Développement du coupleur OASIS4 dans le cadre du projet ANR CICLE - janvier 2008

S. Valcke
M.-P. Moine
L. Coquart

Rapport technique CERFACS
TR/CMGC/08/78

Rapport semestriel d'activité -partenaire
Programme CIGC - Edition 2005

Identification

Acronyme du projet	CICLE
Numéro d'identification de l'acte attributif	ANR-05-CICG-004-02
Coordonnateur (société/organisme)	IPSL
Partenaire (société/organisme)	CERFACS
Période couverte (date à date)	01/07/2007 à 31/12/2007
Période couverte (t0+n mois à t0+m mois)	T0+18 mois à t0+24 mois
Rédacteur (nom, téléphone, email)	Sophie Valcke 05.61.19.30.76 valcke@cerfacs.fr
Date	01/02/2008

Synthèse

Conformité des résultats obtenus aux prévisions (1)	Conformité de la consommation des ressources par rapport aux prévisions (2)	Difficultés particulières (3)
Globalement conforme à quelques réserves près (voir la section Conformité de l'avancement aux prévisions)	Conformes pour les embauches ; en-dessous des prévisions pour les missions (4400 Euros au total au lieu des 8000 Euros prévus).	La remarque faite dans le rapport précédent s'applique toujours. Bien qu'une version parallèle fonctionnelle du coupleur OASIS4 soit disponible depuis 6 mois pour l'assemblage des modèles couplés du projet, le travail de validation détaillée de toutes les interpolations pour tous les types de grilles du projet, en particulier en parallèle, prend plus de temps qu'initialement estimé.

(1) Les résultats sont supérieurs aux prévisions, conformes aux prévisions, inférieurs aux prévisions
(2) Consommation supérieure aux prévisions, conforme aux prévisions, inférieure aux prévisions.
(3) A compléter en particulier si les résultats sont inférieurs aux prévisions et/ou la consommation supérieure aux prévisions.

Faits marquants
Indiquer les résultats et/ou réalisations marquants. Préciser s'ils peuvent ou non faire l'objet de communications externes par l'ANR et la Délégation ANR-CI.

Au cours des 6 derniers mois, la diffusion du coupleur OASIS4 au niveau international s'est poursuivie. Le « Nansen Environmental and Remote Sensing Center » (NERSC, Norvège) a rejoint le groupe des utilisateurs[1] et mis en place un

[1] En plus des partenaires de CICLE, du NERSC et de l'AWI, le groupe des utilisateurs actuels est formé des 3 partenaires du projet européen GEMS (CEPMT, KNMI, Météo-France) pour un couplage 3D entre des codes de dynamique et de chimie atmosphériques, du Swedish Meteorological and Hydrological Institute (SMHI, Suède) pour un couplage océan-atmosphère régional, et du UK Met Office (Royaume-Unis) pour un couplage global océan-atmosphère.

couplage global océan-atmosphère basé sur OASIS4. De plus, l'Alfred Wegner Institute (AWI, Allemagne) coordonnant les activités de recherche sur les régions polaires en Allemagne, a choisi OASIS4 comme coupleur et proposé une collaboration portant sur l'incorporation des grilles non-structurées. Cette collaboration devrait se concrétiser au cours des prochains mois.

Ces résultats peuvent faire l'objet de communications externes par l'ANR et la Délégation ANR-CI.

Description des travaux effectués par le partenaire depuis le dernier rapport d'activité

Faire référence au découpage (tâches) du projet.

La poursuite de la validation détaillée de toutes les interpolations d'OASIS4 pour tous les types de grilles du projet a constitué une bonne partie du travail fourni durant les 6 derniers mois. En particulier, la validation de la parallélisation des interpolations, initiée durant cette période, a bien avancé mais a cependant révélé certaines difficultés pour les grilles Gaussiennes Réduites et l'interpolation conservative 2D. Les détails de cette validation et les problèmes rencontrés sont donnés à la section 1.

De plus, on a depuis lors fait tourné les modèles jouets du modèle couplé IPSLCM4 et du quadri-couplé du CNRM-GAME[2] en activant toutes les fonctionnalités requises pour les véritables modèles couplés. Ces tests ont permis de détecter des problèmes, résolus depuis lors pour certains, en cours de résolution pour d'autres (voir la section 2).

Finalement des premiers tests de performance et de scalabilité du coupleur OASIS4 ont été réalisés et sont présentés à la section 3.

Résultats obtenus / livrables fournis par le partenaire depuis le dernier rapport d'activité

Décrire les résultats obtenus et détailler les livrables (développements, tests, rapports, publications, présentations aux congrès, …).

N'ayant pas de livrable particulier à fournir sur la période visée, le travail a principalement porté sur l'amélioration du coupleur OASIS4. Les sources d'OASIS4 à jour sont disponibles sous http://www.cerfacs.fr/prismsvn/branches/development/prism/. Bien qu'une version parallèle fonctionnelle du coupleur OASIS4 (2e partie du livrable 4.1) ainsi qu'une version pseudo-parallèle d'OASIS3 soient disponibles depuis 6 mois (voir le rapport à t0+18) pour l'assemblage des modèles couplés, la validation détaillée et les tests de qualité des diverses fonctionnalités du coupleur OASIS4[3] sont toujours en cours.

Nous rapportons les résultats obtenus dans les 3 sous-sections suivantes :

1. Tests et validation des interpolations du coupleur parallèle OASIS4
2. Autres fonctionnalités testées avec les modèles couplés jouets
3. Mesure du coût et des performances du coupleur OASIS4

[2] Ces modèles jouets forment une partie du banc d'essai requis pour tester en pratique les fonctions d'OASIS et livrés au mois 18 du projet (1ère partie du livrable 4.1).

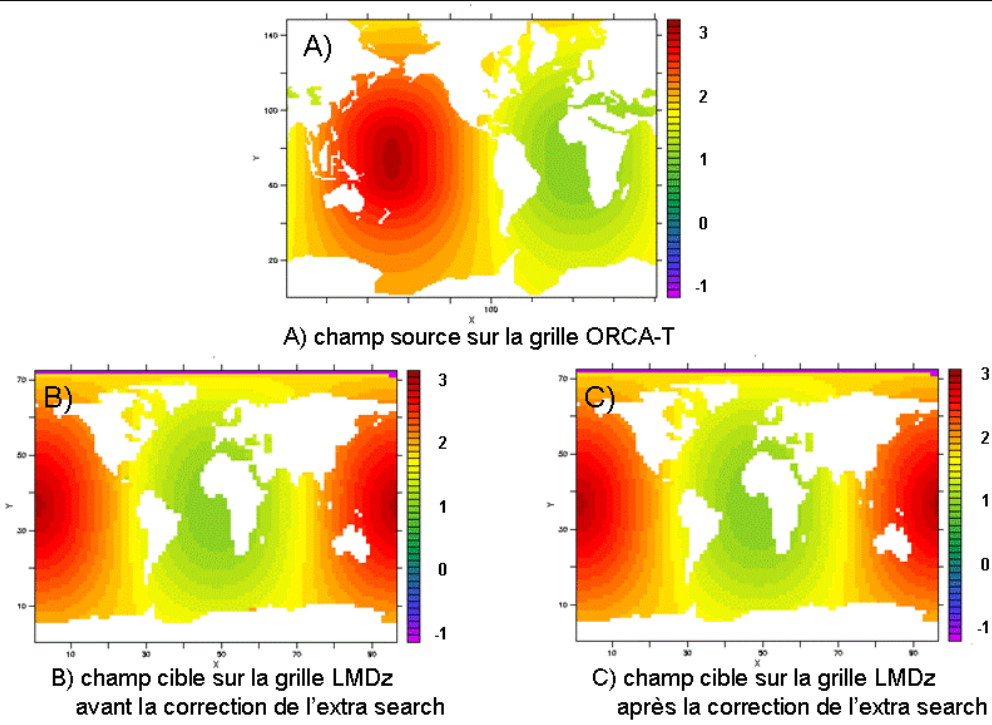
[3] Les fonctions à intégrer dans le coupleur OASIS4 ont été identifiées au mois 12 dans le « Rapport sur les fonctionnalités à intégrer au coupleur OASIS » délivré au mois 12.

1. Tests et validations des interpolations du coupleur parallèle OASIS4

Tous ces tests ont été effectués avec le banc d'essai réalisé pour évaluer en pratique les fonctions d'OASIS4 (1ère partie du livrable 4.2). Pour plus de détails sur son fonctionnement et sur le calcul de l'erreur, voir la section 3 de notre dernier rapport. Note : Afin de faciliter le suivi du travail, la numérotation des paragraphes suivants suit celle de la section « 4. Intégration dans OASIS4 des fonctions identifiées » de notre dernier rapport à t0+18 mois.

1.1 Interpolation parallèle de type bilinéaire et bicubique pour des grilles « longitude-latitude », « logiquement rectangle » et « gaussienne réduite »

- Ces interpolations sont implémentées et validées en mode monoprocesseur. Dans le cas de l'interpolation de la grille ORCAT-LMDZ, il reste cependant un problème particulier en cours d'étude : la ligne de points au pôle Nord de LMDz ne reçoit pas de valeur (voir la figure 1).
- Les problèmes soulignés dans le dernier rapport pour quelques mailles cibles particulières après l'interpolation bilinéaire à partir des grilles ORCA2-T et MED1/2-T ont été résolus. Pour toutes les mailles cibles concernées le calcul de l'interpolation bilinéaire ou bicubique ne peut se faire car ces mailles tombent dans une région de la grille source entièrement masquée. Pour ces mailles, une recherche supplémentaire du « plus proche voisin » source non masqué (appelée « extra-search ») est alors activée. Nous avons mis en évidence que l'extra-search conduisait à un résultat erroné imputable à l'algorithme multi-grilles utilisé. Cet algorithme a depuis été corrigé et l'impact de sa correction est illustré à la figure 1.



A) champ source sur la grille ORCA-T

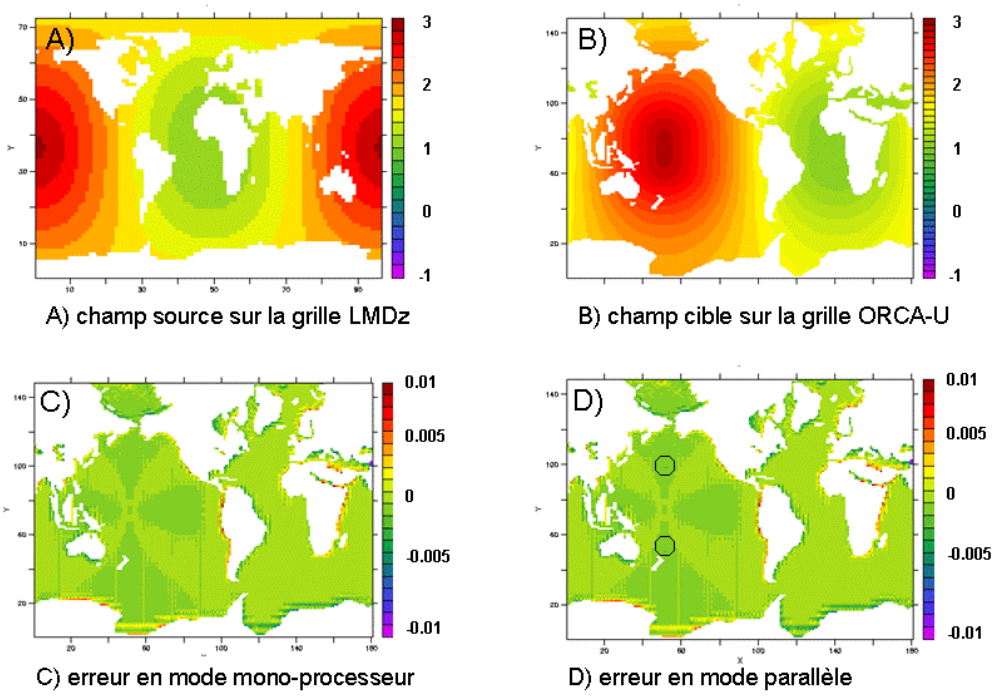
B) champ cible sur la grille LMDz avant la correction de l'extra search

C) champ cible sur la grille LMDz après la correction de l'extra search

Fig. 1: Interpolation bilinéaire de ORCA-T vers LMDz avant et après correction de l'extra search: On observe des améliorations importantes le long des côtes (voir par exemple les mailles orange en Antarctique en B qui n' existent plus en C). De plus, on voit le problème de la ligne de points au pôle Nord qui ne reçoivent pas de valeur (ligne violette au Nord en B et C).

- Ces tests nous ont également permis de noter que les points cibles tombant dans des « trous » de la grille source (i.e. dans une région non-couverte par la grille source) ne reçoivent pas de valeurs. Une solution devra à terme être apportée, soit au niveau du coupleur, soit au niveau du modèle cible après la réception, car le modèle cible a évidemment besoin de champs comportant des valeurs sur toutes ses mailles non masquées.
- Dans le cas parallèle, la recherche globale[4] pour ces interpolations (bilinéaire et bicubique) pour les grilles « longitude-latitude » et « logiquement rectangle » a été implémentée et les premiers tests de validation sont très positifs. La figure 2 montre que l'interpolation bilinéaire parallèle (avec la source et la cible partitionnées en 3 bandes de latitude) donne quasi-exactement les mêmes résultats que l'interpolation monoprocesseur et ce même aux bords des domaines locaux, sauf en 2 points situés dans le Pacifique tropical Nord et Sud, en cours d'étude. La validation va être poursuivie pour d'autres types de partitions et complétée dans les mois qui viennent.

[4] La recherche *parallèle globale* consiste à effectuer la recherche des points sources participant au calcul de l'interpolation de chaque point cible sur tous les processus sources ; dans le cas de la recherche *locale*, la recherche des points sources est faite en considérant seulement le processus gérant le domaine dans lequel tombe le point cible associé. La recherche locale peut entraîner des résultats incorrects sur les bords des domaines locaux.



A) champ source sur la grille LMDz

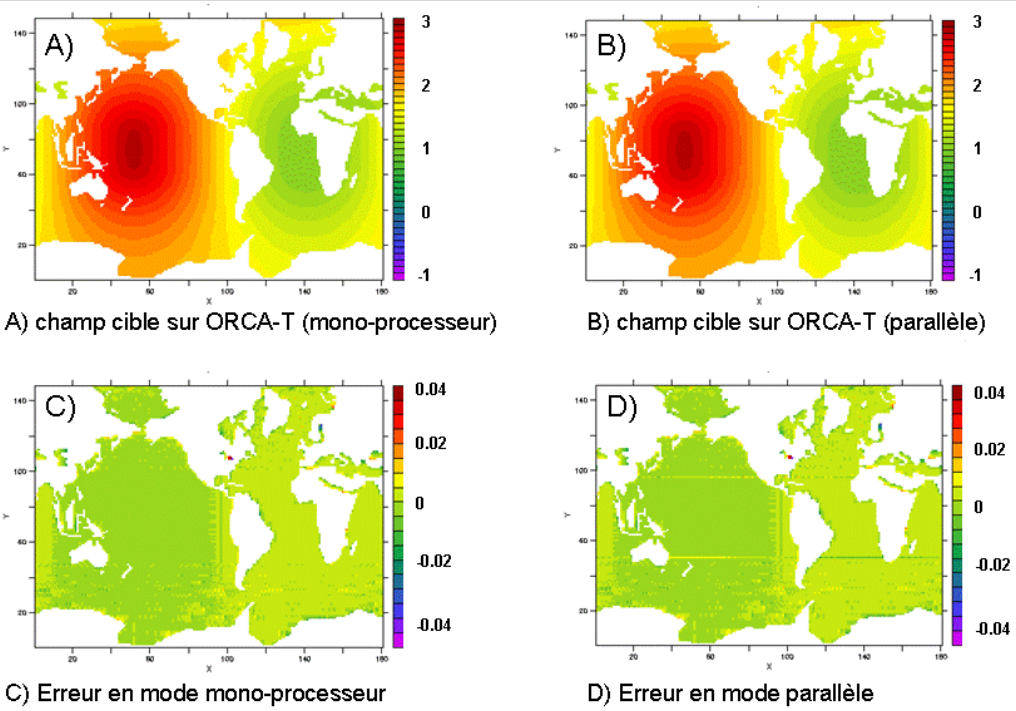
B) champ cible sur la grille ORCA-U

C) erreur en mode mono-processeur

D) erreur en mode parallèle

Fig. 2: Interpolation bilinéaire de LMDz vers ORCA-U en modes mono-processeur et parallèle:
Les champs de l'erreur en modes mono-processeur C) et parallèle D) montrent que l'interpolation bilinéaire en ces deux modes donne quasi-exactement les mêmes résultats et ce même aux bords des domaines locaux, sauf en 2 points situés dans le Pacifique tropical Nord et Sud, en cours d'étude.

- Dans le cas parallèle, ces interpolations (bilinéaire et bicubique) pour les grilles de type « gaussienne réduite » fonctionnent mais la recherche parallèle globale[4] des « voisins » présente encore des problèmes qui sont en cours de résolution. A la figure 3 nous montrons donc ce que produit une telle interpolation parallèle privée de la recherche globale, de façon à mettre en relief l'importance de cette dernière et la nécessité de son implémentation pour tous types de grille. Le cas présenté est celui de l'interpolation bilinéaire de la grille Arpège Gaussienne Réduite (BT42) vers ORCA-T. Les différences au bord des partitions sources ne se remarquent pas directement dans les champs cibles (figure 3 A et B) mais apparaissent clairement dans les champs d'erreur (figure 3 C et D). Bien que cette erreur ne soit pas particulièrement élevée, il est important d'implémenter la recherche globale des voisins, non pas pour la qualité de l'interpolation mais pour s'assurer de la cohérence des résultats en parallèle et en monoprocesseur. .



A) champ cible sur ORCA-T (mono-processeur)

B) champ cible sur ORCA-T (parallèle)

C) Erreur en mode mono-processeur

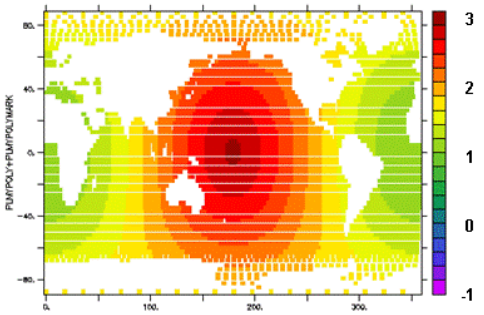
D) Erreur en mode parallèle

Fig. 3: Interpolation bilinéaire de la BT42 vers ORCA-U en modes mono-processeur et parallèle: Les champs de l'erreur en modes mono-processeur C) et parallèle D) montrent bien l'impact de l'absence de recherche globale avec des différences marquées au bord des partitions sources (3 bandes de latitude).

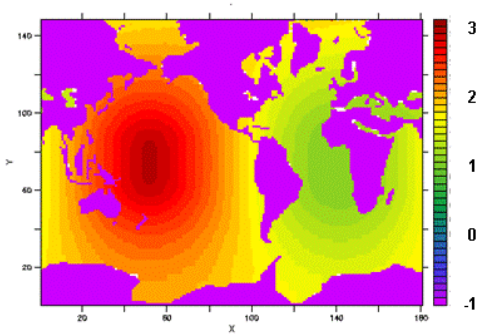
1.2 Interpolation parallèle 2D conservative par intersection de surface des mailles pour des grilles « longitude-latitude » régulière, « logiquement rectangle » et « gaussienne réduite » :

- Tel qu'écrit dans le précédent rapport, l'interpolation 2D conservative a été implémentée et validée en monoprocesseur pour les grilles «longitude-latitude», «logiquement rectangle». Les problèmes spécifiques des grilles régionales mentionnés dans le dernier rapport ont été résolus dans le cas MED1/2-T vers ALADIN mais pas encore dans le cas d'ALADIN vers MED1/2-T.
- Les tests plus détaillés ont cependant identifié le besoin d'implémenter une fonction de type « extra-search » (recherche supplémentaire du « plus proche voisin » source non masqué) pour les mailles cibles tombant dans une région de la grille source entièrement masquée. Cette fonction, déjà développée pour les interpolations bilinéaire et bicubique (cf le paragraphe 1.1), ne devrait pas être difficile à mettre en place pour l'interpolation conservative 2D.
- Tout comme pour les interpolations bilinéaire et bicubique, les points cibles tombant dans des « trous » de la grille source (i.e. dans une région non-couverte par la grille source) ne reçoivent pas de valeurs et il faudra, à terme, y remédier (voir le 3^{e} paragraphe de 1.1 ci-haut).
- Depuis le dernier rapport, la qualité de l'interpolation 2D conservative a également été évaluée en monoprocesseur pour les grilles de type « gaussienne réduite ». Les résultats, illustrés à la Figure 4, montrent

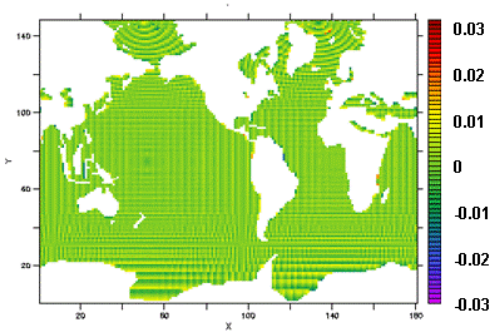
que l'interpolation de la grille ARPEGE Gaussienne réduite vers la grille ORCA-T est globalement bonne avec une erreur dans les bassins inférieure à 0.01 sauf en quelques points dans l'océan arctique et pour j=148 ; ces points spécifiques sont en cours d'étude. (à noter que sur la figure 4, l'échelle de couleur exclut les points problématiques à j=148 qui apparaissent alors en blanc - en haut à droite du domaine). La figure 4 illustre aussi le besoin de l' « extra-search » : en B, on observe des mailles cibles non-masquées apparaissant en blanc le long des côtes car ne recevant pas de valeur (le masque apparaît en violet).



A) champ source sur la grille ARPEGE Gaussienne Réduite (BT42)



B) champ cible sur la grille ORCA-T



C) Calcul de l'erreur

Fig. 4: Interpolation 2D conservative de la BT42 vers ORCA-T en mode mono-processeur

- La recherche parallèle globale des « voisins » pour l'interpolation 2D conservative, nécessaire à son bon fonctionnement en parallèle, n'a pas encore été implémentée dans la librairie du coupleur.

1.3 Implémentation de la conservation parallèle globale avant-après interpolation :

Comme décrit dans le dernier rapport, cette fonctionnalité ne sera pas implémentée dans le cadre du projet car aucun des modèles couplés CICLE ne l'utilisera; l'interpolation conservative 2D sera plutôt utilisée pour assurer une conservation locale.

1.4 Implémentation dans OASIS4 de la possibilité de lire et utiliser un fichier de poids et d'adresses prédéterminé par l'utilisateur

Cette fonctionnalité n'a pas été implémentée. Vu le délai, un consultant sera engagé sur des fonds propres CERFACS pour mener à bien cette tâche d'ici les 6 prochains mois.

Pour les points correspondants aux paragraphes 5., 6., 7. du dernier rapport, voir les conclusions décrites dans ce dernier rapport.

2. Autres fonctionnalités testées avec les modèles couplés jouets

Les modèles couplés jouets livrés au mois 18 du projet (1ère partie du livrable 4.1), servant à tester les capacités du coupleur OASIS4 dans un contexte de représentation spatiale et d'échange des champs de couplage identique à celui des modèles climatiques couplés IPSLCM4 et CNRM-GAME, ont été décrits dans le détail dans le dernier rapport.

Depuis le dernier rapport, on a exécuté ces modèles couplés jouets en mode monoprocesseur et en mode parallèle (avec une partition simple sur 2 bandes de latitude pour chacune des composantes et avec 2 processus pour le coupleur) dans un environnement Linux et sur le NEC SX8 de Météo-France, en activant toutes les fonctionnalités requises pour les véritables modèles couplés. Ces tests se sont exécutés correctement mis à part les deux problèmes décrits ici :

- Ces tests ont permis de détecter un problème dans le modèle source lorsque l'on active à la fois de l'interpolation conservative 2D et d'autres types d'interpolation depuis une même grille source (e.g. LMDz). Ce problème a été partiellement résolu dans le cas d'une configuration simple à deux composantes avec deux champs de couplage l'un interpolé avec un algorithme bilinéaire, l'autre avec l'algorithme conservatif 2D, mais est toujours présent dans la configuration complète des jouets.
- De plus, la fonctionnalité permettant de démarrer une relance[5] de la simulation couplée en lisant automatiquement au premier pas de temps, dans des fichiers dits de « restart » [6], les champs de couplage en entrée des modèles a été activée. Ces fichiers sont créés automatiquement par la librairie de couplage d'OASIS4 au dernier pas de temps de la relance précédente. Pour la toute première relance, des fichiers restarts d'initialisation doivent être générés indépendamment (l'équivalent des conditions initiales pour un vrai modèle). Pour le jouet couplé IPSLCM4, les tests sont concluants, alors qu'un problème dans la librairie des entrées-sorties d'OASIS4 causant l'arrêt de la simulation couplée est observé pour le quadri-couplé. Ce problème devrait être résolu au cours des prochaines semaines : il fait l'objet d'une rencontre prévue début février entre Laure Coquart, ingénieur du CERFACS, et Reiner Vogelsang de SGI Allemagne qui a interfacé la librairie mpp_io du GFDL dans OASIS4 au cours du projet européen PRISM et avec qui la collaboration se poursuit dans le cadre de la « PRISM Support Initiative » (voir http://prism.enes.org/).

[5] Une relance est une partie de simulation s'exécutant en une seule requête sur la machine de calcul.

[6] Cette fonctionnalité de "restart" est requise pour les couplages de type asynchrone où les champs produits par le modèle source durant la période t-1 sont utilisés dans le modèle cible à la période t .

3. Mesures du coût et des performances du coupleur OASIS4

Depuis le dernier rapport, nous avons effectué plusieurs tests pour tenter de quantifier le coût, les performances et la scalabilité du coupleur OASIS4. Rappelons tout d'abord que durant la simulation couplée, OASIS4 agit en tant qu'exécutable séparé, le « Transformeur » (effectuant le calcul des poids des « voisins[7] » d'interpolation et les interpolations proprement dites), et en tant que librairie de couplage, la « PSMILe » attachée aux composantes modèles (effectuant la recherche des « voisins » d'interpolation[8]). Étant donné cette dualité et le fait qu'OASIS4 puisse être utilisé par n'importe quel modèle ayant ses caractéristiques propres de consommation et de performance dans une multitude de configurations différentes (nombre de champs, type de grilles, interpolations, plateforme de calcul, parallélisation des composantes, etc.), il est évidemment impossible de qualifier les performances du coupleur de façon absolue. Nous présentons cependant ici quelques analyses permettant une première évaluation du coupleur dans un environnement particulier.

Pour ces tests, un modèle couplé jouet simplifié basé sur deux composantes jouets a été constitué :

- chacune de ces composantes modèles envoie à l'autre composante puis reçoit de celle-ci un nombre de champs variant selon les tests ;
- les différents champs de couplage dans les composantes sont des « clones » les uns des autres, i.e. ils sont fournis sur une même grille avec même partition et que l'interpolation effectuée par OASIS4 sur ces champs est la même;
- dans chaque test, chaque composante n'utilise qu'une grille, de type « longitude-latitude » pour la première, et de type « longitude-latitude » ou « logiquement rectangle » pour la deuxième[9].
- différents niveaux de parallélisation ont été testés : chaque configuration est identifiable par un triplet X-C-Y où X, C et Y donnent respectivement le nombre de processus de la 1re composante, du Transformeur, et de la 2e composante (e.g. 4-1-4 signifie que les deux composantes tournent chacune avec 4 processus et que le Transformeur tourne avec 1 processus).

Comme les composantes du modèle couplé jouet simplifié sont vides (pas de calculs physiques ni dynamiques), les performances de ce couplé sont un bon indicateur des performances absolues du coupleur ; ses performances relatives en situation réelle de couplage de modèles climatiques ne pourront être que supérieures.

Ces tests ont été réalisés sur le NEC-SX8 de Météo-France. Afin d'assurer la reproductibilité de nos expériences (stabilité des métriques pour une expérience donnée) et d'éviter toutes les perturbations générées par les éventuels conflits d'accès à la mémoire partagée d'un nœud, nous avons choisi de travailler en « nœuds dédiés » ; ceci implique cependant que le dernier nœud réservé (8 processeurs) est bien souvent sous-utilisé.

---

[7] Pour chaque point cible, les « voisins » désignent les différents points sources utilisés pour l'interpolation de ce point.

[8] Pour chaque point cible, les « voisins » désignent les différents points sources utilisés pour l'interpolation de ce point.

[9] Lorsque les composantes utilisent la même grille, l'échange est direct sans interpolation et donc sans passer par le Transformeur.

3.1 Consommation des différentes routines d'OASIS4 (PSMILe et Transformeur)

L'outil de « profiling » FTRACE du NEC-SX8 a premièrement été utilisé dans différents configurations pour identifier les routines les plus consommatrices. Le tableau 1 donne les détails des 5 configurations testées :

Test	Nbr de champs	Taille de la grille	Nbre de pas de temps
C1	1	44x44	24
C2	10	44x44	24
C3	1	490x420	24
C4	10	490x420	24
C4_long	10	490x420	2400

Tableau 1 Tests pour évaluer la consommation des routines OASIS4 (PSMILe et Transformeur)

L'analyse des profils obtenus montre que les routines de lecture XML, appelées au début de chaque relance[5] pour lire l'information de configuration contenue dans les fichiers XML fournis par l'utilisateur, sont de loin les plus consommatrices. Le tableau 2 donne le temps CPU passé dans ces routines de lecture XML ainsi que la proportion de ce temps CPU par rapport au temps CPU total, pour les différentes configurations testées.

	C1	**C2**	**C3**	**C4**	**C4_long**
Temps CPU lecture XML (ms)	344	8050	346	8167	8021
Temps lecture /temps total (%)	19	59	5	27	1

Tableau 2 – Temps CPU absolu et relatif des routines de lecture XML

Tout à fait logiquement, le temps CPU de lecture des fichiers XML augmente avec le nombre de champs de couplage car l'information de configuration XML est particulière à chaque champ de couplage. Comme le CPU utilisé pour le reste des calculs est plus faible pour des petites grilles et pour des relances courtes, la proportion du CPU utilisé pour les routines de lecture XML peut devenir important dans ces cas-là (jusqu'à 59% pour notre cas C2). Ces pourcentages résultant de tests basés sur des composantes jouets « vides », ils ne sont évidemment pas représentatifs d'un modèle couplé réel pour lequel le CPU utilisé dans le calcul de la physique ou de la dynamique sera important. Il conviendra donc de faire le même genre de mesures dans les modèles couplés réels avant de conclure si le coût CPU relatif des routines de lecture XML est important ou pas. Si tel était le cas, ces tests nous permettent de conclure qu'il faudrait optimiser en premier lieu les routines de lecture XML ou de prévoir des façons plus efficaces de stocker et réaccéder à l'information de configuration d'une relance à l'autre.

3.2 Coût des échanges MPI et de la librairie de communication d'OASIS4

Le surcoût des échanges MPI ou de la librairie de communication d'OASIS4 ne peut être établi de façon absolue car, comme expliqué ci-haut, il dépend de la configuration d'un modèle couplé réel donné et de la performance de ses composantes modèles. A titre indicatif, nous donnons quand même dans le tableau 3, le coût mesuré sur le NEC SX8 de la routine d'envoi du PSMILe (PRISM_PUT), de la routine de réception du PSMILe (PRISM_GET), de l'envoi MPI seul (MPI_BSEND appelé par le PRISM_PUT), de la réception MPI seule dans le modèle cible (MPI_RECV appelé par le PRISM_GET), ainsi que les rapports PRISM_PUT/MPI_BSEND et PRISM_GET/MPI_RECV, pour différentes tailles de champ.

	44x44	**96x72**	**260x140**	**490x420**	**700x600**	**825x756**
MPI_BSEND (s)	1,69E-05	1,76E-05	9,03E-05	1,69E-04	2,52E-04	2,95E-04
MPI_RECV (s)	6,09E-06	7,30E-06	1,60E-05	5,45E-05	1,03E-04	1,49E-04
PRISM_PUT (s)	9,60E-05	1,01E-04	1,56E-04	3,90E-04	7,04E-04	9,77E-04
PRISM_GET (s)	9,63E-05	1,06E-04	2,07E-04	6,83E-04	1,27E-03	1,82E-03
PRISM_PUT/ MPI_BSEND	5,7	5,8	1,8	2,3	2,8	3,3
PRISM_GET/ MPI_RECV	15,8	14,5	12,9	12,5	12,3	12,2

Tableau 3 - Coût des routines d'envoi MPI (MPI_BSEND), de réception MPI (MPI_RECV), d'envoi du PSMILe (PRISM_PUT), de réception du PSMILe (PRISM_GET), et rapport des coûts.

Pour avoir une idée du surcoût de la communication par la PSMILe d' OASIS4 et/ou de la communication MPI seule pour un modèle couplé en particulier, il faudra en principe multiplier ce temps par le nombre de champs de couplage et par le nombre de communications. Notons cependant que ces temps peuvent évidemment varier en fonction de la machine, de l'implémentation MPI, du nombre de nœuds utilisés.

On constate cependant que le PRISM_PUT a un coût entre 2 et 6 fois plus élevé que le MPI_BSEND seul et que le PRISM_GET a un coût entre 12 et 16 fois plus élevé que le MPI_RECV. Ce surcoût PRISM vs MPI est plus faible pour les gros messages (ce qui est cohérent). Ce surcoût est lié aux nombreux tests faits sous le PRISM_PUT et le PRISM_GET qui assurent la flexibilité de la communication en fonction des spécifications de l'utilisateur: transformations locales, fréquence de couplage, source/cible de la réception/envoi, etc. Ces chiffres peuvent a priori paraître élevés mais il convient de rappeler que c'est la relation entre le coût absolu de ces routines et le coût des composantes d'un modèle couplé réel qui déterminera le surcoût relatif du couplage.

3.3 Optimisation du traitement des champs de couplage

Comme les différents champs de couplage produits par une composante modèle du couplé jouet simplifié sont des « clones » les uns des autres (voir ci-haut), nous avons profité des tests de performance pour vérifier l'optimisation de leur traitement par le coupleur. Dans ces tests, la première composante utilise une grille de type « longitude-latitude » avec 540x460 points et la seconde une grille de type « logiquement rectangle » avec 489x421 points.

La recherche par la librairie PSMILe des adresses des « voisins[10] » ainsi que le calcul par le Transformeur des poids associés ne devraient être faits qu'une seule fois pour tous les champs « clones ». Ainsi le calcul pour un champ de couplage donné ne devrait donc pas se refaire au 2e échange de ce champ, ni même au 1er échange si ce champ est un « clone » d'un autre champ pour lequel le calcul a déjà été fait.

Nous avons donc mesuré les différents temps passés par le Transformeur dans la routine d'interpolation (prismtrs_interp.F90) pour les différents champs de couplage. Il s'avère que, pour un champ donné, le temps requis à partir du 2e échange (4,2 ms en moyenne) est sensiblement inférieur au temps du 1er échange (5,3 ms en moyenne). Par contre, le temps requis au 1er pas de temps des champs « clones » est également de 5,3 ms ce qui indique un manque d'optimisation des calculs associés aux « clones ». Cette optimisation, importante dans des couplages réels comportant habituellement plusieurs champs clones, sera effectuée aux cours des prochaines semaines.

3.4 Scalabilité

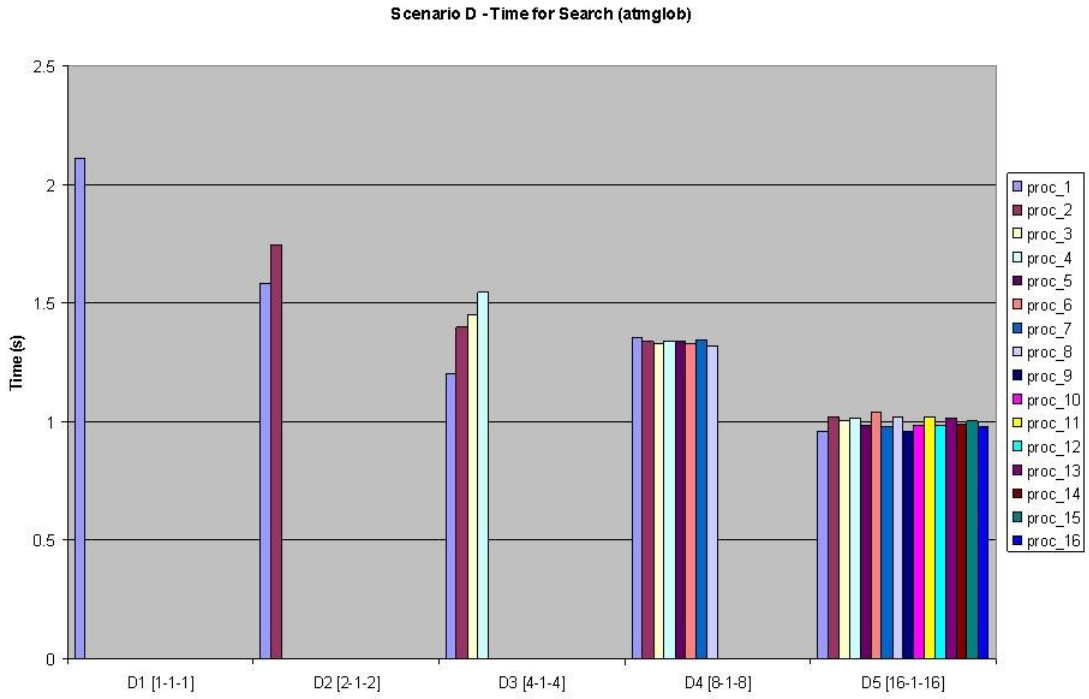
La première composante modèle des tests décrits ici utilise une grille de type « longitude-latitude » avec 540x460 points et la seconde une grille de type « logiquement rectangle » avec 489x421 points.

Scalabilité du PSMILe dans la phase d'initialisation

Le travail le plus important de la librairie de communication PSMILe dans la phase d'initialisation est l'établissement des schémas de communication basés sur la recherche, pour chaque point de chaque partition cible, des voisins d'interpolation sur les partitions sources. La figure 5 donne les mesures du temps passé dans la phase d'initialisation par les différents processeurs d'une composante modèle et ce pour une parallélisation des composantes sur 1, 2, 4, 8 et 16 processus, le Transformeur lui ne tournant qu'avec un seul processus.

On observe effectivement que le temps CPU par processus diminue avec la parallélisation ce qui nous permet de conclure à un bon fonctionnement parallèle de la librairie de communication PSMILe mais le gain (facteur 2 pour une parallélisation d'un facteur 16) n'est pas très élevé, probablement parce que la phase d'initialisation comporte une partie non-négligeable de communications entre les librairies PSMILe sources et cibles (échange des enveloppes des domaines pour le calcul des intersections) et avec le Transformeur (transmission de l'information issus de la recherche des « voisins »), celui-ci agissant comme un goulot d'étranglement de la simulation comme démontré ci-après. Il faudrait aussi refaire ce genre de tests avec des grilles plus grosses donc plus sujette à une parallélisation efficace du point de vue du rapport entre le domaine de calcul local et les communications requises.

10



Nbr de nœuds : 1 1 2 3 5

Figure 5. Temps CPU en secondes passé dans la phase d'initialisation (recherche des « voisins » pour l'interpolation) pour chaque processus d'une composante modèle du couplé jouet simplifié pour une parallélisation avec 1, 2, 4, 8, et 16 processus.

Scalabilité globale du modèle couplé jouet simplifié

La figure 6 donne le temps d'horloge ou « elapse » (courbe rose) pour une simulation de 24 pas de temps, pour une parallélisation sur 1, 2, 4, 8 et 16 processus des composantes modèles, le Transformeur lui ne tournant toujours qu'avec un seul processus. Le temps CPU du Transformeur (« USER Oasis4 » en marron) et le temps CPU moyen des processus des modèles (« USER Psmile Average » en vert) sont également montrés. On se rend compte que si le temps d'horloge (courbe rose) baisse en passant de la configuration 1-1-1 à 2-1-2, il augmente pour 4-1-4 puis se stabilise. L' augmentation entre 2-1-2 et 4-1-4 est très probablement associée au passage de 1 à 2 nœuds utilisés . La communication inter-nœuds passant par réseau interne haut débit (IXS), les communications sont donc moins efficaces qu'au sein du même nœud. Pour les configurations 8-1-8, et 16-1-16 le nombre de nœuds utilisés passe à 3 et 5 ce qui nuit à la performance globale du modèle couplé. Mais de façon plus intéressante encore, on se rend compte que plus la parallélisation des composantes augmente, moins le Transformeur tournant toujours sur un processus est efficace (courbe marron), probablement surchargé par les communications. On en conclut que le Transformeur agit comme goulot d'étranglement de la simulation et que sa parallélisation est indispensable pour assurer la scalabilité du modèle couplé jouet en général.

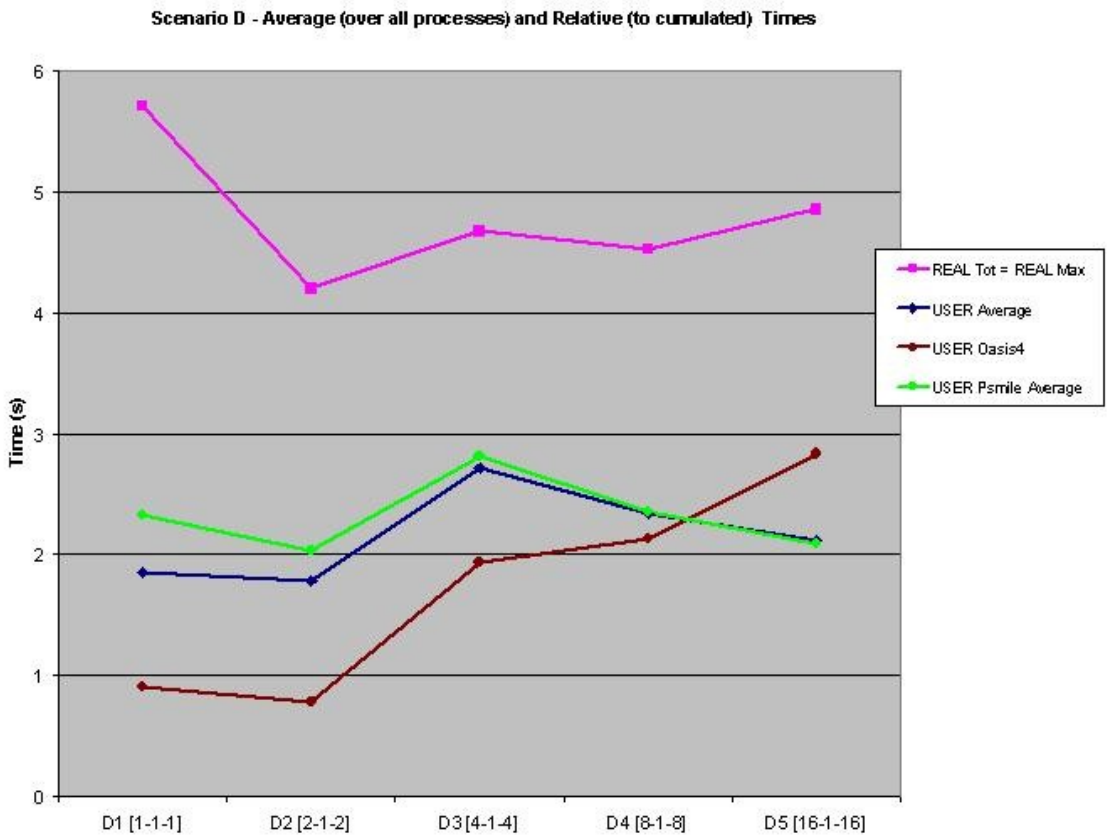


Figure 6 – Temps d'horloge (en rose), temps CPU pour le Transformeur (en marron), temps CPU moyen pour les processus des composantes modèles (en vert) pour différentes configurations de parallélisation : 1-1-1, 2-1-2, 4-1-4, 8-1-8, 16-1-16.

Ce résultat est confirmé par l'étude de l'effet de la parallélisation du Transformeur. Quatre configurations supplémentaires (8-1-8, 8-2-8, 8-4-8 et 8-8-8) ont été testées et les résultats sont reportés à la figure 7

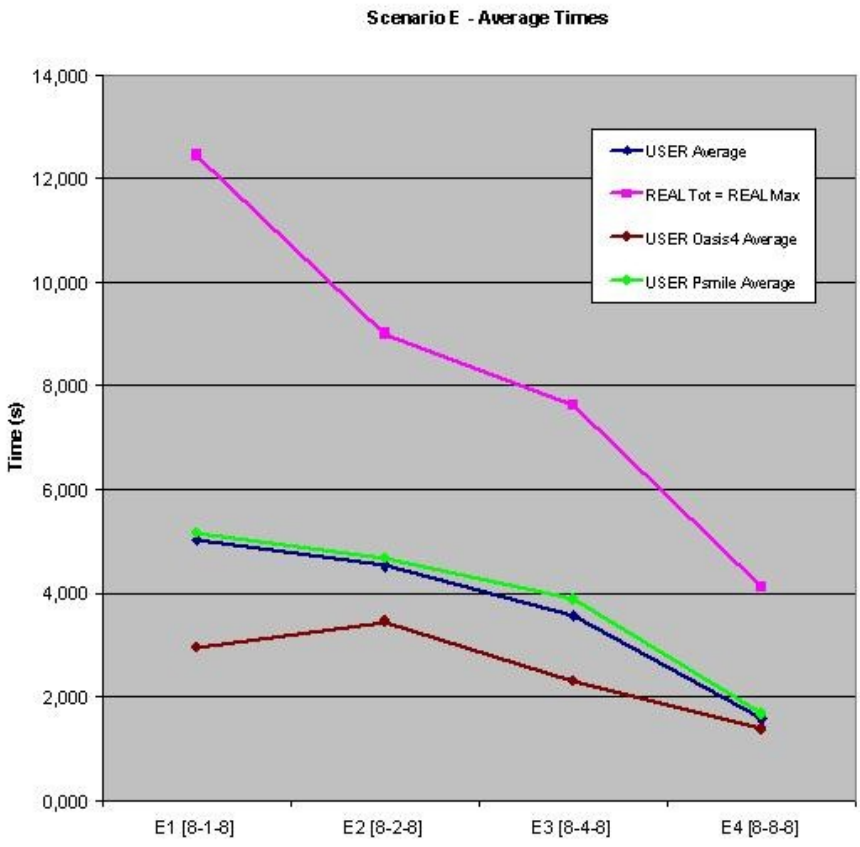


Figure 7 – Temps d'horloge (en rose), temps CPU moyen pour les processus du Transformeur (en marron), temps CPU moyen pour les processus des composantes modèles (en vert) pour différentes configurations de parallélisation : 8-1-8, 8-2-8, 8-4-8, 8-8-8.

Ces configurations ont été choisies car elles permettent de comparer les résultats de simulations tournant toujours sur 3 nœuds dédiés. On voit clairement le bénéfice apporté par la parallélisation du Transformeur faisant sauter le goulot d'étranglement que constituait un Transformeur monoprocesseur : le temps d'horloge diminue d'un facteur 3 en passant de la configuration 8-1-8 à 8-8-8.

De façon cohérente, quand on augmente à la fois la parallélisation du PSMILe et du Transformeur (figure 8), on diminue bien le temps d'horloge du modèle couplé (environ 6 sec pour 1-1-1 puis 4,5 sec pour 2-2-2 ; ou 5,5 sec pour 4-4-4, puis 4,5 sec pour 8-8-8, puis 4 sec pour 16-16-16) sauf quand on passe d'une configuration pour laquelle on n'utilise qu'un seul nœud à une configuration où l' on en utilise plusieurs et où les communications passent donc par le réseau interne, par exemple le passage de 2-2-2 à 4-4-4.

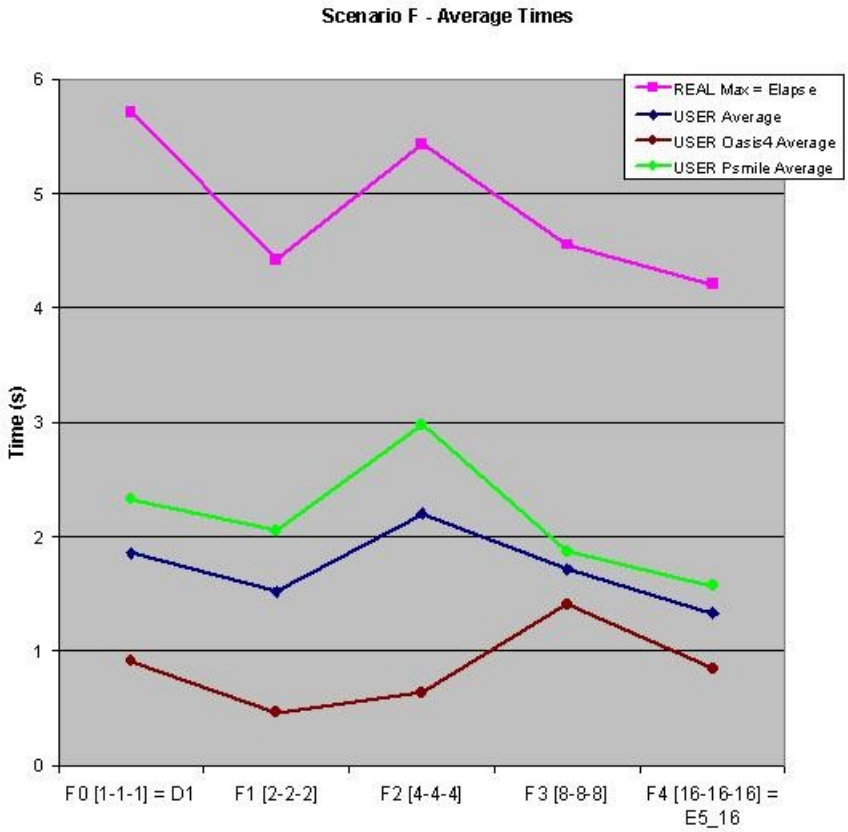


Figure 8 – Temps d'horloge (en rose), temps CPU moyen pour les processus du Transformeur (en marron), temps CPU moyen pour les processus des composantes modèles (en vert) pour différentes configurations de parallélisation : 1-1-1, 2-2-2, 4-4-4, 8-8-8, 16-16-16.

4. Résumé et conclusions

Une première version parallèle du coupleur OASIS4 (2e partie du livrable 4.2), ainsi qu'une version pseudo-parallèle d'OASIS3, sont disponibles depuis 6 mois (voir le rapport à t0+18 mois) pour permettre l'assemblage des composantes des modèles couplés et d'en gérer les échanges et la synchronisation. Les principales fonctions à intégrer dans le coupleur OASIS4 identifiées au mois 12 (livrable 4.1) sont aujourd'hui implémentées sauf :

- La recherche parallèle globale des « voisins » pour l'interpolation 2D conservative (bien que cette interpolation fonctionne correctement en monoprocesseur) : cette tâche est relativement complexe et sera traitée avec priorité dans les prochains mois.
- La possibilité de lire et utiliser un fichier de poids et d'adresses prédéterminé par l'utilisateur : un consultant sera engagé sur des fonds propres CERFACS pour mener à bien cette tâche d'ici les 6 prochains mois.

La qualité de toutes les interpolations implémentées a été analysée dans le détail pour les grilles spécifiques du projet et a révélé, en mode monoprocesseur, certains problèmes en cours de résolution. Ces problèmes sont détaillés à la section 1 ci-

haut. Les premiers tests de validation en parallèle effectués se sont révélés très positifs et n'ont pas montrés de problèmes particuliers supplémentaires pour les interpolations de type bilinéaire et bicubique pour des grilles « longitude-latitude », « logiquement rectangle ». Pour les grilles « gaussiennes réduites », la recherche parallèle globale est en cours de finalisation.

De plus, tel que détaillé à la section 2, on a exécuté les modèles jouets du modèle couplé IPSLCM4 et du quadri-couplé du CNRM-GAME en mode monoprocesseur et en mode parallèle dans un environnement Linux et sur le NEC SX8 de Météo-France, en activant toutes les fonctionnalités requises pour les véritables modèles couplés. Ces tests ont révélé 2 problèmes particuliers, l'un qui apparaît quand on active à la fois de l'interpolation 2D conservative et un autre type d'interpolation pour des champs différents donnés sur une même grille, l'autre se révélant lorsqu'on utilise un nombre important de fichiers de restart dans le modèle jouet du quadri-couplé. Ces problèmes sont en cours de résolution.

Finalement, nous avons procédé à des mesures du coût et des performances du coupleur OASIS4 détaillées dans la section 3. Comme OASIS4 peut être utilisé par n'importe quel modèle dans une multitude de configurations de couplage et matérielles différentes, il est évidemment impossible de qualifier les performances du coupleur de façon absolue. Ces mesures qui ont été effectuées sur le NEC-SX8 de Météo-France nous permettent quand même de porter les conclusions suivantes :

- Les analyses de « profiling » nous permettent de conclure que les premières routines d' OASIS4 à optimiser, si cela s'avérait nécessaire, sont celles de lecture des fichiers de configuration XML (voir la section 3.1).
- Le coût absolu d'une communication MPI et de la surcouche de communication du PSMILe d' OASIS4 a été quantifié pour cette plateforme. Le coût relatif dépend bien évidemment de la configuration du couplage et des caractéristiques des composantes modèles (voir la section 3.2)
- Il est nécessaire d'optimiser le traitement des champs de couplage dans le PSMILe et dans le Transformeur en détectant les champs « clones » et en ne répétant pas les calculs des poids et des adresses pour ces champs (voir la section 3.3)
- Un bon comportement du coupleur en parallèle a été observé : de façon générale, le temps d'horloge d'une relance du couplé jouet simplifié diminue significativement quand le niveau de parallélisation du PSMILe (figure 5) et du Transformeur (figure 7) augmente. Les tests ont également montré que si le Transformeur n'est pas parallélisé, il agit comme goulot d'étranglement de la simulation (figure 6). Sur la machine NEC-SX8, le temps d'horloge augmente cependant significativement quand on passe d'une configuration n'utilisant qu'un seul nœud à des configurations utilisant plusieurs nœuds pour lesquelles les communications inter-nœuds passent par le réseau interne haut débit IXS (figure 8).

Conformité de l'avancement aux prévisions

L'avancement des travaux et la consommation des ressources sont-ils conformes aux prévisions ? Dans la négative, pour quelles raisons ? Quelles mesures ont ou vont être prises pour palier cette situation ? Faut-il revoir le contenu du projet ? Faut-il revoir le calendrier du projet ?

La consommation des ressources pour les embauches est conforme aux prévisions avec le recrutement d'un ingénieur d'études du 1er mars au 31 décembre 2007. Par contre, les dépenses en mission sont en-dessous des prévisions avec un montant

total pour les 2 premières années d' environ 4400 Euros au lieu des 8000 Euros prévus.

L'avancement des travaux est globalement conforme aux prévisions, à quelques réserves près. Les fonctionnalités identifiées dans le rapport OASIS du mois 12 (livrable 4.1) non encore intégrées dans OASIS4 sont (voir la section 1. ci-dessus) :

- La recherche parallèle globale des « voisins » pour l'interpolation 2D conservative: cette tâche est relativement complexe et sera traitée avec priorité dans les prochains mois. Un recul de quelques mois de la date officielle de fin de projet (correspondant à la date effective de démarrage + 36 mois) serait, dans ce contexte, tout particulièrement apprécié.
- La possibilité de lire et utiliser un fichier de poids et d'adresses prédéterminé par l'utilisateur : comme cette tâche n'a pu être effectuée comme prévu dans les 6 derniers mois, un consultant sera engagé sur des fonds propres CERFACS pour mener à bien cette tâche d'ici le prochain rapport.

Pour pouvoir délivrer une version complètement satisfaisante du coupleur, il faudra également compléter la résolution des problèmes résiduels identifiés pour les grilles particulières des modèles du projet et ceux identifiés avec les modèles jouets du modèle couplé IPSLCM4 et du quadri-couplé du CNRM-GAME (voir le résumé à la section 4. ci-dessus).

De toutes façons, l'assemblage des modèles couplés réels IPSLCM4 v3 et du CNRM-GAME peuvent se faire avec les versions actuellement disponible d'OASIS4 et d'OASIS3. La validation complète d'OASIS4 peut s'effectuer en parallèle et ne retarde donc pas le déroulement du projet.

De plus, le travail effectué durant la dernière période sur le coût et les performances du coupleur OASIS4 nous a permis d'aborder des questions d'optimisation, de parallélisme et de scalabilité, aspects tout à fait indispensables dans un projet tel que CICLE financé dans le programme « Calcul Intensif et Grilles de Calcul », même si aucun livrable n'y était directement attaché.

Difficultés rencontrées par le partenaire

La remarque faite dans le rapport à t0+18 est toujours valable : la phase de validation des fonctionnalités d'OASIS4, en particulier celles d'interpolation pour les grilles du projet CICLE, a entraîné plus d'ajustements qu'initialement prévu. Le travail a cependant bien progressé, les problèmes particuliers sont résolus un par un, et les premiers résultats de la validation des interpolations en parallèle sont prometteurs. Ce surplus de travail ne nous a pas permis d'aborder certains aspects sur lesquels nous comptions travailler durant la dernière période, en particulier la possibilité de lire et utiliser un fichier de poids et d'adresses prédéterminé par l'utilisateur, tâche pour laquelle nous prévoyons maintenant d'embaucher un consultant sur des fonds propres CERFACS (la personne est déjà identifiée et un accord de principe a été conclu.)

Prévision des travaux du partenaire pour la prochaine période
Résumer les travaux prévus et les résultats / livrables escomptés. Identifier les risques éventuels.

Tel que détaillé dans ce rapport, nous comptons poursuivre le développement d'OASIS4, en particulier :

- compléter la résolution des problèmes d'interpolation identifiés en mode monoprocesseur pour les grilles particulières des composantes modèles du projet (voir la section 1.);

- finaliser la validation des interpolations parallèles (voir la section 1)
- implémenter la possibilité de lire et utiliser un fichier de poids et d'adresses prédéterminé par l'utilisateur (voir la section 1.4);
- résoudre les 2 problèmes révélés par les modèles couplés jouets, à savoir celui lié à l'utilisation combinée de l'interpolation 2D et celui lié à l'utilisation d'un grand nombre de fichiers de restart (voir la section 2);
- optimiser le traitement des champs de couplage « clones » (voir la section 3.3)
- interagir avec nos collaborateurs de NEC-CCRLE à Sankt-Augustin en Allemagne pour finaliser la recherche des « voisins » parallèle globale pour la grille gaussienne réduite et implémenter celle pour l'interpolation 2D conservative (voir section 1).

Nous comptons de plus porter OASIS4 et en particulier le modèle couplé jouet IPSLCM4 sur le cluster Bull (processeurs Ithanium) du CEA. Laure Coquart, ingénieur du CERFACS, suit actuellement au CEA une formation sur l'utilisation de cette plateforme.

Finalement, nous prévoyons apporter un support utilisateur actif à l'IPSL et au CNRM-GAME pour la réalisation des livrables 2.2 et 3.3

Aspects non scientifiques

Le cas échéant, liste des CDD recrutés par des établissements publics dans le cadre du projet

Nom	Prénom	Qualifications	Date de recrutement	Durée du contrat (en mois)
Moine	Marie-Pierre	Ingénieur d'études	01/03/2007	10
…				

Le cas échéant, modalités d'utilisation du complément de financement « pôles de compétitivité »
N/A

Le cas échéant, équipements achetés par les partenaires dans le cadre du projet
Lister ici tous les équipements achetés depuis le début du projet

Désignation	Date d'achat	Prix d'achat (en Euros)	Part financées par l'aide ANR (en Euros)
DELL écran 20 pouces	**25/08/06**	**470,00**	**235,00**
ICONCEPT Macbook Pro Intel Dual Core	**31/10/06**	**2849,78**	**1424,89**
Poste precision (2) 390 core duo processeur E6400	**08/02/2007**	**3 280,00**	**1 571,00**