D3.5 Full HPCW suite v1.0

Dissemination Level: Public

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 800987

Funded by the European Union

Co-ordinated by ECMWF

ESCAPE2

Energy-efficient Scalable Algorithms for Weather and Climate Prediction at Exascale

Author **Ralf Mueller (DKRZ), David Guibert (BULL), Erwan Raffin (BULL), Mario Acosta (BSC), Daniel Beltran (BSC)**

Date **11/10/2021**

Research and Innovation Action

H2020-FETHPC-2017

Project Coordinator: Dr. Peter Bauer (ECMWF)

Project Start Date: 01/10/2018

Project Duration: 36 months

Published by the ESCAPE-2 Consortium

Version: 1.0

Contractual Delivery Date: 30/09/2021

Work Package/ Task: WP3 / T3.5

Document Owner: DKRZ

Contributors: DKRZ, Bull

Status: Final

Table of Contents

Figures

1 Executive Summary

One of the main objectives of ESCAPE-2 is to establish weather and climate representative benchmarks based on world class European prediction models to enable deployment on energy efficient and heterogeneous HPC architectures towards Extreme-scale Demonstrators.

Here we present the version 1.0 of the HPCW benchmark which tends to be the reference benchmark suite for the Weather and Climate community in Europe. HPCW components, main characteristics and deployment are given. We also present the reference baseline results, time and energy to solution, obtained at Atos. Finally, the future of HPCW is outlined.

2 Introduction

2.1 Background

ESCAPE-2 will develop world-class, extreme-scale computing capabilities for European operational numerical weather and climate prediction systems. It continues the pioneering work of the ESCAPE project. The project aims to attack all three sources of enhanced computational performance at once, namely (i) developing and testing bespoke numerical methods that optimally trade off accuracy, resilience and performance, (ii) developing generic programming approaches that ensure code portability and performance portability, (iii) testing performance on HPC platforms offering different processor technologies.

ESCAPE-2 will prepare weather and climate domain benchmarks that will allow a much more realistic assessment of application specific performance on large HPC systems than current generic benchmarks such as HPL[1] and HPCG. These benchmarks are specifically geared towards the pre-exascale and exascale HPC[2] infrastructures that the European Commission and Member States will invest in through the European Commission Joint Undertaking.

ESCAPE-2 also combines generic uncertainty quantification tools for high-performance computing originating from the energy sector with ensemble-based weather and climate models to quantify the effect of model and data related uncertainties on forecasting – a capability, which weather and climate prediction has pioneered since the 1960s. This collaboration combines user-friendly tools from one community with scientific expert knowledge from another community to achieve economy of scales beyond the scope of each domain.

2.2 Scope of this deliverable

2.2.1 Objectives of this deliverable

The objective of this deliverable is to report the work done in WP3, especially the Task 3.5 “The HPCW Benchmark Suite” and its v1.0 version developed and deployed during the project.

[1] HPL - High Performance LINPACK - https://www.top500.org/project/linpack/

[2] High Performance Conjugate Gradient Benchmark - http://www.hpcg-benchmark.org/

2.2.2 Work performed in this deliverable

The final version of the HPCW benchmark v1.0 collects code from most of Europe's major modelling centers and lead scientific institutions. It is based on the open standard tool CMake to orchestrate building and testing the different benchmark components as well as its individual dependencies.

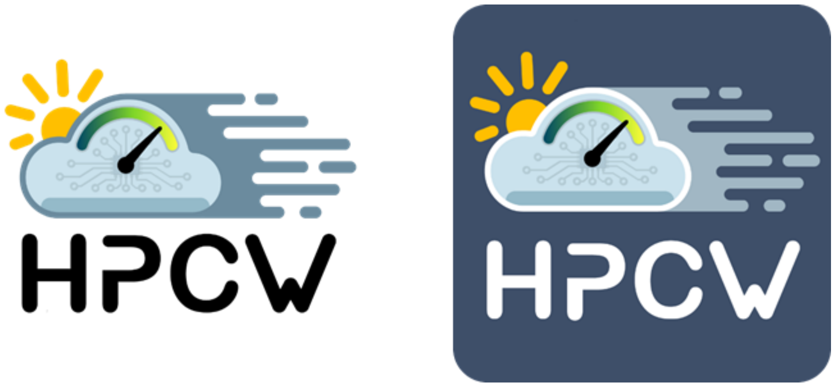


Figure 1 - HPCW logo

WP3 managed to design and develop a highly adaptable framework, that has proven to be portable and customizable to multiple HPC systems: it can integrate existing software packages such as compilers and 3rd party libraries and work with different queueing systems depending on the user's preference.

Because HPCW contains many different codes, each of its components can be built and run separately to allow focusing on specific aspects when it comes to profiling, performance measurement and optimization. Apart from performance HPCW also includes methods to estimate the energy-to-solution if the target HPC systems provides energy-related information.

HPCW components are full Numerical Weather Prediction (NWP) and climate models as well as so called dwarfs focusing on specific algorithm and a pure workload manager. Their tests cover a wide range of applications from low and medium to high resolution available for pure atmosphere, ocean as well as coupled setups. In contrast to models, the dwarfs represent specific compute-intense workloads usually extracted from models for further investigation on optimization. A good example is ECMWF's radiation library ecRad, which has a high impact on overall performance and is already used in many other models. The complete list of the HPCW components, associated tests and the dwarfs are provided in Appendix 5.1, 5.2, and 5.3 respectively.

All tests can be run in an automatic way including a validity check for the input fields. Many tests also come with basic validation of the generated output. This ensures that the tests run within a reasonable parameter space and further analysis makes sense.

An ensemble-based verification toolchain was developed for two of the models (IFS-FVM, ICON) and proved to be very useful for estimating changes that do not lead to bit-identical results. Benchmarkers can benefit from this a lot since critical performance optimization almost always include non-identical model results.

Portability and usability were of special interest during the HPCW development. That's why the benchmark was tested on HPC systems at Atos, DKRZ, ECMWF and BSC. These systems provide a wide range of software and hardware environments which are ideal to work on the targeted flexibility of the framework.

The time and energy to solution obtained at Atos are presented in Appendix 5.4.

HPCW has been used to profile IFS-FVM dwarf on BSC Marenostrum4 HPC system. The results and profile analysis are given in Appendix 5.5.

HPCW also supports multiple hardware platforms. As a first step ICON can be built and run on NVIDIA GPUs. IFS RAPS already support GPU and its integration is ongoing. Others like IFS-FVM and ecRad will support GPUs and further improve the portability of HPCW. As second approach towards portability ESCAPE-2 has put forward WP2 to develop a DSL toolchain prototype. This prototype has been successfully applied to the current ICON release 2.6.3 and was added to the HPCW benchmark in collaboration with WP2 partners.

HPCW v1.0 benchmark software, associated input data, and documentation are accessible via ECMWF portal. The licenses for each component are given in Appendix 5.6.

2.2.3 Deviations and counter measures

WP3 had many links to other work packages because HPCW, its main deliverable incorporates results from them. These strong links come with an increased risk for completing all tasks to 100%. Hence some of the initial goals could not be met.

During the design phase of the project, it seemed applicable to incorporate the work from WP1 and 2, i.e. a semi-implicit, semi-Lagrangian Discontinuous Galerkin model and a DSL-toolchain prototype. Although WP1 achieved remarkable results, its development did not stabilize in time to integrate them into the benchmark and stick to HPCW roadmap at the same time. Hence, we decided to drop the dependency to WP in favor of including other interesting and relevant setups.

The second dependency was the opportunity to incorporate the DLS toolchain developed by WP2. Its development was delayed, but the more severe problem was, that it was designed as a prototype. WP2's only reasonable approach was focusing on the front- and backend and taking a fixed ICON release to incorporate both into a standalone testcase. HPCW instead is meant to provide representative workloads from climate and weather models with a long-term support strategy. This includes regular model updates by scientific partners in the future. In contrast to that the prototype developed by WP2 was never meant to provide a stable interface usable by future versions of ICON. ICON itself is a moving target and influencing the ICON development at that scale is beyond the scope of ESCAPE2. Hence long-term strategy for HPCW and the focus of WP2 to build a protype pose a problem, that WP3 had to cope with.

Our solution was to incorporate the required code from the DSL toolchain together with the ICON source code version used to integrate the toolchain into a special branch of the HPCW repository. This way we can make the DSL available to the customers, provide a full model setup for testing it and keep a stable development environment for the benchmark as a whole in the main development branch. In case the toolchain proves to be successful for future model development HPCW is ready to take it over for more of its components.

2.2.4 Exceeding the plans

As mentioned in the section above WP3 faced some challenges regarding its components for v1.0. But on the other hand, there are a couple of additions that could be added to the benchmark, which will be beneficial for its impact.

Despite the planed ICON tests for ocean and atmosphere, WP3 managed to include a first coupled experiment. Although this is technically a basic setup, it is a highly relevant for the future. ICON has undergone a severe change regarding the coupled model development: Max-Planck-Institute for Meteorology, having solely developed the coupled setup before now has joined forces with the German weather service DWD in this area. It can be assumed that more Weather Prediction Centres will focus not only on pure NWP, but also put seasonal climate projections into their portfolio in the future.

ECMWF is constantly working on modernizing they production model IFS, which was planned and successfully added to HPCW. The new model called IFS-FVM is under development, but WP3 managed to include this component in several tests.

Although it was not part of the initial project plan, HPCW includes an ICON-atmosphere test for NVIDIA GPUs. Developers of ICON and IFS-FVM are working on porting more parts of the model. Having this test is a solid base for future support for multiple hardware platform.

Radiation has proven to be a very compute-intense component of all climate and weather models. This is the reason why the radiation dwarf Acraneb and the ECMWFs radiation library ecRad are part of the HPCW benchmark.

Tracer transport is another model component, that demands a lot compute resource. To cover this HPCW comes with an ICON-ocean-advection dwarf. This component is under development by the ICON developers and will be ported to GPU in the near future. Having both tests in HPCW will let benchmarkers be able to compare CPU and GPU with the same model component in the future.

3 Future of HPCW

The future of HPCW benchmark heavily relies on collaboration with the Weather and Climate community through the ESiWACE2 project. It will take over the further development and maintenance of the HPCW benchmark. This task will be led by DKRZ, which was also co-lead in the WP3. For a smooth transition WP3 started to build up a Continuous Integration (CI) system at the DKRZ facilities. At the moment this includes building automatically almost all components of HPCW with several software environments. This CI already proved to be very useful for the development within WP3. It is planed that ESIWACE2 will add runtime checks for the tests and by that provide a full environment for a stable and productive development of the benchmark.

4 Conclusion

This document presents the HPCW benchmark v1.0 which contains the code from most of Europe’s major modelling centers and leading scientific institution. HPCW allows benchmarkers to build, run, validate and profile its component as well as providing performance and energy measurement. It is a highly adaptable framework, that has proven to be portable and customizable to multiple HPC systems as it has been ported to several systems at Atos, DKRZ, ECMWF and BSC already.

5 Appendix

5.1 HPCW components

5.1.1 Models

- ICON Ocean and Atmosphere
- ICON Atmosphere GPU (from DSL implementation)
- IFS (RAPS)
- NEMO

5.1.2 Dwarfs

- IFS atmosphere FV dwarf (IFS-FVM)
- Radiation dwarf (ACRANEB2)
- ICON ocean advection dwarf
- ecRad

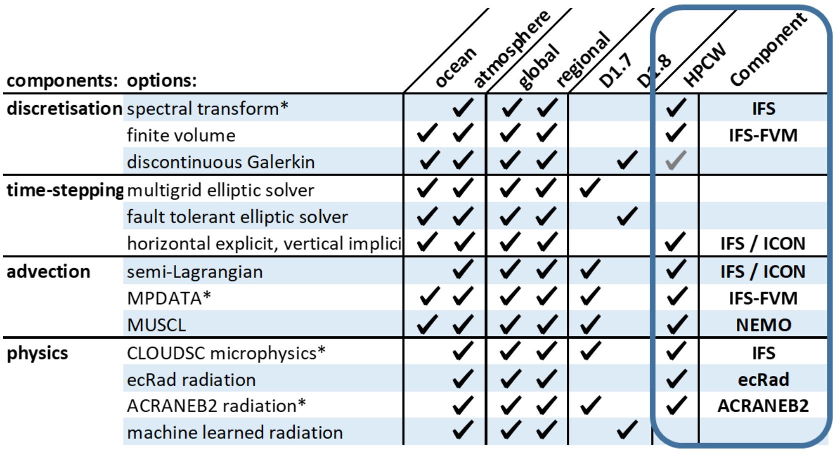
5.1.3 Workload Simulator

- Kronos

5.2 HPCW test configurations

- ICON Ocean: Small (160km), Medium (40km) and big (10km), 3-points on a strong scaling line
- ICON Atmosphere: Small (160km)
- ICON Atmosphere GPU: Small (160km)
- ICON coupled atmosphere and ocean (160km)
- ICON Ocean advection dwarf: Small
- IFS (RAPS): Small (TL159), Medium (TCo639), Big (TCo1999)
- NEMO: Small BENCH1 (ORCA1 like) Medium (ORCA0,25)
- IFS-FVM: Small (O160), Medium (O640), Big (O1280)
- ACRANEB2: Small
- Kronos: Single-serial, single-parallel, multi-serial-events, external-job
- ecRad: Small

5.3 HPCW dwarfs – from “Algorithms & Mathematics” (WP1) to “Weather and climate benchmarks: HPCW” (WP3)



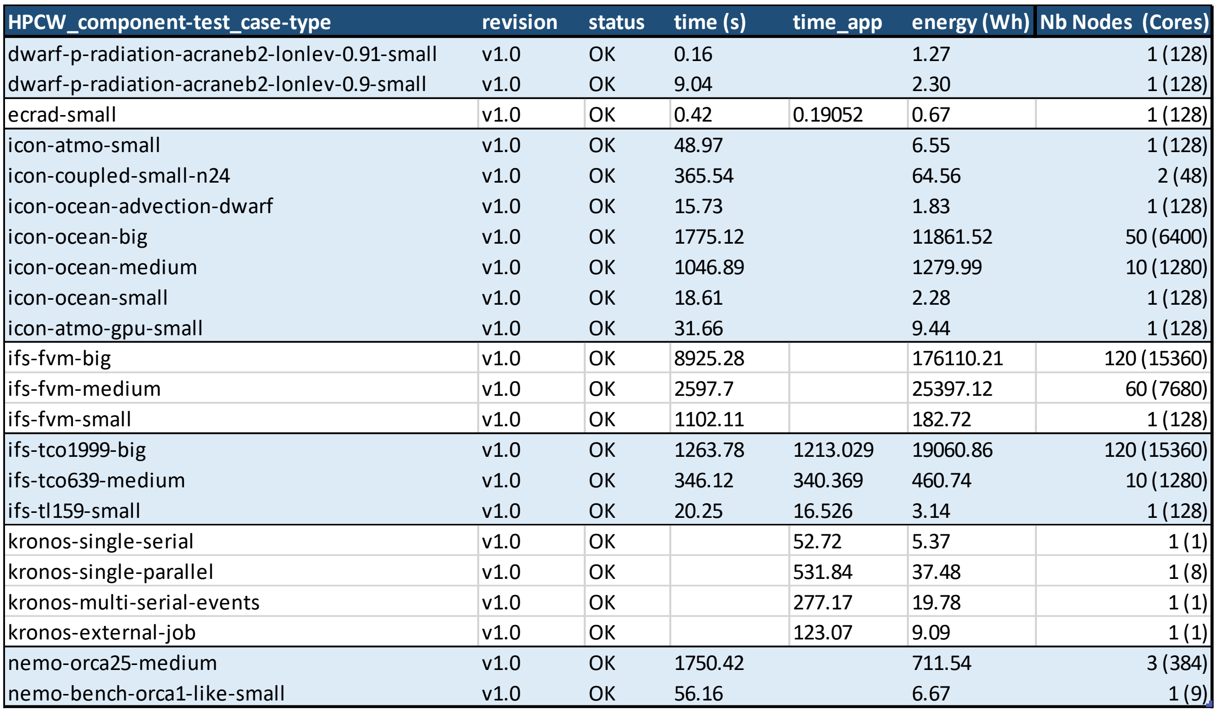
components:	options:	ocean	atmosphere	global	regional	D1.7	D1.8	HPCW	Component
discretisation	spectral transform*		✓	✓	✓			✓	IFS
	finite volume	✓	✓	✓	✓			✓	IFS-FVM
	discontinuous Galerkin	✓	✓	✓	✓		✓	✓	
time-stepping	multigrid elliptic solver	✓	✓	✓	✓	✓			
	fault tolerant elliptic solver	✓	✓	✓	✓		✓		
	horizontal explicit, vertical implici	✓	✓	✓	✓			✓	IFS / ICON
advection	semi-Lagrangian		✓	✓	✓	✓		✓	IFS / ICON
	MPDATA*	✓	✓	✓	✓	✓		✓	IFS-FVM
	MUSCL	✓	✓	✓	✓	✓		✓	NEMO
physics	CLOUDSC microphysics*		✓	✓	✓	✓		✓	IFS
	ecRad radiation		✓	✓	✓			✓	ecRad
	ACRANEB2 radiation*		✓	✓	✓	✓		✓	ACRANEB2
	machine learned radiation		✓	✓	✓		✓		

Figure 2 - HPCW dwarfs and how they are integrated into HPCW (in grey: the work in progress, with star: from ESCAPE 1)

5.4 Time and energy to solution of HPCW v1.0 at Atos

Here are, in Figure 3, the results obtained at Atos using the following setup. Note that the job runtime is noted “time” and application runtime extracted from the application log (if present) is noted “app_time” and is expressed here in seconds. For Kronos, the reported time are only “app_time” because Kronos launch its own jobs for which the time to solution is the most relevant. The energy to solution is measured for the job at node level.

- Hardware setup
  - BullSequana XH2000 (X2410/5 AMD blades)
  - CPU: 2x AMD EPYC 7763 64-Core Processor (MILAN)
  - + GPU: 4x Nvidia A100 GPU when needed
  - Memory: 256GB RAM DDR4 3200 MT/s
  - Interconnect: HDR 100 (fat tree)
- Software setup
  - OS: RHEL 8.3
  - Environments:
    - bull-intel.env.sh
    - bull-intel+mpi+mkl.env.sh
    - bull-intel+openmpi+mkl.env.sh
    - bull-intel19+openmpi+mkl.env.sh
    - bull-spack-icon-nvhpc.env.sh
  - Toolchain:
    - bull-intel.cmake
    - bull-nvhpc.cmake
  - Job launcher
    - bull-job-launcher.sbatch
    - Energy measurement: Bull Energy Optimizer (BEO)



HPCW_component-test_case-type	revision	status	time (s)	time_app	energy (Wh)	Nb Nodes (Cores)
dwarf-p-radiation-acraneb2-lonlev-0.91-small	v1.0	OK	0.16		1.27	1 (128)
dwarf-p-radiation-acraneb2-lonlev-0.9-small	v1.0	OK	9.04		2.30	1 (128)
ecrad-small	v1.0	OK	0.42	0.19052	0.67	1 (128)
icon-atmo-small	v1.0	OK	48.97		6.55	1 (128)
icon-coupled-small-n24	v1.0	OK	365.54		64.56	2 (48)
icon-ocean-advection-dwarf	v1.0	OK	15.73		1.83	1 (128)
icon-ocean-big	v1.0	OK	1775.12		11861.52	50 (6400)
icon-ocean-medium	v1.0	OK	1046.89		1279.99	10 (1280)
icon-ocean-small	v1.0	OK	18.61		2.28	1 (128)
icon-atmo-gpu-small	v1.0	OK	31.66		9.44	1 (128)
ifs-fvm-big	v1.0	OK	8925.28		176110.21	120 (15360)
ifs-fvm-medium	v1.0	OK	2597.7		25397.12	60 (7680)
ifs-fvm-small	v1.0	OK	1102.11		182.72	1 (128)
ifs-tco1999-big	v1.0	OK	1263.78	1213.029	19060.86	120 (15360)
ifs-tco639-medium	v1.0	OK	346.12	340.369	460.74	10 (1280)
ifs-tl159-small	v1.0	OK	20.25	16.526	3.14	1 (128)
kronos-single-serial	v1.0	OK		52.72	5.37	1 (1)
kronos-single-parallel	v1.0	OK		531.84	37.48	1 (8)
kronos-multi-serial-events	v1.0	OK		277.17	19.78	1 (1)
kronos-external-job	v1.0	OK		123.07	9.09	1 (1)
nemo-orca25-medium	v1.0	OK	1750.42		711.54	3 (384)
nemo-bench-orca1-like-small	v1.0	OK	56.16		6.67	1 (9)

Figure 3 - Time to solution and energy to solution of HPCW v1.0 at Atos

5.5 Deployment at BSC and IFS-FVM profile analysis results

Here are the results obtained at BSC-Marenostrum4 for the IFS-FVM profile analysis compiled with HPCW and using the following setup:

- Hardware setup
  - CPU: 2 sockets Intel Xeon Platinum 8160
  - Total: 3456 nodes and 165888 cores
  - Memory: 12x 8GB RAM DIMM 2667 MT/s
  - Interconnect: 100 Gbit/s Intel Omni-Path
- Software setup
  - OS: Linux 4.4.120-92.70-default
  - Details:
    - Branch -> db/bsc-config
    - Intel version -> intel/2018.4 (Kronos: 2018.1)
    - IFS-FVM: Profiled with only MPI
  - Environments:
    - bsc-intel_marenostrum4.env.sh
    - bsc-intel_marenostrum4+kronos.env.sh
  - Toolchain:
    - bsc-intel_marenostrum4.cmake
  - Modifications:
    - kronos-slurm.bsc.py
    - Icon wrapper: marenostrum4.intel-18.0.4
  - Job launcher
    - bsc-job-launcher.sbatch
- BSC profiling tools:

  - **EXTRAE**: A package to instrument the code of the model automatically or/and manually. It generates trace files with HW counters, MPI messages and other info through its API.
  - **PARAVER**: It is a powerful trace browser that understand the traces generated by EXTRAE.
  - **DIMEMAS**: It is a performance analysis tool for message-passing programs that allows to simulate a trace in an optimal environment.
  - **MODELFACTORS**: It is a tool to compare different traces of the same model and obtain a set of scalability factors.

- Traces Configuration:
  - MPI + PAPI counters + Call stacks.
  - 3 different traces: 1node*48tasks, 2nodes*48tasks and 4nodes*48tasks.
- IFS-FVM results:
  - Good computation scalability: The main issue of IPC decreasing shown at figure 4 seems to be communication related and we didn't notice that computing-related factors are being an issue for the model performance.
  - Load balance: It is a measure how globally balanced work is between all process and as shown in figure 5, we see that threads are having different amount of work and starting/ending at different time.
  - MPI overhead may affect IPC as shown on figure 6. Only with 4 nodes we see an decrement on the transfer efficiency and load balance.

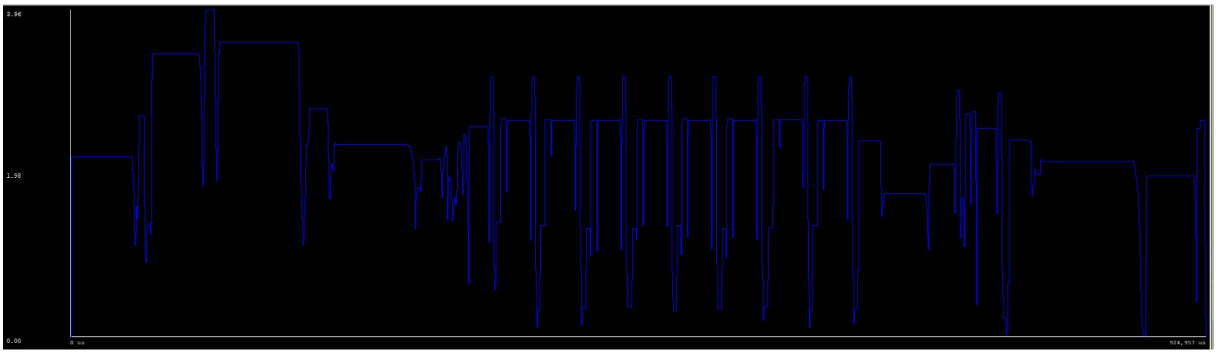


Figure 4 - Useful IPC

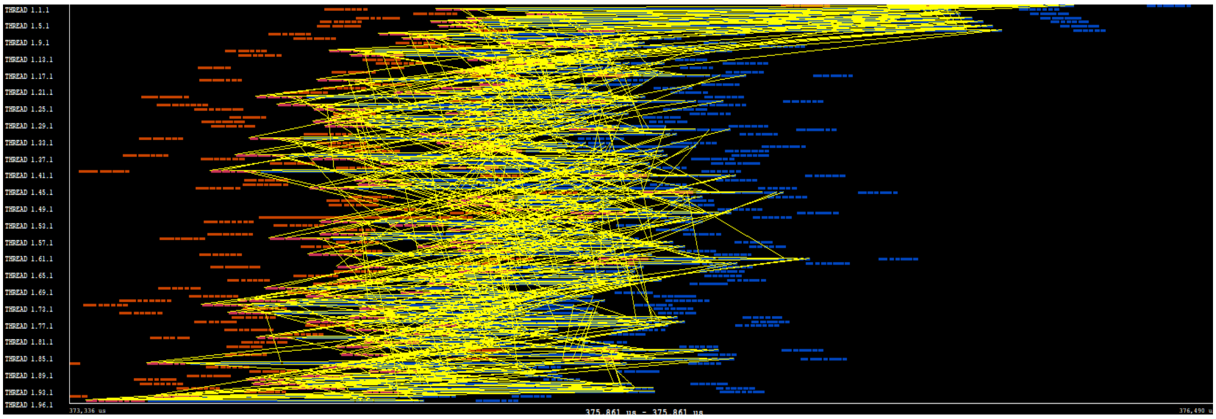


Figure 5 - Threads communication in one step of the model iteration

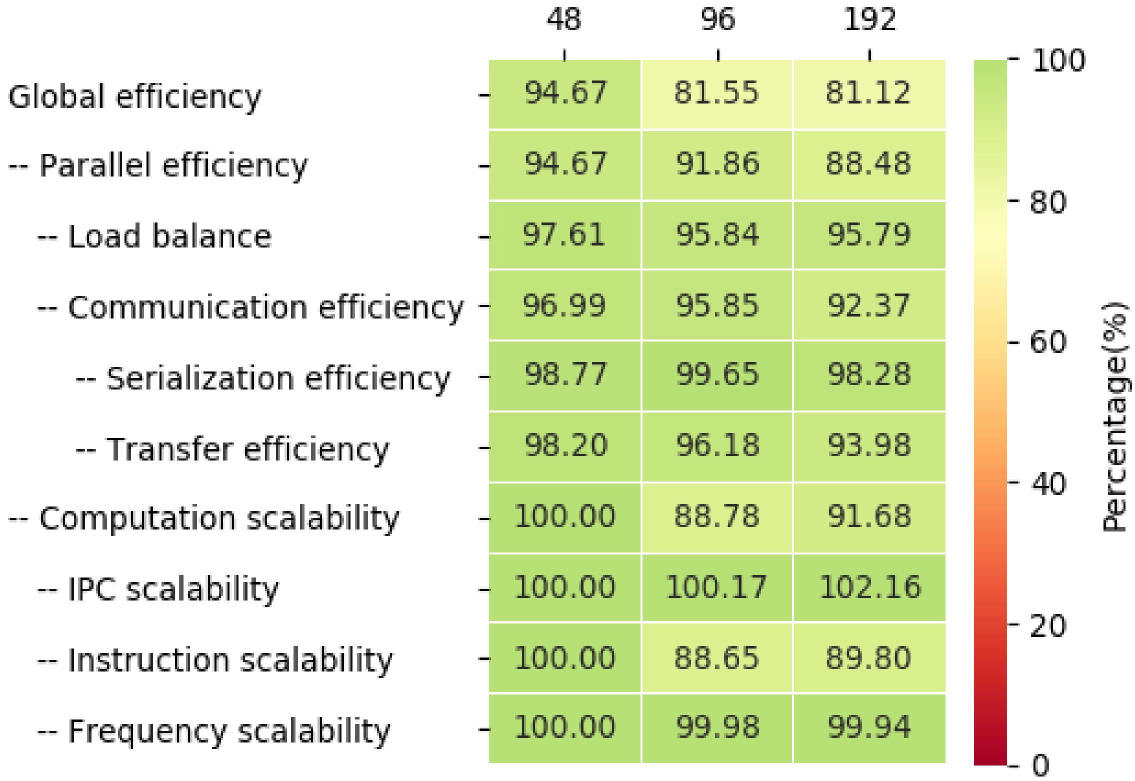


Figure 6 - Model factors numbers of the three MPI scalability cases

5.6 HPCW licenses

HPCW current licenses:

- Open source (Apache, CeCILL v2.0 license (GNU GPL compatible), etc.):
    - HPCW Framework (To be defined in ESiWACE2)
    - NEMO
    - ecRad
    - Kronos
- ESCAPE license:
    - IFS-FVM
    - dwarf-p-radiation-acraneb2-lonlev (ACRANEB2)?
- RAPS license:
    - IFS RAPS
- ICON:
    - ICON licenses

Document History

Version	Author(s)	Date	Changes
0.1	Ralf Mueller (DKRZ), David Guibert (BULL), Erwan Raffin (BULL)	22/09/2021	First draft
0.2	Erwan Raffin (BULL)	27/09/2021	Minor corrections, update results with v1.0 and add effort estimate
0.3	Mario Acosta (BSC), Daniel Beltran (BSC)	05/10/2021	BSC contribution added
1.0	Ralf Mueller (DKRZ), Erwan Raffin (BULL)	11/10/2021	Final version including internal reviewers' comments and requests

Internal Review History

Internal Reviewers	Date	Comments
Mario Acosta (BSC)	05/10/2021	
Italo Epicoco (CMCC)	11/10/2021	

Effort Contributions per Partner

Partner	Efforts
DKRZ	11.03 PM (Person Month)
Bull	17.9 PM
BSC	3 PM
Total	**31,93 PM**

ESCAPE 2

ECMWF Shinfield Park Reading RG2 9AX UK

Contact: peter.bauer@ecmwf.int

The statements in this report only express the views of the authors and the European Commission is not responsible for any use that may be made of the information it contains.