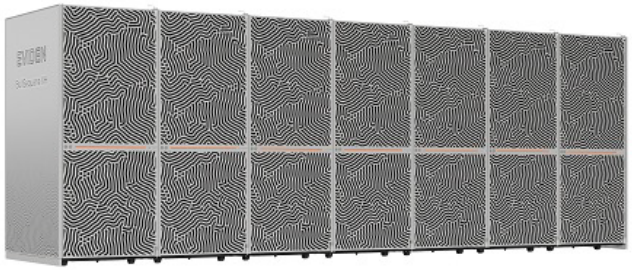
GENCI

AI in France



GENCI, A FRENCH HPC RESEARCH INFRASTRUCTURE

Serving yearly 1700 research projects in HPC and AI (academia, industry)

TGCC/CEA - Ile de France

- **Hosting Site** for the 2nd Exascale system (**EuroHPC**) within Jules Verne consortium (FR, NL)
- Hosting Site for the **1st hybrid HPC + Quantum computing infrastructure** (HQI, HPCQS, EuroQCS-France)

IDRIS/CNRS - Ile de France

- **1st FR converged HPC/AI system** *(#AIForHumanity)*
- Bring **sovereign** computing facilities / services to AI research community
- **>1000 yearly projects in AI allocated in 2023 !**
- **> 3700 GPUs in 2024**

CINES/FU - Montpellier

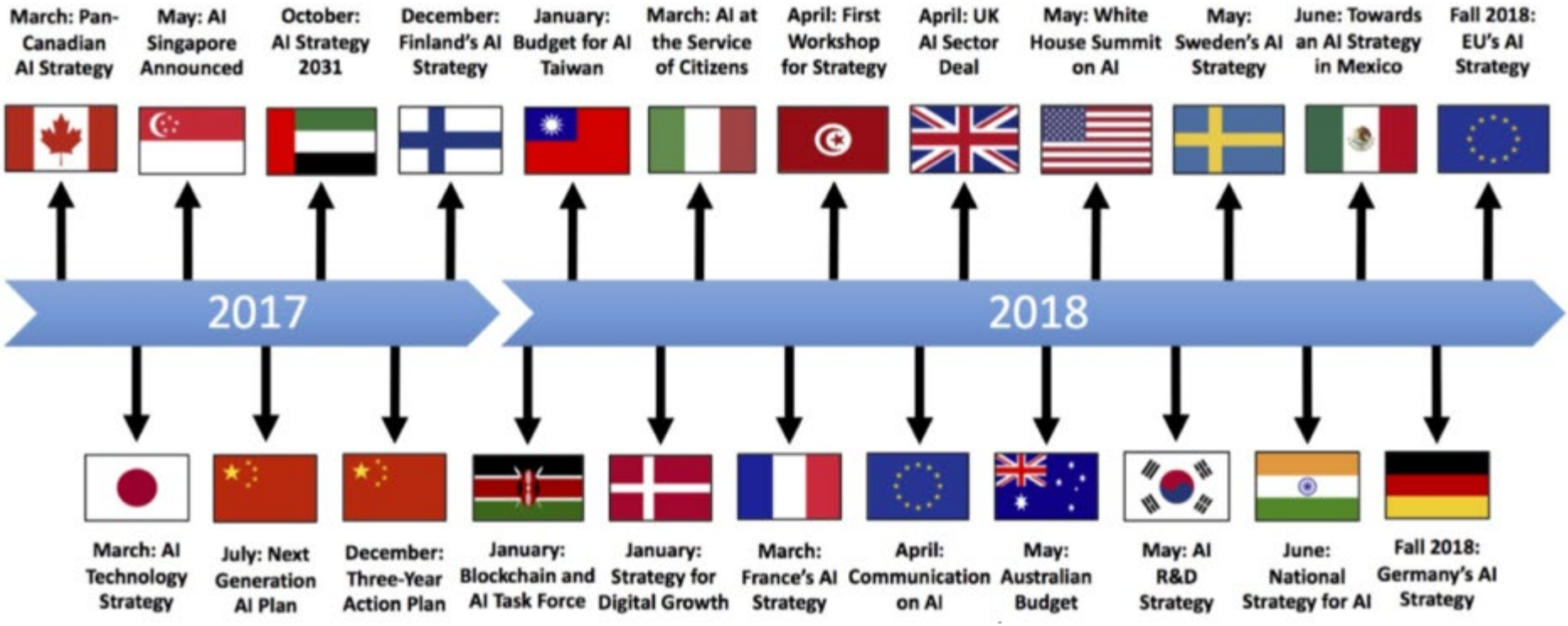
- **> 70 PF** with AMD next gen GPUs (>1400) & CPUs (>100k)

- **Next step before French Exascale system**

#3 THE GREEN 500

THE BEGINNING OF THE STORY FOR US

First AI announcements

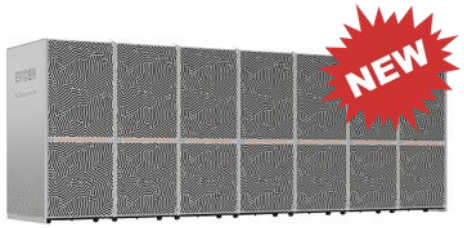


AI FOR HUMANITY
L'INTELLIGENCE ARTIFICIELLE AU SERVICE DE L'HUMAIN.

Phase #2 in 2023 toward Gen AI

ET VOILA JEAN ZAY AT IDRIS (CNRS) !

A converged supercomputer for HPC & AI



❑ Objectives

- Bring sovereign HPC facilities for the FR AI community
- Foster collaboration between HPC and AI

❑ **Converged** supercomputer

- HPC, AI & HPC/AI > 3700 GPUs

Hewlett Packard Enterprise · EVIDEN · NVIDIA · intel

125.9 PF

(FP64)

❑ With **new access modes** and **tools**

- AI software stack, containers, notebooks,
- Repository of models and datasets

❑ Some milestones

- Sept 2019 : Jean Zay in production
- Mid 2020 and Q1 2022 : 2 successive evolutions
- **Q2 2024 : New extension (H100) provided by EVIDEN**

❑ More than **1000 projects in IA** supported in 2023

- NLP, vision, multi modality, explainable AI, robotics…
- AI For Science : biology/health, energy, material science…

❑ Computing facilities

- Scalar partition
  - 720 nodes, 1440 CPU Intel CSL, 28 800 cores, 1x OPA
- Converged partitions
  - 396 nodes quad GPU → 1584 GPU V100 SXM2 16/32GB, 4xOPA
  - 83 nodes octo GPU → 664 GPU V100 or A100, up to 768 GB mem, 4xOPA
  - NEW **364 nodes quad GPU → 1456 GPU H100 SXM5 80GB, 512 GB mem, 4xNDR**

❑ Storage

- 4.3 PB @ 1.2 TB/s full Flash (N1)
- 39 PB @ 300 GB/s rotative (N2)
- Up to 100 PB tapes (N3)

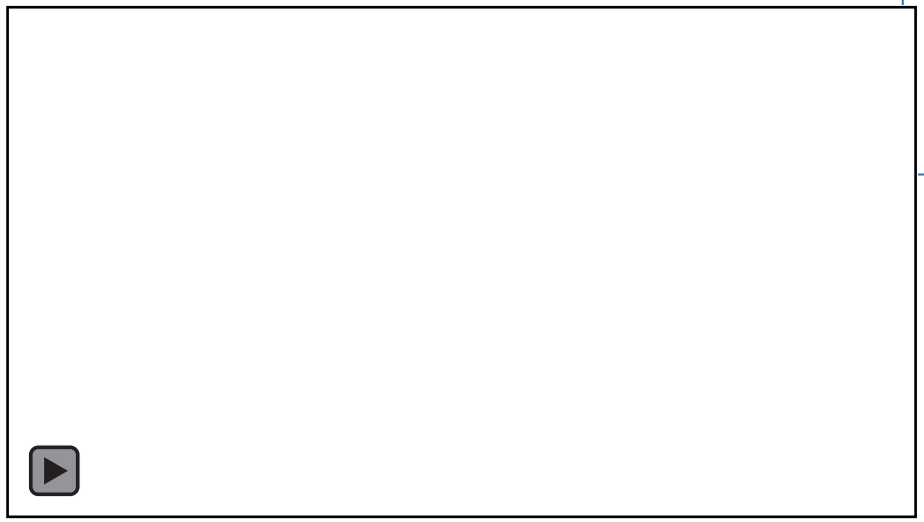
❑ And the support

- **25p with 13p dedicated for AI**

SOME RESULTS ON HPC AND AI CONVERGED WORKLOADS

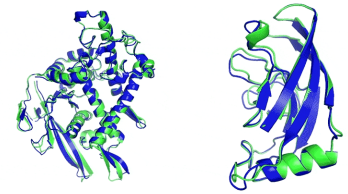
In academia and industry

World first-ever : Full engine combustion model using 13k cores on Joliot Curie



Institut Pasteur

Use of AlphaFold for Identifying new coronaviruses



AI model for playing bridge toward explainable AI

NukkAI

Update: Introducing The World's Largest Open Multilingual Language Model - BLOOM

- Training on Jean Zay of the biggest open NLP model
- Global collaboration (>1500 researchers), 47 natural and 13 programming languages
- 176B parameters, more than 400 GPUs used (3 months)

Hugging Face

NEW ACCESS MODES AS A KEY OF SUCCESS

Dynamic access (and soon strategic access)

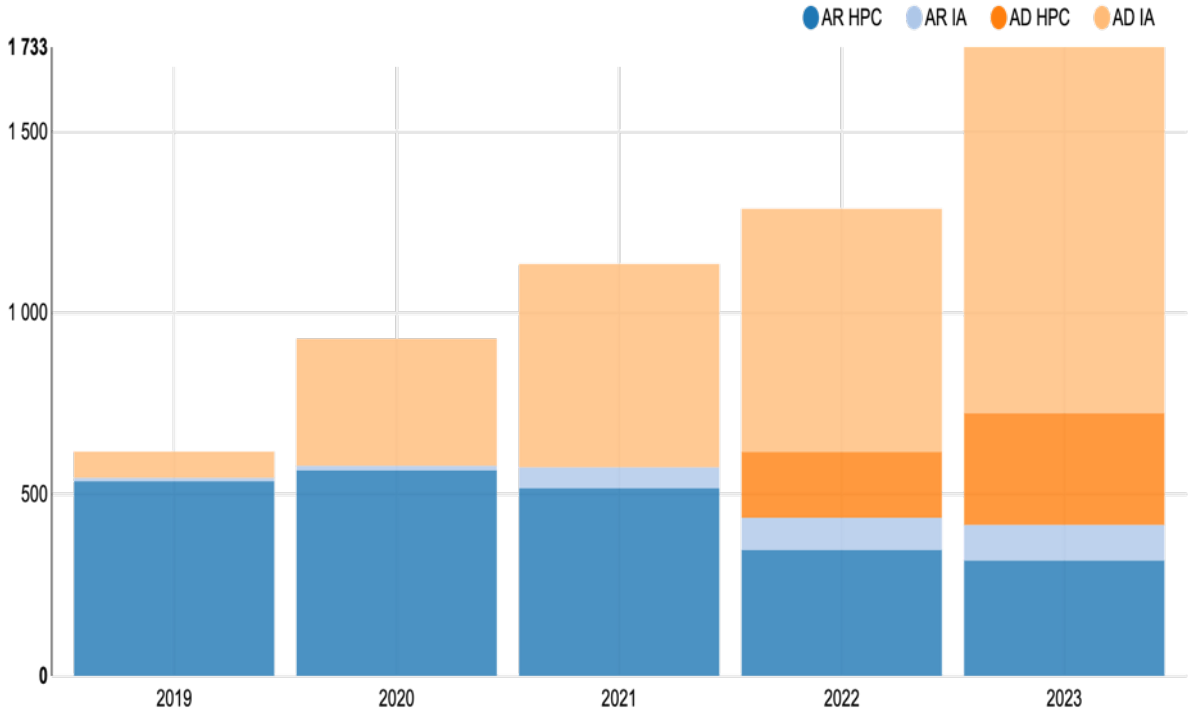
❑Before Jean Zay

- Regular access (twice a year) for large HPC allocations
- Preparatory access (cut-off every month) for code porting / dev.

Not fitted for AI needs

❑With Jean Zay

- Dynamic access : permanent access, in few clicks and few days access to up to 50k GPU hours or 500k CPU hours for 1 year
- IDRIS' user support (HPC/AI) for improved user engagement



Dynamic access extended to HPC in 2022

SOME ONGOING AND NEXT ACTIVITIES

Improving support and fostering synergies between HPC and AI

- ❑ Maintaining and increasing user support (HPC and AI)
- ❑ Developing strategic access
  - Mix of regular and dynamic access : large and fast dedicated allocations
  - Only for limited projects (strategic) per year

- ❑ Implementing a sovereign AI Cloud in France
  - CLUSSTER project federating public (GENCI, CEA, CNRS, Inria) and private actors (OVHCloud, **ATOS**, CS, Qarnot and more to come)
  - Offering a full continuum of facilities and services from learning, fine tuning to inference at scale
  - Use of the NIMBIX (ATOS/Eviden) Cloud layer

- ❑ Going converged HPC / AI (and hybrid Quantum) Exascale with EuroHPC
  - Jules Verne consortium (France and Netherlands)
  - Leveraging from national experience (GENCI, CEA, SURF)
  - 15k to 20k next gen GPUs coupled if possible to EU CPU technologies in 2025/6
  - Strong interested on EC AI Factories

EuroHPC Joint Undertaking