論文審査の結果の要旨

氏名　松本　尚

単体プロセッサの性能を越えた情報処理を可能にするために、並列処理システムもしくは分散処理システムが多く構築されるようになってきている。これらのシステムは通信同期サブシステムによって複数のコンピュータを相互接続した形態を採っている。効率の良い並列分散処理システムのの構築には、効率の良い通信同期方式の確立が不可欠である。

本研究では、操作対象となるデータの論理アドレスを通信同期サブシステムが認識するというメモリベース通信同期方式（Memory-Based Communications and Synchronization scheme: MBCS scheme）というアプローチを採用している。これは共有メモリの考え方を通信同期サブシステムの効率化に使用するアプローチである。従来の通信同期サブシステムはメッセージパッシング方式に基づいており、送信手続きとそれに対応する受信手続きが必要であり、通信同期サブシステムからユーザタスクがデータを受信する時に付加的なデータコピーが必要になる。これに対して、MBCS 方式に基づく通信同期サブシステムは通信要求側のみで通信が完了し、ユーザタスク上のデータを直接アクセスするため、データコピー回数も削減できる。また、直接ユーザタスク上のデータをアクセスするため、読み書き操作以外に不可分操作やキュー操作といった高度な同期機能を実現可能である。また、プロセッサのメモリ管理機構と同様の機構を使用することにより、マルチタスクに対応可能な保護と汎用性を保証することも可能である。そして、これらの利点は通信同期サブシステムがハードウェア実装されているかソフトウェア実装されているかに依存しない。

本論文では、MBCS 方式に基づいた三種類の通信同期機構が提案され、実証実験と考察によりその有効性が示されている。より具体的には、MBCS 方式に基づいたサブシステムを、取り扱うデータの粒度と実装方式の違いによって大きく三つに分類し、それぞれに対して通信同期機構が提案されている。最初の機構は Memory-Based Processor (**MBP**) と呼ばれ、細粒度のデータ通信同期を行うハードウェア実装された機構である。また、MBP はハードウェア分散共有メモリ実現のための機構としての側面を持つ。次の機構は特殊なハードウェアを必要としない中粒度以上のデータ通信同期を行うソフトウェア実装された Memory-Based Communication Facility (**MBCF**) と呼ばれる機構である。三番目の機構は Memory-Based Processor II (**MBP2**) と呼ばれ、中粒度のデータ通信同期を行うハードウェア実装された機構である。

本論文は 9 章よりなっている。第 1 章は序論である。まず、メモリを介して通信同期を行う従来技術であるハードウェア分散共有メモリおよびソフトウェア分散共有メモリの過去の発展を概観する。そして、本論文の主題である、MBCS 方式の導入が行われている。

第 2 章はハードウェア実装の細粒度通信同期機構である MBP について述べられている。MBP はメインプロセッサに対するコプロセッサであり、メインメモリ内に実装され、通信同期のような局所性を活かせない処理を担当する。コプロセッサである MBP の通信粒度はメインプロセッサの外部メモリトランザクションの粒度であり、メインプロセッサの内蔵キャッシュにおけるブロックサイズと同じである。MBP が構成するキャッシュシステムはハードウェア分散共有メモリシステムであるにもかかわらず、ソフトウェア分散共有メモリと同様にメインメモリをキャッシュ領域として使用可能である。このキャッシュシステムを実現するために、MBP は世界に先駆けて二段階のアドレス変換方式を採用している。ハードウェア分散共有メモリシステムのスケーラビリティを向上するのに重要となる「階層マルチキャストと Ack コンバイニング」も MBP の機能の一部として開発されている。これはマルチキャストメッセージに対する Acknowledge-message （Ack）を効率良く回収する方法である。この技術によりマルチキャスト通信の発信元における Ack 回収ボトルネックを解消することができる。

第 3 章はソフトウェア実装の中粒度通信同期機構である MBCF の開発背景、基本概念、基本動作が述べられている。MBCF は、普及品のネットワークインタフェースカードを利用する機構であり、MBP よりも大きな粒度で通信同期が実行される。このため、ネットワークやバスの使用効率を細粒度通信の MBP に比べて大幅に改善することができる。MBCF はネットワークインタフェースの受信割り込みルーチン内において、通信パケット内に含まれた通信対象タスクと操作対象アドレスの情報を利用して、直接ユーザタスクのメモリ空間を操作してメモリベース通信同期を実行する。このアドレス情報を活用した処理方式によって、無駄なデータコピーを 1 回も発生させない。メモリやタスクの保護はメインプロセッサのメモリ管理機構の流用により、新たな付加ハードウェアを必要とせず、ソ

フトウェアオーバヘッドを非常に抑えて実現される。

第4章はMBCFによって実現される機能の詳細が述べられている。取り扱うデータの粒度の差はあるが、本章において記述される機能はすべてのメモリベース通信同期機構に共通のものである。メッセージパッシング型の通信をユーザメモリ空間で実現するMBCF_FIFOや保護されたアクティブメッセージと同じ働きをするMBCF_SIGNALも機能と動作が紹介されている。

第5章はMBCFの高速実装方式に関する議論とMBCFの定性的な他通信同期機構との比較が提示されている。キャッシュを活用したプログラミングテクニックや高性能マイクロプロセッサのTLB等の内蔵機構を活用して、MBCFは非常に低オーバヘッドで実装可能である。また、定性的にMBCFはメッセージパッシング型の通信機構よりも自由度が高く、この自由度の高さのためのペナルティがほとんど存在しない。このため、MBCFの方がシステムが提供する通信同期機構として優れている。アクティブメッセージとMBCFとの比較では、アクティブメッセージがユーザプログラムを受信割り込みで直接走らせるため、保護と仮想化のオーバヘッドによってMBCFよりも実装コストが大きくなることを示している。この他にもFast MessageやPMと言ったソフトウェア実装の高速通信機構との比較を行い、MBCFがこれらの既存高速通信機構よりも定性的に優れていることを明らかにしている。

第6章はMBCFの実装例として、ethernet上へのMBCFシステムであるMBCF/Etherの実装方式とその実証実験について述べられている。1000BASE-SXとULTRA 60の組合せにおける性能測定実験の結果では、最大80.92MB/secのデータ転送能力と9.6μsecの片道レイテンシを達成した。最大転送能力が1000BASE-SXの理論値の1Gbit/sec (125MB/sec)に及ばないのはULTRA 60のハードウェア的な制約から来るものである。市販OSのTCP/IPとのレイテンシの比較では約10倍の性能（1/10の遅延時間）を達成している。

第7章では中粒度のメモリベース通信同期機構のための新しいソフトウェア分散共有メモリ構成法について提案している。MBCFは中粒度のソフトウェア実装の通信同期機構であるため、従来のソフトウェア分散共有メモリ方式を使った場合には、高い頻度のページトラップ発生によるオーバヘッドコストから逃れられない。そこで、ページトラップの流用という従来のソフトウェア分散共有メモリの常識を捨てた新しい方法論を提案している。それは、ユーザレベルでキャッシュエミュレーションを行うコードを最適化コンパイラによって元プログラムに挿入し、本来のコードとキャッシュエミュレーション用に挿入されたコードを一緒にして徹底的な最適化を施す方法である。MBCFと新方式を適用したソフトウェア分散共有メモリが開発され、これに対応した最適化コンパイラが実際に研究開発されている。その結果、最適化によって多くのプログラムでは通信粒度を大きく改善できることが実証実験によって示されている。

第8章ではハードウェア実装された中粒度のメモリベース通信同期機構について議論している。通信処理や通信に伴う暗号処理をメインプロセッサから肩代りするネットワークインタフェースアーキテクチャMBP2をMBCFに基づいて考案し、そのプロトタイプであるMBP2Pの開発を行っている。開発期間の短縮のために、MBP2Pはマルチチップ構成を採り、市販LSIと大規模field programmable gate arrayを組み合わせて開発されている。MBP2PはTLBを内蔵しており、ユーザタスクから直接DMAによってパケットを送信することが可能であり、受信時は直接ユーザタスク内のメモリにパケットの内容を反映させることができる。MBP2は当初の予定通りメインプロセッサの負荷を大幅に軽減することに成功したが、MBCFよりもレイテンシが大幅に悪化した。この原因は市販組込マイクロプロセッサにある。市販組込プロセッサは様々な速度低下要因を持っており、同一クロック周波数のメインプロセッサの数分の一程度の実力しかないことが明らかになった。

第9章では、本論文で提案したメモリベース通信同期機構とソフトウェア分散共有メモリの構成方式の特徴をまとめる。MBCFに関する研究の今後の展開に関し言及し、そして、MBP2アーキテクチャのような用途に使用可能な高性能汎用組込マイクロプロセッサの研究開発について言及している。

以上、本論文は並列分散処理にとって非常に重要な通信同期方式にMBCS方式という新しいアプローチを導入し、大きく異なる三つの実現形態に関して具体的に機構を設計し深く議論している。利用可能なメモリ空間を複数ノードに拡大しようという従来の単純な分散共有メモリの考え方ではなく、MBCS方式は通信同期サブシステムを汎用性を維持しつつ高速化を達成するために共有メモリの考え方を適用している。このMBCS方式は計算機科学にとっても非常に意義のあるアプローチである。よって本論文は理学博士を与えるにふさわしいものであると審査員全員一致で判定した。なお、研究成果の大きな部分は同氏に寄与によるものであり、共同研究者から、学位論文の内容として使用することの承諾を得ていることを確認している。