



BSC Profiling reports outcomes

Carlos Peña and Joan Vinyals

Barcelona Supercomputing Center



Funded by
the European Union

Earth Sciences
Department



**Barcelona
Supercomputing
Center**
Centro Nacional de Supercomputación

BSC's contribution to WP10

Performance Profiling of EERIE Configurations

IFS	HadGEM3	Icon	FESOM
ECMWF	Met Office	MPI	AWI

BSC profiling and performance analysis

Configuration & scaling

How resource allocation and MPI/OpenMP layouts affect performance.

Runtime analysis

Where the model spends time in computation and communication.

Guiding optimisation

Configuration recommendations and code improvement priorities based on the findings.

IFS

Standalone atmosphere and wave model

Performance analysis

HPC SYSTEM

HPC2020

MODEL CYCLE

CY48R1

HORIZONTAL RESOLUTIONS

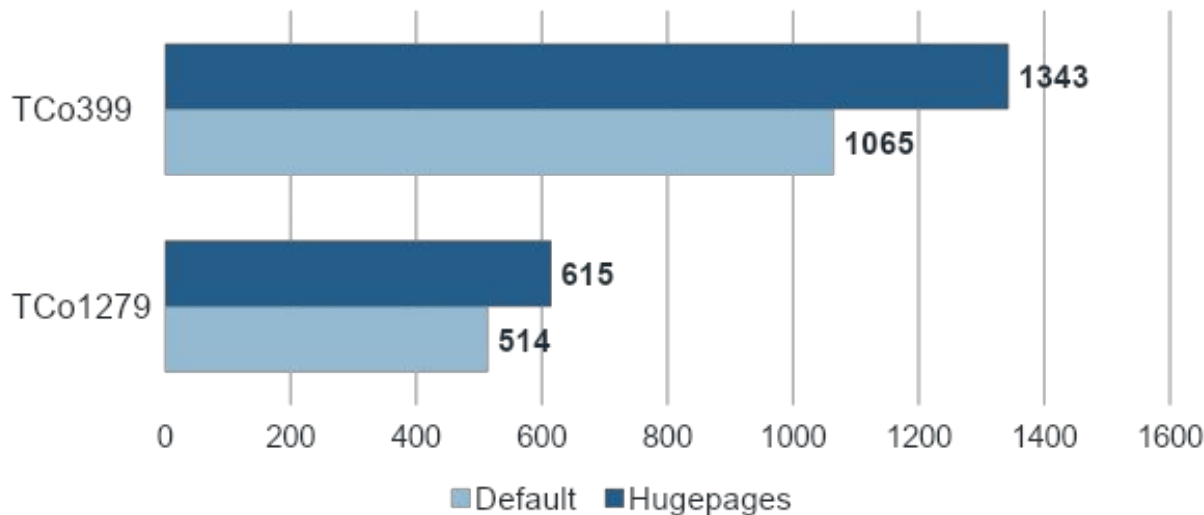
25 km (TCo399)

10 km (TCo1279)

IFS: configuration and throughput



Hugepages improve throughput at both resolutions



+20–26%

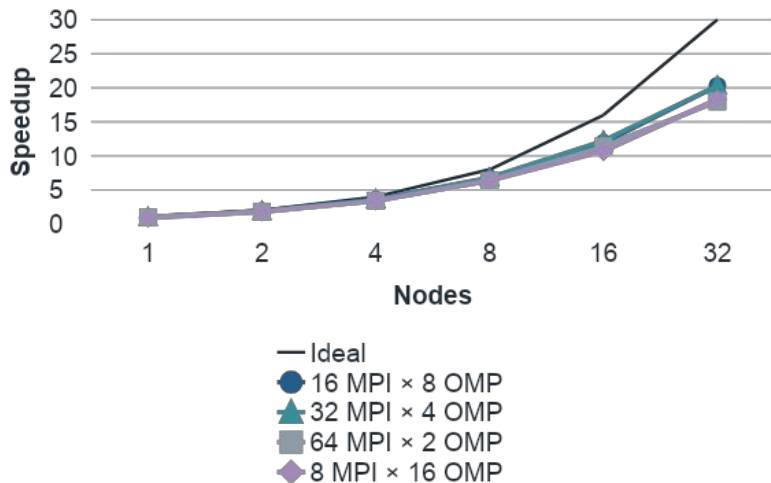
**Throughput gain
with hugepages**

**16 MPI × 8 OpenMP
threads per node**

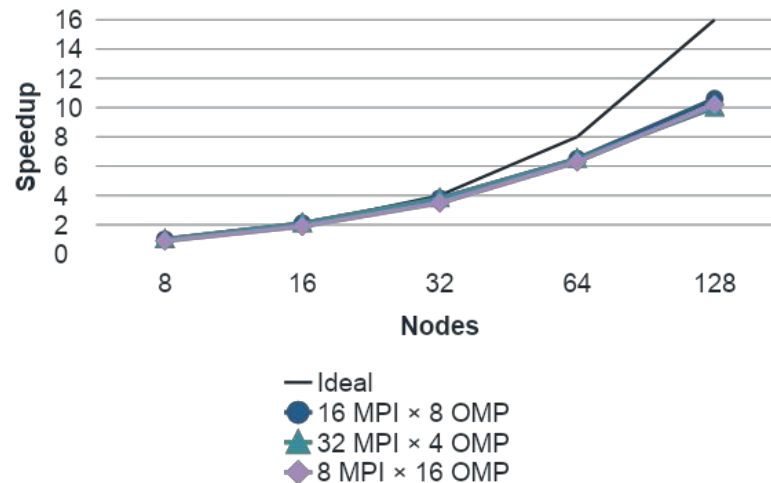
IFS: strong scaling

MPI/OpenMP layouts compared with ideal scaling, without I/O

TCo399



TCo1279



16 MPI × 8 OpenMP threads is a strong choice across both resolutions

IFS: component strong-scaling efficiency

Speedup relative to ideal scaling, with I/O enabled

Component	TCo399 1 to 32 nodes	TCo1279 16 to 64 nodes
Physics	82%	100%
Radiation	79%	95%
Dynamics	54%	80%
TRANS	39%	78%
WAM	58%	78%
I/O	21%	53%

TRANS and dynamics

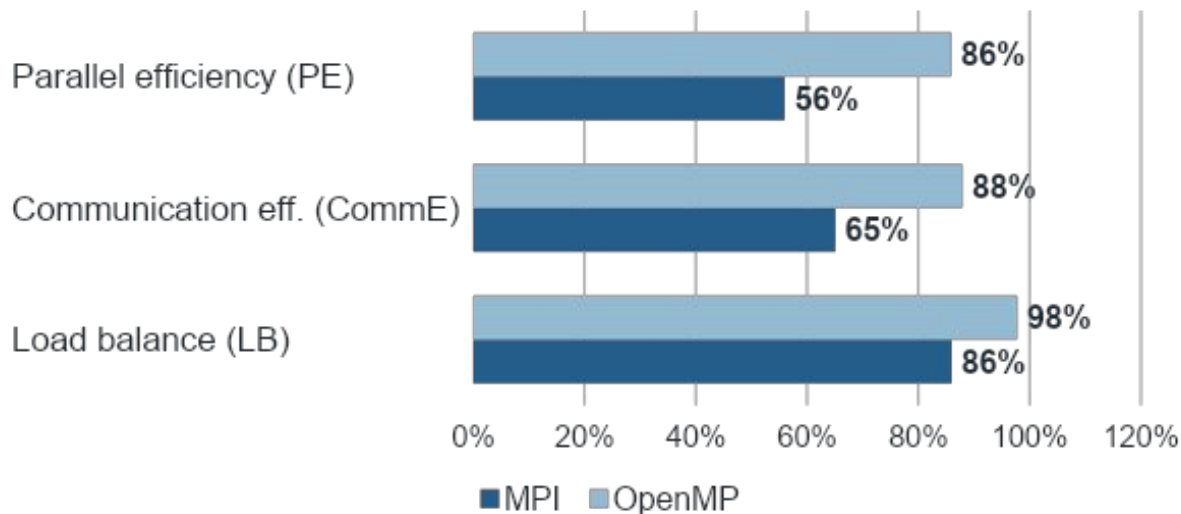
Substantial runtime cost and weaker node scaling

I/O scales worse, but its runtime share is small

Strong-scaling efficiency = speedup / node ratio. 100% = ideal.

IFS: POP performance metrics

TCo1279 on 160 compute nodes, hybrid MPI/OpenMP analysis



48.1%

Hybrid parallel efficiency

MPI communication efficiency falls

79.0% to 65.1%
from 80 to 160 nodes

LB = mean useful time / maximum useful time

HadGEM3

Atmosphere-only

Performance analysis

HPC SYSTEM

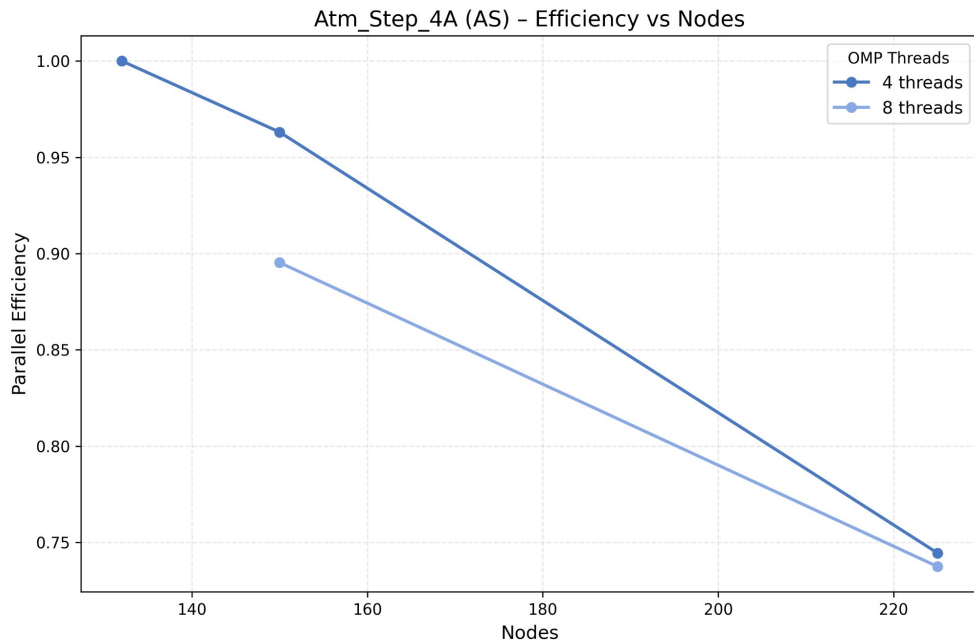
Met Office / Microsoft Azure

HORIZONTAL RESOLUTION

20 km (N640)

HadGEM3: atmosphere scaling

N640 atmosphere only, 132–225 nodes with 128 cores per node



**32 MPI × 4 threads
per node**

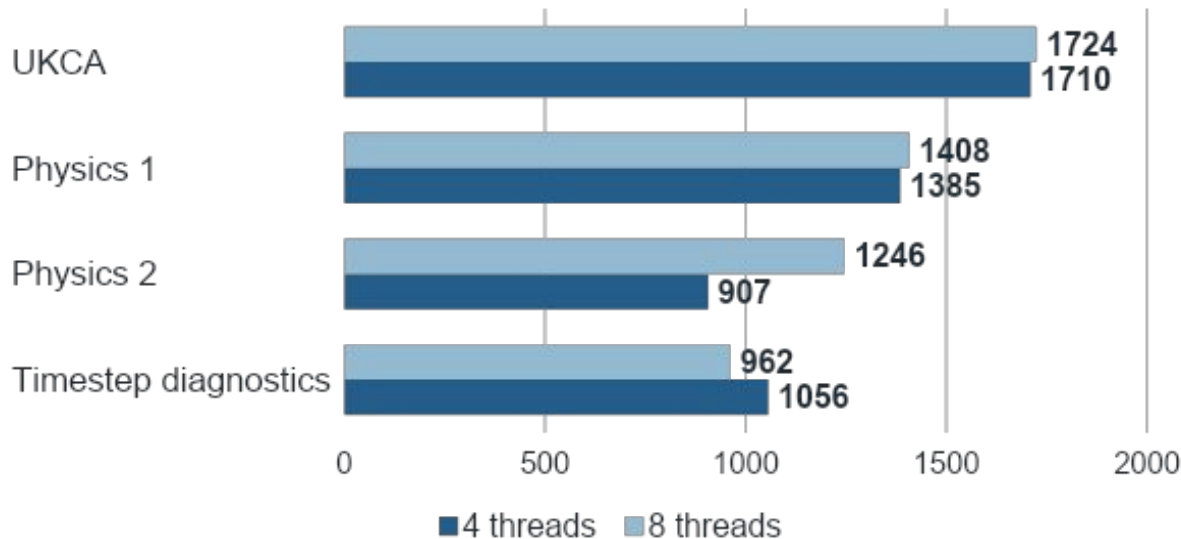
Best absolute performance
among tested layouts

<75%

Relative efficiency at 225
nodes

HadGEM3: output diagnostics scale least

MPI/OpenMP layout comparison at 225 nodes, timer totals in seconds



13–15%

Of the atmosphere
timer
in end-of-step
diagnostics

Little or no gain
from adding
resources

8-thread layout: AP2 timer +37%
AP2 still scales relatively well as nodes increase

Team follow-up identified output
optimisations

ICON

Atmospheric component (ICON-A) and Ocean component (ICON-O)

Performance analysis

HPC SYSTEM

Levante

HORIZONTAL RESOLUTIONS

Atmosphere: 10 km (R2B8)

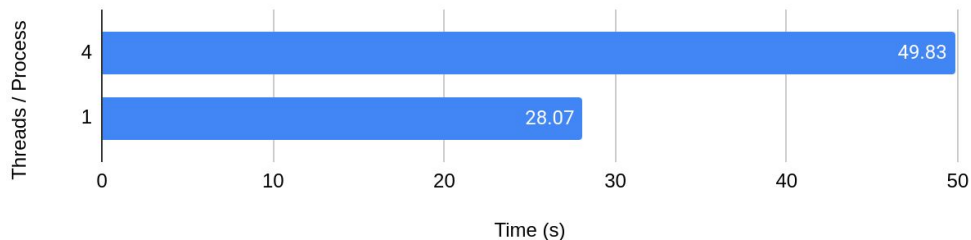
Ocean: 5km (R2B9)

ICON: OpenMP Scalability

Atmosphere

integrate_nh - Elapsed Time

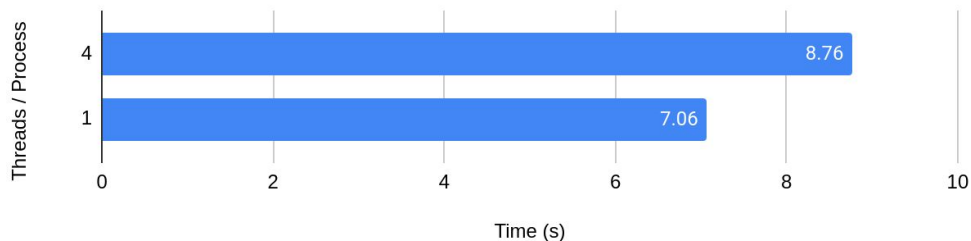
120 nodes



Ocean

solve_ab (without solve_init) - Elapsed Time

120 nodes



Throughput

From **72.66 SDPD**

with 3840 MPI x 4 OpenMP Threads

x1.92

to **139.75 SDPD**

with 15360 MPI x 1 OpenMP Threads

IFS-FESOM

FESOM

Performance analysis

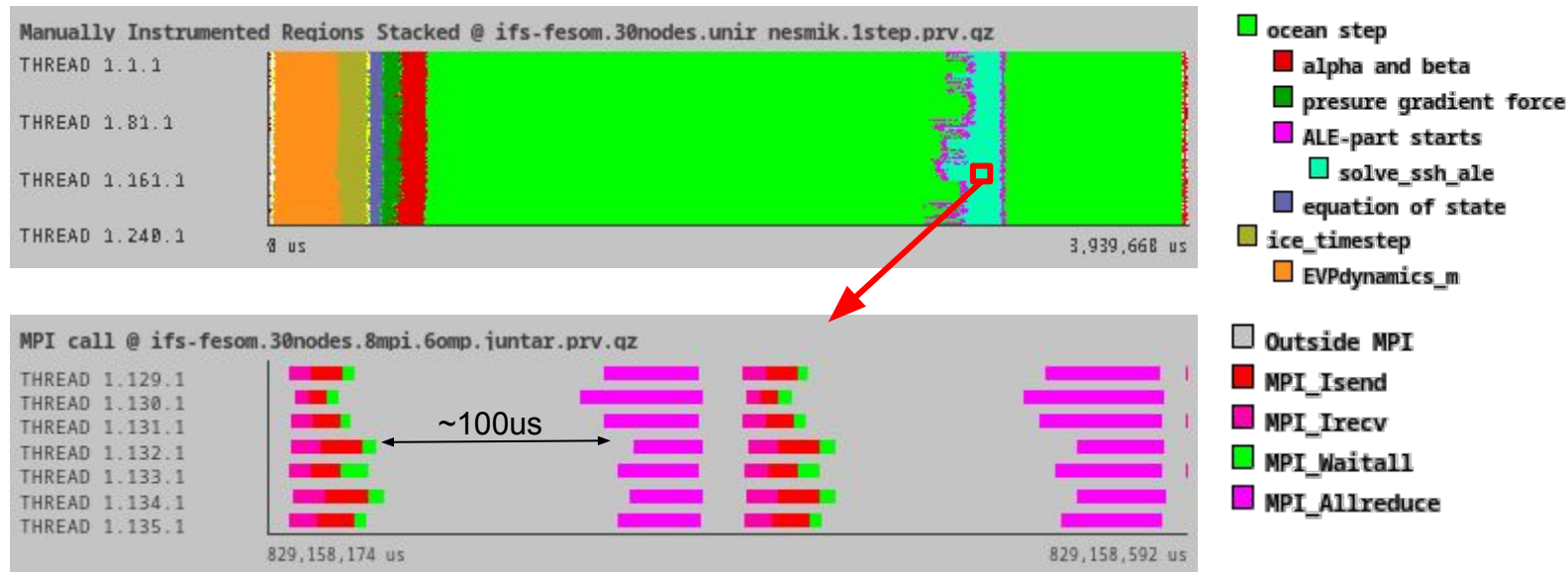
HPC SYSTEM

JUWELS Supercomputer

CONFIGURATION

As used in runs for Phase 1 of the EERIE project

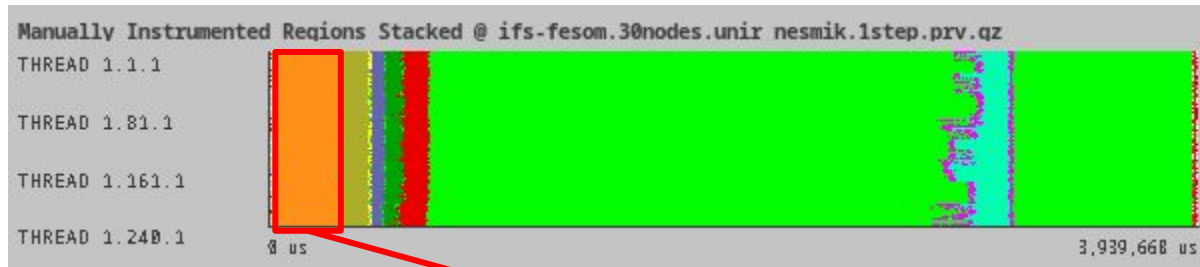
FESOM: solve_ssh_ale



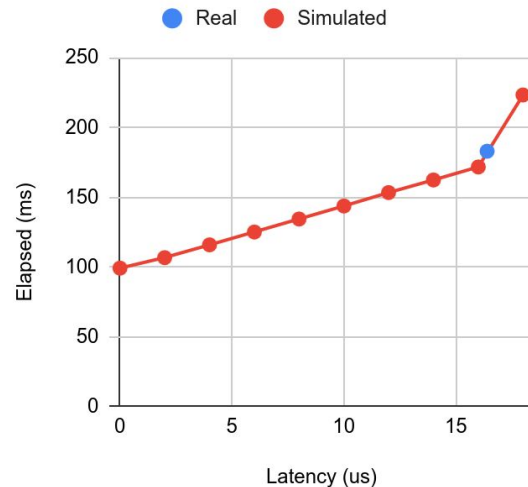
Granularity

This region executes global communications for every computation of ~100us. Granularity is TOO low for this solver.

FESOM: sea-ice (EVPdynamics_m)



sea-ice simulated elapsed time
wrt. MPI Call latency



Load imbalance

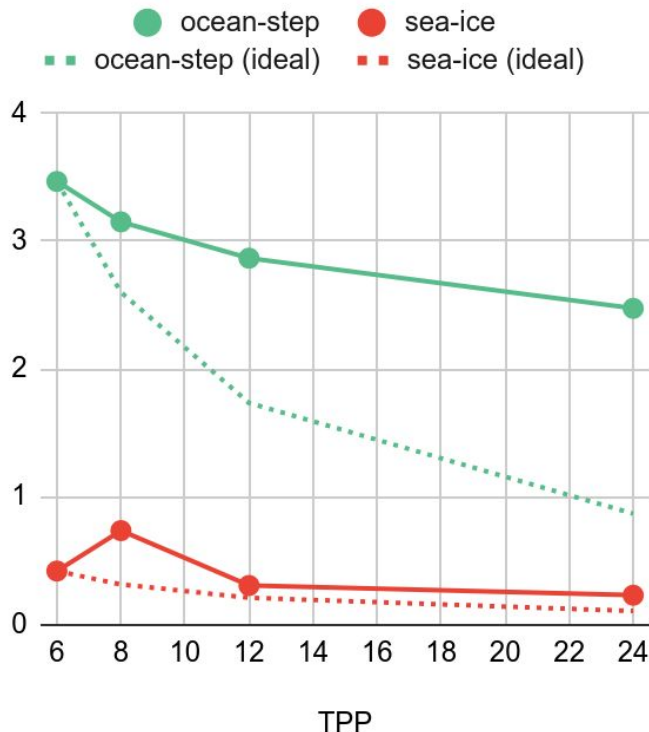
240 MPI Processes
Only 38 with any work.

MPI Call Overhead

Reducing the number of MPI
calls, can improve performance.

FESOM: Profiling OpenMP

Time [s]



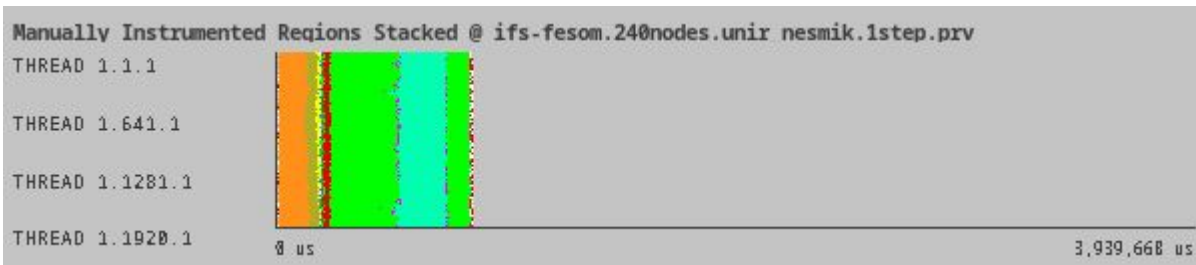
OpenMP Coverage

With Amdahl's law, we infer a 20% of serial code for the *ocean-step*, and a 10% for the *sea-ice*.

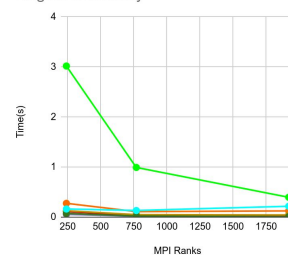
Manual instrumentation

With manual instrumentation we found that the subroutine *cvmix-tke* lacked OpenMP parallelization, which led to it being the main bottleneck when scaling the number of threads.

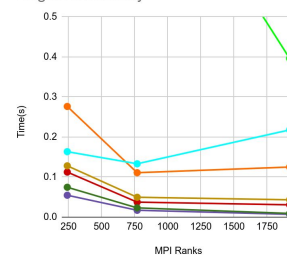
FESOM: Scalability



Region Scalability



Region Scalability



- No region
- output
- restart
- ocean step
 - alpha and beta
 - pressure gradient force
 - ALE-part starts
 - solve_ssh_ale
 - equation of state
- ice_timestep
- EVPdynamics_m