IS-ENES WP4

OASIS Dedicated User Support 2011

Annual report

E. Maisonnave, S. Valcke

TR/CMGC/12/18

Abstract

The two 2011 Dedicated User Support missions have contributed to foster collaborations on climate community, helping ETHZ (Eidgenossische Technische Hochschule, Zürich) and SMHI (Sveriges meteorologiska och hydrologiska institut, Rossby center, Norrköping) to expand their configurations and enhance their performances using the couplers OASIS3 and the newly developed OASIS3-MCT.

At ETHZ, to overcome a performance default of our previous OASIS4 interfaces, the COSMO regional atmosphere model (DWD and consortium) has been coupled again with the CLM land model (CESM, NCAR), but, this time, using OASIS3. Several optimizations made it as fast as the initial integrated COSMO-TERRA model. The relatively low resolution of the configuration justified the use of the low parallelism coupler OASIS3, but could easily be upgraded with OASIS3-MCT in the future when finer resolution, and higher parallelism, will be required (possibly by MeteoSwiss and other COSMO community users).

In complement, taking advantage of both OASIS and CESM modularity, our coupled model easily integrated a version upgrade of CLM (v3.5 to v4), preparing the way for other possible CESM/OASIS couplings. This characteristic of our coupling leaded to the CAM-NEMO coupling related part of the IS-ENES2 WP10/JRA2 proposal. It also largely facilitates the plugging of new components on COSMO such as Parflow hydrological model (Bonn university, coupling supported independently of IS-ENES funding but reported on present document) or NEMO (SGN, Senckenberg Gesellschaft für Naturforschung).

Our second mission (SMHI) focuses on performance enhancements of an HPC designed coupled model (Ec-Earth, high resolution version T799-ORCA025). Based on 2010 Dedicated User Support conclusions, the OASIS3/OASIS3-MCT upgrade was required. Made available on national supercomputer “ekman”, this new configuration is currently used for PRACE IP1/IS-ENES WP8 joint project on PRACE tier-0 machine “curie”. But, for the moment, such demanding configurations do not seem to be widely used for scientific purpose, which should give more time to OASIS developers to perfect coupler functioning.

To be able to accurately measure and compare the coupling extra cost of the standard OASIS3 based version and the presently implemented OASIS3-MCT based version, a portable and OASIS3 pseudo-parallel mode compliant version of our OASIS performance measurement tool has been developed and tested on several SMHI models. A similar development for OASIS3-MCT is planned in 2012.

In complement, different interactions contributed to set up two new OASIS3 couplings with regional models (RCA-NEMO and RCA-RCO). We emphasize the fact that, focusing and isolating the work of both model and coupler specialists on a given time period, the Dedicated User Support Program strongly contributes to quicken couplings set-up, distribute OASIS best practice through laboratories and give us a clearer idea on present and future model community requirements.

Thanks to Uwe Fladrich, Klaus Wyser, Martin Evaldsson, Wang Shiyu, Ralf Döscher, Robinson Hordoir, Colin Jones (SMHI), Edouard Davin, Sonia Seneviradne, Anne Roches (ETHZ), Olivier Fuhrer (MeteoSwiss), Jean-Guillaume Piccinalli (CSCS), Andy Döbler (Frankfort University) Matthieu Masbou, Prabakhar Shresta and Mauro Sulis (Bonn University) for their strong support and the constant interest for our work. Once again, thanks to our patient OASIS developers, Sophie Valcke, Laure Coquart (CERFACS), Moritz Hanke (DKRZ) and Anthony Craig.

Estimated carbon emission diagnostic for those 3 journeys by terrestrial/maritime means of transport: 180 Kg

Mission #7
Jun 20- Jul 21 2011

Host: Edouard Davin
Laboratory: ETH, Zürich (Switzerland)

Main goal: Optimize OASIS interfaces on regional atmosphere and land models

Main conclusion
To overcome a performance default of our previous OASIS4 interfaces, the COSMO model has been coupled again with CLM (CCSM), but using OASIS3. Several optimizations made it as fast as the initial integrated COSMO-TERRA model. In complement, taking advantage of both OASIS and CCSM (CESM) modularity, our coupled model easily integrated a version upgrade of CLM (v3.5 to v4), preparing the way for other possible CESM/OASIS couplings.

Model / machine description

COSMO-CLM (here called COSMO)
This regional atmosphere model (COSMO v4.8, and its climate version, COSMO-CLM v11) is used by a large community in several central Europe countries (from which ETHZ). DWD, MeteoSwiss and several other meteorological agencies host the operational version of the model. Grid size: 109x121x32, 0.44 degrees. Parallelisation reaches 100 MPI tasks on the targeted supercomputer.

CLM
This land model is developed at NCAR (v4). It is used within the integrated CESM climate model. Initially, CLM3.5 was coupled as a stand alone model through OASIS4 with COSMO (see Dedicated User Support #5).

Those models are available on CRAY XT5 supercomputer, with 22,128 compute cores (2 six-core AMD Opteron 2.4 GHz Istanbul processors per node), CRAY SeaStar 2.2 interconnect. Peak performance of 212 Teraflop/s. The machine is located at CSCS, Manno, Ticino, Switzerland.

OASIS3 interface for CLM3.5

Initial issue

The previously developed OASIS4 interfaces on CLM3.5 (part of CCSM climate model) land model reveal a lack of scalability at relative low level of parallelism. As shown on figure 1, the CLM model (as part of the coupled system) response time increases dramatically when parallelism reaches 60 PE (light blue curb). A code tracing revealed that time was mainly spent on the OASIS4 receiving routine (prism_get). Unfortunately, this default could not be reproduced with toys.

Dedicated User Support duration is limited to a few weeks: to switch from OASIS4 to OASIS3 is the quicker solution we found to overcome this scalability issue. It took a few days to adapt interfaces and bypass the issue (orange curb). Scalability of our interfaces is now limited by COSMO-OASIS3 communications cost only (red curb).

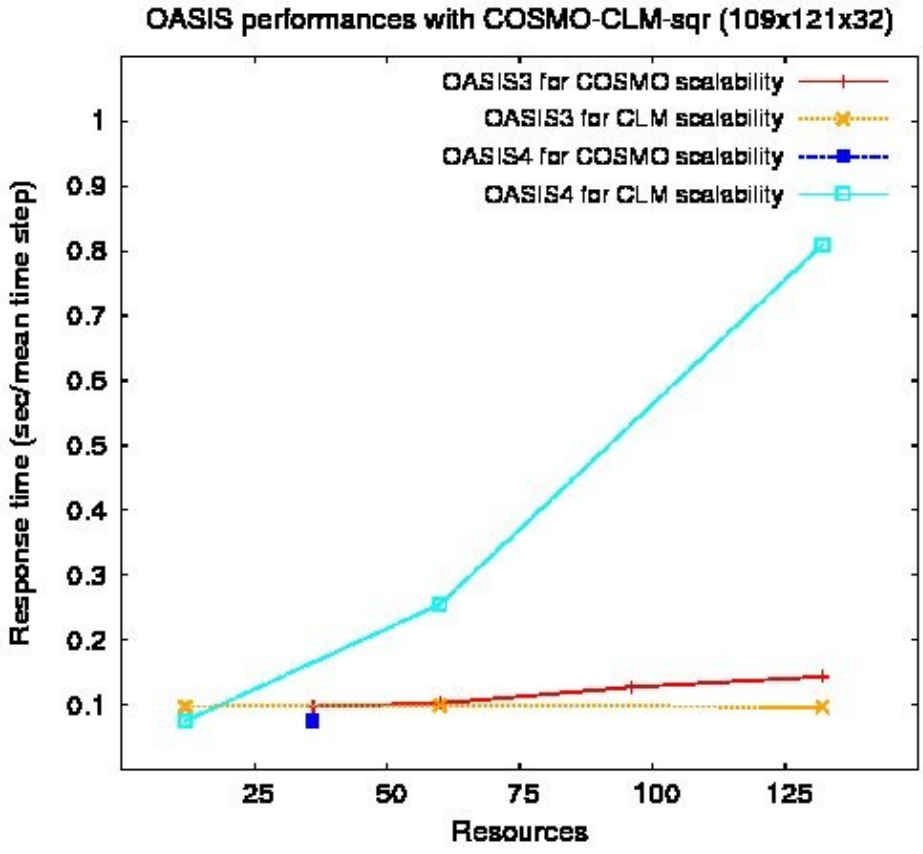

Illustration 1: OASIS3 and OASIS4 interfaces performance

Implementation

Consequently, our new OASIS3 interfaces on both CLM and COSMO models are derived from the previous implementation designed for OASIS4 (see Dedicated User Support report #5). One of the PRISM/OASIS4 specifications was the compatibility of PSMILE interface library (routines called by models) with the OASIS3 one: it explains why it was

quite easy to adapt the previously implemented interface to OASIS3 specificity.

In addition, we took benefit of an existing OASIS3 interface on COSMO, developed by Andy Dobler (Frankfort University) to couple this model with NEMO.

Our final implementation on COSMO differs from U. Frankfort's one:

- coupling fields are different
- we add the possibility to produce auxiliary files (masks, grids, areas) during the definition phase (this characteristic is inherited from our previous OASIS4 interface)

Nevertheless, the similarity of both implementations should encourage the COSMO community to merge them into an unified interface, extending the coupling system to a land/ocean-atmosphere configuration. In this way, a clean package including our interface and the associated input files has been communicated to COSMO-CLM administrator. It should contribute to help COSMO community to build a modular coupled system based on OASIS standard (possibly with IS-ENES help during a new Dedicated User Support).

As the great majority of present supercomputers, CSCS CRAY XT5 requires that a minimum number of PE (12, one node) was allocated for each executable of our coupled system. Consequently, to use OASIS3 on its “pseudo”-parallel mode was mandatory, but has some side effect on the interface implementation[1].

Optimization

Measurement tool

To be able to measure the impact of the following optimizations, the previously developed OASIS option (CPP key “balance”), using MPI_Wtime routine, has been activated (see OASIS Dedicated User Support #4). It delivers informations such as relative duration of each module of the system and OASIS communications + calculations time. But CSCS machine characteristics forbid a simple use of this measurement tool:

- each node has different clock times
- measures writing (fortran WRITE on standard output) at each time step significantly slows down the simulation execution

The first problem has been addressed measuring the clock differences at the beginning of the run but, again, calling an MPI_Barrier on both OASIS and model routines.

The second one makes necessary a complete re-rewriting of our measurement tool: the different informations measured must be synthesized and written at the end of the run (the mean/min/max values), and not at each time step. Due to a lack of time, this development was postponed: the French ANR project “PULSATION”[2] is supposed to address this issue (2012). A first implementation has been designed and is described on the last Dedicated

1 The auxiliary file writing routines, launched during the definition phase of the interface, was not compatible with the pseudo-parallel mode (neither with OASIS performances measurement pre-compiling option). Some code modifications within OASIS were necessary to bypass the issue. On the code interface, MPI_Barrier (on the MPI_COMM_WORLD communicator, which manages the coupled exchanges) has been called. This implementation is temporary, and has to be redefined for an official release.

2 http://www.locean-ipsl.upmc.fr/~pulsation

User Support mission report of this document. For the moment, ratios between the measured quantities are supposed to be the same with or without measurement tool enabling. Absolute values are deduced from one of those quantities (from the total run duration, for example).

Coupling fields number reduction

Five coupling fields (see annex 2) are exchanged from CLM to COSMO, which is less than the number of available OASIS driver PEs. It means that all those coupling fields are processed at the same time by one OASIS instance: the parallelism is almost ideal.

On the way back (COSMO to CLM), the initial coupling fields number was 13. It means that 1 OASIS instance has 2 fields to process, which slowed down the whole coupling sequence. A brief analysis of how coupling fields were used by CLM showed that convective rain and snow, as well as grid scale rain, snow and mist, could be merged into two coupling fields only (total convective and total grid scale precipitations). Gain on performances (on OASIS total time) is about 20%.

Raw performances

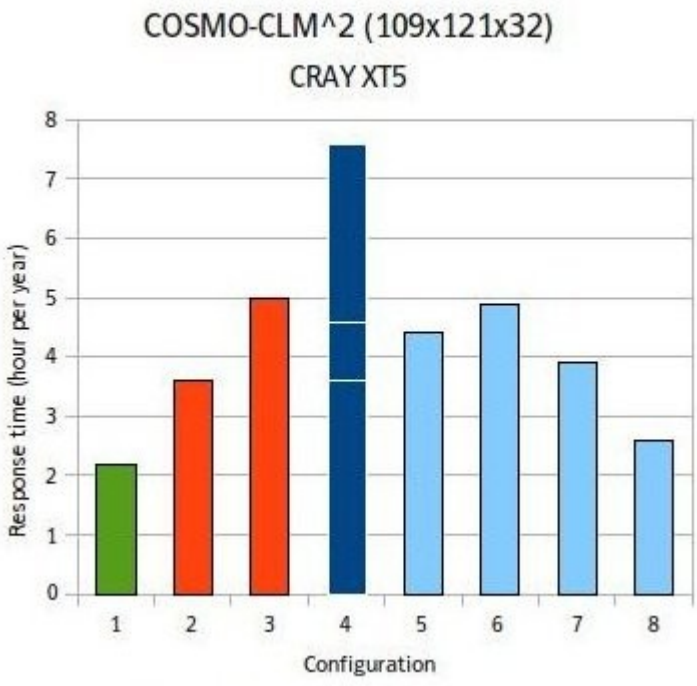

Illustration 2: Performances of various COSMO-CLM couplings (OASIS and non OASIS)

This figure shows the compared performances (elapsed time) between:

- the very first COSMO configuration, without CLM, using the native land model

TERRA, called as a subroutine on the same grid than COSMO (green box, n° 1 on 132 cores).

- the COSMO/CLM existing implementation (developed by ETHZ), where the transformed CLM model is called as a subroutine by COSMO (orange boxes, n°2 on 132 PE and n°3 on 60 cores)
- a serie of COSMO/CLM OASIS coupled configurations, tested during the present Dedicated User Support period. The dark blue box n°4 shows the performances of the very first configuration, using 12 PE for OASIS, 60 for COSMO and 60 for CLM, for a total of 132 cores. White lines represent, from bottom to top, respective COSMO, CLM and OASIS contributions to the total time.

Configurations 1, 2 and 4 are using 132 cores but, on the OASIS-based configuration 4, COSMO and CLM calculations are done on 60 cores only. The extra cost due to OASIS coupling is then deduced from the comparison with configuration n°3, where COSMO and CLM models also compute on 60 cores: this extra cost reaches 50%.

However, on a CPU consumption point of view, it is more suitable to compare the performances of the OASIS-based / OASIS-less configurations on the same number of total resources (configuration 1 and 2): then, the OASIS configuration is more than 2 times slower than the previous COSMO/CLM coupling (configuration 2) and more than 3 times slower than the initial COSMO/TERRA run (configuration 1).

Optimizations are definitely necessary to reduce this important extra cost.

Coupling frequency

For practical reasons, the CLM time step model was initially the same than the atmosphere time step (240s). This similarity could have a scientific justification but an increase of the land model time step could also be tested: CLM time step (and COSMO-CLM coupling) were set to 1h[3].

Designing our interface, we chose to call, at each time step, coupling fields sending and receiving routines. This permits to change easily[4] the value of the coupling frequency.

Blue box n°5 of figure 2 shows a significant improvement: most of the time is now spent on COSMO, essentially because CLM and OASIS are called 15 times less often.

Obviously, this modification must have an impact on model results. Those consequences would have to be analyzed by ETHZ users, if this configuration is chosen.

Coupling sequence

Another heritage of the previous COSMO/CLM “by subroutine” coupled configuration is the

3 When a model is called as a subroutine of the other (coupling previously developed on configuration 2), both models should have the same time step (or buffers have to be implemented to accumulate coupling fields). With OASIS, model time steps are independent and coupling frequency can be changed with a simple directive on parameter file

4 Just modifying “namcouple” OASIS parameter file (second section, coupling period per field and lag index). See OASIS3 user guide. If OASIS “prism_put” sending routine is called at each time step and LOCTRANS-AVERAGE option is activated, OASIS ensures the necessary accumulations of coupled quantities

sequentiality of calls (model calculations are done one after the other).

Again, it is particularly simple to change the OASIS coupled sequence and do both model calculations in parallel[5]. COSMO and CLM could process calculations of a given time step at the same time.

The corresponding performances are shown on blue box n°6 of figure 2. It represents the total duration of the slower model, increased by a fraction of the time necessary for coupling.

As for coupling frequency, model behavior modifications, induced by the new coupling strategy, has to be further investigated.

The last two optimizations can be jointly set and performances enhanced again (blue box n°7 on figure 2).

Compared to the previous “by subroutine” coupling, the OASIS multi executable approach let us choose the best parallelism for each model (according to their own scalabilities). Launching COSMO on 132 cores, CLM on 60 (and OASIS still on 12), we reach the most efficient configuration at this resolution (blue box n°8 on figure 2). The total duration is now comparable to the initial COSMO stand alone configuration.

Current limitations

Limits of our Dedicated User Support exercise forbid the tuning of all the possible parameters of our implementation.

1. An explicit process mapping is possible on CSCS Cray XT5 machine[6]. Given that a sensible spread has been observed in our performance measures, it is possible that a mapping which would take into account communication density between PEs and their position on the machine would change those performances.

2. Theoretically, coupling frequency could be different for each coupling field but some light modification will be necessary on the code to ensure it.

3. OASIS proposes a large variety of interpolations. The conservative one has to be chosen for some quantities (fluxes). Others could be tested to enhance performances.

5 Both models are now using coupling fields calculated by the other model at the previous coupling time step. At the first time step, CLM has now to read the initial coupling fields on a file. This file could be created by COSMO on an previous independent run, activating an optimization option on the OASIS interface (oas_cos_vardef.F90 file). This operation has to be done once: at the end of each run, OASIS creates a restart file with coupling fields of the last coupling time step. This is this file that has to be used at the beginning of the next run.

6 Each process of the coupled configuration could be assigned to one particular core, among nodes reserved through SLURM batch scheduler. Notice that, if the machine has been initially configured for such purpose (on SGI Altix “jade” CINES machine, for example), a multi-threading (using more than one process on one core) could significantly reduce the amount of necessary resources without changing performances: actually, two sequentially coupled models can share the same resources, because calculations are processed one after the other.

But the main limitation affects perspective on resolution increase, particularly for meteorological applications (MeteoSwiss is one possible user of the COSMO-CLM OASIS configuration), given that OASIS3 already exhibits lack of performances on some previously developed configurations (see for example Dedicated User Support report #4 on EC-EARTH high resolution CGCM).

OASIS3 interface for CLM4

Rationale

Coupling modularity is one the most appreciable feature of OASIS. Once an interface is written on a model, due to the implementation non intrusiveness, it is relatively easy to maintain it on the successive versions of the model. In addition, if one model has to be upgraded, nothing has to be done on the other side to keep using the coupled system.

Version 4 of CLM is available through CESM integrated system. The land model stand alone configuration is no longer available, and the whole system (land model + coupler + driver + atmospheric variable forcing module) has now to be coupled with OASIS.

Strategy

Popularity of the OASIS framework mostly relies on its capacity to make the use of an external model as simple as the reading of a forcing dataset.

That is exactly the philosophy of this new CLM-COSMO coupling.

Build from a CLM stand alone configuration case (I_TEST_2003), our CESM coupled model mainly consists on the driver, the prognostic land model and a “data models” (DATM for atmosphere). The main function of data models is to read forcing files. Modules are linked to the driver using the CPL7 internal coupler, which ensures remapping or interpolations, if necessary.

The only CESM code modifications necessary for an OASIS coupling consists in:

- defining DATM module grid on the original CLM grid, through a forcing file which holds variables not provided by COSMO atmosphere (aerosols)
- organizing OASIS exchanges through this DATM module

The CESM code, coupled with OASIS, still consists on its original components. Code modifications (communications with OASIS) mostly take place on DATM module. As shown on figure 3, the red arrows, which represent the OASIS connections, only connect the atmosphere data model (DATM) rectangle.

Implementation

An exhaustive description of our Fortran interface implementation and input files modification/addition is given in annex 1. This paragraph only summarizes the principle of the CESM modifications needed to build the OASIS interface.

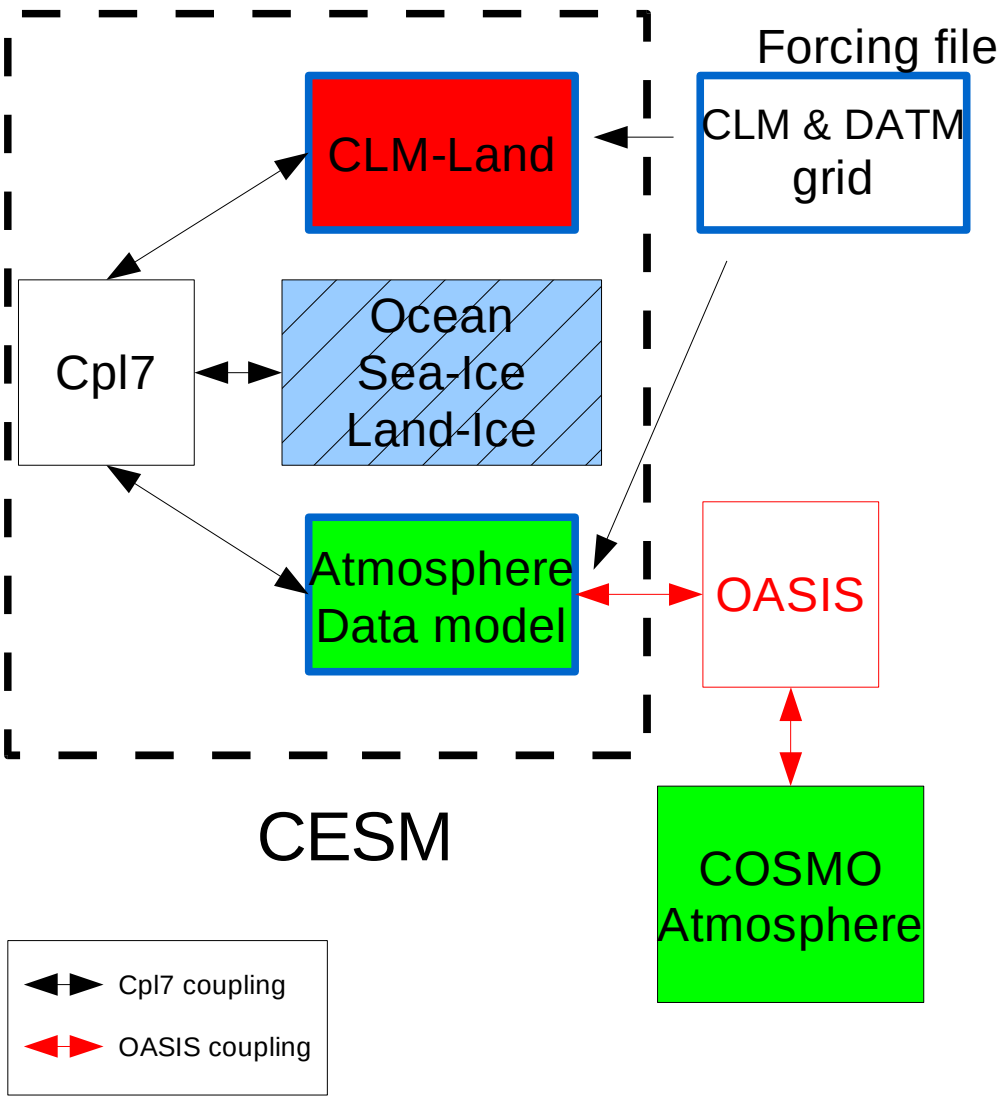

Illustration 3: Description of CLM (CESM) / COSMO coupling using OASIS

We choose to start from a CLM stand alone CESM configuration (I_TEST_2003).As a first step of the OASIS interface implementation, we modify the file which contains the atmospheric forcing fields. We interpolate the input file variables, describing them on the CLM original grid. Aerosols (not given by COSMO) are the only variables actually used by the model: the other variables will be overwritten by the OASIS coupling fields.

In this way, DATM module has now the same spatial discretization than CLM land model (it defines its own grid according to the dimensions read on the forcing files). Then, the CPL7 functions will be limited to remapping (no interpolation between land and data atmosphere models) and only if decompositions of both components (CLM and DATM) are different.

OASIS interpolations (with atmospheric grid) are defined for the CLM discretization:

- On an initialization phase, grid mask and coordinates are communicated to OASIS, at the same time than names of exchanged coupling fields.
- On main temporal loop, and at each time step, OASIS “send” and “receive” primitives are called through the DATM module. To convey land model variables

there, a driver modification is necessary: those variables have to be (potentially) remapped as if the prognostic atmosphere model was active.

Notice that the standard MPI management has to be slightly changed at initialization phase: CESM is not supposed to use the MPI_COMM_WORLD communicator, and its driver is forced to work with a local communicator (provided by OASIS). Consequently, a predefined OASIS routine is called by the CESM driver to let it switch off MPI.

Advantages

1. Simplicity: As previously said, rapidity and non-intrusiveness of implementation are a strength of OASIS. To call a set of initialization, declaration, sending, catching and ending OASIS interface routines adapted to CESM, we only had to modify 2 driver subroutines (ccsm_driver.F90 and ccsm_comp_mod.F90) and 1 DATM file (datm_comp_mod.F90).
2. Modularity: No other modification is required on COSMO and OASIS code or on their input files (in use on the previously set up CLM3.5 / COSMO / OASIS coupled model).
3. Scalability: taking advantage of the internal DATM parallelization, which could be adjusted independently of the CLM one, just changing a namelist parameter, OASIS exchanges could be made on a variable number of PEs. Figure 4 shows that the OASIS exchanges cost remains constant with parallelization (but expected to grow significantly at higher resolution with decomposition of more than 100 sub domains). On the contrary, it appears much more efficient to parallelize the DATM module (less than 0.01s on 122 PE but 0.3s when DATM runs on only 1 PE). Slowing down (reducing parallelism) occurs on remapping between CLM and DATM through the driver (driver_l2c, driver_a2c, driver_c2l, driver_c2a) but, above all, during DATM reading (“strdata_advance”) and scattering (“datm_scatter”).

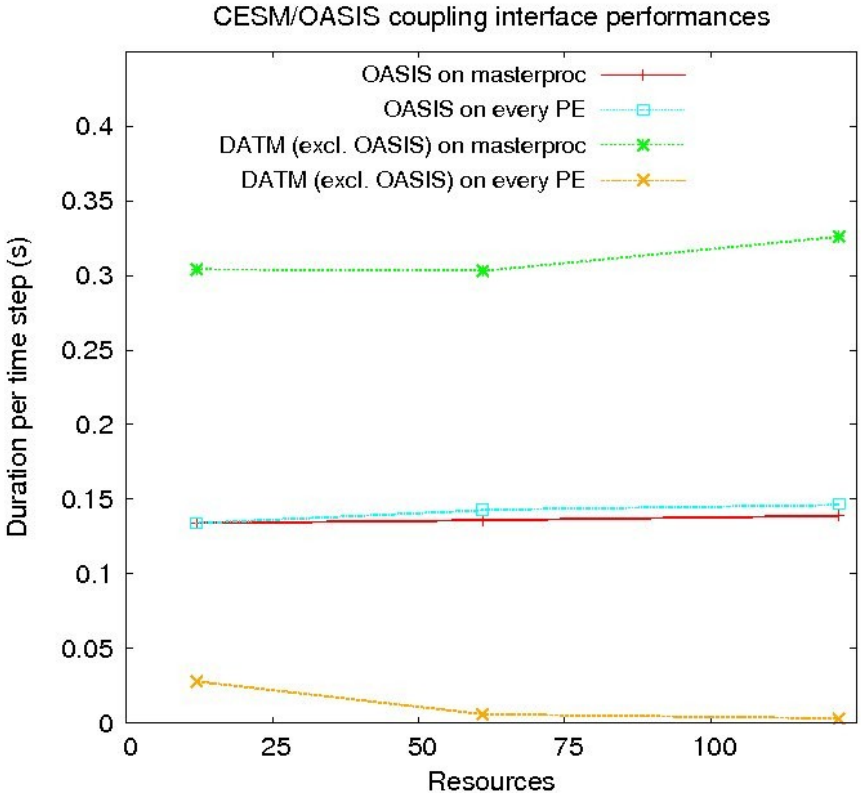

Illustration 4: Performances of OASIS exchanges (communications + interpolations, two ways) and of DATM routines (excluding OASIS send/receive calls)

4. Extensibility: on figure 4, a blue box represents different CESM modules, disabled in the present configuration. Theoretically, the same OASIS coupling interface (on DATM module) should allow us to exchange, with COSMO, information coming from (and given to) ocean, sea-ice or land-ice modules. To go further, the same interface may be implemented on other data modules (like DOCN) to ensure an OASIS coupling of the only CESM module that could not be plugged in the present configuration: the CAM atmosphere model.

Current limitations

The present implementation only addresses problems of version update, allowing ETHZ to keep using their CLM model in an OASIS coupled system, though the new CLM version cannot be used easily without the whole CESM framework.

Considering low size of the targeted configuration, we prefer to focus on implementation facility rather than on performances, in order to facilitate management, by user, of next version updates.

Consequently, the system general design (presented in figure 3) strongly suggests that an further increase of parallelism (with higher resolution model) would lead to a lack of performances.

1. OASIS3 restricted parallelism (one OASIS process per coupling field) is not sufficient when problem size increases
2. CLM/DATM internal coupling, though efficient, increases the total time needed to exchange information between CLM and COSMO
3. As on any other OASIS coupling, MPI process (from the different executables) mapping on the reserved resources could significantly affect performances, which makes mandatory a fine and, possibly, difficult tuning
4. In addition, OpenMP could not be used, at least without code and/or MPI launcher settings modification

It is obvious that extra developments are necessary to be able to increase resolution and parallelism of the system. Will they be sufficient ? This question is the concern of the larger debate of compared advantages of integrated/composite coupling.

Anyway, the OASIS capacity to make the *use* of an external model as simple as the reading of a forcing dataset should not hide that, once a technically validated configuration is available, a substantial work, including modifications of models parametrization, is then necessary to take into account the newly created coupled phenomena.

Annex 1: OASIS3 interface implementation on CESM

Added routines

oas_clm_vardef.F90	CLM/OASIS interface global variable definition
oas_clm_init.F90	Let OASIS organize MPI initilization
oas_clm_define.F90	Communicate model information to OASIS at initial step
oas_clm_finalize.F90	Let OASIS organize MPI ending
send_fld_2cos.F90	Fill arrays with coupling fields and call oas_clm_snd for each coupling field sending
oas_clm_snd.F90	Send one coupling field to OASIS
receive_fld_2cos.F90	Call oas_clm_rcv for each coupling field catching and fill model arrays
oas_clm_rcv.F90	Receive one coupling field from OASIS

Modified routines

To find these modifications on the code, see CPP key “COUP_OAS ”

ccsm_driver.F90 - Let OASIS close MPI communications , calling oas_clm_finalize
ccsm_comp_mod.F90 - Let OASIS define local MPI communicator (instead of MPI_COMM_WORLD) , calling oas_clm_init - Launch internal coupling routines to bring information to DATM module from other modules (force the prognostic atmosphere case) and particularly from land model to be able to use this information, sending it to OASIS
datm_comp_mod.F90 - Communicate model characteristics (grid lat/lon and mask, subdomain distribution per process) to OASIS, calling oas_clm_define - Prepare coupling fields and send coupling field to OASIS (through send_fld_2cos routine) - Receive coupling fields from OASIS (through receive_fld_2cos routine) after reading complementary forcing fields (aerosols) and overwrite appropriate arrays with corresponding information

Compiling on CSCS system

Change CSCS batch_script (/project/s193/emaison/cesm1_0_3/scripts/batch_cscs.sh)

1. start from "I_TEST_2003" CLM configuration
2. indicate 2 new include directory and library (OASIS) to the CESM compile script:

```
export USER_FFLAGS="-DCOUP_OAS -I/users/emaison/oasis3/CRAYXT/build/lib/psmile.MPI1"
```

```
export USER_LDFLAGS="/users/emaison/oasis3/CRAYXT/lib/libpsmile.MPI1.a
                     /users/emaison/oasis3/CRAYXT/lib/libmpp_io.a"
```

3. copy OASIS interface fortran files (and CESM modified routines) into scratch compiling directory from /project/s193/emaison/CLM4/src/ directory to /scratch/rosa/emaison/testclm4/SourceMods/

Running on CSCS system

Launching directory: /users/emaison/COSMO4.8-CLM11-CLM3.5/run/clm4_EXP

Prepare new input files:

1. No need to change any oasis and cosmo parameter and input files (could be the sames than CLM3.5 coupling)

2. Change some values within input_clm/lnd_in parameter file:

```
 finidat = ' ' -> no restart
 fatmgrid              = 'data/surfdata_0122x0276.nc' -> CLM uses Europe grid, modified to match COSMO
mask
 fatmlndfrc            = 'data/surfdata_0122x0276.nc'
 fsurdat               = 'data/surfdata_0122x0276.nc'
```

3. Change values of oatm_input/datm_atm_in parameter file:

```
  dataMode       = 'CLMNCEP' -> same option than initial CESM config
  domainFile     = 'data/surf_datm.nc' -> read DATM data ( = OASIS coupling fields) on the same Europe grid
than CLM, modified to match COSMO
  streams        = 'OASIS.stream.txt 1 1 1 ' -> take same forcing information at any time step from parameter
file OASIS.stream.txt
  vectors        = 'null'
  mapmask        = 'nomask'
  tintalgo       = 'linear' -> those last 3 info for CLM/DATM interpolations (should not be used).
```

4. Build the "oatm_input/OASIS.stream.txt" fake parameter file:
This file allows to:
- read aerosols forcing file
- read other forcing variables. Those forcing values will be replaced by the coupling fields: they could be a simple copy of aerosols (or zero).
- define DATM model grid reading this file
The aerosol file defines the DATM grid. OASIS cpl fields are exchanged following this grid: That means that aerosol file defines the CLM grid (as seen by OASIS).

5. Build netcdf input files:

oatm_input/surf_datm.nc: this file holds lat/lon information for DATM grid. It could be built copying CLM variables from file: /project/s193/emaison/preproc_CLM/surfdata_0122x0276.nc

Original variables	LONGXY	LATIXY	LANDMASK	AREA	LANDFRAC

Copy names	XC	YC	MASK	AREA (converted to radian squared, x2.464E-08)	FRAC

oatm_input/aero_dummy.nc: with correct aerosols data on CLM grid. WARNING: for the moment, aerosols values are not correct. Build them with NCAR tools.

6. Change some values on original drv_in parameter file:

- the start date start_ymd (WARNING: COSMO/CLM calendars could be inconsistent)
- the total duration (in time step and not in days)
- the total task for both CLM and DATM modules. DATM task number could be equal to CLM total tasks (every PE are involved in the OASIS coupling), or equal to 1 (only master PE exchanges information through OASIS). WARNING: the number of PE involved in the coupling must be changed consistently on namcouple parameter files.

7. To build OASIS auxiliary files, DATM task number must be set to 1. Once the files are created, they can be saved and copied on the working drectory before launching the next simulation. Then, the OASIS auxiliary files procedure won't be activated no more. This second production phase appears more efficient if DATM task number is then set to CLM total tasks.

Annex 2: CESM/COSMO coupling fields

Coupling field	**OASIS naming rule (CESM interface)**	**Sent by**
surface temperature	CLMTEMPE	COSMO
surface winds	CLMUWIND, CLMVWIND	COSMO
specific water vapor content	CLMSPWAT	COSMO
thickness of lowest level	CLMTHICK	COSMO
surface pressure	CLMPRESS	COSMO
direct shortwave downward radiation	CLMDIRSW	COSMO
diffuse shortwave downward radiation	CLMDIFSW	COSMO
longwave downward radiation	CLMLONGW	COSMO
total convective precipitations	CLMCVPRE	COSMO
total gridscale precipitations	CLMGSPRE	COSMO
wind stresses	CLM_TAUX, CLM_TAUY	CESM
total latent heat flux	CLMLATEN	CESM
total sensible heat flux	CLMSENSI	CESM
emitted infrared (longwave) radiation	CLMINFRA	CESM
albedo	CLMALBED	CESM

Bonus mission
Nov 3 2011

Host: Matthieu Masbou
Laboratory: Bonn University (Germany)

Main goal: Provide support on the previously designed COSMO-CLM OASIS coupling and tutorial on general OASIS use

Main conclusion
Bonn University users of COSMO-CLM model ended installing their configuration and start coupling Parflow hydrographycal model to the OASIS based system.

Model / machine description

COSMO-CLM (here called COSMO)
This regional atmosphere model (COSMO v4.8, and its climate version, COSMO-CLM v11) is used by a large community in several central Europe countries (from which Bonn University). DWD, MeteoSwiss and several other meteorological agencies host the operational version of the model. Grid: centered on West Germany settlements. High resolution (10km) is targeted.

CLM
This land model is developed at NCAR (v3.5)

ParFlow
Hydrological model developed at Bonn University. Finer resolution are targeted (100m)

The described configuration is developed for the German project TR32 (joining Aachen, Bonn, Braunsweig, Köln and Juelich Universities). TR32 is focused on soil/atmosphere interactions at spatial scale from Km to cm square. Possible extensions could lead to include WRF and ICON to the initial coupled configuration.

Model is available on the Bonn University local cluster.

OASIS3 interfaces for COSMO-CLM and CLM-ParFlow

OASIS3 interfaces on both CLM and COSMO models have been derived by Prabakhar Shresta (Bonn University) from the previous implementation described on Dedicated User Support reports #5 and #6)

He implemented a new functionality on OASIS (COOKING stage) to ensure efficient downscaling between coupling field exchanged between highly different spatial discretization scales (Schomburg at al. 2010). This development has been proposed to the OASIS development team.

He is currently modifying COSMO spatial discretization to match perfectly CLM grid requirements (due to no possibility of grid stretching on CLM model).

For this coupling , IS-ENES support only consists in useful bypasses or advices and corresponding report to OASIS users such as:

- mpp_io / OpenMPI 1.2 mismatch on previously designed cluster (and Intel compiler). Solution consists in disabling mpp_io features and providing Moritz Hanke's bypass (see Dedicated User Support reports #5 and #6)
- NOBSEND option disabling for large buffer exchanges
- incompatibility of OASIS3 pseudo parallel mode and OASIS grid writing functionalities (solved by OASIS3 ETHZ modified version providing)

Concerning CLM-Parflow coupling, a first implementation is currently developed by Mauro Sulis and a quick overview of the implementation state has been done. OASIS3 version use allows parallel coupling on CLM3.5 (instead of master-processor-only coupling, implemented at ETHZ). This option should allows TR32 users to enhance coupling performances when a fully parallel version of OASIS3 will be practically available.

Parflow C-language written code benefits from a C encapsulated version of the PSMILE routines (also developed at Bonn University).

Difficulties have been expressed by developers on topics such as:

- prism_put/get positioning on the newly coupled Parflow model and on CLM (for Parflow exchanged coupling fields)
- OASIS restart functionality
- prism_def_var_proto argument characteristics

The OASIS support gives us opportunities to:

- better explain characteristics of the previously developed OASIS interface and ensure diffusion of IS-ENES realization
- identify usual difficulties consecutive to OASIS interface implementation and parametrization
- report unknown malfunctions
- evaluate Bonn University OASIS related work and possible contributions to OASIS further enhancements

Mission #9
Feb 6- Mar 2 2012

Host: Uwe Fladrich
Laboratory: SMHI, Norrköping (Sweden)

Main goal: Measure and enhance performances of the OASIS3 based Ec-Earth model

Main conclusion
A portable and OASIS3 pseudo-parallel mode compliant version of our OASIS performance measurement tool has been developed and tested on several SMHI models. Thank to it, it could be soon possible to measure and compare the coupling extra cost of the standard OASIS3 based version and the presently implemented OASIS3-MCT based version of the Ec-Earth high resolution model. At the same time, different interactions contributed to set up two new OASIS3 coupling with regional models (RCA-NEMO and RCA-RCO).

Model / machine description

SMHI's coupled model (high resolution version) originally deals with:

- IFS, cycle 36: T799, 843.490 grid points, ˜25Km, 62 vertical levels, time step: 720s
- NEMO, v3.3: ORCA025, 1.472.282 grid points, ˜40Km, 45 vertical levels, time step: 1200s
- OASIS v3 (pseudo parallel)

20 coupling fields are exchanged between the two components at a coupling frequency of 3 hours. The model is available on Ekman supercomputer, 1.268 compute nodes of 2 quadripro AMD Opteron (# 10.144), Infiniband interconnection, located at Royal Institute of Technology (KTH), Stockholm, center for parallel computers (PDC).

OASIS3-MCT upgrade

Set-up during #4 Dedicated User Support[7], the Ec-Earth OASIS3 based configuration was still slowed down by coupler, and its performances supposed to be strongly reduced on machine allowing massive parallelism.

For several reasons, the replacement of OASIS3 by OASIS3-MCT has been preferred to the firstly envisaged OASIS4 upgrade.

7 Maisonnave, E. and Valcke, S.: OASIS Dedicated User Support 2010, Annual Report ,Technical Report, TR/CMGC/11/28, SUC au CERFACS, URA CERFACS/CNRS No1875, France (2011)

Started on the ekman machine, the replacement process consisted in a very few operations:

- into code interfaces (a single mod_prism module has to be called instead of a suite of specialized module)
- on namcouple (simplified due to the fact that OASIS3-MCT is currently not able to calculate interpolation weight, but only to read it on a file, which name must now be specified on namcouple)

Several FORTRAN philosophy related inaccuracies (argument array dimensions) have been corrected on IFS and NEMO coupling interface to be able to reproduce, with OASIS3-MCT, the identical coupling process, including coupling field restart read/write.

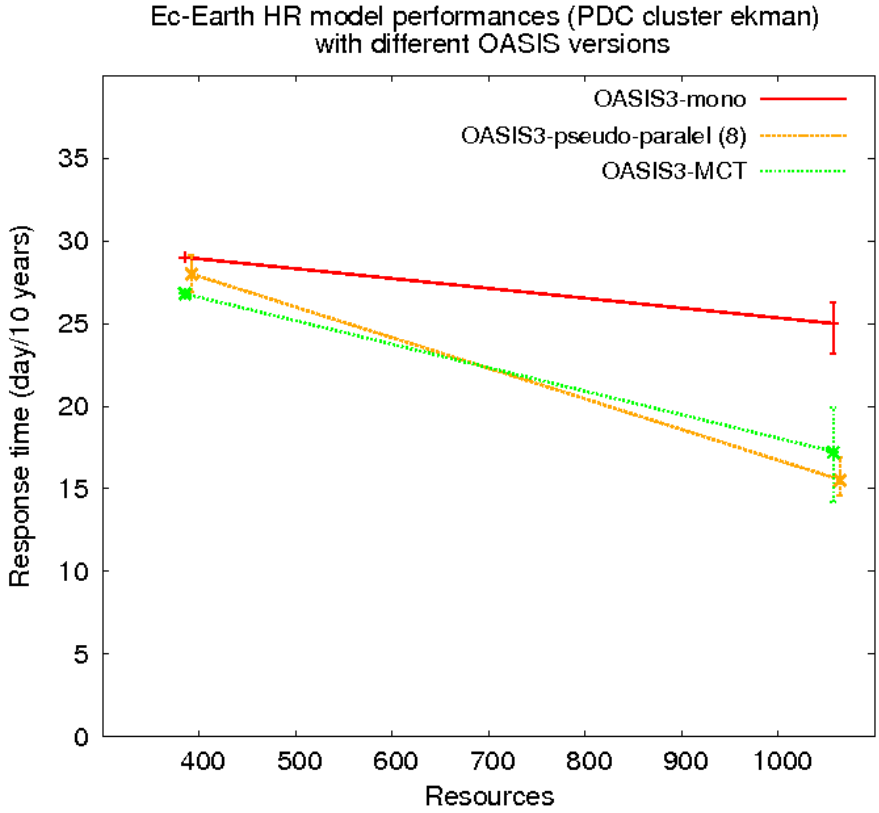

Illustration 5: Ec-Earth performances with coupler upgrade

On Figure 5, measurements of coupled model performances (total simulation time) are shown for both OASIS3 and OASIS3-MCT based configurations. As a reminder, performances taken with an OASIS3 mono-process coupling (all coupling fields are exchanged through a single OASIS3 process) are also displayed.

Due to the Ec-Earth coupling sequence (IFS and NEMO run in parallel), a large amount of the time needed for coupling (MPI message exchanges + interpolations) is hidden by the speed difference between the two components: most of the coupling operations are performed when the fastest model has ended its computations and before the slowest model has ended theirs.

Combined to the relatively low parallelism level of the configuration (reducing the amount

of exchanged messages during coupling and consequently the time needed to perform them), this particularity forbids to get a clear idea of how profitable our coupling technique enhancement is. Nevertheless, a small fastening is observed comparing our o(1000) parallel tests.

To better estimate the OASIS3-MCT benefit, several new experiments have been performed. The first idea was to increase parallelism (to increase coupling exchanges and test both coupler ability to manage them). To do so, a larger machine was required: PRACE tier-0 machine “curie” has been targeted and the model ported on it by John Donners (SARA) through the IS-ENES/PRACE IP1 joint project. Unfortunately, the results of this work has been delay by several “curie” operating defaults (machine upgrade) and can not be presented here.

The second idea was to change the coupling strategy and run atmosphere and ocean sequentially: this technique has the advantage to clearly make appearing all the time needed to perform coupling, as each model is waiting calculation results of the other one and OASIS operations needed to bring the information to its receiving interface. But this configuration could be difficult to set up and has no particular scientific interest for Ec-Earth teams. Consequently, test has been done (on “curie”) using the widely used CERFACS coupled model ARPEGE-NEMIX. Results clearly show the interest of an OASIS3-MCT upgrade at such level of parallelism[8].

Even though such result could be easily extended to the ocean-like and atmosphere-related model Ec-Earth, we preferentially would like to show the reduction of coupling time induced by the OASIS upgrade on the Ec-Earth model itself.

OASIS3 performance measurement toolkit

To reach this goal, a precise evaluation of the coupling communication and interpolation cost is required. Such quantities could be evaluated using the OASIS dedicated support development presently available on current coupler release. Activating the “balance” CPP key during OASIS3 compiling, MPI_Wtime clock time measures are printed on “prt” OASIS output files.

During the simulation, each time than a model sends or receives a coupling field, a clock time is output before and after the corresponding PSMILE library call. Symmetrically, the same measures are printed from coupler side.

On a post processing phase, a shell script (sh_balance) is used to convert the different measures into synthetic informations. Despite several advantages, proved by its capacity to provide all the previous performance related informations published in the previous OASIS Dedicated User Reports, the increasing level of models parallelism, but also coupler parallelism (OASIS3 pseudo parallelism) reveals the limit of a shell based development.

For those reasons, we decided to entirely re-write our tool in FORTRAN-90, ensuring its

8 See PRACE IP1/IS-ENES WP8 joint project web site: https://redmine.dkrz.de/collaboration/projects/prace/wiki/_OASIS4_upgrade_

portability at the same time than its capacity to process results produced on massively parallel systems. Nevertheless, given that each model process, at each coupling time step and for each coupling field, writes a certain amount of ASCII format information on files, performances could be affected by a relatively large amount of disk access.

OASIS instrumentation

To partly avoid such drawback, an simple enhancement on OASIS implementation consists in suppressing FORTRAN “flush” routine call after each write file access call. But possibility must be given to the user to keep this functionality when an on-line analysis is required.

A second enhancement on OASIS implementation is necessary to measure the possible time shift between the different clocks of nodes allocated to our coupled model. This issue is particularly difficult to address, and an exact synchronization impossible to achieve. We assume that a simple measure after the coupling initialization phase MPI_Wait call (common to all process) will fit our precision requirements.

Those two enhancements will be soon available on OASIS3 official distribution.

Post-processing tool

The FORTRAN executable is called through a simple shell script, which ensures portable compiling (-c option) at the same time than execution. Completing the analysis, the graphical tool “gnuplot” is used (if available) to produce a simple EPS format visualisation of the main results.

As previously described, each time than a coupling field is exchanged by any process involved in the coupling, two clock measures are produced on the corresponding “prt” file[9]: one before calling the PRISM sending or receiving routine, and one after.

Those two standard measures are read twice by our FORTRAN program.

On a first step, our program identifies which coupling field is exchanged by which model and counts how many time it is. This first reading allows us to determine the arrays dimensions which will contain the information to process:

- the number of exchanged fields[10]
- how many times are they exchanged

The field exchange sequence (as seen by coupler) is deduced and displayed on standard output[11]. This information will help the user to check whether this sequence matches the sequence defined on the models. If not, it means that buffered MPI communications are

9 There is one “prt” file per coupled model process (model or coupler)

10 Equal to half the number of fields exchanged on all coupler executable (they could be several if OASIS pseudo-parallel mode is enabled)

11 On OASIS pseudo parallel mode, the first fields are those described on namcouple_1 file

probably activated.

Program also checks that each field is received (by coupler or by a model) as often as it is send (by a model or by coupler). If not, a message is displayed to inform the user that simulation did not end correctly. Consequently, further analyses will be systematically done excluding the last two coupling step. Symmetrically, the first coupling step is also excluded to not take into account restart operations duration that could slow down the simulation beginning.

For those reasons, the total simulation elapsed time (as observed with a simple unix “time” command) is greater than the figures given by our program.

On a second step, information available on “prt” files is read again. Now, the purpose is to fill the two different arrays allocated with the previously defined dimensions.

One dimension of those two arrays is 8. This figure matches the measure number necessary to trace every operation done to carry a given coupling field from one model to the other, i.e. before (a) and after (b) source model send, before (c) and after (d) coupler receive, before (e) and after (f) coupler send and before (g) and after (h) target model receive.

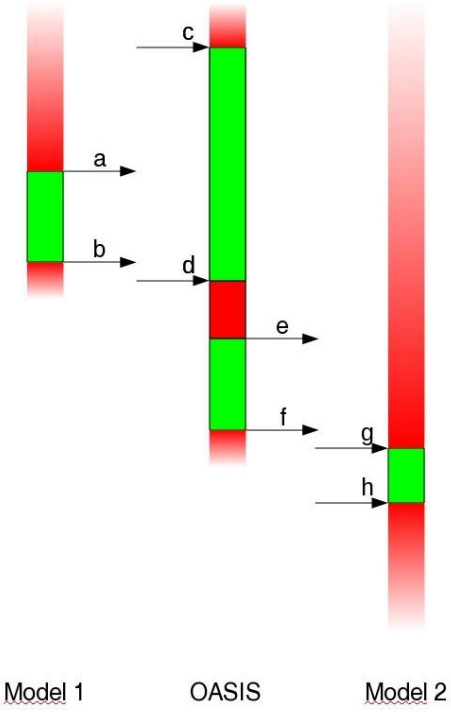

Illustration 6: Naming convention for coupling field exchanges

Load balancing

For each process involved in coupling and for each coupling field, our first data array is filled with the cumulated value, along all the valid coupling steps, of 3 quantities:

- the time needed to send the coupling field: timing (b) – timing (a)
- the time needed to receive the coupling field: timing (h) – timing (g)

– the time needed to perform all the other operations:
  timing (a) – timing_from_previous_cpl_time_step (b) (for source model)
  timing (e) – timing (d) + timing (c) – timing_from_previous_cpl_time_step (f) (for coupler)
  timing (g) – timing_from_previous_cpl_time_step (h) (for target model)

The first two operations gather the time needed to write or read coupling fields on MPI buffers but also the time spend to wait the moment when those exchanges are allowed. After adding quantities for all coupling fields, this time could be seen as the time that models need to exchange the coupling fields. Consequently, the third operation could be seen as a time when coupling independent operations are performed[12]. For more convenience, we will call it “computation” time.

As different model (and coupler) MPI process could start or end the different operations at different moments, we choose to selected the maximum duration from every process.

On coupler side, the important figure lies on this computation time: it measures interpolations and other operations speed. It cumulates the time spent by each coupling field: on OASIS pseudo-parallel mode, it does not take into account the fact that several coupling field are processed at the same time. On any mode, if SEQ namcouple option is the same for each coupling fields, our calculated quantity adds several times the duration when OASIS is performing all the interpolation of fields with same SEQ parameter. For all those reasons, this OASIS calculation time could generally not be considered as the total elapsed time needed to perform the various calculations, but better as a measure of how fast a subset of coupling fields is computed. On pseudo parallel mode, this quantity must be compare with itself for several machines, or for several model parallelism. On OASIS mono-processor mode *only*, it could be seen as the total time needed to perform interpolations, and only when SEQ parameter differs from one coupling field to the other.

Even though adaptations are needed to take into account special (and quite non standard) cases[13], we could test these analysis on two different SMHI OASIS3 based coupled models: Ec-earth and RCA-NEMO. For the first example, simulation has been arbitrarily stopped before the end defined in the different namelists and namcouple. Our tool don't need a complete simulation to give its results and can be launched on working directory even during the simulation.

12 It is not exactly equal to the total time needed by the model to perform a forced simulation, on a stand alone mode because, in this case, coupling field reading is replaced by forcing field reading

13 For example, when coupling time step differs from one model to the other

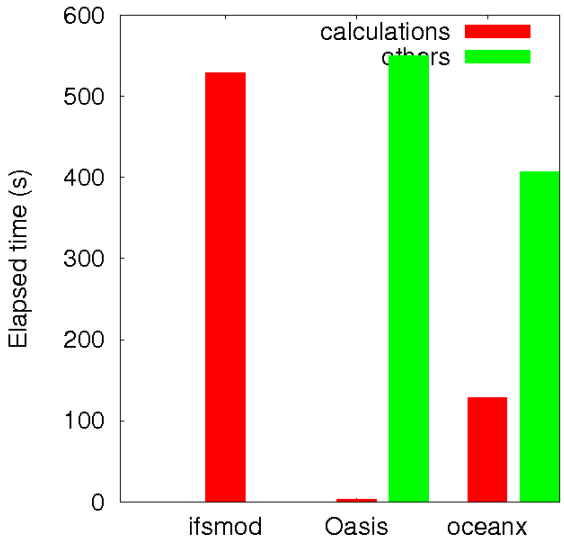

Illustration 7: Balance analysis for IFS/NEMO coupled model

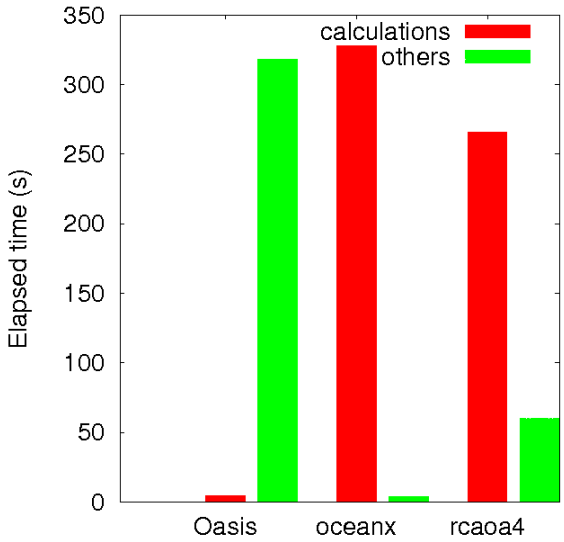

Illustration 8: Balance analysis for RCA/NEMO coupled model

On graphics proposed above[14], red boxes represent "total" calculation time, and green boxes "total" time needed to send, receive or wait the coupling fields.

Coupler calculation time is negligible compared to model computation time. Matrix multiplication (interpolation) of global grid can be quickly processed. On RCA-NEMO configuration, OASIS is used on mono-processor mode and each coupling field has different SEQ parameter: the calculated quantity adds calculation time of each coupling field. On IFS-NEMO, the OASIS pseudo parallel mode is enabled and all coupling field of one coupler instance have the same SEQ parameter. This could explain that IFS-NEMO value of OASIS computation time remains lower than RCA-NEMO one, despite its finer resolution.

Comparing computation time for model of a same configuration allows us to conclude that RCA and NEMO response times are much more balanced than IFS and NEMO ones. To avoid that a model spends too much time waiting to the other, it is recommended to allocate more resources to the slowest model. To make his (her) choice, the user has also to take into account the relative scalability of the two (or more) coupled system components.

One interesting aspect of our tool is that, comparing the different computation performances measured with several resource numbers, one could establish, simultaneously, scalabilities of the different components. This method is much more rapid and accurate than measuring those results with the stand-alone models, mainly because forced and coupled configuration scalabilities could differ.

Coupling efficiency

A second array is produce during the second part of our program.

14 Those pictures are automatically produced by our tool (EPS format), if gnuplot is available on the processing machine

For each coupling time step, and for each coupling field, we select, for each model or for coupler, the latest timing produced by any of its MPI process for the following quantities:

- the time for the source model needed to write the coupling field to MPI buffer + the times needed for the coupler to read it and write the interpolated field + the time for the target model to read it.
- The time spent by any model (or coupler) to wait coupling field, needed to keep calculating.

The first quantity is equal to:

(timing (d) – timing (a)) + (timing (h) – timing (e))

But timing (c) – timing (b) must be subtracted from this quantity each time than timing (b) is prior to timing (c), i.e. each time that OASIS is waiting the model to start interpolations. Identically, timing (g) – timing (f) must be subtracted too from the same quantity each time that timing (f) is prior than timing (g), i.e. each time than source model is waiting the coupler to start its new calculations.

Consequently, in the previously described case, we can increase the total time when target model (and/or coupler) is waiting an information with the quantity:

timing (g) – timing (f) and/or timing (c) - timing (b)

One should notice that, when OASIS is waiting coupling fields from source model, the target model can wait the coupler at the same time. It means that the addition of the two quantities is not a measure of how much the fastest model is slew down by the coupling (and the slowest model)[15] but how much all the coupling operations are slew down by the chosen coupling sequence.

Extension to OASIS3-MCT

Due to time limitation reasons, it has been impossible to design an identical tool for the new OASIS3-MCT version. A preliminary step will be necessary: a coupler instrumentation allowing to print timing on output files.

Consequently, at the end of the Dedicated Support period, it was still not possible to compare OASIS3 and OASIS3-MCT on EC-Earth high resolution model. Nevertheless, CERFACS plans to end this work next time than an efficient OASIS3 / OASIS3-MCT comparison will be necessary (ANR PULSATION project, for example).

Other debugging activities

In addition to EC-Earth, two other OASIS based couplings are currently implemented at SMHI: the regional RCA-NEMO model and the previously OASIS4 based regional RCA-RCO model.

At this occasion, we could expand our data base with OASIS possible misleading features,

15 i.e. timing (g) – timing (f)

from which the possibility given to the user to:

- read OASIS restart file with wrong dimension (unclear failure)
- send and receive halos in addition to working arrays (high possibility of shifts in the coupling field communication)
- open existing interpolation weigh file in unix “write” mode (could be protected)
- spend a lot of time searching the FORTRAN north_thresold non parametrizable variable and its right value (1.6) to avoid weigh calculation anomaly at north pole (SCRIP conservative interpolation)
- hang his/her simulation only when using NOBSEND exchanges at some high parallelism level
- choose between misleading ways to globally conserve (CONSERV operation) spatially weighted and cumulated values
- use mask values (“masks” file) different from those previously used to calculate weights
- choose SCRIP related “bins” number without a clear diagnostic on how seriously an interpolation could be affected near sub-domain boundary if this number is insufficiently high (and, at the opposite, how too much time consuming is an insufficiently low number)

Those different questions, asked by five different persons (mainly with permanent positions) all along the Dedicated User Support period, give a good information on general users strategy chosen to set up an OASIS coupling.