Charges de travail d'apprentissage en profondeur TensorFlow : un coût d'exécution plus avantageux sur les serveurs bare metal d'IBM Cloud que sur Amazon Web Services

Les serveurs bare metal d'IBM Cloud équipés de GPU ont obtenu un nombre plus élevé de trames par seconde (FPS) par dollar sur 5 modèles d'apprentissage automatique TensorFlow.

Les entreprises sont de plus en plus nombreuses à choisir l'apprentissage automatique, un sous-domaine restreint de l'IA, pour tenter de transformer la masse apparemment infinie de leurs données en une ressource exploitable. L'apprentissage automatique (Machine Learning), et en particulier l'apprentissage en profondeur (Deep Learning), résout ces problèmes rapidement et efficacement, mais exige une très importante puissance de calcul. Ces charges de travail nécessitent de puissants GPU en plus des UC pour arriver à bout de leurs tâches. Les fournisseurs de services de cloud public proposent aujourd'hui des options d'accélération via des GPU NVIDIA, mais comment évaluer le niveau de performance et la valeur réellement offerts ?

Nous avons comparé la performance de l'apprentissage automatique TensorFlow de deux solutions de services d'hébergement cloud connus: IBM Cloud™, qui propose à la fois machines virtuelles et serveurs bare metal, et Amazon Web Services (AWS), qui ne propose à l'heure actuelle qu'un hébergement sur machine virtuelle pour l'accélération GPU.

Dans les cinq modèles d'apprentissage en profondeur TensorFlow que nous avons testés, l'hébergement sur serveurs bare metal et machines virtuelles d'IBM Cloud a permis de constater des performances comparables à celles de l'offre d'AWS. La solution bare metal d'IBM Cloud que nous avons testée, cependant, est nettement plus avantageuse sur le plan coût/performance que la solution d'AWS. C'est donc elle la plus intéressante des deux pour l'apprentissage automatique.

Si votre entreprise recherche un hébergement cloud pour les charges de travail d'apprentissage en profondeur TensorFlow, l'IBM Cloud représente l'investissement le plus avantageux.

Optimisez votre investissement

Jusqu'à

17,3 %

de performance en plus par dollar

par rapport à Amazon Web Services

L'avènement des big data et l'utilisation de l'apprentissage automatique pour les analyser

Le cabinet d'analyse Gartner définit le big data comme « ...de gros volumes d'actifs d'information très diversifiés qui exigent des formes innovantes de traitement de l'information, permettant d'améliorer les connaissances obtenues, le processus décisionnel et l'automatisation des processus. »[1] L'analyse d'un tel amalgame de données n'est pas à la portée des êtres humains, en raison de l'extrême quantité des volumes et des liens à établir.

L'intelligence artificielle, en revanche, peut apprendre à créer efficacement des connexions pour donner un sens aux données et permettre aux entreprises de glaner de vraies indications métier à partir des informations qu'elles stockent. La méthode s'est d'ailleurs tellement banalisée que l'IA, les applications intelligentes et l'analytique, ainsi que les objets intelligents, figurent aux trois premières places du Top 10 des tendances technologiques de 2018 établi par Gartner.[2] L'apprentissage automatique est un type particulier d'IA. L'apprentissage en profondeur est un sous-domaine de l'apprentissage automatique qui utilise des réseaux de neurones multi-couches pour permettre au système d'apprendre. Les unités de traitement graphique, ou GPU, sont idéales pour fournir aux charges de travail d'apprentissage en profondeur la puissance de calcul nécessaire à leurs opérations.

Avec l'avènement de l'hébergement sur le cloud public, les fournisseurs cloud d'aujourd'hui proposent des solutions avec accélération GPU qui rendent possible l'apprentissage automatique sur le cloud. IBM Cloud est un service de cloud computing complet qui propose l'infrastructure en tant que service, des services de migration cloud, le développement d'applications cloud, des services de stratégie cloud, et bien plus encore. Pour en savoir plus sur l'offre IBM Cloud, visitez le site www.ibm.com/fr-fr/cloud.

Les avantages du bare metal

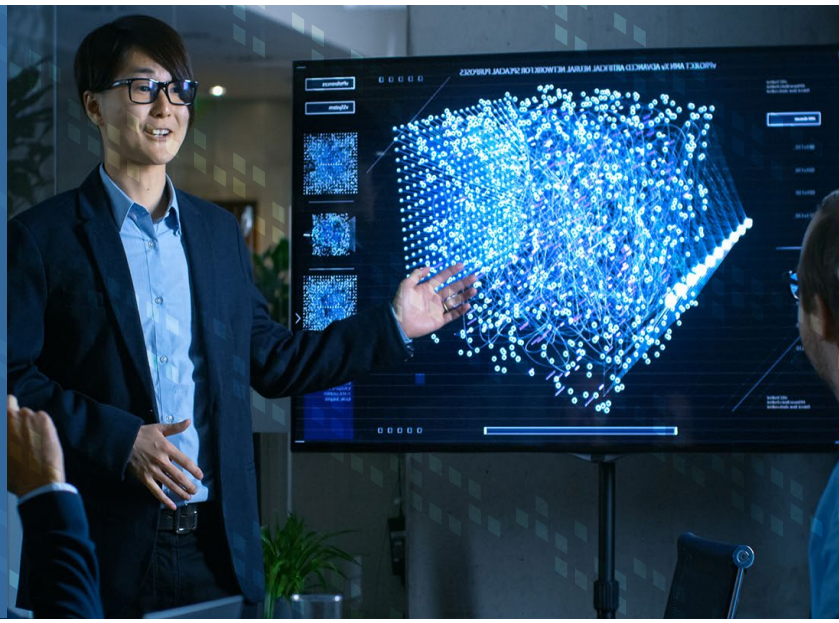
Même si la virtualisation a des atouts certains, le bare metal possède des avantages essentiels capables de répondre aux besoins de certaines entreprises, notamment celles qui souhaitent :

- mieux contrôler la personnalisation,
- ne pas partager de serveur, afin d'éviter que les charges de travail des voisins ne mobilisent régulièrement les ressources et ne dégradent les performances,
- disposer d'accords de niveau de service stricts restreignant les accès, pour des questions de protection des données.

En choisissant l'offre bare metal d'IBM Cloud, vous obtenez des performances similaires à celles d'AWS tout en réalisant des économies sur votre budget de services cloud que vous pouvez réinvestir ailleurs.

A propos de TensorFlow

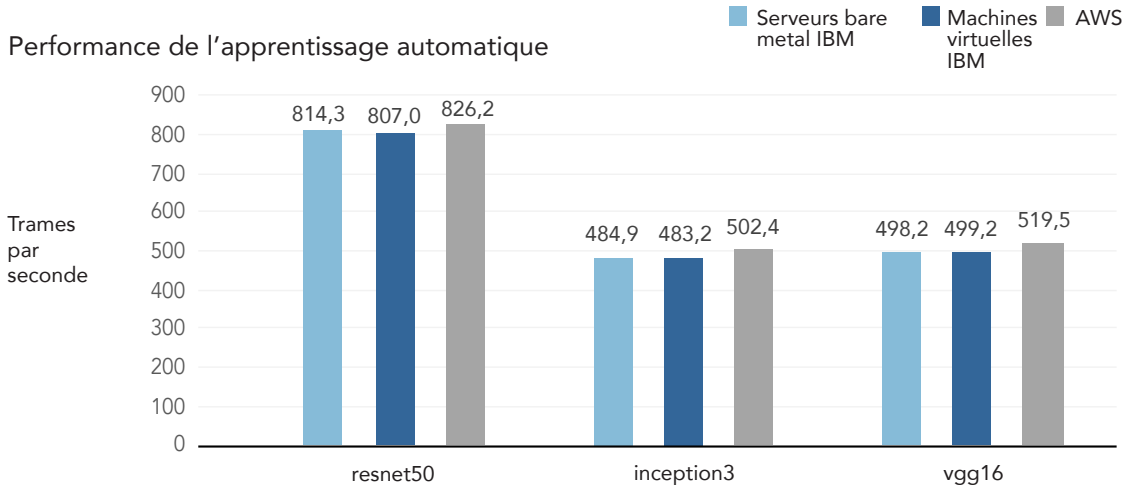
Au cours de nos tests, nous avons utilisé TensorFlow, une bibliothèque open source de modèles d'apprentissage automatique. Nous avons choisi cinq des modèles d'apprentissage en profondeur les plus populaires accessibles publiquement : resnet50, inception3, vgg16, alexnet et googlenet. TensorFlow a mesuré les trames par seconde (FPS) obtenues par les solutions à l'aide de ces modèles, les scores les plus élevés indiquant les meilleures performances pour ces types d'apprentissage automatique. Pour en savoir plus sur TensorFlow, visitez le site https://www.tensorflow.org/.



Comparatif des solutions d'apprentissage en profondeur

Lorsque vous triez une pile de photos, vous pouvez facilement indiquer quelles sont les photos qui représentent un chien, et celles qui n'en comportent pas. Vous en êtes capable parce qu'au fil du temps, vous avez appris ce qu'est un chien, et ce qui le différencie d'un chat, d'un cheval ou d'une poule. Les ordinateurs peuvent aussi apprendre à différencier les chiens d'un autre animal, par exemple une poule. C'est le type d'apprentissage que les modèles TensorFlow ont permis à nos solutions de faire. Les trames par seconde désignent le nombre de trames, ou d'images, que la solution a pu analyser par seconde. Pour notre comparatif, nous avons testé uniquement la rapidité et non l'exactitude de l'apprentissage.

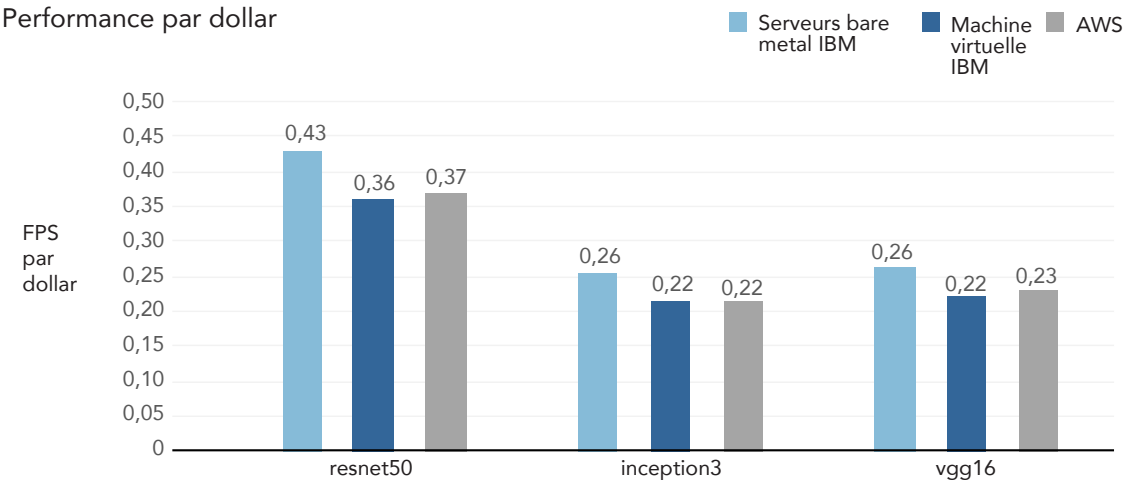
Performances comparables avec une variation de 5% pour chaque modèle



Les modèles d'apprentissage automatique nécessitent des réglages spéciaux pour générer des performances optimales. En gardant cette donnée à l'esprit, nous avons paramétré chaque solution de différentes façons, et nous indiquons ici les meilleurs résultats pour chaque solution. Dans nos cinq modèles d'apprentissage en profondeur, les solutions ont des performances comparables (avec une variation de 5 % ou moins), AWS se montrant légèrement supérieur aux solutions d'IBM® à chaque test. Nous présentons ci-dessus les résultats des trois modèles d'apprentissage en profondeur les plus populaires : resnet50, inception3 et vgg16. En ce qui concerne les résultats pour alexnet et googlenet, consultez les données scientifiques sur lesquelles se base le rapport.

Dans nos tests, la solution bare metal d'IBM Cloud obtient le meilleur rapport coût/performance, avec 17,3 % de FPS (trames par seconde) en plus par dollar[3] avec le modèle resnet50. En d'autres termes, en choisissant l'offre bare metal de l'IBM Cloud, vous obtenez des performances similaires à celles d'AWS tout en réalisant des économies sur votre budget de services cloud que vous pouvez réinvestir ailleurs.

Jusqu'à 17,3 % de performance en plus par dollar



Le tableau ci-dessous indique de façon synthétique les solutions que nous avons comparées. La comparaison de solutions exactement équivalentes peut être difficile lorsque vous avez des recours à des fournisseurs cloud, parce que les options sont limitées. Comme nous comparions des charges de travail sollicitant intensivement le GPU, nous avons choisi les options de référence les plus basses disponibles pour le processeur, la mémoire RAM et l'unité de disque afin d'obtenir un comparatif de prix équitable. Nos réglages au cours des tests de reconnaissance ont montré que ces spécifications n'avaient qu'un impact minimal sur nos charges de travail d'apprentissage en profondeur avec sollicitation intensive du GPU. Le coût de l'offre de serveurs bare metal d'IBM Cloud tel que nous l'avons calculé est de 16 % inférieur à celui de l'offre virtualisée d'AWS. Pour des informations plus détaillées, consultez les données scientifiques sur lesquelles se base le rapport.

Comparatif de prix			
	Serveurs bare metal IBM	**Machines virtuelles IBM**	**AWS**
Processeur	Processeur Intel Xeon E5-2690 v4	Processeur Intel Xeon E5- 2690 v4	Processeur Intel Xeon E5- 2686 v4
Nb de cœurs	28	8	8
Vitesse du processeur (GHz)	2,60	2,60	2,30
RAM (Go)	64	60	61
Disque du système d'exploitation	Disque dur SATA 1 To	SAN 100 Go	SSD EBS 100 Go
GPU	Tesla V100-PCIE-16 Go	Tesla V100-PCIE-16 Go	Tesla V100-SXM2-16 Go
Prix/mois	$1889	$2244,09	$2249,92

Avec IBM Cloud, le coût de gestion de charges de travail exigeantes d'apprentissage en profondeur TensorFlow est plus avantageux

Les entreprises s'intéressent à l'apprentissage automatique qui leur permet d'obtenir des connaissances de leurs données, mais rechignent à lui consacrer un budget trop important. IBM Cloud propose deux solutions d'accélération GPU qui répondent à cette exigence : l'une avec serveurs bare metal, l'autre virtualisée. Nos tests ont démontré que sur les trois solutions analysées utilisant les modèles TensorFlow, la solution bare metal de l'IBM Cloud offrait le meilleur rapport coût/performance. C'est donc une option d'apprentissage en profondeur intéressante pour les administrateurs et les chefs d'entreprise à la recherche d'un maximum de performance pour un coût réduit.

1 Glossaire IT de Gartner, accès le 12 mars 2019, https://www.gartner.com/it-glossary/big-data/.

2 Kasey Panetta, « Garner Top 10 Strategic Technology Trends for 2018, » accès le 12 mars 2019, https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018/.

3 Nous avons arrondi ici à deux décimales, mais basons nos calculs sur les chiffres fournis par TensorFlow. Nos calculs sont basés sur ces chiffres plus précis, que vous pouvez consulter dans les données scientifiques sur lesquelles se base le rapport.

Consultez les données scientifiques sur lesquelles se base le rapport sur le site http://facts.pt/3nk9q5f

▶ Consultez la version originale en anglais de ce rapport à l'adresse http://facts.pt/yoal9hz

Principled Technologies®

Facts matter.®

Ce projet a été commandé par IBM.