So eignet sich Flash über PCIe

von Hermann Strass

eine Publikation von

speicherguide.de
Das Storage-Magazin

powered by

SEAGATE

So eignet sich Flash über PCIe

PCIe wurde ursprünglich nicht für Speicheranwendungen entwickelt. Es ist aber ein sehr schnelles Transportmedium für Daten mit einer niedrigen Latenz. Gruppierungen von Speicherzellen auf einer einzelnen PCIe-Karte können als virtuelle SSDs oder als Hauptspeicher genutzt werden. Dadurch wird der Speicher auf der PCIe-Karte in den virtuellen Speicher des Anwenders eingebunden. Ein Überblick über die Möglichkeiten von Flash über PCIe.

Flash und PCIe sind allgemeine Bezeichnungen. Damit werden keine speziellen Techniken, Produkte Protokolle oder Elemente bezeichnet. Wie in dem früheren Artikel »So lassen sich die Vorteile von Flash richtig nutzen« beschrieben gibt es deutliche Unterschiede sogar in der grundlegenden Flash-Chip-Technik. Diese Unterschiede, weitere Schaltkreise und unterschiedliche Algorithmen sind die Grundlage für sehr unterschiedliche Produkte und Systeme, die unter dem Begriff Flash zusammengefasst werden.

Die herkömmlichen Transport-Techniken, wie in dem Beitrag »SAS oder SATA: Schnittstellen-Wissen für den richtigen Einsatz« beschrieben, gibt es ebenfalls in unterschiedlichen Variationen. Die hier beschriebene PCIe-Technik gibt es in sehr vielen praktischen und technischen Variationen. Einige davon sind freilich derzeit noch Nischenprodukte.

Exkurs in die PCI- und PCIe-Geschichte

»Peripheral Component Interconnect« (PCI) ist als eine parallele Verbindungstechnik (Medium für den Datentransport) ab 1991 (PCI 1.0) eingeführt worden. Die Buslänge war auf 10 cm (4 Zoll) begrenzt. Es gab keine besonderen Verfahren für den Transport von Hauptspeicher- oder Massenspeicherdaten. PCI wurde später um Steckplätze für Grafik- (AGP), Netzwerk- und E/A-Karten erweitert. Die übliche Begrenzung lag bei fünf Steckplätzen, bei CompactPCI für industrielle Anwendungen waren es bis zu acht Steckplätze. Für parallele (Bus-)Transportmedien gibt es mehrere technische Begrenzungen und Probleme.

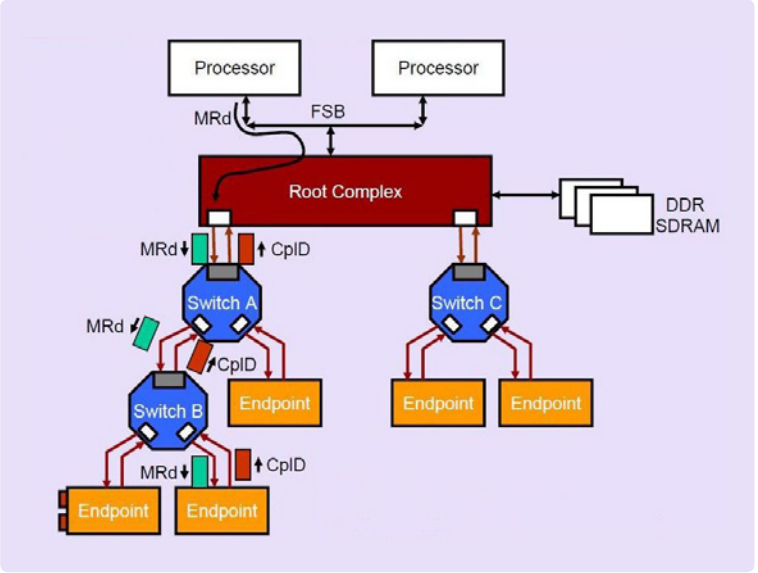


Bild: PCI-SIG

PCIe-Leitungen können durch Switches vervielfältigt werden

Die technischen Probleme bei parallelen Bussen waren die Gründe für den Umstieg in serielle Varianten wie PCIe, nachdem Chips für die benötigten höheren Übertragungsraten verfügbar wurden. Damit konnten die gleichen oder größere Datenmen-

gen wie bei parallelen Varianten übertragen werden. In den meisten Fällen entwickelte sich die serielle Technik aus einer Busstruktur in der sich viele Teilnehmer die Datenleitungen und damit auch die Bandbreite teilen. Eine Anzahl von einzelnen seriellen Verbindungen, bei denen an den Enden nur jeweils ein Gerät angeschlossen ist (Punkt-zu-Punkt-Verbindung), wie bei PCIe (PCI Express), ersetzen einen Parallelbus.

Das serielle PCIe-Protokoll wurde aus Kompatibilitätsgründen praktisch unverändert vom parallelen Protokoll übernommen. PCIe ging aus 3GIO und anderen Vorläufern hervor. PCIe 1.0 wurde 2002 freigegeben. Es gibt auch eine kleine Steckvariante, »ExpressCard« (früher »Newcard«) genannt, ähnlich wie PCMCIA-Karten. Eine weitere kleine Variante, Mini-PCI-Express genannt, ist eine internationale Variante für mobile Computer zur Nutzung als WLAN oder für andere Zwecke. Weitere Varianten sind beispielsweise SATAe, mSATA, XQD card, PCI ExpressModule, XMC, ATCA, AMC, FeaturePac, Thunderbolt, m.2 (früher NGFF), SCSI-over-PCIe (SOP), PCIe/104 und NVDIMM.

Die Unterschiede betreffen die physikalische Größe, die Anzahl der Leitungen (lanes). In einigen Fällen ist auch noch ein begleitendes serielles Medium, wie USB, als Teil der Transportmechanismen integriert. Thunderbolt ist eine serielle Kabelversion von PCIe zusammen mit DisplayPort. Industrielle Versionen von PCI, CompactPCI genannt, haben eine serielle Variante, »CompactPCI Serial« genannt, bei der es eine Anzahl von seriellen Übertragungswegen gibt, wie PCIe, SATA/SAS, USB und Ethernet.

PCIe ist sehr weit verbreitet, weil das Protokoll praktisch unverändert von PCI übernommen wurde. PCI ist weit verbreitet, um Chips über sehr kurze Distanzen auf einer Basisplatine zu verbinden. Diese frühe Version von PCI wurde zu einer (immer noch parallelen) Version mit Steckplätzen weiterentwickelt. So ist es verständlich, dass PCIe das bevorzugte Medium für PC-Anwendungen wurde, weil die Entwicklungsumgebung aus Software-Routinen, Protokollen und weitverbreitetem Expertenwissen über lange Zeit weitgehend unverändert oder kompatibel blieb.

Aus der Sicht von Anwendungen ist PCIe näher an der CPU (root complex) als SAS oder SATA. Das eliminiert Latenzzeiten und Protokollumwandlungsprobleme in der Hardware und Software des Zwischenelements: Steckkarte, Host-Bus-Adapter (HBA), Steuerelement oder ähnlich genannt. Wie noch erklärt wird, gibt es aber physikalische Grenzen bei der Anzahl der Elemente (Speicherkapazität), die an einen einzelnen »Root Complex« (Host, CPU) anschließbar sind.

Das müssen Sie wissen: PCIe-Grundlagen

PCIe wurde von Version 1.0 bis 3.0 weiterentwickelt (4.0 ist für 2015 geplant). Die Versionen erhielten jeweils höhere Transferraten (siehe Tabelle unten) und andere Verbesserungen. Die PCIe-3.0-Version mit 8 GT/s (giga transfers per second) ist doppelt so schnell wie Version 2.0, wozu die Anzahl der Redundanzbits von 8b/10b-Codierung auf 128b/130b-Codierung verringert wurde (siehe Tabelle). So wurde die Transferrate verdoppelt ohne die Anzahl der Transfers ganz zu verdoppeln. Bei seriellen Transfers werden die 8 Bit eines Byte in 10-Bit-Einheiten codiert.

Mit dieser 20-prozentigen Erhöhung der elektrischen Bandbreite werden längere Folgen von NULL-Bits vermieden zur Aufrechterhaltung der Taktung, und die Anzahl der NULL- und EINS-Bits auf der Übertragungsleitung ausgeglichen. Zur Verbesserung der verfügbaren Bandbreite wird in der PCIe-Version 3.0 (und zukünftig 4.0) stattdessen die 128b/130b-Codierung (nur 1,5 Prozent Aufschlag) eingesetzt. Die in der **Tabelle1** angegebenen Werte zeigen die Anzahl der Datentransfers je Leitung (lane). Dieser werden mit der Anzahl der Leitungen (bis zu x32) multipliziert, wenn mehrere Leitungen (lanes) je Verbindung (link) genutzt werden. Das Basiselement für die Datenübertragung ist die Leitung (lane), die aus zwei Paaren verdrillter Drähte besteht. Ein Drahtpaar dient zur Sendung von Daten, das andere zum Empfang von Daten, aus der Sicht des »Root Complex« (CPU). PCIe-Steckkarten gibt es in den für PCs (Stand- oder Tischgerät) üblichen Größen.

Leitungen (lanes) können gemäß den PCI-

Tabelle 1: Technischen Fakten und Funktionen der PCIe-Versionen

PCIe	Code	Transferrate	Nutzbare Bitrate	Nutzbare Byterate
1.0	8b/10b	2,5 GT/s	2 GBit/s	200 MByte/s
2.0	8b/10b	5 GT/s	4 GBit/s	400 MByte/s
3.0	128b/130b	8 GT/s	~ 8 GBit/s	~ 800 MByte/s
4.0*	128b/130b	16 GT/s	~ 16 GBit/s	~ 1600 MByte/s

* 4.0 noch nicht verfügbar

Quelle: Seagate/Hermann Strass

Tabelle 2: Standards für PCIe-Karten bezüglich Steckverbindern und Kartengrößen

Leitungen	Kontakte	Länge (mm)
x1	2 x 18 = 36	25
x4	2 x 32 = 64	39
x8	2 x 49 = 98	56
x16	2 x 82 = 164	89

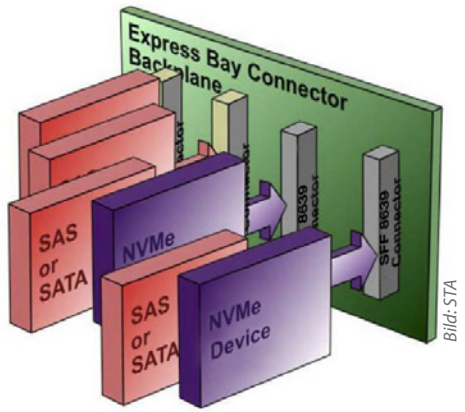
Quelle: Seagate/Hermann Strass

SIG-Standards zu Verbindungen (links) gebündelt werden. Definiert sind Leitungsbündel zu 1, 2, 4, 8, 12, 16 und 32 Leitungen, die entsprechend als x1 bis x32 gekennzeichnet werden. Am beliebtesten sind x1, x2, x4, x8 und x16 (für Grafikkarten). Derzeit gibt es im Handel keine x32-Varianten. Typischerweise verwalten sogenannte »Port Controller« 40 Leitungen, die zu zehn Verbindungen (links) zu je vier Leitungen (lanes) gebündelt werden. Kleinere Bündelungen werden direkt aus dem Root-Complex angeboten. Üblich sind 12-Port-Controller mit je einer Leitung (lane). Jeder Port (lane) kann individuell als PCIe x1, SATA oder USB konfiguriert werden. Doppelchips sind für zweimal 40 (80) Leitungen ausgelegt. Quadchips stellen dementsprechend bis zu 160 Leitungen zur Verfügung. Es gibt aber praktische Grenzen für die Leitungslängen auf der Basisplatine.

Auf PC-Basisplatinen (Motherboards) oder in Servern gibt es typischerweise einen (ersten) PCIe-Switch (Multiplexer), der direkt am Root-Complex angeschlossen ist. Diese Leitungen oder Verbindungen verbinden den Root-Complex mit PCIe-Endpunkten (Geräten) oder weiteren Switches im System, wodurch die Anzahl der Zugriffspunkte für PCIe-Datentransfers vervielfältigt wird. Endpunkte können auch als Brücken zu anderen Geräten oder Übertragungswegen eingesetzt werden.

Leitungen, die von einem Root-Complex ausgehen, können mit unterschiedlichen Geräten verbunden werden oder Daten mit unterschiedlichen Geschwindigkeiten übertragen. Leitungen als Teil von Verbindungen oder Ports arbeiten mit gleicher Geschwindigkeit und sind mit dem gleichen Gerät (endpoint) verbunden, weil die Datenbytes logisch über alle Leitungen verteilt übertragen werden, wodurch die Übertragungsrate vervielfältigt wird. Ein Root-Complex kann bis zu 256 Geräte (Endpoints) adressieren. Die Länge der Datenleitungen kann durch die Zwischenschaltung von Multiplexern (Switches) oder Takt-Regenerierung vergrößert werden. Gruppierungen von Speicherzellen auf einer einzelnen PCIe-Karte können als virtuelle SSDs oder als Hauptspeicher genutzt werden. Dadurch wird der Speicher auf der PCIe-Karte in den virtuellen Speicher des Anwenders eingebunden. Diese Nutzung ist nicht mehr herstellerspezifisch, weil **Intel** dieses Verfahren unterstützt. Dabei gibt es aber starke Einschränkungen, weil dies nicht über einen Pufferspeicher (Cache) möglich ist, und die die Datenmenge auf den halben Pufferbereich begrenzt ist.

Die Verarbeitung im Prozessor muss für die Dauer eines solchen Vorgangs pausieren, bis dieser abgeschlossen ist. Im Falle von Flash wird dadurch eine massive Verlängerung der Latenzzeit ausgelöst. DRAM und andere neuere Techniken haben deutlich niedrigere Latenzzeiten, weil diese in DIMM-Steckplätzen deutlich näher am Prozessor arbeiten als Speicher in PCIe-Steckplätzen.



Große Systeme werden mit Steckkarten in Rückwänden aufgebaut

Neuere Speicherarchitekturen werden den Vorteil von DIMM-Steckplätzen nutzen, weil dieser interne Speicherbus deutlich schneller arbeitet als ein E/A-Bus, wie PCIe oder SAS/SATA. Ein E/A-Bus ist sehr ineffizient, wenn es um Latenz, Geschwindigkeit und Komplexität gehr. Im PCIe-Protokoll sind keine Nutzungsklassen definiert, allerdings kann im Konfigurationsbereich der zugehörige Treiber definiert werden, den der Root-Complex laden soll. Für PCIe-Karten sind im Standard verschiedene Steckverbinder und Kartengrößen definiert, wie in der **Tabelle2** angegeben. Am wichtigsten ist die Länge des Steckbereichs. Auf den Karten werden beidseitig Kontakte an der Kartenkante als Steckverbinder genutzt. Karten mit 32 Leitungen (x32) sind derzeit nicht gebräuchlich. Externe Verkabelung ist nur für niedrige Geschwindigkeiten definiert. Das betrifft die Anbindung von externen Gehäusen, typischerweise mit SFF-8639-Steckverbindern für den Einsatz von Speicher im Format von Laufwerken.

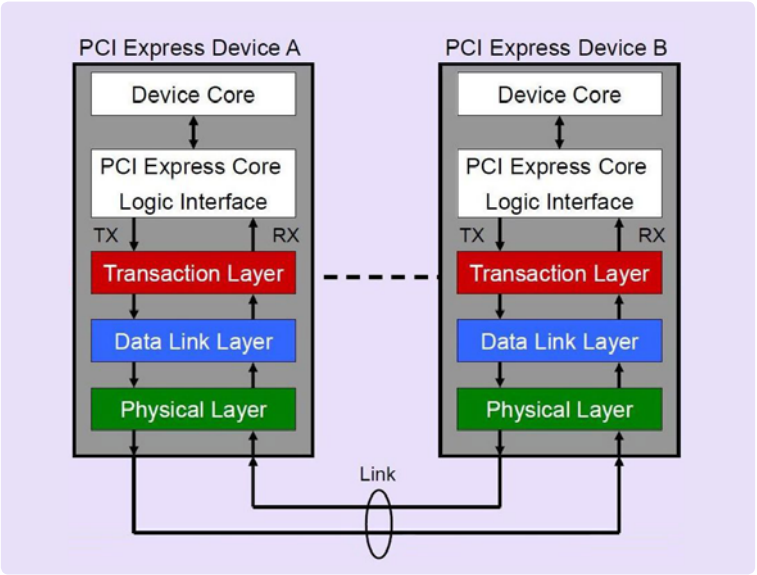


Bild: PCI-SIG

Übersicht PCIe-Protokollebenen

Eine PCIe-Karte passt in einen Steckplatz mit gleicher Größe oder größer. Steckplätze mit großer Länge können mit weniger logischen Leitungen verdrahtet werden, vorausgesetzt die nötigen Masseverbindungen für die insgesamt möglichen Leitungen sind verdrahtet. Das PCIe-Protokoll ermittelt die höchstmögliche Anzahl von gemeinsamen Verbindungen selbstständig.

PCIe in Flash-Systemen

Speichersysteme in großen Rechenzentren benötigen sehr große Speichermengen. Dazu werden PCIe-Karten mit den passenden Steckverbindern in Gehäusen in eine passende Rückwand gesteckt. Die Endpunkte (PCIe-Geräte, Brückenschaltungen, Root-Complex) kommunizieren über Software-Modul-Stufen. Die Daten fließen durch die Sendeseite (TX) nach unten, über die Leitung und auf der Empfangsseite (RX) wieder nach oben. Jede Modulstufe ist für einen bestimmten Teil des PCIe-Protokolls zuständig.

Für den Kartentausch bei laufendem Betrieb muss zuerst ein definierter Programmhalt (orderly shutdown) durchgeführt werden. Bevor eine Steckkarte gezogen wird, muss dem System protokollgerecht mitgeteilt werden, dass eine Karte gezogen wird. Das kann vermieden werden, wenn ein Multiplexer (Switch) zwischen dem Root-Complex und dem Gerät eingebaut ist. Das ist die übliche Vorgehensweise. Möglicherweise kann zukünftig der Switch wegfallen, allerdings wird sicher noch Takt-Regenerierung benötigt. Dieses Problem wird zukünftig noch deutlich größer werden, weil die Leitungslänge für die vierte PCIe-Generation halbiert werden muss. Mit der Takt-Regenerierung wird die nominelle Latenzzeit verlängert, weil die Daten zwischengespeichert und weitergeleitet werden müssen.

Viele Speicheranwendungen auf der Basis von PCIe-Steckkarten nutzen das AHCI-Protokoll, das ursprünglich für SATA entwickelt wurde. AHCI ist sehr gut bekannt, wodurch das Entwicklungsrisiko bei der Anwendung mit PCIe vermindert wird. Es gibt dazu eine Anzahl von Vorschlägen oder Prototypen zur Verbesserung bei der Anwendung mit dem PCIe-Protokoll.

Wie schon früher beschrieben reagiert Flash empfindlich auf Stromausfall. Datenverlust oder -verfälschung gibt es, wenn bei einem plötzlichen Stromausfall nicht genügend Energiereserve verfügbar ist, um den Inhalt von Zwischenspeichern (DRAM oder SRAM) rechtzeitig und vollständig in die Flash-Speicherzellen zu schreiben. Die geringe dafür benötigte Energie wird üblicherweise in einem entsprechend bemessenen »SuperCap« vorgehalten.

Große Speichersysteme sind mit einer unterbrechungsfreien Stromversorgung (USV,

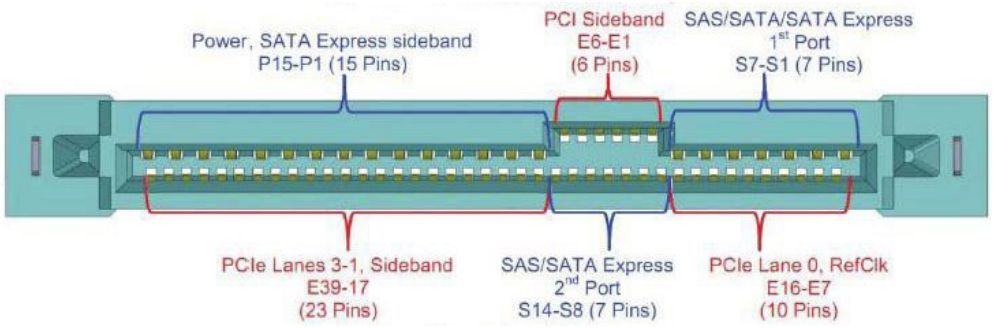


Bild: SFF committee

Bis zu vier Leitungen je Rückwand-Steckplatz

UPS) ausgestattet. Die darin gespeicherte Energie reicht auch für die schnelle Speicherung der temporären Daten im Flash. Aus Kostengründen gibt es weder Super-Caps noch USVs in kleinen Systemen in Heim oder Büro. Es gibt Notebook-Computer mit sehr großen Akkus, die diese Extraenergie liefern könnten, wenn die dafür benötigte Elektronik eingebaut wird. Die Qualität und Leistungsfähigkeit eines Flash-Speichersystems wird vorwiegend durch die Qualität der Speichersteuerung (Controller) im Flash-Gerät bestimmt.

Anwendungen mit PCIe-Flash-Karten

Beschleunigung eines Microsoft-SQL-Servers mit einer Seagate-Nytro-Karte: Der Microsoft-SQL-Server unterstützt unternehmenskritische Anwendungen mit besonders guter Leistung, Verfügbarkeit und mit wichtigen Funktionen für kritische Anwendungen. Die Funktion Pufferspeichererweiterung (»Buffer Pool Extension« = BPE) bietet dem Datenbank-Administrator die passende Funktion zur Erhöhung der Transaktionsleistung. Dazu wird nichtflüchtiger NAND-Flash-Speicher zur Vergrößerung des Pufferspeichers bereitgestellt.

BPE vergrößert sofort die Menge von direkt lesbaren Speicherseiten und vergrößert dabei die Datenintegrität bei Anwendungen mit einer großen Anzahl Leseanfragen. Speicherabhängige Anwendungen und/oder solche, die keinen zusätzlichen RAM-Hauptspeicher zur Verfügung haben, können davon profitieren. Der Nytro-Flash-Beschleuniger bietet kurze Latenzzeit und die Zuverlässigkeitsfunktionen für die optimale Nutzung von BPE.

Das BPE-Merkmal wird als Pufferspeicher auf Ebene 2 (level 2 cache = L2) eingesetzt, wobei der Pufferspeicher hauptsächlich auf Ebene 1 (L1) eingesetzt wird. Bei Anwendungen mit häufig vorkommenden Leseanfragen verlagert der SQL-Server diese Datenseiten automatisch in den L2-Pufferspeicher. Nur nicht mehr benötigte Seiten aus dem L1-Pufferspeicher werden aus Gründen der Datensicherheit übertragen. Mit einem einfachen »Ändere Server-Konfiguration«-Befehl werden der Ort und die Größe des BPE-Speicherbereichs festgelegt. Wird dieser Bereich auf die Nytro-Beschleunigerkarte gelegt, dann wird die Transaktionsleistung bei leselastiger Verarbeitung deutlich verbessert, wenn die Arbeitsbereiche nicht mehr in den L1-Pufferspeicher aus DRAM passen.

Die Nytro-Flash-Beschleunigerkarte ist ein primärer Halbleiterspeicher zur Beschleunigung von SQL-Server-Arbeitsabläufen. Wegen dem geringen Platzbedarf können Datenbank-Administratoren sehr leicht ihre vorhandenen Speichersubsysteme auf ein Halbleiterspeichersystem migrieren. Herkömmliche Festplatten beanspruchen die Ressourcen mit mehr Energie, Kühlung und Platz im Vergleich zu nur einem Steckplatz für Anwendungen, die in den verfügbaren Speicherplatz passen.

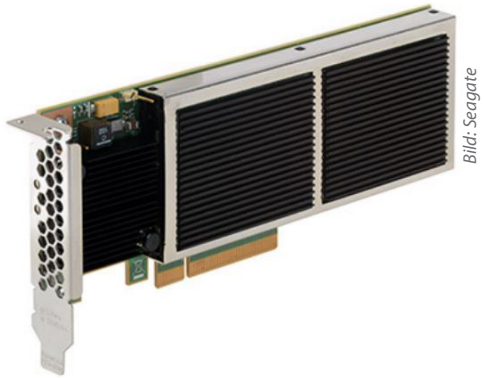


Bild: Seagate

PCIe-x8-Nytro-Speichermodul mit hoher Kapazität von Seagate

Die wichtigsten Merkmale der Seagate-Nytro-Karte

Die Nytro-Karte ist ausgerüstet mit den Funktionen moderner Halbleiterspeicher und einer Technik, die für lange Nutzungsdauer und für hohe Zuverlässigkeit in Anwendungen bei großen Unternehmen ausgerüstet ist. Die »host offload«-Technik (Auslagerung) verringert die Abhängigkeit von der CPU und vom DRAM, wodurch die Ressourcen für den Rechner und SQL-Sever verbleiben. Weitere Merkmale sind:

- In-box-Treiber für Windows Server 2012 R2
- Alle Microsoft-Treiber sind WHQL-qualifiziert
- Wird als Einzellaufwerk ohne Konfiguration durch den Anwender installiert
- DuraClass-Technik von SandForce für verbesserte Flash-Zuverlässigkeit, Nutzungsdauer und Energieeffizienz
- Datensicherheit gegen NAND-Flash-Fehler durch das RAID-ähnliche RAISE
- Dynamische Extraspeicherbereiche (overprovisioning)
- Weniger als fünf Sekunden Wiederanlaufzeit nach Stromausfall
- PCIe-2.0- und PCIe-3.0-Unterstützung
- schnelle Reaktionszeiten bis zu 50 Mikrosekunden
- Hoher Datendurchsatz (bis zu 4 GByte/s mit PCIe-3.0-Produkten)
- professionelle Qualität und Zuverlässigkeit.

Die Datenbank-Administratoren mit SQL Server 2014 können jetzt die Verarbeitungsleistung erhöhen durch Nutzung der Leistungsverbesserungen und Zuverlässigkeit

von nicht-flüchtigem Flash-Speicher. Die niedrige Latenz, hohe Leistung und Energieeffizienz von PCIe-Flash bringt eine sehr hohe Datenbank-Geschwindigkeit.

Zusammenfassung

PCIe wurde ursprünglich nicht für Speicheranwendungen entwickelt. Es ist aber ein sehr schnelles Transportmedium für Daten mit einer niedrigen Latenz. Es gibt eine große Vielfalt an Produkten, Software und Expertenwissen, das ständig, soweit technisch möglich, zur Anwendung in Flash-Speichersystemen verfeinert wird. Aus technischen Gründen ist PCIe bei der Nutzung in großen Systemen auf eine kleine Anzahl von Geräten (Steckkarten, Laufwerke) begrenzt. PCIe eignet sich auch gut als Ersatz für SATA, wenn ein Gerät an einem einzelnen Rechner beim Endanwender angeschlossen ist.

Die Anzahl der PCIe-Leitungen an einem Root-Complex oder Rechner ist ebenfalls begrenzt. Müssen viele Geräte zu großen Speichersystemen gebündelt werden, dann ist SAS die günstigere Architektur und das günstigere Transportmedium. Es hat zwar etwas mehr Latenz, eignet sich aber sehr gut zur Skalierung von Anwendungen in großen Anlagen oder Rechenzentren.

Hermann Strass

Sonderdruck

Veröffentlichung vom 28.04.2015

So eignet sich Flash über PCIe

Powered by:
Seagate Technology GmbH

SEAGATE

Messerschmittstr. 4
D-80992 München
Tel. +49 (0) 89/14 30 50 00
Fax +49 (0) 89/14 30 51 00
E-Mail: DiscSupport@Seagate.com
Web: **www.seagate.de**

speicherguide.de GbR
Karl Fröhlich, Ulrike Haak, Engelbert Hörmannsdorfer
Bahnhofstr. 8, D-83727 Schliersee
Tel. +49 (0) 80 26/928 89 96
E-Mail: redaktion@speicherguide.de

Chefredaktion:
Karl Fröhlich (verantwortlich für den redaktionellen Inhalt)
E-Mail: redaktion@speicherguide.de
Projektleitung:
Engelbert Hörmannsdorfer
Redaktion:
Hermann Strass

Layout/Grafik:
Uwe Klenner, Layout und Gestaltung,
Rittsteiger Str. 104, 94036 Passau,
Tel. 08 51-9 86 24 15
www.layout-und-gestaltung.de
Titelbild: © istockphoto/Henrik5000

Mediaberatung:
Claudia Hesse,
Tel. +41 (0) 41 - 780 04 86
E-Mail: media@speicherguide.de

speicherguide.de
Das Storage-Magazin