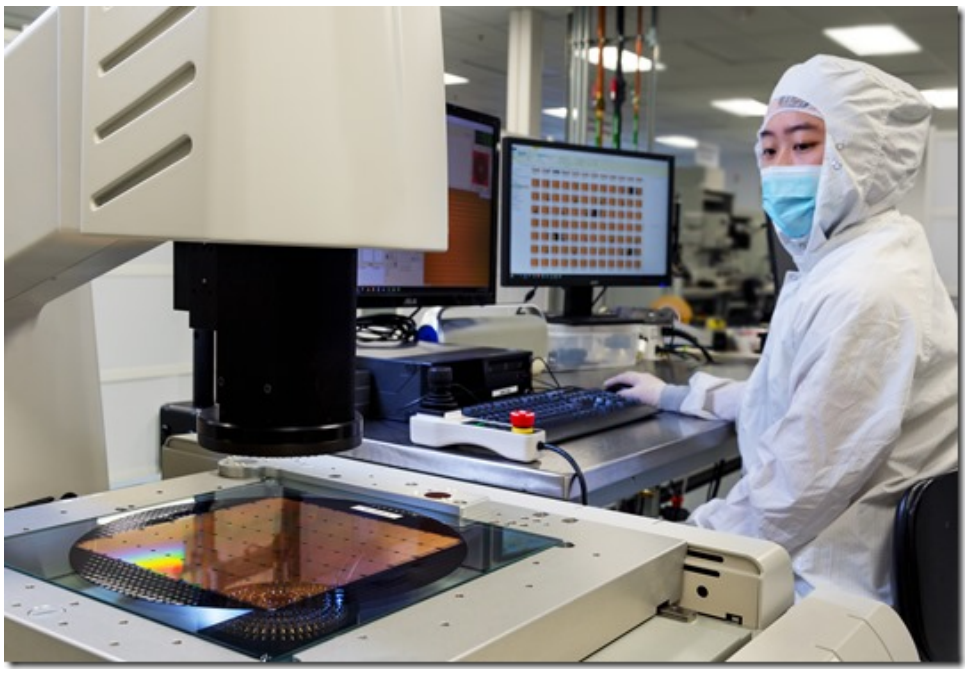
Opinions Libres
le blog d'Olivier Ezratty

# Le plus grand processeur d'IA

Après avoir traité avec grande légèreté le sujet du **management quantique**, je reviens aux choses sérieuses avec une actualité de l'IA qui mérite le détour, l'annonce de Cerebras de création du plus grand processeur d'IA (ou tout court) au monde.



Il est dédié à l'entraînement et à l'inférence de solutions de machine learning et de deep learning. Pour comprendre cette annonce, il faut revenir aux basiques. Le machine learning et sa forme la plus avancée, le deep learning requièrent des masses de calcul impressionnantes, surtout lors des phases d'entraînement des modèles. Ils sont à la base d'une bonne part actuelle des solutions de l'intelligence artificielle, que ce soit en particulier pour le traitement de l'image et celui du langage.

### Fonctions mathématiques d'un processeur d'IA

La principale voie choisie pour gérer ces calculs consiste à créer des processeurs qui exécutent aussi nativement que possible les fonctions mathématiques qui servent à gérer les réseaux de neurones du deep learning. Il faut d'abord entraîner ces réseaux de neurones avec de gros jeux de données comme des bases de millions d'images ou des tombereaux de textes, ce qui constitue la charge la plus lourde, puis les exécuter avec les données de production, ce qui consomme moins de ressources mais peut éventuellement être réalisé simultanément pour un très grand nombre d'utilisateurs.

Plusieurs types de calculs doivent être réalisés par ces processeurs :

- Des **multiplications de matrices et de vecteurs**, en particulier dans les réseaux de neurones convolutionnels de reconnaissance d'images. Ces calculs peuvent être éventuellement optimisés pour des vecteurs et matrices qui sont parfois remplis d'une grande quantité de zéros.
- La capacité à gérer des **fonctions non linéaires** comme les sigmoïdes des fonctions de normalisation dans les neurones, et si possible avec une grande précision, en nombres flottants, notamment lors de l'entraînement des réseaux de neurones.
- Le calcul de **fonctions dérivées** pour la gestion de la descente des gradients lors de l'entraînement de réseaux de neurones. Là encore, une grande précision est préférable pour les phases d'entraînement.
- Ces calculs sont plutôt réalisés en **nombres flottants** sur serveurs avec une grande précision pour l'entraînement et exécutés ensuite en nombres entiers dans les systèmes embarqués, comme les smartphones, pour économiser de l'énergie.
- Le tout doit être réalisé en faisant en sorte que l'**accès à la mémoire** qui contient les "*hyperparamètres*" des réseaux de neurones soit le plus rapide possible. Les architectures rivalisent donc pour rapprocher la mémoire des unités de calcul dans les chipsets du marché.

Une alternative consiste à gérer des **neurones logiciels** avec leurs entrées et sorties, fonctions de calcul et mémoire internes, ces neurones étant organisés en réseaux maillés interconnectés. C'est la forme la plus « pure » de chipset neuromorphique. Elle est notamment favorisée pour les systèmes fonctionnant par apprentissage par renforcement dans des systèmes embarqués.

**Les plus grands chipsets actuels**

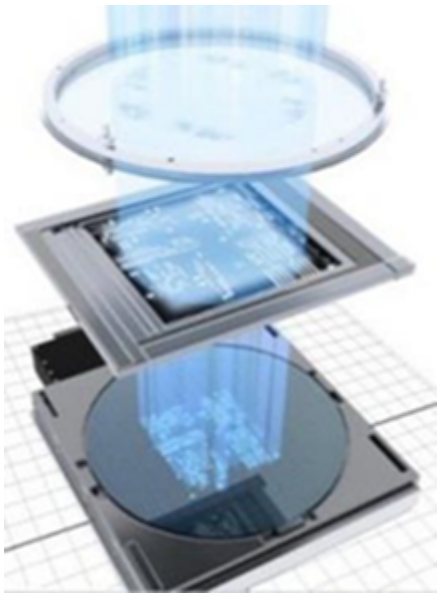
La diversité des architectures matérielles pour réaliser tout ou partie de cela est de plus en plus grande. Le nombre de startups lancées sur ce marché se compte en dizaines, en plus de grands acteurs tels que Nvidia et Intel. J'aurais l'occasion de faire le point dessus dans la prochaine édition de mon ebook **Les usages de l'intelligence artificielle**, édition 2019 d'ici la fin de l'automne.

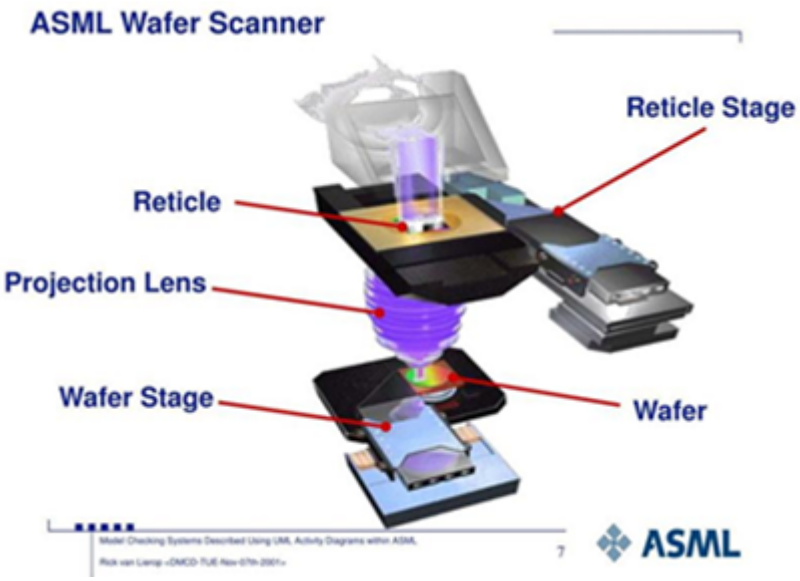
**Intel** mise notamment sur sa famille de chipsets Crest issue de l'acquisition de Nervana. Ce sont des processeurs dédiés à l'entraînement et/ou à l'inférence de réseaux de neurones selon les références. Ils s'appuient sur jusqu'à 24 unités de traitement dites TPC (Tensor Processing Cores) qui chacune comprennent deux gestionnaires de matrices 32×32.

A ce jour, le plus grand processeur du marché commercialisé est le GPU **Nvidia** V100 avec ses 21 milliards de transistors et gravé en technologie 12 nm. Lancé en 2017, il comprend 640 tenseurs qui sont des gestionnaires de matrices 4×4. Chaque tenseur est capable de multiplier deux matrices 4×4 et d'additionner le résultat avec une troisième matrice 4×4, pendant un cycle d'horloge du processeur, le tout en nombres flottants. Son successeur qui ne devrait pas trop tarder doublera probablement son nombre de transistors grâce au passage à une intégration en 7 nm. Le GPU V100 comprend en pratique plusieurs chipsets : le GPU proprement dit qui fait une surface de 815 $mm^2$ et est relié par des fils métalliques à des chipsets de mémoire RAM totalisant

32 Go de mémoire au standard HBM2 avec un débit de 900 Go/s. Ces derniers sont même empilés les uns sur les autres.

Plusieurs raisons techniques limitent la taille des chipsets de calcul. La principale est liée à l'optique des systèmes de lithogravure qui servent à "dessiner" les chemins des circuits et transistors sur le silicium. L'optique éclaire un réticule qui comprend un masque agrandi des dessins à graver. Il existe une limite de taille pour ces réticules. Je n'en connais pas l'origine précise. On pourrait dire que c'est comme ça et que toute la chaîne de production est calibrée comme cela. Mais ça peut aussi être lié à des contraintes optiques de ces systèmes qui fonctionnent dans l'ultra-violet, et, plus récemment, dans l'extrême ultra-violet (lumière UV à plus grande fréquence). Les illustrations ci-dessous issues du Hollandais **ASML**, le leader mondial de la lithogravure de semiconducteurs, permettent de comprendre cela.






L'autre limite vient du fait que plus un processeur est grand, plus les risques de défauts de gravure augmentent. Le taux de rebus peut devenir prohibitif, sauf pour toucher des marchés de niche à budgets quasiment illimités comme dans le militaire ou le spatial. Il existe pourtant des processus de fabrication de chipsets de plus grande taille, pour des capteurs photos. Mais en général, les systèmes de captation d'images les plus avancés, comme dans les télescopes, utilisent des matrices de capteurs CMOS ou CCD. J'avais approfondi le sujet dans l'ebook **De l'astronomie à l'entrepreneuriat** en août 2017. Il se trouve que les processus de fabrication des capteurs photos sont différents de ceux des processeurs et mémoire, du fait d'un niveau d'intégration plus faible, de l'ordre du micron au lieu du nanomètre.

### L'annonce de Cerebras

L'intérêt de l'annonce récente de la startup américaine **Cerebras Systems** (2016, USA, $112M) est de proposer un moyen de contourner ces limitations de taille et de proposer ce qui devient en pratique le plus grand processeur du monde d'un seul tenant. Cette startup a déjà atteint une valorisation de $860M, lui donnant un statut envié de pré-licorne. Elle a opéré pendant trois ans en mode silencieux ("*stealth mode*").

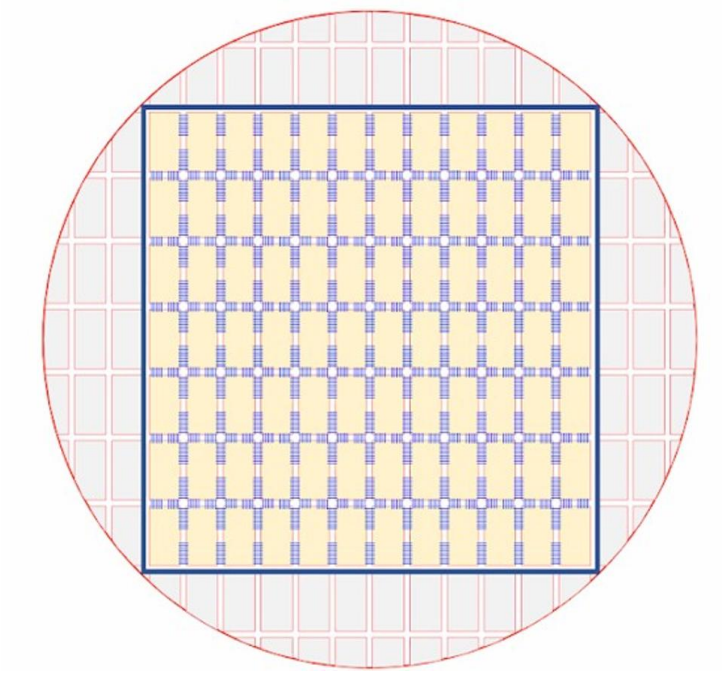
La startup a été créé par des anciens de **SeaMicro**, une startup constructeur de serveurs à basse consommation acquise par AMD en 2012 pour $357M, complétée récemment par un dirigeant d'Intel, Dhiraj Mallick. Lors de sa création, elle annonçait concevoir un chipset ASIC pour de

l'entraînement de réseaux de neurones qui optimise les calculs de matrices faiblement denses (avec beaucoup de zéros).

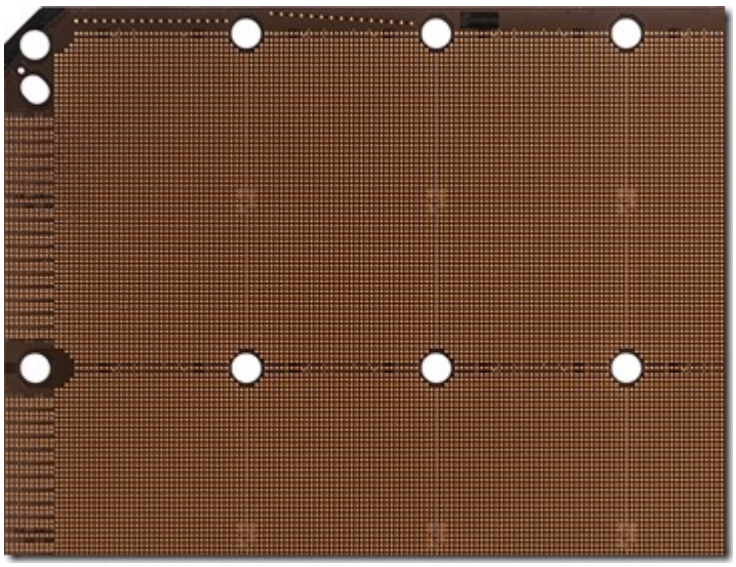
Un **ASIC** est un chipset fabriqué en volume, comme les processeurs Intel et Nvidia courants ou les processeurs polyvalents qui équipent vos smartphones. Ils s'opposent aux **FPGA** qui peuvent être reconfigurés dynamiquement par programmation mais sont moins optimums côté performance en termes de vitesse et de consommation d'énergie. En gros : ASIC = bonne performance mais coût fixe élevé et variable faible et FPGA = moins bonne performance, coût fixe faible et coût variable plus élevé.

En août 2019, Cerebras révélait ce qu'elle concevait en douce : le plus grand chipset au monde, le **Cerebras Wafer Scale Engine.** C'est en fait une matrice de 7*12 donc 84 chipsets gravés sur un wafer et reliés entre eux, totalisant 1,2 trillions de transistors, 18 Go de SRAM et 400 000 cœurs SLAC (Sparse Linear Algebra Cores) sur 8,5 pouces x 8,5 pouces ou 215 mm x 215 mm. Cerebras n'est pas très locace sur ce que font ces SLAC. Ils ont l'air de comprendre une petite unité programmable de calcul matriciel, mais ce n'est pas clair. Ce qu'ils en disent : "*The WSE contains 400,000 Sparse Linear Algebra (SLA) cores. Each core is flexible, programmable, and optimized for the computations that underpin most neural networks. Programmability ensures the cores can run all algorithms in the constantly changing machine learning field*". Et "*To achieve high performance, the SLA cores have a specialized tensor processing engine where full tensors are first-class operands in architecture. The tensor operations are programmable, so the same engine can be programmed to perform a variety of tensor operations such as convolution or matrix multiply. The hardware internally optimizes the tensor processing to achieve datapath utilization three of four times greater than graphics processing units*".

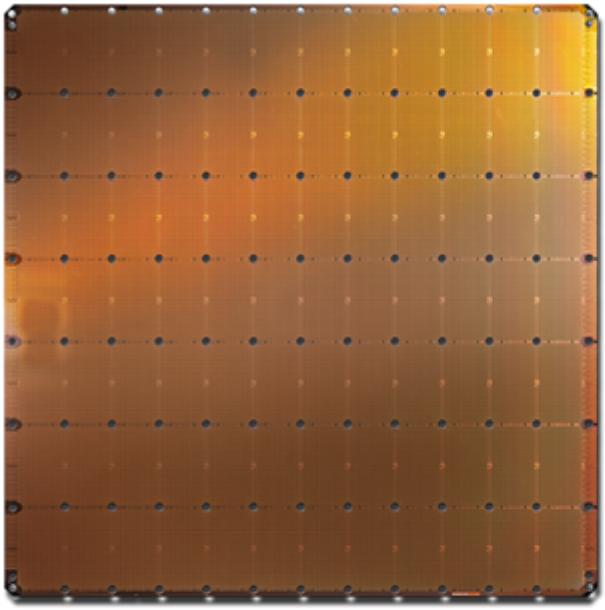
Le carré est le plus grand qu'il soit possible de générer dans un wafer de 30 cm de diamètre. Ces 18 Go de SRAM intégrée dans le composant sont à comparer aux 32 Go de mémoire HBM2 qui sont dans des chipsets externes dans le cas du processeur V100 de Nvidia. Sauf qu'ici, cette mémoire est très proche des unités de traitement et répartie dans les 400 000 cœurs. La latence d'accès sera donc bien meilleure, mais le nombre d'hyperparamètres et la complexité des réseaux de neurones gérés ne devrait donc pas augmenter significativement pour autant. Il faudra voir comment Cerebras ou d'autres optimiseront la répartition des traitements sur plusieurs processeurs de ce type et le type d'interconnexions entre chipsets qui seront possibles (NVLink ou autre).



On peut comparer sa taille *ci-dessous* avec le packaging d'un chipset Nvidia V100. C'est imposant ! La prouesse technique est liée à un partenariat avec le fondeur TSMC de Taïwan (qui est aussi le fondeur de Nvidia…). Ce dernier a dû mettre en place une batterie d'outillages industriels inédits pour fabriquer une telle bête. Ils ont notamment conçu conjointement une connectique de fils très courts reliant les chipsets et permettant une grande vitesse de transfert de l'information. Ils ajoutent pour cela une couche métal conductrice au-dessus des "scribe lines" qui relient traditionnellement les chipsets avant découpage.



Ils doivent aussi gérer le problème de la dilatation thermique d'une grande surface de silicium par rapport à son substrat. Ils utilisent pour cela des connecteurs à géométrie variable mais n'expliquent pas trop comment.



**Cerebras WSE**
1.2 Trillion transistors
46,225 mm² silicon

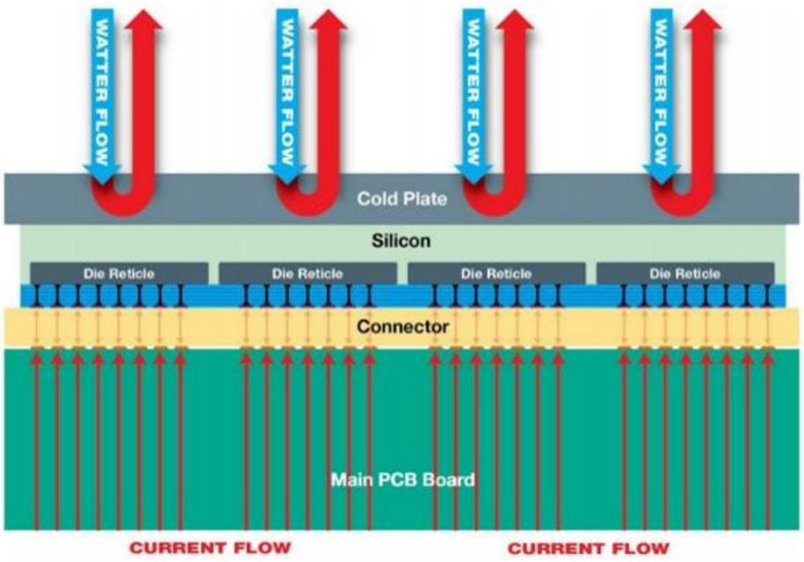
**Largest GPU**
21.1 Billion transistors
815 mm² silicon

Le second point est la gestion des défauts de gravure. Un chipset de la taille d'un wafer entier aura certainement des défauts. L'architecture de l'ensemble est conçue de telle manière que les quelques cœurs défectueux sur les 400 000 en tout sont désactivés et une connectique de remplacement permet de relier les cœurs qui sont autour en contournant "le malade".

Le troisième point clé d'un chipset géant est sa montée en température. Un chipset de type Nvidia V100 consomme plus de 300W, une grande partie étant transformée en chaleur à évacuer. Ici, chaque cœur consomme environ 175W, qu'il faut aussi évacuer, ce qui représente 15 kW. C'est au passage la consommation électrique d'un ordinateur quantique actuel comme les 2000Q de D-Wave, une bonne partie servant à faire fonctionner le système de cryogénie pour amener le processeur à une température très basse de 15 mK.

Pour le Cerebras, le refroidissement utilise un système à eau un peu particulier avec plusieurs radiateurs indépendants avec leur propre alimentation en eau couvrant différentes zones du processeur.

Un datacenter utilisant ce genre de composant sera probablement bien plus compact qu'avec des chipsets traditionnels d'Intel, Nvidia ou consorts. Ainsi, même s'il n'y a pas forcément équivalence, il faudrait environ 80 V100 pour créer une machine équivalente à un seul wafer Cerebras. Un serveur DGX-2 "10U" comprend 16 V100. On peut en mettre 4 par rack de datacenter. L'équivalent Nvidia V100 d'un Cerebras occupera ainsi 1,25 racks (qui font 42U de hauteur). Il est probable qu'un serveur utilisant un Cerebras fera au moins 5U. Mais il faudra également comparer la taille des systèmes de refroidissement qui sont associés.




Il subsiste au moins un point à éclaircir concernant ce processeur géant : sa connectique. Comment est-il relié au monde extérieur ? Par les milliers de pins qui inondent sa surface ? Cerebras n'a rien indiqué de ce côté-là. Il est probable qu'un serveur l'exploitant devra aussi comprendre un processeur classique, genre Intel Xeon Phi, pour pouvoir piloter l'ensemble.

Maintenant que le wafer a été prototypé, il reste à le fabriquer en série et à un coût raisonnable, qui n'a pas encore été communiqué. Mais pas que. Autour, il faut concevoir un packaging complet et un ordinateur avec, puis le système de refroidissement à eau et enfin, créer les briques logicielles qui feront le lien entre le processeur et les principaux SDK de deep learning du marché, TensorFlow et premier. Le SDK optimise le cheminement des données entre les unités de traitement. C'est un enjeu classique d'écosystème d'outils de développement. La soif de puissance des applications du deep learning n'a pas de limites ! Reste aussi à savoir si ce processeur permettra de créer des applications d'IA difficiles à faire autrement. Il est également probable que ce genre d'architecture aura d'autres usages que ceux de l'IA. Dans les supercalculateurs, on fait aussi des calculs par éléments finis qui nécessitent beaucoup de puissance de calcul parallèle et d'accès rapides à de gros volumes de données. Ils pourraient s'inspirer de ce genre d'architecture.

Pour plus de détails, voir **AI Startup Cerebras Develops The Most Powerful Processor In The World** par Kim McGregor dans Forbes ainsi que  le livre blanc de Cerebras : **Cerebras Wafer Scale Engine: An Introduction**, 9 pages.

**L'architecture ultime ?**

L'annonce de Cerebras est significative et pourrait changer la donne du traitement de l'IA dans les data-center. Elle est uniquement tournée vers les ressources du cloud. D'autres startups ou acteurs établis investissent ce domaine avec des approches différentes, en plus d'Intel et Nvidia que nous avons déjà cités :

Côté serveurs, nous avons par exemple :

- **Graphcore** (2016, UK, $310M) conçoit son Intelligence processuing Units (IPU), un chipset adapté à l'exécution d'applications de deep learning côté entraînement et inférence qui comprendrait 1000 cœurs. Ils ciblent notamment le marché autonomobile.

- **Huawei** qui lançait en août 2019 son processeur pour entraînement d'IA sur serveur Ascend 910 doté de 256 TFLOPS et 512 TOPS sur des entiers 8 bits avec une consommation maximale de 310W.

- **Groq** (2017, USA, $62,3M) créée par des anciens de Google qui avaient participé à la conception de leurs TPU. Leur chipset pour serveur est censé générer 400 TFLOPS avec 8 TFLOPS par Watt.

- **Gyrfalcon Technology Inc** ou GTI (2017, USA) est sorti du bois en septembre 2017 avec deux chipsets d'inférences ASIC à basse consommation, l'un pour les serveurs et l'autre pour les objets connectés. La version serveur (Lightspeeur 280x AI Accelerator) est intégrée dans des cartes à 16 composants. Le Lightspeeur 2802M, un ASIC intègre de la mémoire MRAM non volatile (Magnetic RAM) produite en technologie 22 nm chez TSMC.

- **Habana Labs** (2016, Israël, $120M) commercialise une carte PCIe pour serveurs comprenant leur processeur Goya HL-1000 qui peut traiter 15 000 images/second avec seulement 100W contre 3211 pour un Nvidia V100 et 320W. Le tout grâce à l'intégration d'un multiplicateur de matrices, le GEMM (General Matrix et Matrix Multiplication) qui s'appuie sur huit cœurs tensoriels exploitant leur mémoire locale et qui supportent le calcul sur nombres flottants et entiers entre 8 et 32 bits. Le système supporte TensorFlow et le format d'échange ONNX.

- **Wave Computing** (2010, USA, $203M) développe ses Dataflow Processing Units avec 16 000 cœurs produits en ASIC chez TSMC en 16 nm, dédiés à l'entraînement de réseaux de neurones. Ces DPU sont assemblés dans des serveurs par paquets de 16, donnant 128 000 cœurs. Ils n'utilisent par contre que de la DRAM, bien moins performante que la mémoire HBM2 des GPU Nvidia. La startup avait fait l'acquisition de l'activité MIPS de l'Anglais Imagination Technologies.

Et dans l'embarqué, il y a encore plus de monde avec les chipsets de smartphones de Qualcomm, Samsung, Apple et HiSilicon (Huawei) qui intègrent des NPU (Neural Processing Units) comprenant quelques dizaines au plus de petits multiplicateurs de matrices. Et d'autres acteurs tels que les Français Kalray et AnotherBrain, le Chinois Horizon Robotics, puis Vathys, Hailo, Synthiant, Novumind, DeePhi Tech, GrAI Matter Labs, Eta Compute, Cornami et Cognimem. Leur ambition n'est généralement pas de créer "le plus grand processeur d'IA au monde", mais plutôt des chipsets à très basse consommation.

L'indicateur important dans ce marché est le nombre de TOPS (tera-operations per seconds, en nombres entiers) par Watt consommé. Il est situé entre 1 et 3 selon les cas. D'autres font en sorte qu'il soit possible d'utiliser ces chipsets pour réaliser de l'apprentissage par renforcement avec peu de données. Et nous avons aussi les tentatives de création de memristors et les chipsets utilisant ce que l'on appelle des neurones à impulsion. Le neurone reçoit un train d'impulsions dans l'ensemble de ses synapses et génère en sortie un train d'impulsions résultat du calcul. Ces spiking neurones sont difficiles à programmer. C'est la voie choisie par IBM avec ses chipsets TrueNorth ainsi que par Intel avec ses chipsets Loihi.

Bref, ce marché des chipsets pour l'IA a encore du mou sous la pédale et il est passionnant de l'observer de près !

Cet article a été publié le 27 août 2019 et édité en PDF le 2 septembre 2019.