

PC クラスタは Poor Man ' s Supercomputer

姫野 龍太郎、理化学研究所、〒351-0198 埼玉県和光市広沢 2-1、E-mail : himeno@postman.riken.go.jp
重谷 隆之、理化学研究所、〒351-0198 埼玉県和光市広沢 2-1、E-mail : shige@postman.riken.go.jp
黒川 原佳、北陸先端大院、〒923-1292 石川県能美郡辰口町旭台 1-1、E-mail : kurokawa@jaist.ac.jp

Ryutaro Himeno, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama, 351-0198 JAPAN
Takayuki Shigetani, RIKEN, 2-1 Hirosawa, Wako-shi, Saitama, 351-0198 JAPAN
Motoyoshi Kurokawa, JAIST, 1-1 Asahidai, Tatsunokuchi-cho, Ishikawa, 923-1292 JAPAN

要約

PC クラスタは従来の一部マニアのものではなく、簡単に導入できる価格性能比の良い計算機プラットフォームになってきた。最大の問題である MPI による並列化も、小規模なシステムであれば、かなり簡単に性能を引き出すことができる。ここでは小規模なクラスタの導入に関わる手間・コストを示すとともに、姫野 BMT を使って書き直しに必要な手間と性能の関係を調べ、報告する。また、更なる性能を目指したときにボトルネックとなる問題点とその解決策も挙げる。

Key words : PC クラスタ, 並列化, HPF, MPI, himenoBMT

1. はじめに

最近、PC(Personal Computer)やそれを取り巻くデバイス等のハードウェア分野では高性能化、低価格化が急速に進んでいる。そして、ソフトウェアについても OS(Operating System)や開発ツール、ライブラリ等のオープンソース化によって、ソフトウェア利用における金銭的なコストを考慮する必要がほとんどない。これまでも、先端的な研究者や、UNIX や LINUX に詳しい人間のいる研究室で導入している。我々は理研の研究センターとして VPP700E などのスーパーコンピュータを運用し、ユーザーにサービスを行ってきた。それらの HPC 専用のマシンは PC クラスタに比べ、桁違いにコストパフォーマンスが悪い。このままの状況を続けると、一般の研究者・ユーザーとセンター側との意識のずれが拡大し、ユーザーの計算機センター離れが進むのではないかと危惧される。そこで、昨年 100 万円以下という制限を設けてハードウェアを購入し、PC クラスタを構築した。現在は更に 4 台追加して 8 台構成である。これによって、一般の研究者が導入するときの問題点、うまく構成されたときのコストパフォーマンス、HPC 専用機との特性の違いなどを調べた。

PC クラスタになじみのない人間が最初に戸惑うのは、次のような問題であろう。

- ハードウェア、ソフトウェアの選択
- どのようなプログラミングを行えば良いのか
- どの程度の性能が出せるのか
- 他のマシン(商用並列計算機)と比べてどの程度か

そこで、実際に数値流体計算に用いる PC クラスタ環境を構築する場合の導入前後の状況およびベンチマークプログラムによる性能評価の結果を紹介する。

2. ハードウェア・ソフトウェアの選択

ハードウェアの選択の際に条件となるのは、次のようなことである。

- 使用する OS で利用できるデバイスか
- 計算速度(CPU)
- メモリ量
- ネットワーク通信速度
- DISK 容量
- 設置容積(筐体)
- モニタは 1 台だけ

使用する OS 等で利用できるデバイスでなければ、利用に際して苦勞することになるため、一般に広く利用されている物を選択する必要がある。そのため、平凡な機器構成となるように注意す

る。

まず、選択肢が少ないネットワークデバイスを選択する。ネットワーク(通信)は、速ければ速いほど良い。Gigabit Ethernet が徐々に一般的になってきているが、まだコストがかなりかかる。コストと性能のバランスを考えた結果、我々が選択したのはネットワークカードが 100 Base-TX で通信できる製品(今回は、PC に Build-in された物)、ネットワークハブが、100 Base-TX Switching HUB(Non-Intelligent)という構成の製品(Allied Telesis FS708E)になった。次に、拡張性を犠牲にしコンパクトにまとめるか、拡張性を持たせある程度大きくとるかという問題が出てくる。今回は容積を小さくまとめるため拡張性を犠牲にした。最も重要な計算速度(CPU)とメモリの選択肢を考える。この選択は予算との兼ね合いが発生し、多数の選択肢がある問題である。この問題に対処するには、解析する問題のある程度限定し、それに見合った構成にする。例えばメモリが大量に必要ならば CPU の性能を一段落とす事である。また、CPU 性能を落とし、CPU 数を増やすこともできる。この二つのハードウェアの選択がクラスタの構成の最重要部分である。取り扱うデータ量によって影響する DISK 容量についても考慮する必要がある。出力するデータ量に必要な DISK を取り付ける必要があり、並列ファイル I/O を用いれば台数分に比例して出力出来るデータ量が増え、時間も短縮する。

これらの構成を考えた上で時間が許すならば自作するなり、予算内の既製品を購入すれば良い。今回は、設置場所をコンパクトにまとめるため、一体型省スペースの既製品(COMPAQ 製 DeskPro EN SF モデル)を用いた。以下の表 1、2 にハードウェア構成を示す。

Table 1. COMPAQ Hardware Specification

CPU	Pentium II 450MHz
Memory	Master : 256 MB、Slave : 128 MB
Network	100 Base-TX (Intel Pro/100+)
Disk capacity	6.4 GB
Size(W × D × H)	300 × 400 × 80 (mm)

Table 2. Allied Telesis Network Specification

Switching HUB	100Mbps、Full-duplex、8 Port
---------------	----------------------------

次にソフトウェアの選択では、OS、開発ツール及び並列計算用ライブラリは、金銭的なコストをかけないような選択が可能である。今回用いたソフトウェア構成を表 3 に示す。

Table 3. Use of Software

OS	Linux-2.2.5-15
Compiler	egcs 2.91.66 Based gcc & g77
Communication Library	MPICH-1.1.2、LAM/MPI 6.3

3. PC クラスタの価格

上記のクラスタを構成した場合のハードウェア、ソフトウェアの価格を表4示す。

Table 4. Hardware and Software Cost

COMPAQ DeskPro EN SF	150,000 ~ 200,000
Monitor	~ 30,000
Allied Telesis FS708 E	~ 20,000
ネットワークケーブル	?,000
ソフトウェア(OS, compiler, etc.)	Free

PC を4台用いて(PC 4 台の構成で、外部ネットワークとはそのうち1台のみで接続)、PC クラスタを構築する。外部接続されたPCは、サーバ(NFS等)、制御用(Job投入、監視)にも使用し、計算JOBも実行する。また、このPCをMasterとし、その他にPC 4台をSlaveとして用意し、SlaveのみでJOBを実行することで、多少の性能向上が見込まれる。

4. 並列計算

4.1 並列プログラム

現在PCクラスタで用いられている並列プログラムの手法は、言語仕様を変更せず並列計算(データ通信等)に必要な部分をサブルーチンや関数を用いて記述する。並列計算に必要な部分の記述方法としては、MPI(Message Passing Interface)やPMV(Parallel Virtual Machine)という標準API(Application Programming Interface)を用いることが多い。「姫野ベンチマーク」(Fortran 77)をMPIで並列化を行った。MPIとは、並列プログラミングの記述方法等を統一すべくMPIフォーラムで標準化された並列プログラム記述環境である。MPIを用いて記述されたプログラムは、現在稼動しているほとんどの並列計算機で動作する(はずである)。

4.2 姫野ベンチマークの並列化

今回の並列化のコンセプトは以下の通りである。

- (1) とりあえず並列で動作させる
- (2) データ分割はしない、手続き分割のみ
- (3) それなりのパフォーマンスを出す

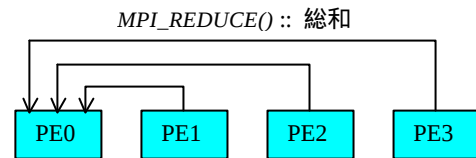
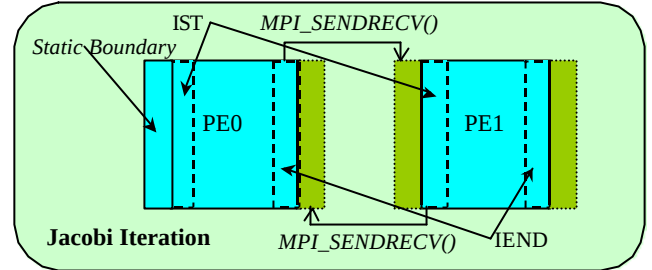
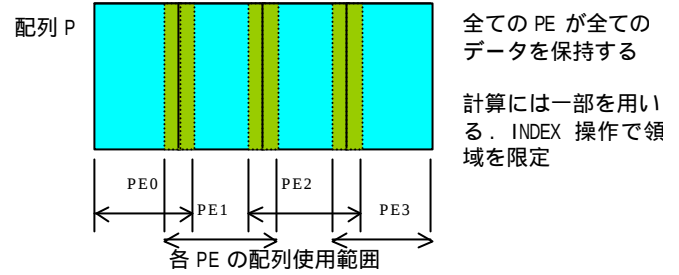
姫野ベンチマークは、3D Poisson 方程式をヤコビ法で解いているため、並列化には領域分割法を用いた。ヤコビ法は、領域ごとに同時に分散計算させても結果に変動はない。分割方法には、最も単純な一次元の分割を用いた。手続き分割のみを行うのでPE(Processing Element : この場合PC 1台)には、無駄なデータ領域を持つことになる。これらの事を行うのに必要なMPIのサブルーチンは以下の通りである。

```

MPI_INIT() :: 並列動作準備
MPI_COMM_SIZE() :: すべての動作 PE 数を把握
MPI_COMM_RANK() :: 動作 PE 中の自分の ID 値を取得
MPI_SENDRECV() :: データ通信(送信と受信)
MPI_BARRIER() :: PE間の同期
MPI_REDUCE() :: 値の縮約演算 (総和、最大、最小、等)
MPI_FINALIZE() :: 並列動作の終了

```

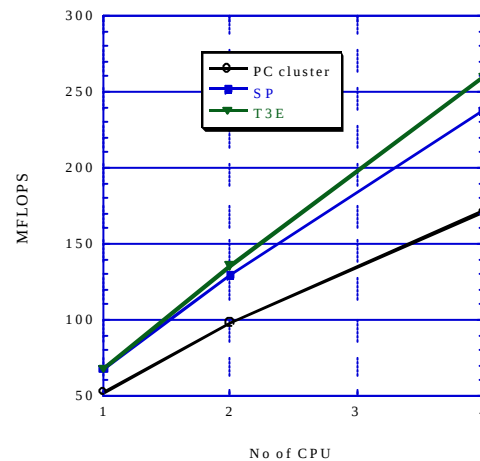
これらサブルーチンの文法と応用に関しては、参考文献(5)を参照して頂きたい。以下にこれらの並列計算の流れを図示したものを示す。ただし、図は2次元の場合とする。



データ分割は、実行 PE 数と MPI_COMM_RANK から得られる ID と配列の大きさ N から、その PE での初めと終わりのインデックスを計算して得る。計算始りのインデックスは $IST = (N/PE) \times ID$ となり、計算終了のインデックスは $IEND = (N/PE) \times ID + (N/PE)$ である。例えば、 $PE=4$ 、 $ID=1$ 、 $N=64$ では、 $IST = (64/4) \times 1 = 16$ 、 $IEND = (64/4) \times 1 + (64/4) = 32$ となる。そして、通信しなければならないデータは、 $IST-1$ と $IEND+1$ のデータである。

5. パフォーマンス

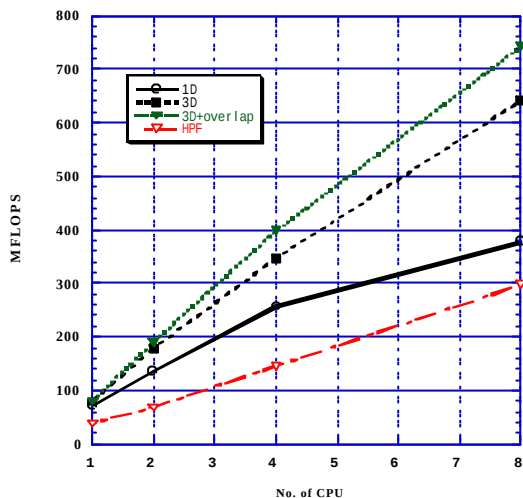
並列化された姫野ベンチマーク(問題サイズ $64 \times 64 \times 128$)をPCクラスタ上で実行した結果を示す。同時に商用計算機の結果を合わせて示す。



6. さらなるパフォーマンスの向上

更にパフォーマンスを向上させるためには何を変更すると効果的か考えた場合

- (1) CPU をより高速な物に換える
- (2) ネットワークを高速な物を用いる(Myrinet)
- (3) 商用コンパイラを用いる (Fujitsu Fortran & C Package)
- (4) ユーザレベル通信を用いる(Score+PM Ver2.4)
- (5) 領域分割を2次元にする
- (6) 台数を増やす
- (7) 通信と計算を同時に行う



結果をまとめると、

富士通製コンパイラはフリーのものに対して、価格に見合う性能向上が認められる
16台までは問題なく、充分速くなる
MPIでのソフトの書き直しは、パターンさえ理解すると簡単で機械的

うまく作ると元のプログラムと差がない

初期・境界条件のサブルーチンに差を凝縮増やしていくと飽和するようになるが、通信の工夫や高速なネットワークにすることで克服可能
領域分割は、分割方法によって速度に違いがある
3次元分割は台数が少ないと効果が低く、むしろ1次元分割が手間の割りに効果が高い

PCクラスタでは通信量より、通信の回数が重要
計算している最中に通信をするよう、計算を分割すると、台数効果が高くなる

一方、問題点としては

性能向上が激しく、6ヶ月後に同じCPUが手に入らない
導入して1年が経ち、2年で更新の必要
自分で構成するには、それ相応の知識(調査)が必要

ターンキーシステムも利用可能だが、その分高い
自分で組むのなら32台まで(できれば16台くらい)が妥当)
64以上だと維持も更新も手間が大変
故障時の予備を持つ必要がある

7. まとめ

小規模なPCクラスタを構築の過程と、その性能をベンチマークプログラムで明らかにした。本報では、既存のコード(問題サイズを含む)を並列化した場合、比較的並列化効果が良い結果が得られた。そのため、パーソナルユース並列計算機としては十分な性能である。また、特別なハードウェアや商用のソフトウェアを用いる、あるいは少しいの工夫で更にパフォーマンスの向上が望める。その結果、ベクトル計算機の1PE程度の性能は十分に得られる。現在、我々の測定で最も速いのはAthlon 1GHzのマシンであるが、姫野BMTで138MFLOPSを実効性能として達成している。理論ピーク性能の何%という計算をしてみれば高い値ではないが、20万円でこの数字はやはり驚嘆すべきものである。300万円出せば実効性能で1GFLOPSを出すことができる。

8. 今後の展望

現在、我々のセンターでも主なジョブは1PEを使う5時間以内の計算である。また、数十%のユーザーがベクトル化率数十%以下で、マシン性能を十分に引き出しているユーザーはほんのわずかである。例えば、もしGAUSSIANなどの市販ソフトがMPIによる並列化とLINUXに対応していれば、全ユーザーの1/3はPCクラスタに移行するのではないかと予想できる。このような状況から今後の計算機センター・ユーザー像を考えると

- 1) PCの数十倍以上の性能を必要とするユーザー
- 2) 分割しても1CPUあたり2GB以上を必要とするユーザー
- 3) アプリケーションソフトを購入できないユーザー
- 4) 計算機のハードやソフトに興味がなく、計算だけに徹したいユーザー

このうち、1)や2)を重視し、3)4)をおざなりにしていると、使いやすさやコストパフォーマンスの点から1)と2)のユーザーは自力で専用のシステムを構築し、逃げて行きかねない。コストの高いHPC専用機は計算機センターとしても今後は選択できなくなってくる。

参考文献

- (1) <http://www-unix.mcs.anl.gov/mpi/mpich/>
- (2) <http://www.mpi.nd.edu/lam/>
- (3) <http://www.beowulf.org/>
- (4) <http://w3cic.riken.go.jp/HPC/HimenoBMT/index.html>
- (5) Gropp, W., Lusk, E. and Skjellum, A.: "Using MPI"; The MIT Press, 1994.

