様式Ｃ－１９

科学研究費助成事業（科学研究費補助金）研究成果報告書

平成　２４年　６月　１３日現在

機関番号：82645
研究種目：基盤研究(C)
研究期間：2009　～　2011
課題番号：21500064
研究課題名（和文）　将来予測に基づくスーパーコンピュータの運用効率化ツールセット構築のための研究
研究課題名（英文）　A research and tool-set development for efficient operation of supercomputers based on workload prediction
研究代表者
伊藤　利佳（ITO　RIKA）
独立行政法人宇宙航空研究開発機構・研究開発本部・主任開発員
研究者番号：70442928

研究成果の概要（和文）：
スーパーコンピュータのハードウェア資源の利用効率を上げながらコストを削減することを目的とし，利用効率向上のためのアルゴリズムの開発と省電力への取組みを行った．
ジョブの経過時間予測や分析を行い，これに数学的手法を組み合わせることにより，利用効率が向上するアルゴリズムの開発を行った．省電力の取り組みとしては，京都大学で平成２４年から稼動予定のスーパーコンピュータシステムの省電力運転のための電源およびジョブスケジューリングの方式を設計した．これにより，消費電力の削減とQOSの維持を両立する方式が得られた．

研究成果の概要（英文）：
The purpose of our project is research and tool-set development for efficient operation of supercomputers and for saving power consumption. We analyzed each job tendency and predicted the elapsed time of jobs, and developed job scheduling algorithms based on them. As for the power saving operation, we designed a power scheduling and a job scheduling for the new supercomputer system whose operation will start from 2012 in Kyoto University. This new design enables us to achieve both of power saving and QOS.

交付決定額

（金額単位：円）

	直接経費	間接経費	合　計
2009 年度	1,000,000	300,000	1,300,000
2010 年度	700,000	210,000	910,000
2011 年度	600,000	180,000	780,000
年度			
年度			
総　計	2,300,000	690,000	2,990,000

研究分野：総合領域
科研費の分科・細目：情報学・　計算機システム・ネットワーク
キーワード：ハイパフォーマンスコンピューティング

１．研究開始当初の背景

現在のスパコンは，複数のノードをネットワークで接続したクラスタシステムが一般的であり，複数のユーザが１つのシステムを共有する運用形態をとるため，システム資源の利用効率向上のための資源配分のみならず，ユーザ間の公平性などのユーザの満足度にも留意する必要がある．また，近年のシステムの大規模化にともなう電力消費の増加が，環境面，コスト面に対して深刻な問題となっている．これをハンドリングするための技術の一つが，ジョブスケジューラである．

ジョブスケジューラは，できるだけ資源が有効に活用されるよう，ユーザから投入されたジョブを，優先順位など多数のパラメータから成るルールに基づきジョブを実行する．

しかし，パラメータ設定によっては，空き資源があるにも関わらず，ジョブがキューに滞留したり，特定のユーザが長い時間待たされたりするという問題が起きる．これは公平性だけでなく，資源活用の観点からも大きな問題である．さらに，ジョブの投入状況には再現性がないため，他のパラメータとの比較によって最適性を検証することができないという困難さのため，置きざりにされてきた課題でもある．この問題を解決するためには，ジョブの投入状況に応じて，システム利用効率やスループットを向上させるようなパラメータ最適化ツールの構築，経過時間予測を含めた統計分析，省電力を目標としたスケジューリング方式の開発などがあげられる．

２．研究の目的

本研究の目的はスーパーコンピュータのハードウェア資源の利用効率を上げながら，同時にコストを削減することを目的とした研究である．研究の大きな概念としては，システムが高負荷である時にはジョブをできる限り稠密させて効率よく資源配分をし，逆に低負荷である時には省電力のための縮退運転を行うというもので，これらを組み合わせることによって，ハードウェア資源の利用効率の向上と省電力を同時に実現するシステムの構築を目的とする．

３．研究の方法

(1)ユーザ分析

統計的手法を用いてユーザ毎の傾向分析やジョブの統計分析を行い，各ジョブの経過時間予測を行うための手法を検討するにあたって，それらの結果を反映させて数値実験をおこなった．また，パラメータ設定の設定変更のタイミングをはかるための実験を行い，設定変更のタイミングについての検討をおこなった．また，各ユーザのジョブの実際の経過時間とジョブ投入時にユーザが設定する要求経過時間との間の乖離に関する詳細な分析を行った結果，両者の間には，特徴的な傾向が見られることが明らかになった．以上の結果を論文としてまとめ，国内外で発表した．

(2)経過時間予測

JAXAスーパーコンピュータにおけるユーザのジョブの傾向分析を行い，ユーザのジョブの経過時間予測に関する数値実験を行った．予測に関していくつかの手法を試みた中で，比較的結果の良かった移動平均法，Bootstrap法，Box-Muller法のそれぞれの方法でジョブの経過時間予測を行った．その結果，経過時間予測に適したアルゴリズムを特定することができ，さらに，経過時間予測をスケジューリングに取り入れることによってシステム効率が向上するかどうかを数値実験によって確認した．

(3)資源の稠密化に関する研究

システム資源をより効率よく活用するために，スーパーコンピュータのハードウェア資源の配分問題を，2 次元詰込み問題として定式化を行い，さらに，できるだけ稠密に資源を配分する改善方法を提案し，数値実験を用いてその有効性を確認した．

４．研究成果

(1)ユーザ分析の実験結果

JAXA におけるユーザのジョブの最大待ち時間と平均待ち時間，システム利用効率の相関分析を行い，３つの要素の相関関係を明らかにした．また，ジョブスケジューラのひとつである NQS のパラメータ設定変更の効果を検証するために，各パラメータの分散分析を行い，各パラメータにおける設定変更がシステム利用効率に大きな影響を与えることを確認した．さらに，ユーザがジョブの投入時に要求するジョブの経過時間（要求経過時間）と実際の経過時間（実経過時間）(log)の標準偏差を調査し，それぞれの時間のばらつきの検証を行った．その結果，要求経過時間の標準偏差に関しては，72%のジョブの標準偏差が小さく，多くのジョブが要求経過時間を変更されずにシステムに投入されていることがわかった．一方，実際の経過時間に関しては 60%以上のジョブの標準偏差が大きく，かなりのばらつきがあることが明らかとなった．図 1 においては要求経過時間の標準偏差を示し，図 2 は実経過時間の標準偏差を示す．

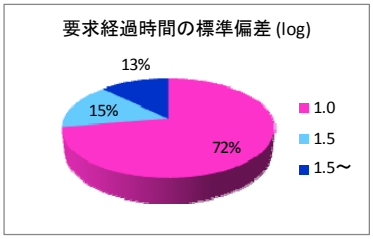

図 1　要求経過時間の標準偏差

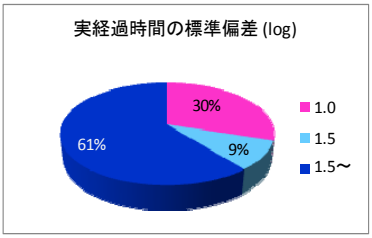

図 2　実経過時間の標準偏差

(2) 経過時間予測の実験結果

現実の運用においては，実際の経過時間が不明なことがジョブのスケジューリングを難しいものにしている，そこで，統計的手法を用いて経過時間予測を行った，比較的結果の良かった移動平均移動平均法，Boot strap 法，Box-Muller 法のそれぞれの方法でジョブの経過時間予測を行った結果，他の方法に比べて Box-Muller 法を用いるともっとも実際の値に近くなるという結果が得られた．この方法には各ユーザの平均および標準偏差を用いる．そのため，ある程度ユーザの特徴がつかめる場合においては，他の方法に比べて，それらの特徴を容易に反映できることが予測に有利に働いたと考えられる．図 3 は経過時間と予測時間の差の平均値を示すもので，図 4 は経過時間と予測時間の差の最大値を示す．どちらの図も値が小さいほうがより予測の精度が高いことを示す．

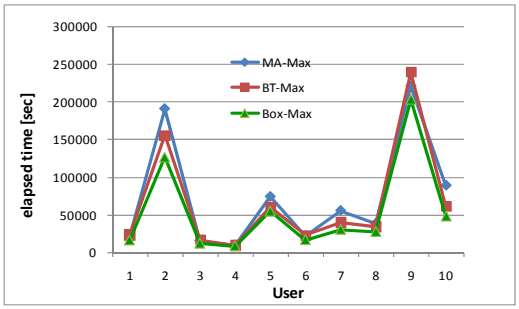

図 3　経過時間と予測時間の差の平均値

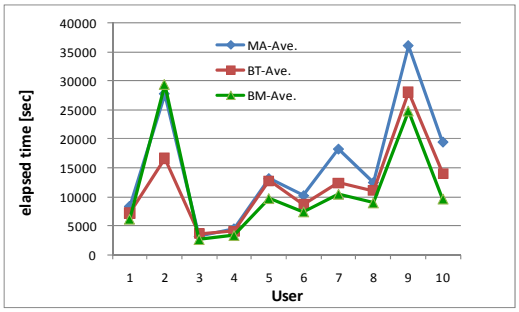

図 4　経過時間と予測時間の差の最大値

図 5 予測に基づくシステム利用効率

また，図 5 は 6 カ月分のデータに関して，FIFO でスケジューリングを行った場合と予測を導入してスケジューリングを行った場合の比較である，600 秒ごとにリスケジューリングを行っているが，予測の精度が高い場合には，およそ 3%の利用効率の向上になることが明らかになった．

(3) 資源の稠密化に関する実験結果

スーパーコンピュータのハードウェア資源をできるだけ有効に稠密させてジョブ実行を行うために，ハードウェアの配分問題を 2 次元詰込み問題として定式化し，数値実験を行うとともに，改良案を提案して同様に数値実験を行った．

2 次元詰込み問題は，ある決められた母材の上に決められた個数の長方形を配置する問題で，　その際に，長方形は互いに重ならないようにすることによって，限られた資源をより有効に活用しようとするための手法である．この手法を用いて実際にあった過去のデータを用いて数値実験を行った結果，従来法で資源配分を行った場合と比較して，提案法を用いた場合にはシステム利用効率が向上し，すべてのジョブが完了するまでの終了時間も短縮することが確認できた．図 6 および図 7 は従来法で配分を行った場合と提案法で配分を行った場合の比較である．

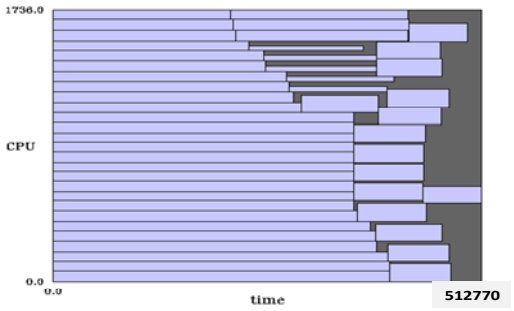

図 6 システム利用効率（従来法）

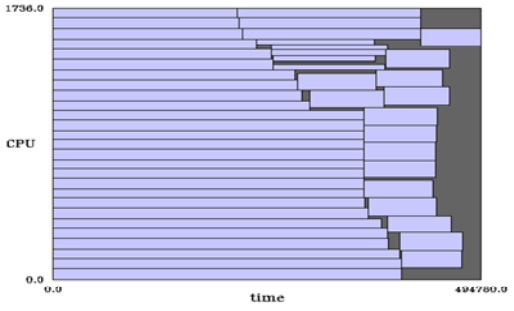

図 7 システム利用効率（提案法）

図 6 と比べ，図 7 においては，図 6 の空き領域に対して，ジョブがより稠密し，全ジョブが終了する時刻も短縮されていることが示されている．

(4) 省電力スケジューリング
省電力のためにスーパーコンピュータの一部の計算ノードの電源を切ることは，一般にQOS の維持と矛盾するが，個々のユーザに対するSLA を高確率で遵守しつつ稼動ノード数を減らすことは必ずしも不可能ではない，京都大学のスーパーコンピュータシステムでは，各ユーザグループに対して一定量の計算ノードを常時保証するSLA に基づくジョブスケジューリングが行われており，これを厳密に解釈・遵守して保証量に見合うノード数を常にhot standby 状態とすると，稼動ノードを減らした省電力運転はほぼ不可能となる，そこで，短時間に限ってSLA に違反するスケジューリングを故意に実施し，それがユーザのジョブ実行に真に影響するか，すなわち本来保証されているはずの資源が得られないために実行不能となるジョブが生じるか否かを確認する実験を１年間にわたって実施した，その結果，ノードの稼働率（ジョブが実行されているノード数の割合）が 85%未満である場合には，１時間以内に解消されるSLA 違反スケジューリングの影響が皆無であることが確認された，そこで，ノード稼働率に基づき，(a) １時間以内に完了するジョブをSLA に違反してスケジューリング可能とする，(b)稼働率が 85%を下回らない範囲で遊休ノードの電源を切る，という電源・ジョブスケジューリング方式を設計し，2012 年５月より稼動予定のスーパーコンピュータシステムに適用することとした，

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 Xavier OLIVE, Hiroshi NAKASHIMA ,Efficient Representation of Constraints and Propagation of Variable-Value Symmetries in Distributed Constraint Reasoning, J.Information Processing, 査読有,Vol.19, pp201-210, 2011.
② Rika. Ito, Kenichi Kikuchi, An Analysis for Parameter Configuration To Find a Trigger of Change, ICIC Express Letter, 査読有, Vol.4, Num.3, 2010, pp1959-1964.
③ Rika. Ito , An Analysis for Parameter Configuration of Job Scheduler Focusing on Users, ICIC Express Letter, 査読有 Vol..3, Num.4, 2009, pp1387-1392.

〔学会発表〕(計 12 件)
① Rika. ITO, Kenichi KIKUCHI, Naoyuki FUJITA,“Job Scheduling Method Based on Estimation of Elapsed Time and User Analysis”,PDP2012, 査読有, 2012.02.17, Munich, Germany.
② Hiroshi Nakashima, New Supercomputer System in Kyoto University, SC11 Conference(Exhibition),2011.11.12-18,Seattle,WA,USA.
③ 伊藤利佳,菊地賢一,“ユーザの要求経過時間における解析的検証(An analytical study for User elapsed time)”, 電子情報通信学会ソサイエティ大会, 2011.09.15, 北海道.
④ Rika ITO, Kenichi KIKUCHI,“Efficient Algorithm for hardware resource management using 2-D strip packing problem”, ISII2011, 査読有, 2011.05.03, 青島, China.
⑤ Rika ITO, Kenichi KIKUCHI, Naoyuki FUJITA , An Analysis for Estimation of Elapsed Time of Job Scheduler, PDP2011, 査読有, 2011.02.11,Ayia Napa, Cyprus.
⑥ Hiroshi NAKASHIMA High-Performance as a Service: Cloudy Supercomputing in Kyoto, 9th AEARU Web Technology and Computer Science Workshop,2011.1.17, 京都.
⑦ Rika. ITO, Kenichi Kikuchi, An Analysis for Parameter Configuration To Find a Trigger of Change, ISII 2010 , 査読有, 2010.09.05,大連, China.
⑧ 伊藤 利佳, コンピュータ資源活用のための解析的検証, 電子情報通信学会総合大会, 2010.03.17, 仙台.
⑨ 中島 浩, T2K@京大とプログラム高度化共同研究, PC クラスタワークショップ, 2010.2.19, 京都.
⑩ Rika ITO, Job scheduler parameter analysis for evaluation of effectiveness, PDP2010. 査読有, 2010.02.19, Pisa, Italy.
⑪ 中島 浩, T2K オープンスパコン@京大とそのコラボレーション, インターネットコンファレンス 2009, 2009.10.26,京都.
⑫ Rika ITO, An Analysis for Parameter Configuration of Job Scheduler Focusing on Users. ISII2009, 査読有, 2009.09.15, 秦皇島, China.

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

〔その他〕なし

６．研究組織

(1)研究代表者

伊藤　利佳（ITO RIKA）
独立行政法人宇宙航空研究開発機構・
研究開発本部・主任開発員
研究者番号：70442928

(2)研究分担者

中島　浩（NAKASHIMA　HIROSHI）
京都大学・学術情報メディアセンター・
教授
研究者番号：10243057

(3)連携研究者

なし　（　　）