Vol.2014-HPC-144 No.8
2014/5/26

地球シミュレータ（ES2）の運用実績とシステム構成

板倉 憲一[1,a)]　上原 均[1,b)]　浅野 俊幸[1,c)]

概要：本稿は、独立行政法人海洋研究開発機構で現在運用している地球シミュレータ (ES2) の実運用において出力される実績データの種類を示し、５年間の実データに基づくシステム全体の運用効率の評価およびユーザリクエストの実行状況を解析する。さらにこのような実システムの運用データの利用法について検討を行う。

The system architecture and operation results of the Earth Simulator (ES2)

Ken'ichi Itakura[1,a)]　Hitoshi Uehara[1,b)]　Toshiyuki Asano[1,c)]

Abstract: This paper describes the system architecture and operation results for five year of the Earth Simulator (ES2) which is operated by Japan Agency for Marine-Earth Science and Technology (JAMSTEC). we show the evaluation of the operating efficiency of the whole system and the execution situation of a user request based on the kind of the track record data for five years which are outputted in the real operation.

1. はじめに

独立行政法人海洋研究開発機構 (JAMSTEC) では、2002 年 2 月から地球シミュレータの運用を行っており [1]、2009 年 3 月にはシステムに更新し現在は NEC SX-9/E で構成される地球シミュレータ (以降、ES2 と記す) を運用している。ES2 の運用期間は 6 年間であり、約 5 年間の運用をした結果をまとめ、運用実績の調査・検討を行ったので報告する。

本論文の構成について述べる。第 2 章では地球シミュレータ ES2 のハードウェアと運用システムについて整理する。次に、第 3 章でシステムの運用実績のデータ収集について説明する。第 4 章では、実際の運用実績データを示しその解析を行う。第 5 章では関連研究について述べる。

2. ES2 システム構成

本章では、始めに ES2 のハードウェア構成について述べ、次にシステムの運用効率と大きく関わりのあるバッチジョブシステムについて述べる。

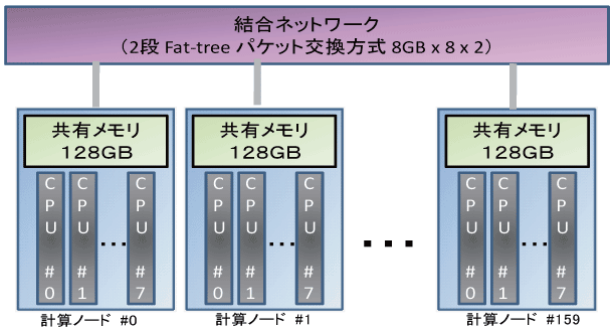

図 1　ES2 システム概要
Fig. 1　ES2 System Outline

2.1 ハードウェア構成

ES2 は 160 台の計算ノードを Fat-Tree 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。各計算ノードは、ピーク性能 102.4GFLOPS のベクトル型計算プロセッサ 8 個が主記憶 装置 128GB を共有する共有メモリ型計算機となる。システム全体では、1280 個のプロセッサによりピーク性能は 131TFLOPS、主記憶容量は 20TB となる。

[1] 独立行政法人海洋研究開発機構
JAMSTEC
[a)] itakura@jamstec.go.jp
[b)] uehara@jamstec.go.jp
[c)] asanot@jamstec.go.jp

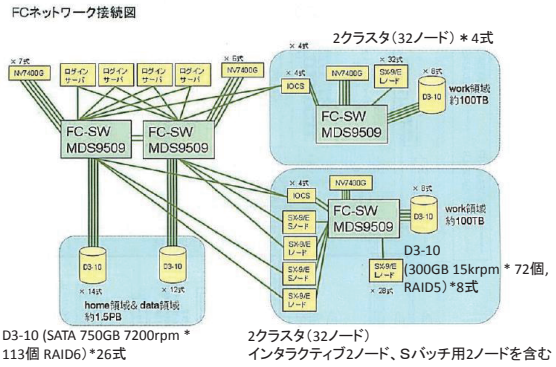

図 2　ストレージシステム概要
Fig. 2　Storage System Outline

ノード間結合ネットワークは 2 段の Fat-Tree で構成され、1 段目の 10 個の各スイッチには 16 ノードが接続され、2 段の 8 個の各スイッチには 1 段目の 10 個のスイッチが全て接続される。それぞれの接続は独自技術の光通信プロトコルを使用しており、1 リンク当たりの転送性能は 8GB/sec を持つ。

ストレージシステムは、パーマネントなデータを保持する大容量ファイルシステム装置と、各ノードに直結して一時的なデータを保持するワークデスク装置によって構成される。大容量ファイルシステムは HOME 領域、DATA 領域で構成され約 1.5PB の総容量を持つ。ワークディスクは 32 ノードで約 100TB のストレージシステムを物理的にはシェアしており、ノードあたり約 3TB の容量を持つ。

2.2　バッチリクエストシステム NQSII

ES2 は基本的には、バッチ処理システムであり、バッチ処理システムとして、Network Queuing System II(NQSII) が使用されている。図 3 に、ES2 のキュー構成を示す。実行キューとして、L バッチキューと S バッチキューと呼ばれる 2 種類のキューが用意されている。L バッチキューは大規模バッチジョブの実行を目的としている。S バッチキューは、L バッチキューで実行される大規模ジョブの初期値データ作成等の前処理や、大規模ジョブの実行結果の後処理での使用を目的としている。ユーザは実行するジョブに最適なキューを選択し、投入可能となっている。

S バッチキューのスケジューリングは CPU 時間で行われ、シングルノードジョブのみを実行することができる。一方、L バッチキューはスケジューリングは、次のようなポリシーを持って行われる。

- CPU 時間ではなく、ユーザが宣言した経過時間をもとにスケジューリングを行う。
- ジョブに割り当てられた計算ノードは、そのジョブが専有する。

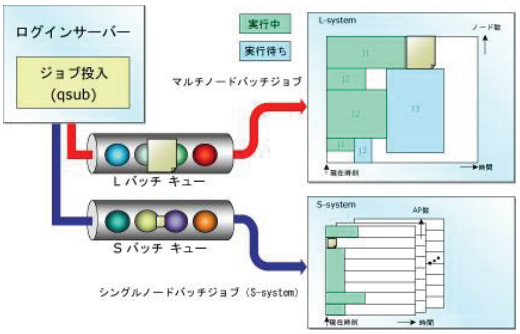

図 3　ES2 のキュー構成
Fig. 3　Queuing System on ES2

これによって、スケジューラはジョブの終了時間を予測することができ、他のリクエストを効率よくスケジュールすることができる。また、ジョブの実行時の効率の向上に寄与している。ジョブはノードを専有することができるので、各ノード上でのプロセスの実行状況を均一化することができるようになり、大規模並列プログラムを効率よく実行することができる。ストレージシステムで述べたように、各計算ノードはユーザディスクに直接アクセスできず、一時ファイル用である各ノード専用のワークディスクにのみを利用する。これは、ジョブ実行時のディスクアクセス競合を防ぎ、コンスタントな性能を確保するためである。このために、ジョブの実行に必要なファイルはジョブの実行前にワークデスクへコピーする必要があり、これをステージインと呼ぶ。このステージインにかかる時間は、ジョブスケジューリングによって隠蔽する。ジョブの実行フェーズは以下の通りである。

(1) ノード割り当て
(2) ステージイン（ファイルをユーザディスクからワークディスクにコピー）
(3) ジョブエスカレーション（再スケジューリング）
(4) ジョブ実行
(5) ステージアウト（ファイルをワークディスクからユーザディスクにコピー）

ジョブが投入されると、スケジューラは使用可能なノードを探し、実行開始時刻を決定する (1)。そして、ジョブの実行ノードが確定するとそのノードのワークデスクに対してステージインを行う (2)。ステージインが完了すると、そのジョブは実行開始時刻まで待機する。この間に、実行開始時刻を早めることができる場合には、再スケジューリング（ジョブエスカレーション）が行われる (3)。実行開始時刻になると、スケジューラはそのジョブを実行する (4)。ユーザが宣言した実行時間を過ぎるか、計算が終了すると、スケジューラはそのジョブを終了させる。ジョブ終了後、スケジューラはステージアウトを実行する。

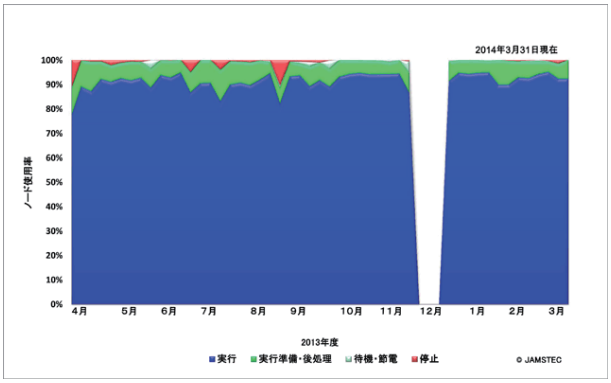

図 4　2013 年度 ノード使用状況
Fig. 4　Node Usage FY2013

3. 運用実績データ

ES2 の運用実績データは主に計算ノード単位に出力される物とリクエスト単位で出力される物がある。計算ノード単位で出力される物には、ノードのステータスの 1 分単位での記録がある。ノードのステータスは以下の通り。

- 実行 (S/L): S または L キューのリクエストを実行中
- 実行準備・後処理: リクエストが割り当てられており、ステージンイン・アウトを含む実況準備・後処理中
- 省電力休止: 長時間リクエストが割り当てられていないため、ノードの電源を落とし省電力休止を行っている
- 待機: リクエストが割り当てられておらず、待機している状態
- 停止 (S/L): S または L キューに割り当てられているが、障害等で停止している状態
- 計画保守 (S/L): S または L キューに割り当てられているが、保守等で計画的に停止している状態

各ノードには複数個のリクエストが割り当てられる可能性があるが、実行状態のリクエストは高々 1 つである。ステータスを測定する 1 分間の中に実行状態のリクエストがあれば、そのノードは実行状態として記録される。ノードに割り当てられているリクエストが全て実行準備・後処理の状態であれば、そのノードは実行準備・後処理として記録される。省電力休止、停止、計画保守の状態ではノードにリクエストは無く、ステータスを測定する 1 分間の中にノードが該当する状態となればその状態で記録される。どの状態にもならなかった場合には、待機として記録される。

この計算ノード単位の状態をグラフ化した図 4 は、一般向けに公式 Web で公開している。

また、リクエスト単位で収集するデータには、以下の項目がある。

- リクエスト基本情報（リクエスト ID、投入ホスト、投入キュー、リクエスト名、終了ステータス）
- ユーザ情報（ユーザ名、グループ名、氏名、責任者名）
- L 系資源情報（L 系ノード数、L 系経過時間、L 系宣言ディスク量）
- S 系資源情報（S 系 CPU 数、S 系メモリ量、S 系 CPU 時間）
- 時刻情報（投入日時、実行開始日時、実行終了日時、pre-run 日時、post-run 日時）
- ステージング情報（ステージイン時間・量、ステージアウト時間・量）
- リクエスト詳細情報（プライオリティ値、rerun カウント、hold カウント、delete 要因）
- 性能情報（I/O バイト量、GFLOPS 値、ベクトル演算率、GOPS 値）

L 系資源情報のノード数、経過時間、ディスク量はジョブスクリプトでユーザが指定する値である。スケジューラは現在から 48 時間先までのノード資源の中でノード数 × 経過時間のリソースがはまる場所を探して配置するが、その時に各ノードに専用で付属しているワークデスクの容量も考慮する必要がある。ワークデスクはノード当たり 3TB 確保しているが、ユーザは最大 1.5TB まで 1 リクエストで使用することができため、最大規模のリクエストが 2 つノードに割り当てられると、時間的には空いていてもそのノードにはリクエストを配置することができなくなる。

表 1　5 年間のノード利用実績
Table 1　Node Usage in 5 years

年度	実行	実行準備・後処理	待機・節電	停止	計画保守
H21	72.75%	10.09%	9.59%	1.48%	6.08%
H22	86.40%	5.61%	0.81%	0.73%	6.45%
H23	84.62%	6.45%	3.94%	1.41%	3.58%
H24	82.43%	8.94%	4.79%	0.67%	3.17%
H25	82.96%	6.59%	0.35%	0.90%	9.19%

4. 5 年間の運用実績

4.1 ノードの利用率

表 1、図 5 に 5 年間のノードの利用率を示す。全体としては 9 割程度のノードの使用率となっており、残りの部分もほとんど実行待ちの状態となっていることが分かる。マシンをリプレースして最初の半年間は、新しいマシンへのプログラムチューニングや結果の検証等のためにマシンが使われ、結果として大規模な温暖化実験や長時間の地震波解析などに関わるプロダクトランが行われなかったため、低い利用率に留まった。また、2012 年の年度前半の落ち込みは、ES の主要な研究グループである地球温暖化研究のプロジェクトが切り替わったタイミングであり、ES の利用低下に繋がった。

2011 年 3 月 11 日の東日本大震災の影響は、速報データによる緊急解析を実施した後に、電力不足に対応するために 3 月末までの運用停止を自主的に決めて実施した。さら

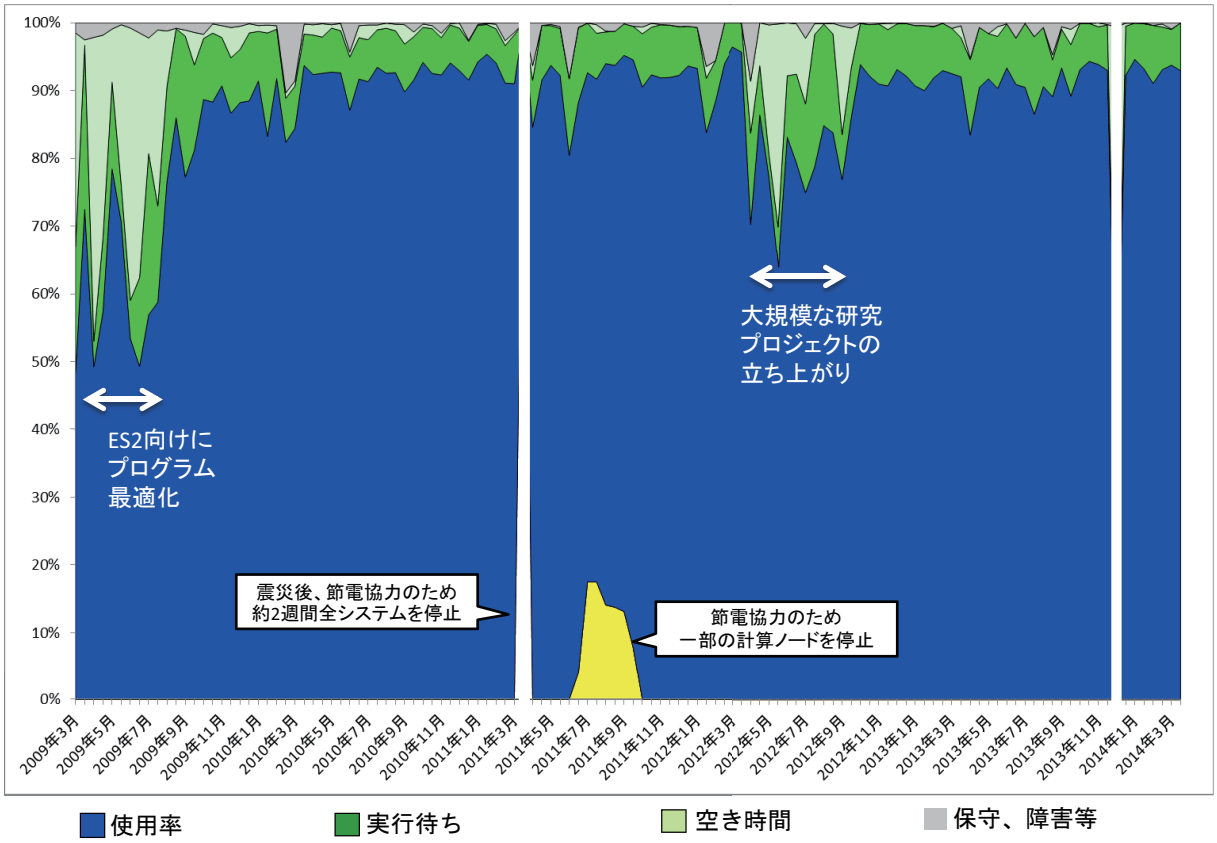

図 5　5 年間の ES2 の運用実績
Fig. 5　Node Usage in 5 years

に、その夏の節電対策では国の施策に従って 15%の電力削減を一部の計算ノードを停止することで実現し、社会的要請と研究推進の両立を図った。

4.2　ユーザの利用状況

ES2 の NQSII リクエストは、ほぼコンスタントに 1 ヶ月あたり約 9,000 件を処理してきた。ここでは、直近で 2013/4/1〜2014/3/31 までの 1 年間に L 系に投入された全 92,530 件のリクエストからメンテナンス等でシステム管理者が投入したリクエストとキャンセルされたリクエストを除いた 77,202 件リクエストについて解析を行う。なお、この間に 12/1 から 12/24 まで大規模な空調設備のメンテナンスの為に運用を停止と 2 ヶ月に一度 1 日程度のメンテナンスを行っており、実質的な運用日数は 338 日である。

図 6、7、8 はそれぞれステージインのデータ量、ステージアウトのデータ量、L 系ワークデスクの使用量のヒストグラムで縦軸はリクエスト数を示す。ステージインの量は、2.5GB から 75GB 未満あたりが多く、全体の約 63%を占める。また、ステージアウトの量も、2.5GB から 75GB 未満あたりが多く、全体の約 65%を占める。これは、ES2 の実行時間が最大 12 時間に制限しているため、長期間のシミュレーションには必要なデータを全てステージングする必要があり、前回の実行のステージアウトしたファイルが

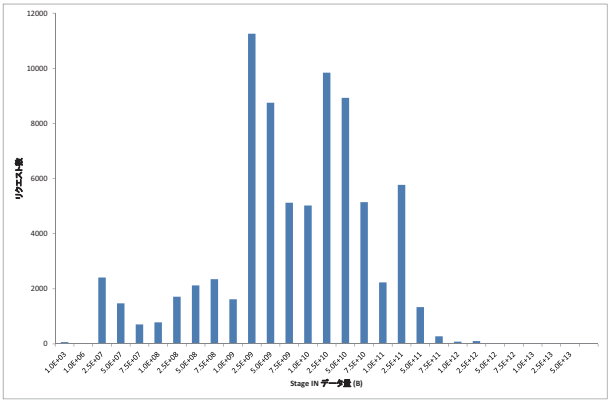

図 6　ステージインのデータ量ヒストグラム
Fig. 6　Histgram of Stage-in data amount

そのまま次回のステージインのファイルになることも一例としてあげられる。L 系ワークディスクの使用量は、25GB から 500GB 未満あたりが多く、全体の約 92%を占める。運用初期の頃は、ストレージ量の見込みがつかず、多めに宣言することが多く見られたが、ワークディスクを多く必要とするリクエストはノードの割り当てが後ろに回されること多く、適切な量を指定することが、TAT（トータルアラウンドタイム）を短くすることに繋がることがユーザコミュニティの中で理解されてきていることが分かる。

次に、図 9、10 はそれぞれステージインのデータ量と転

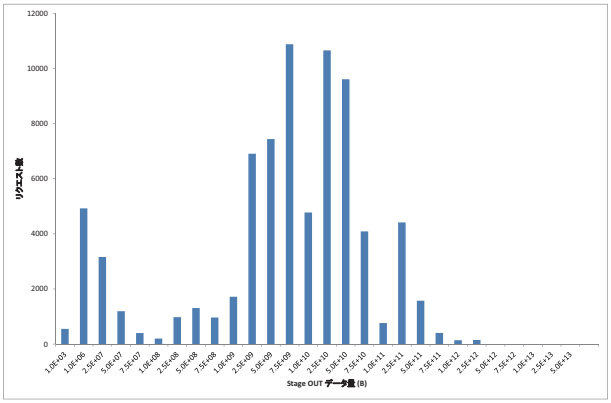

図 7　ステージアウトのデータ量ヒストグラム
Fig. 7　Histgram of Stage-out data amount

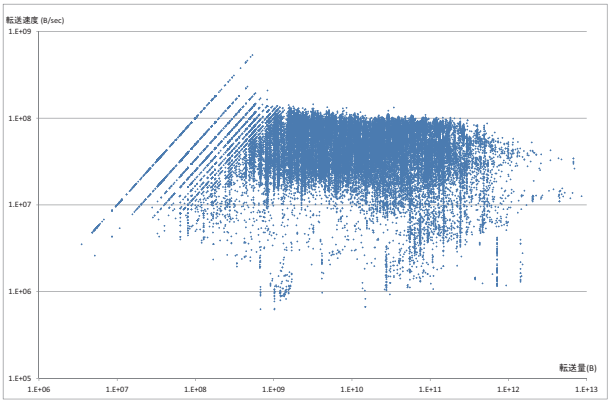

図 9　ステージインのデータ量と転送速度の関係
Fig. 9　Data amount v.s. transfer speed in Stage-in

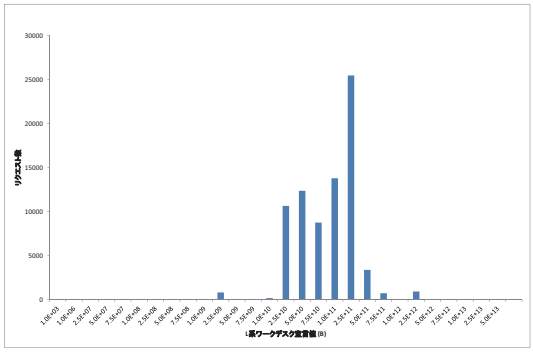

図 8　L 系ワークデスク使用量ヒストグラム
Fig. 8　Histgram of work disk usage in L batch requests

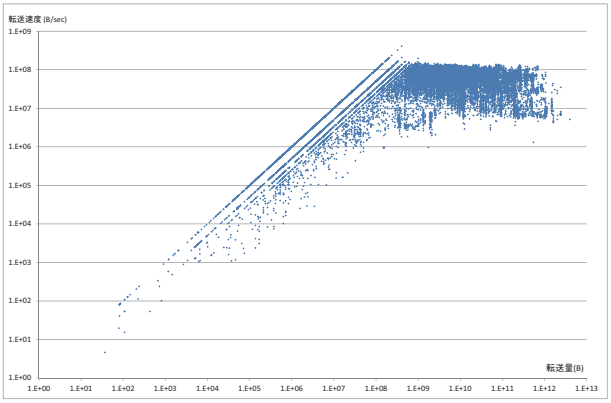

図 10　ステージアウトのデータ量と転送速度の関係
Fig. 10　Data amount v.s. transfer speed in Stage-out

送速度の関係、ステージアウトのデータ量と転送速度の関係を示す。どちらの場合でも、データ量が少ない場合にはファイル I/O に必要となる固定のオーバーヘッド時間が支配的となり、転送時間が数秒の単位で固定となる。現在のリクエスト単位で収集する統計データは秒単位で丸められているため、1 秒から 10 秒未満の時間のリクエストが集中する傾向があり、図で左側の部分にすじ状の分布となって現れる。データ転送量が十分ある場合には、転送スピードは 10MB/sec から 100MB/sec で行われている。但し、特にステージインの場合に数 MB/sec しか出ていないケースも多く見られる。ES2 のステージングファイルはユーザがジョブスクリプトに明記し、その記述に従ってデータ転送が行われる。実際には図 2 のクラスタ内にある IOCS と呼ばれる Linux ワークステーションが HOME、DATA 領域と WORK 領域の間でコピー処理を行う。各ストレージ領域内でのコンフリクトが起きている可能性があり、ユーザのジョブスクリプトに無駄が無いかコンサルティングを行っていく必要がある。

5.　関連研究

直接的に大型計算機の運用結果に基づいたシステム構成の評価は、各運用機関の中で運用主体者とメーカ、ユーザが一体となって行われており、アカデミックな報告としてはあまり例がない。ここでは、本論文で示したような実システムのジョブ履歴情報がジョブ管理システムの改良やシステム評価に利用されている例を示したい。

まず、宇野らは、[2] において、ジョブスケジューリングの方式をソフトウェアシミュレーションによって解析し、特にファイルステージングにおけるファイルアクセス競合によって帯域が減少した場合のスケジューリングに及ぼす影響を検討している。この時には実際の「京」コンピュータの運用を考慮しつつ、複数のタイプのジョブミックスを想定して評価を行っている。ユーザ層が等しいようなシステムの検討では実システムのジョブミックスを利用することも十分に検討できる。

次に、Wong らは、[3] において ESP(Effective System Performance) test というシステムレベルでの大型計算機の性能評価をする手法を提案している。ここでは、単純なプログラムの計算性能だけでなく、ジョブハンドリングの効率やシステムのシャットダウン・ブートアップの時間も含めて評価する取り組みとなっている。ここでもシステムを評価するにあたり、実運用のジョブミックスを用いることは重要と考えられており、そのようなデータは貴重だと

考えられている。

6. おわりに

本論文では、ES2 の 5 年間の運用結果をまとめて報告した。本システムは多くのユーザが利用するマシンであり、ハードウェアの利用状況のログや、ユーザリクエストの実行ログが多く存在する。ここでは、ES2 の主な特徴であるステージング処理に注目してリクエスト結果の解析を行ったが、他にも多くの情報が得ることができると考えられる。また、関連研究で述べたように、このようなユーザリクエストの履歴はジョブ実行系の研究に利用することが可能であり、そのような研究にも積極的に貢献する予定である。さらに、他の大型計算機を運営しているセンターとも情報交換を行い、よりユーザが高性能な計算を行える環境を提供していくことを考えている。

謝辞 運用データをまとめるにあたりご協力頂いた日本電気株式会社 津田義典氏はじめ関係者に深謝する。

参考文献

[1] 村井 均 他: “特集：地球シミュレータ”, 情報処理学会会誌「情報処理」 Vol45, NO.2, 情報処理学会, Feb. 15, 2004.
[2] 宇野 篤也、庄司 文由、横川 三津夫: “ファイルステージングのるジョブスケジューリングの評価”, 情報処理学会研究報告 12-HPC-136(22), 情報処理学会, Oct. 4, 2012.
[3] Adrian Wong, Leonid Oilker, William Kramer, Teresa Kaltz and David Bailey: *“System Utilization Benchmark on the Cray and IBM SP”* The 6th Workshop on Job Scheduling Strategies for Parallel Processing, April 19, 2000.