高速インターコネクトにおけるOSオーバーヘッド削減方式の検討

2H-04

村山 和宏、落合 真一、山口 義一
三菱電機(株) 情報技術総合研究所

1.はじめに

近年、ネットワークハードウェアの進歩に伴い、分散ソフトウェアの通信時のオーバーヘッドが明らかになってきた。このオーバーヘッドの主な原因は、ネットワーク資源のほとんどがカーネルによって管理されており、ネットワークアクセスに関するほとんどの処理がカーネルトラップを必要とするからである。

そこで、OS のオーバーヘッドを削減する通信方式について研究が行われており、Intel、Compaq、Microsoft が 1997 年に発表した Virtual Interface Architecture (VIA) [1] や、UC Berkeley の U-Net[2]、Active Message[3]などがある。それらの中でVIAは、

・仕様を公開しており、今後の高速インターコネクトの主流となる可能性が高い
・MPI(Message Passing Interface)の開発も行われており、従来の並列アプリケーションの移行が容易であるなどの特徴がある。そこで、本稿では、OS のオーバーヘッドを削減した通信機構として VIA を取り上げ、VIA が高速インターコネクトとして実用的なものであるのか、評価を行った。

2. VIA性能測定

VIA は、ネットワークインタフェースをユーザ空間に直接マップし、カーネルのオーバーヘッドなしに NIC を利用できるという特徴を持つ。

表 2-1:測定環境

CPU	DUAL Pentium II Xeon 450MHz
チップセット	440GX
メモリバス	100MHz
PCI バス	32bit
OS	Windows NT 4.0 Workstation
VIA NIC	giganet cLAN 1000 (バンド幅 1.25Gb/sec)

図 2-1 は、表 2-1 における環境で、さまざまなメッセージサイズでの通信性能をグラフ化したものである。今回の測定においては汎用PCを使用し、32ビット PCI バスに VIA NIC を接続したため、PCI バスとネットワークのバンド幅のうち、小さいほうが通信バンド幅となる。32ビット PCI バス、ネットワークの転送理論値はそれぞれ 133MByte/sec、156.25MByte/sec(1.25Gbit/sec)であるため、本測定環境での転送性能の理論値は 133MByte/sec となる。しかし、本測定環境では最高通信性能が84MByte/sec であり、理論値との差が大きい。そこで、この性能劣化の原因について検討を行った。次章以降で検討内容について述べる。

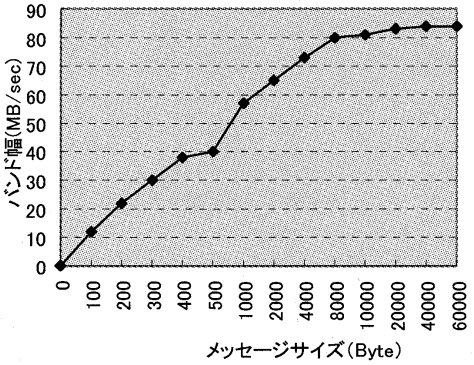

図 2-1:メッセージサイズとバンド幅の関係

3. VIAにおける通信処理解析

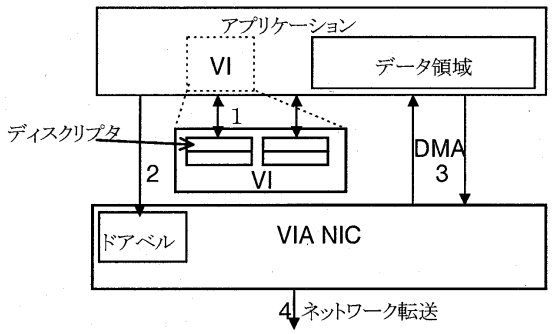

図 3-1:VIA の構造

ここでは、2 章で発見された通信性能劣化の原因について考える。

データ転送は以下の手順で実施される。

1. ディスクリプタをプロセスのユーザ空間上にある送信キューに追加する。(図 3-1 中矢印1)
2. ドアベルによってVIA NIC にデータ送信を指示する。(同 2)
3. アプリケーション空間から VIA NIC 上に DMA 転送する。(同 3)

An evaluation of the Virtual Interface Architecture for high-speed communication
Kazuhiro MURAYAMA, Shinichi OCHIAI,
Yoshikazu YAMAGUCHI
Mitsubishi Electric Corporation

4. VIA 処理を行い、受信側計算機にデータ転送を行う。(同 4)

この一連の処理手順で処理を行う際、ボトルネックとなりうるものとしては、以下の項目がある。

・アプリケーションメモリ空間上処理による遅延
・PCI バスの DMA 転送による遅延
・VI A 処理による遅延

本測定環境においては、VIA 処理は専用ハードウェアにて実現しているため、遅延時間は十分小さいと判断する。そこで、残った要因である、アプリケーションの関数呼び出しによる遅延とPCIバスのDMA転送による遅延について検討を行う。

4.ボトルネックの検討

4.1 アプリケーションによる遅延の検討

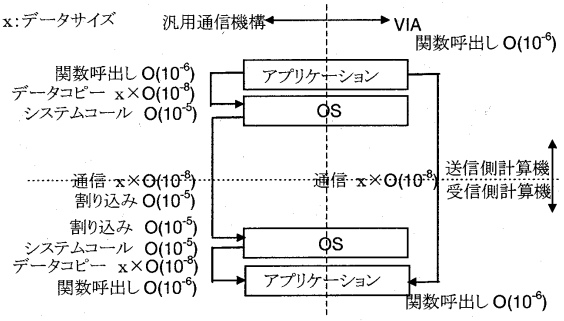

図 4-1:各処理の遅延時間(オーダー)

分散アプリケーション間でのデータ通信に要する処理時間(オーダー、単位:秒)を図 4-1 のように仮定する。

従来の通信機構は図中左側の経路で通信を行うため、トータルの遅延時間は$O(10^{-5})$以上となる。

一方、VIA は、図 4-1 中右側の経路で通信を行うため、データ転送遅延時間は $x * O(10^{-8})$または$O(10^{-6})$となる。ピークの通信性能について考えると、図 2-1 によると、VIA は 20KByte 以上の場合に通信速度が最高(84MByte/sec)であり、その場合の遅延時間は 240 μ 秒以上となる。一方で、アプリケーションの関数呼出しによる遅延時間は$O(10^{-6})$であることから、アプリケーションの関数呼び出しは通信遅延時間よりも十分小さい。このように、VIA では、OS によるオーバーヘッドを排除したことにより、ソフトウェアが性能劣化の原因とはなっていない。

4.2 DMA 転送による遅延の検討

4.1 より、通信速度がピークの場合にはアプリケーションによる遅延が無視できるため、本測定環境での性能劣化の原因は、PCI バス、チップセットにあると考えることができる。そこで、DMA 転送の遅延について考えると、遅延の原因は、主に以下の二つが考えられる。

・PCI バス上のデータ転送
本測定環境では、実際に使用できる PCI バスのバンド幅が理論値よりも極端に小さい可能性がある。
・チップセットの性能
VIA NIC として cLAN 1000 を使用した場合、440LX(メモリバス 66MHz)、440BX(メモリバス 100MHz)の通信性能はそれぞれ95-100MByte/sec、90-95MByte/sec[4]であり、メモリバスの狭い 440LX のバンド幅のほうが広い。チップセットの性能によっても転送データ量が増減する可能性がある。

このような、DMA 転送によるボトルネックを解消するためには 64 ビット PCI バスを使用し、高性能なチップセットを使用する必要がある。

5.おわりに

今回の検討により、以下の結論を得た。

・VIA では、アプリケーションが原因で性能劣化が起きることはなく、VIA 自体は、現状のハードウェア性能をそのまま通信性能に反映している。
・現在の VIA のボトルネックは、PCI バス、チップセットにある。

今後は、分散アプリケーション作成に向け、ライブラリを含めた VIA の評価を行い、また、 PCI バスの 64 ビット化により通信性能の向上を図る。

参考文献

[1] "Virtual Interface Architecture Specification. Version1.0", *Compaq, Intel, and Microsoft Corporations* (1997) .

[2] T. von Eicken, Anindya Basu, Vineet Buch and Werner Vogels, "U-Net: A User-level Network Interface of parallel and Distributed Computing", *Proc. of the 15th ACM.* (1995) .

[3] T. von Eicken, D. E. Culler, S. C. Goldstein, and K. E. Schauser,"Active Messages: a Mechanism for Integrated Communication and Computation", *Proc. of the 19th Int'l Symp on Computer Architecture*, (1992).

[4] Giganet cLAN Installation, Verification and Troubleshooting, CTC Support Training (1999).