ComConsult Research

Der Netzwerk Insider

Mai 15

Systematische Weiterbildung für Netzwerk- und IT-Professionals

Schwerpunktthema

IT-Sicherheit im Zeitalter der Cyberangriffe

von Dr. Melanie Winkler

Jedem Nutzer elektronischer Daten sollte mittlerweile klar geworden sein, dass alle Informationen, welche über ein öffentliches Netz übertragen werden, generell durch Unbefugte abgegriffen werden können. Dies gilt umso mehr seit den Enthüllungen durch Edward Snowden und die dadurch öffentlich bekannt gewordenen Angriffe auf vertrauliche Daten unterschiedlicher Art durch große Hackerangriffe. Daher wird natürlich schon seit einiger Zeit, zum Beispiel durch das BSI, die Empfehlung ausgesprochen, vertrauliche Daten ausschließlich verschlüsselt zu übertragen und zu speichern.

Im Rahmen der Enthüllungen durch Edward Snowden ist jedoch auch bekannt geworden, dass zumindest einige große Hackergruppen - häufig mit staatlicher Unterstützung - über weitere Methoden verfügen, an Informationen zu gelangen, welche für sie von Interesse sind. Es stellt sich Unternehmen und Privatanwender daher immer mehr die Frage, wie sicher ihre Daten wirklich noch sind. Lohnt es sich beispielsweise als Endanwender überhaupt noch Daten zu verschlüsseln oder können diese nicht sowieso mit vergleichbar geringem Aufwand entschlüsselt und gelesen werden?

weiter auf Seite 5

Zweitthema

High Performance Storage

von Dr. Joachim Wetzlar

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.

weiter auf Seite 15

Geleit

Bandbreite erschlägt alles: gilt das immer noch?

auf Seite 2

Aktuelle Veranstaltung

Sommerschule 2015 - Intensiv-Update auf den neuesten Stand der Netzwerktechnik

ab Seite 4

Standpunkt

Monitoring von Voice- und Video-Strömen!

auf Seite 13

Aktuelle Sonderveranstaltung

Voice und Video im WAN

ab Seite 14

Zum Geleit

Bandbreite erschlägt alles: gilt das immer noch?

Wir hatten in der Vergangenheit viele Diskussionen über die Notwendigkeit von Quality of Service in Netzwerken. Und meine Haltung war, dass ein solide geplantes und aufgebautes Netzwerk mit ausreichend Bandbreiten-Reserven keine QoS benötigt. Das wurde auch durch diverse Messungen und Tests in zum Teil sehr großen Kundennetzen immer wieder bestätigt. Insbesondere das Argument, dass VoIP QoS im LAN benötigt, hat sich aus meiner Sicht als nicht belegbar herausgestellt (ich will die Diskussion hier nicht wieder starten, mir ist klar, dass es ganz verschiedene Motivationen für den Einsatz von QoS geben kann).

Nun haben wir ja mit moderner Netzwerk-Technologie sehr viel Bandbreite, seien es 10, 40 oder 100 Gigabit-Ethernet im LAN. Dann hat sich das Thema jetzt ja wohl erledigt, oder?

Tatsächlich hat sich Netzwerk-Technologie zusammen mit den typischen Nutzungs-Situationen in den letzten Jahren verändert. Es gibt dabei mindestens drei Bereiche, die einer besonderen Beachtung bedürfen:

- Der Ost-West-Verkehr zwischen virtuellen Servern und zentralen Speichersystemen stellt hohe und vor allem schwer vorhersehbare Anforderungen an Netzwerke. Auch wenn FCoE nicht in dem prognostizierten Ausmaß gekommen ist, haben wir mit iSCSI und NFS oder ähnlichen Verfahren Datenströme, die ein erhebliches Ausmaß annehmen können. Und was besonders wichtig ist: das kann sich sehr schnell ändern, je nachdem was auf der Speicherseite passiert.

- Der Wunsch, automatisch Anwendungen aus einem Service-Katalog wie aus der Public Cloud bekannt auch im lokalen Umfeld starten zu können, verändert unsere Netzwerk-Welt. Persönlich bin ich zwar sehr skeptisch, ob wir diesem Wunsch wirklich folgen sollten, aber wir reden hier auch über ein politisches Thema, das nicht immer nach rein technischen Kriterien entschieden wird. Die Umsetzung dieses Wunsches erfordert programmierbare und vor allem virtualisierte Netzwerke im Rahmen einer Open Stack-Realisierung.

- Im Umfeld dynamischer virtueller Architekturen gibt es ähnlich dem Konstrukt,

das wir bei der automatischen Inbetriebnahme von Anwendungen haben, den Bedarf, die virtuellen Elemente, die zu einem Service gehören, durch ein abstraktes Netzwerk miteinander zu verbinden. Dieses Netzwerk integriert auch Netzwerk-Elemente wie Switches, Router, Load-Balancer als Software-Elemente. Der Grund für diese Architektur liegt in der Skalierbarkeit durch Parallelisierung von virtuellen Maschinen und der damit verbundenen Beweglichkeit der einzelnen virtuellen Elemente. Die damit erreichte Ortsneutralität hat zudem eine ganze Reihe weiterer Vorteile. Da dabei u.a. Layer-2-Verbindungen über Layer-3-Grenzen hinweg geschaltet werden müssen und Technologien wie LISP hier ein Overkill sind (auch wenn ich persönlich ein LISP-Anhänger bin), brauchen wir Technologien, die Netzwerk-neutral einen solchen Service-Verbund umsetzen.

Was hat das nun mit QoS zu tun? Bitte warten Sie noch einen weiteren Gedankengang, bevor ich darauf zurück komme.

Aus meiner Sicht brauchen wir für die Zukunft der LANs Edge/Core-Technologien, die die Intelligenz mit den Service-Definitionen am Rand umsetzen und den Kern relativ einfach gestalten. Ich habe große Befürchtungen, dass Technologien, die es erfordern Service-Anforderungen in jedem Switch zu programmieren zu einem Komplexität-Desaster führen. Aus diesem Grund bin ich auch von CISCO ACI nicht überzeugt (und bitte, kann endlich mal jemand das CLI-Interface aus solchen Lösungen entfernen? In welchem Jahrhundert leben wir denn eigentlich?).

Edge/Core-Lösungen sind aber direkt mit zwei Fragen verbunden:

1. Ist der Edge immer Software? Also immer eine Software-Switch in einem Hypervisor?

2. Was gilt eigentlich im Core, wie kann oder muss dieser gestaltet werden?

Sonderveranstaltung

Voice und Video im WAN - 22.06.15 in Köln

Wie kann die Übertragung von Sprache und Video im WAN optimiert werden ohne andere Anwendungen zu gefährden? Wie gehen wir mit einem immer größeren Anteil und vor allem der Integration von Video in wesentliche Geschäftsprozesse um? Neue Kollaborations-Lösungen erhöhen zudem den Druck auf die Infrastrukturen. Verkehrslasten werden dabei immer dynamischer und einfache statische Regeln wie traditionelles QoS stoßen an ihre Grenzen.

Die Sonderveranstaltung beleuchtet den Status Quo, die Zukunftsaussichten, mögliche Optionen und eventuellen Investitionsbedarf für den sicheren Betrieb aller Anwendungen.

Referenten: Dipl.-Ing. Martin Egerter, Dipl.-Math. Leonie Herden, Dipl.-Ing. Dominik Zöller
Preis: € 990,-- netto

Buchen Sie über unsere Web-Seite

www.comconsult-akademie.de

Schwerpunktthema

IT-Sicherheit im Zeitalter der Cyberangriffe

Fortsetzung von Seite 1

Dr. Melanie Winkler ist als Beraterin bei der ComConsult Beratung und Planung GmbH in dem Bereich IT-Sicherheit tätig. Dort beschäftigt sie sich besonders mit Sicherheitskonzeptionen nach ISO 27001 und BSI Grundschutz und deren Umsetzung.

In diesem Artikel werden die wesentlichen der aktuell bekannten Gefährdungen dargestellt und es wird aufgezeigt, wie man sich zumindest gegen einige dieser Bedrohungen schützen kann und welche Grenzen hier aktuell bestehen.

1.1 Zugriff auf Daten

1.1.1 Zugriff durch Abfangen von Daten
In vielen Fällen durchlaufen Daten zu einem Zeitpunkt während ihrer Übertragung ein öffentliches Netz und können dann von einem Angreifer an einem Punkt der Verbindung abgefangen und ausgewertet werden. Werden die Informationen unverschlüsselt übertragen, so können diese direkt vom Angreifer abgehört oder umgeleitet, gespeichert und genutzt werden. Ein Abfangen übertragener Daten ist in allen Bereichen von Kommunikationsverbindungen möglich.

Beispielsweise werden bei jedem normalen Telefonat Daten über ein öffentliches Netz übertragen. Dabei handelt es sich hauptsächlich um Sprachdaten, welche zwischen den Teilnehmern ausgetauscht werden sollen. Bei Telefonaten werden die Sprachdaten normalerweise (zumindest auf Teilen der Gesamtübertragungsstrecke) unverschlüsselt übertragen. Das bedeutet, dass ein Angreifer, welcher die Leitung abhört oder ein Gespräch mitschneidet, dieses unmittelbar verstehen kann.

Auch für Informationen, welche über das Internet verschickt werden, ist es möglich, diese mit geeigneten Werkzeugen abzufangen. Dies ist beispielsweise gegeben, wenn per E-Mail kommuniziert wird oder im Internet Daten ausgetauscht werden. Es werden aber auch Daten übers Internet übertragen, wenn ein Mitarbeiter von außerhalb des Büros auf ein Firmennetz zugreift (VPN Zugriff) oder einen Rechner von Extern administriert (Zugriff über SSH).

Darüber hinaus kann auch auf Daten zugegriffen werden, welche lediglich auf dem System gespeichert sind, aber nicht über das Netz übertragen werden. Ein solcher Zugriff ist dann durchführbar, wenn über Schadsoftware, Sicherheitslücken oder sogenannte Backdoors auf ein System und die darauf befindlichen Daten zugegriffen werden kann. Außerdem existiert auch Schadsoftware, welche Daten des Systems direkt an einen Angreifer überträgt. Ein Beispiel für eine solche Schadsoftware ist „EquationDrug“. Mit Hilfe dieser Software können die Personen, welche die Software auf den Rechner eingeschleust haben, Daten vom Rechner abziehen. Sie können so beispielsweise Informationen darüber erhalten, was auf der Tastatur getippt wird oder welche Seiten im Browser geöffnet wurden.

Die hier beschriebenen Zugriffe auf Informationen sind jedoch bei immer weiter steigenden Datenmengen sehr mühsam. Die Größenordnung sowohl der gespeicherten Daten, als auch der übertragenen Daten steigen von Tag zu Tag. Daher sind derartige Vorgehensweisen des Abfangens sehr aufwendig und erfordern einen hohen Einsatz an Zeit und Ressourcen zur Speicherung und Auswertung der abgefangenen Daten.

1.1.2 Gezielte Angriffe auf Systeme
In der Vergangenheit ist es bereits vorgekommen, dass Certificate Authorities, welche die privaten SSL Schlüssel vieler bekannter Hersteller verwalten, durch externe Angreifer attackiert und dabei private Schlüssel gestohlen wurden. Somit ist jede SSL Kommunikation, welche mit den zugehörigen öffentlichen Schlüsseln kodiert wurde, ohne viel Aufwand mit den gestohlenen privaten Schlüssel in Echtzeit zu entziffern. Eine andere Methode verschlüsselte Kommunikation zu entschlüsseln benötigt Eingriffe auf die kommunizierenden Komponenten.

Angriffe auf solche Komponenten zielen beispielsweise darauf ab, die Konfigurationen der Verschlüsselung zu modifizieren. Dabei kann eine Verschlüsselungsart eingestellt werden, die der Angreifer ohne privaten Schlüssel entschlüsseln kann. Solche Modifikationen können an den Konfigurationen der E-Mail Clients, Router oder VPN-Clients vorgenommen werden. Daher gehören Benachrichtigungen und Logging im Falle von Konfigurationsänderungen an Software und Hardware zu wichtigen Sicherheitsfeatures.

Die Angriffe auf Endpunkte sind jedoch wesentlich aufwendiger als Eingriffe in die Kommunikation. Viele der bekannten Angriffe auf Endpunkte benötigen physikalischen Zugriff zum Endpunkt (z.B. USB-Stick in Laptop einschieben). Solche Angriffe können jedoch auch über Backdoors in auf dem Endpunkt installierten Programmen initiiert werden, z.B. durch böswillige E-Mail Anhänge und modifizierte Update-Pakete für das Betriebssystem oder einzelne Programme.

Eine weitere Methode gezielt auf Daten zuzugreifen ist heutzutage weit verbrei-

tet. Sie wird „Advanced Persistent Threat“ (APT) genannt. Ein APT besteht meist aus mehreren aufeinander abgestimmten Angriffsformen und verfolgt ein festumrissenes Angriffsziel (z.B. Industriespionage). Ein APT lässt sich in 5 verschiedene Phasen unterteilen:

1. Auskundschaften / Zielerfassung:
   Bevor ein erster Angriff auf das System erfolgt, informiert sich der Angreifer möglichst genau über den Aufbewahrungsort (z.B. Server, Rechenzentrum, Laufwerk, …) der Daten. Hierzu kann er sowohl auf öffentlich zugängliche Informationen zugreifen (z.B. eine unsicher konfigurierte Fehlermeldungswebseite eines Apache Servers) als auch gezielt Informationen erfragen (Social Engineering, Spear Phishing, …).

2. Eindringen und Erstinfektion:
   Nachdem der Angreifer die gesuchten Informationen grob lokalisiert hat, versucht er gezielt an diese heranzukommen. Ziel dieser Phase ist es als Vorbereitung für einen Zugriff auf die gewünschten Daten, unentdeckt in das System einzudringen. Dies erfolgt beispielsweise darüber, dass im ersten Schritt auf dem Zielsystem installierter Schadcode gezielt weitere Software auf dem System installiert. Bei dieser handelt es sich häufig um einen speziellen Trojaner, der oft ironisch als Remote Administration Tool (RAT) bezeichnet wird und welcher mit dem Command and Control Server (C&C) kommuniziert.

3. Folgeinfektion (en):
   Sollte der Angreifer nach der Erstinfektion noch nicht auf alle für ihn wünschenswerten Daten zugreifen können, dann wird er von seinem ersten Einfallstor aus weitere Systeme infizieren. Dieser Schritt ist häufig wesentlich leichter als die Phase der Erstinfektion, da die Sicherheit nach Außen bei den meisten Systemen wesentlich besser ist als die interne Sicherheit. Dieser Schritt ist auch notwendig, wenn im ersten Schritt zwar eine Infektion eines Systems im Zielnetzwerk (z.B. ein normaler PC), aber keine direkte Infektion des Zielsystems (z.B. Administrations-PC, Server…) möglich war.

4. Datendiebstahl / Sabotage:
   In dieser Phase beginnt die Übertragung interessanter Daten in großen Mengen. Bis zu dieser Phase hat der Eindringling noch keine Daten abgegriffen und keine großen Datenmengen kopiert, um nicht aufzufallen. Diese Phase birgt das größte Risiko entdeckt zu werden, da hier Anomalien in Bezug auf den Datenverkehr und die durchzuführenden Aktionen (Datenbank-Dump, Kopieren von Dateien, …) gemessen werden können. Nach Abschluss dieser Phase hat der Angreifer die gewünschten Daten abgegriffen und der Schaden ist vollständig verursacht.

5. Verwischen von Spuren:
   Sollte der Angriff nicht aufgefallen und unmittelbar blockiert worden sein, so hat der Angreifer noch die Möglichkeit Spuren zu verwischen. Hierzu gehört das Löschen von Log-Dateien, Verbindungsinformationen und des Terminalverlaufs. Außerdem löscht sich die Schadsoftware selbst oder sie löscht einen Teil von sich und wird inaktiv. Wird diese Phase ebenfalls erfolgreich abgeschlossen, so fällt es den Geschädigten wesentlich schwerer den Angreifer und das Ausmaß des Schadens zu bemessen.

1.1.3 Zugriff auf rechtlichem Wege

Ein wesentlich angenehmerer und weniger aufwendigerer Weg an Informationen zu gelangen, ist mit der Unterstützung der Informationshalter. Wenn also Hersteller dazu gebracht werden können, Zugang zu ihrer Software / Firmware bereitzustellen, dann spart dies Zeit gegenüber der Informationsbeschaffung durch Angreifer. Außerdem können die Grenzen, die durch andere Beschaffungsmöglichkeiten vorhanden sind, gegebenenfalls aufgehoben werden. So muss beispielsweise bei einem Abtransport von großen Datenmengen durch die Angreifer nicht vermieden werden, dass der Datenfluss auf Ebene des Netzwerkverkehrs auffällt.

Eine Grundlage für einen solchen rechtlich legitimen Zugriff auf sensible Informationen bietet den Geheimdiensten der USA der Foreign Intelligence Surveillance Act (FISA). Dieses Gesetz regelt insbesondere Fragen zur Auslandsaufklärung und Spionageabwehr. Es beschreibt unter anderem, dass Informationen von verdächtigen Personen, die bestimmten Kriterien entsprechen, für die Regierung der Vereinigten Staaten von Amerika zugänglich gemacht werden müssen. Zu diesen Kriterien gehören beispielsweise, dass die Person kein US-Bürger ist und sich nicht auf dem Territorium der USA aufhält. Jedoch können unter bestimmten weiteren Auflagen auch US-Bürger mit dem FISA durchleuchtet werden.

Laut amerikanischer Rechtsprechung unterliegen dem FISA alle auf amerikanischem Boden agierenden Unternehmen. Hierzu gehören unter anderem Microsoft, Yahoo, Google, Facebook, AOL, Skype, YouTube und Apple. Somit kann beispielsweise bei einem terroristischen Verdacht auf rechtlichem Wege Einsicht in die E-Mail-Konten von verdächtigen Personen erlangt werden.

Um solche Informationen von den betroffenen und interessanten Unternehmen zu erhalten, gab es laut den Dokumenten von Edward Snowden mehrere Treffen zwischen Regierungsvertretern (NSA, CIA, FBA, …) und Vertretern der Unternehmen (inklusive CEOs). Diese Treffen wurden unter der Enduring Security Framework (ESF) Initiative abgehalten. Dabei wurden angeblich sowohl die Art der benötigten Informationen, als auch die Übermittlungswege für Daten besprochen.

Aber nicht nur amerikanische Geheimdienste und damit auch die amerikanische Regierung können auf diesem Weg an Informationen gelangen. Durch Weitergabe können die so erlangten Informationen auch von anderen Geheimdiensten und Regierungen genutzt werden, welche mit der USA zusammenarbeiten. Laut den Dokumenten von Edward Snowden muss mindestens mit einer Weitergabe an die Geheimdienste und Regierungen verbündeter Staaten der USA gerechnet werden. Allerdings ist auch zu beachten, dass es sich hierbei um einen sehr speziellen Weg der Informationsgewinnung handelt. Auf diese Art Informationen zu erlangen können zwar einige der Geheimdienste zurückgreifen, jedoch außerhalb dieser Kreise ist dies nicht möglich. Maßnahmen zu ergreifen, welche gegen diese Art von Angriff nicht schützen können, sind daher trotzdem sinnvoll, um sich vor sonstigen Hackerangriffen zu schützen.

Die Art der Informationen, welche über diesen Weg erlangt werden können, hängen vom jeweiligen Anwendungsfall und von der ausgespähten Applikation ab. So wird in manchen Veröffentlichungen behauptet, die abgegriffenen Informationen würden sich auf Metadaten beschränken. Diese Datensätze beinhalten beispielsweise Informationen darüber, wer mit wem, wann und wie lange kommuniziert hat. Andere Veröffentlichungen beschreiben jedoch die Übertragung vollständiger Datensätze, also auch die Inhalte von E-Mails, Facebook-Profilen, Skype-Kommunikationen und Kalendereinträgen.

Die Wege über welche ein Angreifer an Informationen gelangen kann, hängen von den Applikationen und Szenarien ab. Im Folgenden geben wir einen Überblick.

1.2 Übertragungswege von Daten

Für die Übertagungswege von Informationen an unberechtigte Dritte gibt es unterschiedliche Möglichkeiten. Zum

Zweitthema

High Performance Storage

Fortsetzung von Seite 1

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.

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 Blockbasierten 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 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 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 ver-

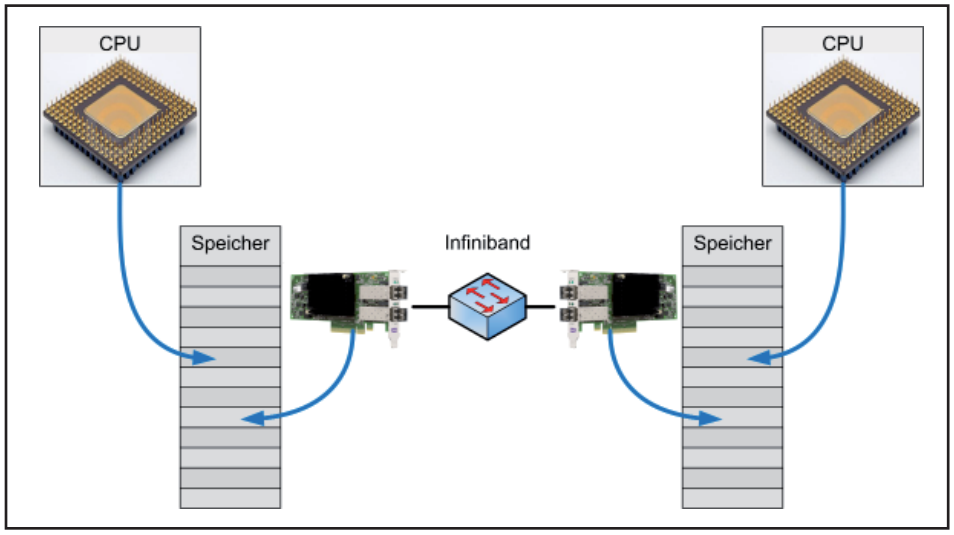

Abbildung 1: Zum Prinzip von RDMA