Ethernet hebt ab

Performancemessungen von Gigabit-Ethernet-Systemen

Kai-Oliver Detken

Schon 1995 wurde bei der IEEE die Arbeitsgruppe 802.3z Gigabit Task Force gegründet, die die Arbeiten des Fast-Ethernet-Standards auf Gigabit-Ethernet (GE) erweitern sollte. Ziel war es, Fullduplex-Übertragung mit 1 Gbit/s zu ermöglichen und dabei dieselben Datenpakete wie Ethernet 10Base-T und Fast-Ethernet 100Base-T zu verwenden. Die Empfehlungen für Media Access Controller, Repeater und Physical Layer lagen sehr schnell vor, da man auf bestehende Spezifikationen zurückgriff. Komplexere Themen wie die Übertragung über Kupferleitungen oder die Flußsteuerung verzögerten den Standard jedoch bis 1999. Nun sind erste GE-Systeme verfügbar und sollen auch in den Weitverkehrs- sowie den Access-Bereich vordringen. Ob sie anderen Techniken wie SDH oder ATM Konkurrenz machen können, muß sich aber noch erweisen.

Kai-Oliver Detken ist Senior IT-Consultant der Decoit sowie freier Autor und Berater in Grasberg/Bremen

Gigabit-Ethernet (GE) verwendet den gleichen Rahmenaufbau und das gleiche Zugriffsverfahren wie seine Vorgänger. Der Gigabit Media Access Frame (GMAF) ist identisch mit der Struktur des klassischen Ethernet. Die sieben Byte der Präambel mit dem Bitmuster 1010...10 dienen der Taktsynchronisation des Empfängers. Danach kommt das Feld Start Frame Delimiter (SFD), das den Rahmenbegrenzer darstellt. Er besitzt ein unverwechselbares Bitmuster (10101011). Die MAC-Quellen- und -Zieladresse (Media Access Controller) schließen sich an. Sie werden wiederum durch das Feld Länge/Typ begrenzt. Anschließend kommen die Nutzdaten, die max. 1500 Byte betragen können.

Rahmen-Erweiterungen

Das Padding-Byte dient zum Auffüllen der Daten auf eine gerade Anzahl von Bytes und muß nicht vorhanden sein. Abschließend wird noch eine Prüfsumme aus dem Datenfeld gebildet, um Fehler erkennen zu können.
Aufgrund von Problemen bei der Erkennung von Kollisionen und im Sinne der Effizienz der Bandbreite waren aber noch folgende Erweiterungen einzufügen:

Das Thema in Kürze

Seite Ende 1999 diskutiert man über eine mögliche nächste Generation von Ethernet. Erste Systeme nach 10-Gigabit-Ethernet sind verfügbar, allerdings völlig proprietär aufgebaut. Ob man die versprochene Performance, die selbst 1-Gigabit-Ethernet heute bereits bietet, überhaupt am Endgerät nutzen kann und welche Einsatzmöglichkeiten GE bietet, soll dieser Artikel, auch anhand eigener Messungen, klären.

- *Carrier Extension (CE):* Sie erweitert das Ethernet-Paket auf eine Mindestlänge von 512 Byte, um Kollisionen noch erkennen zu können;
- *Packet-Bursting:* Sendung von zusätzlichen Paketen nach Ablauf des störungsfreien Wettbewerbs-Intervalls, um die Performance zu erhöhen;
- *Multiple Link Segments:* hat die Aufgabe, mehrere Ethernet-Verbindungen zusammenzuschalten. Diese können dann parallel genutzt werden, um eine grobe Skalierbarkeit und redundante Verbindungen zu bekommen;
- *Buffer Distribution:* soll das Problem der Einschränkung von Bandbreite durch Repeater kompensieren. Dieser ist zwar in der Lage, Pakete zu speichern, kann aber keine Paketanalyse durchführen;
- *Jumbo-Frames:* es können durch den Einsatz von 10-/100-/1000-Mbit/s-Rahmen unterschiedliche Übertragungsraten für eine grobe Skalierbarkeit fest eingestellt werden.

Bei einer Paketgröße von z.B. 64 Byte werden durch die CE 448 Byte verschwendet, wodurch die Leistung des Netzes im schlimmsten Fall auf Fast-Ethernet-Niveau sinken könnte.
Eigentlich wären 640 Byte notwendig gewesen, um Kollisionen in jedem Fall erkennen zu können. Wegen des damit verbundenen Leistungsabfalls hat man aber auf diese Paketgröße verzichtet. Bisherige Untersuchungen zeigen jedoch, daß die minimale Paketgröße zwischen 200 und 640 Byte liegt, so daß dieser Effekt nicht so stark ins Gewicht fällt.
Diese Paketverlängerung hat jedoch einen Engpaß zur Folge, da je nach Paketlänge nur von einer Leistungssteigerung mit dem Faktor zwei bis neun gegenüber Fast-Ethernet ausgegangen werden kann. Ob dieses Verfahren in der Praxis noch weitere Probleme verursachen wird, bleibt abzu-

warten. Auf jeden Fall wird sich die Verzögerungszeit weiter erhöhen.

Funktionsweise von Gigabit Ethernet

Gigabit-Ethernet bietet auch noch das Zugriffsverfahren CSMA/CD an, das die Basistechnologie darstellt. Im Halfduplex-Betrieb können hier Datenkollisionen entstehen, die erst im Fullduplex-Betrieb verschwinden, da hier bidirektional gesendet und empfangen wird und somit die Gesamtbandbreite für Sender und Empfänger gleichzeitig zur Verfügung steht. Oft wird allerdings fälschlicherweise angenommen, daß man durch Fullduplex-Betrieb statt 1 Gbit/s die doppelte Datenrate von 2 Gbit/s zur Verfügung hat.
Allerdings kann es bei Fullduplex-Betrieb zu Verarbeitungsproblemen im Switch kommen. Verwendet werden Punkt-zu-Punkt-Verbindungen, die über die Switche gesteuert werden. In diesem Zusammenhang wird ein einfaches Flußsteuerungsverfahren angeboten, das auf einem Pause-Mechanismus basiert. Das heißt, die empfangende Station kann den Sender durch Aussenden eines XOFF-Pakets beeinflussen. Somit kann der Switch kurz vor Überlastung eines Eingangs-/Ausgangspuffers eine Nachricht an den Sender abschicken. Der Sender verzögert oder stoppt anschließend den Datentransport für den im Paket angegebenen Zeitraum. XOFF-Pakete der Dauer Null können längere Pausen aufheben. Das wird in dem Standard IEEE 802.1x festgehalten. Damit diese einfache Flußregelung zum Einsatz kommen kann, müssen natürlich alle Switche diesen Standard unterstützen.
Um im Backbone gleichzeitig verschiedene aktive Verbindungen zwischen zwei Switchen zur Verfügung zu stellen und eine dynamische Lastverteilung abhängig vom Verkehr und dem Übertragungsmuster zu gewährleisten, ist das Switch Meshing von einigen Herstellern eingefügt worden. Hierbei wird durch die integrierte Redundanz eine höhere Effizienz erreicht als beispielsweise durch das Spanning-Tree-Verfahren.
Das Spanning Tree Protocol (STP) schaltet alle redundanten Wege ab. Erst wenn eine Verbindung ausfällt, wird sie durch einen anderen Pfad ersetzt. Der Ansatz des Switch Meshing setzt hingegen alle existierenden Verknüpfungen für den Datentransport ein. Dabei wird der Verkehr auf alle Pfade verteilt, indem dynamisch die entstehenden Kosten ermittelt werden und daraus der günstigste Weg berechnet wird. Dadurch kann ebenfalls eine Verdoppelung der Bandbreite zwischen den Switchen umgesetzt werden. Ausfallende Verbindungen werden in ca. 1 s durch andere ersetzt.
Bei der Grundstruktur vermaschter Switche wird mindestens ein Port für die Vermaschung konfiguriert, der wiederum mit einem vermaschten Switch verbunden sein muß.
Es ist ebenfalls möglich, mit mehreren vermaschten Switchen in einer Meshing Domain angeordnet zu sein, die auch mit weiteren Switchen außerhalb dieser Domäne verbunden sind. Die Meshing Domain besteht dabei aus einer Gruppe von geswitchten Ports, die über das Meshing-Protokoll Pakete austauschen. Die Pfade zwischen den Switchen können vielfach redundant ausgelegt sein, ohne Broadcast-Stürme zu verursachen. Vermaschte Verbindungen gehören zu den Punkt-zu-Punkt-Verbindungen. Alle vermaschten Ports eines Switches sind dabei in derselben Switch Meshing Domain angeordnet. Es müssen allerdings nicht alle Ports dazugehören. Die Anordnung der Switch Meshing Domain verspricht folgende Vorteile gegenüber herkömmlichen Routing-Protokollen:

- Dynamische Werte werden bei Auswahl der Verbindung zugrunde gelegt, um die günstigste Verbindung zu erhalten. Herkömmliches Routing (OSPF oder RIP) verwendet nur konstante Größen, wie Kosten oder Anzahl der Router-Hops.
- Layer-3-Protokolle können genauso verwendet werden wie nicht-routbare Protokolle.
- Die Konfiguration ist relativ einfach, da nur die Ports definiert werden müssen, die zu einer vermaschten Domäne gehören. Alles andere wird durch den Switch umgesetzt.
- Die Reaktionszeiten sind extrem gering und liegen bei ca. 1 s.
- Verzögerungszeiten sind geringer gegenüber einem Router. Da die Pakete nicht geändert werden, kommen keine Verzögerungen im Netz mehr hinzu.

LAN Aggregation wird ebenfalls oft mit Switch Meshing verglichen, da hier ebenfalls Kapazitätssteigerungen und Redundanz im Netz erreicht werden können. Hier sind mehrere Verbindungen gleichzeitig aktiv, so daß sich Schleifen bilden können. Zusätzlich ist keine dynamische Lastverteilung möglich, da dies statisch vorgenommen wird. Man darf sie nicht mit der Link Aggregation verwechseln, die zur Steigerung der Datenrate bei Punkt-zu-Punkt-Verbindungen eingesetzt wird. Das Port Trunking setzt praktisch auf dieser Technik auf, indem mehrere Ports als einzelner, schneller Port interpretiert werden. Das Ziel des Port Trunking ist dabei ausschließlich die Reduzierung des Verwaltungsaufwands durch mehrere Adapterkarten (Network Interface Controller – NIC) in einem Server. Trotz mehrerer NICs bekommt nämlich der Server nur einen Namen und eine IP-Adresse (Internet Protocol) zugewiesen. Zur Lastverteilung der parallel laufenden Verbindungen werden Einwege- und Zweiwege-Lastverteilungen eingesetzt. Die erste Methode ist die einfachere, da keine Änderung am Switch erforderlich ist. Bei der Zweiwege-Methode müssen die Switche einbezogen werden, da spezielle Link-Aggregation-Funktionalität geleistet werden muß. Dafür sind eine hohe Transparenz bzw. einfache Verwaltung und bidirektionale Lastverteilung vorhanden. Weiterführende Informationen findet man unter der Adresse der 10-GE-Allianz: **www.10gea.org**.

Meßmöglichkeiten von GE-Netzen

Die fehlende Skalierbarkeit bei Gigabit-Ethernet führt schnell zu hohen Auslastungen, da auch immer mehr GE-Adapterkarten in Servern Verwendung finden. Diese belasten dann das Netz mit der maximal möglichen Datenrate, da die Endstationen die Da-

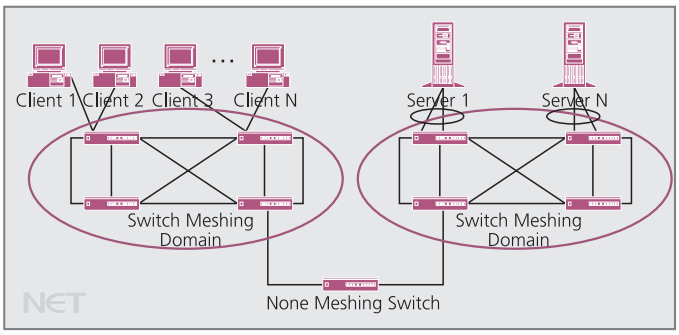

Bild 1: Switch Meshing Domains, gekoppelt mit einem normalen Switch

tenrate bei der GE-Technik nicht begrenzen können. Auf der anderen Seite stellt sich aber auch die Frage nach der Performance einer GE-Verbindung, da für die Datenrate von einem Endsystem zum Server auch weitere Parameter wie Bustyp, interner Bustakt, Betriebssystem, aktive/passive Adapterkarte, Backplane der Switche usw. ausschlaggebend sind. So erreichen selbst 100-Mbit/s-Adapterkarten im Fast-Ethernet-Umfeld bei schlechter Konfiguration oder nicht leistungsfähigen Switchen/Endgeräten nicht immer die mögliche Nettoübertragungsrate von ca. 90 Mbit/s.

Folgende Testmöglichkeiten lassen sich nach den Spezifikationen RFC-1242, RFC-1944 und RFC-2285 nennen, um ein Netz auf wesentliche Kriterien hin überprüfen zu können:

- *Performance (Durchsatz):* Die maximal mögliche Datenrate wird ermittelt, indem der Switch mit unterschiedlichen Paketgrößen und Netzverhaltensmustern konfrontiert wird. Anhand solcher Messungen lassen sich Leistungsengpässe und Probleme relativ schnell erkennen und ggf. beseitigen.
- *Latenzzeit:* Hohe Latenzzeiten sind ebenfalls der Leistung abträglich. Das gilt besonders für die Übermittlung zeitkritischer Signale, etwa bei Voice over IP (VoIP), Audio- oder Videostreams, es kann sich aber auch auf empfindliche Datenanwendungen beziehen. Deshalb muß die Gesamtverzögerung eines Netzes immer beachtet werden.
- *Paketverlustmessung:* Die Messung der Paketverluste stellt den „Gesundheitszustand" eines Netzes auf die Probe. Verliert das Netz übermäßig viel Pakete, ist es überlastet und verliert somit auch wieder an Performance. Fehlerhafte Switche können ebenfalls zu Paketverlusten führen.
- *Many-to-Many-Test:* Bei dieser Messung wird von jedem Port ein Datenstrom an jeden anderen Port eines Switches gesandt. Gleichzeitig muß jeder Port Daten entgegennehmen, und zwar von jedem der anderen an der Messung beteiligten Ports. Dabei werden die Belastung sowie die Paketgröße variiert.
- *Many-to-One; One-to-Many:* Bei GE-Messungen kann man auch diese Methoden einsetzen, um mittels vieler Fast-Ethernet-Verbindungen einen einzelnen GE-Port durchzumessen.
- Broadcast-Belastung: Die Belastung eines Layer-2-Netzes wird oftmals durch sog. Broadcast-Stürme auf die Probe gestellt. Sie verursachen eine hohe Netzbelastung durch kontinuierliches Abfragen einzelner Ports. Manche Layer-2-Switche haben hier Vorkehrungen vorgesehen, die die Broadcast-Stürme begrenzen sollen.
- *Head-of-Line-Blocking:* Dabei wird überprüft, ob eine Überlastsituation an einem Port Störeinflüsse an anderen, unbelasteten Ports nach sich zieht. Dies darf bei ausreichender Dimensionierung der Backplane und ausreichendem Pufferverhalten keinen Einfluß auf heutige Switche haben.
- *Virtual LAN (VLAN):* Das Aufsetzen eines VLAN im Zusammenspiel mit einer Full-Meshed-Anordnung kann ebenfalls zu Leistungstests herangezogen werden, um realistische Werte, bezogen auf ein Gesamtnetz, zu bekommen. Auch hier stehen die Performance im Vordergrund sowie die Interoperabilität mit anderen Switchen in bezug auf den Standard IEEE.802.Q.

Vermeidung von Leistungsengpässen

Um Performance-Messungen durchzuführen, kann man unterschiedliches Meßequipment einsetzen. Dabei ist darauf zu achten, daß man nicht die Applikation oder die Übertragungsrate der Festplatte mißt, sondern wirklich die Netzleistung. Eine Möglichkeit dazu bietet das Tool NETPERF von Hewlett-Packard an, das auf Basis von Kommandozeilen bedient wird. Hier können im Umfeld TCP/UDP-IP einige Parameter wie Puffer- und Paketgröße variiert werden.

Eine Beispielmessung unterstreicht, daß in der Praxis durchaus andere Werte erreicht werden können, als theoretisch denkbar sind. Es wurde eine Punkt-zu-Punkt-Messung zwischen zwei Rechnern durchgeführt, die die gleiche Ausstattung (AMD Duron, 600 MHz, 128 MByte RAM) und das gleiche Betriebssystem (Windows NT, SP6) hatten. Verwendet wurden auch identische 1GE-Adapterkarten von SMC Networks , die den PCI-Bus (32 und 64 bit) unterstützten. Um unterschiedliche Verhaltensmuster zu identifizieren, wurden unterschiedliche Paketgrößen verwendet. Kleine Pakete etwa belasten das Netz sehr stark und werden z.B. häufig bei Datenbankabfragen verwendet. Große Pakete wiederum haben einen geringen Overhead, müssen aber fragmentiert werden, was sich in der Performance niederschlägt.

Wie man aus den Testergebnissen sehen kann, wird unabhängig von der verwendeten Paketgröße ein Durchsatz von 400 Mbit/s nicht überschritten. Zwar bricht die Datenrate bei kleinen Paketen noch stärker ein, da nur ca. 240 Mbit/s ermittelt werden konnten, aber die theoretisch erreichbare Nettodatenrate von 950 Mbit/s wurde nicht ansatzweise erreicht.

Die Gründe sind unterschiedlich. Zum einen konnte nur ein PCI-Bus von 32 bit mit 33 MHz eingesetzt werden, was aber keine Begrenzung darstellt. Trotzdem werden künftige PCI-X-Bustypen mit 1 GByte/s höhere Datenraten anbieten. PCI-X gilt als einer der potentiellen Nachfolger des PCI-Standards. Zum anderen handelte es sich

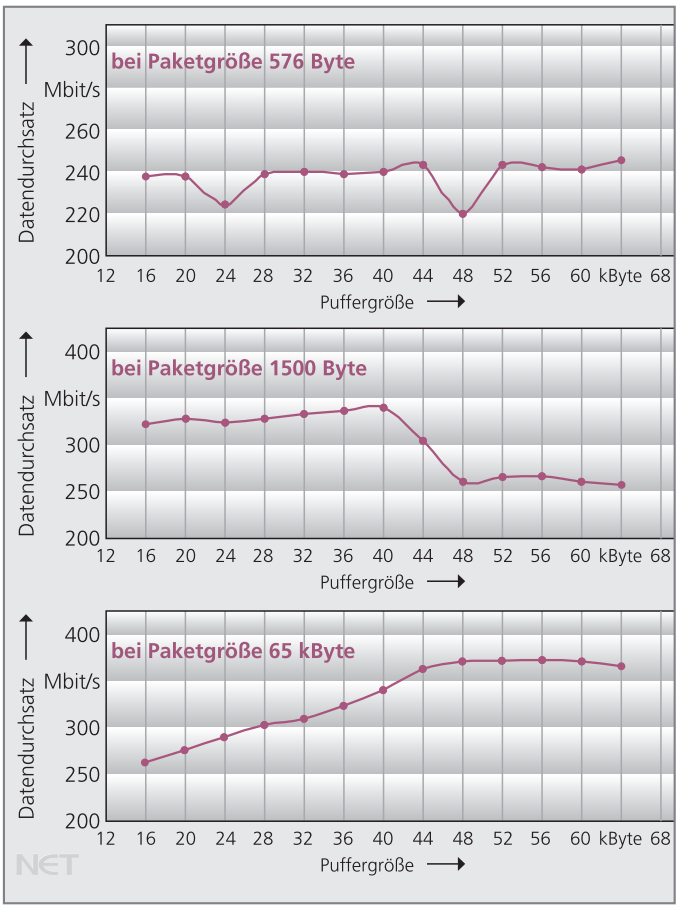

Bild 2: Performancewerte einer GE-Verbindung

um passive Adapterkarten, die die Verarbeitung des Datenverkehrs nicht aktiv unterstützen, sondern die CPU-Leistung in Anspruch nehmen. Hinzu kommt, daß das Betriebssystem Windows NT ebenfalls hohe Anforderungen an die Rechnerperformance stellt, weshalb hier nicht die volle Leistung zur Verfügung stand.

Das Testprogramm NETPERF verwendete als Testprotokoll TCP/IP. TCP ist ein bidirektionales Protokoll, das verbindungsabhängig ist. Um zu verhindern, daß Datensegmente wegen zu kleiner Datenpuffer oder Überlastung des Empfängers nicht verarbeitet werden können, ist der Empfänger in der Lage, den ankommenden Datenfluß vom Sender zu begrenzen. Dabei sind die Gründe, daß Protokolle höherer Schichten oder der Empfänger die TCP-Segmente nicht so schnell verarbeiten können, für eine Verminderung des Datenflusses ebenfalls ausschlaggebend. Das liegt an nicht ausreichend dimensionierten Empfangspuffern. Sind die Empfangspuffer ausgelastet, werden neu empfangene Datensegmente verworfen. Dadurch wird eine Sendewiederholung nötig, die den Durchsatz einer Übertragung empfindlich verringern kann. Aus diesem Grund wurde eine zusätzliche Funktion eingearbeitet, die Sliding-Window-Mechanismus genannt wird. Dieser funktioniert so, daß der Empfänger dem Sender die Datenmenge mitteilt, die er verarbeiten kann. Dabei wird der verfügbare Speicherbereich des Empfangspuffers zur Ermittlung der TCP-Windows-Size verwendet.

Bei jeder Verbindung wird deshalb eine Maximum Segment Size (MSS) für das Netz festgelegt, um eine effektive Kommunikation gewährleisten zu können. Die MSS-Größe legt dabei die maximal zulässige TCP-Segmentgröße fest. Diese ist ebenfalls für die Durchsatzraten entscheidend, da zu kleine MSS-Werte das Netz höher belasten würden. Normalerweise wird dabei die MSS durch das Abziehen des TCP/IP-Headers von der Maximum Transmission Unit (MTU) des Netzes gebildet. Somit kann der Teilnehmer an einer Sitzung den Inhalt der Windows Size ohne Bestätigung an andere Benutzer schicken. Die Fenstergröße wird anschließend immer weiter geschlossen und erst durch eine Quittierung von empfangenen Datenbytes wieder geöffnet. Durch diesen Mechanismus ist ein höherer Datendurchsatz möglich, da nicht nach jeder Datensendung auf eine Empfangsbestätigung gewartet werden muß.

Fazit

Gigabit Ethernet und auch der Nachfolger 10-Gigabit-Ethernet sind leistungsstarke Netztechnologien, die allerdings richtig eingesetzt werden wollen. Nicht immer ist die Auswahl der Endgeräte, Bussysteme, Betriebssysteme usw. adäquat. Ebenfalls dürfen die verwendeten Protokolle nicht außer acht gelassen werden, da beispielsweise FTP durch die Quittungsmechanismen von TCP ebenfalls nicht in der Lage ist, solche Datenraten auszunutzen. Hinzu kommt, daß immer wieder proprietäre Eigenschaften in GE- und vor der Verabschiedung von Standards vor allem in 10GE-Geräten implementiert werden, um die Leistungsfähigkeit zu erhöhen auf Kosten der Interoperabilität mit anderen Herstellern. Diese Engpässe sollte man herstellerunabhängig vor der Konzeption und dem Aufbau eines Netzes evaluieren lassen, um späteren Enttäuschungen oder gar Mehrkosten aus dem Weg zu gehen.

Heutige Ethernet-Technik (GE, Fullduplex und 10GE) hat mit der Grundtechnik des Ur-Ethernet nichts mehr gemeinsam. Während GE auf Fibre Channel basiert, wird 10GE auf SDH-Framing aufbauen. Nur der Name sowie die Rahmengröße stehen noch für Ethernet. Allerdings dringt Ethernet in Bereiche vor, die dieser Technologie vorher fremd waren. So werden sich GE und 10GE immer mehr zusätzlich im WAN sowie Access-Bereich ansiedeln und weitere Möglichkeiten für höhere Datenraten bieten. Dabei muß Ethernet allerdings in den Bereichen Zuverlässigkeit, Redundanz und Ausfallsicherheit noch einiges nachholen, um mit WAN-Techniken wie ATM und SDH konkurrieren zu können.

(we)