名古屋大学情報基盤センターの新スーパーコンピュータの利用方法

2009年6月1日

荻野竜樹

名古屋大学太陽地球環境研究所の計算機利用共同研究で利用している名古屋大学情報基盤センターのスーパーコンピュータが2009年5月18日からその一部が稼動を始めました。更に、2009年10月1日には全システムが稼動する予定です。

新システムは、次の3種類の特徴のあるスカラー並列型スーパーコンピュータで構成されています。

3種類のスーパーコンピュータ

S1（SPARC Enterprise M9000）：M9000 3node x 128core

Host name: sp1.cc.nagoya-u.ac.jp

Large shared memory

3 nodes x 128 cores （1 node is FX1's front-end）

Per node: Performance: 1.28TFlops, memory: 1TB

S2およびアプリケーションサーバ：HX600 160node x 4cpu x 4core

Host name: sp2.cc.nagoya-u.ac.jp

Cluster type computer, node uses shared memory

160 nodes x 16 cores

Per node: Performance: 160GFlops, memory: 64GB

S3：FX1 768node x 4 cores

Large distributed memory computer in connection with

Next Generation Supercomputer

Per node: Performance: 40GFlops, memory: 32GB,

Memory bandwidth: 40GB/s

S1は共有メモリの計算機でFX1のフロントエンドプロセッサになっています。従って、S1とS3を利用する場合はS1に、S2を利用する場合はS2にsshでloginします。また、FX1はメモリバンド幅が大きくてプリペタコンに対応します。

ssh a41456a@sp1.cc.nagoya-u.ac.jp

ssh a41456a@sp2.cc.nagoya-u.ac.jp

また、従来使用していた大容量ファイルは
　/large/a41456a/
の配下に移動しています。ここにa41456aは各自のユーザーIDになります。ジョブの実行状態を見るために、新しくjstatコマンドが追加されました。従来のqstatコマンドは同じ様に使えます。ジョブのキャンセルコマンドはqdelです（qdel –k job_number）。更に、データの転送はsftpを使用することになります。sftpは従来使えたftpに比べてファイル転送はかなり遅くなります（STE研からだと約3倍の時間がかかる）。また、データのBinary Format形式はCPUチップにより、M9000とFX1がビッグエンディアンで、HX600がリトルエンディアンなので、それらの計算機をまたがってデータを利用する場合は変換が必要です。
具体的には、新スパコンの利用は後に示す Homepage を見てください。

また、並列計算速度ですが、M9000、HX600、FX1 のいずれでもそれ相応の速度が出ると思います。前の HPC2500 と比較すると、1.5-2.0 倍の速度が得られることもあります。HX600 と FX1 のどちらが速いかは、プログラムによります。多くのプログラムで HX600 の方が速い結果が得られているようです。最速を狙ってプログラムをうまくチューニングすれば FX1 が速い結果も得られていますが、各自のプログラムで調べてください。

(A) M9000 と FX1 の使い方
先ず M9000 に login します。その後、M9000 と FX1 を利用できます。FX1 はバッチジョブのみです。
　　ssh a41456a@sp1.cc.nagoya-u.ac.jp

パラメータを設定する場合、ノード数、コア数、メモリの上限値にも注意してください。
コア当たりの最大利用可能メモリは 7GB です。
FX1 の場合、1ノード4コアで、
　　ノードのメモリの上限値　28GB
　　コアのメモリの上限値　　7GB
です。

標準的な使い方は、ノード間をプロセス並列（ユーザー並列）、ノード内の4コアを自動並列（Thread 並列や OpenMP 並列など）にして下さい。ジョブ実行のパラメータが合っていない時、エラーメッセージが返されますが、そのメッセージが明瞭でない場合もあります。

コンパイルと実行の例
M9000 でコンパイルして、実行ファイル mearthd3dd2n016.fx1 を作成し、FX1 で６４core を用いて、１６プロセス並列＋４スレッド並列での実行です。

```
mpifrt mearthd3dd2n016.f -o mearthd3dd2n016.fx1 -Kimpact -Z mpilist
cp mearthd3dd2n016.fx1 progmpi
qsub mpiex_fx0064s4.sh
```

```
se000% more mpiex_fx0064s4.sh
# @$-q f64 -lp 4 -lP 16 -eo -o pexecmpi0064s4.out
# @$-lm 8.0gb -cp 1:00:00
cd ./gridtest2/
mpiexec -n 16 ./progmpi
```

また、コンパイルで prefetch の機能を使うこともできます。現在、自動並列機能を用いる場合は次のオプション-Kimpact -Kprefetch_model=FX1 を利用するのが最速を得る方法です。

```
mpifrt prog.f -o prog.fx1 -Kimpact -Kprefetch_model=FX1 -Z mpilist
```

また、flat MPI の場合は-Kimpact は使えませんので次のようにします。

```
mpifrt progmpi712bb4a.f -o progmpi64 -Kprefetch_model=FX1 -Z mpilist
qsub mpiex_fx0064s1.sh
```

```
se000% more mpiex_fx0064s1.sh
#    @$-q f64 -lp 1 -lP 64 -eo -o pexecmpi0064s1.out
#    @$-lm 7.0gb -cp 24:00:00
cd ./vpp05a/mearthb3/
mpiexec -n 64 ./progmpi64
```

flat MPI が速いとは限りませんので、情報基盤センターや富士通の方は自動並列の機能を使うことを奨めてあります。

(B) HX600 の使い方
先ず、HX600 に接続します。

```
ssh a41456a@sp2.cc.nagoya-u.ac.jp
```

シングルのジョブのコンパイルと実行の方法

```
frt -o prog prog.f
qsub exeh16.sh

se000% more exeh16.sh
#   @$-q h16 -lp 16 -eo -o sexec016.out
#   @$-lm 8.0gb -cp 1:00:00
setenv   parallel 16
cd ./mhdta/hx600/
./prog
```

MPI 並列ジョブのコンパイルと実行の方法

128core を用いて、4core を thread 並列で動かす場合

```
mpifrt progmpi.f -o progmpi -Kparallel -Z mpilist
qsub mpiex_fx0128s4.sh

se000% more mpiex_fx0128s4.sh
#   @$-q h128 -lp 4 -lP 32 -eo -o pexecmpi.out
#   @$-lm 12.0gb -cp 1:00:00
cd ./gridtest2/hx600/
mpiexec -n 32 ./progmpi
```

(C) M9000 の使い方

これは、HPC2500 の後継機種なので、従来の HPC と同じように使うことができます。また、TSS も利用できます。FX1 で計算したデータのデータ処理や画像処理は FX1 で直接に処理して出力を出すことができませんので、その処理にも M9000 を利用します。

次に、TSS でジョブを流す場合と MPI 並列プログラムをバッチジョブとして実行する場合の例を示します。

TSS でジョブをコンパイル・実行

```
frt prog.f -o prog
prog
```

MPI 並列プログラムをバッチジョブとしてコンパイル・実行

```
mpifrt progmpi.f -o progmpi -Kparallel -Z mpilist
```

qsub mpiex_m64s4.sh

se000% more mpiex_m0064s4.sh
#　@$-q m64 -lp 4 -lP 16 -eo -o pexecmpi0064s4.out
#　@$-lm 10.0gb -cp 1:00:00
cd ./gridtest2/m9000/
mpiexec -n 16 ./progmpi

荻野

名古屋大学の情報基盤センターの新システムを利用するための
Homepage は以下にあります。

センターのサービス
http://www2.itc.nagoya-u.ac.jp/center/index.html

新システムを利用するためのドキュメント
http://www2.itc.nagoya-u.ac.jp/sys_riyou/manual.htm

ここの中で、

1. 利用の前に
4. 新システム
　HX600 の利用
　FX1 の利用
　M9000 の利用

の PDF ファイルに永井さん、津田さんが書かれたものがあります。
