

TR/CMGC/09/88
Portage d'ARPEGE-NEMO
sur les plateformes scalaires
IBM BG/P et SGI Altix
E. Maisonnave

Table des matières

I. Portage.....	4
I.1. SGI Altix CINES.....	4
I.2. IBM Blue Gene/P IDRIS.....	6
II. Performances.....	7
Annexe.....	11

Construit et optimisé pour les plateformes vectorielles, le modèle communautaire ARPEGE-Climat peut-il être utilisé intensivement sur des machines scalaires du type de celles mises à disposition par les centres nationaux de calcul et GENCI ?

Le présent document relate les difficultés rencontrées lors du portage de notre application sur les noeuds de calculs IBM BlueGene/P de l'IDRIS et SGI Altix du CINES, et détaille les performances comparées scalaires/vectorielles.

Des tests de sensibilité sur le niveau de résolution horizontale de l'application nous permettront d'esquisser une première conclusion quant à l'utilisation des plateformes sus-citées.

I. Portage

Le modèle de climat utilisé conjointement par le CERFACS et Météo-France pour le 5ème exercice du GIEC est bâti autour du modèle d'atmosphère ARPEGE(-IFS). Il est dérivé du modèle opérationnel utilisé en prévision météorologique par Météo-France et en prévision à moyen terme par le Centre Européen de Prévision (ECMWF). On lui adjoint un modèle d'océan (NEMO, modèle communautaire de l'IPSL) par l'intermédiaire du coupleur OASIS.

Plusieurs résolutions du modèle d'atmosphère ont été testées: en troncature 159 (notée T159, maillage d'environ 100Km de côté) et en troncature 359 (notée T359, maillage d'environ 50Km de côté), avec 30 niveaux verticaux. L'océan a une résolution de $\frac{1}{2}$ degré.

ARPEGE a été utilisé dans sa version 5.1, NEMO dans sa version 3.1 et OASIS dans sa version 3.2.

Nous rappelons que le coupleur OASIS s'appuie sur la fonctionnalité MPMD de la librairie de communication MPI. Plusieurs instances de chacun des trois exécutables parallèles qui composent le modèle de climat (atmosphère, océan et coupleur) sont démarrées simultanément (via un lanceur de type mpirun ou mpiexec), OASIS se chargeant de re-définir des sous-communicateurs pour autoriser la parallélisation interne des modèles d'océan et d'atmosphère. OASIS pilote également par MPI les échanges qui se font entre les deux autres exécutables..

Nous décrivons maintenant les problèmes majeurs survenus lors de la mise en service de notre application sur les deux machines accessibles par la demande de moyens DARI 2009. Le rapport des divers aléas techniques qui ont pu être surmontés (problème de compilation, d'édition de lien, d'exécution) sera fourni en annexe.

I.1. SGI Altix CINES

A notre connaissance, seul le modèle océanique avait déjà fait l'objet d'un portage sur ce type de machine. Pour celui-ci, il nous a toutefois été nécessaire de limiter l'optimisation au niveau 2 du compilateur.

Pour le modèle d'atmosphère, plusieurs fonctionnalités ont dû être désactivées, dont le programme

de traçage interne « DrHook ». Pour une optimisation ultérieure, il sera nécessaire de se pencher sur ces dysfonctionnements.

Plus problématique se révèle l'utilisation simultanée du mode couplé du modèle d'atmosphère avec sa parallélisation interne « 2D ». Les améliorations du parallélisme implémentées à l'ECMWF permettent en effet:

1. pour la partie où les équations du code sont résolues en point de grille, un découpage des domaines MPI en rectangles (c'est à dire suivant la latitude et la longitude, et non plus seulement en bandes de latitude), ce qui permet bien sûr d'augmenter la parallélisation.
2. pour la partie où les équations du code sont résolues en mode spectral, la parallélisation se fait sur le nombre d'harmonique et de niveaux verticaux.

Le niveau de parallélisation en mode point de grille et en mode spectral devant être le même, la décomposition limite pour une T159, avec 30 niveaux verticaux, se fera sur environ 3000 processeurs maximum, 7000 pour une T359.

En mode couplé, il apparaît que les routines d'interface avec OASIS n'ont pas été mises à jour pour permettre une parallélisation de type 2D. Notre étude en couplé a donc été limitée par une parallélisation 1D. Le découpage en domaine ne dépasse pas 110 en T159, 250 en T359: cela reste assez éloigné de l'optimum mesuré pour ARPEGE en mode stand-alone.

Les optimisations les plus efficaces du couplage OASIS3 ont été apportées sur cette configuration: parallélisation d'OASIS par champs de couplage (jusqu'à 16 processeurs utilisés), utilisation du mode séquentiel de couplage. L'amélioration est conforme à celle qui avait été précédemment constatée (à plus haute résolution) sur machine vectorielle NEC SX8: le temps passé sur l'ensemble des fonctions de communication et d'interpolation du coupleur reste inférieur à la différence de temps de restitution entre les modèles d'océan et d'atmosphère [1]. **A cette résolution, il n'est donc pas nécessaire d'avoir recours à l'utilisation d'une version assurant une plus grande scalabilité au coupleur (OASIS4).**

Un problème est apparu en comparant les temps de restitution du modèle d'atmosphère en mode couplé et stand-alone. La différence (couplé plus lent) n'apparaît pas seulement aux pas de temps de couplage, mais à chaque pas de temps des modèles.

En mode couplé, chacune des instances des différents modèles (et du coupleur) sont « mappées » sur les processeurs alloués suivant un algorithme qui ne nous est pas accessible. De fait, nous ne pouvons pas nous assurer que des processus « atmosphère » ne partagent pas la mémoire d'un noeud de calcul avec un autre processus (« océan » ou « coupleur »), ce qui pourrait ainsi perturber les communications entre processus du modèle d'atmosphère.

Afin de limiter les risques d'interactions, nous lançons 8 instances de notre coupleur, en espérant qu'elles seront placées par le « launcher » sur les huit coeurs d'un même noeud et qu'elles ne partageront donc pas la mémoire avec les processus des modèles.

En suivant ce protocole, les temps de restitution moyens pour un pas de temps hors IO, hors convection, hors couplage OASIS, avec NEMO sur 64 processus pour les expériences couplées, sont les suivants:

	ARPEGE sur 16 processus	ARPEGE sur 104 processus
ARPEGE seul	1.05	0.195
Couplé avec Oasis sur 1 processus	1.09	0.264
Couplé avec Oasis sur 8 processus	1.10	0.264

Temps de restitution (en secondes) d'un pas de temps d'ARPEGE T159

Le fait de réserver l'équivalent d'un noeud entier (8 coeurs) pour OASIS ne change pas le ralentissement. Celui-ci se renforce avec le nombre de coeurs utilisés pour ARPEGE (5% à 16 coeurs, 25% à 104 coeurs). Mais, sans certitude sur la localisation des différents processus, il ne nous est pas possible d'aller plus avant dans la recherche de l'origine de ce ralentissement.

Néanmoins, malgré ce problème de surcoût non maîtrisé, **les meilleurs temps de restitution possibles avec la configuration t159-1/2 degré sont de l'ordre de 1.5 jour par 10 années simulées sur environ 400 coeurs, ce qui fait de la machine SGI du CINES une bonne candidate à la mise en production de simulations de climat moyenne et longue durée**, même avec les résolutions les plus importantes actuellement utilisées en standard dans le communauté française.

Bien que la force de travail nécessaire à la mise au point d'une simulation complète soit encore importante (post-traitement, enchaînement/checkpointing des simulations mensuelles, validation physique du couplage), et encore plus à son optimisation, il est raisonnable d'envisager au besoin une utilisation de cette machine pour la réalisation de simulations de type « exercice GIEC », sans préjuger toutefois de la similarité du climat simulé avec celui produit sur la machine vectorielle de Météo-France.

I.2. IBM Blue Gene/P IDRIS

Précédemment réalisé sur les machines IBM BG/L (CERFACS) et IBM JS21 Mare Nostrum (BSC), le portage des modèles d'atmosphère et d'océan n'a posé aucun problème particulier.

Une complication rédhibitoire est toutefois survenue lors du couplage des deux composantes. Du fait de certaines caractéristiques de l'environnement logiciel de la machine Blue Gene/P, il apparaît en effet que la version 3 de notre coupleur OASIS ne peut y être utilisé sans modification majeure du code.

L'utilisation du mode MPMD de la librairie de communication MPI propre à la machine comporte en effet deux type de restriction:

1. un même exécutable ne peut être lancé que sur un nombre entier de « pset » (ensemble de noeuds de calcul associés à un noeud d'IO; sur babel, ce nombre de noeuds de calcul est de 64)
2. les numéros de rang MPI associés aux process d'un même exécutable ne se suivent pas.

La version 3 du coupleur OASIS que nous utilisons n'a été adaptée que tardivement à une utilisation en parallèle. Le nombre d'instance MPI de ce coupleur ne peut en effet pas dépasser le nombre de champs de couplage, soit 15 dans notre cas, ce qui n'est pas compatible avec la première contrainte.

Deuxièmement, la définition des sous-communicateurs associés aux modèles couplés impose que les numéros de rang MPI des instances du coupleur commencent au chiffre zéro et se suivent un par un (par exemple, 0,1,2,3,4,5,6,7 pour 8 instances du coupleur). Ce qui ne satisfait pas non plus la deuxième contrainte.

Nous touchons ainsi pour la première fois une limite de notre coupleur « pseudo-parallèle » OASIS3. Pour poursuivre notre étude sur ce type de machine (et probablement des machines massivement parallèles en général), **sur IBM Blue Gene/P, il est désormais indispensable d'avoir recours au coupleur parallèle OASIS4**. Une action dans ce sens est d'ores et déjà entreprise pour définir la nouvelle interface dans le modèle d'océan NEMO. Elle sera poursuivie avec un travail identique dans le modèle d'atmosphère.

D'une manière plus générale, nous devons également nous interroger sur la pérennité de notre approche basée sur le couplage d'exécutables séparés. Cette technique n'est pas celle choisie par nos homologues américains (qui préfèrent construire un seul exécutable avec toutes leurs composantes). On note également que dans la communauté scientifique des utilisateurs de supercalculateurs, cette technique est peu répandue.

Si elle répond à un besoin de modularité et facilite les échanges dans notre communauté scientifique, peut-être est-elle également moins adaptée à la structure des machines massivement parallèles, standardisée pour le plus grand nombre d'utilisateurs ?

II. Performances

Des tests de scalabilité ont été effectués avec le modèle atmosphérique dans deux résolutions différentes (T159 et T359).

La première s'utilise désormais en standard dans certains laboratoires de climat. La seconde est réservée à des simulations frontières, son usage à des fins scientifique est encore expérimental et requiert un volume conséquent d'heures de calculs.

Un point important pour la comparaison des résultats: sur toutes les machines, les IO ont été débranchées, le mode de parallélisme 2D (voir § I.1) pouvant également poser problème dans les routines d'IO sur certaines machines (dont la machine IBM BG/P).

Sur IBM BG/P, les expériences ont été réalisées en mode DUAL; sur SGI Altix, en utilisant les 8 coeurs d'un noeud.

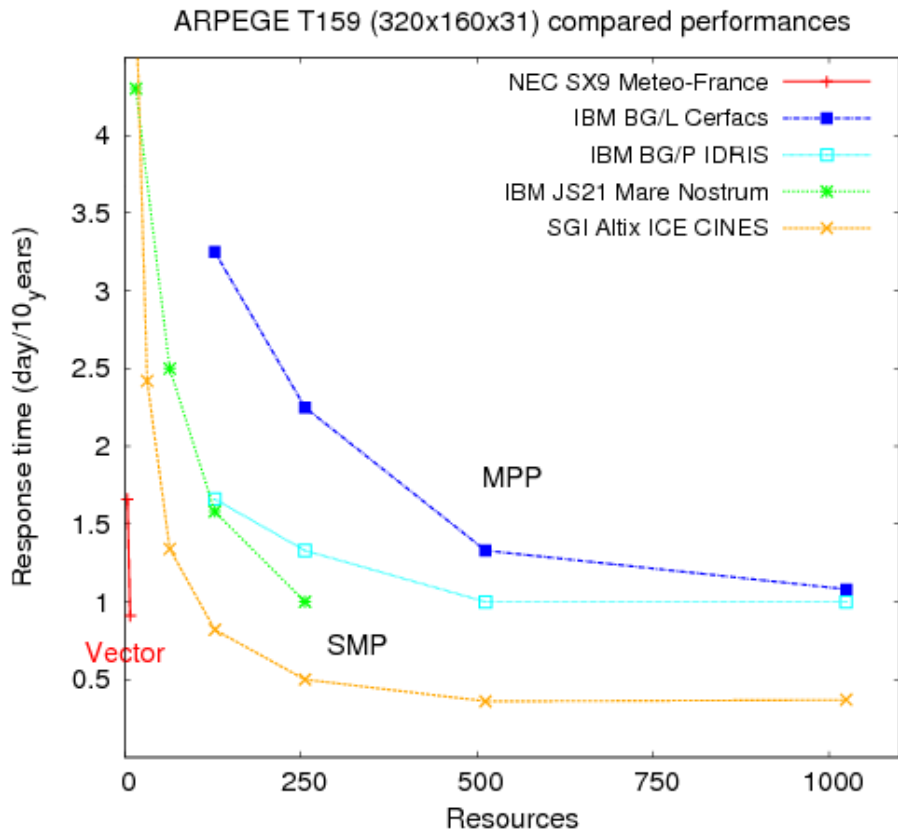


Fig 1: Temps de restitution du modèle d'atmosphère ARPEGE sur différentes plateformes

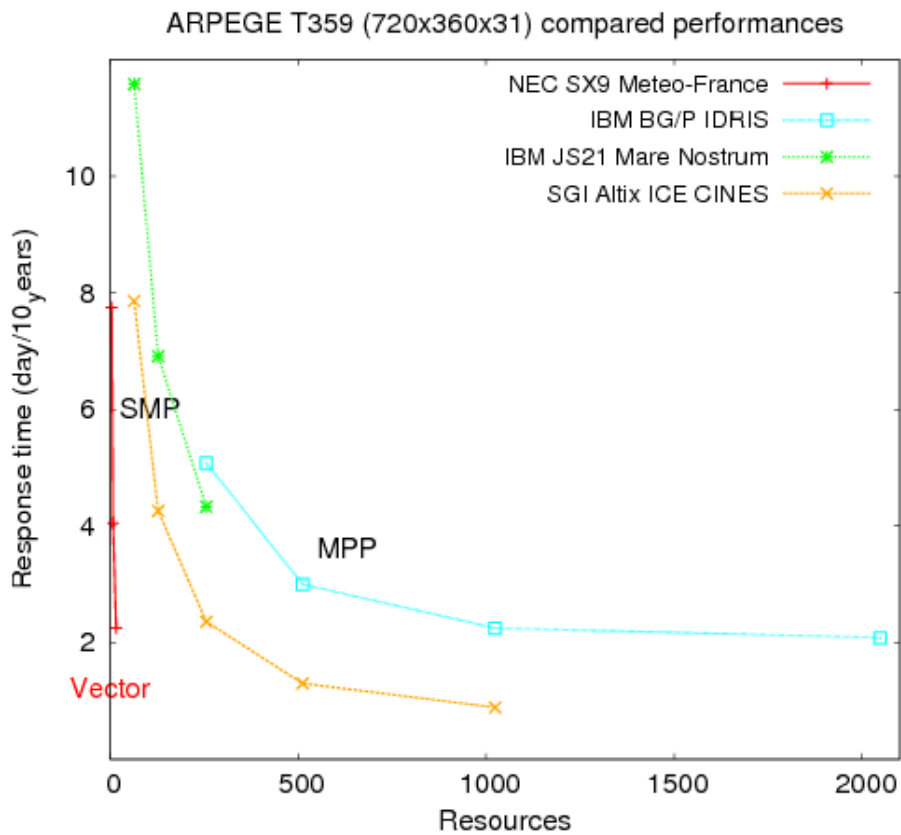


Fig 2: idem avec ARPEGE haute résolution

Ces résultats suggèrent que **les performances du modèle d'atmosphère en résolution standard (sans IO) sur les machines examinées lors de la demande de moyens DARI 2009 sont compatibles avec une utilisation en mode production** (tant en prévision saisonnière qu'en étude longue de variabilité naturelle). Il en est de même **pour la résolution « frontière » T359**, puisque **une centaine d'années peuvent être simulées en moins de 15 jours** (en utilisant 500 coeurs de calcul sur la SGI Altix du CINES).

Pour une utilisation en production, ces bons résultats doivent être confirmés et des mises au point importantes doivent être faites:

- des développements complémentaires doivent être entrepris pour la réactivation des IO (opération toujours préjudiciable sur le plan des performances)
- une amélioration des performances du modèle d'océan doit être conduite en priorité, celui-ci se montrant bien moins scalable qu'attendu, allant même jusqu'à devenir le modèle le plus lent du système couplé, quelque soit le nombre de processeurs lui étant alloués
- une redéfinition de l'interface de couplage de l'atmosphère est nécessaire pour pouvoir permettre l'activation de la parallélisation 2D dans ce modèle

Cette étude de performances comparées de différentes résolutions de nos modèles hautement parallélisés sur machines scalaires nous conduit à suggérer une piste dans la définition des modèles de climat pour les années à venir.

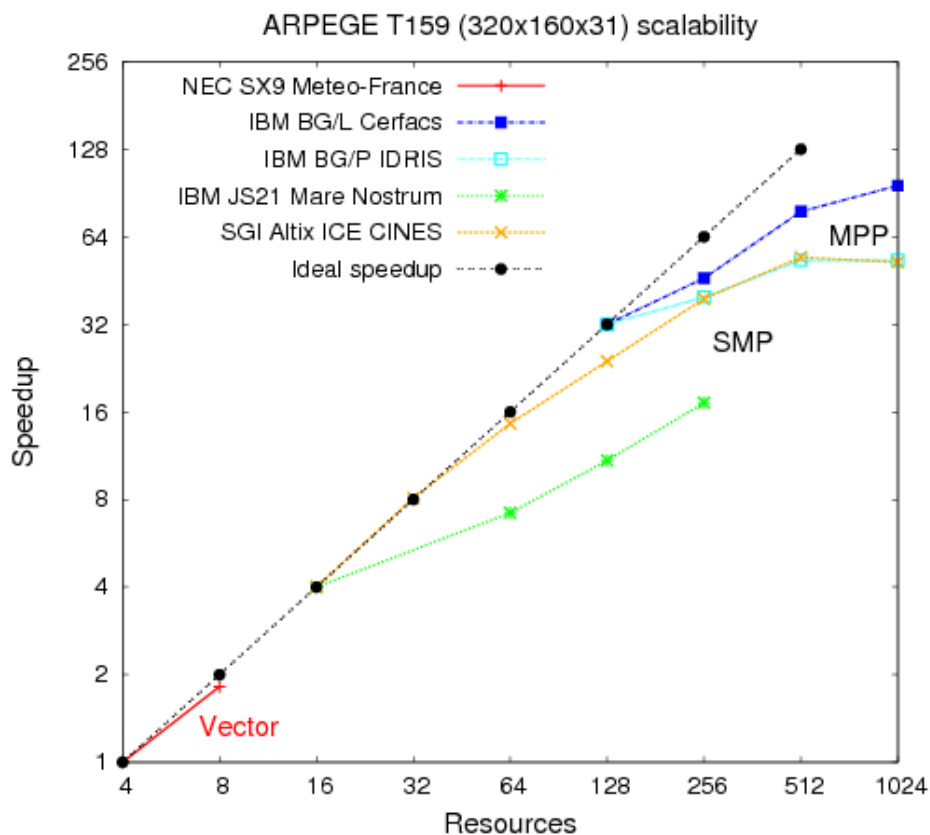


Fig 3: Speedups comparés du modèle d'atmosphère ARPEGE sur différentes plateformes

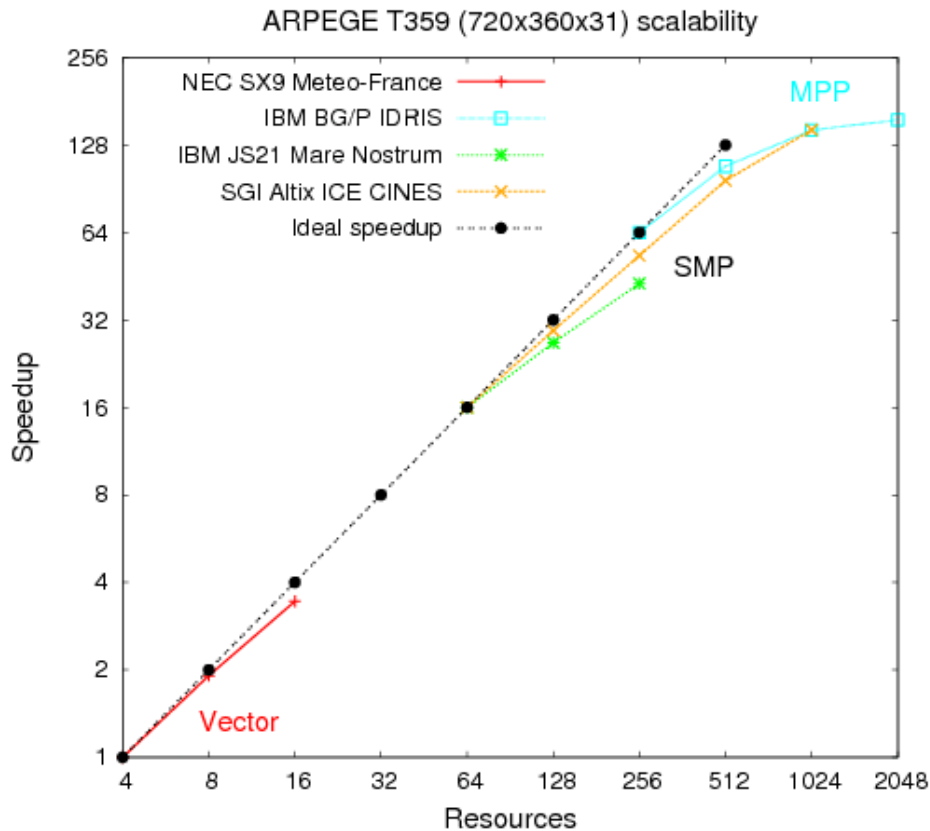


Fig 4: idem pour ARPEGE haute définition

Comme on le voit, sur les figures 3 et 4, le speed-up reste bon au delà de 500 processeurs dans la version haute définition du modèle alors qu'il faiblit déjà à 100 avec la résolution standard.

Ce résultat ne peut bien sûr être extrapolé indéfiniment mais il suggère cependant que le « weak scaling » semble être suffisamment bon avec ce modèle pour envisager de pouvoir tester des parallélisations supérieures à 10.000, voire 100.000 domaines et commencer à se confronter à des problématiques que nous rencontreront d'ici quelques années sur ce type de machines (telles que la tolérance aux pannes).

La scalabilité de notre modèle de climat augmentant avec la résolution, ce sont nos configurations « frontières » qui tireront le meilleur parti des futures plateformes massivement parallèles.

Tous mes remerciements aux équipes de support de l'IDRIS, d'IBM, du CINES et du CERFACS, et en particulier à Isabelle Dupays, Thierry Goldman, Pascal Vezolle, Mathieu Cloirec et Isabelle d'Ast.

Annexe

Symptôme	Origine	Solution
Exécution ARPEGE stand alone: arrêt au premier appel des routines de la librairie de trace « Dr hook »	?	Désactivation dans modules/yomhook.F90
Exécution ARPEGE stand alone: MPI_Comm_group: Invalid communicator	Edition de lien avec MPI SGI (MPT) mais lancement avec MPIIntel	Utilisation de la librairie optimisée MPT
Exécution ARPEGE stand alone: restart illisible	Non correspondance de la longueur des enregistrements Fortran sur la machine qui a produit le restart et sur jade	Option de compilation: -assume byterecl
Exécution ARPEGE couplé: blocage à la définition d'un champ de couplage non défini dans la namcouple	Couplage incomplet	Simplification de l'interface de couplage ARPEGE-OASIS: pas de définition du champ incriminé
Exécution ARPEGE couplé: WIND TOO STRONG !!	Incohérence des flux couplés avec les SST, choc à l'initialisation	Mise à zéro des flux durant la période de prise de performances du modèle
Exécution OASIS: code de plantage -10 dans OASIS lors de la réception du premier champ de couplage	Définition de la partition des champs de couplage	Parallélisation 2D de l'interface de couplage à actualiser (cycle 35 ?)
Exécution OASIS: suspendue	Redéfinition de la partition pour la parallélisation interne d'ARPEGE: incohérence avec la partition déclarée à OASIS	LEQ_REGIONS=.F. Dans la namelist ARPEGE
Compilation NEMO: compilation limrhg_2.F90 impossible	?	Optimisation limitée à -O2 pour la routine
Exécution NEMO stand alone: explosion du modèle (solveur) après 16 pas de temps	?	Optimisation limitée à -O2 pour tout le code

Liste des problèmes rencontrés lors du portage du modèle couplé sur SGI Altix CINES

[1] Proceedings of the OASIS User Meeting 2009, TR/CMGC/09/89