Teleimmersion – Erfahrungen mit der Gigabit-Ethernet-Strecke zwischen Berlin und Hannover

Article · January 2002

CITATIONS
0

READS
47

4 authors, including:

Ralf C. Staudemeyer
University of Applied Sciences Schmalkalden
34 PUBLICATIONS **414** CITATIONS

SEE PROFILE

Teleimmersion

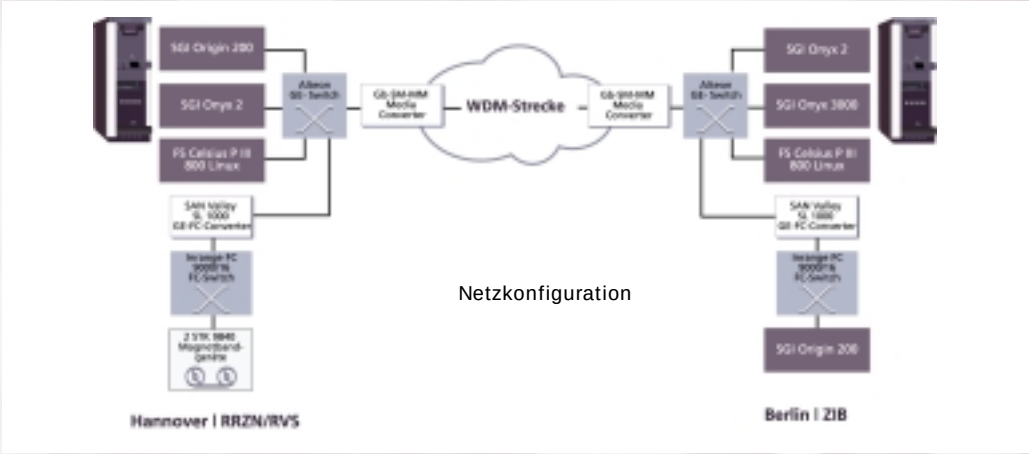
Erfahrungen mit der Gigabit-Ethernet-Strecke zwischen Berlin und Hannover

Seit dem Jahr 2000 sind die Wissenschaftseinrichtungen in Deutschland mit Bandbreiten bis zu 622 Mb/s an das Gigabit-Wissenschaftsnetz G-WiN angeschlossen. Und eigentlich sollten durch die Dimension dieser Anschlüsse alle Wünsche von Netzverbindungen erfüllt sein. Dennoch gibt es im scientific computing Anforderungen, die besondere Netzwerklösungen erfordern. So ist bei der engen Kopplung mehrerer Rechner oder bei der Steuerung von wissenschaftlichen Großgeräten nicht allein die Bandbreite ausschlaggebend, sondern auch Qualitätsparameter der genutzten Internet-Verbindung, zum Beispiel eine geringe Latenzzeit.

Das Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB) und das Regionale Rechenzentrum für Niedersachsen (RRZN) an der Universität Hannover werden ab Mitte 2002 gemeinsam einen verteilten Hochleistungsrechner an ihren Standorten betreiben, dieser wird eine außergewöhnliche Netzverbindung - hohe Bandbreite bei geringer Latenzzeit - erfordern. Um die Machbarkeit einer solch engen Kopplung zweier entfernt betriebener homogener Rechner zu prüfen und um Erfahrungen mit solch einer dedizierten Netzverbindung zu sammeln, betreiben ZIB und RRZN gemeinsam mit dem DFN-Verein seit August 2001 eine Gigabit-Ethernet-Verbindung zwischen Berlin und Hannover. In den DFN-Mitteilungen Heft 57 vom November 2001 wurde das DFN-Projekt "Tele-Immersive Visualisierung mittels 3D-Streamingverfahren im Gigabit-Wissenschaftsnetz" näher beschrieben, im vorliegenden Beitrag wollen wir näher auf einige netzspezifische Eigenschaften dieser Verbindung eingehen.

Die Gigabit-Ethernet-Verbindung Berlin - Hannover

Die zentrale Komponente der ca. 315 km langen Netzverbindung ist der im Rahmen des zwischen dem DFN-Verein und der T-Systems vereinbarten G-WiN-Vertrages bereitgestellte "transparente



Netzkonfiguration

WDM-Kanal -" mit möglichen Übertragungsgeschwindigkeiten zwischen 0,1 und 2,5 Gbit/s. Vereinfacht wurde diese Bereitstellung dadurch, dass die vom DFN-Verein betriebenen G-WiN-Netzknoten 1. Ordnung für Berlin, Brandenburg und Mecklenburg-Vorpommern bzw. für Niedersachsen in Räumen des ZIB bzw. des RRZN untergebracht sind. Taktgeber für diese Fernstrecke und gleichzeitig Endpunkte der Strecke sind je ein SingleMode-MultiMode-Umsetzer vom Typ Transition Networks Gigabit Singlemode to Multimode Media Converter in beiden Zentren. Die lokale Weiterleitung erfolgt über je einen Gigabit-Ethernet-Switch vom Typ Alteon ACESwitch 180 mit der Eigenschaft des Transfers von Jumbo-Frames. An diese Switche sind die verschiedenen Grafikrechner, Streaming-Server, PCs und Fibre Channel Geräte über spezielle Umsetzer angeschlossen. Das ganze Netz bildet ein IP-Subnetz bzw. eine Broadcast-Domain.

Nach längeren Arbeiten seitens der T-Systems unter Hilfestellung von Mitarbeitern des ZIB und des RRZN konnte die Strecke am 15. August 2001 in Betrieb gehen. Erste Messungen noch am selben Tag zeigten hervorragende Round-Trip-Zeiten von IP-Paketen von etwa 4,5 ms (Millisekunden). Üblicherweise rechnet man die Ausbreitungsgeschwindigkeit des Lichts in einer Glasfaser mit 200 km/ms, damit beträgt die Round-Trip-Zeit des Lichtimpulses dieser 315 km langen Strecke etwa 3,15 ms, die übrige Verzögerung von ca. 1,35 ms ist durch die Verarbeitung in den Rechnern und in den Ethernet-Switches bedingt. Die zunächst gemessenen Bandbreiten hingegen waren mit 285 Mbit/s enttäuschend.

Durch spezielle Erweiterungen des Ethernet-Standards (Jumbo-Frames) und spezielle Übertragungsprotokolle (Scheduled Transfer, STP) lassen sich jedoch erheblich bessere Werte und sogar fast die theoretisch mögliche Transferrate von 991 Mbit/s erreichen.

Jumbo Frames

Die Ursache der geringen Datenraten von zunächst nur einem Viertel des theoretisch Möglichen liegt nicht unbedingt in einer zu geringen I/O-Leistung, sondern eher in einer zu geringen IP-Paket-Leistung der Rechner. Da nach dem Ethernet-Standard die maximale Rahmen-Größe auch für Gigabit-Ethernet 1518 Byte beträgt, müssen z. B. gegenüber Fast-Ethernet zehnmal so viele Rahmen (ca. 80000) pro Sekunde übertragen werden, um die volle Leistung zu erreichen. Da viele Operationen des Protokoll-Stacks auf die Header angewandt werden, erzeugen viele kleinere Rahmen eine höhere Belastung als wenige große. Die meisten heutigen Rechner sind damit überfordert.

Der Hersteller Alteon hat eine Abweichung vom Gigabit-Ethernet-Standard propagiert, bei der durch eine Erhöhung der Ethernet-Rahmengröße auf 9018 Byte die Belastung der Rechner signifikant verringert wird. Alteon hat seine Produkte mit diesem Feature, genannt Jumbo-Frames, ausgestattet und viele Hersteller sind diesem Beispiel gefolgt. Ein positiver Nebeneffekt der Jumbo-Frames ist, dass der Protokoll-Overhead geringer ist und sich dadurch die Nutzdatenrate erhöht.

Jumbo-Frames sind allerdings nicht in jedem Szenario sinnvoll und erfordern besondere Anpassungen der Netzwerkhardware, um wirklich Verbesserungen zu bringen. So profitieren Anwendungen, die viele kleine Nachrichten austauschen, nicht davon. Die im Rahmen des DFN-Projekts "Tele-Immersion" zu übertragenden Streaming-Anwendungen sind in diesem Sinne ein idealer Kandidat für Jumbo-Frames, da pro Stream mehrere Gigabyte an Daten anfallen, die mit etwa konstanter Datenrate übertragen werden sollen.

Ein weiterer kritischer Parameter in diesem Zusammenhang ist die TCP-Fenstergröße: Sie bestimmt die Anzahl der Bytes, die ein Sender abschicken darf, bevor eine Bestätigung vom Empfänger erwartet wird. Durch sie erfolgt die Flusssteuerung. Wird sie zu klein gewählt, so legt der Sender Pausen ein, wenn entsprechend viele Daten gesendet wurden, ohne eine Bestätigung erhalten zu haben. Die benötigte Größe lässt sich aus Bandbreite und Verzögerung ermitteln. Die Standard-Fenstergröße ist systemseitig oft mit 16 KByte festgelegt. Für die Verbindung zwischen Berlin und Hannover mit einer Round-Trip-Zeit von ca. 4,5 ms zeigt die Rechnung, dass die Standard-Fenstergröße mit 16 KByte gegenüber der optimalen von über 500 KByte allerdings viel zu klein ist.

Langzeitmessungen zwischen ZIB und RRZN wurden mit der ursprünglich von Hewlett/Packard entwickelten Messsoftware netperf (Version 2.1p3) vorgenommen. Mit netperf können zu übertragende Datenraten beliebig vorgegeben werden, es wird z. B. keine Beeinträchtigung der Messung durch Plattenzugriffe nötig. Die höchsten Transferraten wurden im Rahmen dieser sehr ausführlich durchgeführten Messungen zwischen einer SGI Origin 200 (Hannover) und SGI Onyx 2 (Berlin) erreicht: Bei diesen Rechnern mit ihren hohen I/O-Leistungen wurden bei einer TCP-Fenstergröße von 1280 KByte und einem Socketbuffer von 1024 KByte jeweils 1000 generierte Datenpakete ab einer Größe von 64 KByte übertragen. Auf der dediziert genutzten Datenleitung mit theoretisch möglichen 991 Mbit/s wurden so Datentransferraten über 980 Mbit/s (122 MByte/s) erreicht.

Das Scheduled Transfer Protocol (STP)

Bei Übertragungen im Gigabit-Bereich werden enorme Anforderungen an die Protokollsoftware (an das Betriebssystem) gestellt, da das Ein- und Auspacken der Daten in die Protokolldatenstrukturen mit hoher Geschwindigkeit erfolgen muss. Das führt zu Problemen bei der Verwendung von Standardprotokollen auf Hochleistungsnetzen.

Zur Entwicklungszeit des Transmission Control Protocols (TCP, RFC-793, 1981[!]) waren die Übertragungsnetze fehleranfällig und vergleichsweise langsam. Daher enthält TCP Mechanismen, um auf Datenverluste, Datenverfälschungen und Stausituationen zu reagieren. Heutige Hochleistungsnetzwerke erreichen selbst bei hohen Übertragungsgeschwindigkeiten sehr geringe Fehlerraten, hier reicht ein sehr einfacher Fehlerkorrekturmechanismus aus. Da TCP für diese Netze zu schwerfällig und ressourcenintensiv ist, wurden effizientere Übertragungsprotokolle entwickelt.

Das Scheduled Transfer Protocol (STP) ist aus der Entwicklung des GSN-Interconnects (Gigabyte System Network) hervorgegangen. Hier werden vor der tatsächlichen Datenübertragung mit Hilfe von kleinen Kontrollnachrichten Empfangspuffer beim Empfänger reserviert, das heißt, der Transfer der Daten wird gewissermaßen angekündigt, so dass der Empfänger ihn einplanen (schedulen) kann. Auf diese Art und Weise können die Daten direkt vom Netzwerk in den Speicher des Empfängers übertragen und Stausituationen vermieden werden.

Darüber hinaus ist das Protokoll für eine Ausführung bestimmter Funktionen in der Netzwerkkarte ausgelegt. So kann einerseits die Verarbeitungsbelastung des Betriebssystems drastisch gesenkt werden und andererseits werden bessere Leistungswerte erreicht.

Der Nachteil der Hardwareunterstützung des STP ist, insbesondere bei der Verwendung auf Gigabit-Ethernet (GE), dass nur eine kleine Gruppe von Netzwerkkarten die Voraussetzungen mitbringt, die für einen Leistungsgewinn durch die Auslagerung von Protokollfunktionen in die Karte benötigt werden. Im Gegensatz zu GSN- oder Myrinet-Adaptern, die generell mit Arbeitsspeicher ausgestattet und programmierbar sind, ist diese Eigenschaft bei GE-Interfaces nicht sehr verbreitet.

Zur Zeit existieren Implementierungen des STP von SGI für GSN und GE unter IRIX. Diese sind allerdings nicht öffentlich verfügbar. Daher wurden für die Tests zwei Linux-PCs mit einer als Open-Source verfügbaren STP-Implementierung ausgestattet, die ebenfalls von SGI stammt. Sie kann sowohl mit als auch ohne Hardwareunterstützung benutzt werden, wobei die Verwendung ohne Hardwareunterstützung nur sehr geringe Leistungsvorteile gegenüber TCP bringt. Für die Verwendung der Hardwareunterstützung werden GE-Adapter mit dem Tigon oder Tigon II Chipsatz benötigt. Sie sind von verschiedenen Herstellern erhältlich, so zum Beispiel von 3Com und NetGear. Für unsere Tests wurden Karten vom Typ 3C985B-SX von 3Com verwendet, die mit einem Glasfaser-Anschluss ausgestattet sind.

Während der Tests zeigte sich der klare Vorteil von STP gegenüber TCP bezüglichder Verarbeitungslast während des Datentransfers. Bei Übertragungen mit TCP waren die Testrechner, sowohl beim Senden, als auch beim Empfangen, im

Durchschnitt zu 50 % mit der Verarbeitung der Daten beschäftigt. Durch die Verwendung von hardwarebeschleunigtem STP fiel die CPU-Last auf beiden Seiten auf ca. 5 %. Bei den lokalen Durchsatzmessungen mit Jumbo-Frames zeigte sich ein Vorteil von STP mit 104 MB/s gegenüber TCP mit 85 MB/s. Über den WDM-Kanal erreichte STP dagegen nur 74 MB/s (TCP: 85 MB/s). Offenbar ist STP anfälliger für die größere Latenzzeit der WDM-Strecke.

Die Arbeiten mit dem STP werden fortgesetzt. Insbesondere wurde die Verbesserung der Latenzzeit als Ziel formuliert, die aufgrund der aktuellen Implementierung ca. doppelt so groß ist wie bei TCP.

Verlängerung eines Fibre Channel

Für größere Rechnersysteme eignen sich SAN (Storage Area Network) -Speicher als Platten- und Magnetband-Hintergrundspeicher am bes-ten. Fibre Channel (FC) Protokolle auf Glasfaserleitungen stellen üblicherweise die notwendigen Datenverbindungen dazu her. Für den zukünftigen von ZIB und RRZN gemeinsam betriebenen verteilten Hochleistungsrechner sind jeweils SAN-Systeme vorgesehen. Von großem praktischen Wert ist die Eigenschaft, beide 315km voneinander entfernten Teil-SANs über die Gigabit-Ethernet-Strecke als ein großes SAN zu betreiben. Historisch gesehen lag die Längenbegrenzung eines Fibre Channels zunächst unter 1 km, bei Verwendung von dedizierten Singlemode Strecken um die 10 km und bei Geräten der Firma INRANGE Technologies GmbH (XCAF-Technik) bei 100 km.

Die Firma INRANGE Technologies GmbH erklärte sich bereit, gemeinsam mit dem ZIB und dem RRZN die Machbarkeit der Verlängerung eines Fibre Channel über diese 315 km lange Strecke zu verifizieren und stellte die notwendigen Geräte (2 Fibre Channel - Gigabit Ethernet Converter vom Typ SANValley SL 1000 und 2 Fibre Channel Switche vom Typ Inrange FC 9000/16) sowie das notwendige KnowHow zur Verfügung. Dieser FC-GE-Converter verfügt über einen Durchsatz von mehr als 8 GByte/s bei 4 FC-Verbindungen (bis 100 MByte/s full duplex) und 4 GE-Verbindungen mit Quality of Service Unterstützungen. Der Converter SL 1000 verwendet als Transportprotokoll UDP und nutzt ebenfalls die Eigenschaft der Jumbo-Frames, allerdings mit 3000 Byte pro Paket.

Da die neuen Rechner in ZIB und RRZN mit ihrem SAN noch nicht installiert sind, haben wir als Testkonfiguration eine FC-Verbindung zwischen einer SGI O200 und zwei Magnetbandstationen vom Typ STK 9840 gewählt. Damit beschränkte sich der Test auf Transfers mit einer Bandbreite in der Größenordnung von 30 MByte/s. Der normale Betrieb innerhalb des ZIB sieht vor, dass diese beiden Magnetbandstationen über eine direkte (kurze) FC-Verbindung bedient werden. In mehreren Schritten wurden die von der Fa. INRANGE bereit gestellten Geräte integriert: Zunächst lokal die beiden FC-Switche, dann ebenfalls lokal Auftrennen der FC-Verbindung zwischen den Switchen durch Einfügen der FC-GE-Converter und damit Transport auf IP-Ebene über Gigabit-Ethernet und schließlich Betrieb der Magnetbandstationen einschließlich eines Converters und eines FC-Switches im RRZN Hannover und damit Betrieb der FC-Kopplung über 315 km hinweg. Die Tabelle unten zeigt die gemessenen Transferraten für das Lesen von beiden Magnetbandgeräten und das Schreiben auf beide Magnetbandgeräte.

Beim Lesen der Magnetbänder führt die Konvertierung über IP-UDP Pakete auf dem Gigabit-Ethernet lediglich zu einer Reduktion der Bandbreite um 8 %. Erstaunlich ist die gemessene Gesamttransferrate beim Test der entfernten Anbindung einschließlich eines konkurrierenden Verkehrs von 97 MByte/s zwischen zwei SGI-Rechnern: Mit 989 Mbit/s werden die theoretisch möglichen 991 Mbit/s praktisch erreicht! Der drastische Abfall des Durchsatzes beim Beschreiben der Magnetbänder erklärt sich durch einen entstandenen Start-Stop-Betrieb der Magnetbandstationen; hier reichte die knappe Testzeit nicht aus, dies durch Veränderung einiger Parameter zu verbessern. Für die zukünftige SAN-Kopplung ist diese Eigenschaft nicht relevant.

Wir halten weitere Untersuchungen der in diesem Zusammenhang auftretenden Fragestellungen für dringend erforderlich und sind daran interessiert, weitere Studien durchzuführen. •

Hubert Busch, Sebastian Heidl, Ralf Staudemeyer
Dr. Manfred Stolle (von links nach rechts)
Konrad-Zuse-Zentrum für Informationstechnik Berlin (ZIB)
Takustrasse 7
D-14195 Berlin-Dahlem
e-mail: busch@zib.de

	Bandbreite beim Lesen	Bandbreite beim Schreiben
Lokale direkte Anbindung	28,8 MByte/s	26,4 MByte/s
Lokale Anbindung über FC-Switch	27,6 MByte/s	19,6 MByte/s
Lokale Anbindung mit Umsetzung auf IP über Gigabit Ethernet	27,6 MByte/s	19,6 MByte/s
Entfernte Anbindung auf dedizierter Leitung	27,0 MByte/s	6,0 MByte/s
Entfernte Anbindung mit konkurrierendem Verkehr	26,6 MByte/s	5,6 MByte/s