Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

High Performance Storage

von Dr. Joachim Wetzlar

Dr.-Ing. Joachim Wetzlar ist seit mehr denn 20 Jahren Senior Consultant der ComConsult Beratung und Planung GmbH und leitet dort das Competence Center „Data Center“. Er verfügt über einen erheblichen Erfahrungsschatz im praktischen Umgang mit Netzkomponenten und Serversystemen. Seine tiefen Detailkenntnisse der Kommunikations-Protokolle und entsprechender Messtechnik haben ihn in den zurückliegenden Jahren zahlreiche komplexe Fehlersituationen erfolgreich lösen lassen. Neben seiner Tätigkeit als Trouble-Shooter führt Herr Dr. Wetzlar als Projektleiter und Senior Consultant regelmäßig Netzwerk- WLAN- und RZ-Redesigns durch. Besucher von Seminaren und Kongressen schätzen ihn als kompetenten und lebendigen Referenten mit hohem Praxisbezug.

Die Diskussion Fiber Channel versus Fibre Channel over Ethernet (FCoE) ist inzwischen ein alter Hut. Die FCoE-Euphorie vom Anfang dieses Jahrzehnts ist verflogen; ungeachtet dessen werden heute beide Techniken eingesetzt.

Network Attached Storage (NAS) ist bei vielen unserer Kunden auf dem Vormarsch – auch in Bereichen, die aus Performance-Gründen traditionell mit Fibre Channel SAN betrieben wurden. Die Rede ist z.B. von Datenbanken oder Data Stores der Server-Virtualisierung. Auf der anderen Seite werden die Speicher immer schneller; Storage Tiering mit schnellen Solid State Disks (SSD) in der höchsten Speicherklasse wird zu einer Selbstverständlichkeit. Wenn aber der Speicher hohe Datenraten und Unmengen von Input/Output Operations pro Sekunde (IOPS) unterstützt, muss das auch für die Schnittstellen zu diesem Speicher gelten. Ja, Fibre Channel und geswitchtes Multi-Gigabit Ethernet sind bereits sehr schnell. Wie wäre es also, wenn man nun auf der Betriebssystemseite optimierte? Ein vielversprechender Ansatz dafür ist Remote Direct Memory Access (RDMA). Hier hat es in jüngster Vergangenheit einige interessante Neuentwicklungen gegeben. Sogar (und gerade) Microsoft ist auf den Zug aufgesprungen. Und auch eine altbekannte Nischentechnologie aus dem High Performance Computing kommt wieder zu ihrem Recht, das Infiniband.

Storage Area Networks (SANs) sind von Natur aus auf hohe Bitraten und geringe Latenz getrimmt. Insbesondere die geringe Latenz ist eine Voraussetzung dafür, dass Leitungen überhaupt mit hoher Bitrate ausgelastet werden können. Und für schnelle Datenbanken ist die Latenz gar der einzige Hemmschuh. Dementsprechend besitzen SANs im Allgemeinen eine geringe Ausdehnung mit nach Möglichkeit nur einem Layer-2-Switch zwischen Host Bus Adapter (HBA) des Servers und dem Speicherprozessor.

Data Center Bridging

Das gilt letztlich auch für Ethernet, wenn es für Storage-Zwecke eingesetzt wird. Data Center Switches sind auf geringe Latenzen getrimmt. Das Cut Through Switching der 90er Jahre, das Pakete bereits auszusenden begann, bevor sie vollständig empfangen waren, erlebt eine Renaissance. Und außerdem verfügen Data Center Switches über Techniken, die ein „Lossless Ethernet“ ermöglichen. Denn Paketverluste dürfen in SANs nicht auftreten da den Block-basierten Protokollen die Möglichkeit der Retransmissions fehlt. Und außerdem bremsen Retransmissions die Performance gnadenlos aus. Das kennen Sie von Ihren TCP-basierten Anwendungen in ausgedehnten LANs und WANs. Fibre Channel kennt dieses Problem nicht, hier sorgen eingebaute Layer-2-Flusskontrollmechanismen (insbesondere die Buffer Credits) für Verlustfreiheit.

Über „Lossless Ethernet“ wurde in den Netzwerk-Insidern der vergangenen Jahre

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

bereits ausführlich berichtet. Ich beschränke mich daher an dieser Stelle auf eine kurze Zusammenfassung der drei Mechanismen des Data Center Bridging (DCB).

- Congestion Notification (IEEE 802.1Qau): In einem geswitchten Netz können grundsätzlich alle Switch Ports von einer Überlast betroffen sein, wenn nämlich die entsprechende Warteschlange voll ist. Dasselbe gilt auch für die empfangende Netzwerkkarte. Solche Ports sind Congestion Points (CP) im Sinne des Standards. Der CP meldet Überlast an den Verursacher, also an die sendende Netzwerkkarte, die im Sinne des Standards der Reaction Point (RP) ist. Der RP begrenzt daraufhin die Rate des verursachenden Datenstroms.

- Priority-based Flow Control (IEEE 802.1Qbb): Hierbei wird der im Ethernet altbekannte Mechanismus des Pause Frame gem. 802.3x (oft als „Flow Control“ bezeichnet) um Prioritäten erweitert. Der Switch kann damit die Überlast verursachende Warteschlange seines Nachbarn „bremsen“. Alle übrigen Warteschlangen bleiben davon unbeeinflusst.

- Enhanced Transmission Selection (IEEE 802.1Qaz): Dabei handelt es sich sozusagen um die Spezifikation von Differentiated Services (DiffServ) auf Layer-2. Datenverkehr wird anhand der Prioritäts-Markierung im VLAN Tag in eine Verkehrsklasse eingeordnet. Die entsprechende Wartschlange wird vom Switch auf konfigurierbare Weise bedient, beispielsweise als Priority Queue oder per Rate Queuing.

Infiniband

Während das Data Center Bridging erst in den letzten Jahren entstanden ist, gibt es Infiniband bereits seit 15 Jahren. Die Infiniband Trade Association (IBTA, http://www.infinibandta.org/) wird von den Herstellern Cray, Emulex, HP, IBM, Intel, Mellanox, Microsoft und Oracle angeführt. Sie prüft Infiniband-Produkte auf Einhaltung der Spezifikationen und auf Interoperabilität, letzteres auf halbjährlich stattfindenden „Plugfests“. Die aktuelle Liste kompatibler Produkte („Integrators‘ List“) umfasst mehr als 400 Produkte. Damit spielt die IBTA für Infiniband eine ähnliche Rolle, wie die Wi-Fi Alliance für WLAN.

Infiniband dient – wie Fibre Channel – zur Verbindung zwischen Servern („Hosts“) und ihren Peripheriegeräten. Die Adapter auf Server-Seite heißen dementsprechend Host Channel Adapter (HCA) und die auf Peripherie-Seite Target Channel Adapter (TCA). Dazwischen befindet sich eine geswitchte Fabric, wie bei Ethernet oder Fibre Channel.

Infiniband zeichnet sich durch hohe Performance und sehr geringe Latenzen aus; Round Trip Times liegen bei vielen Produkten im einstelligen Mikrosekunden-Bereich. Dementsprechend lassen sich über ein Infiniband-Netz sehr hohe Befehls-Raten (I/O Operations per Second, IOPS) erzielen. Infiniband gibt es derzeit in verschiedenen Bitraten, beginnend bei der Single Data Rate (SDR) mit 2 Gbit/s über die Double Data Rate (DDR) und Quad Data Rate (QDR) bis zur Fourteen Data Rate (FDR) mit 14,0625 Gbit/s. Eine Enhanced Data Rate (EDR) mit 25 Gbit/s und High Data Rate (HDR) mit 50 Gbit/s sind in der Entwicklung. Außerdem ist in der Infiniband-Spezifikation bereits die Link Aggregation enthalten. Im Server-Bereich eingesetzte Adapter verfügen häufig über 4 parallele „Lanes“ („4x“), leisten also 56 Gbit/s bei FDR. Für Supercomputer gibt es sogar Adapter mit 12 Lanes.
Interessant an der Infiniband Link Aggregation ist, dass die Bits eines Paketes gleichmäßig auf alle verfügbaren Lanes aufgeteilt werden. Dadurch wird die Auslastung aller Lanes immer gleich groß. Und außerdem verringert sich die Zeitdauer zur Übertragung eines Pakets. Im Gegensatz dazu verteilt die Ethernet Link Aggregation die Last pro Paket. Schaltet man also 5 mal 10 Gigabit Ethernet zu einer Link Aggregation zusammen (um etwa die FDR mit 4 Lanes zu erzielen), ist die Bitrate zwar fünfmal so hoch, das einzelne Paket wird aber nach wie vor mit 10 Gbit/s ausgesendet. Demgegenüber sendet die 4xFDR das einzelne Paket de facto mit 56 Gbit/s in einem Fünftel der Zeit aus.

Wie kommt es, dass diese eigentlich so interessante Technik erst jetzt in den Fokus der RZ-Betreiber rückt? Ganz einfach: Die Technik wird nun von der Firma Microsoft eingesetzt, um den Windows Server richtig schnell zu machen. Der Anwender braucht nun nicht mehr auf leistungsfähige „Fremdprodukte“ zurückzugreifen, um z.B. große Farmen virtueller Server zu betreiben. Das Verfahren, das dieser Technik zu Grunde liegt, ist „SMB Direct“. SMB Direct basiert auf einer Technik, die sich „Remote Direct Memory Access“ (RDMA) nennt. Und das wiederum ist eine Erweiterung des schon lange bekannten DMA. SMB Direct benötigt außerdem „SMB Multichannel“. Fangen wir also ganz langsam und von vorne an.

Direct Memory Access

Die Von-Neumann-Architektur, auf der letztlich alle unsere Computer basieren, verfügt bekanntlich über einen zentralen Speicher für Daten und Programme. Ein Prozessor greift auf diesen Speicher zu und verwendet dafür ein Bus-System, das aus Daten und Adressbus besteht. Auch Peripheriegeräte sind an dieses gemeinsame Bus-System angeschlossen. Empfängt ein Peripheriegerät Daten, liest der Prozessor diese Daten über das Bussystem in eines seiner internen Register ein und schreibt sie anschließend über das Bus-System an sinnvoller Stelle in den Speicher. Nehmen wir an, bei dem Peripheriegerät handelte es sich um eine Netzwerkkarte. Dann setzte der Prozessor solche Lese- und Schreibvorgänge dazu ein, um beispielsweise die einzelnen Segmente eines TCP-Datenstroms zusammenzusetzen.

Es ist offensichtlich, dass das Bus-System bei derlei Ein-/Ausgabeoperationen einem Flaschenhals gleichkommt. Direct Memory Access (DMA) schafft hier Abhilfe, indem er dem Peripheriegerät die Möglichkeit gibt, Daten unmittelbar in den Speicher zu schreiben bzw. daraus zu lesen. Die Netzwerkkarte aus dem Beispiel könnte also den gesamten TCP-Datenstrom im Speicher selber zusammensetzen, ohne dass der Prozessor dafür etwas tun müsste. Erst wenn alle Daten im Speicher liegen, läse der Prozessor sie und führte sie damit der Anwendung zu.

Das Peripheriegerät übernimmt beim DMA zeitweise die Kontrolle über das Bus-System. Da der Prozessor zu bestimmten

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

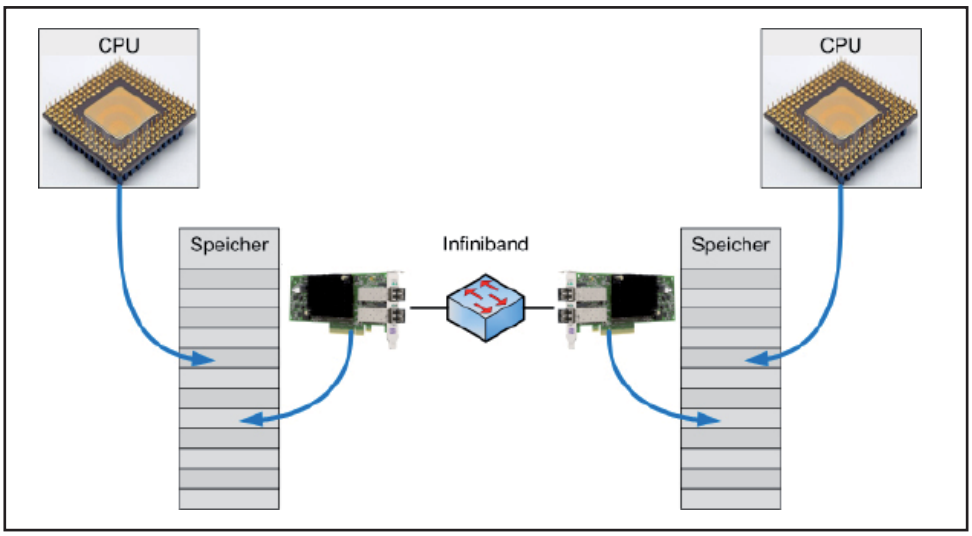

Abbildung 1: Zum Prinzip von RDMA

Zeiten anderweitig beschäftigt ist und nicht auf das Bus-System zugreift, lässt sich DMA so gestalten, dass der Prozessor nicht ausgebremst wird. Letztlich steigt durch DMA die Effizienz des Bus-Systems.

Remote Direct Memory Access

Stellen Sie sich nun zwei Netzwerkkarten vor, die über eine Infiniband Fabric gekoppelt sind. Beide Netzwerkkarten greifen per DMA auf einen Speicherbereich ihres Computers zu und kopieren die Daten. Der Prozessor von Computer 1 legt Daten im RDMA-Speicherbereich der Netzwerkkarte ab. Die Netzwerkkarte im Computer 1 holt diese Daten per DMA ab und sendet sie per Infiniband an die Netzwerkkarte im Computer 2. Diese legt die Daten per DMA im Speicher ab. Von dort kann sie der Prozessor des Computers 2 weiter verarbeiten. Die Abbildung 1 illustriert dieses Prinzip.

RDMA schöpft seinen Performance-Gewinn nicht nur aus der Tatsache, dass die Netzwerkkarten per DMA auf den Speicher zugreifen – in der Tat gibt es das schon sehr lange. Mit entsprechender Unterstützung des Betriebssystems kann dank RDMA die eigentliche Datenübertragung sogar ganz ohne Mitwirkung des Protokollstacks erfolgen. Die Anwendungen auf beiden Computern kommunizieren also direkt miteinander, indem sie auf einen gemeinsamen Speicherbereich schreiben bzw. davon lesen, der mittels RDMA quasi gespiegelt wird.

In Abbildung 2 habe ich den Unterschied deutlich gemacht. Bei der herkömmlichen Kommunikation über TCP Sockets erledigt der Prozessor einen erheblichen Teil der Arbeit. Er muss die Daten der Anwendung in Segmente zerlegen und dabei die Sequenz- und Acknowledge-Nummern des TCP verwalten. Danach verpackt er die Segmente in IP-Pakete und übergibt diese dem Kartentreiber. Der ist aber auch nur ein Stück Software, das wieder vom Prozessor abgearbeitet wird. Letztlich werden die Ethernet-Pakete in Register auf der Netzwerkkarte geschrieben, die danach (ohne weiteres Zutun des Prozessors) für das Aussenden im Netz sorgt. Sicher, so genannte TCP Offload Engines erlauben die Auslagerung dieses Geschäfts auf die Netzwerkkarte. Allerdings habe ich noch keine wirklich gut funktionierende Implementierung von TCP Offloading gesehen. Zu eng ist doch die Abhängigkeit zwischen TCP Stack im Betriebssystem und TCP Stack auf der Netzwerkkarte.

Bei RDMA werden die Daten dagegen gänzlich ohne TCP kopiert. Eine Abhängigkeit zur TCP-Implementierung des Betriebssystems besteht also nicht. Die Anwendungen kommunizieren gleichsam am TCP Stack vorbei (rechte Seite der Abbildung 2). Da in diesem Fall offensichtlich die Fehlerkorrektur-Mechanismen des TCP nicht mehr wirken, muss die Fehlerfreiheit der Übertragung auf andere Weise sichergestellt werden. RDMA greift dazu auf die entsprechenden Mechanismen des Infiniband zurück. Infiniband verfügt nämlich über wirksame Flusskontroll-Mechanismen. So gibt es einerseits eine Flusskontrolle auf Layer 2 – vergleichbar zu den Buffer Credits des Fibre Channel – und andererseits eine Ende-zu-Ende Flusskontrolle zwischen den Adaptern.

RDMA over Converged Ethernet

Eine Technik verkaufen zu wollen, die nur mit einem „Nischenprodukt“ wie Infiniband einzusetzen ist, könnte wohl als kurzsichtig bezeichnet werden. Ethernet ist nun mal der Stand der Technik, allen Nachteilen zum Trotz. Die Infiniband Trade Association hat daher RDMA auch für die Übertragung im Ethernet spezifiziert [1] und ihm dem Namen RDMA over Converged Ethernet (RoCE, gesprochen „Rocky“) gegeben. Infiniband-Pakete werden dabei einfach mit einem Ethernet MAC Header und entsprechender Prüfsumme versehen. Es handelt sich (wie bei IP) um Ethernet Frames des Typs 2. Als Typ Code wird der Wert 8915 hexadezimal verwendet.

Da sich Infiniband auf die Layer-2-Flusskontrolle seiner Fabrics verlässt, muss Ethernet für RoCE also Vergleichbares bereitstellen. RoCE kann letztlich nur mit einem Lossless Ethernet auf der Basis von Data Center Bridging (siehe oben) wirklich gut funktionieren. Interessanterweise verzichtet die Spezifikation explizit darauf, ein solches Lossless Ethernet zu fordern.

Seit einiger Zeit gibt es sogar eine Spezifikation[2] RoCEv2 zur Bereitstellung von RDMA über IP-Netze. RoCEv2 ermöglicht die Enkapsulierung der Infiniband-Pakete in UDP. Im Oktober des letzten Jahres wurde von der Internet Assigned Numbers Authority (IANA) dafür das UDP Port 4791 reserviert. Auch in IP-Netzen, insbesondere bei der Verwendung von UDP, stellt sich die Frage nach der Verlustfreiheit. RoCEv2 empfiehlt den Einsatz von Priority-based

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

Flow Control (IEEE 802.1Qbb, siehe oben), lässt jedoch auch vergleichbare Verfahren zu. Darüber hinaus wird der Einsatz von Quality of Service (QoS) empfohlen, entweder IP-basiert oder mittels Enhanced Transmission Selection (IEEE 802.1Qaz, siehe oben). Und nicht zuletzt empfiehlt die IBTA den Einsatz von Explicit Congestion Notification (ECN) gemäß RFC 3168.

SMB Multichannel

Der Begriff „Server Message Block" (SMB) bezeichnet das Dateitransfer-Protokoll der Windows-Welt. Wie das entsprechende Network File System (NFS) aus der UNIX-Welt existiert SMB schon sehr lange. Microsoft hat es in den letzten Jahren immer wieder optimiert. SMB 2 – mit Windows Vista bzw. Server 2008 eingeführt – brachte ein gestrafftes Protokoll, das sich auf die Übertragung hoher Bitraten auf weiten Entfernungen fokussiert. Mit Windows 8 bzw. Server 2012 kommt nun SMB 3, das neue Features für den Einsatz im Rechenzentrum im Gepäck hat. Eines dieser Features ist „SMB Multichannel".

Die Idee hinter SMB Multichannel ist, einen Ersatz für die meist proprietären Techniken des Adapter Teaming zu bieten. Eine Freigabe (Share) auf einem SMB Server kann dank SMB Multichannel über mehrere Zieladressen erreicht werden. SMB Server werden dazu mit mehreren Netzwerkkarten ausgestattet, die sich in unterschiedlichen IP-Subnetzen befinden. Gleiches gilt auch für die SMB Clients. Typischerweise handelt es sich bei diesen „Clients" um andere Server, die auf die File Server zugreifen, wie beispielsweise Datenbanken oder Virtualisierungs-Hosts. Der SMB Client kann jetzt mehrere TCP Sessions zum Server aufbauen und seine Zugriffe auf die Sessions verteilen. SMB Multichannel funktioniert sogar auf einzelnen Netzwerkkarten, wenn diese das so genannte Receive Side Scaling (RSS) unterstützen. In diesem Fall werden bis zu vier parallele TCP Sessions zu einer entsprechenden Netzwerkkarte im SMB Server aufgebaut. Der Geschwindigkeitsgewinn resultiert daraus, dass die TCP Sessions von unterschiedlichen Prozessorkernen verarbeitet werden können.

SMB Direct

Setzt man im SMB Server und Client Netzwerkkarten ein, die RDMA bzw. RoCE unterstützen, kann SMB Multichannel in einer besonderen Spielart wirken: Zunächst baut der SMB Client per TCP die Verbindung zum SMB Server auf. Danach etablieren SMB Client und Server eine RDMA-Verbindung. Infolgedessen werden alle SMB-Befehle per RDMA übertragen, wodurch sich wegen der verringerten Laufzeit die Performance gegenüber der TCP/IP-basierten Verbindung erhöht. Das ist in Abbildung 3 skizziert. SMB Multicannel bezieht sich bei SMB Direct auf die Parallelität von TCP-Verbindung und RDMA. SMB Direct ist ohne SMB Multicannel nicht denkbar. Selbstverständlich lassen sich dafür beliebige RDMA-fähige Netzwerkkarten einsetzen, also Infiniband oder Ethernet mit RoCE bzw. RoCEv2.

Wer wird nun die bis hierher beschriebene Technik einsetzen wollen? Denn immerhin muss die hohe Performance mit entsprechenden Investitionen erkauft werden. Man benötigt RDMA-fähige Netzwerkkarten und eine dementsprechende Fabric, also Infiniband oder Ethernet mit Data Center Bridging. Microsoft hat hier insbesondere seine eigene Virtualisierungs-Lösung Hyper-V im Visier. Unzählige Virtuelle Maschinen laufen in großen Hyper-V-Clustern und benötigen entsprechende Performance auch im Backend Storage. Mehr noch, es wird eine Speicher-Virtualisierung benötigt, die einem virtuellen Server „seinen" Speicher bereitstellt, unabhängig vom physischen Ort, an dem der virtuelle Server gerade läuft. Verschiebt man den virtuellen Server auf einen anderen Hyper-V Host, so muss insbesondere der Data Store (also die VHDX-Datei) dort verfügbar sein, ohne langwierig Gigabytes Daten verschieben zu müssen. Hierfür bietet SMB eine ideale Plattform. Jeder Hyper-V Host eines Clusters sieht dieselben SMB Shares. Die SMB Server müssen zu diesem Zweck höchst performant und skalierbar sein.

Microsoft Scale Out File Server

Dementsprechend heißt das neue Produkt „Scale Out File Server" (SOFS). Genau genommen handelt es sich dabei um eine zusätzliche Server-Rolle des Windows Server 2012 R2. Die Idee des SOFS ist einfach: Mehrere SMB Server stellen parallel dieselben Shares bereit. Zu diesem Zweck teilen sie sich dieselben Festplatten in einem „Shared Storage". Das können herkömmliche logische Laufwerke (Logical Units, LUNs) in einem Fibre Channel SAN sein. Das können aber auch Festplatten mit Se-

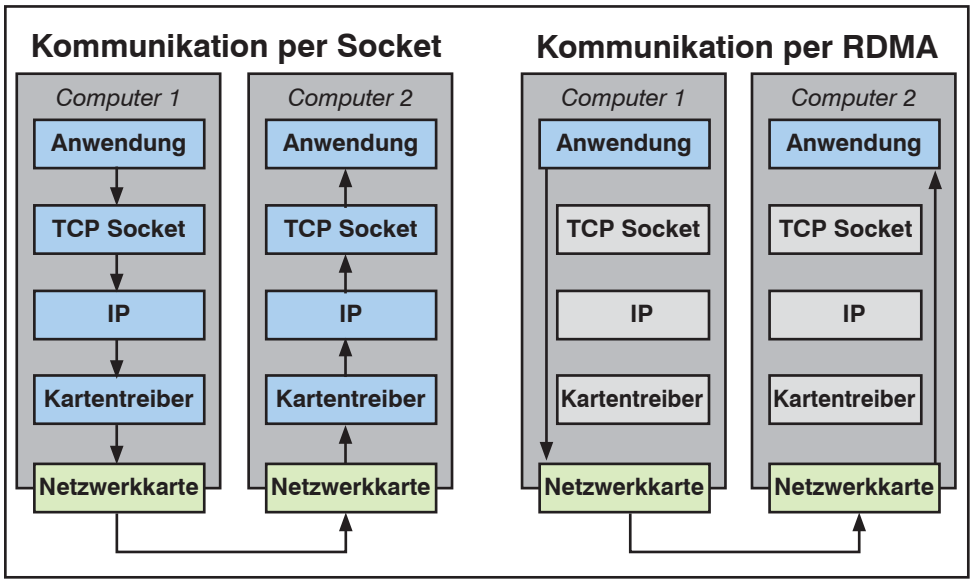

Abbildung 2: Anwendungs-Kommunikation ohne und mit RDMA

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

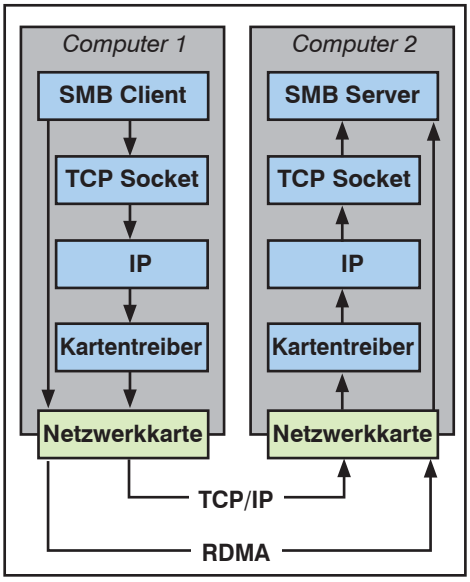

Abbildung 3: Zur Wirkungsweise von SMB Direct

rial Attached SCSI (SAS) in preiswerten Disk Shelves sein, die man gerne als „Just a Bunch of Disks“ (JBOD) bezeichnet. Bis zu 8 Server lassen sich auf diese Weise zu einem SOFS kombinieren.

Stattet man alle Server mit RDMA-fähigen Netzwerkkarten aus, ist für die hohe Performance gesorgt. Die Hyper-V Hosts können dann SMB Direct nutzen. Baut man darüber hinaus zwei unabhängige Ethernet- bzw. IP-Netze auf (vgl. Abbildung 4), wird dank SMB Multichannel die erforderliche Verfügbarkeit auch bei Netzwerk-Störungen oder Adapter-Fehlern sichergestellt.

Zusammenfassung

Remote Direct Memory Access (RDMA) über Infiniband und dessen Ethernet-basierende Variante RoCE sind Techniken, die hohe Bitraten mit geringen Latenzen miteinander verbinden. Diese Techniken erfahren nun Unterstützung durch die aktuellen Server-Produkte von Microsoft. SMB Multichannel und SMB Direct bieten einerseits höhere Performance als mit den herkömmlichen TCP/IP-basierten Zugriffen erreicht werden kann. Andererseits wird hohe Verfügbarkeit durch Redundanzmechanismen erzielt, die unabhängig von

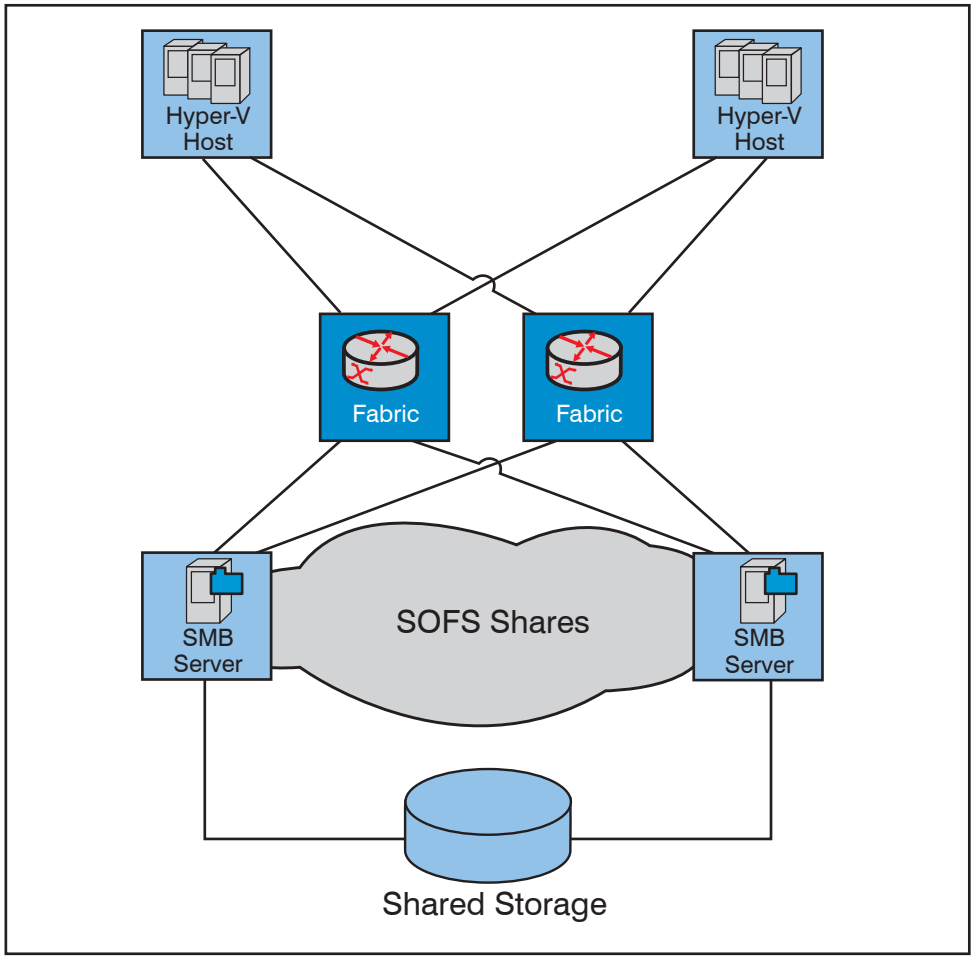

Abbildung 4: Scale Out File Server (SOFS)

speziellen Techniken der Switch- und Adapter-Hersteller sind (sieht man einmal von RDMA ab, das natürlich eine „spezielle Technik“ der Adapter-Hersteller ist).

Microsoft kombiniert diese Techniken zu einem neuen Produkt, dem Scale Out File Server (SOFS) als Teil von Windows Server 2012 R2. Mit dem SOFS bekommt der Kunde eine Lösung zur Realisierung performanten virtuellen Speichers, der insbesondere im Zusammenhang mit der Server-Virtualisierung (Hyper-V) Vorteile bietet. Im SOFS mutiert der Windows Server quasi zu einem Storage Controller.

Letztlich versucht Microsoft mit dem SOFS in den Markt der Speicher-Systeme vorzudringen und den „Platzhirschen“ der Branche Marktanteile abzunehmen. Für den Kunden ist damit die Hoffnung verbunden, zukünftig ein performantes Speichersystem preiswerter aufbauen zu können als es heute mit Fibre Channel SANs oder speziellen NAS Clustern möglich ist.

Ausblick

Wie oben bereits angedeutet, könnte die einstige Nischentechnologie Infiniband dank der Unterstützung von RDMA durch Microsoft eine gewisse Renaissance erleben. Jedoch funktioniert ein Infiniband HCA im Grunde nicht anders als eine Ethernet-Karte oder ein Fibre Channel HBA. Auf dem Adapter befindet sich ein Controller, der auf

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

der einen Seite das entsprechende Protokoll bedient. Der Controller ist also dazu in der Lage, die Daten zu Paketen zusammenzuschnüren und mit den erforderlichen Header-Informationen zu versehen. Auf der anderen Seite verfügt dieser Controller über eine Schnittstelle zum Computer, in den der Adapter hineingesteckt wurde. Das ist heute im Allgemeinen der „Peripheral Component Interconnect Express“, kurz PCIe.

Was ist PCIe eigentlich? Wie wir weiter oben gesehen haben, ist er die Verbindung an das Bus-System des Computers. Es lassen sich also Speicherstellen adressieren (Adressbus) und Daten von diesen Speicherstellen lesen bzw. darauf schreiben (Datenbus). Außerdem hat das Bus-System einige zusätzliche Funktionen wie die Möglichkeit Interrupts zu generieren oder eben DMA-Zugriff zu initiieren.

Frühe Bus-Systeme hatten für jeden der genannten Zwecke eigene Leitungen. Ein 32-Bit-Datenbus brauchte also 32 Drähte. Beim PCIe hat man sich stattdessen für eine serielle Übertragung entschieden. Adressen und Daten werden nacheinander übertragen und in Paketen zusammengefasst. Auch Interrupts und alle anderen Funktionen sind beim PCIe spezielle Pakete. Übertragen werden die Pakete mit vorgegebener Bitrate über Leitungen, die nun „Lanes“ genannt werden. Bis zu 16 Lanes lassen sich über PCIe parallel bedienen, je nachdem wie die Steckplätze ausgestattet sind. Beim PCIe in der Version 4.0 leistet jede Lane 2 GByte/s, umgerechnet also 16 Gbit/s. Ein voll ausgestatteter Steckplatz schafft also sagenhafte 256 Gbit/s.

Die Lanes der PCIe-Steckplätze werden im Computer über spezielle Bausteine mit den übrigen Komponenten verbunden, insbesondere also mit Prozessor und Speicher. Bei diesen Bausteinen – als „Southbridge“ und „Northbridge“ bezeichnet – handelt es sich genau genommen um Switches. Nur vermitteln diese nicht Ethernet oder Fibre Channel Frames sondern PCIe-Pakete. Was läge also näher als das Verlängern der PCIe Lanes aus dem Computer hinaus, etwa über Glasfasern. Verbände man auf diese Weise zwei Computer miteinander,

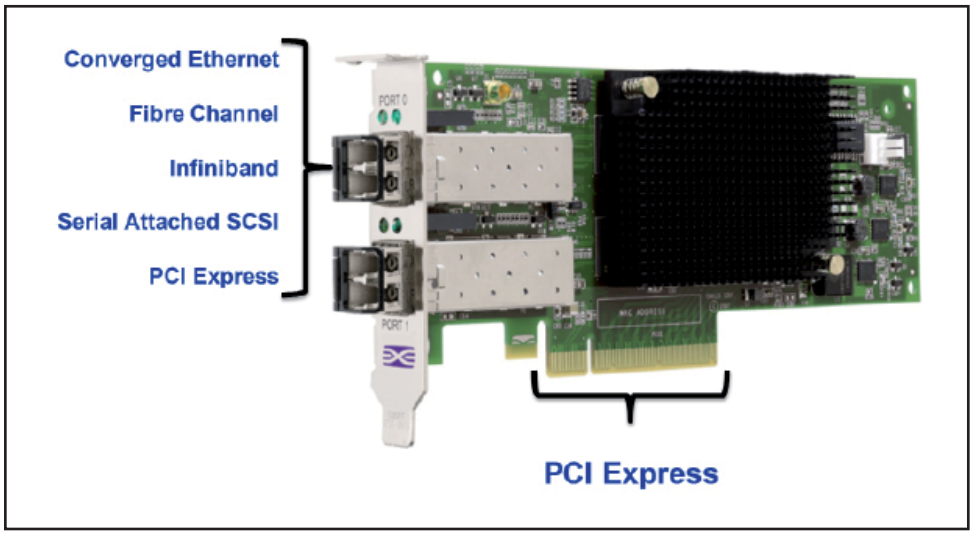

Abbildung 5: Schnittstellen eines Server-Adapters

ließen sich Speicher und Peripherie beider Computer direkt adressieren. Ein Prozessor wäre dann in der Lage, Speicher auf dem anderen Computer direkt zu beschreiben. Im Vergleich zu RDMA hätte man noch einmal Rechenzeit eingespart.

In der Tat hat man bereits in 2011 eine PCIe-Übertagung per Glasfaser demonstriert, immerhin mit 64 Gbit/s. Und auch erste PCIe Switches sind bereits gesichtet worden. Es bleibt abzuwarten, ob eines Tages einer der großen Hersteller diese Technik unterstützt und damit der Marktakzeptanz eine Chance gibt.

Verweise

[1] Spezifikation von RoCE: **https://cw.infinibandta.org/document/dl/7148**

[2] Spezifikation von RoCEv2: **https://cw.infinibandta.org/document/dl/7781**

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

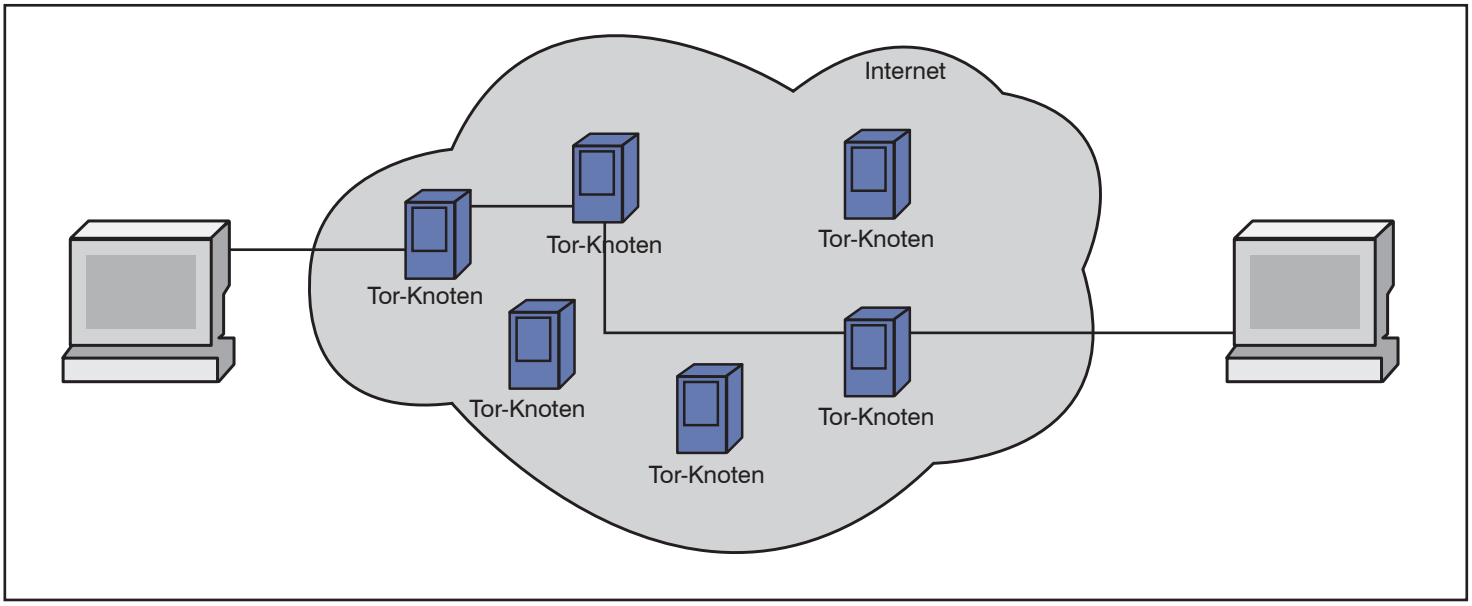

Abbildung 3: Zufällige Wahl einer Route über Tor-Knoten

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals

Der Netzwerk Insider

Systematische Weiterbildung für Netzwerk- und IT-Professionals