Ein paralleles Transportsubsystem für zellenbasierte Hochgeschwindigkeitsnetze

Torsten Braun
Institut für Telematik, Universität Karlsruhe, Zirkel 2, D-76128 Karlsruhe
z.Zt.: INRIA, 2004 route des Lucioles, F-06902 Sophia-Antipolis

1. Einleitung

Im Gegensatz zu den Netzwerken sind die als Bindeglied zwischen Rechnersystem und Netzwerk agierenden Kommunikationssysteme nicht in der Lage, die Anforderungen moderner verteilter Anwendungen in Form von hohem Durchsatz und geringer Verzögerung zu erfüllen. Einerseits können die zur Unterstützung einer sicheren Rechnerkommunikation notwendigen Kommunikationsprotokolle nicht schnell genug verarbeitet werden, andererseits stellt die Kopplung der Kommunikationssysteme an Netzwerke und Rechnersysteme einen gravierenden Engpaß dar. Insgesamt lassen sich die Leistungsschwächen auf unterschiedliche Ursachen wie ungeeigneten Protokollentwurf, ungünstige Implementierungsplattformen und die fehlende Abstimmung der Kommunikationssysteme auf unterliegende Netzwerke zurückführen. Außerdem wird häufig die Integration der Kommunikationssysteme in die Architektur und in die Betriebssystemumgebung von Rechnersystemen erschwert.
Zur Beseitigung des Leistungsengpasses bei der Protokollverarbeitung entwickelten sich im Forschungsbereich der Hochleistungskommunikation zwei klassische Ansätze. Zum einen können Kommunikationsprotokolle durch verbesserte Implementierungsstrategien, die Verwendung von Spezialhardware oder durch die Unterstützung von parallelen Implementierungsarchitekturen leistungsfähiger implementiert werden. Der zweite Ansatz versucht die Protokollverarbeitung durch den Einsatz sogenannter Hochgeschwindigkeitsprotokolle zu beschleunigen, die auf ihren Einsatz in modernen Netzumgebungen mit geringen Fehlerraten hin optimiert wurden.
Das Transportsubsystem PATROCLOS (Parallel Transport Subsystem for Cell-Based High-Speed Networks) kombiniert die beiden klassischen Lösungswege, d.h. es werden verschiedene Aspekte wie die Eigenschaften zellenbasierter Netze und die zur Realisierung auf parallelen Implementierungsarchitekturen geeignete Protokollarchitekturen integrierend betrachtet. Das Transportsubsystem erbringt unter Berücksichtigung zukünftiger Erfordernisse hinsichtlich neuer Dienste und Dienstqualitätsmerkmale einen erweiterten Transportdienst und besitzt dabei nicht nur die Funktionalität konventioneller Transportprotokolle, sondern übernimmt auch Funktionen der Sicherungs- und Vermittlungsschicht. Ein wesentlicher Entwurfsaspekt der entwickelten Protokollarchitektur besteht in der Berücksichtigung von Eigenschaften zellenbasierter Hochgeschwindigkeitsnetze auf ATM-Basis, die zur Unterstützung verschiedenartiger Dienste kleine Übertragungseinheiten fester Größe (Zellen) verwenden. Die Verarbeitung dieser Zellen stellt zwar hohe Leistungsanforderungen an Kommunikationssubsysteme, erlaubt aber auch eine effizientere Übertragung kleiner Datenmengen.

2. Architektur des parallelen Transportsubsystems

Ausgehend von Erfahrungen und Ergebnissen aus der Analyse verschiedener Standard- und sogenannter Hochgeschwindigkeitsprotokolle sowie aus Leistungsuntersuchungen an paralle-

len Implementierungen dieser Protokolle wurden Konzepte für den Entwurf des parallelen Transportsubsystems PATROCLOS erarbeitet. Dieses wurde in orthogonale, funktionale Einheiten in Form von autonomen Automaten unterteilt, womit die Dekomposition in parallel ausführbare funktionsbasierte Einheiten direkt effiziente Implementierungen auf parallelen Architekturen wie hybriden Multiprozessorsystemen unterstützt. Das Transportsubsystem stellt insgesamt ein Automatensystem dar, das sich aus einer Menge innerhalb einer Instanz kooperierender Automaten zusammensetzt. Automaten einer Instanz tauschen zu Synchronisations- und Kooperationszwecken Nachrichten aus.

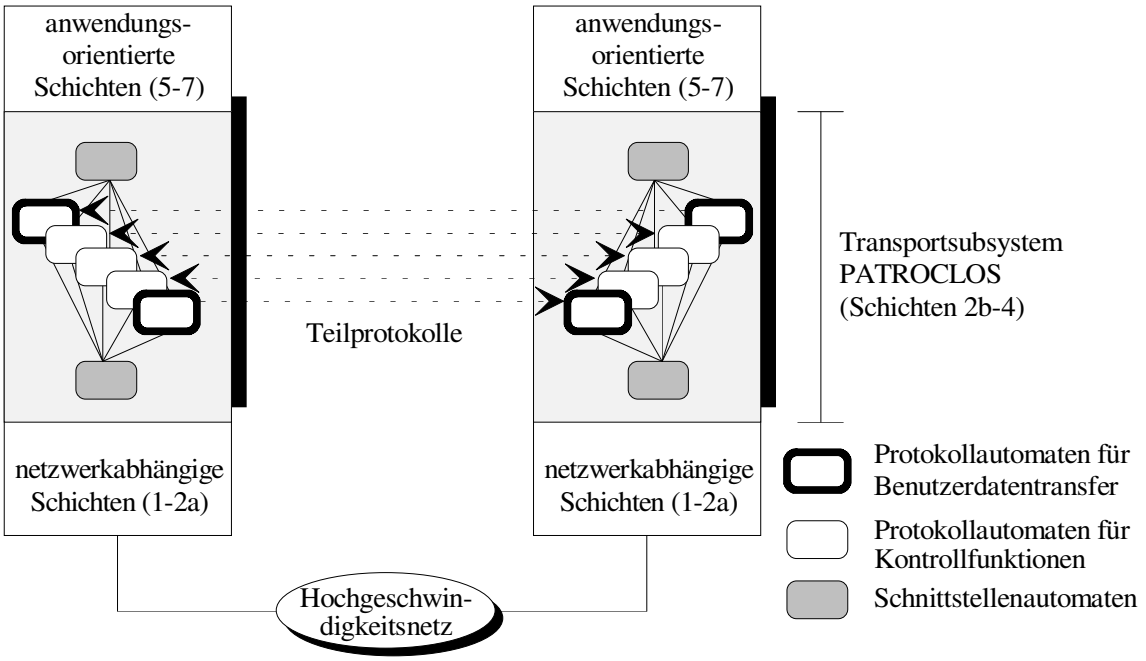

Abbildung: PATROCLOS-Architektur

Das Automatensystem unterscheidet Protokoll- und Schnittstellenautomaten (vgl. Abbildung). Schnittstellenautomaten unterstützen Interaktionen mit dem Transportdienstbenutzer bzw. mit dem unterliegenden Netzwerk. Sie sind lediglich zur lokalen Kommunikation innerhalb eines Systems notwendig und haben im Gegensatz zu Protokollautomaten keinen Automaten in der Partnerinstanz, mit dem sie kommunizieren. Im Transportdienstschnittstellen-Automaten ist dabei der erweiterte Transportdienst festgelegt, der sich stark an dem im OSI95-Projekt definierten Transportdienst orientiert. Die Art des Transportdienstes beeinflußt nur den Transportdienstschnittstellen-Automaten, für die Protokollautomaten ist die Art des Dienstes transparent. Der Transportdienstschnittstellen-Automat läßt sich aber auch leicht an andere Dienstformen anpassen.

Das Kommunikationsprotokoll zwischen Transportsubsystem-Instanzen setzt sich aus mehreren, voneinander weitgehend unabhängigen Teilprotokollen zusammen, wobei die Kommunikation zwischen den einzelnen Instanzen durch die direkte Kommunikation zwischen einzelnen, in der Regel gleichartigen Protokollautomaten erfolgt. Die Teilprotokolle verwenden jeweils spezielle Protokolldateneinheiten, um das Multiplexen von Protokolldateneinheiten verschiedener Protokollautomaten zu vermeiden und Koordinationspunkte zwischen Protokollautomaten einer Instanz zu minimieren. Die verwendeten Datenformate der Protokolldateneinheiten orientieren sich an der Zellengröße in zellenbasierten Netzen. Sie erlauben damit eine effiziente Übertragung und vereinfachen ihre Komposition und Analyse. Der vom Austausch der Benut-

zerdaten entkoppelte Kontrolldatenaustausch erlaubt eine weitgehende Unabhängigkeit von Steuer- und Datentransferfunktionen und damit deren parallele Verarbeitung.
Für jede Steuerfunktion sowie für die Datentransferfunktionen sind separate Protokollautomaten vorhanden. Der Austausch von Quittungen und Verbindungszustandsdaten zwischen Instanzen wird nicht durch gesendete oder empfangene Benutzerdatenpakete veranlaßt, sondern periodisch durchgeführt oder durch explizite Anforderungen bzw. Fehlersituationen initiiert. Da jeweils separate Protokollautomaten für die Sende- und die Empfangsrichtung einer Verbindung existieren, wird die Verarbeitung der beiden Datenflußrichtungen vollständig getrennt und kann damit parallel erfolgen. Die Teilprotokolle der einzelnen Protokollautomaten besitzen soweit erforderlich ihre eigenen Fehlerbehandlungsmechanismen und Zeitgeber, womit die in der Regel aufwendige Zeitgeberverwaltung auf mehrere Protokollautomaten verteilt wird. Die Protokollautomaten unterstützen jeweils eine Menge von Protokollfunktionen, von denen die Anwendung passende Funktionen auswählen und entsprechend parametrisieren kann.
Der *Verbindungsverwaltungs-Automat* übernimmt den impliziten oder Handshake-basierten Aufbau, die Überwachung und den normalen, geordneten oder Zeitgeber-gesteuerten Abbau von Verbindungen. Zwei Automaten unterstützen den *Benutzerdatenaustausch*, wobei das Segmentieren und Reassemblieren von Transportdienstdateneinheiten sowie die reihenfolgetreue Auslieferung wählbar sind. Zur *Quittierung und Übertragungswiederholung* (Go-Back-N-Strategie oder selektive Übertragungswiederholung) existiert jeweils ein Automat und auch die *Flußkontrolle* (Fenstermechanismus, Übertragungsratenkontrolle, Stop-And-Go-Verfahren oder Kombinationen) wird von je einem Automaten auf der Sende- und der Empfangsseite unterstützt. Bei den beiden *Staukontroll*-Automaten werden als Algorithmen zur Erkennung von Stausituationen innerhalb des Netzes die adaptive Zugangskontrolle und ein Paketpaarverfahren verwendet. Beide Algorithmen schätzen aus der Verzögerung von Testpaketen eventuelle Stausituationen innerhalb des Netzes ab.
Innerhalb einer Protokollinstanz kooperieren die einzelnen Automaten durch den Austausch von Signalen. Diese interne Kommunikation wurde auf ein Minimum reduziert, um die Abhängigkeiten zwischen den Automaten zu reduzieren und damit einen höheren Parallelitätsgrad zu erreichen. Beispielsweise verläuft der Signalaustausch zwischen den Automaten während der Datentransferphase in der Regel unidirektional. Signale werden außerdem nicht abhängig von eintreffenden oder zu sendenden Benutzerdatenpaketen, sondern periodisch generiert. Beim Senden von Benutzerdaten muß beispielsweise der Benutzerdatenaustausch-Sender-Automat dem Übertragungswiederholungsautomaten periodisch die im letzten Zeitintervall übertragenen Benutzerdaten mitteilen. Außerdem erhält der Benutzerdatenaustausch-Sender-Automat ebenfalls periodisch vom Flußkontrollautomaten die maximal zum Senden erlaubte Sequenznummer. Auf der Empfangsseite teilt der Benutzerdatenaustausch-Empfänger-Automat in gewissen Zeitabständen dem Quittierungsautomaten die angekommenen Benutzerdaten mit und schickt dem Flußkontrollautomaten Informationen über den Zustand die zur Speicherung der Benutzerdaten verfügbaren Puffer. In einer Erweiterung wurden inzwischen auch Multicast-Funktionen in das Transportsubsystem integriert, wobei hierzu lediglich Änderungen in den Automaten der Sendeseite erforderlich waren.
Ein wesentlicher Vorteil der funktionsbasierten Dekomposition in Protokollautomaten besteht auch in der sich aus der Modularität der Protokollarchitektur ergebenden leichten Konfigurierbarkeit und Anpaßbarkeit des Protokolls an Eigenschaften der unterliegenden Netzwerke sowie an unterschiedliche, sich eventuell dynamisch ändernde Anwendungsanforderungen. Dadurch, daß die einzelnen Protokollautomaten nur dedizierte, streng voneinander abgegrenzte Protokollfunktionen bearbeiten, wird die Selektion und die Parametrisierung einzelner Proto-

kollfunktionen wesentlich erleichtert. Es zeigte sich, daß die Auswahl von Protokollmechanismen und zugehöriger Parameter nur wenige, meist sogar nur einen einzigen Protokollautomaten beeinflußt. Zur Konfiguration können hierzu auch Werkzeuge des funktionsbasierten Kommunikationssubsystems F-CSS eingesetzt werden. Ein Protokollkonfigurator ermittelt dabei aus einer Beschreibung der geforderten Dienstgüte die dazu notwendigen Protokollfunktionen und -parameter für die PATROCLOS-Automaten.

3. Implementierung

Um die Leistungsfähigkeit der entwickelten Protokollarchitektur zu belegen, wurde für das Transportsubsystem eine hybride Multiprozessorarchitektur entwickelt, die sich aus Universalprozessoren für Funktionen auf Protokolldateneinheiten- und auf Transportdienstdateneinheiten-Ebene sowie aus dedizierten Hardwareeinheiten für Funktionen auf Zellenebene und zur Unterstützung der Parallelität zusammensetzt. Mit dem hybriden Ansatz soll eine hohe Leistungsfähigkeit bei gleichzeitig gegebener Flexibilität garantiert werden.
Die in Tabellenform sowie in der Spezifikationssprache PROMELA erstellte, formale Spezifikation des parallelen Transportsubsystemprotokolls wurde direkt in eine prototypische Implementierung umgesetzt, indem die Automaten auf Prozesse abgebildet wurden. Die Prozesse sind auf separaten Prozessoren bzw. Prozessormodulen ablaufbar und über Kanäle miteinander verbunden, welche auf Software- oder Interprozessorkanäle abgebildet werden. Ein Prozessormodul setzt sich dabei aus mindestens einem Prozessor, einer Hardwareeinheit zum Reassemblieren empfangener Zellen und einem gemeinsamen Speicher mit entsprechender Zugriffsteuerung auf den mit den Prozessoren des Moduls gemeinsam benutzten Bus zusammen. Die gemeinsamen Speicher stehen den Prozessoren für zu sendende oder empfangene Benutzerdaten zur Verfügung, insbesondere um Kopiervorgänge zwischen den Prozessoren zu vermeiden. Ein Hardware-Demultiplexer verteilt ankommende Zellen an die für deren weitere Verarbeitung zuständigen Prozessormodule, so daß diese Verarbeitung im folgenden parallel erfolgen kann. Das Segmentieren von Protokolldateneinheiten in Zellen und die entsprechenden Reassemblierfunktionen zum Anschluß des Transportsubsystems an zellenbasierte Netze wurden ebenfalls in Hardware implementiert. Für besonders zeitkritische Funktionen wie z.B. zur Speicherverwaltung wurden weitere Spezialhardwareeinheiten entwickelt, welche die Softwarerealisierungen ersetzen können.

4. Leistungsbewertung

Bei den Leistungsmessungen, die mit einem hybriden Monitorsystem durchgeführt wurden, kristallisierten sich die Prozesse zum Senden und Empfangen von Benutzerdaten mit 126 µs bzw. 159 µs als Engpässe heraus. Die Durchsatzergebnisse hängen jedoch nicht nur von der Zeit zur Verarbeitung eines Benutzerdatenpakets, sondern auch von der Häufigkeit des Austauschs von Kontrollinformationen mit anderen Prozessen ab. Beispielsweise muß der Daten-Sende-Prozeß in gewissen Abständen dem Übertragungswiederholungs-Prozeß Informationen über die gesendeten, d.h. vom Partner zu quittierenden Pakete zukommen lassen. Der Übertragungswiederholungs-Prozeß benötigt diese Informationen bevor er die entsprechende Quittung vom Empfänger erhält, d.h. der Austausch der Informationen sollte mindestens einmal pro Paketumlaufzeit erfolgen. Analog muß der Daten-Empfangs-Prozeß dem Quittierungs-Prozeß periodisch die empfangenen Daten mitteilen, damit dieser Quittungen erzeugen kann.

Beim Senden wird in der Regel ein Durchsatz von über 6000 Paketen/s (maximal 8000 Pakete/s) erzielt. Beim Empfangen werden üblicherweise mehr als 5000 Pakete/s erreicht, der Durchsatz ist hier auf 6300 Pakete/s begrenzt. Bei Duplex-Verkehr begrenzt der Tansportdienstschnittstellen-Prozeß den erreichbaren Durchsatz auf 5500 Transportdienstdateneinheiten pro Sekunde. Im Vergleich zu anderen parallelen Protokollimplementierungen auf Transputern stellen die erzielten Werte ein beachtliches Ergebnis dar. Auch der mit der parallelen Architektur erzielbare Leistungsgewinn gegenüber einer sequentiellen Verarbeitung liegt vergleichsweise hoch und bewegt sich üblicherweise zwischen 2,5 und 3,5 beim Senden sowie zwischen 2 und 2,5 beim Empfangen. Diese Werte übertreffen ebenfalls die Ergebnisse anderer paralleler Implementierungen.
Bei einer häufigen Durchführung von Kontrollfunktionen geht der Durchsatz sowohl bei der parallelen als auch bei der sequentiellen Implementierung zurück. Allerdings macht sich dieser Effekt bei der sequentiellen Implementierung wesentlich stärker bemerkbar. Da bei der parallelen Implementierung die Kontrollfunktionen parallel ausgeführt werden können, wird die Verarbeitung von Benutzerdaten und damit der erzielbare Durchsatz in einem wesentlich geringeren Maß beeinträchtigt. Die unterschiedliche Auswirkung der Kontrollfunktionen ist damit für die variierenden Speed-up-Werte verantwortlich. Gleichzeitig unterstreichen die Ergebnisse den Nutzen des Ansatzes, die Benutzerdaten und die Kontrollfunktionen zu trennen und parallel auszuführen.
Die Hardwarekomponenten lassen sich in einfachen PLDs implementieren und erreichen dennoch gute Leistungswerte. Die Komponenten zum Segmentieren und Reassemblieren können über 470000 Zellen/s verarbeiten, während der Demultiplexer mehr als 700000 Zellen/s and die angeschlossenen Prozessoren verteilen kann. Die hohe Leistung des Demultiplexers kommt dadurch zustande, daß sämtliche Kontrolloperationen parallel zum Speichern und Weiterleiten der eintreffenden Zellendaten vorgenommen werden können, so daß die Leistungsfähigkeit lediglich durch die Zugriffsgeschwindigkeit der eingesetzten Speicherbausteine begrenzt ist.
Zur weiteren Erhöhung der Leistungsfähigkeit der in Software implementierten Teile können schnellere Prozessoren wie digitale Signalprozessoren oder Transputer der neuen Generation eingesetzt werden, welche eine bis zu zehnfache Leistungssteigerung versprechen. Außerdem ist derzeit eine Implementierung auf einem Hochleistungs-Arbeitsplatzrechner in Bearbeitung. Die Trennung von Kontroll- und Datentransferfunktionen vereinfacht in diesem Zusammenhang die Erstellung einer Protokollfunktionsbibliothek, die leicht in Anwendungen eingebunden werden kann.
Desweiteren wird mit Hilfe von VHDL-Werkzeugen die für das Transportsubsystem entworfene VLSI-Implementierungsarchitektur weiterentwickelt, um noch höhere Leistungen zu erzielen und damit dem steigenden Bedarf an hochleistungsfähigen Kommunikationssystemen gerecht zu werden. Die PATROCLOS-Protokollarchitektur besitzt dadurch, daß keine von mehreren Funktionseinheiten, d.h. Automaten, gemeinsam genutzten Strukturen wie gemeinsame Verbindungszustandsdaten, Zeitgeber, Protokolldateneinheiten usw. existieren, bei einer Realisierung wesentliche Vorteile gegenüber konventionellen Protokollen.

Literatur

Braun, Torsten
Ein paralleles Transportsubsystem für zellenbasierte Hochgeschwindigkeitsnetze
Fortschritt-Berichte VDI, Reihe 10: Informatik/Kommunikationstechnik, Nr. 264, VDI-Verlag, Düsseldorf, 1993