プラズマ乱流シミュレーションにおける通信マスク手法開発

井戸村　泰宏[1], 仲田　資季[1], 山田　進[1], 町田　昌彦[1], 今村　俊幸[2]
渡邉　智彦[3], 沼波　政倫[3], 井上　晃[4], 堤　重信[4], 三吉　郁夫[4], 志田　直之[4]

日本原子力研究開発機構[1], 理化学研究所[2], 核融合科学研究所[3], 富士通[4]

5次元ジャイロ運動論モデルに基づくプラズマ乱流シミュレーションは、核融合プラズマにおける乱流輸送現象を解析するための標準的な手法となっている。国際熱核融合実験炉ITERのような将来の大型装置を解析する上で、シミュレーションの適用範囲、特に、装置サイズの規模を拡大することが重要となっており、より大きな計算機資源が必要とされている。しかしながら、現在のペタスケール計算機、あるいは、将来のエクサスケール計算機の能力を引き出すには、1)コア/CPU/ノード/ネットワークの複雑な階層構成、および、2)10万コアを超える並列度、というハードウェアの厳しい要求を満たす必要がある。本研究では、差分法に基づくプラズマ乱流コードGT5D[1]において、これらの要求を満たす並列計算技術の開発を推進している。

1)についてはハードウェアの階層構造に適合するような多階層ネットワークを複数のMPI層とOpenMP層で構築し、多次元領域分割を行う並列化手法[2]を既に報告している。この並列化手法の特徴は全てのMPIコミュニケータのサイズを100程度以下に抑えつつ100万MPIプロセス（1000万コア以上）の並列化を実現できること、および、全ての集団通信を1階層のMPIコミュニケータのみで閉じ、複数の集団通信を同時実行することにより、バイセクションバンド幅を最大限に活用できることである。しかしながら、[2]で報告したストロングスケーリングの並列化率（アムダール則）は約99.996%となっており、10万コアを超える領域の超並列計算は困難であった。そこで、2)を解決するために、今回、通信と演算のオーバーラップにより通信コストを隠蔽する通信マスク手法の開発を行った。

GT5Dにおける主要な通信コストは領域分割した差分演算における袖領域の1対1通信、および、並列化軸の異なるソルバー間のデータ転置のための集団通信で占められている。袖領域通信のオーバーラップは差分演算を袖領域を参照する部分とそれ以外に分割し、後者をMPI_Isend/MPI_Irecvのような非同期1対1通信とオーバーラップすることによって実現される。しかしながら、今回行った数値実験では、このような標準的な実装では通信と演算のオーバーラップが実現しないことが判明した。多くのMPIライブラリにおいてメッセージ長の大きな1対1通信の実装は通信開始前に制御通信の往復を必要とするRendezVouzプロトコルに基いている。上記の実装方法だと制御通信の完了前に演算が始まってしまうために、制御通信の往復が妨げられ、結果的に演算完了後にMPI_Waitが起動されるまで通信が処理されない[3]。この問題は京、FX1、BX900 (Nehalem-EP + InfinibandQDR)、Helios (SandyBridge-EP + InfinibandQDR)における富士通製MPI、Intel製MPI、Bull製MPIにおいて共通の問題であった。

この問題を解決するために、本研究では2つの通信マスク手法を開発した。一つは、演算中にMPI_Testのようなダミー MPI関数を定期的にコールして一時的にMPIプロセスを動かすことにより制御通信を促進する方法、もう一つは非対称なハイブリッド並列モデルによってマスタースレッドを通信完了まで通信処理のみに割り当てる方法である。前者はMPI_Testのオーバーヘッド以外は完全に通信と演算をオーバーラップできるのに対し、後者は実質的に通信コストをマルチコアで分割する、例えば、8コアであれば通信のオーバーヘッドが1/8に減少することになるので、原理的には前者のほうが高速な処理が可能である。一方、後者は非同期1対1通信だけでなく、同期通信や集団通信でも汎用的に適用可能な技術であり、実際、GT5Dのデータ転置処理にも適用された。これらの通信マスク手法の適用によって、GT5Dの並列化率は飛躍的に向上した。図は京およびHeliosという通信機構およびMPIライブラリの全く異なる2つの環境における処理性能のストロングスケーリングを示すが、特に、京では約20万コアまで良好なスケーリングが得られ、従来に比べて一桁以上高い約99.9998%という並列化率を達成した[4]。

[1] Idomura et al., Comput. Phys. Commun 179, 391 (2008).
[2] Idomura et al., SC11; HPCS2011.
[3] Sayantan et al., PPoPP06.
[4] Idomura et al., SC12.

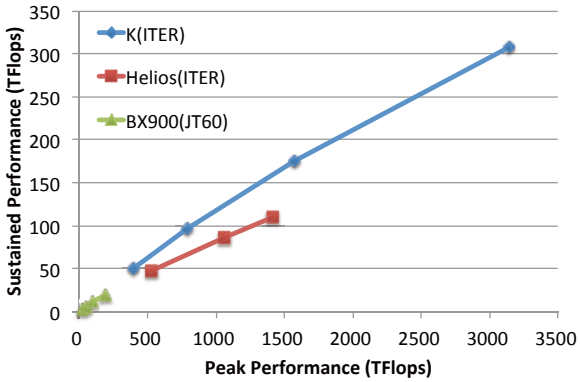

Fig. The strong scaling of GT5D on BX900, Helios, and K. Helios and K use ITER size parameters (N_R, N_z, N_Z, $N_{v//}$, N_m) = (768, 64, 768, 128, 32). BX900 uses JT-60U size parameters (N_R, N_z, N_Z, $N_{v//}$, N_m) = (240, 64, 240, 128, 32).