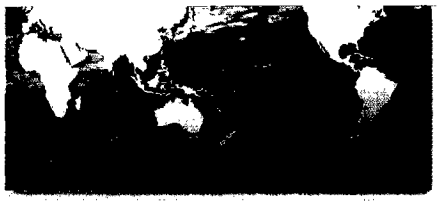


解説

地球シミュレータ計画

スーパーコンピュータで地球の未来を映し出す

〈後編〉

横川 三津夫
谷 啓二
日本原子力研究所
地球シミュレータ開発特別チーム

科学技術庁は，プロセス（基礎科学）研究，観測，計算機シミュレーションの三位一体で地球環境変動予測研究を推進するプロジェクトを平成9年度より推進している．その一環として，大気大循環シミュレーションで実効性能5TFLOPS（ピーク性能40TFLOPS）の超高速並列計算機「地球シミュレータ」を開発中である．この地球シミュレータのハードウェア，基本ソフトウェア，応用ソフトウェアについてその概要を解説する．

●地球シミュレータの全体システム

地球シミュレータ・システムは，①ハードウェア，②オペレーティングシステム（OS），コンパイラ，運用・管理ソフトウェア（センタルーチン）などからなる基本ソフトウェアおよび③応用ソフトウェアから構成される．以下，各システムの概要をまとめる．

●地球シミュレータのハードウェア

現在，気候・気象分野の代表的な大規模シミュレーションでは，4～6GFLOPSの実効速度が実現されている．地球シミュレータでは，同種のシミュレーションにおいてその1000倍である5TFLOPSの実効速度を達成することを目標として開発を進めている．また，それらのシミュレーションでは，現用のベクトル並列計算機で30％程度の実効効率が得られているが，地球シミュレータは，現用の計算機に比べ並列数も格段に多くなることから，実効効率は15～20％に低下すると仮定して，実効速度5TFLOPSを実現するために必要なピーク性能として，30TFLOPS以上は必要となる．一方，シミュレーションモデルの高分解能化のために必要な主記憶容量としては，全球大気10kmメッシュでは，経度方向4000点，緯度方向2000点，鉛直方向150点，各メッシュ点上の変数は問題にも依存するが倍精度で200～400個，海洋もほぼ同程度を仮定して，

$$4000\times2000\times150\times(200\sim400)\times2\times8\ (\text{倍精度})$$
$$=4\sim8\text{TB}$$

が必要になる．

地球シミュレータの完成が見込まれる2002年においても，ピーク性能30TFLOPS以上，主記憶（メモリ）8TB以上を単一プロセッサで実現することは到底不可能であ

り，分散並列化せざるを得ない．問題は，これらCPUとメモリをどのように分散並列化するかにある．計算機の並列化方式は以下の2つに大別される．

①共有メモリ方式：大容量のメモリを複数のプロセッサで共有する方式

②分散メモリ方式：小容量のメモリとプロセッサで独立した要素計算機を構成し，複数の要素計算機を相互に高速ネットワークで結合，一体化する方式

共有メモリ方式は，メモリのプロセッサへのデータ供給能力でシステム全体の大きさが制限され，来世紀初頭の半導体技術を用いても，100GFLOPS程度が限界である．

一方，分散メモリ方式のシステム全体の理論性能（ピーク性能）は，要素計算機のピーク性能×並列数であるため，原理的にはネットワークで接続する要素計算機の数（並列数）を増やしていけばいくらでもシステムを大きくできる．しかし，並列数が多くなりすぎると，システムのピーク性能に対して，実効性能はあまり伸びなくなる．これは，

①並列数増加に伴う要素計算機間のデータ通信時間等のオーバーヘッド

②各要素計算機の負荷のアンバランス

などに起因する計算機の遊びの時間（アイドリング）の増大によるものであり，並列計算機の利用（並列化ソフトウェア）技術が十分成熟していないことにも原因がある．ユーザのここ数年間の並列ソフトウェア技術の進展から判断して，分散メモリ方式の並列数は1000未満が現実的である．

要素計算機を並列に接続するためのネットワークには，メッシュ型，トーラス型，クロスバ・スイッチ型など種々の方式が提案されている．地球シミュレータでは，クロスバ・スイッチ型の一種である，単段クロスバ・ネットワークを採用する．その内容は後述するが，この方式は，他の方式に比べ非常に高性能である一方，非常に多くの高速スイッチ（約並列数の2乗）を必要とする．この単段クロスバ・ネットワークに必要な半導体技術から判断して，並列数は，やはりたかだか1000程度である．

分散メモリの並列数を1000以下とした場合，

1並列当たりの計算機能力＝30TFLOPS/1000
＝30GFLOPS

が必要となる．来世紀初頭では，汎用マイクロプロセッサの動作周波数はGHz時代（現在，0.3～0.6GHz）になっているものと思われるが，この場合，複数命令を同時に処理してもそのピーク性能は数GFLOPS程度であろう．一方，ベクトル計算機では，マイクロプロセッサの同時複数命令処理に対応するのが，ベクトル演算器の複数セット（多重度）である．従来のベクトル計算機では，この多重度は多数のLSIチップ（数十個）を用いて実現されてきたが，これがベクトル計算機の高性能化の障害の1つになっていた．地球シミュレータでは，今後のLSIの高密度化技術に期待し，数十個のLSIを1つの超集積回路（ULSI）に収める1チップの高性能ベクトル計算機の実現を目指す．この場合，チップの大きさが格段に大きくなるため，汎用マイクロプロセッサ並みまではいかないにしても，来世紀初頭には動作周波数は500MHz（現行150MHz）程度が期待できる．しかし，あまり多くのセットの演算器を1チップに詰め込むことの半導体技術上の制限から，単一のベクトルプロセッサとしては10GFLOPS程度（現在，～2GFLOPS）が限界である．以上のことから，単一プロセッサで1並列分（計算ノード）を構成することは困難である．そこで，地球シミュレータでは，複数のベクトル型要素計算機を共有メモリで密に結合させ1つの計算ノードを構成し，そのノードを数百台のネットワークで粗に結合させた（分散メモリ型）システムを採用することにした．

以上，ハードウェアの概要をまとめると（平成10年4月時点での設計値），

(1) アーキテクチャ　：ベクトル計算機を要素計算機とし，複数の要素計算機を共有メモリで密結合して計算ノードを構成．その計算ノードを複数台ネットワークで結合し，システム全体では分散メモリ型並列計算機を構成

(2) 処理性能　　：ピーク性能 40TFLOPS
　実効性能　　：大気大循環シミュレーションで5TFLOPS以上

- 全ノード数　　：640
- 1ノードあたりのプロセッサ数 ：8
- 全プロセッサ数 ：5120
- 要素計算機性能 ：8 GFLOPS
- 計算ノード性能 ：64 GFLOPS

(3) 主記憶容量　　：10 TB

- ノード内共有メモリ：16 GB
- ノードあたりのメモリスループット：256 GB/s

(4) 結合ネットワーク：単段クロスバ・ネットワーク

これらの地球シミュレータのハードウェアの概要を表したものが**図-1**である．

地球シミュレータは，多様な応用ソフトウェアに利用される．このような多種のソフトウェアの並列化に柔軟

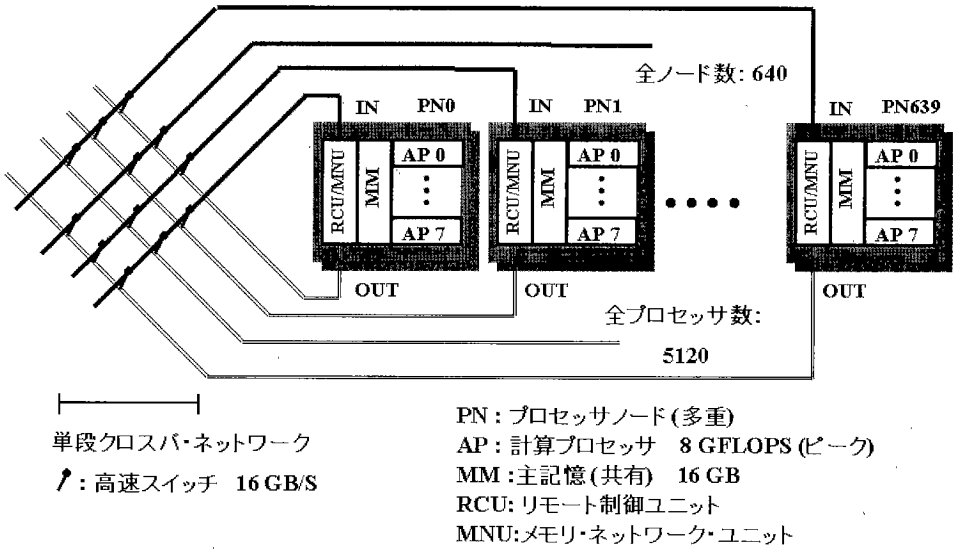


図-1 地球シミュレータのハードウェア概念図（試案）

に対応するため，また，多数のユーザに効率よくシステムを提供するためには，すべてのノードが平等にネットワーク接続される（任意のノード間が中継ノードを経ることなく直接的に接続される）必要がある．このため，地球シミュレータでは，これらの条件を満足する結合ネットワークとして単段クロスバ・ネットワークを採用する．これは，図-1に示すように，各ノードからの入出力ライン（バー）を交差させ，各交点に高速スイッチを設けたものである．任意の2つのノード間のデータ転送は，データ出力（OUT）のバーとデータ入力（IN）のバーの交点のスイッチ1つをオンにするだけでデータ転送路が開ける．また，すべてのノードが平等な（ネットワークの中の接続位置に差異がない）ことから，必要ノード数を確保する場合，どのようなノードの組合せも可能であり，システムの運用にきわめて柔軟に対応できる．そのデータスループットは，1ノード，1方向（IN，あるいはOUT）当たり16GB/s（1ノードの処理速度の1/32）の高速性能の実現を目指している．この高速性能は，来世紀初頭の技術をもってしても，1個のスイッチで実現することは不可能であるため，129本のケーブルと同数のスイッチを用いて並列データ転送することにより実現する．したがって，ケーブルの総本数は640×129＝82,560本，総ケーブル長は約2,890km，ケーブルだけの総重量でも217tに及ぶ．この総ケーブル長は，北は知床から南は石垣島までの，ほぼ日本の南北端間の距離に匹敵する．このように，このネットワークは，膨大なハードウェアのリソースを必要とする．

地球シミュレータ開発の最大の課題は，省スペース，省エネルギー，省コストである．それには，要素計算機の高性能化と複数の要素計算機で構成される計算ノードのコンパクト化が不可欠であり，①半導体技術，②パッケージング技術，実装技術，③冷却技術のすべてにわたり来世紀初頭の最先端技術が要求される．ここで，地球シミュレータのための技術開発の主要な項目についてまとめてみる．

(1) 半導体の高速化：クロックサイクル500MHz以上（現行150MHz），
(2) LSIの高密度化：トランジスタ密度，1000万個／cm^2以上（現行350万個／cm^2），
(3) LSIチップサイズの大型化：2cm×2cm程度：総トランジスタ数，4～5000万個，
(4) 高密度実装：1チップ当たりのピン数，4～5000本（現行1000本程度）．

これらの技術開発を行い，1チップの高速ベクトルプロセッサを開発する（従来は数十個のLSIで構成）．また，このワンチップ・ベクトルプロセッサを高密度に実装することにより，省スペース，省エネルギー，省コストシステムの実現を目指している．**図-2**に地球シミュレータの1ノードと同じ性能の現用スーパーコンピュータとの大きさの比較を示す．このように，来世紀初頭には，現在のスーパーコンピュータに匹敵するものがワークステーションのようにデスクサイドに置けるようになるものと期待される．

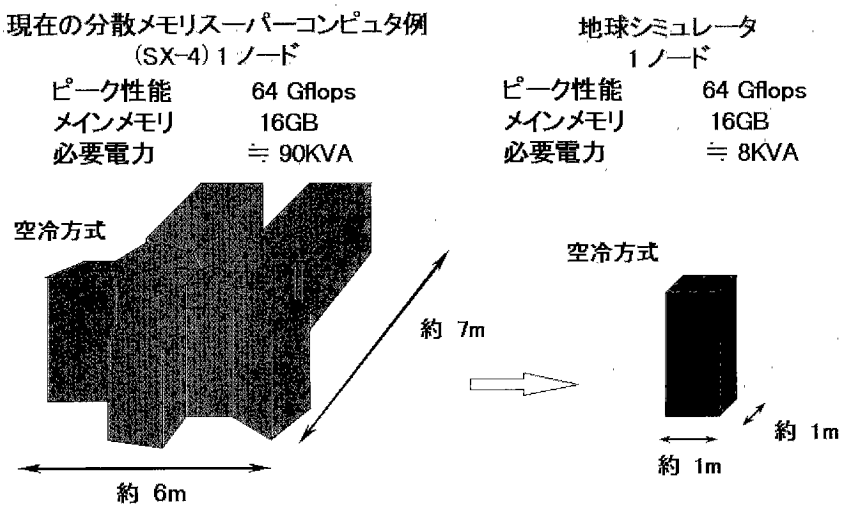


図-2 1ノード用キャビネットのサイズ比較

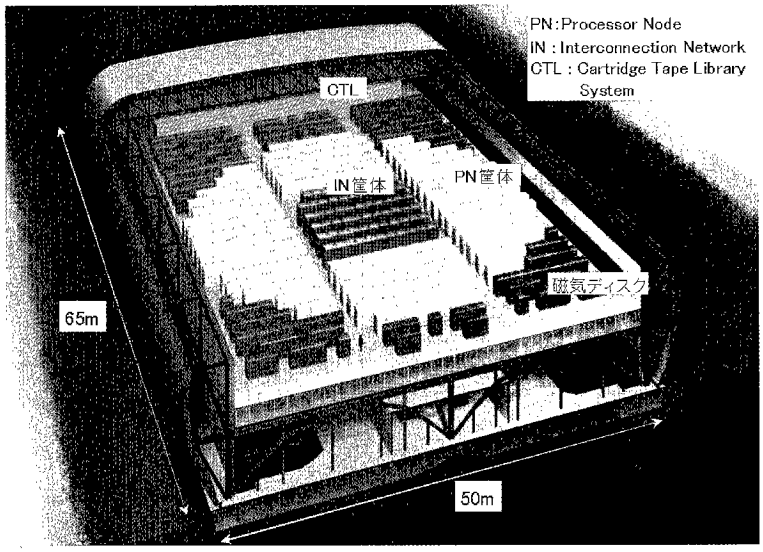
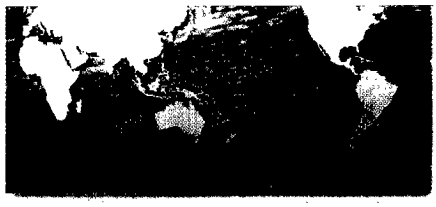


図-3 地球シミュレータ完成予想図

以上の各技術開発をベースに，想定される地球シミュレータ完成のイメージを**図-3**に示す．地球シミュレータは，専用の建家に収納される．建家は3層からなる．第1層部には，計算機システムからの膨大な発熱を除去するための空調機器が設置される．第2層部はフリーアクセスとして，前述の全長3,000kmに達するネットワークケーブルを納めるとともに計算機システムに冷気を分配する空調ダクトスペースとしても利用される．第3層部には，計算ノード筐体，単段クロスバ・ネットワークの高速スイッチを収納した結合ネットワーク筐体，磁気ディスクシステムおよび大容量記憶装置（カートリッジテープライブラリシステム）を設置する．建家の大きさは，約50m×65mである．

地球シミュレータの基本ソフトウェア

基本ソフトウェアは，ハードウェアとユーザの間をとりもつものであり，ハードウェアから近い順に，オペレーティング・システム（OS），コンパイラ，運用・管理ソフトウェア（センタルーチン）と階層構造をなしている．地球シミュレータは，前述のように，現用最大級の計算機が計算ノードとして数百台の規模で高速ネットワークで



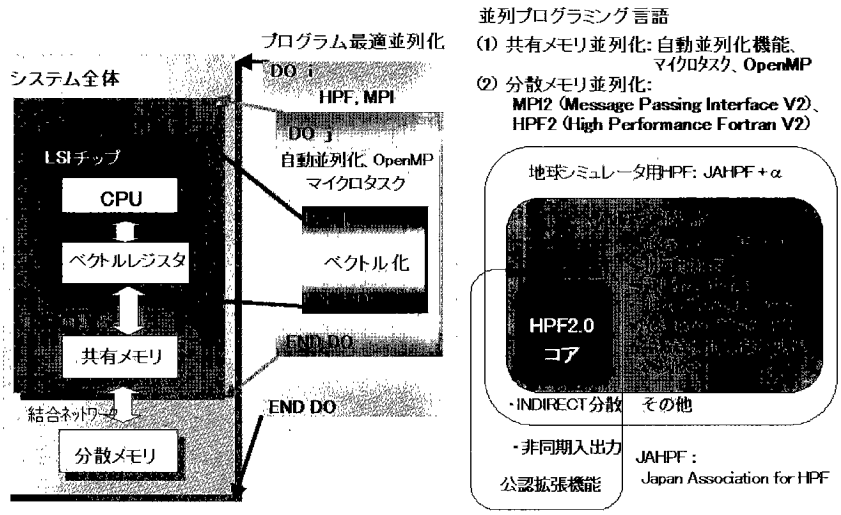


図-4　地球シミュレータのメモリハイアラキ構造と並列プログラミング言語

並列に結合された前代未聞の巨大システムとなる．並列処理にあまり習熟していないユーザでもその性能を十分引き出せるためには，この基本ソフトウェアの性能がきわめて重要になる．地球シミュレータにおいては，基本ソフトウェアの開発は，既存のベクトル型並列スーパーコンピュータ用の標準機能をベースに，地球シミュレータに特有な機能を付加することで開発を進める．以下，その開発の重要課題に触れる．

(1) OS (UNIXベース)：地球シミュレータでは，メーカが自社製品のために開発を行っているものを極力流用し，地球シミュレータに特有な機能に限定して開発する．地球シミュレータでは640台のノードに磁気ディスク装置が分散接続されるため，計算の入出力 (I/O) も並列処理となる．これらの並列ファイルをユーザになるべく意識させることなく，単一ファイルイメージで利用できるようにするための並列I/O機能が最重要開発項目である．

(2) コンパイラ：地球シミュレータは，**図-4** (図-1も参照) に示すように，ベクトルレジスター共有メモリー分散メモリとメモリに階層構造を持っている．システムの性能を最大限引き出すためには，このメモリ構造を十分考慮した並列処理が必要になる．このため，地球シミュレータでは，並列プログラミング言語として，

- 共有メモリ並列化：マイクロタスク，自動並列化，OpenMP，
- 分散メモリ並列化：MPI (Message Passing Interface)，HPF (High Performance Fortran)

が用意される．特に，HPFについては，我が国では計算機ベンダとユーザの合同検討会 (JAHPF：Japan Association for HPF) において拡張仕様を検討し，HPFフォーラムに対してその標準化を提言していく活動を行っているが，地球シミュレータでは，このJAHPF拡張仕様にさらに計算地球科学に必要な機能を追加する．

(3) 運用センタールーチン：地球シミュレータは数千台の全計算機を同時に1つのジョブで使用することもできるし，また，計算機を複数の中小グループに分割してそれらを多数のユーザで同時にシェアすることもできる．このように，必要並列数やCPU時間の異なる多数のジョブを，遊んでいるプロセッサ数を極力少なくし，システム全体の利用効率をよく制御するための高機能スケジューラや，システムの稼働状況，ユーザのジョブの現況 (実行待ち，実行中など) のモニタ機能など新規に開発する．

地球シミュレータの応用ソフトウェア

地球シミュレータ計画では，世界で初めての試みとして，**図-5**および**図-6**に示すような並列最適化プラグイン方式を採用した大気・海洋大循環結合モデルライブラリ・システムを宇宙開発事業団 (NASDA) および海洋科学技術センタ (JAMSTEC) の委託を受け (財) 高度情報科学技術研究機構 (RIST) において開発中である．このソフトウェアは，関係機関の頭文字をとってNJR (NASDA-JAMSTEC-RIST) プログラムと称されている．ここで，プラグインとは，図-5に示すように，プラットフ

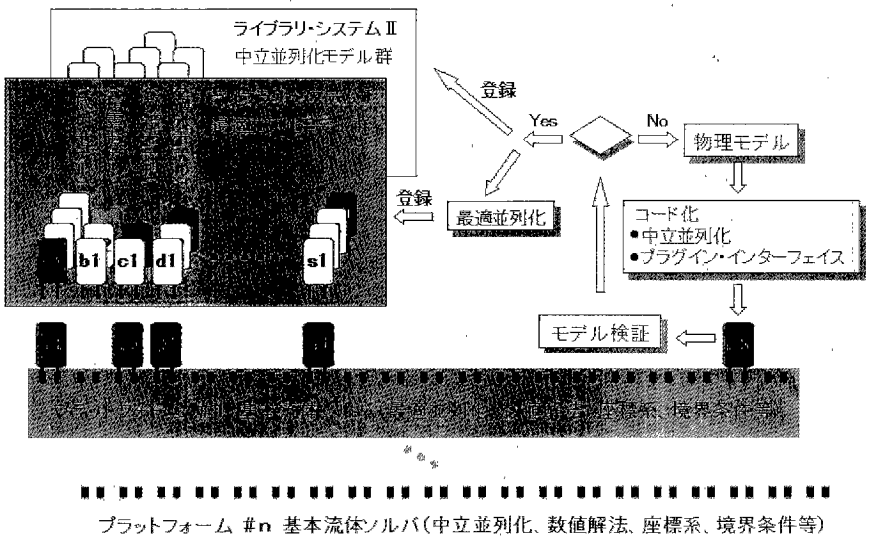


図-5　並列最適化プラグイン方式の概念図

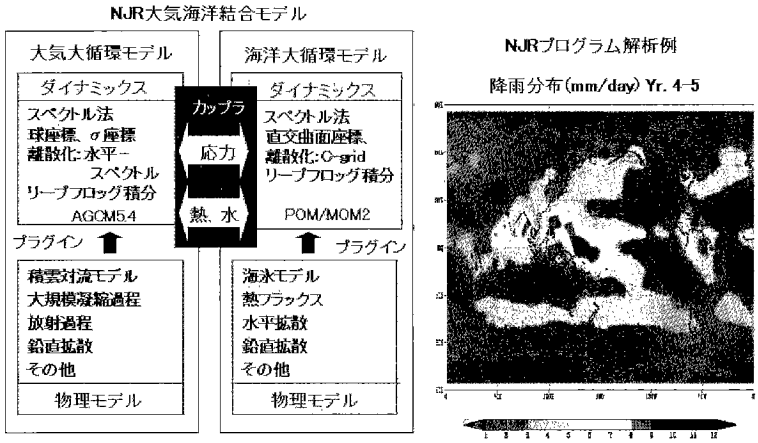


図-6　大気・海洋結合モデルNJRプログラムとグローバルな降雨分布の解析例

ォームと呼ばれる基本流体ソルバであるメインプログラムに，サブルーチンを一定のルールに従ったインタフェースを介して組み込む方式である．このシステムは，地球環境シミュレーションを実行するユーザに広くオープンにされ，新しいモデル開発に威力を発揮するものと期待される．たとえば，新しい物理モデルを開発する場合，プラグイン方式を用いれば，同じ流体ソルバ上で旧モデルと簡単に交換することにより，モデルの比較検討，検証が容易に行えることになる．検証を終えた新モデルコードは，地球シミュレータ上でそのハードウェア構成やコンパイラ等を考慮して最大限に調整し，最適化並列プログラムとしてライブラリ・システムに組み込まれる．このように，ライブラリ・システムが充実し，やがては大気・海洋大循環結合コミュニティ・モデルが形成されていくものと期待される．解析例として，静止大気から計算を開始し，4年目と5年目に得られた1日降雨量の年平均のグローバルな分布を図-6に示す．

固体地球シミュレーションに関しても，最適並列化プラグイン方式を採用し，並列有限要素法をベースとしたGeoFEMと称されるソフトウェアが科学技術庁の振興調整費をもとに開発されている．

謝辞　本テキストは，宇宙開発事業団，旧動力炉・核燃料開発事業団（平成9年度のみ），日本原子力研究所（平成10年度以降）からの委託により（株）日本電気において実施された地球シミュレータの概念設計，基本設計および要素技術設計をベースにまとめたものである．ご尽力をいただいた関係各位に感謝いたします．

（平成11年12月27日受付）