新スーパーコンピュータシステムの概要と利用方法について

永井　亨　津田知子

Ⅰ．はじめに

３月１日から新システムが稼動します。新システムは，スーパーコンピュータ，アプリケーションサーバ，画像処理システム，メールサーバ，媒体変換システム，データアーカイブサーバなどから構成されます。新システムの構成を図１に示します。

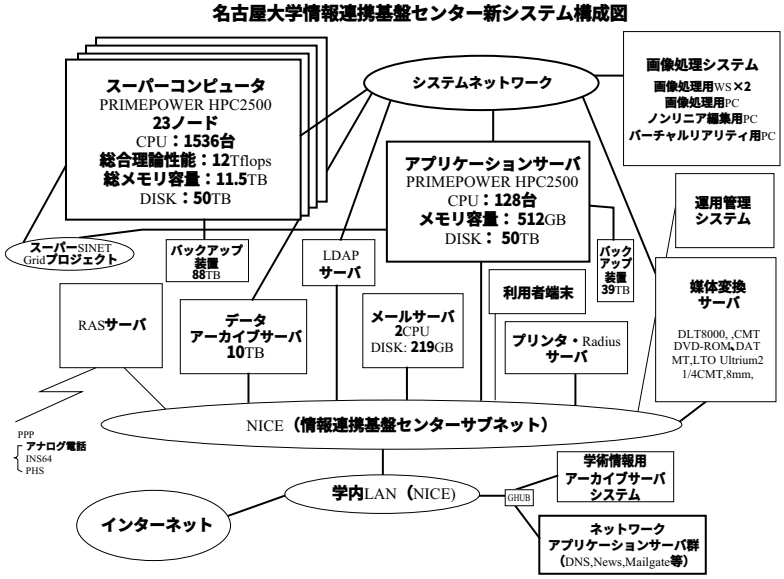

図１　新システム構成図

スーパーコンピュータとアプリケーションサーバの構成はつぎのとおりです。

＜スーパーコンピュータ＞

Fujitsu PRIMEPOWER HPC2500　23ノード

64CPU/512GBメモリ　22ノード

128CPU/512GBメモリ　１ノード

総合理論性能　12Tflops

総メモリ容量　11.5TB

ディスク容量　50TB

＜アプリケーションサーバシステム＞

Fujitsu PRIMEPOWER HPC2500

128CPU/512GBメモリ　１ノード

総合理論性能　１Tflops

ディスク容量　50TB

上に示すようにスーパーコンピュータとアプリケーションサーバは，PRIMEPOWER HPC2500という同じ種類の機種になりました。そこで，本センターでは，スーパーコンピュータとアプリケーションサーバとを一体的に運用し，利用者は，アプリケーションサーバ（ホスト名：hpc.cc.nagoya-u.ac.jp，IPアドレス：133.6.1.153）にloginすることにより，スーパーコンピュータもアプリケーションサーバも利用できるようになります。利用方法は，後で述べるように大きく変わることはなく，現在の利用方法とほぼ同じで，バッチの処理の形態で利用することになります。しかし，スーパーコンピュータは，ベクトル並列型からスカラ並列型へと変わりますので，プログラムの性能向上を目指すプログラミングテクニック面では，大きな変化があります。翻訳時のコンパイルオプションやバッチジョブ投入時のオプションの指定は，変更がありますので，これらのオプション指定には注意してください。

Ⅱ．スーパーコンピュータシステムの概要について

新スーパーコンピュータは23ノードから構成される分散並列スカラ計算機です。各ノードは共有メモリ計算機で，ノード間は高速クロスバネットワーク（４GB／秒）で結合されます。22ノードは64CPU，１ノードは128CPUを搭載し，主記憶容量はそれぞれ512GBです。各CPUのマシンサイクルは２GHz，最大4命令同時実行可能なので理論最大性能は８Gflopsです。したがって，システム全体では理論最大性能12Tflops，主記憶容量11.5TBとなります。また，磁気ディスク容量は，アプリケーションサーバと合わせて100TBです。

気になる性能についてですが，これまでのベクトル並列計算機VPP5000/64の１PEの理論最大性能が9.6Gflopsですから，理論上はHPC2500の１CPUもそれほど遜色ない性能をもっています。しかし，VPP5000ではベクトルレジスタやロード・ストアパイプライン等のベクトル計算機特有のハードウェア機構が装備されていたことなど，特にデータアクセスに関する性能に差があるため，実効的にはVPP5000の１PEとHPC2500の４～８CPUが同程度の性能になると予想しています。ただし，これはいくつかのサンプルプログラムで比較した結果に基づいた予想ですので，VPP5000上で利用者が作成した個々のプログラムがどの程度の性能がでるかは実際にHPC2500で実行してみないとわかりません。ベクトル計算機とスカラ計算機とでは最適なプログラミングも違ってきますし，自動並列化コンパイラにもまだ性能向上の余地があります。チューニング（特に１次・２次キャッシュの効果を意識したプログラミング）によってかなり性能が向上する場合もありますので，性能が出ない場合は，遠慮なくセンターにご相談ください。

主記憶容量は１ノードあたり512GB，全体では11.5TBですから，VPP5000では実行できなかった超大規模計算も可能となります。また，ノード間のデータ転送速度は最大４GB／秒ですから並列計算の高速化が期待されます。

HPC2500でノードをまたがる並列プログラムを実行するにあたって注意を要する点があります。

ノード間のデータ通信にはユーザDTU（data transfer unit）と呼ばれるハードウェアを使用します。ユーザDTUは各ノードに32個あって31個は占有タイプ，１個は共有タイプです。各プロセスはいずれかのタイプのユーザDTUを介して別のノードにあるプロセスと通信を行いますが，占有タイプより共有タイプの方が，通信速度が遅い傾向にあり，特にデータ転送量が小さい場合にこれが顕著になります。したがって，占有タイプを使用する方が計算時間は短くなるのですが，各ノードに31個しか実装されていませんから，１ノードあたり最大31プロセスを起動する並列ジョブしか使用できません。このため，１ノードあたり31プロセスを超える並列ジョブを投入すると資源が不足するため永久に実行されないことになります。１ノードあたり31プロセスを超える並列ジョブでは共有タイプのユーザDTUを使用する必要があります。タイプの選択は実行時のオプションで指定します。ただし，共有タイプのユーザDTUが使用できるのはMPIライブラリを使用するプログラムのみで，XPFortran（VPP Fortranと同等の機能をもつHPC2500の並列処理用Fortran）プログラムは使用できません。

Ⅲ．スーパーコンピュータの利用について

１．hpcで利用できるソフトウェア

Ⅰ章で述べたようにスーパーコンピュータとアプリケーションサーバは，一体的に運用されます。以下では，これらのシステムをhpcシステムと呼ぶことにします。hpcシステムのOSは，Solaris9です。hpcシステムで利用できるソフトウェアの一覧を表１に，アプリケーションサーバでのみ使用できるソフトウェアの一覧を表２に示します。アプリケーションサーバに使用が限定されているソフトウェアは，主にインタラクティブに使用するアプリケーションパッケージです。アプリケーションパッケージは，従来からのものが新システムでも継続利用できます。新しいアプリケーションとしては，SAS/Genetics，MOPAC，ICEM CFD，CADfixなどがあります。

表１　hpcシステムのソフトウェア一覧

種　類	ソフトウェア名
プログラミング言語	Fortran（JIS X3001-1:1998準拠）
	C（ISO/IEC 9899:1999準拠）
	C++（ISO/IEC 14882:1998準拠）
	Java
並列処理言語	XPFortran
	OpenMP(Fortran ver. 2.0，C/C++ ver. 2.0)
数値計算ライブラリ	SSLⅡ（科学用サブルーチンライブラリ）
	並列版SSLⅡ（科学用サブルーチンライブラリ）
	C- SSLⅡ（科学用サブルーチンライブラリ）
	NUMPAC（数値計算ライブラリ）
	BLAS（線形計算ライブラリ）
	LAPACK（線形計算ライブラリ）
	ScaLAPAC（MPI並列処理向き線形計算ライブラリ）
メッセージパッシングライブラリ	MPI（MPI2.0仕様準拠）
アプリケーション	α－FLOW（汎用3次元流体解析システム）*
	STAR-CD（非構造格子汎用熱流体解析ソフトウェア）*
	LS-DYNA3D（非線形動的構造解析ソフトウェア）*
	Material Explorer（材料設計システム）*
	Gaussian（分子軌道計算プログラム）*
	AMBER（分子構造計算プログラム）*
	VisLink（リアルタイム可視化ソフトウェア）
	AVS/Express Developer（対話型３次元ビジュアリゼーションシステム）
	AVS並列版
	eta/FEMB(Finite Element Model Builder)

*印のアプリケーションは，並列版が用意されている。

表２　アプリケーションサーバだけで使用できるソフトウェア

SAS（BaseSAS，SAS/STAT，SAS/GRAPH，SAS/Genetics）
MATLAB（matrix laboratory）
Maple（数式処理システム）
Mathematica（数式処理システム）
IDL（Interactive Data Language）
I-DEAS（総合設計支援システム）
ICEM CFD（汎用メッシュ生成／可視化統合システム）
CADfix（データ検証・修正ツール）
ATLAS（日英・英日翻訳システム）
fastDNAml（最尤法による進化系統樹推定プログラム）
MOLPRO（分子軌道計算プログラム）*
MOPAC（半経験的分子軌道法プログラム）*
α-FLOWのプリ／ポストプロセッサ
STAR-CDのプリ／ポストプロセッサ
Material Explorerのプリ／ポストプロセッサ

*印のアプリケーションは，並列版が用意されている。

２．利用形態とジョブ種別

前述したようにスーパーコンピュータとアプリケーションサーバは，一体的に運用されていますので，スーパーコンピュータを利用する場合は，ネットワークからアプリケーションサーバ hpc.cc.nagoya-u.ac.jp（IPアドレス：133.6.1.153）に接続します。プログラムの編集，翻訳，デバッグなどは，ここで行います。スーパーコンピュータの利用は，従来と同じようにNQSによりバッチジョブを投入して行います。

スーパーコンピュータとアプリケーションサーバで利用できるファイルシステムを図２に示します。これらのシステムで利用できるファイルは，以下に示す/home，/large0（または，/large1），/large_tmp0（または，/large_tmp1）の３種類で，どちらのシステムからも同じように見えます。

/home：loginしたときのホームディレクトリのファイル。

/large0，/large1：高速大容量ファイルSRFS。

/large0_tmp0，/large1_tmp1：高速大容量ファイルSRFS。ただし，バックアップが行われないファイル。

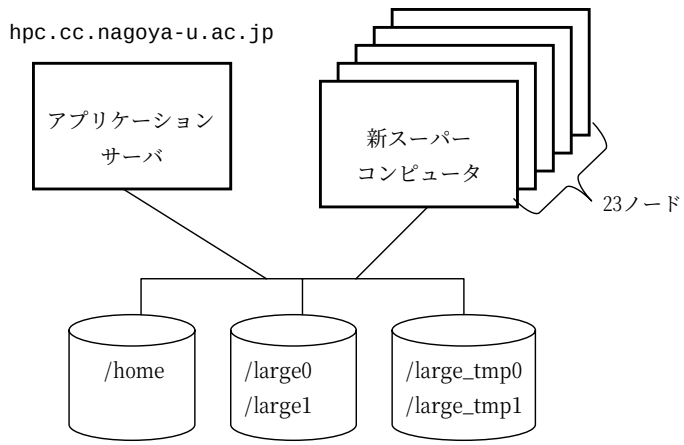

図２　スーパーコンピュータのファイルシステム

/large0と/large1は，従来の/home/dpfsに対応する高速大容量向きのファイルシステムです。/large_tmp0と/large_tmp1は，/large0と/large1と同じで高速大容量向きのファイルシステムですが，バックアップ処理を行いません。それぞれの用途に合わせてご利用ください。

スーパーコンピュータで利用可能なCPU数，CPU時間やメモリ量などの計算資源はジョブの種別により異なります。表３にジョブ種別を示します。

表３　スーパーコンピュータのジョブ種別

	キュー名	最大使用可能CPU数	CPU使用時間		ラージメモリ		経過時間	ユーザDTU
			標準値	制限値	標準値	制限値	制限値	
バッチジョブ	a8	8	10時間	10時間	2 GB	400GB	２時間	利用不可
	p8	8	10時間	無制限	2 GB	400GB	20時間	利用不可
	p16	16	10時間	無制限	2 GB	400GB	100時間	利用可能
	p64	64	200時間	無制限	2 GB	400GB	200時間	利用可能
	p128	128	200時間	無制限	2 GB	400GB	336時間	利用可能
	p256	256	200時間	無制限	2 GB	400GB	336時間	利用可能
	p1024	1024	200時間	無制限	2 GB	400GB	336時間	利用可能
TSS	―	128	２時間	無制限	―	128GB	―	利用不可

注１）CPU使用時間は，各CPUの合計です。
注２）ラージメモリは，プロセスあたりの値です。

このジョブ種別によるCPUやメモリの資源に関するユーザの指定は，従来のスーパーコンピュータVPP5000と大きく変わっています。VPP5000では，CPU時間は，プロセスあたりの値でしたが，新スーパーコンピュータでは，使用した各CPUの総合計の時間となります。また，メモリは，VPP5000では，ジョブあたりの値を指定しましたが，新スーパーコンピュータでは，プロセスあたりの値を指定することとなりますので，ジョブの投入に当たっては，利用の手引きを参考に十分注意して指定してください。

３．翻訳コマンド

hpcシステムで利用できる言語と翻訳コマンドを表４に示します。なお，frtコマンドでは，Fortran95の言語仕様が標準で動きます。

表４　翻訳コマンド

言　　語	コマンド
Fortran	frt
XPFortran	xpfrt
C	fcc
C++	FCC

翻訳コマンドのオプションの詳細は，manコマンドで確認してください。以下では，従来のVPP5000と大きく異なる部分を示します。

（１）自動並列化機能

hpcシステムの１CPUの実効性能は，ベクトル機であるVPP5000に劣ります。プログラムの性能向上には，並列化が大きなキーとなります。一番手軽にプログラムを並列化する方法は，コンパイラの自動並列化機能を利用することです。hpcシステムのFortran，C，C++のコンパイラには，それぞれ自動並列化の機能が備わっています。コンパイルオプションで，`-Kparallel`

と指定することにより，自動並列化されます。例えば，Fortranのプログラムでは，多くの場合DOループの外側が，スレッド並列という形で並列化されます。スレッド並列することで，各スレッドに別々のCPUを割り当てて実行することにより，プログラムの実行性能をあげることが期待できます。しかし，その並列効果は，プログラムに大きく依存しますので，現在お持ちのプログラムを一度自動並列化して実行してみることをお薦めします。なお，自動並列化だけでは，性能を向上させることができない場合には，利用者自身で，OpenMPディレクティブを挿入することにより更なる並列化を行うことも可能です。

（2）ラージページ機能

ラージページ機能とは，メモリ管理の単位であるページサイズを通常の8KBから４MBに拡張し，更に実行に先立ち固定的にメモリを確保することにより，実行性能を向上させるものです。本センターでは，hpcシステムの各コンパイラの標準値としてこの機能を指定していますので，コンパイルされたモジュールは，ラージページのメモリ上でしか動作しないことにご注意ください。バッチ処理の場合は，qsubコマンドの`-lM`オプションでプロセスあたりのラージメモリ領域を指定しますが，ラージメモリ量としては，その指定されたメモリ量にプロセス数を乗じた値が，ジョブが実行された段階で固定的に確保されます。メモリ領域の無駄使いにならないよう`-lM`で指定する値には，妥当な値を指定するようにご留意願います。

４．ライブラリの利用

ライブラリを使用する場合には，翻訳コマンドでオプションの指定が必要です。使用するライブラリとfrtコマンドで指定するオプションを表５に示します。

表５　ライブラリとfrtコマンドでのオプション

ライブラリ名	オプション
NUMPAC：数値計算ライブラリ	`-l numpac`
SSLⅡ：科学用サブルーチンライブラリ	`-SSL2`
SSLⅡ：科学用サブルーチンライブラリスレッド版	
BLAS（線形計算ライブラリ）	
LAPACK（線形計算ライブラリ）	
ユーザ登録のセンターライブラリ	`-l ulib`
図形出力ライブラリ	`-l ps`

５．バッチジョブの投入例

（1）スレッド並列モジュールの実行例

`frt`コマンドの`-Kparallel`オプションを指定して自動並列化を行ったり，OpenMPを用いて並列化した場合には，スレッド並列となります。ここでは，スレッド並列モジュールを8並列で実行する場合の例を示します。実行キューは，`p8`（`8CPU`まで利用可能）を指定します。スレッド数の指定は，`-lp`オプションで行います（ここで，`p`は小文字）。使用するメモリが`2GB`を超える場合には，`-lM`オプションで必要メモリ量を指定します。この例では，実行の終了をメ

ール[1]で通知するように-meオプションを指定しています。また，-lTオプションに使用するCPU時間（この例では12時間30分）を指定しています。なお，-lTオプションで指定するCPU時間は，各CPUの総合計を指定します。

１－１）スクリプトファイルnqs/exec_tp.shの内容

実行可能ファイル名：pg/tpara

```
#  @$-q p8  -lp 8  -eo  -o tpara.out
#  @$-me  -lT 12:30:00   -lM 5gb
cd  pg
./tpara
```

１－２）スレッド並列モジュールの実行依頼

スクリプトファイル：exec_tp.sh

qsubコマンドで依頼する。

```
hpc% qsub exec_tp.sh
```

（２）プロセス並列モジュールの実行例

XPFortranやMPIで並列化した場合には，プロセス並列になります。ここでは，プロセス並列モジュールを32並列で実行する場合の例を示します。実行キューは，p64（64CPUまで利用可能）を指定します。プロセス数の指定は，-lPオプションで行う（ここで，Pは大文字）。１プロセスあたりの使用メモリが2GBを超える場合には，-lMオプションで必要メモリ量を指定します。以下の例では，実行の終了をメールで通知するように-meオプションを指定しています。また，-lTオプションに使用するCPU時間（この例では20時間）を指定しています。なお，-lTオプションで指定するCPU時間は，各CPUの総合計を指定します。

１－１）スクリプトファイルnqs/exec_pp.shの内容

実行可能ファイル名：pg/ppara

```
#  @$-q p64  -lP 32  -eo  -o ppara.out
#  @$-me  -lT 20:00:00   -lM 10gb
cd  pg
./ppara
```

１－２）スレッド並列モジュールの実行依頼

スクリプトファイル：exec_pp.sh

qsubコマンドで依頼する。

```
hpc% qsub exec_pp.sh
```

1　qsubを実行しているホスト以外のシステムにメールで通知したい場合には，qsubを実行しているホストの.forwardファイルにメールアドレスを記述する。

Ⅳ．おわりに

この稿をしたためている時点では，まだ，新スーパーコンピュータはその姿を現していません。そのため，利用の詳細をお伝えすることができません。詳細については，「利用の手引」という形で本センターのWebに掲載する予定ですので，適宜そちらをご覧ください。新スーパーコンピュータは，スカラ並列型であるので，CPUの数が1536と非常に多く，これだけのシステムを運用するのは，センター職員も始めての経験です。このシステムをユーザが使いやすく，かつ，効率よく利用していくためには，ユーザ，センター共々，これからいろいろな経験を積んでいくことが必要だと思われます。システムや利用に関する質問，疑問，要望などは，どしどしsoudan@cc.nagoya-u.ac.jpにお寄せください。

（ながい　とおる：名古屋大学情報連携基盤センター大規模計算支援環境研究部門）
（つだ　ともこ：名古屋大学情報連携基盤センター学術情報開発研究部門）