Kommunikationshürde Speichersubsystem

Thomas M. Warschko, Joachim M. Blum und Walter F. Tichy
Institut für Programmstrukturen und Datenorganisation*
Universität Karlsruhe, Am Fasanengarten 5, D-76128 Karlsruhe

Zusammenfassung

Dieser Artikel beleuchtet den Zusammenhang zwischen der zu erwartenden Kommunikationsleistung, speziell dem Durchsatz, und der Speicherbandbreite einer Rechnerarchitektur beim Einsatz verschiedener Konzepte zur Realisierung von Kommunikationsprotokollen (One-copy, One-copy-pipelining und Zero-copy).

Die Ergebnisse zeigen, daß *Copy*-Protokolle entweder enorm hohe Anforderungen an das Speichersubsystem stellen, oder unter inakzeptablen Leistungswerten leiden. Lediglich *Zero-Copy*-Protokolle ermöglichen es die volle Leistung einer Rechnerarchitektur auszuschöpfen. Einige der untersuchten Systeme können Netzwerke nur bis zu einer maximalen Übertragungskapazität von 300 MBit/s bzw. 500 MBit/s handhaben. Lediglich ausgesuchte Systeme sind mit den Kommunkationsstrukturen heutiger Betriebssysteme in der Lage 1 GBit/s-Netzwerke zu bedienen. Dieses Ergebnis ist insofern von Bedeutung, da die Copy-Problematik mit dem Einzug von Gigabit-Netzwerken in Arbeitsplatzrechner in Kürze stark an Aktualität gewinnen wird.

1 Einleitung

Mit der zunehmenden Verfügbarkeit von Hochgeschwindikeitsnetzwerken im Gigabit-Bereich (z.B: Myrinet [BCF+95], Gigabit-Ethernet [Tho97], FibreChannel [Jur95], MemoryChannel-2 [FG97], SCI [IEE92], 1244Mbit-ATM [PB94], ...) stellt sich die Frage, welchen Datendurchsatz man in realen System erwarten kann. Neben der verwendeten Netzwerkhardware und dem eingesetzten Bussystem, meist dem PCI-Bus, spielt die Konzeption des Kommunikationsprotokolls die entscheidende Rolle für die Kommunikationsleistung des Gesamtsystems. Hierbei können drei Varianten unterschieden werden:

- **One-copy Protokolle:** Die Daten werden zunächst vom Applkationsadreßraum in den Kernadreßraum mittels einer Kopieroperation übertragen und anschließend per DMA-Operation an die Netzwerkhardware weitergeleitet.
- **One-copy-pipelining Protokolle:** Im Gegensatz zu den One-copy-Protokollen wird hier die Kopieroperation des Auftrags k mit der DMA-Operation des Auftrags $k-1$ simultan ausgeführt. Diese Variante stellt damit höhere Anforderungen an das Speichersubsystem als das einfache One-copy-Protokoll.

*Now: Scarasoft AG, Mühlfelder Straße 10, 82211 Herrsching, Germany. Email: {warschko, blum}@scarasoft.com

- **Zero-copy Protokolle:** Die Daten werden direkt aus dem Applikationsadreßraum mittels einer Kopier- oder DMA-Operation in die Netzwerkhardware übertragen (sog. *User-Level-Kommunikation*).

Im folgenden soll die Frage beantwortet werden, wie sich die zu erwartende Kommunikationsbandbreite in Abhängigkeit des Kommunikationsprotokolls, des Bussystems, sowie der Speicherbandbreite verhält. Dazu wird im nächsten Abschnitt ein Modell eingeführt, das all diese Faktoren berücksichtigt, um dann die zu erwartende Leistung der verschiedenen Protokollvarianten zu analysieren. In Abschnitt 3 werden dann die theoretischen Untersuchungen in Bezug zur Leistung realer Systeme gesetzt. Abschnitt 4 faßt die erzielten Ergebnisse noch einmal zusammen.

2 Modellierung

Abbildung 1 zeigt das Architekturmodell anhand dessen alle weiteren Überlegungen durchgeführt werden.

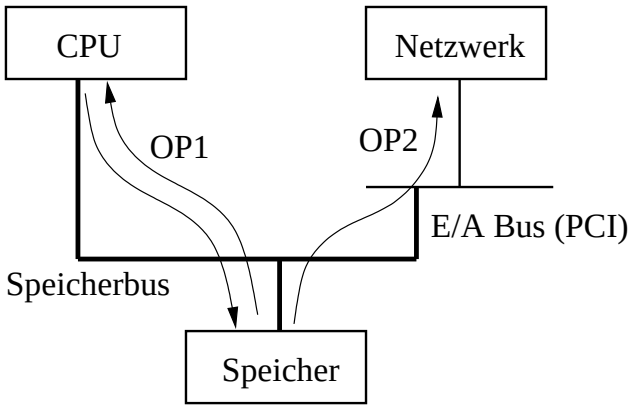

Abbildung 1: Architekturmodell

Die modellspezifischen Parameter umfassen die I/O-Bus-Bandbreite (B_{PCI}), die theoretische Speicherbandbreite (B_{MEM}), sowie die tatsächliche Bandbreite einer Kopieroperation (B_{CPY}). B_{MEM} und B_{CPY} sind insofern korreliert, als daß gilt: $B_{CPY} \leq B_{MEM}/2$, da bei einer Kopieroperation die Daten zweimal über den Speicherbus transferiert werden müssen. Anhand der Operationen *OP1* (Kopieroperation) und *OP2* (DMA-Transfer) aus Abbildung 1 lassen sich die verschiedenen Protokollvarianten charakterisieren:

- One-copy: Die Operationen 1 und 2 werden getrennt nacheinander ausgeführt.
- One-copy-pipelining: Die Operationen 1 und 2 werden parallel zueinander ausgeführt (bei zwei aufeinanderfolgenden Kommunikationsoperationen).
- Zero-copy: Es wird nur Operation 2 ausgeführt.

Im folgenden soll die zu erwartende Kommunikationsbandbreite (B_{COM}) unter den gegebenen Randbedingungen bei variierender Speicherbandbreite (B_{MEM}) untersucht werden. Für die

theoretischen Untersuchungen wird dabei die halbe Speicherbandbreite ($B_{MEM}/2$) als obere Schranke für die Bandbreite einer Kopieroperation (B_{CPY}) verwendet. Ein Vergleich mit real gemessenen Werten findet in Abschnitt 3 statt.

2.1 One-copy Protokolle

Da die beiden Operationen OP1 und OP2 nacheinander ausgeführt werden, berechnet sich die Kommunikationsbandbreite aus dem Kehrwert der Summe der Übertragungszeiten der beiden Einzeloperationen.

$$B_{COM} = \frac{1}{\frac{1}{B_{PCI}} + \frac{1}{B_{CPY}}}$$

Das charakteristische an dieser Funktion ist, daß die Kommunikationsbandbreite kleiner als das Minimum der beiden Einzelbandbreiten ist ($B_{COM} \leq \min(B_{PCI}, B_{CPY})$).

2.2 One-copy-pipelining Protokolle

Bei dieser Protokollvarianten werden die beiden Operationen *OP1* und *OP2* simultan ausgeführt und wie in Fließbandsystemen üblich, bestimmt das schwächste Glied der Kette die Leistung des Gesamtsystems. Somit berechnet sich die Kommunikationsbandbreite als Minimum der PCI-Bus-Bandbreite (OP2), der Bandbreite der Kopieroperation (OP1) und einem Drittel der Speicherbandbreite.

$$B_{COM} = \min(B_{PCI}, B_{CPY}, B_{MEM}/3)$$

Der dritte Parameter ($B_{MEM}/3$) ist notwendig, da bei der Überlagerung der Operationen OP1 und OP2 insgesamt drei Datenströme um den Speicherbus konkurrieren. Erstaunlicherweise stellt dieser Term bei geringen Speicherbandbreiten den beschränkenden Faktor dar.

2.3 Zero-copy Protokolle

In dieser Protokollvarianten werden lediglich Operation OP2 ausgeführt, und die Kommunikationsbandbreite berechnet sich aus dem Minimum der PCI-Bus-Bandbreite und der Speicherbandbreite.

$$B_{COM} = \min(B_{PCI}, B_{MEM})$$

Im Gegensatz zu den beiden anderen Protokollvarianten stellt dieses Protokoll die geringsten Anforderungen an das Speichersubsystem und liefert erwartungsgemäß die besten Durchsatzraten.

2.4 Vergleich der Protokollvarianten

Abbildung 2 zeigt das Verhalten (B_{COM}) der verschiednen Protokollvarianten bei Variation der Speicherbandbreite B_{MEM}.

Wie erwartet steigt die Leistung des Zero-copy Protokolls linear mit der Speicherbandbreite bis zur maximalen PCI-Bus-Bandbreite an. Das One-copy-pipelining Protokoll dagegen steigt lediglich mit einem Drittel der Speicherbandbreite und erreicht das Maximum (133MB/s) folgerichtig ab einer Speicherbandbreite von ca. 400MB/s. Das One-copy Protokoll

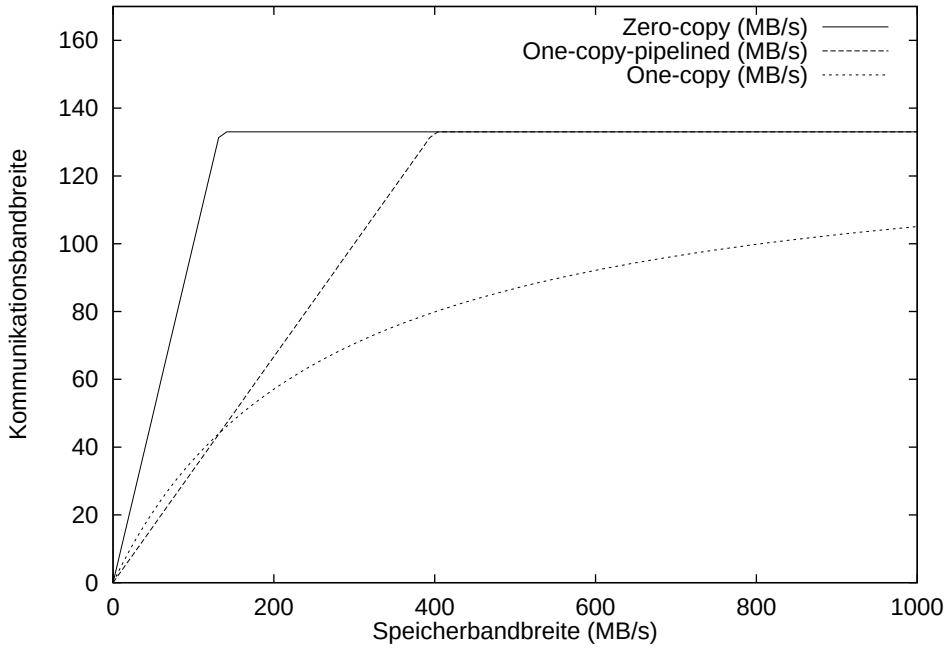

Abbildung 2: Kommunikationsleistung der Protokollvarianten (133MB/s PCI-Bus)

zeigt für kleine Speicherbandbreiten (< 133MB/s) aufgrund der geringeren Anforderungen an das Speichersubsystem eine etwas bessere Leistung als das One-copy-pipelining Protokoll; für größere Speicherbandbreiten jedoch nähert sich die Kurve nur sehr langsam der Maximalleistung an.

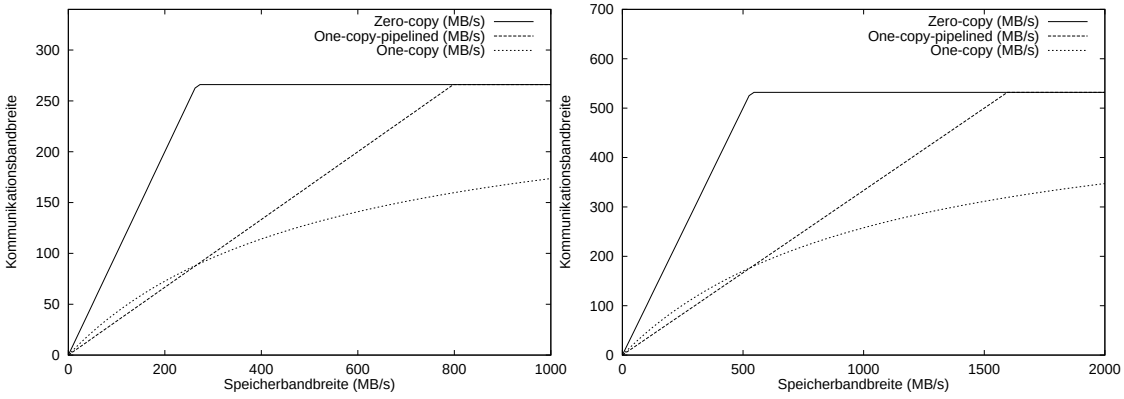

Abbildung 3: Kommunikationsleistung bei 266 MB/s und 532 MB/s PCI-Bus

Die theoretischen Untersuchungen lassen zunächst keinen Leistungsengpaß innerhalb des Speichersubsystems vermuten. Das einfache One-copy Protokoll weist jedoch auch bei extrem hohen Speicherbandbreiten (> 1GByte/s) keine ausreichende Kommunikationsbandbreite auf. Das One-copy-pipelining Protokoll erreicht zwar die maximale PCI-Bus-Bandbreite, benötigt dazu jedoch eine Speicherbandbreite von der dreifachen PCI-Bus-Bandbreite. Bei dem derzei-

tigen PCI-Bus (32bit, 33MHz) mit 133MB/s sollten die benötigten 400MB/s Speicherbandbreite kein Problem darstellen; bei schnelleren Bussystemen (z.B. PCI mit 64 bit und/oder 66Mhz) würden mit 800MB/s bzw. 1600MB/s Speicherbandbreiten benötigt (siehe Abbildung 3), die von den wenigsten Systemen heutzutage erbracht werden können. Für Zero-copy Protokolle dagegen reicht die einfache Speicherbandbreite aus und deshalb ist diese Protokollvariante zur Ansteuerung von Hochgeschwindigkeitsnetzwerken im Gbit-Bereich den anderen Alternativen vorzuziehen.

3 Leistung realer Systeme

Die bisher gewonnenen Erkenntnisse gehen davon aus, daß Kopier- bzw. DMA-Operationen auch die volle Speicherbandbreite des Systems ausnutzen können. Wie sich diese Annahme gegenüber Messungen an realen Systemen verhält, ist Gegenstand dieses Abschnitts.

Tabelle 1 zeigt die gemessenen Leistungswerte dreier unterschiedlicher Systeme: (a) Alpha 164LX mit 600MHz 21164A Prozessor, 10ns SDRAM, 100MHz Boardtakt, 33MHz PCI-Takt, 128bit Speicherbus, (b) Dual Pentium 200MMX, GigaByte GA-586DX Board, 60ns EDO-Ram, 66MHz Boardtakt, 33MHz PCI-Takt, 64bit Speicherbus und (c) Pentium-II 400, ASUS P2BLS Board, 100MHz Boardtakt (BX-Chipsatz), 7ns SDRAM, 33MHz PCI-Takt, 64bit Speicherbus. Aus diesen Systemparametern ergibt sich eine maximale Speicherbandbreite von 1600MB/s für die Alpha (Speicherbusbreite * Boardtakt) sowie 528MB/s bzw. 800MB/s für die PCs. Da alle Systeme den PCI-Bus mit 33MHz takten, beträgt der maximale PCI-Bus-Durchsatz 133MB/s.

System	B_{MEM} peak	B_{CPY} real	DMA-In	DMA-Out
Alpha 600MHz, LX	1600MB/s	125MB/s	67MB/s	120MB/s
DualPentium 200, GA586 DX	528MB/s	42MB/s	40MB/s	40MB/s
Pentium-II 400, BX	800MB/s	125MB/s	124MB/s	124MB/s

Tabelle 1: Leistungswerte realer Systeme

Die gemessenen Werte weichen bei allen Systemen zum Teil erheblich von den Erwartungswerten ab. Erschreckend sind die geringen Bandbreiten der Kopieroperationen[1], die in zwei Systemen nicht einmal 16% des Nominalwertes (800MB/s bzw. 266MB/s) erreichen und beim dritten System bei 32% des Nominalwertes (400Mb/s) liegt. Aber auch die gemessenen DMA-Leistungen bleiben außer beim dem Pentium-II System hinter den Erwartungen zurück. Während die Alpha beim Transfer zum Netzwerk[2] mit 67MB/s die halbe PCI-Bus-Bandbreite und beim Transfer zum Hauptspeicher mit 120MB/s immerhin 90% der PCI-Bus-Bandbreite erreicht, fallen die gemessenen Werte auf dem Pentium-PC mit 40MB/s recht dürftig aus. Dies dürfte am verwendeten Chipsatz (430HX) in den PC's liegen, denn andere Chipsätze erreichen fast ideale PCI-Bus-Bandbreiten.[3] Die Meßwerte des Pentium-II Systems (440BX Chipsatz) bestätigen dies.

[1] Diese Werte wurden mit dem `bw_mem_cp` Programm aus der *lmbench V1.1*-Benchmarksammlung ermittelt.

[2] Diese Werte wurden mit dem `hswap`-Programm aus der Myrinet-Software ermittelt.

[3] zum Vergleich siehe: http://www.myri.com/myrinet/performance/DMAperf.html oder http://www-csag.cs.uiuc.edu/projects/comm/hcl.html

Nimmt man die gemessenen Leistungsdaten als Basis zur Berechnung der Kommunikationsleistung nach Abschnitt 2, so ergibt sich folgendes Bild (siehe Tabelle 2).

System	One-copy	One-copy-pipeling	Zero-copy
Alpha	44MB/s	67MB/s	67MB/s
Pentium	21MB/s	40MB/s	40MB/s
Pentium-II	64MB/s	125MB/s	125MB/s

Tabelle 2: Maximale Kommunikationsleistung

Der recht deutliche Vorteil eines Zero-copy Protokolls aus den theoretischen Betrachtungen spiegelt sich in keinster Weise in den Ergebnissen nach Tabelle 2 wider. Die ist vor allem in der schlechten DMA-Leistung der untersuchten Systemen (Alpha und Pentium) begründet. Das Pentium-II System dagegen verfügt über eine ausgewogene Systemleistung, so daß sich bei einem 33MHz, 32-Bit PCI-Bus kein Leistungsengpaß einstellt. Beim Übergang auf eine schnellere PCI-Bus Variante (266 MByte/s oder 532 MByte/s) wird sich auf diesem System der Vorteil der Zero-Copy Protokolle einstellen, da dann die geringe Speichbandbreite der Kopieroprtationen (125 MByte/s) ihre Auswirkungen zeigt.

Die Ergebnisse aus dem Modell als auch die Meßwerte an konkreten Systemen belegen, daß bei den heute eingesetzten Netzwerken (10Mbit Ethernet, 100Mbit FastEthernet oder 155Mbit ATM) das Speichersubsystem noch keinen Engpaß darstellt. Werden jedoch in Zukunft Gigabit-Netzwerke mit Übertragungsraten von 100MB/s und mehr eingesetzt, dann wird deutlich, daß mit den Transferleistungen heutiger Systeme es nicht möglich sein wird, die Kapazitäten der künftigen Hochleistungsnetzwerke auszunutzen.

4 Konklusion

Die Analyse verschiedener Ansätze zur Konzeption von Kommunikationsprotokollen hat gezeigt, daß (a) einfache One-copy Protokolle im allgemeinen unzureichende Leistungswerte aufweisen, (b) One-copy-pipelined Protokolle bei entsprechender Speicherbandbreite akzeptable Leistungswerte liefern und (c) Zero-copy Protokolle die besten Leistungsdaten aufweisen und zudem noch unabhängig von der Bandbreite von Kopieroperationen agieren.

Die Analyse realer System jedoch hat gezeigt, daß sich die Problematik von Kopieroperationen in den untersuchten Systemen nicht stellt, da diese Systeme weit hinter den Leistungswerten der Datenblätter zurückbleiben. Die erreichten Leistungsdaten reichen aus, um 100Mbit/s bis 200Mbit/s Netzwerke zu unterstützen; dringt man jedoch in den Gigabit-Bereich vor, dann gilt es das einzusetztende Zielsystem sehr genau auszuwählen, falls die Leistung eines Gigabit-Netzwerkes ausgeschöpft werden soll.

Literatur

[BCF+95] Nanette J. Boden, Danny Cohen, Robert E. Felderman, Alan E. Kulawik, Charles L. Seitz, Jarov N. Seizovic, and Wen-King Su. Myrinet: A Gigabit-per-Second Local Area Network. *IEEE Micro*, 15(1):29–36, February 1995.

[FG97] Marco Fillo and Richard B. Gillett. Architecture and Implementation of MEMORY CHANNEL 2. *Digital Technical Journal*, 9(1):27–41, 1997.

[IEE92] IEEE. *IEEE – P1596 Draft Document. Scalable Coherence Interface Draft 2.0*, March 1992.

[Jur95] Clint Jurgens. Hot topics: Fibre Channel: a connection to the future. *Computer*, 28(8):88–90, August 1995.

[PB94] Philip Dumortier, Luc Van Hauwermeiren and Joost Boerjan. Transport of gigabit ATM cell streams over lower order SDH backbone. In *Proceedings of the 13th Annual Joint Conference of the IEEE Computer and Communciations Society on Networking for Global Communciations. Volume 3*, pages 1160–1169, Los Alamitos, CA, USA, June 1994. IEEE Computer Society Press.

[Tho97] Geoffrey O. Thompson. Standards: Work progresses on gigabit Ethernet. *Computer*, 30(5):95–96, May 1997.