1.1. 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, ネットワークや計算機の故障や追加・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 14 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.

研究計画のおおよそのタイムスパンは, 平成 14 年度:実験環境の構築, プロトタイプモデルの開発, 平成 15 年度:プロトタイプシステムの実装と評価, 平成 16 年度:実アプリケーションへの適用手法の開発, 平成 17 年度:実アプリケーションへの適用と評価, 平成 18 年度:総合的な評価と知見のとりまとめ, とした. すなわち, 最初の 2 年間で基礎的な知見を固め, ネットワーク・計算機の不確定・動

的要因による実効性能の低下を緩和する手法の提案と実装を行い, 残りの 3 年間でアプリケーションに適用して並列計算環境における性能（計算資源の利用効率）のロバストネスを実現する並列化手法を実証するという方針である.

2 本年度の成果

本節では, 平成 14 年度の研究成果を中心に, これまでの準備状況やより具体的な研究計画などについて報告する.

研究的な内容としては, (1) 並列数値計算プラットフォームとしてのグリッドの可能性, (2) 耐故障性を実現するためのチェックポイントシステムの検討, (3) 研究拠点形成アシスタントによる研究の進捗状況, の 3 点について報告する. またこれに関連して, (4) 本年度予算で導入した PC クラスタ, (5) 大域ディペンダブル融合プロジェクトの田浦研究室との交流についても報告をおこなう.

2.1 広域計算環境における並列数値処理

近年, 広域に分散された世界中の情報資源を統一的に扱うグリッドと呼ばれる技術が国内外において盛んに研究され, 開発が進められている. グリッド技術は計算機を使用するための新しい概念であり, 電力網を示す “Grid” という言葉に表わされるように, 増大する計算機資源に, 統一的に, また確実かつ安価にアクセスできるようにすることを目標としている. 広域計算は, これらの世界中に分散した異機種計算機環境を仮想的な一つの高性能計算機と見立て, 分散並列計算を行うための技術である.

広域計算に関しては, 現在様々なプロジェクトが進められているが, このうち最も有名なものとして, Globus プロジェクトを挙げることができる. Globus プロジェクトでは, 広域計算を実現するために必要とされる Globus Toolkit と呼ばれるソフトウェアツールを開発しており, グリッド上のサービスを提供するための標準的なミドルウェアとなりつつある. また, 疎粒度な分散並列計算を目的とした Ninf や Netsolve など, 多くのプロジェクトで研究が進められている.

本研究では, 広域計算環境においてネットワーク帯域幅が計算性能に与える影響を調べるため, 仮想的なグリッド環境として Fast Ethernet, Gigabit Ethernet の 2 種類のネットワークで接続された PC クラスタを構築し, 基本的なネットワーク性能を評価した. さらに, メッセージ通信インタフェースの実質的な標準規格である MPI を使用し, Argonne 国立研究所によって開発された MPI 実装の一つである MPICH と, その Globus 環境への拡張である MPICH-G2 を用いて, 直接法による連立一次方程式の求解ライブラリである Linpack の並列計算性能に関する評価を行い, グリッド上での並列科学技術計算の可能性について検討した.

性能性能ツール mpptest により通信性能をテストした結果, Fast Ethernet 上での性能と比較して, Gigabit Ethernet 上では, メッセージサイズが小さい場合に MPICH に対する MPICH-G2 の性能低下がより顕著であった. また Linpack ベンチマークを用いて MPICH と MPICH-G2 の性能を比較したところ, アルゴリズムのパラメータを最適に選ぶことにより, Gigabit Ethernet 上では MPICH と MPICH-G2 の性能がほぼ同等となることが分かった. これは, MPICH-G2 上で最適なアルゴリズムにおいて, MPICH の場合より通信回数が少なく, ネットワーク遅延の影響が小さいことによるものであるが, このことは, 十分な通信帯域幅のある環境では, 通信回数を抑えたアルゴリズムを採用することによって ネットワーク遅延の影響を効果的に軽減できることを示している.

以上の実験は LAN 上の理想的な環境で行ったものであり, WAN 上に分散した環境において, 現時点ではより大きな通信遅延を想定しなければならない. しかしながら, この結果は広域環境上における実用的な科学技術計算の可能性を示唆するものであり, 今後はより高性能な広域計算環境を実現する上で必要となるソフトウェア技術について, クラスタを効果的に活用した研究及び性能評価を行っていきたいと考えている.

2.2 チェックポイントシステム

科学技術計算の並列計算では, 領域分割などでデータ分散を行い, owner-computes rule を基礎として並列処理を行うことが多い. このような手法は通信量が少なく抑えられる傾向にあり, プログラミングもそれほど難しくないなどの利点があるが, 外乱要因に対するロバストネスを達成するためにはいくつかの工夫が必要となってくる.

我々は動的で故障の可能性があるハードウェアを仮定し, その上で最大限に性能を得ることを目標として設定した. そして, 性能のロバストネスを実現する基本方針として (1) 遅延隠蔽, (2) 動的負荷分散, (3) 耐故障性の 3 つを提案した.

特にプロセッサやネットワークの予期しない停止に対してロバストに対応することを考えると, プロセッサに分散格納されているデータを再現することが必要である. プロセッサやネットワークが停止した場合に, プロセッサが格納していたデータを再現するためには, プロセッサのメモリの内容を保存しておくチェックポインティングが一般には必要となる. 我々はチェックポインティングのオーバーヘッドを必要最小限に抑えられるようにユーザーレベルでのチェックポイント制御を行うこととし, メッセージログを併用してリカバリの負荷を軽減することを試みることを提案した. また, 手法の有効性を実証するためのテストベッドシステムの構成について検討を行った.

今後は, テストベッドシステムの構築とチェックポインティングを用いたロバスト並列化手法の実装・評価を進める予定である. アプリケーションとしては, 比較的シンプルな並列姓を持つ科学技術計算から始めて, 数値ライブラリのロバスト化などの応用までを計画している.

2.3 研究拠点形成アシスタント報告

本節は今年度研究拠点形成アシスタントとして雇用された蓬来祐一郎による研究成果報告である.

本年度は, 生命情報科学における幅広い有用性をもつ実アプリケーションとして, タンパク質の立体構造予測法の一つ分子動力学法, ゲノム配列データなどで必要とされる高速な文字列検索を可能とするデータ構造である Suffix Tree, Suffix Array 等を実装し, クラスタ, グリッド環境での並列処理における問題点, 並列化可能性等を検討した.

今後は, 耐故障性, 効率的な負荷分散を達成するためのシステム開発を目指す. その際, 長時間動作し大量にメモリを消費するようなアプリケーションでは, チェックポイント等の際にすべてのメモリイメージをダンプするのでは, 効率が悪い上に, 計算を中断させることなく継続させることは難しい. そのためプログラムの耐故障性を考える際にアルゴリズムの構造を考慮することが有効になることが考えられる.

そこでアルゴリズムの構造が簡明で汎用的な, Branch and Bound アルゴリズムを題材としてこの問題に取り組んでいく. Branch and Bound アルゴリズムは, NP 困難な最適化問題において厳密解を求めるアルゴリズムとして幅広い適用範囲をもち, かつ並列性も高いため, 研究に適した対象である. 具体的には様々な計算機の混在するクラスタ, グリッド上でリアルタイムに変化する解の下界, 上界等のデータを効率的に管理, 共有しつつ, かつ計算機環境の変化への対応に優れた並列分散アルゴリズムの開発及び, それに必要なロバストな並列計算環境の構築をめざす. これにより, 既存の計算機資源を有効に活用することで, 今までは困難であったサイズの問題でも実用的な時間で解ける可能性が増すことが期待でき, また Branch and Bound アルゴリズム以外の様々な並列性をもつ, 並列分散アルゴリズムに対しても適用できるような汎用性が得られることが期待される.

2.4 クラスタ計算機の導入

本プロジェクトではグリッド計算環境を想定した実装実験のために PC クラスタを導入した. 購入したのは IBM 社製クラスタ (IA サーバ xSeries335 [dual Xeon 2.4GHz, PC2100 DDR-SDRAM 512MB, Gigabit Ethernet 2 ポート, Ultra160 SCSI HDD 36.4GB] × 8 ノード) である.

OS には Redhat Linux7.3 をインストールし, その上で SCore-5.4.0, LAM/MPI-6.5.9 などの並列ソフトウェア環境を構築し現在稼働している.

Globus をはじめ他のシステムソフトウェアの導入も必要に応じて検討し, 本クラスタ及び, その他既存の計算機システムを用いて, ソフトウェアの開発, 性能評価を行なっていく計画である.

2.5 田浦研究室との交流

「超ロバスト並列処理」サブプロジェクトは,「超ロバスト計算原理」融合プロジェクトのサブプロジェクトの一つとして位置づけられているが, ネットワーク計算環境を対象とする「大域ディペンダブル情報基盤」融合プロジェクトとも内容的な関連がある.

そこで,「大域ディペンダブル情報基盤」融合プロジェクトに参加している田浦助教授の研究室と小柳研究室・須田研究室で合同のセミナーを行い, お互いの問題意識やアプローチについて発表しあい, 意見の交換を行うことを試みた.

セミナーは平成15年2月18日15時より18時過ぎまで行われた. まず田浦助教授と田浦研の学生により Resource-Reconfigurable Fault-Tolerant Parallel Computing の概念と開発中の Phoenix ライブラリを用いたプログラム例について発表があった. その後, 須田により超ロバスト並列処理のためのチェックポインティングシステムについて発表が行われた. 両者は注目しているアプリケーションの並列性の違いやアプローチの違いがあるが, 問題意識や目標は共通部分が多く, 双方の研究を比較し議論することにより相互に洞察を深めることができ, きわめて有意義な合同セミナーであった.

3 まとめと外部発表

「超ロバスト並列処理」サブプロジェクトでは, ネットワーク計算環境における性能のロバストネスを目指している. 現在は主にグリッドと耐故障性をキーワードとして研究を進めているが, 今後は動的負荷分散やアプリケーションの開発にも力を入れて研究を推進してゆく予定である.

本プロジェクトは開始から半年に満たないため, 純粋に本プロジェクトとしての研究発表には及んでいない. そこで最後に, 研究の準備状況と周辺分野の研究状況として, 本年度の我々の研究グループの研究発表の主なものを報告する.

- Y. Oyanagi, Future of Supercomputing, Journal of Computational and Applied Mathematics, Vol. 149, No. 1, pp. 147–153, 2002.
- 藤井昭宏, 西田晃, 小柳義夫. 領域分割による AMG アルゴリズム, HPCS2003 論文集, pp. 83–90.
- 額田彰, 西田晃, 小柳義夫. 分散共有メモリを用いた並列 FFT とその最適化, HPCS2003 論文集, pp. 63–70.
- 西田晃, 額田彰, 小柳義夫, 分散共有メモリを用いた疎行列アルゴリズムの細粒度並列処理, コンピュータシステム・シンポジウム 2002 論文集, pp. 13-20.
- 武田恵史, 西田晃, 小柳義夫, Globus を用いた Grid 上での並列数値処理とその性能評価, インターネットコンファレンス 2002 論文集, pp. 5–12.
- R. Suda, "Fast spherical harmonics transform of FLTSS and its evaluation", The 2002 Workshop on the Solution of PDEs on the Sphere.
- 西田晃, 小柳義夫, OpenMP を用いた Jacobi-Davidson 法の並列実装とその性能評価, JSPP2002 論文集, pp. 79–86.
- R. Suda, M. Takami, "A Fast Spherical Harmonics Transform Algorithm", Math. Comp., Vol. 71, No. 238, 2002, pp. 703–715.
- S. Itoh, Y. Oyanagi, S.-L. Zhang and M. Natori. Effect on Spectral Properties by the Splitting Correction Preconditioning for Linear Systems that Arise from Periodic Boundary Problems, In Proceedings of Enabling Society with Information Technology, pp. 234–243, Springer-Verlag, 2002.