NEMO: improving computational performance

7th ENES HPC Workshop – Barcelona – May, 9-11 2022

7th ENES HPC Workshop
9-11 May 2022

NEMO: computational performance community

NEMO improvements

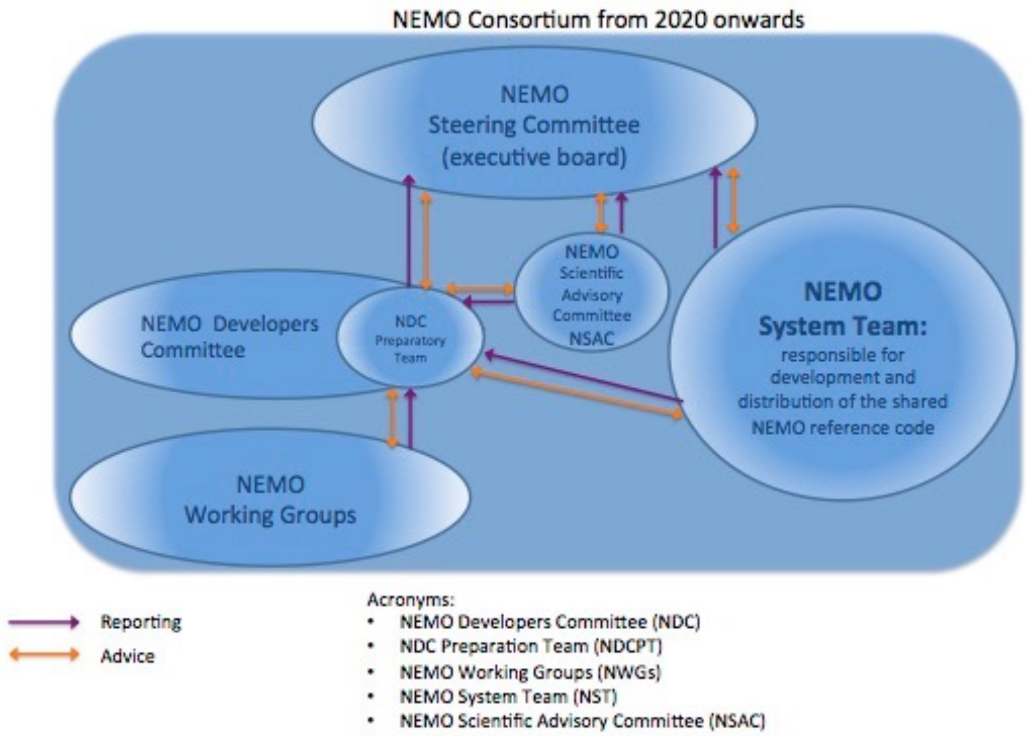
- Single core performance
  - Tiling
  - Loop fusion
  - Mixed precision
- Communication
  - Neighborhood collective communications
- Macro task parallelization
- Multigrid refinement optimization
- I/O
  - Improving read/write with XIOS
  - Online diagnostics
- Support for different architectures
  - GPU
  - Domain Specific Languages

NEMO Consortium organization

- 2003 Building NEMO as European platform
- 2008 NEMO Consortium is formally build
- Consortium members:
  - CNRS-INSU
  - Mercator Ocean International
  - Met Office
  - Natural Environment Research Council NERC-NOC
  - Euro-Mediterranean Center on Climate Change, Foundation - CMCC

NEMO Consortium organization

- NEMO System Team (NST) is responsible for development and distribution of the NEMO reference code
  - New actions are defined in the annual WorkPlan
- NEMO Working Groups articulate and coordinate the exploration of options for development of the NEMO reference code
  - **NEMO HPC-WG** aims at evaluating solutions to improve the computational performance of the NEMO code.



Loop fusion and Tiling

- Efficient exploitation of memory hierarchies and hardware peak performance
- **Loop fusion technique** aims at better exploiting the cache memory by fusing DO loops together

```
DO j=1, n-1
   DO i=1, n
      b (i,j) = in(i,j+1) - in(i,j)
   END DO
END DO

DO j=2, n-1
   DO i=1, n
      out (i,j) = b(i,j) - b(i,j-1)
   END DO
END DO
```

```
DO j=2, n-1; DO i=1, n
      b_0  = in(i,j+1) - in(i,j  ) ! correspond to b(i,j)
      b_m1 = in(i,j  ) - in(i,j-1) ! correspond to b(i,j-1)

      out(i,j) = b_0 - b_m1
END DO; END DO
```

Loop fusion and Tiling

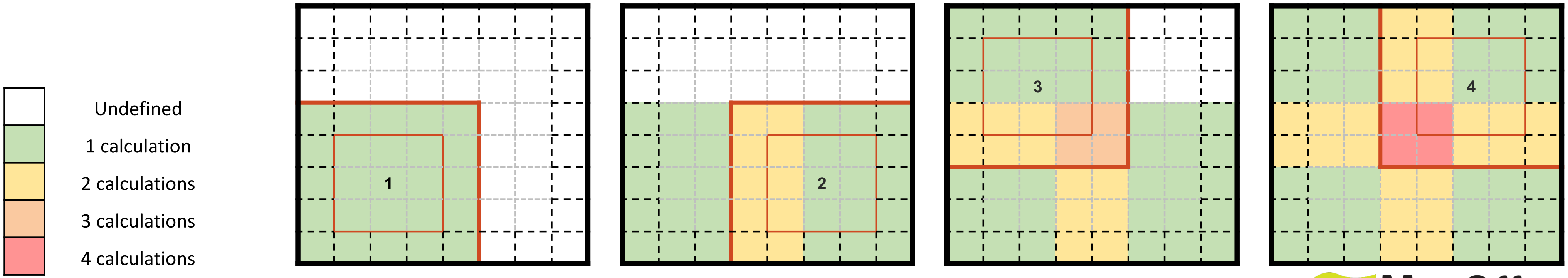
- Tiling allows us to divide the calculation into chunks of work that can remain cache-resident for as long as possible.
- The technique leaves the tile size and shape as tunable parameters, which can be tuned appropriately for cache sizes on any platform.
- Preliminary tests established that the CPU time taken by some typical 3D routines within NEMO using current configurations could be reduced by at least a factor of 2 by 3D tiling

Halo calculations & tiling

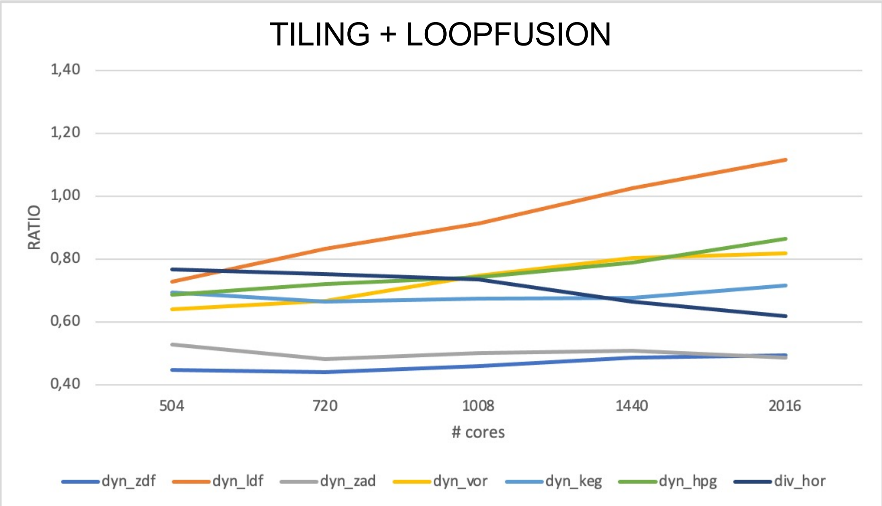
- For a 2D loop over an MPI domain with internal size (*X*, *Y*), halo width *H* and tile size (*x*, *y*), the total number of loop iterations *N* scales as:

$$\frac{N}{XY} = 1 + 2\,\mathrm{H}\,\frac{x+y}{xy} + \frac{4H^2}{xy}$$

- Local working arrays: not preserved in memory, so must calculate all points on a tile
  - Calculations depend on tile and halo size
- Module / allocatable arrays: preserved in memory, so no need to repeat calculations done by other tiles
  - Calculations depend only on halo size



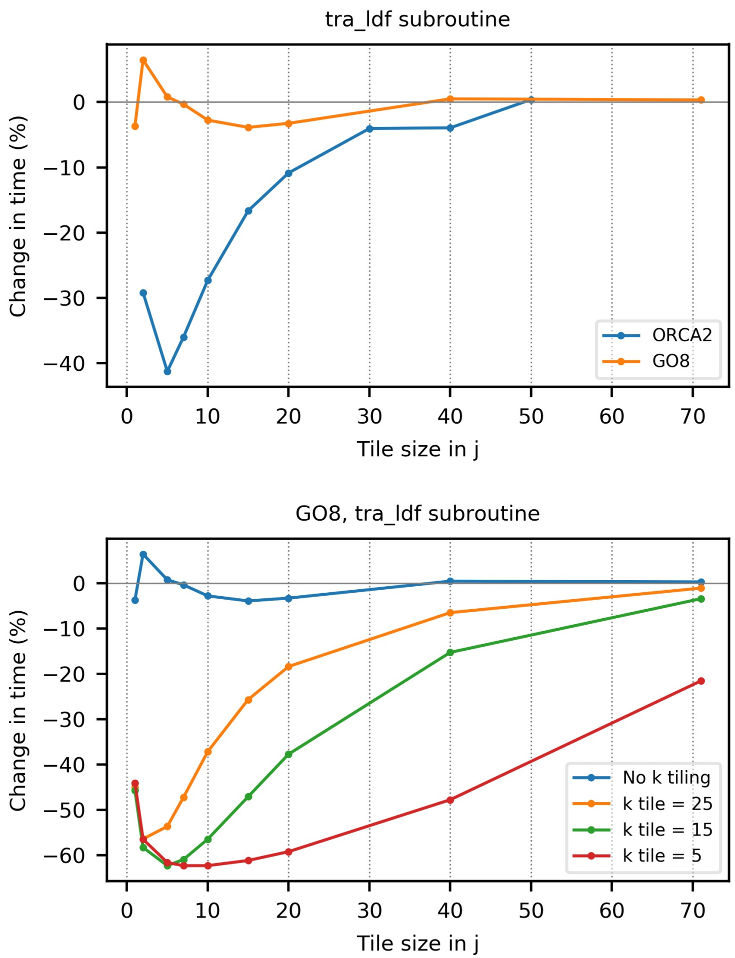
Loop fusion and Tiling



The ratio of the optimized code w.r.t. the baseline is reported changing the number of cores for the key routines of ocean dynamics. Ratio < 1 is good

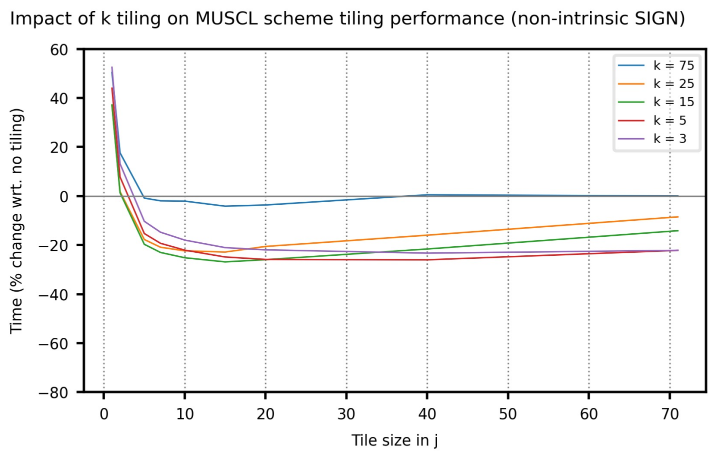
- LoopFusion and Tiling applied only to the Ocean Dynamics and Ocean Tracer
- On average a speedup of 1.4x can be achieved
- The impacts of this optimization strongly depends by the platform and by the configuration

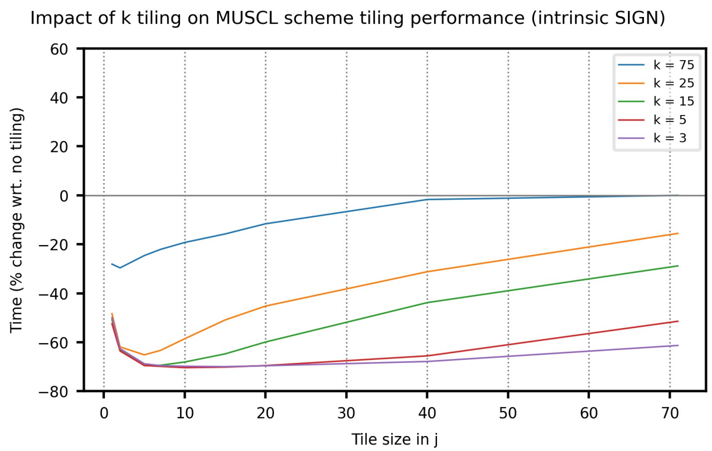
Vertical tiling



- `tra_ldf` scales poorly in GO8 compared to ORCA2
- GO8 domain ~6x larger, optimal tile size ~6x smaller
  - `56i x 75j x 75k` (GO8)
  - `34i x 54j x 31k` (ORCA2)
- Tiling only in the horizontal is not sufficient
  - We must also tile in the vertical for optimal and consistent performance
- However, this has its own unique challenges
  - No existing vertical partitioning to leverage
  - Tridiagonal solvers

Vectorisation and tiling





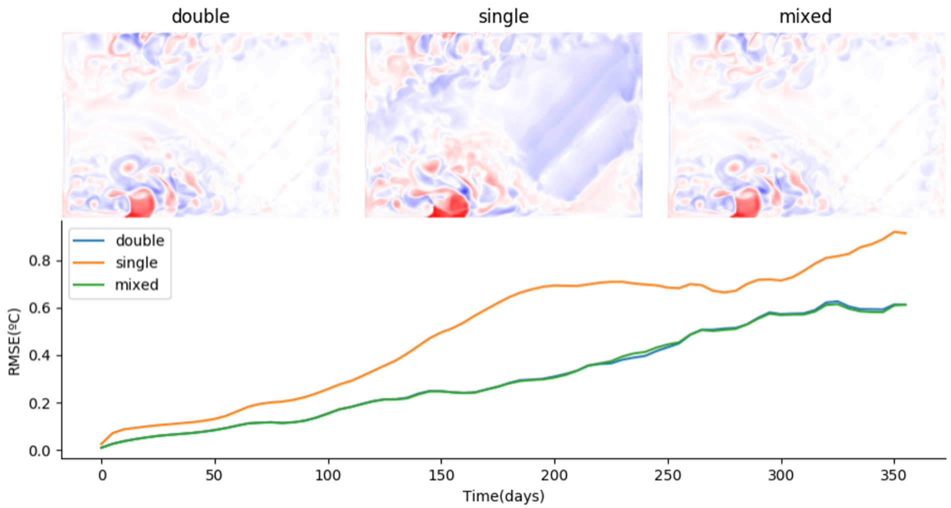
- The MUSCL scheme (`tra_adv_mus`) uses a non-intrinsic `SIGN` function (activated with `key_nosignedzero`)
- Loops containing this function cannot be vectorised- there are many of these in MUSCL
- Using the intrinsic `SIGN` results in better vectorisation coverage and performance
- Tiling performance is also much better- up to 70% faster compared to 25%

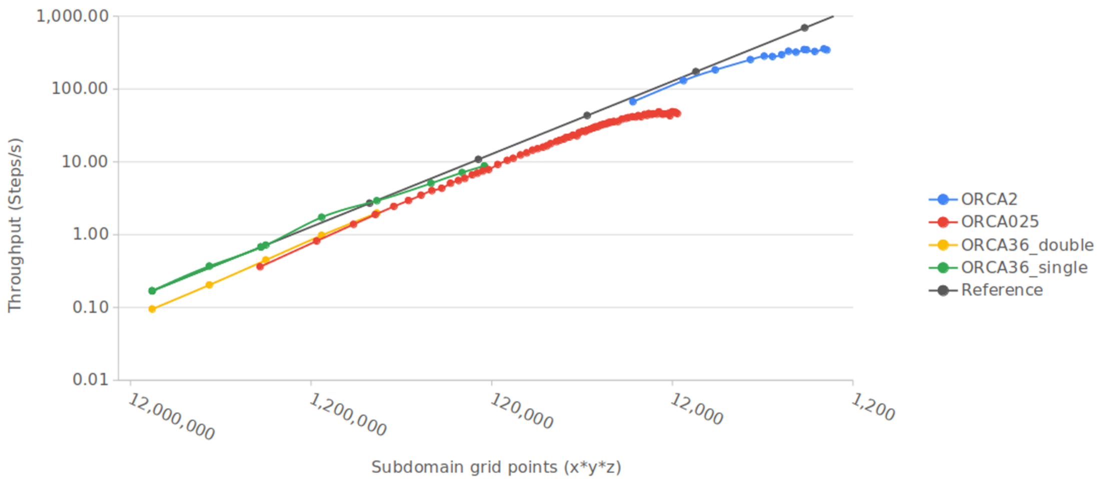
Mixed Precision

- Advantages
  - Reduce the memory footprint
  - Improve the aritmetic intensity which measures the ratio between the number of operations executed and the amount of data moved from main memory to CPU.
  - Reduce the computational cost due to the use of single precision operations
  - Considerably improve the parallel scalability
- With an appropriate tuning of the variabiles in SP vs those in DP, the results accuracy of the mixed precision version is preserved

Mixed Precision

Impact of precision on sea-surface temperature in NEMO4:
comparison of GYRE1/9º simulations using different precisions

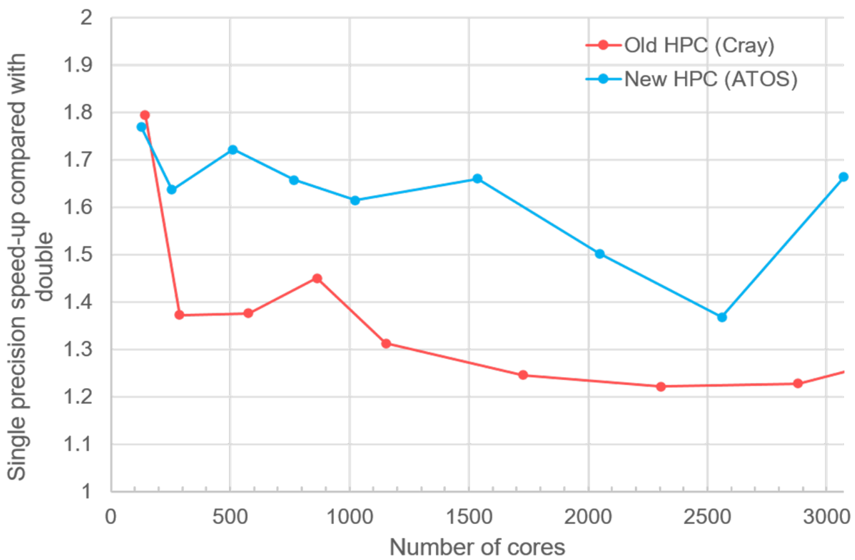




Mixed-precision approaches can provide performance benefits while keeping the accuracy of the results.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

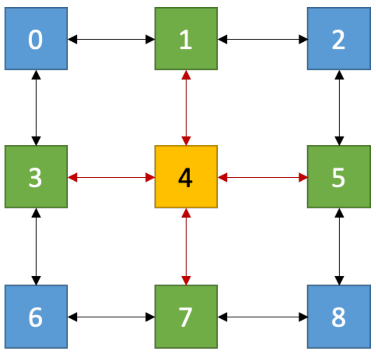
Single Precision at ECMWF

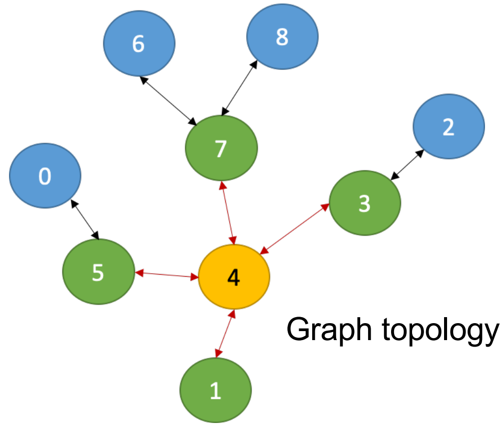


- Fully single-precision coupled atmosphere-wave-ocean forecasts now possible, **including NEMO**
- Tested with eORCA1 ocean and compared with operational reference (DP NEMO) in extended range forecasts
- Mostly skill neutral change
- Speed-up from using single precision in NEMO measured on old (Cray) and new (ATOS) HPC at ECMWF
- Final speed-up depends on I/O server → integration of NEMO with ECMWF I/O server MultIO underway

MPI Communication Neighborhood collectives

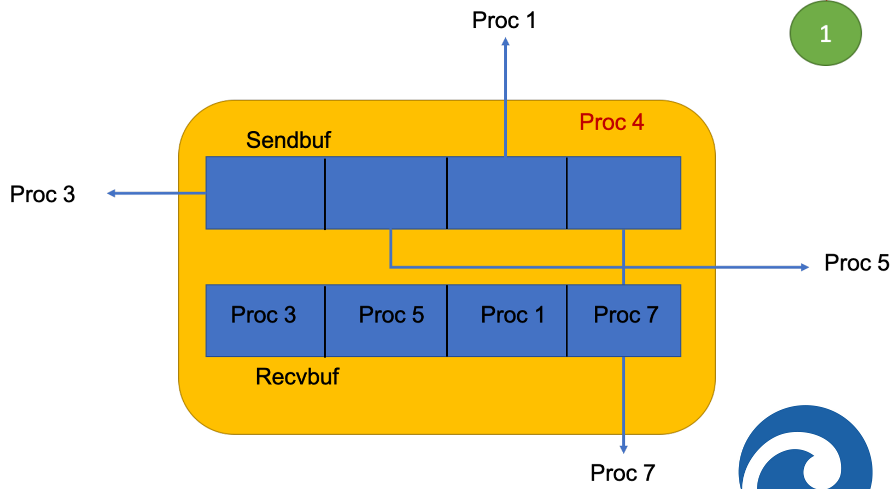
Cartesian topology

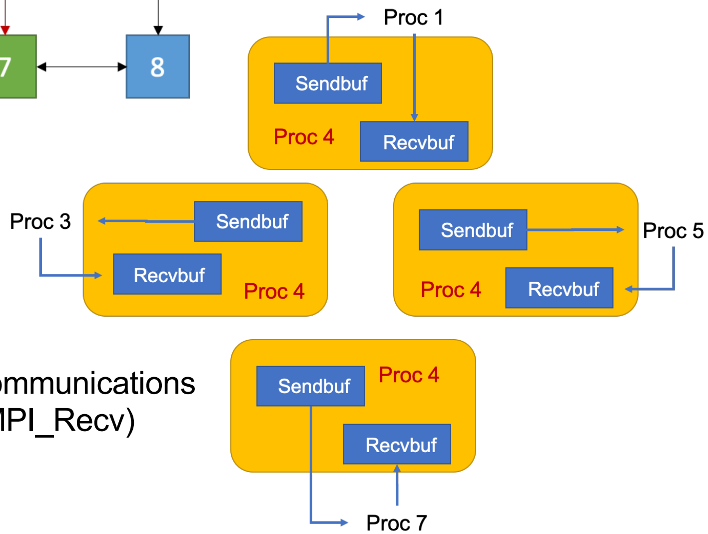




Graph topology

1 collective communication
(MPI_Neighbor_alltoall)



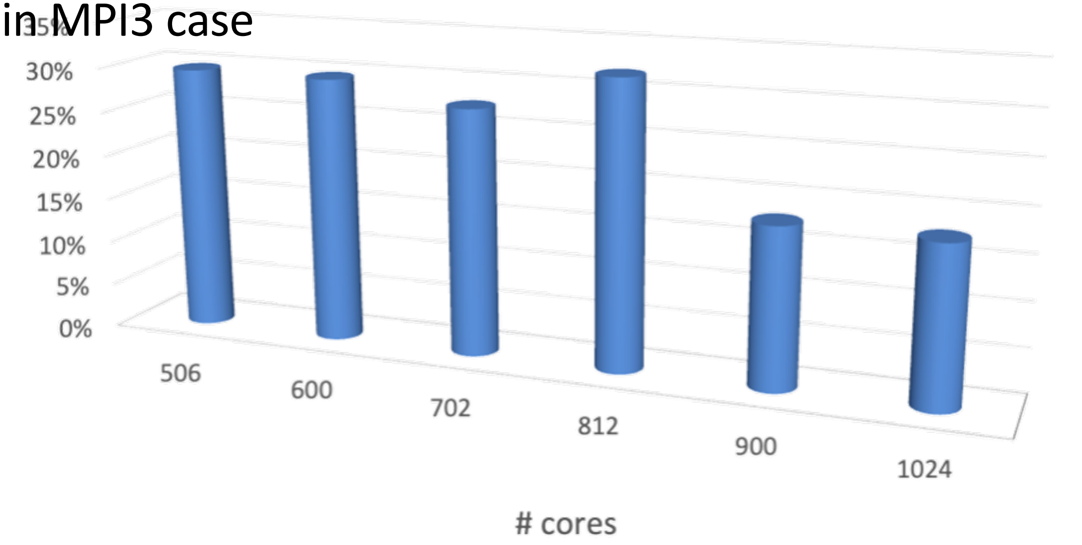


4 Point-2-Point communications
(MPI_Send/MPI_Recv)

MPI Communication Neighborhood collectives

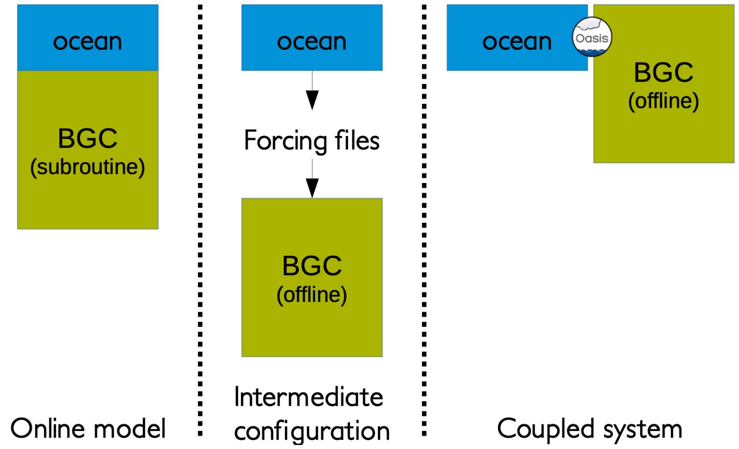
- Extension of LBC (Lateral Boundaries Condition) module to support MPI3 Neighborhood Collectives:
  - New Cartesian communicator
  - Ranks reordering to match NEMO processes order
  - Data buffer handling
  - Implementation of multi field exchange in MPI3 case

- Test on the advection scheme
  - GYRE_PISCES configuration (nn_GYRE=200 → ~6000x4000x31 grid resolution)
  - Communication time improved within a range of 18%-32%



Macro Task Parallelism

- Parallelize OPA (ocean module) and TOP-PISCES (tracer advection biogeochemistry -BGC- module) into two executables and ensure 3D coupled fields exchange via the community coupler OASIS.
- The ocean-BGC coupled model exhibits an improvement of computing performance when the subdomain decomposition leads to computations/communications ratio that put the performance just below the scalability limit
- The coupling cost, caused by OASIS coupling extra cost and load imbalance between components is non negligible (around 20% in our case) but can be reduced
- This contrasted result suggests that the only clear performance gain can only be ensured with the radical cost lowering of the most time consuming component, the BGC model (coarsening)



Sciences de l'environnement
Institut Pierre Simon Laplace

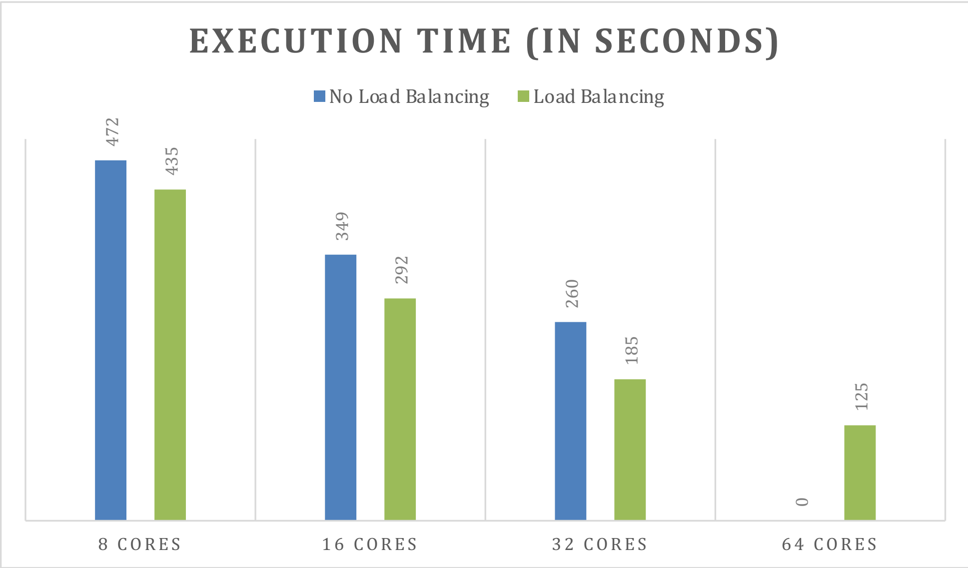
I/O optimization through XIOS

- Improvement on I/O reading initial conditions and reading regridding weights using XIOS
- the XIOS support has also been adopted for reading and writing of the restart files in the SI3 (sea ice model).
- New XIOS version is going to be released with a relevant expected improvements

Met Office

Multigrid capability

- The support for nested multigrid in NEMO is implemented in the AGRIF component
- NEMO model has been updated to provide an estimation of the computational cost of each cell grid; a new load balancing policy has been implemented in AGRIF
- Achieved 1.4x speedup on average

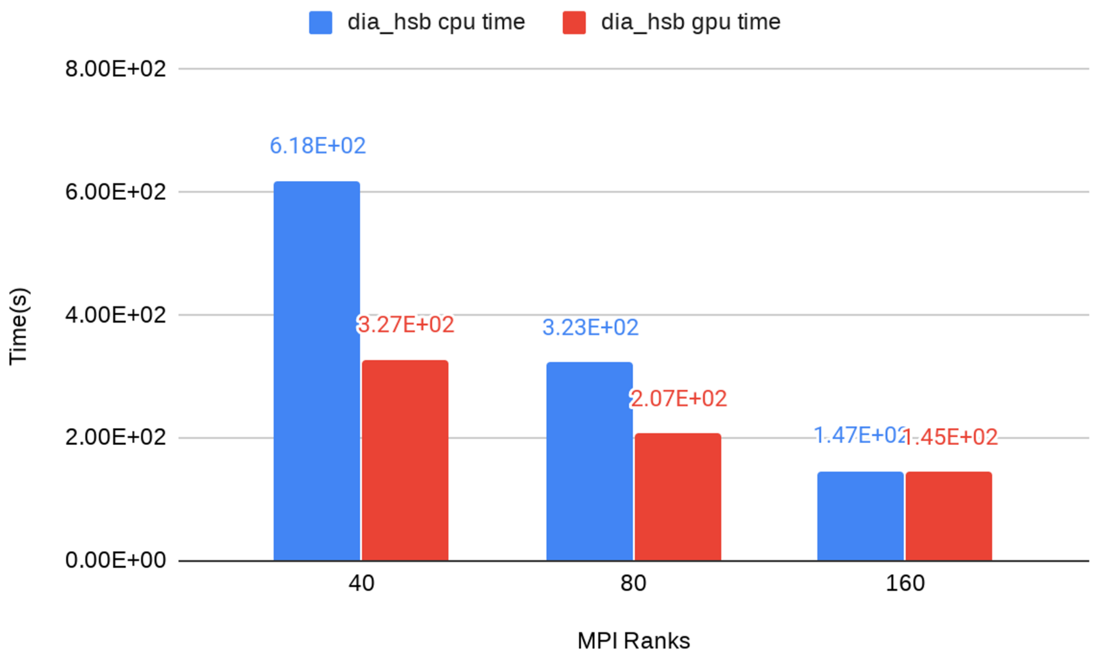


Online diagnostics – GPU based

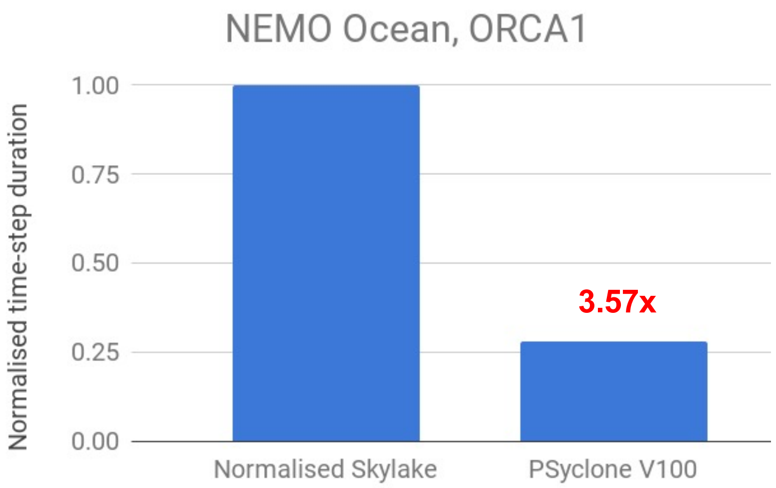
- The rationale of this activity is to improve the NEMO computational performance by offloading the computations for diagnostics on GPU.
- The ocean global heat content, salt content and volume conservation diagnostics (`dia_hsb`) has been chosen as starting point because it is the most expensive.
- The code itself is executed 50x faster than in a single CPU but the data transfer to and from GPU is the main bottleneck.
- Pinned Memory and GPU Directly Attached to the host can be used to mitigate the data transfer penalty
- Asynchronal communications and a memory buffer approach reduce significantly the data transfer penalty

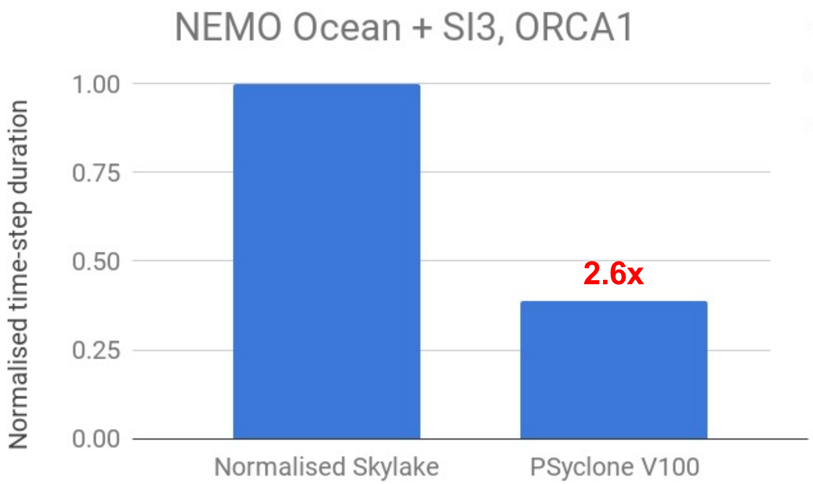
Online diagnostics – GPU based

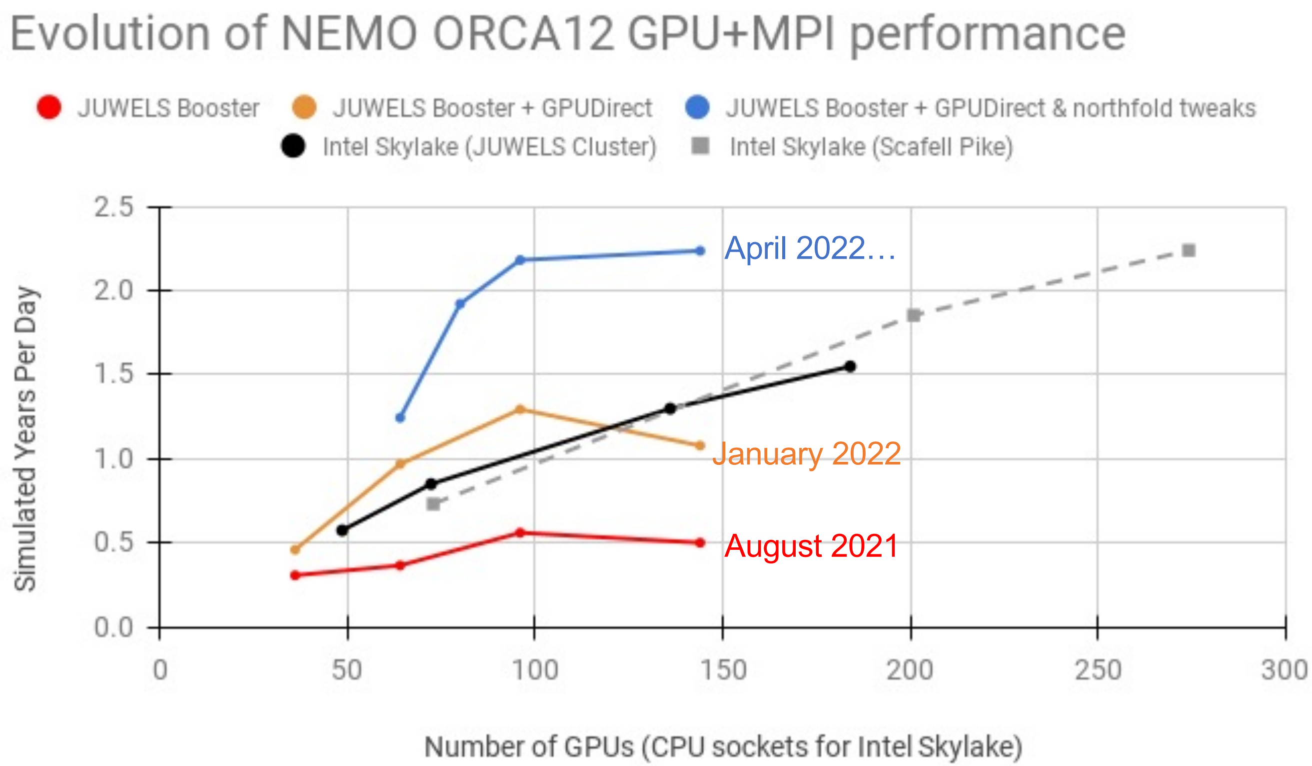
`dia_hsb` scalability



PSyclone for NEMO







Science and Technology Facilities Council
Hartree Centre