Scripts de préparation des fichiers clim et des fichiers de couplage sur le nouveau système de calcul NEC (TORI)

Quelques informations pour les stagiaires qui utiliseront TORI
pour la première fois

Par Rachida El Ouaraini
Maroc-Météo

Avril 2007

1. Informations utiles à savoir !

a) Le nouveau système de calcul de Météo-France NEC

Le nouveau système de calcul intensif de Météo-France NEC comprend une frontale scalaire *TX7* et deux systèmes symétriques vectoriels *SX-8R* de 16 noeuds chacun.
Le premier système « sumo » est dédié aux travaux opérationnels, le deuxième « tori » est dédié aux travaux de recherches et développement.
Chaque noeud est constitué de 8 processeurs vectoriels ayant chacun une puissance crête de 35 Gflops. La capacité de calcul dédiée aux travaux de recherches et développement s'élève donc à :
16 * 8 * 35 Gflops = 4480 Gflops !

La frontale scalaire TX7 est le point d'accès unique des utilisateurs du système NEC. C'est à partir de la frontale qu'on soummettra les travaux sur les noeuds vectoriels et qu'on fera des transferts de fichiers avec « cougar » et avec d'autres machines du réseau Météo-France. La frontale sera aussi utilisée pour les compilations (de préférence en mode « batch ») !
L'OS de la frontale étant Linux, et des noeuds vectoriels Unix système V.

Document sur l'utilisation du NEC

Un document, rédigé par Marion Python sur l'utilisation de la machine « tori », est disponible en français et en anglais (traduction faite par Jean Maziejewski):

- **Français**: Documentation Utilisateur: Introduction à l'utilisation du système NEX « tori » de Météo-France.
- **Anglais**: The User's Guide: Introduction to NEC « tori » Machine at Météo-France.

b) La soumission des jobs sur « tori »

En interactif

La connexion interactive se fait uniquement sur la machine frontale (telnet tori), il n'y a donc pas d 'interactif sur les noeuds vectoriels !
Seuls les travaux courts et peu consommateurs de ressources peuvent tourner en interactif sur la frontale.

En batch

Tous les travaux importants doivent être soumis en « batch » sur les noeuds vectoriels. Le gestionnaire du batch est NQSII.
Toutes les compilations importantes doivent être soumises en « batch » sur la frontale scalaire. Une classe « compile » est dédiée pour ces travaux (qsub -q compile).

La soumission des travaux se fait depuis tori (la frontale). L'utilisateur peut demander pour son job un processeur (monoprocesseur), un noeud (jusqu'à 8) ou plusieurs noeuds !

Contrairement au VPP, l'ordonnanceur des tâches est basé sur le temps réel (elapsed time). Il faut donc obligatoirement préciser ce dernier dans les options de soumission de « qsub ».
Il est important, pour un foctionnement optimal de l'ordonnanceur des tâches de décrire : le nombre de noeuds, le nombre de processeurs par noeud, le temps CPU, le temps elapsed et la mémoire par noeud !

La structure des classes sur « tori »

Plusieurs classes ont été définies sur la nouvelle machine. La commande « qstat -Q » permet de les lister.
Jusqu'ici, seulement 2 classes ont pour machine d'exécution la frontale scalaire: la classe « compile » pour les travaux de compilations importantes et la classe « ft » pour les transferts de fichiers. Toutes les autres classes (mono, multi, express, vector, ...) concernent les noeuds vectoriels.

Les principales commandes standard *NQS*

- qsub *job* : pour la soumission du job.
- qstat : pour voir le(s) job(s) de l'utilisateur qui tourne(nt).
- qstat_all : pour voir tous les jobs.
- qdel *job_id* : pour tuer le job.

Je note que jusqu'à la fin de mon stage, je n'ai pas trouvé l'équivalent de la commande « qcat -f » qu'on utilisait sur le VPP !

c) Plus de « ftget » ou « ftput » dans les jobs de calculs !

Le logiciel « ftserv » a été porté uniquement sur la frontale. Les commandes « ftget » et « ftput » doivent obligatoirement être utilisées à partir de cette machine pour transférer les fichiers entre cougar (ou autre machine du réseau Météo-France) et la frontale « tori ». Les noeuds vectoriels ne sauront pas ce que c'est « ftget » et « ftput ».
Comme tous les travaux de calcul sont soumis sur les noeud vectoriels, ces travaux ne doivent pas contenir de ftget ou ftput !
Donc pour soumettre des travaux batch, il est important que ceux-ci soient décomposés en 3 sous-travaux NQS:

1. pre_job: désarchivage des fichiers d'entrée à partir de cougar. Le job tourne sur la frontale

en classe « ft »:

- acquisition des fichiers
- qsub -q vector job_calcul

2. job_calcul: tourne sur les noeuds vectoriels.
    - calcul
    - qsub -q ft post_job
3. post_job: archivage sur « cougar ». Le job tourne sur la frontale en classe « ft » .

NB: pour des fins de transfert de fichiers, ne pas oublier de créer le fichier « **.ftuas »** sur la machine « tori » ($HOME). Le mot de passe du login sur la machine « cougar » y est stocké de façon cryptée.
La commande qui permet de créer ce fichier sur tori est :

/usr/local/bin/ftmotpasse -u *login* -h cougar-tori

d) le job MULTISTEP avec directives MTOOL

Le MultiStep permet d'avoir au sein d'un même job plusieurs étapes différentes, permettant de profiter au mieux d'un environnement hétérogène (e.g. machines scalaires, vectorielles) ou de contraintes d'exploitation particulières.

Le fait d'être obligé d'avoir 3 jobs, le premier pour le rapatriement des fichiers d'entrée à partir de « cougar », le deuxième pour le calcul et le troisième pour l'archivage sur « cougar » peut s'avérer pour certains lourd. La solution a été portée par l'approche MULTISTEP. En effet, les trois travaux (pre_job, job_calcul et post_job) peuvent être remplacés par un seul travail en mode MULTISTEP !

Ci-dessous un exemple simple qui permet d'illustrer l'utilisation du MULTISTEP avec directives MTOOL :

```
#MTOOL autolog
#MTOOL autoclean
#MTOOL set nnp=2
#MTOOL set node=1
#MTOOL set jobname=e927_Noraf

#MTOOL step id=fetch target=toritx
ftget /chaine/mxpt/mxpt001/arpege/oper/production/$AA/$MM/$JJ/r$RR/icmsharpe+0006$WORKDIR/icmsharpe+0006
ftget /home/m/marp/marp001/pub/const/clim_lace/20km68/clim_lace.20km68.07.m03 $WORKDIR/clim_lace.20km68.07.m03
ftget /home/m/marp/marp001/pub/const/clim_arpege/tl358/clim_arpege.tl358.05.m03 WORKDIR/clim_arpege.tl358.05.m03

#MTOOL step id=compute target=torisx
xmpirun -nn $NN -nnp $NNP ./MASTER -m$MODEL -v$VERSION -e$CNMEXP -c$NCONF -t$TSTEP -f$NSTOP -a$ADVEC

#MTOOL step id=fetch target=toritx
ftput $WORKDIR/PFFPOSLACE+0000
```

Dans l'exemple ci-dessous, on a divisé les travaux en trois étapes (steps). La première étape (#MTOOL step id=fetch target=toritx) s'éxecutera sur la frontale scalaire **toritx** car il s'agit d'un rapatriement de fichiers d'entrée à partir de « cougar » (ftget ...). La deuxième (#MTOOL step id=compute target=torisx) s'éxecutera sur le(s) noeud(s) vectoriel(s): **torisx**. Il s'agit de la tâche de calcul. Quant à la dernière étape, elle s'éxecutera sur la frontale scalaire **toritx**, car on effectue un archivage des fichiers de sortie sur « cougar » (ftput ...).

NB: la soumission d'un job multistep ne se fait pas en utilisant « qsub job », mais en utilisant la commande « mtool_filter.pl »: **mtool_filter.pl job_multistep.**

Document sur l'utilisation du MULTISTEP

Un document sur l'utilisation du MultiStep a été rédigé par ***Eric Sevault : MULTISTEP avec directives MTOOL.***

e) Quelques conseils/normes à utiliser pour les namelists et les scripts

1. Il est conseillé que toutes les commandes « cp », « rm », « cat », « mv » soient précédées par un antislach (\) afin d'éviter les problèmes que peut engendrer l'interférence avec les alias.
2. Pour copier un fichier, il est préférable de spécifier « \cp -b 32768 » au lieu de la commande « cp » afin d'optimiser les opérations de transfert de données.
3. Un certain nombre de normes est à respecter pour les namelists, je note ci-dessous quelques unes:
    - Mettre tous les éléments de la namelist dans l'ordre alphabétique.
    - Inclure tous les éléments existants dans une namelist, y compris ceux qui ne sont pas utilisés par la configuration courante.
    - Supprimer de la namelist tous les éléments obsolètes.
    - Un élément de namelist ne doit apparaître qu'une seule fois, ainsi que les variables (enlever les duplicatas).

L'outil qui permet de mettre tous les éléments d'une namelist dans l'ordre alphabétique et supprime les duplicatas est : **xpnam** (~marp001/public/bin/xpnam)**.**

Un autre outil: **alignnamelist** (~mrpm603/ykproc/alignnamelist) permet aussi de remettre les éléments de la namelist dans le bon ordre alphabétique, supprime les éléments obsolètes et rajoute les éléments manquants à notre namelist. Pour réaliser cela, il faudrait fournir à cet outil une namelist de référence vide.
Un jeu de namelists vides contenant les bons éléments est disponible pour chaque cycle sur la machine Andante : /u/mrpm/mrpm603/ykscar/name-cycle 32T0_jnar

NB1: la machine Andante sera remplacée bientôt par « meru » et « triolet ». Voir la namelist de référence sur l'une de ces deux machines (triolet ??) ou demander à Karim Yessad.

NB2: l'application de l'outil « xpnam » ou « alignnamelist » à la namelist qu'on veut modifier « *namelist* » créé une nouvelle namelist : « *namelist.*new ». C'est celle-ci qui respecte les normes qu'on a cité précédemment et qu'il faudrait utiliser pour la configuration courante.

NB3: Karim Yessad a rédigé une note sur la **« Gestion des Namelist de Mitraillette, d'Olive et de l'Opérationnel: Cycle 32t0 ».** Elle contient des jeux de namelists ainsi que des normes de présentation de ces namelists.

2. Scripts de préparation des fichiers clim: e923

Les scripts ci-dessous permettent l'élaboration des fichiers clim ***modèle*** *et* ***BDAP*** sur la machine « tori » pour différents domaines :

- e923-BOURBON01 : clim BDAP réunion.
- e923-NORAF : clim BDAP Noraf.
- e923-FRANX01 : clim BDAP France.
- e923-lace : clim modèle Lace.
- e923-france : clim modèle France.
- e923-reunion : clim modèle Réunion.
- e923-noraf : clim modèle Noraf.

Ces scripts se trouvent sous : ***/cnrm/gp/mrpa/mrpa667/e923/cy32t1.***

Tous les fichiers d'entrée pour cette configuration se trouvent sur disque, on n'a donc pas besoin de script (pre_e923) pour leur désarchivage à partir de cougar.
En revanche, un script d'archivage (post_e923) a été élaboré afin de stocker les fichiers clim en sortie des « e923-*domaine* » sur « cougar ».
Le script post_e923 (se trouve sous la même arborescence) sert à stocker tous les fichiers clim issus des différents modèles. Il suffit juste de spécifier les variables « cycle » et « pref » pour le cycle utilisé ainsi que le domaine choisi.

3. Scripts d'élaboration de fichiers de couplage: e927

4 scripts de couplage e927 ont été élaborés sur la machine « tori ». Ils permettent l'élaboration de fichiers de couplage pour les domaines : Aladin France, Aladin Noraf, Aladin Réunion et Aladin Lace.

- e927-France
- e927-NORAF
- e927-Reunion
- e927-LACE

Ces scripts se trouvent sous: ***/cnrm/gp/mrpa/mrpa667/e927/cy32t1.***

Deux autres scripts se trouvent sous la même arborescence:

- pre_e927
- post_e927

Le premier permet le rapatriement des fichiers *d'entrée* nécessaires à cette configuration de « cougar » vers « tori ». Le deuxième permet d'archiver les fichiers de couplage *en sorti*e de la e927 sur « cougar ».

Un autre script ***en mode MULTISTEP*** a été réalisé pour le domaine Lace: **e927-LACE-mstep.**
Ce script permet à lui seul de réaliser les trois étapes suivantes:

- le rapatriement des fichiers d'entrée à partir de cougar (ftget).
- Le calcul (exécution du Master).
- l'envoi des fichiers de couplage vers cougar (ftput).

Et donc **e927-LACE-mstep** permet de remplacer les 3 scripts suivants :

1. pre_e927
2. e927-LACE
3. post_e927

Remerciements
Merci beaucoup à Françoise TAILLEFER et Gwenaelle HELLO