Programmation GPU

Spécificités

GPU – Un peu d’histoire

Traditionnellement, les GPU ont été utilisés pour accélérer les calculs gourmands en mémoire pour l'infographie comme le rendu d'image (jeux vidéos) et le décodage vidéo. **Ces problèmes sont sujets à la parallélisation.**
En raison de nombreux cœurs et d'une bande passante mémoire supérieure, un GPU semblait être un élément indispensable du rendu graphique.

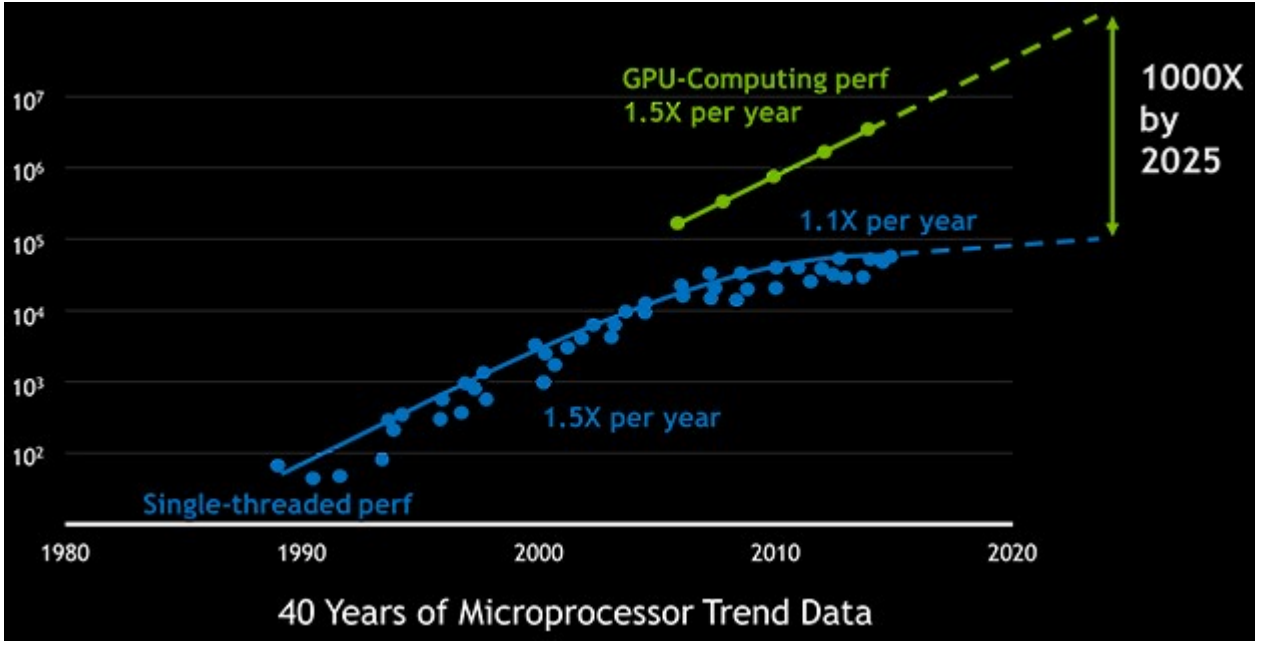
Le "premier GPU au monde" est arrivé en **1999** ! C'est ainsi que Nvidia a fait la promotion de sa **GeForce 256**. Nvidia a défini le terme unité de traitement graphique comme "un processeur monopuce avec transformation intégrée, éclairage, gestion des triangles.

GPU – Utilisation pour le calcul

L’ère des GPU programmable réellement et utilisable pour du calcul à commencer en **2007**.

Cependant, les deux sociétés principales (**Nvidia** et **AMD**) ont pris des voies différentes vers le GPU à usage général (GPGPU). En 2007, Nvidia a publié son environnement de développement CUDA, le premier modèle de programmation largement adopté pour le calcul GPU. Deux ans plus tard, OpenCL est devenu largement pris en charge. Ce cadre permet le développement de code pour les GPU et les CPU en mettant l'accent sur la portabilité. Ainsi, les GPU sont devenus un appareil informatique plus généralisé.

GPU vs CPU



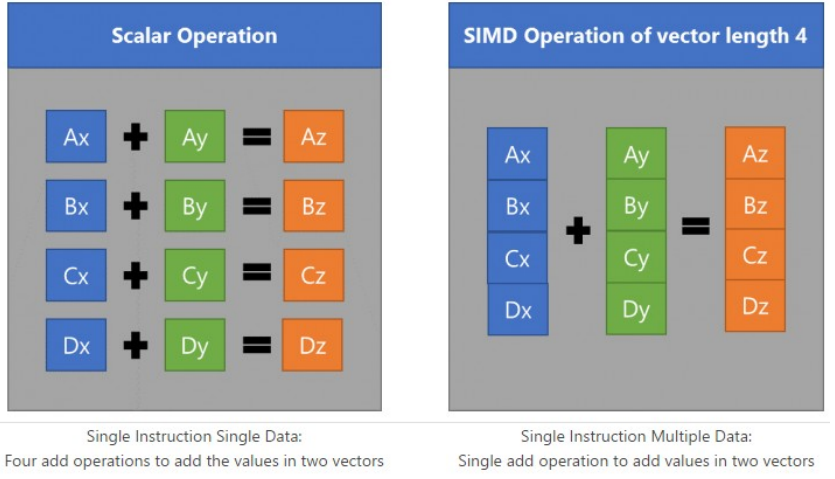
Présentation de la journée des investisseurs Nvidia 2017. La loi de Huang étend la loi de Moore - les performances des GPU vont plus que doubler tous les deux ans.

GPU vs CPU

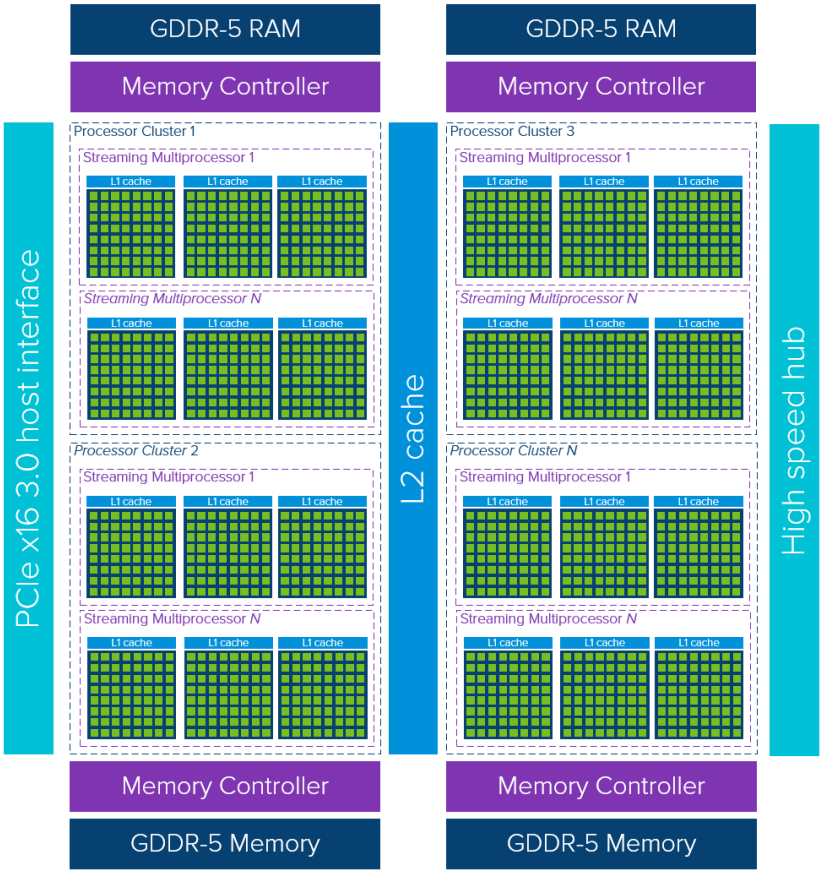
Alors qu'un CPU est orienté vers la latence et peut gérer des tâches linéaires complexes à grande vitesse, un GPU est orienté vers le débit, ce qui permet une énorme parallélisation.

Sur le plan architectural, un processeur est composé de quelques cœurs avec beaucoup de mémoire cache qui peuvent gérer quelques threads logiciels en même temps en utilisant un traitement série séquentiel.

Cependant, il est possible d'avoir un certain niveau de parallélisme par coeur en utilisant les instructions vectorielles (SIMD)



Architecture GPU



Concrètement, un GPU est composé de milliers de cœurs organisés en clusters. Chaque cœur utilise un jeu d'instruction très réduit part rapport à un CPU standard. De même les performances de chaque cœurs sont bien en deçà de ceux d'un CPU.

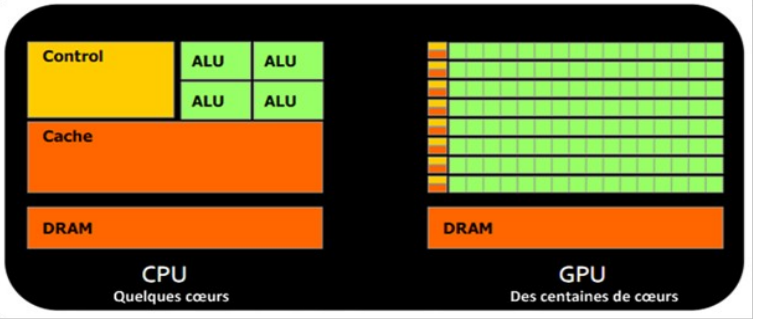
Nvidia A100

Nombre de coeurs	6912
Fréquence max	1410 Mhz
VRAM	40Go HBM2 1,5To/s
TDP	400 W
Perfs FP32	19,5 TFLOPS
Perfs FP64	9,7 TFLOPS

Architecture GPU

Aujourd'hui, les CPU exécutent les instructions en OoO (Out-of-order). Chaque cœur possède des instructions SIMD et sont optimisés pour l'accès à la mémoire de manière aléatoire
(avec certaines limites).

Les Coeurs GPU exécutent les instructions de manière séquentiel. Chaque cœur possède utilise les instruction SIMD. Un GPU est optimisé pour les accès séquentiel à la mémoire.



Architecture GPU

Alors que les CPU utilisent généralement de la mémoire DDR (DDR5 aujourd'hui pour les plus récents. Les GPU utilisent de la mémoire GDDR ou HMB qui disposent de bus de données plus larges et offrent une bande passante très supérieur.

Bande passante mémoire

DDR5	GDDR5	HBM2
~80 GB/s	>480 GB/s	>1,5 TB/s

GPGPU – Domaines d’application

- Météorologie
- Simulation financières
- Simulation par éléments finis
- Système de particule
- Traitement d’images
- FFT
- Algèbre linéaire
- Algorithmes de trie, de recherche
- Cryptographie
- … etc