論文の内容の要旨

論文題目　Research on Network-aware Parallel Computing
（計算機のネットワーク構造を考慮した並列計算の研究）

氏　名　蓬来 祐一郎

近年、並列計算環境はクラスタ・グリッドをはじめとしてより身近になりつつあるが、それを構成する計算機の計算性能・通信性能・ネットワークトポロジーは様々である。このため並列計算環境でよりよい性能を出すには、これら並列計算機の不規則性、不均一性の問題に向き合わなければならない。また、並列化を行った際には、計算の分割が可能な単位においても、要素ごとに必要な計算量や通信量が大きく異なってくるため、計算の割り当て負荷分散には、これらを十分考慮しなくてはならない。

この論文では、このような科学技術計算の持つ共通の課題について、生体分子のシミュレーションを分散メモリ型並列計算機上で行う分子動力学法を対象に研究を行った。周期境界条件を用いない分子動力学法や、真空中もしくは、溶媒の連続体近似を行うシミュレーションの場合には、領域により計算量の大きな偏りがあり、粒子の分布も一定ではない。従来の分子動力学法では、このような不規則性を持った問題の負荷分散はほとんど考慮されて来なかった。この研究では、そのような不規則性を持った問題を扱う。

分子動力学法は反復計算を行い、各ステップ内では粒子への力の計算を行った後、位置もしくは力の情報を通信を行うが、各々の力の計算の順序に依存関係はない。このときの並列化手法としては、全プロセスが全粒子の位置情報の複製を持つ手法と、分割して持つ手法に大別される。２つの手法にはそれぞれ長所と短所があり、どちらが優れていると一概にいうことはできない。そこでこれら２つの並列化手法について、通信時間の削減のために、それぞれ異なるアプローチを提案する。一つは全プロセスが複製を持つ手法で用いられる集団通信の最適化、もうひとつは、小セルによる空間分割法を用いた場合のデータ割り当てによる隣接通信の最適化である。

まず、第１部では、集団通信の最適化について扱う。前述の不規則性へ対処するために従来手法の多くは、プロセス全てに全原子のデータの複製を持たせている。この方法では、粒子とプロセッサ数の増加とともに通信時間が大きくなるが、系が小さい場合や、データの局所性が少ない場合、空間分割法よりも計算負荷を細かく制御でき、通信の効率も近くなるため、全体の効率が良くなる。まずここでは、集団通信のひとつであるBroadcast通信をpoint-to-pointの通信を用いてネットワークに適したスケジューリングを得る手法の研究を行った。集団通信の通信時間はそのスケジューリングに大きく左右され、分子動力学法のように大きなメッセージを通信する場合には、特にネットワーク構造が重要である。一般的にネットワークは多様な性質の回線から構成されており、ヘテロなネットワーク環境で効率的なスケジュールを得る事は難しい。しかし、そのような非均一性が増えつつある今、適切なスケジューリングを得る事が非常に重要になってくる。また、現在の汎用のMPIの実装においてBroadcast内の個々の通信はブロッキング通信で行われているため、複数のノードへの同時送信や受信と送信をオーバーラップすることができていないため、ネットワーク構造を考慮する事によっ

て大きなデータの通信をより高速に行える機会を逸している。しかし、これを効率的に行うには、通信の競合を避ける事が重要である。この研究では、木構造型のネットワークにおけるBroadcastのスケジューリングを扱う。木構造型ネットワークでは、多くのノード間で共通の通信路を使うため、競合する可能性が大きい。このためスケジューリングによる通信時間削減の効果が大きい。ここでは、木構造の対称性から由来する冗長なスケジューリングの探索を削減する手法を提案し、分枝限定法による探索を行う。この手法では、通信ペアの削減により探索を高速化するため、通信のモデルに依らず枝刈りが可能となる。実験によりこのアルゴリズムの効率性を示し、また得られたスケジューリングによるBroadcastを組み込みのBroadcastと比較し、データサイズが大きい場合、大幅に通信時間が減る事を示す。また、このネットワーク構造に最適化されたBroadcastは、全プロセスが粒子の複製情報を持つ、粒子分割法や力計算の分割法による分子動力学法に用いられる集団通信Allgatherや、Allreduceの高速化に貢献することを実験により示し、ネットワークのヘテロ性を考慮する事の重要性を示す。

第２部では、分子動力学法などにおいて空間分割を行う際の新しい分割手法の研究を行った。空間分割法は系の空間が十分大きい場合、データの局所性が高くなり通信量が少ないため高い並列性能を示す。しかし、従来の空間分割法では、粒子の分布する空間が直方体領域をとるなど仮定が大きく、領域の割当もカットオフ半径の大きさの立方体領域単位で行うため粗く、負荷の均衡が難しい。そのため、全プロセスがデータの複製を持つ手法の方が好まれていたようである。一方で、グラフ分割手法は、科学技術計算においてデータの局所性を活かし、並列性を出す手法として貢献してきた。この手法ではデータ依存関係をグラフ表現し、頂点集合を同じ重みのパーティションに分割し、枝カット、つまりは通信コストを最小化する。グラフにより任意のデータ依存関係が表現でき、カットオフ半径よりも細かい領域の分割も容易に扱う事が可能となる。しかし、グラフ分割により通信量の削減は期待できるが、分割データの計算ノードへの割り当てを適切に行わなければ、通信に局所的な偏りが起きるなど期待した並列性能が得られない可能性がある。そこで本研究では、計算とそれに必要な通信をグラフとして表現した計算グラフを定義し、これをもとにプロセッサの処理能力に応じてデータを分割し、ネットワーク構造を考慮した割り当てを行う手法を開発した。提案手法では、グラフデータを並列計算機のネットワーク構造を考慮しながら、再帰的にマルチレベルな二分割を行うことにより、プロセッサへ割当てる。提案手法は、従来手法の以下の点を改善する。まず、従来の研究で用いられる枝カットは、通信量を正しく反映したものではない。そこで、並列計算のモデルとして計算グラフを定義した。次に、枝カットの最小化によって全体の通信量が削減されたとしても、個々のプロセッサの通信量が減らなければ、通信時間は削減されない。この研究では、このような個々の通信時間を反映するような目的関数を提案した。そして、従来は、ネットワーク構造は考慮されないか、考慮されても非常に単純なものになっている。そこで、MDのような通信の多いアプリケーションではバンド幅が通信時間を最も左右するパラメータと考え、ネットワークモデルを構築した。この３点の改良を元に、再帰的な二分割手法を拡張し実装した。この手法を分子動力学法のデータ分割に適用した実験において、負荷の均衡を保ちつつ、通信時間が大きく短縮される事を示した。特に、ヘテロな環境では、直接通信を行った場合で10-20%、中継を用いた通信で20-60%通信時間が削減された。

この研究により、特に、分子動力学法のような通信量の多いアプリケーションでは、ヘテロなネットワークで結ばれたクラスタ環境では、通信時間が大きく、並列効果を上げることが難しかったが、２つのアプローチによる提案手法によりそのような環境での通信時間を大きく抑える事が可能となった。