NCI

NCI
NATIONAL COMPUTATIONAL INFRASTRUCTURE

SCALABILITY OF MOM 5, NEMO, AND MOM 6 ON NCI'S RAIJIN SUPERCOMPUTER

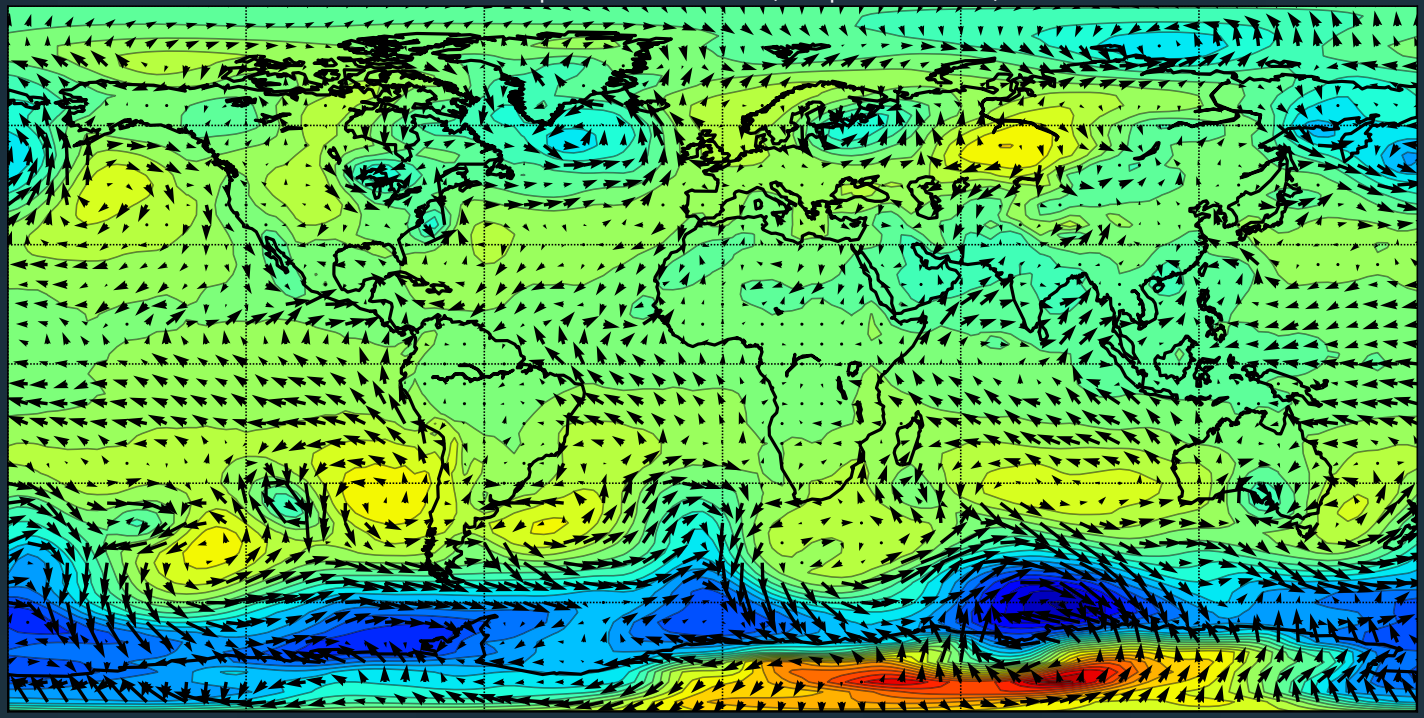
Marshall Ward
National Computational Infrastructure

NCRIS National Research Infrastructure for Australia — An Australian Government Initiative
Australian National University
Australian Government Bureau of Meteorology
CSIRO
Australian Government Geoscience Australia
Australian Government Australian Research Council
nci.org.au
@NCInews

http://marshallward.org/talks/ecmwf2016.html

ATMOSPHERIC SCALES

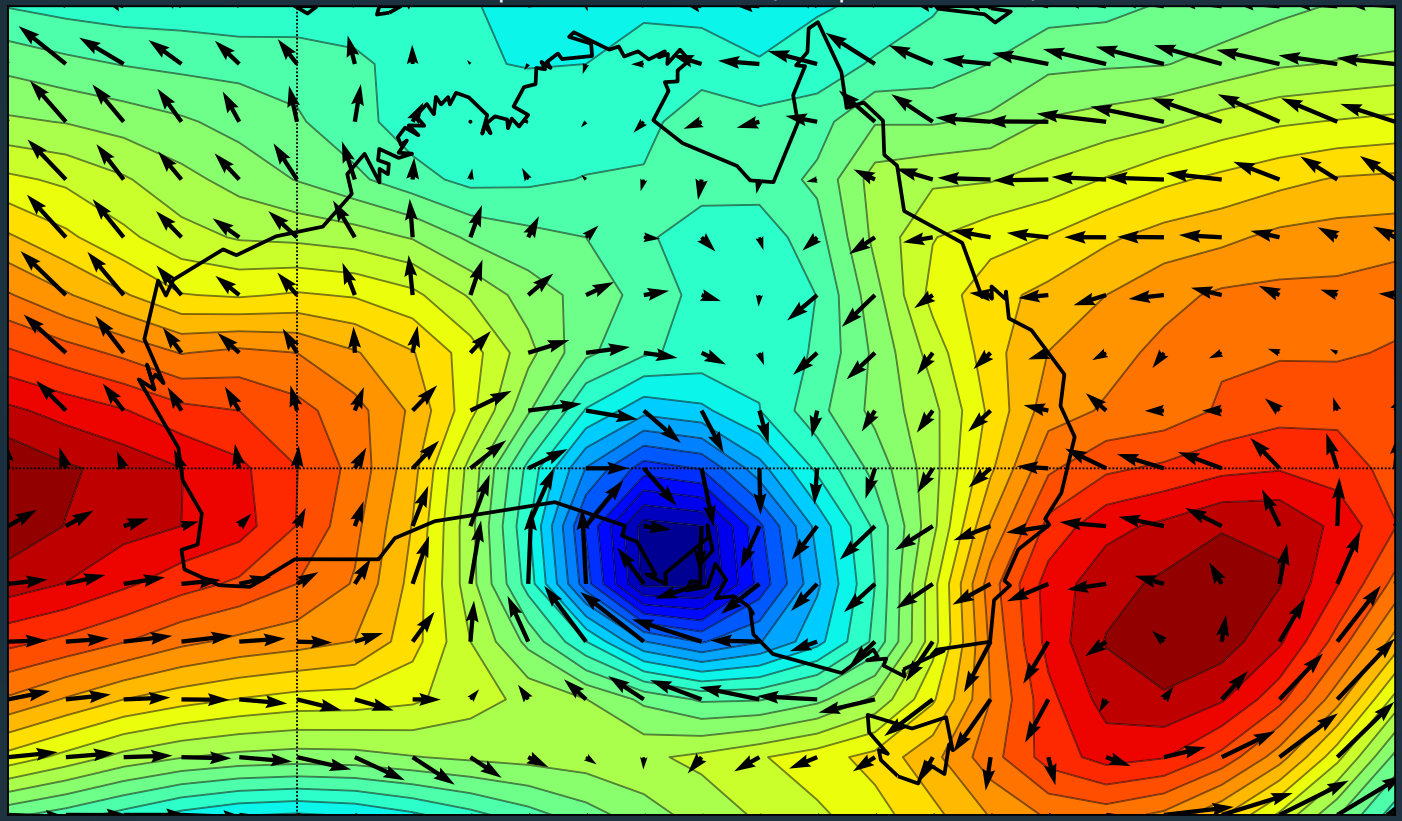
Sea level pressure and winds (2 September 2010)



http://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml

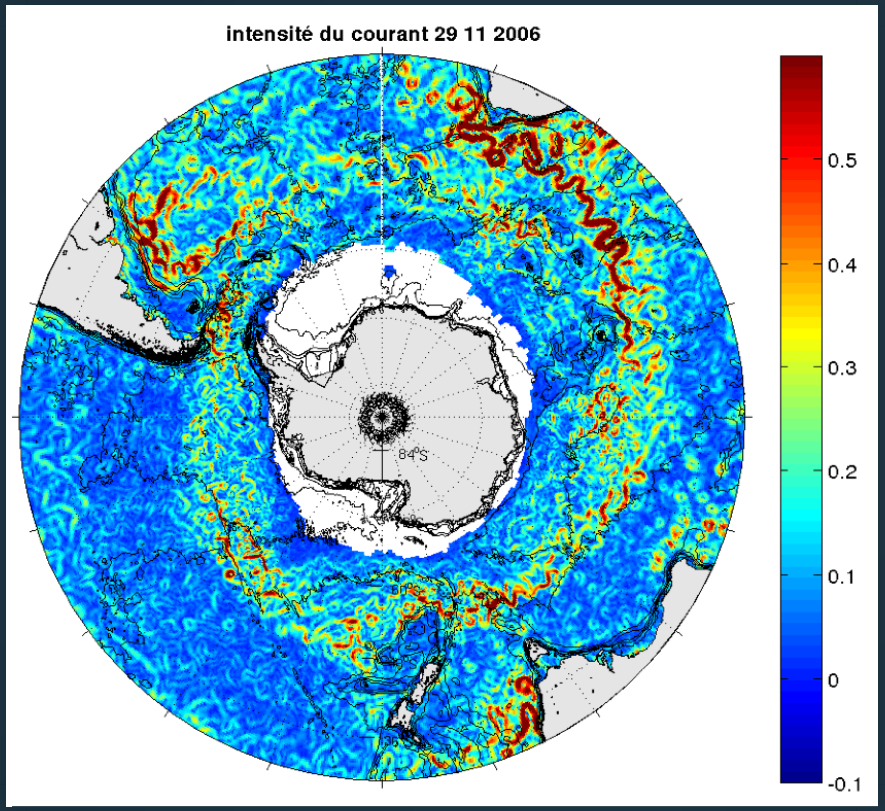
ATMOSPHERIC SCALES

Sea level pressure and winds (2 September 2010)



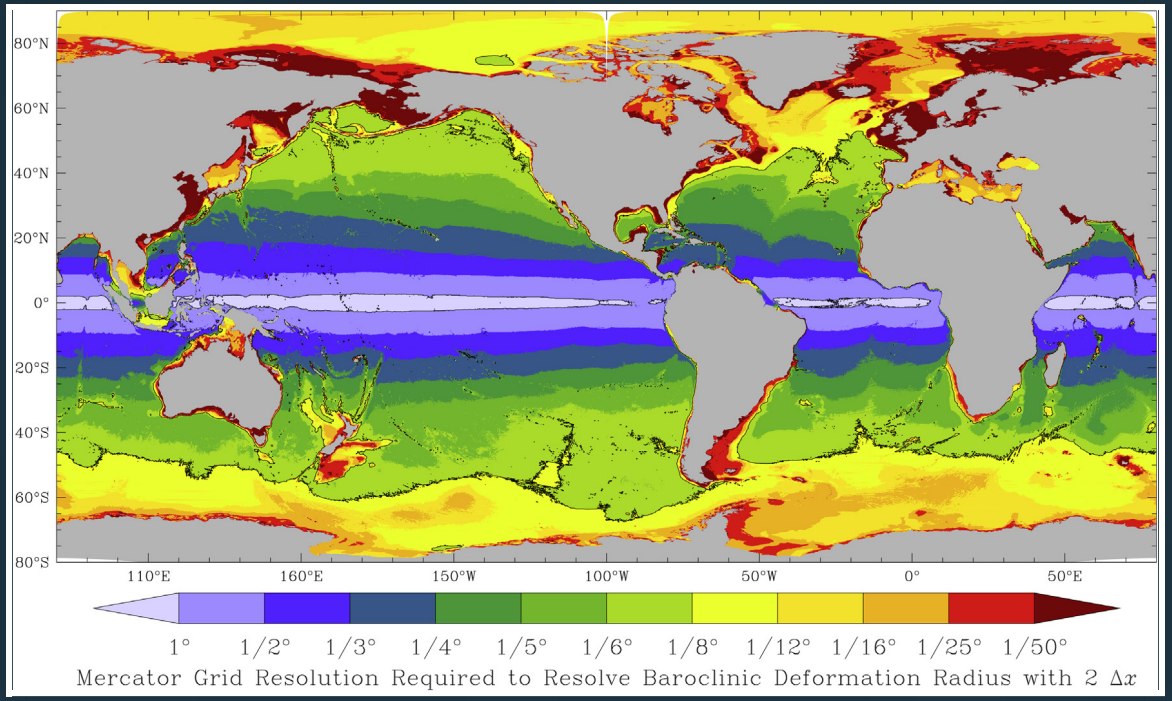
http://www.esrl.noaa.gov/psd/data/reanalysis/reanalysis.shtml

OCEAN SCALES



http://www.pmel.noaa.gov/people/cronin/ARC/ARC.html

LENGTH SCALE OF THE OCEAN



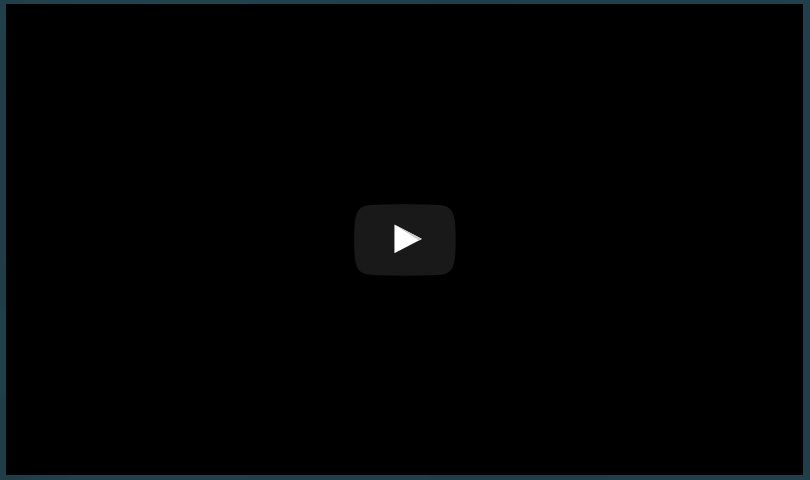
(Hallberg 2013)

OCEAN MODELLING AT NCI

Model	Resolution	Grid	CPUs
ACCESS-CM *(climate)*	1°	360 x 300	40
	0.25°	1440 x 1080	960
OFAM *(forecast)*	0.1° to 2°	1191 x 968	384
	0.1°	3600 x 1500	512
ARCCSS (GFDL) *(dynamics)*	0.25°	1440 x 1080	960
	0.1°	3600 x 2700	9196

OCEAN MODELS

MOM 5, NEMO

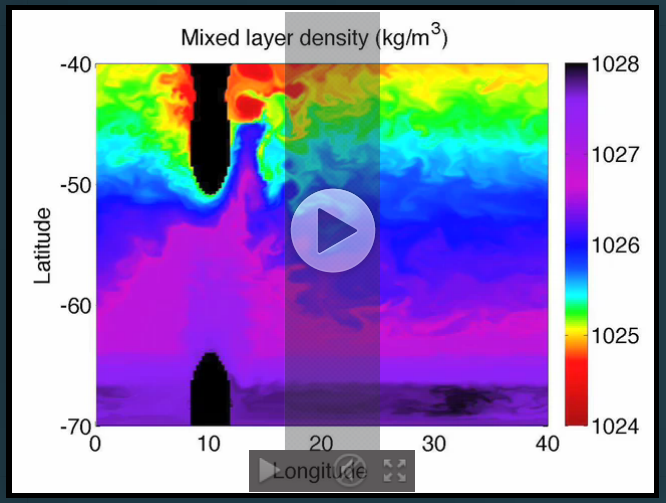


- Bryan-Cox models
- Finite volume advection
- Convective adjustment
- KPP mixing
- Submesoscale param.

(Video: Hogg et al. 2015)

MOM is B-grid, NEMO is C-grid

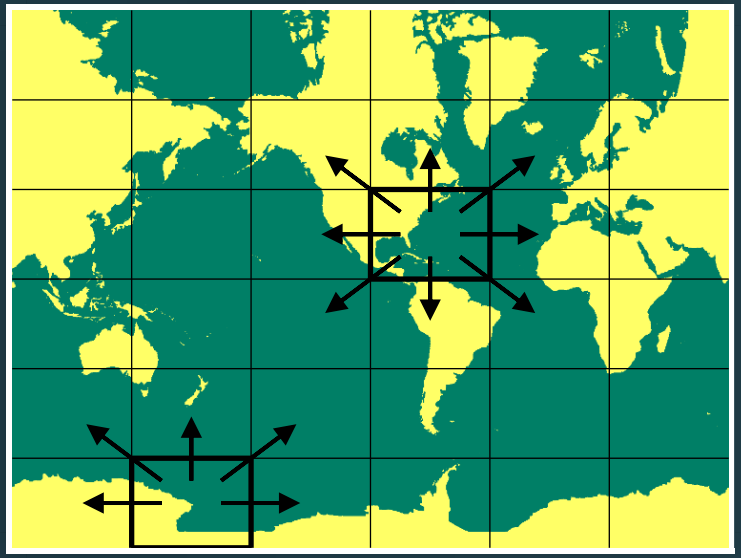
MOM 6

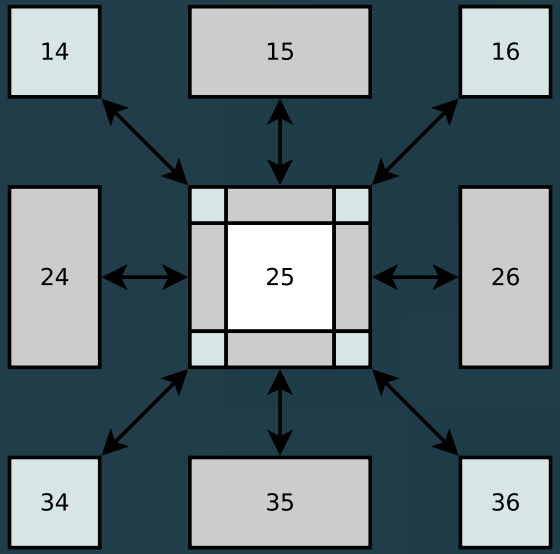


- Isopycnal (layered) dynamics
- Arbitrary Lagrangian-Eulerian
- Horizontal C-grid
- Union of GOLD and MOM 5
- New SIS2 sea ice model

(Video: Morrison et al. 2013)

TILE DECOMPOSITION





- Load-balanced using equal tiles, land-masked arrays
- Unbalanced message sizes (esp. diagonals)
- Land-only tiles removed in operational models

SPLIT TIMESTEPPING

Ocean stratification yields *fast* and *slow* dynamics:

$$\begin{aligned}\mathbf{u} &= \left\{ \frac{1}{H} \int_{-H}^{\eta} \mathbf{u} \right\} + \left\{ \mathbf{u} - \frac{1}{H} \int_{-H}^{\eta} \mathbf{u} \right\} \\ &= \mathbf{U} + \mathbf{u}'\end{aligned}$$

- Depth-averaged variability is ~100x faster
- Accuracy is less important; uses simpler solvers
- Additional filtering required

GENERALISED ORTHOGONAL COORDINATES

(Murray 1996)

Tripolar grids eliminate the "pole problem"

MODEL CONFIGURATIONS

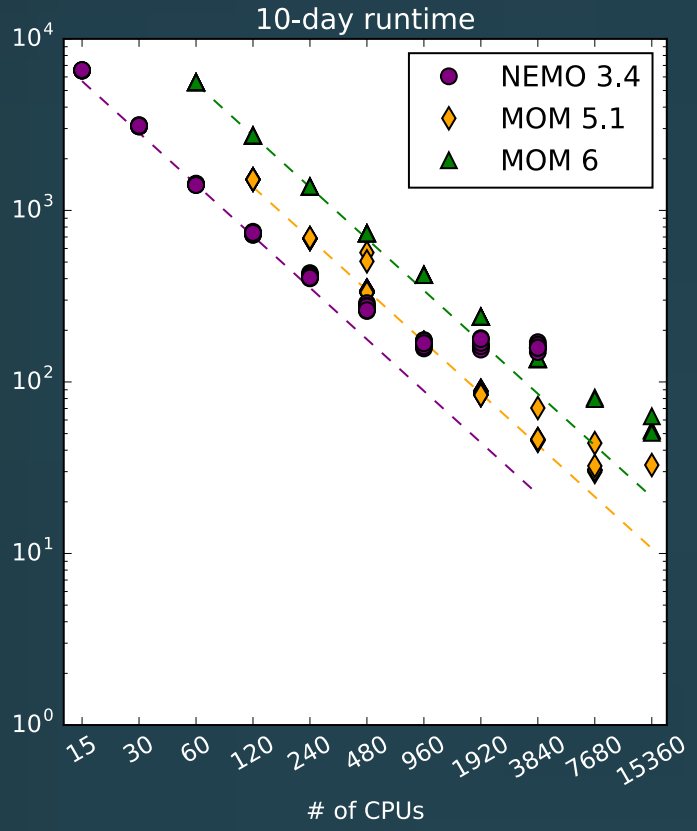
Model	Config	Resolution	Timestep
MOM 5.1	"CM2.5"	1440 x 1080 50 level	1800 s (22.5 s split)
NEMO 3.4	ORCA 0.25°	1442 x 1021 46 level	1440 s (24 s split)
MOM 6	OM4	1440 x 1080 75 layer	1200 s (~20 s split)

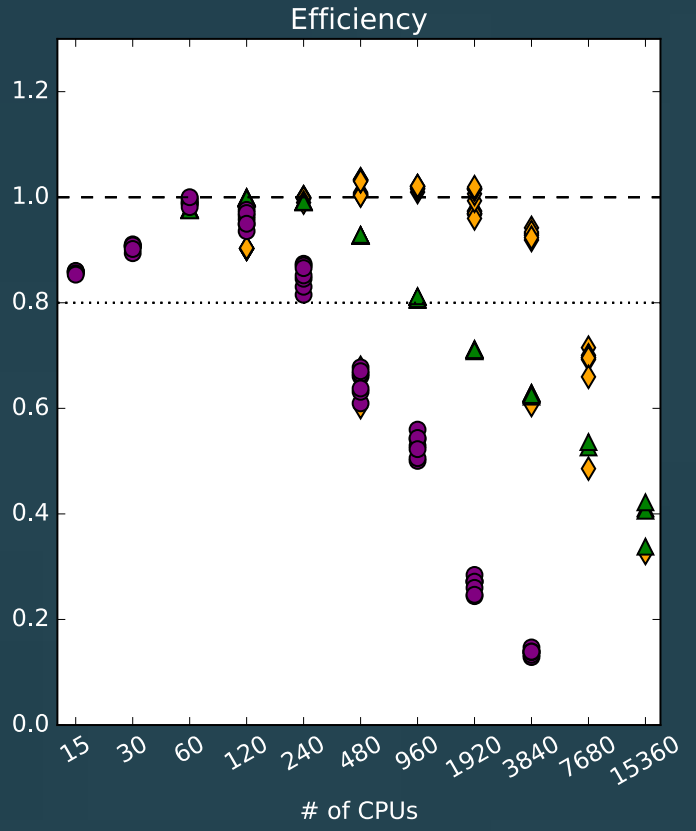
- Serial ocean-ice coupling
- 10-day runtime
- No model output

NCI PLATFORM: RAIJIN (雷神)

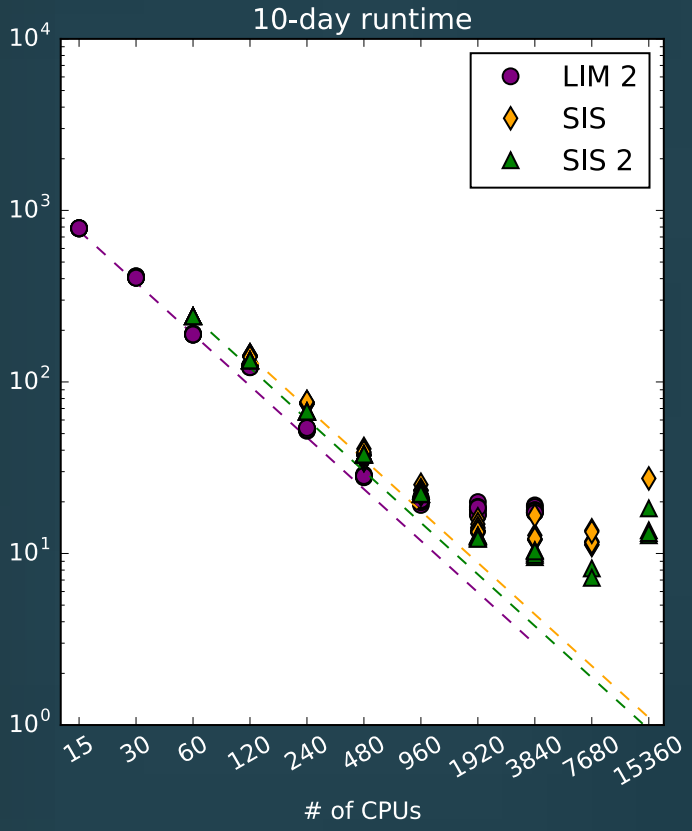
- 57,472 cores (3592 nodes, 16 core / node)
- Intel Xeon (Sandy Bridge), 3 GHz (turbo)
- 32+ GiB per node
- 56 Gb/s Infiniband network
- Two-level switched fabric fat tree
- $R_{\max}$ = 0.978 PFlops

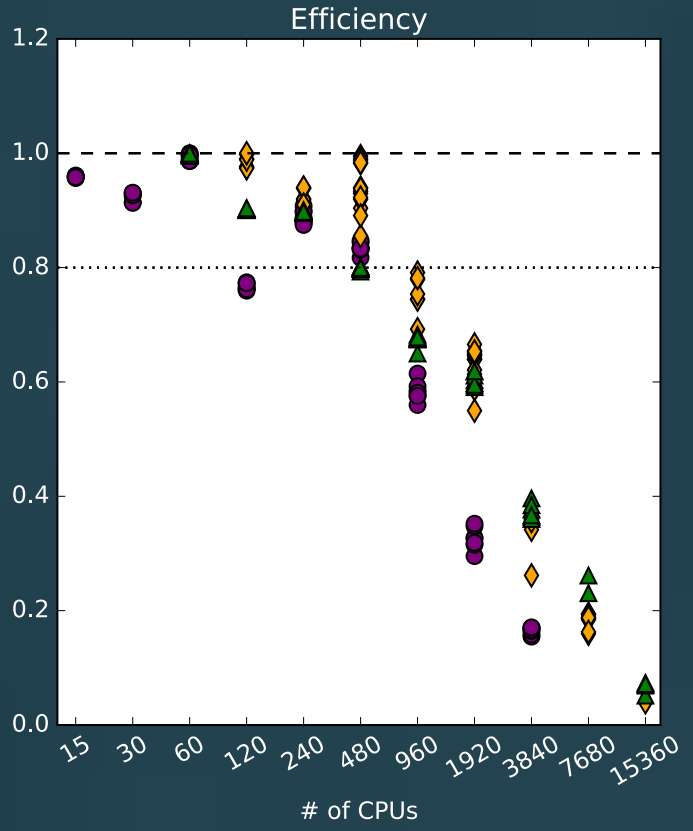
OCEAN RUNTIME



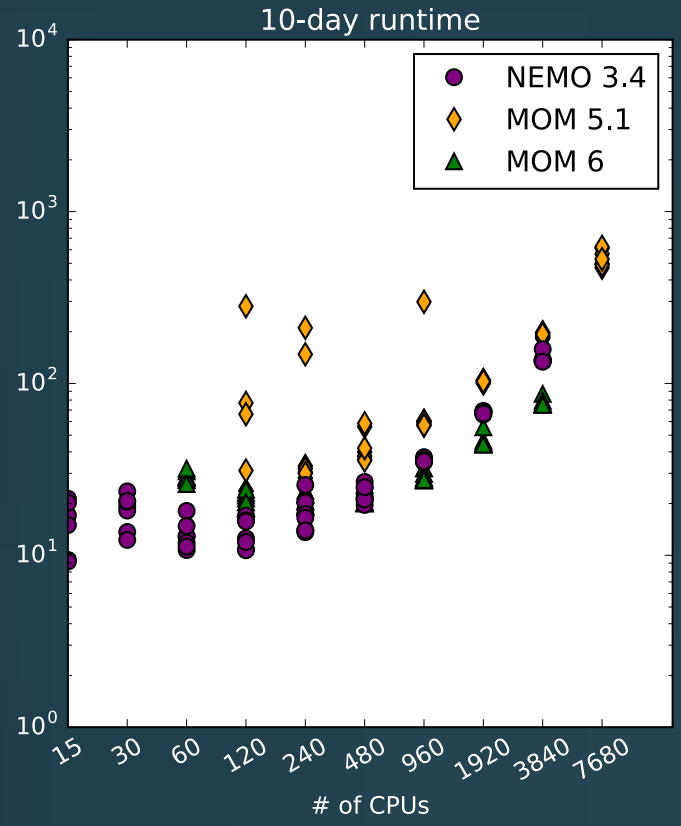


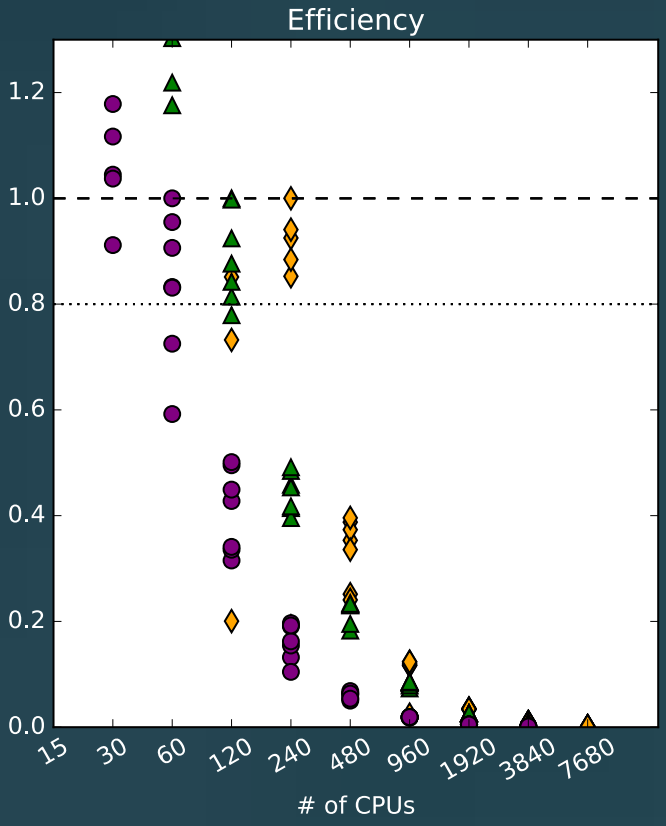
SEA ICE RUNTIME



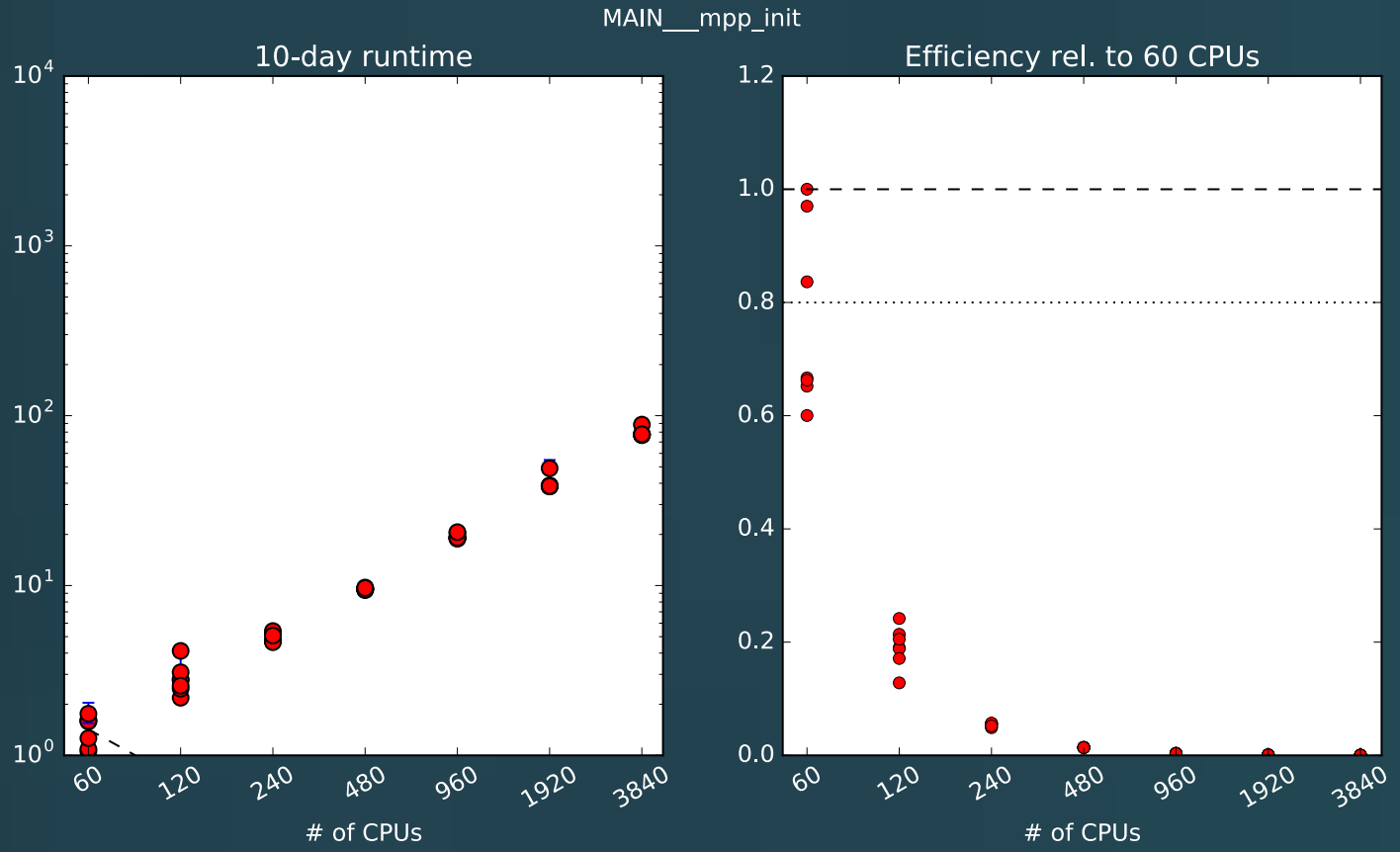


INITIALISATION





MPI INITIALISATION



1-YEAR SIMULATION TIMES

Model	CPUs	CPU Hrs	Time (s)	yr/day
MOM 5	960	1868.0	7088.2	12.1
	1920	2126.3	3986.8	21.7
NEMO	480	1068.0	8043.6	10.7
	1920	1964.2	3753.6	23.0
MOM 6	480	3910.1	29363.1	2.9
	3840	6349.0	6057.2	14.3

1 YEAR, MOM 5 VS NEMO

Model	CPUs	CPU Hrs	Time (s)	yr/day
MOM	480	1832.4	13783.9	12.1
	960	1868.0	7088.2	12.1
	1920	2126.3	3986.8	21.7
	3840	4075.2	3995.6	21.6
NEMO	480	1068.0	8043.6	10.7
	960	1368.8	5132.9	16.8
	1920	1964.2	3753.6	23.0
	3840	3547.8	3413.9	25.3

1 YEAR, MOM 5 VS MOM 6

Model	CPUs	CPU Hrs	Time (s)	yr/day
MOM	480	1832.4	13783.9	12.1
	960	1868.0	7088.2	12.1
	1920	2126.3	3986.8	21.7
	3840	4075.2	3995.6	21.6
MOM 6	480	3910.1	29363.1	2.9
	960	4458.6	16756.5	5.2
	1920	5154.6	9718.3	8.9
	3840	6349.0	6057.2	14.3

MODEL RUNTIME OBSERVATIONS

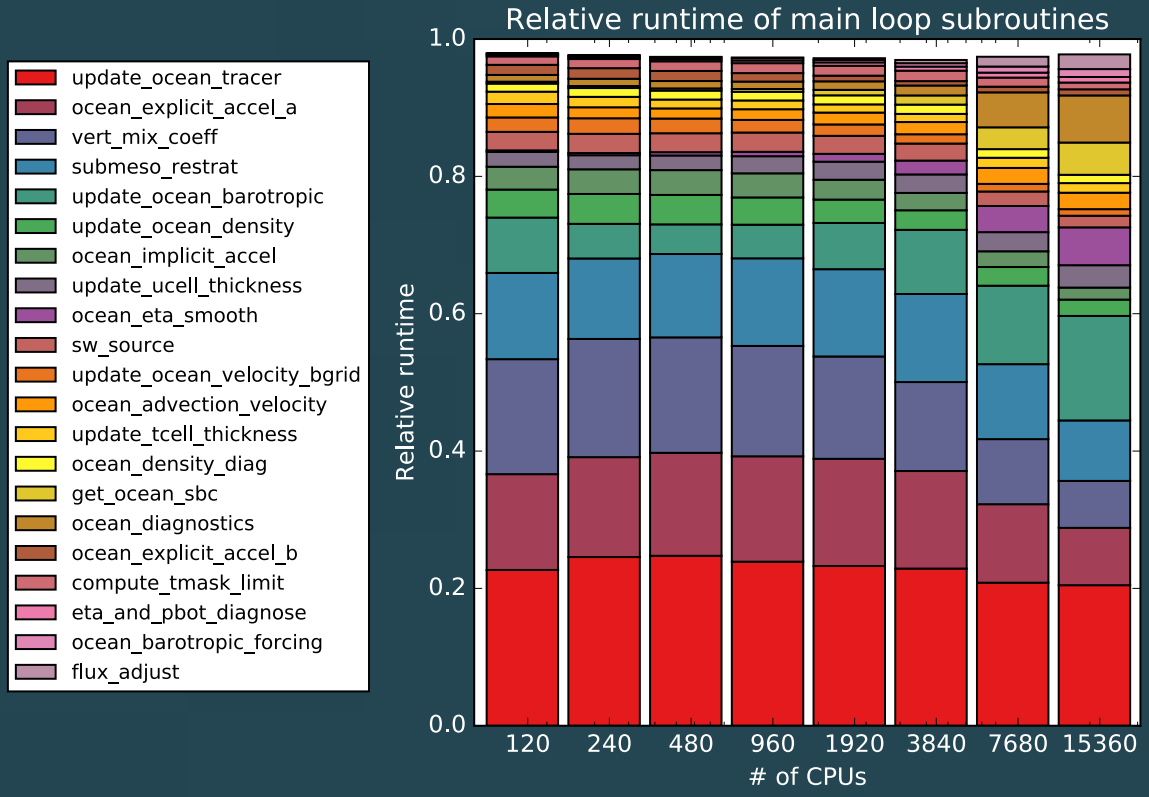
- NEMO
  - Lowest complexity (2.4 CPU hours / day)
  - Lowest memory usage (~50 GiB at 15 cores)
  - Drastic efficiency loss after 960 CPUs
- MOM 5
  - Moderate complexity (4.9 CPU hours / day)
  - Strong scaling up to 3840 CPUs
- MOM 6
  - Highest complexity (10.7 CPU hours / day, 75 levels)
  - Moderate efficiency loss after 240 CPUs

MODEL RUNTIME OBSERVATIONS

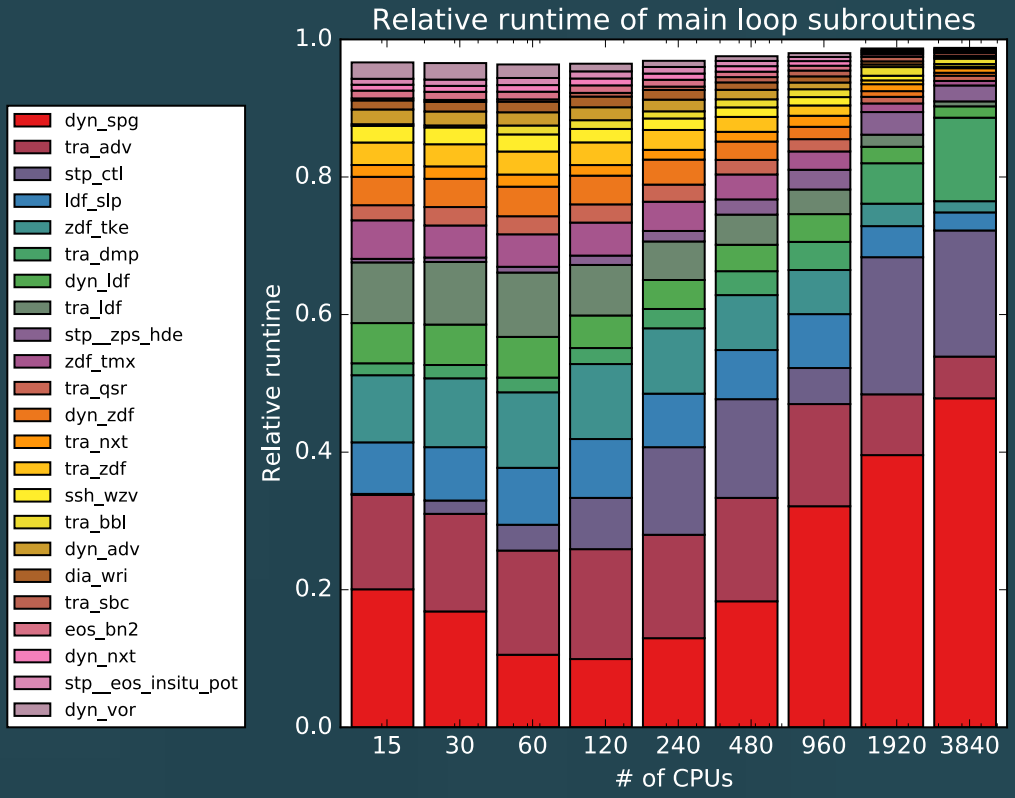
- Sea Ice
  - Scaling is comparable across models
  - Efficiency drop after 960 CPUs
- Initialization
  - All models show higher init times with CPU size
  - MPI initialization is a strong factor

SUBROUTINE ANALYSIS

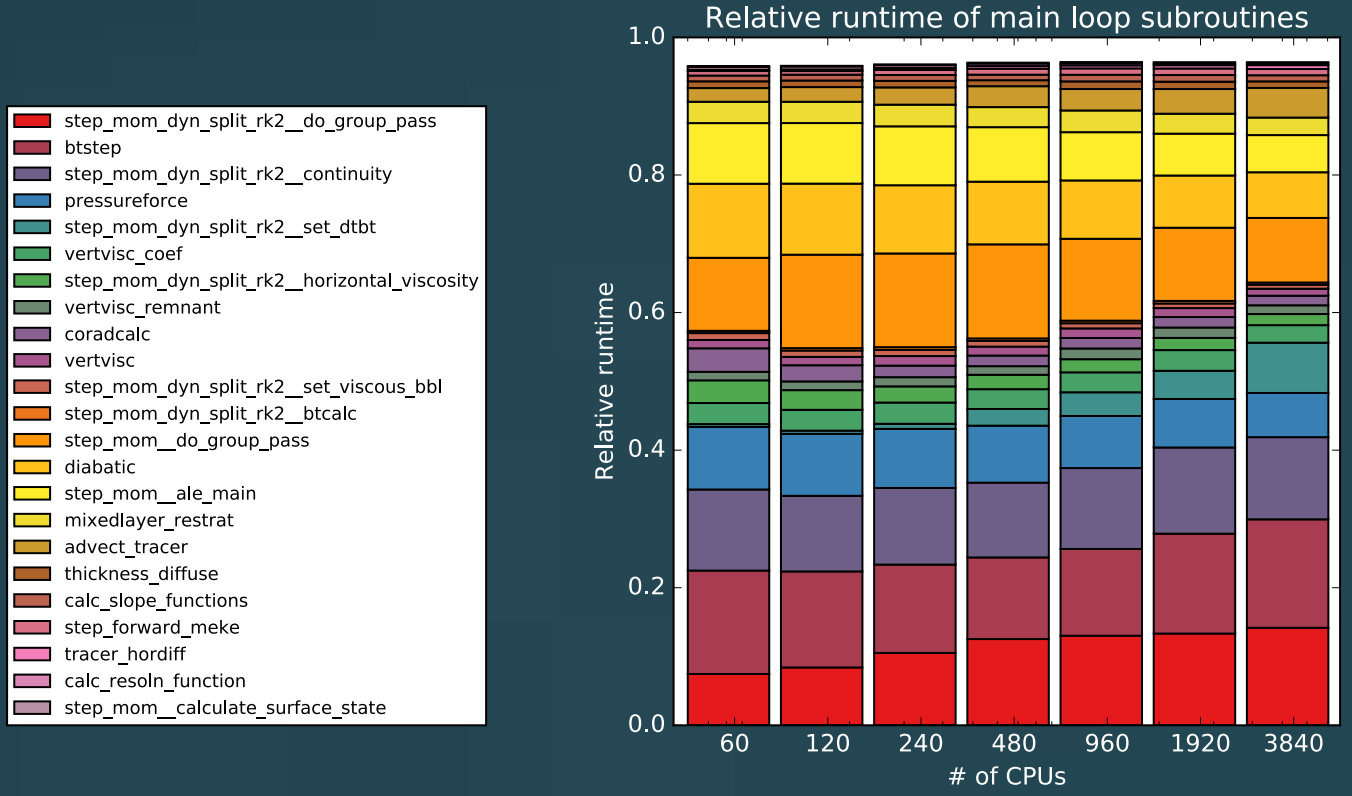
MOM 5 SUBROUTINES



NEMO SUBROUTINES



MOM 6 SUBROUTINES

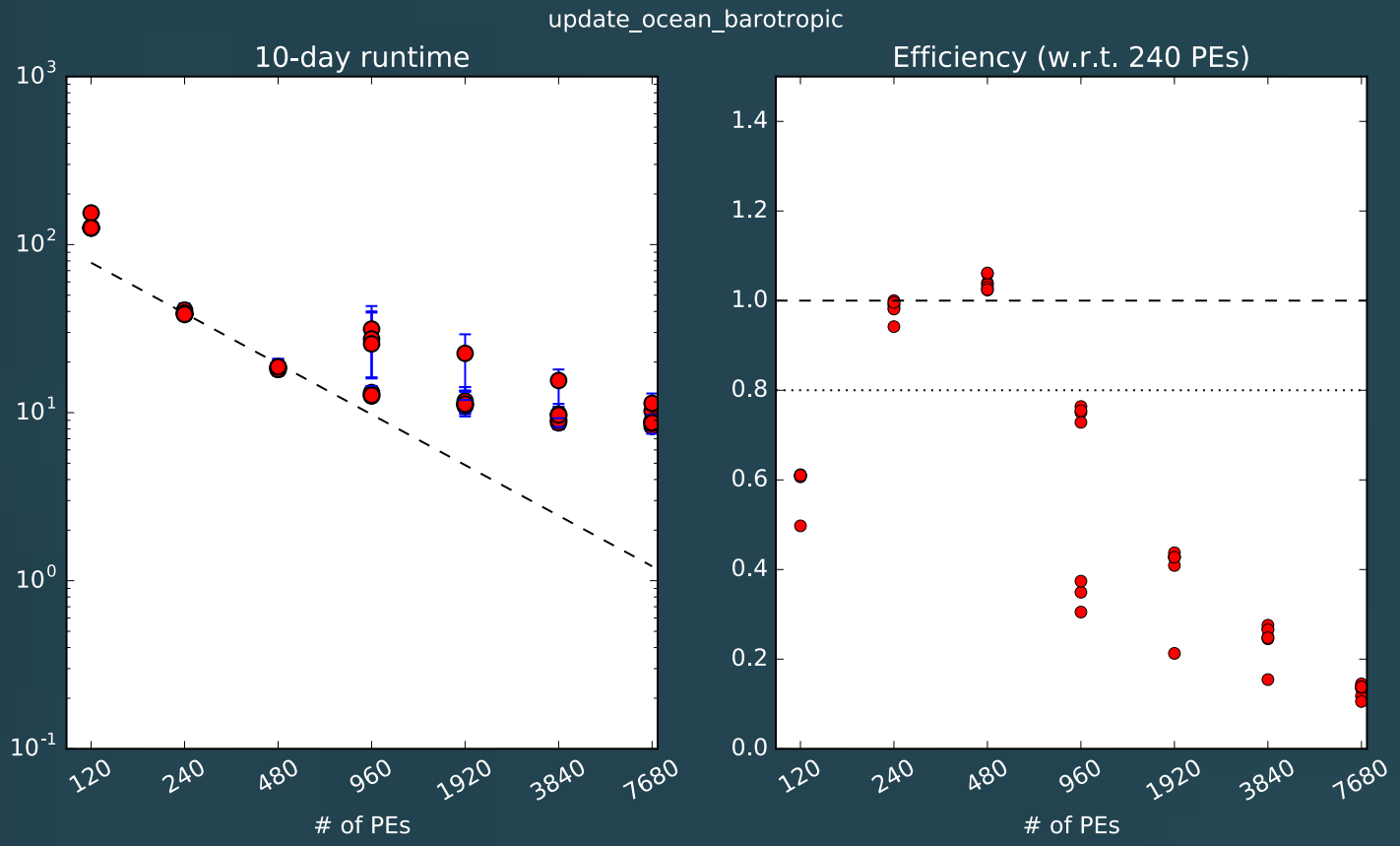


SUBROUTINE COMPARISON

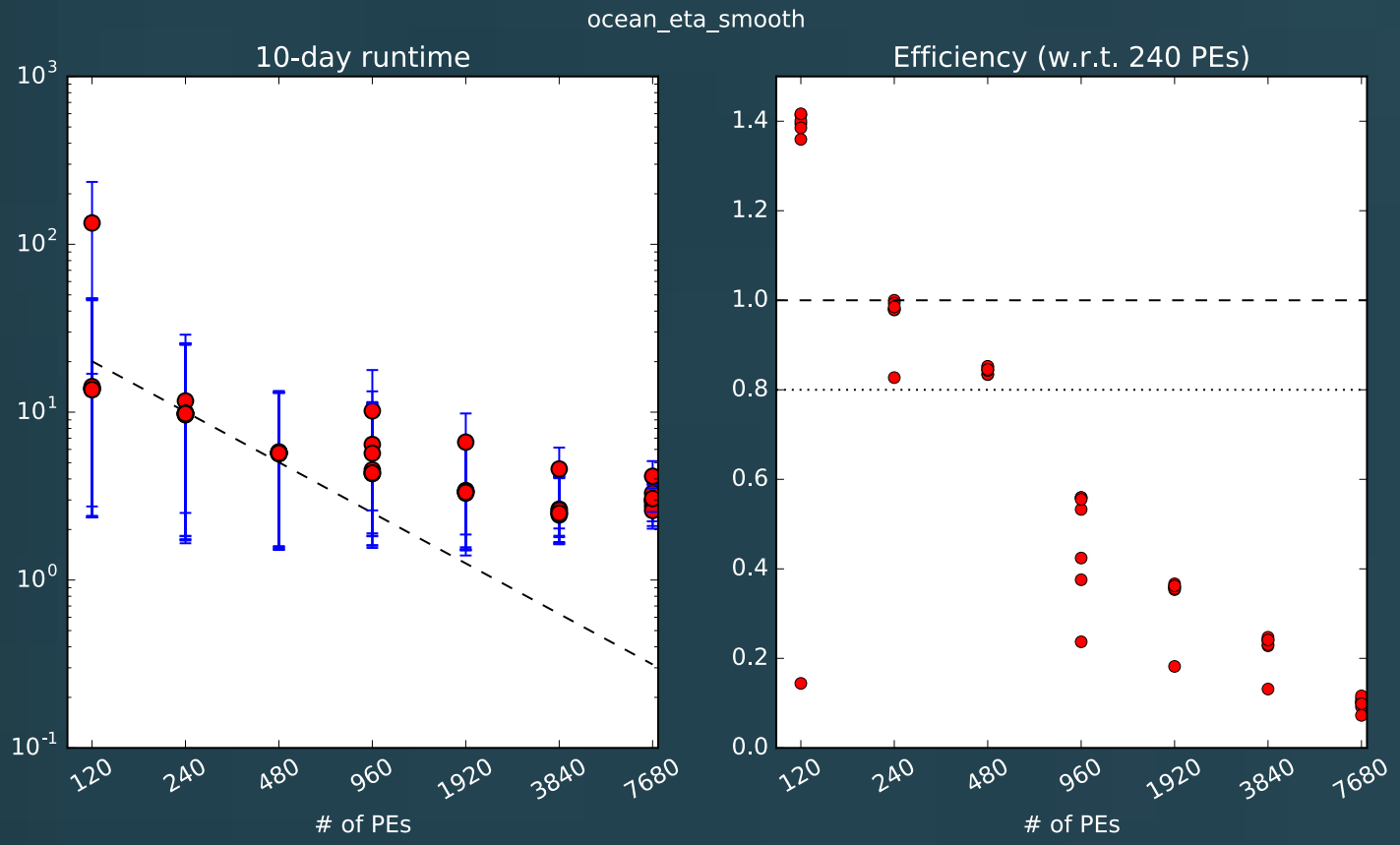
- Poor scaling of MOM 5, NEMO free surface dynamics
  - `update_ocean_barotropic`
  - `ocean_eta_smooth`
  - `dyn_spg`
- MOM 5, NEMO tracer advection is expensive
  - `update_ocean_tracer`
  - `tra_adv`
- MOM 6 communication is expensive
  - `do_group_pass`

FREE SURFACE DYNAMICS

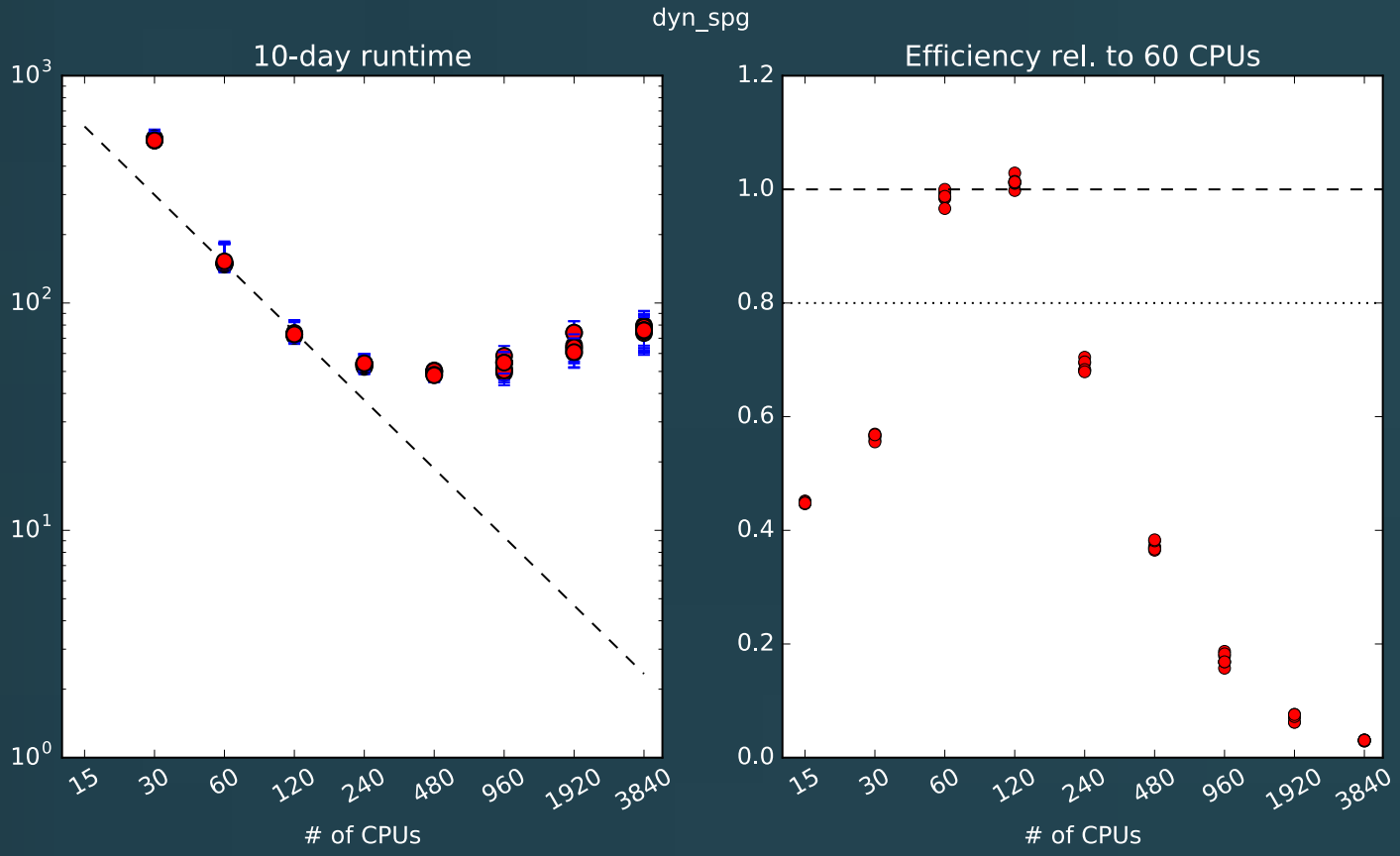
MOM 5: BAROTROPIC SOLVER



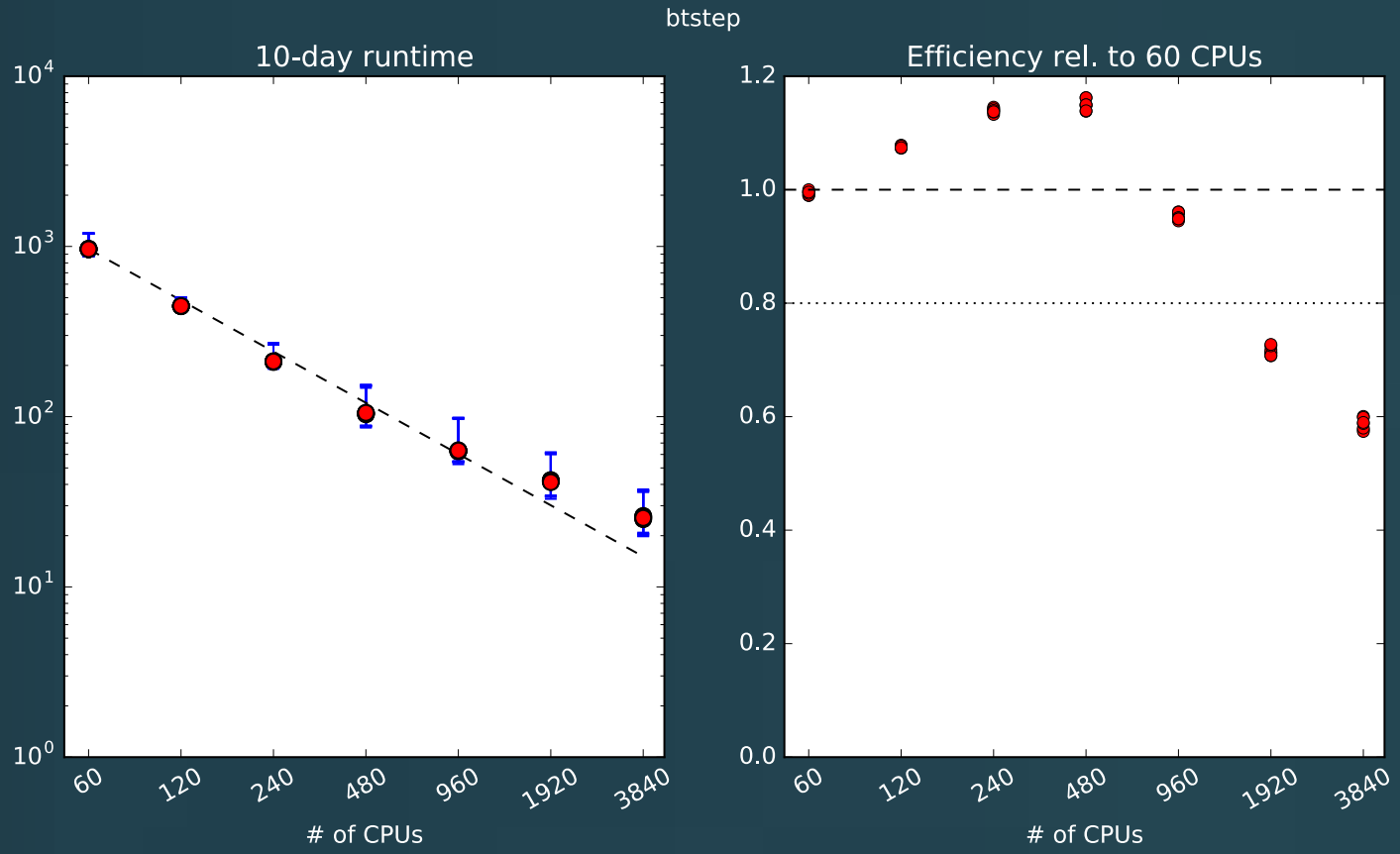
MOM 5: FREE SURFACE SMOOTHING



NEMO: SURFACE PRESSURE GRADIENT



MOM 6: BAROTROPIC TIMESTEP

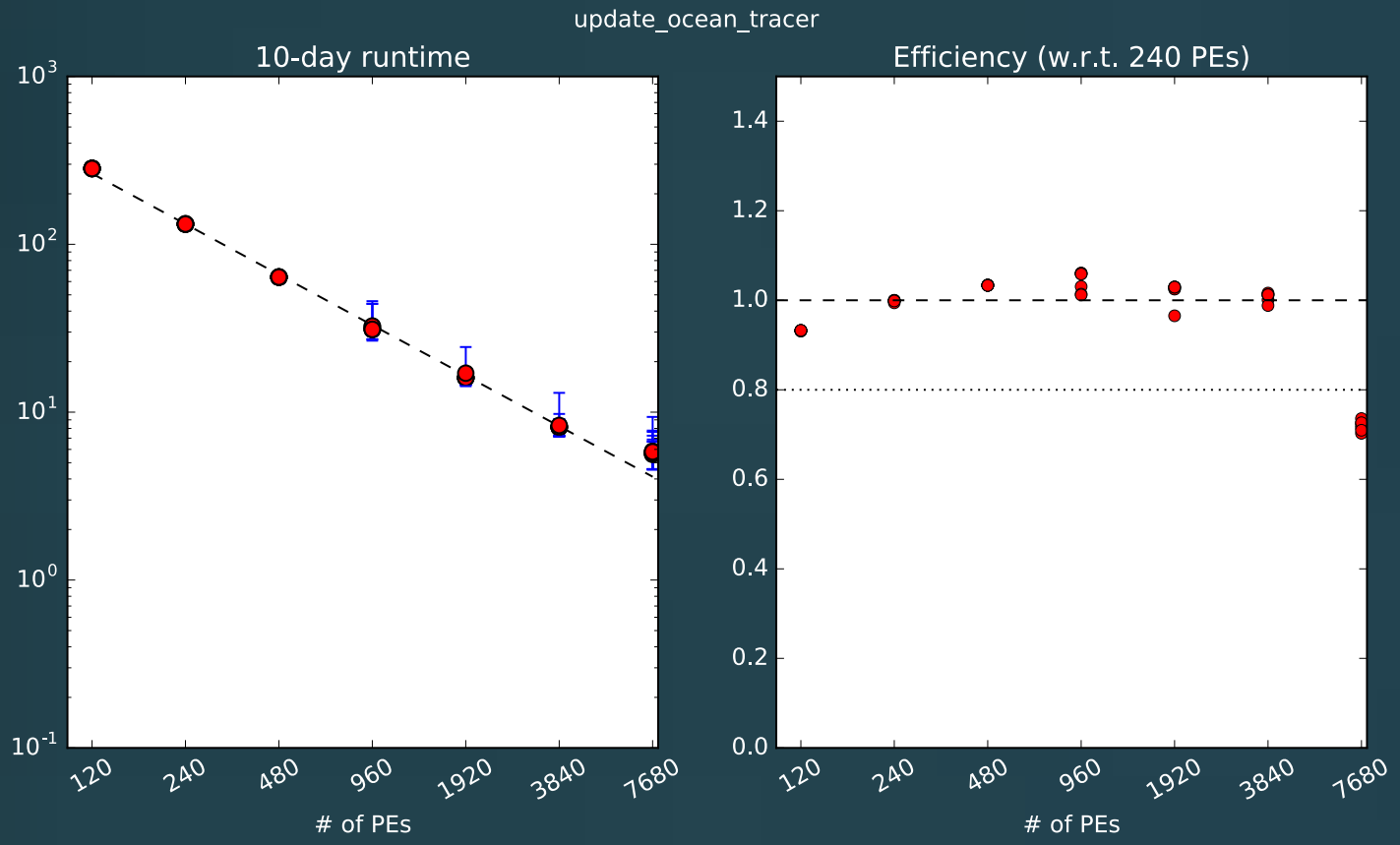


BAROTROPIC SCALING

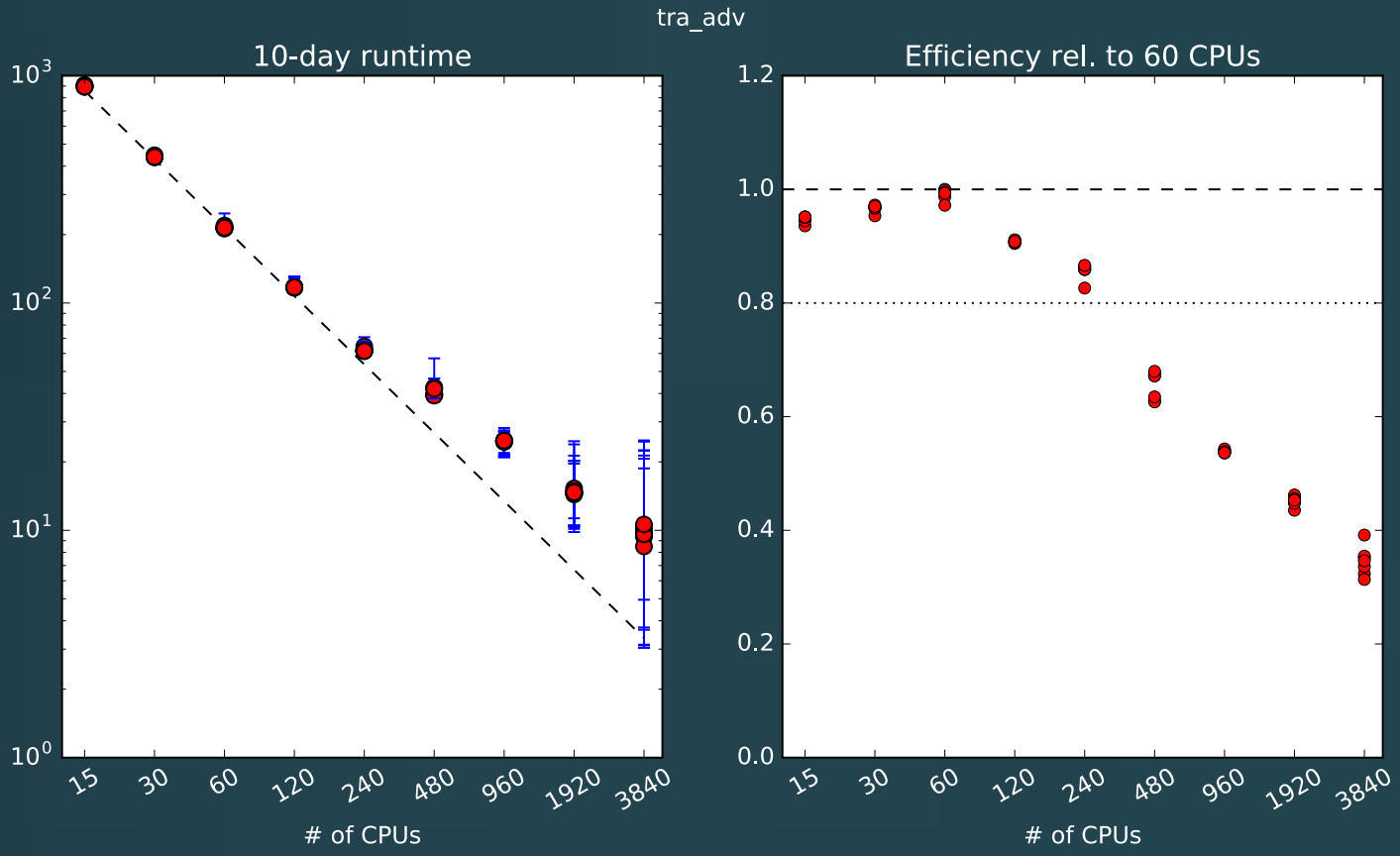
- MOM 5
  - Predictor Corrector Euler timestep
  - Scaling constrained to 960 cores
  - Further constrained by biharmonic filter (B-grid)
- NEMO
  - Elliptic conjugate gradient solver
  - Severe scaling constraint at 240 CPUs
- MOM 6
  - Predictor-corrector Euler timestep
  - Strong scaling within `btstep`
  - Communication unmeasured (`do_group_pass`)

TRACER ADVECTION

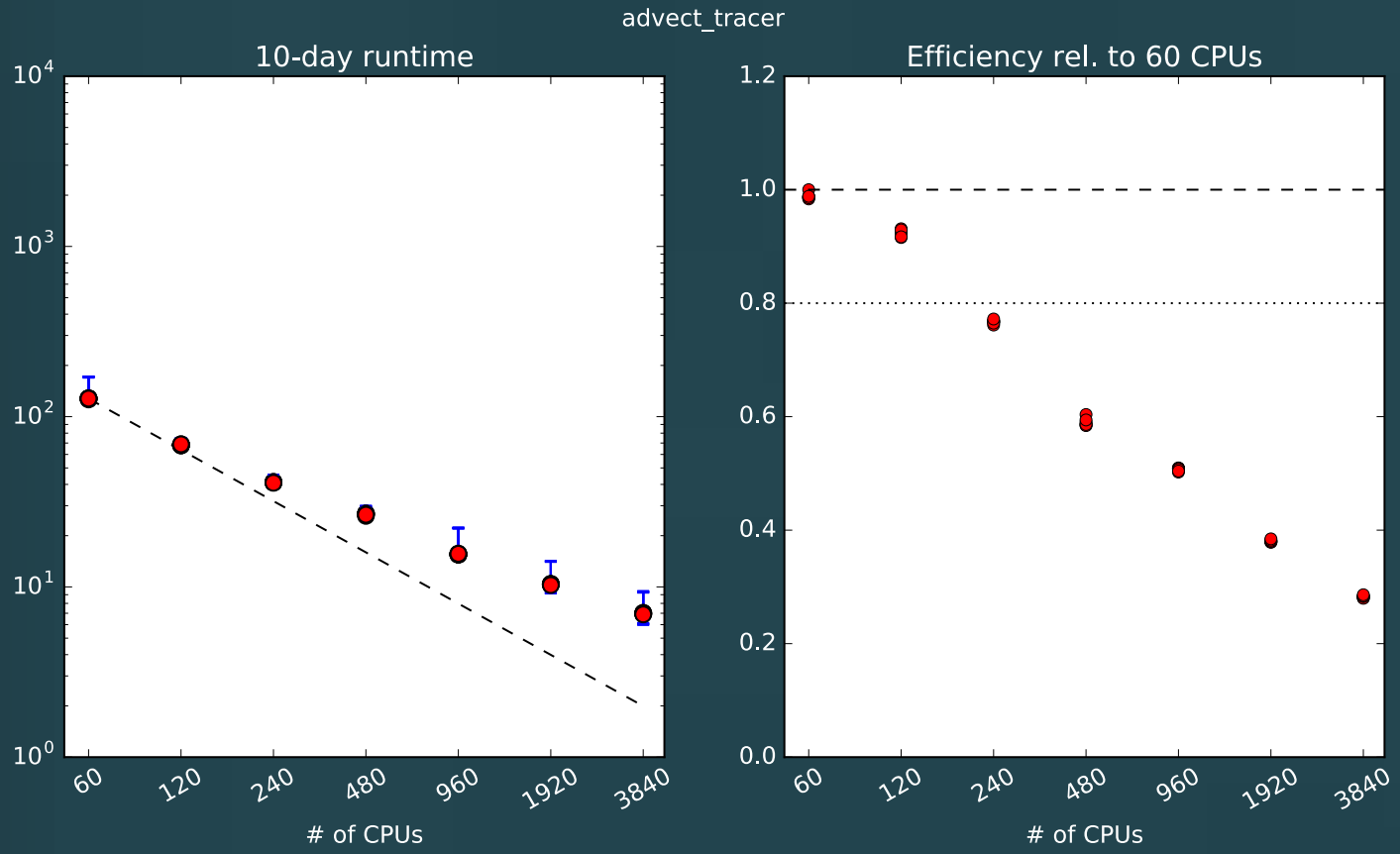
MOM 5: TRACER ADVECTION



NEMO: TRACER ADVECTION



MOM 6: TRACER ADVECTION

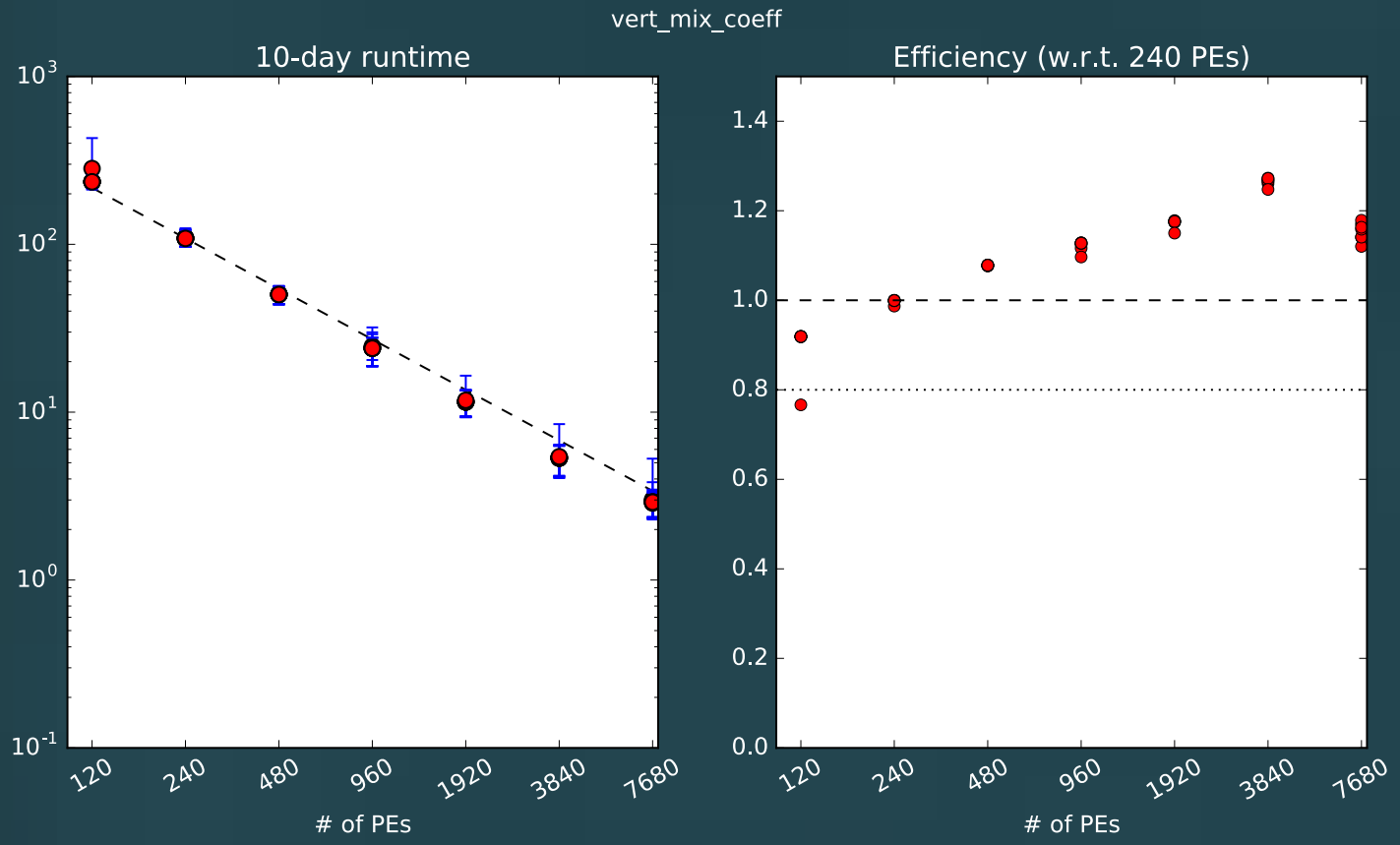


TRACER ADVECTION SCALING

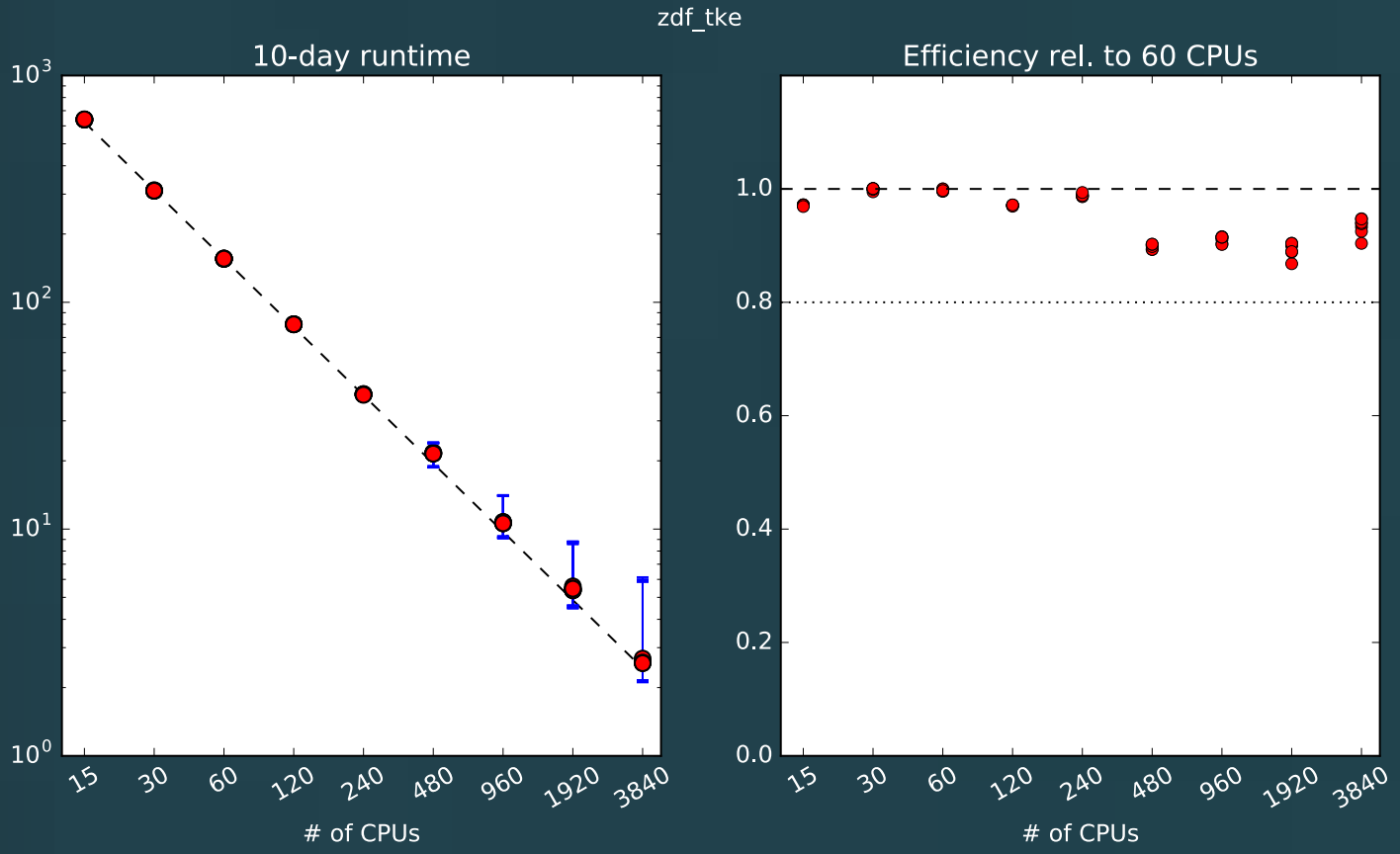
- MOM 5
  - MDPPM finite volume
  - Very strong scaling, at least 3840 CPUs
- NEMO
  - TVD finite volume
  - Inefficient scaling after 240 CPUs
- MOM 6
  - 3rd order Huynh PPM
  - Poor scaling, esp. after 240 CPUs
  - Communication unmeasured

VERTICAL PHYSICS

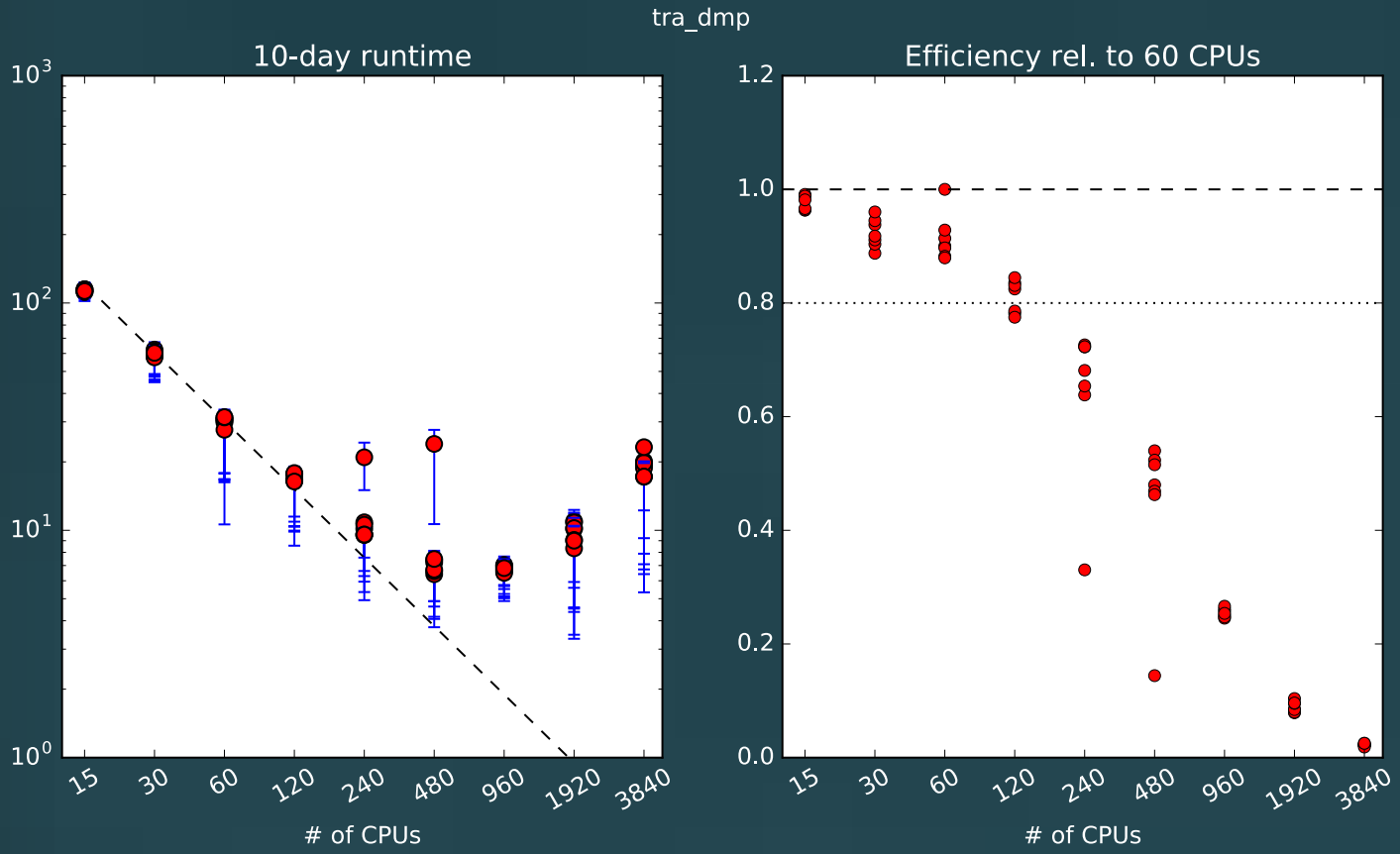
MOM 5: VERTICAL MIXING COEFFICIENT



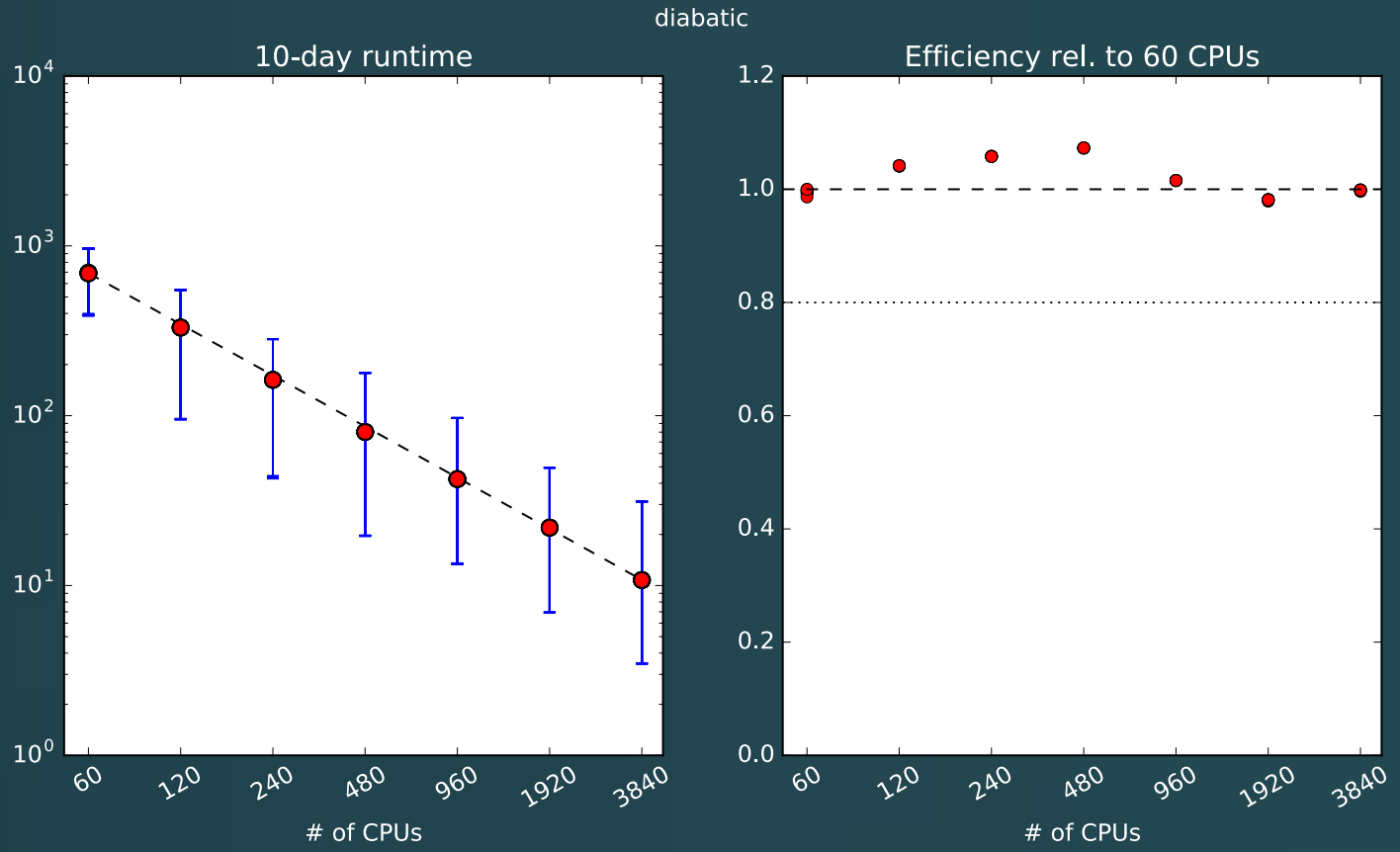
NEMO: VERTICAL MIXING COEFFICIENT (TKE)



NEMO: TRACER RELAXATION



MOM 6: DIABATIC PHYSICS



VERTICAL PHYSICS SUMMARY

- Vertical processes scale well due to horizontal tiling and low communication requirements
- NEMO tracer relaxation strongly constrains scalability, due to interpolation communication

MOM 5 VECTORISATION

Subroutine	FLOP/CPU	% vector
MAIN	2.04e+10	88.1
update_ocean_model	1.87e+10	91.0
update_ocean_tracer	8.40e+09	92.9
vert_mix_coeff	2.04e+09	80.9
ocean_explicit_accel_a	1.95e+09	92.8
update_ocean_barotropic	1.80e+09	94.5
ocean_eta_smooth	1.78e+07	87.5
update_ice_model_slow_dn	7.48e+08	83.2

NEMO 3.4 VECTORISATION

Subroutine	FLOP/CPU	% vector
stp	1.32e+10	55.0
dyn_spg	1.85e+09	96.0
tra_adv	1.31e+09	60.1
tra_dmp	4.18e+07	93.4
ldf_slp	7.47e+08	37.5
sbc	3.90e+09	6.2
sbc_ice_lim_2	3.80e+09	4.4
lim_dyn_2	3.61e+09	1.3

MOM 6 VECTORISATION

Subroutine	FLOP/CPU	% vector
MAIN	5.20e+10	65.3
update_ocean_model	5.00e+10	65.7
step_mom_dyn_split_rk2	3.80e+10	67.7
ale_main	9.85e+07	59.8
diabatic	3.54e+09	47.6
mixedlayer_restrat	1.83e+09	56.9
advect_tracer	1.17e+09	89.2
update_ice_model_slow_dn	1.08e+09	76.9

MOM 6 dynamics vectorisation

Subroutine	FLOPs	% vector
step_mom_dyn_split_rk2	3.80e+10	67.7
btstep	4.35e+09	63.8
do_group_pass	2.40e+03	0.0
continuity	1.51e+07	87.4
set_dtbt	5.07e+04	24.8
pressureforce	1.55e+10	79.5
set_viscous_bbl	2.23e+08	49.0

MOM 5 VECTORISATION IMPACT

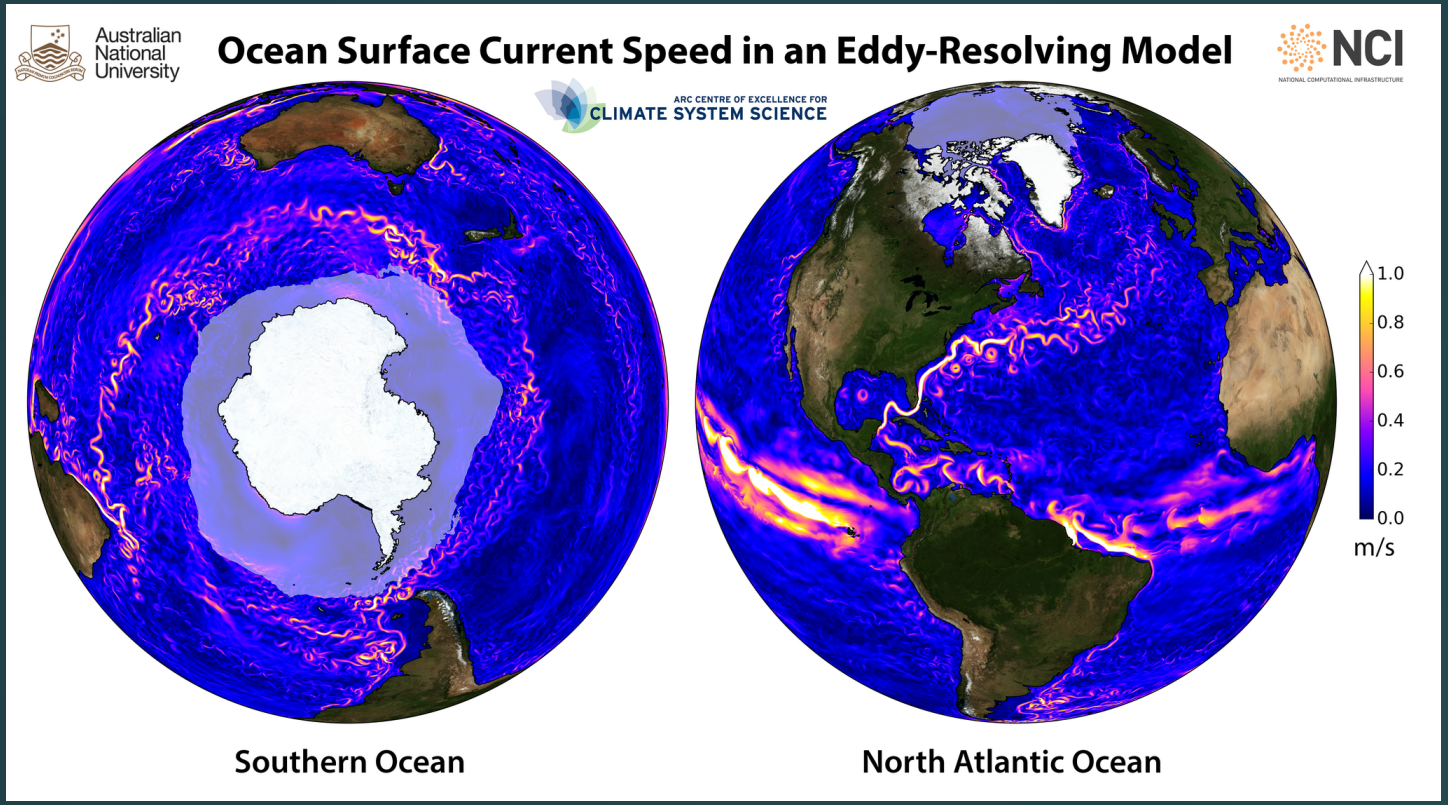
Section	Aligned AVX	Unaligned AVX	Serial
Ocean core	166.7 s	165.4 s	179.2 s
Tracer update	37.5 s	37.2 s	43.5 s
MDPPM	25.2 s	25.1 s	31.4 s
Vert. mix	24.9 s	24.9 s	26.5 s
Bih. frict.	16.1 s	16.0 s	16.7 s

Performance is likely bound by RAM speed

SUMMARY

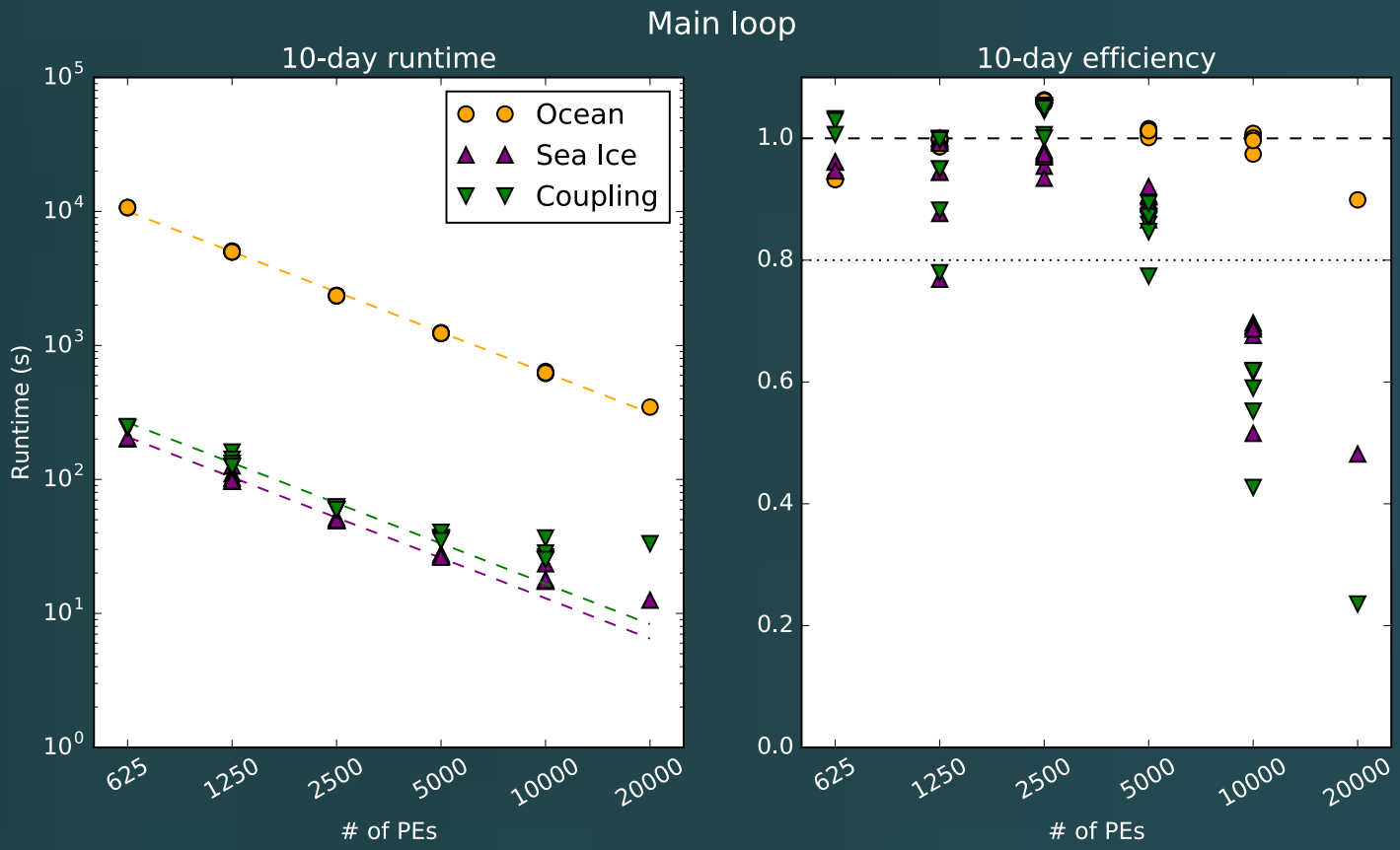
- MOM 5, OM 2.5
  - Strong scalability
  - High vectorisation (~90%), but memory-bounded
  - Significant barotropic scaling limit
- NEMO, ORCA 0.25°
  - Fastest serial computation
  - Lowest memory usage
  - Severe barotropic scaling constraint
- MOM 6, OM 4
  - Moderate efficiency loss after 960 CPUs
  - Moderate vectorisation (~65%)

HIGH RESOLUTION PERFORMANCE



(Stewart et al. 2016)

MOM 5, 0.1° scaling



MOM 5, 0.1° subroutines