RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ENAC

Programmation CUDA.

S. Puechmorel

2023

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

ENAC

Plan

Historique.

Programmation sur GPU.

L'évolution des cartes graphiques.

Années 1980 : contrôleurs video.

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

L'évolution des cartes graphiques.

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 fonctionalités 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.

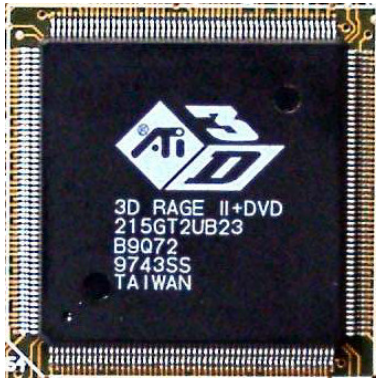

Figure: Contrôleur 3D ATI Rage.

L'évolution des cartes graphiques.

Les processeurs progammables.

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.

Architecture matérielle.

Les multiprocesseurs de flux (SM).

- Héritier du pipeline graphique, le multiprocesseur de flux ("Streaming multiprocessor", SM) est une entité de traitement comportant un séquenceur, plusieurs unités de traitement numérique et une mémoire locale.

Architecture matérielle.

Les multiprocesseurs de flux (SM).

- Héritier du pipeline graphique, le multiprocesseur de flux ("Streaming multiprocessor", SM) est une entité de traitement comportant un séquenceur, plusieurs unités de traitement numérique et une mémoire locale.
- Un processeur graphique (GPU) regroupe plusieurs multiprocesseurs.

Architecture matérielle.

Les multiprocesseurs de flux (SM).

- Héritier du pipeline graphique, le multiprocesseur de flux ("Streaming multiprocessor", SM) est une entité de traitement comportant un séquenceur, 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.

Architecture matérielle.

Les multiprocesseurs de flux (SM).

- Héritier du pipeline graphique, le multiprocesseur de flux ("Streaming multiprocessor", SM) est une entité de traitement comportant un séquenceur, 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.

Architecture matérielle.

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.

Architecture matérielle.

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", qui se voient affecter le même séquenceur d'instructions.

Architecture matérielle.

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", qui se voient affecter le même séquenceur d'instructions.
- Le modèle associé est dit "SIMT" pour "Single Instruction Multiple Thread".

Architecture matérielle.

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", qui se voient affecter le même séquenceur d'instructions.
- Le modèle associé est dit "SIMT" pour "Single Instruction Multiple Thread".
- Les dernières générations de processeurs graphiques tendent à supprimer ces limitations.