PUMA-III における1 Gbps FCネットワークの実現技術

5B－1

新家　正総　陣崎　明
（株）富士通研究所

1. はじめに

ワークステーション（WS）をネットワークで結合し分散／並列処理を行うWSクラスタが注目されている。高速標準ネットワークの普及に伴い、WSクラスタへの期待はより高まっている。

我々は標準ネットワークによる高性能WSクラスタの実現を目指し、1Gbpsファイバーチャネル[1]（FC）を用いたWSクラスタPUMA-III[2]を開発している。

高性能なWSクラスタ実現の鍵はプロセス間通信の高速化である。しかし標準ネットワークの利用では独自ネットワークと比較してオーバーヘッドがかかる点が問題となる。PUMA-IIIの最終的な課題は、標準ネットワークを用い、なおかつ高速なプロセス間通信を実現することである。

今回PUMA-III開発の第一ステップとして、FCの機能のうちFC1以下のレイヤを利用するネットワークアダプタを開発した。本稿ではこのアダプタハードウェアに焦点をあて、検討と開発の結果を述べる。

2. アダプタハードウェアの検討

WSクラスタのプロセス間通信はスループットと遅延の両面で高速性を実現しなければならない。そこでアダプタハードウェアの要件としてはアダプタ内部の処理の高速化とアダプタインタフェースの高速化が重要である。

しかしながらこれらの要件を実現する上で、アダプタを実現する方式には様々な選択肢がありうる。まずアダプタ内部の処理方式については、おおまかには制御CPUを用いるか専用ハードウェアで制御するかの選択がある。またアダプタインタフェースの方式でも、まずDMAとプログラムI/Oの選択肢があり、更にDMAの場合にはDMAの終了認識の方式としてポーリングと割り込みの選択肢がある。以下ではこれらの点のうち、DMAとプログラムI/Oの選択に焦点を絞って考える。

DMAはCPUの負荷なしに効率良く高速転送が可能である。しかしメッセージが非常に短くDMA起動処理時間が無視できない場合や、転送データがDMA領域に置けない場合はプログラムI/Oの方が性能が優るかもしれない。例えばイリノイ大のメッセージ通信方式であるFast Messages[3]（FM）では、短いメッセージの性能に重点を置き、送信にプログラムI/Oを用いている。しかしながら、どの程度の長さのメッセージにおいてプログラムI/OよりDMAが劣るかは、DMAの起動方式、I/Oアクセス性能、DMA性能、ソフトウェアオーバーヘッドなどに依存するため一般に明らかではない。

我々は今回DMA方式でSBusを用いたFCアダプタを開発した。以下ではSBus FCアダプタについて概要を述べ、DMAがどの程度有効なのか評価した結果を述べる。

3. SBus FCアダプタ

SBus FCアダプタ（図1）は、SBusのDVMA空間上のデータをDMAで送受信する。DMAの起動、終了確認は以下のようにして行う。

送信の場合はバッファアドレスと転送サイズをアダプタにセットしDMA起動をかける。

受信の場合は複数のDMAアドレスをアダプタが保持できるように受信アドレスFIFOを用意した。受信アドレスは64個まで設定できる。また受信済バッファごとにDMA末尾アドレスがわかるように受信RESULT FIFOを設けた。受信認識はポーリングで行う。

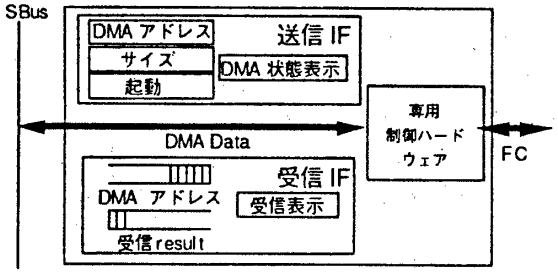

図1　SBus FCアダプタ

アダプタの内部は専用の制御ハードで構成した。制御ハードはファイバーチャネルのプロトコル階層におけるFC0（物理インタフェースとメディア）、FC1（encode、decodeなど）の機能をもつ。パケット形式はFC2と異なる独自方式であり、ホストから受け取ったデータにパケットの先頭、末尾の識別子を付加してネットワークに転送する。

Implementation of 1Gbps FC Network for PUMA-III
Tadafusa Niinomi, Akira Jinzaki
Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki 211-88, Japan

4. SBus FC アダプタの評価

アダプタの評価は、1つの送信バッファから同じデータを送り続ける試験（独自プロトコル[4]による試験）とイリノイ大の Fast Messages を我々のハードウェア上に移植した FC FM [4] を用いて行った。イリノイ大の FM は送信にプログラム I/O を用いているが、我々は新たにユーザ空間からの直接 DMA で実装している。

独自プロトコルではソフトウェアオーバーヘッドがほとんど無い。送信起動のための 3word の I/O アクセスと、ポーリングだけがオーバーヘッドとなっている。FM ではシーケンス番号管理などのソフトウェアオーバーヘッドが若干必要である。

測定に用いた WS は SPARCstation2（CPU Weitek 倍速、SBus 58MB/s）と SPARCstation5（CPU microSPARC、SBus 35MB/s）である。比較として用いた Myrinet[3] によるイリノイ大FM は SPARCstaion20 を使用している。

図2に各種プロトコルのスループットを示す。独自プロトコルでは SS2 で最大 52MB/s、SS5 で 28MB/s の性能を得た。SS5 の方が遅い理由は2つある。1つは SS5 の DMA の最大バーストサイズが SS2 の半分の 32byte のためである。もう一つは SS2 では DMA 転送中は CPU からのポーリングがほとんど入らないのに対して、SS5 ではポーリングで DMA がブロックされているためである。

次に FC FM とイリノイ大 Myrinet FM を比較する。 SS2 ではほぼ 512byte 以上、SS5 ではほぼ 256byte 以上で FC FM が Myrinet FM より高性能であり、2KB での性能はそれぞれ 40MB/s 、25MB/s となる。また Myrinet FM より性能の低い領域でも、SS5 ではほぼ Myrinet FM に近い性能が出せている。このことから、256byte 以下の小さいメッセージサイズでも、DMA でプログラム I/O に近い性能が出せることがわかる。

図3に FC FM の遅延を示す。SS5 で Myrinet FM とほぼ同等の性能が得られた。SS2 では CPU 性能が低いために小さなメッセージサイズで Myrinet FM より遅延が大きくなっているが、512byte 以上では DMA の効果により逆に Myrinet FM よりも小さな遅延となっている。

5. まとめ

DMA 方式の SBus FC アダプタは SS5 で最大 25MB/s 以上の性能が得られ、256byte 以下の小さなメッセージにおいてもイリノイ大 FM に近い性能が得られる。また SS2 での最大性能は 40MB/s 以上である。これらの結果から DMA 方式のアダプタは WS クラスタのネットワークアダプタとして

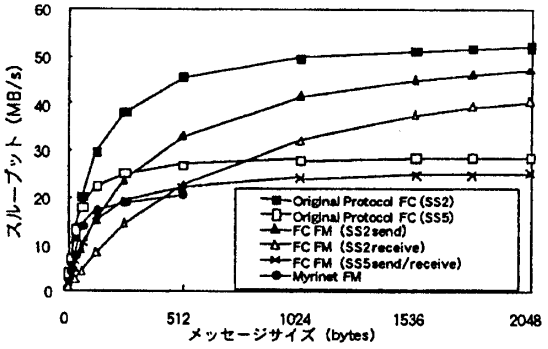

図2 スループット性能

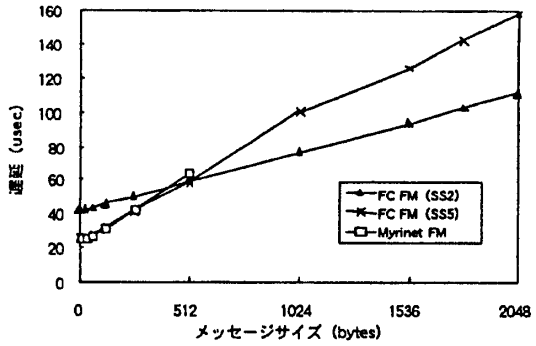

図3 遅延

有望なことがわかった。

6. 今後の予定

今回開発した SBus FC は FC1 レイヤまでのサポートになっており、FC2 以上の複雑な機能を実現することができない。そこで FC2 より上位のプロトコル処理をハードウェアで行う処理エンジンを開発中である。プロトコル処理ハードウェアでは、FC だけではなく TCP/IP などの処理をアシストすることも考えている。

また FC1 以下のアダプタとしても、SBus のボトルネックを解消するために PCI バスを用いた FC ボードを開発中である。PCI の場合理論性能で 125MB/s 程度が得られ、より高速な通信性能が期待できる。

参考文献

[1]FIBRE CHANNEL: PHYSICAL AND SIGNALING INTERFACE (FC-PH) REV4.3, ANSI, June 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,第 50 回情報処理学会全国大会,1H-10,(1995-3)

[3]Scott Pakin 他：High Performance Messaging on Workstations: Illinois Fast Messages (FM) for Myrinet, Supercomputing, Dec. 1995

[4]小林他：PUMA-III における各種メッセージプロトコルの実装と評価, 本大会予稿