R24b　PC クラスタ用超低レイテンシ通信ライブラリの開発と性能評価

○似鳥 啓吾、牧野 淳一郎 (東大理)

我々は、安価なギガビット・イーサネット用のネットワークカードを用いて Myrinet 等の高価な通信ハードウェアに匹敵する性能を実現する通信ライブラリを開発し、並列 N 体シミュレーションコードでその性能を評価した。

汎用の PC を汎用のイーサネットネットワークで接続した PC クラスタは、ノードあたりのコストが低く価格性能比が優れているために近年急速に普及した。しかし、汎用のイーサネットの通信レイテンシ (遅延時間) が大きいことが多くの問題で並列化効率を制限しており、大規模な PC クラスタでは Myrinet, Infiniband 等の高価であるが通信レイテンシの小さいネットワークハードウェアが採用されていることが多い。

我々は、汎用のイーサネットの通信レイテンシのほとんどが通信プロトコル自身、あるいは OS やソケットライブラリの実装によるソフトウェア的なものであることに注目し、ソフトウェアのオーバーヘッドを極限まで小さくすることでレイテンシを削減することを試みた。例えば、OS を経由しないでユーザープロセスが直接にネットワークデバイスを操作し、TCP/IP プロトコルを使わずに低レベルのイーサネットパケットでやりとりする等の手法を使った。

Realtek 社の RTL8169 ネットワークインターフェースチップをターゲットにした実装では、一方向の通信レイテンシが約 5μs となり、通常の TCP/IP の約 60μs の 1/10 以下にすることに成功した。これは大型のクラスタで採用されている専用通信アダプタの速度に匹敵するものである。さらに、独立時間刻みの N 体シミュレーションプログラムを MPI で並列化したもので、レイテンシが問題となるいくつかの通信部分を今回開発したライブラリによるもので置き換えることで、大きな高速化を実現できた。

現在、ライブラリとしての API、ドキュメント等を整備しており、近い将来に公開することを予定している。