2018-00957 - Accelerating Krylov linear solvers with agnostic lossy data compression

**Contract type**: Public service fixed-term contract  
**Renewable contract**: Oui  
**Level of qualifications required**: PhD or equivalent  
**Fonction**: Post-Doctoral Research Visit

About the research centre or Inria department

An important force which has continued to drive HPC has been to focus on frontier milestones which consist in technical goals that symbolize the next stage of progress in the field. After the step of the teraflop machine in the 1990s, the HPC community envision the use of generalist petaflop supercomputers and soon coming exaflop machines in the 2020s. For application codes to sustain petaflop and more using a few millions of cores or more will be needed, regardless of processor technology. Currently, a few HPC simulation codes easily scale to this regime and major code development efforts are critical to achieve the potential of these new systems. Scaling to at this rate will involve improving physical models, mathematical modelling, super scalable algorithms that will require paying particular attention to acquisition, management and visualization of huge amounts of scientific data.

In this context, the purpose of the HiePACS project is to perform efficiently frontier simulations arising from challenging research and industrial multiscale applications. The solution of these challenging problems require a multidisciplinary approach involving applied mathematics, computational and computer sciences. In applied mathematics, it essentially involves advanced numerical schemes. In computational science, it involves massively parallel computing and the design of highly scalable algorithms and codes to be executed on future petaflop (and beyond) platforms. Through this approach, HiePACS intends to contribute to all steps that go from the design of new high-performance more scalable, robust and more accurate numerical schemes to the optimized implementations of the associated algorithms and codes on very high performance supercomputers.

**Context**

This position is open in the framework of the Joint Laboratory for Extreme Scale Computing (JLESC) within a collaboration between Inria and Argonne national laboratory. The joint project will study how lossy compression can be monitored by Krylov solvers to significantly reduce the memory footprint when solving very-large sparse linear systems. The resulting solvers will alleviate the I/O penalty paid when running large calculations using either check-point mechanisms to address resiliency or out-of-core techniques to solve huge problems.

The HiePACS team at Inria Bordeux-Sud Ouest has been studying and developing high performance linear solvers based on Krylov subspaces that are candidates for extreme scale calculation. Theoretical results exist showing that these solvers can accommodate some inexactness in the calculation without preventing the convergence at the originally prescribed accuracy [4, 5].

The extreme resilience team of the Mathematics and Computer Science division at Argonne National Lab is currently developing a comprehensive effort for lossy compression for scientific data in the context of the US Exascale Computing Project (ECP). In particular, the team has developed the SZ lossy compressor [3, 6] that achieves very high compression ratios while respecting strictly user set error controls. The team has shown empirically that SZ can be used to checkpoint some iterative solvers such as GMRES while preserving convergence.

**Assignment**

For the solution of large linear systems of the form \( Ax = b \) where \( A \in \mathbb{R}^{n \times n} \), \( x \) and \( b \in \mathbb{R}^{n} \), Krylov subspace methods are among the most commonly used iterative solvers; they are further extended to cope with extreme scale computing as one can integrate features such as communication hidden in their variants referred to as pipelined Krylov solvers [2]. On the one hand, the Krylov subspace methods such as GMRES allow some inexactness when computing the orthonormal search basis; more precisely theoretical results [4, 5] show that the matrix-vector product involved in the construction of the new search directions can be more and more inexact when the convergence towards the solution takes place. An inexact scheme of that form writes into a generalized Arnoldi equality

\[
[(A + E_1)v_1, \ldots, (A + E_k)v_k] = [v_1, \ldots, v_k, v_{k+1}]H_k. \tag{1}
\]

where the theory gives a bound on \( \|E_k\| \) that depends on the residual norm \( \|b - Ax_k\| \) at step \( k \), where \( x_k \) is the \( k \)th iterate. Such a result has a major interest in applications where the matrix is not formed explicitly, e.g., in the fast multipole (FMM) or domain decomposition (DDM) methods context, where this allows one to drastically reduce the computational effort.

One the other hand, novel agnostic lossy data compression techniques are studied to reduce the I/O footprint of large applications that have to store snapshots of the calculation, for a posteriori analysis, because they implement out-of-core calculation or for checkpointing data for resilience. Those lossy compression techniques...
allow for precise control on the error introduced by the compressor to ensure that the stored data are still meaningful for the considered application. In the context of the Krylov method, the basis \( V_{k+1} = [v_1, \ldots, v_k, v_{k+1}] \) represents the most demanding data in terms of memory footprint, so that, in a fault-tolerant or out-of-core context, storing it in a lossy form would allow for a tremendous saving.

The objective of this postdoc is to dynamically control the compression error of \( V_{k+1} \) to comply with the inexact Krylov theory. The main difficulty is to translate the known theoretical inexactness on \( E_k \) into a suited lossy compression mechanism for \( v_k \) with loss \( \| \delta v_k \| \).

**Main activities**
The successful candidate will share her/his time between Inria Bordeaux and Argonne National Laboratory to work on the activities that will follow the tentative agenda given below:

- **M0-M2 at Inria**: theoretical analysis to translate the perturbation control from \( \| E_k \| \) into a computable norm perturbation control on \( \| \delta v_k \| \) (3 months).
- **M3-M6 at Argonne**: design/tune a lossy compression technique so that the loss will be below \( \| \delta v_k \| \) (3 months).
- **M7-M9 at Inria**: implement/integrate the compression technique into a parallel out-of-core GMRES solver to evaluate the gain on large problems (4 months).
- **M10-M15 at Inria**: extend the methodology to block pipelined Krylov techniques [2] for the solution of linear systems with multiple right-hand sides [1] (6 months).

**Benefits package**
- Subsidised catering service
- Partially-reimbursed public transport

**Remuneration**
2653€ / month (before taxs)

**General Information**
- **Theme/Domain**: Distributed and High Performance Computing
- **Town/city**: Talence
- **Inria Center**: CRI Bordeaux - Sud-Ouest
- **Starting date**: 2018-11-01
- **Duration of contract**: 1 year, 4 months
- **Deadline to apply**: 2018-09-30

**Contacts**
- **Inria Team**: HIEPACS
- **Recruiter**: Giraud Luc / luc.giraud@inria.fr

**About Inria**
Inria, the French National Institute for computer science and applied mathematics, promotes “scientific excellence for technology transfer and society”. Graduates from the world’s top universities, Inria’s 2,700 employees rise to the challenges of digital sciences. With its open, agile model, Inria is able to explore original approaches with its partners in industry and academia and provide an efficient response to the multidisciplinary and application challenges of the digital transformation. Inria is the source of many innovations that add value and create jobs.

**The keys to success**
This position is intended for candidates with a strong background in computational sciences, preferably holding a PhD in applied mathematics or computer science, with some knowledge in numerical linear algebra. A knowledge/experience of parallel programming would also be much appreciated.

**Conditions for application**
Thank you to send:
- CV
- Cover letter
- List of publication

**Defence Security**
This position is likely to be situated in a restricted area (ZRR), as defined in Decree No. 2011-1425 relating to the protection of national scientific and technical potential (PPST). Authorisation to enter an area is granted by the director of the unit, following a favourable Ministerial decision, as defined in the...
decree of 3 July 2012 relating to the PPST. An unfavourable Ministerial decision in respect of a position situated in a ZRR would result in the cancellation of the appointment.

Recruitment Policy:
As part of its diversity policy, all Inria positions are accessible to people with disabilities.

Warning: you must enter your e-mail address in order to save your application to Inria. Applications must be submitted online on the Inria website. Processing of applications sent from other channels is not guaranteed.