WN/CMGC/07/44

Portage et déploiement du modèle OCC17 sur la plateforme distribuée Grid'5000

Eric Maisonnave, Thierry Morel, Sophie Valcke

Abstract

Ce document décrit les différentes étapes du portage sur la grille de calcul Grid'5000 du tout dernier modèle couplé ARPEGE-OPA développé au Cerfacs.
Il détaille les raisons de nos choix en matière d'outils de compilation, de librairies annexes telle qu'OpenMPI, en fonction des particularités de cette architecture.
Il présente comment le coupleur OASIS3 a pu y être utilisé, et dans quelles configurations.

Table of Contents

I. Contexte

Ce travail s'inscrit dans le mouvement progressif de migration des versions vectorielles de nos modèles numériques de climat vers les architectures scalaires.

Disponible dès l'été 2005 sur un système d'exploitation Linux **[1]**, puis sur le cluster CRAY XD1 du CERFACS, notre modèle de climat « OCC17 » vient d'être porté sur la grille de cluster Grid'5000 **[2]**, dans le cadre du projet **ANR-CICG05-11** « LEGO », League for Efficient Grid Operation **[3]**.

Un environnement optimisé de production d'expérience et de gestion du flux d'entrées/sorties est actuellement en co-développement avec le laboratoire de l'informatique du parallélisme (LIP, ENS Lyon).

Les performances du modèle dans un tel contexte seront prochainement détaillées dans un document séparé de celui-ci.

I.1. La plateforme Grid'5000

La plateforme expériementale Grid'5000 rassemble environ 5000 processeurs répartis sur 9 sites en France.

La création de cette machine a été décidée dans le cadre d'une ACI (Action Concertée Initiative) Grid et poursuit ses activités grâce au Ministère de la Recherche, l'INRIA, le CNRS et les universités des différents sites.

17 laboratoires, dont le Cerfacs, sont impliqués dans le projet.

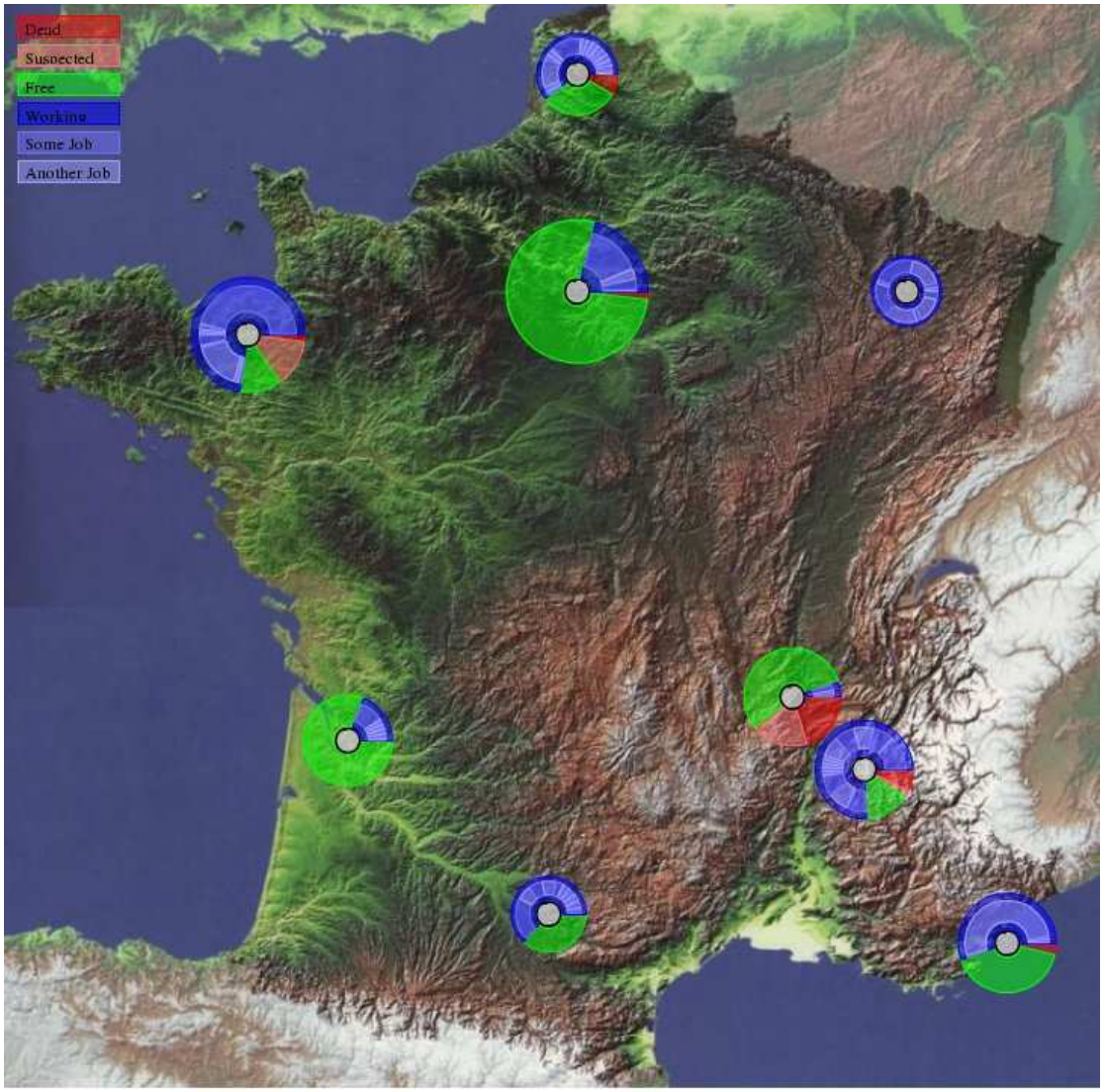

Figure 1: Localisation des sites Grid'5000 (et état des noeuds de calcul)

La plateforme est dédiée à la recherche sur les grilles de calcul, l'e-Science et les cybers infrastructures.

Les différents noeuds qui la composent sont hétérogènes entre eux. C'est une des spécificités majeures de Grid'5000. Nous avons porté notre application sur 7 des 9 sites, soit 10 clusters de plusieurs dizaines de processeurs chacuns, adressant tous leur mémoire en 64 bits mais différant suivant:

- le type de processeurs (AMD Opteron / INTEL Xeon)
- le constructeur (HP, SUN, IBM, DELL)
- le réseau inter-processeurs (Myrinet, Infiniband)

Le réseau inter-sites est un réseau dédié RENATER 2Gb/s appelé également VPN (Virtual Private Network). Certains sites sont reliés entre eux par RENATER-4 à 10Gb/s.

I.2. Le modèle de climat

La 66ème sous version du modèle de climat OCC17 comprenant l'atmosphère ARPEGE v4.6 (physique dite « intermédiaire », grille t63), l'océan OPA v9.1 (NEMO, grille ORCA 2 degrés) et le modèle de routage de fleuves TRIP est couplée par OASIS v3 (technique de communication MPI).

Ce modèle a été implémenté sur le cluster CRAY XD1 Linux du Cerfacs. L'analyse de ses caractéristiques a mis en évidence d'importantes dérives sur la température globale de surface, ainsi que des biais régionaux. Ces erreurs étant susceptibles d'affecter gravement les résultats de nos expériences, une correction de flux (sur les tensions de vents) est envisagée.

C'est la version parallèle du modèle d'atmosphère qui a été implémentée. Cela permettra, à chaque soumission batch, de choisir le meilleur ratio temps elapse/processeurs en fonction des ressources disponibles sur la grille.

En plus de la compilation des codes sources des trois modèles et du coupleur, une installation partant de zéro doit comprendre la génération avec gfortran des éléments suivants:

- les librairies de calcul BLAS et LAPACK (pour ARPEGE) **[4]**
- la librairie d'I/O Netcdf 3.6.0 (pour tous les modèles) **[5]**
- les outils de traitement de données Netcdf NCO **[6]**
- l'outil de conversion de données SAFO (format ARPEGE vers format Netcdf) **[7]**
- la suite d'outil d'analyse statistique Statpack **[8]**
- une librairie de communication MPI (nous avons choisi OpenMPI) **[9]**

II. Portage

II.1. Image Kadeploy

La mise en place du modèle « patrimonial » de climat sur l'ensemble des 10 machines n'a nécessité qu'un seul et unique portage.

En effet, toutes les machines sont capables de démarrer (reboot) sur une seule configuration standardisée (ou « image »), comprenant principalement le système d'exploitation Debian4all, un système de montage NFS, les librairies de compilation gcc et gfortran (version 4.1.2) et la librairie de communication OpenMPI.

Notez la puissance de cette approche: une seule campagne de portage, au lieu de 10, est nécessaire. De plus, la compatibilité des codes est assurée en cas d'expérience distribuée sur plusieurs machines. Par contre, il nous sera nécessaire de redémarrer le

système d'exploitation des processeurs à chaque redémarrage du modèle (checkpointing).

Le logiciel kadeploy **[10]**, développé à l'IMAG, a permis la création d'une image adaptée à nos besoins et déployable sur un grand nombre de machines de Grid'5000 (excluant toutefois PowerPC et Intel Itanium, marginales en nombre de processeurs disponibles).

Notons que cet outil est disponible en dehors de Grid'5000 et que son utilisation est possible sur un grand nombre d'architecture (linux, BSD, Windows, Solaris pour x86 et 64 bits). Aucune autre manipulation que le redémarrage des machines sur cette image n'est nécessaire pour y lancer notre simulation du climat.

II.2. Particularités de la compilation gfortran

Utilisé auparavant sur de nombreuses autres plateformes de calcul, le modèle couplé ARPEGE-OPA n'avait jamais été compilé avec gfortran **[11]**.

La particularité de ce compilateur est de ne plus supporter les standards Fortran 77 et 90 et de s'en tenir à la norme fortran95.

C'est la version 4.1.2 qui a été choisie, pour le compilateur c gcc comme pour le compilateur gfortran.

II.2.1. Options de compilation

Trois options de compilation ont été utilisées:

-fsecond-underscore : pour compatibilité avec les options par défaut du compilateur gcc (utilisé pour construire la librairie de communication netcdf)

-fconvert=big-endian : pour compatibilité avec la représentation des données des fichiers binaires d'entrée des modèles, issus de simulations effectuées sur des architectures (PC 32 bits) où le stockage en mémoire est différent

-frecord-marker=4: pour lecture formatée fortran de données écrites sur des architectures 32 bits

II.2.2. Problèmes rencontrés à la compilation

§

Routine/ modèle	***Description***	***Solution***
Toute routine utilisant SIGN(A,B) avec B=0.	La norme fortran95 signe les zéro. Cela peut changer le comportement de la routine fortran SIGN. Effet: perte de 10 degrés celsius à la surface de l'océan	Ajouter epsilon au deuxième argument de la routine SIGN
wrcpl/ARPEGE	Appel de la primitive Oasis prism_get avec un tableau de taille différente de celle qui a été déclarée. Erreur à la compilation	Recoder
usnunp/ARPEGE	Initialisation de variables avec des nombres dépassant la précision de la machine. Erreur à la compilation	Commenter l'appel de la routine (inutile)
rrtm/ARPEGE	Erreur dans l'appel des arguments. Erreur à la compilation	Commenter l'appel de la routine (inutile)
su0phy, suecrad15, sumcc / ARPEGE	Erreur de syntaxe dans l'écriture de FORMAT fortran. Erreur à la compilation	Recoder
sufpc /ARPEGE	Fonction fortran « system » indisponible	Commenter son appel
mpl_end_mod /ARPEGE	MPI_Buffer_Detach provoque une erreur système à l'exécution	Commenter son appel. MPI_Finalize fonctionne quand même
scan2mdm / ARPEGE	En configuration CDCONF(3)='0', la variable IFLDSTGO n'est pas initialisée, alors qu'elle est utilisée pour une allocation de mémoire. Erreur à l'exécution.	Forçage de la définition de IFLDSTGO par appel de la fonction SC2CGAP en configuration CDCONF(3)='0'

Oasis3 n'a posé aucun problème à la compilation et à l'exécution

II.3. OpenMPI

Un des résultats corollaires de ce portage est la qualification de la librairie de communication OpenMPI comme librairie MPI utilisable par Oasis3.

Cette librairie a été choisie pour sa grande compatibilité avec les compilateurs gcc /gfortran.

Cette librairie s'utilise comme MPICH. La clef de précompilation LAM_MPI ne doit donc pas être activée. C'est la version MPI-1 qui doit être choisie à la compilation et à l'exécution (namcouple).

Pour le lancement des exécutables, des arguments de l'utilitaire mpirun permettent de préciser, pour chaque noeud:

- le nom du noeud concerné (option -host)
- le modèle à y lancer (oasis, opa, arpege ou trip)
- le nombre de processeurs utilisés pour ce modèle (en cas de modèle parallèle)

Exemple;
mpirun -np 1 --host helios40.sophia.grid5000.fr $RUN_DIR/oasis : -np 4 --host helios41.sophia.grid5000.fr $RUN_DIR/arpege : -np 4 --host helios42.sophia.grid5000.fr $RUN_DIR/arpege : -np 4 --host helios43.sophia.grid5000.fr $RUN_DIR/arpege : -np 4 --host sol20.sophia.grid5000.fr $RUN_DIR/arpege : -np 1 --host helios40.sophia.grid5000.fr $RUN_DIR/opa9 : -np 1 --host helios40.sophia.grid5000.fr $RUN_DIR/trip

Dans cet exemple, oasis, opa et trip sont lancés sur 3 processeurs du noeud 40 de la machine helios. Arpege est lancé en mode parallèle sur 4 processeurs des noeuds 41, 42 et 43 d'helios, ainsi que sur 4 noeuds de la machine sol.

Oasis a été compilé avec la clef use_key_noIO. La librairie d'entrés sorties netcdf mpp_io n'a donc pas été testée sur cette architecture.

Notons enfin que l'environnement de compilation d'Oasis3 présente une inadéquation avec le concept de « compilation distribuée». Si le code peut-être compilé sur n'importe quelle machine de la grille, les scripts de création des fichiers makefile s'attendent, eux, à ce qu'à un nom de machine corresponde une configuration donnée.

Les fichiers de paramètres du répertoire prism/util/compile/frames/include_$HOSTNAME doivent être dupliqués autant de fois qu'il y a de machines sur la grille si l'on veut garder la possibilité d'une même compilation partout.

III. Simulation distribuée

Des essais de couplage inter-noeuds ont été effectués. Sur un même site (Sophia), les exécutables d'une simulation peuvent être lancés en distribué sur des machines différentes (azur, sol et helios).

OpenMPI est également utilisé pour la parallélisation interne du modèle d'atmosphère. Là encore, une distribution sur plusieurs machines est possible. Une série complète de tests doit maintenant nous permettre d'évaluer la scalabilité des modèles, les surcoûts induit par l'utilisation du réseau RENATER ou des réseaux intra-sites.

Références:

[**1**] Arpege PC: Note technique relative au portage, *Technical Report* **TR/CMGC/04/99**

[2] Grid'5000 web site:
https://www.grid5000.fr/mediawiki/index.php/Grid5000:Home

[3] LEGO: Grid Compliant Climate Model Analysis, *Technical Report* **TR/CMGC/07/16**

[4] LAPACK -- Linear Algebra PACKage: http://www.netlib.org/lapack/

[5] NetCDF (network Common Data Form):
http://www.unidata.ucar.edu/software/netcdf/

[6] Netcdf Operators Homepage: http://nco.sourceforge.net/

[7] SAFO: Guide Utilisateur, *Technical Report* **TR/CMGC/05/16**

[8] P. Terray, Statpack: outil d'analyse statistique,
http://www.cerfacs.fr/%7Emaisonna/Statpack/statpack.html

[9] Open MPI, Open source high performance computing:
http://www.open-mpi.org/

[10] Kadeploy Web Site: **http://kadeploy.imag.fr/**

[11] gfortran — the GNU Fortran compiler, part of GCC:
http://gcc.gnu.org/wiki/GFortran