GROUPE DE PROJET MESONH
Compte-rendu de la réunion technique du 24 septembre 2008
Rédacteur J.-P. Chaboureau, LA

Présents : J.-P. Chaboureau, J. Escobar, D. Gazen, C. Lac, J. Payart, G. Tanguy.

1. Portage du code Méso-NH : état des lieux des moyens de calcul

Portage du code Méso-NH sur centres CNRM, IDRIS et CEPMMT

- Au CNRM. Sur la NEC-SX8R (tori), les versions 4.8, 4.7, 4.6 et 4.5 sont disponibles. Sur la NEC SX9 (yuki), ces mêmes versions sont disponibles. La SX9 calcule en théorie 4 fois plus vite que la SX8R, mais seulement 2 fois au mieux en pratique. Les options d'optimisation agressive (hopt et vopt (linéarisation des boucles i, j, k en même temps)) apporteraient un facteur 2 de performance supplémentaire, mais les calculs peuvent être faux. Aussi, l'option d'optimisation Vsafe est actuellement utilisée. Les différentes versions sont également disponibles sur les PC du CNRM et gérées par les procédures. L'installation sur le cluster est en cours.
- Aux centres du GENCI, les versions 4.7 et 4.8 sont désormais compilées par Makefile. Sur la NEC-SX8 (brodie, IDRIS), les versions 4.8, 4.7, 4.6, 4.5 sont disponibles. Sur IBM-SP6 (vargas, 3000 processeurs, IDRIS), les versions 4.7 et 4.8 sont utilisées par plusieurs personnes. Sur IBM BlueGene (babel, IDRIS), l'utilisation est actuellement déconseillée due à la limite basse en mémoire par processeur (cf. ci-dessous). Sur SGI/ICE (jade, 8000 cœurs, CINES), les versions 4.7 et 4.8 sont utilisées par plusieurs personnes. Sur BULL X (titane, 10000 processeurs, CRCT), la version 4.7 est actuellement testée.
- Au CEPMMT, les versions 4.5 et 4.6 étaient disponibles sur hpcf, qui disparaît au 30 septembre. Sur la nouvelle IBM SP6 (c1a), les versions 4.7 et 4.8 sont disponibles sous Makefile. Une demande a été faite pour augmenter l'espace disponible non purgé afin que chaque utilisateur puisse compiler le code avec suffisamment de place et sans risque de voir le répertoire de compilation détruit.

Portage sur autres machines

Le portage de MESO-NH se fait à l'aide de Makefile. Cela nécessite une adaptation du Makefile à l'architecture de la machine et au compilateur (ceux utilisés jusqu'à présent sont ifort, nag, g95 et gfortran pour les machines à base de PC).

- PC individuels (32 et 64 bits). Portage courant des versions 4.5 à 4.8
- IBM-AIX à l'université de La Réunion. La version 4.3 est disponible.
- CRAY-Opteron au CERFACS (éq. PC linux 64 bits). Méso-NH disponible.
- SGI Altix Itanium au LTHE (machine similaire au CICT). Méso-NH disponible.
- Cluster Opteron à l'INRA Bordeaux. Méso-NH disponible.
- Cluster d'Opteron au LA. Méso-NH disponible.
- MacOS10 chez V. Masson. Méso-NH disponible.

2. Evolution vers les grandes grilles

Bilan
Dans une version modifiée (notamment dans la décomposition en trois dimensions du domaine de résolution du solveur de pression), le code MESONH a été testé avec succès sur jade (CINES) jusqu'à 8192 cœurs, une grille de 4096 x 4096 x 128 pour une puissance de calcul atteignant le Téraflops. Le code montre une excellente scalabilité (autour de 100%).

Travaux en cours
- PREP_PGD appelle beaucoup de routines d'interpolation horizontale par spline. En conséquence, PREP_PGD ne tourne qu'en monoprocesseur sur des machines dotées d'une mémoire importante : la génération de grande grille ne se fait que sur vargas (jusqu'à 256 Go de mémoire). PREP_PGD fait aussi de nombreux accès en lecture (par exemple, les 200 tableaux cover sont lus 5 fois), ce qui augmente le temps de restitution de manière importante. Une action de parallélisation et d'optimisation de la librairie Surfex est projetée par le CNRM et le CERFACS.
- Dans PREP_REAL, la lecture des champs GRIB a été parallélisée, ce qui permet de gagner en mémoire. Un travail d'optimisation reste à faire pour finaliser cette amélioration. Par contre, le traitement des champs en lfi pour PREP_REAL et SPAWNING doit encore être parallélisé.
- Le programme MESONH, parallélisé depuis longtemps, nécessite une adaptation aux machines massivement parallèles. Le travail sur le solveur de pression réalisé par J.Escobar est intégrable, après un travail de mise en forme. La lecture et l'écriture d'entrée/sortie (I/O) en format LFI ne se font actuellement que sur un seul processeur, limitant les I/O à la mémoire du processeur. L'écriture des champs 3D en N_Z matrices 2D permet de dépasser la limite mémoire du processeur et pouvoir écrire avec N_Z processeurs. Une écriture des champs 3D par blocs est à envisager. Par ailleurs, la taille des fichiers LFI est restreinte à 16 Go due au passage du codage d'adresse en 32 octets (l'utilisation de 64 octets permet de s'affranchir de cette limite). Une alternative au format LFI serait les formats HDF5 et NETCDF. Mais les performances de parallélisation de l'écriture et la lecture dans ces formats sont très dépendantes des machines. Au-delà, si ces formats permettent la création d'un fichier de 500 Go, son utilisation est néanmoins problématique.
- Le programme DIAG de calculs diagnostiques est parallélisé.

3. Libtools

Les dernières versions disponibles au CNRM et à l'IDRIS sont en phase (version étiquetée LIBTOOLS-CNRM-4-8-a sur le dépôt CVS). Les tools sont des programmes monoprocesseur qui devront être adaptés pour faire face à la problématique des grandes grilles.

4. Procédures

Le lancement de MESONH sous OLIVE (l'interface web du CNRM pour le lancement d'AROME notamment) est en test. Les quelques bogues identifiés, dont la lecture du format libre des fichiers namelists (par ex les ZHAT des PRE_REAL1.nam), sont en train d'être corrigés. Olive permet un lancement répétitif ;

il a vocation à remplacer la procédure prep_experiment. Il est également prévu à terme de pouvoir lancer les cas idéalisés par OLIVE.

La commande cpio est utilisée pour regrouper et dégrouper les fichiers des et lfi (shells fm2deslfi et deslfi2fm) afin de limiter le nombre de fichiers dans la machine d'archivage. Cette commande, coûteuse en temps de calcul, ne fonctionne pas dès que la taille des fichiers dépasse 2 Go. Aussi il est nécessaire de paramétrer dans les procédures l'emploi de cette commande. A terme, ce problème pourrait être résolu par l'inclusion du fichier des dans le fichier lfi.

5. Gestionnaire de sources

Le gestionnaire de sources CVS impose un dépôt centralisé des sources. Comme la connexion réseau entre CRNM et LA est mauvaise pour des contraintes de sécurité propres à Météo-France, le partage des sources entre CNRM et LA se fait avec difficulté. Par ailleurs, l'écriture sous CVS impose l'écriture inconfortable de tags. Aussi le remplacement du gestionnaire CVS est envisagé.

Le gestionnaire Git est utilisé pour le développement de grands codes, notamment le noyau linux. Git montre plusieurs avantages, celui d'avoir des dépôts distribués mis à jour par ssh ou mail, et l'identification de la mise en consigne (commit) de manière unique pouvant se substituer aux TAG CVS. Un autre intérêt d'avoir des dépôts distribués est une gestion locale des sources pour chaque utilisateur, ce que ne permet un dépôt CVS centralisé. L'interface graphique associée Gitk est aussi plus conviviale et « intelligente » que celle de CVS, permettant notamment la visualisation complète de l'arborescence des modifications. Le gestionnaire Git est actuellement testé pour la gestion des libtools. A terme, il devrait être utilisé pour la gestion des sources du code MESONH.

6. Site WEB

Le site WEB comporte une partie historique mise à jour par Juan Escobar et Jean-Pierre Chaboureau et deux interfaces wiki (pour des raisons de sécurité), l'un pour l'équipe de développement (Team) et l'autre pour les utilisateurs extérieurs (User), permettant l'écriture et la modification de pages web de manière communautaire. Une action de hiérarchisation de l'information est en cours d'étude afin de permettre le développement du site web en totalité en wiki.

L'ensemble de ces points seront repris et exposés lors de la prochaine réunion des utilisateurs (12 et 13 octobre 2009).