様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月　 ６ 日現在

機関番号：１４６０３
研究種目：挑戦的萌芽研究
研究期間：2013～2014
課題番号：２５５４０００７

研究課題名（和文）スケールアウト型並列計算パラダイムに向けたリスク理論的性能評価法

研究課題名（英文）Risk-Theory Based Performance Analysis for Scale-Out Parallel Computing Systems

研究代表者
笠原　正治（Kasahara, Shoji）

奈良先端科学技術大学院大学・情報科学研究科・教授

研究者番号：２０２６３１３９

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：本研究課題では，莫大な数のタスクから構成されるジョブを高速かつ高効率に処理するタスク・スケジューリングの評価に向けたリスク理論的基礎研究を行った．研究成果として，処理開始時点でタスクの複製を生成して並列処理を行うリプリケーション法については，複製数が３程度でジョブ応答時間の改善効果が十分得られることが判明した．また，ある閾値以上の時間が経過しても処理が終了しないタスクを別のワーカマシンに複製・実行することで処理を冗長化するバックアップ・タスクについて極値理論を応用した性能解析を行い，タスク処理時間が裾の重い分布に従うときは小さい閾値の方が電力消費量削減の観点から効果的であることを明らかにした．

研究成果の概要（英文）：This study considered the risk-theory-based analysis of task scheduling for large-scale parallel distributed computing. It was found that in task replication, the optimal number of replications which achieves the shortest task-processing time mainly depends on the coefficient of variation of the worker-processing time, and that three replications are sufficient to guarantee a low variance of the task-processing time. In terms of backup-task scheduling, we analyzed system performance by extreme value theory. It was found that a small deadline time is effective for reducing energy consumption when the subtask-processing time follows a heavy-tailed distribution. In case of light-tailed subtask-processing time, on the other hand, energy consumption can be reduced by a large deadline time.

研究分野：待ち行列理論

キーワード：情報基礎　クラウド・コンピューティング　極値理論　タスク・スケジューリング　性能解析

１．研究開始当初の背景

Google や Amazon, Salesforce.com に代表されるクラウド・コンピューティングでは，非常に多くの低コストサーバマシンで構成した大規模クラスタ上でスケールアウト型の並列計算サービスを提供している．具体的には検索処理やテキスト処理等の高い並列性を有する計算ジョブを非常に多くのタスクに分割し，大規模なワーカマシン群で並列処理を行うことで，高ジョブスループットを実現している．しかしながら，非常に多くのワーカマシンでジョブをタスクレベルで分散実行する場合，ハードディスクやメモリ・CPU といったハードウェアの故障やソフトウェア的な不具合が頻発し，タスク処理時間に大きなばらつきが生じてジョブレベルの応答性能が悪化するという落伍者の問題(Issue of Strugglers)が知られている．今後は 1 千万台オーダーのサーバマシン群からなるデータセンターの運用を計画している Google・Spanner プロジェクトに代表されるように，非常に多くのワーカマシンから構成される超大規模データセンターによるスケールアウト型コンピューティング・サービスが世界規模で展開されることが予想され，このような超大規模データセンターで効率的なジョブ管理やタスク・スケジューリングを実現するためには，巨大なデータ処理を要求するジョブに対する効率的な計算資源の割当てや落伍者の問題を考慮したタスク・スケジューリングの開発が欠かせない．

２．研究の目的

本研究では非常に多くのタスクから構成されるジョブを高速かつ高効率に処理するタスク・スケジューリングの開発評価に向けた，リスク理論に基づく基礎的研究を行う．リスク理論の一分野である極値理論は標本の極値（最大値・最小値）が母集団の大きさとともにどのように変化するのかという漸近的挙動を取り扱う確率統計理論であり，土木工学では大地震の最大震度予測や河川の洪水発生予測に応用され，建築物の設計に役立ててきた．ここでは各ワーカマシンにおけるタスク処理時間の最大値の挙動を極値理論で取り扱うことにより，ワーカマシンの台数がスケールしたときの計算性能劣化予測，消費電力と計算性能のトレードオフ特性の把握，落伍者の問題に対するリプリケーションやバックアップタスク等の対策による性能改善効果について，厳密な理論展開の下で新しい知見の獲得を目指すと同時に，スケールアウト型並列計算のワークフローに忠実なタスク処理モデルについても研究を展開する．

３．研究の方法

(1) 通常並列処理とリプリケーション・スケジューリングによる並列処理の性能解析

研究の第一段階として，並列タスク処理を一回行って全体のジョブ処理が完了する場合に焦点を当て，通常の並列処理の応答時間性能と，各ワーカマシンが処理するタスクの複製を別のワーカマシンにも実行させるリプリケーション法の応答時間性能についての理論解析を行う．具体的には各ワーカマシンにおけるタスク処理時間を独立同一な分布に従うと仮定し，通常並列処理ではタスク処理時間の最大値が従う分布を考え，一方リプリケーション法では同一タスクを二台のワーカマシンで並列実行して早く終了した方の処理時間をタスク処理時間としたときのジョブ応答時間分布の導出を試みる．ここではワーカマシン単体でのタスク処理時間の確率分布として，故障発生間隔モデルとして利用されるワイブル分布やパレート分布に着目し，これらの分布を仮定したときの応答時間分布の陽公式導出を試みる．性能比較の数値実験においては，全体の台数を一定にした場合のリプリケーション法の効果について，ワイブル分布とパレート分布の形状パラメータを様々な場合に設定した数値実験を徹底的に行い，ジョブ応答時間の変化を詳細に調査する．また，リプリケーション法での並列度を二台以上にしたときのジョブ応答時間についても解析を行い，並列度による応答時間性能の改善効果についても検討を行う．

次に極値理論を応用し，台数が非常に大きくなったときの応答時間の漸近分布を通常並列処理とリプリケーション処理の二つについて解析する．ここでは先に導出した厳密解析による分布と漸近分布の差異に対するワーカマシン台数の影響について，タスク処理時間分布の形状パラメータを様々に変えた数値実験を徹底的に行って定量的に比較検証し，極値理論によるアプローチが効果的となるワーカマシン台数やタスク処理時間分布についての知見獲得を目指す．

(2) バックアップタスク・スケジューリングのジョブレベル応答時間改善効果

Google の代表的なソフトウェアサービスフレームワークである MapReduce では，ジョブレベルの応答性能を劣化させる落伍者の問題に対処するため，個々のタスクの処理状況をステップ単位で管理し，ある閾値以上の処理時間が経過しても終了しないタスクの残りステップを別のワーカマシンにも並列的に処理させることで，全体のジョブ応答時間の短縮化を図るバックアップタスクと呼ばれる性能改善手法を導入している[Dean 他 2008]．ここでは単一の並列タスク処理に対

するバックアップタスクの性能改善効果について，極値理論に基づくジョブ応答時間の近似解析を行う．具体的には，各ワーカマシンにおけるタスク処理時間に対して閾値を設け，タスク処理を行っているワーカマシン群の内，タスク処理にかかる時間が閾値を超えたワーカマシンに対してそのタスク全体の複製を代替ワーカマシンで実行させる場合を検討する．このモデルに対して極値理論に基づくジョブ応答時間の漸近分布解析を行い，タスク処理時間に対する閾値が全体のジョブ応答時間に与える影響を数値実験により精査する．一方，処理を行ったすべてのワーカマシンのタスク処理時間を積算した総タスク処理時間はワーカマシン群で消費される総電力量と密接に関連することから，総タスク処理時間分布についても解析を行い，ジョブ応答時間と総消費電力量の間のトレードオフ関係に与えるタスク処理時間閾値の影響についても数値実験を通して徹底的に調査を行う．

(3) データ・センターレベルの電力消費量削減を目指したサーバ電源管理法

近年のクラウド・コンピューティング・サービスの急速な普及により，大規模データセンターの需要が高まりつつある．データセンターでは，莫大な数のサーバ運用に伴う電力消費が問題となっている．この電力消費を抑える機構として，米国カリフォルニア大学バークレー校で開発された BEEMR (Berkeley Energy Efficient MapReduce) と呼ばれる MapReduce ワークロード・マネージャがある．BEEMR では，莫大な計算リソースを必要とするバッチ・ジョブをキューに一度蓄積し，一定の制御周期毎に一斉にサービスを開始する．周期内に全ジョブへのサービスが終了すると，サーバの電源を切断することで節電を図る．ここでは BEEMR 型電源管理機構を持つデータ・センターにおけるジョブの応答時間や電力消費量について性能解析を行い，制御周期やサーバ配置法が性能評価量に与える影響を定量的に検討する．

４．研究成果

本研究で得られた成果は以下の三点である．

(1) 大規模分散並列処理環境におけるタスクリプリケーション法の性能改善効果

タスク・リプリケーション法では，大規模なジョブを複数のタスクに分割した後，各タスクの複製を生成して別々のワーカマシンで処理をさせることにより，落伍者の問題を回避している．ここではタスク・リプリケーション法のジョブ応答時間に対する改善効果について，確率モデルを用いた分析を行った．具体的には，複数の確率変数列の最大値が漸近的に従う極値分布の理論を応用し，分割されたタスクの処理時間が従う確率分布として超指数分布，ワイブル分布，パレート分布の三種類を考え，それぞれの分布に対してジョブ応答時間の期待値と分散を近似的に導出した．モンテカルロ・シミュレーションとの比較実験により，ワーカマシンの台数が多いときには近似式が高精度の数値を与えることを確認した．また数値実験結果より，タスク複製の効果はワーカマシンの処理時間分布に大きく依存すること，及び複製数が３の場合でタスク処理時間の分散を極めて低く抑えることができることを確認した．

(2) バックアップタスク・スケジューリングにおけるデッドライン時間の効果

MapReduce では，ワーカマシンでタスクの処理時間がある制限時間（デッドライン時間）を超えても処理が完了しない場合，別のワーカマシンで残りのタスクをバックアップ的に実行するバックアップタスクと呼ばれるスケジューリングが行われている．一般に，デッドライン時間が短いとバックアップが早く開始されることになり，ジョブ全体の処理時間は短縮できる一方で電力消費量が増大する．逆に長いデッドライン時間では電力消費量を抑制できる一方でジョブの応答時間が増大する．このバックアップタスクにおけるデッドライン時間の性能改善効果を確率モデルによる分析を通じて解明した．具体的にはタスク・スループットとシステム全体の電力消費量に相当する全ワーカ稼働時間について，極値理論を応用した近似式を導出し，数値実験より近似式の妥当性を確認した．数値例より，サブタスク処理時間が裾の軽い分布に従っているときは，デッドライン時間を大きくした方が電力消費量を削減できること，一方で裾の重いサブタスク処理時間分布のときには小さいデッドライン時間の方がスループットを改善し，かつ電力消費量も抑制できることが判明した．

(3) 電力消費量削減を目指したサーバ電源管理法

BEEMR と呼ばれる電源管理法では，サーバ群はインタラクティブ型とバッチ型の二種類に分類され，インタラクティブ型サーバ群では処理時間の短いジョブが連続的に処理される一方で，バッチ型サーバ群では一定の制御サイクル期間でバッチ型ジョブを処理する．バッチ型サーバ群ではバッチ型ジョブの処理がすべて終了した時点から制御サイクル期間終了時点までの期間，すべての電源が切断され，次の制御サイクル期間開始時点で新規バッチジョブが到着している場合には電源をオンにし，処理を再開する．この BEEMR 型電源管理法の性能を解析するため，二種類のサーバ群を独立な二つの待ち行列

系でモデル化し，ジョブ応答時間と平均電力消費量を導出した．具体的には，インタラクティブジョブを処理するサーバ群は，プロセッサシェアリングサービス規範を持つ単一サーバ待ち行列でモデル化を行い，一方バッチジョブを処理するサーバ群は一定時間毎にゲートの開閉を行うプロセッサシェアリング型単一サーバ待ち行列でモデル化した．これらの確率モデルよりインタラクティブジョブとバッチジョブの平均応答時間，および平均電力消費量を導出した．数値例より，制御サイクル期間およびバッチ型サーバ群のサーバ数がインタラクティブジョブの応答時間と平均電力消費量に与える影響を定量的に明らかにした．

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計 3 件）

1. Kato, M., Masuyama, H., Kasahara, S., and Takahashi, Y., “Effect of Energy-Saving Server Scheduling on Power Consumption for Large-Scale Data Centers,” Journal of Industrial and Management Optimization 採録. （査読有）
2. Hashimoto, K., Masuyama, H., Kasahara, S., and Takahashi, Y., “Performance Analysis of Backup-Task Scheduling with Deadline Time in Cloud Computing,” Journal of Industrial and Management Optimization, vol. 11, no. 3, pp. 867-886, July 2015.（査読有）
DOI:10.3934/jimo.2015.11.867
3. Hirai, T., Masuyama, H., Kasahara, S., and Takahashi, Y., “Performance Analysis of Large-Scale Parallel-Distributed Processing with Backup Tasks for Cloud Computing,” Journal of Industrial and Management Optimization, vol. 10, no. 1, pp. 113-129, 2014.（査読有）
DOI:10.3934/jimo.2014.10.113

〔学会発表〕（計 8 件）

1. Kato, M., Masuyama, H., Kasahara, S., and Takahashi, Y., “Performance Analysis of Energy-Saving Server Scheduling Mechanism for Large-Scale Data Centers,” The 9th International Conference on Queueing Theory and Network Applications (QTNA2014), Bellingham (USA), pp. 28-35, 18-21 August, 2014.（査読有）
2. Hashimoto, K., Masuyama, H., Kasahara, S., and Takahashi, Y., “Effect of Deadline Time on Job Completion Time for Backup-Task Scheduling in Cloud Computing,” The 8th International Conference on Queueing Theory and Network Applications (QTNA2013), Taichung (Taiwan), pp. 23-29, 30 July - 2 August, 2013.（査読有）
3. 笠原正治, “大規模分散並列処理系におけるバックアップタスク型ジョブ・スケジューリング,” 日本ＯＲ学会 「不確実性システムにおける意思決定」研究部会 第６回研究会, 関西学院大学・大阪梅田キャンパス（大阪府大阪市）2014.4.12.
4. 加藤将隆，増山博之，笠原正治，高橋豊, “MapReduce フレームワーク用サーバ管理機構 BEEMR の省電力効果解析,” インターネット技術第 163 委員会 (ITRC) 新世代ネットワーク構築のための基盤技術研究分科会 (NWGN) ワークショップ (ITRC-NWGN 2013), 紀三井寺ガーデンホテルはやし（和歌山県和歌山市）, 2013.9.5.
5. 平井嗣人，増山博之，笠原正治，高橋豊, “大規模分散並列処理システムにおけるハートビート式故障検知の性能解析,” インターネット技術第 163 委員会 (ITRC) 新世代ネットワーク構築のための基盤技術研究分科会 (NWGN) ワークショップ (ITRC-NWGN 2013), 紀三井寺ガーデンホテルはやし（和歌山県和歌山市）, 2013.9.5.
6. 橋本恭佑，増山博之，笠原正治，高橋豊, “大規模分散並列処理システムにおけるデッドライン・タイム型バックアップ・タスクの性能改善効果,” インターネット技術第 163 委員会 (ITRC) 新世代ネットワーク構築のための基盤技術研究分科会 (NWGN) ワークショップ (ITRC-NWGN 2013), 紀三井寺ガーデンホテルはやし（和歌山県和歌山市）, 2013.9.5.
7. 加藤将隆, 増山博之, 笠原正治, 高橋豊, “データセンタにおける消費電力低減を目的としたサーバ管理スケジューリング方式の性能解析,” 日本オペレーションズ・リサーチ学会 2013 年度秋季研究発表会, 徳島大学・常三島キャンパス（徳島県徳島市）, アブストラクト集, pp.30-31, 2013.9.11-12.
8. 佐嘉田智之, 笠原正治, “サービス受付時の背後過程に依存したサービス時間をもつ複数サーバ待ち 行列,” 2014 年度確率モデルシンポジウム, 東北大学・片平さくらホール(宮城県仙台市), pp. 122-131, 2015.1.23.

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計　0　件）

○取得状況（計　0　件）

〔その他〕
ホームページ等
http://www-lsm.naist.jp/~kasahara/index.html

６．研究組織
(1)研究代表者
笠原　正治　（KASAHARA SHOJI）
奈良先端科学技術大学院大学・情報科学研究科・教授
研究者番号：20263139

(2)研究分担者
該当者なし

(3)連携研究者
増山　博之（MASUYAMA HIROYUKI）
京都大学・大学院情報学研究科・准教授
研究者番号：60378833

橘　拓至（TACHIBANA TAKUJI）
福井大学・大学院工学研究科・准教授
研究者番号：20415847