Life Is On | Schneider Electric

Pourquoi le Cloud Computing nous oblige à repenser la résilience du edge

Livre blanc 256

Révision 0

par Kevin Brown et Wendy Torell

Résumé général

Les entreprises ont de plus en plus recours au Cloud Computing. Une plus grande dépendance vis-à-vis des applications du cloud implique que les entreprises doivent repenser le niveau de redondance des équipements de leur infrastructure physique (alimentation, refroidissement, mise en réseau) qui restent sur site.
Dans ce livre blanc, nous décrivons et analysons les pratiques d'infrastructure physique courantes d'aujourd'hui, proposons une méthode d'analyse de la résilience nécessaire et discutons des meilleures pratiques qui garantiront que les employés resteront connectés à leurs applications professionnelles critiques.

Introduction

L'essor continu de l'Internet des objets (IoT - Internet of Things), le volume croissant du trafic internet et l'adoption toujours plus importante du Cloud computing sont des tendances technologiques essentielles qui font évoluer le paysage des datacenters.

Les grands ou très grands datacenters du cloud hébergent désormais beaucoup d'applications critiques pour les entreprises qui, à une époque, résidaient dans leurs datacenters, sur site. Toutes les applications n'ont certes pas été déplacées dans le cloud et les motifs sont divers, notamment les réglementations, la culture d'entreprise, les applications propriétaires et la latence – pour n'en citer que quelques-uns.

Par conséquent, il nous reste ce que nous appelons, dans le présent document, un « environnement de datacenter hybride ». C'est-à-dire, un environnement composé d'un mélange de (1) datacenters cloud centralisés, (2) datacenters régionaux de taille moyenne et (3) datacenters localisés, plus petits, sur site. Voir **Figure 1**.
Ce qui, à une époque, était un datacenter de 1 MW sur site, dans une succursale d'entreprise, peut désormais se composer de quelques racks d'équipements informatiques qui exécutent les applications critiques et/ou assurent la connectivité réseau avec le cloud. La diminution de l'encombrement et la capacité du datacenter sur site ne doivent pas être assimilées à une moindre criticité. En fait, dans de nombreux cas, ce qui reste sur site devient plus important.

Dans ce document, nous décrirons les pratiques courantes observées dans les trois types de datacenters susmentionnés, discuterons de la manière dont les attentes en matière de disponibilité ont évolué, proposerons une méthode pour évaluer le niveau de résilience nécessaire pour les datacenters sur site pour garantir l'atteinte des objectifs professionnels et décrirons les meilleures pratiques pour la mise en œuvre de micro datacenters à proximité de la source de données (Edge Computing).

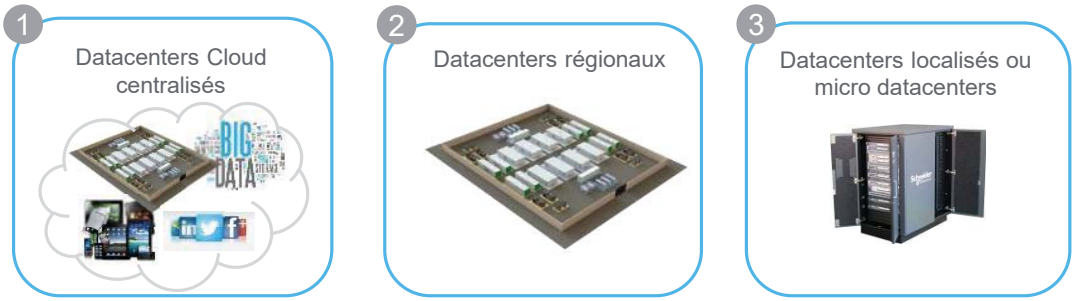

Figure 1
Les trois types de datacenters. Ce document se concentre sur les datacenters de proximité "Edge".

Types de datacenters

Le cloud centralisé a été conçu, à l'origine, pour certains types d'applications, par exemple la messagerie électronique, le personnel et les médias sociaux.
Il s'agissait d'applications dans lesquelles le timing n'était absolument pas crucial. Mais, au fur et à mesure que les applications critiques se sont déplacées vers le cloud, il est devenu évident qu'il fallait tenir compte de la latence, des limites de bande passante, de la sécurité et d'autres exigences réglementaires. Pensez à l'application des voitures autonomes. Cette application requiert une quantité importante de calculs pour fonctionner correctement, et toute latence est inacceptable, tout comme le fait que des personnes soient victimes d'accidents. La santé est une autre application critique, notamment en ce qui concerne les capteurs qui enregistrent les données des patients ou les outils chirurgicaux qui transmettent aux chirurgiens un retour d'informations per-opératoires en temps réel.

Le besoin de rapprocher la puissance de calculs au plus proche de l'usage numérique est vite apparu évident.

La distribution du contenu très haut débit est une autre application qui bénéficie du rapprochement du contenu de son point d'utilisation. Les coûts de la bande passante sont réduits et la diffusion en temps réel est améliorée.

Pour de nombreuses entreprises, il y a souvent un besoin (ou un souhait) de conserver certaines des applications critiques de l'entreprise sur site. Cela permet de conserver un plus grand niveau de contrôle, notamment en répondant aux exigences réglementaires et aux besoins en termes de disponibilité. Parfois, ces applications sont copiées dans le cloud pour une question de redondance.

Le livre blanc 226, de Schneider Electric, Les tendances et les bénéfices du « Edge Computing », explique plus en détail ces applications qui nous mènent vers un écosystème qui comprend plus de datacenters régionaux et localisés. Dans cette section, nous allons présenter chacun de ces types de datacenters et discuter des pratiques d'infrastructures physiques types déployées dans chacun d'entre-eux.

Datacenter centralisé

Les grands datacenters centralisés de plusieurs mégawatts, qu'ils fassent partie du cloud ou qu'ils appartiennent à une entreprise, sont généralement considérés comme essentiels et, en tant que tels, sont conçus en ayant toujours la disponibilité à l'esprit. Des meilleures pratiques éprouvées ont été déployées pendant de nombreuses années pour s'assurer que ces datacenters ne tombent jamais en panne. Les services généraux et le personnel informatique exploitent ces sites avec comme objectif numéro un le fonctionnement de tous ces systèmes en continu, 24h/24, 7 jours/7. De plus, ces sites sont généralement conçus et parfois homologués selon les normes de l'Uptime Institute, Tier 3 ou Tier 4. Les hébergeurs et professionnels du cloud font souvent l'éloge de ces attributs de conception de haute disponibilité comme arguments commerciaux.

Meilleures pratiques couramment observées :

- **Systèmes critiques redondants** – les systèmes d'alimentation et de refroidissement critiques sont conçus avec un dispositif de redondance (souvent 2N) pour éviter les temps d'arrêt en raison de pannes ou d'activités de maintenance.
- **Hauts niveaux de sécurité physique** – On voit souvent des capteurs biométriques au niveau des portes, ces sas de sécurité et des systèmes de vidéosurveillance et de protection 24h/24 pour garantir la sécurité des systèmes et veiller à ce que seules les personnes autorisées puissent y accéder.
- **Racks et rangées organisés** – Outre le verrouillage des racks, les câbles d'alimentation et réseau sont organisés de manière à réduire les risques d'erreur humaine : il faut limiter le risque de tirer sur les mauvais câbles, de brancher les doubles alimentations dans le même schéma de puissance, etc. La distribution de l'air est planifiée et on utilise des dispositifs tels que des joints balais et des panneaux d'obturation pour réduire les points chauds.
- **Surveillance** – Des capteurs et des compteurs sont déployés pour que les systèmes de gestion des infrastructures de datacenters (DCIM) et les systèmes de gestion technique des bâtiments (BMS) puissent gérer, contrôler et optimiser tous les systèmes de datacenters.

La **Figure 2** illustre les types de pratiques de sécurité communes à tous ces datacenters :

Capteurs biométriques

Sas

Gardes de sécurité

Figure 2
Pratiques de sécurité communes aux datacenters de colocation et de cloud centralisés

Datacenter régional

Les datacenters régionaux sont plus proches des points finaux (c.-à-d. des endroits où les données sont créées et utilisées) et plus petits que les grands datacenters centralisés. Comme décrit plus haut, ces datacenters existent pour rapprocher les applications sensibles en termes de latence ou de bande passante du point d'utilisation. Ils ont une position stratégique, afin de pouvoir gérer les besoins importants en volume. Ces datacenters peuvent être considérés comme le « pont » entre les datacenters centraux et les datacenters installés sur site.

À l'instar des grands datacenters centralisés, les datacenters régionaux sont généralement conçus avec l'esprit de la sécurité et la disponibilité. On déploie parfois des approches de conception préfabriquée et des conceptions de référence sont proposées comme point de départ (voir l'exemple sur la **Figure 3**).

Figure 3
Exemple de conception de référence comme point de départ pour l'élaboration de datacenters centralisés ou régionaux

Datacenter localisé

Un datacenter localisé est un datacenter installé au même endroit que les utilisateurs du datacenter. Plusieurs termes sont utilisés pour décrire ces datacenters, notamment **datacenter sur site** ou **micro datacenter**. La taille des datacenters localisés peut varier entre 1-2 kW et 10-20 kW. Alors que les entreprises sous-traitent de plus en plus leurs applications professionnelles à des prestataires du cloud ou de colocation, ces datacenters ont tendance à être plutôt à la petite extrémité, avec parfois juste quelques racks laissés dans une petite pièce ou un placard.

Dans bon nombre de ces datacenters moins encombrants, les pratiques de conception actuelles sont souvent équivalentes à une conception de Tier 1, avec très peu de cas accordé à la redondance ou à la disponibilité. On constate souvent les éléments suivants dans ces petits datacenters sur site :

- **Manque de sécurité** – Les salles ne sont souvent pas sécurisées ; les racks sont souvent ouverts (pas de portes)
- **Racks inorganisés** – On ne réfléchit souvent à la gestion des câbles qu'a posteriori, ce qui entraîne une accumulation de câbles, des obstructions de la circulation d'air à l'intérieur des racks et une augmentation des erreurs humaines lors des ajouts/déplacements/modifications. Voir **Figure 4**.
- **Pas de redondance** – Les systèmes d'alimentation (onduleur, bandeaux de prises) sont souvent en 1N, ce qui a un impact sur la disponibilité et la capacité à garder les systèmes en fonctionnement pendant la maintenance.
- **Pas de refroidissement dédié** – Ces petites pièces et armoires s'appuient généralement sur l'air climatisé du bâtiment, ce qui peut entraîner une surchauffe des équipements.
- **Pas de surveillance des systèmes de gestion des infrastructures de datacenters (DCIM)** – Ces salles ne sont généralement pas gérées, et ne disposent pas de personnel dédié ni de logiciels pour gérer les équipements ou garantir que les temps d'arrêt sont évités.

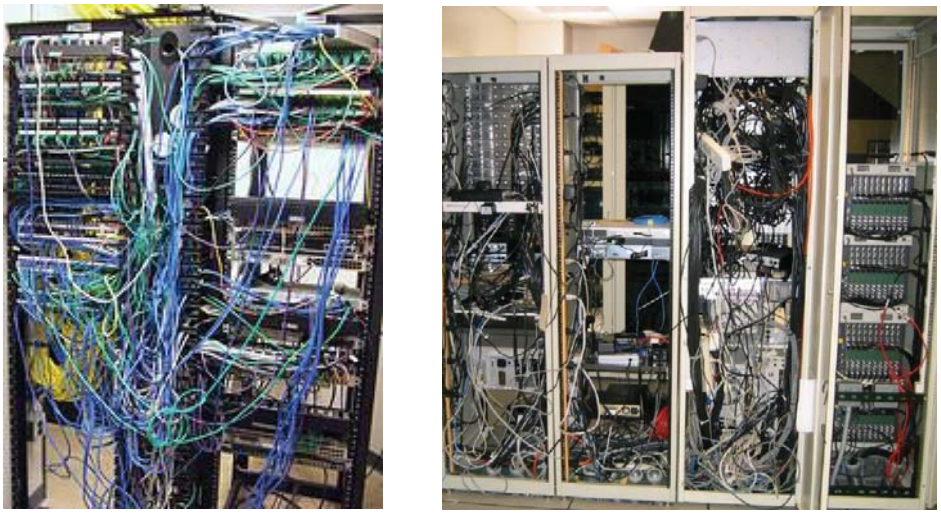

Figure 4
Exemples de petits datacenters sur site avec une mauvaise gestion et une sécurité faible.

Les sites finissent souvent par ressembler à cela dans la mesure où, au fur et à mesure que les entreprises se déplacent vers le cloud ou la colocation, les quelques racks qui restent sont considérés moins importants. On a tendance à se concentrer sur la garantie de la disponibilité des plus grands datacenters. Cette logique est toutefois bancale, dans la mesure où souvent, les racks restants sont tout autant critiques, voire plus.

Réfléchissez à ce qui reste généralement sur site : (1) applications propriétaires, critiques et (2) connectivité réseau avec le cloud. Quels sont les impacts sur la productivité de mon entreprise si je ne peux plus accéder à mes applications ? Si l'on part du principe qu'il reste le même nombre de personnes sur un site particulier dans lequel il ne reste que quelques racks, cela **augmente en fait l'importance de chaque rack**. Les équipements locaux sont critiques pour la connectivité avec les applications professionnelles de tous les jours. Vu qu'il se passe de plus de plus de choses dans le cloud, quand ce point d'accès est en panne, les employés ne sont plus productifs.

Cela suggère qu'il est nécessaire de changer la manière dont nous concevons ces petits datacenters sur site. Nous ne pouvons plus nous concentrer uniquement sur les datacenters centraux et régionaux. Il faut sans aucun doute davantage se concentrer sur les sites localisés dans la mesure où ils sont, à l'heure actuelle, les maillons faibles. Nous décrivons, plus loin dans ce document, les meilleures pratiques à déployer sur ces sites pour garantir une activité très connectée et productive.

Une mesure plus complète de la disponibilité

Dans le cadre de cet environnement hybride interconnecté, voici une question importante que nous devons nous poser : devons-nous repenser la manière dont nous parlons de criticité et de redondance ? Les outils que nous utilisons en tant qu'industrie du datacenter de nos jours se concentrent sur la manière dont je garantis qu'un datacenter est le plus robuste possible. Les niveaux « Tier » nous aident à concevoir un site particulier, afin d'atteindre un niveau de disponibilité particulier (nombre de 9). Une panne se définit par toute interruption de n'importe quel équipement informatique dans un datacenter particulier.

Les outils et les mesures n'envisagent pas la dépendance de plusieurs datacenters, le nombre d'utilisateurs concernés par une panne, la criticité des fonctions professionnelles concernées ou le basculement des applications (logiciels). Nous pensons qu'il est nécessaire d'avancer.

L'évolution des attentes en termes de disponibilité

Les attentes des employés d'aujourd'hui sont différentes de celles des anciennes générations. Alors que la main-d'œuvre vieillit et compte une plus grande part de personnes nées après l'an 2000, les attentes évoluent. Cette génération a été élevée dans une mentalité de « toujours disponible, toujours connecté », selon laquelle les appareils et les systèmes informatiques doivent fonctionner en continu. La tolérance vis-à-vis des interruptions de service est faible. La technologie est importante pour eux dans leur vie quotidienne, y compris au travail. En fait, 82 % des personnes de la génération Y pensent que la technologie disponible sur leur lieu de travail a une influence sur leur décision d'accepter un nouvel emploi.[1]

Si nous partons du principe que cette tendance va se poursuivre, il est essentiel d'envisager des manières plus holistiques de rendre compte de la résilience des datacenters qui nous donnent la visibilité nécessaire pour procéder aux modifications de conception appropriées. Comme le dit le dicton, « on ne peut gérer ce qu'on ne peut mesurer ». Les mesures de la résilience doivent évoluer pour s'adapter aux attentes professionnelles d'aujourd'hui.

Une approche différente

Un autre point de vue sur la disponibilité entraînera des actions différentes. Le **Tableau 1** illustre la comparaison entre le paradigme d'aujourd'hui (ancien) et le nouveau paradigme que nous pensons nécessaire pour prendre les mesures qui s'imposent.

[1] http://www.dell.com/learn/us/en/uscorp1/press-releases/2016-07-18-future-workforce-study-provides-key-insights (dernier accès le 31/10/2016)

Tableau 1
Décalage du paradigme de la panne de datacenter

Ancien paradigme	Nouveau paradigme
Concentré sur le datacenter centralisé	Concentré sur l'environnement hybride
On parle de panne quand l'équipement informatique en rack est touché	On parle de panne quand l'expérience utilisateur est touchée
Ne comprend pas les sites distants ou les personnes/fonctions	La criticité est touchée par le nombre d'employés et les fonctions professionnelles concernés

Pensez à une société de service public (électricité) et à la manière dont elle envisage la disponibilité. Elle ne considère pas uniquement ses installations de production et ses lignes HT (son « datacenter centralisé »). Elle taille des branches d'arbres, assure la maintenance de transformateurs montés sur des poteaux et, au final, mesure la réussite en fonction de la fourniture d'électricité à ses clients (leurs datacenters « de proximité »). L'industrie du datacenter doit évoluer en direction de ce modèle de service public, dans lequel le edge est aussi important (voire plus) que les datacenters centralisés.

La disponibilité de deux systèmes en série, ce qui signifie que vous dépendez de la disponibilité des deux, se calcule de la manière suivante :

$$\text{Disponibilité}_{\text{système}} = \text{Disponibilité}_1 * \text{Disponibilité}_2$$

Commençons par réfléchir à un seul utilisateur qui a besoin que son datacenter sur site local et son datacenter central soient disponibles ou productifs. Pour calculer la disponibilité des datacenters de son point de vue, nous utilisons donc cette formule. Si, par exemple, le datacenter central a une disponibilité de 99,98 % (datacenter de Tier 3, avec 1,6 heure de temps d'arrêt) et que son datacenter sur site a une disponibilité de 99,67 % (datacenter de Tier 1, avec 28,8 heures de temps d'arrêt), le temps d'arrêt total du point de vue de cet utilisateur-l) serait de 99,98 %*99,67 % ou 99,65 % (30,7 heures de temps d'arrêt).

Si l'on adopte le point de vue du PDG, comment évaluer l'impact de la totalité de mon écosystème de datacenters sur la productivité et la connectivité de mon entreprise ? Tous les datacenters n'ont pas besoin que tous les autres datacenters soient en fonctionnement pour fonctionner. Ainsi, une succursale de Marseille ne dépend pas d'une succursale de Lyon, mais elles peuvent toutes deux dépendre d'un datacenter central à Paris.

Tous les datacenters n'ont pas le même impact sur l'activité. Le nombre d'employés touchés est un facteur. Ainsi, un datacenter sur site de 1 000 employés pourrait être considéré plus critique qu'un de 10 employés. Le **Tableau 2** illustre la quantité de personnes-heures d'arrêt d'un exemple d'écosystème avec un datacenter central de Tier 3 et 10 datacenters localisés de Tier 1, chacun de 100 employés. Il est évident, au vu de ce tableau, que les datacenters de proximité de Tier 1 sont responsables de la totalité des temps d'arrêt. Plus le nombre de sites de proximité est important, plus le nombre d'heures de sites sans arrêt est petit.

Tableau 2

Disponibilité de 10 datacenters de proximité et de 1 datacenter centralisé, en tenant compte du nombre de personnes touchées

Disponibilité du datacenter						
Description	**Disponibilité**	**Temps d'arrêt (en heures)**	**Nombre de sites**	**Nombre de personnes par site**	**Nombre total de personnes touchées**	**Personnes-heures de temps d'arrêt par an**
Datacenter de proximité de Tier 1	99,67 %	28,82	10	100	1 000	28 820
Datacenter central de Tier 3	99,98 %	1,58	1	0	1 000	1 580
				Total personnes-heures de temps d'arrêt		30 400
					Disponibilité	99,65 %

Le tableau ci-dessus était un scénario simple avec 2 Tiers de datacenters dans lesquels 1 000 personnes sont touchées par ces deux Tiers. En présence de plus de datacenters, chacun avec des niveaux de disponibilité et des nombres de personnes touchées différents, les calculs ne sont pas aussi évidents. Cela est, en outre, incomplet dans la mesure où est exclue une évaluation de chaque site par fonction commerciale, réalisée sur les différents sites. Un site qui assure les fonctions de service à la clientèle ou de fabrication est probablement plus critique qu'un site qui héberge des administrateurs qui pourraient travailler à distance en cas de panne de leur réseau.

Analyse de la criticité

L'analyse de la criticité qualitative est une méthode éprouvée pour évaluer les risques et classer les mesures correctives (elle est également appelée « analyse des modes de défaillance, de leurs effets et de leur criticité (AMDEC) »). Cette analyse comprend la notation de la gravité des effets d'une panne avec un Indice de priorité de risque (IPR). L'IPR repose sur 3 facteurs : (1) gravité de la panne, (2) probabilité de l'occurrence et (3) détection de la panne.[2]

Nous proposons comme meilleure approche d'évaluation de tous les sites de manière holistique le recours à un tableau de bord, comme dans l'exemple du **Tableau 3**. Cela aidera les PDG et les responsables de datacenters à identifier les sites prioritaires sur lesquels concentrer les améliorations. Ce tableau de bord comporte la disponibilité et le temps d'arrêt associé de tous les sites dans l'environnement de datacenters hybrides (mesuré idéalement) et, le plus important, une note de la criticité pour chaque site. Voir l'**encadré** pour en savoir plus au sujet de l'aspect scientifique de ces notes de criticité.[2] Dans le cas de ces datacenters, la « criticité des effets de la panne » de chaque site repose sur :

- Nombre de personnes touchées
- Fonction exécutée

Une échelle de 1 à 5 est fréquente, 1 étant le plus petit impact sur l'activité en cas de panne du site et 5 le plus grand impact. Bien qu'il s'agisse d'un système de notation qualitatif, cela offre une approche systématique de l'étude de tous les sites dans l'écosystème des datacenters professionnels. Notez que, selon les entreprises, ces dernières peuvent avoir des préférences différentes pour ce qui est de l'exploitation des valeurs utilisées ici. L'essentiel est d'avoir une méthode pertinente pour évaluer tous les sites.

Dans cet exemple, il y a cinq datacenters qui composent l'écosystème hypothétique. Le temps d'arrêt annuel de chacun d'entre eux est multiplié par la note définie de « criticité des effets de la panne » de chacun pour obtenir la note pondérée.

[2] http://www.weibull.com/hotwire/issue46/relbasics46.htm (dernier accès le 31/10/2016)

À partir de là, il vous suffit de trier les sites par note, la note la plus élevée étant celle de la plus haute priorité pour les améliorations des datacenters. Vous pouvez aussi calculer le pourcentage de la note pour chaque site (comme le montre l'exemple) pour ce qui est de l'« impact du site sur la note », et les sites qui obtiennent le pourcentage le plus élevé sont ceux de plus haute priorité.

Tableau 3

Exemple de tableau de bord aidant à classer par priorités les améliorations des datacenters

Tableau de bord du datacenter					
Nom du site	**Disponibilité**	**Temps d'arrêt annuel (en heures)**	**Criticité des effets de la panne (1 à 5)***	**Score (en fonction de la criticité)**	**Impact du site sur le score**
1	99,98 %	1,752	2	3,5	0,4 %
2	99,2 %	70,08	4	280,3	30,0 %
3	99,6 %	35,04	1	35,0	3,7 %
4	98,6 %	122,64	5	613,2	**65,5 %**
5	99,98 %	1,752	2	3,5	0,4 %
			Score général de criticité	*935,6*	

Il s'agit d'une approche « step and repeat » (dupliquer et déplacer). Une fois que la disponibilité du site 4 de cet exemple est améliorée, un nouveau site apparaîtra en début de liste comme étant le plus important. Grâce à ce cycle d'amélioration continue, les sites ayant le plus d'impact seront améliorés.

Une fois la méthode de génération de rapports au sujet de la disponibilité appropriée en place, il sera évident de déterminer où les améliorations de conception sont nécessaires pour garantir la meilleure productivité et le meilleur retour sur investissement. **Dans la plupart des cas, le fait de procéder à cet exercice démontre que les datacenters du edge, qui ont souvent une disponibilité plus basse, ont souvent l'impact le plus important sur l'activité.**

Meilleures pratiques Edge Computing

Une fois les mesures et les méthodes appropriées mises en place, le besoin de repenser la conception des systèmes de datacenters de proximité deviendra évident. Les pratiques de conception typiques edge (telles que décrites plus tôt) ne sont pas appropriées en raison de la nature critique de ces sites. Les améliorations doivent se concentrer sur les domaines suivants :

- Sécurité physique
- Management (DCIM), pratiques d'exploitation, surveillance à distance
- Alimentation et refroidissement redondants
- Connectivité double réseau

Dans les sections suivantes, nous décrivons les meilleures pratiques essentielles à déployer en réponse au Edge computing. Le livre blanc 174, de Schneider Electric *Options pratiques pour le déploiement d'équipement informatique dans les petites salles serveurs et succursales*, présente plus en détail comment procéder à des améliorations réalistes de l'alimentation, du refroidissement, des racks, de la sécurité physique et de la surveillance dans les petites salles serveurs et les succursales ayant une charge informatique de 10 kW maximum.

Environnement sûr et sécurisé

Les petits datacenters locaux sont souvent placés dans une pièce très accessible, comme un espace de bureau partagé. Il n'y a souvent pas d'espace dédié, si bien que les racks ouverts ne sont pas sécurisés. Cela représente un risque en matière de sécurité, lié aux activités malveillantes ou accidentelles.

Parmi les meilleures pratiques pour réduire ces risques, on inclut :

- Le déplacement des équipements dans une pièce verrouillée ou une ou plusieurs armoires verrouillées.
- La mise en place d'un système de contrôle d'accès (biométrique ou autre).
- Pour les environnements difficiles, le placement des équipements, en sécurité, dans une armoire qui les protège du feu, des inondations, de l'humidité, du vandalisme et des effets des CEM.
- Le déploiement d'une surveillance de la sécurité et de l'environnement 24h/24 et 7 jours/7 et d'un système de vidéosurveillance

Vous trouverez des exemples d'armoires sécurisées sur la **Figure 5**. Ces dernières sont souvent préfabriquées et comportent toute l'infrastructure de support nécessaire.

Figure 5
Exemples de micro datacenters par Schneider Electric

Gestion du datacenter

Le protocole de gestion et d'exploitation est souvent différent d'un site de proximité à un autre (à condition qu'un protocole soit tout bonnement en place). La gestion de centaines ou de milliers de sites edge-of-network peut être coûteuse et prendre beaucoup de temps et la disponibilité, sur de nombreux sites, dépend de systèmes d'infrastructures partagés sur le lieu de l'installation, tels que les groupes électrogènes, les appareillages électriques et les refroidisseurs.

Parmi les meilleures pratiques pour réduire ces risques, on note :

- Faire l'inventaire des méthodes et systèmes de gestion existants.
- Consolider dans une plate-forme de surveillance centralisée tous les équipements de tous les sites.
- Déployer un système de surveillance à distance quand les ressources sont limitées. Voir le livre blanc 237, La surveillance à distance numérique et comment elle révolutionne l'exploitation et la maintenance des datacenters, pour en savoir plus sur la manière dont la surveillance à distance peut aider à réduire les temps d'arrêt.

Alimentation et refroidissement

Les systèmes d'infrastructure d'alimentation et de refroidissement (tels que les onduleurs et les climatiseurs) sont généralement déployés sur les sites de proximité (edge), sans redondance. Cela entraîne l'existence de points de défaillance, ainsi que l'incapacité à assurer la maintenance des systèmes simultanément. Dans certains cas, aucun système de refroidissement dédié n'est prévu pour les pièces en question, ce qui entraîne une surchauffe des équipements. Les systèmes d'infrastructure sont souvent partagés avec le reste du bâtiment polyvalent, si bien que la disponibilité du datacenter de proximité dépend de la disponibilité de ces ressources partagées.

Parmi les meilleures pratiques pour réduire ces risques, on inclut :

- Mesurer la température et l'humidité pour comprendre le niveau de refroidissement nécessaire (c.-à-d. le flux d'air passif, le flux d'air actif ou le refroidissement dédié).
- Envisager des schémas de puissance redondants pour pouvoir effectuer, simultanément, la maintenance sur les sites critiques.
- S'assurer que les circuits critiques sont dotés d'un groupe électrogène d'urgence.

La **Figure 6** montre un exemple de micro datacenter de Tier 3 composé d'une solution intégrée préfabriquée dans une armoire 42U unique, et d'onduleurs et d'un système de distribution électrique redondants.

Figure 6
Exemple de micro datacenter comportant 1 rack et à redondance intégrée

Connectivité réseau

Comme évoqué plus tôt, la connectivité avec le cloud est cruciale pour les sites edge. Pourtant, souvent, il n'y a qu'un prestataire de service Internet qui assure cette connexion. Cela constitue un point de défaillance. Le chaos parmi les câbles dans les armoires réseau multiplie les risques d'erreurs humaines.

Parmi les meilleures pratiques pour réduire ces risques, on inclut :

- Envisager d'ajouter un deuxième prestataire de service réseau pour les sites critiques.
- Organiser les câbles réseau avec des dispositifs de câbles pour la gestion réseau (circuits, systèmes d'acheminement, attaches, etc.).
- Étiqueter et coder avec des couleurs les lignes réseau pour éviter les erreurs humaines.

Conclusion

L'adoption du Cloud computing encourage de plus en plus d'entreprises à adopter des environnements de datacenters hybrides avec des datacenters reposant sur le cloud et sur les réseaux de proximité edge. Bien que ce qui demeure sur site occupe de moins en moins de place physique, les équipements qui restent sont plus critiques que jamais. Cela s'explique par les éléments suivants :

- Sachant qu'il y a de plus en plus d'applications qui reposent sur le cloud, la connectivité au cloud est cruciale pour la poursuite des activités de l'entreprise.
- Les employés exigent de plus en plus une technologie « disponible en continu » et ne tolèrent pas le moindre temps d'arrêt.

Malheureusement, la plupart des datacenters edge d'aujourd'hui se heurtent à des pratiques de conception inadaptées qui entraînent des temps d'arrêt coûteux. Une approche systématique de l'évaluation de la disponibilité de tous les datacenters dans un environnement hybride est donc nécessaire pour s'assurer que les investissements sont effectués là où ils permettent d'obtenir le plus grand retour sur investissement.

Une approche de tableau de bord permettant aux cadres et aux responsables d'avoir une vue holistique de leur environnement a été proposée. Elle tient compte du nombre de personnes et des fonctions professionnelles de chaque datacenter. Cette méthode identifie les sites les plus critiques sur lesquels se concentrent les investissements.

Les micro datacenters préfabriqués sont un moyen simple d'assurer un environnement sécurisé, hautement disponible adapté au edge computing. Les meilleures pratiques telles que les onduleurs redondants, un rack sécurisé ordonné, des pratiques appropriées pour les flux d'air et la gestion des câbles, un dispositif de surveillance à distance et une connectivité double réseau garantissent que les sites les plus critiques peuvent atteindre le niveau de disponibilité dont ils ont besoin.

À propos des auteurs

Kevin Brown est le Directeur de la Technologie de la division Datacenter de Schneider Electric. Il est titulaire d'une licence en génie mécanique de l'université de Cornell. Avant d'occuper son poste actuel chez Schneider Electric, Kevin a été Directeur du développement des marchés chez Airxchange, un fabricant de produits et de composants de ventilation et de récupération d'énergie du secteur HVAC (CVC - chauffage, ventilation, climatisation).
Il a précédemment occupé divers postes de direction chez Schneider Electric, notamment celui de Directeur du groupe de développement logiciel et Vice-président senior des Solutions de datacenters.

Wendy Torell est Analyste de recherche senior au sein du Data Center Science Center de Schneider Electric. À ce poste, elle étudie les meilleures pratiques en matière de conception et d'exploitation des datacenters, publie des livres blancs ainsi que des articles, et développe des outils, les TradeOff Tools, pour aider les clients à optimiser la disponibilité, l'efficacité et le coût de leur environnement de datacenter. Elle interroge également les clients sur leurs méthodes techniques en termes de disponibilité et sur leurs pratiques de conception afin de les aider à atteindre les objectifs de performances de leurs datacenters. Elle est titulaire d'un diplôme de génie mécanique de l'Union College de Schenectady, dans l'État de New York, et d'une maîtrise de l'Université de Rhode Island. Wendy Torell est ingénieur en fiabilité, certifiée par l'American Society for Quality.