Accelerating Krylov linear solvers with agnostic lossy data compression

Post-doctoral project - 2018/2019
HiePACS*project-team, Inria Bordeaux-Sud Ouest
Jointly with JLESC/Argonne National Lab.
Offer number: 2018-00957

July 18, 2018

1 Scientific 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 Bordeaux-Sud Ouest has been studying and developing high performance linear solvers based on Krylov subpaces that are candidate 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.

2 Objectives of the work

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

*https://team.inria.fr/hiepacs/

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

$$[(A + E_1)v_1, \ldots, (A + E_k)v_k] = [v_1, \ldots, v_k, v_{k+1}]\bar{H}_k. \quad (1)$$

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 mutipole (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 $\|\delta v_k\|$.

3 Worplan

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).

4 Background

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.

5 Supervision and starting date

The postdoc will work closely with Emmanuel Agullo and Luc Giraud on the Inria side and with Franck Cappello and Sheng Diat Argonne.

This 16 month position is planned to start on November 1st, 2018 at the latest.

In order to apply, send a CV, reference letter and the contact details of 2 or 3 academic references to cappello@mcs.anl.gov or luc.giraud@inria.fr.

References

[1] E. Agullo, L. Giraud, and Y. F. Jing. Block GMRES method with inexact breakdowns and deflated restarting. *SIAM J. Matrix Analysis and Applications*, 35, 4, p. 1625–1651, 2014.

[2] S. Cools, Y. F. Yetkin, E. Agullo, L. Giraud, and W. Vanroose Analyzing the Effect of Local Rounding Error Propagation on the Maximal Attainable Accuracy of the Pipelined Conjugate Gradient Method. *SIAM J. Matrix Analysis and Applications*, 39 (1), pp.426 – 450, 2018.

[3] S. Di and F. Cappello. Fast Error-bounded Lossy HPC Data Compression with SZ, *International Parallel and Distributed Processing Symposium (IEEE/ACM IPDPS)*, 2016.

[4] L. Giraud, S. Gratton, and J. Langou. Convergence in backward error of relaxed GMRES. *SIAM J. Scientific Computing*, 29(2):710–728, 2007.

[5] V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM Journal Scientific Computing*, 25:454–477, 2003.

[6] D. Tao, S. Di, F. Cappello. A Novel Algorithm for Significantly Improving Lossy Compression of Scientific Data Sets, *International Parallel and Distributed Processing Symposium (IEEE/ACM IPDPS)*, 2017.