オペレーティング・システム 35－3
（１９８７．６．１２）

１００Ｍｂ／ｓ光ファイバリングネットワークを用いたプロセッサ間結合方式

木村行男、星子隆幸
ＮＴＴ 電気通信研究所

センタシステムにおける計算機の処理量増大、システムの段階的成長、高信頼化等の要望に応えて、多数台の疎結合マルチプロセッサシステムを構築するため、伝送速度１００Ｍｂ／ｓの光ファイバリングネットワークを開発した。

本ネットワークのノードであるプロセッサ間結合装置の通信効率に関連する特徴的な機能とその適用効果は次の通りである。（１）複数サブチャネルを用いて、個々の通信プロセス間で直接転送を行うことにより、通信処理プログラムのダイナミックステップ数を削減（２）通信待ち合せ機能適用によるアテンション割込み処理時間の削除（３）分割転送方式により、長データをチャネル速度で転送し、プログラムに対するデータ長制限を撤廃

本稿では、上記プロセッサ間結合装置の機能およびＰＣＩの実効データ転送能力を紹介し、上記効果を上げるためにソフトウェアで対処すべき事項を述べる。

Processor to Processor Communication System Using a 100Mb/s Optical Token Ring Network

Yukio Kimura and Takayuki Hoshiko
NTT Electrical Communications Laboratories
1-2356, Take, Yokosuka-shi, Kanagawa-ken, 238-03 Japan

For the purpose of constructing a loosely couppled multiprocessor system which can meet rapid increases in demands for increased processing power, step-by-step growth of systems, and higher reliability in center systems, we have developed a 100Mb/s optical fiber ring network.

Principal functions of this ring network node (processor-to-processor communication interface unit : PCI) and its application effects are as follows;

(1)Reduction of dynamic steps of a communication control program by communicating directly between any pairs of communication processes using multiple subchannels.

(2)Deletion of time for processing attention interruption by preissueing a communication synchronizing command.

(3)Reduction of time for transferring long data and removal of data length limit by using the split transfer method.

This paper describes the above functions, actual data transfer ability, and ways of deriving the effects of the above functions by software.

１．はじめに

データ通信サービスの高度化、システムの段階的成長にともない、センタシステム内のプロセッサ（ＨＯＳＴ，ＦＥＰ等）間の高速な接続が要求されるようになってきている。この要求に応えて多数台の疎結合マルチプロセッサシステムを構築するため、伝送速度１００Ｍｂ／ｓ光ファイバリングネットワークを開発した。

本稿では、１００Ｍｂ／ｓ光ファイバリングネットワークのノードであるプロセッサ間結合装置（ＰＣＩ；Ｐｒｏｃｅｓｓｏｒ－ｔｏ－ｐｒｏｃｅｓｓｏｒ Ｃｏｍｍｕｎｉｃａｔｉｏｎ Ｉｎｔｅｒｆａｃｅ ｕｎｉｔ）について、その特徴的な動作の概要および性能評価結果を紹介する。またこのＰＣＩを用いて通信性能向上を図るためにプログラムで対処すべき事項について述べる。

２．１００Ｍｂ／ｓ光ファイバアリングネットワークの概要

本稿で述べる１００Ｍｂ／ｓ光ファイバネットワークは、以下のねらいのもとに開発した。

（１）センタシステムの処理能力拡大に対して柔軟に対処可能なシステム拡張性の提供

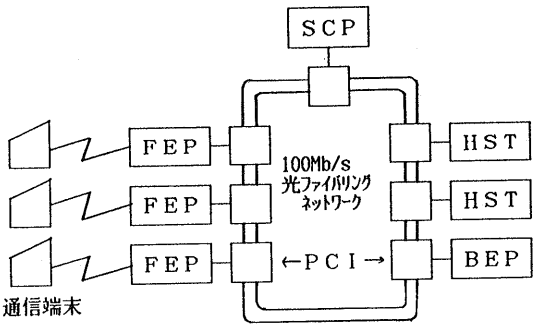
（２）経済的なセンタ内プロセッサ間通信の実現

（３）センタシステムの長時間連続運転とシステム障害時の高速リカバリの実現

本ネットワークを用いたセンタシステム構成例を図１に示す。

２．１　主な物理仕様及びデータリンク仕様

本ネットワークのネットワークノード（ここではＰＣＩとよぶ）とプロセッサとの接続については、中大型汎用計算機において多く使われている高速で標準的なチャネルインタフェース（Ｉ／Ｏインタフェース）を用いている。ネットワーク伝送路については、伝送速度１００Ｍｂ／ｓで、センタ内接続の他、隣接センタ間接続も考慮してＰＣＩ間距離が



ＰＣＩ：ﾌﾟﾛｾｯｻ 間結合装置

ＳＣＰ：ｼｽﾃﾑ制御ﾌﾟﾛｾｯｻ

ＨＳＴ：ﾎｽﾄ ﾌﾟﾛｾｯｻ

ＦＥＰ：ﾌﾛﾝﾄｴﾝﾄﾞﾌﾟﾛｾｯｻ(通信制御ﾌﾟﾛｾｯｻ)

ＢＥＰ：ﾊﾞｯｸｴﾝﾄﾞﾌﾟﾛｾｯｻ

図１　光ファイバリングネットワークによるセンタシステム構成例

最大２ｋｍまで可能な光ファイバ伝送方式を用いている。ネットワークトポロジーは、１００Ｍｂ／ｓの高速転送を多数台（数１０台以上）のＰＣＩ間で光を用いて信頼性良くかつ経済的に実現可能な、リング形式を用いた。ネットワークアクセス方式は、伝送効率の良いトークンパッシング方式を採用している。

リングネットワークは、ＰＣＩまたは伝送路障害がシステム全体の障害につながる性質をもっているため、本ネットワークでは伝送路を２重化するとともに代替パス切り替え、ループバック切り替えによる障害箇所自動切り離し、送信権自動回復を行っている。これらの動作は特定のＰＣＩによる集中制御ではなく完全分散制御であり、ハードウェアで自動的におこなわれる。

２．２　プロセッサ間通信方式

高速のプロセッサ間通信実現のためには、伝送路の効率を向上させるだけでなく、伝送路上を流れるトラヒックの制御の改善、すなわち通信制御に要するオーバヘッドを削減し、ハードウェア・ソフトウェア処理双方の高速化が必要である。そのため、本方式では通信制御に要するソフトウェア処理高速化に有効

な機能をＰＣＩでサポートしている。

この機能は、複数サブチャネルを用いた多重通信方式による通信プロセス間直接転送機能であり、１対１チャネル結合のプロセッサ間通信で使用されている機能をＮ対Ｎプロセッサ間通信用に拡張したものである。（図２参照）

以下に本機能ならびに本機能をＮ対Ｎ環境で用いるための通信先アドレス指定およびデータ転送方法、分割転送方式等特徴的な機能について述べる。

（１）サブチャネル多重による通信プロセス間直接転送機能

本機能は、通信プロセス毎にサブチャネルを用いて通信パスを特定し（サブチャネル多重通信方式）、送受信バッファを個々の通信プロセス側で用意して、該バッファ間でのデータの直接転送を行うものであり、同一送受信バッファの連続使用が可能なため、バッファエリアのフィックス化／フロート化を最適な契機で実施することが可能である。すなわち、高トラヒック、長データ転送の環境下で送受信バッファ管理や割り込み処理、データ移送等の通信処理に要するダイナミックステップ数を削減することができる。

なお、多重通信の効率化のために、各サブチャネル対応にチャネルを早期に解放し、アテンション割り込みを伴わずに相手からの通信起動を待ち合わせる通信待ち合わせ機能も併せてサポートしている。

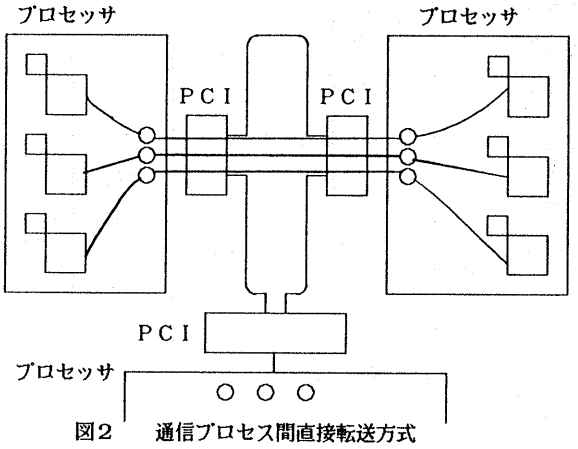


図２　通信プロセス間直接転送方式

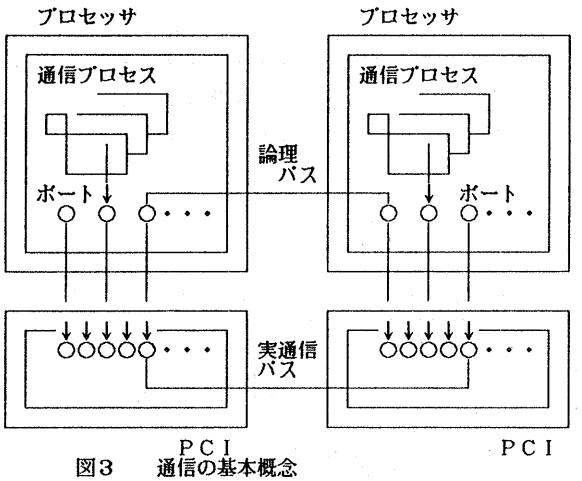


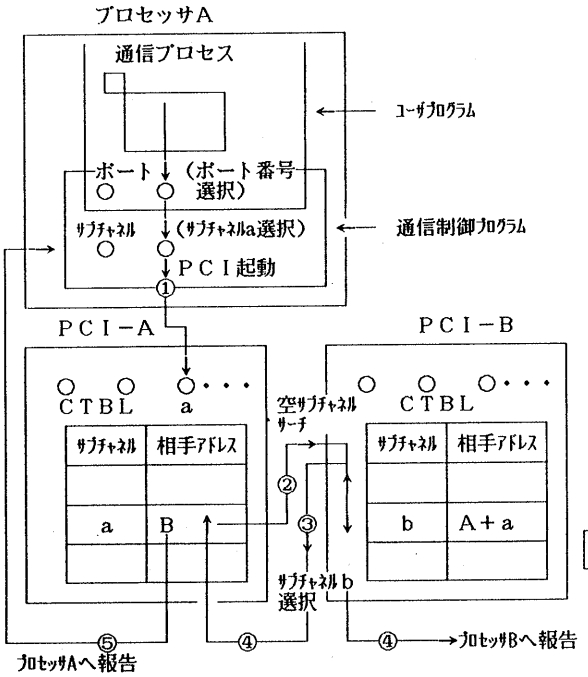
図３　通信の基本概念

（２）通信先アドレス指定

サブチャネル多重による通信プロセス間直接転送機能をＮ対Ｎ通信環境で用いる場合の対処として通信相手のノードアドレスとサブチャネルアドレスを、チャネル仕様下で効率的に指定する通信先アドレス指定法を以下に示す。

本プロセス間通信は、図３に示すように通信プロセス間の仮想回線である論理パスと論理パスの窓口であるポート及びＰＣＩ間のデータ転送を行う実通信パスを用いて行われるものとする。個々の通信プロセスは論理パスを指定して通信を行うが、通信に先立って論理パスを開設する必要がある。各プロセッサ内に論理パスを開設するための窓口として、ポートを設定する。論理パスはポート名の対で意識される。ポート名はＰＣＩアドレス＋ポート番号で指定される。ＰＣＩ間のデータ転送は実通信パスを指定して行われる。実通信パスは送受信双方のＰＣＩアドレス＋ＰＣＩサブチャネルアドレスで識別される。実通信パスの指定にあたり、現行チャネル仕様では自ＰＣＩのサブチャネルをデバイスとして指定できるのみなので、前もって実通信パスを設定し、自ＰＣＩサブチャネル指定のみで通信できれば効率的である。このため、実通信パスを記憶し管理するパス管理テーブル（ＣＴＢＬ）をＰＣＩ内に設置した。通信時の

前処理として、専用コマンドによりＣＴＢＬを設定し実通信パスを確立する。通信をともなうコマンド（Ｗｒｉｔｅコマンド、Ｒｅａｄコマンド等）発行時、相手アドレスはＣＴＢＬ検索によりＰＣＩハードが自動設定する。この実通信パス設定の動作例を図４に示す。



（実通信パス設定手順）

（１）通信制御プログラムがＰＣＩ－Ｂとの実通信パス設定用に選択したＰＣＩ－Ａのサブチャネルａに対して通信パス設定コマンド（ＳＰＡ）により、相手ＰＣＩアドレス情報Ｂを転送する。

（２）サブチャネルａに対応するＣＴＢＬエントリの相手アドレス域にＢを設定し、上記ＳＰＡをＰＣＩ－Ｂに転送する。

（３）ＳＰＡを受信したＰＣＩ－Ｂは、ＣＴＢＬ内で相手アドレスが確立していない任意のサブチャネルｂを選択し、ｂの相手アドレス域にＡ＋ａを設定する。

（４）ＰＣＩ－Ｂのサブチャネルｂに対応する相手アドレスが確立したことをＰＣＩ－Ｂの上位プロセッサおよびＰＣＩ－Ａに報告

（５）ＰＣＩ－ＡはＰＣＩ－Ｂからの上記報告を基に、サブチャネルａの相手アドレスとしてｂを追加設定し、一連のＳＰＡコマンド処理が完了したことをＰＣＩ－Ａの上位プロセッサに報告する。

図４　　実通信パス設定動作例

（３）データ転送

チャネル結合されたプロセッサ間通信においては、送受信双方のプログラムが発行する通信用コマンドで同期をとってデータ転送が行われる。このため、ＰＣＩ－ＰＣＩ間のデータリンク手順としては、単に上記通信用コマンド、データ、関連応答情報のみを伝達するだけの手順を用いている。ＰＣＩで採用している通信用コマンドレベルのデータ転送動作例を図５に示す。

（４）分割転送方式

ＰＣＩ－ＰＣＩ間のデータ転送は、長データ時のメッセージ通過時間を短くするために、分割転送方式を用いている。分割転送方式は、図６に示すようにＰＣＩ内の交替バッファを用いて、プロセッサ－ＰＣＩ間のデータ送受信とＰＣＩ－ＰＣＩ間のデータ送受信を並行動作させるものである。転送データ長が長くなると、同期化に要する時間がデータ転送時間に比べて無視できる程小さくなるため、チ

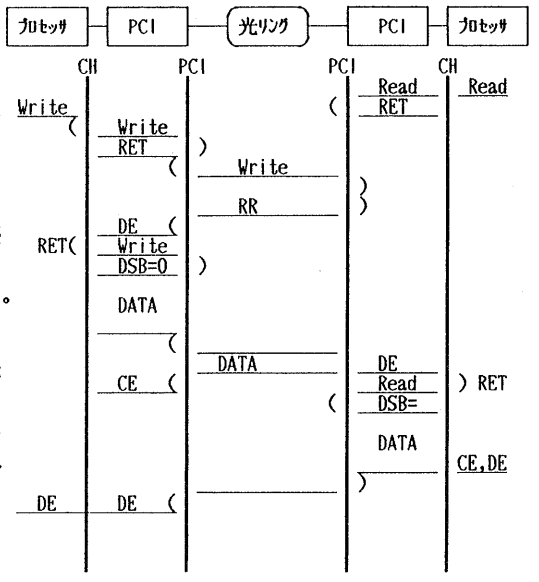
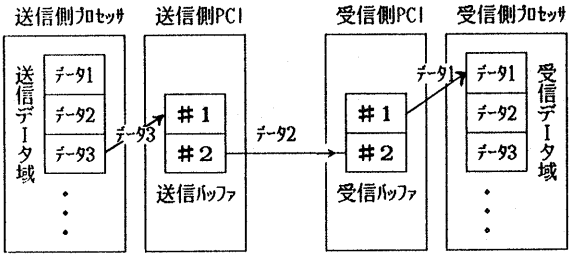


図５　データ転送動作

ャネル転送速度でデータ転送を行うことが可能になる。この方式は、ハードウェアが一定長にデータを分割するために、プログラムが指定するデータ長に制限をつける必要がない。



（注）ＰＣＩ内バッファは交替バッファ方式をとっている。

図6　分割転送方式

図８に示す。

図７は転送データ長を変化させた場合の性能との関係である。チャネルとＰＣＩ間のデータ転送時間を除くＰＣＩハード動作時間が変わらないため、転送データ長の増加に伴い性能は向上し、各々チャネル転送速度で飽和する。

図８は短データ（２００Ｂ）転送時のソフト走行時間に対するＰＣＩ性能との関係である。高速のプロセッサの場合は、ソフト走行時間増加の影響は少ないが、低速プロセッサでは影響が大きい。低速プロセッサにＰＣＩを接続する場合にはソフトウェアダイナミックステップ数の削減が重要であることを示している。

３．ＰＣＩ性能

ＰＣＩの性能として、送受信一対のＰＣＩ間での単位時間内の転送データ長をとると、性能は次式で近似される。

$$\text{ＰＣＩ性能} = \frac{\text{転送データ長}}{\text{ソフト走行時間}+\text{ＰＣＩハード動作時間}}$$

ここで転送データ長とは、１回のコマンドで転送されるデータの長さである。またＰＣＩハードウェア動作時間（Ｔ）は、１コマンドを処理する際にＰＣＩのハードが占有される時間であり、転送データ長の関数として次式で示される。

$$T = A + B + C + D$$

Ａ：ＰＣＩファームウェア処理時間
Ｂ：チャネル転送時間（転送データ長に比例）
Ｃ：伝送路送信待ち時間
Ｄ：リング伝送路遅延時間

また、ソフト走行時間は送受信の際にＰＣＩハードと直列に走行する通信制御プログラムの走行時間を示す。

上式に基づいたＰＣＩ性能算出結果を図７、

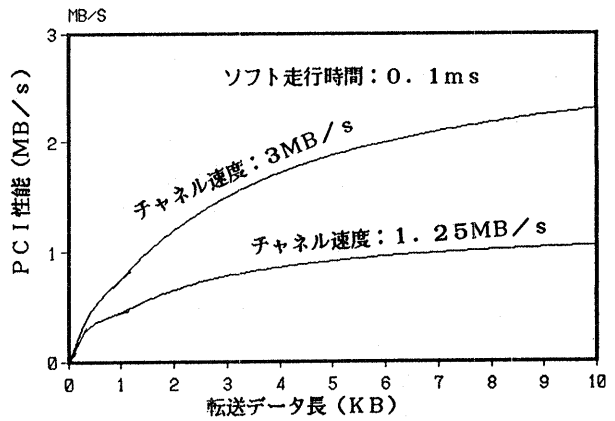


図7．　転送データ長に対するＰＣＩ性能

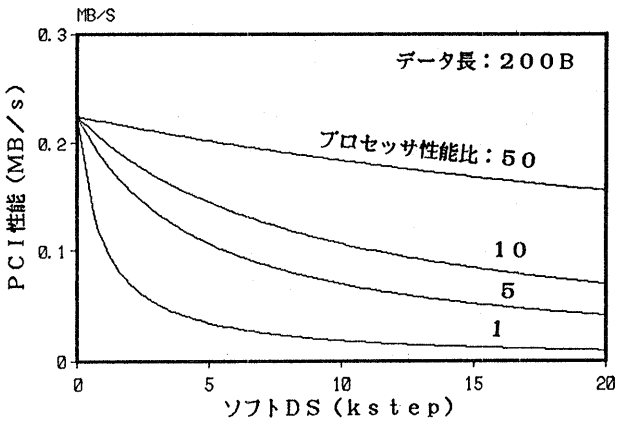


図8．　ソフトＤＳに対するＰＣＩ性能

４．光ファイバリングネットワークを用いたプロセッサ間結合システムの効率的使用方法

ＰＣＩを介した通信の効率化を図るため、以下の点に留意する必要がある。

（１）Ｗｒｉｔｅコマンド発行回数の削減

ＰＣＩでは、性能向上の点から、１Ｗｒｉｔｅコマンドあたりのデータ長は、なるべく長い方が好ましい。分割転送方式の採用により、データ長の増大によるＰＣＩハード動作時間の増大はチャネル転送時間の増大分だけである。短い電文を多数転送する場合には、短電文を一まとめにして１回のコマンドで転送するデータチェインを使用したほうがよい。

この多数の短データを一まとめで転送する効果を、転送データ長をパラメータとして図９に示す。

（２）Ｒｅａｄコマンドの先行発行

ＰＣＩでは、アテンション割り込みをなくし、複数サブチャネルを用いた効率的な多重動作を行うためにコマンド待ち合わせ機能を用いている。このため、Ｒｅａｄコマンドを Ｗｒｉｔｅコマンドに先行して発行できるプログラムの作りにしておくことが有効である。すなわち、各サブチャネルに対し最初に先行Ｒｅａｄをまとめて発行し、各Ｒｅａｄ動作が終了次第すぐに先行Ｒｅａｄ発行を行うようにすることで効率化が図れる。

（３）多数サブチャネルの使用

２．２節で述べたように、多数サブチャネルの使用によって通信処理に要するダイナミックステップ数の削減が可能である。

またＰＣＩハードウェアでも、チャネル早期解放により、動作中のチャネル空き時間（コマンド発行時のＲＥＴ（再試行要求）からＲＲ受信時まで、及びチャネルデータ転送後のＣＥからＡＣＫ受信まで：図５）を他サブチャネルの動作に用いることにより、多重動作が可能である。

従って、ＰＣＩを用いてプロセッサ間通信を行う場合、多数のサブチャネルを使用することで、通信効率向上が図れる。

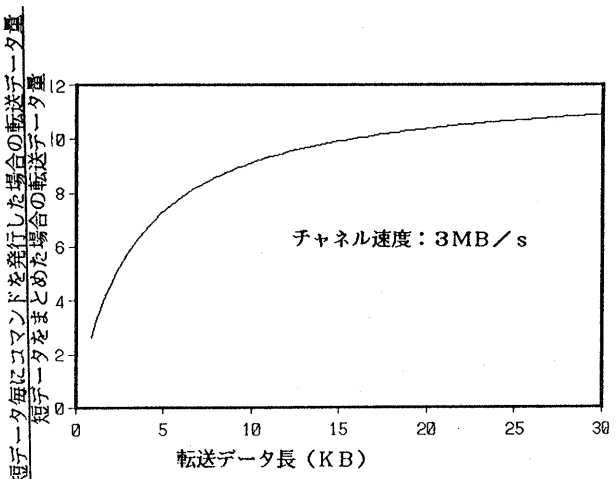


図９．　短データをまとめて転送する効果

５．おわりに

１００Ｍｂ／ｓ光ファイバリングネットワークを用いたプロセッサ間通信における特徴的な動作（サブチャネル多重、長データ分割転送等）および性能について述べた。また、これら機能を生かすための使用方法を述べた。

本ネットワークは、既に十数システムのＤＩＰＳセンタに導入され、商用に供されている。

参考文献

（１）Stuck,B.W.:Calculating the Maximum Mean Data Rate in Local Area Networks, Computer,Vol.16,No.5,pp72-76(1983)

（２）Bux,W.:Local-Area Subnetworks:A Performance Comparison,IEEE Trans.Commun.,Vol.COM-29,No.10,pp1465-1473(1981)

（３）小柳津他：ＤＩＰＳ複合構成システムプロセッサ間結合装置、通研実報，３５，Ｎｏ．３、ｐｐ２８３－２９３（１９８６）

（４）星子他：１００Ｍｂ／ｓ光トークンリングを用いたプロセッサ間結合システム，情処論文誌、第２７巻、第４号、４３５頁－４４４頁（１９８６）
（５）中野、森：疎結合計算機システムにおける高速計算機間通信方式、情処第３２回計算機アーキテクチャ研究会（１９８１）
（６）星子他：共通バスを介したプロセッサ間通信効率化の検討、情処第２４回全国大会、６Ｈ－６（１９８２）