Les tendances et les bénéfices du « Edge Computing »

Livre blanc n° 226

Révision n°0

par Steven Carlini

Synthèse analytique

L'utilisation d'Internet a tendance à évoluer vers des contenus gourmands en bande passante et un nombre croissant de «choses» associées. Parallèlement, les réseaux de télécommunication mobile et les réseaux de données convergent vers une architecture de cloud computing. Pour répondre aux attentes d'aujourd'hui et de demain, la puissance informatique et le stockage sont déplacés en périphérie du réseau afin de faire diminuer le temps de transport des données et d'augmenter la disponibilité. Le «edge computing» permet de rapprocher de l'utilisateur ou de la source de données, les contenus gourmands en bande passante et les applications sensibles à la latence. Ce livre blanc présente les facteurs du «edge computing» et explore les différents types de «edge computing» disponibles.

APC by Schneider Electric White Papers font désormais partie de la bibliothèque de livres blancs Schneider Electric produite par le Data Center Science Center de Schneider Electric.
DCSC@Schneider-Electric.com

Schneider Electric

Définition du « edge computing »

Le « edge computing » rapproche de l'utilisateur final, l'acquisition des données et les fonctions de contrôle, le stockage des contenus gourmands en bande passante et les applications. Il est placé à un point final logique d'un réseau (Internet ou réseau privé), dans le cadre d'une architecture plus importante de cloud computing.

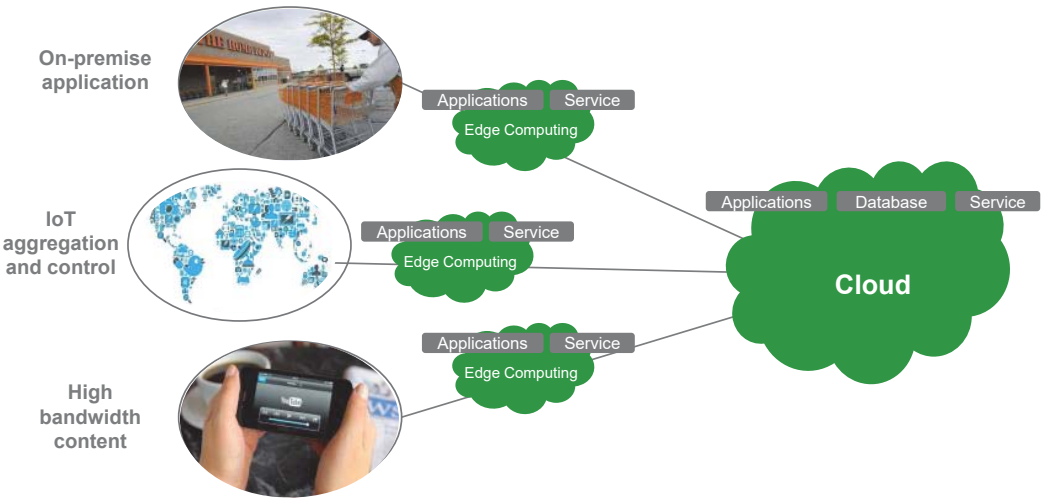

Figure 1

Diagramme de base du cloud computing avec des appareils de pointe (edge)

Le « edge computing » comporte trois applications principales que nous aborderons dans ce livre blanc.

1. Un outil visant à rassembler de nombreuses informations provenant de « choses » locales, et qui sert de point d'agrégation et de contrôle.
2. Un stockage local et une solution de livraison de contenu gourmand en bande passante dans le cadre d'un réseau de distribution de contenu.
3. Un outil d'application et de production pour répliquer les services de cloud et isoler le datacenter du cloud public.

Mais avant d'aborder les applications et les solutions, expliquons comment la mise en réseau et Internet fonctionnent

Comment Internet fonctionne

Transmission de données « est-ouest »

Les données source sont converties en paquets qui sont transmis sur le réseau via le protocole de mise en réseau IP (Internet Protocol). Le routage Internet est géré par un autre protocole appelé BGP (Border Gateway Protocol). Internet a été conçu pour survivre à des pannes massives et contourner les problèmes. Le protocole BGP n'anticipe pas le temps nécessaire au routage des données. Il se contente d'observer le nombre de sauts entre deux réseaux qui tentent de communiquer. Ces sauts peuvent être véritablement encombrés ou le routage peut emprunter une route physiquement longue, avec moins de sauts plutôt qu'une route très courte et de nombreux sauts. La **Figure 2** présente une carte des nombreux sauts longue distance aux États-Unis.[1] Si le protocole BGP fonctionne particulièrement bien en termes de fiabilité et constitue une technologie fondamentale sur laquelle repose Internet, il est vraiment sous-optimal du point de vue des performances en matière de latence (retards, gigue et gel d'image).

[1] http://conferences.sigcomm.org/sigcomm/2015/pdf/papers/p565.pdf

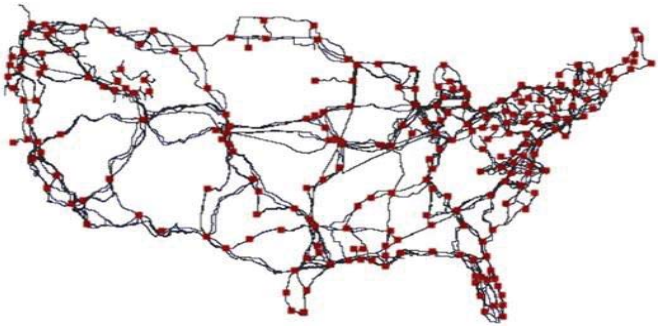

Figure 2

Carte des différents sauts de réseau aux États-Unis

Transmission de données « nord-sud »

Comme le montre la **Figure 3**, de l'intérieur du réseau de datacenter de cloud typique vers l'extérieur, le flux des données part d'une interface de serveur physique à travers des commutateurs empilés horizontalement ou des commutateurs en rangées. Au départ de chaque commutateur empilé, les données passent à travers un commutateur d'agrégation et les commutateurs d'agrégation acheminent les données à travers un commutateur principal qui est la principale entrée et sortie du datacenter. Chacun de ces commutateurs transfère les données et est considéré comme un saut de réseau avec ses ralentissements de données associés et la possibilité d'encombrement du réseau. En cas de sollicitation trop importante de n'importe quelle couche du réseau (c.-à-d. que la bande passante n'est pas dimensionnée pour la puissance de sortie de crête), il y a un autre risque de ralentissements supplémentaires pendant ces périodes d'intense utilisation.

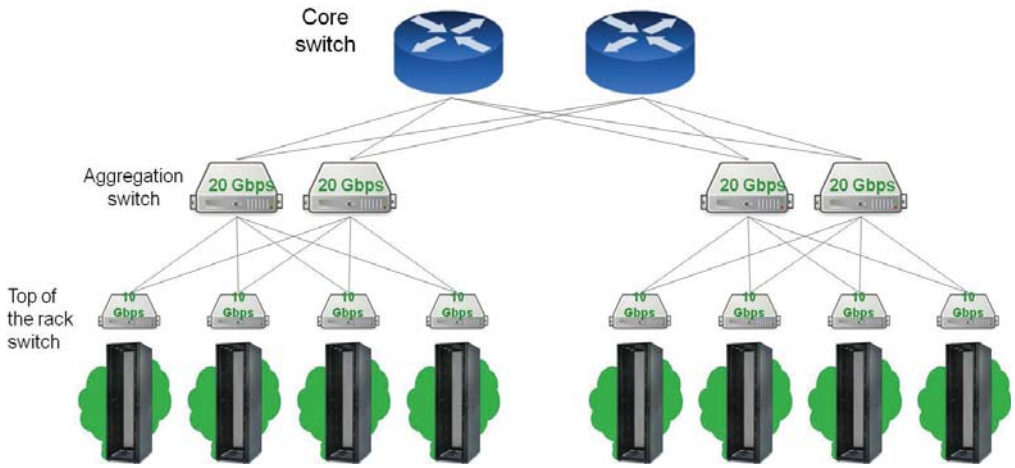

Figure 3

Réseau du datacenter

Application n° 1 : distribution de contenu haute définition

La latence est le temps entre le moment auquel un paquet de données est transmis et le moment auquel il atteint sa destination (un sens) et qu'il revient (aller-retour). Bien que la plupart des données ne voyagent que dans un sens, cela est pratiquement impossible à mesurer. C'est la raison pour laquelle le temps de l'aller-retour à partir d'un point est la méthode de mesure la plus fréquemment utilisée pour la latence. Les latences aller-retour inférieures à 100 millisecondes (ms) sont typiques et l'objectif visé est une latence inférieure à 25 ms.

La bande passante désigne la vitesse de transmission des données sur le réseau. Les vitesses maximales des équipements de mise en réseau sont publiées par leurs fabricants. Toutefois, la vitesse réelle obtenue dans un réseau donné est pratiquement toujours inférieure à la puissance de pointe. Une latence excessive crée des embouteillages qui empêchent les données de remplir le réseau au maximum de sa capacité. L'impact de la latence sur la bande passante réseau peut être temporaire (durée de quelques secondes), comme un feu tricolore, ou constante, comme un pont à une file seulement. La plus grande probabilité d'encombrement du réseau est due à un contenu vidéo à large bande passante. Comme nous le voyons sur la **Figure 4**, la VoD, la TV 4K et la diffusion vidéo sont les applications à large bande passante dont la croissance est la plus rapide[2].

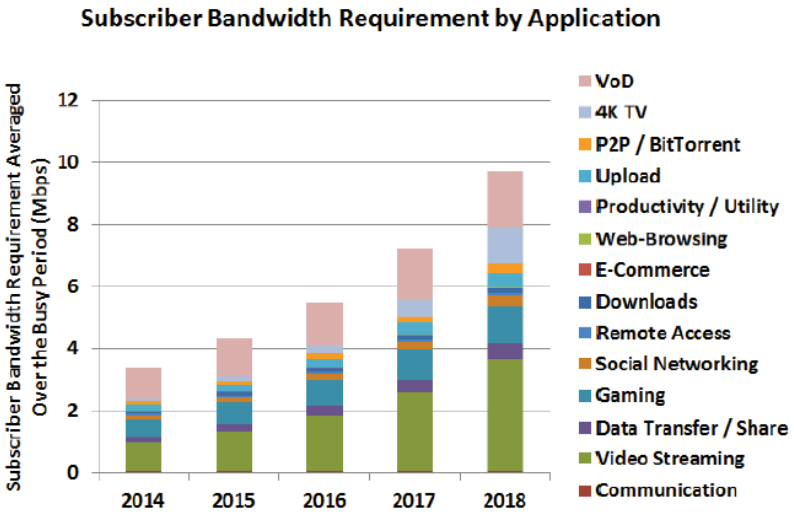

Figure 4

Croissance des applications à large bande passante

Pour soulager l'encombrement du réseau afin d'améliorer la diffusion du contenu à large bande passante aujourd'hui et à l'avenir, les prestataires de services interconnectent un système d'ordinateurs sur Internet qui met en mémoire cache le contenu, plus proche de l'utilisateur. Cela permet au contenu d'être déployé rapidement à destination de nombreux utilisateurs en dupliquant le contenu sur plusieurs serveurs et en dirigeant le contenu vers des utilisateurs en fonction de leur proximité. Ces ordinateurs utilisés pour mettre en mémoire cache du contenu sont un exemple du « Edge Computing » (**Figure 5**).

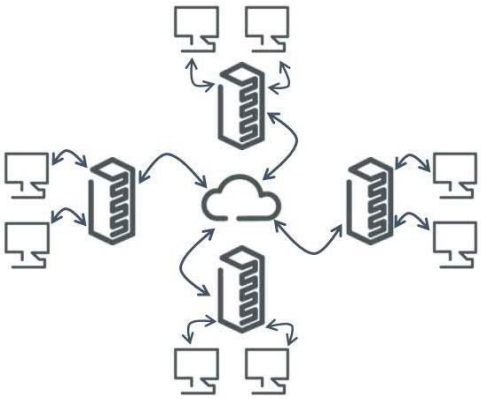

Figure 5

Diagramme d'un réseau de distribution de contenu simple (CDN)

[2] ACG Research, The value of content at the edge (La valeur du contenu à la périphérie), 2015, p.4

Application n° 2 : Le « Edge Computing » en tant que point de contrôle et d'agrégation de l'Internet des objets (IoT)

Les technologies qui seront derrière le caractère « intelligent » de tout – villes, agriculture, véhicules, santé, etc. – à l'avenir, requièrent le déploiement en masse de capteurs de l'Internet des objets (IoT). Un capteur IoT est un nœud ou un objet autre qu'un ordinateur avec une adresse IP connectée à Internet.

Dans la mesure où le prix des capteurs est en constante baisse, le nombre d'objets IoT connectés va exploser. Selon Cisco, l'Internet des objets devrait représenter 50 milliards d'appareils connectés à Internet d'ici 2020[3] L'Internet des objets permet d'automatiser l'exploitation en :

- Récoltant automatiquement des informations au sujet des équipements physiques (machines, équipements, appareils, installations, véhicules) pour en surveiller l'état ou le comportement.
- Utilisant ces informations pour bénéficier d'une visibilité et d'un contrôle pour optimiser les processus et les ressources.

Le concept de Machine to Machine, en anglais, (M2M) fait référence aux technologies qui permettent aux systèmes avec et sans fil de communiquer avec d'autres appareils du même type. Le M2M est considéré comme une partie intégrante de l'Internet des objets et il apporte plusieurs avantages à ce secteur et cette activité, en général, dans la mesure où il trouve de nombreuses applications dans la ville intelligente (Smart City, en anglais).

L'Internet industriel des objets (IIoT) qui implique la valorisation des données des capteurs, le contrôle des communications machine-to-machine et les technologies d'automatisation, génère de grandes quantités de données et de trafic réseau. Les systèmes informatiques industriels propriétaires et les technologies de mise en réseau migrent vers des systèmes informatiques commerciaux conventionnels qui communiquent via les réseaux IP (Internet Protocol).

L'exploration pétrolière et gazière est un exemple de cette application IIoT. Une flotte de drones appelés « robots de collecte aérienne de données » permet d'examiner les sites de travail pendant l'exploration pétrolière et génère de grandes quantités de données sous forme d'une vidéo haute définition. Ces sites de travail sont difficiles à coordonner avec les nombreuses flottes de camions énormes, de grues et d'excavatrices rotatives. Les méthodes plus anciennes de gestion du trafic avaient recours à des hélicoptères pilotés par des humains pour la surveillance vidéo. Les drones peuvent photographier les sites de travail 24 heures sur 24 ce qui permet aux responsables de site de bénéficier d'une vue de pointe de la manière dont leurs ressources sont déployées. Le fait d'opter pour le « edge computing » permet aux drones de transmettre les données en temps réel et d'obtenir des instructions en temps opportun.

Figure 6

Exploration pétrolière et gazière : les drones collectent d'énormes quantités de données sur les champs pétrolifères et utilisent le « edge computing » pour permettre le transfert de données en temps réel et les instructions de mouvement

[3] Dave Evans, The Internet of Things (L'Internet des objets) : How the Next Evolution of the Internet Is Changing Everything (Comment la prochaine évolution d'Internet change tout), Cisco Internet Business Solutions Group, p. 3

Application n° 3 : Applications sur site

Le besoin de conserver ou d'augmenter la disponibilité de l'informatique et des réseaux est pratiquement toujours une préoccupation de premier plan. Le cloud computing a toujours été une architecture centralisée. Le « edge computing » transforme le cloud computing en une architecture plus distribuée. Le principal avantage est que tout type d'interruption se limite à un seul point du réseau plutôt qu'à la totalité du réseau. Une attaque par déni de service distribué (Distributed Denial of Service, DDoS) ou une longue panne de courant, par exemple, se limitera ainsi à l'appareil de « edge computing » et aux applications locales sur cet appareil contrairement à toutes les applications exécutées sur un datacenter de cloud centralisé.

Les entreprises qui ont opté pour un cloud computing hors site peuvent profiter du « edge computing » pour de meilleures redondance et disponibilité. Les applications critiques pour les entreprises ou les applications nécessaires pour exploiter les fonctions cœur de métier de l'entreprise peuvent être dupliquées sur site. Pour illustrer ces faits, prenons une petite ville qui a recours à un très grand approvisionnement en eau comme source principale, tel qu'illustré sur la **Figure 7**. Si cet approvisionnement en eau était interrompu à cause d'une perturbation de l'approvisionnement principal ou du réseau de distribution, une cuve d'eau d'urgence est prévue dans la ville.

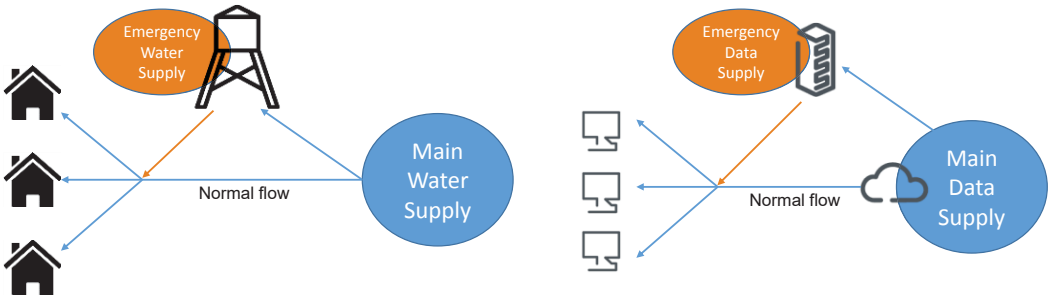

Figure 7

Un système d'alimentation en eau d'une ville, comme métaphore du « edge computing ».

Types de « edge computing »

En général, il existe trois types de « edge computing », tel qu'illustré sur la **Figure 8**.

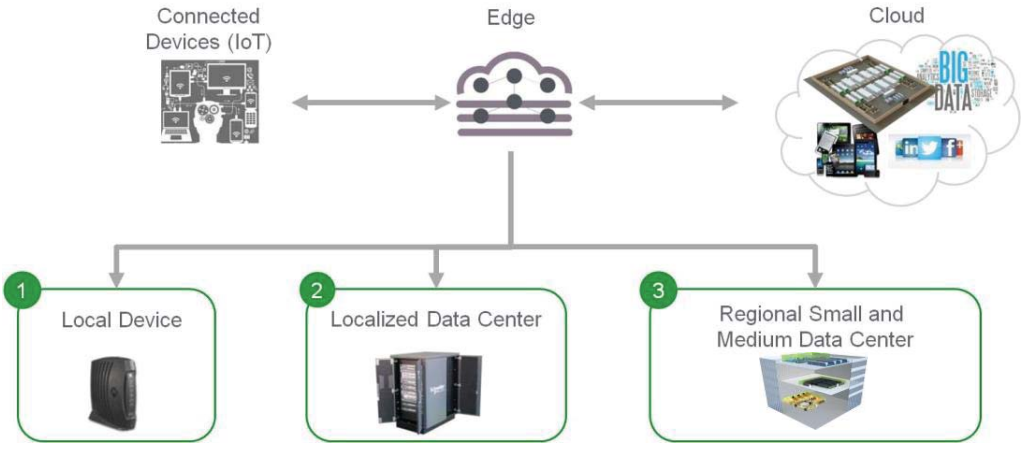

Figure 8

Types de « edge computing »

Appareils locaux :
Appareils dimensionnés pour être utilisés pour un objectif défini et spécifié. Le déploiement est « immédiat » et ils sont adaptés pour les applications domestiques ou les petites entreprises. L'exécution du système de sécurité pour le bâtiment (appareil Intel SOC) ou le stockage de contenu vidéo local sur un DVR en sont des exemples. Un autre exemple est une passerelle de stockage cloud qui est un appareil local et est généralement un dispositif réseau ou un serveur qui convertit les API de stockage cloud telles que SOAP ou REST. Les passerelles de stockage cloud permettent aux utilisateurs d'intégrer le stockage cloud dans des applications sans déplacer les applications dans le cloud à proprement parler.

Datacenters localisés (1-10 racks) :
Ces datacenters fournissent des capacités de traitement et de stockage considérables et sont rapides à déployer dans les environnements existants. Ces datacenters sont souvent disponibles sous forme de systèmes configurés à la commande qui sont préfabriqués, puis assemblés sur site, tel qu'illustré sur la **Figure 9** (gauche). Une autre forme de datacenter localisé sont les micro datacenters préfabriqués qui sont assemblés en usine, puis déposés sur site, tel qu'illustré sur la **Figure 9** (droite). Ces systèmes dans une seule enceinte peuvent être équipés dans des boîtiers robustes, qui résistent à la pluie, à la corrosion, au feu, etc. ou des boîtiers informatiques normaux pour un environnement de bureau.
Les versions à un seul rack peuvent exploiter le bâtiment, les systèmes de refroidissement et d'électricité existants, ce qui permet de réaliser des économies en termes d'investissement, plutôt que d'avoir à construire un nouveau site dédié. Il faut alors, lors de l'installation, choisir un emplacement à proximité de la source d'électricité et de fibre du bâtiment.
Les versions à plusieurs racks sont plus puissantes et flexibles de par leur taille, mais nécessitent plus de temps pour la planification et l'installation, ainsi que leur propre forme de refroidissement dédié. Ces systèmes de 1 à 10 racks sont adaptés à une multitude d'applications qui requièrent une faible latence et/ou une large bande passante et/ou une sécurité ou une disponibilité supplémentaires.

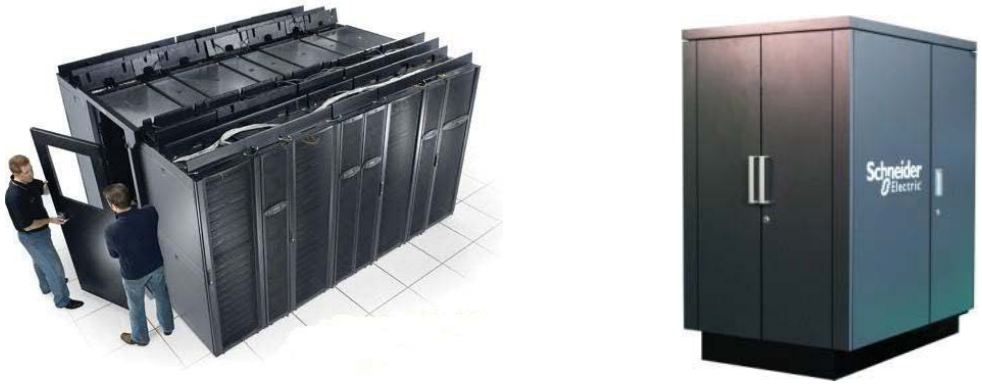

Figure 9

Un exemple de datacenter configuré sur commande (gauche) et de micro datacenter préfabriqué (droite)

Datacenters régionaux
Les datacenters qui possèdent plus de 10 racks et qui se trouvent plus près de l'utilisateur et de la source de données que les datacenters de cloud centralisés sont appelés des datacenters régionaux. Compte tenu de leur étendue, ils disposent de capacités de traitement et de stockage supérieures à celles des datacenters de 1 à 10 racks localisés. Bien qu'ils soient préfabriqués, ils sont plus longs à construire que les datacenters localisés en raison des éventuels problèmes liés à la construction, aux autorisations et à la conformité locale. Ils auront également besoin de sources d'électricité et de refroidissement dédiées. La latence dépendra alors de la proximité physique des utilisateurs et des données, ainsi que du nombre de sauts intermédiaires.

Conclusion

Le « edge computing » peut résoudre les défis liés à la latence et permettre aux entreprises de mieux tirer profit des opportunités d'exploitation d'une architecture de cloud computing. Les charges de travail générées par la vidéo en continu qui exige une grande largeur de bande provoquent encombrement du réseau et latence. Les datacenters de type « edge » permettent d'apporter le contenu gourmand en bande passante plus près de l'utilisateur final et les applications sensibles à la latence plus près des données. La puissance informatique et les capacités de stockage sont insérées directement à la périphérie du réseau pour faire diminuer le temps de transport et améliorer la disponibilité. Parmi les types de « edge computing », on note les appareils locaux, les datacenters localisés et les datacenters régionaux. Ceux qui améliorent la vitesse de déploiement et la capacité, en fonction des demandes futures des applications de l'Internet des objets sont les versions de 1 à 10 racks localisées. Il est possible de les concevoir et les déployer rapidement et facilement, avec des variantes configurées à la commande ou préfabriquées.

À propos de l'auteur

Steven Carlini est le Directeur marketing de la division Data Center Solutions chez Schneider Electric. Il est à l'origine de nombreuses solutions innovantes qui ont changé le paysage et l'architecture du datacenter tout au long de sa carrière. Il est titulaire d'un BSEE de l'Université de l'Oklahoma et d'un MBA en Affaires internationales de l'Université de Houston. Il est un expert reconnu dans le domaine et prend régulièrement la parole lors de conférences et d'événements dans le secteur du datacenter.