Informacje o technologii

Porównanie interfejsów dysków SSD klasy korporacyjnej

Wprowadzenie

PCI Express (PCIe) to interfejs magistrali ogólnego przeznaczenia stosowany zarówno w klienckich i korporacyjnych aplikacjach obliczeniowych. Istniejące interfejsy pamięci masowej (SATA, SAS) podłącza się do komputera za pośrednictwem adapterów hostów, które z kolei podłączane są do interfejsu PCIe.

Interfejs SATA został zaprojektowany jako interfejs do dysków twardych (HDD), natomiast interfejs SAS zaprojektowano zarówno jako interfejs urządzeń oraz interfejs/infrastruktura podsystemu pamięci masowej. Wraz z ewoluowaniem dysków twardych i wymagań systemowych, niosącym za sobą zapotrzebowanie na szybsze interfejsy i nowe funkcje, interfejsy SATA i SAS przeszły kilka zmian.

Dyski półprzewodnikowe (SSD) szybko ustanowiły względem tych interfejsów nowe istotne wymagania eksploatacyjne, gdyż transfer danych z dysków SSD wzrósł z kilkudziesięciu MB/s do setek, a obecnie tysięcy MB/s. Oprócz zwiększenia szybkości transmisji danych, brak ruchomych części mechanicznych w dysku SSD również wpłynął na zwiększenie liczby operacji wejścia/wyjścia na sekundę (IOPS), które takie urządzenia pamięci masowej mogą wykonywać.

Ten rozwój skutkował potrzebą poprawy wdrażania istniejących standardów, a także rozszerzenia istniejących standardów interfejsów w celu sprostania nowym wymogom w zakresie wydajności, przy jednoczesnym zachowaniu kompatybilności z istniejącą architekturą systemu.

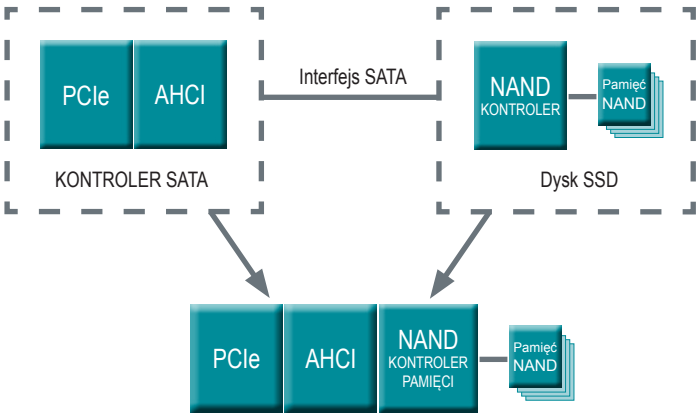
Niniejszy artykuł omawia różne interfejsy i prezentuje porównanie różnych napotykanych kompromisów w zakresie wydajności i kompatybilności.

Porównanie interfejsów dysków SSD klasy korporacyjnej

Interfejs SATA

SATA jest tanim interfejsem zaprojektowanym do punktowego połączenia za pośrednictwem kabla lub płytki drukowanej (PCB). Host podłączany jest do zaawansowanego interfejsu kontrolera hosta (AHCI), które zazwyczaj znajduje się w chipsecie hosta jako adapter hosta na magistrali PCIe. Istnieją pewne problemy projektowe z takim interfejsem, które mogą powodować obciążenie magistrali na poziomie 1 µs (lub więcej) dla każdego polecenia. Nie jest to poważny problem w przypadku dysków, w przypadku których transfer w formacie 4KB jest rzędu 10 µs, lecz dyski SSD mogą przesyłać 4KB danych w ciągu 2 µs (lub krócej) – w ten sposób powstaje istotne obciążenie, a interfejs SATA traci na znaczeniu jako interfejs o wysokiej wydajności przeznaczony dla pamięci masowej.

SATA jest nadal przydatny jako tani interfejs do dysków SSD, w przypadku których głównym czynnikiem decydującym jest koszt, a nie wydajność. Architekturę SATA można również konsolidować w adapter hosta, który zarządza zespołem poleceń interfejsu SATA bez faktycznego włączania fizycznego interfejsu SATA (PHY) (rys. 1).



Źródło: Seagate Technology, 2011

Rysunek 1. Konsolidacja architektury

Interfejs SAS

SAS jest również szeregowym interfejsem, dołączanym do komputera za pośrednictwem adaptera hosta, jednak występują istotne różnice, które sprawiają, że nadaje się on jako interfejs dysków SSD:

- mniejsze obciążenie sprzętowe,
- szybszy transfer,
- szerokie porty,
- wydajne interfejsy sterownik-kontroler.

Ponadto, interfejs SAS ma cechy, których nie ma interfejs SATA, a które zwiększają niezawodność i dostępność urządzeń podłączonych do interfejsu:

- niezawodny protokół szeregowy,
- obsługa wielu hostów,
- całościowa integralność danych,
- podwójny port,
- wysokie stopnie współbieżności i agregacji.

Mniejsze obciążenie sprzętowe

Nie ma uniwersalnego interfejsu hosta dla SAS, który byłby ekwiwalentny z kontrolerem SATA AHCI. Zamiast tego, wielu dostawców konkuruje na rynku adapterów hosta SAS, gdzie kluczowym czynnikiem jest wydajność – nie tylko do łączenia poszczególnych dysków HDD, lecz także różnych systemów RAID, w których prędkość transferu wielu dysków HDD jest grupowana w celu zwiększenia prędkości transferu. Ponadto, adaptery hosta SAS są przeznaczone do zarządzania wydajnymi dyskami SSD i dyskami HDD (np. tzw. *short-stroked* pracującymi z prędkością 15 tys. obr./min). Ponieważ adapter hosta sprzętu i sterownik urządzenia zarządzający tym adapterem hosta zaprojektowano jako system, nowe konstrukcje zoptymalizowane dla dysków SSD stają się dostępne i jeszcze bardziej zwiększają nie tylko prędkość transferu, lecz także liczbę operacji IOPS.

Wyższe prędkości przesyłu

Porty SAS obsługują obecnie do 6 Gb/s prędkości przesyłu danych. Firmy takie jak LSI i PMC-Sierra badają opracowywane obecnie konstrukcje, które mają obsługiwać prędkość przesyłu danych na poziomie 12 GB/s i ponad 2 mln operacji IOPS, z możliwością obsługiwania w przyszłości 24 GB/s.

Szerokie porty

Nieodłącznym elementem architektury SAS jest koncepcja szerokich portów, w której można sumować wiele łączy, aby zapewnić wiele jednoczesnych ścieżek pomiędzy jednym lub większą liczbą hostów a urządzeniem. Obecne złącze SAS napędu określa dwa porty na napęd. W kwestii wyboru konstrukcji, obecne dyski twarde nie obsługują szerokich portów, tylko podwójne porty, przy czym każdy port ma inny adres SAS, który zapobiega konfiguracji jako szeroki port.

Akceptowane propozycje dla *SAS-3* (12 GB/s) umożliwiają zwiększenie liczby portów na dysku do czterech, z których wszystkie można połączyć z tą samą jednostką lub parami do różnych jednostek. Oprócz podwójnego portu, na urządzeniu wyposażonym w dwa porty bardzo ograniczona liczba dysków SSD może obsługiwać szeroki port.

Porównanie interfejsów dysków SSD klasy korporacyjnej

Potężny protokół szeregowy

Protokół szeregowy SAS zapewnia wdrażanie szeregowych nadajników i odbiorników. Poprawia to jakość sygnału na kablu lub płycie bazowej poprzez kompensację długości kanału, niedopasowania impedancji i zakłóceń między symbolami. Protokół szeregowy SAS zarządza również wykrywaniem błędów i retransmisją na poziomie sprzętowym. Pozwala to na szybsze przywracanie po sporadycznych problemach sygnałowych.

Obsługa wielu hostów

Interfejs SAS i pole komutowane umożliwiają wielu hostom dostęp do tego samego urządzenia. Tej funkcji można używać do zarządzania podczas awarii hosta, a także awarii ścieżek danych w celu poprawy dostępności danych.

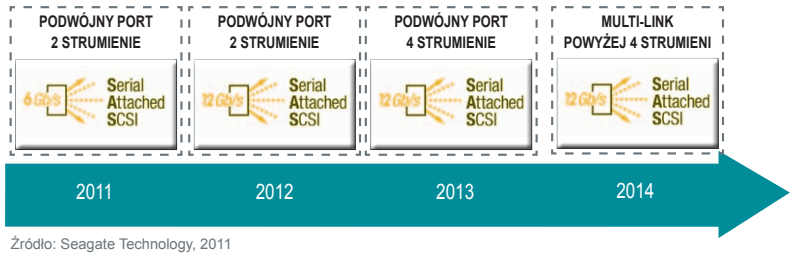
Całościowa integralność danych

Interfejs SAS może weryfikować integralność danych poprzez cykliczne kontrole nadmiarowości (CRC) danych każdorazowo po ich utworzeniu w buforze danych hosta, poprzez przesłanie interfejsem PCIe oraz interfejsem SAS, aż zostaną zachowane w urządzeniu i ponownie odczytane i przesłane do bufora danych hosta. Zapewnia to wiele punktów kontrolnych na ścieżce od aplikacji poprzez kontrolery RAID do urządzeń. Tę funkcję czasem określa się mianem ochrony informacji (protection information, PI).

Podwójny port

Docelowe urządzenia SAS obsługują operacje wykonywane przez podwójny port. Umożliwia to tworzenie dwóch tzw. „fault domain" (niezależnych sprzętowo jednostek, których równoczesne uszkodzenie jest bardzo mało prawdopodobne) i zapewnia większą dostępność. Nawet w przypadku uszkodzenia na jednej ze ścieżek do portu uniemożliwiającego uzyskanie dostępu na tej ścieżce, dostęp do urządzenia jest nadal możliwy przy użyciu drugiego portu.

Historycznie patrząc, firma Seagate wprowadziła przyjęcie interfejsu na rynku. Seagate współpracuje z SCSI Trade Association (STA) i innymi liderami w branży nad wykorzystaniem powszechnie stosowanej i istniejącej infrastruktury SAS (rys. 2). Tabela 1 pokazuje korzyści, które zapewnia interfejs SAS 12 GB/s i architektura multi-link konstruktorom systemów i organizacjom będącym użytkownikami końcowymi.



Źródło: Seagate Technology, 2011

Rysunek 2. Ewolucja interfejsu SAS

Interfejs PCI Express

PCI Express (PCIe) jest podstawowym interfejsem, który łączy urządzenia peryferyjne z procesorem komputera, a za pośrednictwem kontrolera pamięci z architekturą pamięci w systemie. Oba omówione wcześniej interfejsy SATA i SAS podłącza się za pomocą interfejsu PCIe (lub adaptera hosta) do procesora komputera i pamięci.

Tabela 1. Zalety i proponowane udoskonalenia typu multi-link interfejsu SAS

Wiele łączy (BW)	X4 (4x600 MB/s)
Dostępna moc	25 W (2,5 cala)
Całkowite opóźnienie	bardzo małe
Protokół multi-host	**tak**
Wysoka dostępność	**tak (podwójny port)**
Skalowalność	**doskonała**
Potężny, niezawodny stos protokołu	**tak**
Obsługa z możliwością wymiany w trakcie pracy	**tak**
Zgodny z istniejącym oprogramowaniem do zarządzania	**tak**

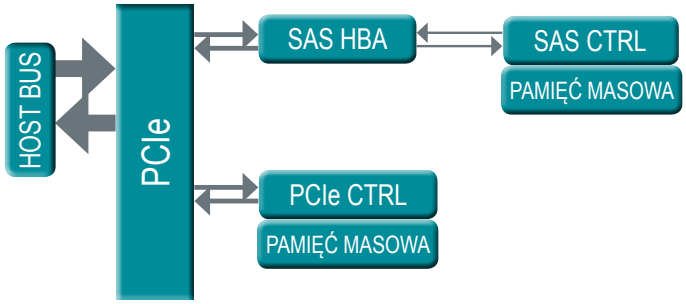
Źródło: Seagate Technology, 2011

Interfejs PCIe to szeregowa realizacja oryginalnego interfejsu PCI, która zapewniała równoległe połączenie adresu/danych pomiędzy urządzeniami peryferyjnymi a procesorem/pamięcią hosta. Interfejs PCIe komunikuje się za pośrednictwem jednego lub więcej *ścieżek*, na które składa się jeden nadajnikowy i jeden odbiornikowy interfejs szeregowy dla każdej ścieżki. W celu podłączenia hosta do urządzenia można wykorzystać do 32 ścieżek. Szeregowa prędkość transmisji danych po każdej ścieżce zależy od wersji realizowanego standardu PCIe, obecna wersja to 3.0, a prędkość transmisji danych wynosi około 1 GB/s.

W przypadku serwera 1U interfejs PCIe umożliwia korzystanie z jednego złącza w płycie głównej (kliencie) lub dwuzłączowego adaptera kątowego w płycie głównej (serwerze). Dostępny jest również system okablowania (choć jest on rzadko używany). Serwer 2U, 4U lub 7U ma znacznie więcej slotów PCIe, podobnie do wdrożeń klienckich. Specyfikacja PCIe również wykorzystuje wdrożenie nadajnika (i odbiornika) w celu dostosowania się do zmian impedancji w konfiguracji, lecz jest ukierunkowana jako kanały transmisji o krótszej długości niż SAS.

Switch PCIe może obsługiwać wirtualizację SR-IOV i wirtualizację MR-IOV – metody stosowane w celu poprawy wydajności kontrolera w systemach wirtualnych (hypervisor) z pojedynczymi lub wieloma hostami. Wirtualizacja SR-IOV dopiero staje się ogólnie dostępna w adapterach; jednak VMware nie może z niej jeszcze korzystać. Wirtualizacja MR-IOV zwykle nie jest obsługiwana w adapterach.

Porównanie interfejsów dysków SSD klasy korporacyjnej



Źródło: Seagate Technology, 2011
Rysunek 3. Ewolucja interfejsu SAS

Urządzenia pamięci masowej, które łączą się za pomocą interfejsu PCIe, dokonują tego albo poprzez bezpośrednie połączenie rejestru, albo poprzez adapter hosta, który następnie łączy się z urządzeniem za pośrednictwem dodatkowych kabli lub interfejsu w formie płyty bazowej.

Obecnie dostępnych jest wiele różnych implementacji obu architektur. SATA korzysta z implementacji adaptera magistrali hosta na chipsecie systemu *(mostek południowy)* – Intel lub AMD AHCI – wymagających różnych sterowników AHCI, ale odwzorowujących kompatybilne starsze implementacje IDE. Te interfejsy realizują również różne funkcje zarządzania RAID.

SAS ma wielu sprzedawców kart HBA, z dostępnymi dodatkowymi rozszerzeniami i kontrolerami RAID, które korzystają z przypisanych sterowników i BIOS-u w celu zaspokojenia różnych potrzeb w zakresie wydajności i konfigurowalności.

Interfejs PCIe sterownik-kontroler jest implementowany w specyfikacji NVM Express i w proponowanej specyfikacji SCSI over PCIe (SOP).

Opisana powyżej połączona architektura SATA stanowi kolejny przykład bezpośredniego połączenia rejestru PCIe.

Dyski SSD bazujące na interfejsie PCIe obecnie

Istnieją dwa podstawowe rodzaje dysków SSD bazujących na interfejsie PCIe na rynku: natywny i agregatorowy (zbiorczy). Natywny sterownik jest podłączany do magistrali PCIe hosta, a następnie bezpośrednio steruje wieloma szynami pamięci flash. Na ogół używają interfejsu oprogramowania, który jest przypisany do producenta i wykorzystywany wyłącznie dla określonego urządzenia. Część z tych implementacji przenosi ciężar translacji adresów i innych funkcji na hosta jednostki centralnej i pamięci. To z kolei powoduje zmniejszenie zasobów systemowych dla aplikacji, gdy urządzenia są wykorzystywane przy dużych obciążeniach. Dodatkowo, będąc stosunkowo nowymi rozwiązaniami na rynku, te unikalne napędy i kombinacje sprzętowe mają czasami skłonność do niestabilności, ponieważ ich ekosystemy są nadal rozwijane.

Model agregatorowy stosuje inne podejście do konstrukcji. Podejście to wykorzystuje istniejący kontroler SAS lub SATA RAID, do którego przyłączonych jest wiele dysków SSD z interfejsem SAS lub SATA. Są one łączone razem na jednej karcie PCIe. Kontroler RAID agreguje działanie wielu urządzeń zapewniając wysoką wydajność. Oparte o istniejący sprawdzony sprzęt klasy korporacyjnej i interfejsy oprogramowania konstrukcje są bardzo stabilne. Ponadto, te konstrukcje wykorzystują inteligentne kontrolery, które realizują operacje translacji adresów i inne funkcje, umożliwiając pełne wykorzystanie cykli systemu procesora i pamięci przez aplikacje, nawet przy dużych obciążeniach operacjami we./wy.

Przyszłość dysków SSD bazujących na interfejsie PCIe

Oba podejścia SOP i NVMe mają identyczną architekturę. Jednak NVMe jest tworzony przez branżową grupę roboczą, natomiast SOP przez uznane forum otwartych standardów. NVMe jest przeznaczony wyłącznie do użytku w urządzeniach posiadających pamięć nieulotną, natomiast SOP jest przeznaczony do użytku w adapterach HBA i sterownikach RAID z funkcjami mostkowania pomiędzy różnymi urządzeniami SOP. Dodatkowo, SOP intensywnie wykorzystuje istniejące branżowe architektury i funkcje, natomiast NVMe wykorzystuje nowy, bardzo ograniczony zestaw instrukcji i interfejs kolejkowania.

Zalety i problemy interfejsów

Każda z opisanych architektur pamięci masowych ma swoje zalety, jak również problemy. W zależności od ogólnej konstrukcji systemu, zalety wynikające z zastosowania określonej architektury mogą przeważyć nad problemami związanymi z tą architekturą, zatem niezbędna jest staranna analiza w celu podjęcia odpowiedniej decyzji. Taka decyzja musi także uwzględniać kompatybilność z istniejącą konstrukcją systemu.

Na przykład, modernizacja systemu komputerowego w laptopie, w którym znajduje się 2,5-calowy dysk twardy z interfejsem SATA, z wykorzystaniem dysku SSD powiedzie się tylko wtedy, gdy dysk SSD będzie mieć identyczne rozmiary fizyczne i taki sam (lub nowszy) interfejs SATA. W tym przypadku wystąpi ograniczenie prędkości działania dysku SSD; przekroczenie prędkości istniejącego interfejsu SATA hosta nie przyczyni się do lepszego działania systemu.

W podobnej sytuacji serwer klasy korporacyjnej wykorzystujący dysk short-stroked pracujący z prędkością 15 tys. obr./min z interfejsem SAS do przechowywania indeksu bazy danych można rozbudować przy użyciu dysku SSD z interfejsem SAS, zwiększając ogólną wydajność systemu, lecz tylko w takim stopniu, w jakim inny format systemu stanie się nowym wąskim gardłem (jednostka centralna, pamięć, sieć, karty itp.).

W nowej architekturze systemu dodanie dysku SSD może znacznie zwiększyć wydajność systemu, lecz tylko w takim stopniu, w jakim pozostała architektura systemu jest w stanie obsługiwać zwiększoną

Porównanie interfejsów dysków SSD klasy korporacyjnej

Seagate

Tabela 2. Porównanie dysków SSD z interfejsem macierzystym i zbiorczym PCIe		
	natywny	**agregatorowy**
Polecenia/Transport	Przypisany (FTL[1] w hoście/ pamięci głównej) 	SCSI lub SATA (wiele dysków SSD, kontroler na karcie) 
Komisja	brak	brak
Oparte o standardy	nie	**tak**
Wydajność z pamięcią Flash	wysoka	wysoka
Obciążenie procesora	wysokie	**niskie**
Opóźnienie z krótką kolejką	bardzo niskie	niskie
Opóźnienie przy długiej kolejce	średnia	**niskie**
Rozszerzalność przypadku użycia	nie	**tak** (RAID, HBA itp.)
Stabilność	ewolująca	**oparta o sprawdzone architektury**
Zestaw funkcji klasy korporacyjnej (PI, bezpieczeństwo, zarządzanie itp.)	nie	**zależy od implementacji**

[1] FTL: Flash Translation Layer

Źródło: Seagate Technology, 2011

Tabela 3. Porównanie interfejsów SOP i NVMe PCIe SSD		
	SOP[1]	**NVMe[2]**
Polecenia/Transport	SOP/PQI[3] (FTL w kontrolerze) 	NVMe/NVMe (FTL w kontrolerze) 
Komisja	**T10/INCITS[4]**	Branżowa Grupa Robocza
Oparte o standardy	**tak (ANSI/ISO)**	nie
Wydajność z pamięcią Flash	bardzo wysoka	bardzo wysoka
Obciążenie procesora	niskie	niskie
Opóźnienie przy krótkiej kolejce	bardzo niskie	bardzo niskie
Opóźnienie przy długiej kolejce	niskie	niskie
Rozszerzalność przypadku użycia	**tak** (RAID, HBA itp.)	nie (tylko NVM)
Stabilność	**oparta o sprawdzone architektury**	do ustalenia
Zestaw funkcji klasy korporacyjnej (PI, bezpieczeństwo, zarządzanie itp.)	**pełna obsługa**	ograniczony

[1] SOP: SCSI over PCI Express

[2] NVMe: Nonvolatile Memory Express

[3] PCIe Interfejs kolejkujący

[4] INCITS: International Committee for Information Technology Standards (Międzynarodowy Komitet Standardów Technologii Informacyjnych)

Źródło: Seagate Technology, 2011

www.seagate.com

prędkość transferu danych oraz przepustowość danych. Szybszy transfer danych w dyskach SSD wymaga dostarczenia większej ilości energii do urządzenia i rozproszenia większej ilości ciepła zawsze w przypadku zamontowania dysku SSD.

Innym czynnikiem jest czas sprawdzania dostępności sterowników systemów operacyjnych urządzeń i obsługa systemu BIOS dla takich nowych interfejsów dysków SSD, a także początkowa niezawodność oprogramowania.

Fakty dotyczące interfejsów i opóźnienia pamięci Flash w dyskach półprzewodnikowych (SSD)

Istnieje wiele błędnych przekonań na temat tego, jakie czynniki zwiększają opóźnienie i na ile mają faktycznie wpływ na działanie aplikacji. Przy analizie tego aspektu ważne jest, aby skupić się na ogólnym obrazie, a nie tylko na jednej jego części.

Czynnikami, które w przeważający sposób przyczyniają się do opóźnień dysków SSD, są same komponenty pamięci flash. Czasy dostępu do pamięci SLC wynoszą powyżej 25 µs; czasy dostępu do pamięci MLC wynoszą powyżej 50 µs, w obu przypadkach przy założeniu braku konfliktu dostępu. Ponieważ kolejki ulegają zwiększeniu, konflikt dostępu do elementów pamięci flash może w znacznym stopniu dodatkowo przyczyniać się do opóźnień.

Gdy część pamięci flash rozpoczyna uzyskiwanie dostępu, inne wywołania do tej samej części muszą czekać. Aż osiem struktur półprzewodnikowych pamięci flash dzieli wspólną magistralę, przez co struktury półprzewodnikowe czekają na swoją kolej za pomocą magistrali. Działania porządkowe dodatkowo zwiększają opóźnienia (translacja adresów, zbieranie śmieci, niwelacja zużycia itp.).

Innym aspektem jest system operacyjny, który zwiększa opóźnienia niezależnie od protokołu dostępu i wzajemnych połączeń. Należą do nich system plików, menedżer głośności, sterowniki klasy oraz koszty przełączania kontekstu.

Różnice w protokołach i łącznikach mają niewielki wpływ na opóźnienie postrzegane przez aplikację (ułamki mikrosekundy).

AMERYKA PŁN. I PŁD. Seagate Technology LLC 10200 South De Anza Boulevard, Cupertino, California 95014, USA, +1 408 658 1000
AZJA/PACYFIK Seagate Singapore International Headquarters Pte. Ltd. 7000 Ang Mo Kio Avenue 5, Singapur 569877, +65 6485 3888
EUROPA, BLISKI WSCHÓD I AFRYKA Seagate Technology SAS 16-18 rue du Dôme, 92100 Boulogne-Billancourt, Francja, +33 1 41 86 10 00