Supernova クラスタにおける HPL Benchmark の計測

Measurement of HPL Benchmark in Supernova Cluster System

荒久田 博士
Hiroshi ARAKUTA

Abstract: HPL is one of the implemention of Linpack Benchmark for distributed-memory parallel computer. It has very numerous parameter, and the optimal parameter depends on the composition of the system. Because of it, to get the optimal parameter is very difficult. In order to get the optimal parameter, we execute parameter tuning in Supernova Cluster System. This paper shows process of parameter tuning to find the results of optimal parameter in Supernova Cluster System.

1　はじめに

近年，超並列計算機に代わりパーソナルコンピュータやワークステーションといった一般に利用されているコンピュータをネットワークで繋ぎ，1 つの計算機として利用する PC クラスタシステムが注目されている．Myricom 社 [1] の Myrinet のようにスループットが 1Gbps を超え，従来の超並列計算機と同等の低レイテンシを実現するネットワークを利用して，大規模なクラスタも構築されている．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 Supercomputer Sites[2] においても，クラスタシステムはスーパーコンピュータにひけを取らない性能を見せている．TOP500 にランクインすることは，所有機関にとって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．そのため，ハイパフォーマンスコンピュータのユーザやベンダ，大規模計算機センターにとって大きな興味の対象であり，TOP500 はこの種のリストの中で最大のものとなっている．TOP500 へのランクインを目指し，本研究室では Fig. 1 に示す Supernova Cluster System を導入した．

Fig. 1　Supernova Cluster System

2　Supernova Cluster System

Supernova は，AMD 社 [3] の 64bit CPU である Opteron プロセッサ 512CPU により構築されている大規模なクラスタシステムである．主なハードウェア構成，ネットワーク構成は Table 1 のとおりである．

Table 1　Supernova のハードウェア構成

#node	256
CPU	AMD Opteron 1.8GHz × 2
L1/L2 キャッシュ	128 KB / 1 GB
Chipset	AMD 8131+8111
Memory	PC2700 Registered ECC 2GB
OS	Turbolinux 8 for AMD64
通信ライブラリ	mpich-1.2.5 build by gcc3.2
通信プロトコル	TCP/IP
通信媒体	Gigabit Ethernet

#：number of

Opteron プロセッサでは統合メモリ・コントローラ，HyperTransport リンクという特徴を備えている．統合メモリ・コントローラはメモリのボトルネックを解消し，HyperTransport リンクは I/O ボトルネックの解消・低減，バンド幅の向上とレイテンシの低減によるシステム全体の性能向上を可能とする．Opteron 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）を有しており，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．このことより Supernova のピーク性能値（Rpeak）は式 (1) より 1.8432TFlops となる．

$$Rpeak = \#CPU \times ClockFrequency \times \#FPU \quad (1)$$

また，Supernova ではノード間通信を行うためのスイッチとして Force10 Networks 社 [4] の E1200 を使用している．E1200 は，1.44 Tbps のバックプレーンを持ち，毎秒 5 億パケットの処理速度を有する．E1200 の使用により，スケーラブルな性能向上を期待することが出来る．

3 HPL

HPL[5]（High-Performance LINPACK Benchmark）は，LINPACK ベンチマーク実装の一つである．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一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を行うことができ，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．

3.1 アルゴリズム

HPL ではまず，Fig. 2 のように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ブロックサイクリックに並べ，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，Fig. 3 のように Panel Factorization，Panel Broadcast 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．

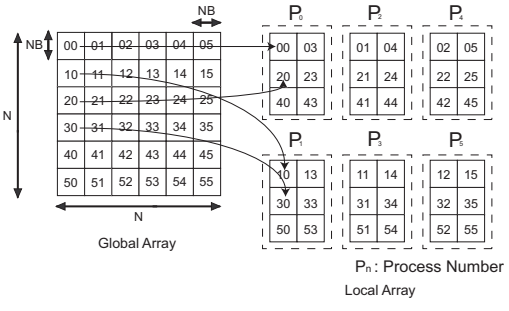


Fig. 2　ブロックサイクリック分割

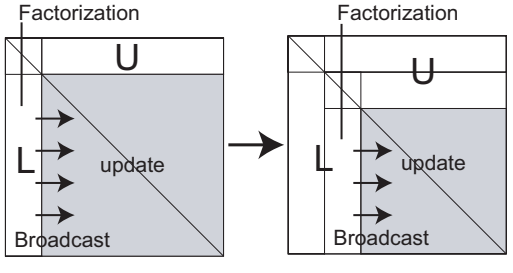


Fig. 3　更新計算

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 [6] 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 HPL の主要パラメータ

LINPACK においてシステムの最大実行性能を得るためにはシステムの特性にあった最適なパラメータを設定する必要がある．前節で述べたパラメータの内，計測値に大きく影響を及ぼすものについて調査を行い，HPL の最適なパラメータを検討した．HPL の主要パラメータについて概説する．

4.1 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．一般的に N の値が大きくなる程良い結果を得られるが，N の増加に伴いメモリの使用量は増加する．

4.2 ブロックサイズ NB

ブロックサイズ NB は，HPL で解く問題の粒度である．NB が大きくなると通信量は減少する一方，ロードバランスが悪くなる．逆に，NB が小さくなると通信量は増加する一方，ロードバランスは良くなる．また良好な結果を示す NB の値があれば，その値の整数倍も良好な結果を示すことがある．

4.3 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ものである．P と Q の積が実行ノード数となる．一般的に，P，Q の値は等しい，もしくは P の値より Q の値が大きい方が良い結果が得られる．

4.4 Panel Broadcast のトポロジー

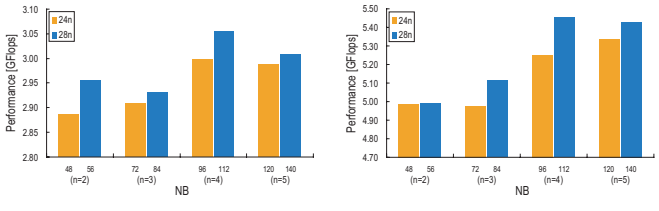
Panel Broadcast のトポロジーには Increasing-1ring，Increasing-2ring，Bandwidth-reducing の 3 種類と，次の Panel Factorization を行うプロセスにメッセージ送信をさせない modified 版がそれぞれ 3 種類の計 6 種類が存在する．normal 版と modified 版のトポロジーの流れは，次のとおりである．

- normal 版
  メッセージ受信 → メッセージ送信
  → Update → Panel Factorization
- modified 版
  メッセージ受信
  → Update → Panel Factorization

5　パラメータの検討

5.1　ブロックサイズ NB 値の検討

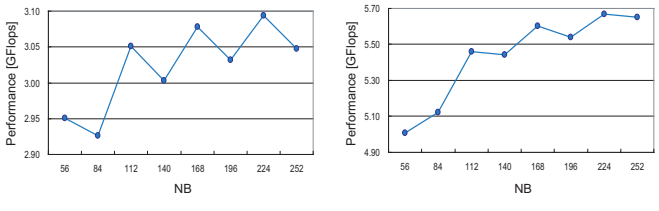
NB は通常，HPL のパラメータを決定する際に最も設定が困難であるとされている．最適な NB の値を求めるため，ATLAS[7] (Automatically Tuned LinearAlgebra Software) が CPU のキャッシュサイズを認識する際に導き出した値より得られた 24 の倍数と 28 の倍数に関する計測を行った．NB 以外の主なパラメータは N:10000, BCAST:1ring であり，用いたコンパイラは gcc3.2，最適化オプションは- formit-frame-pointer -O3 -funroll-loops，演算ライブラリは atlas-3.5.6 である．結果を Fig. 4 に示す．



(a) 1CPU 使用時　　(b) 2CPU 使用時

Fig. 4　NB 値の検討

Fig. 4 の結果より，NB は 28 の倍数が良好な結果を示していることが分かる．28 の倍数に関する計測を続けて行い，最適な NB 値の検討を行った．計測に用いたた NB 以外の主なパラメータ，コンパイラ，最適化オプションは先と同様である．結果を Fig. 5 に示す．



(a) 1CPU 使用時　　(b) 2CPU 使用時

Fig. 5　最良 NB 値の検討

Fig. 5 より，最適な NB の値は 224 であることが分かる．また，良好な結果が得られる NB は 28 の倍数ではなく 56 の倍数であると考えられる．

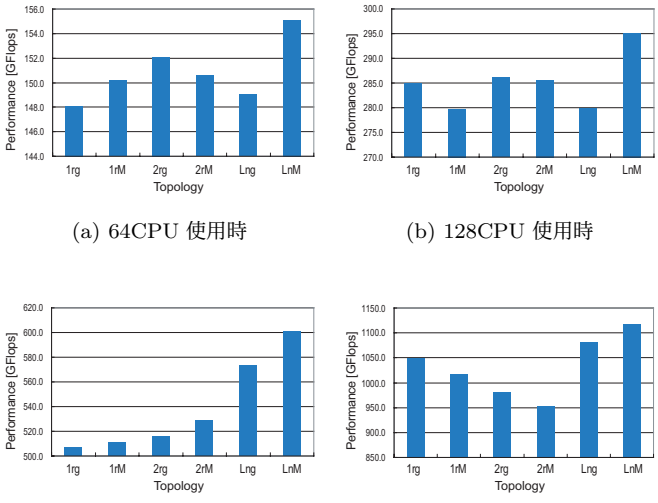
5.2　Panel Broadcast のトポロジーによる比較

Panel Broadcast の各トポロジーに関して計測を行った結果を Fig. 6 に示す．計測の際に利用した BCAST 以外の主なパラメータを Table 2 に示す．用いたコンパイラは gcc3.2，最適化オプションは-fomit-frame-pointer -O3 -funroll-loops，演算ライブラリは atlas-3.5.6 である．

Fig. 6 より，BCAST には使用 CPU 数の違いにより様々な傾向があることが分かる．本来，normal 版と modified 版のトポロジーではメッセージ送信が無い分 modified 版を利用した場合が normal 版に比べ良好な結果を示すはずである．しかし Fig. 6 を見ると，normal 版が modified 版より良好な結果を示しているものがある．

Table 2　BCAST 検証の際に用いたパラメータ

	64cpu	128cpu	256cpu	512cpu
N	80000	110000	160000	220000
NB	224			
(P，Q)	(8，8)	(8，16)	(16，16)	(16，32)



(a) 64CPU 使用時　　(b) 128CPU 使用時

(c) 256CPU 使用時　　(d) 512CPU 使用時

Fig. 6　Panel Broadcast の検証

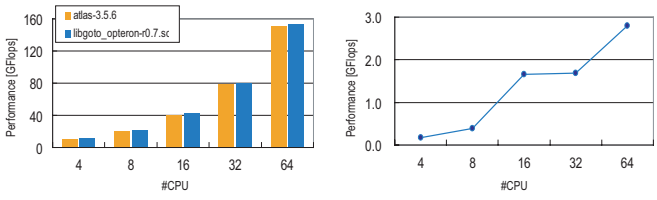
様々な傾向を示しているが共通して言えることは，使用 CPU 数の違いに関わらず，Supernova において BCAST は Long（bandwidth reducing modified）が最も良い性能を出すということである．Long 方式は，一度の通信で大きなメッセージを送信するのに適しており，Supernova のようにノードの処理速度が速く，比較的ネットワークが遅い環境に適している．

5.3　使用ライブラリによる比較

HPL 実行の際に使用するライブラリとして，ATLAS を用いた場合と goto-library[8] を利用した場合の計測結果を Fig. 7 に示す．計測の際に用いた主なパラメータは，Table 3 のとおりである．用いたコンパイラは gcc3.2，最適化オプションは-fomit-frame-pointer -O3-funroll-loops である．

Table 3　library による検証の際に用いたパラメータ

	4cpu	8cpu	16cpu	32cpu	64cpu
N	20000	28000	40000	56000	80000
NB	224				
(P，Q)	(2，2)	(2，4)	(4，4)	(4，8)	(8，8)
library	atlas-3.5.6，libgoto_opteron-r0.7.so				
BCAST	Increasing-1ring				



(a) 異なる library 使用時　　(b) 性能誤差の推移

Fig. 7　使用 library の検討

Fig. 7 より，goto-library が優れた結果を示しており，使用するプロセッサ数が増えると goto-library がより効果的であると考えられる．

5.4　問題サイズ N の検討

N の値は大きくなる程良好な結果を得られ，HPL の結果に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の 80%以上を使用するよう設定する．この設定を目標に求めた値は，226274 となる．しかし N を 226000 で実行した際，計算機への負荷が高くなりすぎ，結果が出る前にプロセスが停止してしまうことが分かった．そこで N を 200000 より 5000 ずつ増加させ，計測を行った．計測の際に利用した N 以外の主なパラメータは，NB：224，(P，Q)：(16，32)，BCAST：LnM である．用いたコンパイラは gcc3.2，最適化オプションは- fomit-frame-pointer -O3 -funroll-loops，演算ライブラリは goto-library である．

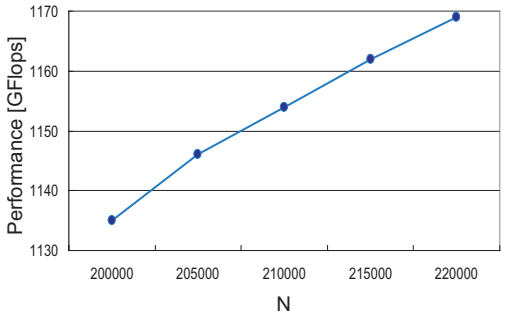


Fig. 8　HPL 最高値の検討

Fig. 8 より，問題サイズが 220000 で最も良好な結果を示していることが分かる．220000 を超える問題サイズではスワップを生じ，実行完了前にプロセスが停止してしまった．これ以上問題サイズを大きくしても，良好な結果は得ることは出来ないと考えられる．これまでの計測により得られた最適なパラメータは Table 4 となる．

Table 4　Supernova における主なパラメータの最適値

N	220000
NB	224
(P，Q)	(16，32)
BCAST	Long（bandwidth reducing modified）
library	libgoto_opteron-r0.7.so

6　まとめ

5 節で最適なパラメータに関する検討を行った．検討により得られた最適なパラメータを用い数回の計測を行った結果，Fig. 9 に示す 1.169 TFlops という値を最高値として計測することが出来た．これは，ピーク性能値の約 63.4%である．これらの検討を通じて，HPL のパラメータチューニングに関する様々な知見を得ることが出来た．1.169TFlops という結果で Supernova は，Fig. 10 に示すとおり 2003 年 11 月度の TOP500 において 93 位にランクインした．またこの結果は国内では 6 位，国内の PC クラスタとしては 1 位である．これらの結果より，Supernova は国内だけでなく世界に誇ることの出来る高速計算機であると言える．

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
W15R2C4       220000   224    16    32            6072.75              1.169e+03
--------------------------------------------------------------------------------
||Ax-b||_oo / ( eps * ||A||_1  * N        ) =        7.3316029 ..... PASSED
||Ax-b||_oo / ( eps * ||A||_1  * ||x||_1  ) =        2.9053690 ..... PASSED
||Ax-b||_oo / ( eps * ||A||_oo * ||x||_oo ) =        0.4970550 ..... PASSED
================================================================================
```

Fig. 9　最高記録

Rank	Site Country/Year	Computer / Processors Manufacturer	Rmax Rpeak
91	CSC (Center for Scientific Computing) Finland/2002	**pSeries 690 1.1GHz** / 512 IBM	1170 2253
92	Florida State University United States/2002	**pSeries 690 1.1GHz** / 512 IBM	1170 2253
93	Doshisha University, Intelligent Systems Design Laboratory Japan/2003	**Opteron 1.8 GHz, Gig Ethernet** / 512 Visual Technology	1169 1843.2
94	**Government** United States/2001	**T3E1200** / 1900 Cray Inc.	1166 2280
95	ERDC MSRC United States/1999	**T3E1200/900** / 1792 Cray Inc.	1166 1999
96	**Semiconductor Company (C)** United States/2003	**xSeries Xeon 2.8 GHz, Gig-Ethernet** / 602 IBM	1146.35 3371.2

Fig. 10　2003 年度 11 月度の Top500

参考文献

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

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

3) Advanced Micro Devices Home Page
http://www.amd.com/us-en/

4) Force10 Networks Home Page
http://www.force10networks.com/

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

6) 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

7) Automatically Tuned Linear Algebra Software
http://math-atlas.sourceforge.net/

8) High-Performance BLAS by Kazushige Goto
http://www.cs.utexas.edu/users/flame/goto/