IS-ENES - WP 7 - D7.4

Original title in the DoW: **OASIS4 including high priority developments as identified in the User Survey.**
Revised title: **OASIS3-MCT parallel coupler**

Abstract:

OASIS3-MCT_2.0, the new version of the OASIS coupler interfaced with MCT (Model Coupling Toolkit) released in February 2013, offers today a fully parallel implementation of coupling field regridding and exchange, while supporting most of OASIS3.3 previous functions. The scalability tests done with OASIS3-MCT_2.0 at high number of cores and the fact that it supports unstructured grids allow us to conclude that this coupler offers today a fully parallel and efficient coupling solution answering the short and mid term needs of the European climate modelling community. Therefore, OASIS3-MCT_2.0 fulfils the objective of Task 2 of IS-ENES WP7/JRA1 and is proposed as deliverable 7.4 as it covers all of its original objectives, even if it does not build on the OASIS4 software originally targeted.

Grant Agreement Number:	**228203**	**Proposal Number:**	**FP7-INFRA-2008-1.1.2.21**
Project Acronym:	**IS-ENES**		
Project Co-ordinator:	**Dr Sylvie JOUSSAUME**		

Document Title:	OASIS3-MCT parallel coupler			**Deliverable:**	D 7.4
Document Id N°:		**Version:**	1.0	**Date:**	March 4th 2013
Status:	Draft				
Filename:	ISENES_Deliverables_7.4.docx				
Project Classification:	Public				

Approval Status		
Document Manager	**Verification Authority**	**Project Approval**

REVISION TABLE

Version	Date	Modified Pages	Modified Sections	Comments
0.1	2013/02/17			First version by S. Valcke
0.2	2013/02/28			Revisions by S. Valcke including M. Carter's comments
1.0	2013/03/04			Revisions by S. Valcke including M.-A. Foujols' comments

Executive Summary

Deliverable 7.4 is the last deliverable of IS-ENES WP7/JRA1 Task 2, which objectives were to deliver a fully parallelised and optimized coupler for current coupled climate components, assess the coupler performance and scalability, and include the support of unstructured grids identified as a priority in the WP7/JRA1 User Survey.

In the previous deliverable 7.2 “Fully parallelised and optimised version of OASIS4 answering needs of current coupled climate components”, we reported on the work done for the development and validation of OASIS4 and provided a general description and results of scalability tests performed with the version delivered in July 2011, OASIS4_1beta. Although specific versions of OASIS4 have been or even are still used in real coupled system, a deeper analysis of the remaining problems and efforts already devoted to the OASIS4 development, led us to conclude that OASIS4 source code was not a stable basis to build on to answer future climate modelling coupling needs, in particular with respect to the support of unstructured grids. Consequently, no additional effort has been invested in the OASIS4 development since July 2011.

Two options were first considered for future development, the OASIS4 user-defined regridding functionality and the interfacing of MCT in OASIS3 and it became clear very rapidly that the later was the solution to invest in. The OASIS3-MCT_2.0 version released in February 2013 offers today a fully parallel implementation of coupling field regridding and exchange, while supporting most of OASIS3.3 previous functions. The scalability tests done with OASIS3-MCT at high number of cores and the fact that it supports unstructured grids allow us to conclude that this coupler offers today a fully parallel and efficient coupling solution answering the short and mid term needs of the European climate modelling community.

Therefore, OASIS3-MCT_2.0 fulfils the objective of Task 2 of IS-ENES WP7/JRA1 and is proposed as deliverable 7.4 as it covers all of its original objectives.

1. Introduction

Deliverable 7.4 is the last deliverable attached to IS-ENES WP7/JRA1 Task 2, which objective was to deliver a fully parallelised and optimized coupler offering 2D/3D linear and cubic interpolations and 2D conservative remapping for current coupled climate components. Assessment of the coupler performance and scalability was also mentioned in the DoW. As stated in its original title, deliverable 7.4 also had to include the priorities identified in the User Survey (WP7/JRA1 Task 1); in this respect, the only issue that came out was the support of unstructured grids.

In the previous deliverable 7.2 "Fully parallelised and optimised version of OASIS4 answering needs of current coupled climate components", we reported on the work done for the development and validation of OASIS4 and provided a general description and results of scalability tests performed with the version delivered in July 2011, OASIS4_1beta. Although specific versions of OASIS4 have been or even are still used in real coupled system[1], a deeper analysis of the remaining problems and efforts already devoted to the OASIS4 development, led us to write:

"We therefore conclude here that the original objective, which was to deliver a fully operational coupler performing online parallel calculation of the interpolation weights and addresses, is out of reach given the current resources" and *"Furthermore, our analysis is that the current OASIS4 code base is not a stable basis to build on as it has reached a point where it is too complex to evolve easily to answer current and future climate modelling coupling needs. In particular, the support of unstructured grids was not included in the original design and it now would be very difficult to add it in the current code."*

Since July 2011, no additional effort has been invested in the OASIS4 development. At that time it was also quite clear that the OASIS3 version of the coupler, still giving satisfactory results for the resolution used in operational coupled climate components, was a good basis to build on. However, it was also evaluated that its limited parallelism would soon become a bottleneck in coupled simulations running on more than O(1000) cores. The next steps in IS-ENES had to address this issue. It was then proposed to evaluate two alternative solutions to remove the foreseeable OASIS3 bottleneck. These two solutions supposed that the regridding weights and addresses were pre-computed offline and implement parallel regridding of the coupling fields and parallel exchanges of these fields directly from the source to the target component processes. Indeed the fully parallel calculation of the regridding weights and addresses, implemented in OASIS4 with only limited success, is certainly not mandatory for our climate components in the short and mid term (~5 years, next IPCC report); this functionality will actually only be needed when the component components will run on adaptive grids (which possibly change at each time step) or when the sequential pre calculation of the weights and addresses will no longer be possible, because the memory of one core will not be sufficient to hold the entire global grid. The first solution was the OASIS4 user-defined regridding functionality[2] and the second one was to interface the American Model Coupling Toolkit (MCT) developed by the Argonne National Laboratory (see http://www.mcs.anl.gov/research/projects/mct/) in OASIS3 communication library. The two solutions were considered equivalent in principle; it was then decided to compare the two solutions and choose the best one in terms of efficiency, robustness and ease of use.

We describe here how the two options were first considered and how it became clear very rapidly that OASIS3-MCT was the solution to invest in. We then give some details on OASIS3-MCT functioning and describe what are the differences and similarities with the previous OASIS3.3

[1] In particular: for 3D coupling between atmosphere and atmospheric chemistry component -France in the framework of the EU GEMS project; at BoM (Australia) for ocean-atmosphere limited area coupled model; at SMHI (Sweden) for regional ocean-atmosphere coupling applied to the Arctic region.

[2] This functionality completely bypassed OASIS4 parallel calculation of the interpolation weights and addresses, which had reached a point where it was too complex to build on.

version. Finally we conclude with some scalability tests done with OASIS3-MCT at high number of cores and discuss how this new version of the OASIS coupler probably satisfies the short and mid term needs (~5 years, next IPCC report) of the European climate modelling community.

2. Initial investigation of OASIS4 user-defined and OASIS3-MCT options

When we started this investigation, we decided to migrate from the Trac web-based management environment used for previous OASIS4 developments (see https://oasistrac.cerfacs.fr/) to the Redmine web application (see www.redmine.org) as Redmine became the standard project management tool at CERFACS. The work done to test the OASIS4 user-defined functionality and to develop OASIS3-MCT is described in detail in a series of Redmine tickets on the respective Redmine sites[3].

OASIS4 user-defined functionality was first tested with two "toy" coupled components[4]: the "low-resolution" and "high-resolution" versions used a Gaussian Reduced grid for the first component (respectively a T42 with 6232 points and a T359 with 181724 points) and a logically rectangular grid for the second component (respectively 2 degree with 182x149 points and 0.25 degree with 1021x1442 points). To use the OASIS4 user-defined functionality, we first described, for each set of source and target grids, the links associating specific source grid points with specific target grid points; for each link, we provided the index of the source point (in the total source grid dimension) and the index of the target point (in the total target grid dimension), and the weight associated to that link. This information was produced using the OASIS3.3 coupler in mono-process interpolation-only mode, which internally uses the SCRIP library to calculate these sets of regridding weights and addresses. The OASIS4 coupling library reads these indices and weights provided and automatically defines on each side a non-geographical (gridless) grid with one point for each link. The multiplication of the source field values by the appropriate weights is implemented in the coupling library on the source side and the parallel redistribution of the results to be done directly between the source and the target processes.

After numerous adjustments, these first tests performed with "low-resolution" and "high-resolution" toy components showed satisfactory results with some compilers but memory allocation problems were observed with others, especially with the PGI compiler. Furthermore, a close collaboration between CERFACS and the MetOffice aiming at implementing the OASIS4 user-defined regridding functionality in the MetOffice real coupled system assembling the UM atmosphere model (192x145 points) to the NEMO ocean model using the ORCA1 configuration (360x292 points) revealed additional problems[5] when an important number of coupling fields were exchanged (38 in this case) or when the parallelism was increased. Even if workarounds were found to temporarily solve these specific problems, it once again showed that OASIS4 was not a robust basis onto which to build further developments.

At the same time, the interfacing of Model Coupling Toolkit (MCT) in OASIS3 was progressing rapidly. MCT embodies a generic approach for creating coupled applications. Its design philosophy, based on flexibility and minimal invasiveness, is close to the OASIS approach. MCT uses distributed objects to store the coupling data and pre-computed regridding weights as well as a "domain decomposition descriptor" (DDD) to describe the parallel decomposition of the

[3] OASIS4 user-defined Redmine page is https://inle.cerfacs.fr/projects/oasis4-user-defined-interpolations ; OASIS3-MCT Redmine page is https://inle.cerfacs.fr/projects/oasis3-mct . These pages are accessible, after registration, to all users interested in following the developments.

[4] A "toy" coupled model is a coupled system composed of two programs simulating no dynamics and no physics but performing realistic coupling exchanges and therefore providing dimensioning tests of the coupler functions.

[5] Problems that emerged were related to pointer arrays no longer defined as their targets have been deallocated, or global reading of the regridding weights and addresses by all processes, etc. They are all described on OASIS4 user-defined Redmine page.

components. MCT computes parallel communication patterns based on the source and target DDDs. To use MCT, OASIS3 coupling library loads the coupling data into MCT data types and calls MCT parallel matrix multiplication (for regridding) and communication methods. Parallel data transfer is then accomplished by pairs of send/receive methods with coupling data and communication pattern as inputs. MCT is, most notably, the underlying communication layer in cpl7, the coupler used in NCAR Community Earth System Model 1 (CESM1) as detailed in Craig 2012 and Dennis 2012 (see also Valcke 2012). A first analysis revealed that the information transferred from the component code to the OASIS3.3 coupling library through its Application Programming Interface (API) contained all the necessary information to use MCT underneath to perform the computation of the parallel communication patterns, the parallel matrix multiplication for regridding and the parallel data transfer. One important advantage of this solution identified at the time is that it is therefore be totally transparent for current OASIS3.3 users to migrate to OASIS3-MCT as the coupling library API remains the same. The user has full control on the regridding, as it is based on predefined weights and addresses he or she provides; this allows any form of regridding to be performed, including special cases such as the regridding of a runoff field from a land model to an ocean model into which each discharge land point needs to be associated to a specific ocean region around it.

In December 2011, a first prototype of OASIS3-MCT was already available and used in a series of tests with toy components on Curie Bullx platform, which produced very promising results in terms of performance (see also “OASIS3-MCT performance” paragraph below). It was therefore decided at that point to invest all further development efforts in OASIS3-MCT. A first official version, OASIS3-MCT_1.0, was released in August 2012. The last version OASIS3-MCT_2.0, released in February 2013 offering all OASIS3.3 regridding methods and most of its additional features is described in details hereafter.

3. OASIS3-MCT

OASIS3-MCT is a portable set of Fortran 77, Fortran 90 and C routines offering a Fortran API. Low-intrusiveness, portability and flexibility are OASIS3-MCT key design concepts as for all previous OASIS versions. An important difference with respect to previous OASIS3.3 is that there is no longer a separate coupler executable: OASIS3-MCT acts as a coupling library that needs to be linked to the components, with the main function of interpolating and exchanging the coupling fields between these components. OASIS3-MCT supports coupling of general two-dimensional fields. Unstructured grids are also supported using a one dimension representation of the two or three dimensional structures. Thanks to MCT, all transformations, including regridding, are executed in parallel on the set of source or target component processes and all couplings are now executed in parallel directly between the components via Message Passing Interface (MPI). OASIS3-MCT also supports file I/O using NetCDF, allowing an easy switch between the coupled and forced modes. In the current version, the implementation of this functionality is however non parallel with the reading/writing of the fields performed by the master process only followed/preceded by a global redistribution/gathering of the local parts of the fields to/from the component parallel processes.

In spite of the significant changes in underlying implementation, usage of OASIS3-MCT in the component code has largely remained unchanged with respect to OASIS3.3. To communicate with another model, or to perform I/O actions, a component model needs to include a few specific calls to the OASIS3-MCT coupling library, using the same API as in OASIS3.3. The namcouple configuration file is also largely unchanged relative to OASIS3, although several options are either not used or not supported. There is a new transformation in namcouple i.e. MAPPING which allows a user to specify a mapping file generated externally. Some details about OASIS3-MCT_2.0 are given in the next subsections with bigger emphasis put on the differences with the previous OASIS3.3 version. For a more complete description of OASIS3-MCT_2.0, the reader is referred to the User Guide available on-line at https://verc.enes.org/oasis/oasis-dedicated-user-support-

1/documentation (see also Valcke, Craig, Coquart 2013).

OASIS3-MCT sources

OASIS3-MCT, and in particular OASIS3-MCT_2.0 sources, are available from CERFACS SVN server. OASIS3-MCT is released under a LGPL licence. Anyone can download the sources after fill in the registration form at https://verc.enes.org/oasis/download/oasis-registration-form, which then provides download instructions. A copyright statement about OASIS3-MCT, MCT itself and the SCRIP1.4 library (that can be used to calculate the regridding weights and addresses, see below) is distributed with the sources and can be found in the User Guide.

OASIS3-MCT API

To communicate with another model, or to perform I/O actions, a component model needs to include a few specific calls to the OASIS3-MCT coupling library; the API used in the components is unchanged with respect to OASIS3.3. These include calls for the coupling initialisation, grid data file definition, partition definition, field declaration, field sending and receiving, and termination.

The API use statement has been updated and now requires a single “use mod_oasis” statement instead of the various use statements required in prior OASIS3 versions. A few auxiliary routines *prism_put_inquire*, *prism_put_restart_proto, prism_get_freq* are not supported yet but new routines *oasis_get_debug* and *oasis_set_debug* are now available to respectively retrieve the current internal debug level or to change it.

As for OASIS3.3, the OASIS3-MCT sending (“put”) and receiving (“get”) calls support:

- Automatic sending and receiving actions at appropriate times following user’s choice indicated in the namcouple
- Lagged communications
- Time average or accumulation of the coupling fields
- Some transformations such as regridding (interpolation)
- I/O actions from/to files

In addition, the sending and receiving actions have the following notable characteristics:

- When the source and target components are parallel, OASIS3-MCT now performs fully parallel communication and regridding. This was not the case with OASIS3.3 where the coupling fields were gathered from the source parallel component to one separate coupler process where it was regridded and then redistributed to the target parallel component; this global gathering implied extra communications and represented a potential bottleneck in the data exchanges.

- OASIS3-MCT can now flexibly couple a component using a subset of its processes only (which was not the case with OASIS3.3). This can be particular useful for model in which some processes are dedicated to a specific task (like I/O) and are not involved in the exchange of the coupling fields; or if the coupling domain is a subset of the global domain and is therefore treated only by a subset of the component processes. To do so, new routines oasis_create_couplcomm and oasis_set_couplcomm are now available to create or set a coupling communicator for the subset of the component processes participating in the coupling exchanges (see User Guide section 2.2.3).

- Support for coupling multiple fields via a single communication.

  This is supported through colon delimited field lists in the namcouple, for example

  ATMTAUX:ATMTAUY:ATMHFLUX TAUX:TAUY:HEATFLUX 1 3600 3 rstrt.nc EXPORTED

in a single *namcouple* entry. All fields will use the *namcouple* settings for that entry. In the component model codes, these fields are still sent ("put") or received ("get") one at a time. Inside OASIS3-MCT, the fields are stored and a single mapping and send or receive instruction is executed for all fields. This is useful in cases where multiple fields have the same coupling transformations; aggregating multiple fields into one single communication reduces the communication overhead by reducing the latency costs.

- Matching one source field with multiple targets

  A coupling field sent by a source component model can be associated with more than one target field and model (get). In that case, the source model needs to send ("put") the field only once and the corresponding data will arrive at multiple targets as specified in the *namcouple* configuration file. Different coupling frequencies and transformations are allowed for different coupling exchanges of the same field. The inverse feature, which could be useful when an input field should come from the combination of different source fields, is not allowed i.e. a single target (get) field cannot be associated with multiple source (put) fields.

OASIS3-MCT configuration

The configuration file, *namcouple,* is a text file that must be written by the user before the run to define all information necessary to configure a particular coupled run. The *namcouple* configuration file of OASIS3-MCT is mostly backward compatible with OASIS3.3.

However, several *namcouple* keywords have been deprecated. They may still appear in the file but the information below these keywords will not be read nor used:

- $SEQMODE: It is still possible to define indices of sequence for the different coupling fields. They will be used to detect a deadlock before it happens. It is however not needed anymore to indicate the maximum sequence index under the $SEQMODE keyword as before.
- $CHANNEL: This keyword is not needed anymore as only MPI1 start mode is allowed (see User Guide section 2.2.2). This should not be an issue as OASIS3 users did not, as far as we know, use the MPI2 mode in which each component was spawned by the OASIS3.3 separate application allowing it to keep its own MPI_COMM_WORLD communicator as in the standalone mode.
- $JOBNAME: This keyword was not used by OASIS3.3 but just printed for information in the coupler log file.
- $MODINFO: This keyword, linked to binary coupling restart files, was already deprecated in OASIS3.3.
- $INIDATE and $CALTYPE: These keywords are not used anymore as they were linked to operation FILLING not supported anymore (see below).

The following transformations should no longer appear in the *namcouple* as they are not supported anymore : ONCE, REDGLO, INVERT, MASK, EXTRAP, CORRECT, INTERP, MOZAIC, FILLING, SUBGRID, MASKP, REVERSE, GLORED. All these transformations were either already deprecated in OASIS3.3, or have alternatives that should be used instead, or imply field combination that is not supported in OASIS3-MCT (see last paragraph in "OASIS3-MCT transformations and interpolations" below).

OASIS3-MCT transformations and interpolations

As OASIS3.3, OASIS3-MCT supports the following transformations and interpolations:

- Time accumulation or averaging
- Addition or multiplication by a scalar
- Forced global conservation (with same options as for OASIS3.3, see User Guide section

4.4)

- Interpolation/remapping:
  - With the MAPPING operation, OASIS3-MCT has the ability to read a predefined set of weights and addresses (mapping file) specified in the *namcouple* to perform the interpolation/remapping of the coupling fields. The user also has the flexibility to choose the location and the parallelization strategy of the remapping with specific MAPPING options.
  - As OASIS3.3, OASIS3-MCT can generate the mapping file using the SCRIP library (Jones 1999) on one process of the model components. The N-nearest-neighbour, bilinear, bicubic and 2D conservative remapping schemes from SCRIP available in OASIS3.3 are still supported.
  - The SCRIP bicubic interpolation and 2nd order conservative remapping involve more than one weight per source grid point. For example a SCRIP bicubic remapping applies a weight to the value of the field at the grid point but also to the value of its gradients into two directions and to the value of its cross gradient. OASIS3-MCT will not calculate the gradient and cross-gradient fields as OASIS3.3 could do for logically-rectangular grids. To perform these higher order transformations, additional fields (e.g. the gradients and cross-gradient) have to be transferred as optional arguments to the sending call ("prism_put_proto" or "oasis_put") and in that case the additional weights will automatically be applied to these additional fields. Up to 5 weights for 5 higher order terms are allowed.

A few transformations available in OASIS3.3 are not yet supported in OASIS3-MCT. This should not be a big issue for current OASIS3 users as these transformations have alternatives or were, as far as we know, rarely used. These transformations are all described in detailed in Appendix B.3 of the OASIS3-MCT User Guide:

- Transformation involving the combination of coupling fields or external fields (e.g. BLASOLD, BLASNEW, CORRECT, FILLING, SUBGRID)
- Time transformation LOCTRANS/ONCE as it is equivalent to defining a coupling period equal to the total runtime
- The transformations that were already deprecated in OASIS3.3 (REDGLO, INVERT, REVERSE, GLORED)
- MASK and EXTRAP as they can be replaced by the more efficient option using the nearest non-masked source neighbour for target points having their original neighbours all masked.
- INTERP interpolations are not available; SCRIPR should be used instead.
- MOZAIC is not available, as MAPPING should be used instead.

OASIS3-MCT test suite

As was the case with OASIS4 (see deliverable D7.2), OASIS3-MCT developments are continuously tested and validated on different computers with a test suite under Buildbot, which is a software written in Python to automate compile and test cycles required in software project.

The first step in the set up of the test suite is to define a reference state of some "toy" components once they are entirely running successfully. Then, at each modification of the coupling library, we use Buildbot to test that the results obtained with the new version are the same as for the reference state (or possibly improved/fixed by the modification if that should be the case). The different toy components are routinely tested on a Linux PC "tioman" (PGI Fortran compiler and mpich message passing), and an HP AMD cluster "corail" (Intel Fortran compiler and message passing version 4.0.0.028) at CERFACS, even if OASIS3-MCT itself has been compiled and run on many other platforms (see the complete list at oasis.enes.org under "USER SUPPORT"). The following toy coupled systems with the following characteristics are used:

- *test_prism_oasis* : 2 components, tests backward compatibility between oasis/prism

routines, tests multiple transformations, with or without restart files, with constant field or analytical function constant in time.

- *mapchk* : 2 components, tests the options of MAPPING (bfb/opt/sum), no restart files.
- toy_eric_pulsation : 2 high-resolution components, constant fields, with or without restart files.
- *toy_eric_echam_cosmo_cottbus*: 2 components, 3D exchanges, analytical function varying in time, no restart files.
- *test_simple_options* : 2 components, very simple set-up to do as many tests as possible but with an analytical function varying in time; tests in particular the creation of the remap files with the SCRIP library.
- *no_pes_coupling* : 3 components with 1 component not coupling; reproduces the coupling ARPEGE/NEMO/XIOS where XIOS is not coupling but exchanges data with NEMO; tests the creation of a specific MPI communicator between NEMO and XIOS.
- *tc3a* : 3 components, more complex toy to test different combinations of operations.
- *pes_coupling* : same as *tc3a* but with a component coupling only on a subset of processors.
- *maphot* : tests high order remapping with pre-defined remapping weights files of a field defined by analytical function varying in time.
- *testr4r8* : tests mixed single precision and double precision; coupling field defined by an analytical function varying in time.

The compilation of OASIS3-MCT is performed every morning on each computing platform described above thanks to a “cron” defined on a local Linux platform. The different toy coupled systems are tested at night with Buildbot (installed also on the local Linux platform) if a modification in the sources of OASIS3-MCT is detected via SVN since the last build. The configuration file of Buildbot contains calls to scripts written to test the different tasks performed by the toys. The different tests are launched and run in parallel for the different toys. All tests are first run in monoprocessor mode and then in parallel and some files are created with the results. Then the script verifies that the test ran until completion, that the metrics written in OASIS3-MCT log files are good (comparing the files created by the run to the ones of the reference state), and that the exchanged fields are the same as the ones of the reference state. The exchanged fields produced by the parallel runs are also compared to the monoprocessor case.

OASIS3-MCT current users

Many groups started to use OASIS3-MCT for relatively high-resolution couplings. In all these coupled systems, OASIS3-MCT shows a very satisfactory behaviour, although only few quantitative analyses have been performed up to now. OASIS3-MCT is used at:

- CERFACS (France) for an ocean-atmosphere coupled model based on the ocean NEMO using the ORCA025 configuration (1021x1442 grid points horizontally) and the atmosphere ARPEGE running on a Gaussian Reduced T359 grid (181724 grid points horizontally). This coupled model is used for seasonal prediction experiments in SPRUCE, a PRACE project that was granted 27 Mhours on tier-0 Bullx Curie platform at TGCC near Paris (France). It will also be run for decadal experiments in HiResClim, another PRACE project granted 22 Mhours on tier-0 Mare Nostrum platform at BSC in Barcelona (Spain) and in the European SPECS project (www.specs-fp7.eu).
- IPSL (France) for coupling WRF atmosphere model and NEMO ocean model, both embedding two-way nested zooms, for a resolution ranging from 27 km to 9 km. This coupled model is developed in the framework of PULSATION funded by the French Agence Nationale de la Recherche (ANR) and was granted 22 Mhours on PRACE tier-0 Bullx Curie platform. A first analysis showed that OASIS3-MCT cost is in this case negligible as no specific overhead was detected when comparing a coupling time step to another regular

time step.

- MPI-M (Germany) for all their MPI-ESM versions, in particular MPI-ESM-XR, developed in the framework of the STORM project, into which ECHAM6 T255L95 (768x384 grid points, ~50km, 95 vertical levels) is coupled to MPIOM TP6ML40 (3602x2394 grid points, ~10km, 40 vertical levels); 17 fields are exchanged at a coupling frequency of 1h.
- MetOffice (UK) for relatively high-resolution ocean-atmosphere coupling between the UM global atmosphere model (N512, 1024x769) and NEMO ocean under the ORCA025 configuration (1021x1442). A detailed analysis led to the conclusion that coupling interpolations are done about 7 times faster with OASIS3-MCT than with the previous OASIS3.3 (see also IS-ENES deliverable D4.7).
- BTU-Cottbus (Germany) for a 3D coupling between ECHAM global atmosphere (T63, 192x96x47) and the regional atmosphere COSMO-CLM (221x111x47, ~2 degree resolution); this configuration also includes coupling to MPI-OM ocean model (254x220). A 6% coupling overhead was observed for exchanging six 3D fields between ECHAM and COSMO-CLM every ECHAM time step (see also IS-ENES deliverable D4.7).
- BoM (Australia) for a coupled limited area model (CLAM) based on MOM4p1 ocean (0.1 degree) and the UK Met Office UM6.4 atmosphere (~12 km resolution), over the [140o E – 160o E, 5o S – 25o S] region for tropical cyclone studies.

OASIS3-MCT performance

The benefit of OASIS3-MCT is, in particular, that all steps of the coupling exchanges are now performed in parallel in the source or target component coupling library. OASIS3-MCT should therefore be much more efficient than the previous OASIS3.3 where all the coupling fields were gathered onto one coupler process and interpolated, causing extra communication and a bottleneck in the processing of the transformations. We verify here that OASIS3-MCT removes the coupling bottleneck that was observed with OASIS3.3 when coupling high-resolution components running on a high number of cores (see Valcke 2013 for details).

To have an exact measure of the time required by the OASIS3-MCT, we performed measurements of the coupling initialisation and of the coupling exchanges on the Bullx Curie thin nodes at the “Très Grand Centre de Calcul” (TGCC) in Bruyères-le-Châtel near Paris (http://www-hpc.cea.fr/en/complexe/tgcc-curie.htm). Compared to fat nodes[6], thin nodes[7] that follow the actual trend of less cores and less memory per node should allow us to better identify potential problems linked to the current evolution of computer architectures. We used a toy coupled model composed of two components simulating no dynamics and no physics but performing realistic “ping-pong” coupling exchanges (i.e. one field in each direction) between a 0.25 degree logically rectangular grid (1021x1442 grid points) for the first component and a Gaussian Reduced T799 grid (843 000 grid points) for the second component, for different number of cores. The codes ran on Curie thin nodes and were compiled with Intel Fortran 12.1.7.256 and Bullx MPI 1.1.16.5, which is based on OpenMPI.

[6] 4 eight-core Intel® Xeon® processors and 128 Go memory per node
[7] 2 eight-core Intel® Sandy Bridge EP (E5-2680) 2.7 GHz and 64 Go memory per node

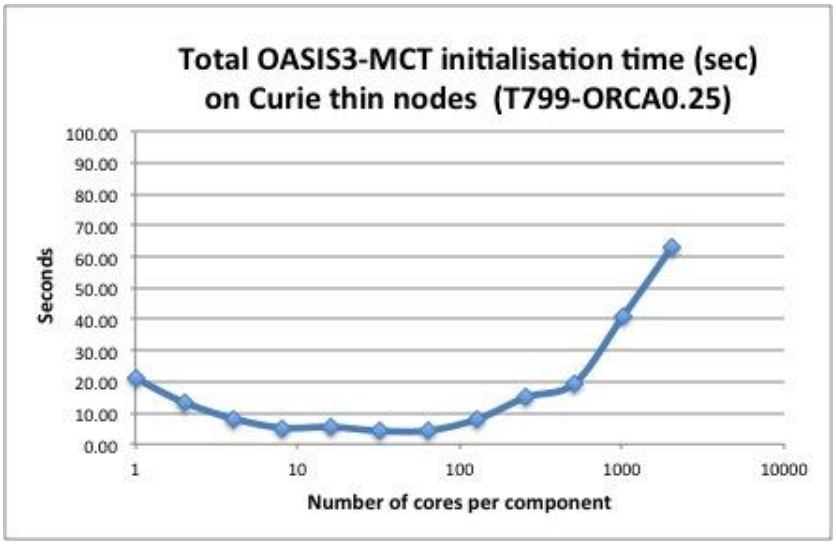

Figure 1: Total coupling initialisation elaspe time on Bullx Curie thin nodes for a coupling between a component running on a 0.25 degree logically rectangular grid (1021x1442 grid points) for the first component and another component running on a Gaussian Reduced T799 grid (843 000 grid points) performing ping-pong exchanges of one field in each direction as a function of the number of cores used to run each component (from 1 to 2048).

Fig 1 shows the total coupling initialisation as a function of the number of cores used to run each component. The initialisation contains all coupling related tasks that must be done only once at the beginning of each run, i.e. in particular the computation of the parallel communication patterns based on the source and target decompositions, the parallel reading of the weight-and-address file, the initialisation of the parallel sparse matrix vector multiplication that will be used during the run to remap the source fields on the target grid. As shown on Fig. 1, the initialisation time slightly decreases from 1 to O(100) cores and then regularly increases for higher number of cores. This is most likely linked to the fact that the computation of the parallel communication patterns, which was observed to take most of the initialisation time when the number of cores is greater than O(100), becomes more complex as the components are parallelised on a higher numbers of cores even if more processes are available to do this computation. We consider that the initialisation time is still very reasonable being of the order of one minute for one coupling field exchanged in each direction when each component is run on ~2000 cores. Furthermore, in a real coupled system, the initialisation time will not be proportional to the number of coupling fields exchanged as the communication pattern will be reused for all fields sharing the same source and target grids and the same regridding. In a typical coupled system, the number of different combinations of source grid, target grid and regridding between these grids is typically much less than the number of coupling fields itself. However, additional measures have to be performed to quantify this cost more precisely and deduce a typical overhead for our coupled systems.

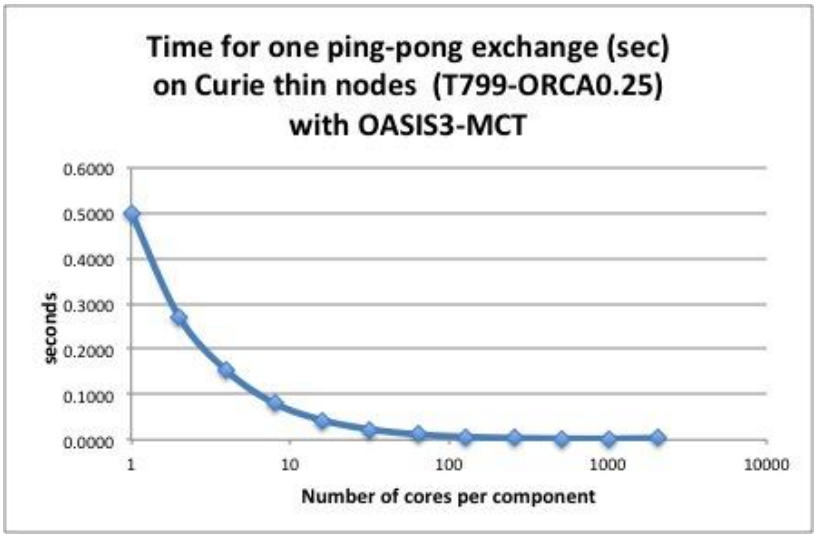

Figure 2: Elapse time (in second) on Bullx Curie thin nodes for one ping-pong exchange with OASIS3-MCT between a component running on a 0.25 degree logically rectangular grid (1021x1442 grid points) for the first component and another component running on a Gaussian Reduced T799 grid (843 000 grid points) as a function of the number of cores used to run each component (from 1 to 2048).

Fig 2 shows the time needed for a ping-pong exchange as a function of the number of cores used to run each component. In a “ping-pong” exchange, the first component sends a coupling field to the second component that receives it and sends it back to the first component. The interpolation (i.e. the multiplication by the weights and the addition to get the target values) is done for each exchange on the source processes and the redistribution is automatically done from the source to the target processes. The elapse time of the ping-pong exchange is measured in the first component as the difference between the time just before the sending and the time just after the reception of the coupling field. The total time for 100 exchanges was measured and then divided by 100. We see that the time needed for a ping-pong exchange nicely decreases from 0 to O(100) cores. For higher number of cores, the time seems to stabilize; the exact measures for 256, 512, 1024 and 2048 cores per component are 0.0037, 0.0024, 0.0023, 0.0036 seconds respectively. OASIS3-MCT coupling exchanges are therefore very satisfactory. Of course, its behaviour at even higher number of cores still needs to be investigated.

Based on these numbers, we can extrapolate that the coupling cost of a one-year long simulation performed on Bullx Curie platform in which one coupling field would be exchanged every hour in each direction between two components, each one running on ~2000 cores and using a grid with O(1 M) grid points would be about 60 seconds for the coupling initialisation and about 30 seconds for data exchange, which a priori seem totally acceptable.

4. Discussion and next steps

Based on the description and results detailed above, we can say with confidence that OASIS3-MCT will most likely provide a satisfactory solution for fully parallel coupling of our climate models at the resolutions currently used and at the ones targeted for the next 5 years or so. During IS-ENES2, we plan to maintain the coupler sources and environment (tutorial, web site, Redmine development site, etc.) and to provide active user support. Some additional developments are foreseen, in particular the support of combination of coupling fields so to offer the same panel of transformations as OASIS3.3. Integration of libraries other than SCRIP for performing the calculation of the regridding weights-and-addresses, such as ESMF regridding, and the

development of a Graphical User Interface to build the *namcouple* configuration file will also be considered. Some efforts will also be devoted to analyse and optimise the calculation of the communication patterns and to reduce even further the communication costs.

We recall here that compared to what was targeted for OASIS4, OASIS3-MCT does not perform real-time parallel calculation of the interpolation weights and addresses. Use of OASIS3-MCT supposes that these weights and addresses are pre-calculated. However, on the longer term, we have to prepare for the time when real-time fully parallel calculation of the regridding weights and addresses will become a clear requirement. This functionality will be needed when the component models will run on adaptive grids (which grid point locations change during the run) or when the sequential calculation of the weights and addresses will not be possible anymore because the memory of one core will not be sufficient to hold the definition of the entire source grid. Even if this is some years off for most climate modelling groups, we are currently evaluating in how Open-PALM, another coupler developed at CERFACS could answer future coupled climate modelling needs. Open-PALM was originally designed to perform the communication and synchronisation of the software components of a data assimilation suite; it therefore addresses the particular issue of “dynamic” coupling in the sense that the software components to be coupled can be started and stopped “dynamically” during the run. Open-PALM has proven to be a flexible and powerful dynamic coupler and it is now used by about 40 different groups in France for different multi-disciplinary coupling in different domains, such as aeronautics and space, computational fluid dynamics, combustion but also atmospheric chemistry, hydrology and oceanography. Since January 2011, Open-PALM is developed in collaboration with ONERA, the French Aerospace Laboratory. In particular, the geometrical interpolation library CWIPI developed at ONERA and based on previous work done at EDF (Electricité de France) is interfaced in Open-PALM since April 2011. The CWIPI library is designed for finite elements (unstructured) grids in the 3D space and offers online parallel computation of the weights and addresses for linear interpolations. Open-PALM and its CWIPI library have already shown good performance for up to 12000 cores (Duchaine 2011) but it is obvious that they do not cover all needs of the climate modelling community at this point. In particular, no conservative remapping or 2nd order interpolation are currently available in CWIPI. Also, it is currently not possible to use a set of weights and addresses pre-calculated off line, which is in some cases essential (for example, when manual input is required to better model the discharge of water runoffs into specific regions of the ocean as a coupling exchange between a river routing model and an ocean model). Evaluation of the work required to adapt Open-PALM to climate modelling requirements is therefore on going. It will be compared with the work required to further develop OASIS3-MCT in order to fit future (5-10 year timeframe) climate modelling needs. At that point, which we should aim to reach within 2 or 3 years, we will have to decide if further development efforts should go in OASIS3-MCT or in Open-PALM or possibly in a merge of the two couplers, and we will need to find resources in the community to proceed to such developments.

Conclusions

The performance analysis and the first implementations done in real coupled systems detailed above, allow us to say with confidence that OASIS3-MCT fulfils the objective of Task 2 of IS-ENES WP7/JRA1 which was to “deliver a fully parallelised and optimized coupler offering 2D/3D linear and cubic interpolations and 2D conservative remapping for current coupled climate components”. This report also presents a first positive assessment of the coupler performance and scalability, as was mentioned in the DoW. The support of unstructured grids, which was really the only issue identified as a priority in the User Survey (WP7/JRA1 Task 1), is also covered by OASIS3-MCT as local partitions of these grids can be expressed in a 1D global index space covering the whole grid. The release of the last version, OASIS3-MCT-2.0, is therefore proposed as deliverable 7.4 as it covers all of its original objectives.

References

1. *[Craig 2012] Craig, A. P., Vertenstein, M., and Jacob, R.: A New Flex- ible Coupler for Earth System Modeling developed for CCSM4 and CESM1, Int. J. High Perform. C, 26-1, 31–42, doi:10.1177/1094342011428141, 2012.*
2. *[Dennis 2012] Dennis, J. M., M. Vertenstein, P. Worley, A. A. Mirin, A. P. Craig, R. Jacob, and S. Mickelson, 2012: Computational Performance of Ultra-High Resolution Capability in the Community Earth System Model, Int. J. High Perf. Comput. Appl. , February 2012 26, 5-16, doi: 10.1177/1094342012436965*
3. *[Duchaine 2011] Duchaine, F., Morel, T., and Piacentini, A.: On a first use of CWIPI at CERFACS, CERFACS Technical Report TR/CMGC/11/3, Toulouse, France, 15 pp., 2011.*
4. *[Jones 1999] Jones, P.: Conservative remapping: First- and second-order conservative remapping, Mon. Weather Rev., 127, 2204-2210, 1999.*
5. *[Valcke 2012] Valcke S, V. Balaji, A. Craig, C. Deluca, R. Dunlap, R. Ford, R. Jacob, J. Larson, R. O'Kuinghttons, G. Riley, M. Vertenstein, 2012: Coupling technologies for Earth System Modelling. Geosci. Model Dev., 5, 1589-1596, doi:10.5194/gmd-5-1589-2012.*
6. *[Valcke 2013] Valcke S., 2013: The OASIS3 coupler: a European climate modelling community software. Geosci. Model Dev. Discuss., 5, 2139-2178, doi:10.5194/gmdd-5-2139-2012.*
7. *[Valcke, Craig, Coquart 2013] Valcke, S., Craig, T. and Coquart, L. 2013. OASIS3-MCT User Guide, OASIS3-MCT_2.0, Technical Report TR/CMGC/13/17, Cerfacs, France.*