Artigo sobre tecnologia

Comparação das interfaces de SSD empresariais

Introdução

PCI Express (PCIe) é uma interface de barramento de propósito geral para uso em aplicações computacionais empresariais e clientes. As interfaces de armazenamento em massa existentes (SATA, SAS) são conectadas ao computador host por meio de adaptadores de host, que, por sua vez, são conectados à interface PCIe.

A interface SATA foi projetada como interface de disco rígido (HDD) e a interface SAS foi projetada tanto como interface de dispositivo quanto como interface/infraestrutura de subsistema de armazenamento. Conforme os HDDs e os requisitos de sistema evoluíram, exigindo interfaces mais rápidas e novos recursos, as interfaces SATA e SAS sofreram diversas revisões.

As unidades de estado sólido (SSDs) rapidamente trouxeram novos requisitos de desempenho significativos a essas interfaces, visto que as taxas de dados das SSDs passaram de dezenas de MB/s para centenas e, agora, milhares de MB/s. Além desse aumento nas taxas de dados, a falta de movimento mecânico nas SSDs também elevou o número de operações de entrada e saída por segundo (IOPS) que esses dispositivos de armazenamento podem realizar.

Esse fato criou uma necessidade de implementações aprimoradas dos padrões existentes, bem como melhorias dos padrões de interface existentes, para lidar com os novos requisitos de desempenho e manter a compatibilidade com a arquitetura de sistema atual.

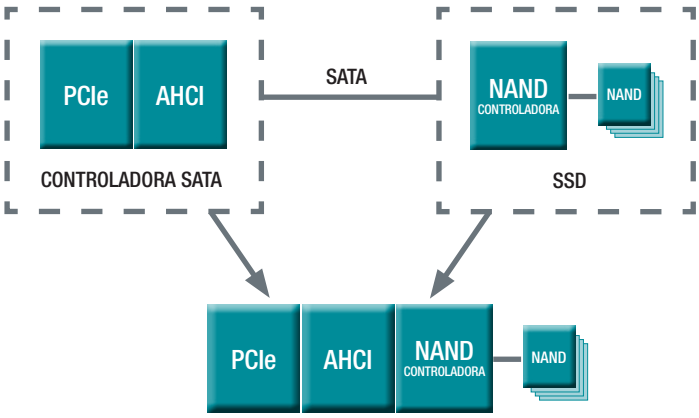
Este artigo discute as diferentes interfaces e contrasta as várias compensações de compatibilidade e desempenho encontradas.

Comparação das interfaces de SSD empresariais

Interface SATA

SATA é uma interface de baixo custo projetada para conexão ponto a ponto, seja por cabo ou por placa de circuito impresso (PCB). A conexão host é feita com uma interface avançada de controladora de host (AHCI), geralmente localizada no chipset do host como um adaptador de host no barramento PCIe. Há alguns problemas de projeto com essa interface, que podem criar uma sobrecarga de barramento de 1 µs (ou mais) por comando. Isso não é grave para HDDs, nos quais uma transferência de 4 KB fica na ordem de 10 µs, mas SSDs podem transferir 4 KB de dados em 2 µs (ou menos). Assim, a sobrecarga se torna significativa e a interface SATA fica menos interessante como interface de armazenamento em massa de alto desempenho.

A SATA ainda é adequada como interface de SSD de baixo custo, quando o preço, não o desempenho, é um fator de decisão importante. A arquitetura SATA também pode ser consolidada em um adaptador de host que gerencie o conjunto de comandos SATA sem realmente incluir a interface SATA física (PHY) (Figura 1).



Fonte: Seagate Technology, 2011
Figura 1. Consolidação de arquitetura

Interface SAS

SAS também é uma interface serial, conectada ao host por meio de um adaptador de host, mas com diferenças significativas que fazem com que ela seja adequada como uma interface de SSD:

- Menos sobrecarga de hardware
- Taxas de transferência mais rápidas
- Portas amplas
- Interfaces de controladora de driver eficientes

Além disso, a SAS inclui recursos não encontrados na SATA, que aumentam a confiabilidade e a disponibilidade dos dispositivos conectados à interface:

- Protocolo serial robusto
- Aceitação de vários hosts
- Integridade de dados ponta a ponta
- Recurso de porta dupla
- Altos graus de simultaneidade e agregação

Menos sobrecarga de hardware

Não há uma interface de host universal para SAS que seja equivalente à controladora AHCI SATA. Contudo, vários fornecedores competem no mercado de adaptadores de host SAS, no qual o desempenho é um fator importante não só para HDDs independentes de interface mas também para diversos sistemas RAID, cujas taxas de transferência de várias rotações de HDD são agregadas para garantir mais velocidade de transferência. Além disso, os adaptadores de host SAS são projetados para gerenciar SSDs e HDDs de desempenho mais alto, como os discos *short-stroked* (em que somente uma parcela da unidade é usada) de 15.000 RPM. Como o adaptador de host de hardware e o driver do dispositivo que gerencia esse adaptador de host são projetados como um sistema, novos designs otimizados para SSDs estão começando a ser lançados, com melhores taxas de transferência e IOPS.

Taxas de transferência mais rápidas

As portas da SAS sustentam atualmente taxas de dados de até 6 Gb/s. Empresas como a LSI e PMC-Sierra estão testando designs em desenvolvimento para sustentar taxas de dados de 12 Gb/s e IOPS superior a 2 milhões, com a possibilidade de chegar a 24 Gb/s no futuro.

Portas amplas

O conceito de portas amplas é inerente à arquitetura SAS, em que vários links podem ser agregados para possibilitar diversos caminhos simultâneos entre um ou mais hosts e um dispositivo. O conector atual da unidade SAS define duas portas para a unidade. Por uma opção de design, os HDDs atuais não aceitam porta ampla, somente porta dupla, e cada porta tem um endereço SAS diferente que impede a configuração como uma porta ampla.

As propostas aceitas para *SAS-3* (12 Gb/s) aumentam o número de portas no disco para quatro, todas podendo ser conectadas ao mesmo domínio ou em pares a domínios diferentes. Uma quantidade muito limitada de SSDs aceita porta ampla além da porta dupla em um dispositivo de duas portas.

Comparação das interfaces de SSD empresariais

Protocolo serial robusto

O protocolo serial SAS promove o treinamento de transmissores e receptores seriais. Isso melhora a qualidade do sinal no cabo ou no backplane, compensando o comprimento do canal, a disparidade de impedância e a interferência entre símbolos. O protocolo serial SAS também gerencia a detecção de erros e a retransmissão no hardware. Isso possibilita a recuperação mais rápida durante a ocorrência de problemas de sinalização intermitente.

Compatibilidade com vários hosts

A interface SAS e a matriz de comutação (switching fabric) possibilitam que vários hosts acessem o mesmo dispositivo. Esse recurso pode ser usado para gerenciar falhas de host, bem como falhas no caminho de dados para aprimorar a disponibilidade de dados.

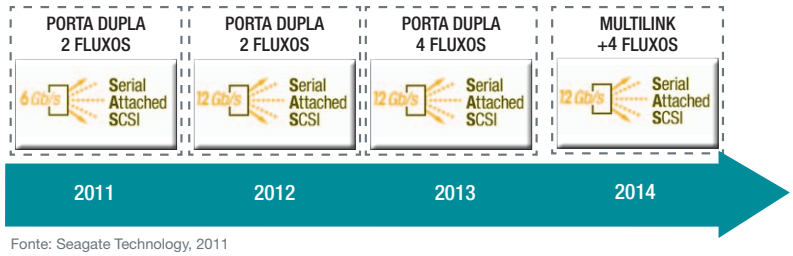
Integridade de dados ponta a ponta

A interface SAS pode confirmar a integridade de dados por meio de verificações de redundância cíclicas (CRC) dos dados do momento em que eles são criados no buffer de dados do host, durante a transferência pela interface PCIe e interface SAS até serem armazenados no dispositivo, e novamente quando são lidos e transferidos para o buffer de dados do host. Assim, há vários pontos de verificação ao longo do caminho, de aplicativos a controladoras RAID e nos dispositivos. Esse recurso às vezes é chamado de informações de proteção (PI).

Recurso de porta dupla

Dispositivos alvo de SAS aceitam operação de porta dupla. Com isso, é possível criar dois domínios de falhas e aumentar a disponibilidade. Mesmo se uma falha ocorrer em um dos caminhos a uma porta, impedindo o acesso ao longo desse caminho, o dispositivo continuará acessível pela segunda porta.

Historicamente, a Seagate determinou a adoção da interface no mercado. A Seagate está trabalhando com a STA (SCSI Trade Association, Associação de Comércio SCSI) e outros líderes do setor para aproveitar a infraestrutura SAS já amplamente implantada (Figura 2). A Tabela 1 mostra como a SAS de 12 Gb/s e o multilink beneficiam os desenvolvedores de sistemas e as organizações de usuários finais.



Fonte: Seagate Technology, 2011

Figura 2. A evolução da interface SAS

Interface PCI Express

PCI Express (PCIe) é a interface fundamental que conecta dispositivos periféricos ao processador host e de uma controladora de memória à arquitetura de memória do sistema. As interfaces SATA e SAS discutidas anteriormente conectam-se por meio de uma interface PCIe (ou adaptador de host) à memória e ao processador de host.

Tabela 1. Vantagens da SAS e da SAS com aprimoramentos multilink propostos

Vários links (LB)	X4 (4 x 600 MB/s)
Potência disponível	25 W (2,5 polegadas)
Latência total	Muito baixa
Protocolo de vários hosts	**Sim**
Alta disponibilidade	**Sim (porta dupla)**
Escalabilidade	**Excelente**
Pilha de protocolo robusta comprovada	**Sim**
Hot swap	**Sim**
Compatível com SW de gerenciamento existente	**Sim**

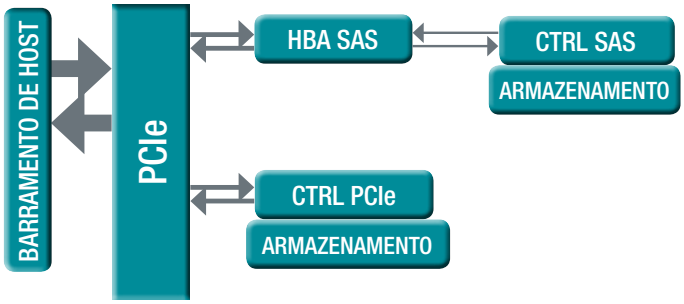
Fonte: Seagate Technology, 2011

A interface PCIe é uma implementação serial da interface PCI original, que fornecia um endereço paralelo/conexão de dados entre periféricos e a memória/processador de host. A interface PCIe se comunica por meio de uma ou mais *pistas (lanes)* que consistem em uma interface serial de transmissão e uma de recepção para cada pista. Até 32 pistas podem ser usadas para conectar um host a um dispositivo. A taxa de dados seriais de cada pista depende da versão do padrão PCIe implementado. A versão atual é a 3.0 e a taxa de dados é de aproximadamente 1 GB/s.

Para um servidor 1U, a interface PCIe é projetada para utilizar um único conector em uma placa-mãe (cliente) ou um adaptador de ângulo reto de dois conectores em uma placa-mãe (servidor). Um sistema de cabeamento também está disponível (porém é raramente usado). Um servidor 2U, 4U ou 7U tem muito mais slots PCIe, semelhante às implementações clientes. A especificação PCIe também usa treinamento de transmissor (e receptor) para adaptar-se às variações de impedância de uma configuração, mas é voltada a canais de transmissão de comprimento mais curto do que a SAS.

Switches PCIe podem acomodar virtualização de E/S de raiz única (SR-IOV) e virtualização de E/S de várias raízes (MR-IOV), que são métodos usados para aprimorar o desempenho da controladora em sistemas virtuais (hypervisor) com um ou vários hosts. A SR-IOV está começando a ser disponibilizada mais amplamente agora em adaptadores. Entretanto, a VMware ainda não pode tirar vantagem dela. Geralmente, a MR-IOV não é compatível com adaptadores.

Comparação das interfaces de SSD empresariais



Fonte: Seagate Technology, 2011
Figura 3. A evolução da interface SAS

Os dispositivos de armazenamento conectados pela interface PCIe o fazem por meio de uma conexão de registro direta ou de um adaptador de host, que depois é conectado ao dispositivo por meio de um cabeamento adicional ou uma interface do tipo backplane.

Atualmente, há uma série de implementações das duas arquiteturas. A SATA usa uma implementação de adaptador de barramento de host no chipset do sistema *(southbridge)*, HCI Intel ou AMD, que requer drivers AHCI diferentes, mas mapeamento a implementações IDE legadas compatíveis. Essas interfaces também implementam diversos recursos de gerenciamento RAID.

A SAS tem vários fornecedores de HBAs, com expansores e controladoras RAID adicionais disponíveis, todos usando BIOS e drivers de dispositivos patenteados para satisfazer as variadas necessidades de desempenho e configurabilidade.

A interface driver-controladora PCIe é implementada na especificação NVM Express e na especificação SCSI sobre PCIe (SOP) proposta.

A arquitetura consolidada da SATA descrita acima é outro exemplo da conexão de registro direto PCIe.

SSDs PCIe hoje

Há dois tipos principais de SSDs PCIe atualmente no mercado: a nativa e a agregadora. A controladora nativa é ligada ao barramento PCIe de host e controla diretamente vários barramentos de memória flash. Geralmente, é usada uma interface de software patenteada do fabricante que é específica ao dispositivo. Algumas dessas implementações impõem uma carga extra para a memória e CPU do host, exigindo tradução de endereço e outras funções. Isso, por sua vez, causa uma redução nos recursos do sistema disponíveis para aplicativos, quando os dispositivos são usados com cargas de trabalho pesadas. Além disso, por serem relativamente novas no mercado, essas unidades e combinações de hardware exclusivas tendem, às vezes, a apresentar instabilidades, visto que seus ecossistemas ainda estão em desenvolvimento.

O modelo agregador tem uma abordagem diferente quanto ao design. Essa abordagem utiliza uma controladora RAID SAS ou SATA existente à qual são conectadas várias SSDs SAS ou SATA. Elas são reunidas em uma única placa PCIe. A controladora RAID agrega o desempenho de vários dispositivos para oferecer altos níveis de desempenho. Baseados em interfaces de software e hardware de classe empresarial comprovadas, esses designs são muito estáveis e maduros. Além disso, esses designs usam controladoras inteligentes, que executam traduções de endereço e outras funções, possibilitando uso total dos ciclos de CPU e da memória do sistema pelos aplicativos, mesmo com cargas de trabalho de E/S intensas.

O futuro das SSDs PCIe

As duas abordagens, SOP e NVMe, têm arquiteturas semelhantes. Entretanto, a NVMe está sendo desenvolvida em um grupo de trabalho do setor, enquanto a SOP está sendo desenvolvida em um reconhecido fórum de padrões abertos. A NVMe é voltada somente para uso em dispositivos de memória não-volátil; a SOP também está sendo direcionada ao uso em adaptadores de barramento de host e controladoras RAID com recursos para fazer a ponte entre vários dispositivos SOP. A SOP também aproveita amplamente as arquiteturas e recursos existentes no setor, enquanto a NVMe usa um conjunto de instruções e interface de fila muito limitado.

Benefícios e problemas das interfaces

Cada uma das arquiteturas de armazenamento descritas tem benefícios e problemas. Dependendo do projeto do sistema geral, os benefícios de usar uma arquitetura específica pode superar os problemas associados a essa arquitetura. É necessária uma análise cuidadosa para que seja tomada a decisão certa. Essa decisão também deve levar em consideração a compatibilidade com o sistema existente.

Por exemplo, atualizar um sistema de laptop que tenha um HDD SATA de 2,5 polegadas existente com um SSD só daria certo se fosse usado um SSD com as mesmas dimensões e a mesma interface SATA (ou uma mais recente). Nesse caso, haverá um limite para a velocidade que o SSD poderá alcançar. Uma interface SATA de host mais rápida do que a existente não aumentará o desempenho do sistema.

Em uma situação parecida, um servidor empresarial que esteja usando um HDD SAS "short-stroked" de 15.000 RPM para armazenar um índice de banco de dados pode ser atualizado com um SSD SAS, que aumentará o desempenho geral do sistema, mas somente até o ponto em que outro fator do sistema se torne o novo gargalo (CPU, memória, rede, adaptadores etc.).

Em uma arquitetura de sistema mais nova, a adição do armazenamento de estado sólido pode aumentar significativamente o desempenho do sistema, mas somente na medida em que o resto da arquitetura do sistema possa acomodar a maior taxa de dados e largura de banda de dados. As taxas de dados mais rápidas nos SSDs também exigem mais alimentação do dispositivo e uma maior dissipação de calor, independente de onde o SSD seja montado.

Comparação das interfaces de SSD empresariais

Tabela 2. Comparação de SSD PCIe nativa e agregadora		
	Nativa	**Agregadora**
Comandos/transporte	Patenteado (FTL[1] em memória principal/host) 	SCSI ou SATA (várias SSDs, controladora na placa) 
Comitê	Nenhuma	Nenhuma
Baseado em padrões	Não	**Sim**
Desempenho com flash	Alto	Alto
Sobrecarga de CPU	Alto	**Baixo**
Latência com fila curta	Muito baixa	Baixa
Latência com fila profunda	Moderada	**Baixa**
Usa extensibilidade de caso	Não	**Sim** (RAID, HBA, etc.)
Maturidade	Em desenvolvimento	**Baseado em arquiteturas comprovadas do setor**
Conjunto de recursos para empresas (PI, segurança, gerenciamento etc.)	Não	**Depende da implementação**

[1] FTL: Flash Translation Layer (camada de tradução flash)
Fonte: Seagate Technology, 2011

Tabela 3. Comparação de SSD SOP e NVMe PCIe		
	SOP[1]	**NVMe[2]**
Comandos/transporte	SOP/PQI[3] (FTL em controladora) 	NVMe/NVMe (FTL em controladora) 
Comitê	**T10/INCITS**[4]	Industry Working Group
Baseado em padrões	**Sim (ANSI/ISO)**	Não
Desempenho com flash	Muito alto	Muito alto
Sobrecarga de CPU	Baixa	Baixa
Latência com fila curta	Muito baixa	Muito baixa
Latência com fila profunda	Baixa	Baixa
Usa extensibilidade de caso	**Sim** (RAID, HBA etc.)	Não (somente NVM)
Maturidade	**Baseado em arquiteturas comprovadas do setor**	A ser definido
Conjunto de recursos para empresas (PI, segurança, gerenciamento etc.)	**Compatibilidade total**	Limitado

[1] SOP: SCSI over PCI Express (SCSI sobre PCI Express)
[2] NVMe: Nonvolatile Memory Express (memória não volátil expressa)
[3] PCIe Queuing interface (interface de fila PCIe)
[4] INCITS: International Committee for Information Technology Standards (Comitê Internacional para Padrões de Tecnologia da Informação)
Fonte: Seagate Technology, 2011

Outro fator é o tempo para que haja compatibilidade entre os drivers de dispositivo do sistema operacional e BIOS e essas novas interfaces de SSD, bem como a confiabilidade inicial do software.

Fatos sobre a latência do SSD flash e interfaces

Há muitos equívocos sobre quais fatores adicionam latência e quanto eles realmente afetam o desempenho da aplicação. Ao analisar esse aspecto, é importante observar o panorama geral, não apenas parte dele.

O que mais contribui para a latência em SSDs são as próprias partes flash. Os tempos de acesso da SLC são 25 µs ou mais; os tempos de acesso da MLC são 50 µs ou mais, ambos supondo que não haja nenhuma contenção de acesso. À medida que a profundidade da fila aumenta, a contenção do acesso às partes flash podem aumentar substancialmente a latência.

Uma vez que uma parte flash inicia seu acesso, outras solicitações para a mesma parte devem esperar. Oito chips flash compartilham um mesmo barramento, fazendo com que os chips tenham que esperar sua vez para usar o barramento. Atividades de manutenção acrescentam latência adicional (tradução de endereço, coleta de lixo, nivelamento de desgaste etc.).

Outro aspecto é o sistema operacional, que acrescenta latência independentemente do protocolo de acesso e da interconexão. Aí estão incluídos o sistema de arquivos, gerenciador de volumes, drivers de classe e sobrecargas de troca de contexto.

Diferenças nos protocolos e interconexões têm efeitos insignificantes sobre a latência, como visto por um aplicativo (frações de um microssegundo).

www.seagate.com

AMÉRICAS Seagate Technology LLC 10200 South De Anza Boulevard, Cupertino, California 95014, Estados Unidos, +1 408 658 1000
ÁSIA/PACÍFICO Seagate Singapore International Headquarters Pte. Ltd. 7000 Ang Mo Kio Avenue 5, Cingapura 569877, +65 6485 3888
EUROPA, ORIENTE MÉDIO E ÁFRICA Seagate Technology SAS 16-18 rue du Dôme, 92100 Boulogne-Billancourt, França, +33 1 41 86 10 00

TP625.1-1203PT, março de 2012