Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

The NEMO ORCA36 configuration and approaches to increase NEMO4 efficiency

Miguel Castrillo

BSC-ES Performance Team, Computational Earth Sciences

6th ENES HPC workshop

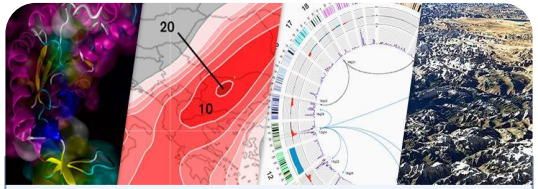
25/05/2020

The Barcelona Supercomputing Center

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%
Catalan Government	30%
Univ. Politècnica de Catalunya (UPC)	10%

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 investment

EuroHPC Joint Undertaking

Hosting Consortium:

Spain Portugal Turkey Croatia

ESiWACE2 H2020

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.
- **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.

IMMERSE & IS-ENES3 H2020

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

BSC-ES involvement in IMMERSE

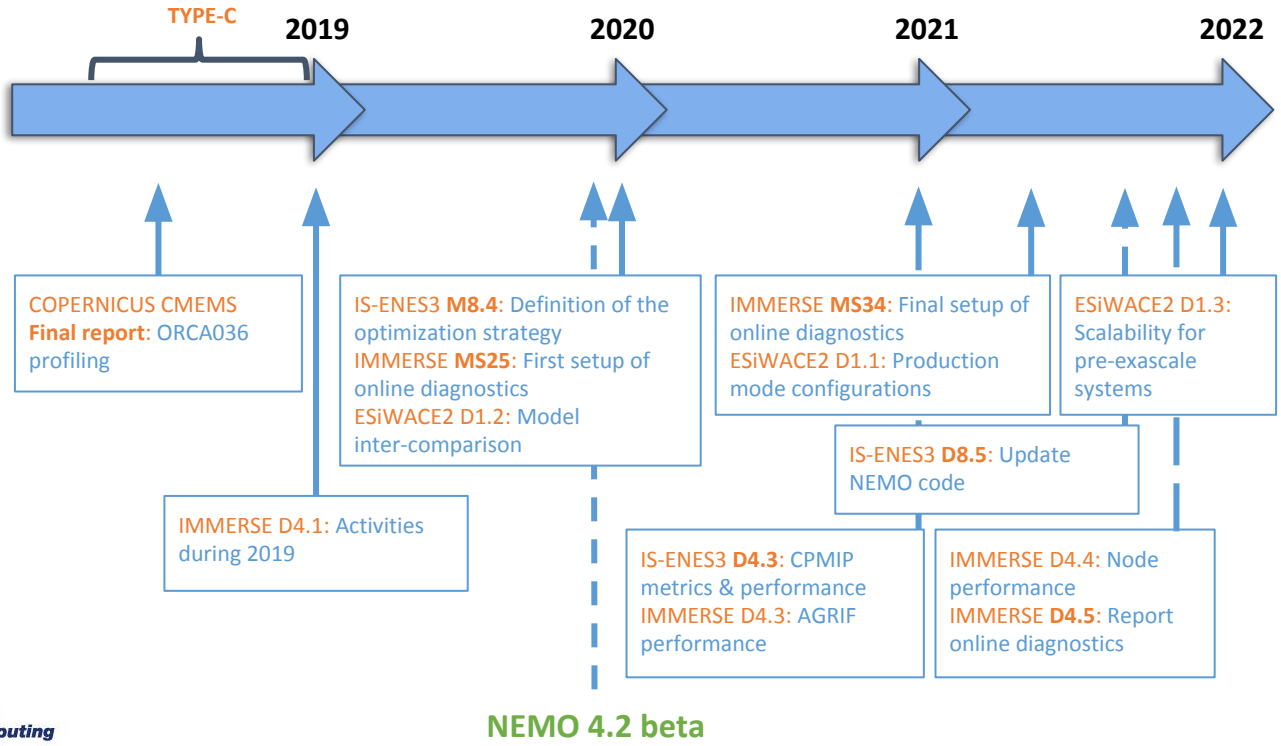
- **IS-ENES3 T8.1:** Improving Nemo computational performance
  - **Routine level scalability: Communication impact, workload balance, mixed precision analysis**
- **IMMERSE T4.1:** Efficient exploitation of memory hierarchies and hardware peak performance
  - **Assessment** of the **performance impact**
- **IMMERSE T4.3:** Efficient IOs and diagnostics for operational systems
  - **Offload** NEMO model **diagnostics to GPUs**
- **IMMERSE T4.4:** Load balancing for AGRIF massive multigrid capability
  - **Efficiency assessment** for **high-resolution** configuration

is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

immerse
IMPROVING OCEAN MODELS FOR THE COPERNICUS PROGRAMME

NEMO timeline in BSC-ES performance



From ORCA2 to ORCA36

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

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

ORCA36 in MareNostrum4

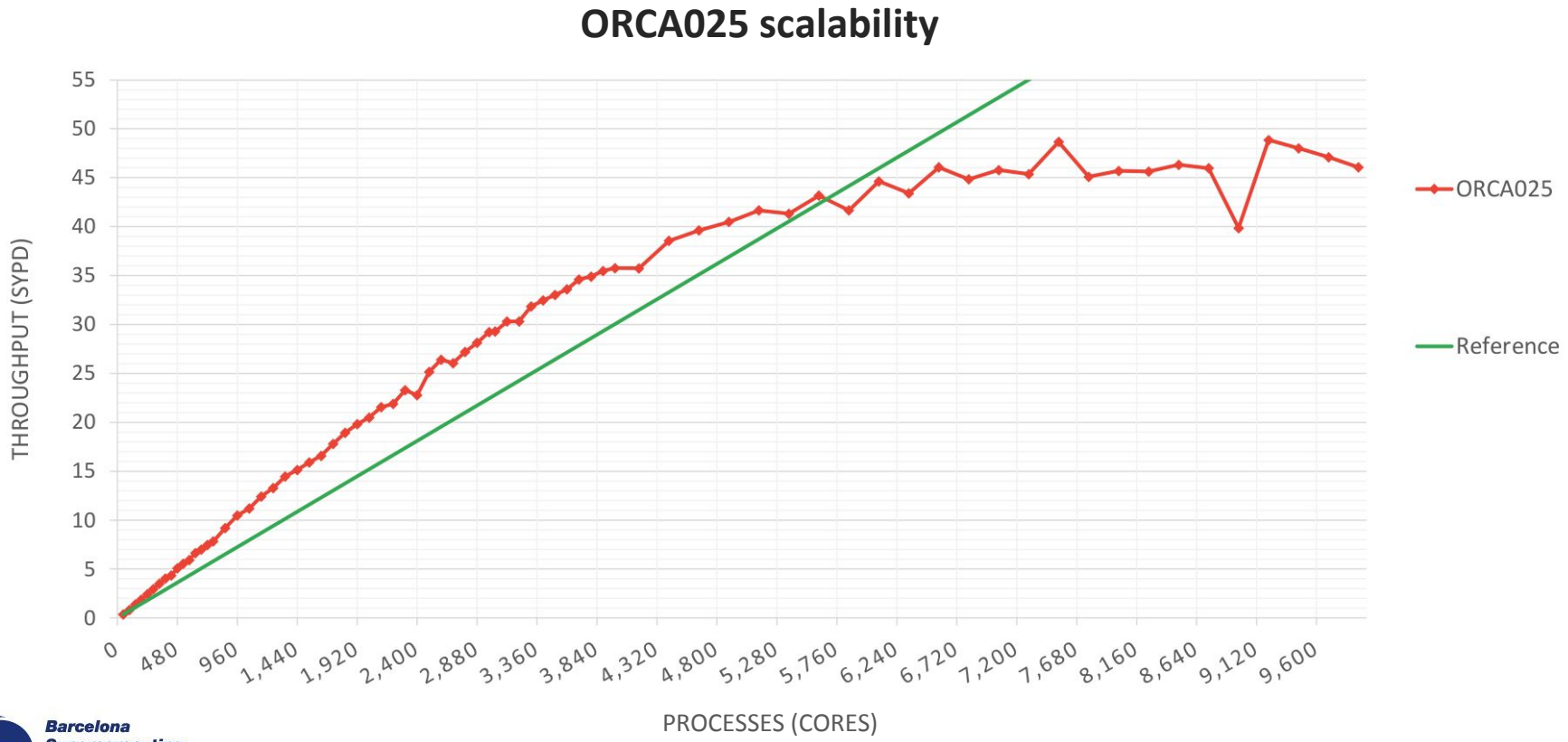
Resources constraints

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

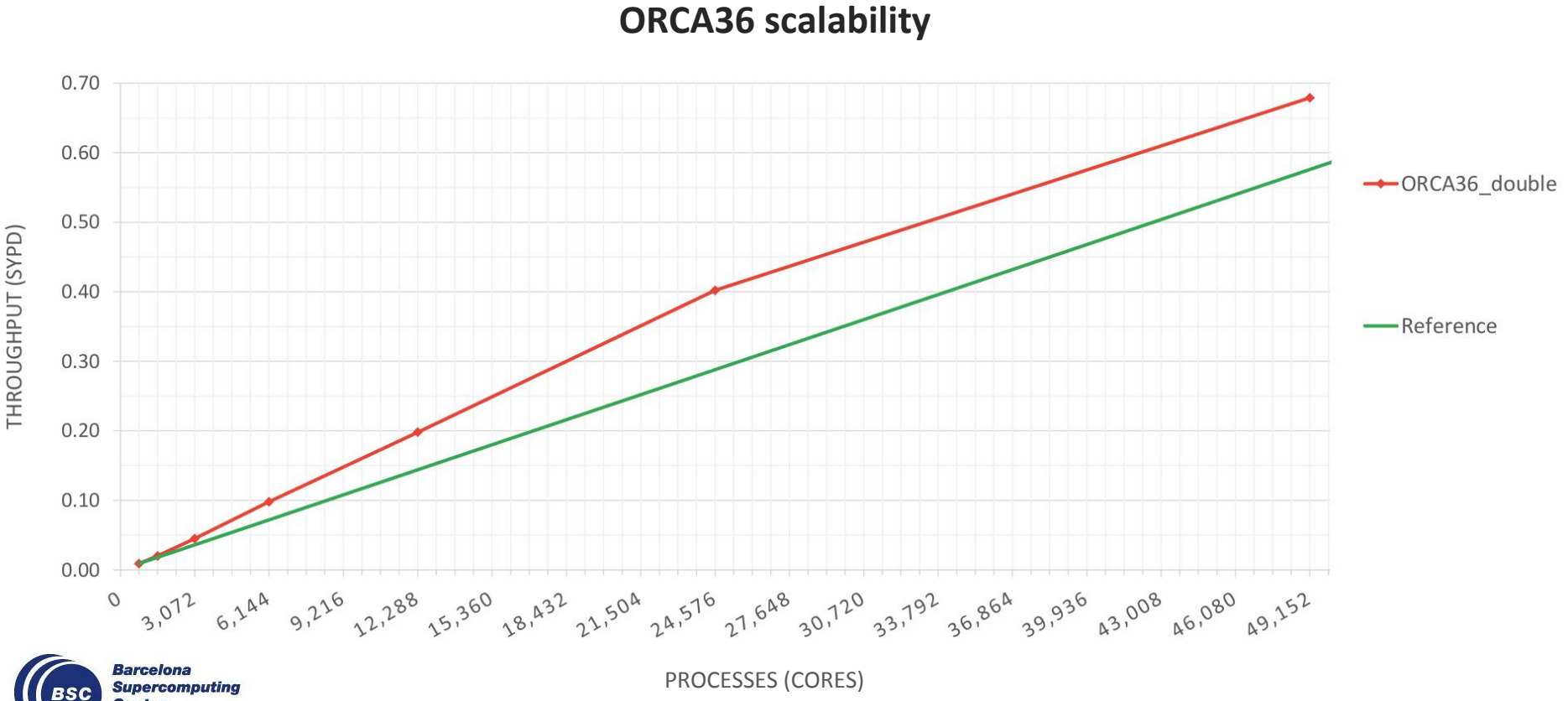
ORCA36 scaling

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

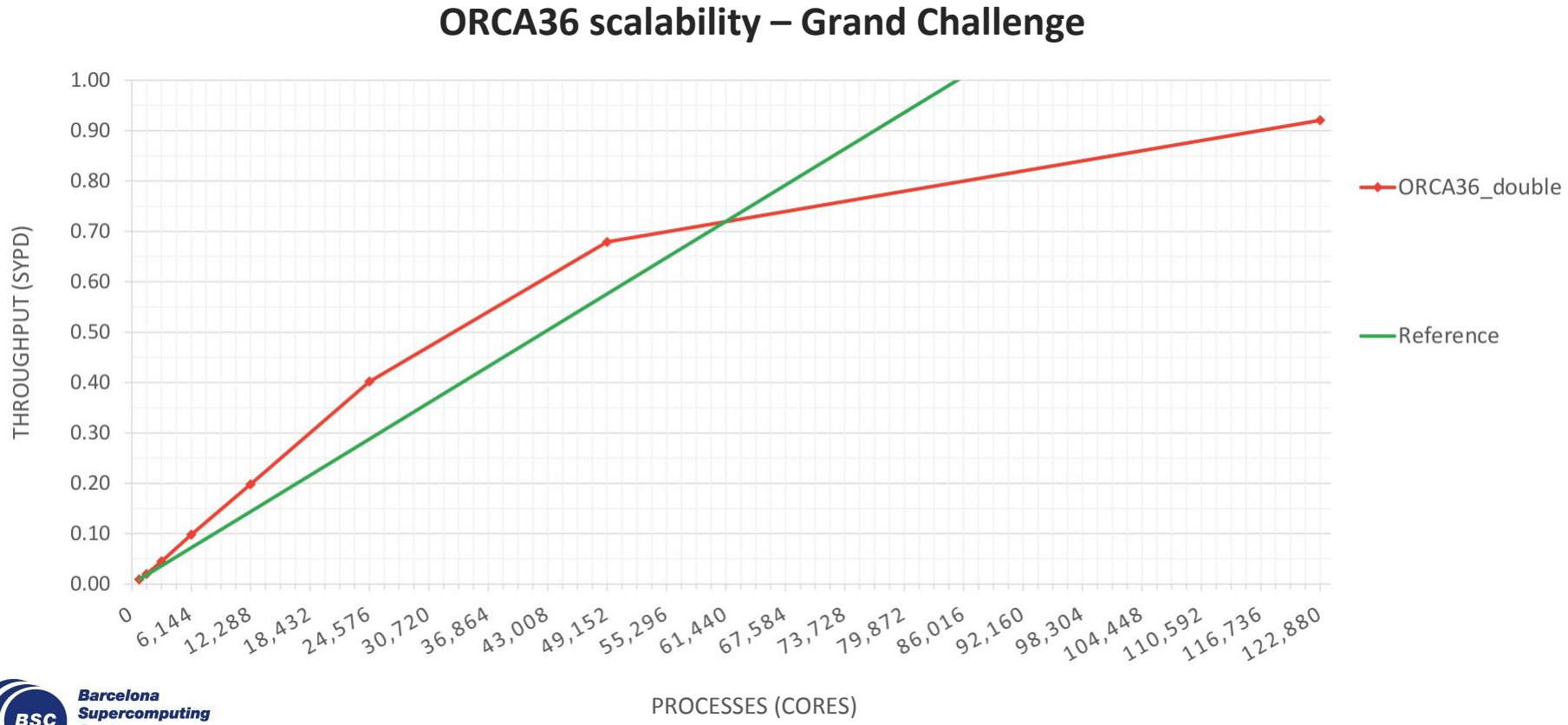
ORCA025 scalability (MN4)



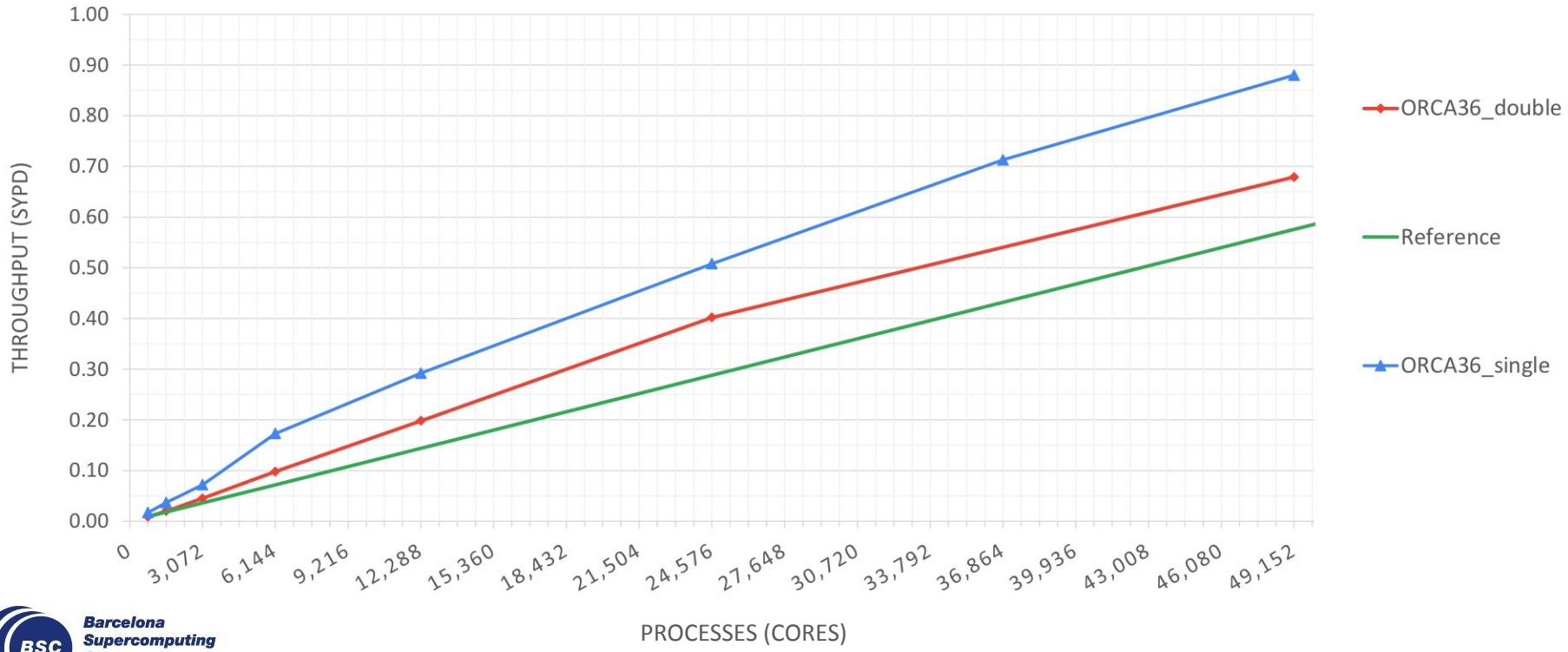
ORCA36 scalability (MN4)



ORCA36 scalability (MN4)

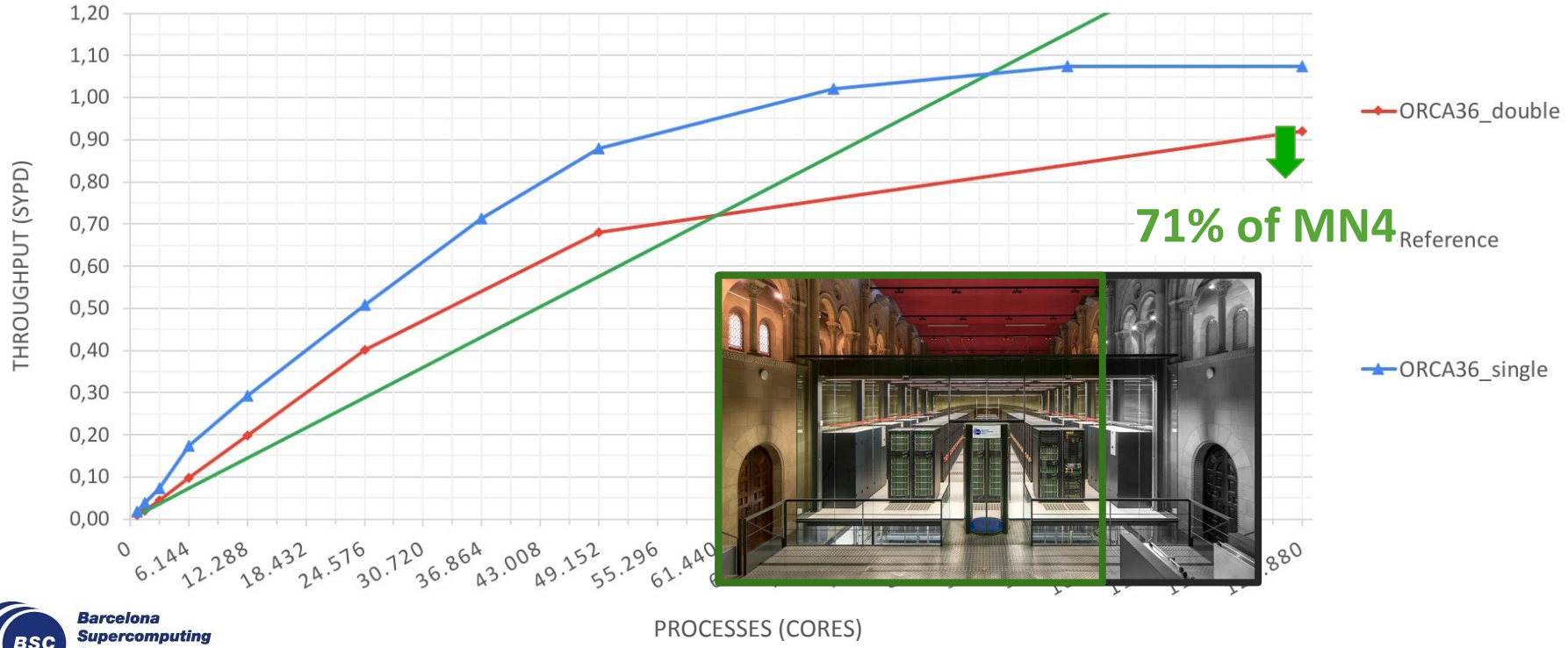


ORCA36 scalability (MN4)

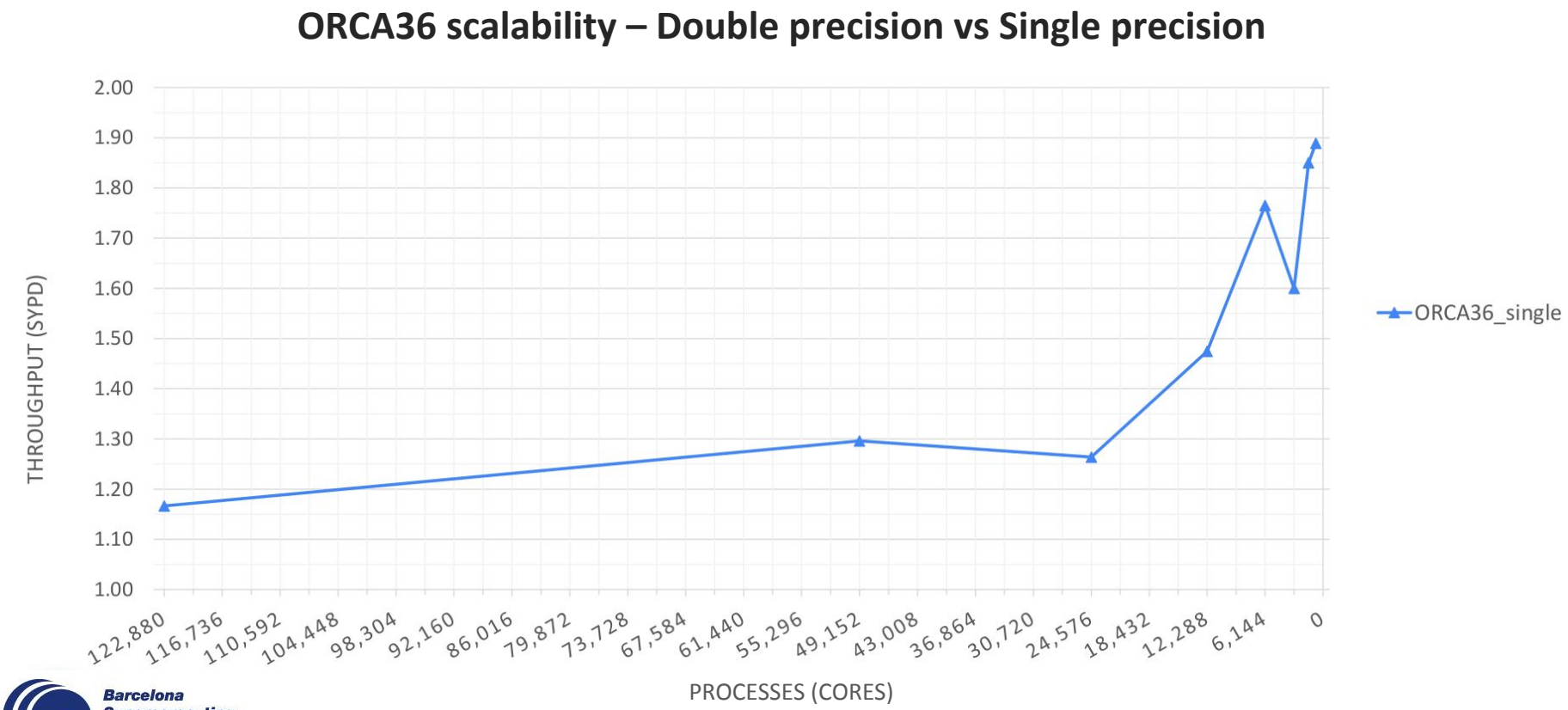


ORCA36 scalability (MN4)

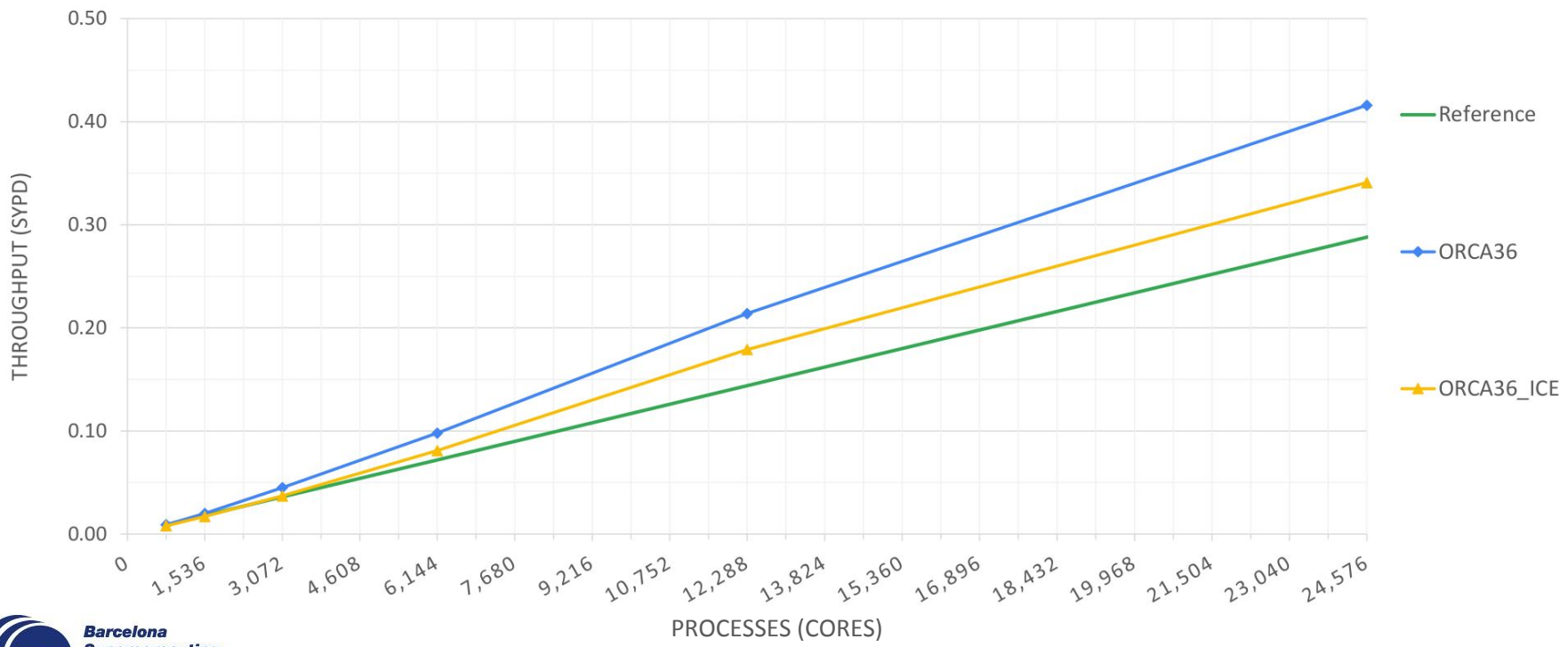
ORCA36 scalability – Double precision vs Single precision – Grand challenge



ORCA36 scalability (MN4)

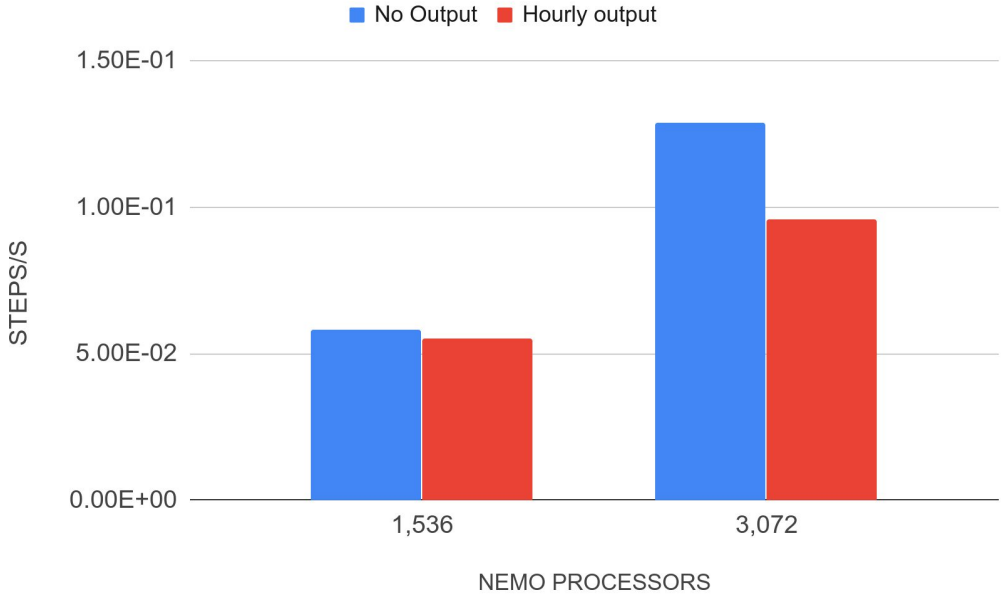


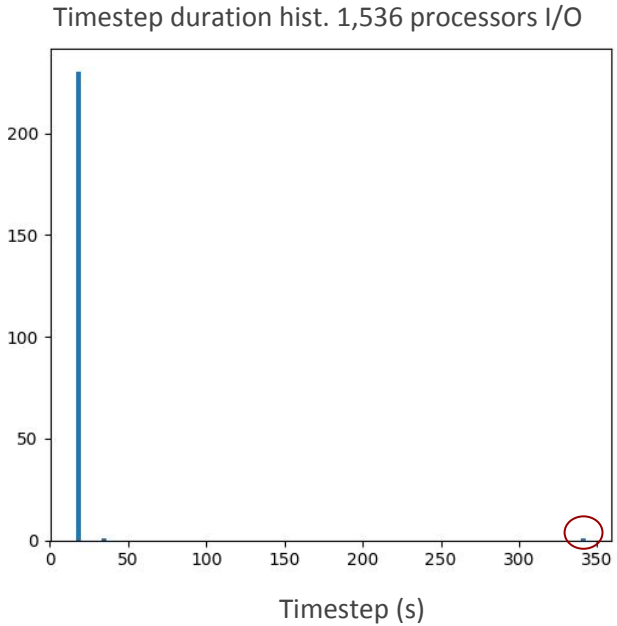
ORCA36 scalability (MN4)



ORCA36 I/O (MN4)

ORCA36 I/O tests with XIOS2.5



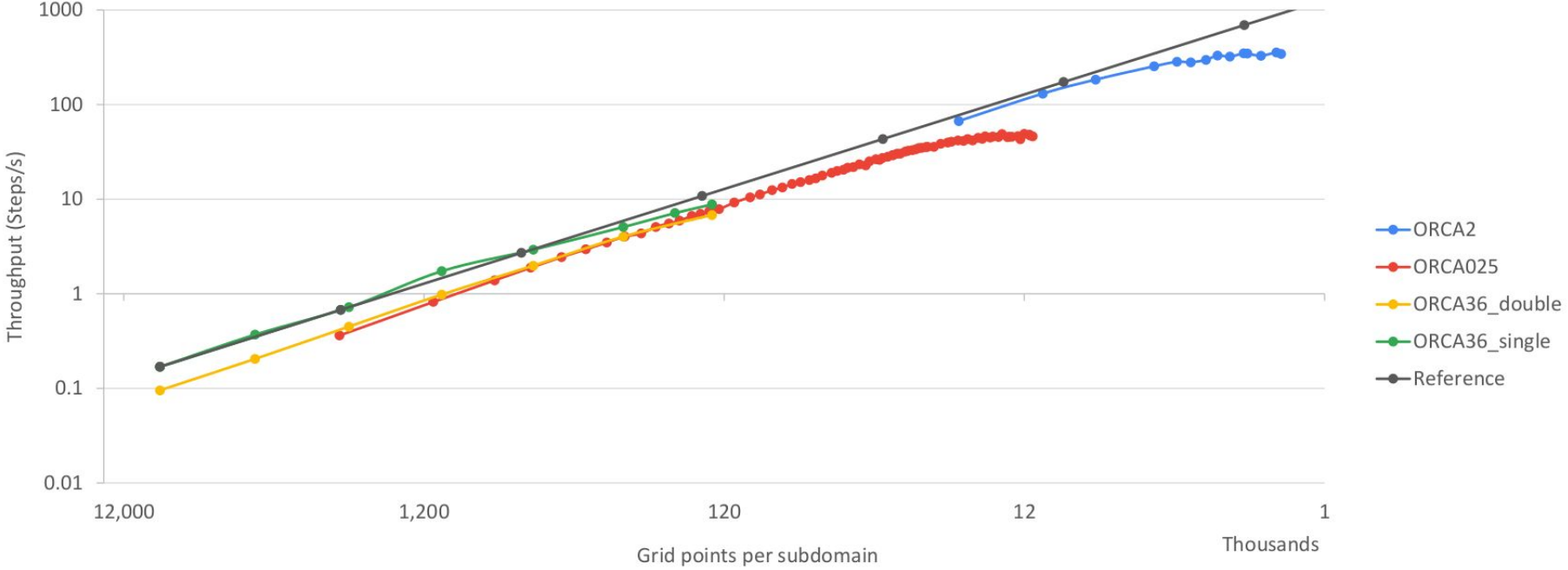


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

ORCA weak scaling (MN4)

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

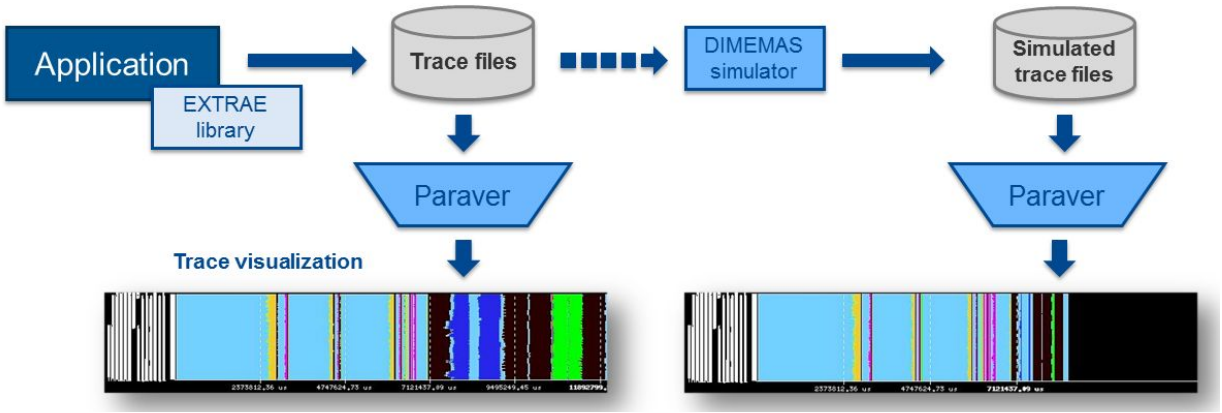


ORCA36 Performance analysis

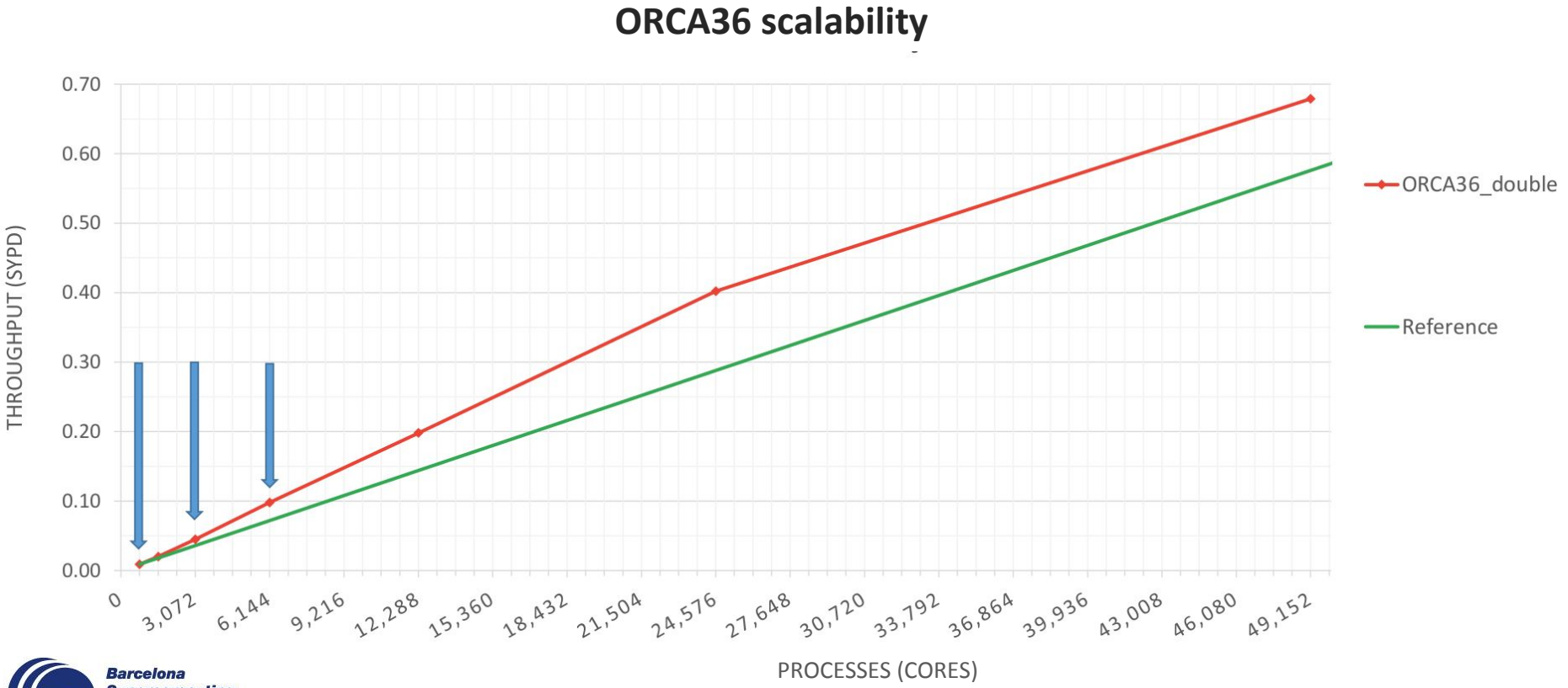
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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



ORCA36 scalability (MN4)

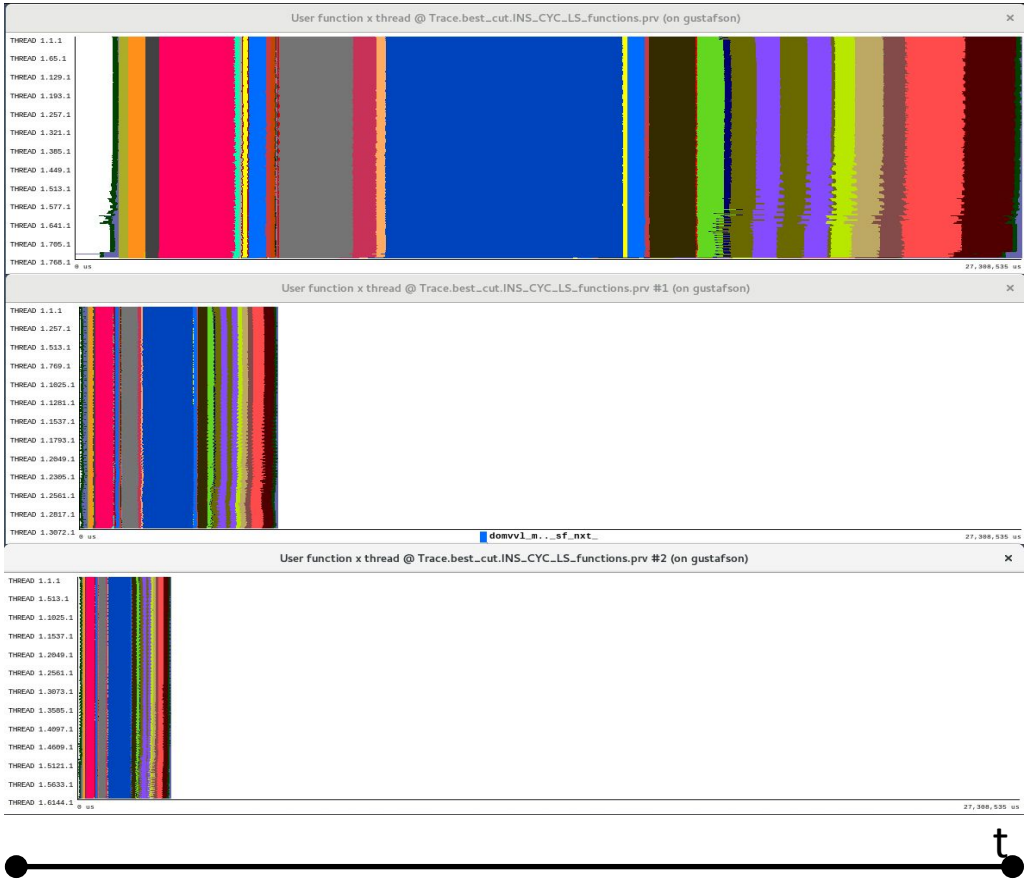


ORCA36 functions view

768

3,072

6,144

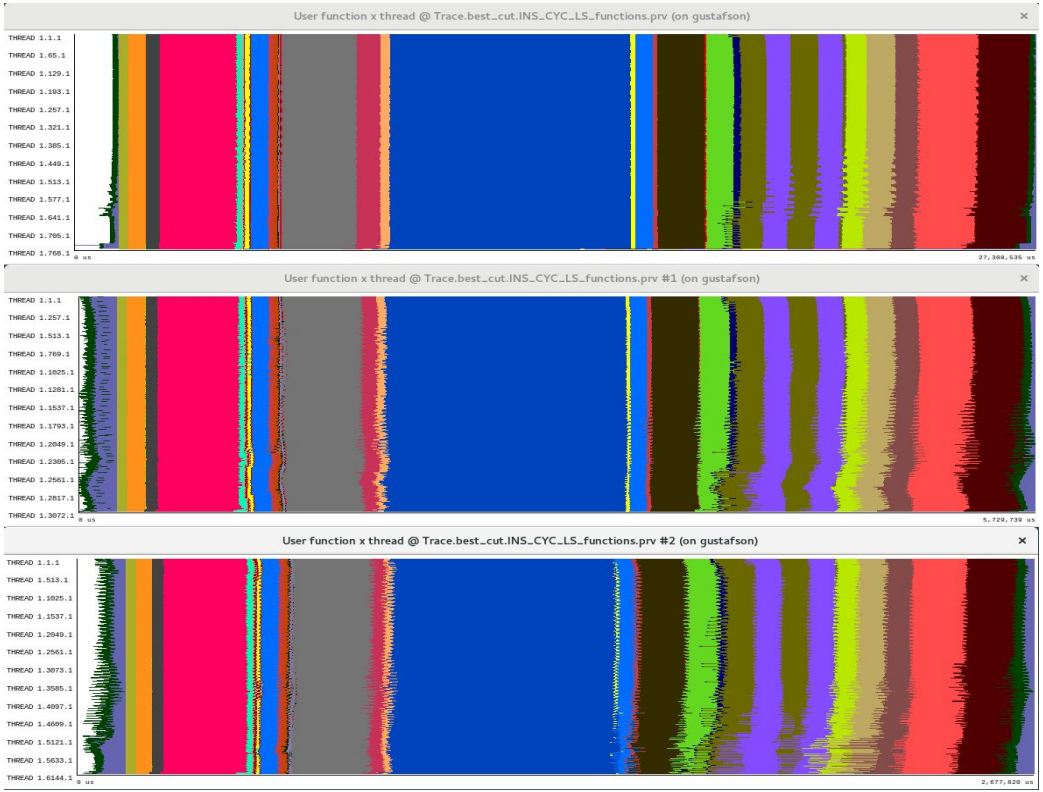


ORCA36 functions view

768

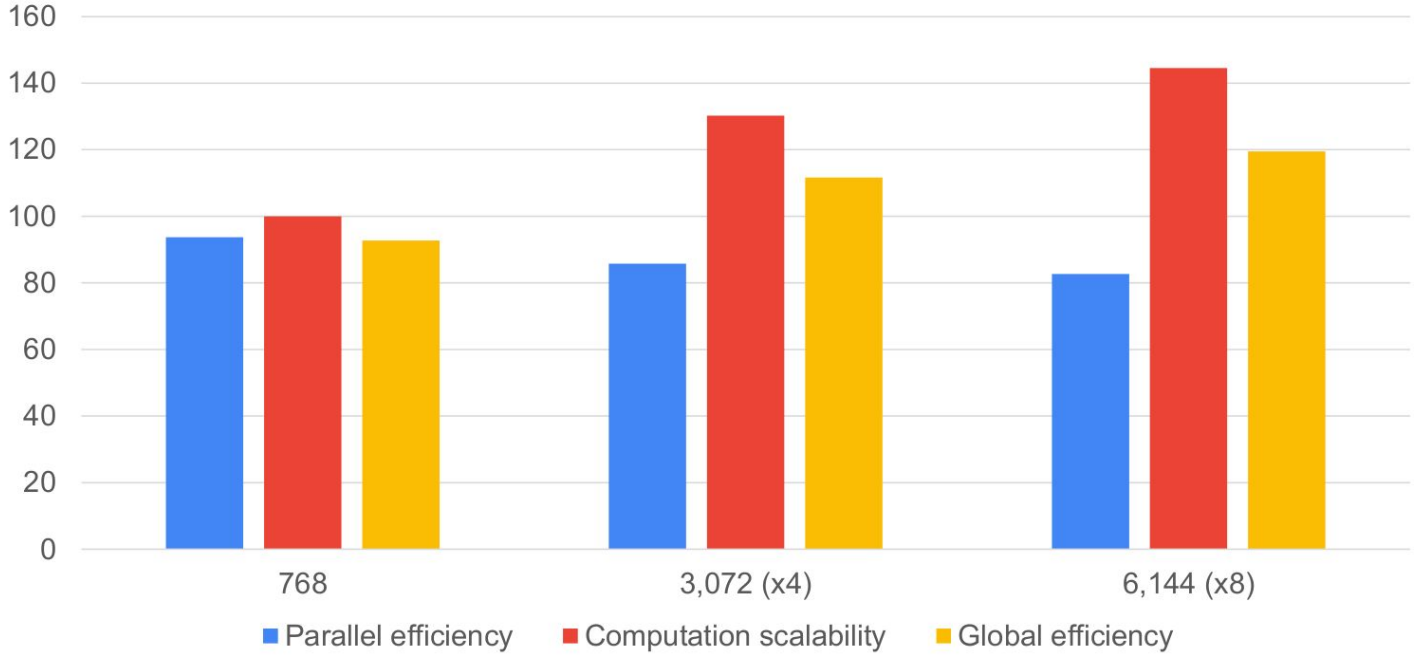
3,072

6,144

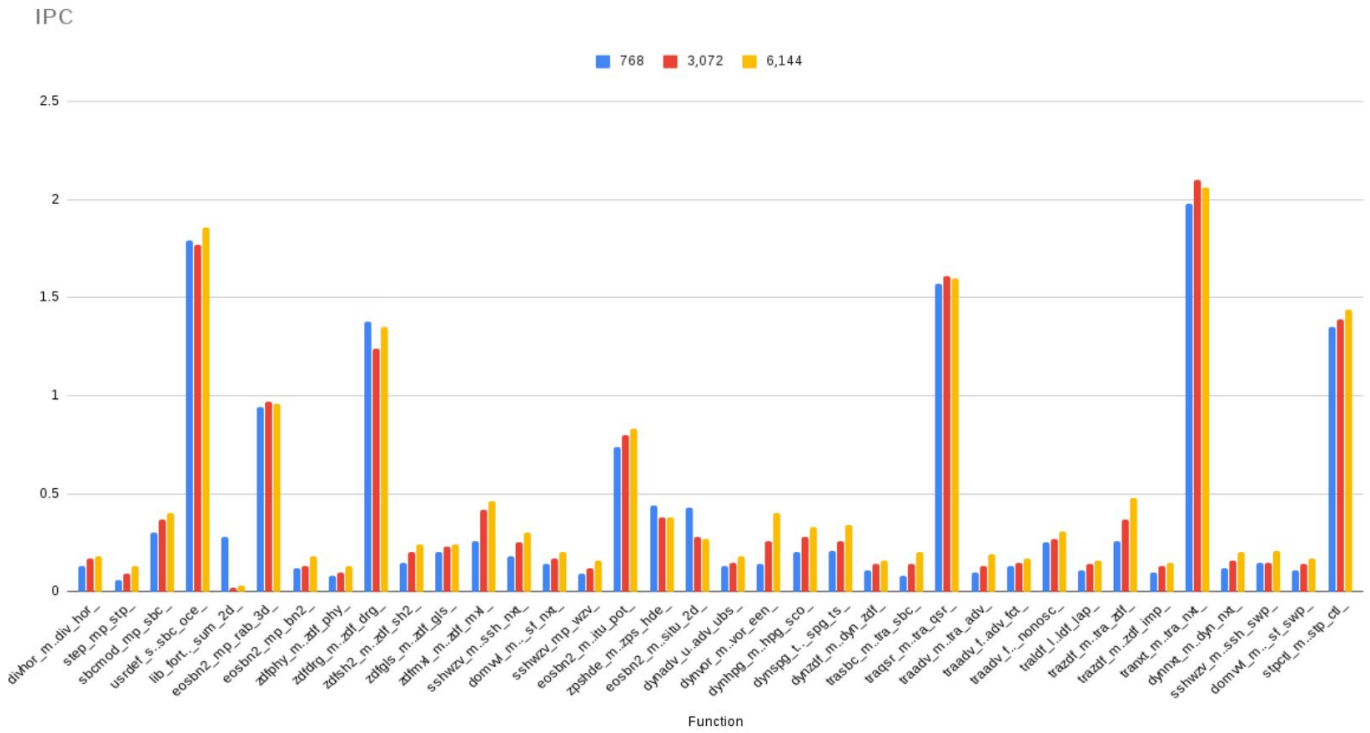


ORCA36 scalability

Model factors explaining scalability on 16, 32 and 64 nodes



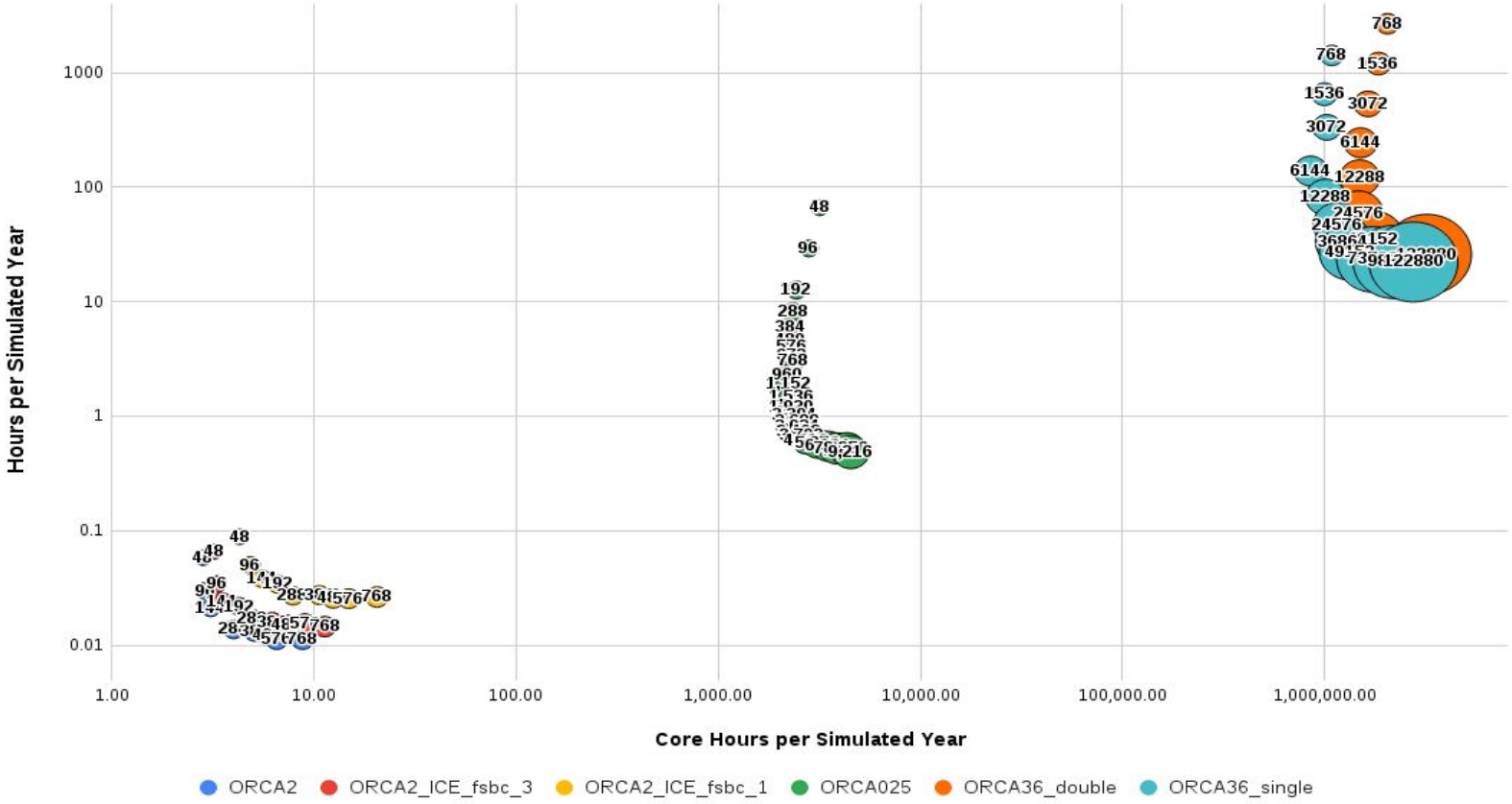
ORCA36 IPC per function



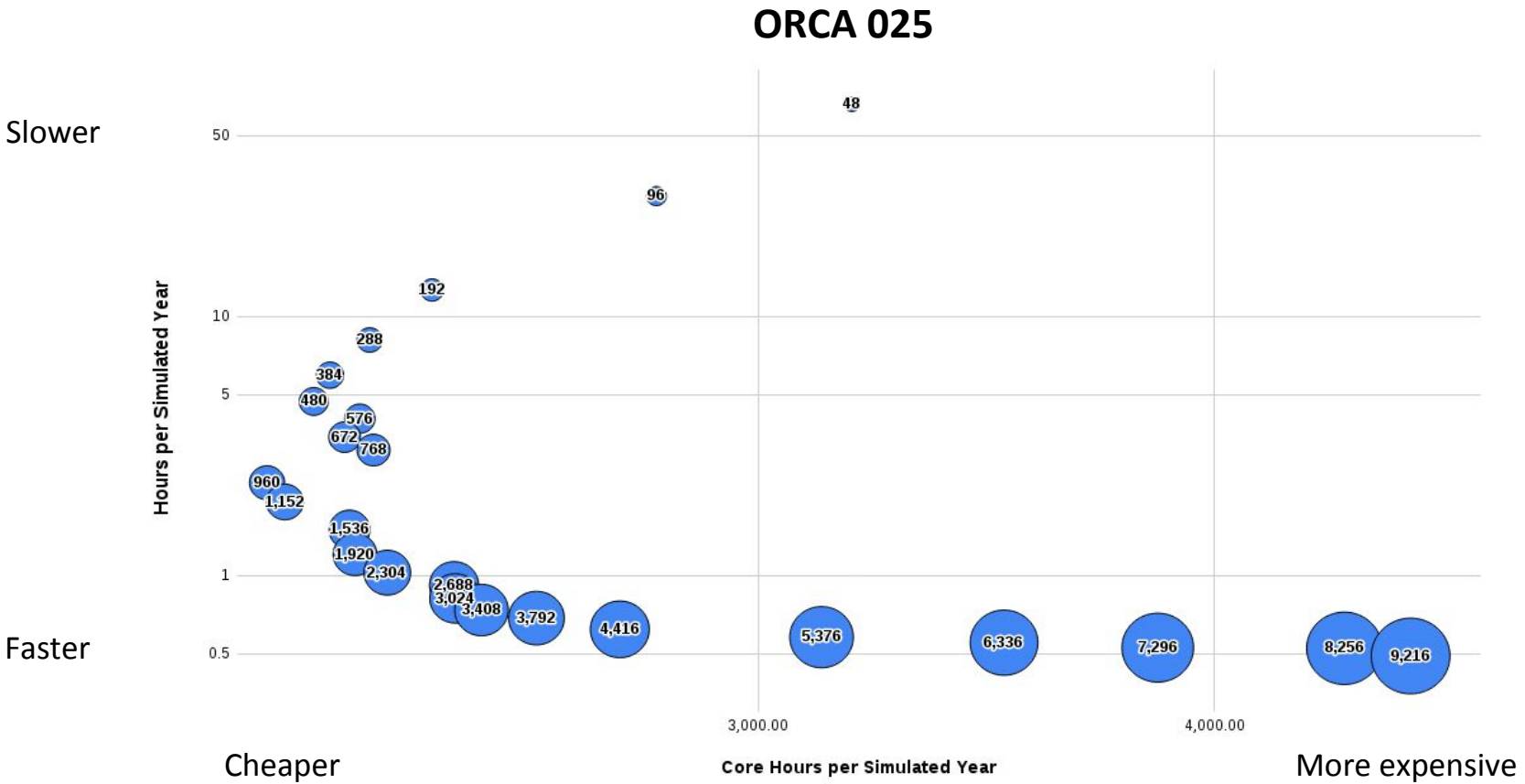
NEMO4 time vs cost

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NEMO4 time vs. cost



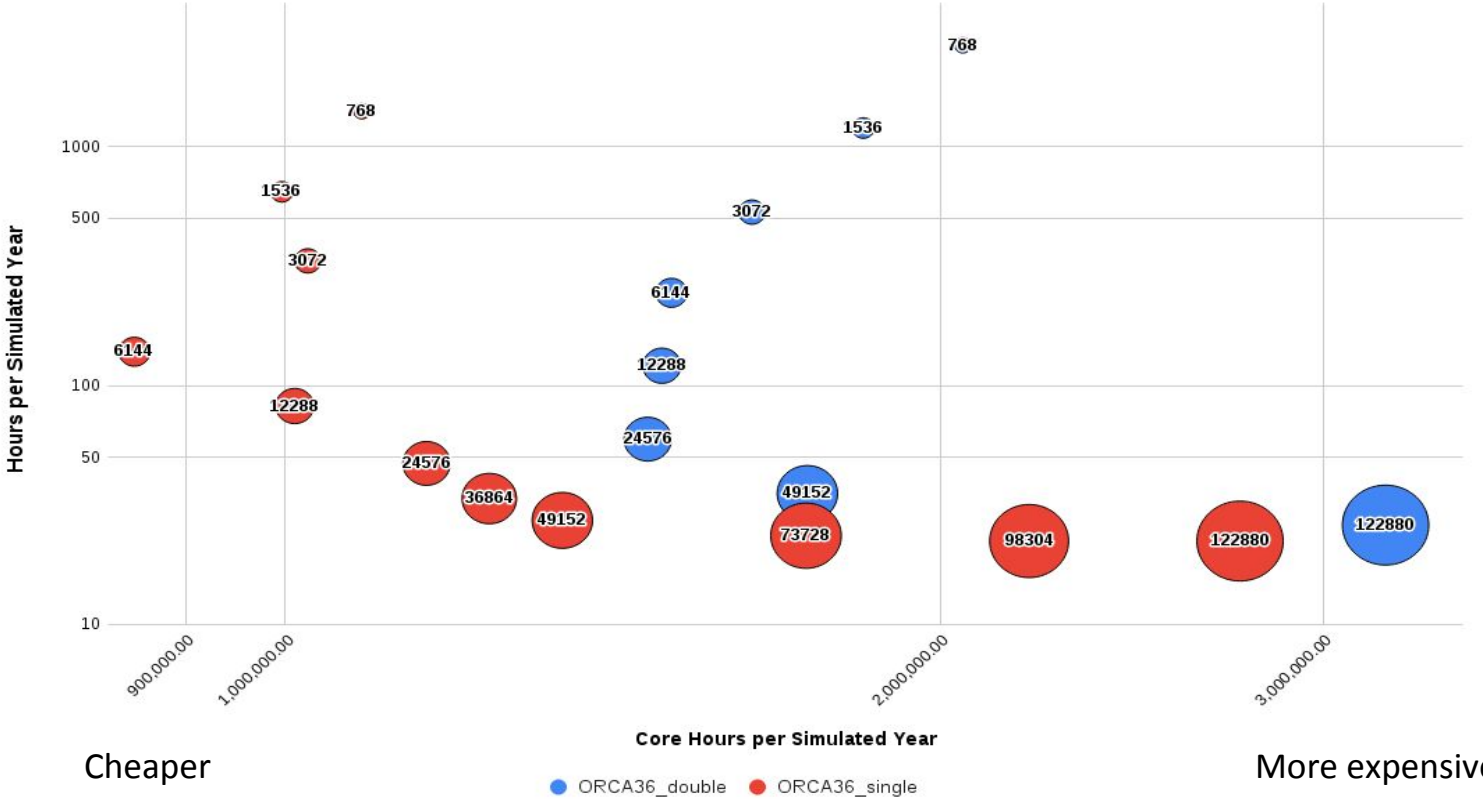
NEMO4 time vs. cost



NEMO4 time vs cost

ORCA 36

Slower



Faster

Cheaper

More expensive

Conclusions

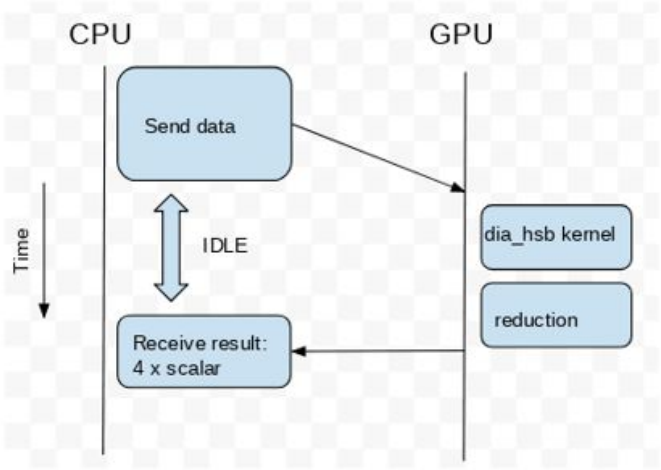
- **NEMO scalability** is good when maintaining subdomain size over 15x15. Max. throughput achieved at 10x10. With **very large** configurations (and many more PE's) this may not be true.
- **Using mixed precision** in NEMO may allow to achieve **1SYPD** with 3km global resolution on current architectures. 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:** ~1Tb size restarts, ~360GB per simulation hour.
- **Production throughput depends on many variables:** time step size, I/O frequency, I/O size, diagnostics computation, coupling, namelist parameters……

Porting NEMO diagnostics to GPUs

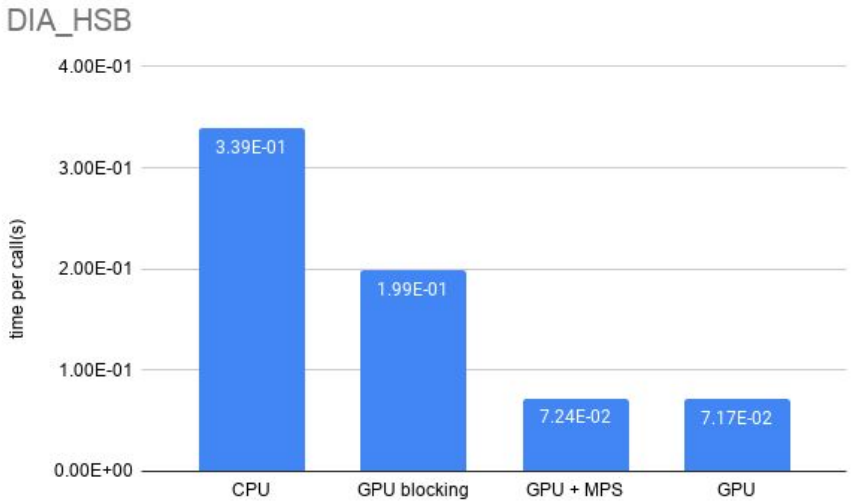
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

IMMERSE: Porting diagnostics to GPUs

The diagnostics dia_hsb kernel

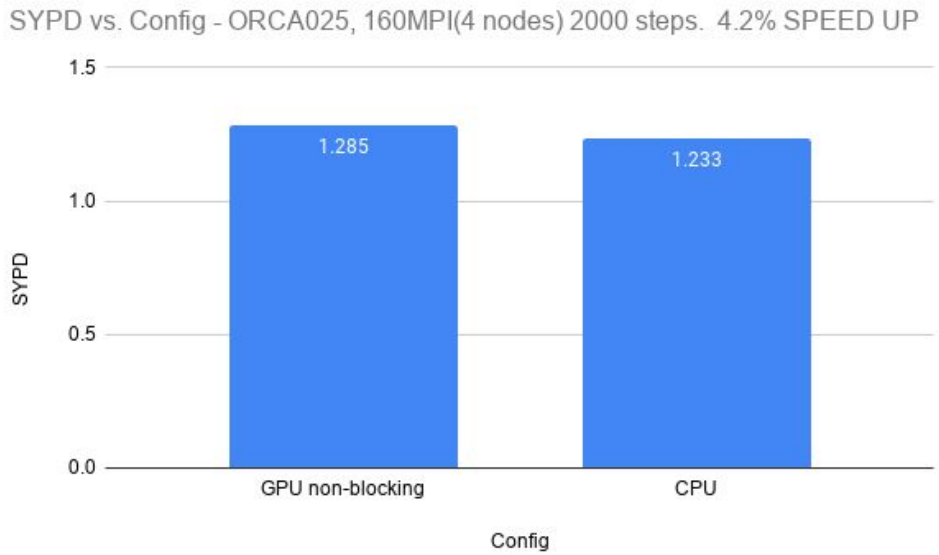


DIA_HSB diagnostics time, using one Power9 node



NEMO's GPU Diagnostic Strong Scaling

ORCA 025



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Maicon Faria (BSC)

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