BSC-ES Performance team updates from IMMERSE and ESiWACE2 H2020 projects

Miguel Castrillo

BSC-ES Performance Team, Computational Earth Sciences

NEMO HPC Group

22/04/2020

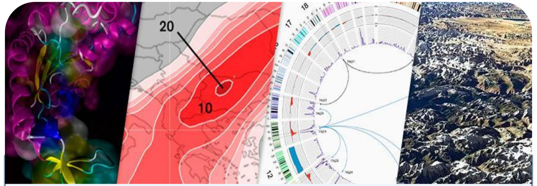
The Performance Team in BSC Earth Sciences

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-CNS objectives

Supercomputing services
to Spanish and EU researchers



R&D in Computer, Life, Earth
and
Engineering Sciences

PhD programme, technology
transfer, public engagement

BSC-CNS is a consortium that includes

Spanish Government	60%	Gobierno de España, Ministerio de Ciencia, Innovación y Universidades
Catalan Government	30%	Generalitat de Catalunya Departament d'Empresa i Coneixement
Univ. Politècnica de Catalunya (UPC)	10%	Universitat Politècnica de Catalunya BarcelonaTech

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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 investiment

EuroHPC Joint Undertaking

Hosting Consortium:

Spain **Portugal** **Turkey** **Croatia**

European Commission

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

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

BSC-ES involvement in ESiWACE2

- **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 (~8 km) ocean
- **Task 1.3: Port models to pre-exascale EuroHPC systems**
  - Port EC-Earth ~10km to MareNostrum5

EC-Earth4

- Development of a **mixed precision mode for NEMO 4.2**. Port **EC-Earth3 Ocean** configurations.
- Set up **PISCES for NEMO 4** (plus Age and CFCs). Pending on funding, develop **PISCES-lite** for HR.
- **XIOS integration into OpenIFS 43r3**. XIOS **benchmarking** and **computational evaluation** to improve I/O efficiency.
- Scientific and computational **evaluation** (in collaboration with other institutions) of **EC-Earth4** in **MP mode** (OpenIFS-SP, NEMO-MP).
- **Profiling studies** to ensure that the eventual main bottlenecks of EC-Earth4 are highlighted and their solution studied.

NEMO Mixed Precision implementation

a) BSC developing a **prototype** based on NEMO 4.0.1; Further testing done at ECMWF with this version.

b) Provide demonstrator for NEMO ST (how? => Scientific Publication, Branch in NEMO rep.) (BSC)

c) Prepare minimal list of changes for merge party that could enter next NEMO version (BSC+ ECMWF together)

d) Prepare **tools for automation** to make sure that all the workflow that goes from identifying sensitive variables to a final implementation can be reapplied to any version of the code. (BSC, April 2020)

e) Provide a **mixed-precision branch** at the NEMO SVN starting after the merging party. (BSC, April 2020)

f) Provide **scientific evaluation and progress** idealised test cases motivated by single precision evaluation and beyond (throughout 2020) (ECMWF); (Discussion on intercomparison test cases at Commodore Workshop, January 2020).

g) By end 2020, ECMWF anticipates to have a **working version of NEMO + SI3 in mixed precision** for DA, medium-range and extended-range ocean prediction in forced and coupled mode. Potential changes will be included into the Mixed-precision branch. (ECMWF+BSC task).

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NEMO Mixed Precision implementation

After the sensitive variables have been identified it is time to think on the actual implementation of the code.

Highlights:

- We are using a new **key_single:**
  - Activating this key will set the **working precision to single**.
  - Compiling without this key will yield a version of the code in double precision, that **shall pass all the sette tests**.

IMMERSE

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

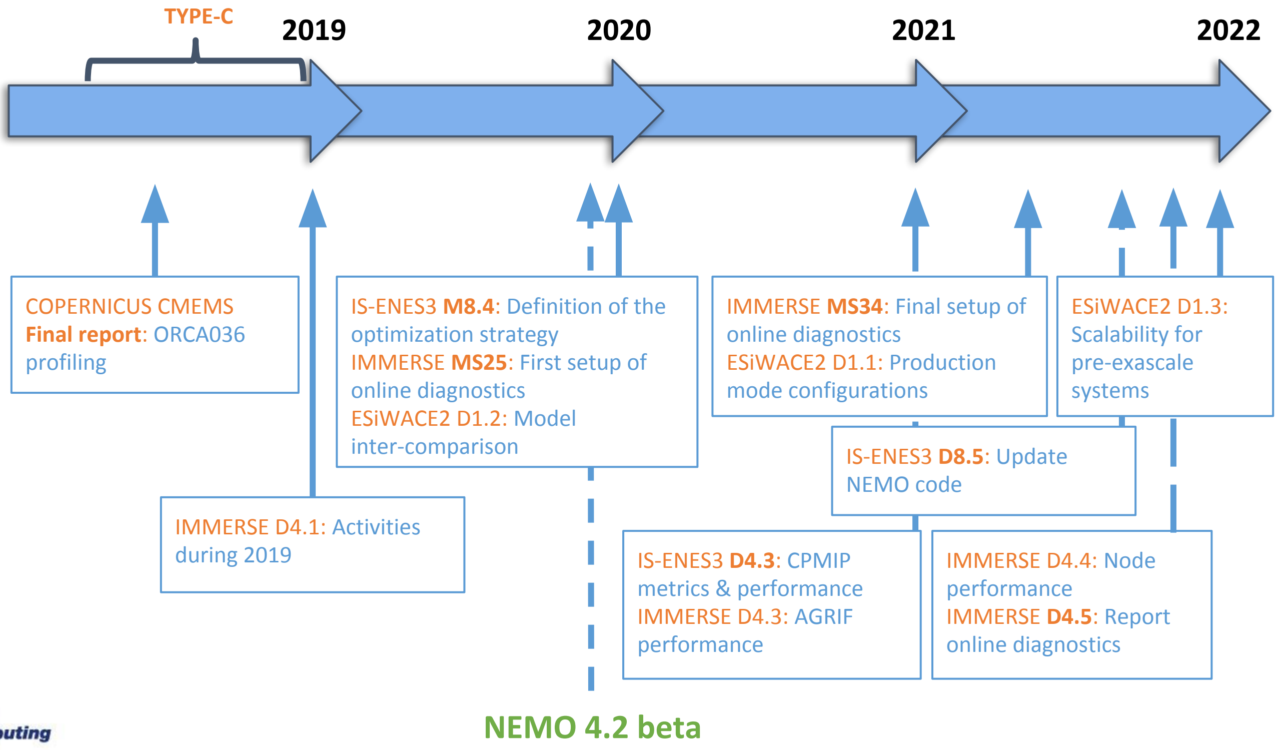
BSC-ES involvement in IMMERSE

- **T4.1:** Efficient exploitation of memory hierarchies and hardware peak performance
  - **Assessment** of the **performance impact**
- **T4.3:** Efficient IOs and diagnostics for operational systems
  - **Offload** NEMO model **diagnostics to GPUs**
- **T4.4:** Load balancing for AGRIF massive multigrid capability
  - **Efficiency assessment** for **high-resolution** configuration

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

immerse
IMPROVING OCEAN MODELS FOR THE COPERNICUS PROGRAMME

NEMO timeline in BSC-ES performance



* Bold indicates BSC is the lead

From ORCA2 to ORCA36

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

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

MERCATOR OCEAN INTERNATIONAL

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 »

IGE 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

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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 in MareNostrum4

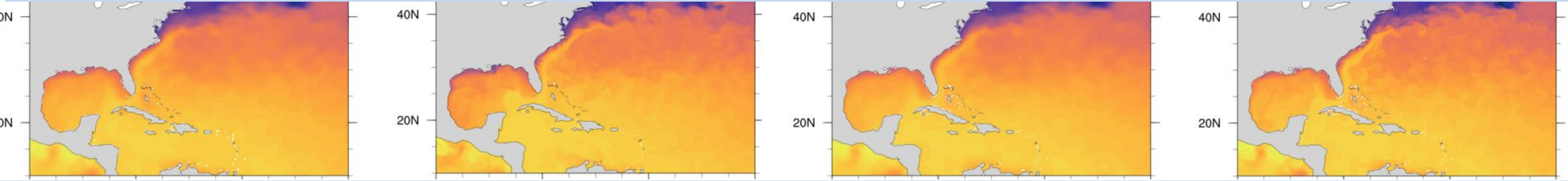
Resources constraints

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

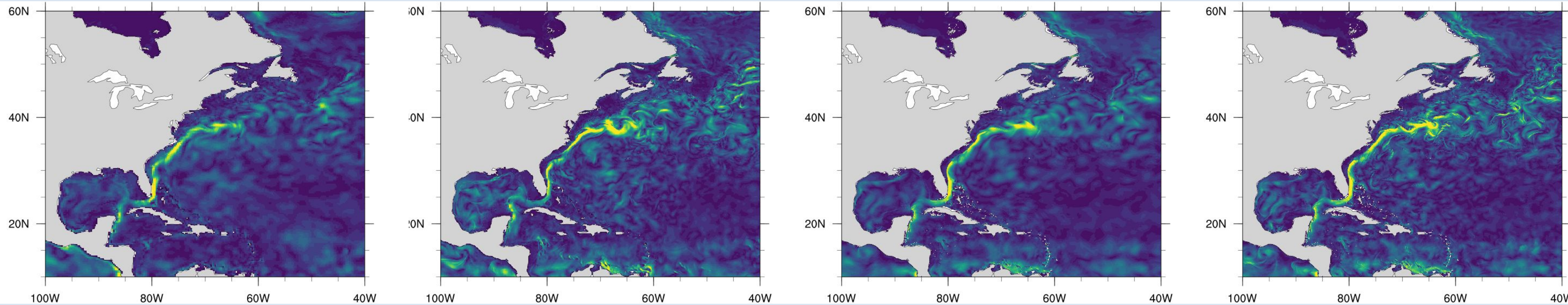
MERCATOR OCEAN INTERNATIONAL

Interpolation from G2V4 to grid model for CI

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



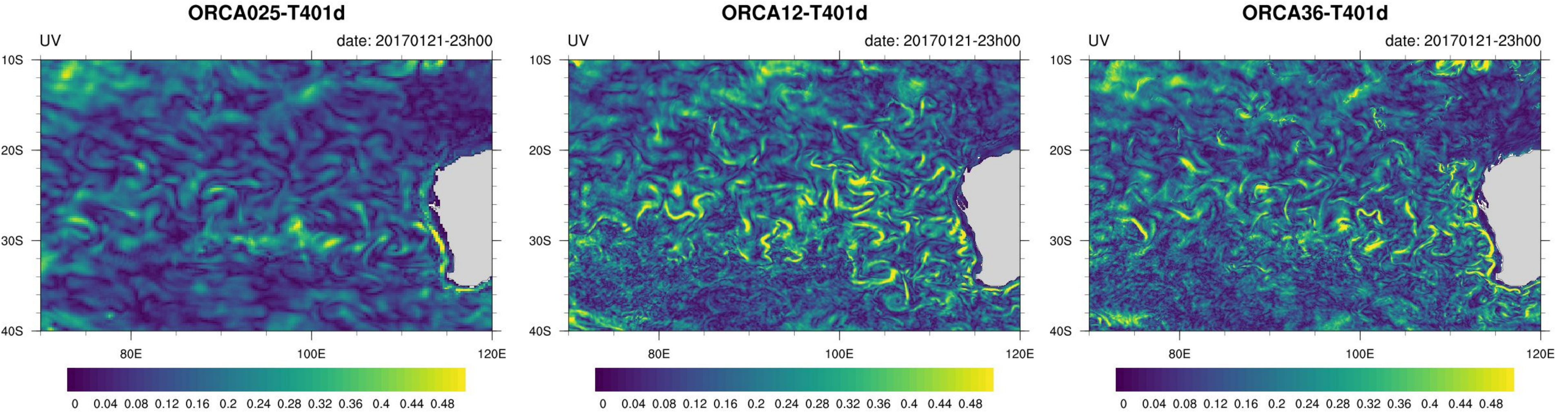
SST after 1hour



MOD(UV) after 7 days (hourly)

Surface velocities after 3 weeks (1 hour outputs)

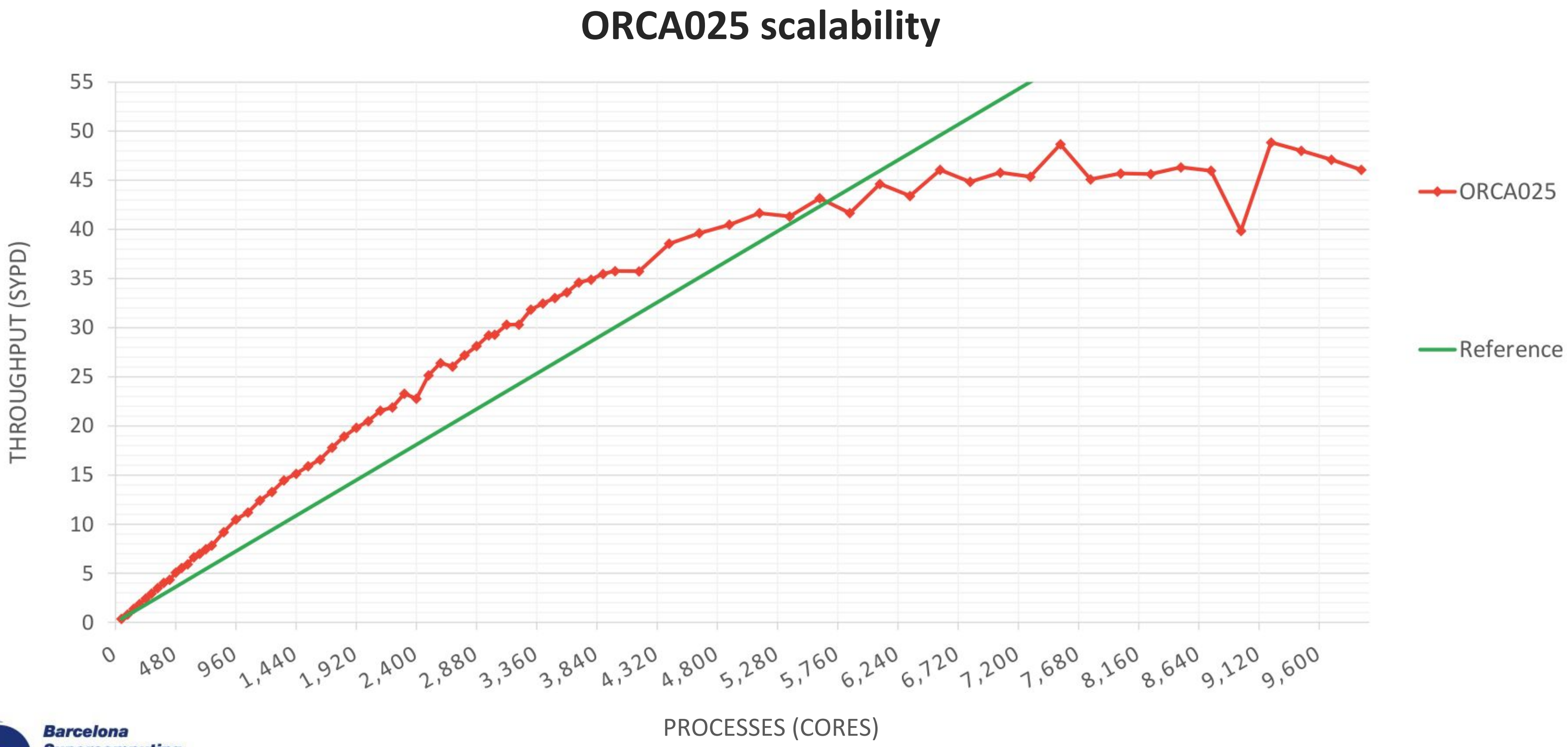
global ¼° | global 1/12° | global 1/36°



ORCA36 scaling

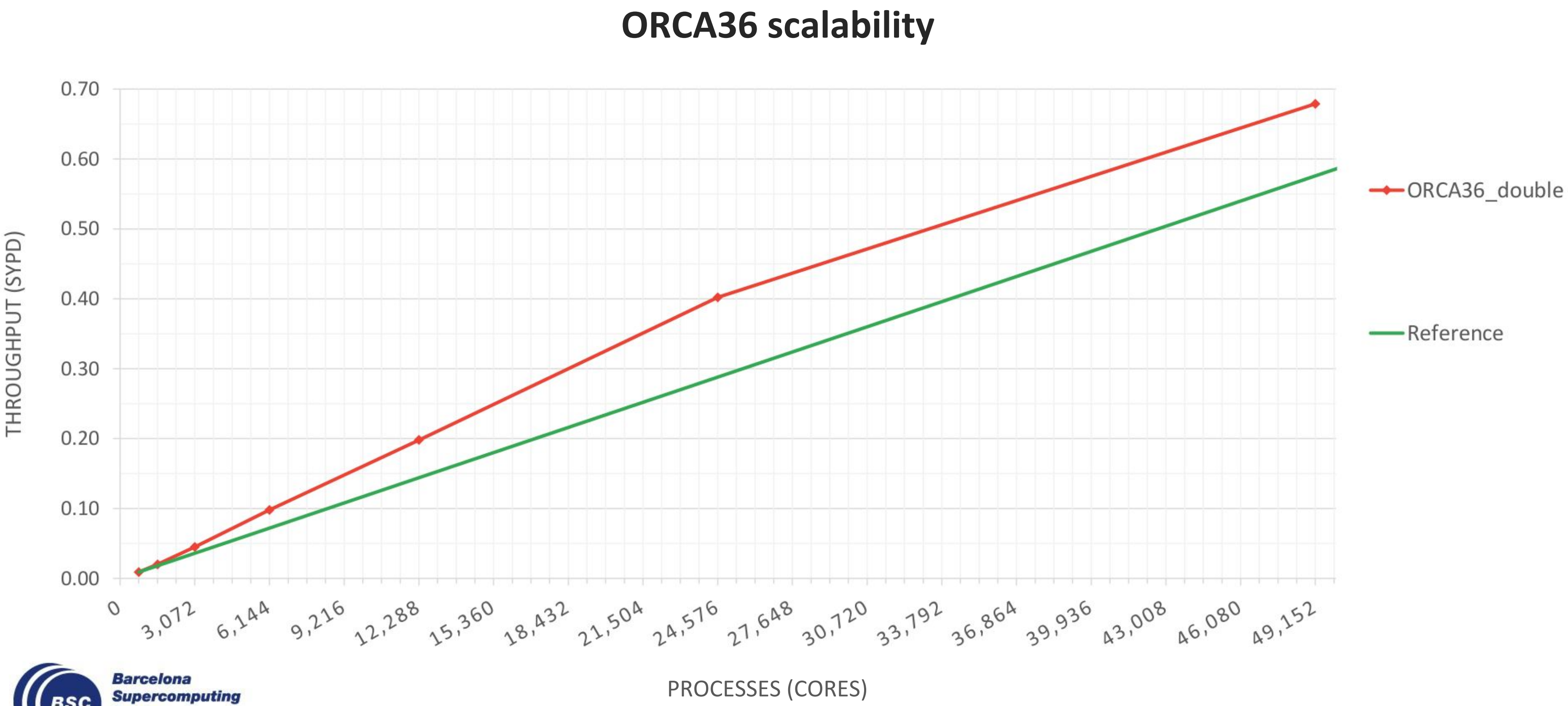
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA025 scalability (MN4)

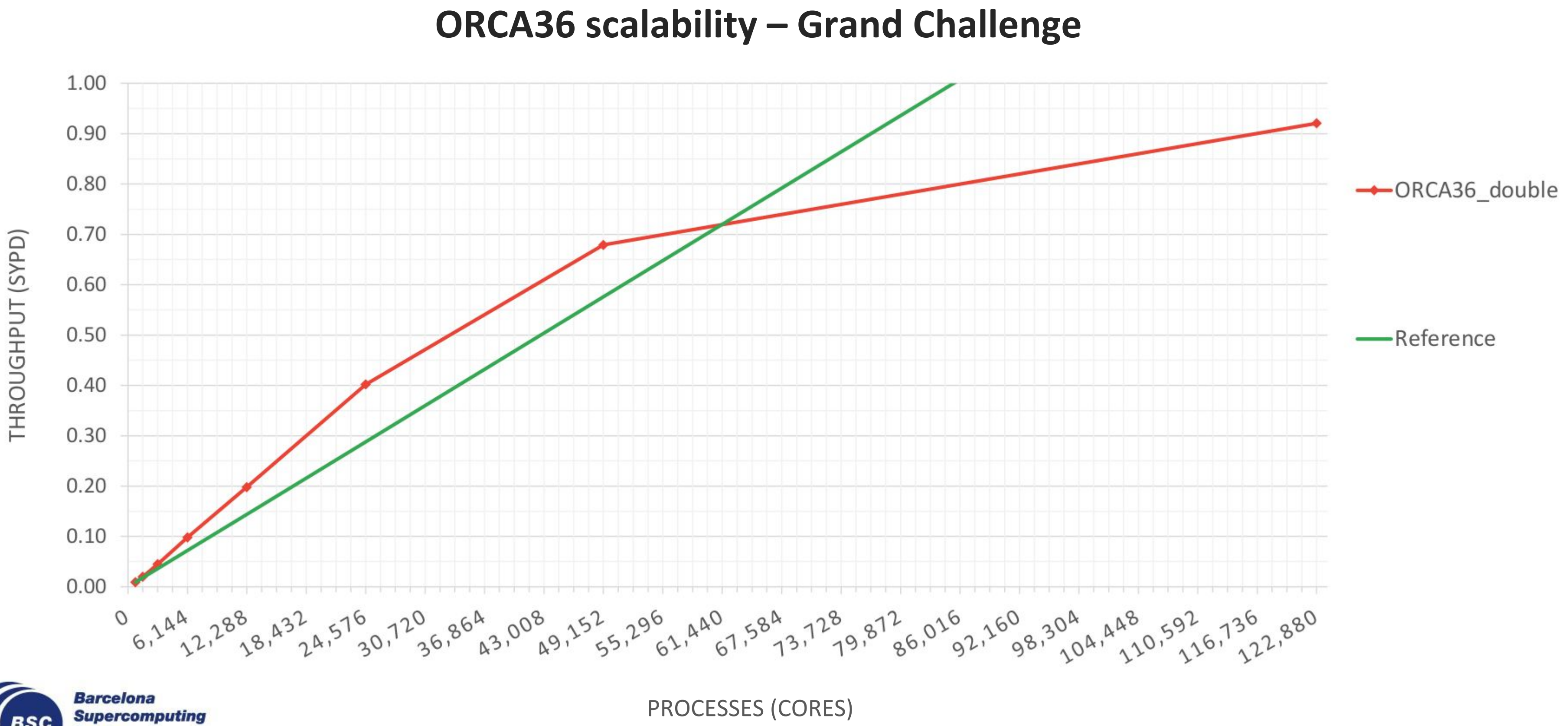


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability (MN4)

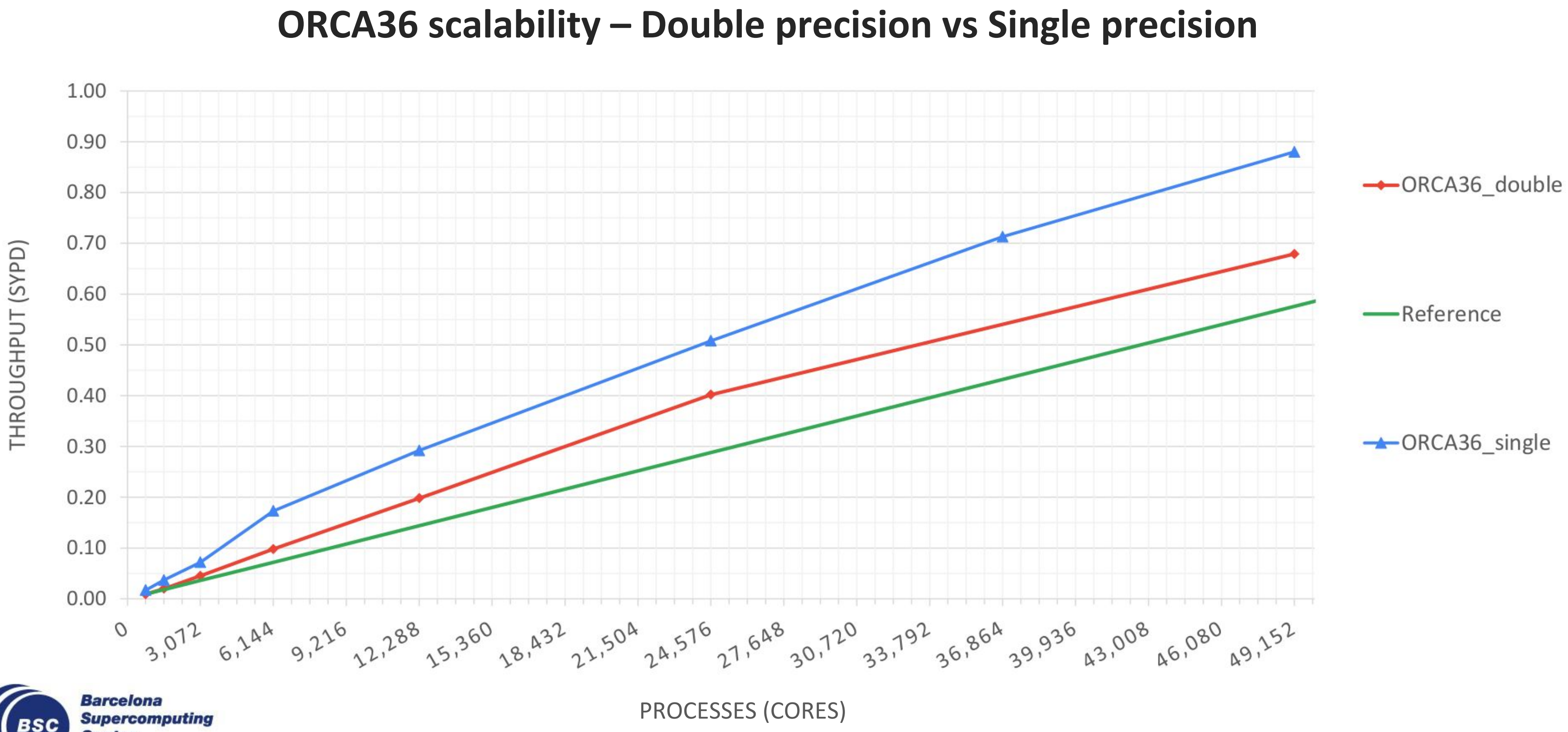


ORCA36 scalability (MN4)



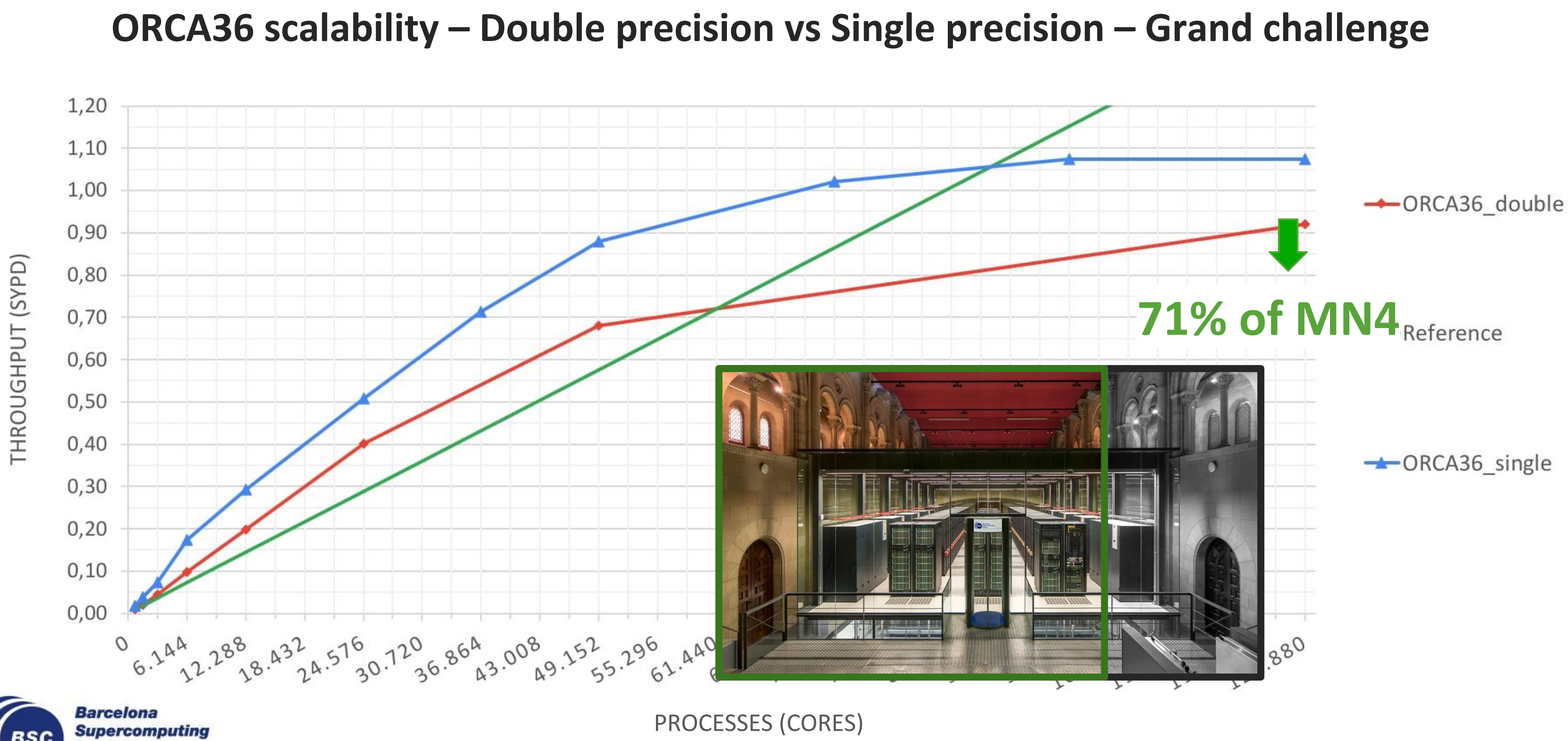
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability (MN4)



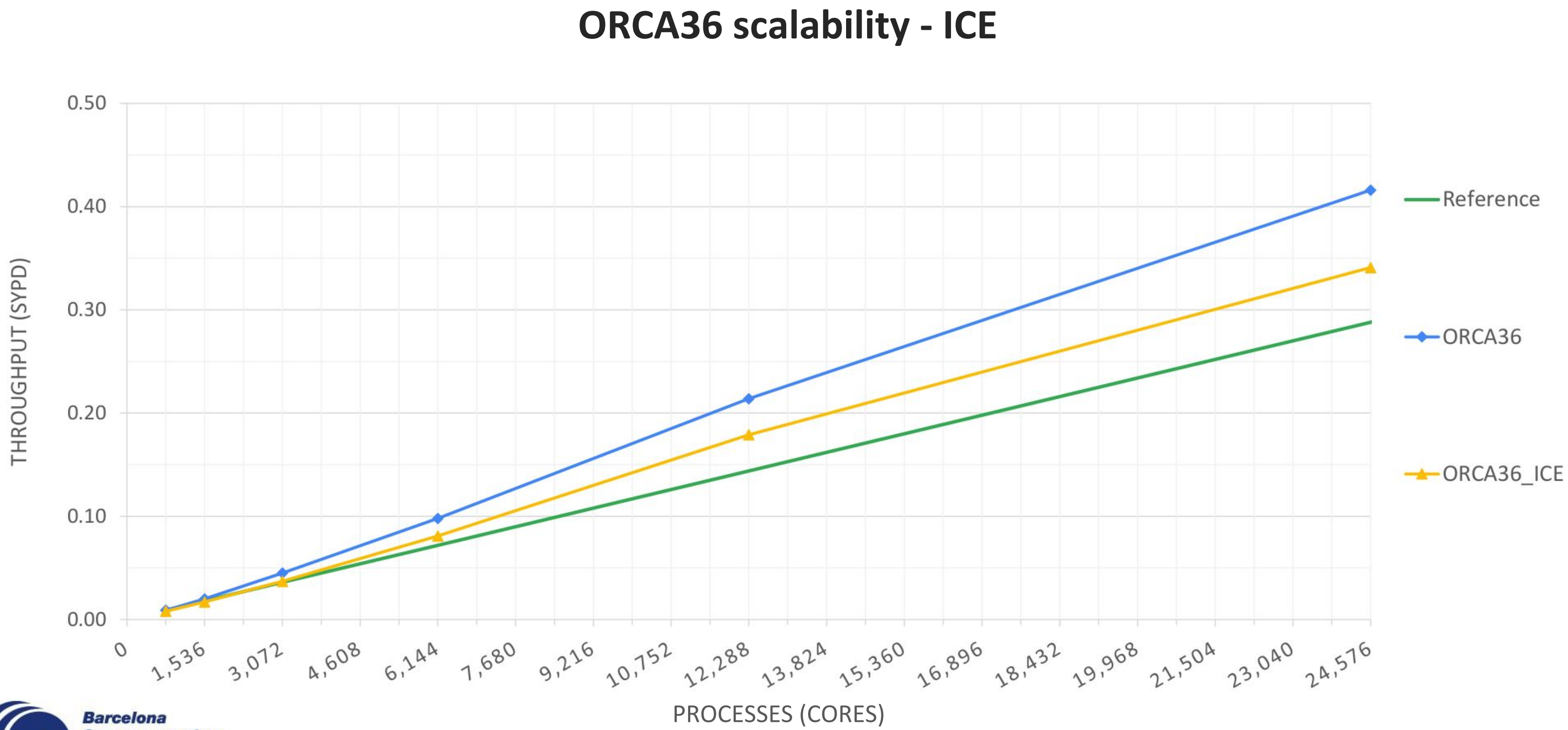
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability (MN4)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

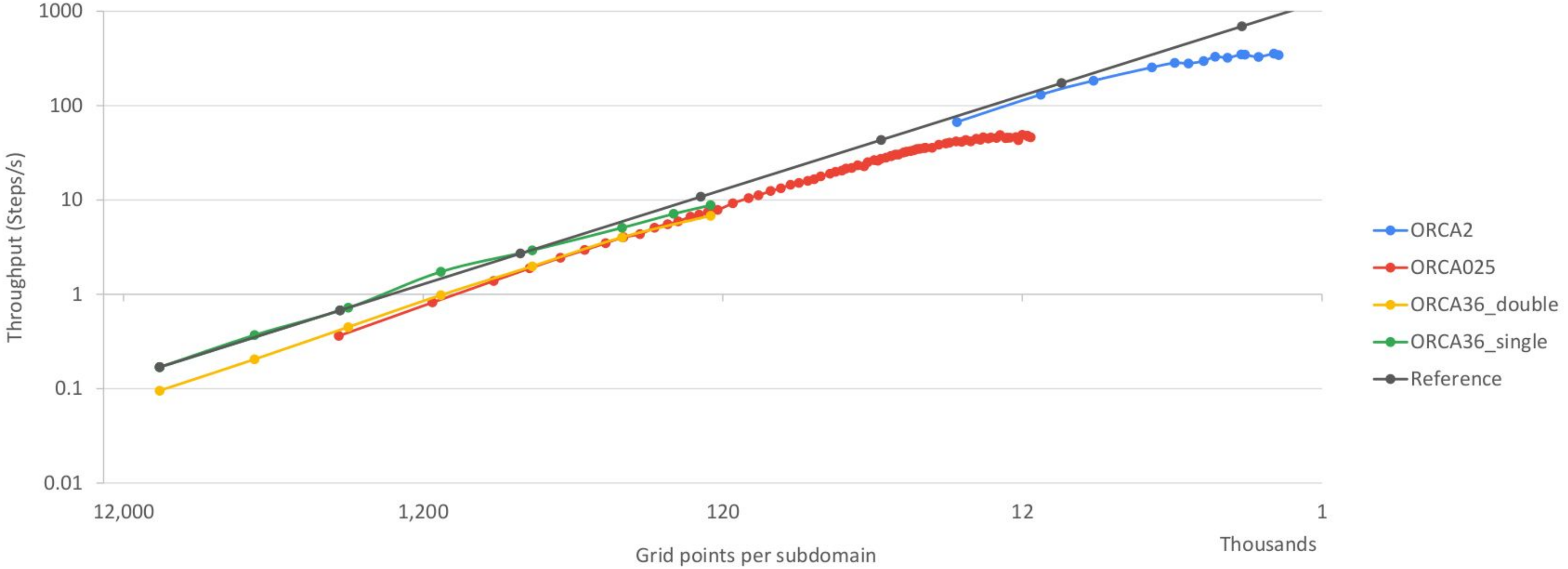
ORCA36 scalability (MN4)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA weak scaling (MN4)

ORCA2, ORCA025 and ORCA36 scalability. Steps per second per subdomain size



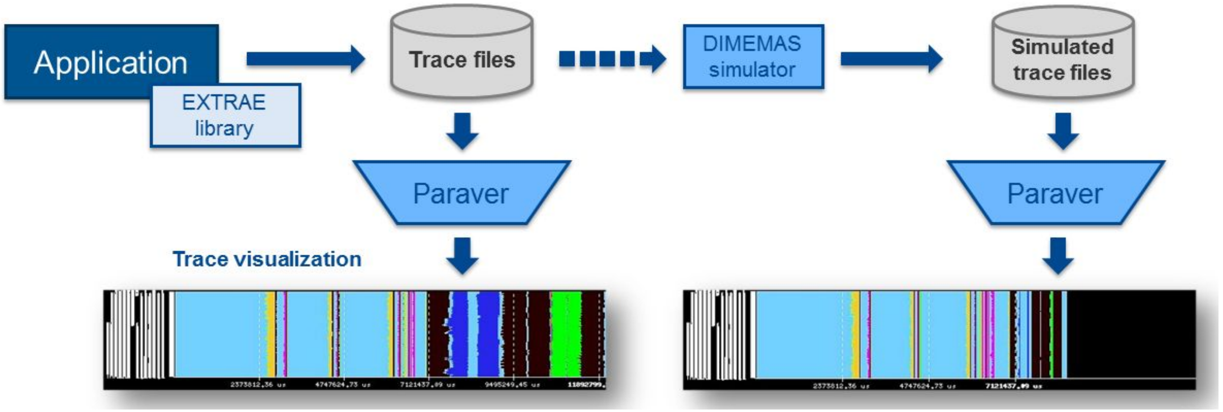
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 Performance analysis

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Performance analysis

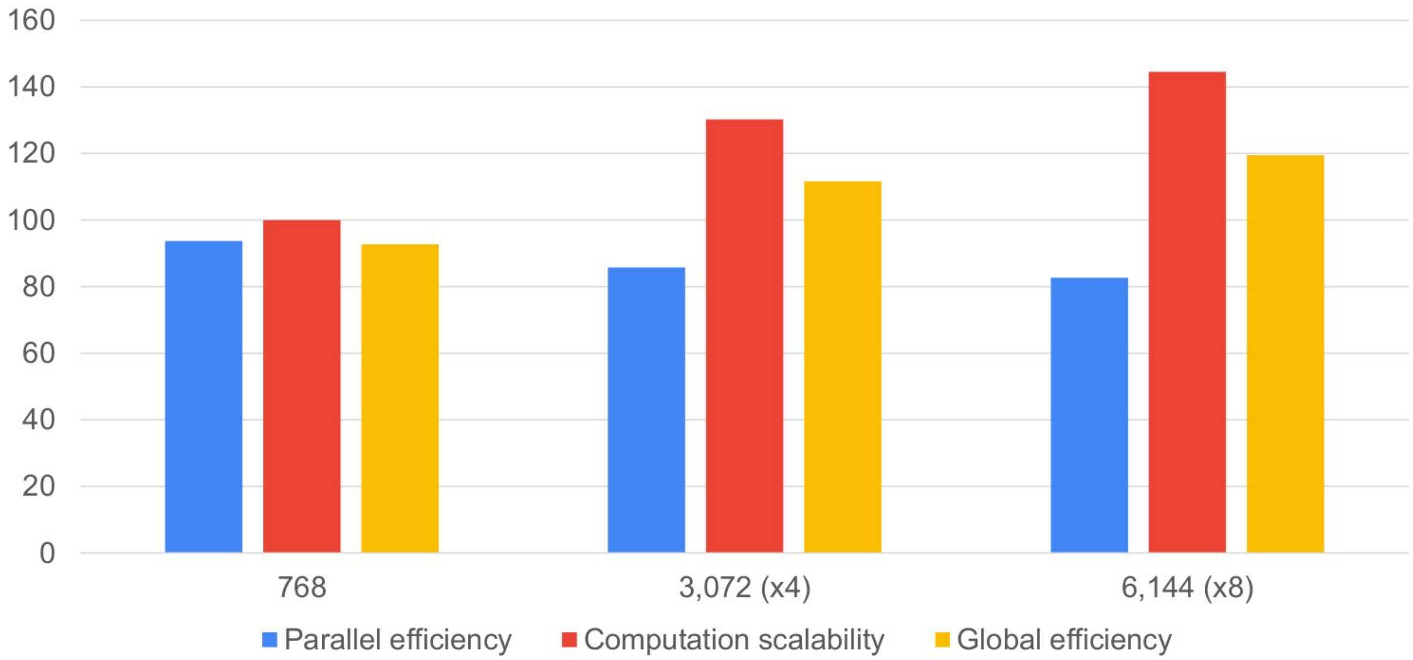
- Since 1991
- Based on **traces**
- Open Source: https://tools.bsc.es
- **Extrae**: Package that generates Paraver trace-files for a post-mortem analysis
- **Paraver**: Trace visualization and analysis browser
- **Dimemas**: Message passing simulator



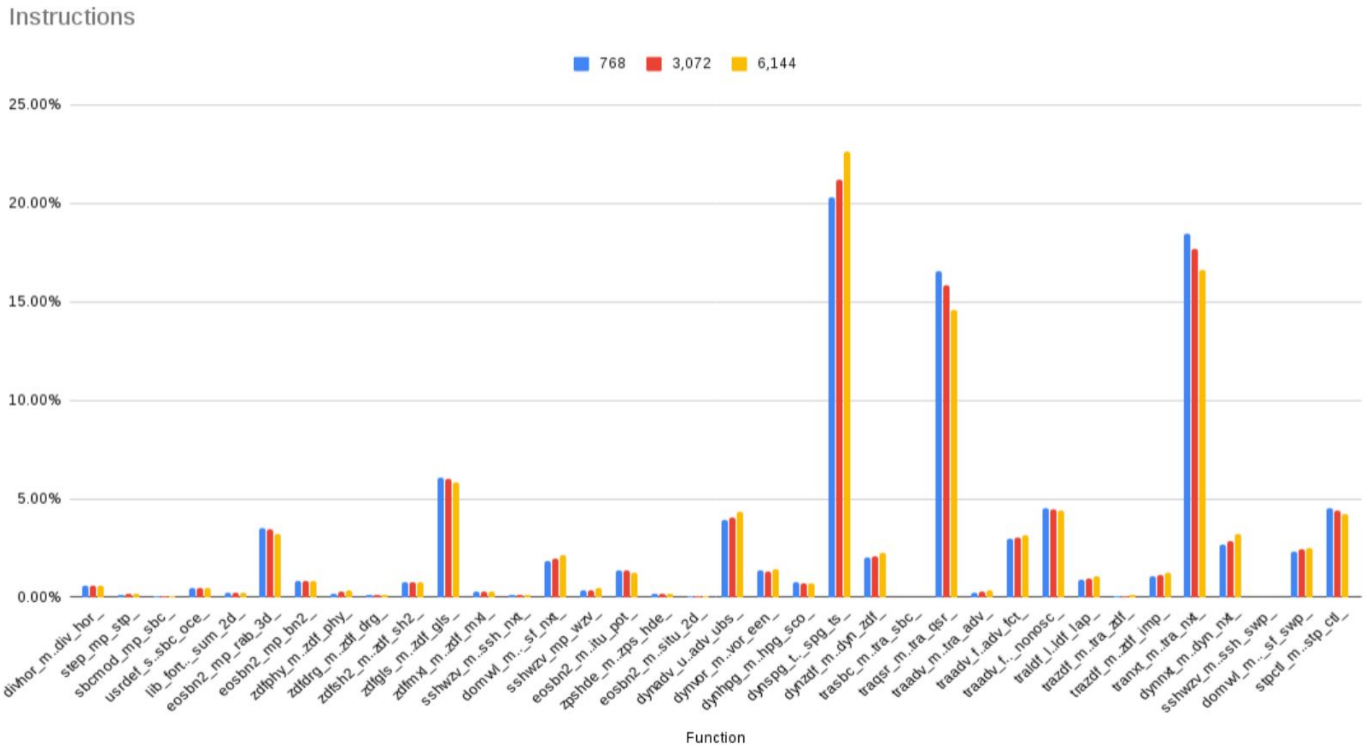
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability

Model factors explaining scalability on 16, 32 and 64 nodes

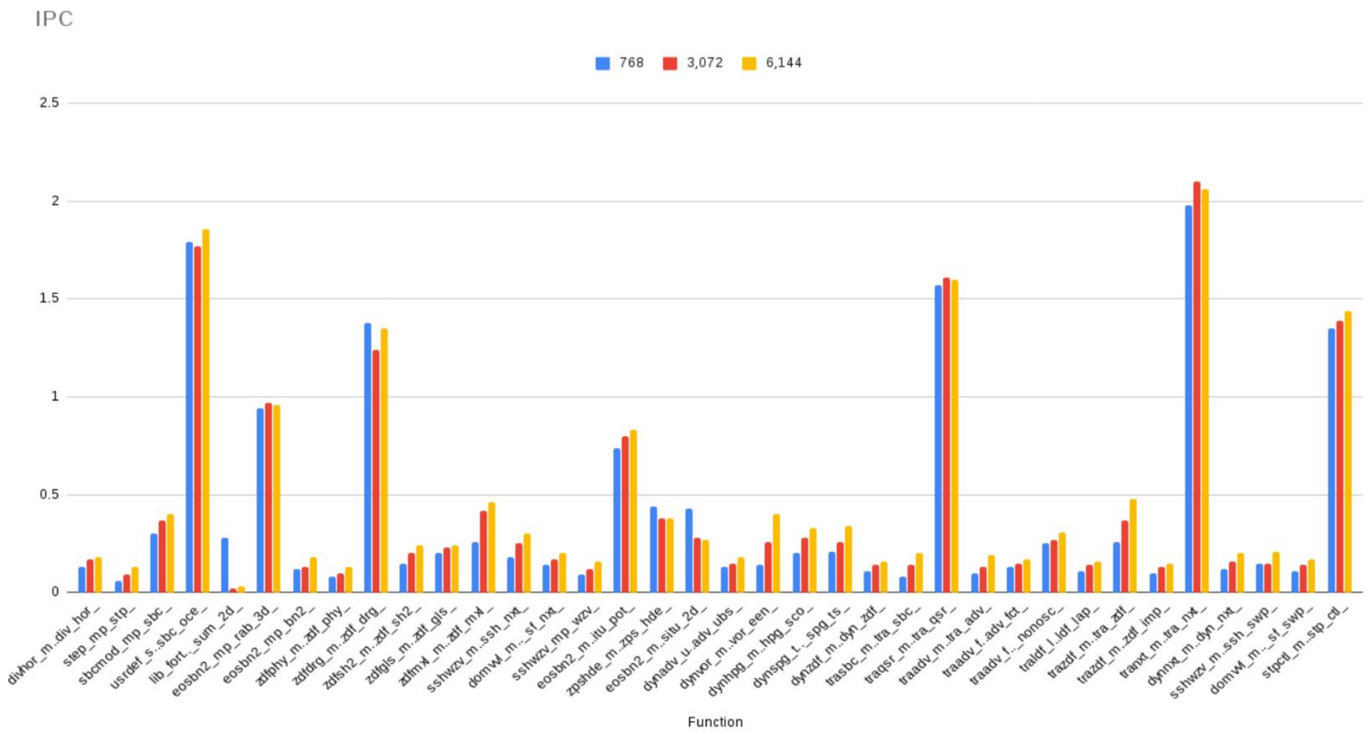


ORCA36 instructions breakdown



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 IPC per function

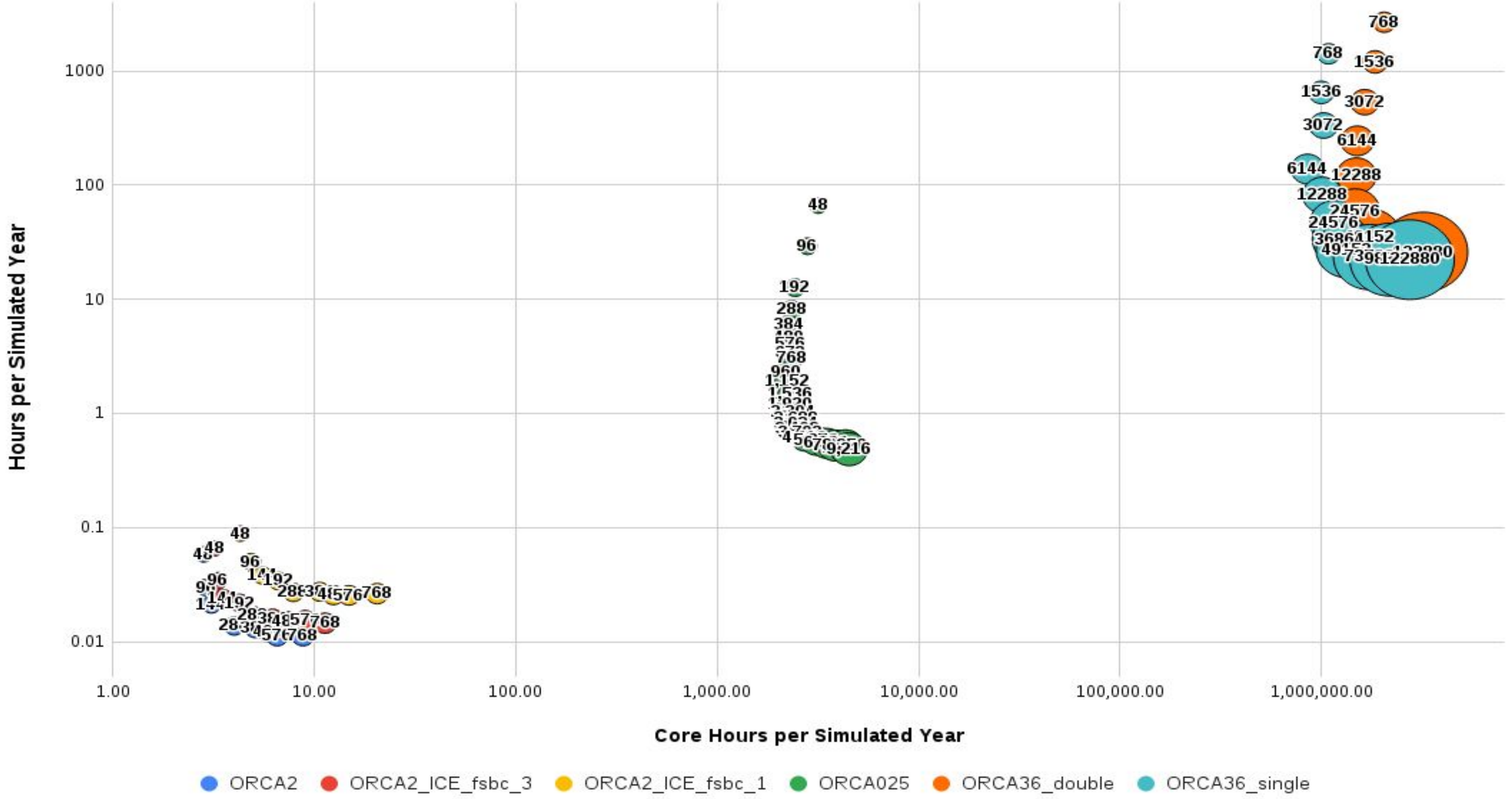


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NEMO4 time vs cost

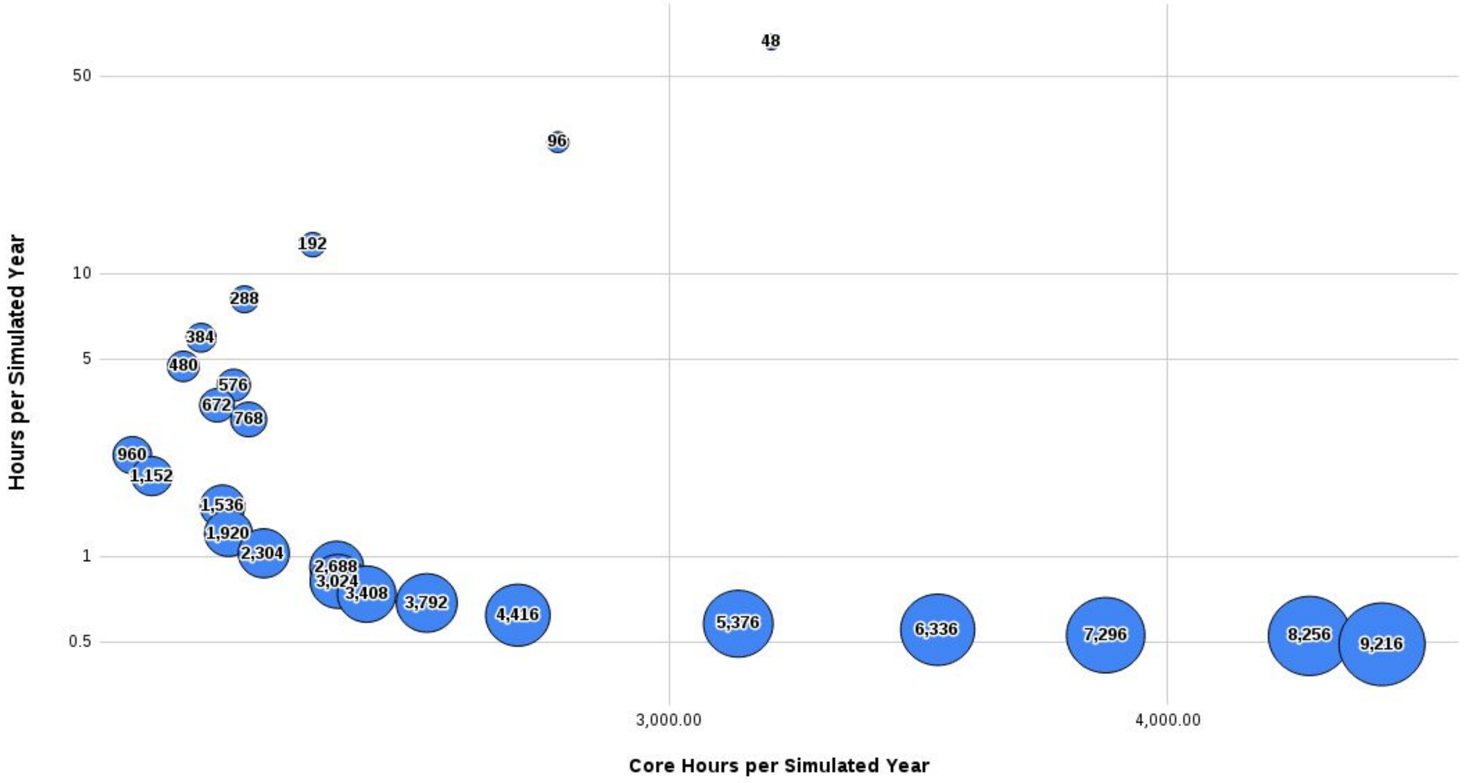
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NEMO4 time vs. cost

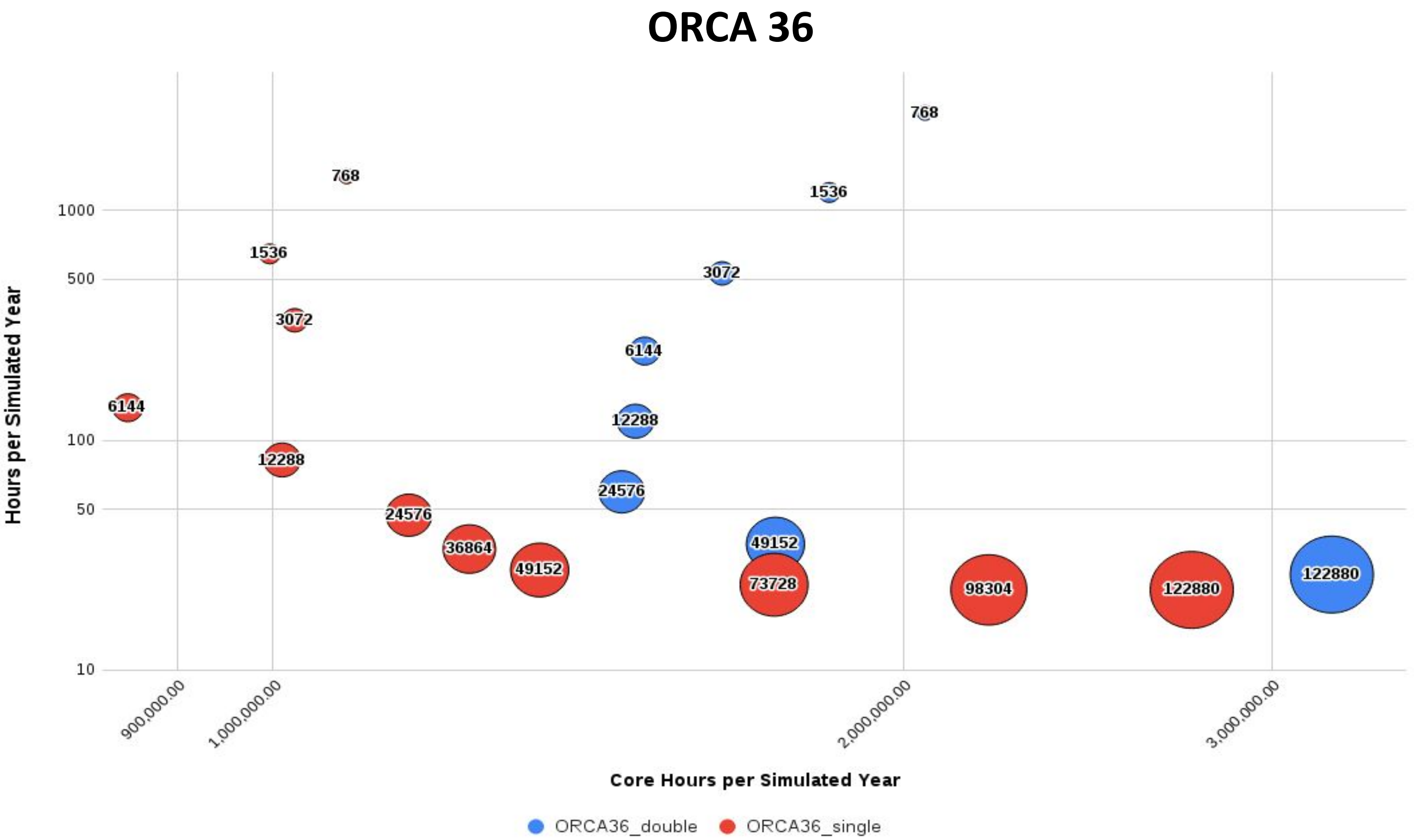


NEMO4 time vs. cost

ORCA 025



NEMO4 time vs cost



Conclusions

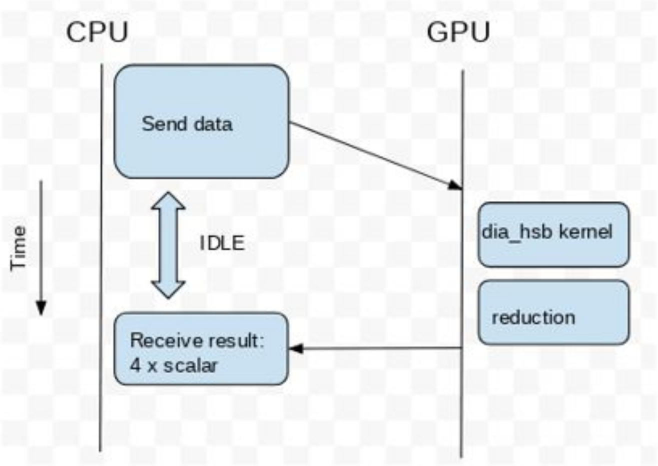
- **NEMO scalability** is good when maintaining subdomain size over 15x15. Max. throughput achieved at 10x10. With **very large** configurations (and many more PE’s) this may not be true.
- **Using mixed precision** in NEMO may allow to achieve **1SYPD** with 3km global resolution on current architectures. Up to **x1.9 speedup** on memory bandwidth bound configurations.
- NEMO **memory usage** is not scaling: **online interpolations** in ORCA36 make impossible to run the model on standard nodes.
- **Data is an issue:** restarts of ~1Tb size.

Porting NEMO diagnostics to GPUs

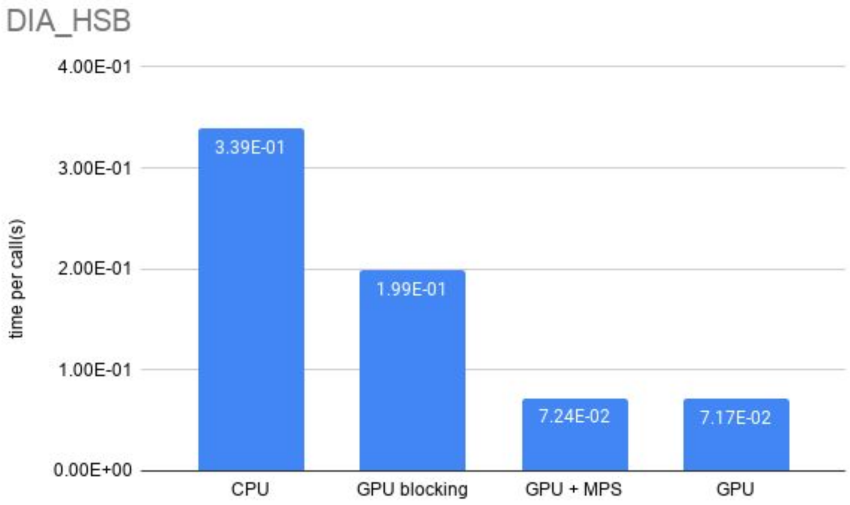
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

IMMERSE: Porting diagnostics to GPUs

The diagnostics dia_hsb kernel



DIA_HSB diagnostics time, using one Power9 node

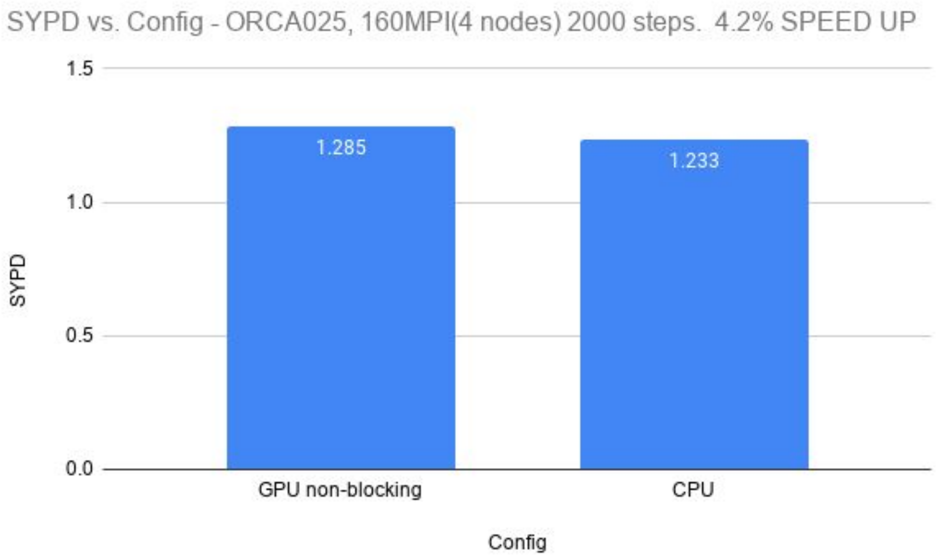


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Maicon Faria (BSC)

NEMO's GPU Diagnostic Strong Scaling

ORCA 025



Thank you

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 821926.

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