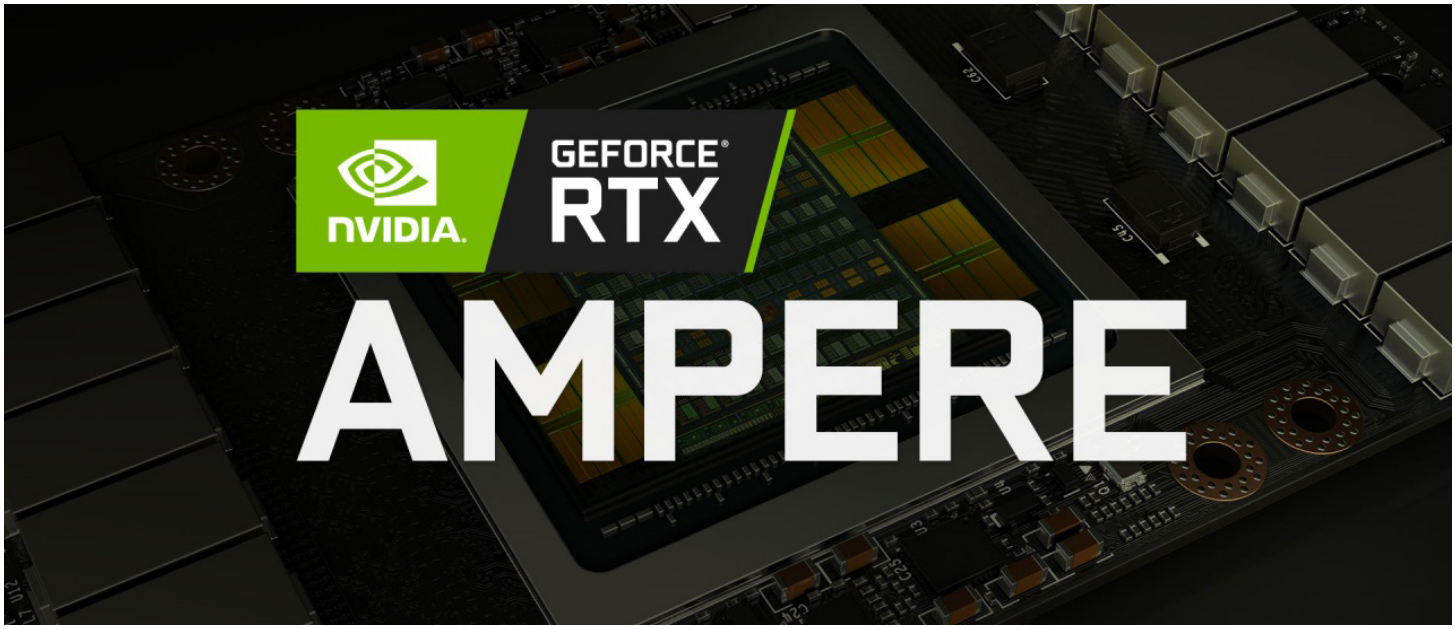
Ampere est la dernière architecture de jeu NVIDIA. Le plus important du livre blanc



Depuis l'invention de son premier GPU en 1999, NVIDIA est à l'avant-garde des graphiques 3D et de l'informatique accélérée par GPU. Chaque architecture NVIDIA est soigneusement conçue pour offrir des niveaux révolutionnaires de performances et d'efficacité.

L'A100, le premier GPU avec l'architecture NVIDIA Ampere, est sorti en mai 2020. Il fournit une accélération considérable pour la formation à l'IA, le HPC et l'analyse des données. L'A100 est basé sur la puce GA100, qui est purement informatique et, contrairement au GA102, n'est pas encore un jeu.

Les GPU GA10x sont basés sur l'architecture GPU NVIDIA Turing. Turing est la première architecture au monde à offrir un traçage de rayons en temps réel haute performance, des graphiques accélérés par l'IA et un rendu graphique professionnel, le tout dans un seul appareil.

Dans cet article, nous analyserons les principaux changements dans l'architecture des nouvelles cartes vidéo NVIDIA par rapport à leur prédécesseur.

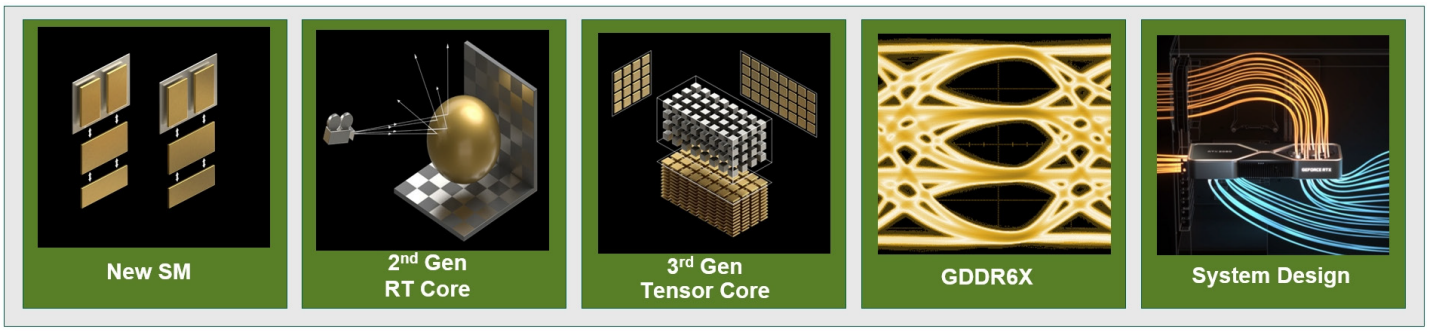


Figure 1. Architecture Ampere GA10x

Principales caractéristiques du GA102

Le GA102 est fabriqué à l'aide de la technologie 8 nm propriétaire de NVIDIA - 8N NVIDIA Custom. La puce contient 28,3 milliards de transistors sur une puce de 628,4 mm2. Comme tous les GeForce RTX, le GA102 est basé sur un processeur qui contient trois types de ressources informatiques différents:

- Noyaux CUDA pour l'ombrage programmable;
- RT-, (BVH) ;
- , .

Ampere

GPC, TPC SM

Comme ses prédécesseurs, le GA102 se compose de grappes de traitement graphique (GPC), de grappes de traitement de texture (TPC), d'unités de rastérisation multiprocesseurs en continu (SM), d'opérateurs raster (ROP) et de contrôleurs de mémoire. La puce complète comprend sept unités GPC, 42 TPC et 84 SM.

Le GPC est le bloc de haut niveau dominant qui contient tous les graphiques clés. Chaque GPC dispose d'un moteur Raster dédié et dispose désormais de deux sections ROP de huit blocs chacune, ce qui est une innovation dans l'architecture Ampere. En outre, le GPC contient six TPC, chacun contenant deux multiprocesseurs et un moteur PolyMorph.

Figure 2. Compléter le GPU GA102 avec 84 blocs SM

À son tour, chaque SM du GA10x contient 128 cœurs CUDA, quatre cœurs Tensor de la troisième génération, un fichier de registre de 256 Ko, quatre unités de texture, un cœur de traçage de rayon de deuxième génération et 128 Ko L1 / mémoire partagée, qui peuvent être configurés pour différentes capacités. en fonction des besoins des tâches informatiques ou graphiques.

Optimisation ROP

Dans les GPU NVIDIA précédents, les ROP étaient liés à un contrôleur de mémoire et à un cache L2. À partir de GA10x, ils font partie du GPC, qui améliore les performances des opérations raster en augmentant le nombre total de ROP.

Au total, avec sept GPC et 16 ROP dans chaque GPC, le GPU GA102 se compose de 112 ROP au lieu de 96, par exemple, dans le TU102. Tout cela a un effet positif sur l'anti-crénelage multi-échantillons, le taux de remplissage des pixels et le mélange.

NVLink troisième génération

Les GPU GA102 prennent en charge NVIDIA NVLink de troisième génération, qui comprend quatre voies x4, chacune fournissant 14,0625 Go / s de bande passante entre deux GPU dans les deux sens. Les quatre canaux réunis donnent 56,25 Go / s de bande passante dans chaque direction et un total de 112,5 Go / s entre les deux GPU. Ainsi, en utilisant NVLink, deux GPU RTX 3090 peuvent être connectés.

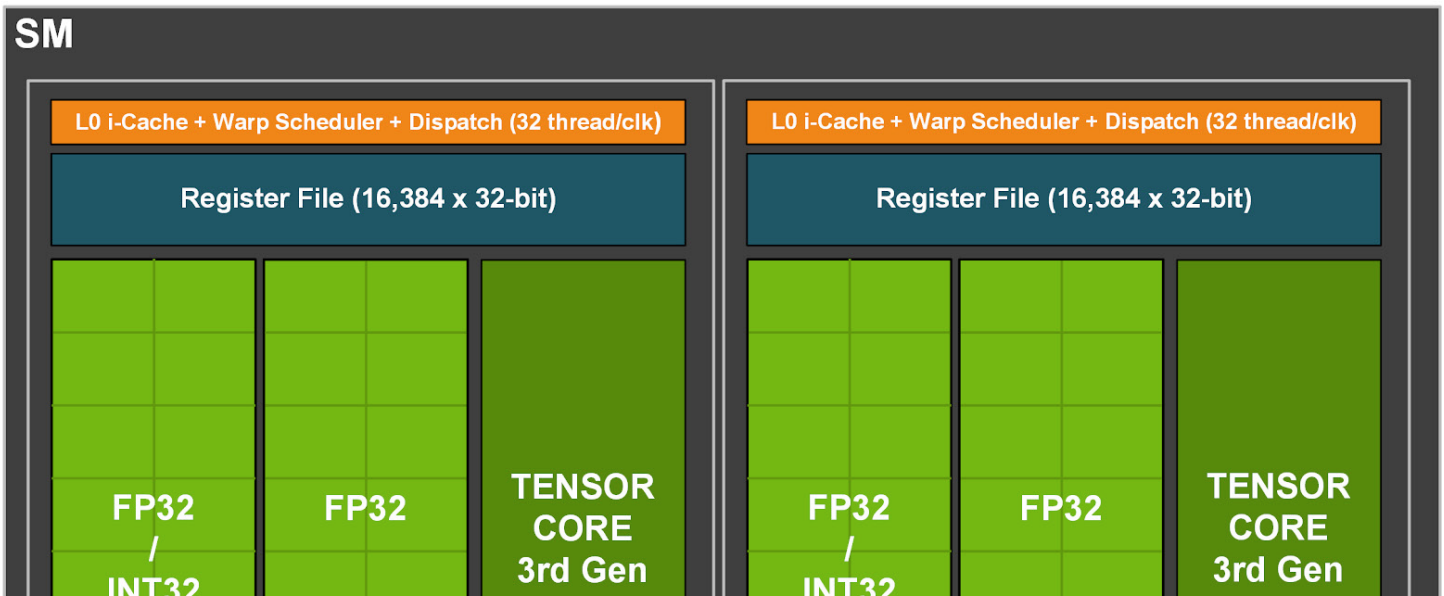
PCIe Gen 4

Les GPU GA10x sont équipés de PCI Express 4.0, qui offre deux fois la bande passante de PCIe 3.0, des taux de transfert jusqu'à 16GTransfers par seconde et, grâce au slot x16 PCIe 4.0, la bande passante maximale atteint 64 Go / s.

Architecture multiprocesseur GA10x

L'architecture multiprocesseur de Turing a été la première chez NVIDIA à avoir des cœurs séparés pour accélérer les opérations de lancer de rayons. Ensuite, Volta a introduit les premiers noyaux tensoriels, et Turing a introduit les noyaux tenseurs avancés de deuxième génération. Une autre innovation de Turing et Volta est la possibilité d'exécuter simultanément des opérations FP32 et INT32. Le multiprocesseur du GA10x prend en charge toutes les fonctionnalités ci-dessus et possède également un certain nombre de ses propres améliorations.

Contrairement au TU102, qui possède huit cœurs tenseurs de deuxième génération, le multiprocesseur GA10x dispose de quatre cœurs tenseurs de troisième génération, chaque noyau tensoriel GA10x deux fois plus puissant que Turing.



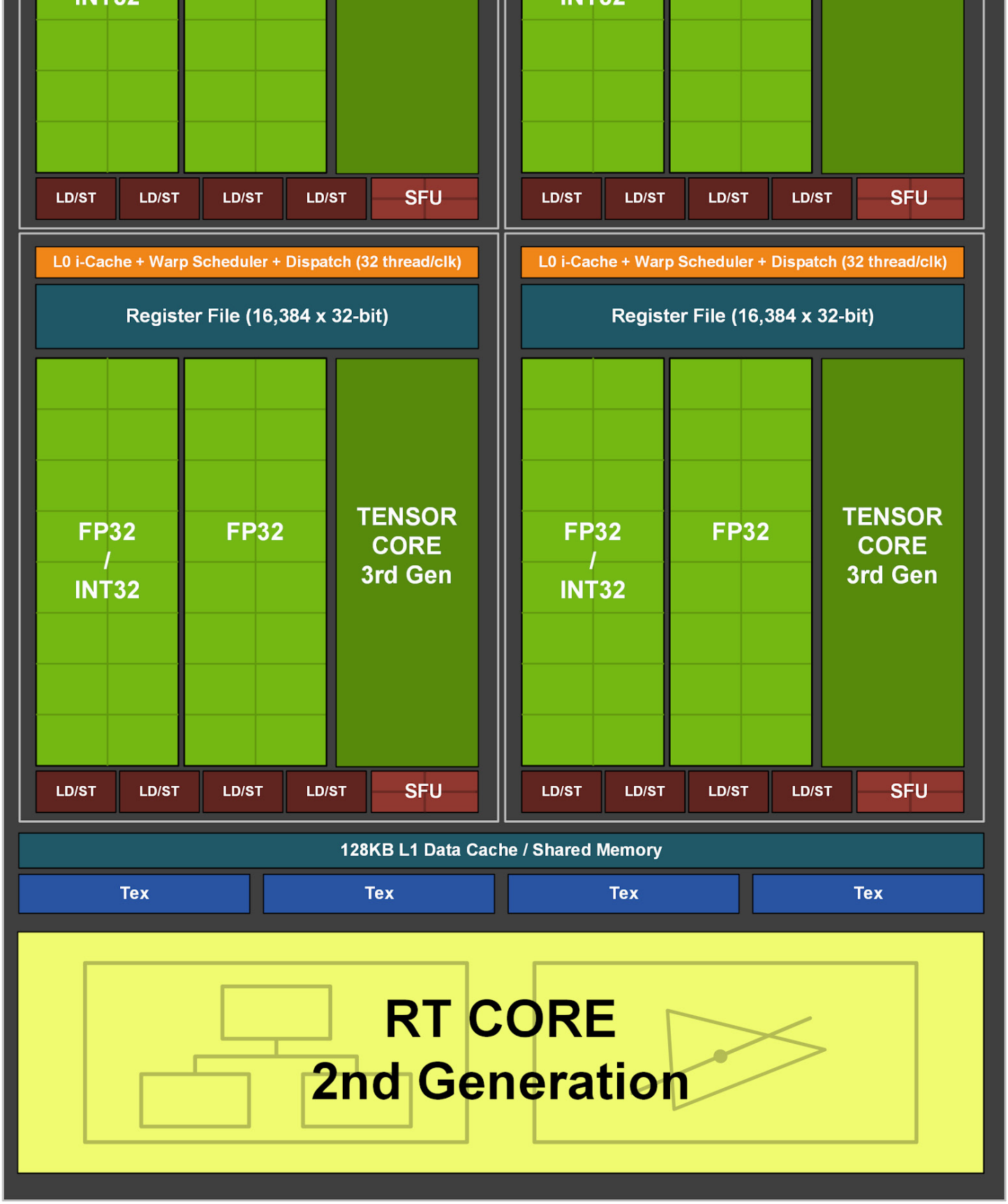


Figure 3. Multiprocesseur de streaming GA10x

Doublez la vitesse de calcul FP32

La plupart des calculs graphiques sont des opérations en virgule flottante 32 bits (FP32). Le multiprocesseur de streaming Ampere GA10x offre deux fois la vitesse des opérations FP32 sur les

deux canaux de données. En conséquence, dans le cadre de FP32, la GeForce RTX 3090 fournit plus de 35 téraflops, soit plus de 2 fois les capacités de Turing.

Le GA10X peut exécuter 128 opérations FP32 ou 64 opérations FP32 et 64 opérations INT32 par horloge, soit le double de la vitesse des calculs de Turing.

Les tâches de jeu modernes ont un large éventail de besoins de traitement. De nombreux calculs nécessitent un ensemble d'opérations FP32 (telles que FFMA, l'addition en virgule flottante (FADD) ou la multiplication en virgule flottante (FMUL)), ainsi que de nombreux calculs d'entiers plus simples.

Les multiprocesseurs GA10x continuent de prendre en charge les opérations FP16 (HFMA) à double vitesse, qui étaient également prises en charge dans Turing. Et, comme pour les GPU TU102, TU104 et TU106, dans le GA10x, les opérations FP16 standard sont également gérées par des cœurs de tenseurs.

Mémoire partagée et cache de données L1

GA10x possède une architecture unifiée pour la mémoire partagée, le cache de données L1 et le cache de texture. Cette conception unifiée peut être modifiée en fonction de la charge de travail et des besoins.

La puce GA102 contient 10752 Ko de cache L1 (contre 6912 Ko dans le TU102). En dehors de cela, le GA10x a également le double de la bande passante de la mémoire partagée par rapport à Turing (128 octets / cycle contre 64 octets / cycle). La bande passante totale L1 pour la GeForce RTX 3080 est de 219 Go / s contre 116 Go / s pour la GeForce RTX 2080 Super.

Performance par watt

Toutes les architectures NVIDIA Ampere sont conçues pour améliorer l'efficacité - de la logique, de la mémoire, de l'alimentation et de la gestion thermique à la conception de circuits imprimés, aux logiciels et aux algorithmes. Au même niveau de performances, les GPU Ampere sont jusqu'à 1,9 fois plus écoénergétiques que les appareils Turing comparables.

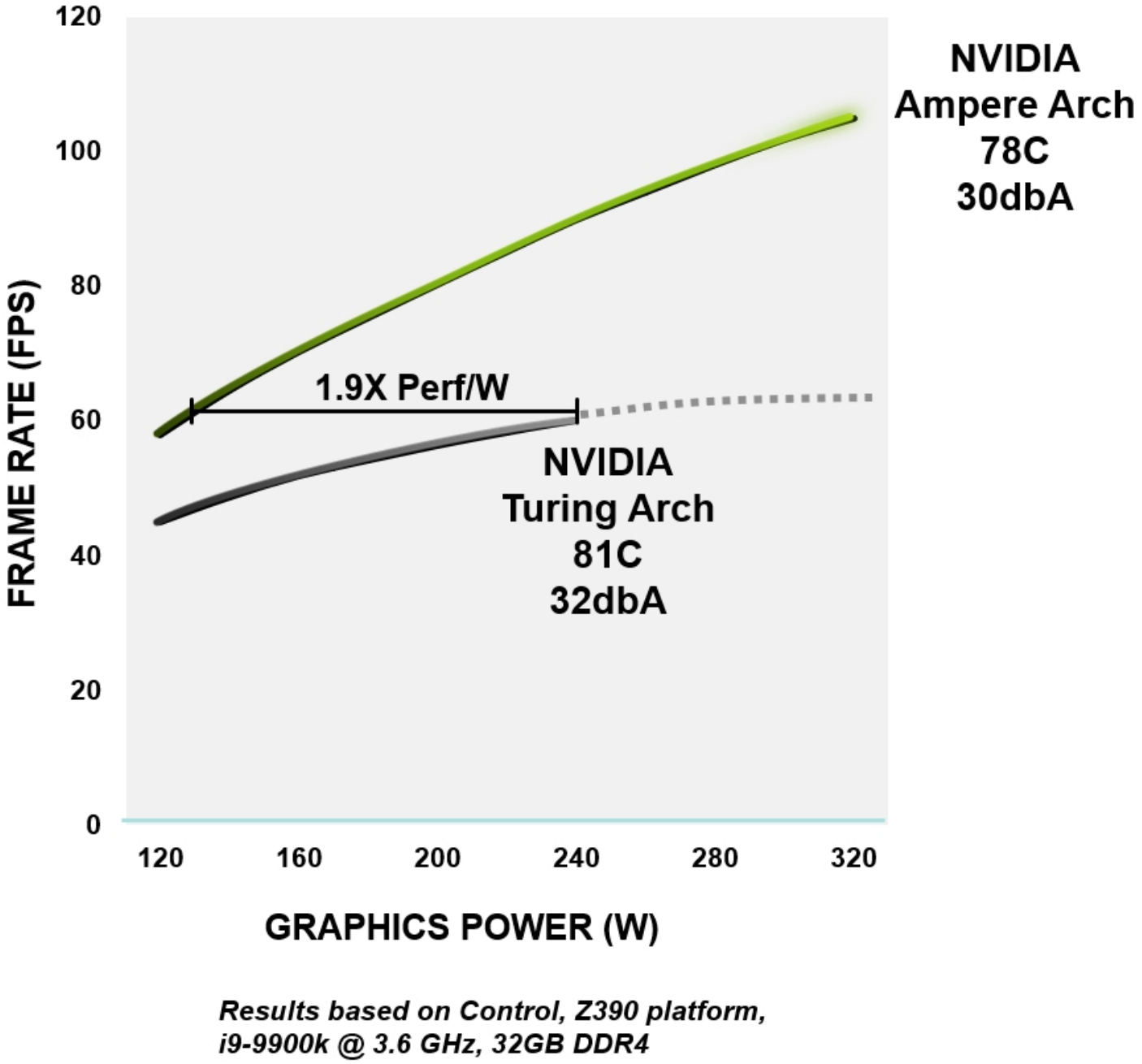


Figure 4. Efficacité énergétique du RTX 3080 par rapport à l'architecture GeForce RTX 2080 Super

Cœurs RT de deuxième génération

Les nouveaux cœurs RT présentent un certain nombre d'améliorations qui, combinées à des systèmes de mise en cache mis à jour, doublent efficacement les performances de lancer de rayons des processeurs Ampere par rapport à Turing. De plus, le GA10x permet à d'autres processus d'être exécutés simultanément avec le calcul RT, accélérant ainsi considérablement de nombreuses tâches.

Traçage de rayons de deuxième génération dans GA10x

Les GeForce RTX basés sur l'architecture Turing ont été les premiers GPU avec lesquels le lancer de rayons cinématographique est devenu une réalité dans les jeux PC. Le GA10x est équipé d'une technologie de traçage de rayons de deuxième génération. Comme Turing, les multiprocesseurs du GA10x ont des blocs matériels spécialisés pour vérifier les intersections de rayons avec les BVH et les triangles. Dans le même temps, les cœurs des multiprocesseurs Ampere ont deux fois la vitesse de test de l'intersection des rayons et des triangles par rapport à Turing.

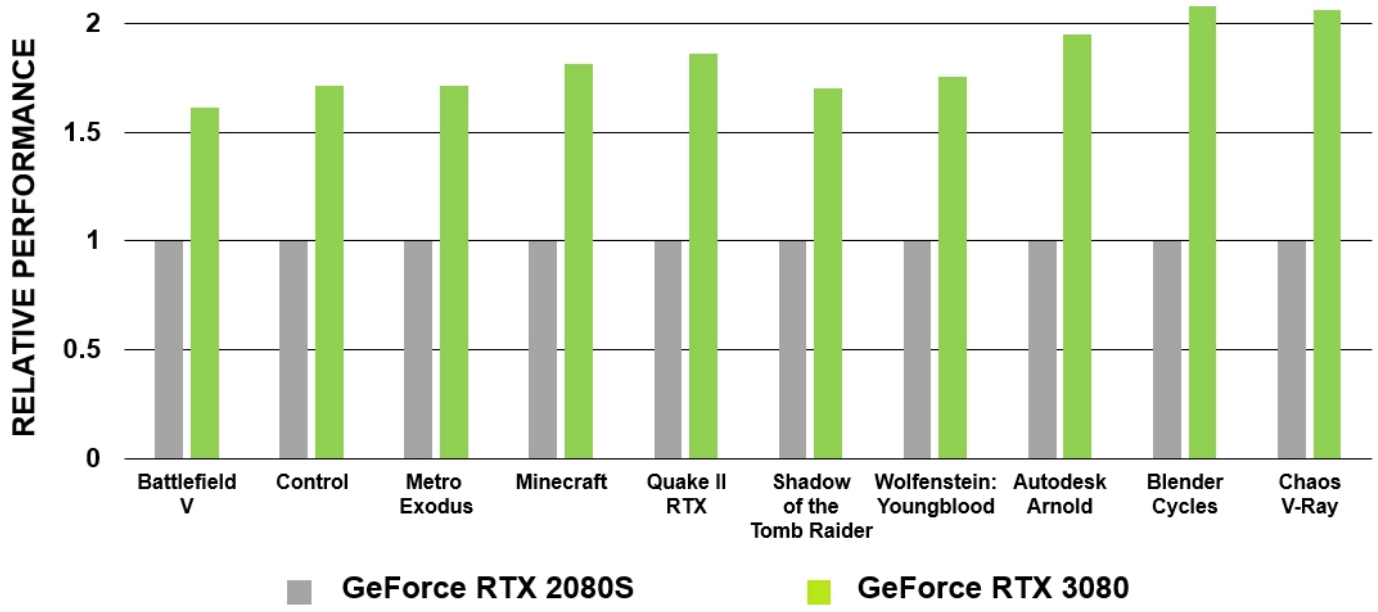


Figure 5. Comparaison des performances des cœurs RT de GeForce RTX 3080 et GeForce RTX 2080 Super

Le multiprocesseur GA10x peut effectuer des opérations simultanément et n'est pas limité au calcul et aux graphiques uniquement, comme c'était le cas avec les générations précédentes de GPU. Ainsi, par exemple, dans le GA10x, l'algorithme de réduction du bruit peut être exécuté simultanément avec le lancer de rayons.

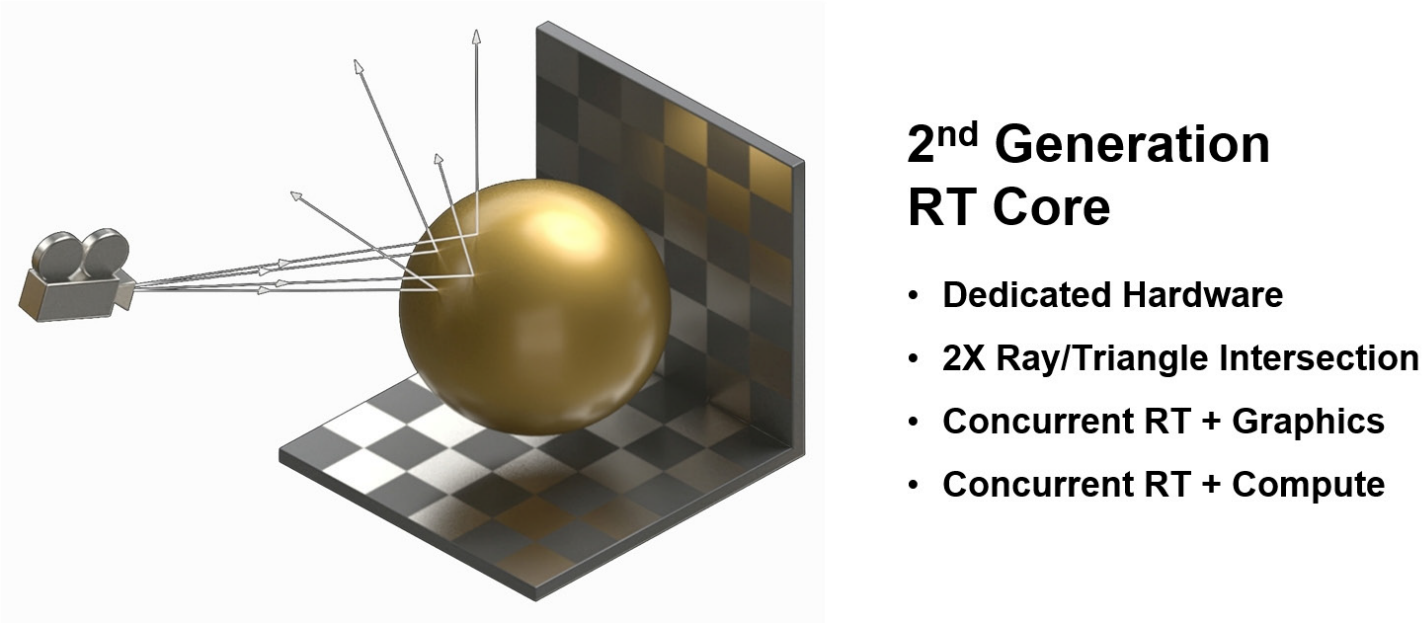


Figure 6. RT Core de deuxième génération dans les GPU GA10x

Notez que les charges de travail intensives en RT n'augmentent pas de manière significative la charge sur les cœurs multiprocesseurs, permettant ainsi à la puissance de traitement multiprocesseur d'être utilisée pour d'autres tâches. C'est un gros avantage par rapport aux autres architectures concurrentes qui n'ont pas de cœurs RT dédiés et doivent donc utiliser leurs blocs de construction pour les graphiques et le lancer de rayons.

Processeurs Ampère RTX en action

Le lancer de rayons et les shaders nécessitent beaucoup de calculs. Mais il serait beaucoup plus coûteux de tout exécuter avec des cœurs CUDA seuls, donc l'inclusion des cœurs tensor et RT permet d'accélérer considérablement le traitement. La figure 7 montre un exemple de jeu Wolfenstein: Youngblood avec le lancer de rayons activé dans divers scénarios.

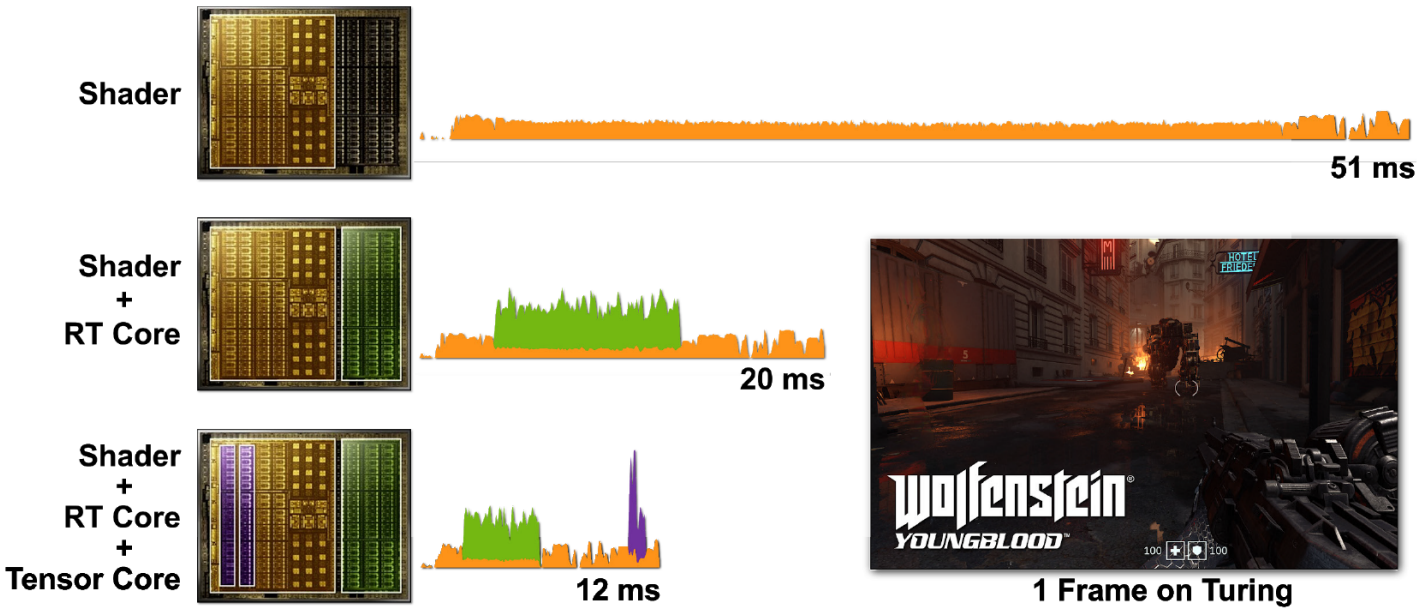


Figure 7. Rendu d'une seule image de Wolfenstein: Youngblood sur un Super GPU RTX 2080 en utilisant a) des cœurs de shader (CUDA), b) des cœurs de shader et des cœurs RT, c) des cœurs de shader, des cœurs de tenseur et des cœurs RT. Notez les temps de trame qui diminuent progressivement au fur et à mesure que vous ajoutez la puissance des différents cœurs de processeur RTX.

Dans le premier cas, il faut 51 ms (~ 20 ips) pour démarrer une image. Lorsque les cœurs RT sont activés, l'image est rendue beaucoup plus rapidement - en 20 ms (50 ips). L'utilisation de DLSS sur des cœurs de tenseur réduit le temps de trame à 12 ms (~ 83 ips).

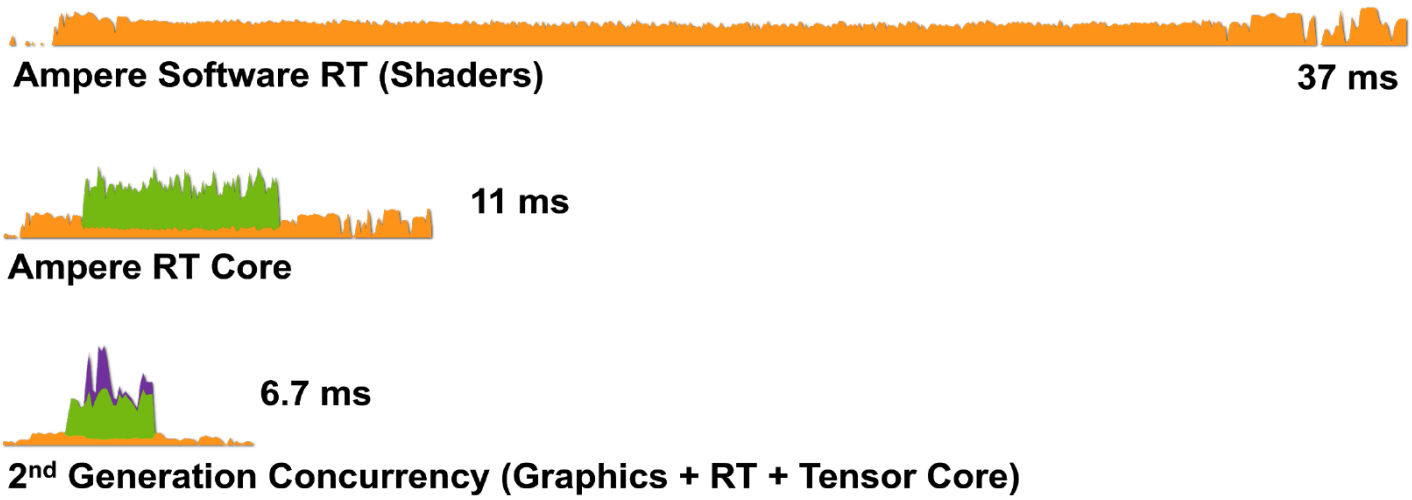


Figure 8. Rendu d'une seule image de Wolfenstein: Youngblood sur un RTX 3080 en utilisant a) des cœurs de shader (CUDA), b) des cœurs de shader et des cœurs RT, c) des cœurs de shader, des cœurs de tenseur et RT.

Ainsi, la technologie RTX avec l'architecture Ampère est encore plus efficace pour gérer les tâches de rendu: le RTX 3080 restitue une image en 6,7 ms (150 ips), ce qui est une énorme amélioration par rapport au RTX 2080.

Traçage de rayons accéléré par le matériel utilisant le flou de mouvement

Le flou de mouvement est un mouvement souvent utilisé en infographie. Une image photographique n'est pas créée instantanément, mais en exposant le film à la lumière pendant une durée limitée. Les sujets se déplaçant assez rapidement par rapport au temps d'exposition de l'appareil photo apparaîtront sur la photo sous forme de stries ou de taches. Pour que le GPU crée un flou de mouvement d'aspect réaliste lorsque les objets d'une scène se déplacent rapidement devant une caméra statique, il doit être capable de simuler le fonctionnement de la caméra et du film avec de telles scènes. Le flou de mouvement est particulièrement important dans la réalisation de films car les films sont lus à 24 images par seconde et une scène sans flou de mouvement apparaîtra nette et saccadée.

Les GPU Turing font un très bon travail d'accélération du flou de mouvement en général. Cependant, dans le cas d'une géométrie en mouvement, la tâche peut être plus difficile, car les informations sur le BVH changent avec la position des objets dans l'espace.

Comme vous pouvez le voir sur la figure 9, le noyau Turing RT effectue une traversée matérielle de la hiérarchie BVH, en vérifiant l'intersection des rayons avec BBox et des triangles. Le GA10x peut faire tout de même, mais il dispose en outre d'un nouveau bloc Interpoler Triangle Position, qui accélère le flou de mouvement dans le lancer de rayons.

Les cœurs Turing et GA10x RT implémentent l'architecture MIMD (Multiple Instruction Multiple Data), qui permet de traiter plusieurs faisceaux simultanément.

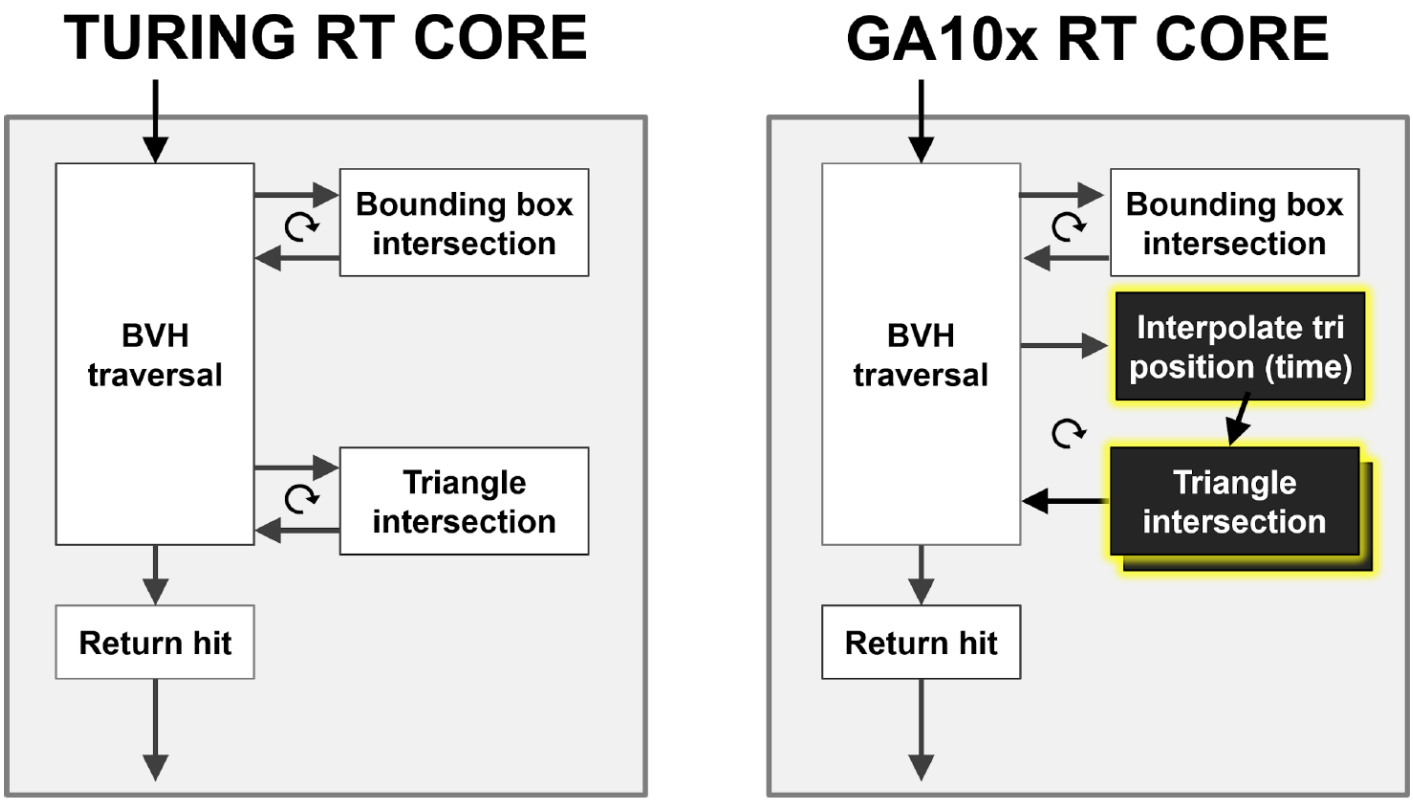
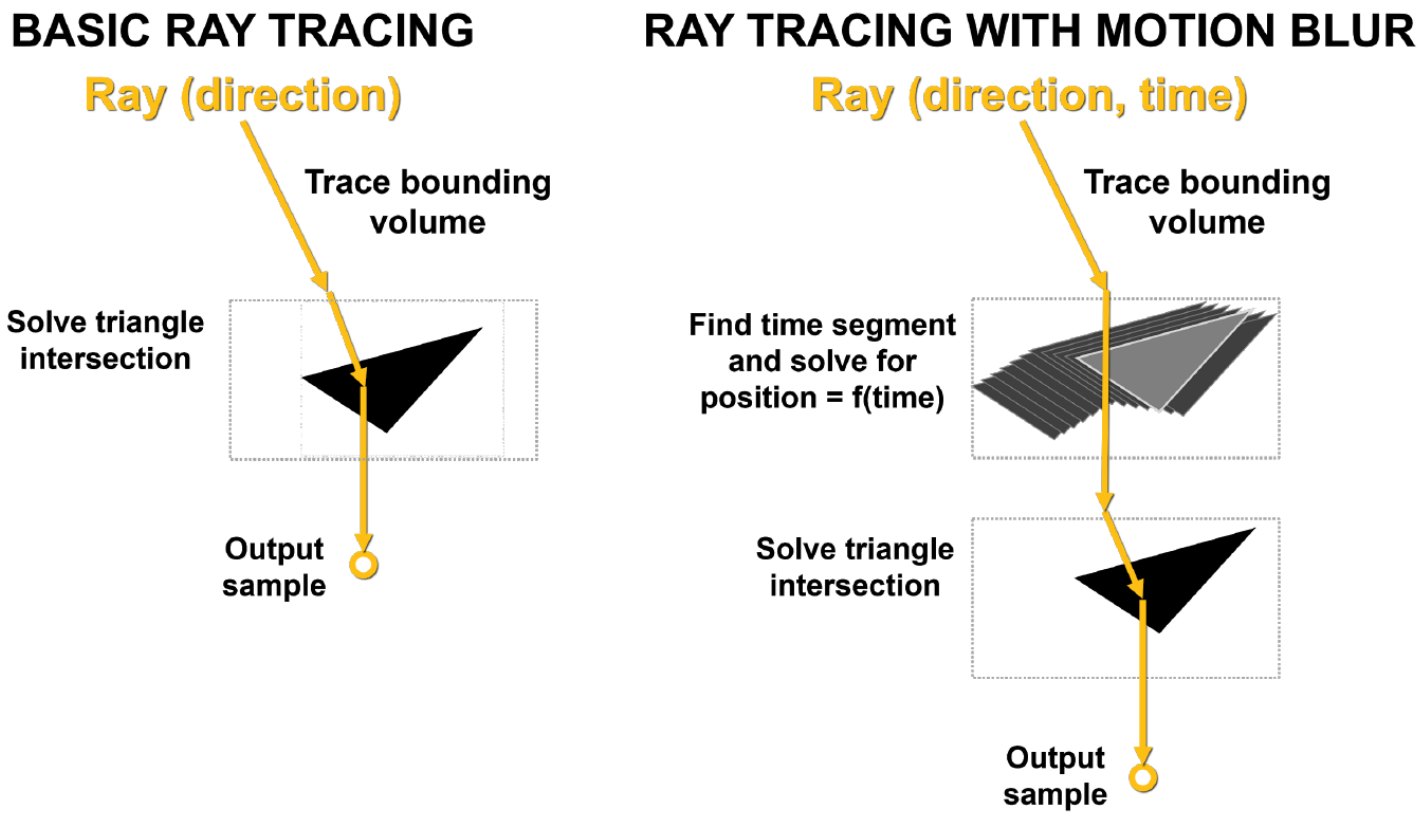


Figure 9. Comparaison de l'accélération matérielle du flou de mouvement dans le cas de Turing et Ampère

Le principal problème du flou de mouvement est que les triangles de la scène ne sont pas fixés dans le temps. Dans le lancer de rayons de base, des tests d'intersection statiques sont effectués, et lorsqu'un rayon atteint un triangle, il renvoie des informations sur ce coup. Comme le montre la figure 10, avec le flou de mouvement, aucun des triangles n'a de coordonnées fixes. Chaque rayon est horodaté pour indiquer son temps de suivi, et la position du triangle et l'intersection du rayon sont déterminées à partir de l'équation BVH.

Si ce processus n'est pas accéléré par le matériel, il peut vraiment causer beaucoup de problèmes, notamment en raison de sa non-linéarité.



Dessin. 10. Lancer de rayons de base et lancer de rayons avec flou de mouvement

Sur le côté gauche de la figure 11, les rayons envoyés à une scène statique ont frappé le même triangle en même temps. Les points blancs indiquent le lieu de l'impact, ce résultat est renvoyé. Dans le cas du flou de mouvement, chaque rayon existe à son propre moment dans le temps. Chaque faisceau se voit attribuer au hasard un horodatage différent. Par exemple, les rayons orange essaient de traverser les triangles orange en même temps, puis les rayons vert et bleu font la même chose. À la fin, les échantillons sont mélangés, produisant un résultat flou plus mathématiquement correct.

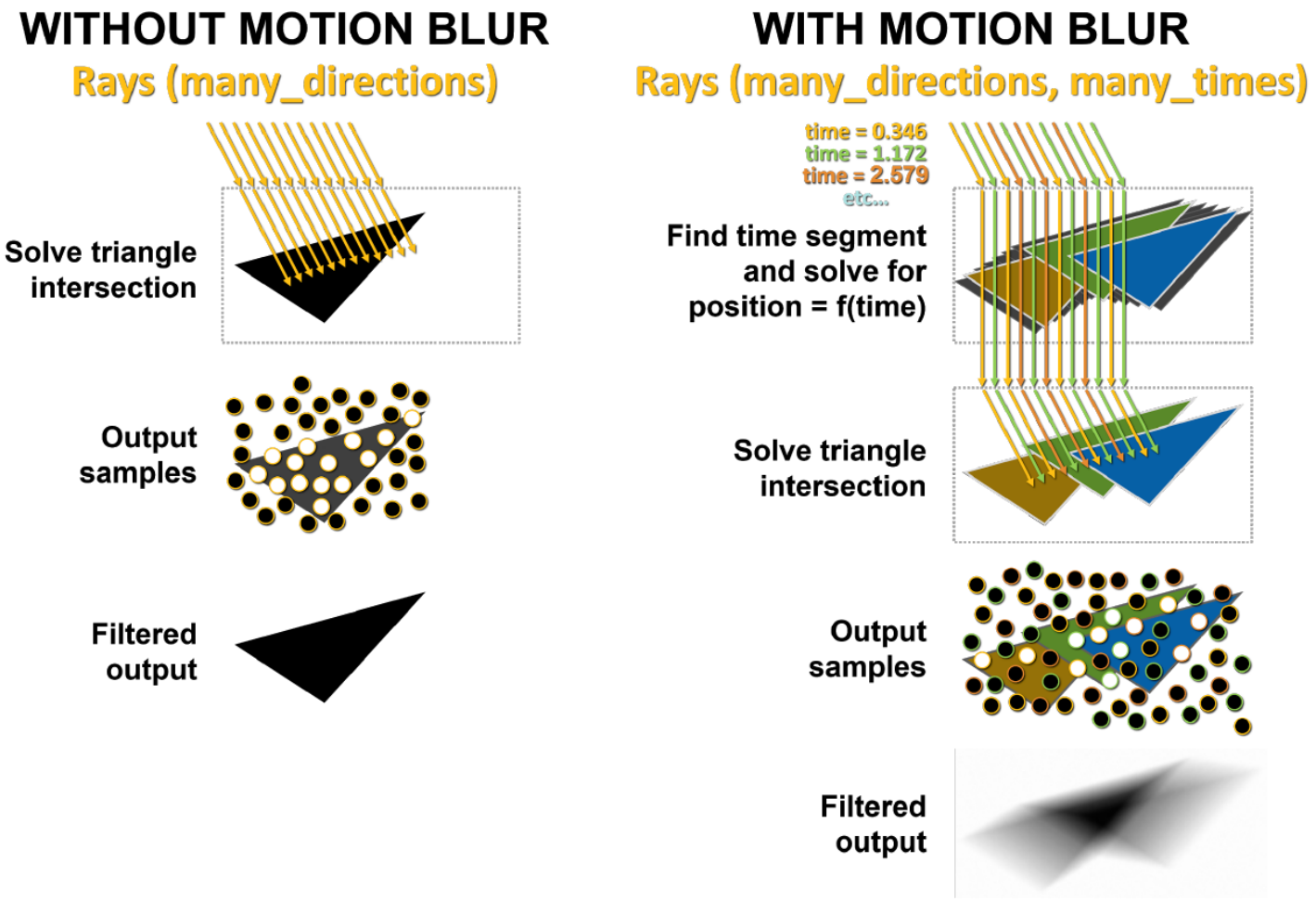


Figure 11. Rendu sans flou de mouvement et avec flou dans GA10x

Le bloc Interpoler Triangle Position interpole les triangles dans BVH entre les triangles déjà existants en fonction du mouvement de l'objet, de sorte que les rayons les coupent aux emplacements attendus aux moments spécifiés par les horodatages des rayons. Cette approche permet un rendu précis du flou de mouvement par lancer de rayons jusqu'à huit fois plus rapide que Turing.

Le flou de mouvement accéléré par le matériel GA10x est pris en charge par Blender 2.90, Chaos V-Ray 5.0, Autodesk Arnold et Redshift Renderer 3.0.X à l'aide de l'API NVIDIA OptiX 7.0.

La vitesse de rendu du flou de mouvement est jusqu'à 5 fois plus rapide avec le RTX 3080 par rapport au RTX 2080 Super.

Tensor Cores de 3e génération dans les GPU GA10x

Le GA10x intègre de nouveaux cœurs Tensor NVIDIA de troisième génération, offrant une prise en charge de nouveaux types de données, des performances, une efficacité et une flexibilité de

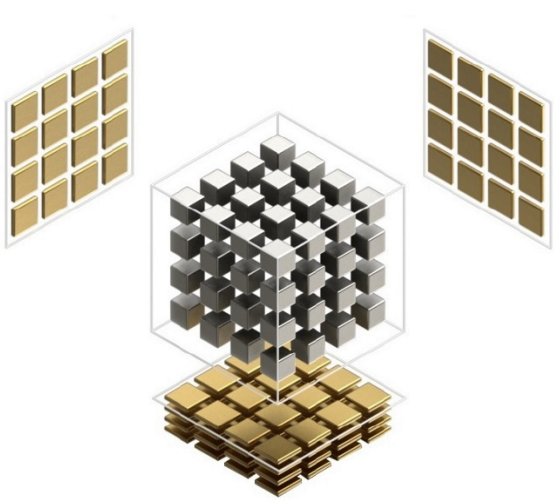
programmation améliorées. La nouvelle fonction de parcimonie double les performances des cœurs Tensor par rapport à la génération précédente de Turing. Les fonctions AI telles que NVIDIA DLSS pour la super-résolution AI (maintenant avec prise en charge 8K), NVIDIA Broadcast pour le traitement de la voix et de la vidéo et NVIDIA Canvas pour le dessin sont également plus rapides.

Les noyaux Tensor sont des unités d'exécution spécialisées conçues pour effectuer des opérations tensorielles / matricielles - la principale fonction de calcul de l'apprentissage profond. Ils sont nécessaires pour améliorer la qualité graphique avec DLSS (Deep Learning Super Sampling), la suppression du bruit basée sur l'IA, la suppression du bruit de fond dans les conversations vocales de jeu avec RTX Voice et bien d'autres applications.

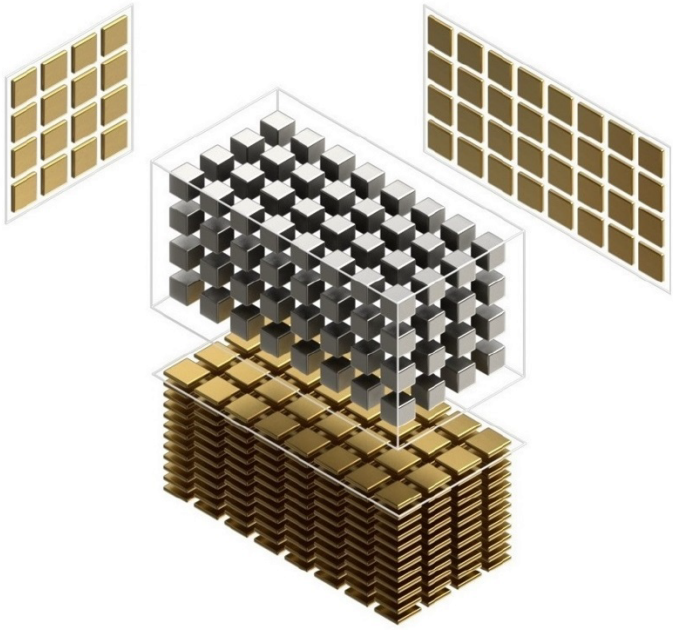
L'introduction des cœurs Tensor dans les GPU de jeu GeForce a permis pour la première fois un apprentissage profond en temps réel dans les applications de jeu. La conception du noyau de tenseur de troisième génération des GPU GA10x augmente encore les performances brutes et exploite de nouveaux modes de précision de calcul tels que TF32 et BFloat16. Cela joue un rôle important dans les applications de services neuronaux NVIDIA NGX basées sur l'intelligence artificielle pour améliorer les graphiques, le rendu et d'autres fonctionnalités.

Comparaison des cœurs de Turing et Ampere Tensor

Les noyaux Ampere Tensor ont été réorganisés sur Turing pour améliorer l'efficacité et réduire la consommation d'énergie. L'architecture du cœur Ampère SM a moins de cœurs tensoriels, mais chacun est plus puissant.



TURING ARCHITECTURE TENSOR CORE
(GeForce RTX 2080 Super)



AMPERE ARCHITECTURE TENSOR CORE with Sparsity
(GeForce RTX 3080)

Figure 12. Cœurs Tensor avec architecture Turing et Ampère. La GeForce RTX 3080 offre une bande passante FP16 Tensor Core 2,7 fois plus rapide que la GeForce RTX 2080 Super

Clarté structurée à grain fin

Avec le GPU A100, NVIDIA introduit la Sparsity Structurée Fine-Grained, une nouvelle approche pour doubler la bande passante de calcul pour les réseaux de neurones profonds. Cette fonctionnalité est également prise en charge par les GPU GA10x et permet d'accélérer certaines opérations de rendu graphique basées sur l'IA.

Étant donné que les réseaux d'apprentissage profond peuvent adapter les pondérations grâce à l'apprentissage par rétroaction, en général, les contraintes structurelles n'affectent pas la précision des modèles entraînés.

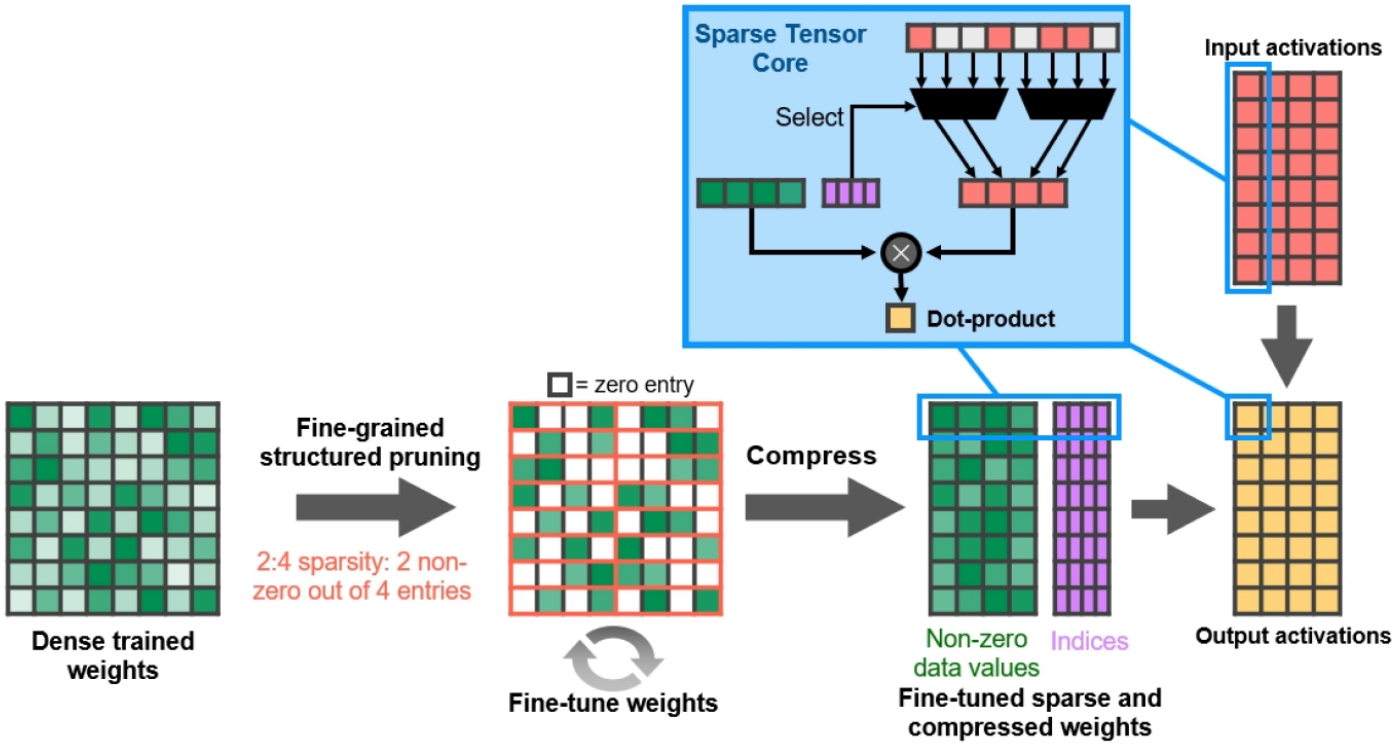


Figure 13. Clarté structurée à grain fin

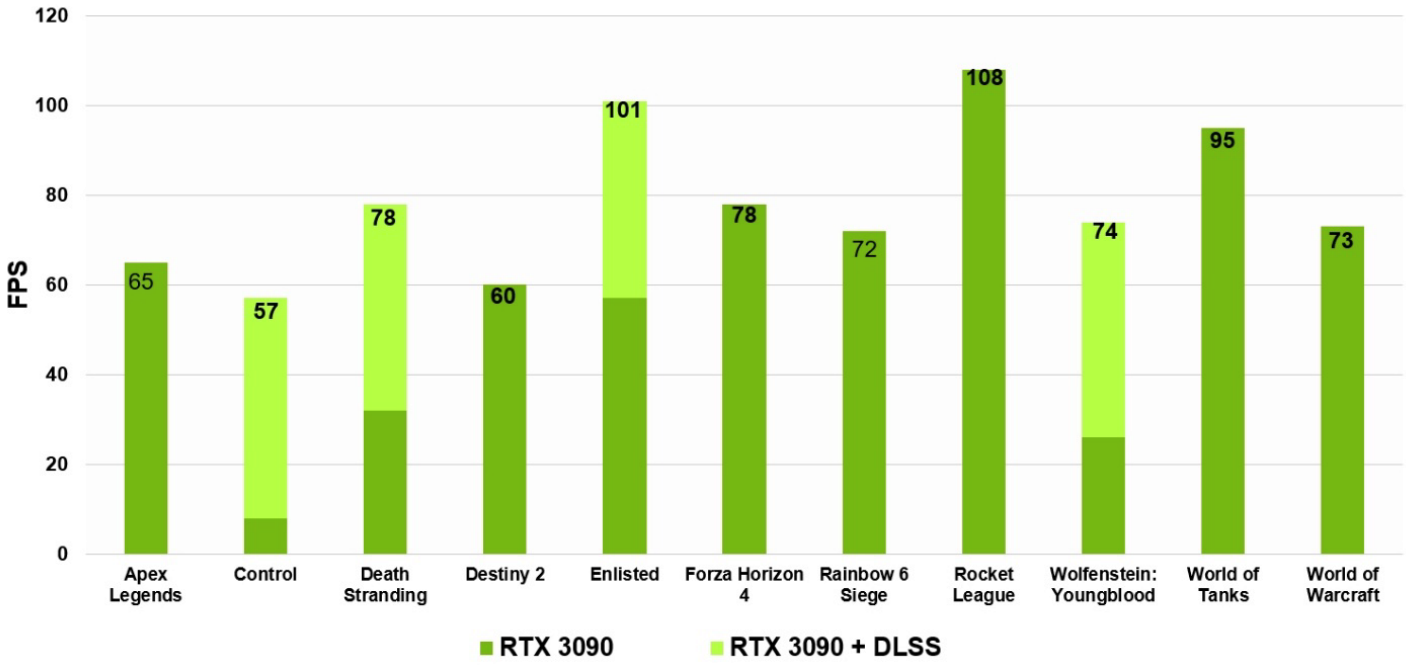
NVIDIA a développé un algorithme de parcimonie de réseau neuronal profond simple et polyvalent utilisant un modèle de parcimonie structuré 2: 4. Le réseau est d'abord formé avec des poids denses, puis un élagage structuré à grain fin se produit, après quoi les valeurs nulles peuvent être rejetées et les mathématiques restantes sont compressées pour augmenter le débit. L'algorithme n'affecte pas la précision du réseau formé pour l'inférence, il ne fait que l'accélérer.

NVIDIA DLSS 8K

Le rendu d'une image avec le lancer de rayons à une fréquence d'images élevée est extrêmement coûteux en calcul. Avant l'avènement de NVIDIA Turing, on pensait que sa mise en œuvre prendrait des années. Pour résoudre ce problème, NVIDIA a créé le Deep Learning Supersampling (DLSS).

Figure 14. Watch Dogs: Legion avec DLSS à 1080p, 4K et 8K. Notez que le texte et les détails plus nets fournis par DLSS dans 8K

DLSS ne font que s'améliorer sur NVIDIA Ampere grâce à l'utilisation de Tensor Cores de troisième génération et d'un facteur de mise à l'échelle de super-résolution 9x, qui pour la première fois permet d'exécuter un jeu de lancer de rayons à 8K à 60 ips.



15. GeForce RTX 3090 60 fps 8K DLSS . , . Core i9-10900K

GDDR6X

Les jeux PC modernes et les applications créatives nécessitent beaucoup plus de bande passante mémoire pour gérer une géométrie de scène de plus en plus complexe, des textures plus détaillées, le traçage de rayons, l'inférence IA et bien sûr l'ombrage et le suréchantillonnage.

GDDR6X est la première mémoire graphique à dépasser 900 Go / s. Pour ce faire, une technologie de signalisation innovante et une modulation d'amplitude d'impulsion à quatre niveaux (PAM4) ont été utilisées, révolutionnant collectivement la façon dont les données sont déplacées en mémoire. Avec l'algorithme PAM4, GDDR6X transmet plus de données à un rythme beaucoup plus rapide, déplaçant deux bits de données à la fois, ce qui double le débit de données d'E / S du schéma PAM2 / NRZ précédent.

GDDR6X prend actuellement en charge 19,5 Gbps pour la GeForce RTX 3090 et 19 Gbps pour la GeForce RTX 3080. Grâce à cela, la GeForce RTX 3080 offre 1,5 fois les performances mémoire de son prédécesseur, le RTX 2080 Super. ...

La figure 16 montre une comparaison de la structure de GDDR6 (à gauche) et GDDR6X (à droite). GDDR6X transmet les mêmes données à la moitié de la fréquence de GDDR6. Ou bien, GDDR6X peut doubler sa bande passante effective tout en conservant la même fréquence.

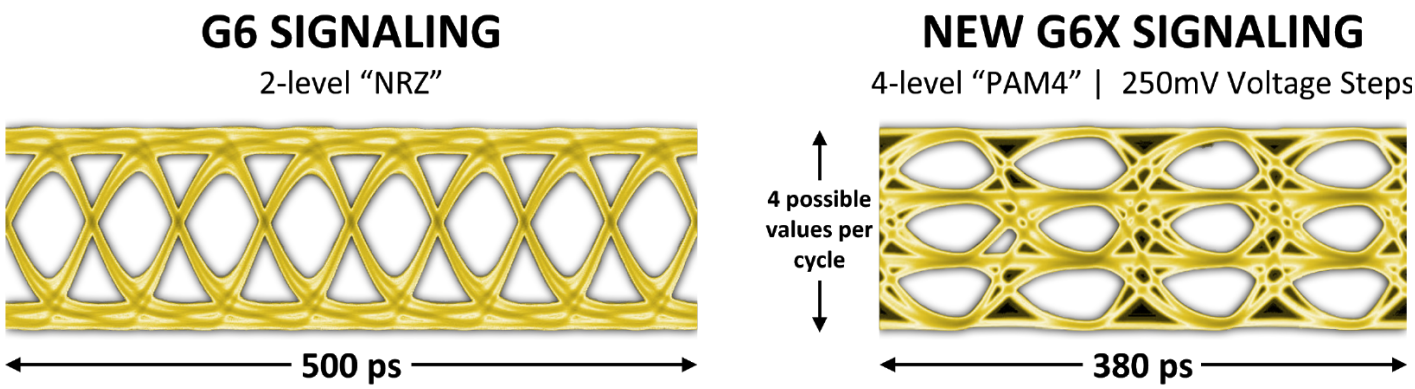


Figure 16. GDDR6X utilisant des signaux PAM4 montre de meilleures performances et efficacité que GDDR6

Un nouveau schéma de codage MTA (Maximum Transition Prevention) a été développé pour résoudre les problèmes de SNR associés à la signalisation PAM4. Le MTA empêche les signaux à haut débit d'aller du plus haut au plus bas et vice versa.

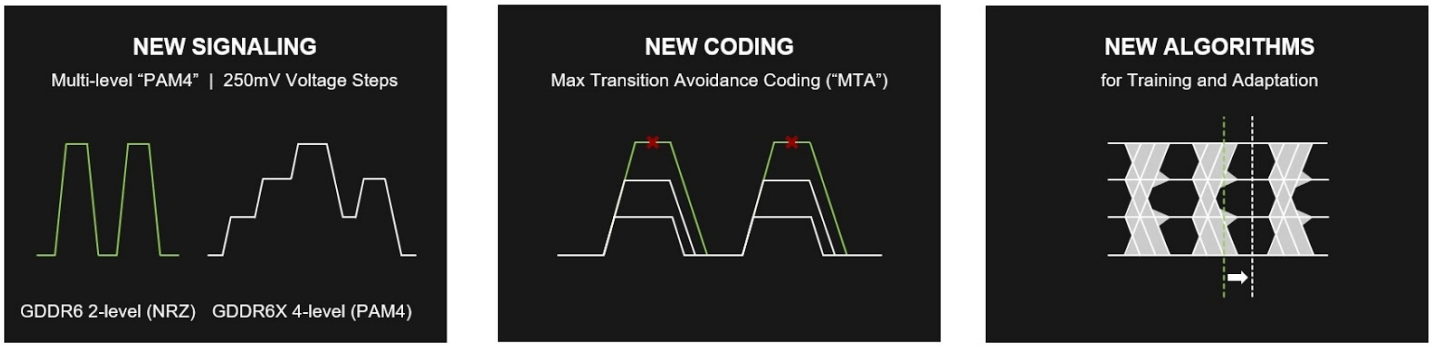


Figure 17. Nouveau codage dans GDDR6X Prenant en charge

des débits de données jusqu'à 19,5 Gbit / s sur les puces GA10x, GDDR6X offre une bande passante mémoire de pointe allant jusqu'à 936 Go / s, soit 52% de plus que le GPU TU102 utilisé dans GeForce RTX 2080 Ti. GDDR6X a le plus grand saut en bande passante en 10 ans après les GPU de la série GeForce 200.

RTX IO

Les jeux modernes contiennent des mondes immenses. Avec le développement de technologies telles que la photogrammétrie, elles imitent de plus en plus la réalité et, par conséquent, sont contenues dans des fichiers de plus en plus volumineux. Les plus grands projets de jeux occupent plus de 200 Go, soit 3 fois plus qu'il y a quatre ans, et ce nombre ne fera qu'augmenter avec le temps.

Les joueurs se tournent de plus en plus vers les SSD pour réduire les temps de chargement des jeux: alors que les disques durs sont limités à une bande passante de 50 à 100 Mo / s, les derniers SSD M.2 PCIe Gen4 lisent les données jusqu'à 7 Go / s.

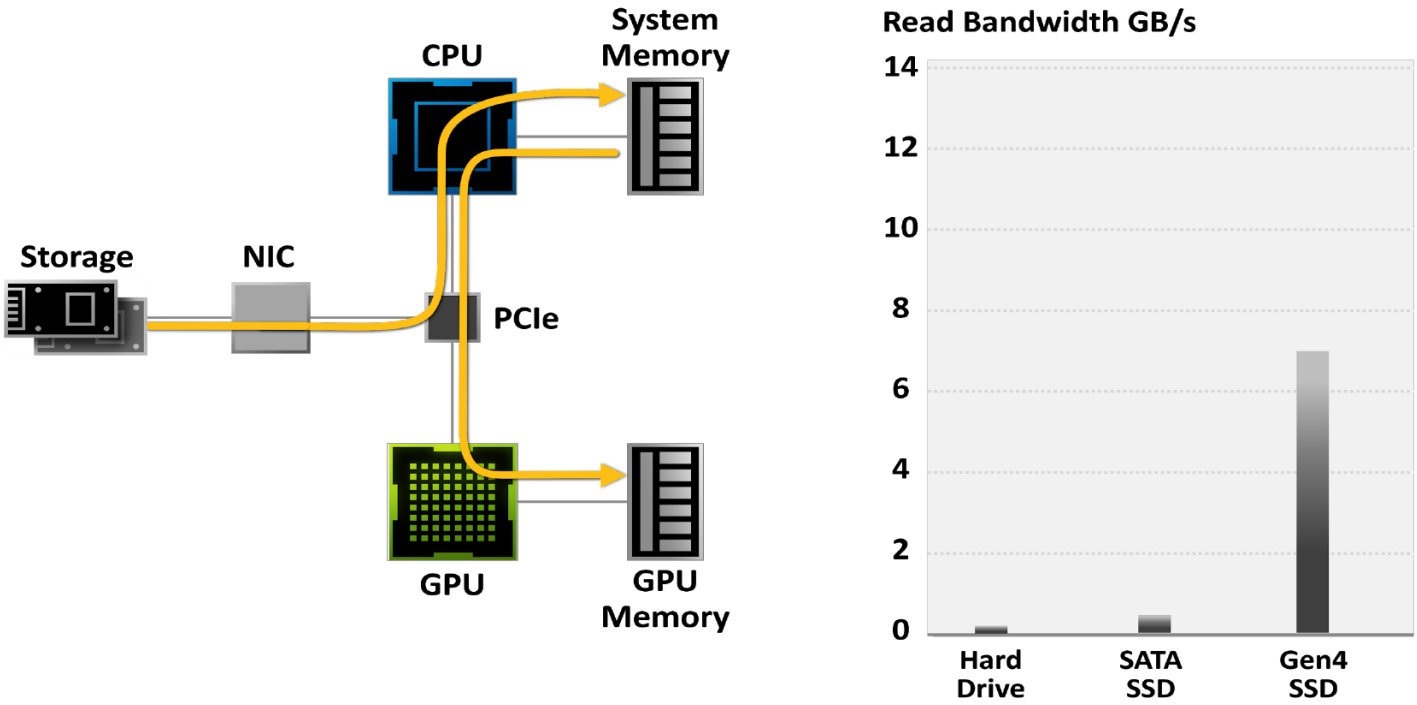


Figure 18. Jeux limités par les systèmes d'E / S traditionnels

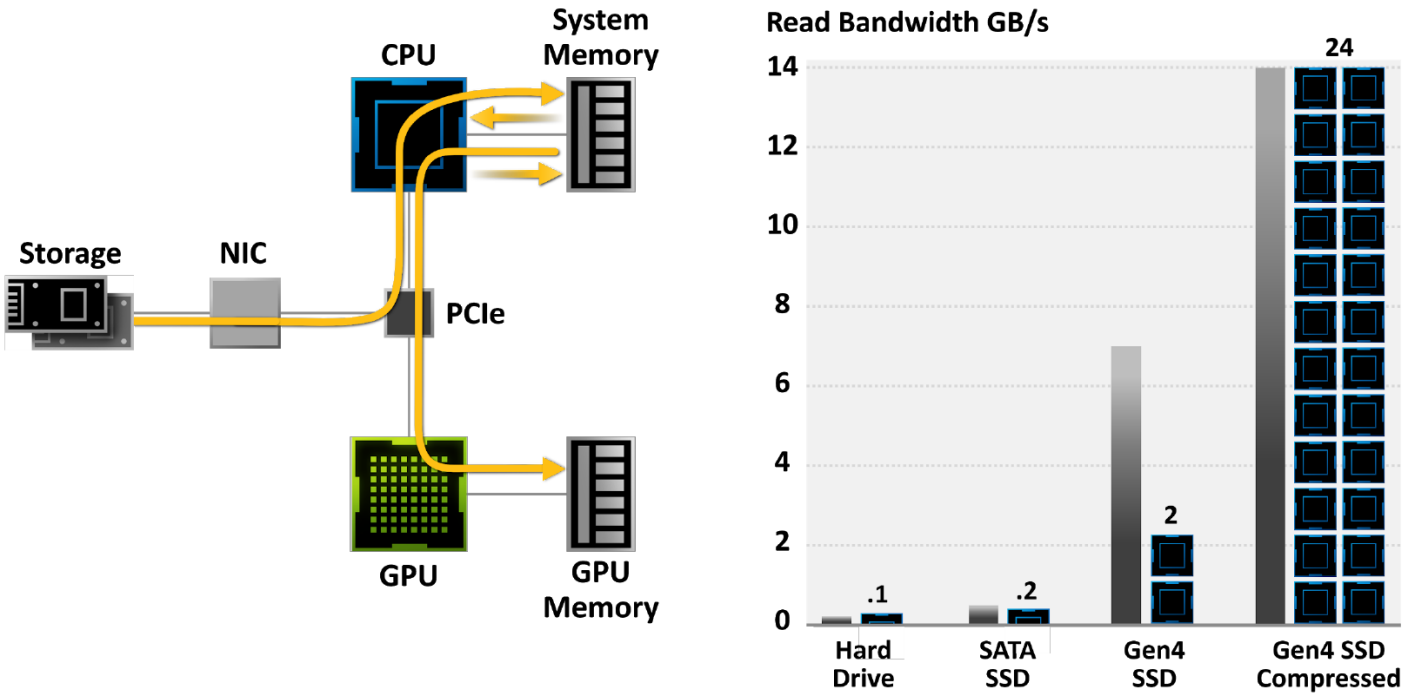


Figure 19. En utilisant le modèle de stockage traditionnel, le déballage d'un jeu peut prendre les 24 cœurs de processeur. Les moteurs de jeu modernes ont surpassé les capacités des API de stockage traditionnelles. C'est pourquoi une nouvelle génération d'architecture d'E / S est nécessaire. Ici, des barres grises indiquent le taux de transfert de données, des blocs noirs et bleus - les cœurs de processeur nécessaires pour cela.

NVIDIA RTX IO est un ensemble de technologies qui permettent un chargement et un déballage rapides des ressources GPU et offrent des performances d'E / S jusqu'à 100 fois plus rapides que les disques durs et les API de stockage traditionnelles.

NVIDIA RTX IO est alimenté par l'API Microsoft DirectStorage, un stockage de nouvelle génération spécialement conçu pour les PC de jeu SSD NVMe d'aujourd'hui. NVIDIA RTX IO offre une décompression sans perte, permettant aux données d'être lues sous forme compressée via DirectStorage et livrées au GPU. Cela décharge la charge du processeur en déplaçant les données du stockage vers le GPU sous une forme compressée plus efficace et en doublant les performances d'E / S.

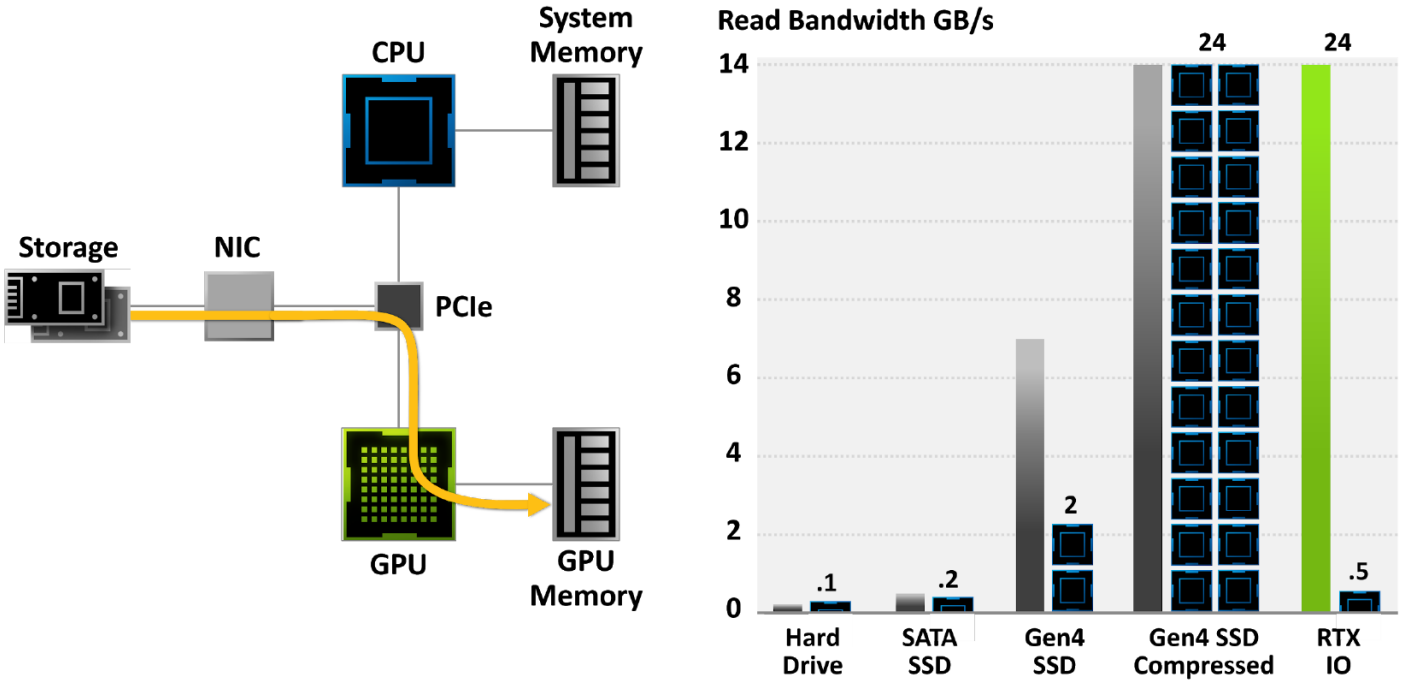


Figure 20. RTX IO fournit 100 fois la bande passante et 20 fois l'utilisation du processeur. Des barres grises et vertes indiquent la vitesse de transmission, des blocs noirs et bleus sont nécessaires pour ce cœur de processeur.

Moteur d'affichage et vidéo

DisplayPort 1.4a avec DSC 1.2a

La marche vers des résolutions toujours plus élevées et des fréquences d'images plus élevées se poursuit, et les GPU NVIDIA Ampere s'efforcent de rester à la pointe de l'industrie pour offrir les

deux. Les joueurs peuvent désormais jouer sur des écrans 4K (3820 x 2160) à 120 Hz et 8K (7680 x 4320) à 60 Hz, soit quatre fois le nombre de pixels de 4K.

Le moteur d'architecture Ampere est conçu pour prendre en charge de nombreuses technologies émergentes incluses dans les interfaces d'affichage les plus rapides actuellement disponibles. Cela inclut DisplayPort 1.4a, qui fournit 8K @ 60Hz avec VESA Display Stream Compression (DSC) 1.2a. Les nouveaux GPU Ampere peuvent être connectés à deux écrans 8K 60Hz avec un seul câble par écran.

HDMI 2.1 avec DSC 1.2a

L'architecture NVIDIA Ampere ajoute la prise en charge de HDMI 2.1, la dernière mise à jour de la spécification HDMI, pour la première fois pour les GPU discrets. HDMI a augmenté la bande passante maximale à 48 Gbit / s, ce qui permet également des formats HDR dynamiques. La prise en charge de 8K @ 60Hz avec HDR nécessite une compression DSC 1.2a ou un format de pixel 4: 2: 0.

NVDEC de 5e génération - Décodage vidéo accéléré par le matériel

Les GPU NVIDIA incluent le décodage vidéo accéléré par le matériel (NVDEC) de 5e génération, qui fournit un décodage vidéo matériel complet pour une variété de codecs populaires.

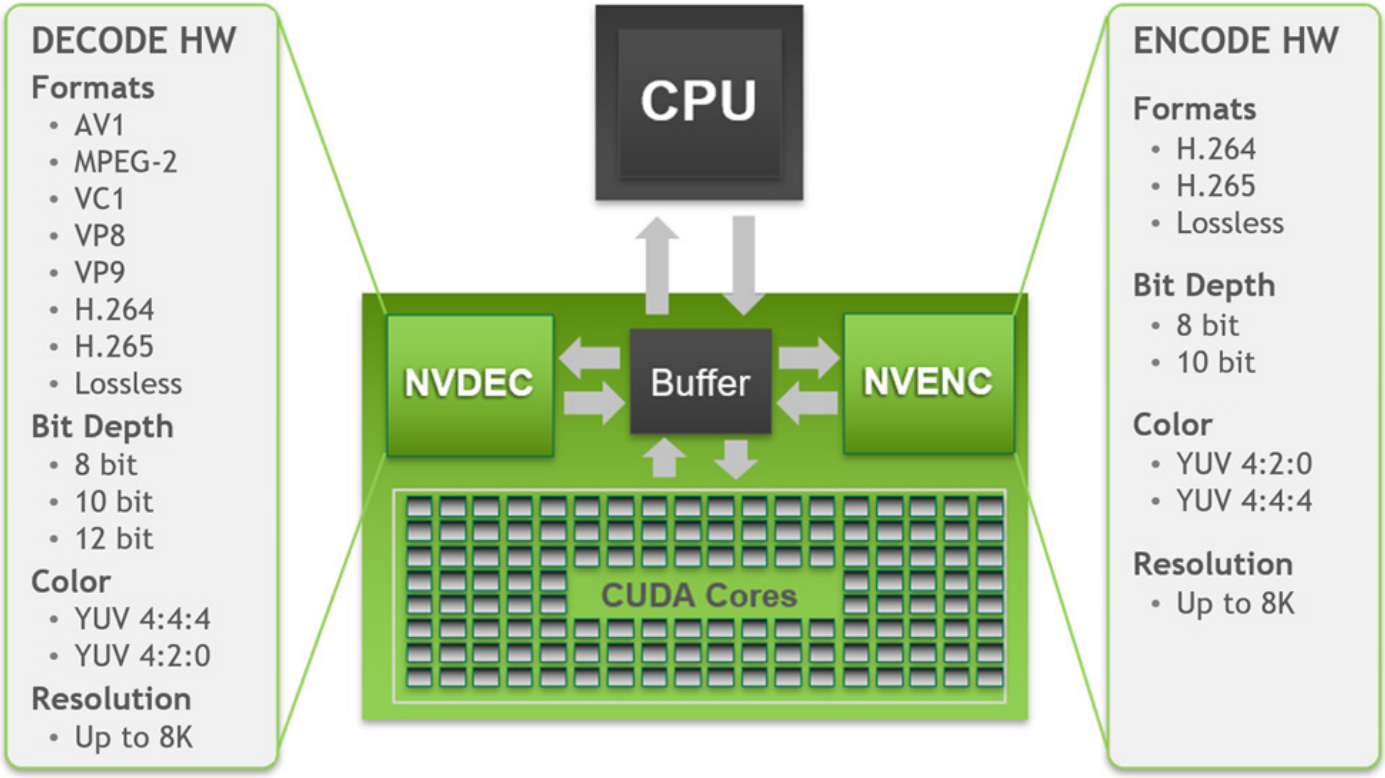


Figure 21. Formats d'encodage et de décodage vidéo pris en charge par les

GPU *GA10x* Le décodeur NVIDIA de cinquième génération du GA10x prend en charge le décodage accéléré matériel des codecs vidéo suivants sur les plates-formes Windows et Linux: MPEG-2, VC-1, H.264 (AVCHD), H.265 (HEVC), VP8, VP9 et AV1.

NVIDIA est le premier fabricant de GPU à fournir un support matériel pour le décodage AV1.

Décodage matériel AV1

Bien qu'AV1 soit très efficace pour compresser la vidéo, son décodage demande beaucoup de calculs. Les décodeurs logiciels modernes entraînent une utilisation élevée du processeur et rendent difficile la lecture de vidéos ultra haute définition. Lors des tests NVIDIA, le processeur Intel i9 9900K affichait en moyenne 28 images par seconde sur YouTube en 8K60 HDR, avec une utilisation du processeur dépassant 85%. Les GPU GA10x peuvent lire AV1 en passant le décodage à NVDEC, qui est capable de lire jusqu'à 8K60 de contenu HDR avec une très faible utilisation du processeur (~ 4% sur le même processeur que dans le test précédent).

NVENC 7e génération - Encodage vidéo accéléré par le matériel

L'encodage vidéo peut être une tâche de calcul complexe, mais si vous le téléchargez sur NVENC, le moteur graphique et le processeur sont libérés pour d'autres opérations. Par exemple, lors de la diffusion de jeux sur Twitch.tv à l'aide du logiciel Open Broadcaster (OBS), le déchargement de l'encodage vidéo vers NVENC permettra d'allouer le moteur GPU pour le rendu du jeu et le processeur pour d'autres tâches utilisateur.

NVENC permet:

- Encodage et streaming de haute qualité à latence ultra-faible de jeux et d'applications sans utiliser le processeur;
- encodage de très haute qualité pour l'archivage, le streaming OTT, la vidéo web;
- Encodage ultra-faible puissance par flux (W / flux).

Avec des paramètres de streaming partagés pour Twitch et YouTube, l'encodage matériel basé sur NVENC dans les GPU GA10x surpasse les encodeurs logiciels x264 utilisant le préréglage Fast et est comparable à x264 Medium, un préréglage qui nécessite généralement la puissance de deux ordinateurs. Cela supprime considérablement l'utilisation du processeur. L'encodage 4K représente une charge de travail trop importante pour une configuration de processeur typique, mais l'encodeur GA10x NVENC fournit un encodage haute résolution transparent jusqu'à 4K en H.264 et même 8K en HEVC.

Conclusion

Avec chaque nouvelle architecture de processeur, NVIDIA s'efforce d'offrir des performances révolutionnaires à la prochaine génération tout en introduisant de nouvelles fonctionnalités qui améliorent la qualité d'image. Turing a été le premier GPU à introduire le traçage de rayons accéléré par le matériel, une fonctionnalité autrefois considérée comme le Saint Graal de l'infographie. Aujourd'hui, des effets de lancer de rayons incroyablement réalistes et physiquement précis sont ajoutés à de nombreux nouveaux jeux PC AAA, et le lancer de rayons accéléré par GPU est considéré comme un must pour la plupart des joueurs sur PC. Les nouveaux GPU NVIDIA GA10x Ampere offrent les fonctionnalités et les performances dont vous avez besoin pour profiter de ces nouveaux jeux par lancer de rayons avec des fréquences d'images jusqu'à 2 fois plus rapides que celles actuellement disponibles.Une autre caractéristique de Turing - le traitement amélioré de l'IA accéléré par le processeur qui améliore la suppression du bruit, le rendu et d'autres applications graphiques - est également portée au niveau supérieur grâce à l'architecture Ampere.

Enfin, un lien vers le document complet .

More articles:

- Système de gestion des commentaires d'arbre hiérarchique pour Laravel
- Communication efficace pour les programmeurs introvertis
- Lecteur de Pocketbook VS Amazon Kindle VS Cool Reader
- Travailler chez Amazon WorkSpaces: expérience de déploiement et de configuration
- Technologie Apphost: un univers alternatif de microservices dans Yandex
- Comment une base de données mal configurée nous a permis de capturer un cloud entier avec 25000 hôtes
- Réécrire l'historique du référentiel de code, ou pourquoi vous pouvez parfois git push -f
- Les joies de posséder une adresse e-mail courte
- Implémentation de CI / CD et DevOps en Entreprise (dans notre cas, Rostelecom)
- Nos résultats pour un an de migration de GitLab.com vers Kubernetes

All Articles

Geek Cheese Tech (GCT)
17, rue Sadi Carnot
89000 AUXERRE, France
03.91.45.16.35