PUMA-IIIにおける各種メッセージプロトコルの実装と評価

5B－2

小林　伸治　　陣崎　明
(株) 富士通研究所

1　はじめに

PUMA-IIIは1Gbps Fibre Channel (FC)で結合したワークステーションクラスタである。PUMA-IIIのSBus FCアダプタ[1]は約58MB/sの帯域を有する。このように高速なネットワークを最大限に活用するためには、オーバヘッドの少ない通信プロトコルが不可欠である。そこで、PUMA-III上にいくつかのプロトコルを実装して性能を評価し、高速ネットワーク向けにプロトコルの拡張を行った。

ネットワークの性能を計る指標としては帯域(Bandwidth)と遅延(Latency)を採用した。これらは一般に転送サイズの関数となる。

2　各種プロトコルの実装

今回実装したプロトコルのモデルを図1に示す。

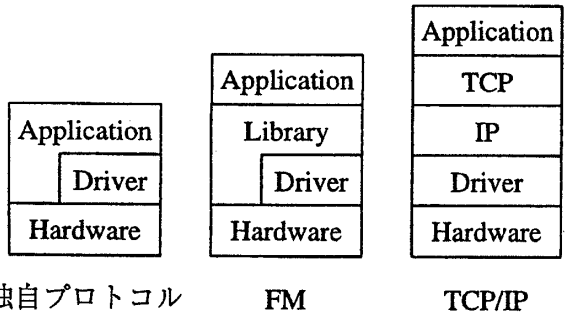

図1. プロトコルの階層モデル

2.1　独自プロトコル

ハードウェアが持つ帯域を最大限に活かす方式として、ユーザ空間にネットワークアダプタを直接マッピングして操作する方式を実装した。データバッファはDMA領域にもマッピングし、直接DMAが行えるようにする。この方式を独自プロトコル方式と呼ぶ。独自プロトコル方式では、DMA転送中であることを示すハードウェアレジスタをビジーウェイトでチェックして次のDMA転送が可能になるまで待つ。この方式はハードウェアが持つ送信帯域を最大限に利用できるが、アプリケーション側の負担が大きい。

2.2　Illinois Fast Messages

Illinois大学が開発したFast Messages (FM)[2]は、ワークステーションクラスタなどで並列計算を行うためのメッセージパッシング方式のプロトコルである。SIMD(Single Instruction stream Multiple Data stream)のプログラミングモデルに基づき、アクティブメッセージなどと同様、受信側で起動されるハンドラを送信側が指定する。FMは小さいメッセージを低遅延で転送することを主眼とし、ネットワークアダプタを直接ユーザ空間にマッピングする、送信にはDMAを用いない、受信側で割り込みを使わないといった方針を採っている。

送信にDMAを用いない理由は、Sunのアーキテクチャでの通常のDMAは遅延が大きいためである。Sunのアーキテクチャではdmaは仮想アドレスに対して行われ、その範囲は仮想アドレス空間の上位1MBに限られている。そのため、ユーザ空間上にあるデータバッファをDMA転送するにはこの領域にコピーし直すかマッピングする必要がある。Illinois大学はこれらを避けるために送信にはDMAを用いないアプローチを選択している。

これに対して、PUMA-IIIにおけるFMの実装では送信にもDMAを用いることにした。SBusはDMA転送を用いると約61MB/s(20MHzクロックの場合)の帯域を持つのに対して、プログラムI/Oの帯域は約20MB/s(同上)に制限されてしまうためである。しかし、DMAのためにメモリ間コピーを行っていては遅延も大きくなり、帯域を活かすことができない。そこで、FMにデータ領域割り当てのプリミティブを追加し、メッセージ転送を行うデータはこの領域上に確保したデータに限るようにプロトコルを変更した。FMのライブラリはDMA領域にもマップされたバッファを確保し、ユーザプロセスにマッピングしてからその先頭アドレスを返す。送信時はデータコピーやマッピングの操作が不要で、DMAのオーバーヘッドはハードウェアに対するレジスタアクセスのみとなる。上記の拡張を行ったFMをゼロコピーDMA方式FMと呼ぶ。

Implementation and evaluation of message protocols on PUMA-III
Shinji Kobayashi, Akira Jinzaki
Fujitsu Laboratories Ltd.
4-1-1 Kamikodanaka, Nakahara-ku, Kawasaki 211-88, Japan

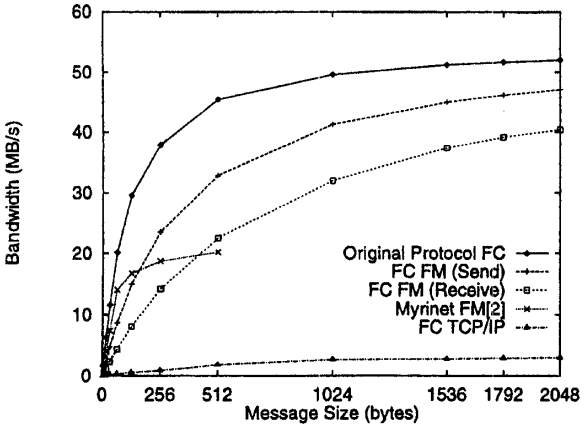

図 2. 各種プロトコルの帯域

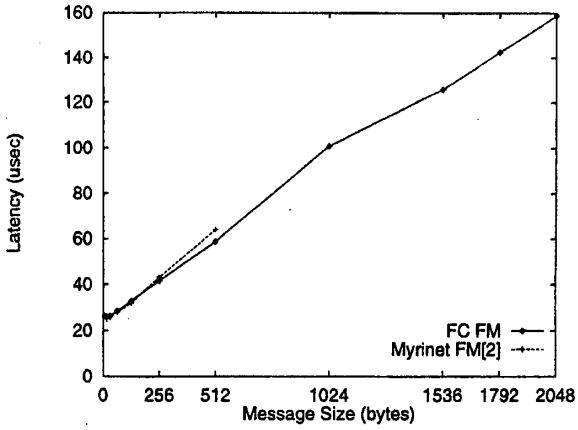

図 3. FM の遅延

2.3 TCP/IP

TCP/IP はプロトコル処理が重いので 1Gbps FC のような高速ネットワークの性能を発揮できないが、比較のために実装を行った。

3 各種プロトコルの評価

各種プロトコルの帯域を図 2に示す。測定は、CPU を Weitek 倍速プロセッサに載せ替えた SPARCstation 2 を用いて行った。SBus クロックは 20MHz である。Myrinet FM の測定には SPARCstation 20 が使用されている。

独自プロトコルではハードウェアの送信帯域に近い、最高 50MB/s 以上の性能を実現している。

ゼロコピー DMA 方式 FM では、DMA による送信を採用しながら低遅延を実現できた。このため最高 40MB/s 以上の帯域が得られている。これに対して Myrinet FM は送信をプログラム I/O で行っているために SBus の帯域が制限され、最高 20MB/s 程度である。

なお、FM は送信要求後にデータバッファの変更が自由に行えるセマンティクスになっているため、ゼロコピー DMA 方式では送信完了まで送信要求から返ることができない。送信完了を知る手段を別途用意し、送信要求から返っても送信完了まではデータバッファを変更できないセマンティクスに変更すればさらに効率化できる。

遅延に関しては、図 3に示すように Myrinet FM と同等以上の性能が得られた。遅延データの測定には SPARCstation 5 を用いた。

TCP/IP では、ユーザ空間からシステム空間内の mbuf 構造体、mbuf 構造体から DMA 領域へと 2 回のデータコピーが必要になる。TCP のコネクション状態管理など、データコピー以外のプロトコル処理オーバヘッドも大きい。このため、TCP/IP の帯域はハードウェア帯域に比べて非常に低くなっている。

4 まとめ

今回実装したゼロコピー DMA 方式 FM は、Myrinet 版 FM と比較してより大きな帯域と同等以上の遅延を実現できた。1Gbps FC などの高速ネットワークを活かすためには、FM のような軽いプロトコルが欠かせない。さらに、送信要求と送信完了検知とを分離した API を採用すれば、より効率化できる。ワークステーションクラスタなど高速な通信が必要とされるアプリケーションに広く応用できる。

TCP/IP はデータコピーやプロトコル処理のオーバヘッドが非常に大きい。TCP/IP を 100MB/s クラスのネットワークに適用するためには、プロトコル処理の一部をハードウェアで実現するなど、大きな改良が必要である。

参考文献

[1] 新家他:PUMA-III における 1Gbps FC ネットワークの実現技術、本大会予稿

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