Comment bien choisir un équipement IA de périphérie de réseau (edge)

L'edge computing est devenu l'une des tendances les plus discutées dans le domaine technologique, et, avec toute cette effervescence, de nombreuses entreprises envisagent d'investir dans des technologies de périphérie (edge) intelligentes pour leur réseau IoT (Internet des objets). Dans cet article, ADLink rappelle ce qu'est l'edge computing et ce que cette approche permet, avant de lister les bonnes questions qu'il faut se poser pour déterminer si une application peut bénéficier des technologies edge.

L'edge computing peut ajouter beaucoup de flexibilité, de vitesse et d'intelligence aux réseaux de l'Internet des objets (IoT), mais il est important de comprendre que les équipements de périphérie de réseau dotées d'intelligence artificielle (edge IA) ne sont pas la panacée pour tous les défis auxquels sont confrontées les applications réseau intelligentes. A la fin de cet article, après avoir déterminé si les technologies edge sont adaptées à une application donnée, nous aborderons les principales caractéristiques et considérations que les acheteurs doivent prendre en compte lors de l'évaluation des équipements edge IA.

AUTEUR

Toby McClean, vice-président, AIoT Technology & Innovation, ADLink Technology.

L'edge computing, c'est quoi ?

L'edge computing permet de hisser l'Internet des objets à un niveau supérieur à la périphérie des réseaux, là où les données brutes prennent toute leur valeur en temps réel. Cette technologie accroît l'importance et la gouvernance des nœuds, points d'extrémité et autres appareils intelligents connectés, en redistribuant l'effort de traitement des données sur l'ensemble du réseau.

L'edge computing est pratiquement l'exact opposé du cloud computing, là où les données arrivent de réseaux distribués pour être traitées dans des centres de données centralisés, les résultats étant souvent retransmis au réseau distribué d'origine pour déclencher une action ou effectuer une modification. Toutefois, le transfert de grandes quantités de données sur de longues distances engendre des coûts. Ces coûts peuvent être mesurés financièrement, mais aussi selon des critères critiques comme, par exemple, la consommation ou le délai de traitement.

C'est là où l'edge computing intervient. Lorsque la consommation, la bande passante et la latence sont vraiment des critères importants, l'edge computing peut être la solution. Contrairement au cloud computing centralisé, où les données peuvent parcourir des centaines de kilomètres pour être traitées, l'edge computing permet de traiter les données en périphérie de réseau, à l'endroit même où les données sont captées et créées, ou là où elles résident (photo A). Cela signifie que la latence de traitement devient presque négligeable et que les besoins en énergie et en bande passante sont généralement très réduits. L'un des principaux moteurs de l'edge computing est lié à la manière

● A.- Contrairement au cloud computing centralisé, où les données peuvent parcourir des centaines de kilomètres pour être traitées, l'edge computing permet de traiter les données en périphérie de réseau, à l'endroit même où les données sont captées et créées, ou là où elles résident.

dont les fabricants de semi-conducteurs augmentent la capacité de traitement des puces sans augmenter la consommation d'énergie de manière considérable. Cela signifie que les processeurs situés en périphérie peuvent faire plus avec les données qu'ils acquièrent, sans pour autant consommer plus d'énergie. Dès lors, cela permet à davantage de données de rester en périphérie, plutôt que d'être transférées vers le cœur des infrastructures. En plus de réduire la consommation système globale, cette approche diminue les temps de réponse et améliore la confidentialité des données.

Parmi les technologies qui bénéficient de cette évolution, on retrouve l'intelligence artificielle (IA) et l'apprentissage automatique (ML), mais ces technologies sont également liées à la réduction du coût d'acquisition des données et à l'augmentation du niveau de confidentialité des données. Ces deux critères peuvent être pris en compte par le traitement en périphérie de réseau. Certes, en ce qui concerne l'IA et le Machine Learning, ces deux technologies ont typiquement requis des ressources gigantesques, bien plus que ce qui serait normalement disponible dans un point d'extrémité ou dans un appareil intelligent. Mais aujourd'hui, grâce aux progrès réalisés tant au niveau du matériel que des logiciels, il est possible d'intégrer ces technologies dans des dispositifs plus petits et plus limités en ressources qui sont déployés en périphérie de réseau (photo B).

● B.- L'intelligence artificielle et l'apprentissage automatique ont typiquement requis des ressources gigantesques, bien plus que ce qui serait normalement disponible dans un point d'extrémité ou dans un appareil intelligent. Mais aujourd'hui, grâce aux progrès réalisés tant au niveau du matériel que des logiciels, il est possible d'intégrer ces technologies dans des dispositifs plus petits et plus limités en ressources qui sont déployés en périphérie de réseau (ici une plate-forme edge AI DLAP de la société ADLink).

Evaluer l'edge IA

Le choix d'une plate-forme capable d'effectuer des traitements en périphérie, avec éventuellement l'exécution d'algorithmes d'IA ou de moteurs d'inférence ML, nécessite une évaluation précise. De fait, les capteurs et actionneurs simples, même ceux qui sont parties intégrantes de l'IoT, peuvent être réalisés à l'aide de circuits intégrés relativement modestes. L'augmentation de la capacité de traitement en périphérie nécessitera une plate-forme plus puissante, utilisant probablement des architectures hautement parallèles. Souvent, cela implique l'usage d'un processeur graphique GPU, mais si cette plate-forme est trop puissante, elle commencera à peser sur les ressources limitées qui existent généralement en périphérie de réseau.

Il est également important de se rappeler qu'un équipement edge est fondamentalement une interface qui communique avec le monde réel, et qu'il devra probablement intégrer certaines technologies d'interface traditionnelles comme, par exemple, des ports Ethernet, GPIO, CAN, série ou USB. Il peut également être nécessaire de prendre en charge des périphériques, tels que des caméras, des claviers et des écrans (photo C).

L'environnement edge peut aussi s'avérer très différent de celui d'un centre de données confortable et bien climatisé. L'équipement edge peut être exposé à des conditions extrêmes de température, d'humidité, de vibration ou même d'altitude. Cela aura un impact sur l'équipement choisi, ainsi que sur la façon dont il est packagé ou hébergé.

Un autre aspect important à prendre en compte est celui des exigences réglementaires. Tout appareil qui utilise des fréquences radio (RF) pour communiquer sera soumis à des réglementations et nécessitera éventuellement une licence pour fonctionner. Certaines plates-formes seront conformes d'emblée « out-of-the-box », mais d'autres nécessiteront plus de travail. Une fois en service, il est peu probable que ces plates-formes puissent bénéficier d'une mise à niveau matérielle, de sorte que la puissance de traitement, la capacité mémoire et la capacité de stockage doivent toutes être soigneusement déterminées lors de la conception afin de permettre une augmentation future des performances.

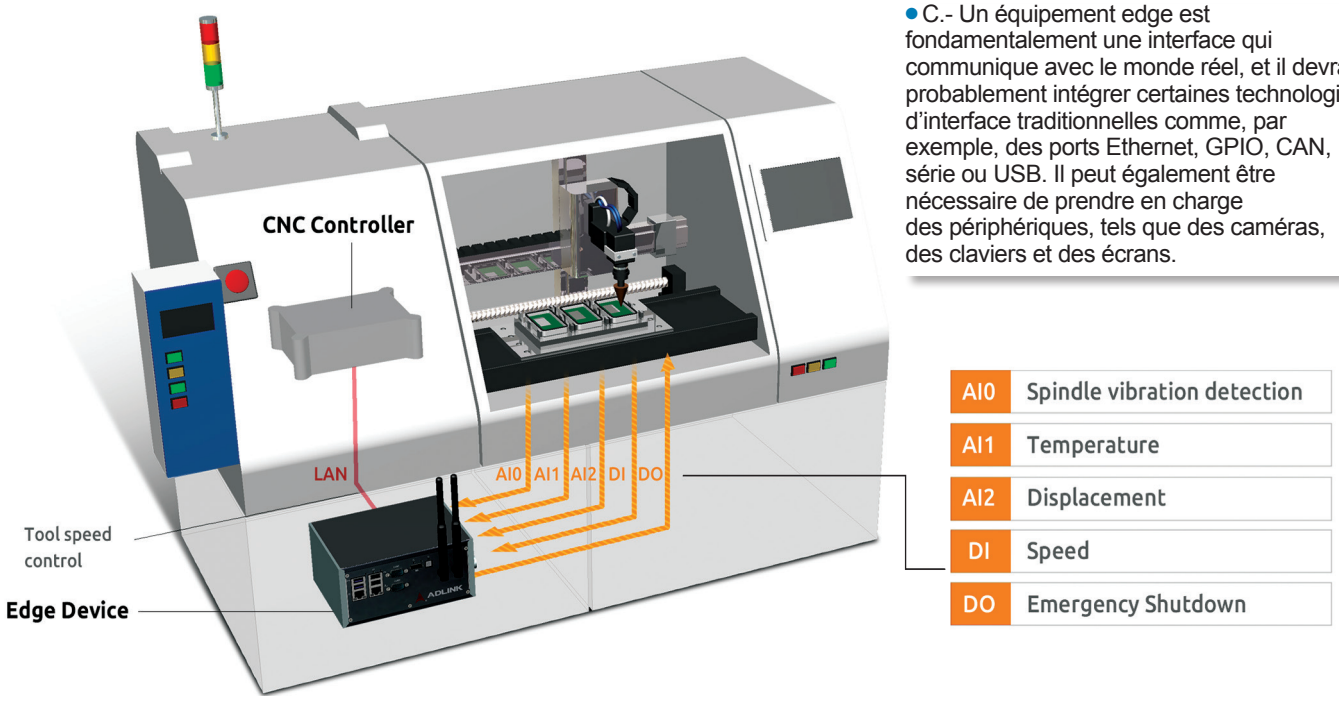
Cela inclut les mises à jour de logiciels. Contrairement au matériel, il est possible de déployer des mises à jour de logiciels alors qu'un équipement est sur le terrain. Ces mises à jour over-the-air (OTA) sont désormais très courantes et il est probable que tout équipement edge devra être conçu pour prendre en charge ce type de mises à jour.

Choisir la bonne solution implique donc une évaluation minutieuse de tous ces points généraux, mais aussi un examen attentif des exigences spécifiques de l'application. L'équipement doit-il traiter des données vidéo ou peut-être audio ? Gère-t-il uniquement la température ou surveille-t-il également d'autres aspects environnementaux ? Doit-il rester allumé en permanence ou se met-il en veille pendant de longues périodes ? Sera-t-il déclenché par un événement extérieur ? Un grand nombre de ces questions vaut pour toutes les technologies déployées en périphérie, mais à mesure que le niveau des traitements augmente et que les attentes à l'égard des résultats s'accroissent, il sera nécessaire d'élargir la liste des exigences.

Les bénéfices de l'edge computing

Il est désormais techniquement possible d'intégrer l'IA et le Machine Learning dans les équipements edge et les nœuds intelligents, ce qui ouvre de nombreuses perspectives. Cela signifie que non seulement le moteur de traitement est plus proche de la source des données, mais que ce moteur peut faire beaucoup plus avec les données qu'il collecte.

Il y a de réels avantages à procéder ainsi. Premièrement, cela peut accroître la productivité, c'est-à-dire l'efficacité avec laquelle les données sont utilisées. Deuxièmement, cela simplifie l'architecture réseau, car il



● C.- Un équipement edge est fondamentalement une interface qui communique avec le monde réel, et il devra probablement intégrer certaines technologies d'interface traditionnelles comme, par exemple, des ports Ethernet, GPIO, CAN, série ou USB. Il peut également être nécessaire de prendre en charge des périphériques, tels que des caméras, des claviers et des écrans.

y a moins de données à déplacer. Troisièmement, la proximité du centre de données devient moins essentielle. Ce dernier point peut ne pas sembler trop important si le centre de données est situé dans une ville et proche de l'action, mais cela peut faire une grande différence si la périphérie de réseau est un endroit éloigné, comme une ferme ou une station de traitement des eaux.

Il est indéniable que les données circulent rapidement sur Internet. Beaucoup de personnes seraient surprises d'apprendre qu'une requête de recherche peut faire deux fois le tour du globe avant que les résultats ne soient affichés sur l'écran. Le temps total écoulé peut être une fraction de seconde, ce qui est, pour nous, pratiquement instantané. Mais pour les machines et les autres équipements intelligents qui constituent l'Internet des capteurs et des actionneurs connectés, intelligents et souvent autonomes, chaque fraction de seconde peut sembler durer une heure.

Cette latence aller-retour est une réelle préoccupation pour les fabricants et les développeurs de systèmes temps réel. Le temps que les données mettent pour aller et venir d'un centre de données à un autre n'est pas sans importance et n'est en aucun cas instantané. La réduction de cette latence est un aspect clé de l'edge computing. Cet aspect entre en conjonction avec des réseaux plus rapides, et c'est là que la 5G pourra jouer son rôle. Mais le déploiement de réseaux plus rapides ne pourra pas compenser la latence cumulée à laquelle nous pouvons nous attendre, du fait que de plus en plus d'équipements vont être mis en ligne.

Les analystes prévoient qu'il pourrait y avoir jusqu'à 50 milliards d'appareils connectés en ligne d'ici à 2030. Si chacun de ces appareils exigeait une large bande passante vers un centre de données, les réseaux seraient en permanence saturés. Si un grand nombre d'entre eux fonctionnent au sein d'un pipeline et attendent que les données arrivent de l'étage précédent, le retard total deviendra très vite évident. L'edge computing est vraiment la seule solution pratique pour désengorger les réseaux.

Toutefois, si le besoin en edge computing en général est bien réel, les avantages spécifiques de l'edge computing dépendront beaucoup de l'application. C'est là que les lois de l'edge computing s'appliquent. Ces lois aideront les équipes d'ingénieurs à décider si l'edge computing est la bonne alternative pour telle ou telle application spécifique.

Les 4 lois de l'edge computing

La première loi est celle de la physique et elle est immuable. L'énergie RF se déplace à la vitesse de la lumière, tout comme les photons dans un réseau de fibres optiques. Ça, c'est la bonne nouvelle. La mauvaise nouvelle, c'est qu'ils ne peuvent pas voyager plus vite que cette limite. Dès lors, si le temps d'aller-retour n'est toujours pas assez rapide, l'edge computing peut être le bon choix.

La latence ne dépend pas pour autant entièrement du mécanisme de transport. Il y a des encodeurs et des décodeurs à chaque extrémité, des couches physiques qui doivent convertir les électrons en n'importe quelle forme d'énergie utilisée, puis réaliser l'opération inverse. Tout cela prend du temps et, même avec des processeurs fonctionnant à des vitesses mesurées en gigahertz, le temps est limité et dépend de la quantité de données qui circulent.

La deuxième loi est la loi de l'économie. Celle-ci est peut-être un peu plus souple, mais avec la montée en flèche de la demande en ressources de traitement et de stockage, elle en est aussi moins prévisible. Les marges sont toujours faibles, mais si le coût du traitement des données dans le cloud augmente soudainement, cela peut faire la différence entre bénéfices et pertes.

Le traitement des données en périphérie n'est pas soumis à ce type de coût variable. Une fois que le coût initial de l'équipement a été engagé, le coût supplémentaire du traitement de n'importe quelle quantité de données en périphérie est pratiquement nul.

Les données ont une valeur parce qu'elles signifient ou représentent

quelque chose. Cela nous amène à la troisième loi, qui est la loi du pays. Quiconque saisit des informations peut désormais être soumis aux lois sur la confidentialité des données qui existent dans la région où ces données ont été collectées. Cela signifie que même si vous êtes le propriétaire légal du dispositif qui capture les données, vous ne serez peut-être pas autorisé à déplacer ces données au-delà des frontières géographiques.

Le traitement en périphérie permet de contourner ce problème. En traitant les données à la périphérie, elles n'ont pas besoin de quitter l'équipement. La confidentialité des données est de plus en plus importante dans les appareils portables grand public ; la reconnaissance faciale sur les téléphones mobiles utilise une IA locale pour traiter l'image de la caméra, de sorte que les données ne quittent jamais l'appareil. Il en va de même pour les systèmes de vidéosurveillance et autres systèmes de surveillance de la sécurité. L'utilisation de caméras pour surveiller les espaces publics implique généralement que les images soient transférées et traitées par des serveurs de données dans le nuage, ce qui pose des problèmes concernant la confidentialité des données. Le traitement des données dans la caméra est à la fois plus rapide et plus sûr, ce qui permet de supprimer ou de simplifier les mesures de protection des données (photo D).

Enfin, nous devons tenir compte de la loi de Murphy, selon laquelle, si quelque chose peut mal tourner, cela tournera mal à coup sûr. Bien sûr, il y a toujours quelque chose qui peut mal tourner, même dans les systèmes les plus soigneusement conçus au monde. Le traitement en périphérie peut éliminer une grande partie des points de défaillance possibles associés au déplacement des données sur un réseau, à leur stockage dans le cloud et au recours à des centres de données pour la puissance de traitement.

Poser les bonnes questions sur l'edge computing

Même si une application peut techniquement bénéficier du traitement en périphérie, il faut d'abord se poser les bonnes questions. Voici un aperçu des questions les plus pertinentes :

- Sur quel type d'architecture de processeur votre application s'exécute-t-elle ? Le portage d'un logiciel sur un jeu d'instructions différent peut s'avérer coûteux et entraîner des retards ; monter en niveau n'est donc pas sans risques.
- De quel type d'entrées/sorties avez-vous besoin ? Il peut s'agir d'un nombre plus ou moins important d'interfaces câblées et/ou sans fil. Les ajouter plus tard serait inefficace, il faut donc s'en occuper dès le début.
- Quel est l'environnement opérationnel ? Est-il extrêmement chaud, extrêmement froid, ou les deux ? La mission vers Mars est un bon exemple, même s'il est extrême, de « traitement en périphérie » où l'environnement opérationnel est extrêmement variable !
- Votre matériel doit-il être conforme aux réglementations ou être certifié ? La réponse à cette question est presque toujours positive et le choix d'une plate-forme précertifiée permet de gagner du temps et de l'argent.
- De quelle puissance électrique votre application aura-t-elle besoin ? L'alimentation d'un système est coûteuse, tant en termes de coût unitaire que d'installation, et il peut être très utile de savoir quelle quantité d'énergie est considérée comme « suffisante ».
- L'équipement en périphérie doit-il loger dans un encombrement donné ? Cet aspect est plus important dans les traitements en périphérie que dans de nombreux autres déploiements ; il doit donc être pris en compte dès le début du cycle de conception.
- A quel type de cycle de vie a-t-on affaire ? S'agit-il d'une application industrielle qui devra fonctionner pendant de nombreuses années, ou le cycle de vie se mesure-t-il en mois ?
- Quelles sont les exigences de performance système, en termes de puissance de traitement, ou peut-être en nombre d'images par seconde ? Quelles sont les exigences en matière de mémoire ? Quel est le langage de l'application ?
- Le critère coûts est-il à considérer ? Il s'agit d'une question délicate, car la réponse est toujours affirmative, mais le fait de savoir dans quelle mesure les coûts sont limités vous aidera dans le processus de sélection.

Le choix de la bonne plate-forme peut être facilité par le choix du bon partenaire technologique. A cet égard, ADLink dispose d'un large portefeuille de solutions d'edge computing et est associé à un grand nombre de sociétés proposant des technologies complémentaires. En entrant dans un écosystème qui a été développé autour de l'edge computing, vous aurez toutes les chances d'opter pour la bonne plate-forme de traitement en périphérie pour votre application dopée à l'intelligence artificielle. ■

● D.- L'utilisation de caméras pour surveiller les espaces publics implique généralement que les images soient transférées et traitées par des serveurs de données dans le nuage, ce qui pose des problèmes concernant la confidentialité des données. Le traitement des données dans la caméra au niveau edge est à la fois plus rapide et plus sûr, ce qui permet de supprimer ou de simplifier les mesures de protection des données.