

**ENEA**

**Atos** CINECA *Inria*



université de **BORDEAUX**



UNIVERSITÀ DI PISA



UNIVERSITÀ  
DEGLI STUDI  
DI TORINO



Consiglio Nazionale delle Ricerche



POLITECNICO  
MILANO 1863



**Fraunhofer**

**E4**

COMPUTER  
ENGINEERING



Istituto Nazionale di Fisica Nucleare

**QUATTRO**

**textarossa**

# The TEXTAROSSA Project: Cool all the Way Down to the Hardware

**Antonio Filgueras (BSC)**

DSD 2024, Paris, France

# Partners

- 11 partners
- 6 linked institutions
- 5 countries
- Lead by Massimo Celino (ENEA)



# Project Motivation

- Performance and energy efficiency remain main HPC challenges
  - Users are in demand of higher performance
  - Power often limits the available performance
- Heterogeneous systems try to address these challenges
  - Increased complexity
  - Large knowledge gap with domain experts

# Objectives

- Increase performance while keeping energy bounds
  - Hardware stack redesign
    - Infrastructure improvements (2-phase cooling and thermal management)
    - Experimental Hardware platforms (GPU and FPGA based)
  - Software stack redesign
    - Use application-specific accelerators
    - Efficient multi-device and multi-node runtime support
- Lower the entry barrier for new users to heterogeneous HPC systems
  - Provide a set of application-specific IP blocks for different tasks
  - Develop tooling for leveraging these IPs
  - Provide tools for resource management



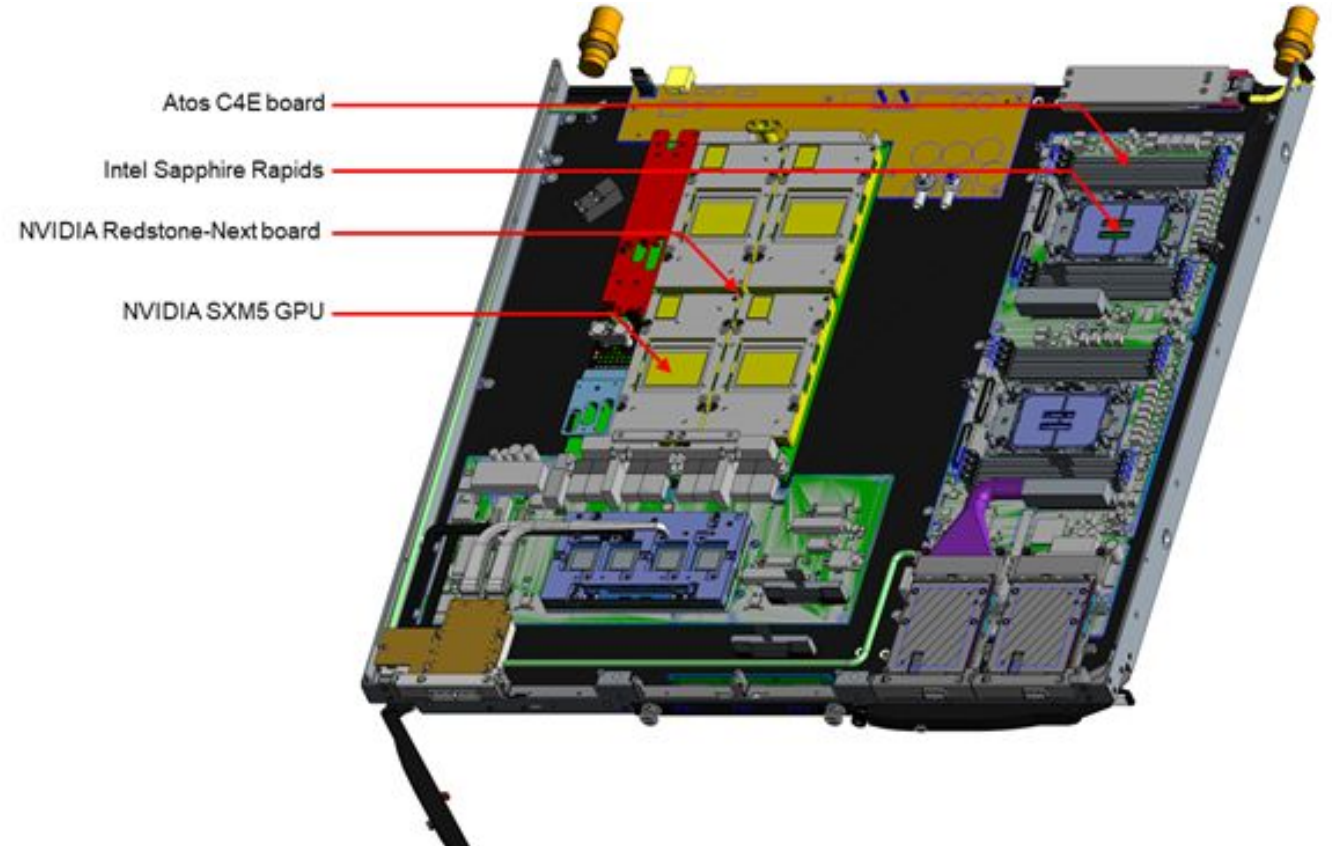
# Hardware prototypes

**textarossa**



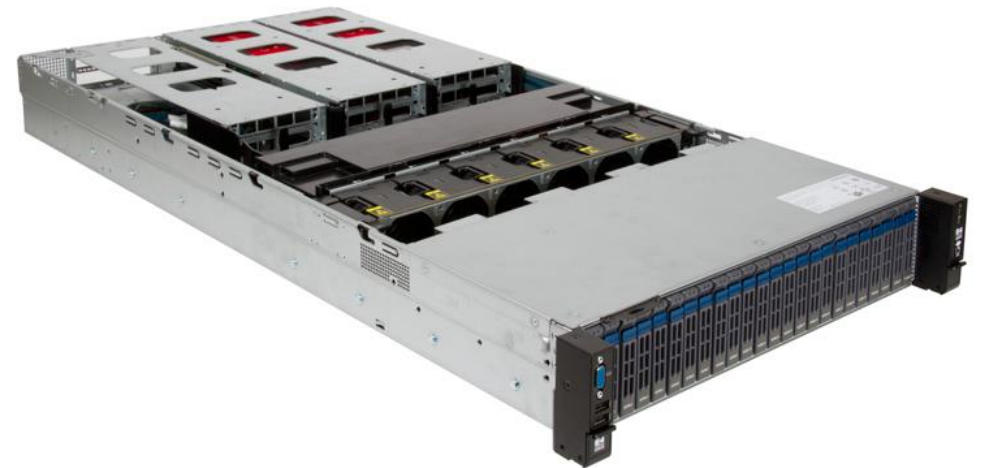
# Hardware platforms: IDV-A

- Developed by **Atos**
- 4 Nvidia H100 GPUs
- 2 Intel Xeon 8470 CPUs (2x54 cores)
- 2-phase cooling system
- >3500W Thermal Design Power



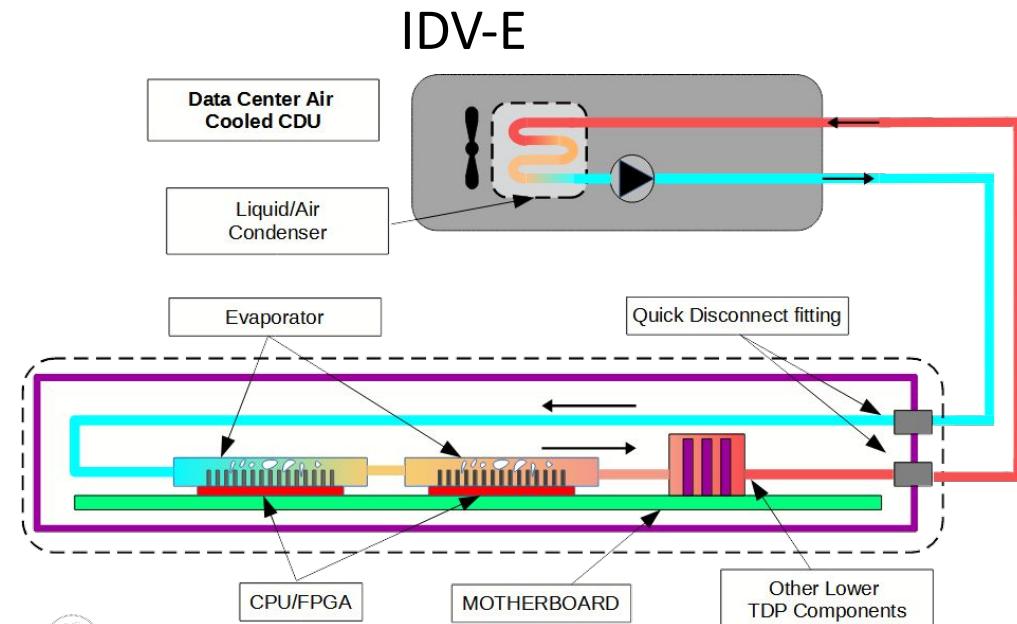
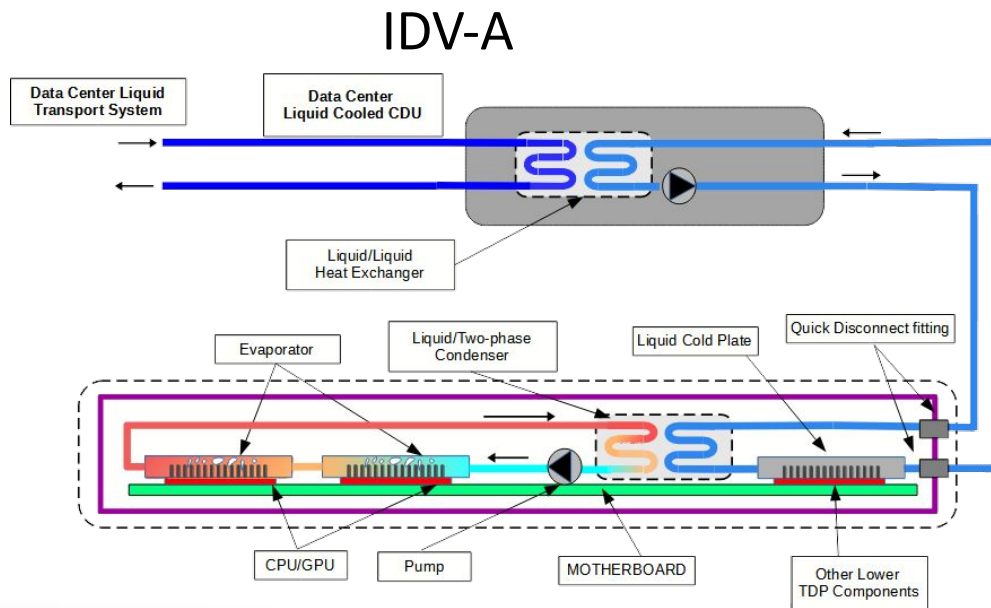
# Hardware platforms: IDV-E

- Developed by **E4**  
(based on Ampere Mt.Collins)
- 2 Ampere Altra Max CPU  
(2x128 ARMv8 cores)
- 2 AMD Alveo U280 Accelerator cards
- 2-phase cooling system
- 950W TDP



# Evaporative Cooling (two-phase cooling)

- Use fluid phase change for energy exchange
  - Electronic device cooling (evaporation)
  - Waste heat reject (condensation)
- Impact on thermal control strategies

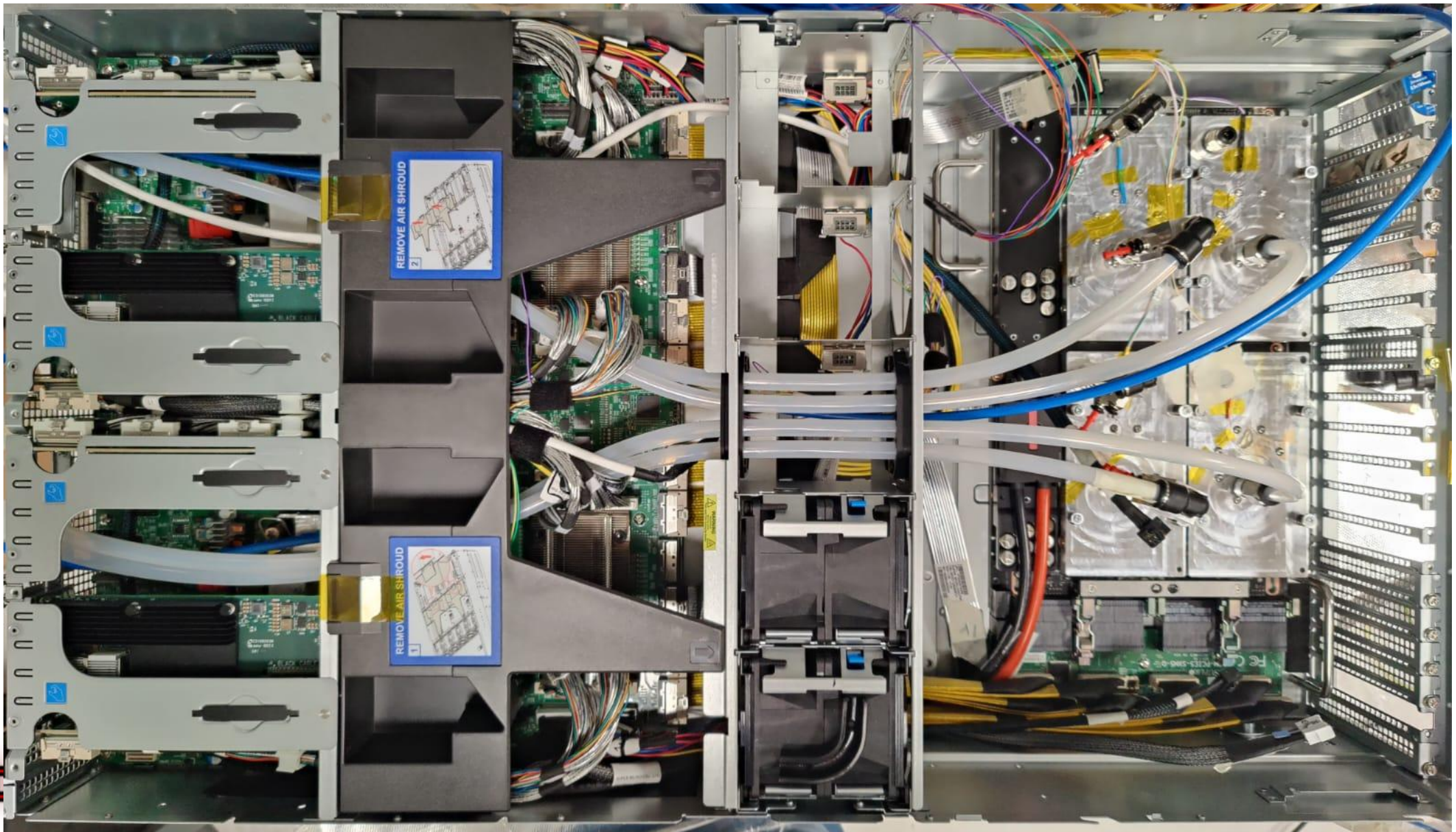




# Thermal Management

- Thermal control prevents ICs (CPUs, GPUs, FPGAs) from burning
- Use Dynamic Frequency and Voltage Scaling to reduce power (and heat)
- Heat spikes are quick (~10ms)
- Cooling actuators (pumps, fans) have to adapt to load changes
- Power actuators (DVFS) need to take into account thermal mass and actuator inertia
  - Fan vs. pump, heatsink vs. block + fluid, etc.
- Transient models of the evaporative cooling loop have been developed
- Hierarchical thermal controller has been designed

# IDV-A Prototype



text

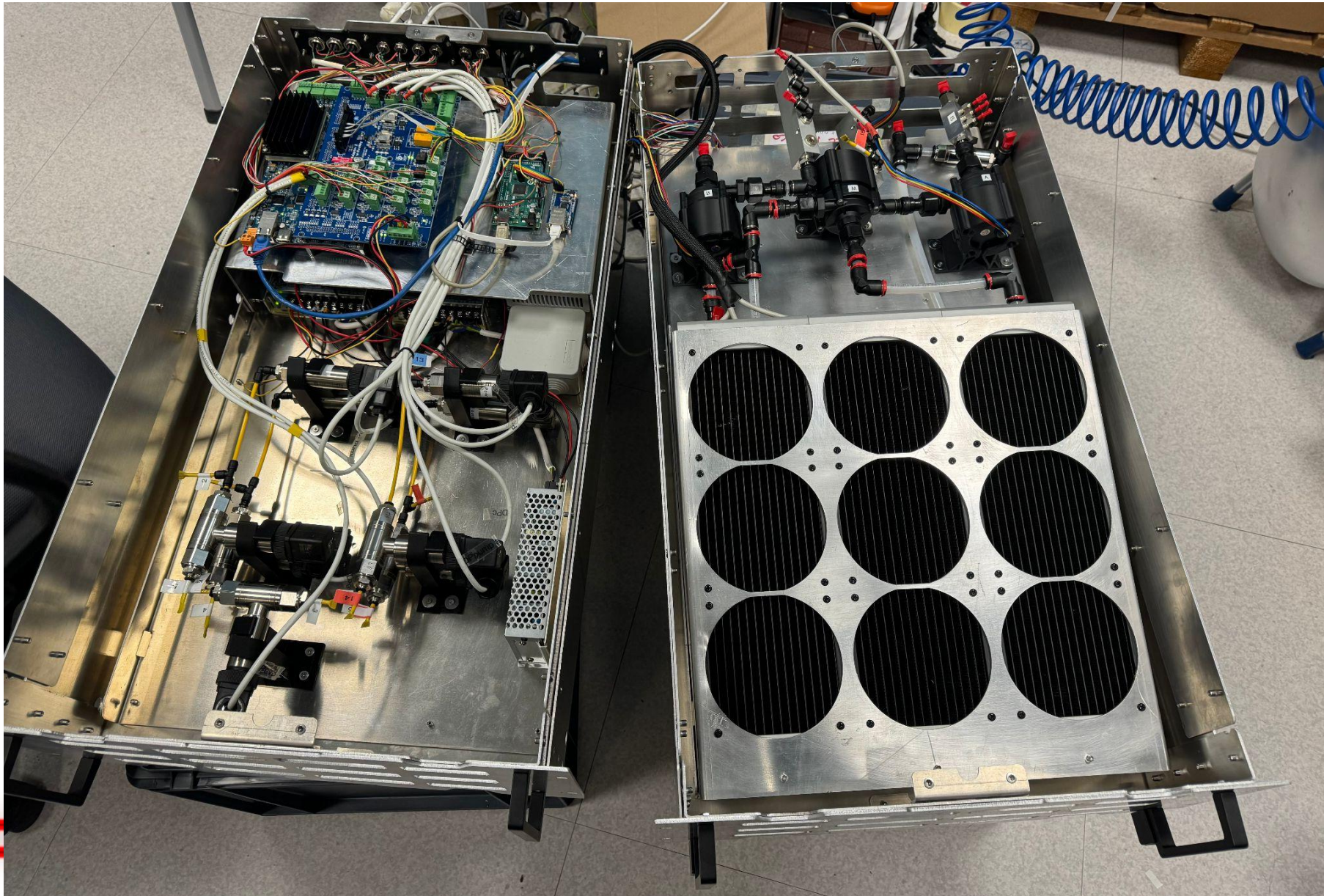


# IDV-A Prototype



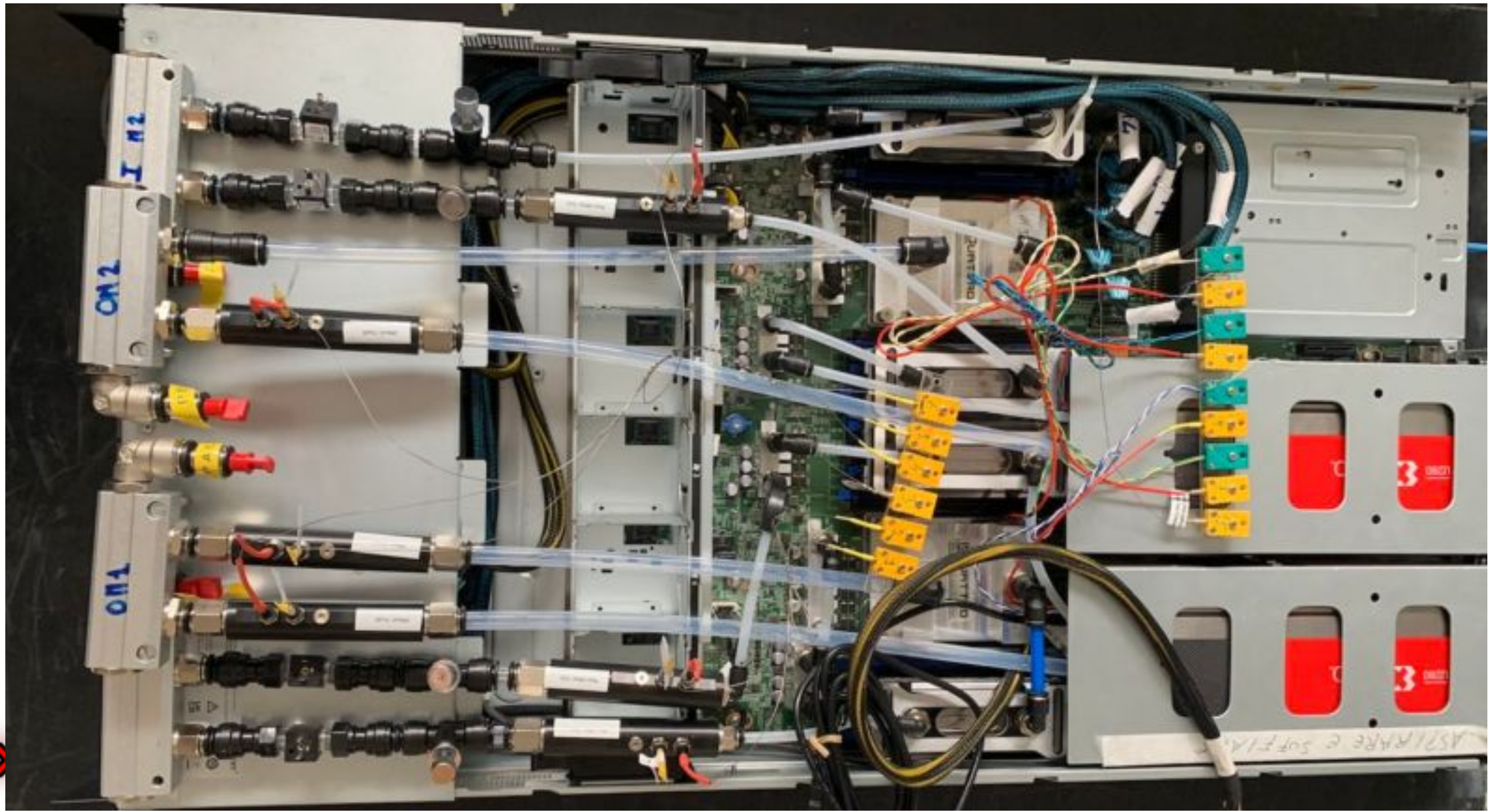


# IDV-A Prototype



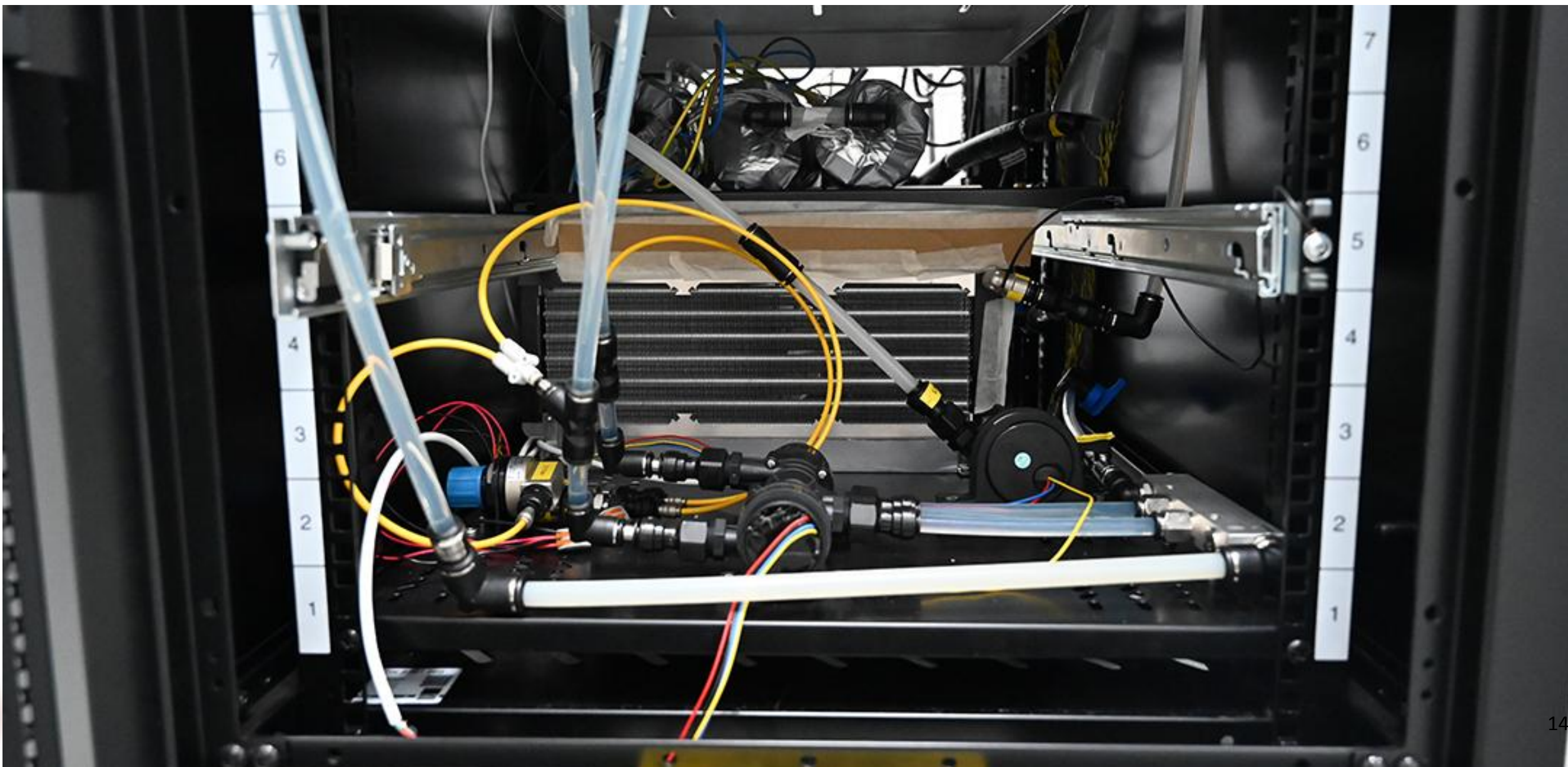


# IDV-E Prototype





# IDV-E Prototype









# IP project contributions

**textarossa**

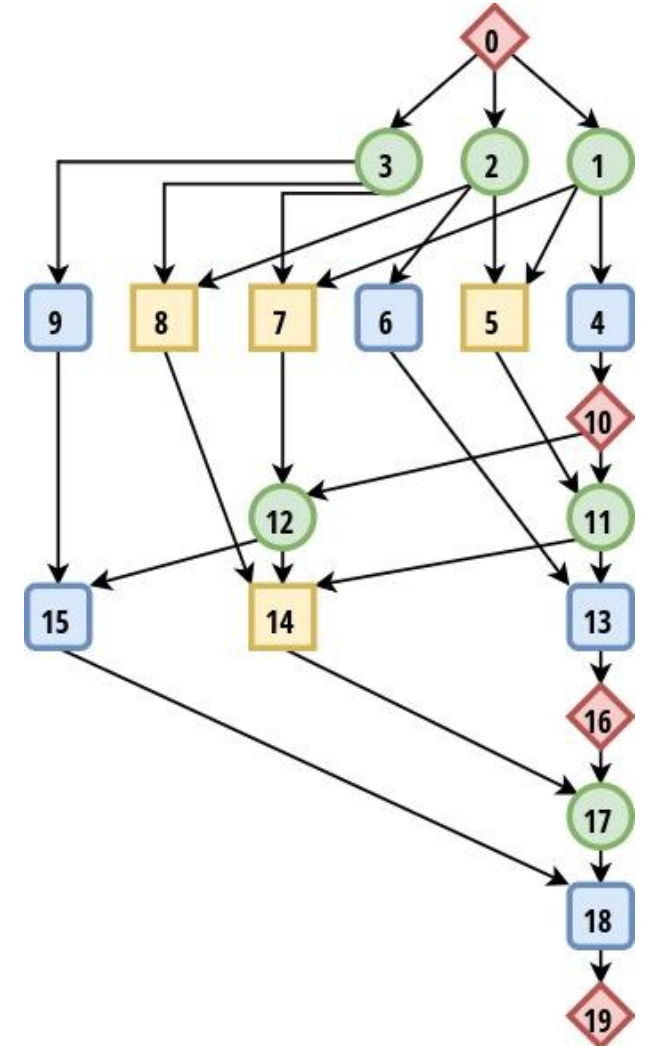
# OmpSs@FPGA

## OmpS-2 task-based programming model

### Cholesky source code

```
void cholesky_blocked(const int nt, float *A[nt][nt])  
{  
  for (int k = 0; k < nt; k++) {  
    #pragma oss task inout(A[k][k])   
    potrf( A[k][k] );  
    for (int i = k+1; i < nt; i++) {  
      #pragma oss task in(A[k][k]) inout(A[k][i])   
      trsm( A[k][k], A[k][i] );  
    }  
    for (int i = k+1; i < nt; i++) {  
      for (int j = k+1; j < i; j++) {  
        #pragma oss task in(A[k][i], A[k][j]) inout(A[j][i])   
        gemm( A[k][i], A[k][j], A[j][i] );  
      }  
    }  
    #pragma oss task in(A[k][i]) inout(A[i][i])   
    syrk( A[k][i], A[i][i] );  
  }  
}
```

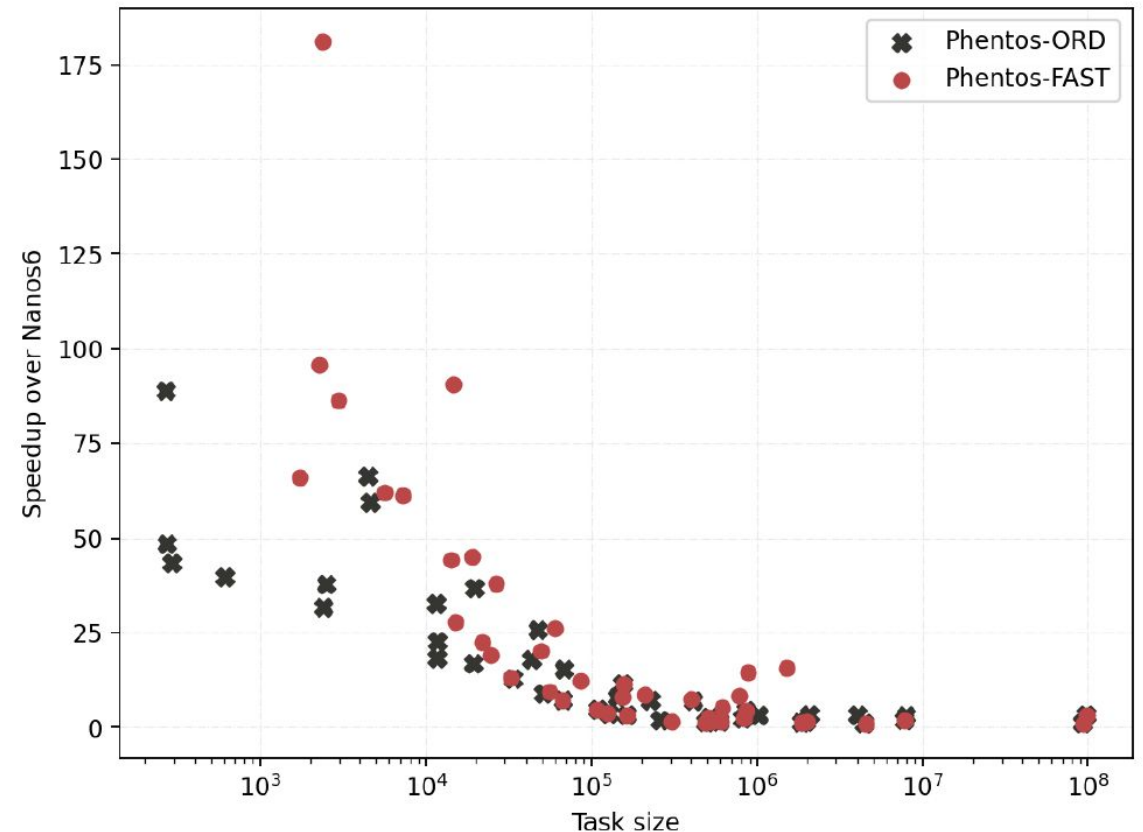
### Task graph



# Hw IP for Task Scheduling

- A HW Fast Task Scheduler IP allows OmpSs@FPGA and many-core nodes to schedule tasks with negligible overhead
- In many-core RISC-V nodes results in over 100x speedup in task scheduling
- In OmpSs@FPGA allows near perfect scalability (as shown later)

Speed-up over SW scheduling (30 RISC-V cores)









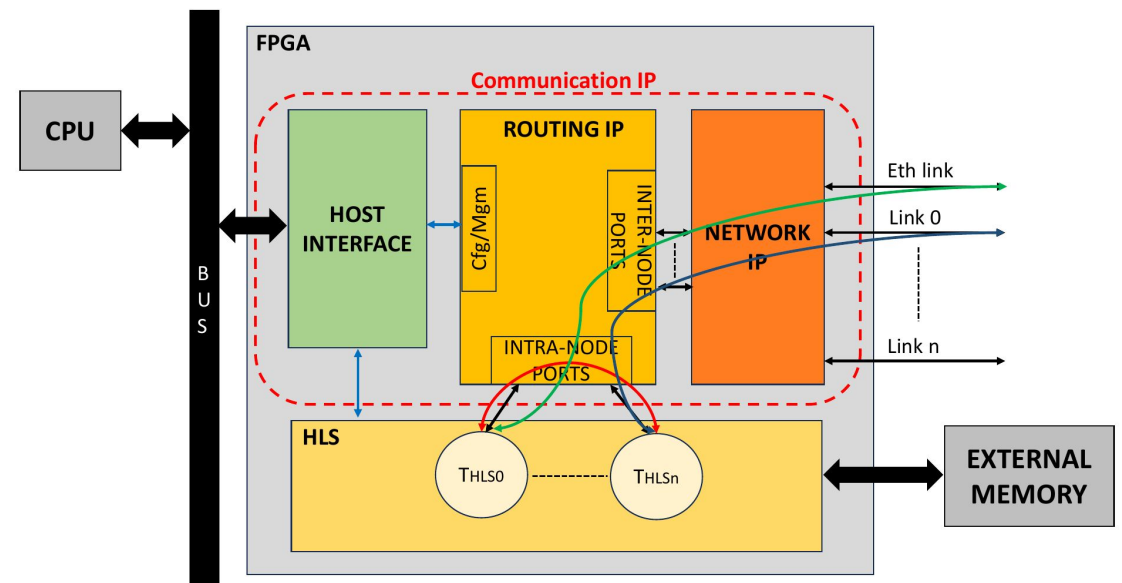
# APEIRON: Low-Latency FPGA communication

- Communication between kernels in the same or different device
- 3D mesh routing
- HLS kernel interface

## Kernel API

```
size_t send(void* data, size_t size, int dest_node,  
           int kernel_id, int channel,  
           message_stream_t message_data_out[N_CHANNELS]);  
size_t receive(void *data, int channel,  
              message_stream_t message_data_[N_CHANNELS]);
```

## IP Architecture



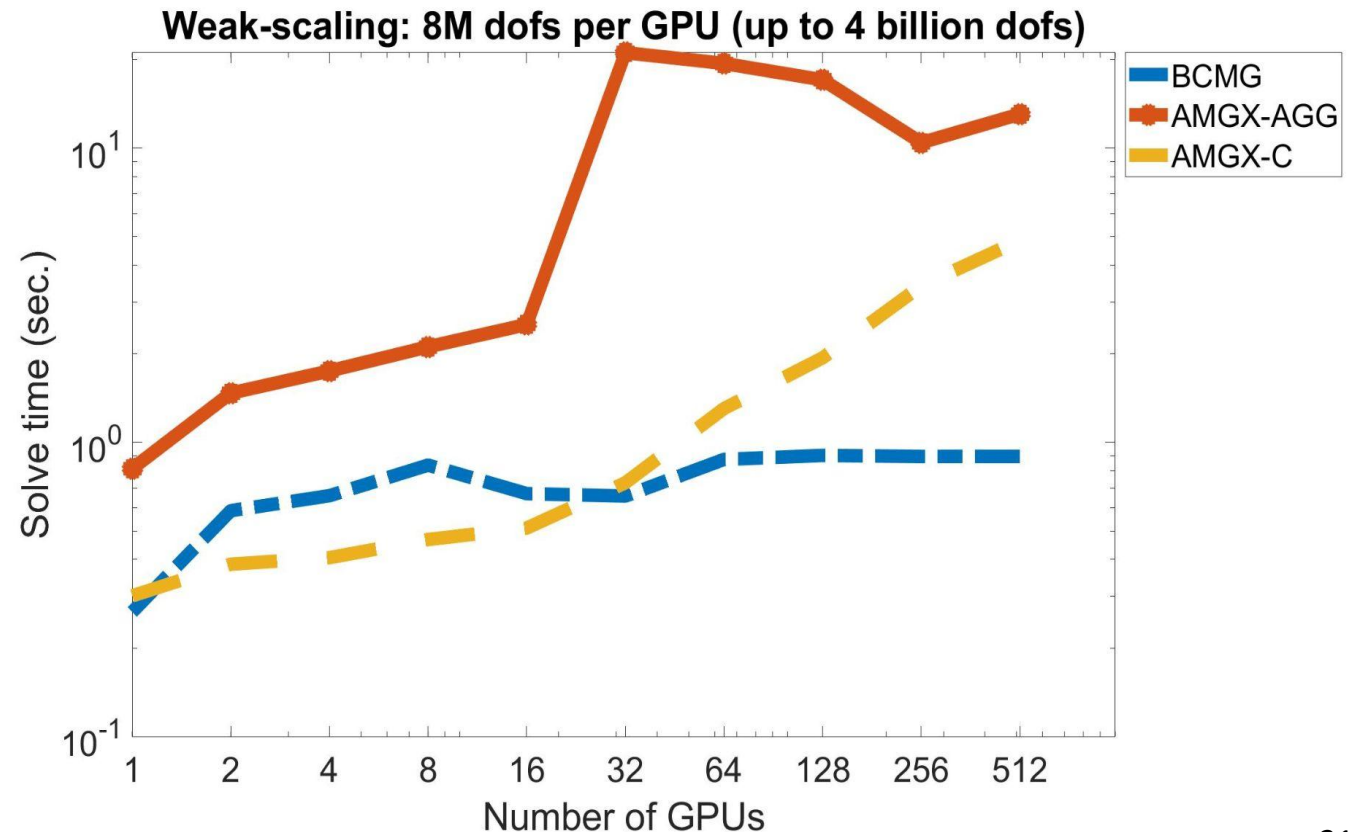


# Results

**textarossa**

# Math library (IDV-A)

- A communication-reduced CG solver for extreme-scale sparse linear systems on GPU-accelerated supercomputers
- An innovative AMG preconditioner (BCMG) developed in Textarossa shows significant benefits if compared with the state of the art (Nvidia AmgX library) on a HPCG-like benchmark



# TAFFO

- A suite of compiler passes integrated with LLVM to automatically tune the computation precision.
- Employs value range analysis from input data ranges provided by the programmer via attributes

Imagenet V2 classification accuracy with different precisions

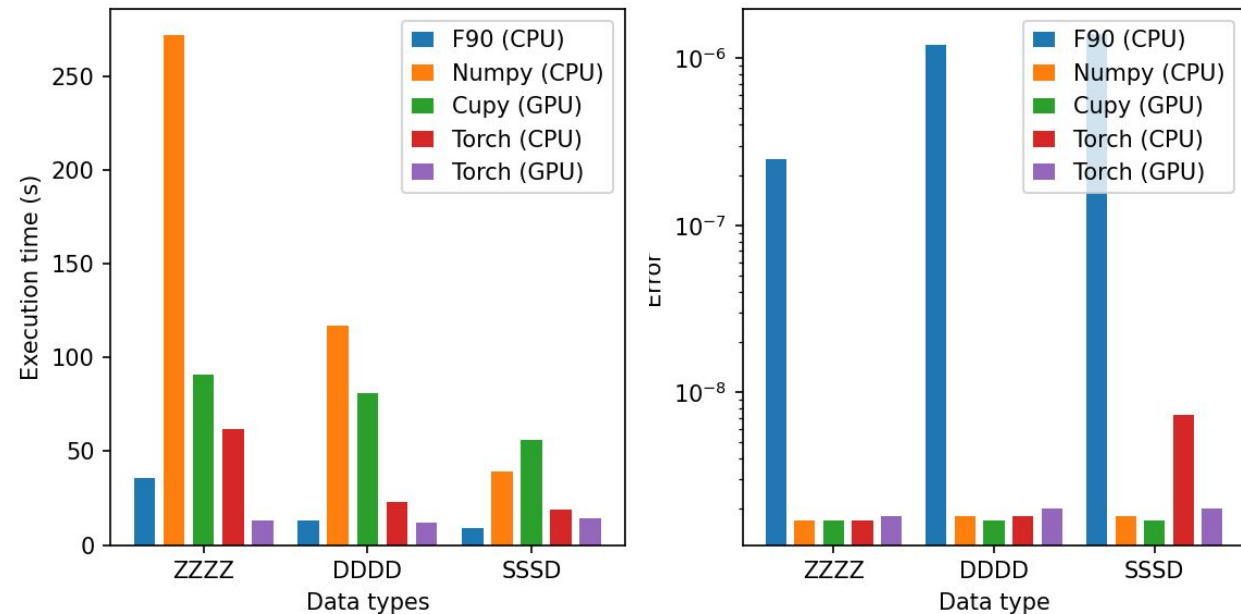
Numeric format	Accuracy (%)									
	Resnet50		Renet101		Resnet152		VGG16		VGG19	
	top1	top5	top1	top5	top1	top5	top1	top5	top1	top5
p16e2	<b>57</b>	75	<b>60</b>	77	<b>61</b>	<b>80</b>	<b>55</b>	73	52	<b>74</b>
p8e2 (mixed)	39	58	36	59	45	61	49	72	46	65
fp16 (e5m10)	48	67	52	76	43	65	54	<b>74</b>	<b>53</b>	73
fp32 (e8m23)	56	<b>76</b>	59	<b>79</b>	<b>61</b>	<b>80</b>	54	<b>74</b>	<b>53</b>	<b>74</b>

# Quantum TEA (IDV-A)

- Quantum TEA simulates quantum systems using Tensor Network Models

- Via precision tuning of the different layers of the TNM we can obtain performance gains while maintaining precision

Precision tuning (lower is better)



S: Single precision real, D, Double precision real,  
Z: Double precision complex

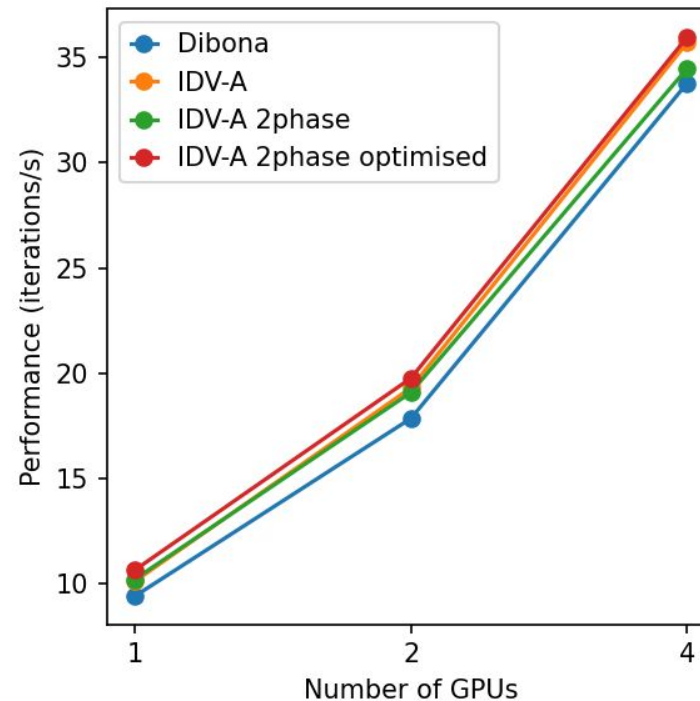


# UrbanAir (IDV-A)

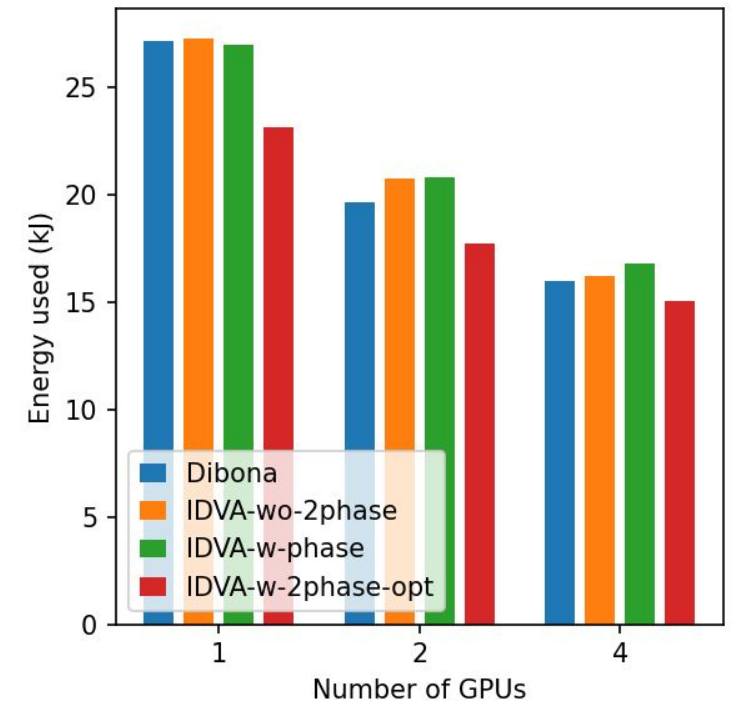
- Predicts air quality in in complex urban areas using a multi-scale model

- Parallelized using cuda
- 11% higher performance (IDV-A vs. Dibona)
- 25% further improvement by using mixed precision

GPU scalability



Energy (lower is better)

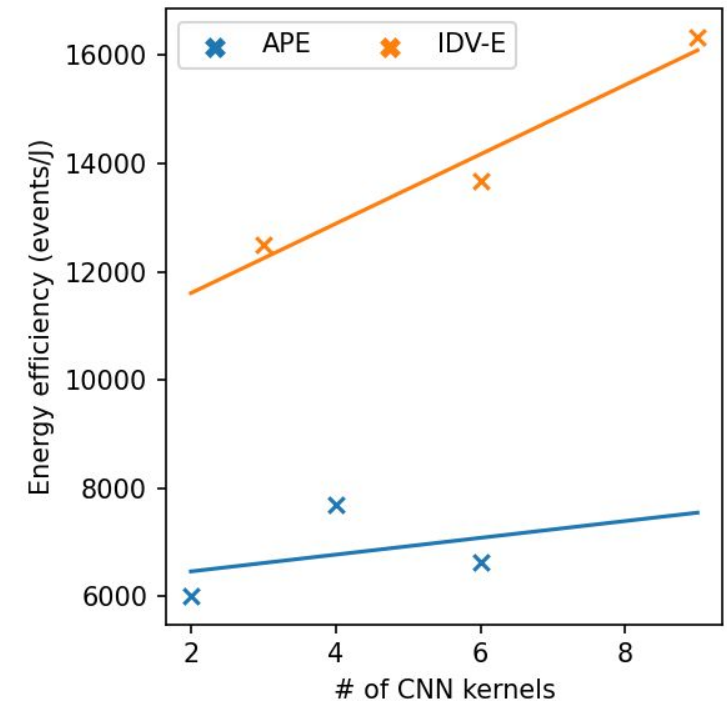
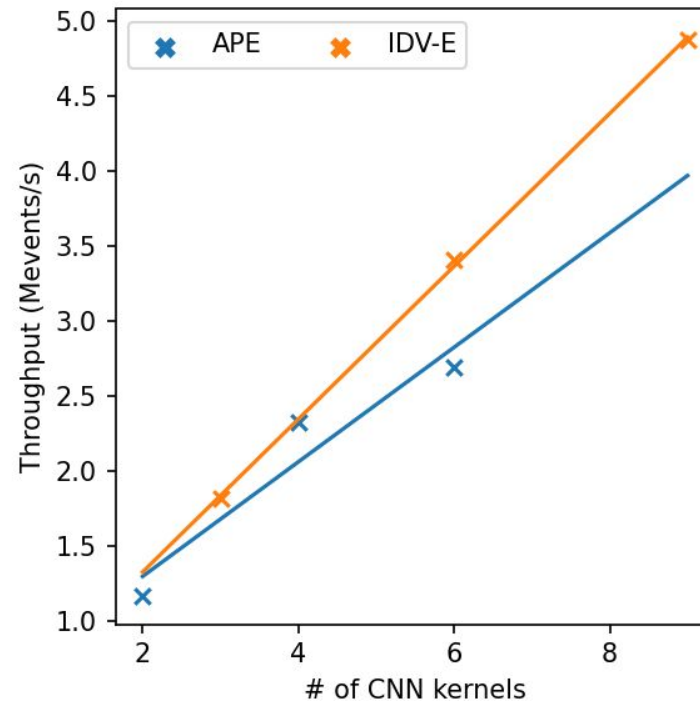


# Raider/APEIRON (IDV-E)

- Predict the number of charged particles in each RICH detector physics event (@10MHz) using a CNN on FPGA

- Distribute processing across multiple FPGA
- Use APEIRON for low-latency communications

Performance and energy efficiency (higher is better)

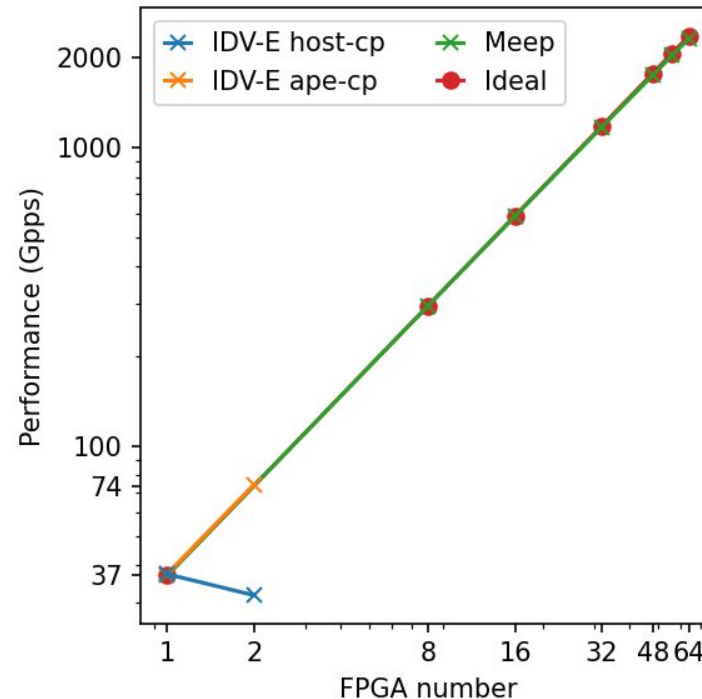


# Nbody/OmpSs (IDV-E)

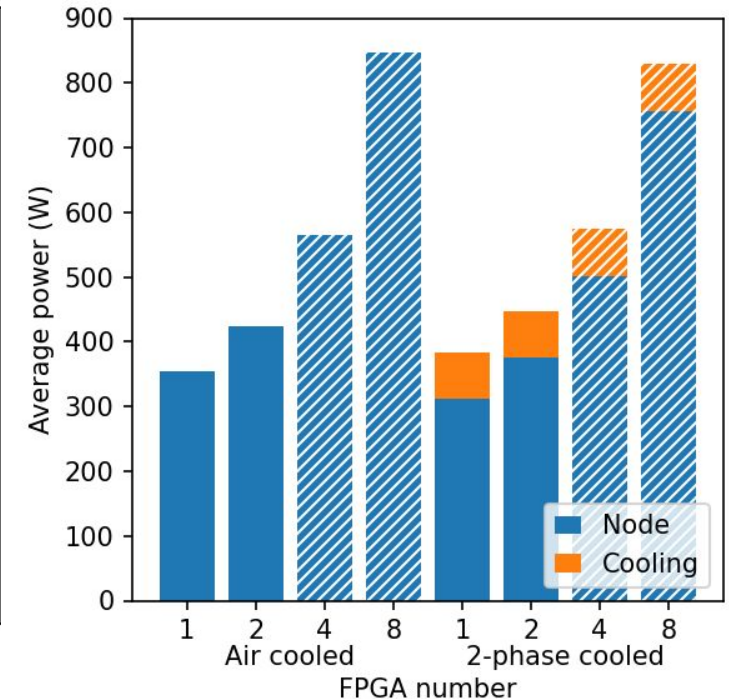
- Simulate gravitational interactions between particles

- Distributed processing using OmpSs@FPGA & OMPIF
- Near-ideal scalability in large number of nodes
- Energy efficiency improvements due to cooling technology

### Scalability



### IDV-E Power



# Conclusions

- Two-phase cooling is a new reliable and sustainable solution for effective thermal management of exascale systems.
  - It is feasible to effectively reject heat in warm climates without using chillers or cooling towers
- Thermal management has a big impact on energy efficiency
- Algorithm redesign is often required to exploit large scale complex architectures
  - Developer tools (programming models and toolchains) are key
- Multi-FPGA architectures are interesting for particular workloads
  - Fast inter-FPGA communication is needed
  - Programmability is still an issue
- Two-phase cooled IDV-A and IDV-E provide a starting point for greener HPC







**ENEA**

**Atos** CINECA *Inria*



université de **BORDEAUX**



UNIVERSITÀ DI PISA



UNIVERSITÀ  
DEGLI STUDI  
DI TORINO



Consiglio Nazionale delle Ricerche



POLITECNICO  
MILANO 1863



**Fraunhofer**



Istituto Nazionale di Fisica Nucleare

**E4**

COMPUTER  
ENGINEERING

**textarossa**

# The TEXTAROSSA Project: Cool all the Way Down to the Hardware

antonio.filgueras@bsc.es

Coordinator: massimo.celino@enea.it

[textarossa.eu](http://textarossa.eu)