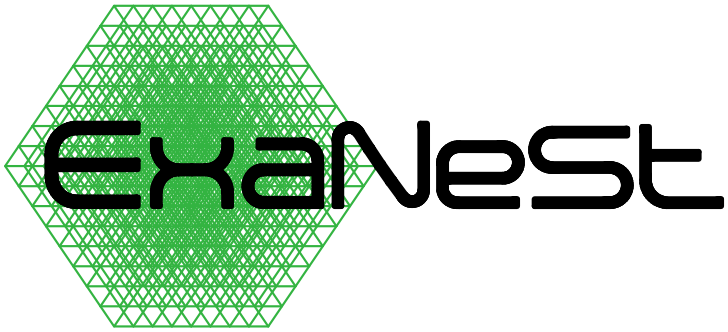


European Exascale System Interconnect & Storage

www.exanest.eu

Peter Hopton (peter.hopton@iceotope.com)

Iceotope, UK

The Foundations For ExaScale

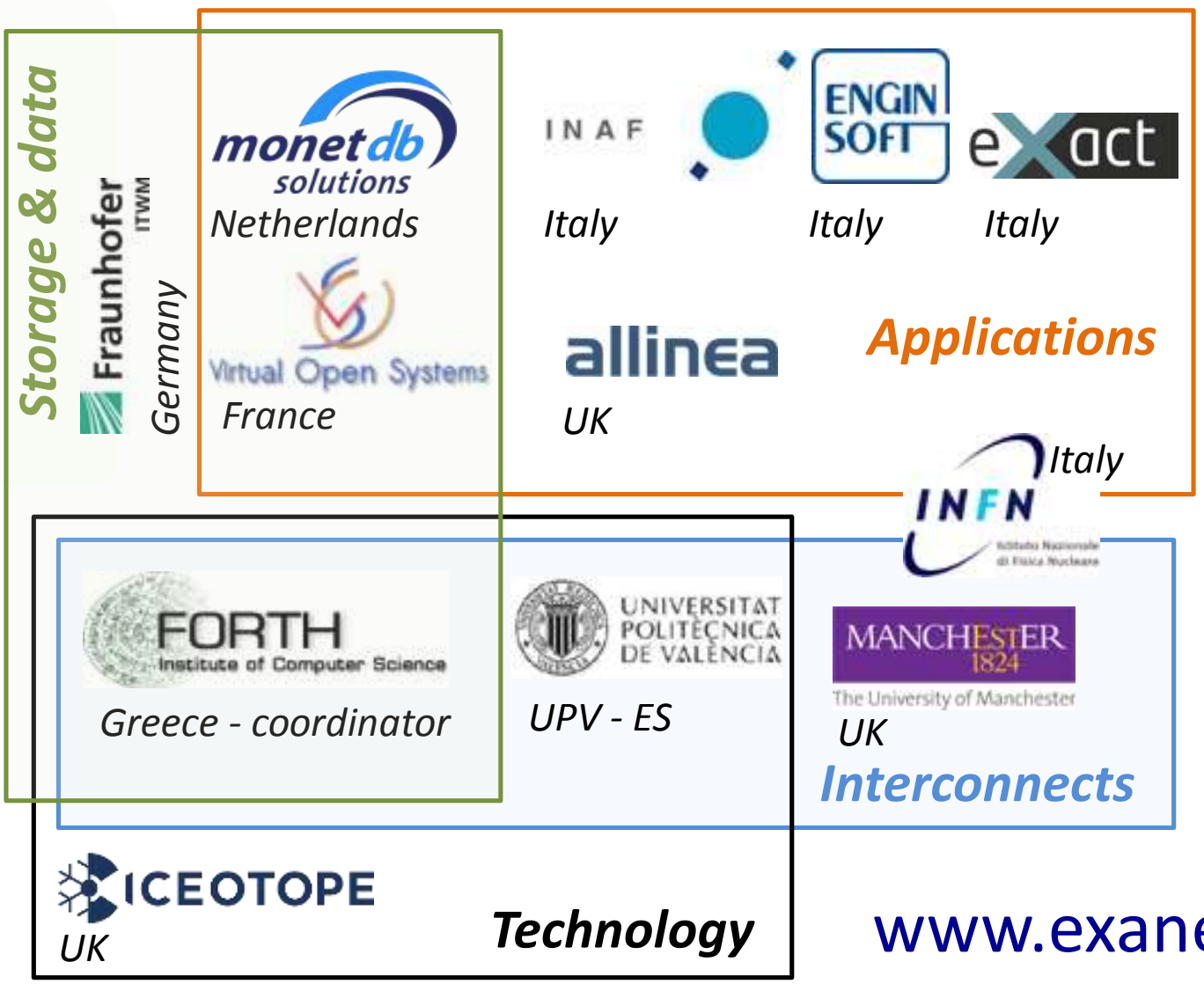
Project Group Landscape

- “EuroExa” Group Of Projects
- Historic Input Effort (from 2010)
  - EuroServer, Mont Blanc, Encore
- Currently Running Projects (2014-2018)
  - ExaNeSt
  - EcoScale
  - ExaNode
- Follow Through Project (2017-2020)
  - To Be Announced... ;-)

What ExaNeSt is about

- ARMv8, *UNIMEM* Partitioned Global Address Space (PGAS)
  - low energy compute
  - low overhead communicate
  - FPGA-assisted acceleration
  - working closely with *Other Projects…*
- ***Network***: *unified* compute & storage, low latency
- ***Storage***: distributed, *in-node* non-volatile memories
- Real ***Applications***: Scientific, Engineering, Data Analytics
- Data Centre ***Infrastructure*** (Power and Cooling): Total Liquid Cooling
- ***Prototype***: 1K cores, 16GBytes DDR4 per FPGA

The ExaNeSt Consortium



www.exanest.eu

Network

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 TOR switch using 2×2/4×4 building blocks

Storage

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

Applications

Applications, Traces

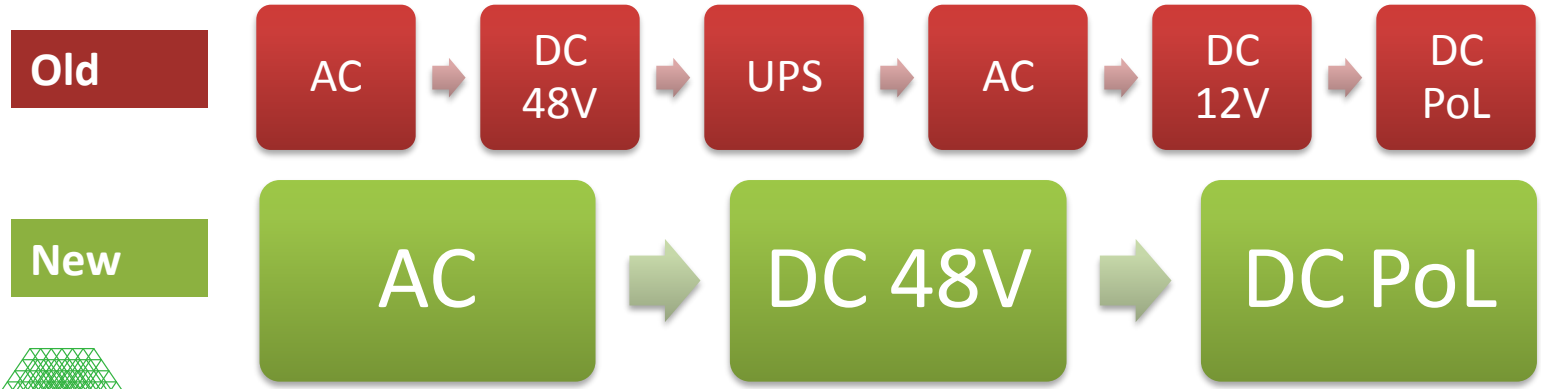
Main Applications:

- *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

Infrastructure

Power

- 48V DC to POL Converter on Daughter Board
- 3 Phase 415V to 48V DC Liquid Cooled Power Conversion
- 48V DC Power Distribution
- Ambition to integrate renewable energy on-site generation and energy storage at DC to remove conversion cycles and give “green power” (*Separate Project)*



Cooling

Combination of 3 Different Patented Cooling Technologies

- Immersion
  - Natural Convection “Cells”
  - Forced Convection “Fountains”
- Conduction
- Total Liquid Cooling
- Zero Airflow
- Potential for High Temperatures and Heat Capture

Prototypes

Our Nodes

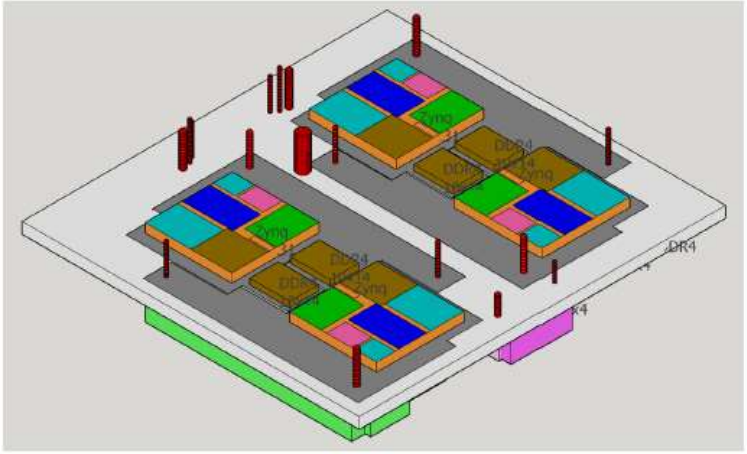


Figure 2: Iso view

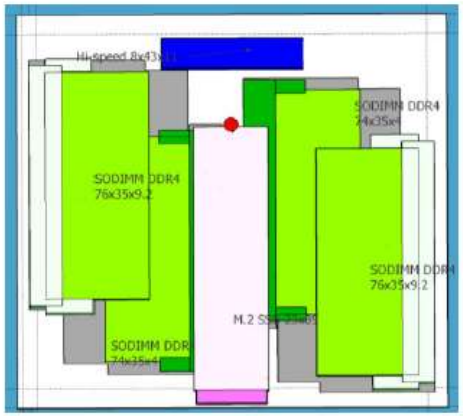


Figure 3: Bottom view

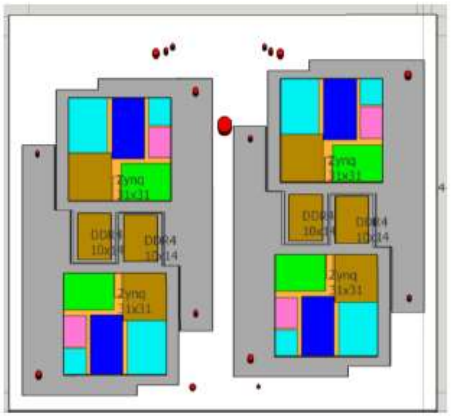
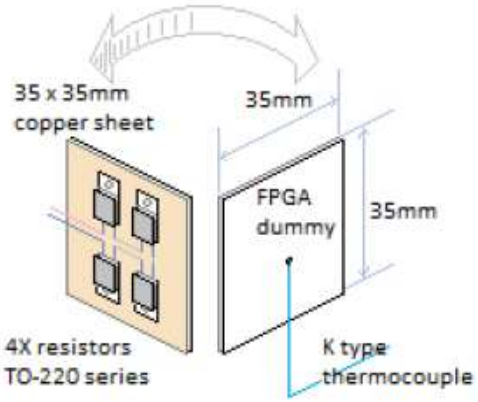
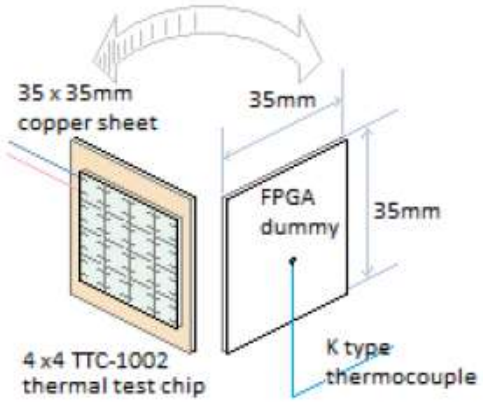
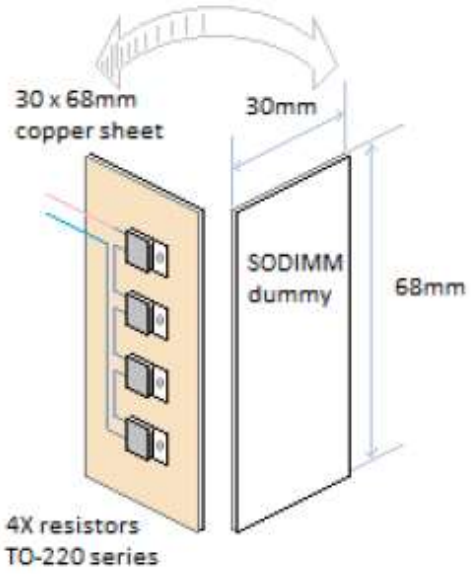


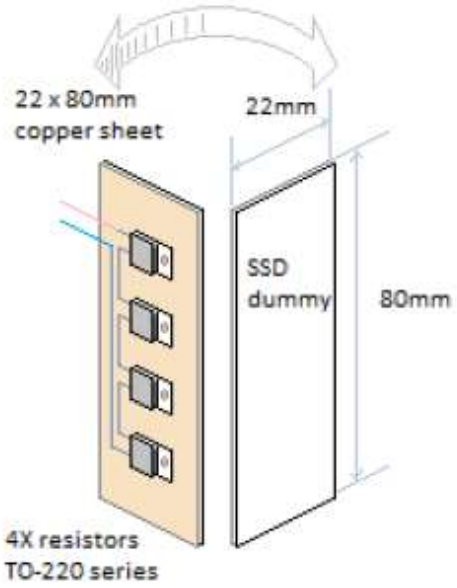
Figure 4: Top view

Thermal Proxies

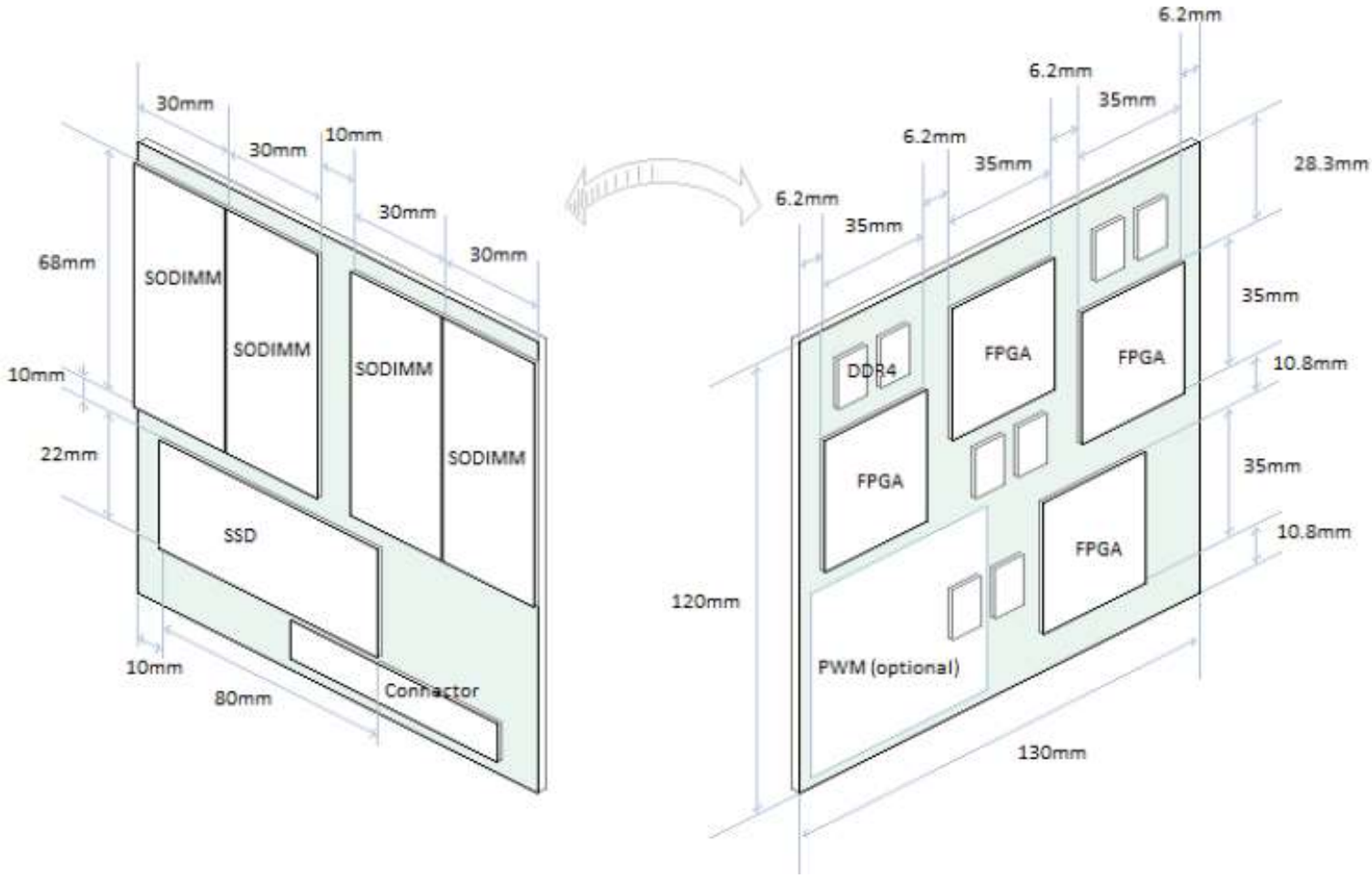








Thermal Proxies

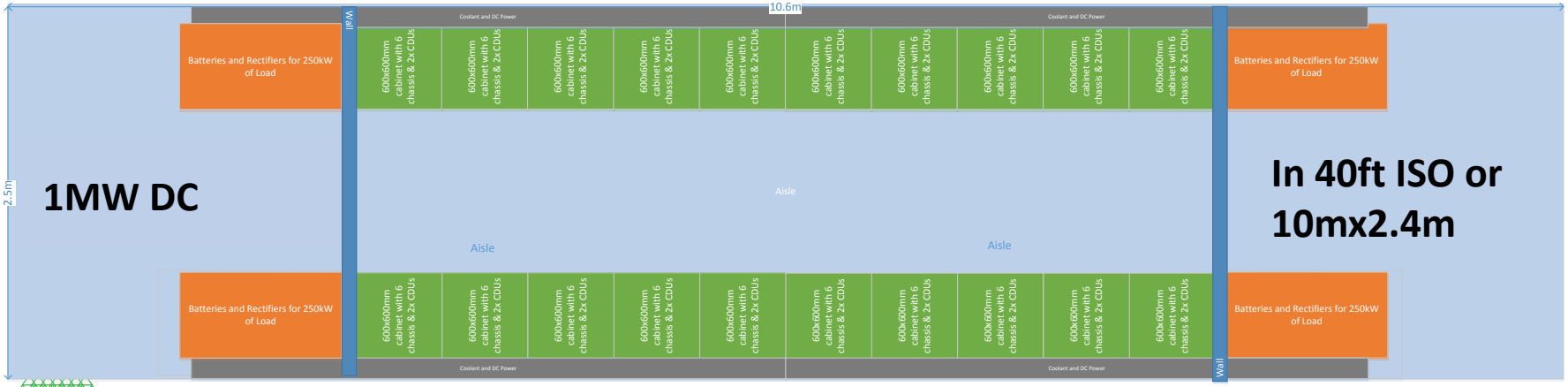


The 1st Major ExaNeSt Prototype (2017)

- Using Xilinx Zynq UltraScale+ FPGAs:
  - Quad-core 64-bit ARM A53 per FPGA
  - Cache-coherent low-latency I/O port
- On 120×130 mm² Daughter Boards
  - 4 FPGA's
  - 1 TB SSD
  - 10× 16Gb/s I/O's
- 1st Prototype
  - Up to 8 DB's per Blade, 9 Blades
- In Early Test Now
  - First deployment = within 2017

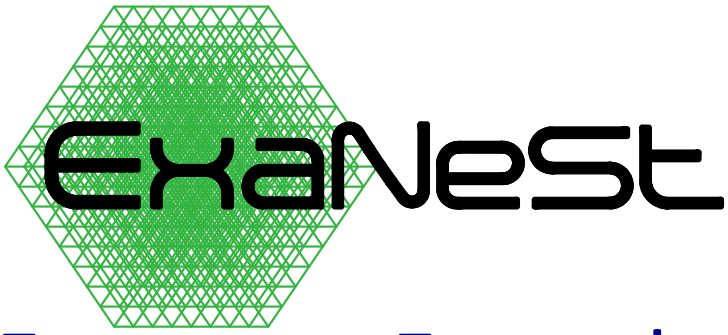
The 2nd Major ExaNeSt Prototype (2018)

- Same FPGAs, Same Quad FPGA Daughter Board
- 16 DB's, 1 Blade, 3.2kW in 1u, Small Footprint Cabs
- Our Ambition
  - Hot Water Cooling (>50C), Heat Capture Functions
  - 100kW cabinets, 1.0x PUE, >0.9 ERF Potential
  - DC Power Distribution
  - Potential for High Density Data Centres or Modular Facilities:



What ExaneSt offers to the Ecosystem

- Use of the Prototypes by ExaNoDe and EcoScale
- Use of the Prototype by the Ecosystem (2018 onwards)
- HPC technology components (2018 onwards):
  - ARM/Unimem
  - Fine-tuned Applications
  - The Ultra-Dense Data Centre for Scaling HPC
  - Packaging & Cooling
  - Distributed NVM / Storage
  - Interconnects
  - DB compute-node prototype



European Exascale System Interconnect & Storage

- Interconnection **Ne**twork
- In-node **St**orage
- Data Centre Infrastructure
- Advanced Cooling & Power
- Real Applications

www.exanest.eu Project Coordinator:

Prof. Manolis Katevenis (kateveni@ics.forth.gr)