DÉFINIR QUATRE ARCHÉTYPES EDGE ET LEURS EXIGENCES TECHNOLOGIQUES

Introduction

Au cours des dernières années, "l'edge computing" est devenu l'une des tendances les plus discutées en informatique, et ceci pour une bonne raison. Grand Valley Research prévoit un **TCAC de 41 % pour l'edge computing** entre 2018 et 2025. Presque toutes les industries reconnaissent les limites de la prise en charge des utilisateurs et des technologies émergentes à travers des infrastructures informatiques centralisées et rapprochent le stockage et l'informatique au plus près des utilisateurs et des appareils.

Ce changement devient nécessaire en raison de la connectivité accrue des appareils et des utilisateurs, ainsi que des énormes volumes de données qu'ils génèrent et consomment. D'après le **Cisco Visual Networking Index** (indice de réseau visuel Cisco), le trafic IP mondial a été de 1,2 zettaoctet en 2016. En 2021, ce chiffre devrait tripler pour atteindre 3,3 zettaoctets. Ainsi d'ici 2021, Cisco prévoit que le nombre d'appareils connectés à des réseaux IP sera trois fois supérieur à la population mondiale. Cela représente plus de 23 milliards d'appareils connectés en seulement trois ans. **D'autres entreprises font des projections semblables**: d'ici 2020, Gartner prévoit 20,8 milliards d'appareils connectés, IDC 28,1 milliards et IHS Markit 30,7 milliards.

Une grande partie de ces données IoT seront des données de capteurs mobiles devant être transmises sur des réseaux sans fil ou mobiles plutôt que des connexions Internet câblées, ce qui crée une charge sur l'infrastructure des réseaux mobiles. **Le trafic IP mobile devrait être multiplié par sept d'ici 2021**, deux fois plus vite que la croissance du trafic IP fixe.

Les changements de l'infrastructure informatique et de stockage requis pour prendre en charge un futur intelligent et connecté, surtout au niveau local, seront profonds.

Cependant, quand on étudie les informations actuellement disponibles à propos de l'edge computing, on découvre qu'il existe peu voire pas de ressources qui donnent un aperçu complet de l'écosystème edge. Une analyse approfondie du marché révèle une grande variété de cas d'utilisation actuels et émergents et, bien qu'ils partagent certaines similarités basées sur la large définition de l'edge computing, ils sont également distincts de manières significatives.

Vertiv a analysé les cas d'utilisation constituant l'écosystème edge afin d'avoir une meilleure compréhension de ces différences et de leurs implications pour l'infrastructure de prise en charge. Suite à cette analyse, nous avons identifié quatre archétypes principaux pour les applications edge:

- Grand volume de données
- Sensibilité à la latence humaine
- Sensibilité à la latence machine à machine
- Critique pour la vie humaine

Ce papier présente une description de chaque archétype avec des exemples des cas d'utilisation ayant le plus d'impact, ainsi qu'un aperçu de leurs exigences de connectivité avec les hubs locaux, urbains et régionaux, qui représentent le cœur et la couche de transmission périphérique et sont parfois distingués en edge, fog et cloud computing.

Comprendre les cas d'utilisation edge

Pour identifier les quatre archétypes, il a d'abord été nécessaire de comprendre les cas d'utilisation de la technologie edge. L'équipe de recherche Vertiv a identifié et examiné plus de 100 cas d'utilisation de la technologie edge et affiné sa liste initiale aux 24 qui auront l'impact le plus important sur l'infrastructure IT pour une analyse plus détaillée.

L'analyse a examiné les exigences de performance de chaque cas en termes de latence, disponibilité et croissance projetée, ainsi que d'exigences de sécurité telles que le besoin de cryptage, d'authentification et de conformité réglementaire. Le besoin d'intégration avec les applications existantes ou anciennes et autres sources de données ainsi que le nombre d'emplacements potentiels requis pour prendre en charge le cas d'utilisation ont également été évalués.

En particulier, l'équipe a étudié les caractéristiques des données des cas d'utilisation et a constaté que les applications qui soutiennent chacun ont un ensemble d'exigences en matière de charge de travail axées sur les données en plus de leurs exigences de disponibilité et de sécurité. Celles-ci incluent le volume de données, le moyen d'accès aux données, les exigences de transmission des données, l'intégrité des données et l'analyse des données. Cette approche axée sur les données, filtrée à travers les exigences de disponibilité et de sécurité, est essentielle pour comprendre et catégoriser les exigences de divers cas d'utilisation.

Une liste de 24 cas d'utilisation, organisée par archétype, se trouve dans la Figure 1.

L'écosystème Edge

GRAND VOLUME DE DONNÉES	SENSIBILITÉ À LA LATENCE HUMAINE	SENSIBILITÉ À LA LATENCE MACHINE À MACHINE	CRITIQUE POUR LA VIE HUMAINE
• Connectivité restreinte • Villes intelligentes • Usines intelligentes • Maisons et bâtiments intelligents • Distribution de contenus HD • Informatique Haute Performance • Réalité Virtuelle • Digitalisation Pétrole et Gaz	• Sécurité intelligente • Smart Grid • Distr. de contenus à faible latence • Arbitrage Marché • Analyses en temps réel • Simulation de Forces Armées	• Santé numérique • Véhicules Autonomes / Connectés • Drones • Transport intelligent • Robots autonomes	• Optimisation de site internet • Réalité Augmentée • Vente au détail Intelligente • Traitement Automatique du Langage Naturel

Figure 1: *Archétypes*

Archétype Un: Grand volume de données

Bande passante	Latence	Disponibilité	Sécurité
Élevée	Moyenne	Élevée	Moyenne

L'archétype de grand volume de données représente les cas d'utilisations où la quantité de données rend difficile le transfert par le réseau directement sur le cloud, ou du cloud vers le point d'utilisation, en raison de problèmes de volume de données, de coût ou de bande passante.

L'exemple d'application edge de grand volume de données le plus largement discuté est probablement la livraison de contenu haute définition. **En 2016, la vidéo a constitué 73 % du trafic IP et cela devrait monter à 82 % d'ici 2021** avec la croissance de la diffusion de vidéo et de réalité virtuelle. Les principaux fournisseurs de contenus, tels que Amazon et Netflix, concluent activement des partenariats avec des fournisseurs de colocation afin d'agrandir leurs réseaux de diffusion et de rapprocher la diffusion vidéo intensive en données des utilisateurs afin de réduire les coûts et la latence.

Aujourd'hui déjà, **35% du contenu auquel un utilisateur d'Internet en Amérique du Nord accède est envoyé de la zone municipale où se trouve l'utilisateur**. Cela devrait augmenter à 51% d'ici 2021, alors que les fournisseurs de contenu poursuivent l'extension de leurs réseaux vers la périphérie. Cependant, cela ne représente que la première vague d'informatique du cœur vers la périphérie. Alors que la demande de vidéo haute définition continue d'augmenter, les hubs locaux prendront de plus en plus en charge les hubs urbains actuels afin de réduire encore les coûts de bande passante et les problèmes de latence.

Un autre exemple important de l'archétype de grand volume de données est l'utilisation de réseaux IoT pour créer des maisons, bâtiments, usines et villes intelligents. Une enquête de 2018 par 451 Research et Vertiv a déterminé qu'alors que seuls 33 %des 700 organisations étudiées avaient largement déployé l'IoT, 56 pour cent ont indiqué qu'au moins 25 % de leur capacité informatique prenait en charge l'IoT actuellement. Même si l'IoT en est toujours à ses débuts, les organisations ont déjà du mal à gérer le volume de données généré.

Dans ce cas, le défi est le contraire de celui présenté par la livraison de contenu haute définition. Au lieu de rapprocher les données des utilisateurs, ces applications doivent déplacer les énormes quantités de données générées par les appareils et systèmes à la source vers un emplacement central pour le traitement. Ceci nécessitera l'évolution d'une architecture de réseau du périphérique au cœur.

L'IoT et l'Internet des objets industriels (IIoT) représentent un grillage de capteurs qui génèrent d'énormes volumes de données chaque heure. Ces données prennent en charge une boucle « capter-déduire-réagir » offrant la visibilité et le contrôle de tout, des appareils domestiques à l'équipement industriel. Seul un sous-ensemble de ces données est transmis à un datacenter local, régional ou sur le cloud pour traitement, ce qui signifie que d'énormes quantités de calculs seront requises à l'extrémité du edge afin de permettre aux appareils et systèmes de prendre des décisions et d'agir en fonction des données fournies par les capteurs.

Les plus simples de ces applications, la maison intelligente, doivent prendre en charge plusieurs appareils et systèmes intensifs en données, y compris le divertissement, les systèmes CVC et la sécurité.

Grand volume de données
D'après IHS Markit, **le marché mondial des appareils domestiques connectés va augmenter de plus de 100 millions d'unités en 2017 à environ 600 millions d'unités en 2021**.

Les villes et usines intelligentes reprennent les défis relatifs aux données inhérentes aux maisons intelligentes et les amplifient. De nombreuses villes expérimentent ou évaluent déjà la technologie de ville intelligente pour améliorer les flux de trafic, soutenir les services d'urgence et réduire les coûts.

Les usines intelligentes, qui tirent parti de la convergence de l'IoT, des systèmes cyber-physiques et du cloud pour permettre aux fabricants d'utiliser les données en temps réel afin d'augmenter l'efficacité, réduire les coûts et s'adapter aux changements de la demande, sont annoncées comme la prochaine révolution industrielle. D'après McKinsey, les usines et autres environnements de production peuvent réaliser le plus gros impact financier en appliquant l'IoT. Ils prédisent que l'IIoT génèrera une **valeur économique entre 1,2 millier de milliards $ et 3,7 milliers de millions $** d'ici 2025. Cette valeur viendra de nouvelles efficacités énergétiques, de la productivité des laboratoires, de l'optimisation des inventaires et d'une meilleure sécurité des travailleurs. Mais il faudra une infrastructure locale solide pour y arriver.

Dans l'industrie du pétrole et du gaz, la numérisation a déjà créé de vastes améliorations dans l'efficacité des processus d'exploration et d'extraction, mais a également introduit d'immenses défis en matière de gestion des données. Une

seule plateforme de forage peut générer des téraoctets de données chaque jour.

D'autres cas d'utilisation relevant de l'archétype de grand volume de données incluent la réalité virtuelle, l'informatique haute performance et les environnements à connectivité limitée, tels que les zones où des opérations de récupération ont lieu suite à une catastrophe naturelle ou une cyberattaque.

Le point commun de tous ces cas d'utilisation est le besoin de déplacer de grands volumes de données vers les utilisateurs, où elles peuvent être consommées, ou à partir des appareils et systèmes où elles sont générées vers un répertoire central.

Archétype Deux: Sensibilité à la latence humaine

Bande passante	Latence	Disponibilité	Sécurité
Moyenne	Élevée	Moyenne	Moyenne

L'archétype de la sensibilité à la latence humaine couvre les cas d'utilisation où les services sont optimisés pour la consommation humaine. Comme son nom l'indique, la vitesse est la caractéristique qui définit cet archétype.

Le défi de la latence humaine se voit dans le cas d'utilisation de l'optimisation de l'expérience client. Dans les applications telles que l'e-commerce, la vitesse a un impact direct sur l'expérience de l'utilisateur ; les sites Web optimisés pour la vitesse avec l'infrastructure locale se traduisent directement en une hausse des vues de pages et des ventes.

Sensibilité à la latence humaine
Google a constaté qu'ajouter un retard de 500 millisecondes aux temps de réponse des pages entraînait une baisse de 20 % du trafic, alors que Yahoo a observé qu'un retard de 400 millisecondes entraînait une baisse du trafic de 5 à 9 %.

Cet effet s'étend aussi au traitement des paiements. Amazon a constaté qu'un retard de 10 millisecondes dans le traitement des paiements entraînait une baisse de 1 pour cent des revenus obtenus. En moyenne, l'approbation centralisée par mot de passe prenait 7 secondes. Un déplacement vers un traitement local a diminué le temps à 600 millisecondes, une amélioration de 6400 millisecondes, chaque tranche de 100 millisecondes pouvant augmenter de 1 % des revenus obtenus.

Un autre exemple émergent d'application de la sensibilité à la latence humaine est le traitement du langage naturel. La voix sera probablement la principale forme d'interaction avec les applications informatiques quotidiennes à l'avenir. Le traitement automatique du langage naturel pour Alexa et Siri se fait actuellement sur le cloud. Cependant, alors que le volume d'utilisateurs, d'applications et de langues pris en charge augmente, il sera nécessaire de rapprocher ces capacités des utilisateurs.

D'autres cas d'utilisation de latence humaine identifiés incluent la vente au détail intelligente, telle que les magasins Amazon Go sans caissier, ainsi que les technologies immersives telles que la réalité augmentée, où de petits retards de latence peuvent faire la différence entre le plaisir et la nausée. Dans chaque cas, les retards de livraison des données affectent directement l'expérience technologique de l'utilisateur, comme avec le traitement automatique du langage naturel et la réalité augmentée, ou les ventes et la rentabilité d'un commerçant, comme avec l'optimisation Web et la vente au détail intelligente. Tandis que ces cas d'utilisation se multiplient, le besoin de hubs de traitement des données locaux en fera de même.

Archétype Trois: Sensibilité à la latence machine à machine

Bande passante	Latence	Disponibilité	Sécurité
Moyenne	Élevée	Élevée	Élevée

L'archétype sensible à la latence machine à machine couvre les cas d'utilisation où les services sont optimisés pour la consommation de machine à machine. Comme les machines peuvent traiter les données beaucoup plus vite que les humains, la vitesse est la caractéristique principale de cet archétype. Les conséquences de l'impossibilité de fournir les données aux vitesses requises peuvent être encore pires dans ce cas que dans l'archétype de la sensibilité à la latence humaine.

Par exemple, les systèmes utilisés dans les transactions financières automatisées, telles que les transactions sur les matières premières et les actions, sont sensibles à la latence. Dans ces cas, les prix peuvent changer en quelques millisecondes et les systèmes n'ayant pas les dernières données quand il le faut ne peuvent pas optimiser les transactions, transformant des gains potentiels en pertes.

Sensibilité à la latence machine à machine
D'après une étude du groupe Tabb, un courtier pourrait perdre **pas moins de 4 millions $ en revenus par milliseconde** si sa plateforme de transaction électronique avait 5 millisecondes de retard par rapport aux concurrents.

La technologie smart grid fait également partie de cet archétype. Cette technologie est déployée sur le réseau de distribution électrique pour équilibrer automatiquement l'offre et la demande et gérer l'utilisation d'électricité de manière durable, fiable et économique. Elle permet aux réseaux de distribution une autoguérison, l'optimisation des coûts et la gestion des sources d'alimentation intermittentes, en supposant que les bonnes données sont disponibles au bon moment.

Il existe d'autres applications sensibles à la latence machine à machine, notamment les systèmes de sécurité intelligents qui se basent sur la reconnaissance des images, les simulations de forces armées et l'analyse en temps réel.

Archétype Quatre: Critique pour la vie humaine

Bande passante	Latence	Disponibilité	Sécurité
Moyenne	Élevée	Élevée	Élevée

L'archétype critique pour la vie humaine inclut les cas d'utilisation ayant un impact direct sur la santé et la sécurité humaines. Dans ces cas d'utilisation, la vitesse et la fiabilité sont essentielles.

Les meilleurs exemples de l'archétype critique pour la vie humaine sont probablement les véhicules et les drones autonomes, qui offrent de grands avantages quand ils fonctionnent comme prévu ; cependant, s'ils prennent de mauvaises décisions, ils peuvent mettre en danger la santé des humains.

Les véhicules autonomes ont progressé plus vite que beaucoup le pensaient, de nombreuses entreprises automobiles et de technologie testent déjà activement des véhicules aujourd'hui. La plupart de ces véhicules ont un humain dans le siège conducteur, prêt à remplacer les commandes automatiques en cas de problèmes, afin de réduire le risque pour la santé humaine. Mais dans un avenir proche, des véhicules de livraison et systèmes de transport sans chauffeur seront sur la route. Si ces systèmes n'ont pas les données dont ils ont besoin quand ils en ont besoin, les conséquences peuvent être désastreuses.

Il en va de même pour les drones. Nous pourrions facilement entrevoir un avenir où des centaines de drones de livraison volent au-dessus d'une ville à tout moment.

Critique pour la vie humaine
Les grandes entreprises d'e-commerce et de livraison de paquets telles qu'Amazon et DHL font déjà des expériences avec des drones pour la livraison de paquets.

L'utilisation accrue de la technologie dans les soins représente aussi un archétype critique pour la vie humaine. Les dossiers médicaux électroniques, la cyber-médecine, la médecine personnalisée (cartographie du génome) et les dispositifs d'auto surveillance modifient les soins et génèrent d'énormes volumes de données.

D'autres exemples incluent les transports intelligents et les robots autonomes. Les industries des transports et de logistique recherchent des solutions axées sur les données afin d'améliorer la sécurité des conducteurs et des passagers, l'efficacité du carburant et la gestion des actifs. La technologie dans cette espace inclura les systèmes de transport intelligents, la gestion de la flotte et la télématique, les systèmes de guidage et de contrôle, les applications de divertissement et commerciales pour les passagers, les systèmes de réservation, de taxes et de billets, ainsi que les systèmes de sécurité et de surveillance.

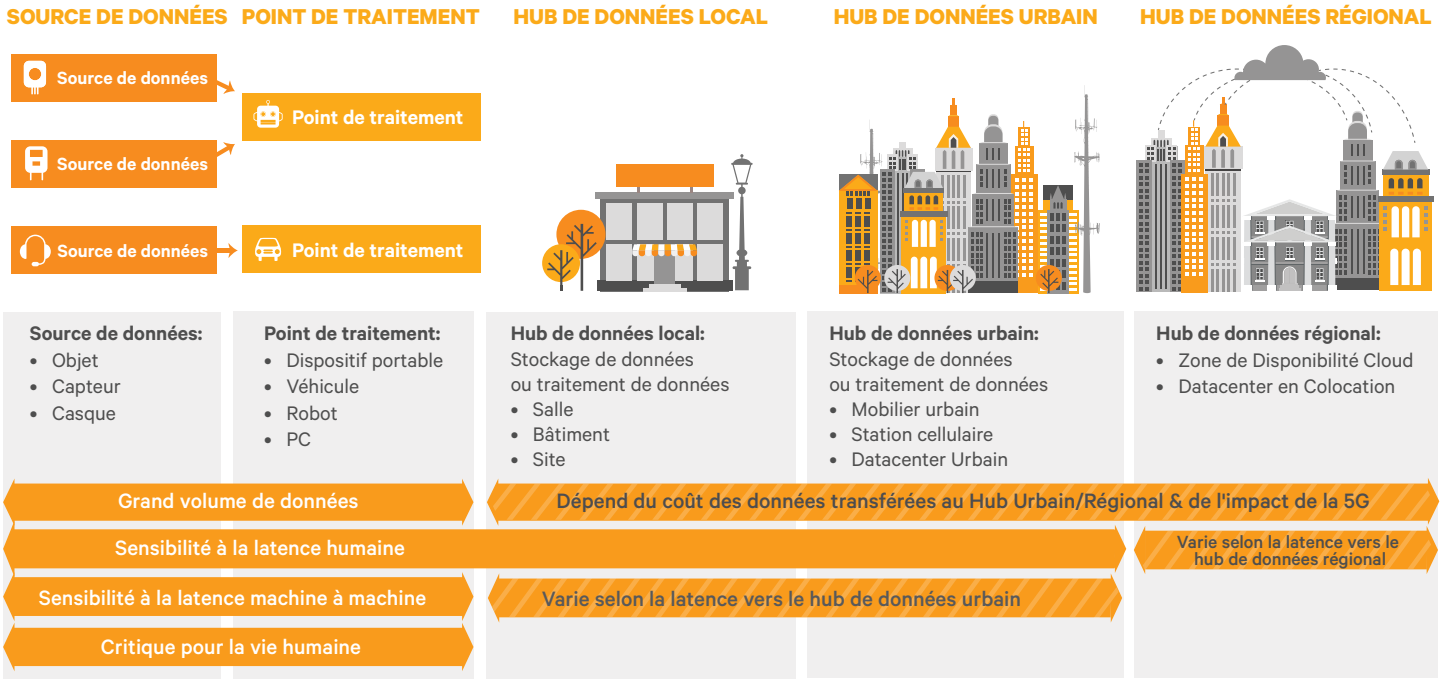
Exigences technologiques pour les hubs locaux et régionaux

L'infrastructure requise pour prendre en charge ces cas d'utilisation actuels et établis comprend quatre couches de stockage et d'informatique en plus de l'infrastructure de communication requise pour déplacer les données entre les couches.

À la source, il y a généralement un appareil qui génère ou consomme des données et un point de traitement. L'appareil peut être un capteur qui surveille tout, l'état d'alimentation d'une lampe, l'accès à une porte, la température d'une pièce ou d'autres informations souhaitées. Le point de traitement peut être aussi simple que le PC ou la tablette sur lequel un consommateur diffuse une vidéo, ou pourrait être les microprocesseurs intégrés dans les automobiles, les robots ou dispositifs portables. Ces composants dépendent de l'application et sont généralement intégrés par le fabricant de l'équipement ou ajoutés aux appareils existants.

Chaque archétype, à l'exception de l'archétype critique pour la vie humaine, pourrait être placé dans le hub de données local en fonction de l'application. Le hub de données local fournit un stockage et un traitement à proximité de la source. Dans certains cas, le hub local peut être un datacenter autonome. Plus couramment, ce sera un système en baie ou en rangée fournissant 30-300 kW de capacité dans un boîtier intégré pouvant être installé dans tout environnement.

Ces systèmes de boîtiers en baie et en rangée intègrent la communication, le calcul et le stockage avec une protection de l'alimentation, des contrôles environnementaux et une sécurité physique appropriés. Pour les archétypes



nécessitant un haut niveau de disponibilité, tels que la sensibilité à la latence machine à machine et Critique pour la vie humaine le hub local doit inclure des systèmes d'alimentation de secours redondants et être équipé pour une gestion et un suivi à distance. De nombreux cas d'utilisation requièrent aussi le cryptage des données et d'autres fonctions de sécurité dans le hub local.

Pour tous les archétypes à l'exception de Critique pour la vie humaine, le hub urbain et/ou régional pourrait être utilisé pour prendre en charge les cas d'utilisation en fonction des coûts de transfert des données, de la bande passante permise par le déploiement de la 5G et de la latence avec l'emplacement du datacenter physique. Le hub urbain tire parti de l'infrastructure de télécommunications établie pour fournir des capacités d'informatique et d'infrastructure. Il sera conçu selon les normes de télécommunications, y compris l'alimentation CC et le refroidissement à l'air libre, supportant une plage de températures et d'humidité bien plus large qu'habituellement dans les datacenters traditionnels. Le hub régional pourrait être un datacenter sur le cloud ou en colocation fonctionnant dans la même région que les hubs local et urbain.

Pour les hubs urbain et local, des conceptions modulaires capables de s'adapter facilement au-delà des caractéristiques de conception initiales doivent être envisagées afin de tenir compte des pics de demande inattendus. Ces installations doivent aussi être conçues pour s'adapter en termes d'intensité. Les applications sensibles aux images, telles que la réalité virtuelle, ainsi que les applications à traitement intensif, telles que l'analyse et l'apprentissage machine, nécessiteront probablement des densités de baies dépassant la conception 10 kW habituelle. Dans pratiquement tous les cas, ces hubs doivent fournir un niveau identique ou supérieur au hub local en matière de redondance et de sécurité.

Pour aller plus loin

En identifiant les besoins de charge de travail pour les vingt-quatre cas d'utilisation traités, quatre archétypes principaux ont émergé, pouvant orienter les décisions relatives aux exigences d'infrastructure et de configuration pour les cas d'utilisation analysés, ainsi que ceux qui apparaîtront au cours des prochaines années. Vertiv se basera sur ce travail initial sur les archétypes afin de définir des exigences et configurations technologiques spécifiques pour chaque archétype.

VertivCo.fr | **Vertiv France,** Bâtiment Liège, 1 Place des Etats Unis 94150, Rungis, France

SL-11490-FR (R04/18)