計算機アーキテクチャ 132−10
システムソフトウェアとオペレーティング・システム 80−10
ハイパフォーマンスコンピューティング 75−10
（1999. 3. 4）

地球シミュレータ用性能評価システムの開発

横川 三津夫[†1] 新宮 哲[†1] 萩原 孝[†2]
磯部 洋子[†3] 高橋 正樹[†1] 河合 伸一[†4]
谷 啓二[†1] 三好 甫[†5]

地球シミュレータは，640台の計算ノードをクロスバネットワークで結合した分散メモリ型並列計算機である．それぞれの計算ノードは，16GBの主記憶装置を8個のベクトルプロセッサで共有しており，全体のピーク性能は40Tflop/s，主記憶容量は10TBである．

地球シミュレータによるプログラムの実効性能を推定するため，地球シミュレータとそれに類似の計算機の動作をシミュレートできるソフトウェアシミュレータ（GSSS）を開発した．いくつかの基本的なDOループ群をSX-4の単体プロセッサで実行した場合の処理時間について，実測時間とGSSSによる推定時間を比較した結果，平均で約1%の相対誤差が得られた．また，地球シミュレータの単体プロセッサによるDOループ群の実効速度を推定した結果，平均で4.18Gflop/sが得られた．

Development of Performance Estimation System for the Earth Simulator

MITSUO YOKOKAWA[†1], SATORU SHINGU[†1], TAKASHI HAGIWARA[†2], YOKO ISOBE[†3], MASAKI TAKAHASHI[†1], SHINICHI KAWAI[†4], KEIJI TANI[†1] and HAJIME MIYOSHI [†5]

The Earth simulator is a distributed memory parallel system which consists of 640 processor nodes connected by a crossbar network. Each processor node is a shared memory system which is composed of eight vector processors. The total peak performance and main memory are 40Tflop/s and 10TB, respectively.

A software simulator(GSSS) for the Earth Simulator and its similar computers has been developed to estimate the sustained performance of programs. To validate an accuracy of the software simulator, the processing times for some kernel DO loops estimated by the GSSS are compared with the ones measured on an SX-4. It is found that the absolute relative error of the processing time is about 1% in average. The sustained performance of the kernel loops on the Earth Simulator has been estimated by the GSSS and a performance of 4.18Gflop/s in average is obtained.

1. はじめに

近年，地球温暖化やエルニーニョ現象等の地球規模の現象が注目されている．温室効果ガス等の排出は，地球温暖化に甚大な影響を及ぼしていると言われ，排出削減目標が設定されるなど我々の日常生活にも大きな影響が予想される．また，エルニーニョ現象は，局所的な集中豪雨や干ばつなどの被害をもたらしていると言われ，それらの被害を軽減するためにも現象のメカニズム解明が求められている．しかし，地球規模の現象は，時間的，空間的にもスケールの極端に異なる現象が複雑に絡み合って生じており，これらの複雑な現象のメカニズム解明は非常に困難である．

科学技術庁においては，地球規模の複雑な現象を解明することが重要であるという認識の下に，地球変動プロセスに関する基礎研究，地球規模の観測，数値シミュレーションの三位一体による研究体制を推進している．この内，数値シミュレーション分野の推進では，地球規模の複雑な諸現象をシミュレートするための超高速並列計算機システム「地球シミュレータ」開発と高度なソフトウェア開発を目標とする「地球シミュレータ」計画が策定された[1)]．この中では，近年の科学技術分野に特化したスーパーコンピュータのめざましい発展とそれに基づく計算科学の進展について述べられ，観測困難な現象や実験不可能な現象を解明するための

†1 日本原子力研究所
Japan Atomic Energy Research Institute
†2 日本電気（株）
NEC Corporation
†3 甲府日本電気（株）
NEC Kofu Ltd.
†4 宇宙開発事業団
National Space Development Agency of Japan
†5 地球シミュレータ研究開発センター
Earth Simulator Research and Development Center

強力なツールとして，数値シミュレーションによる方法が非常に重要であるとされている．すなわち，計算地球科学分野においては，数値シミュレーション精度が大幅に向上できれば，現在シミュレーションで捉えられていない種々の現象が解明されることが期待されている．このため，気象・気候分野の代表的なシミュレーションにおいて，約 5Tflop/s の実効性能を有する大規模並列計算機システム開発が目標とされた．

宇宙開発事業団と日本原子力研究所は，この報告書に基づき平成 13 年度の完成を目標に地球シミュレータの開発を開始したところである．現在，地球シミュレータ開発は，ハードウェアの基本設計が終了し，本体製作に必要な要素技術の試作評価が進められている．

開発過程においては，設計されたハードウェア性能の予測・評価や設計へのフィードバックを行うため，種々のプログラムに対する地球シミュレータ上での実効性能の評価が非常に重要である．また，応用ソフトウェア開発に資する上でも重要である．このため，地球シミュレータ用性能評価システムの開発を行っている．本稿では，まず地球シミュレータの概要について述べ，ハードウェアの主要部分の動作をシミュレートするソフトウェアシミュレータの開発について述べる．さらに，ソフトウェアシミュレータの一部の機能について，その動作検証とそれを用いた地球シミュレータの性能評価について述べる．

2. 地球シミュレータの概要

地球シミュレータの目標性能と完成時期の技術動向を踏まえ，基本設計を実施した．ここでは，基本設計に基づく地球シミュレータの概要について述べる☆．

地球シミュレータは，640 台の計算ノードをクロスバネットワークで結合させた分散メモリ型並列計算機である．各計算ノード（PN:Processor Node）は，ピーク性能 8Gflop/s のベクトル型計算プロセッサ（AP:Arithmetic Processor）8 台が主記憶装置 16GB を共有する共有メモリ型並列計算機となっている．したがって，全体では AP が 5120 台，ピーク性能は 40Tflop/s，主記憶容量 10TB となる（図 1）．各 PN は，8 台の AP，32 台の主記憶ユニット（MMU:Main Memory Unit），リモート制御装置（RCU:Remote Control Unit）及び入出力プロセッサ（IOP:I/O Processor）から構成されている（図 2）．主要な LSI には CMOS テクノロジーを，冷却方式には空冷を採用した．

AP では，ベクトル処理部（VU:Vector Unit），スカラ処理部 (SU:Scalar Unit)，プロセッサネットワークユニット（PNU:Processor Network Unit）及びアドレス制御部（ACU:Address Control Unit）を 1 つの LSI 上に実装し，主要クロック周波数 500MHz で動作させる（図 3）．SU は，4 ウェイのスーパースカラであり，128 個の汎用レジスタ，2 ウェイセットアソシアティブ方式の命令キャッシュとデータキャッシュをそれぞれ 64KB づつ実装している．また分岐予測や投機実行の機能を持つ．VU は，6 種類（加算，乗算，除算，論理，ビット列論理，ロード/ストア）のベクトル演算器と 72 個のベクトルレジスタからなるベクトル演算器セット 8 個で構成され，最大 8Gflop/s の性能

☆ ここで述べるハードウェアの構成は，今後の開発の進展により変更される可能性がある．

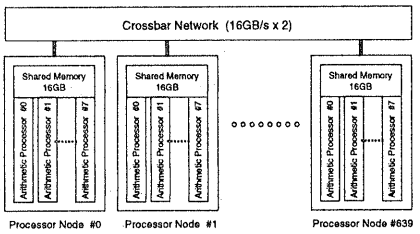

図 1 地球シミュレータの全体構成
Fig. 1 Configuration of the Earth Simulator

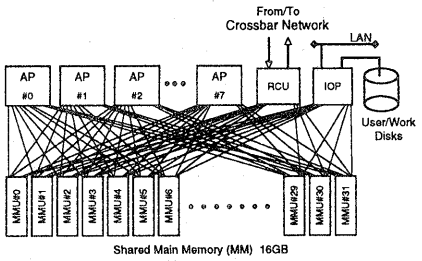

図 2 プロセッサノードの構成
Fig. 2 Configuration of the processor node (PN)

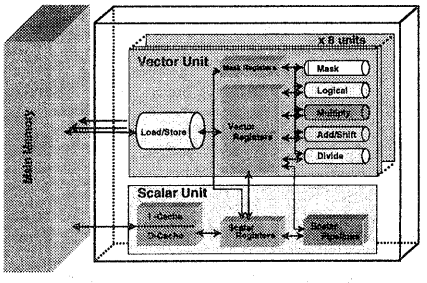

図 3 計算プロセッサの構成
Fig. 3 Configuration of the arithmetic processor (AP)

を有している．32台のMMUには，主記憶素子としてDRAMベースの128Mbitの高速RAMを採用し，2048バンク構成とした．各々のAPは，主記憶システムとの間に32GB/sのバンド幅を持っており，1PNで256GB/sを確保している．理想的には，主記憶素子としてSSRAM，演算性能とメモリスループットの比は1対1が望ましいが[2]，地球シミュレータのハードウェア規模と目標性能を考慮し，コストパフォーマンスを高めるために，主記憶素子として高速RAM，演算性能とメモリスループットの比を2対1とした．この構成において，計算地球科学分野の連続体モデルに基づく応用ソフトウェアや基本カーネルの処理性能を検討した結果，目標性能を達成できることを確認している[3]．

RCUはクロスバネットワークと直接接続され，クロスバを介した送信，受信をAPと独立に動作させることができる．クロスバネットワーク（IN:Internode Network）は，複数のセルフルーティングクロスバスイッチを並列に実装したデータパス部（XSW:Internode Crossbar Switch）と制御部（XCT:Internode Crossber Control Unit）から構成されている（図4）．XCTは，PN間の転送経路の予約，競合調停等を行い，XSWを介して実際のデータが転送される．IN及びRCUは，PN間データ転送機能，PN間同期等の機能を持つ．データ転送機能では，同期型転送と非同期型転送がサポートされており，主記憶上の連続した領域のデータを転送するブロック転送の他に，非同期型転送としてストライド付きベクトル転送，リストベクトル転送が可能である．データ転送性能は，レイテンシがソフトウェアのオーバヘッド込みで3〜6μsec，スループットはハードウェアの最大性能で送信，受信とも16GB/sである．

一方，地球シミュレータ本体の運用，管理を行うために，16台のPNをクラスタと呼ぶ論理単位に分割している（図5）．各クラスタに，クラスタ制御装置（CCS）を設置し，クラスタに属するPNの電源投入及び切断等各クラスタ内の制御を行う．全体で40台のCCSのうち，1つのCCSを特にスーパークラスタ制御装置（SCCS）とし，全クラスタの運用，管理を統括させる．

地球シミュレータのシステムソフトウェアは，既存のUnix系オペレーティングシステム（OS）をベースに地球シミュレータのクラスタ構造を管理，制御するための種々の機能を実現する予定である．特に，大規模分散並列処理に対応するために，高速ファイルアクセスが可能なファイルシステムを導入するとともに，並列ファイル入出力機能を提供する予定である．

運用におけるジョブ配置では，SCCSのあるクラスタをTSSクラスタ，その他の39個のクラスタをバックエンドクラスタとし，TSSクラスタでは会話型処理，小規模バッチ処理を行い，バックエンドクラスタでは，大規模バッチ処理を行う．地球シミュレータ上の並列プログラミングでは，AP内ベクトル処理，PN内並列処理（共有メモリ型）及びPN間並列処理（分散メモリ型）が可能な3階層の並列実行機能を有効に活用できる環境を用意する予定である．基本的なプログラミングスタイルとして，Fortran90またはCをベースにメッセージパッシングライブラリMPI2を利用した並列処理形態を用意する．また，ユーザの利便性と従来のプログラム資産を考慮し，HPF2を拡張した言語体系も用意する予定である[4),5)]．

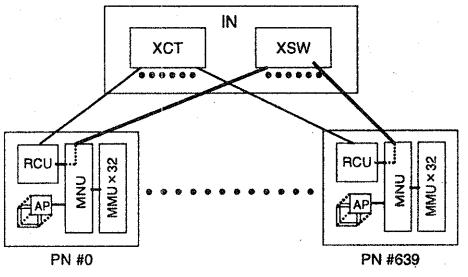

図4　クロスバスイッチの構成
Fig. 4　Configuration of the crossbar switch (IN)

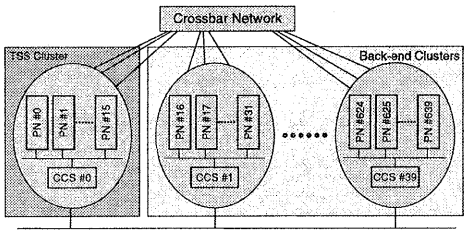

図5　クラスタ構成の概念
Fig. 5　Concept of cluster system in operation

3. 地球シミュレータ用性能評価システム

地球シミュレータの性能を予測するために，地球シミュレータの主要な部分の動作を忠実にシミュレートするソフトウェアシミュレータ(GSSS)を開発した．GSSSは，地球シミュレータの動作だけでなく，同様のアーキテクチャを持つマシン，例えばNEC SX-4などの動作も入力パラメータを変更することによってシミュレートできる．

本節では，GSSSの概要，性能推定に用いたベンチマークプログラム，及びGSSSの一部の機能であるGSSS_APの精度検証結果について述べる．

3.1 ソフトウェアシミュレータ (GSSS) の概要

地球シミュレータの性能を予測する上で重要な構成要素は，AP 部，メモリ部，及び結合ネットワークを介したデータ転送部である．このため，GSSS は，AP 部分をシミュレートする GSSS_AP，複数の計算プロセッサからのメモリアクセス動作をシミュレートする GSSS_MS，結合ネットワーク部分におけるクロスバスイッチからの非同期データ転送のタイミングシミュレーションを行う GSSS_IN の 3 つの部分から構成される (図 6)．

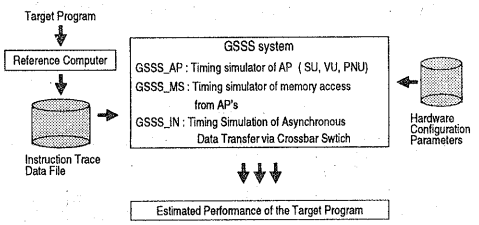

図 6 GSSS システムの構成
Fig. 6 Configuration of the GSSS system

これらの各部分は，入力パラメータとして，ベクトルパイプライン数，AP 台数などのハードウェア構成情報やクロック周波数，メモリアクセスレイテンシ，バンク構成，結合ネットワークのレイテンシ，スループット等のハードウェアコンフィグレーション情報を与えることが可能である．以下では，GSSS_AP について詳細に説明し，計算プロセッサ単体の性能推定について議論する．

GSSS_AP は，スカラ処理部 (SU), ベクトル処理部 (VU) とプロセッサネットワークユニット (PNU) の一部のシミュレータから構成され，単一 AP の動作，特にベクトル命令の処理動作に関するタイミングシミュレーションを，3 種類の入力データを用いて行う．すなわち，ハードウェアコンフィグレーションファイル，命令データベースファイルの記述を動作パラメータとし，プログラムから生成した機械語の命令トレースデータを入力として，命令実行をシミュレートし，実行に要するクロック数を出力する．

ハードウェアコンフィグレーションファイルでは，ベクトル演算器の数や種類，ベクトルレジスタの容量等ハードウェアの物量を記述する．命令データベースファイルでは，各アセンブラ命令毎にベクトル演算器のレイテンシやスループット等を記述する．これら 2 つのファイルは，推定対象のハードウェア毎に固定的に用意されるものであり，パラメータを変更することにより地球シミュレータに類似した処理方式，構成，命令体系を有するマシンのシミュレーションを行うことができる．

GSSS_AP の入力としての機械語の命令トレースデータは，地球シミュレータが SX シリーズのアーキテクチャと類似しているので，SX-4 の Fortran90 コンパイラで作成したロードモジュールを SX-4 上で実行させたときに出力されたトレースデータを用いることにした．ただし，地球シミュレータの性能を発揮させるための命令の並べ替え等は考慮している．以下では，GSSS_AP 内の各部の動作についてさらに詳しく述べる．

SU では，演算器構成，各演算器のレイテンシなどが全てパラメータ化されており，指定したデコードレートや命令発行レートに従って命令処理を行なう．また，アドレス加算器，整数演算器，浮動小数点演算器のショートカットパスが指定可能であるとともに，命令間のレジスタ依存関係（データ依存関係）についてもハードウェアと同様に管理する機能を備えている．ただし，GSSS_AP ではベクトル性能の評価に重点がおかれているため，キャッシュや分岐予測など複雑な処理の機能は備えていない．したがって，キャッシュミスヒット等シミュレートできない部分に要する時間は，SX-4 で実測した時間をクロック周波数などの相違点を考慮して補正し，処理時間に加えることとした．

VU では，各種ベクトル演算器，ベクトル演算レジスタ，ベクトルデータレジスタ，ベクトルマスクレジスタの動作をシミュレートする．ベクトル演算器のレイテンシ，スループット等はパラメータで指定される．また，ベクトルレジスタやベクトル演算器などの各リソースのビジー状態を管理しており，命令発行時にその命令が使用するすべてのリソースがビジーでないことを確認してから処理を行う．ビジー状態のセット時間はリソース毎に指定できるとともに，チェイニング動作等に対してもシミュレートできる．

PNU では，メモリアクセスレイテンシおよびメモリアクセスレートを指定できる．シミュレーションでは，指定したメモリアクセスレイテンシを用い，命令発行後一定時間後にメモリ部からリプライが返ると仮定する．メモリアクセスレートは，メモリアクセスの要素間距離によって変化するので，命令トレースデータの情報を用いてメモリアクセスレートを計算する．特にリストベクトル処理では，命令トレースデータの当該処理の部分に理論的なメモリアクセスレートを指定して行う．また，バンクサイクルによるバンクアクセスの遅れも考慮に入れられている．

3.2 カーネルループベンチマーク

地球シミュレータの性能を評価するために，地球科学分野における代表的なシミュレーションコードである CCSR/NIES AGCM(Atmospheric General Circulation Model) や POM(The Princeton Ocean Model) などの応用ソフトウェアの中から幾つかの特徴的な

DO ループを選び，ベンチマーク用のカーネルループとした．

CCSR/NIES AGCM は東京大学気候システムセンタ (CCSR) と国立環境研究所 (NIES) で共同開発されたスペクトル法を用いた大気大循環モデルであり，POM はプリンストン大学で開発された有限差分法を用いた海洋大循環モデルである[6),7)]．

これらのカーネルループは，単一プロセッサでの実行のみを対象にしており，3 つのグループに分類される．すなわち，Group A は四則演算のみを含む単純なループ群，Group B は IF 文あるいは組込み関数を含むループ群，Group C は配列へのメモリアクセスとして間接アクセス（リストベクトル）を含むループ群である．なお，これらのカーネルループは，配列サイズ（メモリ使用量）や最内側ループ長をパラメータとして指定でき，メモリ上の配列の位置，ベクトル長などを変化させたときの実効性能への影響を評価できるようになっている．本検討では，メモリ使用量を 128MB，最内側の平均ループ長を 256 に固定して評価を行った．

カーネルループの浮動小数点演算 (flop) 数は以下のように求めた．一般に flop 数は，対象とするマシンの浮動小数点演算命令の種類やコンパイラの最適化レベルなどによって異なる．したがって，ソースレベルで flop 数をカウントすると実際に処理される命令数よりも多くなる場合があり，実効性能（flop/s）値がピーク性能を越えてしまう場合がある．本ベンチマークでは，SX-4 の Fortran90 コンパイラが生成するオブジェクトリストについて，最適化を考慮しつつ解析を行い，標準的な最適化を想定した flop 数を算出した．

以下では，Group A に属する 21 個のカーネルループを対象にして GSSS_AP の精度検証を行った結果について述べる．

3.3 GSSS_AP の精度検証

GSSS_AP の推定精度を検証するために，カーネルループを SX-4 の単一 AP で実行したときの実行時間（T_{SX4}）と，SX-4 のハードウェア構成をパラメータとした GSSS_AP でシミュレーションした推定実行時間（$T_{GSSS(SX4)}$）とを比較した．表 1 に実測時間と推定時間を示す．最左欄の番号は Group A のループ番号である．GSSS_AP では，ループ処理に要するクロック数が出力されるので，SX-4 のクロック 8nsec を掛けた推定時間に，SX-4 で実測したキャッシュミスヒット時間を加算したものを推定時間とした．表からほぼ等しい処理時間が得られていることがわかる．

また，式 (1) で計算した実測時間と推定時間の相対誤差 E_r (%) を，表 1 の最右欄に示した．

$$E_r = \frac{T_{GSSS(SX4)} - T_{SX4}}{T_{SX4}} \times 100 \ (\%) \qquad (1)$$

この結果，半分以上のループに対して絶対値で 1% 以下の相対誤差が得られた．相対誤差の絶対値最大のものは，ループ A16 の −6.89% である．A16 の最内側ループでは，ベクトルロード命令 5 個，ベクトルストア命令 7 個，ベクトル演算命令 2 個の処理が行なわれ，メモリアクセスに対して演算量が少ない．GSSS_AP では，MS 部へのアクセス時間について一定の値を仮定しているため，メモリアクセスの多いループでは精度が低下しているものと考えられる．

Group A 全体としては，相対誤差の絶対値の平均で，1.38% と良好な推定結果を得ることができ，実効性能の予測が十分行えることが明らかとなった．

表 1 Group A の処理時間と相対誤差
Table 1 Execution times for the Group A and theirs relative errors

No.	T_{SX4} (μsec)	$T_{GSSS(SX4)}$ (μsec)	E_r (%)
A01	2.941	2.918	-0.78
A02	1.472	1.464	-0.53
A03	4.728	4.703	-0.52
A04	1.200	1.208	0.67
A05	91.114	94.137	3.32
A06	10.982	11.013	0.28
A07	3.968	3.856	-2.81
A08	2.439	2.368	-2.90
A09	3.120	3.115	-0.16
A10	2.954	2.833	-4.11
A11	3.977	3.969	-0.20
A12	2.347	2.321	-1.10
A13	1.388	1.384	-0.26
A14	2.174	2.149	-1.17
A15	1.411	1.401	-0.69
A16	1.960	1.825	-6.89
A17	0.512	0.512	0.02
A18	27.185	26.803	-1.41
A19	7.089	7.074	-0.21
A20	6.465	6.442	-0.36
A21	2.688	2.672	-0.58

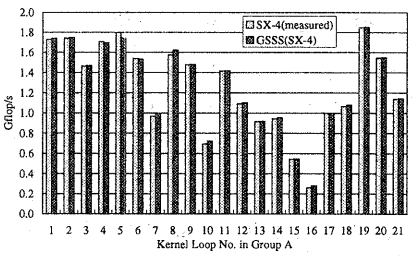

図 7 SX-4 に対する Group A の実測性能と推定性能
Fig. 7 Measured and estimated performance of Group A for the SX-4

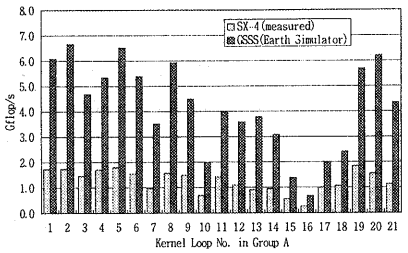

図 8　地球シミュレータに対する Group A の性能推定
Fig. 8　Estimated performance of Group A for Earth Simulator

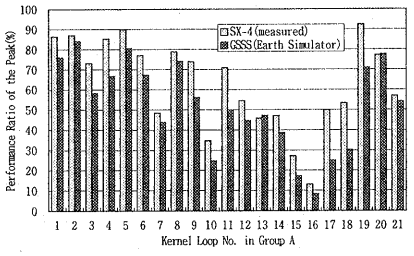

図 9　ピーク性能に対する実効性能に比
Fig. 9　Performance ratio of the peak performance

各ループの flop 数と経過時間から計算した flop/s 値を図 7 に示す．ループ A19 では，SX-4 の単体プロセッサのピーク性能 2Gflop/s と比較して約 90%の性能が得られている．A16 は最も flop/s 値が低いが，これは浮動小数点演算の比率が低いためである．平均で 1.26Gflop/s の実効性能が得られた．

4. 地球シミュレータの性能推定

GSSS に対して，地球シミュレータのハードウェア構成と命令データベースファイルを設定し，Group A に対する性能推定を行った．地球シミュレータの性能推定値を図 8 に示す．また，SX-4 の実測値に基づく実効性能を比較のために示した．地球シミュレータの単一 AP のピーク性能（8Gflop/s）の向上に伴い，実効性能が高くなっている．Group A の平均実効性能は，SX-4 で 1.26Gflop/s であるのに対し，地球シミュレータで 4.18Gflop/s が得られ，約 3.3 倍の性能向上が得られた．

また，SX-4 と地球シミュレータのピーク性能に対する実効性能の比率を図 9 に示す．SX-4 の対ピーク性能比は平均で約 63%，地球シミュレータでは約 52%であり，メモリ構成と演算性能/メモリスループット比を考慮すれば，約 9%の性能低下に留まり，地球シミュレータの構成で十分な性能が得られることが分かった．

5. ま と め

本稿では，地球シミュレータのハードウェア構成と，その性能を推定するための性能評価システムについて述べた．Group A の単純なループ群に対して，ソフトウェアシミュレータ GSSS の推定精度の検証を行った結果，GSSS が SX-4 の実効性能を十分な精度で予測できることがわかった．また，GSSS を用いた地球シミュレータの単一 AP の推定性能は，平均で 4.18Gflop/s と十分な性能が得られた．

今後は，カーネルループの他のグループに対する性能評価や結合ネットワーク転送を含む並列計算を行うプログラムに対する性能評価を実施するとともに，スカラ並列計算機との比較を行う予定である．

参 考 文 献

1) 「地球シミュレータ」計画の推進について，科学技術庁計算科学技術推進会議，平成 9 年 7 月．
2) 西川 岳, 萩原 孝, 安藤憲行, 磯部洋子, “スーパーコンピュータ SX-4 におけるデータ供給能力,” 情報処理, Vol.38, No.6, pp.472-478 (1997).
3) M. Yokokawa, et al., “Performance Estimation of the Earth Simulator,” Proceedings of the ECMWF Workshop, November (1998).
4) “High Performance Fortran Language Specification Version 2,” High Performance Fortran Forum, January (1997).
5) “HPF/JA1.0 仕様,” http://www.tokyo.rist.or.jp/jahpf/spec/jahpf-j.htm/.
6) A.Numaguti, et al., “Description of CCSR/NIES Atmospheric General Circulation Model,” CGER's Supercomputer Monograph Report, Center for Global Environmental Research, National Institute for Environmental Studies, No.3, 1-48 (1997).
7) G. L. Mellor, “Users guide for a three-dimensional, primitive equation, numerical ocean model(June 1996 version),” Program in Atmospheric and Ocean Sciences, Princeton University (1996) (http://www.aos.princeton.edu/WWWPUBLIC/htdocs.pom).