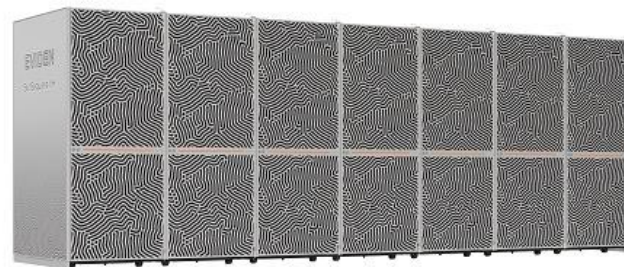


EPI Forum

Barcelona, 9-10.10.2024.



AI in France from an infrastructure point of view



S. Requena (GENCI)



GENCI, A FRENCH HPC RESEARCH INFRASTRUCTURE

Serving yearly **1700 projects** in HPC and AI (academia, industry) in 2023



EuroHPC
Joint Undertaking



France
Universités



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**

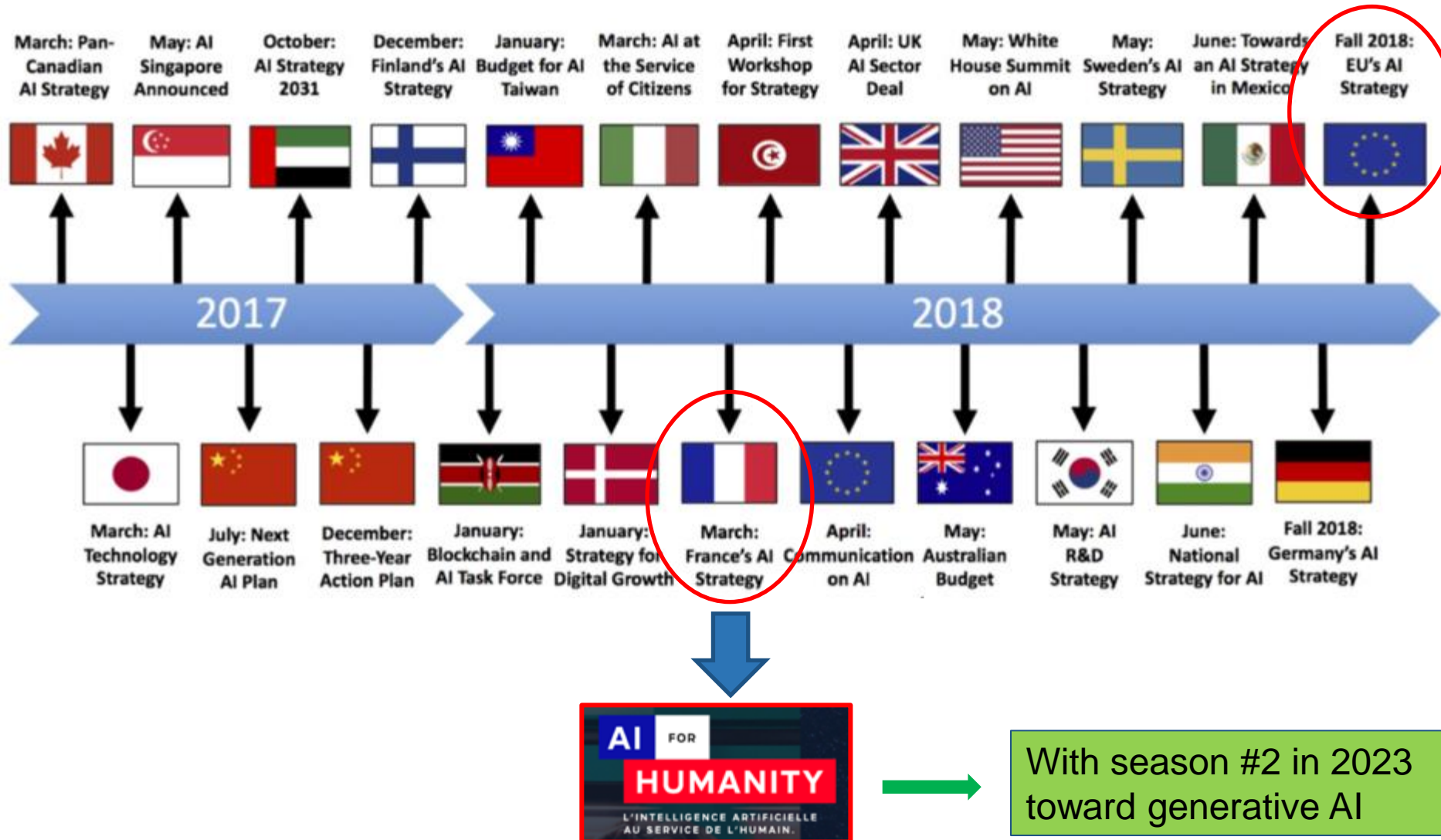
CINES/FU - Montpellier

- **> 70 PF** with AMD next gen GPUs (>1500) & CPUs (>100k)
- **Next step before French Exascale system**

#7 **THE GREEN 500**

EVOLUTION OF THE CONTEXT

In particular in AI



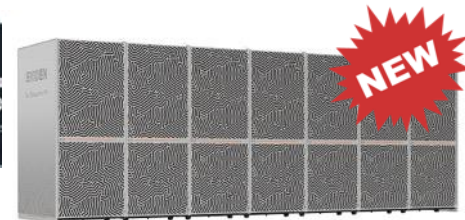
ET VOILA JEAN ZAY AT IDRIS (CNRS) !



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



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
 - 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

ET VOILA JEAN ZAY AT IDRIS (CNRS) !

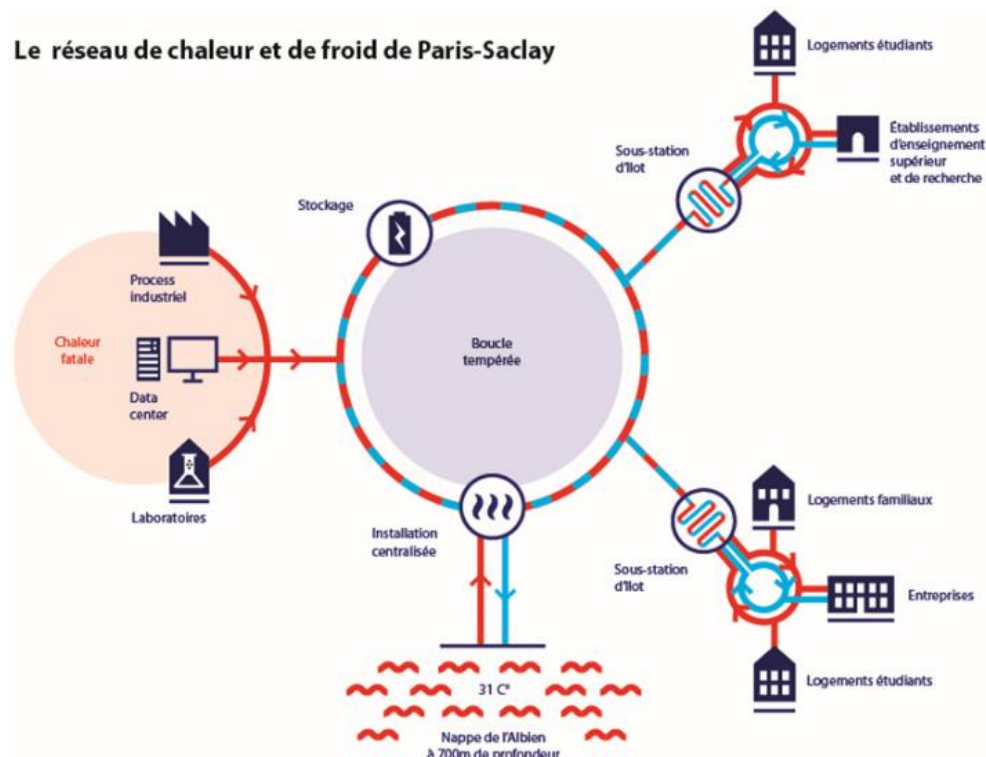
And one of the most eco-efficient in Europe

- Agreement between EPAPS and CNRS in February 2021



Philippe van de Maele, directeur général de l'EPA Paris-Saclay, et Alain Schuhl, directeur général délégué à la science du CNRS signent la convention CNRS/EPA Paris-Saclay visant à récupérer la chaleur fatale du supercalculateur Jean Zay. Photo CNRS IdF Gif-sur-Yvette

Le réseau de chaleur et de froid de Paris-Saclay



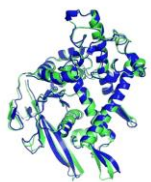
- Goal : Reinject the fatal energy of Jean Zay to heat a whole urban campus (Paris Saclay)
- Operational since 2023 → 4000 MWh/year = 1 000 houses

SOME RESULTS ON HPC AND AI CONVERGED WORKLOADS

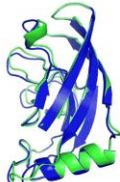
In academia and industry



Use of AlphaFold for
Identifying new coronaviruses



TI037 / 6vr4
90.7 GDT
(RNA polymerase domain)



TI049 / 6vr4f
93.3 GDT
(adhesin tip)

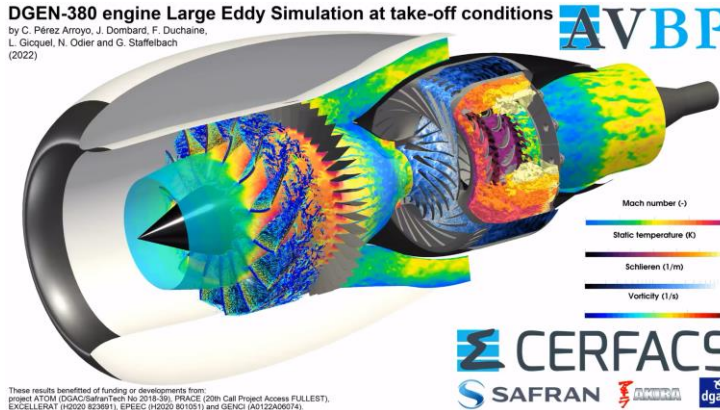
● Experimental result
● Computational prediction

AI model for playing bridge
toward explainable AI



DGEN-380 engine Large Eddy Simulation at take-off conditions

by C. Pérez Arroyo, J. Dombard, F. Duchaine,
L. Gicquel, N. Odier and G. Staffelbach
(2022)



These results benefited of funding or developments from:
project ATOM (DGAC/SafranTech No 2018-36), PRACE (20th Call Project Access FULLEST),
EXCELLERAT (H2020 823691), EPEEC (H2020 801051) and GENCI (A012246674).

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



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)



- Multimodal Cognitive AI Agents to
create Human-Machine Interfaces
(a la Neuralink)



SERVING AI ACROSS ACADEMIA AND PUBLIC SERVICES

From a wide range of organisations



+ many engineering schools



**PREMIER
MINISTRE**

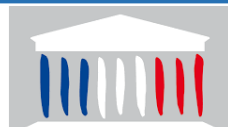
Liberté
Égalité
Fraternité

Direction interministérielle
du numérique



**MINISTÈRE
DE LA CULTURE**

Liberté
Égalité
Fraternité



**ASSEMBLÉE
NATIONALE**



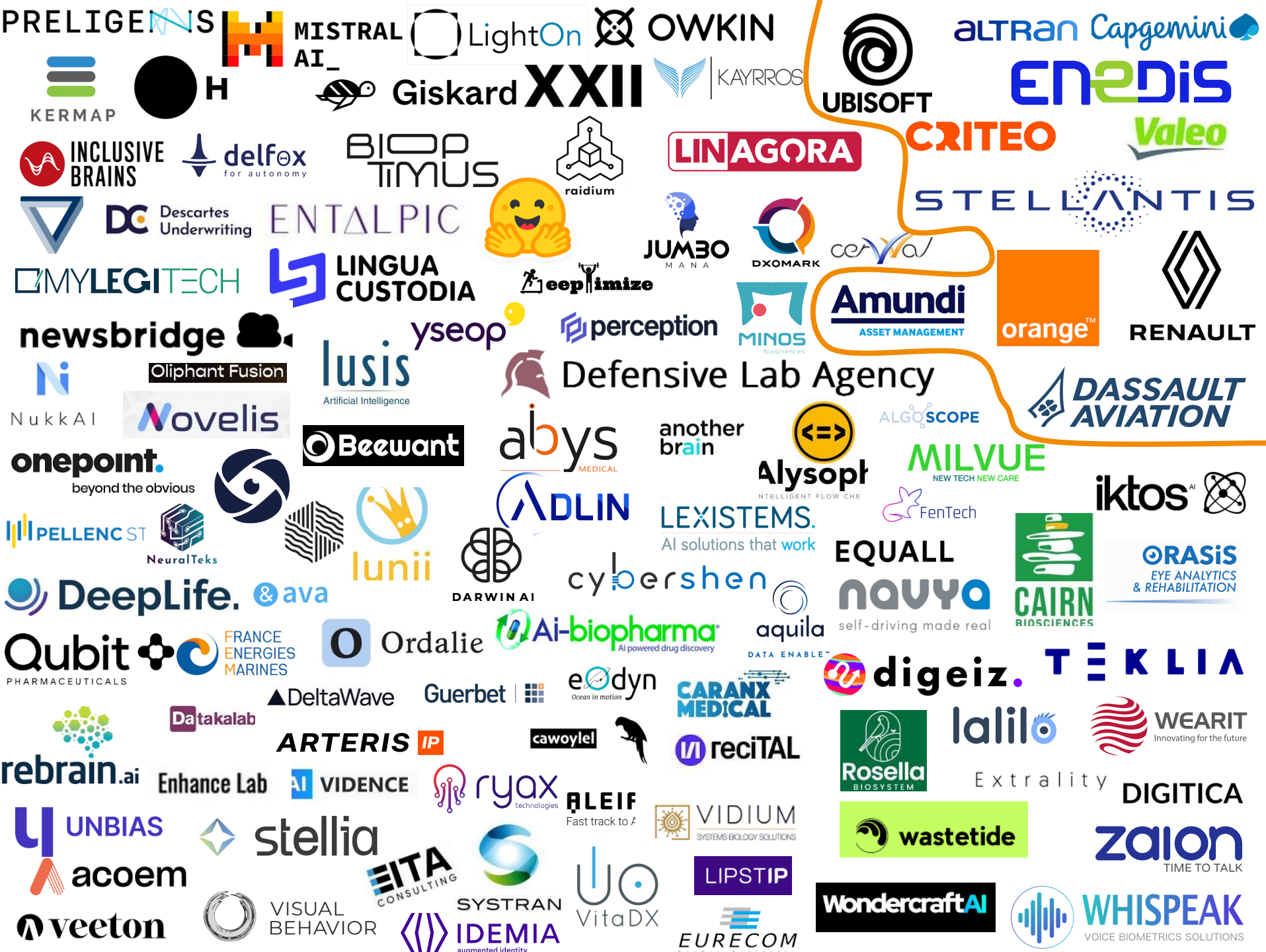
GOVERNEMENT

Liberté
Égalité
Fraternité

Pôle d'Expertise
de la Régulation Numérique



Serving AI across industry
startups (many), SMEs & large groups



NEW ACCESS MODES AS A KEY OF SUCCESS

Dynamic access (and soon strategic access)

❑ Before Jean Zay

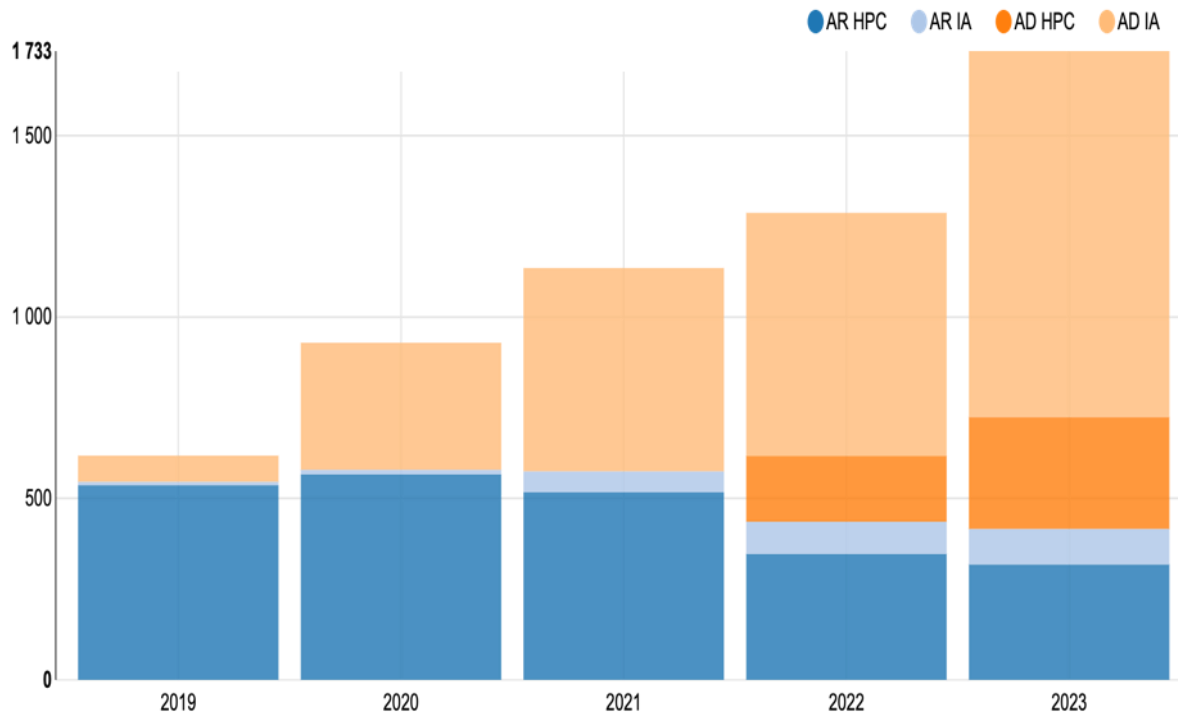
- Regular access (twice a year) for large HPC allocations
- Preparatory access (cut-off every month) for code porting / dev.



Not fitted for AI needs

❑ With Jean Zay

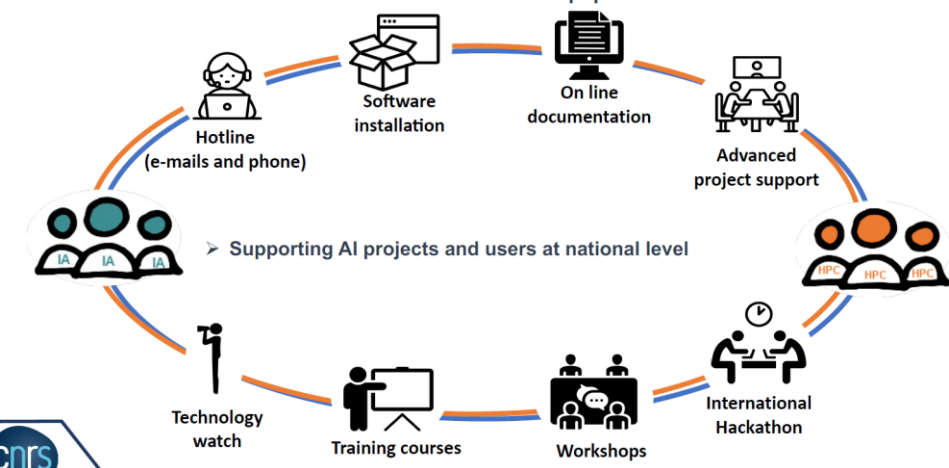
- Dynamic access : permanent access, in few clicks and few days access to up to 50k GPU hours or 500k CPU hours for 1 year
- Increase of IDRIS' user support (HPC/AI) for user engagement



Dynamic access
extended to HPC and
to all 3 computing centers
in 2022

The other key of success

Missions of user support teams



Jean Zay computing environment



- Shared disk space accessible to all users

- Storage of voluminous data bases (in size or number of files)
- Storage of 🧠 **huggingface_hub** models
- More than 220 models and data bases
- More than 1,4 Po data
- More than 600 millions files and directories

Training courses since 2021

- Practical Introduction to Deep Learning (IPDL)
 - 5 sessions, 100 people
- Optimised Deep Learning on Jean Zay
 - 5 sessions, 100 people
- FIDLE (Introduction to Deep Learning)
 - 3rd season, 20 sequences of 2 hours, 40 000 hours watched for the 2nd season
- IDRIS-NVIDIA International Hackathon for IA and HPC
 - 3 sessions, 27 projects, next session :
 - <https://www.openhackathons.org>



IDRIS OPEN HACKATHON
March 18-28, 2024
Application Deadline: January 15, 2024
Hybrid Event



LOOKING BACK IN THE MIRROR

Some feedback after 4 years of production

□ Things we did

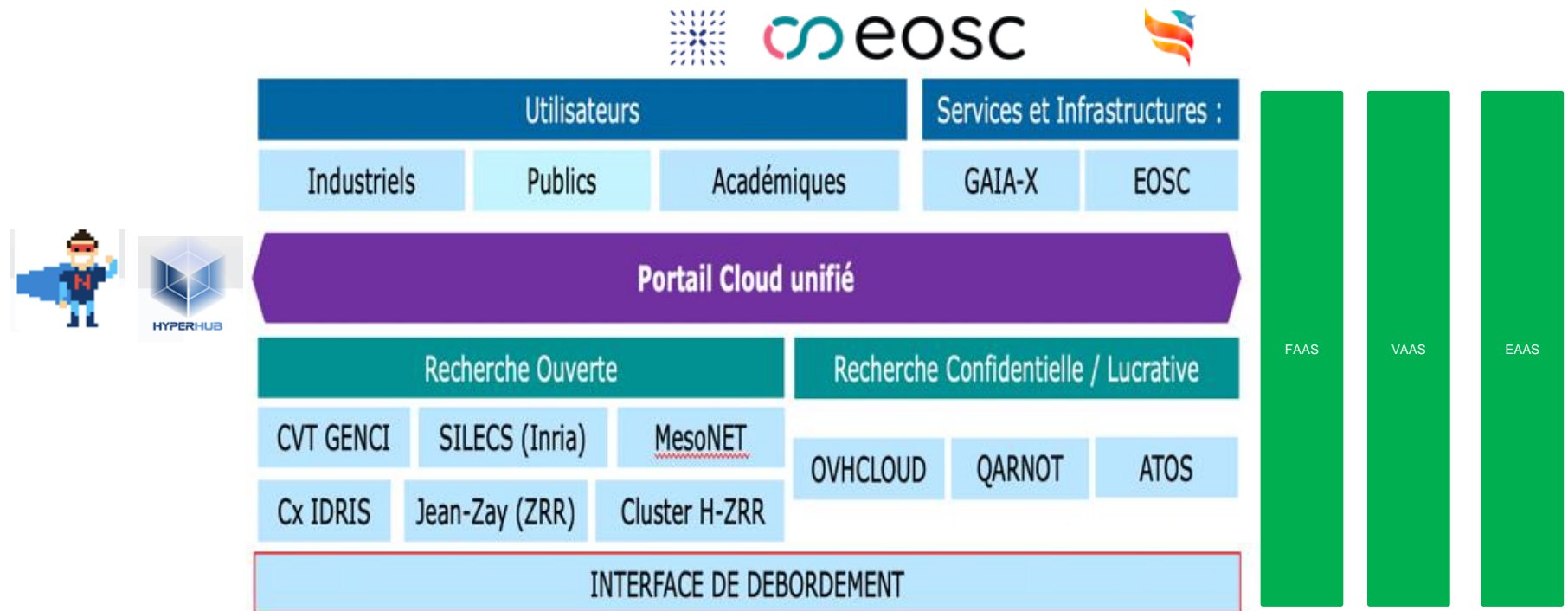
- 4 updates in 4 years ! → High demand and new GPU features to support
- New access modes and HPC/AI user support (25p)
 - Strong efforts from IDRIS and CNRS 🙏
 - But major stakes on attracting / retaining skills (EuMasterHPC need a 10x increase !)
- Support of containers and Jupyter notebooks
- Served AI research from academia and industry (many startups and groups)
- Paved the path to Cloud access with CLUSTER following the FR Cloud Strategy
- Setup a repository of models and datasets (>500 models)
 - Allow to expose models and reduce download needs (bw, security rules...)

□ Things to improve or develop

- Having Kubernetes and SLURM together
- Integrate object based storage alongside with Lustre / ephemeral storage serv.
- Better support of workflows & MLOps services in a Cloud based env.
- Support of longer batch queues ? And dedicated allocations ?
- Inference services : AI for science only or for any inference ?

□ Unified sovereign and secure portal for AI (+ HPC and Quantum)

- Federating existing infrastructures from private and public operators
- Continuum Open Research, Confidential Research & Commercial Activities
- Academic Sector, Industry and public authorities
- Integration European ecosystem: GAIA-X, European Open Science Cloud and Fenix



WHERE NOW WE GO ?



- Computing power for AI is doubling every 100 days and the energy required to run AI is accelerating with an annual growth of 26%-36%.

Sources: International Energy Agency, 2023 / Intelligent Computing: The Latest Advances, Challenges, and Future, 2023. / WEF, 2024.

Hyperion Research Announces a 36.7% Increase in the HPC/AI Market Size

By Doug Eadline

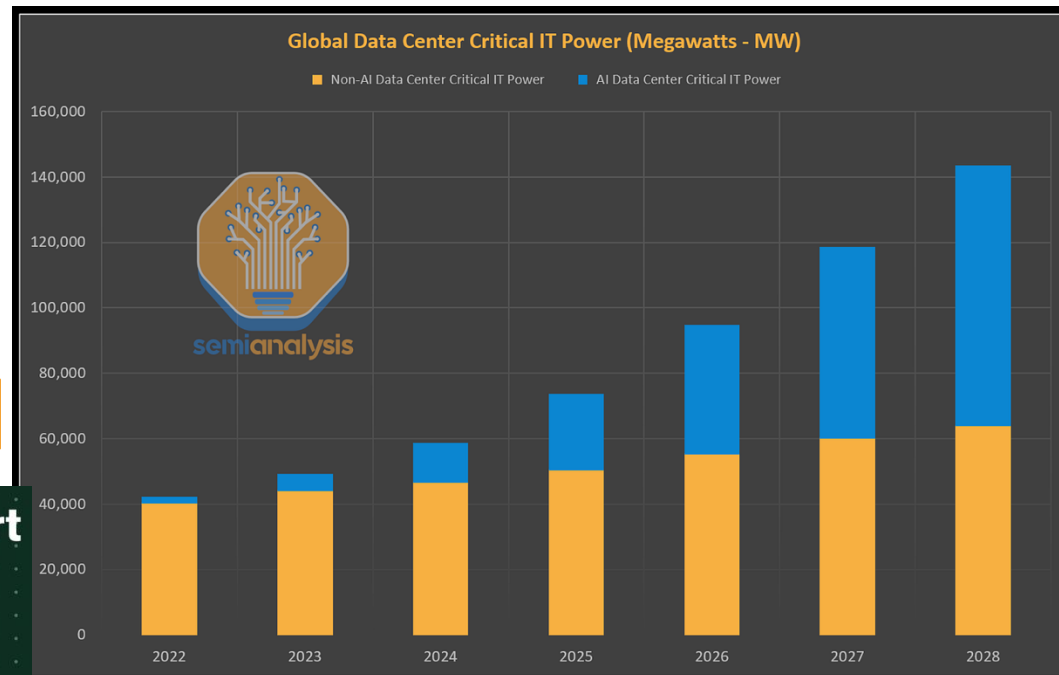
Generative AI to Become a \$1.3 Trillion Market by 2032, Research Finds

June 01, 2023

Bloomberg Intelligence: New Report Finds That the Emerging Industry Could Grow at a CAGR of 42% Over the Next 10 Years

Rising demand for generative AI products could add about \$280 billion of new software revenue

UBS predicts the AI industry's revenue will grow at a ~72% CAGR from \$83B in 2024 to \$420B in 2027.



Why Microsoft made a deal to help restart Three Mile Island

A once-shuttered nuclear plant could soon return to the grid.

By Casey Crownhart

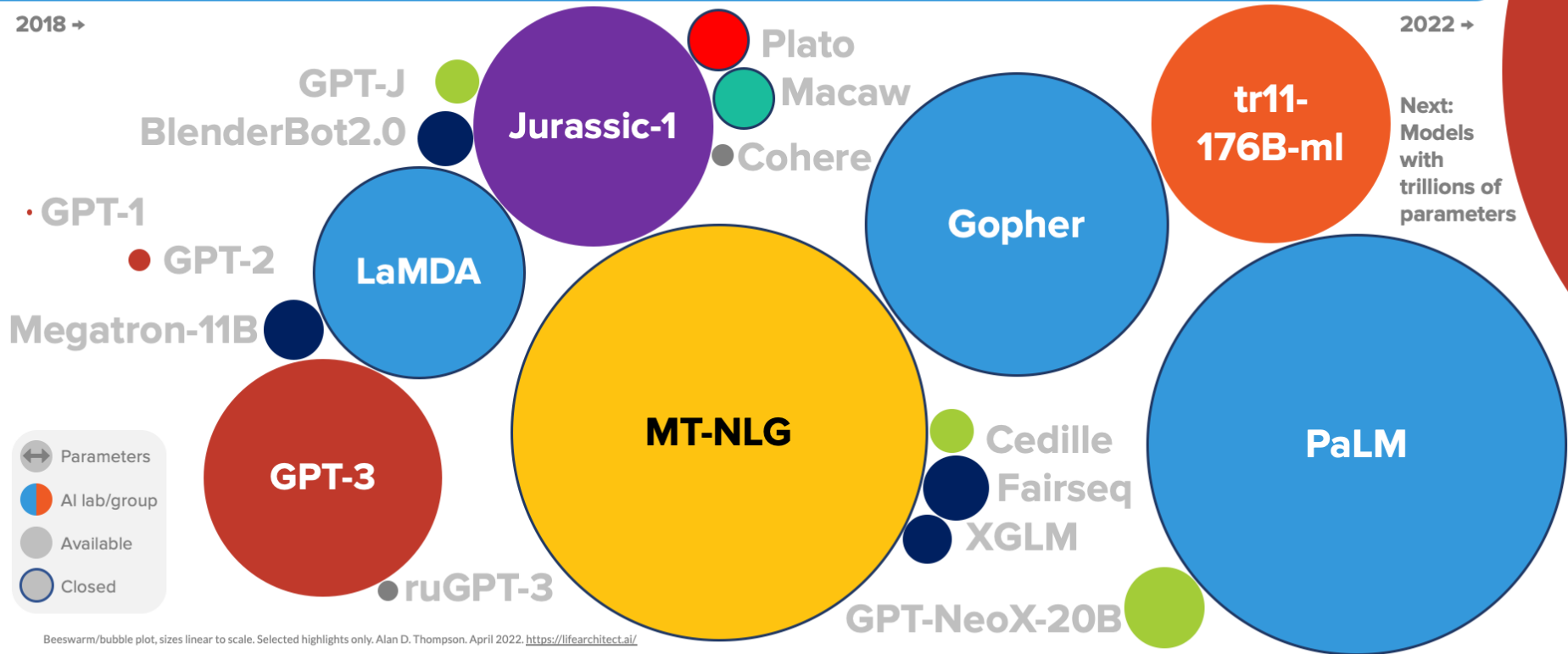
September 26, 2024

SOME ONGOING BATTLES INTO THE AI ECOSYSTEM

LLM : explosion of the number of models since 2022

LANGUAGE MODEL SIZES TO APR/2022

OR: WHILE YOU WERE SLEEPING,
AI SIZES WERE EXPLODING



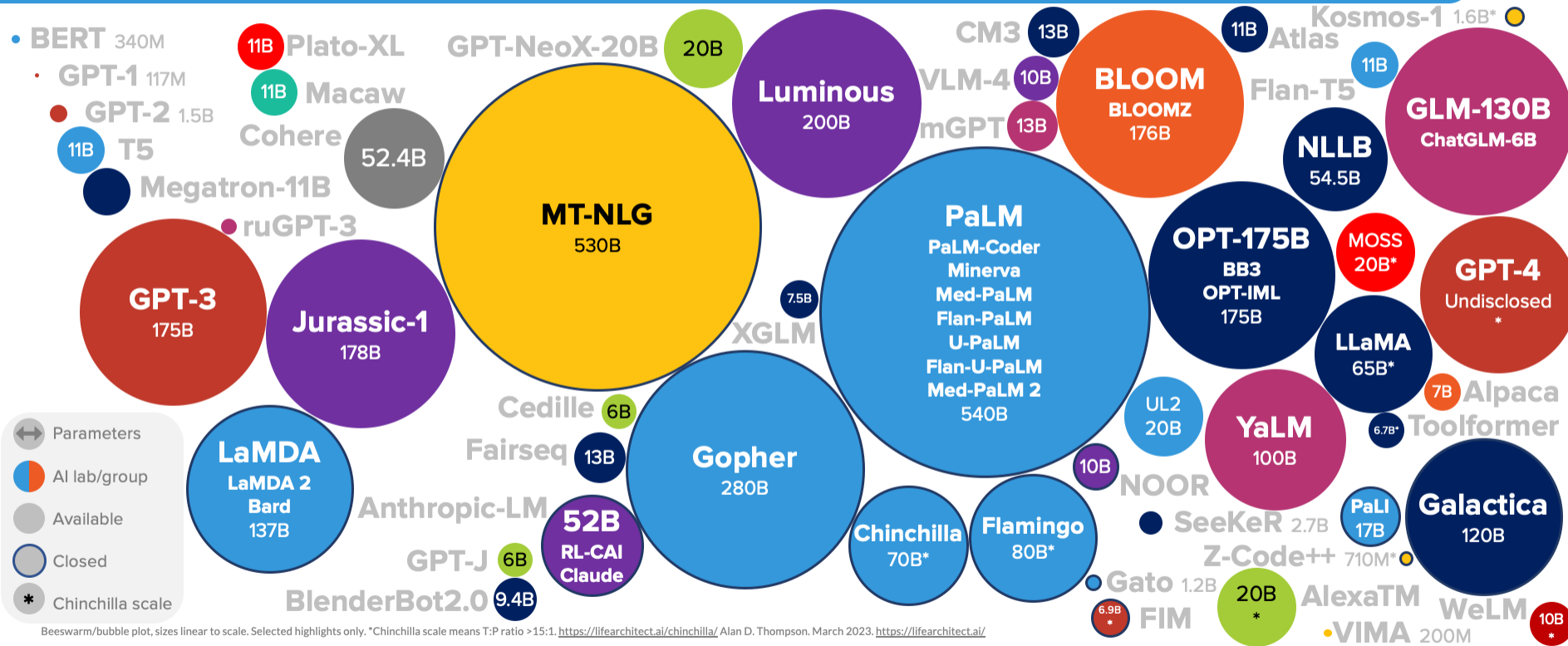
LifeArchitect.ai/models

- Almost only closed models available (MT-NLG, GPT3, LaMDA, PaLM...)

SOME ONGOING BATTLES INTO THE AI ECOSYSTEM

LLM : explosion of the number of models (in 2023)

LANGUAGE MODEL SIZES TO MAR/2023



LifeArchitect.ai/models

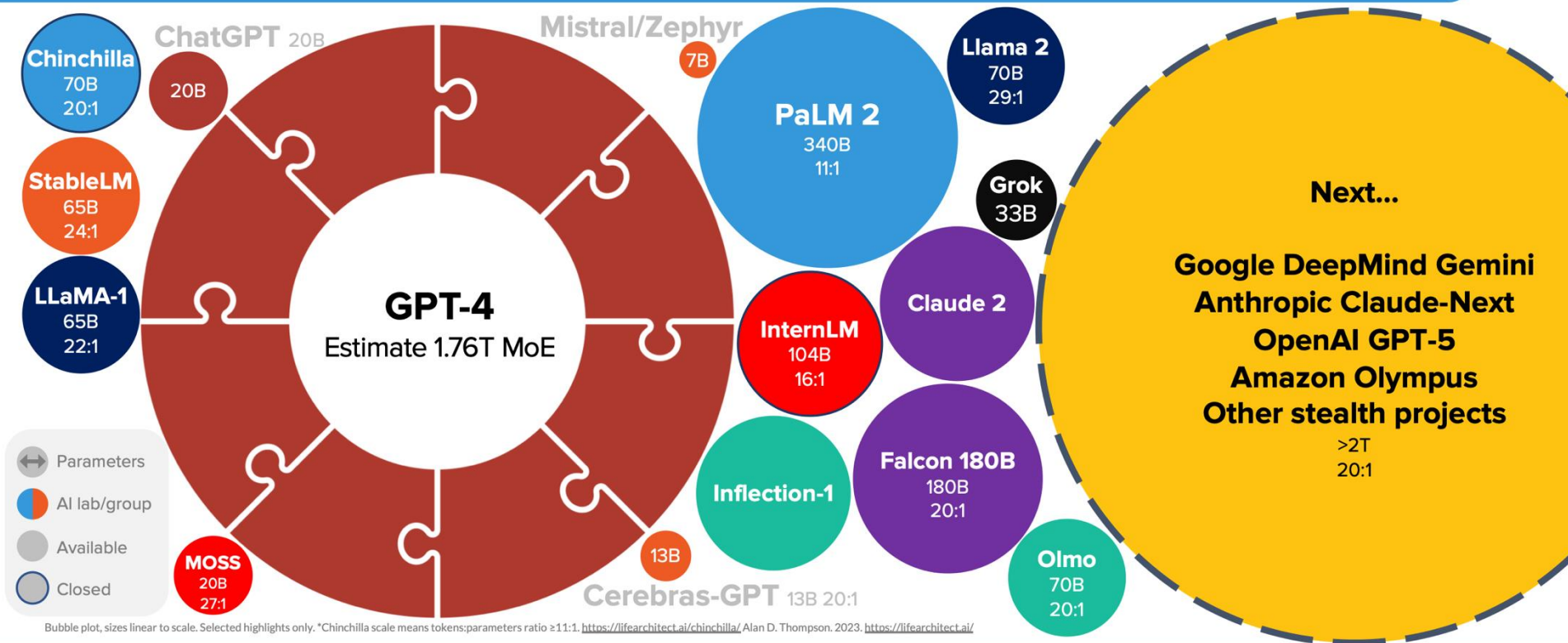
- Open source (Bloom, OPT, Luminous...) vs proprietary models (GPT3, PALM, LaMBDA...)

SOME ONGOING BATTLES INTO THE AI ECOSYSTEM

LLM : explosion of the number of models (in 2023/4)

2023-2024 OPTIMAL LANGUAGE MODELS

NOV/
2023



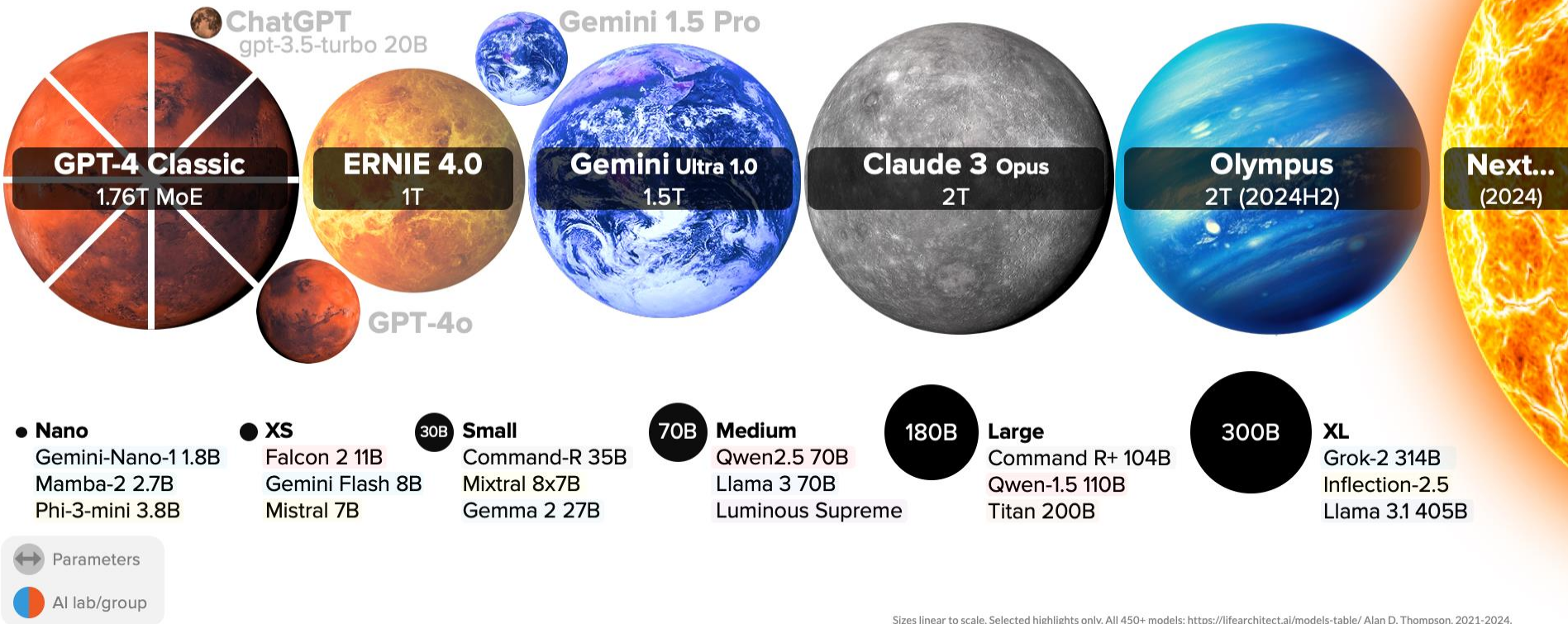
LifeArchitect.ai/models

- Open source (Llama2, Olmo, Falcon...) vs proprietary models (GPT4, PaLM2, Claude2...)
- **Large models vs over-trained smaller models (Chinchilla laws)**

SOME ONGOING BATTLES INTO THE AI ECOSYSTEM

LLM : explosion of the number of models (in 2024)

LARGE LANGUAGE MODEL HIGHLIGHTS (OCT/2024)



LifeArchitect.ai/models

& 450+ more models at LifeArchitect.ai/models-table

- Open source (Llama3, Mistral, Falcon...) vs proprietary models (GPT4, Gemini, Claude, Bard...)
- Large models vs over-trained smaller models and **SLMs (Llama3.2 1B better than Llama2 13B)**
- **Multimodal models, AI Agents, MoE**
- **RAG and long context windows**

ANOTHER CHALLENGES TO CONSIDER

AI is THE dominant market

No GPUs available, what now?

How to tackle the shortage of GPUs?

March 2023 UPDATE

Anouk Dutrieu



Announcing General Availability of OCI Compute with AMD MI300X GPUs

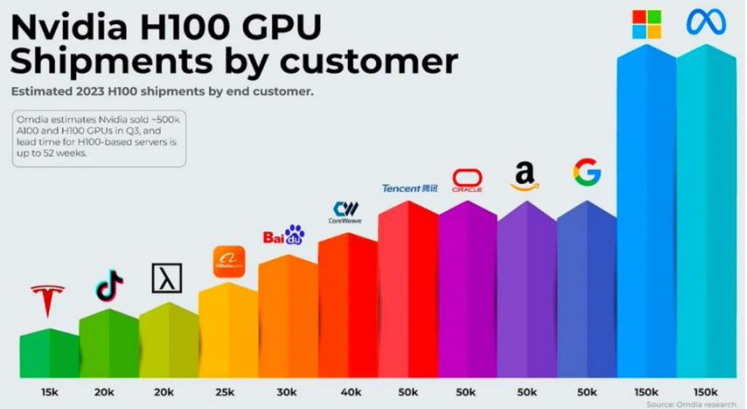
Nvidia Shipped 3.76 Million Data-center GPUs in 2023, According to Study

By Agam Shah

Nvidia H100 GPU Shipments by customer

Estimated 2023 H100 shipments by end customer.

Omdia estimates Nvidia sold ~500k A100 and H100 GPUs in Q3, and lead time for H100-based servers is up to 52 weeks.



Are you GPU poor? 🙌🤔

Calculate GPU memory requirement and token/s for any LLM

❑ May recent GPU shortages with (big) customers ready to pay 150%

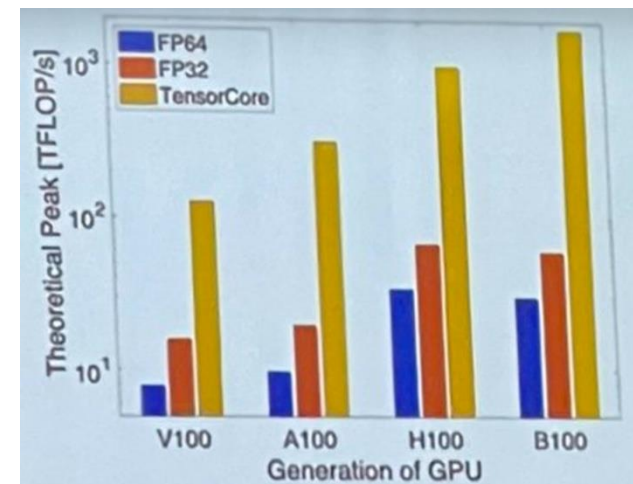
❑ New GPU features dedicated to AI → « *HPC is converging with AI* »

❑ HPC market need to adapt

- Exploiting low precision operators / FP64 emulation for HPC apps
- Integrating AI and HPC

❑ And Europe too

- AI Factories for reaching the critical mass
- EU Chips Act with our own xPU technologies



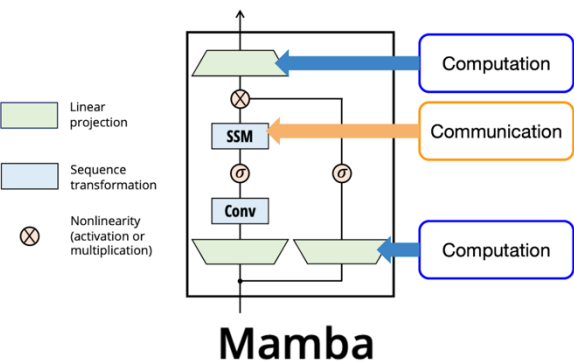
ANOTHER CHALLENGES TO CONSIDER

Impacting infrastructures and services

❑ Today HPC/Cloud infrastructures are hyper provisioned for Transformers

- Scale up level with 8 to tens GPUs/node, HBM and (LP,MR)DDR memory
- Injection bandwidth per node (400/800 Gbps/GPU), ...
- Quadratic complexity : when input data doubles -> compute/mem is 4x

❑ But new alternatives are rising



Megalodon

7B

Transformer Alternative

Jamba

AI21's Hybrid SSM -Transformer Model

xLSTM

Extended Long Short-Term Memory

Liquid Foundation Models: Our First Series of Generative AI Models

Published September 30th, 2024

Kolmogorov-Arnold Networks (KAN)

2024.5.1



RWKV LANGUAGE MODEL

❑ Importance to understand the impacts in a mid-term view?

→ A Scientific / Industrial Case on AI is needed !



SOME ONGOING AND NEXT ACTIVITIES FOR GENCI

Improving support and fostering synergies between HPC and AI

□ Developing strategic access

- Mix of regular and dynamic access : large and fast dedicated allocations
- Request from the French state, only for few projects (strategic) per year

□ Implementing a sovereign AI Cloud in France

- Offering a full continuum of facilities & services from learning, fine tuning to inference at scale
- From open research, confidential research to commercial activities

□ Going converged HPC / AI (and hybrid Quantum) Exascale with EuroHPC

- Jules Verne consortium (France and Netherlands)
- Leveraging from national experience (GENCI, CEA, SURF)
- Alice Recoque Exascale system will aim 15k next gen GPUs strongly coupled if possible to EU CPU tech. in 2025/6
- Strong interested on EuroHPC AI Factories call



EPI FORUM



EuroHPC
Joint Undertaking

PLATINUM SPONSORS



GOLD SPONSORS

