Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Facilitating higher resolution Ocean simulations in EC-Earth4: NEMO mixed precision and I/O at 3km global

Miguel Castrillo

BSC-ES, Computational Earth Sciences

EC-Earth meeting

09/02/2021

MareNostrum 4

Total peak performance: **13,7** Pflops		
General Purpose Cluster:	**11.15 Pflops**	**(1.07.2017)**
CTE1-P9+Volta:	**1.57 Pflops**	**(1.03.2018)**
CTE2-AMD:	**0.52 Pflops**	**(2020)**
CTE3-Arm V8:	**0.5 Pflops**	**(2020)**

PRACE

Access: prace-ri.eu/hpc_acces

RES
RED ESPAÑOLA DE SUPERCOMPUTACIÓN

Access: bsc.es/res-intranet

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MareNostrum 1
2004 – 42,3 Tflops
1st Europe / 4th World
New technologies

MareNostrum 2
2006 – 94,2 Tflops
1st Europe / 5th World
New technologies

MareNostrum 3
2012 – 1,1 Pflops
12th Europe / 36th World

MareNostrum 4
2017 – 11,1 Pflops
2nd Europe / 13th World
New technologies

MareNostrum 5. A European pre-exascale supercomputer

- **200 Petaflops** peak performance (200 x 10^{15})
- **Experimental platform** to create supercomputing technologies “made in Europe”
- **223 M€** of investment

EuroHPC Joint Undertaking

Hosting Consortium:

Spain Portugal Turkey Croatia




Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Generalitat de Catalunya
Departament d'Empresa i Coneixement

Gobierno de España
Ministerio de Ciencia, Innovación y Universidades

Universitat Politècnica de Catalunya
BarcelonaTech

ESiWACE2 WP1 - GCM 10 km

Excellence in Simulation of Weather and Climate in Europe

- **Task 1.1: Develop infrastructure for production-mode configurations**
  - Introduce **XIOS** in EC-Earth, NEMO **Mixed Precision**...
- **Task 1.2: Develop production-mode configurations**
  - **EC-Earth: 16 km** (TL1279) atmosphere coupled to a **1/12 degree (9 km) ocean**
- **Task 1.3: Port models to pre-exascale EuroHPC systems**
  - Port EC-Earth ~10km to MareNostrum5

NEMO 4

- **New Sea-Ice** component (SI3)
- **AGRIF compatible** with sea-ice and z* coordinate
- **Aerobulk** package for atmospheric **forcing**
- **Wave coupling** to external wave model
- Passive tracer module (**TOP**) **re-designed** (modular)

- **MPI** communications **reduced**
- Removal of **wrk_alloc**'s
- Automatic **land** sub-domains **removal**
- **Simplification** & **robustness**

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NEMO 4

- **New Sea-Ice** component (SI3)
- **AGRIF compatible** with sea-ice and z* coordinate
- **Aerobulk** package for atmospheric **forcing**
- **Wave coupling** to external wave model
- Passive tracer module (**TOP**) **re-designed** (modular)

- **MPI** communications **reduced**
  - North pole **folding**
  - **SI3:** Group comms., remove globals
- Removal of **wrk_alloc**'s
- Automatic **land** sub-domains **removal**
- **Simplification** & **robustness**

Mixed precision calculations in NEMO

In a nutshell

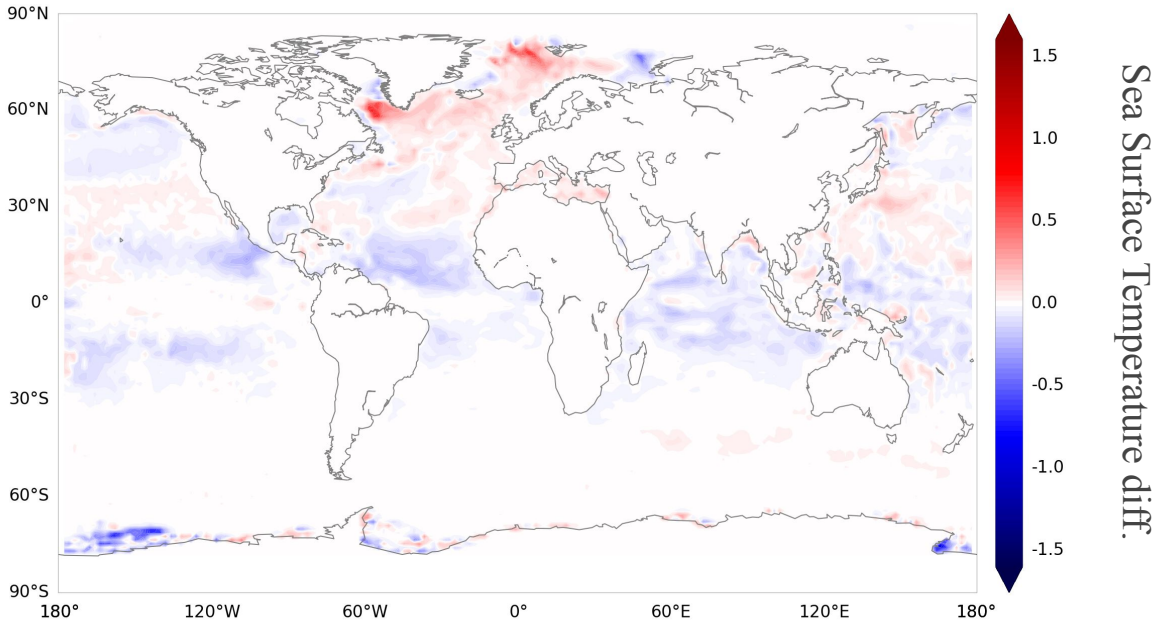
- Different computations might require a **different** level of **precision**.
- Customized precision has emerged as a **promising approach** to improve power / performance trade-offs.
- We developed a **method** useful to **determine** the **precision needed** for the different variables in a model.
- Branch in NEMO 2020 WP ⟶ Aiming for merge in **NEMO 4.2**.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

2019: "How to use mixed precision in ocean models: exploring a potential reduction of numerical precision in NEMO 4.0 and ROMS 3.6."
Tintó Prims, O., Acosta, M.C., Moore, A.M., Castrillo,M., K. Serradell, A. Cortés and F.J. Doblas-Reyes

Discriminating accurate results in nonlinear model

*Difference between **double** and **single** after one month*

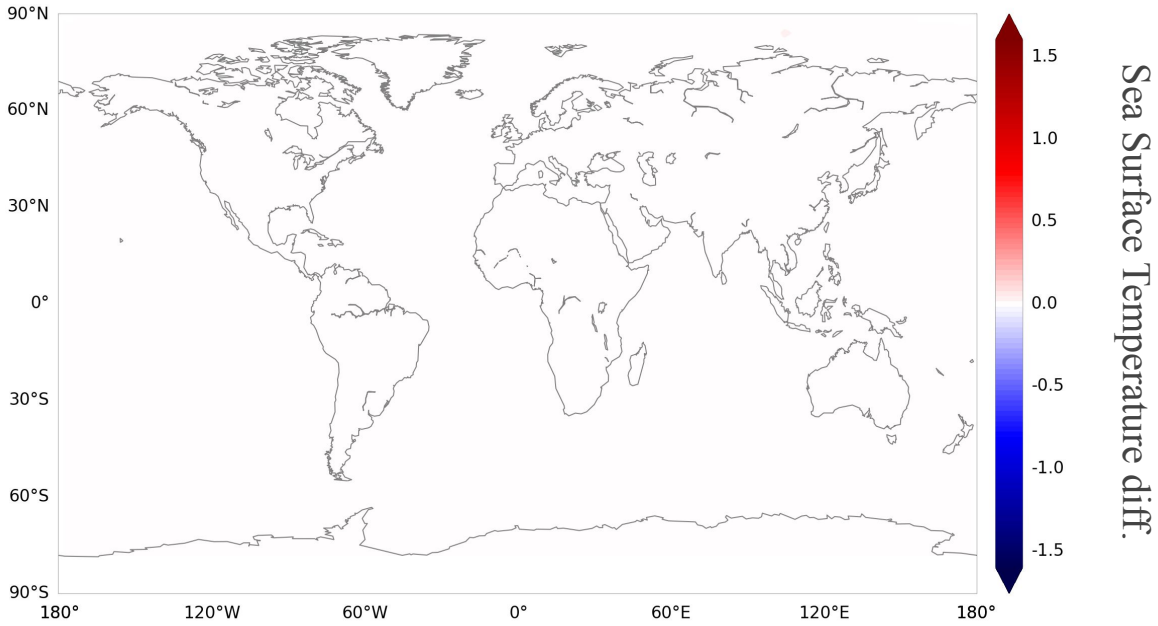


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Oriol Tintó

Discriminating accurate results in nonlinear model

*Difference between **double** and **mixed** after one month*

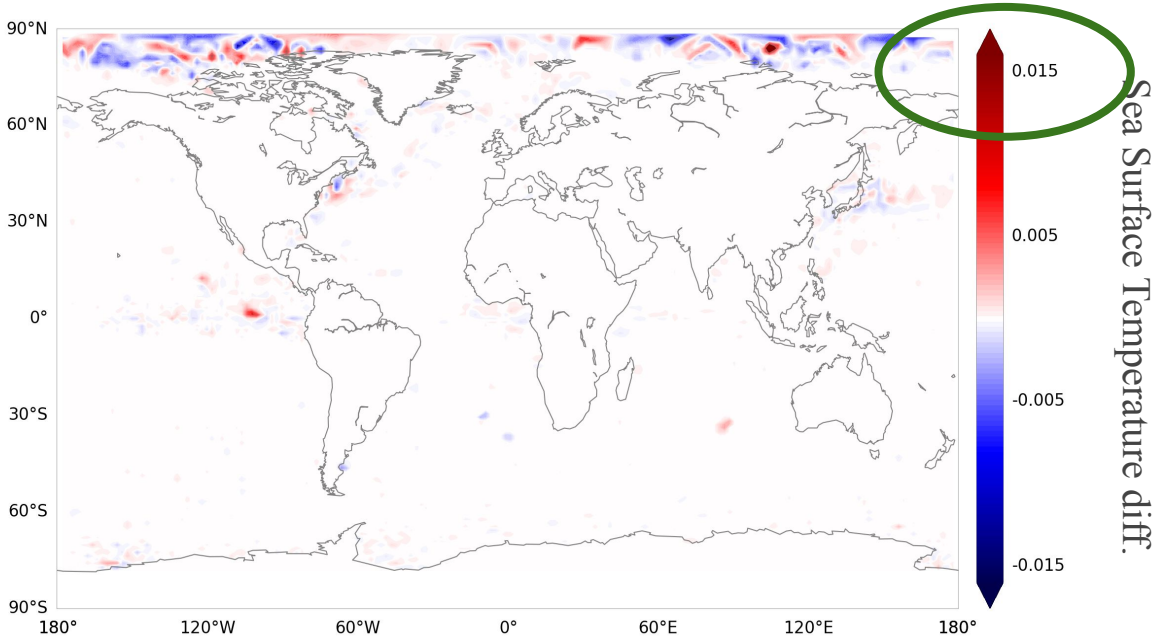


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Oriol Tintó

Discriminating accurate results in nonlinear model

*Difference between **double** and **mixed** after one month*



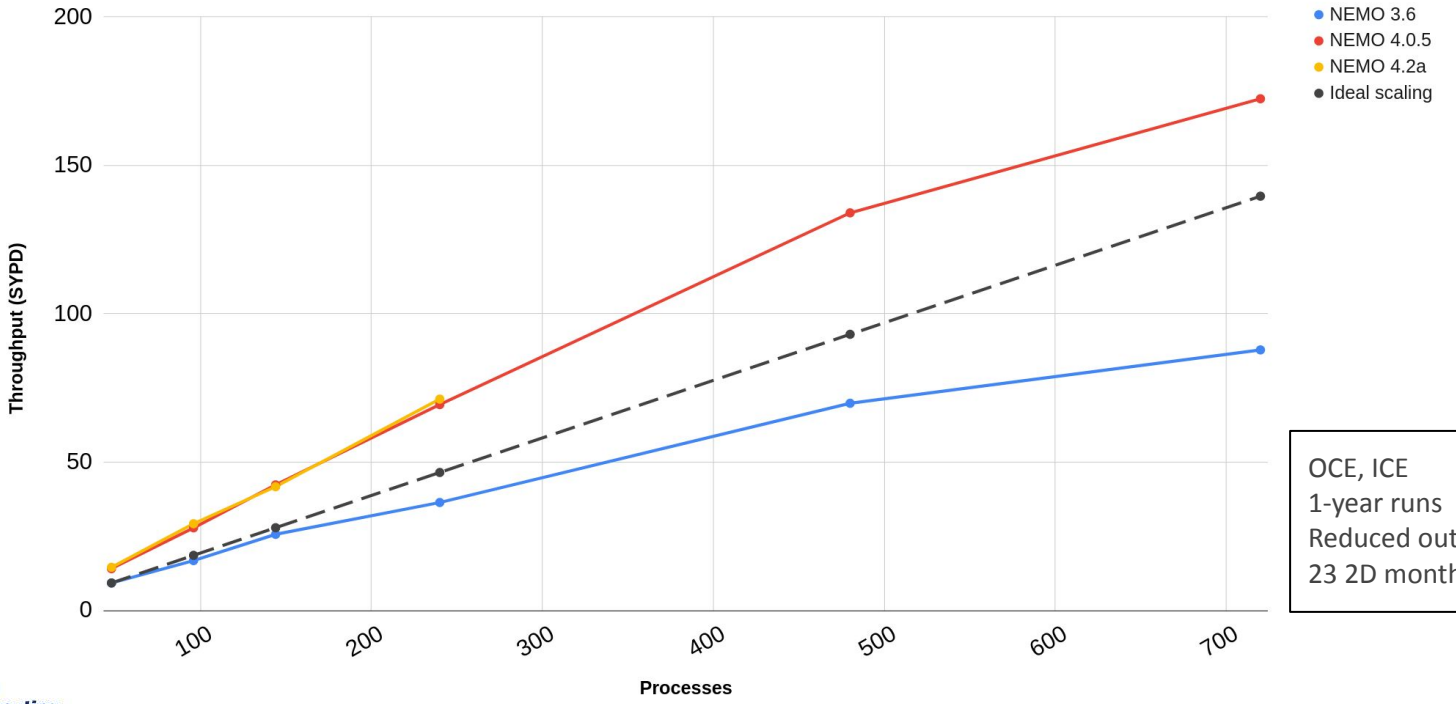
The range is a hundred times smaller!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Oriol Tintó

NEMO efficiency evolution

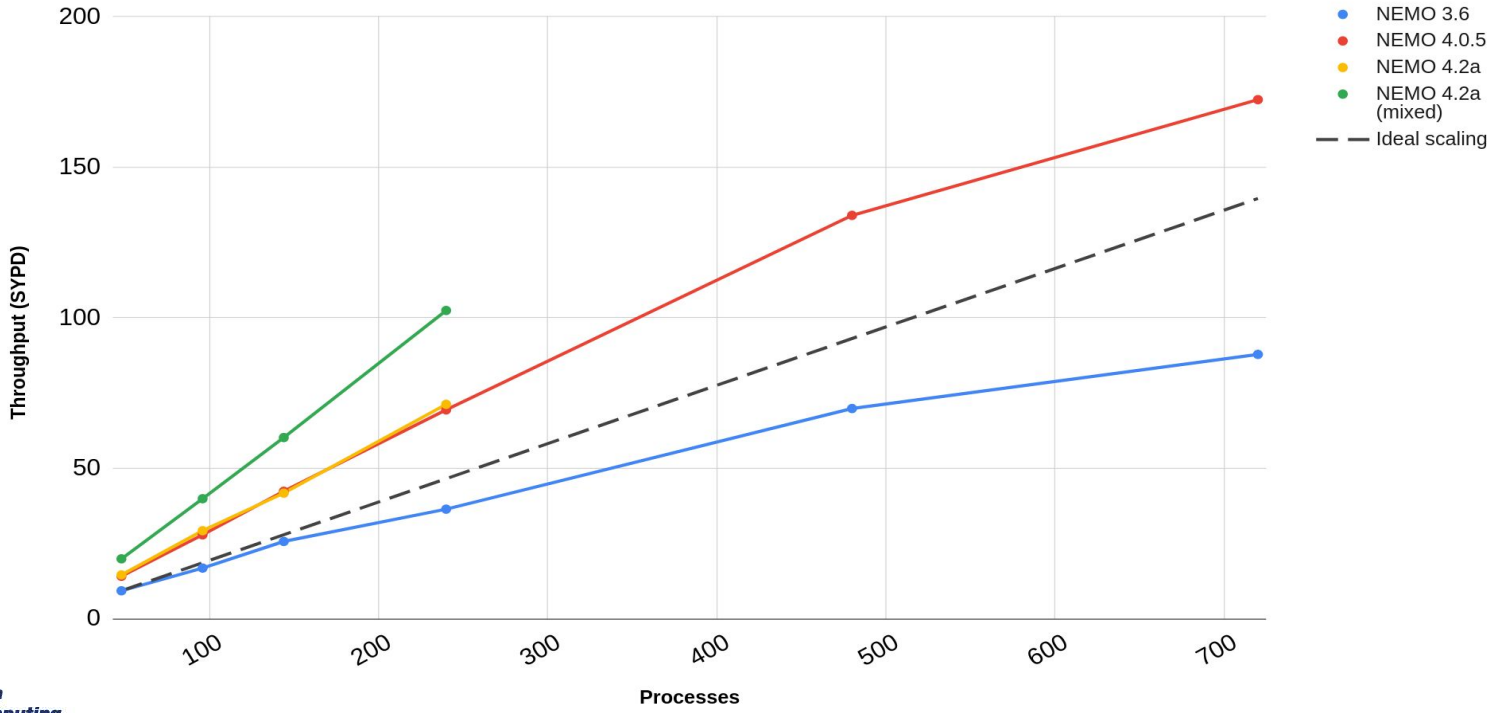
ORCA1 scalability (MareNostrum4)



OCE, ICE
1-year runs
Reduced outclass (4 3D, 23 2D monthly avgs)

NEMO efficiency evolution

*ORCA1 scalability - **Mixed precision** branch (MareNostrum4)*



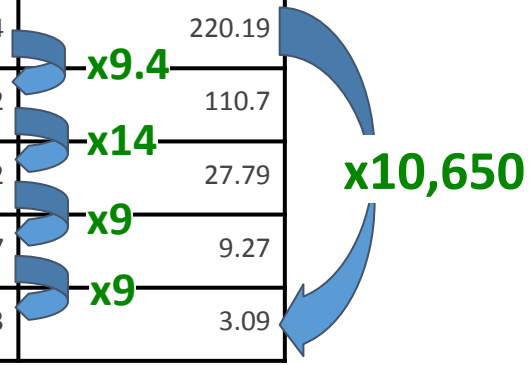
The ORCA36 configuration

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

From ORCA2 to ORCA36

- **ORCA:** Curvilinear tripolar grid family without singularity point inside the computational domain. It has two north mesh poles placed on lands.

name	jpiglo	jpjglo	jpk	size (million vertices)	resolution (km)
ORCA2	182	149	31	0.84	220.19
ORCA1 (SR)	362	292	75	7.92	110.7
ORCA025 (HR)	1,442	1,021	75	110.42	27.79
ORCA12 (VHR)	4,322	3,059	75	991.57	9.27
ORCA36 (VVHR?)	12,962	9,173	75	8,917.53	3.09



Global 1/36° (ORCA36): context

- Model configuration for future CMEMS/MOI global forecasting and reanalysis systems
- Based on NEMO 4

IMMERSE (EU H2020):

demonstrator for developments in NEMO 4 (HPC developments)

ESIWACE2 (EU H2020):

demonstrator for « production runs at unprecedented resolution on pre-exascale supercomputers »

CMEMS contract with BSC:

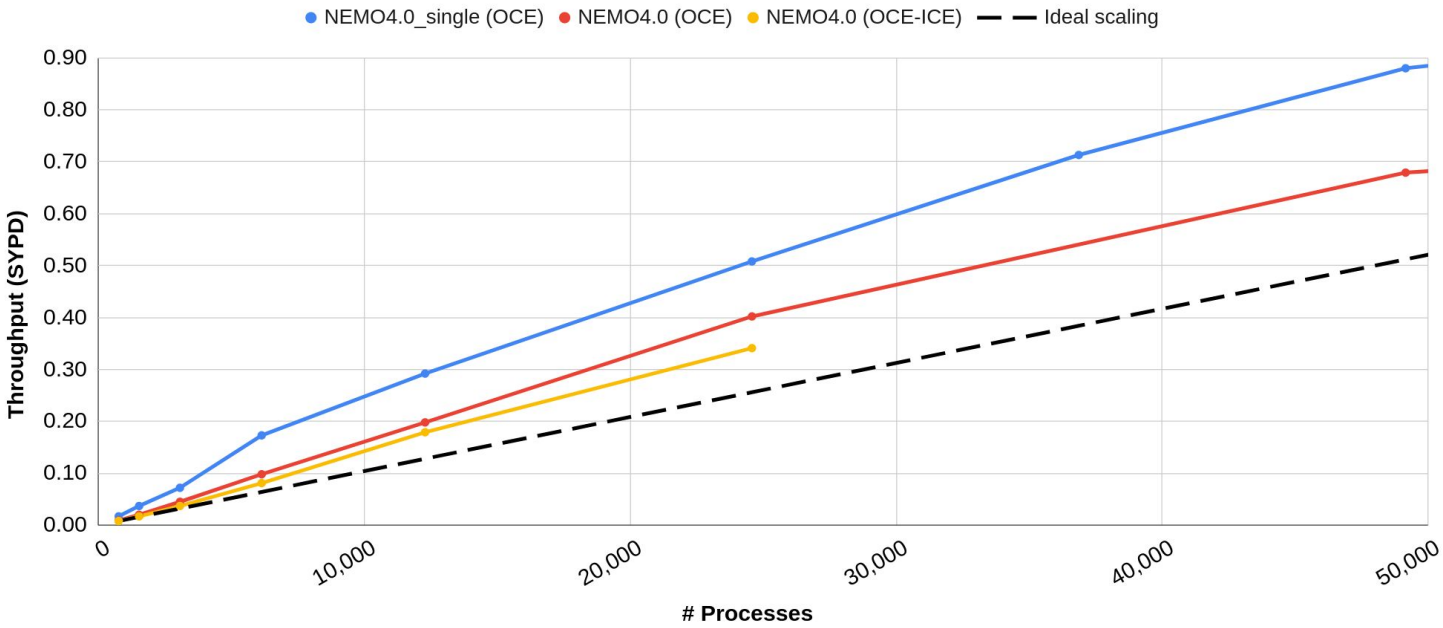
« 87-GLOBAL-CMEMS-NEMO: EVOLUTION AND OPTIMISATION OF THE NEMO CODE USED FOR THE MFC-GLO IN CMEMS » :

NEMO HPC performance, global 1/36°

Clement Bricaud (MOI)

Scaling ORCA36 - Grand Challenge 2019

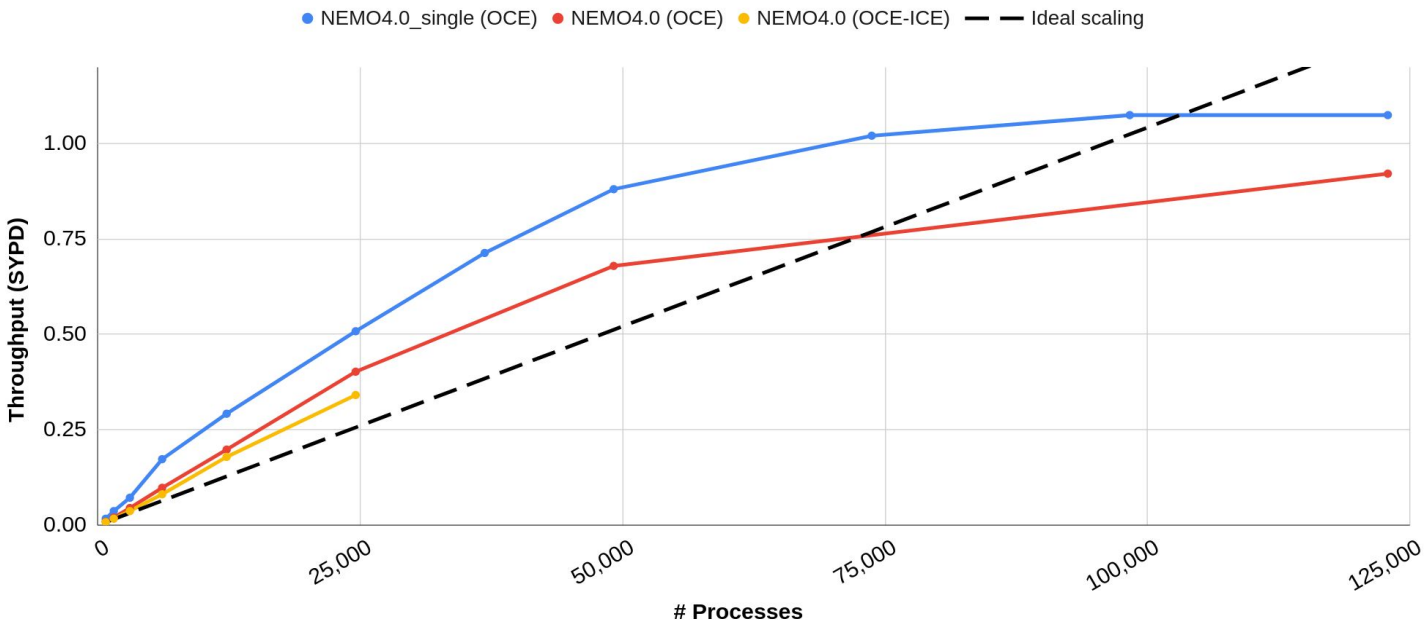
ORCA36 scalability (MareNostrum4)



NEMO r4.0
30s. timestep
300 steps
No output

Scaling ORCA36 - Grand Challenge 2019

ORCA36 scalability (MareNostrum4)



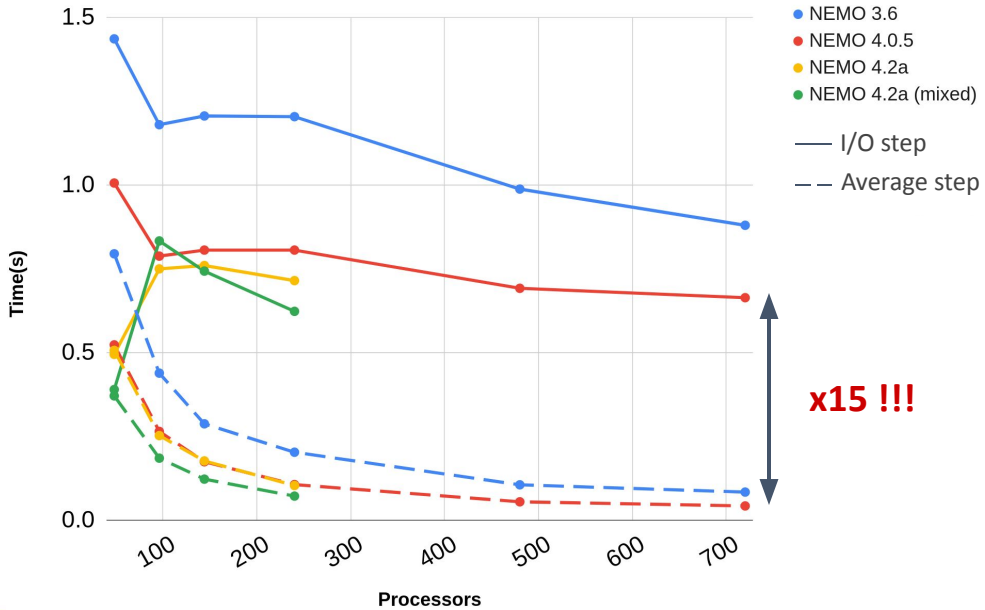
MN4 total = 165,888 cores!!

NEMO I/O


BSC
Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

NEMO I/O scalability

ORCA1 - XIOS scalability (MareNostrum4)



- Need to **scale up XIOS** with NEMO
- Performance at **higher scales**?
- **One-file** mode usable?

Constant number of I/O servers
48 XIOS servers (1 node)
Writing monthly averages
One-file mode

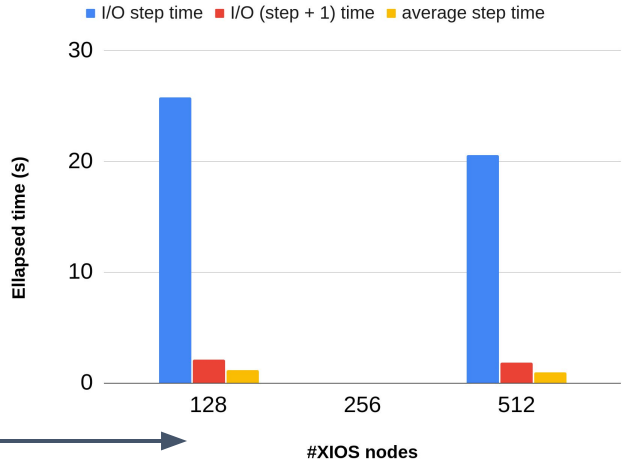
Scaling ORCA36 - Grand Challenge 2020

ORCA36 - XIOS scalability (MareNostrum4)

Performance mode
(large buffer)

- **Multiple-file** mode. One-file mode takes **minutes**!
- **High memory** needs
- XIOS not scalable

Memory mode
(large buffer)



NEMO using 512 nodes
Scaling I/O servers
5-hour 3D output (340 GB per hour)
Multiple-file mode

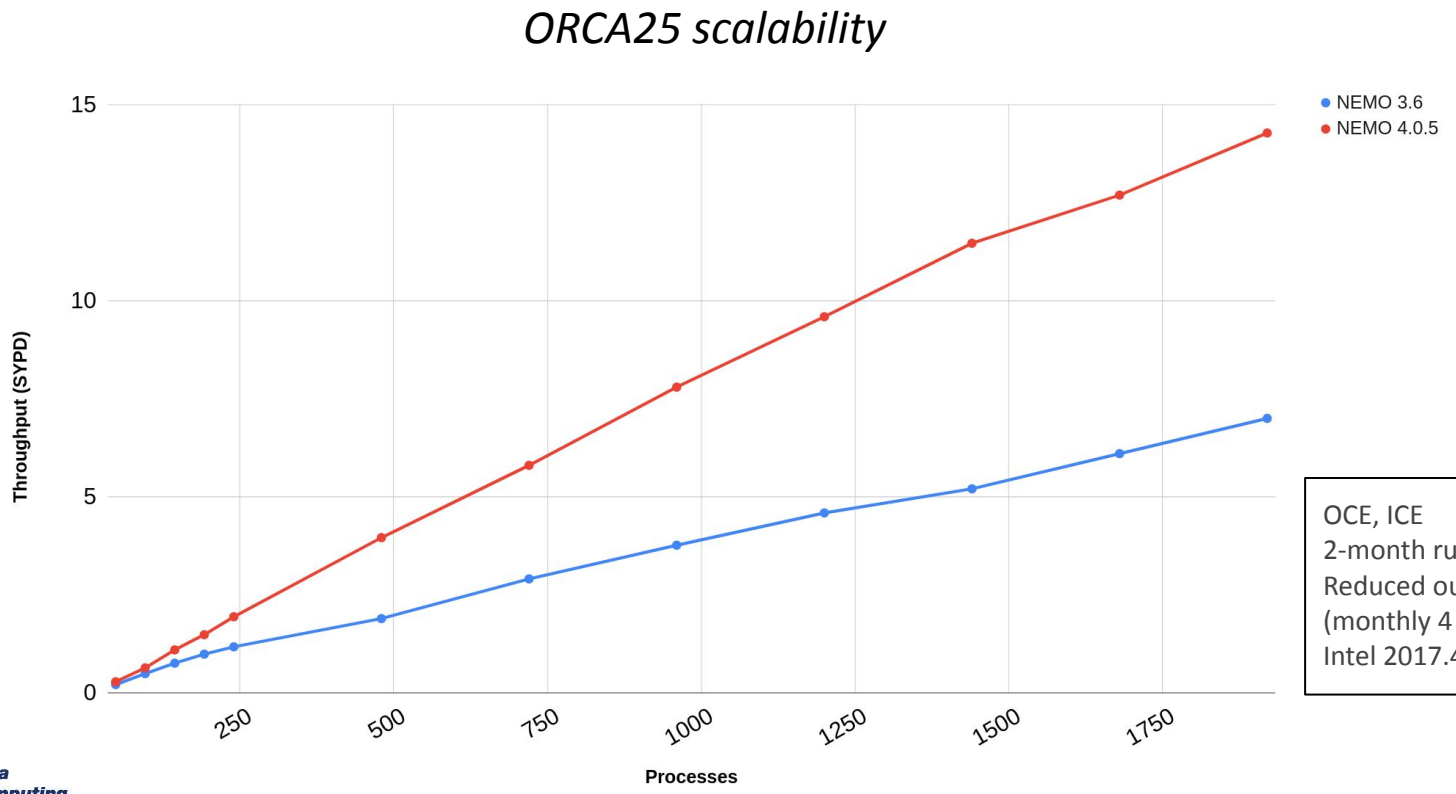
Take home messages

- **NEMO scalability** improved over the last years.
  - Max throughput at 15x15 / 10x10 subdomain size.
  - Not with **very large** configurations! (hardware limitations).
- **Using mixed precision** can help to optimize computational and memory resources:
  - ORCA 1: up to **x1.35-1.5** speedup.
  - ORCA36: possible to achieve **1 SYPD** on current architectures.
- **Production throughput depends on diverse factors:** time step size, I/O frequency, I/O size, diagnostics computation, coupling, namelist parameters…...

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 823988.

miguel.castrillo@bsc.es

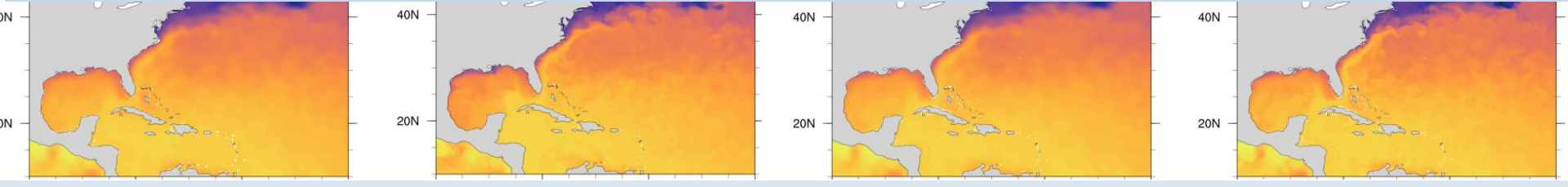
NEMO efficiency evolution



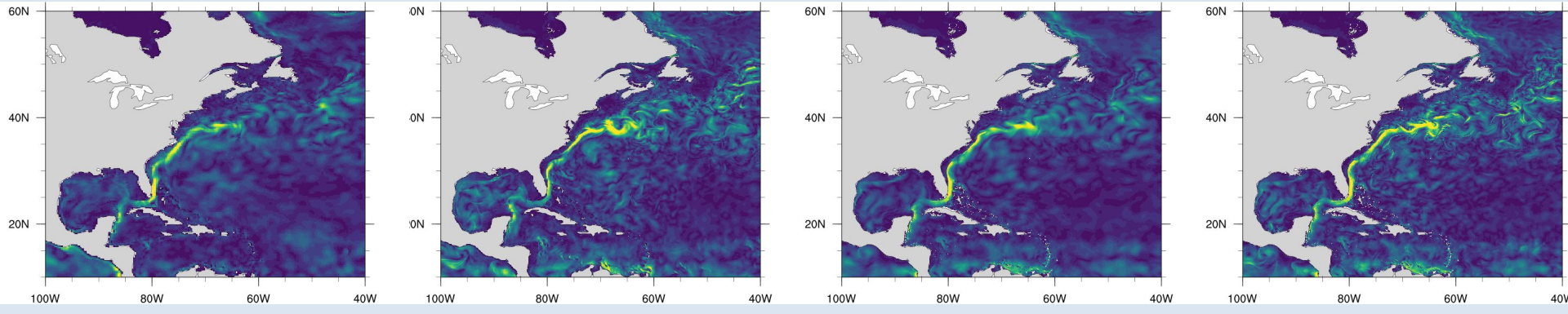
OCE, ICE
2-month runs
Reduced outclass
(monthly 4 3D, 23 2D)
Intel 2017.4; IMPI 2018.4

Interpolation from G2V4 to grid model for CI

¼° (ORCA025) | 1/12° (ORCA12) | 1/36°(ORCA36) **IC smooth** | 1/36°(ORCA36) **IC no smooth**



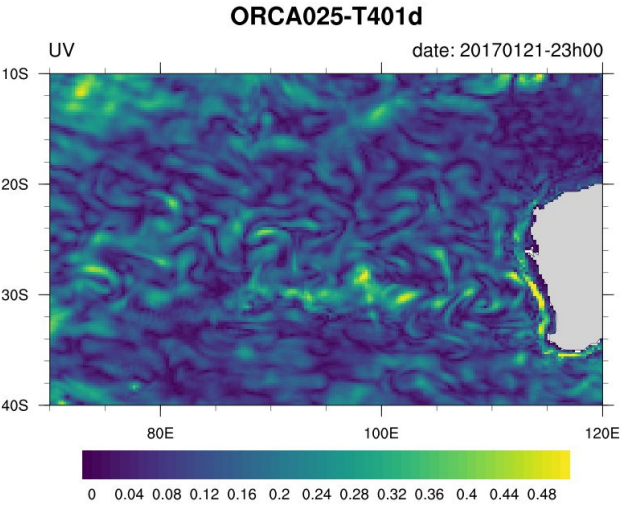
SST after 1hour



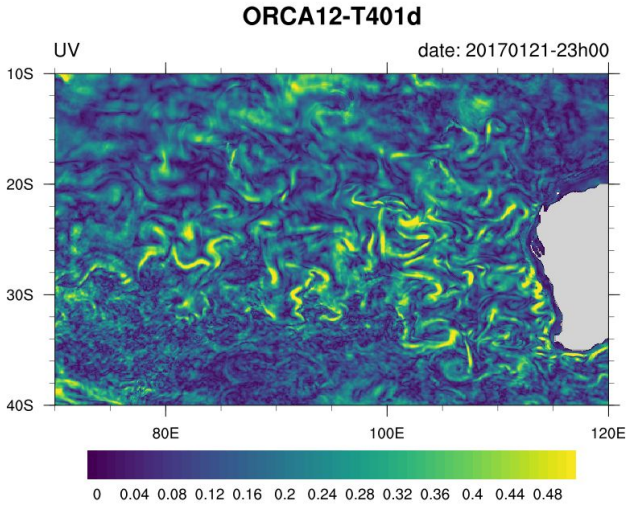
MOD(UV) after 7 days (hourly)

Surface velocities after 3 weeks (1 hour outputs)

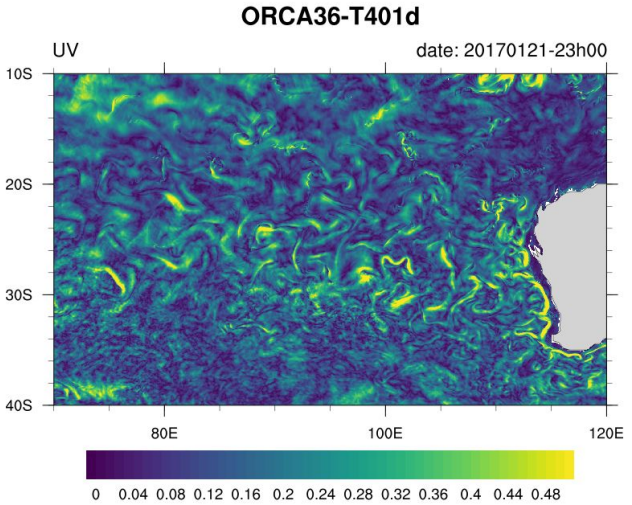
global ¼°



global 1/12°



global 1/36°



ORCA36

Configurations

Code	Step	Init T&S	Atmospheric Forcing	ICE	Runoff	Geothermal heating	QSR
O36-I	90	F	F	F	F	F	F
O36-II	90	F	512x256	F	F	F	F
O36_ICE	90	F	512x256	T	F	F	F
O36_FULL*	30	9,173x12,962	512x256	T	9,173x12,962	360x180	9,173x12,962

ORCA36 in MareNostrum4

Resources constraints

Configuration	Minimum resources standard nodes (96GB)	Minimum resources high-mem nodes (384GB)
O36-I	64 nodes, 6TB memory	16 nodes, 6TB memory
O36-II	64 nodes, 6TB memory	16 nodes, 6TB memory
O36_ICE	64 nodes, 6TB memory	16 nodes, 6TB memory
O36_FULL*	-	16 nodes, 6TB memory

The key_single

To enable compilation in mixed precision:

```
[bsc32402@login0: NEMO-4.0.5 (svn/NEMO/branches/2020/dev_r14116_HPC-04_mcastril_Mixed_Precision_implementation_final)]
$ ./makenemo -r ORCA2_ICE_PISCES -n ORCA2 -m X64_MN4 -d OCE key_add 'key_single'

You are installing a new configuration ORCA2 from ORCA2_ICE_PISCES with sub components: OCE
Creating ORCA2/WORK = OCE for ORCA2
MY_SRC directory is : ORCA2/MY_SRC
```

The key enables the following code:

par_kind.F90

```
# if defined key_single
   INTEGER, PUBLIC, PARAMETER ::   wp = sp
# else
   INTEGER, PUBLIC, PARAMETER ::   wp = dp
# endif
```

single_precision_substitute.h90

```
#if defined key_single
#   define CASTWP(x) REAL(x,wp)
#   define CASTDP(x) REAL(x,dp)
#else
#   define CASTWP(x) x
#   define CASTDP(x) x
#endif
```

ORCA36 scalability with I/O

3D hourly output

One file mode

NEMO proc.	XIOS proc.	NEMO step time	XIOS step time	**Steps/second**
1536	1536	~18s	~366s	**0.05**
3072	1536	~8s	~348s	**0.097**
3072	1920	~8s	~376s	**0.095**

Multiple file mode

NEMO proc.	XIOS proc.	NEMO step time	XIOS step time	**Steps/second**
1536	1536	~18s	~17s	**0.056**
3072	1536	~8s	~17s	**0.122**