Core クラスタにおける Benchmark 計測

Measurement of Benchmark in Core Cluster System

高畑 泰祐
Taisuke KOHATA

Abstract: In recent years, PC clusters has been receiving increased attention. The Core Cluster System in Keihanna is composed of IBM PowerPC 970. This paper shows process of parameter tuning to find the results of optimal parameter in the cluster to enter TOP500 Supercomputer sites, and feature of the cluster.

1　はじめに

近年，科学技術の発達に伴った問題の大規模化・複雑化により，高性能な計算資源の需要が高まっている．これまででは高い性能を実現するために設計された専用計算機が必要であったが，近年では PC クラスタシステムが従来の超並列計算機に代わって注目されている．PC クラスタシステムとは，一般的に利用されているコンピュータをネットワークで接続し，１つの計算機として利用できるようにしたシステムのことである．汎用の製品がそのまま使用でき，ソフトウェアもフリーのものが数多く提供されているので非常に高いコストパフォーマンスを得ることができる．これによって一部の研究機関や企業しか所有することができなかった高性能な計算機を，研究室やグループ単位でも所有することが可能となってきた．世界中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 Supercomputer sites においても，その約 3 分の 2 をクラスタシステムが占めており，スーパーコンピュータにひけをとらない性能を示していることがわかる．本稿ではこの TOP500 へのランクインを目指して Fig. 1 に示す Core Culster System において Benchmark の計測を行った．

Fig. 1　Core Cluster System

2　Benchmark とは

Benchmark とはコンピュータのハードウェアやソフトウェアの処理性能を計測する試験である．Benchmark にはユーザの主観が入らないので客観的な数値としてシステム性能の一つの指標とすることができ，また，Benchmark によっては計算機の能力を限界まで使用するので動作テスト・耐久テストとしても使用できる．しかし各 Benchmark によって処理は異なるので，用途に応じた評価の指標を決定し，それに基づいて評価を行わなければならない．したがって Benchmark がコンピュータの性能を表す絶対的なものではない．

計測に用いたクラスタに関する並列計算 Benchmark は，TOP500 の評価基準にもなっている Linpack Benchmark と，姫野 Benchmark，NAS Parallel Benchmarks の 3 種類である．

2.1　HPL

HPL(High-Performance Linpack Benchmark) とは，LinpackBenchmark の実装の一つであり，分散メモリ型並列計算機用の Benchmark ソフトウェアである．密行列連立一次方程式を解く際の実行時間によって性能評価を行う．この Benchmark は TOP500 で採用されており，問題サイズやブロックサイズなどのパラメータを独自に設定することができる．

2.2　姫野 Benchmark

理化学研究所の情報基盤センター長の姫野龍太郎氏が非圧縮流体解析コードの性能評価のために考えたもので，ポアッソン方程式解法をヤコビの反復法で解く場合に主要なループの処理速度を計るものである．姫野 Benchmark では反復計算に三次元配列を用い，その配列を x，y，z 方向に分割 (グリッド分割) して各 CPU に割り振って計算している．その際の配列のサイズは Table 1 に示す 5 パターンの指定ができ，分割パターンは分割数の積が計算に使用する CPU の数と同数であればよい．

Table 1　配列サイズと分割要素

配列サイズ	分割要素
XL	1024 × 512 × 512
L	512 × 256 × 256
M	256 × 128 × 128
S	128 × 64 × 64

2.3　NAS Parallel Benchmarks

NAS Parallel Benchmarks(NPB) は，NASA Ames Research Center で開発された，並列コンピュータのための Benchmark である．NPB は，5 つの Parallel Kernel Benchmarks と，3 つの Parallel CFD(Computational Fluid Dynamics) Application Benchmarks から構成されている．Table 2 にその構成を示す．

Table 2　NAS Parallel Benchmarks の対象問題

Parallel Kernel Benchmarks	
EP	乗算合同法による一様乱数，正規乱数の生成
MG	簡略化されたマルチグリッド法のカーネル
CG	正値対称な大規模疎行列の最小固有値を求めるための共役勾配法
FT	FFT を用いた 3 次元偏微分方程式の解法
IS	大規模整数ソート
Parallel CFD Application Benchmarks	
LU	Synmetric SOR iteration による CFD アプリケーション
SP	Scalar ADI iteration による CFD アプリケーション
BT	5 × 5 block size ADI iteration による CFD アプリケーション

また各問題には，問題サイズの異なる 5 つのクラス (A，B，C，W，S) が定義されている．それらの相違点は，基本的には問題サイズや反復回数の違いにある．

3　Core Cluster System

計測を行ったクラスタは「けいはんなプラザ」内に設置されており，126 台の IBM Blade Center JS20 で構築されている．Core Cluster System の主な構成を Table 3 に示す．

PowerPC 970 は 1 プロセッサにつき 2 つの浮動小数点演算ユニット (FPU) を搭載しており，各ユニットは 1 クロックで 1 回の演算を行うことが可能である．これから Core Cluster System のピーク性能値 (Rpeak) は式 (1) より 1.6128TFlops となる．

$$Rpeak = CPU \times ClockFrequency \times FPU \qquad (1)$$

Table 3　Core のハードウェア構成

ノード数	252
CPU	IBM PowerPC 970　1.6 GHz × 2
メモリ	ECC DDR SDRAM　2.5 GB
OS	SuSE Linux Enterprise Server 8.0 ServicePack 3.0a for PPC
通信ライブラリ	mpich-1.2.5
通信媒体	Myrinet
コンパイラ	GCC 3.2.2

HPL の計測には行列演算ライブラリを用いるが，一般に GOTO BLAS のパフォーマンスがその他の BLAS に比べて高いとされているので，計測は今回はこれを用いた．

4　計測結果

今回行った 3 種類の Benchmark のそれぞれの計測結果を順に以下に示す．

4.1　HPL

HPL では 17 種類のパラメータを設定することが可能であるが，計測値に最も大きな影響を与える次の 3 種類について調査・検討を行った．

4.1.1　ブロックサイズ (NB)

1CPU のみを用いて NB10〜319 までの 310 パターンで計測を行った結果を Fig. 2 に示す．問題サイズ (N) は 14000 を用いている．

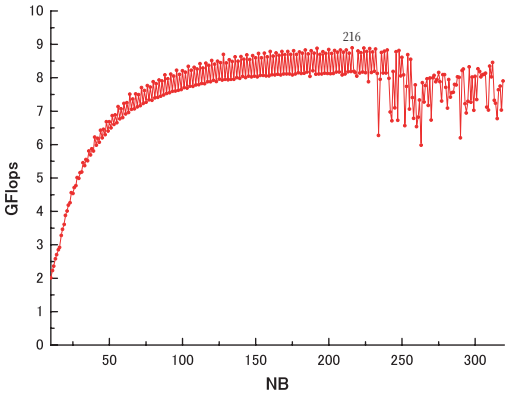


Fig. 2　NB の検討

これより最適な NB の値は 216 であるとわかる．

4.1.2　Panel Broadcast のトポロジーによる比較

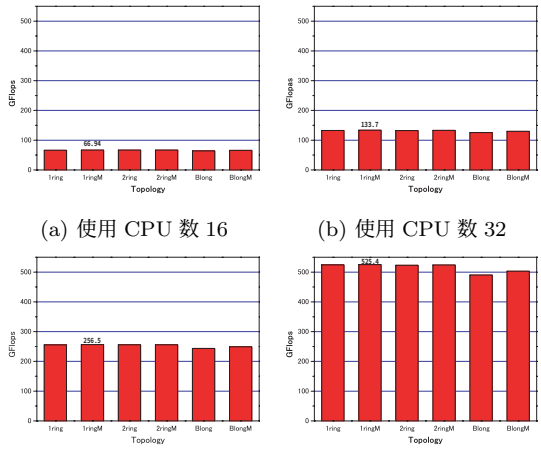
使用 CPU 数を 16，32，64，128 にした時の Panel Broadcast の各トポロジーに関して計測を行った結果を Fig. 3 に示す．その他のパラメータは Table 4 に示すものを用いた．

この結果を見ると，どの CPU 数においても Panel Broadcast のトポロジーには 1ringM が最適である．

Table 4　BCAST 以外のパラメータ

# CPU	16	32	64	128
N	44000	63000	89000	126000
NB	216			
(P,Q)	(4,4)	(4,8)	(8,8)	(8,16)

#:number of



(a) 使用 CPU 数 16　(b) 使用 CPU 数 32

(c) 使用 CPU 数 64　(d) 使用 CPU 数 128

Fig. 3　BCAST の比較

4.1.3　問題サイズ (N)

N の値は大きくなるほど良好な結果が得られる．そこで全メモリの約 80 %を使用するように計算すると，175000 となる．よって 160000〜180000 まで 5000 刻みで 7 パターンの計測を行った結果を Fig. 6 に示す．また今回は計算機のロードバランスを考慮した場合の以下の条件で求めた N のパターンも計測し Fig. 6 に加えた．その他のパラメータは Table 5 に示すものを用いた．

ロードバランスを重視した N の条件

- N が P × NB で割りきれる
- N+1 が Q × NB で割りきれる

上の条件では NB=216 の場合，P × NB=3024，Q × NB=3888 であり，3024 と 3888 の最小公倍数は 27216．この 27216 の倍数で 160000 近辺の値である 163296 から 1 を引いた値の 163295 が N の最適な数字となる．

Table 5　N 以外のパラメータ

NB	216 or 224
(P,Q)	(14,18)
BCAST	1ring

4.1.4　最適パラメータ

以上の 3 種類の計測結果から得られた最適なパラメータを Table 4 に示す．またそのパラメータを用いて計測を行った結果を Fig. 4 に示す．この計測より，ピーク性能値の 62.6 %である 1.009TFlops という高い実行性能値が得られた．

Table 6　N の検討

N	NB	GFlops
160000	224	999.5
165000	224	998.0
170000	224	960.5
175000	224	838.3
180000	224	893.7
169343	224	939.5
163295	216	1005.0

Table 7　最適パラメータ

N	163295
NB	216
BCAST	1ringM

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WR01L2L2      163295   216    14    18            2878.26              1.009e+03
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) =        0.0012819 ..... PASSED
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) =        0.0017404 ..... PASSED
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) =        0.0002993 ..... PASSED
================================================================================
```

Fig. 4　最高記録

4.2　姫野 Benchmark

CPU 数 16，配列サイズ M の場合において，グリッド分割およびコンパイルオプションの検討を行った．グリッド分割 15 パターン，コンパイルオプション 5 通りの全 75 パターンについて Table 8 にまとめた．

4.3　NAS Parallel Benchmarks

今回は全 8 種類のうちの 3 種類の計測を行った．それらの問題サイズ A，B，C クラスにおける結果を Table 9 から Table 11 に示す．

5　まとめ

TOP500 へのランクインを目指して HPL のパラメータチューニングを行ったが，最終的には 1TFlops の大台を達成することができ，24th TOP500 List でのランクインは確実と思われる．しかし，今回計測に用いた Core Culster System の設置されていた環境は決していいものではなかった．十分な熱対策が施されておらず，Fig. 5 に示すように強制的に排熱処理を行いながらの計測となった．そのため，室温が上がりすぎると計測を中断しなければならず，思い通りの計測にはならなかった．また，実行性能にも何らかの影響がでていたものと思われる．大規模になるほどクラスタの性能は向上するが，そ

Table 8　グリッド分割およびコンパイルオプション

	コンパイルオプション				
グリッド	none	-O1	-O2	-O3	-Os
(1,1,16)	1034	3281	3384	3440	3213
(1,2,8)	939	3475	3515	3608	3448
(1,4,4)	937	3438	3555	3622	3388
(1,8,2)	936	3472	3529	3573	3409
(1,16,1)	987	3610	3450	3564	3439
(2,1,8)	871	3436	3618	3633	3458
(2,2,4)	812	2635	2769	2854	2702
(2,4,2)	829	2735	2665	2778	2697
(2,8,1)	860	3731	3697	3772	3630
(4,1,4)	868	3387	3580	3622	3427
(4,2,2)	912	3562	3648	3654	3441
(4,4,1)	852	3655	3705	3737	3554
(8,1,2)	889	3211	3244	3231	3071
(8,2,1)	844	3250	3291	3324	3267
(16,1,1)	869	3085	3101	3106	3063

Table 9　IS Benchmark

	問題サイズ			
# CPU	A	B	C	W
1	32.98	32.80	32.55	35.13
2	49.04	50.59	50.00	52.02
4	75.95	68.89	53.42	79.69
8	101.61	95.51	90.63	110.76
16	159.44	153.13	139.81	195.90
32	287.42	219.82	201.46	276.35
64	422.18	283.40	249.90	205.47
128	632.02	607.81	385.42	159.41

#:number of

Table 10　LU Benchmark

	問題サイズ			
# CPU	A	B	C	W
1	388.92	354.01	359.37	384.92
2	603.27	317.21	600.83	801.74
4	1244.66	525.35	1169.19	1604.00
8	3047.43	1385.41	2403.01	2892.00
16	6133.82	4683.49	4637.73	5791.91
32	11090.02	10841.29	8770.14	9467.14
64	19991.97	21556.12	20340.41	13965.90
128	33352.47	36687.46	1838.89	計測不能

#:number of

Table 11　SP Benchmark

	問題サイズ			
# CPU	A	B	C	W
1	181.04	178.08	159.35	176.10
4	163.17	533.10	556.84	505.72
9	1104.18	1135.00	1243.08	1289.77
16	1969.53	1992.36	2113.65	2277.31
25	3542.31	2995.69	3167.11	3103.60
36	5034.43	4360.71	2950.02	4054.90
49	6413.99	6030.25	5891.33	4731.64
64	7606.61	8626.71	7543.84	5927.08
81	9263.31	8923.23	9436.31	5939.40
100	10963.29	12777.29	11879.40	7245.41
121	12371.82	15131.19	14707.35	7233.62
144	12921.37	17353.11	17049.27	7242.43
169	15008.57	19928.92	21916.26	計測不能
196	15869.96	20971.86	22325.98	計測不能
225	16493.20	23724.19	29335.80	計測不能

#:number of

の反面メンテナンスに多くのコストがかかる．そのバランスをうまくとった運営・管理計画が欠かせない．

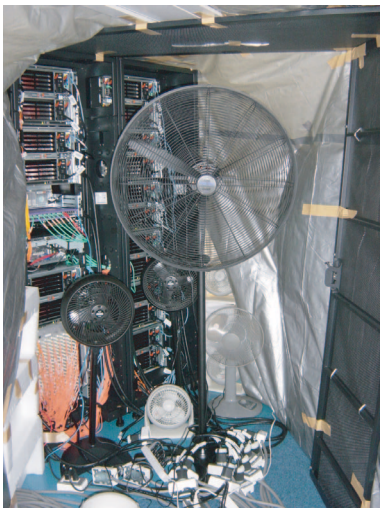


Fig. 5　排熱のための扇風機群

参考文献

1) TOP500 SUPERCOMPUTER SITES
http://www.top500.org/

2) The Netlib
http://www.netlib.org/

3) 理化学研究所・情報基盤センター
http://accc.riken.jp/