超ロバスト並列計算

小柳義夫　須田礼仁　西田晃

情報理工学系研究科コンピュータ科学専攻

概 要

サブプロジェクト「超ロバスト並列計算」では, グリッドのようにネットワークや計算機の構成・性能が動的に変動する並列計算環境において, 資源の故障や追加，負荷の変動などの外乱に対しロバストに適応し，計算能力を効率よく引き出す手法を開発する. 数値アルゴリズムと並列化手法の両面から, 性能（計算機資源の利用効率）のロバストネスを導くことが研究の目的である. 本報告では, 平成 15 年度の成果の概要と今後の計画について略述する.

1　研究の背景と計画

21 世紀に入って, 計算機ネットワークが社会の隅々まで行き渡り, 成熟の度合いを高めてきている. ネットワークの性能向上は速く, 大容量のデータを短時間でやり取りすることも容易になってきた. これに伴い, 並列計算環境としてクラスタが一般的になり, プロセッサ数が数百台規模のクラスタもいくつも構築されるようになってきた. また, 複数の組織のネットワークをまたぐ計算機のインフラストラクチャとしてグリッドの概念が一般に認知されるようになり, グリッドシステムとともにアプリケーションの開発も進められてきている.

このようなネットワークで接続された計算機の計算能力と記憶容量はきわめて大きく, それらを並列計算という形で活かしたいというのは自然で現実的な要求である. しかし, 現在のネットワーク計算環境は, 従来のスーパーコンピュータ型の並列計算とはかなり異なる特質を持っている. ネットワークに関しては, 従来型の並列計算環境では専用の結合網を利用していたものが, インターネットなどの汎用のネットワークを利用するようになるため, 性能の保証はもちろん予測さえも容易ではなくなってしまう. また, 要素計算機に関しては, 従来型の並列計算環境では均一な仕様のものが一般的であったが, 特にグリッドなどではすべての計算機が同一仕様ということはほとんど期待できない. また, 多くのネットワーク組織をまたがった計算環境では, 計算機が追加されたりダウンしたり, ネットワーク構成が変更されたりするといったこともかなりの頻度で起こるものと考えられる. これに対して, 従来の並列計算環境を想定した並列化手法では, このような新しいネットワーク並列計算環境においては, 正常に動作することを保証することはできないし, ましてや十分な性能を引き出すことは不可能である.

我々のサブプロジェクト「超ロバスト並列計算」では, 上述のような新しいネットワーク計算環境のもつ処理能力をロバストにかつ最大限に引き出すことができる新しい並列化手法を開発する. ネットワークや計算資源の増減や負荷の変動といった外乱要因に対してロバストに適応する手法を数値アルゴリズムと並列化手法の両面から開発する計画である.

本プロジェクトでは，平成 14 年度に実験環境の整備と既存研究の調査を中心に活動し，本 COE の融合プロジェクトのひとつである大規模ディペンダブルプロジェクトの田浦研究室との研究交流などを行った. 平成 15 年度には，本報告で述べるように，並列化手法として集団通信とデータ再分散について研究し，簡易かつロバストに並列性能を引き出す手法として ERXPP を提案した. こ

れまでの研究により，動的環境の一つの要因であるヘテロな環境に対する効率的な並列化技術の見通しが立ったと考えている．今後はヘテロ並列化手法の実装と評価を進めるとともに，故障を含む予期せぬ性能の変化に対応する手法を開発してゆく予定である．

2　本年度の成果

本節では, 平成 15 年度の研究成果と，それに関する今後の研究計画などについて報告する．

まず，並列処理に欠かせない集団通信について，グリッドをも想定した一般的なネットワークトポロジーにおける最適化について報告する．

次に，高度に適応的な数値ライブラリを用いることにより簡単に並列性能を利用する手法 ERXPP を提案し，その予備評価とデータ分散の最適化について報告する．

2.1　ロバストな集団通信

様々なグリッド等の汎用の並列計算環境で性能を最大限引き出すためには、効率的な通信の実装が必要である．本研究では，様々な環境に適応した集団通信のスケジューリングを得る手法の研究を行った．

2.1.1　ネットワークと通信のモデル化

集団通信のスケジューリングにおいては，まずネットワークと通信をモデル化する必要がある．従来の集団通信においてはネットワークのモデルを固定したり，通信性能を無視したりすることが多かった．それは，複雑なネットワークモデルでの最適化という問題は、解くことが非常に困難な NP 困難というクラスに属することが多いためである．これに対し本研究では，性能を出すためにはより詳細なモデル化が必要であると考え，ハブや通信遅延，バンド幅も含めたモデル化を行った．インターネットなどのネットワークのトポロジーに関しては，通常，階層的に構築されていくため，木構造または，それに非常に近い構造をとることが多い．また，ルーティングプロトコルにも全域木を用いることがあるほど，ネットワークがある程度密でなければ，故障時以外は迂回路を利用するメリットも多くはない．このため，木構造のネットワークにおける通信のスケジューリングを考えた．これには，動的に変化し，制御の難しいルーティングを考慮せずにすむメリットがある．通信のモデルにおいては，１つのノードが複数のノードに同時に通信をすることを許容している．これは，非同期通信を有効に活用するために不可欠で，最大マッチングを用いるような手法では，カバーできない問題となる．

2.1.2　最適スケジューリング

本研究課題として、集合通信の中で最も代表的な Broadcast のスケジューリングを取り上げた．Broadcastは，ある１つのノードが持つデータを他の全てのノードに知らせる問題である．まず，上記のネットワークモデルにおいてこの Broadcast の最小通信時間のスケジューリングを探索する問題は、3-PARTITION という NP 完全な問題から多項式還元可能であることを示し，非常に難しい NP 困難な問題のクラスに属することを示した．しかし，現在の通信ライブラリの性能は極めて不十分であり，規模を制限してでも最適解を得ることは必要であると考えられる．そこで，実用的な範囲の規模の並列計算機における Broadcast のスケジューリングの最適解を求めるために，木の同型判定と下限計算による効率的な枝刈りを行う分枝限定法を用いたアルゴリズムを提案、実装した．さらに，実際の並列計算機の構成を変えて，提案アルゴリズムで最適スケジューリングを求めた．さらにえられたスケジューリングに従う Broadcast の通信時間を実機で計測し，ほぼスケジューリング通りの性能がでることを確認した．また，既存のライブラリの Broadcast の通信時間と比較し，優位になることを示した．提案アルゴリズムは，Broadcast と通信パターンが逆である Reduce 演算にも適用可能である他，Gather とそれと通信

パターンが逆の Scatter にも類似の手法が適用可能であると思われるため，今後実装し，検証していく．

2.1.3 今後の課題

本研究で，最適なスケジューリングを得ることは非常に難しいということがわかったが，1つの通信が集団通信全体に及ぼす影響に関するある程度の知見を得られた．これら得られた知見をもとに，大規模な並列計算においても，最適なスケジューリングの通信時間に近い近似解の計算が可能な高速なアルゴリズムの開発を目指していく．

研究を進めて行く上で，通信相手がごく限られるような並列プログラムが非常に多くあることがわかった．このようなプログラムの個々のプロセスをネットワーク上のどのマシンに割り当てるかという問題は，スケジューリングの問題同様，プログラムの修正を最小限に留めつつ，グリッド環境上で性能を出すためには，重要である．今後，両者を合わせて汎用性のある並列計算環境の構築を目指す．

2.2 ERXPP — 数値ライブラリを利用したロバスト 並列計算

2.2.1 ERXPP の提案

本プロジェクトが目標としている，不確定で動的な要因を含む並列環境に適応した並列プログラミングには，これまでの並列化技術の単なる延長ではない手法が必要である．しかし新たな手法を用いて並列化をやりなおすのは無視できない労力を必要とする作業となってしまう恐れがある．

そこで簡単にロバスト並列処理の恩恵を受けることができる枠組みとして，我々は数値ライブラリに注目をした．数値ライブラリは科学技術計算のさまざまなアプリケーションにおいて計算時間の大部分を消費する．そのような場合には，さまざまな並列環境に高度に適応した数値ライブラリを使用することにより，アプリケーションが容易に並列性能を活用することができると期待される．我々はこのアイデアを ERXPP (Easy and Robust Extension of Parallel Performance) と呼んで，ロバスト並列処理のひとつのありかたとして提案した．

ERXPP により有効に並列計算資源を利用できると期待される状況には以下のようなものが考えられる．

- 呼び出し側（アプリケーション）は逐次か，SMP 上で並列化されているが，ライブラリはネットワーク上の資源も利用する．
- 呼び出し側はホモなクラスタ上で従来手法により並列化されているが，ライブラリはヘテロを含むより大きな環境を利用する．
- 呼び出し側は静的なデータ分散を用いるが，ライブラリは動的負荷分散を行って性能向上を図る．
- 呼び出し側は小規模な安定したシステムで動作しているが，ライブラリはより信頼性の低い計算資源をも利用する．

これまで数値計算ライブラリは高い性能を提供するべく開発されてきたものであるが，ERXPP はそれに加えて高い適応性と耐故障性をも提供することを提案するものである．

2.2.2 惑星大気の流体シミュレーションによる予備評価

我々はまず提案する ERXPP がどの程度現実的か，簡単なアプリケーションを用いて予備評価を行った．アプリケーションとしては惑星大気の流れのシミュレーションを行うプログラムを用い，その中で用いられる球面調和関数変換について ERXPP の理念に従う改良を施した．但し，現状では耐故障性についてはシステム用件が十分でないため実装には至らなかった．また，クラスタとしてはホモな構成のものを使用した．

アプリケーション（呼び出し側）が逐次の場合，また使えるプロセッサ数よりも少ない台数で並列化してある場合について，ライブラリはデータ再

分散を行って使用できる最大限のプロセッサを利用するようにプログラムした．また，再分散と計算結果の収集の通信コストも考慮してデータ分散とスケジューリングを最適化した．

その結果，ギガビットイーサ程度の通信性能があればデータの再分散のコストも含めて高速化が達成されることを確認した．しかし 100 Mbps のネットワークでは通信性能が不足して高速化が達成されず，ERXPP のためには高性能ネットワークの存在が本質的であることが明らかとなった．また，プロセッサの 1 台に計算負荷をかけてみたところ，計算所要時間は短縮されたものの，全体としては所要時間が大幅に増加してしまった．これは他のプロセスの存在により OS レベルのスケジューリングが影響を受けたことによるもので，集団通信がブロックされて所要時間に大きな影響が出たことがわかった．

今後の課題として，故障を含む予期せぬ性能の変化に対応できる手法の開発が急務であることが明らかとなった．

2.2.3 Multi-master divisible load の提案と漸近最適スケジューリング

上述の予備実験において，ヘテロプロセッサに対するデータの再分散が必要となる．ここで，データの再分散と計算結果の収集にかかる通信と，計算の所要時間をあわせてスケジューリングしなければ，かえって性能を低下させてしまう場合もあることが判明した．そこで我々はデータ再分散の最適スケジューリングについて研究を行った．

計算のモデルとして最も単純と思われる divisible load theory (DLT) のモデルを採用した．しかし DLT はもともとマスタースレーブ型の並列処理しか想定していないので，我々はこれをデータ再分散問題に拡張し，multi-master divisible load (MMDL) としてモデルを提案した．

さらに，MMDL に対するデータ再分散の最適アルゴリズムについて研究を行った．その結果，プロセッサ性能がヘテロであることは何の困難も引き起こさないこと，その一方でネットワーク性能が非一様である場合には問題が複雑になることを見出した．そしてネットワーク性能が一様なクラスタ（プロセッサ数を p とする）について，所要時間の下限（すなわち処理性能の上限）を $O(p \log p)$ の計算量で求めるアルゴリズム，およびこの下限に（タスク量が多くなった場合に）漸近するスケジューリングアルゴリズムを提案した．

提案手法を実装してシミュレーションで評価したところ，従来の通信を考慮せずに負荷のみを均衡化する場合や動的負荷分散に比較して，高い性能が安定して得られることが確認された．

3 まとめ

本報告では超ロバスト並列処理プロジェクトの研究開発状況について報告を行った．これまでの研究で，ヘテロ環境における計算性能を引き出す手法については目処が立ったと考えている．また，MMDL の高速なスケジューリングと効率的なデータ再分散の研究により，計算資源の増減に対応する手法も開発が進んでいる．一方で，予期せぬ性能低下や突然の故障に対応するための手法については，現在アイデアはあるものの，具体化は進んでいない．来年度以降の研究においては，特に故障をはじめとする予期せぬ性能変化に対応する手法の具体化を急ぐ予定である．

以下は，本報告に直接関係する主な外部発表である．

- 蓬来祐一郎, 西田晃, 小柳義夫, 「木構造型ネットワークにおける最適ブロードキャストスケジューリング」, 情報処理学会 ACS 論文誌 (掲載予定).

- 須田礼仁,「ERXPP — 数値ライブラリにより並列計算性能を簡易かつ適応的に引き出す方式の提案」, 情報処理学会研究報告 2003-HPC-96, Oct. 2003, pp. 19–24.

- 須田礼仁, 「Multi-master divisible load model における漸近最適スケジューリング」, 情報処理学会研究報告 2004-ARC-149/HPC-97, Mar. 2004, (掲載予定).