WN/CMGC/08/98

Enjeu et problématique du portage d'ARPEGE-NEMO sur calculateurs super-scalaires

Eric Maisonnave

Table des matières

Le bouleversement provoqué par l'apparition des machines massivement parallèle dans le monde des super-calculateurs, jusqu'à occuper les places du top 500 jusque là réservées aux machines vectorielles, posent à la communauté des modélisateurs du climat un problème d'adaptation important.

En effet, l'ensemble des codes utilisés dans un modèle de climat ont tous été conçus dès l'origine pour le travail sur plate-forme vectorielle. Le travail que nous nous proposons d'initier sur les machines du CNRS va, dans un premier temps, nous aider à nous assurer de la viabilité de nos codes actuels sur ce type d'architecture. Il doit également nous permettre de mieux identifier les points de blocage en vue de proposer à moyen terme de nouvelles solutions algorithmique.

Ce document se propose de décrire la méthodologie qui doit nous permettre de créer le plus rapidement possible une version de notre modèle couplé capable de répondre aux demandes les plus urgentes de la communauté (prochain exercice du GIEC sur le réchauffement climatique des 50 prochaines années). En même temps, cette tâche nous donnera la possibilité d'évaluer concrètement de le degré d'adéquation des machines super-scalaires avec la simulation du climat mais aussi de mettre à disposition de la recherche en informatique et algorithmique une plate-forme de test pour les optimisations futures.

Ce travail s'inscrit dans le projet européen Is-ENES, où, à l'échelle du continent, toute la communauté des modélisateurs du climat va s'efforcer de rendre compatibles ses codes avec les nouvelles architectures de l'infrastructure PRACE. Dans ce projet, le CERFACS recevra l'aide de deux partenaires (BSC, Barcelona Supercomputing Centre et NSC, Centre Scientifique Finlandais) pour effectuer des portages similaires sur deux super-calculateurs des tiers 0 et 1.

Algorithme utilisé, adaptation à la plate-forme visée

Le modèle de climat proposé au portage, à l'optimisation et à la validation sur les plate-formes IBM BG/P de l'IDRIS et SGI ALTIX du CINES est composé de deux codes principaux d'atmosphère et d'océan. Ces deux sous-modèles échangent leurs conditions d'interface via le coupleur OASIS. Une description plus précise de chacune de ces composantes est faite au paragraphe 4.

L'adaptation de ce modèle de climat aux deux plate-formes visées constitue la première partie du projet. Le lecteur se reportera aux documents **[1]** et **[2]** pour une description des expériences de prévision du climat qui seront réalisées dans un deuxième temps lorsque le modèle aura été porté, optimisé mais aussi validé.

Car en effet, la simple compilation, puis le travail plus poussé d'optimisation (décrit au paragraphe 2) ne sera pas suffisant pour pouvoir commencer les expériences scientifiques de géophysique proprement dites.

Le type d'expérience visé dans la deuxième partie du projet s'appuie sur des 'analyses probabilistes: un jeu (ensemble) de plusieurs simulations différant légèrement par leurs conditions initiales doit nous permettre d'estimer l'erreur sur nos résultats. Les ressources d'une seule machine ne pouvant être suffisantes, plusieurs plate-formes seraient donc mises à contribution pour réaliser la totalité de l'expérience: L'effet produit par un changement hardware devrait alors être mesuré et devrait rester significativement inférieur à la dispersion due à la seule physique de notre modèle.

Des simulations de climat de plusieurs dizaines d'années sont indispensables à la mesure de cet effet, d'où le montant des heures demandées (460.000 heures au CINES, 900.000 heures à l'IDRIS).

Ces chiffres représentent 200 ans de simulation du climat, compte tenu des résultats préliminaires obtenus sur la machine IBM Blue Gene/L du CERFACS (où une heure elapsed est nécessaire sur 370 processeurs pour accomplir une simulation climatique d'un mois, voir table 1). Sur la SGI ALTIX, un ratio ½ a été appliqué compte tenu des performances atteintes par un autre code du CERFACS (ELSA) sur ces deux machines.

	NEC SX8R	***IBM BG/L***	***CRAY XD***	***Grid'5000***
BR	17 (2p)	26 (32p)	22 (8p)	25 (12p)
HR	32 (16p)	60 (370c)	non porté	non porté

Table 1: Temps de restitution minimaux (en minutes) obtenus sur 4 super-calculateurs différents pour simuler 1 mois de climat avec un modèle basse résolution (BR) ARPEGE t63-NEMO ORCA2 et un modèle haute résolution (HR) ARPEGE t159-

NEMO ORCA05. Entre parenthèses, le nombre nécessaire de ressources (processeurs vectoriels ou coeurs scalaires).

D'autres plate-formes ont également été testée dans des configurations plus légères (ARPEGE+couche de mélange NEMO sur ordinateur portable DELL Latitude) ou intégrant des composants allogènes (ECHAM sur Earth Simulator).

Particularités du portage

Tout changement d'OS implique la recompilation. Cette étape exigeante en temps et en patience pourrait néanmoins être évitée sur tout système linux autorisant l'installation du logiciel kadeploy et la mise en place de l'image grid'5000 **[3]**.

Dans le cas contraire, les options de compilation devront être choisies et définies dans chacun des trois makefiles des différentes briques du système couplé: ARPEGE/compile.options, NEMO/WORK/Makefile, IOIPSL/src/Makefile et prism/util/compile/frames/include_machine.

On s'assurera de la compatibilité de la librairie MPI Version 1. ARPEGE, OASIS et NEMO fonctionnent avec MPICH, LAMMPI et OPENMPI. On vérifier que ce MPI supporte bien le mode MPMD (des exécutables différents peuvent être lancés sur des processeurs différents).

Le coupleur OASIS possède une librairie d'IO parallèles fort utile mais pouvant poser quelques problèmes au portage. A condition de ne pas avoir à utiliser les options d'écritures de fichiers (pour contrôler les champs de couplage uniquement, les lectures de fichiers de redémarrage sont faits de toutes façons), on peut activer manuellement l'option use_key_noIO=yes dans le script de compilation des librairies OASIS (prism/util/COMP_libs.machine)

La librairie IOIPSL et le modèle NEMO doivent avoir les mêmes options de compilation. En particulier, l'option de précompilation D_P, qui détermine la taille par défaut des entiers et des réels. Il est plutôt conseillé de garder cette taille par défaut dans ARPEGE et OASIS.

Toujours en ce qui concerne la taille des entiers et réels, une ligne en tout début du fichier de log d'ARPEGE renseigne sur la validité de ceux-ci:

```
 --- Set up machine-specific constants----------------------------------
 NINTLEN=  4 NREALEN=  8 NLOGLEN=  8 NDBLLEN= 16
```

En cas d'incompatibilité avec les options de compilation, il faudra modifier le fichier ARPEGE/Sources/headers/lficom0.h

Un problème peut se poser à la lecture Fortran des fichiers non formatés. Prendre garde à l'option d'endianness. Ce problème se révèle dès l'utilisation des exécutables de préparation du fichier de redémarrage d'ARPEGE (la lecture de ces fichiers est impossible). Il existe dans la plupart des compilateurs des options permettant de traiter l'un ou ou l'autre des deux modes de codage.(**-fconvert=big-endian** ou **little-endian** pour gfortran, par exemple).

Ces conseils n'éviterons certainement pas d'avoir à ré-écrire diverses parties du code. Préférer les modifications sous clefs CPP afin de pouvoir réutiliser le code sur d'autres machines.

Modalités d'optimisation (vectorisation, optimisation super-scalaire, parallélisation)

ARPEGE

En aval de cette action de portage, un effort d'optimisation tout particulier sera fourni dans le domaine de la modélisation de l'atmosphère. La scalabilité du modèle spectral ARPEGE compte effectivement parmi les problèmes les plus intéressants soulevés par ce portage.

Les calculs parallèles de la dynamique sont limités par le nombre de fonctions de la décomposition spectrale prises en compte (troncature). Pour repousser cette limite, la parallélisation sur les fonctions spectrales peut être doublée d'une parallélisation sur la verticale. Dans le cas d'étude qui nous intéresse (troncature 159), le nombre de processeurs ne serait donc plus limité à un peu plus d'une centaine mais aux environ de 3500.

De même, en ce qui concerne les calculs de physique, une parallélisation en X doublerait celle en Y et permettrait également d'utiliser plusieurs milliers de ressources.

Dans ces deux cas, les routines développées sur IFS au Centre Européen de Prévision à Moyen Terme (ECMWF) seraient activées dans le modèle ARPEGE. Un travail non négligeable de phasage est attendu pour rendre conforme l'ensemble des routines d'ARPEGE à l'inclusion de ces nouveaux modes de parallélisation.

Des tests préliminaires effectués au CERFACS sur la machine Blue Gene/L nous confortent dans notre projet:

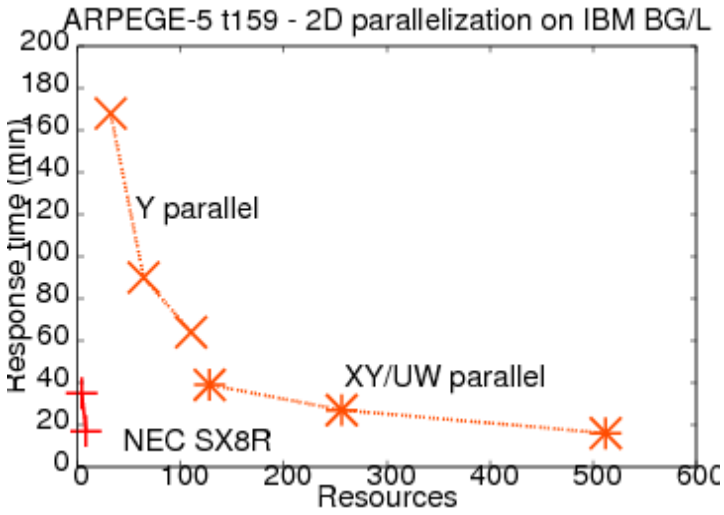

Fig 1: Comparaison des performances du modèle d'atmosphère seul ARPEGE-v5 (t159) pour une simulation d'un mois de climat sur machine vectorielle (en rouge, NEC SX8R) et super-scalaire (en orange, IBM BG/L). La deuxième partie de la courbe orange est obtenue avec les options de parallélisation en deux dimensions (latitude/longitude pour la physique, nombre d'onde/altitude pour la dynamique)

Pour aller plus loin, des essais de parallélisation hybride avec Open MP, en s'appuyant toujours sur les précédents développements de l'ECMWF, et devraient nous permettre de mieux estimer les potentialités de ce type de parallélisation pour les coeurs de calculs d'un même noeud. Cela sera particulièrement intéressant sur la machine du CINES possédant des bi-processeurs quadri-coeurs.

Côté algorithmique, différent type de transformées pourront être testées, ainsi que le mode d'itération dans le calcul des trajectoires optimales du modèle.

Pour ce qui est des entrées/sorties, qui peuvent se révéler coûteuses sur ce type de machines, plusieurs stratégies devront être mise à l'épreuve. En particulier, le calcul du nombre optimal de processeurs à impliquer dans les écritures comme dans les lectures devra être réalisé.

Enfin, une bonne estimation des coûts de calculs des différentes parties du code devra être faite afin de mieux définir les actions futures d'amélioration de ses algorithmes (physique, dynamique, code radiatif, post-processing et communications).

NEMO

La composante océanique utilisée dans notre modèle est le modèle communautaire NEMO. L'adaptation de ce modèle (en mode océanique non couplé) aux plate-formes superscalaires bénéficie du travail de toute la communauté océanographique française (LEGI, LOCEAN, IFREMER) mais aussi européenne (Met Office, Hadley Center). Il a notamment été utilisé dans sa version très haute définition au 1/12e de degré global lors de la VSR de la SGI Altix du CINES. De nombreuses interactions en géophysique comme sur un plan plus technique entre le CERFACS et les laboratoires sus-cités nous assurent que les solutions de parallélisation déjà mises en oeuvre dans des configurations océaniques forcées pourront être facilement transférées et testées dans notre mode couplé.

OASIS

Le coupleur OASIS développé au CERFACS fait l'objet d'un suivi important de son

équipe de développement en matière d'adaptation aux nouvelles plate-forme. Dans ce cadre, un test poussé de la parallélisation de ses interpolation pourra être mené. De même, une validation de sa toute nouvelle option de parallélisation « par champ de couplage » sera implémentée dans notre modèle couplé haute résolution. Les résultats de ces tests seront diffusés auprès de la nombreuse communauté des utilisateur du coupleur OASIS (quasi totalité des grands centres européen de modélisation du climat) avec un impact certain.

Structure du programme

ARPEGE

ARPEGE est l'acronyme de « Action Recherche Petite Echelle Grande Echelle ». Ce modèle est utilisé à la fois à Météo-France et à l'ECMWF (où il se nomme IFS). ARPEGE est un modèle global spectral, avec une grille Gaussienne pour les calculs en point de grille.

La discrétisation verticale est faite suivant un système de coordonnées hybride en pression. Une grille horizontale irrégulière dite « étirée » est utilisable: le changement dans la représentation horizontale est défini par un changement du pôle (pouvant alors différer du pôle Nord géographique) suivi par une transformation conforme (suivant Schmidt, 1977).

ARPEGE contient différents modèles (aux équations primitives 3D, modèle non-hydrostatique, modèle shallow water) et un schéma de post-processing interne.
L'assimilation de données est possible suivant différents schémas (interpolation optimale, 3D ou 4D-VAR). Le modèle peut être initialiser de différentes manières (mode adiabatique ou par itération).

ARPEGE/IFS peut s'appuyer sur différentes paramétrisations physiques. Une même dynamique spectrale est utilisée à la fois dans sa version Climat (portée ici) et Prévision Météorologiques.

NEMO

Le moteur océanique de NEMO (Nucleus for European Modelling of the Ocean) est un modèle aux équations primitives de la circulation océanique régionale et globale. Il se veut un outil flexible pour étudier sur un vaste spectre spatio-temporel l'océan et ses interactions avec les autres composantes du système climatique terrestre (atmosphère, glace de mer, traceurs biogéochimiques...).

Les variables pronostiques sont le champ tridimensionnel de vitesse, une hauteur de la mer linéaire ou non, la température et la salinité. La distribution des variables se fait sur une grille C d'Arakawa tridimensionnelle utilisant une coordonnée verticale z à niveaux entiers ou partiels, ou une coordonnée s, ou encore une combinaison des deux.

Différents choix sont proposés pour décrire la physique océanique, incluant notamment des physiques verticales TKE et KPP .

A travers l'infrastructure NEMO, l'océan est interfacé avec un modèle de glace de mer, des modèles biogéochimiques et de traceur passif.

OASIS

Le coupleur OASIS est constitué d'une librairie de communication, permettant de synchroniser les codes à coupler et d'échanger les champs de couplage à l'interface de ces modèles, et d'une librairie d'interpolation, permettant d'effectuer les transformations requises pour exprimer, sur la grille du code cible, les champs de couplage fournis par le code source sur sa propre grille.

Les concepts fondamentaux de modularité, flexibilité, portabilité et parallélisme sur lesquels est fondé OASIS lui ont permis d'acquérir une reconnaissance internationale ; OASIS est aujourd'hui en effet utilisé par une vingtaine de groupes de recherche en modélisation climatique en France et en Europe mais aussi aux Etats-Unis, au Canada, au Japon et en Australie.

Deux versions d'OASIS sont actuellement disponibles: OASIS3, version stable pseudo-parallèle du coupleur ne traitant que des champs 2D, et OASIS4, nouvelle version complètement parallélisée encore en cours de développement, traitant des champs 3D. Cette nouvelle version pourra être éventuellement testée au cours du portage.

Pour plus d'information sur la structure des codes, se reporter aux documentations établies par leurs développeurs:

ARPEGE
http://www.cnrm.meteo.fr/gmgec/arpege-climat/ARPCLI-V5.1/doca/arp51ca.pdf

NEMO
http://www.nemo-ocean.eu/index.php//content/download/2702/18843/file/NEMO_book.pdf

OASIS
http://www.prism.enes.org/Publications/Reports/oasis3_UserGuide_T3.pdf

Logiciels nécessaires, langages et bibliothèques utilisées

Les codes écrits en Fortran 77, 90 et C ne nécessitent pas de logiciels particuliers pour s'exécuter. Seules les bibliothèques BLAS/LAPACK (pour ARPEGE) et Netcdf sont requises à l'édition de lien.

En ce qui concerne la parallélisation, la librairie de communication est utilisée à 2 niveaux. Dans les modèles ARPEGE et NEMO pour leur parallélisation interne, mais aussi par le coupleur OASIS, qui dirige les communications inter-modèles (échanges de données aux interfaces physiques. Ici: surface de la mer).

Ce deuxième type de communications MPI rend indispensable l'utilisation du mode MPMD (Multiple Programs Multiple Data). Ce mode doit impérativement être disponible sur les plate-forme visées. Dans le cas contraire (IBM Blue Gene/L par exemple), un programme en C doit pouvoir se substituer à l'utilitaire de lancement mpirun. Ce programme doit pouvoir appeler la fonction C « execv » afin de lancer un process par coeur réservé, ce process pouvant être soit une instance du modèle d'atmosphère, soit une de celles de l'océan, ou bien encore une de celles du coupleur.

Bibliographie

[1] Dossier de demande d'aide en ressources informatique 2009, DARI, Présentation générale.
[2] Comprehensive Modelling of the Earth System for Better Climate Prediction and Projection (COMBINE), FP7 Proposal, 2008
[3] Maisonnave, E., Morel, T., and Valcke, S., 2007: Portage et déploiement du modèle couplé OCC17 sur la plateforme distribuée Grid 5000 , Working Note, **WN/CMGC/07/44**, Cerfacs, France