Vol.2020-HPC-177 No.6
2020/12/21

大規模環境の動的実行領域分割ジョブスケジューリング手法の一検討

高山　沙也加[†1]　関澤　龍一[†2]　鈴木　成人[†3]　山本　拓司[†3]　小口　正人[†1]

概要：近年の HPC システムでは利用者の増加と利用者層の拡大に伴い, 投入されるジョブ数および必要ノード数は増加しており, また, 多様化している. そのためシステム規模は増加傾向にあり, 効率的な運用がより重要となる. システムの運用において, ユーザの立場ではジョブ投入から実行されるまでの待ち時間が, 運用の立場では利用可能な計算機資源のうち実際に利用された割合である充填率が重視される. 特にあらゆる規模のジョブが投入される環境では, 小規模ジョブが大規模ジョブの割り当ての妨げとなり待ち時間が著しく長い大規模ジョブが生じてしまうことがあり, この軽減が大きな課題となる.
本研究では, 大規模ジョブの待ち時間の削減を目的として, 過去のジョブ分布やジョブ実行状況を分析し, 自動で実行領域の分割数と分割ノードサイズ, 実行ジョブのノード数範囲を決定するアルゴリズムの提案と効果検証を行う.

キーワード：ジョブスケジューリング, HPC システム, 領域分割

1. はじめに

近年の HPC システムでは利用者の増加と利用者層の拡大に伴い, 投入されるジョブ数およびその必要ノード数は増加しており, また, 多様化している. 2019 年に計算資源の共用を終了したスーパーコンピュータ「京」のジョブ規模別計算資源利用状況を見てみると, 必要ノード数が 1000 を超えるジョブのノード割当時間積が半数以上を占めている [1]. このように, ジョブ規模の増大化に伴い, システム規模は全体的に増加傾向にある. システムの運用において, 図 1 に示すようにユーザの立場ではジョブ投入から実行されるまでの待ち時間が, 運用の立場では利用可能な計算機資源のうち実際に利用された割合である充填率が重視される. 待ち時間はシステム側で十分なノード数を用意することで, 充填率は必要ノード数に応じてジョブ受け入れに制限を設けることで改善可能だが, 一方を優先させることで他方が悪化する可能性がある. 特にあらゆる規模のジョブが投入される環境では, 小規模ジョブが大規模ジョブの割り当ての妨げとなり待ち時間が著しく長い大規模ジョブが生じてしまうことがあり, この軽減が大きな課題となる. 投入ジョブの必要ノード数に応じてシステムとキューにパーティションを設け 2 分割することで, 充填率と待ち時間の改善を図るという手法は実環境の運用で既に導入されている. しかし, 大規模環境では 2 分割の制御では十分な効果を得られない場合がある.

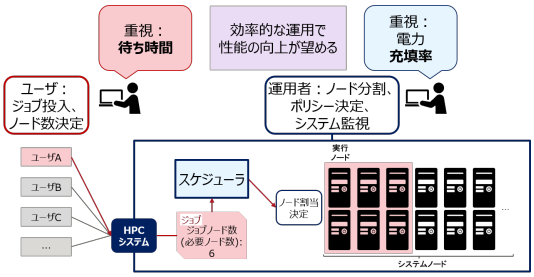

図 1　HPC システムの想定図

本研究では, 大規模ジョブの待ち時間の削減を目的として, 過去のジョブ分布やジョブ実行状況を分析し, 自動で実行領域の分割数と分割ノードサイズ, 実行ジョブのノード数範囲を決定するアルゴリズムの提案と効果検証を行う.

2. 関連研究

システムの効率的な運用を目的としたジョブスケジューリング手法は既にいくつか提案されている. 多様なジョブ受け入れを前提とした, 不均一なマシンで構成された HPC システムの利用率とジョブの待ち時間の最適化を目的とし

†1　現在, お茶の水女子大学
Presently with Ochanomizu University
†2　現在, 富士通株式会社
Presently with Fujitsu Ltd.
†3　現在, 株式会社富士通研究所
Presently with Fujitsu Laboratories Ltd.

た研究として，履歴ワークロードを用いた Portable Batch System（PBS）ベースのクラスタのジョブシミュレーションを実行する方法が提案されている [2]．この検証ではジョブスケジューリングのシミュレーション機能を持たない PBS リソースマネージャの代わりに，マウイスケジューラを用いて大規模な PBS ベースのクラスタのジョブシミュレーションを行っている．

スケジューリングのコストと効率のコントロールを目的として，時間軸方向をいくつかの区間に区切り，区間単位でジョブスケジューリングを行う区間スケジューリングが提案されている [3]．この検証では，充填率は直近の未来では短いスケジュール区間，遠い未来には長いスケジュール区間を使った場合に最も高く，待ち時間はスケジュール区間が長いほどノード数の少ないジョブの待ち時間が短くなることが確認されている．

また，HPC システムのジョブスケジューリングでは後述の Backfill が用いられることが多いが，その効率はジョブ実行時間の見積の正確さに依存する．利用者のジョブ実行時間見積の正確性を示す指標を導入し，その指標に応じた優先度で Backfill する手法によって通常の Backfill と比較して最大で 6％ の処理性能が向上したことが示されている [4]．

以前の研究では，異なるシステム規模，ジョブ分布条件でジョブスケジューリングを行い，充填率と待ち時間の評価を行った [5]．その結果，ジョブの必要ノード数とその分布がジョブスケジューリングの充填率と待ち時間の違いに大きく影響を与えており，効率的なシステム運用のためには実行領域の動的な変更が必要となることが明らかになった．

本研究ではジョブの必要ノード数の分布や待ち時間に注目し，自動で実行領域の分割数と分割ノードサイズ，実行ジョブのノード数範囲を決定するアルゴリズムの提案と効果検証を行う．

3. 実行領域分割アルゴリズム

HPC ユーザが最も重視する指標は自身のジョブが投入されるまでの時間である待ち時間である．特にあらゆる規模のジョブが混在する環境では小規模ジョブが大規模ジョブの割り当ての妨げになることがある．そのため大規模ジョブの待ち時間は長期化しやすく，これを短くするのが課題となる．そこで，小規模ジョブが大規模ジョブの実行の妨げとなるのを防ぎ，ジョブ規模毎に生じる待ち時間の大きなばらつきの軽減を目的とした実行領域分割アルゴリズムを提案する．アルゴリズムの概要を以下に記す．

提案アルゴリズムでは，過去の投入ジョブ分布とジョブの待ち時間，そして運用者が設定したジョブの最大最小待ち時間の差から，実行領域の分割数，各実行領域が受け入れるジョブの必要ノード数範囲，各実行領域のサイズを決定する．アルゴリズムはジョブスケジューリングから独立して動作し，一定間隔で適用する．初めに，各実行領域における最大最小待ち時間の差と運用者が設定した最大最小待ち時間の差を比較し，実行領域の分割数を決定する．いずれの実行領域でも最大最小待ち時間の差が運用者の設定を下回っていれば分割数は-１，そうでなければ＋１する．次に，各実行領域が受け入れるジョブの必要ノード数範囲を決定する．この時，各実行領域のジョブ粒度が揃うよう分割する．例えば，最大必要ノード数が 10000 として分割数を２とすると，実行領域１には１から 100，実行領域２には 101 から 10000 の必要ノード数のジョブを投入する．最後に，決定した分割数とジョブの必要ノード数範囲で，各実行領域の大きさを決める．ジョブの必要ノード数範囲別にノード時間積の総和を求め，その比に合わせてシステムノードを分割し，各実行領域とする．この時，ジョブの必要ノード数に対して実行領域があまりに小さいとキューが詰まりやすくなるため，最大ジョブノード数の２倍より大きくない実行領域があった場合，該当する実行領域のノード数を最大ジョブノード数の２倍とし，他の実行領域はシステムノードから該当実行領域のノード数を差し引き割合で分割した大きさとする．アルゴリズムの流れを図２に示す．

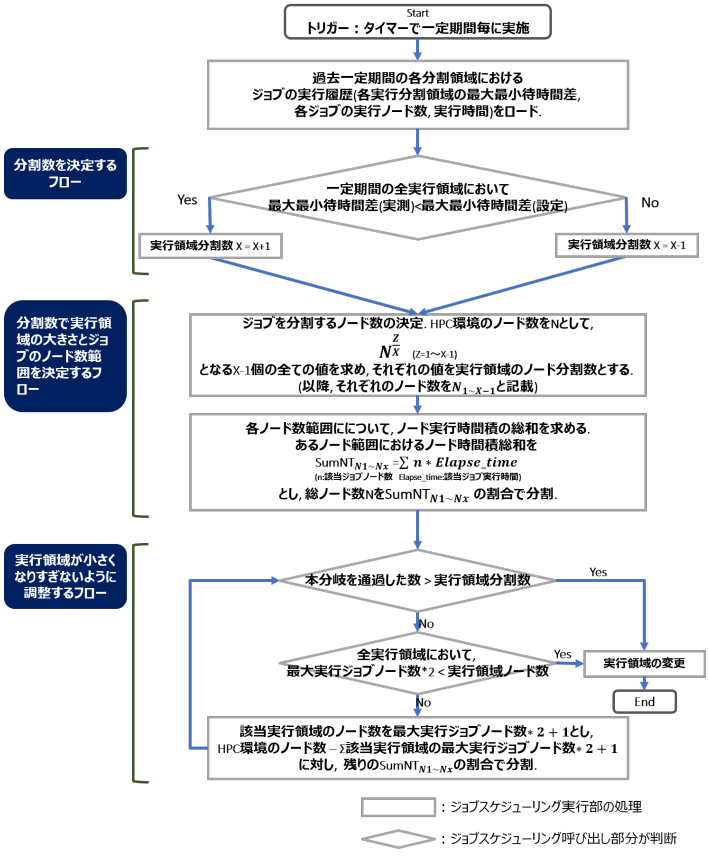

図 2 提案アルゴリズムのフローチャート

4. 実験概要

提案アルゴリズムの効果検証のために，あるジョブ条件

でアルゴリズムを適用し, 分割数, 各実行領域が受け入れるジョブの必要ノード数範囲, 各実行領域の大きさを変えた際の充填率と待ち時間の評価を行う. 本来のアルゴリズムでは各実行領域における最大最小待ち時間の差と 運用者が設定した最大最小待ち時間の差を比較し分割数を決定するが, 実験では効果確認のために最大最小待ち時間の差に関係なく分割数を増やし実験する.

ジョブスケジューラには Slurm Workload Manager (Slurm)[6] を用いる. ジョブデータには, 後述のジョブミックスを利用する. ジョブ密度をシステムノードに対する投入ジョブ規模の多寡の指標とし, 次式で求めた値とする.

$$Input_Ratio = \frac{\sum_{i=0}^{n} NodeJob_i \times ElapseTime_i}{TotalSubmit \times SystemNode} \times 100$$

この時, NodeJob はそのジョブが必要とするノード数, ElapseTime はジョブの実行時間, TotalSubmit は総投入期間, SystemNode はシステム全体のノード数を表す.

シミュレーションにおいて定めたシステム側の条件を表 1 に, ジョブ条件を表 2 に記す. ノードごとの性能差やネットワークによる不都合がない環境を前提とする. また, システムを構成するノードの扱いはいずれも平等とする.

表 1 システム側の条件

システムノード数	9216
backfill	なし
離散割当	なし
Fair share	なし
トポロジ	Tree

表 2 ジョブデータの条件

ジョブの最大必要ノード数	1000
ジョブの取得期間	7 日間
ジョブ密度	95%

4.1 ジョブミックス

ジョブミックス [3] とは, 「京」ジョブ統計情報から最小必要ノード数と最大必要ノード数, そしてジョブ密度を設定し, ノード時間積がジョブ密度に達するまでジョブをランダムピックアップして生成されたジョブ群である. ジョブミックスの生成の際, 「京」のジョブデータは必要ノード数, 実行時間で分類されている. それぞれのグループの実際の実行時間やジョブ数, ジョブの投入頻度や分散といった統計情報を用いて, より実データに近い条件でジョブデータは生成される.

4.2 スケジューリングアルゴリズム

本研究での実験でスケジューリングアルゴリズムとして利用した First-Come-First-Served (FCFS) では, ジョブは割り当て優先順位が到着順に定められ, 処理が行われる. FCFS の利点としてはアルゴリズムがシンプルであること, ジョブ処理が公平であることが挙げられるが, ジョブの投入タイミングによっては充填率が著しく悪くなる可能性がある.

そのため, Backfill を有効にすることで, 割り当て優先順が低い場合でも優先度の高いジョブを遅延させずに, かつ前方に実行可能領域を確保できる場合は優先度の高いジョブを追い越して割り当てできるようになる. Backfill の導入の有無によるスケジューリングの違いを図 3 に示す.

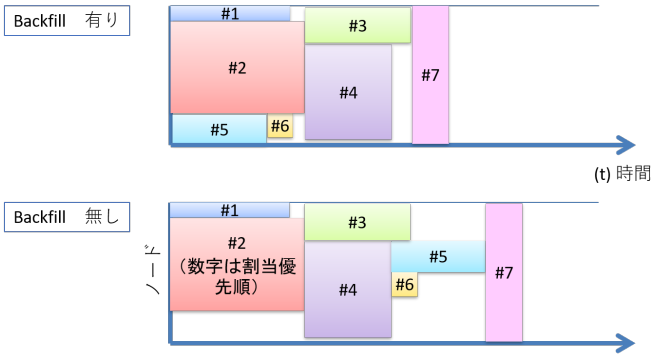

図 3 Backfill の有無によるスケジューリングの違い

4.3 Slurm Workload Manager

Slurm Workload Manager[6] はあらゆる規模の HPC システムで利用されているジョブスケジューリングシステムで, Sequoia (Lawrence Livermore National Laboratory) や Piz Daint (Swiss National Supercomputing Centre) のような TOP500 のスーパーコンピュータにも用いられている. 一般的なリソース管理プログラムと同様にジョブのスループット, システムの利用率, およびジョブの待ち時間に大きく影響する可能性のある多くのパラメータ設定が可能である. ただし, 実験条件やポリシーの変更が実稼働 HPC システムにおけるスケジューラのパフォーマンスに与える影響を確認するには数日から数週間かかる場合があるため, パラメータを変更することでシステムパフォーマンスを調整したり, 新しいポリシー選択を実装したりすることは実用的ではない.

そこで, 本研究では Slurm Simulator[7] を用いる. このシミュレータでは小規模なクラスタ構成であれば, ジョブの構成とパラメータ条件によっては 1 時間で 17 日分のスケジューリングシミュレーションが可能である.

5. 実験結果

5.1 アルゴリズムを適用したジョブスケジューリング

実行領域分割アルゴリズムを適用しなかった場合と適用した場合とでジョブスケジューリングの結果を比較する．表 2 の条件で生成したジョブミックス (以降, Job1 とする) を利用した．アルゴリズムを適用して決定した実行領域条件を表 3 に示す．また, 表 4 に各分割数でのジョブ粒度を示す．分割数を増やすほど各実行領域での最大最小必要ノード数の差は小さくなるため, ジョブ粒度は大きくなる．

表 3 Job1 の実行領域条件

	実行領域サイズ	受け入れノード数上限
2 分割	193/9023	-30/-1000
3 分割	19/1081/8116	-9/-99/-1000
4 分割	11/190/1487/7528	-5/-31/-177/-1000
5 分割	7/66/636/1252/7255	-3/-15/-63/-251/-1000

表 4 Job1 の各分割数でのジョブ粒度

	ジョブ粒度
分割なし	0.001
2 分割	0.030
3 分割	0.101-0.111
4 分割	0.178-0.200
5 分割	0.252-0.333

分割アルゴリズム適用による充填率への影響を確認する．実行領域を分割せずにジョブスケジューリングを実行した際の充填率を図 4 に, 分割アルゴリズムを適用して実行領域を 2 分割した際の充填率を図 5 に示す．この 2 つを比べると分割アルゴリズム適用前後で変化はほとんどないことがわかる．また, 分割数を 5 つに増やしても図 6 に示すように 2 分割と違いはほとんど見られない．これらの結果から, 分割アルゴリズムによる充填率への悪影響はほぼなく, 分割数を増やしても同様であると考えられる．

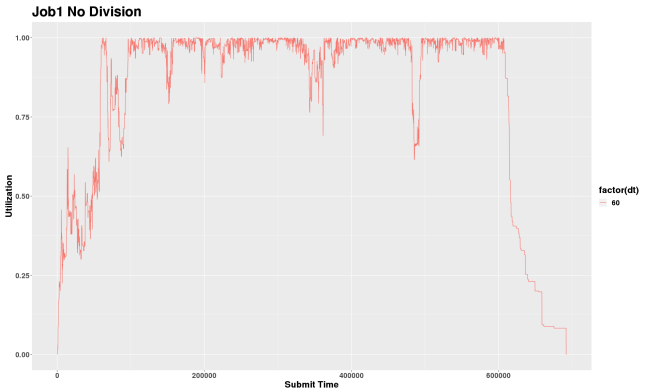

図 4 分割なしの充填率

次に, Job1 の各分割数での最大最小待ち時間の差を図 7 に示す．分割した場合では, 最大最小待ち時間の差が最も大きい分割領域の結果を出力している．アルゴリズムを適用した場合では 2，3，4 分割では分割なしと比べて最大最小待ち時間の差は大幅に小さくなっている．この結果から, 特定のジョブのみ待ち時間が大幅に増加してしまう問題はアルゴリズムの適用によって軽減できていることが確認できる．ただし，5 分割では他の実験結果と比べて最大最小待ち時間の差が大幅に増えてしまっている．この結果を見ると, 分割数を増やせば必ずしも良くなるわけではないことがわかる．

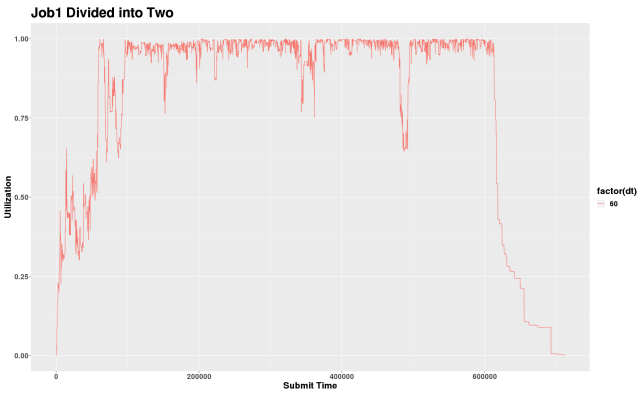

図 5 2 分割の充填率

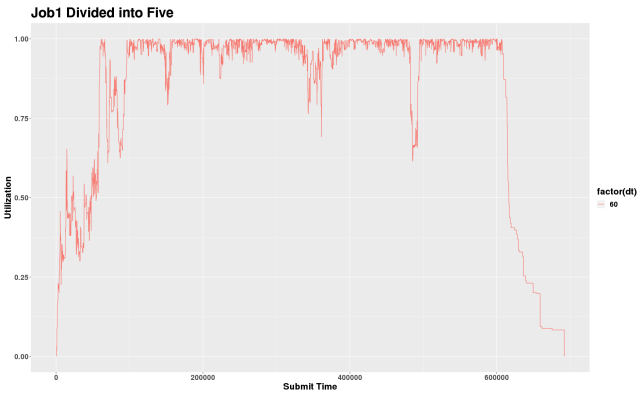

図 6 5 分割の充填率

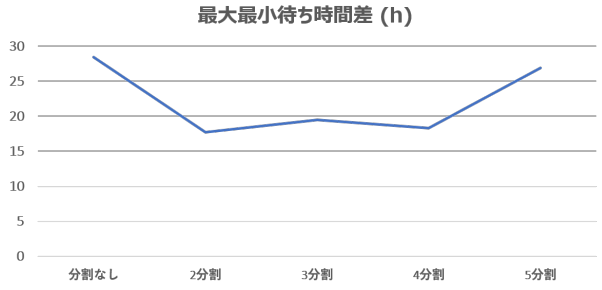

図 7 最大最小待ち時間の差

実行領域を分割せずにジョブスケジューリングを行った場合と, 提案アルゴリズムに則って実行領域を 2 分割した場合での最大最小待ち時間の差を図 8 に示す．受け入れる

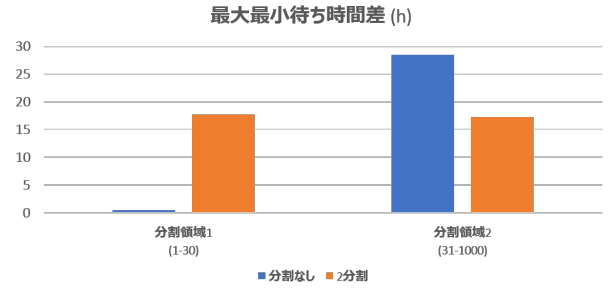

図 8 分割なしと 2 分割の最大最小待ち時間の差

ジョブの必要ノード数範囲が小さい順に番号を振っており, 必要ノード数に応じて投入領域を分けられた各ジョブ群での最大最小待ち時間の差を評価の対象とする. また, 分割なしの実験結果でも, 比較のため分割した場合と同様のジョブの必要ノード数範囲でジョブを分けて最大最小待ち時間の差を見た. 図 8 を見ると, 分割なしの実験結果では, 実行領域 1 の小規模ジョブ群の待ち時間の差は著しく小さいのに対し, 実行領域 2 の大規模ジョブ群では大幅に待たされているジョブが生じているのが確認できる. アルゴリズムに則って実行領域を 2 分割した場合では分割なしと比べると実行領域 1 , 2 の最大最小待ち時間の差の違いは小さく, アルゴリズムの目的であるジョブ規模毎に生じる待ち時間の大きなばらつきの軽減は達成できていると言える. 分

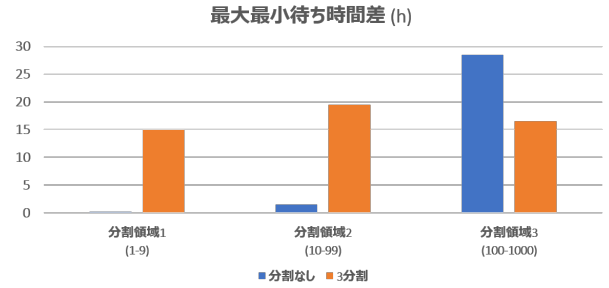

図 9 分割なしと 3 分割の最大最小待ち時間の差

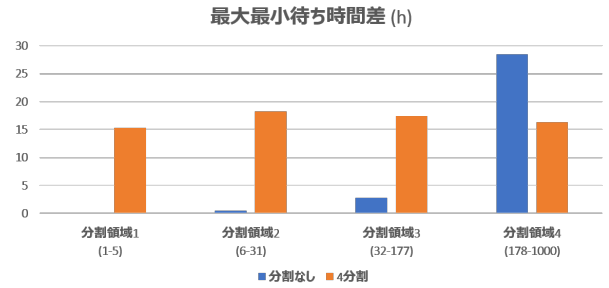

図 10 分割なしと 4 分割の最大最小待ち時間の差

割数 3 , 4 , 5 の最大最小待ち時間の差を図 9 - 11 に示す. いずれの分割数でも分割なしと比べてジョブ規模毎の最大最小待ち時間の差の違いは小さいが, 分割数 5 では最大最小待ち時間の差のばらつきが大きい. これは, 分割数を増

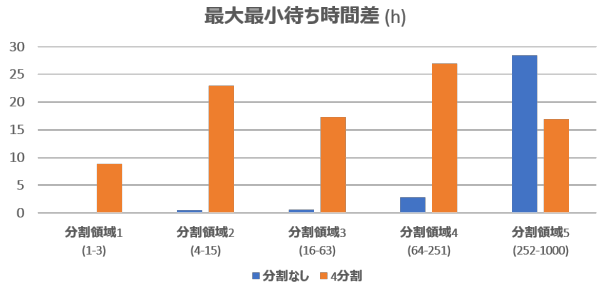

図 11 分割なしと 5 分割の最大最小待ち時間の差

やしたことで各分割領域のジョブ粒度のばらつきが大きくなってしまったためであると考えられる.

5.2 ジョブミックス比較

ジョブ密度と最大最小必要ノード数, 投入期間を Job1 と同じ条件で生成したジョブミックス (以降, Job 2 , Job 3 とする) で同様に実験を行った. ただし, 生成したジョブミックスはそれぞれジョブの投入タイミングや総投入数が若干異なる. 表 5 に生成した各ジョブミックスの詳細を記す.

表 5 各ジョブミックスの詳細

	ジョブ投入数	平均必要ノード数	最大ノード実行時間積
Job1	2515	164.58	23040
Job2	2521	172.81	19008
Job3	2461	172.57	21888

各ジョブミックスの最大最小待ち時間の差を図 12 に示す. Job 2 , Job 3 では分割なしと比べて分割アルゴリズム

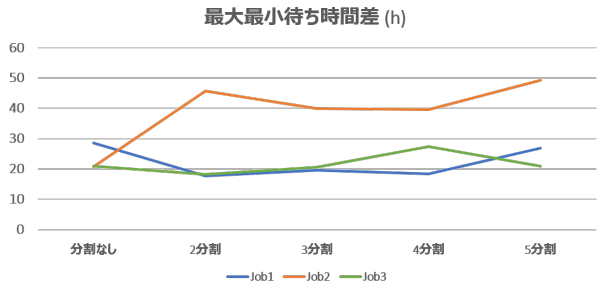

図 12 各ジョブミックスの最大最小待ち時間の差

適用によって最大最小待ち時間の差は小さくなっている. ただし, 最大最小待ち時間の差が最も小さくなる分割数は異なっている. また, Job 1 は Job 2 , Job 3 とは異なり分割しない方が最も最大最小待ち時間の差が短い. これら結果から, 投入ジョブの必要ノード数分布や実行領域, ジョブ密度が同じでもジョブの投入タイミングによって最適な分割数は異なり, 想定されるジョブ規模に大きな変化がなくても定期的にアルゴリズムを適用し分割領域と分割条件を変える必要があると言える.

5.3 分割がうまく機能しないジョブミックスの分析

本項では Job 1 で提案アルゴリズムがうまく機能しなかった原因の分析を行う．図 13-15 は 2 分割適用時の，それぞれ Job 1，Job 2，Job 3 の各ジョブの投入時間と待ち時間をを表している．Job 1 と Job 3 は待ち時間が長いジョブでも 20 時間程度で収まっているのに対し，Job 2 ではジョブ投入開始から 150000 秒頃に待ち時間が 40 時間近いジョブが発生しており，このジョブが最大最小待ち時間の差の悪化の原因であると考えられる．

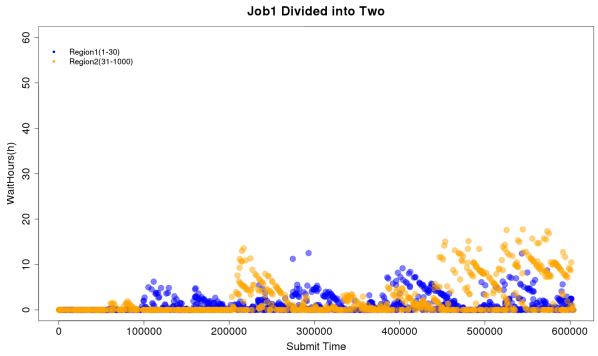

図 13 Job1 のジョブの投入時間と待ち時間

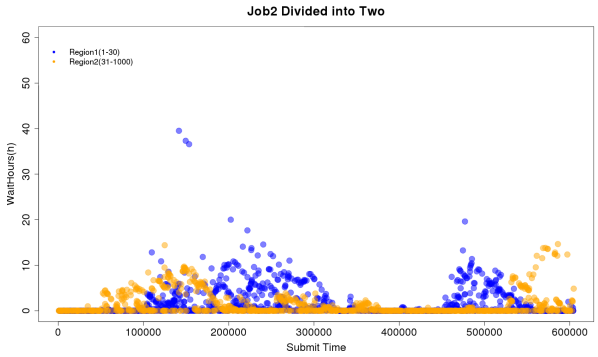

図 14 Job2 のジョブの投入時間と待ち時間

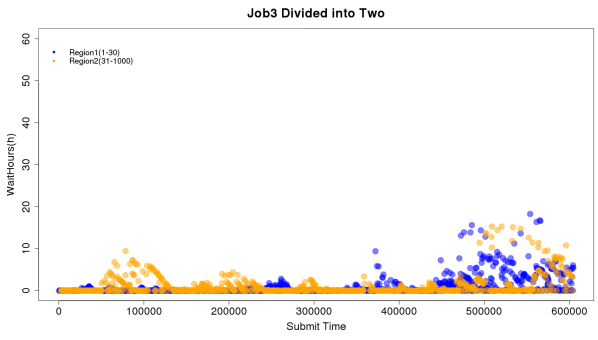

図 15 Job3 のジョブの投入時間と待ち時間

次にジョブのジョブの投入タイミングと投入密度を確認する．Job 2 において最大最小待ち時間の差が大きかった実行領域 2 に限定して，

$$\frac{\sum_{i=0}^{n} NodeJob_i \times ElapseTime_i}{NodeSize \times Interval}$$

の式から一定時間隔での実行領域における投入密度を求める．式において，NodeJob はそのジョブが必要とするノード数，ElapseTime はジョブの実行時間，NodeSize は実行領域のノード数，Interval はジョブを集計する時間隔を表す．Interval は 20000 秒とする．図 16 に 20000 秒間隔で集計した各ジョブの投入密度を示す．Job 2 は 100000 秒と

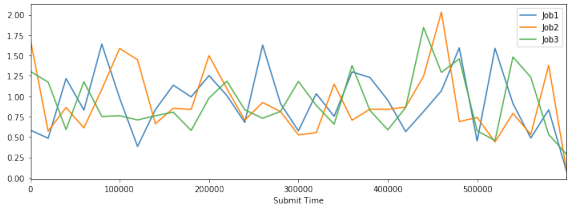

図 16 各ジョブの投入密度

120000 秒で投入密度が高くなっており，実行領域に対してジョブ投入が集中している．また，Job 2 において最も投入密度が高いのは 460000 秒だが，次の時間隔では 0.75 程度まで下がっている．そのため，ある程度の期間投入密度が高くなければ著しく待ち時間が長いジョブは発生しづらいと考えられる．提案アルゴリズムででは分割される実行領域のサイズはジョブの投入間隔を考慮していないため，Job 2 のように投入タイミングによっては最大最小待ち時間の差の改善が見られないことがある．そのため，投入タイミングに偏りが発生した場合の対処を考える必要がある．

6. まとめと今後の予定

あらゆる規模のジョブが投入される環境における大規模ジョブの待ち時間の削減を目的として，過去のジョブ分布やジョブ実行状況を分析し，自動で実行領域の分割数と分割ノードサイズ，実行ジョブのノード数範囲を決定するアルゴリズムの提案と効果検証を行った．提案アルゴリズムでは，過去の投入ジョブ分布とジョブの待ち時間，そして運用者が設定したジョブの最大最小待ち時間の差から，実行領域の分割数，各実行領域が受け入れるジョブの必要ノード数範囲，各実行領域のサイズを決定する．実験結果から，アルゴリズムの適用によって充填率の悪化なく最小待ち時間の差が改善されることが示された．今後の課題として，投入されるジョブ分布や密度が同じ条件でもジョブの投入タイミングに偏りがあると待ち時間が長いジョブが発生し得ることが挙げられる．

そのため，ある程度投入タイミングに偏りがあっても最小待ち時間の差を一定以下に抑えられるようアルゴリズムの改善を検討したい．また，今回利用した「京」のジョブ統計情報から生成されたジョブミックスとは異なるジョブ分布での検証も行いたい．

謝辞

本研究の一部はお茶の水女子大学と富士通研究所との共同研究契約に基づくものである.

参考文献

[1] 「京」の稼働状況. *https://www.r-ccs.riken.jp/jp/k/machine-status/*, Accessed: November 2020.
[2] Georg Zitzlsberger, Branislav Jansík, and Jan Martinovič. Job simulation for large-scale pbs based clusters with the maui scheduler. *BIG DATA ANALYTICS, DATA MINING AND COMPUTATIONAL INTELLIGENCE 2018 THEORY AND PRACTICE IN MODERN COMPUTING 2018*, p. 137.
[3] 山本啓二, 宇野篤也, 関澤龍一, 若林大輔, 庄司文由ほか. 区間スケジューリングを用いたジョブスケジューリングの性能評価. 研究報告ハイパフォーマンスコンピューティング (HPC), Vol. 2014, No. 3, pp. 1–5, 2014.
[4] 滝澤真一朗, 高野了成ほか. 正確な実行時間指定を促すジョブスケジューリング. 研究報告ハイパフォーマンスコンピューティング (HPC), Vol. 2018, No. 2, pp. 1–9, 2018.
[5] 高山沙也加, 関澤龍一, 鈴木成人, 山本拓司, 小口正人. 投入ジョブ情報を踏まえたシステムノードのパーティション分割の考察. 第 12 回データ工学と情報マネジメントに関するフォーラム 論文集, Vol. 2020, pp. H4–4, 2020.
[6] Slurm Workload Manager. *https://slurm.schedmd.com/*, Accessed: November 2020.
[7] Nikolay A Simakov, Martins D Innus, Matthew D Jones, Robert L DeLeon, Joseph P White, Steven M Gallo, Abani K Patra, and Thomas R Furlani. A slurm simulator: Implementation and parametric analysis. In *International Workshop on Performance Modeling, Benchmarking and Simulation of High Performance Computer Systems*, pp. 197–217. Springer, 2017.