Immersion Cooling

De la préparation de machines génériques pour l' *Immersing Cooling*

… à une première analyse du comportement d'un GPU à l'immersion

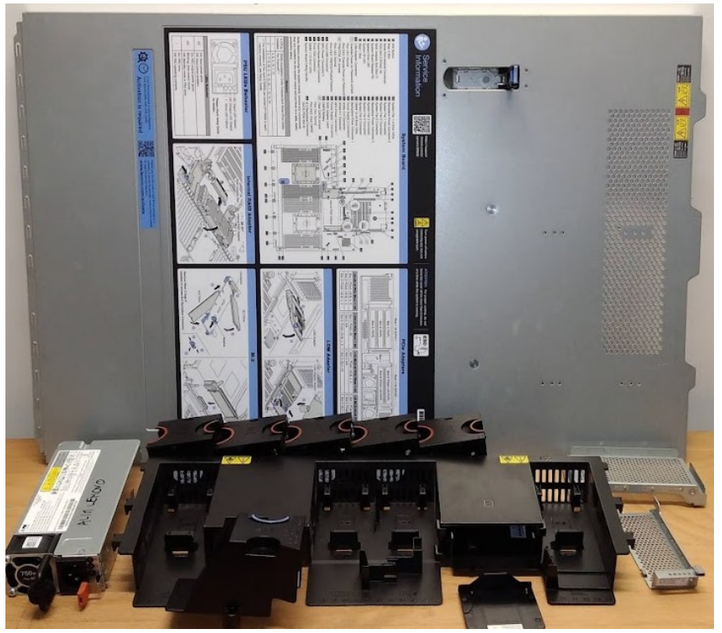
Emmanuel Quémener

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)

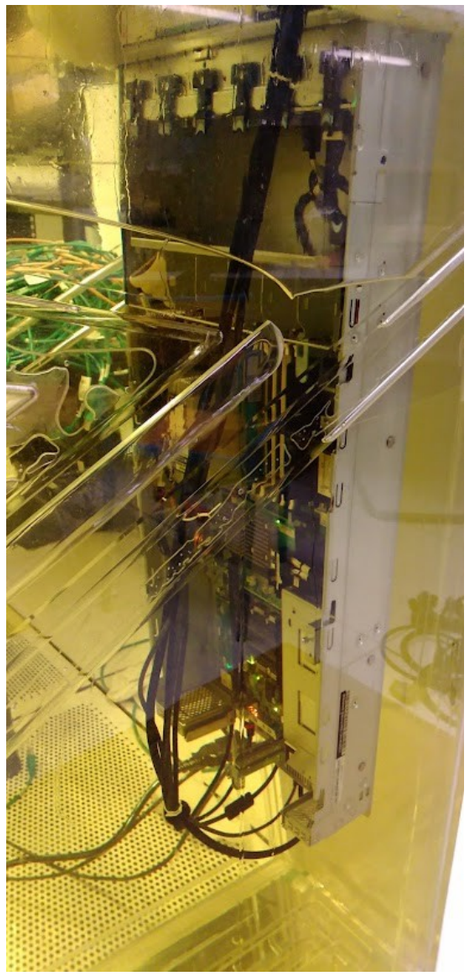
Préparation pour l'immersion : supprimer « tout ce qui bouge »

- Dans l'air : évacuer les « calories » :
  - On multiple la surface de contact : x200 pour un radiateur
  - On diffuse la chaleur par conduction ou évaporation (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



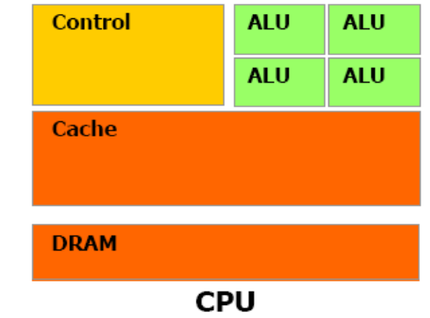
L’immersion sans modification (logicielle)
Sans ventilateur, ça démarre & ça tourne...

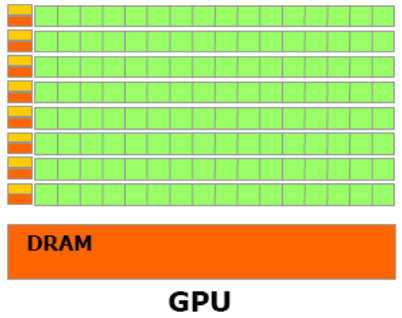
- Au repos : oil 74 W, air 68 W
  - Températures : oil 24°C, air 25°C
- En charge : oil 142 W, air 142 W
  - Températures : oil 36°C, air 38.5 °C
  - Mais en IPMI et AC : 140 W pour les deux
  - Mais en IPMI et DC : oil 115 W, air 120 W
- Côté performance :
  - 1 % de différence…
- Mais : processeur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



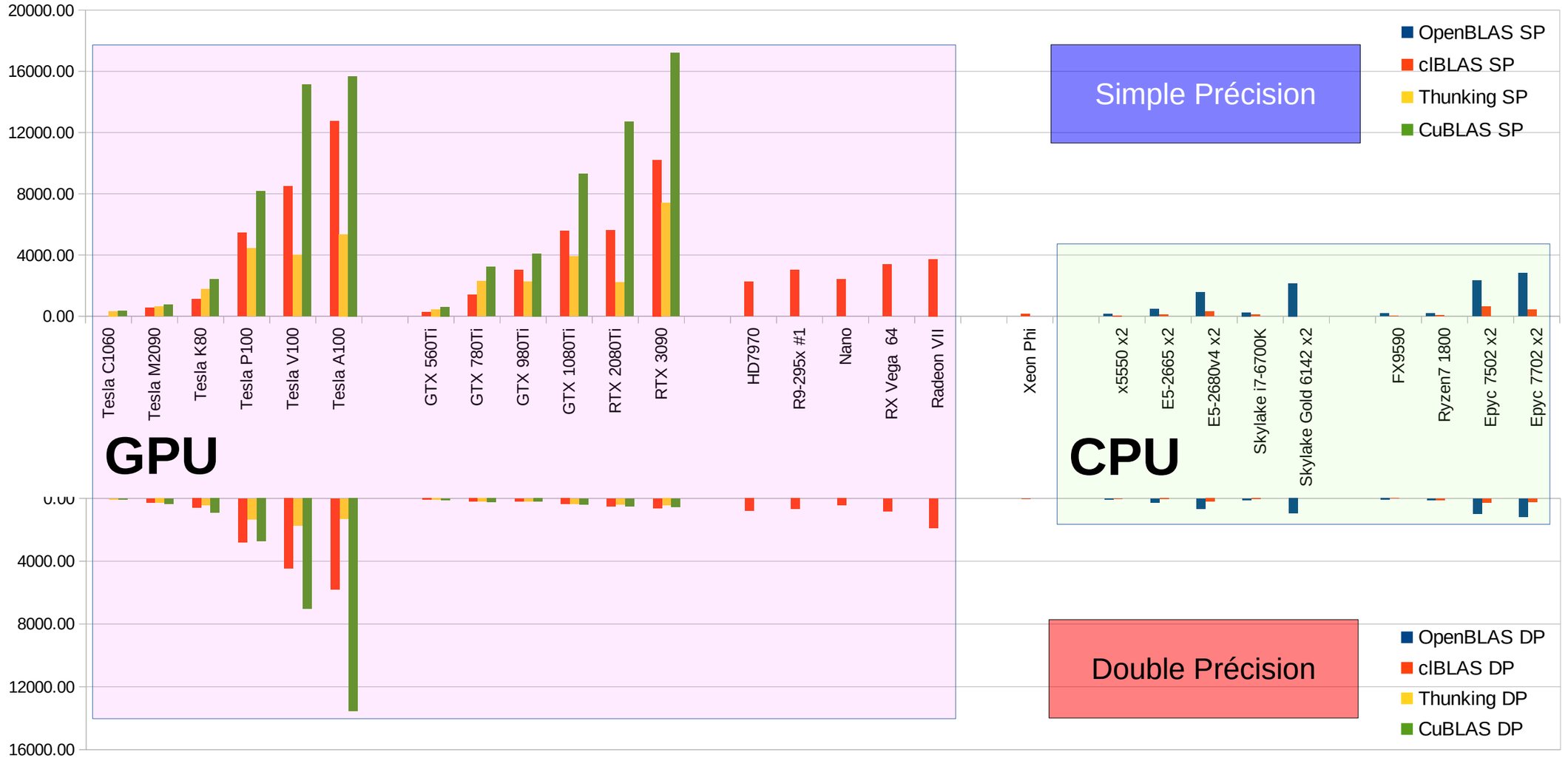
Pourquoi le GPU est-il si puissant ? Parce que il dispose :

- 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)

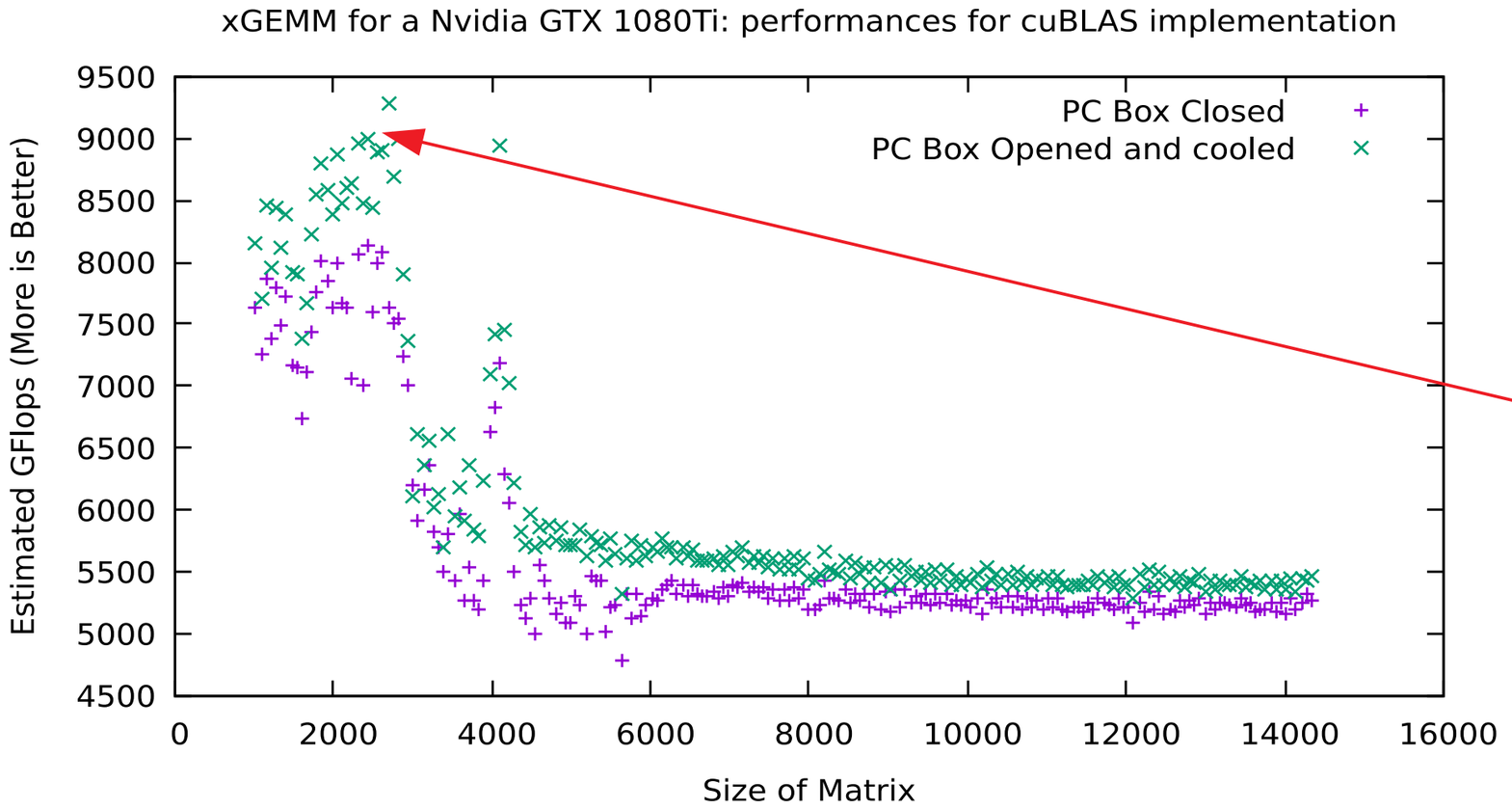




La puissance comparée entre GPU & CPU illustrée sur un cas simple (mais orienté...)

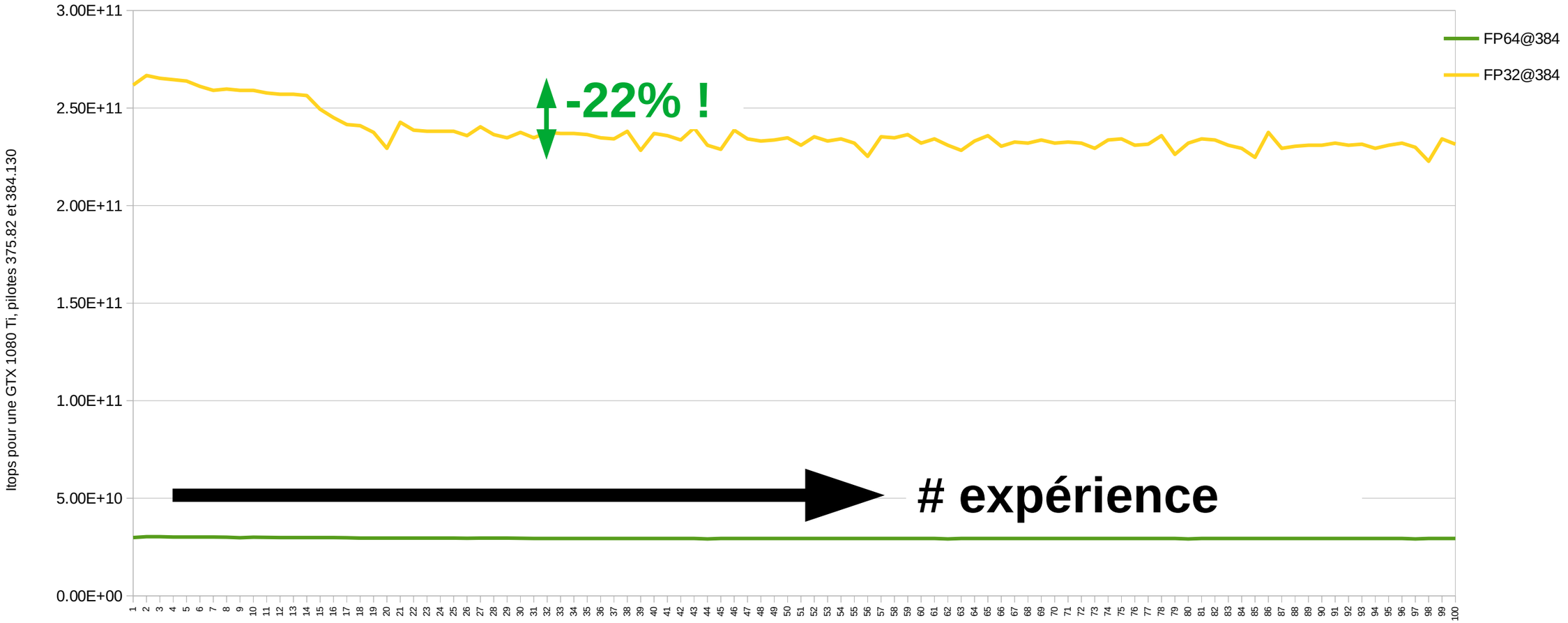


Un mariage impossible : Performance & Température



- 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 »



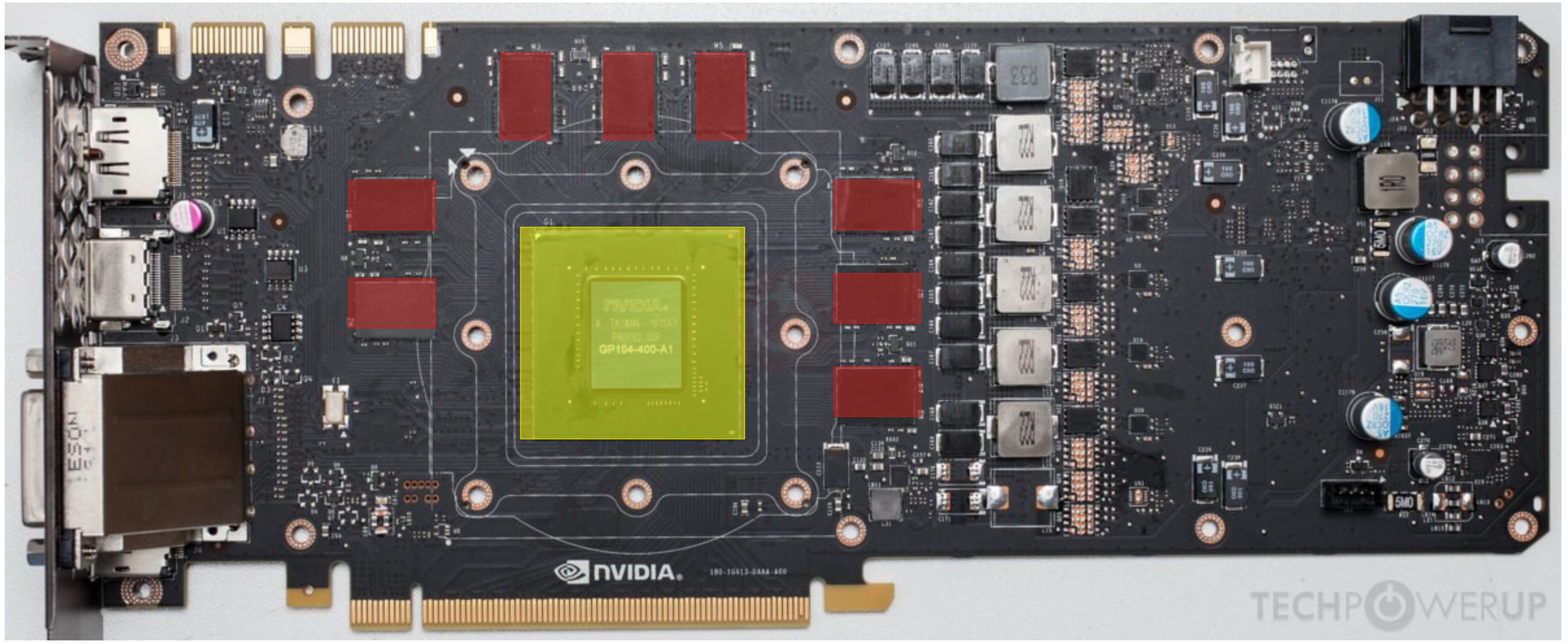
A bien prendre en compte dans l'expérience !

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

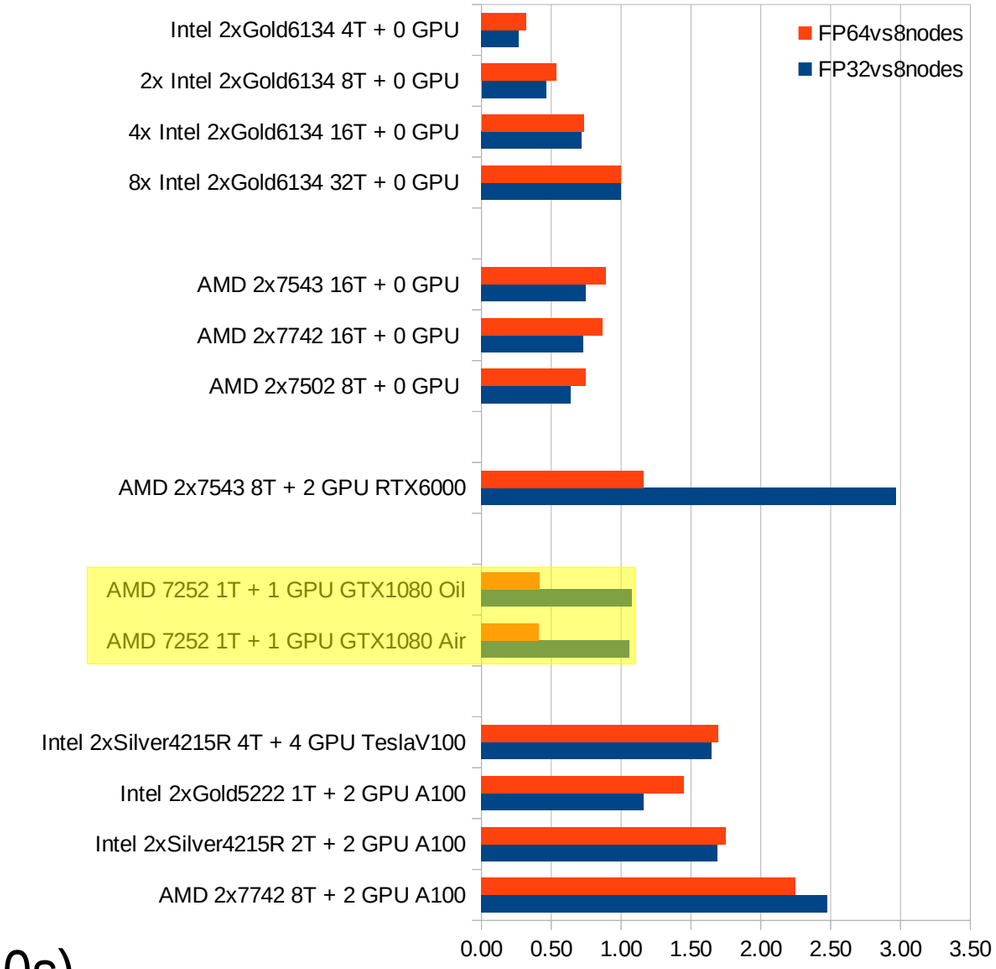
Un refroidissement nécessaire sur GPU **&** circuits DRAM

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 : e
  - Performances comparables en FP32 : (32104s/32050s)
  - Performances meilleures de 2 % en FP64 : (109066s/110860s)



Et la suite ?

- Evaluer plus précisément la consommation électrique :
  - Les pinces ampèremétriques disponibles étaient inopérantes…
  - Investigations sur dispositifs « simples » à déployer et à interfacer
- 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)