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

Quand les futurs datacenters Azure ouvriront leurs portes quelque part en 2017, il comportera en son sein un composant peu courant dans les centres de données : des FPGA, des puces reprogrammables utilisées pour accélérer certains traitements. Le recours à ces auxiliaires n'est pas nouveau chez Microsoft : le premier éditeur mondial s'est intéressé très tôt à leur emploi dans des configurations Cloud. C'est ce que Redmond a appelé le projet Catapult, lancé par un de ses chercheurs, Doug Burger. D'abord voués à l'accélération des recherches sur le moteur de recherche maison Bing, les FPGA sont toutefois en passe de prendre une nouvelle dimension dans les datacenters Microsoft.

Dans un article de recherche, une vingtaine de chercheurs de Microsoft décrivent comment les puces reprogrammables s'intègrent dans le design des nouveaux datacenters Azure, un projet qu'ils baptisent Captapult v2. Plutôt que de déployer les FPGA par grappes de 48 unités regroupées dans un rack relié à un réseau secondaire – le choix retenu lors de Catapult v1 -, les équipes d'Azure optent cette fois pour une architecture bien plus ambitieuse. Celle-ci couple les accélérateurs directement au réseau des datacenters, en plaçant les FPGA entre les serveurs et les commutateurs Ethernet. Bref, plutôt que de reléguer les puces reprogrammables à des tâches subalternes qui lui sont envoyées par d'autres composants, elles sont désormais placées en amont des serveurs, et voient arriver les messages qui leur sont destinés. Les FPGA peuvent ainsi prendre des décisions sur la façon de gérer ces messages et réaliser des tâches, sans même impliquer les processeurs principaux. « *Ce que nous avons fait, c'est transformer les FPGA en porte d'entrée* », résume Derek Chiou, qui dirige l'équipe chargé de l'accélération du Cloud Azure.

Haas : transformer les FPGA en pool de ressources

Avec ce design, où chaque FPGA est aussi relié à un serveur via PCIe, les puces reprogrammables peuvent à la fois servir à accélérer des tâches locales – comme des requêtes Bing -, mais peuvent être également agrégées pour réaliser des tâches plus consommatrices de ressources. C'est ce que Microsoft appelle le Haas (Hardware-as-a-service). « *En permettant aux FPGA de parler directement aux commutateurs réseau, chaque FPGA peut communiquer directement avec tout autre FPGA, au sein du datacenter ou sur le réseau, sans aucune intervention CPU*, écrivent les chercheurs dans leur article de recherche. *Cette flexibilité permet d'agréger des groupes de FPGA dans des pools de ressources.* » Pour ce faire, Microsoft emploie un protocole de communication inter-FPGA (dénommé LTL pour Lightweight Transport Layer) qui offre des temps de latence « *comparable à l'état de l'art* », assure Microsoft. La conséquence est majeure pour un prestataire de Cloud : « *les services sont libérés de la contrainte imposant un ratio fixe de coeurs de CPU par FPGA et peuvent donc allouer (ou acheter, dans le cas du Iaas) uniquement les ressources de chaque type dont ils ont besoin* », résume l'étude.

L'avantage de cette architecture, que Redmond baptise le Configurable Cloud ? Elle offre une justification économique au déploiement en masse d'accélérateurs. Comme l'écrivent les chercheurs, « *sortir du ratio 1 pour 1 entre CPU et FPGA est essentiel pour casser le syndrome de la poule et*

de l'oeuf, dans lequel les accélérateurs ne peuvent être ajoutés (aux architectures des datacenters, NDLR) faute d'un nombre suffisant d'applications les employant et dans lequel les applications ne vont pas exploiter les FPGA tant qu'ils ne sont pas présents massivement dans les architectures. »

Un supercalculateur indépendant

Dans son article de recherche, sur la base de tests sur 5 760 serveurs, Microsoft explique que le design peut s'adapter à une multitude d'usages : soutien aux CPU (par exemple pour les recherches Bing, avec des performances multipliées par plus de deux), accélération réseau (par exemple, chiffrement/déchiffrement en amont du serveur, ce qui permet d'économiser 5 coeurs de processeur avec le protocole AES 128 bits en 40 Gbit/s), mais aussi service d'accélération disponible à l'échelle d'Azure. Microsoft teste l'emploi de ce concept de nouveau avec l'accélération des recherches Bing. Selon les résultats publiés, la flexibilité offerte par la mise en réseau des accélérateurs ne dégrade pas les temps de latence : ceux-ci sont similaires aux performances obtenues avec le design de Catapult v1. « *Ainsi, les serveurs dotés de FPGA peuvent donner leur ressource d'accélération en toute sécurité à un pool de ressources avec un impact minimal sur les performances logicielles* », écrivent les chercheurs. Qui précisent tout de même que LTL est associé à des limitations de bande passante pour éviter de voir les accélérateurs ralentir le réseau.

Au-delà de cette comparaison entre les deux design Catapult sur l'accélération Bing, la nouvelle architecture ouvre donc la voie à une multiplication des usages des FPGA au sein des datacenters Azure. « *Ce design transforme les FPGA distribués dans les datacenters Azure en un supercalculateur indépendant* », tranchent les chercheurs.

Amener de la flexibilité dans les datacenters

Ainsi, avant la fin de 2016, Microsoft prévoit de déployer des réseaux de neurones profonds sur ses accélérateurs Catapult afin d'améliorer les résultats de recherche sur Bing. « *Avec, pour conséquence, des résultats plus pertinents* », promet Sitaram Lanka, un des auteurs de l'étude sur Catapult v2. Et nul doute que Microsoft n'entend pas se cantonner à des usages internes. Lors de son passage à Paris, début octobre, Satya Nadella a présenté Azure comme le premier supercalculateur pour l'IA dans le Cloud. En faisant un lien direct avec les capacités des accélérateurs Catapult.

Plus largement, la généralisation des FPGA permet à un acteur du Cloud, qui investit des centaines de millions dans son réseau de datacenters, de s'offrir un minimum de flexibilité vis-à-vis des vagues technologiques. Il y a six ans, par exemple, bien peu nombreux auraient été les experts capables de prédire le rôle central que jouerait le Deep Learning dans de nombreuses applications analytiques. Des chocs que les FPGA permettent d'encaisser. En reprogrammant ces composants, un prestataire peut proposer des services adaptés sans avoir à en passer par du logiciel, structurellement moins efficace, ou sans avoir à attendre la disponibilité d'une puce spécifiquement pensée pour ces nouveaux besoins.

Basé sur des cartes Stratix V D5 fournies par Altera (une société aujourd'hui dans le giron d'Intel), le design Catapult v2 est actuellement déployé dans les datacenters Microsoft situé dans 15 pays dans le monde, signalent les chercheurs dans leur article. Sans toutefois donner davantage de précisions. Environ 150 000 FPGA seraient toutefois déjà à l'oeuvre dans les datacenters Azure (hors de toute accélération Bing donc) avec l'architecture Catapult v1, un design appelé à être supplanté par la nouvelle version au gré du remplacement des serveurs et du lancement des nouveaux datacenters Azure. Par ailleurs, Redmond utilise également les puces reprogrammables dans les routeurs employés sur son Cloud, des systèmes maison au sein desquels les FPGA apportent leur flexibilité. Une fois de plus.

A lire aussi :

Satya Nadella : « Azure est le premier supercalculateur pour l'IA »

FPGA : l'arme secrète d'OVH pour parer les attaques DDoS