Thesis subject

Laboratory: Laboratory of the Physics of molecular and ionic interactions
PIIM UMR 7345 AMU CNRS – http://piim.univ-amu.fr

Thesis supervisor : David ZARZOSO-FERNANDEZ (david.ZARZOSO-FERNANDEZ@univ-amu.fr)

Co-supervisor :

Title of the thesis subject : Deep learning and artificial intelligence for numerical treatment and physics analysis of kinetic data: towards exascale fusion plasma simulations.

Description of the thesis subject:

Nuclear fusion aims at producing on Earth the energy of the stars, by confining the fuel (called plasma). However, a fusion plasma is an extremely complex nonlinear system, characterised by instabilities developing on disparate spatiotemporal scales, ranging from the electron Larmor radius to the size of the tokamak and from the microsecond to the confinement time (of the order of several seconds), which can lead in nonlinear regimes to turbulent transport. It is well-known that turbulence can limit the performance of fusion devices. Therefore, understanding, predicting and controlling turbulence and the induced transport and losses of particles is of prime importance for nuclear fusion and represents an extremely challenging research activity that is crucial for the ITER project, under construction at Cadarache.

In this context, the numerical simulations are a precious tool to support experiments on ITER. Indeed, current machines are not able to produce plasma scenarios equivalent to those that we will find in ITER. Therefore, the optimization, performance and risk minimization in each ITER scenario must be validated numerically. The international progress regarding the fusion performance relies on our ability to understand, predict and control the turbulent transport and confinement properties of the plasma, including thermal and energetic particles, and the transport of particles from the core to the edge. Numerical simulations are an essential tool to numerical validate the optimization, performance and risk minimization in each ITER scenario.

In tokamak, plasmas are characterized by low collisionality regimes, so that conventional fluid models are questionable and kinetic descriptions are more appropriate. In such kinetic descriptions of plasmas, the six dimensional evolution equation for the distribution function -Vlasov or Fokker-Planck equations- is solved for each species, coupled to the self-consistent equations for the electromagnetic fields, namely Maxwell's equations. Fortunately, as far as turbulent fluctuations are concerned, they develop at much lower typical frequencies than the high frequency cyclotron motion. Therefore, this 6D problem (3D in space and 3D in velocity) can be reduced to a 5D (3D in space and 2D in velocity), known as the gyrokinetic model. But even with this dimensionality reduction, the development process leading to a 5D gyrokinetic code reveals extremely challenging and requires state-of-the-art high performance computing (HPC). There exist about a dozen of gyrokinetic codes in the world, five being European.

The 5D GYSELA [1] (for GYrokinetic SEmi-LAgrangian) code is developed at IRFM/CEA for 15 years through national and international collaborations with a strong interaction between physicists,

mathematicians and computer scientists. Since 2009 a large effort has been dedicated to improving the efficiency of the parallelization (hybrid MPI/OpenMP programming), thus leading to an excellent scalability [2] (> 90% of relative efficiency at 458752 cores). The GYSELA code uses frequently from 8k to 64k cores for simulations which often run during several weeks. The annual time consumption on supercomputing facilities is currently of 150 million of core-hours. The code already benefits from petascale computational power of the largest national and European high-performance computers. Because of the multi-scale physics at play and because of the duration of discharges, we already know that ITER core-edge simulations will require exascale HPC capabilities.

The GYSELA code produces very large data sets as outputs, typically several TeraBytes for standard simulations. A typical 5D mesh is a mesh of 275 billion of points, leading to 5D distribution functions of the order of 2 TB to track at each time iteration. Given 10 000 to100 000 iterations for each simulation, it is not conceivable to store the time evolution of the 5D distribution functions. These are only saved in checkpoint files as required to deal with simulations running for several weeks. Inside a run, some embedded post-processing tools are responsible for dimensionality reduction to get quantities that are physically relevant. These predefined diagnostics (ranging from 0D to 3D) are stored at fixed time intervals. Such data can represent from 500GB up to few TeraBytes of permanent storage per simulation. The 0D to 2D quantities are post-processed with Python scripts (Fourier projections, correlation lengths, correlation times, time evolution of reduced quantities such as fluid moments, fluxes…). Regarding 3D visualization and movies, they are currently mainly performed with VISIT tool for communication purposes.

However, such large amount of 3D data fields, even under the assumption that they all can be permanently stored, is so far under-exploited, due to the fact that post-processing of high-dimensionality data is sometimes prohibitive using conventional brute-force-based techniques. In that sense, Data Science is now mature enough to provide us with new and efficient techniques which can help improve our physical interpretation of simulation data extracting the maximum information in minimum time. Also, much progress can be made regarding the analysis of data which cannot be permanently stored. Indeed, the 3D moments are saved at fixed time steps, which implies that some parts of the simulations are saved but do not contain necessary any relevant information for physics purposes. The same applies to the 5D distribution function. New tools to detect regions in phase-space with relevant events exhibiting rich underlying nonlinear physics become mandatory nowadays. Finally, the extremely expensive simulations required for ITER modelling can lead to numerical crashes, reducing therefore their validity and use for predictive purposes. New ways to detect such events in advance would be beneficial towards the optimal use of HPC resources and data storage. Within this context, a database of GYSELA simulations (~50-100 simulations) is currently under construction with Python-SQLite.

The proposed thesis will be organized following two parts:

1. Implement AI-based pattern-recognition diagnostics for 2D to 5D data recognition to detect automatically structures of the electrostatic potential and the distribution function. The goal of this task is twofold. First, we intend to automatically detect non-physical events and predict when a simulation will crash in order to stop it sufficiently in advance. This crash detection technique will be compared to the non-intrusive PoPe approach [3,4] in order to find the best way to prevent the codes from running while solving incorrectly the equations. This automatic detection will have a strong impact in terms of saved CPU time. Second, we intend to detect physical nonlinear patterns and rare events and adjust automatically the data saving modules (increase the time resolution or focus the diagnostic on a reduced region in phase-space) in

order to capture the most meaningful data in terms of physics. This diagnostic rationalization will have a positive impact on the use of memory for data saving.

2. Implement embedded dimensionality reduction, clustering techniques, Bayesian selection algorithms and missing-data inference. The goal of this task is twofold. First, it will complement the previous one in terms of memory optimization, by enabling the inference of non-stored data (a.k.a. missing data) from the permanently stored data. Second, classification algorithms will be applied to the nonlinear events detected in the previous part as well as to the trajectories of particles in GYSELA. Both nonlinear events and trajectories clusters will be linked to each other using artificial neural networks. This will lead to significant breakthrough on the understanding of the physical mechanisms in turbulent plasmas responsible for the de-confinement of particles, which is an essential question for ITER.

All these developments are part of the global strategy of the team to prepare the GYSELA code for ITER-like exascale simulations. The PhD student will be fully involved in this strategy and work in a dynamic and collaborative scientific environment. The GYSELA code is internationally recognized in the fusion community (more than 70 papers in international journals and 14 PhD theses). Concerning the mathematic and computing part, the work will be performed in strong collaboration with the CEA Cadarache, with the division of numerical methods in Plasma Physics of IPP Garching (Germany) led by E. Sonnendrücker and with our partner Maison de la Simulation at CEA Saclay (Paris).

References :

[1] V. Grandgirard et al., Computer Physics Communication 207 (2016).
[2] G. Latu et al., SBAC-PAD 2018 proceedings, https://hal.archives-ouvertes.fr/hal-01719208.
[3] T. Cartier-Michaud, Thèse Aix-Marseille (2015).
[4] T. Cartier-Michaud et al., Physics of Plasmas, 23(2), 020702 (2016).