scalability issues

S. Masson for the NEMO system team
F. Vigilant (Bull)
E. Maisonnave (CERFACS)

OUTLINE

- NEMO scalability: State of the art & bottleneck
- “Exascale” project for NEMO
- IO performances
- Other components: Sea-Ice, AGRIF, TOP
- XeonPhi, GPU…
- Climate: coupler
- Conclusion

State of the art

The “project funding” paradox :

Always more groups and projects working or proposing to work on NEMO performances…

but

- Still no clear ideas of the issues related to NEMO performances
- Still no real quantifications of the bottlenecks
- Is there only one configuration profiling ?
- Sensitivity of these figures with domain size and core #?

A marketing problem?

Architect of an Open World™

eXaScale Projet for

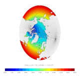
NEMO

center for excellence in parallel programming

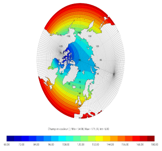
center for excellence in parallel programming

Co-design

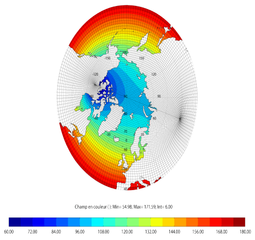
Contribute to the next generation of NEMO for eXaScale



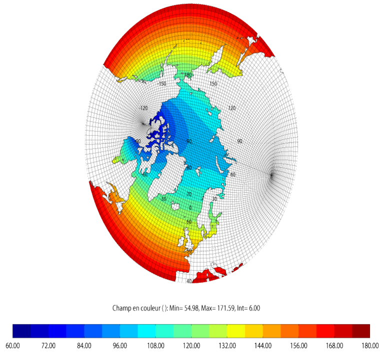
ORCA 2
550 MB of memory
8 CPU hours
10 Gigabytes of output
(daily)



ORCA ¼
47 Gigabytes of memory
3500 CPU hours
120 Gigabytes of output
(daily)



ORCA 1/12
414 Gigabytes of memory
90 000 CPU hours
1 Terabyte of output
(daily)



ORCA 1/36
> 1 Terabytes of memory
~4 000 000 CPU hours
> 5 Terabytes of output
(daily)

Science improvements may be driven by mesh refining involving more and more grid points but also more and more parameters in the models. **Optimisation** & **Scalability** are thus key to compute efficiently.

center for excellence in parallel programming

Co-design

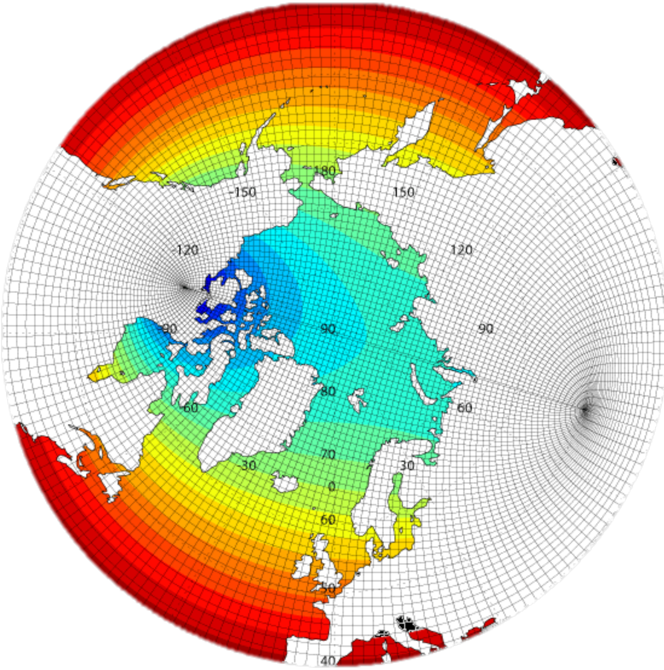
Bull Architect of an Open World

NEMO

Start from the basics:
- Benchmaking, timing
- Isolate the parameters impacting NEMO scalability
  - domain size (GYRE6/GYRE144, ORCA2/ORCA12)
  - Use of sea-ice model
  - North-pole folding

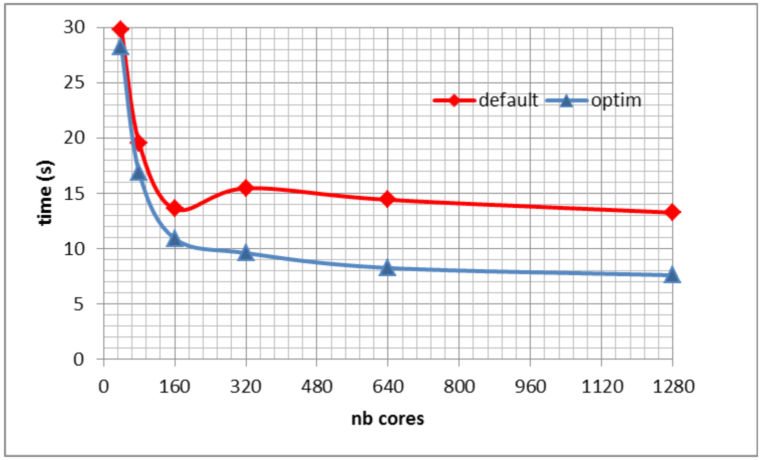
Improve existing model at limited cost:
- Suppress all global communications (time splitting)
  NO MORE SOLVER !!!
- Gather communications
- Point out sequential parts of the model
- Improve vectorisation
- MPI Communication improvement

Longer term work:
- Kernel optimization
- Hybrid MPI/OpenMP
- Intel Xeon Phi and GPU testing



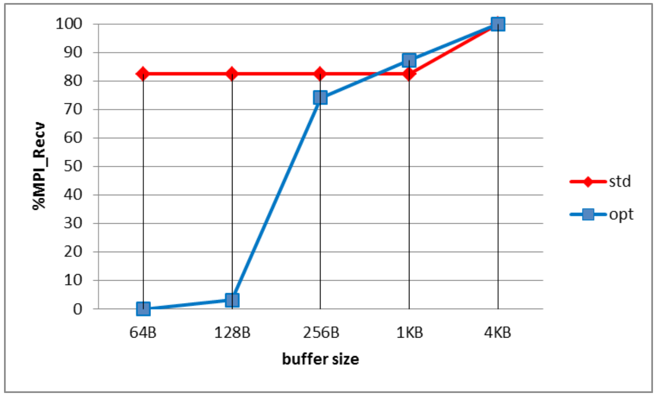
Configuration: Ideal case – GYRE6

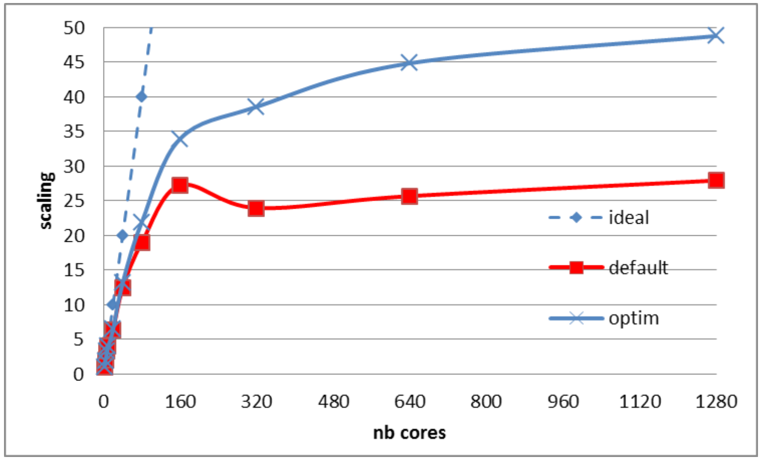
- domain size 182 x 122 x 31
- scales up to subdomain size of 20x10
- point-to-point MPI communications explode **(no more global)**



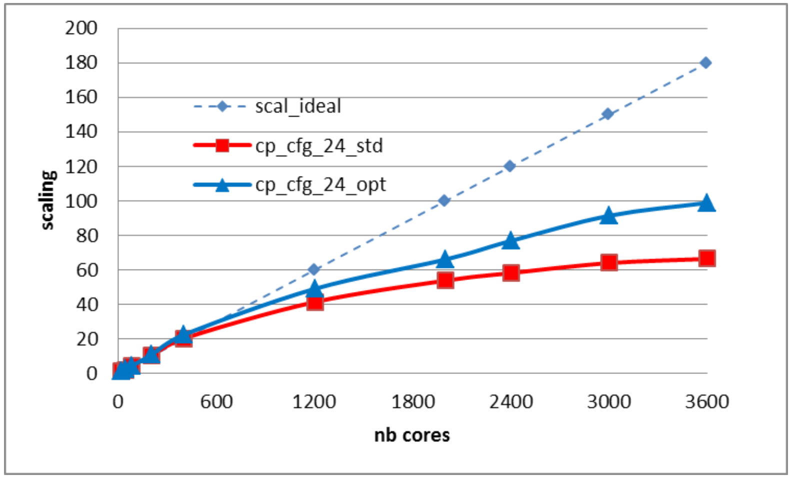
Improve MPI efficiency:

- maximize volume of data to be sent
- minimize MPI calls
- scales up to subdomain size of 7x6



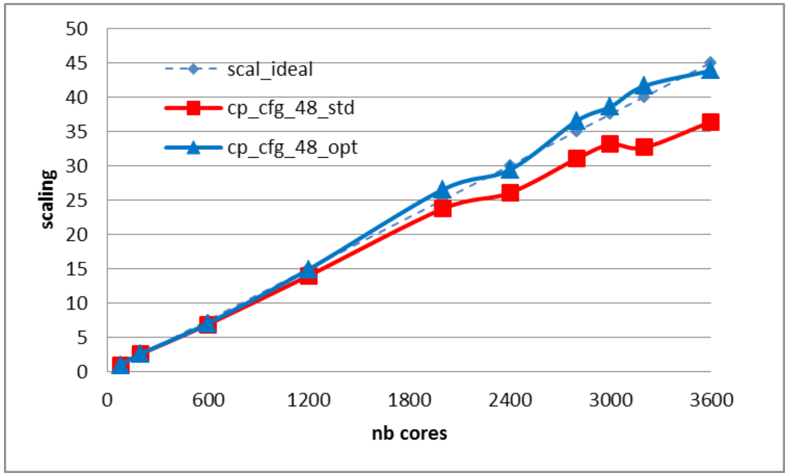


Configuration: Ideal case – GYRE_24



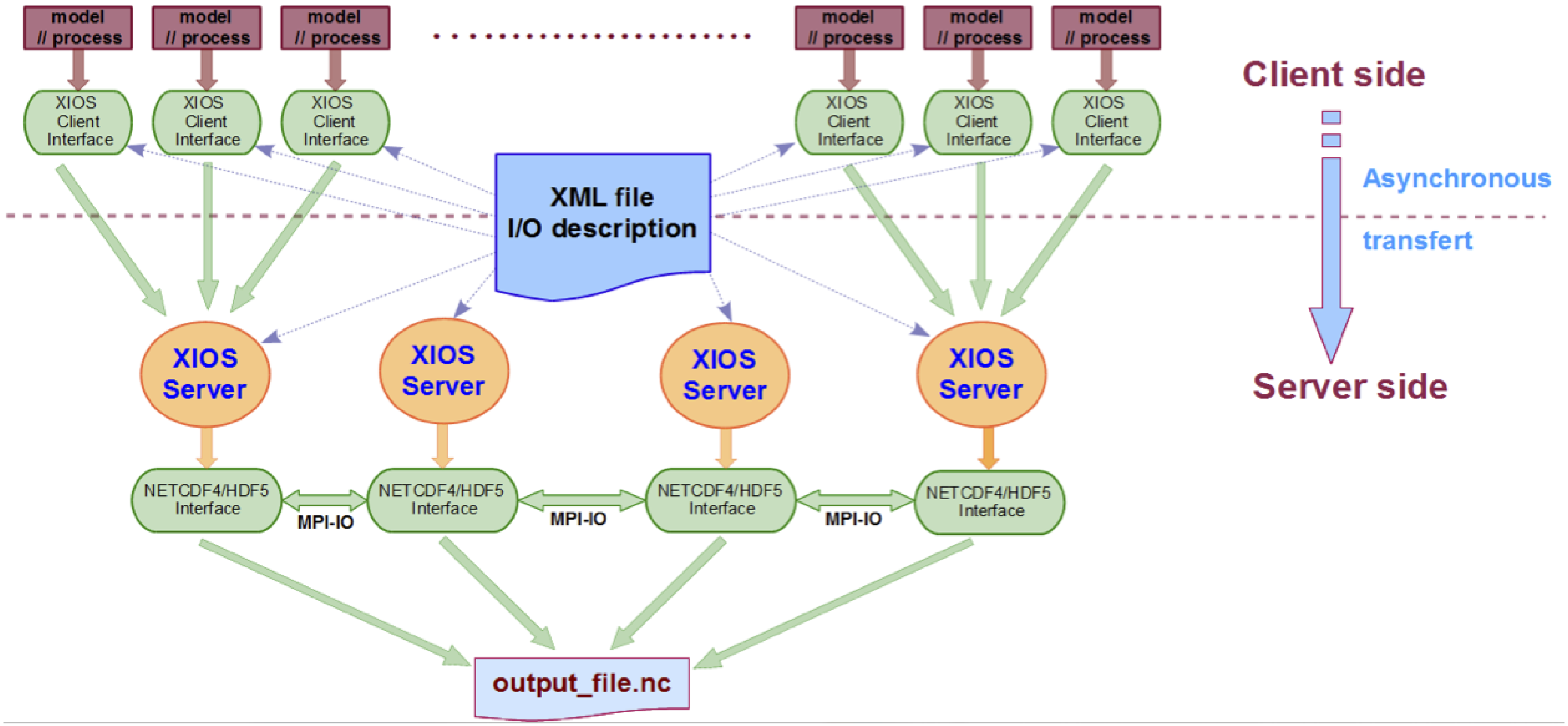
- domain size 722 x 482n x 31
- scales up to subdomain size of 12x12
- scalability is improved

Configuration: Ideal case – GYRE_48



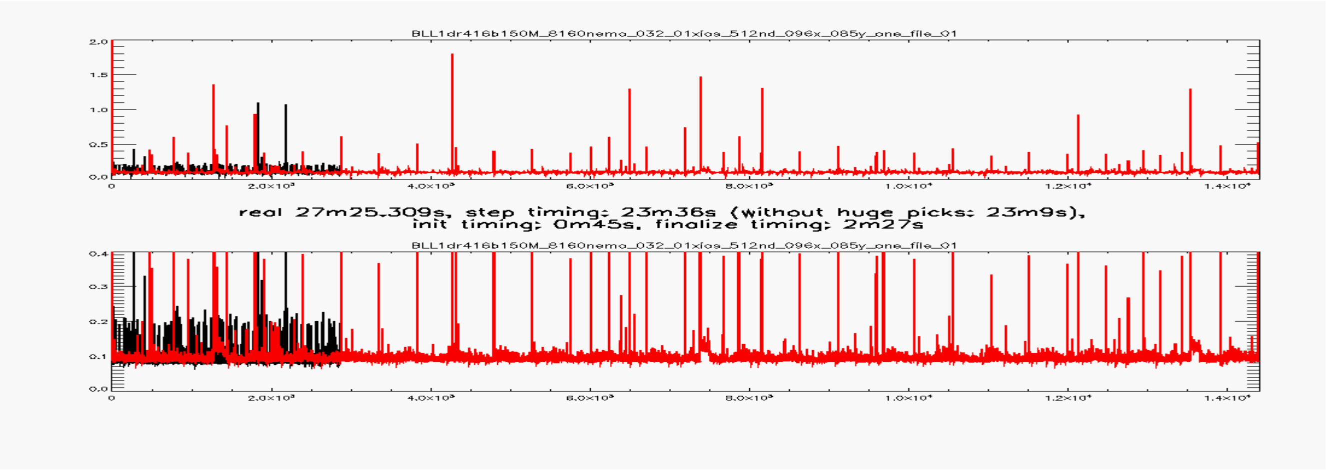
- domain size 1442 x 962 x 31
- scales up to subdomain size of 22x22
- ideal scalability on the experiment range

IO
Output diagnostic files: based on XIOS



XIOS
BIG output benchmark
daily mean outputs (one_file mode)

example: GYRE 144 (4322*2882*31)
30d simulation (14400 time steps):
in red: with daily outputs (every 480 step, total: **235G**)
in black: no outputs (enable = false)

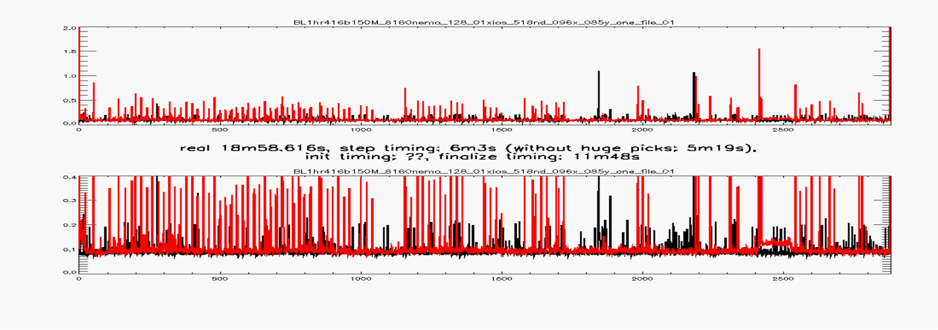


8160 nemo
+ 32 xios
=> +1.5% for IO

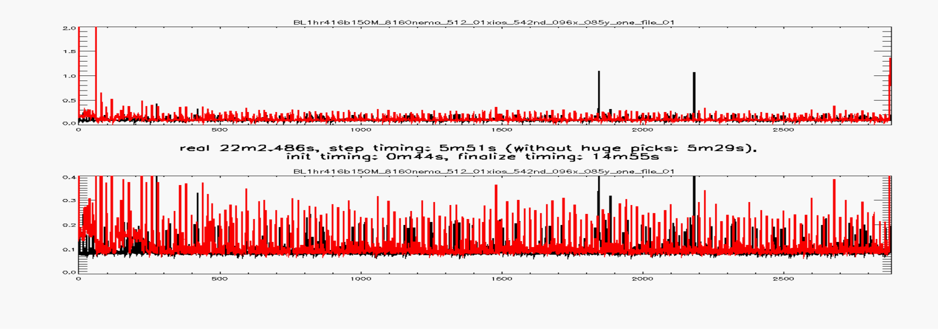
XIOS
HUGE output benchmark
hourly mean outputs (one_file mode)

example: GYRE 144 (4322*2882*31)
6d simulation (2880 time steps):
in red: with hourly outputs (every 20 step, total: **1.1T**)
in black: no outputs

8160 nemo + 128 xios



8160 nemo + 512 xios



3.1-3.6G/s => 10-12T/h...

8160 nemo: **+ 15~20% for IO**

IO

XIOS extremely efficient and convenient to output diagnostics

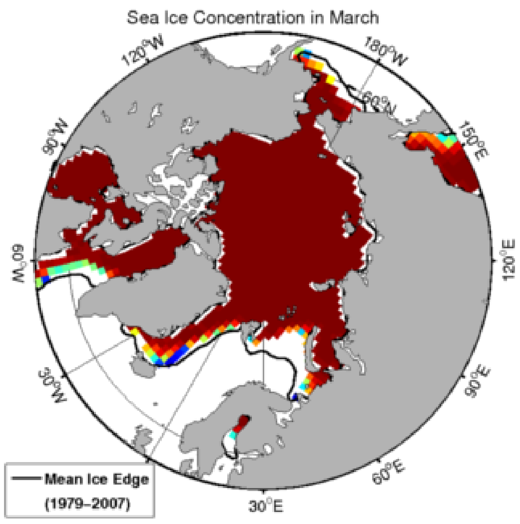
Remaining (future) bottlenecks:

- **Input** files: read by each MPI subdomain ask (but with on-the-fly interpolation)
- **Restarts** files are written/read by each MPI subdomain

Further development already planned for XIOS

- Input file and restart with XIOS
- Improve even more the scalability
- Optimise the usage and the size of buffers
- Allow grib format (?)

Sea-Ice

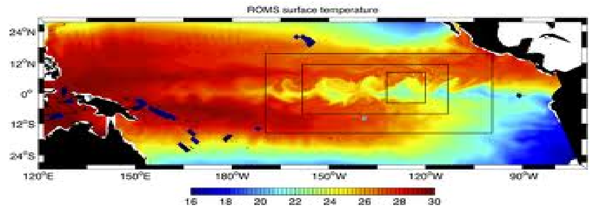


More and more expensive…
- LIM2 -> LIM3 with ice categories, active salinity
- future: more complex rheology

Expected issue for scalability:
- unbalance between points with/without sea-ice
- solver in the sea-ice rheology

Proposed solutions for future developments
- Again, start with a clear and quantitative benchmaking
- asynchronous integration of ocean and sea-ice
  -> dedicated cores for sea-ice
- different grid resolution of ocean and sea-ice
  -> coupling with oasis?
- Replace solver by time-splitting as for the ocean

AGRIF



Again, no clear and quantitative benchmaking of AGRIF…
Impact of the interpolation between the different grids on performance and scalability?
On going work: run several nests at the same level in parallel

TOP (PISCES)

Again, no clear and quantitative benchmaking of TOP…
More computation, not so many communications…
Should help for the scalability…

OpenMP
Xeon Phi
GPU
Vector again ?

First step: add OpenMP
Ongoing work by
- CMCC (Italo Epicoco, Silvia Mocavero)
- BULL (Franck Vigilant, Cyril Mazauric)

Second step: check vectorization

To go further ? H2020 “CHANCE” lead by CMCC
Parallel-in-time NEMO ?

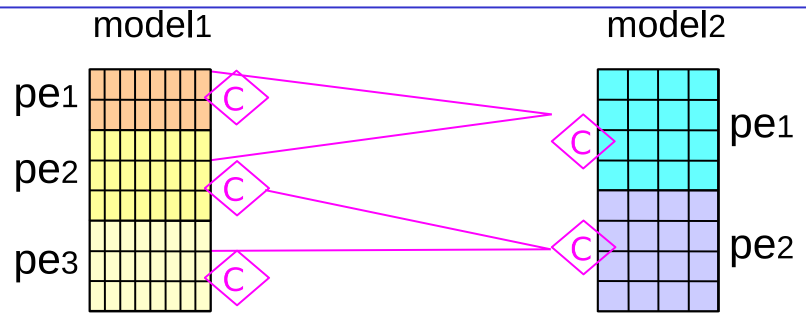
CONCLUSION

- Need a clear and quantitative benchmaking of NEMO to sort out key issues.
- Co-design: a key of the success if involving HPC and NEMO experts
- Still a large scope for scalability improvement before rewriting everything
- Clear roadmap for the IO part.
- But need to start now to work on long term developments

… and what about ocean-atmosphere coupling ?

OASIS3-MCT

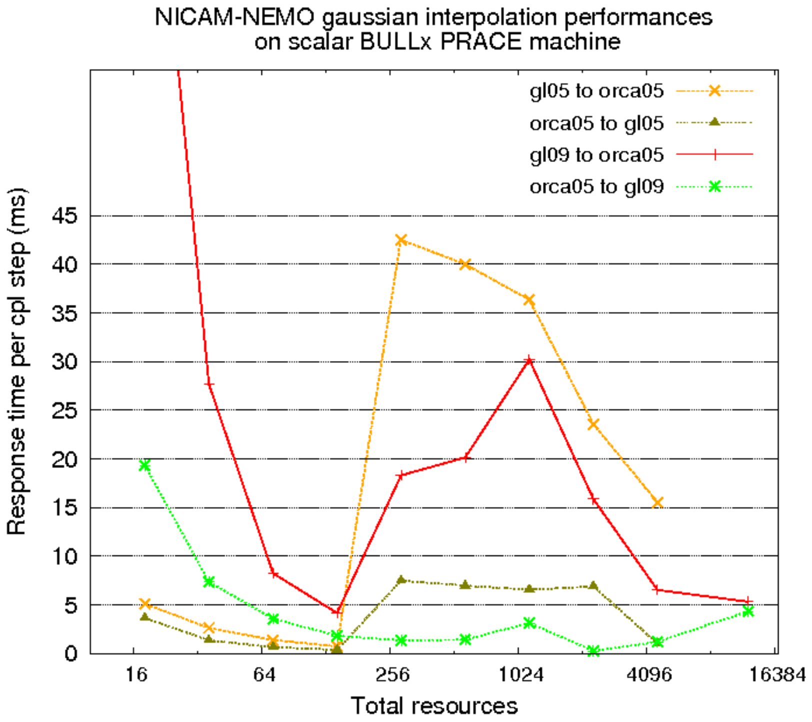
- Developed by CERFACS since 1991 with CNRS since 2005 and many others
- Written in F90 and C; open source license (LGPL)
- Last OASIS3-MCT version based on MCT
- Public domain libraries: MPI; NetCDF; LANL SCRIP
- Large community of users: ~35 climate modelling groups world-wide,rapidly growing



- Sequential weights calculation (SCRIP library)
- Parallel regridding on source or targets processes (MCT)
- Redistribution of coupling data (MCT)

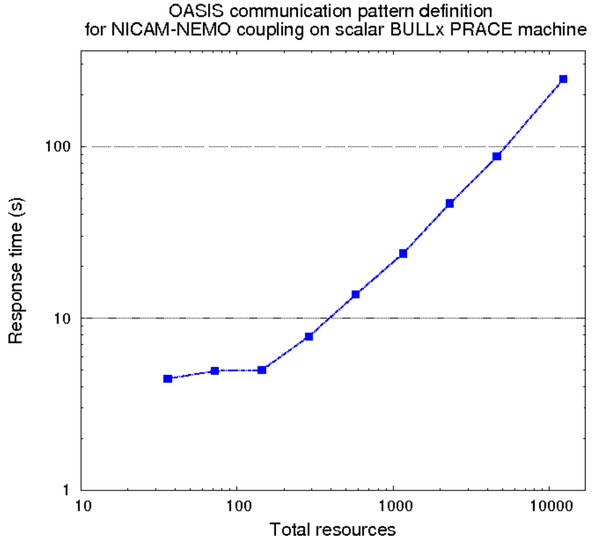
OASIS3-MCT Success Stories

1. NICAM-NEMO
(JAMSTEC-IPSL)



10 interpolations (1 coupling time step) from/to NICAM icosahedral grid (12Km)

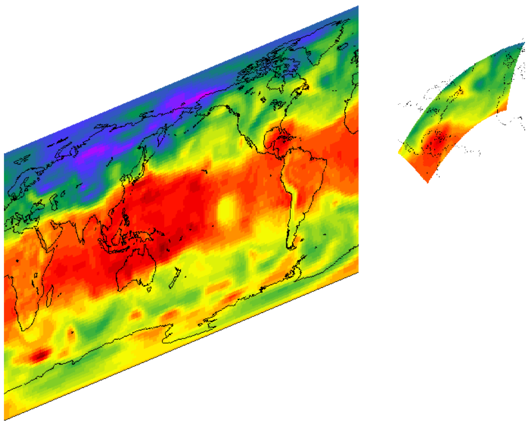
Next bottleneck:
initialisation at o(10,000)



OASIS3-MCT Success Stories

2. ECHAM-COSMO
(BTU Cottbus, FU Berlin)

OASIS3-MCT coupling between global & regional grid
- with 6 47-levels 3D fields (2 way nesting) = 287 2D fields
- at each ECHAM time step
- includes ECHAM-MPI-OM (ocean) coupling



<u>Main results</u>
- Efficiency: overhead = few %
- Modularity: can be coupled with CLM
(Community Land Model, as part of CESM, NCAR)

Conclusion:
OASIS is scalable again, and still good for modularity