PCクラスタの性能評価と今後の計算機シミュレーション

北沢, 充弘
九州大学応用力学研究所

矢木, 雅敏
九州大学応用力学研究所

伊藤, 早苗
九州大学応用力学研究所

https://doi.org/10.15017/4784012

出版情報：九州大学情報基盤センター年報. 3, pp.15-20, 2003-03. 九州大学情報基盤センター
バージョン：
権利関係：

PCクラスタの性能評価と今後の計算機シミュレーション
Performance and Efficiency of PC Cluster for Computer Simulation

北沢充弘　Atsuhiro Kitazawa
矢木雅敏　Masatoshi Yagi
伊藤早苗　Sanae-I. Ito

九州大学応用力学研究所
Research Institute for Applied Mechanics, Kyushu University

要旨　自然現象を解明する手段として、数値シミュレーションは大きな存在となっている。近年のPCハードウェアの性能向上とネットワーク技術の発展により、PCクラスタを用いた自然科学数値シミュレーションが注目されている。ここでは、我々のシミュレーションや利用計算機環境の紹介を行ない、現状でのPCクラスタを用いた計算機シミュレーションの有効性を考えた。

Abstract　In order to investigate a natural phenomenon, the numerical simulation is an efficient method. Due to the rapid improvement of PC hardware and the progress of network technology, numerical simulations with PC clusters are widely noticed in the area of the natural sciences. We introduced our simulations and computer environments, and considered the effectiveness of the numerical simulation with PC clusters.

1　はじめに

自然現象を解明する手段として、数値シミュレーションは大きな存在となっている。我々の研究グループでは、核融合プラズマに代表される高温プラズマの物理について研究しているが、プラズマにおける乱流や自己組織化・構造形成などの非線形な性質についての研究では、現象を再現するだけでなく、その背後に隠れた性質を解き明かすのに数値シミュレーションが重要な役割をはたしている。

近年、Pentiumプロセッサーに代表されるような安価なチップが市場にでまわるようになり、３～４年前ならスパーコンピュータでしかできなかったシミュレーションが自前のPCで行えるようになった。またチップ性能も毎年１．５～２倍程度向上しており、高速な演算性能となっている。

PCによる数値計算では、PCクラスタはコストパフォーマンスの高い並列計算機環境である。PCクラスタによる並列化プログラムの実行では、処理を各ノードに分散させることにより計算時間が短縮され、高い演算性能が期待される。また、分散メモリー型のシステムの為、各マシンにあるメモリは分散される処理に必要な量だけが要求され、１台の計算機ではハード的な制約で行なえない大規模な数値計算も可能となる。ノード間の通信では、MPIやPVMなどの並列通信ライブラリがTCP/IPにより行なわれており、最近普及してきた1000Base-Tによる高速なEthernetにより、通信によるオーバーヘッドは小さくなっている。

ここでは、我々の研究室でおこなっているシミュレーションや利用計算機環境を紹介し、PCクラスタでの計算機シミュレーションの有効性を考えたい。

2　計算機利用環境

我々の研究室ではプロジェクト型研究として高温プラズマにおける異常輸送のシミュレーション研究を行っている。使用している計算機は核融合科学研究所（以下NIFSと呼ぶ）のSX4/64M2、SX5/6B、応用力学研究所（以下RIAMと呼ぶ）のVPP5000/2、ES40(3node)、九州大学情報基盤センター（以下センターと呼ぶ）のVPP5000/64等である。NIFSの計算機は課金がかからないという大きなメリットがあるが、巨大なデータをやりとりするには今のSINETでは負荷が大きい。またSX4/64M2に対して長時間ジョブを投入するためにはNIFSのゲートウエイ内部のマシンから投入する必要があり遠隔地から使用するには

不便である。一方、課金はかかるがセンターやRIAMのVPP5000はより使いやすい環境にある。しかし、プロジェクトジョブを大量に流すことを考えるとディスク容量が十分ではない。我々の研究室では、ここ数年、閑散期にはセンターのCPU定額利用制度を利用し（１０万円で５０万円分利用できる制度）、年度末はRIAMやNIFSの計算機を利用していた。それに並行して昨年よりXeonクラスターを研究室に導入し運用を始めたが、クラスターの性能が見えてくるにつれてスーパーコンピュータのコストパフォーマンスを再評価する必要があるのではと感じている。

3 Xeonクラスターの性能評価

平成１３年度よりLHD計画共同利用研究（研究代表者　福山淳（京都大学））及び応用力学研究所全国共同利用研究の支援を受けて核融合理論グループによるグリッドコンピューティングプロジェクトを開始した。これはNIFS、RIAM、京都大学、山口大学に異なるシステムの並列型計算クラスタ(Xeon, Alpha, PowerPC G4, etc)を設置し、Super SINETでの接続から計算機資源の有効利用を行うとともに、将来的にはLHDの実験データを遠隔からリアルタイムで解析しようとするものである。例えばデータの解析はNIFSのクラスターで行い、画像処理はRIAMのクラスタで行うような複合型クラスターシステムを構築していく予定である。グラッフィクライブラリー（http://p-grp.nucleng.kyoto-u.ac.jp/ fukuyama/gsaf/index.html）の並列化及びOpenGL化は京都大学で現在進行中である。RIAMにはXeon 1.7GHz（4ノード、GbE接続）を導入し、並列型計算機クラスタの有効性を検証するためのベンチマークテストを行った。

我々のクラスタシステムの環境を表1にまとめた。(Pentium 4はNISおよびNFSのサーバとなっており、これらのサービスは100Base-TXのEthernetで行なわれている。) 通信性能向上のためSCore 5.0.1 [1]を導入した。SCoreはワークステーションおよびPCクラスタ用の高性能並列プログラミング環境であり、クラスタコンピューティング専用通信ライブラリ、プログラミング環境、ジョブスケジューリング処理等が提供されている。SCore用のMPI (MPICH-SCore)はMPICH-1.2.0に低レベルメッセージパッシングシステムとしてPMを使っている。

CPU	Pentuium 4 1.7GHz ×1 Xeon 1.7GHz (dual CPU) ×4
メモリ	1024Mbyte
SCore	5.0.1
ベースクロック	400 MHz
NIC	Intel PRO/1000 MT Intel EtherExpressPro 100
スイッチ	NETGEAR GS516T
OS	RedHat Linux 7.3 (kernel 2.4.18)
MPICH	1.2.4

表 1: 評価環境

3.1 通信性能の評価

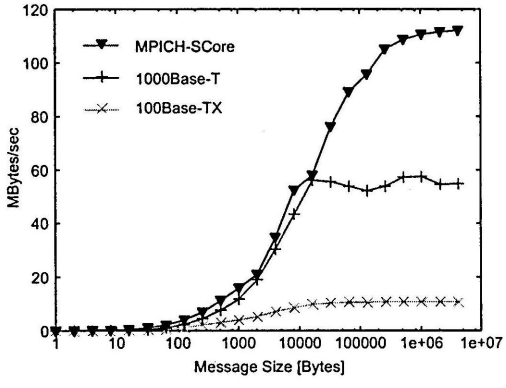

図 1: ピンポン転送によるバンド幅の測定

PCクラスタの通信性能を評価するために、2ノード間のピンポン転送によるスループットの計測結果を図1に示す。ピンポン転送では、送信ノードから受信ノードにメッセージを送信し、そのメッセージを受信ノードが受信後、同じサイズのメッセージを送信ノードに送信し、そのラウンドトリップ時間を計測する転送方法を用いている。この測定では通信ライブラリとしてMPICHおよびMPICH-SCoreを用い、メッセージサイズを1バイトから4Mバイトまで変えながら行ない、経過時間からスループットを求めた。ここでは、NICとしてGigabit Ethernet (1000Base-T, Intel MT/1000 MT)とFast Ethernet (100Base-TX, Intel EtherExpressPro 100)を用いた結果を示す。(Gigabit Ethernetは PCI-64 (64 bit/66 MHz)に接続している。)

100Base-TXの場合は、スループットは64 KB転送

Xeon	P II	ES40	VPP
1.7GHz	400MHz	667MHz	5000
PGI	Fujitsu	Compaq	VPP
Fortran	Fortran	Fortran	Fortran
1000Base	100Base	Crossbar	Crossbar
-T	-TX	SW	SW

表 2: ベンチマークに使用したマシン、Fortran コンパイラー及び NIC（ネットワーク接続形態)

あたりで飽和し約 10MB/sec となった。この時のレイテンシは、6ms であった。1000Base-T の場合は、スループットは 16 KB 転送あたりで飽和し、55 MB/sec (レイテンシは 280μs) となり、100Base-TX の場合の数倍の性能となった。しかし、1000Base-T の理論性能は 125MB/s であり、TCP/IP による通信では、十分なスループットを享受できない事がわかった。SCore を用いた 1000Base-T での測定では、データサイズが 4MB で 110 MB/sec となり大きなデータサイズでの性能向上が見られた。MPICH-SCore の結果で、データサイズが 8 KB での段差はデータ転送処理をメッセージ転送からリモートメモリライトによるデータ転送に切替えているためである。

3.2 姫野ベンチによる評価

姫野ベンチとは、理化学研究所の姫野龍太郎氏が開発した非圧縮性流体のソルバーによる、ベンチマークテストであり、ポアッソン方程式をヤコビの反復法で解く場合の主要なループの処理速度を計るものである [2]。サイズは medium 256x128x128 で行った。表 1 にベンチマークに使用したマシン、Fortran コンパイラ及び NIC（ネットワーク接続形態）を示し、表 2 にそのベンチマーク結果を示す（単位は FLOPS)。Xeon 1.7GHz は、4 年程前に導入した Pentium II 400MHz の 5〜6 倍程の性能を出している。ノード数が増加すると通信負荷も高くなるためグラフの傾きは次第に小さくなっている。8CPU での 1CPU あたりの性能比は、1CPU での値を基準とすると Xeon で 0.55、Pentium II で 0.65 となり、Pentium II の方が並列処理の効率は良い結果となった。

3.3 NAS パラレルベンチマーク

NAS パラレルベンチマークは、NASA Ames Research Center で開発された、並列コンピュータのためのベンチマークであり、5 つのカーネルベンチマーク (EP, MG, CG, FT, IS) と 3 つの流体シミュレーション (LU, SP, BT) からなる。

CPU	Xeon	P II	ES40	VPP
/PE	1.7GHz	400MHz	667MHz	5000
1	593M	95M	201M	4173M
2	1112M	159M	334M	8235M
4	1936M	282M	676M	15668M
8	2608M	494M	-	22157M

表 3: 姫野ベンチマークの結果（単位 FLOPS)

MPICH 1.2.4 および MPICH-SCore を用いた NAS パラレルベンチマーク (NPB 2.3)[3] の結果が図 2 である。ここでは、CG、MG、LU のベンチマークの結果のみを示す。

CG:	正値対称大規模疎行列の最小固有値の共役勾配法での近似計算。
MG:	簡略化されたマルチグリッド法で三次元ポアソン方程式を解く。
LU:	5×5 ブロックの上下三角行列システムを SSOR 法で解く。

NIC は 1000Base-T を用い、コンパイラは PGI Fortran 3.2 を使った。ベンチマークのクラスはサイズの小さい W と大きい B である。CG ではノード間通信の影響が大きく、サイズが小さいクラス W では、4CPU で性能が飽和しているが、サイズの大きいクラス B では、8CPU クラスタでも性能の向上が見られる。MPICH-1.2.4 と MPICH-SCore を比べると、クラス W では通信性能の向上により、MPICH-SCore の方が性能低下が小さい事がわかる。LU と MG では通信よりも演算処理時間が長いため、クラス B ではノード数を増加させても、MPICH-1.2.4 と MPICH-SCore で大きな差は生じなかった。よりノード数の大きな PC クラスタでは、MPICH-SCore の方が性能低下は小さいと思われる。

3.4 大きなクラスタ

現状でどの程度の CPU 数が PC クラスタシステムにとって有効であるかを見るため、京都大学院工学研究科原子核工学専攻 福山研のグループの PC クラスタの結果を示す (図 3 および図 4)。評価環境は、Xeon (2.2GHz, dual CPU) 16 台を 1000Base-T で接続し、

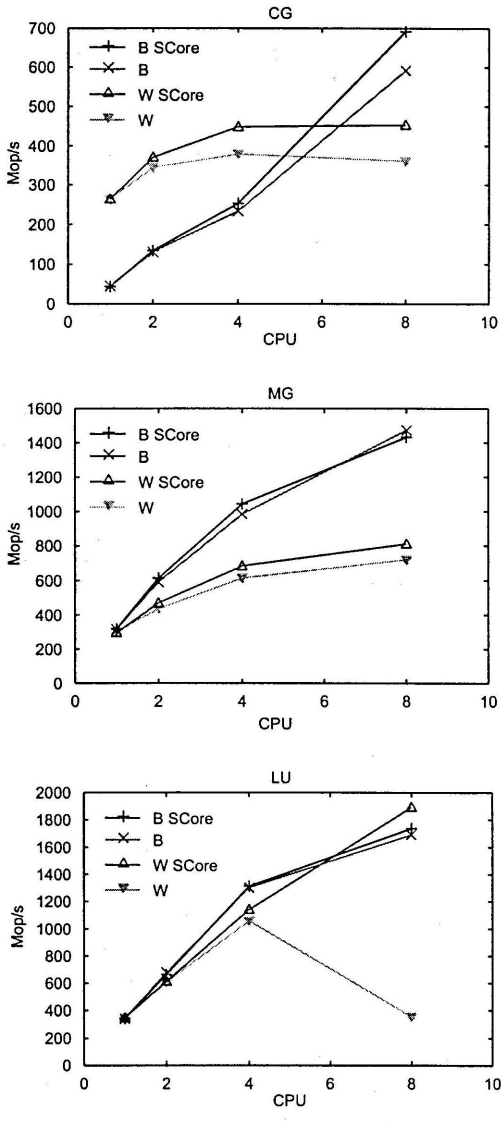

図 2: NAS パラレルベンチマーク。サイズは、CG: W(7000), B(75000), MG: W(64×64×64), B(256×256×256), LU: W(33×33×33), B(102×102×102) である。

OS は、RedHat Linux 7.3 である。通信ライブラリとして MPICH-1.2.4 を用いて、コンパイラは PGI Fortran 3.2 を使った。

CPU 数の増加とともに、1 CPU あたりの性能は低下する傾向にある事がわかる。これは、通信のレイテンシの増加が主な理由であり、最近の高速な CPU では、通信性能の向上が必要であると思われる。

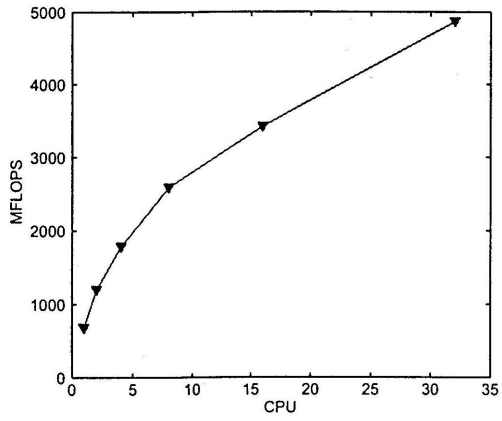

図 3: 姫野ベンチマーク (サイズ M)

3.5 LU 分解を用いたスペクトルコードの評価

3.2 節での計測から、Xeon 1.7GHz 8CPU は VPP5000 1PE の 60% の性能を出していることがわかる。8 CPU/PE で比較すると VPP5000 は Xeon の１０倍近い性能を出している。この結果だけを見ると VPP5000 の性能はかなり高いと結論されるが一般のコードではどうであろうか?次に VPP5000 で用いていたコードを Xeon へ移植しベンチマークテストを行った結果を紹介する。

異常輸送のシミュレーション研究に用いている流体コードは空間に関しては有限差分とスペクトル法を混在させており、時間に関しては線形部分を LU 分解を用いて陰解法で解き、移流項は予測修正子法で進めるアルゴリズムを用いている。今回のテストでは、DO LOOP を並列化し MPI_ALLREDUCE でデータを寄せ集める非常に初歩的な手法を用いた。表 4 に結果を示す。配列の大きさは 128x512 (約 500MB) に選んだ。1CPU/PE の比較から明らかなようにこのコードのベクトル化効率は悪い。LU 分解のドライバーを Lapack.vp に変えて試してみたがスピードがわずかに遅くなった。CPU 数を増加させた場合、MPICH に比べて MPICH-SCore では性能向上が見られた。

この配列サイズに対しては Xeon では 4CPU で使用するのがもっとも効率がよさそうである。同じ 4CPU/PE で VPP5000 と比較すると、MPICH-1.2.4 を用いた場合 57% の、MPICH-SCore を用いた場合

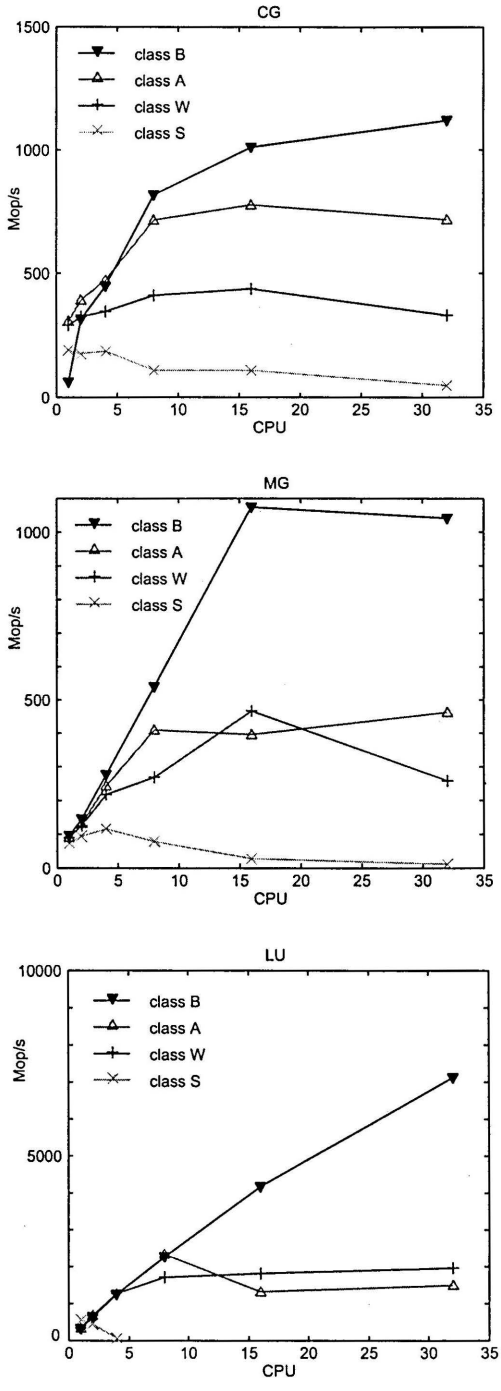

図 4: NAS Parallel Benchmark

で 64% の性能がでていることがわかる。

CPU /PE	MPICH -1.2.4	MPICH -SCore	VPP 5000
1	179.34 sec	178.01 sec	134.67 sec
2	100.08 sec	96.25 sec	69.22 sec
4	63.44 sec	57.33 sec	36.74 sec
8	51.895 sec	45.16 sec	20.77 sec

表 4: スペクトルコードのベンチマーク結果

扱ったコードの大きさは 500MB 程度であり 1CPU の場合に比べて 4CPU で約 2.8 倍程度速度が向上している。しかしコストパフォーマンスからは 4 CPU に対して並列化していないコードを 4 本独立に走らせた方がいいのは明らかである。並列化のメリットは各 CPU に搭載されているメモリーを有効に使用することが可能となる点である。マスターとスレーブで各 CPU に搭載されたメモリーを最大限使用するようなプログラムを書けばクラスターを有効利用することが可能となろう。

4 並列化のメリットと計算機のコストパフォーマンス

最後に PC クラスタによる並列化のメリット・デメリットと計算機利用のコストパフォーマンスについて検討してみたい。

最近の PC を用いた小規模な PC クラスタでも、並列処理により計算時間が短縮され、スーパーコンピュータ 1 台相当の性能となる事がわかった。また、スカラ実行によりベクトル化が難しい処理などでは、スーパーコンピュータよりも PC クラスターを用いた方が良い性能を示すと思われる。

PC クラスタの利点は以下のものが上げられる。まず、ハードやソフトの面で安価にシステムを構築でき、独自に拡張が可能である事が上げられる。京都大学の PC クラスタ (Xeon 2.2GHz × 16、Gigabit Ethernet on board) では、予算 500 万円で VPP5000 1 台相当の実高性能を達成している。OS に Linux 等のフリーな PC Unix を用いるので、OS やコンパイラ、数値計算ライブラリ等のソフトの更新が簡単である。ハードの面でも汎用の部品を使用しているためコストパフォーマンスが高い。また、計算に大容量のメモリが必要で、1 台の PC では不可能な場合でもメモリの有効利用が可能となる。

しかし、PC クラスタを用いる場合、一般ユーザーにとっては閾の高い並列用のプログラムを作成する必要があり、プログラマーの負担は大きくなる。これまで作成してきたプログラムを並列計算用に作りなおす場合、局所的な変更ではすまなく、有効に CPU を利用する戦略が必要となる。また、ベンチマークの結果から、現在の PC クラスタでは通信のレイテンシが問題となる。これは、規模を大きくするほど顕著になり、より高速な通信手法が必要となる。しかし、Myrinet などの高速通信媒体の利用はコストパフォーマンスとの兼ね合いとなる。運用面では、場所をとり、大規模にすると管理が大変となり、電気代などの運用経費も上昇する点が上げられる。

並列計算機とベクトル計算機の利用ではスカラ処理にするかベクトル処理にするかで使い分ける必要があり、非並列コードを大容量のメモリサイズ走らせたいユーザーはベクトルプロセッサを、コード開発に労力を惜しまず並列化コードを巨大メモリーサイズで走らせたいユーザーは並列計算機を使うことになる。現在のところ、研究室単位で PC クラスタを導入し、並列計算機シミュレーションを行なう場合、16〜32 CPU の小規模なシステムが計算能力やコストの面で有効なのではないだろうか。

5 謝辞

この原稿を書くにあたり有益なコメントをいただきました京都大学の福山淳教授、核融合科学研究所の伊藤公孝教授に感謝します。この研究の一部は、LHD 計画共同研究および応用力学研究所共同利用研究の支援によるものです。

参考文献

[1] http://www.pccluster.org/

[2] http://w3cic.riken.go.jp/HPC/HimenoBMT/

[3] http://www.pccluster.org/