ＷＳ向け高速分散共有メモリシステムの試作と評価
1B－5　－ハードウェアアーキテクチャー

新家正総　揚野祐三†　マックス・モリス　陣崎　明
（株）富士通研究所　†富士通ディジタル・テクノロジ（株）

１．はじめに

既存のUNIXワークステーション（WS）に接続可能な分散共有メモリPUMA-IIを試作し性能評価した。PUMA-IIは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、最高11.5MB/sのメモリ間ページ転送速度を実現する。分散共有メモリ制御にネットワーク仮想記憶（NET-VMS）方式[1]を採用することにより、高速転送性能とWSソフトウェアの簡略化を実現している。本稿ではPUMA-IIのハードウェアアーキテクチャと基本性能の評価結果について述べ、更に高速化検討を行う。

２．PUMA-II開発の狙い

分散共有メモリが実用的であるためにはメモリアクセス性能の高速性が必須条件である。この高速性を実現するために我々はNET-VMS方式を提案し、複数の試作によって有効性の検証を行ってきた[1]。しかしこれらの試作はメモリ性能を重視するため計算機と分散共有メモリを一体化しており、既存のWSに接続できなかった。一方、分散共有メモリの適用分野を拡大し、有益なソフトウェアの開発を促進するためには既存のWSで利用可能な汎用的な分散共有メモリの実現が重要な課題となる。そこでPUMA-II開発ではIEEE標準バスであるVMEバスベースで、既存のWSに実用的な性能を提供できる分散共有メモリの実現を狙いとした。

３．ハードウェアアーキテクチャ

PUMA-IIノード（図１）はWS（SUN-SS2）、VMEサイズの制御ボード、VMEメモリからなる[図２]。制御ボードのハードウェアは次の特徴をもつ。

(1)分散共有メモリ制御をハードウェアで高速に行う。
(2)UNIXなどタイムシェアリングOSから分散共有メモリを効率的に制御できる。

Implementation and Evaluation of a High-Speed Distributed Shared Memory System for Workstations (Hardware Architecture)
Tadafusa Niinomi, Yuzo Ageno†, Max Morris, Akira Jinzaki
Fujitsu Laboratories Ltd.
1015 Kamikodanaka, Nakahara-ku, Kawasaki 211, Japan
†Fujitsu Digital Technology Limited
2-3-9, Shin-Yokohama, Kohoku-ku, Yokohama 222, Japan

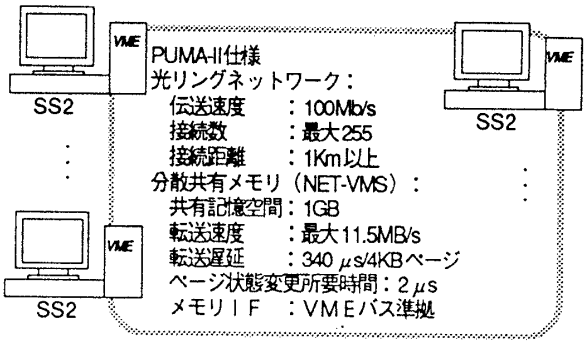

図1　システム概要

(1)についてはNET-VMS方式[1]で実現した。分散共有メモリでは、ネットワーク上に分散したメモリ間でコンシステンシを維持するため、データの移動、複写、削除を行う必要がある。これらの操作をハードウェアで効率的に行うため、NET-VMS方式では共有メモリ管理テーブル内部にメモリページの状態を表すタグ（ページ状態タグ）を設ける[図２]。通信・メモリコンシステンシ制御回路はネットワークからのデータ操作要求に対してページ状態タグを参照し、パケットがネットワークを通過する間に要求の受付判断、データの転送制御、ページ状態変更を行う。これらの機能によりハードウェア的にコンシステンシ制御をサポートする。

(2)についてはUNIXからのPUMA-II制御のインターフェースとしてコマンドFIFO（C-FIFO）と応答FIFO（R-FIFO）を設けた。C-FIFOの深さは410、R-FIFOの

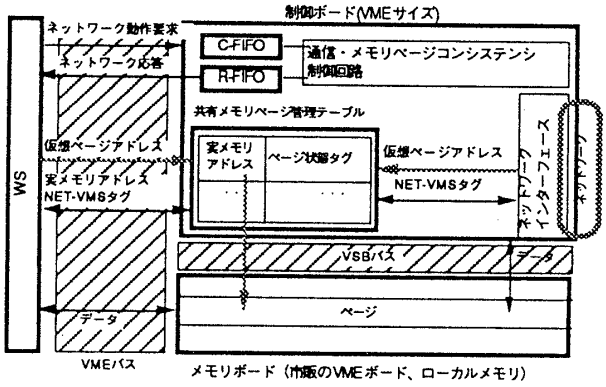

図2　PUMA-IIノードの構成

深さは512である。FIFOへのアクセスは1プロセッサ命令で行える。また、任意の命令が実行された直後やR-FIFOに一定数の応答が格納された時にWSへ割り込む機能がある。

4．基本性能評価

基本性能評価は、SUN-SS2上でテストプログラムを動作させて行った。

表1にPUMA-IIのネットワーク処理性能を示す。ページ転送（COPY）の実効転送速度11.5MB/s（物理転送速度の92%）、4KBページ転送遅延340μsを実現した。リモートページの無効化（UNIFY）は3μsである。C-FIFOに格納した410ページ分（1.6MB分）のページ転送要求を0.14秒で処理可能である。

図3はワードカウント（WC）プログラムを実行した時の処理時間を示す。40MBの処理データをWSのメインメモリ、PUMA-IIのローカルメモリ、PUMA-IIのリモートメモリに置いた場合について調べた。測定の結果、ローカルメモリ、リモートメモリでの処理時間はメインメモリのそれぞれ1.2倍、1.3倍であった。この結果から当初の狙い通り、SS2のメインメモリ性能に匹敵する分散共有メモリを実現できたことがわかる。

5．高速化の検討

近年WSの性能は急速に向上している。より高速なWSに対応するには、PUMA-IIを更に高速化する必要がある。そこで今回の評価結果を基に高速化の検討を行う。

図3において、ローカルメモリとメインメモリの差はVMEバスのオーバーヘッドであり、ローカルメモリとリモートメモリの差はネットワークオーバーヘッドである。高速化の為にはこれらのオーバーヘッドを短縮する必要がある。

まずVMEバスのオーバーヘッドは、VMEバスをSBusなどの高速な標準バスに置き換えれば削減できる。SBusは20MB/s程度であるから、WC処理ではVMEバスのオーバーヘッドを少なくとも63%改善できる。

次に今回の性能測定結果をもとに、ネットワークを高速化した場合の実効通信性能を予測した[図4]。例えば1Gb/sのファイバーチャネルを用いれば74.4MB/sの性能が得られることがわかる。この場合、WC処理ではネットワーク処理を84%以上改善できる。

以上述べた改善を行うと、WC処理で分散共有メモリのオーバーヘッドを8秒から2.4秒以下に削減できることがわかった。

6．おわりに

本論文ではPUMA-IIのハードウェアアーキテクチャを述べ、基本性能の評価と高速化検討を行った。その結果、PUMA-IIはSS2クラスのWSで実用的な分散共有メモリであることを確認した。またバスとネットワークの高速化により、さらに高速なWSにも適用可能な見通しを得た。今後は1Gbpsファイバチャネルを用いた高速化を進める予定である。

表1　ネットワーク処理性能

	転送遅延	転送速度(MB/s)
COPY	340μs/4KB	11.5MB/s
UNIFY	3μs	—

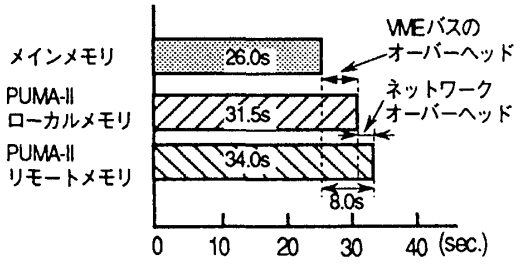

図3　ワードカウントプログラム実行時間
(40MBデータ処理、4B単位でアクセス時)

[参考文献]

[1]A Jinzaki:"A Fast Distributed Shared Virtual Memory System: NET-VMS", FUJITSU Sci. Tech. J., Vol. 29, No.3, Sep. 1993, pp.286-295.

[2]小林他：WS向け高速分散共有メモリシステムの試作と評価　-ソフトウェア性能評価-、本大会予稿

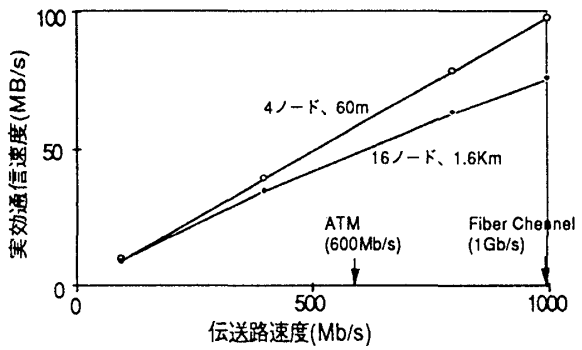

図4　ネットワーク高速化時の性能予測
(4KBページ転送時の実効通信性能)