Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Scaling NEMO4 I/O with the new ORCA36 configuration

Miguel Castrillo

BSC-ES Computational Earth Sciences

Telco on XIOS current developments

10/12/2020

The NEMO4 ORCA36 benchmark

Global 1/36° (ORCA36): context

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

IMMERSE (EU H2020):

demonstrator for developments in NEMO 4 (HPC dvpts)
with CMCC and Ocean-Next

ESIWACE2 (EU H2020):

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

Courtesy of C. Bricaud (MOI)

Global 1/36° (ORCA36): context

- Collaborations:

CMEMS contract with BSC:

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

NEMO HPC performances, especially with global 1/36°

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

CMEMS contract with CNRS/IGE/MEOM team:

« 2-GLO-HR Evolution of CMEMS Global High Resolution MFC »

Institut des Géosciences de l'Environnement

- sensitivity of NEMO solutions to numerical and parametric choices in realistic configurations

an Atlantic (20S-81N) 1/12° configuration with AGRIF zooms (1/12° to 1/48° and 75 to 200 vertical levels)

- Definition of metrics to assess resolved fine-scale structures

Small scale vorticity variance, KE wavenumber spectra, regularity of resolved fields at the grid scale,submesoscale vertical buoyancy flux, fine scale horizontal gradient of surface buoyancy

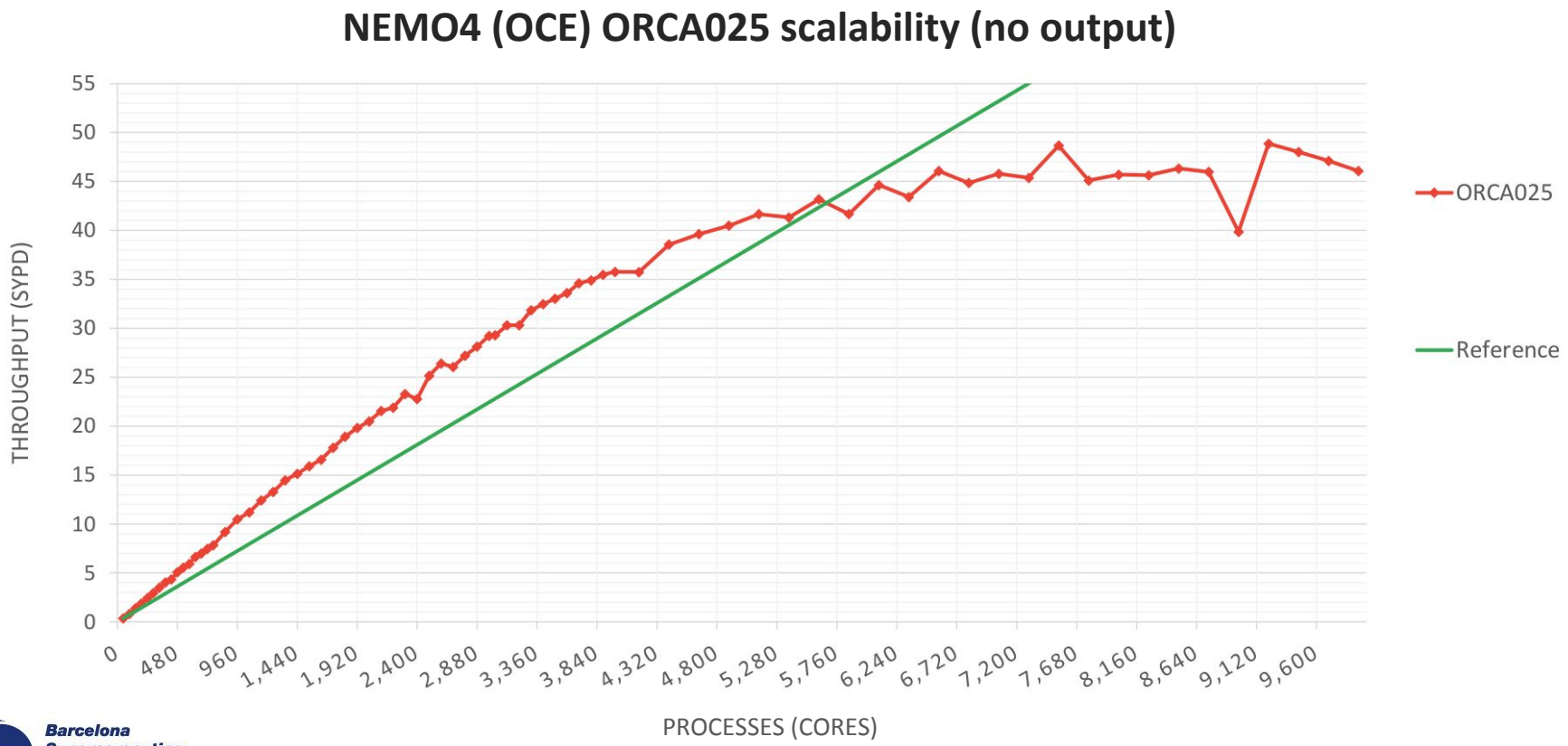
From ORCA2 to ORCA36

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

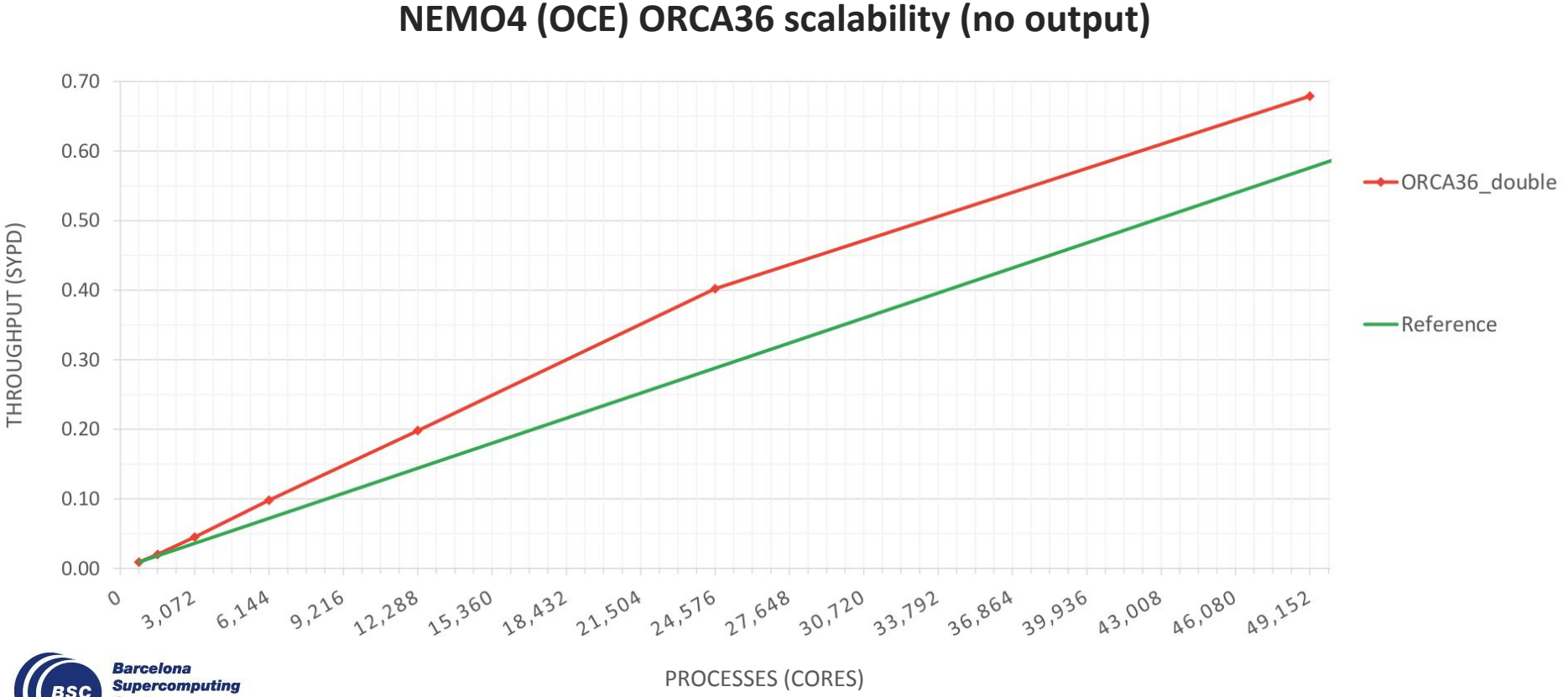
name	jpiglo	jpjglo	jpk	size (million vertices)	resolution (km)
ORCA2	182	149	31	0.84	220.19
ORCA1 (SR)	362	292	75	7.92	110.7
ORCA025 (HR)	1,442	1,021	75	110.42	27.79
ORCA12 (VHR)	4,322	3,059	75	991.57	9.27
ORCA36 (VVHR?)	12,962	9,173	75	8,917.53	3.09

x9.4
x14
x9
x9
x10,650

NEMO4 scalability in MareNostrum4

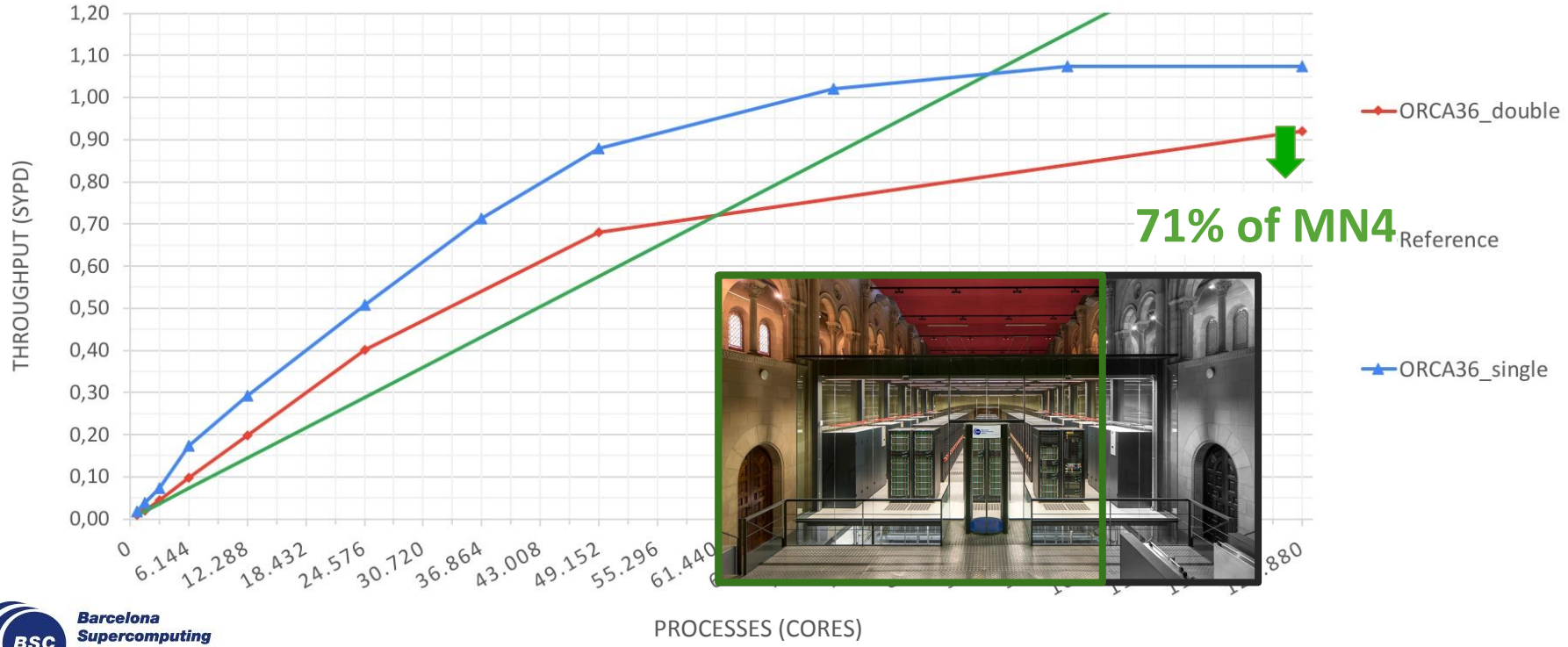


NEMO4 scalability in MareNostrum4



NEMO4 scalability in MareNostrum4

NEMO4 (OCE) ORCA36 scalability – Double vs Single precision – Grand challenge 2019



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

x100 ORCA025, but only scaling up to x10 resources

Adding output to the ORCA36 benchmark

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

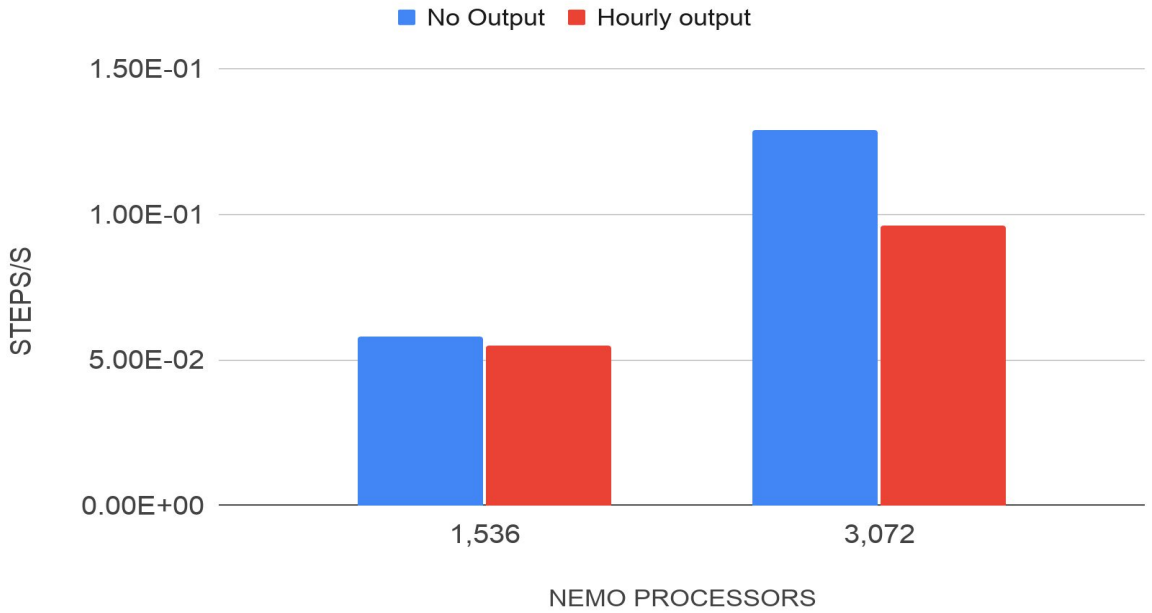
ORCA36 scalability with I/O

Test description

- NEMO 4.0 running with XIOS 2.5.
- **OCE** and **ICE** modules (Blue Ocean and Ice in the poles being simulated).
- MareNostrum4 supercomputer, Intel 2017.4 compiler and Intel MPI 2018.4.
- ORCA36 **configuration** provided by Mercator International, CMEMS project.
- **30 seconds** timestep for NEMO. (Clement B. using 120s in “production mode”).
- 2-hour tests (240 steps).
- **Memory mode** used for XIOS (conservative approach, smaller buffer).
- XIOS and NEMO running on independent high-memory nodes (they do not share nodes).

ORCA36 scalability with I/O

NEMO-XIOS ORCA36 scalability. No output vs 3D hourly output. First results.



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Hourly output runs: 1,536 XIOS servers in 32 high-memory nodes (8GB/core). One-file mode. 340GB per simulated hour.

ORCA36 scalability with I/O

No output / 2D output

No output

NEMO proc.	XIOS proc.	NEMO step time	XIOS step time	**Steps/second**
1536	1536	~17s	-	**0.058**
3072	1536	~8s	-	**0.129**

2D variables (one file mode)

NEMO proc.	XIOS proc.	NEMO step time	XIOS step time	**Steps/second**
1536	1536	~17s	~43s	**0.058**
3072	1536	~8s	~34s	**0.126**

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**

ORCA36 scalability with I/O

Some questions to answer

Multiple file mode reduces the overhead significantly. But NEMO time per step can still be much smaller (by factor of 10 in MareNostrum4):

- May we scale by adding **more processing elements** (servers)?
- Can we reduce the wait (XIOS step) by using **performance** mode?
- Can we run NEMO and XIOS processes in the **same nodes** and reduce the overhead (less inter-node comms)? Memory may be an issue.
- Can we speed up the executions by writing in the **local disk** instead of using GPFS?
- Can we benefit from using **Level-2** servers?

Grand Challenge 2020

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability with I/O

Grand challenge executions (2020)

- NEMO **4.0.2** and XIOS 2.5 r1903.
- From 24k to **100K** cores.
- **Intel** MPI and **Open** MPI environment.
- **Multiple-file** mode.
- Not using high-memory nodes anymore.
- Test if the I/O overhead can be reduced by **adding more servers** and/or using **performance mode**.

Outcome of the tests (detailed)

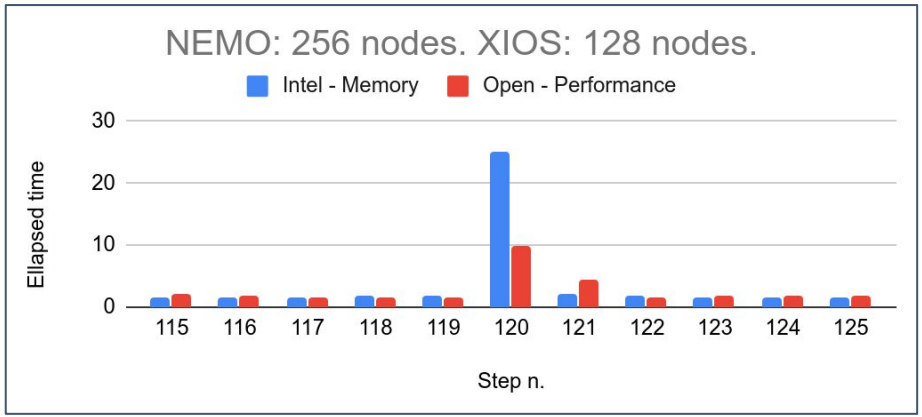
NEMO nodes (proc)	XIOS nodes (proc)	MPI	Total runs	Start	OK
256 (12,288)	128 (256)	Intel	4	4 (100%)	0
256 (12,288)	128 (256)	Open	2	2 (100%)	2 (100%)
256 (12,288)	256 (512)	Intel	4	2 (50%)	0
256 (12,288)	256 (512)	Open	2	1 (50%)	1 (50%)
512 (24,576)	128 (256)	Intel	3	3 (100%)	0
512 (24,576)	128 (256)	Open	2	2 (100%)	2 (100%)
512 (24,576)	256 (512)	Intel	4	1 (25%)	0
512 (24,576)	256 (512)	Open	2	1 (50%)	1 (50%)
512 (24,576)	512 (1,024)	Intel	4	4 (100%)	1 (25%)
512 (24,576)	512 (1,024)	Open	4	3 (75%)	2 (50%)

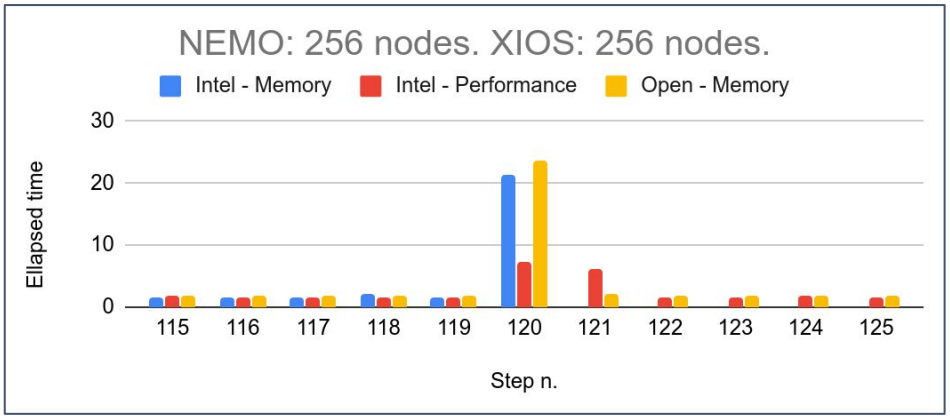
128 nodes = 12T memory. This is the minimum possible for this configuration (=32 memory nodes).

Outcome of the tests (detailed)

NEMO nodes (proc)	XIOS nodes (proc)	MPI	Total runs	Start	OK
1,024 (49,152)	512 (1,024)	Intel	2	0	0
1,024 (49,152)	512 (1,024)	Open	2	0	0
1,024 (49,152)	1,024 (2,048)	Intel	2	0	0
1,024 (49,152)	1,024 (2,048)	Open	2	0	0

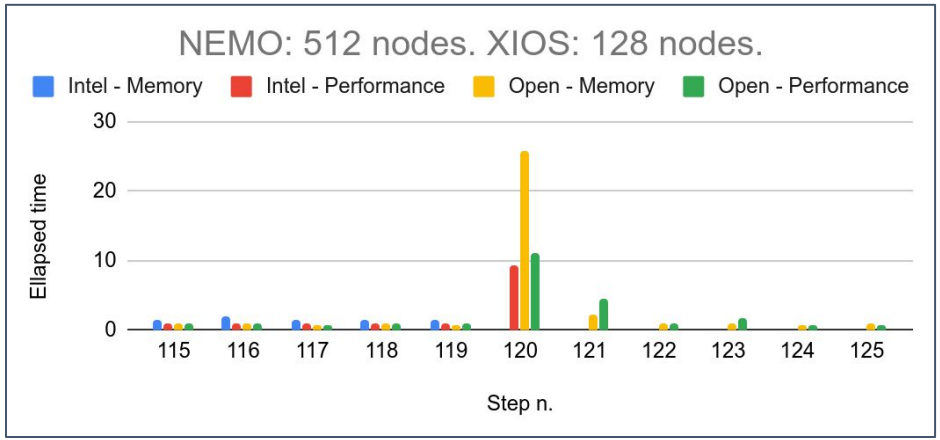
NEMO: 256 nodes (12,288 processes)

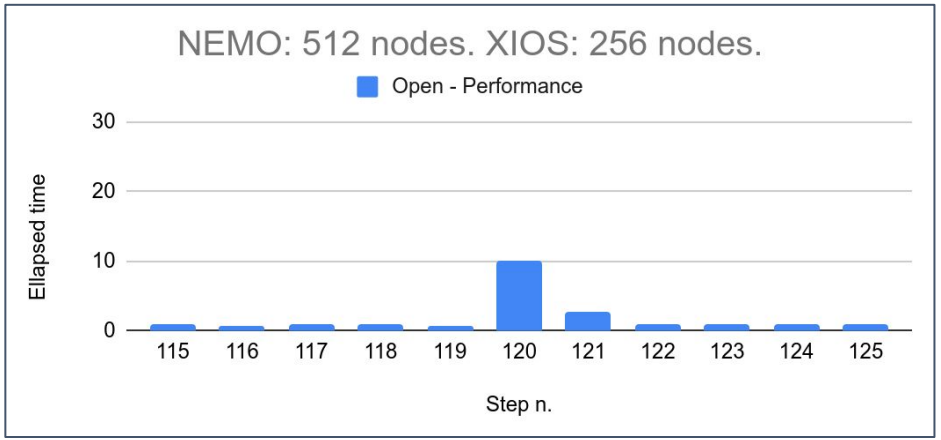


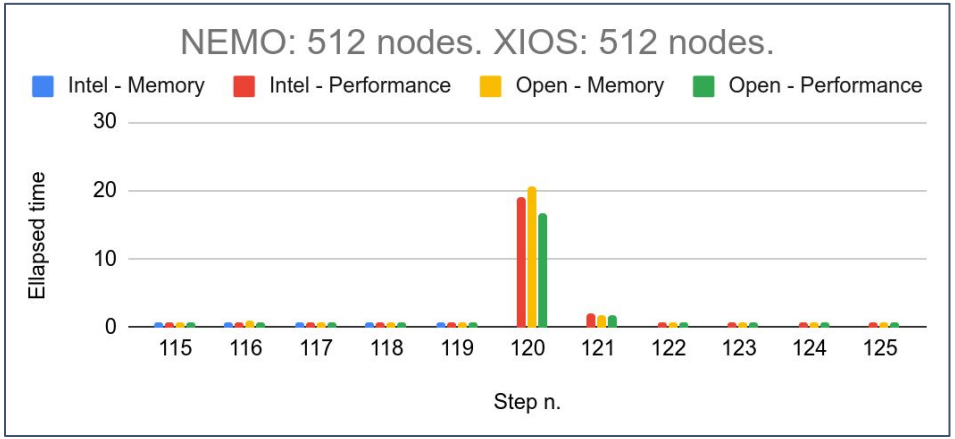


Intel MPI slightly better than Open MPI. Performance mode reduces time by x1.5 - x2.

NEMO: 512 nodes (24,576 processes)

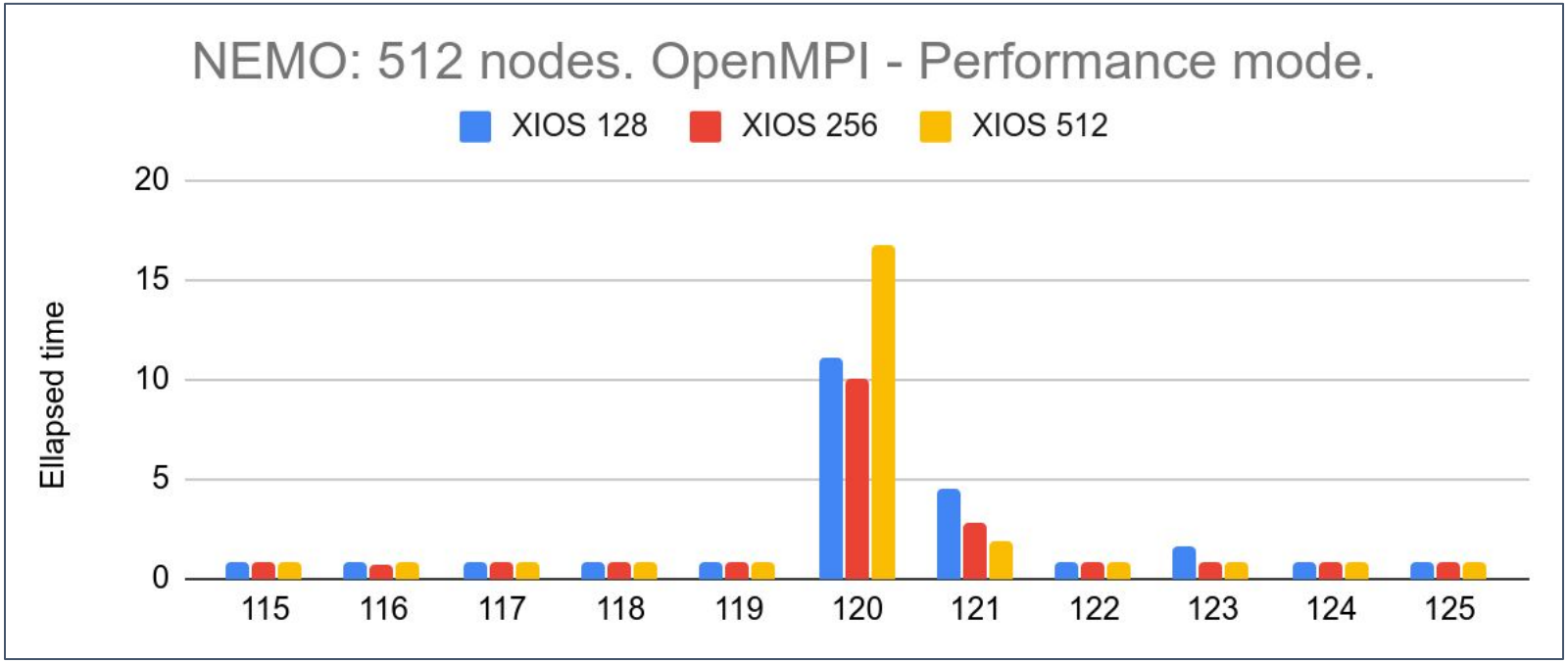






BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NEMO: 512 nodes (24,576 processes)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Adding 12-17 seconds every 120 steps (0.85 s per step). 12-17% overhead.
Adding 12-17 seconds every 30 steps (assuming +50% per step). 23-45% overhead??

Conclusions

- We are **not ready** to run NEMO ORCA36 - XIOS in MN4 with so much nodes / cores (it is, with a more modest number like ~64 highmem nodes).
- An issue in XIOS 2.5 **is the memory needs.** Communications involved it's also a factor to take into account. It was not possible to run with 49,152 NEMO processes (100T memory for NEMO and 100T for XIOS).
- Writing time can be reduced using a **bigger buffer**.
- At this point it seems difficult to reduce I/O time by just adding more resources.
- Using an efficient NEMO configuration (512 nodes). I/O overhead → 20-40%.
- More tests are needed, maybe in different conditions to see if these results stand, including affinity tests, using local storage, etc.

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

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

esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER AND CLIMATE IN EUROPE

miguel.castrillo@bsc.es

Outcome of the tests

- Total runs: 44
- Started: 22
- Completed: 8
- Failed: 36
  - Initialization: 22
    - Reading input files: 6
    - Memory: 6 (Always **1,024 nodes**)
    - Hung (time limit): 4
    - Aborted (no error): 4
    - Connection issues: 2
  - Writing step: 10
    - Connection issues: 9
    - Memory: 1
  - Other step: 2