Xenia クラスタにおける HPL の計測とパラメータチューニング

Measurement and parameter tuning of HPL in Xenia cluster system

釘井 睦和
Yoshikazu KUGII

Abstract: HPL is one of the implementation of Linpack benchmark, and the parameter is very numerous. Since the optimal parameter depends on the composition of the system, it is very difficult to get the optimal parameter. Therefore, the parameter is considered through the parameter tuning in Xenia cluster system that introduced newly.

1　はじめに

近年，超並列計算機に代わってパーソナルコンピュータやワークステーションなど単一で稼働するコンピュータをネットワークで繋ぎ，1 つの計算資源として利用することができる PC クラスタが注目されている．また，CPU 性能の向上や Myricom 社 [1] の Myrinet のようにリンク速度が 1.26Gbps と超並列計算機のネットワークと同等，もしくはそれ以上の性能のものが一般に利用できるようになり，大規模なクラスタも構築されるようになった．そのため，連立一次方程式の直接解法を基本とした Linpack Benchmark の公式結果から，世界の高性能コンピュータの上位 500 位をリストアップしている Top500[2] においても，クラスタシステムの活躍はめざましいものがある．Top500 に挑戦することは，高性能の計算サーバを有することを世界にアピールする最大の機会である．こういった高性能コンピュータのランキングはハイパフォーマンスコンピュータのユーザやベンダにとって大きな興味の対象であり，Top500 はその中でも最大のものとなっている．

そこで，本研究室にも Top500 ランクインを目指し，PC クラスタ Xenia(Fig. 1) が導入された．本発表ではこの Xenia の Linpack Benchmark の計測結果およびそのパラメータチューニングについて述べる．

Fig. 1　Xenia クラスタ

2　Xenia

PC クラスタ Xenia は，IBM 社のサーバ用ワークステーション IntelliStation M Pro[3] を 64 台用いた PC クラスタである．主なハードウェア構成，ネットワーク構成は Table 1 の通りである．

Table 1　Xenia のハードウェア構成

CPU	Intel Xeon 2.4GHz × 2 × 64
Memory	1GB × 64(計 64GB)
OS	Red Hat Linux 7.3
通信ライブラリ	MPI
通信プロトコル	GM，TCP/IP
通信媒体	Myrinet 2000，Fast Ethernet

Intel Xeon プロセッサの特徴は以下の Table 2 の通りである．Xeon プロセッサは Pentium4 とアーキテクチャが類似しているが，デュアルプロセッサに対応しているという点で異なる．

Table 2　Xeon プロセッサ概要

L1 キャッシュ	8KB
L2 キャッシュ	512KB
クロック当たりの命令発行数	6
整数パイプライン	4
浮動小数点パイプライン	2
システムバス速度	400MHz
3D 拡張命令	SSE2

Xeon プロセッサでは 8KB のデータキャッシュ以外に実行トレースキャッシュを備え， デコード済みのマイクロオペレーションをプログラムの実行順に最大 12KB 格納することができ，高速な演算が実現する．SSE2 (ストリーミング SIMD 拡張命令 2) 命令セットを用いることにより，クロック周波数の 2 倍の浮動小数点演算が可能となる．このことから Xenia のピーク性能値は 614.4GFlops となる．

3 HPL

HPL[4] (High-Performance LINPACK Benchmark) は，分散メモリ型並列計算機用のベンチマークソフトウェアであり，ガウス消去法を用いた密行列連立 1 次方程式の求解における実行時間により性能を評価する．HPL は様々なパラメータを計算機の特性に合わせて設定したり，高度に最適化された行列演算カーネルを組み込むことで，より高い性能を得ることができるようになっている．

3.1 アルゴリズム

HPL では，まずプロセスをプロセスグリッドという 2 次元配列の格子状に並べる．次に，係数行列を複数の正方形に分解してプロセスグリッド上に割り当てる．LU 分解処理は Panel Factorization，Panel Broadcast，Update，Backward Substitution というフェイズから構成される．それぞれにおいてパネル列 LU 分解，分解済みパネルの送信，未分解小行列の更新計算，後退代入演算による求解を行う．

3.2 パラメータ

HPL では以下の 16 項目についてのパラメータ [5] を設定できる．性能に大きく影響を与えるものは問題サイズ N，ブロックサイズ NB，プロセスグリッド (P，Q)，Broadcast のトポロジーなどである．

- 問題サイズ N
- ブロックサイズ NB
- プロセスグリッド (P，Q)
- 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 Panel Factorization のアルゴリズム
- 再帰的 Factorization における subpanel 数
- 再帰的 Factorization における subpanel 幅の最小値
- Panel Broadcast のトポロジー
- Look-ahead の深さ
- Update における通信トポロジー
- 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

4 コンパイラ

Xenia には gcc，Intel，pgi の 3 種類のコンパイラを用意した．それぞれの特徴について述べる．

4.1 gcc

gcc[6] (GNU Compiler Collection) は C，C++，Objective C，Fortran などで書かれたプログラムをコンパイルすることが可能である．GNU プロジェクトに使用されているため，UNIX 系 OS では現在最も広く普及している．

4.2 Intel

Intel C++/Fortran Compiler[7] は Intel が開発している C++/Fortran 用のコンパイラである．特徴としては，Pentium 系の CPU に最適化したバイナリを生成することがあげられる．

4.3 pgi

pgicompiler[8] は，Portland Group により開発されているコンパイラで，HPC(High Performance Computing) において，最適化能力が優れていると評されている．

5 計測結果

5.1 最適パラメータ

前節で述べた計測に大きく影響するパラメータそれぞれについて，結果から得られた Xenia の最適なパラメータについて述べる．

5.1.1 問題サイズ N

問題サイズ N は HPL で解く問題の大きさである．つまり，HPL では N 次元連立方程式を解くことになる．N は HPL の結果に最も大きな影響をもたらす．N の値は，計算対象となる計算機の全メモリ容量の 80 %を使用するように設定する．これより，導き出された最適な N の値は 82897 となる．この値付近で，N を変動させ HPL の計測を行った結果を Fig. 2 に示す．用いたコンパイラは gcc-2.96，最適化オプションは -fomit-frame-pointer -O3 -funroll-loops で，N 以外の主なパラメータは NB:80，(P，Q):(8，16)，BCAST:1ringM である．

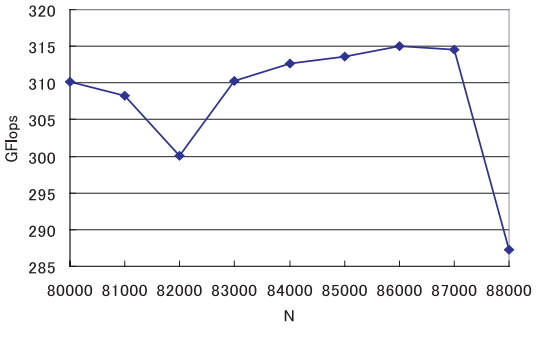

Fig. 2　N による比較

この結果を見ると，問題サイズが 86000 の時に最も良い値となっている．計算で求めた最適な N である 82897 よりも大きくなっているが，これはメモリの空き容量を

増加させるために様々なサービスを停止しているので多くのメモリを使用することが可能になっているためである．86000 ではメモリの 95 %以上が使用されスワップが起こっているのでこれ以上問題サイズを上げると，極端に結果が悪くなる．

5.1.2 ブロックサイズ NB

ブロックサイズ NB は問題をどのような大きさに分けるかを決める粒度のことである．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．また，良い NB の値があればその整数倍も良い結果をもたらすことがある．NB は通常，HPL のパラメータを決定する際に最も設定が難しいとされている．最適な NB を求めるために ATLAS のインストールログから得られた 40 の倍数と 48 の倍数を，2 の累乗である 32 の倍数の合計 3 通りの NB を用いて計測を行った結果を Fig. 3 に示す．用いたコンパイラは gcc-2.96，最適化オプションは -fomit-frame-pointer -O3 -funroll-loops で，NB 以外の主なパラメータは N:80000，(P，Q):(8，16)，BCAST:1ringM である．

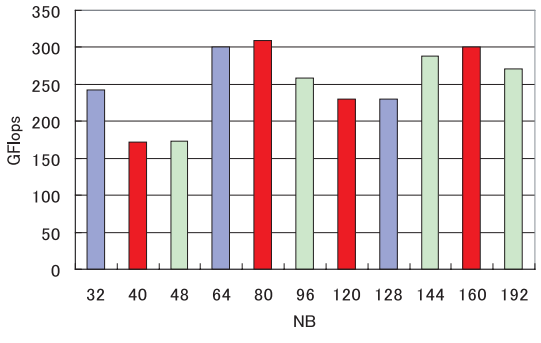

Fig. 3　NB による比較

この結果より，40 の倍数である 80 が最も良い NB であるといえる．この 40 という値は，ATLAS が CPU のキャッシュサイズを認識する際に導き出す NB の値である．

5.1.3 プロセスグリッド（P，Q）

プロセスグリッド (P，Q) は問題の行列をそれぞれのプロセスにどのように分割するかを示す．必然的に P と Q の積が実行ノード数となる．P と Q は等しいか，P より Q が大きい方が良いとされている．Xenia の実行プロセッサ数は 128 なので，この条件に合う (P，Q) は (8，16) となる．

5.1.4 Panel Broadcast のトポロジー

Panel Broadcast のトポロジーには increasing-1ring，increasing-2ring，Bandwidth-reducing の 3 種類と，次の Panel Factorization を行うプロセスにメッセージ送信をさせない modified 版がそれぞれ 3 種類の計 6 種類が存在する．それぞれの方法に関して計測を行った結果を Fig. 4 に示す．用いたコンパイラは gcc-2.96，最適化オプションは -fomit-frame-pointer -O3 -funroll-loops で，BCAST 以外の主なパラメータは N:80000，NB:64，(P，Q):(8，16) である．

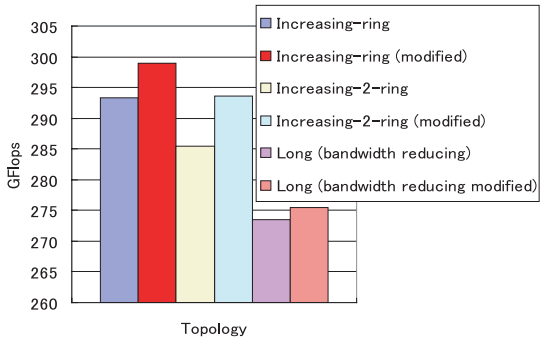

Fig. 4　Panel Broadcast による比較

この結果より，Xenia における BCAST は Increasing-ring (modified) が最も良い性能を出すことが分かる．

5.2 コンパイラによる比較

gcc，Intel，pgi の 3 種類のコンパイラを用いて計測比較を行った．各コンパイラの最適化オプションは以下の通りである．

- gcc

```
CCFLAGS      = -fomit-frame-pointer
                  -O3 -funroll-loops
LINKFLAGS    = $(CCFLAGS)
```

- Intel

```
CCFLAGS      = -O3 -axKW
LINKFLAGS    = $(CCFLAGS)
```

- pgi

```
CCFLAGS      = -fast -Mvect=sse
                  -DMAIN_=main
LINKFLAGS    = -fast -Mvect=sse
                  -DMAIN_=main -Mnomain
```

gcc ではデフォルトで設定されているオプションを，Intel では Xeon プロセッサと類似したアーキテクチャをもつ Pentium4 に最適化するオプションを，pgi では SSE2 をサポートするオプションをそれぞれ用いている．用いたパラメータは 5.1 節で導き出した Table 3 の通りである．

Table 3　Xenia における最適パラメータ

N	86000
NB	80
(P，Q)	(8，16)
Broadcast	Increasing-ring (modified)

計測結果は Fig. 5 のようになった．この結果より，HPL の計測に関しては pgi が最も良い値を出すことが分かる．

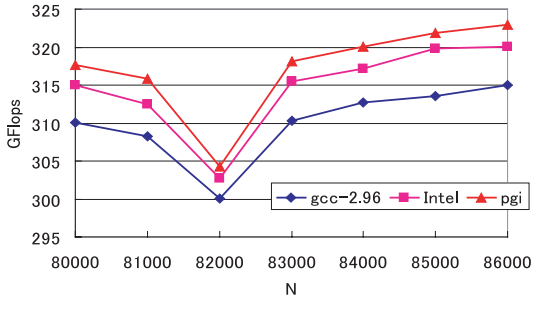

Fig. 5　コンパイラによる比較

6　まとめ

5 節で最適なパラメータおよびコンパイラについて検討した．このパラメータ，コンパイラを用いて Panel Factorization のアルゴリズム 3 通り，再帰的 Panel Factorization のアルゴリズム 3 通りを組み合わせた 9 通りで数回計測を行い，9 月 24 日現在で 323.3GFlops(Fig. 6) という値を出すことができた．この値は，2002 年 6 月度の Top500(Fig. 7) において 146 位，国内では 25 位という成績である．

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WR01L2L2       86000    80     8    16            1311.64              3.233e+02
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) =        0.0038852 ...... PASSED
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) =        0.0038544 ...... PASSED
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) =        0.0006775 ...... PASSED
================================================================================
```

Fig. 6　最高記録

Xenia の導入により，本研究室も世界有数の高速計算サーバを利用できる環境になり，研究の発展が期待される．また，Top500 のランクインが実現されれば多数の企業や研究所との共同研究の確立や，ソフトウェアやツールの交換などといったコミュニティの確立が期待できる．

Rank	Manufacturer	Computer	R_{max}(GFlops)	Installation Site	Country	Year	Installation Type
141	IBM	Netfinity Cluster PIII 1.13 GHz - Eth	326.00	Saudi Aramco	Saudi Arabia	2001	Industry
142	IBM	Netfinity Cluster PIII 1.13 GHz - Eth	326.00	WesternGeco	USA	2002	Industry
143	IBM	Netfinity Cluster PIII 1.13 GHz - Eth	326.00	Compagnie Generale de Geophysique (CGG)	UK	2002	Industry
144	IBM	Netfinity Cluster PIII 1.13 GHz - Eth	326.00	Compagnie Generale de Geophysique (CGG)	UK	2002	Industry
145	IBM	Netfinity Cluster PIII 1.13 GHz - Eth	326.00	WesternGeco	Egypt	2001	Industry
146	Fujitsu	VPP700/160E	319.00	Institute of Physical and Chemical Res. (RIKEN)	Japan	1999	Research
147	SGI	ORIGIN 3000 400 MHz	315.50	CSAR at the University of Manchester	UK	2001	Academic
148	SGI	ORIGIN 3000 400 MHz	315.50	ERDC MSRC	USA	2001	Research
149	SGI	ORIGIN 3000 400 MHz	315.50	Manufacturing Company	Sweden	2002	Industry
150	SGI	ORIGIN 3000 400 MHz	315.50	NASA/Ames Research Center/NAS	USA	2001	Research
151	SGI	ORIGIN 3000 400 MHz	315.50	NASA/Goddard Space Flight Center	USA	2001	Research
152	SGI	ORIGIN 3000 400 MHz	315.50	Silicon Graphics	USA	2001	Vendor
153	SGI	ORIGIN 3000 400 MHz	315.50	US Army Research Laboratory (ARL)	USA	2000	Research
154	IBM	pSeries 690 1.1GHz GigEth	315.00	I2 Technologies Inc.	USA	2002	Industry
155	NEC	SX-6/40M5	311.70	Communications Res. Lab. (CRL)	Japan	2002	Research
156	IBM	SP Power3 222 MHz	307.60	ERDC MSRC	USA	2000	Research

Fig. 7　2002 年度 6 月度の top500

7　今後の課題

- lam/mpi での測定
- 問題サイズ 82000 での性能劣化の検討
- 他のクラスタのパラメータチューニングとの比較

参考文献

1) Myricom Home Page
http://www.myri.com

2) TOP500 Supercomputer Sites
http://www.top500.org/

3) IntelliStation M Pro
http://www-6.ibm.com/jp/pc/intellistation/ismpr27/ismpr27a.html

4) HPL Algorithm
http://www.netlib.org/benchmark/hpl/algorithm.html

5) 笹生 健，松岡 聡．HPL のパラメータチューニングの解析．ハイパフォーマンスコンピューティング．91-22．2002

6) GCC Home Page
http://gcc.gnu.org

7) Intel Compilers
http://www.intel.com/software/products/compilers/

8) The Portland Group Compiler Technology
http://www.pgroup.com/