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

1H－10

新家正総　揚野祐三†　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

標準的なワークステーションをネットワークで結合し、分散／並列処理可能な大規模システムを構築するワークステーション（WS）クラスタが注目されている。PVMなど分散処理プログラミング環境の開発、ワークステーションの高速化、そしてファイバーチャネルやATMなどの高速ネットワークの普及に伴い、WSクラスタへの期待は非常に強まっている。

我々は現在、1Gbpsのファイバーチャネルを用いた分散共有メモリによる高性能WSクラスタPUMA-IIIを開発している。PUMA-IIIでは高速なネットワークを活かし、分散型マルチプロセッサのように利用できる高性能WSクラスタの実現を目指す。本稿ではPUMA-IIIの技術的ポイントについて検討した結果を述べ、目標性能を示す。

2．分散共有メモリによるWSクラスタ

WSクラスタ開発の重要なポイントは、標準技術をベースにネットワーク経由のプロセス間通信を高速に実現することである。しかし現状の通信技術では、ネットワークソフトウェアのオーバーヘッドのために、ユーザレベルの通信で物理ネットワークの通信性能を100%活用できない問題がある。

この問題を解決するために、我々は分散共有メモリベースで高速なプロセス間通信を行うことを提案し、分散共有メモリをハードウェアサポートにより高速に実現するネットワーク仮想記憶方式（NET-VMS方式）を開発した。NET-VMS方式の試作システムPUMA-IIでは100Mbpsのネットワークを用いてSparcStation2を結合し、ハードウェアレベルで最高11.5MB/s、UNIXユーザレベルで最高5.6MB/sという高速なプロセス間通信性能を実現している[1]-[3]。

この高速性の要因は大きく二つある。まず分散共有メモリの実現に必要な通信機能は比較的単純化できるためハードウェアサポートが容易である。PUMA-IIのハードウェアレベル性能がネットワーク伝送性能の90%以上に達している理由は基本的な通信機能のハードウェア化にある。

PUMA-III: High Speed Distributed Shared Memory System Using 1Gbps Fibre Channel
Tadafusa Niinomi, Yuzo Ageno †, 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

もう一つの要因は分散共有メモリではユーザ空間に直接マップされた共有メモリを介してプロセス間通信を行うため、通信に伴うデータのコピーやUNIXシステムの介在を削減することができる点である。PUMA-IIのユーザレベル性能がsocketを用いた場合の2～3倍となっている理由はこのオーバヘッド削減にある。

このようにファイバーチャネルやATMなどの高速なネットワークの性能を活かし高性能なWSクラスタを実現するための基盤として、分散共有メモリは非常に有望と考えられる。

3．高速ネットワークによる分散共有メモリ

ここでは現在使用可能な最高速の標準ネットワークである1Gbpsファイバーチャネルを用いることを想定してNET-VMS方式を実現するための基礎検討を行う。

NET-VMS方式は、分散共有メモリを実現するためのメモリコヒーレント制御をブロードキャストを用いて高速に行うことが特徴である。PUMA-IIはこのブロードキャストを行うのにトークンリングネットワークを用いている（図１）。このネットワークではリング上の１回のフレーム周回でブロードキャストでき、かつ戻ってきたフレームを見て通信の成否を確認できるので高速で確実なメモリコヒーレント制御ができる。

そこでファイバーチャネルで同じように効率的なブロードキャストと送達確認が可能かが重要となる。ファイバーチャネルではファブリックによるスター型のネットワーク接続とファブリックを用いないリング型の接続方式がある。どちらの場合もブロードキャストや送達確認は、FC-2（フレーミング、フローコントロールを行う）とFC-3（複数ポートにまたがる機能を実現する）のレイヤで規定される。しかしながら、ブロードキャストでは送達確認の機能が仕様上無い。

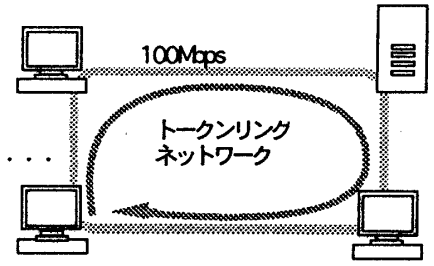

図１　PUMA-IIのネットワーク

従ってファイバーチャネルを用いて高速な分散共有メモリを実現するためには、どちらのトポロジの場合もFC-2、FC-3レイヤの一部の機能を変更して送達確認の可能なブロードキャストを行う必要がある。

次に問題となるのはI/Oバスの性能である。ファイバーチャネルの性能は双方向通信で200MB/sであり、I/Oバスもそれに見合った速度にしなければネットワークの速度を活かせない。ところが現状の標準的なワークステーションバスではこの速度に対応できない。例えばSBusは32bit転送時のピーク性能で80MB/sである。新しい標準であるPCIバスは最高132MB/sだが、それでもファイバーチャネルの速度に対応できない。1Gbpsクラスのネットワークを活用するには最低でも200MB/s以上のI/Oバスが必要である。

また1Gbpsネットワークではメインメモリとのデータ転送にDMAを用いることが必須となるため、DMA機能も重要である。例えばSBusではDVMA（Direct Virtual Memory Access）をサポートしているが、DVMA空間の大きさが制限されるためメインメモリの任意の領域にDMAするためにはページマップの切り替えを必要とする。このような仕様は性能低下の要因となりうる。

4．PUMA-III

上記の検討に基づき1Gbpsファイバーチャネルによる分散共有メモリを用いた高性能WSクラスタPUMA-IIIの目標仕様を決定した（表1）。ブロードキャストと送達確認についてはファイバーチャネルFC-2、FC-3レイヤの一部の機能を変更することとした。当面はループ構成で実現し、最終的にはファブリック構成で接続数の増大を実現する（図2）。

I/OバスとしてはプラットホームのSBusを用いることとした。現時点で1Gbpsネットワークに対応できる標準的なバスは存在しないため、一般性を基準に選んだ。WSクラスタは標準的なワークステーションによって構築することが重要なポイントである。SBusは先に述べたように性能上、機能上の問題をもつが、このような問題に対応することも必要である。

図3にPUMA-IIIワークステーションの内部構成を示す。性能的にはSBus DMA性能およびDVMA空間の制御オーバーヘッドやデータ転送がネックとなる。性能検討の結果、ハードウェアレベルの実効スループットは50MB/s、ユーザレベルではSuperSparc(50MHz）マシンでやはり50MB/sと予測している。この性能はネットワーク性能と比較すると十分なものではないが、絶対性能的には専用ネットワークを用いた超並列マシンの通信性能を超えるものであり、高性能WSクラスタ実現の第一歩としては十分魅力的と考えている。

表1　PUMA-IIIの目標仕様

●WS：	SBusをもつWS（Sun-WSなど）
●OS：	UNIX
●並列ソフトウェア：PVMなど	
●メインメモリ間スループット：50MB/以上	
●分散共有メモリハードウェア：	
- ネットワーク共有仮想記憶空間	1GByte
- ページサイズ	2KByte
- 物理メモリ	Main Memory使用
- ページ状態変更遅延	2μs（3ノードループ時）
- 最大接続数	127
- 使用ネットワーク	Fibre Channel
- 接続形態	リング型／スター型
- ネットワーク転送速度	100MB/s
- 接続I/Oバス	SBus
- 最高DMA転送速度	80MB/s（32bit転送時）

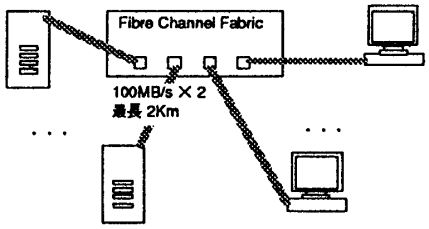

図2　PUMA-III: Fibre ChannelベースNET-VMS

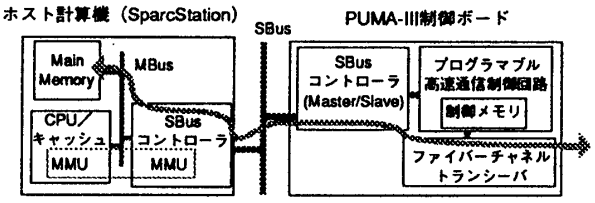

図3　PUMA-IIIノードの構成

5．まとめ

1Gbpsのファイバーチャネルを用いた分散共有メモリの実現について検討した。高速な分散共有メモリ実現のためにPUMA-IIIではファイバーチャネルの仕様の一部を独自仕様とし、I/Oバスは一般性を考慮しSBusを採用することとした。PUMA-IIIのメモリ間データ転送性能は50MB/sと見積もっている。

PUMA-IIIは現在ループ構成のハードウェアをSBusシングルスロットアダプタとして設計中である。

[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ーハードウェアアーキテクチャー,第48回情報処理学会全国大会, 1B-5,(1994-3)

[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ーソフトウェア性能評価ー,第48回情報処理学会全国大会, 1B-6,(1994-3)

[3]小林他：PUMA-II:分散共有メモリを用いたPVMの性能評価,本大会予稿,1H-9