Perspective

Boostez vos applications d'IA

Comment la bonne infrastructure vous aide à éviter les saturations de données et à libérer tout le potentiel de l'IA.

verizon business

Avant-propos

Pour concrétiser vos projets d'IA, votre écosystème doit reposer sur un réseau à la hauteur de ces nouveaux enjeux.

David Bailey
Global Solutions Executive

Depuis plus de 20 ans, David Bailey aide de grandes entreprises internationales et des acteurs du secteur public à définir et à satisfaire leurs besoins en matière d'infrastructure et de sécurité. Ses domaines de compétences comprennent notamment les réseaux filaires et sans fil pour l'Internet des objets (IoT), ainsi que les infrastructures agiles et intelligentes qui sous-tendent les applications les plus exigeantes, notamment l'IA.

L'IA est en plein essor. Face à l'évolution de ses modèles et à des besoins en puissance de calcul toujours plus importants, les technologies qui sous-tendent ces nouveaux outils doivent s'adapter. À commencer par les réseaux. Dans ce domaine, haute performance et faible latence sont devenues des critères indispensables.

Les technologies d'intelligence artificielle et de machine learning (IA/ML) prolifèrent et entrent dans nos habitudes. Dans tous les secteurs, les Comex se posent la même question : comment exploiter l'IA pour gagner en compétitivité, augmenter les bénéfices, transformer les opérations et améliorer l'expérience client ? En réponse, on voit apparaître des matériels toujours plus puissants et des logiciels toujours plus sophistiqués, chacun conçu dans le but de satisfaire les exigences de l'IA et de concrétiser ses promesses.

Les fabricants de semi-conducteurs (Nvidia, AMD, Intel, Qualcomm, etc.) cherchent à prendre l'avantage sur le marché des puces IA, optimisées pour les workloads d'intelligence artificielle. Mais certains acteurs de la tech préfèrent s'affranchir de ces fournisseurs. C'est le cas de Google, d'Amazon et de Microsoft qui développent leurs propres processeurs. Grâce à ces équipements sur mesure, ces géants du cloud peuvent désormais fournir des services d'IA publics capables de traiter d'énormes charges en à peine quelques millisecondes.

Un excellent moyen de créer et de développer une solution performante. Mais pour transporter toutes ces données jusqu'au cloud ou au data center (généralement via Internet), vous aurez besoin d'un réseau suffisamment robuste. De la qualité de cette connectivité dépendra la réactivité de votre application. Vous aurez beau investir en masse pour booster la performance de vos puces IA, si votre réseau est trop lent, tous vos efforts seront vains.

Prenons un exemple. L'heure du dîner arrive, et vous voulez vous faire livrer un burger. Cela tombe bien, le restaurant peut préparer votre menu en quelques minutes (l'équivalent de votre puce IA capable de traiter un workload en une poignée de millisecondes). Mais si le livreur (c.-à-d. votre connexion Internet) met une demi-heure à vous apporter votre repas, vous mangerez froid. Peu importe si le chef s'est démené aux fourneaux.

Cet article aborde ces thèmes essentiels : l'importance de la latence (le délai de livraison), les solutions pour la réduire et les mesures à mettre en place pour concrétiser vos projets d'IA.

Les entreprises sont de plus en plus nombreuses à mettre au point leurs propres puces IA pour booster la performance. C'est ainsi qu'est née MTIA (Meta Training and Inference Accelerator). Grâce à la supériorité de sa puissance de calcul et de sa bande passante mémoire, cette puce de seconde génération booste la performance des applications IA de Meta, par exemple les moteurs de notes et de recommandations sur Facebook et Instagram.

L'utilisation de puces IA sur mesure dans des technologies hors cloud et hors data center se normalise. Voici quelques exemples de cas d'usage :

- Dans le secteur de la santé, certains équipements sophistiqués (scanners, IRM, TEP-scan, etc.) ont besoin de traiter de grandes quantités de données, très rapidement.
- Aujourd'hui, les véhicules sont bardés de capteurs pour éviter les collisions et renforcer la sécurité du conducteur et des usagers de la route. Ces capteurs sont connectés à des puces surpuissantes qui permettent au véhicule de prendre des décisions critiques en quasi temps réel.

Dans les deux exemples précités, la majorité des données sont traitées localement (à bord du véhicule ou dans l'équipement médical), même si certaines de ces données sont envoyées dans le cloud après leur traitement.

Mais il existe de nombreux autres cas dans lesquels le traitement intervient hors site :

Finance
- L'agrégation et l'analyse de vastes quantités de données historiques et temps réel permettent de prendre des décisions rapides là où chaque milliseconde compte. Cette réactivité est essentielle pour optimiser les résultats financiers.
- L'analyse des données biométriques et comportementales facilite la détection des fraudes. Contrairement aux méthodes traditionnelles basées sur des questions de sécurité, cette approche est beaucoup plus efficace et a l'avantage de préserver l'expérience utilisateur.

Industrie
- Grâce à l'analyse automatisée des données vidéo captées en temps réel sur les lignes de production, chaque produit peut être contrôlé à différentes étapes du processus, là où la méthode habituelle se contente de contrôles qualité sporadiques en fin de ligne. Les défauts sont détectés plus rapidement, réduisant ainsi le gaspillage et le risque d'endommagement des machines.

Retail
- L'analyse des comportements et de l'historique d'achat des clients ouvre la voie à des expériences in-store plus qualitatives et personnalisées. Un souci du détail qui améliore à son tour la valeur vie client.

Industrie, agriculture, services... tous les secteurs peuvent bénéficier de ces avantages. Les puces IA sont d'ailleurs omniprésentes. Lorsqu'une entreprise choisit de migrer ses applications et son calcul vers le cloud, son réseau joue alors le rôle d'intermédiaire entre les lieux de collecte, de traitement et de stockage vers lesquels transitent les données.

La problématique

La généralisation rapide du cloud computing a redéfini le modèle économique de l'IT en entreprise, incitant les équipes achat et les DSI à optimiser leurs coûts de réseau. La plupart du temps, les entreprises achètent séparément les composantes underlay (accès réseau) et overlay (routeurs et commutateurs). Dans un souci d'économies, elles sélectionnent les offres les moins chères et se retrouvent parfois avec des dizaines de fournisseurs différents.

Mais la réduction des coûts se fait souvent au détriment de la performance et de la qualité de service. À l'origine de cette problématique, la certitude que tous les accès Internet se valent.

Ce n'est pas le cas. La fiabilité, la performance et, plus particulièrement, la latence de ces accès varient de façon significative d'une offre à l'autre.

J'explore ce sujet plus en détail dans l'article ci-dessous.

Intégrer ou ne pas intégrer ?
Dans la même série, cet article apporte des éléments de réponse à la question que tout le monde se pose : faut-il confier l'overlay et l'underlay au même fournisseur, ou les séparer ?

Temps de lecture : 15 min

L'importance de la latence

La latence, autrement dit le délai entre l'envoi de la requête et la réponse, se décompose en trois éléments, présentés en détail ci-dessous. La latence telle que nous la percevons est la somme de ces trois composantes.

Une latence trop élevée peut avoir des conséquences désastreuses : opportunité manquée, impossibilité de bloquer une transaction frauduleuse à temps, vente non finalisée, etc. Et dans le cas d'un véhicule autonome ou dans un contexte médical, c'est la vie des usagers ou des patients qui est en danger.

Pour les IA « embarquées », autrement dit contenues dans un système précis (un véhicule, une machine IRM, etc.), le réseau est intégré à l'équipement. La distance que les données doivent parcourir est minime et la latence ne dépend d'aucune infrastructure externe. Elle reste donc elle aussi très faible.

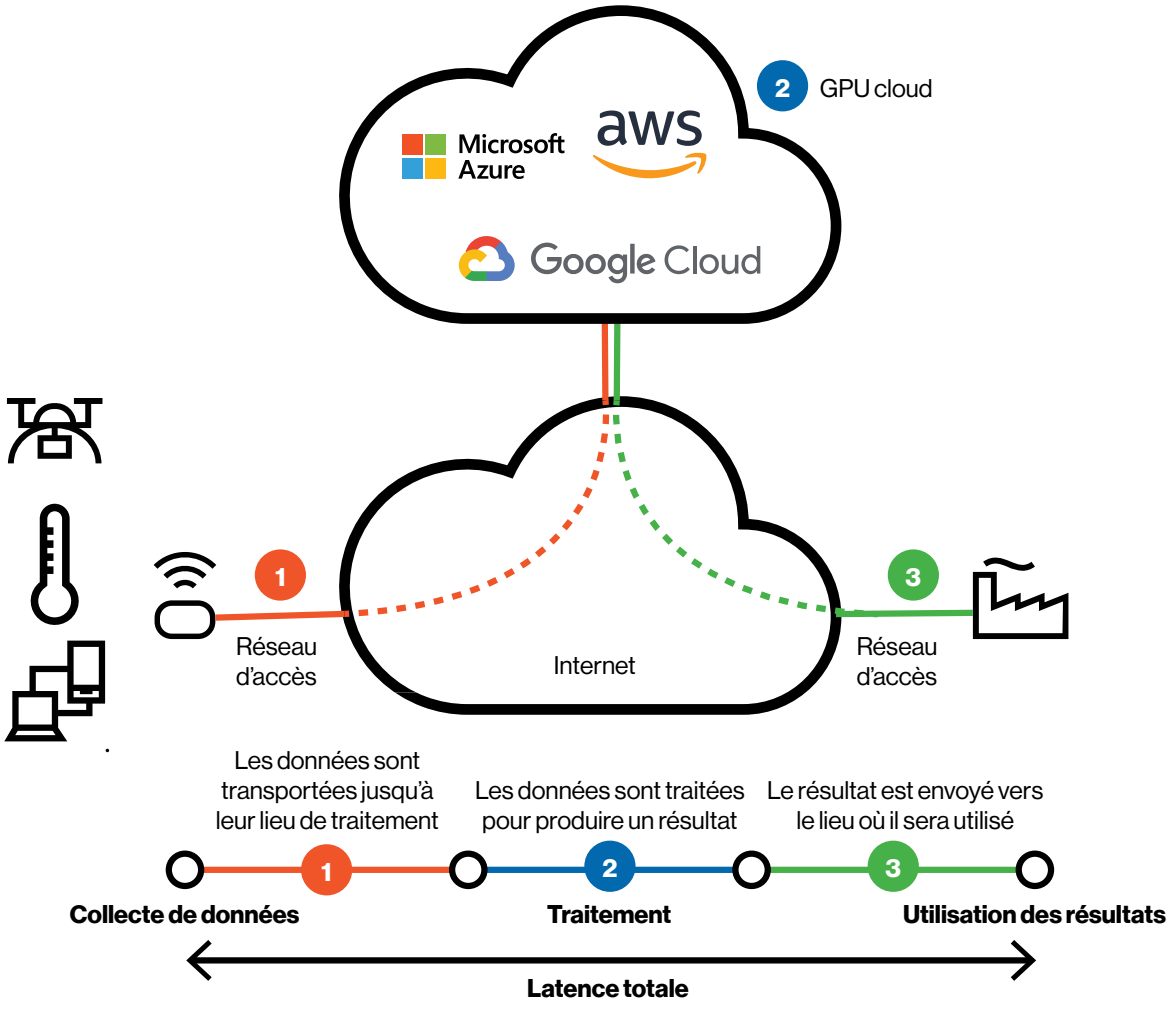
Dans d'autres cas, l'IA et ses capacités de calcul se limitent à l'échelle d'un site et de ses équipements (les machines d'une usine, par exemple). Une infrastructure LAN (fixe ou sans fil) parfaitement conçue et gérée se révèle alors essentielle pour réduire la latence.

Certaines entreprises utilisent leur IA dans plusieurs sites à la fois, un cas de figure qui nécessite le déploiement d'une infrastructure WAN. La latence de leurs applications d'IA dépend donc des fournisseurs de cette infrastructure.

Plusieurs facteurs doivent être pris en compte :

- La quantité de données traitées, qui détermine la capacité totale de l'architecture de calcul
- L'origine des données et la distance qu'elles doivent parcourir pour leur traitement, dans le but de définir les besoins en latence
- Avec qui l'application doit échanger (humain ou système) et pour quelle finalité, afin de cerner l'impact de la latence sur les résultats

Dans la mesure où Internet est constitué d'une myriade de réseaux, et où les données peuvent emprunter différents itinéraires, la latence n'est généralement pas garantie. Et le risque de subir des latences trop élevées est encore plus grand lorsque l'entreprise fait appel à plusieurs prestataires pour gérer les accès dans chaque pays ou sur chaque site. Il en va de même pour les organisations qui se fournissent en composants underlay (accès) et overlay (routeurs et commutateurs managés) auprès de fournisseurs différents : la gestion de la latence s'en retrouve fragmentée.



Toute personne impliquée dans la gestion d'un environnement IT vous le dira : lorsque la performance d'une application n'est pas à la hauteur, l'utilisateur final se moque de savoir d'où vient le problème. Tout ce qu'il veut, c'est qu'il soit résolu. De la même manière, quand une entreprise dépense des sommes astronomiques pour acquérir des puces IA dans l'espoir de booster la performance de ses applications, la priorité n'est pas de comprendre pourquoi les résultats ne sont pas au rendez-vous. Le plus important est de corriger la situation pour capitaliser sur son investissement.

Un autre facteur à l'équation

Face à des problèmes de performance, il est possible de mettre les serveurs à niveau, d'optimiser les puces ou d'augmenter les capacités de calcul, et ce presque instantanément. Par contraste, mettre à niveau le réseau peut prendre des mois, voire plus si l'infrastructure WAN s'étend sur de multiples sites.

46 %

des preuves de concept (PoC) probantes de l'IA n'atteignent pas la phase de mise en production[1].

Prenons un exemple. Un collaborateur en télétravail tente tant bien que mal de suivre une réunion en visioconférence, mais la vidéo n'arrête pas de couper. Ce genre de dysfonctionnements est généralement dû à une dorsale Internet sous-performante. Lorsque vos données IA critiques transitent via Internet, vous ne disposez que d'une visibilité limitée (quand visibilité il y a) sur les problèmes de routage et les goulets d'étranglement qui nuisent à la performance de vos applications d'IA. Pour plus d'informations à ce sujet, explorez l'article mentionné en bas de cette page. Tous les efforts déployés pour optimiser votre application d'IA, en investissant dans des solutions d'hébergement spécialisées et des puces personnalisées, seront vains si votre réseau ne fournit pas la performance nécessaire et si les problèmes ne peuvent pas être résolus de façon simple et rapide.

Plus vous injectez de données dans votre modèle d'IA, plus vous aurez besoin de puissance de calcul, d'espace de stockage… et d'un réseau performant. Si vous commencez petit, ce qui est tout à fait possible, vous devrez vérifier régulièrement que votre infrastructure est suffisamment robuste pour absorber les nouvelles charges.

Il est tout aussi important de bien évaluer vos fournisseurs réseau. Posez-leur ces questions essentielles :

- Êtes-vous directement connecté à une dorsale Tier 1 au cœur d'Internet ? Ou l'accès que vous fournissez s'arrête-t-il à la périphérie du réseau, en laissant les données emprunter un chemin aléatoire ?
- Êtes-vous directement connectés aux grands hyperscalers comme AWS, Microsoft Azure et Google Cloud Platform ?
- Offrez-vous des garanties en termes de latence et de performance ?
- Êtes-vous en mesure de prioriser certaines applications sensibles à la latence, comme dans beaucoup de cas d'usage IA ?
- Assurez-vous une visibilité de bout en bout sur la performance (du site du client jusqu'à l'hyperscaler) ?

Pour une performance optimale, le réseau doit être aussi efficace que les puces dans votre data center ou l'infrastructure de votre opérateur cloud.

Le réseau mondial de Verizon est justement conçu pour répondre aux exigences des applications d'IA. Nos experts vous accompagnent pour concevoir et déployer une infrastructure à la hauteur de vos projets.

Accès réseau : explorer les options

Cet autre article de la série vous invite à découvrir à quel point le choix du fournisseur influence les performances et les résultats pour votre entreprise. Une décision d'autant plus cruciale pour les applications sensibles à la latence, très présentes dans les cas d'usage de l'IA.

Temps de lecture : 15 min

Dorsale multi-service Verizon : une solution taillée pour les enjeux de l'IA

Verizon fait de l'optimisation de son infrastructure réseau une priorité. Nous avons investi 176 milliards $ depuis 2000 et prévoyons de débloquer entre 17,5 milliards et 18,5 milliards $ supplémentaires avant la clôture de l'exercice en cours. Les réseaux Verizon reposent sur une dorsale multi-service (MSB) internationale, garante d'une priorisation granulaire des applications, y compris sur le réseau Internet mondial.

Cette approche nous permet d'adosser nos offres à des SLA définissant des garanties contractuelles en termes de latence et de performance. Bref, notre offre d'accès Internet est conçue et gérée pour s'adapter à votre puissance de calcul IA dans le cloud.

Autre avantage : nos services s'intègrent aux systèmes d'exploitation avancés de multiples fournisseurs (dont Cisco, Juniper et Aruba) pour étendre ces contrôles jusqu'à la périphérie du réseau grâce aux technologies LAN et de routage managées. Notre réseau est équipé de son propre moteur IA qui permet de déplacer les ressources d'un opérateur cloud à un autre de façon dynamique et intelligente, suivant ainsi les variations de consommation de puissance de calcul de vos applications entre différents opérateurs.

Nous avons enchâssé ces capacités dans le cœur même de notre MSB, ce qui nous permet de prioriser les applications sensibles à la latence, le tout encadré par des SLA adaptés. Si nous pouvons proposer ces boosters de performance et réduire ainsi la latence, c'est grâce à une combinaison de toutes ces capacités :

- Verizon détient et gère sa propre infrastructure
- Nous pouvons équilibrer les charges entre différents opérateurs cloud
- Nous établissons un peering direct avec ces derniers

Nous avons déjà évoqué la stratégie des géants Meta, Google, Amazon et Microsoft, lesquels développent leurs propres puces optimisées pour l'IA. Un investissement considérable qui leur permet de réduire la latence tout en boostant la performance. Ces entreprises considèrent leur dorsale réseau comme un élément essentiel à la réussite de leurs projets d'IA. D'où un investissement conséquent dans des architectures de premier plan.

Verizon compte parmi ses équipes des experts réseau aguerris et des spécialistes de renom intervenant en première ligne. Tous ces professionnels travaillent sans relâche pour développer nos réseaux mondiaux et répondre aux besoins d'aujourd'hui et de demain. Notre mission : assurer un niveau de performance à la hauteur des exigences de l'IA.

Intégrés à la fabric même de notre architecture mondiale, nos boosters de performance réduisent la latence et vous permettent de bénéficier d'un réseau rapide, sécurisé et résilient. L'avantage est double : vous corrigez les problèmes de performance en toute simplicité et donnez un nouvel élan à vos projets d'intelligence artificielle.

Échangeons

Notre équipe d'experts intervient à vos côtés pour libérer tout le potentiel de vos applications d'IA et concrétiser vos projets dans ce domaine.

IoT, analyse en temps réel, intelligence artificielle… Verizon et ses technologies avancées vous aident à surveiller, gérer et optimiser vos opérations pour garder une longueur d'avance dans cette nouvelle ère. Des questions ? Cliquez ici pour nous contacter :

verizon.com/business/fr-fr/contact-us

Série Network Procurement

Cet article s'inscrit dans le cadre d'une série qui explore les pressions croissantes qui pèsent sur le réseau et dresse la liste des questions essentielles à poser aux fournisseurs durant le processus d'achat. Il vise ainsi à guider les entreprises dans le choix d'une solution réellement à la hauteur de leurs enjeux présents et futurs.

Prêt pour le déferlement annoncé ?

Données **IoT** **IA**

Cet article explore les principaux ressorts de l'explosion des volumes de données collectées par les entreprises et ses implications pour la planification réseau.
verizon.com/business/resources/articles/iot-genai-data-explosion.pdf

Accès réseau : explorer les options

Performances

De nombreuses décisions entrent en jeu lors de l'achat de solutions réseau. Pour évaluer les options disponibles, il est impératif de bien comprendre les trois couches qui composent Internet. Cet article explique leur importance pour la performance et la sécurité des réseaux.
verizon.com/business/resources/articles/tier-1-isp-enterprise-connectivity.pdf

Peering réseau

Cloud **Performances** **Fiabilité**

Composante essentielle aux performances du réseau, le peering est tout autant indispensable pour les applications d'entreprise, notamment dans le cloud. Malgré cela, ce sujet figure rarement au cahier des charges des processus d'achat. Ce court article vous aide à rectifier le cap.
verizon.com/business/resources/articles/network-peering.pdf

Quelle visibilité sur vos performances ?

Données **Performances** **Simplicité de gestion**

Découvrez dans cet article comment un achat séparé des composants physiques (underlay) et logiciels (overlay) du réseau peut impacter la performance, la visibilité et la gestion de votre réseau.
verizon.com/business/resources/articles/overlay-underlay-network-procurement.pdf

L'union fait la force

Sécurité **Performances** **Simplicité de gestion**

Dans un contexte de prolifération et de sophistication des cybermenaces, cet article offre six raisons de mieux intégrer la cybersécurité et les réseaux, parmi lesquelles une protection renforcée, une baisse des coûts et une réduction de la charge de travail.
verizon.com/business/resources/articles/unified-network-security-services.pdf

Boostez vos applications IA

IA **Performances**

L'intelligence artificielle (IA) s'annonce comme la technologie la plus disruptive depuis la démocratisation d'Internet, il y a près de 30 ans. Cet article souligne l'importance cruciale des performances réseau pour concrétiser les promesses des applications d'IA.
verizon.com/business/resources/articles/network-infrastructure-ai-platforms.pdf

1. Verizon, Déployer l'IA à grande échelle, 2025

verizon business