European Exascale System Interconnect & Storage

www.exanest.eu

Manolis Ploumidis *(ploumid@ics.forth.gr)*

Foundation for Research & Technology - Hellas (FORTH)

ARM: On the Road to HPC (hosted by the Mont-Blanc Project)

January 16-17, 2017, UPC, Barcelona

What ExaNeSt is about

- ARMv8, *UNIMEM* Partitioned Global Address Space (PGAS)
  - low energy compute
  - low overhead communicate
  - FPGA-assisted acceleration
  - working closely with *ExaNoDe*, *EcoScale*, (& EuroServer)
- *Network*: *unified* compute & storage, low latency
- *Storage*: distributed, *in-node* non-volatile memories
- Extreme Compute *Density*: totally-liquid cooling
- *Prototype*: 1K cores, 16GBytes DDR4 per FPGA
- Real *Applications*: Scientific, Engineering, Data Analytics

The ExaNeSt Prototype (2016 – 17)

- Using Xilinx Zynq UltraScale+ FPGAs:
  - Quad-core 64-bit ARM A53 per FPGA
  - Cache-coherent low-latency I/O port
- On 120×130 mm^2 Daughter Boards
  - 4 FPGA's
  - 0.5 to 1 TBy SSD
  - 10× 16Gb/s I/O's
- Prototype track-1
  - 8 DB's per Blade, Dozen Blades
- Prototype track-2
  - 16 DB's per Blade
- DB Design completed
  - First deployment = within 2017
- SW dev. now on EuroServ. Prototy.

Interconnection Network

- Now: Simulations, Studies:
  - at the Packet/flit level, for protocol behavior and interactions (using INSEE and Omnet+);
  - Traffic Inputs: Synthetic models, real App Traces, or running App's.
- Later: Experiments on real Prototype running real App's
- Packaging & interconnect considered in tandem
  - Hierarchical interconnect
- Design Goals:
  - unified network for compute & storage
  - flow prioritization: heavy / storage versus short / sync (compute)
  - throttle congestive flows at network edges
  - resiliency: error detect/correct, monitor links, multipath routing
  - Zero-copy, user-level RDMA
  - Global address space
  - all-optical proof-of-concept switch using 2×2/4×4 building blocks

Applications, Traces

Main Applications:

- *Material science*: LAMMPS
- *Climate change*: REGCM
- *Engineering CFD*: openFoam, SailFish
- *Astrophysics:*
  - *Large-scale high-resolution simulations of cosmic formation and evolution*
  - Gadget, Pinocchio, Changa, Swift
- *Neuroscience – brain simulation*: DPSNN
- *High Energy Physics*
  - Lattice Quantum Chromodynamics simulations - LQCD
- *Data Analytics*: MonetDB
- Porting selected App's to ARM

Storage: current Design work

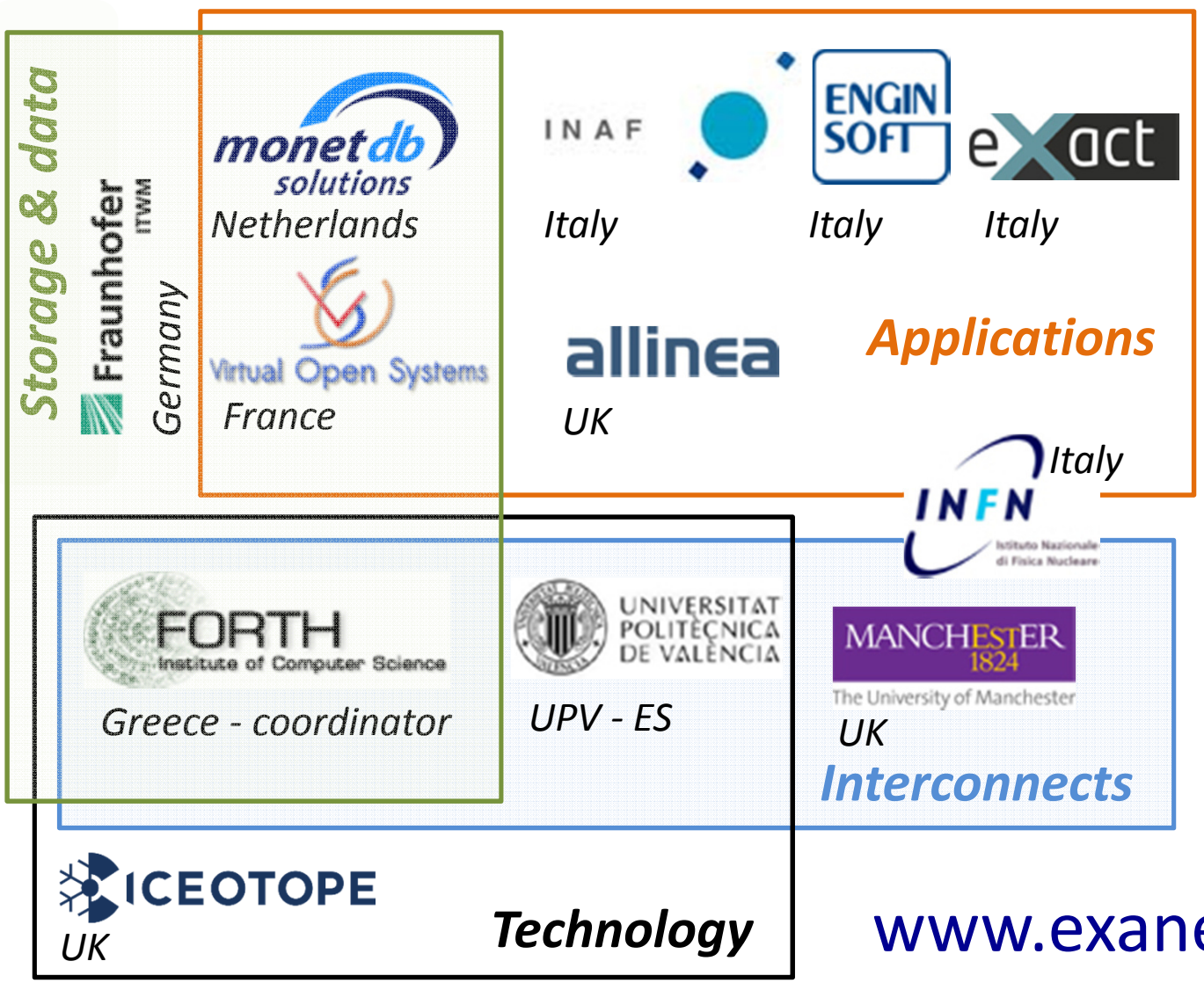
Global Storage Layer +
+ per-job SSD/NVM on-demand Parallel Cache Layer

- Based on the *BeeGFS* parallel filesystem (open source), with caching and replication extensions
- Low-latency memory-mapped storage access path in Linux
- Virtualization: RDMA from within VM's; MPI remoting
- Acceleration for Host-to-VM and VM-to-VM interactions

What ExaneSt offers to the Ecosystem

- Use of the Prototype by ExaNoDe and EcoScale (2016 – 17)
- Use of the Prototype by the Ecosystem (2018 onwards)
- HPC technology components (2018 onwards):
  - ARM/Unimem
  - Fine-tuned Applications
  - Packaging & Cooling
  - Distributed NVM / Storage
  - Interconnects
  - DB compute-node prototype

The ExaNeSt Consortium



www.exanest.eu

European Exascale System Interconnect & Storage

- Interconnection **Ne**twork
- In-node **St**orage
- Advanced Cooling
- Real Applications

www.exanest.eu

Project Coordinator:
Prof. Manolis Katevenis (kateveni@ics.forth.gr)