IA : Intel passe à l'offensive sur le Deep Learning

Après être passé au travers de la révolution mobile, Intel multiplie les initiatives – et les rachats – pour éviter de se trouer une nouvelle fois dans la mutation naissante vers l'IA. Pour ce faire, le géant du silicium disposera, sous peu, d'un portefeuille de produits très large. Avec, comme porte-étendard, de nouvelles solutions pour le Deep Learning issues du rachat de Nervana, acquisition annoncée en août dernier. Rappelons qu'Intel y avait mis le prix fort, déboursant pas moins de 400 millions de dollars environ pour cette start-up employant moins de 50 personnes.

Aujourd'hui directeur de l'activité solutions IA dans son ensemble, le co-fondateur de Nervana, Naveen Rao, a détaillé, lors d'une conférence à Paris, l'extension de l'offre d'Intel avec les solutions issues de la recherche de son entreprise. « *L'IA est comme un bulldozer pour la donnée, il permet d'en manipuler de quantités paraissant jusqu'alors hors de portée* », explique ce scientifique formé à la fois à la science de la donnée et aux neurosciences. Et de rappeler la place du Deep Learning (ou apprentissage profond) dans la prise de conscience du rôle que peut jouer l'intelligence artificielle. « *En 2012, le test Imagenet, consistant à catégoriser des millions d'images, a accéléré cette prise de conscience*, explique le nouveau dirigeant d'Intel (sa société a été officiellement absorbée par le groupe en octobre dernier). *Avant l'utilisation du Deep Learning, le taux d'erreur de la machine dans la reconnaissance de ces images atteignait 26 %, là où un humain entraîné parvenait à descendre à 5 % seulement. Après utilisation du Deep Learning, la proportion d'erreurs est brutalement descendue à 16 % ; elle est même aujourd'hui passée sous les 3 % !* »

Le Deep Learning au cœur même des Xeon

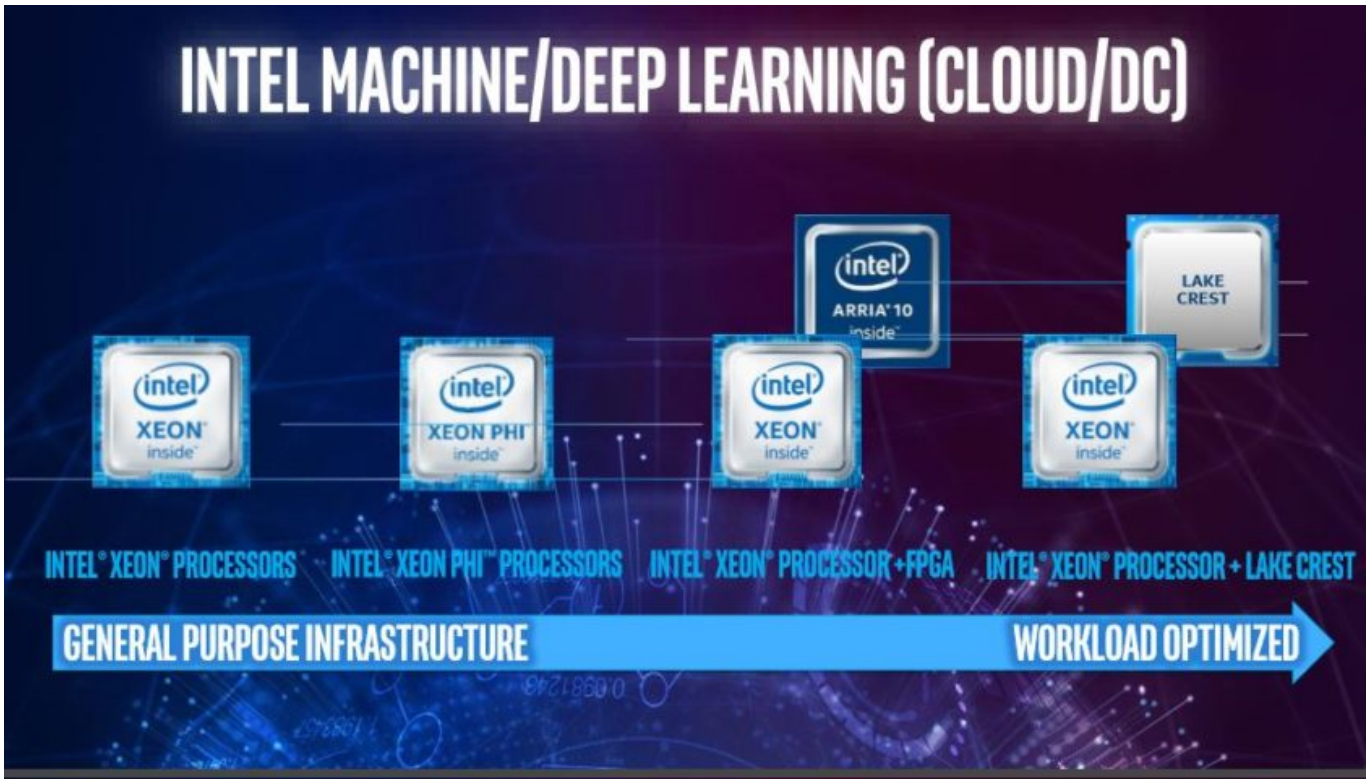
Si le Deep Learning reste minoritaire dans l'IA – il serait employé sur 37 % des serveurs mettant en œuvre une forme ou une autre d'intelligence artificielle -, il connaît une croissance significative : + 66 % par an selon des chiffres présentés par Intel. Un marché aujourd'hui largement servi par les GPU, fournis par AMD et surtout Nvidia. « *Mais ce sont des puces qui ont été reconfigurées pour ces besoins, là où Nervana a développé son silicium d'entrée de jeu pour le Deep Learning. Ce qui permet d'arriver à des performances supérieures* », assure Naveen Rao.

L'ambition de Nervana – qui revient à « *extraire les principes de fonctionnement du cerveau pour les appliquer dans le silicium* », doit trouver sa première concrétisation en 2017, avec la sortie d'un accélérateur fonctionnant de concert avec les Xeon d'Intel : Lake Crest. Ce co-processeur, pensé pour doper la parallèlisation des opérations, sera cadencé à 2 GHz et disposera de 32 Go de mémoire. Lake Crest, dont les premiers exemplaires de tests doivent sortir au début de cette année, promet aussi un accès ultra-rapide à la mémoire, avec des débits allant jusqu'à 8 Tbit/s. La puce intègrera aussi 12 liens à haute vitesse afin de constituer des architectures en grappes. Ces liens sont annoncés comme 20 fois plus véloces que le PCI Express. Ce premier développement sera prolongé par l'intégration des technologies de Nervana au sein même du silicium des Xeon, un projet qu'Intel appelle Knights Crest. Si l'agenda de sortie de ce Xeon dopé au Deep Learning n'a pas été détaillé pour l'instant, Intel promet, avec les technologies issues de Nervana, une

accélération par 100 des performances actuelles des applications d'apprentissage profond d'ici à 2020.

Aux côtés des Xeon Phi et des FPGA

« *Nous travaillons à la convergence entre les puces d'usage général et les accélérateurs dédiés*, confirme Naveen Rao. *Avec Knights Crest, nous cherchons à amener ces fonctions au plus proche du cœur, afin de diminuer la complexité qu'amène la prise en compte des accélérateurs. Sans oublier les pertes de performances qui découlent des liens PCI Express, via lesquels ils sont raccordés.* »



L'arrivée de Lake Crest puis de Knights Crest viendra compléter le portefeuille d'Intel, qui compte déjà quelques puces adaptées à l'IA. A commencer par les co-processeurs Xeon Phi, une puce massivement multi-cœur adaptée aux « *usages standards en matière de Machine Learning* », selon Naveen Rao. Le Xeon Phi doit prochainement bénéficier d'une mise à jour technologique, avec la sortie de la génération Knights Mill, qui promet des performances multipliées par 4 selon Intel. Enfin, depuis le rachat d'Altera en 2015, le géant de Santa Clara dispose d'une offre solide en matière de puces reprogrammables (FPGA), « *très pertinentes pour les inférences* ». Avec un certain succès, puisque Microsoft a retenu cette architecture pour accélérer certaines tâches dans les datacenters de son Cloud Azure et que le Français OVH mise sur ces puces pour participer à la lutte contre les attaques par déni de service (DDoS).

A lire aussi :

Quand le Deep Learning permet aux aveugles d'écouter leur environnement

IA : on ne joue plus… ou alors à se faire peur