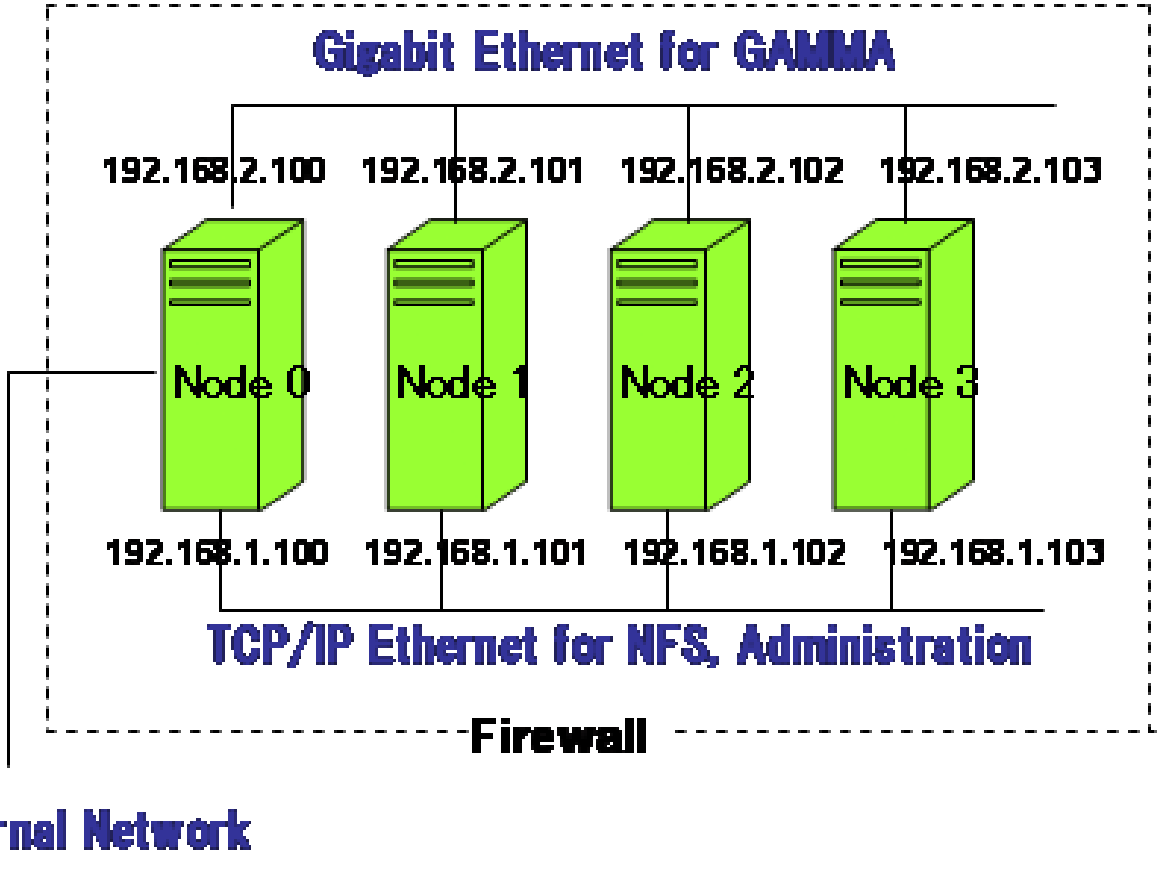
手軽に作れる研究室専用スーパーコンピュータ
高速通信ソフトウェアで動くPCクラスター計算機



田中基彦

日本物理学会誌 vol.59, No.12, pp.898-902 (2004)

手軽に作れる研究室専用スーパーコンピュータ
——高速通信ソフトウェアで動くPCクラスター計算機——

田中基彦　〈核融合科学研究所連携研究推進センター　509-5292 土岐市下石町 322-6　e-mail: mtanaka@nifs.ac.jp〉

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の速い演算速度とメモリ容量を実現する．しかしその並列利用においては，TCP/IPによるPC間通信の遅さが障害である．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，あわせてリナックス (Linux) 非標準の高速C/Fortranコンパイラを実際の研究で用いる複雑な応用プログラムに初めて適用した．その結果，PCクラスターの潜在能力をわずかな費用で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を1日単位で安定に長時間計算できること，その演算速度が最大で約5割アップすることを確かめた．並列利用時でPentium 4 (3.0 GHz) が，価格で10倍異なる同数の並列RISCマシン (1.5 GHz) に近い実効速度をもつ．また，多くのプロセッサを1つのジョブに比較的容易に割り当てられるため，領域分割可能な応用プログラムを並列で高速処理することができる．効率的な適用例として巨大行列の並列解法，物質科学では古典および第一原理（量子力学的）分子動力学が挙げられる．

1. PCクラスター高速化の問題点

『多数のパーソナルコンピュータ (PC) によりスーパーコンピュータを実現する』，ひと昔前の夢物語はいまや現実のものとなった．実際に，PCやワークステーションの余剰能力を活用するためそれらをネットワークで結び，長時間・大容量の計算をさせるグリッド計算 (grid computing) が世界各国で試みられている．これは1990年代アメリカでの，分散メモリをもつ多数の専用PCによりクラスター（並列）計算機を構成し高速計算を行う研究[1,2]の形をかえた発展といえる．日本でも大学の研究室を中心に，PCクラスター計算機の導入が行われてきた．[3,4]

PCクラスター計算機はベウルフ (Beowulf) クラスターマシンとも呼ばれ，その特徴は，スカラー型（逐次）演算器と分散メモリをもつ多数個のPCを協調して作業させ，全体として高い演算速度と大きなメモリ容量を実現することにある．このため汎用品を利用して低コストでスーパーコンピュータ並みの演算性能をもつ計算機が構築できる．私のグループでもリナックスを処理系にMPI (Message Passing Interface)[5]でPC間の通信を行うクラスター計算機を製作して，物質科学の研究に利用している．[6,7]

PCクラスター計算機に適するプログラムは，演算要素を領域（空間）分割できるものであり，多数プロセッサを用いた分割処理の効果が顕著に現れる．一般例としては，理工学の多くの分野で現れる巨大行列を空間分割により並列処理する解法が挙げられ，これに類して領域分割が可能なプログラムでも高速化が期待できる．物質理工学では重力や電磁場下での粒子（分子）動力学法がある．

さて，一般に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CPU（プロセッサ），(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．PCでは需要に応じて急速にCPUの高速化が成し遂げられ，われわれ研究者もその恩恵にあずかっている．しかし，複数のプロセッサをネットワークでつなぐ高速化は，用途が計算サーバなど科学や工学の研究分野に限られるため，PCにおいては (1) (2) に比べて発展が遅かった．もちろん，利用目的が明確で予算が豊富なスーパーコンピュータではこれらの技術は早く確立され，ワークステーションでも一部の高速計算は専用ハードウェアで実現されている．その例としてCPUでは重力場やクーロン場など中心力の計算に特化した超高速なGrapeチップ，[8] 分散メモリ間の高速通信ではMyrinet[9]が有名である．ただ，これらのハードウェアはPCクラスター自身より高価であり，研究室で保有するPCクラスターへの導入は経済性の点で疑問がある．

ところで，単体で非常に優れたコストパフォーマンスで高速計算を行うPCであるが，それをクラスターとして並列利用するための問題点はあまり認識されていない．それはOS（処理系）であるリナックスに実装されたTCP/IPプロトコルのもつPC間通信の遅さである．この原因は通信の安定性を図るために挿入される応答待ち時間に由来し，これは短いメッセージをプロセッサ間で頻繁に交換する現実の応用プログラムのクラスター計算機上での速度を大きく左右する．この状況下では，100 Mbits/sから1,000 Mbits/sに通信経路を拡張しても，演算性能の向上は見られない．

たとえば，同じCPU性能をもつPentium 4とRISCマシンについて，頻繁にプロセッサ間通信が発生するプログラムでベンチマークテストを行うと，その差は歴然である．前者では通信時間が（プロセッサの稼動を示す）CPU時間の半分を占めることもあるが，後者ではごくわずかである．例として，電子の空間相関を表すバンド行列の解法が主要な計算である緊密結合（原子基底）型の第一原理分子動力

表1　異なるPC間通信による，密度汎関数第一原理分子動力学コードSiesta[17]の実行時間．測定環境は，Pentium 4 (3.0 GHz) とギガビットイーサネット (3Com996) を備えたPCの4台並列，およびFortranコンパイラpgf90[14]である．経過時間は1ステップ (SCFループを1回) の計算に要する実時間，オーバーヘッドは経過時間とCPU時間の差，比は経過時間とCPU時間の比である．MPI TCP/IPはTCP/IP通信を利用したMPI,[19] MPI/GAMMAはここで採用した非TCP/IPの通信[10]によるMPIであり，*1と*2は通信のフロー制御がonとoffの場合である．比較のため最下段に，クラスターで運用されている標準的なRISCマシン (IBM Power 4, 1.5 GHz) を4台でMPIを用いた場合を示す．

	経過時間	CPU時間	オーバーヘッド	比
MPI TCP/IP	93 s	67 s	26 s	1.39
MPI/GAMMA	66 s*1	66 s	0.1 s	1.00
	115 s*2	98 s	17 s	1.17
RISCマシン	59 s	59 s	0.1 s	1.00

学 (first-principle molecular dynamics) 計算において，従来のTCP/IPを用いたPCクラスター計算機では表1上段に示すように，経過時間 (wallclock time) は4 CPUの並列時でCPU時間の約1.4倍である．

以下の節では，多くの研究室の手持ち資産であるパソコンとイーサネットを活用し低コストでその能力を100%発揮させるため，非TCP/IP通信で動作するフリーソフトウェアのGAMMA (Genoa Active Message Machine)[10]を導入，あわせて市販の高速C/Fortranコンパイラを適用する道筋を紹介する．その結果，PCクラスター計算機から通信の応答待ち時間 (latency) を取り除き，ベンチマークテストのレベルを越え実際の物理研究で用いる応用プログラムを，高速で長時間安定に計算処理できることを初めて示した．一方，異種PCを統合利用するSCoreプロジェクトは，極限帯域幅のベンチマークで高いスコアを記録しているが，イーサネット下での通信待ち時間は全般に大きい．[11]

2. 高速な通信ソフトウェアとコンパイラの利用

この節では，OSを経由せずにプロセッサ間で高速な通信を行うアクティブメッセージ (active message)[12]通信法に基づき，ジェノバ大学で作成された非TCP/IP通信ソフトウェアのGAMMAをインストール，高速コンパイラとあわせて利用する方法を要約する (詳細な手順は文献7を参照)．この方法では，無料または安価なリナックスOSとNIC (ネットワークインターフェース・カード)，スイッチングハブが利用できる．ただし，高い通信効率のためには，GAMMA通信専用のギガビットネットワークと，遠隔 (異なるPC上) ファイルの参照を可能とするNFS (ネットワークファイルシステム) と管理用を兼ねたTCP/IPネットワークの2系統を設置することが推奨される (図1)．

必要なプログラムは，GAMMAとMPI-1.2,[5] このプラットフォーム上でMPIの利用を可能にするインターフェースの3種である．[10] サポートされているNICはギガビットイーサネットで数種類，リナックスのカーネル (OSの基本部分，ここではgccのアセンブラ) は現在は2.4.21版である．必要な場合，カーネルのソースをコンパイルしてPC環境に組み込むが，カーネルの再構築はある程度の試行錯誤を要する．このとき旧カーネルをデュアルブート環境として保存することで再起動後のトラブルに対処することができる．GAMMAのインストールはカーネルの再構築に比べるとやさしく，コンパイルした後でその結果をカーネルに組み込む．このとき，default設定とは異なり，通信のフロー制御はオンにするべきである (第3節を参照)．インストール後にいくつかの設定を行うが，鍵はGAMMA通信に参加するすべてのPCでgamma.conf設定ファイルを作成し，PCのホスト名とNICのMAC (固有) アドレスをペアにして記載することである．[7] その後で添付のテストプログラムを走らせ，正しくネットワークが機能していればその環境下での通信速度が表示される．

次に，応用プログラムからGAMMAのプロセッサ間通信を利用するため，MPIプログラムとGAMMAとのインターフェースプログラムをインストールする．GAMMAライブラリとの整合性を図るため，ここでもリナックス添付のGNU C/Fortranでコンパイルする．GAMMA通信はプロセッサ間のブロッキング通信 (転送が完了してから次の演算動作に移る) によるデータ転送の正確さを保障して

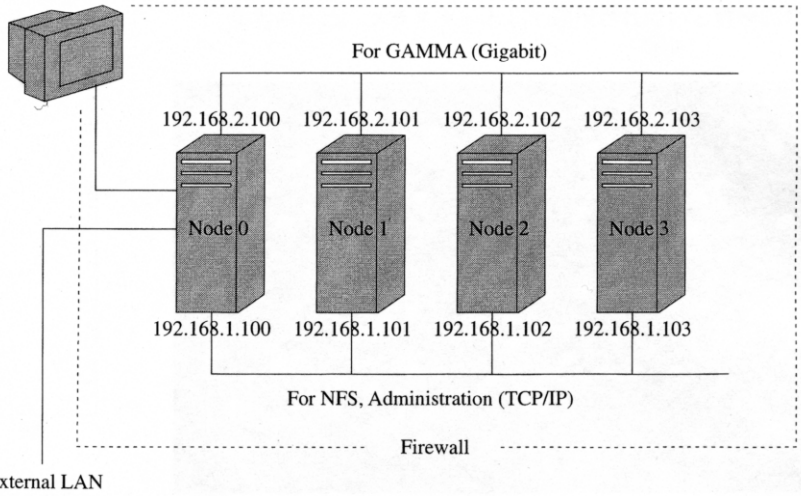


図1　高速通信システムGAMMAをPCクラスター計算機で利用するときに推奨されるネットワークの構成．1番目の系統はGAMMAによるデータ通信専用のギガビットネットワーク，2番目の系統はNFSマウントにより異なるPC上のファイル (実行バイナリを含む) 参照やタスクの管理に使用するTCP/IPネットワークである．

おり，実際に1日単位の長時間連続運用に耐えることを確認した．しかし，運用中にタスクを強制終了（異常終了）したときは，ノード間で状態の同期が失われた可能性があるため，gammaresetall で状態復帰させるのが賢明である．

ここまではGNU C/Fortranを使用したが，一般的には高速性の追求や応用プログラムの書式（たとえばFortran 90の利用）のため，市販のC/Fortranコンパイラを利用することが多い．そこでは，ユーザーの応用プログラムをコンパイルした後に生成されるオブジェクトとMPI/GAMMAライブラリの整合性を図ることが必須である．具体的には，GNU C/Fortranにあわせ一般のコンパイラでも生成されるオブジェクトに2個のアンダースコア(__)を添付，線形計算ライブラリとその並列計算への拡張版BLACS, SCALAPACK[13]を先の条件でコンパイル，引数参照を仲介するfarg.fをリンク，[14] 論理定数の定義の整合性を確認する，[15] などである．さいごの要件は，論理定数の真(.true.)に対応させる整数値が処理系により異なるためであり，発見が難しい思わぬ落とし穴である．

3. 高速通信による処理性能の向上

はじめに基礎データとして，通信速度が送信データ量とともに向上する様子を図2に示す．[16] これはGAMMAに添付のpingpongプログラムを用いての2つのプロセッサ間通信の測定であり，[転送処理能力]＝[データ量]/[所要時間] の値をデータ量に対して描いてある．使用した3Com996はギガビットイーサネットのNICである．送信データ量が小さいときは，通信の応答待ち時間が転送処理能力を小さく抑え，データ量1バイトでの0.6 Mbits/sは待ち時間15 μsに対応する．しかし，TCP/IP通信における大きな応答待ち時間100 μsが大きく減少するため，小さなデータ転送が頻繁に発生する現実の応用プログラムでは，GAMMAの通信性能が顕著に現れる（表1で再び言及）．送信データ量が大きくなるにつれて処理能力は向上し，10^5 バイト付近で飽和する．非常に大きいデータ量に対する極限帯域幅は706 Mbits/sであり，ギガビットネットワークの最大処理能力の70%に達している．

GAMMAのホームページ[10]に記載されている最高値は，ハードウェアのMyrinet (1.28 Gbits/s)＋BIP通信利用で応答待ち時間と極限帯域幅はそれぞれ4.3 μsと1,005 Mbits/sであり，ギガビットイーサネット利用のGAMMA＋Netgear GA621 NICでそれぞれ8.5 μsと976 Mbits/sである．多次元の巨大行列を並列計算で解く場合などではPC間通信が頻繁に発生するため，通信の応答待ち時間の短さは高速演算の鍵といえる．

次に，実際に物理研究の場で使われる応用プログラムである，密度汎関数法に基づく第一原理分子動力学コードSiesta[17]におけるGAMMAの実効性能を示そう．これは原子基底を用いた緊密結合 (tight binding) 型の量子力学的分子動力学コードであり，主要な計算は電子密度行列の対角化である．ここでは電気伝導度が大きく燃料電池に使われるイオン液体のイミダゾール（炭素，窒素，水素からなる5員環分子の液体，図3）[18] で，とくに20分子＋プロトンの場合のランタイムを表1に示す．注目すべきは，TCP/IP通信によるMPI[19]使用の場合（上段）と，通信のフロー制御をオンにしたMPI/GAMMA（中段）との比較で，通信に関わるオーバーヘッドが26秒から0.1秒に急減し，そのため，経過時間が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）．もしGAMMA利用時に通信の衝突を防ぐフロー制御をオフにすると，失われたデータの再送信が生じるため経過時間は大きく増加する．

参考のため，クラスターとして運用されている標準的な

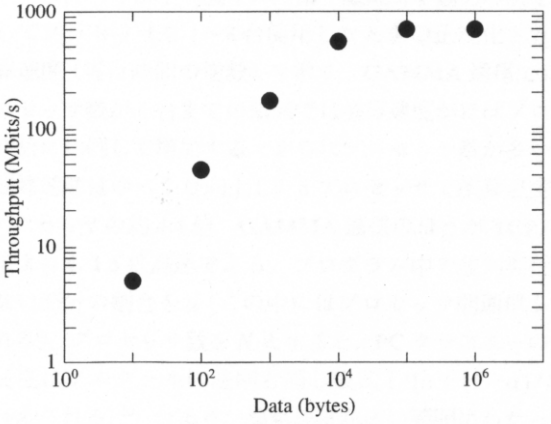


図2 GAMMA通信における，プロセッサ間で転送するデータのサイズ（単位：バイト）と転送速度 (Mbits/s) の関係．Pentium 4 (3.0 GHz) と3Com996 NIC（ギガビットイーサネット）を使用．この測定では，通信の待ち時間は15 μs，大きなデータに対する極限転送速度は706 Mbits/sであり，これは最大性能の約70%にあたる．

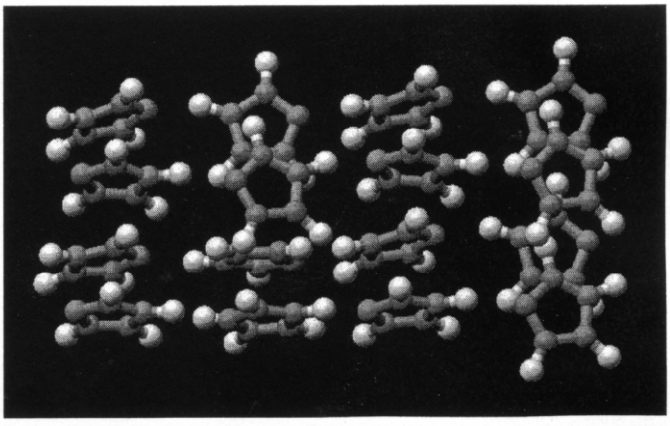


図3 密度汎関数・第一原理分子動力学による，イオン液体イミダゾールのシミュレーション．この分子は炭素3個，窒素2個，水素4個からなる5員環であり，プロトンが付加した分子（最下段，左から2番目）の周囲では隣り合う分子の配向が同一面内になる傾向がある．

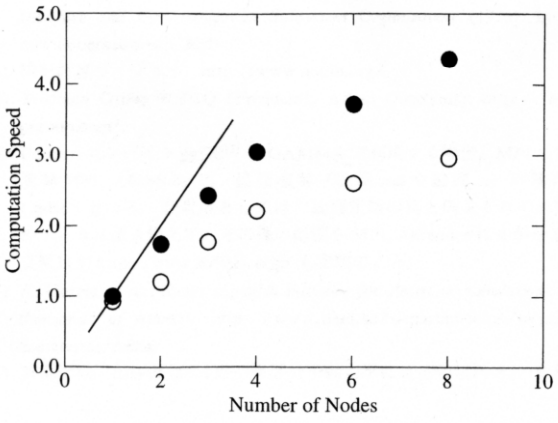


図4　GAMMA 通信 (●) および TCP/IP 通信 (○) 下での，密度汎関数・第一原理分子動力学コード Siesta の正規化された演算速度（プロセッサ数 1 の能力を 1 とした相対値）とプロセッサ数の関係．測定は Pentium 4 (3.0 GHz), 3Com996 NIC（ギガビットイーサネット），pgf90 v5.1 コンパイラを用いており，直線は正規化された演算速度がプロセッサ数に等しい理想的な場合を表す．演算速度は並列度とともに向上するが，その度合いは GAMMA 通信のほうが TCP/IP 通信よりも大きい．

RISC マシンである IBM Power 4（Itanium 2 の Altix3000 は少しだけ高速）を先と同じ条件下で同数台用いたときの測定結果を示す．表 1 は，RISC マシンである Power 4 (1.5 GHz) と Pentium 4 (3.0 GHz)＋MPI/GAMMA の実効性能がほぼ同じことを示している．これらで通信のオーバーヘッド時間が小さい点も共通である．

ところで，PC クラスター計算機はスーパーコンピュータを上回る演算性能を示すことがある．それは多くのプロセッサを容易に 1 ジョブに割り当てられること，また複雑な DO ループのためベクトル化が容易でない応用プログラムの場合である．後者にあたる第一原理分子動力学コードでは，単体でもスーパーコンピュータを凌ぐ演算性能を示す．

それでは使用するプロセッサ数に応じて，実計算速度は向上するだろうか？　図 4 に，第一原理分子動力学法を例としてプロセッサを 1〜8 台使用したときの正規化された演算速度（実行時間の逆数）を示す．GAMMA 通信では，プロセッサ数が 4 台までの領域では演算速度がほぼプロセッサ数に比例して増加する．さらにプロセッサ数が多いとき演算速度はゆっくり向上し，8 プロセッサで演算速度は 1 プロセッサの約 4.4 倍，GAMMA 通信のほうが TCP/IP 通信よりも 1.5 倍高速である．プログラム中で並列実行できない部分の割合を α（この中にはプロセッサ間通信も含まれる），プロセッサ数を N とすると，PC クラスターの演算速度はアムダールの法則と同じ表式 $1/[\alpha+(1-\alpha)/N]$ で与えられる．[20] つまり，演算速度が広い範囲にわたってプロセッサ数に比例するためには，非並列化部分が小さいこと $\alpha \ll 1$ が必要である．図 4 のデータは上式によく適合し，$\alpha_{\mathrm{GAMMA}} \sim 0.10$ (10%)，$\alpha_{\mathrm{TCP}} \sim 0.23$ (23%) である．この相違は，分割実行できない通信の応答待ち時間の差から生じている．[21]

4. まとめ

この記事では，PC クラスター計算機の演算性能向上のボトルネックであった遅いプロセッサ間通信が，リナックス標準実装の TCP/IP 通信に由来し，それが通信ソフトウェア GAMMA の導入によりイーサネット利用下で解消されることを示した．さらに，リナックス非標準の高速コンパイラをこの GAMMA システムと一緒に利用する方法を述べた．その結果，PC クラスター計算機が同数の並列 RISC マシンに近い演算性能をもち長時間安定に稼動すること，そして複雑な量子力場を計算する第一原理分子動力学に限れば，PC クラスター計算機の能力はベクトル型並列スーパーコンピュータを上回ることを確かめた．その半面，PC クラスター計算機の運用とメインテナンスはユーザー自身で行う必要があり，これは実験装置の運用とよく似ている．結論として，PC クラスター計算機はコストパフォーマンスに優れた計算機シミュレーションの道具であり，その可能性を読者が試し広げることで，これを用いての多くの豊かな理工学研究の成果が生み出されていくことだろう．

GAMMA ソフトウェアの PC クラスター計算機へのインストールと初期運用については Dr. Giuseppe Ciaccio から親切な助言とサポートをいただいた．第一原理分子動力学法の整備と PC クラスター計算機の製作は，善甫康成氏との共同研究により行った．両氏に心から感謝いたします．なお本研究の遂行では文部科学省科学研究費（特定領域 16032217 (2003-5)）の支援を受けた．

参考文献

1) Purdue 大学の Beowulf プロジェクト：http://www.psych.purdue.edu/~beowulf/
2) D. Becker たちの Beowulf プロジェクト (NASA): http://www.beowulf.org/
3) 青山学院大学理工学部の PC クラスタープロジェクト：http://www.phys.aoyama.ac.jp/~aoyama/
4) 同志社大学工学部の PC クラスタープロジェクト：http://www.is.doshisha.ac.jp/SMPP/
5) M. Snir, S. Otto, S. Huss-Lederman, D. Walker and J. Dongara: *MPI—The Complete Reference, Vol. 1 and 2* (MIT Press, Cambridge, 1998)
6) 善甫康成，田中基彦：『第一原理分子動力学コードの整備と応用』，核融合科学研究所 Annual Review (2001).
7) M. Tanaka: *PC Cluster Machine Equipped with High-Speed Communication Software* (Los Alamos Arxiv: physics 0407152 (2004)); 最新の情報は以下の URL を参照：http://dphysique.nifs.ac.jp/
8) 国立天文台 Grape: http://www.astrograpе.org/
9) Myrinet: http://www.myri.com/
10) G. Chiola and G. Ciaccio: *GAMMA Project: Genoa Active Message Machine*（Genoa 大学）：http://www.disi.unige.it/project/gamma/
11) SCore Consortium: http://www.pccluster.org/
12) A. Mainwaring and D. Culler: *Active Message Applications Programming*

Interface and Communication Subsystem Organization (1996): http://now.cs.berkeley.edu/AM/
13) 線形計算ライブラリ：http://www.netlib.org/.
14) Portland Group の FAQ (Frequently Asked Questions): http://www.pgroup.com/.
15) Fortran コンパイラ pgf90[12] を GAMMA で利用する場合，MPI の総和演算 MPI_Allreduce で，論理演算（論理 and や論理 or）が常に偽 (.false.) となる．回避策としては，論理定数の真と偽をそれぞれ整数の 1 と 0 に置き換えて，その後の処理を MPI_Allreduce にまかせる．
16) 文献11や http://www.softek.co.jp/ も参照のこと．
17) A. Garcia, *et al.*: *Siesta* (*Spanish initiative for electronic simulations with thousands of atoms*): http://www.uam.es/departamentos/ciencias/fismateriac/siesta/
18) 善甫康成，田中基彦：『第一原理分子動力学による物質科学』，核融合科学研究所ニュース No. 2/3 (2004).
19) MPICH (MPI Chameleon)—1 & 2, Argonne National Laboratory: http://www.mcs.anl.gov/
20) プロセッサ数が非常に多い場合，通信相手先が増えるため α はもはや定数とはみなせず，演算性能が最大となるプロセッサ数が出現する．これは応用プログラムのデータサイズとハードウェア（ハブ等）の能力に依存する．
21) 巨大（数10万）次元の線形連立方程式で表されるポアソン方程式の並列解法（たとえば共約勾配法）において，演算性能のプロセッサ数依存性は図 4 と同様となる．このとき Power 4 (1.5 GHz) は Pentium 4 (3.0 GHz) より約 1.5 倍高速である．

（2004 年 7 月 5 日原稿受付）