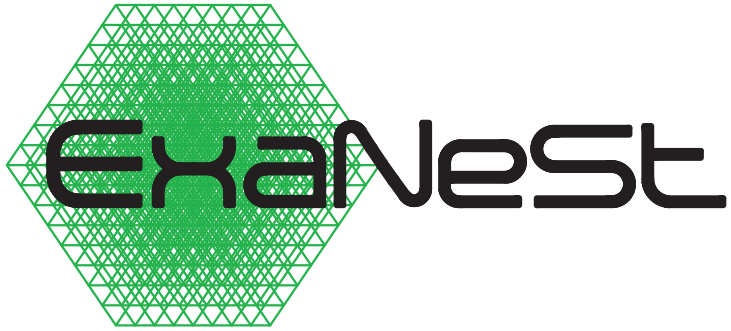


European Exascale System Interconnect & Storage

Giuliano Taffoni
INAF - OATs

ICT Meeting INAF – Cefalù 8 Ottobre 2015

COSA È ExaNeSt

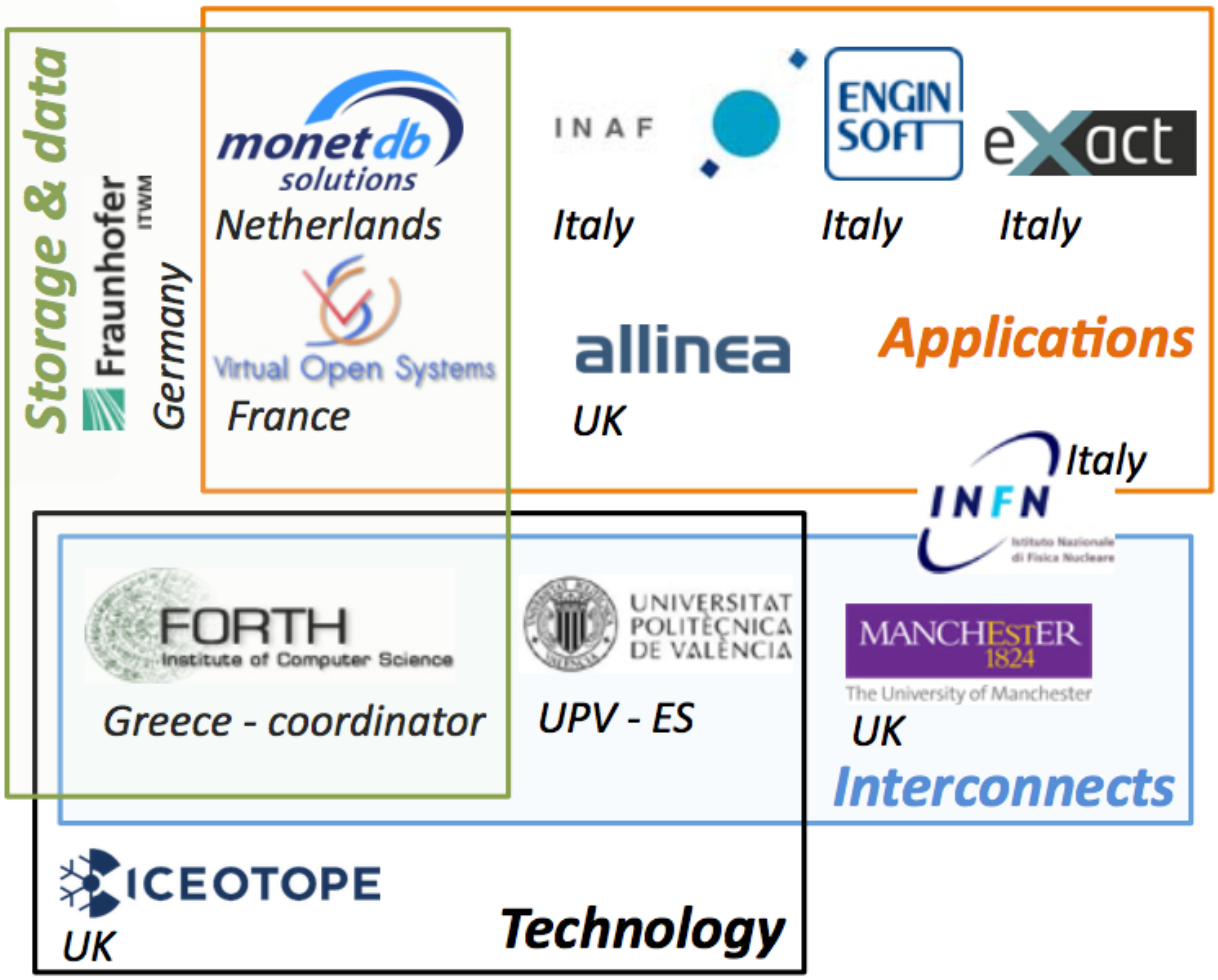
- Progetto Finaziato dalla EU nel programma H2020-FETHPC-1-2014
- Finanziamento totale di circa 7 Meuro
- 12 Partners in tutta Europa (6 partners industriali)
- Coordinatore Manolis Katevens FORTH (*Foundation for Research & Technology – Hellas)*

Quale obiettivo

"Overall strategy to develop an 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)
- **Extreme compute-power density**
  - Advanced *totally-liquid* **Cooling** technology (ICETOPE)
  - Scalable packaging for 64-bit ARM-based Microservers
- **Real scientific and data-center applications**
  - Applications used to identify system requirements
  - Tuned versions will evaluate our solutions

Il Consorzio



Ecosistema per exascale HPC

- ***EuroServer:*** *Green Computing Node for European microservers*
  - UNIMEM address space model among ARM compute nodes
  - Storage and I/O shared among multiple compute nodes
- ***ExaNoDe:*** *European Exascale processor-memory Node Design*
  - ARM-based Chiplets on silicon Interposer
- ***ECOSCALE:*** *Energy-efficient Heterogeneous Computing at exaSCALE*
  - Heterogenous infrastructure (ARM + FPGAs), programming, runEmes
- ***Kaleao:*** *Energy-efficient µServers for Scalable Cloud Datacenters*
  - Startup company, interested to commercialize many of the results

Verso l'exascale: infrastruttura hw

Hierarchy	Scale	Performance	DRAM	Storage	Maximum Power
Chiplet (Compute Unit)	Heterogeneous CPU/GPU compute unit	8 CPU 200 GFLOPS	Up to 6x 8GB	virtualized	15 W (16 GB)
Interposer (3D-IC)	4 × Chiplet	32 CPU 800 GFLOPS	64 GB	virtualized	70 W
Compute Node (Shared IO & Acceleration)	2 × Interposer, I/O + OpenCL FPGA	64 CPU 3.5 TFLOPS	128 GB	Host SSD 400-3400 GB	140 W + 20 W for I/O
Compute Element (daughter board PCB)	2 × Nodes	128 CPU 7 TFLOPS	256 GB	6.8 TB	320 W
Mezzanine (mother-board for Elements)	4 × Elements	512 CPU 28 TFLOPS	1 TB	27 TB	1.28 kW + 120 W Interconnect
Blade (deployment unit / hot-swap)	3 × Mezzanine	1536 CPU 84 TFLOPS	3 TB	81 TB	4.2 kW + 0.8 kW cooling
Rack (metal frame)	72 × Blades	110,592 CPU 6 PFLOPS	221 TB	5.8 PB	360 kW + 1 kW TOR switch
Example HPC System	100 × Racks	11 M CPU 600 PFLOPS	22 PB	58 PB	36 MW
ExaScale Level	167 × Racks	1 ExaFLOPS 18.5 M CPU	37 PB	1 ExaByte	60 MW

The UNIMEM Architecture, usata in ExaNeSt

- **Realistic rack-level shared-memory based on Unimem**
  - Single owner per page: every memory page in at most one node's cache
    - *no system-level hardware coherence traffic*
    - owner can be any node - not just the (local) one adjacent to DRAM
- **Example usage models**
  - Remote memory accesses, remote mailbox, remote interrupts:
    - for fast synchronization & processing of distributed (read-only) data
  - Remote-page borrowing for memory disaggregation
  - Zero-copy remote direct memory transfer (RDMA)
    - sockets over zero-copy RDMA
    - MPI over sockets

Racks using total liquid cooling

ICEOTOPE current solutions

- Immersed liquid cooled systems based on convection flow
- 52 kW / rack
  - 720 Watt / blade

Enhancements planned during ExaNeSt

- Hybrid : phase-change (boiling liquid) and convection flow cooling
- target 360 kW/ rack

System prototype

- 3 Chassis within an Iceotope Rack
- 27 blades
- 81 mezzanine board with 4 connectors
- 324 daughter cards that contains an EuroServer
- About 4500 ARM 64 bit Cores
- 20% of the budged invested in the prototype.

Quali applicazioni

- Cosmological n-Body and hydrodynamical code(s) suited to perform large-scale, high-resolution numerical simulations of cosmic structures formation and evolution.
- Lattice QCD simulations. Lattice QCD is a non-perturbative numeric approach to solve the theory of “strong interactions” at sub-nuclear level (quarks and gluons).
- Brain Simulation. Generate spiking behaviours and synaptic connectivity that do not change when the number of hardware processing nodes is varied
- weather and climate simulation
- Material science simulations
- Workloads for database management on the platform and initial assessment against competing approaches in the market,

WP2: Exascale HPC Application

- Leader INAF- OATs
- Requirement, testing and Porting
- “The applications will be re-designed for a many-core, large shared-memory address space architecture, so as to be able to identify networking tasks, storage tasks and computing tasks. “

- “Full re-engineering and optimization of the selected applications and associated networking, storage and computing algorithms. To get the better possible performance, the selected applications need to be structurally re-engineered, using an asynchronous parallelization paradigm.”