

論文要旨

市販の PC をノードとし、高性能なネットワークでノード間を相互接続して並列分散処理を行うクラスタコンピューティングは、安価で実用的な計算資源として、従来の大型機に代わるハイパフォーマンスコンピューティングの主要なプラットフォームとなりつつある。一般に、ハイエンドなクラスタコンピューティングは、ラックなどに高密度に集積した PC 間を Myrinet に代表される System Area Network (SAN) や高性能な Ethernet などを用いて相互接続した環境で行われる。

RHiNET は、オフィスなどにおいて日常業務で用いられている PC の余剰計算力を利用して、このようなハイエンドなクラスタコンピューティングを行うことを目的に開発された独自ネットワークである。RHiNET では、SAN に匹敵する高い通信性能と Ethernet のような接続性を両立させることで、このようなクラスタコンピューティングを実現する。

Martini は、この RHiNET 用に開発されたネットワークインタフェースコントローラである。Martini は、高い通信性能を実現するために、Remote Direct Memory Access (RDMA) を用いた通信機構をハードウェアで提供する。また、Martini は On-the-fly 通信機構と呼ばれる低遅延なパケット送出機構や、乗っ取り機構と呼ばれる協調処理機構などの実験的な機構を多数備える。

本研究では、Martini のコアロジック部および周辺のソフトウェアを中心に実装し、これを用いたシステムの構築・評価を行った。実機上で基本通信性能を評価した結果、Martini は 2 ノード間のメモリコピーにおいて 470Mbyte/sec の双方向スループットと 1.74 μ sec の最小レイテンシを示した。レイテンシに関しては最新鋭の SAN に匹敵する値を実現しており、RDMA をハードウェア実装したことが効果的であることが確認された。また、乗っ取り機構の有効性を検証するために、これを用いた通信機構を新たに実装した。評価の結果、乗っ取り機構の利用により、ソフトウェア単体で通信処理を行う場合に比べ高い通信性能が得られることが確認された。

並列分散処理システム下での Martini の評価は、RHiNET 上に既存のクラスタシステムソフトウェアである SCore を移植することで行った。SCore の低レベル通信ライブラリである PM はメッセージ通信を必要とするが、メッセージ通信は Martini のハードウェアに実装されていない。そこで、上位レイヤを用いて、Martini の RDMA を利用したメッセージ通信の実装を行った。評価の結果、16 ノード規模の環境においてアプリケーションレベルで台数に応じた性能向上が見られた。一方で、Martini が単純な RDMA を用いた通信しかハードウェアで提供していないことにより、システムが大規模化した際に、メッセージ通信のレイテンシが増大してしまうという問題点が明らかになった。