6D-5

疎結合計算機クラスタのインターコネクト性能評価

落合 真一、村山 和宏、山口 義一

三菱電機(株) 情報技術総合研究所

1. はじめに

音声・ビデオデータ処理、センサデータ処理などは大量なデータに対する大規模な演算を必要とする。近年、演算処理の高度化と、汎用マイクロプロセッサの急速な性能向上から、このような処理を汎用プロセッサの並列計算機により実現したいという要求が高まっている。疎結合の計算機クラスタを利用すれば、システムのスケーラビリティ、構成の柔軟性を高めることができる。本稿では大量データ処理向け計算機クラスタのシステム設計検討のため、疎結合並列演算で問題となるプロセッサ間のデータ交換について、ギガビットネットワーク、バックプレーンバス、汎用 LAN などの種々の通信路を比較し性能評価を行う。

2. システムモデル

計算機クラスタのプロセッサ間の通信路には種々のものがあり、それぞれ得失がある。システム要求に応じた選択が必要となる。

- バス: ネットワーク、スイッチなどの外部装置が不要
  VME バス、コンパクト PCI バスなど
- システムエリアネットワーク(SAN): 高速通信が可能
  ファイバーチャネル、Myrinet など
- 汎用 LAN: 多種の装置を接続可能
  イーサネットなど

システム構成の例を図 1に示す。

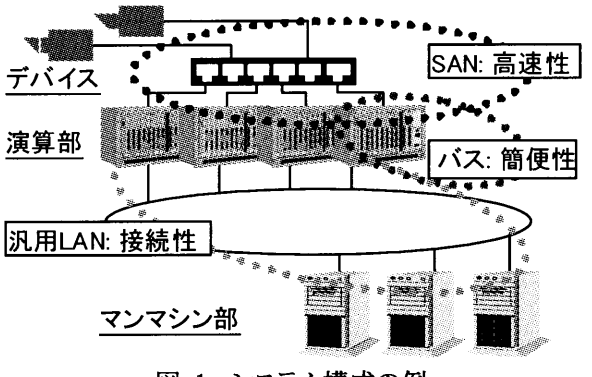

図 1: システム構成の例

本稿ではこのようなシステムを構築する際の検討項目の一つとして、各種インターコネクトの通信オーバヘッドの評価を行う。

3. 評価項目

本稿では特に、データ交換による応答遅延時間に着目して性能測定を行い、考察する。性能測定における評価の観点を以下に挙げる。

(1) メッセージサイズによる応答遅延時間の変化
(2) 通信路の種別による応答遅延時間の差異
(3) 通信プロトコルによる応答遅延時間の差異

本稿での評価対象には以下を選択した。

a) 通信路の種別:
   100M イーサネット、1G イーサネット、バックプレーンバス通信
b) 通信プロトコルの種別:
   TCP、UDP、VIA(ポーリング)、VIA(ブロッキング)

評価の H/W 構成を表 1に示す。

表 1: 測定環境

CPU	Pentium III 600MH z
100M イーサネット	100BaseT ハブ(半二重)
1G イーサネット	ギガネット社 cLAN(VIA 対応)
BP バス通信	コンパクト PCI バス 33MHz、32 ビット

4. 測定結果

4.1. 応答遅延時間

1 対 1 のプロセッサ間通信で、固定長のメッセージを交換した場合の片道の応答遅延時間を測定した。各種通信路での測定結果を図 2に示す。

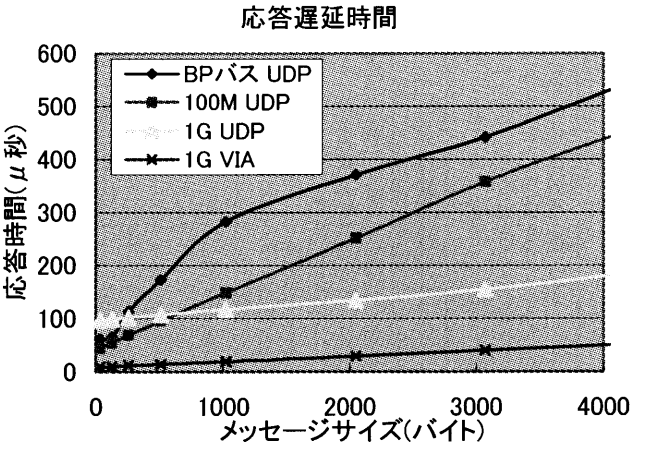

図 2: 応答遅延時間の比較

An Evaluation of Various Interconnect for Parallel Computing Clusters
Shinichi OCHIAI, Kazuhiro MURAYAMA, Yoshikazu YAMAGUCHI
Mitsubishi Electric Corporation

UDP による 32 バイトメッセージの応答遅延時間は、100Mイーサネット45μ秒、1Gイーサネット97μ秒、バックプレーン通信 65μ秒で、通信路の物理的転送性能差が応答遅延時間に現れていない。これはプロトコル処理のオーバヘッドが性能を決定する要因となっているためである。それに対しVIAを1Gイーサネットで使用して通信を行った場合は 9μ秒と通信による遅延を大きく短縮できる。VIAにより通信に対する OS オーバヘッド削減の効果が得られる。

データ交換のメッセージサイズが大きくなると、応答遅延時間に対する通信路の物理転送性能の影響が現れる。留意が必要なのは BP バス通信で、物理転送性能は 133M バイト/秒ともっとも高いにも関わらず、プロセッサ間通信で使用した場合の応答遅延時間は100Mイーサネットと同程度以下と良くない。これはプロセッサによるデータ転送では、バストランザクションの単位データサイズが小さいため効率が悪いからである。性能を向上させるためには、プロセッサのデータコピーによらないバス上のデータ転送を実装する必要がある。

4.2. 通信方式の比較考察

各種通信方式の応答遅延時間の比較を図 3に示す。

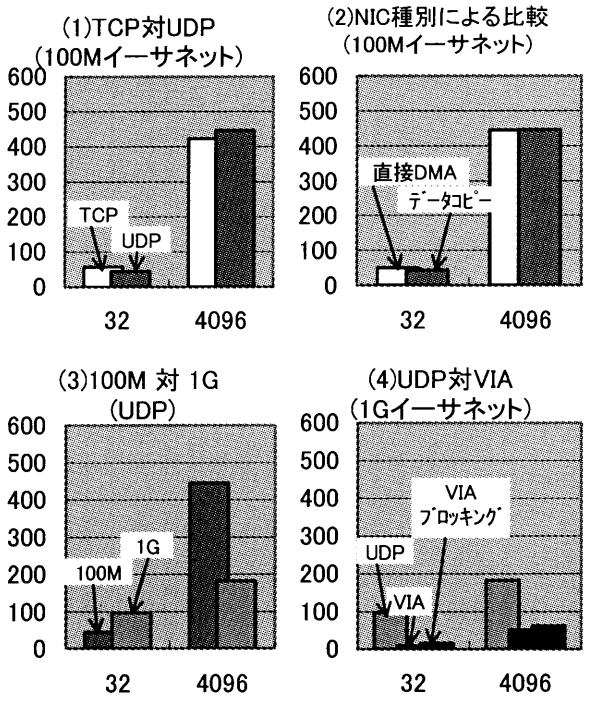

図 3: 各種通信方式による応答遅延時間の差
(縦軸 応答時間(μ秒)、横軸 メッセージサイズ(バイト))

(1) TCP と UDP の遅延時間比較

TCP (NO DELAY オプション設定)とUDPの応答遅延時間を比較すると、メッセージサイズが小さい場合は UDP の方が高速であるが、メッセージサイズが大きい場合はウィンドウ制御によるデータ転送の最適化により TCP の方が高い性能を示す。これは通信路が高速であるほど顕著である。ただし、通信処理にかかる CPU 負荷はTCPの方が2倍近く高く、本評価環境では40%を超えるCPU時間を消費する。

(2) NIC 種別による応答遅延時間比較

NIC にはデータ送信時に、プロトコル処理バッファから直接 DMA 転送を行うものと、専用領域にデータコピーしてからDMA転送を行うものがある。100 M イーサネットの物理転送性能ではデータコピー回数を減らすことによる応答遅延時間短縮の効果は大きくない。しかし、送信処理に対するCPU負荷は1/2に削減できる。

(3) 100M イーサネットと 1G イーサネットの比較

UDP による通信では、メッセージサイズが小さいと通信路の物理転送性能が 10 倍になっても応答遅延時間は短縮しない。メッセージサイズが 2K バイトを超えれば通信路の高速化の効果が得られるが、その場合でもプロセッサのプロトコル処理がボトルネックとなり、物理転送性能 10 倍に対し、応答遅延時間は 1/2〜1/3 である。

(4) UDP と VIA の比較

プロセッサでプロトコル処理を行う UDP と比較して、VIAを適用すると通信路の物理転送性能に応じた遅延時間短縮が得られる。ポーリングを行えばコンテキストスイッチがなくなるために応答遅延時間がもっとも短くなるが、プロセッサは通信処理に占有される。ブロッキングの場合には、100μ秒以下の間隔で発生する割込みに対して確実に応答できるリアルタイム OS が必要となる[1]。

5. おわりに

本評価により以下の知見を得た。

- 汎用プロトコルを用いた通信では、メッセージサイズが小さいならば通信路の物理転送性能の高速化による応答遅延時間短縮の効果は小さい。
- PCI バスのような大容量通信路を用いても、CPU によるデータ転送では応答遅延時間は短縮されない。
- ギガビットを超える高速ネットワークではVIAのようなゼロコピー、ユーザレベル通信の技術を利用することにより通信路の物理転送性能を有効化できる。

今後、この評価結果をもとにシステム設計を行っていく。

参考文献

[1] 落合、村山、山口、"VI Architecture による分散リアルタイム環境の構築"、情報処理学会 DPS 研究会、2000 年 3 月