Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Present models & machines running future resolutions. The ORCA36 configuration and approaches to increase NEMO4 efficiency

Miguel Castrillo

BSC-ES Performance Team, Computational Earth Sciences

ECMWF – Reading

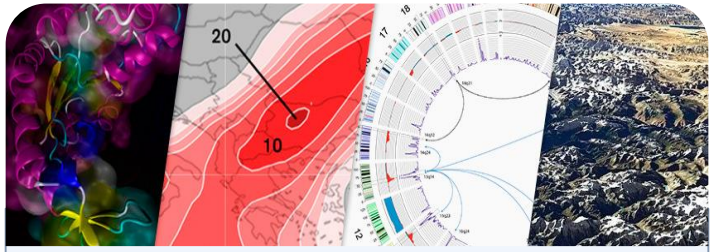
27/11/2019

The Performance Team in BSC Earth Sciences

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	(1.11.2019)
CTE3-Arm V8:	0.5 Pflops	(????)

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

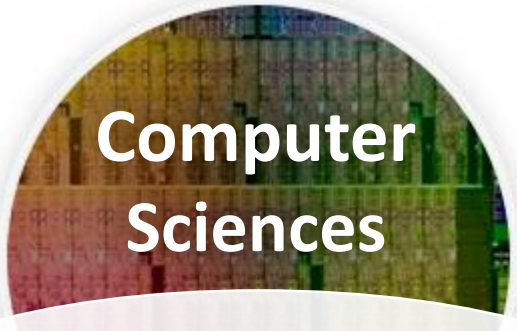
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Generalitat de Catalunya
Departament d'Empresa i Coneixement

Gobierno de España
Ministerio de Ciencia, Innovación y Universidades

Universitat Politècnica de Catalunya
BarcelonaTech

Mission of BSC Scientific Departments



To influence the way machines are built, programmed and used: programming models, performance tools, Big Data, computer architecture, energy efficiency

To develop and implement global and regional state-of-the-art models for short-term air quality forecast and long-term climate applications

To understand living organisms by means of theoretical and computational methods (molecular modeling, genomics, proteomics)

To develop scientific and engineering software to efficiently exploit super-computing capabilities (biomedical, geophysics, atmospheric, energy, social and economic simulations)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Earth Science Department

Environmental modelling and forecasting, with a particular focus on weather, climate and air quality

Director: **Francisco Doblas-Reyes**

- >100 people
- Leading: H2020 project, COPERNICUS contract, ERC Consolidator Grant and hosts an AXA Chair

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Earth Sciences

Environmental modelling and forecasting, with a particular focus on weather, climate and air quality

Modeling
air quality and
Sand and Dust Storms
Processes from urban to Global and the impacts on weather, health and ecosystems

Research and Forecasts

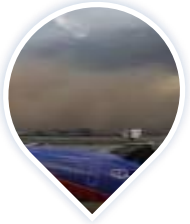
Climate
predictions system from subseasonal-to-decadal forecasts

Service Users Sectors

Infrastructures

Solar Energy

Urban development



Transport

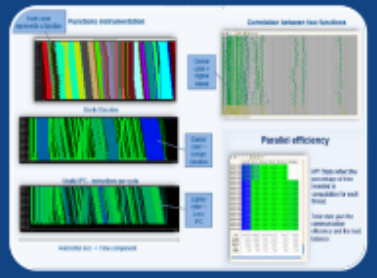
Wind Energy

Agriculture

Insurance

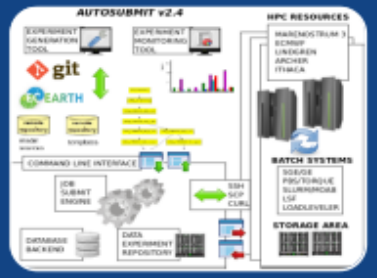
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Computational Earth Sciences



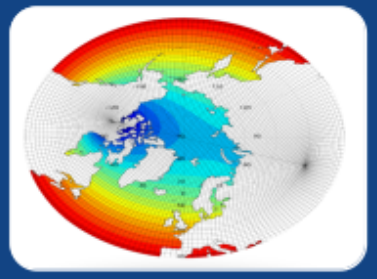
Performance Team

- Provide HPC Services (profiling, code audit, ...) to find main bottlenecks of our operational models
- Research and apply new computational methods for current and new platforms



Models and Workflows Team

- Development of HPC user-friendly software framework
- Support the development of atmospheric research software



Data and Diagnostics Team

- Big Data in Earth Sciences
- Provision of data services
- Visualization

Performance Team

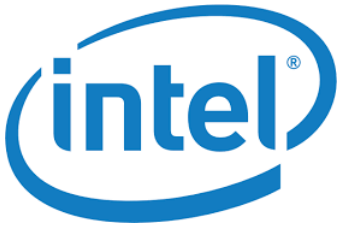
Knowledge about the **mathematical** and **computational** side of Earth System applications

Knowledge about the specific needs in **HPC** of the Earth System applications

Researching about **HPC methods specifically used** for Earth System applications

Performance Team

- Necessary refactoring of numerical codes gaining a lot of attention and stirring many discussions
  - Computational performance analysis and new optimizations are needed for actual numerical models.
  - Studying new algorithms for the new generation of high performance platforms (path to exascale).
- Collaborating with several institutions on different projects at different scale



High Performance Computing in Earth Sciences

- Earth System Models (ESMs) are sophisticated tools with continuously increasing complexity:
    - More components of Earth System are included
    - Finer Spatial and Temporal resolutions
- This increase in complexity could be developed thanks to the important parallel advances in HPC

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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**

ORCA2

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.

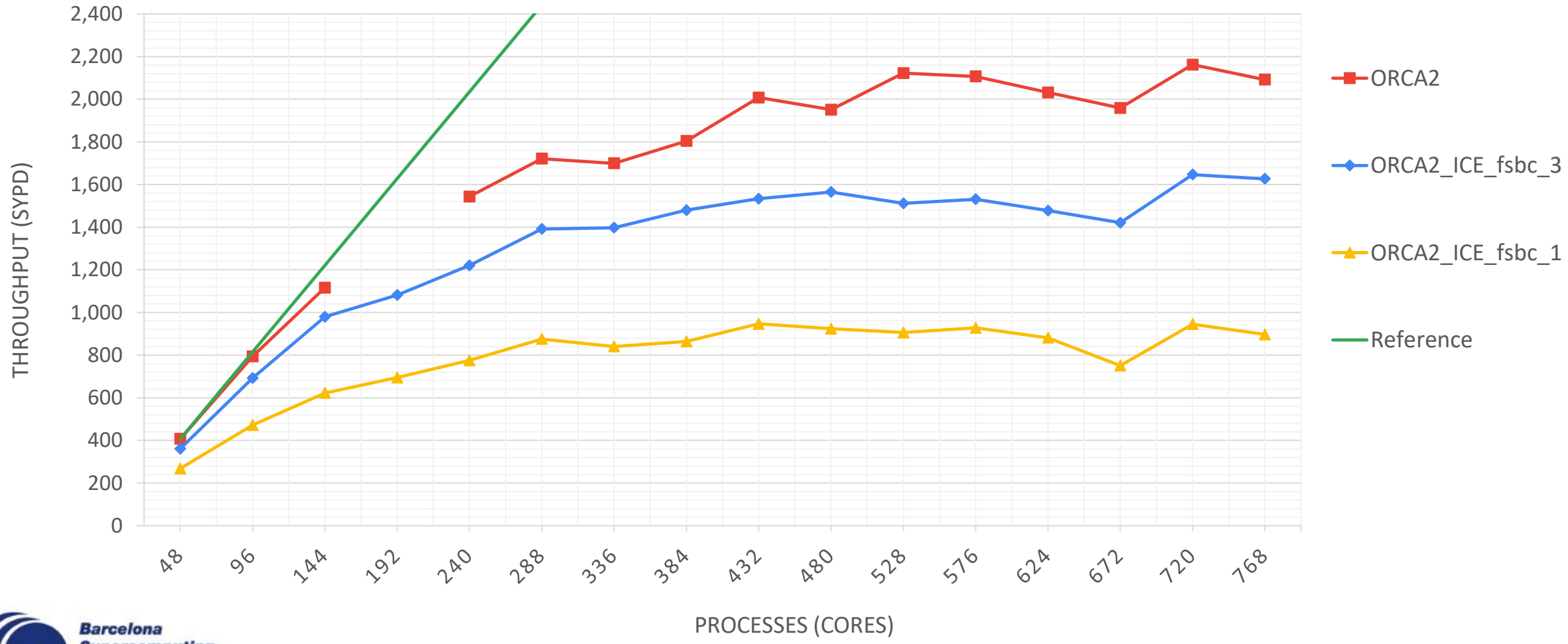
name	jpiglo	jpjglo	jpk	size (million vertices)	resolution (km)
ORCA2	182	149	31	0.84	20.19
ORCA1 (SR)	362	292	75	7.92	11[illegible]7
ORCA025 (HR)	1,442	1,021	75	110.42	27.79
ORCA12 (VHR)	4,322	3,059	75	991.57	9.[illegible]7
ORCA36 (VVHR?)	12,962	9,173	75	8,917.53	[illegible].09

x9.4
x14
x9
x9

x10,650

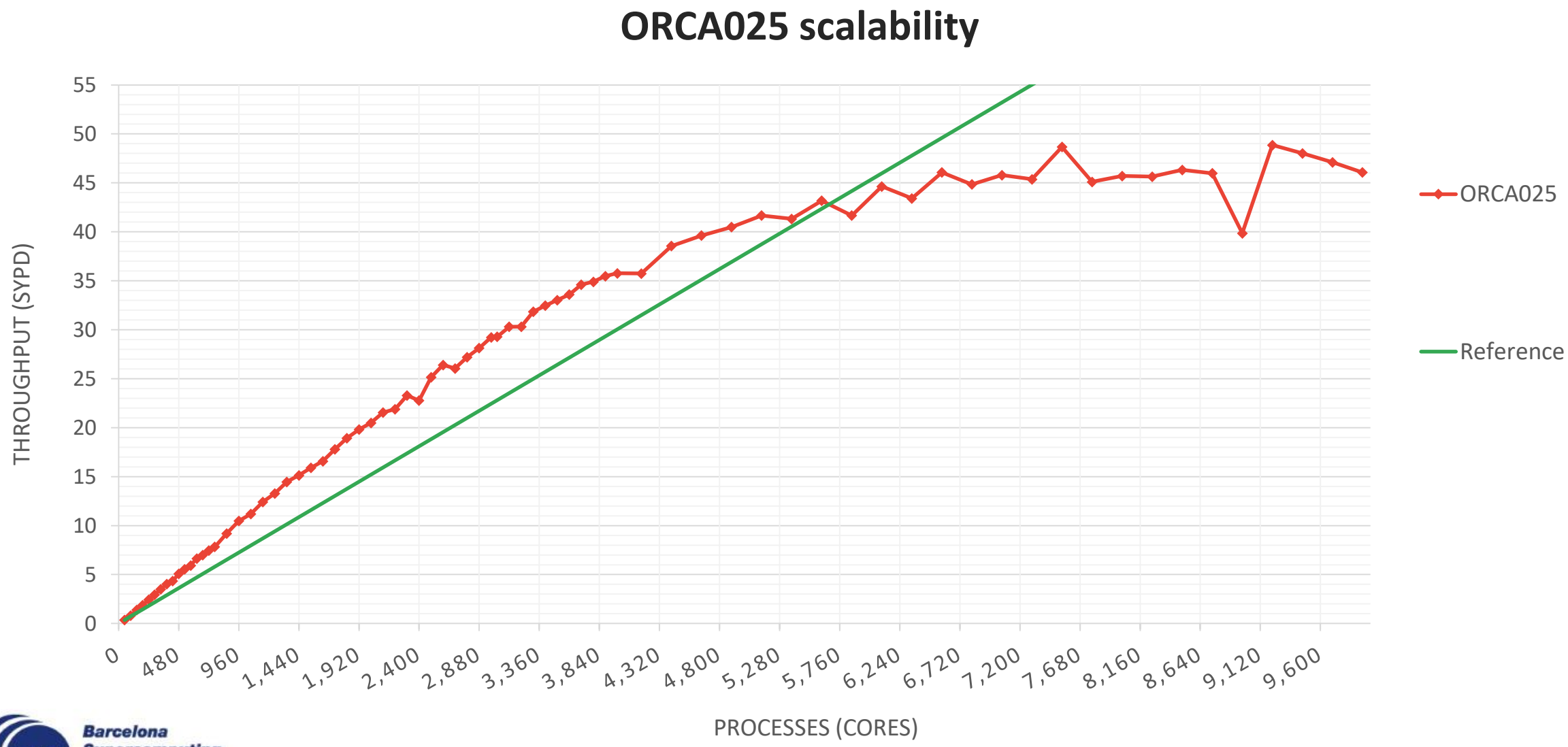
ORCA2 scalability

ORCA2 scalability and SI3 coupling frequency impact



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA025 scalability (MN4)



ORCA36

Configurations

Code	dom	tsd	sbc				qsr	lbc	bbc	traldf
	rdt	init	usr	blk dm2dc	ice	rnf	rgb/ chldta 2bd	shlat	trabbc	hor/triad
O36-I	90	F	T	F	F	F	T/F	0.0	F	T/F
O36-II	90	F	F	T**	F	F	T/F	1.0	F	T/F
O36_ICE	90	F	F	T**	T	F	T/F	1.0	F	T/F
O36_FULL*	30	T	F	T**	T	T	F/T	0.0	T	F/T

** rn_bt_max = 0.8 (instead of 0.6) and nn_baro=60 (instead of 30)*

** ln_NCAR = true

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

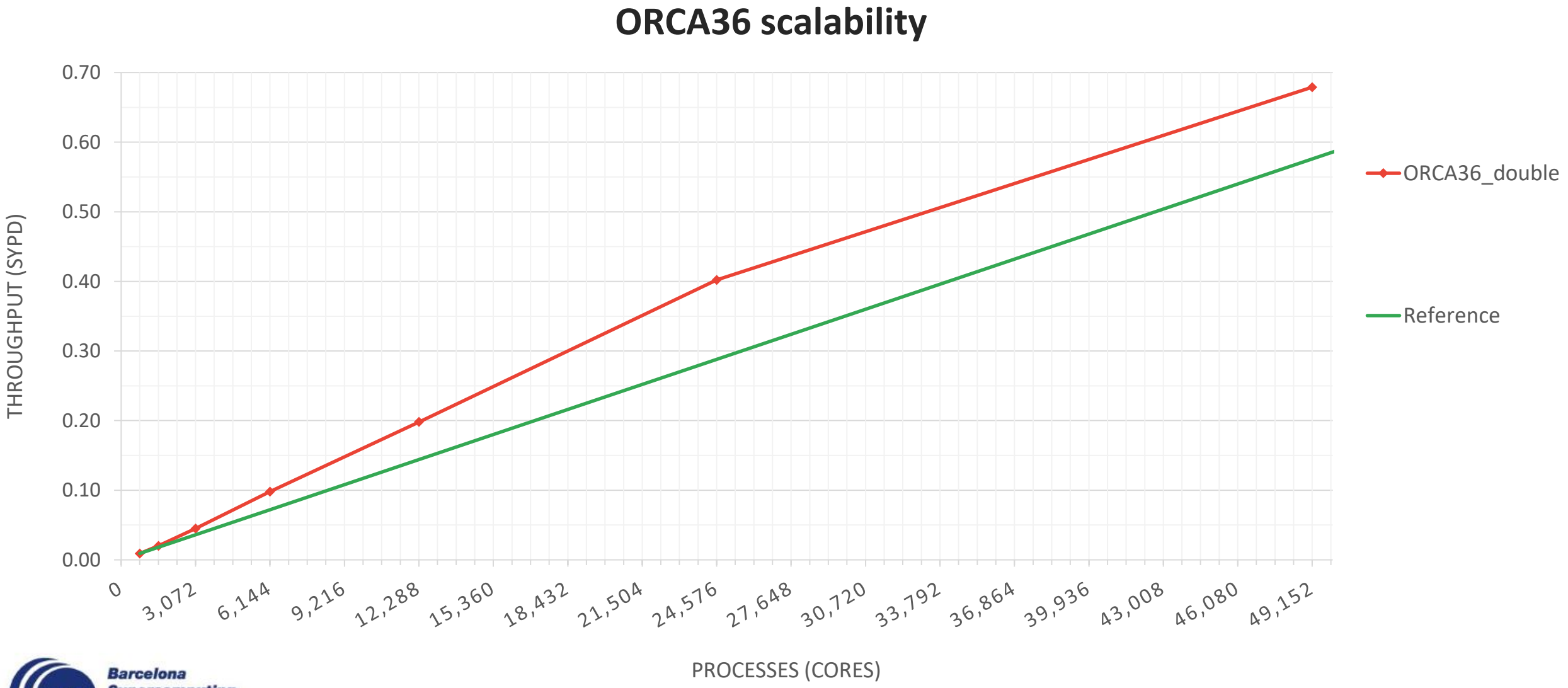
Resources constraints

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

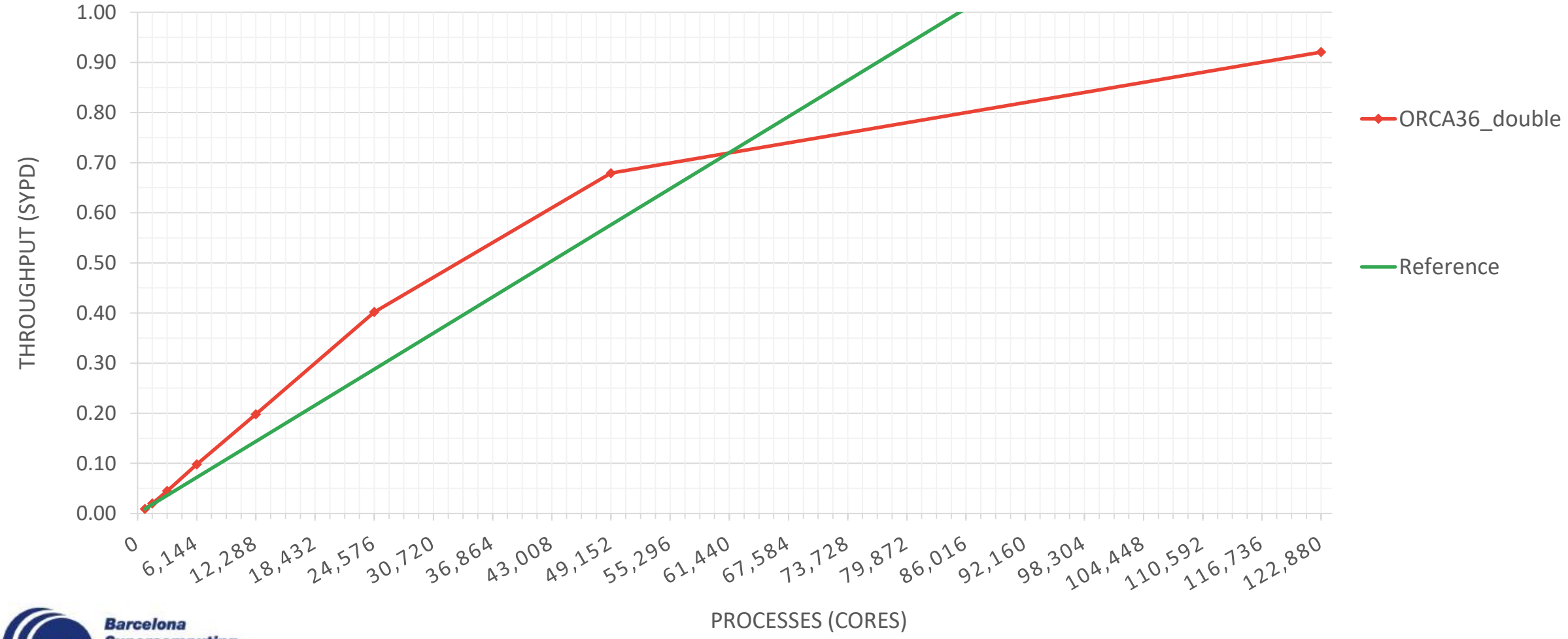
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability (MN4)

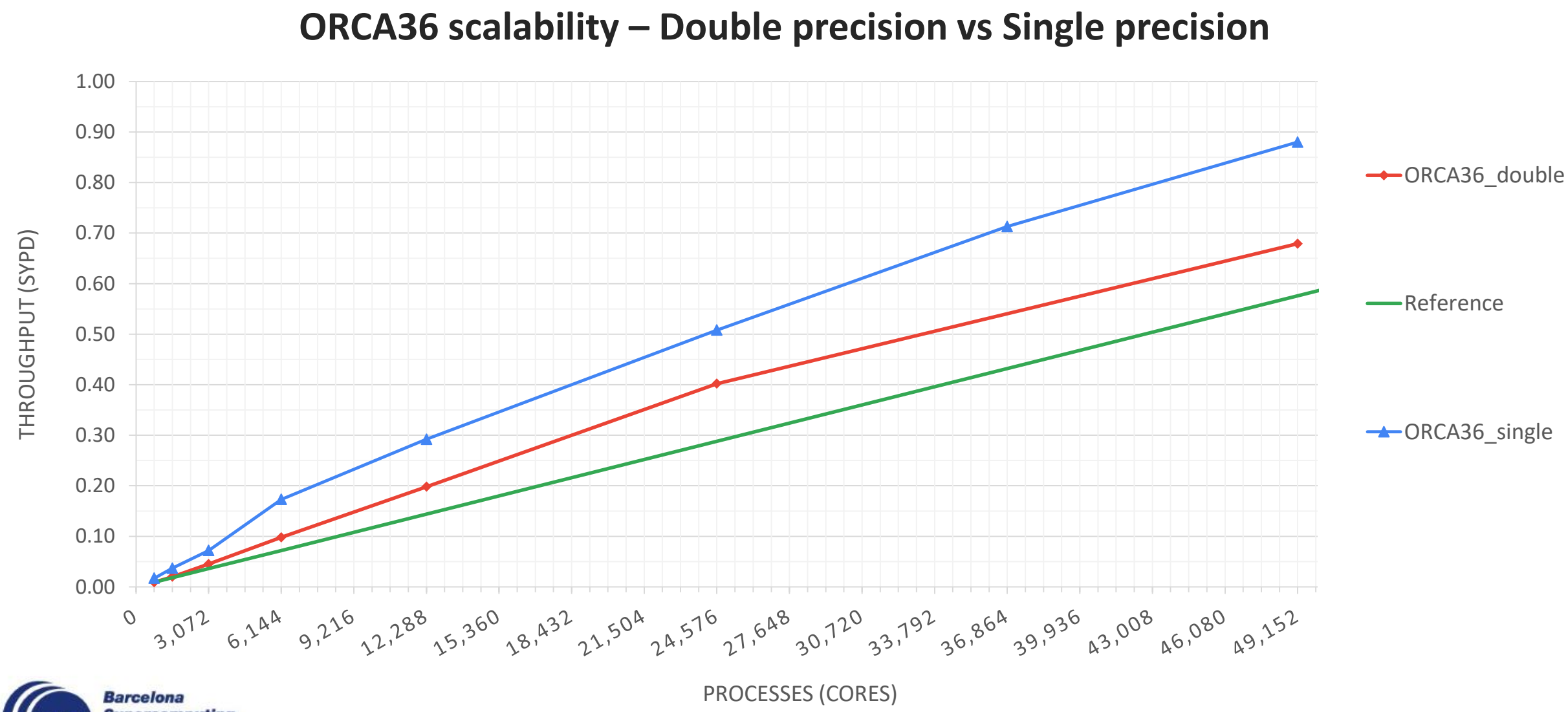


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability (MN4)

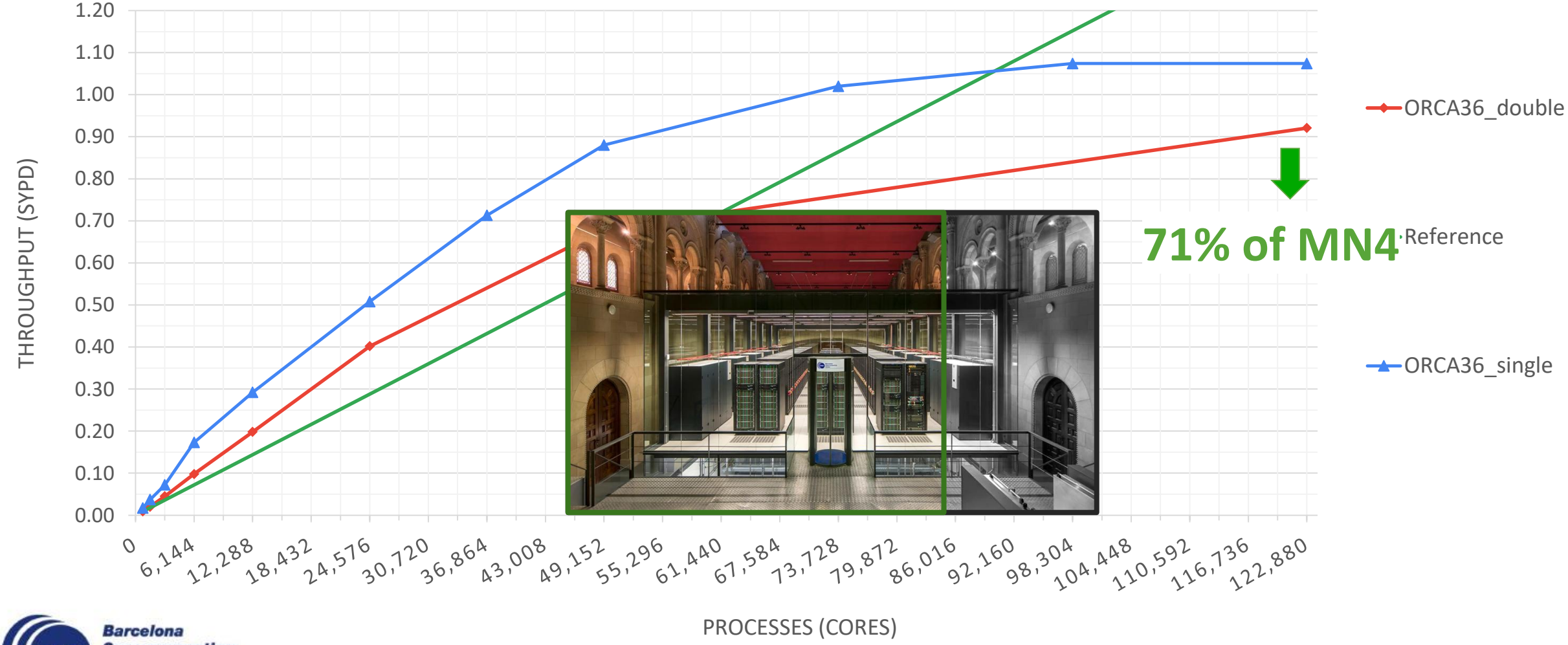


ORCA36 scalability (MN4)



ORCA36 scalability (MN4)

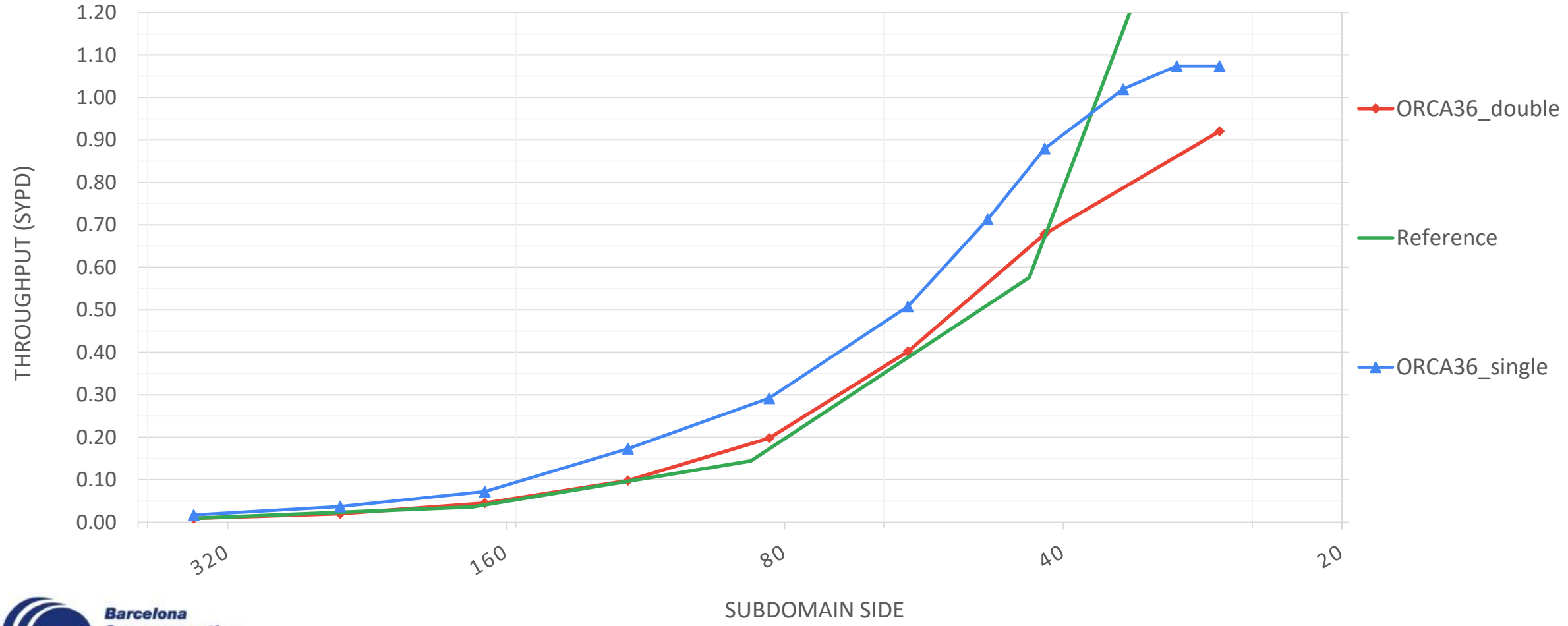
ORCA36 scalability – Double precision vs Single precision – Grand challenge



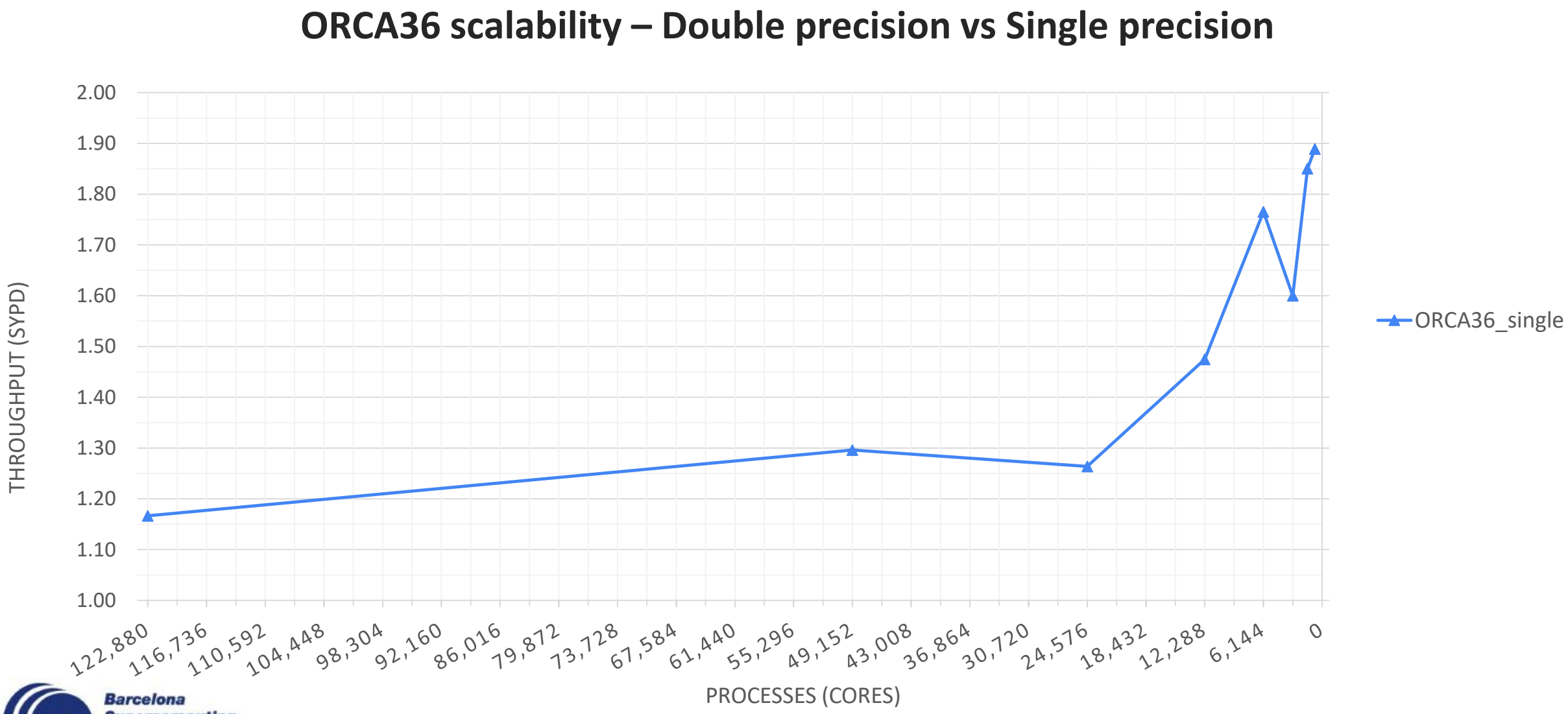
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability (MN4)

Throughput per subdomain side (subdomain area = side²)

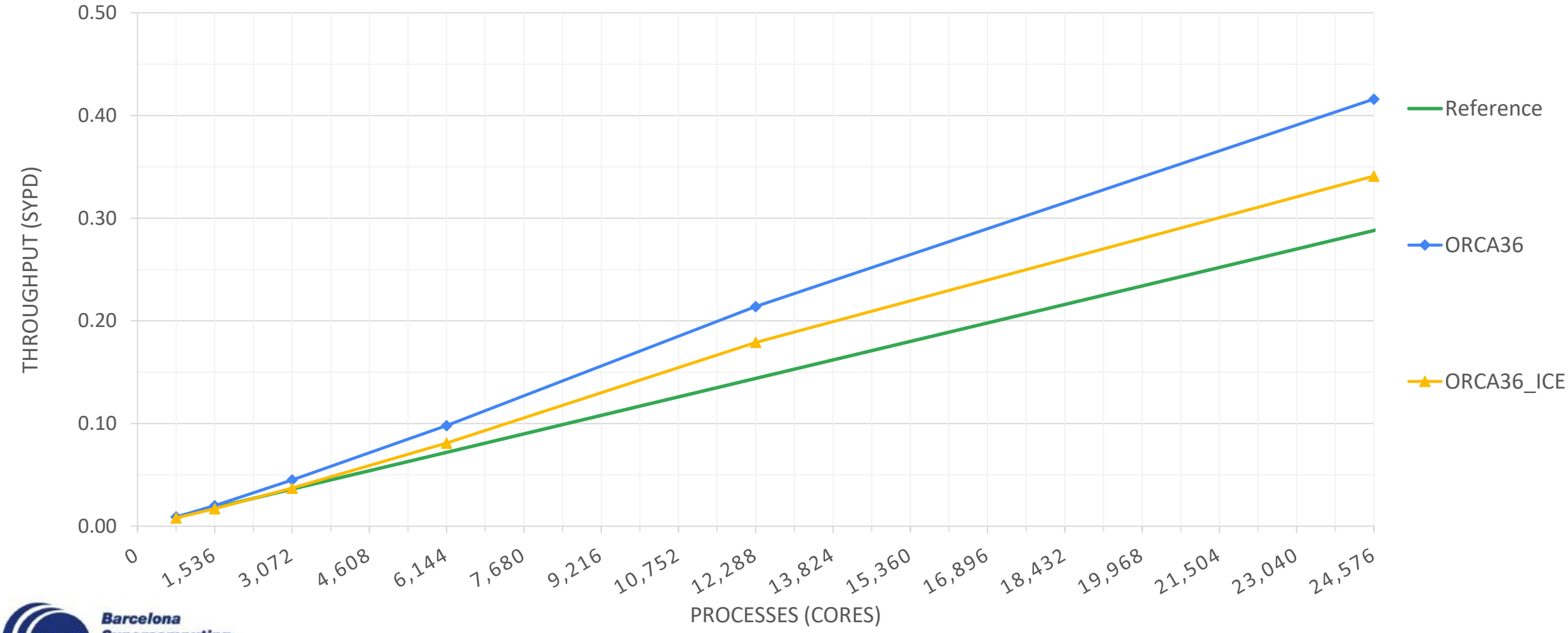


ORCA36 scalability (MN4)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability (MN4)



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

NEMO4 weak scaling

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

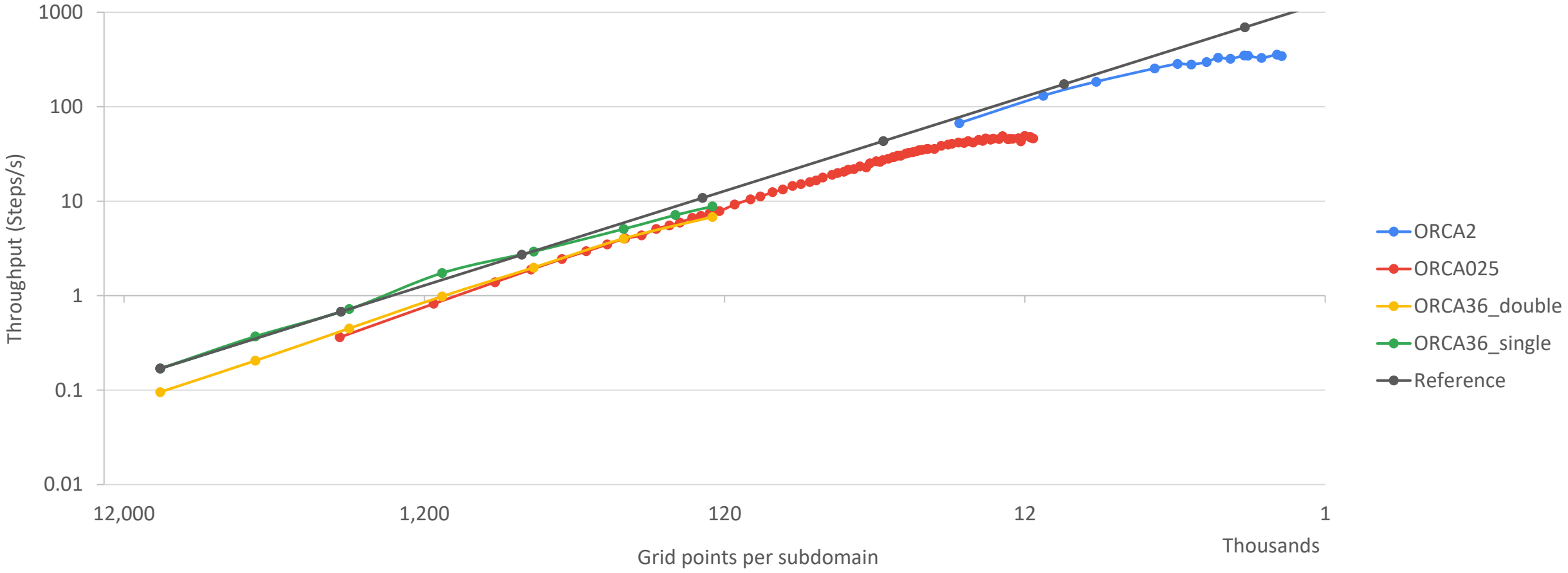
ORCA weak scaling (MN4)

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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA weak scaling (MN4)

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



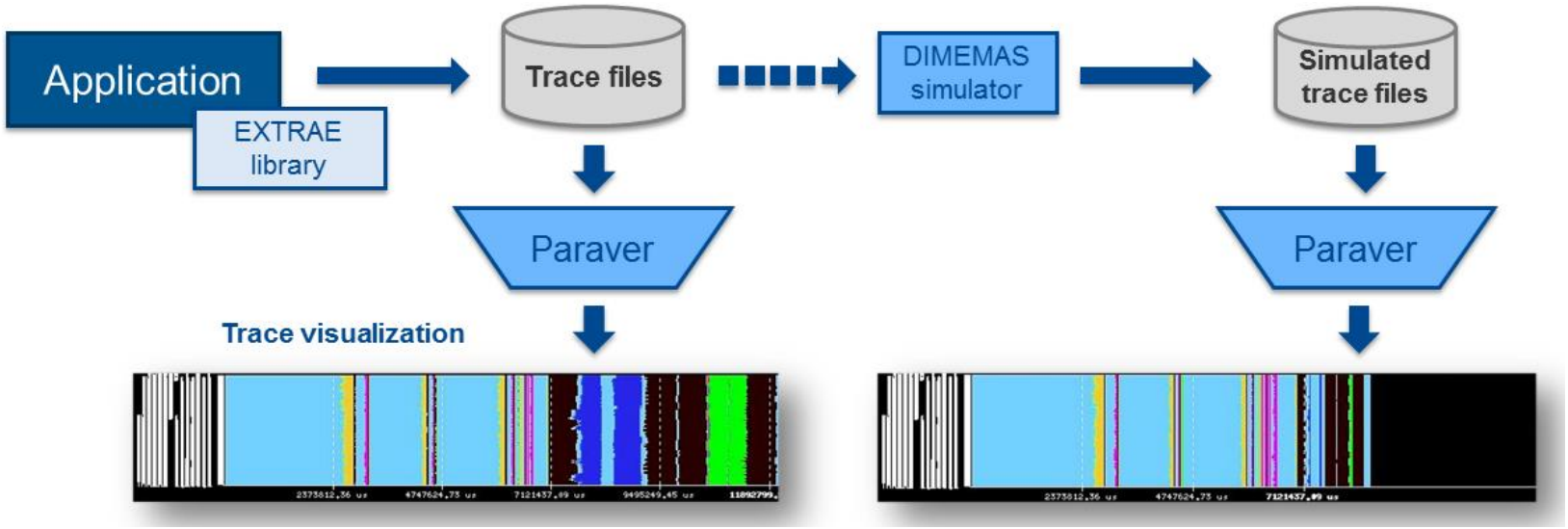
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 Performance analysis

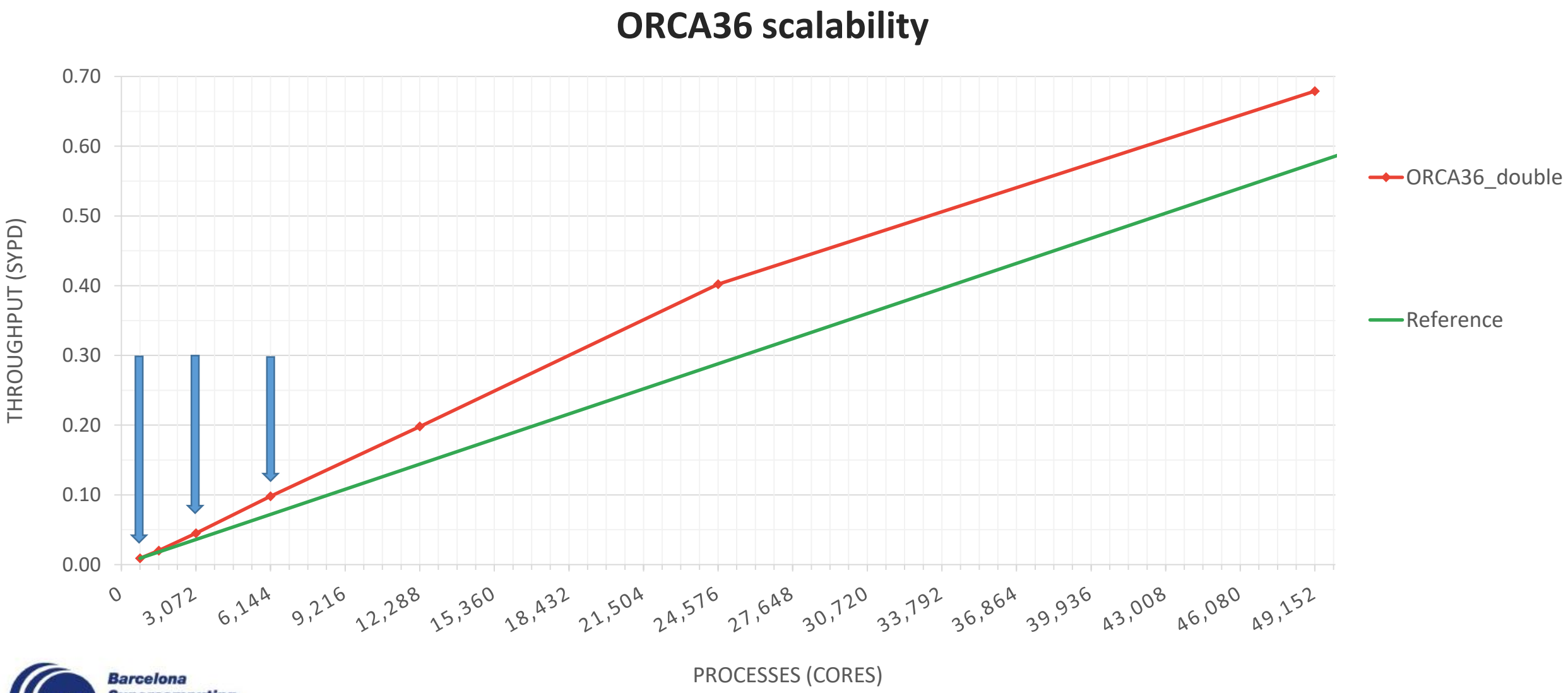
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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



ORCA36 scalability (MN4)



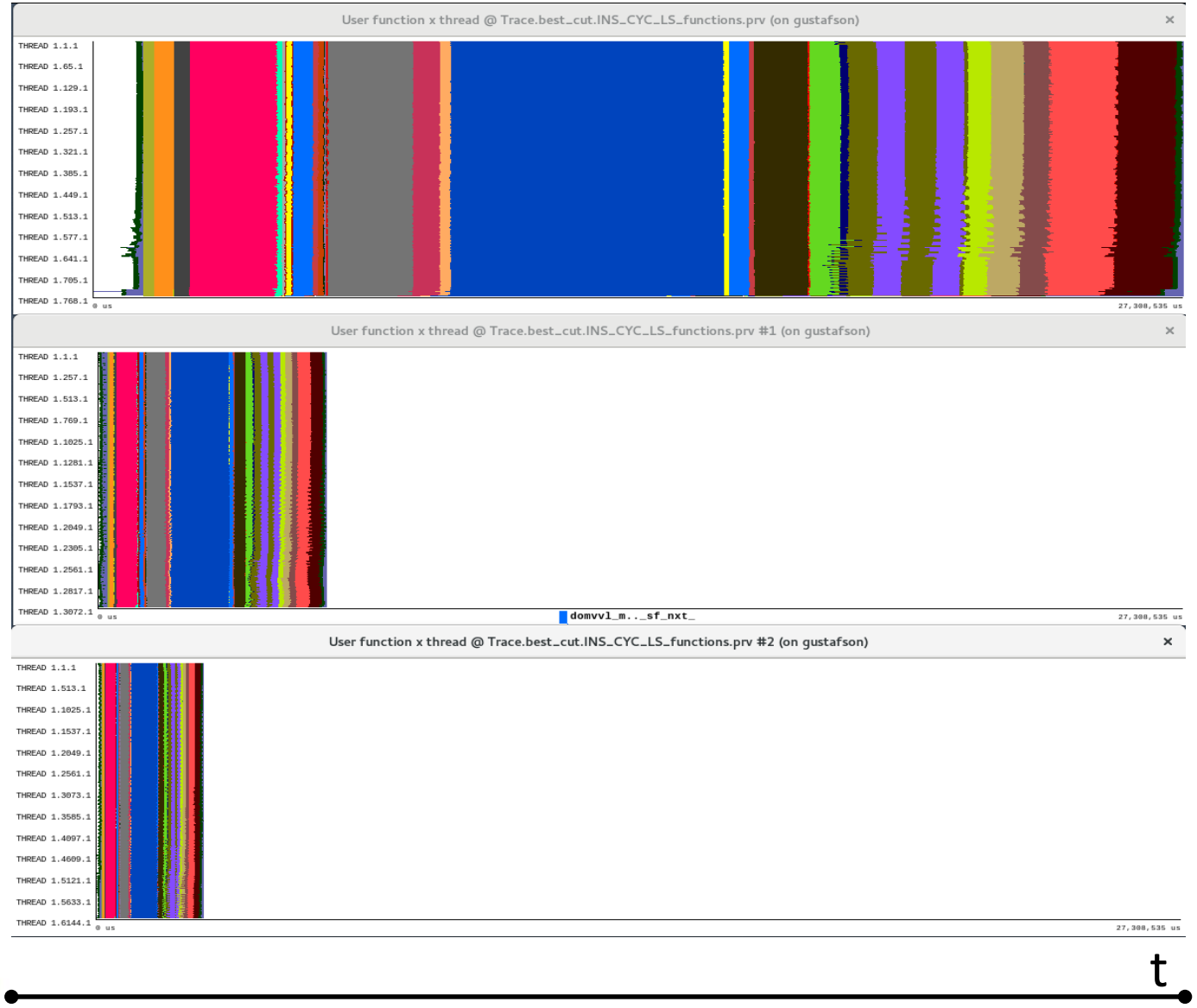
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 functions view

768

3,072

6,144

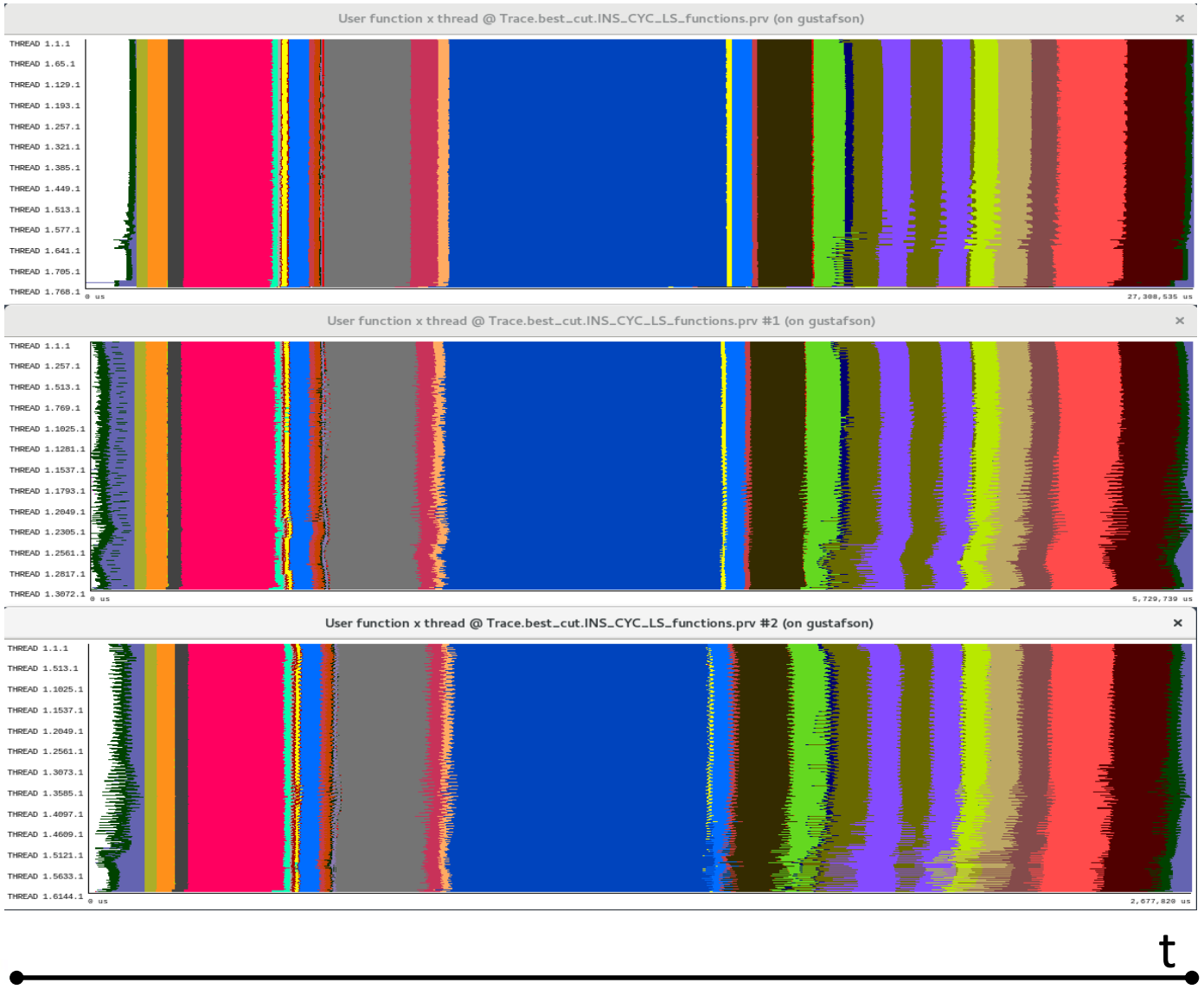


ORCA36 functions view

768

3,072

6,144

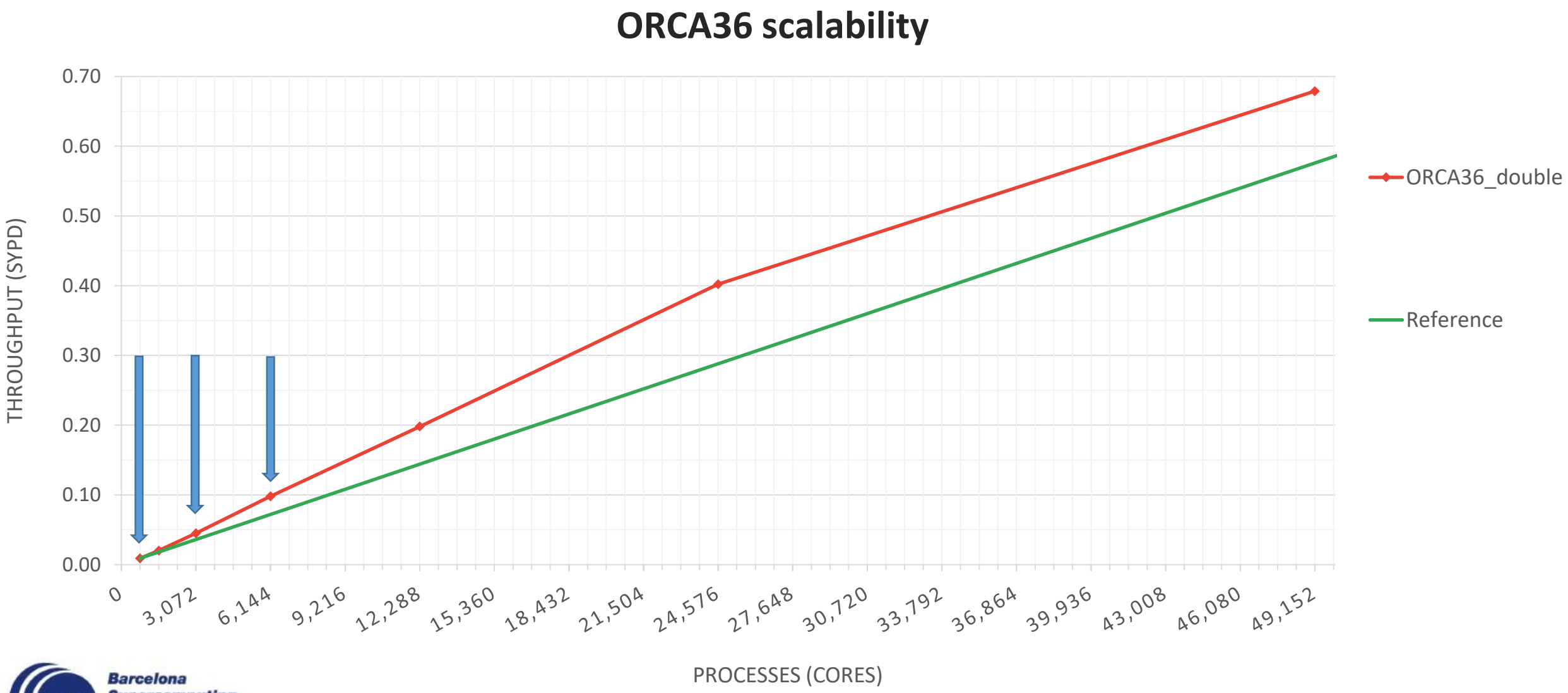


ORCA36 model factors

Model factors explaining scalability on 16, 32 and 64 nodes

Number of processes	**768**	**3,072 (x4)**	**6,144 (x8)**
Parallel efficiency	93.72	85.74	82.65
Load balance	97.42	92.4	92.74
Communication efficiency	96.2	92.79	89.12
Computation scalability	100	130.25	144.55
Global efficiency	**92.72**	**111.67**	**119.47**
IPC scalability	100	123.47	145.99
Instruction scalability	100	102.63	97.18
Frequency scalability	100	102.78	101.88
Speedup	**1.00**	**4.77**	**10.21**
Average IPC	0.29	0.35	0.42
Average frequency (GHz)	2.09	2.15	2.13

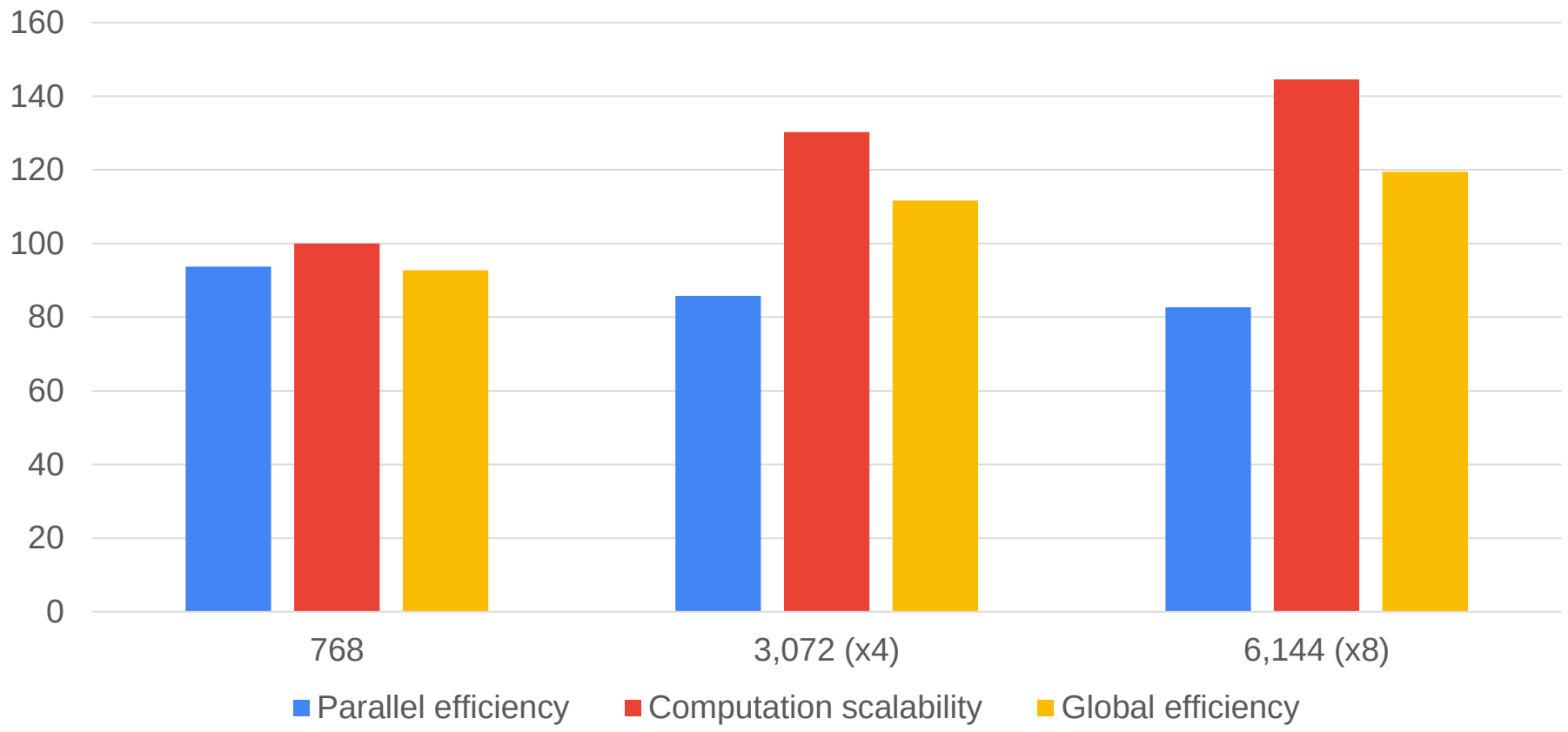
ORCA36 scalability (MN4)



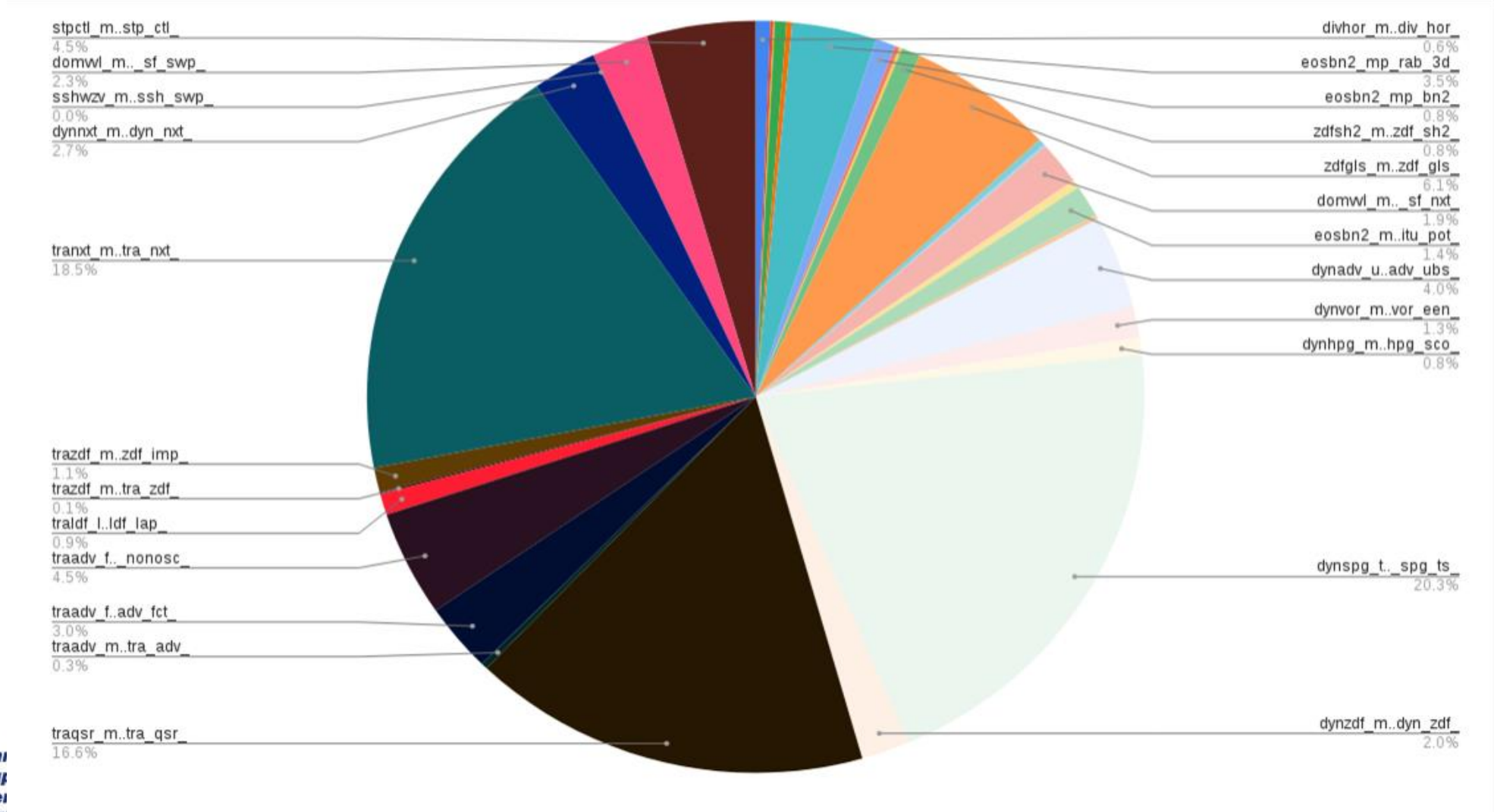
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

ORCA36 scalability

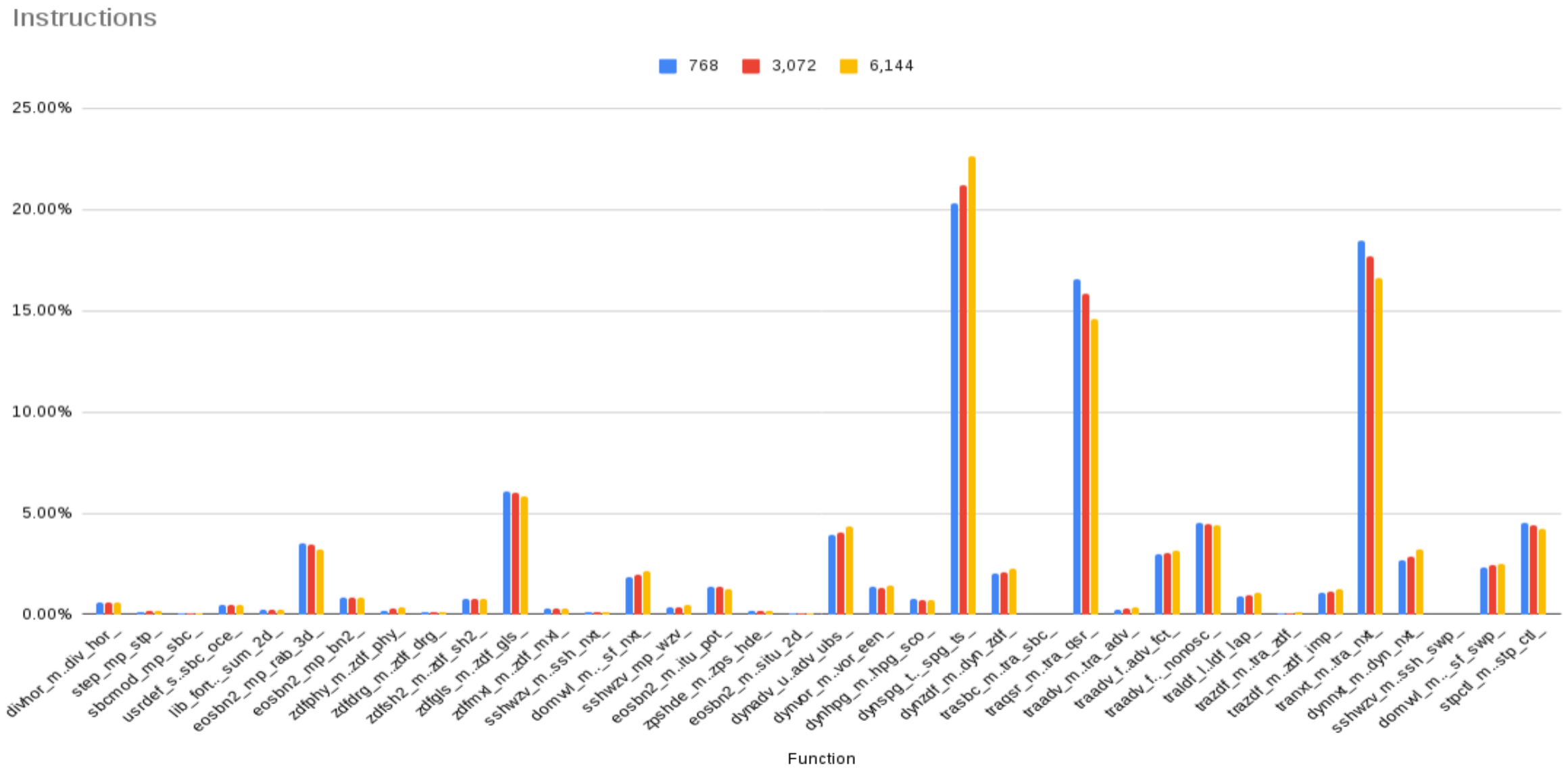
Model factors explaining scalability on 16, 32 and 64 nodes



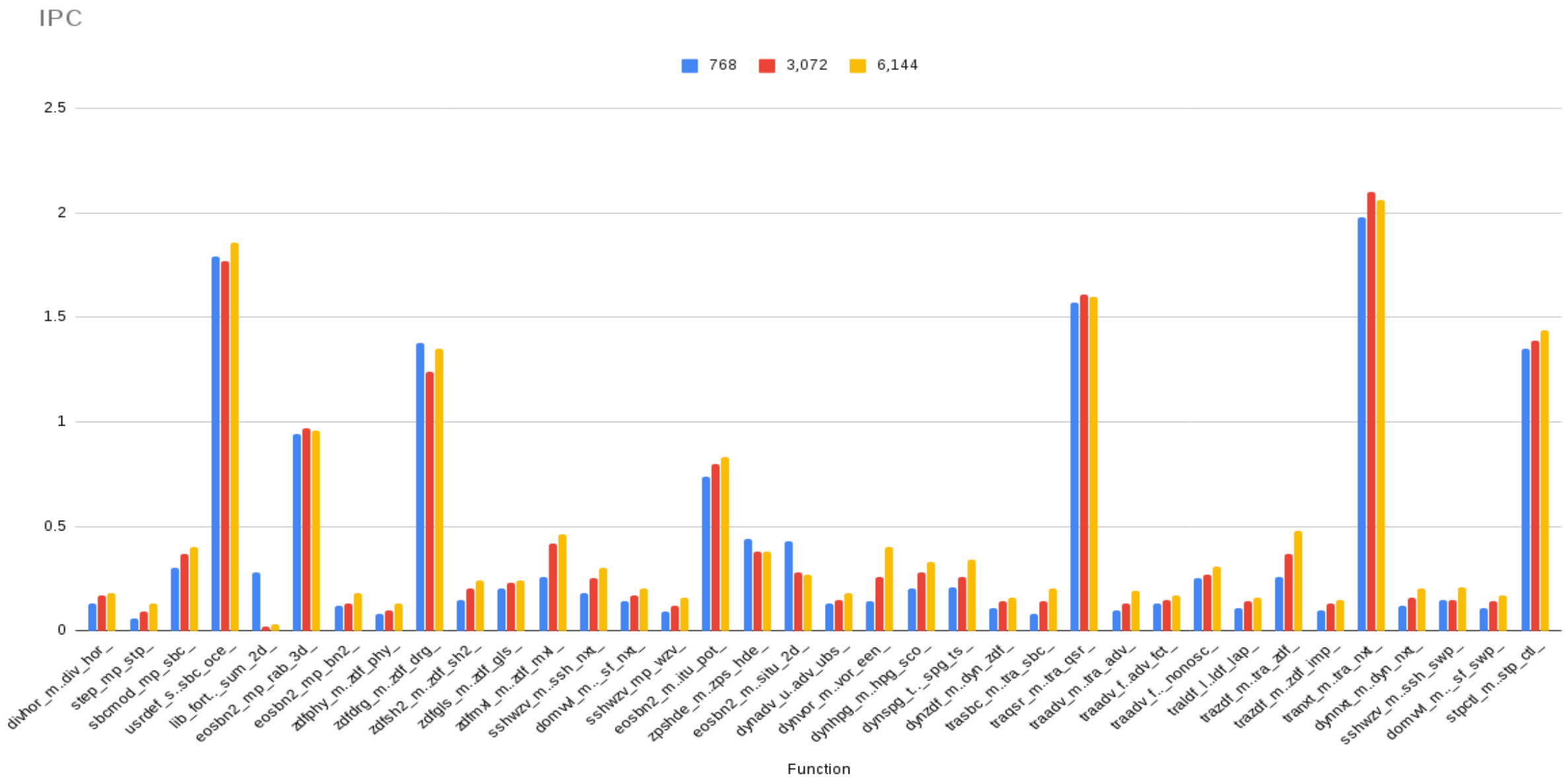
ORCA36 instructions breakdown



ORCA36 instructions breakdown



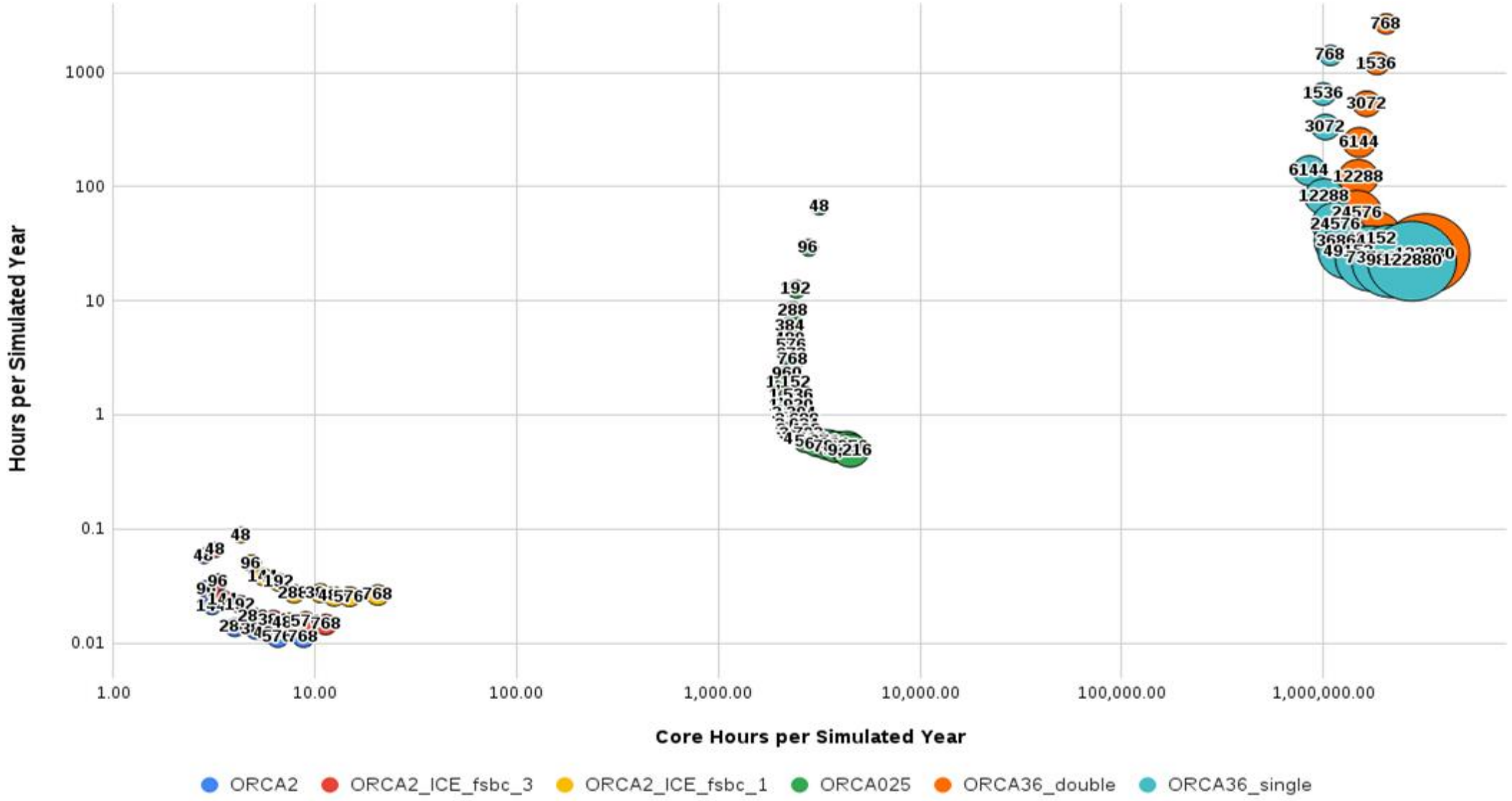
ORCA36 IPC per function



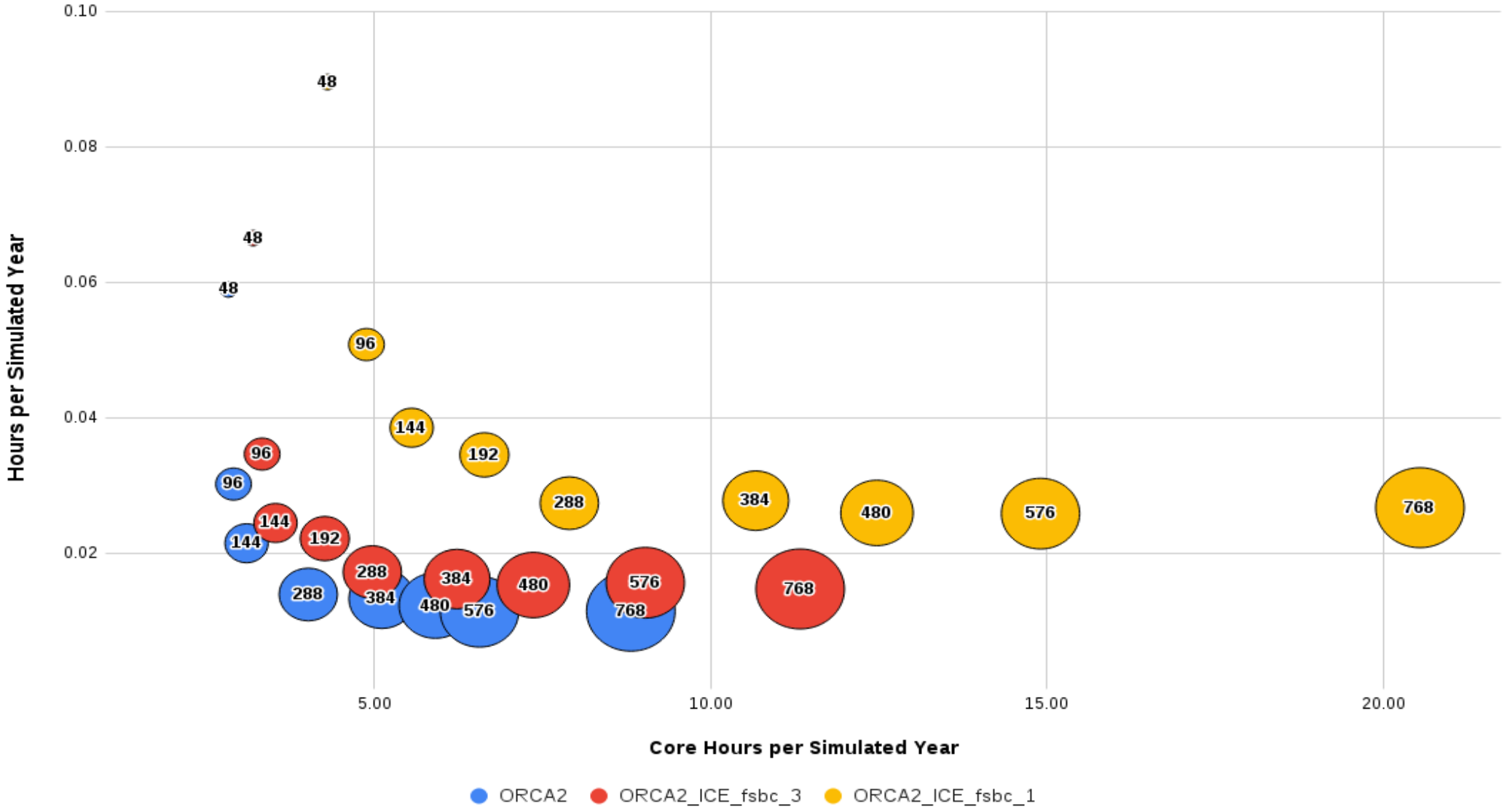
NEMO4 time vs cost

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

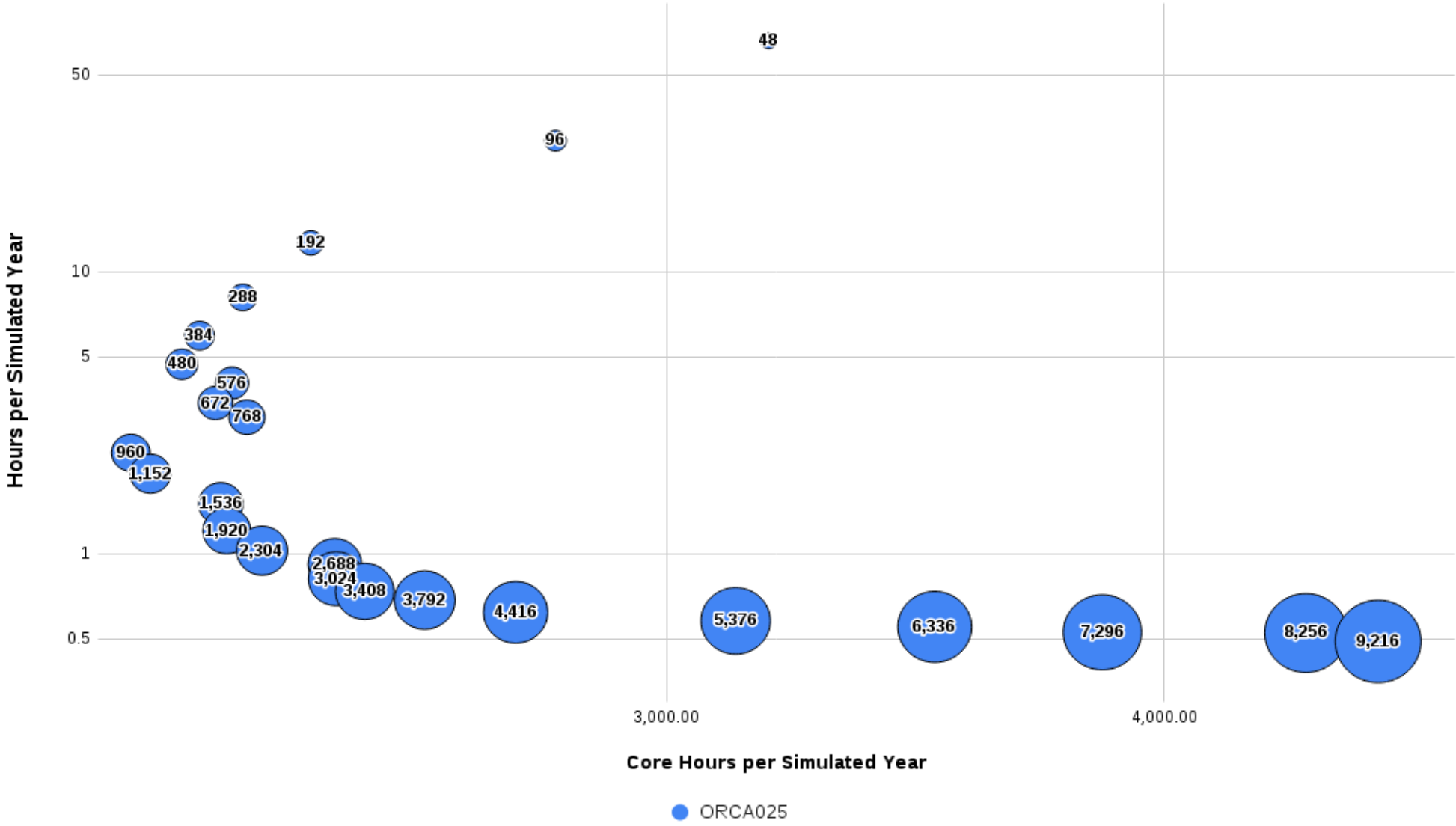
NEMO4 time vs cost



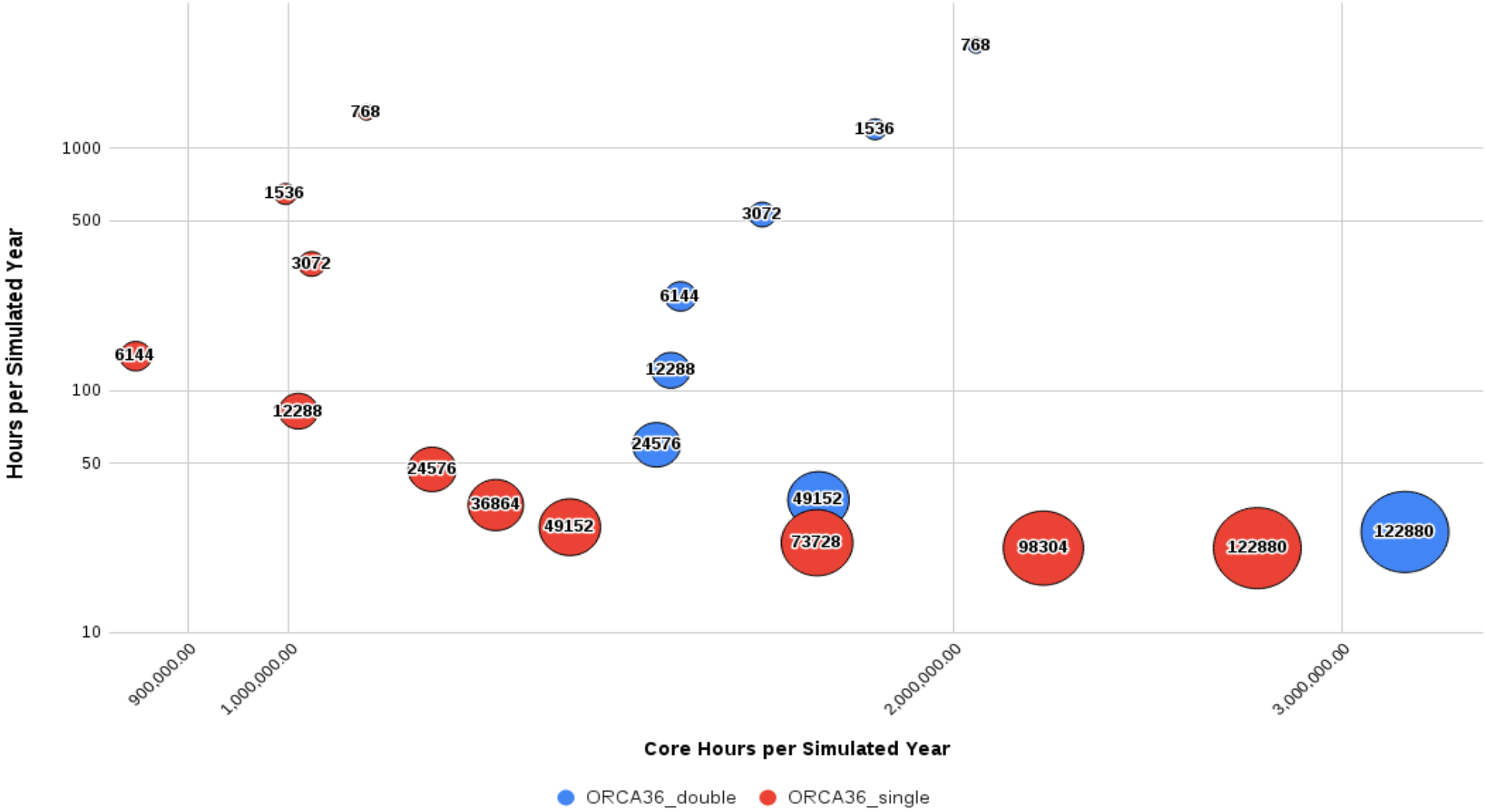
NEMO4 time vs cost



NEMO4 time vs cost



NEMO4 time vs cost



Conclusions

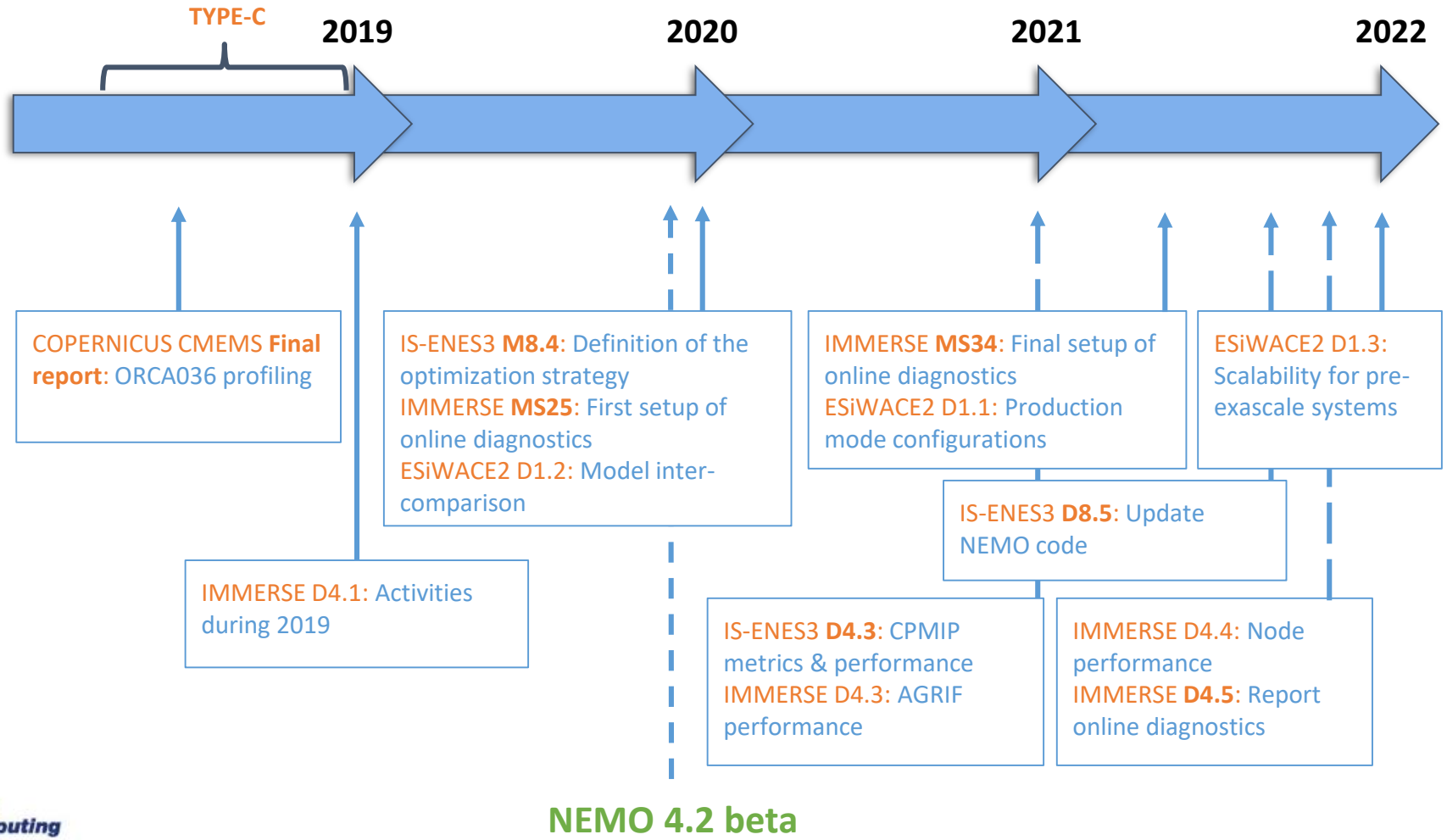
ORCA36 “final” configuration

- **Bathymetry:** 12,962x9,173 (domain_cfg.nc 485G)
- **Initialisation:** T&S: Files in O36 grid (no interpolation).
- **Forcing:**
  - **Surface** boundary condition:
    - **Ice** coupled every step
    - **Bulk** formulae: ERAinterim files (512x256). Online interpolation to O36.
    - **Runoffs**: Enabled. Input files in O36 grid (no interpolation).
  - **Solar** radiation: Files in O36 grid (no interpolation).
  - **Bottom** boundary condition: Files in O36 grid (no interpolation).
- **Timestep:** 30 (90) secs

Conclusions and ideas for the future

- **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.
- **XIOS is also an issue.** It was very difficult to run the new configuration with XIOS2.5.

NEMO timeline in BSC-ES performance



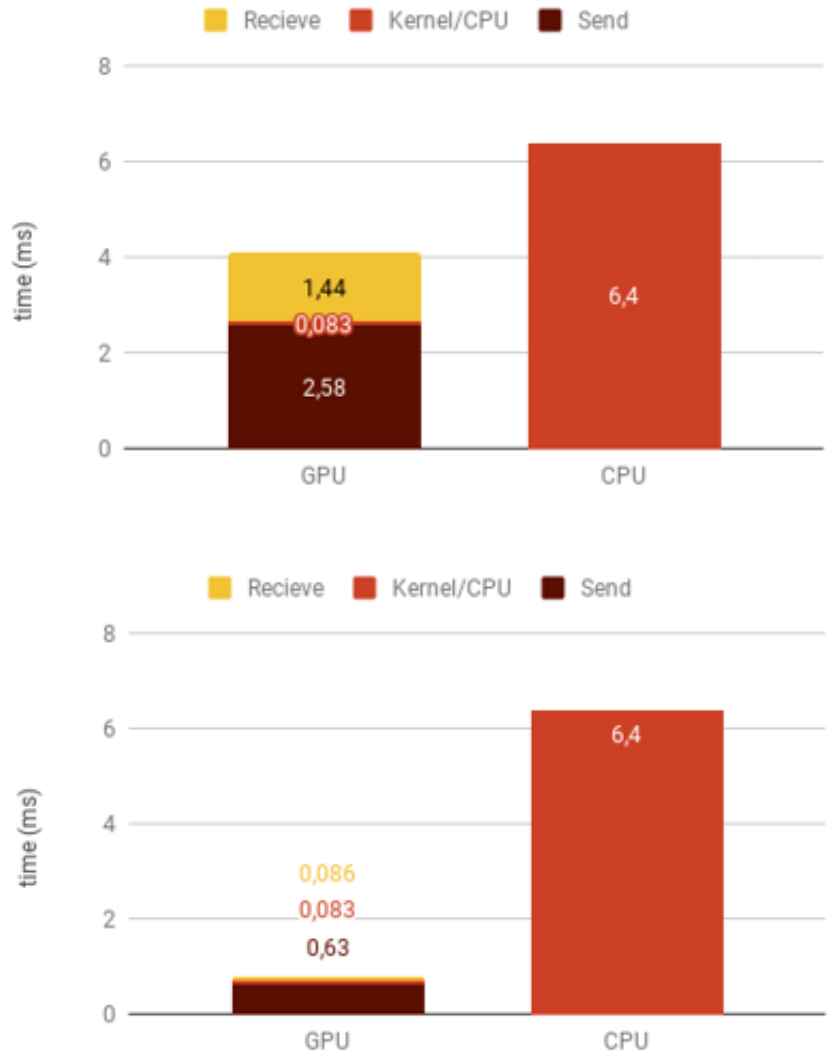
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

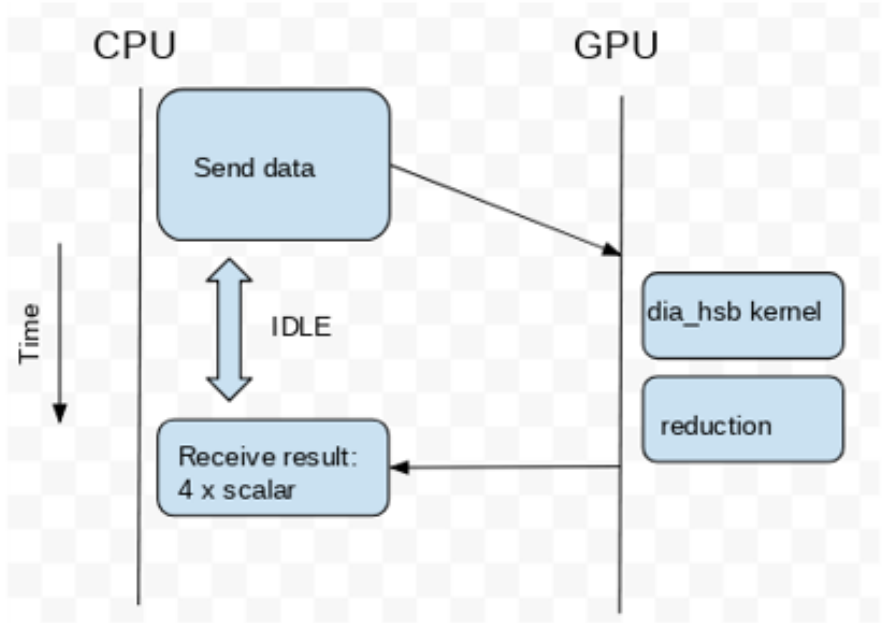
* Bold indicates BSC is the lead

Porting NEMO diagnostics to GPUs

IMMERSE: Porting diagnostics to GPUs

The diagnostics dia_hsb kernel





Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

MERCATOR OCEAN INTERNATIONAL

immerse
IMPROVING OCEAN MODELS FOR THE COPERNICUS PROGRAMME

Thank you

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

miguel.castrillo@bsc.es