Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Automatic approach to balance coupled ESMs:

Saving energy and time on coupled ESMs

Sergi Palomas, Dr. Mario Acosta, Dr. Etienne Tourigny

January, 2023

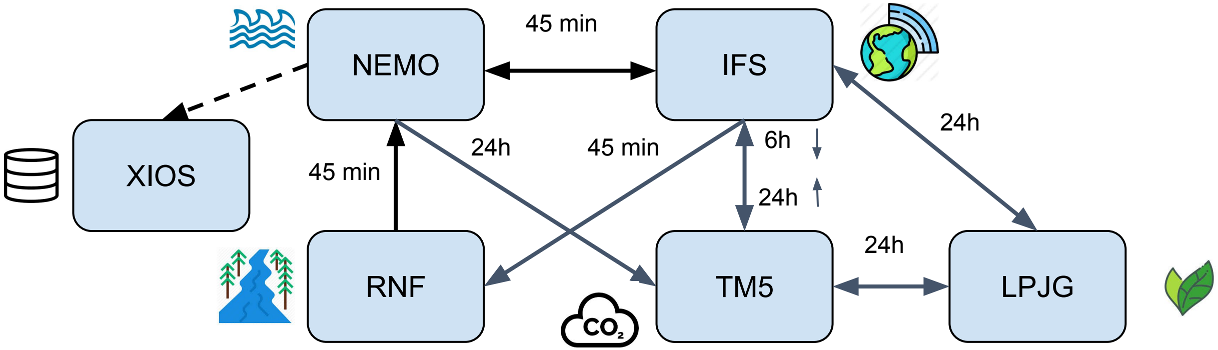
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Introduction

ESMs are commonly built from different independent components

Each simulates a specific natural phenomenon

Multi-Program Multi-Data (MDMP) application

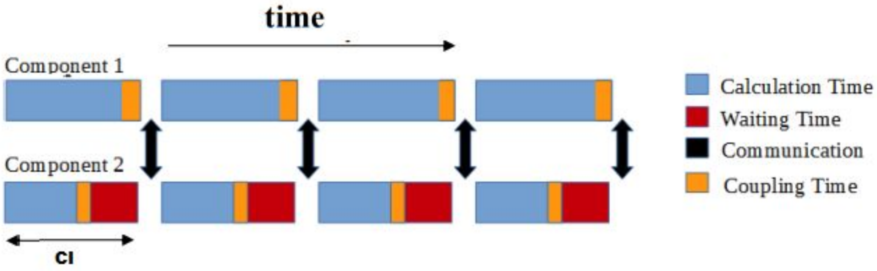


Introduction

ESMs are commonly built from different independent components

Each simulates a specific natural phenomenon

Multi-Program Multi-Data (MDMP) application



Components exchange information during the simulation → the fastest components wait for the slowest

Introduction

Coupled ESMs parallel efficiency is reduced due to the load-imbalance:

- **Dependencies**

  Components have to be synchronized to exchange data during the run. Different component calculation times, irregular ts and the algorithms used to regrid the data make the interactions between them complex

- **Parallelization**

  Components may not be able to run at their optimal scalability point but rather at one that is better for the whole coupled execution

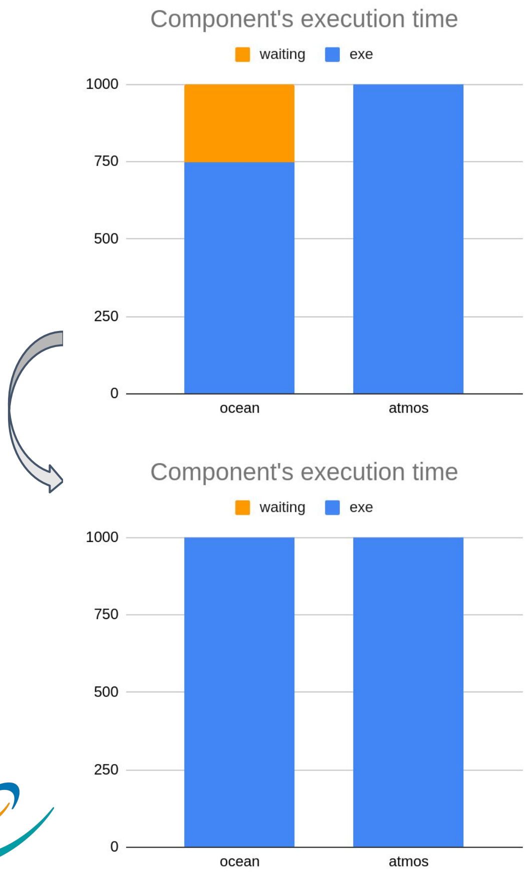
Introduction

Typical approach

Coupled ESMs parallel efficiency is reduced due to the load-imbalance:

- **Dependencies**

  A setup where all component's execution time is the same "ensures" that the load-imbalance is minimized



Introduction

Typical approach

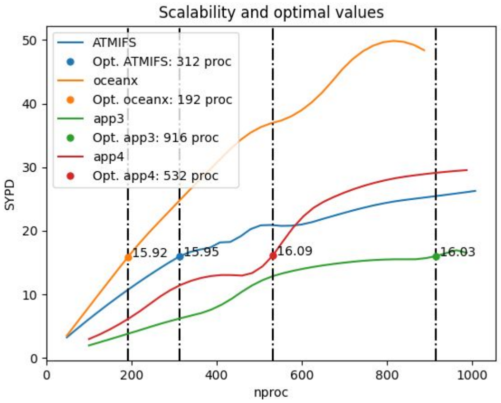
Coupled ESMs parallel efficiency is reduced due to the load-imbalance:

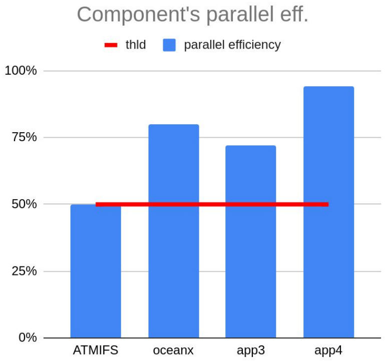
- **Dependencies**

  A setup where all component's execution time is the same "ensures" that the load-imbalance is minimized

- **Parallelization**

  The total number or PEs to give to each component is the maximum as far as all components **parallel efficiency** is kept **over 50%**





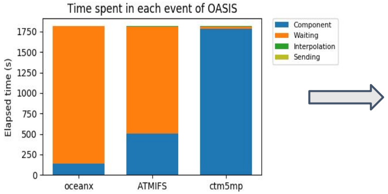
Objectives

Current methods to find the optimal number of PEs to assign to each component in coupled ESMs give suboptimal solutions

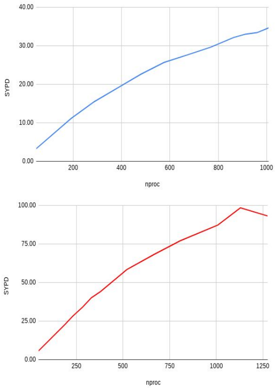
- Define a **metric** to evaluate the **performance of coupled ESMs** and control the time / energy tradeoff
- Create a **methodology to find the best resource configuration** → No changes to the sources of the models but only how many PEs are allocated to each one
- **Automatize the steps** (workflow manager) to require the minimum user intervention

Objectives

An experiment with a poor load-balance



Get component's scalability properties



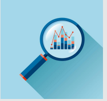
Load-balancing method

Prediction script

Automatic iterative process

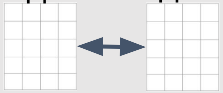
1. Run the ESM on the HPC

2. Get performance metrics

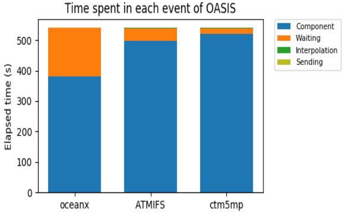


3. Modify the processors allocation

App. 1 App. 2



A balanced experiment



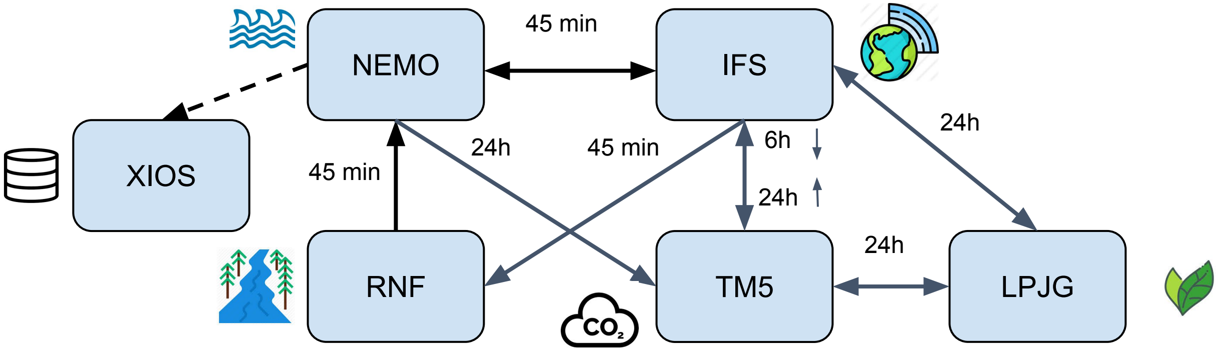
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

EC-Earth

The method has been used to optimize different configurations of EC-Earth3 in MareNostrum4 and ECMWF HPC machines

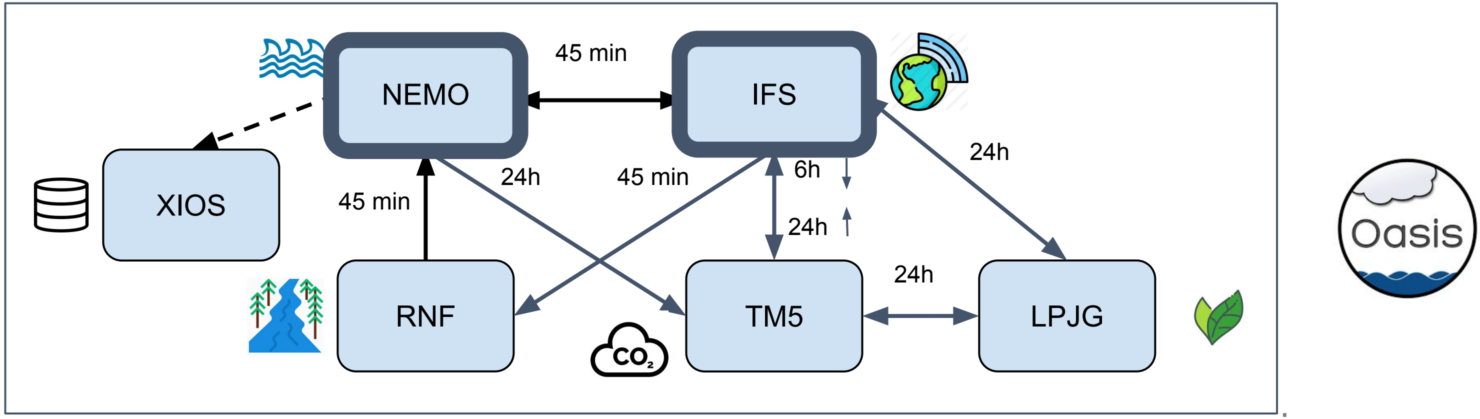
EC-Earth is a global coupled climate model made of multiple components and developed by a consortium of European institutions



EC-Earth

The method has been used to optimize different configurations of EC-Earth3 in MareNostrum4 and ECMWF HPC machines

EC-Earth is a global coupled climate model made of multiple components and developed by a consortium of European institutions



Performance metrics

Performance metrics

CPMIP: Model execution time and cost

- **SYPD**: total number of simulated years (SY) per 24 h of execution

  Metric of the Time-To-Solution (**TTS**)

- **CHSY**: the core-hours per SY

  Metric of the Energy-To-Solution (**ETS**)

$$CHSY = \frac{24 \cdot P}{SYPD} \qquad \text{Equation (1)}$$

P being the parallelization of the run

Performance metrics

CPMIP: Coupling overhead

- **Coupling cost:** total execution cost overhead due to coupling events (waiting, regridding, sending)

$$Cpl\_cost = \frac{T \cdot P - \sum_c T_C P_C}{T \cdot P}$$

T and *P* are the runtime and parallelization for the whole model, and *Tc* and *Pc* are the same for each component

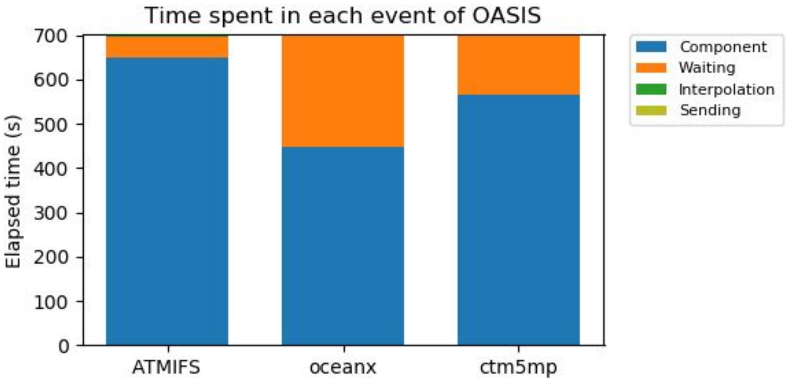
Performance metrics

CPMIP: Coupling overhead

- **Coupling cost:** total execution cost overhead due to coupling events (waiting, regridding, sending)

$$Cpl\_cost = \frac{T \cdot P - \sum_c T_C P_C}{T \cdot P}$$

T and *P* are the runtime and parallelization for the whole model, and *Tc* and *Pc* are the same for each component



Cpl_cost = 20.81%

Performance metrics

CPMIP: Coupling overhead

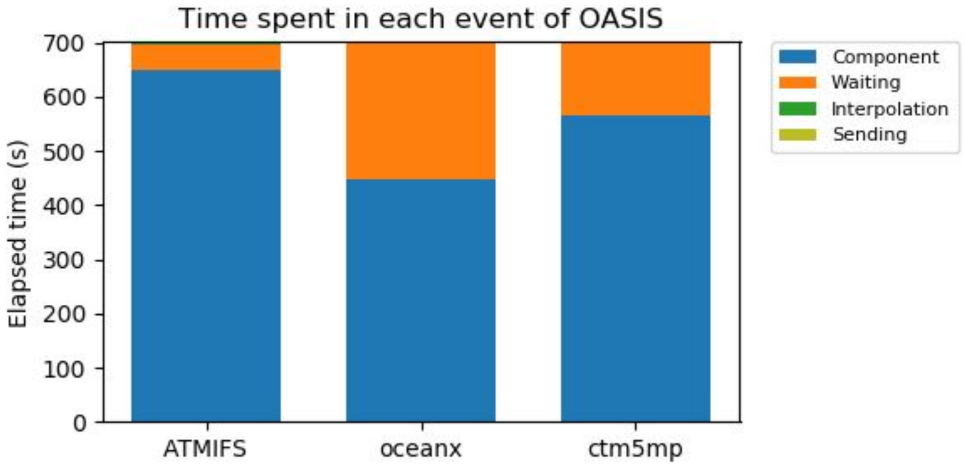
- **Coupling cost:** total execution cost overhead due to coupling events (waiting, regridding, sending)

$$Cpl\_cost = \frac{T \cdot P - \sum_c T_C P_C}{T \cdot P}$$

T and *P* are the runtime and parallelization for the whole model, and *Tc* and *Pc* are the same for each component

- Component coupling cost: how much each component adds to the total *Cpl_cost*

$$Component\_cpl\_cost = \frac{(T - T_C)\, P_C}{T \cdot P}$$



ATMIFS = 2.48
oceanx = 7.49
ctm5mp = 10.84

Cpl_cost = 20.81%

Performance metrics

Energy-Time tradeoff: Fittingness metric

With non-perfectly scalable models, if we want an application to run faster we will increase the number of PEs and, consequently, the execution cost (i.e energy)

- Energy-Delay Product (EDP) → $EDP = \frac{Speedup}{Efficiency}$

- **Fittingness metric** (FN): new metric that allows to have control over the Energy-Time tradeoff

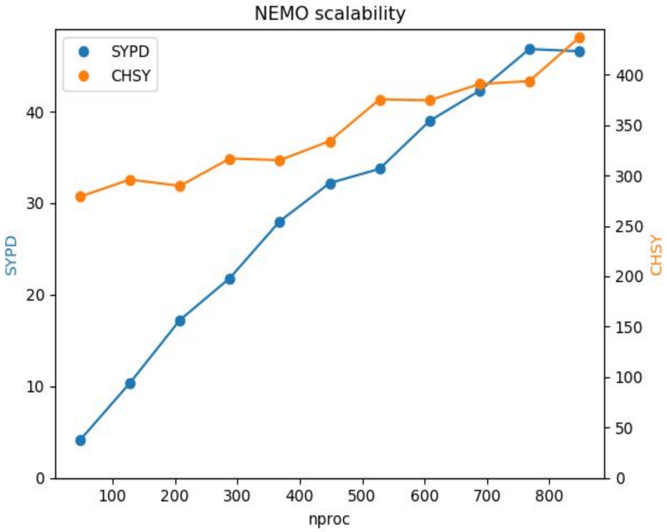
$$TTS_r + ETS_r = 1$$

$$FN = TTS_r\ SYPD_n + ETS_r\,(1 - CHSY_n)$$

Performance metrics

Energy-Time tradeoff

nproc	SYPD
48	4.1
128	10.4
208	16.2
288	22.8
368	27
432	32.2
528	35.8
608	39
688	43.3
768	46.8
848	46.6



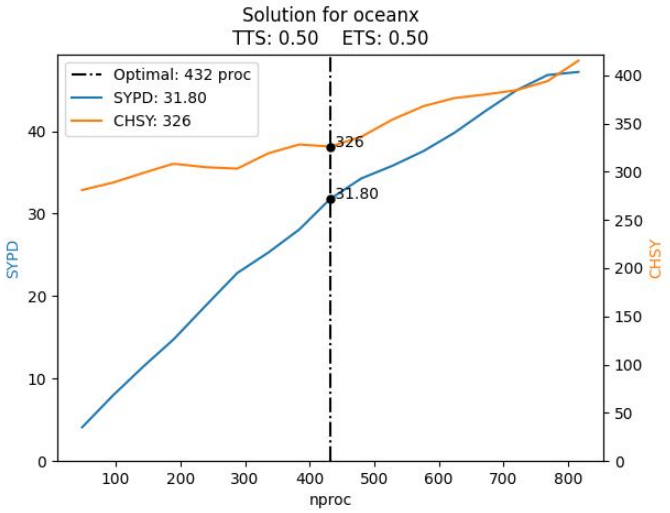
Performance metrics

Energy-Time tradeoff

TTSr = 0.5

ETSr = 0.5

nproc	SYPD	FN
48	4.1	0.50
128	10.4	0.53
208	16.2	0.55
288	22.8	0.65
368	27	0.62
432	32.2	0.68
528	35.8	0.64
608	39	0.61
688	43.3	0.64
768	46.8	0.64
848	46.6	0.50

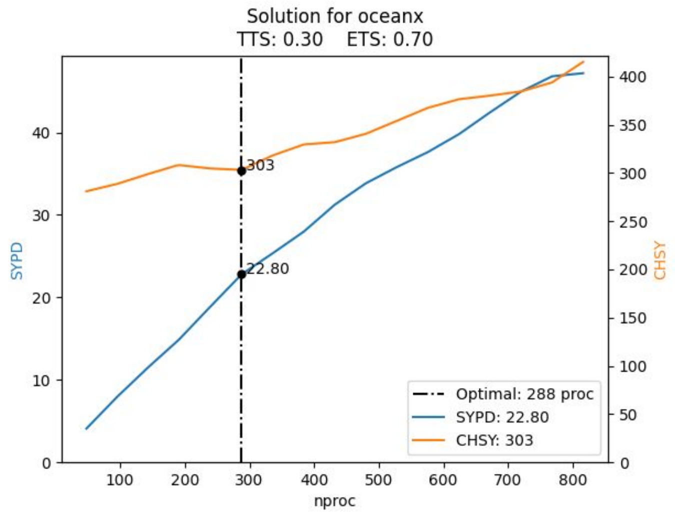


Performance metrics

Energy-Time tradeoff

TTSr = 0.3
ETSr = 0.7

nproc	SYPD	FN
48	4.1	0.70
128	10.4	0.68
208	16.2	0.66
288	22.8	0.73
368	27	0.65
448	32.2	0.66
528	35.8	0.59
608	39	0.53
688	43.3	0.52
768	46.8	0.49
848	46.6	0.30



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

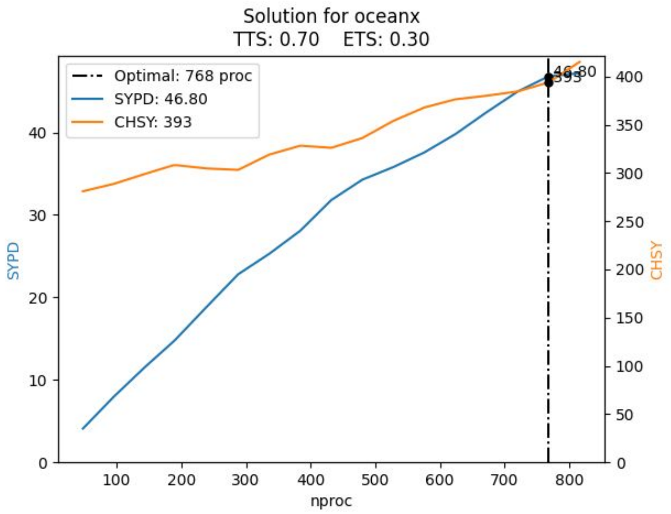
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

Performance metrics

Energy-Time tradeoff

TTSr = 0.7
ETSr = 0.3

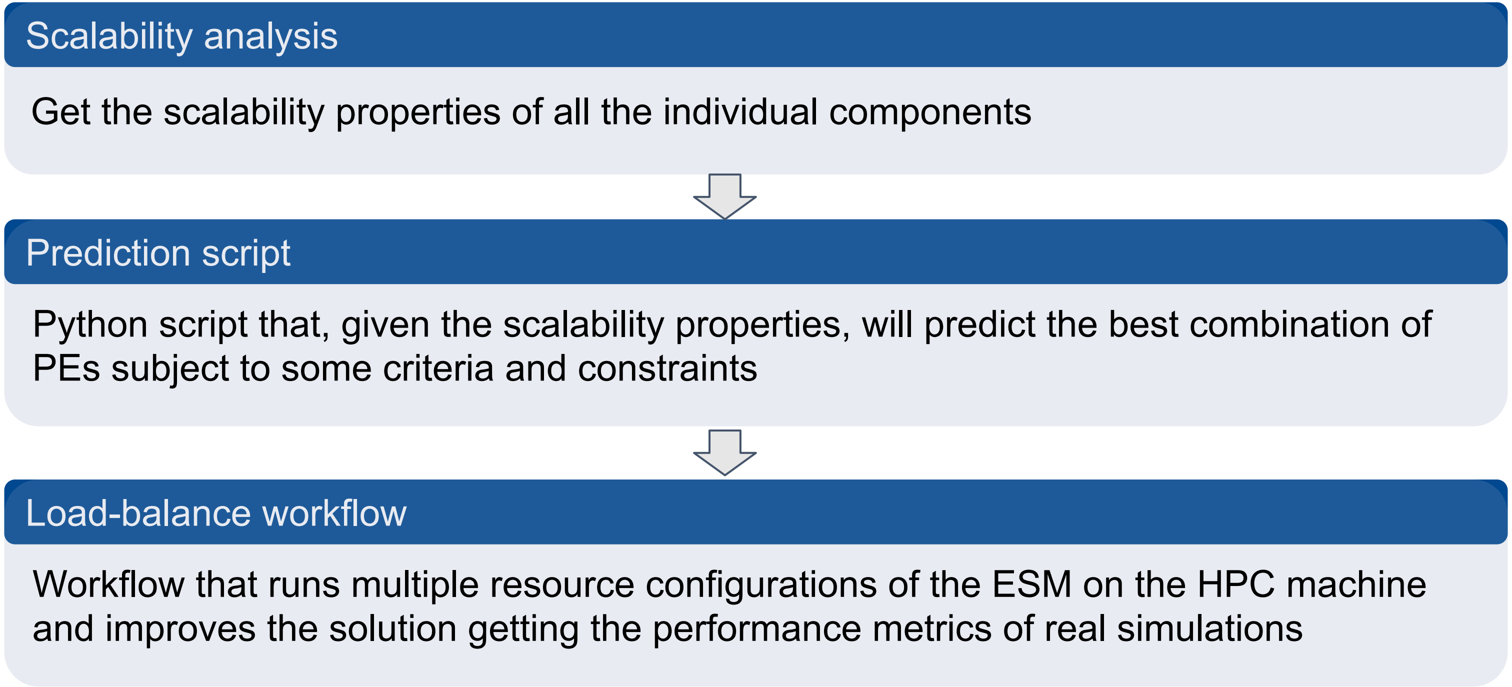
nproc	SYPD	FN
48	4.1	0.30
128	10.4	0.38
208	16.2	0.45
288	22.8	0.56
368	27	0.59
448	32.2	0.66
528	35.8	0.68
608	39	0.69
688	43.3	0.75
768	46.8	0.78
848	46.6	0.70



Automatic load-balance method

Auto-lb

Approach



Scalability analysis

Get the scalability properties of all the individual components

Prediction script

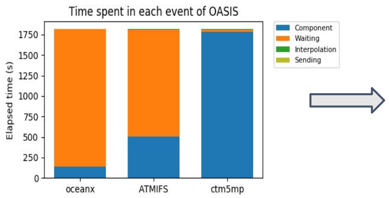
Python script that, given the scalability properties, will predict the best combination of PEs subject to some criteria and constraints

Load-balance workflow

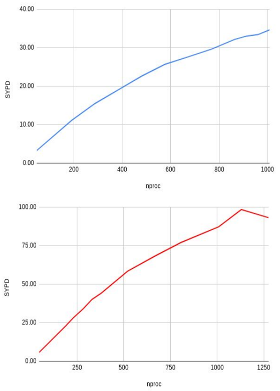
Workflow that runs multiple resource configurations of the ESM on the HPC machine and improves the solution getting the performance metrics of real simulations

Auto-lb

An experiment with a poor load-balance



Get component's scalability properties



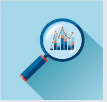
Load-balancing method

Prediction script

Automatic iterative process

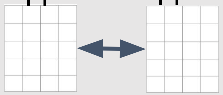
1. Run the ESM on the HPC

2. Get performance metrics

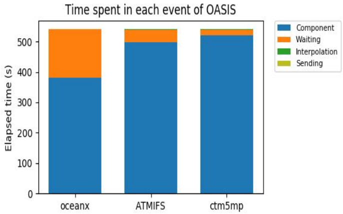


3. Modify the processors allocation

App. 1 App. 2



A balanced experiment



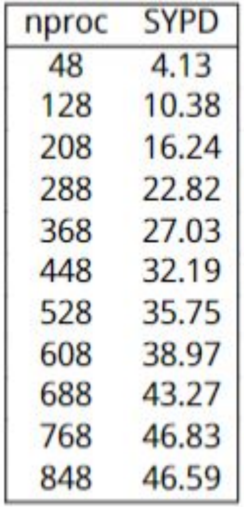
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

Auto-lb

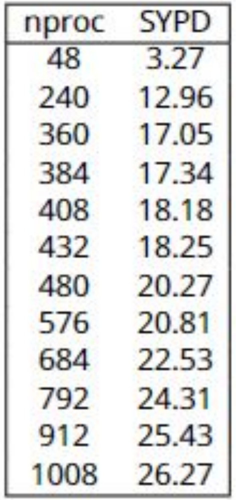
Scalability analysis

NEMO scalability



nproc	SYPD
48	4.13
128	10.38
208	16.24
288	22.82
368	27.03
448	32.19
528	35.75
608	38.97
688	43.27
768	46.83
848	46.59

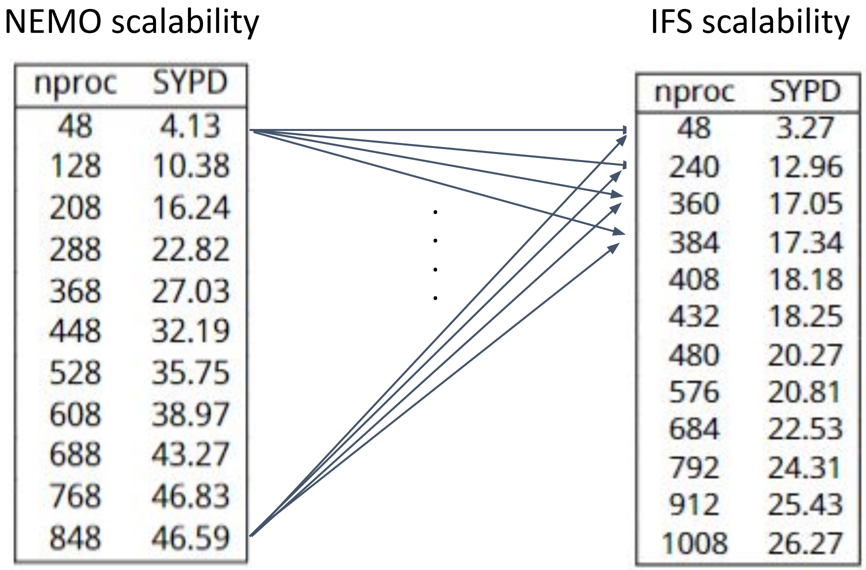
IFS scalability



nproc	SYPD
48	3.27
240	12.96
360	17.05
384	17.34
408	18.18
432	18.25
480	20.27
576	20.81
684	22.53
792	24.31
912	25.43
1008	26.27

Auto-lb

Prediction script



NEMO scalability

nproc	SYPD
48	4.13
128	10.38
208	16.24
288	22.82
368	27.03
448	32.19
528	35.75
608	38.97
688	43.27
768	46.83
848	46.59

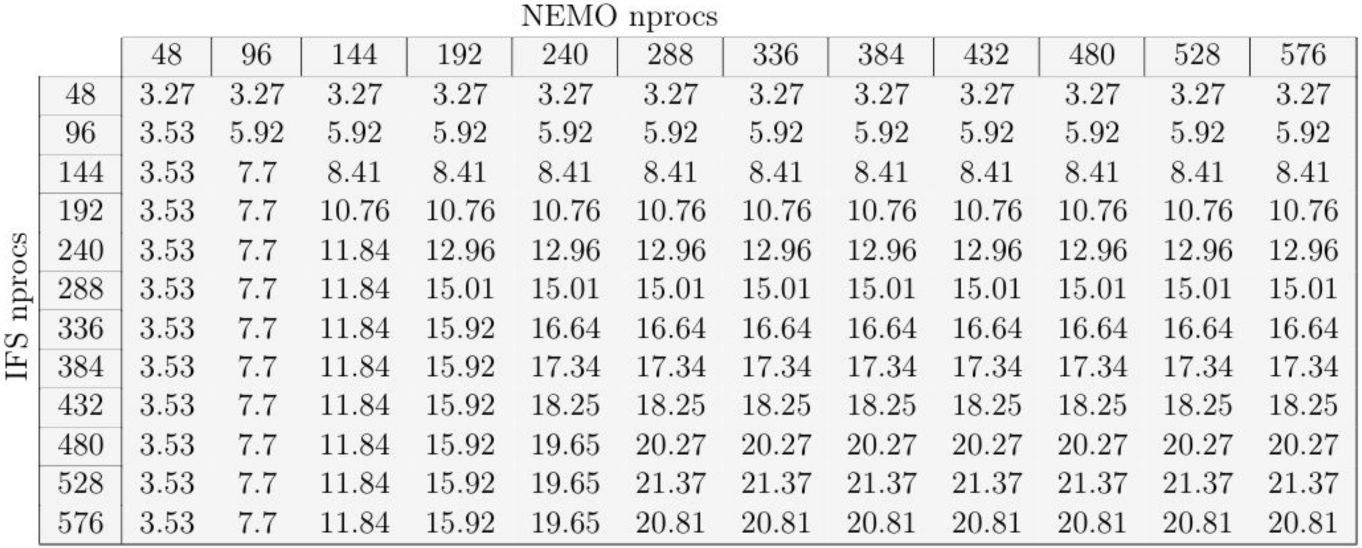
IFS scalability

nproc	SYPD
48	3.27
240	12.96
360	17.05
384	17.34
408	18.18
432	18.25
480	20.27
576	20.81
684	22.53
792	24.31
912	25.43
1008	26.27

Auto-lb

Prediction script

We assume that the **coupled execution** speed is expected to be **as fast as the slowest component**



NEMO nprocs

IFS nprocs	48	96	144	192	240	288	336	384	432	480	528	576
48	3.27	3.27	3.27	3.27	3.27	3.27	3.27	3.27	3.27	3.27	3.27	3.27
96	3.53	5.92	5.92	5.92	5.92	5.92	5.92	5.92	5.92	5.92	5.92	5.92
144	3.53	7.7	8.41	8.41	8.41	8.41	8.41	8.41	8.41	8.41	8.41	8.41
192	3.53	7.7	10.76	10.76	10.76	10.76	10.76	10.76	10.76	10.76	10.76	10.76
240	3.53	7.7	11.84	12.96	12.96	12.96	12.96	12.96	12.96	12.96	12.96	12.96
288	3.53	7.7	11.84	15.01	15.01	15.01	15.01	15.01	15.01	15.01	15.01	15.01
336	3.53	7.7	11.84	15.92	16.64	16.64	16.64	16.64	16.64	16.64	16.64	16.64
384	3.53	7.7	11.84	15.92	17.34	17.34	17.34	17.34	17.34	17.34	17.34	17.34
432	3.53	7.7	11.84	15.92	18.25	18.25	18.25	18.25	18.25	18.25	18.25	18.25
480	3.53	7.7	11.84	15.92	19.65	20.27	20.27	20.27	20.27	20.27	20.27	20.27
528	3.53	7.7	11.84	15.92	19.65	21.37	21.37	21.37	21.37	21.37	21.37	21.37
576	3.53	7.7	11.84	15.92	19.65	20.81	20.81	20.81	20.81	20.81	20.81	20.81

Auto-lb

Prediction script

The coupled execution cost is computed using Equation (1):

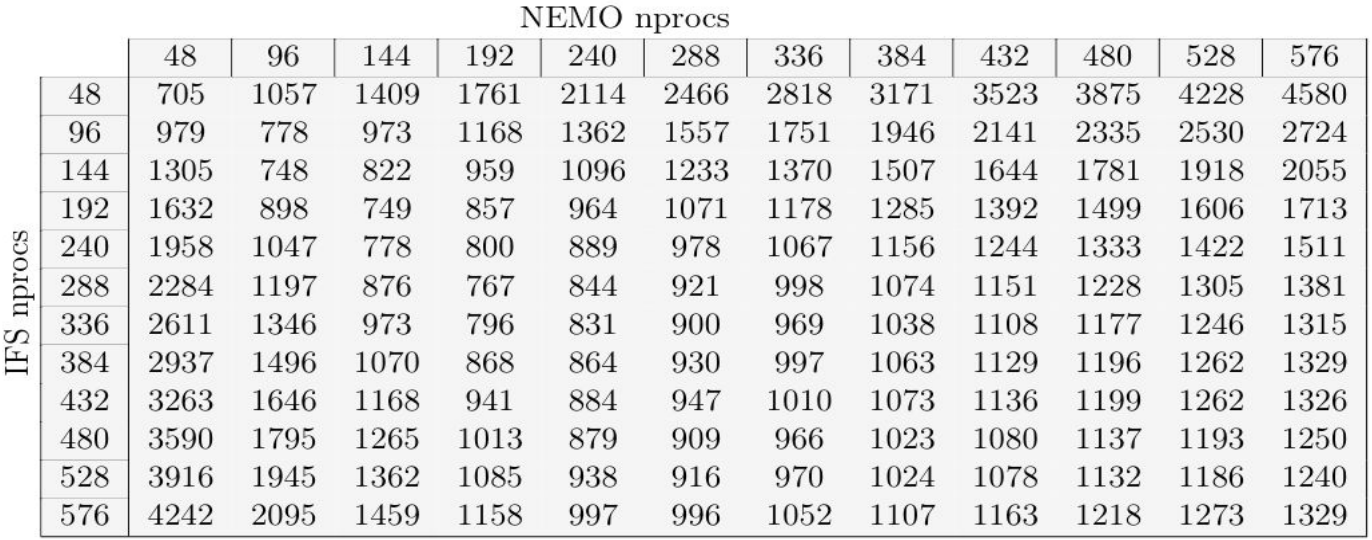
$$CHSY = \frac{24 \cdot NP}{SYPD}$$

Auto-lb

Prediction script

The coupled execution cost is computed using Equation (1):

$$CHSY = \frac{24 \cdot NP}{SYPD}$$



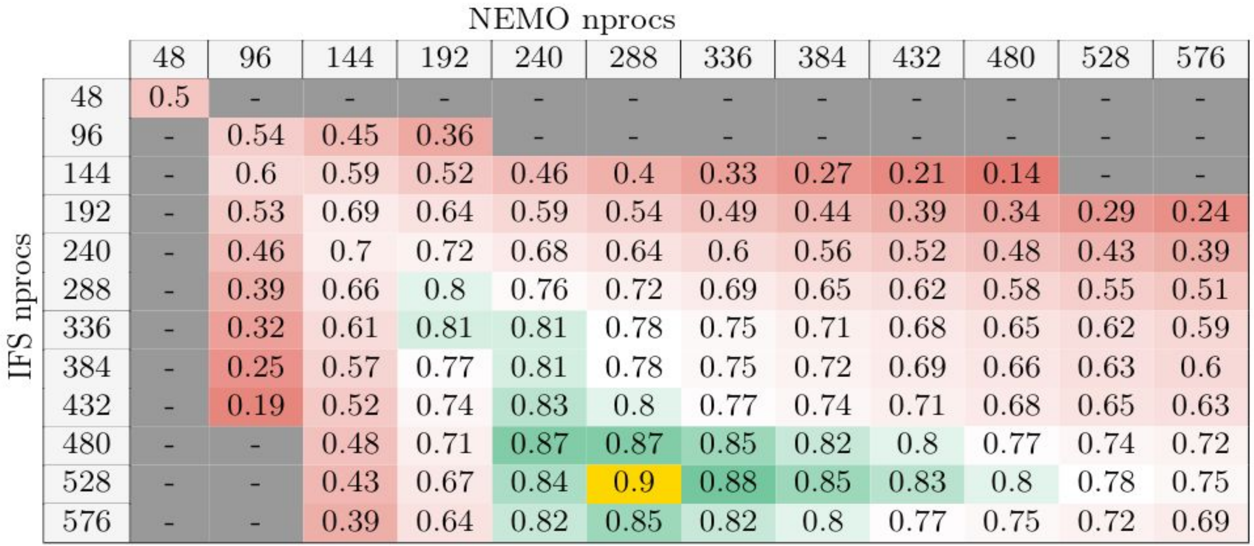
	NEMO nprocs											
IFS nprocs	48	96	144	192	240	288	336	384	432	480	528	576
48	705	1057	1409	1761	2114	2466	2818	3171	3523	3875	4228	4580
96	979	778	973	1168	1362	1557	1751	1946	2141	2335	2530	2724
144	1305	748	822	959	1096	1233	1370	1507	1644	1781	1918	2055
192	1632	898	749	857	964	1071	1178	1285	1392	1499	1606	1713
240	1958	1047	778	800	889	978	1067	1156	1244	1333	1422	1511
288	2284	1197	876	767	844	921	998	1074	1151	1228	1305	1381
336	2611	1346	973	796	831	900	969	1038	1108	1177	1246	1315
384	2937	1496	1070	868	864	930	997	1063	1129	1196	1262	1329
432	3263	1646	1168	941	884	947	1010	1073	1136	1199	1262	1326
480	3590	1795	1265	1013	879	909	966	1023	1080	1137	1193	1250
528	3916	1945	1362	1085	938	916	970	1024	1078	1132	1186	1240
576	4242	2095	1459	1158	997	996	1052	1107	1163	1218	1273	1329

Auto-lb

Prediction script

Finally, compute the Fittingness using Equation (3): $FN = TTS_r\ SYPD_n + ETS_r\,(1 - CHSY_n)$

$$TTSr = ETSr = 0.5$$



NEMO nprocs

IFS nprocs	48	96	144	192	240	288	336	384	432	480	528	576
48	0.5	-	-	-	-	-	-	-	-	-	-	-
96	-	0.54	0.45	0.36	-	-	-	-	-	-	-	-
144	-	0.6	0.59	0.52	0.46	0.4	0.33	0.27	0.21	0.14	-	-
192	-	0.53	0.69	0.64	0.59	0.54	0.49	0.44	0.39	0.34	0.29	0.24
240	-	0.46	0.7	0.72	0.68	0.64	0.6	0.56	0.52	0.48	0.43	0.39
288	-	0.39	0.66	0.8	0.76	0.72	0.69	0.65	0.62	0.58	0.55	0.51
336	-	0.32	0.61	0.81	0.81	0.78	0.75	0.71	0.68	0.65	0.62	0.59
384	-	0.25	0.57	0.77	0.81	0.78	0.75	0.72	0.69	0.66	0.63	0.6
432	-	0.19	0.52	0.74	0.83	0.8	0.77	0.74	0.71	0.68	0.65	0.63
480	-	-	0.48	0.71	0.87	0.87	0.85	0.82	0.8	0.77	0.74	0.72
528	-	-	0.43	0.67	0.84	0.9	0.88	0.85	0.83	0.8	0.78	0.75
576	-	-	0.39	0.64	0.82	0.85	0.82	0.8	0.77	0.75	0.72	0.69

Auto-lb

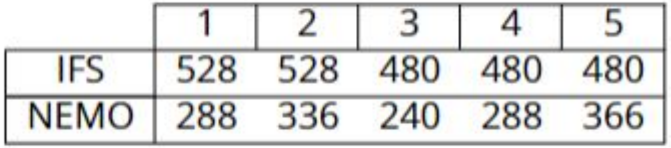
Prediction script

Optimal result is predicted based only on the scalability properties:

- Real coupling interactions are not taken into account
- Variability expected when running on the HPC platform

Instead of boldly selecting the setup which maximizes the FN, the top 5 solutions are taken as potential optimal ones

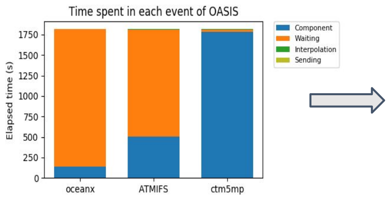
top5 →



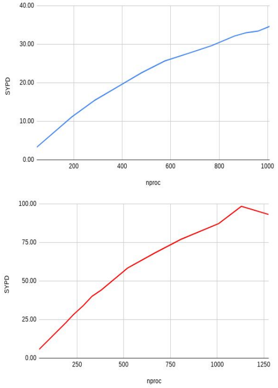
	1	2	3	4	5
IFS	528	528	480	480	480
NEMO	288	336	240	288	366

Auto-lb

An experiment with a poor load-balance



Get component's scalability properties



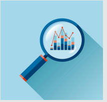
Load-balancing method

Prediction script

Automatic iterative process

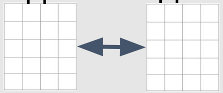
1. Run the ESM on the HPC

2. Get performance metrics

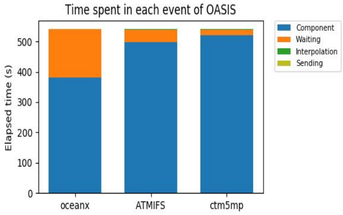


3. Modify the processors allocation

App. 1 App. 2



A balanced experiment



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

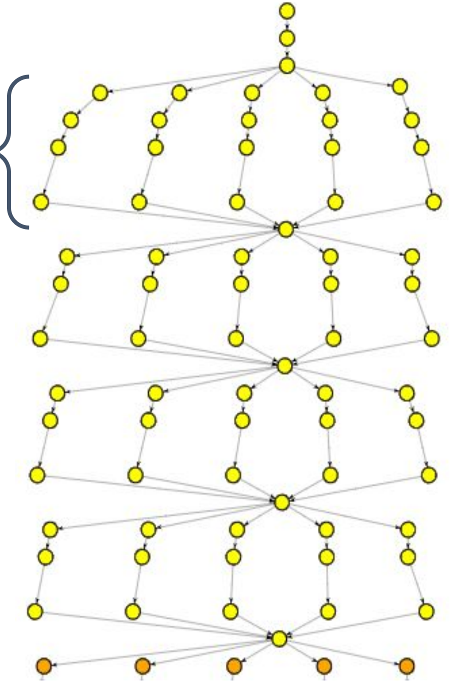
is-enes
INFRASTRUCTURE FOR THE EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

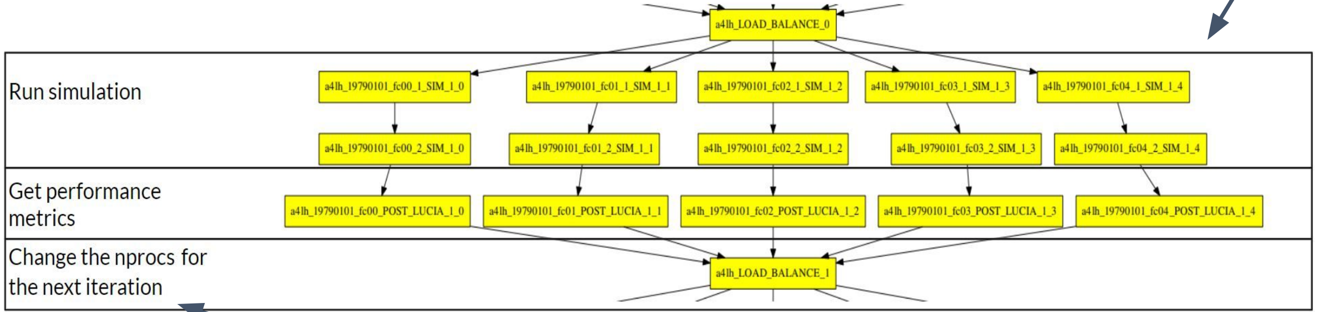
Auto-lb

Load-balance workflow

Automatic iterative process:

1. Run the top5 setups
2. Get the performance metrics (Lucia)
3. Modify the setup to minimize the coupling cost




a4lh_LOAD_BALANCE_0
a4lh_19790101_fc00_1_SIM_1_0
a4lh_19790101_fc01_1_SIM_1_1
a4lh_19790101_fc02_1_SIM_1_2
a4lh_19790101_fc03_1_SIM_1_3
a4lh_19790101_fc04_1_SIM_1_4
a4lh_19790101_fc00_2_SIM_1_0
a4lh_19790101_fc01_2_SIM_1_1
a4lh_19790101_fc02_2_SIM_1_2
a4lh_19790101_fc03_2_SIM_1_3
a4lh_19790101_fc04_2_SIM_1_4
a4lh_19790101_fc00_POST_LUCIA_1_0
a4lh_19790101_fc01_POST_LUCIA_1_1
a4lh_19790101_fc02_POST_LUCIA_1_2
a4lh_19790101_fc03_POST_LUCIA_1_3
a4lh_19790101_fc04_POST_LUCIA_1_4
a4lh_LOAD_BALANCE_1

Run simulation

Get performance metrics

Change the nprocs for the next iteration

Component_coupling_cost

Results

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

Results

1. CMIP6 Standard Resolution experiment in MN4

   Finding a better parallelization and overall load-balance

2. SR experiment in ECMWF HPCF

   Using the Fittingness metric to find different TTS and ETS setups to the same experiment

Results

SR CMIP6

Original	Optimal
SYPD: 15	SYPD: 16.8
CHSY: 1135	CHSY: 1074
Cpl_c: 13.8%	Cpl_c: 13.6%
PEs: 624	PEs: 672

TTSr = ETSr = 0.5

The original setup used less resources than the optimal point

The optimal configuration is 12% faster and 5% less costly than the original

This experiment was used to simulate 14000 years and consumed 15M core-hours in MN4

Results

SR CMIP6

Original	Optimal	Same SYPD
SYPD: 15	SYPD: 16.8	SYPD: 16
CHSY: 1135	CHSY: 1074	CHSY: 1100
Cpl_c: 13.8%	Cpl_c: 13.6%	Cpl_c: 17.4%
PEs: 624	PEs: 672	PEs: 672

TTSr = ETSr = 0.5

Using the right parallelization but following the same SYPD strategy leads to a suboptimal setup

Compared to the Original setup, the coupling cost is higher but it is 6% faster and 3% less costly

Results

SR in ECMWF machine

Results

SR in ECMWF machine

Original	Optimal
SYPD: 11.2	SYPD: 17.6
CHSY: 928	CHSY: 1230
Cpl_c: 15.8%	Cpl_c: 11.2%
PEs: 432	PEs: 900

The optimal setup is better balanced, 56% faster but 32% more costly

This is a more **TTS oriented setup**

Results

SR in ECMWF machine, ETS solution

Original	Optimal TTS	Optimal ETS
SYPD: 11.2	SYPD: 17.6	SYPD: 13.9
CHSY: 928	CHSY: 1230	CHSY: 939
Cpl_c: 15.8%	Cpl_c: 11.2%	Cpl_c: 8.3%
PEs: 432	PEs: 900	PEs: 540
	TTSr = ETSr = 0.5	TTSr = 0.2, ETSr = 0.8

Setting a TTSr = 0.2 (ETSr = 0.8) and rerunning the auto-lb workflow, we get a more **ETS oriented setup**

The new configuration is **24% faster and has the same execution cost** as the original

Conclusions

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

Conclusions

- The load-balance is one key limiting factors of coupled ESMs performance
- Current approaches to find the best resource allocation can not find the optimal setup
- Manually finding the optimal setup is a repetitive, tedious and prone to error task
- The auto-lb method has achieved better setups for experiments used in big projects such as CMIP6, for multiple resolutions and on HPC platforms

Future work

- ESMs will keep growing in complexity and in the number of components they include
- Without the proper tools and metrics, traditional approaches would not even achieve suboptimal solutions
- The auto-lb method has proven to work with two components but it will be extended to handle more complex simulations

Thank you!

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

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

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Automatic approach to balance coupled ESMs:

Saving energy and time on coupled ESMs

Sergi Palomas, Dr. Mario Acosta, Dr. Etienne Tourigny

January, 2023

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