

Programmation CUDA.

S. Puechmorel

2023

Historique.

Programmation sur GPU.

L'évolution des cartes graphiques.

Années 1980 : contrôleurs vidéo.

Ces circuits permettaient d'afficher sur un tube cathodique des informations stockées en mémoire. Ils fournissaient des fonctionnalités de base, essentiellement orientées autour de la gestion de la mémoire vidéo et de la génération des signaux de synchronisation.



Figure: Le contrôleur vidéo MC6845. ¹

¹ <https://commons.wikimedia.org/w/index.php?curid=976920>

Contrôleurs graphiques

Apparus vers la fin des années 1980, ils apportent des fonctionnalités graphiques, telles le tracé de segment, et gèrent une mémoire distincte de celle de l'unité centrale.



Figure: Contrôleurs graphiques.

L'évolution des cartes graphiques.

Les processeurs graphiques 3D.

Disponibles pour le grand public depuis le milieu des années 1990, ils incluent des fonctionnalités d'affichage en trois dimensions. Parallèlement, des bibliothèques logicielles font leur apparition (OpenGL, Direct3D). L'affichage d'une scène est réalisé à travers un pipeline graphique: transformation de sommets, projection, traçage.



Figure: Contrôleur 3D ATI Rage.

Les processeurs programmables.

En 2001, la société NVIDIA introduit sur le marché la gamme de processeurs graphiques GeForce 3 qui permettent de programmer les étapes du pipeline graphique.



Figure: Contrôleur 3D programmable.

L'évolution des cartes graphiques.

Le calcul.

Fin 2006, NVIDIA lance la gamme GeForce 8 et l'environnement de développement CUDA qui permet d'exploiter la puissance de calcul des cartes graphiques pour des applications générales. Les performances théoriques sont impressionnantes, de l'ordre de celles obtenues avec un superordinateur.



Figure: Carte graphique CUDA.

Les multiprocesseurs de flux (SM).

- ▶ Héritier du pipeline graphique, le multiprocesseur de flux ("Streaming multiprocessor", SM) est une entité de traitement comportant des séquenceurs, plusieurs unités de traitement numérique et une mémoire locale.
- ▶ Un processeur graphique (GPU) regroupe plusieurs multiprocesseurs.
- ▶ Les multiprocesseurs exécutent des blocs de processus de façon **indépendante** et peuvent accéder à une mémoire partagée.
- ▶ Pour un développeur sur une architecture conventionnelle, un multiprocesseur s'apparente à un cœur de calcul vectoriel.

Les multiprocesseurs de flux (SM).

- ▶ À l'intérieur d'un multiprocesseur, les processus s'exécutent de façon concurrente, mais peuvent communiquer via la mémoire locale ou être synchronisés.
- ▶ Les processus sont regroupés par blocs, appelés "warps" (chaînes), qui se voient affecter le même séquenceur d'instructions.
- ▶ Le modèle associé est dit "SIMT" pour "Single Instruction Multiple Thread".

Les unités de calcul.

- ▶ Le nombre d'unités de calcul par multiprocesseur dépend des générations de cartes. Pour l'architecture Ampère, on trouve 64 ou 128 unités flottantes 32bits, 32 ou 2 unités flottantes 64bits, 64 unités de calcul entier sur 32 bits, 16 unités spéciales (fonctions transcendantes), 4 cœurs de calcul tensoriel et 4 ordonnanceurs.
- ▶ Un ordonnanceur est affecté à une chaîne. Tous les processus de la chaîne exécutent la même Instruction au même moment.
- ▶ En cas d'instruction conditionnelle, il peut y avoir divergence de code à l'intérieur d'une chaîne, ce qui se traduit par la mise en attente d'un ou plusieurs processus dont l'exécution se poursuivra après celle de la branche principale.

La mémoire.

- ▶ Chaque multiprocesseur possède une mémoire locale très rapide, pouvant être partagée entre les processus d'un même bloc. Elle est organisée en 32 banques pouvant être utilisées simultanément. Idéalement, chaque processus d'une chaîne accède à sa propre banque. La capacité de cette mémoire varie entre 64kB et 228kB selon les générations.
- ▶ Les multiprocesseurs partagent une mémoire globale, plus lente, mais en mesure de stocker beaucoup plus de données. Les cartes de dernière génération, comme la RX4090, embarquent 24Gb de RAM.
- ▶ L'architecture 9.0 introduit la notion de cluster de blocs et de mémoire partagée distribuée.