HPC チャレンジの測定

吉田 高志
Takashi YOSHIDA

1 はじめに

近年コモディティコンピュータをネットワークで接続して高性能並列計算機とする PC クラスタシステムがトレンドである．PC クラスタシステムがトレンドとなっている理由は，汎用のコンピュータとネットワークを用いることによる，高いコストパフォーマンスと導入の容易性である．しかし，アプリケーションの種類によっては実効性能が出ないという問題がある．理論的な性能値と実際の性能値が異なるためである．

現在 PC クラスタが多勢を占める TOP500 での性能評価では，LINPACK が評価基準に採用されている．しかし Linpack は浮動小数点演算の性能を計測するプログラムであり，システム全体の性能を評価できない．そういった背景より近年，HPC チャレンジベンチマークというものがある．HPC チャレンジベンチマークは LINPACK を含む，7 項目のテストから構成されるベンチマークセットである．演算性能だけでなくネットワーク性能やメモリアクセス性能など，システムを多角的に評価することが可能である．

本報告では自作マシン，及び研究室の所有する Satuki というマシンで HPC チャレンジの計測を行うことで，マシンによって高性能を出せる分野が違うことを示した．

2 HPC チャレンジ

HPC チャレンジとは，全 7 項目のテストから構成されるベンチマークセットである．HPL，DGEMM，STREAM，PTRANS，Random Access，FFTE，Communication bandwidth and latency から構成される．なお，パラメータチューニングが可能な項目は，HPL と PTRANS である．Communication bandwidth and latency はパラメータ固定，他の項目に関しては，HPL のデータサイズ N に依存する．

2.1 HPC チャレンジの評価方法

現在，複数の CPU が同等な立場で処理を分担する SMP システムによるメモリ共有型計算機，MPI を用いた分散メモリ型計算機，そして 2 つを組み合わせたハイブリッド計算機がある．特にハイブリッド計算機においてはメモリ共有されている単位で 1 ノードとすることが多く，それをネットワークで結合したものが 1 つのシステムとなっている．このようなハイブリッド計算機では全体と部分の評価が必要になるため，HPC チャレンジベンチマークには以下の 3 つの評価方法がある．

- Global system performance(G)
  全ノードによる総合性能を評価する．あらゆるシステムが対象となる．
- Single Environment(SN)
  各ノードの性能を評価するものであり，単一プロセスでの性能評価である．特にハイブリッド計算機システムでの 1 ノードの評価をする．
- Embarrassingly Parallel(EP)
  SN と同じく，ハイブリッド計算機システムの 1 ノードの評価をする．各ノードにおいて，複数プロセスを同時に使い，メモリ使用量の奪い合いによる性能の劣化の程度を評価するテストである．

2.2 HPL

2.2.1 HPL の概要

HPL(High-Performance Linpack) は本来単一 CPU 用のライブラリである LINPACK を分散メモリ型並列計算機用に並列化したものである．LINPACK は米テネシー大学の Jack J. Dongarra 博士らが開発した性能計測プログラムであり，LU 分解に基づく N 次元の連立一次方程式の解を求めるプログラムであり，浮動小数点演算の性能を計測することができる．HPL では全ノードを使った演算性能評価が行われる．

2.2.2 パラメータ

HPL では以下の 16 項目についてパラメータを決めることができる．性能に大きく影響を及ぼすのは，問題サイズ N，ブロックサイズ NB，プロセスグリッドである．

- 問題サイズ N
- ブロックサイズ NB
- プロセスグリッド
- 解のチェックにおける残差の境界値
- Panel Factorization のアルゴリズム
- 再帰的 PanelFactorization のアルゴリズム
- 再帰的 Factorization におけるサブパネル数
- 再帰的 Factorization におけるサブパネル幅の最小値
- Panel Broadcast のトポロジ
- Lookahead の深さ
- Update における通信トポロジ
- Long における U の平衡化処理の有無

- mix における行数の境界値
- L1 パネルの保持の仕方
- U パネルの保持の仕方
- メモリの alignment

問題サイズ N は Linpack の中で最も影響を及ぼす．通常，N が大きくなるほど良い結果が得られるが，N が大きくなるほどメモリ使用量は増加する．通常 OS とデーモンが起動していると，メモリ容量の 10 %〜20 %が使用されるため，総メモリ量の 80 %程度を利用するように N の設定を行う．行列の各要素は double 型である．このような条件を満たす N は式 (1) より求めることができる．

$$N = \sqrt{MemorySize(byte) \times 0.8/8} \qquad (1)$$

ブロックサイズ NB は，粒度のことである．NB が大きくなると，通信量が減るがロードバランスが悪くなり，NB が小さくなると，通信量が増えるがロードバランスが良くなる．NB の値を 32〜256 にすると，良い結果が得られる．また良好な結果を示す NB があれば，その値の整数倍も良好な結果を示すことがある．

2.3 DGEMM

実数行列の積を計算するプログラムで，演算性能を評価する．LINPACK では，演算の主要部分が DGEMM となり，性能を決める重要な要素となる．ノード単体での性能評価するシングル環境と多重負荷環境でのテストを行う．DGEMM のデータサイズは，HPL のデータサイズ N と MPI のプロセス数から式 (2) で求められる．また，今後式の中での Process とは MPI のプロセス数である．

$$Size = N^2/(\sqrt{Process} \times 2) \qquad (2)$$

DGEMM は，特に CPU コアの演算性能を測る．単位は GFLOPS である．

2.4 STREAM

メモリバンド幅の性能を評価を行う．複写 (Copy)，定数倍 (Scale)，総和 (Add)，積和 (Triad) という 4 つの評価がある．全ノードの性能評価には依存せず，ノード単体での性能に依存する．そのためシングル環境と多重負荷環境での上記の 4 つの項目について，計 8 項目についてテストを行う．STREAM の配列データサイズは，HPL のデータサイズ N と MPI のプロセス数から式 (3) で求められる．

$$Size = N^2/(Process \times 3) \qquad (3)$$

STREAM は上記の 4 つの評価方法を，メモリから CPU にデータを読み込むことと，CPU からメモリに書き込むことによってバンド幅を計測している．計測したメモリバンド幅が大きいほど，一度に送れるデータ量が増える．単位は GB/ s となる．

2.5 PTRANS

行列の転置で全ネットワーク転送性能を評価を行う．行列のデータはメモリ領域に入っており，転置を行うとメモリ領域でデータが全部入れ替えることになる．さらに分散して行列を格納している場合，転置を行うと分散していたデータ全部の入れ替えをネットワークを介して行うことになる．このときネットワーク転送に多大な負荷がかかるため，ネットワーク転送性能のが分かる．全ノードを使用したネットワーク性能のテストを行う．

データサイズは，HPL のデータサイズ N の同じ，もしくは PTRANS 独自にデータサイズ N を決められる．転置されるデータサイズは N/2 となる．また NB も独自に決めることができる．

2.6 Random Access

RandomAccess は，ノード単体テストでは，メモリのランダムアクセス性能を評価する．乱数で生成されたテーブルに基づき，配列データを順次格納することでメモリのランダムアクセスを評価する．全ノードテストでは，ノード間の MPI 転送性能の評価する．性能の単位は（Gup/s：Giga updates per second）であり，1 秒間に更新する要素数である．シングル環境と多重負荷環境，全ノードを使用した総合性能のテストを行う．RandomAccess のテーブルサイズは，HPL のデータサイズ N と MPI のプロセス数から式 (4) で求められる．

$$Size = N^2/Process \qquad (4)$$

2.7 FFTE

高速フーリエ変換の性能評価を行う．シングル環境と多重負荷環境，全ノードを使用した総合性能のテストを行う．FFTE で用いられるデータサイズは，以下の計算式を超えない最大の 2 のべき乗となっている．なお式 (5) がシングル環境と多重負荷環境，式 (6) が総合環境でのデータサイズである．

$$Size < N^2/(Process \times 2 \times (fftw_complexsize)) \qquad (5)$$

$$Size < N^2/(Process \times 3 \times (fftw_complexsize)) \qquad (6)$$

FFTE は，科学技術計算におけるフーリエ変換の重要性を考慮に入れ，FFT の演算性能を評価するものである．

2.8 Communication bandwidth and latency

データ転送能力を評価を行う．一般にデータ転送時間 T は，データ転送立ち上がり時間とデータ量をバンド幅で割ったものを足したものである．もしデータ量が大きくなければ，転送時間 T は立ち上がり時間に依存す

Table 1　自作マシンのシステム構成

CPU	Intel Pentium4 3.0GHz
L1/L2	16KB/1MB
Memory	DDR-SDRAM 512MB
OS	WindowsXP Professional
通信ライブラリ	mpich-1.2.7p1 by gcc3.3

Table 2　Satuki

CPU	AMD Athlon64X2 2.0GHz
L1/L2	128KB × 2/512KB × 2
Memory	DDR-SDRAM 1GB
OS	WindowsXP Professional
通信ライブラリ	mpich-1.2.7p1 by gcc3.3

る．一方データ量が大きくなると，T はバンド幅に依存することになる．バンド幅とレイテンシの 2 点について評価を行う．バンド幅，レイテンシの評価スキームは Ping-Pong 転送スキームと Ring 転送スキームの 2 種類である．Ring 転送スキームには，MPI のランク順に転送する方法とランダムに転送する方法がある．バンド幅テストでは 2M バイトのデータを，レイテンシテストでは 8 バイトのデータを転送することで評価を行う．

3　実行結果

今回，HPC チャレンジを実行したのは，自作マシン，および研究室の所有する Satuki である．各マシンの構成を Table.1，Table2 に示す．なお，本報告において計測しているのは 1 ノードのみであり通信が発生しないため，ネットワークが性能に関係しているテスト項目については省略する．このため今回掲載する結果は，HPL，DGEMM(SN)，STREAM(SN)，RandomAccess(SN)，FFTE(SN) とする．

3.1　HPL

3.1.1　自作マシン

自作マシンでは，メモリが 512MB であるため，(1) 式より導かれる N 値は 7155 である．通常 NB が 24 以下である場合，あまり良い性能が得られない．得られた N 値 7100 に固定し，NB 値を 28 から 4 刻みで 48 まで計測したところ，32 および 48 が最適な値であることが得られた．2.2.2 項で述べたとおり良好な NB 値の整数倍が良い値を示すことがあるため，32 と 48 の整数倍の値を計測した．その結果を Fig.1 に示す．

Fig.1 より最適な NB 値は 288 である．最適な N 値を求めるため，式 (1) で導かれた 7100 から 100 刻みで計測した．Fig.2 に結果を示す．最も高い性能を得られた

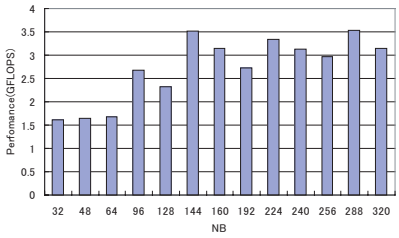

Fig. 1　最良な NB 値の検討:自作マシン

のは，N が 7200 のときの 3.550GFLOPS という計測結果になった．

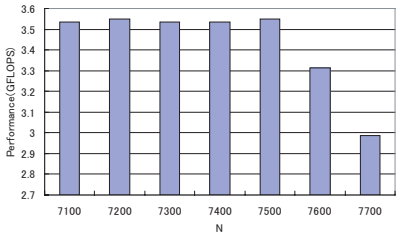

Fig. 2　最良な N 値の検討：自作マシン

3.1.2　Satuki

Satuki は，メモリが 1GB であるため，(1) 式より導かれる N 値は 10000 である．得られた N 値 10000 に固定し，NB 値を 28 から 4 刻みで 48 まで計測したところ，40 が最適な値であることが得られた．40 の整数倍の値を計測した．その結果を Fig.3 に示す．

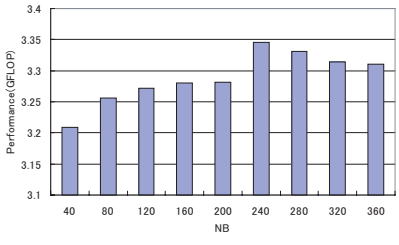

Fig. 3　最良な NB 値の検討：Satuki

Fig.3 より最適な NB 値は 240 である．最適な N 値を求めるため，式 (1) で得られた 10000 から 200 刻みで計測した．Fig.4 に結果を示す．最も高い性能を得られたのは，N が 10600 のときの 3.344GFLOPS という計測結果になった．

3.2　DGEMM

HPL 以外の項目では良好な性能を計測するデータサイズの決め方が，HPL ほど明確でない．そこでまず HPL で計測した N の値で検証すると，HPL で計測した周辺の N の値では，HPL 以外の項目では性能に大きな違いが存在しなかった．このため HPL で計測した様に，N の値を小さな間隔で変更しても性能に差が出ないと予測して，データサイズ N を 1000 刻みという大きな間隔で，8000 まで計測を行った．次節以降の STREAM，RandomAccess，FFTE も同様のデータサイズで計測を

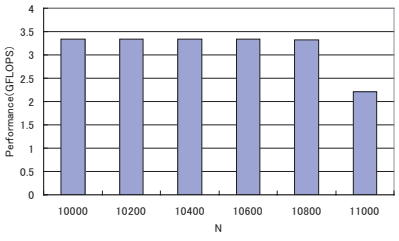

Fig. 4　最良な N 値の検討：Satuki

行っている．

DGEMM の計測結果を Fig.5 に示す．計測結果全体を比較すると，Pentium4 を搭載した自作マシンのほうが良好な結果である．これは CPU コア単体の性能は Pentium4 が Athlon より優れていることを示している．自作マシンではデータサイズ N が 8000 の所で急激に性能が落ちているが，これはメモリ容量が Satuki のほうが大きいためであると考えられる．

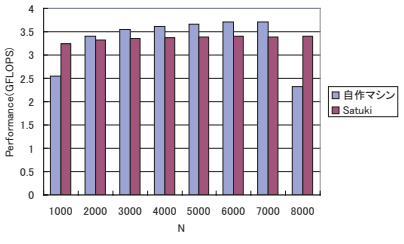

Fig. 5　DGEMM の計測

3.3　STREAM

STREAM の計測結果を Fig.6，Fig.7 に示す．自作マシン，Satuki とも N が 1000 のときの計測結果が良好である．それ以降は横ばいになっている．自作マシンにおいて N が 8000 のとき，大幅に計測結果が落ちている．これはメモリにデータが収まらずスワップが起こったためと考えられる．

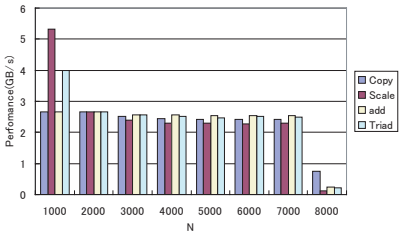

Fig. 6　STREAM の計測:自作マシン

3.4　RandomAccess

RandomAccess の計測結果を Fig.8 に示す．データサイズが小さいとき良好な値をとることが分かる．これは，データサイズが小さなときのほうがアクセスが速いためであると考えられる．

また，自作マシンより Satuki のほうが全体的に良好な結果が出ている．これはメモリコントローラが CPU 側についており，AMD の CPU のほうがメモリアクセスが良好であるためと考えられる．

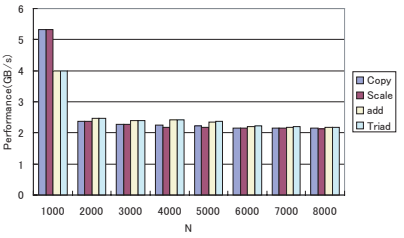

Fig. 7　STREAM の計測:satuki

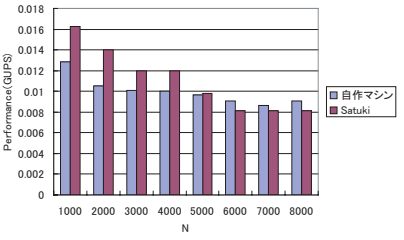

Fig. 8　RandomAccess の計測

3.5　FFTE

FFTE の計測結果を Fig.9 に示す．Satuki がある程度以上のデータサイズになると性能が劣化しており，自作マシンは性能が劣化していない．この相違点は，キャッシュサイズの違いが性能差になっていると考えられる．

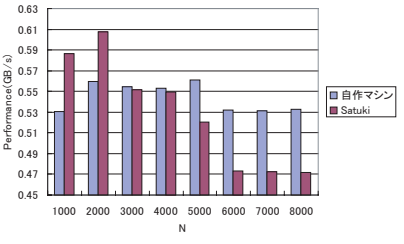

Fig. 9　FFTE の計測

4　今後の課題

本報告で計測したデータは 1 ノードのみである．HPC チャレンジが真の効果を発揮するのは，複数ノードで計測した場合である．また，未だに調査不足しているベンチマークもある．今後はそれらのベンチマークの調査，及び SuperNova および Xenia クラスタの HPCC 計測を行う．その上でどのような実アプリケーションで大きな性能を出すことが可能か，検証していく．

参考文献

1) HPC Challenge
http://icl.cs.utk.edu/hpcc/
2) HPC チャレンジでの SX システムの性能評価
`http://www.cc.tohoku.ac.jp/refer/pdf_data/v38-1p5-28.pdf`