PUMA-II: 分散共有メモリを用いたPVMの性能評価

1H－9

小林　伸治　陣崎　明
(株) 富士通研究所

1　はじめに

PUMA-IIはUNIXワークステーション(WS)のVMEバスに搭載したメモリ(分散メモリ)を100Mb/sのトークンリングネットワークで結合し、ハードウェアサポートによる高速な分散共有メモリを提供するシステムである(図1)[1]。

我々はPVM(Parallel Virtual Machine)[3]をPUMA-IIに移植中であるが、基本部分の開発と性能評価を完了したので報告する。今回の実装ではUNIX System Vの共有メモリ機構をPUMA-II分散共有メモリに拡張し、その上に共有メモリ対応のPVMを移植するという方針を取った。

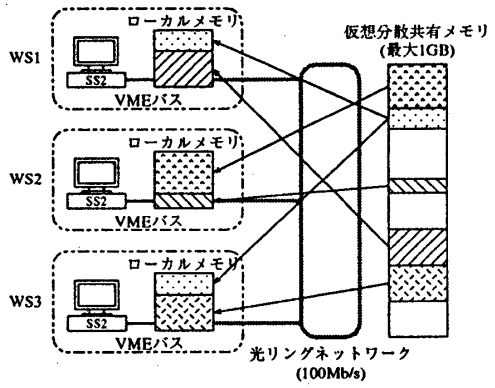

図1. PUMA-IIシステム

2　PVM

PVMはネットワークで接続された複数のWSを1つの並列計算機のように扱うWSクラスタを実現するソフトウェアである。プログラミングモデルとしてメッセージパッシングを採用し、アプリケーションはメッセージの送受信により通信を行う。

PVMはライブラリとデーモンから成り、PVMアプリケーションはライブラリ内の関数を用いてプロセス間通信を行う。PVMの多くの実装ではメッセージ通信をソケットによって実現している。ソケットを用いることで移植性が高い反面、通信オーバヘッドが大きい点が問題である。

この問題を改善するため、同一マシン内での通信にUNIX System Vの共有メモリ機構を用いたPVMの実装があり、マルチプロセッサで用いられている。そこでUNIX System Vの共有メモリ機構をPUMA-II分散共有メモリに拡張するライブラリ(System V IPC互換ライブラリ)を作成し、このライブラリを用いてPVMを移植することとした。

3　分散共有メモリによるPVMの実装

3.1　System V IPC互換ライブラリ

UNIX System Vの共有メモリ機構は、IPC(Inter Process Communication)の1つとして実装されている。shmgetシステムコールを用いて共有メモリの識別子を獲得し、shmatシステムコールでプロセスの仮想空間にマップする。shmgetへの引数として同一のキーを与えたプロセスが同一のメモリを共有することができる。我々のライブラリでは現在のところSystem V IPCの内、共有メモリとセマフォを実装し、ユーザIDによるアクセス権チェックなどは省略している。

3.2　PUMA-IIデバイスドライバ

PUMA-IIは分散共有メモリ方式としてネットワーク仮想記憶方式(NET-VMS方式)を採用している。NET-VMS方式では各ノードに分散したメモリの間でページ単位のコピーや無効化などのネットワーク操作を行い、全体を1つの共有メモリとして構成する。

PUMA-IIノードは仮想的な分散共有メモリ空間のページごとに、そのページが有効であるか、他ノードがコピーを持っているか、他ノードからのネットワーク操作を受け付けるか、自ノードのどの物理アドレスに対応しているかといった情報を含むタグを持つ。また、プロセッサはページ単位でCopyin、Copyout、Unifyという3種類のネットワーク操作を要求できる。タグの構成とネットワーク操作を表1にまとめる。

NET-VMS方式では、あるページにアクセスするためにはあらかじめロックを取得する必要がある。アプリケーションプログラムがPUMA-IIデバイス

PUMA-II: Evaluation of a PVM implemented on Distributed Shared Memory
Shinji Kobayashi, Akira Jinzaki
Fujitsu Laboratories Ltd.
1015 Kamikodanaka, Nakahara-ku, Kawasaki 211, Japan

表 1. PUMA-II のタグとネットワーク操作

タグビット	意味
VLD	ページが有効である
CPY	コピーを他ノードが持っている
PVT	プライベートページである
ALC	ページが割り当てられている
LCK	ネットワーク操作を受け付けない

ネットワーク操作	意味
Copyin	ページを他ノードから取得する
Copyout	ページを他ノードに供給する
Unify	他ノードのページを無効化する

ドライバに対してロックの取得を要求すると、デバイスドライバは必要に応じて Copyin や Unify などのネットワーク操作を行った後に、タグのロックビットを立てて他ノードからのネットワーク操作を受け付けないようにする。

ロック取得によるブロックを最小限に抑えるため、書き込み用のロックと読み出し用のロックとを別に用意した。書き込み用のロックは 1 つのプロセスしか同時には取得できないが、読み出し用のロックは同一ノード内で複数のプロセスが同時に取得することができる。

PUMA-II では、通信処理やタグ管理処理のハードウェアサポートによりデバイスドライバレベルで毎秒 2500 回以上のネットワーク操作が可能であり、分散メモリ間のデータ転送性能は伝送速度 100Mbps のネットワーク上で 10MB/s 以上と高速である。

4 性能評価

PUMA-II 分散共有メモリ版 PVM(PUMA-II-PVM) の性能を予測するため、System V IPC 互換ライブラリの基本性能評価を行った。測定には SparcStation2(SunOS4.1.1) を用いた。

2 ノード間でデータの送受信を行った場合の性能を図 2 に示す。伝送速度 100Mbps のネットワーク上で最高 5.59MB/s のユーザレベル通信性能を達成した。PUMA-II と同等のネットワーク伝送性能をもつ FDDI で Socket を用いた場合 [4] と比較して 2 ～ 3 倍の実効性能を実現しており、プロトコルの軽い分散共有メモリベースの通信が高速であることがわかる。

PUMA-II-PVM は現在デバッグ中であるが、PVM のオーバヘッドを考慮しても 3 ～ 4MB/s のユーザレベル通信性能が得られる見込みである。

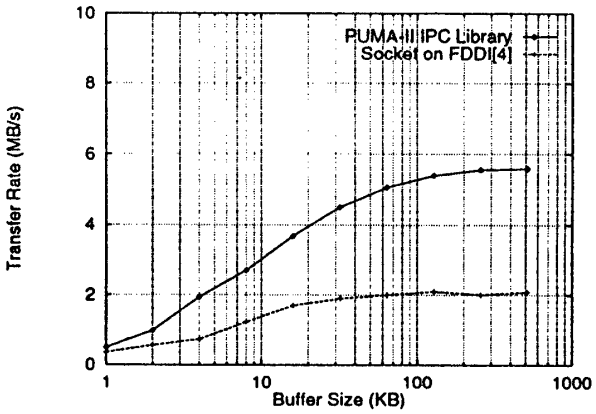

図 2. System V IPC 互換ライブラリの性能

5 まとめ

UNIX System V の共有メモリ機構を分散共有メモリに拡張した。これを PUMA-II 上にライブラリレベルで実装し、性能測定した結果、Socket に対して 2 ～ 3 倍のユーザレベル通信性能を実現できることを確認した。本システムで分散共有メモリ版 PVM を実現すれば 100Mbps の汎用的なネットワークを用いて 3 ～ 4MB/s という、CM-5 など高速な専用ネットワークを持つ MPP の PVM 性能をも凌駕する高速なメッセージ通信性能を達成することが期待できる。

現在 PUMA-II-PVM の移植と並行して、ワークステーションを 1Gbps のファイバーチャネルで結合した PUMA-III[2] を開発中である。PUMA-III ではハードウェアレベルで 50MB/s の実効通信性能を目標としており、今回開発した分散共有メモリライブラリや PVM などの開発環境を用いて分散を意識せずマルチプロセッサのように利用できる高性能ワークステーションクラスタの実現を目指す。

参考文献

[1] 新家他:WS 向け高速分散共有メモリシステムの試作と評価 — ハードウェアアーキテクチャ —、情報処理学会第 48 回全国大会 (1B-5), 1994

[2] 新家他:PUMA-III: 1Gbps ファイバーチャネルを用いた高速分散共有メモリシステム, 本大会予稿 (1H-10), 1995

[3] V.S.Sunderam 他:The PVM concurrent computing system: Evolution, experiences, and trends, Parallel Computing 20, 1994

[4] M.Lin 他:Distributed Network Computing over Local ATM Networks, IEEE Journal on Selected Areas in Comm., 1995(予定)