Demande d'allocation de temps de calcul pour A9 11/2020-10/2021 05/09/2020

Dossier technique IPSL

Introduction

La plateforme de modélisation climatique de l'IPSL est un ensemble de composantes modélisant différentes parties du système climatique. Ces composantes peuvent être utilisées séparément, ou couplées ensemble, proposant ainsi une grande variété de configurations possibles en fonction des besoins scientifiques.

Ces composantes sont : le modèle de circulation atmosphérique, LMDZ, le modèle de circulation océanique, NEMO, incluant la dynamique océanique, la glace de mer et la bio-géochimie marine, le modèle de surfaces continentales, ORCHIDEE, et les modèles de chimie atmosphérique, INCA pour la troposphère et REPROBUS pour la stratosphère. Les interactions entre les composantes sont décrites dans la Figure 1. Le système est conçu pour être évolutif et d'intégrer ou d'activer facilement de nouvelles composantes, par exemple les calottes de glace. Il inclut également le coupleur OASIS. Les entrées-sorties sont gérées par les bibliothèques IOIPSL (version historique) et XIOS qui est massivement parallèle et inclue un serveur d'IO débranchable. Le système comprend également un ensemble d'outils de gestion des versions (basés sur svn), de compilation, MODIPSL, et un ensemble de scripts d'exécution et de post-traitement, libIGCM, incluant depuis 2017 le *workflow* CMIP. Ces outils sont partagés par l'ensemble des composantes.

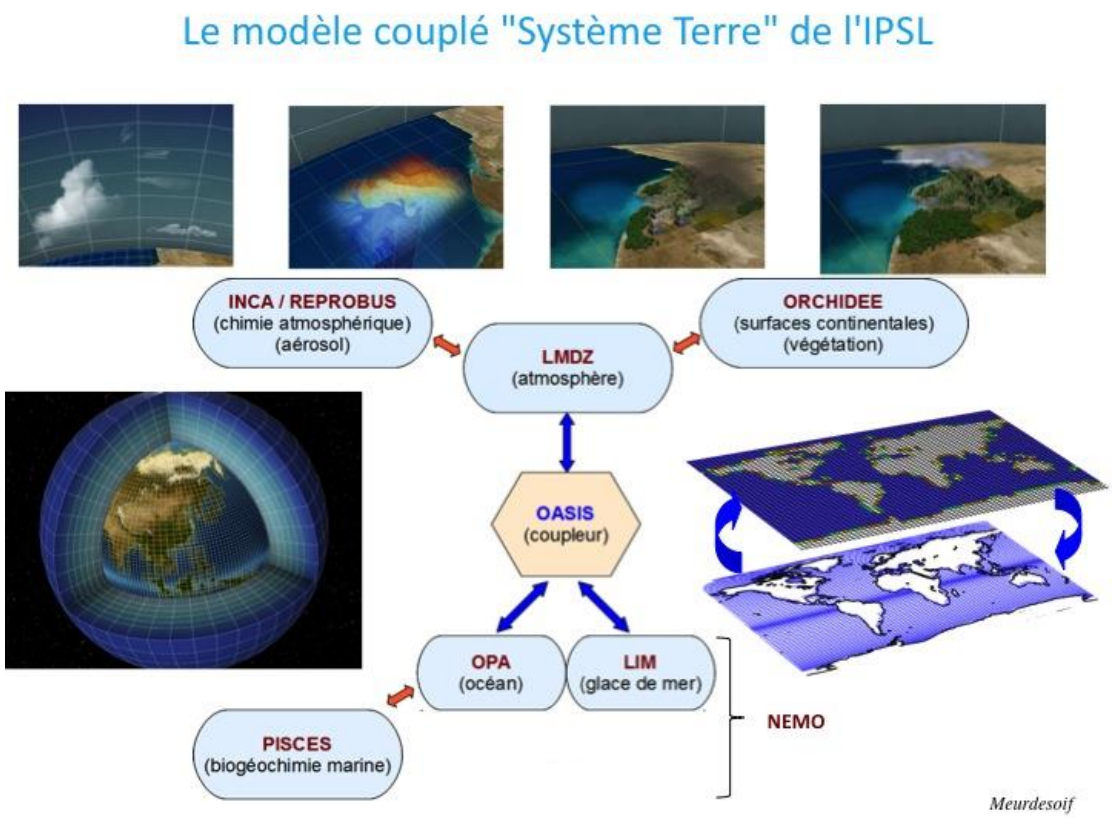


Figure 1.Le modèle système climatique de l'IPSL.

Les configurations supportées aujourd'hui sont :

- IPSLCM6 : le couplé de référence pour les simulations de type CMIP incluant l'atmosphère, les surfaces continentales, l'océan, la glace de mer et le cycle du carbone. Cette configuration propose également deux expériences de type ESM avec un calcul interactif du cycle du CO_2 et un calcul interactif des aérosols atmosphériques.
- IPSLCM5A2 : une version à plus basse résolution que le couplé IPSLCM6, utilisé en particulier pour des simulations longues de paléoclimat.
- LMDZOR : la version atmosphère-continents correspondante.
- LMDZORINCA : la version atmosphère-continents-chimie troposphérique correspondante.
- LMDZREPR : la version atmosphère-chimie stratosphérique correspondante.
- IPSLCM5CHS : le couplé de référence avec chimie stratosphérique.
- IPSLCM5A2CHT : un couplé à très basse résolution correspondant à IPSLCM5A2 mais permettant d'avoir une chimie troposphérique interactive.
- NEMO : la version océanique correspondante incluant la glace de mer et la bio-géochimie marine.
- ORCHIDEE_OL : la version surfaces continentales seules
- ICOLMDZOR_v7 : la version atmosphère-continents incluant le nouveau cœur dynamique sur grille icosaédrique DYNAMICO.

Entre 2018 et 2020, 11 versions de la configuration IPSLCM6.1, ne variant plus en terme de physique et de résultats mais incluant les corrections nécessaires aux variables du *workflow* CMIP6, ont été mises à disposition pour test et évaluation. La configuration IPSLCM6.1.11, figé en décembre 2019, inclut en particulier :

- NEMO v3-6_STABLE et LMDZ6
- La bibliothèque de couplage OASIS3-MCT
- XIOS 2.0 pour toutes les composantes
- La parallélisation hybride MPI/OpenMP pour l'atmosphère, les surfaces continentales et la chimie
- Les deux résolutions associées à IPSLCM5A2 (comme pour IPSLCM5A-LR utilisé pour CMIP5), à IPSLCM6-LR utilisée pour CMIP6 et à IPSLCM6-MR1

Des optimisations ont été intégrées au fur et à mesure de ces versions (appels des paramétrisations physiques à la fréquence optimale, élimination des processeurs sans points océan dans l'océan, version récente du compilateur et options de compilations agressives testées et validées).

Algorithmes utilisés

LMDZ et NEMO/OPA

Ce sont les composantes dynamiques du système. Les deux modèles résolvent les équations dites primitives des fluides géophysiques :

- Equations de Navier-Stokes en repère tournant pour la vitesse horizontale, avec diffusion horizontale turbulente pour représenter les phénomènes sous-maille.

- Pour la verticale : approximation hydrostatique et diffusion turbulente verticale dans la couche limite.
- Conservation de la masse, de l'énergie, du sel. Conservation de la vorticité pour l'écoulement horizontal.
- Pour l'atmosphère : équation de conservation de la vapeur d'eau avec paramétrisation des phénomènes de condensation; également paramétrisations des transferts radiatifs.

Dans les configurations couplées NEMO/OPA-LMDZ, l'essentiel du temps de calcul est consommé par LMDZ lorsque l'océan est à la résolution 2° ou 1° (ORCA2 ou eORCA1). Cela n'est plus le cas pour les résolutions océaniques plus hautes comme le 1/4° (eORCA025) ou plus.

NEMO/PISCES

NEMO/PISCES traite des traceurs océaniques. Il est composé d'un module de transport de traceurs océaniques, qui consomme l'essentiel du CPU, et un module de biogéochimie océanique. La parallélisation de NEMO/PISCES suit celle de la partie dynamique NEMO/OPA. NEMO/PISCES peut être utilisé *off-line* (la dynamique est lue dans des fichiers) ou *in-line* (un seul exécutable pour océan et biogéochimie marine qui interagissent directement).

NEMO/LIM

LIM traite de la dynamique (rhéologie et cinématique) et de la thermodynamique de la glace de mer. NEMO_LIM peut être utilisé *off-line* (la dynamique est lue dans des fichiers via OASIS et le Standalone Surface module) ou *in-line* (un seul exécutable pour océan et biogéochimie marine qui interagissent directement).

INCA

INCA traite de la chimie et des aérosols atmosphériques. Il calcule les réactions chimiques (modèle Eulérien 3D, choix entre méthode explicite et/ou implicite), l'effet des aérosols sur le système radiatif ainsi que les émissions, les transformations photochimiques et les dépôts secs et humides des espèces chimiques et des aérosols. Il est couplé au module de transport du modèle LMDZ. La parallélisation d'INCA suit celle de la partie physique du modèle LMDZ.

REPROBUS

REPROBUS traite de la chimie et des aérosols stratosphériques.

ORCHIDEE

ORCHIDEE rassemble trois composantes : SECHIBA traite des transferts d'eau et d'énergie entre l'atmosphère, la végétation et le sol ; STOMATE traite du cycle du carbone dans la végétation et les sols ; LPJ traite de la dynamique de la végétation. ORCHIDEE consomme peu de temps CPU par rapport aux autres composantes. ORCHIDEE peut aussi être utilisé en offline, sur un seul point, sur un ensemble de points ou sur l'ensemble de la planète, en étant piloté par des champs climatologiques précalculés.

XIOS

XIOS (XML-IO-server) est une nouvelle bibliothèque capable de gérer les entrées-sorties des modèles climat avec souplesse et performance. Le pilotage des sorties se fait grâce à un fichier XML. Les performances sont atteintes grâce aux processeurs supplémentaires dédiés. C'est la version 2.0 qui est utilisée dans tous les modèles IPSL. Voir http://forge.ipsl.jussieu.fr/ioserver

Structure des programmes

Pour les deux modèles dynamiques NEMO et LMDZ, les équations sont discrétisées en différences finies, sur une grille C (Arakawa 1972) et l'intégration temporelle est réalisée par un schéma saute-mouton. Pour éviter la divergence des pas de temps pairs et impairs due au schéma saute-mouton, un filtre Asselin est utilisé dans le cas de l'océan alors que pour l'atmosphère un schéma Matsuno (avant-arrière) est appliqué tous les cinq pas de temps. Pour satisfaire le critère de stabilité CFL, le modèle d'atmosphère a un pas de temps de six minutes et le modèle d'océan de une heure trente aux résolutions de la configuration IPSLCM5A2. Pour IPSLCM6-LR le pas de temps de la dynamique atmosphérique est de 2,15 mn et de 45 mn pour la dynamique océanique. Dans le modèle d'atmosphère, les termes de diffusion et de sources d'énergie sont calculés toutes les quinze minutes et toutes les six heures pour le rayonnement.

Parallélisation

LMDZ

LMDZ est parallélisé par découpage de domaine, chaque partie étant découpée de façon particulière.

Dynamique

Le découpage se fait par bande de latitude et par niveaux verticaux. Le découpage en latitude est une parallélisation en mémoire distribuée, avec l'utilisation de la bibliothèque *MPI* de 2 bandes de latitudes par tâche MPI. Le découpage vertical se fait grâce à une parallélisation en mémoire partagée, avec l'utilisation d'*OpenMP*. Toutefois, cette parallélisation possède des limites intrinsèques de l'ordre de quelques centaines de cœurs de calcul aux résolutions usuelles.

Transport

La méthode est identique à celle de la dynamique, mais les bandes de latitudes sont réparties de façon égale entre les processeurs.

Physique

Découpage mono-dimensionnel : on répartit les points également sur les processeurs, en mélangeant les parallélisations *MPI* et *OpenMP* pour conserver le même nombre de tâches *MPI* et de thread OpenMP que dans les parties dynamique et transport. Il faut noter que cette partie est *embarassingly* parallèle et s'adaptera aisément à toute autre module dynamique comme DYNAMICO.

LMDz et DYNAMICO

Le maillage latitude-longitude utilisé par LMDZ possède deux points singuliers aux pôles qui limitent, de fait, la scalabilité du modèle. DYNAMICO, le nouveau cœur de calcul de la dynamique, a été conçu en utilisant un maillage icosaédrique sur lequel les équations de la dynamique ont été discrétisées en utilisant des schémas numé-

riques nouveaux conservant l'énergie ou la vorticité potentielle. Le développement du code DYNAMICO, démarré en 2012, permet maintenant des simulations avec la physique LMDZ et les surfaces continentales ORCHIDEE. A la résolution la plus basse (comparable à celle de IPSLCM5A2) la scalabilité est excellente jusqu'à 3 000 cœurs, là où LMDZ stagnait à partir de 256 processeurs.

ORCHIDEE

Lorsqu'il est couplé à LMDZ (via une interface Fortran), la parallélisation d'ORCHIDEE est identique à celle de la partie physique de LMDZ à l'exception des tâches associées à des domaines sans aucun point de terre qui ne font que participer aux barrières globales.

NEMO

NEMO est parallélisé par découpage de domaine dans les directions X et Y. C'est une parallélisation à mémoire distribuée utilisant la bibliothèque *MPI*. Pour limiter le nombre de coeurs utilisés, il est possible d'éliminer les processeurs ne calculant que sur des points terre. Cela est le cas dans le modèle couplé IPSLCM de référence.

Précisions informatiques

Temps CPU, mémoire utilisée et volumes de fichiers

Volume de fichiers (taille et inodes, en 2020)

Sur le serveur de fichiers (STORE à l'IDRIS et CCCSTOREDIR au TGCC) **une simulation de 100 ans avec IPSLCM6-LR**, en sorties volontairement limitées (Output, Analyse et RESTART/DEBUG), représente 4.1 TB et 660 inodes.

Les ATLAS et MONITORING représentent également un volume de 3.9 GB et 8000 inodes, stockés sur WORK à l'IDRIS ou sur CCCWORKDIR au TGCC.

Enfin, le workflow CMIP6, lorsqu'il est activé, ce qui est possible seulement au TGCC, représente un peu plus de 9 TB et 1 800 inodes.

	TGCC	**IDRIS**	**Volume**	**Nb inodes**
Sources	HOME_GENCMIP6/CCCWORKDIR	WORKDIR	1.6 GB	40 000
Output/Analyse/RESTART/DEBUG	CCCSTOREDIR	STORE	4.1 TB	660
ATLAS/MONITORING	CCCWORKDIR	WORK	3.9 GB	8000
WF CMIP6	CCCWORKDIR	N/A	9 TB	1 725

Langages de programmation utilisés

Tous les modèles sont programmés en *Fortran 90*. Cependant, pour des raisons de performances, peu des spécificités nouvelles du Fortran 90 sont utilisées. La bibliothèque d'entrée-sortie XIOS et son serveur détachable sont écrits en C++.

Bibliothèques et logiciels tiers utilisés

Pour l'éxécution des codes de calculs

- NetCDF
- Pour les multi-executables lancés en mode MPMD :
  - mpirun sur curie et ccc_mprun sur irene au TGCC
  - Environnement Intel sur ada à l'IDRIS

En post-traitement

- evince
- cdo
- NetCDF
- ferret
- Firefox
- ghostscript
- imagemagick
- nco
- netpbm
- Python, avec les modules NetCDF4, numpy, cdat, cartopy, xarray
- subversion
- tetex-latex

État de préparation des programmes

Le modèle couplé et toutes ses composantes sont exploités de façon intensive depuis plus de vingt ans sur les différents calculateurs du CEA et du CNRS. Ils ont régulièrement été optimisés sur ces machines.

La production scientifique se fait actuellement en utilisant quelques dizaines à un, voire deux milliers de cœurs.

La bascule vers le nouveau cœur dynamique DYNAMICO va permettre l'utilisation, en production, de plusieurs milliers de cœurs de calcul.

Checkpoint restart

Tous les codes composant le modèle couplé disposent d'un système de checkpoint-restart. Un ensemble de scripts permet de décomposer les simulations longues (plusieurs dizaines ou centaines d'années). Généralement, l'exécutable est relancé chaque année de simulation, avec écriture et sauvegarde d'un checkpoint. Chaque job enchaîne plusieurs années, de façon à avoir des travaux en machine environ 24h. Les jobs sont enchaînés sur plusieurs semaines pour atteindre la longueur de simulation voulue.

Performances

Performances du modèle couplé et des configurations forcées associées

Ces chiffres sont issus de programmes en production scientifique, avec en particulier des conditions réalistes d'entrées/sorties. Les post-traitements se font sur des partitions dédiées (`jean-zay-pp` à l'IDRIS et les noeuds `xlarge` au TGCC).

Rappelons que la chaîne de calcul n'est pas opérationnelle au CINES. Seules les compilations et les exécutions simples sont supportées par l'IPSL à ce jour.

Tableau des performances sur Irène-SKL

Configuration	Nb total de coeurs	Temps réel pour 10 ans	**Temps à demander pour 10 ans**
IPSL-CM6.2-MR1	1200	28 h	**33 600 h**
IPSL-CM6.2-MR025	4720	22 h	**103 900 h**
IPSL-CM6.1.11-LR	976	15 h	**14 800 h**
IPSL-CM5A2-VLR	437	2.5h	**1 200 h**
eORCA1-LIM3-PISCES	433 (432 nemo + 1 xios)	12 h	**5 200 h**
LMDZOR_v6.1.10-LR	576	12 h	**7 000 h**

Tableau des performances sur Irène-AMD

Configuration	Nb total de coeurs	Temps réel pour 10 ans	**Temps à demander pour 10 ans**
IPSL-CM6.1.11-LR	1952 (dépeuplement x 2)	10 h	**20 000 h**
IPSL-CM5A2-VLR	604	2.5h	**1 500 h**
eORCA1-LIM3-PISCES	640 (502 nemo + 1 xios noeud dédié)	10.2h	**6600**
LMDZOR_v6.1.10-LR	1136 (dépeuplement x 2)	10 h	**11 000 h**

Tableau des performances sur JeanZay-CPU

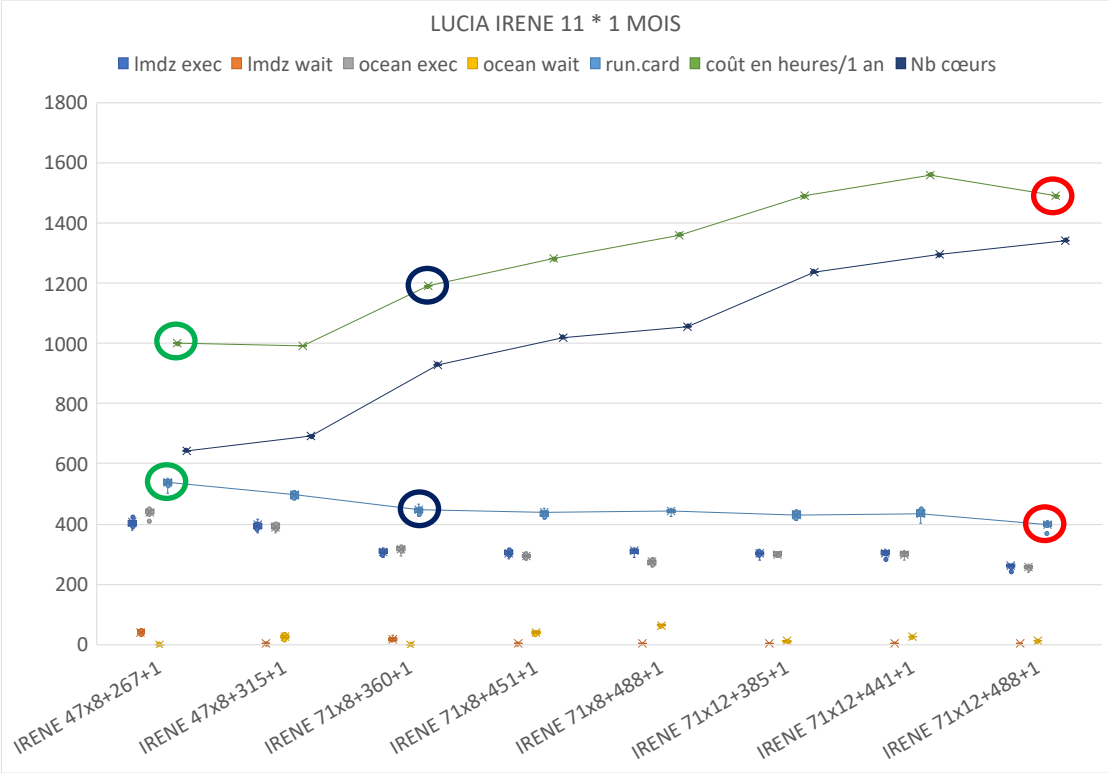
Configuration	Nb total de coeurs	Temps réel pour 10 ans	**Temps à demander pour 10 ans**
IPSL-CM6.1.11-LR	1071	10 h	**10 800 h**
IPSL-CM5A2-VLR	399	2.6h	**1 100 h**
eORCA1-LIM3-PISCES	428	6 h	**2 600 h**
LMDZOR_v6.1.10-LR	711	10.5 h	**7 500 h**

Equilibrage des charges sur Irène-SKL pour IPSLCM6-LR :

Grâce à l'outil LUCIA inclus dans OASIS3-MCT, il est possible d'étudier l'équilibrage des charges de la configuration IPSLCM6-LR, par exemple sur Irene-SKL. L*a* figure

suivante montre l'équilibrage des charges (temps d'attente de chacun des exécutables) pour des répartitions différentes. Les mesures sont faites sur des ensembles de 11 simulations de 1 mois à chaque fois. La figure montre également le temps d'exécution tel qu'il est enregistré dans le fichier run.card ainsi que le nombre de cœurs de calcul utilisés. Enfin, le coût calcul pour un an de simulation est représenté en vert. Elles permettent de repérer pour chaque machine une répartition optimale selon le critère retenu : coût calcul, temps d'exécution ou le meilleur compromis. Pour 10 ans de simulation, en retenant le meilleur compromis coût calcul/temps de simulation, on retrouve les chiffrages de 14 000 h pour Irène-SKL avec la répartition usuelle conseillée : 71 tâches MPI et 8 threads OpenMP pour l'atmosphère LMDZ, 360 tâches MPI pour l'océan NEMO et une tâche pour le serveur XIOS. Ces chiffrages sont avec des sorties limitées.

Figure 1 : Les points représentent les temps d'exécution et d'attente des deux exécutables sur ***Irène-SKL****. La courbe bleu clair représente les temps d'exécution :* ***entre 540 secondes et 400 secondes selon les répartitions****. La courbe bleu foncé représente le nombre de cœurs utilisés et enfin la courbe verte représente le coût calcul en heures pour une année de simulation. Avec 47 tâches MPI pour LMDZ, c'est la configuration 47x8+315+1 qui coûte le moins cher comme l'indique les cercles verts. Avec 71 tâches MPI pour LMDZ, c'est la configuration conseillée 71x8+360+1 qui coûte le moins cher (cercles bleus) et la configuration 71x12+488+1 qui va le plus vite (cercles rouges).*



Post-traitement

La chaîne de calcul climat est composée de nombreux travaux qui s'enchaînent les uns après les autres, ou parallèlement suivant les cas. Pour une simulation d'une centaine d'années, il y a quelques dizaines de jobs de calcul et plusieurs centaines de jobs de post-traitement. L'utilisation de XIOS a simplifié et fiabilisé cette chaîne en supprimant les travaux de rebuild/mask puisque les fichiers sont alors directement

générés depuis les jobs de calcul. Ce qui supprime 150 jobs de post-traitement, soit 15%, par rapport à l'ancienne méthode.

TGCC

Les jobs de calcul tournant sur Irène-SKL utilisent Irène-xlarge pour les post-traitements tandis que les jobs de calcul tournant sur Irène-Rome utilisent Irène-Rome comme machine de post-traitement (car il n'est actuellement pas possible de lancer des travaux sur Irène-xlarge depuis Irène-Rome). Le passage des jobs se fait selon une priorité dont le calcul tient compte de l'ancienneté du job, de la consommation récente et du type de projet (PRACE ou GENCI). Il n'est pas possible d'indiquer qu'un job appartient à une longue chaîne de calcul qui va rester plusieurs semaines en machine. Dans certains cas, nous pourrions être amenés à demander des priorités afin de garantir des temps de passage compatibles avec les échéances des projets.

Les fichiers intermédiaires sont stockés en SCRATCHDIR.

Certains fichiers (produits de simulations pour la communauté) sont destinés à être visibles depuis les serveurs thredds. En 2017, la mise en place avec succès de l'export des files system (WORK et STORE standards, WORK et STORE CMIP6) vers le mésocentre ESPRI IPSL a permis de n'utiliser que le seul datanode IPSL : https://vesg.ipsl.upmc.fr/thredds/ L'optimisation du trafic réseau a permis d'optimiser les temps de transferts.

IDRIS

Les jobs de calcul utilisent Jeanzay. Les jobs de post-traitements utilisent `jean-zay-pp`.

Les fichiers intermédiaires sont stockés WORKDIR, et l'archivage pérenne sur STORE.

Certains fichiers sont également destinés à être visibles depuis le serveur thredds : http://www.idris.fr/jean-zay/pre-post/jean-zay-thredds.html

Logiciels de post-traitements nécessaires

- Envoi de Mail depuis les serveurs de calcul
- Serveurs svn (oasis, lmdz et forge)
- evince
- cdo
- NetCDF
- ferret
- Firefox
- ghostscript
- imagemagick
- nco
- netpbm
- Python 2.6.2 au moins, avec les modules NetCDF4, numpy, xarray, cartopy, cdms2
- subversion
- tetex-latex