Internship offer:

Adaptive precision Krylov methods for solving large sparse linear systems

Contact: Théo Mary (theo.mary@lip6.fr), Emmanuel Agullo (emmanuel.agullo@inria.fr), Luc Giraud (luc.giraud@inria.fr), Fabienne Jézéquel (fabienne.jezequel@lip6.fr), Pierre Jolivet (pierre.jolivet@lip6.fr)

Context:
Solving sparse linear systems is one of the fundamental problems in scientific computing. Iterative methods based on Krylov subspaces (conjugate gradient, GMRES, etc.) are capable of tackling very large problems, but require costly linear algebra operations to ensure their fast convergence: preconditioning, orthogonalization, etc. The frugal and reliable solution of large sparse linear systems is therefore one of the major current challenges in the field. This internship seeks to address this challenge by developing mixed-precision methods [1], that is, exploiting several different precision levels in order to optimize their resource consumption as much as possible. Indeed, modern computing architectures have several precisions implemented in hardware, including double (64 bits), single (32 bits), and half (16 bits) precision. Calculations performed in low precision (32 or even 16 bits) are much faster and more energy-efficient. However, most scientific computing applications require a solution accuracy equivalent to 64 bits.

Main objectives:
The main objective of the internship is therefore to develop mixed-precision Krylov methods that reduce precision only at certain well-chosen locations. We will focus on adaptive precision methods, which consist in adapting the precision of each operation dynamically at runtime, according to the data provided as input [1, section 14]. An adaptive precision method has for example been proposed for the matrix-vector product [2], which stores each matrix coefficient in a potentially different precision. Inexact Krylov methods [3] have also been proposed, which gradually reduce the precision of the operations as the iterations progress. A first objective of this internship will be to combine both approaches and perform a numerical and performance study on a range of large scale sparse problems. This internship may also tackle more exploratory avenues of research such as developing adaptive precision schemes for more complex kernels such as orthogonalization or preconditioning.

This internship involves both a mathematical component, which aims at performing theoretical error analyses to rigorously determine where precision can be reduced without impacting the final quality of the solution, and a computer science component, which aims at developing efficient implementations of these new methods to exploit modern parallel computing architectures.

Work environment:
The internship will take place in the PEQUAN team of the LIP6 laboratory (Sorbonne University, Paris), and will be carried out in close collaboration with the CONCACE team of Inria Bordeaux.

- Internship location: LIP6 laboratory (Paris), with potentially a stay or two in Bordeaux and/or Toulouse.
- Duration: 5/6 months.

- Remuneration: to be specified according to the training.
- Required knowledge:
  - Master's level Research or final year of engineering school.
  - FORTRAN/C and MPI/OpenMP programming, UNIX environment.
  - HPC, linear algebra.

Continuation with a PhD thesis:
Depending on the results of the avenues raised by this internship, a PhD thesis could be proposed in the context of the national project NUMPEX (https://numpex.org/en/). The goal of NUMPEX is to design and develop the software components that will equip the future exascale supercomputers with applications to climate, energy transition, health, AI, and industry.

References:
[1] N. Higham, T. Mary. Mixed precision algorithms in numerical linear algebra, https://hal.archives-ouvertes.fr/hal-03537373
[2] S. Graillat, F. Jézéquel, T. Mary, R. Molina. Adaptive precision sparse matrix-vector product and its application to Krylov solvers, https://hal.science/hal-03561193
[3] L. Giraud, S. Gratton and J. Langou. Convergence in backward error of relaxed GMRES, https://doi.org/10.1137/040608416