ハイパフォーマンス
コンピューティング 57－1
（１９９５．８．２４）

STAFF-Linkを用いたワークステーションクラスタ上へのPVMの実装とその評価

高橋 淳, 中條 拓伯, 小畑 正貴, 金田 悠紀夫

神戸大学 工学部 情報知能工学科

e-mail:taka@jedi.seg.kobe-u.ac.jp

ワークステーションクラスタ技術により，並列処理が身近なものになってきた．しかしながらワークステーションクラスタを構成するノード間の通信速度がシステム全体の性能を大きく左右する．我々はSTAFF-Linkと呼ばれる高速なシリアルリンクでワークステーション間を接続した並列分散処理環境を構築し，その上に並列処理ライブラリPVMの実装を試みる．STAFF-Linkを用いることによってさまざまなトポロジーでバンド幅の広いネットワークを構築することができ，柔軟な処理環境を提供できる．本稿ではSTAFF-Linkによって通信を行なうPVMの実装方針について述べ，その通信性能を評価した結果について述べる．

Implementation and Evaluation of PVM on Workstation Cluster using STAFF-Link

Atsushi Takahashi, Hironori Nakajo, Masaki Kohata, Yukio Kaneda

Department of Computer and Systems Engineering, Faculty of Engineering, Kobe University

e-mail:taka@jedi.seg.kobe-u.ac.jp

A workstation cluster technology has made parallel processing familiar to end users. However, the performance of whole system is greatly influenced by communication speed between nodes which comprise workstation cluster. Therefore we try to construct parallel and distributed environment which consists of workstations connected via high-speed serial links called STAFF-Link, and implement parallel processing library PVM on it. By using STAFF-Link we can organize the broad bandwidth network of various topologies and can provide flexible parallel processing environment. In this paper we describe the policy of implementing PVM communication via STAFF-Link and evaluate its processing performance.

1 はじめに

近年の高速RISCプロセッサを搭載した高性能ワークステーションの低価格化はめざましく，それらのワークステーションを並列処理に利用するための，ワークステーションクラスタと呼ばれる技術が注目されている．ワークステーションクラスタに対応した並列処理ライブラリとしてPVM，p4，Express，MPIなどが開発され，広く利用されている．このワークステーションクラスタはネットワークを用いて利用率の低いワークステーションの演算能力を集約して利用するため，コストパフォーマンスが高いという利点がある．

しかしワークステーションクラスタや並列計算機などの並列処理システムが解決しなければならない問題点として，通信スループットの向上が挙げられる．各ノード間の通信性能がシステム全体のスループットに大きく関わってくるので，通信時間を短縮することの重要性は高い．

近年になり，光ファイバケーブルなどの高速ネットワークメディアも利用されるようになってきたが，LAN(Local Area Network)の事実上の標準であるEthernetでは大量のデータ転送に対しては通信容量不足である．また，Ethernetの媒体アクセス制御方式であるCSMA/CD方式(Carrier Sense Multiple Access with Collision Detection)では，データの送信時に衝突が検出されると，あるランダムな時間間隔をおいて再送信を行なうようになっており，Ethernetの利用率があまり高くない場合にはこの方法がうまく機能するが，通信が頻繁に発生すると急激に応答時間やスループットが悪くなることがある．そして，LANの基本プロトコルであるTCP/IPプロトコルは，広域通信に対応するための信頼性を重視しているため，プロトコルのオーバヘッドが大きい．

そこでこれらの問題を解決するために，研究室内などのある程度狭い範囲内での環境を想定し，Serial Transparent Asynchronous First-in First-out Link(STAFF-Link)と呼ばれる高速なシリアルリンクを用いてワークステーション間を接続し，通信プロトコルのオーバヘッドを軽減した形態の並列処理環境を構築することを提案する．そしてSTAFF-Linkで接続された分散環境において並列処理を行なうためのライブラリとしてPVMを実装する．以下，このSTAFF-Linkを介して通信を行なうPVMのことをSTAFF-PVMと呼ぶ．

STAFF-Linkを用いてワークステーション間を接続することによってさまざまなトポロジーのネットワークを構築することができ，ネットワークのトポロジーに依存するような問題に対しても柔軟に対応できるようになる．

本稿では，STAFF-PVMの実装の方針について述べ，STAFF-PVMの基本的な通信性能の予測を行なう．まず2章でワークステーション間を接続するSTAFF-Linkについて説明し，3章では並列処理ライブラリPVMの説明とSTAFF-PVM実装の方針について述べる．4章でSTAFF-PVMの通信性能を予測し，5章でまとめと今後の課題について述べる．

2 STAFF-Link

STAFF-Linkとは，高速シリアル通信用LSIと送信/受信用FIFO，及びそれらを制御するための通信コントローラとからなる通信ブロック間を，ツイストペアケーブルで接続することによって構成されるシリアルリンクである．STAFF-Linkは最高17.5Mバイト/秒でデータ転送を行なうことができる．以下にSTAFF-Linkの特徴を挙げる．

通信スループットの向上

一般的にシリアル通信は，1.データの書き込み，2.パラレル→シリアル変換，3.データ転送，4.シリアル→パラレル変換，5.データの読み出し，の5つのフェーズに分けることができる．従来のシリアル通信ではパラレル→シリアル，シリアル→パラレルの変換の際に通信遅延が生じ，通信スループットが上がらなかった．STAFF-Linkでは，3～5のフェーズを高速シリアル通信用LSIで処理し，FIFOメモリを使用したバッファによって5つのフェーズをオーバラップさせ，通信スループットを向上させている．

簡易なインタフェース

内部の通信コントローラがFIFOが溢れないようにフロー制御を自動的に行なってくれるので，STAFF-Linkに対するデータの読み書きは，通常のFIFOメモリへの読み書きと同様に行なうことができる．

柔軟なネットワーク環境の構築

シリアルリンクであるのでポート数を比較的容易に増加できる．それにより1台のワークステーションから多数のワークステーションへ接続することができ，メッシュやハイパーキューブなどいろいろなトポロジーのネットワークを構築することができる．

3 STAFF-PVM

3.1 PVMの概要

PVM(Parallel Virtual Machine) はネットワーク上のUNIXワークステーション群を互いに通信させることにより，それらを仮想的に単一の分散メモリ型並列計算機として利用するシステムのことである．PVMの下ではワークステーションの集まりをユーザが定義し，一つの大きな分散メモリ型計算機として表す．この論理的な計算機のことをバーチャルマシンといい，バーチャルマシンを構成する各ワークステーションをホストという．また，PVMにおいて並列に実行されるプログラムの単位をタスクという．

PVMソフトウェアシステムの構成は大きく次の2つに分けられる．

PVMデーモン (pvmd)

バーチャルマシンを構成する全てのホスト上に常駐し，互いに通信し，協調しながらPVMアプリケーションを実行する．タスク間の通信に関してメッセージのルータ，コントローラの役割を果たすプロセスである．

PVMライブラリ

メッセージ通信，プロセスの生成，タスクの協調及びバーチャルマシンの再構成などのためのライブラリルーチンを提供する．ユーザはこのライブラリによってマシン間の複雑な通信プロトコルを気にすることなくプログラムを作成することができる．

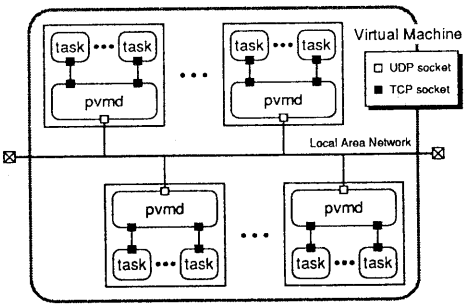

図 1: PVMの概念図

3.2 PVMの通信

PVMではタスク間の通信はPVMデーモンを介して行なわれる．具体的なプロトコルは，PVMデーモン間ではUDPソケットを通じて通信が行なわれ，PVMデーモンとタスクの間ではTCPソケットを通じて通信が行なわれる．ただし特別に直接通信の指示があったときのみ，タスク間でTCPソケットを通じて通信が行なわれるようになっている．UDPは信頼性が低いので信頼性のあるパケット配送システムがUDPの上位に実装されている．

タスク間の基本的な通信パターンを簡単な例を用いて説明する．異なるホスト上にあるタスクAからタスクBにデータを送信する場合を考える (図2)．まずタスクAからPVMデーモンAにTCPソケットを通じてデータが送信される．次にPVMデーモンAは受け取ったデータをUDPソケットを通じてPVMデーモンBに送信する．PVMデーモンBは受け取ったデータをタスクBにTCPソケットを通じて送信し，タスクAからタスクBへのデータ送信は完了する．

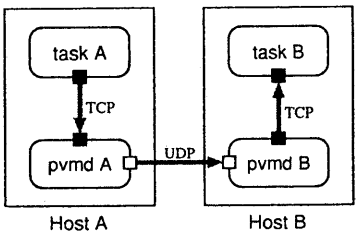

図 2: タスク間の基本的な通信

3.3 STAFF-PVMの実装

PVMデーモンとタスクの間の通信スループットを上げるために，TCPソケットによる通信を共有

メモリを用いて実装することも行なわれている．本研究ではネットワーク (Ethernet) を介して行なわれるPVMデーモン間のUDPソケットによる通信をSTAFF-Linkを用いて実装する．

PVMはバーチャルマシン全体で一意なタスクの識別子 (TID) を提供している．タスク間で通信を行なう場合にはTIDによって送信先を指定する．PVMデーモンはタスクから受け取ったデータの宛先が自分のホスト内のタスクであるならそのタスクにデータを送信し，そうでなければ該当するタスクのあるホストのPVMデーモンにネットワークを介してデータを送信する．STAFF-PVMではこのデーモン間の通信を次のようにして行なう．PVMデーモン間でやりとりするパケットに送信元のIPアドレス，データ長などをSTAFF-Link用のヘッダとして付加して新しいパケットを生成し，それをSTAFF-Linkを介して送信する．具体的なSTAFF-Linkへのアクセスは次のようにして行なう．

- リードアクセス

  PVMデーモンからはSTAFF-Linkの状態を示すフラグを見ることができる．フラグには通信エラーを示すフラグと受信用FIFOが空であることを示すEmpty flag，送信用FIFOの半分が満たされていることを示すHalf full flagがそれぞれ4つずつある．リードアクセスはEmpty flagを見てどのポートにデータが送られてきているか調べる．続いて，該当するポートからヘッダ部分を読み込み，データ長の情報からその分だけデータを読み込む．

- ライトアクセス

  Half full flagを見て，該当するポートにデータを書き込めるかどうか調べ，書き込めるならフラグを見ながらデータを書き込む．

現在は4台のワークステーションをSTAFF-Linkで完全結合させたシステムの実装を進めている．

4 性能評価

基本的なPVMの通信パターンの通信性能を測定し，PVMデーモン間の通信をSTAFF-Linkを用いて行なったとき，どの程度の性能向上が見込めるかを予測する．

4.1 実験環境

表1に示すワークステーションとネットワークを用いて測定を行なった．

表1: ワークステーションの仕様

SPARC station 5	
CPU	microSPARC-II
クロック	70MHz
メインメモリ	16Mbyte
OS	日本語 Solaris2.3
ネットワーク	Ethernet(10BaseT)

4.2 PVMの通信とUDP通信

まずPVMのタスク間の通信においてネットワークを介して行なわれる通信がどの程度の割合を占めるのかを調べるために，図2で示したような1対1のタスクの通信に要する時間と，単にUDPソケットによる通信に要する時間とを比較する．いろいろなサイズのデータを100往復させ，それに要した時間を200で割って1回当たりの通信時間を測定した．その結果を図3に示す．

図3からわかるようにPVMのタスク間の通信においておよそ半分の時間はネットワークを介した通信に費やしている．データサイズが小さい時に通信の割合が少ないのはメッセージを送信するために必要なPVMのセットアップ時間がメッセージ長に関係なく同じくらい必要になるからであると思われる．

UDPによる通信時間がSTAFF-Linkによって大幅に短縮できれば，PVMのタスク間の通信スループットの向上が見込める．

4.3 アクセス競合時のPVMの通信性能

次にEthernetの利用率が高くなった場合に通信時間がどのように変化するかを調べるために，前述のPVMのタスク間でデータを100往復させるプログラムを，同時に最大8台の別々のホスト上で2組から4組まで実行し，1回当たりの通信時間を測定した．その結果を図4に示す．

図4から，通信するデータのサイズが大きくなり，Ethernetへのアクセスの競合が増えるにつれて通信時間が長くなっているのが分かる．特に3000バイトを越えた当たりから急激に応答時間が悪くなって

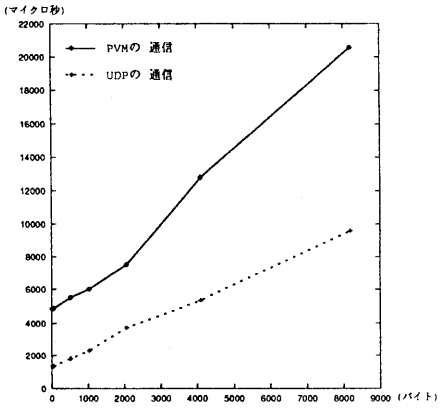

図 3: PVM と UDP の通信時間

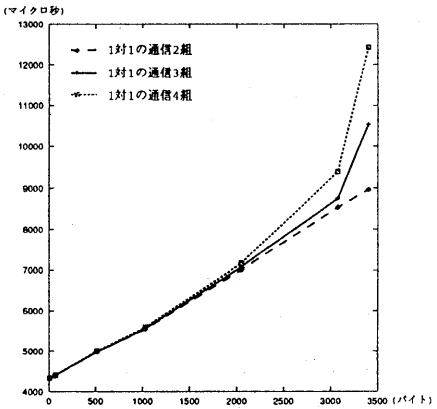

図 4: 衝突時の PVM の通信時間の比較

いる．これは Ethernet がバス型のネットワークであり，ある時刻にはワークステーション間で通信の競合が生じているためであると考えられる．

STAFF-Link でワークステーション間を接続することによって同時に複数の通信を行なうことができ，ネットワーク全体のバンド幅は STAFF-Link 単体のバンド幅よりも向上すると考えられる．

4.4 STAFF-Link の通信性能

最後に STAFF-Link の通信性能を測定した．また，UDP による通信時間と DMA を用いた STAFF-Link の通信時間とを併せて図 5 に示す．この結果より，粒度の大きい通信には DMA を用いた連続転送が有効であり，粒度の小さい通信には mmap システムコールを用いたバイト転送が有効であることが分かる．転送速度としては，バイト転送としては 500K バイト/秒程度しか出ていないが，DMA 転送を用いた場合は 2.6M バイト/秒程度の転送速度を示している．しかしながら少量の転送には DMA 転送をセットアップするためのオーバヘッドが大きい．これは STAFF-Link のデバイスドライバをチューニングすることによってある程度改善される．

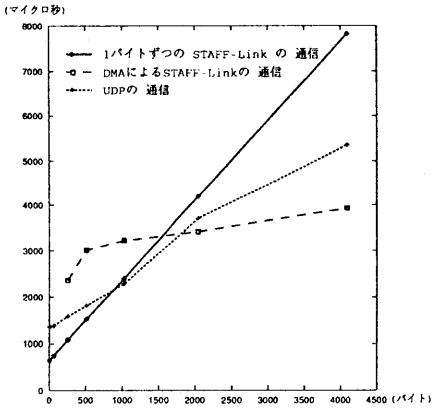

図 5: STAFF-Link の通信時間

次に，データサイズが小さい場合と大きい場合に分けて，PVM の UDP 通信に要する時間を STAFF-Link で通信する時間に置き換えて STAFF-PVM の通信性能の予測を行なう（図 6，図 7）．転送粒度が小さい場合は 1 バイトずつの転送の方が高速にデータ転送が行なえる．粒度の大きい転送を行なう場合には DMA 転送の効果が効いてくる．したがって転送するデータサイズによって通信方式を換えることによって効率良くデータ転送が行なえる．

5 現状と今後について

現在，PVM のソースコードを解析し，STAFF-Link で 4 台を完全結合させた STAFF-PVM の実装を行なっている．STAFF-Link の性能を十分に引き出すためには DMA 転送が不可欠である．現在，DMA 転送を行なうデバイスドライバにバグがあり，高速通信に対応できておらず，原因を究明中である．

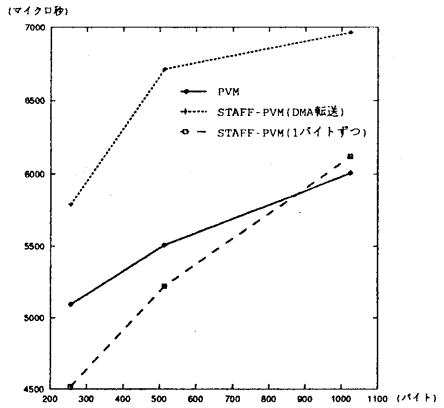

図 6: データサイズが小さい場合

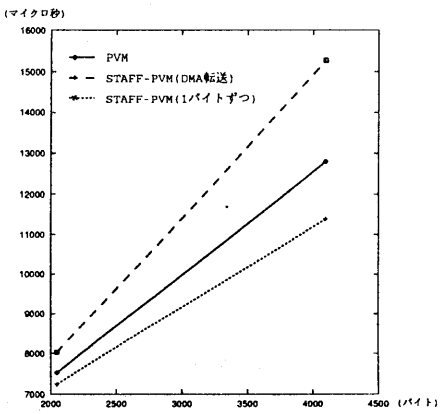

図 7: データサイズが大きい場合

これからの予定としては，まず 4 台完全結合の STAFF-PVM を実装し，種々のベンチマークプログラムを用いて STAFF-PVM の性能を評価していきたい．その後はワークステーションクラスタを構成するマシンを増やしていく．そのためにはパケットのルーティングについて考慮する必要があり，STAFF-Link にルータをハードウェアで実装するか，ソフトウェアで実装するかについて考えていかなければならない．ハードウェアで実装すると高速な経路制御が可能となり，ソフトウェアで実装すれば処理は少し遅くなるものの，柔軟な経路制御が可能となる．

現在，PVM をベースにした並列分散ファイルシステム [5] が提案され，評価されている．今後，これをもとに STAFF-PVM 上に並列ファイルシステムを実装し，超並列計算機のファイルシステムへの応用を考えている．具体的には文部省科学研究費補助金・重点領域研究において開発されている超並列計算機 JUMP-1[6] の初期ファイルシステムとして利用する予定である．

謝辞

本研究の一部は文部省科学研究費（重点領域研究（1）課題番号 04235130「超並列ハードウェア・アーキテクチャの研究」）によります．

参考文献

[1] Al Geist, Adam Beguelin, Jack Dongarra, Weicheng Jiang, Robert Manchek, Vaidy Sunderam, "PVM 3 USER'S GUIDE AND REFERENCE MANUAL", ORNL/TM-12187, Sep 1994.

[2] 中條 拓伯, 松田 秀雄, 金田 悠紀夫, "超並列計算機におけるワークステーションクラスタ・ファイルシステム", 情報処理学会計算機アーキテクチャ研究会報告 ARC107-24, Jul 1994.

[3] 岩下 茂信, 村上 和彰, "KU PVM3/AP1000 の性能評価", 情報処理学会ハイパフォーマンスコンピューティング研究会報告 HPC52-16, Jul 1994.

[4] 弘中 哲夫, "大規模ワークステーション・クラスタにおける PVM の性能評価", 情報処理学会ハイパフォーマンスコンピューティング研究会報告 HPC55-12, Mar 1995.

[5] Steven A. Moyer, V. S. Sunderam, "Characterizing Concurrency Control Performance for the PIOUS Parallel File System", Department of Math and Computer Science Emory University, Computer Science Technical Report CSTR-950601, Jun 1995.

[6] 文部省重点領域研究「超並列原理に基づく情報処理基本体系」第 6 回シンポジウム予稿集, pp4-42-4-49, Mar 1995.