地球シミュレータの利用促進の取り組み

甲斐 恭、齋藤 友一、上原 均

海洋研究開発機構 地球情報基盤センター

tadashik@jamstec.go.jp

Promote the use of the Earth Simulator.

Tadashi Kai, Yuichi Saito, Hitoshi Uehara

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

概要

国立研究開発法人海洋研究開発機構で運用するベクトル型スーパーコンピュータ「地球シミュレータ」の利用促進の取り組みについて紹介する。

1 はじめに

海洋研究開発機構で運用するベクトル型スーパーコンピュータ「地球シミュレータ」は、2015年にNEC社製のSX-ACEで構成された第3世代のシステムに更新した。計算ノードは総計 5,120 ノードで、総理論演算性能が 1.31PFLOPS と前世代の131TFLOPSの10倍となっている。これらの計算ノードはクラスタ単位で管理されており、512ノード構成の基本クラスタ 6 台と 2,048 ノード構成の拡張クラスタ1台がある。

この第3世代の地球シミュレータを効率的に運用するために、利用促進の取り組みを行っている。

本稿では、利用促進の取り組みとして実施した低優先キュー、穴埋めリクエスト、上半期・下半期制の導入について紹介する。

2 低優先キュー

2.1 低優先キュー

低優先キューは、利用促進を目的に作成した、スケジューリングの優先度が低い代わりに計算資源（ノード時間積）を消費しないキューである。また、同時実行リクエスト数についても通常キューよりも少なく設定している。

低優先キューは、以下のいずれかの条件を満たした利用グループが利用できる。

- 月あたりの使用ノード時間積による条件
  毎月 1 日からの使用ノード時間積が割当ノード時間積を利用期間の月数で割った値を超える
  ただし、4 月については 上記の値の半分を超える
- 期ごとの累積使用ノード時間積による条件
  利用期間内の累積ノード時間積が割当ノード時間積を利用期間の月数で割った値に利用期間の開始月からの経過月数を掛けた値を超える

2.2 低優先キューの実現

低優先キューの運用を実現するにあたり、判定・設定・通知方法に課題があった。

そこで、これらの課題に対応したシェルスクリプトとcronによる定期実行とを組み合わせることで低優先キューの運用を実現した。

判定を行う上での課題は、判定の実現方法にあった。低優先キューの利用条件を満たしているかを判定する機能はキューを制御するジョブスケジューラには搭載されていない。そこで、ログインサーバ上にて利用状況を確認するコマンドなどをシェルスクリプトによって組み合わせることで判定を行う機能を実現した。

設定を行う上での課題は、低優先キューを使用できる利用グループを制限する方法にあった。ジョブスケジューラの管理コマンドから特定のキューを利用できるグループを登録する機能があり、判定を行うシェルスクリプトと組み合わせることで、低優先キューを使用できる利用グループを制御する機能を実現した。

通知方法では、条件を満たした利用グループへの通知方法に課題があった。そこで、ログイン

サーバの sendmail コマンドを判定を行うシェルスクリプトと組み合わせて、低優先キューを利用可能となった利用グループに対して定型メールでの通知を行う機能を実現した。

2.3 低優先キューの効果

2016 年度は 25 グループ（全所内課題と一部公募課題）を対象として、約 600 万ノード時間積、2017 年度は 47 グループ（全所内課題と全公募課題）を対象として、約 500 万ノード時間積が低優先キューの実行に使用された。

実際に低優先キューを使用した利用グループ毎の使用計算資源量と低優先キューで実行した資源量（ノード時間積）を低優先キューでの実行資源量が多い順に並べると以下の通りとなる。縦軸は資源量（ノード時間積）を示している。

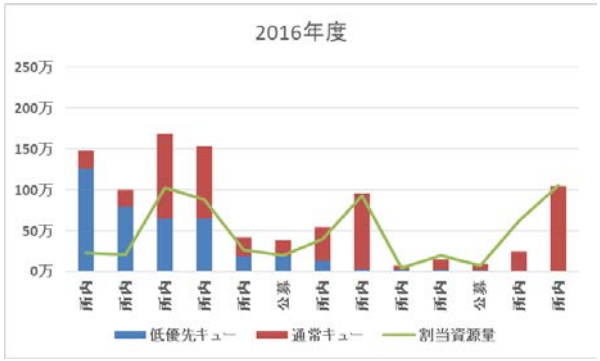

図 1 低優先キュー・通常キューの使用量および割当資源量(2016 年度)

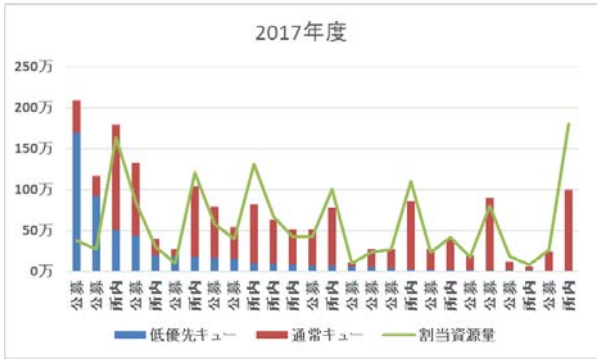

図 2 低優先キュー・通常キューの使用量および割当資源量(2017 年度)

2016 年度と 2017 年度のいずれも、上位の数グループが積極的に低優先キューを使用する傾向がある。また、上位 2 グループが割当資源量を大幅に超えて低優先キューを使用している。特に 2017 年度の公募課題では、選定時に希望する計算資源（ノード時間積）が必ずしも割り当てられない場合があり、その救済措置として利用される面がある。

ただし、低優先キューと全体の、利用バランスについて、最適なパラメータを探る必要がある。

3 拡張クラスタの穴埋めリクエスト

3.1 拡張クラスタ

拡張クラスタは最大 2,048 ノードのリクエストを実行可能な大規模リクエスト用クラスタである。拡張クラスタで実行するリクエスト専用の L2 キューは 256 ノード以上 2,048 ノード以下のリクエストのみを投入可能としている。

3.2 穴埋めリクエスト

使用するノード数や投入されるリクエスト数の関係でタイミングによって、L2 キューで投入されるリクエストだけでは拡張クラスタで待機中となる計算ノードが発生する場合があった。待機中の計算ノードを有効利用するため、穴埋めリクエストとして、1 ノード以上 512 ノード以下のリクエスト用の L キューに投入された、小規模で宣言経過時間の短いリクエストを拡張クラスタでも実行するようにした。

また、L キューに投入されたリクエストが拡張クラスタで実行されることで基本クラスタの混雑を緩和する効果も期待される。

3.3 対象リクエストの調整

穴埋めリクエストの運用は 2016 年 4 月にスタートした。その後、256 ノード未満で 1 時間以下(2016 年 4 月)、256 ノード未満で 4 時間以下(2017 年 1 月)、256 ノード未満で 6 時間以下(2017 年 2 月)に対象を見直している。なお、2017 年 8 月から 2017 年 9 月にかけては特別のリクエストの実行に対応するために 512 ノード以下で 3 時間以下、同 6 時間以下、同 8 時間以下に対象を変更したが、その後は現在（2018 年）まで 256 ノード未満で 6 時間以下のリクエストを対象としている。

3.4 穴埋めリクエストの効果

穴埋めリクエストの運用を取り入れたことにより、小規模なリクエストが拡張クラスタで実行され、スケジューリングの状況が改善した。

一例であるが、穴埋めリクエスト実施前後での実際の拡張クラスタのスケジューリング状況

を示す。

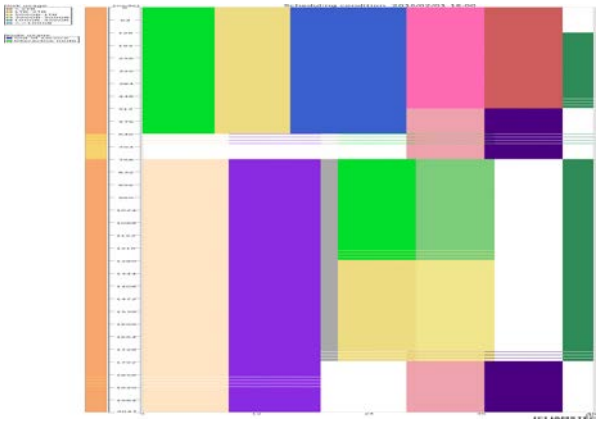

図 3 拡張クラスタのスケジューリング状況 (2016 年 2 月 1 日 18:00)

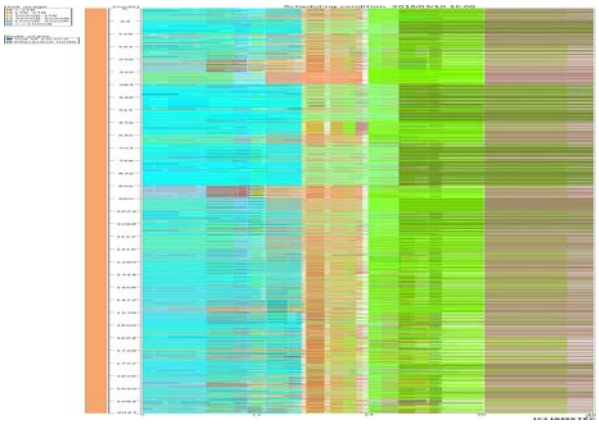

図 4 拡張クラスタのスケジューリング状況 (2018 年 3 月 10 日 10:00)

これらの図の縦軸はノード、横軸は待ち時間を示しており、スケジューリングされたリクエストを色分けして表示している。図の左端に接しているリクエストが現在実行中であることを示す。また、白色はリクエストが無い状態を示す。

実施前の 2016 年 2 月 1 日 18:00 時点では、実行が 93.75%、待機が 6.25%であった。一方、実施後の 2018 年 3 月 10 日 10:00 時点では、実行が 99.80%、待機が 0.20%であった。この 2 点においては、使用されずに待機となるノード数が削減されている。また、リクエストが無いことを示す白色も実施後のスケジューリング状況では減少している。

4 上半期・下半期制度の導入

4.1 上半期・下半期制度導入の背景

2016 年度までは年度を通じて単一の利用期間での利用であった。運用の課題として年度前半は利用が進まず、年度後半の年末から年度末にかけて利用が集中する点があげられる。

4.2 上半期・下半期制度

年度前半の利用を促進するために、利用期間を 4 月から 9 月までの上半期と 10 月から 3 月までの下半期に分けて、利用期間毎に計算資源を設定し、上半期で使用しなかった資源は下半期に持ち越せない運用を実施した。

4.3 上半期・下半期制度の効果

以下に 2016 年度上半期と 2017 年度上半期のノード使用状況を示す。

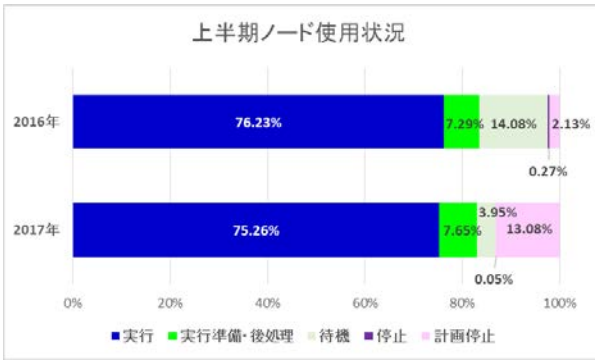

図 5 2016 年度上半期と 2017 年度上半期のノード使用状況の比較

ノード使用状況から、以下に示す使用率の算出式にもとづいて、使用率を比較すると、2016 年度上半期は使用率 85.33%、2017 年度上半期は使用率 95.39%となった。

$$\text{使用率：提供した計算機資源が使用された割合} = \frac{\text{実行＋実行準備･後処理}}{\text{全体－計画停止}}$$

図 6 使用率の算出式

上半期・下半期制にしたことで年度前半の利用が促進されている。

5 その他の対策

その他、特別推進課題の利用期間について、年度をまたいで設定することで、4 月にも計算を継続して実施する利用グループを確保した。また、

メンテナンス作業によりシステムの運用を停止する際に投入済みのリクエストをクリアせず、リクエストの情報を残したままメンテナンスを実施して、運用再開後に残していたリクエストをユーザが再投入することなく再スケジューリングして実行した。

5　まとめ

2015 年から運用している第 3 世代の地球シミュレータにおいて、低優先キュー、穴埋めリクエスト、上半期・下半期制の導入などの利用促進策を実施した。低優先キューでは、2016 年度は約 600 万ノード時間積、2017 年度は約 500 万ノード時間積が低優先キューを利用しての実行に使用された。穴埋めリクエストでは、大規模リクエストのみではリクエストのノード数や経過時間の設定によって発生する実行待機中の計算ノードで小規模リクエストが実行されるようになり、実行待機中の計算ノードの低減につながった。上半期・下半期制の導入では、年度前半の利用が促進された。

年間の使用状況では、2016 年度は待機が 10.40%であったが、2017 年度は最終的にノード使用状況の実績で待機が 2.71%まで低減される結果となった。

提供しているすべての計算ノードで計算が実行されている状態は、その計算機が最も有効に利用されている状態であるといえる。この状態のとき、時間あたりに処理できる計算量も最大となり、利用者にとってもリクエストの処理が最も効率的に行われる状態であるといえる。待機している計算ノードが少ないということは、この最も効率的な状態に近づくことである。

そのため、2017 年度に待機が減少したことは、より効率的な運用を実現できたといえる。

参考文献

[1]　地球シミュレータ
http://www.jamstec.go.jp/es/jp/