OASIS Dedicated User Support 2012
Annual report

E. Maisonnave[+], R. Hill[o], O. Jamil[o], S. Valcke[+]

[+]CERFACS, [o]The Met Office

TR/CMGC/13/18

Abstract

Three new Dedicated Support periods, respectively at BTU (Brandenburgische Technische Universität, Cottbus), UBO (Université de Bretagne Occidentale, Brest) and the MO (Met Office, Exeter), concluded the IS-ENES 4-year-long collaboration program focused on coupling techniques for climate modelling, between CERFACS and European laboratories.

At BTU and UBO, coupling field exchanges with sophisticated components has been made possible taking benefit of OASIS modularity and efficiency, respectively a two way nesting 3D coupling between global (ECHAM) and regional (COSMO) atmospheric models and a global coupling between atmosphere (ARPEGE) and ocean (NEMO) including an AGRIF zoom focused on Northern Atlantic region.

Even though ECHAM-COSMO 3D coupling still needs additional developments to allow vertical interpolation between grid components and a thorough scientific validation, our study proves the efficiency of the new OASIS3-MCT coupler for low-resolution coupling at frequency (global model time step) never reached before.

With ARPEGE-NEMO special grid (stretched Gaussian grid on atmosphere, parent/child double grid on ocean), we tested the limits of OASIS3 standard interpolations. Nevertheless, it was possible to provide a solution that avoid creating artificial gradient at parent/child frontiers.

Both works (in addition to ETHZ previous Dedicated Support) reveal that some further developments would be necessary to make OASIS more friendly-user for regional configurations.

Our collaboration with the MO leads to the conclusion that, at least in the medium term, OASIS3-MCT offers a suitable coupling solution for the planned high resolution models. This new version of the coupler is able to drive reproducible simulations, running interpolations 7 times faster than the previous OASIS3.3 version.

The particular nature of the Dedicated User Support once again gave us the opportunity to give advices and help overcoming set-up issues on new configurations such as adding a wave model on the Met Office climate model or assembling a coupled regional configuration.

Thanks to Andreas Will, Stefan Weiher, Eberhard Schaller (BTU), Markus Thuerkow, Ingo Kirchner (FUB), Edouard Davin (ETHZ), Jennifer Brausch (DWD), Jean-Guillaume Piccinalli (CSCS), François-Xavier Bocquet, Mick Carter, Catherine Guiavarch, Chris Harris, Adrian Hines, Mike Hobson, Matthew Mizielinski, Steve Mullerworth, David Pearson, Jean-Christophe Rioual (the Met Office), Anne Marie Tréguier, Claude Talandier (UBO), Julie Deshayes, Eric Machu (IRD), Clément de Boyer Montaigut (Ifremer), Stéphane Sénési, Silvana Buarque (Météo-France), Sébastien Masson and Rachid Benshila (LOCEAN) for their support and their interest to our work. For the last time, thanks to the non nomadic and essential part of our team, Sophie Valcke, Laure Coquart (CERFACS) and Anthony Craig.

Estimated carbon emission diagnostic for those 3 journeys by terrestrial and public means of transport: 300 Kg

Mission #10
Aug 14- Sep 7 2012

Host: Andreas Will
Laboratory: BTU, Cottbus (Germany)

Main goal: Set-up OASIS3-MCT interfaces in global and regional atmosphere models to enable a 2-way nesting

Main conclusion

Based on OASIS3 previously developed interfaces (for MPI-OM, CLM and NEMO coupling), a 3 dimensional OASIS3-MCT coupling between COSMO and ECHAM has been set up. Observed performances (without advanced optimisation, the coupling overhead on a mono-node IBM Power6 simulation is about 6% of the total elapse time) reveal that an OASIS3-MCT tight 3 dimensional coupling (the coupling frequency is equal to the ECHAM time step) is inexpensive at low resolution, and therefore most probably acceptable at high resolution.

Model / machine description:

COSMO-CLM (here called COSMO)

This regional atmosphere model (COSMO v4.8. In its Climate Mode: COSMO-CLM v11) is used by a large community in several central Europe countries, from which Eidgenössische Technische Hochschule Zürich (ETHZ). Deutscher Wetterdienst (DWD), MeteoSwiss and several other meteorological agencies host the operational version of the model. The grid size is 221x111x47, i.e. ~2 degrees. A decomposition on 32 MPI tasks was used for tests on the target supercomputer.

ECHAM/MPI-OM

The Max Planck Institut für Meteorologie (MPI-M) global atmospheric model is here used in its 6th version. The starting configuration already includes an ocean model (MPI-OM), coupled at Deutsche Klimarechenzentrum (DKRZ) through OASIS3-MCT. The grid size is 192x96x47 (T63) for ECHAM and 254x220 for MPI-OM. A decomposition on 32 MPI tasks was used for tests on the targeted supercomputer.

OASIS:

Both OASIS3 (version 3.3) and OASIS3-MCT (version 1.0) has been used during this work.

Those models are available on IBM Power6 supercomputer, which has 8,448 compute cores (16 dual-core processors per node) and an Infiniband interconnect (peak performance of 158 Teraflop/s). The machine is located at DKRZ, Hamburg, Germany.

Rationale

An increasing number of studies presently require models with more accurate horizontal and vertical resolution. But global high resolution CGCM are, in many cases, not affordable in terms of required human and computing resources. The solution consisting on a local increase of resolution on regions of interest can satisfy most of the present scientific project requirements.

Such zoom can be defined on a sub-domain of the model grid, where calculations are refined. WRF atmosphere or NEMO ocean models, for example, both proposed integrated solutions to allow one way (the inner model is forced by boundary conditions provided by the largest model) or two way nesting (in addition, the largest model considers updated information produced by the inner model).

Researchers of the Cottbus Brandenburgische Technische Universität (BTU), in collaboration with the Berlin Frei Universität (FUB), both belonging to the COSMO community, proposed to use two different models for global (ECHAM) and regional (COSMO) modelling and to take benefit of OASIS to exchange the information necessary for a 2 ways nesting between these models.

Independently of various scientific issues (buffering, filtering ...) not addressed in the present support, a clear challenge of such coupling is the huge amount of information (3D fields) exchanged at a very high frequency (largest model time step). This requirement (exchange of large volume of data) is the major reason why integrated coupling, i.e. the two models are merged into one executable and the data is exchanged through the memory, is presently preferred instead of an external OASIS coupling, in different laboratories such as Météo-France (coupling between ARPEGE atmosphere and SURFEX land surface) or LOCEAN (for the coupling between ocean and LIM sea ice, or PISCES Biogeochemistry).

OASIS interfaces

Both COSMO and ECHAM models already included OASIS3 interfaces.

Previous OASIS Dedicated User Supports (see missions #5 and #7) led to the implementation of OASIS3 interfaces in the Community Land Model (CLM) and in NEMO ocean model (see mission #9).

In ECHAM, a recent DKRZ upgrade to OASIS3-MCT of the atmosphere-ocean coupling with MPI-OM gave a base for an extension to the presently described ECHAM-COSMO coupling.

First implementation

A preliminary study was hosted by the Centre Suisse de Calcul Scientifique (CSCS), gathering three COSMO community members, Edouard Davin (ETHZ), Jennifer Brausch (DWD) and Andreas Will (BTU) as COSMO coordinator.

In order to ensure COSMO capability (I) to exchange simultaneously coupling information with three different models (NEMO or MPI-OM ocean, CLM land surface and ECHAM AGCM) and (ii) to possibly extend coupling to other models, it was decided to implement a single and modular OASIS3 interface for all coupling.

Following NEMO example, the two coupling operations (collecting information to send out and filling model variables with received information) are achieved by common routines for all the coupling fields. The different operations are launched following the type of coupling chosen by namelist (ocean coupling, land surface coupling and/or atmospheric 2 ways nesting). Unfortunately, despite numerous but probably unclear advices of the author, developers did not follow NEMO example, and coupling choices are hard coded with respect to a limited set of models. Consequently, it is not possible to select by namelist different coupling fields for a given type of coupling. For example, the planned coupling between COSMO and MPI-OM ocean model will not be possible through interface implemented for the coupling to NEMO ocean model. A duplication of the code implementing the interface will then be unavoidable.

The simultaneous use of several models in our coupled system imposes the choice of the same version of the coupler for all components (given that OASIS3 and OASIS3-MCT cannot be used together in one coupled configuration). For performance reasons, it seemed more promising to choose OASIS3-MCT as a common version of coupler.

Nevertheless, the existing OASIS3 interfacing was kept as an option in COSMO, through a CPP key. This ensures, for the moment, the possibility of coupling COSMO and version 4 of the Community Land Model.[1] The high similarity between OASIS3 and OASIS3-MCT Application Programming Interface (API) makes the amount of additional code necessary to propose this modularity very limited.

After one week of joint developments with A. Will, S. Weiher (BTU), M. Thürkow (FUB), J. Brausch (DWD), E. Davin (ETHZ), J.-G. Piccinali (CSCS) for model performance measurement and the author for OASIS support, a unified interface is now available in COSMO for coupling to ECHAM, NEMO and CLM.

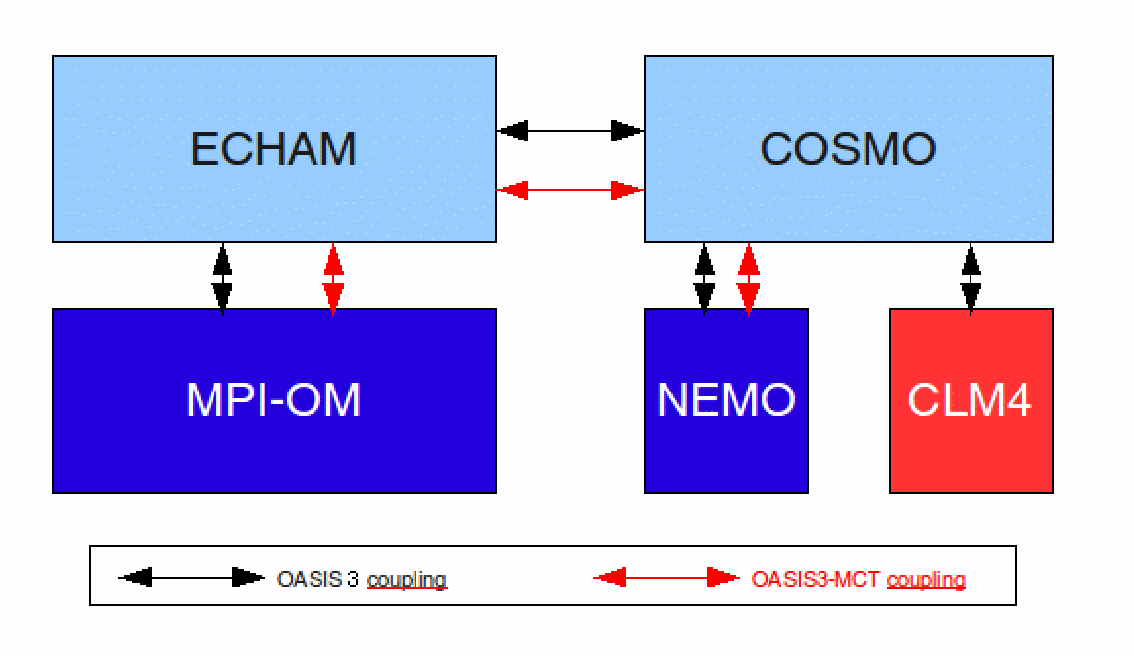


Fig 1: COSMO-CLM coupled constellation currently implemented

[1] CLM4 is a module of the CESM coupled model, which also uses MCT for internal parallelism. For the moment, no clear solution was found to allow simultaneous use of MCT library in CESM and OASIS.

Two ways nesting

The ECHAM interface

The targeted global/regional coupled model also includes coupling with the MPI-OM ocean, thanks to the pre-existing ECHAM-MPI-OM coupling based on OASIS3-MCT. In ECHAM, the existing OASIS3-MCT interface was extended to exchange necessary information with COSMO.

Duplication of the existing interface was chosen by ECHAM developers. In addition to the existing MPI-OM related mo_couple.f90 routine, a similar routine was created for COSMO coupling (mo_couple_two_ways_nesting.f90)

As previously reported (see mission #1), the existing interface does not take advantage of OASIS per-coupling-field accumulation option and accumulates independently the coupling field values according to a single pre-defined coupling frequency.

A mixed system of coupling field declaration (oasis_vardef) depending on namcouple (MPI-OM part) and model namelist (COSMO part) allows the simultaneous use of the two interfaces. Nevertheless, a high level of code duplication and the impossibility to use this code in the previous configuration (ECHAM/MPI-OM coupled model without COSMO) strongly foresees the future (and advisable) rewriting of a common interface, following NEMO or, now, COSMO examples.

Launching script

As usual in such coupling exercise, when models are coming from different communities, one of the two compiling and running environments must be chosen and adapted to the other model needs.

Concerning compiling, due to time limitations, both existing environments were kept, even though it appeared clearly that ECHAM's “Integrating Model and Data Infrastructure” environment (IMDI) is much more adapted to a common production workflow than to such development/debugging exercise. MPI-M team must be contacted to discuss a solution.

Thanks to Ingo Kirchner (FUB) developments, a COSMO style launching script (template) was delivered during the OASIS support period and a COSMO/ECHAM/MPI-OM coupled simulation could be quickly launched: after a first COSMO stand alone run, ECHAM and MPI-OM input files and adapted namelists are created in the COSMO working directory, and the coupled simulation is launched with the COSMO additional component.

A modification was necessary in the coupled template script to take into account the additional two-way nesting coupling fields in the initial ECHAM/MPI-OM namcouple. Optionally, our script allows to exchange the 3D fields as single 2D fields at each level or, together, via a single communication.

Coupling strategies

In our configuration, regional and global models perform their calculations at the same time: the two models are running sequentially (see fig 2.) It means that, unlike most of the presently implemented OASIS coupling in our climate community, any slow down during the coupling operation has an equivalent impact on the total coupled model restitution time. It is then particularly important to try to minimize any coupling overcost.

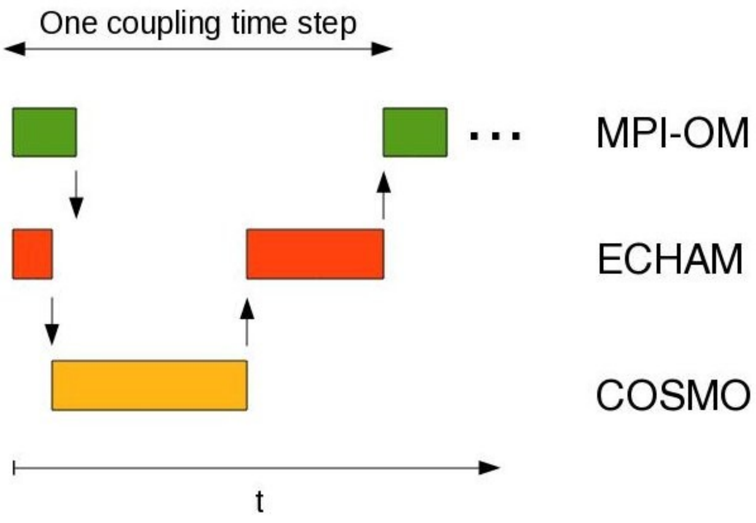


Fig 2: COSMO/ECHAM/MPI-OM coupling sequence

Pseudo 3D coupling

Present COSMO/ECHAM main coupling characteristic is the 3 dimensional size of most of its exchanged fields (see Annex 1). A clear concern of such configuration is the time needed to interpolate and exchange the whole information at each time step of the global model. Compared to a standard coupling, several improvements were necessary to reach an acceptable level of performances.

One usual way of improving the coupling performance is to reduce the amount of exchanged information. Considering that models have a different number of vertical levels, one improvement could be to perform the vertical interpolation by the source processes before the exchange if the target model has less vertical levels, or by the target processes after the exchange if the source model has less vertical levels in order to minimise the amount of information transferred. For the moment, the described coupling configuration does not include this vertical interpolation but only *n* horizontal interpolations, with *n* equal to the number of ECHAM vertical levels (47).

On a preliminary intent of coupling optimisation, an OASIS3-MCT namcouple parameter was adapted to our particular case. Its functionality consists in coupling multiple fields via a single communication (see §B.4 of OASIS3-MCT User guide). Inside OASIS3-MCT, fields are stored together and a single mapping and a single send or receive instruction are executed for the whole group of fields. We will call this technique a “pseudo 3D coupling”.

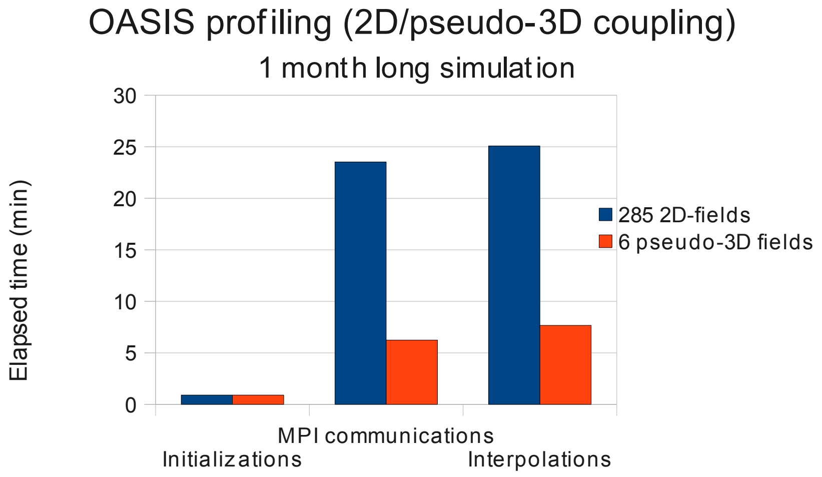


Fig 3: OASIS profiling

The benefit of such improvement is clearly visible in the given profiling (fig. 3) made with the embedded OASIS3-MCT measurement tool (routine set from mod_oas_timers file). Measurements can be turned on by changing a code variable (TIMER_debug) before compiling. With TIMER_debug=1 or 2 timing files are produced at the end of the simulation (one for the root process of each model, one for the other processes of each model). In each file, one paragraph summarizes the total elapsed time spent by several group of OASIS3-MCT routines, that can be gathered into 4 categories:

- initialization (map_definition, cpl_setup, cpl_smatrd and advance_init groups, called only once at the beginning)
- communications (psnd_* and prcv_*, one per coupling field)
- interpolations and other operations (pmap_*, pcpy_* and pavg_*)
- coupling restart writing (wrst_*) if any (it is no more the case in our coupled configuration)

These diagnostics are made on each process. Minimum, maximum and mean values, considering all process values, are displayed in the root process result file. We choose to only consider the maximum value, even though this strategy could lead to over-estimate the OASIS3-MCT impact on performances[2].

Initialization is negligible in our case. Interpolations and communications during coupling exchanges share the total time spend on OASIS (see figure 4).

We compared such quantities for 2 one-month long simulations that only differs by the way the coupling fields are exchanged: 285 2D-fields or 6 pseudo 3D-fields, each pseudo 3D field being composed of 47 2D fields (+ 3 2D-fields). Considering figure 3, it clearly appears that the huge number of MPI messages required in the first case strongly slows down the simulation: communications then takes 24 min. The solution that consists in grouping the small messages into a bigger one must be preferred; the communication then takes only 6 min. Similarly, an interpolation (matrix multiplication) performed with bigger arrays seems faster (7 min) than several small sequential operations (25 min).

[2] Profiling of “receive' configuration is ambiguous for some coupling fields: it can both include communication time and waiting time (the target model is waiting for the source model to finish its calculations and provide the coupling fields)

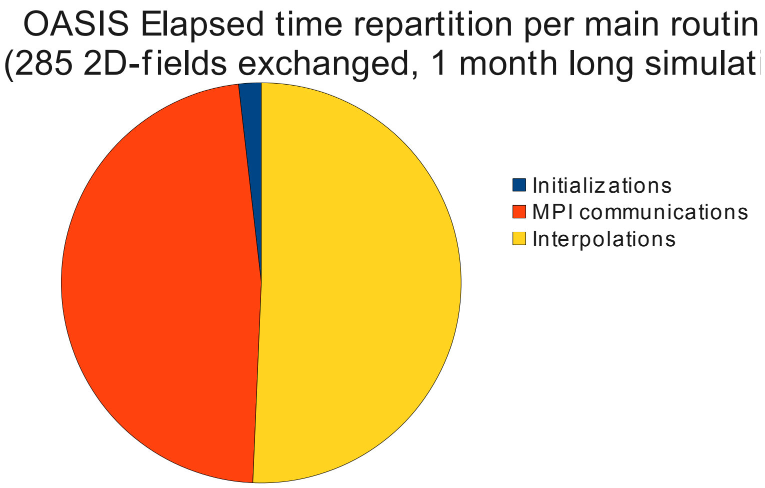


Fig 4: Profiling repartition between OASIS routine categories

The total extra-cost of OASIS3-MCT coupling could be deduced doing a difference between:

- a coupled run where exchanges are limited to only 1 2D-coupling field (where OASIS coupling cost is reduced to less than one second)
- and our full 285 2D-fields or 6 pseudo 3D-fields configurations.

From figure 5, we can evaluated the OASIS3-MCT overhead to 50% in the first case (285 2D-fields) and 6% in the second (6 pseudo 3D-fields). This last figure tends to prove that OASIS3-MCT is quite suitable for such 3D coupling necessary at each time step of our 2 way nesting. This result must be extended to any OASIS3-MCT coupling composed of many similarly interpolated fields (and not necessarily only different vertical levels of the same variable).

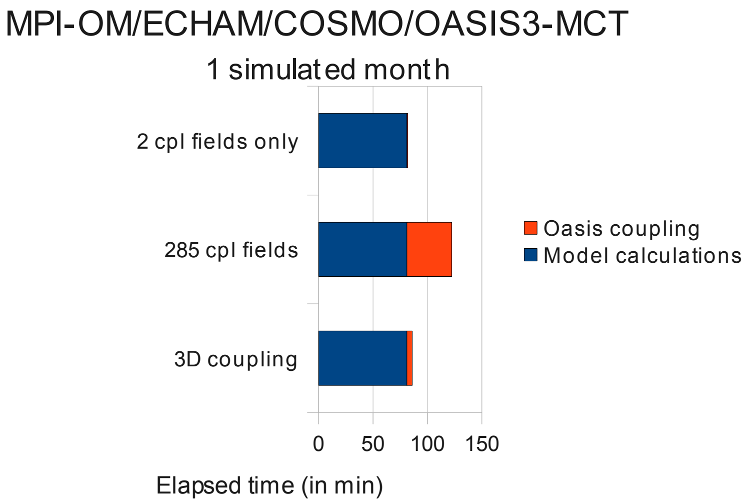


Figure 4: OASIS extra cost

An additional experiment made on more than a single node (64 processes spread on two different 32-cores nodes instead of 64 processed launched on the same 32-cores node) does not exhibit notable increase in OASIS communication (nor interpolation) time. It suggests that, even though we would increase parallelism for higher resolution configurations, using greater number of nodes, our experimental OASIS setup should still keep its good performances[3].

Hyper-threading

Another information can be deduced from this complementary experiment: an hyper-threading of our model processes gives comparable performances than allocating a core to each of them. This could be explained by the sequencing of our coupled system (see fig. 2).

MPI-OM and ECHAM run in parallel: the exchanged coupling fields are calculated by the other model at the previous coupling time step. A contrario, ECHAM and COSMO run sequentially: COSMO is called as a subroutine of ECHAM. These characteristics implies that models are idle half of the time. With hyper-threading (32 cores are allocated for 64 processes), 2 processes are sharing the same resource and using it alternatively. Any extra resource (for example when 2 nodes are allocated) is then useless and there is no gain in restitution time.

Current limitation, further developments

Interpolations choice

Default distance weighted nearest-neighbors interpolation (SCRIPR/DISWGT) was chosen in both ways. We approximately calculated that a minimum number of 4 neighbors[4] is needed from ECHAM to COSMO and 10[5] from COSMO to ECHAM. Those figures can be tuned, following the best compromise between quality and performances. On Fig 6 is shown the general aspect of a coupling field (high atmosphere temperature) before and after ECHAM/COSMO interpolation.

[3] Obviously, this also depends on machine interconnect network and MPI implementation.

[4] Minimum number of neighbours for a DISWGT interpolation: less will lead to produce strong gradient on the target grid at source grid point limits

[5] This figure is deduced from the average ratio between source and target grid point areas.

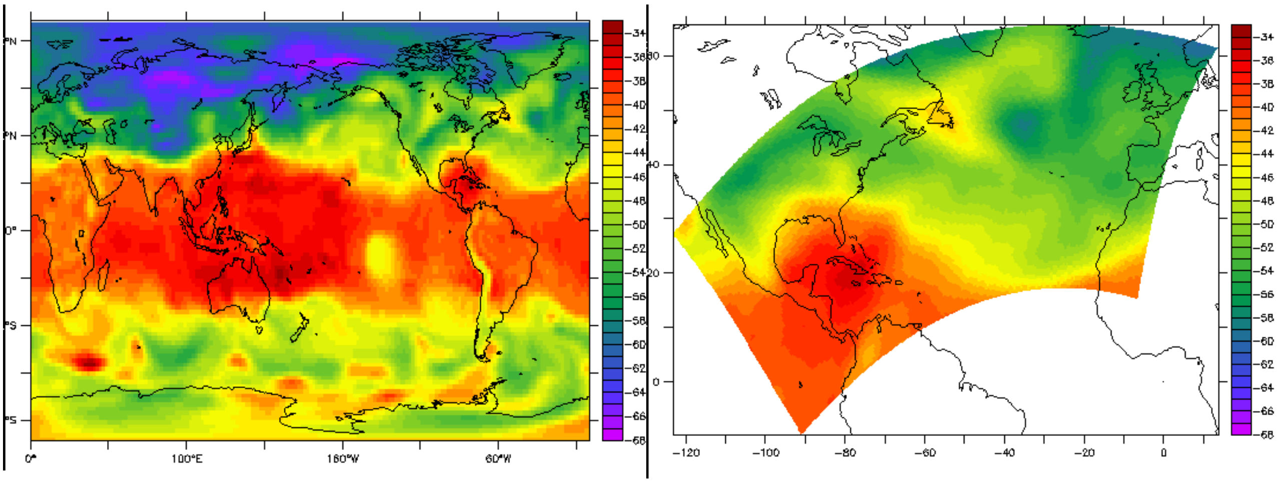


Fig 6: Temperature L28, on ECHAM grid (left) and after 4 neighbours DISTWGT interpolation on COSMO grid (right)

The standard interpolation strategy could be slightly modified to increase even more performances, and, particularly, reduce the amount or the number of MPI communications necessary to exchange the coupling fields.

The easiest improvement consists of changing the location (source or target model) where interpolations are performed, using the MAPPING option, newly offered with OASIS3-MCT. It is expected that performing this operation such that data expressed on the coarser grid (instead of data expressed on the finer grid) is exchanged will reduce the number of exchanges; as observed previously, increasing the size and limiting the number of MPI messages improve the exchange speed.

Another modification is strongly recommended to reduce event further the communication load and the coupling overhead: to pre-select only the ECHAM grid points from the region involved in the coupling, i.e. corresponding to the extension of the COSMO grid. This can be done by building a new coupling mask for ECHAM and adding it on the list of existing variables of masks.nc OASIS auxiliary file. This modification will strongly limit communications and calculations in COSMO-ECHAM exchanges as these will be performed only for ECHAM grid points inside the COSMO region.

Another way to reduce the cost of ECHAM-COSMO exchanges is to declare an ECHAM partition so that only a subset of ECHAM sub-domains, corresponding to the COSMO region, will be involved in the coupling.

Suggestion of OASIS3-MCT improvements

A set of modifications on OASIS3-MCT code were necessary to couple ECHAM and COSMO. Most of them are related to hard coded limits reached in our configuration, due to its unusual high number of coupling fields:

- maximum number of coupling field on mod_oasis_var.F90 and mod_oasis_coupler.F90

- namcouple line length on mod_oasis_namcouple.F90 and mod_oasis_kinds.F90
- number of profiling measures on mod_oasis_timers.F90

Those modifications were reported to the OASIS developers team. A future release should replace hard coded limits by dynamically allocated arrays. This release will then fully support the present COSMO-ECHAM 2 ways nesting coupling.

Some improvements should be investigated by the OASIS developers team to enhance performances of the code. In particular, on the “oasis_coupler_setup” initialization routine, which main task is to found correspondence between model declared and namcouple declared coupling fields. It appears that the routine cost increases with number of coupling fields (up to 1 minute for 285 fields). A better matching algorithm would contribute to speed up our initialization phase and avoid substantial delays at higher parallelism levels.

This support clearly demonstrates that tight 3D coupling is possible with OASIS3-MCT, which should encourage in the future more global to regional couplings. But the current OASIS setup does not really take into account a specificity of such configuration. Given that one domain is larger than the other, interpolation and communication should be limited to the surroundings of the regional area. This cannot be automatically done by OASIS and needs user's intervention.

But this intervention should be, at least, guided. The user should be informed of the non negligible slow down than an automatic creation of interpolation weights (with SCRIPR) can produce. Some advices should be given on how to build an adequate mask on the greater grid and limit exchanges to a subset of global grid sub-domains. Additionally, a tool could be provided to help the user to build this mask. Ideally, an explicit option should be implemented in OASIS so to limit the remapping to the intersection of the two grid domains.

Mission #11
Oct 29- Nov 23 2012

Host: Richard Hill
Laboratory: Met Office, Exeter (UK)

Main goal: Optimise OASIS-MCT performances in the Met Office high resolution configuration

Main conclusion

OASIS3-MCT implementation in the Met Office high resolution global coupled model has been shown to produce bit-reproducible simulations. Coupling interpolations go about 7 times faster than the previous OASIS3 ones.

OASIS3-MCT appears to offer a suitable coupling solution, at least in the medium term, for planned Met Office high resolution models.

In addition, a support has been provided to overcome starting issues in adding a wave model on the Met Office climate model.

Model / machine description

Unified Model (UM)
Met Office's global atmosphere model includes JULES soil module. Grid size: 1024x769x85, N512. Parallelism has reached 1500 MPI tasks on targeted supercomputer with more than 50% parallel efficiency.

NEMO
The widely used European ocean model is here associated in a single executable to the Los Alamos National Laboratory CICE (sea-ice) model. They share the same ORCA025 grid (1442x1021) on a 75 vertical levels configuration and are used without NEMO IO server. NEMO parallelism is constrained by load balancing with respect to the UM.

A wave model, called WaveWatch II, is about to be integrated into the OASIS coupled configuration.

Those models are available on IBM supercomputer, which has 33,792 compute cores (4 height-core 3.8 GHz Power 7 processors per node) with Host Fabric Interconnect. 160 extra nodes (Monsoon) are accessible from Cerfacs for a total peak performance of 1.194 Petaflop/s. Machines are located at Met Office, Exeter.

Reproducible simulations with OASIS3-MCT

For reasons that cannot be detailed in this report[6], reproducibility is an important requirement for many laboratories, like the Met Office. Any coupled model cannot satisfy this characteristic if any one of its different components, including the coupler, does not.

For this reason, Met Office staff perform extensive regression tests on their successive coupled models to ensure that results are not unexpectedly altered when introducing new versions of the components, including the coupler. Performing the same verification with the new OASIS3-MCT coupler has been one of our first tasks.

To do so, two simple 3-day long runs were performed with two different atmospheric decompositions and their results compared.

A careful check of each component behaviour proved:

- OASIS3-MCT reproducibility, as “bfb” map strategy is set by default for any mapping interpolation used in our namcouple
- UM reproducibility, with any tested parametrization or processor arrangement
- NEMO reproducibility with -O2 optimization option, or -O3 if SOR solver is used instead of PGC. With version 3.4 of our ocean model, a -O3 option combined with the use of PGC solver should give reproducible results if pre-compilation uses the CPP key mpp_rep[7]. This still has to be checked on Met Office supercomputers.

Load balancing

The post-processing tool “lucia”, associated to OASIS for measurement of the coupled model load balancing has been installed on Met Office HPC platform, and saved in svn (FCM at the Met Office) repositories, for both OASIS3[8] and OASIS3-MCT [9].

This tool was initially developed in February[10] 2012 at SMHI (see Oasis Dedicated Users Support # 9) for OASIS3, then adapted to our new OASIS3-MCT version (and called “lucia-mct”).

Both versions require OASIS instrumentation by means of the same CPP key (“balance”) that prints MPI_Wtime based measurements in log files[11], before and after MPI send and

[6] Even though climate simulation reproducibility is considered crucial by a major part of our community, the OASIS Dedicated Support staff still keeps on thinking that the reason why it is so is not clearly established from a scientific point of view.
From the Met Office point of view reproducibility is a very useful tool which aids development by allowing bugs to be identified more easily and reducing massively the amount of time spent on testing and validation. E.g. if results change then traceability of results in climate science is such an important issue that we have to run extensive scientific comparisons using significant computing (and staff) resources in order to demonstrate that overall scientific evolution has not been affected.

[7] following Rachid Benshila (LOCEAN) instructions.

[8] svn://fcm2/PRISM_svn/PRISM_UKMO/branches/dev/ojamil/r899_load_balance_tool /utilities/oasis3/load_balancer/lucia/lucia

[9] svn://fcm2/PRISM_svn/PRISM_UKMO/branches/dev/ojamil/r899_load_balance_tool /utilities/oasis3-mct/load_balancer/lucia/lucia-mct

[10] a few weeks after the locally popular St Lucia day

[11] “*.prt” files for OASIS3, “debug.*” files for OASIS3-MCT

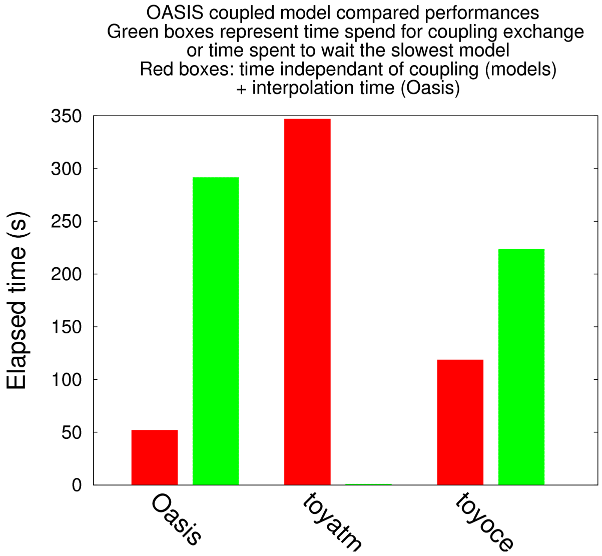
receive actions. This key is included in the OASIS3 release, but not yet in the OASIS3-MCT one. Necessary modifications were included in the Met Office dedicated version[12].

The compiling (-c option) and launching (from the working directory that includes log and namcouple files) of “lucia” and “lucia-mct” are the same but their results differ slightly.

Lucia and lucia-mct graphical output are shown on Fig 1 and 2. Pairs of timings are plotted in red (calculations) and green (waiting or coupling communication time). Where lucia produces three pairs of timing data (one per model and one for OASIS separate executable), lucia-mct only produces two, i.e. one per component which now includes the interpolation time.: what we call calculation time, for the OASIS3 executable, measures the time needed to perform interpolation(fig 1, first box). This time cannot be calculated by lucia-mct, because there is no coupler executable any more: this interpolation time is split into model calculation times. To be able to compare “lucia” and “lucia-mct” measurement, another performance tool measurement must be enabled in OASIS3-MCT (see next §).

On the Met Office IBM Power 7 machine, it has been observed that the production of ASCII output linked to the “balance” option in OASIS log file can significantly change execution time[13]. Timings are supposed to be printed at the very end of the run, which is not the case on this machine (even though no “flush” command is called). This is possibly a consequence of MPI environment variable options, that could be set up later.

For further use of our lucia tools, we emphasise that OASIS3-MCT load balancing cannot be processed with more than 2 models. It is theoretically possible, although untested, with lucia in OASIS3. Use of lucia for OASIS3 in pseudo-parallel mode is not supported either.



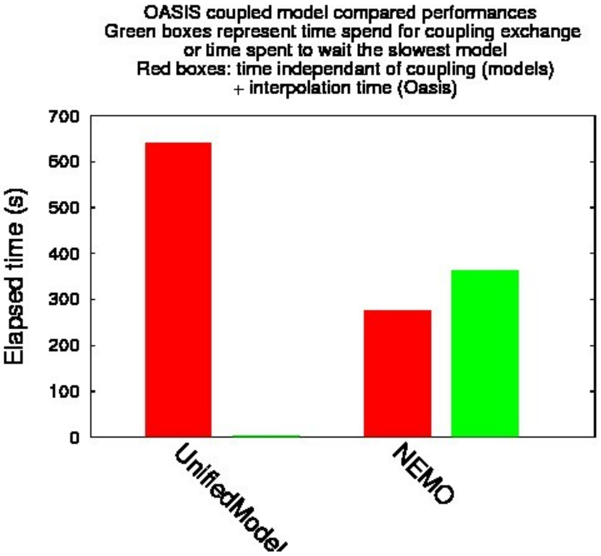


Fig1: OASIS3 coupling

Fig2: OASIS3-MCT coupling

Load balancing of High Resolution UM/NEMO coupled model

Nevertheless, load balancing with the same high resolution UM-NEMO model[14] has been done for two versions of the coupler. In both cases, the UM atmosphere was decomposed

[12] /data/cr/ocean/emaison/oasis3mct-cottbus on hpc2f supercomputer.

[13] A reproducible and significant reduction of elapse time about 10% was measured in high resolution simulations. The fact that enabling timing print makes the simulations faster suggests that synchronization of OASIS communications can probably be enhanced.

[14] Set up and made available by Richard Hill and Omar Jamil

on 20x91 partitions and NEMO on 10x12. Fig 1 shows total time for 5 exchanges, fig 2 for 12.

A clear concern is the total time needed by OASIS3 (52 seconds) to perform its interpolations, although this tends to be significantly reduced in the pseudo-parallel mode with 8 coupler instances (see next §).

Load balance could not be improved in either cases due to component model limitations in scalability and memory requirements: at this resolution, with the present optimisations, the UM cannot go faster. At the same time, NEMO cannot be decomposed on fewer than 100 sub-domains (which would slow it down) without exceeding the memory limit of a Met Office machine node.

OASIS performance measurement and comparisons

OASIS3-MCT interpolation and communication cost can also be evaluated by in-place measurement in the coupler interface library[15]. It can be enabled in released versions by setting the "TIMER_debug" variable to 1 (to obtain details for the master processor and an overall summary) or 2 (separate measurements for each MPI task) in file mod_oasis_timer.F90.

Time spent on groups of coupling operations, such as interpolations, are displayed in [model_name].timers_0000 files, for the master process. The minimum and maximum duration of all processes is also reported.

The released version of OASIS3-MCT routines at first caused deadlock when these time measurements were activated and gave wrong timing results on Met Office IBM Power7 machine. Bugs were reproduced with a high resolution toy model[16]. This toy was made available to OASIS developers, opening an external access to the "Monsoon" Met Office/NCAS part of the IBM Power 7 system. In parallel, to let us produce our measurements, a workaround was implemented in the Met Office dedicated version of the coupler.

A first analysis of OASIS3-MCT internal timing measurement, done with an N512/ORCA025 high resolution configuration, showed that interpolations (map_smat index) and send/receive operations (psnd_00x/grcv_00x) are significantly bigger than any other action. In particular, initialisation time was found one order of magnitude smaller than the total time needed to exchange coupling field during the simulation. Excluding the receive operation of the first coupling field (which gathers MPI communication and model load unbalance), communication are one order of magnitude smaller than interpolation times.

But overall, in an 8 coupling exchange simulation, which execution time, excluding restart and termination operations, is about 15 minutes, the total cost of the most significant part (interpolations) of OASIS3-MCT, i.e. 0.3 seconds, is clearly negligible. Any further coupling strategy improvement described below has no major impact on simulation execution time.

[15] See OASIS3-MCT updated documentation, "Time statistics files" §).

[16] /data/cr/ocean/emaison/oasis3mct-cottbus/examples/toy_eric_pulsation

Interpolation cost was also measured with “lucia” on a similar configuration coupled through OASIS3 (see previous §). This configuration is based on OASIS3 running onto only one process and therefore differs from the most efficient pseudo-parallel case into which OASIS3 runs onto more than one process. An estimation of pseudo-parallel mode performance can be made by dividing the mono-process timing by the ratio of the total coupling field number over the maximum number of coupling fields on a single coupler process, which is about 5 in our case.

Given that the interpolation of 5 coupling exchanges with in the mono-process case takes 52 seconds (see fig 1), we can estimate that interpolation for one coupling exchange in OASIS3 pseudo-parallel mode would take about 2.3 seconds, i.e. 7 times more than the same operation performed by OASIS3-MCT.

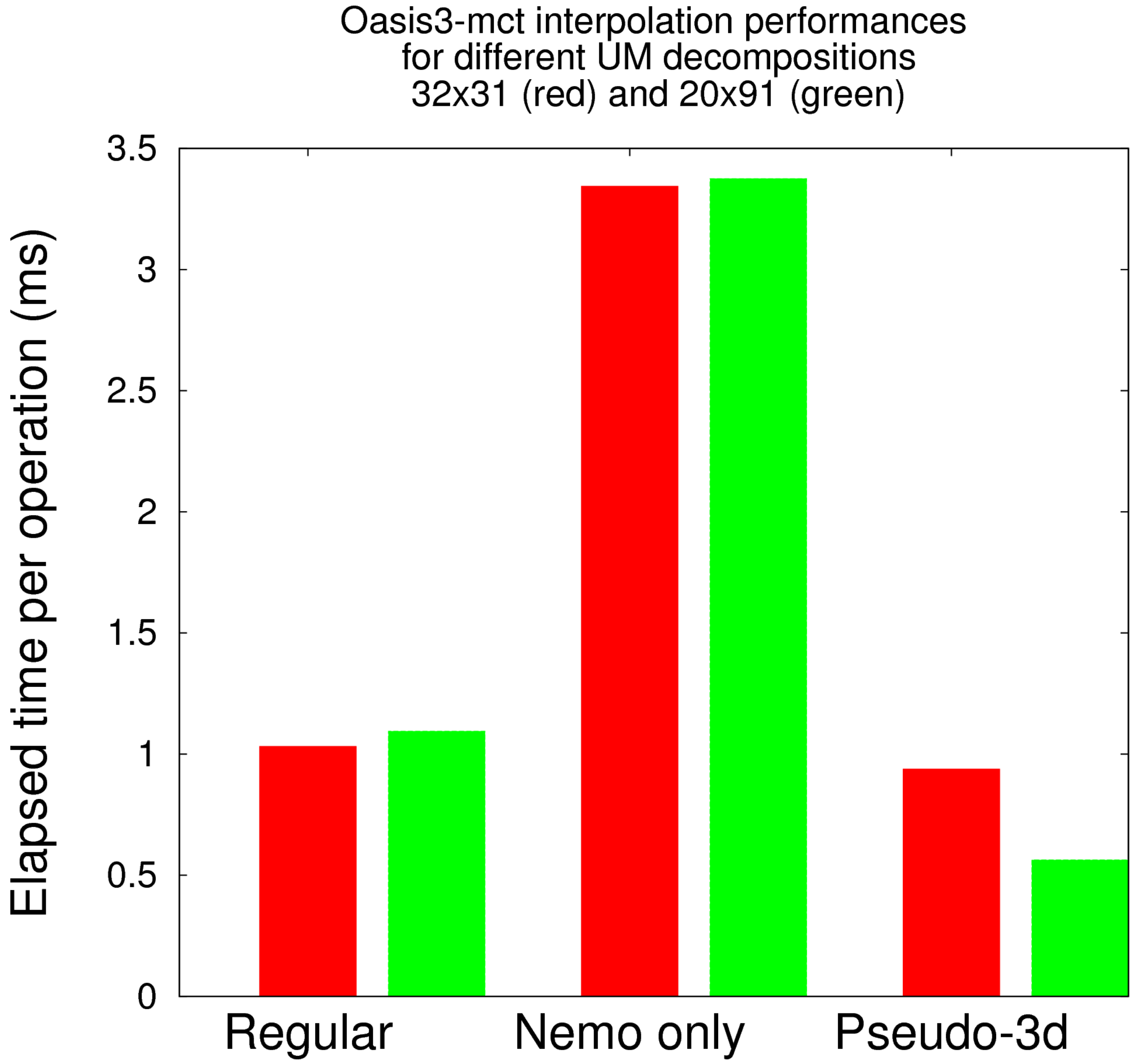


Fig 3: Single interpolation duration for 3 different coupling strategies

In the next step, we tried to further reduce OASIS3-MCT interpolation cost changing two parameters in our namcouple file.

As we know, there is no longer any dedicated coupler executable in an OASIS3-MCT coupling, which means that interpolations, like any coupling operations, are now performed by the models themselves (to be precise, by the linked OASIS3-MCT library routines). These operations can be performed either by the source or destination model processes. This can be defined by the user in the namcouple file (dst or src option of MAPPING operation, src being the default).

Since NEMO always completes its calculations before the UM, our first approach to parameter tuning consisted in performing all interpolations on the NEMO side. Different single interpolation timings (mean of 608) are shown on fig. 3, for two different atmospheric decompositions. Values for default namcouple parameters (the interpolation is done on the source model processes) are indicated by the first two boxes ("Regular"). The following two boxes show duration of interpolations all performed on the NEMO side. Their higher values can be explained by the decomposition difference between the two models: 992 (red box) or 1820 (green) for the atmosphere against 120 for the ocean. When half of the interpolations cannot be done by the most parallel model processes, performance worsens.

A new OASIS3-MCT functionality, so called "pseudo-3D", already tested successfully on the 3D regional/global coupling implemented with COSMO and ECHAM[17], allows several coupling fields sharing the same kind of interpolation to be grouped together in order to perform calculations and field exchanges in bigger arrays[18].

This has been activated for different ice related coupling fields (thickness, cover, temperature …) allowing a reduction in the number of coupling fields from 58 to 22. Results illustrated by the last two boxes of fig. 3 indicate that this tuning slightly speeds up interpolation operations, but only with the higher decomposition. This suggests that it is worth increasing array sizes when UM parallelism gets above a certain level (in this case, gets higher than 992 sub-domains). Better said, we suppose that when parallelism increases, the sub-domain array size gets too small for optimal computing or that the number of MPI messages gets too numerous for the interconnect network capacity. Over some limit, it should be interesting to switch on this OASIS3-MCT pseudo-3D option. But this limit, in our case estimated to be 992 sub-domains or more, needs to be investigated more precisely.

However, the impact of these adjustments was not visible in overall simulation elapsed times. We suppose that the OASIS3-MCT extra cost will be more significant at higher spatial and time resolution, and only if model load balancing can be reached on such future configurations.

Jean-Christophe Rioual started instrumenting OASIS3-MCT with "Dr Hook" measurement routines. This will allow an alternative method of assessing and understanding the different OASIS3-MCT task costs.

Various questions about OASIS

This support also contributed to identifying difficulties and features with OASIS3-MCT and

[17] See Oasis Dedicated User Support #10

[18] For details, see OASIS3-MCT User Guide, Appendix B (Changes between previous OASIS3.3 and new OASIS3-MCT

publicising to key Met Office staff some established OASIS3 solutions which may be adapted for use in new model coupling work currently planned or in progress at Met Office.

Global model

OASIS3/OASIS3-MCT upgrade caused an increase in model master processor memory requirement: at high resolution, about 88% of the available memory on the node against 76%. A memory leak has been found and fixed by Richard Hill a few weeks after the presently described OASIS user support.

OASIS3-MCT behaviour could slightly differ from the existing OASIS3. For example, it has been observed that coupling fields output with EXPOUT option are appended to output from previous runs if files pre-exist on the working directory. We suggest to overwrite them instead. The fact that, with OASIS3-MCT, files produced by the EXPOUT option contain data which is immediately viewable even if the model run ultimately crashes is seen as an extremely useful facility, compared with the equivalent situation under OASIS3 whereby such data was only visible if and when the job had completed successfully.

To conclude, even though not actually tested with UM-NEMO, we reported that OASIS3-MCT - Netcdf4 was already used on ARPEGE(or WRF)-NEMO configurations. Armed with this information, the Met Office aims to upgrade all relevant components to use Netcdf4 during 2013.

Regional models

We took the opportunity of a regional model group discussion to outline and clarify a preferred strategy for coupling regional configurations of atmosphere (with non regular grid) and ocean models. We agreed that conservative interpolation must be chosen to take into account the rapidly varying area of atmospheric grid points. We emphasised that interpolation at boundary can be an issue or, at least, lead to difficulties (see Oasis Dedicated User Support #10). For this OASIS3 based coupling, we informed the regional model group that the “lucia” tool is installed at the Met Office (see first §) for load balancing analysis.

Wave model

After participating to an OASIS training session, François-Xavier Bocquet started an OASIS3 coupling of the WaveWatch II (WW) model with NEMO. As a first step, the ocean model has been replaced by a toy, to check WW-OASIS interface (partition and coupling field declaring, exchange order and frequency, interpolation). After an initial issue due to a non explicit variable naming on both model and OASIS sides, WW interface has been validated.

The next two steps have been defined (NEMO and UM coupling) which clearly pave the way to the first tri-model coupled configuration ever implemented at Met Office.

Mission #12
Nov 26- Dec 21 2012

Host: Anne Marie Tréguier
Laboratory: Université de Bretagne Occidentale, Brest (Brittany)

Main goal: To couple global + regional ocean/sea-ice model to atmosphere

Main conclusion

Model interfaces have been upgraded and special interpolations defined to properly exchange coupling fields between parent/child ocean and atmosphere models. A load balancing analysis of the coupled configuration has been performed during a one month long validation run.

Model / machine description

ARPEGE
Météo-France's AGCM is used in its stretched version (rotated pole centred on Baffin Bay, refined grid around North Atlantic area with stretched factor of 2.5). Grid size: T127 truncation (24,572 grid points), 31 vertical levels, 5.4 version, including SURFEX land model.

NEMO
The popular European ocean model (3.4 version) is here associated to LIM (sea-ice) version 2. This coupled model particularity lies in an so called AGRIF[19] ocean zoom, centred in the North Atlantic region (ERNA[20] configuration). NEMO is defined on 2 different grids: the global (or parent), ORCA05 with 722x511x64 grid points and the regional (or child), 1/8 degree, 724x632x64 grid points. The same computing resources (MPI processes) are used to perform sequentially the computations on the parent and child grids[21]. Boundary conditions of the child grid are provided by the parent grid (through interpolations independent from OASIS). In the other direction, parent grid points belonging to the child area are updated by child calculations at each parent time step.

Before starting the User Support period, those models had been ported on “vargas” IBM supercomputer, 3,584 compute cores (112 thirty-two-cores 4.7 GHz Power 6 processors per node), Infiniband x4DDR Interconnect. Total peak performance of 67.3 Teraflop/s. This machine is located at IDRIS CNRS computing centre, France.

[19] Adaptive Grid Refinement In Fortran, http://www-ljk.imag.fr/MOISE/AGRIF/
[20] Eddy Resolving North Atlantic
[21] In our case, after one time step on parent grid (2160 sec), three time step are performed on the child one (3x720s)

Model interfaces

A preliminary step dealt with an upgrade of the existing model interfaces to allow ocean -atmosphere field exchanges not only with the global ocean but also with its regional zoom.

In NEMO, we mainly relied on an existing implementation provided by LOCEAN experts, developed for a WRF-NEMO high resolution coupling on a tropical belt area (45S-45N)[22]. Our contribution consisted in allowing atmosphere-ocean exchanges from and to the regional zoom in both ways, including for ice related coupling fields.

Implementation principle is relatively simple. As any other NEMO routine in an AGRIF configuration, each OASIS primitive can be called twice, first by the parent and then by the child). Coupling implementation only consisted, as explained in more details below, in choosing to call them twice or once and, in this case, from the parent or from the child.

At initialisation, our sequence of OASIS primitive calls is:

INIT_COMP	GET_LOCALCOMM	DEF_PARTITION	DEF_VAR	def_partition	def_var	enddef

In the time step loop:

PRISM_GET	prism_get	prism_get	prism_get	PRISM_PUT	prism_put	prism_put	prism_put

And at the end:

TERMINATE

Calls from parent grid are shaded in blue and upper-case, from child in red and lower-case. Parent and child share the same MPI process, which means that init_comp, enddef and terminate, that must be called only once per MPI process, must be invoked by one or the other. At the opposite, as parent and child do not have the same grid (even with different dimensions), they both must call def_partition. Consequently, parent and child coupling fields[23] must be also declared separately. As we have chosen to receive and send fields on both grids, prism_get and prism_put primitives are called at each time step of parent and child components[24].

For further extension to configurations including more than one AGRIF zoom, it will be necessary to limit enddef call to the last child only.

Two corrections of the NEMO release version, both related to sea-ice coupling, were necessary (and reported to the NEMO system team): fr1_i0 and fr2_i0 arrays initialisation must be done before (on lim_sbc_init_2) sbc_cpl_init call and dynamical allocation of wind stress on sea-ice must be done in any case, even if this coupling field is not explicitly

[22] PULSATION project (ANR-11-MONU-0010), http://www.agence-nationale-recherche.fr/en/anr-funded-project/?tx_lwmsuivibilan_pi2[CODE]=ANR-11-MONU-0010

[23] In our case, parent or child coupling fields quantities are the same but one could imagine to exchange a different number, or even different coupling fields, from parent and grid child. This is already possible, just changing coupling field configuration on the NEMO parent or child namelist parameter files

[24] According to namcouple defined coupling frequency, MPI coupling exchanges are triggered only on a subset of model time step (see OASIS3-MCT documentation, “Sending a coupling (or I/O) field”). In our case, both parent and child coupling fields are exchanged at the same dates.

provided by the atmosphere model.

Both OASIS3 and OASIS3-MCT interfaces have been updated (and validated) at the same time, considering that they only differ on a few CPP key controlled lines (key_oasis3 and key_oasis_mct).

Fewer corrections were mandatory on the atmosphere routines, all related to the duplication of outgoing coupling fields, bound to ocean child grid. As soon as an OASIS3-MCT coupling will be possible, according to Météo-France schedule, those modifications will not be necessary any more: our new coupler is supposed to be able to send one coupling field to two different targets, performing different interpolations to different grids, which is exactly what is needed here.

Interpolations

Grids description

OASIS auxiliary files gathering all grid descriptions must be built up first. In NEMO, the information about the parent and child grids was deduced from the “meshmask” output file previously produced by an ERNA stand alone simulation. ARPEGE related data were provided by CNRM from previous coupled configurations.

A first issue appeared when considering the limit between parent and child regions. Child grid consists in two different zones: the main inner 1/8 degree part and a buffer zone, which is usually filled with interpolated information coming from the parent grid.

As suggested by LOCEAN experts, we decided to include the field values coming from this buffer zone into the data sent by the child ocean to the atmosphere: smoothed SST and other ocean/ice coupled quantities coming from this transition zone between parent and child regions are then provided to the atmosphere. We will check that this technique leads to avoid artificial gradient source in this zone[25]. In the other direction, fluxes coming from the atmosphere will also be interpolated on this boundary zone.

Global interpolations

On a second step, interpolations weights for coupling the atmospheric grid with the parent ocean grid must be calculated. Due to the ARPEGE stretched grid characteristics (gaussian grid, with less grid points near poles, combined with a stretched coefficient), meshes have extremely different areas from one pole to the other. SCRIP 'CONSERV' is the only standard OASIS interpolation associating to the target points a different number of source neighbours according to target/source mesh areas ratio and, consequently, to provide relevant source information to any target grid point when this ratio is much bigger than 1.

Unfortunately, SCRIP algorithms are not able to perform 'CONSERV' weights calculations with ARPEGE stretched grid and a new kind of interpolation has to be developed to give a satisfactory solution to this problem. In the meantime, a simple 4 neighbours Gaussian

[25] This coupled model configuration is precisely set up to study the impact of physical gradients. That’s why it is particularly crucial to avoid artificial numerically-created gradients at the limit of the zoom.

interpolation was preferred, to give more accurate results near the zoomed area than at the antipode. Weights were automatically produced by OASIS at first simulation (taking a few minutes on our supercomputer).

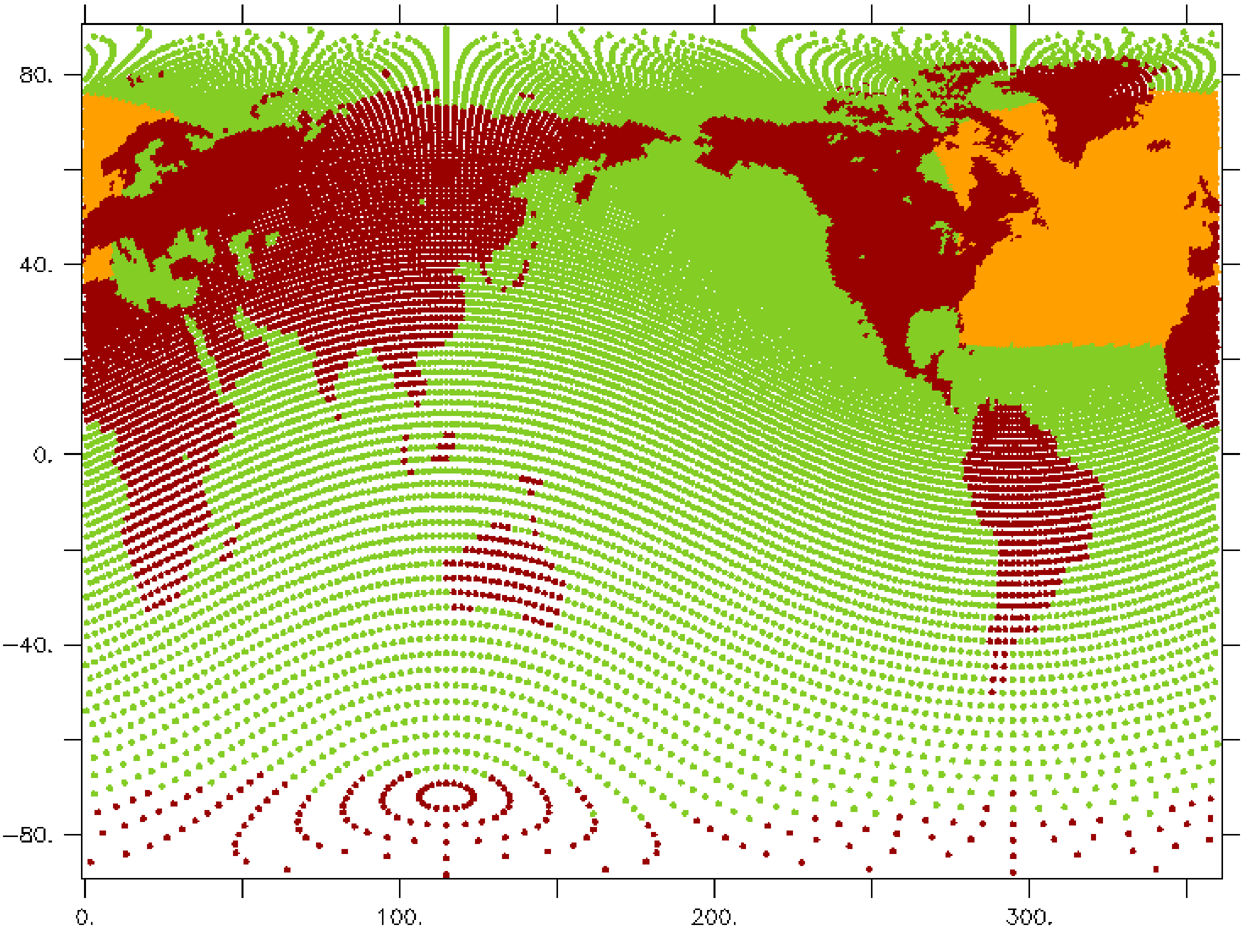


Fig 1: ARPEGE T127-stretched grid, with land mask (brown) and points interpolated from NEMO-AGRIF parent (green) and child grid (orange)

Regional interpolations

The same Gaussian interpolation was chosen to provide information from/to the ocean child grid. As mentioned during previous Dedicated User Supports, OASIS automatic weight calculation was not specially designed for regional coupling. In particular, its algorithm will try to entirely fill a global grid with the information coming from a regional grid. On the other direction (from global to regional), depending on interpolation choice and masking configuration, source grid points located at large distance of regional grid boundaries could be used in calculating coupling field values send to the target grid. It is then strongly recommended to take care on how OASIS calculates its interpolation weight automatically and to modify its behaviour when necessary.

The strategy described below is similar to the one already used with COSMO/CLM and COSMO/ECHAM coupling. OASIS is launched first with initial grid masks. Interpolated fields are then post-processed to better define a mask that reduce exchanges to a smaller region of the global grid.

From the atmosphere (global) to the child ocean (regional), all atmosphere points can be used by the automatic OASIS weight calculation. The closest 4 atmosphere neighbours are associated to each child ocean point. Considering that the target/source grid point area ratio in this region is not too far from 1, far remote atmosphere points are not used at the boundary of the regional grid.

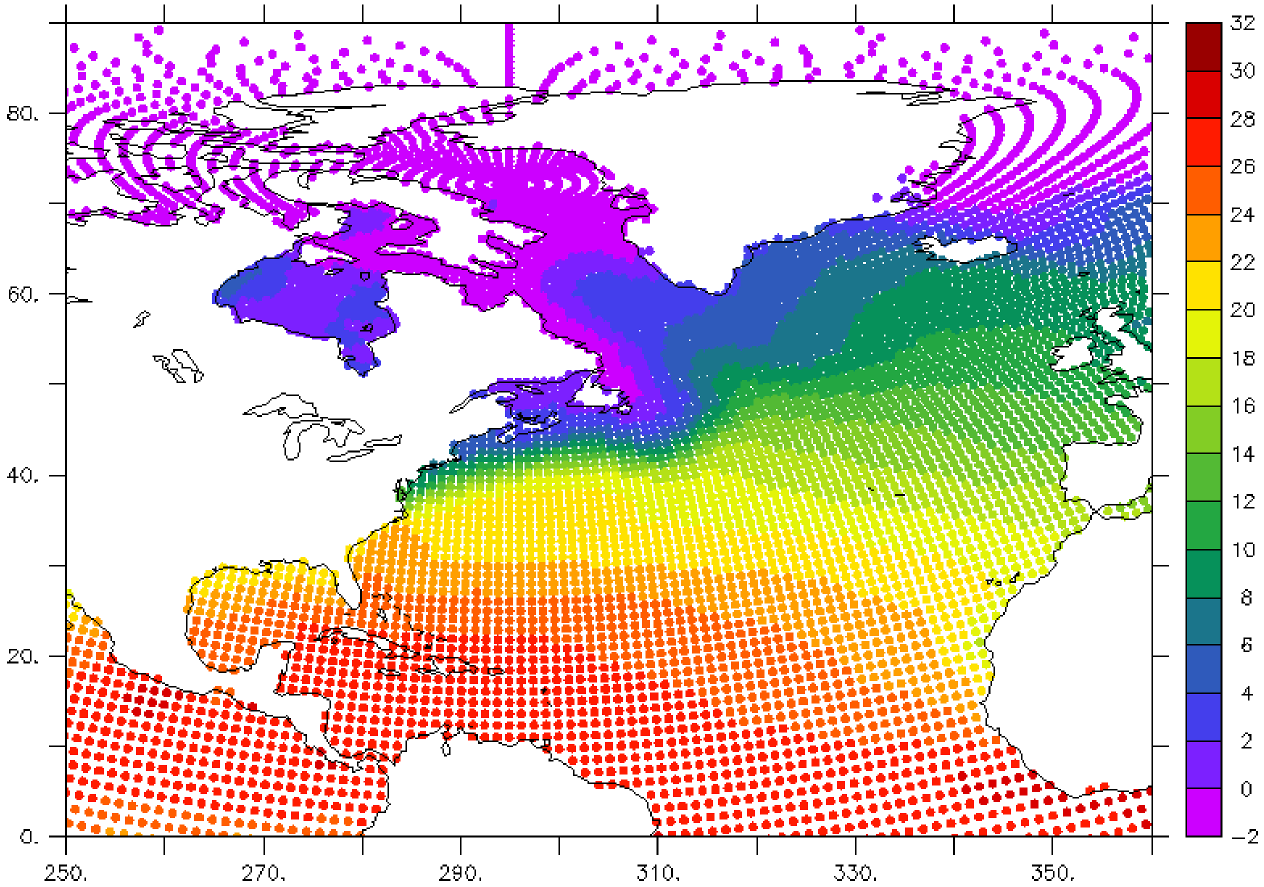


Fig 2: Example of combination on ARPEGE T127-stretched grid of 3h mean SST coupling fields interpolated from NEMO parent and child grids

But in the other direction, OASIS will automatically fill all global atmosphere point with information coming from the ocean zoom region, unless an appropriate mask is defined for the target grid. Different operations must be performed to avoid such problem:

1. A special mask is temporary defined on the child ocean grid: all grid points are unmasked, except those from first and last lines (and columns)
2. A bilinear interpolation is calculated identifying 4 source neighbours (2 in each directions) per target grid point[26]
3. A uniform coupling field is send to OASIS
4. The interpolated field[27] is equal to the source constant value where atmosphere grid points have 4 neighbours on child ocean grid. They are equal to zero on atmosphere masked points or on all grid points slightly outside the limit defined in (1).

We used this information to build two new atmosphere masks as shown on Fig 1. In

[26] OASIS3 bilinear interpolation was slightly modified before as the original algorithm would have used, for the atmosphere points falling outside the regional ocean domain, the non masked points among the 4 nearest neighbours in the regional domain and would have given them a value.

[27] Produced with a restart field with OASIS3 interpolator mode or on an output field with EXPOUT namcouple option

addition to the original mask (masked points in brown) covering the global domain:

- a mask defining only atmosphere sea points filled by interpolated quantities coming from the child ocean grid (unmasked points in yellow)
- a mask defining only atmosphere sea points filled by interpolated quantities coming from the parent ocean grid (unmasked points in green)

With a simple OASIS3 BLASNEW operation[28], it is now possible to rebuild a complete global coupling field combining information from the child and parent grids[29] (see example on Fig 2)

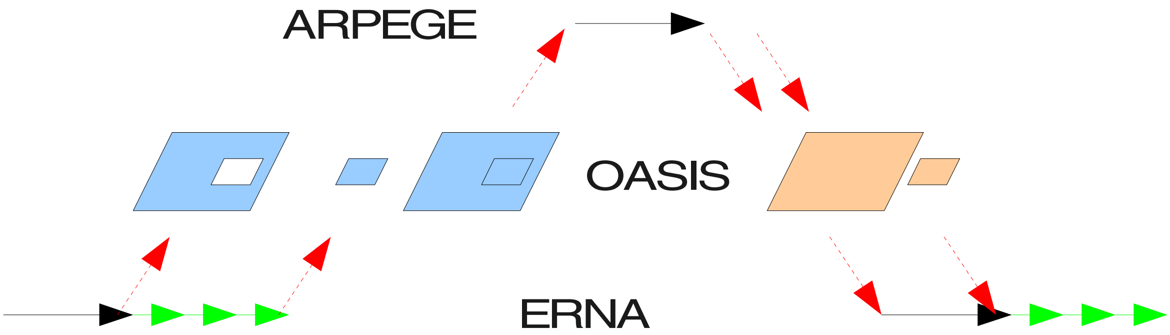


Fig 3: Coupling sequence. Child ocean time steps are represented in green, regular time step in black, OASIS exchanges in red

Namcouple parameter file must be defined accordingly to this new strategy:

- all original coupling fields must now be exchanged twice (in the two directions). Our naming convention follows the NEMO-AGRIF one: the first letter of the global field is replace by the number of the child ocean grid (here 1)
- each child coupling field coming from ocean is added to the corresponding field coming from the parent grid (with BLASNEW option)
- Coupling frequency must be the same on child and parent grid[30].

Parent/child grid re-sticking impact

To evaluate the quality of our interpolated field gathering parent and child grids information, we have compared it, on a 3 hours mean surface temperature (sea/ice mean), with the field coming from the parent grid only (including on the child covered area) and interpolated on the whole atmosphere grid.

Differences are shown on Fig 4. As wished, it clearly appears that no artificial gradient is created at AGRIF zoom domain boundary. One can also noticed that main differences are located on high gradient areas, like Gulf Stream or ice field boundary.

[28] This operation is not yet available on OASIS3-MCT

[29] No modification is then needed in the atmosphere interface compared to when only global field is received

[30] Be careful that LAG value must match model time step of the model from where coupling field is coming (child/parent ocean, or atmosphere, see sequence on Fig 3)

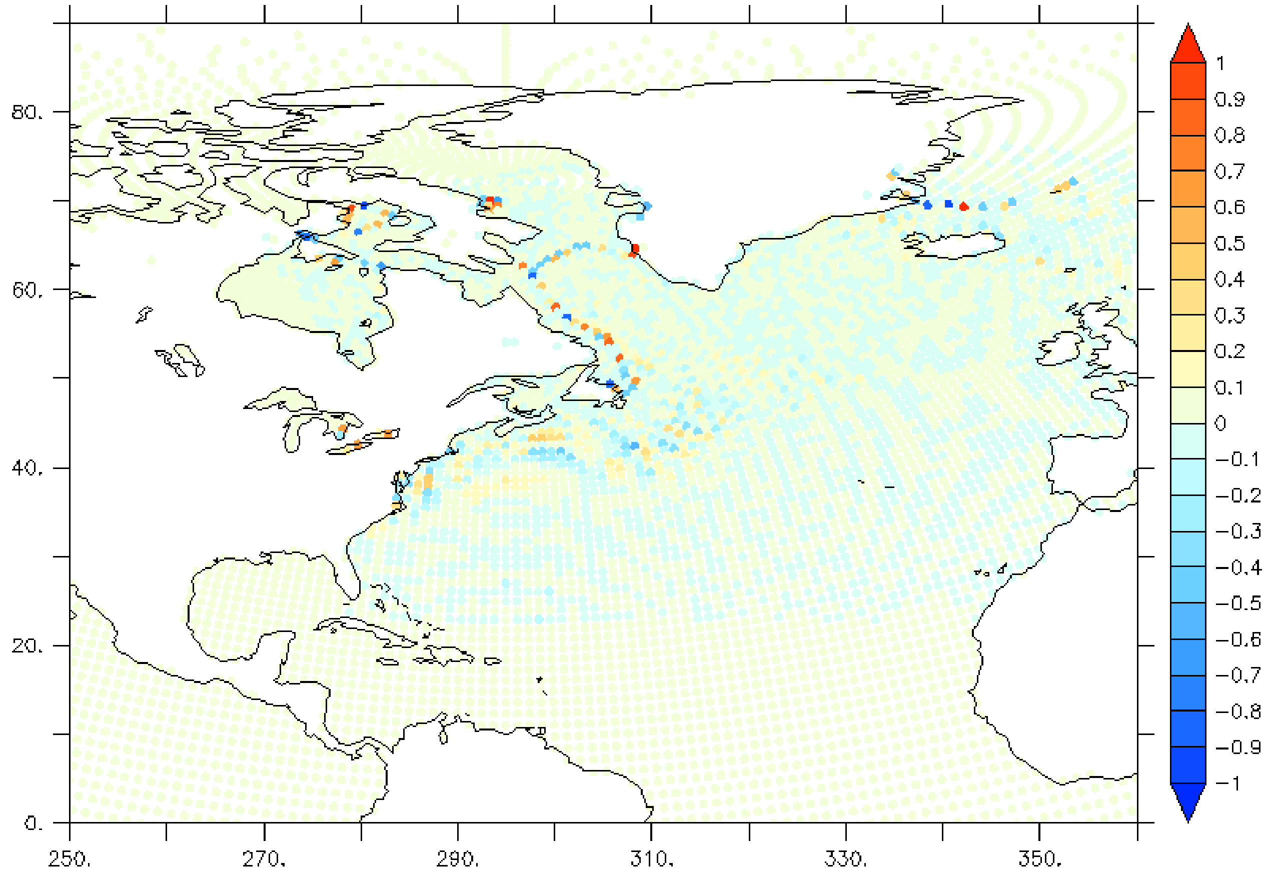


Fig 4: Difference on ARPEGE T127-stretched grid between 3h mean surface temperature field interpolated from parent grid and combination (same as fig 3) of SST fields interpolated from parent and child grid

Validation on model

Due to a lack of time, it was not possible to set up the targeted configuration including stretched gridded atmosphere model. A simple ARPEGE configuration on regular T127 grid (and corresponding OASIS auxiliary and interpolation files) was used to test our coupling.

A one month long run was performed. It still exhibits a pre-existing issue of unrealistic ice temperature biases. Consequently, it was unfortunately not possible to fully certify integrity of our interpolations and interface implementations.

Nevertheless, this regular T127 configuration includes the same grid point number than the stretched one. The load balancing analysis we produced (Fig 5) with the previously developed “lucia” tool[31] will then be equivalent on the targeted configuration[32].

If the main information it shows is robust (speed ratio between ARPEGE on 10 cores and ERNA on 127 is about 3), some information such as OASIS calculation time must be corrected. Our measurement algorithm did not anticipate that during BLASNEW operation (coupling field combination), OASIS has to wait information coming from the child ocean.

[31] For further information on Lucia load balancing analysis, see Dedicated User Support #9
[32] Except if atmosphere time step is different when grid is stretched

Consequently, the OASIS “interpolation time” (Oasis related red box) includes, in our case, the waiting time, at each coupling time step, between parent and child field receiving. Corrected from this bias, the OASIS3 interpolation and exchange times can be considered as negligible compared to the total simulation time.

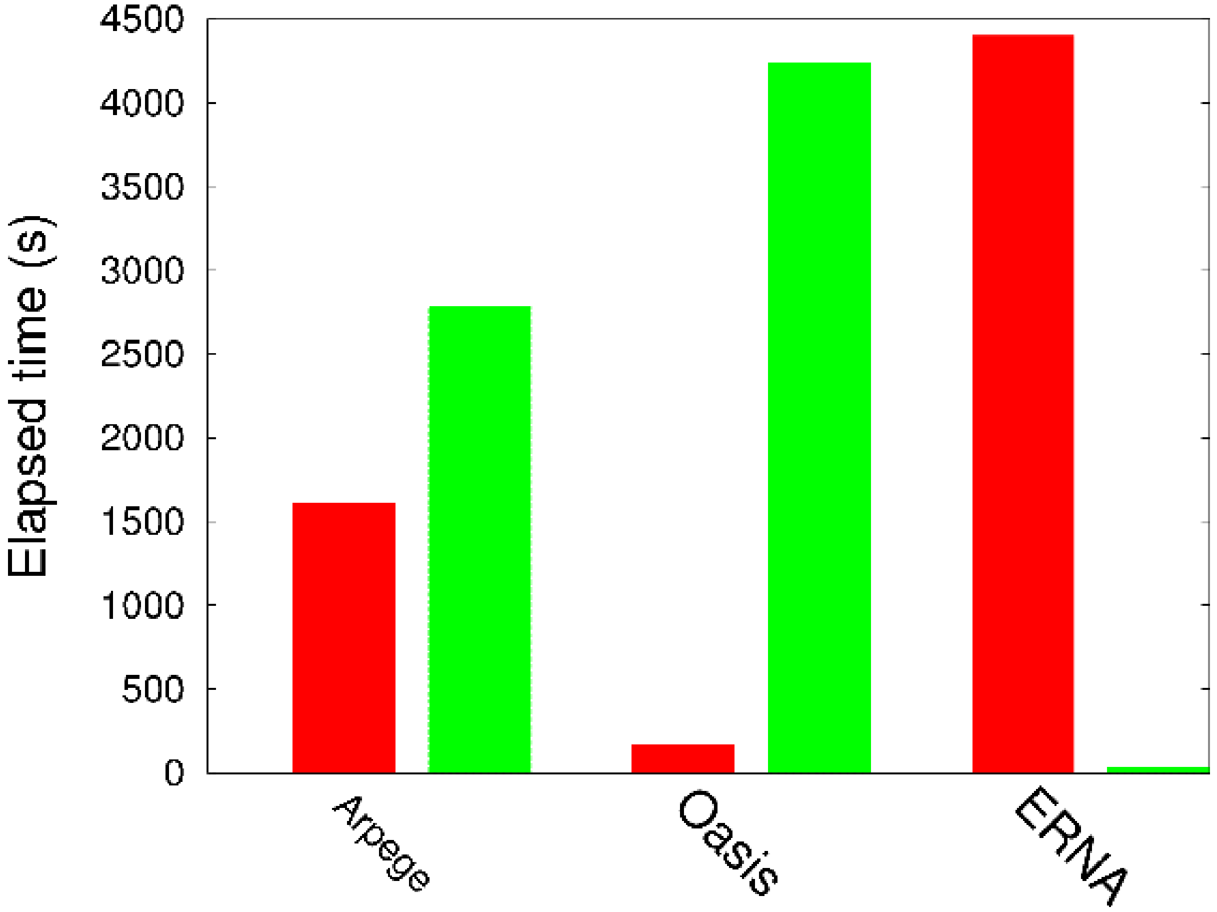


Fig 5: “Lucia” load balancing analysis performed on 26 exchanges (one per day)
ARPEGE T127 on 10 cores, ERNA on 127, OASIS3 on 1

Annex 1: COSMO/ECHAM two ways nesting coupling fields

Coupling field	**Sent by**
Geopotential height	ECHAM
Surface pressure	ECHAM & COSMO
Temperature (47 levels)	ECHAM & COSMO
Zonal wind (47 levels)	ECHAM & COSMO
Meridional wind (47 levels)	ECHAM & COSMO
Specific humidity (47 level)	ECHAM & COSMO
Specific liquid water content (47 levels)	ECHAM & COSMO
Specific solid water content (47 levels)	ECHAM & COSMO