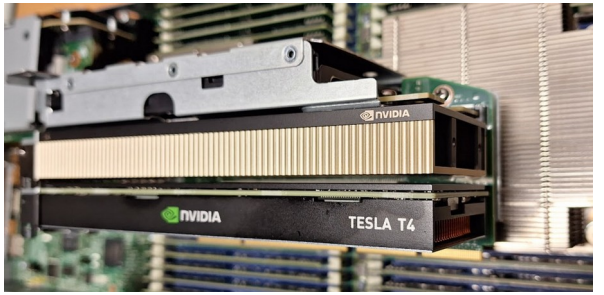
CBPsmn

Centre Blaise Pascal de Simulation et de Modélisation Numérique

JCAD 2025

Internalisation de ressources IA : Comment concilier (au mieux) Sobriété environnementale & Souveraineté technologique

Emmanuel Quémener



Eléments de l'intervention : un Retex

- Le « cadre » de travail : une structure « 100 % internalisée »
- De l'empreinte environnementale des équipements : le CO_2
- Utiliser du « vieux » pour offrir de nouveaux services
  - Pertinent ? Mais Comment ?
- « Utiliser l'IA » : le nouveau « utiliser l'ordinateur »
  - De l'inférence à l'entraînement : de la diversité des usages au matériel
- La GPU comme pierre angulaire, mais pas seulement :
  - Le @TheEdge comme paradigme
- L'IA « soutenable » : de la réexploitation de tout partout
- L'IA « souveraine » : des dépendances multiples

Informatique Scientifique ENS-Lyon
CBPsmn : fusion de CBP & PSMN

Centre Blaise Pascal

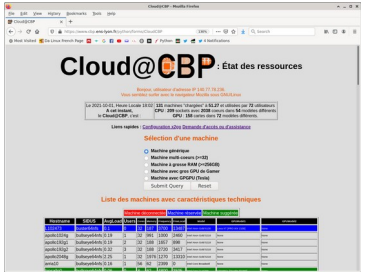
- Hôtel à projets, équipes, ...
- …, conférences et formations
- Centre d'essais (& opérationnel)
  - Reproductibilité
  - Adaptabilité
  - Diversité
  - Intéractivité
  - Simplicité

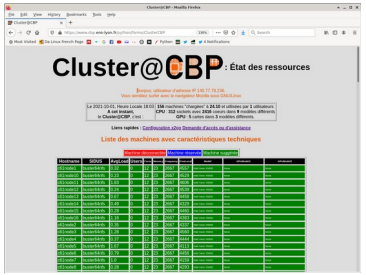
Pôle Scientifique de Modélisation Numérique

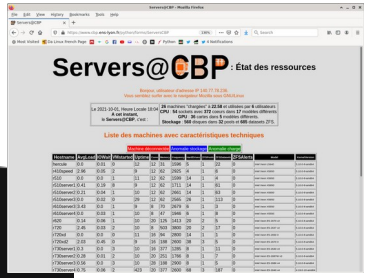
- Mésocentre de Calcul : Tier-2
- *High Performance Computing*
- Localisé au Data Center ENS-Lyon
- Calcul par « Batch » : non interactif
- Entièrement internalisé depuis 2012
- Entièrement Open Source depuis 2012
- ~600 nœuds, ~30000 cœurs
- ~11 PetaBytes de stockage

Au service de toutes les disciplines à l'ENSL (et ailleurs)...

Centre Blaise Pascal @ ENS-Lyon
au service des Recherche & Enseignement

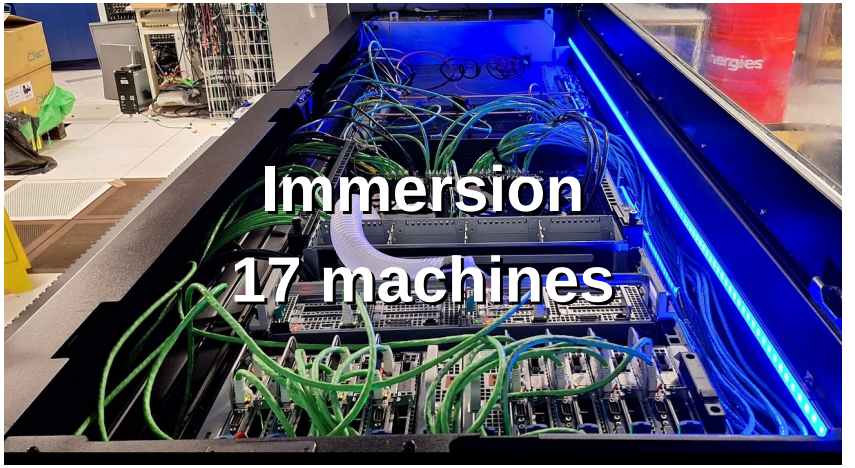






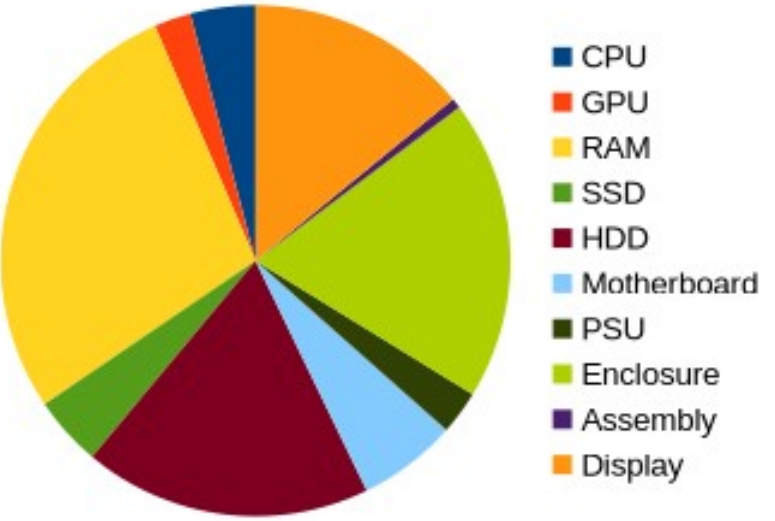
- Près de 330 machines en internalisation complète
  - Près de 8000 coeursCPU, plus de 11000 coeursGPU,
  - 72 TiB RAM, près de 6PB dans 1300 HDD & 113 SSD,
- 70 % d'équipements d'occasion (toutes sources)
- 90 % **hors garantie**
- Disponibilité supérieure à 99.8 % ... et 1 BOFH...
- Coût licences logicielles de l'unité : 0€
- Socle Open Source : **Linux Debian** sous SIDUS

La « meute » de machines du CBP



Quelle empreinte environnementale en CO_2 ? Au JCAD l'an passé...

- **Fabrication** : 337 machines (2025Q2)
  - 10 % sous garantie, 1/3 achetées neuves
  - 7934 coeursCPU, 11255 coeursGPU,
  - 72 TiB RAM, 5.8 PB, 1297 HDD, 113 SSD
  - 352 tonnes CO_2 à la fabrication (approche #3)
  - **RAM+Chassis+HDD+Ecrans : 80 % du total !**



- **Exploitation** : 330 avec 150W H24 7j/7
  - 10 tonnes de CO_2 par an soit une française (20 g/Kwh 2024)
- « Amortissement » carbone : ratio de 30 pour 1…

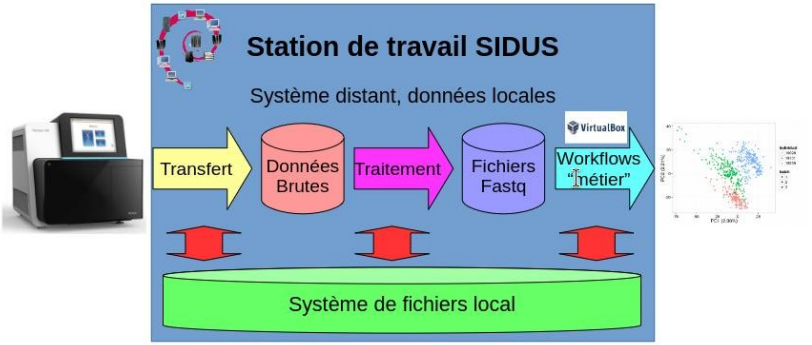
Intérêt certain à exploiter le **plus longtemps** possible...

Déjà, aux JCAD 2018
La solution HPDA@TheEdge...

Où allons-nous ?
Prendre un certain contre-pied...

- Rapprocher les ressources de traitement de la production
- Adapter les systèmes à la véritable nature des traitements
- Assurer la reproductibilité des traitements
- Rationaliser l'exploitation de ressources locales
- Limiter le temps d'administration système de postes individuels
- Limiter l'empreinte énergétique
- Limiter les délais liés au transfert réseau
- Simplifier l'accès à l'archivage
- Recentrer les informaticiens sur des fonctionnalités spécifiques.

Cas d'usage : séquenceur Illumina
Approche « HPDA@TheEdge »



- ***WIP*** : pipeline de Illumina uniquement sous Windows 10

Au coeur de l'IA, la donnée : ces spécifications en héritage

Artisans de l’IA : évolution des GPU

GPU : une « inflation » récente

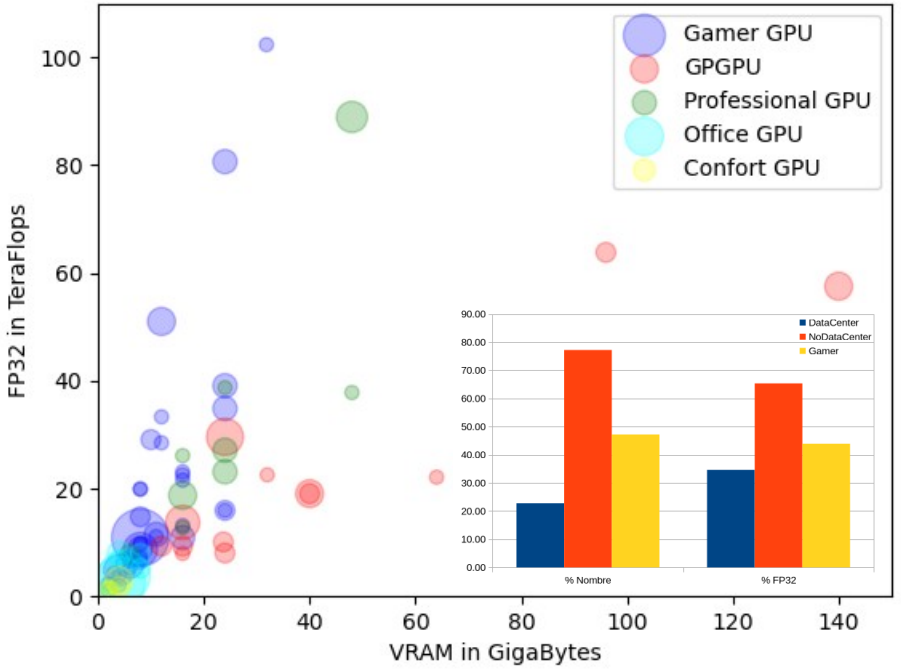
Entre 2019 et 2025 :

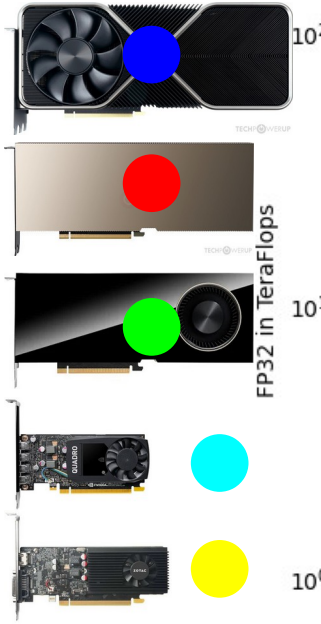
- RTX 2080 Ti à RTX 5090
- De Turing à Blackwell : +4
- De 1237 à 3372 cm³ : ×2.5
- De 250W à 580W : ×2.3
- De 1000€ à 2000€ : ×2
- Hôte 600W à 1200W : ×2
- Performances FP32 : ×4.5

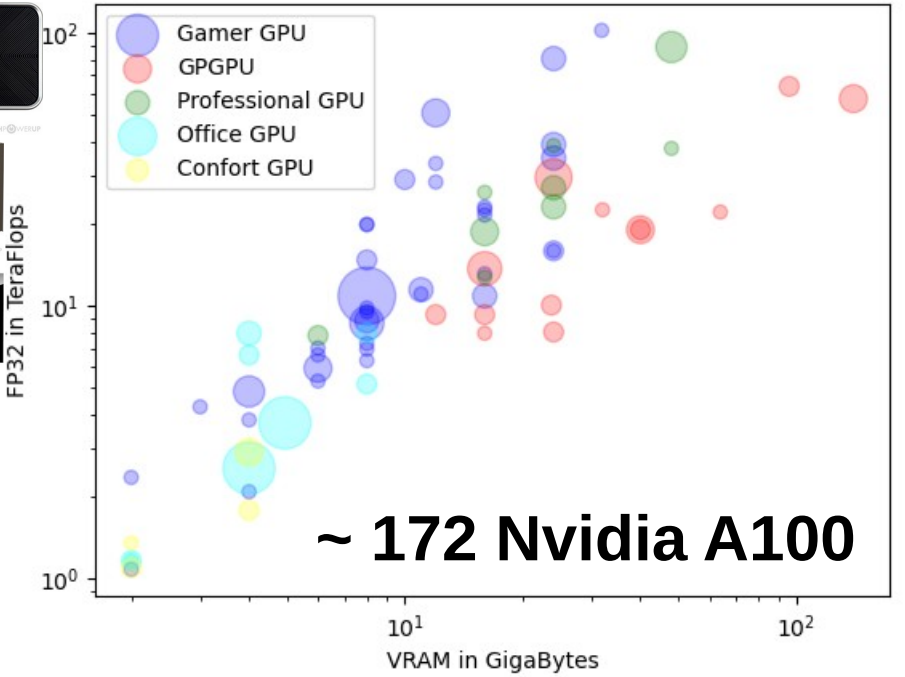
Plus d’intégration de « Gamer » dans « Data Center »...

Artisans de l'IA : les GPU au CBP

180 GPU & 5 catégories différentes !



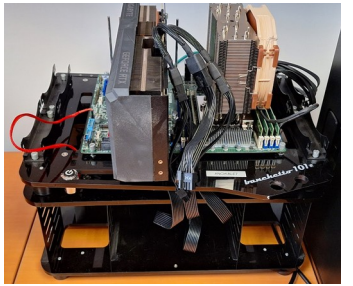




Des caractéristiques différentes pour des usages différents

3 approches pour du @TheEdge : 3 socles, 2 GPU

- Une « grosse » station Dell ou HPE :
  - La nécessité d'un capot « ouvert »
- Une machine « *home made* » : « *gamer like* »
  - Des boitiers à des tarifs imposants,
  - Des ventilateurs en pagaille
- Une machine « ouverte » :
  - Parfaite pour accueillir les plus volumineuses des GPU
  - Plus compliqué pour y mettre des disques durs



Offrir du LLM : ollama « online »

- Simplicité d'installation & de mise à jour
- Stockage « local » des modèles : prévoir de l'espace !
- Exploitation via le terminal (donc SSH, copier/coller, etc) :
  - Mais APIs, notamment Python, disponibles (donc tu « pip » es...)
- Fonctionnement universel :
  - Nvidia (driver supérieur à 500), AMD (avec ROCM), CPU
- Vocation pédagogique
  - Associer la VRAM disponible à la dimension du modèle…
  - Avec « nvidia-smi dmon » : visualiser la charge GPU à l'interrogation
  - Avec « dstat » : juger des entrées/sorties pour la récupération du modèle
  - Avec « RaplWattQueue.sh » : voir la consommation processeur en RT

GPU de « Gamer » : pertinentes ? Comment les « comparer » en IA ?

- Avec une question à un LLM :
  - Exploitation des métriques systèmes nvidia-smi
  - Impossibilité d'évaluer une pertinence de la réponse « rapidement »
- Choix d'une IA générative : une image par un prompt
  - Facilité de programmation, instantanéité de vérification de pertinence
  - Efficacité fonction du nombre d'inférences
- Les modèles : la référence des « stable-diffusion »
  - Trois modèles : 1.5 (5 GB), 2.1 (5 GB), 3.5-medium (16 GB)
- Le comparatif : nombre d'inférences par seconde
- Les autres critères : les « coûts » énergétique & financier

Le banc de test : un prompt simple & des résultats intrigants...

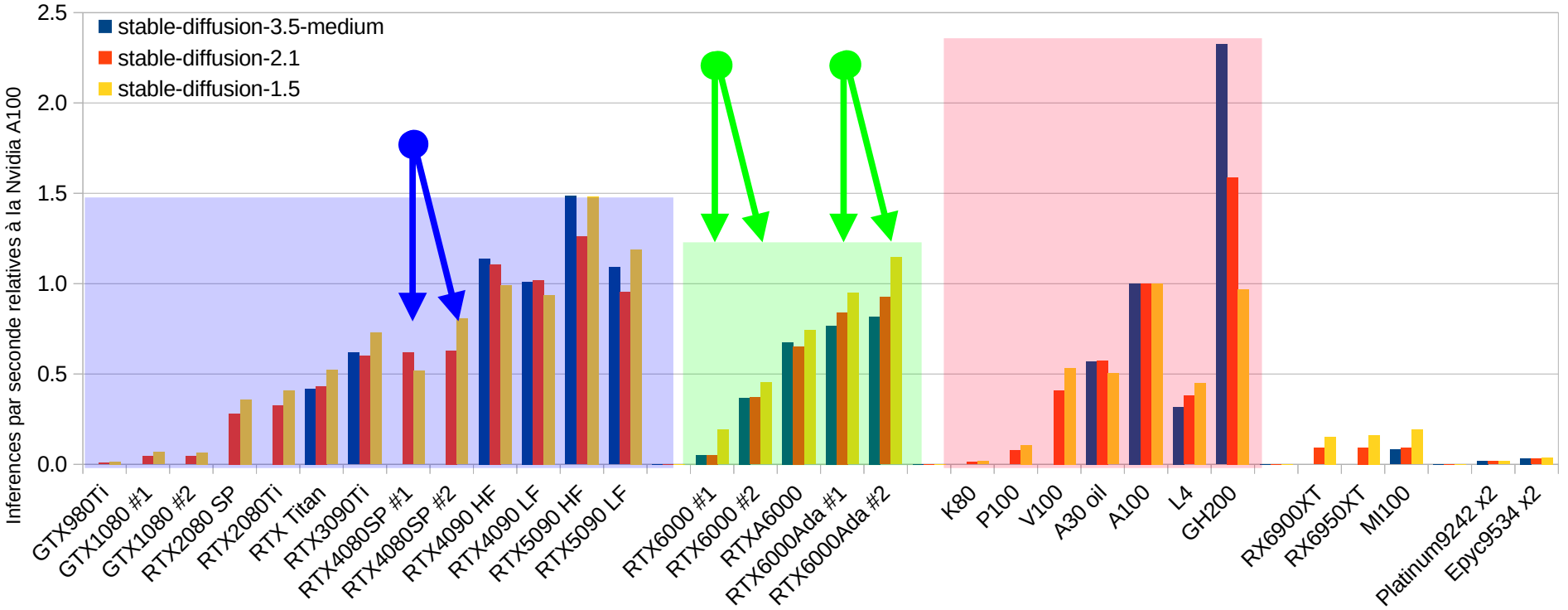
Le prompt : « *The Mona Lisa, painted by Leonardo da Vinci, depicts a woman with an enigmatic smile, seated with her hands clasped against a distant landscap. Her eyes seem to follow the vewer, with subtle details and delicate sfumato.* »

Stable Diffusion 1.5

Stable Diffusion 2.1

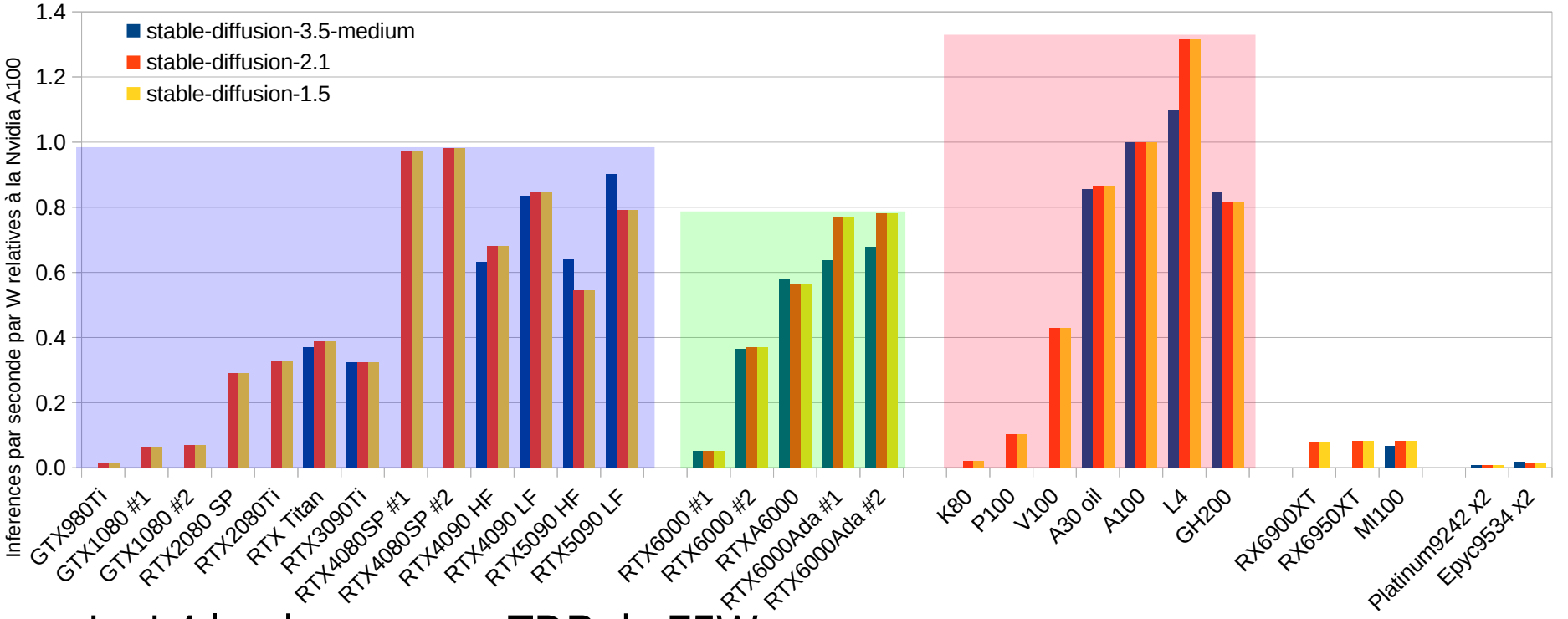
Stable Diffusion 3.5 medium

Inférences/s relatives à A100



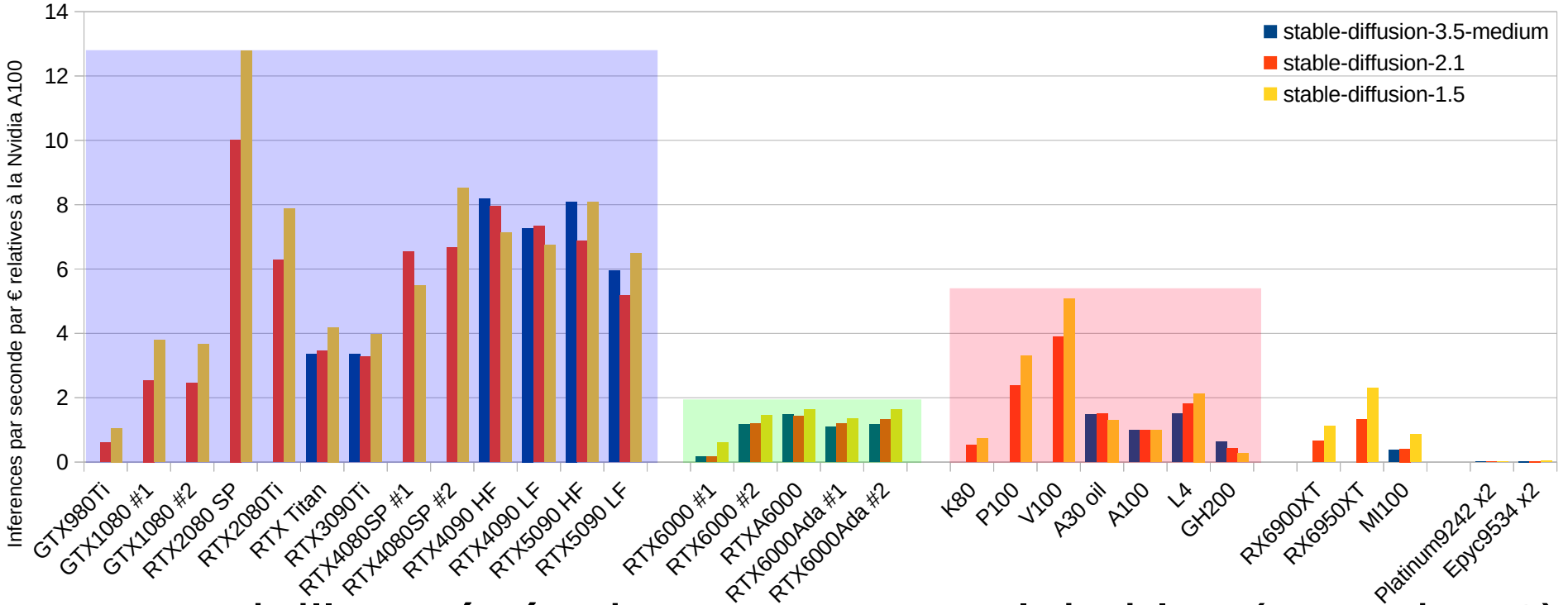
- Des machines exclues (VRAM < 24GB)
- La bonne résistance des cartes de « Gamer »
- La nature de la VRAM (si HBM) comme critère
- La nature du « socle » importante suivant le cas...
- Les GPU AMD hors course
- Les machines à « grosses » CPU indigentes
- L'utilisation de « vieilles » GPU pertinentes
- Les fréquences avec une certaine influence

Inférences/Joules relatives à A100



- La L4 leader avec sa TDP de 75W
- Les RTX4080 à 800€ comparables à une A100
- Les « grosses » professionnelles avec leur limitation de TDP

Inférences/s/€ relatives à A100

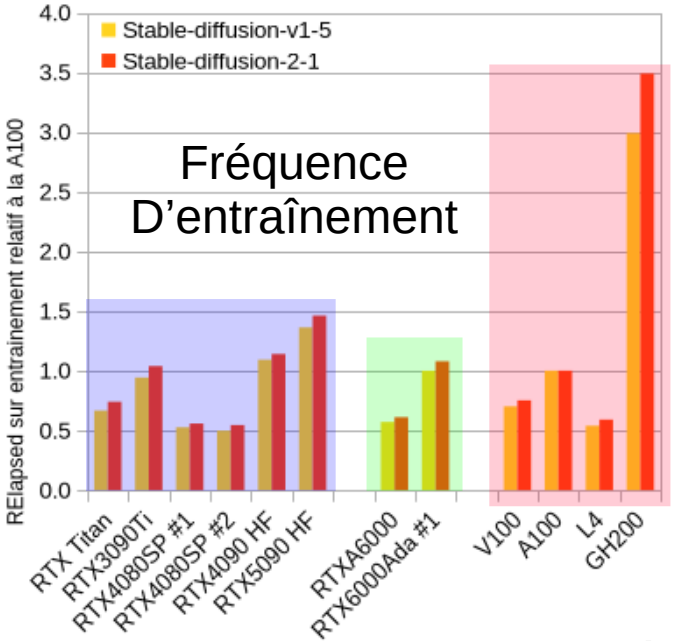


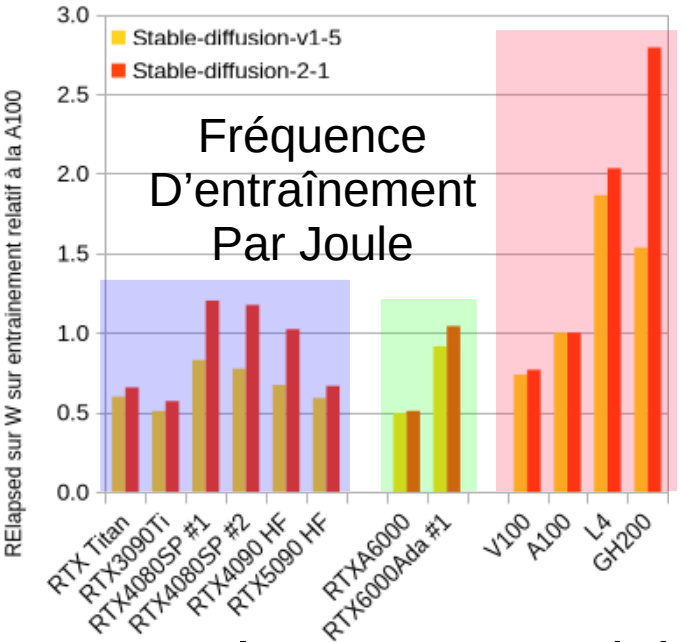
- Les vieilles générations encore exploitables (occasion ?)
- Les grosses cartes de Gamer comme bonnes alternatives

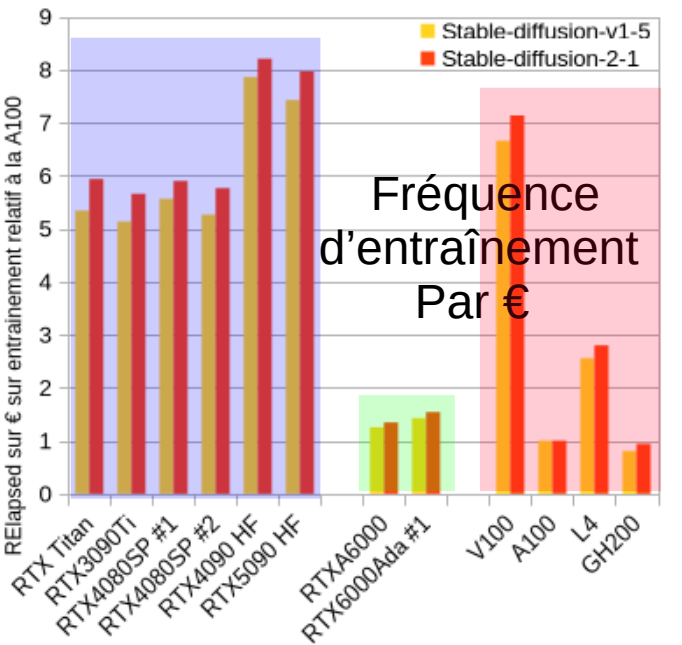
Et pour entraîner un modèle ?

- Dans notre contexte : générer des images sur une base
  - Regrouper des images & les annoter : ici 250
  - Les appliquer aux modèles stable-diffusion 1.5 et 2.1
    - Pas disponible (à l'époque) sur stable-diffusion 3.5
- Les conditions d'application :
  - L'empreinte VRAM : pour 5 GB, 24 GB de VRAM nécessaires
    - Exclusion de facto des machines avec moins de 16 GB de VRAM
    - Possibilité d'apprentissage avec 16 GB avec des astuces (coûteuses en qualité)
  - L'optimisation en vitesse d'exécution :
    - Si VRAM disponible : parallélisation sur la « batch size » :
    - batch_size=1 pour 24 GB, batch_size=128 pour 96 GB

Et pour des entraînements ?







- Beaucoup de GPU exclues : 16GB minimum
- La GH200 & sa HBM2e en tête, mais à quel prix
- Les cartes de « gamer » étonnemment groupées
- La nature du modèle « compte » dans l'efficacité

Conclusion : internaliser son IA

Possible en étant « sobre », mais ...

- Les vieux chassis, serveurs ou stations : upgrader CPU, RAM, HDD
  - Dans le **serveur** : des GPGPU Nvidia L4 avec 24GB à 75W de TDP (2000€)
  - Dans la **station travail** « musclée » :
    - une GPU « gamer » (VRAM 32GB, 2200€) ou une GPU « pro » (VRAM 48GB à 7000€)
  - Dans un **poste de travail** (salle de formation) :
    - une GPU « office » (VRAM 4GB à 200€)
- Les nouveaux serveurs (à GPGPU) : (VRAM 140GB à 22k€)
  - Associer des baies SAS avec de vieux disques SAS : 10k€ d'économies...
- Les nouvelles « anchiales » : boîtiers ouverts
  - Intégrer des disques SAS sur les contrôleurs disponibles

Favoriser la diversité et leur appropriation !

Appel aux dons pour la Computhèque de l'ENS-Lyon

- Tout vieil équipement informatique :
  - Les vieux 8 bits des années 1980 :
    - Sinclair ZX, Commodore, Oric, ...
  - Les vieux PC avec des cartes ISA : 80286, 80386
  - Tout équipement informatique un peu exotique :
    - Machines de technologie : Dec Alpha 21264, HPPA, ...
  - Vieux laptops, …
- Vocation pédagogique & scientifique
  - Fête de la science, ateliers 3IP, …
  - Préservation patrimoine scientifique & technique...