EVIDEN
an atos business

Forum TERATEC

Accelerated Computing in the Exascale Era

Jean-Pierre Panziera
01/06/2023

Eviden ID & leadership in HPC

Circa €5B revenue

An Atos business that will become an independent company in late 2023

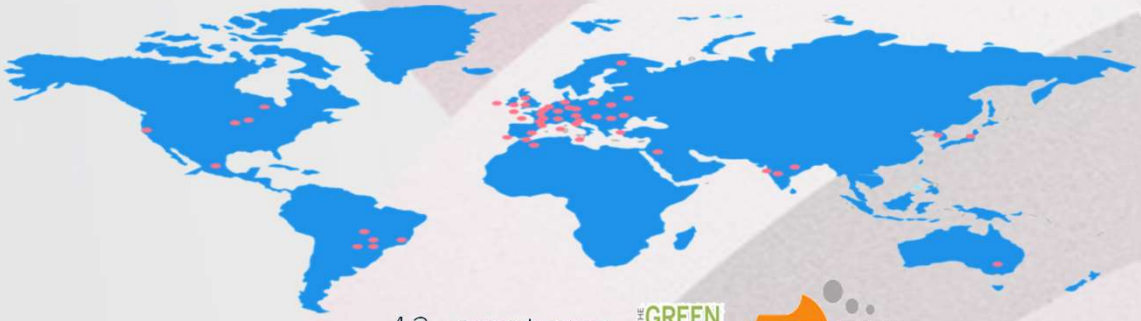
6 Business segments

Digital Transformation, Smart Platforms, Cloud, Advanced Computing, Digital Security and Net Zero

57,000 engineers and problem-solvers in 45 countries

Supercomputing leadership

#1 Europe, India and S. America
#3 worldwide



43 systems

THE GREEN 500

TOP 500 The List

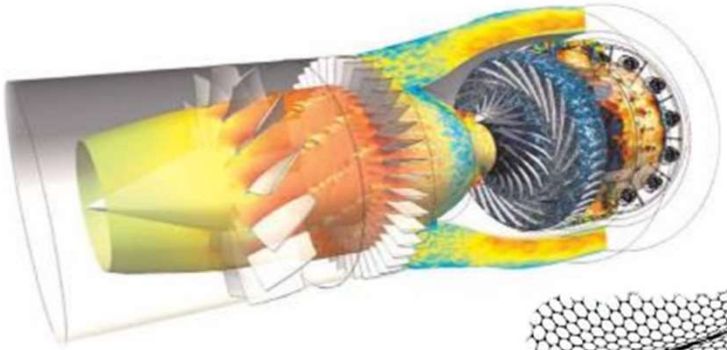
Solving the BIG challenges

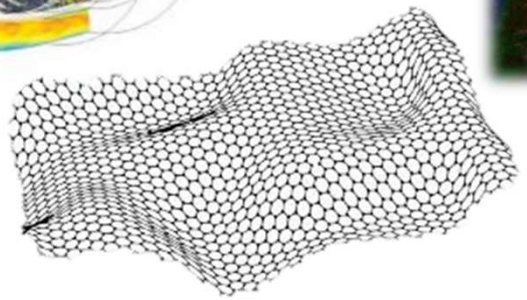
HPC to help meet BIG societal and industrial challenges

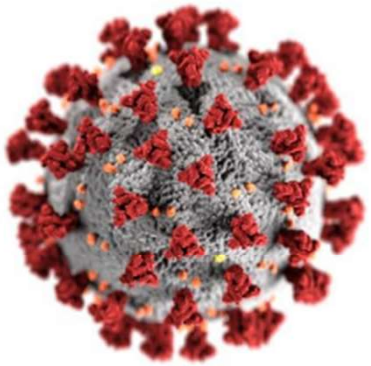
Public Science, top challenges:

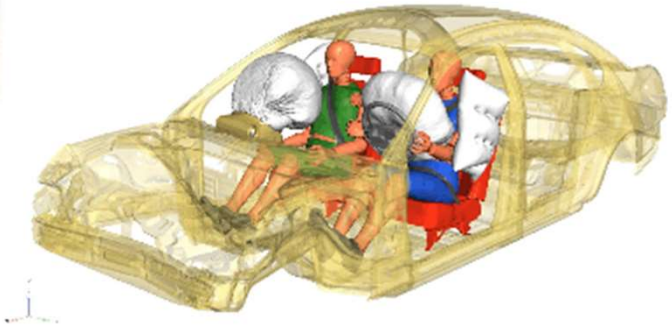
- Climate
- Personal Medicine
- Renewable Energy … and storage
- …

Industry use HPC for better/safer/cheaper products, new materials, new drugs …





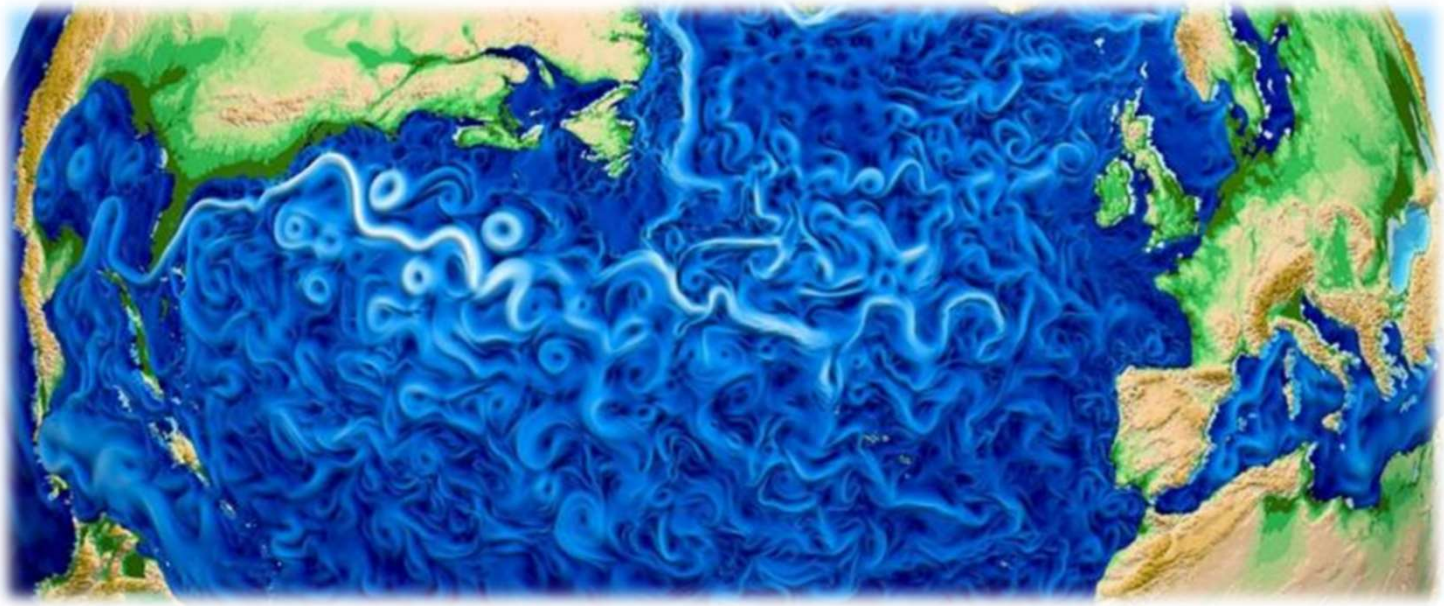




Need to scale solutions BIG times

E.g. Destination Earth project: building an Earth Digital Twin

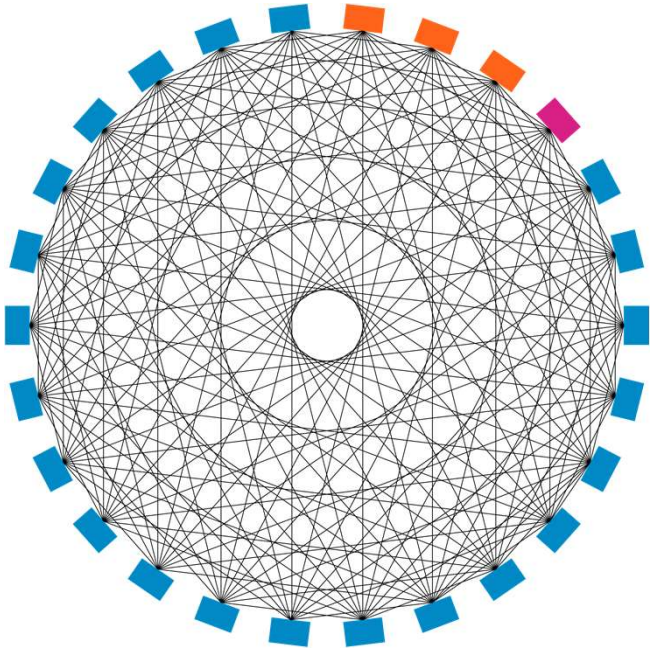
- Today ECMWF run “ensemble simulations” with a 9km resolution grid
- For an accurate solution(clouds, topography…), a 1km resolution is needed
- The 9x grid refinement translates into a **~700x** (9^3) performance improvement
- To close the gap, looking at **much larger systems** and **accelerators**



Simulation **Observations**

CINECA – Leonardo

Hybrid Pre-Exascale Supercomputer



10 ExaFlops FP16 AI workload

320 PFlops (Rpeak)

250 PFlops (Linpack)

4 PFlops (HPCG)

EuroHPC Joint Undertaking

"The Leonardo supercomputer is the result of our long-term commitment to pushing the boundaries of what a modern exascale supercomputer can be."
Sanzio Bassini, Director of the HPC department at CINECA

Atos

NVIDIA

intel. XEON

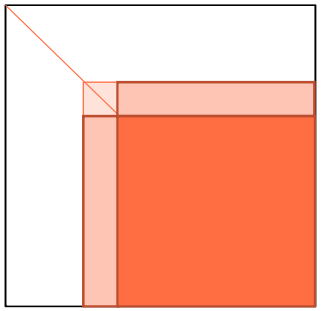
ddn

EVIDEN

Performance Flops vs Bytes

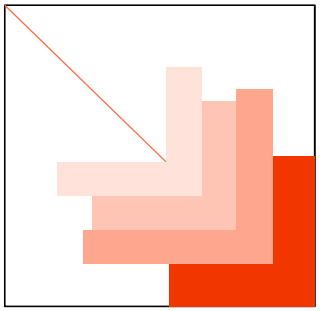
Solving PDEs with Linear Algebra Solvers

Direct solver
Dense Matrix
O(N^3)



Dense blocks
k^2 memory access
k^3 computation

Direct solver
Banded Matrix
O(N^2.7)

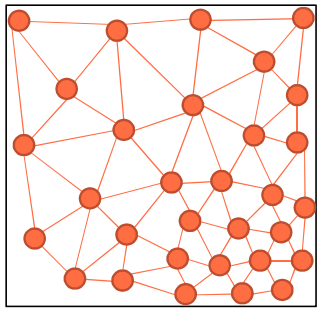


Dense blocks
k^2 memory access
k^3 computation

Linpack benchmark

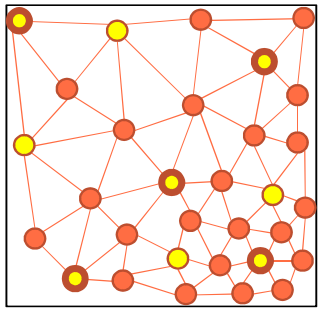
Core intensive

Iterative solver (CG)
Sparse Matrix
O(N^1.5)



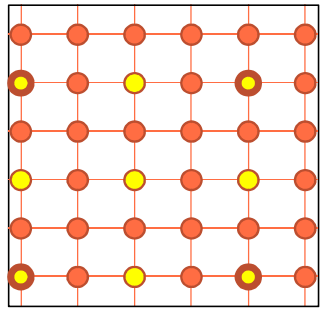
Sparse
N memory access
N computation

Iterative solver (AMG)
Sparse Matrix
O(N.log(N)



Sparse
N memory access
N computation

Spectral solver (FFT)
Regular grid
O(N.log(N)



Sparse
N memory access
N computation

HPCG benchmark

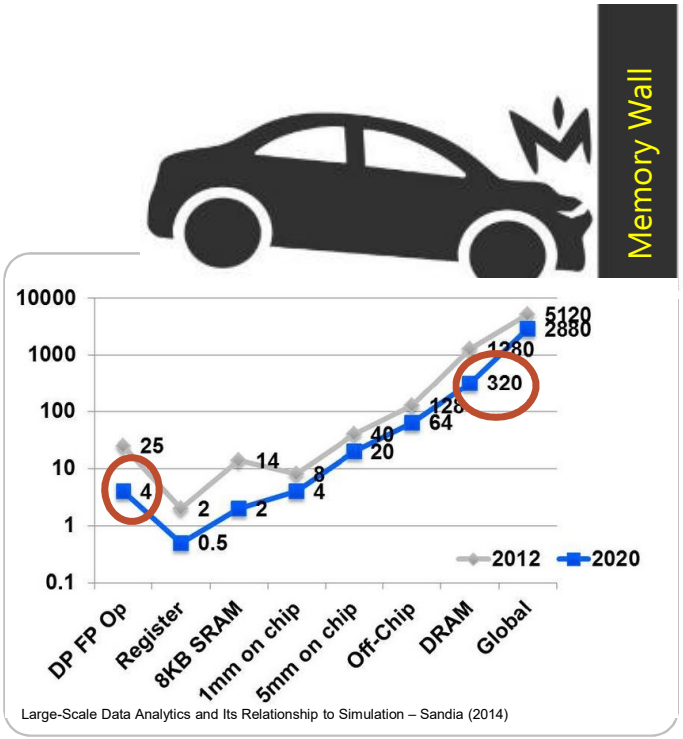
Memory intensive

FASTER computing elements

CPUs, GPUs, Accelerators …

- Top500 Flops rate are impressive: ExaFlops, 100s PetaFlops …
- … but HPC applications do not run at Flops speed
- HPC applications are limited by the **memory bandwidth**
- HPCG benchmark better represents application behavior

	Peak Pflops	HPL Pflops	HPL eff	HPCG Pflops	HPCG eff
Frontier	1 686	1 102	65%	14,1	1,3%
Fugaku	537	442	82%	16,0	3,6%
Lumi	429	309	72%	3,4	1,1%
Leonardo	256	239	68%	3,1	1,5%
Summit	201	149	74%	2,9	2,0%

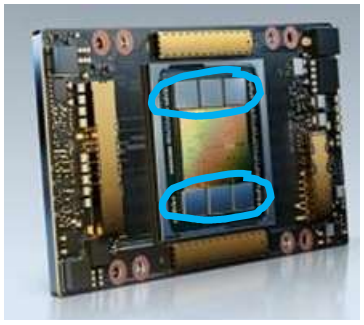


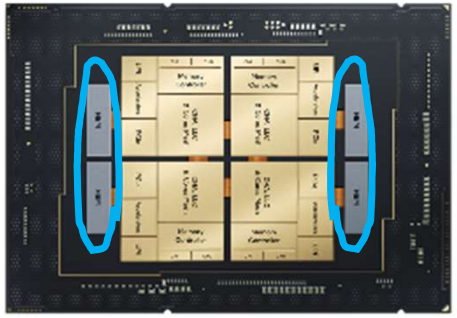
Large-Scale Data Analytics and Its Relationship to Simulation – Sandia (2014)

- GPUs offer highest BW: **2TB/s** for Nvidia A100 with HBM
- vs **200GB/s** for fastest CPU AMD Millan with 8xDDR4 … **10x**

But

- GPUs are "*challenging*" to program … you might get lucky
- CPU can integrate HBM memory to boost their memory BW hence performance


Nvidia H100


Intel SPR+HBM

EVIDEN

Faster memory

for higher performance

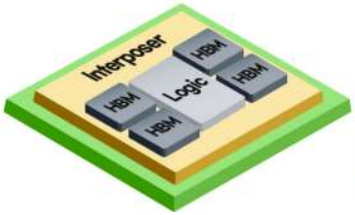
Various strategy to provide more memory Bandwidth (→ more performance to HPC applications):

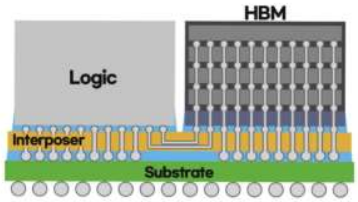
SDRAM:

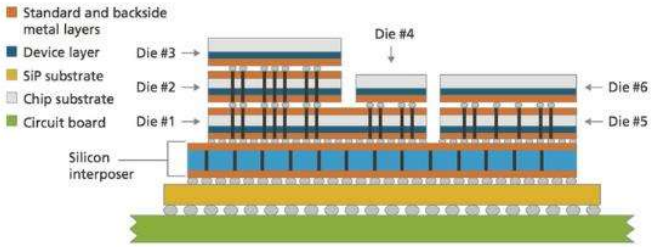
- Increase the number of DDR channels (4 → 6 → 8 → 12 … → 16?)
- DDR5: 4800MTs … 9600
- MRDIMM: 2x capacity & BW vs DDR5
- LPDDR5x (8533MTs … but with BGA)

HBM:

- HBM2, HBM2e, HBM3 in package

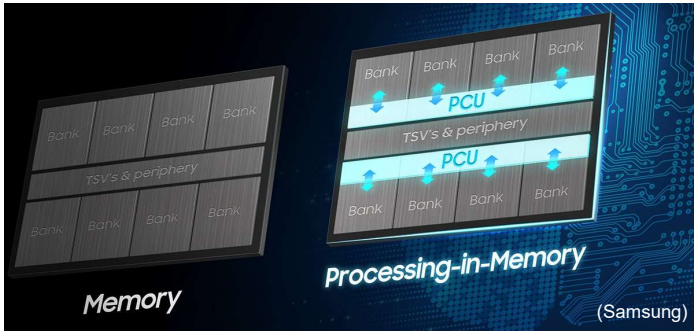






Next big thing:

- 3D stacking
- Memory on top of xPUs for 10-40x performance gain
- Processing-in-memory



AI, AI, AI

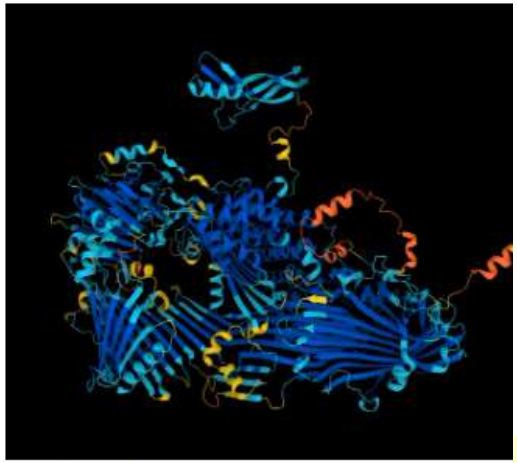
AI have BIG impact on some traditional HPC workloads

NEWS | 28 July 2022 | Correction 29 July 2022

‘The entire protein universe’: AI predicts shape of nearly every known protein

DeepMind’s AlphaFold tool has determined the structures of around 200 million proteins.

Ewen Callaway



The structure of the vitellogenin protein — a precursor of egg yolk — as predicted by the Alp tool. Credit: DeepMind

From today, determining the 3D shape of almost any protein known to science will be as simple as typing in a Google search.

communications biology

ARTICLE

https://doi.org/10.1038/s42003-023-04773-7 OPEN

Does AlphaFold2 model proteins’ intracellular conformations? An experimental test using cross-linking mass spectrometry of endogenous ciliary proteins

Caitlyn L. McCafferty[1,2], Erin L. Pennington[1], Ophelia Papoulas[1], David W. Taylor[1,2] & Edward M. Marcotte[1,2]

A major goal in structural biology is to understand protein assemblies in their biologically relevant states. Here, we investigate whether AlphaFold2 structure predictions match native protein conformations. We chemically cross-linked proteins in situ within intact Tetrahymena thermophila cilia and native ciliary extracts, identifying 1,225 intramolecular cross-links within the 100 best-sampled proteins, providing a benchmark of distance restraints obeyed by proteins in their native assemblies. The corresponding structure predictions were highly concordant, positioning 86.2% of cross-linked residues within Cα-to-Cα distances of 30 Å consistent with the cross-linker length. 43% of proteins showed no violations. Most inconsistencies occurred in low-confidence regions or between domains. Overall, AlphaFold2

… inconsistencies … in low-confidence regions or between domains

Huge projects

Generative AI

Large Language Models

LLaMA

a BigScience initiative

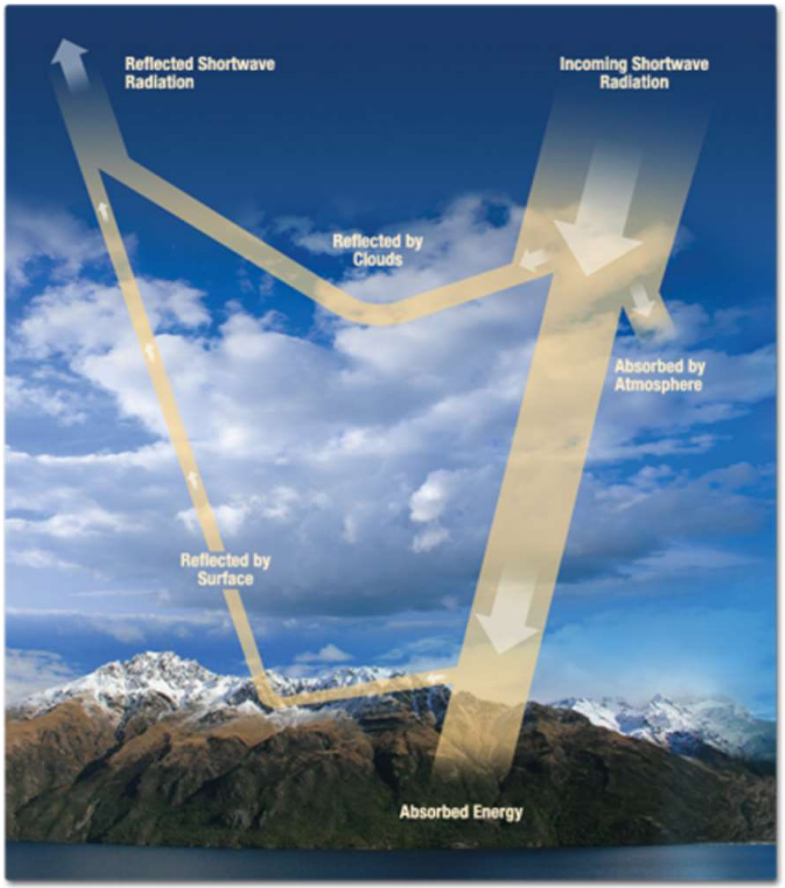
BLOOM

176B params · 59 languages · Open-access

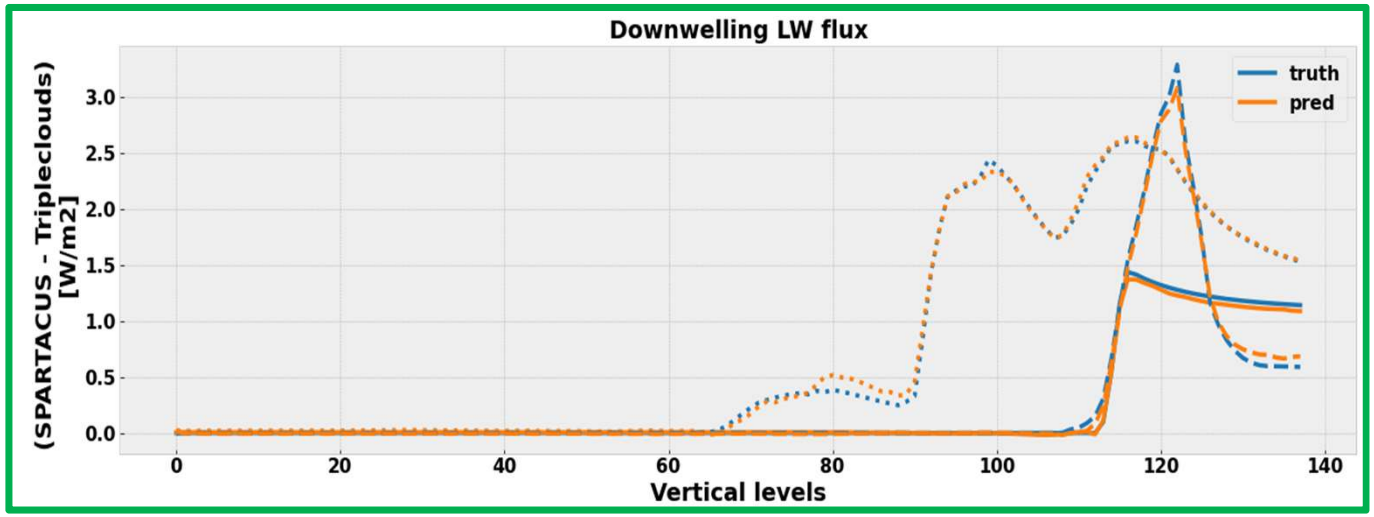
AI acceleration: Artificial Intelligence for Simulation

(AIAS'IM) framework

e.g. Improving the weather forecast radiative scheme with machine learning



Emulate a complex and costly simulation of 3D radiative effects of clouds with simpler model and **AI** correction



Christophe Bovalo, Rémi Druilhe – *Atos/Eviden*
Matthew Chantry, Peter Düben – *ECMWF*

Good fit ☺

Quantum Computing

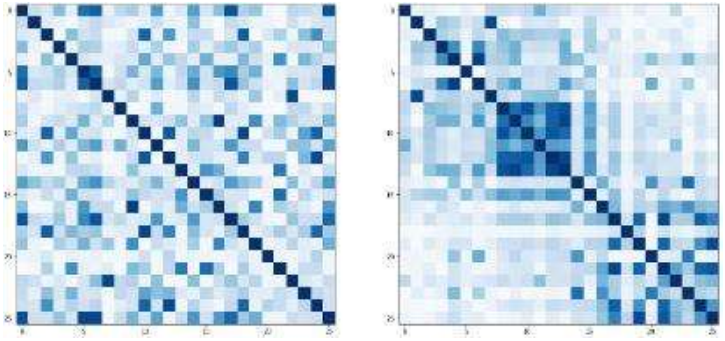
QLM (Quantum Learning Machine) and soon real Qbits

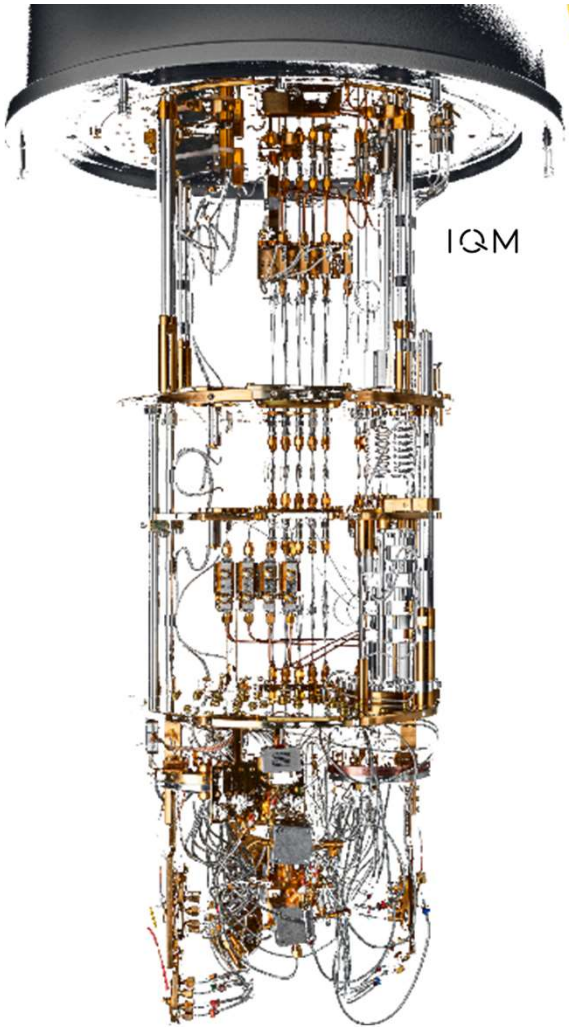
3 main applications domains:

- Universal Quantum Computer
  Solving any inverse problem, including Integer Factorization (RSA)

- Quantum for Quantum problems
  Emulating molecules on Quantum devices

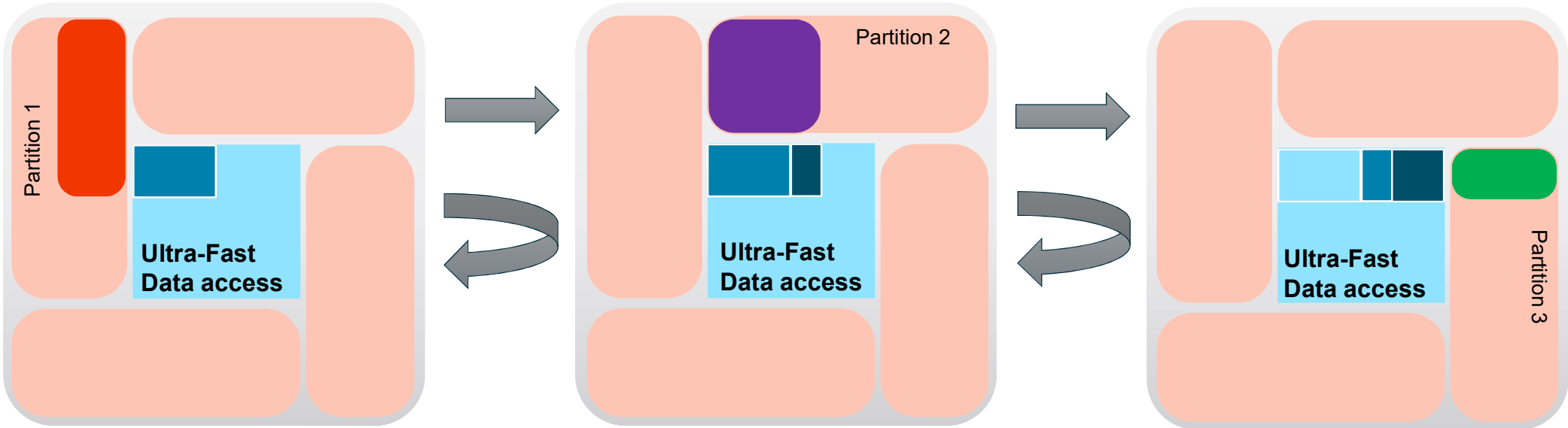
- Combinatorial optimization
  Finance, Energy grid, Crew assignments
  Software solutions, e.g. Quantum Approximate Optimization Algorithm (QAOA)

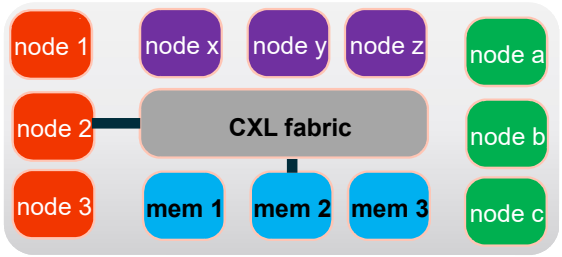


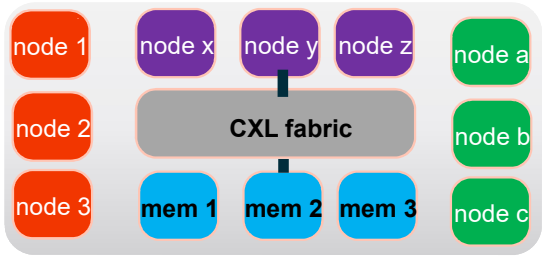


Multi-Partition systems for HPC workloads

Ultra fast “disaggregated” Data access







Disaggregated architecture with CXL/PCIe protocol for

Reconfigurable

Ultra-Fast – IO style – data access

Accelerated Computing in the Exascale Era

The next BIG thing in Exascale HPC is :

- Data, Data, Data …
- Exascale Bandwith → orders of magnitude improvement required
- Energy efficiency → Specialized accelerators / Multi-Partition systems
- New Algorithms for HPC Applications → Includes AI software acceleration
- New computing paradigms → Analogic, Neuromorphic, Quantum …
- Data, Data, Data …

EVIDEN

EVIDEN Atos

Forum TERATEC

Thank you

jean-pierre.panziera@atos.net
jean-pierre.panziera@eviden.com