EUROPEAN COMMISSION

Community research

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

IS-ENES2 DELIVERABLE (D -N°: 10.3)

Report on benchmark suite for evaluation of coupling strategies

CERFACS TECHNICAL REPORT TR-CMGC-17-87

Author(s): Sophie Valcke, Gabriel Jonville, Rupert Ford, Mike Hobson, Andrew Porter, Graham Riley

Reviewer(s): Bryan Lawrence, Giovanni Aloisio

Reporting period: ***01/04/2016 – 30/03/2017***

Release date for review: 06/03/2017

Final date of issue: 10/05/2017

Revision table			
Version	**Date**	**Name**	**Comments**
V0	2017-02-17	S. Valcke	
V1	2017-02-27	S. Valcke	Including comments from G. Jonville, M. Hobson, G. Riley, A. Porter
V2	2017-03-03	S. Valcke	Including comments from R. Ford
V3	2017-05-04	S. Valcke	Modifications to answer comments by reviewers B. Lawrence and G. Aloisio

Abstract

This document describes the work done to develop a first version of a community coupling technology benchmark. Today, stand-alone components running on 4 different grids and coupled test-cases based on components running on the regular latitude-longitude grid and using either the OASIS3-MCT, OpenPALM, ESMF, MCT or YAC coupling technologies are publicly available. As a proof of concept, these coupled test cases were run in different configurations on three different platforms: Bullx at CINES in France, Cray XC40 at the UK MetOffice, and the Broadwell partition of Marconi at CINECA in Italy. The coupled components exchange coupling fields defined on grids of different sizes decomposed in parallel partitions with different aspect ratios and different orientations. The timings obtained for the coupling initialisation and coupling exchanges for the different tests are detailed, primarily to demonstrate the versatility of this benchmarking environment. However, these first results are not yet suitable to draw any firm conclusions on the relative performance of the coupling technologies used.

Project co-funded by the European Commission's Seventh Framework Programme (FP7; 2007-2013) under the grant agreement n°312979		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants including the Commission Services	
RE	Restricted to a group specified by the partners of the **IS-ENES2** project	
CO	Confidential, only for partners of the IS-**ENES2** project	

Table of contents

Executive Summary

This deliverable D10.3 “Report on benchmark suite for evaluation of coupling strategies” contains the final results of WP10 task 3 “Evaluation of coupling strategies”. The objective of this task was to define and implement a suite of coupled benchmarks based on simplified model components that capture the essence of the coupling challenges in climate models without the complexities of the science.

Today, stand-alone components running on 4 different grids and coupled test-cases based on components running on the regular latitude-longitude grid and using either the OASIS3-MCT, OpenPALM, ESMF, MCT or YAC coupling technologies are publicly available. These stand-alone components and coupled test cases form the first version of the IS-ENES2 coupling technology benchmark.

As a proof of concept, these coupled test cases were run in different configurations on three different platforms: Bullx at CINES in France, Cray XC40 at the UK MetOffice, and the Broadwell partition of Marconi at CINECA in Italy. The coupled components exchange coupling fields defined on grids of different sizes (Low Resolution (LR) - 100x100, High Resolution (HR) - 1000x1000 and Very High Resolution (VHR) - 3000x3000 grid points) decomposed in parallel partitions with different aspect ratios and different orientations.

The timings obtained for the coupling initialisation and coupling exchanges for the different tests are detailed, primarily to demonstrate the versatility of this benchmarking environment. However, these first results should not be used to draw any conclusions on the relative performance of the coupling technologies used. More work is required to evaluate the robustness of these results before firm conclusions can be inferred.

1. Introduction

1.1 Objectives

This deliverable D10.3 "Report on benchmark suite for evaluation of coupling strategies" contains the final results of WP10 task 3 « Evaluation of coupling strategies ». The objective of this task was to define and implement a suite of coupled benchmarks based on simplified model components that capture the essence of the coupling exchanges in climate models without the complexities of the science.
After defining the main features of the benchmark suite in milestone M10.1 « Definition of the benchmark suite for evaluation of coupling strategies » and presenting the planned test cases in milestone M10.4 « Implementation of the benchmark suite for evaluation of coupling strategies", we now present the stand-alone components and the coupled test cases forming the first version of the IS-ENES2 coupling technology benchmark suite.
Five different coupling technologies - OASIS3-MCT, OpenPALM, ESMF, MCT and YAC - were used to implement coupling exchanges between components running on regular latitude-longitude grids with 100x100 (LR), 1000x1000 (HR) and 3000x3000 (VHR) grid points. In this report we detail the results obtained while running these coupled test cases on three different platforms: Bullx at CINES in France, Cray XC40 at the UK MetOffice and the Broadwell partition of Marconi at CINECA in Italy. The times for the coupling initialisation and for the coupling exchanges were measured for these test cases running with up to O(10,000) cores.
All results are detailed here, mainly to show the versatility of this benchmarking environment. However, at this point, these first results should not be used, as is, to draw any firm conclusions on the relative performance of the coupling technologies used. More work is required to evaluate the significance and the robustness of these first results before any conclusions can be inferred.

1.2 Context

In milestone M10.1 published in October 2014, we first described the possible functions of coupling technologies and the characteristics of Earth System Models (ESMs) supported by these coupling technologies in a series of mindmaps: "CouplingTechnology", "Components", "Metadata", "Composition" and "Deployment". This community work on the characterization of ESM coupling started during the « 2nd Workshop on Coupling Technologies »[1] held in Boulder in February 2013 and was finalized by IS-ENES2 partners interacting with colleagues of the US project Earth System Bridge. The resulting mindmaps are available at https://earthsystemcog.org/projects/es-fdl/mindmaps.

The next step was to prioritize key coupling characteristics to benchmark; these were identified as:

- the component model grids,
- the number of MPI ranks used to run the component models,
- the number of fields exchanged between the components,

1 https://www.earthsystemcog.org/login/?next=/projects/cw2013/

- the frequency of exchange.

It was also mentioned that code intrusion, development time and issues met during development, which are all aspects of 'ease of use', should also be evaluated and reported on, although they are difficult to quantify.

It was then established that the benchmark suite would first consist of a number of pre-coded stand-alone components running on different grids; these components would then be assembled thanks to the different coupling technologies in coupled test-cases to evaluate the priority characteristics identified.

In the following milestone M10.4 produced in March 2016, the full specifications of the stand-alone components were presented and a possible hierarchy of benchmark test-cases to evaluate the main coupling characteristics were detailed. The specifications of these test-cases covered:

- the grid type of the coupled components
- the grid size of the coupled components
- the number of processes used to run the coupled components
- the layout of the components on the available computing cores
- the number of coupling fields exchanged between the components
- how the coupling fields are matched between the components
- the schedule of components, i.e. their concurrent and sequential execution

A priority was then established from the above characteristics and, as a first step, it was decided to implement cases testing the impact of: 1) the number of cores per component, 2) the grid sizes, and 3) having different numbers of cores for different components.

1.3 Content of the deliverable

In this document, we first describe in Section 2 the stand-alone components and coupled test-cases implemented with five different coupling technologies, forming the current version of the IS-ENES2 benchmark suite :

- OASIS3-MCT : version OASIS3-MCT_3.0 available under SVN at https://oasis3mct.cerfacs.fr/svn/branches/OASIS3-MCT_3.0_branch/oasis3-mct/ (see also https://verc.enes.org/oasis/) ;
- OpenPALM: version v4.2.1 (see http://www.cerfacs.fr/globc/PALM_WEB/) ;
- ESMF : version 7.0.1 (see https://www.earthsystemcog.org/projects/esmf/) ;
- MCT :,version v2.9.0 (see http://www.mcs.anl.gov/research/projects/mct/) ;
- YAC: v1.2.0_p10 (see https://doc.redmine.dkrz.de/YAC/html/)

The stand alone components and coupled test-cases are publicly available as a tar file on the ENES portal at https://verc.enes.org/computing/performance/benchmarks/coupler-benchmarks.
In Section 3, the results of the current coupled test cases are reported. The time for the coupling initialisation and for the coupling exchanges for these coupled configurations running with up to O(10,000) cores on three different platforms: Bullx at CINES in France, Cray XC40 at the UK MetOffice and the Broadwell partition of Marconi at CINECA in Italy, are detailed.

2. Stand-alone components and coupled test cases implemented.

2.1 Available stand-alone components

The stand-alone components consist of simple, individual model codes, containing no physics or dynamics, but representative of real models in term of coupling characteristics. The stand-alone components implemented follow the specifications established in milestone M10.4:

- They are coded in Fortran90 and implemented as a subroutine, or hierarchy of subroutines, wrapped into a driver that enables their stand-alone execution.
- They are parallel MPI codes internally.
- They do not refer to any coupling technology.
- They define a number of 2D Real Fortran arrays, which represent coupling fields that can be received (in) and sent (out) by the model when it is deployed in a coupled context; these coupling fields appear as:
    - IN and OUT arguments of the subroutine declared in the driver (rda_field1 below)
    - arrays in shared modules (rma_field1 below)
    - local data declared at a particular, possibly deep, level in the subroutine call tree (rla_field1 and rla_field2 below)
- They run for 100 time steps.
- In the stand-alone mode, the coupling fields are initialised from a utility library routine or from a NetCDF file.

The only M10.4 specification not satisfied is that the current stand-alone components are not parameterised in terms of the number of coupling exchanges, but this should be quite straightforward to include in a next step.

Four stand-alone components were implemented and are available in the benchmark tar file at https://verc.enes.org/computing/performance/benchmarks/coupler-benchmarks on the ENES portal. The coupling fields of these components are defined on the following specific grids used in real climate model components with the following parallel decomposition (the names used hereafter are the names of the respective sub-directories):

- slatlon: Self-generated regular latitude-longitude grid allowing arbitrary resolutions to be used. This flexibility in resolution is useful for very high-resolution tests on large systems. This grid can be split in MxN rectangular partitions of different aspect ratios. In the tests described below, grids with
    - 100x100 (LR),
    - 1000x1000 (HR) and
    - 3000x3000 (VHR) grid points were used;

  The 1000x1000 grid corresponds to what is currently considered to be a high-resolution (HR) model (e.g. the NEMO ORCA025 configuration, 1442x1021 points) and the 3000x3000 grid corresponds to what is currently considered to be a very high-resolution (VHR) model (e.g. ORCA12 configuration, 4322x3059 points).,The 100x100 grid represents a relatively low resolution (LR) grid (e.g. ORCA2 configuration with 182x149 points).
- stretchlatlon: Irregular, stretched and rotated latitude-longitude mesh, following the ORCA configuration of the NEMO ocean model. As with the slatlon grid, this grid can be split in MxN rectangular partitions of different aspect ratios.
- icosa: Quasi-uniform icosahedral mesh, following the atmospheric DYNAMICO model. This grid is naturally partitioned into diamonds and sub-diamonds

- cubesphere: quasi-uniform cubed sphere mesh. The partitioning is achieved by splitting the "cube" into its six square panels; each panel is then sub-partitioned, into MxN rectangular partitions. This results in a total of 6xMxN partitions.

A stand-alone component on the Gaussian Reduced mesh, identified in M10.4, was finally not developed in this first version of the IS-ENES2 coupling benchmark suite, because of a lack of time.

2.2 Available coupled test-cases and timings

Different coupled test-cases were assembled based on the components using the 4 different grids described above but only the coupled test-cases assembling components using self-generated regular latitude-longitude grids are currently fully validated and were used for the timings reported in Section 3.
The specifications of the coupler test cases are as follows.

- The test cases implement the coupling with the 5 tested technologies listed in section 1.3 (see also respective subdirectories in the available tar file in isenes2wp10/src/coupled/lonlat-lonlat).
- In the components, the receive and send actions are implemented at the beginning and at the end of each time step, respectively.
- The following composition is implemented between the different coupling fields of the two components (see section 2.1 for the type of the different fields). The subroutine local field rla_field2 is sent from the first component to the second one that receives it in the shared module array rma_field1; the second component then sends back a global field rda_field1 that is received by the first component in a local field array rla_field1.
- A sequential schedule of the components is chosen to implement "ping-pong" exchanges. The first component uses some priming mechanism (here a simple initialization by an analytical function based on the spatial position of each grid point) to define its input coupling field rla_field1 at the beginning of its first time step, calculates its output coupling field rla_field2 (with the simple relation rla_field2 = rla_field1 + 1) and sends it to the second component. The second component receives it at the beginning of its first time step as rma_field1, calculates its output coupling field rda_field1 (with the relation rda_field1 = rma_field1 + 1) and sends it back to the first component that receives it at the beginning of the second timestep as rla_field1, and so on.
- The benchmark timings are implemented for:
  - the full initialisation including the coupling initialisation;
  - the first time step;
  - the average over 97 additional time steps (to give reproducible results);
  - the last time step

Figure 1 illustrates the ping-pong exchanges between the two components, the relations between the different fields, and how the different timings are implemented. It can be seen that the timing of the ping-pong exchanges includes the operation "rda_field1 = rma_field1 + 1" and "rla_field2 = rla_field1 + 1" . Additional tests performed without these operations demonstrated that they have a negligible effect on the ping-pong timings and it was decided to include them so to keep a relation between the different of coupling fields (so that a final check on the last coupling field received validates all the chain, see below).

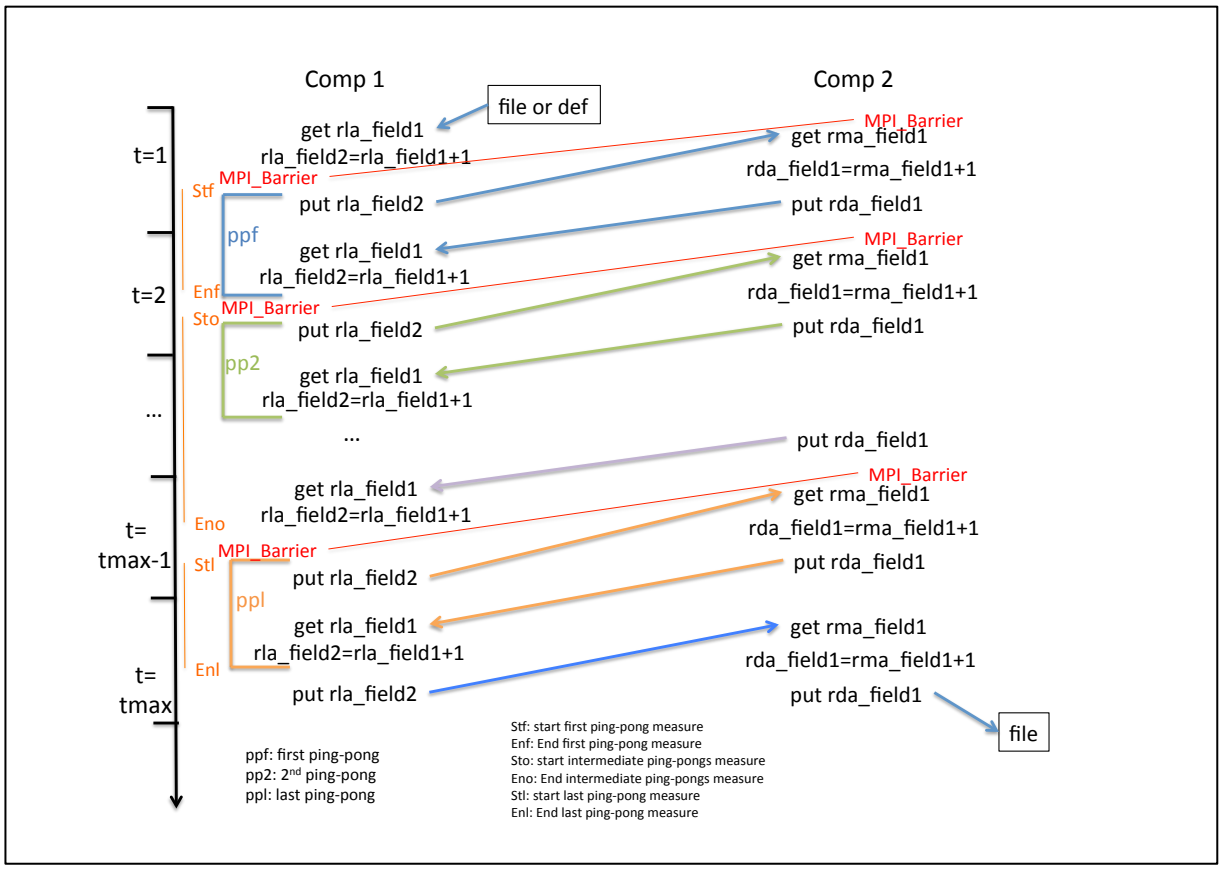

Figure 1 – Ping-pong exchanges between two coupled components. The first component (Comp 1 on the left) defines its input coupling field rla_field1 at the beginning of its first time step, calculates its output coupling field rla_field2 (with rla_field2 = rla_field1 + 1) and sends it to the second component. The second component receives it at the beginning of its first time step as rma_field1, calculates its output coupling field rda_field1 (with the rda_field1 = rma_field1 + 1) and sends it back to the first component that receives it at the beginning of the second timestep as rla_field1. The placement of the timing measures and related MPI_Barrier is also shown: Enf-Stf, Eno-Sto, and Enl-Stl respectively for the first, 97 intermediate, and last ping-pong exchange.

To have significant and robust results, it was decided to return the maximum value over all the component processes for each timing. It was also decided to perform each run 3 times and to analyse the spread to make sure it is relatively small (i.e. to ensure that the results are not affected by external perturbations such as the work load of the platform, punctual MPI or network problem, etc.). In the results below, the 3 timings and their mean are shown.

To validate the correct execution of the ping-pong exchanges, an error field is also calculated in the second component. This error is calculated as the difference between the last coupling field received and the value of the analytical function used to calculate the first input coupling field in the first component incremented by the number of time steps. If the coupling exchanges unfold as planned, this error will be very small compared to the coupling field itself, being linked only to the remapping accuracy (when a remapping is

needed, which is only for the LR-HR case, see below) or to the round-off errors of the coupling fields when the grids are the same.

3. Coupling benchmark results

The coupled test cases using the 5 coupling technologies based on components running on regular latitude-longitude grids were run for different grid sizes, on 3 different platforms for different number of cores. This section presents the results obtained. Preliminary remarks are first presented in section 3.1. The results per platform are then respectively detailed in sections 3.2, 3.3 and 3.4. Section 3.5 finally compares the results for all 3 platforms for each coupling technology.

3.1 Preliminary remarks

3.1.1 Additional specifications

To have meaningful and comparable results between the 3 platforms additional specifications were followed:

- Runs are performed on a number of cores per component varying between O(1) and O(10^4), up to the available number of cores on the platform and as long as the number of grid points per process is at least 50.
- In that range of cores, each test will be run on a number of cores filling entirely a specific number of nodes; e.g. if running on a 24-core node machine, the tests will be run on multiples of 24 cores.
- The processes of each component should be distributed in a "packed" manner, i.e. first filling completely a certain number of nodes with the first component processes and than fill a certain number of additional nodes with the second component processes.

3.1.2 Coupling technology effectively tested on the different platforms

Table 1 shows which coupling technology test case was effectively run on which platform, i.e. Occigen, the Bullx at CINES in France, on the Cray XC40 at the Met Office in the UK, and on the Broadwell partition of Marconi at CINECA in Italy.

	OASIS3-MCT	OpenPALM	ESMF	MCT	YAC
Occigen Bullx	X	X	X		X
Cray XC40	X	X	X	X	
Marconi Broadwell	X	X	X	X	X

Table 1 – For each computing platform considered and each coupling technology, an X indicates that the corresponding test case was effectively run

3.1.3 Important remarks on OASIS3-MCT and MCT

An important point to note is that the calculation of the remapping weight-and-address file is never included in the initialisation time for the OASIS3-MCT coupler. Indeed, this file was pre-calculated offline and simply initially read in, as it would have taken hours for the OASIS3-MCT sequential algorithm to generate it. The only exception is for the LR-HR grids

(section 3.2.3) : with these relatively lower resolution grids, the time needed to calculate the remapping weights-and-address file in OASIS3-MCT was reasonable and is included in the initialisation timings.

Similarly, for MCT, the remapping weight-and-address file is also always pre-calculated offline and initially read in as MCT simply does not provide this functionality. Therefore, one has to remember that initialisation time never includes either the calculation of the remapping weight-and-address file for MCT.

For all other coupling technologies, the initialisation time always includes the time to generate the remapping weight-and-address file.

Another specificity of the current MCT benchmark implementation (but not of MCT itself) is that it would not support grids of different sizes; therefore the LR-HR test case could not be run with the MCT benchmark. Work is going on to generalize the MCT implementation so to support grids of different sizes.

3.2 Results on Occigen Bullx at CINES in France

Occigen is the Bullx platform operated at CINES in Montpellier, France. Occigen has Xeon E5-2690v3 12C 2.6GHz processors and an Infiniband FDR interconnect. With 50544 cores, Occigen has a theoretical peak performance of 2,102.63 TFlop/s and a Linpack performance of 1,628.77 TFlop/s. The coupling benchmarks were compiled with Intel compiler 15.0.3.187, and run using bullxmpi 1.2.9.2 and NetCDF 4.3.3-rc2_fortran-4.4.1 libraries. An allocation of 480 000 core-hours was granted to Cerfacs to perform the tests.

On Occigen,a lack of time and resources prevented us from running the test case with MCT The specifications of these tests and a first analysis of the results are presented next.

3.2.1 HR-HR grids with same decomposition

In this first series of tests, the 1000x1000 (HR) grid resolution and the same rectangular decomposition were used for both components. Runs were performed on a number of cores/component varying between 1 and 6912. Table 2 presents the number of cores/component and the decomposition and resulting partition size for the grids (which are the same for both grids). The partition size given is approximate; for example, for a 3x2 decomposition of the 1000x1000 grid, the partition size is written as 333x500 whereas it really results in 4 partitions of 333x500 grid points and 2 partitions of 334x500 grid points.

Number of cores/component	1	6	24	96	216	432	864	1728	2304	2880	3456	6912
Grid 1 & grid 2 decomposition	1x1	3x2	6x4	12x8	18x12	24x18	36x24	48x36	48x48	60x48	72x48	96x72
Size of grid 1 & grid 2 partitions	1000 x 1000	333 x 500	167 x 250	83 x 125	56 x 83	42 x 56	28 x 42	21 x 28	21 x 21	17 x 21	14 x 21	10 x 14

Table 2 - Number of cores/component, decomposition and resulting partition size for each grid for the test cases on Occigen with HR (1000x1000) grids with same rectangular decomposition on each side.

Figure 2a shows the time for the coupling initialisation and Figure 2b the average ping-pong time (over 97 ping-pongs).

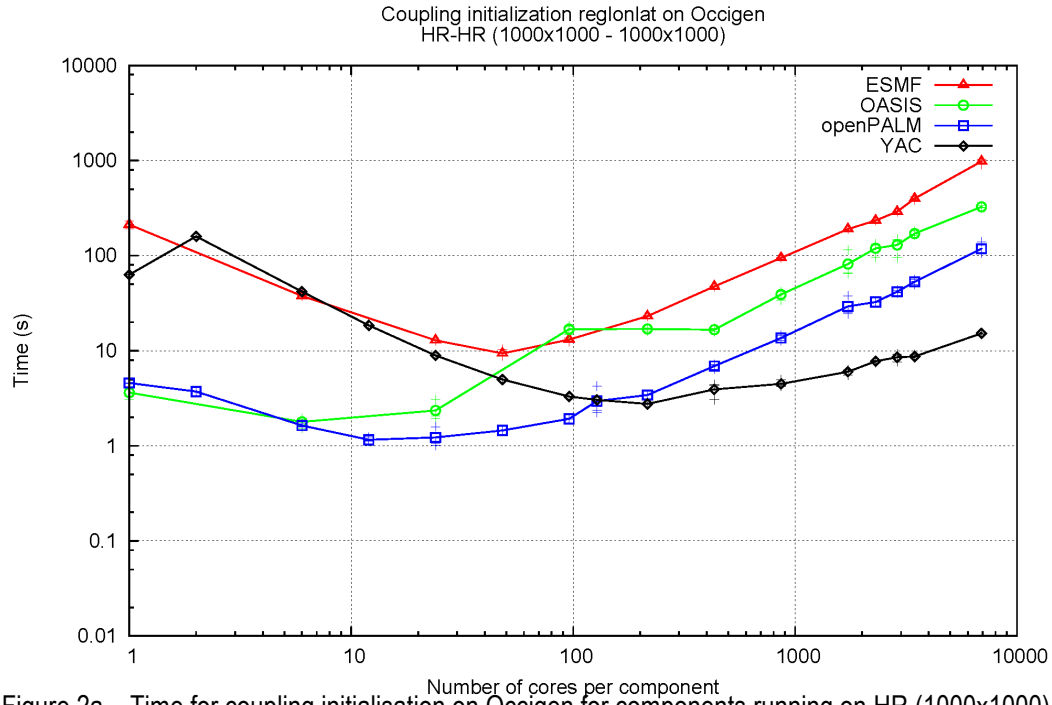

Figure 2a – Time for coupling initialisation on Occigen for components running on HR (1000x1000) grids with same decomposition on both sides.

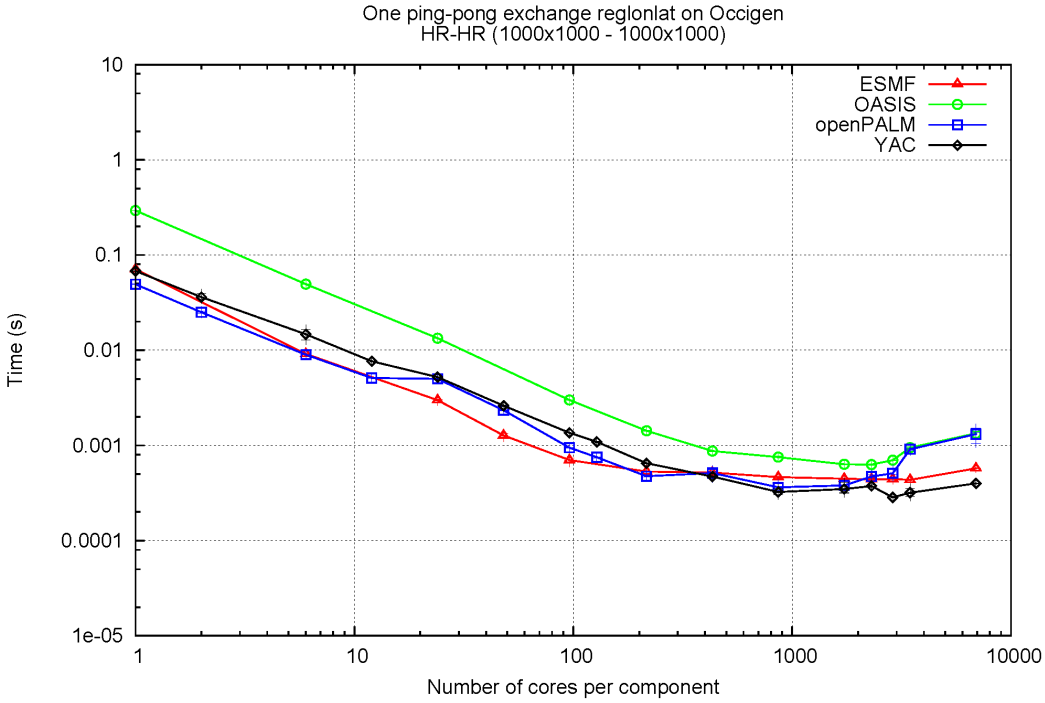

Figure 2b – Average time for one ping-pong exchange on Occigen for components running on HR (1000x1000) grids with same decomposition on both sides

A first analysis shows that the time required for the initialisation generally decreases for up to 10-100 cores/component and then increases with increasing number of cores/component. In all cases, the initialisation time remains below 1000 seconds, which can be considered reasonable compared with the time that a real model at such a resolution would take for a full job. Regarding the time for the ping-pong exchanges, all coupling technologies scale well for up to ~2000 cores but then the scalability curves flatten. One can also note that OASIS3-MCT is systematically about 5 times slower than the other couplers; the reason for this should be investigated and OASIS3-MCT communication schemes should be optimised.

3.2.2 VHR-VHR grids with same decomposition

Results of tests with VHR (3000x3000) grids are presented on Figure 3a for the coupling initialisation time and Figure 3b for the ping-pong average time (over 97 ping-pongs).

The number of cores used and the decomposition are the same as for the 1000x1000 grids. Table 3 shows the number of cores/component, and the decomposition and resulting partition size for the grids. At the highest number of cores (96x72=6912), each partition of the grid still has 31x42 grid points.

Number of cores/component	1	6	24	96	216	432	864	1728	2304	2880	3456	6912
Grid 1 & grid 2 decomposition	1x1	3x2	6x4	12x8	18x12	24x18	36x24	48x36	48x48	60x48	72x48	96x72
Size of grid 1 & grid2 partitions	3000 x 3000	1000 x 1500	500 x 750	250 x 375	167 x 250	125 x 167	83 x 125	63 x 83	63 x 63	50 x 63	42 x 63	31 x 42

Table 3 - Number of cores/component, decomposition and resulting partition size for each grid for the test cases on Occigen with VHR (3000x3000) grids with same rectangular decomposition on each side.

We were not able to run the test cases on 2 and 6 cores/components for ESMF and on 2 cores/components for OASIS3-MCT. These runs would either abort or deadlock without producing any results. The reasons remain unclear at this point and more time would be needed to understand these specific problems.

A first analysis shows that the respective behaviour of the different coupling technologies is similar than for HR-HR grids (Figure 2) but that the ping-pong timings are about 5-10 bigger; this is expected as the communication load of the VHR-VHR case is 9 times bigger than for the HR-HR case. For a number of cores greater than 2000, ping-pong times seem to stabilise around 0.001 second for both HR-HR and VHR-VHR grids; at this point, the exchanges involve many small messages and are presumably becoming latency bounded.

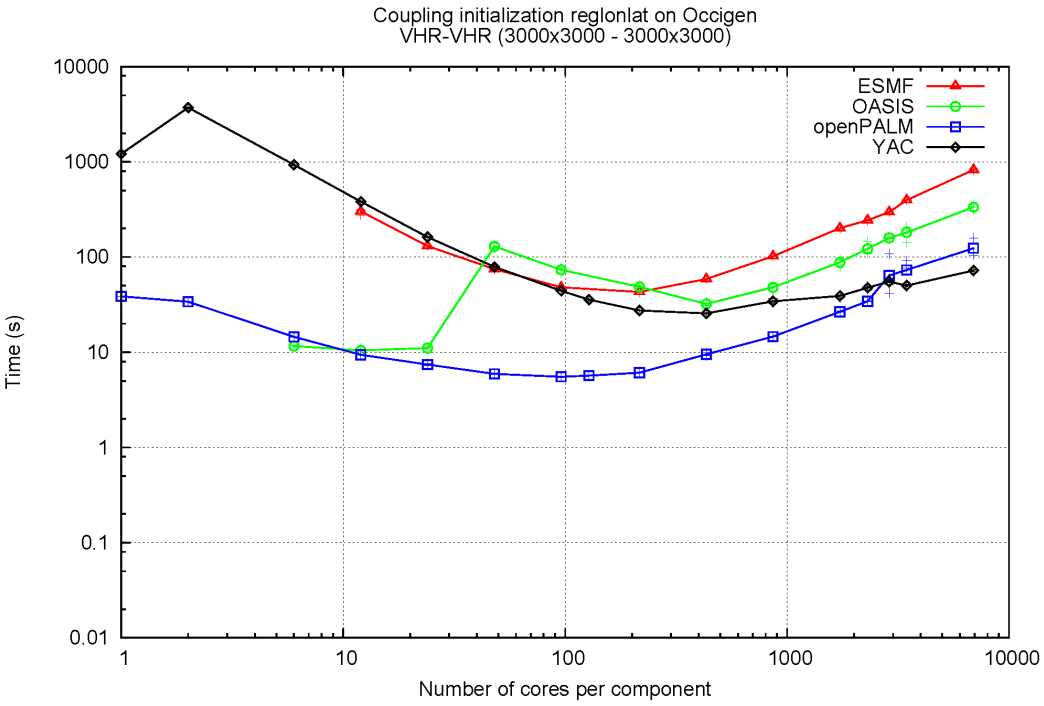

Figure 3a – Time for coupling initialisation on Occigen for components running on VHR (3000x3000) grids with same decomposition on both sides.

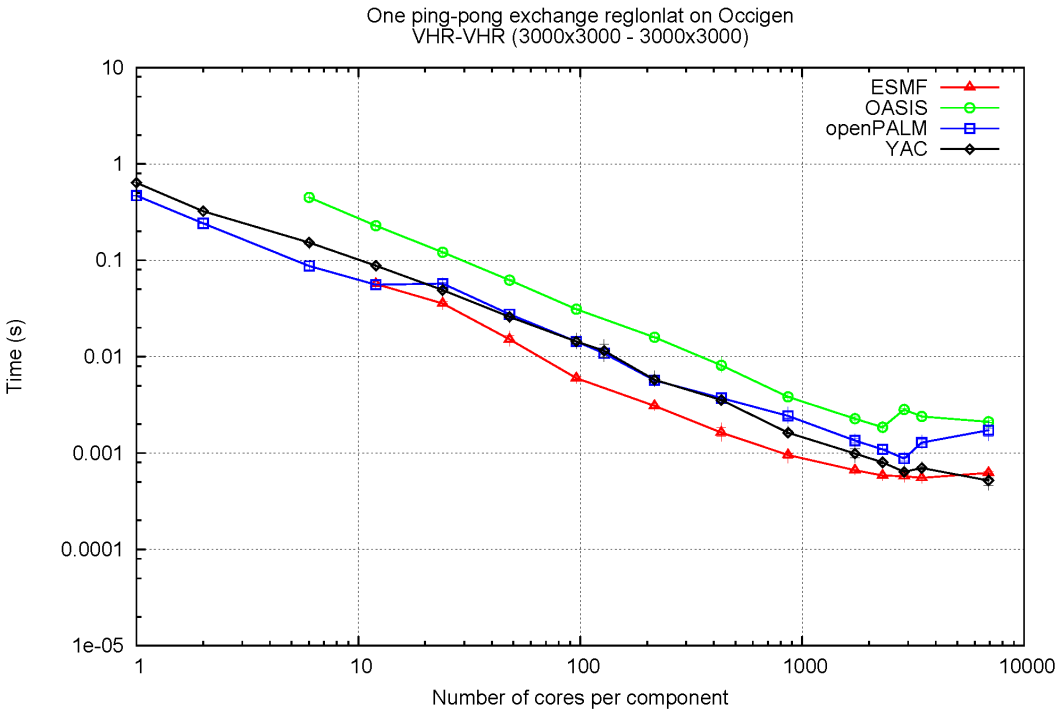

Figure 3b – Average time for one ping-pong exchange on Occigen for components running on VHR (3000x3000) grids with same decomposition on both sides

3.2.3 LR-HR grids with analogous decomposition

The test case was then repeated for a combination of LR (100x100) and HR (1000x1000) grids. Results are presented in Figure 4a and 4b. For this case, remapping was required and a one-nearest neighbour interpolation was applied.

The number of cores per component was limited to 2304 by the LR grid to maintain at least 2 grid points in each direction in the partition. Note that this violates the rule proposed earlier that the number of grid points per process should be at least 50 (see section 3.1.1). Table 4 shows the number of cores/component, and the decomposition and resulting partition size for each grid.

Number of cores/component	1	6	24	96	216	432	864	1728	2304
Grid 1 decomposition	1x1	3x2	6x4	12x8	18x12	24x18	36x24	48x36	48x48
Grid 2 decomposition	1x1	3x2	6x4	12x8	18x12	24x18	36x24	48x36	48x48
Size if grid 1 partitions	1000 x 1000	333 x 500	167 x 250	83 x 125	56 x 83	42 x 56	28 x 42	21 x 28	21 x 21
Size if grid 2 partitions	100 x100	33 x 50	17 x 25	8 x 13	6 x 8	4 x 6	3 x 4	2 x 3	2 x 2

Table 4 - Number of cores/component, decomposition and resulting partition size for each grid for the test cases on Occigen with LR (100x100) - HR (1000x1000) grids with rectangular decomposition on each side.

The results are very similar to the HR-HR results except that the time is always smaller, which is expected as the communication load is decreased.

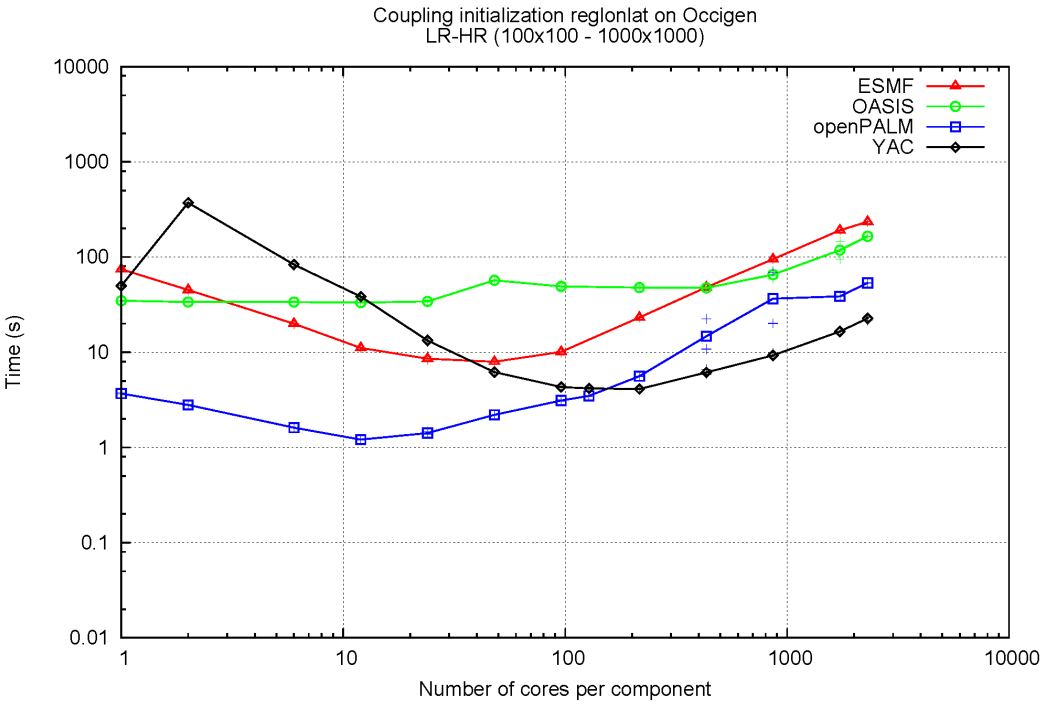

Figure 4a – Time for coupling initialisation on Occigen for components running respectively on LR (100x100) and HR (1000x1000) grids with same decomposition on both sides

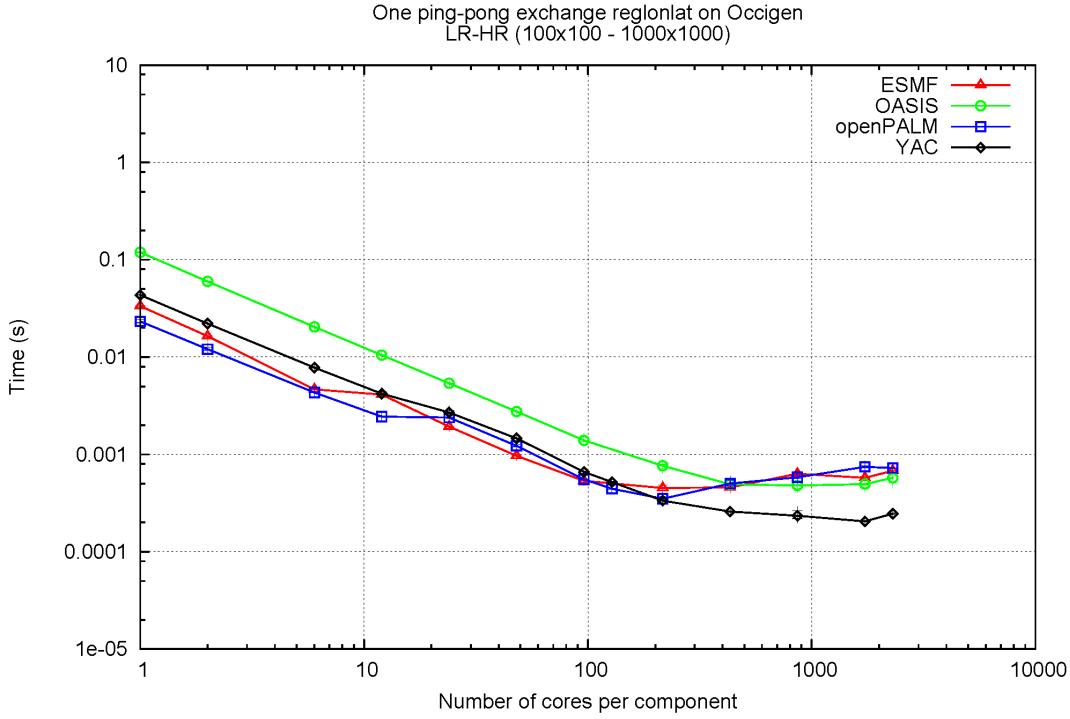

Figure 4b – Average time for one ping-pong exchange on Occigen for components running respectively on LR (100x100) and HR (1000x1000) grids with same decomposition on both sides

3.2.4 VHR-VHR grids with opposite decompositions

The tests were repeated for the VHR-VHR grids but imposing "opposite" decompositions for the two components. The number of cores was varied between 1 and 6912 as in 3.2.2 but the partitions were defined with an aspect ratio as big as possible and as "opposite" as possible for the two grids. Table 5 gives the total number of cores/component, the decomposition, the resulting size of the partitions and the aspect ratio for the grid partitions.

Total number of cores/component	2	6	24	96	216	432	864	1728	2304	2880	3456	6912
Grid 1 decomposition	1x2	1x6	1x24	1x96	1x216	1x432	1x864	2x864	3x768	3x960	4x864	6x1152
Grid 2 decomposition	2x1	6x1	24x1	96x1	216x1	432x1	864x1	864x2	768x3	960x3	864x4	1152x6
Size if grid 1 partitions	3000 x 1500	3000 x 500	3000 x 125	3000 x 31	3000 x 14	3000 x 7	3000 x 3	1500 x 3	1000 x 4	1000 x 3	750 x 3	500 x 2
Size if grid 2 partitions	1500 x 3000	500 x 3000	125 x 3000	31 x 3000	14 x 3000	7 x 3000	3 x 3000	3 x 1500	4 x 1000	3 x 1000	3 x 750	2 x 500
Grid 1 & 2 partition aspect ratio	2	6	24	97	214	428	1000	500	250	333	250	250

Table 5 - Number of cores/component, decomposition, resulting size of the partitions and aspect ratio for the grid partitions for the VHR-VHR grids with opposite decompositions

Figures 5a and 5b present the time for the coupling initialisation and for the ping-pong average time (over 97 ping-pongs) respectively. As for the VHR-VHR case with matching decompositions (3.2.2), we were not able to run the test cases on 2 and 6 cores/components for ESMF and on 2 cores/components for OASIS3-MCT.

We see here that up to 96 cores, the results are very similar to the 3.2.2 case when the VHR grids have matching decompositions. But above 96 cores, the fact that the grids have opposite decomposition and that the partitions have big aspect ratios strongly influence the results. This is intuitively expected as non-matching (opposite) decompositions imply more communication; and this becomes increasingly true as the aspect ratio of the grid partitions increases. For example, in the test running on 864 cores/component, when the decompositions are "opposite" with partitions of 3000x3 and 3x3000 grid points respectively for the source and the target, each of the 864 source processes communicate with all 864 target processes, while with matching partitions of 83x125 grid points for both grids (in 3.1.2), each of the 864 source processes communicates only with the corresponding target process.

In particular, one can note here the nice behaviour of YAC for all number of cores and of ESMF for number of cores greater or equal to 2880. A hypothesis is that these coupling technologies implement some clever multiple-step data redistributions avoiding one to all communications. And we currently have no explanation for the strange behaviour of ESMF from 216 to 2304 cores/component currently, even after discussing the issue with ESMF developers.

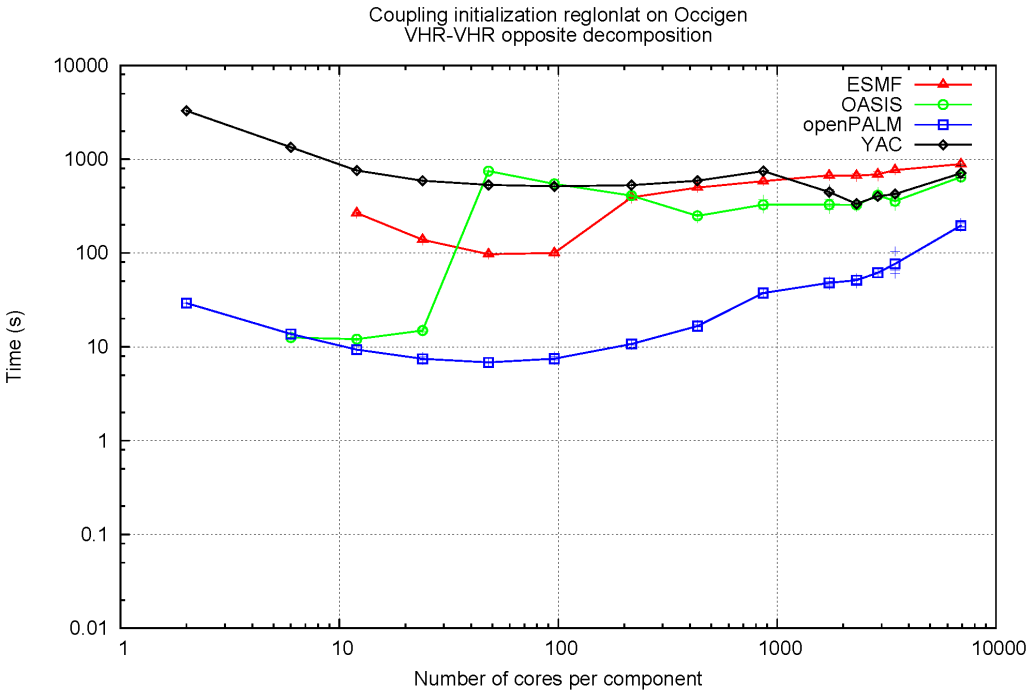

Figure 5a – Time for coupling initialisation on Occigen for components running on VHR (3000x3000) grids with opposite decompositions.

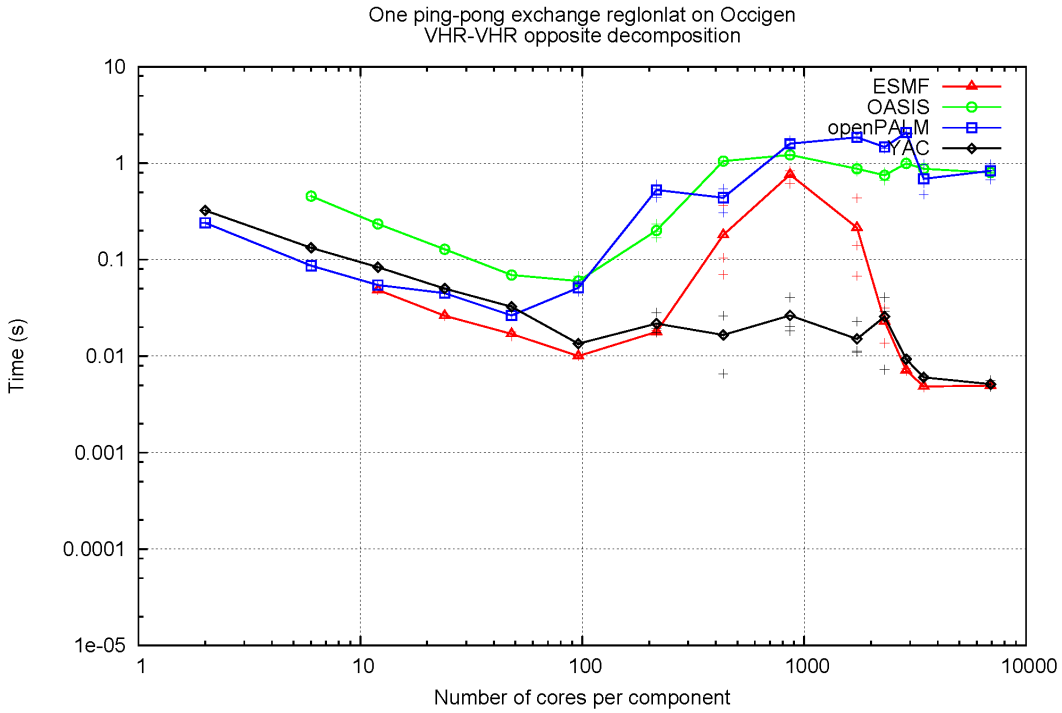

Figure 5b – Average time for one ping-pong exchange on Occigen for components running on VHR (3000x3000) grids with opposite decompositions

3.3 Results on the Cray XC40 at the UK MetOffice

The coupled benchmark test-case with components running on regular latitude-longitude grids of different sizes was also run on the Cray XC40 at the MetOffice.

The Met Office Cray, XCS, is situated in Exeter, UK. For its compute nodes, it uses Intel Xeon E5-2695v4 (18-core) processors running at 2.1GHz. The nodes are connected together through Cray's Aries Network interconnect (for more details, see http://www.cray.com/sites/default/files/resources/Cray-XC-Interconnect-Network.pdf). There are a total of 242496 cores with a theoretical peak performance of 16 PFlop/s. The coupling benchmarks were compiled with the Cray Compiler Environment 8.3.4, using cray-mpich 7.0.4 and NetCDF 4.3.2_fortran-4.4.0.

The specifications and a first analysis of the results for OASIS3-MCT, OpenPALM, ESMF, and MCT coupling technologies are presented next. Unfortunately, we were not able to run the test cases with YAC: for more than 1 core/component, the run aborted with a "segmentation fault" message and we were not able to fix that problem in due time to include YAC results in this deliverable.

3.3.1 HR-HR grids with same decomposition

The test case with regular latitude-longitude HR (1000x1000) grids and same rectangular decomposition on both sides was first run on the Cray XC40 for a number of cores per component varying between 1 and O(10^4). Table 6 shows the number of cores/component, decomposition and resulting partition size for the grids for each test.

Number of cores/component	1	12	108	1008	3456	10008
Grid 1 & grid 2 decomposition	1x1	4x3	12x9	36x28	64x54	139x72
Size of grid 1 & grid 2 partitions	1000 x 1000	250 x 333	83 x 111	28 x 36	16 x 18	7 x 14

Table 6 - Number of cores/component, decomposition and resulting partition size for each grid for the test cases on Cray XC40 with HR (1000x1000) grids with same rectangular decomposition on each side.

Results for the coupling initialisation time and for the ping-pong average time (over 97 ping-pongs) are presented on Figures 6a and 6b respectively. Regarding the ping-pong time, OASIS3-MCT, ESMF and OpenPALM show a relatively similar behaviour to that on Occigen with a ping-pong varying between 0.1 and 0.001 seconds. We note however that the extra cost of OASIS3-MCT as compared to other coupling technologies is much less noticeable on the Cray XC40 than on Occigen. MCT scales better than the other coupling technologies reaching ~0.0001 second for O(10^4) core/components.

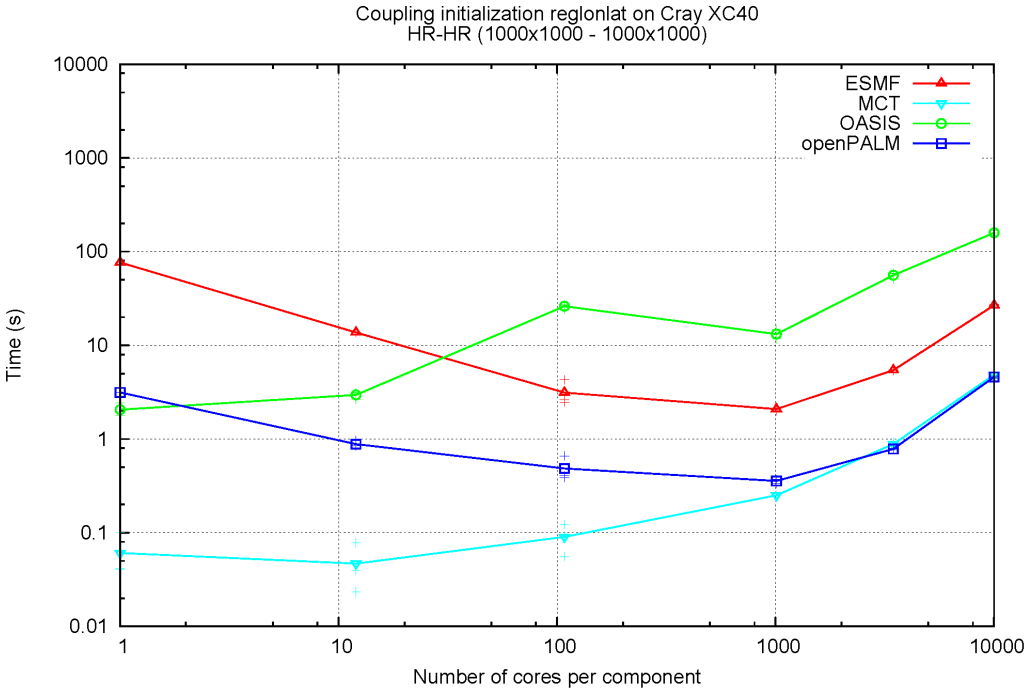

Figure 6a – Time for coupling initialisation on the Cray XC40 for components running on HR (1000x1000) grids with same decomposition on both sides.

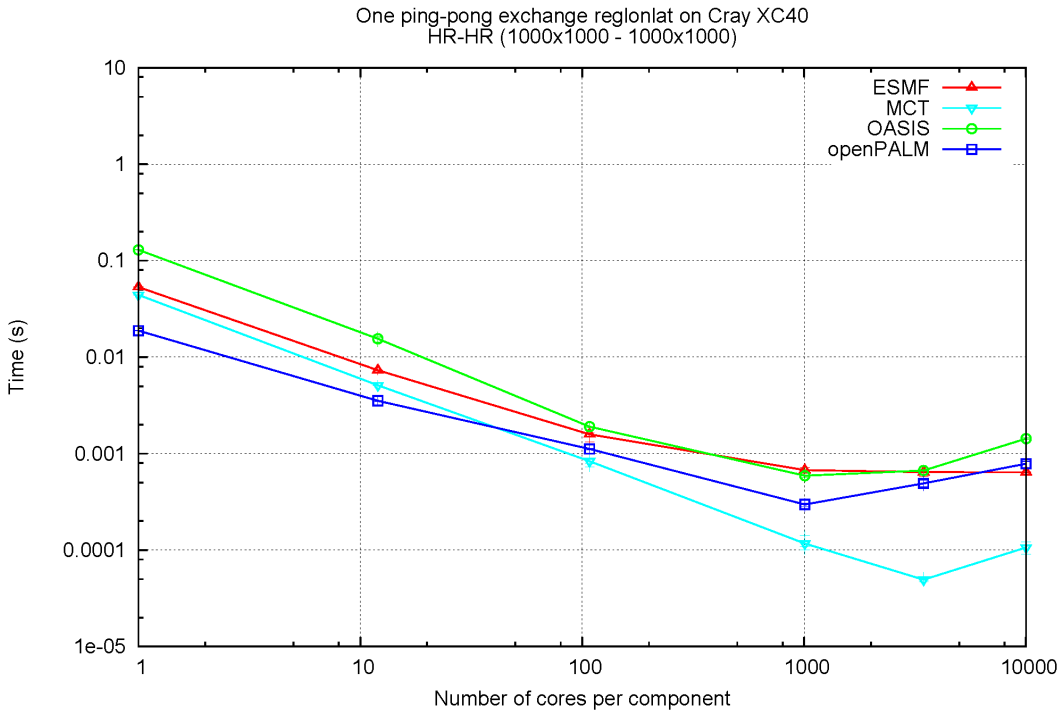

Figure 6b – Average time for one ping-pong exchange on the Cray XC40 for components running on HR (1000x1000) grids with same rectangular decomposition on both sides

3.3.2 VHR-VHR grids with same decomposition

The test case was repeated for VHR (3000x3000) grids. Table 7 shows the number of cores/component, decomposition and resulting partition size for the grids.

Number of cores/component	1	12	108	1008	3456	10008
Grid 1 & grid 2 decomposition	1x1	4x3	12x9	36x28	64x54	139x72
Size of grid 1 & grid 2 partitions	3000 x 3000	750 x 1000	250 x 333	83 x 107	47 x 56	22 x 42

Table 7 - Number of cores/component, decomposition and resulting partition size for each grid for the test cases on Cray XC40 with VHR (3000x3000) grids with same rectangular decomposition on each side.

Results for the coupling initialisation time and for the ping-pong average time (over 97 ping-pongs) are presented on Figures 7a and 7b respectively. As on Occigen, all coupling technologies behave similarly than with HR grids with the timings being somewhat bigger but stabilizing 0.001 seconds for both HR and VHR grids. Again MCT still scales better than the other coupling technologies reaching ~0.0002 second for $O(10^4)$ core/components.

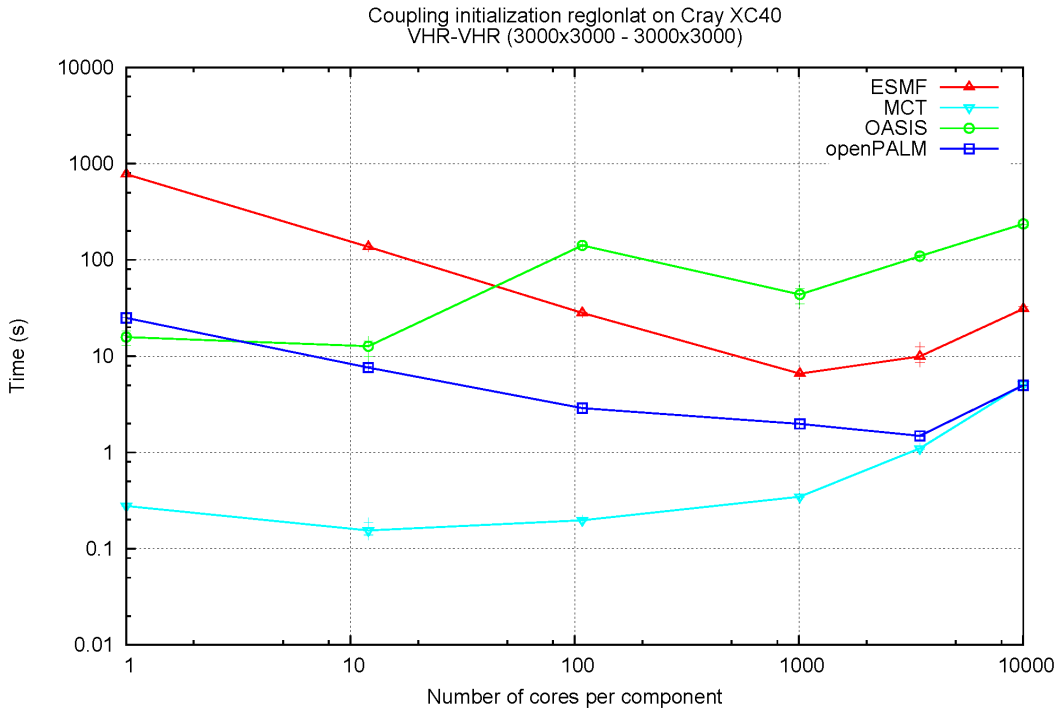

Figure 7a – Time for coupling initialisation on the Cray XC40 for components running on VHR (3000x3000) grids with same rectangular decomposition on both sides.

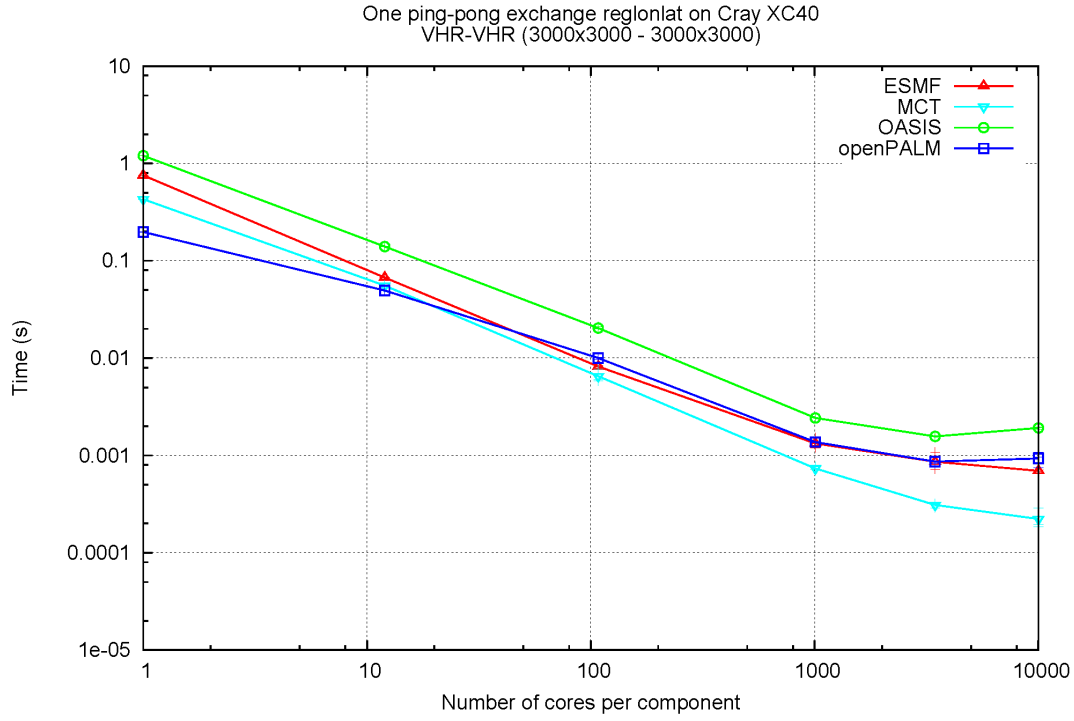

Figure 7b – Average time for one ping-pong exchange on the Cray XC40 for components running on VHR (3000x3000) grids with same rectangular decomposition on both sides

3.3.3 LR-HR grids with analogous decomposition

Results for the coupling initialisation time and for the ping-pong average time (over 97 ping-pongs) for a combination of LR (100x100) and HR (1000x1000) grids on the Cray XC40 are presented on Figures 8a and 8b respectively but only for 1, 12 and 108 cores/components. Unlike on Occigen, the rule to keep at least 50 grid points in each partition (see section 3.1.1) was strictly respected and so number of cores per component was limited to 108 because of the LR grid. As detailed in 3.1.3, these tests could not be run with MCT as the current MCT benchmark implementation does not support grids of different sizes.

For that case, remapping was needed and a one-nearest neighbour interpolation was applied. With these relatively lower resolution grid, the time needed to calculate the remapping weights-and-address file in OASIS3-MCT was reasonable and is included in the initialization timings.

As for Occigen, the results are very similar to the HR-HR results in the range of cores tested, except that the time is always smaller, which is expected as the communication load is decreased. We note also here that OpenPALM is significantly f aster than OASIS3-MCT and ESMF for unknown reasons that need further investigation.

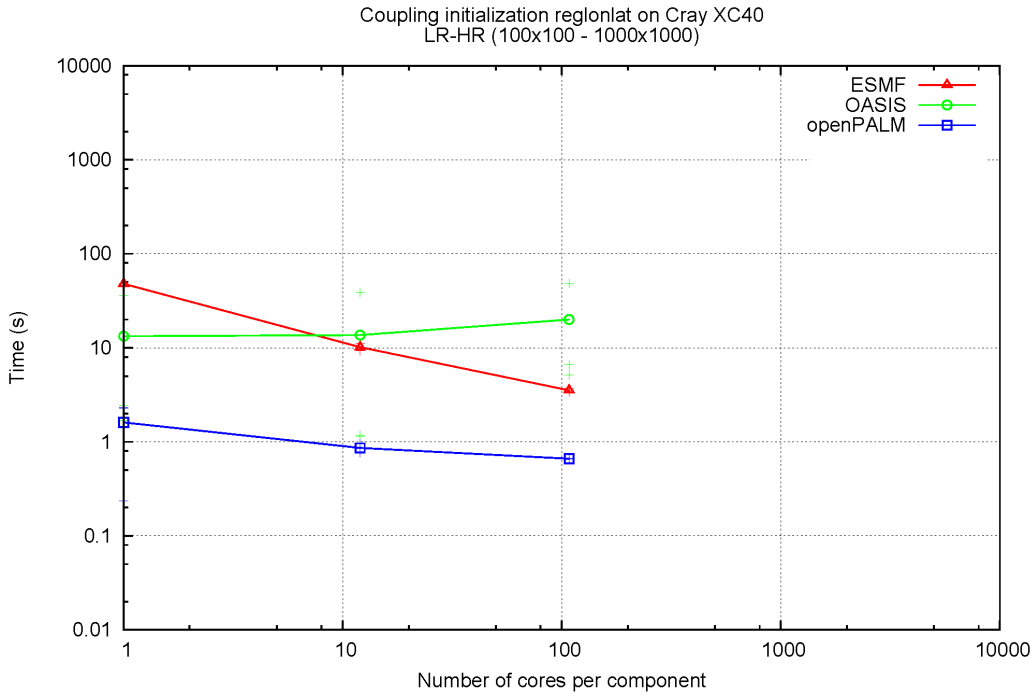

Figure 8a – Time for coupling initialisation on the Cray XC40 for components running respectively on LR (100x100) and HR grid(1000x1000) grids with rectangular decomposition on both sides

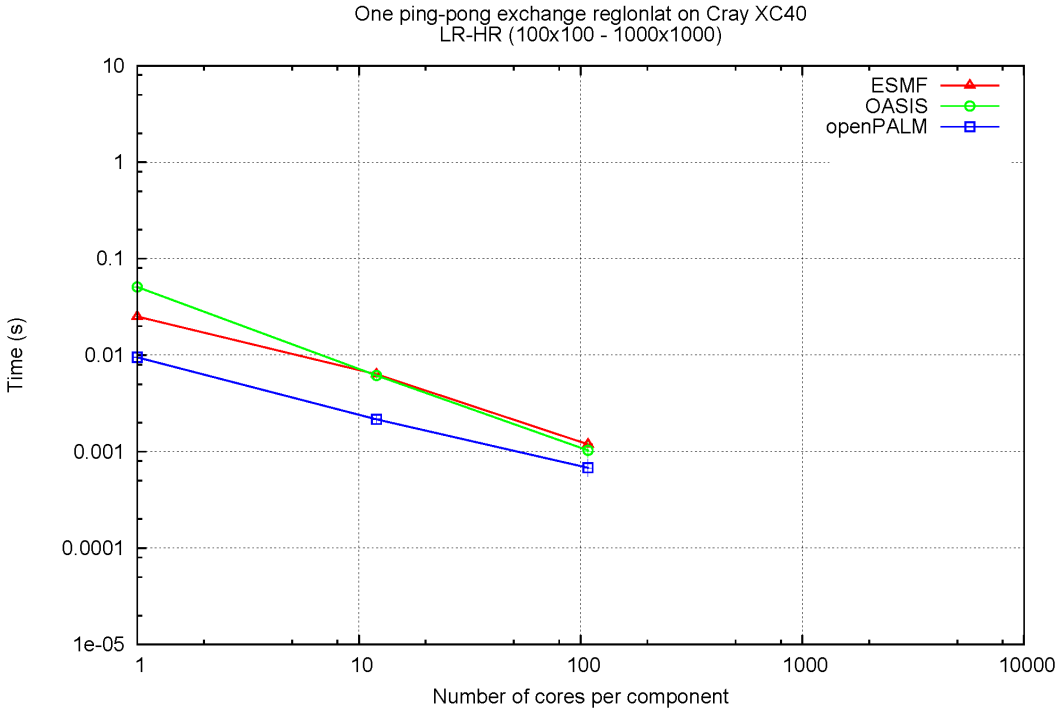

Figure 8b – Average time for one ping-pong exchange on the Cray XC40 for components running respectively on LR (100x100) and HR grid (1000x1000) grids with rectangular decomposition

3.4 Results on Broadwell partition of Marconi at CINECA in Italy.

The coupled benchmark test cases with components running on regular latitude-longitudes grids of different sizes was also run on the Broadwell partition of Marconi at CINECA in Italy. Marconi has Xeon E5-2697 v.4 (Broadwell) processors with a clock speed of 2.3 GHz. A node consists of two Xeon CPUs, each with 18 cores and the machine has 1,512 nodes in total. The interconnect is Intel Omnipath,100 Gb/s. The network topology is Fat-tree with 2:1 oversubscription tapering at the level of the core switches only. There are five OPA Core Switches ("Sawtooth Forest"), each with 768 ports. There are 216 OPA Edge Switches ("Eldorado Forest"), each with 48 ports. This gives a maximum system configuration of 5(OPA) x 768 (ports) x 2 (tapering) = 7680 servers. The coupling benchmarks were compiled with v.16.0.3 of the Intel compiler and v.5.1 of the Intel MPI library. NetCDF library version 4.4.1 compiled with Fortran v.4.4.4 with pnetcdf enabled and HDF5 version 1.10.0-patch1 were used.

To run the benchmarks on Marconi we used an allocation of 19000 core-hours in the framework of the ESiWACE Centre of Excellence.

3.4.1 HR-HR grids with same decomposition

As for the other platforms, HR grids with 1000x1000 points and same rectangular decomposition on both sides were used in a first series of tests. Table 8 shows the number of cores/component, decomposition and resulting partition size for the grids. On Marconi, a single job is limited to 6000 cores so a single component in principle to 3000; to obey the rule that the number of cores/component needs to be a multiple of 18 and to keep increasing the number of cores by one order of magnitude between the different tests, the tests were limited to a maximum of 1800 cores/component (other multiples of 18 between 1800 and 3000 could in principle have been tested but were not because of a lack of time and resources).

Number of cores/component	1	18	180	1800
Grid 1 & grid 2 decomposition	1x1	6x3	15x12	50x36
Size of grid 1 & grid 2 partitions	1000x1000	166x333	66x83	20x27

Table 8 - Number of cores/component, decomposition and resulting partition size on Marconi with HR (1000x1000) grids with same rectangular decomposition on each side.

Results for the coupling initialisation time and for the ping-pong average time (over 97 ping-pongs) are presented on Figures 9a and 9b respectively. As can be seen by the dispersion of the different colour crosses, the spread of the results (for anyone technology on any specific number of cores) is much larger on Marconi than on other platforms. Identifying “outliers” as proposed above is therefore difficult and somewhat subjective. The results presented here are therefore the ones of 3 runs without identification of outliers.[2] Here, YAC seems to scale not as well as the other coupling technologies; given the variability, more tests have to be done to evaluate the robustness of these results.

[2] This is true but for OpenPALM for 1800 cores based only on 2 runs, as the third one giving a value of about 0.2 second was clearly an outlier compared to the 2 others.

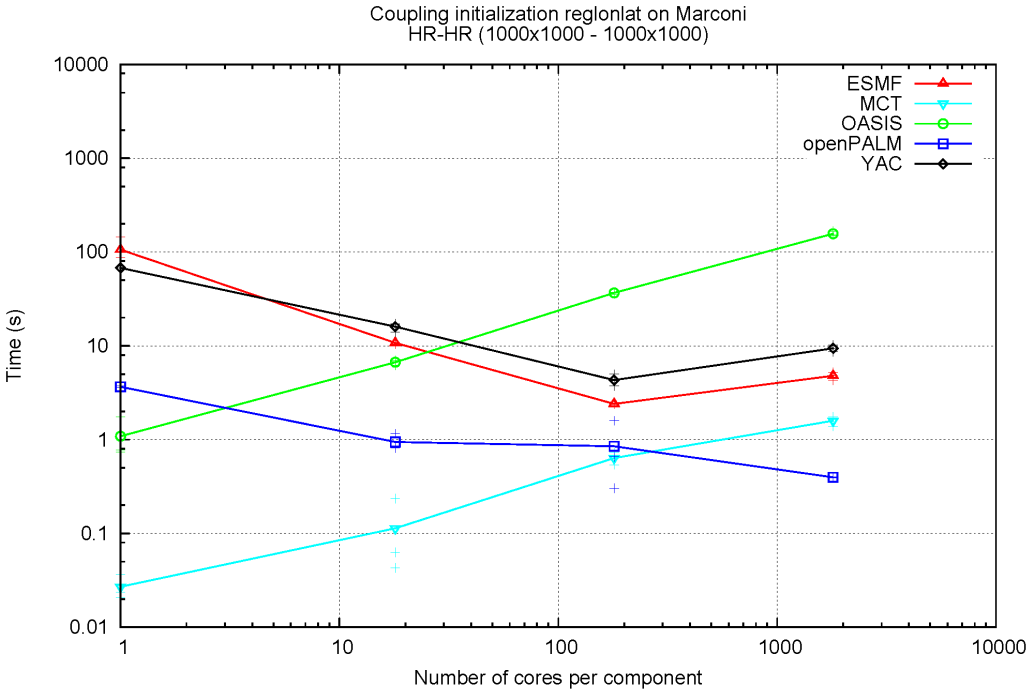

Figure 9a – Time for coupling initialisation on Marconi for components running on HR (1000x1000) grids with same rectangular decomposition on both sides.

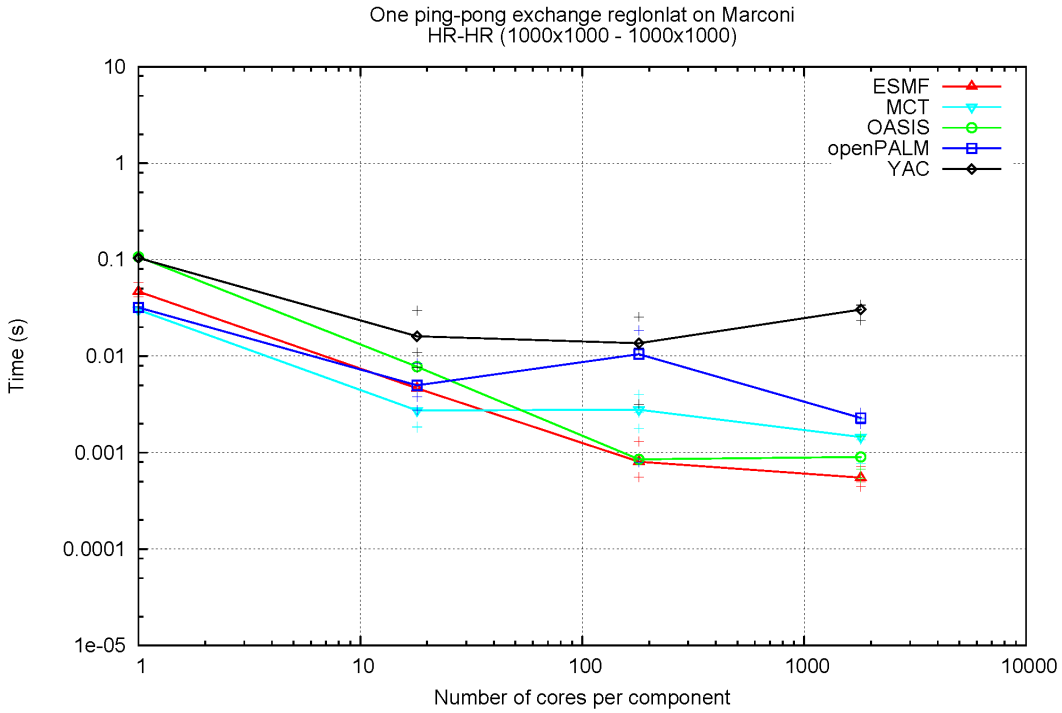

Figure 9b – Average time for one ping-pong exchange on Marconi for components running on HR (1000x1000) grids with same rectangular decomposition on both sides

3.4.2 VHR-VHR grids with same decomposition

The test case with regular latitude-longitude grids and same rectangular decomposition on both sides was repeated for VHR (3000x3000) grids. Table 9 shows the number of cores/component, decomposition and resulting partition size for the grids.

Number of cores/component	1	18	180	1800
Grid 1 & grid 2 decomposition	1x1	6x3	15x12	50x36
Size of grid 1 & grid 2 partitions	3000x3000	500x1000	200x250	60x83

Table 9 - Number of cores/component, decomposition and resulting partition size for each grid for the test cases on Marconi with VHR (3000x3000) grids with same rectangular decomposition on each side.

Results for the coupling initialisation time and for the ping-pong average time (over 97 ping-pongs) are presented on Figures 10a and 10b respectively. Again, it is hard to draw any conclusions regarding the ping-pong time, as the spread of the timings is relatively large. As for the HR case, YAC seems not to scale well for more than 1000 cores but this now seems to also be the case for OASIS3-MCT; this should be investigated further with more tests before being able to draw any firm conclusion.

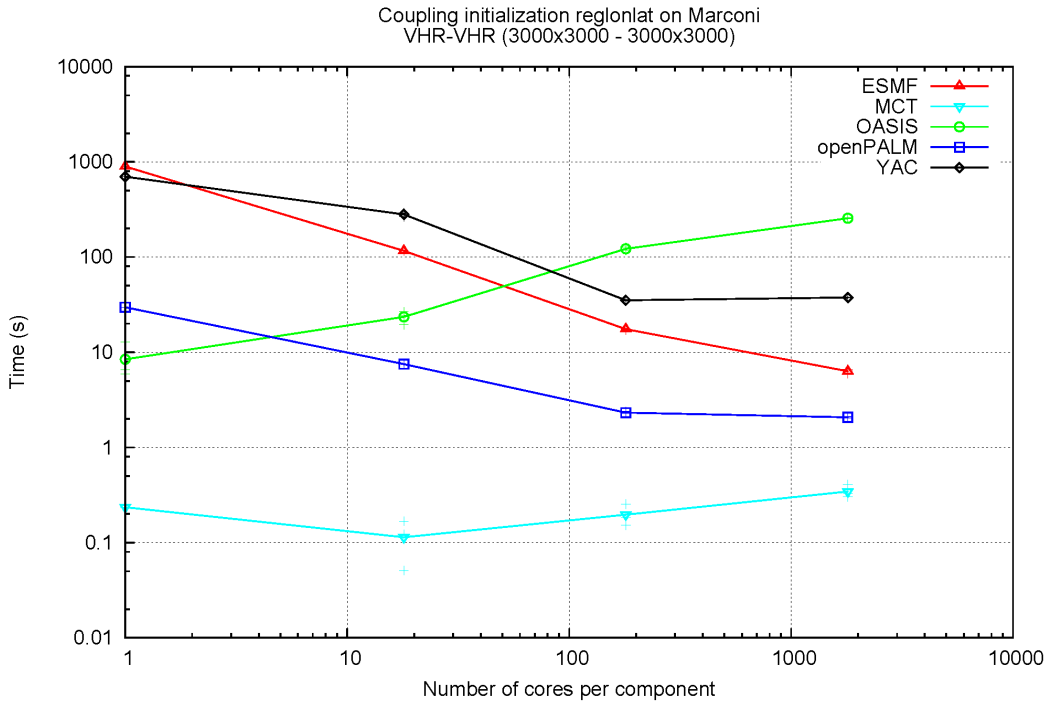

Figure 10a – Time for coupling initialisation on Marconi for components running on VHR (3000x3000) grids with same rectangular decomposition on both sides

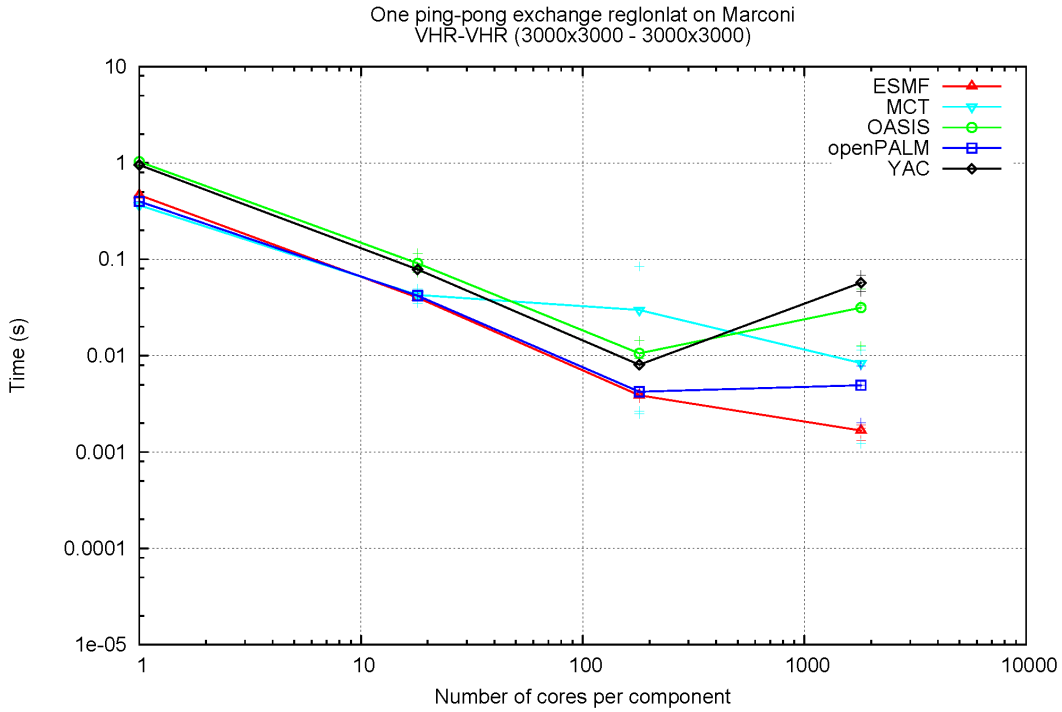

Figure 10b – Average time for one ping-pong exchange on Marconi for components running on VHR (3000x3000) grids with same rectangular decomposition on both sides

3.4.3 LR-HR grids with analogous decomposition

First results for the coupling initialisation time and for the ping-pong average time (over 97 ping-pongs) for a combination of LR (100x100) and HR (1000x1000) grids on Marconi are presented on Figures 11a and 11b respectively for the number of cores/component, decomposition and resulting partition size for the grids shown in Table 10.

Number of cores/component	1	18	180	1800
Grid 1 decomposition	1x1	6x3	15x12	50x36
Grid 2 decomposition	1x1	6x3	15x12	50x36
Size of grid 1 partitions	100x100	16x33	6x8	2x2
Size of grid 2 partitions	1000x1000	166x333	66x83	20x27

Table 10 - Number of cores/component, decomposition and resulting partition size for each grid for the test cases on Marconi with LR (100x100) - HR (1000x1000) grids with rectangular decomposition on each side.

For this case, remapping was needed and a one-nearest neighbour interpolation was applied. With the relatively lower resolution grid,. As already mentioned above, these tests could not be run with MCT as the current MCT benchmark implementation does not support grids of different sizes. We also note here that we could not get any result for the OpenPALM test with 1800 cores/component as the run would hang without producing any results; more time would be needed to investigate this specific problem.

Regarding the ping-pong time, the coupling technologies seem to behave similarly to HR-HR grids (see section 3.4.1) , with YAC now scaling much better. It is however again very hard to draw any conclusions as the dispersion of results is relatively large, especially for YAC this time.

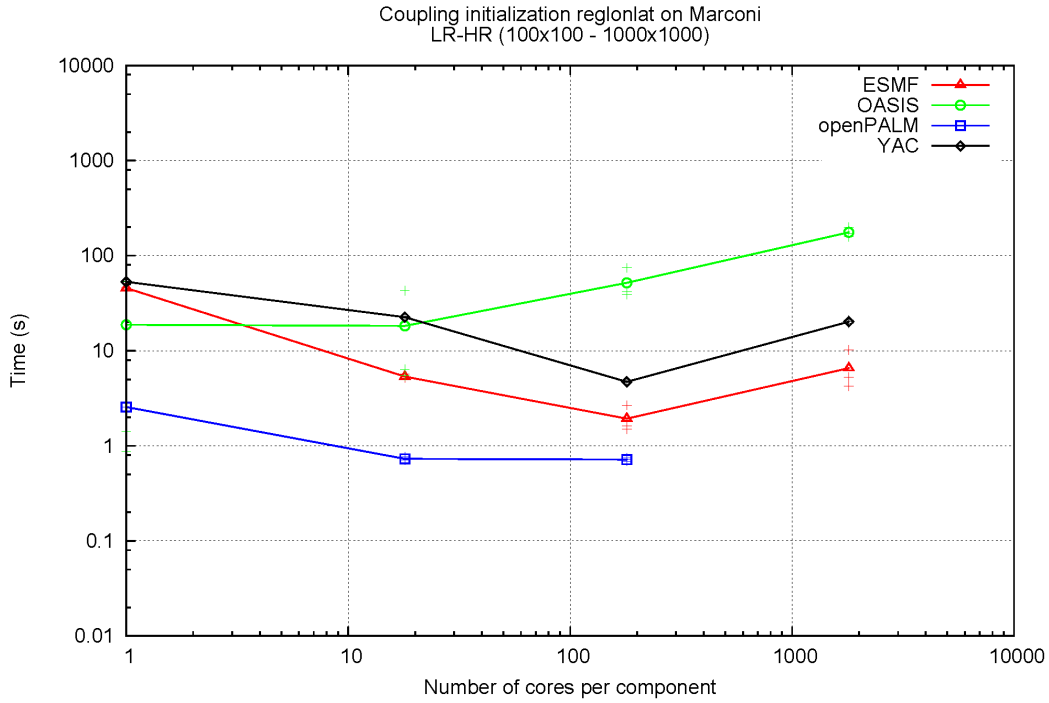

Figure 11a – Time for coupling initialisation on Marconi for components running respectively on LR (100x100) and HR grid(1000x1000) grids with rectangular decomposition on both sides

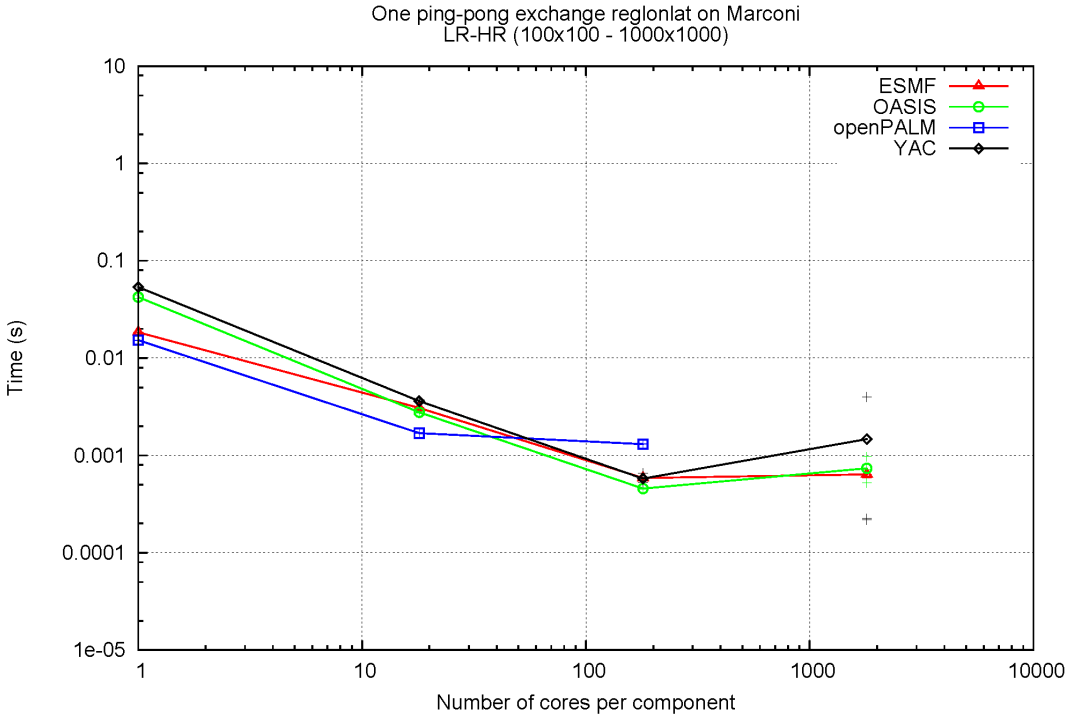

Figure 11b – Average time for one ping-pong exchange on Marconi for components running respectively on LR (100x100) and HR grid (1000x1000) grids with rectangular decomposition.

3.5 Comparison of the results on the different platforms

Figure 12 a, b, c, d and e compare the results obtained for each coupling technology, respectively for OASIS3-MCT, OpenPALM, ESMF, MCT and YAC for the 3 platforms tested, for the VHR-VHR case with same rectangular decomposition on both sides.

A first analysis leads to two general conclusions. Firstly, based on OASIS3-MCT, OpenPALM and ESMF test cases (Figs 12 a, b and c), one can conclude that results compare well on the Bullx and on the Cray XC40 for all number of cores.

Another conclusion is that the scalability curves for all coupling technologies, besides maybe for ESMF, seem to break down on Marconi for more than 1000 cores. One can therefore suspect here a problem specific to the MPI implementation, to the interconnect, or to the machine load on Marconi for high number of cores and not a default of the coupling technologies themselves. However, this observation has to be taken with care given the large variability of the results on Marconi (even if the variability *per se* can be considered as a problem). If one looks at the best of the 3 runs, its result is close to the results for the other platforms at least for OpenPALM (Fig 12b), ESMF (Fig 12c) and MCT (Fig 12d).

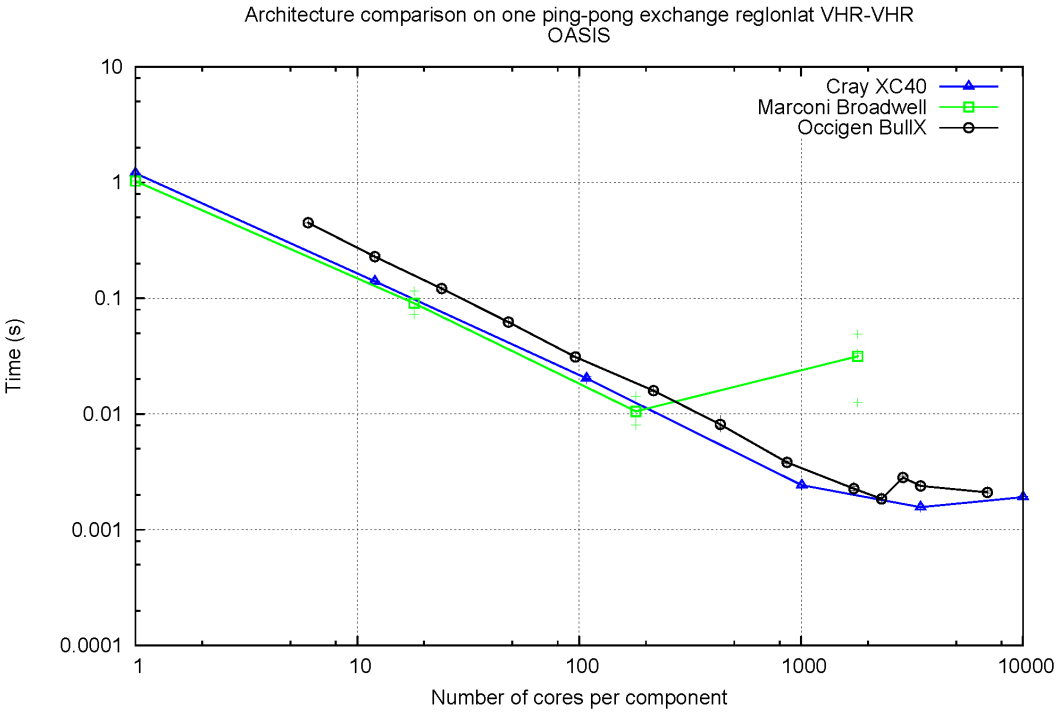

Figure 12a - Average time for one ping-pong exchange for OASIS3-MCT on Occigen Bullx, Cray XC40, and Marconi Broadwell partition for components running on VHR (3000x3000) grids

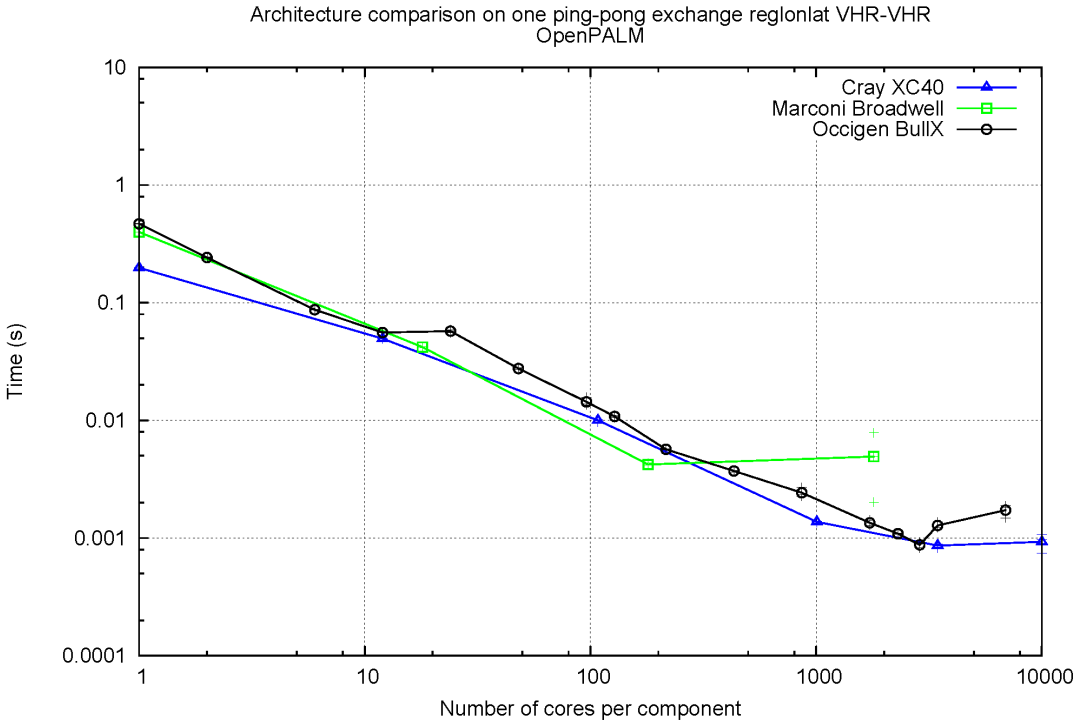

Figure 12b - Average time for one ping-pong exchange for OpenPALM on Occigen Bullx, Cray XC40, and Marconi Broadwell partition for components running on VHR (3000x3000) grids

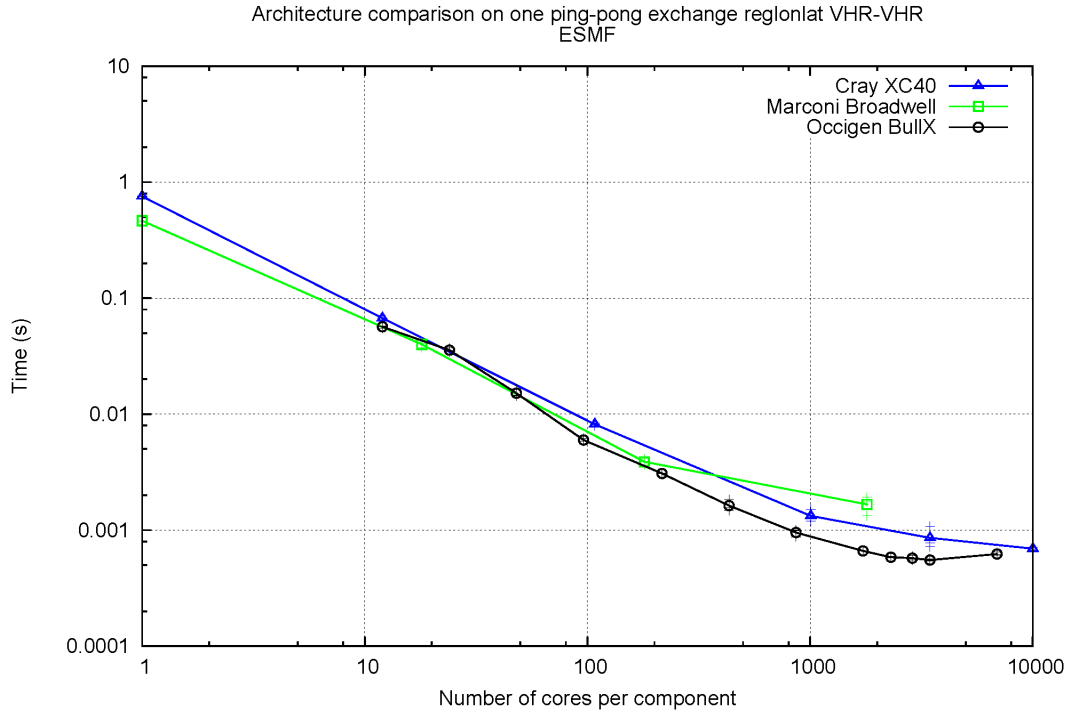

Figure 12c - Average time for one ping-pong exchange for ESMF on Occigen Bullx, Cray XC40, and Marconi Broadwell partition for components running on VHR (3000x3000) grids

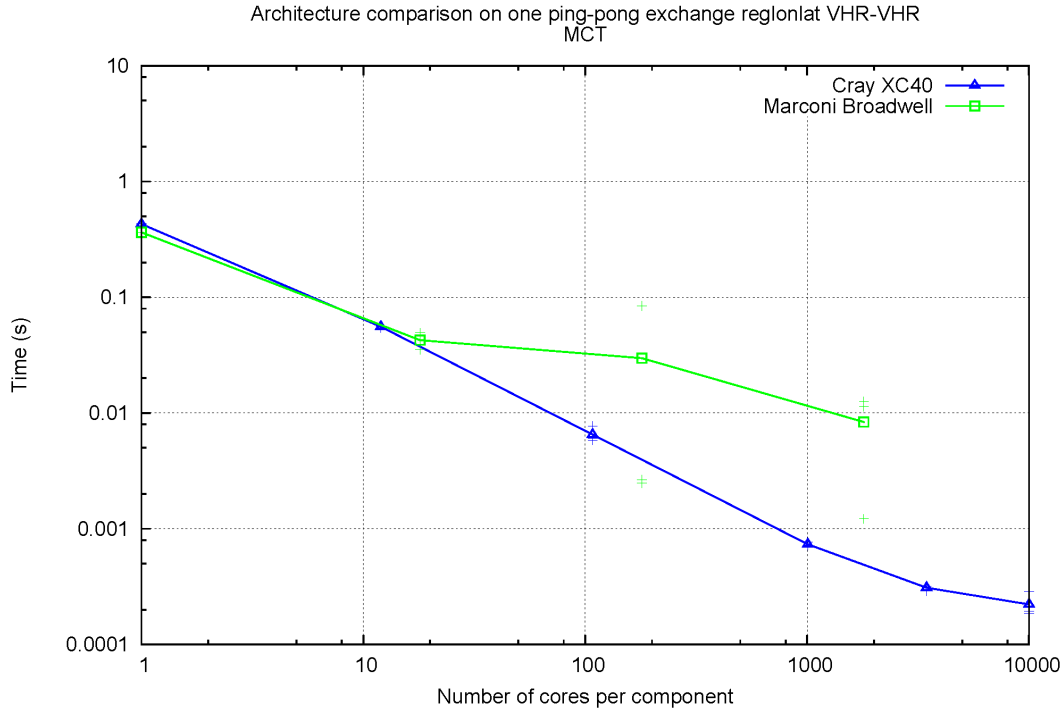

Figure 12d - Average time for one ping-pong exchange for MCT on Cray XC40, and Marconi Broadwell partition for components running on VHR (3000x3000) grids.

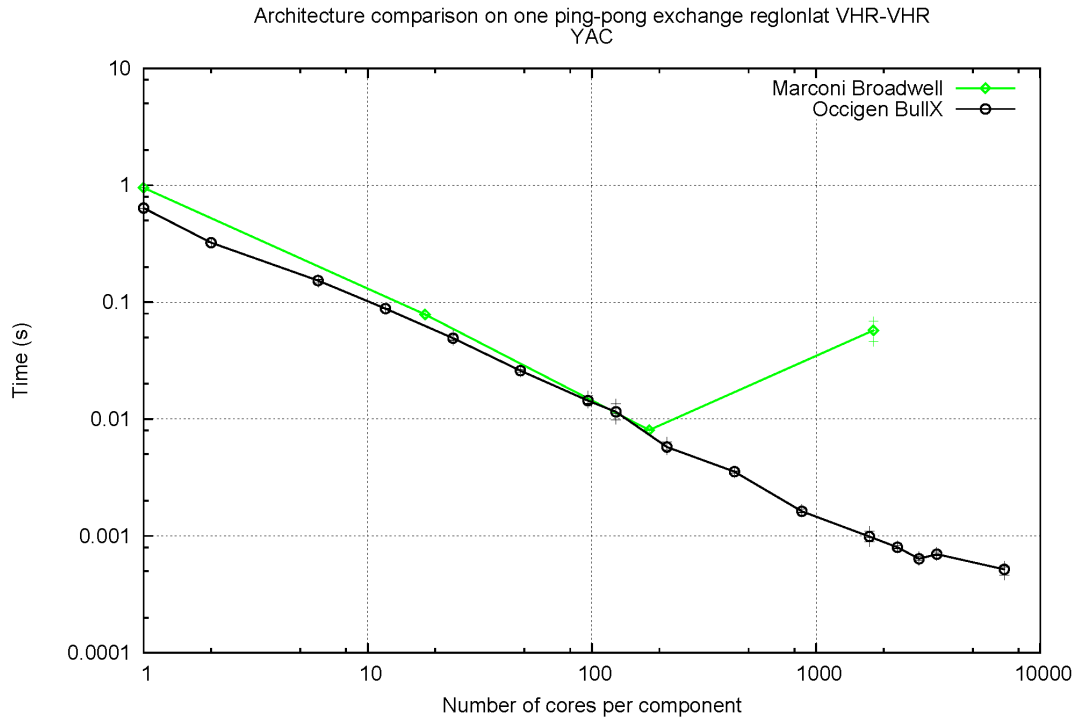

Figure 12e - Average time for one ping-pong exchange for YAC on Occigen Bullx and Marconi Broadwell partition for components running on VHR (3000x3000) grids.

4. Summary and Perspectives

This deliverable presents the work performed in the IS-ENES2 European project toward the establishment of a standard benchmarking suite for coupling technologies. Beyond the efforts devoted to that task by different partners in the project, this work forms part of a wider international community effort to characterize, in a standard way, coupling technologies used in climate modelling, and in that sense exemplifies the maturity this community has reached interacting and working together.

This work started in 2013 with the description of the possible functions of coupling technologies and the characteristics of Earth System Models (ESMs) supported by these coupling technologies during the « 2nd Workshop on Coupling Technologies » in Boulder. These characteristic were then classified by IS-ENES2 partners interacting with colleagues of the US project Earth System Bridge in a series of mindmaps, which are available at https://earthsystemcog.org/projects/es-fdl/mindmaps. The next step for the benchmark development was to define a priority of coupling characteristics to benchmark and the general specifications of the benchmark suite. It was established that the benchmark suite would include a number of pre-coded stand-alone components running on different grids; these would then be assembled thanks to the different coupling technologies in coupled different coupled test cases. All these steps are detailed in IS-ENES2 milestones M10.1 and M10.4 and summarized in this document.

A first version of the IS-ENES coupling technology benchmark suite is available today on the ENES portal. The stand-alone components consist of simple, individual model codes containing no physics or dynamics but representative of real models in term of coupling characteristics and following precise specifications. Four stand-alone components are available running on the following different grids: 1) a self-generated regular latitude-longitude grid, 2) an irregular, stretched and rotated latitude-longitude mesh, following the ORCA configuration of the NEMO ocean model, 3) a quasi-uniform icosahedral mesh, following the atmospheric DYNAMICO model, and 4) a quasi-uniform cubed sphere mesh. The stand-alone component using the self-generated regular latitude-longitude grid was used to assemble toy coupled models using five different coupling technologies, OASIS3-MCT, OpenPALM, ESMF, MCT and YAC.

Test cases were run with this first version of the IS-ENES benchmark to evaluate the impact of: the number of cores per component, the grid sizes, and using different ratio in the number of processes used for each component on the performance of the different coupling technologies. These tests were run on three different platforms: Bullx at CINES in France, Cray XC40 at the UK MetOffice, the Broadwell partition of Marconi at CINECA in Italy and the time for the coupling initialisation and for the coupling exchanges for these coupled configurations running with up to O(10000) cores are detailed in respectively sections 3.2, 3.3, and 3.4 of this document. Finally, a comparison of results obtained on the three platforms for each coupling technology is done in section 3.5.

These results are presented in this deliverable as a demonstration of how this benchmark environment could be used. However, these first results should not be used, as is, to draw any firm conclusions on the performance of the coupling technologies. Much work is needed at this stage to study the significance and the robustness of these results and understand why in few specific cases, some coupling technologies show much better or much worse behaviour than the others, or why some tests abort. It should also be stressed that these results are of

course valid only for the very specific test case implemented here with its specific coupling and environment characteristics (type of grid, type of decomposition, platform, etc.)

Extensions of this first IS-ENES coupling technology benchmark suite are almost infinite. Of course, the same test cases could be repeated for toy coupled models based on components running on the different grids included in the stand-alone components. Also, as detailed in milestone M10.4, the impact of the schedule and layout and the impact of the number of coupling fields could be tested. One aspect that has not been touched either is the type of remapping used when the grids are different in the two components. At this point, anyone is welcome to use or extend the IS-ENES coupling technology benchmark suite. However it would be essential to do so in order to continue the community work that started in 2013, reporting back on extensions and keeping the IS-ENES community informed about results obtained.

.