JCAD 2022

Refroidissement par immersion de serveurs : premiers retours opérationnels

Emmanuel Quémener

Back2Basics : pourquoi immerger ?
L'« effet Joule commandé » a un coût

- Grandeur et décadence de la fréquence
  - Entre 1981 et 1999 : de 4 MHz à 400 MHz x100 en ~20 ans
  - Entre 1999 et 2004 : de 400 MHz à 3 GHz x~10 en 5 ans
  - Entre 2004 et 2009 : de 3 GHz à 2 GHz
- ***Thermal Design Power*** : enveloppe thermique de dissipation maximale
  - **TDP = ½ C V² f** avec C = **Capacitance**, f = **fréquence**, V = **tension d'alimentation** (fonction de f !)
  - Capacitance = Finesse² . Nb Transistors . Constante de Mylq (~ 0.015)
- TDP pour un processeur : jusqu'à 350 W (sur 12 cm²)
  - Densité de chaleur d'une plaque à induction !
- TDP devient le facteur limitant de puissance (de traitement)

Ceux qui ont la chance de ne pas me connaître … ne connaissent pas mon appétance...

… pour l'histoire (des sciences) !

Il y a une génération (humaine)... Un film de série B en 1984

- 1984 : The Last Starfighter
  - 27 minutes d'images synthétiques
  - ~ 30.10^9 operations par image
  - Utilisation d'un Cray X-MP (130 kW)
  - 68 jours (en fait, 1 année nécessaire)

- 2020 : RTX 3090 (350 W)
  - 33 secondes
  - Comparison RTX 3090 / Cray
    - Performance : 178 000 !
    - Consommation ~ 66 000 000 !

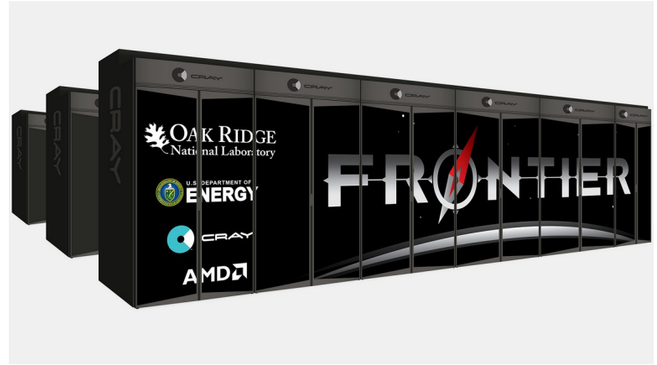
Emmanuel QUÉMENER CC BY-NC-SA
December 6, 2021

CBP

ENS

- Avec **Cray X-MP** :
  - 1 Gflops, 130 kW
- Avec **HPE Frontier** :
  - 1 Eflops, 21100 kW
  - 9248 nœuds
  - 36992 MI250X de 500 W
  - 94 % de Rpeak dans GPU
  - 87 % du total de la consommation annoncée...

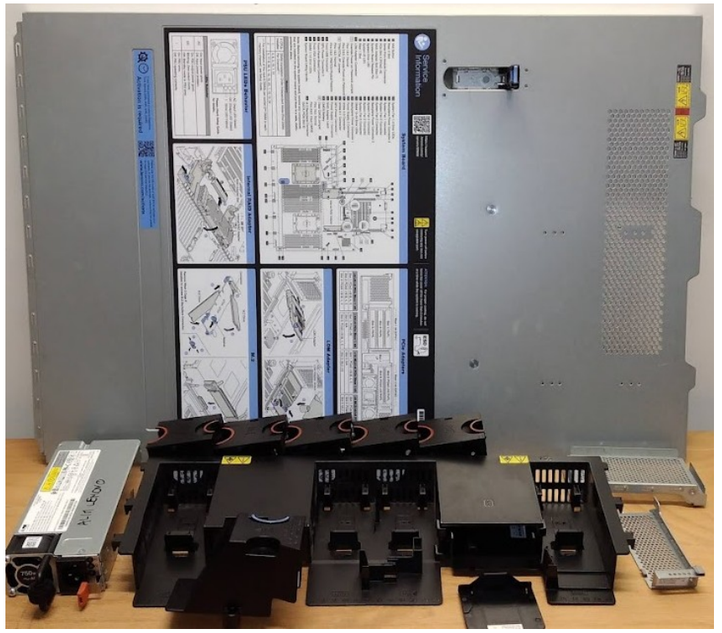
> The dense concentration of components requires special cooling techniques to overcome the accompanying problems of heat dissipation. A proven, patented cooling system using liquid refrigerant maintains the necessary internal system temperature, contributing to high system reliability and minimizing the need for expensive room cooling equipment.



Retour aux « sources » : refroidir les GPU dans un fluide?

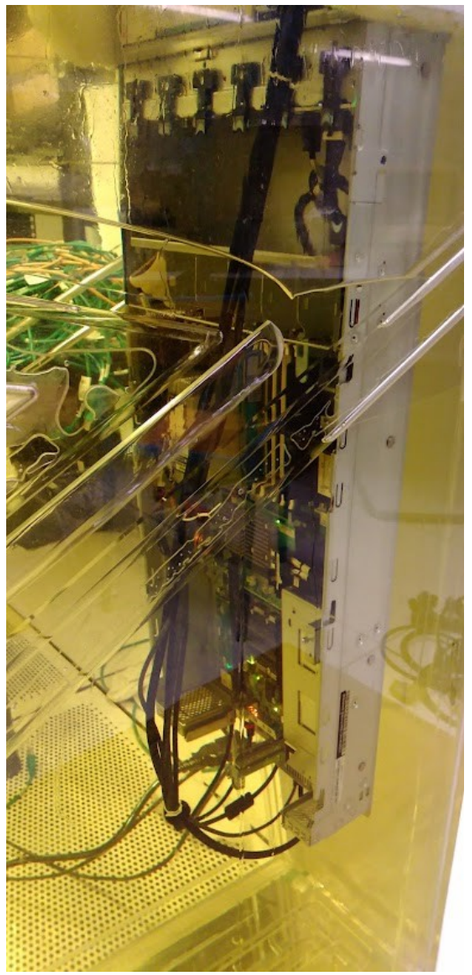
Préparation pour l'immersion : supprimer « tout ce qui bouge »

- Dans l'air : évacuer les « calories » :
  - On multiple la surface de contact : x200 pour un radiateur
  - On diffuse la chaleur par conduction ou convection (caloduc)
  - On chasse au ventilateur l'air chauffé :
    - Dans un serveur : ventilateur de 4 à 6 cm, rotation de 3000 à 20000 tours/min
    - Dans une station : 8 à 16 cm, rotation de 500 à 1000 tours/min
- Pour préparer les machines :
  - Supprimer la pâte thermique : processeur et radiateur en contact direct
  - Supprimer les ventilateurs (là en le conservant dans l'alimentation)
  - Supprimer les « guides » plastiques



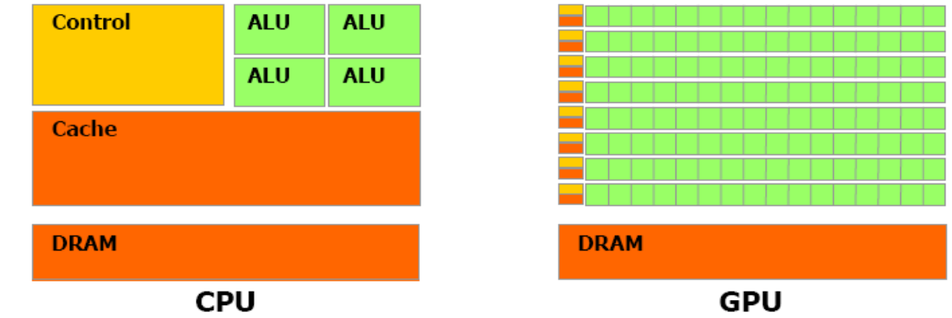
L’immersion sans modification (logicielle)
Sans ventilateur, ça démarre & ça tourne...

- Au repos : oil 74 W, air 68 W
  - Températures : oil 24°C, air 25°C
- En charge : oil 142 W, air 142 W
  - Températures : oil 36°C, air 38.5 °C
  - Mais en IPMI et AC : 140 W pour les deux
  - Mais en IPMI et DC : oil 115 W, air 120 W
- Côté performance :
  - 1 % de différence…
- Mais : processeur « petit », mémoire « faible »

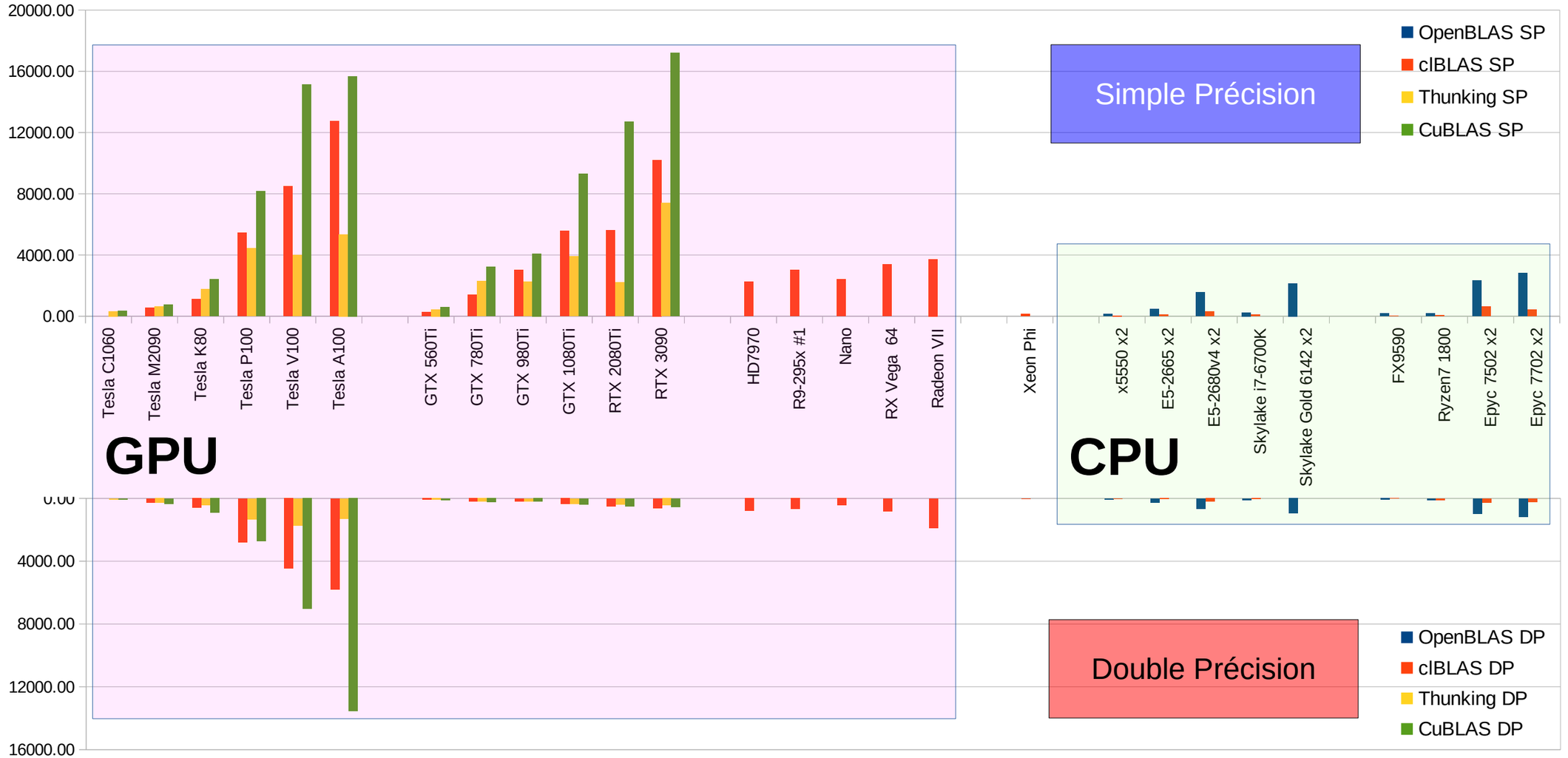


Pourquoi le GPU est-il si puissant ? Parce que il dispose :

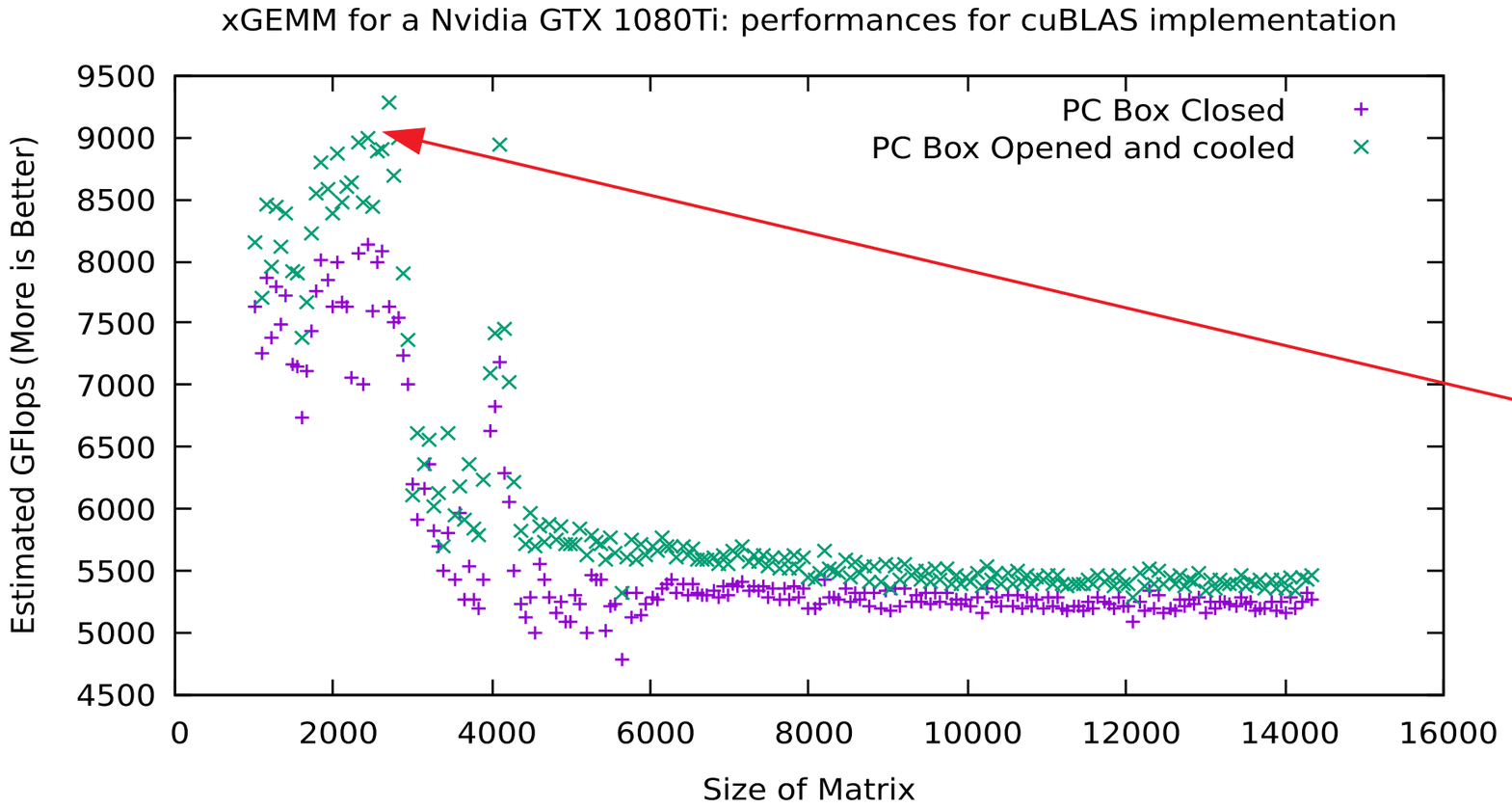
- de milliers d'ALU (unités arithmétiques & logiques)
  - Un CPU, 64 coeurs, 16 ALU par coeur. Le GPU, une myriade (~10000) d'ALU
- d'une RAM une bande passante énorme
  - Un CPU, RAM de ~100 GB/s. Un GPU : ~1 TB/s
- D'une TDP en croissance constante :
  - Un CPU, de 80W à 225W. Un GPU : plus de 350W (déjà 208W en 2009)



La puissance comparée entre GPU & CPU illustrée sur un cas simple (mais orienté...)

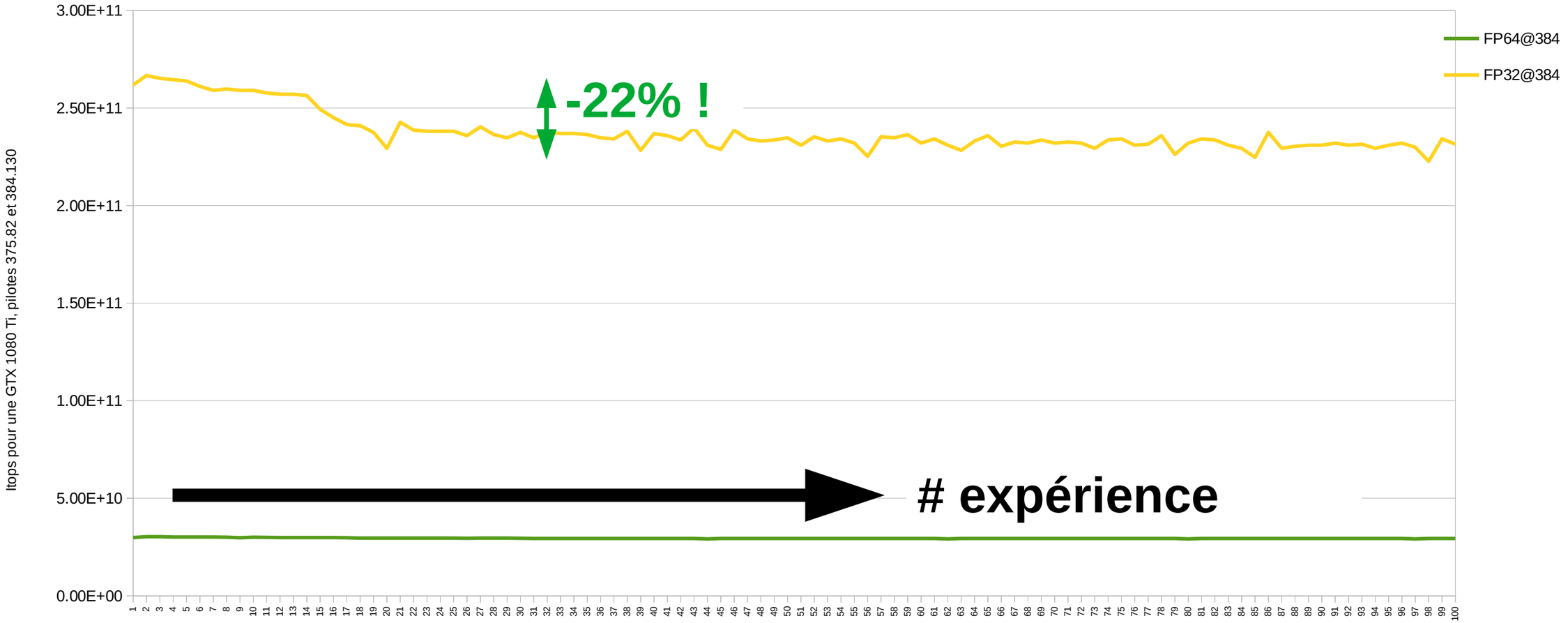


Un mariage impossible : Performance & Température



- Les mêmes socles, cartes, systèmes, et 20 % de différence !
  - De l'importance des conditions climatiques durant l'expérimentation...

L'enveloppe thermique, « Je chauffe donc je ralentis »



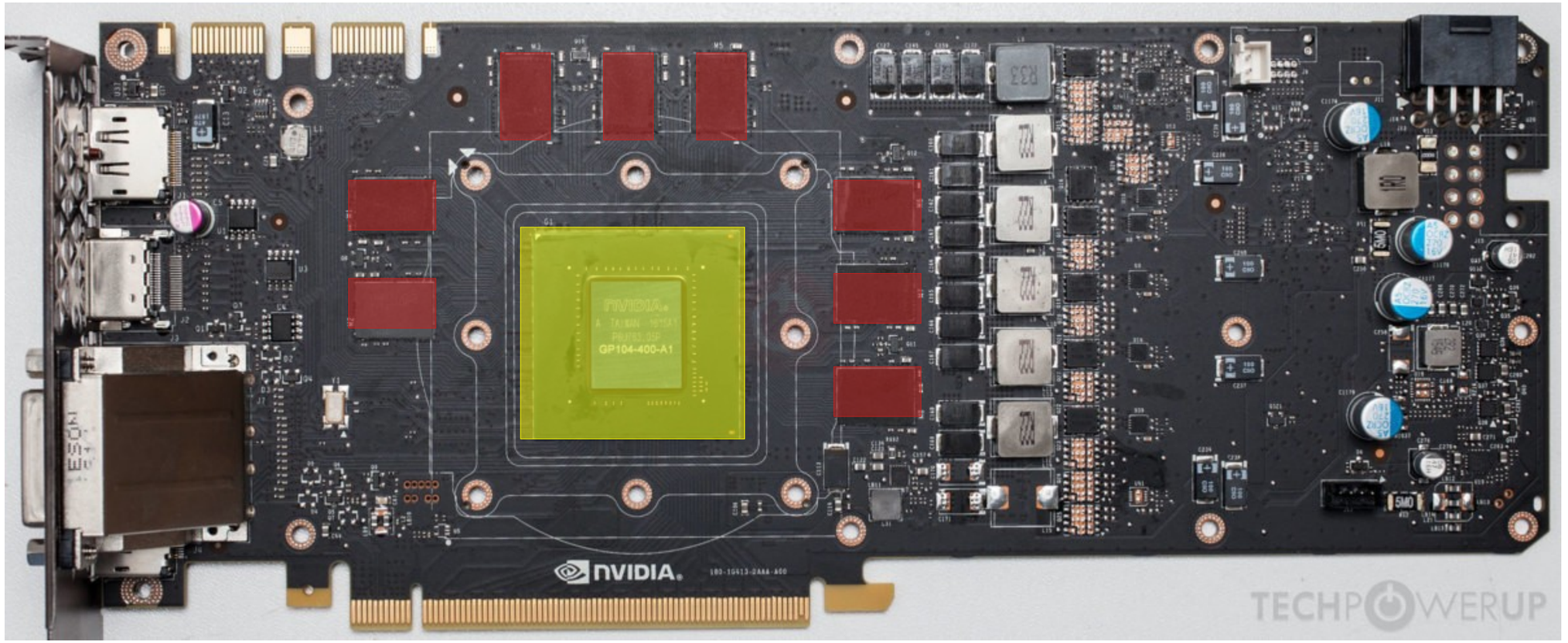
A bien prendre en compte dans l'expérience !

Expérience GPU : le banc d'essais

- Socle matériel : 2 machines Oil/Air (donc **référence**) comprenant :
  - 1 CPU Epyc 7252 : 8 coeurs Rome à 2.8 GHz
  - 2 barrettes de 32GB de RAM
  - **1 GPU Nvidia GTX 1080 : circuit Pascal (génération N-4)**
  - 1 Stockage local SSD de 256GB

- Socle logiciel : SIDUS pour une **reproductibilité parfaite** (et simple, ...)
- Applications de « test » : gros grain, grain fin, métier
  - Gros grain : Pi Monte Carlo (toutes tâches indépendantes, Python/OpenCL)
  - Grain fin : calcul N-Corps (tâches indépendantes à chaque pas, Python/OpenCL)
  - Application métier : Genesis (programme Trhybride : MPI, OpenMP, CUDA)

Le GPU : à « préparer » comme une machine



GPU

DRAM

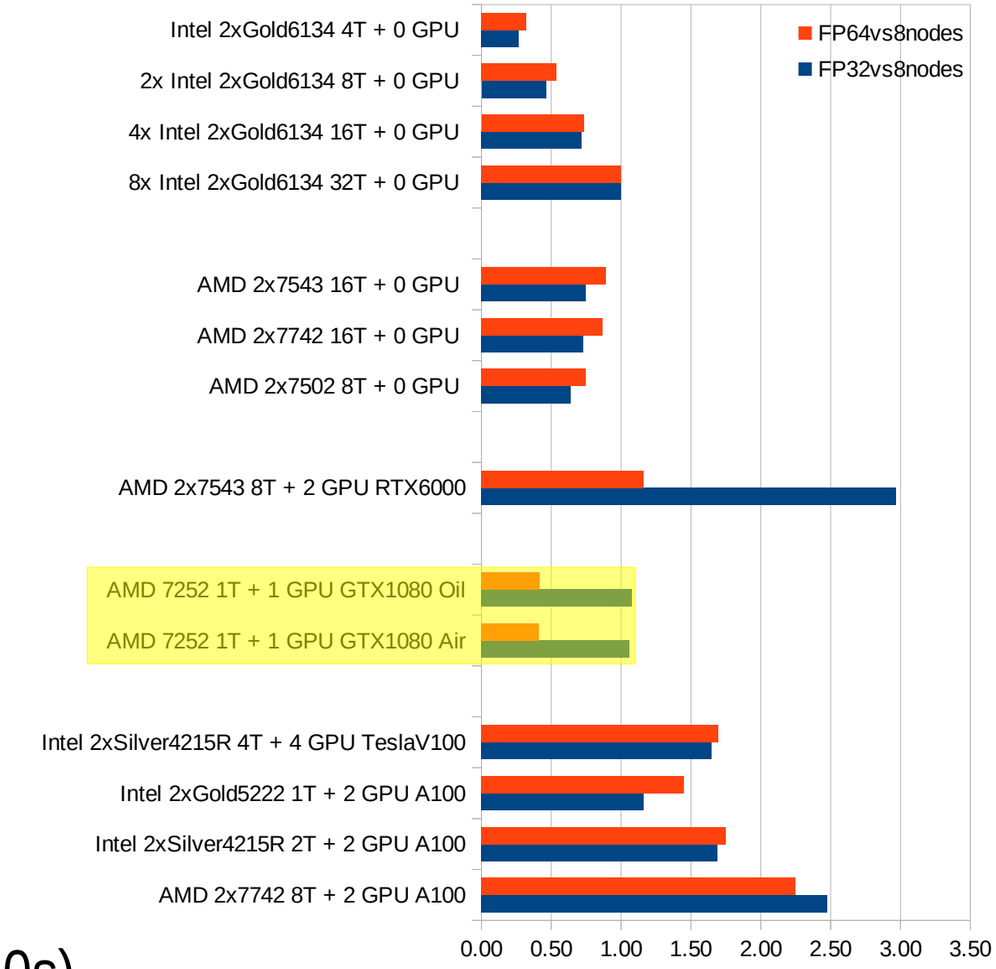
Un refroidissement nécessaire sur GPU **&** circuits DRAM

Les expériences : avec un Pi Monte Carlo… Un code « gros grain » sans accès mémoire...

- Au repos, air (30°C et 10W), oil (30°C et 14 W) : le GPU/oil consomme plus
- Phase 1 : GPU sans radiateur dans Oil
  - Montée rapide en température (jusqu'à 90°C)
  - Plantage du GPU après 3 secondes (mise en sécurité)
- Phase 2 : GPU avec radiateur (sans pâte thermique)
  - Très bonne stabilité à la charge, 12°C de moins entre oil & air
  - Consommations & Températures : Oil/Air : 178W en soutenu, mais Oil/Air : 70°C/82°C
  - Fréquence : 1809 MHz en Oil mais 1759 MHz en Air
  - Performances : comparable en FP32 (516s/519s), 2 % plus rapide en FP64 (466s/474s)
  - Descente de température très lente dans l'air : 1m pour Oil, plus de 10m pour air

Expérience sur code métier trhybride : GENESIS pour comparer CPU/GPU/Cluster

- Un code Open Source
- Trhybride : MPI/OpenMP/CUDA
- Un cas d'usage « stressant » :
  - Grosse sollicitation mémoire
  - Bonne exploitation GPU
  - Référence d'exécution sur « grand centre »
- Pour les machines Air/Oil :
  - Performances comparables en FP32 : (32104s/32050s)
  - Performances meilleures de 2 % en FP64 : (109066s/110860s)



Des consommations plutôt stables… Sur serveurs Lenovo

- Une consommation comparable (modulo les variations de tensions d'entrée)

- Mais un « facteur de puissance » stable, en progression avec la charge

Des consommations plutôt stables…
Sur les « Epyc à GPU »

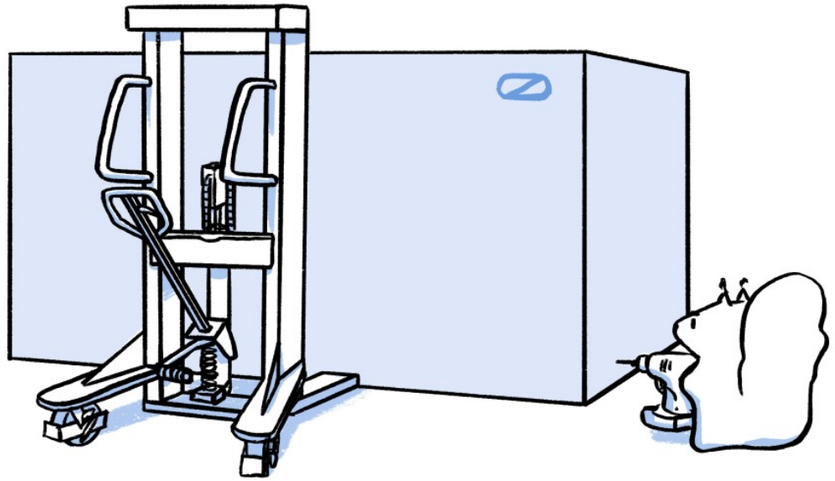
- Une consommation toujours comparable (modulo les variations de tensions)

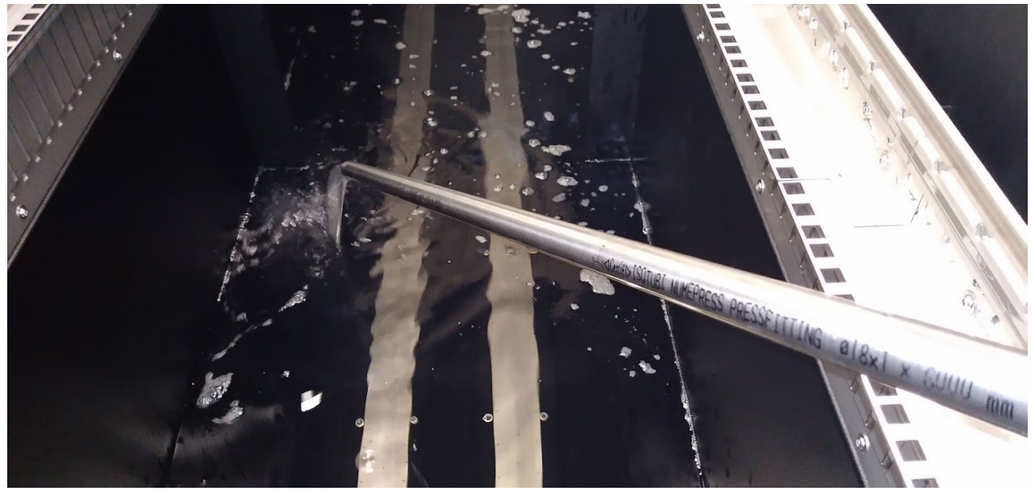
- Mais de grosses variations sur le « facteur de puissance » (et donc...)

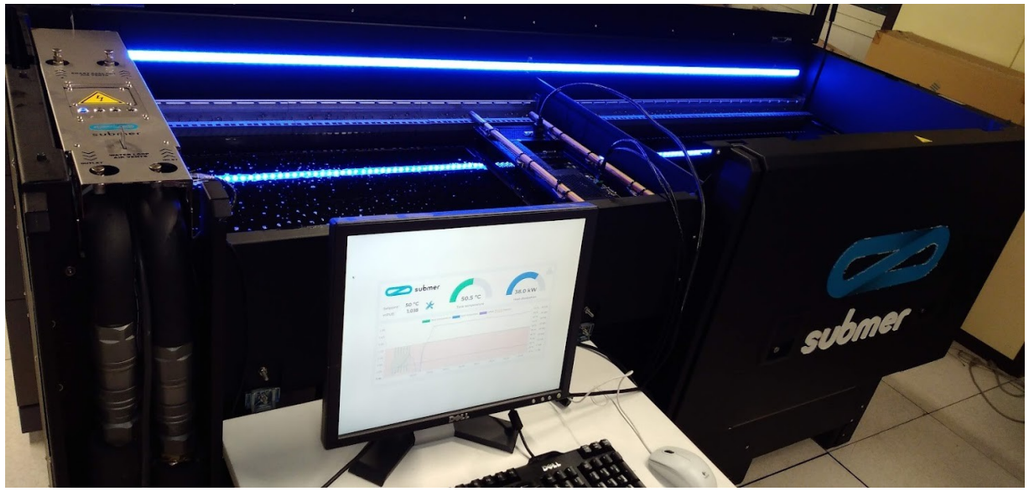
Etude sur nœuds de cluster par le PSMN
Laisser la fréquence monter (et se stabiliser)

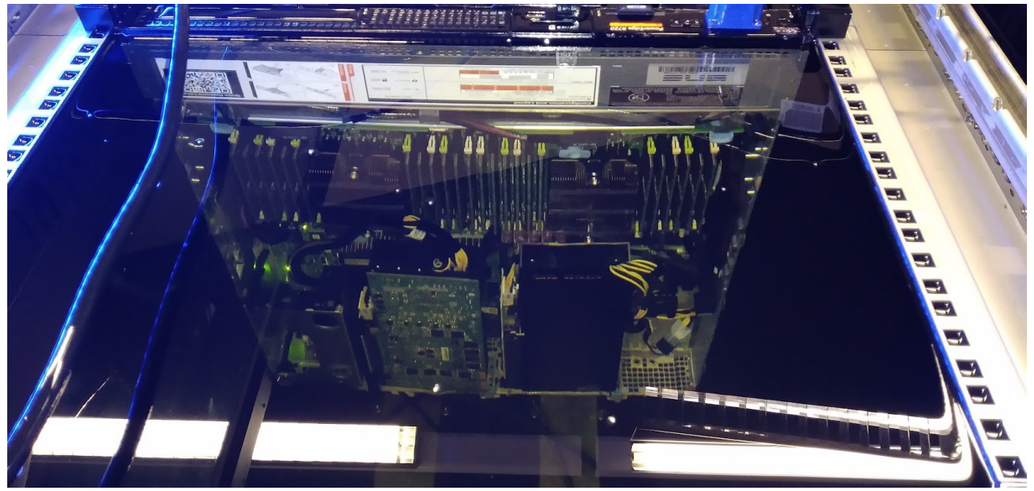
- Banc d'essai : 2 chassis avec chacun 4 lames HPE Apollo avec 2 Intel Xeon 5218
  - Dans l'air, température de 20°C (allée froide), forçage de la fréquence à 2.8 GHz
  - Dans l'huile, températures de 30 à 55°C, fréquence « laissée » variable, de base à 2.3 GHz
- Expérimentations :
  - Menées sur 48h
  - Consommations : Huile (CPU à 2.8 GHz) 1240W, Air (CPU à 2.3 GHz) 1370 W
- Conclusions :
  - Consommation en baisse de 10 % pour une performance en hausse de 20 % !
  - Support des processeurs dans l'huile jusqu'à 80°C (sauvegarde technique)
  - Plateau d'efficacité pour de l'huile à 40°C
  - Fréquence « laissée » de 2.3 GHz permettant de « rester » à 2.8 GHz dans l'huile

Avec un bac Submer, premières analyses





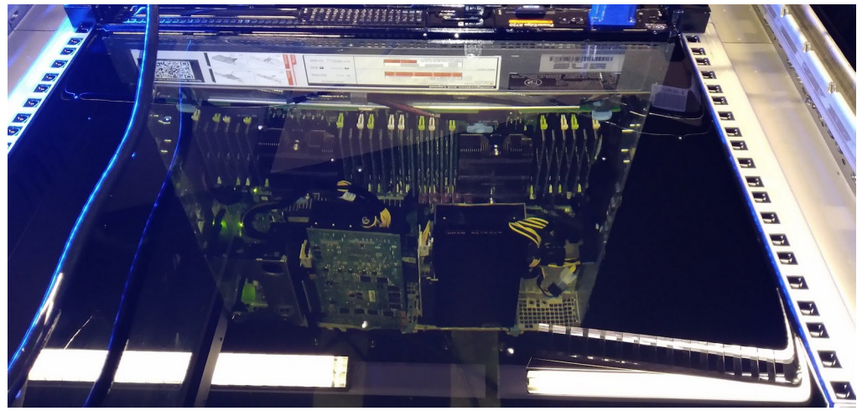




Nouveau bac, nouvelle huile, nouvelles machines, GPU et études !!!

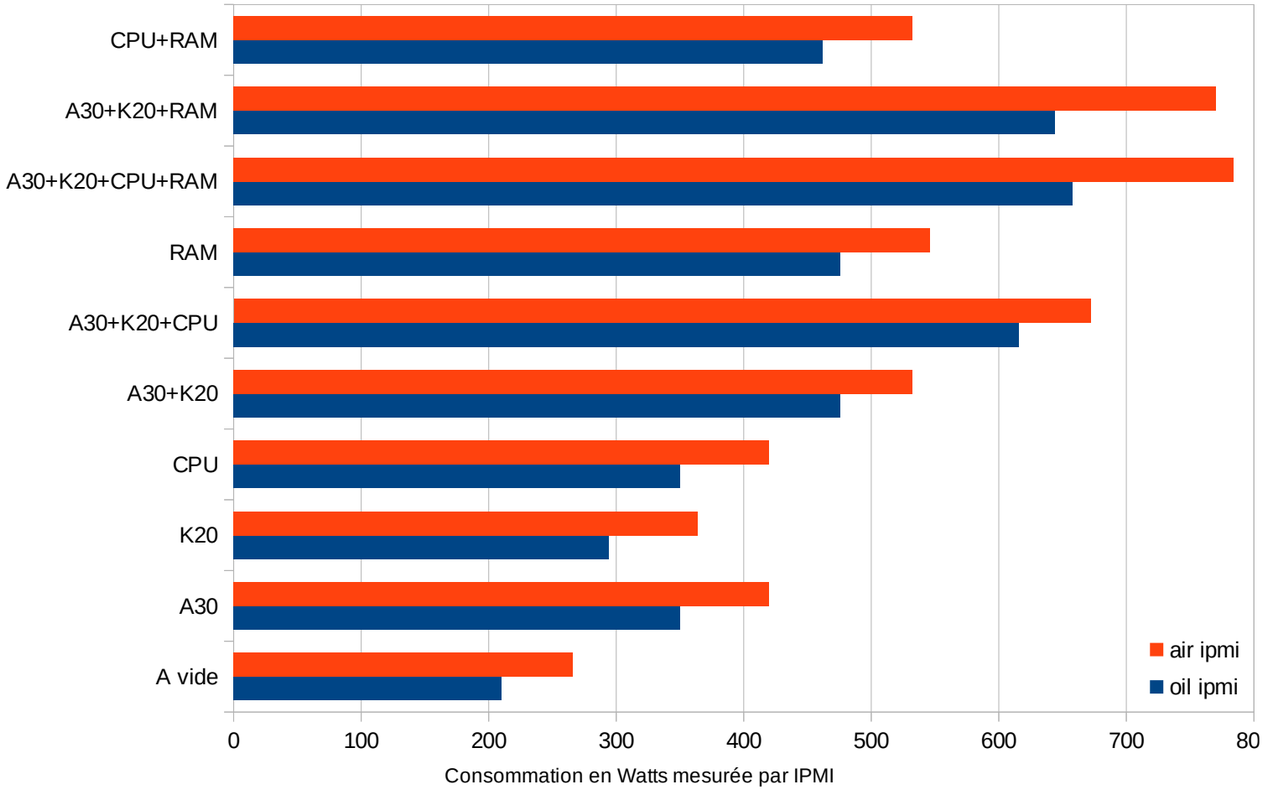
Etudes sur GPU, CPU & RAM
Premiers résultats intéressants (et intriguants)

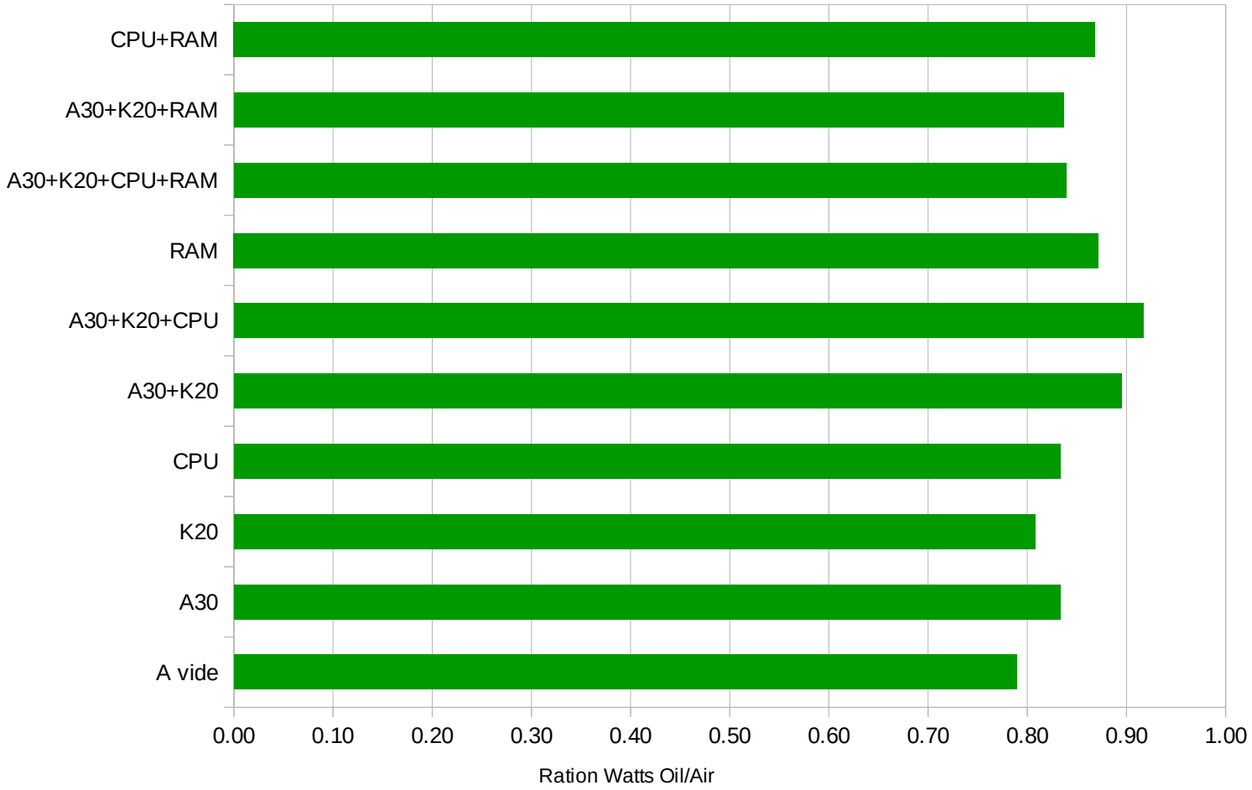
- Banc d'essai : 2 Dell R720 équipés de...
  - CPU : 2 Intel E5-2670 8 coeurs, 95 W de TDP
  - RAM : 384 GB de RAM (24x16 GB)
  - Carte H310 (pour l'étude du stockage ;-))
  - 2 GPU : un Nvidia K20m (Equip@Meso, 225 W) et un Nvidia A30 (prêt Nvidia, 165 W)



- Expérimentations : Nbody & PiDartDash pour CPU/GPU, mbw pour RAM
  - A vide… (réglage des BIOS en « performance » pour les CPU, boost possible, mode HT)
  - Sans la RAM : A30, K20m, CPU, A30+K20m, A30+K20m+CPU
  - Avec la RAM : mbw, mbw+A30+K20m, mbw+CPU, mbw+A30+K20m+CPU
- Métriques : performances, consommation (relève via IPMI)

Dans le bac Submer, pour les R720, ça donne ? Ca consomme « moins » mais côté RAM...





- A vide, 21 % de moins.
- En charge, entre 8 % et 19 % de moins. Influence significative de la RAM !

Et la suite ?

- Poursuivre les investigations avec le LIP dans un contexte de recherche appliquée :
  - T. Arabal, L. Betencour, E. Caron, and L. Lefevre. Setting up an experimental framework for analysing an immersion cooling system. In IEEE 34nd International Symposium on Computer Architecture and High Performance Computing. SBAC-PAD 2022., Bordeaux, October 5-8 2022. To appear.
- Evaluer plus précisément les consommations électriques :
  - Les pinces ampèremétriques disponibles apportent leur lot d'interrogations également...
  - Les wattmètres « simples » à déployer difficiles à interfacer (en masse)
- Evaluer des configurations avec des processeurs à beaucoup de coeurs
  - Seulement des systèmes à seulement 2x16 coeurs
- Evaluer plus de GPU
  - Des GPGPU : une montée en fréquence est-elle possible ?
  - Des GPU : une meilleure intégration (plus compacte) des séries 3000 et supérieures, ou AMD
- Appel à collaboration mais dans une approche scientifique (2 machines)