様式 C-19

科学研究費補助金研究成果報告書

平成２１年　５月　８日現在

研究種目：若手研究（B）
研究期間：2007～2008
課題番号：19700052
研究課題名（和文）　有機的に連携した負荷・実行時間予測と予測を利用する負荷分散システムの構築
研究課題名（英文）　Cooperative Load and Runtime Prediction and Load Balancing System exploiting the Predictions

研究代表者
菅谷 至寛（SUGAYA YOSHIHIRO）
東北大学・大学院工学研究科・助教
研究者番号：80323062

研究成果の概要：効率的な並列・分散処理のためには適切な資源管理が重要だが，環境によっては見かけの計算能力が変動することがある．効率的な負荷分散を実現するために，本研究では負荷予測及びタスク実行時間予測を利用する資源管理システムを構築した．負荷時系列による予測とタスク実行時間予測に基づく負荷予測を併せて用いることで，様々な状況で従来手法よりも高精度な負荷予測を実現できる．さらに，計算資源の分散管理手法や協調型ジョブ管理手法についても検討を行った．

交付額

（金額単位：円）

	直接経費	間接経費	合　計
２００７年度	900,000	0	900,000
２００８年度	2,000,000	600,000	2,600,000
年度			
年度			
年度			
総　計	2,900,000	600,000	3,500,000

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：負荷予測，実行時間予測，負荷分散，並列処理，グリッドコンピューティング

１．研究開始当初の背景

近年の計算機やネットワークの低価格化と普及により，多数の計算機を用いて計算を行うクラスタコンピューティングやグリッドコンピューティングが注目されている．そのような環境では，計算機やネットワークの構成が非均質になることが少なくない．効率的な並列計算のためには，適切な資源管理に基づくタスク割当てが重要である．また，このような計算機ネットワークでは負荷が動的に変化することも考えられるため，負荷変動に耐えるタスク割当てやスケジューリングが必要である．特に，計算機資源を他のユーザーと共有する環境の場合，他のユーザーによって投入されたタスクの影響による負荷変動や，それによる見かけの計算能力の変化を考慮する必要がある．

そのためには，負荷時系列予測やタスク実行時間予測を行い，それに基づいた資源管理や負荷分散を行うことが有効であると考え

られる．負荷予測を含む従来の資源管理システムとしては，Network Weather Service (NWS) や Resource Information Server (RIS)などが挙げられる．NWS では基本的に短期予測のみを行っており，測定間隔を補完する程度に過ぎない．RIS では短期予測に加えて長期予測も行っているが，過去の負荷時系列パターンのみによって予測を行っているため十分な精度が得られていない．

２．研究の目的

本研究は，非均質構成の計算機環境においても効率の良い並列計算を可能にするための複合的な負荷分散システムの構築を目的とし，「負荷予測とタスク実行時間予測」および，それらを利用した資源管理・負荷分散システムを開発する．負荷予測は１ステップ先だけではなく，数ステップ先までの負荷変動傾向を予測すること（長期予測）を目的とする．ある程度の精度で比較的長期の負荷変動傾向を予測できれば，この情報をスケジューリングやタスク割当てに用いたり，あるいはユーザーや管理者に提供したりすることで，より適切な資源利用が可能になると期待される． ただし，予測にはおそらく限界があり，いつでも高い精度で予測できるとは限らないと思われる．そこで，予測精度の推定についても検討を行う．

３．研究の方法

前節で述べた目的に照らし，複合的な負荷分散システムの構築のために必要な要素技術として，以下の５つの項目に関する研究・開発を行った．

(1) タスク実行時間予測アルゴリズム

本研究で提案している手法は Smith らの手法を基礎とし，改良・発展させたものである．予測精度の向上を目指し，類似性テンプレートの導出，仮予測結果の統合手法，実行中に変化するタスク属性の利用などについて検討および改良を行った．

(2) 長期負荷予測アルゴリズム

我々は，類似法を改良したプロセス検索法を提案しているが，オンデマンドシステムとして実装するためには不十分な点があった．プロセス検索法がオンデマンドに応答できるようにするための検討と改良を行った．

(3) 負荷予測・タスク実行時間予測システムの実装と評価

(1)，(2)の成果を踏まえ，負荷・タスク実行時間予測システムの実装を行った．本システムは予測対象マシンに常駐するため，本システム自身が必要とする計算資源や発生する負荷は，無視できる程度でなければならない．実装にあたり，その点に考慮した設計を行った．

(4) 計算資源の分散管理手法

地理的に分散した多数の資源に対し適切なジョブ配置を行うためには，ホスト間遅延などの地理的情報を考慮しつつ，ホストの性能や負荷情報などを共有することが必要になる．しかし，時間的に変化する非常に多くの資源を効率的かつスケーラブルに管理するのは容易ではない．集中的な管理手法では対故障性やスケーラビリティに関して問題が生じる可能性がある．そこで本研究では，計算資源を P2P ネットワークによって自律的に分散管理する手法についての検討を行った．

(5) 協調型ジョブ管理手法

並列分散環境では複数のユーザーが資源を共有することがあり，資源の割当てのためにジョブ管理システムが用いられることが多い．しかし，一般にジョブ管理システムはシステム全体の効率の向上を目的としてシステム管理者が設置するものであり，各ユーザーのニーズに従った柔軟な割当てに応えることは難しい．本研究では，ユーザーの要求に配慮した柔軟な運用を目的とし，ユーザー間の協調によってスケジューリングを行うジョブ管理エージェントについて検討を行った．

４．研究成果

(1) タスク実行時間予測アルゴリズム

本手法は多くの従来手法と同様に，想定する環境下において，属性が類似しているタスクの実行時間は同程度ある場合が多いという事実に基づいている．実行時間を予測したいタスクに類似しているタスクの履歴を取得し，平均などの統計量を求めることで実行時間を推定することができる．

これらの処理において，予測精度向上のためには「タスクの類似性」を求めるためのテンプレート（タスク属性の集合）の構築が重要となる．従来手法では予測を実際に多数回行って，貪欲法または遺伝的アルゴリズムでテンプレート求めていたが，時間がかかるため事前に求めておく必要がある．提案手法では，相互情報量を利用して複数のテンプレートを導出することで複数の仮予測を行い，テンプレートに含まれる属性数と信頼度に基づいて，最終的に採用する仮予測結果の選択を行う．信頼度はt分布から算出され，仮予測の選択に利用するだけでなく，信頼性が低いと推定される予測結果を予めリジェクトするために用いることもできる．また，本手法でのテンプレート生成は比較的高速であるため，予測システム実行中にセミオンラインでテンプレートを更新可能である．シミュ

レーション実験によって，従来手法で最も精度の高い手法であるIBL（Sengerら，2004 年）よりも，本手法の方が良い結果が得られることを確認した（表 1）．

また，メモリ使用量のようにタスク実行中に変化してしまう情報は，類似タスクを見つけるための情報として，これまでは利用できなかった．本研究では，各タスクのメモリ使用量の変動を利用するための方法を開発し，これを用いることで予測精度が向上することを実験によって確認した．

予測手法	誤差率分布(%)	
	0 - 20 %	90 - %
Tp16p	73.28	9.25
IBL	70.50	10.31

表 1　CPU 実行時間予測結果．予測誤差率がそれぞれ 0〜20%の範囲，および 90%以上であった割合を示す．Tp16p は提案手法，IBL は従来手法である．

(2) 長期負荷予測アルゴリズム

我々が提案している長期負荷予測アルゴリズムの一つであるプロセス検索法が，オンデマンドに応答できるようにするための検討および改良を行った．

プロセス検索法は，長期予測が可能な従来手法の一つである類似法を改良したもので，実行中タスク集合および負荷時系列が，現在の窓内でのそれに似ているものを，履歴から類似度順に複数個取得する．取得された類似窓の直後の系列を平均した系列が，本手法による長期予測結果となる．これまで，本手法では予測開始時刻とサンプリング時刻が一致することを暗黙のうちに仮定していたが，サンプリングは一定間隔ごとに行われるのに対し，予測要求は任意に発生するため，実際の応用ではそれらが一致しないことも考えられる．予測システムを実装する際に問題となるため，その場合の対策と影響についての検討を行った．実験の結果，予測要求時に新たにサンプリングを行ってサンプリング間隔のずれを単に無視しても，予測精度への影響は少ないことが分かった．

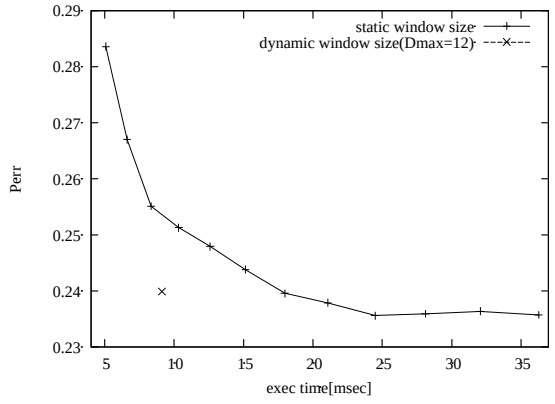

図 1　窓幅と予測誤差の関係．“static window size” の各点は，窓幅を 1,2,…,12 と変化させたときの平均予測誤差と時間を示す．また，“dynamic window size”は動的窓幅決定法（最大窓幅は 12）による平均予測誤差と時間である．横軸は予測にかかる時間を，縦軸の Perr は有効な予測区間内での平均誤差を表す．

また，予測に使用するパラメータの一つである窓幅の自動決定法を開発した．実行中プロセスの集合に着目し，予測開始時刻と同じ状態である過去の範囲を窓幅として設定する．予測時間と平均予測精度はトレードオフの関係にあるが，この手法により，高速な予測と高精度な予測がおおむね両立できることが評価実験によって確認された（図 1）．

(3) 負荷予測・タスク実行時間予測システムの実装と評価

(1)，(2)の成果を踏まえ，負荷・タスク実行時間予測システムを C++と STLport によって実装した．本システムは情報収集スレッドと予測スレッドから構成される．予測対象の計算機（Linux マシン）上で動作させることで予測に必要な負荷やタスク情報を蓄積し，ネットワーク経由での予測要求に応えて即座に予測結果を返答することができる．

実装にあたっては，使用メモリ量や応答時間，予測システム自体が発生する負荷などに考慮し，データ構造などの検討を行った．必要な一定期間の過去の負荷履歴は，PackBits コーディングによってメモリ上に全て格納され，予測の要求から 10ms 程度の遅延で予測結果を返答することができる．本予測システムでは，負荷のサンプリング間隔が 5 分，予測期間は 60 分であるから，これは十分に高速であると言える．

また，実装した長期負荷予測システムを用いて基礎的な負荷分散実験を行った．その結果，本システムによる長期予測の結果を利用した方が，現在の負荷をそのまま利用したタスク割当てよりも効率的であることが確認できた．なお，タスク割当てのアルゴリズムとしては，Yang ら（2003 年）の手法を長期予測結果が利用できるように改良して用いた．さらに詳細な実験や割当てアルゴリズム，スケジューリングアルゴリズムの改良は今後の課題である．

(4) 計算資源の分散管理手法

地理的に分散し，時間的に変化する非常に多くの資源を，効率的かつスケーラブルに管理するのは容易ではない．本研究では，計算

資源をP2Pネットワークによって自律的に分散管理する手法を検討した．

P2P ネットワークを構成するオーバーレイネットワークには，大きく分けて，非構造化オーバーレイネットワークと構造化オーバーレイネットワークがある．一般に，非構造化オーバーレイは柔軟な探索が可能だが検索漏れが生じやすいのに対し，構造化オーバーレイは，単一の条件に一致するものを確実に探索可能だが，柔軟な条件で探索することが困難だという性質がある．そこで本研究では，２種類のオーバーレイを相補的に利用することで，検索の柔軟性と探索効率を両立させる．

① 非構造化オーバーレイ

距離が近いホストの探索と利用を容易にするために，全ホストが参加し，ネットワーク上での距離が近いホスト同士でクラスタリングされた非構造化オーバーレイを構築する．この非構造化オーバーレイは，以下に示すように，クラスタ内でのホスト間遅延を指標としたクラスタの再構築を繰り返し行うことで自己組織化される．

あるホストから，同一クラスタ内の全てのホストに対する遅延の和を全遅延指標と定義する．クラスタを２分割したとき，全遅延指標のクラス間分散が最大なるように分割点を決定し，また，その値から実際に分割するかどうかを判断する．分割する場合の移動先クラスタも，同様に全遅延指標によって決定する．

② 構造化オーバーレイ

構造化オーバーレイを Kademlia によって構築する．このネットワークには各クラスタの代表ホストが参加し，全てのホストの属性データ(IP アドレスや計算能力など）を登録する．これによって，ユーザーが所望する性能のホストを効率的に探索可能とするとともに，非構造化オーバーレイ構築のヒントとしても利用する．

本システムでの非構造化オーバーレイではクラスタの自己組織化が行われるが，参加時に非常に遠いクラスタにたまたま所属してしまうと，最適化がなかなか進まないことがある．この問題は，前述の構造化オーバーレイに参加ホストのサブドメイン名やネットワークアドレスを登録しておき，同一ネットワークにあると推定される既参加ホストを参加時に探索することで，ある程度解決できる．

Overlay Weaver を用いて提案手法を実装し，シミュレーションによる評価実験を行ったところ，ホスト間距離が比較的小さなクラスタが生成されることが確認できた．また，参加ホスト数が増加しても生成されるクラスタの質が変化せず，本手法のクラスタ再構築手順がホスト数の増加に対して耐性があることが分かった．

(5) 協調型ジョブ管理手法

クラスタやグリッドなどのような並列分散環境では，複数のユーザーが同一の計算資源を共有する場合がある．資源を効率よく利用するためには，公平かつ効率的に資源を割り当てる必要がある．

このような資源割当てを実現するためにジョブ管理システムが広く利用されている．しかし，多くのジョブ管理システムでは，ユーザー間での公平性と，システム全体での効率の最大化を目的としている．通常，スケジューリングのポリシーは管理者によって決定され，個々のユーザーの要望に沿った柔軟な運用を行うことは難しい．

そこで本研究では，エージェントが他のユーザーのエージェントと交渉を行うことによって，ジョブの投入順の貸し借りを行う協調型のジョブ管理システムを提案した．本システムでは，基本的には到着順でのスケジューリング(FCFS)を行うが，エージェント間の調停によって実行順を交換することができる．ユーザーは自分のジョブの中で優先して実行したいジョブとそうでないジョブを柔軟に設定でき，これによって，非優先ジョブの実行順を他のユーザーに譲る代わりに，優先ジョブの実行順を早めることが可能になる．また，提案システムは管理者権限を必要とせず，一般ユーザーの権限だけで導入・動作させることができる．本システムによって，システム全体での効率にそれほど影響を与えずに，ユーザーの要望を反映したスケジューリングができることをシミュレーション実験によって確認した．

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 1 件）

① 丹野祐樹，菅谷至寛，阿曽弘具，プロセス情報を利用した実行時間予測と信頼度による予測選択手法，情報科学技術レターズ，Vol.6，pp.21-23，2007，査読有

〔学会発表〕（計 7 件）

① 中村貴，菅谷至寛，阿曽弘具，遊休資源相互共有システムアーキテクチャの検討，電子情報通信学会 2009 年総合大会，2009 年 3 月 18 日，松山市

② 丹野祐樹，菅谷至寛，阿曽弘具，相互情報量と信頼度による予測選択を用いた

実行時間予測手法，情報処理学会 ハイパフォーマンスコンピューティング研究会，2008 年 12 月 17 日，福岡市

③ 小林道治，菅谷至寛，阿曽弘具，プロセス情報を利用したリアルタイム計算機負荷予測手法，情報処理学会 ハイパフォーマンスコンピューティング研究会，2008 年 12 月 17 日，福岡市

④ Yoshihiro Sugaya, Hiroshi Tatsumi, Mitiharu Kobayashi, Hirotomo Aso, Long-Term CPU Load Prediction System for Scheduling of Distributed Processes and Its Implementation, The IEEE 22nd International Conference on Advanced Information Networking and Applications, 2008 年 3 月 28 日，宜野湾市

⑤ 氏家武志，菅谷至寛，阿曽弘具，動的負荷分散システムのための自律的オーバーレイネットワーク，ハイパフォーマンスコンピューティングと計算科学シンポジウム，2008 年 1 月 17 日，東京

⑥ 石山和也，菅谷至寛，阿曽弘具，分散並列環境における協調型ジョブ管理エージェント，ハイパフォーマンスコンピューティングと計算科学シンポジウム，2008 年 1 月 17 日，東京

⑦ 小林道治，菅谷至寛，阿曽弘具，プロセス検索法を用いたリアルタイム長期的負荷予測システムの構築，電気関係学会東北支部連合大会，2007 年 8 月 23 日，弘前市

〔その他〕
http://www.aso.ecei.tohoku.ac.jp/research/prediction/index.html

６．研究組織
(1)研究代表者
菅谷 至寛（SUGAYA YOSHIHIRO）
東北大学・大学院工学研究科・助教
研究者番号：80323062
(2)研究代表者
なし
(3)連携研究者
なし