Développement du système couplé IPSL et partage de données
Institut Pierre Simon Laplace

DESCRIPTION DU PROJET DE RECHERCHE

Résumé

Notre demande de ressources informatiques nationales est un projet technologique complémentaire des projets scientifiques des laboratoires de l'IPSL déposant des demandes séparément. Il est composé de 2 parties : une composante développement/validation du modèle climat, de son environnement et de la diffusion de ses résultats depuis les centres de calcul et une composante gestion des données communes. Depuis 2013, nous intégrons dans ces données les résultats des exercices d'intercomparaison de type CMIP5.

2- Présentation générale

L'IPSL est une fédération de laboratoires travaillant sur l'environnement et depuis 2012 un Labex. Ses études se structurent autour des observations, des études de processus et la modélisation. La mise à disposition de **grands jeux de données** sur les centres de calcul nationaux est une composante essentielle de ses activités.

Le pôle de modélisation du climat de l'IPSL fédère une centaine de scientifiques autour de la modélisation du climat. Le **modèle climat de l'IPSL** en est une composante structurante. Ce modèle inclut régulièrement les développements scientifiques et techniques nécessaires à la modélisation du climat. C'est un travail de longue haleine qui permet la mise au propre des versions successives du modèle, la réalisation de simulations de référence, leurs validations et leurs diffusions.

Le projet déposé ici permet d'avoir quelques ressources pour le développement des outils et pour les tester en grandeur réelle. Il demande également l'accès aux différents centres de calcul nationaux pour anticiper les difficultés de portage et mettre en place un environnement de calcul cohérent sur les différents calculateurs utilisés. Il se place en support des projets scientifiques comme gen6178 et gen2211 qui exploitent pleinement le modèle climat en particulier pour les exercices de type CMIP5 de coordination internationale des simulations climatiques (GIEC AR5).

Le comité scientifique du pôle de modélisation souhaite favoriser une stratégie informatique qui permette :

- des analyses croisées des simulations,
- des comparaisons avec des jeux de données importants,
- une diffusion des résultats lors de coopération scientifique nationale et internationale,
- un renforcement des coopérations interdisciplinaires

L'expérience a montré que le mode de fonctionnement consistant à analyser les expériences en local, tout en maintenant un archivage sécurisé au niveau des grands centres est le meilleur mode de fonctionnement possible. Il restait à trouver le moyen le plus simple, le plus efficace et le plus sûr de rendre accessibles les résultats de simulation présents à l'IDRIS, au CCRT auprès des clusters de laboratoires.

Il a été décidé par le passé de favoriser le déploiement de serveurs OpenDAP/DODS au sein de l'IPSL (LMD Polytechnique et Jussieu, IPSL Jussieu) et de demander au CCRT et à l'IDRIS dans un premier temps de mettre en place un accès par OpenDAP/DODS à certains résultats sélectionnés. Les serveurs dods.idris.fr et dods.extra.cea.fr existent, les résultats sont concluants en terme de performance et de qualité de services, deux types d'accès étant possibles : accès large ou accès restreint à l'IPSL seulement.

La technique a évolué et la méthode de diffusion des données depuis les centres de calcul repose maintenant sur les sur-couches logicielles du projet Américain Earth System Grid Federation (ESGF). Dans le cadre du projet Prodiguer (resp. S Denvil), deux datanodes ont été installé au CCRT et à l'IDRIS.

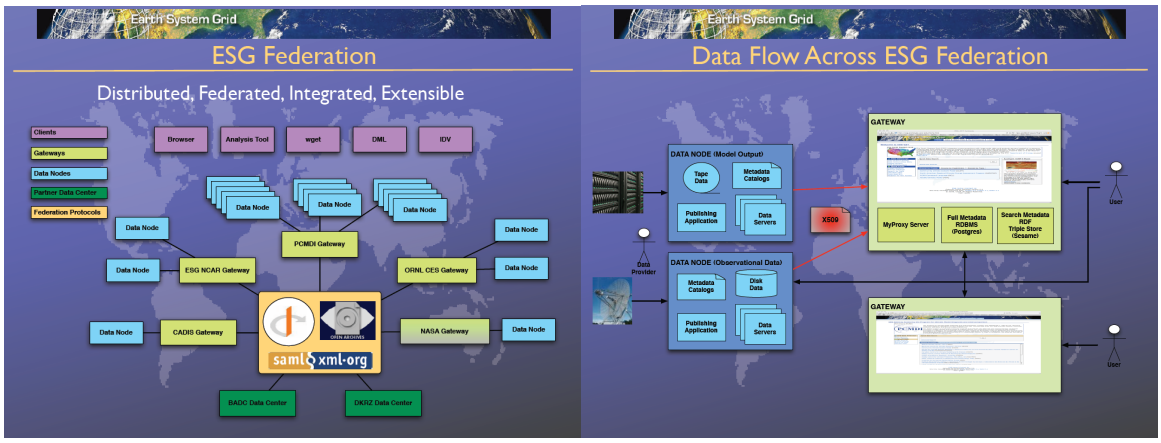

Figure a: une fédération de site distribuée, dont le nombre de nœuds de données est extensible.
Figure b: des données accessibles à travers différents portails communautaire sécurisé.

La sur-couche logicielle ESGF présente sur les nœuds de données comprend :

- une base de données PostGreSQL,
- un serveur apache/tomcat dont dépendent maintenant les serveurs OpenDAP et Thredds (qui gèrent le contrôle d'accès et la sécurité en complément d'un serveur myproxy),
- les « web applications » OpenDAP et Thredds,
- un serveur MyProxy de management des autorisations,
- un serveur gridftp,
- une application permettant de publier vers plusieurs portails les caractéristiques des données disponibles et de la configuration modèle ayant produit ces données.

La gestion des autorisations d'accès se fait à travers l'émission de certificat X509 par le serveur MyProxy en fonction des droits associés à une identité OpenID.

Cette stratégie présente de nombreux avantages :

- valorisation des résultats stockés à l'IDRIS et au TGCC/CCRT (environ 40% des fichiers stockés à l'IDRIS et au TGCC/CCRT sont accessibles par OpenDAP/DODS/Hyrax, sans recopie) à travers une visibilité accrue,
- fédération de sites géographiquement distribués accessible via plusieurs portails dont l'IPSL,
- contrôle d'accès aux données, notion de communauté virtuelle,
- utilisation de standards incontournables en sciences de l'environnement,
- base de données interrogeable par plusieurs communautés identifiées,
- élargissement sensible de la communauté des utilisateurs,
- suivi avec un haut niveau de granularité de la diffusion des résultats,
- coopération accrue entre les centres,
- utilisation scientifique du réseau déployé par RENATER

Les ressources demandées dans ce projet permettront les activités suivantes :

1) Aide au développement des composantes

- maintien de MODIPSL/libIGCM structure portable de développement des modèles
- gestion des évolutions logicielles
- aide à la mise en place des environnements
- suivi et diffusion des nouveautés techniques (compilateurs, évolutions des calculateurs, ...)
- suivi de la qualité numérique
- parallélisation en mode MPI (mémoire distribuée), OpenMP (mémoire partagée) et mixte MPI/OpenNP.

2) Couplage des modèles existants

- études et choix des interfaces
- mise en place de scripts de lancements
- maintien de la portabilité sur les plates-formes accessibles
- optimisation
- mise en œuvre de nouvelles résolutions horizontales et verticales

3) Environnement de production

- Développement, consolidation et maintien de libIGCM : bibliothèque de script de lancement de simulation
- Développement et maintien des scripts d'atlas de post-traitements systématiques et de leur base commune (fast, surcouche ferret)
- Réalisation de simulations 'trusting' permettant de suivre les évolutions du centre de calcul et d'évaluer leur impact. Tous les 2 jours : 3 mois de simulation tournent sur chaque machine de chaque centre. Toutes les semaines : 10 ans de simulations, post-traitement inclus, tournent sur les machines adaptées à la production

4) Réalisation de quelques simulations en production afin de quantifier les améliorations

- Réalisation de quelques simulations en production (100 à 200 ans) afin de détailler les améliorations des différentes évolutions.

5) Diffusion des résultats

- mise à disposition des résultats (et de leurs descriptifs) par OpenDAP/DODS/hyrax et par ESGF au TGCC/CCRT et à l'IDRIS.
- Duplication des résultats vitaux, transferts en masse entre le TGCC et l'IDRIS, demande d'utilisation de la liaison très haut débit DEISA/PRACE.

Travail prévu en 2013

1) Modèle climat :

- Maintien en état de marche performante de la version courante de la chaine couplée : IPSLCM5A basée sur les modèles d'océan (NEMO), d'atmosphère (LMDZ), de glace (LIM2 et LIM3 au choix) et de surfaces continentales (ORCHIDEE) y compris les configurations forcées de chacune des composantes
- Déploiement sur tous les calculateurs disponibles, analyses des performances, diffusion des bonnes pratiques
- Validation des versions incluant de nouvelles composantes : chimie complète, ...
- Mise en œuvre de différentes résolutions
- Portage et optimisation sur ada
- Portage et optimisation sur turing

2) Environnement de production/post-traitement

- Consolidation des chaines de calcul pour une meilleure robustesse
- Mise en œuvre standardisées des simulations d'ensembles. La soumission, les analyses de ce type de simulation impliquent des traitements de données en plus grande masse puisque au lieu d'une simulation, il s'agit d'un jeu d'une dizaine de simulations lancées en ensemble qu'il faut analyser de même.
- Installation et validation des outils de post-traitement basés sur Python/CDAT.
- Intégration dans l'environnement de production des simulations (libIGCM) des informations nécessaires à ESGF : mise à jour de la base de données, extraction de l'information utile (modèle utilisé, typologie de la simulation, données disponibles). C'est de fait une extension aux fonctionnalités Dods déjà présentes.
- Portage et optimisation sur ada
- Portage et optimisation sur turing
- Poursuite de la réalisation de simulations 'trusting' permettant de suivre les évolutions du centre de calcul et d'évaluer leur impact. Tous les 2 jours : 3 mois de simulation tournent sur chaque machine de chaque centre. Toutes les semaines : 10 ans de simulations, post-traitement inclus, tournent sur les machines adaptées à la production. Cela représente un total de 550 ans de simulation par machine.
- Mise en œuvre du trusting sur ada.
- Réalisation d'une simulation de production à la résolution 144x142x39

3) Diffusion des résultats

- Choix des résultats à diffuser par OpenDAP/DODS/Hyrax,
- Poursuite de la diffusion des résultats CMIP5

- Duplication des résultats vitaux
- Maintenance et installation des outils ESGF
- Transfert des fichiers CMIP5 du TGCC vers l'IDRIS pour duplication/curation.

Détail des besoins sur les différentes machines IDRIS :
Vargas et estimation ada :
10 ans de simulation en 96x95x39-ORCA2 prend 1300 h sur 32 procs de vargas
550 ans de **trusting** avec IPSLCM5A-LR : 1300h*55 = 71500 h + 25% post = **89375 h**
10 ans de simulation en 144x142x39-ORCA2 prend 4500 h sur 64 procs de vargas
200 ans de **simulation** à la résolution 144x142x39-ORCA2 : 4500h*20 = 9000 h +25%post = **112500 h**

Babel et estimation turing: utilisation possible pour l'océan en forcé seulement, environnement libIGCM pas encore porté. Besoin d'accès pour des tests d'algorithmie

Ressources calcul demandées à l'IDRIS:
201 875 h IBM ada dont 25% pour les post-traitements
100 000 h Blue Gene turing

Autres ressources demandée à l'IDRIS:
- avoir accès aux fichiers **gaya depuis ada** et turing, au moins en lecture
- avoir des espaces de type **WORKDIR de l'ordre de 10 To** pour installer la fonctionnalité pack de regroupement des fichiers limitant le nombre de fichiers produits
- pouvoir lancer **300 jobs de post-traitements mono** sur ada
- utiliser la **liaison très haut débit entre l'IDRIS et le TGCC**

Utilisation du CCRT/TGCC :
En 2013, nous souhaiterions continuer d'utiliser le TGCC/CCRT et l'IDRIS de façon similaire : trusting systématique fréquent et une simulation longue. Sur titane, aucun développement n'est prévu et seules les simulations de trusting sont demandées jusqu'en juin 2013. Nous souhaiterions également faire quelques essais très préliminaires avec les GPUS.

Détail des besoins sur les différentes machines TGCC/CCRT:
Titane (jusque juin 2013):
10 ans de simulation en 96x95x39-ORCA2 prend 1600 h sur 32 procs de titane,
275 ans de **trusting** avec IPSLCM5A-LR : 1600h*27,5 = 44000 h + 25% post = **55000 h**

Curie nœuds fins et nœuds larges :
10 ans de simulation en 96x95x39-ORCA2 prend 1000 h sur 32 procs de curie nœuds fins et 250h sur curie nœuds larges en post-traitements.
550 ans de **trusting** avec IPSLCM5A-LR : 1000h*55 = **55 000 h nœuds fins** et 25% post : **13 750 h nœuds larges**

10 ans de simulation en 144x142x39-ORCA2 prend 2500 h sur 64 procs de vargas et 600h de curie nœuds larges en post-traitements.
200 ans de **simulation** à la résolution 144x142x39-ORCA2 : 2500h*20 = **50 000 h nœuds fins** et **12 500 h nœuds larges**

Post-traitements sur les résultats à distribuer : 150 ans de simulations à la résolution144x142x39-ORCA2 : **1000 heures curie noeuds larges**. On prévoit une **cinquantaine** de simulations à post-traiter sur curie nœuds larges. Ces post-traitements se faisaient auparavant sur cesium. L'arrêt de cesium reporte mécaniquement ce travail sur les machines de calcul. Ceci est donc une première estimation qu'il faudra affiner l'an prochain.

Ressources calcul demandées au CCRT:
55 000 h titane dont 25% pour les post-traitements
10 000 h GPU

Ressources calcul demandées au TGCC :
105 000 h curie noeuds fins pour les calculs
76 250 h curie noeuds larges pour les post-traitements

Autres ressources demandée au TGCC/CCRT:

- avoir des espaces de type **SCRATCHDIR de l'ordre de 20 To**
  - 10 To pour les post-traitements
  - 10 To pour créer des répertoires TMPDIR et mimer leur fonctionnement type IDRIS
- pouvoir lancer **300 jobs de post-traitements** sur curie noeuds larges
- adapter **les procédures batch de curie pour permettre la production** (un grand nombre de jobs qui s'enchainent) et le développement (un petit nombre de jobs qui doivent démarrer vite). Le fonctionnement actuel est très bien adapté au développement mais la production est quasi impossible une fois qu'on atteint de l'ordre de 10 000 h de consommation : un login qui a consommé autant ayant une priorité plus faible.
- **Avoir les quotas suivants,** voir détail de l'utilisation actuelle en annexe :
  - **p86maf , 500 000 sur CCCSTORE (hors déménagement DMFDIR), 3 000 000 sur CCCWORK**
- utiliser la **liaison très haut débit entre l'IDRIS et le TGCC**

Utilisation du CINES :
En 2013, nous souhaiterions poursuivre le portage du système couplé complet IPSL et de ses outils sur la plates-forme de calcul SGI du CINES. Le grand challenge nous avait donné l'occasion d'une première expérience avec un modèle couplé à plus haute résolution mais incomplet en terme de physique. D'autre part, des discussions et des études préliminaires vont démarrer afin d'étudier la faisabilité de l'installation d'un noeud de données ESGF au CINES.

Etat des portages des modèles IPSL sur la machine jade du CINES:
Jade : compilation OK, couplé haute résolution avec composantes réduites, OK. Il reste à construire et valider l'environnement

Ressources demandées au CINES:
50 000 h SGI jade
50 To pour les fichiers communs et les résultats de ce projet

Rappel sur les accords IPSL-IDRIS pour les espaces de stockage
En 1996, l'attribution d'un espace de stockage commun sur le serveur fichier de l'IDRIS et l'ouverture de comptes « support » ont été consenties par l'IDRIS pour les projets IPSL (projet **960826**). Ces attributions ont été reconduites, par l'IDRIS, tacitement les années suivantes (projet actuel **20080110826**).

Le quota de stockage de cet espace est actuellement de 245 To dont 100 To pour le partage de données ERA; il permet de stocker les données nécessaires aux calculs intensifs effectués sur les machines de l'IDRIS par les différentes équipes appartenant à la communauté des laboratoires de l'Institut. Ces équipes impliquées dans plusieurs projets de l'IDRIS ont décidé de regrouper les données utilisées de façon à éviter toute duplication, consommatrice en stockage ; parmi ces données, un effort important a été accompli pour créer une archive commune de données du CEPMMT, utilisées notamment pour le forçage et la validation des modèles couplés et l'assimilation de données. Cet espace est géré par deux ingénieurs responsables de la gestion de ces données et de leur attribution (création d'un super-groupe « subipsl »).
Les données de réanalyses ERA40, ERA-Interim et un sous-ensemble des analyses opérationnelles du CEPMMT (données 2D et 3D) sont archivées et ont été reformatées en Netcdf en utilisant les moyens de calcul de l'IDRIS (ulam). L'ensemble des données du CEPMMT (ERA40, ERA-Interim, analyses opérationnelles) sont mises sous OpenDAP/DODS/hyrax pour la communauté des laboratoires IPSL (filtrage par plage d'adresse IP des domaines des laboratoires IPSL).

A partir de 2013, les résultats CMIP5 seront inclus dans cet espace de stockage collectif. Leur distribution se fait au travers de la nouvelle technologie décrite : ESGF.

Parallèlement les comptes « support » bénéficiaient d'une attribution modérée de temps CPU appelée « crédit calcul » sur chacune des plates-formes de l'IDRIS.

Travail prévu en 2013

- L'archivage d'une partie d'une nouvelle réanalyse américaine a débuté courant 2012 (couplée ocean-atmosphère-land surface-glace de mer, à haute résolution, période allant de 1979 au présent) : CFSR (NCEP Climate Forecast System Reanalysis), au format GRIB2. Il faudra consolider cet archivage en fonction des besoins de la communauté IPSL et assurer un reformattage en Netcdf des paramètres nécessaires.
- Gestion du stock des données sur la machine « Gaya », et reconduite du stockage mensuel des données opérationnelles produites par le CEPMMT (niveau 2D et 3D basses et hautes résolutions) soit 34 Go/ mois,
- Traitement des données stockées :
  - mise au format NetCDF,
  - interpolation,
  - calcul de nouveaux paramètres de stockage : vorticité potentielle, norme du vent, …
- Fourniture des données aux ayant-droits IPSL,
- Post-traitement des données déjà archivées.
- Distribution des résultats CMIP5
- Duplication des résultats vitaux
- Transferts en masse des résultats entre le TGCC et l'IDRIS
- Extension à 250 To de l'espace réservé pour les résultats CMIP5, 150 To pour le moment, soit une **augmentation de 100 To.**

Ressources totales demandées en stockage à l'IDRIS : 345 To

Ressources en stockage demandées au TGCC/CCRT :
La distribution des résultats de type CMIP5 correspond à **250 To** stockés et archivés sous le login p86cmip5. Cela représente **500 000 inodes.**

3 Méthode(s)

Algorithmes :

Les modèles utilisés sont basés sur des différences finies. La parallélisation se fait en décomposition de domaines (MPI) et selon la verticale (OpenMP) pour l'atmosphère.

Modalité d'optimisation :

Etudes sur des cas tests, recherche des parties peu performantes, étude des améliorations possibles, diffusion à la communauté utilisatrice des optimisations.

Structure du programme :

Une chaîne couplée est basée sur l'exécution simultanée de 3 commandes (process) : l'atmosphère, l'océan et le coupleur échangeant régulièrement des informations (champs de surface) grâce à une bibliothèque basée sur MPI-2 ou MPI-1. Chacun des modèles peut être parallèle (MPI ou hybride MPI/OpenMP) et lancé sur des nombres différents de processeurs et de tâches.

Logiciels et bibliothèques nécessaires :

Gestionnaire de sources cvs et subversion

Bibliothèque NETCDF

Echange de messages avec le coupleur OASIS : Bibliothèque MPI-2 et/ou MPI-1

Fortran 95

Classe batch permettant le lancement de 3 process simultanés, chacun étant parallèle (MPI et/ou hybride MPI/OpenMP) ou non

Structure de batch autorisant le lancement de 300 jobs mono par utilisateur

Logiciel de dialogue entre IDRIS, CCRT/TGCC et CINES et le serveur fichiers IPSL

Serveurs OpenDAP/DODS/hyrax

Python (sur frontale et supercalculateurs)

libxml (version 2)

autoconf, automake

Outils indispensables de post-traitement et de visualisation, sur ulam principalement :

- netcdf/3.6.3 + 4.1, udunits/2.1.5, ferret/6.1, NCAR NCL, CMOR2
- CDO/1.4.0, NCO/4.0 : version netCDF3 et netCDF4, CDAT/5.1, GRADS/2.0, R/2.8.1
- netpbm, imagemagick, tetex-latex
- RSYNC, VTK, Paraview, gnuplot

Outils pour ESGF :

- une base de données postgres,
- un serveur tomcat dont dépendent maintenant les serveurs OpenDAP (ex DODS) et thredds qui gèrent la sécurité et le contrôle d'accès,
- un serveur gridftp,
- une application permettant de publier vers un serveur portail les caractéristiques des données disponibles.

Quotas p86maf au 15/10/2012

Au 15 octobre 2012, le compte p86maf occupe 133 000 inodes sur l'espace STORE et 984 000 inodes sur l'espace WORK.
Ces fichiers correspondent à la production sur SX9, titane et curie entre les mois d'avril et d'octobre 2012 soit 6 mois de production environ. Le quota actuel de 500 000 inodes sur STORE doit permettre d'accueillir la production de la fin 2012 et de 2013. Rappelons qu'au TGCC/CCRT on a un seul login par personne et qu'on peut choisir sur quel projet imputer les heures de calcul. La production du compte p86maf est ainsi rattachée aux projets gen2211 et gen6178.
Le compte p86maf a un quota de 500 000 inodes sur STORE et de 3 000 000 sur WORK jusqu'au 31/12/2012. Nous souhaitons conserver ces quotas en 2013 et auxquels il faudra ajouter au moment opportun les quotas utilisés par le déménagement DMFDIR.
Sur DMFDIR ce login possède 3 541 000 inodes. Le ménage des fichiers inutiles y a été fait très régulièrement. La taux de réduction prévu par l'outil utilisé pour le déménagement de DMFDIR vers CCCSTOREDIR sera de 12 environ, on peut donc déjà prévoir un besoin de 300 000 inodes supplémentaires pour accueillir le déménagement/package des fichiers DMFDIR de p86maf.

Résumé de l'occupation sur STORE du compte p86maf au 15/10/2012 :

Machine de calcul	Configuration utilisés	Production ou test	Nb années simulées	Nombre de fichiers	Nombre d'inodes
SX9	IPSLCM5A-MR	PROD	700 ans	36 200	36 500
	IPSLCM5A	PROD	700 ans	25 300	25 500
Titane	IPSLCM5A-MR	TEST	20 ans	2 500	2 700
	IPSLCM5A	PROD	450 ans	19 000	19 300
Curie	IPSLCM5A-MR	PROD	130 ans (en cours)	9 400	9 700
	IPSLCM5A	PROD	650 ans	36 700	37 300
Toutes machines	Simulations refaites pour refaire des fichiers perdus	REDO		700	900
Toutes machines	LMDZOR	TEST		200	400
Autres fichiers				600	700
Total				130 000	132 300