Skalierung von Ethernet in Rechenzentren

Moderne Netze

In großen Rechenzentren muss Ethernet heute flexibler arbeiten können, als es der Standard bislang vorgesehen hat. Aus diesem Grund ist es logisch, dass Konzepte aus anderen Netzsystemen – etwa Infiniband – auch in ein modernes Ethernet Einzug halten.

Die Architektur von Rechenzentren entwickelt sich ständig weiter. In mehreren Schlüsselbereichen sind derzeit wesentliche Veränderungen zu beobachten:

- Konsolidierung: Größere, von mehreren Kunden gleichzeitig benutzte Rechenzentren ersetzen viele kleinere,
- Fokussierung: Anwendungen und Business-Services bestimmen die IT-Prozesse, die zunehmend automatisiert und weniger manuell gesteuert sind,
- Virtualisierung: Server, I/O, Storage, Netzwerke und Applikationen arbeiten virtualisiert und von der Hardware entkoppelt, und
- Konvergenz: Netzwerk- und Storagebezogene Daten sowie die Interprozesskommunikation mehrerer Anwendungen bewegen sich auf ein und derselben Verbindung.

Diese Trends haben einen großen Einfluss auf die Architektur von Switching Fabrics. Diese müssen nun größere Layer-2-Netzwerke (L2) unterstützen, da die Virtualisierung und Mobilität von Servern, neue Storage-Protokolle und die so genannte Interprozesskommunikation mit kürzesten Latenzzeiten in der selben L2-Domäne angesiedelt sind. Eine neue RZ-Infrastruktur muss zudem den Grad der Management-Komplexität reduzieren und von Grund auf das Thema Virtualisierung adressieren. Die Data Center Bridging Group des IEEE beschäftigt sich mit der Erweiterung von Ethernet, um die genannten Trends zu unterstützen. Viele von Infiniband bekannten Techniken wurden übernommen, wie etwa Class Isolation, Low Latency, I/O- und Switch-Virtualisierung, außerdem Lossless Traffic Flows, Congestion Control und das Multipath L2 Routing. Diese neuen Techniken sind unter dem Begriff Converged Enhanced Ethernet (CEE) oder Data Center Ethernet (DCE) zusammengefasst.

Um die Effizienz von RZs weiter zu erhöhen, beschäftigen sich viele Unternehmen mit der Frage der Wirtschaftlichkeit. Mehrere kleinere RZs lassen sich zu größeren Einheiten konsolidieren. Oft mag es auch wirtschaftlicher sein, eigene RZs zu kommerziellen, so genannten Cloud Providern hin zu verlagern. Solche Unternehmen betreiben mehrere virtuelle RZs an einem physischen Standort. Daher wird die Zahl der öffentlichen und privaten Clouds vermutlich weiterhin rasant zunehmen.

Eine Option für größere RZ-Einheiten ist es, die Server-Dichte zu erhöhen. Neueste Entwicklungen näheren sich der Marke von 100 Systemen mit mehr als 1.000 CPUs pro Rack. Um so wichtiger sind dadurch Themen wie Stromversorgung, Kühlung, Switch-Verkabelung und Port-Dichte. Bereits ein bis zwei Racks mit 10GbE-Servern können mehr Anschlüsse haben als die größten heute im Markt verfügbaren 10GbE-Switches bereitstellen. Die Schlüsseltechniken auf dem Weg zu einer höheren RZ-Effizienz sind Virtualisierung und Automatisierung. Damit lassen sich typische wiederkehrende Aufgaben – wie die Bereitstellung neuer IT-Services, die Durchführung von Wartungsarbeiten und die Lastverteilung – zwischen unterschiedlichen Hardwareplattformen weitgehend automatisieren. Während die RZs größer und dichter werden sowie zunehmend virtualisiert und automatisiert arbeiten, wächst die Last der darunter liegenden Switching-Einheiten drastisch an. Folgende Trends sind zu beobachten:

- Die zunehmende gemeinsame Nutzung externer Speichereinheiten (NAS oder SAN) verursacht ein erhebliches Verkehrsaufkommen innerhalb des RZs, wobei die Anforderungen an Zuverlässigkeit, Service-Qualität und Durchsatz eher steigen,
- die Server-Virtualisierung bedingt höhere Kapazitäten über einen physikalischen Knoten, was wiederum mehr Leistung bei den beteiligten NICs und Netzwerken mit sich bringt, und
- die Mobilität von Servern und Applikationen verursacht mehr Verkehrsaufkommen zwischen unterschiedlichen physikalischen Segmenten.

Zusätzliche Last entsteht durch die Migration von Servern und den damit verbundenen Umzug von virtuellen Servern von einem physischen System auf ein anderes. Der Bedarf an größerer Effizienz führt zu einer starken Verbreitung der Cluster-Technik. Bespiele von verbundenen Servern, die eine bestimmte logische Funktion erfüllen, finden sich bei Web-Clustern, Datenbank-Clustern oder Dateisystem-Clustern. Alle Cluster-Techniken nutzen in hohem Maße Interprozesskommunikation

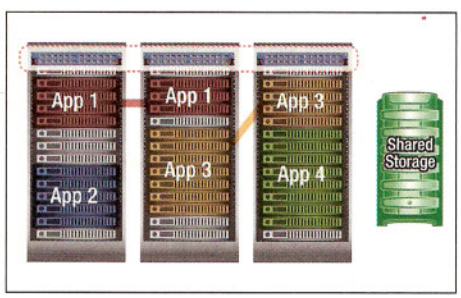

Bild 1. RZ-Architektur der nächsten Generation.

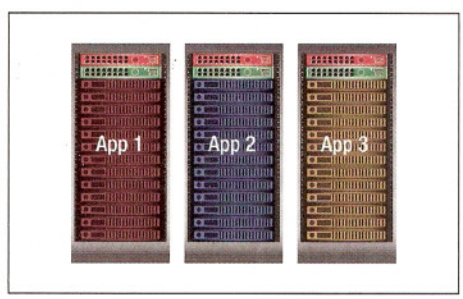

Bild 2. Herkömmliche RZ-Architektur.

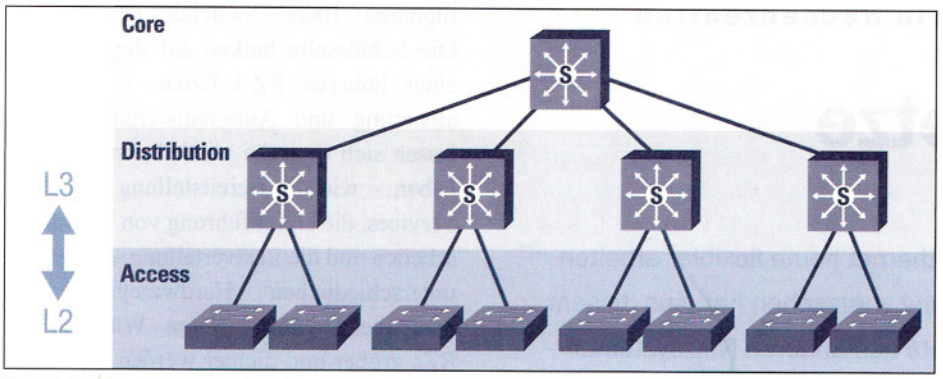

Bild 3. Herkömmliches dreistufiges Netzwerkdesign.

und erzeugen ein erhebliches Verkehrsaufkommen innerhalb des Clusters. Dadurch wachsen die Anforderungen an die beteiligten Switches bezüglich Latenzzeit, verlustloses und vorhersehbares Verhalten sowie den Durchsatz.

Dieser Verkehr ist auf dieselbe L2-Netzwerkdomäne limitiert. Eine virtuelle Maschine migriert mit ihrer IP-Adresse und kann nicht ohne Weiteres in eine anderes IP-Segment wandern. Hinzu kommt, dass einige Storage- und Messaging-Protokolle (wie beispielsweise FcoE, RoCE oder PXE) auf einen L2-Bereich begrenzt sind. Damit wird klar, dass neuere RZ-Netzwerke größere L2-Strukturen unterstützen müssen als in der Vergangenheit. Eine L3-Kommunikation zwischen den Racks ist nicht mehr möglich.

Zusammenfassend unterscheiden sich die Netzwerke zukünftiger RZs von heutigen durch folgende Eigenschaften:

- Anschlüsse für eine sehr große Anzahl von Rechnerknoten bei hoher Dichte,
- geringere Leistungsaufnahme,
- höhere Bandbreite pro Port,
- weniger Bandbreitenaggregation zwischen Endpunkten,
- geringere Latenzzeiten und vorhersehbares Verhalten,
- mehrere große L2-Segmente mit weniger Switching-Ebenen,
- Segmentisolierung und Class of Service (CoS) sowie
- Virtualisierung.

Neben den technischen Anforderungen spielen die Kosten für die Investition und den Betrieb eine entscheidende Rolle. Viele der heute existierenden Ethernet-Produkte wurden nicht mit den zuvor genannten Anforderungen im Fokus entwickelt und sind daher nur bedingt für den Einsatz in den neuen RZs geeignet. Eine neue Kategorie skalierbarer und RZ-optimierter Switches ist erforderlich.

Herkömmlichen RZs bestehen häufig aus so genanten Silos. Damit sind physikalische Racks mit Servern und anderen Hardwarekomponenten gemeint, denen in der Regel eine bestimmten Aufgabe (Applikation) zugeordnet ist. Ein Großteil des Verkehrsaufkommens (Transaktionen, Interprozesskommunikation) bleibt innerhalb des Racks. Lediglich ein kleinerer Teil des Gesamtverkehrs wird – aus der Sicht des Silos – mit der Außenwelt abgewickelt (Bild 2).

Die herkömmliche Architektur wie in Bild 3 ist typischerweise mit Top of Rack (TOR) Access Layer Switches realisiert, die mit wenigen Uplink-Ports mit der Distribution- oder der Core-Ebene verbunden waren. Nicht selten kommt es zu hohen Überbuchungen. In vielen Fällen sind den einzelnen Silos (Racks) eigene IP-Subnetze zugewiesen, wobei die darüber liegende Switching-Ebene (Distribution Layer) das L3/L4 Routing erledigt.

In herkömmlichen Umgebungen sind Core- und Distribution-Switches immer so ausgelegt, dass viele komplexe Netzwerkaufgaben (etwa große Routing- und Adresstabellen, Paketanalyse, Verschlüsselung etc.) erfüllt werden können. Daher haben diese Switches einen höheren Preis pro Port und einen höheren Leistungsbedarf als reine L2-Switches. Core Switches sind primär für den „nordwärts" gerichteten Verkehr (Client zu anderem Netzwerk oder Internet) zuständig, wo es weniger auf Interprozesskommunikation und Anbindung von Storage Einheiten ankommt. Kriterien wie geringe Latenzzeit, Vermeidung von Überbuchungen oder zuverlässig vorhersehbare Übermittlung sind weniger wichtig.

In Ethernet-Umgebungen gibt es traditionell eine hohe Überbuchungsrate (von 5:1 bis 10:1) zwischen Server und Uplink-Ports, insbesondere, wenn preisgünstige Access Switches zum Einsatz kommen. Die zunehmende Virtualisierung in Verbindung mit der Mobilität virtueller Systeme generiert mehr und mehr „Ost-West"-Verkehr zwischen Server-Racks und zwischen Servern und Storage-Subsystemen. Daraus folgt dann: Aggregation Switches mit überbuchten Uplink-Ports sind nicht mehr akzeptabel.

Skalierung von RZ Netzwerken

Bisherige Netzwerk-Switching- und -Softwarelösungen sind nicht besonders skalierbar, denn sie wurden nicht mit dieser Forderung entwickelt. Wachsende Netzwerke führen zu exponentiell wachsenden Kosten, Effizienzeinbrüchen und einer steigenden Management-Komplexität. Bild 4 zeigt den typischen Effekt größer werdender Netze: steigende Kosten bei sinkender Effizienz und komplexerem Management.

Wie zuvor beschrieben, sind herkömmliche Netze häufig mit kostengünstigen Blade oder TOR Switches und wesentlich teureren Aggregation Switches (mit höherem Leistungsbedarf) realisiert. Wachsende Netze erfordern zusätzliche Aggregati-

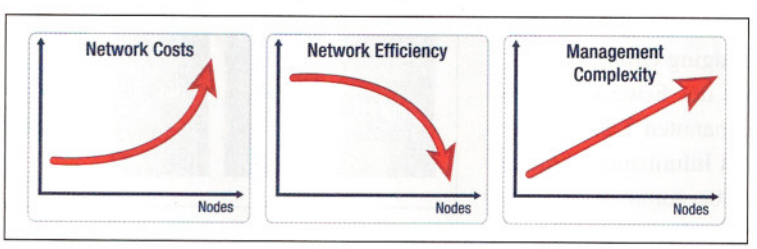

Bild 4. Ergebnisse der geringen Skalierbarkeit von Ethernet.

on Layer, um kleinere Netzwerksegmente zu verbinden. Dies führt dazu, dass für jeden Server-Port mehrere Ports für die Anbindung von Switches genutzt werden. Das Verhältnis zwischen Server-Ports und Netzwerk-Ports verschlechtert sich, so dass mehr und mehr teure Aggregation-Ports erforderlich sind.

Die Veränderungen in den RZs erfordern sowohl niedrigere Überbuchungsraten als auch den Übergang von 1GbE auf 10GbE, was die Kosten drastisch in die Höhe triebe, wenn das bisherige hierarchische Baummodell beibehalten würde. Bis vor Kurzem begegnete man den Kosten- und Skalierbarkeitsproblemen mit einer relativ hohen Überbuchung zwischen

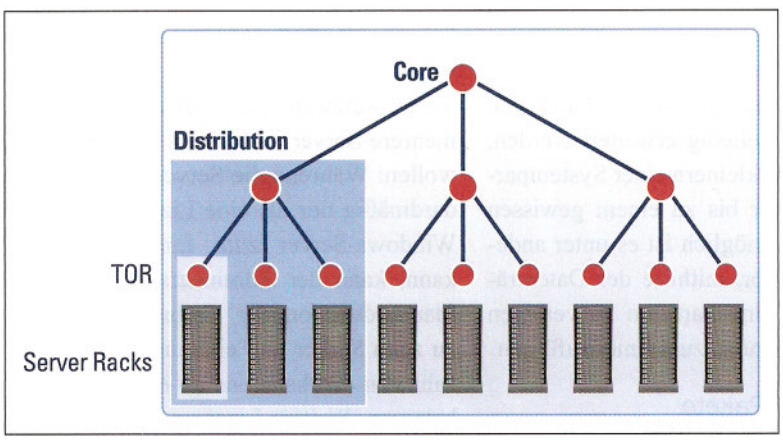

Bild 5. Das hierarchische Baummodell von Ethernet.

den verschiedenen Ebenen (Distribution, Core Layer). Beispielsweise sind an einen 48-Port-Switch mit vier Uplink-Ports 32 bis 36 Server angeschlossen. Der Core Switch sei mit 256 Ports voll bestückt. Damit können etwa 2.000 Server betrieben werden. Soll ein weiterer Ausbau erfolgen, muss eine weitere Hierarchieebene (Bild 5) entstehen.

Derzeitige Netzwerke sind nicht nur teuer zu skalieren, sie verlieren dabei auch an Effizienz. Die hierarchische Struktur schafft Engpässe, sobald Verkehrsströme ein Rack verlassen und auf dem Weg zu anderen Racks mehrere Aggregation Switches durchlaufen. Dies führt zu einer wesentlichen Erhöhung von Latenzzeiten und birgt die Gefahr von Verkehrsstaus. In solchen Fällen werden Datenpakete von Switches verworfen, was zwangsweise einen negativen Einfluss auf die Applikationsleistung hat.

Die herkömmlichen Ethernet-Bridging-Protokolle (Spanning Tree) entstanden, um Schleifen zu vermeiden, auch wenn es dadurch zu Leistungseinbußen kommt. Falls ein Netzwerk mehrere Wege zwischen zwei Endpunkten bereitstellt, schaltet Spanning Tree alle Ports ab, die zum selben Ziel führen könnten. Dieses Verhalten beeinflusst die Skalierbarkeit von Netzen und führt zu einer wachsenden Ineffizienz, da ein wesentlicher Teil der theoretisch verfügbaren Bandbreite nicht zur Verfügung steht.

Anders als bei Ethernet erlauben die Switching-Einheiten (Fabrics) von Infiniband und Fibre Channel Mehrfachwege zwischen zwei Endpunkten. Typischerweise sorgt ein Fabric-Manager für eine optimale Ausnutzung aller Ports. Mehrfachwege werden parallel benutzt, sodass die verfügbare Bandbreite mit der Anzahl paralleler Verbindungen linear wächst. Selbst vermaschte Topologien sind möglich. Um ein ähnliches Verhalten auch für Ethernet zu ermöglichen, arbeiten IEEE und IETF derzeit an einer Erweiterung des Ethernet-Standards. Allerdings wird es sicher einige Jahre dauern, bis ein herstellerübergreifender Standard entwickelt und im Markt eingeführt ist.

Vertikale vs. horizontale Skalierung

Wird mehr Kapazität in Rechenzentren benötigt, so gibt es dazu zwei mögliche Ansätze, nämlich erstens die vertikale Skalierung (Scale-up), was einfach den Einsatz einer größeren monolithischen Maschine bedeutet, und zweitens die horizontale Skalierung (Scale-out). In diesem Fall wachsen mehrere kleinere Maschinen in einem logischen Cluster zusammen.

Der bisher häufig gegangene Weg der vertikalen Skalierung hat einige entscheidende Limitationen. Größere Maschinen kosten meist viel mehr als mehrere kleinere Maschinen mit in der Summe gleicher Kapazität. Die Skalierbarkeit ist begrenzt, denn selbst die größte Maschine ist in manchen Fällen nicht groß genug. Es ist teuer, komplex und manchmal unmöglich, die Kapazität einer Maschine nach der Inbetriebnahme zu verändern, und bei einem Systemausfall ist der gesamte Service unterbrochen.

Converged Enhanced Ethernet

Infiniband (und teilweise auch Fibre Channel) wurde von Anfang an für die horizontale Skalierung entwickelt. Ziel war eine möglichst einfache Switching-Infrastruktur mit Unterstützung vermaschter Topologien und Mehrfachwegen sowie einem zentralem Discovery- und Policy-Management.

Seit ein paar Jahren arbeitet die IEEE daran, Ethernet mit Infiniband-ähnlichen Eigenschaften zu erweitern, um damit das neue Converged Enhanced Ethernet (CEE) zu formen. Dieser Standard würde es möglich machen, Ethernet als einheitliche Technik für horizontal skalierbare Rechenzentren zu nutzen.

Hersteller wie Voltaire mit langjährigen Erfahrungen aus der Infiniband-Welt bieten bereits heute Ethernet-Lösungen an, die beliebig horizontal skalierbar sind und dies ohne Performanceeinbußen und ohne ein lineares Kostenmodell zu verlassen. Voltaires Infiniband- und 10Gig-Ethernet-Switches werden über einen zentralen Fabric-Manager verwaltet. Sie unterstützen die Migration der Rechenzentren der nächsten Generation zu virtualisierten und automatisierten Service-Zentren.

Hans Bley/jos

Hans Bley ist Sales and Marketing Manager bei Atlantik Systeme in Planegg.

■ Info: Atlantik Systeme
Tel.: 089/89505229
Web: www.atlantiksysteme.de