1. 地球シミュレータ・システム

地球シミュレータのハードウェア

NEC
幅田 伸一
s-habata@ay.jp.nec.com

地球シミュレータ研究開発センター
（現在 産業技術総合研究所グリッド研究センター）
横川 三津夫
m.yokokawa@aist.go.jp

海洋科学技術センター 地球シミュレータセンター
北脇 重宗
kitawaki@jamstec.go.jp

開発ターゲット

地球シミュレータシステムは，地球環境問題をコンピュータシミュレーションにより解明するためのツールとして開発した超大型スーパーコンピュータシステムである．開発当初（1997 年）に設定した処理性能の到達目標は，地球上の大気の変動をシミュレートする大気大循環モデルを当時使用されていたスーパーコンピュータ（米国 Cray 社の C90）が処理できる規模の 1,000 倍以上（5TFLOPS 以上）に高めることとした．**表 -1** に示すように，大気大循環シミュレーションの 1 つである CCM2 を実行した時のベクトル型スーパーコンピュータとスカラ型スーパーコンピュータのピーク演算性能に対する実効性能の比率は，ベクトル型スーパーコンピュータが優れており，地球シミュレータはベクトル型とした．

大気大循環モデルを使用したシミュレーションのモデル規模と実効性能の関係より，1,000 倍以上の規模のモデルを実行するために必要な主記憶容量を求めた結果が**表 -2** である．システムに必要な主記憶容量は 8TB 以上，当時使用可能な高速メモリ素子の中で要求仕様に最も近い高速 RAM をベースとした地球シミュレータ専用のフルパイプラインメモリ（128Mbit，8 バンク，動作周波数 133MHz）を開発し，10TB の主記憶を実現した．

実効性能 5TFLOPS を満足するシステムのピーク演算性能は，表 -1 の実行効率よりプロセッサ数が増加することによる効率の低下を考慮し，実行効率を 12.5% と想定し，40TFLOPS とした．

プロセッサ

ピーク演算性能 40TFLOPS，主記憶容量 10TB のスーパーコンピュータをいかに実現するか．これは開発／製造を委託した NEC が，当時（1997 年）プロセッサのピーク演算性能 8GFLOPS の SX-5（1998 年

研究機関	使用計算機	プロセッサ数	T42L18 実行時	T170L18 実行時
NCAR	CRAY C90（ベクトル型）	1	362MFLOPS（効率 38%）	400MFLOPS（効率 42%）
		16	4,200MFLOPS（効率 28%）	5,300MFLOPS（効率 35%）
	CRAY T3D（スカラ型）	64	608MFLOPS（効率 6.3%）	
	TM CM-5（スカラ型）	256	628MFLOPS（効率 1.9%）	
		512	742MFLOPS（効率 1.1%）	
Oak-Ridge & Argonne	Intel Paragon（スカラ型）	512		1,710MFLOPS（効率 4.4%）
		1,024		3,181MFLOPS（効率 4.1%）
	IBM SP2（スカラ型）	128		2,270MFLOPS（効率 6.6%）

1999 年 11 月 ESRDC 報告より

表 -1 Community Climate Model Version 2（CCM2）の実行効率

例	大気大循環モデル	現在	地球シミュレータ	計算量比
経緯度	グローバルモデル	50 ～ 100km	5 ～ 10km	約 100 倍
メッシュ	地域モデル	20 ～ 30km	1km	数百倍
鉛直階数		数十	100 ～ 200	～ 10 倍
時間積分メッシュ		1	1/10	約 10 倍

グローバルモデルにおける必要計算量比（現用計算機に対する） 数千倍

必要メモリ容量 約 8TB

$$4{,}000 \times 2{,}000 \times 200 \times 300 \times 2 \times 8 = 7.68\text{TB}$$

↑ グリッドの数　↑ 階数　↑ バイト / 語

グリッドあたりのデータ量（語）

表 -2 地球シミュレータが具備すべき要件

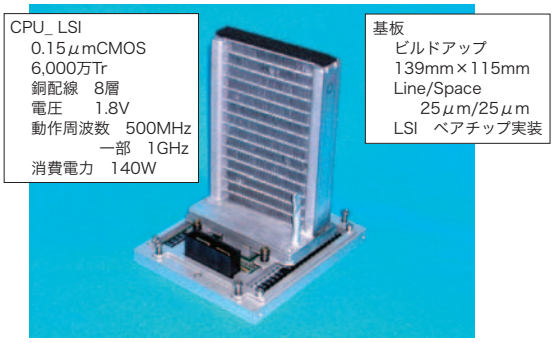

図 -1 プロセッサ

発表）を開発していたことから，プロセッサのピーク演算性能は 8 ～ 16GFLOPS，プロセッサ数は 2,560（16GFLOPS/CPU 採用時）から 5,120（8GFLOPS/CPU 採用時）とした．使用できる LSI テクノロジ（0.15 ～ 0.18μmCMOS）の集積度，ゲート速度からプロセッサの物理サイズを概算，SX-5 の開発に NEC が用いた 32 個の LSI を使用してプロセッサを実現するアプローチを採用し 16GFLOPS のプロセッサ 2,560 個のシステムとするか，1 チッププロセッサを実現するため SX-5 と同じピーク演算性能 8GFLOPS のプロセッサ 5,120 個のシステムとするかを検討した．その結果は，前者と後者のプロセッサカードのサイズ比が 3：1，システム全体のサイズとしては後者が優れていると判断，8GFLOPS の 1 チッププロセッサ 5,120 個からなるシステムの検討を進めた（**図 -1**）．

5,120 台のプロセッサをどのように接続するか．プロセッサ間を接続するネットワークの規模から，8 ～ 16 台のプロセッサが主記憶を共有するプロセッサエレメント（地球シミュレータでは計算ノード（PN）と呼ぶ）をクロスバースイッチに接続する構成とした．計算ノード間を接続するネットワークを多段スイッチではなく，クロスバースイッチとしたのは，大気以外にもいろいろな地球環境問題を解明するためのシミュレーションを計画しており，多段のネットワークではプログラムにより適／不適がハッキリするのに対し，クロスバースイッチはすべての計算ノード間の論理的な距離が均一であり，プログラムを開発する上で使いやすいシステムを実現できると考えたからである．

計算ノード

計算ノードの構成をどうするか．8 プロセッサが主記憶を共有する計算ノード 640 台をクロスバースイッチに接続する構成とするか，16 プロセッサが主記憶を共有する計算ノード 320 台をクロスバースイッチに接続する構成とするか．プロセッサカードと主記憶カード間の接続，実装上の配置を検討し，8 プロセッサ構成とするとと CPU カードと主記憶カードを対向配置にできるのに対し，16 プロセッサ構成では四面配置となり，設置面積が 2 倍以上増加することが分かり，8 プロセッサ構成とした．

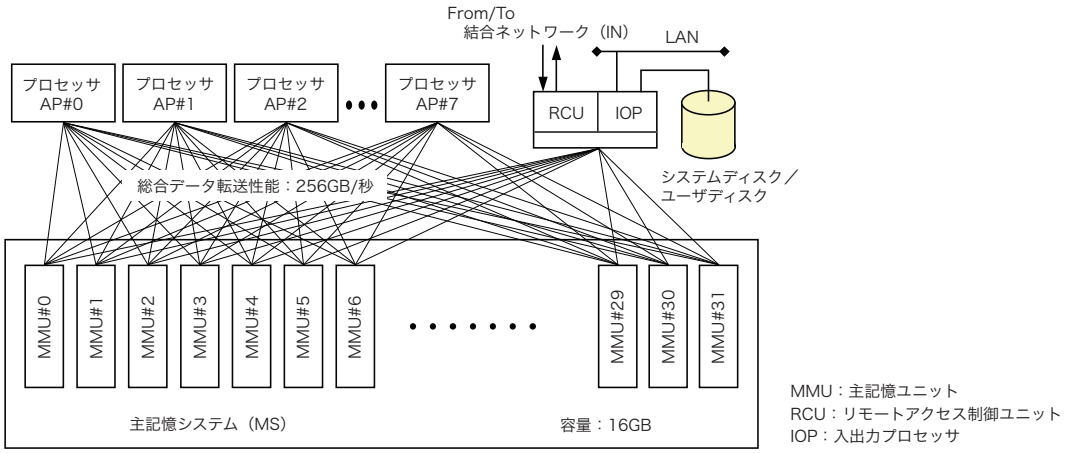

図 -2　計算ノード（PN）の構成

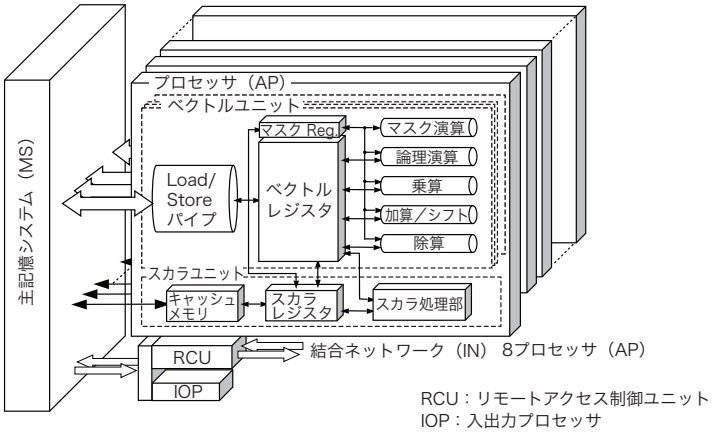

図 -3　プロセッサ（AP）の構成

計算ノードの詳細な構成を**図 -2** に示す．8 台のプロセッサが 32 個の主記憶ユニット（MMU）に接続されており，主記憶にインタリーブ方式を採用することにより，プロセッサあたり 32GB/ 秒のデータアクセスを可能とした．主記憶は MMU あたり 64 個，全体で 2,048 個のバンクに分かれており，8 台のプロセッサが同時に 32GB/ 秒のデータアクセスを行うことができる 256GB/ 秒の総合データ転送能力を備えている．計算ノード外部とのインタフェース制御を行うユニットとして計算ノード間の通信制御を行うリモートアクセス制御ユニット（RCU）と入出力プロセッサ（IOP）がある．この 2 つのユニットは主記憶アクセス制御部を共用しており，主記憶と 2 つのユニット間のデータ転送能力は 16GB/ 秒である．

プロセッサ（AP）の構成を**図 -3** に示す．ベクトルユニットとスカラユニットから構成され，ベクトルユニットには論理演算，乗算，加算／シフト，除算の 4 種の演算パイプラインセットがある．乗算と加算／シフト・パイプラインセットは，各々が 1 システムクロックサイクル（2 ナノ秒）あたり 8 組の演算処理を受け付けることができ，プロセッサとして 8GFLOPS の演算能力を備えている．この他に，演算結果の条件判定を高速に実行するためのマスク演算用のパイプラインセットがある．

計算ノードの実装構造は装置サイズを小さくするため 2 つのノードを 1 台の筐体に実装する方式とした．電源，冷却などのユニットを 2 つのノードが共有することにより，個々のノードを独立した筐体に実装した場合よりシステム全体の設置面積を小さくできるからである．地球シミュレータシステムは 320 台の計算ノード筐体から構成される．

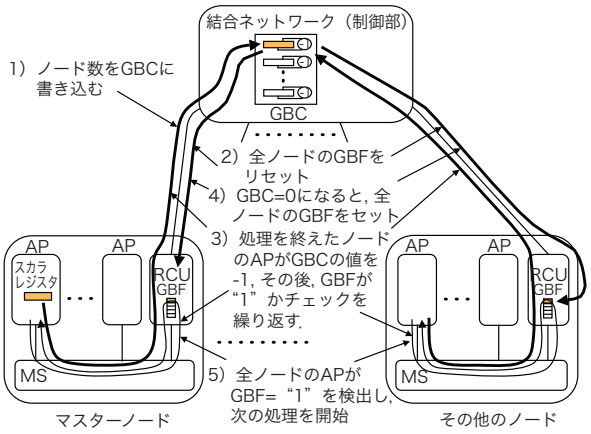

図 -4 バリア同期機構（GBC/GBF）の動作

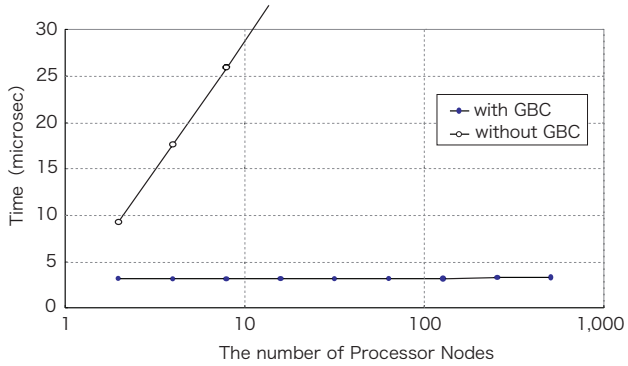

図 -5 MPI_Barrier の実行時間

結合ネットワーク

640 台の計算ノードを接続するクロスバースイッチ（地球シミュレータでは結合ネットワーク（IN）と呼ぶ）をどのように構成するか．計算ノード（以後，ノードと呼ぶ）あたりのデータ転送性能は，ノード内のプロセッサが主記憶からデータをアクセスする性能（32GB/ 秒）の 1/4 以上（8GB/ 秒以上）を開発目標とした．使用可能な信号伝送技術の中から 1.25Gbps の電気インタフェースをノードあたり 128 本使用することにより，ノードあたり 12.3GB/ 秒のデータ転送能力を実現している．

IN の総合データ転送能力は 7.87TB/ 秒である．このデータ転送能力を実現するため，データ転送系のケーブル 81,920 本，制御系のケーブル 1,280 本，合計 83,200 本のケーブルを IN に接続する必要があった．

640 × 640 のクロスバースイッチかつ，接続されるケーブルの本数 83,200 本の IN はデータ系をバイトスライス方式により 128 のユニットに分割，2 つのユニットを 1 台の筐体に実装する構造を採用した．これにノード間のトラフィック制御を行うユニットを実装した 1 台の筐体を加え，IN は 65 台の筐体，14m × 13m のフロアを必要とする巨大な装置になった．

ノード間インタフェースに使用した 83,200 本のケーブルの総長は約 2,400km，重量は約 140t になり，ケーブル敷設に約 3.5 カ月を要した．

データ系に 81,920 対のシリアルインタフェースを使用することから，シリアルインタフェースにおけるデータエラーの発生を考慮した装置設計を行う必要があった．それは，ノード間を転送するデータに ECC（Error Check and Correction）コードを付加し，受信側ノードが ECC コードをチェックすることにより，データ転送中のエラーを検出，軽微なデータエラーの場合は受信ノード側でエラー訂正を行うことにより，データエラーによるノード間通信異常を回避することである．IN がバイトスライス方式により 128 のユニットに分割されているため，ECC の生成／チェック／訂正はすべてノード内の RCU により行われている．

640 台のノードを構成する 5,120 台のプロセッサが備える 40TFLOPS のピーク演算性能を有効に使うため，640 台のノードが同期を確保するための専用ハードウェア Global_Barrier_Counter（GBC）と Global_Barrier_Flag（GBF）を用意した．GBC は IN の制御部にある 128 要素のカウンタである．GBF は全ノードの RCU 内にあり，GBC と同じ 128 要素からなる．GBC と GBF によるバリア同期の動作を**図 -4** に示す．

並列処理に使用するノード数がマスターノードの AP 内（スカラレジスタ）にセットされており，その AP がスカラレジスタの値（ノード数）を IN 制御部内の GBC にセットする．IN 制御部は，ノード数をセットされた GBC に対応する GBF をリセットする．GBF は全ノードの RCU 内にある．並列処理に使用されているノードは個々の処理を行い，バリア同期ポイントに達すると IN 制御部内の GBC をディクリメントし，GBF が“1”になるまでポーリングする．次々とノードが同期ポイントに到達し，IN 制御部内の GBC をディクリメントし，GBC の値が“0”になると，IN 制御部は全ノードの GBF をセットし，全ノードに同期ポイントに達したことを通知する．

ノードは GBF が“1”になると次の処理を開始する．**図 -5** に GBC/GBF を使用した場合と使用しない場合の MPI_Barrier の実行時間を示す．GBC/GBF を使用した

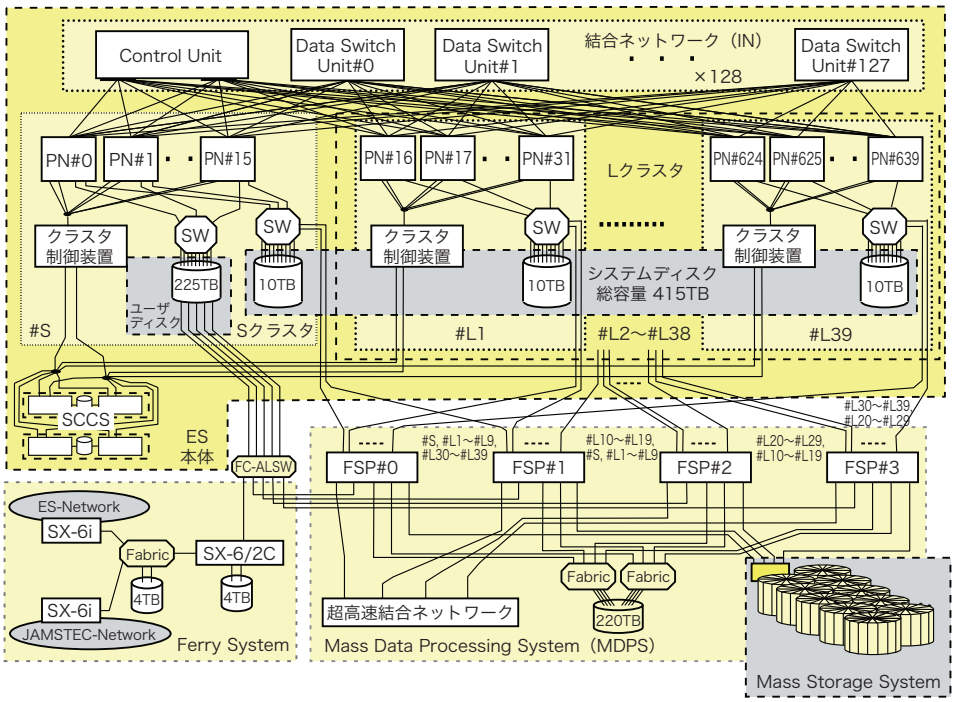

図 -6　地球シミュレータシステムの構成

計算ノード（PN）		プロセッサ数　8 ピーク演算性能　8GFLOPS × 8cpu ⇒ 64GFLOPS 主記憶容量　16GB メモリバンド幅　32GB/ 秒 /cpu × 8cpu ⇒ 256GB/ 秒
	PN 筐体	サイズ　1.4m（W）× 1.0m（D）× 2.0m（H） 重量　950kg
結合ネットワーク（IN）		データ転送性能　12.3GB/ 秒 /PN 総合データ転送性能　7.87TB/ 秒
	IN 筐体	サイズ　1.2m（W）× 1.3m（D）× 2.0m（H） 重量　860kg
システム		PN 数　640，　プロセッサ数　5,120 ピーク演算性能　8GFLOPS × 5,120cpu ⇒ 40TFLOPS 主記憶容量　16GB × 640PN ⇒ 10TB
システム設置諸元		フロア面積　1,640m^2（40m × 41m） PN 筐体　320 台，　IN 筐体　65 台 重量　360t（PN 筐体と IN 筐体） 140t（ノード間ケーブル） 消費電力　5,500 ～ 6,000KVA

表 -3　地球シミュレータシステムの諸元

場合，ノード数が増えても実行時間は約 3.3μ 秒と一定であるのに対し，使用しない場合はノード数が増加すると実行時間が急激に増加し，10 ノードあたりでグラフにプロットできなくなっている．

システムの構成

地球シミュレータシステムは，地球シミュレータ本体（ES 本体），MDPS（Mass Data Processing System），MSS（Mass Storage System），Ferry System の 4 システムから構成されている．**図 -6** に地球シミュレータシステムの構成，**表 -3** に諸元を示す．

ES 本体は，640 台のノード（PN）とそれを結合する結合ネットワーク（IN）からなり，640 台の PN は，運用管理単位であるクラスタと呼ぶ 40 のグループに分けている．各クラスタは 16 台の PN，運用管理プログラムが動くクラスタ制御装置（CCS），約 10TB のシステムディスクから構成されている．

40 のクラスタは，1 つの S クラスタと 39 の L クラ

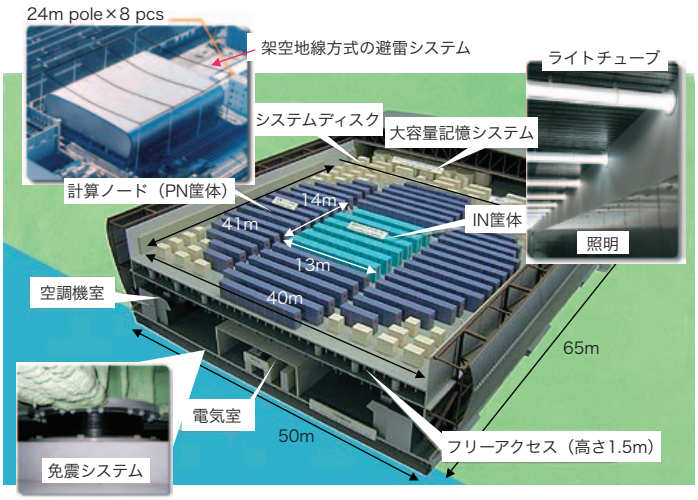

図 -7 シミュレータ棟の構造

スタに分けられ，L クラスタは大規模バッチジョブの実行に，S クラスタはプログラムの開発，小規模ジョブの実行に使用している．

S クラスタはプログラム開発などに使用するため，システムディスクの他に 225TB のユーザディスクを備えている．

40 のクラスタを制御するため，スーパークラスタ制御装置（SCCS）を用意している．SCCS は各クラスタの CCS を管理することにより，40 のクラスタからなる ES 本体を 1 つのまとまったシステムとして動作させている．

MDPS は，容量 1.5PB の MSS が格納するユーザデータを各クラスタのシステムディスクにプリロードしたり，プログラムの実行結果をシステムディスクから MSS に格納する処理を行う．4 台の FSP（File Service Processor）と 220TB のディスクからなり，低速の MSS と ES 本体の間に位置し，MSS 上のデータアクセスを高速化している．

MSS は，1.5PB の容量を備えた，25,000 巻のテープ，96 台のテープドライブからなるテープライブラリ装置である．

Ferry System は，ES 本体，MDPS，MSS が接続されている LAN（ES-LAN）と外部のネットワーク（ES-Network）との間のデータ転送を行うシステムである．3 台のサーバ（1 台の SX-6/2C と 2 台の SX-6i）がディスクを共有し，このディスクを経由して ES 本体が処理した計算結果を外部からアクセス可能にしている．

施　設

シミュレータ棟は地球シミュレータシステムを設置するために建設した専用の建物である．**図 -7** にシミュレータ棟の構造を示す．

電磁ノイズ対策として，アルミめっき鋼板などを使用した三重の電磁シールド構造を採用，システムが発する電磁ノイズおよび外部からの電磁ノイズを遮断している．

落雷によるシステムの誤動作，損傷を防ぐため，建物から独立した高さ 24m の避雷塔を 8 本使い，架空地線方式の避雷システムを採用している．

マシン室の照明に関しては，照明器具が発する電磁ノイズの影響を考慮し，照明の光源はマシン室外に設置するライトガイド方式を採用している．直径 255mm，長さ 44m のライトチューブ 19 本を使用して約 2,600m^2 のマシン室を照明している．

地震対策としては，積層ゴムのアイソレータをシミュレータ棟の床下に配置した免震システムを採用している．

成　果

地球シミュレータは，2002 年 2 月末に稼働を開始，その 2 カ月後に，LINPACK で 35.86TFLOPS の世界記録を達成，さらに，大気大循環モデルでは 26.58TFLOPS の実行性能を記録するなど，数々の成果を生み出しており，今後の活躍が期待されている．

（平成 15 年 12 月 5 日受付）