Figure 1 : Vue générale d'un data center – Source : ABB.

Les data centers

PAR JEAN-PIERRE HAUET
ASSOCIATE PARTNER KB INTELLIGENCE - MEMBRE ÉMÉRITE DE LA SEE

ABSTRACT

Data centers are a backbone of the new economy. From server rooms supporting midsize business activities, to gigantic server farms hosting data and applications of major players in the "Internet economy" and "Cloud Computing", the concept of data center covers a wide spectrum of facilities whose growth is one of the most striking phenomena over the last 15 years. Their number is set to grow and especially their size and capacity, to meet the requirements of Cloud Computing, Internet of Things, Big Data... without forgetting the traditional needs of companies and public bodies.

From about 2005, it was found that the environmental footprint of data centers was a problem and that their growth could not be extrapolated.

This article discusses how data centers are designed and where are localized energy consumptions. It explains the solutions that can be today implemented to reduce consumption of auxiliary equipment and lower the PUE index as low as 1.1. It then discusses more prospectively the changes that may occur over the medium term, thanks to a substantial research and development effort, in order to reduce the data processing energy consumptions

Introduction

Les "data centers" (« centres de données » en français) constituent une épine dorsale de la nouvelle économie. Depuis les salles de serveurs supportant les activités d'entreprises de taille moyenne, jusqu'aux gigantesques fermes de serveurs où sont hébergées les données et les applications des grands acteurs de la « Net économie » et du "cloud computing" (Amazon, Facebook, Google, Apple et autres) la notion de data center recouvre un spectre d'installations très large dont la croissance est l'un des phénomènes les plus marquants de ces 15 dernières années. Nés de la bulle Internet du début des années 2000, ils en sont les survivants les plus vivaces. Leur nombre n'est pas connu de façon précise mais il est destiné à croître, ainsi surtout que la puissance des installations, pour faire face aux exigences du cloud computing, de l'Internet des objets, du Big Data… sans oublier les besoins conventionnels des entreprises et des organismes publics qui trouvent dans les data centers, qu'ils leur soient propres, dédiés ou mutualisés, un moyen d'être plus efficaces, de réduire leurs coûts et de proposer de nouveaux services.

Ces data centers ont grandi très vite et, alors que les technologies de l'information et de la communication sont supposées permettre l'avènement d'une société sobre en énergie et respectueuse de l'environnement, on s'est aperçu, à partir de 2005 environ, que leur empreinte environnementale faisait problème et que leur croissance ne pouvait pas être extrapolée en l'état.

Le présent article explique la problématique à laquelle les data centers sont ainsi confrontés et expose les solutions qui sont aujourd'hui apportées. Il aborde ensuite de façon plus prospective les évolutions qui pourraient intervenir sur le moyen terme au prix d'un effort important de recherche-développement.

Figure 2 : Racks à l'intérieur d'un container prêts à accueillir les serveurs – Source : OVH.

Figure 3 : Containers remplis de racks et de serveurs – Source : OVH.

Qu'est-ce qu'un data center ?

Un data center est un ensemble d'équipements comprenant essentiellement des serveurs informatiques, des dispositifs de stockage des données en mémoire de masse et des dispositifs de communication. Le stockage est assuré pour l'essentiel par des disques durs magnétiques ou par des disques statiques (SSD : Solid-State Drive) lorsque la recherche de performances justifie le supplément de prix. L'ensemble est rassemblé en un même lieu et comporte des équipements d'infrastructure qui viennent en soutien. Les data centers sont conçus pour les ordinateurs, pas pour les hommes. Les bâtiments qui abritent les équipements informatiques sont en règle générale aveugles, ils ont l'allure de grands hangars et la circulation d'air y est conçue plus pour le bien être des serveurs que pour celui du personnel. Certains data centers sont d'ailleurs à présent complètement automatisés et opèrent dans l'obscurité (dark data centers).

Les équipements informatiques

Les **serveurs informatiques**, appelés familièrement « boites à pizza », sont généralement d'une hauteur de façade de 1U (44,45 mm) et sont empilés dans des racks ou baies de 19" disposés le long de corridors conçus pour permettre la circulation d'air tout en ménageant les possibilités d'accès pour la maintenance (figure 1).

Figure 4 : Racks refroidis par eau à Gravelines. Source : OVH.

Le nombre de serveurs informatiques peut être impressionnant. Dans son site le plus récent de Gravelines près de Dunkerque, OVH prévoit d'installer 400 à 600 000 serveurs.

Les racks peuvent être rangés dans des containers, souvent en aluminium afin de mieux organiser la circulation d'air de refroidissement qui est alors insufflé soit à partir du double plancher, soit latéralement, et qui est extrait grâce à des ventilateurs situés en partie haute (figures 2 et 3). Un container peut contenir à lui seul plusieurs milliers de serveurs.

Le refroidissement des serveurs peut se faire par air ou par eau. Cette dernière technique (figure 4), plus délicate à mettre en œuvre, permet d'évacuer quelque 70 % de la chaleur produite par les équipements informatiques. Une circulation d'air extérieur organisée au travers des serveurs se charge du complément. Il est alors possible de renoncer à la climatisation proprement dite du bâtiment, plus facilement lorsque le data center est localisé dans une zone au climat froid ou tempéré.

Niveau	Disponibilité	Description générale
Tier I	99,67 %	28,8 heures d'interruption/an Aucune redondance
Tier II	99,75 %	22 heures d'interruption/an Redondance partielle mais alimentations électriques et climatisation non redondantes
Tier III	99,82 %	1,6 heures d'interruption/an Redondance N+1 – Alimentations électriques doublée mais fonctionnent en actif/passif
Tier IV	99,995 %	0,8 heures d'interruption/an Idem Tier III mais alimentations électriques et climatisation fonctionnent en actif/actif

Tableau 1 : Spécifications générales des niveaux de disponibilité selon l'Uptime Institute.

D'une façon générale les techniques de « refroidissement libre » ou "free cooling" se réfèrent aux solutions fondées sur l'utilisation directe du milieu extérieur (air ou eau) dans ses conditions ambiantes pour assurer en totalité ou en partie le refroidissement. Si l'air est suffisamment sec, il peut être refroidi avant admission en passant dans des chambres d'évaporation. Ces systèmes permettent de limiter l'usage des compresseurs et sont donc bénéfiques sur le plan des économies d'énergie.

Les **équipements de communication** constituent une partie essentielle des data centers. Les salles réseaux des grands data centers accueillent des routeurs de très grande capacité, capables chacun de desservir plusieurs milliers voire plusieurs dizaines de milliers de serveurs et de convoyer l'information vers les réseaux extérieurs. Pour les très grands data centers, les débits à traiter s'expriment en dizaines de térabit/s. La plupart disposent de leur propre réseau en fibre optique leur assurant une liaison directe (en « appairage » ou "peering") avec les fournisseurs de services réseaux.

Des équipements de cybersécurité sont également indispensables : pare-feu, systèmes VPN, systèmes de détection d'intrusion, etc.

La **continuité de service** est un facteur essentiel. L'organisation "Uptime Institute" classe les data centers en quatre niveaux : Tier I, II, III et IV. Ces niveaux correspondent à un certain nombre de garanties sur le type de matériel déployé dans le centre de données en vue d'assurer sa disponibilité.

Le niveau IV exige une infrastructure totalement redondante, à tolérance de panne, alimentation en énergie en mode actif/actif. Ce type de centre de traitement atteint et dépasse un taux de disponibilité de 99,99 % (soit moins de 24 minutes d'arrêt cumulé par an). Il ne nécessite pas d'arrêt des systèmes, même pour des opérations de maintenance logistique ou de remplacement d'équipements actifs.

La disponibilité, comme la sécurité, a un prix, en euros bien entendu, mais aussi en consommation d'énergie et en émissions de carbone dans la mesure où elle implique l'installation de dispositifs de redondance qui doivent, pour atteindre un très haut niveau de performances, fonctionner en 7/24/365.

Elle constitue aussi un frein important à la modernisation des data centers, en vue notamment de les rendre énergétiquement plus efficaces, car les responsables sont extrêmement hésitants à engager des transformations qui pourraient, lors de leur mise en œuvre, avoir un effet, fût-il temporaire, sur les performances du centre. La qualité de la conception initiale du data center conditionne donc fortement sa performance sur toute sa durée de vie.

Les équipements d'infrastructure

Les équipements d'infrastructure viennent en support des activités de traitement de l'information. Ils assurent des fonctions auxiliaires mais néanmoins essentielles et sont, comme on le verra, à l'origine d'une part importante des consommations d'énergie. Les systèmes les plus généralement rencontrés sont les suivants :

- systèmes d'alimentation et de conversion de l'énergie électrique, y compris dispositifs d'alimentation de secours (généralement des groupes électrogènes) ;
- systèmes de refroidissement des processeurs (à air ou à eau) permettant de maintenir les équipements informatiques aux environs de 20 à 25 °C ;
- conditionnement éventuel des bâtiments ;
- filtration de l'air ;
- éclairage ;
- installations de sécurité ;
- salles de contrôle.

La base installée

Une typologie des data centers

Les data centers peuvent être propres à une entreprise et être éventuellement localisés dans ses locaux, mais la tendance depuis quelques années est de développer de grandes infrastructures partagées accessibles par les réseaux et offrant des services à une clientèle multiple : ce sont les data centers en « colocation ». Ces services peuvent se situer à différents niveaux : fourniture de l'infrastructure seule, fourniture de la plate-forme complète ou fourniture de services. Des modèles hybrides, dans lesquels les services offerts en cloud computing viennent en renfort des ressources propres de l'entreprise,

sont possibles. La figure 5 résume les quatre modèles d'affaires usuellement rencontrés.

La segmentation du marché

Une étude américaine récente [1] distingue cinq segments sur le marché :

- Les data centers des petites et moyennes entreprises ;
- Les data centers des grandes entreprises et des groupes ;
- Les data centers partagés ;
- Les installations des grands fournisseurs de services de cloud computing (Amazon, Facebook, Apple, Google…) ;
- Les installations des opérateurs de calculs de haute performance (universités, centres de recherche).

Cette étude fournit une analyse du marché américain en 2011 selon trois critères (tableau 2) :

- Le pourcentage de serveurs se rattachant à chaque segment au sein de la base installée (soit 12,2 millions de serveurs aux États-Unis) ;
- Le taux moyen d'utilisation de ces serveurs ;
- Le pourcentage de la consommation totale d'électricité imputable à chaque segment.

Ces chiffres montrent clairement que réduire l'analyse du problème des data centers aux seuls centres gérés par les leaders de l'Internet conduirait à passer à côté d'une grande partie du problème. En effet, ces data centers, fortement médiatisés, ne représentent que quelques 7 % de la base installée et sont en outre mieux gérés que la moyenne des installations, ce qui se traduit par une responsabilité dans la consommation d'énergie totale qui n'excède pas 4 %. A contrario, les centres de dimensions plus modestes sont très nombreux et se signalent par un taux d'utilisation très faible ce qui est l'un des facteurs expliquant leur poids dans la consommation totale.

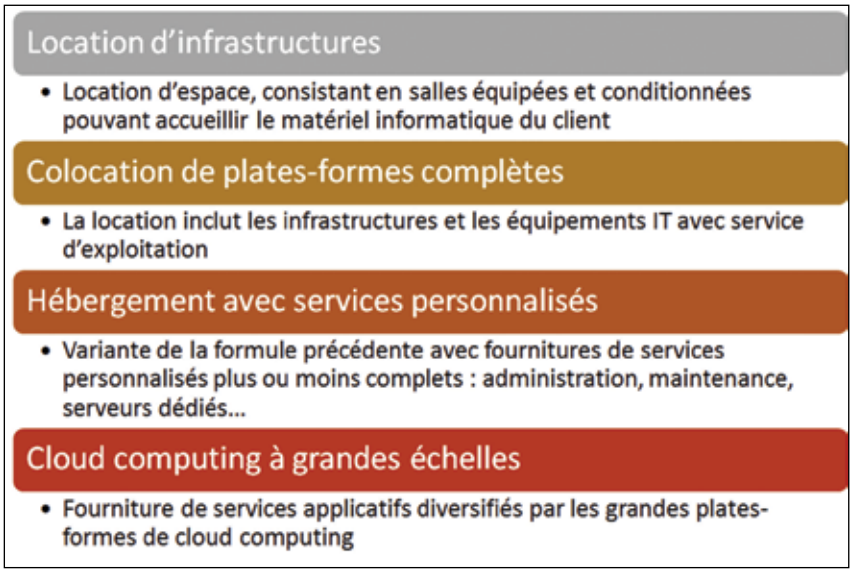

Figure 5 : Les principaux modèles d'affaires des data centers.

Etats-Unis 2011	% de la base installée de serveurs	Taux d'utilisation des serveurs	% de la consommation d'électricité totale
Petites et moyennes entreprises	40 %	10 %	49 %
Grandes entreprises	30 %	20 %	27 %
Data centers partagés	22 %	15 %	19 %
Grands fournisseurs de services	7 %	40%	4 %
Calculs de haute performance	1 %	50 %	1 %

Tableau 2 : Répartition de la base installée des serveurs des data centers aux États-Unis selon différents critères – Source : NRDC – Data Center Efficiency Assessment (2014).

Un trafic en forte croissance

Le nombre de data centers en opération dans le monde ne peut être évalué avec précision car les statistiques ne renvoient pas toujours au même concept. Certaines ne retiennent en effet que le concept de data centers partagés alors que d'autres ont une acception plus large. L'étude [1] estime le nombre de data centers en service aux États-Unis à 3 millions en 2013 ce qui pourrait correspondre à environ 6 à 7 millions au niveau mondial. Une étude publiée en 2014 par IDC [2], estime que ce nombre pourrait atteindre 8,6 millions en 2017 avant de décroître à partir de cette date.

En effet, le mouvement de concentration des data centers devrait se poursuivre, de plus en plus d'entreprises considérant qu'il est de leur intérêt de recourir à des fournisseurs de services dotés d'infrastructures dont la taille ira en croissant. Les volumes de données traitées continueront à croître à un rythme soutenu, la gestion des données et des données massives en particulier devenant l'un des axes majeurs autour duquel s'organise l'économie moderne.

Cisco [3] fait la même analyse et prévoit une croissance forte du trafic de données supporté par les data centers (+ 23 % en moyenne sur 2013-2018) avec une part de plus en plus importante

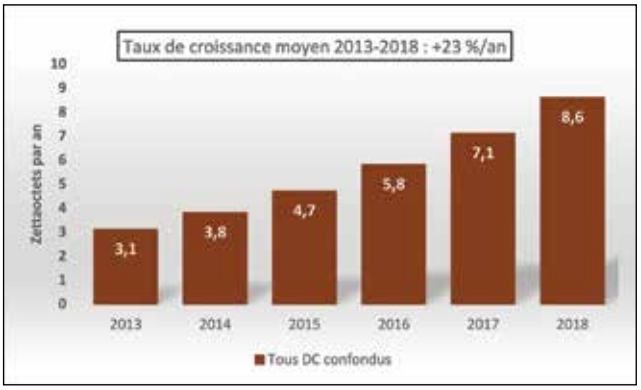

Figure 6 : Évolution du trafic supportés par les data centers de 2015 à 2018 en zettaoctets.

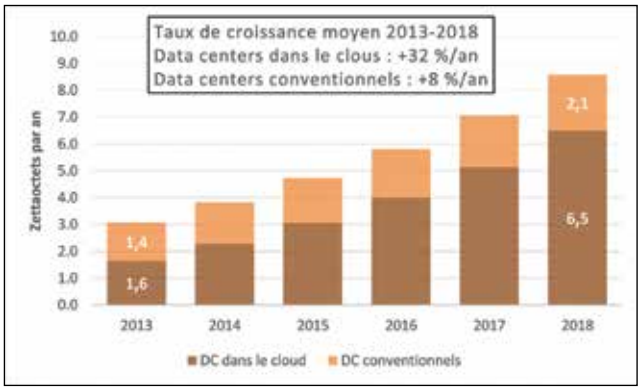

Figure 7 : Évolution du trafic supporté par les data centers de 2015 à 2018 par catégorie.

Nota : Les données des figures 6 et 7 sont en zettaoctets (1 Zo = 10^{21} octets). Elles intègrent le trafic interne aux data centers et entre data centers. Le trafic vers les utilisateurs finaux représentent environ 17 % du total – Source : Données Cisco [3].

supportée par les data centers dans le cloud (figures 6 et 7). Cette évolution vers le cloud s'accompagne d'un élargissement de la base installée exploitée en cloud public qui pourrait représenter 31 % en 2018 contre 22 % en 2013 en termes de charge supportée.

La consommation d'énergie et les émissions de carbone

Des consommations en forte croissance

Les data centers sont les usines des temps modernes et sont très consommateurs en électricité. Il a été calculé par Steve Greenberg que leur consommation d'électricité était 40 fois supérieure à celle de bâtiments de bureaux usuels [4]. Sans négliger la question de l'approvisionnement en eau et de sa gestion, le problème de l'empreinte environnementale des data centers est donc avant tout celui de leur consommation d'électricité et des émissions de CO_2 qui en résulte. Si Internet était un pays, il serait aujourd'hui le 11e ou le 12e pays consommateur d'électricité avec une consommation totale (en 2010) de l'ordre de 1,1 à 1,5 % de la consommation mondiale d'électricité. Aux États-Unis, l'étude [1] conclut que la consommation des data centers a atteint 91 TWh en 2013 soit 2 % de la consommation totale, émettant ainsi quelque 50 Mt d'émissions de CO_2.

Malgré les efforts faits pour les réduire et malgré un infléchissement par rapport aux tendances antérieures, ces consommations, et donc les émissions correspondantes de CO_2, vont continuer à croître du fait de l'expansion des data centers prévue au cours des 10 prochaines années. L'étude [1] prévoit une consommation des data centers de 140 TWh aux États-Unis en 2020 (soit le ¼ de la consommation d'électricité française) alors que la Commission européenne prévoit 104 TWh en 2020 (contre 56 en 2007) [5]. La Commission a émis en 2008 un « Code de bonne conduite pour les data centers » [5] qui est une initiative volontaire à laquelle les exploitants de data centers peuvent souscrire pour ramener leur consommation dans les limites acceptables. 107 participants adhèrent aujourd'hui à cette action pilotée par le Centre de recherche commun – Institut pour l'énergie et les transports (JRC-IET)[1].

En 2012, l'association Greenpeace a publié un rapport "*How clean is your cloud*" [6] appelant l'attention sur les consommations et les émissions générées par les data centers et aussi sur les pratiques de certains exploitants affichant des résultats contestables et pratiquant à peu de frais le "greenwashing" de leurs installations. Ce rapport a contribué à accroître la sensibilisation au problème des data centers et a notamment amené Apple à réagir en annonçant en mars 2013 qu'il s'engageait à alimenter à 100 % ses data centers en énergie renouvelable. Google lui a emboîté le pas en annonçant que les consommations de son centre finlandais d'Hamina seraient entièrement couvertes par un contrat de 10 ans passé avec le fournisseur d'électricité d'origine éolienne O2.

En fait ces annonces peuvent induire en erreur. Les data centers sont en permanence en fonctionnement, de jour comme de nuit et toute l'année. Ils constituent un débouché idéal pour les centrales fonctionnant en base, le nucléaire en particulier, mais ne peuvent évidemment pas se satisfaire de productions intermittentes et aléatoires. La couverture dont il est fait état, doit donc s'entendre au sens de « neutralité carbone » ou de « bâtiment à énergie positive », notions qui font intervenir des bilans sur l'année, sans se soucier de l'ajustement à tout instant de la ressource par rapport au besoin.

[1] Voir http://iet.jrc.ec.europa.eu/energyefficiency/ict-codes-conduct/data-centres-energy-efficiency

Sans réfuter pour autant l'apport des énergies renouvelables, on voit bien que le problème de fond des data centers est celui de la réduction de leur consommation d'énergie afin de se rapprocher de l'optimum thermodynamique dont on est aujourd'hui extrêmement éloigné. Il n'y a évidemment, comme pour toute installation industrielle, aucune solution miracle. La réduction des consommations ne peut venir que d'un ensemble de mesures dont la définition nécessite en premier lieu que l'on ait une vision claire de l'origine de ces consommations.

D'où viennent les consommations d'énergie ?

Un data center est un système complexe dans lequel beaucoup de fonctionnalités, principales ou annexes, sont à l'origine de consommations d'énergie, très majoritairement de consommations d'électricité. La répartition de ces consommations entre leurs différents postes est variable selon la taille, la localisation, l'âge des data centers et bien évidemment selon les efforts qui ont été faits pour les réduire. De façon schématique, on distingue les consommations absorbées par les équipements informatiques et celles qui sont générées par les auxiliaires, en premier lieu les systèmes de refroidissement des serveurs et la climatisation des locaux (figure 9). Cette distinction est à la base d'un indicateur de performances dénommé Power Usage Effectiveness (PUE) qui est explicité plus loin.

De façon approximative, on peut décomposer la consommation moyenne d'un data center conformément au diagramme de la figure 10. Cependant cette répartition des consommations est très approximative et, comme on le verra plus loin, les data centers les plus modernes ont fait de gros efforts pour réduire les consommations des auxiliaires, stimulés par la nécessité de

Figure 8 : Apple a investi dans de grandes fermes solaires afin de réduire son empreinte carbone. Source : Apple.

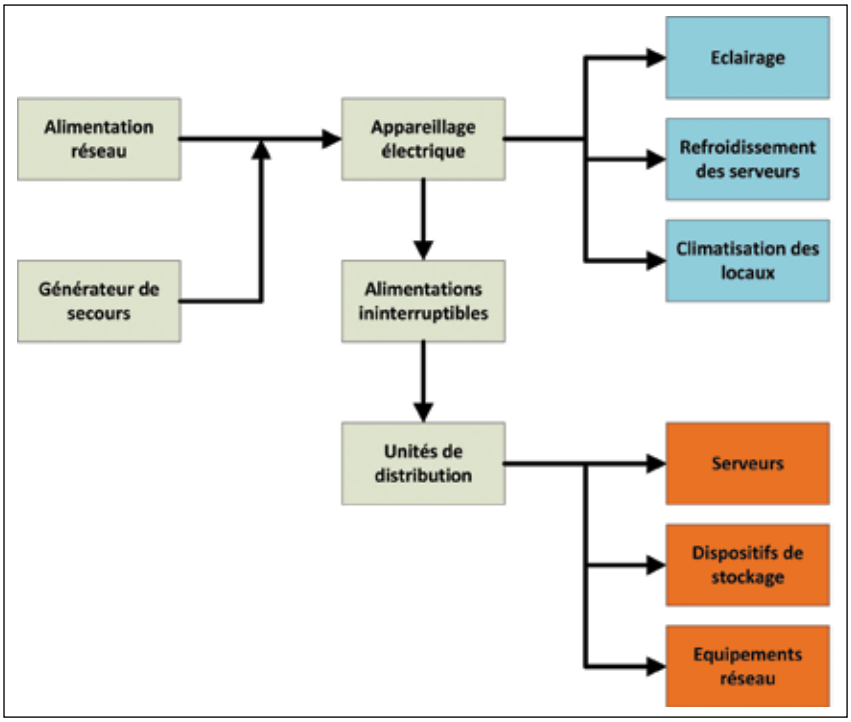

Figure 9 : Schéma des flux d'énergie électrique dans un data center.

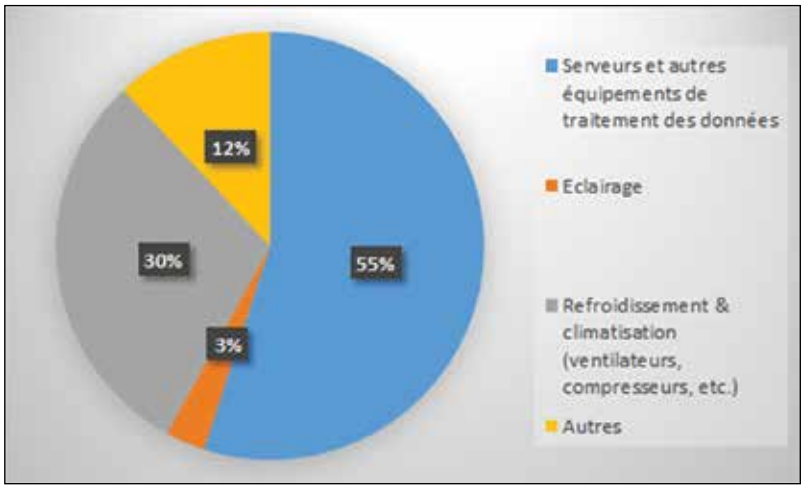

Figure 10 : Répartition approximative des consommations d'électricité dans un data center. Source : d'après EPA [8].

« verdir » leurs installations et par le fait que le poste « Energie » représente une fraction importante de leur prix de revient, évaluée entre 20 et 40 % de leur coûts opérationnels [7].

Le Power Usage Effectiveness (PUE)

Le Power Usage Effectiveness (PUE) ratio est un concept développé par le consortium international Green Grid[2] qui est maintenant largement pris en considération comme standard d'efficacité énergétique dans le monde des data centers. C'est le ratio entre l'énergie totale consommée par le centre et la seule énergie consommée par les équipements de traitement de l'information. Un PUE de 2 signifie que pour chaque kWh consommé par les équipements informatiques, un autre kWh est consommé par l'infrastructure pour le conditionnement, l'alimentation électrique, l'éclairage, etc.

Les PUE ont progressivement décru au cours des dernières années. Cependant des PUE très supérieurs à 2 sont encore fréquents dans les data centers des entreprises de taille moyenne alors que les meilleures performances annoncées par les grands acteurs de l'Internet se situent aux environs de 1,1 (OVH annonce même 1,09 pour ses data centers les plus récents).

L'inconvénient majeur du PUE est de se focaliser sur l'infrastructure sans prendre en compte ce qui constitue la raison d'être des data centers, c'est-à-dire les traitements informatiques. C'est un indicateur destiné plus aux responsables de la logistique et de la maintenance qu'aux responsables informatiques. Un autre indicateur a donc été proposé conjointement par l'Uptime Institute et McKinsey afin de prendre en compte de façon plus holistique l'ensemble des consommations d'un data center. Il s'agit du CADE : Corporate Advantage Data Center Efficiency.

Cet indicateur combine dans une même formule l'efficacité des équipements informatiques avec celle des équipements d'infrastructure :
CADE = Facility Efficiency (FE) x Asset Efficiency (AE).

Plus récemment, au niveau européen, l'ETSI a présenté en juin 2014 un nouvel indicateur composite, le DCEM Global KPI (DCEM : Data Center Energy Management), développé au sein de l'ETSI's OEU (Operational Energy efficiency for Users) Industry Specification Group et visant à qualifier l'efficacité énergétique et environnementale des data centers.

Le DCEM Global KPI combine deux indicateurs : l'un prenant en compte différents gabarits de data centers (S, M, L ou XL), l'autre introduisant neuf niveaux de performance, comparable à l'échelle utilisée pour mesurer l'efficacité énergétique des appareils domestiques. Le DCEM Global KPI vise à permettre l'évaluation de l'efficacité environnementale des data centers et à permettre le benchmarking des data centers dans différents secteurs.

Le DCEM Global KPI est basé sur une formule définie dans le standard de ETSI ES 205 200-2-1 intégrant quatre KPI particuliers :

- Consommation d'énergie : KPI_{EC} ;
- Efficacité opérationnelle : KPI_{TE} ;
- Energie réutilisée : KPI_{REUSE} ;
- Utilisation des énergies renouvelables : KPI_{REN}.

Il reste à savoir si cette publication entraînera un mouvement d'intérêt suffisamment puissant de la part des utilisateurs pour qu'une majorité des hébergeurs s'y rallie.

Aucun indicateur n'est parfait et l'absence d'une batterie d'indicateurs universellement reconnus reste aujourd'hui un handicap pour apprécier et améliorer le niveau de performance atteint par les data centers.

Comment maîtriser les consommations d'énergie ?

Le rendement thermodynamique des data centers est aujourd'hui très faible

Les data centers sont aujourd'hui, d'un point de vue purement thermodynamique, extraordinairement inefficaces. Leur vocation est en effet de manipuler des bits d'information et l'on sait que le basculement d'un bit d'information nécessite une énergie minimale, désignée limite de Landauer, qui a été déterminée théoriquement et expérimentalement comme égale à kT x Ln(2), où k est la constante de Boltzmann et T la température du système physique considéré. A température ambiante, cette limite est d'environ 2,75 zeptojoules (2,75 10^{-21} J).

Un serveur 1U, parmi les plus performants actuellement sur le marché[3], offre une puissance de calcul de 35 téraflops en simple précision mais au prix d'une puissance consommée pouvant atteindre 1 500 W. Malgré les performances affichées, un calcul rapide montre que le rendement thermodynamique d'un tel équipement est encore à des ordres de grandeur de la limite de Landauer. Nous ne sommes donc même pas encore au stade où en était la chandelle au Moyen-âge dans le domaine de l'éclairage ! C'est dire que des progrès énormes surviendront nécessairement dans les décennies à venir, au prix de ruptures technologiques sur les processus mêmes de traitement de l'information et de stockage, ruptures que

[2] Le Green Grid est un consortium créé aux États-Unis, ayant le statut d'organisation à but non lucratif, dont l'objet est de développer, standardiser et promouvoir des méthodes et des indicateurs visant à améliorer l'efficacité énergétique dans les technologies de l'information et notamment dans les data centers – Parmi les membres du Green Grid, on trouve : Siemens, Cisco, IBM, Dell, HP, Schneider Electric, Intel. Voir : http://www.thegreengrid.org.

[3] Il s'agit en l'occurrence du serveur Dell PowerEdge C4130 optimisé pour le calcul à hautes performances.

l'on peine à décrire aujourd'hui car elles mettront nécessairement en œuvre des phénomènes physiques nouveaux.

Quand on parle d'améliorer l'efficacité énergétique des data centers, il faut donc préciser à quel horizon de temps on se place. Nous examinerons dans cette section les actions pouvant être menées sur le court terme avant de donner quelques pistes sur les évolutions à plus long terme.

Action d'ordre général : l'hygiène énergétique

Avant d'aborder les mesures spécifiques à chaque grand poste de consommation, il convient de souligner que les data centers doivent en premier lieu mettre en œuvre les mesures de gestion rationnelle de l'énergie qui sont communes à tous les secteurs industriels :

- réalisation d'un audit détaillé des postes de consommation et mise en place d'un système de suivi ;
- mise en place d'un système de management du data center (DCEM). Un tel système (figure 11), contrôlant notamment la charge des processeurs, la puissance appelée et la température des serveurs, est encore souvent perçu comme un luxe inutile alors qu'il est usuel dans l'entreprise traditionnelle ;
- éclairage basse consommation ;
- alimentations performantes ;
- moteurs à haut rendement ;
- déconnexion des serveurs inutilisés.

Ce dernier point peut paraître trivial mais dans les les faits on constate que les serveurs les plus anciens devenus inutiles du fait d'un accroissement de capacité et plus généralement les serveurs devenus « comateux » quelle qu'en soit la raison, ne sont pas nécessairement déconnectés, par négligence, ignorance ou par peur de prendre des décisions ayant des conséquences imprévues sur certains traitements. Selon l'étude NRDC [1], 20 à 30 % des serveurs en place dans les grands data centers pourraient aujourd'hui être retirés du service sans inconvénient.

Figure 11 : Centre de gestion d'un data center – Source : Décathlon/ABB.

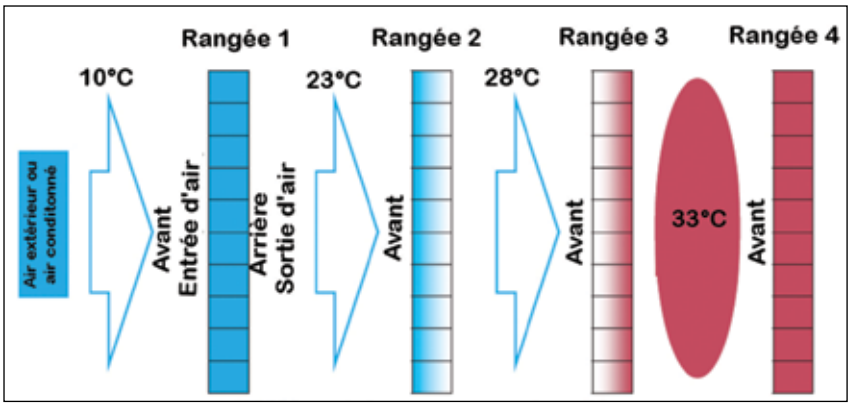

Figure 12 : Circulation d'air à éviter dans un data center – D'après EPA – Energy Star.

Ces mesures sont les "low hanging fruits" et trouvent leur rentabilité à très court terme. Une fois qu'elles sont prises, il faut passer à des dispositions plus complexes qui vont souvent mettre en cause la conception même du data center.

La réduction des consommations des auxiliaires

Réduire la consommation des auxiliaires, c'est en premier lieu réduire la consommation liée au refroidissement des serveurs et des switches.

Celle-ci se fait très généralement, en totalité ou en partie, grâce à des circulations d'air frais qui traverse les racks et permet de maintenir la température de l'environnement entre 18 et 27 °C (règle ASHRAE). L'air ambiant est aujourd'hui utilisé de façon quasi-systématique et ceci conduit à promouvoir la localisation des data centers dans les pays arctiques. L'Islande fait ainsi la promotion de ses installations, avançant que les data centers installés en Islande offrent des coûts opérationnels de plus de deux fois inférieurs à ceux situés en Europe occidentale, avec des temps de latence inférieurs à 20 ms vers l'Europe et 40 ms vers les États-Unis.

Il demeure essentiel dans tous les cas de bien organiser les circulations d'air afin que l'air frais ne vienne pas se mélanger de façon intempestive avec l'air chaud. Des configurations telles que celles de la figure 12 sont typiquement à éviter [9] .

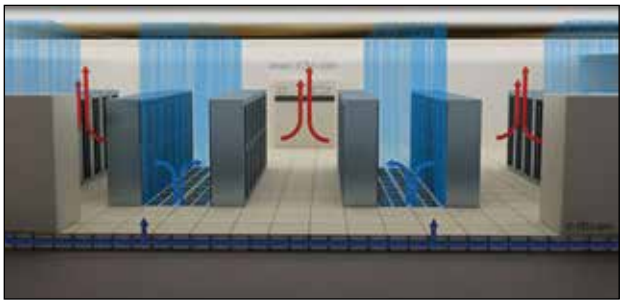

Figure 13 : Arrangement évitant, à l'aide de rideaux flexibles, le mélanges de flux - Source : EPA – Energy Star.

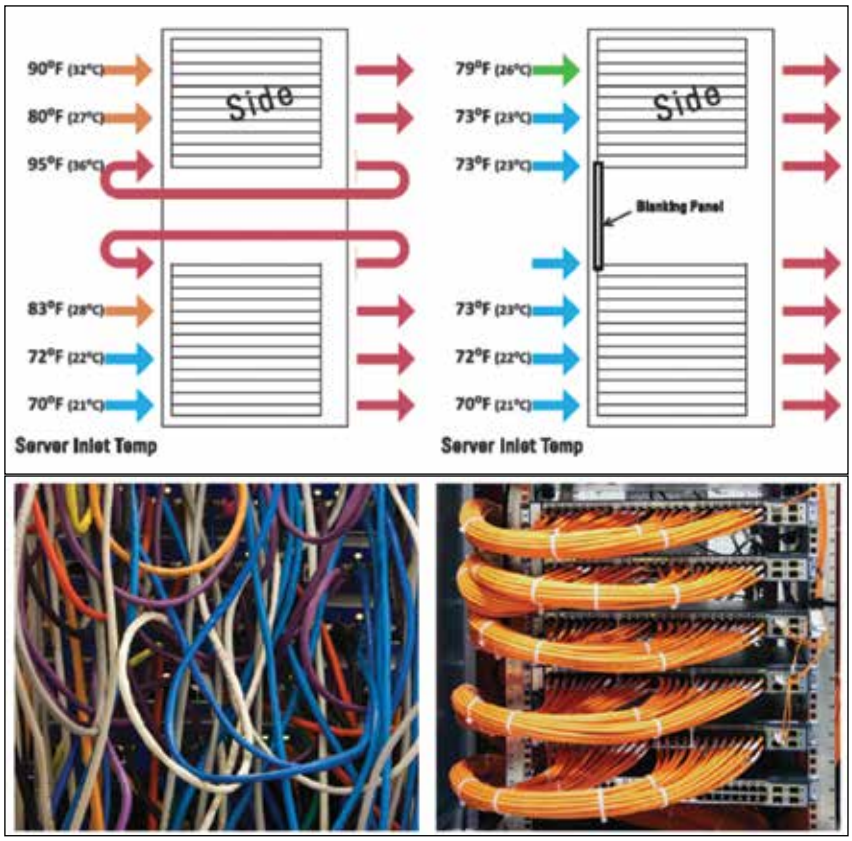

Figure 14 : Deux détails pour améliorer la circulation d'air : obturer les slots laissés vacants (photo du haut) – Soigner le câblage à l'arrière des baies – Source : EPA – Energy Star.

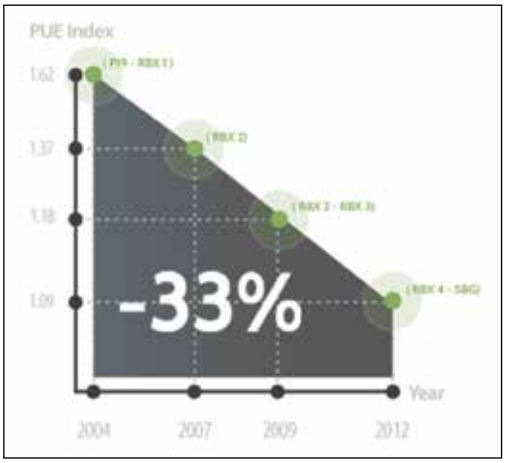

Figure 15 : Évolution de l'index PUE dans les data centers OVH – Source : OVH.

A contrario, des arrangements tels que celui de la figure 13, évitent le mélange entre air chaud et air frais, en guidant les flux par des rideaux ou en installant les baies dans des containers.

Les détails sont importants : il faut penser à obturer à l'avant des racks les slots laissés vacants et veiller, à l'arrière, à ce que le câblage ne porte pas atteinte à l'évacuation de l'air (figure 14).

Comme indiqué précédemment, il est possible de recourir à une circulation d'eau pour maintenir les serveurs en température. Dans ce cas, le refroidissement par air devient accessoire et il est possible de se contenter de l'air extérieur sans avoir à climatiser le bâtiment. Il faut cependant veiller à maintenir le taux d'humidité en dessous de 60 % pour éviter des condensations sur les composants. A contrario, l'air ne doit pas être trop sec pour limiter le risque d'électricité statique.

S'il a été décidé de recourir à l'eau, pour usage direct ou pour refroidir l'air, cette eau peut être fournie à partir des ressources locales et récupérée à des fins diverses. C'est le cas du data center Equinix à Amsterdam qui puise son eau fraîche dans le sol à 180 mètres de profondeur, l'eau réchauffée étant ensuite utilisée pour les besoins de chauffage de l'université d'Amsterdam.

Ces mesures, complétées par un choix approprié de composants, permettent de réaliser des économies d'énergie très appréciables, souvent de l'ordre de 10 à 40 %. Le centre d'Amsterdam d'Equinix annonce un PUE de 1,13 cependant qu'OVH déclare avoir ramené le PUE de ses data centers les plus récents à 1,09 en 2012 (figure 15). Google annonce de son côté avoir ramené son PUE moyen à 1,12 (figure 16)

La consommation des équipements informatiques

Les équipements informatiques sont pour leur immense majorité des équipe-

ments électroniques, y compris les stockages SSD qui permettent le stockage des données sur des dispositifs flash composés de semi-conducteurs. Les data centers bénéficient donc des progrès constants qui sont réalisés en matière de puissance, vitesse de traitement et efficacité énergétique par l'industrie des semi-conducteurs et qui s'inscrivent dans la recherche « d'une électronique verte ». Le maître d'œuvre d'un data center peut ainsi sélectionner de façon préférentielle des serveurs à faible consommation énergétique. Cependant, dans la plupart des cas, il reste confronté à un arbitrage entre consommation d'énergie et performances.

Pour réduire les consommations des équipements informatiques d'un data center, un concepteur et un exploitant peuvent jouer sur d'autres facteurs. Les facteurs principaux sont le dimensionnement du data center, généralement déterminé en fonction du besoin de pointe, et, critère qui va de pair, le taux d'utilisation. Le tableau 2 a montré clairement que beaucoup de serveurs étaient sous-utilisés alors qu'ils tournent 7/24/365. Il doit donc être possible d'en réduire le nombre et d'accroître le taux de charge de chacun.

Lorsque les serveurs sont simplement hébergés par le data center et restent directement opérés par les utilisateurs, il est possible à ces derniers de lutter contre les inefficacités dans les développements logiciels et d'organiser la répartition de la charge entre les différents serveurs sous leur contrôle.

C'est évidemment plus difficile lorsque les serveurs sont en colocation et opérés par l'exploitant du data center. Mais les techniques de virtualisation apportent une réponse. La virtualisation des serveurs consiste à faire fonctionner plusieurs serveurs virtuels sur un même serveur physique, les serveurs virtualisés étant remplacés par leur équivalent virtuel. L'objectif est de mutualiser les

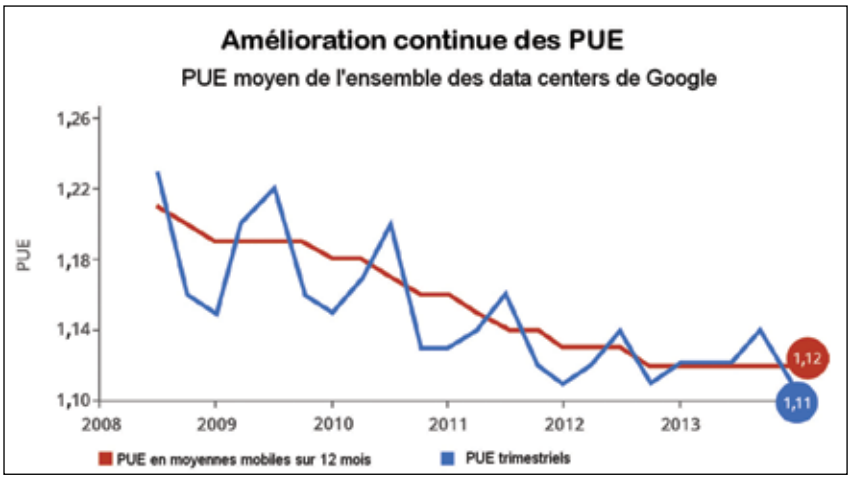

Figure 16 : Évolution du PUE moyen des data centers de Google – Source : Google.

capacités de chaque serveur, en vue de réduire les investissements physiques et de réaliser des économies d'énergie. C'est une bonne stratégie dès lors que les pics de charge afférents à chaque secteur virtualisé ne sont pas synchrones. Elle permet de faire fonctionner différents systèmes d'exploitation et applications liées, sans avoir à faire un choix par ordinateur ou serveur. Cette virtualisation est organisée par une plate-forme logicielle, supportée ou non par une base matérielle spécifique, qui est appelée hyperviseur.

La virtualisation est à présent entrée dans les mœurs. C'est une technique usuelle du cloud computing. Elle n'est pas sans inconvénient : elle entraîne une baisse de performances et le risque de voir s'effondrer simultanément tous les serveurs virtualisés doit être compensé par une redondance active performante. Cependant, il est admis que la virtualisation permet de réaliser des économies d'énergie allant de 10 à 40 %.

Vers des évolutions plus fondamentales

La généralisation du cloud computing et l'arrivée du Big Data posent un réel défi aux data centers, lié au volume mais aussi à la variété des données. Il ne s'agit pas de données relationnelles traditionnelles mais de données brutes, semi-structurées voire non structurées. Ce sont des données complexes provenant du Web ou d'ailleurs, au format texte ou images. Elles peuvent être publiques ou relever de la propriété des consommateurs. Tout ceci les rend difficilement utilisables avec les outils traditionnels. Il est peu probable, pour diverses raisons (performances, coûts, empreinte environnementale) que l'on puisse continuer à aligner, comme on le fait actuellement, les rangées de serveurs pour satisfaire les besoins de stockage, de traitement et d'échanges de données. La problématique posée est celle du stockage de pétaoctets de données avec des temps d'accès de l'ordre (a minima) du terabit/s.

On peut, sans trop de risques, faire le pari que nous aurons dans 50 ans le même regard sur nos data centers d'aujourd'hui que celui que nous portons sur les mémoires à tores magnétiques des années 60/70 qui coûtaient, par octet, 15 milllions de fois plus cher que les mémoires actuelles sans en avoir et de très loin les performances.

Dans cette évolution, il est tout à fait probable que l'optique jouera un rôle essentiel et que le photon détrônera progressivement l'électron. L'optique est aujourd'hui présente dans les data centers sous forme de liaisons assurant les communications entre les racks ou clus-

ters de serveurs et avec les stockages de données, liaisons pour lesquelles le cuivre n'offrirait ni des débits suffisants ni les taux d'erreur requis. Cependant les liaisons optiques débouchent sur des transducteurs électrooptiques associés à des commutateurs de paquets conventionnels (switches) assurant le routage du trafic. Il peut exister dans un data center, en fonction de sa complexité, plusieurs niveaux de switches. Ces switches induisent des temps de latence dans les traitements de 10 à 20 ms et sont fortement consommateurs d'énergie. Mais à partir du moment où ils constituent le goulet d'étranglement du data center, c'est la performance de l'ensemble qui s'en ressent.

Un axe d'évolution important, présenté à l'Optical Society of America en 2011 [10], consiste à remplacer les commutateurs de paquets par des commutateurs optiques basés sur la technologie MEMS (microsystèmes électromécaniques gravés sur le silicium) assurant le routage optique de flux transmis en multiplexage de longueur d'onde. Cette technologie permet de réaliser des progrès très importants en temps de réponse et en consommation d'énergie et de construire des "software defined networks" reconfigurables en temps réel en fonction des besoins de commutation.

Du côté du stockage, la solution classique reste celle des disques durs magnétique, remplacés lorsque les performances l'exigent par des mémoires flash. Pour faire face aux exigences croissantes en capacité et en temps d'accès, le stockage optique des données est évidemment une solution vers laquelle on se tourne naturellement. Cependant, la capacité des moyens actuels de stockage, du type Blu-ray, qui opèrent en microscopie optique à partir d'un seul faisceau laser, est limitée par la diffraction de la lumière et se situe bien en dessous du To par disque[4]. En outre, le mécanisme d'écriture séquentiel, bit par bit, limite la capacité de transfert à quelques dizaines de Mbit/s. Cependant de nouvelles techniques voient le jour et on trouvera dans le Flash Info de ce numéro de la REE un article consacré à une technologie de "Superresolution Photoinduction-Inhibited Nanolithography" qui, avec une technique de convergence multifocale, est susceptible de permettre la réalisation de stockages optiques de très haute capacité (30 To) avec des vitesses d'accès allant jusqu'au Gbit/s [11].

Il restera la question la plus délicate, celle des processeurs. Sur ce point, des progrès importants sont réalisés sur certains aspects essentiels du défi soulevé par un hypothétique ordinateur photonique. La REE les relate régulièrement dans sa rubrique Flash Info. La nano-électronique combinée à la nanophotonique offre des perspectives intéressantes pour de nombreuses applications, y compris les calculs à très haute vitesse. Le lecteur pourra se reporter à la REE 2014-4 pour comprendre comment les antennes « plasmoniques » pourraient être utilisées pour réaliser des liaisons à très haut débit en champ lointain entre composants optoélectroniques intégrés, sans risque d'interférence. Il existe également des travaux très importants, aux États-Unis mais aussi en France[5] sur l'intrication des photons et des électrons qui pourraient conduire d'ici quelques décennies à des ordinateurs quantiques, extraordinairement puissants mais dont on ne sait pas dire pour l'instant s'ils seront plus économes en énergie que ceux que nous utilisons aujourd'hui.

Mais le stockage magnétique n'a pas dit son dernier mot et le lecteur trouvera dans la REE 2013-5 un Flash Info sur les « skyrmions », tourbillons magnétiques générés par de minuscules courants émis depuis la pointe d'un microscope qui pourraient être alignés dans un milieu magnétique en couche mince à des distances de l'ordre de quelques nm et augmenter ainsi considérablement la capacité des disques magnétiques.

Nous sommes ici dans le domaine de la recherche mais celle-ci se trouve stimulée par l'enjeu considérable que revêtent les data centers et l'arrivée du Big Data. Face à des machines superpuissantes, il faudra disposer, pour les exploiter, d'un savoir-faire à la hauteur. On ne confie pas une berline de 600 CV à un conducteur novice ! Certains pensent que la solution pour l'exploitation des data centers se trouve dans l'intelligence artificielle et le machine learning. C'est le cas de Google qui a publié en mai 2014 un Livre blanc intitulé "Machine learning application for data center optimization" [12]. Cette étude montre qu'il est possible de modéliser le système très complexe de consommation d'énergie des data centers par un réseau neuronal enrichi par la confrontation entre les prévisions et les réalisations. Ce modèle permet de prévoir le PUE d'un data center avec une précision de 0,004 c'est-à-dire 0,4 % pour un PUE de 1,1. Il ne s'agit pour l'instant que d'un modèle permettant d'optimiser les consommations représentatives du PUE. Cependant, on peut penser que l'approche pourra être étendue à l'optimisation du traitement de l'information. Face aux requêtes à satisfaire et aux objectifs qui lui seraient assignés, le data center pourrait ainsi être rendu intelligent et déterminer de lui-même des stratégies optimales.

[4] Le disque Blu-ray offre dans sa version la plus performante, une capacité de 128 GB. Des disques optiques de 300 GB sont annoncés pour la fin 2015.

[5] Il faut rappeler le très beau Prix Nobel de physique 2012 décerné à Serge Haroche, professeur au Collège de France, sur ce sujet.

Conclusion

Tirant leurs racines des "main frames" les data centers, après deux décennies où les microordinateurs ont fait florès, marquent un retour vers une informatique où les ressources sont partagées afin de pouvoir disposer de capacités de traitement et de stockage considérables sans avoir à en supporter seul le coût.

Ce mouvement va se poursuivre car même si l'expansion des data centers ne se fait pas au rythme imaginé vers les années 2005/2007, la généralisation du cloud computing et l'arrivée du Big Data les rendent incontournables.

Les data centers sont aujourd'hui les usines des temps modernes. Leur complexité est masquée par des bâtiments couleur muraille. Mais à l'intérieur, la recherche de la performance, de la disponibilité et du moindre coût sont de très forts stimulants pour l'innovation technologique. En matière d'empreinte environnementale, un gros effort a été fait pour réduire les consommations des infrastructures supportant le process proprement dit. Les data centers des grands opérateurs de la Net économie sont arrivés à ramener leur PUE au voisinage de 1. Il faut à présent généraliser ces acquis à l'ensemble de la population des data centers dont beaucoup sont très en retard par rapport aux leaders. Il faut également passer à la phase 2, c'est-à-dire trouver des solutions technologiques, matérielles et logicielles, qui, dans les prochaines décennies, permettront d'améliorer considérablement le rendement thermodynamique des centres de traitement de l'information. C'est un défi technologique et aussi philosophique considérable, qui verra les data centers se mesurer au cerveau humain.

Références

[1] N. R. D. Council, "Data center Efficiency Assessment", août 2014.

[2] IDC, "Worldwide Datacenter Census and Construction 2014-2018 Forecast: Aging Enterprise Datacenters and the Accelerating Service Provider Buildou", 2014.

[3] CISCO, "Cisco Global Cloud Index: Forecast and Methodology, 2013–2018", 2014.

[4] S. E. M. B. T. P. R. &. B. M. Greenberg, "Best Practices for Data Centers: Lessons Learned from Benchmarking 22 Data Centers", chez ACEEE Summer Study on Energy Efficiency in Buildings, Asilomar, 2006.

[5] E. C. - J. r. c. - I. f. Energy, "Code of Conduct on Data Centres Energy Efficiency", 2008.

[6] Greenpeace, "How clean is your cloud", 2012.

[7] Broadgroup, "Iceland's competitive advantages as a global Data Centre location".

[8] U. E. P. Agency, "Report to Congress on Server and Data Center Energy Efficiency", 2007.

[9] U. E. P. Agency, "Programme ENERGY STAR - TOP 12 Ways to decrease the energy consumption of your data center".

[10] A. Vahdat, "The emerging optical data center", chez OSA/OFC/NFOEC, San Diego, 2011.

[11] Y. C. N. T. L. F. e. M. G. Xiangping Li, "Multifocal optical nanoscopy for big data recording at 30 TB capacity and gigabits/second data rate", Optica Vol 2 N°6, 6 June 2015.

[12] J. Gao, Machine learning fo data center optimization, Google, 2014.

[13] J. G. K. - S. University, "GROWTH IN DATA CENTER ELECTRICITY USE 2005 TO 2010", Analytics Press,, 2011.

L'AUTEUR

Jean-Pierre Hauet est ingénieur au corps des Mines. Il est associate partner de KB Intelligence. Au cours de sa carrière, il a dirigé les Laboratoires de Marcoussis du groupe Alcatel-Alsthom et a été Chief Technology Officer du Groupe ALSTOM. Il est membre émérite de la SEE et rédacteur en chef de la REE.

Efficacité énergétique des data centers : le point sur les initiatives en cours en France et en Europe

Le développement de l'économie numérique de demain, e-santé, e-éducation, villes intelligentes, données massives, etc., repose sur la disponibilité de data centers pour faire fonctionner ces applications 24 h sur 24. Bien que la consommation des data centers ne représente que 20 % de la consommation totale de l'Internet, les acteurs de ce monde ont toujours été proactifs pour améliorer l'efficacité de ces infrastructures dans l'utilisation des ressources en général et de l'énergie en particulier, dans le souci notamment de réduire les émissions de gaz à effet de serre.

En France et en Europe, les actions des industriels se sont développées suivant trois axes :

- **Amélioration des rendements énergétiques des composants des data centers**, tels que les alimentations sans interruption (UPS[1]), les équipements de refroidissement ou les serveurs. De nombreuses initiatives européennes ont vu le jour en la matière : code de conduite européen, projets pilotes *Product Environmental Footprint* (PEF) avec la Commission européenne, éco-conception des serveurs et des UPS dans le cadre de l'application de la Directive ErP[2], etc. Répondant aux normes et aux incitations, les constructeurs ont mis sur le marché des produits plus performants. L'efficacité des UPS a ainsi atteint 95 %, contre 85 % en moyenne dans les années 1990. Ces actions sur le rendement de chaque composant entrent dans le cadre de ce qui est aussi appelé efficacité énergétique passive.
- **Amélioration de l'intelligence de fonctionnement des data centers, ou actions portant sur la gestion active**, en créant tout d'abord des métriques de mesure d'efficacité simples à mettre en œuvre. Le PUE (*Power Usage Effectiveness*), métrique de mesure de l'efficacité des data centers créée par le consortium *The Green Grid* est désormais utilisé par la quasi-totalité des 500 plus gros data centers européens en raison de sa simplicité de mise en œuvre. Des logiciels de supervision ont été développés (DCIM[3]) pour introduire davantage de modularité, c'est-à-dire pour adapter plus finement, en temps réel, le fonctionnement de l'ensemble du data center (serveurs, équipements) à la demande de traitement de données. Parallèlement, constructeurs et opérateurs ont travaillé en commun sur des systèmes de refroidissement actifs au plus près des machines générant de la chaleur. L'efficacité énergétique des data centers s'en est trouvée nettement améliorée, le PUE moyen passant de plus de 2,5 à 1,8 à ce jour.
- L'axe de travail le plus récent, dans le contexte d'une forte croissance du besoin de traitement de données, consiste à développer des logiciels pour une **intégration optimale des data centers dans les villes intelligentes**, dont ils assureront le fonctionnement sûr et efficace. La mise en réseau des data centers permettra de transférer la demande de traitement de données d'un data center à l'autre en fonction de la disponibilité d'énergies renouvelables, par le biais de contrats de services négociés avec les acteurs de l'écosystème (fournisseurs d'énergie, etc.). Dans la mesure du possible, la chaleur émise par les data centers sera récupérée pour chauffer les logements et bureaux situés à proximité. C'est la direction prise par les projets de R&D actuellement cofinancés par la Commission européenne (DC4Cities, GreenDataNet, Geyser, Genic, Renew IT ou encore Dolfin), qui souhaite que 80 % de l'énergie consommée par les data centers provienne de sources renouvelables à l'horizon 2030.

André Rouyer
Délégué Data Centers du Gimélec
Syndicat professionnel de l'équipement électrique, le Gimélec travaille depuis plus de cinq ans sur le sujet de la performance énergétique des data centers et a conclu des partenariats avec *The Green Grid* et le *Joint Research Center* de la Commission européenne.

[1] Uninterruptible Power Supply.
[2] Energy Related Products.
[3] Data Center Infrastructure Management.