スーパーコンピュータシステム　地球シミュレータ（ES4）の紹介

中川 剛史，石黒 駿，上原 均，大倉 悟，齋藤 友一，今任 嘉幸，甲斐 恭

海洋研究開発機構　付加価値情報創生部門　地球情報基盤センター

tnakagawa@jamstec.go.jp

Introduction of Supercomputer System Earth Simulator (ES4)

Tsuyoshi Nakagawa, Shun Ishiguro, Hitoshi Uehara, Satoru Okura, Yuichi Saito,
Yoshiyuki Imato, Tadashi Kai

Center for Earth Information Science and Technology (CEIST),
Japan Agency for Marine-Earth Science and Technology (JAMSTEC)

概要

海洋研究開発機構では、地球シミュレータをこれまでの 10 倍以上の演算性能を持つ第 4 世代へと更新し、2021 年 3 月より運用を開始した。本稿では、そのシステムの概要を紹介する。

1　地球シミュレータの概要

海洋研究開発機構（以下、「機構」）が運用するスーパーコンピュータである地球シミュレータは、2002 年に運用を開始した初代地球シミュレータ以降、継続的にシステム更新を行っており、地球科学分野のみならず様々な研究開発分野の年々高まる計算需要に対応している。

この度運用を開始した第 4 世代の地球シミュレータ(ES4)は、複数のアーキテクチャを組み合わせたマルチアーキテクチャ型のスーパーコンピュータである（外観写真：図 1）。将来における機構の計算科学の発展を見据えて、汎用性の高いシステム及び複数のアーキテクチャを備えたシステムとした。また、大規模データの高速処理を行うシステムとの親和性も重視している。

ES4 システムの総合性能は、総演算性能 19.5 PFLOPS（第 3 世代地球シミュレータ ES3 の 15 倍）、総メモリ容量 556.5 TiB（同 1.7 倍）、共有ストレージ容量 61.4PB（同 5 倍）であり、これまでのシステムより飛躍的な性能向上となった。

図 1　ES4 外観

設置および設計については、既設システムを並行稼働させながら、2020 年 9 月より横浜研究所シミュレータ棟へ機器の搬入を開始し、据え付け調整や運用設計を進め、同年 12 月よりユーザデータ 5 PB の移行を差分同期で実施した後、2021 年 3 月 1 日より GPU 搭載ノードを除いた大部分のシステムの運用を開始した（3 月は先行利用期間。GPU 搭載ノード部は 6 月より運用開始）。また並行して、重点プログラムの移植サポートも実施し、速やかに成果を創出できるように利用環境の移行を進めた。

以下では、地球シミュレータの 3 つのノード種別とストレージシステム、運用設定や計算資源の割当について紹介する。

2　計算ノード

計算ノードは、CPU のみを搭載した「CPU ノード」、NEC 社製 SX-Aurora TSUBASA を搭載した「VE 搭載ノード」、NVIDIA 社製 GPU A100 を搭載した「GPU 搭載ノード」から構成される。それぞれのノードは同一のインターコネクトネットワーク(IB HDR　200 Gbps)上で接続されており、複数アーキテクチャにまたがった計算や全ノードを利用した計算の実行も可能となっている（図 2、

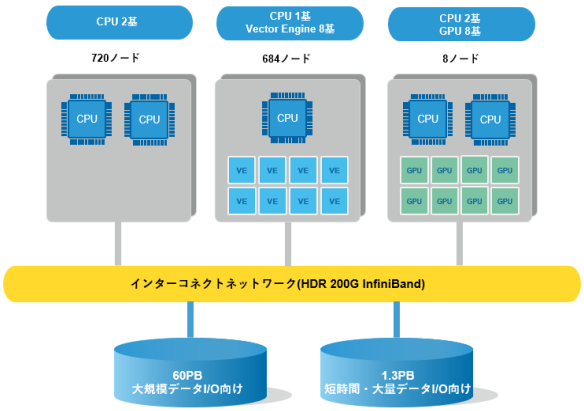

図 2　システム構成イメージ

表 1）。

2.1 CPU ノード部

CPU ノード部は、720 ノードの HPE Apollo2000 から構成され、ピーク性能は 3.3 PFLOPS、メモリ容量は 180 TiB である。各 CPU ノードは、2 基の AMD EPYC 7742 と 256 GiB DDR4 を搭載している。この CPU ノード部は、x86 アーキテクチャに最適化されたアプリケーションや商用アプリケーションなど様々なアプリケーションに対応する。

2.2 VE 搭載ノード部

VE 搭載ノード部は、684 ノードの NEC SX-Aurora TSUBASA B401-8 から構成され、5472 基の Vector Engine により、新システムの中で最も強力な演算性能、最も大きなメモリ容量、最も広いメモリ帯域を提供する。VE 搭載ノード部は、現行地球シミュレータの代表的なワークロードに対応する。

各 VE 搭載ノードは、AMD EPYC 7742 と 128 GiB DDR4 および 8 基の NEC SX-Aurora TSUBASA Vector Engine (VE) Type20B を搭載している。VE は 2.45 TFLOPS の演算性能と 48 GiB の HBM2 を搭載する。各ノードの性能は 21.9 TFLOPS で、全 684 ノードで 14.9 PFLOPS の計算能力を有する。また、新たな利用方法として、ホスト CPU と VE を連携させたハイブリッドジョブ実行も可能となっている。

2.3 GPU 搭載ノード部

GPU 搭載ノード部は、8 ノードの HPE Apollo6500 から構成され、各 GPU 搭載ノードは 2 基の AMD EPYC 7742 と 4 TiB のメモリと 8 基の NVIDIA A100 及び 7 基の 3.2 TB NVMe SSD を搭載する。この GPU 搭載ノード部は、GPU アプリケーションや ML/DL アプリケーションのワークロードのみならず、大容量メモリや高速な I/O 性能が必要なアプリケーションにも対応する。

2.4 インターコネクトネットワーク

すべての計算ノードおよびストレージ、フロントエンド装置は、HDR 200Gb/s の InfiniBand ファブリック（fat-tree 構成）に接続されている。さらに、VE 搭載ノードは演算通信のみで利用するインターコネクトとして、専用の HDR 200 Gb/s InfiniBand ファブリック（3 つの fat-tree で構成される Dragonfly+）で接続されている。

2.5 冷却システム

VE 搭載ノード部の CPU ノード部のプロセッサ部分の冷却方法は、水冷システムを採用している。一方で、GPU 搭載ノード部は、空冷となっている。冷却設備は、計算機室床下に設置されており、計算ノードと熱交換器およびポンプが配管およびホースで接続され、水温 20℃(往)・23℃(復)・流速

表 1　計算ノードの仕様

計算機ノード種別		CPUノード	VE搭載ノード	GPU搭載ノード
ノード数		720	684	8
ノード単体	CPU名	AMD EPYC 7742		
	CPU数(コア数)	2(128)	1(64)	2(128)
	OS	CentOS 8		
	メモリ容量(ホスト)	256GiB	128GiB	4TiB
	アクセラレータ	-	NEC SX-Aurora TSUBASA Type 20B	NVIDIA A100
	アクセラレータ数	-	8VE (VEあたり8コア)	8 GPU
	メモリ容量 (アクセレータあたり)	-	48GiB	40GiB

240L/h で循環する冷却水で計算に伴う発熱を取り除いている。

2.6 異なるアーキテクチャ間の連携

上記の 3 つのアーキテクチャのノード部は、同一の HDR InfiniBand ファブリックに接続されており、ストレージも共有している。また、いずれのノードも AMD EPYC CPU を搭載し、Linux がインストールされており、そのため、全ノードを管理するバッチ処理システムを利用して、複数アーキテクチャにまたがったジョブの実行や全ノードを利用したジョブの実行が可能となっている。

3 ストレージ装置

DDN 社製のストレージで構成され、ユーザ環境を保存する「ホーム領域」、従来のシミュレーションからの大規模データ I/O 向けの「データ領域」、シミュレーションとデータ解析の融合を目指す高速データアクセス用の「ワーク領域」の 3 つの部分から構成されている。ES3 まで採用されていたファイルステージング機能は、ストレージ性能の向上に伴い、ES4 では採用されていない。

ファイルシステムは全て、DDN 社製の Lustre ファイルシステム（DDN EXAscaler）を採用している。

3.1 ホーム領域

ホーム領域は、2 ファイルシステム合計 120TB の容量を有し、利用者のホームディレクトリ用途として、設定ファイルやプログラムのソースコードなどを格納する共有領域である。本領域はフルバックアップを週一度実施している。

3.2 データ領域

データ領域は、合計 60PB の容量を有する ES4 のメイン共有ストレージである。ストレージは 2 パーティション（1 パーティション 30PB）に分割運用されており、パーティション毎に MDS サーバ 8 台、OSS サーバを 20 台構成となっている。パーティション毎の IO 性能は、IO バンド幅 400 GB/s、メタデータ性能として 630 kIOPS を有する。

3.3 ワーク領域

ワーク領域は、NVMe SSD から構成される 1.3PB の容量を持つ共有ストレージである。こちらは 1 パーティションでの運用となり、MDS サーバ 4 台、OSS サーバ 20 台構成となっている。IO バンド幅で 400GB/s、メタデータ性能として 315 kIOPS の性能を有し、特に 4K のランダムアクセス性能では、30M IOPS の性能を有しており、データ解析・ML/DL などの IO インテンシブなワークロードに対応する。

4 ソフトウェア環境

それぞれのアーキテクチャに合わせたプログラミング環境、ライブラリ、プロファイラとデバッガーを中心に利用環境を整備している。CPU 用には、AMD コンパイラ、Intel コンパイラ、GNU コンパイラ等が利用可能で、VE 用には NECSDK、GPU では NVIDIA HPC SDK が利用可能である。利用環境は module コマンドで切り替えて利用できる。

5 運用設定

5.1 利用資源単位

多数の CPU コア、VE および GPU を搭載するノードでシステムが構成されるため、その計算ノード内の計算資源の有効活用を目的として、従来のノード時間での課金に変えて、ノード内を細かく等分割した「リソースセット時間」という単位を用いて課金している。たとえば、VE 搭載ノードではノード内を 8 分割した、8CPU コア＋1VE＝1 リソースセットと定義し、CPU ノードでは 2 分割した 64CPU＝1 リソースセット、GPU 搭載ノードでは 8CPU コア＋1GPU＝リソースセットとして細分化している。また、メモリに関しても、同様に分割した値が利用可能となっている。

ストレージ資源に関しては、課題申請時の希望値を審査の上でクォータ上限として割り当てており、利用量に応じた課金は実施していない。

5.2 バッチシステム

バッチシステムは、NEC の NQSV を採用している。それぞれの計算ノード種別ごとに、インタラクティブに利用可能なインタラクティブキューとバッチスクリプトを用いて利用可能なバッチキューに分けられている。バッチキューは、デバッグ用途のデバッグキュー、小サイズプログラム用の S キュー、高並列実行用の L キューから構成される。また、複数のアーキテクチャを同時に利用可能なマルチアーキテクチャ利用キューを用意している。

ジョブを実行する際には、キュー名、先ほど説明したリソースセットとその個数およびジョブの実行時間の指定が必要となる。

6 計算資源の配分

機構では、競争等による計算資源の追加やチャレンジ利用を推進する仕組みを取り入れ、これらに柔軟に対応するための「機構戦略枠」を令和3年度に導入し、「所内枠」と「公募枠」、「機構戦略枠」の3つの利用枠を設けている（図3）。

6.1 利用枠課題の詳細

・所内課題

機構内の役職員等を対象に募集する課題。主に機構の中長期計画に直結するテーマである。

・公募課題

我が国の海洋地球科学と関連分野の研究を推進するため、広く地球シミュレータ利用の機会を開くもので、機構外を対象に機構が募集する課題。

・機構戦略課題

① チャレンジ利用課題

挑戦的な利用や大型計算機の利用推進等による利用を目的として、機構の内外を問わず募集する課題。機構の中長期計画に直結しないテーマも受け付ける。

② 受託等による利用課題

国等からの委託、補助金等を受け、機構が実施または、第三者に実施させる課題。統合的気候モデル高度化研究プログラムやHPCI（革新的ハイパフォーマンス・コンピューティングインフラ）、成果専有型有償利用の課題が該当する。

6.2 利用課題への資源追加

計算資源の当初配分比率は所内課題で20%、公募課題で10%だが、見込まれる成果をもとに審査を行い、十分な成果が見込まれる所内課題、公募課題等へ優先的に計算資源を充当する仕組み。

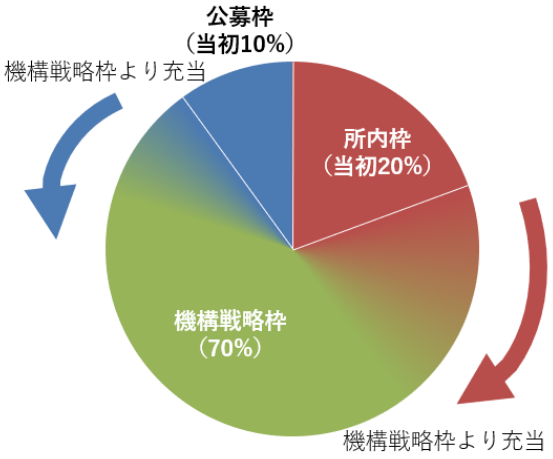

図3　令和3年度計算資源配分

7 まとめ

本稿では、新しい地球シミュレータの概要を紹介した。新システムは、これまでの10倍以上の演算性能とストレージ容量を有し、多種多様なプログラムを活用できるように、マルチアーキテクチャ型のシステムを採用している。汎用性の高いCPUと、これまでの地球シミュレータで培われたソフトウェア資産を活用でき、高い実行性能を持つベクトルプロセッサ、さらにAI研究などで有用なGPUを組み合わせることで、従来研究のさらなる発展とAI研究などの新規研究課題実施の両立を目指している。