CERFACS

EUROPEAN CENTRE FOR RESEARCH AND ADVANCED TRAINING IN SCIENTIFIC COMPUTING

# Latest developments of the OASIS3-MCT coupler for improved performance

S. Valcke, L. Coquart, A. Craig, G. Jonville, E. Maisonnave, A. Piacentini

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

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

# Outline

- Some concepts about code coupling
- OASIS3-MCT overview: history, community, generalities
- Use of OASIS3-MCT: code interfacing, configuration
- OASIS3-MCT parallel communication
- Latest developments: OASIS3-MCT_4.0
- Summary and conclusions

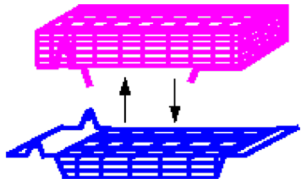
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

# Some concepts about code coupling

## Why couple atmosphere, land, ocean, sea-ice models?

- Of course, to treat the Earth System globally



## What does "coupling of codes" imply?

- Exchange and transform information at the code interface
- Manage the execution and synchronization of the codes

## What are the constraints?

- ✓ Coupling should be easy to implement, flexible, efficient, portable
- ✓ Coupling algorithm dictated by science (sequ. vs conc. coupling)
- ✓ Start from existing and independently developed codes
- ✓ Global performance and load balancing issues are crucial
- ✓ Platform characteristics (OS, CPU, message passing efficiency, …)

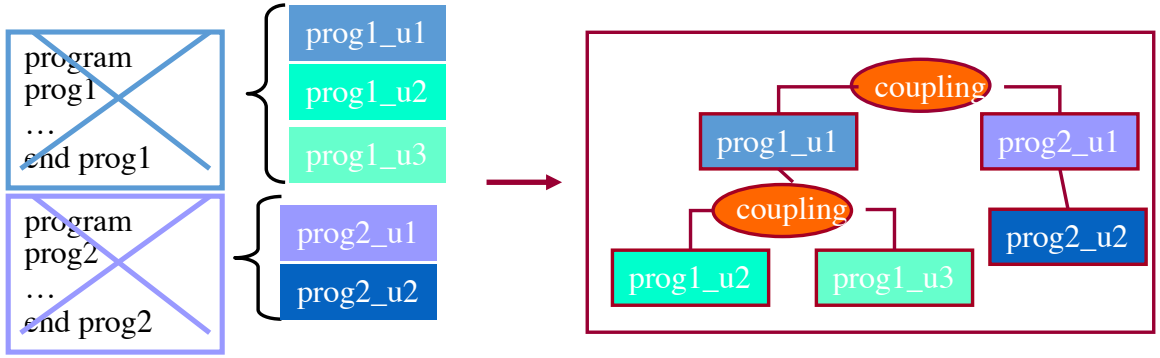
# Some concepts about code coupling

## Two main classes of coupling technologies

### 1. coupling framework

ESMF FMS(GFDL) CESM (NCAR)

- Split code into elemental units
- Write or use coupling units
- Use the library to build a hierarchical merged code
- Adapt code data structure and calling interface




☺ efficient
☺ sequential and concurrent components
☺ use of generic utilities ( parallelisation, regridding, time management, etc.)

☹ existing codes
😐 (easy)

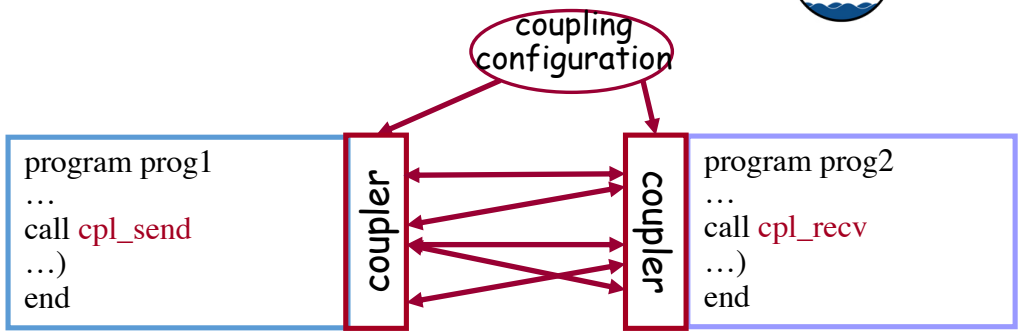
→ probably best solution in controlled development environment

# Some concepts about code coupling

## Two main classes of coupling technologies

### 2. coupler or coupling library

Oasis  O-Palm  mct  YAC  C-Coupler

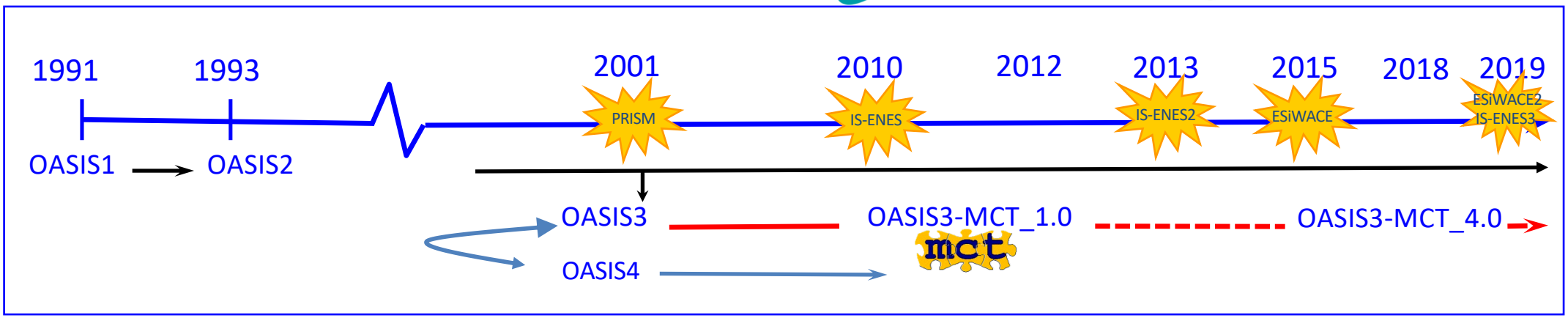



- ☺ existing codes
- ☺ use of generic transformations/regridding
- ☺ concurrent coupling (parallelism)

- ☹ efficient
- ☹ multi-executable: more difficult to debug; harder to manage for the OS

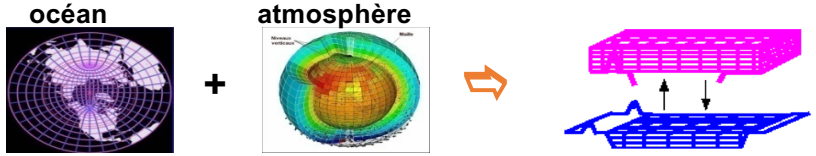
→ probably best solution to couple independently developed codes

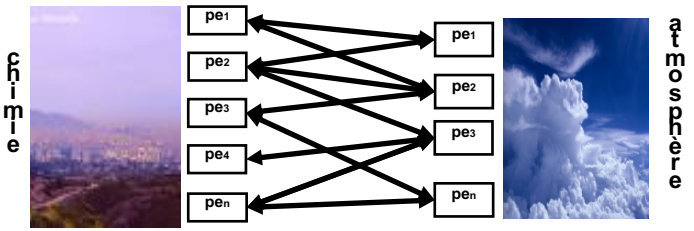
# OASIS3-MCT overview: history




- **OASIS1 -> OASIS2 -> OASIS3:**
  - **2D ocean-atmosphere coupling**
  - **low frequency, low resolution :**
  - **→ Flexibility, 2D interpolations**

- OASIS4 / **OASIS3-MCT:**
  - **2D/3D coupling of high-resolution parallel components**
  - **→Parallelism, performance**



# OASIS3-MCT overview: the community

More than ~45 groups in France, Europe and all over the world use OASIS, mainly for climate modelling, e.g.:

- France: CERFACS,CNRM, LOCEAN, LMD, LSCE, LA, LEGOS, LGGE, IFREMER, ENSTA
- Europe: ECMWF + EC-Earth community
- Germany: MPI-M, IFM-GEOMAR, HZG, U. Frankfurt, BTU-Cottbus
- UK: MetOffice, NCAS/U. Reading, ICL,
- Denmark: DMI
- Norway: U. Bergen
- Sweden: SMHI, U. Lund
- Irland: ICHEC, NUI Galway
- Netherlands: KNMI
- Belgium: KU Leuven
- Switzerland: ETH Zurich
- Italy: INGV, ENEA, CASPUR
- Czech Republic : CHMI
- Spain: IC3, BSC, U. Castilla
- Tunisia: Inst. Nat. Met
- Saudi Arabia: CECCR
- Japan: U. Tokyo, JMA, JAMSTEC
- China: IAP-CAS, Met. Nat. Centre, SCSIO
- Corea: KMA
- Australia: CSIRO, BoM, ACT, NCI
- New-Zealand: NIWA, NCWAR
- Canada: Fisheries and Oceans, U. Waterloo, UQAM
- USA: Oregon St. U., Hawaii U., JPL, MIT
- Peru: IGP

+ downloads from du Nigeria, Colombia, Singapour, Russia, Thaïland, …

OASIS3-MCT is used in 5 of the 7 European ESMs participating to CMIP6

# OASIS3-MCT overview: generalities

- All sources are written in F90 and C
- Uses the Model Coupling Toolkit (MCT) from Argonne National Lab
- Open source product distributed under a LGPL license
- All external libraries used are public domain (MPI, NetCDF) or open source (LANL SCRIP, MCT)
- Current developers are:
  - 1.5 permanent FTEs (CERFACS, CNRS)
  - 2 consultants: A. Craig (also developing for NCAR and ESMF), A. Piacentini

ESiWACE H2020 EU Centre of Excellence
- ESiWACE1 (2015-2019): 18 pm
- ESiWACE2 (2019-2022): 16 pms

IS-ENES EU FP7 project
- IS-ENES2 (2014-2017): 27 pm
- IS-ENES3 (2019-2022): 35 pms


# Use of OASIS3-MCT

At run time, OASIS3-MCT acts as a communication library linked to the models.

To use OASIS3-MCT:

- Download the sources, compile and run the tutorial on your platform
- Identify your component models, grids, coupling fields to be exchanged
- Identify the transformations to go from the source to the target component models
- Use the "test_interpolation" environment (offline) to test the quality
- **Adapt your codes i.e. insert calls to OASIS3-MCT library**
- Choose the other parameters (source and target, frequency, field trans -formations, etc.) and **create the *namcouple* configuration file**
- Compile OASIS3-MCT, your components **with same compiler**, and link the components models with OASIS3-MCT library
- Start the models and let OASIS3-MCT manage the coupling exchanges

# Use of OASIS3-MCT: code interfacing

- Initialisation: `call oasis_init_comp(...)`
- Grid definition: `call oasis_write_grid (...)`
- Local partition definition: `call oasis_def_partition (...)`
- Coupling field declaration: `call oasis_def_var (...)`
- End of definition phase: `call oasis_enddef (…)`
- **Coupling field exchange**:
  - in model time stepping loop

```
call oasis_put (…, date, var_array. …)
call oasis_get (…, date, var_array, …)
```

  - user defines externally the source or target
  - sending or receiving at appropriate time only
  - automatic averaging/accumulation if requested
  - automatic writing of coupling restart file at end of run
- Termination: `call oasis_terminate (…)`

# Use of OASIS3-MCT: coupling exchange configuration

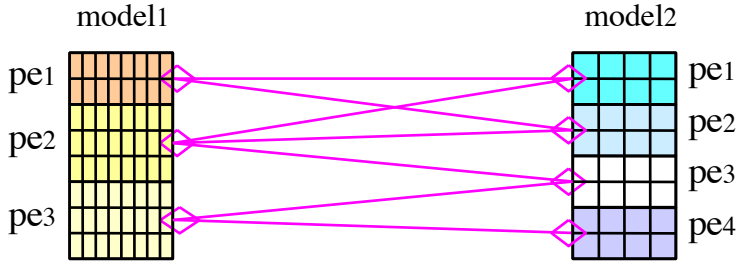
Configuration in a **text** file *namcouple*

- general characteristics of a coupled run
  - total duration
  - components
  - …
- for each exchange of coupling field :
  - source and target symbolic name (end-point communication)
  - exchange period
  - transformations/interpolations

```
# Field 3: model2 to model1
FSENDATM  FRECVOCN  1 10800  1  fdatm.nc EXPOUT
96 72 182 149 lmdz torc LAG=+1800
P  0  P  2
SCRIPR
BILINEAR LR SCALAR LATLON 1
```

# OASIS3-MCT parallel communication

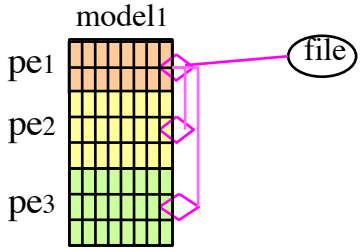
•Fully parallel communication
based on Message Passing Interface (MPI)
between **parallel models running concurrently**
or between **two components running sequentially within one same executable**




*If required, the interpolation weights and addresses are calculated in parallel (new in OASIS3-MCT_4.0) on model processes*

*Interpolation per se from the source grid to the target grid is done in parallel on the source or on the target processes*

•I/O functionality
(switch between coupled
and forced mode):



# Latest developments: OASIS3-MCT_4.0

## OASIS3-MCT_4.0 released in July 2018:

### Few additional functionalities:

- Bundle fields
- Automatic writing of coupling restart files
- Check of consistency between the number of weights and fields

### Optimisation and bugfixes

- Bypass of matrix-vector multiplication for identical grids – impact on IS-ENES2 benchmarks
- Optimising the coupling initialisation
- Upgrade of MCT library
- New more performant algorithms for the global CONSERV operation
- Optimisation of the communication using the mapping weights
- Hybrid MPI+OpenMP parallelisation of the SCRIP library

**Publications:**

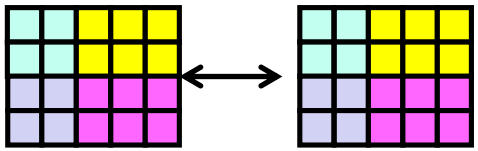

- A. Craig, S. Valcke, L. Coquart, 2017: Development and performance of a new version of the OASIS coupler, OASIS3-MCT_3.0, Geosci. Model Dev., 10, 3297-3308
- S. Valcke, A. Craig and L. Coquart, 2018. OASIS3-MCT User Guide, OASIS3-MCT4.0, CECI, Université de Toulouse, CNRS, CERFACS - TR-CMGC-18-77, Toulouse, France

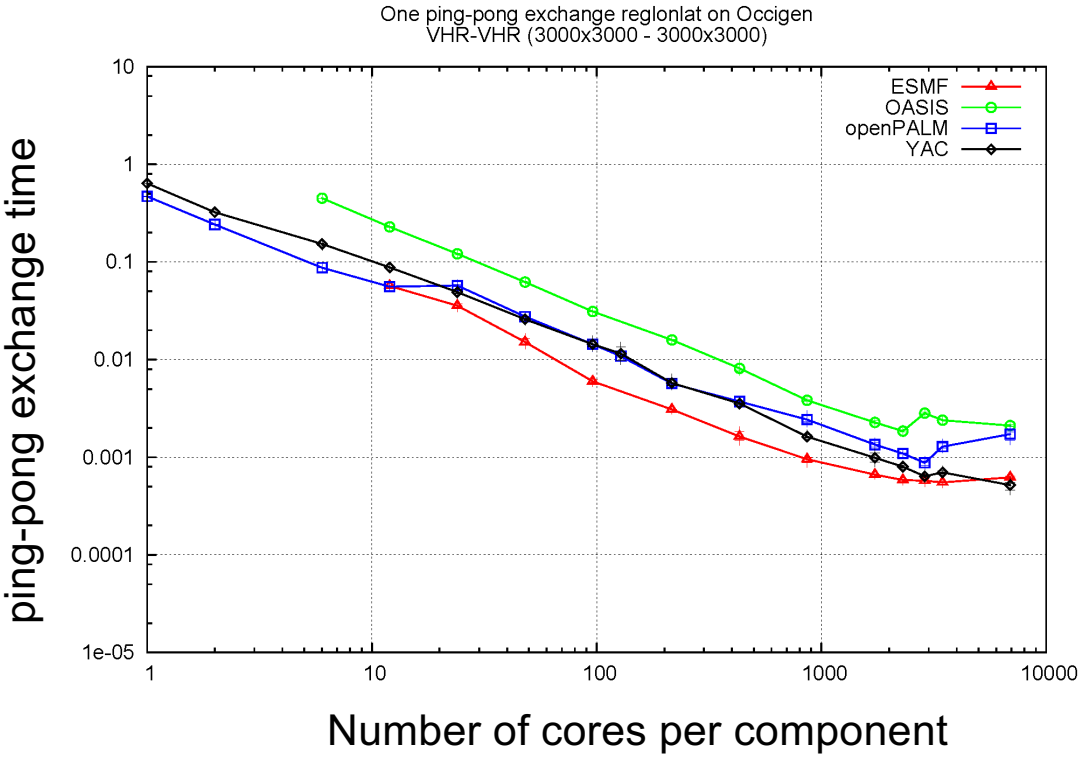
# Latest developments: OASIS3-MCT_4.0

## Bypass of matrix-vector multiplication for identical grids and impact on IS-ENES2 benchmark

- First IS-ENES2 benchmark results:

IS-ENES2 benchmark VHR:
ping-pong exchange between
3000x3000 regular lat-lon grids
same decomposition



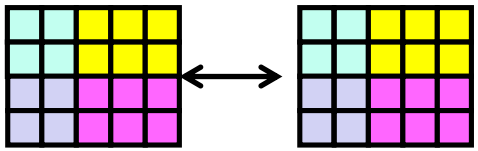


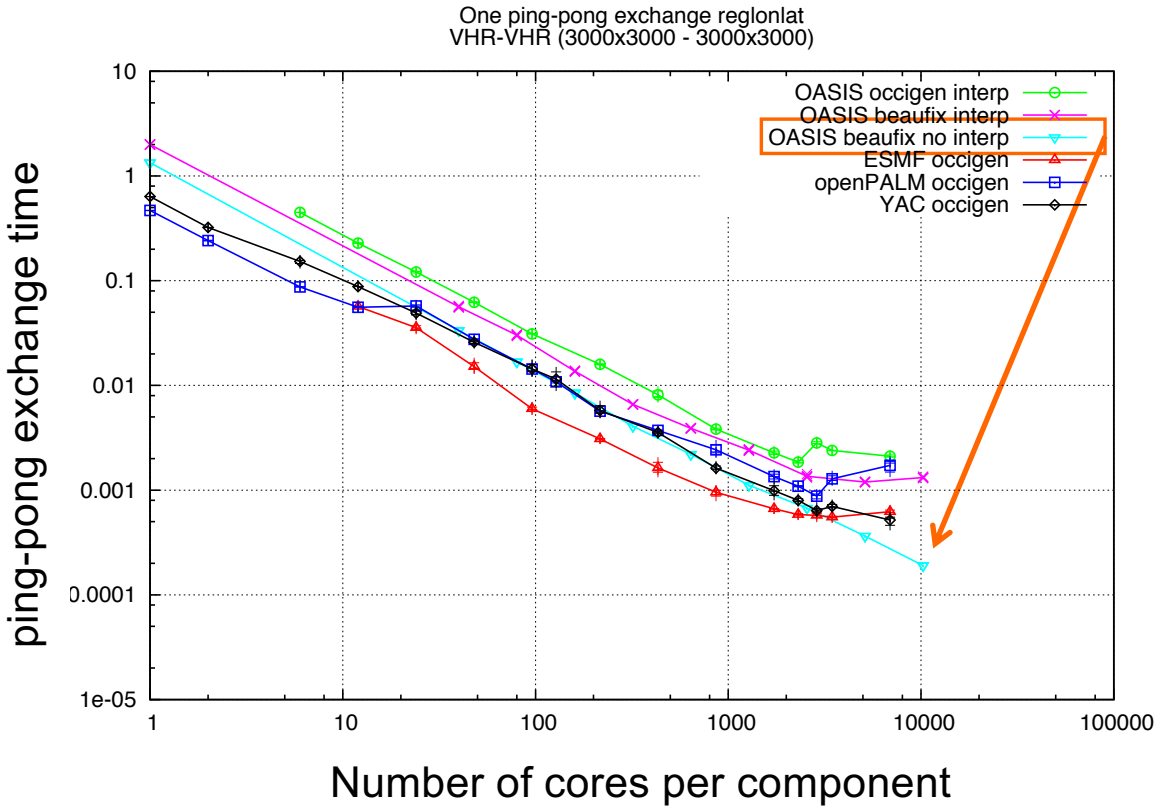
# Latest developments: OASIS3-MCT_4.0

## Bypass of matrix-vector multiplication for identical grids and impact on IS-ENES2 benchmark

- With “nointerp” : bypass of the (identity) matrix-vector product for identical grids

IS-ENES2 benchmark VHR:
ping-pong exchange between
3000x3000 regular lat-lon grids
same decomposition

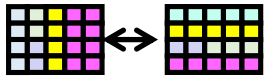


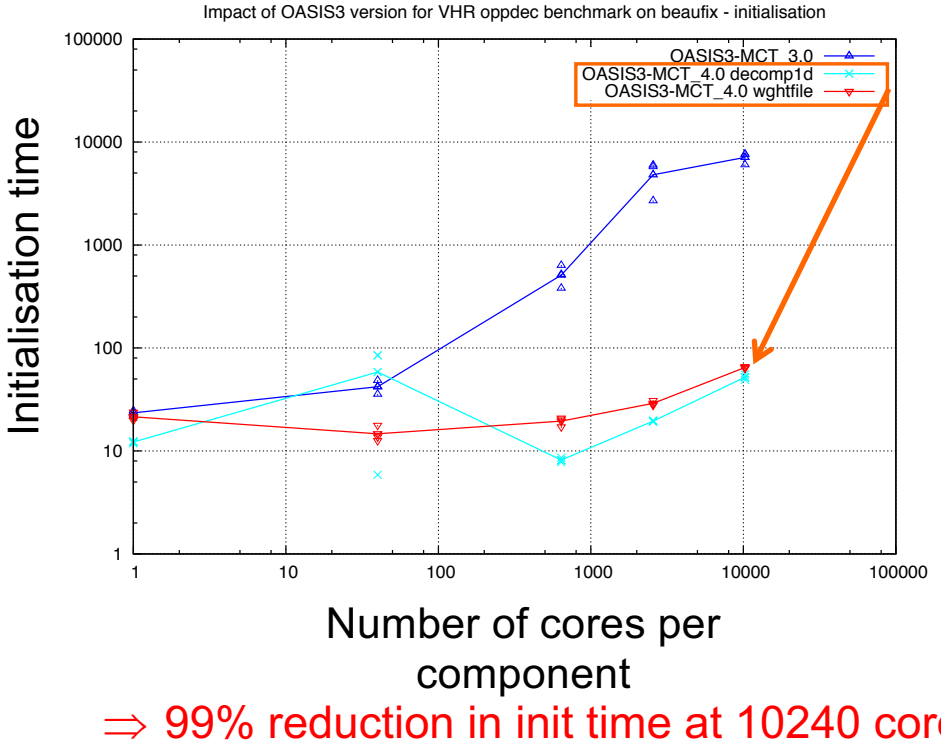


# Latest developments: OASIS3-MCT_4.0

- Removal of concurrent writing into the OASIS3-MCT debug files at initialization

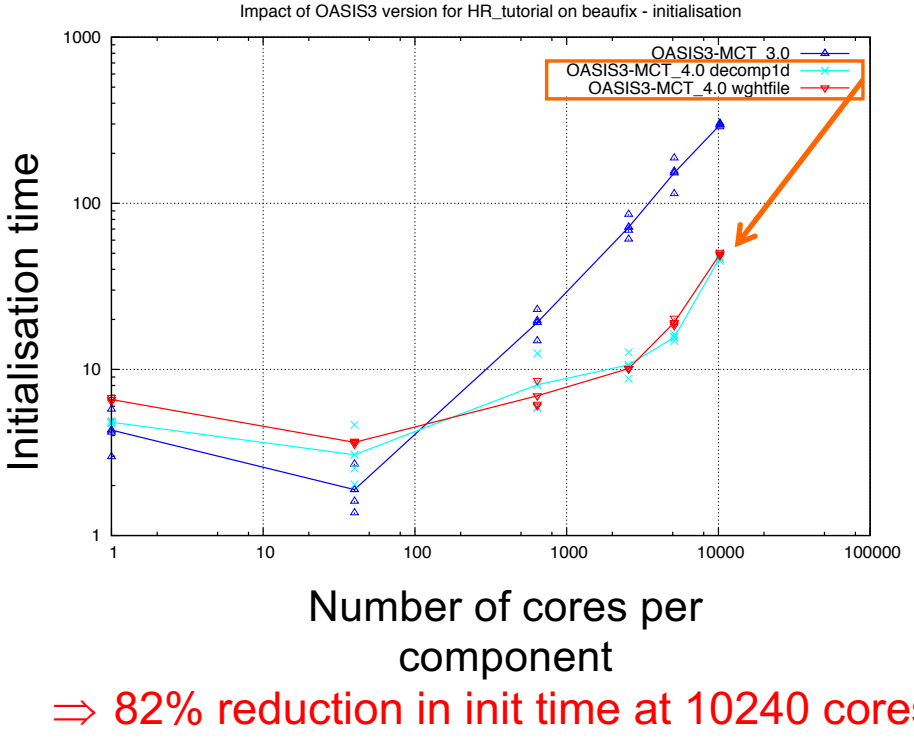
IS-ENES benchmark VHR: 3000x3000
reg lat-lon grids, opposite decompositions






⇒ 99% reduction in init time at 10240 cores

NEMO ORCA025 grid (1021x1442) –
Gaussian Reduced T799 grid (843 000)



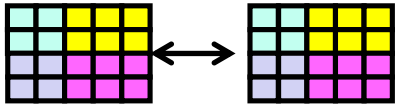

⇒ 82% reduction in init time at 10240 cores

# Latest developments: OASIS3-MCT_4.0

- Update of MCT library from version 2.8 to 2.10.beta1

Reduces by O(10) – O(100) the MCT router initialization cost
e.g. for the IS-ENES2 benchmark VHR test case:

3000x3000
regular lat-lon grids
same decomposition



- cost to compute the router between source and mapping decomposition :
  - 19 -> 0.5 sec (1600 tasks/comp)
  - 41 -> 0.7 sec (3600 tasks/comp)
- cost to compute the router between mapping and target decomposition :
  - 60 -> 6-7 sec (1600 tasks/comp)
  - 124 -> 5-7 sec (3600 tasks/comp)

# Latest developments: OASIS3-MCT_4.0

➢ New algorithms for the global CONSERV operation (forces the global conservation of the coupling field)

In OASIS3-MCT_3.0:

- *bfb* : entire field gathered and summed on the master process, result broadcasted to all other processes
  - ➢ bit-for-bit reproducibility
- *opt*: local sum by each process sent to all other processes, then global sum is performed by all
  - ➢ more efficient but no bit-for-bit reproducibility

In OASIS3-MCT_4.0:

- *gather <-> bfb*
- *lsum8 <-> opt*

- *lsum16 :* as *lsum8* but with quadruple precision
  - ➢ more costly but higher chance of reproducibility
- *reprosum*: fixed point method (Mirin &Worley, 2012)
  - ➢ expected to produce bit-for-bit results except in extremely rare cases
- *ddpdd:* parallel double–double algorithm with single scalar reduction (He &Ding, 2001)
  - ➢ should behave between lsum8 and lsum16

| cores, mapping | CONSERV unset | CONSERV *lsum8* | CONSERV *lsum16* | CONSERV *ddpdd* | CONSERV *reprosum* | CONSERV *gather* |
|---|---|---|---|---|---|---|
| 48, *src* | 4.00 | 8.27 | 16.78 | 10.65 | 17.34 | 117.72 |
| 48, *dst* | 4.39 | 8.02 | 16.59 | 10.42 | 16.98 | 142.12 |
| 180, *src* | 1.25 | 2.21 | 4.59 | 2.87 | 4.85 | 126.91 |
| 180, *dst* | 1.56 | 2.26 | 4.62 | 2.92 | 4.90 | 130.01 |

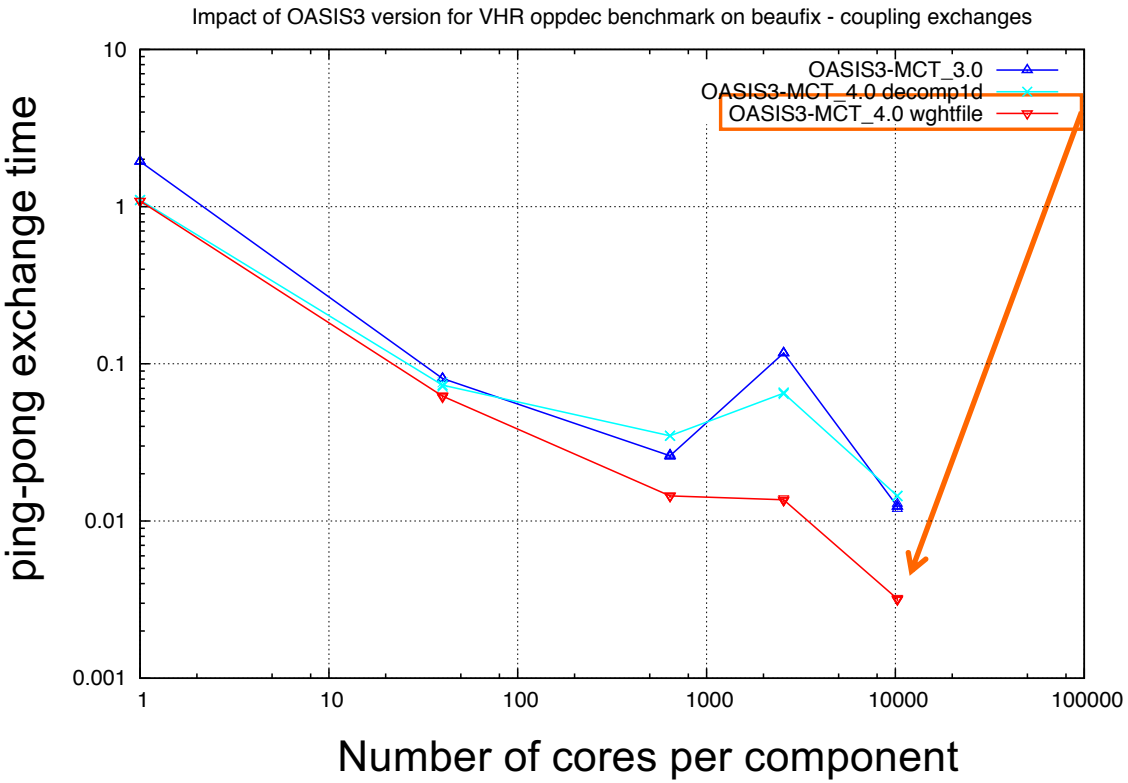
ORCA025 - T799
Cerfacs Lenovo

➢ *lsum8* is the fastest
➢ *reprosum* probably the best choice for bit-for-bit reproducibility as only slightly more expensive than *lsum16*

# Latest developments: OASIS3-MCT_4.0

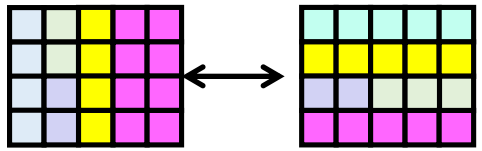
➢ Optimisation of the mapping of the target grid on the source tasks using the mapping weights

- *decomp_1d :* each target grid point is assigned to a source task in a trivial 1-D way (as in OASIS3-MCT_3.0 )
- *decomp_wghtfile:* a target grid point is associated with the source task which holds the source grid points needed for the calculation of its interpolated value




Results

IS-ENES benchmark VHR: 3000x3000
reg lat-lon grids, opposite decompositions

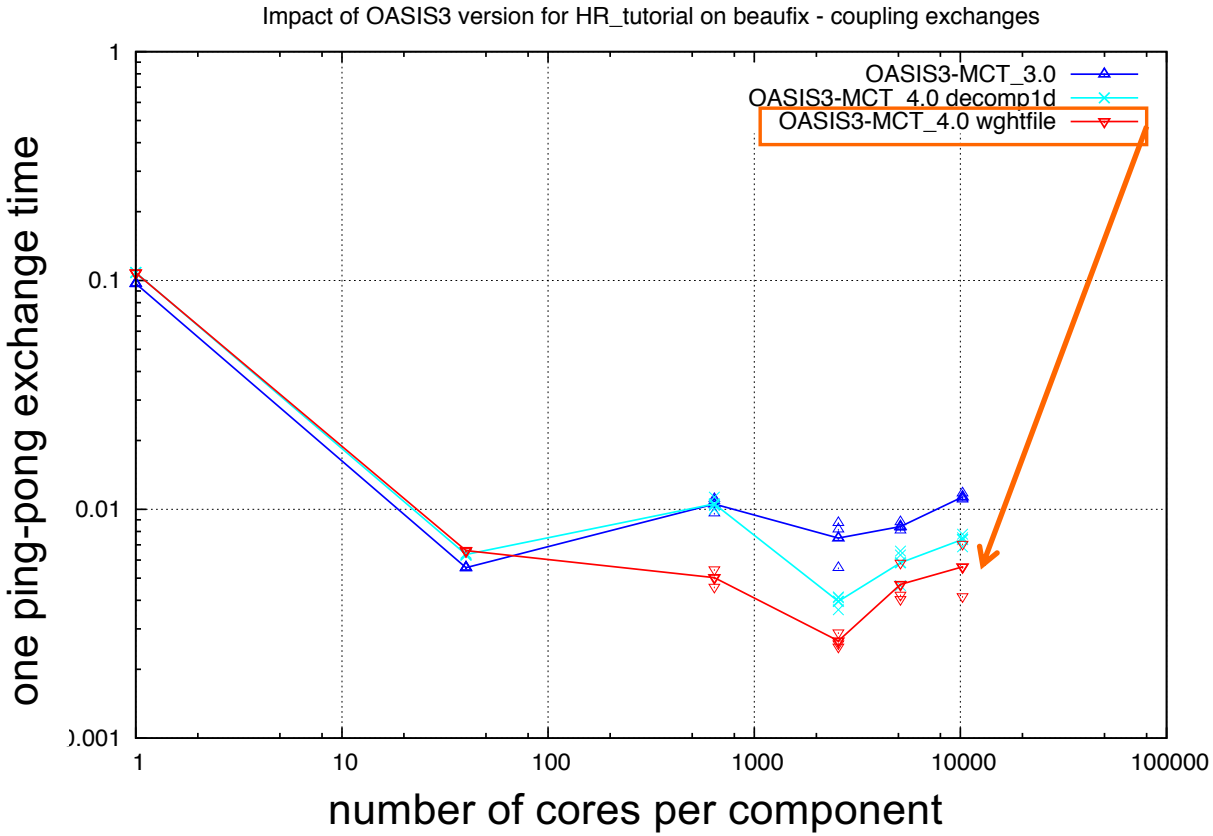


⇒ 75% reduction in exchange time at 10240 cores for that case

# Latest developments: OASIS3-MCT_4.0

➢ Optimisation of the mapping of the target grid on the source tasks using the mapping weights

- *decomp_1d :* each target grid point is assigned to a source task in a trivial 1-D way (as in OASIS3-MCT_3.0 )
- *decomp_wghtfile:* a target grid point is associated with the source task which holds the source grid points needed for the calculation of its interpolated value




Results

NEMO ORCA025 grid (1021x1442) – Gaussian Reduced T799 grid (843 000)

⇒ 24% reduction at 10240 cores for "decomp_wghtfile" wrt "decomp_1D"

# Latest developments: OASIS3-MCT_4.0

## ➢ hybrid MPI+OpenMP parallelisation of the SCRIP library is now implemented

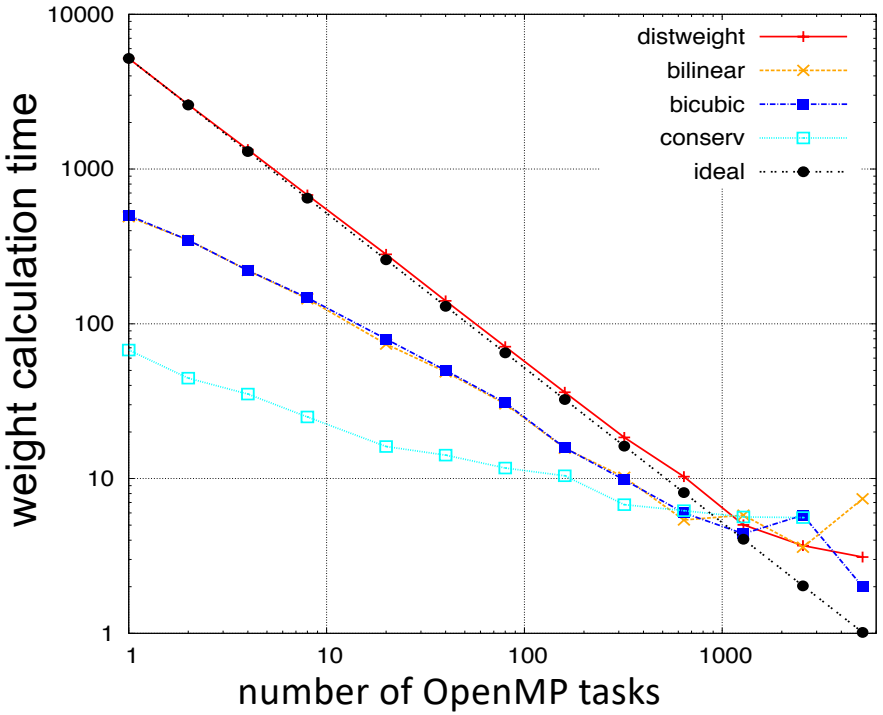
- hybrid MPI+OpenMP parallelisation:
  - one MPI process per node
  - OASIS_OMP_NUM_THREADS OpenMP threads per node (recommended: = number of cores/node)
- Parallelisation
  - over the outer loop on N target grid points for bicubic, distance-weighted and nearest-neighbour
  - over two outer loops over source and target grid cells for mesh border intersection for conservative remapping

**Code optimisation in sequential mode**

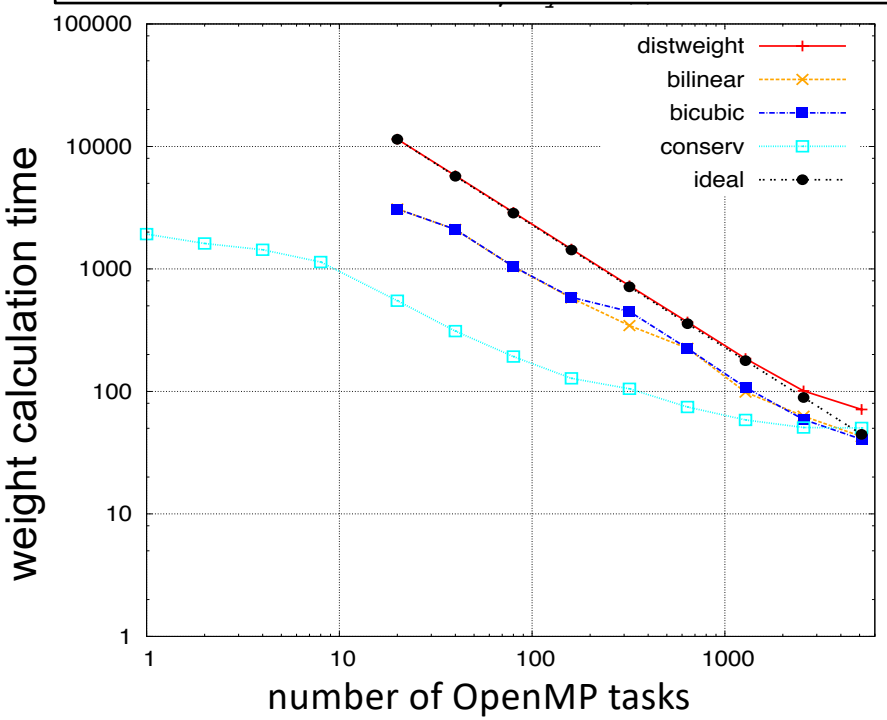
- detection of overlapping points (-DTREAT_OVERLAY): original algorithm $O(n^2)$ -> new algorithm $O(n \log(n))$
  - ➢ orca025 to t359 remapping : 731 sec -> 0.4 sec
- complementary non-masked nearest neighbour for target cells without in any conservative link (FRACNNEI) :
  - ➢ T359 to ORCA025 coupling : 293 sec -> 5.9 seconds

# Latest developments: OASIS3-MCT_4.0

ORCA025 (1442x1050) – T359 (181724)



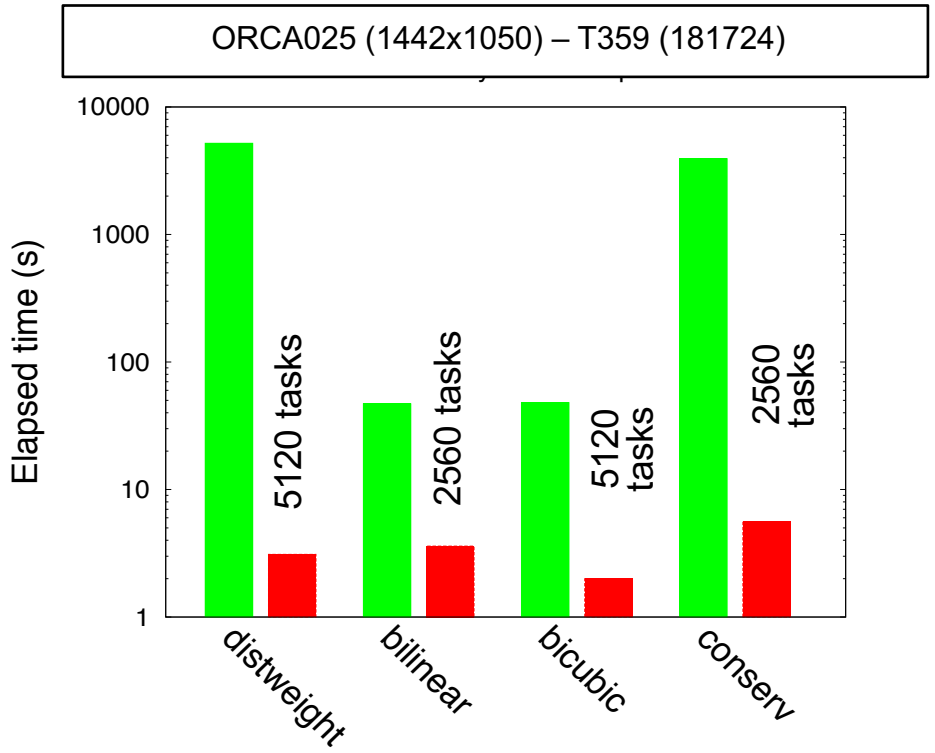

ORCA12 grid (4322x3147) – T799 grid (843490)

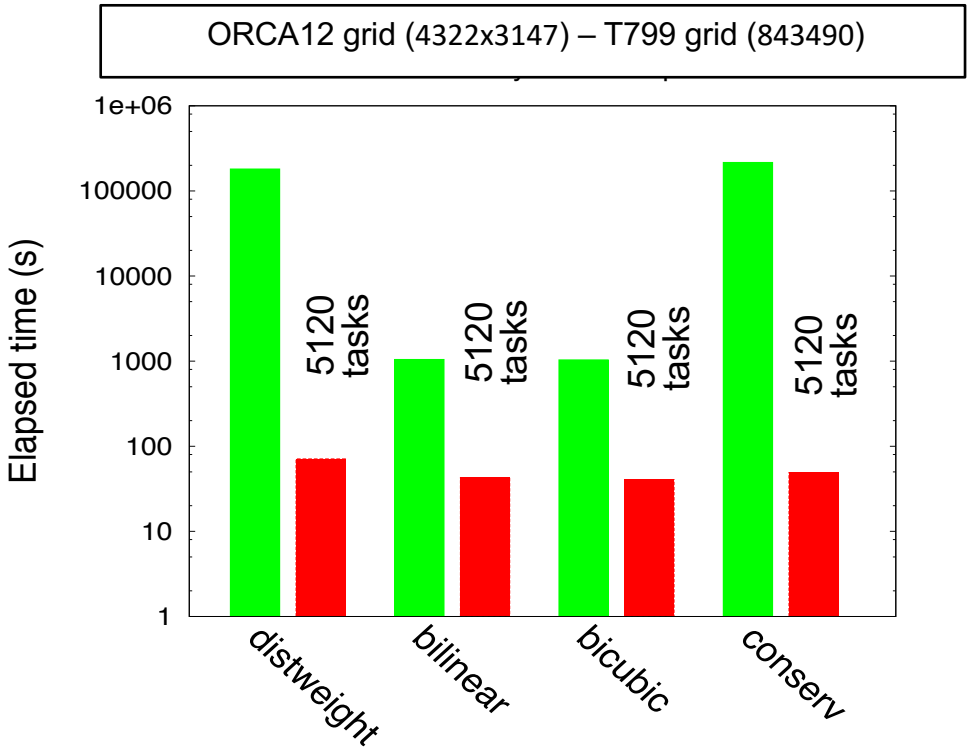



- Almost perfect scalability for nearest-neighbour and bilinear (-> 1280 tasks for HR;-> 2560 tasks for VHR)
- Good scalability for bicubic remapping
- Less scalability for conservative remapping, due to better sequential performance (bin restriction)

# Latest developments: OASIS3-MCT_4.0

OASIS3-MCT_3.0 vs OASIS3-MCT_4.0 MPI+OpenMP hybrid best performance




- reduction in the weight calculation time of 2 or 3 orders of magnitude

# Summary and conclusions

## OASIS3-MCT_4.0 released in July 2018:

- "nointerp" option: bypassing the identity matrix multiplication:
  - OASIS3-MCT as good as other coupling technologies for the IS-ENES2 benchmark VHR test case
- Removal of concurrent writing into the OASIS3-MCT debug files at initialization
  - drastic reduction of the initialisation cost
- Upgrade from MCT 2.8 to MCT 2.10.beta1
  - significant reduction of the initialisation cost
- New global conservation method reprosum:
  - ~bit-for-bit reproducibility, O(10) less costly than previous *bfb* method
- New way to define the intermediate mapping decomposition based on remapping weights (decomp_wghtfile)
  - significant gain at run time
- Hybrid MPI/OpenMP parallel SCRIP library
  - reduction in the weight calculation time of 2 or 3 orders of magnitude for high-resolution grids
  - opens the door to runtime weight computation dynamical coupling with OASIS3-MCT

## Planned within IS-ENES3/ESiWACE2:

- Analysis of other interpolation libraries : ESMF, XIOS, YAC, CWIPI, MOAB/TemestRemap
- First steps toward dynamic calculation of weights during the simulation


Thank you for your attention!