Proposition de thèse CIFRE EDF-LIP6-Mumps Technologies

Algorithmes de factorisation Block Low-Rank multiprécisions pour la résolution de grands systèmes linéaires avec application à des simulations industrielles d'EDF

Contexte industriel:

EDF garantit la maîtrise technique et économique de ses moyens de production, de transport et de distribution d'électricité, de la conception à la fin de vie. Les exigences de sûreté et de disponibilité nécessitent d'étayer ses décisions d'exploitation par la simulation numérique. La prédiction et l'analyse du comportement des infrastructures et des sites impliquent notamment la maîtrise de toute une gamme de modélisations en physique des champs. Pour ce faire, EDF développe ou co-développe plusieurs **codes et plate-formes de simulations** : code_aster [1a], code_carmel [1b], TELEMAC [1c], code_saturne [1d], Salomé [1e], etc.

Or **ces outils**, basés principalement sur la méthode des éléments ou des volumes finis, restent complexes et **gourmands en capacité de calcul**. Notamment du fait des consommations en temps et en mémoire d'une étape algorithmique récurrente, celle de **résolution de grands systèmes d'équations**. Dans ceux-ci on cherche à trouver le plus efficacement possible, le vecteur solution $\mathbf{u}$ d'un système matriciel du type

$$\mathbf{Ku=f} \quad (1)$$

avec $\mathbf{K}$, une matrice creuse de grande taille et $\mathbf{f}$, un vecteur, tous les deux connus.

C'est souvent cette étape d'algèbre linéaire qui **dimensionne les besoins machine** et **les temps de calcul** des simulations, à l'ingénierie comme dans la recherche, voire qui **conditionne leur faisabilité**. Suivant sa célérité et sa consommation mémoire, elle influe grandement sur le niveau de finesse et le degré de prédictivité que se permettent les ingénieurs et les chercheurs conduisant les études. Parfois, ceux-ci doivent même brider leurs besoins en modélisation, en complétude d'analyse et en richesse phénoménologique du fait de cette seule limitation.

Pour résoudre ces systèmes, ces logiciels de simulation proposent souvent à leurs utilisateurs d'avoir recours à une **méthode multifrontale**, soit en tant que méthode directe, soit en tant que préconditionneur d'une méthode itérative. Car celle-ci allie le niveau de **robustesse**, de **généricité**, de **souplesse d'utilisation** et de **précision** requis par les **contraintes logicielles et opérationnelles** de ces simulations : mélange de physiques et de modélisations, difficultés numériques intrinsèques aux méthodes d'analyses et aux méthodologies « métiers », besoin d'une brique logicielle « boite noire » automatique et adaptée à toutes les plates-formes (PC de bureau, cluster…).

De plus, cette **méthode doit aussi pouvoir traiter tous les systèmes réguliers**, sans hypothèse mathématique supplémentaire, s'adapter à de multiples besoins de résolutions (un seul système linéaire, multi-second membres, etc.) fournir des résultats complémentaires au seul vecteur solution (calcul du critère d'inertie, détection de singularité…) et rester efficace et pertinente sur une **large gamme de tailles de systèmes** (de 10^4 à 10^7 voire 10^8 inconnues).

Par contre, par rapport à d'autres méthodes, plus ciblées et moins robustes, **elle peut impliquer des coûts de calcul** plus importants et, surtout, qui **s'accroissent rapidement** (plus que linéairement) **avec la taille des systèmes** à résoudre. Et toutes les techniques usuelles du domaine (parallélisme distribué, déchargement sur disque, renumérotation…) ne permettent pas toujours de s'affranchir de cette difficulté.

Afin de lever au mieux ce verrou, de plus en plus sophistiqué et loin du cœur de métier de l'entreprise, les codes d'EDF construisent leurs briques logicielles de résolution en **s'interfaçant avec des bibliothèques externes optimisées**, telles que le solveur **MUMPS** [2]. Ce dernier est ainsi utilisé quotidiennement par des centaines d'utilisateurs R&D et Ingénierie à EDF *via* les simulations de ses codes.

Ce défi est d'autant plus prégnant que les grands défis actuels de la simulation (modèles plus prédictifs, multiéchelle, multiphysique, stochastique, jumeau numérique, IA…) requièrent des résolutions de systèmes beaucoup plus efficaces et permettant d'adresser des modèles de tailles beaucoup plus importantes. Et ce, **sans sacrifier** aux impératifs de robustesse, de précision, de périmètre et d'ergonomie déjà cités, voire, en essayant de **mieux utiliser les ressources machines** disponibles et en conservant un maximum de **portabilité**.

Contexte scientifique :

Dans la dernière décennie, deux grandes classes de méthodes ont vu le jour pour essayer de faire face à ces défis.

D'une part, des **techniques de compression** utilisant des **approximations de rang faible** ont été proposées pour exploiter la structure « data sparse » présente dans de nombreuses applications. A l'instar des formats MP3 ou pdf de notre quotidien, ces compressions cherchent à réduire les coûts calcul en réduisant les manipulations les plus coûteuses. Parmi les différents formats de compression proposés, le **format dit « Block Low Rank » (BLR)** est utilisé au sein du solveur MUMPS depuis 2010 [3]. Il est issu des thèses de Clément Weisbecker[1] (2010-2013) [4] et de Théo Mary[2] (2014-2017) [5], effectuées en collaboration avec EDF.

Cette méthode BLR consiste à identifier de nombreux blocs de rang numérique faible dans la matrice et à remplacer ces blocs par des **approximations** de rang faible **à précision contrôlée**. En stockant et en opérant directement sur la forme de rang faible de ces blocs, la méthode BLR permet de réduire fortement le coût de résolution, à la fois en temps de calcul et en consommation mémoire [6]. Qui plus est, sa complexité s'améliore asymptotiquement avec la taille de matrice : cette approche est donc particulièrement importante pour le passage à l'échelle de simulations de très grande taille [7]. La perte de précision liée à l'utilisation des blocs BLR est rigoureusement contrôlée par un paramètre numérique de seuil, noté ε [8].

D'autre part, des **algorithmes multiprécisions** sont apparus ces dernières années afin d'exploiter de **nouvelles arithmétiques** de calcul à virgule flottante potentiellement **plus efficaces**. En effet, alors que la grande majorité des codes scientifiques (dont MUMPS, code_aster, etc.) n'utilisaient jusqu'ici que les **précisions double et simple**[3], de nouvelles arithmétiques en **précision réduite** commencent à être disponibles sur les architectures de calcul modernes et offrent ainsi de nouvelles opportunités d'accélération.

Notamment, le **format demi-précision** (16 bits) **fp16** du standard IEEE est supporté par les GPUs NVIDIA depuis la génération Pascal. Il offre une accélération en temps allant jusqu'à un facteur x16[4] par rapport à la précision simple grâce à la technologie Tensor Cores [9]. Le format demi-précision **bfloat16**, proposé par Google sur ses processeurs TPUs, sera lui également disponible sur la prochaine génération des processeurs Intel (Cooper Lake).

Cependant, ces précisions réduites ne peuvent être utilisées telles quelles. A elles seules, elles ne peuvent pas fournir de solution de qualité acceptable. Pour les exploiter efficacement, il est nécessaire de **concevoir de nouvelles variantes multiprécisions des algorithmes** classiques qui combinent hautes et faibles précisions afin d'atteindre une bonne performance tout en conservant une précision acceptable : ces algorithmes ont par exemple connu un franc succès dans le cadre de la résolution de systèmes linéaires classiques [10].

Objectifs et méthodologie :

L'objectif principal de cette thèse est de **combiner les deux approches précédentes**. Il s'agit de développer des **variantes multiprécisions de la méthode BLR** qui soient **efficaces**, **fiables**, **et capables de passer à l'échelle sur des problèmes industriels difficiles**. Contrainte supplémentaire, elles doivent idéalement rester compatibles et pertinentes avec toutes les autres techniques numériques et HPC utilisées dans ce contexte : pivotage pour la stabilité numérique, parallélisme à mémoire partagée ou distribuée, appel à des bibilothèques mathématiques de bas niveau, calculs connexes (déterminant, inertie), etc.

Dans sa **forme actuelle**, la **méthode BLR** n'utilise en effet **qu'une seule précision** dont le choix est dicté par la valeur spécifiée du paramètre de seuil ε : par exemple, pour la plupart des problèmes d'EDF, la valeur la plus grande de ε qui n'impacte pas le comportement numérique de la simulation est de l'ordre de $\varepsilon=10^{-9}$. La précision machine de l'arithmétique en précision simple étant supérieure à cette valeur (de l'ordre de 10^{-8}), c'est uniquement la précision double qui est actuellement utilisée. Or, c'est souvent inutile, car la compression BLR fait justement apparaître des besoins différents dans la représentation de ces blocs compressés. **Chacun peut être stocké et manipulé avec une précision réduite adaptée à ses besoins**. L'utilisation de techniques de compression fait donc office de filtre et permet de décider quelles opérations peuvent être effectuées en précision réduite, et ce, en contrôlant l'impact sur la qualité globale de la résolution.

Ce chantier algorithmique et logiciel s'avérera aussi utile pour le portage sur GPU de certaines parties algorithmiques du code car celles-ci sont plus efficientes lorsqu'on réduit la précision de l'arithmétique employée. Il va donc être conduit en cohérence avec d'autres travaux de recherches menés par l'équipe MUMPS sur ce sujet.

---

1 Prix L.Escande 2013.

2 Prix G.Kahn 2018.

3 Les formats fp64 et fp32 définis par le standard IEEE.

4 Sur la génération à venir dite 'Ampère'.

Plus généralement et à long terme, le développement de variantes multiprécisions de la méthode BLR consistera en un triple volet visant à relever trois défis complémentaires :

Le premier volet est **informatique et algorithmique** : il est nécessaire d'adapter les algorithmes de résolution utilisant la méthode BLR pour répondre aux évolutions matérielles des architectures de calcul accompagnant l'émergence de ces arithmétiques à précisions réduites : accélérateurs avec technologies de type Tensor Cores, etc. Pour cela, outre l'introduction de précisions multiples, il semble prometteur de revoir les structures de données du format BLR, par exemple avec l'extension du format BLR à sa version multiniveau (MBLR) [11].

Le second volet est **mathématique et numérique** : il est indispensable d'accompagner les modifications algorithmiques proposées par une analyse mathématique garantissant la **fiabilité des nouveaux algorithmes**, notamment en **bornant** rigoureusement **l'erreur numérique** introduite par l'utilisation de précisions multiples. Il est aussi crucial de prévenir les risques de sous-passement et de débordement associés à l'utilisation d'arithmétiques avec des plages d'exposants bien plus étroites, comme le format fp16 (qui ne peut représenter que des nombres entre environ 0.00006 et 65500), par exemple en introduisant des stratégies de *scaling* [12].

Le troisième volet est **applicatif** : il est crucial de **valider l'efficacité et la pertinence** des méthodes proposées sur une ample gamme de **problèmes industriels**, notamment des problèmes issus des applications d'EDF en mécanique des structures et en électromagnétisme qui sont particulièrement difficiles du point de vue numérique (nécessitant par exemple des techniques de pivotage avancées). Les différentes méthodes seront jugées sur leur capacité à **passer à l'échelle** sur des problèmes industriels de grande taille **sans remettre en cause l'utilisabilité actuelle de cette brique logicielle**.

Résultats préliminaires et étapes envisagées de la thèse :

La première idée mentionnée dans la section précédente, consistant à utiliser des précisions réduites de manière sélective sur certaines entrées de la matrice et opérations, a été validée lors d'un stage de Master (débuté en mars 2020) en développant un prototype MATLAB. Les **résultats préliminaires** obtenus sont extrêmement **encourageants** : par exemple, sur une sous-partie (dite « séparateur racine ») du problème perf009d issu d'une application d'EDF [13], on estime que l'utilisation de trois précisions (fp64/fp32/fp16) réduirait la consommation mémoire du solveur d'un facteur x2.5 et le nombre d'opérations d'un facteur x3, et ce, *sans aucune perte de précision* par rapport à l'utilisation du solveur BLR (avec $\varepsilon=10^{-9}$) entièrement en précision double.

Qui plus est, **ce gain se cumulerait** naturellement avec celui provenant des **compressions BLR** : ainsi, par rapport au solveur MUMPS classique (en précision double et 'full-rank'), la méthode BLR en trois précisions pourrait théoriquement réduire la consommation mémoire d'un facteur x6 et le nombre d'opérations d'un facteur x12.

Reste bien sûr à traduire concrètement ces estimations de gains en accélération en temps et en réduction de pic RAM effectifs ! C'est précisément un des objectifs techniques de la thèse proposée ici.

En s'inspirant du retour d'expérience engrangé lors de la mise au point de ce prototype et de ses premiers résultats, on prévoit d'échelonner la thèse **en trois temps** :

- Dans la première année de thèse (***T_0 +1***), compte-tenu des premiers résultats du stage, en parallèle d'une étude bibliographique et d'une analyse mathématique plus approfondie, il semble possible de concrétiser le **gain en stockage** promis par cette stratégie.
- Dans la seconde année (***T_0 +2***), on se focalisera plus sur le **gain en nombre d'opérations** car celui-ci requiert des développements plus pointus, plus conséquents voire une analyse mathématique complémentaire. Si les résultats le permettent, cette année pourra aussi être l'occasion de **participer plus concrètement à l'activité académique internationale** en soumettant à des séminaires, à des congrès, en échangeant avec des chercheurs étrangers (séjours, chercheurs invités) voire en proposant un premier article à une revue.
- La troisième année (***T_0+3***) sera, elle, centrée sur une étape souvent négligée dans ce type de travaux, la conversion **des gains « théoriquement attendus »** (en nombre d'opérations, en consommation mémoire) en **« gains effectifs »** (en temps elapsed, en pic RAM) sur des simulations industrielles « jauges » ; soit probablement d'importants efforts d'optimisations et d'adaptations algorithmiques supplémentaires requis qui devraient aboutir à un prototype plus exploitable (pour des simulations tests) et plus évolutif/lisible (pour la continuité de ce chantier logiciel et algorithmique). Cette période sera aussi bien sûr dévolue à la rédaction du manuscrit de thèse, à l'organisation et à la préparation de la soutenance, ainsi qu'à la continuation des activités de publications et de présentations.

Par ailleurs, à chaque étape, ces **maquettes et prototypes** essaieront d'estimer **leurs apports ou leurs difficultés**, non seulement sur des cas académiques, mais aussi sur des problèmes issus **d'études industrielles EDF**. Enfin, ces travaux se focaliseront d'abord sur les **deux précisions actuellement disponible**s en standard : la simple et la double. L'introduction d'une troisième précision (par exemple, la demi-précision) s'effectuera en accord avec les évolutions matérielles et les éventuelles opportunités qui pourraient émerger en cours de thèse.

Encadrement, environnement de travail, collaborations internationales :

La thèse se déroulera entre les sites d'EDF Lab Paris-Saclay et du laboratoire d'Informatique de Paris 6 (LIP6).

D'un point de vue académique, la thèse sera supervisée par Théo Mary et Fabienne Jézéquel (HDR), (LIP6) et, d'un **point de vue industriel**, par Olivier Boiteau (EDF R&D/PERICLES). Bien sûr, elle s'effectuera en étroite collaboration avec **l'équipe de développement du solveur MUMPS** (Patrick Amestoy et Jean-Yves L'Excellent, Mumps Technologies SAS) et Alfredo Buttari (CNRS-IRIT). Pour ce faire, une collaboration avec Mumps Technologies SAS et des déplacements ponctuels dans ses locaux à l'ENS Lyon sont prévus (voir lettre de recommandation).

Par ailleurs, la thèse pourra faire l'objet de plusieurs **collaborations internationales**, notamment avec **l'Université de Manchester**, UK (groupe de Nicholas Higham) et avec les différents acteurs présents sur la côte Ouest des USA (**LBNL, Berkeley** et **ANSYS/LST, Livermore**). Des mobilités, dont une longue (1 à 3 mois) et quelques autres plus courtes (1 ou 2 semaines), sont donc à prévoir afin de faire fructifier ces collaborations.

Coordonnées des intervenants :

Encadrement industriel :
Olivier BOITEAU (olivier.boiteau@edf.fr)
EDF Lab Paris-Saclay, Département PERICLES-I23
7 bd Gaspard Monge ; 91120 PALAISEAU
Support administratif : Alexei MIKCHEVITCH (alexei.mikchevitch@edf.fr), chef du projet P-QUASI hébergeant la thèse à EDF.

Encadrements académiques :
Théo MARY (theo.mary@lip6.fr)
Fabienne JEZEQUEL (Fabienne.Jezequel@lip6.fr)
4 place Jussieu, F-75252 Paris Cedex 05

Collaboration et participation à l'encadrement
Patrick AMESTOY (patrick.amestoy@mumps-tech.com)
MUMPS Technologies (Incubateur ENS Lyon)
46, allée d'Italie ; 69364 Lyon Cedex 07

Références :

[1a] http://www.code-aster.org
[1b] http://code-carmel.univ-lille1.fr
[1c] http://opentelemac.org
[1d] http://code-saturne.org
[1e] http://salome-platform.org
[2] http://www.mumps-solver.org
[3] Patrick Amestoy, Cleve Ashcraft, Olivier Boiteau, Alfredo Buttari, Jean-Yves L'Excellent and Clément Weisbecker. *Improving Multifrontal Methods by means of Block Low-Rank Representations.* SIAM J. Sci. Comput., 37(3), A1451–A1474 (2015).
[4] Clément Weisbecker. *Improving multifrontal solvers by means of algebraic Block Low-Rank representations.* Ph.D. thesis, Institut National Polytechnique de Toulouse (2013).
[5] Theo Mary. *Block Low-Rank multifrontal solvers: complexity, performance, and scalability.* Ph.D. thesis, Université Toulouse Paul Sabatier (2017).
[6] Patrick Amestoy, Alfredo Buttari, Jean-Yves L'Excellent & Theo Mary. *Performance and Scalability of the Block Low-Rank Multifrontal Factorization on Multicore Architectures.* ACM Trans. Math. Softw., 45(1), 2:1–2:26 (2019).
[7] Patrick Amestoy, Alfredo Buttari, Jean-Yves L'Excellent & Theo Mary. *On the Complexity of the Block Low-Rank Multifrontal Factorization.* SIAM J. Sci. Comput., 39(4), A1710–A1740 (2017).
[8] Nicholas Higham & Theo Mary. *Solving Block Low-Rank Linear Systems by LU Factorization is Numerically Stable.* MIMS EPrint 2019.15.
[9] https://devblogs.nvidia.com/nvidia-ampere-architecture-in-depth/
[10] Pierre Blanchard, Nicholas Higham, Florent Lopez, Theo Mary & Srikara Pranesh. *Mixed Precision Block Fused Multiply-Add: Error Analysis and Application to GPU Tensor Cores.* SIAM J. Sci. Comput., 42(3), C124–C141 (2020).
[11] Patrick Amestoy, Alfredo Buttari, Jean-Yves L'Excellent & Theo Mary. *Bridging the Gap between Flat and Hierarchical Low-*

Rank Matrix Formats: the Multilevel Block Low-Rank Format. SIAM J. Sci. Comput., 41(3), A1414–A1442 (2019).

[12] Nicholas Higham, Srikara Pranesh & Mawussi Zounon. *Squeezing a Matrix Into Half Precision, with an Application to Solving Linear Systems*. SIAM J. Sci. Comput., 41(4), A2536–A2551 (2019).