30E20P

ペタスケールジャイロ運動論シミュレーションのための通信マスク手法開発
Development of communication overlap techniques for Peta-scale gyrokinetic simulations

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]
Y. Idomura[1], M. Nakata[1], S. Yamada[1], M. Machida[1], T. Imamura[2]

日本原子力研究開発機構[1], 理化学研究所[2]
Japan Atomic Energy Agency[1], Riken[2]

Gyrokinetic simulations are essential and powerful tools in studying turbulent transport in fusion plasmas. In order to enhance the capabilities of gyrokinetic simulations, further improvements both in simulation models and in computing power are needed. In particular, the plasma size scaling of turbulent transport is one of critical issues, which require extreme scale computing. However, it is not so easy to extract high computing power from modern Peta-scale supercomputers because of 1) very complicated hierarchy of core, CPU, node, and network architectures, 2) extreme concurrency beyond 100k cores. In order to satisfy these technical needs, novel parallel computing techniques are developed on a gyrokinetic toroidal 5D Eulerian code GT5D [Idomura et al., Comput. Phys. Commun. 2008; Nucl. Fusion 2009].

In order to fit the code structure to hardware architectures, multi-layer hybrid parallelization models with multi-dimensional domain decomposition are designed following the physical symmetry properties of each operator in basic equations. In this model, all the communications within and between the solvers (operators) are implemented on a single layer of the hierarchical network consisting of multiple MPI layers and a SMP layer. This avoids collective communications among all the MPI processes, makes extensive use of the bi-section bandwidth, and suppresses the size of each MPI communicator below ~100 up to ~10^6 MPI processes. In addition to this global network design, novel communication overlap techniques are developed to achieve the extreme concurrency. The techniques use either non-blocking communications with additional control sequences or communication threads with heterogeneous OpenMP implementations. Our communication overlap techniques are based on conventional MPI libraries, and work on most of existing platforms based on dedicated and commodity networks.

Performances of newly developed techniques are tested on two Peta-scale platforms, the K-computer at Riken (K), and the Helios at the IFERC (Helios). A Torus network on K naturally fit to the domain decomposition used in GT5D, and therefore, basic communication performances and their scaling against the number of nodes are better than Helios. By applying the communication overlap techniques, communication costs are dramatically reduced both on K and on Helios, and the strong scaling on K achieved the parallel efficiency of ~99.9998% (See Fig.). The parallel computing techniques accelarate GT5D significantly, and enable us to access new physics regimes both in plasma sizes and in time scales.

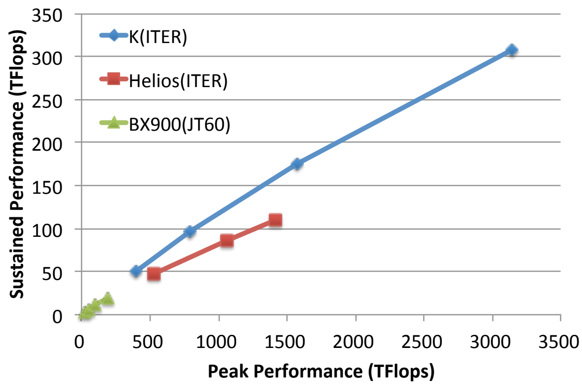

Fig. The strong scaling of GT5D on BX900, Helios, and K. Helios and K use ITER size parameters (N_R, N_z, N_Z, $N_{v//}$, N_m) = (768, 64, 768, 128, 32). BX900 uses JT-60U size parameters (N_R, N_z, N_Z, $N_{v//}$, N_m) = (240, 64, 240, 128, 32).