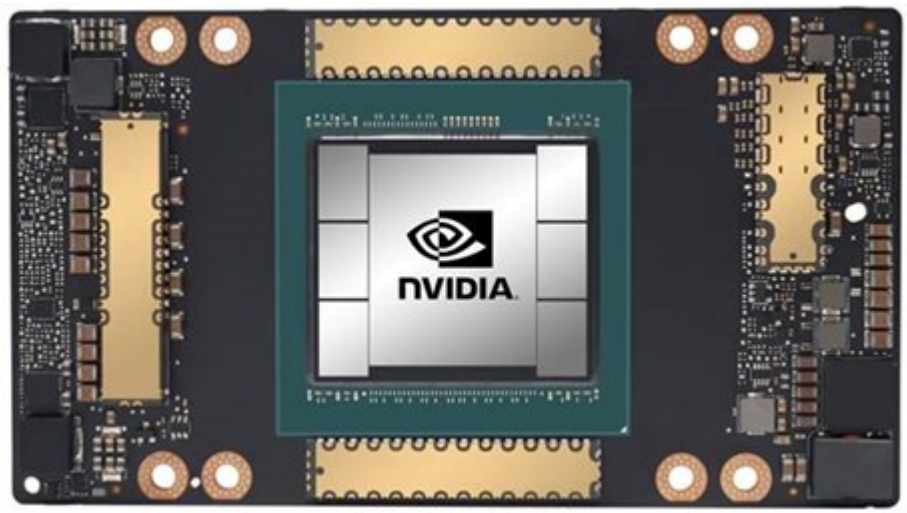
Le nouveau GPU A100 de Nvidia

Trois ans après le lancement de son GPU V100 de génération Volta qui est devenu un standard de l'équipement de data-centers pour alimenter nombre d'applications de deep learning et machine learning, Nvidia lançait enfin son successeur, le A100 Ampere, le 14 juin 2020. Cela faisait un bout de temps que je l'attendais.

J'anticipais ce lancement au printemps 2020 dans le dernier **Rapport du CES 2020**. Le voici donc enfin arrivé !



Comme j'aime bien dépiauter les processeurs extrêmes, celui-là est passé à la moulinette ! Cela vous changera des qubits de l'informatique quantique que j'ai l'habitude de couvrir depuis deux ans (mais nous y reviendrons, ne vous inquiétez pas…) ! Au passage, je tente de répondre dans cet article à quelques de questions de newbie que je me posais sur les calculs de matrices et d'expressions mathématiques à base de nombres flottants.

Un contexte défavorable

La crise du covid-19 a évidemment perturbé ces plans de lancement. À la fois au niveau de sa préparation par les équipes de Nvidia et par celles de ses partenaires et notamment des fabs de TSMC à Taiwan. Mais aussi, j'ai l'impression, parce que tout simplement, pendant le confinement et la montée exponentielle des décès, il était inconvenant de lancer quoi que ce soit qui ne soit pas relié au covid quasiment dans tous les domaines.

Qui plus est, l'IA n'était franchement pas "à la mode" pendant cette crise covid-19. On a découvert que l'IA et le machine learning étaient loin de fournir toutes les réponses face à l'inconnu. Voir à ce sujet l'excellent billet **Mais où est passée l'intelligence artificielle ?** de Benoit Raphaël, 16 mai 2020.

Et pour cause, l'IA *deep learninesque* manquait de données d'entraînement ! Comme le machine learning est un rétroviseur géant, lorsque la route et les règles de conduite changent presque entièrement, il est complètement perdu. Pour prédire les conséquences d'une épidémie, on a plus besoin de réseaux multi-agents permettant de modéliser sa propagation que de machine learning basé sur des données d'entraînement. Or les réseaux multi-agents sont un peu le parent pauvre de l'IA d'aujourd'hui. Depuis cinq ans, on n'a d'yeux que pour le deep learning qui est censé tout faire.

L'IA a toutefois joué un rôle dans des domaines très précis de la recherche comme pour accompagner le criblage de molécules, cette technique qui sert à identifier des pathologies pouvant être traitées par des médicaments existants pour lesquels ils n'avaient pas été initialement conçus. Elle a aussi servi à analyser des radios du poumon, dans la lignée des nombreux systèmes d'imagerie à base d'IA qui existent depuis quelques années, sans d'ailleurs être encore déployés à grande échelle.

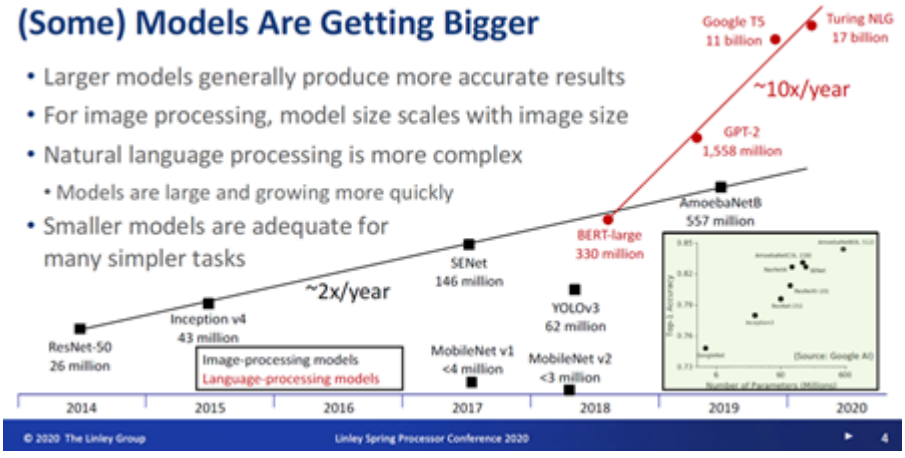
Mais l'empirisme a eu le dessus et on a surtout entendu parler de l'hydroxychloroquine qui n'avait pas du tout été sélectionnée avec des méthodes à base d'IA. Le covid a surtout mis en évidence les biais humains et en particulier les biais de corrélation simplificateurs, portant notamment sur la comparaison des méthodes de limitation de la propagation du virus (confinement ou pas, …). Un coup, on mettait en avant la distanciation sociale, puis les tests, puis les masques, puis la pollution, puis le groupe sanguin, puis les variantes du virus, puis la densité de la population et son âge moyen. Et pour cause, nous sommes face à un système multi-variant complexe dont on découvre progressivement les variables !

Une autre raison mettait l'IA en veilleuse : elle n'est pas la panacée dans une situation où le chômage complet ou partiel explose, surtout pour les cols blancs, qui sont les plus menacés par l'automatisation. Ce n'était pas le moment de faire empirer la situation !

La soif de puissance des processeurs de l'IA

Ce qui ne m'a pas empêché de récupérer les contenus de la conférence **Linley Spring Processor Conference 2020** tenue virtuellement entre le 6 et le 9 avril 2020 (le mot de passe des PDF s'obtient en fournissant ses coordonnées). Ses 496 slides permettent de faire le point sur l'actualité des nombreuses technologies de processeurs dédiés à l'IA, le thème central de cette édition de cette conférence annuelle sur les processeurs.

J'y ai déniché ce slide évocateur de la croissance de la taille des modèles de deep learning qu'il faut entraîner. Cette taille est liée au fait que des modèles avec un grand nombre de paramètres (en gros, un nombre de neurones dans un réseau de neurones) sont plus fiables, comme pour la reconnaissance d'images. Plus ils sont grands, plus est lourde la tâche de leur entraînement. C'est lié au besoin d'ajuster tous les paramètres des nombreux neurones mais aussi parce que cela s'accompagne d'une augmentation du nombre d'objets de la base d'entraînement.



Tout ceci nous rappelle un point clé du deep learning : les tâches d'entraînement sont très lourdes alors que, comparativement, une inférence pour reconnaître un objet, est bien plus rapide. De plusieurs ordres de grandeur. C'est la raison pour laquelle on peut facilement déléguer ces tâches d'inférence à des processeurs embarqués dans les objets ou proches des objets (edge AI).

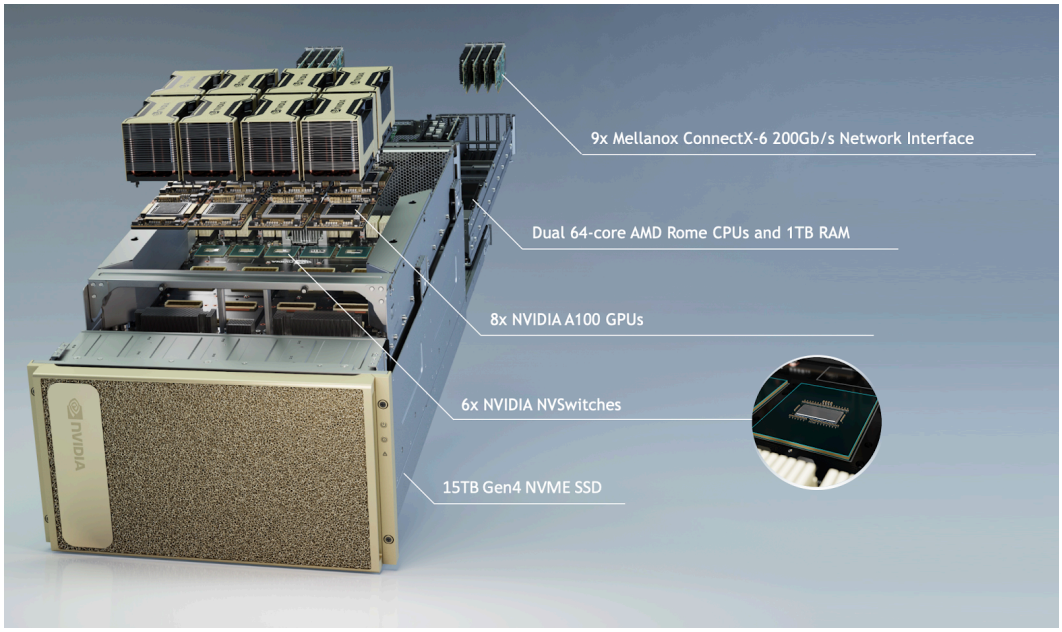
Le GPU A100 de Nvidia dont il est question ici sert à la fois aux tâches d'entrainement mais aussi aux

inférences réalisées en volume sur des serveurs de data-centers. Mais il est aussi déployable dans l'edge AI, pour traiter les données au plus proche des besoins sans passer par le cloud. Il succède au V100 de génération Volta lancé en 2017 et qui était devenu le standard de l'équipement dans le cloud pour l'IA. On le trouve par exemple chez OVH comme chez Scaleway (ex Online.net, filiale d'Iliad) ou CleverCloud, sans compter les géants Amazon AWS et Microsoft Azure. Il équipe aussi une bonne part des supercalculateurs récents comme l'IBM Summit et ses 27 648 GPU V100 complétés de 9216 CPU Power9 maison d'IBM à 22 cœurs, totalisant 200 petaFLOPS et consommant 13 MW.

Un lancement complet

Nvidia n'y est pas allé de main morte dans le lancement du 14 mai 2020 puisque le processeur A100 était annoncé accompagné de toute une panoplie de matériel :

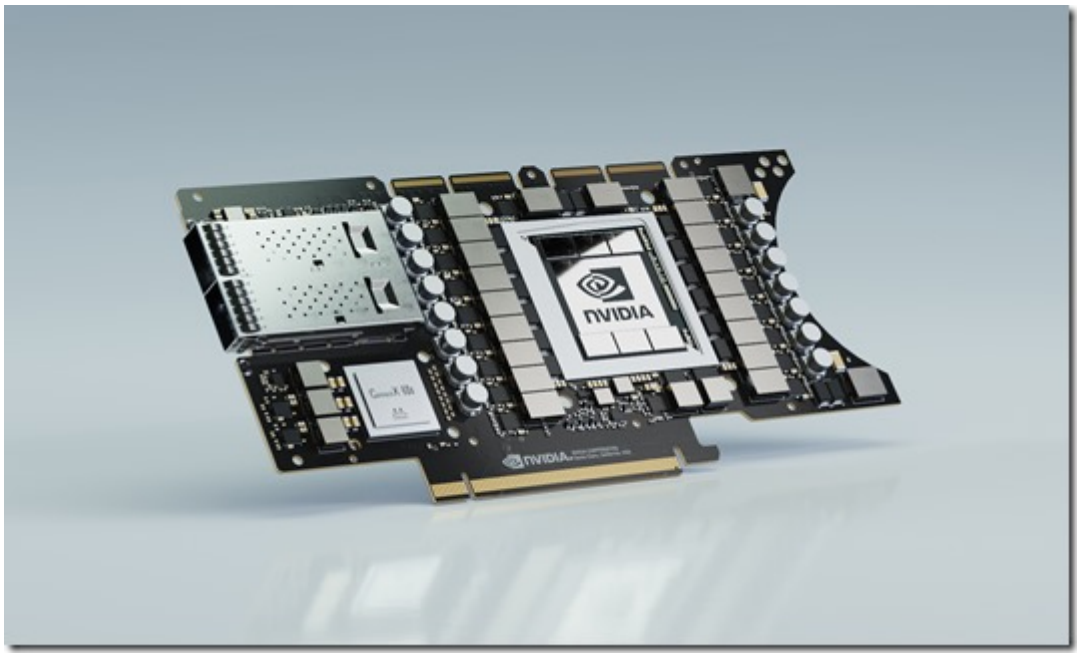
- Des serveurs Nvidia **DGX A100** de 5 petaflops (évalués en FP16) intégrant 8 GPU A100 complétés de 15 To de SSD, de neuf cartes réseaux Mellanox ConnectX-6 VPI de 200 GBits/s et de deux CPU AMD EPYC 7742 Rome, dont nous allons aussi parler plus loin et qui remplacent des processeurs Intel des précédents serveurs DGX. Notamment, parce qu'ils sont les premiers à supporter le bus PCI 4.0 qui est plus rapide que le 3.0, en particulier pour les accès SSD. Le DGX A100 est lancé à $200K ! Rappelons que l'israélien **Mellanox** a été acquis par Nvidia en 2019. Nous avons donc ici une intégration de son offre de connectivité réseau très haut débit dans l'ensemble de l'architecture de serveur Nvidia qui prend tout son sens. En plus, celle-ci est point à point avec de nombreux serveurs, permettant une répartition optimale des traitements en calcul distribué, un point clé pour le calcul haute performance.



- Des serveurs de partenaires constructeurs comme **Gigabyte** et **SuperMicro** qui reprennent la spécification du DGX A100.

- Le **supercalculateur DGX A100 SuperPOD**, comprenant jusqu'à 140 serveurs DGX A100 avec 1120 GPU A100 et 4 Po de stockage. Il faut la salle blanche dédiée qui va avec ! A raison de 4 DGX par rack, cela nécessite donc au minimum 35 racks.

- Une déclinaison du A100 pour l'embarqué avec la carte Jetson **EGX A100** qui a l'air de contenir un GPU A100 normal. Elle peut notamment servir à analyser les images provenant de dizaines de caméras de surveillance. Et par exemple, de compter la proportion des gens qui portent des masques dans les lieux publics surveillés ! Le covid revient par la fenêtre !



- Enfin, l'ensemble est accompagné d'une nouvelle version 11 de la **bibliothèque logicielle CUDA** qui est supporté par les principaux frameworks de deep learning du marché dont TensorFlow, MXNet et PyTorch.

Les détails sur le processeur sont disponibles dans **NVIDIA Ampere Architecture In-Depth** par Ronny Krashinsky, Olivier Giroux, Stephen Jones, Nick Stam et Sridhar Ramaswamy, mai 2020. Olivier Giroux est un canadien, "distinguished architect" et par ailleurs coordinateur d'un groupe de travail sur le parallélisme du calcul à l'ISO. Il est chez Nvidia depuis 2002.

Nvidia a aussi publié le livre blanc **NVIDIA A100 Tensor Core GPU Architecture** (84 pages) pour ceux d'entre vous qui souhaiteraient creuser tout cela.

C'est sur le processeur que je vais me focaliser ici même. Vous pouvez aussi visualiser la **vidéo du lancement** du CEO de Nvidia, Jensen Huang, tournée dans sa cuisine.

La gravure en 7 nm

L'un des sauts technologiques clés de ce processeur A100 est le passage d'une gravure de transistors de 12 nm pour le V100 de 2017 à 7 nm, ce qui a permis à surface presque égale de plus que doubler le nombre de transistors (on passe de 815 mm² à 826 mm²). Cette technologie 7 nm existe depuis plus d'un an. Elle est déjà utilisée pour produire les dernières générations de chipsets pour smartphones comme ceux de Qualcomm, HiSilicon (qui équipent les smartphones Huawei et Honor) et Samsung. Elle est aussi employée dans les dernières générations de CPU et GPU d'AMD.

On se retrouve avec un processeur de 54,4 milliards de transistors pour 21 milliards dans le V100. C'est un record pour une puce monolithique de cette catégorie. Il bat le record précédent du CPU AMD Zen 2 Epyc Rome 7H12 avec ses 39,54 milliards de transistors fabriqués en 7 nm chez TSMC et ses 64 cœurs généralistes. Mais celui-ci était un assemblage de neuf puces : huit en 7 nm qui contiennent les cœurs qui sont agencés autour d'une neuvième puce de gestion des entrées/sorties et est gravée en 14 nm pour des raisons économiques. La fréquence d'horloge de base de ce processeur AMD haut de gamme est de 2,60 GHz, extensible à 3,3 GHz, le tout avec une consommation (TDP) de 280 W. Sa mémoire cache est de 256 Mo. Et la bande passante mémoire est de 204,8 Go/s. Son prix de vente au lancement mi 2019 était supérieur à $7K.

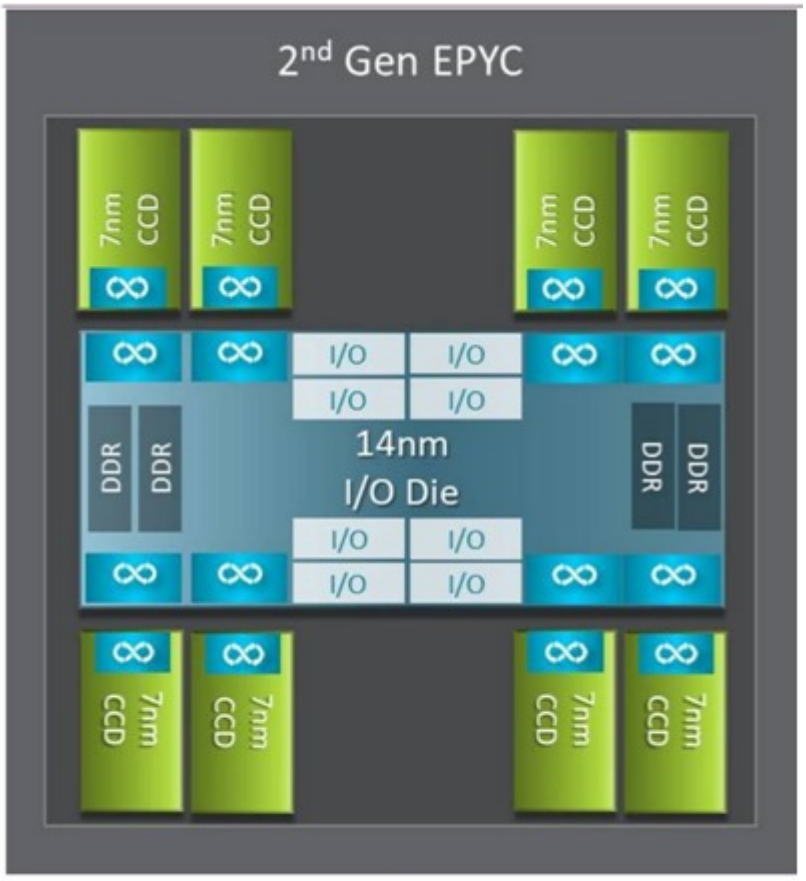


Figure 1. EPYC 7002 Hybrid Multi-Die package configuration.

Mais ce chipset AMD concurrence surtout les CPU Xeon d'Intel pour les serveurs généralistes de data-centers. Il ne joue pas spécifiquement dans la cours des serveurs de deep learning comme les GPU de Nvidia. Et on le retrouve donc dans le serveur DGX A100 en lieu et place d'un Intel Xeon.

Par comparaison, le A100 est fourni dans un package de 7 chipsets dont le processeur lui-même avec ses 54,4 milliards de transistor, entouré de 6 chipsets de mémoire d'accès rapide au standard HBM2E.

Architecture détaillée du GPU

Nvidia a conservé cette architecture hybride associant des unités de traitement arithmétiques classiques et des

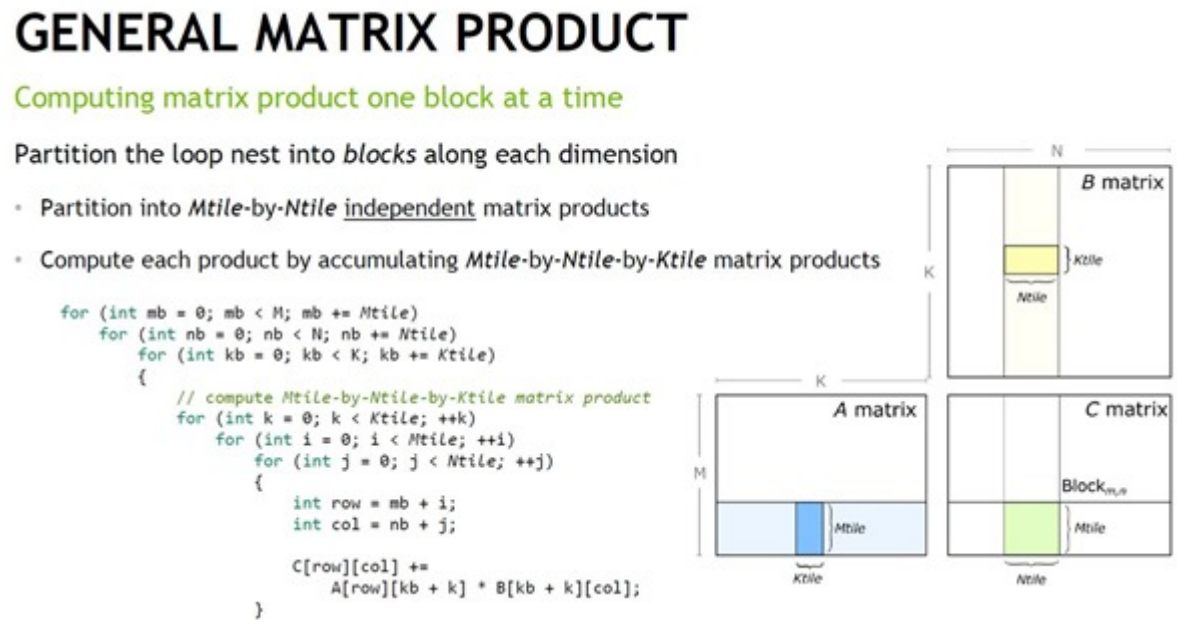
tensor cores. La plupart des concurrents qui visent le marché du deep learning ont tendance à ne retenir que les derniers ou tout du moins des architectures de calcul moins hétérogènes. Le choix de Nvidia lui permet sans doute de viser un marché large englobant à la fois les besoins du machine learning, du deep learning et du calcul haute performance. Il semble ainsi que les grands supercalculateurs qui intègrent des milliers de GPU V100 de 2017 comme le Jean Zay du GENCI à Orsay ou l'IBM Summit à l'Oark Ridge National Laboratory du Département de l'Energie aux USA servent surtout à du calcul haute performance.

Grâce à l'augmentation de la densité du processeur, le nombre d'unités de traitement est modifié. On passe ainsi de 13 440 à 20 480 unités de traitement arithmétique entre le GV100 et le GV100. Ces cœurs sont des ALU (arithmetic logic units) de trois types pour faire des calculs en nombres entiers 32 bits, flottants 32 et 64 bits. Dans la nomenclature Nvidia, seuls les ALU FP32 sont considérés comme étant des cœurs CUDA.

On passe ensuite de 732 à 512 tensor cores. Ces données comparent en fait les GPU GV100 et GA100 qui sont légèrement plus puissants que les V100 et A100. La différence entre ces deux versions des GPU est un peu *confusing* car il n'est pas évident de trouver comment la version "G" plus puissante est distribuée vu que les serveurs DGX n'en contiennent pas. Nvidia m'a confirmé que la version commerciale du GA100 n'existait pas encore. Voir **GPU Tensor Cores for fast Arithmetic Reductions** par Cristobal A. Navarro & al, 2020 (16 pages) qui décrit bien les principes généraux des tensor cores de la génération V100.

Est-ce une régression pour les tensor cores dont le nombre est en baisse ? Pas du tout. Les cœurs du V100/GV100 manipulaient des matrices de 4×4 alors que ceux du A100/GA100 ont l'air de pouvoir manipuler des matrices quatre fois plus grandes de 8×8 flottants.

J'avais lu en 2019 quelque part que Nvidia prévoyait d'augmenter la taille de ses matrices pour optimiser le traitement du deep learning. En effet, les modèles de deep learning qui sortent de bibliothèques comme TensorFlow de Google doivent convertir leurs calculs matriciels qui manipulent souvent des images de 227 ou 224 pixels de côté en calculs adaptés aux tailles de matrices gérées dans ces processeurs. Les filtres utilisés dans les convolutions des réseaux de neurones de reconnaissance d'image sont de taille variable, allant de 3 à environ 11 pixels. En supportant des matrices un peu plus grandes, le support de ces tailles variables de filtres est facilité et le parallélisme des traitements potentiellement amélioré. Avec le GPU V100, la bibliothèque CUDA comprenait déjà des opérations de multiplications de matrices 16×16, qui étaient donc converties en multiplications de matrices 4×4 pour l'exécution en parallèle dans plusieurs tensor cores du GPU. Ce genre de GPU permet en effet du calcul massivement parallèle ! (**source** du schéma *ci-dessous*)



La hiérarchie des unités de traitement du GA100 comprend 8 GPC (GPU processing clusters) qui contiennent chacun 8 TPC (texture processing clusters) qui contiennent chacun 2 SM (streaming multiprocessor). Ce qui fait un total de 128 streaming processors. Mais nuance, le processeur existe en deux versions : le GA100 avec 128 SM et 512 tensor cores et le A100 avec 108 SM et 432 tensor cores.

Autre différence dans les tensor cores : dans le V100 de 2017, la calcul matriciel côté multiplication se faisait sur un flottant 16 bits, et l'addition avec une troisième matrice du flottant 32 bits (schéma *ci-dessous*, **source**). Il se trouve que le "mixed precision deep learning training" est apprécié car il est très efficace en ressources machine utilisées.

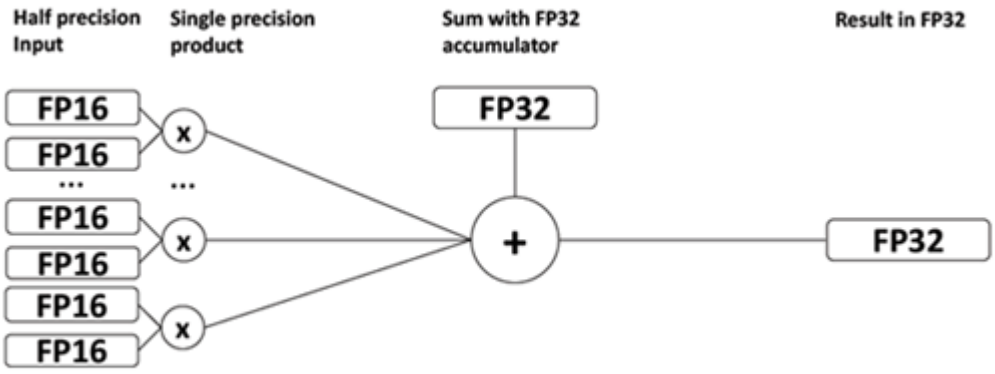
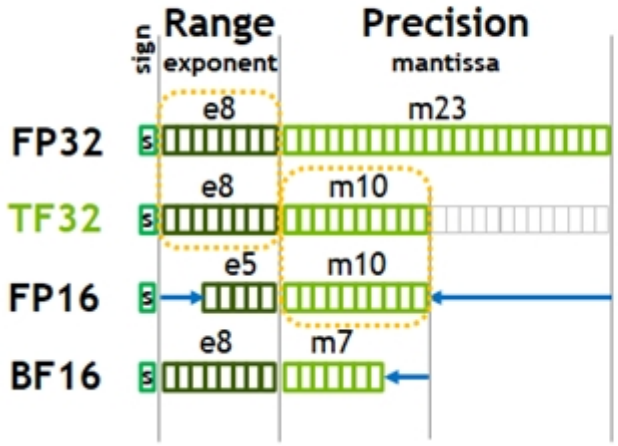
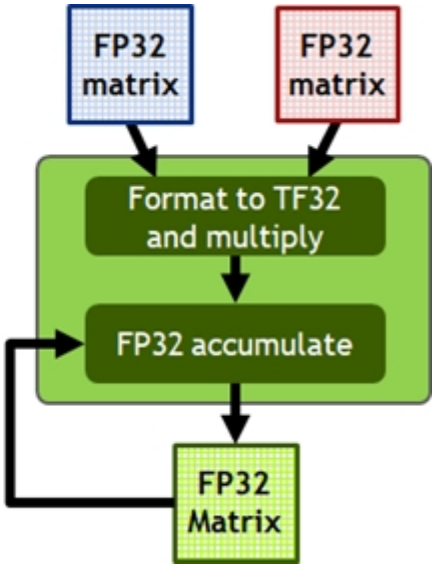


Fig. 3: FMAs in NVIDIA Tensor Cores.

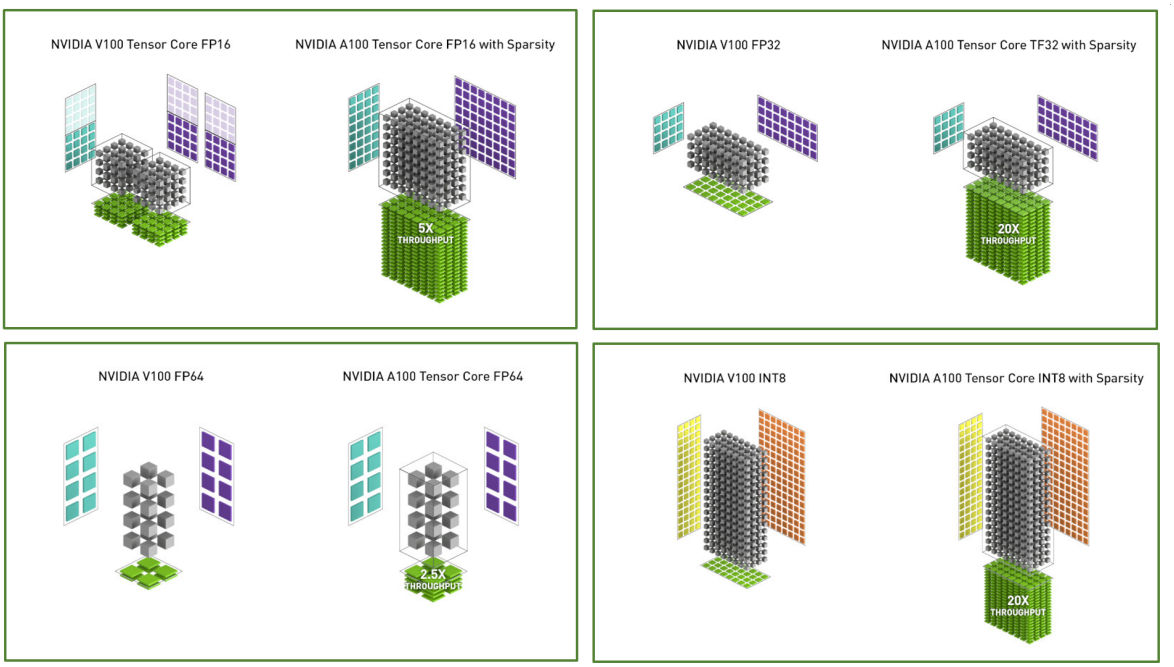
Dans le A100, la multiplication matricielle est aussi réalisable en nombre flottant 32 bits, ce qui peut améliorer la précision… si besoin est. Cela devrait être plus utile pour le calcul scientifique haute performance qui a généralement besoin de précision plus que pour l'entraînement de réseaux de neurones où l'on cherche à s'en passer.

Ces tenseurs font du FMA pour "fused multiply-add" : une multiplication entre deux matrices puis une addition avec une troisième. On passe d'une architecture qui associait deux modules gérant 64 opérations simultanément (4x4x4 fois deux, en haut à gauche du schéma ci-dessous) à un nouveau qui gère 256 opérations d'un coup, avec une flexibilité d'organisation qui est pilotée par logiciel via la bibliothèque CUDA 11.

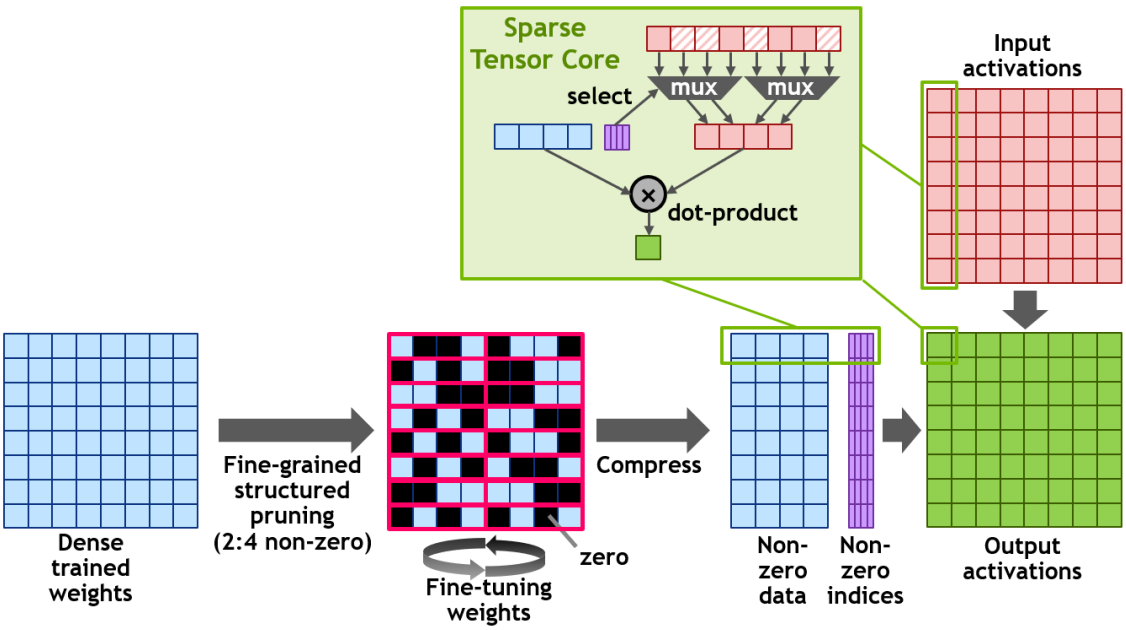




Le GPU A100 supporte un nouveau type de donnée, le TF32 pour Tensor Float 32. Il est exploité dans les multiplications de matrices avec la mantisse de 10 bits du FP16 (nombres après la virgule) et l'exposant 8 bits du FP32 (cf *ci-dessus*). Cela permet de faire des calculs avec une bonne précision et moins de charge que du calcul flottant 32 bits complet (FP32).



Ensuite, ils ont mis en place un système d'optimisation du calcul matriciel qui permet de gérer des "sparse matrix", des matrices partiellement remplies dites clairsemées ou matrices creuses qui sont courantes dans les réseaux de neurones et en particulier dans ceux qui traitent du langage. Tout cela génère des gains de performance variés décrits dans le schéma *ci-dessus* et expliqués *ci-dessous*.



Je me suis posé une question pratique de béotien : comment les fonctions mathématiques avancées comme les exponentielles, les logarithmes et les fonctions trigonométriques sont-elles mises en œuvre dans ces GPUs ? Je pense notamment à ces fonctions d'activation non linéaires de type sigmoïdes et tanh qui sont exploitées abondamment dans les réseaux de neurones après addition/multiplication des poids et valeurs des neurones.

Ces fonctions sont évidemment disponibles dans la bibliothèque logicielle CUDA de Nvidia qui est ensuite utilisée par les bibliothèques de gestion de neurones et par les frameworks de haut niveau comme TensorFlow de Google. A l'échelon proche du matériel, les fonctions mathématiques sont transformées en de longues séries d'opérations de base en calcul flottant (multiplications et additions). Ainsi, un cosinus est-il approximable par une série d'addition et de multiplications selon la formule ci-joint. On peut aussi gérer cela avec de simples tables de correspondance ("lookup table") qui sont encore plus économes en temps de calcul. D'autres nombreuses astuces peuvent être utilisées pour optimiser à la fois la performance et la consommation d'énergie. C'est particulièrement vrai dans le calcul embarqué mais est aussi réutilisé dans ce type de processeur.

$$\cos(\theta) = \sum_{n=0}^{10} \frac{(-1)^n \theta^{2n}}{2n!} \text{ for } -2\pi \leq \theta \leq 2\pi$$

Pour calculer une puissance (x^y), on peut passer par plusieurs étapes : un logarithme, une multiplication puis une exponentielle, qui à chaque fois, pour le log et l'exponentielle, vont faire appel à une évaluation polynomiale exploitant un tas d'opérations élémentaires ou à d'autres astuces. Chez Nvidia, la sauce secrète est située dans la bibliothèque CUDA qui fait le lien entre les primitives mathématiques et l'organisation des processeurs supportés. Le coût en calcul et en énergie rentre aussi en considération lorsque l'on utilise des calculs en nombres entiers pour les inférences des les smartphones et objets connectés.

Pour terminer dans les capacités du GPU A100, passons au côté mémoire et gestion des entrées/sorties :

- La mémoire vive intégrée évolue de 16 Go (2017) / 32 Go (2018) à 40 Go au standard HBM2E entre le GV100 et le GA100. C'est une augmentation modeste.
- Le débit de l'accès à la mémoire passe de 900 Go/s à 1555 Go/s grâce au standard HBM2E. Ces puces de mémoire HBM2E sont notamment proposées par les coréens SK Hynix et Samsung.

- On passe par contre de 6 à 40 Mo pour la mémoire cache L2 (intermédiaire, entre la mémoire HBM2E et le cache L1) ce qui est intéressant pour remplir plus rapidement les matrices de calcul.

- Le bus pour la carte comprenant le processeur passe au PCI Express 4.0 qui double la bande passante à 31,5 Go/s. pour une liaison x16 (que l'on peut agréger). Donc, cela double la vitesse d'accès aux supports de stockage SSD, ce qui permettra d'accélérer aussi de ce côté-là l'entraînement d'un réseau de neurones devant scanner une grande base d'entraînement.

- La liaison NVLink inter GPU/CPU supporte un débit 600 Go/s vs 300 Go/s avec le GV100. Elle est utilisée au sein de serveurs comme le HGX de Nvidia.

- Le GPU supporte aussi des interfaces réseaux allant jusqu'à 200 Gbits/s au standard InfiniBand, qui est supporté par les cartes Mellanox de Nvidia.

Il devient compliqué avec tout ça d'évaluer l'évolution de la performance du GA100 vis à vis du GV100 ou du A100 vis à vis du V100. Elle dépend du traitement à réaliser. Nvidia a choisi d'utiliser le benchmark BERT qui n'est pas, contrairement à l'habitude, un benchmark de traitement de l'image genre ResNet. BERT est un benchmark de deep learning pour du traitement du langage (Bidirectional Encoder Representations from Transformers) provenant de Google. Ils obtiennent dans ces conditions 3x plus de puissance pour l'entraînement et 7x plus pour de l'inférence en nombre de séquences par seconde. Avec d'autres benchmarks de calcul parallèle de type HPC (high-performance computing), ils obtiennent une accélération de d'environ 1,5x à 2x. Bref, à la fin, chacun fera son benchmark pour ses propres applications !

Une dernière fonction clé est celle de la **virtualisation** (MIG = *Multi-Instance GPU*) qui s'appuie sur un partitionnement du processeur sur 7 instances et est très utile pour partager les ressources du GPU pour des services en cloud.

Les deux principaux marchés visés pour le A100 sont le calcul haute performance et le machine learning/deep learning. Au bout du compte, on peut se demander si un tel GPU sert véritablement de GPU au sens "gaming" du terme. Il n'atterrira pas dans un PC de gamer, c'est certain. Des GPU spécifiques leurs sont dédiés, notamment de la famille RTX, qui donnent la part belle à des unités de traitement graphiques spécifiques pour le rendu des jeux vidéos, notamment pour le shading dans la génération d'images. Mais il n'est pas impossible que les A100 puissent tout de même servir à bâtir des offres de cloud gaming.

Énergie, transistors et concurrence

A pleine puissance ce processeur a une consommation maximale qui passe de 300 W (GV100) à 400W (GA100), avec une fréquence en légère baisse à 1,41 GHz vs 1,53 GHz, résultat d'un compromis pour limiter la consommation thermique. C'est en tout cas fait pour les serveurs refroidis par eau.

Avec une cote mal taillée, j'ai essayé de calculer le nombre de transistors par unité de traitement. En gros, cela nous fait 425 millions de transistor pour chacun des Streaming Multiprocessors. Si l'on considère que les unités arithmétiques utilisent 45% des transistors, cela nous fait plus d'un million de transistors par unité. C'est assez surprenant. Un million de transistors pour faire un calcul sur deux nombres flottants ? Cela semble énorme. Dans la littérature sur les processeurs, j'ai tout de même vu qu'il fallait au moins quelques dizaines de milliers de transistors pour de telles unités (ALU) [PS: un **papier d'Intel** qui a l'air de dater de 2010 évoque le fait que le calcul flottant dans un cœur classique occupe un million de transistors. Cela converge bien !].

Il n'est pas impossible que certaines fonctions mathématiques de haut niveau soient mises en œuvre directement dans le layout des portes logiques de ces ALU pour maximiser la performance et exploser quelques

benchmarks ! La loi de Moore permet d'en profiter ! Mais si vous avez le décompte précis du nombre de transistors par unité de ce gros GPU, je suis preneur ! Ce a l'air d'être un secret bien gardé.

Le Nvidia GA100 n'est pas le plus grand processeur dédié à l'IA puisque nous avons le monstre de **Cerebras** de 1,2 trillions de transistors, 400 000 unités de traitement (dont le détail n'est pas évident à récupérer) et 18 Go de mémoire cache. Il avait été annoncé en août 2019 et décrit dans **un article dédié sur ce site**. Il consomme 15 kW, presque autant qu'un ordinateur quantique à qubits supraconducteurs nécessitant un cryostat à 15 mK ! Dans ce dernier cas, 15 kW correspond surtout au coût énergétique du cryostat et en particulier de son gros compresseur.

Et tout cela n'est qu'une partie de l'actualité des processeurs dédiés à l'IA. ll faut notamment intégrer l'abandon progressif de l'architecture Nervana par Intel au profit de celle de Habana. Toutes les deux sont liées à des acquisitions de startups, l'une de 2016 et l'autre de 2019. Un train en cache donc un autre chez Intel qui a bien du mal à stabiliser sa stratégie dans le domaine de l'IA.

Pour terminer, j'aurais l'occasion de brosser un tour d'horizon de ce marché des processeurs dédiés à l'IA lors d'une réunion virtuelle organisée par le **Hub France IA** le 18 juin 2020 entre 18h et 20h (**lien d'inscription**). Nous y traiterons aussi du vaste champ des processeurs d'IA pour l'embarqué : smartphones et objets connectés, dans ce que l'on agrège habituellement et abusivement sous le vocable de Edge AI.

Cet article a été publié le 19 mai 2020 et édité en PDF le 18 mars 2024.