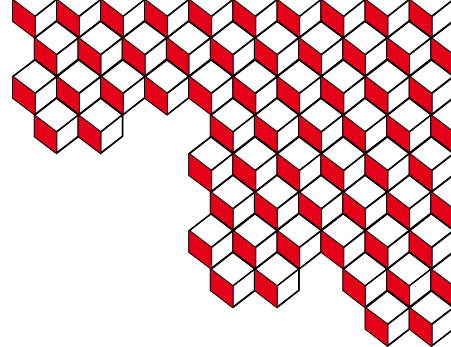




Introduction sur le calcul intensif (+ Alice Recoque)

Pascal Tremblin — Maison de la Simulation / CEA
(+ équipe projet AST)

Septembre 2026



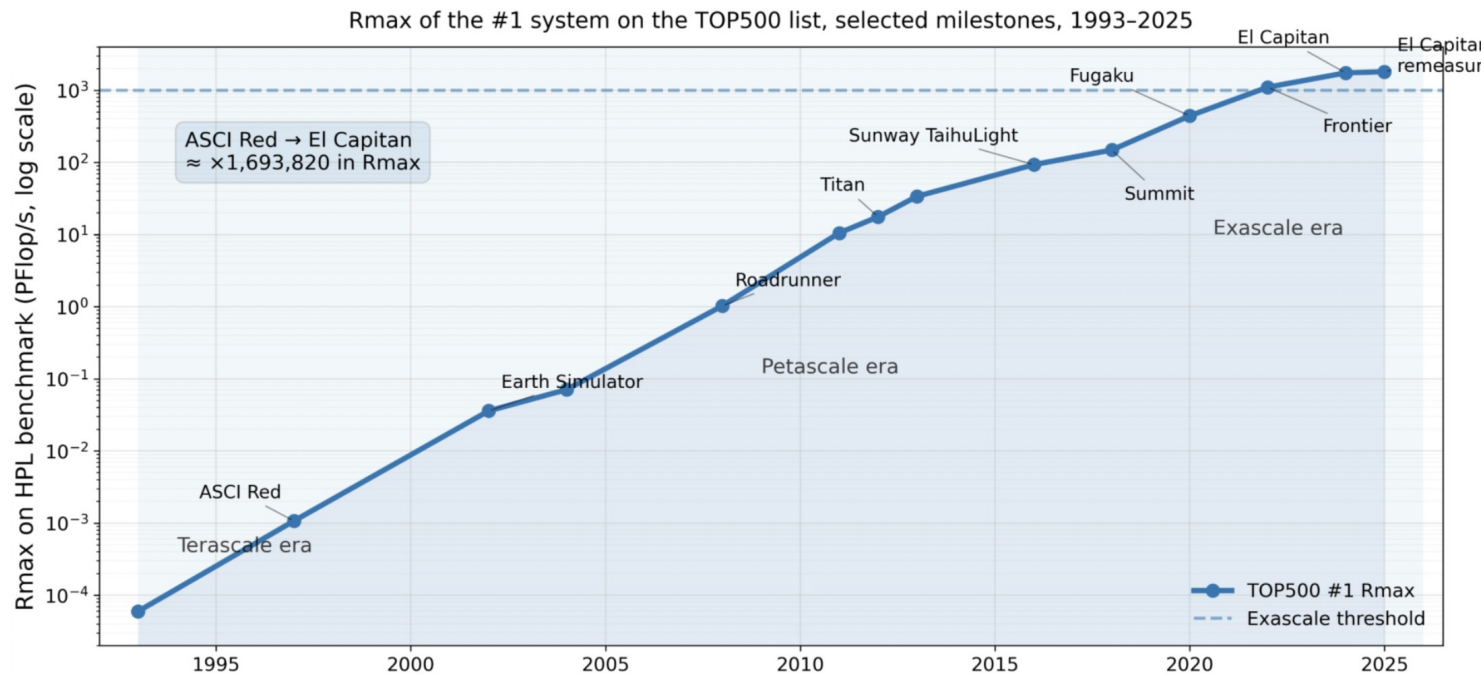
Le blob

Une interview d'un (très 😊) grand spécialiste du domaine
(pendant la canicule)

<https://www.youtube.com/watch?v=0qr0FnkHP54>



Du téraflop à l'exaflop en trois décennies



1997

ASCI Red
 ≈ 1 Tflop/s

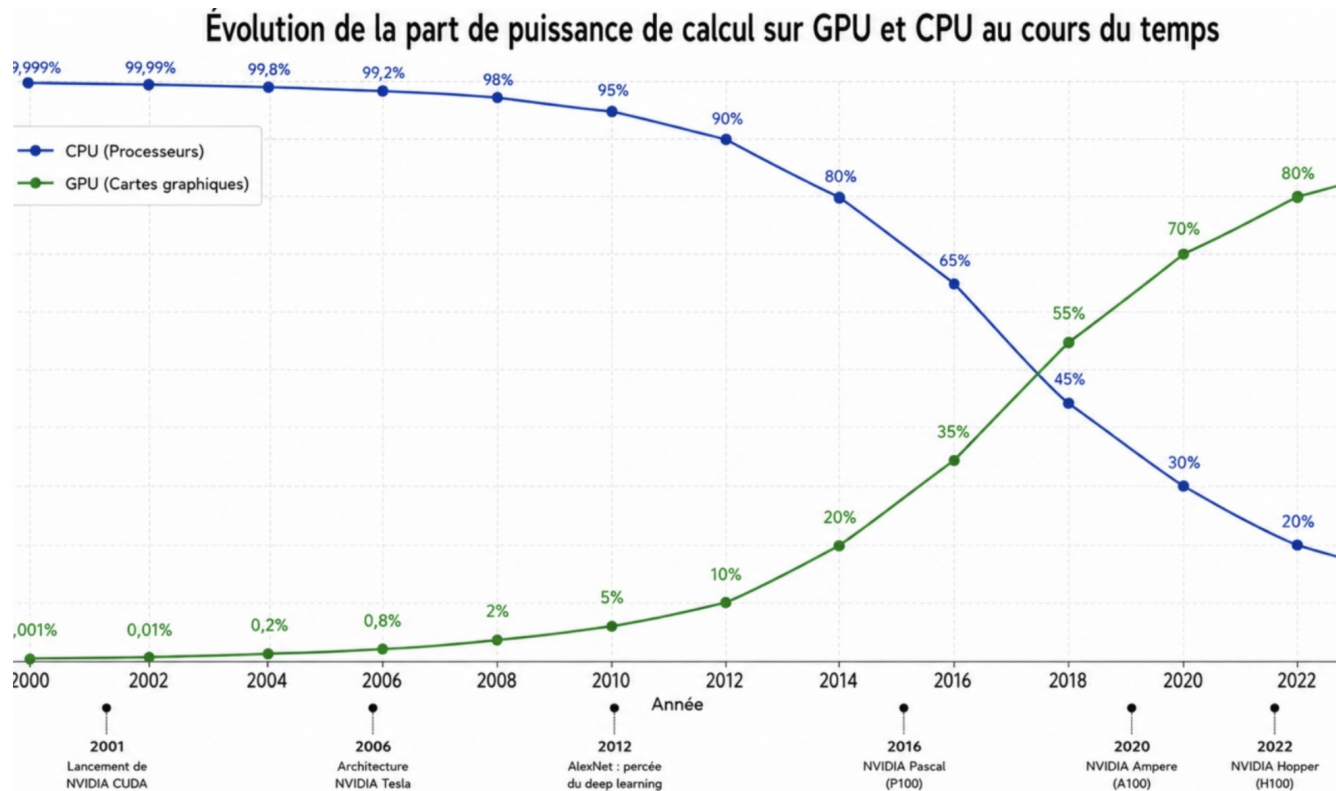
2008

Roadrunner
 ≈ 1 Pflop/s

2022+

Frontier puis
ère exaflopique

Avec un changement majeur d'architecture : les GPU sont devenus centraux (mais ...)

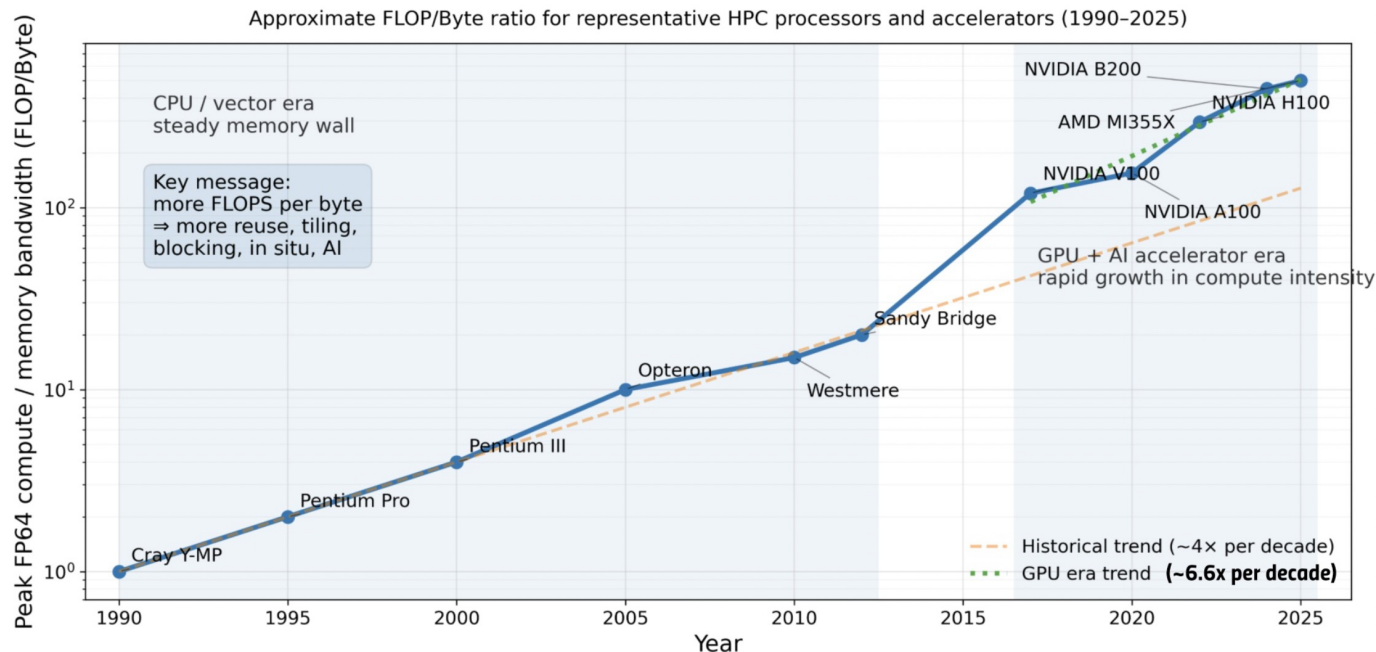


anford AI Index Reports (2019, 2021, 2023)

ports semiAnalysis, Epoch AI – Estimations basées sur les FLOPs utilisés pour l'entraînement des modèles d'IA de pointe

Avec un changement majeur d'architecture : les GPU sont devenus centraux

Compute-to-memory balance keeps rising

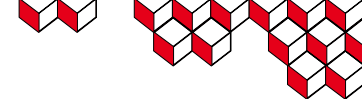


1. Beaucoup plus de parallélisme

2. Le calcul progresse plus vite que la mémoire

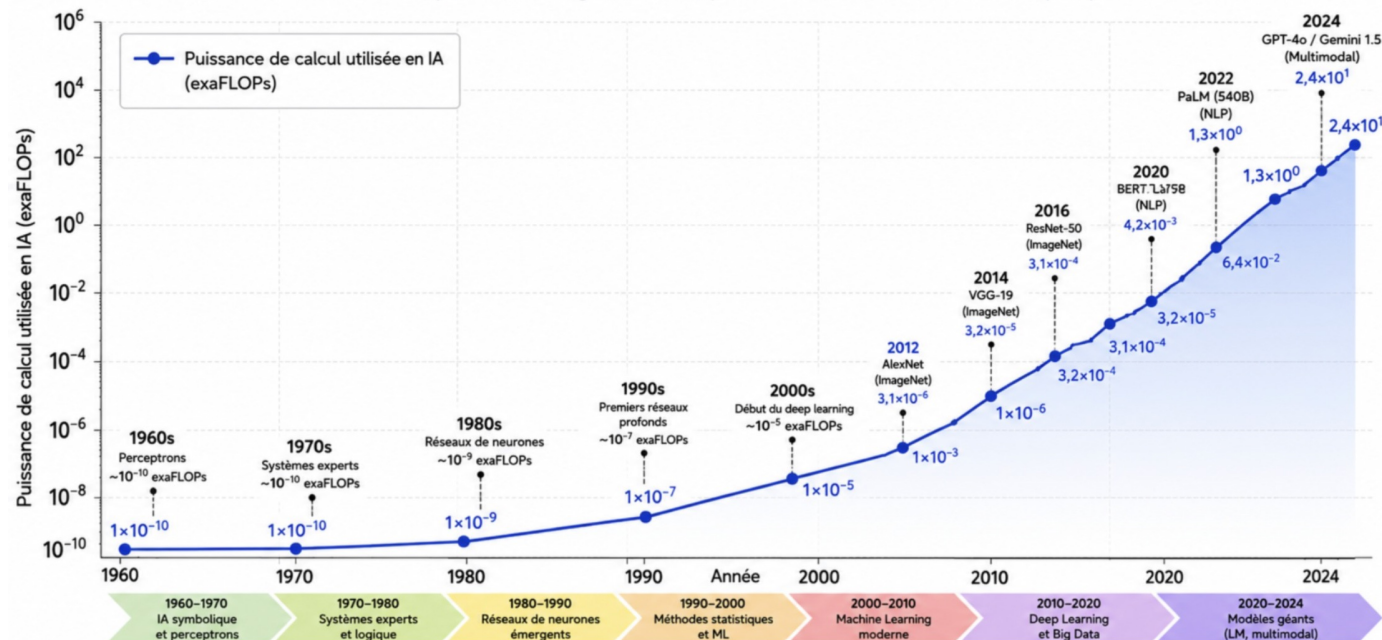
3. Il faut donc réutiliser les données et minimiser les mouvements

HPC, données et IA convergent



Évolution de la puissance de calcul utilisée en IA au cours du temps (depuis 1960)

en exaFLOPs (10^{18} opérations en virgule flottante) pour l'entraînement des modèles les plus performants



Notes : exaFLOPs (EF) = 10^{18} opérations en virgule flottante. Les valeurs représentent des estimations de la puissance de calcul totale utilisée pour l'entraînement des modèles d'IA les plus avancés chaque année.

Sources : AI Index Report (Stanford, 2019-2024) – Epoch AI Compute Database – OpenAI, Google DeepMind, Anthropic, Meta AI publications

Les entraînements de pointe consomment désormais des ressources plus importantes que des très grandes simulations.

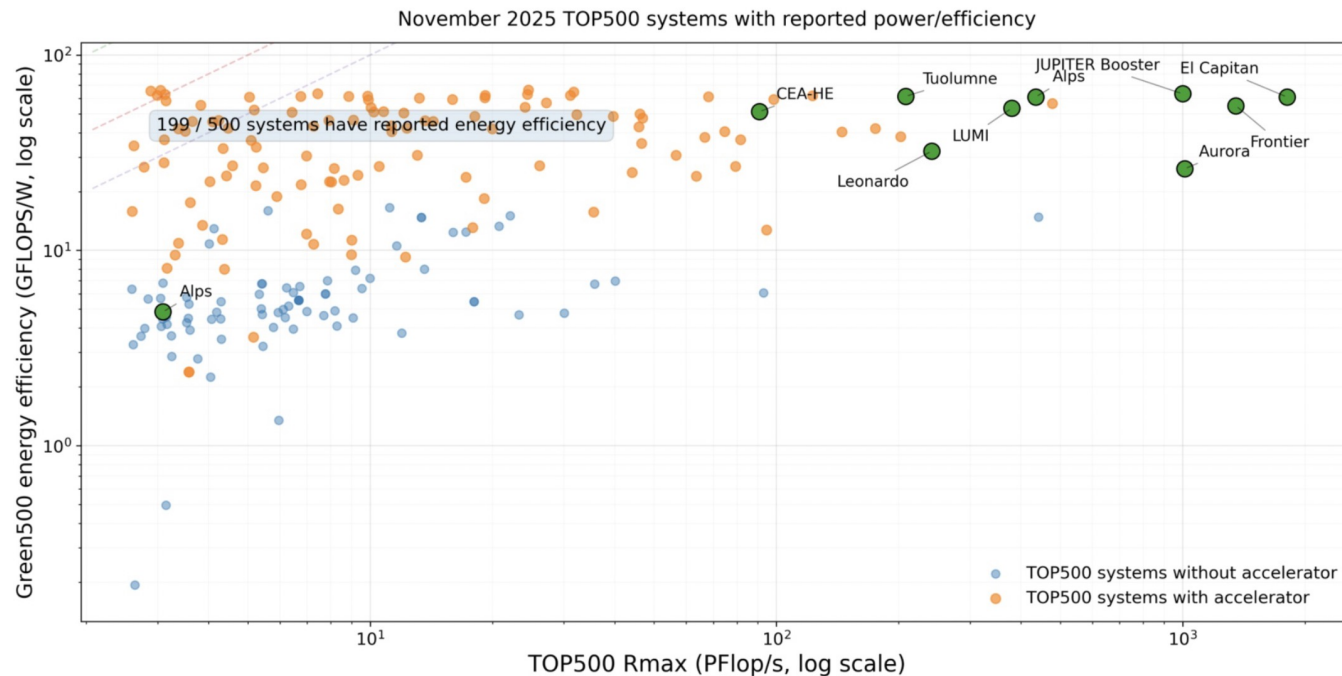
Les mêmes accélérateurs servent à la simulation, à l'IA et à l'analyse de données.

Deux grosses tendances AI4Science :

- Modèle de fondation pour les sciences
- Workflow agentique couplant HPC/IA

La contrainte énergétique devient structurante

TOP500 vs Green500: performance and energy efficiency



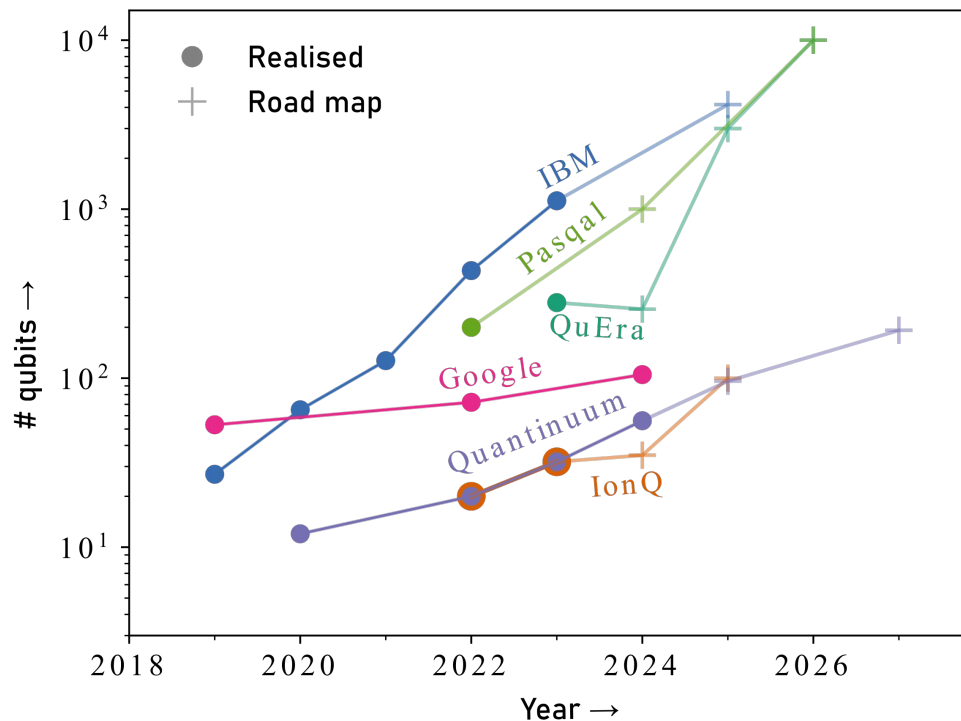
Le coût énergétique limite la puissance installable.

Les accélérateurs dominent les systèmes les plus efficaces.

Optimiser = réduire le temps de calcul, mais aussi les transferts et les accès mémoire.

**Performance
/ watt**

Emergence du quantique ?



POINT CLÉ

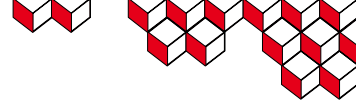
Comparer uniquement le nombre de qubits est trompeur : fidélité, connectivité et correction d'erreurs comptent aussi.

2025

Quantinuum Helios : 98 qubits physiques et démonstrations de qubits logiques corrigés / détectés.

AU TGCC

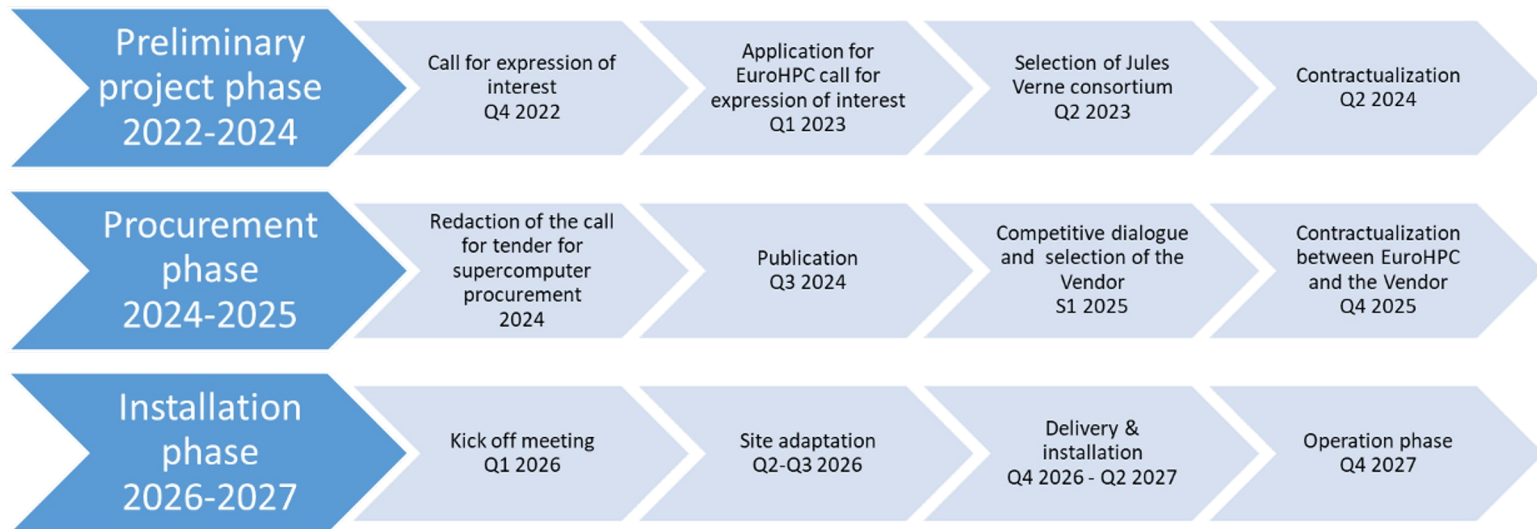
Ruby / Lucy : systèmes à atomes neutres et photoniques. Arrivée prochaine de Kitty (Alice&Bob)



ALICE RECOQUE
BEYOND HPC

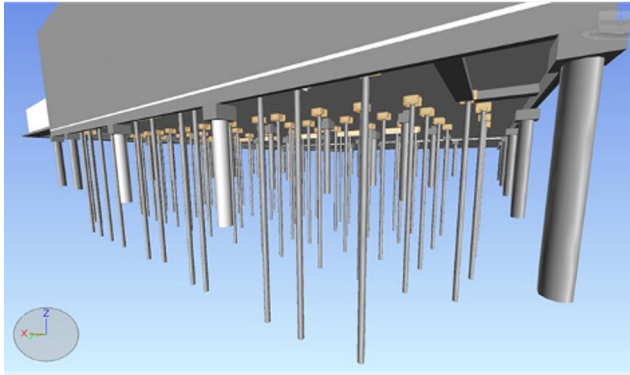
Overall timeline

- A multi-year project that began in 2022 !
- Anticipation of the adaptation work of TGCC (Very Large Computing centre of CEA)

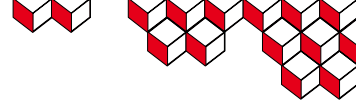


AR infrastructure

- Increasingly dense cabinet => need to reinforce the concrete slab ($1 \text{ t/m}^2 \rightarrow 2.8 \text{ t/m}^2$)
 - In the basement: Creation of a metal frame to sustain the computer room floor (67 micropiles)
 - In the computer room: Add local reinforcement, application of a resin to waterproof the slab
- Done while having many HPC systems in production (including Joliot Curie) !



Where are we today ?



- **Computer room ready to host** Alice Recoque supercomputer
- Final water and power distribution to be realized by Bull
(started Q3/26)
- Delivery of first racks planned in Q4/26
- Alice Recoque ready for production end 27

Key contributors

Jules Verne consortium is composed of

- France represented by GENCI (Hosting Entity) and CEA (Hosting site)
- The Netherlands represented by SURF
- Greece represented by GRNET



Alice Recoque will be delivered

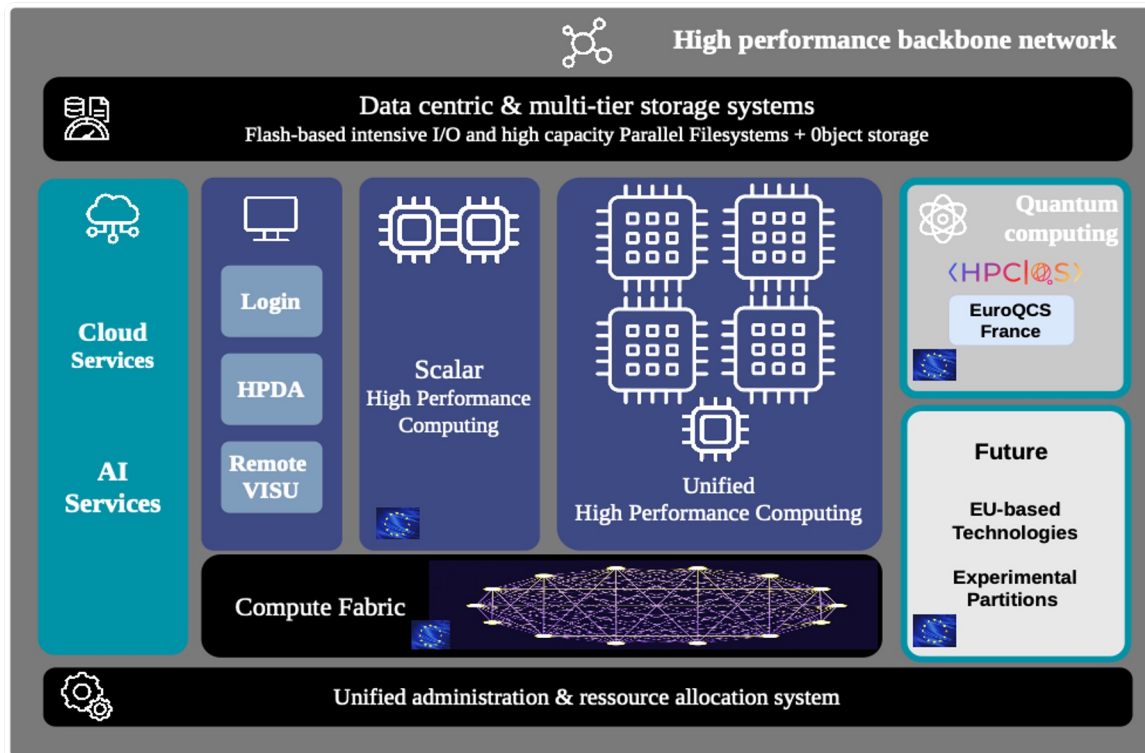
- by Bull
 - with AMD unified partition
 - and SIPEARL CPUs



Funding by

- EuroHPC
- France 2030 and French Ministry of Research for France
- Ministry of Education for the Netherlands
- Ministry of Digital Governance for Greece

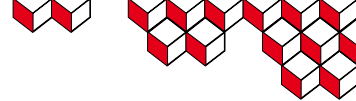
Overall architecture



Modular architecture addressing :

- Needs of HPC/AI/QC communities
- New services
 - Resources combination for complex user scenarios
 - Relying on CEA cybersecurity expertise

CPUs & GPUs for HPC and AI



AMD EPYC “Venice” CPU

- x86 Zen 6 arch, 256 cores
- 4 GB/core => 1TB/node
- CPU/GPU unified memory
- Interconnect BXlv3 400Gb/s per CPU
- CPU socket for scalar and accelerated workflows => **Multi tenant mode**

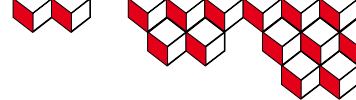
AMD Instinct MI430X GPU

- 288 Tflops FP64 - 432 GB HBM4 memory
- High level performance of native FP64
- Advanced AI capabilities
- Previous generations deployed at CINES (FR), LUMI (FI), LLNL, .. (MI250X / MI300A)
- Interconnect 2x BXlv3 = 800Gb/s per GPU

SiPearl Rhea2 CPU partition : ARM architecture

BULL's newest BullSequana XH3500 platform

Memory and Bandwidth



Unified Node

- 4x432Go HBM4 @ 23,3 To/s (GPUs) + 1 To RDIMM@8000 (CPU)
- Memory coherent access GPUs/CPU
- Enabling large AI models & HPC

Interconnect

- High injection BXIV3 4 x (2x400) Gb/s (GPUs) & 400 Gb/s (CPU Venice)
- Dragonfly topology with low blocking factor

I/O

- LUSTRE & Object Storage (on going procedure)

AST-AR: Application Support Team Alice Recoque



A machine – along with advanced support for applications



+

User support level 3

“[...] for **porting** and the **optimization** of HPC/AI applications, providing **training** services, with support lasting **few days to few weeks, or even longer** per request [...]”



AST – AR

AST-AR: a collaborative approach

- Coordination



- Technical expertise



- Effort : 12 – 14 FTE

- Companion projects

- Partners

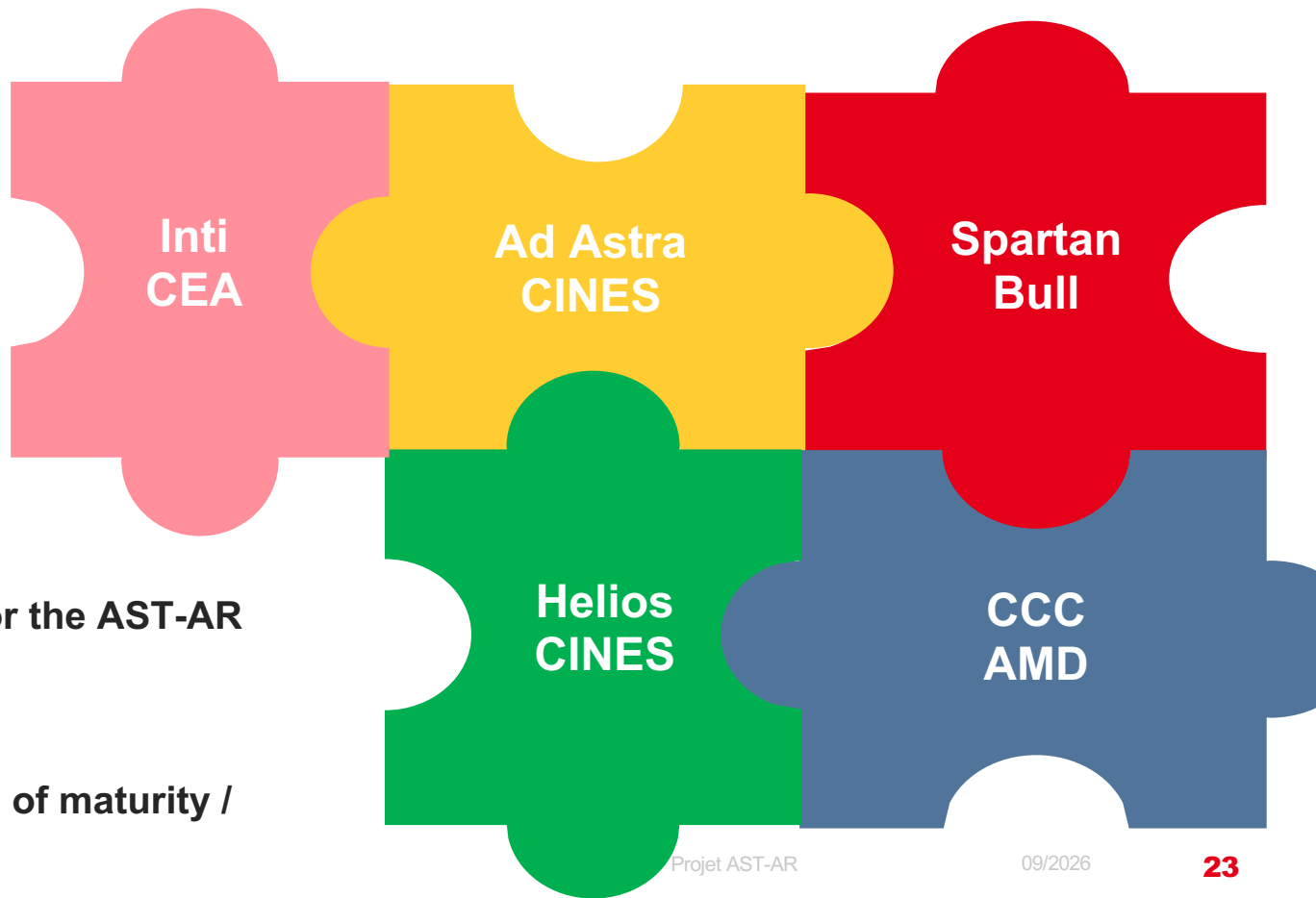


RESEARCH
PROGRAM

DIGITAL
FOR EXASCALE



AST infrastructures



Set of HPC clusters for the AST-AR team, based on

- different platforms
- diverse hardware
- and different levels of maturity / availability

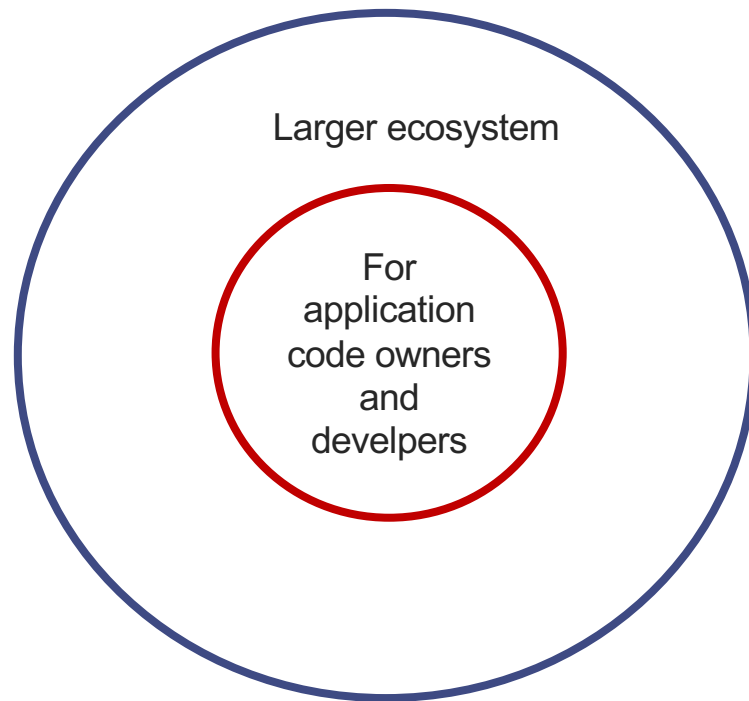
Service offering

For code owners and developers

1. Support for porting and adapting application codes
2. Training and exchange via a network of experts (AMD, Bull, ...)

Larger ecosystem

- Training (including hackathons)
- Network of experts
- Seminars and workshops



Ad 1: Support for application codes



Domain	Description
HPC	« Classical HPC issues » Parallelisation, porting of code to GPU/ new hardware
IA	AI-specific topics Data preparation, AI-specific tool deployments , pre-training of models, fine tuning, ...
HPDA	Data specific topics Data organisation and efficient data access, pre/post-treatment from for AI and HPC applications, covering HPC storage as well as Cloud based storage solutions

Categories	Description
N3.1 – Analysis and recommendations	Main deliverable: technical report With detailed analysis on performance, possible bottlenecks, along with technical recommendations
N3.2 – Integration and optimisation	Main deliverable : adapted application code Integration, reconfiguration or adaption of the application code I
N3.3 – Porting and rewriting	Main deliverable: technically improved application code Application code adapted to new architecture, programming paradigm or parallelisation approach

Ad 2: Training offering

- Rich French ecosystem to build on - void « reinventing the wheel »

- **HPC**

- Advanced usage of Alice Recoque supercomputer
- Efficient parallel IO
- Node level performance engineering
- Parallel programming OpenMP and MPI
- Containers usage on supercomputer

- **DATA – IA**

High performance with Python, R , Julia

- **GPU**

- GPU programming basic training
- GPU programming advanced trainings: Kokkos, Cuda, OpenACC, OpenMP, ROCm

- **IA**

- IA for dummies
- IA with Alice Recoque supercomputer

How does it work ?

1. Evaluation phase of the support request. If positive:
2. Detailed analysis
 1. Jointly with the code owners, organisation of the work to be done in subtasks and short development cycles (2 – 4 weeks at most)
 2. Agree on an objectives for the next 3 months
3. A team (2) of AST engineers assigned to the project
 1. They are supported by a senior engineer (“referent”)
 2. Weekly meetings with the senior engineers and the code owners and their engineers
 3. The AST engineers work jointly with the other application developers in the main code base
4. Joint training and synchronisation sessions amongst all AST engineers and all senior engineers

How to get support form the AST-AR team ?

- **Later:**
 - Via access request to the Alice Recoque machine (which can include AST-AR support)
 - Via the eDari portal of GENCI (for the French community)
- **Today**
 - Focus on some priority codes selected by GENCI's partners
 - If you seek support for your application to run on AR GPU, contact bruno.froge@cea.fr, chloe.azria@cea.fr, or maike.gilliot@cea.fr



Equipe projet AST
CEA/DAM- IdF

Emplacement logotypes financeurs/partenaires