計算機アーキテクチャ 117－6
（1996. 3. 5）

並列計算機システムFOLONの通信ライブラリの設計と評価

上田 学[†] 菊地 賢太郎[†] 土山 了士[††]
小林 一成[††] 大原 拓三[††]
外山 孝伸[††] 上田 和紀[††]

FOLONは安価なPCをPEとした共有メモリを持つ並列計算機システムである．共有バスにはVMEバスを利用し，ハードウェア工作をまったく行うことなく並列計算機を実現している．既存のハードウェアが持たないタイムアウト処理や排他制御についてはソフトウェアで実現している．FOLONでは、共有メモリ機能，メッセージパッシング型プロセス間通信機能をライブラリの形で提供している．本稿ではこれらのライブラリの設計と通信性能について述べる．

Design and evaluation of the communication libraries of the parallel computer system FOLON

GAKU UEDA,[†] KENTARO KIKUCHI,[†] RYOJI TSUCHIYAMA,[††]
KAZUNARI KOBAYASHI,[††] TAKUZO OHARA,[††]
TAKANOBU TOYAMA[††] and KAZUNORI UEDA[††]

FOLON is a shared memory parallel computer system which uses high cost-performance PCs as processing elements. FOLON employed VMEbus and no dedicated hardware was developed to make the whole FOLON system. Functions not supportd by the hardware, such as time-out and mutual exclusion management, and the interprocess communication capabilities are provided as libraries. This paper describes the design and communication performance of these libraries.

1. FOLONとは

FOLONは共有メモリを持つ並列計算機システムであり，記号処理をはじめとする均質でない細粒度並列処理に向いたシステムを目指している．

FOLONの設計方針は以下の通りである．

- 価格性能比の高いパソコンをPE(processing element)として使うことにより，システム全体を安価に仕上げる．
- 既存のハードウェアを使い，新規のハードウェア製作を行なわない．
- 並列計算機として動作するために必要であり，ハードウェアが持たない機能はOSやライブラリレベルで実現する．
- 細粒度並列処理に向いたプラットホームを実現するためEthernetで構成されたワークステーションクラスタよりもはるかに高い通信性能を発揮させる必要がある．特に少量のデータ転送の高速化を図る．

インターコネクションネットワークの製作を行わず、既存のハードウェアを利用した並列計算機システムの研究には文献2があるが，本研究では上記の方針にしたがって，実際には次のように実装されている．

- PEとしてIBM-PC互換機(以降PCと呼ぶ)，Pentium 120マシンを使う．
- 外部にVMEバスを用意し，これを共有バスとして使う．PEとVMEバスとの接続はPCI-VMEバス変換アダプタを利用する．
- PCI-VMEバス変換アダプタはBiT3社製PCI-VMEバスアダプタ Model 617を採用している．
- OSは，FreeBSD 2.1.0-RELEASEを用い，カーネルレベルに組み込むデバイスドライバおよびユーザーレベルのライブラリによって，VMEバス上のメモリを共有メモリとしたシステムを構築する．
- 共有メモリアクセスの高速化を図るため，OSのバイパス，DMA転送の利用を行う．

本稿ではまずFOLONの構成について述べ，各レベ

† 早稲田大学大学院理工学研究科
Department of Information and Computer Science, Graduate School of Science and Engineering, WASEDA University
†† 早稲田大学理工学部情報学科
Department of Information and Computer Science, School of Science and Engineering, WASEDA University

ルにおける通信性能の測定とその結果について述べる.

2. ハードウェア構成

FOLONはIBM PC/AT互換アーキテクチャのパーソナルコンピュータ(以降PCと記述する)をPEとしている. 1996年1月現在, FOLONは6台のPC(Pentium 120MHz)から構成される.

外部にVMEバスを用意しこれを共有バスとして利用している. VMEバスシャーシには各PEとケーブルで接続されたバス変換アダプタカード(VME側)が接続されている. このバス変換アダプタカード上にはデュアルポートメモリがぞれぞれ128KBずつ, 計896KB☆が実装されている(図1).

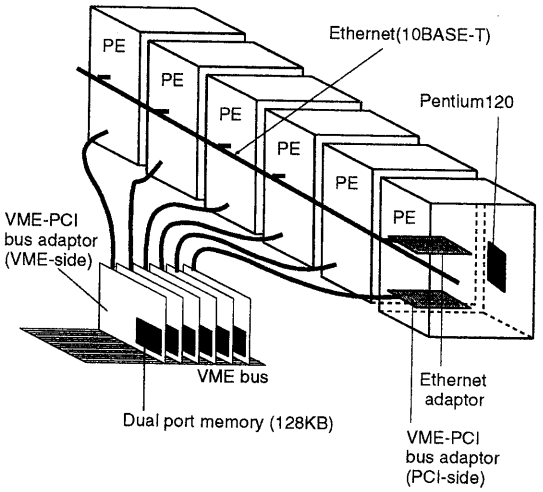

図1 FOLONのハードウェアシステム構成図

VMEバスに接続されたバス変換アダプタカードは, PC内部のPCIバスに接続されたPCI側のカードに接続されている. このほかに, 各PEはEthernetにより接続されいる☆☆.

各PE上ではFreeBSD☆☆☆が動作しており, 共有メモリを使わず, Ethernet経由でアクセスすれば通常のUNIXマシンとして利用できる.

VMEバス上には, 各PEのメモリアドレス空間とは別にVMEメモリアドレス空間が存在する. VMEバスは32bitアドレスのメモリ空間を持ち, 4GBまでのアドレッシングが可能である. FOLONではVMEバスを共有バスとして利用する. VMEバス上のデュアルポートメモリと, 各PEのメインメモリの一部をVME空間にマップした部分を, 共有メモリとして使うことができる(図2).

VME空間でアドレッシング可能なメモリ領域は,

☆ PE6台と監視用の1枚
☆☆ FOLONでは最大20台のPEで稼働することができる.
☆☆☆ FreeBSD 2.1.0-RELEASE

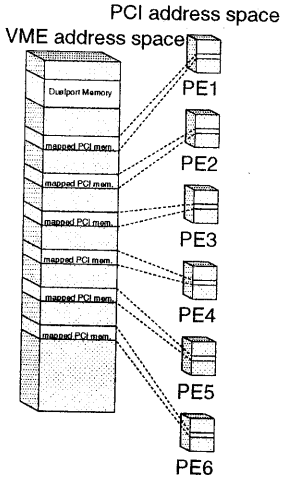

図2 VMEアドレス空間

PCIメモリ空間にもマップさせ, Test & Set命令が使えないことを除けば通常のメモリアクセスと同じ方法でアクセスすることができる.

また, 大量のデータに対してはDMA転送を用い, より高速な処理を行なう. これにより共有メモリを通信路として使い, Ethernetよりもレイテンシが小さくスループットが大きい通信を実現している.

3. ソフトウェア構成

3.1 ライブラリ階層

FOLONのPEはBSD系UNIXであるFreeBSDで動作する. FOLONの並列計算機システムとしての機能はFreeBSD上に構築されたデバイスドライバ, ユーザーアプリケーションにリンクするライブラリで実現されている. FOLONのハードウェアおよびOSは本来並列計算機用途に開発されたものではなく, 不足している機能は既存の機能の組み合わせあるいはソフトウェアアルゴリズムによって補う必要がある.

並列計算機システムとしての機能を, FOLONでは大きく3つの階層に分けインプリメントしている.

デバイスドライバ PE間を接続しているVMEバスおよびバスを操作するハードウェアの制御を可能にするドライバ

VCL(VME Control Library) VME資源のコントロールおよび共有メモリアクセス機能を実現するライブラリ

FCP(FOLON Control Protocol) ライブラリ マルチプロセス対応プロセス間通信を実現するプロトコルライブラリ

階層構造を図に表わすと図3のようになる. VCLは, VME上の共有メモリに対するアクセスをOS/カーネルをバイパスさせ直接操作することで高速化(主にレイ

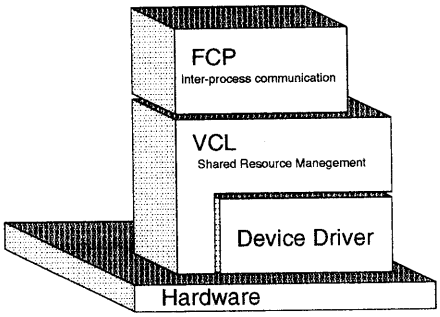

図3 FOLON のライブラリ階層

テンシ性能の向上) を図っている.

3.2 デバイスドライバ

デバイスドライバはVMEバスアダプタのコントロールの他, メモリマッピング, FOLON プロセスの管理, 割り込み処理, DMA 転送を担当している.

3.2.1 メモリマッピング

FOLON では, 各 PE のメモリ空間上の空き領域にデュアルポートメモリのメモリイメージをマップすることで, 共有メモリを実現している. 2 つの PE がこの領域をアクセスすることにより, VME バス経由での通信が実現されている. さらに PE の PCI メモリの一部を VME 空間にマップすることで, VME メモリ空間を増やしている. このメモリをマップド PCI メモリと呼び, ローカルな PE からは, 通常のメモリと同じ速度でアクセスできるという特性を持つ.

3.2.2 VME バスアダプタのコントロール

VME バスアダプタの状態は I/O ポートを経由して取得することができる. また VME バスアダプタの制御も I/O ポートを操作することで行う. VME バスの初期化など VME バスを操作する処理は, 実際はドライバが I/O ポートを操作することで実現している.

3.2.3 DMA 転送の実現

VME バスアダプタは DMA 転送機能を持っている. この機能をドライバを経由して使うことで, 大量のデータ転送の効率をあげることができる. DMA 転送は以下のような理由でドライバ内部で実現されている.

- I/O ポートのうち多くのレジスタを使って初期設定を行う必要がある.
- DMA 転送中は PE の転送データの含まれる物理メモリのスワップアウトを禁止する必要がある.
- 1PE 内で同時に行える DMA 転送は 1 つであり, 共有メモリにアクセスするプロセス間での調停が必要となる.

3.2.4 FOLON プロセスの管理

VME バスアダプタ自体は複数プロセスの同時使用を考えて設計されているわけではない. そのため I/O ポートやメモリマップ情報など, FOLON としての動作に関わる共有資源はデバイスドライバで一括して管理している. さらに FOLON を使用するプロセスに関する情報を管理しており, ソフトウェア割り込み処理のために使用している.

3.2.5 割り込み処理

VME バスで発生する割り込みは PE のバスアダプタによって, 一般割り込みとエラー割り込みの二種類の割り込みとして, PE のデバイスドライバに伝えられる. 現在は一般割り込みのうち DMA Done 割り込みとエラー割り込みを監視している.

3.2.5.1 エラー割り込み

VME バスへのアクセス時にエラーが発生した場合, VME バスアダプタは PE にエラー割り込みをかける. この割り込みはドライバ内で処理され, 必要ならばプロセスに通知される.

現在のドライバでは

- インターフェースパリティエラー
- リモートバスエラー
- インターフェースタイムアウト

の 3 種類のエラーを検出している. 発生したエラーの識別は割り込みハンドラがステータスレジスタを参照することで行っている.

3.2.5.2 DMA Done 割り込み

この割り込みは DMA 転送の成功, 失敗にかかわらず終了時に発生する. DMA 転送中は PE 上の他のプロセスは VME 資源のアクセスを禁止されているのでアクセスできずに待っている. ドライバはこの割り込みを確認すると DMA 転送の後処理を行った後, 他のプロセスに使用許可をソフトウェア割り込みを使って通知する.

3.2.5.3 タイムアウト対策

後述する VCL ライブラリで説明する通り, VME アクセス時に 30μs 以上の遅延があった場合, VME バスはタイムアウトエラーによる割り込みを発生させる. 現在の仕様ではタイムアウトエラーの場合は再送を行うことになっており, ドライバがタイムアウトエラーを確認した場合には, FOLON を使用中のユーザープロセスに対してドライバはソフトウェア割り込みを使って通知する. 実際に再送する機能は後述する VCL ライブラリ内で実現されている.

3.3 VCL (VME 制御ライブラリ)

VCL(VME Control Library) は, FOLON の持つ共有メモリ機能を実装するライブラリである.VME バス上の共有メモリに対するアクセスはすべてこのライブラリを通して行われる.

VME 上の共有メモリに対するアクセス速度の向上のため, 共有メモリアクセスは OS のバイパスによってユーザープロセスから直接行う. デバイス利用の競合の解決を VCL レベルで行う.

このライブラリでは次の機能を実装している.

- 共有メモリに対するさまざまな種類のアクセス
- デバイスドライバで実現されている機能のインターフェース (DMA 転送など)
- 共有メモリに対する Test & Set 機能

● ユーザーからのシステムの保護

共有メモリに対するアクセスの種類は限られており，それらすべてのアクセスをVCLでインプリメントする．上位プログラムは必然的にこれらのいずれかの機能を利用するため，中間にライブラリが存在することによるオーバーヘッドは問題にならない．

3.3.1 VCLが提供する共有メモリ空間

FOLONで利用できる共有メモリには，VMEバス上のデュアルポートメモリとVMEメモリ空間上にマップされた他のPEのPCIメモリがある．

自PEのPCIメモリ空間の一部にはデュアルポートメモリ，他のPEのマップドPCIメモリがまとめられている．自PEのマップドPCIメモリ(他からアクセスできる領域)は，マップされたメモリとは別の領域にカーネルが確保しページングの対象外となっている．

VCLは，個々のPEごとに異なるPCIアドレス空間を，すべてのPEから同じ共有メモリアドレス体系でアクセスできるようFOLONアドレス空間を提供し，ライブラリ内でアドレス変換を行っている(図4)．

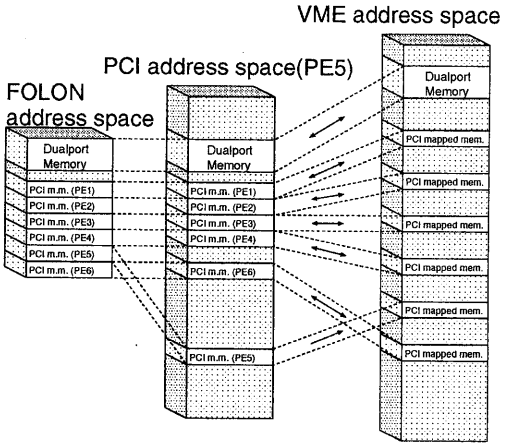

図4 各メモリ空間の間の関係(PE5の例)

3.3.2 共有メモリに対するTest & Setの提供

VCLはPE内の複数のプロセスから同時に利用することが可能なマルチプロセス対応型ライブラリである．FOLONではレイテンシ改善のため，共有メモリへのアクセスをOS(デバイスドライバ)を経由させず，直接ユーザープロセスからアクセスできるようにしている．このためバス変換アダプタが複数のプロセスから同時に操作されることがあるため，Test & Setの実現には次のような方法をとっている。

Test & Setを実現するには，同一プロセスが処理を掌握しなければならない．同一プロセスによる処理を保証するために，まずPE内のプロセス排他を行い，成功すればVMEバスにバスロックをかけPE間の排他を行うことで，Test & Setが実現できる(図5)．

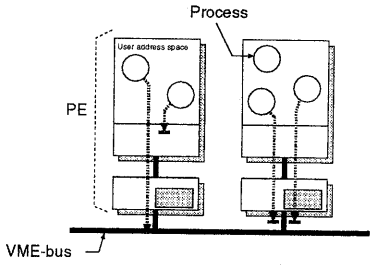

図5 排他制御の流れ

3.3.3 タイムアウト処理

VMEバスアダプタ経由のメモリ読み込み/書き込みが一定時間実行できなければVMEバスアダプタのインターフェースタイムアウトエラーが発生し，タイムアウトを発生させたPEに割り込みがかかる．

割り込みがカーネルに通知され，割り込み処理が開始されるまでに，VCLのメモリ操作処理が先に進んでしまう☆という問題が発生する．

この問題はPE(PCIバス)においてPosted Write[5]という手法が取られていることが原因であり，FOLONで利用しているマザーボードでは5ダブルワード分のバッファが設定されている．

割り込みはVMEデバイスドライバに捕捉され，PE内のVCLを利用するすべてのプロセスに通知される．VCLでは，メモリ操作関数に対し5命令分の時間を待ってから処理を終了することにより，割り込みを確実にメモリ操作関数内で捕捉することを可能にし，割り込みがかかってきた時点でメモリ操作を行っていれば，現在の書き込みを中止し最初から再度書き込みを行うことによりタイムアウトの処理を実現している(図6)．

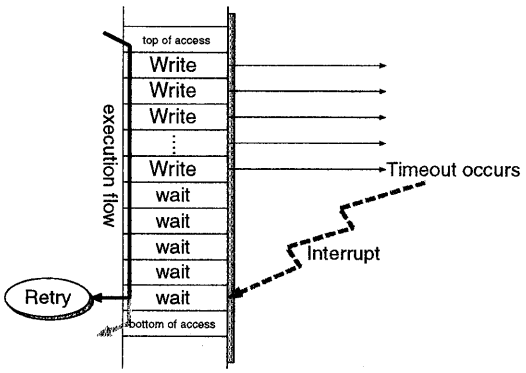

図6 タイムアウト処理

3.4 FCP通信ライブラリ

FCP(FOLON Control Protocol)ライブラリは、VCL上に構築されたメッセージパッシングモデルの通

☆ 最大4ダブルワード先まで書き込まれる場合がある

信ライブラリである．FCPはマルチプロセスに対応したことが最大の特徴である．

FCPの設計方針は以下の通りである．

- UDPと同等の通信をより高速に行う方式を実装．
- PE内のマルチプロセスに対応．
- バスの混雑をできるだけ軽減．
- 今後のFOLONの拡張に対する移植性の確保．

3.4.1 高速な通信の実現

FCPではVCLを利用しOSを介さず共有メモリを経由した通信が可能になるため，Ethernet経由の通信に比べレイテンシの小さい通信が可能となる．FCPは，メッセージの大きさによって最適なメモリアクセス方法を選択し，大きなメッセージに対してはDMAを利用することにより最適化を図っている．

3.4.2 PE内マルチプロセスへの対応

FCPでは同一PE内のプロセスが同時にPE間通信に参加することが可能となっている．FOLONの通信に参加するプロセスは初期化時にVME空間の共有メモリ上にそのプロセスの受信バッファが割り当てられる．送信プロセスがそのバッファに書き込むことにより通信が成立する．共有メモリアクセスをクリティカルセクションにせずに，FCPのプロトコルレベルで同時アクセスに対しても動作することを理論的に保証し，排他制御を最低限に押さえている．

3.4.3 バスの混雑の緩和

FOLONに利用している共有メモリは，デュアルポートメモリ，マップドPCIメモリともPEがローカルに持つメモリに対しては共有バスを経由せずにアクセスができる．この特性を活かすためFCPでは受信側のポーリングがバス上に出ていかないよう受信バッファをローカルにある共有メモリに用意し，書き込み側のメモリアクセスのみがバス上に現れるようになっている．

3.4.4 移植性の向上

FCPは，PVMやMPIなどの並列ライブラリや低レベルでの操作の必要な並列アプリケーションから利用することを主眼におき，UDPと同等の通信を実装し，これらを利用するアプリケーションの移植が容易なよう設計されている．FCPを利用する場合は共有メモリ空間に対する考慮は不要となる．

FCPは，FCPを利用するアプリケーションからFOLONのハードウェアに依存する部分を取り除き，ハードウェア構成や共有メモリマップの変更，さらにより高速な転送方法が実装された場合にもアプリケーションのソースコードに変更を加えることなくそのメリットを享受できるようになっている．

4. 性能評価

FOLONの通信性能を評価するため，ハードウェアレベル，VCLレベルでベンチマークテストを行った．

4.1 FOLONの基本性能

まずハードウェアの持つ基本性能を測定した．各種ライブラリではこの性能に対し，ソフトウェアオーバーヘッドが加算されるため，最大限この値に近づけるようインプリメントすることを目標としている．

レイテンシは，4バイトのデータをPEのユーザーメモリ空間から別PEの共有メモリへ転送し，転送の命令を終了した時間までを測定した．スループットは4096バイト転送の場合の値である(表1および表2)．

送信先	アクセス	レイテンシ	スループット
デュアルポートRAM (ローカル)	read	1	1.89
	write	2	1.91
デュアルポートRAM (リモート)	read	1	1.75
	write	1	1.74
マップドPCIメモリ (リモートのみ)	read	6	0.87
	write	1	0.93

表1 PIO性能(レイテンシ単位 μs, スループット単位MB/s)

送信先	アクセス	レイテンシ	スループット
デュアルポートRAM (ローカル)	read	136	8.70
	write	134	8.77
デュアルポートRAM (リモート)	read	137	7.55
	write	134	7.50
マップドPCIメモリ (リモートのみ)	read	142	8.79
	write	145	7.46

表2 DMA性能(レイテンシ単位 μs, スループット単位MB/s)

4.2 VCLレベルのベンチマーク

共有メモリアクセスを提供するVCLレベルでのベンチマークテストは，データが送信側PEのプロセスから受信側PEのプロセスに到達するまでの時間を，往復にかかる時間から算出する．通信に利用するメモリがマップドPCIメモリの場合とデュアルポートメモリの場合について測定した．

4.2.1 PIO転送

まず，共有メモリに対するアクセスをPIOで行う性能について測定した．表3の値は，PEのユーザーメモリ空間から別のPEのユーザーメモリ空間までの片道の転送時間を往復の転送時間から算出した値である．データの経由する経路は，クライアントプロセス→サーバのバッファ(PCIまたはデュアルポートメモリ)→サーバプロセス→クライアントのバッファ(PCIまたはデュアルポートメモリ)→クライアントプロセスである(図7, 図8)．レイテンシは1バイト転送，スループットは4096バイト転送の場合の値である．今回の測定ではタイムアウト処理は行っていない．

PIO転送では，Ethernet経由の通信に比べレイテンシ性能を50倍程度向上させることが可能になるがスループットに関してはEthernet経由UDPと同等の転送速度しか得られない．

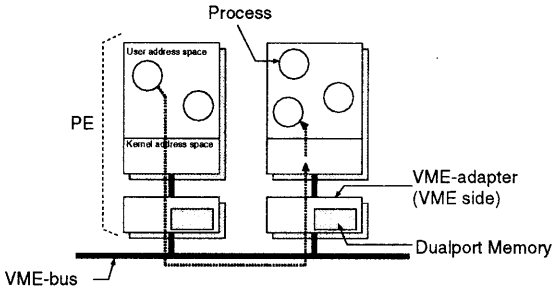

図7 マップドPCIメモリ経由の通信の経路

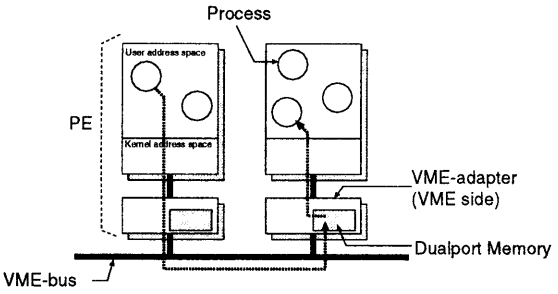

図8 デュアルポートメモリ経由の通信の経路

経路	レイテンシ	スループット
マップドPCIメモリ経由	5	0.43
デュアルポートメモリ経由	6	0.38
Ethernet経由UDP/IP	320	0.39

表3 VCLレベルにおける通信性能(レイテンシ単位 μs, スループット単位MB/s)

4.2.2 DMA転送

PIO転送の欠点である大量のデータ転送の遅さを解決するため, FOLONではバスアダプタの持つDMA転送を利用したメモリアクセス関数をサポートしている. ベンチマークは, PEのユーザーメモリ空間→別PEのマップドPCIメモリ→別PEのユーザーメモリ空間→PEのマップドPCIメモリ→PEのユーザーメモリ空間の転送時間を測定し, その片道の通信のレイテンシを最小の通信である4バイトの通信, スループットを4096バイトの通信から算出した.

経路	レイテンシ	スループット
マップドPCIメモリ経由	480	1.00
Ethernet経由UDP/IP	320	0.39

表4 VCLレベルにおけるDMA転送性能(レイテンシ単位 μs, スループット単位 MB/s)

DMA転送の場合はDMA転送用のレジスタ設定などのオーバーヘッドのため, レイテンシ性能はEthernet経由UDPに比べて悪いが, 160ダブルワード(640バイト)以上の転送でPIOアクセスよりも早くデータ転送ができるようになり, スループットも2.5倍程度あり, メッセージ長の増加に対する通信時間の増加幅をPIO転送に比べ大幅に押さえることができる.

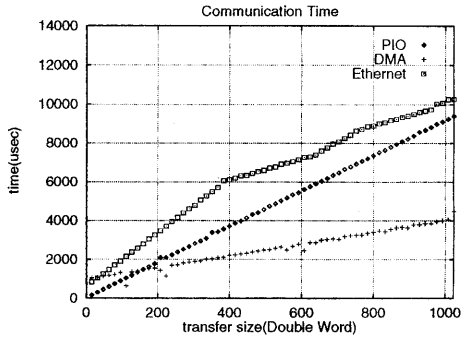

図9 VCLレベルにおける通信性能

5. まとめ

FOLONは, PEとしてPCを使いデバイスドライバ, VCL, FCPでPE間通信を実現している. FOLONのVCLレベルでは, レイテンシ性能でUDPの50倍程度を実現している。また, 大量のデータに関してはDMA転送を用い, スループット性能でUDPの2.5倍程度の性能を発揮する.

プロセス間通信を実現するFCP上へのPVMの移植は現在進行中であり, KL1言語処理系でありPVMを利用する並列版KLICを動作させる予定である.

また今回通信性能の測定ができなかったFCPレベルでの転送性能の測定, VCLにタイムアウト処理を追加した場合の転送性能の測定を行うとともに割り込みなどを活用し, 既存のVCL, FCPのパフォーマンスチューニングを行う予定である.

参考文献

1) BIT3 Computer Corporation: "Bus-to-Bus Adaptor Model 617 Adaptor Hardware Manual", BIT3 Computer Corporation, Minneapolis, Minnesota, 1994.
2) 瀧和男, 小倉毅, 小西健三: "ワークステーション複合体による並列処理システム", 情報処理学会研究報告, 93-PRG-13, pp. 49–56, 1993.
3) Janet I. Egan Thomas J. Teixeira, 野中浩一, 大西照代(訳): "改訂版UNIXデバイスドライバ", アスキー出版局, 1989.
4) Samuel J. Leffler: "The Design and Implementation of the 4.3BSD UNIX Operating System", Addison-Wesley, 1988.
5) 滝誠一: "PCIバスの詳細と応用へのステップ", OpenDesign, No. 7, CQ出版社, 1995.