Sources :  
<https://www.intel.ca/content/www/ca/en/products/docs/processors/what-is-a-gpu.html>  
<https://computer.howstuffworks.com/graphics-card.htm>  
<https://www.tomshardware.com/reviews/gpu-graphics-card-definition,5742.html>  
<https://blogs.nvidia.com/blog/2009/12/16/whats-the-difference-between-a-cpu-and-a-gpu/>

<https://www.rastergrid.com/blog/gpu-tech/2021/07/gpu-architecture-types-explained/>

<https://www.cherryservers.com/blog/everything-you-need-to-know-about-gpu-architecture>

<https://www.techtarget.com/whatis/definition/NVIDIA#:~:text=Nvidia%20Corporation%20is%20a%20technology,headquartered%20in%20Santa%20Clara%2C%20Calif>.

<https://www.cherryservers.com/blog/everything-you-need-to-know-about-gpu-architecture>

<https://www.nvidia.com/en-us/geforce/ada-lovelace-architecture/>

<https://www.tomshardware.com/features/nvidia-ada-lovelace-and-geforce-rtx-40-series-everything-we-know>

<https://www.digitaltrends.com/computing/what-is-a-teraflop/>

<https://www.amd.com/en/technologies/zen-core>

<http://www.ordinateur.cc/Mat%C3%A9riel/Ordinateurs-de-bureau/28225.html>

<https://www.topflightpc.com/blog/sli-and-crossfire>

<https://hothardware.com/reviews/amd-radeon-rdna-3-architecture-overview>

<https://www.amd.com/fr/technologies/rdna#RDNA%E2%84%A2-3>

<https://www.tomshardware.com/news/amd-rdna-3-gpu-architecture-deep-dive-the-ryzen-moment-for-gpus?utm_campaign=socialflow&utm_medium=social&utm_source=twitter.com>

**Introduction**

Ce travail a pour objectif d’identifier les caractéristiques des différents principaux fabricants de processeur graphique. Plus précisément, les particularités au niveau logiciel et architecturale de chaque manufacturier de puce graphique. Cependant, quelques connaissances de base seront nécessaires à la compréhension de ce document. En premier lieu il est essentiel de comprendre ce qu’est un GPU, aussi connu sous le nom de Graphics processing unit, ou en français, processeur graphique.   
  
**Qu’est-ce qu’un processeur graphique?**  
  
Simplement, un GPU a pour objectif d’accélérer la vitesse de traitement d’image. Il a pour tâche d’afficher les images qu’il produit sur un moniteur. Ce qui rend cette pièce d’équipement spécial, c’est sa capacité à traiter plusieurs données à la fois, on appelle ce processus traitement parallèle. Il est très utile dans plusieurs domaines, que ce soit pour produire du contenu créatif à l’aide de logiciel de montage, pour développer de l’intelligence artificielle, ou bien encore simplement pour rendre les jeux vidéo plus fluide. Le processeur graphique n’est qu’une petite pièce d’un plus grand tout, soit la carte graphique ou parfois même le CPU, il est alors connu sous le nom de processeur graphique intégré.   
  
Lorsqu’il fait partie d’une carte graphique, cette dernière a pour but de le relier au reste du système à travers un port PCIe. Elle a aussi pour rôle de réguler sa consommation en watt et de lui offrir un voltage adapté à son bon fonctionnement. Elle lui offre aussi accès à de la mémoire vive (VRAM) directement intégrée sur la carte, cela a pour effet d’en augmenter la vitesse, puisqu’elle se trouve à être physiquement plus proche de la pièce qui l’utilise, soit le GPU.   
  
La différence entre un CPU et un GPU est souvent difficile à comprendre. La différence principale est le nombre de cœur que chacun d’entre eux possèdent. D’un côté, le CPU n’en utilise que quelques un pour effectuer des séries de processus séquentiel, alors que le GPU en possède des milliers qu’il utilise pour réaliser plusieurs tâches à la fois. Il utilise ses cœurs pour décomposer des problèmes complexes en milliers de tâches individuels qu’il réalise toutes en même temps.   
  
En résumé, et en simplifiant beaucoup, le GPU décide ce que chaque pixel de l’écran affichera et s’occupe d’accomplir des milliers d’instructions en un même instant. Il faut maintenant établir la différence entre chaque producteur de processeur graphique, selon leur architecture respective.   
  
**Mais qu’est-ce qu’une architecture?**  
  
Il en existe plusieurs types, mais pour les besoins de la cause, nous expliqueront rapidement ce que fait la version la plus simple et la plus connu, soit le « immediate-mode rendering » ou IMR. Toutes les architectures ont plus ou moins la même disposition ce qui en fait un bon exemple. En fait, c’est un peu un long tunnel possédant plusieurs arrêts où certaines tâches devront être accomplis. Chaque commandes complexes sera subdivisé en courtes instructions. Cela, avant que toutes celles-ci ne soient traités en un seul voyage à travers l’entièreté de cette « pipeline ». Ces arrêts sont placés à l’intérieur de deux groupes majeurs. Celui qui s’occupe de traiter les commandes en langage machine, de les traduire, de les diviser en instructions simples et d’aller chercher les ressources qui seront nécessaires, tel que les textures, plus tard dans le traitement. Le deuxième groupe commence avec la rastérisation, qui renvoi plusieurs fragments de l’image à une unité qui les traite et les combine. Finalement les valeurs de couleur de chaque fragment sont envoyées à l’écran pour que ceux-ci soient affichées par les pixels.   
  
**Ada Lovelace**  
  
Nvidia est la compagnie productrice de GPU discret la plus connu, ainsi que la plus répandu à travers les machines de consommateurs autour du globe. Elle fut fondée en 1993. Depuis cette époque, le géant corporatif ne cesse d’innover. Sa création la plus connu est probablement le « CUDA core ». C’est ainsi que la corporation fait appeler les cœurs de ses GPU. En parlant très généralement, plus un GPU à de cœur, plus il sera rapide. Le principale avantage des cœurs CUDA est la simplicité avec laquelle il est possible de les programmer. Ils possèdent leur propre langage de programmation, connu sous le même nom. Ce langage est très similaire au C++, il permet au programmeur de donner des instructions directement au GPU, et cela, au lieu de passer par une bibliothèque graphique 3D. Cela permet d’écrire des programmes complexes à exécution parallèle qui fonctionneront beaucoup plus rapidement que si écrit autrement.  
   
L’architecture la plus récente produite par la compagnie est ADA Lovelace. Elle utilise une puce monolithique produite sur le dernier processus 4 nanomètres de TSMC. Il est important de ce souvenir, que plus les composantes électroniques sont proches les unes des autres, plus ils pourront communiquer rapidement. Cela s’applique aussi aux différents transistors d’un GPU. La version la plus rapide provenant de ce processus est AD102 avec un total de 76.3 milliards de transistors, ainsi que 16384 cœurs CUDA.   
  
Cette quantité élevée de transistors et de cœurs à pour résultat un maximum de 82.6 téraflops. Pour ceux qui ne serait pas ce qu’est un téraflop. C’est la capacité qu’a un processeur à calculer mille milliards d’opération en virgule flottante par secondes. Cela offre aux nouveaux GPU provenant de Nvidia la capacité de pousser des taux de rafraichissement élevé, et cela même à une résolution aussi grande que 4K. On voit maintenant apparaitre des CPU « bottleneck » même à cette résolution, ce qui était impensable auparavant.

Cette nouvelle architecture offre pour la première fois des performances acceptable en tracement de rayon ou « ray tracing ». Cela, grâce à la troisième génération de cœur RT introduite avec cette génération de carte graphique. Le tracement de rayon est une technologie utilisée depuis très longtemps en animation 3D. La différence est ici que les rayons de lumières sont tracés en temps réel. Selon Nvidia, ces nouveaux cœurs exécutent leurs calcules jusqu’à deux fois plus vite, que ceux de la génération précédente.

Avoir autant de rayons devant être tracés aussi vite à pour conséquence de ralentir la vitesse à laquelle les « shaders » peuvent calculer les niveaux appropriés de lumière, d’ombre et de couleur qui devront être affichés à l’écran. C’est ici que rentre en jeu une nouvelle technologie introduite avec cette architecture, soit le réarrangement de l’ordre d’exécution des « shaders ». Changer l’ordre d’exécution en fonction de la scène et de ce qui doit être calculé en premier, à pour conséquence d’accélérer la production d’image complète par jusqu’à 25 pourcents.

La dernière technologie intéressante introduite par ces nouvelles cartes graphiques est DLSS 3. Celle-ci fait en soit la même chose que son prédécesseur. Donc prendre une image de basse résolution et, grâce à des algorithmes, augmenter cette dernière tout en reconstruisant l’image. Ce qui offre parfois des images de meilleures qualités que l’original. La nouveauté est ici la capacité qu’on les cœurs Tensor a aussi généré de nouvelles images grâce à de l’intelligence artificielle. L’IA va analyser l’image grâce à des vecteurs de mouvement, pour ensuite déterminer et prédire à quoi ressemblera la prochaine image qui devra être affiché à l’écran. Ces nouvelles images générées par l’IA sont affichées en alternance avec d’autres produite plus traditionnellement. Cette nouvelle version de DLSS a pour effet d’augmenter grandement le taux de rafraichissement.

**RDNA 3**

AMD est le principal compétiteur d’Nvidia dans le domaine des cartes graphiques. « Adavanced Micro Devices » est une compagnie fondée aux États-Unis en 1969. Elle est surtout reconnue comme un fabriquant de CPU, surtout depuis la sortie de son architecture Ryzen. Celle-ci se différencie de la compétition grâce à son approche de « chiplet ». Elle consiste à produire plusieurs puces différentes pour plus tard les connectés sur un même circuit imprimé. Cela permet de réduire les coûts de production, car le taux d’échec est beaucoup plus bas. Mais elle permet aussi une grande variété de configuration possible, et cela, plus facilement qu’avec un processus monolithique traditionnel.   
  
Avant de parler des nouveaux processeurs graphiques de la compagnie. Il nous faut au préalable expliquer ce qu’est SLI ou aussi connu sous le nom de Crossfire. La seule différence est que SLI est le nom que donne Nvidia à la technologie et Crossfire celui qu’AMD lui donne. Cette technologie permet de relier deux GPUs à l’aide d’un lien externe. Laissant les deux cartes communiquer plus rapidement que s’il ne faisait que transférer leurs données par leur port PCIe respectif. Ce lien permet aux cartes de se séparer leurs respectives tâches, accélérant ainsi la vitesse des calculs devant être fait. Malheureusement, la technologie était considérée trop lente et problématique par les deux géants de « Silicon Valley ». Elle n’est à présent plus supporter par les deux compagnies.

Vient alors AMD avec le dévoilement récent de leur nouvelle architecture pour processeur graphique : RDNA 3. Ce qui rend la corporation le plus fiers est son efficacité watt pour watt. Promettant 54 pourcents plus de performance que RDNA 2 lorsque le même budget de courant est utilisé. Mais ce qui redéfinit l’histoire est l’introduction des premiers GPU utilisant des « chiplets ».  
  
Tel le faisait SLI ou Crossfire avant, on tente ici de rejoindre plusieurs GPU ensemble dans le but d’accélérer les calculs de l’ordinateur. La différence est que l’on a plus besoin de connecteur externe. Tout se fait sur le PCB. AMD a choisi d’essayer ce nouveau processus, car par le passé, tout amélioration se produisait généralement en réduisant la taille de la puce en elle-même. Malheureusement, le progrès que cela apportait a grandement ralentit dans les dernières années, raison pour laquelle il nous faut trouver de nouvelles solutions. L’utilisation de « chiplets » était plus simple car la vitesse requise pour relier les différentes puces sans latence est moindre. Pour un GPU, on a besoin d’exponentiellement plus de bande passante. La solution est de réduire la taille de ce qui peut être réduit et de conserver le reste sur une grande puce. Cela donne naissance au GCD fait sur un processus de 5 nanomètres, qui contient ce à quoi l’on s’attendrait d’un GPU traditionnel, mise à part une interface pour la mémoire et de la cache L3. Cela se situe à l’intérieur de 6 MCD fait sur un processus de 6 nanomètres. Les MCD sont reliés au GCD à l’aide de qu’AMD nomme « Infinity Fanouts ». Ce lien personnalisé permet d’obtenir des vitesses de transfert entre les différentes puces allant jusqu’à 5.3TB/sec. Cela se trouve à être douze fois plus vite que ce qui est nécessaire pour les « chiplets » présentent sur un CPU.