The 53rd Meeting of Society of Polymer Science Japan (2004)

高速通信ソフトウェア搭載の PC クラスターによる高度並列計算

核融合研（連携研究センター） 田中基彦

[緒言] PC クラスター計算機は低コストで、スーパーコンピュータ並みの高い演算性能を発揮する。しかし、その多重並列化において顕在化する、TCP/IP プロトコルによるプロセッサ間通信の大きな待ち時間（Latency）はあまり認識されていない。実際これが全計算時間の 30－40%を占めることも珍しくない。筆者は、イーサネット下で機能する非 TCP/IP 通信のフリーソフトウェア GAMMA（Genoa Active Message Machine)を採用し、さらにリナックス(Linux) 非標準の高速 C/Fortran コンパイラを実際の応用プログラムに適用した。この方法により、お金をかけずに PC クラスターの潜在能力を 100%引き出し、プロセッサ間データ通信が頻繁に発生する応用プログラムにおいてその演算性能を最大で約５割アップすることに成功した。実際の並列環境で Pentium 4 (3.0GHz)は、価格で 10 倍異なる同数の並列 RISC マシン（1.5GHz）と同じ実効性能をもち、とくに量子多体系の複雑な力を計算する緊密結合型の第一原理分子動力学計算では、ベクトル型並列スーパーコンピュータに並ぶ性能を示す。

[方法] 高速演算を行う並列計算機のハードウェアには、(1) 高速のプロセッサ(CPU)、(2) CPU への高速なデータ供給、(3) 分散メモリをもつ CPU 間の高速な通信、のすべてが要求される。ここでは、Pentium 4 を利用することで条件(1)と(2)に対応し、条件(3)についてはソフトウェアである GAMMA と MPI [1] をインストールして用いる（詳細な手順は文献[2] に記載あり）。サポートされている NIC(ネットワークカード)は数種類で、場合によってはリナックス・カーネルのアップグレードが必要となる。さらに、高速性の追求と Fortran９０利用のため、リナックス非標準の市販コンパイラを利用するが、このとき GAMMA のライブラリとこのコンパイラ間で注意深く整合性を取ることが必要である[2]。

[結果] Fig.１は、GAMMA 通信において 2 つのプロセッサ間の通信速度が送信データ量とともに向上する様子を示す。ここでは Pentium 4 (3GHz)と 3Com996 NIC を装備したＰＣを測定に用いた。送信データ量が小さいときは、通信開始の遅延時間があるためにデータ送信能力は小さく、データ量１バイトでの 0.6Mbits/s は遅延時間の 15μs に対応している。TCP/IP プロトコルでの大きな遅延時間 100～150μs がここで 15μs に減少する。つまり小さなデータ転送が頻繁に発生するプログラムで

A Beowulf Cluster Machine Equipped with High-Speed Communication Software

Motohiko Tanaka (National Institute for Fusion Science, Toki 509-5292, Japan)

Email: mtanaka@nifs.ac.jp URL: http://dphysique.nifs.ac.jp/

Keywords: PC cluster machine, non-TCP/IP communication, GAMMA, small latency and high throughput, fast C/Fortran compilers

Abstract: A high performance PC cluster computer installed with Linux operating system and MPI (Message Passing Interface) for interprocessor communication has been constructed using a communication software GAMMA (Genoa Active Message Machine) instead of the standard TCP/IP protocol. Fast C/Fortran compilers have been exploited with the GAMMA communication libraries. This method has resulted in drastic reduction of the communication overhead and significant increase in the computation performance of real application programs including the first-principle molecular dynamics simulation code.

は GAMMA の通信性能が顕著に現れる。例えば、多次元の巨大行列を並列計算で解く場合には PC 間通信が頻繁に発生するため、通信待ち時間の短さが高速演算の鍵となる。

送信データ量が大きくなるにつれて転送能力は向上し、10^5 バイト付近で飽和する。非常に大きい

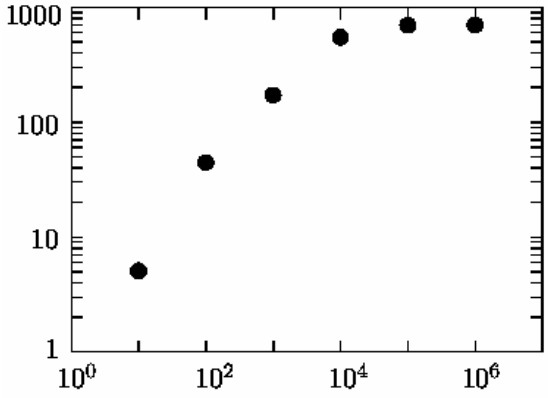

Fig.1 Relation between the transmitted data size (horizontal axis, in bytes) and the transmission speed (vertical axis, in Mbits /sec) for the GAMMA point-to-point communications. Pentium 4 (3GHz) and 3Com996 NICs are used for this timing measurements.

データ量での極限帯域幅はこの測定では 706Mbits/s であり、Gigabit NIC の最大処理能力の 70%に達し、ネットワークの能力を非常に良く利用している。GAMMA のホームページに記載されている最高値は、ハードウェアの Myrinet+BIP 通信で待ち時間と極限帯域幅はそれぞれ 4.3μs と 1005Mbits/s であり、ソフトウェアの GAMMA＋Netgear GA621 NIC でそれぞれ 8.5μs と 976Mbits/s である。

次に Table１は、異なる PC 間通信に対するプログラムの実行時間を示す。ここでは、ナノスケール量子系を扱う第一原理分子動力学コード Siesta [3] を用い、測定環境は Pentium 4 (3.0GHz) とギガビットイーサネットカード(3Com996) を備えた４台の並列 PC、および Fortran コンパイラ pgf90 である。Wallclock 時間は 1 ステップ (SCF ループ)の計算時間、Overhead 時間は Wallclock と CPU 時間の差、比は Wallclock と CPU 時間の比である。TCP/IP 通信による MPI 利用と比較して、MPI/GAMMA では、通信のオーバーヘッドが 26 秒から 0.１秒に急減、Wallclock 時間が 93 秒から 66 秒に大きく減少する(両者の CPU 時間はほぼ同じ)。さらに、最下段に示したように、MPI/GAMMA は標準的な RISC マシン（IBM Power 4、1.5GHz）の同数台並列と同じ能力をもつことがわかる。

ところで、PC クラスター計算機はプログラムによっては、（同じプロセッサ数の）スーパーコンピュータを上回る速さを示す。それは複雑な do ループをもちベクトル化が容易でない応用プログラムに対してであり、この第一原理分子動力学コード Siesta はそのひとつである。

Table 1. Timing of GAMMA (middle row) and other protocol and the RISC machine. Wallclock time is the real computation time, and Ratio = [Wallclock time]/ [CPU time].

	Wallclock time	CPU time	Overhead time	Ratio
MPI TCP/IP	93 sec	67 sec	26 sec	1.39
MPI/GAMMA	66 sec	66 sec	0.1sec	1.00
RISC machine	64 sec	64 sec	0.4 sec	1.01

[1] G.Chiola and G.Ciaccio, ”GAMMA Project: Genoa Active Message Machine” (Genoa 大学) http://www.disi.unige.it/project/gamma/

[2] M.Tanaka, ”Plasma and Ionic Condensed Matters by Molecular Dynamics Simulations”, http://dphysique.nifs.ac.jp/

[3] A. Garcia et al., Siesta (Spanish Initiative for Electronic Simulations with Thousands of Atoms) http://www.uam.es/departamentos/ciencias/siesta/