Stockage: IBM croit à la bande et mise sur la déduplication

Avec Diligent vous utilisez à présent des technologies de déduplication avancées. Quels arguments ont plaidé en faveur de ces choix technologiques?**Cincy Grossman, iBM:** Nous cherchions à disposer d'une technologie et nous étudiions le "*hash based*", le "*content aware*" ou la comparaison "byte by byte".

Pour des raisons de rapidité d'accès au marché [time2market], nous avons opté pour l'acquisition de Diligent [en avril 2008 *[cf .article IBM signe l »acquisition de Diligtent*] qui proposait déjà ce type de solutions depuis 2005 *[NDLR : globalement "byte by byte", octet par octet].*

Nous avons choisi la technologie de déduplication in-line plutôt que post-process. Car les informations sont ainsi stockées après traitement de déduplication. Et la taille du cache n'impacte pas les performances, car l'index reste en mémoire. Ce qui procure un avantage supplémentaire : l'évolutivité (scalability). En effet, les variations de cache liées à la montée en charge restent linéaires et ne dépendent que des capacités matérielles.

[NDLR : Avec la technologie in-line (parfois dite synchrone), le processus vérifie si la donnée à déjà été écrite (en mémoire). Si oui, il inscrit l'adresse du pointeur, sinon toute la donnée. En mode post-process (ou asynchrone), la donnée est entièrement écrite, puis un autre processus distinct (asynchrone) et plutôt sur un autre serveur vérifie si la donnée existe déjà. Si oui, il la détruit et la remplace par le pointeur, sinon, il n'intervient pas.]

Certains de nos concurrents proposent de petites boîtes dont l'accumulation montre vite les limites. Car plus leur nombre augmente, plus leur administration devient complexe, et consommatrice d'énergie et d'espace.

Le '*byte-by-byte*' présente un avantage par rapport au pur "hash-based", car il supprime le risque de collision. Nous avons aussi effectué ce choix technologique pour des raisons de performance, d'évolutivité, et pour son taux élevé d'intégrité des données.

Le segment de la bande n'est-il pas qu'un marché de remplacement ? Comment se répartissent les données stockées de vos clients sur ces supports ?

C.G.: Notre vision ne s'exprime pas en répartition de pourcentage de données sur bande ou sur disque. Nos clients s'équipent encore fortement en bandes, et en VTL (virtual Tape Library).

Certes, certains acteurs du marché prônent le "no-band". Mais, cela ne traduit nullement la réalité des entreprises. Des questions pertinentes seraient plutôt : « *Après combien de temps accédez-vous peu ou plus du tout à ces données ou ce type de données ?*» ou encore, « *De combien le volume de vos données croît-il par an ?*».

Outre l'explosion due aux usages numériques et aux applications, les obligations réglementaires obligent à conserver un volume de plus en plus conséquent. Et dans ce dernier cas, l'accès à cette information est très peu fréquent. Par ailleurs, la bande est plus écologique. Non seulement elle consomme moins d'énergie, mais grâce à la déduplication et ses taux de compression elle devient

encore plus intéressante et toujours aussi économique, avec de meilleures performances.

Par ailleurs, on rencontre souvent plusieurs duplications des informations : les données de production, la sauvegarde, et une copie de sécurité supplémentaire exigée par certaines compagnies d'assurance. Or, très souvent, une partie importante de ces données (bien plus de 50 %) ont très peu – ou pas – besoin d'être consultées au-delà de trois mois.

Les archives et le backup participent à l'explosion des équipements VTL, de + 35 % depuis plusieurs trimestres. Cependant, il ne s'agit pas uniquement d'un marché de remplacement des systèmes de stockage sur bande installés. On retrouve du VTL sur dans les environnements mainframes, mais aussi dans les environnements ouverts, et de plus en plus dans les environnements Windows. Car ces technologies proposent aujourd'hui une sauvegarde rapide, mais aussi et surtout une restauration très performante.

Bien entendu, le coût du stockage plaide en faveur de la bande. Ainsi, des environnements d'entreprises uniquement en disque reviennent 33 fois plus cher que leurs équivalents uniquement sur bande. Et si on ajoute la déduplication, on atteint un ratio de 5 pour 1 !

Vos solutions de déduplication Diligent ne sont finalement destinées qu'à la sauvegarde sur bande ? Et sur disques?

C.G.: Nous utilisons les technologies Diligent, mais également d'autres technologies IBM. En outre, la déduplication peut être utilisée pour un stockage primaire. Cependant, le ratio (et donc le gain d'espace) devient alors moins important, car il faut alors conserver un bon niveau de performances. En outre, certains applicatifs réclament de très hautes performances pour les données de production. C'est d'ailleurs un avantage supplémentaire de travailler avec des données dédupliquées avant stockage (post-process), une technique plus performante et évolutive. De plus, cela réduit fortement le besoin en bande passante. Or, les entreprises apprécient cette caractéristique, car elles possèdent très souvent des infrastructures de stockage ou de sauvegarde multisites.

À partir des technologies de Diligent, nous travaillons d'ailleurs sur des procédés concernant le *mirroring* sur disque, utilisant la déduplication pour l'envoi de données sur un autre site. Nous devrions annoncer des choses en ce sens en 2009.