ハイパフォーマンス
コンピューティング 67－8
（1997. 8. 19）

量子色動力学プログラムを用いた並列計算機の性能評価

益口摩紀† 長嶋雲兵†
関口智嗣†† 佐藤三久†††

本研究では量子色動力学をシミュレートするQCDMPIプログラムを用いて，Alpha Stationを用いたワークステーションクラスタetlwizとNECのCenju-3，富士通AP1000+,AP1000の分散メモリ型並列計算機システムの予備的な性能評価を行なった．QCDMPIプログラムの実行に関しては，etlwizはPE数が大きくなるにつれ経過時間の80%が通信に費やされる．逆にCenju-3は90%以上を計算時間が占める．AP1000+，AP1000は30～50%を通信時間が占めることがわかった．

Preliminary Performance Evaluation of Parallel Computer Systems using QCDMPI

MAKI MASUGUCHI,† UMPEI NAGASHIMA,† SATOSHI SEKIGUCHI†† and MITSUHISA SATO†††

Using QCDMPI, a quantum chromo dynamics simulation program, a preliminarily performance evaluation has been carried out for four distributed memory parallel computing systems: a dedicated workstation cluster with Alpha station(etlwiz), NEC Cenju-3, Fujitsu AP1000 and AP1000+. More than 80% of elapsed time is spent for data communication in the 32 processor elements case of etlwiz. Data communication time of Cenju-3 is less than 10% of elapsed time. Data communication of AP1000 and AP1000+ is about 50 ~ 30% of elapsed time.

1. はじめに

ユーザの大規模かつ高速演算に対する要求と科学技術の急速な発展によって並列・ベクトル型計算機をはじめとしたコンピュータシステムの多様化が進んでいる．こうしたコンピュータシステムの特徴と性質を明らかにするための性能評価技術はますます重要となってきている．このような研究は国内では盛んに行なわれている．本研究ではその研究の一端として日置によるポータビリティの高いQCDMPI[1]を様々な並列計算機に移植し，これをベンチマークプログラムとして用いて並列計算機システムの性能評価を行なった．

2. QCDMPI

2.1 QCDとQCDMPIについて

原子核を構成する陽子や中性子はさらにクォークという基本粒子から合成される複合粒子だと考えられている．このクォークの色電荷の間に働くクォーク間力に関する量子理論がQCD(Quantum Chromo Dynamics)である[2),3)]．QCDでは格子ゲージ理論を元に，有限化された格子空間内における格子点と隣接する2つの格子点を結ぶリンクに対してファインマンの経路積分を適用する．QCDのような複雑で大規模な問題に対し解析的手法を適用することは不可能である．モンテカルロ法を用いた数値計算によってゲージ場の更新を繰り返すことにより，陽子や中性子の質量等の物理量に対する近似解が得られる．

本研究で用いたQCDMPIは，1995年にSan Diegoで行なわれたSuper Computing95でGordon Bell Prizeを獲得した日置によって作成されたプログラムであり，Fortranでの記述，MPI[4]を用いたメッセージ通信，SPMDモデルに基づいている．特徴はプログラム上で演算と通信を全く分離することにより，プロセッサ数や格子の分割方法に依存しないポータブルなコードになっている．このため，計算機の演算性能・通信性能をそれぞれ独立に評価するのに適している．計算に必要なCPU時間は格子空間のサイズに比例するが，並列処理効果により計算の高速化が実現できる．

2.2 QCDMPIのアルゴリズム

図1に示すように指定された4次元の格子空

† お茶の水女子大学
Ochanomizu University
†† 電子技術総合研究所
Electrotechnical Laboratory
††† 新情報処理開発機構
Real World Computing Partnership

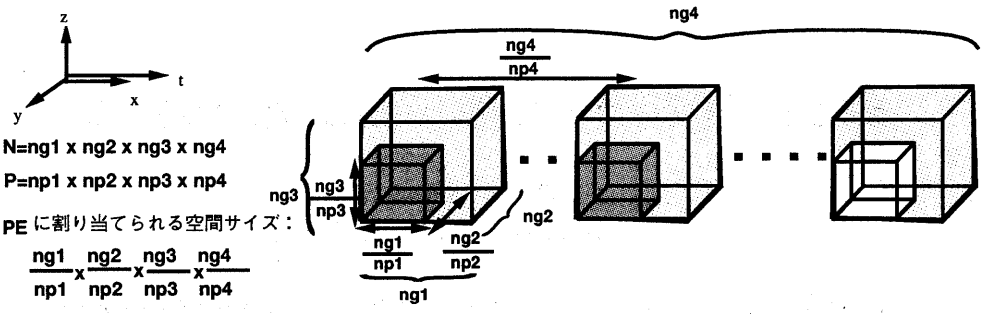

図 1　空間の分割

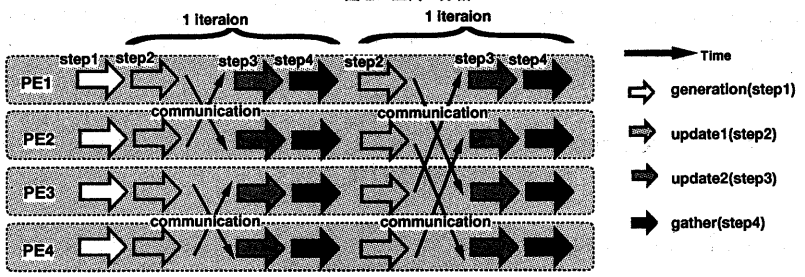

図 2　並列実行フローチャート

間 $N = ng1 \times ng2 \times ng3 \times ng4$ をプロセッサ数 $P = np1 \times np2 \times np3 \times np4$ で分割する．各プロセッサは一斉にそれぞれ割り当てられた空間内の処理を行なう．各プロセッサに割り当てられる空間サイズは，$\frac{ng1}{np1} \times \frac{ng2}{np2} \times \frac{ng3}{np3} \times \frac{ng4}{np4}$ となり，計算するリンクの数は空間サイズに次元数 4 を掛けたものとなる．

図 2 に並列実行の流れを示す．図中の step1 ～ step4 は，以下のとおりである．

[step1： **generation**] 各プロセッサは与えられた格子空間内の全ての格子点とリンクに対して重み因子によりゲージ場の初期値を設定する．

[step2： **update1**] 全てのリンクに対して，それぞれのリンク上のゲージ場に対する作用を各次元毎に計算する．具体的には行列同士の積または和を解く．並列実行の場合，割り当てられた空間の隣接するプロセッサ間で境界情報を交換するための通信を行なう．

[step3： **update2**] 熱浴法に従って，乱数から生成したゲージ場における変数の値をある確率で更新しながら，ユークリッド化された時空間の下で経路積分を行ない各経路に対する作用を計算する．

[step4： **gather**] それらの結果を集め，総和を取ることにより連続的な時空間における物理量 (plaquette energy) の近似値を得る．ここでゲージ場の更新回数 *iteration* は事前に与える．

step2 において，1 リンクずつ実行したのでは効率が悪いため，独立なリンクは全て同時に処理する．1 つの次元を分割する場合，1*iteration* 当たり 12 回の通信を必要とする．低次元分割の場合，通信の頻度は低いのに対し，1 通信当たりの転送量は大きくなる．一方，高次元分割により転送量を軽減できるが，通信回数は増加する．4 次元分割を行なう場合に最大値を取り，その時の通信回数は 12 回 × 4 *dim* となる．計算機の特性に合った次元分割を選択することにより，経過時間の短縮化が望める．MPP においては，4 次元分割の場合に通信性能を最大限に引き出せることが経験上わかっている．本研究では，対象とした全ての並列計算機に対し，4 次元分割に固定した計測を行なった．問題サイズに対して要素プロセッサ (PE) 数を 2 から 32 まで変化させた時の 1PE における通信回数と一回の通信当たりの通信量を表 1 に示す．

3. 並列計算機の性能評価

表 2 に計測に用いた並列計算機システムの仕様を示す．

実行時間の計測は，通信ライブラリの組み込み関数である MPI_Wtime() を用い，問題サイズは，$N = 1024$～32768 とした．QCDMPI の標準的なプログラムにおいて 1 リンクの計算に必要な浮動小数点演算数は 5700 である．単位時間当たりの浮動小数点演算数を *QCDFlops* とする．ただし，空間サイズによって与

表1　1PE での 1 iteration 当たりの通信回数と 1 通信当たりの転送量[KByte](1dim : 2dim : 3dim : 4dim)

N ($ng1 \times ng2 \times ng3 \times ng4$)	2PE[通信 12 回] ($2\times1\times1\times1$)	4PE[通信 24 回] ($2\times2\times1\times1$)	8PE[通信 36 回] ($2\times2\times2\times1$)	16PE[通信 48 回] ($2\times2\times2\times2$)	32PE[通信 48 回] ($4\times2\times2\times2$)
4096 (8× 8× 8× 8)	18 : 0 : 0 : 0	9 : 9 : 0 : 0	4 : 4 : 4 : 0	2 : 2 : 2 : 2	2 : 1 : 1 : 1
8192 (16× 8× 8× 8)	18 : 0 : 0 : 0	9 : 18 : 0 : 0	4 : 9 : 9 : 0	2 : 4 : 4 : 4	2 : 2 : 2 : 2
16384(16×16× 8× 8)	36 : 0 : 0 : 0	18 : 18 : 0 : 0	9 : 9 : 18 : 0	4 : 4 : 9 : 9	4 : 2 : 4 : 4
20480(16×16× 8×10)	46 : 0 : 0 : 0	23 : 23 : 0 : 0	11 : 11 : 23 : 0	5 : 5 : 11 : 9	5 : 2 : 5 : 4
24576(16×16×12× 8)	55 : 0 : 0 : 0	27 : 27 : 0 : 0	13 : 13 : 18 : 0	6 : 6 : 9 : 13	6 : 3 : 4 : 6
27648(16×12×12×12)	62 : 0 : 0 : 0	31 : 41 : 0 : 0	15 : 20 : 20 : 0	7 : 10 : 10 : 10	7 : 5 : 5 : 5
32768(16×16×16× 8)	73 : 0 : 0 : 0	36 : 36 : 0 : 0	18 : 18 : 18 : 0	9 : 9 : 9 : 18	9 : 4 : 4 : 9

表2　各並列計算機システムの仕様

マシン	プロセッサ (PE 数)	クロック周波数	メモリ	キャッシュ	ネットワーク	バンド幅
etlwiz	Alpha21164(32PE)	333 MHz	128MB	8KB,96KB,2MB	100Base/switch	100Mbps
Cenju-3	VR4400(64PE)	75 MHz	64MB	32KB,1MB	多段接続網	40MB/sec
AP1000+	Super SPARC(64PE)	50 MHz	64MB	128KB	2 次元トーラス	25MB/sec
AP1000	SPARC(64PE)	25 MHz	16MB	128KB	2 次元トーラス	25MB/sec

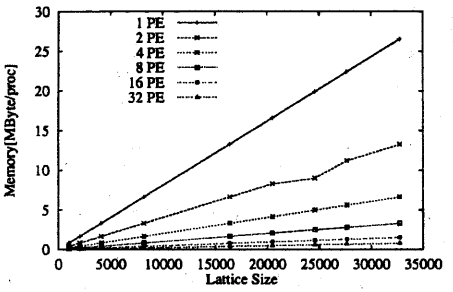

図3　演算に必要なメモリ量

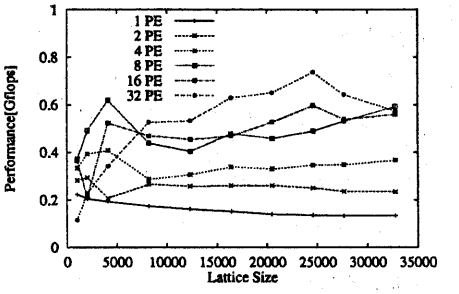

図4　etlwiz における性能 (平均値)

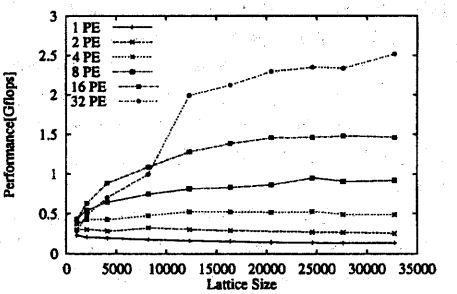

図5　etlwiz における性能 (最高値)

えられるリンク数によって,

$QCDFlops = 5700 \times$ (リンク数)/(実行時間)

としている．このようにして得られた Giga$QCDFlops$ を以下では Gflops として評価指標に用いる．

また，それぞれの PE 数と問題サイズのセットに対して必要な 1PE 当たりのメモリ量は図 3 に示すように空間サイズに対して $O(N)$ で増加する．PE 数が大きくなるにつれて 1PE 当たりに必要なメモリ量は減少する．

3.1 etlwiz

etlwiz は Alpha Station 500 を Ether switch で結合したワークステーションクラスタである．

図 4 に etlwiz における PE 数毎の問題サイズに対する性能の変化を示す．性能は 10 回の *iteration* の平均値により求めた．最高性能値は 32PE，問題サイズ 24576 の時の 0.73Gflops となった．4PE までは問題サイズによらずほぼ安定した性能値を保つのに対し，8PE を越えると $N \leq 8192$ の値が乱れている．これは極めて短い経過時間に対する測定のバラつきが影響している．16, 32PE で $N \geq 24576$ のときの性能は低下しており，8 ~ 32PE における $N = 32768$ のときの性能値はほぼ同じ値となっている．通信の負荷により，並列度を高くしても平均性能値は向上しない結果となった．

一方，10 回の *iteration* の最短時間により求めた実効性能を図 5 に示す．etlwiz では各 *iteration* における通信時間が不安定なために．実行性能の最高値と平均値に大差が見られた．最高性能値では，32PE で $N \geq 16384$ のとき 2Gflops 越える性能が得られている．$N \leq 8192$ の場合 32PE の性能が 16PE の性能より劣るのは，問題サイズが小さくなり過ぎて通信のオーバーヘッドが大きく実行時間に影響するためである．32PE，$N = 32768$ の時に最高値 2.5Gflops となった．

etlwiz の特徴として，333MHz という高速なク

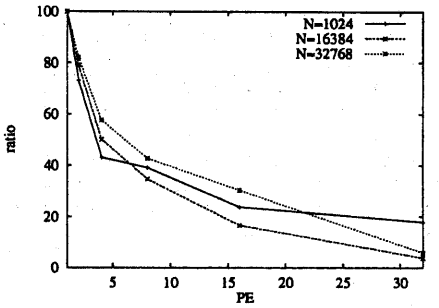

図 6 etlwiz における通信時間と演算時間の割合

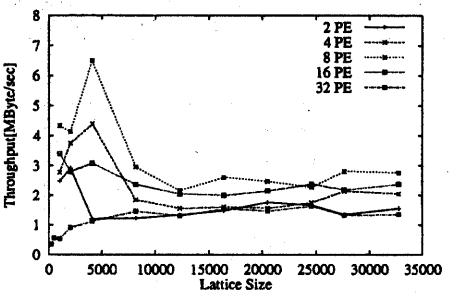

図 7 etlwiz における通信スループット

ロック周波数を持つ CPU を保持する一方で，PE が 100BaseTX と Ether switch で結合されているため，MPP と比較すると通信が遅いことが挙げられる．図 6 は実行時間に占める演算時間と通信時間の比率を示しており，グラフの下側が演算時間，上側が通信時間となっている．この図から etlwiz では，PE 数が増加するに従って全実行時間に占める通信の比重が高くなり，PE 数を増加させた時の性能は通信性能に大きく依存することがわかる．

図 7 に各 PE 数毎の空間サイズに対する通信スループットを示す．$N \geq 10000$ のとき PE 数毎に安定した値をとり，etlwiz では 8PE の場合に最も通信スループットが高い．一方 $N \leq 10000$ のとき値が安定していないのは，実際の通信時間が極めて短いために測定のバラつきが大きくスループットに影響するためである．

3.2 Cenju-3

図 8 に Cenju-3 の各 PE 数毎の問題サイズに対する性能の変化を示す．性能は 10 回の *iteration* の平均値より求めた．最高性能値は $N = 20480$，64PE の時の 1.35Gflops である．Cenju-3 は CPU 性能に比べて，高速な通信が実現されているので，粒度の大きな計算に対してはスケーラブルであることがわかる．Cenju-3 はその優れた通信性能により etlwiz に比べて通信時間も安定している．10 回の *iteration* における経過時間の最短時間と平均値に大差はなく，最短時間による最高性能も

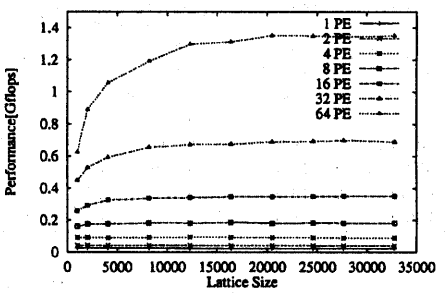

図 8 Cenju-3 における性能 (平均値)

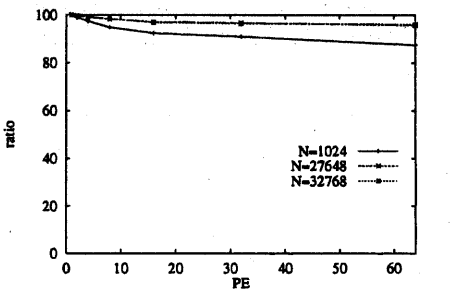

図 9 Cenju-3 における通信時間と演算時間の割合

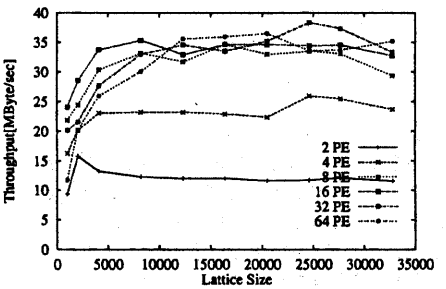

図 10 Cenju-3 における通信スループット

$N = 20480$，64PE の時の 1.35Gflops となった．

図 9 に示したグラフは演算時間と通信時間の比率を表しており，etlwiz における図 6 に対応する．Cenju-3 ではプロセッサ間通信を統制する専用インターフェースハードウェアの開発により高速なプロセッサ間通信が実現されているため，図 9 に示されているように Cenju-3 の通信は QCDMPI の実行にほとんど影響を及ぼさない．

図 10 に各 PE 数に対する通信スループットを示す．2PE，4PE の場合のような通信量が多く通信回数が少ない時はスループットはあまり伸びないが，8PE 以上の場合のような細かい通信を繰り返し行なう時には高いスループット値となることがわかる．最高値は 32PE を用いた時の 38MB/sec 程度で，実効速度 30MB/sec よ

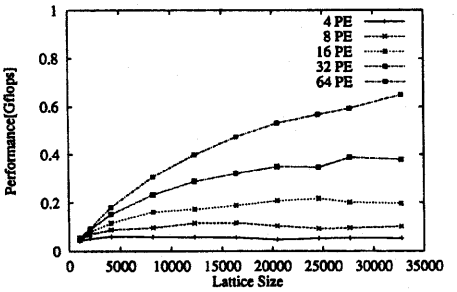

図 11 AP1000+ における性能 (平均値)

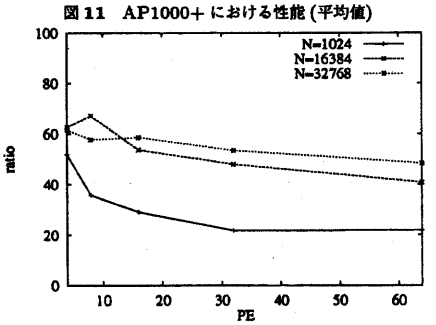

図 12 AP1000+ における通信時間と演算時間の割合

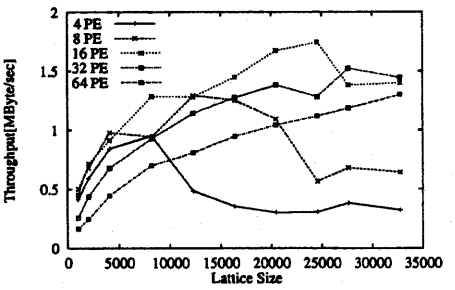

図 13 AP1000+ における通信スループット

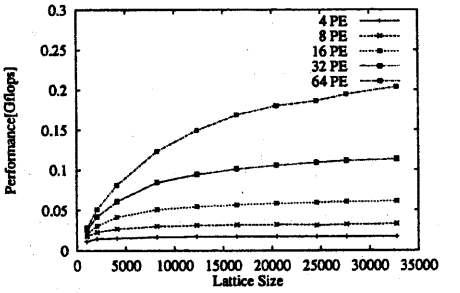

図 14 AP1000 における性能 (平均値)

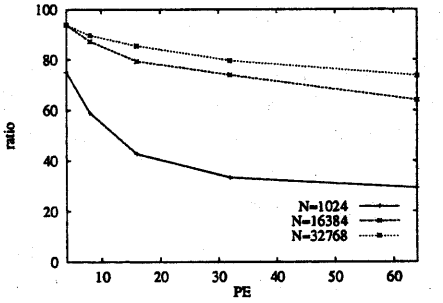

図 15 AP1000 における通信時間と演算時間の割合

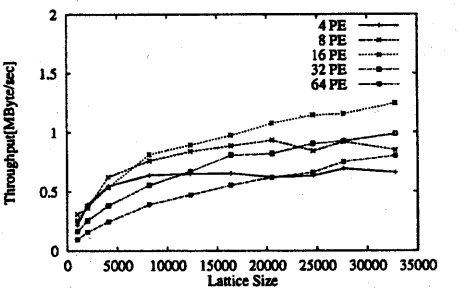

図 16 AP1000 における通信スループット

りも高い結果となった. 表 1 との比較により各ディメンションにおける通信量が 8KB 程度の時に最も効率の良い通信が実現されていることがわかる.

3.3 AP1000+

AP1000+ では 4PE ~ 64PE を評価対象とした.

図 11 に各 PE 数毎の問題サイズに対する性能を示す. 性能は 10 回の *iteration* の平均値より求めた. $N = 32768$, 64PE の時, 最高値 0.64Gflops となった. AP1000+ も Cenju-3 同様, 通信の安定した MPP である. 10 回の *iteration* における経過時間の最短時間と平均値に大差はなく, 最短時間による最高性能は $N = 32768$, 64PE の時の 0.65Gflops となった

図 12 に実行時間に占める演算時間と通信時間の比率を示す. 演算量が大きくなるにつれ, 演算の割合が増加する. PE 数の変化に対する比率の変化は Cenju-3 に比較すると大きいが etlwiz に比べると小さい.

図 13 に各 PE 数に対する通信スループットを示す. 4PE, 8PE ではスループットにピークが見られる. 表 1 に示した通信量と比較すると, 1 回の転送量が 9KB 程度の時に通信性能のピークがあることが予測される. この要因として, バッファサイズに対して転送量が溢れるとスループットが劣化することが挙げられる.

3.4 AP1000

AP1000 では, 4PE ~ 64PE を評価の対象とした.

図 14 に問題サイズに対する PE 数毎の性能を示す. 性能は 10 回の *iteration* の平均値より求めた. $N = 32768$, 64PE の時, 最高値 0.20Gflops となった. また実行時間は安定しており, AP1000 においても実効性能の最高値と平均値は等しい結果となった.

表 3　1iteration 当たりの演算時間と通信時間 [sec]

マシン	4 PE comp/comm	8 PE comp/comm	16 PE comp/comm	32 PE comp/comm	64 PE comp/comm
etlwiz	1.16/0.35	0.49/0.28	0.23/0.27	0.15 /0.14	-
Cenju-3	8.40/0.07	3.98/0.06	2.05/0.06	1.04/0.03	0.53/0.02
AP1000+	8.75/3.54	4.36/1.33	2.20/1.17	1.03/0.81	0.55/0.59
AP1000	40.50/2.65	20.18/2.38	10.34/1.76	5.19/1.32	2.66/0.96

実行時間に占める演算時間と通信時間の比率を図 15 に示す．$N = 1024$ のとき，16PE を越えると通信の比重が全体の 50% 以上を占める．$N = 16384, 32768$ のように演算量が多いとき演算の比重が増加し，その実効性能は CPU 性能に大きく依存することがわかる．

図 16 に通信スループットを各 PE 数毎に示す．4PE と 8PE の場合 AP1000+ と同様に問題サイズの増大によるスループットの減少が見られるが AP1000+ ほど顕著ではない．最もスループットが高いのは 16PE のときである．

4. 各種並列計算機の性能比較

図 17 に本研究の対象とした 4 種の並列計算機の実効性能を比較した結果を示す．実行性能は 10 回の *iteration* の中で最短時間となったデータをもとに求めた．$N = 32768$ の時の PE 数に対する性能比較となっている．Cenju-3 の 64PE における最高性能値 1.35Gflops よりも etlwiz，16PE の 1.45Gflops が勝っていた．8PE 以下のとき，etlwiz では経過時間の 60% 以上を演算時間が占め，高速な CPU 性能を十分に発揮して他の MPP と比べて良い立ち上がりを見せている．

表 3 に 10 回の *iteration* の中で最短時間となったときの 1*iteration* 当たりの演算時間と通信時間を示す．$N = 32768$ の時の PE 数に対する比較となっている．etlwiz は通信が安定していないため，平均時間の場合，32PE，$N = 32768$ の時の通信時間は経過時間の 9 割以上を占めていたのに対し，最短時間の場合には通信の割合が 5 割以下にまで減少している．Cenju-3 と AP1000+ では同程度の演算時間を費やすのに対し，通信時間は数十倍も Cenju-3 の方が高速である．図 17 の実効性能値にもその差が見られる．

5. ま と め

本稿では QCD シミュレーションプログラム QCDMPI を用いて，etlwiz と Cenju-3, AP1000+,AP1000 の性能評価を行なった．

etlwiz では通信性能と比較して CPU 性能が非常に高い．本研究において，プログラム実行時間における通信時間の割合は，平均時間の場合，PE 数の増加により大幅に上昇し，32PE で問題サイズを最大とした時の通信時間は全体処理時間の約 94% を占めていた．よって並列度の高い時の etlwiz の実効性能は通信性能に左右される．一方，最高性能値に関してはクロック周波数 333MHz という高速な CPU 性能により，今回研究の対象とした 4 種の並列計算機の中で最高値を示した．

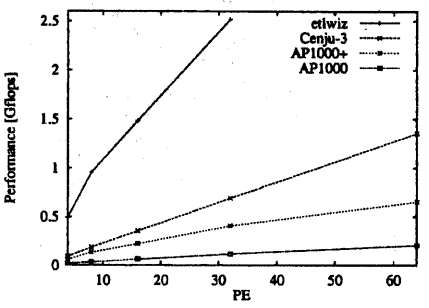

図 17　4 種の並列計算機における性能比較 (最高値)

Cenju-3 は CPU 性能に比べて，高速な通信が実現されているので，経過時間に占める通信時間の割合は 10% 以下であった．そのため QCDMPI のような粒度の大きな計算に対してはスケーラブルとなった．

AP1000+,AP1000 は 64PE で問題サイズ最大の時の通信時間が経過時間の 30 ～ 50% を占めた．AP1000+ では，4PE，8PE において問題サイズの増加に関して通信スループットに性能低下が見られた．

バランスの良い通信性能が効率の良い並列処理には不可欠であるが，本研究で用いた QCDMPI のような粒度の大きな計算に対し高い性能が維持されるのは，より CPU 性能の優れた計算機の方であった．

今後の課題としては，より多くの並列計算機の性能測定を行なうことと QCDMPI の実行モデルの作成が挙げられる．

謝辞　QCDMPI を御提供頂き，また多くの有益な御助言を頂いた帝塚山大学の日置慎治助教授に深く感謝致します．

参 考 文 献

1) 日置慎治:
http://insam.sci.hiroshima-u.ac.jp/QCDMPI .
2) 原康夫: 量子色力学とは何か, 丸善 (1986) .
3) 南部陽一郎: クォーク 素粒子物理の最前線, 講談社 (1981).
4) MPI: メッセージ通信インターフェース標準 (日本語訳ドラフト) (1996).
http://tupc3472.tezukayama-u.ac.jp/mpi-j-html/.