6ZA-07 PCクラスタシステムFireClusterにおける効率的な大容量通信の実現と評価

佐藤 喬　兵頭 和樹　中山 泰一
電気通信大学 情報工学科

1 はじめに

近年、PCやEthernetなどのコモディティハードウェアを用いたPCクラスタシステムの開発が増えており、その柔軟性や優れたコストパフォーマンスが注目されている[1]。

我々は汎用高速シリアルバスのIEEE 1394を用い、コストパフォーマンスの向上をはかったPCクラスタシステム**FireCluster**[2]の実装をしている。

現在、**FireCluster**はユーザレベル通信により低遅延通信を実現している。これはカーネル空間に存在する固定サイズの通信バッファをユーザ空間にマッピングすることで、カーネルの介在なしに通信が行えるよう実現されている。

従来の**FireCluster**上で大容量のデータ通信を行った場合、通信バッファへ頻繁にデータコピーが発生し、ハードウェア性能を活かしきれていないと考えられる。

本稿では、通信バッファへのデータコピーによるオーバヘッドを低減するため、Pin-downキャッシュ[3]を用いたゼロコピー通信を実現する。また実験により効率的な大容量通信が可能になることを示す。

2 Pin-downキャッシュ

FireClusterはIEEE 1394のPhysical Writeと呼ばれる、DMAを用い直接相手の物理メモリにデータを書き込む機能により通信している。通信中のデータは常に同一物理メモリ上に存在する必要があるため、データの仮想アドレスと物理アドレスとのマッピングを固定し、ページングされないようにするPin-downと呼ばれる操作をしなければならない。

現時点では図1のように、Pin-downされた通信バッファを通信するノードごとに割り当て、それを介してPhysical Writeを行っている。この方式では大量のデータを通信する場合、通信バッファへのデータコピーが頻繁に発生し、性能低下の原因になる。

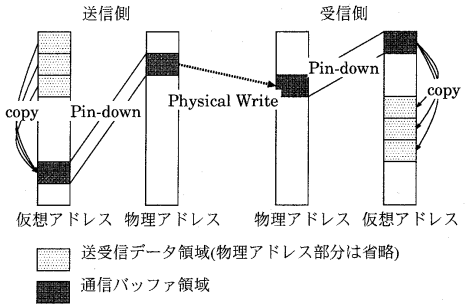

図 1 現在の通信方式

この解決策としてゼロコピー通信が考えられる。これはデータ領域を直接Pin-downし、通信バッファを介さずに通信する方法である。まず送信に使用するデータ領域を送受信双方でPin-downし、受信側が送信側へデータ領域の物理アドレスを教える。送信側は教えられた受信側の物理アドレスにPhysical Writeし、その後Pin-downを解除する。ただし、Pin-down操作にはカーネルコールが必要となるため、オーバヘッドが発生する。毎回Pin-down操作が発生することは、ユーザレベル通信の利点を損ねてしまう。そこでカーネルコールを減少させるためPin-downキャッシュと呼ばれる技術が有効である[3]。

図2のようにPin-downキャッシュでは、Pin-downされた領域の情報を保持しておく。通信終了後のPin-down解除操作では、使用済の印をつけるだけで実際の解除を行わない。Pin-downされた領域がユーザに許された一定量を超えた場合に解除を行う。通信を行う場合には、データ領域が既にPin-downされているかを調べ、Pin-down操作が必要か判定する。Pin-downされた領域の情報はユーザプロセスに公開しているため、Pin-downが必要かどうか判定をすることにカーネルコールは必要無い。これによりカーネルコールを減少さ

Implementation and evaluation of an effective high bandwidth communication on "FireCluster" the PC cluster system

Takashi SATOU, Kazuki HYOUDOU and Yasuichi NAKAYAMA
Department of Computer Science, The University of Electro-Communications
E-mail: satou-t@igo.cs.uec.ac.jp

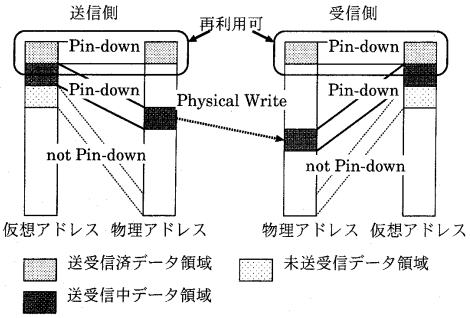

図 2 新しい通信方式

表 1 ノード PC の仕様

OS	Linux 2.0.36
プロセッサ	Intel Celeron 366MHz
メモリ	64MByte
IEEE 1394	IOI-1394TTO 400Mbps
Ethernet	DEC 21140 チップ搭載 100Mbps

せる。

3 実験

通信ノードに用いた PC の仕様を表 1 に示す。FastEthernet はファイルシステムの共有など、システム運用に必要な通信に使用する。

以上のシステムを用いて、1 対 1 片方向送信時のスループットを測定する。従来の通信と Pin-down キャッシュを用いた通信とを比較した結果を図 3 に示す。

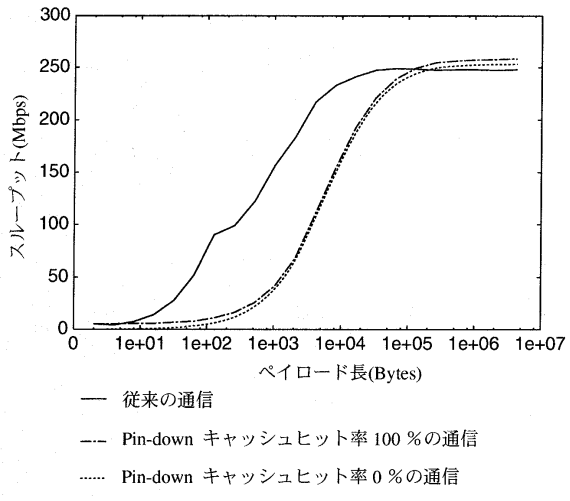

図 3 ペイロード長 対 スループット

Pin-down キャッシュのヒット率とは、通信データ領域が既に Pin-down されている割合を示す。ヒット率が 100%の場合には、通信時に Pin-down 操作のためのカーネルコールが発生しない。ヒット率が 0%の場合には、通信時に Pin-down 操作のためのカーネルコールが発生する。

ペイロード長が 256KB を超えた状態では、Pin-down キャッシュのヒット率に関係なく Pin-down キャッシュを用いた通信の方が高いスループットを得ている。ペイロード長が 4096KB でのスループットは、従来の通信では 248.1Mbps であるのに対して、Pin-down キャッシュのヒット率 100%において 254.5Mbps、ヒット率 0%において 253.1Mbps であった。

ペイロード長が 256KB 以下においては、従来の通信の方が高いスループットを得ている。従来の方式と本稿で述べたゼロコピー通信をペイロード長により切替える方法で、効率の良い通信を実現できると考えられる。

4 おわりに

本研究では、通信バッファへのデータコピーによるオーバーヘッドを低減するため、Pin-down キャッシュを用いたゼロコピー通信を **FireCluster** に実装し、実験結果を報告した。

ペイロード長が 256KB を超えた状態では、Pin-down キャッシュのヒット率に関係なく Pin-down キャッシュを用いた通信の方が高いスループットを得ることができた。

今後の予定として、以下のことが挙げられる。

- Pin-down 領域を送受信同士で効率良く通達しあうプロトコル
- Pin-down 領域の情報を高速に検索できるデータ構造とアルゴリズム
- Pin-down 領域を効率良く解除するアルゴリズム

参考文献

[1] 石川: コモディティハードウェアを用いた並列処理技術, 情報処理,Vol.39,No8,pp.784-791(1998).

[2] 兵頭, 佐藤, 中山: IEEE 1394 を用いた PC クラスタシステムの設計, コンピュータシステム・シンポジウム論文集,pp.169-176(1999).

[3] Tezuka et al.: Pin-down Cache: A Virtual Memory Management Technique for Zero-copy Communication, *Proc. Int'l Parallel Processing Symp.*, IEEE CS Press, Los Alamitos, Calif., pp.308-314(1998).