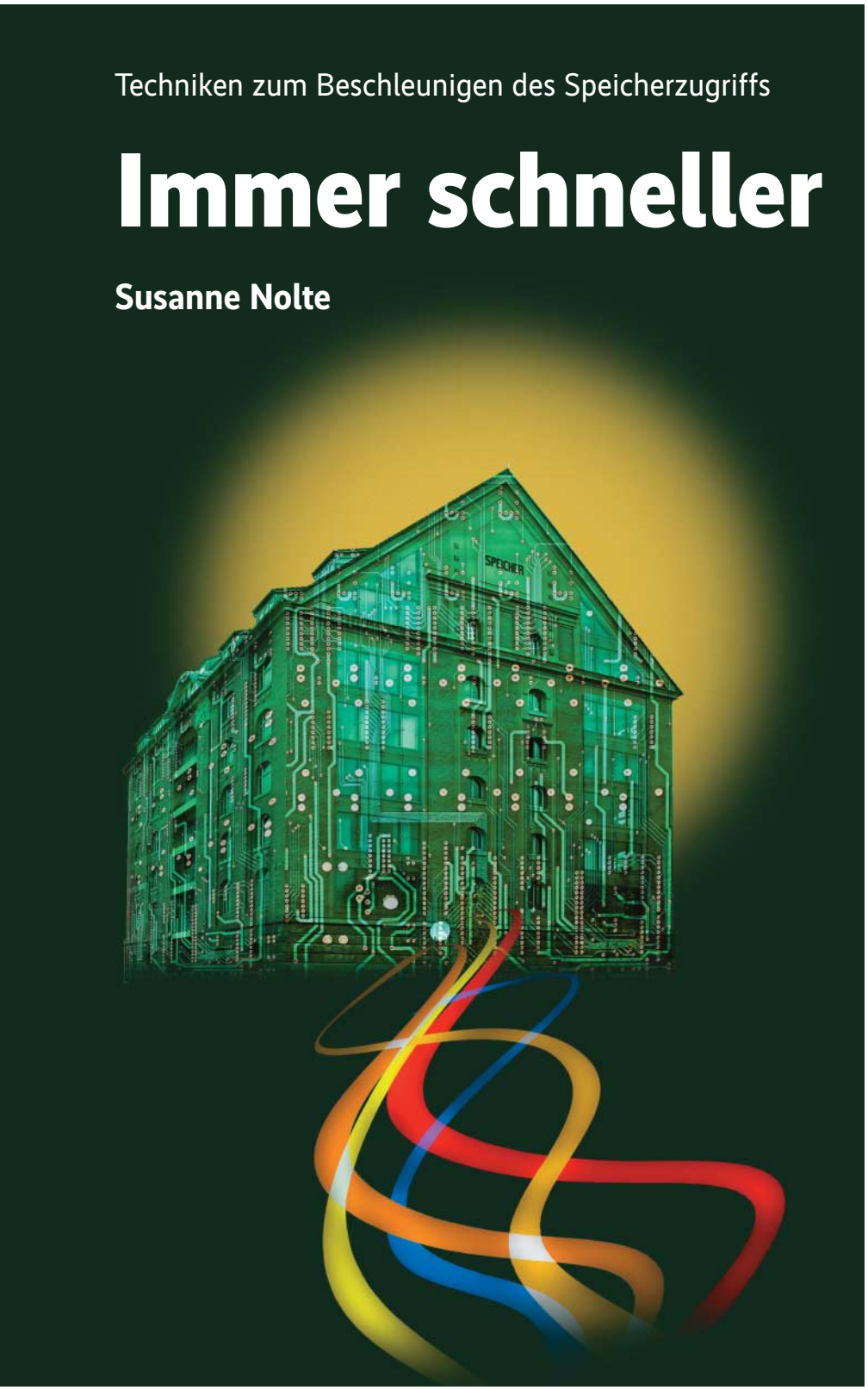
Techniken zum Beschleunigen des Speicherzugriffs

Immer schneller

Susanne Nolte



Trends wie SSDs, SDS, Cloud- oder Object-Storage verändern die Speicherlandschaft in vielfältiger Weise. Dabei verfolgen alle Konzepte nur ein Ziel: Storage günstig bereitzustellen und schnell zugreifbar zu halten.

Vorbei die Zeiten, in denen Speichermedien nicht groß genug sein konnten. Heute sind Kapazitätsengpässe beim Storage selten, und mit den heutigen Public Clouds auch ad hoc zu überbrücken.

An Grenzen stoßen Administratoren eher beim Datentransport, dem IO. Die Herausforderung heute liegt also im Erhöhen der Zugriffsgeschwindigkeit. Dabei ist es unerheblich, ob die Anwendungen auf SSDs, Platten oder Bänder zugreifen wollen, performant und verzögerungsfrei muss es vor allem vonstattengehen.

Doch sind die Medien nur ein Puzzlestein, und zwar ein immer kleinerer. Denn je mehr Schichten sich zwischen sie und die Anwendung schieben, desto mehr nehmen sie Einfluss auf den Datentransport. Deshalb bilden Schreib- und Lesegeschwindigkeiten respektive die berühmten IOPS nur die Basis. Einen reibungslosen Datenverkehr erreicht man über die vielen kleinen Stellschrauben darum herum. Genau darum geht es in den drei folgenden Artikeln.

Dass SSDs den Administrator diesem Ziel näher bringen, liegt auf der Hand: hohe IOPS-Zahlen, die vor allem Datenbanken und Umgebungen mit vielen kleinen konkurrierenden Zugriffen etwa durch viele Anwender oder viele VMs entgegenkommen.

Solid State Drives fürs Rechenzentrum

Bewegung kommt in die Schnittstellen und Formate: Klassiker waren bisher SATA, SAS und – im geringen Umfang – PCIe, nun soll der PCIe-Aufsatz NVMe (Non-volatile Memory Express) den Durchbruch schaffen. Bisherige Karten konnten ihre Stärken kaum ausspielen. Nun beherrschen Intels Skylake 48 und AMDs Epyc 128 PCIe-Lanes. Die Chips der Skylake- und Epyc-Generation sind mittlerweile für NVMe optimiert und haben Funktionen wie NVMe-Software-RAID an Bord.

Mit großer Spannung erwarten die Hersteller die ersten Produkte mit NVMe over Fabric (NVMe OF), also NVMe über Fibre Channel, InfiniBand, RoCE v2 oder iWARP (siehe unten). Parallel existieren Flash-Systeme mit proprietären Modulen wie IBMs FlashSystem 900, die durch zusätzliche Funktionen wie eigene ASICs zur Inline-Datenkomprimierung oder interne Crossbar-Switches herausstechen. Auch solche Systeme bekommen Konkurrenz von Markteinsteigern wie AIC, die mit Mission Peak ein Referenz-

© Copyright by Heise Medien
Persönliches PDF für alle Leser der iX aus 30625 Hannover

system für Karten in Samsungs eigenem Next Generation Standard Form Factor (NGSFF) vorgestellt haben.

Wer NAND-Flash-Medien mit den bisherigen Standardschnittstellen in seinen Servern oder seinem Software-defined Storage einsetzen will, muss einige Dinge beachten. Zuerst muss sich die Umgebung auf SSDs umstellen, denn viele Standardkonfigurationen setzen noch Festplatten voraus. Was etwa bei RAID-Adaptern zu beachten ist, zeigt der Artikel „Schnellspeichern“ zur Optimierung von SSD-RAID-Verbünden auf Seite 46.

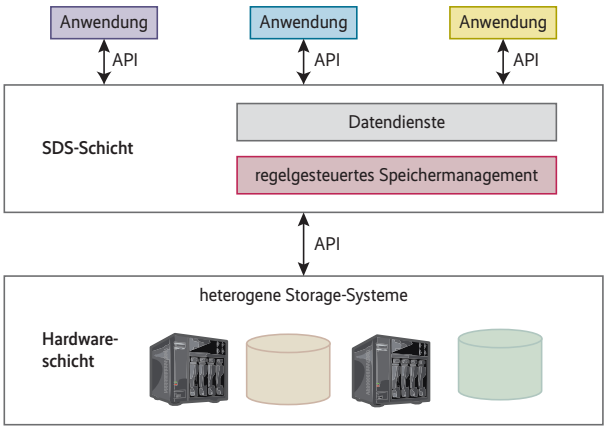
Vor allem müssen aber NAND-Flash-Medien enterprisefähig sein. Ein von Festplatten bekanntes Kriterium ist die MTBF (Mean Time Between Failure), die bei Enterprise-SSDs 1 Million Stunden nicht unterschreiten sollte. Ein wichtiges SSD-spezifisches Kriterium ist dagegen das Over-Provisioning (OP), also die Größe des überdimensionierten Bereichs, den die SSD als Ersatz für defekte Zellen verwendet.

Ein weiterer Punkt ist die Uncorrectable Bit Error Ratio (UBER). An dieser Datenfehlerrate lässt sich die Fähigkeit des SSD-Controllers ablesen, diese Bitfehler zu korrigieren. Hingegen bezeichnen die BER (Bit Error Ratio) und RBER (Residual Bit Error Ratio) die Rate, mit der Bitfehler im NAND-Flash ohne Ausgleich durch den ECC (Error Correction Code) auftreten.

Zusätzlich können Enterprise-SSDs über RAID-5-ähnliche Techniken verfügen, die die beschädigten Blöcke über Paritätsdaten wiederherstellen. Ein End-to-End-Datenschutz samt CRC (Cyclic Redundancy Check) und ECC kann die Datenintegrität vom internen Cache zum NAND-Speicherbereich prüfen.

Zudem enthalten SSDs der Enterprise-Klasse Schaltungen zum Erkennen von Stromverlust, die die Stromspeicherkondensatoren auf den SSDs verwalten. Ein solcher Powerfail-Support überwacht den eingehenden Strom, während die Kondensatoren das Abschließen der Schreibvorgänge bei einem überraschenden Stromverlust gewährleisten.

Software-defined Storage entkoppelt Anwendungssysteme und Speicherhardware, kann aber mit den zusätzlichen Zwischenschichten neue Performanceengpässe erzeugen (Abb. 1).



S.M.A.R.T.-Reportings sollten über die Standardangaben hinausgehen und nicht nur Werte wie den aktuellen Write Amplification Factor (WAF) ausgeben, sondern auch rechtzeitige Warnungen vor Ereignissen wie einem Stromausfall, Bitfehler am Port oder ungleichmäßigen Verschleiß enthalten. Zudem sollten die Hersteller ein entsprechendes Dienstprogramm zum Auswerten der Reports bereitstellen.

Nicht alles, was neu ist, eignet sich

Grundsätzlich sind hochkapazitive MLC- (Multi-Layer Cells) mit 256 GBit und TLC-Chips (Triple-Layer Cells) mit 384 GBit heute durchaus enterprisefähig. Der Grund für die Kapazitätssprünge und damit auch des Preisverfalls der letzten Jahre liegt in der Entwicklung der 3D-Technik. Dabei ordnen die Hersteller die Zellen nicht nur nebeneinander, sondern zusätzlich übereinander an. SSD-Chips im 16-nm-Verfahren mit 32 vertikalen Schichten sind auch bei RZ-tauglichen SSDs State of the Art. Im Sommer 2018 ist bereits mit der nächsten Generation 3D-NAND-Flash fürs Rechenzentrum zu rechnen – mit doppelter Kapazität.

Die kommenden 96 Layer und QLCs (Quadruple-Level Cells) mit vier Bits in einer Zelle bleiben aber noch eine Weile den Consumer-Produkten vorbehalten. Dasselbe gilt für PCM-Medien (Phase Change Memory) wie Intels viel beachtete Optane-Karte. Zu PCM existieren im Rechenzentrum noch keine Erfahrungen, ihre Tauglichkeit müssen sie also in Tests der nächsten ein bis zwei Jahre unter Beweis stellen.

Relativ wenig tut sich bei den Festplatten. Energiegestützte Aufzeichnungsverfahren wie Heat-assisted Magnetic Recording (HAMR) und Microwave-assisted Magnetic Recording (MAMR) sind noch nicht spruchreif und ein Performancesprung ist derzeit weder bei den Festplatten noch bei ihren Schnittstellen zu erwarten [1].

Durch den Preisverfall schneller Flash-Chips hat jedoch die Cache-Größe in Disks deutlich zugenommen. Sogenannte Hybridfestplatten oder SSHDs (Solid-State-Hybridfestplatten), die einen kleinen schnellen Speicherbereich aus Halbleitern mit einem großen magnetischen Bereich kombinieren, finden in Rechenzentren aber nur selten Verwendung. Eher kombiniert man enterprisefähige Festplatten mit SSDs in einer eigenen hierarchischen Speicherumgebung.

Veränderungen diesseits der Festplatte

Ebenfalls der Entwicklung des Flash-Marktes geschuldet ist die Neuerung bei den RAID-Controllern. Statt des Random Access Memory (RAM), den zusätzliche Akkus vor einem Datenverlust bei Stromausfall schützen müssen, setzten Hersteller bei neuen Modellen Flash

iX-TRACT

- Zu den großen Aufgaben heutiger IT-Umgebungen gehört die Sicherstellung eines performanten Datenflusses. Zugleich sind es immer weniger die Speichermedien, die das Nadelöhr bilden.
- Immer mehr Zwischenschichten ziehen zwischen die Anwendung und die Speicherhardware ein. Hier sind die meisten Stellschrauben zum Beschleunigen des Datentraffics zu finden.
- Obwohl Festplatten in ihrer Entwicklung immer weiter hinter Halbleiter- und Bandspeicher zurückfallen, verändert sich die diskzentrierte Speicherarchitektur nur wenig.

© Copyright by Heise Medien
Persönliches PDF für alle Leser der iX aus 30625 Hannover

ein. Der ist auf der einen Seite schnell genug als Schreibcache, benötigt als nichtflüchtiger Speicher aber keine eigene Stromversorgung, um die Daten zu halten. Waren bei den RAM-Modellen die Akkus meist optional, sind es nun die Flash-Caches selbst.

Veränderungen sind dagegen bei den sie umgebenden Architekturen zu verzeichnen. Zum einen ist es Mode, Festplatten als Object-Storage Device anzusprechen, zum anderen werden klassische Disksysteme von modernen Storage-Clustern abgelöst, allen voran von selbst konfigurierten nach dem SDS-Konzept (Software-defined Storage).

Beide Verfahren ziehen aber besagte zusätzliche Zwischenschichten zwischen Medien und Anwendung, die als Bremse wirken können (siehe Abbildung 1). Deshalb sind hier die meisten Schrauben zur Performanceoptimierung zu finden. Wie das im Detail aussieht, zeigt der Artikel „In Bestform“ anhand eines Ceph-Clusters auf Seite 38.

Die Technik, die in den letzten Jahren die größten Änderungen im Rechenzentrum verursacht hat, ist wohl die Cloud. Welchen Einfluss sie auf die Storage-Performance hat, die selbst je nach Internetanbindung eigentlich durch begrenzten Durchsatz und hohe Latenzen auffällt, erschließt sich nicht sofort, und dennoch gibt es einige Ansätze.

Neue Ideen nutzen die Cloud dabei nicht als Datenlager, das Festplatten- oder Tape-Systeme ersetzt, sondern als Rechenressource. Ein Ansatz nutzt die Cloud fürs Storage Resource Management. SRM gehört zu den SAN-Werkzeugen (Storage Area Network) und untersucht dort nicht nur die Speicherressourcen, sondern auch die Daten und analysiert beispielsweise das Verhältnis von Kapazität und Nutzung oder von Medientyp innerhalb der Speicherhierarchie und Relevanz oder Zugriffshäufigkeit der darauf befindlichen Daten. Auf diese Weise versucht SRM, zu wenig oder nicht sinnvoll genutzte Volumes zu identifizieren. Darüber hinaus kann sie Tendenzen aufzeigen, bei der Planung und der Automatisierung von Aufgaben helfen, etwa bei Backups und Performanceanalysen.

Analyseressourcen von außen

Die Großen der SAN-Branche bieten ihre SRM-Software oft im Rahmen ihrer großen Management-Frameworks an, etwa Dell EMC ViPR SRM, HPE Storage Essentials Suite, HDS Command Suite, IBM Tivoli Storage Productivity Center, NetApp OnCommand Insight oder Symantecs Veritas InfoScale Operations Manager. Andere bieten Stand-alone-Software an. Dazu gehören Aptare StorageConsole, CA Storage Resource Manager, Northern Parklife SRM und SolarWinds Storage Resource Monitor. Eine solche SRM-Software benötigt einen eigenen Server, auf dem sie ihre Arbeit verrichten kann, oder man lagert sie eben in die Cloud aus. IntelliMagic bietet mit Vision bereits eine SRM as a Service an, andere, vor allem die großen Anbieter, wollen folgen.

Einen Schritt weiter geht der zweite Ansatz, die Cloud für die Performanceoptimierung beim Datenschaufeln einzuspannen. Er kombiniert die Management- mit Analytics-Verfahren und versucht, Vorhersagen nicht nur über die Auslastung der Volumes und den Verlauf der Zugriffe zu treffen, sondern auch über mögliche Ausfälle und Performanceengpässe.

Als Newcomer versuchen sich etwa Arxscan mit seiner Arxview Data Center Analytics Engine und Visual Storage Intelligence mit seiner Storage Analytics Software ihren Platz im Markt des SAN Analytics sichern, aus den Reihen der Big Player hat etwa IBM mit Spectrum Control Storage Insights sein erstes Produkt zum „Analytics driven Data Management“ vorgestellt. Letzteres ist nur als Cloud-Dienst zu haben.

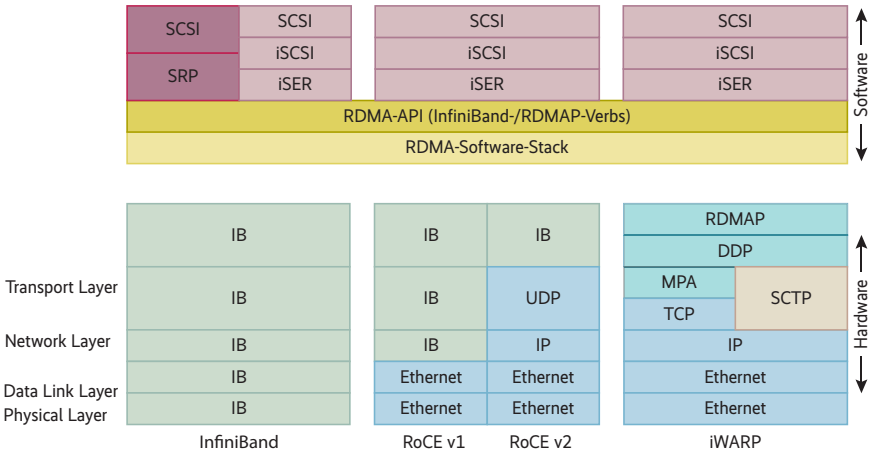
Wiederentdeckt haben Anbieter auch die klassischen Ansätze des hierarchischen Storage-Management (HSM), bei denen der Cloud-Speicher entsprechend seinen Zugriffslatenzen seinen Platz in der eigenen Storage-Landschaft hat. Da die Latenz von Internetverbindungen recht hoch ist, die Zugriffszeiten also über der von lokal installierten Festplattensystemen und unter der von Bandbibliotheken liegen, sind entfernte Cloud-Speicher auch genau zwischen diesen beiden einsortiert.

Zudem positionieren sich Cloud-Anbieter als Backup-Provider. Sei es, dass sie ihre Ressourcen für Point-in-Time-Kopien oder für asynchrone Replikationen bereitstellen oder als Backup-Backend. In diesem Fall schiebt der Kunde nur die alten Backups ins entfernte Rechenzentrum, solche, die ihre Existenz in erster Linie rechtlichen Vorgaben und firmeninternen Policies verdanken und die Administratoren höchstens anfassen, wenn ein Anwender vor Wochen gelöschte Daten wiederhaben will.

Einfluss auf die Protokollwahl

Indirekt und auf ziemlich unerwartete Weise nimmt die Cloud außerdem Einfluss auf die IO-Entwicklung. Ihre Provider haben für Fibre-Channel-SANs wenig übrig. Sie möchten ihre Storage-Systeme, ob nun mit SSDs, Disks oder Bändern bestückt, über Ethernet, beispielsweise über iSCSI, anbinden. Lieber wäre ihnen iSER (iSCSI Extensions for RDMA), also iSCSI mit RDMA. Remote Direct Memory Access oder Remote DMA benötigt aber einen passenden Unterbau. Zur Wahl stehen momentan InfiniBand (IB), iWARP (Internet Wide-Area RDMA Protocol) oder RoCE (RDMA over Converged Ethernet), genannt „Rocky“.

RDMA ist unter Performancegesichtspunkten insofern besonders interessant,



Neuere Storage-Protokolle setzen auf Remote DMA, das den direkten Zugriff auf den entfernten Arbeitsspeicher ohne Umkopieren der Daten gewährt. Zur Überbrückung zwischen den Rechnern verwenden sie bereits definierte Schichten des InfiniBand- oder des TCP/IP-Stack (Abb. 2).

© Copyright by Heise Medien
Persönliches PDF für alle Leser der iX aus 30625 Hannover

als es den von der Grafikkartenkommunikation bekannten Direktzugriff auf den Arbeitsspeicher ohne aufwendiges Umkopieren der Speicherseiten von einem entfernten Rechner aus erlaubt. RDMA ist bisher vor allem bei InfiniBand im Einsatz. Dadurch und durch den schlanken Aufbau kann IB seine Latenzen extrem gering halten.

Innerhalb eines InfiniBand-Netzes benutzt man in der Regel SRP (SCSI RDMA Protocol), das einem SCSI-Gerät über InfiniBand den direkten Zugriff auf den Arbeitsspeicher des entfernten Computers gewährt. SRP ist absichtlich auf das Nötigste beschränkt und hält die Latenzen niedrig.

Einen etwas größeren Overhead bringt iSER mit sich, das mit iSCSI als Schnittstelle zwischen SCSI und TCP/IP Funktionen wie das Target Discovery und den Simple Name Service beinhaltet. Will man diese Funktionen nutzen oder SCSI-Daten per RDMA übers Ethernet transportieren, bietet sich iSER an.

Nicht so einfach wie gedacht

Erste Wahl der iSER-Designer ist iWARP. Die ursprüngliche Intention von iWARP war es, RDMA-Zugriffe direkt über TCP (Transmission Control Protocol) zu bewerkstelligen. Letztlich gestaltete sich das Unterfangen schwieriger als angenommen.

Zuerst musste das Kernstück, das Direct Data Placement Protocol (DDP), zur Anwendungsseite hin durch ein separates RDMA Protocol (RDMAP) ergänzt werden. Zudem kann das verbindungsorientierte Protokoll TCP nicht mit der für RDMA notwendigen Begrenzung von Nachrichten umgehen. Die Vermittlung übernimmt deshalb ein Marker PDU Aligned oder MPA (siehe Abbildung 2).

Etwas eleganter lässt sich der gewünschte Protokollstapel mit der TCP-Alternative SCTP (Stream Control Transmission Protocol) bauen – dies auch die Empfehlung der IETF (Internet Engineering Task Force). Nachdem die ersten fünf RFCs 5040 bis 5044 von 2007 zu viele Fragen offen ließen, sollten die Ergänzungen RFC 6580 (2012), 6581 (2011) und 7306 aus dem Jahr 2014 Klarheit schaffen.

Als iWARP-Konkurrent gilt RoCE. Es kommt mit weniger Schichten aus und entlastet Leitungen und Arbeitsspeicher dadurch, dass es ohne TCP und Konsorten auskommt. In der Version 1 setzt RoCE den InfiniBand-Network- auf den Ethernet-Link-Layer, RoCE v2 ersetzt den InfiniBand-Network-Layer durch UDP/IP (siehe Abbildung 2). Allerdings lässt vor allem die RoCE v1 einige Fragen offen.

Allen DMA-Ansätzen übers Netz ist gemeinsam, dass sie möglichst viel der Protokollarbeit von den HCAs erledigen lassen wollen und damit nicht nur die Host-CPU entlasten, sondern auch Umwege in den Betriebssystemschichten vermeiden wollen. Auch das verringert die Latenz und erhöht damit den Durchsatz.

Auch am hinteren Ende der Speicherhierarchie, bei den Offline-Medien, spielt die Performance eine Rolle. Während optische Medien in Rechenzentren ausgedient haben, erlebten Tapes – entgegen allen Unkenrufen noch quicklebendig – eine Renaissance in der Fachpresse – Ransomware sei dank. Für Rechenzentren relevant sind heute nur noch drei Tape-Serien: StorageTeks T10000 ist über Sun bei Oracle gelandet und gehört wie IBMs Serie 3592, genannt Jaguar, zu den Laufwerken für große Rechenzentren. Das sogenannte High-End-Segment, also Serverräume und Rechenzentren, deckt das LTO-Format (Linear Tape Open) inzwischen allein ab. Andere Formate spielen keine Rolle mehr.

Oracle ist beim Modell StorageTek T10000D angelangt, das die normalen 8- und kurzen 1,6-TByte-Bänder mit maximal 252 MByte/s über das 16 Gbit/s schnelle Fibre Channel oder FICON beschreibt. Für IBMs aktuelles Jaguar-Modell TS1150 sind Cartridges mit 10, 7 und 2 TByte native verfügbar, die es durchgehend mit 360 MByte/s und im Burst-Modus mit 800 MByte/s beschreiben kann.

Auffällig ist die ganz eigene Tuning-Strategie bei den RZ-Laufwerken. Jenseits der Geschwindigkeitserhöhung beim Lesen, Schreiben und Spulen sollen kurze Bänder schneller zum Ziel, hier zur gewünschten Datei, führen, da sie die Such- und Spulzeiten deutlich verkürzen. IBMs Jaguar-Modelle beherrschen außerdem eine Partitionierung, wodurch sich Dateien, auf die schneller zugegriffen werden muss, auf kurze und damit übersichtliche Abschnitte legen lassen.

Für alle zum Pflichtprogramm gehört das vom LTO stammende LTFS (Linear Tape File System), das Bänder als Blocklaufwerke darstellt und damit bis zum Dateimanager durchreicht. Das macht Bandlaufwerke auch diesseits von Backup- und Archivsystemen handhabbar.

Ein echter Paradigmenwechsel steht 2018 an. Dann will IBM mit der nächsten Jaguar-Generation TS1160 eine Ethernet-Version liefern. Grund ist die große Zahl der Cloud-Provider unter den Tape-Kunden, die vor allem auf Ethernet setzen.

LTO hat in der nun achten Generation die Kapazität von 6 auf 12 TByte unkomprimiert pro Cartridge verdoppelt. Lässt das Datenformat eine hohe Kompressionsrate zu, erhöht sich die gespeicherte Datenmenge um das entsprechende Vielfache.

Anders als bei Festplatten erhöht sich die Schreib- und Lesegeschwindigkeit – eine andere als die serielle ergibt bei einem Streaming-Medium keinen Sinn – in ähnlicher Weise wie die Kapazität. Auch das Einschalten der Kompression beschleunigt das Lesen und Schreiben der Nutzdaten entsprechend der Kompressionsrate.

Allerdings übten die Entwickler diesmal Zurückhaltung. Auch aufgrund einiger technischer Neuerungen, etwa der Umstellung von GMR (Giant Magnetoresistance) auf TMR (Tunneling Magnetoresistance), stieg er nur vorsichtig von 300 auf 360 MByte/s, bei den halbhohen Laufwerken blieb er sogar unverändert. Zum Vergleich: Beim Wechsel von LTO-6 auf LTO-7 verdoppelte sich der Durchsatz beinahe von 160 auf 300 MByte/s.

Entwicklung mit klarer Zielsetzung

Pünktlich zur Markteinführung von LTO-8 erweiterte das LTO Consortium die Roadmap bis zur zwölften Generation und legte sie damit für die nächsten zehn Jahre fest. Demnach sollen bei LTO-9 bis zu 24 TByte, bei LTO-10 bis zu 48 TByte, bei LTO-11 maximal 96 TByte und bei LTO-12 gegebenenfalls 192 TByte native auf ein Band passen. Zwar hält sich das LTO Consortium mit den Durchsatzangaben neuerdings zurück, doch dürfte LTO die Marke von 1000 MByte/s in etwa acht Jahren geknackt haben – dann dauert das Schreiben eines TByte nicht mehr knapp 47, sondern weniger als 18 Minuten.

Zur Einordnung: Festplatten erzielen heute etwa 250 MByte/s ohne Aussichten auf große Performancesprünge. Was sie noch am Leben hält, nachdem Halbleiter- und Bandmedien sie abgehängt haben, sind andere Kriterien – etwa die gewohnte Handhabung und eine diskzentrierte Rechnerarchitektur. (sun@ix.de)

Literatur

[1] Susanne Nolte; Storage; Hürdenlauf; Kurz erklärt: Alternativ magnetisierte Festplatten; *iX* 1/2018, S. 106

© Copyright by Heise Medien
Persönliches PDF für alle Leser der iX aus 30625 Hannover