Using Open MP in OASIS3-MCT for the N-nearest-neighbor remapping

Coquart L., Maisonnave E., Valcke S.

CECI, Université de Toulouse, CNRS, CERFACS, Toulouse, France - TR-CMGC-18-19

Table of Contents

1. Introduction

As the power of computational platforms is increasing, coupled climate models use more and more high resolution grids in the climate simulations. The OASIS3-MCT coupler, developed at Cerfacs in collaboration with the CNRS and the Argone National Laboratory, is widely used in the climate community. It includes the SCRIP library routines (*Jones, P.W. 1998*) to offer to the users the possibility to calculate the weight and address files associated to the interpolations between the grids of their models, in spherical coordinates. The problem is that this step in OASIS3-MCT is still done on the master processor of each model so it represents a bottleneck at high or very high resolution and it can still take multiple days to calculate one remapping file for these grids, even on a recent computer.

Eric Maisonnave showed in his technical report of 2015 (*Maisonnave, E. (2015)*) that it was possible, for the N-nearest-neighbor remapping, to parallelize the loop over the unmasked target points using Open MP (N being the number of source neighbor used, specified in the configuration file namcouple). In this loop, the N-nearest-neighbor grid source points are searched for each unmasked target point to calculate the associated weights and addresses. In his tests, the N-nearest-neighbor remapping was performed between 3 sets of increasing resolution grids on different computers, and the results were presented for two couple of grids (Orca2-Reg3°) and (Orca025-Reg0.5°). They showed an important diminution of the computational time with the number of threads.

As we have access to higher resolution grids and new computers, we decided to perform the same kind of tests for a 4-nearest-neighbors remapping between 4 couples of increasing resolution grids on the Cerfacs computer nemo_Lenovo. The computer is composed of 288 bi-socket Intel Hashwell nodes each with 24 cores and with 64 Go of memory. The aim is to give the Open MP routine to the users with the next official release of OASIS: OASIS3-MCT_4.0.

We decided to focalize our study on the impact of using Open MP on the remapping time calculation when the resolution of the grids increases, using a standard Open MP launching when working with 2 couple models running in mode MPMD. We did not perform any tests of the impact on the results of different threads and/or nodes configurations on nemo_lenovo (use until 23 cores for model1 and only 1 core for model2 on 1 node, use of 1 node for each model instead of 1 node for 2 models, …).

The next part presents the SCRIP N-nearest-neighbor remapping routine modified with the Open MP instructions (*Maisonnave, E. (2015) ; Chapman, B. et al (2008)*). Then, the environment developed to perform the tests is described. The fourth part of the document presents the results we obtained, compared to the ones of Eric Maisonnave and we finally conclude in the last part.

2. The Open MP algorithm for the N-nearest-neighbor interpolation in the SCRIP library

Following what did Eric Maisonnave in the SCRIP N-nearest-neighbor routine in the old official version OASIS3-MCT_2.0 (*Valcke, S. et al. (2013) ; Maisonnave, E. (2015)*), we added the same Open MP instructions in the same SCRIP routine in the trunk of OASIS3-MCT, which will become the next official version OASIS3-MCT_4.0.

The Open MP instructions implemented in the routines remap_distwgt and store_link_nbr, both contained in the module remap_distance_weight, are given in the Appendix A.

All the Open MP instructions begin with "***!$***" to be able to use the same routine both compiled with or without openmp. Before beginning the Open MP loop ***("!$OMP DO SCHEDULE(RUNTIME)***") over the unmasked target grid points to calculate the weights and addresses in parallel, the arrays of the routine have to be defined private to each thread (using "***!$OMP PARALLEL DEFAULT(FIRSTPRIVATE)***"). Only the coordinates of all the source points, that have already been calculated, can be used and shared between all the threads ***("!$ SHARED(coslat,coslon,sinlat,sinlon)***").

For each unmasked target point of the loop, the routine store_link_nbr is called to store the source and target addresses of each link (here 4 links) attached to each unmasked target point. To be able to have reproducibility results between results with and without Open MP, a barrier had to be implemented in store_link_nbr to be sure that each thread was written correctly its results and in the good order ("***!$OMP CRITICAL***").

3. Computing environment developed for the tests on Cerfacs computer nemo_Lenovo

A toy coupled model, *test_mono_rmp_withopenmp*, was used to perform the tests. Model1, running on grid1, sends at t=0 a field defined by an analytical function on its grid to model2, running on grid2. Regridding files can be computed online using the SCRIP library and it was used to evaluate the time of calculation of the regridding files by model1 with or without Open MP for different definition and resolution of the source and target grids. This environment is very close to the one that users can find in the official version OASIS3-MCT_3.0 to test the quality of their interpolation (*Valcke, S. et al. (2015) ; Senhaji, H. (2016) ; Craig, A. et al. (2017))* and it will be distributed with OASIS3-MCT_4.0.

As the remapping file is only calculated by the master process even if the model is parallelized in MPI, each model runs in monoprocessor. They are both compiled with the same options used to compile OASIS3-MCT on nemo_Lenovo with intel and impi. These compilations options are:

- Without Open MP : F90FLAGS = **-O2 -xCORE-AVX2 -I. -assume byterecl**
- With Open MP : F90FLAGS = **-O2 -xCORE-AVX2 -I. -assume byterecl –openmp**

As we did not study the influence of the threads and/or nodes distribution on the results, we wrote a submission script when running 2 couple models running in mode MPMD with Open MP ***on one node***. The submission script is given in the Appendix B (8.1). The "***#SBATCH***" commands correspond to the way we want to allocate the resources on the computer. We never share our node to avoid performance bias (with or without using Open MP).

So we choose to run the 2 coupled models, running in MPMD, on one node ("***#SBATCH –N 1***"). Then it is necessary to describe how we will use the resources on this node.

To be able to run Open MP tasks on the node, it is necessary to define two granularity levels on the node: the MPI tasks per nodes ("***#SBATCH --ntasks-per-node=2***" as there are 2 sockets by node) and the Open MP tasks by "MPI task" which on nemo_lenovo is 12 as there are 24 cores on 2 sockets ("***#SBATCH --cpus-per-task=12***").

The number of OMP_NUM_THREADS effectively used in the job is defined below in the script using :

- ***export OMP_NUM_THREADS=X[1-12]***

We did not used more than 12 threads for model1 as the curves seems to converge at 12 (see results below).

We also put, in Appendix B (8.2), the submission script for 2 coupled models running in mode MPMD on one node, ***not using Open MP***. In this case there is only one level of granularity to define, the "MPI tasks" per node ("***#SBATCH --nodes=1 --ntasks-per-node=24***").

4. Results

We tested 4 couple of grids:

- ***Orca2 (182x149x1 points)*** for model1 and ***bt42 (6232x1 points)*** for model2 (Low-Resolution)
- ***Orca1 (294x362x1 points)*** for model1 and ***t127 (24572x1 points)*** for model2 (Medium-Resolution)
- ***Orca025 (1442x1021 points)*** for model1 and ***t359 (181724x1 points)*** for model2 (High-Resolution)

- ***Orca12e (4322x3059x1 points)*** for model1 to ***t799 (843490x1 points)*** for model2 (Very-High-Resolution)

The results of the time calculation of the weights and addresses file for a 4-nearest-neighbors remapping as a function of the number of thread (cores) for model1 are given below, for each couple of grids, on nemo_Lenovo:

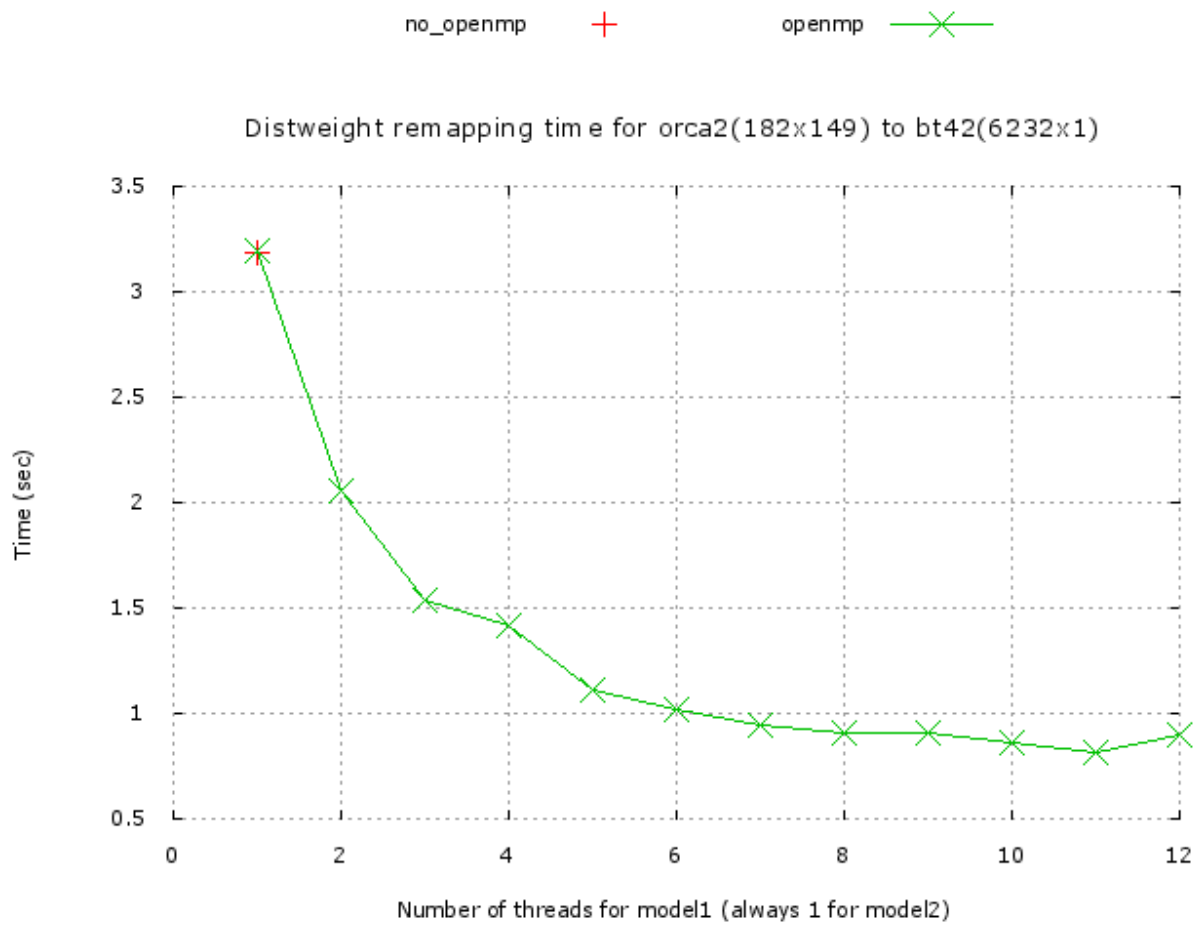

Figure 1: Distweight remapping time between Orca2 to Bt42

The results on ***Figure 1*** obtained for a very low resolution are closed to the one obtained by Eric Maisonnave on a Sandy Bridge with the Orca2-Reg3° (6912x1 points) configuration (*Maisonnave, E. (2015)*). Anyway, the calculation takes less than 5 seconds with or without Open MP so it is not so useful to use Open MP at these resolutions.

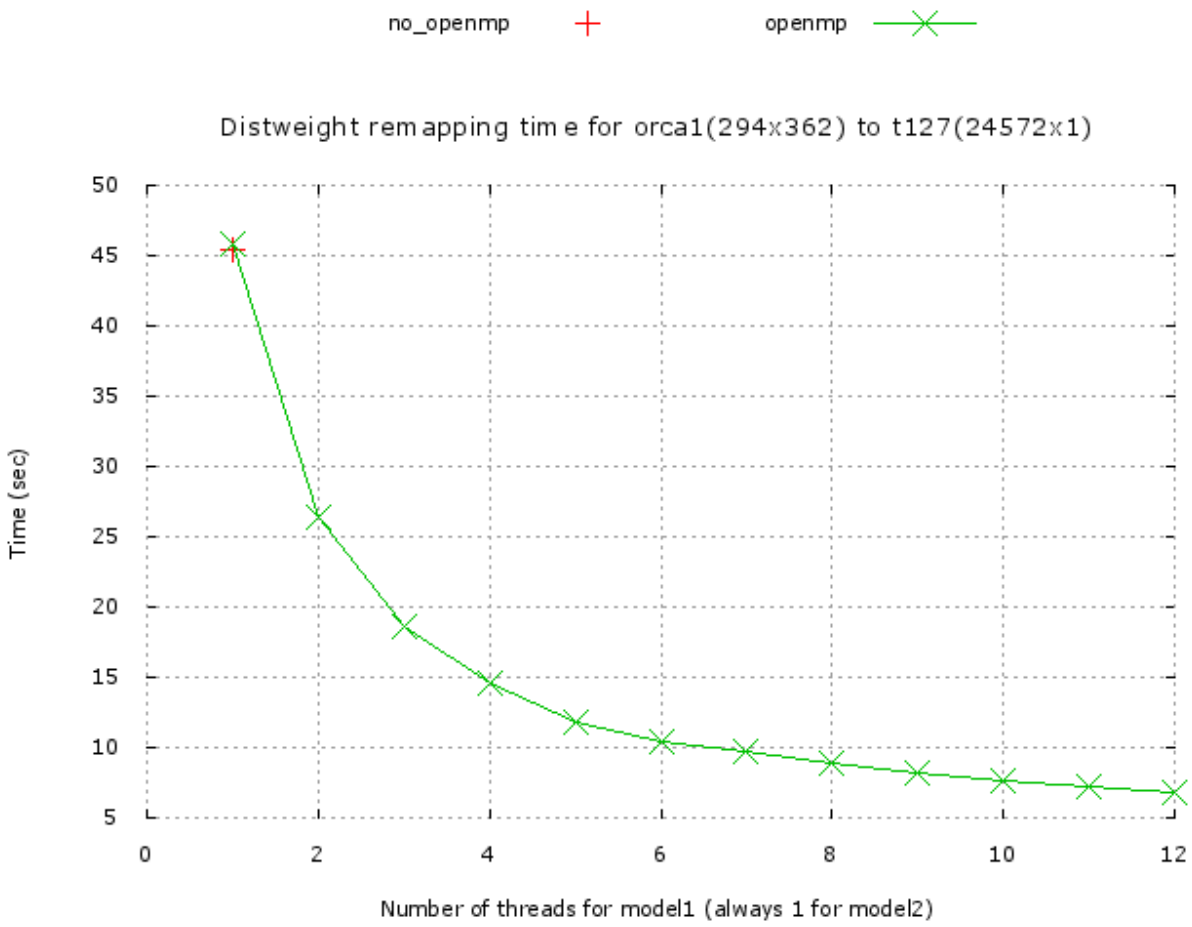

Figure 2: Distweight remapping time between Orca1 and T127

There is no results in Eric Maisonnave's report for the configuration tested on ***Figure 2***. It can be observed that even if the time calculation of the remapping file is less than a minute, it is of benefit to use Open MP as the time is almost divided by 10 when using 12 threads (cores) for model 1.

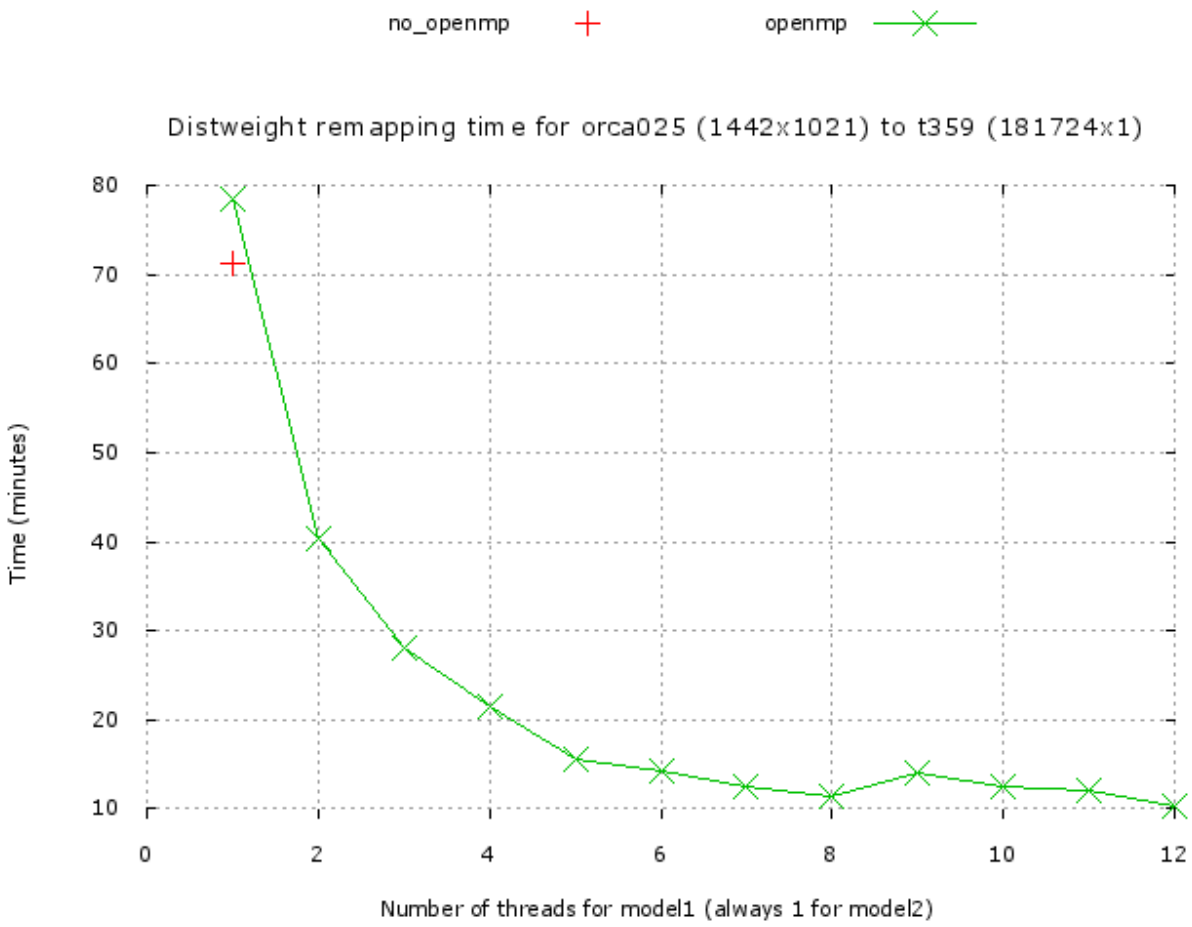

Figure 3: Distweight remapping time between Orca025 and T359

The closest configuration to this one in Eric Maisonnave's report (*Maisonnave, E. (2015)*) is Orca025-Reg05° (260281x1). Our results are similar to what he obtained on the Sandy Bridge. We see on ***Figure 3*** that it takes 4800 seconds on one core to calculate the remapping file between Orca025 and T359 while on 12 cores it takes only 600 seconds. So it is very useful to use Open MP to calculate the remapping files for these high-resolution grids.

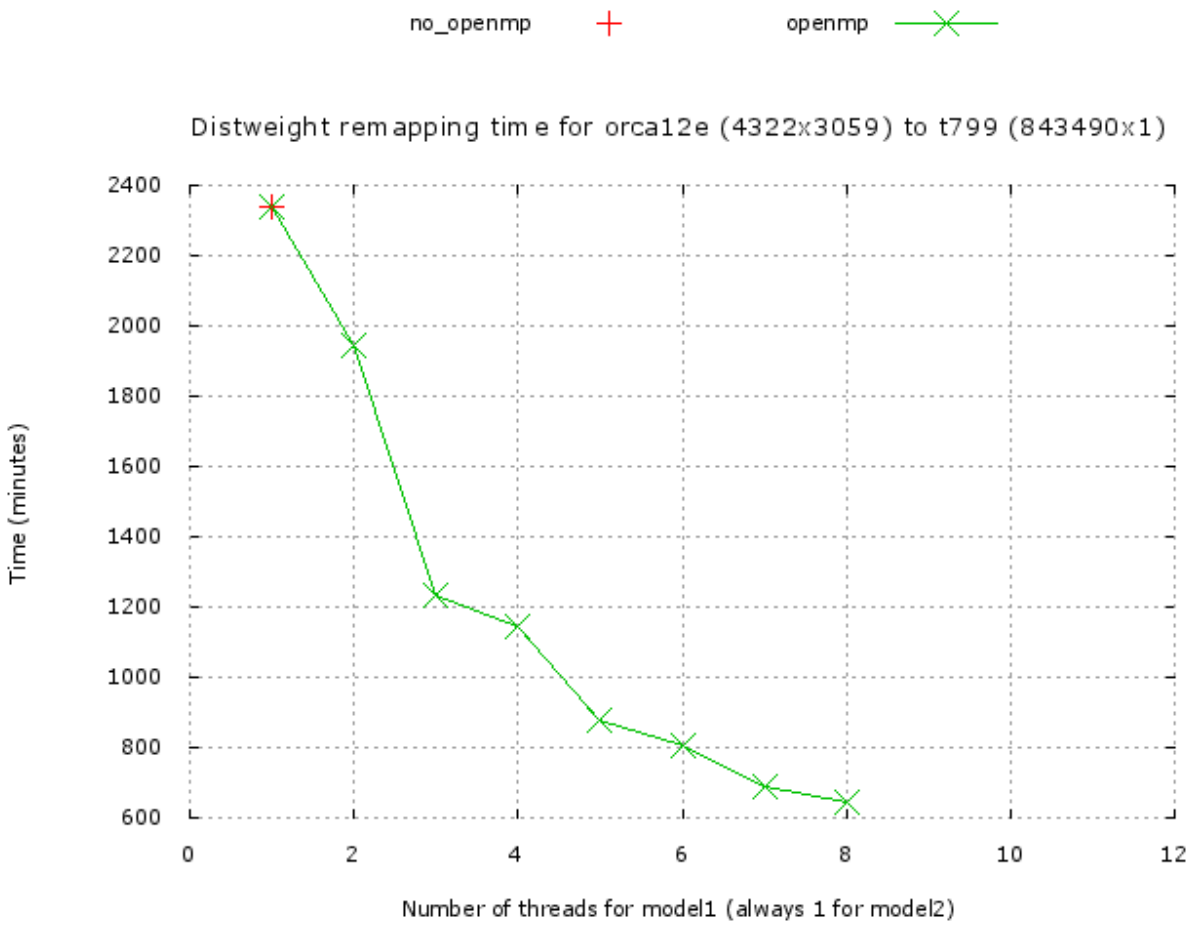

Figure 4: Distweight remapping time between Orca12e and T799

We could not use more than 8 cores to perform the tests. With 9 threads the job is cancelled due to time limit (more than 48 hours) but we did not find any easy explanation to these results. The results on ***Figure 4*** show that even if the time computing of the remapping file is still high with 8 cores (almost 10 hours) it is much less than on one core (39 hours) so it becomes also interesting to use Open MP at these resolutions.

5. Conclusion

Thanks to Eric Maisonnave, we could implement some Open MP instructions in the N-nearest-neighbor routine of the last version of OASIS3-MCT.

The first part of the report presents the environment used on the Cerfacs computer nemo_Lenovo to test the time calculation of the weights and addresses file for a 4-nearest-neighbors remapping using or not using Open MP. We did not analyze the influence of the threads and/or nodes configuration on the results.

The tests were done for 4 couples of grid from low to very high resolution. The results show that the time for the calculation of the weight and addresses file are of the same order of the one obtained by Eric Masionnave on a Sandy Bridge. Globally, it becomes interesting to use Open MP with our medium-resolution couple of grids up to the very-high-resolution grids.

Andrea Piacentini (Cerfacs) and Eric Maisonnave (Cerfacs) have started a new study from this one to further improve the nearest-neighbor routine as well as the bilinear and the bicubic routines of the SCRIP, using OpenMP and MPI.

Acknowledgements

We would like to thanks Isabelle D'Ast from Cerfacs who helped us to create the submission scripts for the tests on Cerfacs computer nemo_Lenovo.

This research was supported by the ESiWACE H2020 European project, grant agreement no. 675191 (www.esiwace.eu) and the CON- VERGENCE project funded by the French National Research Agency: ANR-13-MONU-0008.

6. References

Craig, A., Valcke, S. and Coquart, L. (2017) Development and performance of a new version of the OASIS coupler, OASIS3-MCT 3.0, Geoscientific Model Development, 10, pp. 3297-3308, doi:10.5194/gmd-10-3297-2017

Senhaji, H. (2016), Evaluation et comparaison des interpolations numériques des librairies SCRIP et ESMF, UMR 5318 CECI, CERFACS-CNRS, WN-CMGC-16-227, Toulouse, France , working note

Maisonnave, E. (2015), Interpolation weights calculation distributed with OpenMP, URA SUC 1875, CERFACS/CNRS, TR-CMGC-15-27172, Toulouse, France , Technical report

Valcke, S., Craig, A. and Coquart, L. (2015), OASIS3-MCT User Guide, OASIS3-MCT3.0, URA SUC 1875, CERFACS/CNRS, TR-CMGC-15-22804, Toulouse, France , Technical report

Valcke, S., Craig, T. and Coquart, L. (2013), OASIS3-MCT User Guide: OASIS3-MCT2.0, URA SUC 1875 CERFACS/CNRS, TR-CMGC-13-22805, Toulouse, France, technical report

Chapman, B., Jost, G. and Van der Ruud, P. (2008) Using OpenMP: Portable Shared Memory Parallel Programming, MIT Press, Cambridge, MA

Jones PW (1998) A user's guide for SCRIP: a spherical coordinate remapping and interpolation package, Los Alamos National Laboratory. http://oceans11.lanl.gov/trac/SCRIP

7. Appendix A: Open MP instructions in the nearest-neighbor routine

```
   module remap_distance_weight
!$    USE OMP_LIB
   contains
   subroutine remap_distwgt ()
!  Parallelization of loop on the all the unmasked target points
!$OMP PARALLEL DEFAULT(FIRSTPRIVATE)  ! some initial values set in a non parallel section have to be copied before the parallel section into private array
!$   & SHARED(coslat,coslon,sinlat,sinlon) ! except the geometric data of the grid source needed to find the source nearest neighbors needed by all threads and not modified during the loop
      !$OMP DO SCHEDULE(RUNTIME)
    do dst_add = 1, grid2_size     ! grid2_size : target grid size
          call store_link_nbr()
    end do
    !$OMP END DO
!$OMP END PARALLEL
    end subroutine remap_distwgt
    subroutine store_link_nbr()
! Barrier to ensure reproducibility of the remapping field when running on 1 socket and on threads (cores) >= 1
     !$OMP CRITICAL
      num_links_map1  = num_links_map1 + 1
      if (num_links_map1 > max_links_map1)
    &     call resize_remap_vars(1,resize_increment)
      grid1_add_map1(num_links_map1) = add1
      grid2_add_map1(num_links_map1) = add2
      wts_map1    (:,num_links_map1) = weights
     !$OMP END CRITICAL
    end subroutine store_link_nbr
    end module remap_distance_weight
```

8. Appendix B: Submission script for 2 couple models running in mode MPMD using Open MP and not using Open MP

8.1 Submission script for 2 coupled models running in mode MPMD using Open MP on nemo_Lenovo:

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
#!/bin/bash -l
# Name of the job
#SBATCH --job-name openmp
# Time limit of the job
#SBATCH --time=12:00:00
#SBATCH --output=$rundir/$casename.o
#SBATCH --error=$rundir/$casename.e
# Number of nodes
#SBATCH -N 1
# Number of MPI processes by node = number of sockets by
#SBATCH --ntasks-per-node=2
# Number of openmp threads by MPI socket = cores number by socket
#SBATCH --cpus-per-task=12
#SBATCH --mail-user=coquart@cerfacs.fr
#SBATCH --mail-type=END
cd $rundir
export KMP_STACKSIZE=1GB
export I_MPI_PIN_DOMAIN=socket
export KMP_AFFINITY=verbose,granularity=fine,compact
export OMP_NUM_THREADS=X (X=1 to 12)
time mpirun -genv I_MPI_DEBUG 5  -np 1 ./$exe1 : -np 1 ./$exe2
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

8.2 Submission script for 2 coupled models running in mode MPMD not using Open MP on nemo_Lenovo:

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
#!/bin/bash -l
# Name of the job
#SBATCH --job-name openmp
# Time limit of the job
#SBATCH --time=12:00:00
#SBATCH --output=$rundir/$casename.o
#SBATCH --error=$rundir/$casename.e
# Number of nodes and cores (processes)
#SBATCH --nodes=1 --ntasks-per-node=24
#SBATCH --distribution cyclic
#SBATCH --mail-user=coquart@cerfacs.fr
#SBATCH --mail-type=END
cd $rundir
ulimit -s unlimited
time mpirun -np 1 ./$exe1 : -np 1 ./$exe2
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```