DATACORE

COMMENT TIRER PARTI DE L'HYPERCONVERGENCE

pour les bases de données et les applications d'entreprise de niveau 1

Par Scott D. Lowe

ActualTech Media

Sommaire

Introduction

Nous vivons dans un monde où les données sont devenues essentielles. Leur quantité, tout comme notre besoin d'un volume de données toujours croissant, explosera dans les années à venir. Or, tout cela repose sur des bases de données qui aident les entreprises à ne pas se laisser déborder. Au même moment, l'infrastructure qui supporte ces bases de données est en pleine transformation à mesure que les entreprises cherchent à simplifier des systèmes informatiques complexes pour en réduire le coût et accélérer le rythme de leur activité.

L'infrastructure hyperconvergée a émergé ces dernières années parce qu'elle donnait aux entreprises un outil d'une puissance incroyable pour endiguer la folie dans les centres de données. L'adoption de cette technologie a été explosive (ce qui est une bonne nouvelle !) et les entreprises en retirent des gains sensibles en termes d'efficacité opérationnelle et de réduction des coûts. Suite à certaines décisions précoces sur des cas d'utilisation, prises par les fournisseurs à l'origine du marché de l'infrastructure hyperconvergée, et aux limitations des premières plateformes, nombreux sont ceux qui voient encore dans l'hyperconvergence une solution uniquement destinée aux postes de travail virtualisés (virtual desktop infrastructure ou VDI) et à de petites applications traitant des charges légères.

Bien que l'hyperconvergence ait de très loin dépassé ces premiers stades, ces cas d'utilisation continuent de mettre en exergue ce qui fait la puissance de cette option d'architecture pour les centres de données. Prenons, par exemple, l'infrastructure VDI, il s'agit d'une charge de travail à l'expansion véritablement linéaire. À mesure que vous ajoutez des utilisateurs, vous avez besoin davantage de puissance de traitement. Avec l'hyperconvergence, à mesure que vous ajoutez des utilisateurs VDI, il vous suffit d'ajouter des boitiers jusqu'à ce que votre déploiement soit achevé. Pour ce type d'utilisation, le calcul est très simple. Il suffit de multiplier et de diviser. Même si vos besoins de VDI changent et qu'il vous faut élargir la taille de chaque ordinateur virtuel, le calcul reste facile. Vous ajoutez quelques hôtes et le tour est joué.

Les applications à faible charge (celles qui ne nécessitent pas d'énormes ressources) ont également été jugées par le passé tout à fait adaptées aux solutions hyperconvergées. Plus récemment, des applications plus musclées ont également montré qu'elles fonctionnaient bien dans des scénarios hyperconvergés tant qu'elles restaient cantonnées à un seul serveur. D'ailleurs, outre le VDI, ces types d'application ont été accueillis sur des architectures hyperconvergées.

Pour autant, le monde des applications d'entreprise compte bien plus que des applications VDI ou de petite taille. L'une des principales difficultés rencontrées dans les scénarios hyperconvergés concerne les applications qui exigent des niveaux d'I/O élevés sur des hôtes uniques. Par exemple, un grand nombre de systèmes de bases de données d'entreprise et d'applications associées ont du mal à dépasser ce stade de l'hôte unique, même si cela est techniquement possible. Les bases de données ont besoin de davantage d'I/O (un ordre de grandeur de plus) que les autres types d'application et sont souvent mal supportées par les architectures évolutives. Ces applications peuvent souvent nécessiter des centaines de milliers, voire des millions d'IOPS, selon leur taille, contre seulement des milliers ou des dizaines de milliers pour des applications de niveau 2 ou inférieur.

Pour les applications de bases de données, l'avantage de l'évolutivité est limité, voire nul dans certains cas, et vous n'avez pas d'autre solution pour évoluer que d'ajouter des ressources à chaque nœud. Le principal problème est que la plupart des systèmes ne peuvent évoluer que jusqu'à un certain point et des goulots d'étranglement peuvent apparaître.

Sachant cela, comment persévérer dans vos plans d'infrastructure hyperconvergée tout en garantissant que vos applications de bases de données pourront continuer à croître ?

Pourquoi même se soucier de l'hyperconvergence ?

La première question que l'on pourrait se poser à ce stade est la suivante : « Pourquoi même se soucier d'une infrastructure hyperconvergée si elle est mal adaptée à l'ensemble des applications que j'exécute ? ». C'est une question tout à fait légitime. Pour de nombreuses entreprises, il existe un désir de simplifier les centres de données et rendre plus facile l'ajout de ressources à la demande et d'une manière non-disruptive. L'hyperconvergence a fait ses preuves en tant qu'architecture au service des applications d'entreprise, grandes et petites. Connaissant les avantages que peut apporter sa mise en place, il importe de s'assurer qu'elle fera correctement fonctionner les applications.

Dans un premier temps, l'hyperconvergence pourra vous apparaître comme un espèce de bloc

monolithique. Vous déployez un cluster, puis vous tentez d'y faire entrer vos charges de travail. Bien que cela puisse fonctionner correctement pour les petites entreprises, vous aurez probablement davantage de succès en déployant plusieurs clusters dédiés aux applications. Ce faisant, vous éviterez de mélanger vos charges de travail de niveau 1 avec les autres et vous pouvez laisser à vos applications critiques suffisamment de ressources pour fonctionner, tout en conservant la méthodologie d'attribution des ressources « juste à temps » qui fait l'intérêt des infrastructures hyperconvergées. Vous cherchez donc à pouvoir attribuer juste assez d'infrastructure aux applications qui résident dans le cluster, cela sans entraîner trop de gaspillage.

Décupler les performances de l'hyperconvergence

Ce livre blanc s'intéresse surtout aux bases de données et aux applications de niveau 1, pour lesquelles les caractéristiques de performances sont absolument cruciales. Comme vous le savez, les performances peuvent être mesurées de diverses manières, chaque application présentant des besoins très différents.

L'élément que chaque caractéristique de performances partage avec toutes les autres lorsqu'il s'agit de la plateforme sous-jacente est précisément cette plateforme elle-même. Les systèmes d'infrastructures hyperconvergées présentent des caractéristiques de performances intéressantes imposées par l'architecture : les nœuds qui génèrent les charges d'I/O sont précisément ceux qui doivent les gérer.

Lorsque les serveurs et le stockage sont séparés, les serveurs génèrent la charge d'I/O et le stockage, assuré par un dispositif distinct, fait intervenir sa puissance de calcul pour pouvoir traiter cette charge. Avec l'hyperconvergence, le même nœud fait les deux, ce qui augmente la pression sur la plateforme, qui doit alors exécuter les applications et gérer les I/O. C'est l'une des raisons pour lesquelles les applications à faible charge ou à charge évolutive se comportent bien dans les scénarios d'infrastructure hyperconvergée.

Pourtant, ce goulot d'étranglement potentiel n'est pas le seul du genre. Pour une infrastructure hyperconvergée, il existe un certain nombre de points susceptibles d'affecter les performances globales des charges de travail. De même, il existe un certain nombre de moyens pour stimuler les performances de ces systèmes mais, dans de nombreux cas, chacune de ces méthodes présente un inconvénient.

Commençons par poser le postulat fondamental selon lequel les goulots d'étranglement touchant les performances des applications sont en général à imputer aux I/O. Quelque part entre l'application et les divers composants de l'infrastructure, il se produit un bouchon qui doit être éliminé pour que l'application fonctionne bien. Étudions maintenant certaines des méthodes qui permettent d'améliorer les performances des applications.

Opter pour le tout-flash

La façon peut-être la plus courante de bénéficier instantanément d'un gain de vitesse consiste à se débarrasser des disques rotatifs pour tous les remplacer par des périphériques flash flambant neufs. Vous avez en effet toutes les chances d'aller plus vite. Le principal inconvénient est que cette accélération peut s'accompagner d'un énorme surcoût. Au moment de la rédaction de cet article, on peut trouver dans le commerce un disque dur Seagate Enterprise Capacity de 4pour à

peine plus de 150 USD. Un SSD Samsung 850 EVO de 4 To coûte aujourd'hui un peu plus de 1 500 USD. La version de 1 To du Samsung 850 EVO vaut 300 USD. Tous les tarifs ne sont évidemment pas aussi disparates, mais l'objet ici était de montrer qu'opter pour le tout-flash restait encore bien plus onéreux que de conserver ses disques.

Mais même cette solution comporte aussi des inconvénients. La plupart des solutions de stockage tout-flash intègrent également des technologies exhaustives de réduction des données sous la forme de fonctions de déduplication et de compression. Associés aux charges de travail les mieux adaptées, ces services peuvent ramener le coût du flash plus près de celui d'un disque brut. Malgré l'encre qui a coulé pour arguer du contraire ces dernières années, le disque n'est pas mort et n'a aucune raison de l'être. Combiné à la bonne application et au bon logiciel, il en est même loin et, en termes de capacité brute, reste en tête de la guerre du stockage dans les centres de données.

Cela ne signifie pas pour autant qu'il faut totalement se priver de la technologie flash. Je conseillerais plutôt d'envisager le meilleur des deux mondes et de penser à un stockage hybride pour vos solutions d'infrastructure hyperconvergée. En combinant le flash et les disques rotatifs, vous obtenez le meilleur de ce que ces deux technologies ont à offrir.

En outre, toutes les applications n'ont pas besoin de flash. De même, toutes les données ne sont pas cruciales en permanence. En réalité, seule une fraction des données figurant dans vos baies de stockage est utilisée à un instant donné, ce qui fait du stockage hybride un choix idéal pour de nombreuses entreprises. Par ailleurs, les solutions flash vous feront bénéficier d'un grand nombre d'IOPS et d'un large débit, mais que se passera-t-il si les problèmes de performances de vos applications ont une autre origine ?

Évoluer

Si vous avez besoin d'hôtes de plus grande taille, pourquoi ne pas les construire ? Qu'est-ce qui vous empêche d'ajouter tout simplement de la RAM et des dispositifs de stockage plus rapides à vos hôtes hyperconvergés ? Pourquoi ne pas simplement remplacer tout le stockage existant dans vos hôtes par une interface NVMe, par exemple, pour augmenter encore vos performances de stockage ?

Tout d'abord, la capacité ou non de mettre vos nœuds à niveau dépend de ce que vous avez acheté. Si vous avez acheté un dispositif d'infrastructure hyperconvergée, vous ne pouvez probablement pas le mettre à niveau vous-même. Vous devez au moins faire appel au fournisseur afin de ne pas annuler votre contrat de support technique. Si vous avez opté pour un produit hyperconvergé à base de logiciel, tel qu'une solution d'infrastructure hyperconvergée de DataCore, vous avez beaucoup plus de choix dans les mises à niveau matérielles applicables à chaque nœud. Vous pouvez ajouter de la RAM, du stockage flash, une interface NVMe, plus de puissance processeur ou tout ce dont vous pouvez avoir besoin pour traiter vos charges de travail.

L'évolutivité suffit-elle à résoudre vos problèmes de performances pour les charges de travail ? Le problème avec l'évolutivité est que toutes les charges de travail ne peuvent en réalité pas bénéficier de cette puissance supplémentaire si des goulots d'étranglement subsistent ailleurs dans le système. Même s'il est probable que certaines mises à niveau matérielles amélioreront quand

même les choses, il est possible qu'elles ne donnent pas tous les résultats que vous attendez de ce type d'investissement.

Quels chiffres prendre en compte ?

Vous savez que l'ajout de mémoire flash augmentera le nombre d'IOPS disponibles pour votre application. Vous savez qu'en ajoutant des disques, vous augmenterez le débit. Vous savez qu'un plus grand nombre de cœurs de processeur vous permettra peut-être d'augmenter la densité de certaines charges de travail.

Mais quelle est la mesure stratégique qui les gouverne toutes ? La latence.

Tous les autres indicateurs auxquels vous pouvez penser se réduisent en fin de compte à cette mesure fondamentale de l'expérience utilisateur, qui se définit comme le temps pendant lequel un utilisateur, qu'il s'agisse d'un employé, d'un sous-traitant ou d'un client, doit attendre pour qu'une opération soit exécutée. Il existe toutes sortes de mesures techniques de la latence, comme le temps de recherche sur les disques rotatifs ou encore quelque chose d'aussi profond que le temps de disponibilité des processeurs sur les machines virtuelles. Au bout du compte, tout cela se réduit au temps pris pour l'exécution d'une opération, ce qui se mesure en regardant un minuteur et… en attendant. Toutes ces petites latences accumulées dans tout le système peuvent s'ajouter au fil du temps.

Il existe d'innombrables outils d'évaluation pour mesurer les différents chiffres de performances. Pour tenter de comparer les solutions et les fournisseurs de façon équitable, le Storage Performance Council (SPC) a créé la norme d'évaluation des performances SPC-1, qui aide les utilisateurs finaux à comparer les chiffres de performances système entre différents fournisseurs. Cet indicateur de performance est utilisé par les fournisseurs de systèmes de stockage afin d'offrir une comparaison standard des performances en termes d'IOPS, de coût par IOPS et de latence, tous étant des facteurs de décision critiques au moment d'acheter une nouvelle infrastructure.

Surtout, lorsqu'il s'agit de charges de travail de bases de données, la norme SPC-1 vise à se rapprocher de l'utilisation concrète des bases de données OLTP.

La place de DataCore dans l'évaluation SPC-1

Au moment de la rédaction de ce livre blanc, il n'existe qu'un seul fournisseur d'infrastructures hyperconvergées qui ait soumis des produits aux tests de la norme SPC-1 : DataCore. Les autres fournisseurs du marché s'appuient sur leurs propres tests ou sur d'autres indices d'évaluation. Enfin, ce qui est une excellente nouvelle pour les utilisateurs d'infrastructures DataCore hyperconvergées, les résultats des tests SPC-1 montrent que les performances des solutions hyperconvergées de la marque surpassent celles des systèmes tout-flash proposés par de grands fournisseurs de solutions de stockage comme Dell/EMC, NetApp ou HPE.

Comme vous pouvez le voir dans la figure ci-dessous, le chiffre global de latence de DataCore est très nettement inférieur à celui d'autres systèmes de stockage d'entreprise testés. Sous chaque barre apparaît également un coût. Il s'agit du coût par IOPS de la solution testée. Ici aussi, DataCore est en tête, avec moins de la moitié du coût des autres solutions testées.

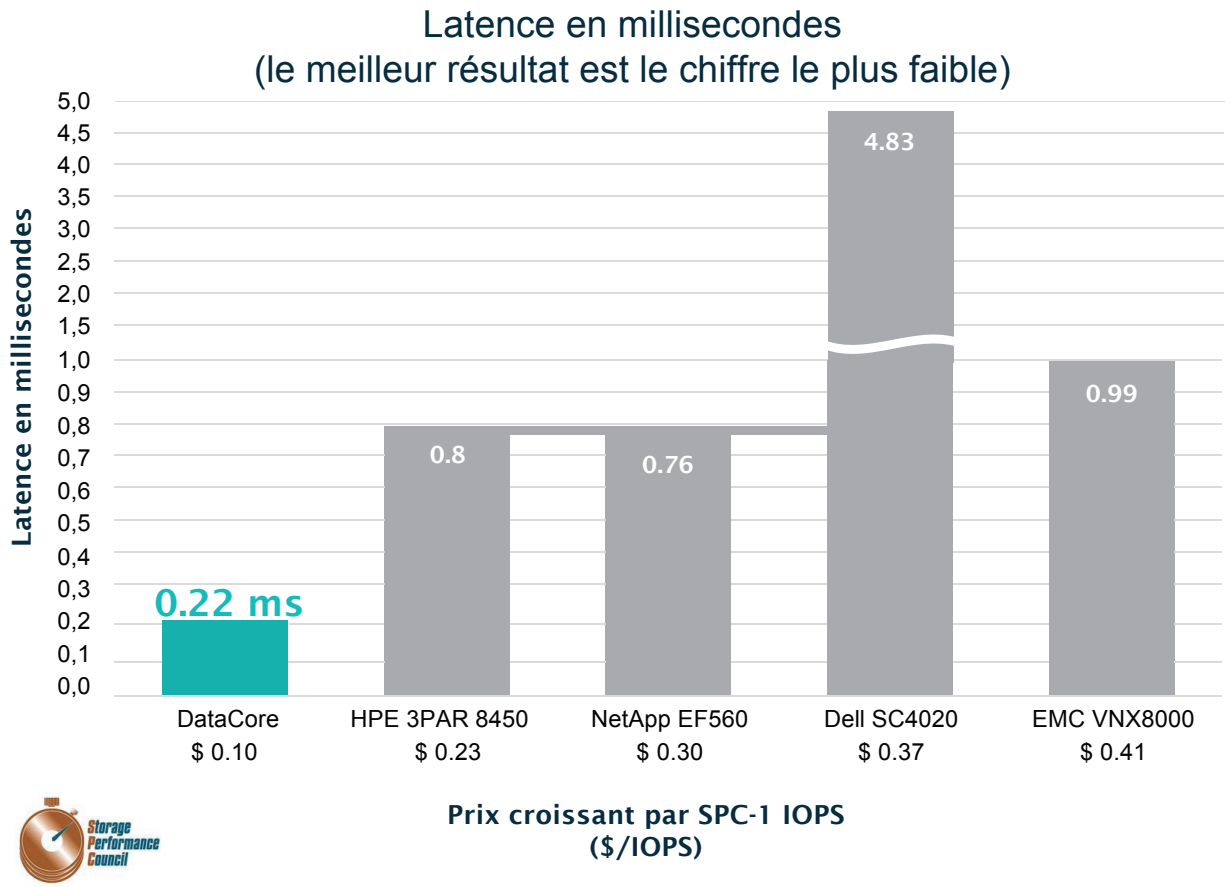

Figure 1. Comparaison de DataCore et de la concurrence:
LatenceSource: DataCore et le Storage Performance Council

Parallel I/O, une solution gagnante

Comment la société DataCore parvient-elle à surpasser ces autres solutions aussi facilement ? L'ingrédient secret est une approche innovante des performances d'I/O. De nombreux systèmes comportent des points de contention uniques qui obligent à gérer les éléments de données de façon séquentielle. Cela signifie que la solution globale ne peut pas aller plus vite que son composant le plus lent.

DataCore envisage la question des I/O sous l'angle d'un traitement parallèle, ce qui permet aux utilisateurs de traiter plusieurs éléments d'I/O en parallèle, améliorant ainsi considérablement les performances globales du système. En outre, même si le nombre d'IOPS n'est pas à lui seul un indicatif suffisant des performances d'une application, il joue néanmoins un rôle indéniable dans la détermination de la latence globale. Comme le montre le diagramme ci-dessous, DataCore est parvenue à créer une solution capable de prendre en charge plus du double d'IOPS que son concurrent le plus proche.

Mais le plus impressionnant est que ces statistiques de performances sont tirées de la solution hyperconvergée de DataCore, comparées ici à des produits de stockage dédiés d'autres fournisseurs. La capacité des systèmes d'infrastructure hyperconvergée de DataCore à gérer les performances plus près de l'application jouent un rôle non négligeable dans ces résultats impressionnants.

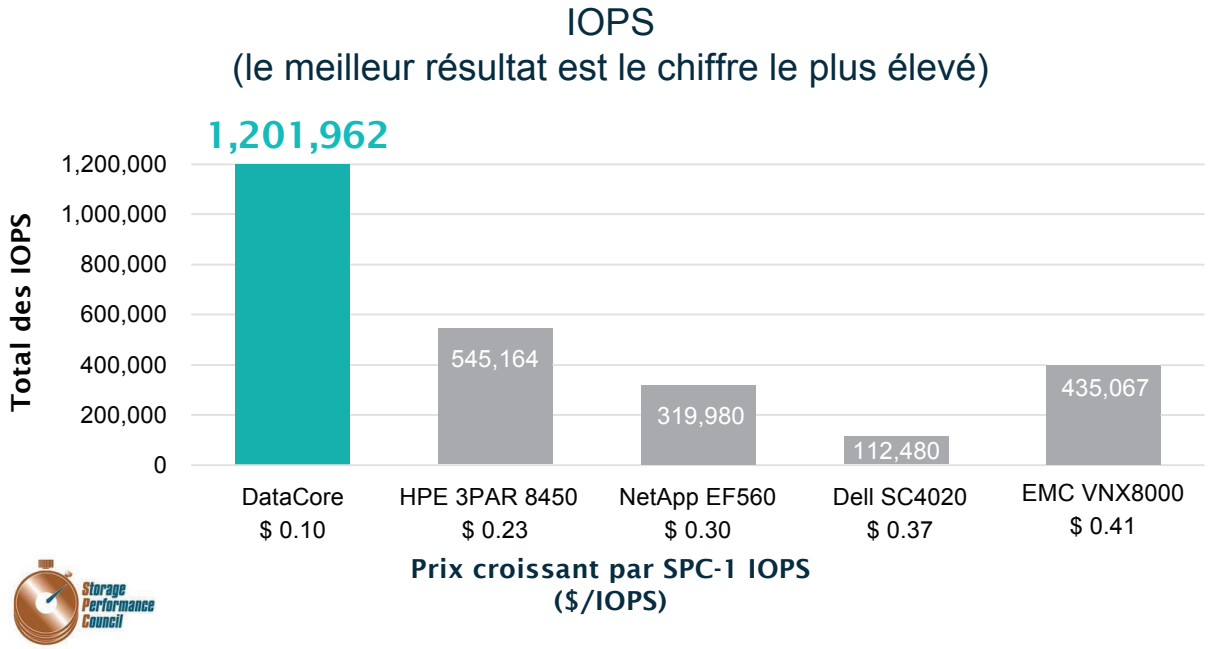

Figure 2. Comparaison de DataCore et de la concurrence: IOPS

Source: DataCore et le Storage Performance Council

La capacité des logiciels DataCore à utiliser efficacement le matériel sous-jacent apporte également un certain nombre d'autres avantages, dont certains ne sautent pas tout de suite aux yeux. Tout d'abord, il faut prendre en compte les questions de disponibilité. DataCore peut prendre en charge des clusters métropolitains (des volumes de données géographiquement dispersés mis en miroir de façon synchrone) n'ayant que deux nœuds, avec un seul nœud sur chaque site. Ensuite, chaque nœud peut prendre en charge de nombreuses charges de travail avec des débits élevés. Cette fonctionnalité de cluster métropolitain à deux nœuds diminue considérablement le coût global de la solution DataCore par rapport à la plupart des autres produits hyperconvergés, qui exigent en général au moins deux ou trois nœuds de chaque côté. Au total, cela signifie que les autres solutions peuvent être contraintes de commencer avec au moins quatre à six nœuds, ce qui coûte beaucoup plus cher.

Un autre aspect essentiel de la gestion des charges de travail est la redondance. Le matériel n'est pas infaillible et il tombera en panne à coup sûr. De nombreux systèmes sont bâtis sur le concept de redondance N+1, ce qui signifie que lorsqu'un nœud tombe en panne, les N nœuds restants doivent pouvoir gérer toute la charge, y compris celles du nœud défaillant. Si les nœuds ne sont pas efficaces, alors, même si la taille minimale du cluster est N du point de vue de la disponibilité, il sera préférable de conseiller une taille de cluster de N+2 afin que la panne d'un nœud n'entraîne pas de diminution des performances.

Avantages pour l'entreprise

Quel rapport tout cela a-t-il avec la capacité de DataCore de permettre aux utilisateurs de faire évoluer des nœuds individuels ? Comme nous l'avons vu précédemment, toutes les solutions ne sont pas à même d'optimiser efficacement l'utilisation de matériel supplémentaire. En offrant la possibilité de servir les I/O de façon parallèle, DataCore permet aux utilisateurs d'exploiter toute la puissance contenue dans la couche de calcul afin de répondre aux besoins d'I/O des applications.

La technologie Parallel I/O de DataCore permet de bénéficier d'un certain nombre d'avantages, notamment les suivants :

- Des nœuds nettement plus efficaces dans le cluster d'infrastructure hyperconvergée
- Extension facile des ressources sur chacun des nœuds
- Nombre de nœuds nécessaires réduit, ce qui diminue encore la complexité et les coûts d'investissement
- Des bases de données aux performances incroyables, puisque ces applications sont une charge de travail qui peut bénéficier considérablement de la technologie Parallel I/O

Résumé

Grâce à sa fonctionnalité de traitement parallèle des I/O, DataCore peut aider ses clients à exécuter les charges de travail des bases de données les plus lourdes sur des hôtes uniques. Si les nombreux avantages liés à l'utilisation d'une infrastructure hyperconvergée vous intéressent et si vous envisagez d'exécuter des bases de données sur ce type de plateforme, vous devez trouver le produit présentant le niveau idéal de performances, de disponibilité et de prix. Nous vous conseillons vivement d'étudier la façon dont DataCore peut aider votre entreprise à combler ses besoins d'infrastructure et même d'aller au-delà.

À propos de DataCore

Les approches classiques de l'infrastructure informatique sont inefficaces et leur coût est disproportionné par rapport à leur valeur à long terme. C'est pourquoi DataCore a ouvert la voie de la virtualisation du stockage basé sur les logiciels. Aujourd'hui, l'entreprise apporte flexibilité et réactivité à la gestion du stockage hétérogène, tout en élevant les performances et la disponibilité à des niveaux qui changent réellement et durablement la donne pour ses clients.

L'approche de DataCore en matière de stockage défini par logiciel, de SAN de serveurs et d'infrastructure hyperconvergée a été choisie par des milliers de clients qui ont compris qu'il était absurde de changer sans cesse de matériel.

Et ce sont des clients fidèles à 95 % parce qu'à mesure que le matériel sous-jacent change et que de nouvelles technologies émergent, DataCore offre une continuité d'activité, des performances et une valeur ajoutée sans précédent.

Avec DataCore, des technologies telles que la mémoire flash, les interfaces NVMe, le cloud, les conteneurs (et tout ce qui leur succédera) peuvent être introduites sans perturbation et intégrées de façon transparente à l'environnement existant déjà en place. DataCore a également appliqué sa technologie Parallel I/O à de nouveaux domaines, notamment l'optimisation des bases de données à travers son produit MaxParallel™ for Windows Server.

Le siège social de l'entreprise se trouve à Ft. Lauderdale, en Floride, et la société possède des bureaux en Amérique du Nord, en Europe et en Asie. Pour en savoir plus, adressez un e-mail à infofrance@datacore.com.

À propos de l'auteur

Scott D. Lowe

Scott Lowe est vExpert et partenaire chez ActualTech Media. Il travaille dans le domaine de l'informatique depuis plus de 20 ans et a occupé pendant 10 ans le poste de directeur informatique de diverses entreprises. Que ce soit sous forme de livres, d'articles ou publications sur des blogs, Scott a écrit des milliers de documents. Aujourd'hui, au sein de l'équipe multidisciplinaire d'ActualTech Media, il se consacre à la formation des acheteurs en informatique et à la mise en contact de sociétés informatiques avec des clients potentiels. Pour en savoir plus sur ActualTech Media, vous pouvez consulter le site www.actualtechmedia.com.