jh190065-NAHI

Modernizing and accelerating fusion plasma turbulence codes targeting exa-scale systems

Yuuichi ASAHI（Japan Atomic Energy Agency）

Abstract
We aim to explore a performance portable approach for fusion plasma turbulence simulation codes employing a kinetic model such as GYELA. For this purpose, we have encapsulated the key features of GYSELA such as the high dimensionality and the semi-Lagrangian scheme, and encapsulate them into a mini-application which solves the similar but a simplified Vlasov-Poisson system (2D space and 2D velocity space). We implement the mini-app with a mixed OpenACC/OpenMP and Kokkos, where we suppress unnecessary duplications of code lines for the better readability, portability and productivity. For a reference problem size of 128^4, the Skylake (Kokkos), P100 (OpenACC), and P100 (Kokkos) versions achieved an acceleration of 1.45, 12.95 and 17.83, respectively, with respect to the baseline Skyalke OpenMP version. It turned out that it is important to use the multidimensional range policy for the nested multidimensional loops for the better vectorization and cache effects. For plasma physics, we have performed the linear benchmark of hybrid electron model that are newly introduced in GYSELA. The benchmark against GT5D gives the pretty good agreement in the real eigen frequency, but some discrepancy in the linear growth rate, particularly in the trapped electron mode dominant region.

1. Basic Information

(1) Collaborating JHPCN Centers

Tokyo Institute of technology
Nagoya University

(2) Research Areas

Very large-scale numerical computation

(3) Roles of Project Members

Project representative Yuuichi ASAHI works for a performance portable implementation of a GYSELA mini-app. The deputy representative Shinya Maeyama performs the plasma turbulence simulations and investigate the characteristics of turbulent transport with the presence of a 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 of 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 Research

Performance portability is considered to be an inevitable requirement in the coming exa-scale supercomputing era. A high diversity in computer architectures is expected for the upcoming supercomputers, including Intel/Cray CPU/GPU machine Aurora and AMD/Cray CPU/GPU machine Frontier and Fujitsu ARM CPU machine Fugaku in Japan. In order to use these architectures efficiently, we have to establish code refactoring approaches to easily get a good performance on them.

Our goal is to explore a performance portable approach for fusion plasma turbulence simulation codes employing a kinetic model like the GYSELA code [V. Grandgirard et al, CPC 2016]. For this purpose, we extract the key features of GYSELA such as the high dimensionality and the semi-Lagrangian scheme, and encapsulate them into a mini-application which solves the similar but a simplified Vlasov-Poisson system (2D space and 2D velocity space). We implement the mini-app with a mixed OpenACC/OpenMP and Kokkos, where we suppress unnecessary duplications of code lines for the better readability, portability and productivity. In the mixed OpenACC/OpenMP version, we have changed only the directives by switching macros and kept the code structure. In the Kokkos version, the single codebase works on both CPUs and GPUs, meaning the great portability. By comparing these two approaches, we clarify the advantages and disadvantages of the directive based-approach (mixed OpenACC/OpenMP) and the higher level abstraction approach (Kokkos) from the view point of performance portability, readability and productivity.

The plasma physics targets in FY2019 are the analysis of the impact of poloidal asymmetries on transport processes and the verification of a hybrid electron model. Experimentally, poloidal asymmetries on transport processes are found to impact on transport processes, but there are no theoretical explanations for that. This is a global effect which can be modeled with the GYSELA code. Secondly, we will introduce the hybrid kinetic electron model [Y. Idomura, J. Compute. Phys. 313, 511 (2016)] to GYSELA. We will perform the linear benchmark with the GT5D code.

3. Significance as JHPCN Joint Research Project

Establishing a performance portable approach with a good scalability is a very challenging task which needs a good understanding of the large variety of architecture characteristics and programming models. Japanese group can offer the GPU optimization strategies with Kokkos and directives, and French group can offer the optimization techniques for many-core CPUs, namely ARM CPUs. Using the multi-platforms offered by JHPCN framework helps to test the performance portability over different types of supercomputers.

4. Outline of Research Achievements up to FY2018

In FY2018, we have carried out the following subjects: appling the task level parallelization to a mini-app and investigating a performance portable implementation of 4D Vlasov-Poisson code, and investigating synergies between turbulence and collisional transport through the poloidal asymmetry.

<u>Testing the task level parallelization</u>

In order to investigate the performance of communication and computation overlapping with task level parallelization, we have developed a mini-app solving the 2D Vlasov and Poisson (1D space and 1D velocity space) system with Semi-

Lagrangian method, parallelized with MPI and OpenMP task pragmas.

Contrary to the overlapping with loop-based OpenMP (associating the master thread to MPI communication), we can reduce the thread synchronization costs in task-based OpenMP by localizing the synchronization based on the task dependencies. We also found that task submission costs are not necessarily negligible. By minimizing the amount of dependencies per tasks and reducing the amount of tasks, we have obtained the better scalability with the task-based OpenMP. Adding priorities on the MPI tasks also improves the performance, since they should be executed as soon as possible.

<u>Performance portable implementation with Kokkos framework</u>

In order to establish the performance portable approach for the coming exa-scale computing, we tested the Kokkos framework as a new way of parallelization. We have succeeded to port our 4D mini-app with Kokkos but the performance on CPUs was insufficient. At that moment we have concluded that the Kokkos gives reasonable performance on GPUs but not on CPUs. Since the single codebase with Kokkos works on both CPUs and GPUs, this approach gives the great portability followed by readability and productivity. Thus, we continued further investigation for Kokkos framework in FY2019.

<u>Synergies between turbulence and collisional transport</u>

Using the French gyrokinetic code GYSELA, we investigated the impact of convective cells on transport processes. We perform two simulations with and without numerical filter for poloidal asymmetries. By comparing the results, we found that the collisional energy flux is reduced by a factor of about 2 once numerical filter is applied. Since the turbulence generates the poloidal asymmetry, it can be concluded that the turbulence and collisional transport can interact though the poloidal asymmetry. This work has been published as a peer-reviewed article in FY2019.

5. Details of FY2019 Research Achievements

<u>High Performance Computing</u>

In Q1 and Q2, we have developed the 4D Vlasov-Poisson code with the higher level abstraction Kokkos and the directives OpenACC/OpenMP. After measuring the performance, it turned out that more than 95 % of the elapsed time comes from the advection kernels in Vlasov solver (1D advection along each direction with Largrange interpolation) and the integral kernel in Poisson solver. Thus, in the present work, we only measure the performance of these kernels.

Figure 1. The obtained performance of the OpenACC/OpenMP and Kokkos versions.

We have measured the performance on Intel Skylake (Skylake), Nvidia Tesla P100 (P100), and ThunderX2 (Arm). Figure 1 shows the obtained performance of the OpenACC/OpenMP and Kokkos versions, with the problem size of 128^4. After some investigations, it turned out that the usage of 3D (or higher) MDrange policy to parallelize the nested loop is a key to accelerate Kokkos kernels both on CPUs and GPUs (see https://github.com/yasahi-hpc/vlp4d for the detailed implementation). This promotes the vectorization on CPUs and better cache effects on GPUs. As a result, the Kokkos versions achieved the better performance than the mixed OpenACC/OpenMP version (some of the kernels have achieved almost the ideal performance). The Skylake (Kokkos), P100 (OpenACC), P100 (Kokkos) versions achieve an acceleration of 1.45, 12.95 and 17.83, respectively, with respect to the baseline OpenMP version on Intel Skylake. Our objective to investigate the intranode performance portability with directives and the higher level abstraction has achieved in Q1 and Q2. This work was presented in the international conferences and the workshops [2, 4, 8]. The French collaborate will present this work in the international conference next year [3].

In Q3, we have further tried the OpenMP4.5 for GPU offloading.

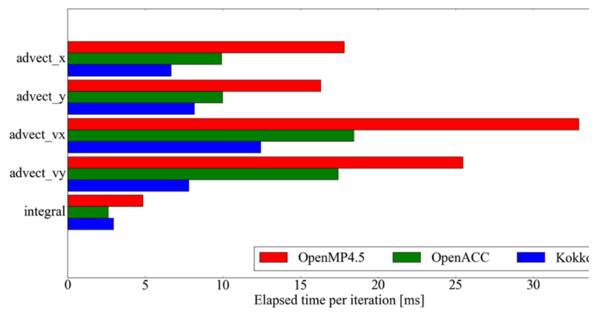

Figure 2. The elapsed time of the OpenMP4.5, OpenACC, Kokkos versions for one iteration on Nvidia Tesla V100.

For OpenMP4.5, we use the xl/16.1.1 compiler with compiler options "-qsmp=omp -qoffload -Xptxas -v". With this setting, we have found that the STREAM Triad bandwidth is 540 GB/s which is 30% slower than OpenACC or Kokkos versions. This partially explains the slower performance of OpenMP4.5. We will test other compilers and compiler options to use OpenMP4.5 more effectively in FY2020.

<u>Plasma physics</u>

Our plan to verify the hybrid kinetic electron model in GYSELA is partially achieved by the linear benchmarking with the GT5D code. Figure 3 shows the ion temperature gradient R_0/L_{ti} dependence of the most unstable mode with n = 15 computed by GYSELA and GT5D. If kinetic electrons are appropriately treated, it is expected that the most unstable linear stable mode changes from the trapped electron mode (TEM) to the ion temperature gradient (ITG) driven mode as the ion temperature gradient increases. As expected, the sign of the eigen real frequency changes from positive to negative at around $R_0/L_{ti} = 5.5$ as shown in Fig. 3 (b).

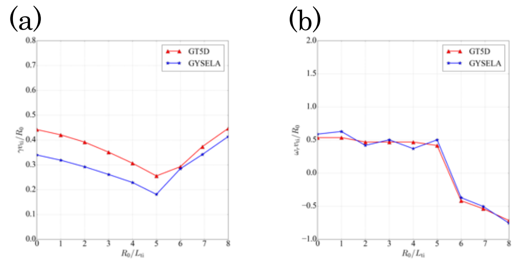

Figure 3. The R_0/L_{ti} dependence of (a) the linear growth rate and (b) the eigen frequency for the Cyclone like parameters. The growth rate and frequency are computed at r/a = 0.5.

In contrast to the good agreement in the eigen frequency, we find the discrepancy in the linear growth rate particularly in the TEM dominated region with $R_0/L_{ti} < 5$. We are currently investigating the cause of difference, which may be related to the treatment of passing-trapped boundary of kinetic electrons.

The analysis of the impact of poloidal asymmetries on transport processes has been almost completed in FY2018. This work is published in FY2019 as a peer reviewed article [Y. Asahi et al, PPCF, 2019]. The results are also presented in the international conferences [6, 9].

In FY2019, we newly conducted the simulation with the hybrid kinetic electron model using the GT5D code [Y. Idomura et al, CPC, 2008]. In the hybrid kinetic electron model proposed in [Y. Idomura, J. Compute. Phys. 313, 511 (2016)], the numerical filter is applied to the poloidally asymmetric components of electrostatic potential so as to avoid the numerical instability caused by the passing electrons, and thus we cannot investigate the impact of poloidal asymmetries with this model. To handle the poloidal asymmetry under the presence of kinetic electrons, we implemented in GT5D the new hybrid electron model proposed by Lanti [E. Lanti, CPC, 2019], where the numerical instabilities due to passing electrons are avoided by the averaging operation on the passing electron density without applying the numerical filter to the electrostatic potential. The comparison of simulations with the electron models described above highlight the impact of poloidal asymmetries under the presence of kinetic electrons.

In line with the study without kinetic electrons [Y. Asahi et al, PPCF, 2019], we compared the collisional electron heat transport level with and without poloidal asymmetries for the turbulence driven by trapped electron mode (TEM). It should be noted that TEM can be modeled only with the kinetic electron model. Figure 4 shows the spatio-temporal evolutions of the collisional electron heat transport with and without poloidal asymmetries. As shown, the collisional transport has been reduced by a factor of about 2 or more without poloidal asymmetries. This result is consistent with the previous result obtained by the GYSELA code in the adiabatic electron limit [Y. Asahi et al, PPCF, 2019]. We will further investigate the detail in FY2020. This work is presented in the international conference [7].

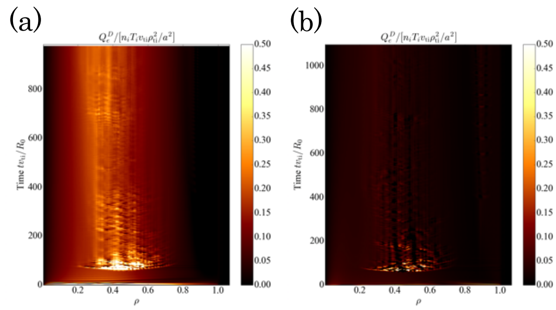

Figure 4. Spatio temporal evolutions of the collisional heat transport with (a) Lanti's model and (b) Idomura's model. Lanti's (Idomura's) model represents the simulation with (without) polodial asymmetries.

6. Progress during FY2019 and Future Prospects

High performance computing

In Q1-Q3, we have completed the intra-node performance implementation of GYSELA mini-app with OpenMP4.5, OpenACC, and Kokkos.

In Q4, we have parallelized the mini-app with MPI and updated the interpolation algorithm from Lagrange to Spline. As a preparation for using the hierarchy grids, we employed the Unbalanced Recursive Bisection (URB) algorithm which gives the similar communication patterns as those with the hierarchy grids. Table 1 shows the obtained performance with the problem size of 128^4, parallelized with 8 GPUs (2 nodes) on TSUBAME3.0. We use the CUDA-Aware MPI for MPI communication.

Table 1 The performance of the MPI version of mini-app on GPUs

Kernel	Elapsed time [s]
Pack & boundary	0.042
Unpack	0.0042
Halo Comm	1.23
Advection 2D	0.023
Advection 4D	0.06
Poisson	0.797
Spline xy	0.0597
Spline vxvy	0.05209

As easily found, the communication costs (halo communication costs in Halo Comm and all reduce communication costs in Poisson) are much larger than computation costs by the significant amount. Although the code gives the correct result, something is wrong with the communications called in-between the Kokkos::initialize and Kokkos::finalize. It may be coming from the inconsistent setting of Kokkos arguments with MPI parallelization. We will resolve this issue in FY2020.

Plasma physics

In FY2019, Q1 and Q2, we have completed the implementation and initial tests of the hybrid electron model in GYSELA.

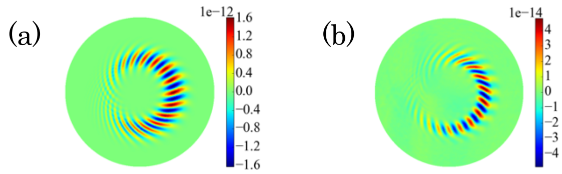

Figure 5. Linear simulations of (a) Ion temperature gradient (ITG) mode and (b) trapped electron mode (TEM). The ITG mode shows the tilting into the upper direction, while the TEM shows the tilting into the lower direction.

Fig. 5 shows the global eigen structure of linear ITG mode (a) and TEM (b). These mode structure shows the appropriate tilting directions for the ITG mode and TEM.

In Q3 and Q4, we have performed the linear benchmark of GYSELA code against GT5D code, where both codes employ the equivalent kinetic electron model in the linear limit. We find pretty good agreement in the eigen frequency but find discrepancy in the growth rate particularly in the TEM region. We will investigate the cause of the difference in FY2020.

International collaboration

From 26/June to 6/July, the project representative Yuuichi Asahi visited CEA France to discuss the plasma physics part with Xavier Garbet and the HPC part with Guillaume Latu. He also took part in an international conference [8] to present the results in JHPCN 2018 project.

With Xavier Garbet, we have intensively discussed the theorical analysis of phase space dynamics in plasma turbulence. We also initiated a collaboration to extract the phase space (space + velocity space) structure from the time series of 5D distribution function using dimensionality reduction techniques. We will continue this collaboration in FY2020.

In addition to the face-to-face meeting, we had visio-meetings dedicated to physics and HPC for every two months in FY2019. We will continue to hold visio-meetings in FY2020.

7. List of Publications and Presentations

(1) Journal Papers (Refereed)

[1] Yuuichi Asahi, Guillaume Latu, Julien Bigot, Shinya Maeyama, Virginie Grandgirard, Yasuhiro Idomura, “Overlapping communications in gyrokinetic codes on accelerator-based platforms”, Concurrency and Computation Practice and Experience, Vol. 32, e5551 (2020)

(2) Proceedings of International Conferences (Refereed)

[2] Yuuichi Asahi, Guillaume Latu, Virginie Grandgirard, Julien Bigot, “Performance portable implementation of a kinetic plasma simulation mini-app”, WACCPD 2019 (SC19), November, Denver, US

(3) International conference Papers (Non-refereed)

[3] Virginie Grandgirard, Yuuichi Asahi, Julien Bigot, et al, “How to prepare the GYSELA gyrokinetic code to future exascale Edge-Core simulations”, to be presented in PASC21, The plat-form for Advanced Scientific Computing Conference, June 2021, Geneva, Swiss. (postponed by the effect of Covid-19.)

[4] Yuuichi Asahi, Guillaume Latu, Virginie Grandgirard, Julien Bigot, “Accelerating a kinetic plasma simulation mini-app keeping performance portability”, 5th US-Japan Joint Institute for Fusion Theory (JIFT) Exascale Computing Collaboration, October 2019, Kobe, Japan

[5] Shinya Maeyama, “Current status of GKV code”, 5th US-Japan Joint Institute for Fusion Theory (JIFT) Exascale Computing Collaboration, October 2019, Kobe, Japan (to be presented)

[6] Y. Sarazin, P. Donnel, Y. Asahi, E. Caschera, G. Dif-Pradalier, X. Garbet, Ph. Ghendrih, C. Gillet, V. Grandirard, G. Latu, C. Passeron, L. Vermare, “Impact of Asymmetries and Anisotropy on Transport”, 17th EFTC Conference, September 2019, Thessaloniki, Greece

[7] Yuuichi. Asahi, Yasuhiro Idomura, Masatoshi, “Impact of poloidal convective cells on transport processes with kinetic electrons”, 16th Technical Meeting on Energetic Particles in Magnetic Confinement Systems – Theory of Plasma Instabilities, September 2019, Shizuoka, Japan

[8] Yuuichi. Asahi, Guillaume Latu, Virginie Grandgirard, Julien Bigot, “Accelerating a plasma simulation code with portable frameworks: OpenACC and Kokkos”, OpenACC annual meeting, September 2019, Kobe, Japan

[9] Yuuichi Asahi, Virginie Grandgirard, Yasuhiro Idomura, Guillaume Latu, Yanick Sarazin, Guilhem Dif-Pradalier,

Xavier Garbet, Peter Donnel, "Turbulence/Neoclassical interaction through poloidal convective cells", 10th Festival de Theorie, July 2019, Aix-en-Provence, France

[10] _Yuuichi Asahi_, Takuya Ina, Yasuhiro Idomura, _Shinya Maeyama_, _Tomohiko Watanabe_, "MPI+CUDA/OpenACC in plasma simulation codes", 19th The platform for Advanced Scientific Computing (PASC) Conference, June 2019, Zurich, Swiss

[11] _Yuuichi Asahi_, Takuya Ina, Yasuhiro Idomura, _Shinya Maeyama_, _Tomohiko Watanabe_, "Accelerating Eulerian gyrokinetic codes on the latest GPU platforms", 19th The platform for Advanced Scientific Computing (PASC) Conference, June 2019, Zurich, Swiss

(4) Presentations at domestic conference (Non-refereed)

(5) Other (patents, press releases, books and so on)