Project ID: jh180081-NAHI

Optimisation of Fusion Plasma Turbulence Code toward Post-Petascale Era III

Project Representative's Name（Affiliation）

Yuuichi Asahi (National Institutes for Quantum and Radiological Science and Technology)

Abstract

In this project, we improve the implementations of plasma turbulence simulation codes and use them for physical analysis. As a new approach for further reduction of communication costs, we have tested a task-level parallelization in a mini-application of GYSELA code. So far, we have ported the mini-app to Xeon Phi KNL and GPUs to find that the large amounts of computational tasks are not ideal for GPUs, while they do not harm the performance on KNLs. On GPUs, tasks should be decomposed in more coarse-grained way, where a single task includes some amounts of computations or communications. For physical analysis, we have shown that turbulence can generate poloidal asymmetry which can affect the collisional energy transport. In parallel, we have developed a hybrid kinetic electron model for more realistic simulations and completed the basic verifications.

1. Basic Information

(1) Collaborating JHPCN Centers

Tokyo Institute of technology

Nagoya University

(2) Research Areas

- ☑ Very large-scale numerical computation
- ☐ Very large-scale data processing
- ☐ Very large capacity network technology
- ☐ Very large-scale information systems

(3) Roles of Project Members

Project representative Yuuichi Asahi works for the optimization and development of simulation codes. The deputy representative Shinya Maeyama performs the plasma turbulence simulations and investigates the characteristics of turbulent transport with the presence of kinetic electrons. The deputy representative Guillaume Latu contributes on the task level parallelization of GYSELA mini-app. Xavier Garbet works for theoretical analysis of non-local transport processes. Prof. Watanabe gives comments on characteristics of local transport processes. Prof. Ogino supports the optimization on FX100, particularly for the effective usage for assistant cores. Prof. Aoki gives advices about the usage of NV-link on TSUBAME3.0 to minimize the communication costs.

2. Purpose and Significance of the Research

Five dimensional Gyrokinetic simulations are powerful tools to analyse and predict turbulent transport in magnetic confined plasmas. Since huge computational resources are needed for production runs, it is attractive to take advantage of the high computational performance of the state-of-the-art many core architectures including Fujitsu FX100, Xeon Phi KNL and Tesla P100.

We have already introduced communication and computation overlapping to GYSELA and GKV in FY2017. We found performance degradation of GYSELA with 512 KNLs. Detailed profiling shows that the performance is degraded mainly by Poisson solver which is a complex mixture of

communications and computations. There is no straight forward way to optimize this kind of kernel manually, since there are quite many possible combinations of overlapping. The purpose of computational research in FY2018 is testing the applicability of the task-level parallelization to achieve an efficient overlap. Once tasks are defined appropriately, we do not have to think about the detailed implementations of overlapping, rather computations and communications are overlapped automatically at run-time. Since the task-level parallelization could be applicable to conventional fluid simulations, our work will also contribute to the other stencil-based fluid simulations.

The plasma physics targets in FY2018 are the analysis of the impact of poloidal asymmetries on transport processes and the verification of a hybrid electron model. Recently, a lot of attentions have been paid to the role of poloidal asymmetries in magnetic confined plasmas. Although there are some experimental evidences that they could have significant impacts on transport processes, there are no theoretical expiations for that. In this work, we investigate the drive of poloidal asymmetries and their impacts on transport processes with the global gyrokinetic code GYSELA. Global gyrokinetic codes are the most promising tools to predict the plasma profile formation and turbulence behavior consistently.

Secondly, we introduced a hybrid kinetic electron model in GKV and completed the basic verifications. The introduction of kinetic electrons to GYSELA will allow a more realistic and reliable modeling. Unfortunately, it is more complicated to introduce full kinetic electron model in global codes than in local codes like GKV. So far, we have tested the so-called hybrid-kinetic electron model [Y. Idomura et al., J. Compute. Phys. 313, 511 (2016)] in GYSELA. In order to understand its limitation, we would like to compare the hybrid-kinetic electron model with the full kinetic electron model by implementing both models in GKV code. We can improve the reliability of kinetic electron model in global gyrokinetic code in this way.

3. Significance as a JHPCN Joint Research Project

A JHPCN Joint Research Project offers state-of-the-art platforms like FX100 and TSUBAME3.0, which is essential for our multi-platform development. In FY2017, we have compared the implementation of transpose communication in GYSELA and GKV code to find that the GYSELA implementation is more efficient. In the GPU version of GKV code, the transpose communication is implemented in the GYSELA pattern and the communication cost is reduced by the factor of three. In FY2018, we have updated the CPU version of GKV code to use the improved communication pattern. This gives the factor of 2 speed up of convolution part. Based on the JHPCN Joint Research framework, we can share the optimization strategies which benefits each side.

Another advantage is the international collaboration. The French group has focused on global aspect of plasma turbulence with GYSELA code and the Japanese group has worked for the local

plasma turbulence with GKV code. French group can offer the physical understanding of non-local transport. Japanese group can offer the details of different hybrid-kinetic electron models and their limitations. This collaboration benefits both sides. In addition, we can share some ideas for better implementation of task-level parallelization.

4. Outline of the Research Achievements up to FY2017

High Performance Computing

The communication and computation overlapping has been introduced to 5D gyrokinetic codes GYSELA and GKV. In order to anticipate some of the exa-scale requirements, these codes were ported to the modern accelerators, Xeon Phi KNL and Tesla P 100 GPU. On accelerators, a serial version of GYSELA on KNL and GKV on GPU are respectively 1.3x and 7.4x faster than those on a single Skylake processor (single socket).

We have measured the scalability of GYSELA on Xeon Phi KNL from 16 to 512 KNLs (1024 to 32k cores) and GKV performance on Tesla P100 GPU from 32 to 256 GPUs. By applying overlapping, the GYSELA 2D advection solver has achieved a 33-92 % speed up and the GKV 2D convolution kernel has achieved a factor of 2 speed up with pipelining. The task-based approach gives 11-82 % performance gain in the derivative computation of the electrostatic potential in GYSELA. This work is submitted to a peer reviewed journal [1].

As an algorithm level optimization, implicit collision solver for multi-species collisions is developed on FX100. The implicit collision solver allows larger time step sizes than explicit solver at the cost of several iterations for convergence. More importantly, the iterative operations can be performed without communications (transpose communications are needed only once before iterations). This work is published as a peer reviewed article in FY2018 [2].

Plasma Physics

For Plasma physics, we have completed the linear and nonlinear analyses of high pressure plasma focusing on the microtearing mode. It is revealed that electron-scale turbulence could suppress the ion-scale micro tearing mode through the cross-scale interactions. This work is published in a peer reviewed journal [S. Maeyama et al, Phys. Rev. Lett. 119, 195002, (2017)].

5. Details of FY2018 Research Achievements

In FY2018, we have carried out the following subjects: (5.1) task level parallelization, (5.2) developing a 4D (space 2D and velocity space 2D) mini-app with Kokkos framework [https://github.com/kokkos/kokkos] for better portability, (5.3) investigating synergies between turbulence and collisional transport through global structure and (5.4) developing a hybrid electron model.

5.1 Testing the task level parallelization

In Q1 and Q2, we have analysed the performance of a mini-app solving the 2D Vlasov and Poisson system with Semi-Lagrangian method, which is parallelized with MPI and OpenMP task pragmas. In

this mini-app, the 2D plane consists of multiple tiles which include a part of distribution function f. A single task is defined as a kind of operation acting on a tile including computations and communications.

In order to understand how tasks are executed, we monitored the thread activities by libkomp software [https://gitlab.inria.fr/openmp/libkomp] tracer which are then visualized by vite tool [http://vite.gforge.inria.fr/index.php] as shown in Fig.1. This kind of figure gives us an intuitive understanding about the task execution order in each thread. The basic strategy is to reduce the idle parts represented by the gray colors.

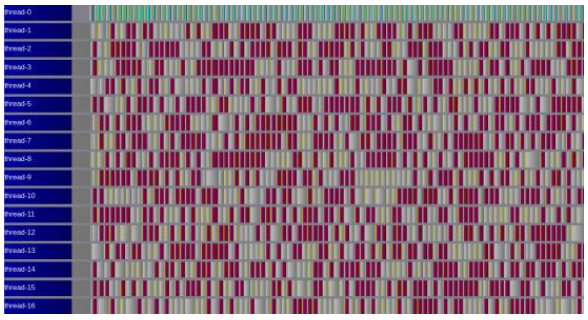

Fig 1. Gantt chart of GYSELA mini app on 32 KNL nodes. The red parts correspond to internal computations. Thread 0 works for communications. Thread level parallelization is implemented with OpenMP task pragmas.

Although the parallelization with OpenMP task pragmas works well on CPUs (OpenMP task version sometimes outperforms classic OpenMP version), it is not the case for GPUs. Through the performance analysis of GPU version, we found that the large amounts of computational tasks are not ideal for GPUs, while they do not harm the performance on CPUs. On GPUs, tasks should be decomposed in more coarse-grained way, where a single task includes some amounts of computations or communications. This view should be kept in mind, for the future development of a more complicated app.

5.2 Developing 4D Vlasov + Poisson mini-app with Kokkos framework

In Q3 and Q4, we have started to develop a more complicated and realistic mini-app in order to test several types of communication/computation patterns. For this purpose, we have developed a 4D (space 2D and velocity space 2D) Vlasov and Poisson mini-app, which is more relevant to our 5D (space 3D and velocity space 2D) production codes. From the numerical point of view, this app employs Semi-Lagrangian scheme to solve 4D Vlasov equation and spectral method is used to solve 2D Poisson equation (thus periodic boundary condition).

As a new challenge with this application, we have implemented a code based on a new parallel programming framework, Kokkos, in order for a better performance portability. Performance portability is a critical issue in the near future, since upcoming exascale machines can be either GPU or ARM based machines, both of which require quite different optimization techniques. Without a portability, we have to keep at least two different versions leading to almost doubled developing costs.

Kokkos framework may offer a reasonable solution. At least, a porting cost from Kokkos OMP to Kokkos CUDA or Kokkos CUDA to Kokkos OMP can be

reduced significantly. The new mini-app (so far without MPI parallelization, 1000 code lines) is parallelized with OpenMP. It took us 2 days to port it to Kokkos CUDA background, which is the toughest part of this porting. Surprisingly, a porting cost from Kokkos CUDA to Kokkos OMP is around 10 minutes, where we can reuse all the codes except for FFT (Fast Fourier transform) wrappers. It should be noted that we have already developed 2D FFT wrappers in CUDA (cufft) and OpenMP (fftw) for GKV code.

We measured performances with Kokkos CUDA, Kokkos OMP and OpenMP versions on TSUBAME3.0. With the problem size of (32x32x32x32) and 128 iterations, it took 0.15s, 4.5s and 1.8s with Kokkos CUDA on (P100 GPU), Kokkos OMP (Broadwell CPU, single socket) and OpenMP (Broadwell CPU, single socket) versions, respectively. Since the OpenMP version is highly optimized for CPUs, it is difficult to achieve the same performance with Kokkos OMP version. To explore further optimizations, we can replace the bottlenecked kernels with more CPU dedicated ones.

5.3. Synergies between turbulence and collisional transport

In Q1 and Q2, we have investigated the impact of convective cells (poloidal asymmetry) on transport processes. Firstly, we have confirmed that poloidal convective cells are driven by plasma turbulence through the frequency analysis of turbulent Reynolds stress tensor. It is also confirmed that the in-out asymmetry of convective cell (Fig 2.) agrees with a feature predicted by a theoretical analysis [4].

In order to understand the impact of convective cells, we have applied a numerical filter to them and compared the simulation results with and without the filter. As shown in Fig 3., it turned out that the collisional transport can be reduced by a factor of about 2 once the numerical filter is applied. Since the origin of the convective cells are turbulence, the impact of convective cells on the collisional transport can be interpreted as a synergy between turbulence and collisional dynamics. This is a completely new perspective, which can be modeled by a global gyrokinetic code, which solve turbulent and collisional transport consistently (no scale separation).

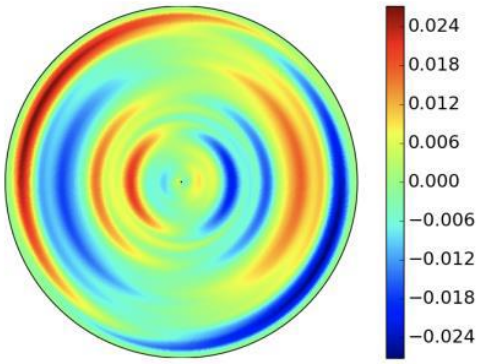

Fig 2. Poloidal cross-section of meso-scale mode structure in the electrostatic potential computed by GYSELA [3]. The structure clearly shows a cosine like (in-out) asymmetry.

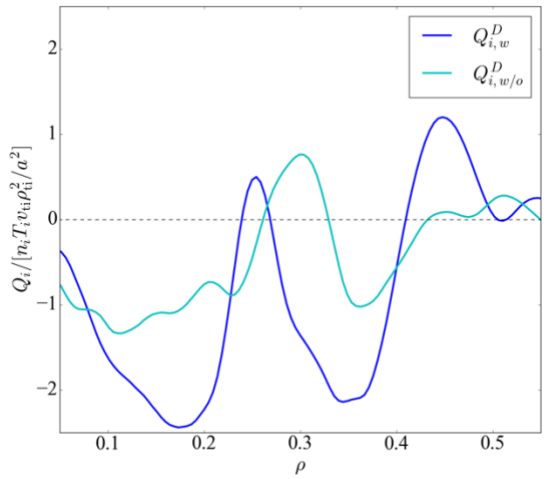

Fig 3. The radial profiles of the collisional energy flux Q^D with and without poloidal convective cells [3].

This work is presented as invited talks in the international conferences [6], [10] and published as a peer reviewed article [3].

5.4. Development of a hybrid electron model

In Q3 and Q4, we have introduced a hybrid electron model in GKV and GYSELA. In magnetic confined plasmas, plasma particles can be decomposed into trapped and passing particles. Particle trapping is an intrinsic phenomenon in magnetic confined plasmas, caused by the so-called mirror effects of magnetic field. In general, it is often considered to be a good assumption that passing electrons are highly mobile and respond adiabatically. In contrast, trapped electrons are considered to behave kinetically. In our hybrid electron model, passing electrons are assumed to respond adiabatically, while trapped electrons are assumed to respond kinetically.

For verification, we have computed linear growth rate using GKV code with different fractions of passing and trapped electrons as shown in Fig. 4. Physically speaking, the fraction is determined by magnetic configuration but we change it for numerical tests. Low value of kappa means almost no trapping and high value of kappa means electrons are almost fully trapped. As expected, the growth rate is quite close to the adiabatic case if we choose kappa = 0.2 (red circle and black dotted line). With the high value of kappa (kappa = 10), the growth rate approaches to the value with the kinetic case (purple triangle and black solid line). Thus, it is confirmed that hybrid electron model has been introduced successfully in GKV code.

Although the same model is also introduced in GYSELA, we have not completed the verification tests due to the high computational costs. This remains as a future task.

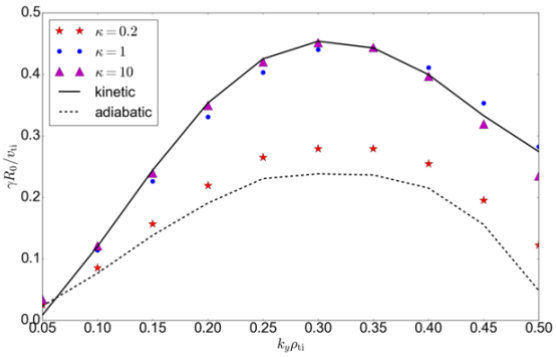

Fig 4. Dependence of linear growth rate of Ion temperature gradient mode on the fraction of passing and trapped electrons.

6. Progress of FY2018 and Future Prospects

High performance computing

As planned, we have developed a mini-app using task-level parallelization and tested it on different platforms including P100 GPU, Xeon Phi KNL and Xeon SKL. The performance analysis tools for thread

activities have also been introduced. Through the performance analysis of GPU version, we found that the large amounts of computational tasks are not ideal for GPUs, while they do not harm the performance on CPUs. On GPUs, tasks should be decomposed in more coarse-grained way, where a single task includes some amounts of computations or communications.

After basic testing of OpenMP-task in Q1 and Q2, we have started to develop a more complicated and realistic 4D mini-app in order to test several types of communication/computation patterns. By using the Kokkos framework, we have confirmed that the exactly same code works on both CPUs and GPUs. Unfortunately, the Kokkos OMP version is 2 times slower than the OpenMP version optimized for CPUs. As a next step, we will replace the slow Kokkos kernels with more CPU dedicated ones. Then, we will parallelize the code with MPI and apply the task-level parallelization.

Plasma physics

We have elucidated the role of poloidal asymmetries in turbulent and collisional transport processes. We have first revealed that plasma turbulence can generate the poloidal asymmetry, which is consistent with the theoretical prediction. In order to understand the impact of poloidal asymmetries, we have applied a numerical filter to them and compared the simulation results with and without the filter. It turned out that the collisional transport can be reduced by a factor of about 2 once the numerical filter is applied. This can be interpreted as a synergy between turbulence and collisional dynamics. This work is published as a peer reviewed article [3] in FY2018.

In parallel, we have implemented a hybrid electron model in GKV. Through the sensitivity test with respect to the fraction of passing and trapped electrons, it is confirmed that the growth rate approaches to the adiabatic limit if the trapping fraction is low and vice versa.

Although the same model is introduced in GYSELA, we have not completed the verification tests due to the high computational costs. This remains as a future task.

International collaboration

From 5/November to 9/November, the project representative Yuuichi Asahi had visited CEA France to discuss the plasma physics part with Xaveri Garbet and the HPC part with Guillaume Latu. With Xavier Garbet, we have intensively discussed the theoretical analysis part with respect to the poloidal asymmetry. With Guillaume Latu, we have discussed the code development strategy toward the exa-scale computing, where we concluded the performance portability is of utmost importance. The representative Yuuichi Asahi thanks the financial support from JHPCN for this face-to-face meeting.

In addition to the face-to-face meeting, we had visio-meetings every two months in FY2018. We will continue to hold visio-meetings in FY2019.

7. List of Publications and Presentations

(1) Journal Papers

[1] (Submitted) Yuuichi Asahi, Guillaume

Latu, Julien Bigot, Shinya Maeyama, Virginie Grandgirard, Yasuhiro Idomura, "Overlapping communications in gyrokinetic codes on accelerator-based platforms", submitted to Concurrency and Computation: Practice and Experience (2018)

[2] S. Maeyama, T.-H. Watanabe, Y. Idomura, M. Nakata, and M. Nunami, "Implementation of a gyrokinetic collision operator with an implicit time integration scheme and its computational performance", Comput. Phys. Commun. (2018), https://doi.org/10.1016/j.cpc.2018.07.015.

[3] Yuuichi Asahi, Virginie Grandgirard, Yanick Sarazin, Peter Donnel, Xavier Garbet, Yasuhiro Idomura, Guilhem Dif-Pradalier, Guillaume Latu, "Synergy of turbulent and neoclassical transport through poloidal convective cells", Plasma Physics and Controlled Fusion (2019), https://dx.doi.org/10.1088/1361-6587/ab0972

[4] P. Donnel, X. Garbet, Y. Sarazin, Y. Asahi, F Wilczynski, E Caschera, G Dif-Pradalier, P Ghendrih[1]and C Gillot, "Turbulent generation of poloidal asymmetries of the electric potential in a tokamak", Plasma Physics and Controlled Fusion (2018), https://doi.org/10.1088/1361-6587/aae4fe

[5] G. Latu, Y. Asahi, J. Bigot, T. Fehér, V. Grandgirard. Scaling and optimizing the Gysela code on a cluster of many-core processors. *SBAC-PAD 2018, WAMCA workshop*, Sep 2018, Lyon, France. https://hal.inria.fr/hal-01719208

(2) Conference Papers

(3) Oral Presentations

[6] (Invited) Y. Asahi, V.Grandgirard, Y. Idomura, G. Latu, Y. Sarazin, G. Dif-Pradalier, X. Garbet, P. Donnel, "Benchmarking of flux-driven full-F gyrokinetic simulations", 2nd Asia-Pasific Conference on Plasma Physics, Kanazawa, Japan, November 14, 2018.

[7] Y. Asahi, S. Maeyama, T.-H. Watanabe, Y. Idomura,"Accelerating and modernizing delta-f gyrokinetic codes on GPUs", The HPC Workshop, Rokkasho, Aomori, October 11, 2018 (visio).

[8] Y. Asahi, S. Maeyama, T.-H. Watanabe, Y. Idomura,"Porting GKV on Tsubame", the 4th US-Japan JIFT Exascale Computing Workshop, Princeton, US, July 31, 2018.

[9] (Invited) S. Maeyama, "Roles of sub-ion-scale structures on cross-scale interactions in Tokamak plasma turbulence", 11th Plasma Kinetic Working Meeting, Vienna, Austria, July 30, 2018.

[10] (Invited) X. Garbet, Y. Asahi, E. Caschera, P. Donnel, G. Dif-Pradalier, P. Ghendrih, V. Grandgirard, Ö. Gürcan, G. Latu, P. Hennequin, Y. Sarazin, A. Smolyakov, L. Vermare, "Impact of flow poloidal asymmetries on transport in tokamaks", 45thConference on Plasma Physics, Prague, Czech Republic, July 2018.

(4) Others