European Exascale System Interconnect & Storage

www.exanest.eu

Manolis Katevenis, Coordinator

Foundation for Research & Technology - Hellas (FORTH)

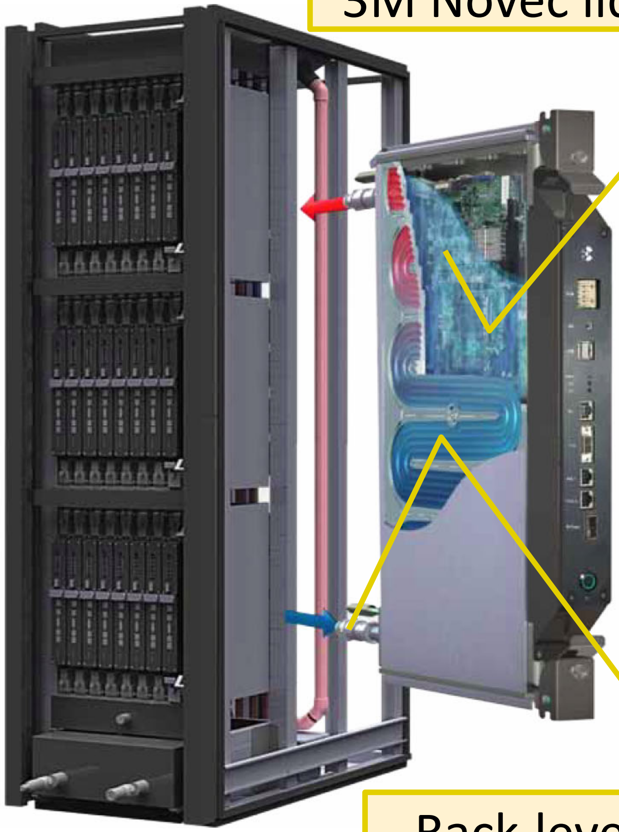
European HPC Summit Week, Prague, 10 May 2016

What ExaNeSt is about

- ARMv8, *UNIMEM* Partitioned Global Address Space (PGAS)
  - low energy compute
  - low overhead communicate
  - heterogeneous: FPGA accelerators
  - 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, 4 Tby DRAM, 40 Tby SSD, 0.5 M DSP sl's
- Real *Applications*: Scientific, Engineering, Data Analytics

The ExaNeSt Prototype (2016 – 17)

- Using Xilinx Zynq UltrScale+ FPGAs:
  - Quad-core 64-bit ARM A53 per FPGA
  - 2.5 K DSP slices (~1 TFLOPS equiv.)
  - Cache-coherent low-latency I/O port
- On 120×130 mm^2 Daugther Boards
  - per DB: 4 FPGA's, 64 GBy DDR4, 0.5 to 1 TBy SSD, 10× 16Gb/s I/O's
- 8 DB's per Blade, Dozen Blades
- DB Design completed; deploy first few DB's in Fall'16, many in 2017
- SW dev. now on EuroServ. Prototy.



Electronics immersed in 3M Novec liquid

Rack-level water circulation

Interconnection Network

- Now: Simulations, Studies:
  - at the Packet/flit level, for protocol behavior and interactions (using INSEE and Omnet+);
  - later at the Flow level: large-scale effects in exascale topologies.
  - Traffic Inputs: Synthetic models, real App Traces, or running App's.
- Later: Experiments on real Prototype running real App's
- 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
  - all-optical proof-of-concept switch using 2×2/4×4 building blocks

Applications, Traces

Traces generated:

- *Scalasca* profiling tool:
- MPI calls instrumented,
- several GBytes per trace,
- filtered down to tens of Mbytes by keeping what our network simulators will need;
- generally, to be made publicly available.

Main Applications:

- *Material science*: LAMMPS
- *Climate change*: REGCM
- *Engineering CFD*: openFoam, SailFish
- *Astrophysics*: Gadget, Pinocchio, Changa, Swift
- *Neuroscience*: DPSNN
- *High Energy Physics*: LQCD
- *Data Analytics*: MonetDB

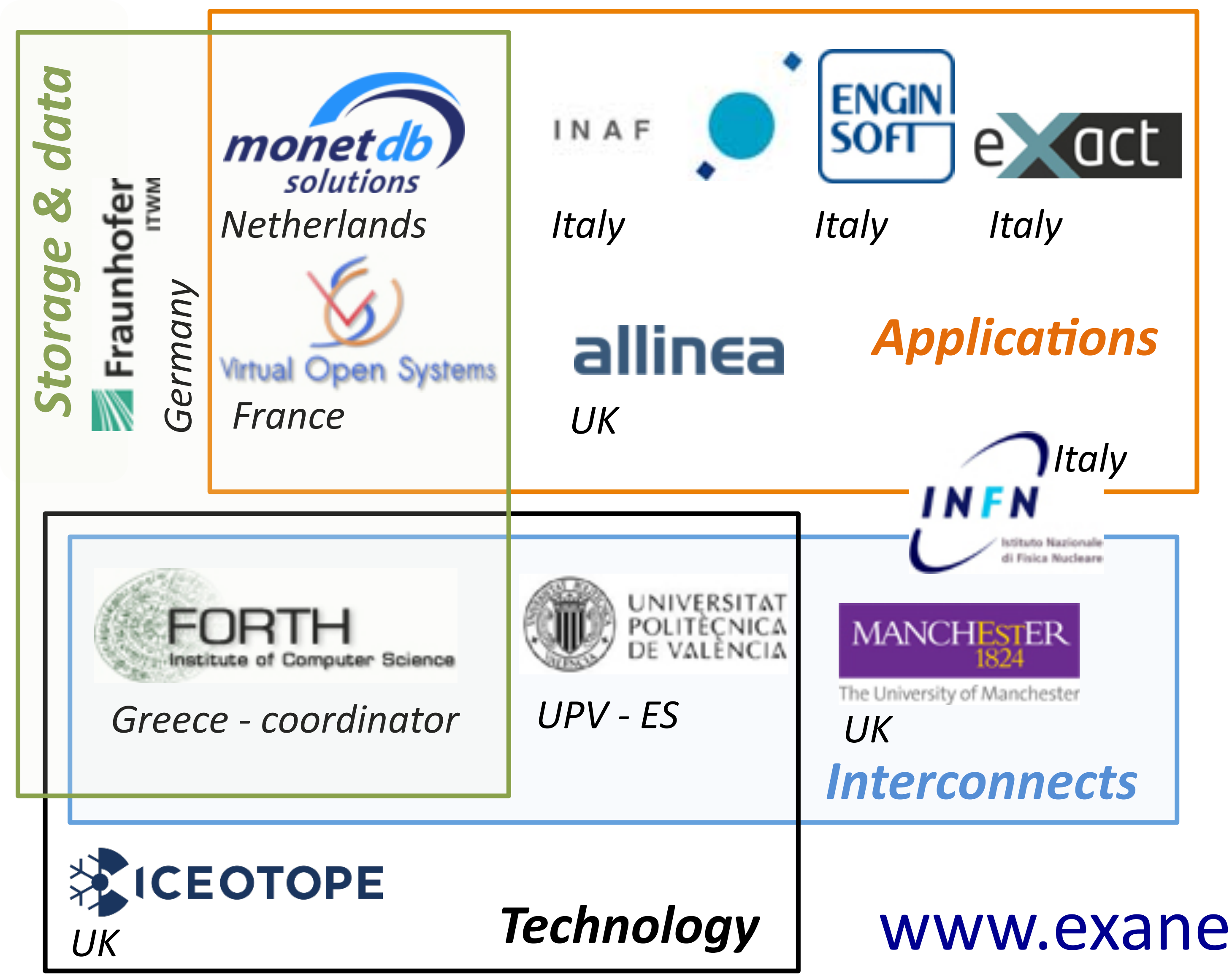
- Next – Applications Porting & Tuning:
  - currently porting selected App's to ARM, on the EuroServer Prototype

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

The ExaNeSt Consortium


monetdb solutions
Fraunhofer ITWM
INAF
ENGIN SOFT
eXact
Virtual Open Systems
allinea
INFN Istituto Nazionale di Fisica Nucleare
FORTH Institute of Computer Science
UNIVERSITAT POLITÈCNICA DE VALÈNCIA
MANCHESTER 1824 The University of Manchester
ICEOTOPE

Storage & data

Germany

Netherlands

France

Italy

Italy

Italy

UK

Applications

Italy

Greece - coordinator

UPV - ES

UK

Interconnects

UK

Technology

www.exanest.eu

ExaNeSt

European Exascale
System Interconnect & Storage

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

www.exanest.eu

ExaNeSt