Parallélisation et distribution de H2D2 : Speed up et choix de configuration

Rapport de recherche R-1146

mars 2010

PARALLELISATION ET DISTRIBUTION DE H2D2 : SPEED UP ET CHOIX DE CONFIGURATION

par

Yves SECRETAN
Dikra KHEDHAOUIRIA

Institut National de la Recherche Scientifique, INRS-ETE
Quebec (Quebec), Canada
G1K 9A9

Présenté à
Environnement Canada

Rapport de recherche NºR-1146

Mars 2010

Pour fins de citations:
Secretan, Y et Khedhaouiria, D, (2010).
Rapport de recherche INRS-ETE R-1146, 13pp

Tables de matières

Liste des figures

Introduction

Il est ici réalisé des tests de parallélisme effectués en mémoire distribuée (librairie MPI) et en mémoire partagée (directive OMP) appliqués à un tronçon Grondines - Île-aux-Grues du fleuve St Laurent. Le but est d'observer les meilleures combinaisons de configuration d'un point de vue rapidité de résolution tout en s'assurant d'une cohérence entre les résultats. Il sera également vu l'influence de la taille du maillage sur la répartition des temps de calculs, pour se faire un maillage avec 8 fois plus d'inconnu sera utilisé.

Ces tests viennent compléter le travail exposé dans le rapport de recherche R-1052 (MATTE et SECRETAN, 2009). La différence réside dans l'utilisation du cluster de calcul Lachine dont les capacités sont supérieures.

1. Tests de parallélisme sous Lachine

1.1. Objectifs

- Déterminer les speed-up pour différentes configurations MPI ;
- Déterminer les speed-up pour différentes configurations OMP ;
- Observer l'impact de la taille du maillage sur la durée des itérations de Newton ;
- Évaluer les configurations optimums et les seuils d'efficacité ;
- Comparer les tests ci-présents à ceux présentés dans le rapport R-1052.

1.2. Protocole

Le maillage est composé de 95 547 nœuds et de 46 174 éléments avec 215 695 inconnus. Le problème à simuler est le suivant : SV2D_Conservatif_CDYS_NN, avec un choix de résolution sur 2 heures et un critère d'arrêt (norme maximum) en erreur absolue=10^{-20}. De cette façon, le système ne converge pas et la comparaison des résultats suivant les différentes configurations devient cohérente.

En effet, la même charge de calcul est assurée à chaque pas de temps et pour chaque configuration. Un schéma d'Euler implicite avec un pas de temps de 5 minutes est choisi pour la résolution temporelle. L'algorithme de Newton est utilisé avec 8 itérations, pour un total de 24 itérations.

Afin de voir l'influence de la taille du maillage sur la durée moyenne d'une itération de Newton, un maillage plus raffiné est utilisé. Ce dernier représente le tronçon Trois Rivières – Québec et est composé de 655 313 nœuds et de 320 946 éléments, soit 6 fois plus grand et 2 fois plus dense que le précédant. Le nombre d'inconnu passe alors à 1.5 millions.

Par soucis de clarté, les deux maillages seront nommés Petit Maillage et Grand Maillage dans la suite du rapport. Le Grand Maillage ne sera pris en compte que dans la partie 1.3.3.

Le choix du solveur matriciel est le solveur distribué MUMPS. Les simulations sont faites sur le cluster Lachine composé de 8 nœuds de calculs à 2 processeurs 4-core, AMD Opteron et de 8GB de mémoire vive. Cependant, le travail se fait de façon locale sur un nœud de calcul afin de ne pas introduire les temps de réseau. Le compilateur de travail est Intel 11.1.

Les simulations sont réalisées avec les configurations suivantes :

- MPI=1 & OMP=1 ;
- MPI=1 & OMP=2 ;
- MPI=1 & OMP=3 ;
- MPI=1 & OMP=4 ;
- MPI=2 & OMP=1 ;
- MPI=2 & OMP=2;
- MPI=2 & OMP=3;
- MPI=3 & OMP=1;
- MPI=3 & OMP=2;
- MPI=3 & OMP=3 ;
- MPI=4 & OMP=1;
- MPI=4 & OMP=2;
- MPI=4 & OMP=3;
- MPI=5 & OMP=1;
- MPI=5 & OMP=2;
- MPI=5 & OMP=3;
- MPI=8 & OMP= 1;
- MPI=14 & OMP= 1[1]

1.3. Résultats

1.3.1. Temps de calculs en mémoire partagée

La figure 1 illustre les temps de calcul en configuration OMP sur un process. Une diminution du temps de calcul total est observé lorsque le nombre de tâches OMP augmente et ce jusqu'à 4 tâches.

[1] Cette configuration est utilisée uniquement pour le Grand Maillage.

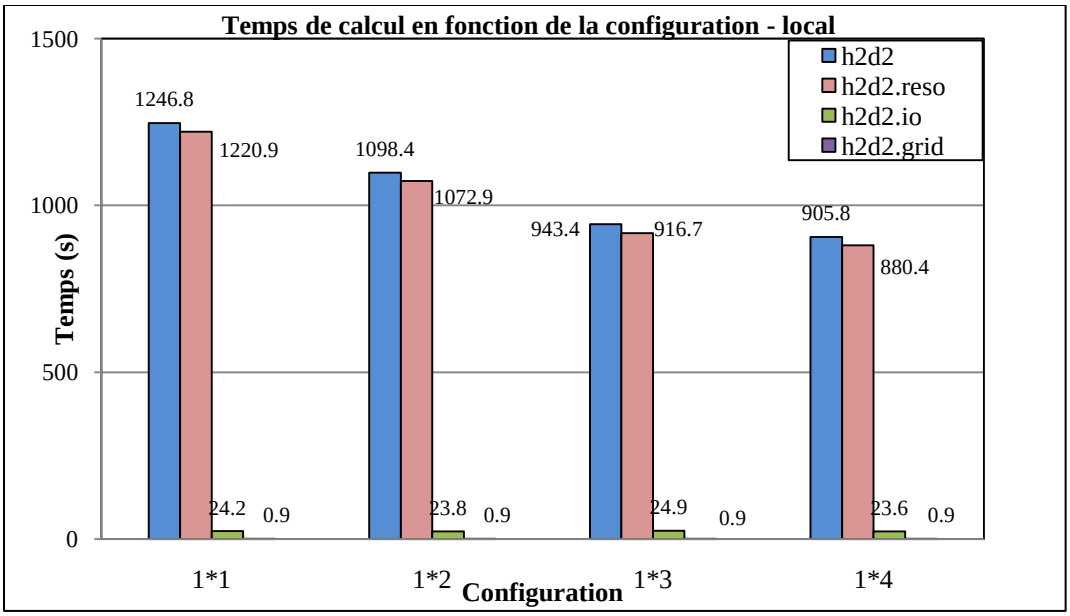

Figure 1 : Temps consacré au calcul total (h2d2), à la résolution (h2d2.reso), à l'écriture sur les disques (h2d2.io) et à la construction du maillage (h2d2.grid) pour les configurations multitâches sous MUMPS

Le temps attribué à l'échange d'information et à la construction du maillage sont faibles et constants en fonction des configurations en mémoire partagée.

Le solveur utilisé, MUMPS, est plus spécifique au travail en mémoire distribué. Pour le calcul en mémoire partagé, d'autre méthode comme Pardiso ou SuperLU sont plus efficaces[2]. Cependant, les résultats obtenus en termes de temps d'assemblage sont satisfaisant. En effet, entre la configuration 1x1 et 1x4, le temps d'assemblage est divisé par trois environ.

[2] Rapport R-1052

1.3.2. Temps de calcul en mémoire distribuée

La figure 2 illustre les temps de calcul des simulations, uniquement en mémoire distribuée.

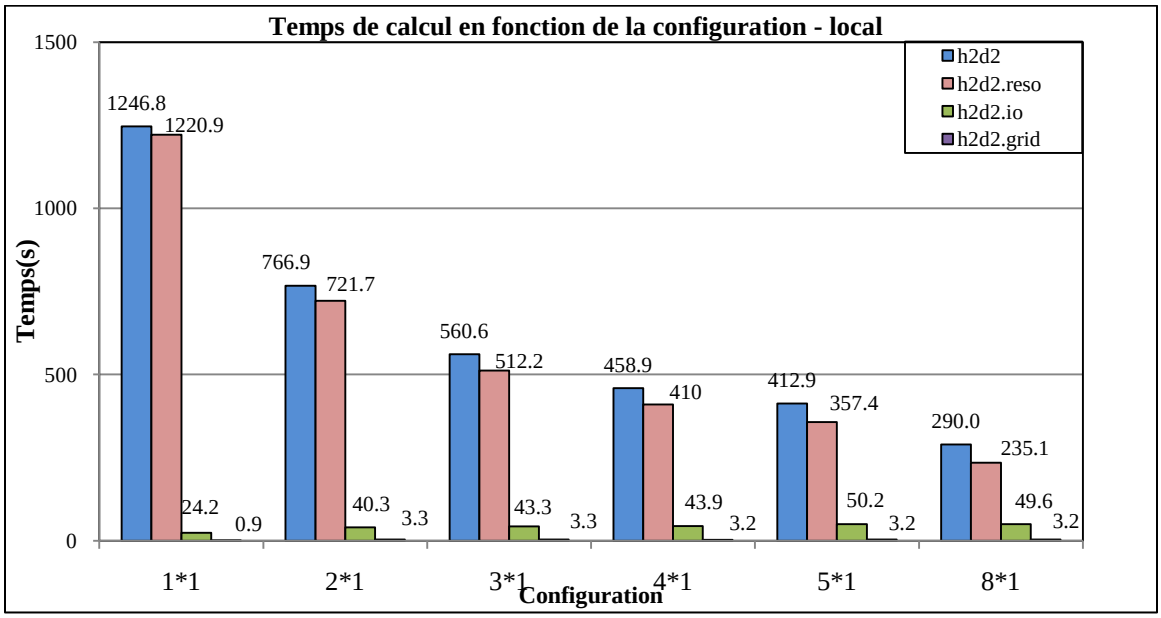

Figure 2 : Temps consacré au calcul total (h2d2), à la résolution (h2d2.reso), à l'écriture sur les disques (h2d2.io) et à la construction et au partitionnement du maillage (h2d2.grid) pour les configurations multiprocess sous MUMPS

Il apparaît clairement que les temps de calcul sont bien inférieurs à ceux énoncés précédemment en mémoire partagée. Plus le nombre de processeurs est important plus le temps de calcul diminue. En effet, avec un travail sur 5 processus les temps de calculs sont divisés par 3 et sont de l'ordre de 410 secondes.

Par ailleurs, les simulations sur 8 nœuds sont plus rapides encore même si un léger ralentissement est observé à partir de la configuration 4x1.

Les temps d'entrée/sortie (IO) et le partitionnement du maillage se stabilisent à partir d'un travail sur 2 processeurs.

1.3.3. Impact de la taille du maillage

a) Temps de calcul sur le Grand Maillage

L'impact de la taille du maillage est évalué avec les configurations 4x2, 8x1, 14x1. La figure 3 suivante illustre les temps de calcul de simulation.

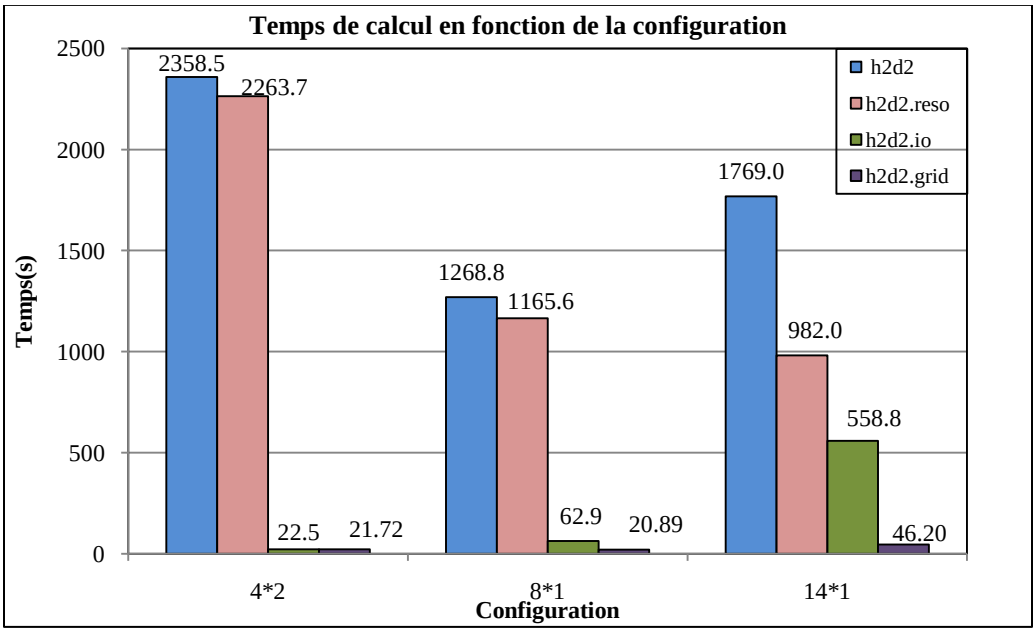

Figure 3 : Temps consacré au calcul total (h2d2), à la résolution (h2d2.reso), à l'écriture sur les disques (h2d2.io) et à la construction et au partitionnement du maillage (h2d2.grid) pour le Grand Maillage.

La simulation avec la configuration 8x1 permet de diviser par deux le temps total de calcul par rapport à la configuration 4x2.

Par ailleurs, pour les deux premières configurations (4x2 et 8x1), c'est la résolution qui impacte le plus temps total de calcul. Elle occupe 95% du temps à chaque fois.

Pour la configuration 14x1, lancée sur 2 nœuds de calcul, le temps de résolution est meilleur que pour la configuration 8x1. Les temps d'échange réseau commencent à être visibles car la diminution des temps de résolution est moins rapide. Le temps total, lui, reste plus important que pour la configuration 8x1, causé par un temps d'écriture sur les disques supérieur aux autres configurations. En effet, il occupe 30% du temps contre environ 5% pour les autres configurations.

Tous ces test ont été effectués avec des fichiers ASCII, le process maître lisant les données et les distribuant ensuite aux autres process. Les tests de comparaison avec la lecture de fichier binaire utilisant la librairie MPI- ROM-IO reste à faire.

b) Comparaison en fonction du maillage par rapport à la durée d'une itération

Le tableau 1 suivant récapitule les durées moyennes sur une itération de Newton, pour les deux maillages.

Petit Maillage /46 174 éléments/ 95 547 nœuds		Grand Maillage/320 946 éléments/ 655 313 nœuds	
4x2	*8x1*	*4x2*	*8x1*
1.8s	1.15s	16.7s	8.7s

Tableau 1 : Durée moyenne en seconde pour les deux maillages sur une itération de Newton.

Ainsi avec un maillage deux fois plus dense et un nombre d'équation 7 fois plus grand, les durées sont multipliées par 9 pour la configuration 4x2 et par 7 pour la configuration 8x1, soit une augmentation linéaire, ce qui est remarquable.

Par ailleurs, sur le Grand Maillage, le même phénomène est observé sur une itération et sur l'ensemble de la simulation. En effet, la durée sur une itération est divisée par deux en passant de la configuration 4x2 à la configuration 8x1, tout comme le temps total de simulation (Figure 3).

1.3.4. Speed up, configuration OMP

La figure 4 présente les speed up calculés à partir des temps totaux en fonction du nombre de tâche OMP.

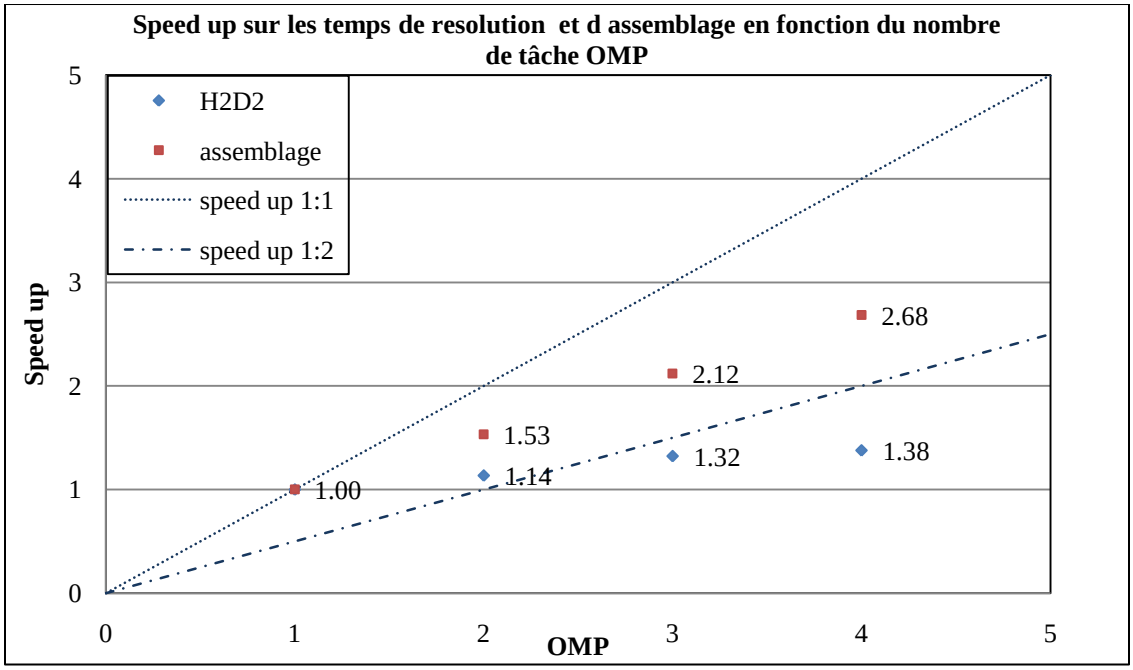

Figure 4 : Speed up calculé pour les temps totaux de résolution et à l'assemblage en fonction du nombre de tâches OMP

Les speed up pour l'assemblage atteignent la valeur de 1,53 pour 2 tâches OMP. Au-delà, les speed up augmentent toujours mais restent de l'ordre de 1:2.

Il en est de même pour le comportement des speed up sur les temps totaux de H2D2, ces derniers sont de l'ordre de 1:2 voire inférieurs à partir de 3 tâches OMP.

Il est clair que le temps d'assemblage est minime par rapport à la résolution. En réalité, pour les configurations en mémoire partagée, la résolution occupe le temps le plus important suivit du temps solveur et du temps d'assemblage (Figure 5).

1.3.5. Speed up, configuration MPI

La figure 5 présente les speed up calculés à partir des temps totaux, d'assemblage et de résolution en fonction du nombre de processus MPI.

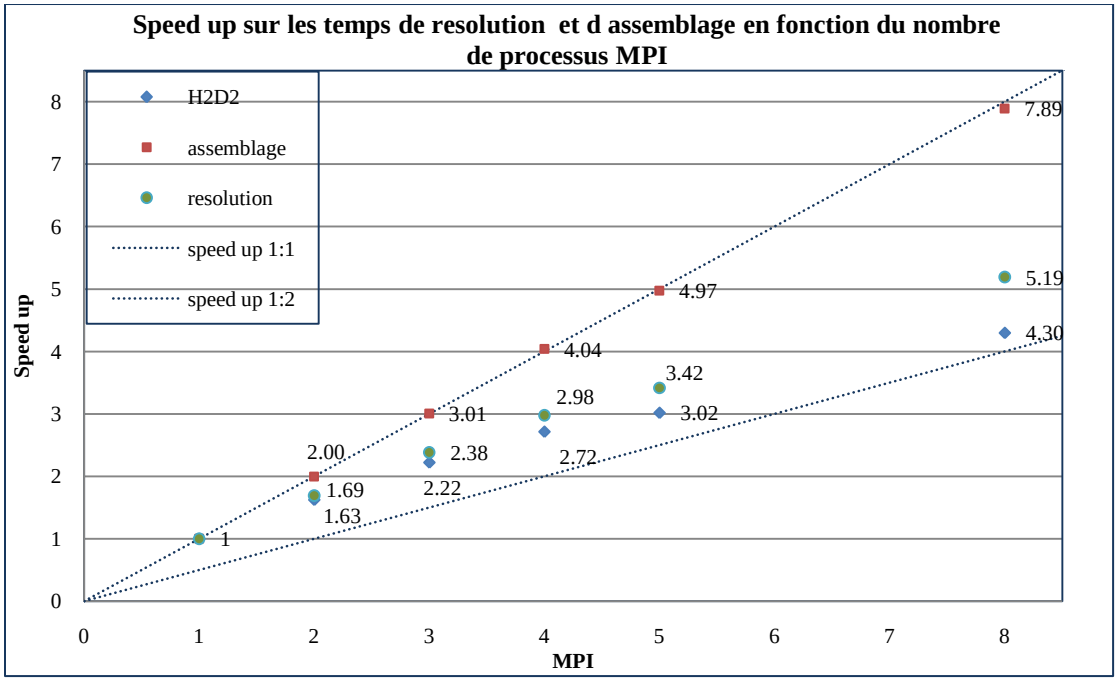

Figure 5 : Speed up calculé pour les temps de résolution totaux, à l'assemblage et à la résolution par MUMPS en fonction du nombre de processus MPI

Les speed up pour l'assemblage sont de l'ordre 1:1, le travail d'assemblage est donc bien parallélisé. Plus le nombre de processeurs est important plus l'assemblage des données est rapide.

La résolution sous MUMPS présente des speed up qui atteignent 1,69 pour 2 processus MPI. Ces derniers continuent de croître en fonction du nombre de processeurs et sont de l'ordre de 3:4 jusqu'à 5 processeurs. Au-delà, les speed up ont tendance à se rapprocher de l'ordre 1:2.

Pour le temps de calcul total sous H2D2, les speed up atteignent 1,63 pour 2 processus MPI. Au-delà, les speed up continuent de croître et restent supérieur 1:2.

Les speed-up, aussi bien pour les configurations OMP et MPI, ont la même allure que dans le rapport R-1052. En effet, les speed-up augmentent jusqu'à 2 processus MPI, ou 2 tâches OMP puis se stabilisent. Les valeurs sont, elles, différentes car les machines utilisées pour les simulations sont différentes.

1.3.6. Configuration MPI et OMP

Les temps en configuration MPI et OMP peuvent encore être améliorés en utilisant des combinaisons de ces dernières, comme l'illustre la figure 6 suivante.

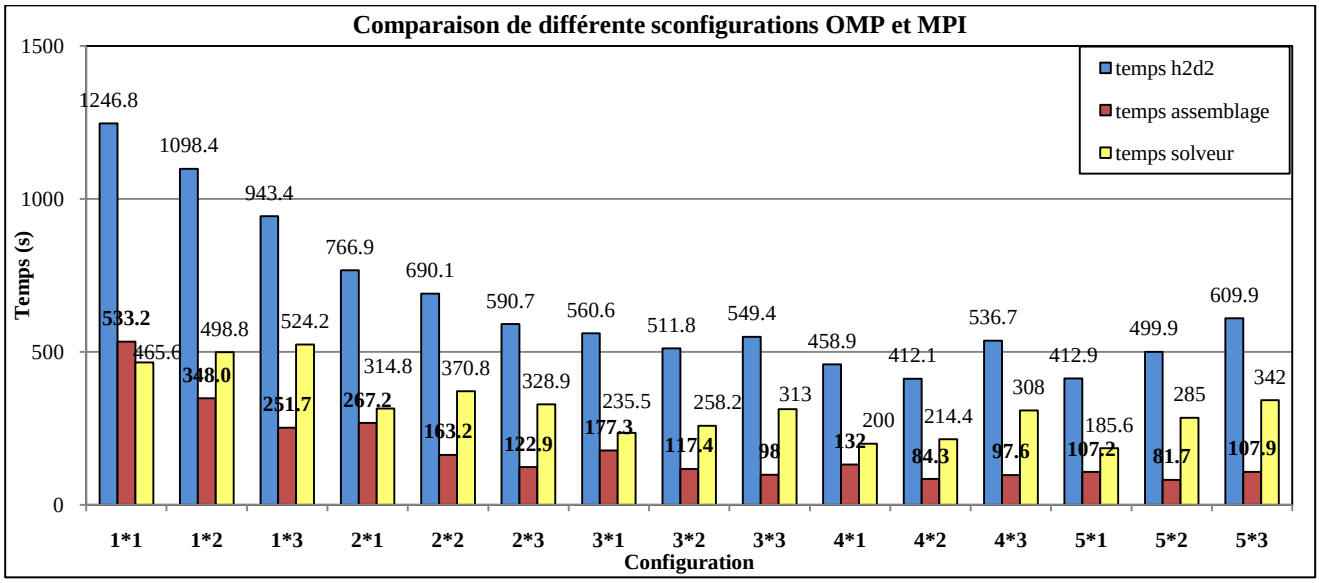

Figure 6 : Temps de calcul totaux, à l'assemblage et à la résolution pour plusieurs configurations sous MUMPS

Certaines configurations permettent de rendre le temps de calcul vraiment optimum. Ainsi il est plus intéressant de travailler en configuration 4x2 plutôt que 3x2. Pour le problème testé ici, les configurations les plus efficaces sont respectivement : 4x2, 5x1 et 5x2.

Le graphique ci-dessus montre une nette diminution du temps de calcul en augmentant le nombre de processus et le nombre de tâches jusqu'à la configuration 2x3. On observe que pour les configurations suivantes les temps de calculs totaux se stabilisent autour de 400/500 secondes.

Par ailleurs, à partir de trois processus MPI, les temps de calculs totaux deviennent plus importants pour les configurations avec 3 tâches que celles avec 1 et 2 tâches.

1.4. Validation des résultats

Afin que les comparaisons précédentes soient acceptables, il faut s'assurer que les fichiers de résultats de la simulation et la solution au problème se comportent de la même manière indépendamment de la configuration de calcul.

Dans les fichiers de résultats, la différence entre les fichiers est de l'ordre de 10^{-13}. Les comportements globaux sont les mêmes, la seul différence notable s'opère toujours au même endroit, pas de temps 2/24 itération 8/8 et pour les fichiers qui ont une combinaison OMP différente (ex : 1x1 et 1x3, 2x1 et 2x3…), erreur autour de 24% sur des chiffres de l'ordre de 10^{-14}.

Les fichiers solutions comportent les débits spécifiques et les hauteurs d'eau au niveau des nœuds du maillage. Ici encore, peu de différences sont observées entre les configurations les erreurs sont comprises entre 0 et 10^{-17}.

1.5. Comparaison à des tests antérieurs (R-1052)

Les figures 7, 8 et 9 illustrent les différences au niveau du temps total, temps d'assemblage et du temps solveur entre les simulations avec Lachine et un autre cluster testé dans le rapport R-1052(deux ordinateurs à un processeur double cœur chacun), avec le solveur MUMPS.

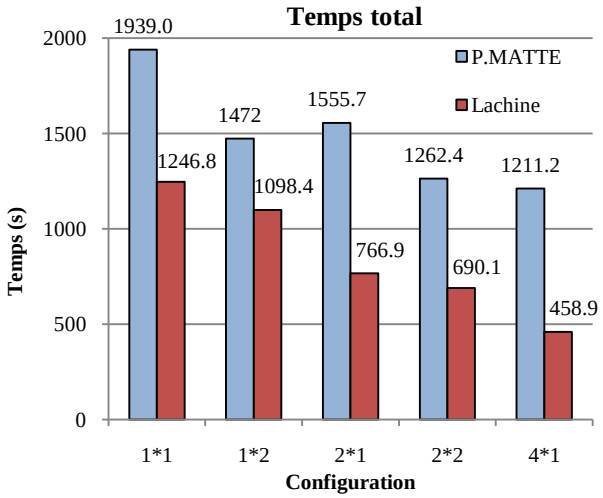

Figure 7 : Temps totaux de calcul

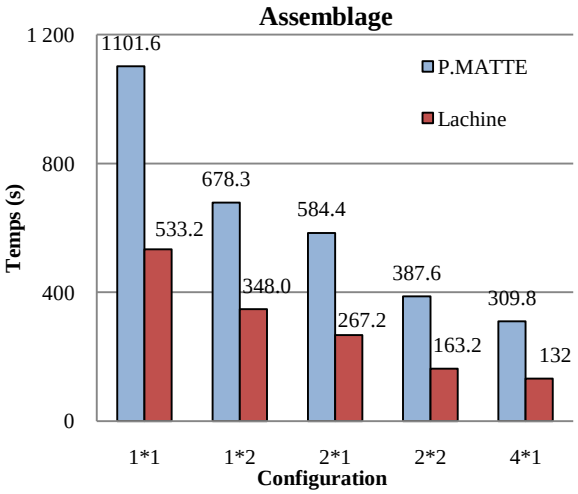

Figure 8 : Temps consacré à l'assemblage

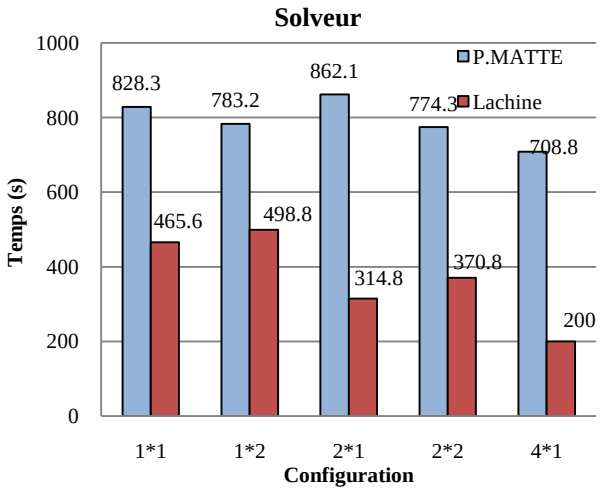

Figure 9 : Temps consacré à la résolution

Il est clair que tous les temps sont bien inférieurs avec l'utilisation de Lachine. Les gains de temps maximum sont observés au niveau de l'assemblage. De façon globale, les temps sont divisés par deux.

Conclusion

Il en ressort de ces tests, avec le solveur MUMPS, les constations suivantes :

- En mémoire partagée, les temps de calcul diminuent lorsque le nombre de tâche augmentent jusqu'à 4 tâches ;
- Les speed up en mémoire partagée sont de l'ordre de 1,14 pour 2 tâches OMP ; au-delà les speed up se stabilisent et restent supérieur à l'ordre 1:2 ;
- Les temps de calcul en mémoire distribuée sont eux inférieurs aux temps de calcul en mémoire partagée. Les temps sont en effet jusqu'à 2,5 fois inférieurs (comparaison entre les configurations 1x2 et 5x1);
- Les speed up pour l'assemblage en mémoire distribuée sont de l'ordre de 1:1 jusqu'à 8 processus, preuve que le travail d'assemblage est bien parallélisé. Les speed up, pour le temps total H2D2, sont de l'ordre de 3:4. Les speed up sont à 1.63 pour 2 processeurs et 3.02 pour 5 processeurs ;
- Ces tests montrent qu'il est plus efficace de travailler en multiprocesseur qu'en multitâche ;
- La taille du maillage a une influence sur le temps de calcul. Sur une itération, pour le passage Petit Maillage-Grand Maillage, les temps sont multipliés par 9 pour la configuration 4x2, et par 8 pour la configuration 8x2 ;
- Les temps peuvent encore être améliorés en choisissant des configurations OMP et MPI particulières. En configuration MPI et OMP, les configurations 4x2, 5x1 et 5x2 sont les plus optimales ;
- Le comportement des speed up sont comparables à ceux issus du rapport R-1052.