地球シミュレータの運用状況

大倉 悟, 中川 剛史, 甲斐 恭, 板倉 憲一

独立行政法人海洋研究開発機構 地球情報基盤センター

情報システム部 基盤システムグループ

okura@jamstec.go.jp

概要：独立行政法人海洋研究開発機構(JAMSTEC)では、2002 年 3 月より世界最大規模のベクトル並列型スーパーコンピュータ「地球シミュレータ」を運用している。2009 年 3 月の地球シミュレータ(ES2)へのシステム更新を経て、2015 年 3 月には、後継システムの運用を開始する予定である。本稿では、地球シミュレータ(ES2)の概要及び運用状況と、後継システムの概要について述べる。

1 はじめに

初代地球シミュレータは、故・三好甫プロジェクトリーダーの元、旧・宇宙開発事業団（現・宇宙航空研究開発機構)、旧・日本原子力研究所(現・日本原子力研究開発機構)、旧・海洋科学技術センター（現・海洋研究開発機構）の三法人によって開発された、世界最大規模のベクトル並列型スーパーコンピュータである。大気大循環モデルの実行性能で 5TFLOPS 以上という開発目標を大きく上回る 26.58TFLOPS を達成するとともに、LINPACK ベンチマークにおいても 35.86TFLOPS を達成した[1]。2009 年 3 月にはシステムを更新し、現在は NEC 製 SX-9E を中心としたシステム(ES2)が運用されている。

本稿では、ES2 の概要と運用状況について紹介する。また、2015 年 3 月に稼働が予定されている、後継システムの概要について紹介する。

2 ES2 の概要

ES2 の概要を、図 1 に示す[2]。160 台の計算ノードは、Fat-Tree 構成のノード間ネットワークにより接続される。計算ノードは、インタラクティブノード(1 台)、プリ・ポスト処理やデバッグ処理用の S 系バッチノード(3 台)、大規模並列処理用の L 系バッチノード(156 台)の、三種類の用途に使い分けられる。各計算ノードには、バッチジョブ実行中のテンポラリ領域としてのワークディスクが接続されており、パーマネント領域であるユーザディスクとの間で、バッチジョブの実行前後にファイルのステージング処理が行われる。

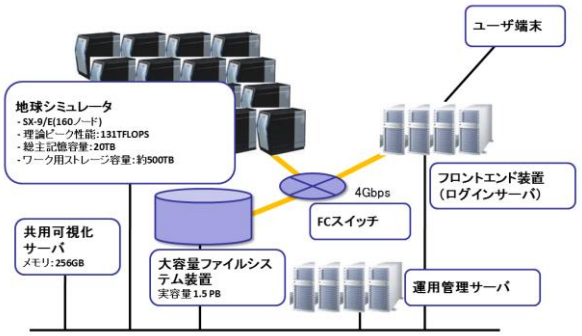

図 1　ES2 システムの概要

2.1 計算資源

ES2 では、計算資源の割当てを、プログラムが使用する計算ノードの数と、それを実行した経過時間の積を用いて行う。例えば、計算ノード 1 台を 1 時間使用した場合、「1 ノード時間」を使用したものとする。L 系バッチノード数に、各年度において最大利用可能な時間数を乗じた値を、配分可能な総ノード時間とする。図 2 に 2014 年度の資源配分割合を示す。

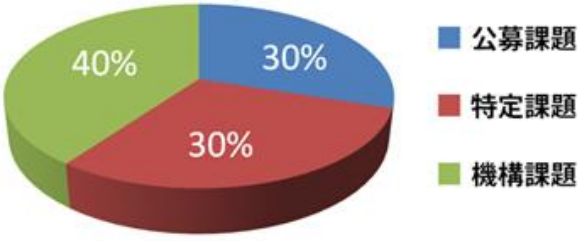

図 2　ES2 の資源配分割合

現在、公募による利用(公募課題)に 30%、国等からの委託・補助等による利用(特定課題)に 30%、

JAMSTEC が主導する研究課題による利用(機構課題)に 40%が充てられている。

2.2 バッチ処理環境と運用管理ソフトウェア

ES2 の運用では、高並列プログラムの効率的な実行と、システム全体の利用率の向上を、同時に実現することを目指してきた。初代地球シミュレータでは、ジョブスケジューラを含む運用管理ソフトウェアを独自開発したが、その基本的な考え方は、現在運用中の ES2 にも受け継がれている。

ES2 におけるプログラムの実行は、バッチ処理システムを介して行われる。バッチ処理システムとして NEC 製「NQSII」が、ジョブスケジューラとして同じく NEC 製「JobManipulator」が使用されている。図 3 に、ES2 におけるバッチ処理概念図を示す。

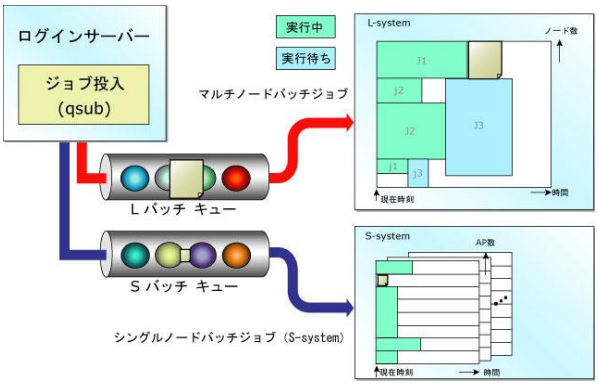

図 3　バッチ処理概念図

2.3 S 系バッチジョブ

S 系バッチノード及び L 系バッチノードは、それぞれ単一のバッチキューにより管理される。利用者は、S 系バッチキューまたは L 系バッチキューのどちらかを選択する。

S 系バッチジョブは、1 台の計算ノード内で実行される。S 系バッチジョブのジョブスケジューリングには、使用する CPU 数と CPU 時間、メモリ使用量が用いられる。S 系バッチジョブの実行時、バッチスクリプト中では以下の指定を行う。

- 使用 CPU 数
- CPU 時間
- 使用メモリ量

S 系バッチジョブでは、8CPU まで使用した並列処理が可能であり、最大使用可能なメモリ量は約 127GB である。計算ノードは複数のバッチジョブで共有される場合があるが、搭載する CPU 数またはメモリ量を超えてバッチジョブがアサインされることはない。また、S 系バッチノードではユーザディスクを NEC GFS によりマウントしており、ジョブ実行に際してステージングは行われない。

2.4 L 系バッチジョブ

L 系バッチジョブは、複数台の計算ノードを用いた実行が可能である。1 つの計算ノードに複数のバッチジョブがアサインされることはなく、計算ノードを専有利用できる。運用上、L 系バッチジョブで最大使用可能なノード数は 128 ノードとしている。利用者が L 系バッチジョブの実行をシステムにリクエストする際、バッチスクリプト中に以下の指定を行う。

- 使用ノード数
- 経過時間
- ワークディスク容量
- ステージインファイル（ロードモジュールや入力ファイル等）
- ステージアウトファイル（出力ファイル等）

リクエストされたバッチジョブは、バッチ処理システムと密接に連携する運用管理ソフトウェアによって、研究課題で保有するノード時間との比較やステージインファイル指定の検査が行われた後に、ノード数と経過時間、及びワークディスク容量を元に、ジョブスケジューラが計算ノードへのアサインを行う。図 4 に、L 系バッチジョブの状態遷移概念図を示す。

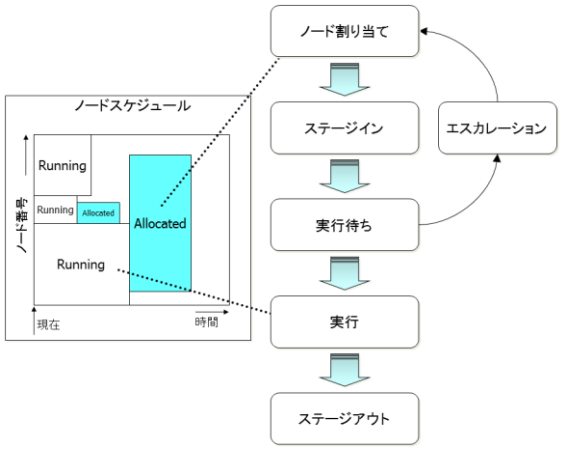

図 4　L 系バッチジョブ状態遷移概念図

計算ノードの割当てには、FIFO + バックフィルのスケジューリングポリシーを用いる。計算ノードの使用効率を高めるためと、計算ノードに外乱を与えないために、ステージング処理は、IOCSと呼ぶ専用の装置によってジョブの実行に先んじて行われる。計算ノードへのバッチジョブ割当て状況は、スケジューリングマップと呼ぶ図によって表す。図５に、48 時間後までのスケジューリングマップを示す。

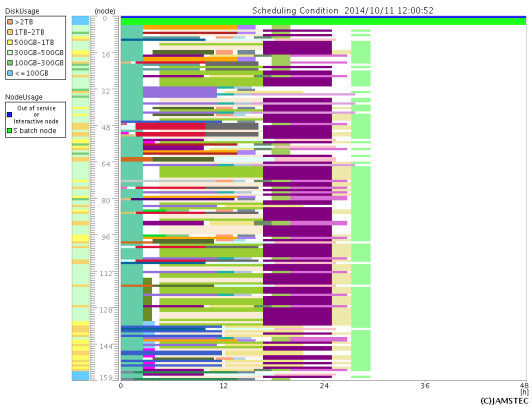

図５　スケジューリングマップ

2.5 ノード使用状況

ES2 では、計算ノードの状態を以下のように定義する。

- RUN　ジョブ実行中
- SAA　ジョブ実行準備・後処理中
- ECO　省電力休止中
- SBY　待機中
- STP　停止中
- MTN　計画保守

運用管理ソフトウェアは、バッチ処理システムと連携し、定期的に計算ノードの状態を記録する。計算ノードへの外乱を排除するため、この処理に際して計算ノードへの問合せは行わない。図６に、計算ノードの使用状況を示す。これは 2014 年 4 月 1 日から 10 月 26 日までのノード状態を表している。計算ノードは、90%の割合でジョブ実行に利用されている。

スケジューリングマップや計算ノード使用状況は、計算資源の計画的な利用の一助となるよう、地球シミュレータ利用者サポートウェブページから閲覧可能としている。

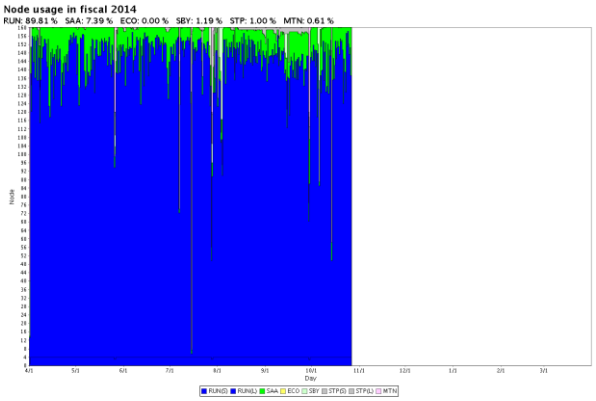

図６　計算ノード使用状況

2.6 大規模共有メモリシステム

ES2 の周辺システムとして、2014 年 4 月から大規模共有メモリシステムの利用が可能となっている。大規模共有メモリシステムは、SGI 製 UV2000 を中心としたシステムである。商用アプリケーションを利用した大規模計算など、主に産業界からの利用を目的として導入された。CPU には Intel E5-4650v2 プロセッサを採用し合計 2560CPU コアを搭載するとともに、32TB の巨大な共有メモリ空間を利用可能である。表１に大規模共有メモリシステムの概要を示す[3]。

表１　大規模共有メモリシステムの概要

理論性能	49.152TFLOPS
総プロセッサー(コア)数	256CPU socket (2,560core) 【Intel Xeon E5-4650v2(2.4GHz, 10core)】
総メモリ容量	32TB
OS	SUSE Linux Enterprise Server 11 SP3
ノード内接続	SGI NUMAlink6

大規模共有メモリシステムの利用者情報や計算資源は、ES2 の運用管理ソフトウェアによって一元管理される。バッチ処理での利用を基本とし、ES2 のバッチ処理システムによって、統一的なジョブ管理と操作を実現している。

3 ES2 後継システム

現行システムである ES2 の運用は、2009 年 3 月から 2015 年 2 月末までの 6 年間の予定で行われている。2015 年 3 月からは、NEC 製 SX-ACE

を中心とした後継システムの運用を開始する予定である。表２に、後継システムの概要を示す。また、図５にシステム概念図を示す。後継システムの運用では、大規模共有メモリシステムを含む周辺システムとの連携機能についても強化を図る予定である。

表２　後継システムの概要

システム		SX-ACE
CPU数(計算ﾉｰﾄﾞ数)		5120CPU・ﾉｰﾄﾞ
総ラック数(計算ﾉｰﾄﾞ部)		80
ノード構成	CPU名×数	専用CPU×1
	CPU周波数	1GHz
	CPUコア数	4
	コア最大性能	64GFLOPS
	CPU最大性能	256GFLOPS
	メモリ容量	64GB
	メモリバンド幅	256GB/s
システム総計	総合演算性能	1.3TFLOPS
	総メモリ容量	320TB
	総メモリバンド幅	1.3 PB/s

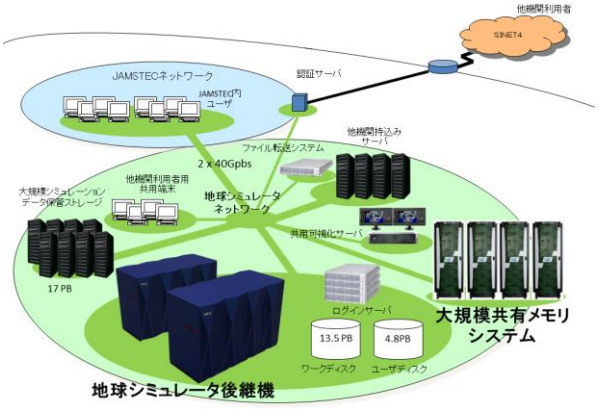

図５　地球シミュレータ後継システム概念図

後継システムの選定は、総合評価方式による一般競争入札によって行われた。総合評価方式とは、システムの性能・機能と入札価格を点数化し、最も高得点を獲得した者が落札者となる方式である。後継システムでは、アプリケーションプログラムの実行性能を重視し、性能評価試験の結果に最も多くの配点を与えている。ベンチマークプログラムには、ES2 で実行されている代表的な以下の 7 本のアプリケーションプログラムを用いた。プログラムの実行に当たっては、提案システムへの最適化を許可している。

- 全球大気海洋結合気候モデル（低解像度）
- 全球大気海洋結合気候モデル（高解像度）
- 全球雲解像大気大循環モデル
- ３次元地震伝搬解析コード
- 地震発生サイクルコード
- 大気海洋結合モデル
- データ同化コード

この性能評価試験の結果から、後継システムでは、単体プログラムの実行性能において現行システムの 2 倍から 2.5 倍、スループットにおいて 4 倍、システム全体の計算能力として 8 倍から 10 倍程度の向上が期待されている。

4 おわりに

地球シミュレータ(ES2)では、ES2 向けに最適化された運用管理ソフトウェアとジョブスケジューラにより、高い計算ノード利用率を実現してきた。後継システムにおいても、システム規模の拡大と性能の向上にあわせ、貴重な計算資源を有効活用し、かつ利用者にはより使いやすいシステムを構築するべく、準備が進められているところである。

参考文献

[1] 独立行政法人海洋研究開発機構地球シミュレータ開発史編集チーム、「地球シミュレータ開発史」、2010 年 12 月
[2] 地球シミュレータ、http://www.jamstec.go.jp/es/jp/index.html
[3] 大規模共有メモリシステム、http://www.jamstec.go.jp/es/jp/uv/2000index.html