# Centre Informatique National de l'Enseignement Supérieur

**Calcul Intensif** | **Hébergement** | **Archivage**


MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE

France Universités

GENCI
Le calcul intensif au service de la connaissance

JCAD 2024

# Adastra: une architecture de calcul exascale au service de la recherche nationale en HPC et IA

4 nov 2024

Gabriel Hautreux

Responsable du département calcul intensif

# L’un des 3 centres de calcul nationaux: 45 ans d’histoire et d’expertise HPC


… 1979 IBM 2004

GENCI
Le calcul intensif au service de la connaissance

2022

1999 sgi 2015

JADE

Atos

OCCIGEN

Hewlett Packard Enterprise CRAY AMD AMD

2023 2024 …


ADASTRA
GPU MI250
CPU GENOA EPYC 9654

ADASTRA 2
GPU MI300A

CINES Centre Informatique National de l'Enseignement Supérieur → CiNES


AMD
HPE Cray EX system
AMD GPU + CPU
#10 Top 500 (June 22)
#3 Green 500 (Nov 22)
POWERING THE VERY TOP OF THE TOP500*
FRONTIER
1
LUMI
3
GENCI
INES
10
AMD
EPYC | INSTINCT
Hewlett Packard Enterprise

# Supercalculateur Adastra (installé en 2022)

75 Pflop/s


## Partition GPU

356 noeuds GPU nodes :

- 8 AMD MI250X GCD avec **64G** HBM2/GCD

*Même technologie que:*

- **Frontier (#1 monde)** et **LUMI (#1 Europe)**

## Partition CPU

544 noeuds scalaires:

- 2 AMD Genoa EPYC 9654 96 coeurs @ 2.4 GHz, 768G DDR5-5200 par noeud

## Réseau et stockage haut débit

- Réseau Slingshot 200Gb/s
- Stockage ClusterStor 2Po SSD +12Po HDD

## Classements

| | June 2022 | Nov 2022 | June 2023 | Nov 2023 | June 2024 |
|---|---|---|---|---|---|
| The GREEN 500 | **4** | **3** | **3** | **3** | **9** |
| TOP 500 The List. | **10** | **11** | **12** | **17** | **20** |

Hewlett Packard Enterprise CRAY AMD AMD GENCI Le calcul intensif au service de la connaissance CINES

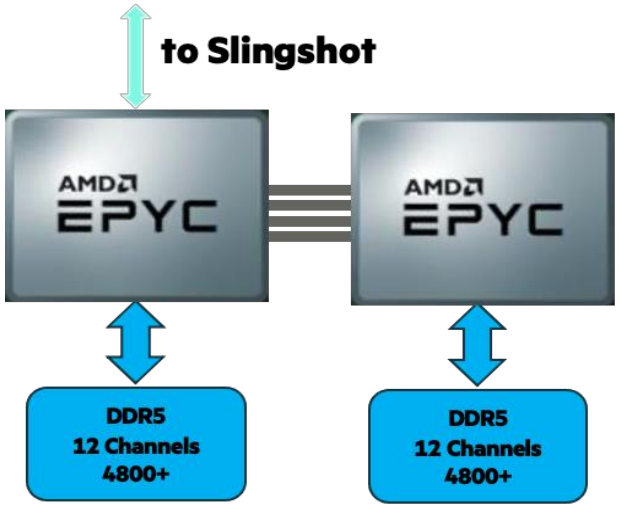
# Zoom sur la partition CPU

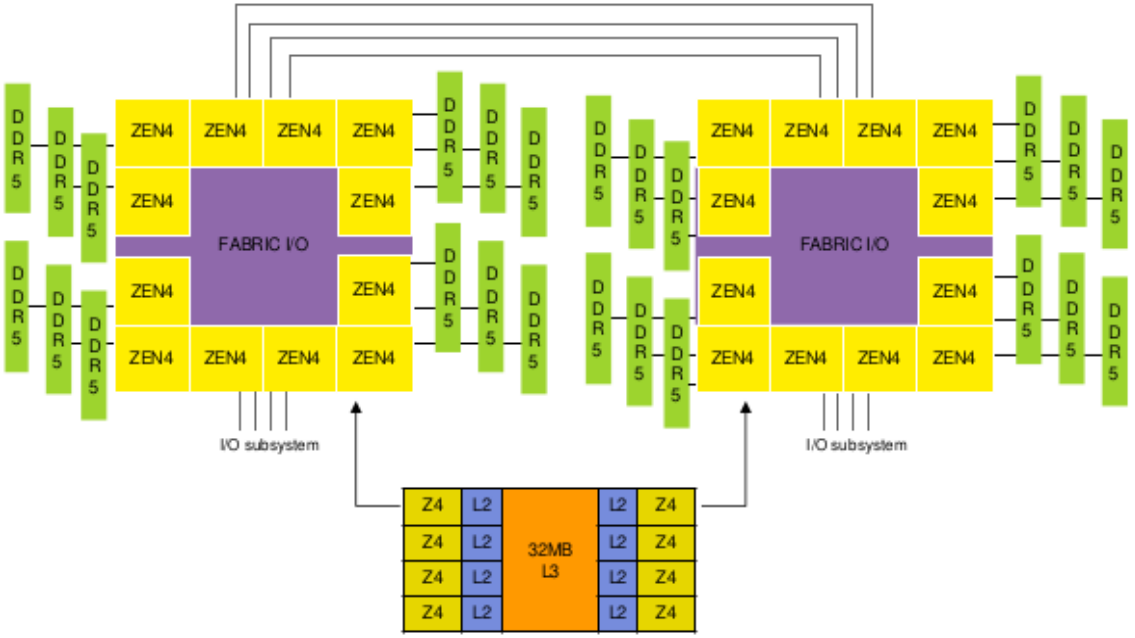

## Peak performance de 3,9 Pflops, équivalent à la machine précédente (Occigen, 3,5PFlops)

A atteint 3,5 PFlops (HPL), **en consommant 3 fois moins d'énergie qu'Occigen**.

### Bi-socket AMD Genoa 2x96 coeurs, 2.4 GHz (TDP 360W) + 768 Go DDR5-4800

- 536 nodes → ~100k coeurs Zen4 et 402To de mémoire aggrégée !
- 4Go de mémoire par coeur
- Puissance ~1kW par noeud (~6,5 GFlops/W)



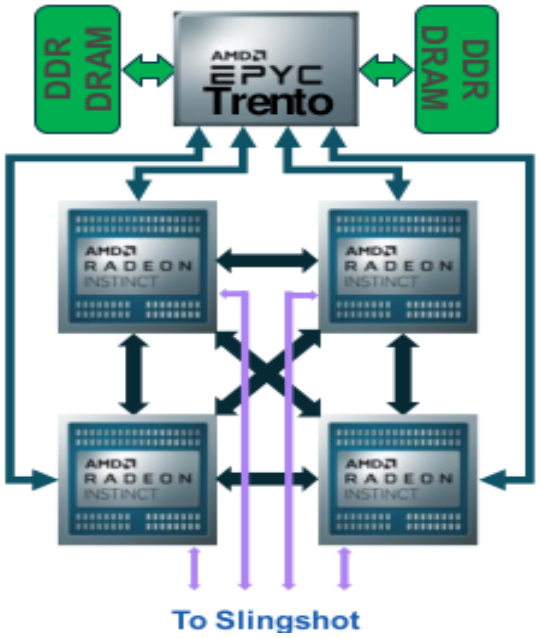



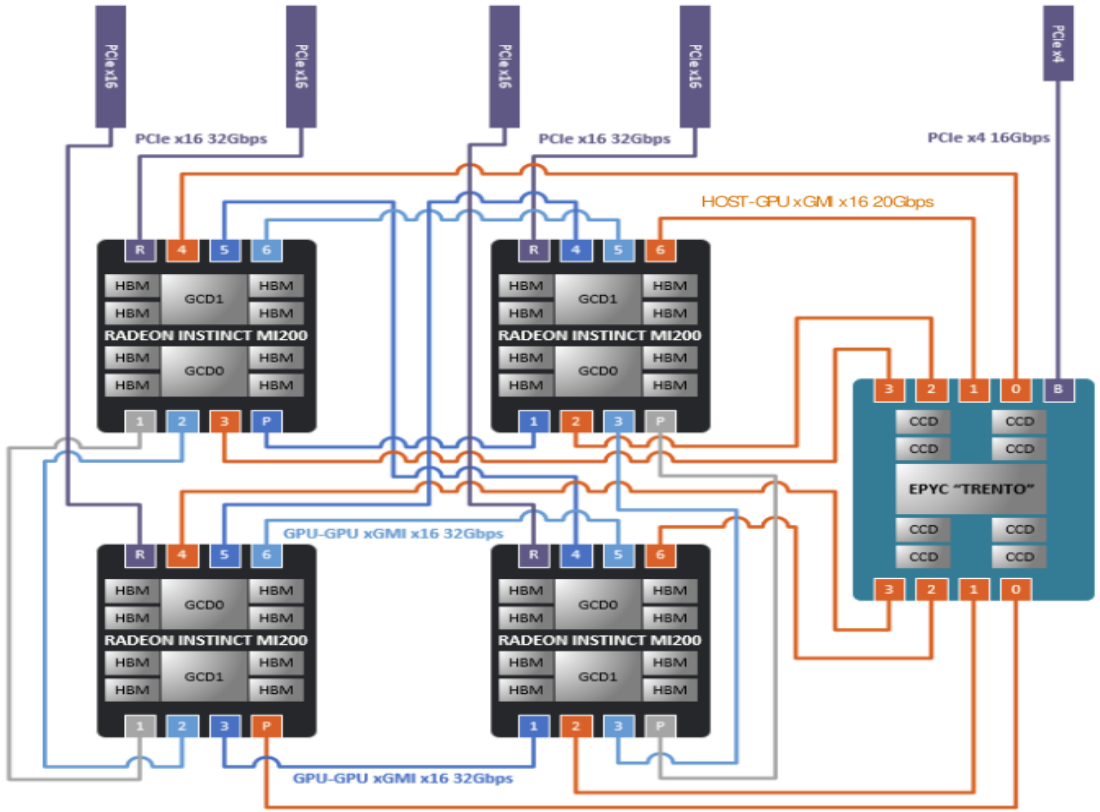
# Zoom sur la partition GPU


## Un équivalent LUMI/Frontier avec une performance peak de ~71 PFlops

## AMD Trento 64 coeurs, 2.4 GHz, 256GB DDR4-3200
## + 4 GPU AMD MI250X, 4x128GB HBM2, 4 Slingshot 200 Gbps per node

- Infinity fabric intra-noeud + 4 liens slingshot GPU direct
- ~200Tflops par noeud
- ~3kW par noeud
- ~60GFlops/W



# Supercalculateur Adastra-2 (2024)


## Nouveau : partition APU

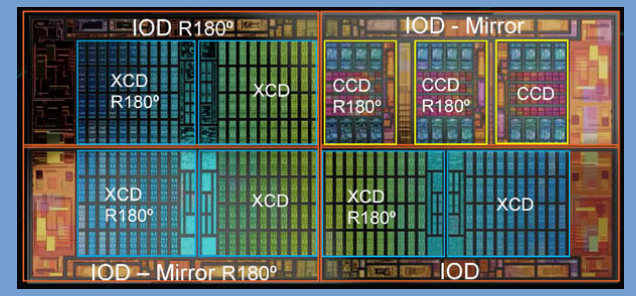
28 noeuds convergés :
- 4 APU AMD MI300A avec **128G** HMB3 / GPU
- Réseau 800Gb/s

- Même technologie que
- **El Capitan (sera #1 au monde en Nov 2024)**

| Adastra 1 | Adastra 2 | Adastra Next ? |
|---|---|---|
| 75 Pflops<br>1,3MW | 13 Pflops<br>0,084MW | 300 Pflops?<br>2-3MW? |

## Sobriété

*Refroidissement eau tiède : 30°C en entrée, 45°C en sortie*




**>13 Pflop/s**

Soit ~90 Pflops/s pour Adastra 1+2


**Intégration de la chaleur générée au réseau de chaleur Montpellier Nord**

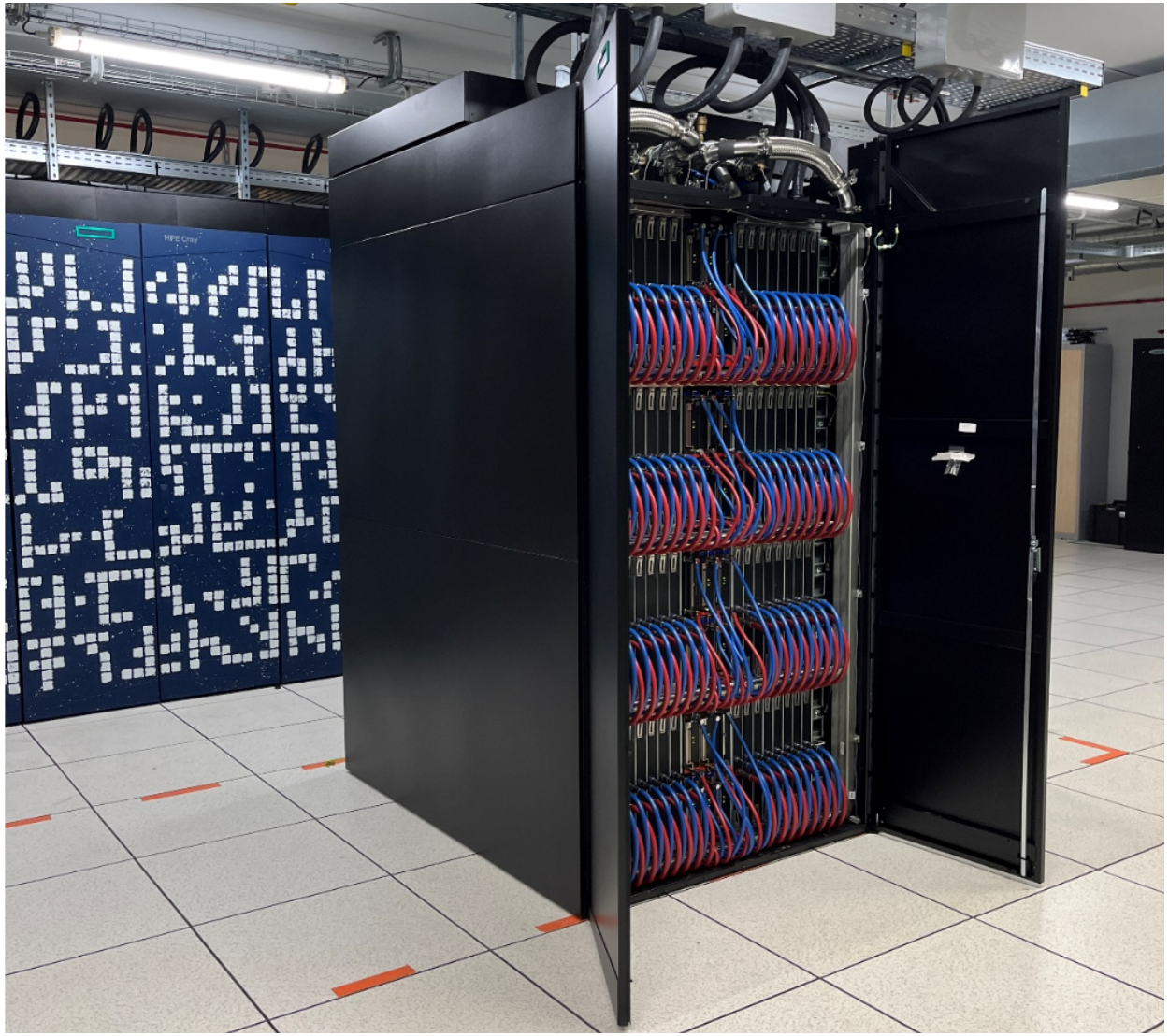
# Supercalculateur Adastra-2 (2024)

## Une technologie novatrice
- Convergence entre CPU et GPU
- Mémoire totalement unifiée

## Une technologie frugale
- Très peu de composants
- Plus aucune barrette de RAM
- Diminution du taux de panne

## Une technologie “scalable”
- Possibilité de continuer à faire évoluer la configuration

## Une technologie efficace
- **Très bon rendement énergétique** (classement green 500 à venir !)
- Des **accélérations en IA <u>observées</u> jusqu’à x3** par rapport à la technologie précédente!
- Des **accélérations en HPC <u>observées</u> jusqu’à x2.5** par rapport à la génération précédente



Hewlett Packard Enterprise · CRAY · AMD · AMD · GENCI · CINES

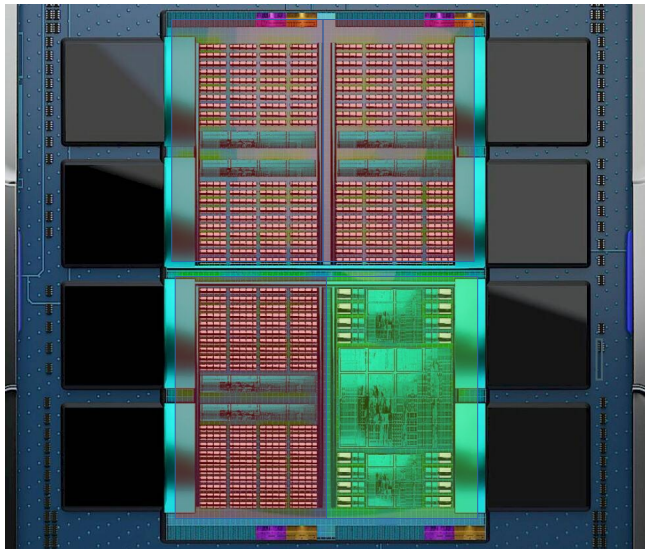
CINES

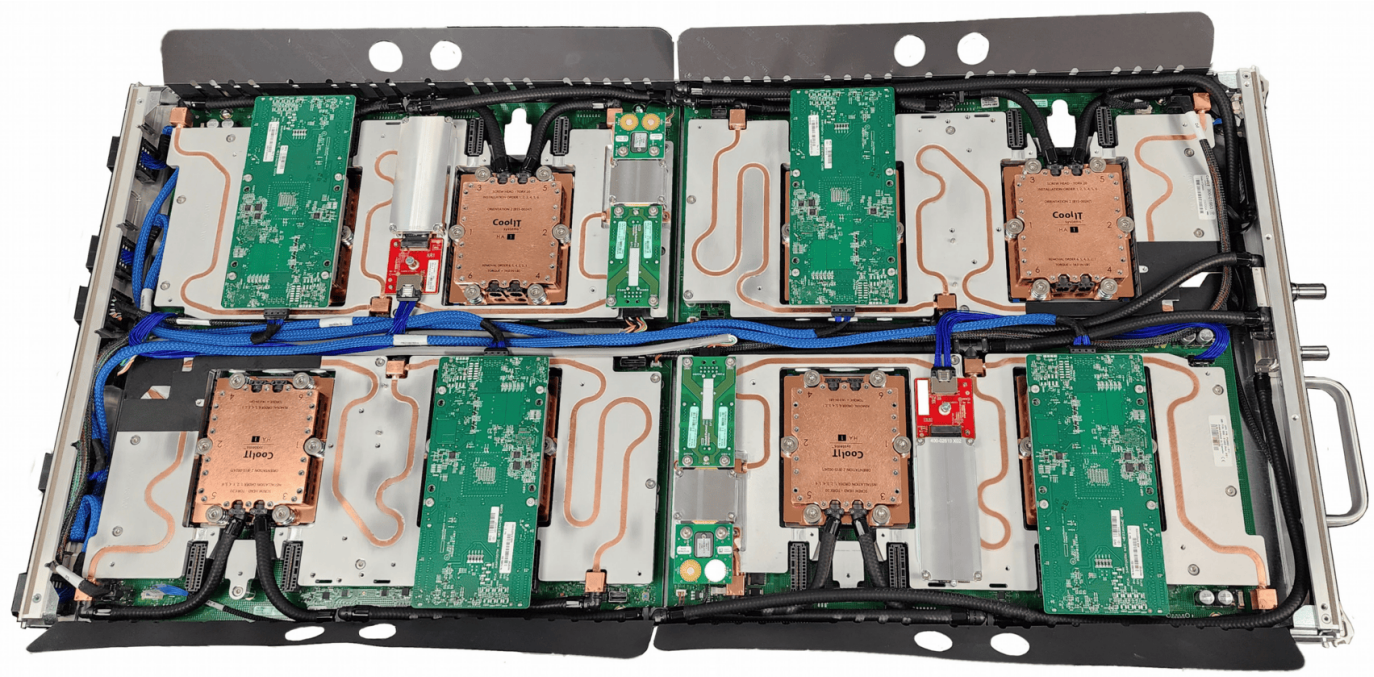
# Zoom sur la partition APU


## Un équivalent El Capitan avec une performance peak de ~13 PFlops

## AMD MI300A 24 coeurs 3,7 GHz + 228 unité de calcul GPU + 4x128GB HBM3, 4 Slingshot 200 Gbps par noeud

- **<u>Une mémoire unique pour les unités de calcul CPU et GPU</u>**
- FP64 matrix: ~480Tflops par noeud
- ~2,8kW par noeud
- Xxx Gflops/W → Voir liste Green500 Nov. 24






CINES

# Technologie MI300, enfin un challenger pour l'IA?

## AMD fait partie de la Pytorch foundation

- https://www.amd.com/en/press-releases/2022-09-12-amd-joins-new-pytorch-foundation-founding-member-to-promote-broader

## AMD intègre ML et IA directement dans Rocm

- https://www.amd.com/fr/graphics/servers-solutions-rocm-ml

## AMD offre des très bonnes performances mémoire par rapport à la concurrence

- Le MI300A (resp. MI300X) offre 128Go de HBM (resp. 192Go!).
- Le MI300A/X offre une bien meilleure bande passante que la concurrence

| | Single GPU | | | | | | |
|---|---|---|---|---|---|---|---|
| | Tflop/s | | | | Watt | Gio | Gio/s |
| | Brain16 | Float16 | Float32 | Float64 | Power | Memory | HBM throughput |
| V100 (NVLink) | 125 | 125 | 15.7 | 7.8 | 300 | 32 | 900 |
| A100 (SMX) | 312 | 312 | 19.5 | 9.7 | 500 | 80 | 2039 |
| H100 (SMX) | 989 | 989 | 67 | 34 | 700 | 80 | 3350 |
| MI250X (1GCD) | 191.5 | 191.5 | 23.95 | 23.95 | 280 | 64 | 1638.4 |
| MI300A | 980 | 980 | 122.6 | 61.3 | 760 | 128 | 5300 |
| MI300X | 1300 | 1300 | 163.4 | 81.7 | 750 | 192 | 5300 |

| | Node normalized | | | | | | |
|---|---|---|---|---|---|---|---|
| | Tflop/s | | | | Watt | Gio | Gio/s |
| | Brain16 | Float16 | Float32 | Float64 | Power | Memory | HBM throughput |
| V100 (NVLink) | 1000 | 1000 | 125.6 | 62.4 | 2400 | 256 | 7200 |
| A100 (SMX) | 2496 | 2496 | 156 | 77.6 | 4000 | 640 | 16312 |
| H100 (SMX) | 7912 | 7912 | 536 | 272 | 5600 | 640 | 26800 |
| MI250X (1GCD) | 1532 | 1532 | 191.6 | 191.6 | 2240 | 512 | 13107.2 |
| MI300A | 3920 | 3920 | 490.4 | 245.2 | 3040 | 512 | 21200 |
| MI300X | 5200 | 5200 | 653.6 | 326.8 | 3000 | 768 | 21200 |

| | GPU per node |
|---|---|
| V100 (NVLink) | 8 |
| A100 (SMX) | 8 |
| H100 (SMX) | 8 |
| MI250X (1GCD) | 8 |
| MI300A | 4 |
| MI300X | 4 |

## Après 2 ans d'usage IA au CINES

- Nvidia a toujours l'avantage logiciel
- Mais la quasi totalité des cas d'usages sont fonctionnels et performants sur Adastra

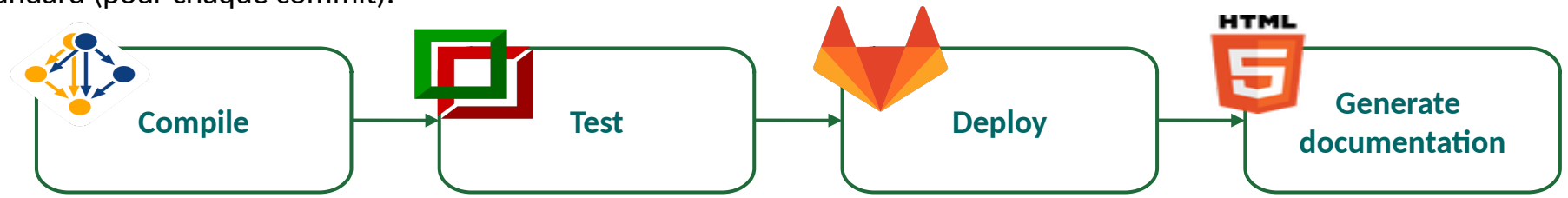
# Une stack logicielle de classe exascale, pensée pour la robustesse

## GAIA, Gestionnaire Automatisé pour l'Installation des Applications : assurer la reproductibilité et la qualité de service pour les environnements utilisateur

**Objectifs** : Déploiement automatique et versionné des piles logicielles d'Adastra, basé sur :

- **GITLAB** : sources/configurations , utilise Gitlab-CI (CI/CD) pour l'automatisation et la documentation (Pages)
- **SPACK** : base pour le déploiement logiciel, peut être complétée par d'autres logiciels
- **REFRAME** : tests de fonctionnalités et performances de la stack

Workflow standard (pour chaque commit):




**Challenges** :être exhaustif ! Surtout avoir toutes les recettes et tests pour les partitions GPU (qui plus est, AMD!)

→ **SharingPartage d'expertise avec les autres opérateurs MI250X mondiaux**
**+ standardisation et efforts internationaux sur Spack + Reframe**

**Des discussions sont aussi en cours avec NUMPEX pour définir les bonnes pratiques de déploiement de stacks logicielles, afin de répondre au mieux au besoins des utilisateurs et développeurs de logiciels de simulation.**

# Portage et optimisation d'application

## Contrat de progrès : pendant l'installation de la machine

- 5 applications ont été adaptées
- Des speed-up jusqu'à **5x**
- Speed-up entre:
  - Noeud Treno MI250 (génération N)
  - Noeud Genoa CPU (génération N+1)

Meso-NH
mesoscale non-hydrostatic model

TRUST

MUMPS

Magic

GYSELA-X

## CINES est aussi membre AST (Application Support Team) EuroHPC pour le HPC et l'IA

Le portage et l'optimisation représente >50% de notre activité support
Ce n'était que ~10% sur les machines classiques CPU précédentes (e.g. Occigen, Intel CPU)

## Enfin, le CINES organise 1 à 2 fois par an un hackathon de portage N'hésitez pas à vous renseigner !

- **Prochaine session → 4 au 9 novembre 2024**
- **Session suivante → Printemps 2025**


CINES

# HPC, exascale, IA, quels enjeux pour le CINES?

## Open Science

La science et la recherche pour tous, par tous

- Open Data, principes FAIR
- Des moyens disponibles, des accès simplifiés

## Le HPC et l'IA pour la société

Des grands enjeux de société

- La santé
- Le climat
- La justice

## Les enjeux de sobriété et de frugalité

Minimisation de l'impact environnemental

- Diminution des consommations énergétiques
- Bilan carbone exhaustif
- Diminution des consommations d'eau

# HPC, exascale, IA, quels enjeux pour le CINES?

## Collaborations et visibilité

A l'international

- ORNL, LLNL (USA), LUMI (Finlande), Pawsey (Asutralie)

En Europe

- 2 projets EuroHPC en cours (Epicure, Minerva)
- D'autres à venir

## Expertise

Partager notre expertise, publier, open sourcer

Consolider les compétences

- Expertise du calcul intensif, à propager à la communauté IA
- Continuer à renforcer les équipes et leurs compétences dans les domaines de pointes
- Assurer une veille technologique permanente



# Focus sur l'énergie/sobriété

## Catégorisation des jobs

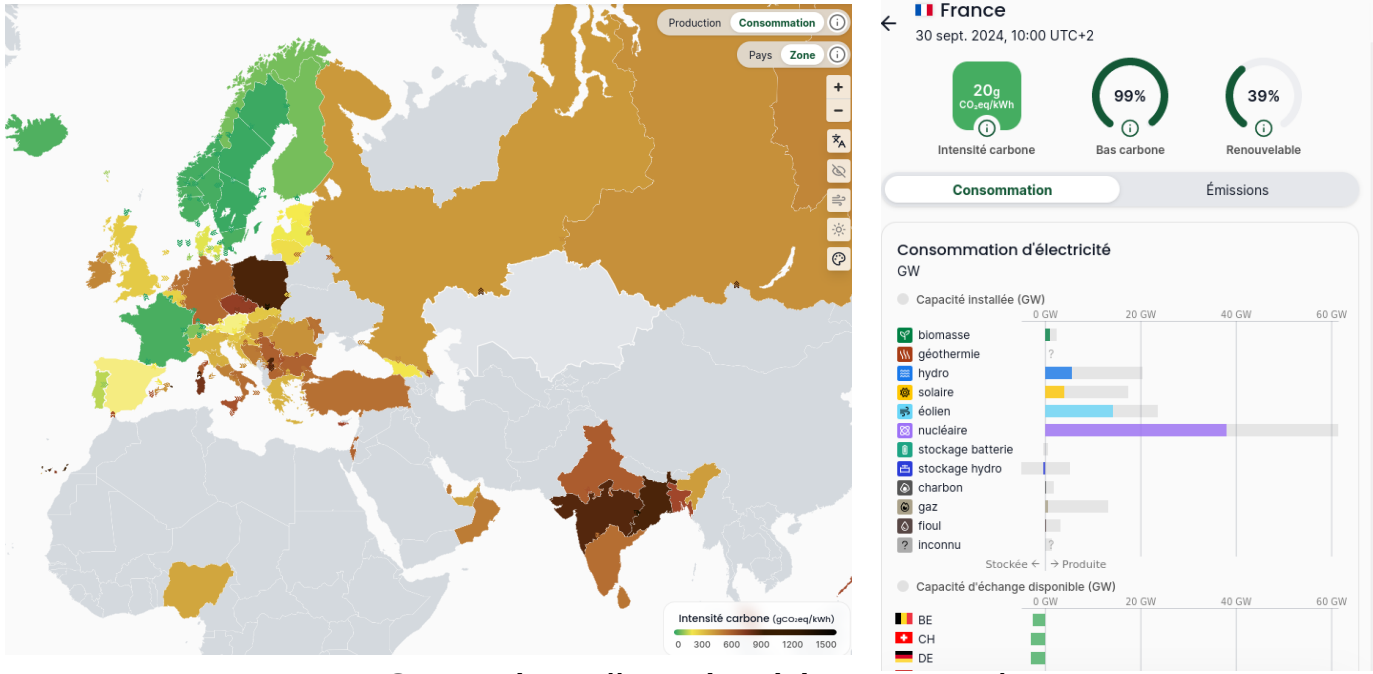
- Surveillance de la conso des jobs, information renvoyée automatiquement à l'utilisateur
- Catégorisation des jobs par conso moyenne au niveau du noeud de calcul

## Baisse de la fréquence

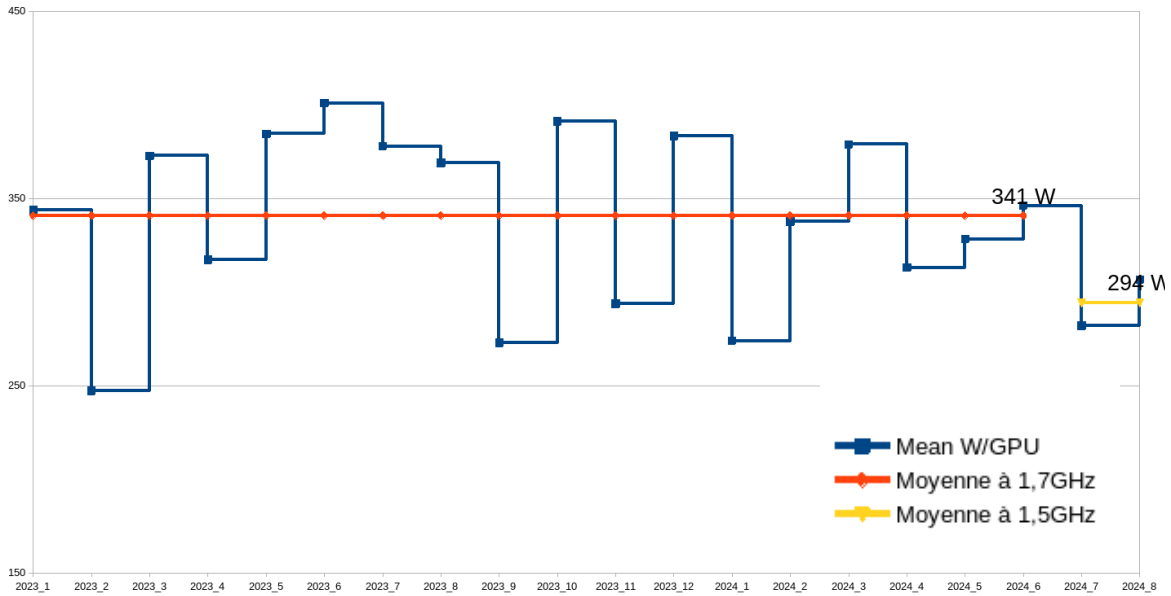
- Effectué le 03 juillet : 1.7GHz → 1.5GHz
- Raison : fortes chaleurs + changement de fluide, limiter l'impact dans un premier temps
- Pas de retour négatif utilisateur
  - HPC, publication SC23 : https://www.cines.fr/wp-content/uploads/2023/11/papier_energy_v3.03-1.pdf
    - Perf : -3 %, Energie consommée -6 %
  - IA, pas encore publié
    - Premiers test : impact en perf 0 %, gain énergétique malgré tout

```
CINES Job Report:
-----------------
o Estimated energy consumption: 826751 Joules
(representing ~ 84% of the maximum nodes utilization)
```

**Exemple de sortie de job, ajout CO2e à l'étude**




**Source: https://app.electricitymaps.com/map**




**Consommation moyenne par GPU pour les jobs Adastra (noeuds idles non comptabilisés)**

# Centre Informatique National de l’Enseignement Supérieur

**Calcul Intensif** | **Hébergement** | **Archivage**


MINISTÈRE DE L’ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE

France Universités

GENCI

Le calcul intensif au service de la connaissance

## Merci!

## Questions?

Gabriel Hautreux

Responsable du département calcul intensif