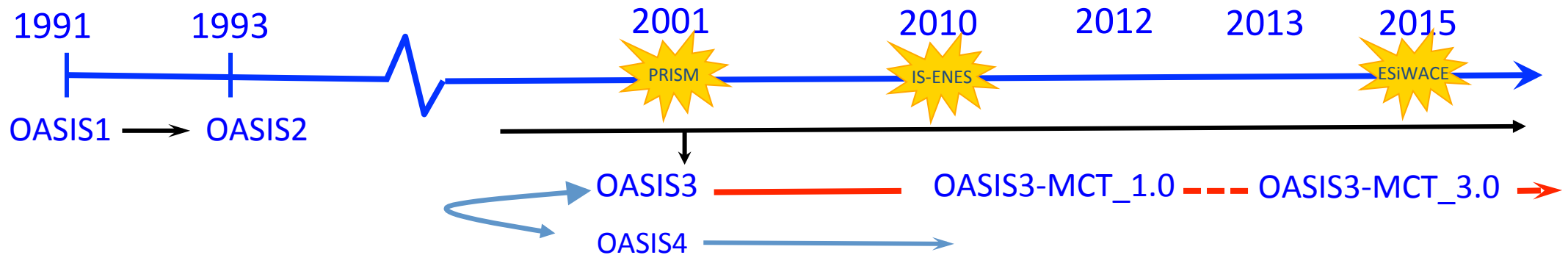


Latest developments of the OASIS3-MCT coupler for improved performance

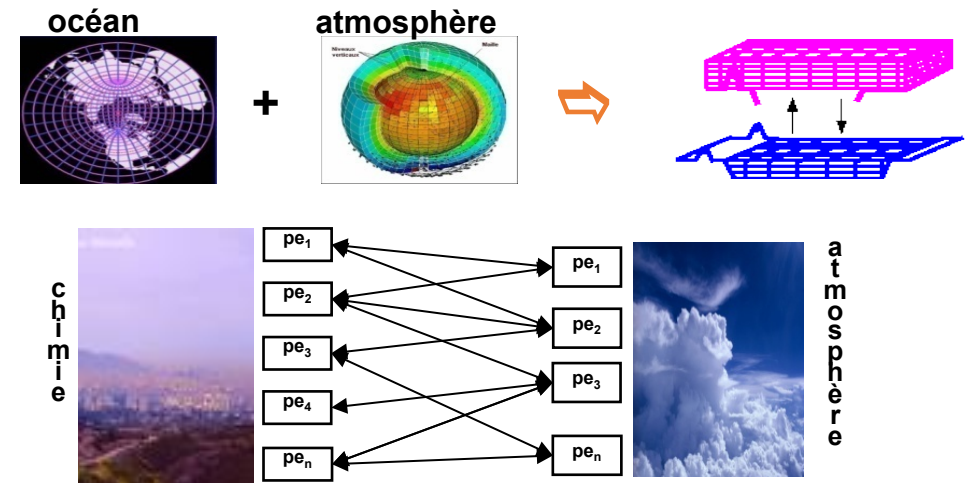
S. Valcke, L. Coquart, A. Craig, G. Jonville, E. Maisonnave, A. Piacentini

The OASIS coupler - Introduction



• 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**



➤ F90 & C, LGPL licence LGPL, public domain libraries (MPI, NetCDF, SCRIP, MCT)



The OASIS coupler - Introduction

More than 40 groups in France, Europe and all over the world use OASIS for climate modelling and seasonal prediction, 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
 - Ireland: 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
 - Korea: 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 Nigeria, Colombia, Singapore, Russia, Thailand, ...

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

The OASIS coupler - Introduction

OASIS3-MCT_4.0 released soon!

Few additional functionalities:

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

Optimisation and bugfixes

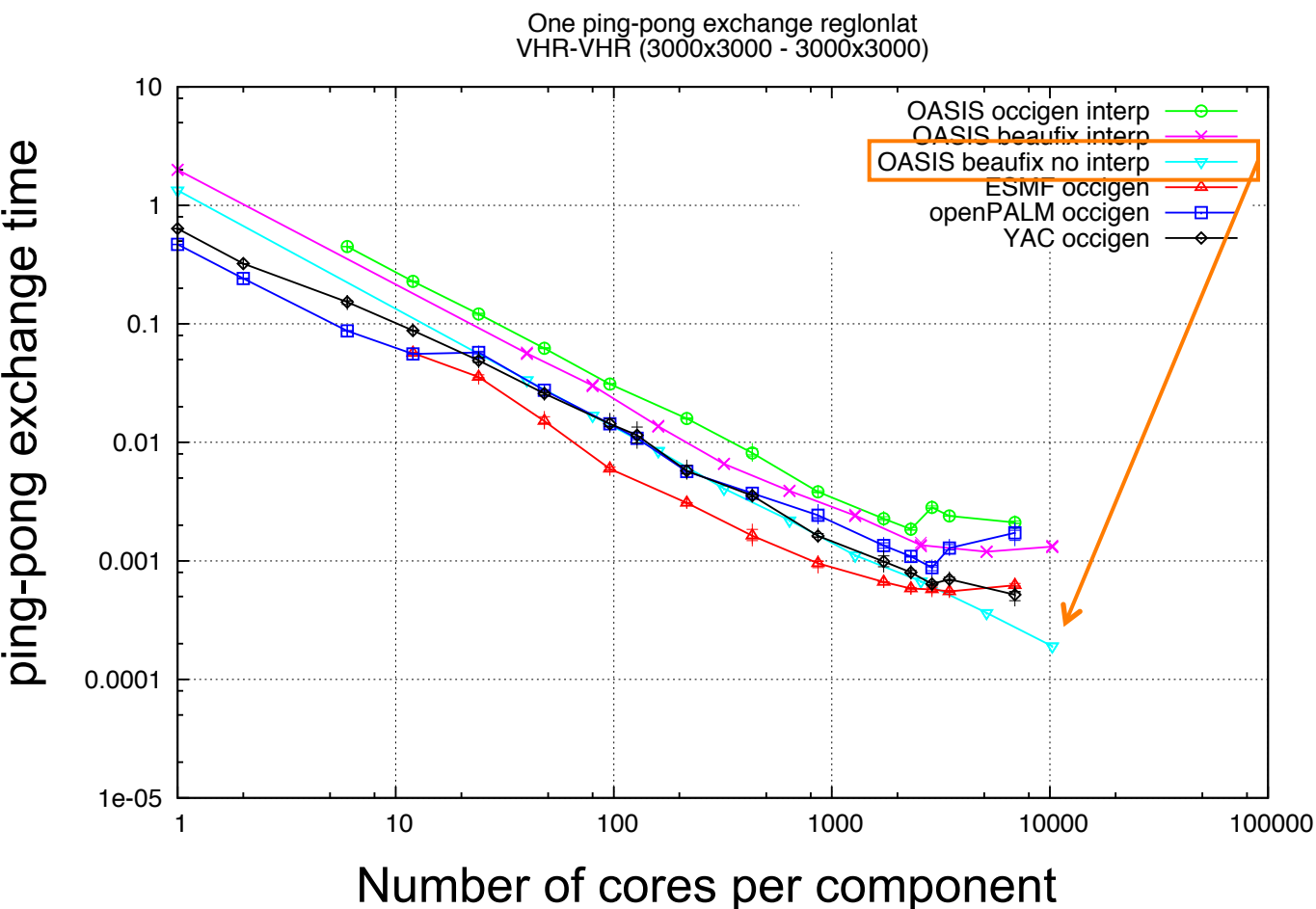
- Activation of « nointerp » for identical grids – impact on IS-ENES2 benchmarks
- New more performant algorithms for the global CONSERV operation
- Upgrade of MCT library
- Debugging of the coupling initialisation
- Optimisation of the communication using the mapping weights
- Hybrid MPI+OpenMP parallelisation of the SCRIP library

Publications:

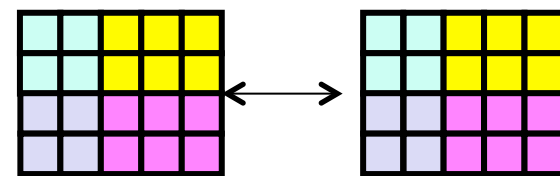
- S. Valcke, L. Coquart, A. Craig, G. Jonville, E. Maisonnave, A. Piacentini, ESiWACE D2.3 “Multithreaded or thread safe OASIS version including performance optimizations to adapt to many-core architectures”
- A. Craig, S. Valcke, L. Coquart, 2017: Development and performance of a new version of the OASIS coupler, OASIS3-MCT_3.0, Geosci. Model Dev., 10, 3297-3308

Activation of "nointerp" for identical grids - impact on IS-ENES2 benchmark

➤ "nointerp" : bypass matrix-vector multiplication for identical grids (identity matrix)



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



New algorithms for the global CONSERV operation

CONSERV: forces the global conservation of the coupling field before and after the remapping

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

OASIS3-MCT_4.0:

- *gather* : as *bfb*
- *lsum8*: as *opt*
- *lsum16* : as *opt/lsum8* but uses quadruple precision
 - more costly but higher chance of reproducibility than *opt/lsum8*
- *reprosum*: fixed point method based on ordered double integer sums(Mirin &Worley, 2004)
 - expected to produce bit-for-bit results except in extremely rare cases
- *ddpdd*: parallel double-double algorithm using a single scalar reduction (He &Ding, 2004)

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

ORCA025 -
Cerfacs Ler

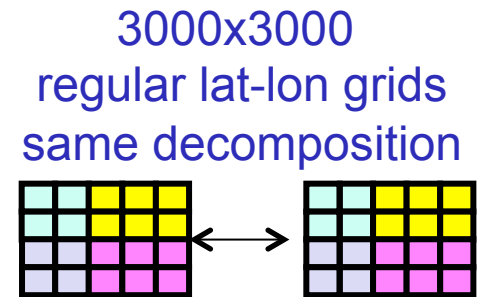
- *lsum8* is the fastest
- *reprosum* probably the best choice for bit-for-bit reproducibility as only slightly more expensive than *lsum16*

Upgrade of MCT library

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:



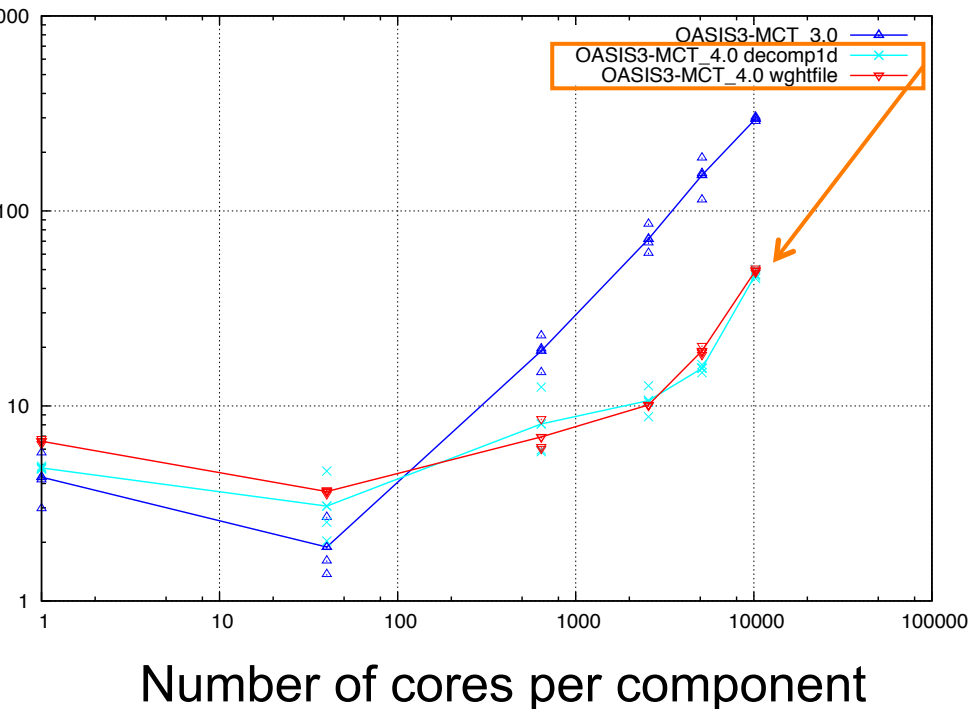
- cost to compute the router between the source and mapping decomposition :
 - 19 sec -> 0.5 sec on 1600 tasks/component
 - 41 sec -> 0.7 sec on 3600 tasks/component
- cost to compute the router between the mapping and target decomposition :
 - 60 sec -> 6-7 sec on 1600 tasks/component
 - 124 sec ->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

Debugging of the coupling initialisation

bugfix: removal of concurrent writing into the OASIS3-MCT debug files at initialization

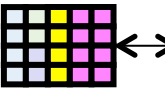
NEMO ORCA025 grid (1021x1442) –
Gaussian Reduced T799 grid (843 000)

Impact of OASIS3 version for HR_tutorial on beaufix - initialisation

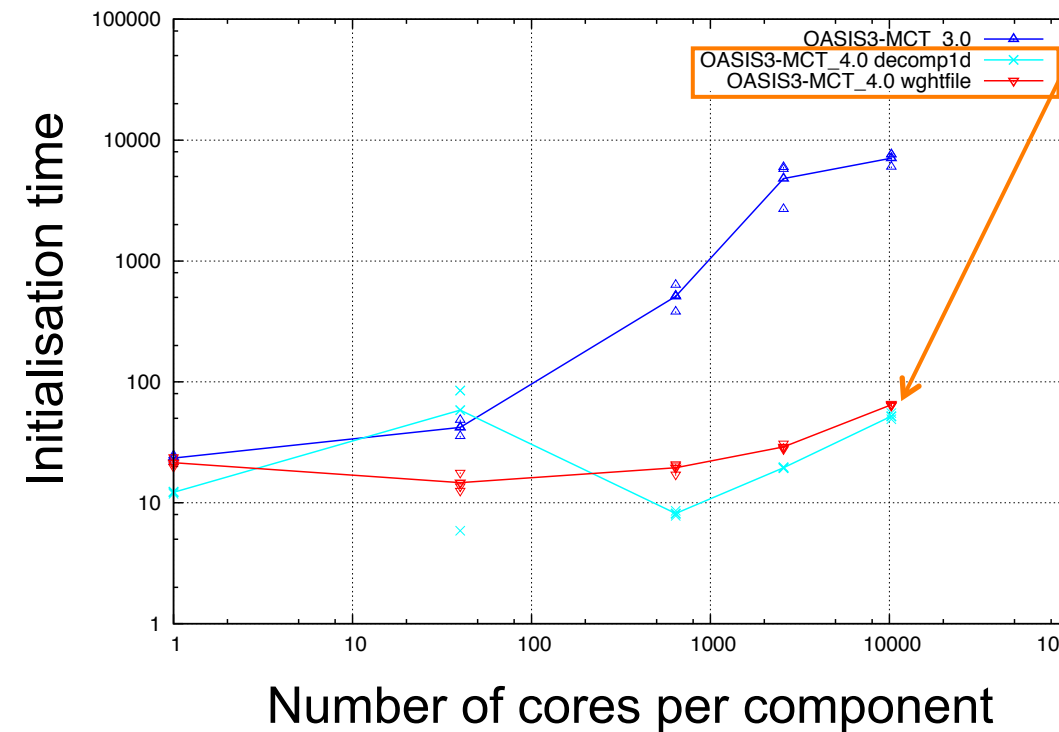


⇒ 82% reduction in init time at 10240 cores

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



Impact of OASIS3 version for VHR oppdec benchmark on beaufix - initialisation



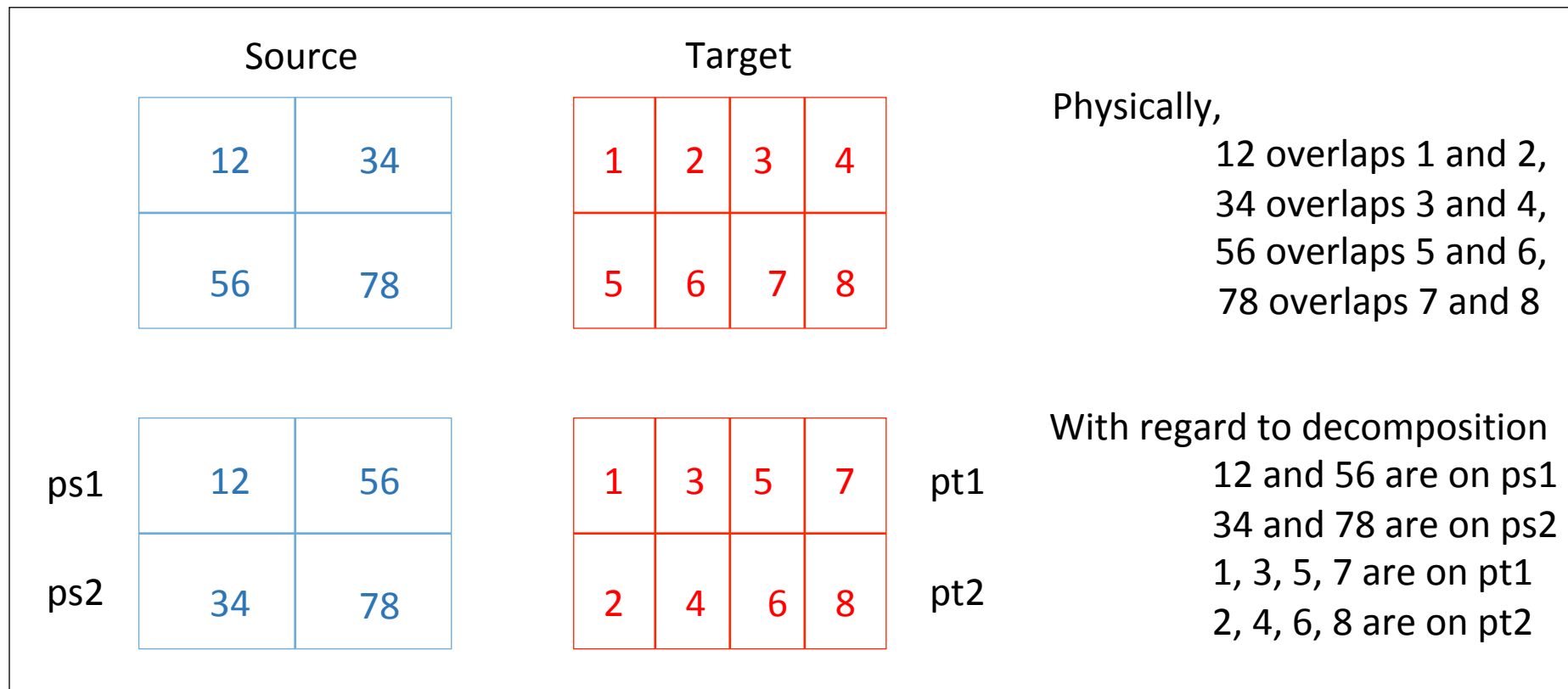
⇒ 99% reduction in init time at 10240 cores

Optimisation of the communication using the mapping weights

to perform the remapping, a “mapping decomposition” of the target grid on the source tasks is created:

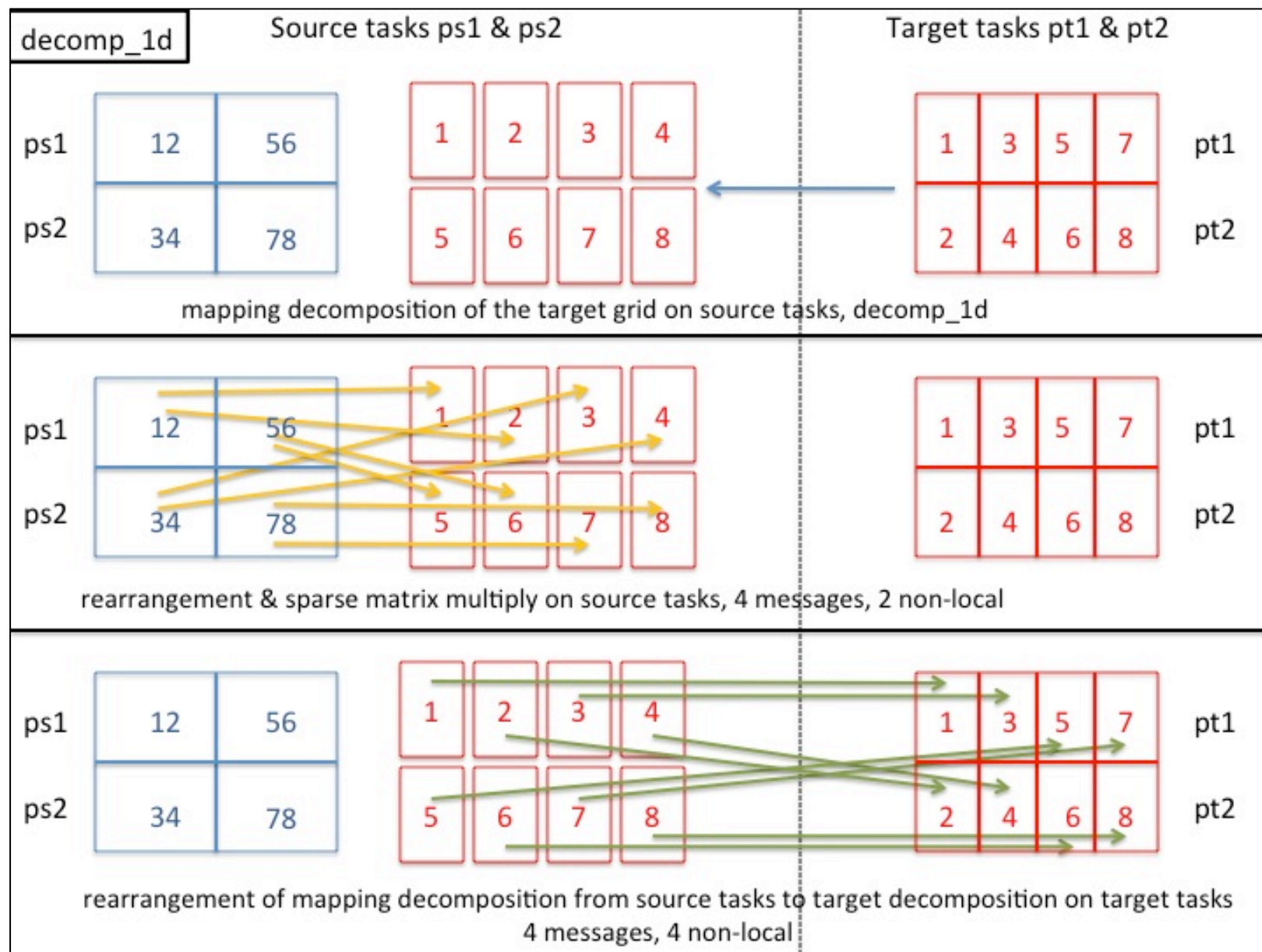
decomp_1d : each target grid point is assigned to a source task in a trivial 1-D way (as in OASIS3-MCT_3.0)

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



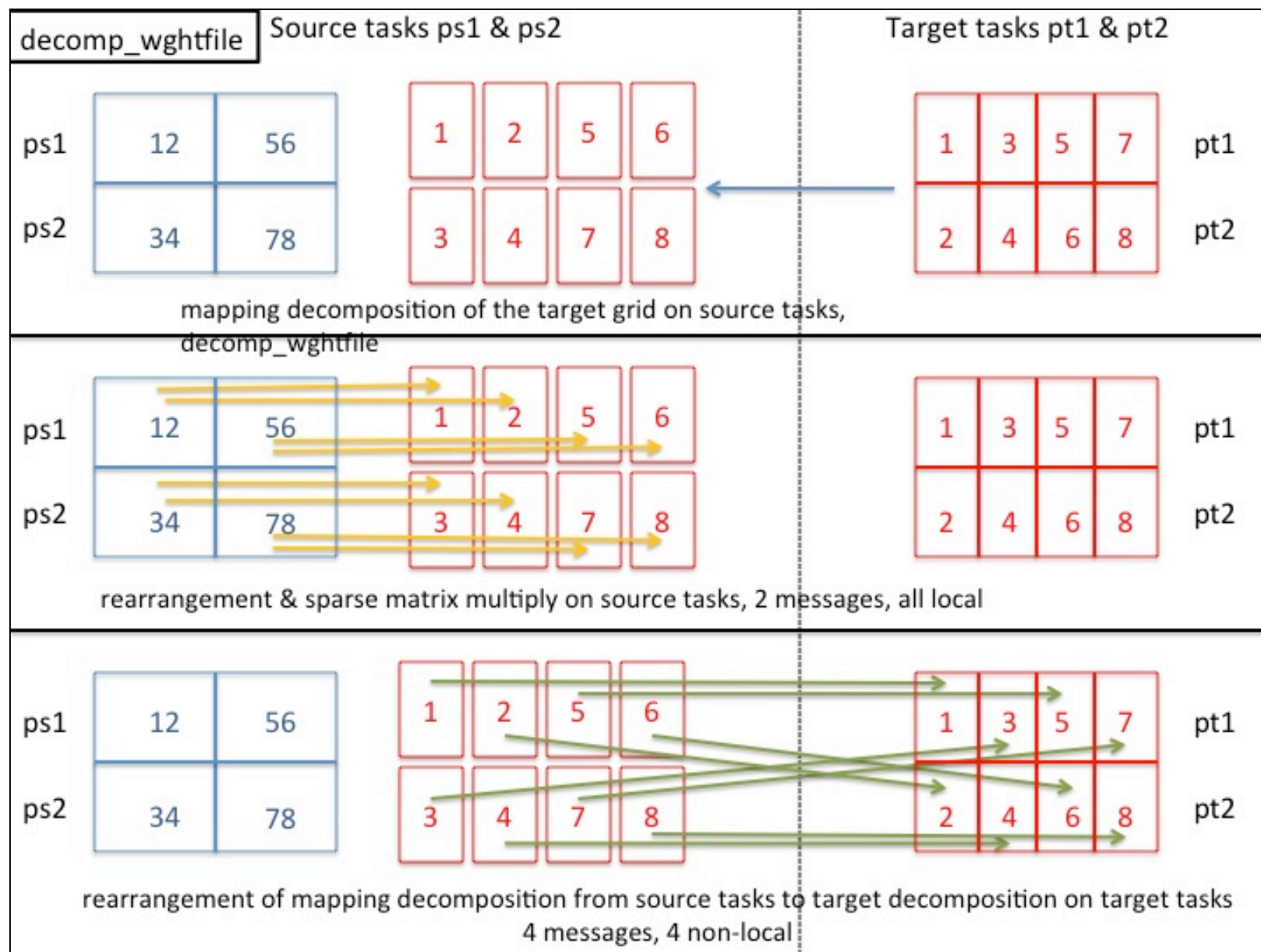
Optimisation of the communication using the mapping weights

decomp_1d



Optimisation of the communication using the mapping weights

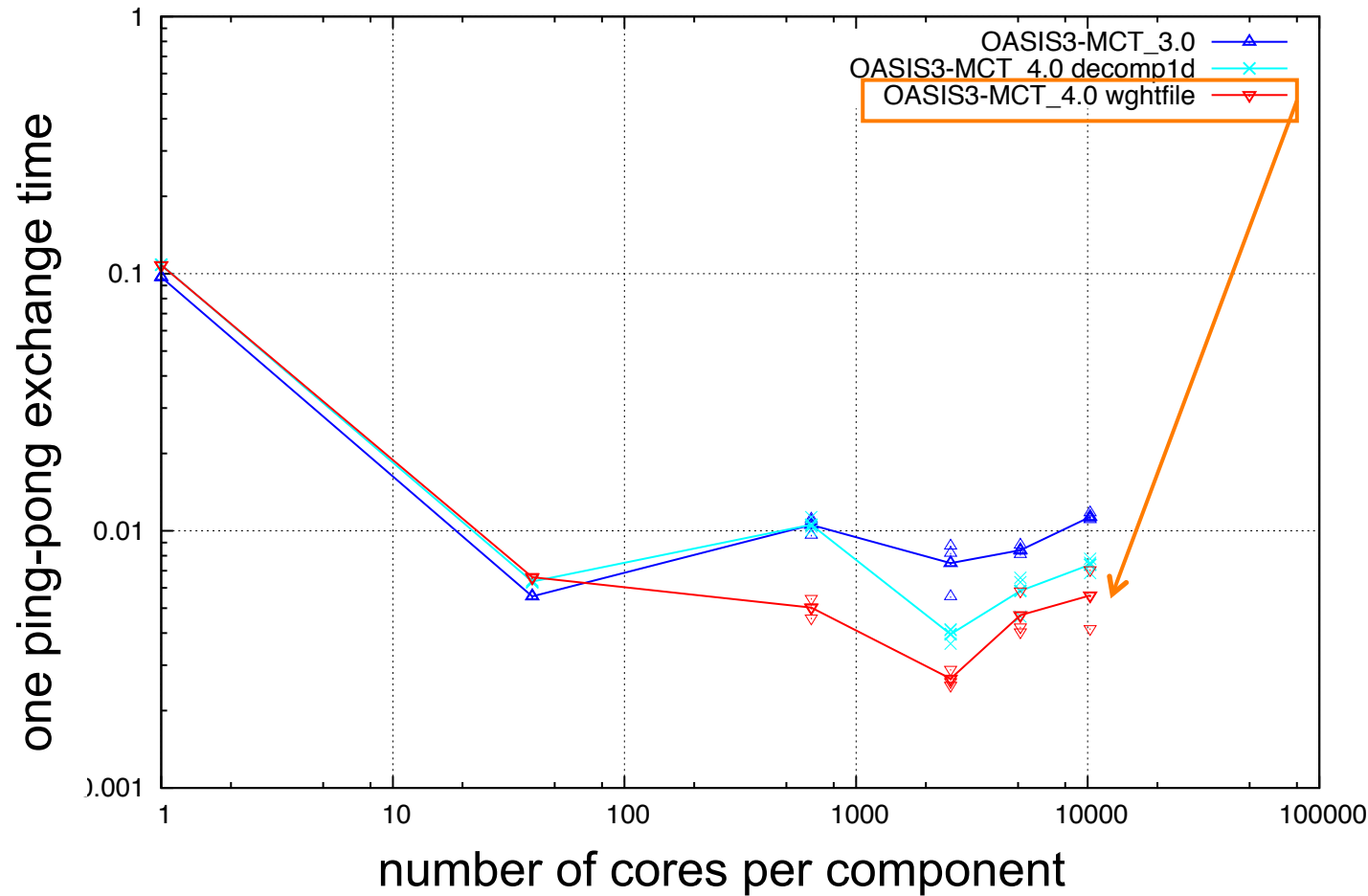
comp_wghtfile



Optimisation of the communication using the mapping weights

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

Impact of OASIS3 version for HR_tutorial on beaufix - coupling exchanges



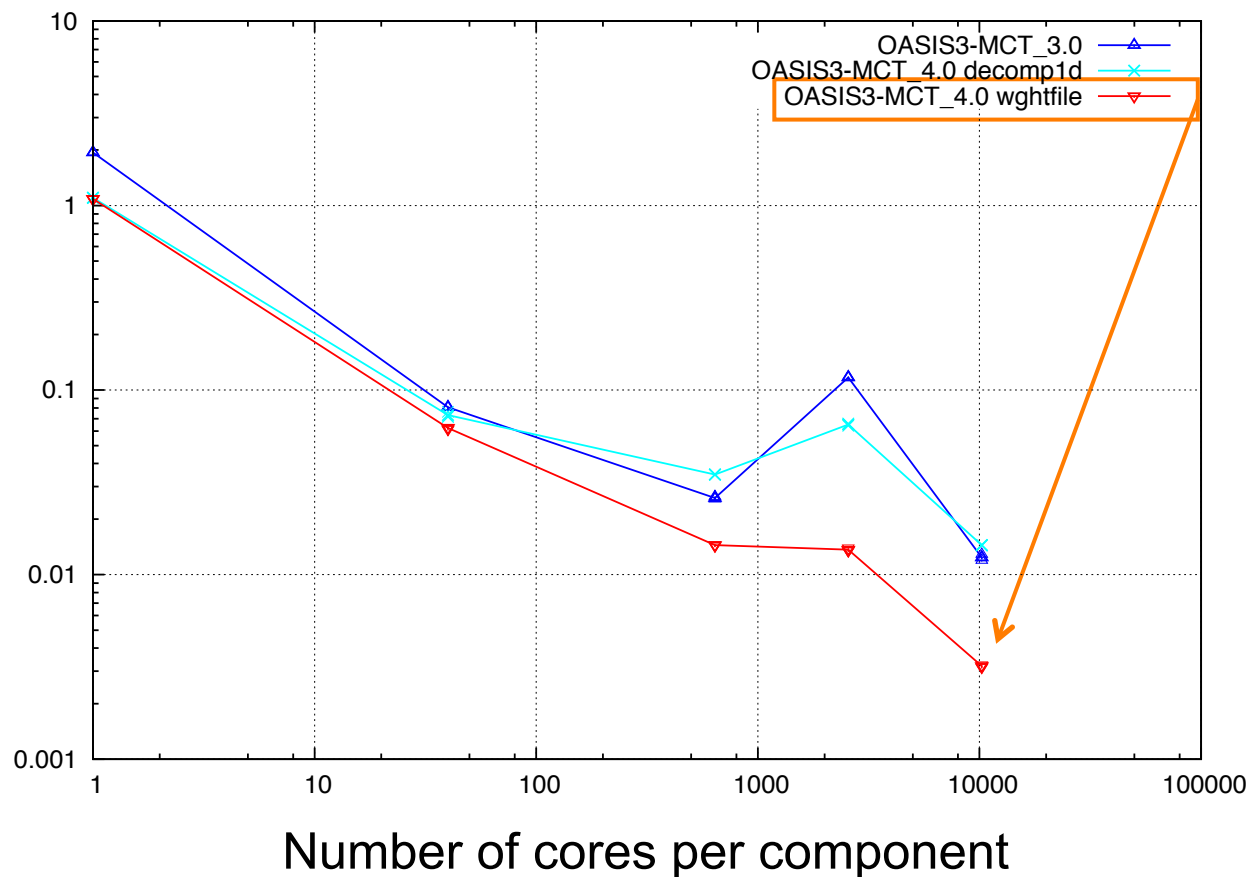
Results

⇒ 24% reduction at 10240 cores
“decomp_wghtfile” wrt “decomp_”
which is already 35% faster th
OASIS3-MCT_3.0

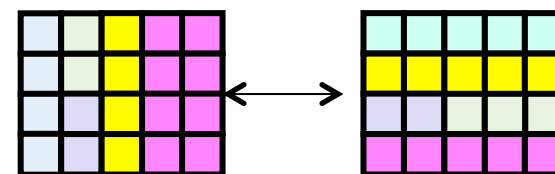
Optimisation of the communication using the mapping weights

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

Impact of OASIS3 version for VHR oppdec benchmark on beaufix - coupling exchanges



Results



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

Hybrid MPI+OpenMP parallelisation of the SCRIP library

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

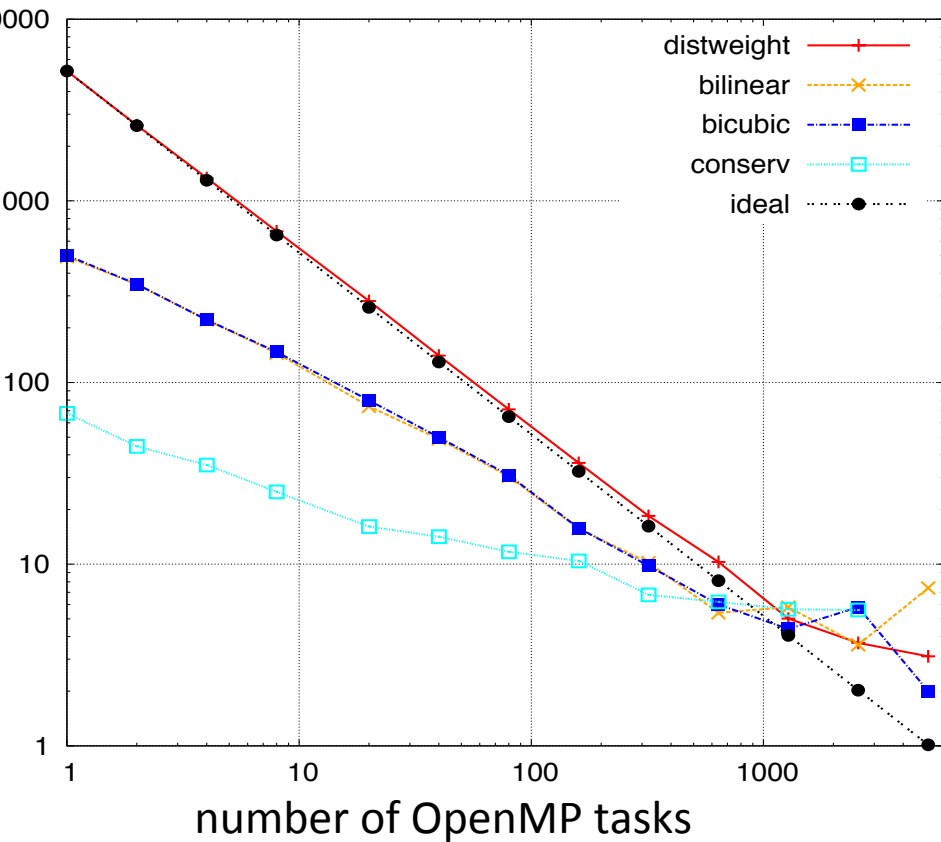
- hybrid MPI+OpenMP parallelisation:
 - one MPI process per node
 - OASIS_OMP_NUM_THREADS OpenMP threads per node (recommended: = number of cores/node)
- Parallelisation
 - over the outer loop on N target grid points for bicubic, distance-weighted and nearest-neighbour
 - over two outer loops over source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

de optimisation in sequential mode

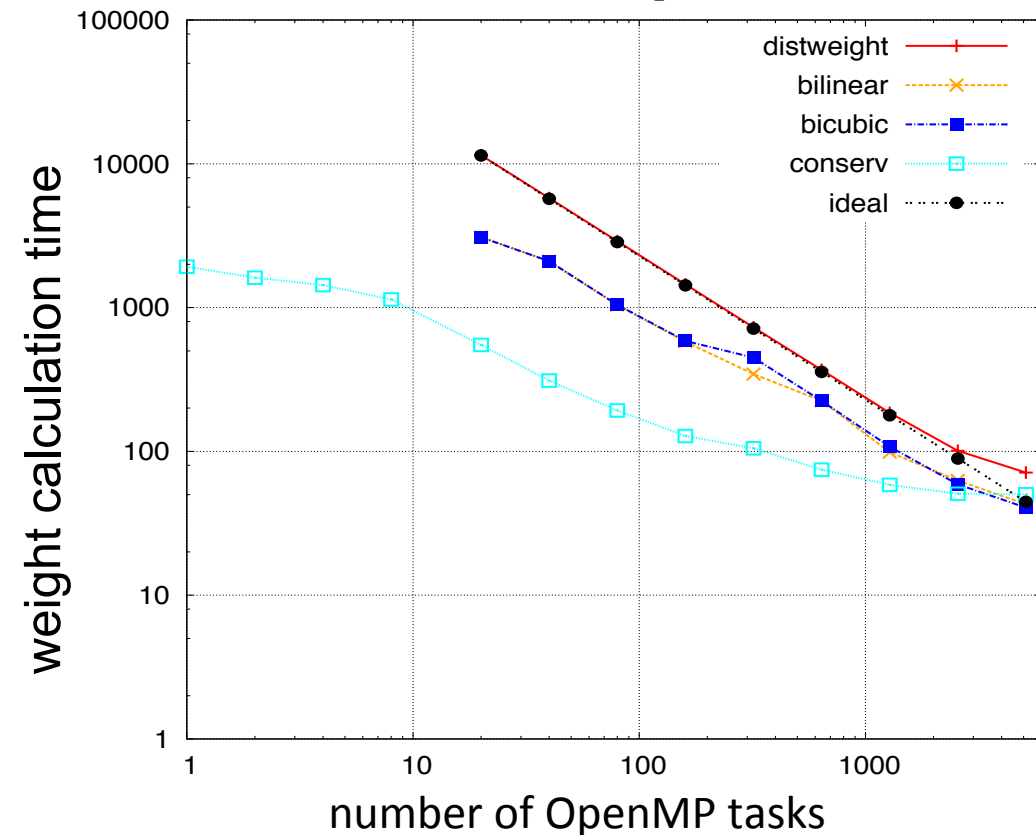
- detection of overlapping points (-DTREAT_OVERLAY): original algorithm $O(n^2)$ -> new algorithm $O(n \log(n))$
 - orca025 to t359 remapping : 731 sec -> 0.4 sec
- complementary non-masked nearest neighbour for target cells without in any conservative link (FRACNNEI) :
 - T359 to ORCA025 coupling :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

Hybrid MPI+OpenMP parallelisation of the SCRIP library

ORCA025 (1442x1050) – T359 (181724)



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

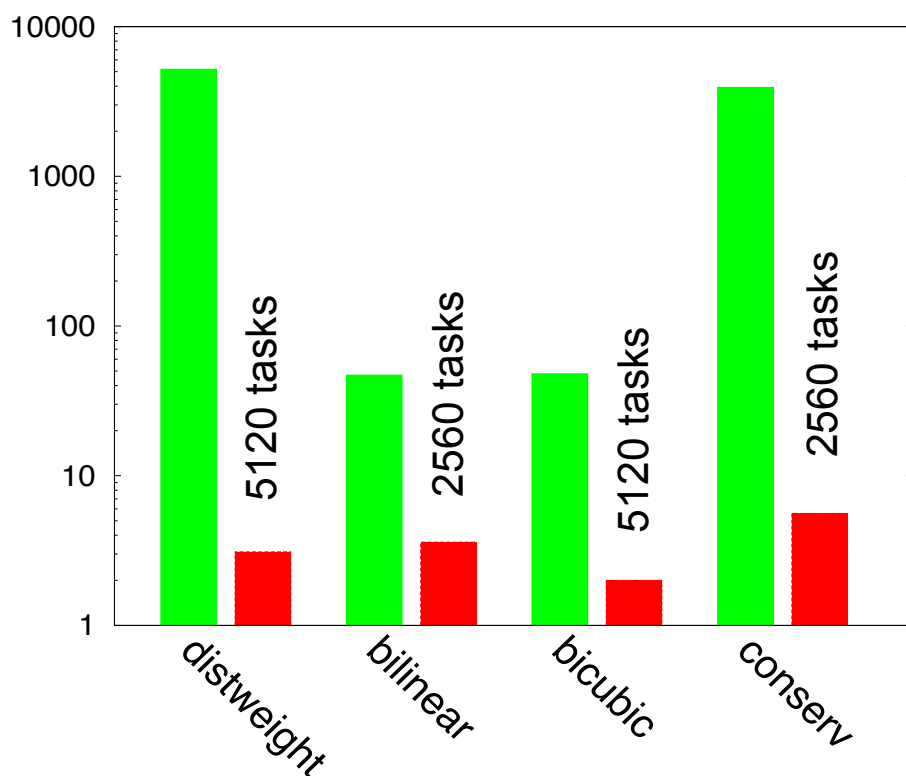


- Almost perfect scalability for nearest-neighbour and bilinear (-> 1280 tasks for HR;-> 2560 tasks for V)
- Good scalability for bicubic remapping
- Less scalability for conservative remapping, due to better sequential performance (bin restriction)

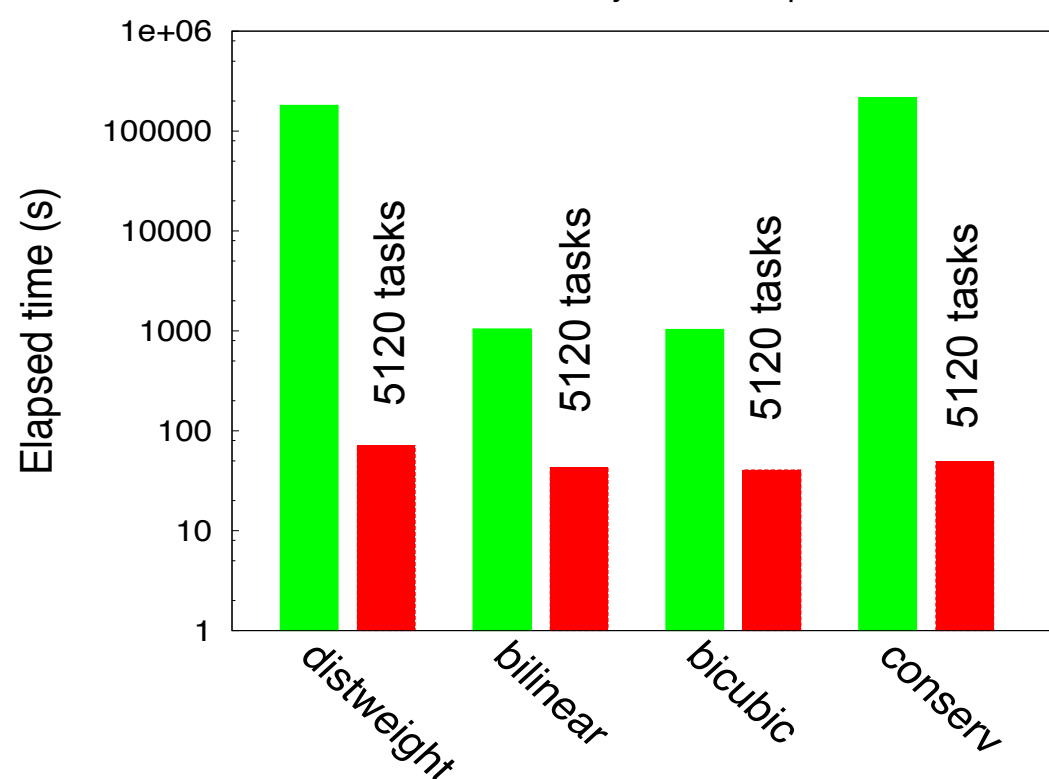
Hybrid MPI+OpenMP parallelisation of the SCRIP library

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

ORCA025 (1442x1050) – T359 (181724)



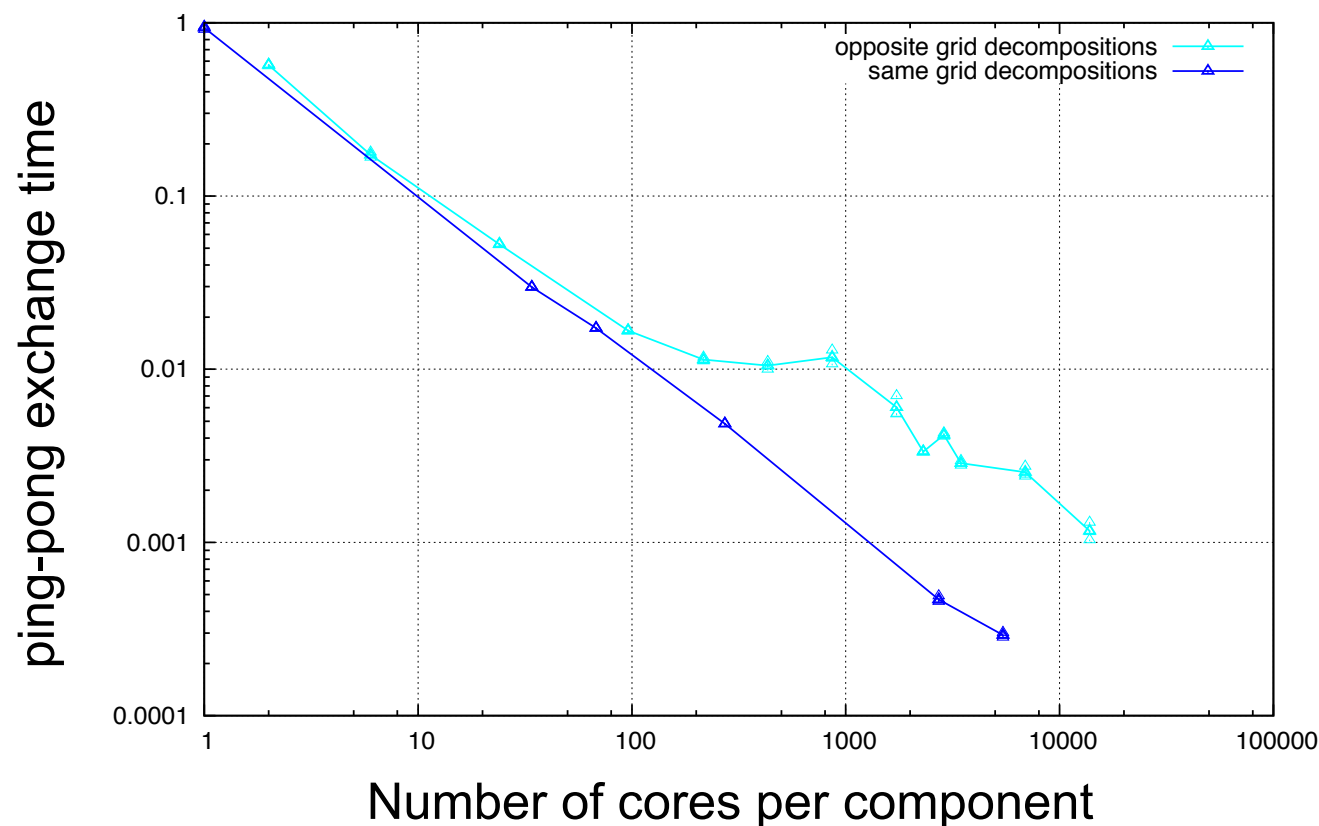
ORCA12 grid (4322x3147) – T799 grid (843490)



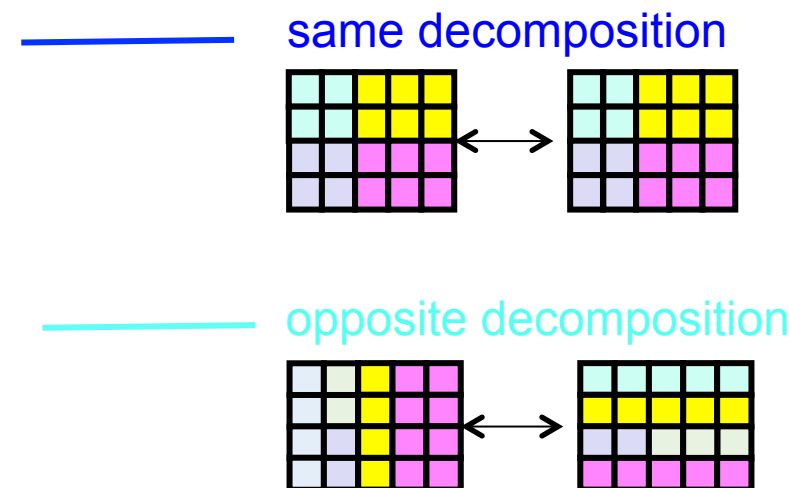
➤ reduction in the weight calculation time of 2 or 3 orders of magnitude

IS-ENES2 coupling benchmarks on Marconi KNL

performance of IS-ENES2 coupling benchmark with OASIS3-MCT_3.0 on Marconi KNL
allocation of 390 000 core hours granted to ESiWACE CoE



VHR: 3000x3000 reg lat-lon grid



- almost perfect scalability for up to $O(10^4)$ cores/tasks per component for VHR
- very reasonable behaviour for the VHR_oppdec;
- would probably be better with OASIS3-MCT_4.0 decomp_wghtfile

From OASIS3-MCT_3.0 to OASIS3-MCT_4.0

OASIS3-MCT_4.0 released before the summer with many performance improvements:

“nointerp” option: bypassing the identity matrix multiplication:

- OASIS3-MCT as good as other coupling technologies for the IS-ENES2 benchmark VHR test case

New global conservation method *reprosum*:

- ~bit-for-bit reproducibility, $O(10)$ less costly than previous *bfb* method

Upgrade from MCT 2.8 to MCT 2.10.beta1

- significant reduction of the initialisation cost

Bugfix : removal of concurrent writing into the OASIS3-MCT debug files at initialization

- drastic reduction of the initialisation cost

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

Thank you for your attention !