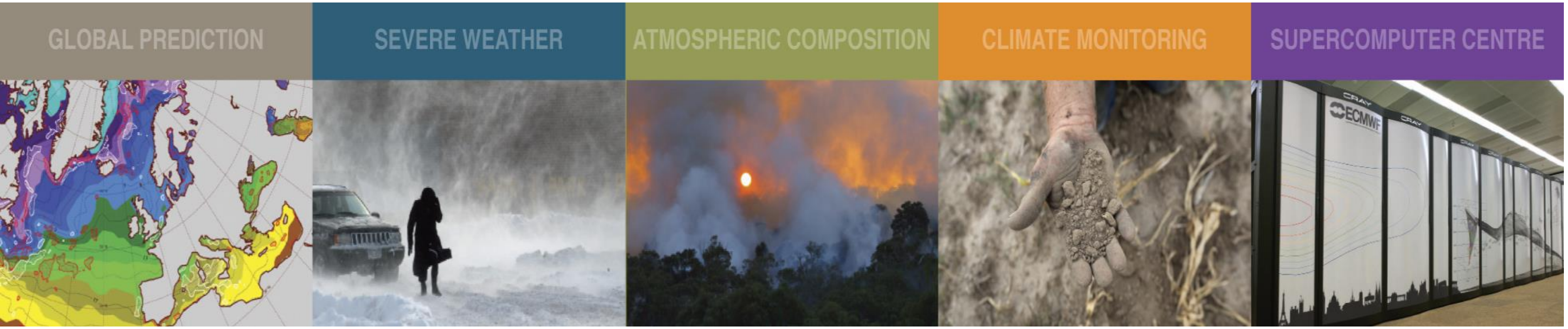
Atmosphere and ocean scalability and HPC readiness

Nils P. Wedi, European Centre for Medium-Range Weather Forecasts (ECMWF)



Many thanks to: ESCAPE partners, Kristian Mogensen, Phil Jones, Silvia Mocavero, Eric Maisonnave, Mike Bell, Alan Wallcraft, V. Balaji, Peter Bauer, Marshall Ward, Simon Marsland, Michel Rixen, …

Introduction

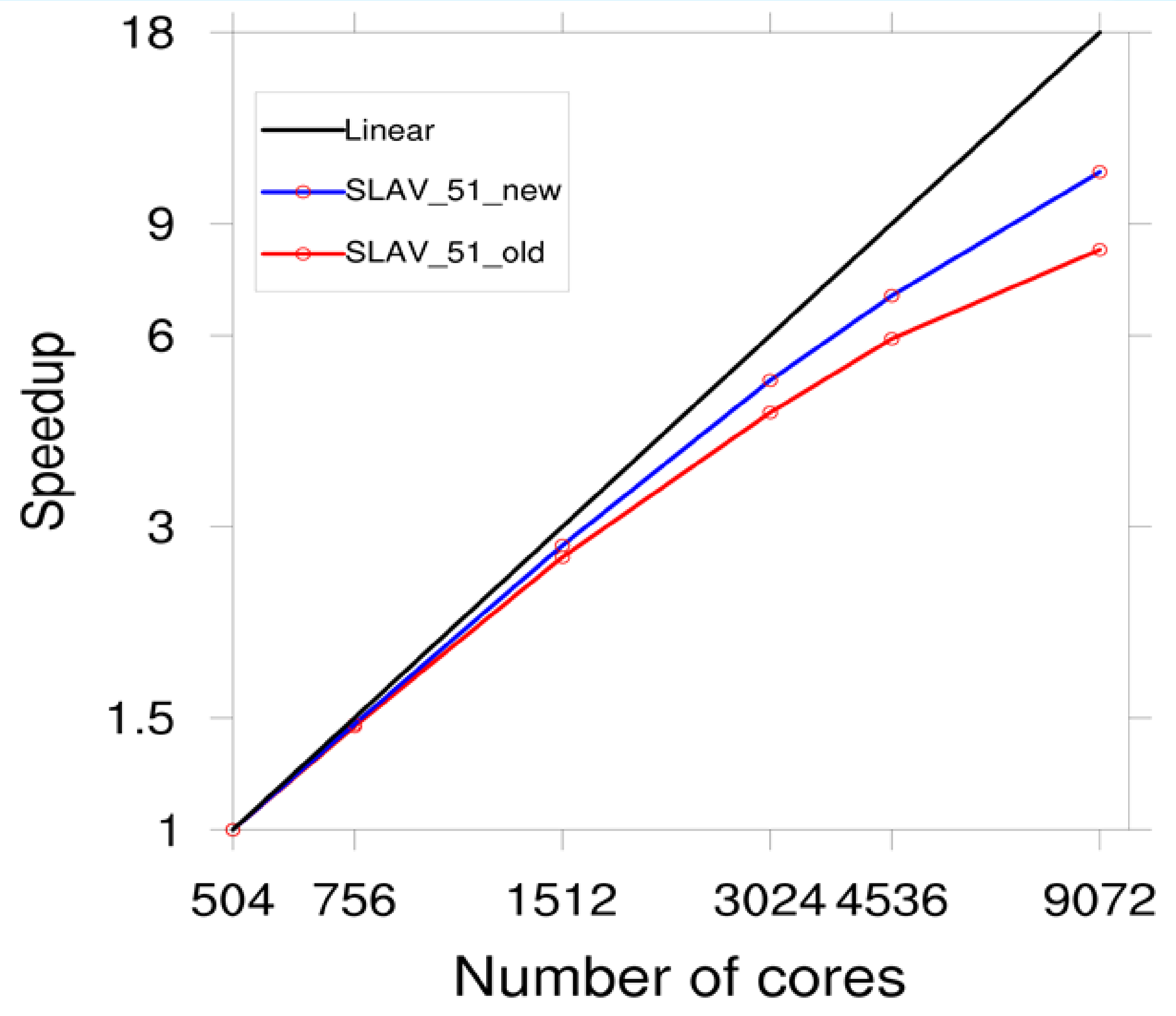
- Information provided by
  - CMC, DWD, ECMWF, Hydromet Russia, NOAA/NCEP, NRL, UK MetOffice
  - JMA, reduced complexity testing (shallow water, 2D, 3D)
  - Wits GCI (South Africa), advection module on cubed-sphere (dwarf) for scalability testing

Hydrometeorological Centre of Russia – SL-AV

- (4th-order) finite-difference, semi-implicit, semi-Lagrangian (on lat-lon, optionally reduced grid)
- Hybrid parallelization optimisations using MPI/OpenMP
- Parallel efficiency of semi-Lagrangian (and elliptic solver) reduced at high core counts

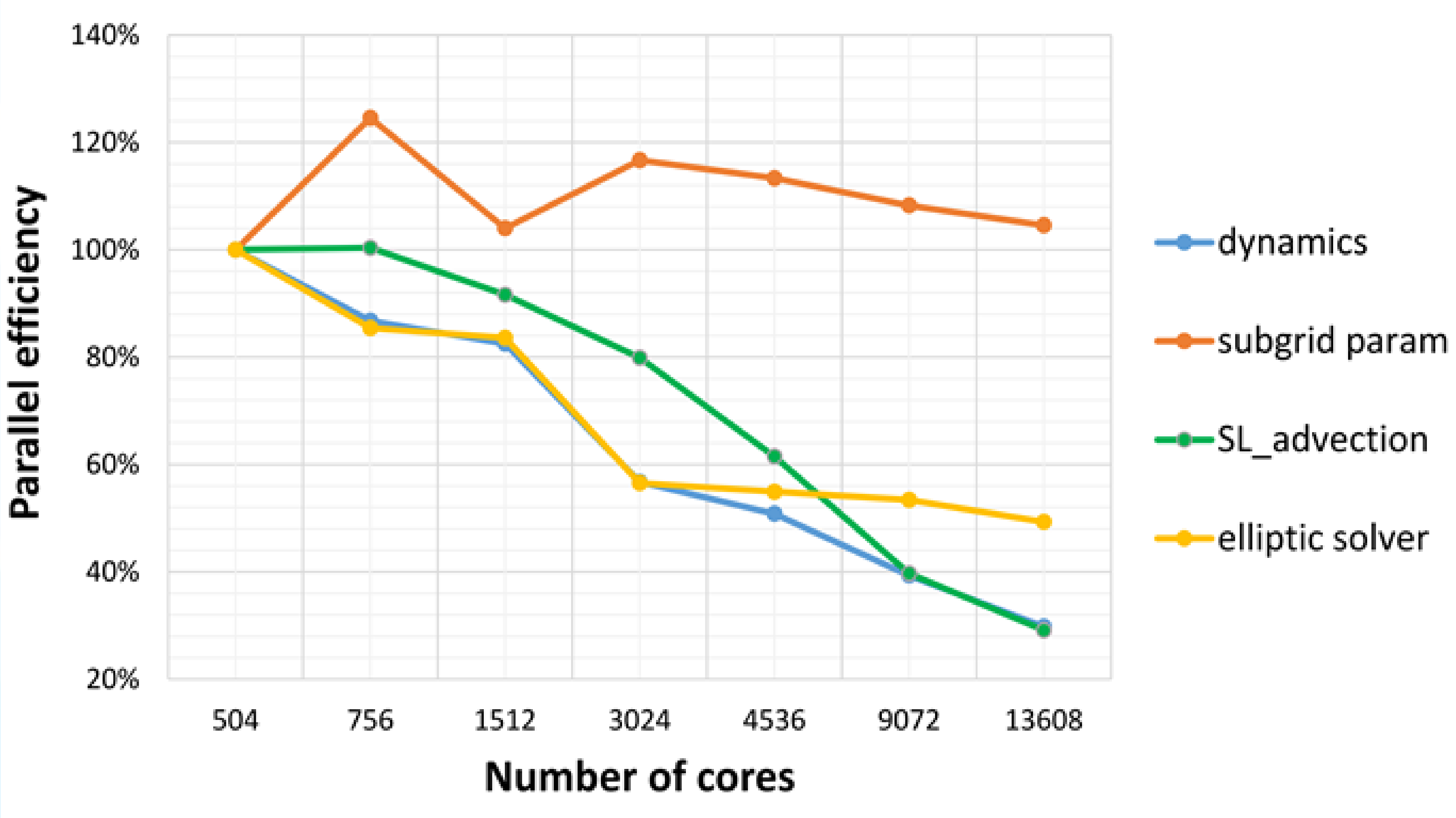
Tolstykh Mikhail, Goyman Gordey, Fadeev Rostislav,
Shashkin Vladimir, Loubov Sergei

Results of optimizations: parallel speedup w.r.t. 504 cores at Cray XC40

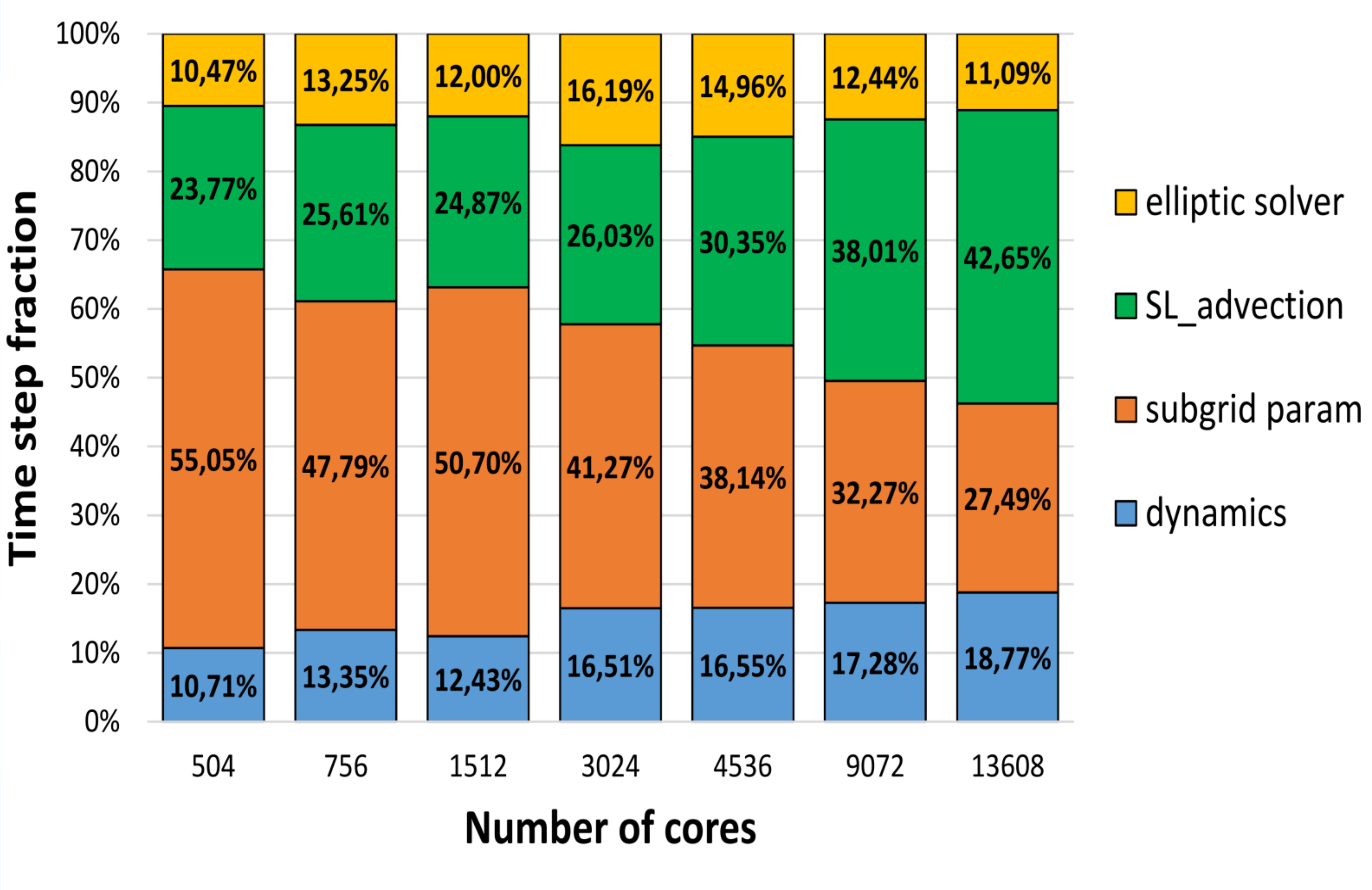


Grid: 3024x1513x51 (~10^8)

Parallel efficiency of different parts of the model code



Percentage of different dynamics part in elapsed time vs. # of cores



MPI optimizations

- The parallel implementation of the elliptic problem solvers requires data transpositions, i.e. global redistribution of data between processes.

- Code modifications allowed to reduce number of transpositions from 4 to 2 per time step

OpenMP code optimizations (1)

- Some parts of the model code used OpenMP to parallelize loops along the same direction as MPI decomposition (latitudinal index)

- Available parallelism is exhausted when N_mpi*N_openMP=N_lat-1

- These parts of code were modified to use OpenMP parallelization along additional index (longitude/wavenumber/vertical)

OpenMP code optimizations (2)

- Subgrid-scale parameterizations block is the most time-consuming part of the model.

- Computations in this block have only vertical index dependencies, so optimal arrays indices arrangement in terms of vectorization:

Array(horizontal dimension, vertical dimension)

UKMetOffice

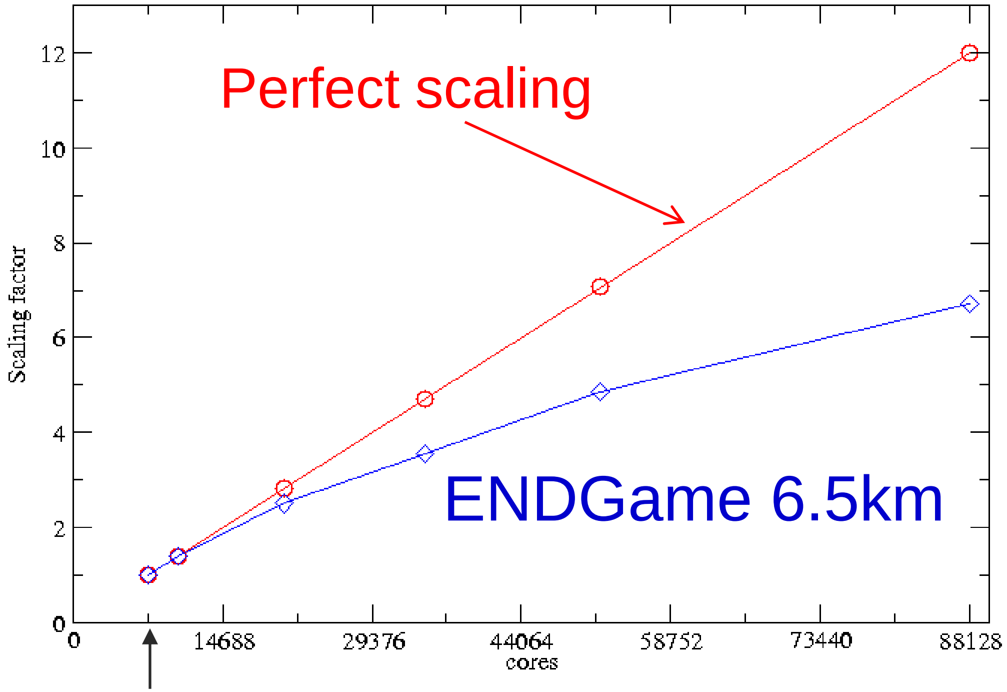
- Scalability limits of global ENDGAME due to poles in lat-lon grid
- Gung-Ho: New model in 2024, mixed FE/FV, uniform-mesh with cubed sphere
- Separation of concerns and code generation with PSyKal
- Leading HPC scalability efforts on NEMO
- Coupling via OASIS
- JEDI for ocean (and atmosphere?) data assimilation

Keith Williams, Andy Malcolm, Paul Selwood, Chris Maynard, Rupert Ford, … Gung-Ho/LFRic Team

Met Office

Scalability = critical element

T_{7344}/T_N N2048



7344 cores Cores

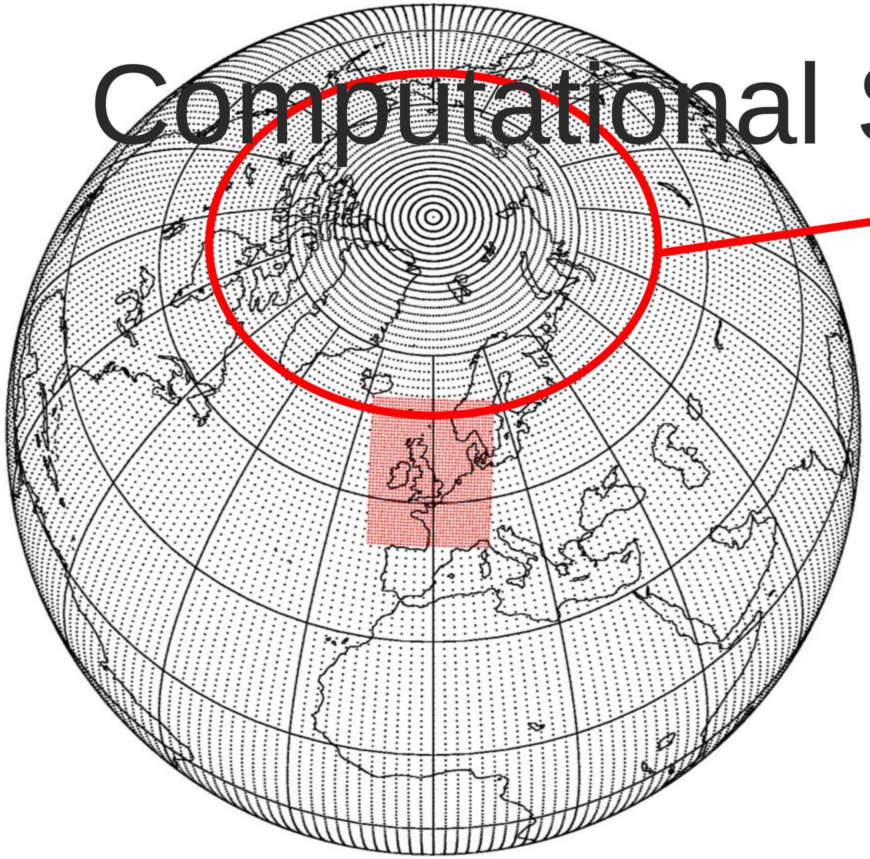
(1 node=36 cores)

Andy Malcolm, Paul Selwood

Met Office

What are the barriers?

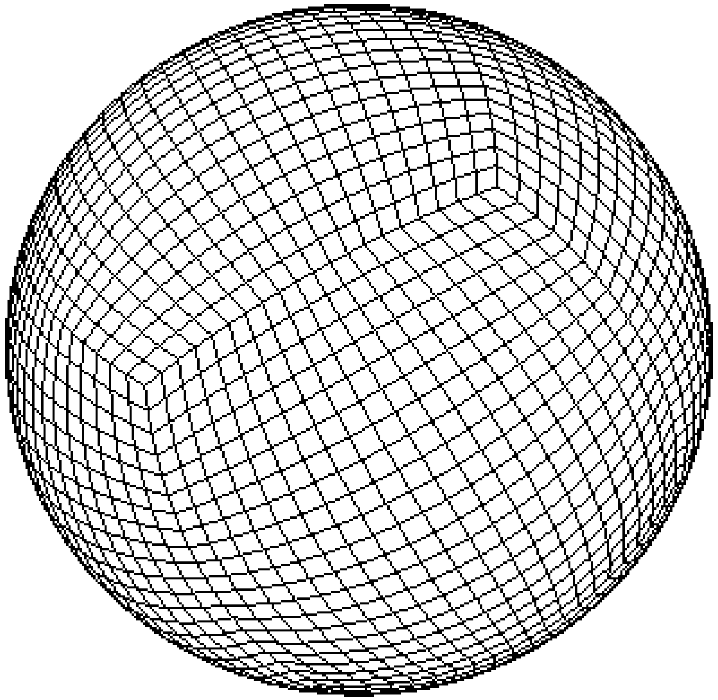
Computational Science



- At 10 km spacing near poles is 13 m
- At 1 km it reduces to 13 **cm**!
- **Communications** increase dramatically

⇒ Seek uniform mesh

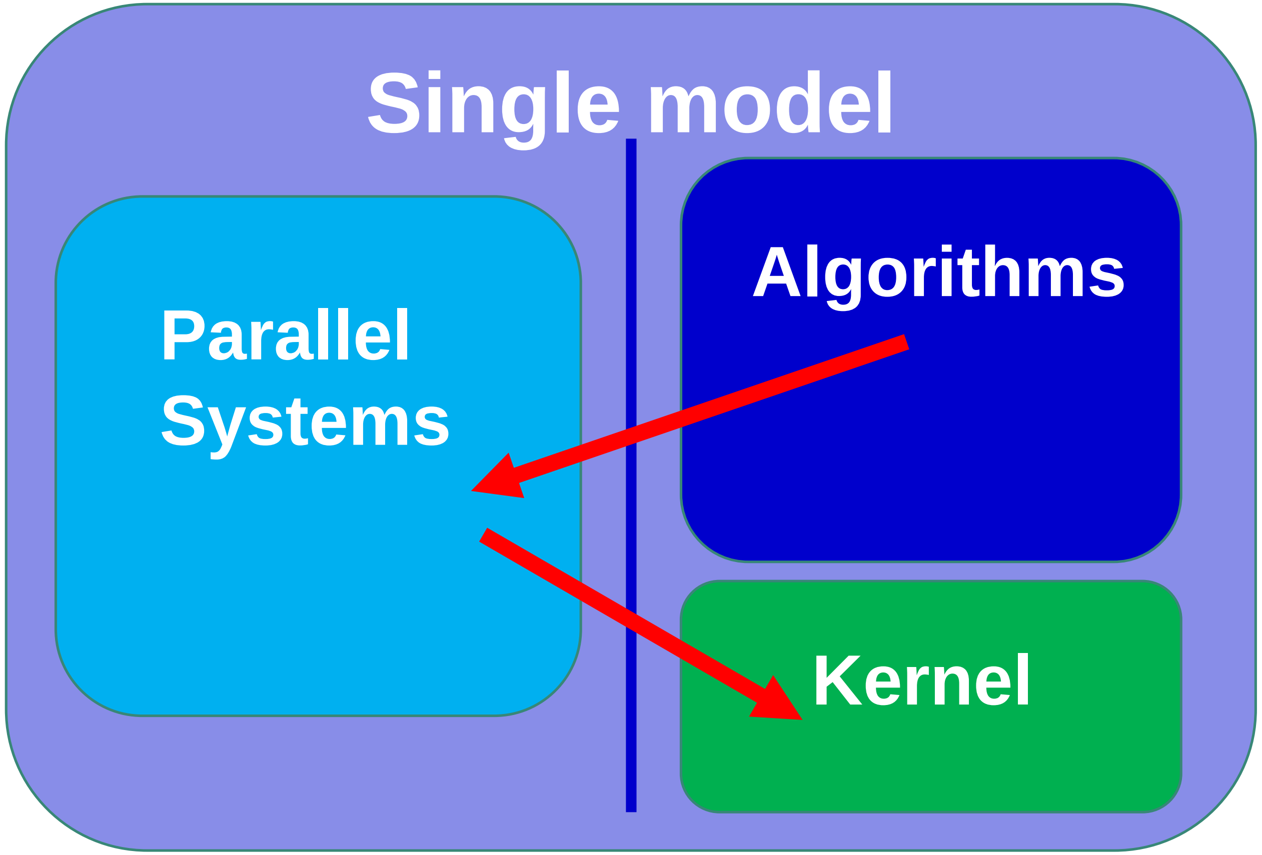
Natural Science



- Cube-sphere selected

(GungHo = Globally Uniform next generation Highly optimized)

The separation of concerns: PSyKAl



- Indirect addressing for horizontal
- Vertical loop inner most
- F2003
- Code auto generation

Parallel Systems Kernel Algorithm = PSyKAl

Met Office

Current code

(Coriolis terms $2\boldsymbol{\Omega} \times \boldsymbol{u}$)

All written by scientist

- Science code
- Horizontal looping
- Shared memory parallelism
- Distributed memory parallelism

```
! Compute work1 = (<vstar>^xi2)*f3_star - (<wstar>^eta)*f2_star
!$OMP PARALLEL DO DEFAULT(NONE) SCHEDULE(STATIC) PRIVATE(i,j,k)       &
!$OMP& SHARED(model_levels,pdims,vstar,f3_star,wstar,f2_star,work1)
  DO k=1, model_levels
    DO j=pdims%j_start, pdims%j_end
      DO i=pdims%i_start, pdims%i_end
        work1(i,j,k) = 0.5*(vstar(i,j,k)+vstar(i,j-1,k))*               &
                        f3_star(i,j,k) - 0.5*(wstar(i,j,k)+wstar(i,j,k-1))*     &
                        f2_star(i,j,k)
      END DO
    END DO
  END DO
!$OMP END PARALLEL DO

CALL swap_bounds(work1,                                                &
                 pdims_s%i_len - 2*pdims_s%halo_i,                     &
                 pdims_s%j_len - 2*pdims_s%halo_j,                     &
                 pdims_s%k_len,                                        &
                 pdims_s%halo_i, pdims_s%halo_j,                       &
                 fld_type_p,swap_field_is_vector, do_west_arg=.TRUE.)

!$OMP PARALLEL DO DEFAULT(NONE) SCHEDULE(STATIC) PRIVATE(i,j,k)       &
!$OMP& SHARED(model_levels,udims,u,r_u,beta_u_dt,work1,h1_xi1_u,dxi1_u)
  DO k=1, model_levels
    DO j=udims%j_start, udims%j_end
      DO i=udims%i_start, udims%i_end
        r_u(i,j,k) = u(i,j,k) + r_u(i,j,k) + beta_u_dt *               &
                     0.5*(work1(i,j,k)+work1(i+1,j,k)) /               &
                     ( h1_xi1_u(i,j,k)*dxi1_u(i) )
      END DO
    END DO
  END DO
!$OMP END PARALLEL DO
```

Met Office

The magic behind the curtain

PSy layer: autogenerated

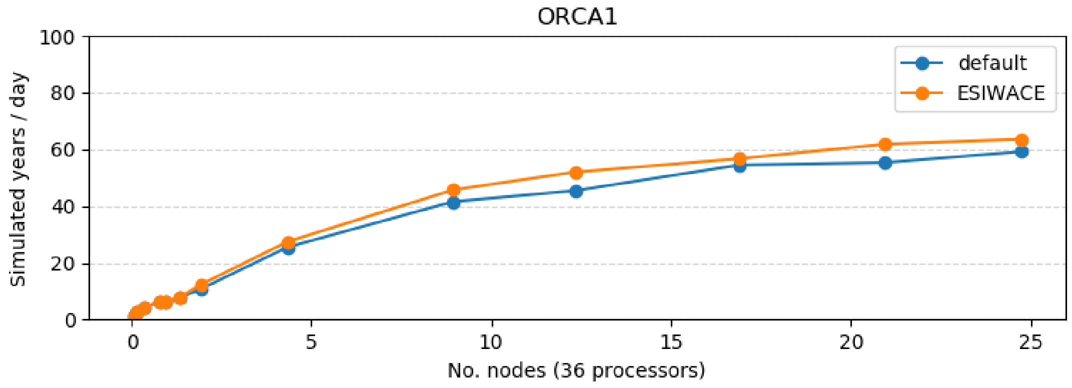
- Shared memory parallelism
- Distributed memory parallelism

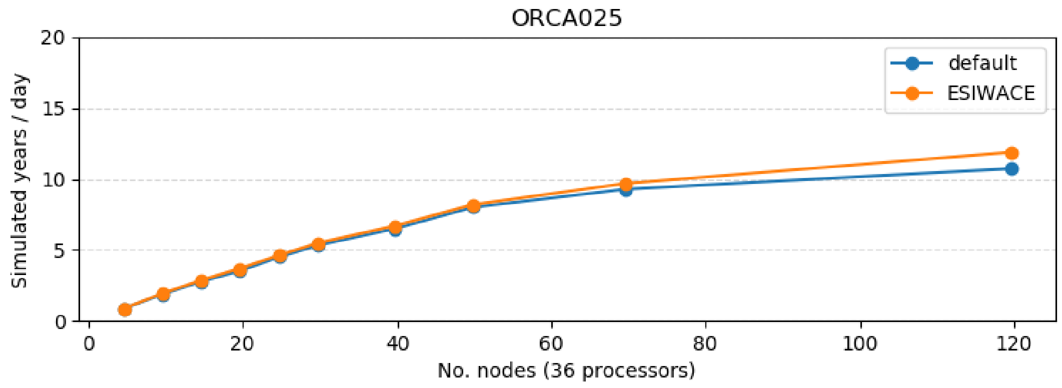
```
...
!
! Call kernels and communication routines
!
IF (rhs_tmp_proxy%is_dirty(depth=1)) CALL rhs_tmp_proxy%halo_exchange(depth=1)
IF (u_proxy%is_dirty(depth=1))       CALL u_proxy%halo_exchange(depth=1)
IF (chi_proxy(1)%is_dirty(depth=1))  CALL chi_proxy(1)%halo_exchange(depth=1)
IF (chi_proxy(2)%is_dirty(depth=1))  CALL chi_proxy(2)%halo_exchange(depth=1)
IF (chi_proxy(3)%is_dirty(depth=1))  CALL chi_proxy(3)%halo_exchange(depth=1)
!
! Look-up colour map
!
CALL rhs_tmp_proxy%vspace%get_colours(ncolour, ncp_colour, cmap)
!
DO colour=1,ncolour
  !$omp parallel default(shared), private(cell)
  !$omp do schedule(static)
  DO cell=1,ncp_colour(colour)
    !
    CALL rotation_code(nlayers, rhs_tmp_proxy%data, u_proxy%data, &
                       chi_proxy(1)%data, chi_proxy(2)%data, chi_proxy(3)%data, &
                       ndf_w2, undf_w2, map_w2(:,cmap(colour, cell)), basis_w2, &
                       ndf_any_space_9_chi, undf_any_space_9_chi, &
                       map_any_space_9_chi(:,cmap(colour, cell)), basis_any_space_9_chi, &
                       diff_basis_any_space_9_chi, nqp_h, nqp_v, wh, wv)
  END DO
  !$omp end do
  !$omp end parallel
END DO
!
! Set halos dirty for fields modified in the above loop
!
CALL rhs_tmp_proxy%set_dirty()
!
...
```

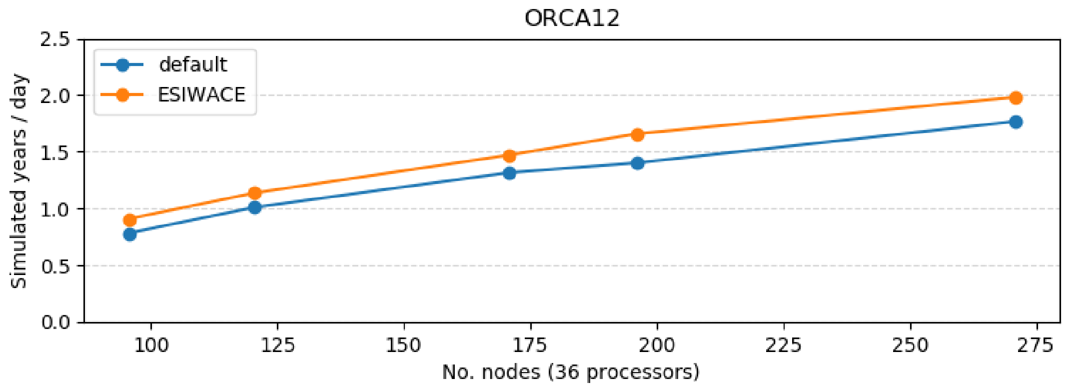
The Future: Ocean & sea ice

Status of NEMO

- NEMO currently runs fairly efficiently on a large number of nodes (>250 for ORCA12)
- Several improvements to time-step algorithm, mixed precision, tiling & OpenMP expected in next 2 years
- A reasonably efficient port to GPUs is achievable (outcome of hackathon with Nvidia)
- Science and Technology facilities council (STFC) has a prototype parser which translates NEMO into code which Psyclone can operate on







Met Office

Status of other components

- SI3 (sea-ice model)
  - is quite efficient on current CPU machines
  - Expected to port to GPUs similarly to NEMO
  - To avoid compromises, the facility to couple through OASIS is desirable

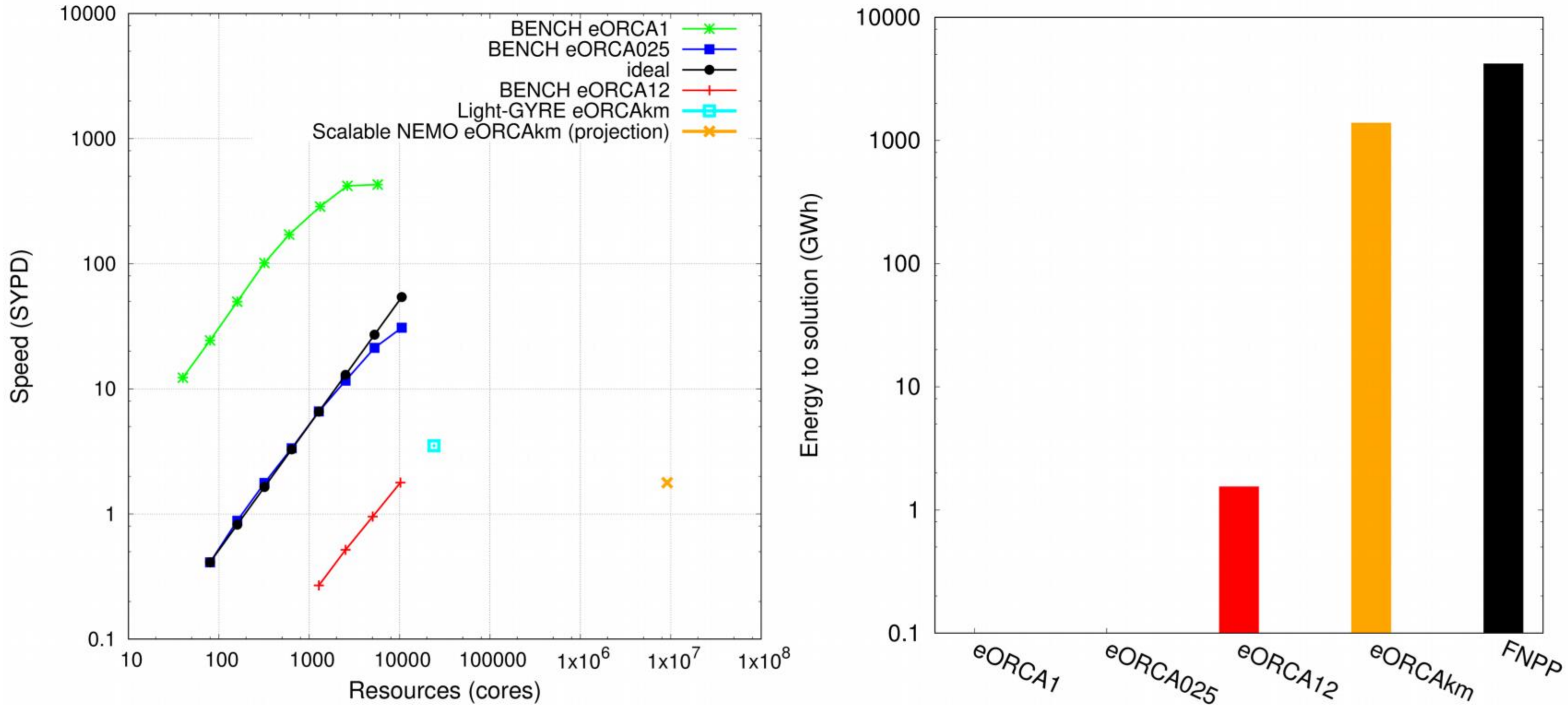
- NEMOVAR (data assimilation for NEMO)
  - Diffusion operator for large correlation scales will use coarsened grid
  - Several other optimisations are in progress
  - Workshop on broadening collaborations; consideration of JEDI

- Biogeochemical models (MEDUSA and ERSEM)
  - Grid coarsening is the first priority
  - Port of ERSEM to GPUs needs to be considered

NEMO HPC working group

- Regular telecons, Mike Bell (UKMO) NEMO HPCchair
- ECMWF participation
- H2020 funded efforts: IMMERSE, ESiWACE, ESCAPE-2, IS-ENES
- Main areas:
  - Improve Inter-node communication and latency
  - 2-time-level timestepping scheme
  - Memory layout to facilitate cache efficiency and threads via tiling approach
  - Mixed precision
  - Experiment with DSL toolchain using NEMO dwarfs
  - Design test problems for ocean dwarf or ocean model intercomparisons
  - Advantages/Disadvantages of unstructured ocean models
  - Review scalability bottleneck of barotropic mode coupling

NEMO tests



Scalability (left) and energy consumption (right) of several NEMO 3.6 model configurations, measured on `beaufix2` Météo-France supercomputer, Intel Broadwell processors. BENCH and GYRE configurations (I) have 75 vertical levels, (ii) exclude realistic bathymetry (no effect on performance), sea-ice, bio-geo-chemistry and output but (iii) include TOP tracers, appropriate physics at each resolution and polar grid folding (BENCH only). Horizontal resolution varies from 1 degree (eORCA1) to 1 km (eORCAkm). Scalable NEMO eORCAkm performances are extrapolated from measurements of a simplified GYRE km scale configuration, assuming a perfect scalability until a 10 million MPI subdomain decomposition. Right figure compares the total production of one reactor of a power plant (FNPP) similar to Fessenheim, France and the energy consumption of a 1,000 year long simulation led with the four NEMO configurations (eORCAkm: projection) at maximum scalability, approximated as suggested in Balaji et al. 2017, assuming `beaufix2` consumption E = 2.15e12 J/month and total capacity A = 5.2e7 CH/month

NRL - NEPTUNE

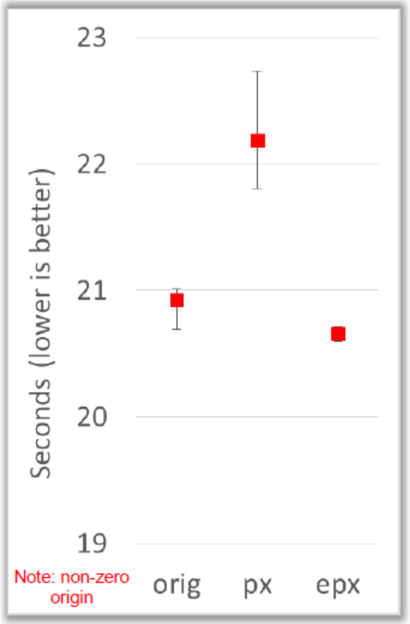
- Goal of <5 km global and <1 km regional NWP by ~2025
- High-order continuous Galerkin, cubed sphere
- Good locality, highly scalable, constant-width halo
- Performance analysis and testing with kernels
- Performance portable restructuring of data layout and loop nesting, kernels and full code
- Challenges: Physics coupling to new dynamical core, grey zone of convection, multi-scale data assimilation (JEDI), exascale computing, coupling (ESMF)

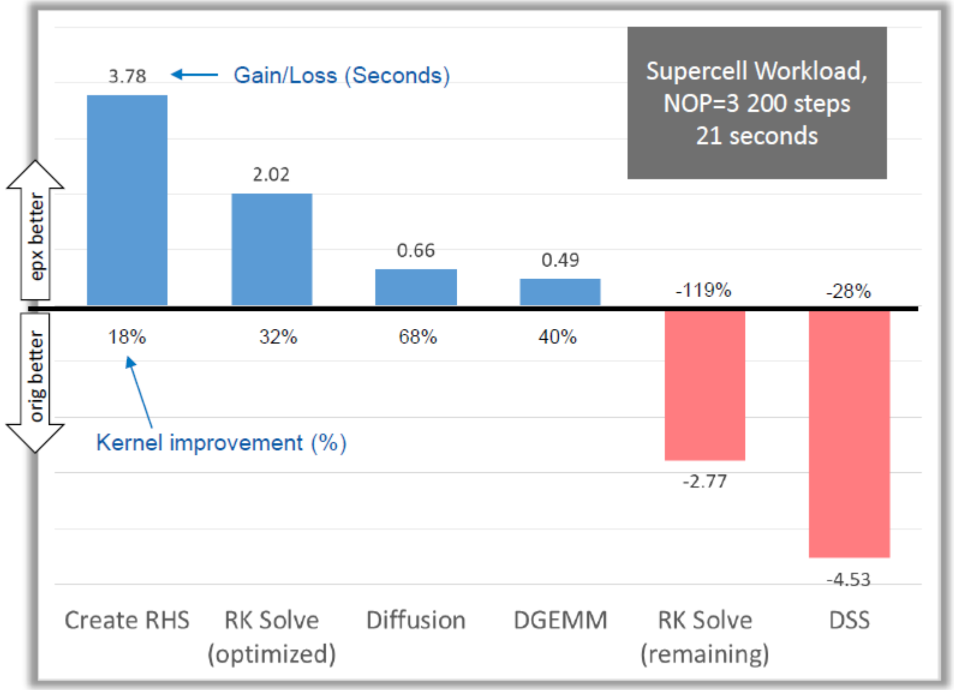
John Michalakes, Alex Reinecke, Jim Doyle, et al

NRL – substantial individual kernel improvements

Full Dynamics

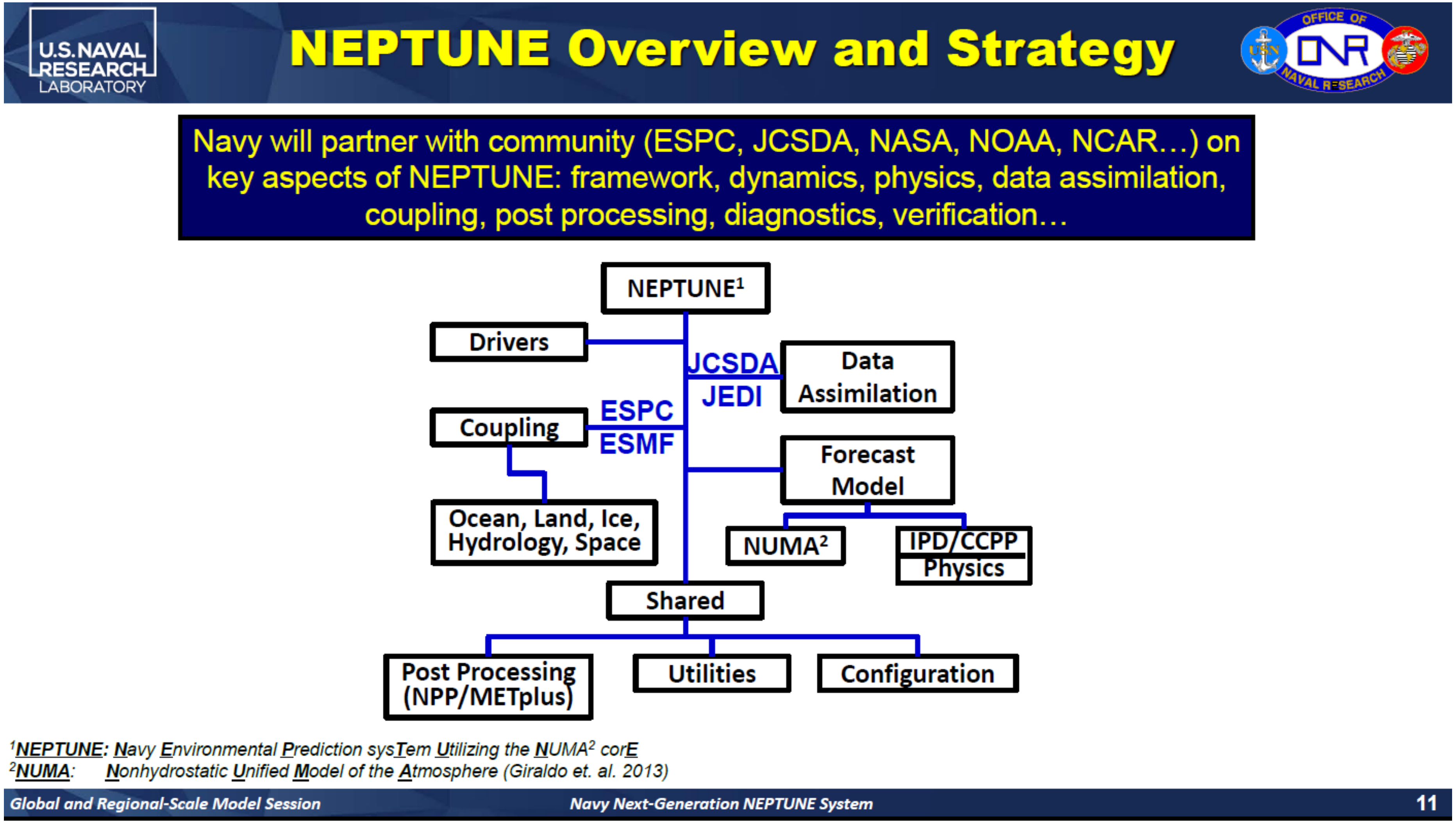
- Applying PX and EPX optimizations to full dynamics
- Testing on 36 cores, Skylake (Xeon Gold 6140, 2.3 GHz)
- GPU next





8

NRL – community coupling strategy



CMC - GEM

- Performance and portability
  - Portability (no vendor lock-in) is of paramount importance and we test our code with gfortran on different platforms and GPUs (P9, x86_64, V100, Telsa card, GeForce, Quadro). The exception to this rule (avoiding CUDA) will be made at very low level for specific optimizations within limited pieces of computationally expensive code.
  - Directive based approaches directly in the code (OpenMP 4.5+, OpenACC, the latter more mature)
  - A version of GMRES modified to run on GPUs had a 2.5x speedup incl all overheads, unclear how this translates when put back in full model.
- Scaling: anticipate a required speed-up of 2-4 in the next 5 years
  - Transposes in the direct solver vs iterative solver at high core counts. Iterative solvers in GEM don't outperform the direct solver in high-rank problems. More efforts could/should be put into preconditioning for the iterative solvers to enhance scalability at high rank.
  - high-resolution (2-4 km grid spacing) regional ensemble prediction system, with a significant increase in the use of available cores. At the global scale, we would like to reach 9-10 km grid spacing.
- Data assimilation: no immediate performance concerns
  - recent incremental refactoring of the DA system has led to a code base that is much more modular (easier to optimize specific problem areas) and testable (implementation of numerous unit tests that will facilitate any future optimization efforts.)

Ron McTaggart-Cowan

DWD - ICON/COSMO

- Performance and portability
  - COSMO model using a DSL for the dycore, openACC for the rest.
  - ICON: ongoing developments, currently based on openACC only; dycore, tracer transport and some physics schemes shared with the COSMO model are already ported.
  - Incremental developments of ICON; no resolution increase is planned for the global deterministic system because preparatory tests indicated a disappointing cost-benefit ratio. Investments into improving the physics package have been found to be much more promising.
  - Coupling to waves and atmospheric composition, ocean not planned for NWP
  - MeteoSwiss to implement a DSL-based (GridTools) version of the dycore, but at DWD, we are still somewhat reluctant about this approach due to our experience with the COSMO model, where the DSL-based dycore (using STELLA) implementation is much too disruptive w.r.t. further scientific development. We are working on designing a more practicable interface for the scientific developers, but an open issue is how to proceed when algorithmic motifs are needed to test new ideas that are not covered by the DSL. We think that a Fortran backend is a necessary requirement in this context, also to facilitate debugging and to have a fallback if no backend is available for a certain architecture.
- Concern: next DWD machine is a NEC Aurora vector computer, which will likely entail many optimization conflicts with GPUs.

Günther Zängl

DWD - ICON/COSMO

- Speed-up of factor 6 expected in 2022 compared to today
  - Higher resolution for the global EPS (40 -> 26 km), accompanied by more ensemble members for the assimilation cycle (LETKF)
  - Higher resolution for the convection-permitting EPS (incl. deterministic run), probably accomplished by inserting a 1-km nest into the 2-km domain; forecast lead time extended to 48 h (currently 27 h) 8x/day
  - Additional rapid-update convection-permitting EPS with 12h-forecasts every hour, probably using the same domain configuration
- Scaling capabilities of ICON are already sufficient for the expected NWP requirements in the next 10-15 years (climate applications generally rely more strongly on strong scaling than NWP)
  - ICON in a configuration with 52 km L90 (184320 x 90 grid points) and standard NWP physics schemes, achieved reasonable scaling up to 640 36-core Broadwell nodes reaching 133 SYPD.
- Concerned about the scaling capabilities (both w.r.t. memory and compute time) of our data assimilation code, which are already hitting limits on our current HPC system, substantial investments expected in the next years, more resources required

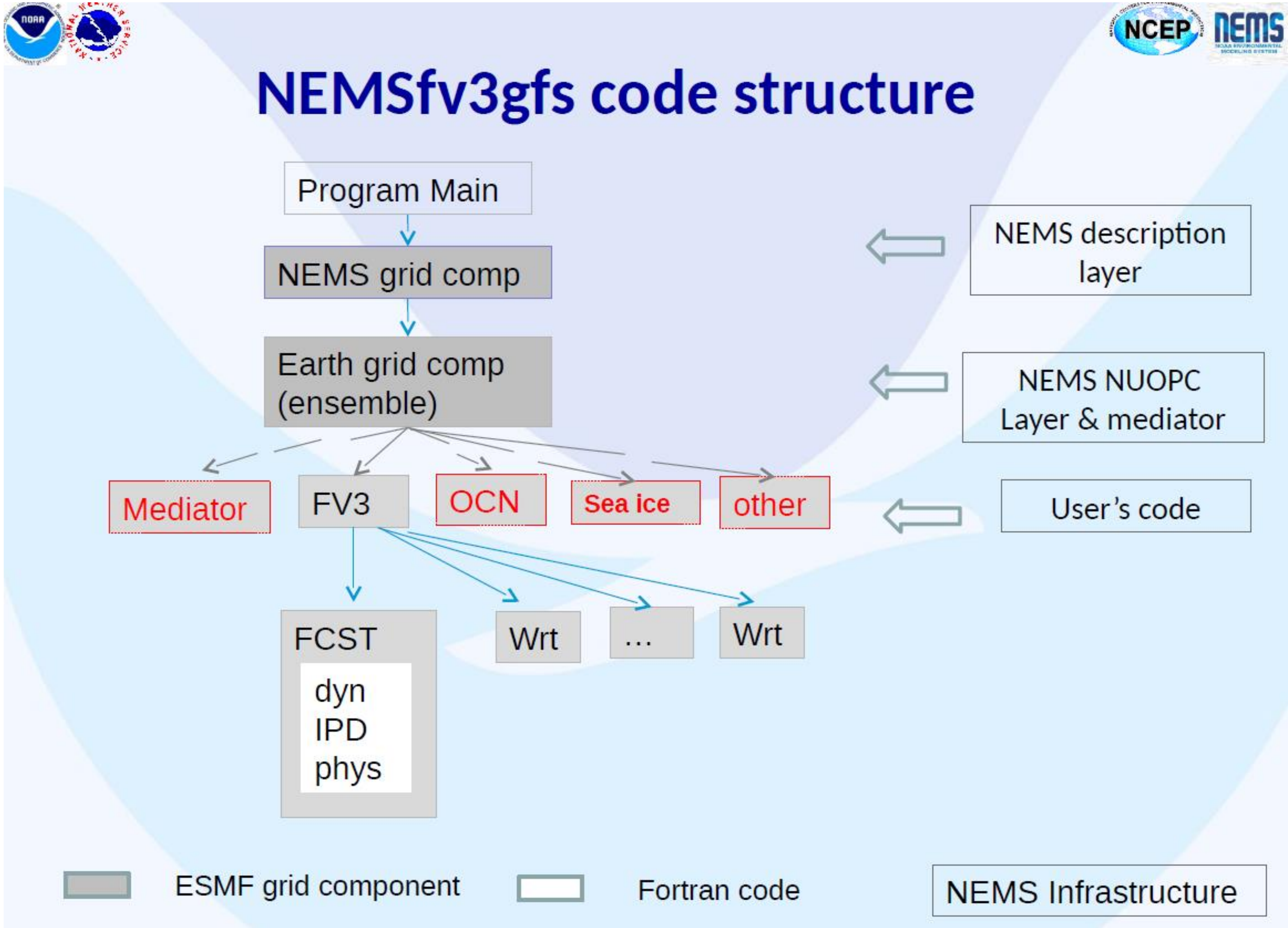
Günther Zängl

NOAA/NCEP – FV3

- NOAA is developing a fully coupled atmosphere-ocean-ice-wave-chemistry model (UFS) for future weather and S2S applications.
- Unified Forecast System (UFS) https://ufscommunity.org/index.html
- Joint Effort for Data assimilation Integration (JEDI)*
- A concern is on the coupling of the subcomponents to achieve the best performance.
- “I/O server” write grid component for cubed sphere tiles of FV3 (and conversion to Gaussian grid)
- Happy with scaling performance of FV3 model (when I/O separated out)
- Scalability benchmark report (incl. nesting capability with higher resolution over the US)
  - https://www.earthsystemcog.org/site_media/projects/dycore_test_group/20160701_AVEC_Computing_Benchmarking_Final_Report.pdf

Fanglin Yang

NOAA/NCEP-FV3



NOAA/NCEP- FV3

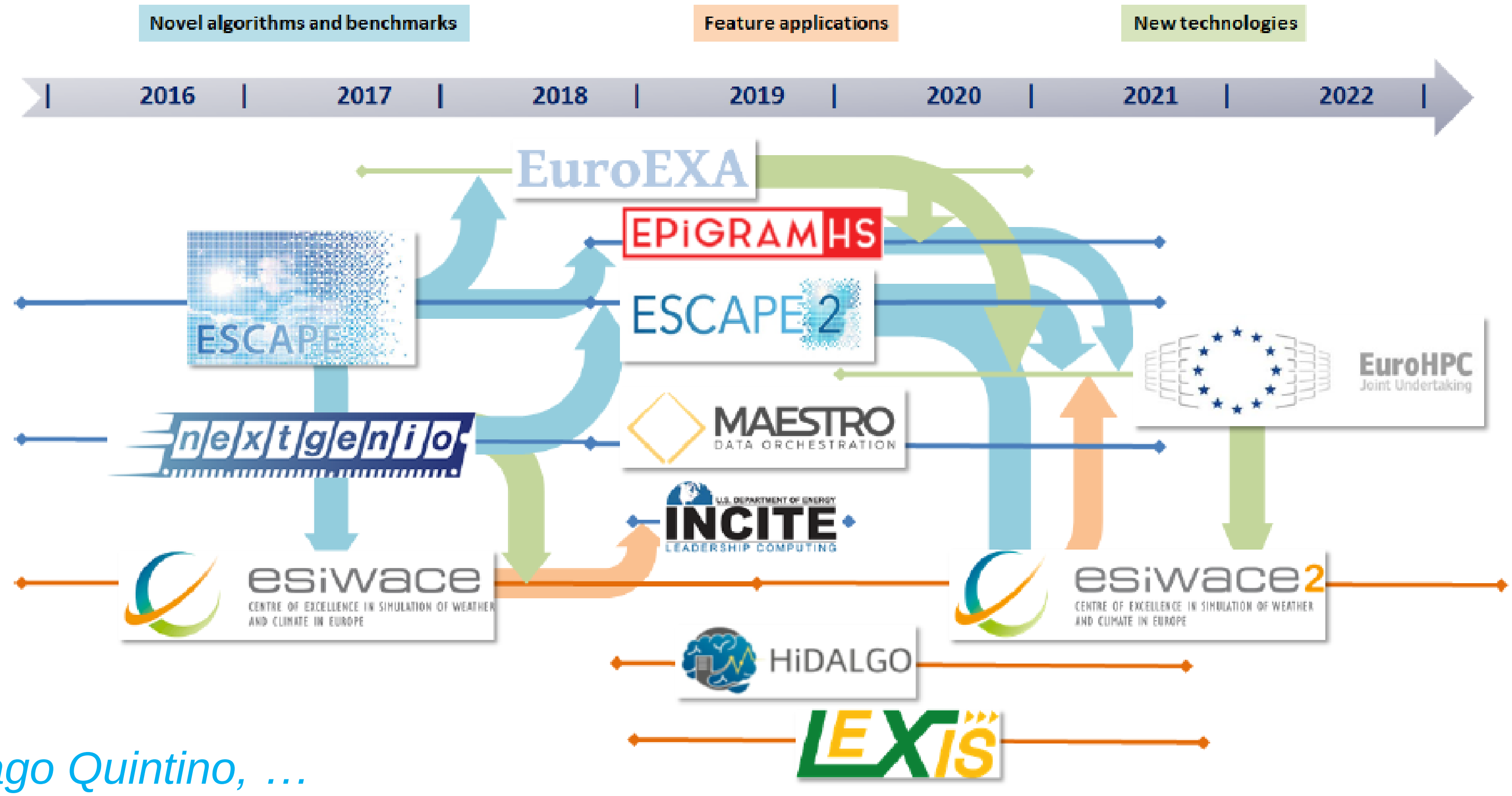
Status report on NEMS FV3 write grid component

Jun Wang, Gerhard Theurich, James Taft, Dusan Javic

Acknowledgement: Fanglin Yang, Shrinivas Moorthi, Jeffrey Whitaker, Philip Pegion, Robert Oehmke, Hui-ya Chuang, Vijay Tallapragada, Arun.Chawla, Mark Iredell, Cecelia DeLuca

ECMWF – IFS

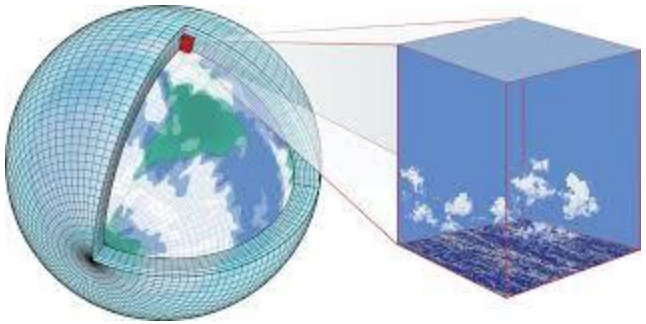
- Spectral, semi-implicit, semi-Lagrangian (octahedral, reduced Gaussian grid)
- Scalability Programme (the first 5 years)
- H2020



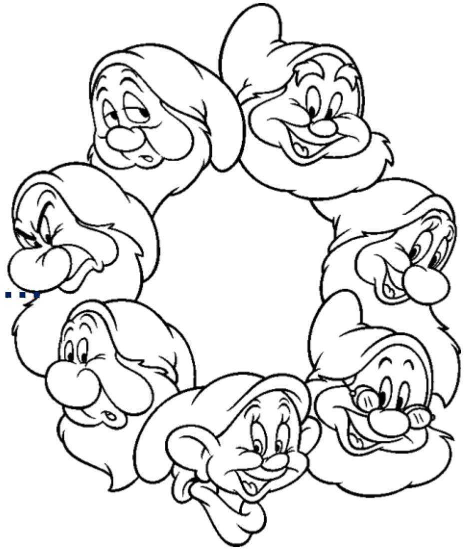
Peter Bauer, Nils Wedi, Tiago Quintino, …

Weather & Climate Dwarfs

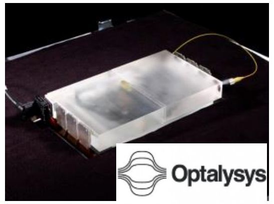
(hpc-escape.eu)

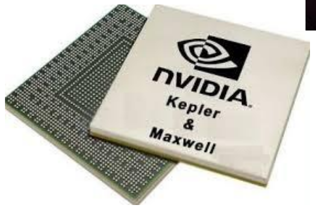


Extract model dwarfs...

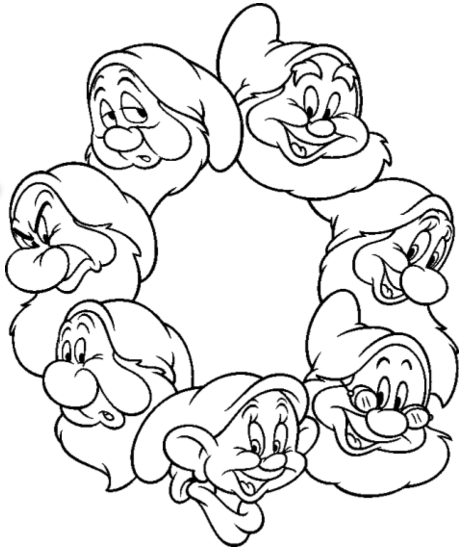


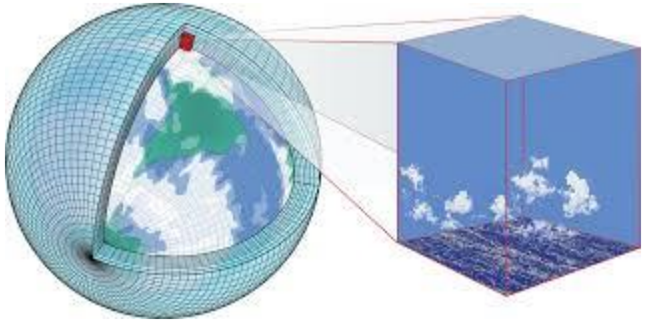
… hardware adaptation …





… explore alternative numerical algorithms …

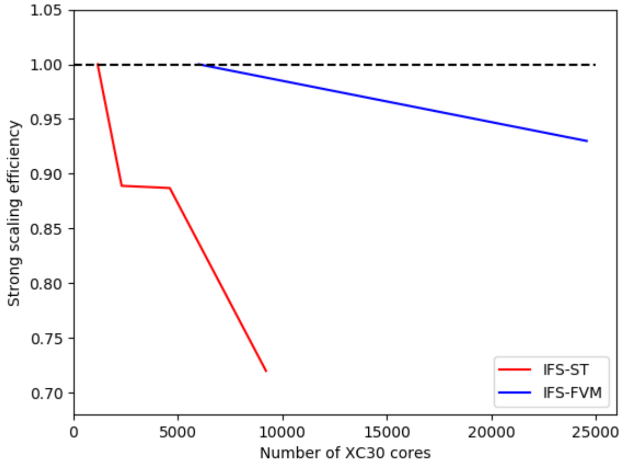




… reassemble model and benchmark

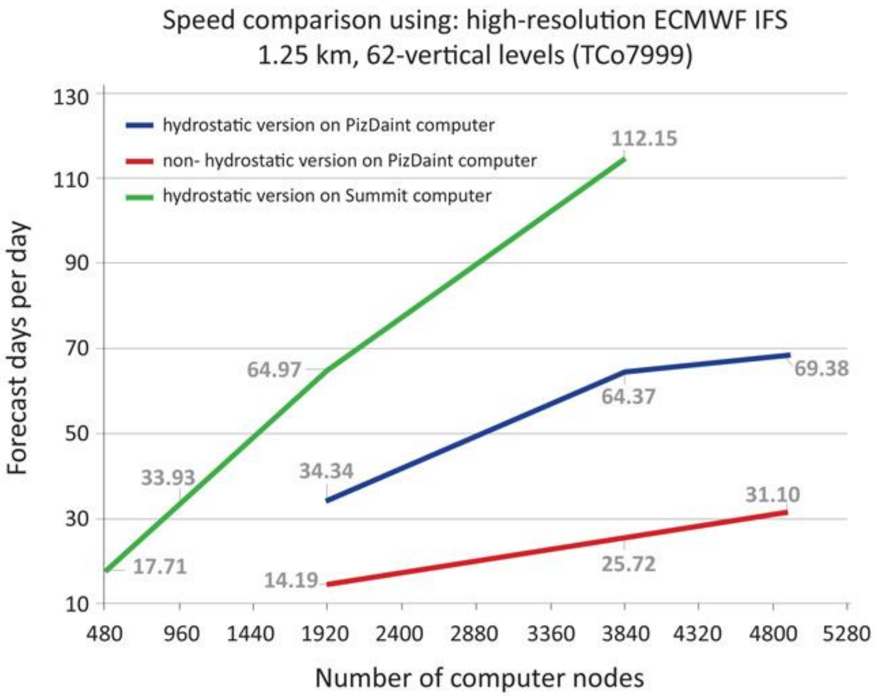
ECMWF · MeteoSwiss · DMI · RMI · NVIDIA · METEO FRANCE Toujours un temps d'avance · ICHEC · Deutscher Wetterdienst Wetter und Klima aus einer Hand · Optalysys · Bull · PSNC · Loughborough University

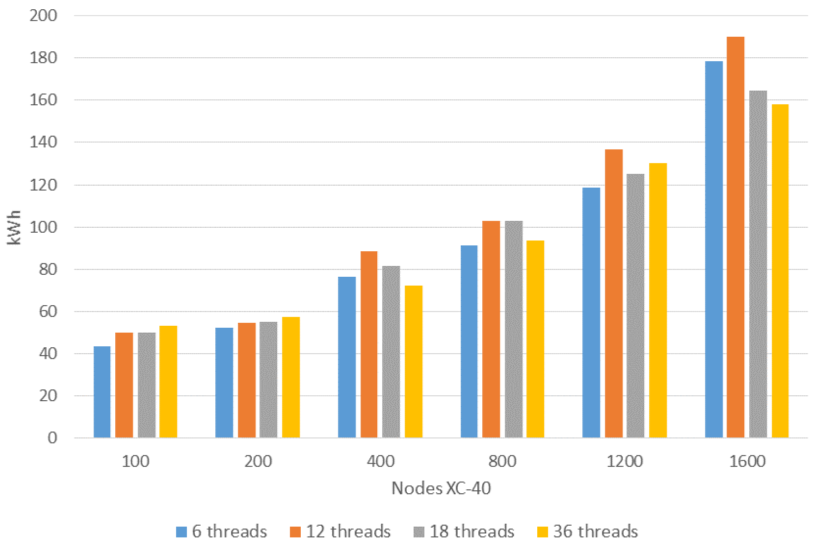
Scalability concerns



Strong scaling

Time-to-solution





Energy-to-solution

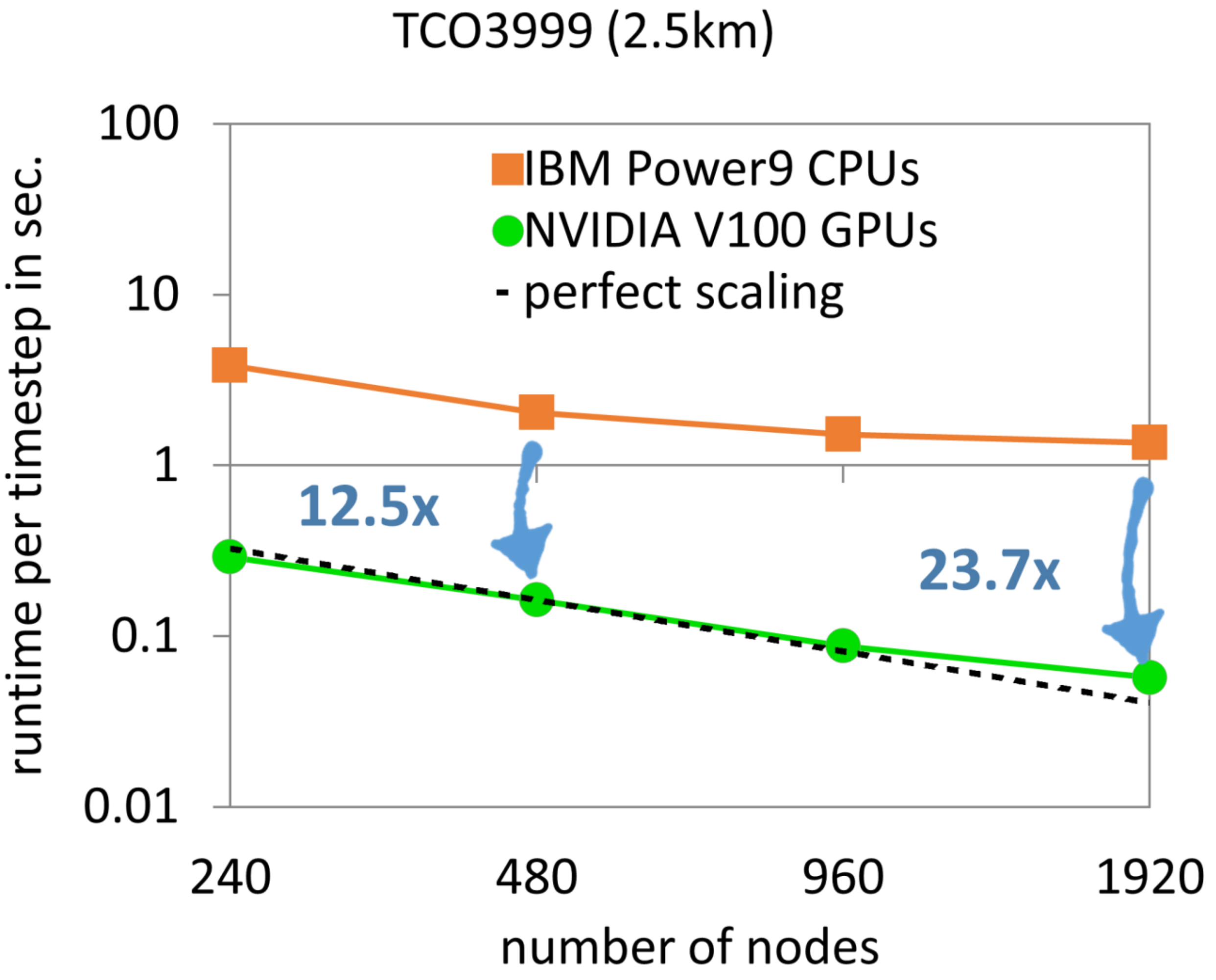
Funded by the European Union

Spectral transform dwarf

- 2019: access to Summit (fastest supercomputer in the world)
- done:
  - ☑ make dwarf run on Summit
  - ☑ compare CPU and GPU
  - ☑ run high resolution on GPUs
  - ☑ scaling efficiency on GPUs and CPUs
  - ☑ move GPU code into trans library

  - ☐ apply GPU optimisations to current trans
  - ☐ optimise vorticity/divergence for GPU
  - ☐ new GPU trans inside RAPS on Summit
  - ☐ avoid zero operations

- 1920 nodes = 11520 GPUs (42% of Summit)
- energy of GPUs = 6x energy of CPUs per node
- fast alltoallv thanks to fat-tree network (like one electrical group across entire machine)

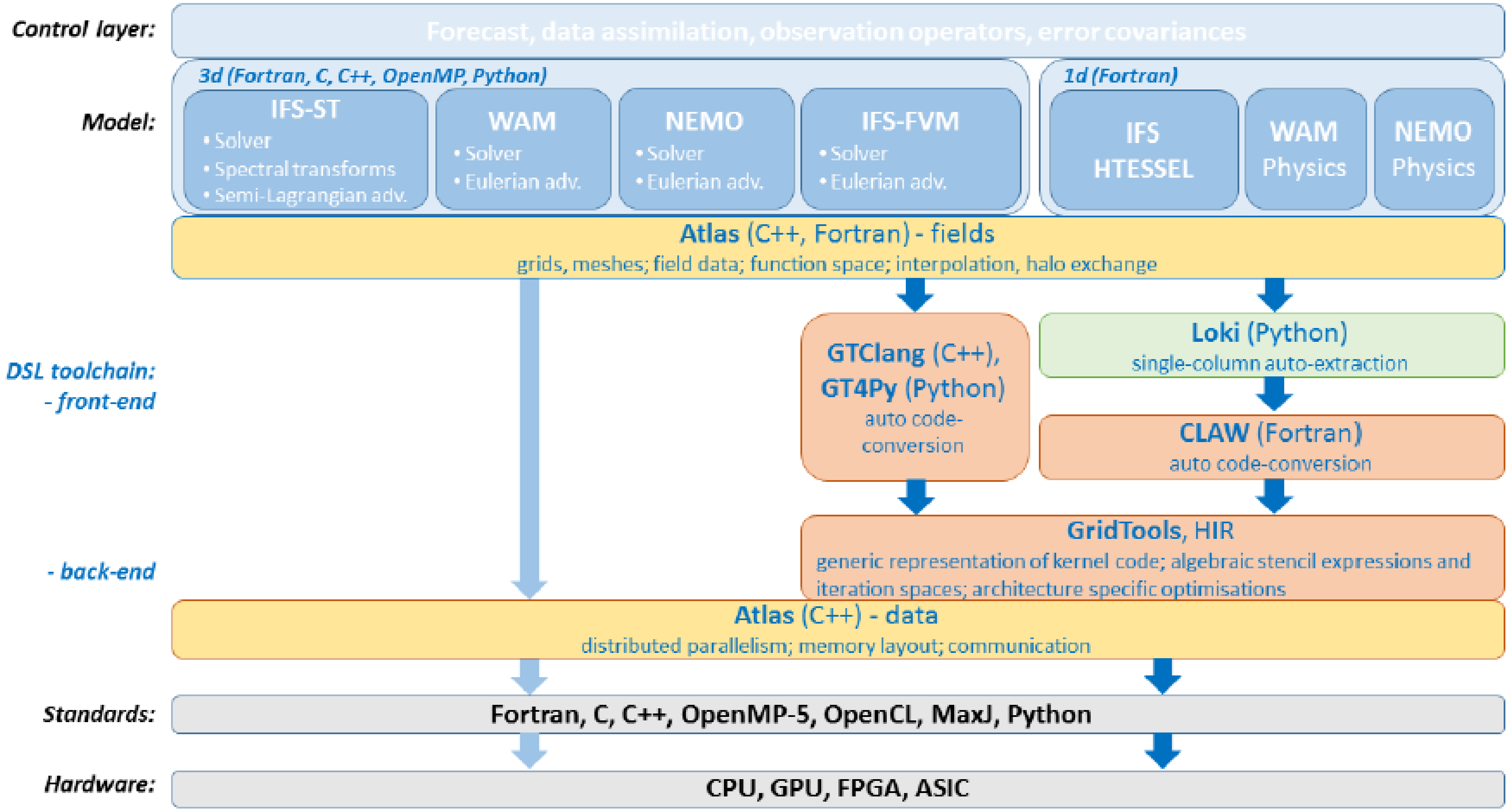


A. Mueller

This research used resources of the Oak Ridge Leadership Computing Facility, which is a DOE office of Science User Facility supported under contract DE-AC05-00OR22725.

Performance and portability

M. Lange, O. Marsden



Structure and components necessary for the transition of the IFS to separate applied science from hardware sensitive code level

ECMWF

Preparing for low-power acceleration

M. Lange, O. Marsden, B. Reuter

EuroEXA project

EuroEXA project: Goals and opportunities

- Demonstrate roadmap towards "European Exascale m
- ECMWF is a software partner with primary focus on FF
- Opportunity to evaluate methods to target different HP(
- Examine "dataflow computing" in principle and with reg

Dataflow computing and FPGA accelerators

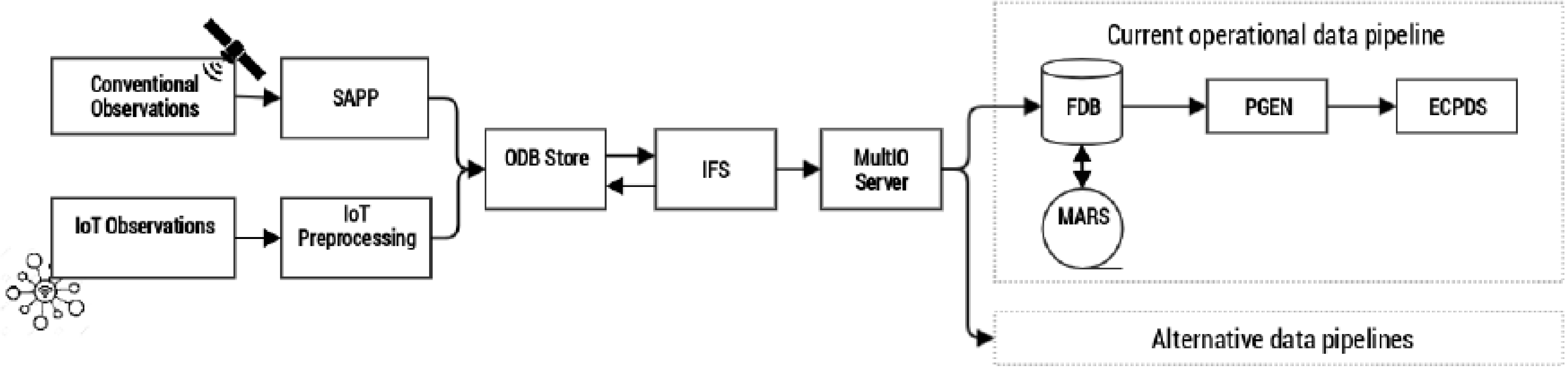
- Power consumption in HPC architectures (CPU and GI is dominated by data movement (memory, cache, regis
- Instead of general-purpose arithmetic units, an FPGA chip has reconfigurable circuitry, that allows bespoke arithmetic pipelines to be configured (dataflov
- As a result data "flows" through the chip, while perform compute in place with lightweight arithmetic units, thus using much less power

Porting CLOUDSC to an FPGA

CLOUDSC: Semi-automatic FPGA port via Maxeler toolchain

- Auto-generated bit-identical C code from Fortran via Loki
- Hand-ported to MaxJ dataflow environment in collaboration with James Targett from Imperial College
  - 1.9% of logic resources, 53.1% of DSP blocks
  - 43.7% of BRAM18 memory blocks and 17.1% of URAM memory blocks
  - → **Main point: It fits on the chip!**
- Validated bit-reproducible single-precision FPGA kernel
  - **Paper in preparation…**
- Performance extrapolation with simulator and Max5 DFE accelerator:
  - Extrapolated equivalent FPGA performance of **133.6 Gflops/s**
    - Benchmarks run out of on-board memory (ignore PCIe overhead)
  - Single-socket performance:
    - **3x faster than Haswell (LXG) and >3x lower power usage**
    - **2.5x faster than Cray XC40 (Broadwell) socket**

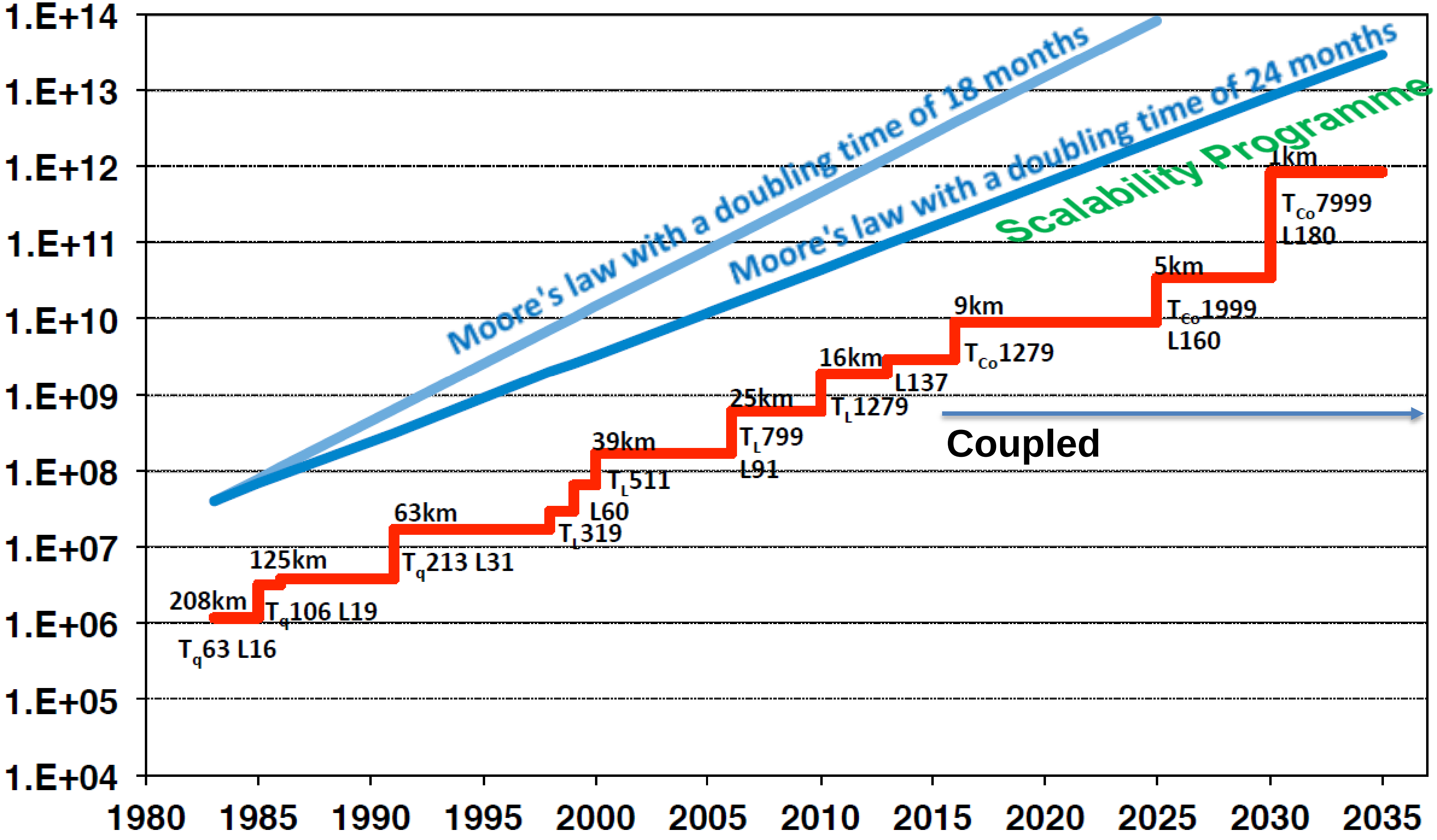
A vision of data-centric workflows at ECMWF



- Harmonised data model beyond individual applications
- Establishing a cloud-based data handling philosophy
- IFS == OOPS DA system

T. Quintino, S. Smart, et al

Computational power drives spatial resolution



(Schulthess et al, 2019)

- Performance increase at reduced energy cost

ECMWF's progress in degrees of freedom
(levels x grid columns x prognostic variables)

Summary:

- Kernel/Dwarf improvements through
  - More traditional reduction of MPI tasks through hybrid MPI/OpenMP
  - Novel algorithms, quasi-uniform mesh (cubed sphere adopted by several)
    - Brings back the old problems: coupling to the physics and to other component models
  - Overlapping communication and computation, data layout restructuring
  - Reduced precision
- Performance and portability
  - GPUs, directives programming (which ?)
  - Separation of concerns, code generation and DSLs, disruptive approach
  - Performance portable restructuring of data layout
- Data-centric workflows
  - Separation of (large) I/O problem beyond individual applications
  - AI facilitator
- Coupling frameworks and data assimilation
  - ESMF, OASIS, single executable, same grid, other ?
  - JEDI, OOPS (won't address the scalability problem of DA per se)

Concerns/Needs:

DSLs too disruptive ?

GPU adaptation worth it ?

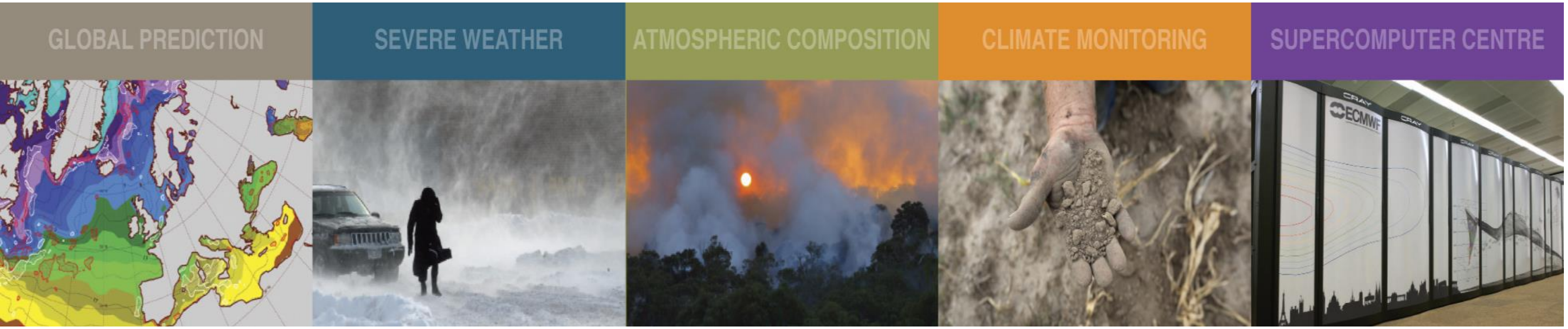
"Traditional" approaches enough ?

Speed-up required/expected: 4-8 in the next 5 years (at a constant or much smaller increase in energy cost)

Additional slides

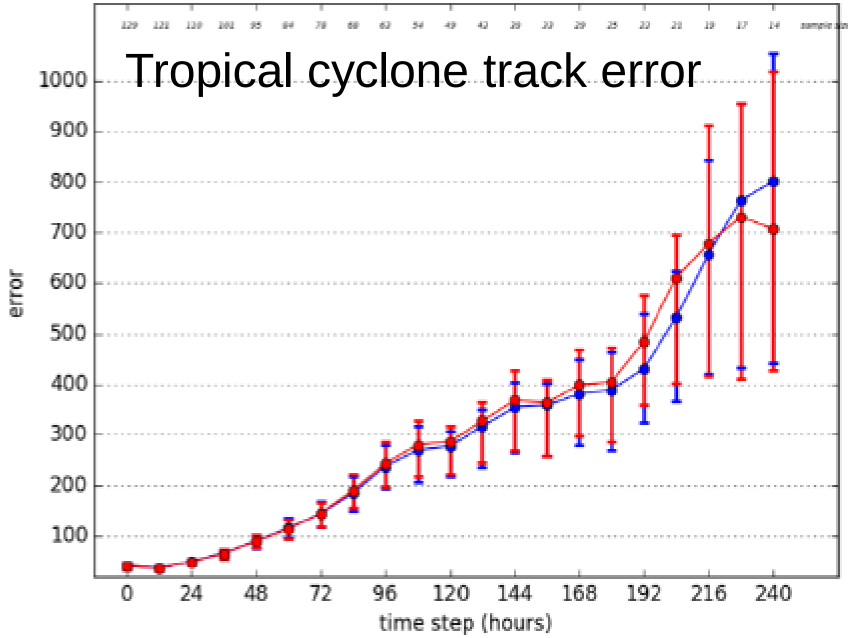
Mixed Precision

Nils P. Wedi, European Centre for Medium-Range Weather Forecasts (ECMWF)

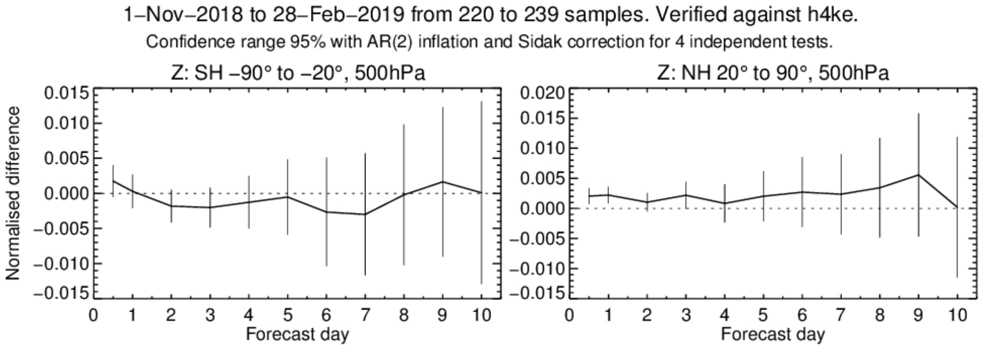


ECMWF - IFS

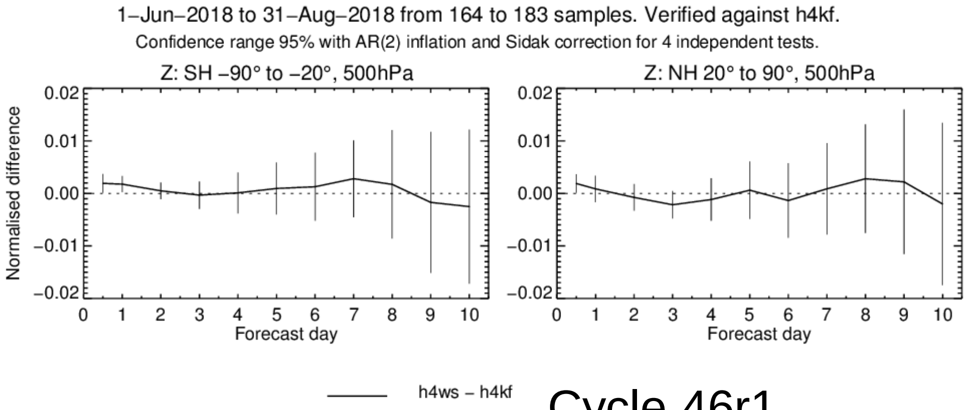
- Single precision everywhere (atmosphere, surface, waves) except in very selected parts of the code (eg. Precomputations of Legendre transforms, some geometric, trigonometric computations in SL, vertical integral with finite-elements, …), saving about 40% (Vana et al, 2017)
- Dominant impact: smaller number of cache misses and better vectorisation
- Global mass conservation impacted (simple fixer at the moment)
- NEMO in SP next
- Planned for operational implementation in 2021



Winter



Summer



Cycle 46r1

NOAA/NCEP – FV3

- At NCEP EMC, we are running the FV3-based global forecast system (GFS) for operational medium-range weather forecast at ~13km horizontal resolution with 32-bit precision for the dycore and 64-bit resolution for physics.

- The configuration reduces computing cost by about 30%. Currently there is no plan to rewrite the physics component to run at 32-bit precision.

DWD – ICON / COSMO

- For ICON, we've been using a mixed precision approach for several years in operational production.

- It employs single precision for coefficient fields and most intermediate storage fields in the dycore and to a lesser extent in the tracer transport code, but the prognostic variables and the physics parameterizations are still DP. Overall, this saves about 20% of total computing time for our global configuration, it will be a bit more for the upcoming convection-permitting limited-area configuration because some physics parameterizations (particularly radiation) are called less frequently in a relative sense and thus make a smaller contribution to total computing time.

- For the COSMO model, an option for using SP in the whole model is available, which saves a bit more than 40% of total computing time but still has unresolved issues in the assimilation cycle. We will start using it for forecast runs in the near future. It was primarily developed by our Swiss colleagues and required much larger developer resources than ICON's MP option (my estimate would be about two orders of magnitude) because many adaptations were necessary in the physics parameterizations in order to reduce the sensitivity of the numerical algorithms to truncation errors.

CMC - GEM

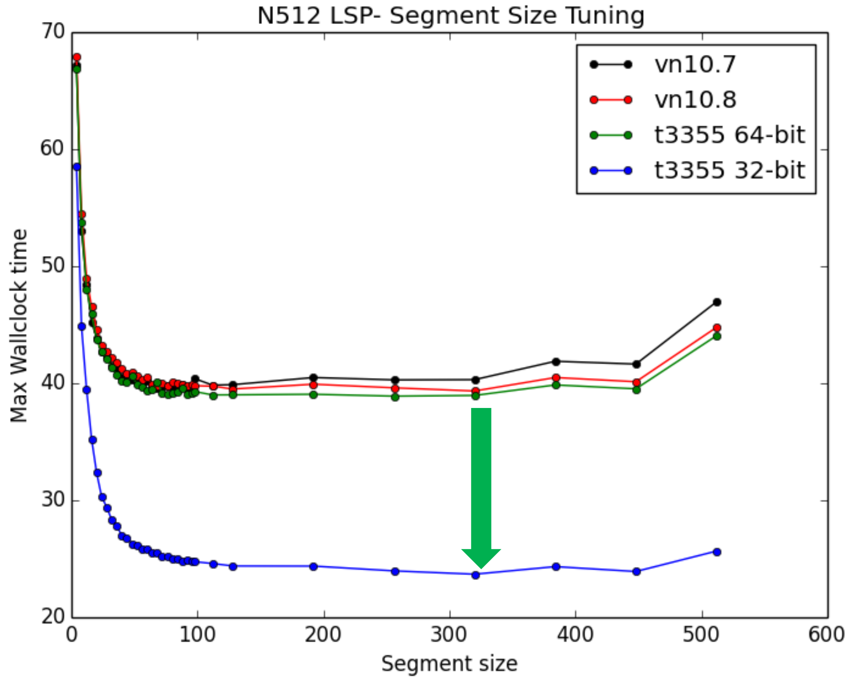
- The main 3D storage for GEM is single precision. Most computation in the dynamical core is either promoted to double or directly written in double.
- The elliptic solver is written in double precision. In a dynamics largely memory bound it would probably be catastrophic in terms of performance to go any further in term of increasing precision.
- In the physics, the majority of calculations are done in single precision except within specific algorithms generally involving sums and root finding. (We recently moved this year from double to single precision in our gwd/blocking scheme with algorithmic changes to yield a neutral result in terms of forecast skill.)

UKMetOffice – UM

- Prototype testing 37% faster at neutral impact, due to improved cache use
- Considerable effort for existing UM, not currently planned
- LFRic design will allow double, single, or less precision of individual components

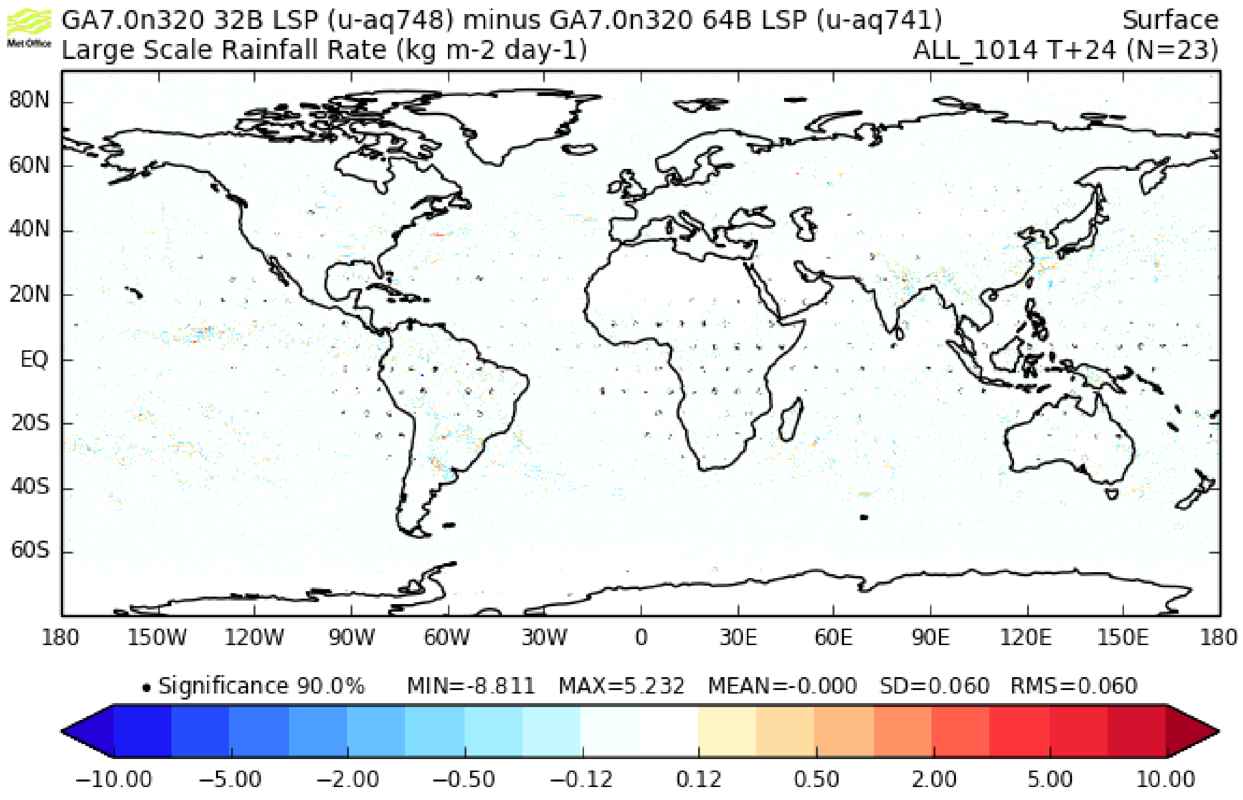
Computational Performance

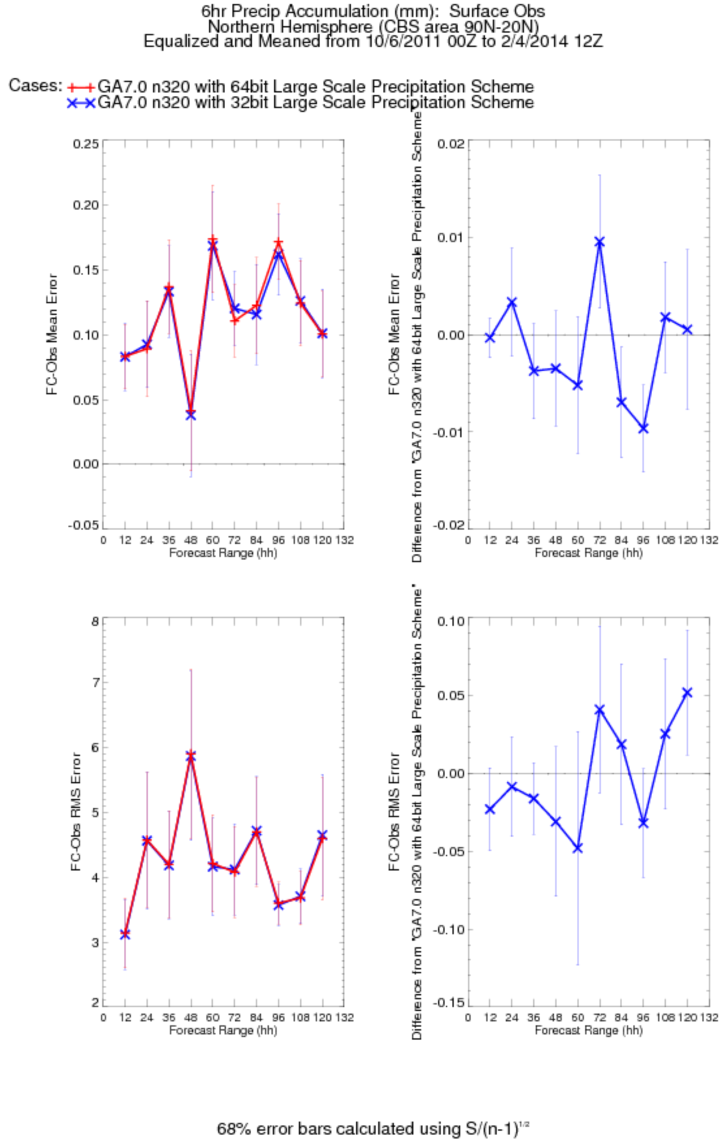
- LSP uses segments to improve cache use
  - 17 Nodes
  - 2 OpenMP threads
- Scheme ~37% faster at 32-bit



Scientific Performance
-Essentially neutral impact

N320 GA7 case studies





Scientific Performance

-Essentially neutral impact