Chapitre 4
Système de stockage intelligent

Pour bien fonctionner, les applications cruciales pour l'entreprise exigent de la part du système de stockage un haut niveau de performance, de disponibilité, de sécurité et d'évolutivité. Le disque dur est l'élément de principal du système de stockage et c'est lui qui dicte en quelque sorte les performances de celui-ci. Les plus anciennes technologies de regroupement de disques ne peuvent plus satisfaire aux contraintes de performance modernes à cause des limitations des disques durs et de leurs éléments mécaniques. La technologie RAID a contribué à améliorer les performances et la fiabilité des systèmes de stockage, mais elle aussi ne suffit plus aux exigences des applications modernes.

Encore une fois des améliorations technologiques ont donné naissance à une nouvelle famille d'équipements de stockage que l'on appelle *système de stockage intelligent*. Ces nouveaux systèmes, que nous verrons dans ce chapitre, sont des groupements RAID avec beaucoup de nouvelles fonctionnalités qui permettent d'optimiser au maximum le traitement des requêtes I/O. L'environnement opérationnel de ces systèmes (genre d'OS) contrôle la gestion, l'allocation et l'utilisation des ressources de stockage. Ces systèmes possèdent une très grande mémoire temporaire nommée *cache* et utilisent des algorithmes sophistiqués afin de satisfaire la vitesse de traitement des I/O exigée par les applications cruciales.

4.1 Les composants d'un système de stockage intelligent

Un système de stockage intelligent est composé de quatre éléments: la partie frontale, la mémoire cache, la partie avale et les disques physiques. Ces éléments et leurs interconnexions sont illustrés à la Figure 4-1. Une requête I/O en provenance d'un serveur est reçue par la partie frontale pour être ensuite dirigée vers le cache. Elle passe ensuite à la partie aval et finalement elle arrive aux disques physiques pour la

sauvegarde ou la récupération des données. Si la donnée se trouve déjà dans le cache, la requête n'a pas besoin d'aller jusqu'aux disques.

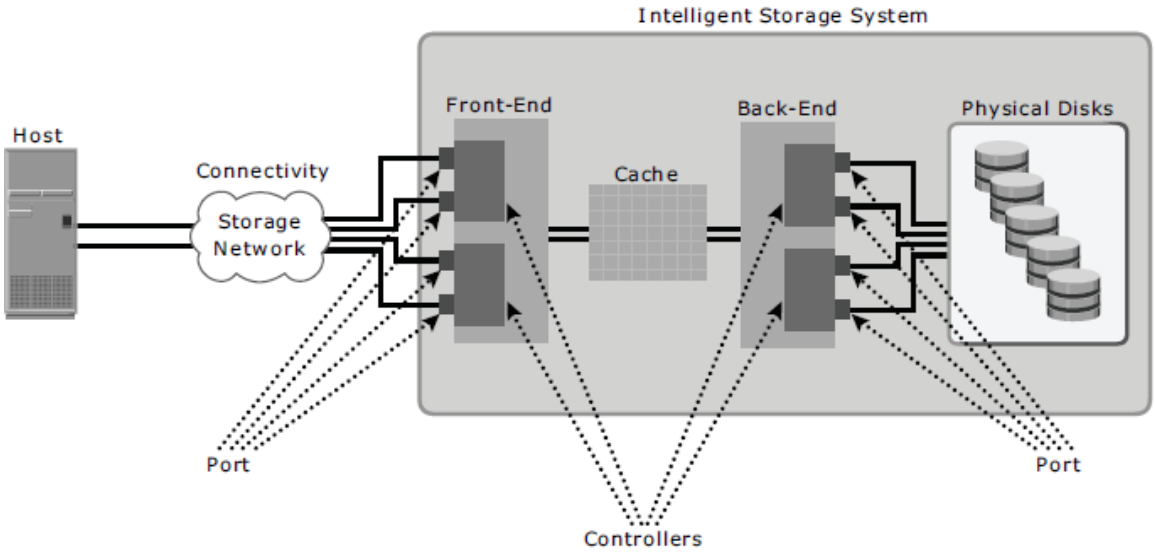

Figure 4-1: Components of an intelligent storage system

4.1.1 La partie frontale (front-end)

La partie frontale sert d'interface entre l'hôte et le système de stockage. Elle est constituée des ports frontaux et des contrôleurs frontaux. Les *ports frontaux* servent aux connexions physiques entre les hôtes et le système de stockage intelligent. Chaque port possède sa propre logique pour exécuter le protocole de transport approprié tel que SCSI, Fibre Channel ou iSCSI. Généralement les ports frontaux sont redondants afin de fournir une plus grande disponibilité des données.

Les contrôleurs frontaux acheminent les données entre les ports frontaux et le cache via un bus interne, et ce dans les deux sens. Lorsque le cache reçoit des données à écrire, il envoie un accusé de réception à l'hôte qui en a fait la requête. L'optimisation du traitement des requêtes I/O est assurée par des algorithmes de mise en queue des requêtes.

Mise en queue des commandes

La mise en queue des commandes est une technique implémentée sur les contrôleurs frontaux servant à déterminer l'ordre d'exécution des commandes reçues. Il est ainsi possible de réduire les mouvements inutiles des têtes de lecture des disques durs et d'augmenter la performance de l'ensemble. Quand une commande est reçue aux fins d'exécution, l'algorithme lui assigne une étiquette déterminant sa place dans l'ordre des

exécutions. Cette technique permet l'exécution de plusieurs commandes simultanées. Le nombre dépend de l'emplacement des données sur les différents disques, mais est indépendant de l'ordre dans lequel ces commandes ont été reçues. Les algorithmes les plus utilisés sont:

- **Premier arrivé premier sorti (FIFO):** C'est l'algorithme par défaut, les commandes sont exécutées dans l'ordre dans lequel elles sont reçues (Figure 4-2 [a]). En terme de performance il est inefficace, car il n'y a aucun réarrangement des demandes donc pas d'optimisation.
- **Optimisation du temps de positionnement (Seek Time):** Les commandes sont exécutées selon un ordre qui optimise les mouvements des têtes de lecture/écriture, ce qui peut nécessiter le réarrangement de l'ordre des commandes. Sans optimisation de temps de positionnement, les commandes sont exécutées selon leur ordre de réception. Par exemple, à la Figure 4-2 (a), les commandes sont exécutées dans l'ordre A, B, C et D. On peut voir que le mouvement radial requis par la tête pour exécuter A et ensuite C est inférieur à celui exigé pour exécuter A puis B. Dans le cas de l'optimisation de temps de positionnement, l'ordre d'exécution des commandes serait A, C, B et D, comme indiqué à la Figure 4-2 (b).

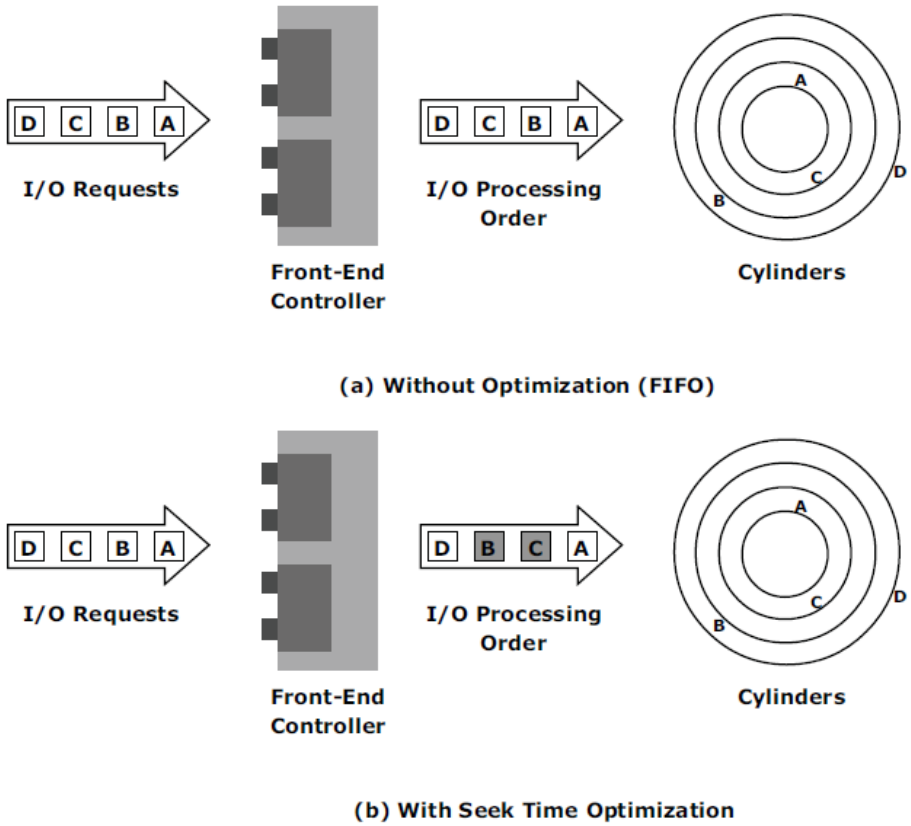

Figure 4-2: Front-end command queuing

- **Optimisation du temps d'accès:** Afin d'optimiser encore plus les performances, les commandes sont exécutées selon une combinaison de l'optimisation de temps de positionnement et d'une analyse de la latence rotationnelle.

La mise en queue des commandes peut également être implémentée sur les contrôleurs de disques, ce qui complémente la mise en queue des contrôleurs d'entrée. La mise en queue des commandes est déjà implémentée sur quelques modèles de contrôleurs de disques durs SCSI et Fibre Channel.

4.1.2 Le cache

Le cache est un élément essentiel pour l'amélioration des performances I/O d'un système intelligent de stockage. Le cache est une mémoire à semi-conducteurs où les données sont placées temporairement afin de réduire le temps requis pour répondre aux demandes I/O en provenance des serveurs.

Le cache améliore la performance du système de stockage en isolant l'hôte des retards mécaniques liés aux composants les plus lents du système que sont les disques physiques. L'accès aux données d'un disque physique prend habituellement quelques millisecondes en raison du temps de positionnement du bras et de la latence de rotation. Si on doit aller jusqu'au disque physique pour chaque opération I/O, on doit mettre les requêtes en queue à cause de la lenteur du disque, ce ajoute un délai à la réponse. L'accès aux données dans le cache prend quant à elle moins d'une milliseconde. Les données à écrire sont placées dans le cache puis écrites sur le disque. Quand les données sont placées dans le cache et sécurisées, l'hôte en est avisé immédiatement.

Structure du cache

Le cache est constitué de pages (ou plages). Ce sont les plus petites unités d'attribution d'espace dans un cache. La taille d'une page de cache est configurée selon la taille des requêtes I/O de l'application. Le cache comprend le *data store* et le *tag RAM.* Le data store contient les données tandis que le tag RAM mémorise l'emplacement des données dans le data store et sur le disque (cf. Figure 4-3).

Les entrées dans le tag RAM indiquent où les données se trouvent dans le cache et où celles-ci sont situées sur le disque. Le tag RAM inclut un drapeau (dirty bit), qui indique si les données dans le cache ont été sauvegardées sur le disque ou pas. Il contient également des informations comme le temps du dernier accès. On s'en sert pour identifier les données contenues dans le cache n'ayant pas été accédé pendant une longue période et qui peuvent être libérées.

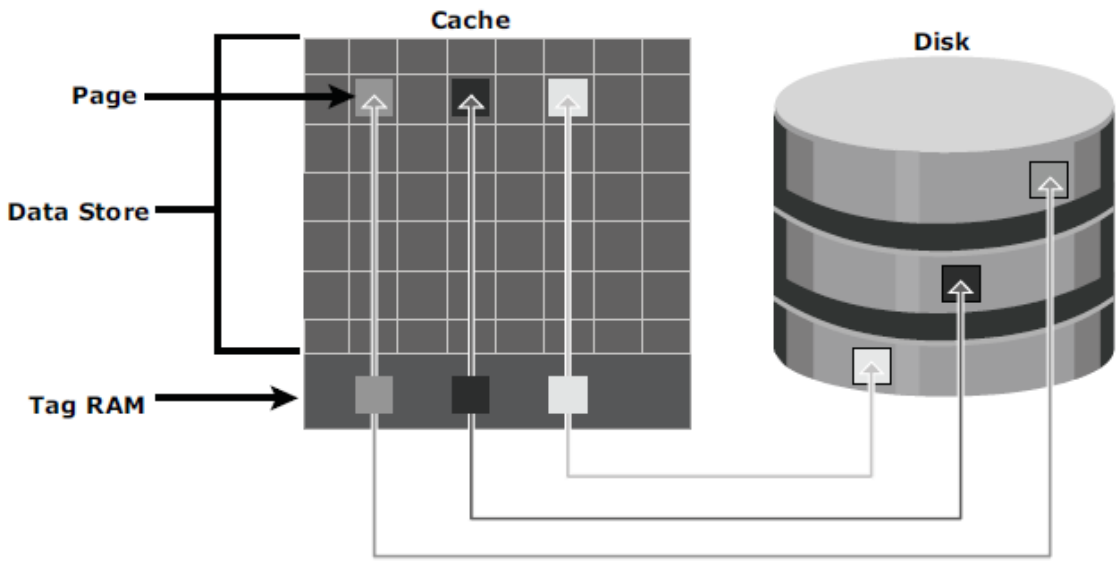

Figure 4-3: Structure of cache

Opération de lecture avec le cache

Quand un hôte fait un requête en lecture, le contrôleur frontal consulte le tag RAM pour savoir si l'information demandée est disponible dans le cache. Si c'est le cas, cela s'appelle un "*cache hit*" ou "*read hit*" et les données sont envoyées directement vers l'hôte sans qu'il y ait une opération sur les disques (cf. Figure 4-4[a]). Le temps de réponse pour l'hôte est d'environ une milliseconde. Si la requête ne se trouve pas dans le cache, on a alors un "*cache miss*" et les données doivent être lues sur le disque (cf. Figure 4-4[b]). Le contrôleur aval doit alors récupérer les données sur le disque et les placer dans le cache. Le contrôleur frontal les acheminera ensuite vers l'hôte. Les "cache miss" augmentent le temps de réponse des requêtes I/O.

Une *lecture anticipée* (pre-fetch ou read-ahead) est un algorithme utilisé lors de requêtes en lecture séquentielles. Lors d'une lecture séquentielle, un ensemble contigu de blocs associés est lu et placé dans le cache. Ainsi plusieurs blocs d'informations non encore demandés mais ayant un lien avec la requête originale sont placés dans le cache dans l'espoir qu'ils seront les prochains. Si c'est le cas et que l'hôte fait appel à eux, on a un cache hit, on sauve des appels aux disques et on améliore les performances. La grosseur des blocs de lectures anticipée dans les systèmes de stockage intelligents peut être fixe ou variable. Dans le cas de la *lecture anticipée de grosseur fixe*, une quantité fixe de données est lue et mise en cache. Cette méthode est surtout utile quand le requêtes I/O sont toutes de la même taille. Dans le cas de la méthode *variable*, la quantité de données récupérée est un multiple de la taille de la requête faite par l'hôte. On fixe généralement un maximum au nombre de blocs de données pouvant être récupéré par

anticipation afin de ne pas trop accaparer les disques et ainsi nuire aux autres requêtes auxquelles ils doivent répondre.

Les performances en lecture se mesurent en termes de proportion de succès de lectures anticipées (read hit ratio ou hit rate) exprimée généralement en pourcentage. Le ratio est égal au nombre de succès par rapport au nombre total de requêtes en lecture reçues. Plus le ratio est élevé, meilleure est la performance.

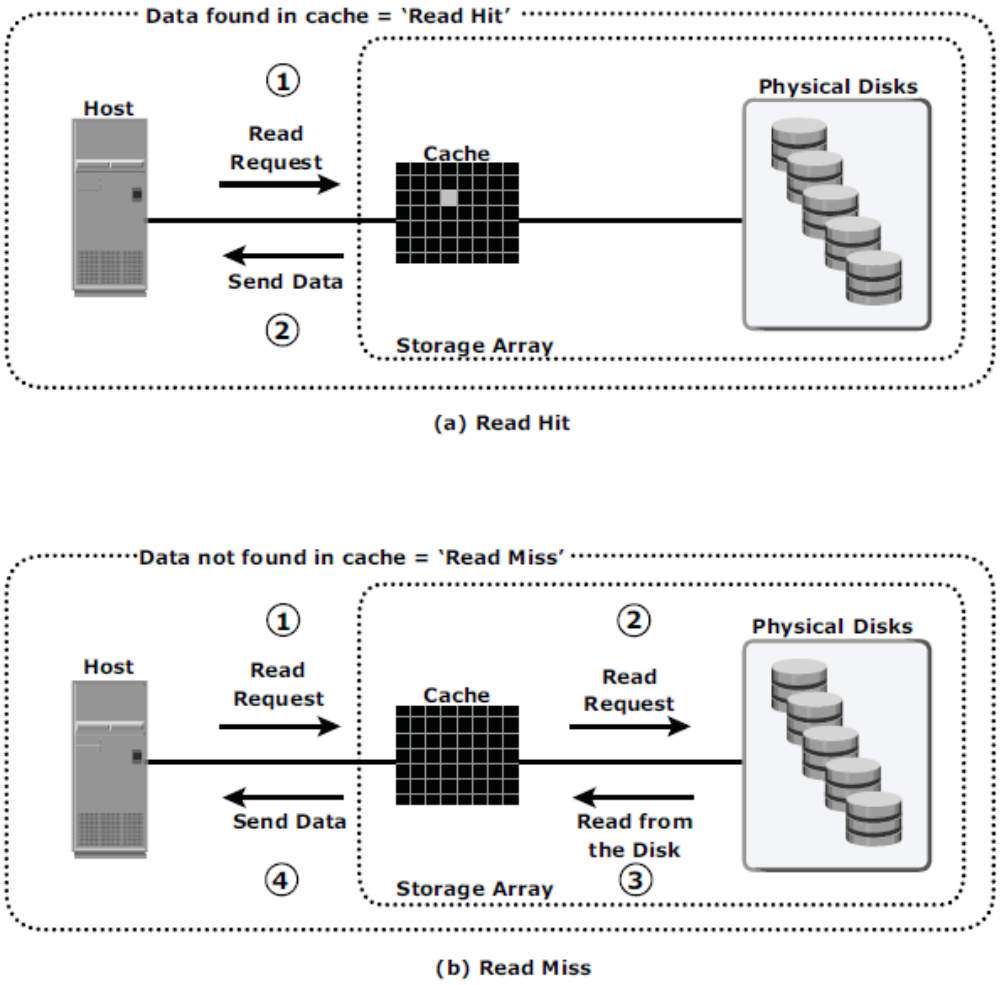

Figure 4-4: Read hit and read miss

Opération d'écriture avec le cache

Utiliser un cache lors des opérations d'écriture plutôt que l'écriture directe sur les disques améliore les performances. Écrire les données dans le cache et envoyer un accusé de réception prend moins de temps que l'écriture sur le disque physique ainsi pour l'hôte, qui ne voit que la partie cache, le temps de réponse semble plus rapide. L'écriture séquentielle permet aussi d'optimiser un peu plus en regroupant plusieurs

petites écritures ensemble pour former un seul gros bloc dans le cache qui sera transféré en une seule opération séquentielle sur le disque.

Il y a deux types d'implémentation de l'écriture avec cache:

- **Write-back cache:** Les données sont placées dans le cache et un accusé de réception est envoyé immédiatement vers l'hôte. Plus tard, les données de plusieurs requêtes d'écriture sont envoyées et sauvegardées sur le disque (committed ou de-staged). La vitesse d'écriture est beaucoup plus rapide, car le délai de la portion mécanique est séparé de la portion cache. Cependant les données qui n'ont pas encore été sauvegardées sur le disque peuvent être perdues advenant une défaillance du cache.
- **Write-through cache:** Les données sont placées dans le cache et immédiatement envoyées vers le disque. Ensuite seulement l'accusé de réception est envoyé vers l'hôte. Comme les données sont à la fois dans le cache et sur le disque, la perte de données est faible, mais le temps de réponse et plus long à cause des délais du disque physique.

Le cache peut être contourné sous certaines conditions comme des requêtes I/O en écriture de très grandes tailles. Dans ce type d'implémentation, si la taille d'une requête I/O dépasse la taille permise, appelée *write aside size*, les opérations d'écritures sont envoyées directement vers les disques pour réduire l'impact causé par un trop grand espace du cache accaparé par de gros fichiers d'écriture. Cette technique sert surtout dans les environnements où le cache est restreint et qu'il doit rester disponible pour les petites requêtes I/O qui surviennent aléatoirement.

Implémentation d'un cache

Le cache peut être dédié ou global. Dans le cas d'un cache dédié, des blocs de mémoire séparés sont réservés pour les opérations de lecture et d'écriture. Quant au cache global, les opérations de lecture et d'écriture peuvent utiliser les mêmes blocs mémoires. La gestion du cache en mode global est plus efficace, car une seule table d'adresse est utilisée.

Le mode cache global permet à l'usager de spécifier le pourcentage de cache disponible pour la gestion des opérations de lecture/écriture. Le cache de lecture est généralement petit, mais il devrait être agrandi si l'application est surtout orientée en lecture. Dans d'autres implémentations du mode global, le ratio entre l'espace pour la lecture et l'écriture est ajusté dynamiquement selon la charge de travail.

Gestion du cache

Le cache est une ressource de grosseur finie et dispendieuse qui nécessite une gestion appropriée. Même si un système de stockage intelligent peut être configuré avec une grande capacité de cache, quand toutes les pages sont remplies il faut faire de la place en libérant des pages pour faire de la place aux nouvelles. Il existe plusieurs algorithmes de gestion de cache qui permettent de maintenir de façon proactive un nombre de pages libres ainsi qu'une liste de pages pouvant être effacées en cas de besoin:

- **Least Recently Used (LRU):** Un algorithme qui surveille constamment l'accès aux données du cache et identifie les pages du cache qui n'ont pas été accédées depuis longtemps. LRU libère ces pages ou les marques pour la réutilisation. Cet algorithme est fondé sur l'hypothèse que des données qui n'ont pas été accédées depuis un moment ont peu de chances d'être demandées par l'hôte. Cependant, si une page contient des données à écrire, mais que celles-ci ne sont pas encore sur le disque, ces données seront d'abord inscrites sur le disque avant que la page soit réutilisée.
- **Most Recently Used (MRU):** Cet algorithme est l'inverse de LRU. Dans MRU, les pages qui ont été accédées récemment sont libérées ou marquées pour la réutilisation. Cet algorithme est fondé sur l'hypothèse que ce qui à été récemment utilisé ne devrait pas nous être utile avant un bon un moment.

Pendant que le cache se remplit, le système de stockage doit s'occuper des *dirty pages* (données écrites dans le cache, mais pas encore écrites sur le disque) afin de maintenir une bonne disponibilité. Le délestage (flushing) est le processus de forcer l'écriture des données du cache sur les disques. Des niveaux de *filigranes*(Watermarks) basés sur le taux et l'historique des accès I/O, permettent de contrôler le processus de délestage du cache. *Le haut filigrane (HWM)* est le niveau d'utilisation de cache auquel le système de stockage commence le délestage obligatoire et accéléré des données du cache. *Le bas filigrane (LWM)* est le point auquel le système de stockage arrête le délestage accéléré ou obligatoire et revient au comportement de délestage au ralenti. Le niveau d'utilisation du cache dicte le mode de délestage à utiliser (cf. Figure 4-5) :

- **Délestage au ralenti (idle):** Se produit sans interruption, à un taux modeste, quand le niveau d'utilisation du cache est entre le filigrane haut et bas.
- **Délestage accéléré (High):** Activé quand l'utilisation de cache atteint le haut filigrane. Le système de stockage consacre alors des ressources additionnelles au délestage. Ce type de délestage a un impact minimal sur le traitement des requêtes I/O de l'hôte.
- **Délestage forcé (forced):** Se produit lors d'une arrivée massive de requêtes et que le cache atteint 100 pour cent de sa capacité. Ceci affecte de manière significative le temps de réponse. Dans le cas du délestage obligatoire, des "dirty pages" sont délestées de force vers le disque.

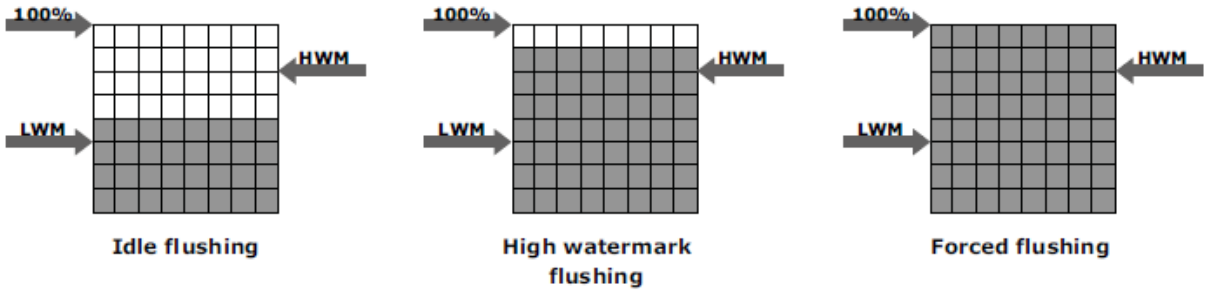

Figure 4-5: Types of flushing

Protection des données du cache

Le cache est une mémoire volatile ce qui implique qu'une panne de courant ou autre défaillance du cache entrainera la perte des données non encore inscrites sur le disque. Ce risque peut être atténué en utilisant le *cache mirroring* et le *cache vaulting* :

- **Cache mirroring:** Chaque écriture dans le cache est placée à deux endroits différents en mémoire sur deux cartes de mémoire indépendantes. En cas d'échec d'un des caches, les données sont disponibles dans le cache miroir et peuvent toujours être envoyées vers le disque. En lecture, les données vont du disque vers le cache donc, en cas de défaillance du cache, les données peuvent toujours être accédées à partir du disque. Pour une meilleure utilisation du cache disponible, seulement les données du cache en écriture font l'objet d'un miroir.
- **Cache vaulting:** Les données non engagées du cache sont vulnérables en cas de panne de courant. Ce problème peut être abordé de diverses manières comme alimenter la mémoire avec une batterie jusqu'au retour du courant alternatif ou utiliser l'énergie d'une batterie pour écrire le contenu de cache sur le disque. En cas de panne de courant prolongée, l'utilisation de batteries n'est pas une option viable parce que dans les systèmes intelligents de stockage, de grandes quantités de données peuvent devoir être sauvegardées sur de nombreux disques et les batteries peuvent ne pas être en mesure de fournir la puissance suffisante et durer assez longtemps pour écrire chaque donnée sur le disque prévu. Par conséquent, les fournisseurs de stockage emploient un ensemble de disques physiques pour vider le contenu du cache pendant la panne de courant. Ceci s'appelle *cache vaulting* et les disques s'appellent *vault drives*. Au retour de la panne, les données de ces disques sont réintroduites dans le cache puis sauvegardées sur les disques prévus.

4.1.3 La partie aval (back-en)

La partie aval sert d'interface entre le cache et les disques physiques. Elle se compose de deux éléments : les ports en aval et les contrôleurs en aval. La partie aval contrôle les échanges entre le cache et les disques physiques. Du cache, les données sont envoyées à la partie aval puis dirigées vers le disque de destination. Les ports sur la partie aval sont reliés à des disques physiques. Il y a communication entre le contrôleur de la partie aval et les disques physiques lors des opérations de lectures et d'écritures. Le contrôleur fournit aussi un espace de stockage de données additionnel provisoire, mais très limité.

Les algorithmes implémentés sur les contrôleurs sur la partie aval permettent la détection et la correction d'erreur ainsi que la fonctionnalité RAID.

Comme pour la partie frontale, les systèmes de stockage utilisent la redondance de contrôleurs et les ports multiples afin de protéger les données et d'augmenter la disponibilité de données. Une telle configuration fournit une voie alternative vers les disques physiques en cas de défaillance d'un contrôleur ou d'un port. Cette fiabilité est encore augmentée si les disques sont également accessibles par deux ports. Dans ce cas, chaque port de disque peut être relié à un contrôleur différent. L'utilisation de contrôleurs multiples facilite également l'équilibrage de la charge de travail.

4.1.4 Le disque physique

Un disque dur physique emmagasine les données de façon permanente. Les disques sont reliés à la partie aval au moyen d'une interface SCSI ou Fibre Channel (discutée dans des chapitres subséquents). Un système intelligent de stockage permet l'utilisation d'un mélange de disques SCSI, Fibre Channel et IDE/ATA/SATA.

SOLID-STATE DRIVES

Flash-based solid-state drives (SSDs) are a recent innovation for delivering ultra-high performance for mission-critical applications. Solid-state Flash drives utilize Flash memory to store and retrieve data. Unlike FC or SATA drives, Flash drives have no moving parts, and leverage semiconductor-based block storage devices, resulting in minimized response time and less power requirements to run. Flash drives are constructed with nonvolatile semiconductor memory to support persistent storage and they use either single-level cell (SLC) or multi-level cell (MLC) to store bits on each memory cell. SLC stores one bit per cell and is used in high-performance memory cards. MLC memory cards store more bits per cell and provide slower transfer speeds. The advantage of MLC over SLC memory cards is the lower manufacturing cost.

Flash drives that use SLC technology combined with sophisticated controllers can behave like virtual HDDs through a traditional storage interface (such as Fibre Channel) to achieve ultra-fast read/write performance, high reliability, and data integrity. Flash drives have been tested and qualified to withstand the intense workloads of high-end enterprise storage applications.

Flash storage technology is ideally suited to support applications that need to process massive amounts of information very quickly, such as currency exchange and electronic trading systems, real time data feed processing, mainframe transaction processing, and many others. Storage systems with enterprise-class Flash drives can deliver single-millisecond application response times, up to 10 times faster than those with traditional 15K RPM Fibre Channel disk drives.

In a storage array, Flash drives can store a terabyte of data using 38 percent less energy than traditional mechanical disk drives. It would take 30 15K RPM Fibre Channel disk drives to deliver the same performance as a single Flash drive, which translates into a 98 percent reduction in power consumption in a transaction-per-second comparison.

Logical Unit Number (LUN)

Les disques physiques ou les groupes de disques protégés par RAID peuvent être divisés en volumes logiques, généralement désignés sous le nom des *numéros*

d'éléments logique (LUNs). L'emploi de LUNs permet une meilleure utilisation du disque. Par exemple, sans l'utilisation de LUNs, un hôte exigeant seulement 200 GB pourrait se voir assigner la totalité du disque physique de 1TB. Si on utilise les LUNs, seulement les 200 GB demandés seraient assignés à l'hôte, permettant aux 800 GB restant d'être assignés à d'autres hôtes.

Dans le cas de groupe de disques en RAID, les LUNs sont des tranches du RAID réparties sur l'ensemble des disques qui forme le RAID. Le LUN peut aussi être vu comme une partition logique du RAID qui serait présenté à l'hôte comme étant un disque physique. Par exemple, à la Figure 4-6 on montre un groupe RAID formé de cinq disques qui ont été découpés (ou partitionnés) en plusieurs LUNs. Seuls les LUNs 0 et 1 sont montrées sur le schéma.

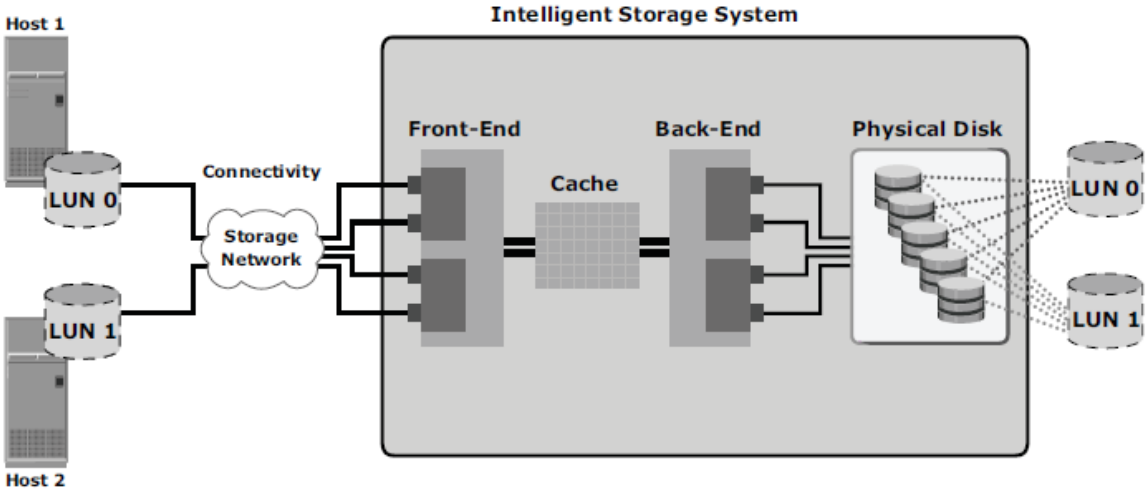

Figure 4-6: Logical unit number

Notez comment une partie de chaque LUN réside sur chaque disque physique dans l'ensemble RAID. Les LUNs 0 et 1 sont présentés aux hôtes 1 et 2, respectivement, en tant que volumes physiques pour stocker et récupérer des données. La capacité utilisable des volumes physiques est déterminée par le type de RAID implémenté.

La capacité d'un LUN peut être augmentée en agrégeant d'autres LUNs. Le résultat de cette agrégation est un LUN de plus grande capacité, appelé *méta-LUN*. Le mappage des LUNs à leurs emplacements physiques sur les disques durs est contrôlé par l'environnement opérationnel (OS) du système de stockage intelligent.

LUN Masking

Le masquage de LUN est un processus de gestion d'accès aux données en définissant à quel LUN un hôte droit d'accéder. La fonction de masquage de LUN est généralement appliquée sur le contrôleur de la partie frontale. On s'assure ainsi d'un contrôle approprié de l'accès au LUN par les serveurs en empêchant leur usage non autorisé ou accidentel dans un environnement distribué.

Par exemple, considérons une baie de stockage avec deux LUNs qui stockent les données des ventes et des finances. Sans le masquage de LUN, les deux départements peuvent facilement voir et modifier des données de l'un comme de l'autre, causant une grande menace à l'intégrité et à la sécurité des données. Avec le masquage de LUN, les LUNs sont accessibles seulement aux hôtes à qui ont a donné les droits.

4.2 Baie de stockage intelligente

Les systèmes de stockage intelligents tombent généralement dans une des deux catégories suivantes :

- Les systèmes de stockage haut de gamme
- Les systèmes de stockage milieu de gamme

Traditionnellement, les systèmes de stockage haut de gamme ont été implémentés avec des *matrices actives-actives,* tandis que les systèmes de stockage de milieu de gamme, utilisés typiquement par les petites et moyennes entreprises, ont été implémentés avec des *matrices actives-passives*. Les matrices actives-passives fournissent des solutions de stockage optimales à coûts moindre. Les entreprises se servent de cet avantage monétaire pour implémenter des matrices actives-passives afin de répondre à des exigences spécifiques telles que la performance, la disponibilité, et l'évolutivité. Il existe de moins en moins de distinctions entre ces deux implémentations.

4.2.1 Système de stockage haut de gamme

Les systèmes de stockage haut de gamme, appelés *matrices actives-actives*, sont généralement destinés aux grandes entreprises pour centraliser les données corporatives. Ces matrices sont conçues avec un grand nombre de contrôleurs et de caches. Une matrice active-active implique que le hôte peut envoyer de requêtes I/Os vers son LUN par tous les chemins disponibles (voir la Figure 4-7).

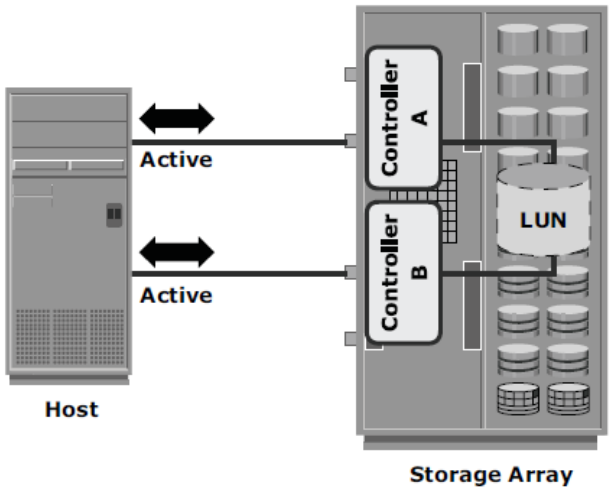

Figure 4-7: Active-active configuration

Pour combler les besoins en stockage de données de l'entreprise, ces matrices fournissent les fonctionnalités suivantes:

- Grande capacité de stockage
- Cache de grande capacité pour optimiser le traitement des requêtes I/O des hôtes
- Architecture avec tolérance aux fautes pour améliorer la disponibilité de données
- Connectivité avec les ordinateurs centraux et aux systèmes ouverts

- Disponibilité de plusieurs types de ports d'entrée et de protocoles d'interface pour satisfaire un grand nombre d'hôtes
- Disponibilité de plusieurs types de contrôleurs de disques en aval comme SCSI RAID ou Fibre Channel
- Évolutivité pour améliorer la connectivité, la performance et la capacité de stockage
- Capacité de manipuler des grands nombres d'I/O simultanés en provenance de plusieurs serveurs et d'applications
- Soutien pour la réplication locale et à distance

En plus de ces dispositifs, les matrices haut de gamme possèdent quelques dispositifs et fonctionnalités uniques nécessaires aux applications cruciales des grandes entreprises.

4.2.2 Système de stockage milieu de gamme

Les systèmes de stockage de milieu de gamme, également *appelés matrices actives-passives*, sont mieux adaptés pour les petites et moyennes entreprises. Dans une matrice active-passive, un hôte peut faire des requêtes I/O vers un LUN uniquement par les chemins du contrôleur qui possède le LUN. Ces chemins s'appellent les *chemins actifs*. Les autres chemins vers ce LUN sont dit passifs. La Figure 4-8 montre que l'hôte peut faire des opérations de lecture et d'écriture vers le LUN seulement via le chemin du contrôleur A, car le contrôleur A est le propriétaire de ce LUN. Le chemin via le contrôleur B reste passif et il n'y a aucune activité d'entrée-sortie sur ce chemin.

Les systèmes de stockage de milieu de gamme sont typiquement conçus avec deux contrôleurs qui contiennent chacun des interfaces pour les hôtes, un cache, des contrôleurs RAID et des interfaces pour les disques durs.

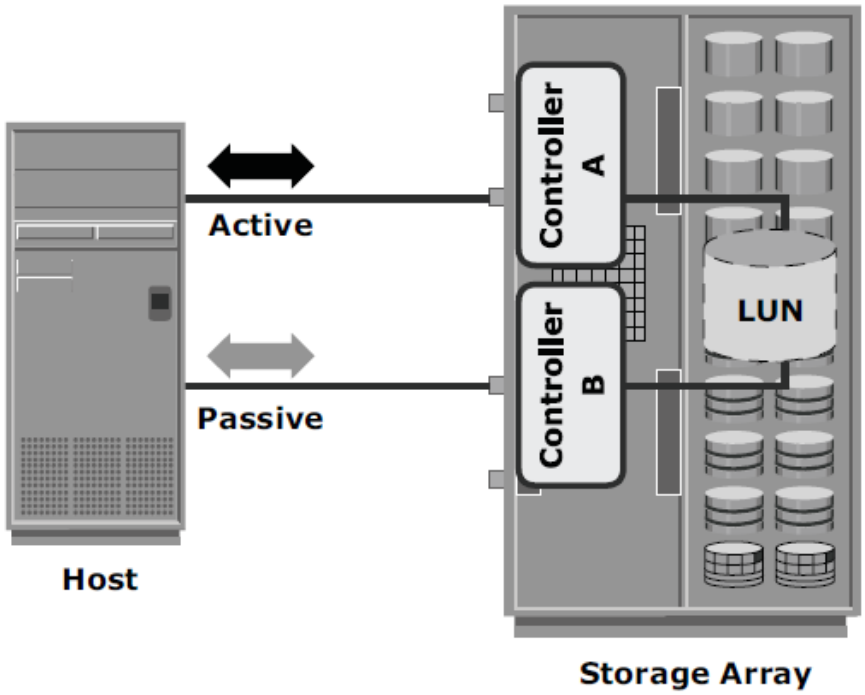

Figure 4-8: Active-passive configuration

Les matrices de milieu de gamme sont conçues pour répondre aux exigences des petites et moyennes entreprises donc elles possèdent moins de capacité de stockage et de cache global que les matrices actives-actives. Il y a également moins de ports d'entrée pour le raccordement aux serveurs. Cependant, elles assurent une bonne redondance et une haute performance pour les applications tout en ayant une charge de travail prévisible. Elles permettent également la réplication locale et à distance.

4.3 Mise en pratique: EMC CLARiiON en Symmetrix

Afin d'illustrer les concepts précédemment énoncés, cette section montre l'implémentation des baies de stockage intelligentes de la compagnie EMC.

La baie de stockage CLARiiON est une matrice active-passive conçue par EMC. Il s'agit de la baie de stockage de milieu de gamme offerte par EMC offrant des fonctionnalités de qualités industrielles. Elle est spécialement bien adaptée aux applications dont la charge de travail est prévisible et qui ont besoin d'un temps de réponse allant de moyen à élevé.

La baie de stockage en réseau d'EMC se nomme Symmetrix et est une matrice active-active. Symmetrix est une solution pour les clients qui exigent un niveau de service et de performance sans compromis. Elle se veut aussi la solution idéale pour une continuité d'affaires plus avancée en supportant d'énormes charges de travail non constantes. Symmetrix fournit également les dispositifs intégrés de sécurité de haute qualité et fait une des plus efficaces utilisations de l'énergie et du refroidissement dans le domaine.

Pour les dernières informations sur CLARiiON et Symmetrix, aller à la page *http://education.EMC.com/ismbook*.

4.3.1 Baie de stockage CLARiiON

La série CX4 est la plate-forme de stockage de quatrième génération de CLARiiON CX. Chaque génération a ajouté des perfectionnements à la performance, à la disponibilité, et à l'évolutivité à la génération précédente tandis que l'architecture des niveaux supérieurs demeure la même. La Figure 4-9 montre une baie de stockage CLARiiON d'EMC.

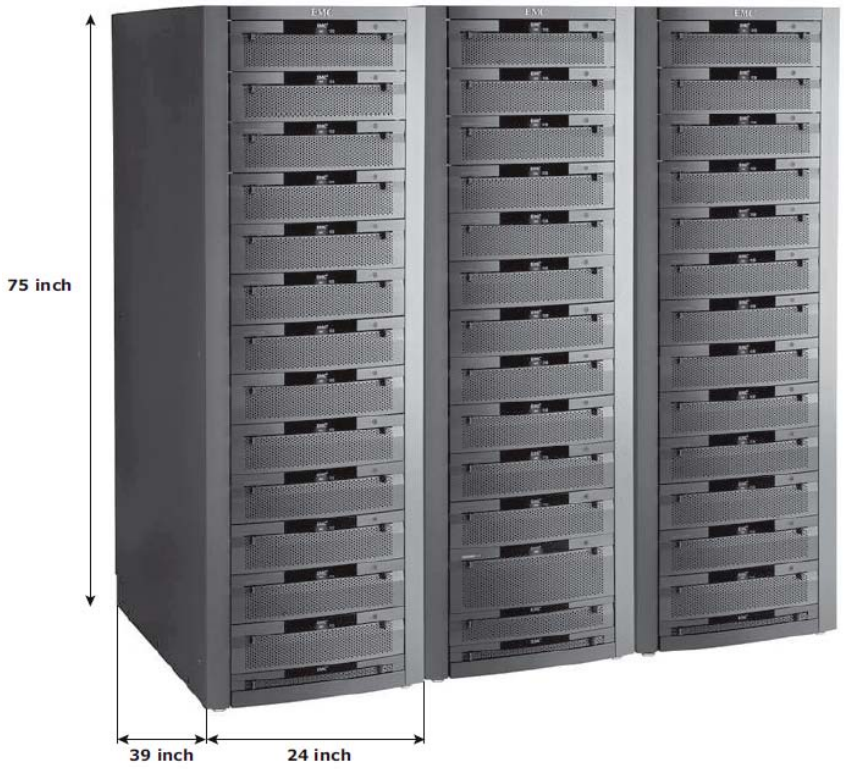

Figure 4-9: EMC CLARiiON

CLARiiON est construit avec des blocs modulaires et sans point de défaillance unique. CLARiiON CX4 est la première baie de stockage de milieu de gamme qui supporte les disques flash pouvant avoir un IOPS 30 fois plus élevé. Les autres fonctionnalités de CLARiiON sont:

- Technologie *UltraFlex* pour les protocoles de connectivité dual, expansion en ligne par l'intermédiaire des modules I/O, et prêt pour les futures technologies telles que iSCSI 10GB/s et Fibre Channel 8 GB/s
- Expansion jusqu'à 960 disques durs
- Supporte différents types et différentes tailles de disques ainsi que différents types de RAID (0, 1, 1+0, 3, 5, 6)
- Supporte jusqu'à 16 GB de cache par contrôleur (SPE)
- Augmente la disponibilité grâce à la mise à niveau et le recouvrement d'erreur sans arrêt de service
- Assure la protection des données à l'aide du cache mirroring et du cache vaulting
- Fournis l'intégrité des données par le nettoyage du disque. Le procédé de vérification fonctionne continuellement en sourdine et lit tous les secteurs de tous les disques. Si un bloc est illisible, le système de gestion d'erreur récupère les données des secteurs défectueux à partir de la parité ou des données miroirs.
- Supporte la réplication locale et à distance pour la sauvegarde de secours et la récupération après désastre à l'aide des logiciels SnapView et MirrorView.

4.3.2 Architecture du CLARiiON CX4

Le Storage Processor Enclosure (SPE) et le *Disk Array Enclosure (DAE)* sont les blocs modulaires principaux constituant une baie CLARiiON. Une DAE contient jusqu'à 15 unités de disques, deux cartes de contrôle de lien (LCCs), deux alimentations en énergie, et deux modules de ventilation. Un SPE contient deux processeurs de stockage, chacun se composant d'un module CPU et de fentes pour des modules I/O. La Figure 4-10 montre l'architecture du CLARiiON CX4.

L'architecture de CLARiiON supporte les composants redondants remplaçables à chaud. Ceci signifie que le système peut survivre avec un composant défaillant qui peut être remplacé sans être réinitialisé. Les composants importants du système de stockage de CLARiiON incluent :

- **Processeur de stockage intelligent (SP) :** Le PS intelligent est le composant principal de l'architecture du CLARiiON. Le PS est configuré en paires pour un maximum de disponibilité. Le PS fournit la connectivité frontale pour l'hôte et avale pour les disques physiques. Le PS inclut également la mémoire dont la majorité est employée pour le cache. Selon le modèle, chaque PS inclut un ou deux CPU.
- **Interface de transmission de messages de CLARiiON (CMI) :** Les SP communiquent entre eux grâce au CMI qui transporte entre les SP les commandes, les informations de statut et les données à écrire dans le cache miroir. Le CLARiiON utilise PCI-Express en tant que CMI à grande vitesse. L'architecture de PCI Express possède une grande largeur de bande par fil, a des propriétés de routage supérieures et améliore la fiabilité.
- **Standby Power Supply (SPS):** En cas d'une panne de courant, le SPS maintient une alimentation en énergie dans le cache assez longtemps pour permettre au contenu d'être copié dans le cache vault.
- **Link Control Card (LCC):** Le LCC fournit des services aux modules de disques dont la possibilité de contrôler des fonctionnalités des modules et la surveillance des variables environnementales. Chaque module de disque possède deux LCC. Les autres fonctions du LCC sont le contrôle de la configuration en boucle, le contrôle de la reprise après incident, le contrôle du marqueur LED, le contrôle individuel de port de disque, détection de présence de disque et information sur statut de la tension.
- **FLARE Storage Operating Environment:** FLARE est un logiciel spécial conçu pour le CLARiiON. Chaque système de stockage est livré avec une copie complète du système d'exploitation FLARE installé sur ses quatre premiers disques. Quand CLARiiON est mis sous tension, chaque PS démarre et le système d'exploitation FLARE. FLARE gère l'attribution des ressources et exécute les autres tâches de gestion dans la matrice.

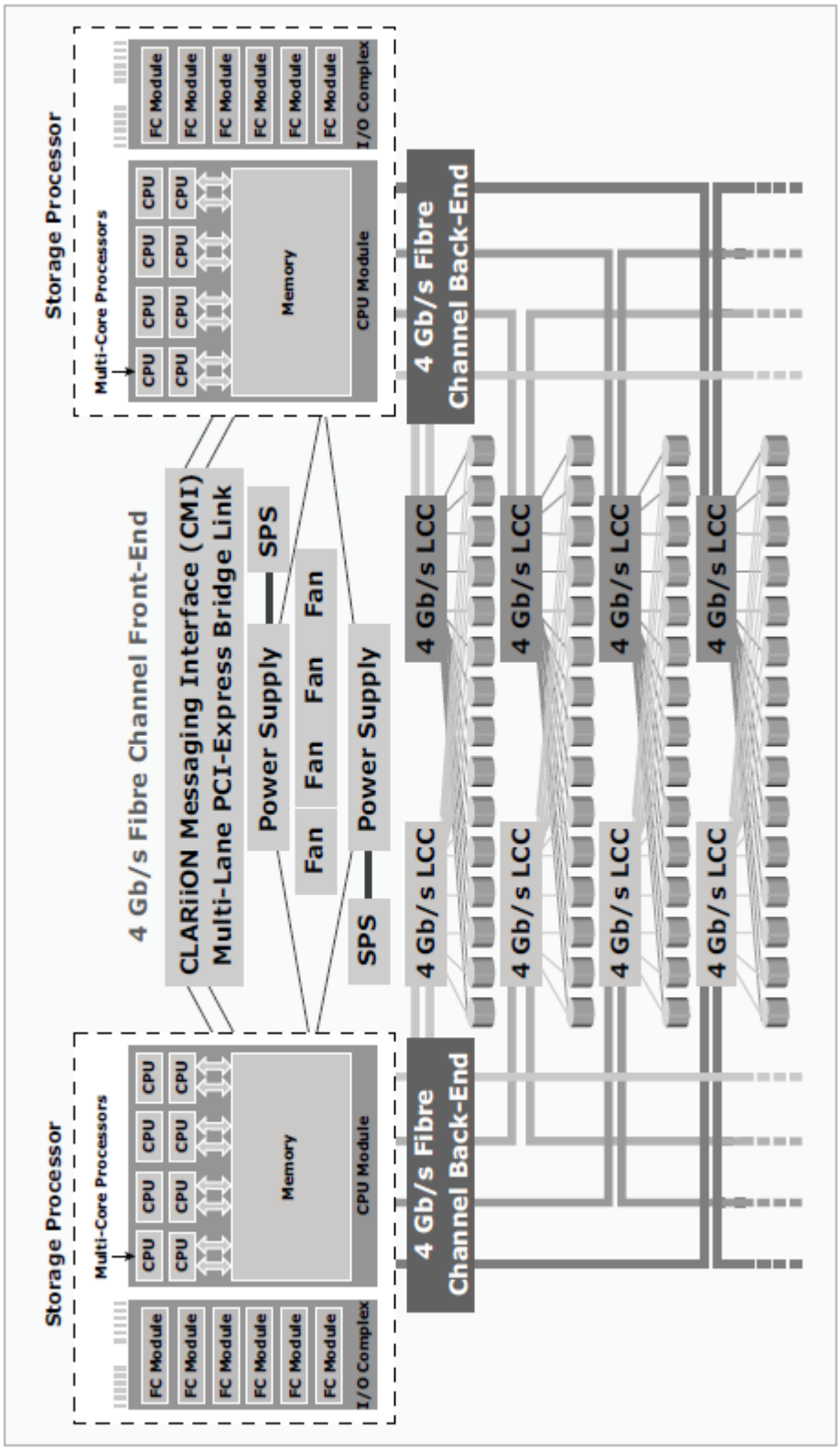

Figure 4-10: CLARiiON architecture

4.3.3 Gestion d'un CLARiiON

Pour la gestion, CLARiiON supporte l'interface de ligne de commande (CLI) et l'interface utilisateur graphique (GUI). *Naviseccli* est un outil de gestion CLI. Les commandes pour la gestion du système peuvent être envoyées à partir du système hôte directement branché ou d'un serveur à distance via Telnet/SSH.

Le logiciel de gestion ***Navisphere*** est une suite d'outils GUI qui permet la gestion centralisée des systèmes de stockage CLARiiON. Ces outils sont utilisés, pour surveiller, configurer, et contrôler des baies de stockage CLARiiON. La suite de gestion de Navisphere inclut ce qui suit :

- Gestionnaire Navisphere: Un outil GUI pour la gestion centralisée d'un système de stockage qui sert à configurer et contrôler CLARiiON. C'est une interface utilisateur WEB qui aide à gérer de façon sécuritaire des systèmes de stockage de CLARiiON localement ou à distance à l'aide d'un lien IP tout en utilisant un navigateur ordinaire. Le gestionnaire Navisphere offre la flexibilité nécessaire pour contrôler un ou plusieurs systèmes.

- Analyseur Navisphere: Un outil d'analyse de performance pour des composants matériels de CLARiiON.

- Agent Navisphere: Un outil résident sur l'hôte qui fournit un moyen de communication pour la gestion du système et permet l'accès à l'aide de CLI.

4.3.4 Baie de stockage Symmetrix

La baie de stockage Symmetrix de EMC offre les niveaux de performances et de capacité les plus élevés pour une solution de stockage d'informations corporative et est reconnue comme plate-forme de stockage la plus fiable de l'industrie. La Figure 4-11 montre la baie de stockage Symmetrix DMX-4 d'EMC.

Symmetrix emploie "Direct Matrix Architecture" et incorpore une conception avec tolérance aux défaillances. Les autres fonctionnalités du Symmetrix sont:

- Évolutif par incrément jusqu'à 2 400 disques
- Supporte les disques flash (SSD)
- Cache global dynamique (16 GB - 512 GB)
- Puissance de calcul accrue (jusqu'à 130 PowerPC)
- Grand nombre de routes disponibles pour les transferts I/O (32 - 128 routes)
- Largeur de bande très élevée pour les traitements des données (jusqu'à 128 GB/s)
- Protection des données RAID 1, 1+0 (ou 10), 5 et 6
- Réplication locale et distante pour la continuité des affaires avec les logiciels TimeFinder et SRDF

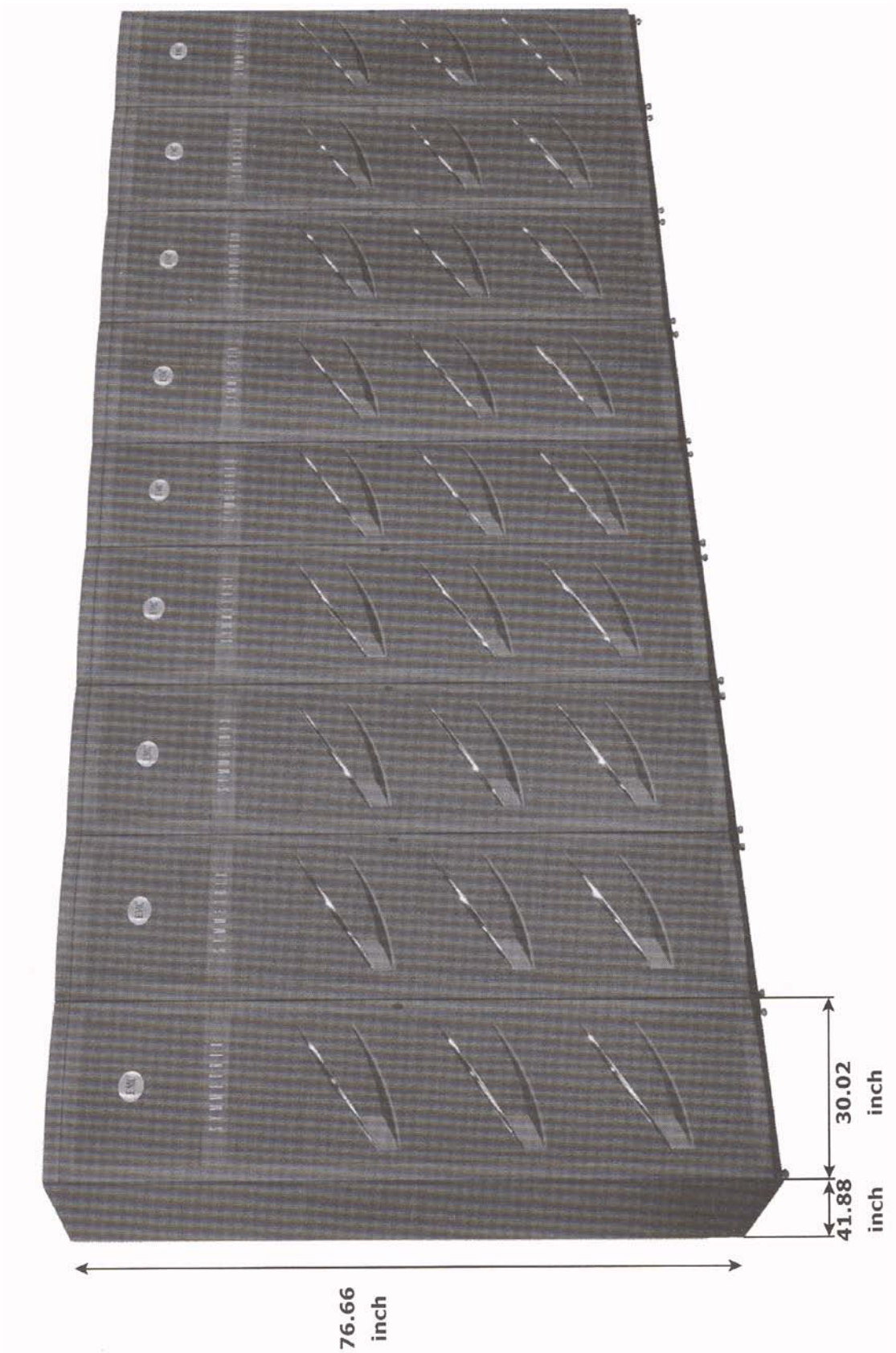

Figure 4-11: EMC Symmetrix

4.3.5 Survol des composants d'un Symmetrix

La Figure 4-12 montre le schéma fonctionnel d'une baie Symmetrix.

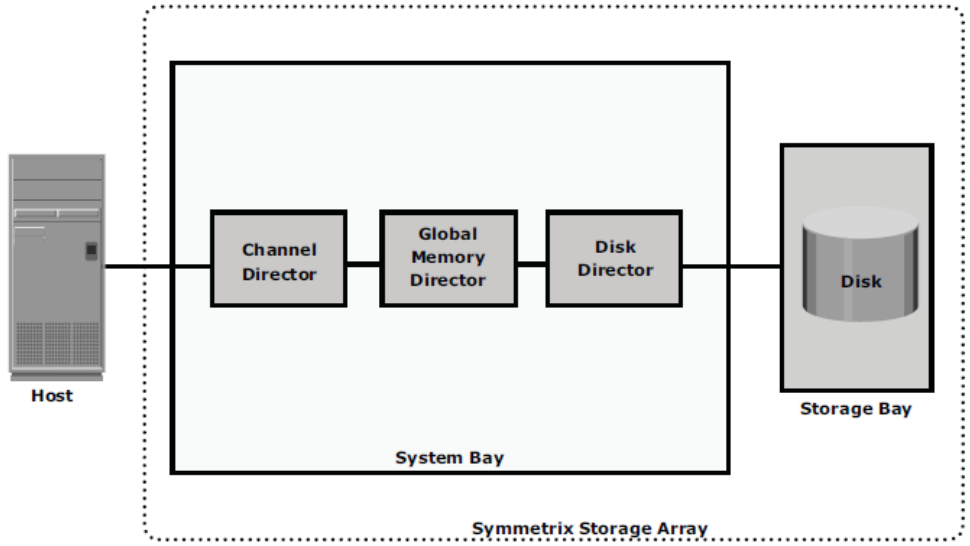

Figure 4-12: Basic building blocks of Symmetrix

Le système Symmetrix se compose des contrôleurs d'entrée (appelés les *Channel Directors*) pour la connectivité avec l'hôte, un très gros cache global (appelé le *Global Memory Director [GMD]*) et les contrôleurs aval (appelés *Disk Directors*) pour la connectivité avec les disques. La baie de stockage est un cabinet qui peut loger 240 disques durs. La Figure 4-13 montre les composants de la baie système et de la baie de stockage d'un Symmetrix.

La baie système se compose d'un compartiment contenant des fentes d'expansions permettant d'insérer des cartes contrôleurs. Les 24 fentes à l'avant contiennent le GMD, le disk director el le channel director. Les fentes arrière contiennent des adaptateurs de canaux, des adaptateurs de disques et des cartes de modules de contrôle d'environnement.

La baie système contient jusqu'à 12 channel directors et adaptateurs frontaux. Les channel directors permettent plusieurs options de connectivité pour Fibre Channel, ESCON, FICON, iSCSI ou GigE avec les hôtes ou le réseau.

Le système de Symmetrix DMX-4 comporte également jusqu'à huit cartes de GMD dans le compartiment de cartes. Les memory directors sont disponibles de 2 GB à 64 GB. Symmetrix emploie la mémoire dynamique synchrone (DDR SDRAM), une des dernières générations de la technologie de SDRAM.

Jusqu'à huit directeurs/adaptateurs de disques, en paires, fournissent connectivité de 2 GB/s en aval aux unités de disques Fibre Channel. Chaque directeur de disque peut supporter un maximum de 240 disques.

Deux modules de communication et de contrôle d'environnement, également appelés les *cross communication modules* (XCMs), sont fournis pour la configuration et autres communications. Les XCMs contiennent l'interface Ethernet entre les directeurs (canal, mémoire, et disque) et le processeur de service.

Le processeur de service contient un clavier, un écran, une souris, et un serveur relié

au sous-système de Symmetrix par l'interface Ethernet privée. Le processeur de service peut être configuré avec un modem externe pour communiquer avec le centre de support à la clientèle d'EMC. La baie de stockage peut comporter jusqu'à 240 unités de disques répartis dans plusieurs boitiers. Chaque boitier contenant un bloc d'alimentation redondant, des modules de refroidissement pour les unités de disques, deux LCCs et 4 à 15 disques durs Fibre Channel.

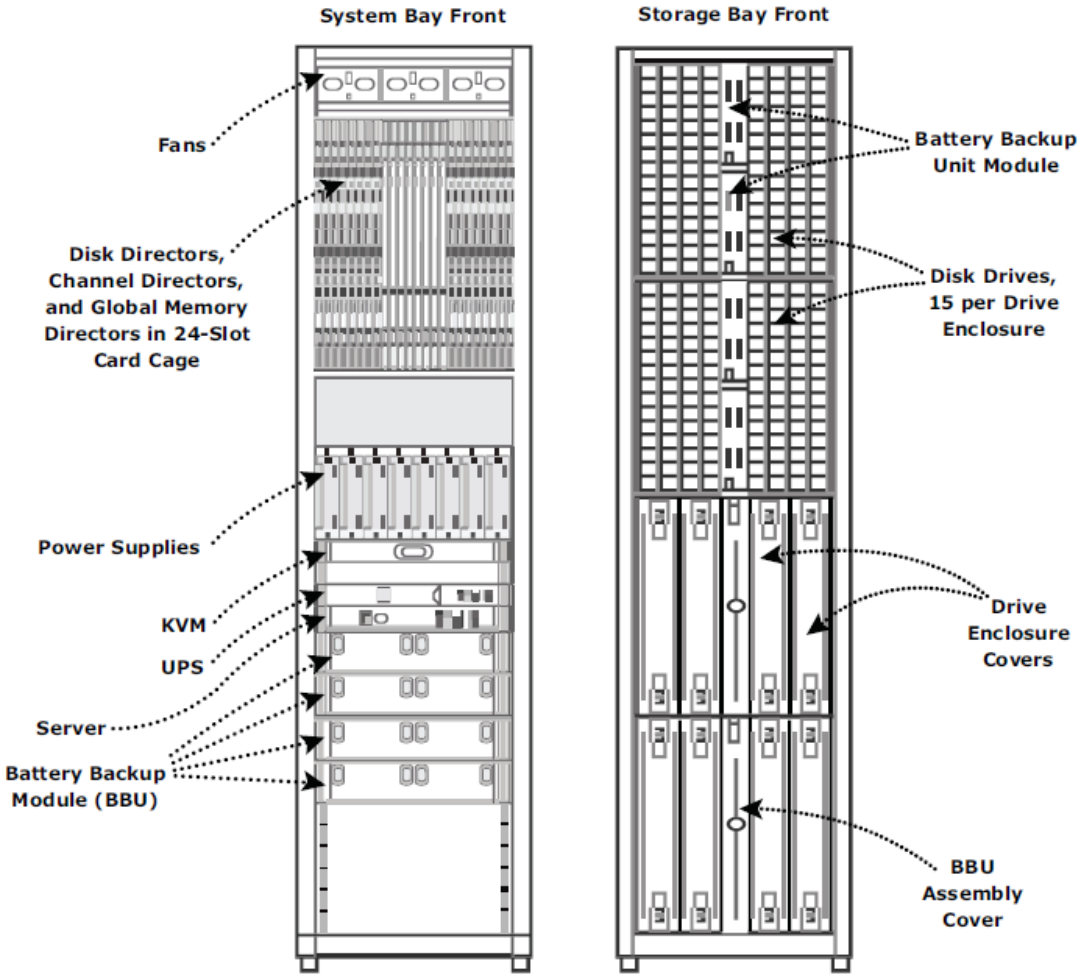

Figure 4-13: Symmetrix system bay and storage bay

4.3.6 "Direct Matrix Architecture"

Symmetrix emploie Direct Matrix Architecture composée de chemins dédiés pour le transfert de données entre la partie frontale, la mémoire globale, et la partie avale. Les composants clés de Symmetrix DMX sont:

- Partie frontale : L'hôte se relie au Symmetrix par l'intermédiaire d'un port d'entrée sur le channel director. Plusieurs directeurs sont configurés, chacun avec des ports multiples afin de fournir une bonne connectivité et de la redondance à

l'hôte. Chaque channel director de Symmetrix supporte huit liens internes vers la mémoire globale. Les transferts de données entre l'hôte et la mémoire globale peuvent s'exécuter simultanément à travers les multiples ports d'un directeur (cf. Figure 4-14).

- Partie avale: Les directeurs de disques aval contrôlent l'interface vers les disques et sont responsables du transfert des données entre les unités de disques et la mémoire globale. Chaque directeur de disque d'un système Symmetrix supporte 8 liens internes vers la mémoire globale.

- La mémoire globale: La mémoire globale est le composant le plus important du Symmetrix. Toutes les opérations de lecture et d'écriture sont exécutées via la mémoire globale. Les requêtes I/O de l'hôte sont reçues par la partie frontale et traitées par la mémoire globale à des vitesses beaucoup plus grandes que pour le transfert impliquant des disques (électronique vs mécanique). Les directeurs de mémoire globale travaillent par paires. L'écriture se fait d'abord sur un directeur global de mémoire et ensuite elle est faite sur le directeur global de mémoire secondaire qui agit comme miroir pour la protection des données. Chaque directeur global de mémoire a 16 ports avec des connexions séries bidirectionnelles (full-duplex) entre le directeur global de mémoire et les directeurs de canaux ou de disques (un total de 16 directeurs) via le *direct matrix*. Chacun des 8 ports directeurs sur les 16 directeurs se relie à un des 16 ports mémoires sur chacun des 8 directeurs globaux de mémoire comme indiqué à la Figure 4-14. Ces 128 différents raccordements points à point permettent jusqu'à 128 opérations de cache simultanées dans le système, fournissant ainsi une largeur de bande ultra-grande pour le traitement d' I/O.

- XCM : XCM est l'agent de communication entre le processeur de service et tous les nœuds de traitement (canal, disque, et directeur de mémoire) dans le système. Les liens avec le monde externe du processeur de service permettent la télésurveillance et les diagnostics à distance à travers des lignes téléphoniques. Le XCM permet également l'envoi de commandes à distance vers le director boards, les directeurs globaux de mémoire, et vers lui-même. Ces commandes peuvent être envoyées à partir du processeur de service ou à distance par le centre de support à la clientèle d'EMC.
- Symmetrix Enginuity : C'est l'environnement de fonctionnement pour le Symmetrix de EMC. Enginuity contrôle et assure le flux optimal des données ainsi que l'intégrité des informations à travers les divers composants matériels du système de Symmetrix. Il gère toutes les opérations de Symmetrix ainsi que les toutes les ressources du système pour optimiser les performances intelligemment. Enginuity assure la disponibilité du système par la surveillance et la détection d'erreurs. Il permet la correction d'erreurs et fournit des fonctions d'entretien. Il offre également une base pour les logiciels spécialisés dans le recouvrement de données en cas de désastre, la continuité des affaires et la gestion du stockage.

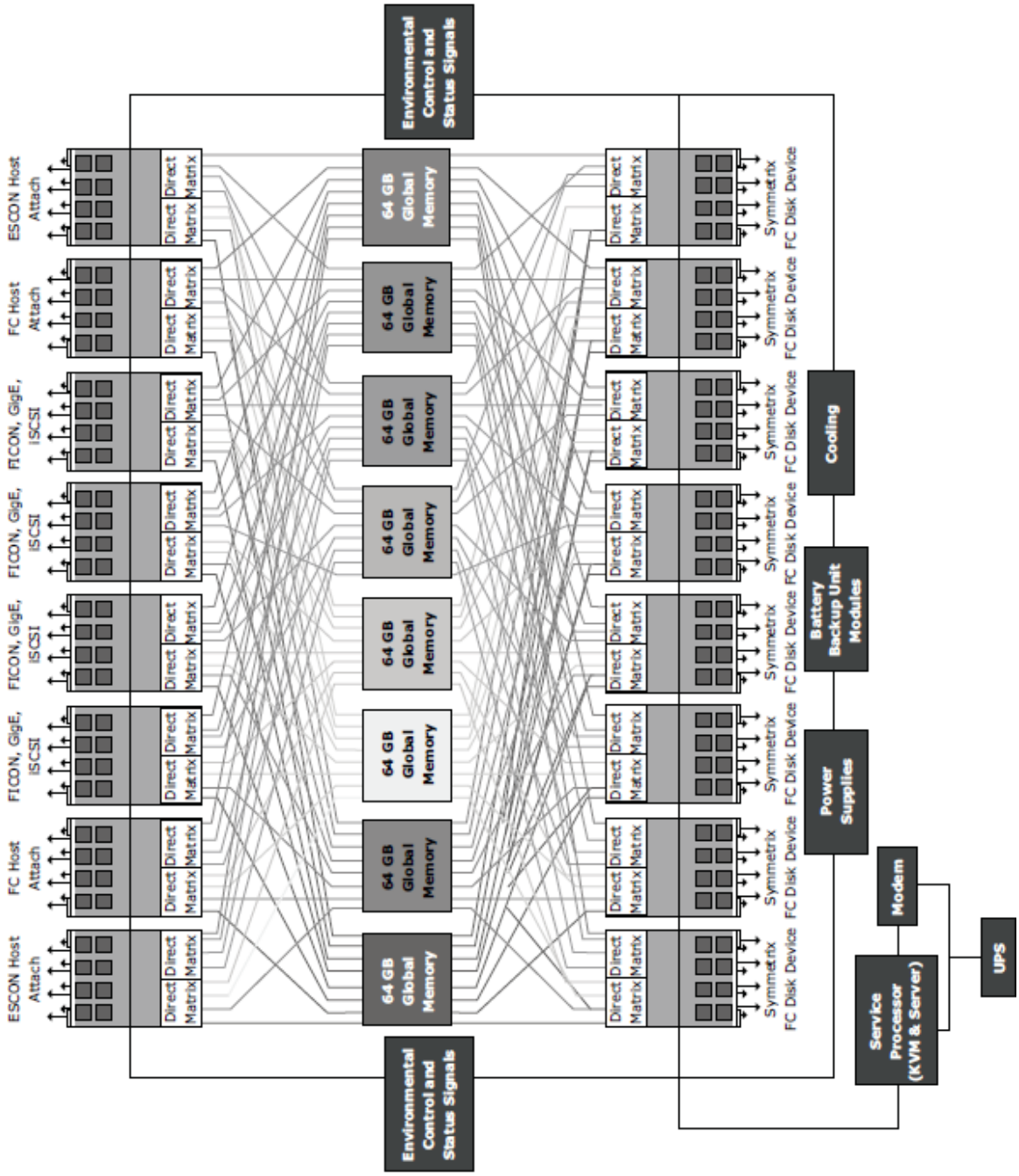

Figure 4-14: Direct matrix architecture

Sommaire

Ce chapitre a montré en détail les dispositifs et les composants d'un système de stockage intelligent - partie frontale, cache, partie arrière et disques physiques. L'implémentation active-active et active-passive des systèmes de stockage intelligents a été également décrite. Un système de stockage intelligent offre les bénéfices suivants à une organisation :

- Capacité accrue
- Performances améliorées
- Gestion de stockage simplifiée
- Disponibilité des données améliorée
- Évolutivité et flexibilité améliorée
- Meilleure continuité des affaires
- Plus grande sécurité et contrôle d'accès

Un système intelligent de stockage est maintenant une partie intégrale d'un centre de traitement de données corporatif. Bien qu'un système de stockage intelligent haut de gamme réponde aux attentes en terme de stockage de données, il lance un défi de taille aux administrateurs, celui de rendre le partage de l'information facile et sécuritaire à toutes les branches de l'entreprise.

La mise en réseau du stockage est une stratégie flexible centrée sur l'information qui prolonge la portée des systèmes de stockage intelligents à la grandeur de l'entreprise. Elle fournit une manière commune de gérer, partager et protéger l'information. Les réseaux de stockage sont détaillés dans la prochaine section.

EXERCISES

1. **Consider a scenario in which an I/O request from track 1 is followed by an I/O request from track 2 on a sector that is 180 degrees away from the first request. A third request is from a sector on track 3, which is adjacent to the sector on which the first request is made. Discuss the advantages and disadvantages of using the command queuing algorithm in this scenario.**
2. **Which type of application benefits the most by bypassing write cache? Why?**
3. **An Oracle database uses a block size of 4 KB for its I/O operation. The application that uses this database primarily performs a sequential read operation. Suggest and explain the appropriate values for the following cache parameters: cache page size, cache allocation (read versus write), pre-fetch type, and write aside cache.**
4. **Download Navisphere Simulator and the lab guide from `http://education.EMC.com/ismbook` and perform the tasks listed.**