Title : On the use of H-matrix arithmetic for the design of hierarchical parallel sparse linear solvers.

Responsables : Mathieu Faverge, Pierre Ramet
Téléphone : +33 5 24 57 40 39
Courriel :
Mathieu.Faverge@inria.fr
Pierre.Ramet@inria.fr

Context :

Over the past few years, parallel sparse direct solvers have made significant progress [1]. They are now able to solve efficiently real-life three-dimensional problems having in the order of several millions of equations. The ongoing hardware evolution exhibits an escalation in the number, as well as in the heterogeneity, of the computing resources.

PaStiX is a parallel sparse direct solver, based on a dynamic scheduler for modern hierarchical architectures. [2] Recently, a comparative study of the performance of the PaStiX solver over generic DAG-based schedulers has been performed. The analysis demonstrates that these runtimes provide a uniform and portable programming interface across heterogeneous environments, and are, therefore, a sustainable solution for hybrid architectures (multiple CPUs and GPUs). [3]

Nevertheless, the complexity and the need of a large amount of memory are still a bottleneck in these methods. Different solutions exists to reduce the memory requirement of these solvers with different level of impact to the numerical accuracy of the solution.

Incomplete factorization techniques are a solution to get a approximate solution of the problem. The idea consist in dropping non-zero elements of the matrix: either due to some structural dependencies within the matrix (ILU-k), or to numerical value under predefined threshold (ILU-t). Those technics usually relies on scalar implementations and thus does not benefit from superscalar effects provided by modern high performance architectures, and these methods are difficult to parallelize efficiently. We have implemented an ILU-k method that exploits the parallel blockwize algorithmic approach used in the framework of high performance sparse direct solvers in order to develop robust parallel incomplete factorization based preconditioners for iterative solvers. [4] On the numerical side, we are now studying how the data sparseness that might exist in some dense blocks appearing during the factorization can be exploited using different compression techniques based on H-matrix (and variants) arithmetics.

Some first attempts have already been investigated, in a recent work, X.S. Li and colleagues have considered the HSS-matrix representation in the context of a sparse multifrontal factorization technique to design an efficient sparse parallel direct solver for the solution of 3D Helmholtz equations. [5] The MUMPS and CHOLMOD solvers [6,7] also investigate the use of low-rank approximations. [6,7]

This research activity is conducted in the framework of the FastLA associate team and will naturally irrigate the hybrid solvers that are also investigated and will closely interact with the other research efforts where similar data sparseness might be exploited. [8]

Scientific program :

- In a first step, we will generate a Schur complement, or some top-level blocks in the elimination tree of the PaStiX solver, to check the compression ratio we could obtain using compression techniques based on H-matrix (and variants) arithmetics. Thus, we will be able to define an upper bound for memory and computation gains and select the blocksize from where the compression could be profitable. Several matrices, from different applications, will be studied in this analysis.

- In a second step, we will develop a prototype of a direct solver where H-matrix compression can be applied on larger supernodes without trying to preserve the initial tree of the compression during the updates of the factorization. This approach is clearly not optimal in terms of memory reduction, so, as first attempt, the data structures and the kernels could be duplicated in order to provide statistics but also to obtain quickly the main functionality of the solver.

- Finally, we will also investigate a way to preserve the tree of the compression all over the factorization. This requires to build a smart coupling between the H-matrix compression with the nested dissection ordering used to minimize the fill-in of the direct methods.

Keywords : Sparse linear algebra, HPC, H-matrix, low rank approximation, nested dissection, direct method.

Comments : This is a preliminary work for the PhD.

References :

[1] Anshul Gupta. Recent progress in general sparse direct solvers. Lecture Notes in Computer Science, 2073:823–840, 2001.

[2] Pascal Hénon, Pierre Ramet, and Jean Roman. PaStiX: A High-Performance Parallel Direct Solver for Sparse Symmetric Definite Systems. Parallel Computing, 28(2):301-321, January 2002.

[3] Xavier Lacoste, Pierre Ramet, Mathieu Faverge, Yamazaki Ichitaro, and Jack Dongarra. Sparse direct solvers with accelerators over DAG runtimes. Research Report RR-7972, INRIA, 2012.

[4] Pascal Hénon, Pierre Ramet, and Jean Roman. On finding approximate supernodes for an efficient ILU(k) factorization. Parallel Computing, 34:345-362, 2008.

[5] X.S. Li. Towards an optimal-order approximate sparse factorization exploiting data-sparseness in separators. Workshop Celebrating 40 Years of Nested Dissection, July 22-23, 2013, Waterloo.

[6] Patrick Amestoy, Alfredo Buttari, Guillaume Joslin, Jean-Yves L'Excellent, Mohamed Sid-Lakhdar, Clément Weisbecker, Michele Forzan, Cristian Pozza, Rémy Perrin, Valène Pellissier. Shared memory parallelism and low-rank approximation techniques applied to direct solvers in FEM simulation. To appear in IEEE Transactions on Magnetics, Extended selected short papers from Compumag 2013 conference.

[7] David S. Bindel and Jeffrey N. Chadwick. An Efficient Solver for Sparse Linear Systems Based on Rank-Structured Cholesky Factorization. Workshop Celebrating 40 Years of Nested Dissection, July 22-23, 2013, Waterloo.

[8] FastLA is an associate team between INRIA project-team HiePacs, Scientific Computing Group in the Computational Research Division in Lawrence Berkeley National Laboratory and the Institute for Computational and Mathematical Engineering and Stanford University, funded from 2012 to 2013. (see http://people.bordeaux.inria.fr/coulaud/projets/)FastLA_Website/