超ロバスト集団通信

蓬来祐一郎

情報理工学系研究科コンピュータ科学専攻

概 要

近年、安価に構築可能な並列計算機環境として、既存の PC や並列計算機ををネットワークで結んだヘテロクラスタやグリッドと呼ばれる計算環境の研究が盛んに行われている．このような計算性能やネットワーク性能が非一様な環境で計算性能を引き出すには，プログラムをこれら新しい環境に適した手法に変更するなどの必要が出てくるが，繁雑なうえ，また新たな計算機環境に変わると性能がでなくなってしまう可能性もある．そこで我々は，メッセージ型通信ライブラリで用いられる集団通信を環境ごとに適したものにすることで，このような問題を解消し，安定した性能を確保する手法の開発を目指す．これにより，既存のプログラムの改変も必要なく，また既存のネットワークにおいても性能の向上が見込まれる．

1　はじめに

近年，PC の高性能化やネットワーク性能の向上に伴って，比較的安価に構築可能なヘテロクラスタやグリッドと呼ばれるような，不均一な並列計算機環境の研究が盛んに行われている．このようなネットワークを介した並列計算機環境においては，一般にメッセージ通信型並列モデルが用いられており，その代表的な仕様が MPI と呼ばれるライブラリとして広く使われている．MPI などを用いたメッセージ通信型の並列プログラムにおいて，プログラムを簡便にするため，複数のノードが参加する頻繁に用いられる通信パターンを集団通信ライブラリとして提供している．

MPI を使用可能な並列環境も一般に普及しつつあり，MPI で実装された並列プログラムの公開も盛んに行われている．しかし，これらのプログラムがグリッドなどの新しい計算機環境で期待通りの性能で動くかというと，必ずしもそうはならない．これは，計算や通信の負荷の不均衡や，通信時間の増加によるところが大きい．特に，計算は通常，台数が増えるほど高速になるため、メッセージ通信型の並列計算において通信時間を減らすことは欠かせない．また，一般に集合通信は，計算ノードが増えるほど負荷も増え，ネットワークデバイスの高バンド幅化による遅延の減少も緩やであることが予想されるため、ノードの増加や計算機の性能向上につれて、より顕著な問題になっていくことが予想される．

このような問題点を考慮すると，既存の並列プログラムを活かすには，通信ライブラリをより高速なものに置き換える方法が効率的であると思われる．専用ハードウェアもしくは専用プロトコルを用いて１対１の通信性能を最大限に上げる研究及び開発は盛んに行われているが，複数の計算ノードが同時に通信に参加する集団通信をネットワーク構造に合わせてスケジューリングを行う研究はまだ少なく，これらを用いた通信の大幅な性能の向上が見込まれる．そこで本研究では，並列計算においてユーザが使用する計算機やネットワークを意識することなく、計算性能を引き出すことができる並列計算環境の開発を目指し，集団通信のネットワーク構造への最適化を行うための研究を行った．これにより既存の並列プログラムを変更をせず、あるいは少ない変更で、計算機環境を変更しても性能を活かした並列計算が行え，また，既存のコンピュータ資源をより有効に活用することが可能になることが期待できる．また，このよ

うにして得られた集団通信の最適スケジューリングは、結果的に上位層の回線を用いる通信を削減するため、他の通信の影響を受けにくいことが期待されプログラムの安定性にもつながることが予想される．

2　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，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ネットワークのモデルは、固定したり，通信性能を無視することが多かった．それは，通常このような複雑なモデルで最適化を行う問題は、解くことが非常に困難な問題クラスである NP 困難に属することが多いためであるが，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．

3　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．そこで、このような難しい問題でも、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考案し、実装した．そして，実際の並列計算機を用い，構成を変えてモデル化し，提案アルゴリズムを用いてそれぞれのモデル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．

4　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．また，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，この問題と合わせ，汎用性のある並列計算環境の構築を目指す．