FPGA : l'arme secrète spéciale IA du Cloud Azure de Microsoft

Lors de sa récente conférence Build, qui se tenait du 10 au 12 mai, Microsoft a confirmé qu'il entendait élargir la place laissée aux FPGA, ces puces reprogrammables que le premier éditeur mondial a récemment greffées à son Cloud Azure. Et qu'il entend bien prochainement proposer comme un service à ses clients.

Rappelons que l'arrivée de ces puces reprogrammables dans l'architecture d'Azure résulte d'un projet (dénommé Catapult) initié dès 2010 par un chercheur maison appelé Doug Burger. Un projet visant, à l'origine, à étudier le potentiel de ces composants pour accélérer l'indexation sur Bing, le moteur de recherche de Microsoft. Des travaux qui ont donné naissance à un premier design, implémenté pour Bing et Azure, utilisant des grappes de 48 FPGA regroupées dans un rack relié à un réseau secondaire. Puis à son évolution, dévoilée en octobre 2016 par Redmond.

FPGA agrégés en un pool de ressources

Ce Catapult v2, implémenté dans tous les nouveaux datacenters Azure, peut à la fois servir à accélérer des tâches locales – comme des requêtes Bing -, mais aussi à réaliser des tâches plus consommatrices de ressources, via l'agrégation de FPGA, désormais vus comme un pool de ressources à part entière. La conséquence ? Le ratio entre FPGA et CPU sur une application donnée peut être décidé en fonction des besoins réels de ladite application, et non plus de la configuration matérielle des serveurs. Une architecture qui offre de nouvelles perspectives, en particulier dans l'IA.

En septembre 2016, lors de sa conférence Ignite, le premier éditeur mondial expliquait que ce design lui amenait une puissance de l'ordre de 10 puissance 18 dans le calcul d'inférences sur des réseaux de neurones profonds. La firme explique aujourd'hui avoir encore dopé cette capacité, employée pour les services cognitifs maison. On retrouve ici des fonctionnalités de traduction, de reconnaissance vocale ou de reconnaissance visuelle offertes par Skype, Cortana ou Bing. Les FPGA sont également employés pour accélérer le réseau sur Azure et pour bâtir les index de Bing.

Pour le Deep Learning

Mais, comme l'a expliqué Mark Russinovich, le directeur technique d'Azure lors de Build, l'objectif est désormais de mettre ce pool de ressources, qui grossit peu à peu à mesure que Microsoft met à jour les serveurs de ses datacenters, à disposition des clients du Cloud de Redmond. En particulier, Microsoft envisage de permettre aux entreprises d'installer leurs propres réseaux de neurones sur cette architecture pour des applications de Deep Learning. Russinovich n'a toutefois pas indiqué quand l'éditeur comptait ouvrir ce service.

Au-delà de ce premier service commercial, Microsoft travaille aussi à confier de nouvelles tâches à ses puces reprogrammables. En particulier l'entraînement des réseaux de neurones, un

apprentissage qui mobilise d'importantes capacités de calcul. La tâche revient aujourd'hui aux GPU de Nvidia au sein d'Azure. Les tests menés par Microsoft, et décrits par Mark Russinovich montrent la capacité à employer 6 FPGA pour soutenir les besoins de 22 threads issus de réseaux neuronaux tournant sur des CPU classiques, pour un gain d'efficacité de 73 %. C'est ce modèle, dans lequel les FPGA sont agregés indépendamment du reste de l'architecture Cloud pour créer une nouvelle couche de services dans Azure, que Redmond entend généraliser. La démarche ressemble à celle de Google, avec ses puces dédiées à l'IA, les TPU (Tensor Processing Units).

FPGA : le bon compromis performances/flexibilité ?

Dans l'optique de Microsoft, les FPGA occupent une place intéressante entre les CPU ou GPU et les Asic, offrant un bon compromis entre flexibilité et performances. Contrairement aux Asic, ils peuvent être reprogrammés. Et si les FPGA sont plus complexes à programmer que des processeurs standards, en particulier dans des architectures hyperscale, Microsoft semble trouver que le jeu en vaut la chandelle. La firme assure constater, sur des calculs d'inférences, une multiplication par un facteur 150 à 200 du débit de données et une amélioration par un facteur 50 de l'efficacité énergétique par rapport à un CPU classique.

Au passage, la démarche de Microsoft et celle de Google (qui a conçu lui-même sa TPU) constitue plus qu'un avertissement pour Intel, mais aussi pour les concepteurs de GPU comme Nvidia. La taille des architectures hyperscale permet aux géants du Cloud de s'orienter vers des design spécifiques, au besoin conçus par leurs soins. Même si, pour l'instant, Microsoft commande ses FPGA chez Altera, une société passée dans le giron d'Intel.

A lire aussi

Google muscle sa puce TPU dédiée au Machine Learning

Pour son Cloud, Microsoft met ARM en concurrence avec le x86

Cloud : comment Microsoft repense l'architecture de ses datacenters Azure