Cloud : Amazon EKS couvre les instances EC2 Inf1

Amazon Elastic Kubernetes Service (EKS) prend désormais en charge les instances EC2 Inf1. Dévoilées l'an dernier, les instances Inf1 d'Amazon Elastic Compute Cloud (EC2), le service d'hébergement cloud évolutif du fournisseur américain, sont conçues pour prendre en charge des applications d'inférence de machine learning.

La reconnaissance d'images, la reconnaissance vocale, le traitement du langage naturel ou encore la détection des fraudes font partie des applications possibles.

Les instances EC2 Inf 1 disposent pour les supporter d'un socle technique regroupant jusqu'à 16 puces AWS Inferentia et des processeurs Intel Xeon Scalable de 2e génération. Les instances Inf1 fournissent ainsi un débit « jusqu'à 3 fois plus élevé et pour un coût par inférence jusqu'à 40% inférieur à celui des instances Amazon EC2 G4 », selon AWS.

EKS et SDK AWS Neuron

Le kit de développement logiciel (SDK) AWS Neuron permet aux développeurs d'optimiser les performances d'inférence d'apprentissage automatique des puces Inferentia. Les frameworks comme TensorFlow, PyTorch et MXNet sont supportés.

En outre, l'exécution de conteneurs est facilitée par la mise à jour de l'AMI (Amazon Machine Image) optimisée pour EKS avec les packages nécessaires du kit AWS Neuron.

« Avec EKS et le plug-in AWS Neuron pour Kubernetes, il est facile de combiner plusieurs appareils Inferentia dans votre cluster pour exécuter des charges de travail d'inférence hautes performances », a déclaré le fournisseur de services cloud dans un communiqué.

Les instances EC2 Inf1 peuvent être utilisées « sur tous les clusters EKS exécutant la version 1.14 et ultérieures » dans les régions couvertes, a précisé AWS. Dans un premier temps, seuls les groupes de nœuds *self-managed* sont supportés. Ils peuvent être lancés en utilisant eksctl, CloudFormation ou l'interface de ligne de commande AWS CLI. La prise en charge des groupes de nœuds ménagés EKS sera ajoutée dans une prochaine version.