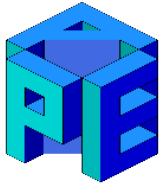
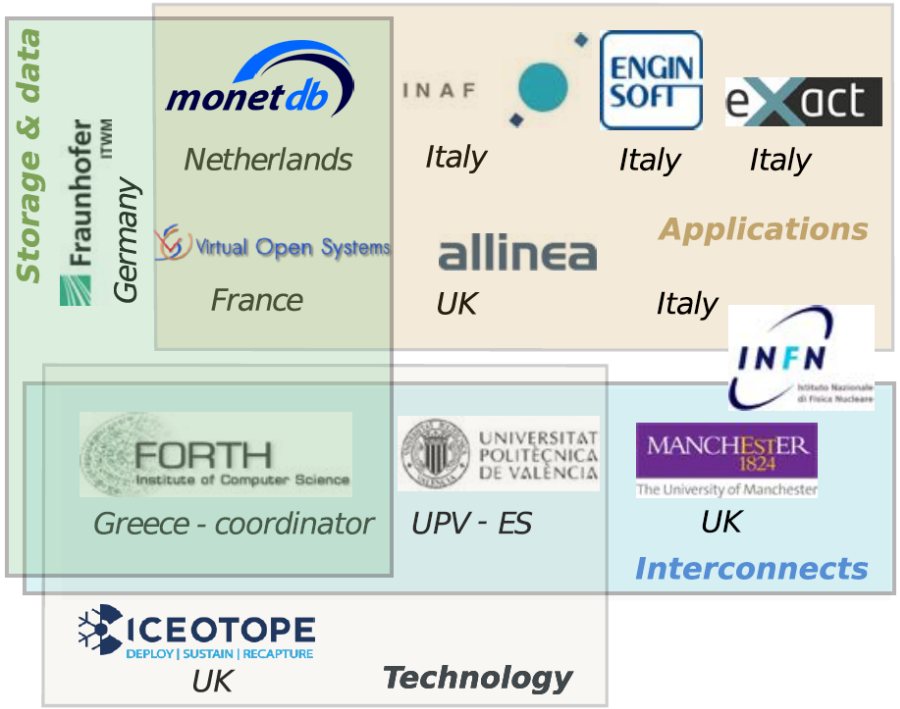


Low latency network and distributed storage for next generation HPC systems: the ExaNeSt project

Andrea Biagioni
INFN – Sezione di Roma
On behalf of ExaNest Consortium

Conference on Computing in High Energy and Nuclear Physics
10 – 14 october 2016

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

Exanest Objectives

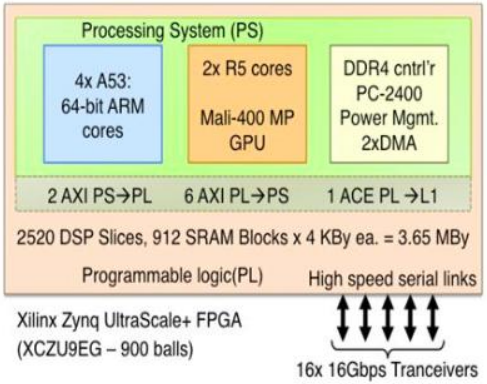
- System architecture for datacentric Exascale-class HPC
  - Storage Low-latency unified Interconnect (compute & storage traffic)
    - RDMA + PGAS to reduce communication overhead
  - Fast, distributed in-node non-volatile-memory
- Extreme compute-power density
  - Advanced totally-liquid cooling technology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

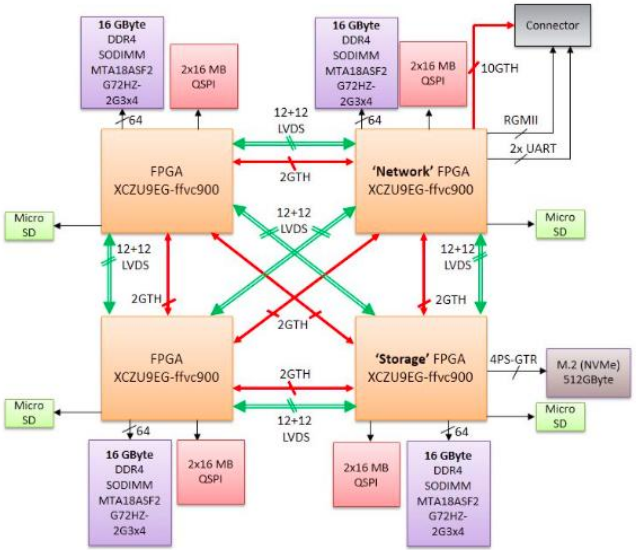
- APE supercomputer: VLSI, system design, high density packing
- APEnet: FPGA-based NIC for clusters (low-latency, high-throughput)

ExaNeSt Node (Tier 0)



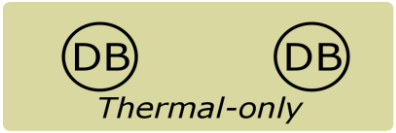
- Xilinx Zynq Ultrascale+ FPGA
  - Four 64-bit ARM Cortex- A53 cores @1.5 GHz
  - High throughput communication
    - 16 High Speed Serial link @16Gbit/s (32 GB/s)
  - Programmable logic: 2.5K DSP units @300MHz
    - 1.4 TFLOPS

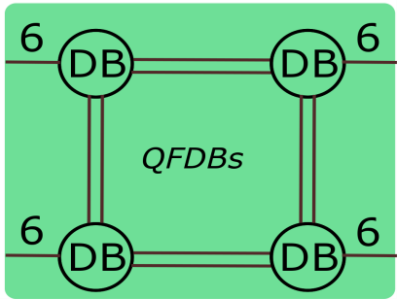
- Quad-FPGA Daughter-Board (QFDB)
- 4 Ultrascale+ FPGAs (16 cores)
  - all-to-all connectivity (2 x HSS + 16 x LVDS)
- 64 GBytes DDR4 (16 GB/FPGA @ 160 Gb/s)
- 512 GBytes SSD/NVMe
  - 4x PCIe v2 (8 GBytes/s)
- 10 HSS links to remote
- 120mm x 130mm (in fabrication)



ExaNeSt Blade/Mezzanine (Tier 1)

- Track 1
  - 4 QFDBs
  - 2 KALEAO
  - 2 Thermal-only
- Passive intra-mezzanine (local) QFDB-QFDB direct network
- 32 SFP+ connectors for inter-mezzanine (remote) network
  - different topologies





The ExaNeSt Prototype

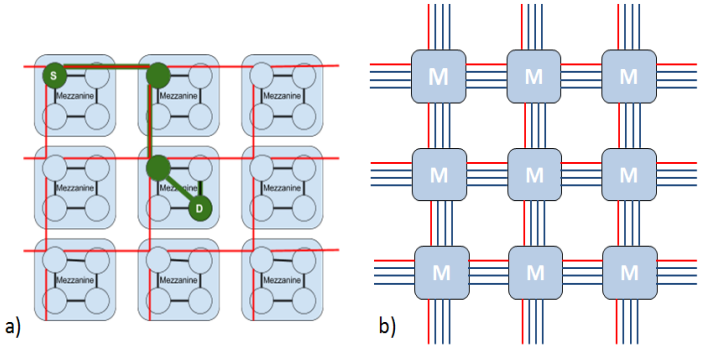
	Track 1	Track 2
Core (Node) per Blade	64 (4)	256 (16)
Blade per chassis	9	6
Core (Node) per Chassis	**576** (36)	**1536** (96)

- Totally liquid cooling
  - Track 1: immersed liquid cooled system based on convection flow
  - Track 2: phase-change (boiling liquid) and convection flow cooling (up to 350 kW of power dissipation capability)
- ~ 7 PFlops per racks and 20 Gflops/W
- ExaNeSt-based Exascale system
  - 140 racks, 21M ARM, 50 MW

ExaNeSt: Interconnection Network

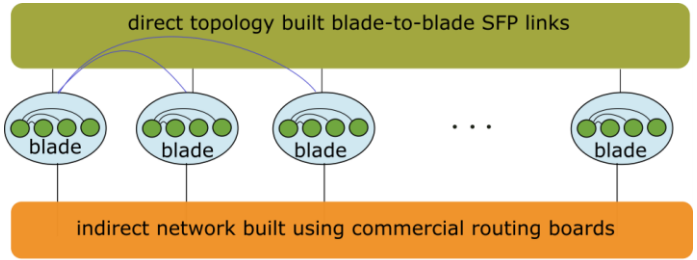
- Multi-tiered network: hierarchical infrastructure of separate networks interacting through a suitable set of communication protocols.
- Evaluate network architecture, topologies and related high performance technologies
- Unified approach:
  - Low latency RDMA
  - PGAS architecture
  - Merge heavy 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

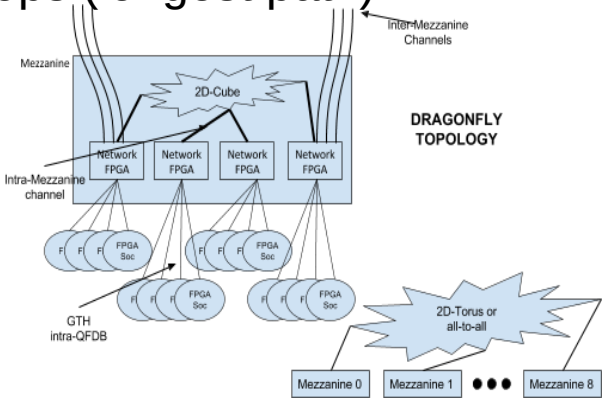
Configurable Interconnect Topologies



- 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

Application

- Co-design
  - Apps define quantitative requirements for the system under design
  - Apps evaluate the hw/sw system
    - Synthetic benchmark: Traces used to test I/O and network capability
    - Re-engineering of real application
- Astrophysics: Gadget, Pinocchio, Changa, Swift
  - Cosmological n-Body and hydrodynamical code(s)
- Neuroscience: DPSNN (Brain Simulation) INFN
  - Large scale spiking behaviours and synaptic connectivity
- Weather and climate: REGCM
- Material science: LAMMPS
- Data Analytics: MonetDB (database management)
- Engineering CFD: openFoam, SailFish

Conclusions

- The race toward ExaScale is started and Europe is trying to compete with established and emerging actors (USA, Japan, China,...)
- Many challenging issues require huge R&D efforts: power, interconnect, system packing and effective software frameworks
- ExaNeSt 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

- Exanest will deliver a fully working prototype able to be scaled up to the ExaFlops in the next years

THANK YOU!!!