

ESiWACE – NEMO working group

1km feasibility– 10/04/2019

BENCH1 – Strong scalability

NEMO – BENCH1 – Routines comparisons

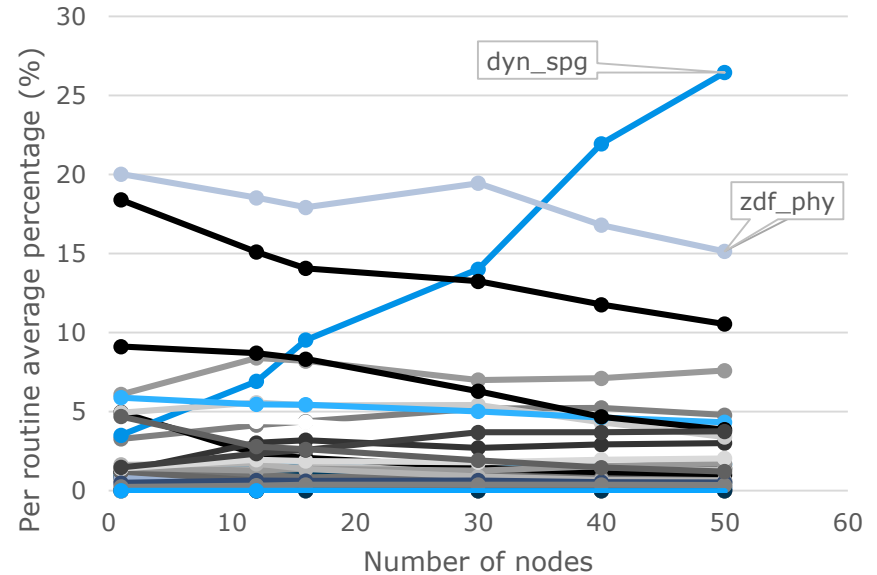
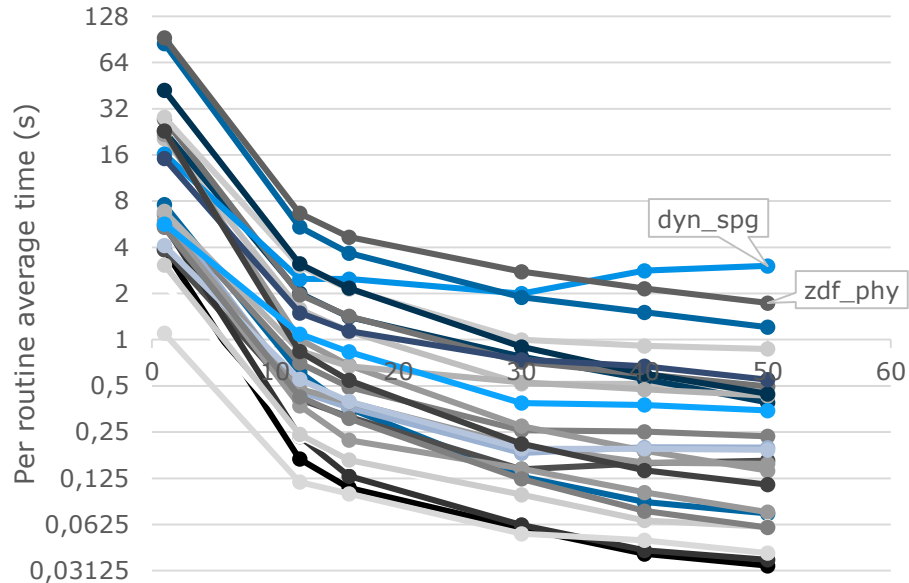
Strong scalability (up-to 10x10x75 subdomains)

Without init and restart – no stpctl – Square decomposition - nperio 7

2x SKL20c per node - 1000 iterations

Turbo - THP - no MMAP - no HT

Intel MPI 18.2.199 – MPI bounds to physical cores

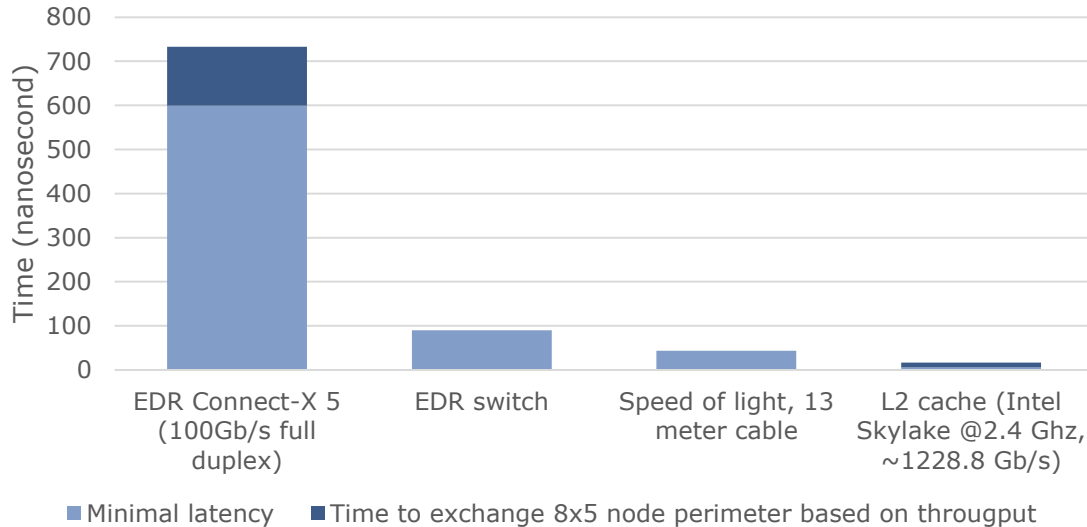


Communication/computation comparison

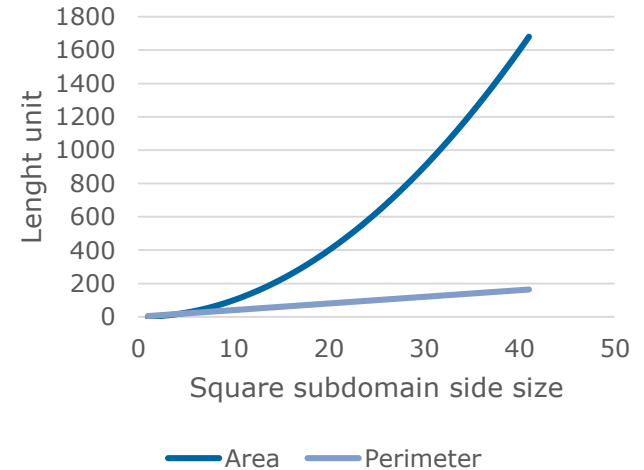
Networks - Orders of magnitude overview

Lower is better

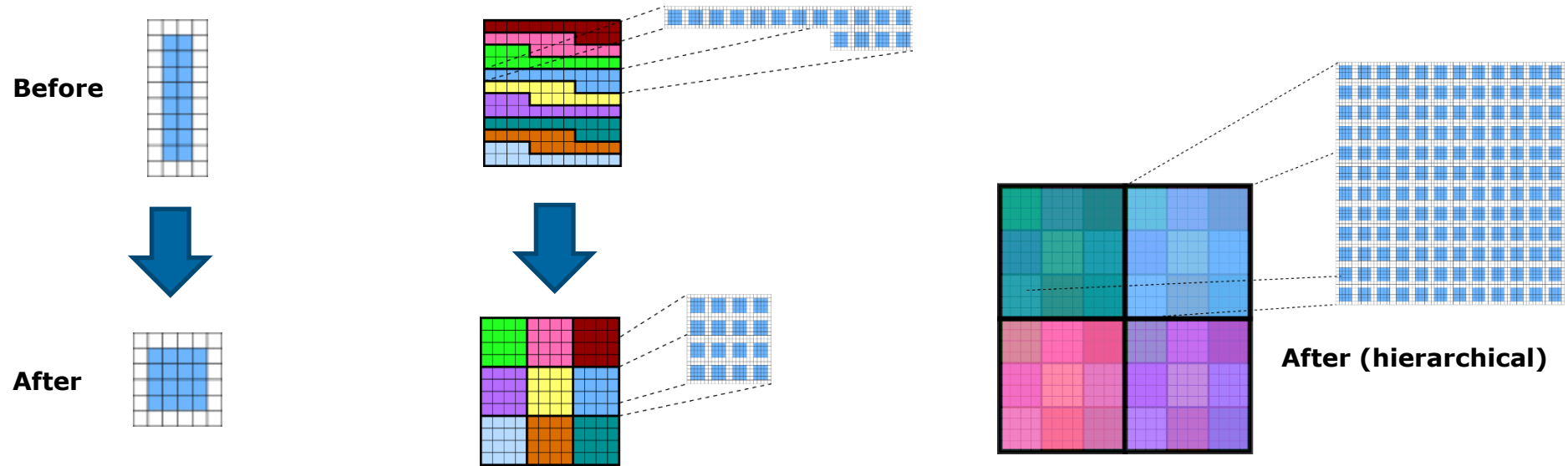
based on theoretical latency and bandwidth
10x10 subdomains w/ 8x5 squared placement per node
Inter-node perimeter of $(8 \times 2 + 5 \times 2) \times 64 = 1664$ Bytes



Trivial comparison between area and perimeter



Square for large boundaries exchanges

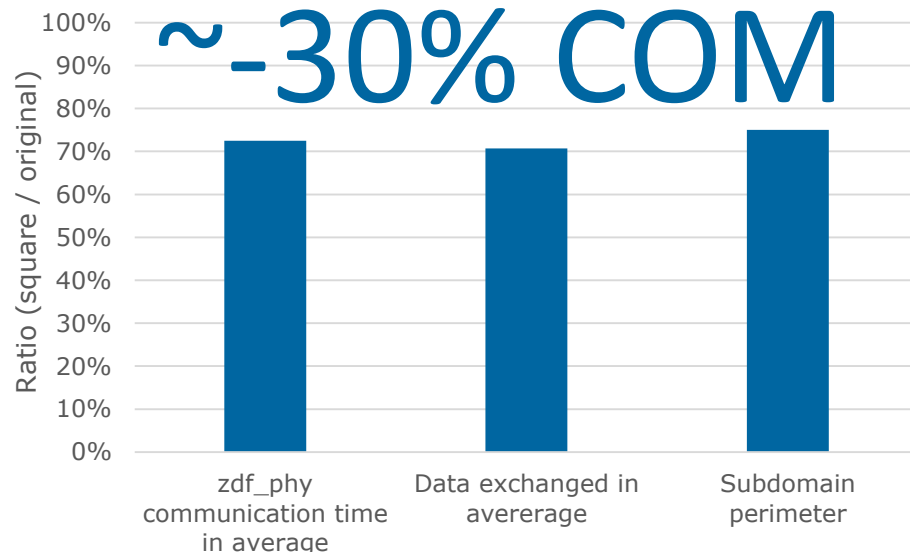
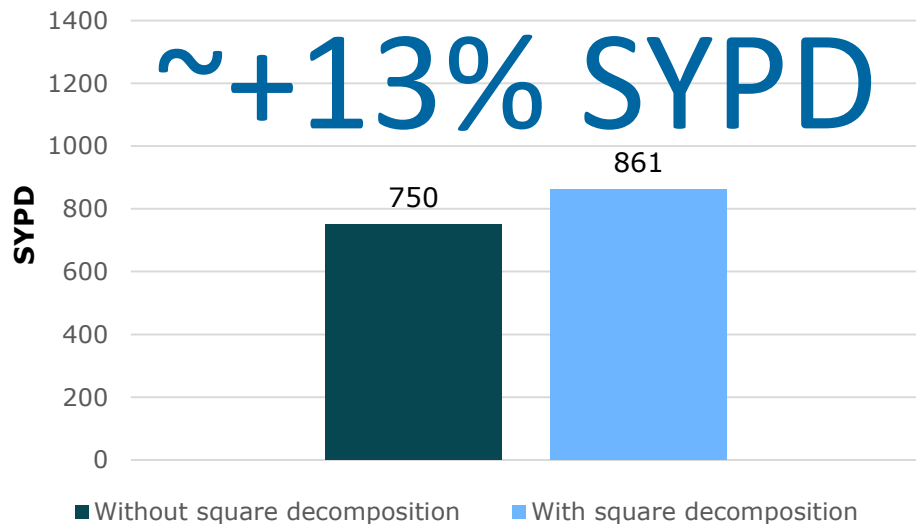


Square for large boundaries exchanges

NEMO ocean – BENCH1

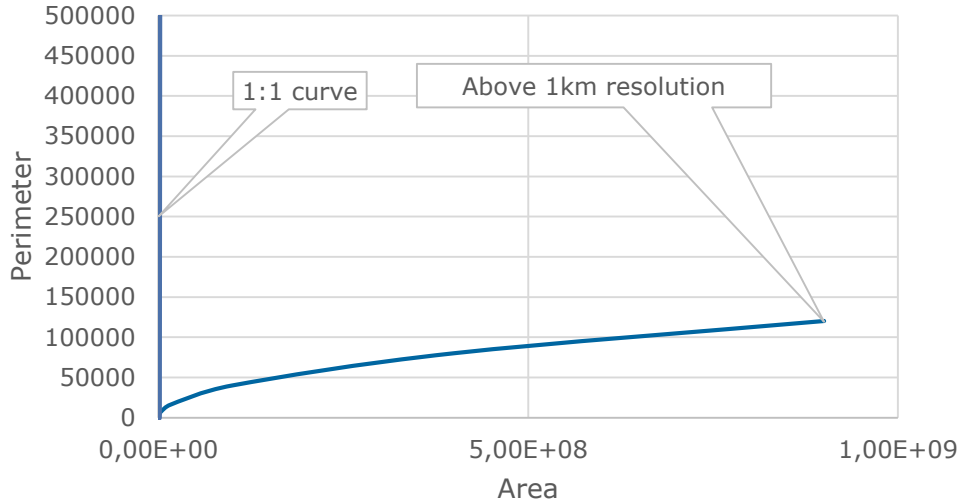
Square subdomains

362x332 – 1000 iterations – 50 nodes – same node for both
Without init and restart – no stpctl – 11x10 instead of 5x23 subdomains
Each node w/ 2x Intel SKL20c 6148@2.4Ghz- 192GB memory
Mellanox EDR connect-x-4 (100 gb/s) – fat-tree topology
Turbo – THP – no MMAP – no HT – Intel MPI 18.2 bounded to physical cores



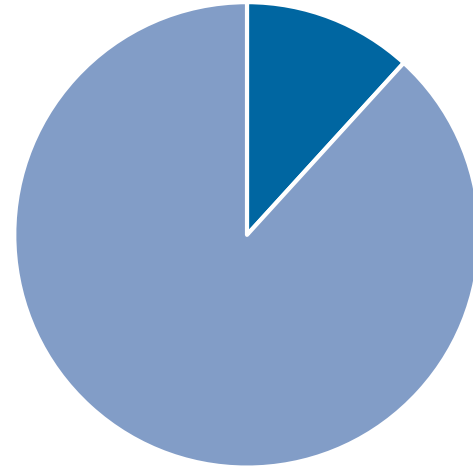
Dense node for feasibility

Square - area and perimeter



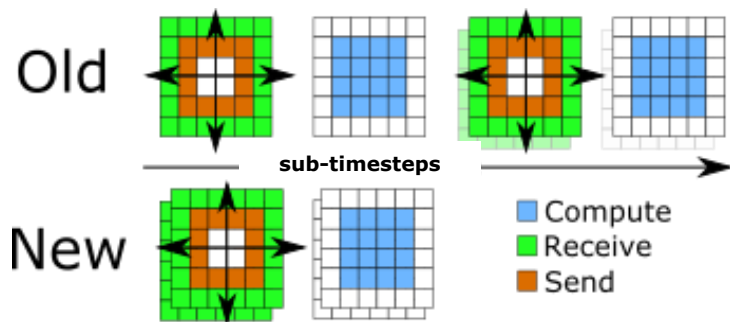
Earth in one node

Theoretical compute and exchange in number of points



- inter-node boundary exchange volume (x100 000)
- Compute area

Aggregate 2D layers for small exchanges



WIP - NEMO optimis - dynvor

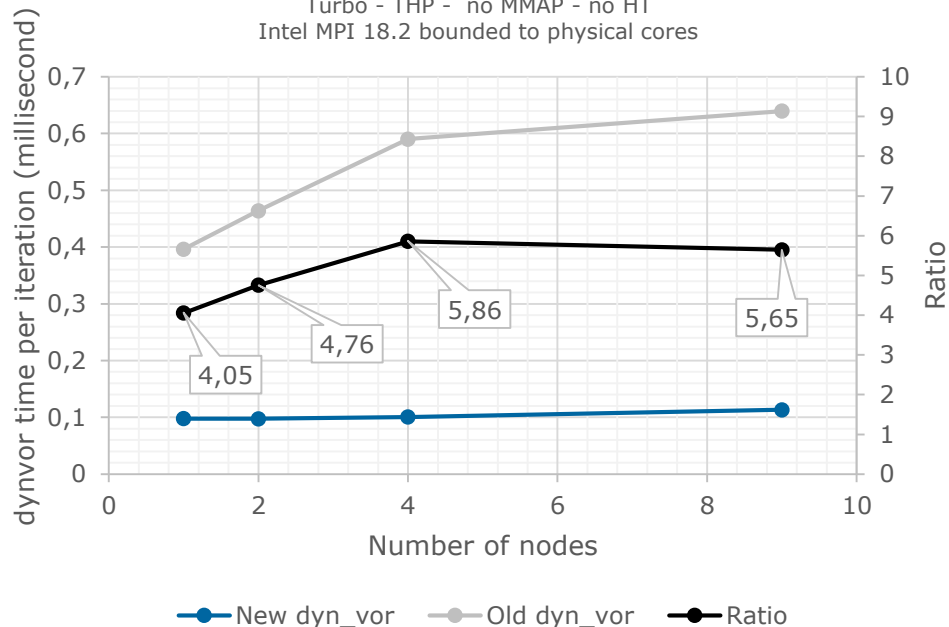
Routine time

Without init and restart - no stpctl - nnprio 7- JPI/JPJ square
each node w/ 2x SKL20c - 192GB memory per node

50 000 iterations

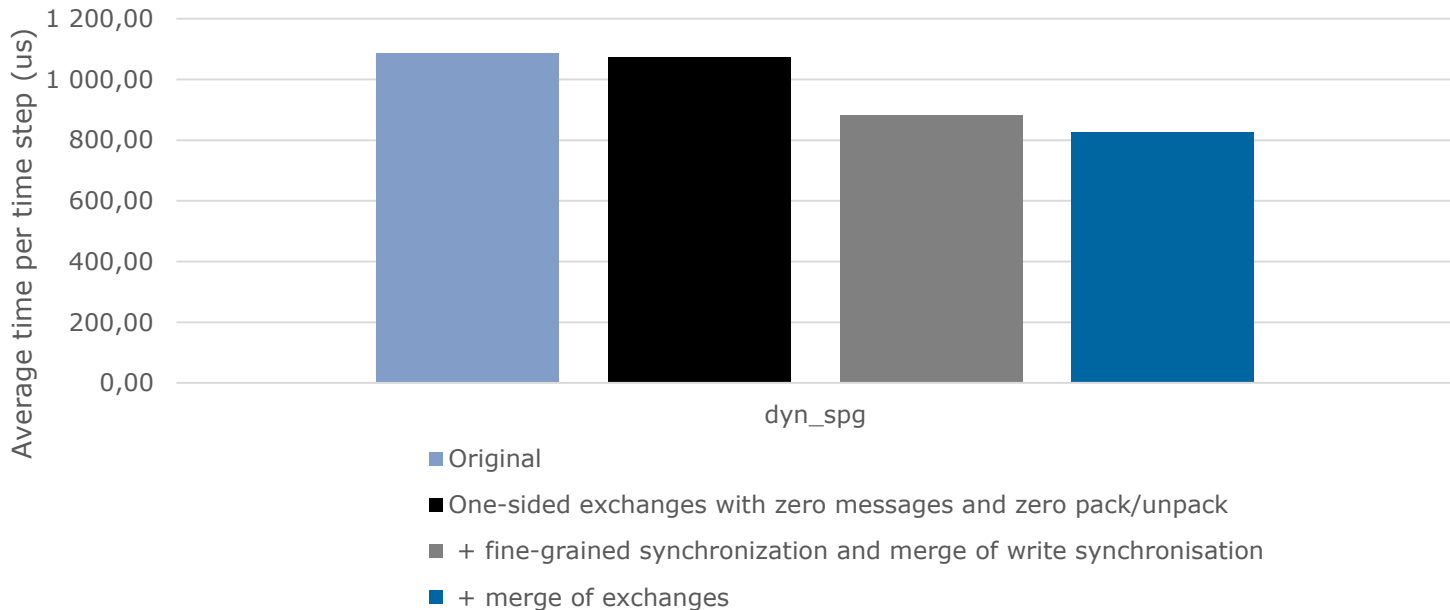
Turbo - THP - no MMAP - no HT

Intel MPI 18.2 bounded to physical cores



Low latency for small exchanges - PoC

NEMO ocean - BENCH1 physics - Lower is better
Single-node fine-grained shared memory exchanges PoC
Without init and restart - no stpctl - JPI fixed(8) - nnperio 7 - 10x10x75 subdomains
1 node w/ 2x SKL20c - 192GB memory - 1000 iterations
Turbo - THP - no M



BENCH at 1km – Extrapolation

NEMO ocean – BENCH1 - Extrap SYPD 1km

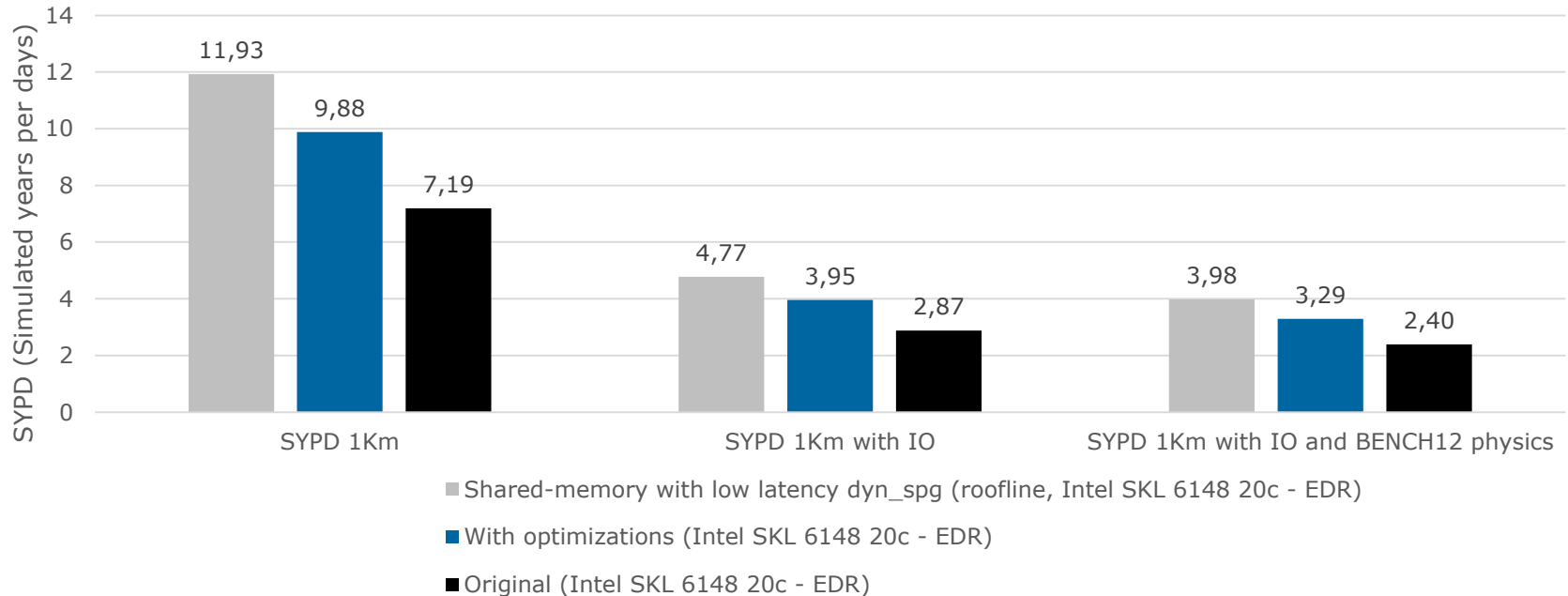
Extrapolation of 36000x24000 points on 9,120,000 cores

Without init and restart – no stpctl - 36s per simulated timestep

bi-socket per node – based on 5000 iterations – nnprio 7

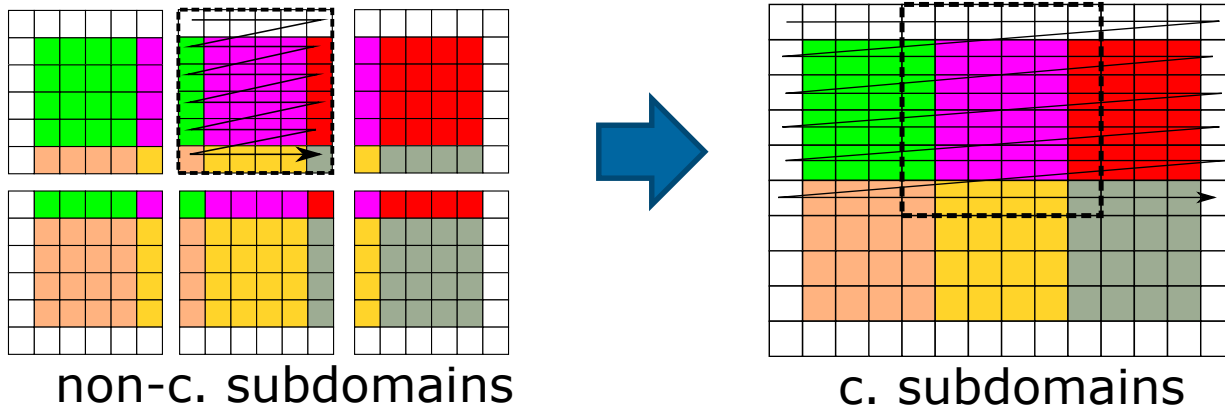
Turbo - THP - no MMAP - no HT

Intel MP



Still On-Going...

- ▶ MPI Shared Memory contiguous (identical to a coarse grain OpenMP approach)
 - **zero copy**
 - **zero ghost zone** → decrease memory usage
 - no loop modification, transparent (with fortran array shape)
 - but false sharing (to real small size: $10 \times 10 < \text{Page size}$)



esiwace
CENTRE OF EXCELLENCE IN SIMULATION OF WEATHER
AND CLIMATE IN EUROPE



This work has been funded by ESIWACE which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 675191

Thanks

For more information please contact:

david.guibert@atos.net

louis.douriez@atos.net

erwan.raffin@atos.net

Atos, the Atos logo, Atos Codex, Atos Consulting, Atos Worldgrid, Bull, Canopy, equensWorldline, Unify, Worldline and Zero Email are registered trademarks of the Atos group. October 2018. © 2018 Atos. Confidential information owned by Atos, to be used by the recipient only. This document, or any part of it, may not be reproduced, copied, circulated and/or distributed nor quoted without prior written approval from Atos.

Bull
atos technologies