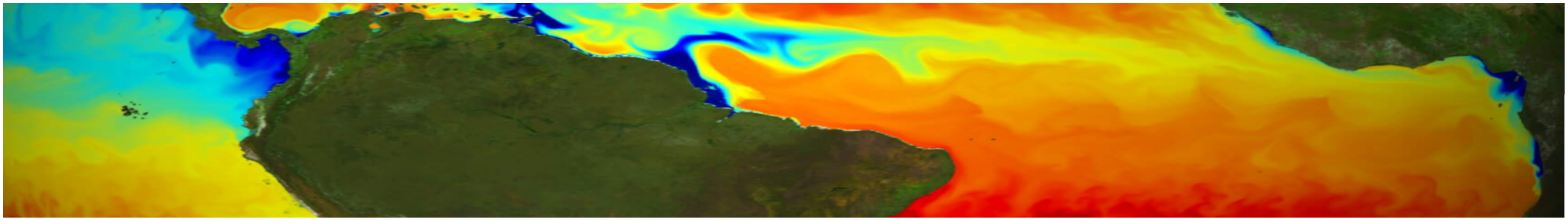


High Performance Computing
for

NEMO

https://forge.ipsl.jussieu.fr/nemo/wiki/WorkingGroups/NEMO_HPC

Met Office
Mercator Ocean
Ocean Forecasters
cnrs
cmcc
Centro Euro-Mediterraneo sui Cambiamenti Climatici
Science & Technology Facilities Council
National Oceanography Centre
NATURAL ENVIRONMENT RESEARCH COUNCIL
BSC
UNIVERSITY OF EXETER
ECMWF
intel
CRAY
NVIDIA
Atos

On going actions

- Tiling
- NEMO on GPU with PSyClone
- Extended halo
- Mixed precision
- Offloading diagnostics on GPU
- Loop fusion
- Neighbouring collective communications
- Improving the use of XIOS

Tiling

Description:

The full processor domain (jpi x jpj) is split into one or more tiles in order to enhance the cache memory reuse.

1. Modifying the DO loop macros to instead use the tile bounds
2. Declaring SUBROUTINE-level arrays using the tile bounds
3. Looping over tiles at the timestepping level
4. A new namelist (namtile) to configure the tile shape
5. Replacing subscripts with a DO loop macro where appropriate

Status:

- Tiling has been implemented for most of the code called in the “active tracers” part of the timestepping subroutine.
- At present, many routines cannot be fully tiled and tiling must be locally disabled to preserve the results.
- Results from a GYRE benchmark show a reduction in cost of 35-70% for tra_ldf, 35-50% for tra_adv, ~65% for tra_zdf and ~35% for tra_sbc.

NEMO on GPU

Description:

Use PSyclone to automatically insert openACC directives into the code

Status:

UK Met Office 'ExCALIBUR' project began in June 2020:

- STFC developing PSyclone and applying to NEMO OCE and SI^3
  - PSyclone-processed NEMO OCE (ORCA1) on V100 now 1.5x faster than Skylake (with more work to do)
  - SI^3 working on GPU but performance not there yet.
- NOC applying PSyclone to MEDUSA
  - MEDUSA has been incorporated into BENCH. Has been processed with PSyclone and executed on GPU but no optimisation work performed yet.
- Reading (NCAS) applying PSyclone to NEMOVAR
  - Mini-app has been optimised on GPU. Work ongoing to support Fortran derived types in PSyclone.

Extended Halo

Description:

A wider halo=2 can be used to reduce the communications and move the lbc_lnk calls forward in the code.

1. Suppress the halo from the inputs and outputs
2. Modifying the domain size
3. Modifying the message size exchanged during the lbc_lnk call
4. Clean up the code removing the useless lbc_lnk calls

Status:

- halo=1 and halo=2 will be supported for the next years
- All the subroutines in the TRA module have been updated

Future:

- The final implementation aims at handling different halo sizes in different part of the code (i.e. a wider halo could be more efficient when timesplitting is used)
- The halo exchange should happen only at the end of the time iteration

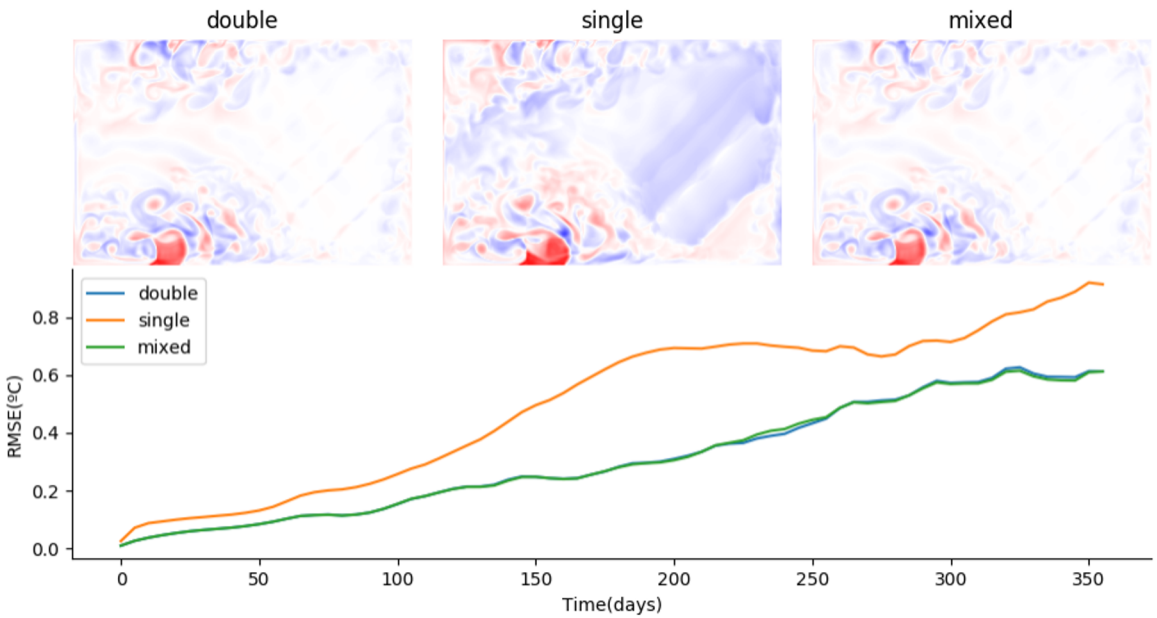
Mixed Precision (BSC)

Description:

An optimization of the numerical precision can help to reduce data movement and help to better exploit vectorization, bringing performance improvements while maintaining the accuracy of the results.

Status:

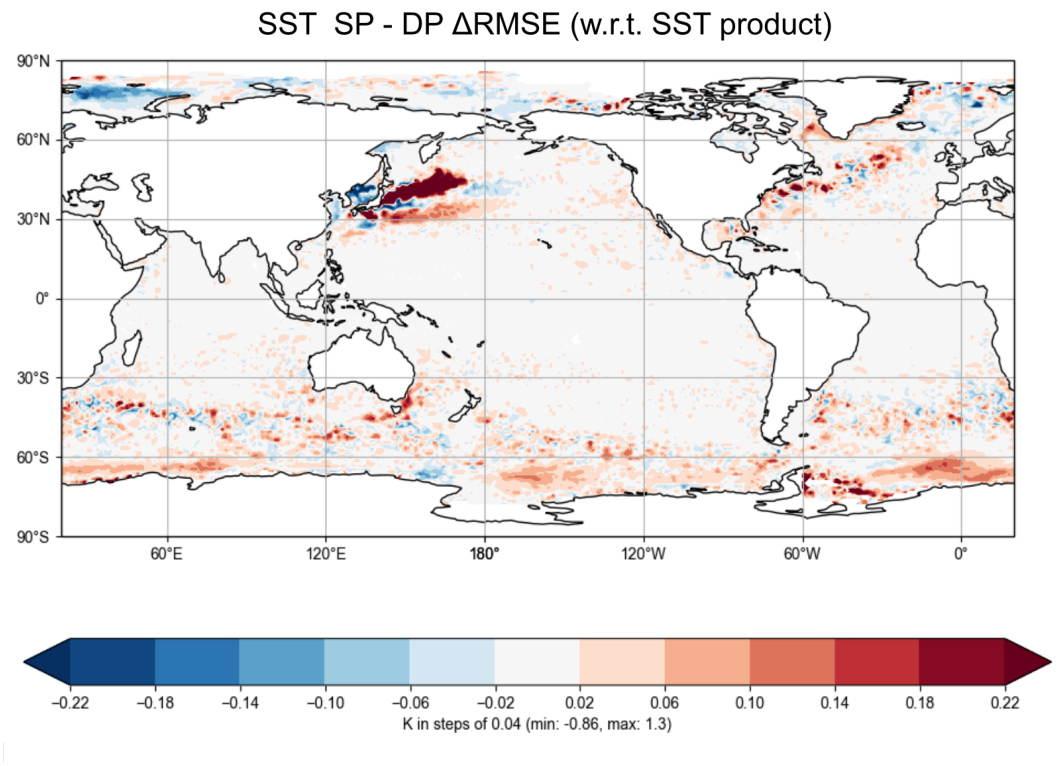
- Everything should be ready for the next merge party.



Impact of the mixed precision on SST using GYRE 1/9 config.

Mixed Precision (ECMWF)

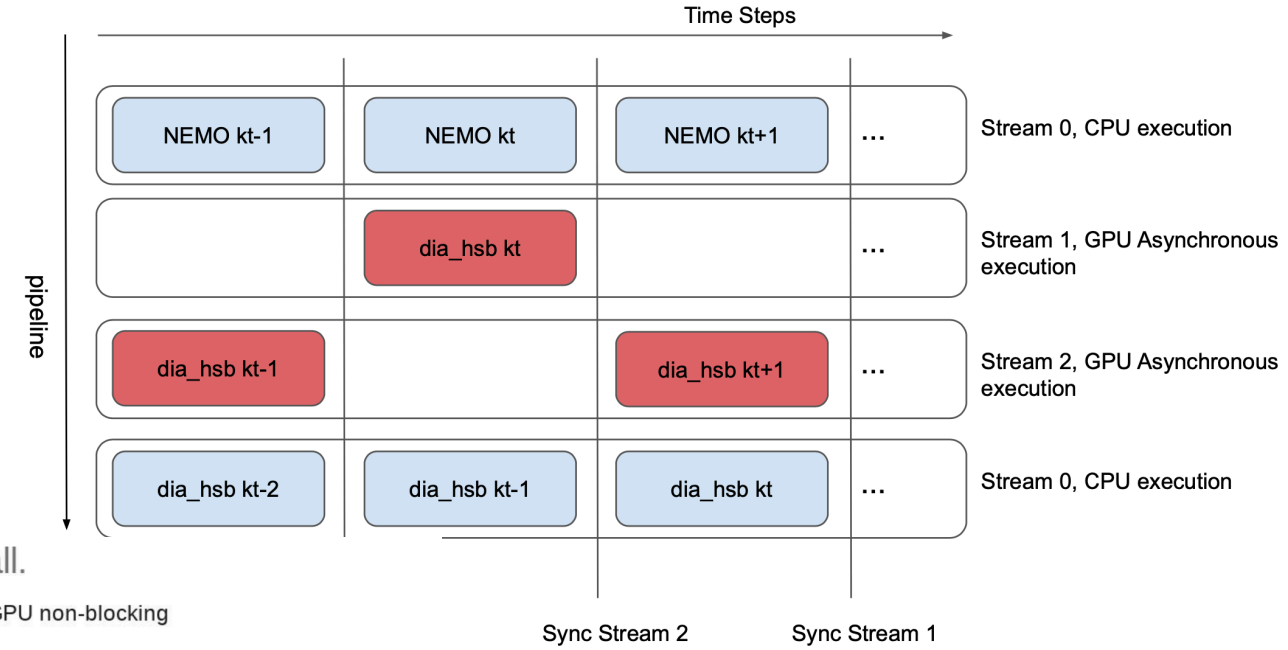
- Testing of BSC's single-precision NEMO underway at ECMWF (including SI3)
- SP in ORCA1 gives ~no change in error w.r.t. DP (compared with real obs.)
- SP in ORCA025 mostly a neutral change except for extra ~1K warm bias over Kuroshio extension
- Speed-up up to 1.7x w.r.t. DP
- Next steps:
  - Fix Kuroshio problem
  - Fully SP ocean-atmosphere runs
  - Rigorous benchmarking of performance gains

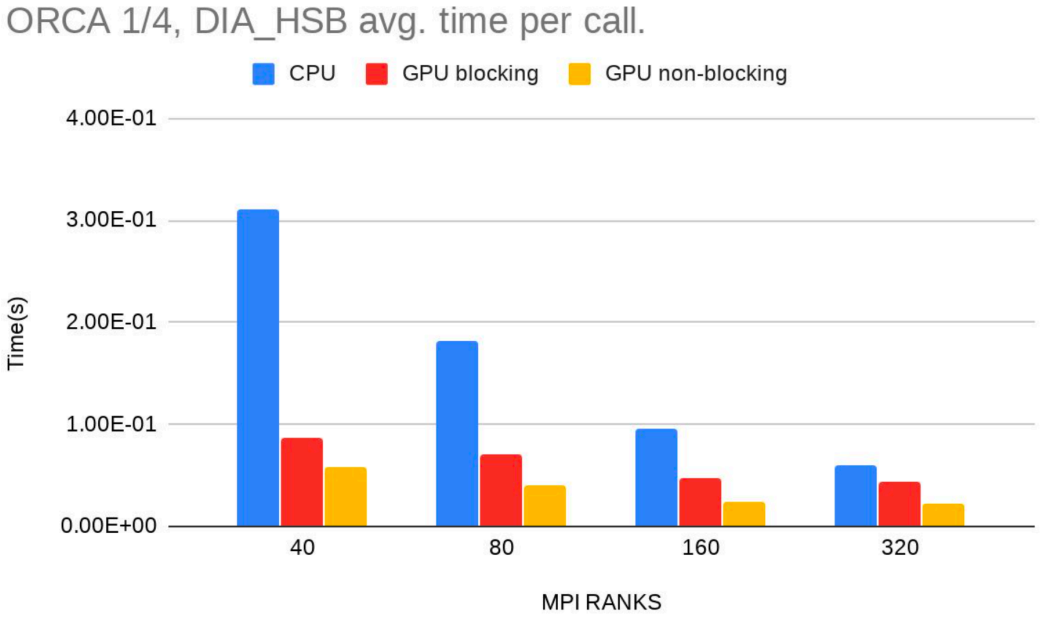


Change in sea-surface temperature error when switching from double to single-precision, NEMO eORCA025_Z75

Offload diagnostics on GPU

The rationale of this activity is to improve the NEMO computational performance by offloading the computations for diagnostics on GPU.



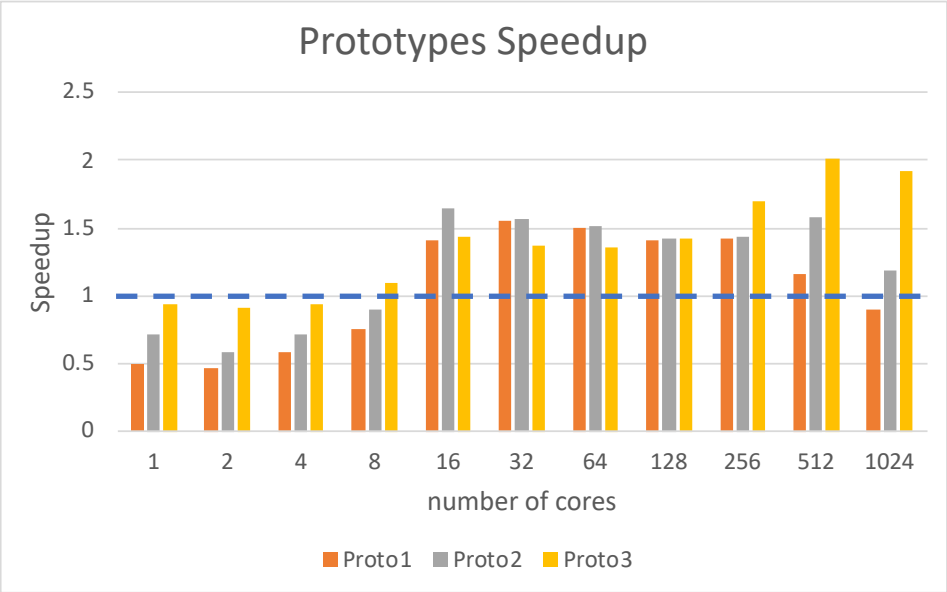


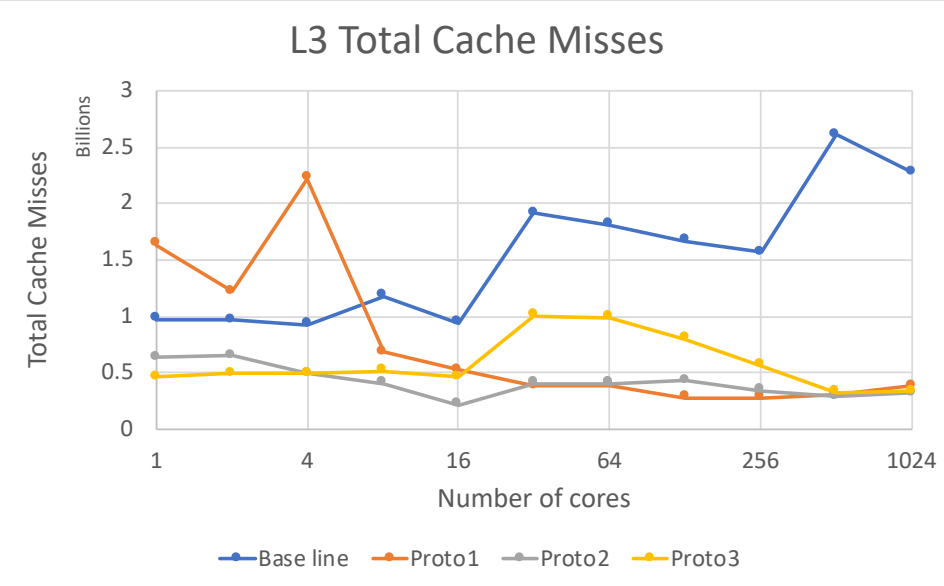
Loop Fusion

Loop fusion 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
```

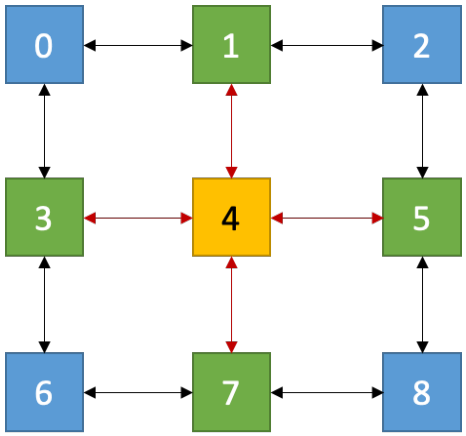
```
DO j=2, n-1; DO i=1, n
      b_0  = in(i,j+1) - in(i,j  ) ! correspond to b(i,j)
      b_m1 = in(i,j  ) - in(i,j-1) ! correspond to b(i,j-1)

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



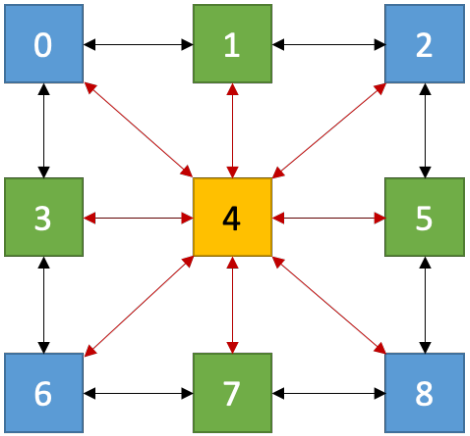


Neighbouring Collective Communications

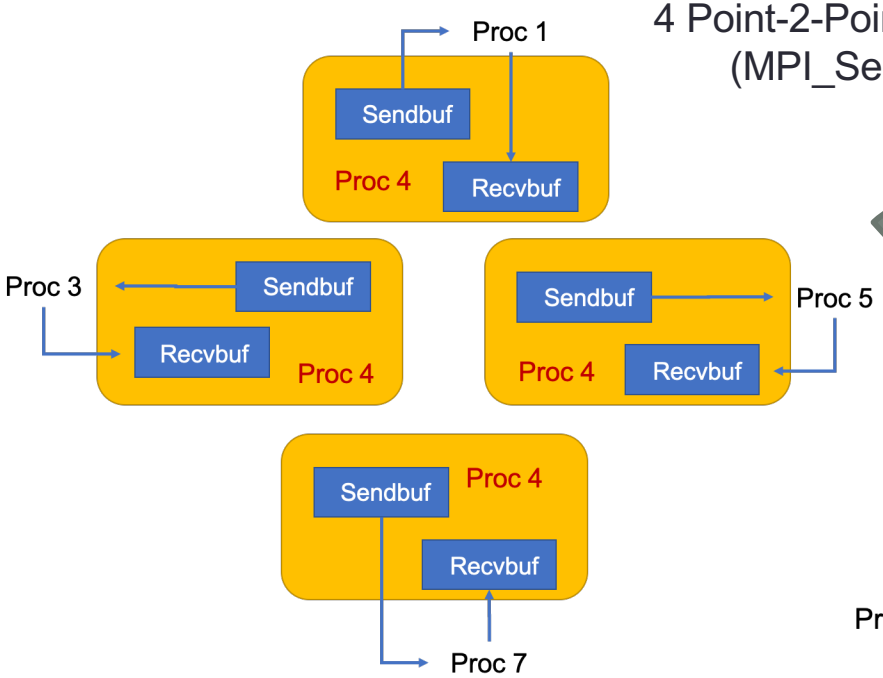
5-points stencil



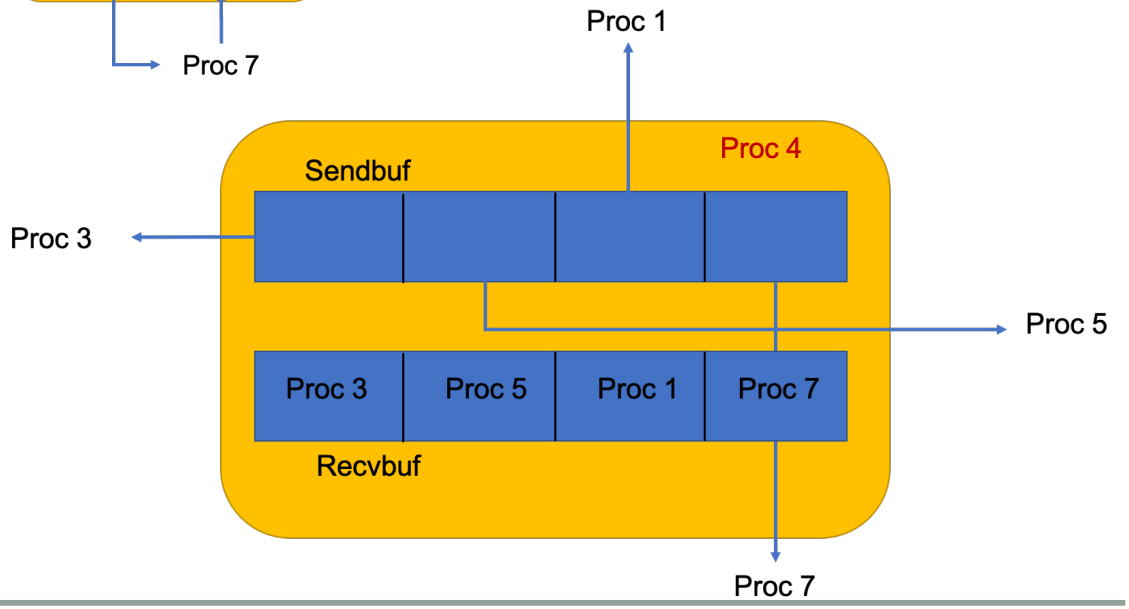
9-points stencil



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

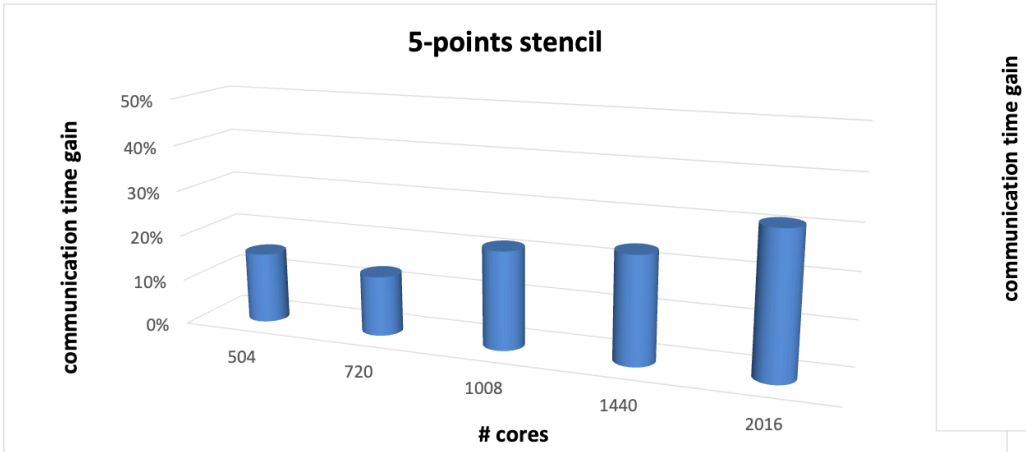


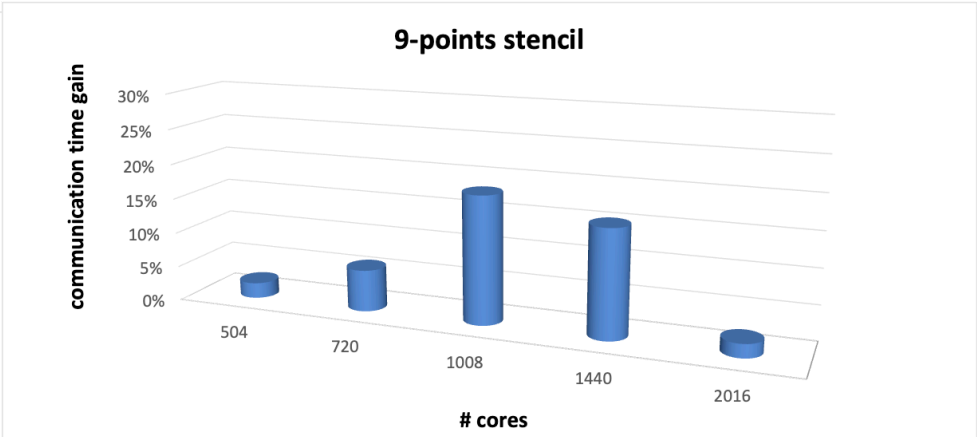
1 collective communication
(MPI_Neighbor_alltoall)



Neighbouring Collective Communications

- Extension of the LBC module to support MPI3 Neighborhood Collectives halo exchange:
  - Use of graph instead of cartesian topology to support halo exchange also when:
    - 9-points stencil is needed
    - Land domains exclusion is activated
  - Implementation both versions of 5-points and 9-points stencil exchanges
- Replacement of point-to-point communications with collective ones in the whole NEMO code
  - 9-points version (done)
  - 5-points version will replace 9-points one if data dependency is satisfied (to be completed in 2021)
- Introduction of the key_mpi3 to activate/deactivate new communications
  - preserving the old point-to-point exchange version to be used on architectures where MPI3 is not supported
- Performance evaluation in communication time using 5-points and 9-points exchanges
  - GYRE_PISCES configuration (nn_GYRE=100 → ~3000x2000x31 grid resolution)





Improve the use of XIOS

This work aims at improving the use of XIOS for reading/writing in NEMO

- Extend restart read write to SI3 and TOP (tracers): Ready and Tested

- Read ancillary data: Ready and tested

- Use of XIOS into fldread
  - XIOS doesn’t support reading of split files
  - There is also problem with jumping between time records in XIOS

Development Strategic Plan 2018-22

Strategic Plan	On Going Actions
3.3.1 Internode communications - Extending the halo size - Overlapping communications and computations	HPC-08_epico_Extra_Halo HPC-07_mocavero_mpi3
3.3.2 Shared Memory Parallelism - Tiling - Use of OpenMP / OpenACC	HPC-10_mcastril_HPDAonline DiagGPU
3.3.3 Single core performance - Better exploitation of cache memory (Tiling) - Enhancement of vectorization level	HPC-02_daley_Tiling HPC-09_epico_Loop_fusion
3.3.4 Designing a user-friendly code structure - Performance portability - Separation of Concerns - PSyClone	HPC-01_daley_GPU (PSyClone)
3.4 Additional - Macro task parallelism - Mixed precision	HPC-04_mcastril_Mixed_Precision TOP-06_emalod_OASIS_btw_TOP_NEMO

Overall considerations

- The NEMO HPC-WG gathers a wider community beyond the members of the System Team
  - BSC, ECMWF, NVIDIA, ATOS
- The HPC-WG meets quite regularly once every two months
- All the recommendations of the Dev Strategic Plan are fully covered
- Almost all the current developments will be completed in 2021 or early 2022
- The new HPC improvements can be included in the NEMO trunk with some more restrictive requirements
  - The accuracy of the model must not be “compromised”
  - The developer’s interfaces must be kept easily understandable even by not hpc experts
- Some possible issues are related to the maintenance of the code (i.e. debugging, ticketing, …) not developed by any of the System Teams members
- All the activities are funded with projects at National or at European level
  - (e.g. IS-ENES3, ESiWACE2, IMMERSE, ESCAPE2, 'ExCALIBUR, …)