Cambria 及び Gregor での Linpack の実行結果

Linpack result on Cambria and Gregor and Gregoria

下神納木 淳，松山 靖彦
Atsushi SHIMOKOUNOKI , Yasuhiko MATSUYAMA

Abstract:
This document shows the process of Cluster-Group executing The Linpack Benchmark to rank in TOP500. This time,we mesured three Cluster-system named Cambria,Gregor and Gregoria(Cambria and Gregor).There are many troubles in running the three Cluster-system.We wish that Gregoria will rank in TOP500.

1 はじめに

TOP500 ランクインを目指し，Cambria と Gregor を連結し，Linpack を計測した．最終的に，78.62Gflops という結果がでた[1]．計測にあたって以下のような問題点があった．

1. LAM で 225 台以上起動しない．
2. メモリ増設による不具合
3. Gregor での不具合．
4. 全マシンで動かした時に不安定になってしまう．

2 TOP500 について

TOP500[2]は，世界のパフォーマンスの高いコンピュータのリストである．パフォーマンスの高いコンピュータを有する組織や設置場所，システムに使用されているソフトウェアについての情報などを提供するために作成されている[3]．TOP500 は，1993 年 6 月から作成され，毎年 2 回ずつ更新されてきた．今回エントリしたのは，第 18 回である．

3 Linpack Benchmark について

Linpack Benchmark(以下 Linpack) は，米国のテネシー大学の J.Dongarra 博士によって開発された LU 分解にもとづく連立一次方程式の解法プログラムである．TOP500 では，ベストな性能に達するために，ユーザが問題サイズを決めたり，コンパイラ等のソフトウェアを最大限に利用することができる．問題が規則的なので，結果として得られる性能はかなり高い．Linpack は，実行結果に CPU，メモリ，そしてネットワークの性能が反映される総合ベンチマークである．[4]

[1]エントリ時のシステム名は Gregor.
[2]http://www.top500.org
[3]http://www.top500.org/lists/intro.html
[4]http://www.top500.org/lists/linpack.html 参照

4 Linpack のパラメータ

Linpack 計測結果を報告する前に，Linpack におけるパラメータについて述べる．

Linpack には 17 のパラメータが存在する．それらの中で，計測結果に大きく影響を及ぼすものについて説明する．

4.1 問題サイズ (N)

問題サイズ (N) は，Linpack で解く問題の大きさである．つまり，Linpack では N 元連立方程式を解くことになる．大体において，N が大きくなるほどよい結果が得られるが，N が大きくなるほどメモリ使用量が増える．最適な結果をもたらす N の値について，http://www.netlib.org[5]に以下の記述がある．

1. N は，メモリの約 80%を使用するように設定すると良い結果が得られる．
2. 上のような N の値は，以下の式によって求められる．

$$N = \sqrt{TotalMemory * 0.8/8}$$

上の式から計算した N の値のいくつかを Table 1 に示す．

Table 1　N とメモリ使用量の関係

ノード数	1 ノードあたりのメモリ容量 (MB)	N
16	128	約 14310
128	256	約 57243
256	256	約 80954
384	256	約 99148

[5]Linpack プログラム HPL はここから入手した．

4.2 ブロックサイズ (NB)

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，よい結果が得られる[6]．

4.3 プロセスグリッド (P と Q)

プロセスグリッド (P と 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 が大きい方がよい[7]．

5 Cambria

現時点 (2001 年 9 月) での Cambria の主な仕様を Table 2 に示す．

Table 2　Cambria の仕様

CPU	PentiumIII 800MHz 256 プロセッサ
Memory	256MB (計 65.536GB)
Network	Fast Ethernet
O.S	Debian/GNU Linux
Kernel Version	2.4.9
Communication	lam-6.6b1

なお，以下に述べる計測結果は，2001 年 8 月〜10 月に渡って計測した結果である．その間，システムにいくつかの変更がなされた．8 月と 9 月でのシステムの違いを以下に述べる．

1. lam-6.5.2 から lam-6.6b1 になる (256 ノードで実行できるようになる)．
2. SSE に対応 (同じパラメータでの計測結果が上る．)．
3. 1 ノードあたりのメモリが，128MB から 256MB に増設された (メモリ使用量と関係のあるパラメータ N を大きくできる)．

このように変更がなされたので，計測結果には，計測した月を併記しておく．

5.1 計測結果 (Cambria)

Cambria での Linpack 実行結果を通して，Linpack のパラメータが結果に与える影響について示す．

[6]http://www.netlib.org 参照
[7]http://www.netlib.org 参照

5.1.1 N を変化させる

N 以外のパラメータを Table 3 のように固定し，N を変化させて，計測結果の変化を調べた．計測結果を Fig. 1 に示す．

Table 3　パラメータ

NB	32
P	15
Q	15
etc	W00L2L2

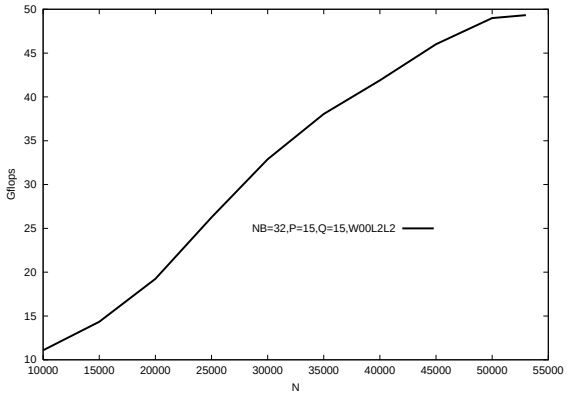

Fig. 1　N による計測結果の変化 (8 月)．

Fig. 1 を見る限り，N が大きくなる程結果がよくなっている．Fig. 1 は，N=53000 までしか計測していない．これ以上 N を大きくすると，メモリ不足 (8 月時点では，1 ノードあたり 128MB) となり，実行できない[8]．N=50000 以上になるとグラフの傾きが小さくなっているが，ピーク性能に達したとは言いがたい．つまり，メモリ容量がネックとなって，性能が頭打ちになっている可能性がある．

5.1.2 スワップの影響

Fig. 1 では，メモリ容量が，ネックとなっている可能性がある．Cambria の大半のノードはディスクレスなので，スワップできないが，もしスワップが起きた場合，実行結果にどのような影響があるのかを計測した．Cambria の 16 のノードにはハードディスクがある．その 16 ノードを使用して，計測を行った．パラメータは Table 4 の通りである．計測結果を Fig. 2 に示す．

16 ノードで実行する場合，N=15000 になったところでスワップが起きる．Fig. 2 をみると分かる通り，計測結果は大きく落ち込んでいる．

[8]Cambria のノードの大半はディスクレスなのでスワップできない

Table 4　パラメータ

NB	32
P	4
Q	4
etc	W00L2L2

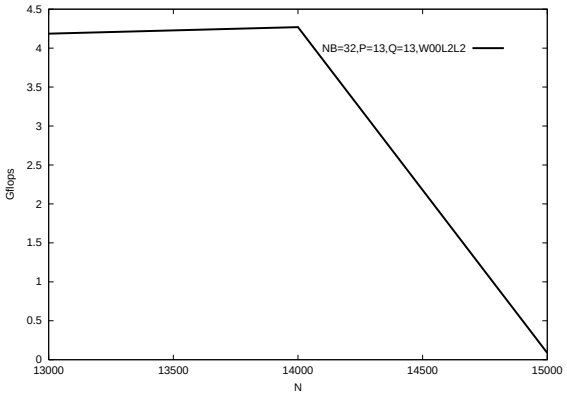

Fig. 2　スワップの影響 (8 月).

5.1.3　NB を変化させる

NB 以外のパラメータを Table 5 に固定し，N を変化させて，計測結果の変化を調べた．計測結果を Fig. 3 に示す.

Table 5　パラメータ

N	20000
P	12
Q	12
etc	W00L2L2

Fig. 3 から，NB は大きければ (もしくは小さければ) よいと言うわけではなく，ロードバランスと通信量のバランスを保った値に設定するとよい結果が得られる.

6　Gregor

6.1　仕様

2001 年 9 月時点における Gregor の主な仕様 Table 6 を以下に示す.

9 月の中頃から Cambria と Gregor が連結された．そして，そのシステムのネットワークは **Ethernet** となった．その際に Gregor の測定結果がかなり下がるという問題が出てきた．よって，本章では **Myrinet2000** と **Ethernet** における Linpack 実行結果を比較する.

また，16 ノードのそれぞれのノードを 1CPU で動かした結果と 8 ノードで 2CPU を動かした結果を比較し，NIC を介して計算を行っているのかを確かめる.

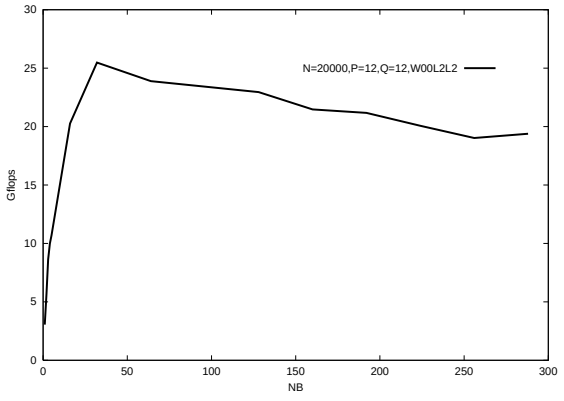

Fig. 3　NB による計測結果の変化 (10 月).

Table 6　Gregor の仕様

CPU	PentiumIII 1GHz 128CPU （64 ノードデュアルプロセッサ）
Memory	512MB (計 32.768GB)
Network	Myrinet2000
OS	Kondara HPC（Red Hat 系）
Compiler	gcc ，g77
Communication	MPICH1.2/GM1.3
Peak 性能	128Gflops

6.2　計測結果

6.2.1　Myrinet2000 と Ethernet の比較

60 ノード（120CPU）において，Table 7 に示すパラメータを使用し， **Myrinet2000** と **Ethernet** のそれぞれで測定を行った.

測定方法としては 10 回試行を行い，測定結果としては一番良い結果とした．パラメータおよび測定結果を Table 7 に示す.

Table 7　60 ノード（120CPU）におけるパラメータおよび測定結果

N	20000
NB	32
P	10
Q	12
etc	W00L2L2
Myrinet2000	43.80Gflops
Ethernet	15.60Gflops

Table 7 を見ると，明らかに **Myrinet2000** が速いことがわかる．この測定結果を踏まえると，Linpack ベン

チマークのプログラムはネットワークに依存するということが考えられる．

6.2.2　デュアルプロセッサ間の通信

今回の Linpack ベンチマーク測定のために **LAM** をデュアルプロセッサ対応にコンパイルした．しかし，ノード内の通信がどのように行われているのかが明確でない．よって，デュアルプロセッサで起動した場合にメモリ上で通信を行っているのか，それとも NIC を介して通信を行っているのかについて考える．今回は，**8 ノード 16CPU** と **16 ノード 16CPU** の場合で比較することにする．なお，計測にはネットワークに **Ethernet** を使用する．また，測定方法は 10 回試行とし，測定結果はその中で一番良い数値とした．パラメータおよび測定結果を Table 8 に示す．

Table 8　16CPU におけるパラメータおよび測定結果

N	10000
NB	32
P	4
Q	4
etc	W00L2L2
8 ノード 16CPU	4.077Gflops
16 ノード 16CPU	4.862Gflops

Table 8 の測定結果を比べると，ほぼ同じ測定結果となっていることがわかる．これは，おそらく NIC を介して通信を行っているからであると考えられる．また，デュアルプロセッサでの実行結果 16 ノード 16CPU での測定結果の約 84%となっているのは，NIC を介して同じメモリで測定を行っていることが原因であると考えられる．

このようなことから，Linpack ベンチマークでは Ethernet より通信帯域の広い Myrinet2000 を使用する方がよいとわかる．

6.2.3　参考資料

今回，TOP500 に向けての測定を行う前に，参考資料にするために Myrinet を使用して 1 ノード 2CPU の Peak 値を測定した．その測定結果を以下に示す．

1.N を変化させる N 以外のパラメータを以下のように固定し，N だけを変化させて計測結果の変化を調べた．パラメータを Table 9 に示す．また測定結果を Fig. 4 に示す．

Fig. 4 を見ると，N=7500 前後で極端に性能が落ちている．この性能が落ちている原因は **スワップ** であると考えられる．スワップしていると分かった理由は，計測中にコマンド **top** で調べた時にスワップしていること

Table 9　N を変化させたパラメータ

NB	32
P	1
Q	2
etc	W00L2L2

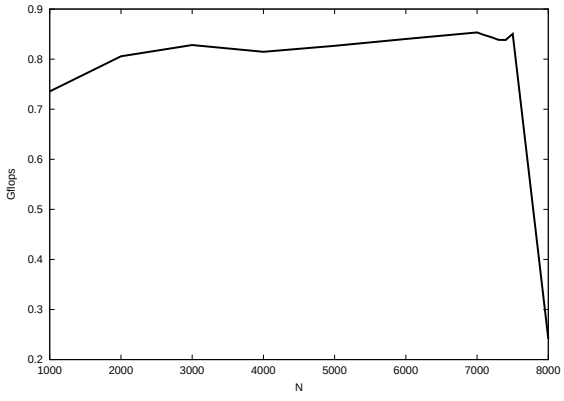

Fig. 4　1 ノード 2CPU における測定結果

を発見したからである．また，スワップが起こると極端に計測時間も増えることがわかった．

Fig. 4 から考えると、全体的にあまり性能が上がっていないことから N=7000 前後が Peak であると考えられるので，メモリを増設して N の値を増加させてもあまり良い結果は期待できないと考えられる．

2.NB を変化させる

NB 以外のパラメータを以下のように固定し，NB だけを変化させて計測結果の変化を調べた．尚，N の値は上記において一番良かった値を使用した．パラメータを Table 10 に示す．また測定結果を Fig. 5 に示す．

Table 10　NB を変化させたパラメータ

N	7000
P	1
Q	2
etc	W00L2L2

Fig. 5 を見ると NB=208 あたりが MAX 値となった．以上 N と NB の値を変化させた結果の MAX 値は **1.185 Gflops** とわかった．

また，パラメータの一つである **etc** を適当に変化させて測定した結果，Table 11 パラメータで **1.187 Gflops** という一番良い計測結果が出た．しかし，このパラメータはどのように変化させたらよいかということはまだ

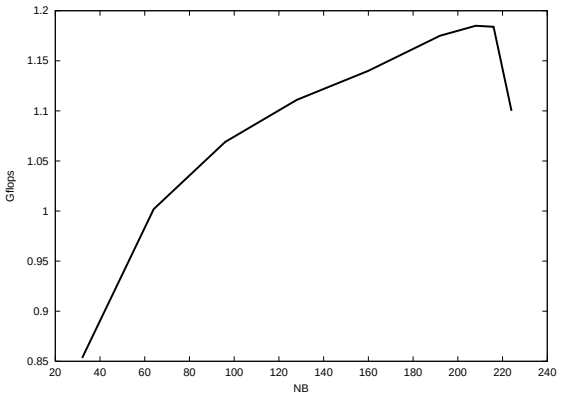

Fig. 5　1 ノード 2CPU における測定結果

はっきりしていない.

Table 11　適当なパラメータにおける測定結果

N	7000
NB	208
P	1
Q	2
etc	W00C2C8

6.3　今後の問題点

1. 確実に LAM がデュアルプロセッサで対応しているのか.
2. デュアルプロセッサにおける通信問題の改善
3. 大多数のマシンにおいて大きな問題サイズで実行する（通信量の増加）と測定結果に [**FAILED**] となってしまう．これはなぜか.

7　Gregoria

Gregoria は Cambria と Gregor を連結したシステムである．Cambria に関して大きな変更は無いが，Gregor に関してはネットワークが **Myrinet** から **Ethernet** に変更された.

7.1　計測結果

Cambria+Gregor を Top500 にエントリした (システム名は Gregor)．このシステムで，TOP500 ランクインを目指し計測した．最終的に，以下のような結果になった.

Linpack では，実行の最後に計算結果の答え合わせをする．残念ながら，Table 12 の 84.58Gflops という値は，計算結果が間違っており (FAILED)，TOP500 にエントリできる値ではない．エントリした値は，Table 13 の 78.62 Gflops である.

Table 12　パラメータと実行結果

N	95000
NB	64
P	16
Q	24
etc	W03L2L2
	84.58 Gflops(FAILED)

Table 13　パラメータと実行結果

N	95000
NB	64
P	16
Q	24
etc	W00L2L2
	78.62 Gflops(PASSED)

7.2　まとめ

TOP500 へのエントリは，最終的に 78.62 Gflops となった．この値は，前回の TOP500 では，432 位の値である．当初の目標は 100Gflops であったが，遠く及ばなかった．その原因として考えられるものを以下に挙げる.

1. 全てのパラメータを調整したわけではない.
2. Gregor の性能が伸びなかった.
3. よい結果がでても計算間違いだった.

これらの問題を解決しつつ，今後は，Cambria 単体での Linpack の新記録を出したい.

また，計測中にノードが落ちるということがしばしばあり，計測時間を消費した．クラスタの管理を快適にする重要性を身を持って感じた.

8　謝辞

今回の Linpack 計測にあたり，研究室の皆様には，クラスタのメモリを借して頂くなどの御協力を頂きました．最後になりましたが，厚く御礼申し上げます.