IEEE 1394 による PC クラスタシステムの設計

3F−1

山之内 暢彦　兵頭 和樹　南 将朝　中山 泰一
電気通信大学 情報工学科

1 はじめに

複数の計算機を何らかの通信手段で接続し、仮想的に並列計算機を構成して高い計算処理能力を引き出す、クラスタシステムの開発事例が近年増えてきた。とりわけ、従来は大型計算機が必要であった場面で、PC や Ethernet などのコモディティハードウェアから作られた PC クラスタが利用され、そのシステムの柔軟性やコストパフォーマンスの高さが注目されている [1]。

本研究で提案する PC クラスタ "FireCluster" は、クラスタの構成要素となる各計算機ノードに PC を使用し、汎用高速シリアルバスの IEEE 1394 [2] でそれらを接続する。この IEEE 1394 は、PC やマルチメディア機器の間を接続する通信方式の次世代標準とされ、400Mbps の高速な通信速度を実現する。

既存の PC クラスタに関する研究においては、Myrinet や ATM といった高速な通信方式を前提としているものが多く見られるが、機器が高価で必ずしも一般的とは言えない。それに対し IEEE 1394 は、いくつかのトレードオフがあるものの、低コストかつ容易に PC クラスタを構築することができる。

本稿では、FireCluster の概要を述べるとともに、通信性能に関する予備実験を通して IEEE 1394 がクラスタシステムに十分適用できることを示す。

2 FireCluster の概要

クラスタシステムにおいては、各計算機ノード間で頻繁に通信を行うため、そのレイテンシやスループットがクラスタ全体の性能を大きく左右する。従って、ハードウェア、ソフトウェアの両面で高い通信性能を持つことが必要とされる。

FireCluster は、IEEE 1394 のコントローラに対応したデバイスドライバ、通信プリミティブ、そして、メッセージパッシング機構を提供する汎用通信インタフェースの 3 つのレイヤで構成される（図 1）。特に、以下の機能を持つ。

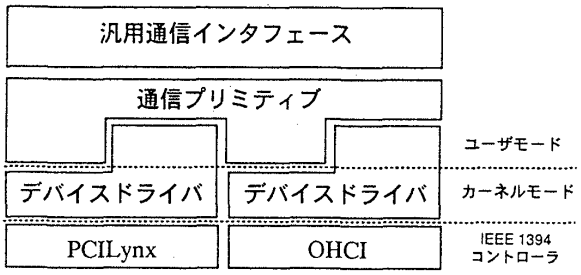

図 1 FireCluster の基本構成

- **ユーザレベル通信**

ユーザレベル通信 [3] とは、通信コントローラのレジスタやバッファ領域などをユーザプロセスのメモリ空間に直接マッピングすることで、OS のカーネルを介さずにデータ通信を行う技術である。これにより、通信時のアドレス空間の切り替えなどが起きずレイテンシを抑えることができ、クラスタシステムにおいて非常に有効である。また、ドライバのカーネルモード部分への呼び出しは OS を介さず高速に行う。

- **DMA 転送・ゼロコピー通信**

ホストメモリと IEEE 1394 バスとの間のデータコピーにおいて、PC のプロセッサを使わず IEEE 1394 コントローラ上の DMA 機能を利用する。これにより、プロセッサ資源をより有効に活用できるようになる。さらに、OHCI [4] の機能を利用して、デバイスドライバとユーザプロセスとの間のデータコピーを省くゼロコピー通信 [1] の実現を目指す。

3 IEEE 1394 の特徴

計算機ノード間の通信手段として見た場合、IEEE 1394 は以下のように特徴付けられる。

- **価格性・入手性**

PC 99 仕様 [5] の中で IEEE 1394 は "Required" とされており、近い将来大半の PC に標準搭載されることが予想される。従って、何ら他

A Design of PC Cluster System Employing IEEE 1394
Nobuhiko YAMANOUCHI, Kazuki HYOUDOU, Masatomo MINAMI and Yasuichi NAKAYAMA
Department of Computer Science, The University of Electro-Communications
E-mail: bon@igo.cs.uec.ac.jp

の機器を用意することなく、ケーブルのみで PC 間を接続することが可能になる。

- **高速性・将来性**

十分低コストでありながら、Fast Ethernet の 100Mbps を大きく上回る 400Mbps（半二重）の通信速度と低いレイテンシを実現しており、クラスタシステムの通信手段として適したデバイスと言える。さらに、800Mbps や 1.6Gbps、3.2Gbps の規格（IEEE 1394.B）が現在ドラフト段階にある。

- **最大ノード数の制限**

IEEE 1394 は、Ethernet などと比べ最大ケーブル長が短く、ノードの数が原則的に 63 に限られる。このため、計算能力を高めるためには、ノード数の増加による通信性能の低下を抑え、同時に SMP 型計算機への対応などが必要となる。

- **非インテリジェントなコントローラ**

IEEE 1394 のコントローラは、Myrinet のようにプログラマブルではないため、通信に関する処理の多くをソフトウェアで行わなければならない。従って、通信のレイテンシやプロセッサの使用率を低く抑えるためにソフトウェア上の工夫が求められる。

4 通信性能に関する予備実験

作成したデバイスドライバを用い、IEEE 1394 の基本的な通信性能を計測した。使用したマシンは Intel Celeron プロセッサ（300MHz）搭載の PC/AT 互換機、OS は Linux 2.0.36。IEEE 1394 のコントローラには、リンク層に TI PCILynx を使用した Unibrain 社製の PCI ボード（最高 200Mbps）を 2 枚用いた。

通信におけるレイテンシの指標となるパケットのラウンドトリップ時間（図 **2**）は、ペイロード（パケット中のユーザデータ部分）の長さが最も小さい 4 バイトのとき 52μ 秒である。同様の環境で 100BASE-T（コントローラは DEC 21140A）を用いて測定した場合は 48μ 秒であり、これよりやや遅い程度になった。

スループット（図 **3**）は、最大ペイロード長の 1,024 バイトのとき 114Mbps であったが、この性能に達するには 100BASE-T と比べてペイロード長を大きく取る必要がある。短いデータが頻繁に行き交う通信では帯域をフルに利用できない。

レイテンシは 100BASE-T 並みの性能を得たが、スループットは必ずしも良い結果が出たとは言えず、今後の課題である。

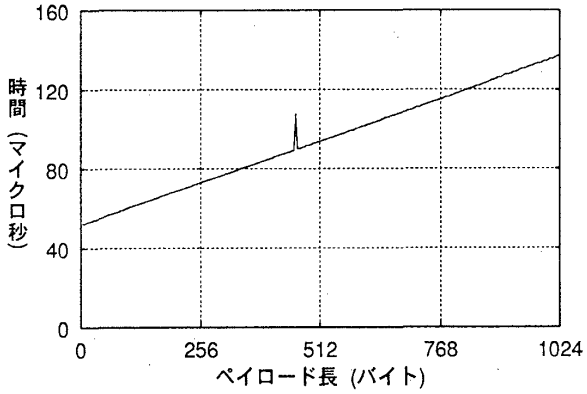

図 2 ペイロード長 対 ラウンドトリップ時間

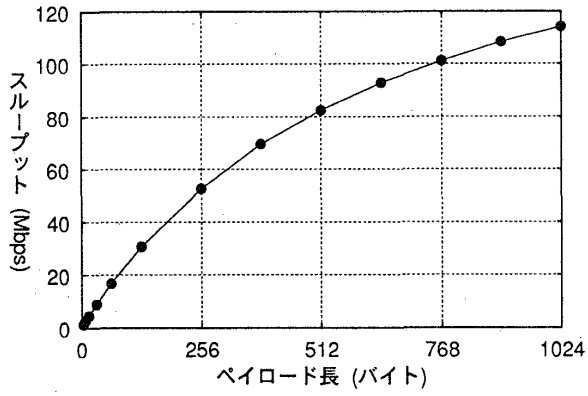

図 3 ペイロード長 対 スループット

5 おわりに

本稿では、容易に PC クラスタを構築する手段として IEEE 1394 の利用を提案した。今後はデバイスドライバの改良やメッセージパッシング機構の実装を行い、クラスタシステム全体としての性能を評価していく予定である。

参考文献

[1] 石川 裕: 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, No.8, pp.784–791 (1998).

[2] Don Anderson, MindShare, Inc.: *FireWire System Architecture*, 2nd Edition, Addison-Wesley (1998).

[3] Matt Welsh *et al.*: ATM and Fast Ethernet Network Interfaces for User-level Communication, *Proc. 3rd International Symposium on High Performance Computer Architecture*, IEEE (1997).

[4] The Promoters of The 1394 Open HCI: *1394 Open Host Controller Interface Specification*, Release 1.00 (1997).

[5] Intel and Microsoft: *PC 99 System Design Guide*, Version 1.0 (1998).