Article technique

Comparaison de l'interface des disques SSD d'entreprise

Introduction

PCI Express (PCIe) est une interface bus générique qui est employée dans les applications de traitement client et d'entreprise. Les interfaces de stockage en masse (SATA, SAS) se connectent à l'ordinateur hôte par le biais d'adaptateurs hôtes, eux-mêmes connectés à l'interface PCIe.

L'interface SATA a été conçue comme une interface de disque dur ; l'interface SAS a, pour sa part, été conçue comme une interface de périphérique et une interface/infrastructure d'un sous-système de stockage. Pour faire face à l'évolution des disques durs et des configurations système requises, qui font désormais appel à des interfaces plus rapides et à de nouvelles fonctionnalités, les interfaces SATA et SAS ont fait l'objet de différentes révisions.

Les disques SSD (Solid State Drive) ont rapidement ajouté de nouvelles exigences de performances à ces interfaces : les taux de transfert de ces disques sont ainsi passés de dizaines de Mo/s à des centaines de Mo/s pour atteindre aujourd'hui des milliers de Mo/s. Parallèlement à l'augmentation de ces taux de transfert, l'absence de mouvement mécanique au sein des disques SSD a également entraîné la hausse du nombre d'opérations d'entrée/sortie par seconde assuré par ce type de périphérique de stockage.

Un tel développement a nécessité de meilleures mises en œuvre des normes en vigueur et l'amélioration des normes d'interface actuelles pour répondre aux nouvelles exigences de performance tout en conservant la compatibilité avec l'architecture système existante.

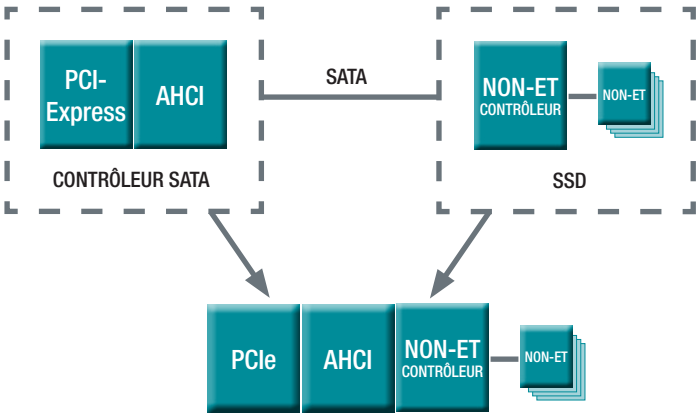
Ce document présente les différentes interfaces et détaille les différents compromis en termes de performances et de compatibilité.

Comparaison de l'interface des disques SSD d'entreprise

Interface SATA

L'interface SATA est une interface peu coûteuse conçue pour les liaisons point à point par câble ou circuit imprimé. La connexion hôte s'effectue au niveau d'une interface avancée du contrôleur hôte (AHCI), qui réside généralement sur la puce hôte en tant qu'adaptateur hôte sur le bus PCIe. Certaines contraintes de conception de cette interface peuvent causer une perte de temps au niveau du bus de 1 µs (ou plus) pour chaque commande. Il ne s'agit pas là d'un problème majeur pour les disques durs puisque le transfert de 4 Ko y est de l'ordre de 10 µs. En revanche, le problème se pose pour les disques SSD pour lesquels la vitesse de transfert de ce même volume de données est de 2 µs (voire inférieure) : le temps perdu devient alors important et l'interface SATA se révèle moins intéressante en tant qu'interface de stockage de masse hautes performances.

L'interface SATA reste malgré tout intéressante en tant qu'interface SSD peu onéreuse lorsque le principal facteur de décision est le coût, et non les performances. L'architecture SATA peut également être consolidée au sein d'un adaptateur hôte qui gère le jeu de commandes SATA sans inclure réellement l'interface SATA physique (Figure 1).



Source : Seagate Technology, 2011
Figure 1. Consolidation de l'architecture

Interface SAS

L'interface SAS est également une interface série, reliée à l'hôte via un adaptateur hôte, mais plusieurs différences majeures la rendent intéressante en tant qu'interface SSD :

- Perte de temps matérielle moindre
- Vitesse élevée de transfert des données
- Ports Wide
- Interfaces pilote-contrôleur efficaces

L'interface SAS propose des fonctions que n'offre pas l'interface SATA et qui améliorent la fiabilité et la disponibilité des périphériques qui lui sont connectés :

- Protocole série puissant
- Prise en charge de plusieurs hôtes
- Intégrité des données de bout en bout
- Fonction double port
- Hauts niveaux de simultanéité et d'agrégation

Perte de temps matérielle moindre

Il n'existe aucune interface hôte universelle pour SAS comparable au contrôleur AHCI SATA. Au lieu de cela, plusieurs fournisseurs s'affrontent sur le marché des adaptateurs hôtes SAS où les performances sont un facteur majeur, non seulement pour servir d'interface à plusieurs disques durs, mais également à différents systèmes RAID dans lesquels les débits de plusieurs broches HDD sont agrégés pour augmenter la vitesse de transfert. Les adaptateurs hôtes SAS sont, par ailleurs, conçus pour gérer des disques durs et des disques SSD hautes performances comme des disques 15 000 tr/min après un *short-stroking* (réduction de la capacité pour des performances accrues). Comme l'adaptateur hôte matériel et le pilote du périphérique qui le gère sont conçus en tant que système, de nouvelles conceptions optimisées pour les disques SSD commencent à voir le jour et contribuent à améliorer les vitesses de transfert ainsi que le nombre d'opérations d'entrée/sortie par seconde.

Vitesses de transfert plus élevées

Les ports SAS prennent actuellement en charge des vitesses de transfert pouvant aller jusqu'à 6 Gbits/s. Des sociétés comme LSI et PMC-Sierra échantillonnent les conceptions actuellement en développement pour prendre en charge plus de 2 millions d'opérations d'E/S par seconde et des vitesses de transfert de 12 Gbits/s, voire de 24 Gbits/s à l'avenir.

Ports Wide

L'architecture SAS s'appuie sur un concept qui lui est propre : les ports Wide. Grâce à eux, plusieurs liaisons peuvent être agrégées pour autoriser plusieurs chemins simultanés entre un ou plusieurs hôtes et un périphérique. Le connecteur actuel du disque SAS définit deux ports pour le disque. Pour des raisons de choix de conception, les disques durs actuels ne prennent pas en charge les ports Wide, mais uniquement la fonction double port dans laquelle chaque port possède une adresse SAS différente qui empêche de les configurer en tant que port Wide.

Les propositions admises pour *SAS-3* (12 Gbits/s) permettent de passer le nombre de ports du disque à 4, ces derniers pouvant tous se connecter au même domaine, ou à différents domaines par paire. Un nombre très limité de disques SSD peut prendre en charge les ports Wide en plus de la fonction Double port sur un périphérique à deux ports.

Comparaison de l'interface des disques SSD d'entreprise

Protocole série puissant

Le protocole série SAS permet de former les transmetteurs et récepteurs série. Il en résulte une amélioration de la qualité du signal sur le câble ou le support par compensation de la longueur du canal, du défaut d'adaptation d'impédance et du brouillage entre symboles. Le protocole série SAS gère également la détection et la retransmission des erreurs au niveau matériel. Une reprise plus rapide est ainsi possible en cas de problèmes de signalisation intermittente.

Prise en charge de plusieurs hôtes

L'interface SAS et la matrice de commutation permettent à plusieurs hôtes d'accéder au même périphérique. Cette fonction peut servir à gérer les défaillances des hôtes et celles des chemins de données pour assurer une meilleure disponibilité des données.

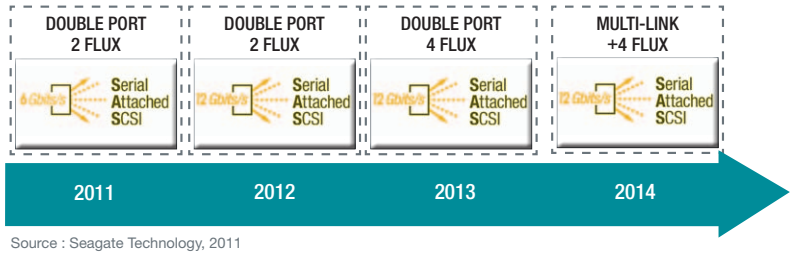
Intégrité des données de bout en bout

L'interface SAS peut vérifier l'intégrité des données au moyen de contrôles de redondance cyclique (CRC) des données, de leur création dans le tampon de données hôte à leur transfert via les interfaces PCIe et SAS, jusqu'à leur stockage sur un périphérique, puis leur lecture et de nouveau leur transfert dans le tampon de données hôte. Plusieurs points de contrôle sont ainsi possibles tout au long du chemin, des applications aux contrôleurs RAID et jusqu'aux périphériques. Cette fonction est parfois appelée Informations de protection.

Fonction double port

Les périphériques cibles SAS prennent en charge le fonctionnement en mode double port. Il est ainsi possible de créer deux domaines de défaillance et d'accroître la disponibilité. Même en cas de panne sur l'un des chemins menant à un port et bloquant l'accès sur ce chemin, un périphérique reste accessible via le second port.

Historiquement, Seagate a favorisé l'adoption de cette interface sur le marché. Seagate collabore avec la SCSI Trade Association (STA) et d'autres leaders du secteur pour exploiter l'infrastructure SAS existante et largement déployée (Figure 2). Le tableau 1 indique comment une interface SAS 12 Gbits/s et un système multiliaison profitent aux fabricants de systèmes et aux organisations d'utilisateurs.



Source : Seagate Technology, 2011

Figure 2. Évolution de l'interface SAS

Interface PCI-Express

L'interface PCI-Express (PCIe) est l'interface fondamentale qui connecte les périphériques au processeur hôte et, via un contrôleur mémoire, à l'architecture mémoire du système. Les interfaces SATA et SAS présentées ci-dessus peuvent se connecter via une interface PCIe (ou un adaptateur hôte) à la mémoire et au processeur hôte.

Tableau 1. Avantages des disques SAS et suggestions de disques SAS avec améliorations multiliaisons

Liens multiples (LB)	X4 (4 x 600 Mo/s)
Puissance disponible	25 W (2,5 pouces)
Latence totale	très faible
Protocole multihôte	**Oui**
Haute disponibilité	**Oui (Double port)**
Évolutivité	**Excellente**
Pile de protocoles robuste et éprouvée	**Oui**
Remplacement sous tension	**Oui**
Compatible avec les logiciels de gestion existants	**Oui**

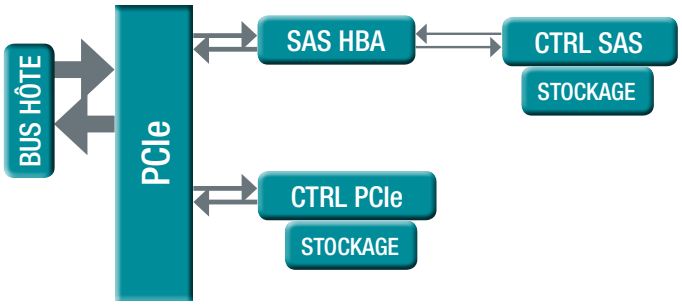
Source : Seagate Technology, 2011

L'interface PCIe est une implémentation série de l'interface PCI qui fournissait une adresse parallèle/connexion de données entre les périphériques et le processeur/la mémoire hôte. L'interface PCIe communique sur une ou plusieurs *voies*, chacune étant dotée d'une interface série de transmission et d'une de réception. Jusqu'à 32 voies peuvent être utilisées pour connecter un hôte à un périphérique. Sur chaque voie, la vitesse de transfert série dépend de la version de la norme PCIe implémentée. La version actuelle est la version 3.0 et la vitesse de transfert avoisine 1 Go/s.

Pour un serveur 1U, l'interface PCIe est conçue pour utiliser un seul connecteur sur une carte mère (client), ou un adaptateur coudé à deux connecteurs sur une carte mère (serveur). Un système de câblage est également disponible (bien que rarement utilisé). Un serveur 2U, 4U ou 7U offre beaucoup plus d'emplacements PCIe, comme les implémentations clientes. La spécification PCIe utilise également une formation transmetteur (et récepteur) pour assurer l'adaptation aux variations d'impédance d'une configuration, mais elle est conçue pour des canaux de transmission plus courts qu'une interface SAS.

Les commutateurs PCIe peuvent prendre en charge la virtualisation E/S monoracine (SR-IOV) et multiracine (MR-IOV). Ces méthodes permettent d'améliorer les performances du contrôleur dans des systèmes virtuels (hyperviseurs) mono ou multi-hôtes. La première méthode (SR-IOV) commence à se généraliser dans les adaptateurs. Or, VMware n'en tire probablement pas encore avantage. En général, la seconde méthode (MR-IOV) n'est pas prise en charge sur les adaptateurs.

Comparaison de l'interface des disques SSD d'entreprise



Source : Seagate Technology, 2011
Figure 3. Évolution de l'interface SAS

Les périphériques de stockage qui se connectent au moyen de l'interface PCIe s'appuient pour ce faire sur un accès registre direct ou sur un adaptateur hôte qui se connecte ensuite au périphérique via un câblage supplémentaire ou une interface de type fond de panier.

Actuellement, il existe plusieurs sortes d'implémentations de ces deux architectures. L'interface SATA utilise une implémentation avec adaptateur de bus hôte dans la puce système *(contrôleur d'extension)* — AHCI Intel ou AMD — qui requiert différents pilotes AHCI mais permet un mappage à des implémentations IDE existantes compatibles. Ces interfaces mettent également en œuvre différentes fonctions de gestion RAID.

Pour l'interface SAS, plusieurs fournisseurs d'adaptateurs HBA (Host Bus Adapter) ayant d'autres expandeurs et contrôleurs RAID à leur disposition, utilisent des pilotes de périphériques propriétaires et le BIOS pour répondre aux différents besoins en termes de performances et de facilité de configuration.

L'interface du pilote-contrôleur PCIe est mise en œuvre dans les spécifications NVM Express et SCSI sur PCIe (SOP).

L'architecture consolidée SATA décrite précédemment constitue un autre exemple d'accès registre direct PCIe.

Disques SSD PCIe actuels

Actuellement, sur le marché, on trouve deux grands types de disques SSD PCIe : natif et agrégateur. Le contrôleur natif se connecte au bus PCIe hôte et contrôle directement plusieurs bus mémoire flash. Ces derniers utilisent généralement une interface logicielle propre au fabricant et réservée à un périphérique donné. Certaines de ces mises en œuvre basculent la traduction des adresses et d'autres fonctions sur le processeur et la mémoire hôtes. Ce transfert cause à son tour une baisse des ressources système allouées aux applications lorsque les périphériques sont employés dans le cadre de fortes charges de travail. Par ailleurs, du fait de leur récente apparition sur le marché, ces combinaisons uniques alliant matériel et disques ont tendance à présenter des instabilités dans la mesure où leur écosystème est encore en cours d'évolution.

Le modèle agrégateur implique une autre approche de conception qui s'appuie sur un contrôleur RAID SAS ou SATA, auquel sont connectés plusieurs disques SSD SAS ou SATA. Tous sont regroupés sur une même carte PCIe. Le contrôleur RAID consolide les performances de plusieurs périphériques pour garantir de meilleurs niveaux de performance. Basées sur des interfaces logicielles et matérielles d'entreprise éprouvées, ces conceptions sont très stables et abouties. Par ailleurs, elles utilisent des contrôleurs intelligents qui effectuent des traductions d'adresses et d'autres fonctions, ce qui assure une utilisation optimale des cycles du processeur et de la mémoire système par les applications, même en cas de fortes charges d'E/S.

L'avenir des disques SSD PCIe

Les approches SOP et NVMe sont toutes deux similaires sur le plan architectural. Quelques différences les distinguent toutefois : l'approche NVMe est développée dans un groupe de travail de l'industrie tandis que l'approche SOP est développée dans un forum de normes ouvertes reconnu. L'approche NVMe n'est destinée qu'aux périphériques à mémoire rémanente, tandis que l'approche SOP est également destinée aux adaptateurs de bus hôtes et aux contrôleurs RAID avec des fonctions permettant de basculer entre différents périphériques SOP. Par ailleurs, l'approche SOP s'appuie largement sur les architectures et fonctions sectorielles existantes, tandis que l'approche NVMe utilise un nouveau jeu d'instructions très limité et une interface de mise en file d'attente.

Avantages et inconvénients des interfaces

Chacune des architectures de stockage décrites ici présente des avantages et des inconvénients. En fonction de la conception globale du système, les avantages liés à l'utilisation d'une architecture spécifique peuvent l'emporter sur les problèmes qui lui sont liés. Une analyse approfondie est donc nécessaire pour prendre la décision la plus appropriée. Cette décision doit également tenir compte de la compatibilité avec une conception système existante.

Par exemple, la mise à jour du système d'un ordinateur portable doté d'un disque dur SATA 2,5 pouces avec un disque SSD ne fonctionne que si ce disque SSD présente les mêmes dimensions et la même interface SATA (ou une plus récente). Dans ce cas, la vitesse du disque SSD est limitée. Une vitesse supérieure à celle de l'interface SATA hôte existante n'accroît pas les performances du système.

Dans une situation similaire, un serveur d'entreprise utilisant un disque dur SAS 15 000 tr/min, dont la capacité a été réduite en vue d'augmenter les performances, pour stocker l'index de base de données, peut être mis à niveau avec un disque SSD SAS, qui augmentera les performances système globales, mais uniquement dans la mesure où certains facteurs système génèrent de nouveaux engorgements (processeur, mémoire, réseau, adaptateurs, etc.).

Dans la nouvelle architecture système, l'ajout d'un disque SSD peut accroître considérablement les performances système, mais uniquement si le reste

Comparaison de l'interface des disques SSD d'entreprise

Tableau 2. Comparaison des disques SSD PCIe de type natif et agrégateur		
	Natif	**Agrégateur**
Commandes/Transport	Propriétaire (FTL[1] dans la mémoire hôte/principale) 	SCSI ou SATA (Plusieurs disques SSD, Contrôleur intégré) 
Comité	Aucun	Aucun
Basé sur des normes	Non	**Oui**
Performances avec Flash	Élevées	Élevées
Temps processeur perdu	Élevé	**Faible**
Latence avec courte file d'attente	Très faible	Faible
Latence avec longue file d'attente	Modérée	**Faible**
Extensibilité des cas d'utilisation	Non	**Oui** (RAID, HBA, etc.)
Maturité	En évolution	**Basée sur des architectures sectorielles éprouvées**
Fonctions d'entreprise (IP, Sécurité, Gestion, etc.)	Non	**Dépend de la mise en œuvre**

[1] FTL : Flash Translation Layer (Couche de traduction Flash)
Source : Seagate Technology, 2011

Tableau 3. Comparaison des disques SOP et SSD PCIe NVMe		
	SOP[1]	**NVMe[2]**
Commandes/Transport	SOP/PQI[3] (FTL dans le contrôleur) 	NVMe/NVMe (FTL dans le contrôleur) 
Comité	**T10/INCITS[4]**	Groupe de travail de l'industrie
Basé sur des normes	**Oui (ANSI/ISO)**	Non
Performances avec Flash	Excellentes	Excellentes
Charge du processeur	Faible	Faible
Latence avec courte file d'attente	Très faible	Très faible
Latence avec longue file d'attente	Faible	Faible
Extensibilité des cas d'utilisation	**Oui** (RAID, HBA, etc.)	Non (NVM uniquement)
Maturité	**Basée sur des architectures sectorielles éprouvées**	À définir
Fonctions d'entreprise (IP, Sécurité, Gestion, etc.)	**Prise en charge complète**	Limitée

[1] SOP : SCSI over PCI Express
[2] NVMe : Nonvolatile Memory Express
[3] Interface de mise en file d'attente PCIe
[4] INCITS : International Committee for Information Technology Standards
Source : Seagate Technology, 2011

www.seagate.com

de l'architecture peut prendre en charge une vitesse de transfert et une bande passante supérieures. La hausse des vitesses de transfert des disques SSD implique un apport de puissance supérieur au niveau du périphérique et nécessite une meilleure dissipation de la chaleur lorsque le disque SSD est monté.

Un autre facteur à prendre en compte est le créneau de disponibilité des pilotes des périphériques du système d'exploitation et la prise en charge du BIOS de ces nouvelles interfaces SSD, ainsi que la fiabilité initiale des logiciels.

Interfaces et faits de latence des disques SSD (Flash)

De nombreuses idées erronées circulent à propos de la nature des facteurs responsables de latence, et de l'impact de ces facteurs sur les performances des applications. Il est donc important de se concentrer sur l'ensemble du problème et non sur une partie uniquement.

Les principaux générateurs de latence dans les disques SSD sont en réalité les composants flash eux-mêmes. Les temps d'accès SLC et MLC sont supérieurs à 25 µs et 50 µs, respectivement, en l'absence de contention d'accès pour les deux. Lorsque la longueur des files d'attente augmente, la contention de l'accès aux composants Flash peut renforcer considérablement la latence.

Une fois que l'un de ces composants débute son accès, les autres demandes d'accès à ce composant doivent patienter. Jusqu'à huit puces flash partagent un même bus : chacune d'elles doit donc attendre son tour pour pouvoir se servir du bus. Les activités de gestion interne renforcent la latence (traduction d'adresse, nettoyage de mémoire, gestion de l'usure, etc.)

Il convient également de tenir compte du système d'exploitation qui ajoute de la latence indépendamment du protocole d'accès et de l'interconnexion. Sont ainsi concernés le système de fichiers, le gestionnaire de volume, les pilotes de classe et les temps perdus en changement de contexte.

Des différences en termes de protocole et d'interconnexion n'ont que de faibles effets sur la latence perçue par une application (exprimée en fraction de microseconde).

AMÉRIQUES Seagate Technology LLC 10200 South De Anza Boulevard, Cupertino, California 95014, États-Unis, +1 408 658 1000
ASIE/PACIFIQUE Seagate Singapore International Headquarters Pte. Ltd. 7000 Ang Mo Kio Avenue 5, Singapour 569877, +65 6485 3888
EUROPE, MOYEN-ORIENT ET AFRIQUE Seagate Technology SAS 16-18 rue du Dôme, 92100 Boulogne-Billancourt, France, +33 (0)1 41 86 10 00