Atmosphere and ocean scalability and HPC readiness - 2020

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

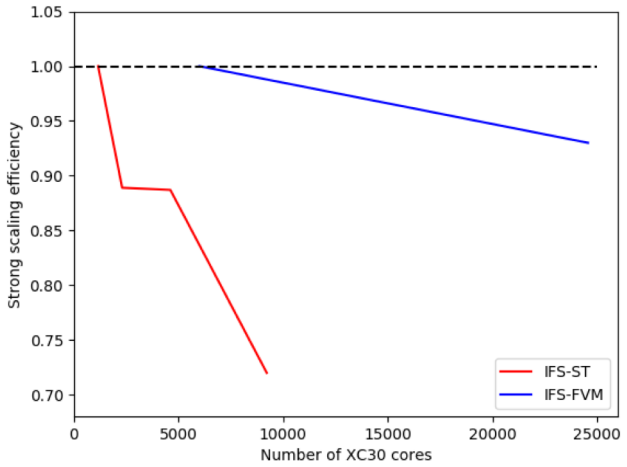
Many thanks to WGNE members for providing the requested information

Introduction

- Information provided by
  - CMA, CMC, DWD, ECMWF, Hydromet Russia, NOAA/NCEP, *JMA*, MeteoFrance, *NRL, UK MetOffice, Wits GCI (South Africa)*
  - ECMWF Annual Seminar 2020 on Numerical Methods
  - HPCwire background:

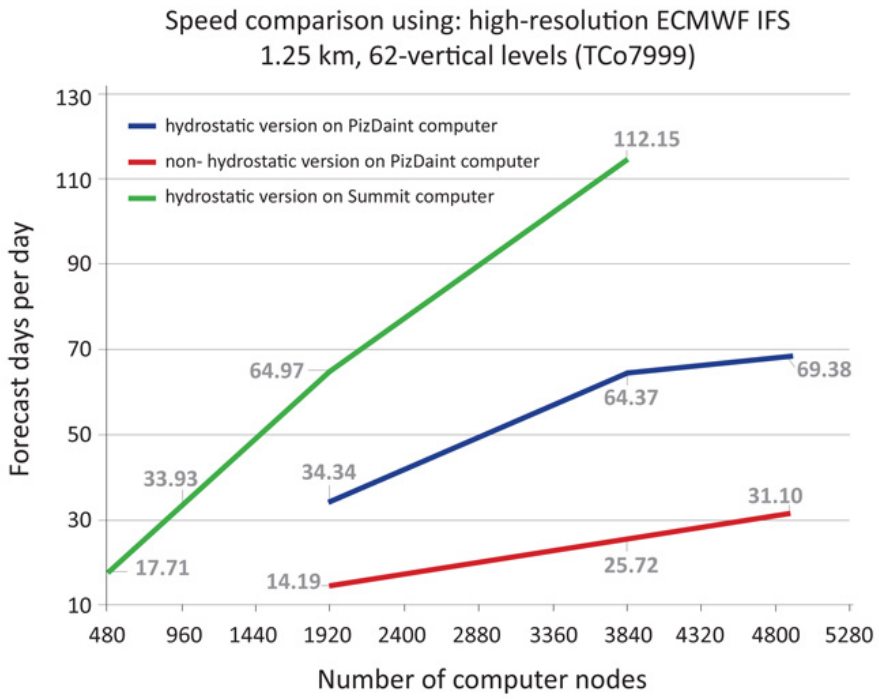
    - *Fugaku, Riken*: ARM 2 nodes x 48 cores, no GPUs
    - *Frontier OLCF*: AMD CPU + 4x AMD GPU acceleration
    - *LUMI EuroHPC*: AMD CPU + Xx AMD GPU acceleration
    - *Next Earth Simulator, JAMSTEC*: NEC 8 vector engines powered by AMD CPU
    - 4 x (s*maller) EuroHPC systems*: nodes with Xx NVIDIA A100 GPUs

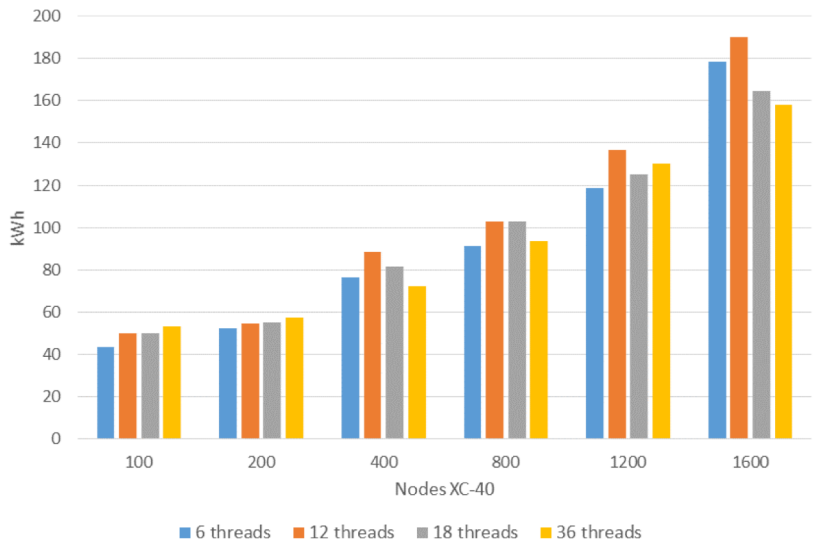
Scalability concerns



Strong scaling

Time-to-solution





Energy-to-solution

2019 WGNE Report summary

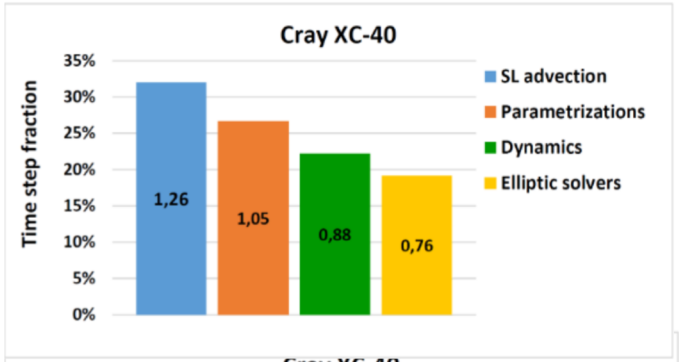
- Hybrid parallelisation MPI/OpenMP
- Targeting the efficiency of advection and other key algorithmic patterns, often with the help of dwarfs, and directive-based porting to GPUs
- Evolving new dynamical cores
- Focus on ESM components, e.g. ocean and wave model scalability
- Use of reduced precision
- Separation of concerns (DSLs, PsyClone, GT4PY, kokkos)
- Targeting I/O performance
- Starting to see efforts to use AI methodologies for accelerating time-critical models (none used in operational contexts)

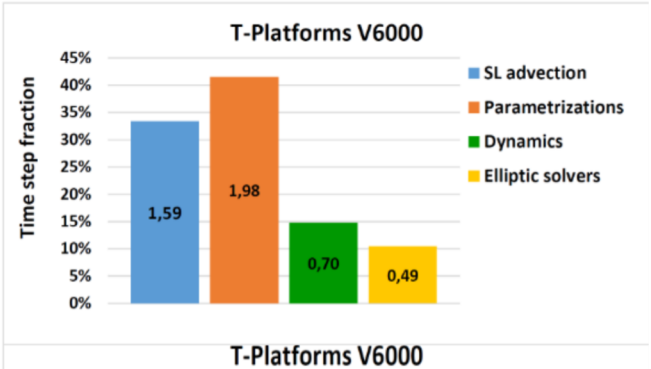
Hydrometeorological Centre of Russia – SL-AV

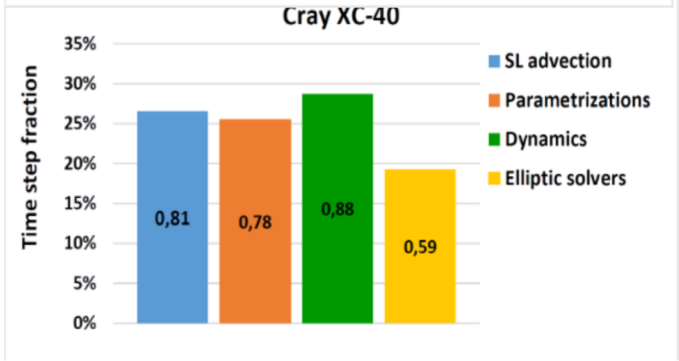
- Switching most time-consuming part to single precision (semi-Lagrangian advection)
- Optimizing the vector length in parameterizations of subgrid scale processes
- Reduction of data amount in transpositions by making them single precision instead of double precision

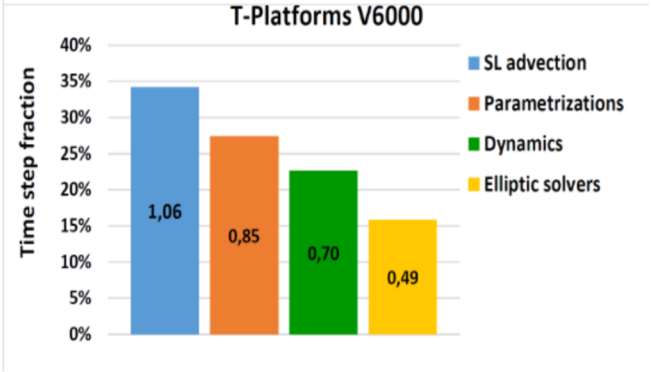
Percentage of time used in different parts of SL-AV model code while using 3888 cores at Cray XC40 (left) and 3880 cores at T-Platforms V6000 (right); before (top) and after (bottom) optimizations.
~23% reduction of wall clock time per 24hrs forecast

Number inside the column denotes the wall-clock time of respective code part (in seconds).









Tolstykh Mikhail

UKMetOffice

- Move towards a quasi-uniform mesh
- Improve conservation and scalability
- Gung-Ho: New model in 2024, mixed FE/FV, uniform-mesh with cubed sphere
- Separation of concerns and code generation with PSyClone
- Leading HPC scalability efforts on NEMO
- Coupling via OASIS
- JEDI for ocean (and atmosphere?) data assimilation

- Move to advective form of momentum equation
- Use of multigrid preconditioning for elliptic equation
- Measures to improve stability of new dynamical core (see *Tom Melvin*'s talk at ECMWF Annual Seminar 2020)

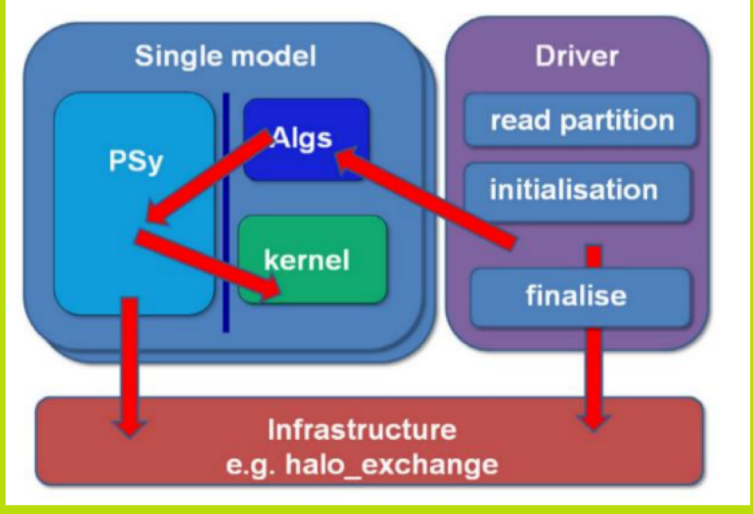
Gung-Ho/LFRic Team

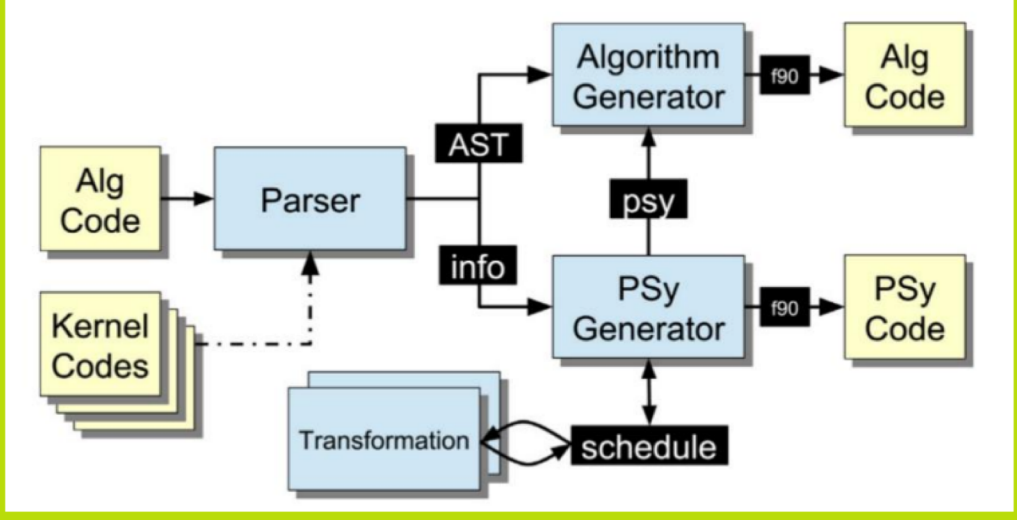
Met Office

PSyclone

Code Generation: PSyclone

- LFRic [3] uses the PSyclone [4] code generation tool to create the parallel layer
- PSyclone reads the code and generates interfaces based upon defined rules
- Parallelisation & optimisation strategies are defined via a simple script





[3] *LFRic: Meeting the challenges of scalability and performance portability in Weather and Climate models*, Adams et al J. Par. Dist. Comp. 19

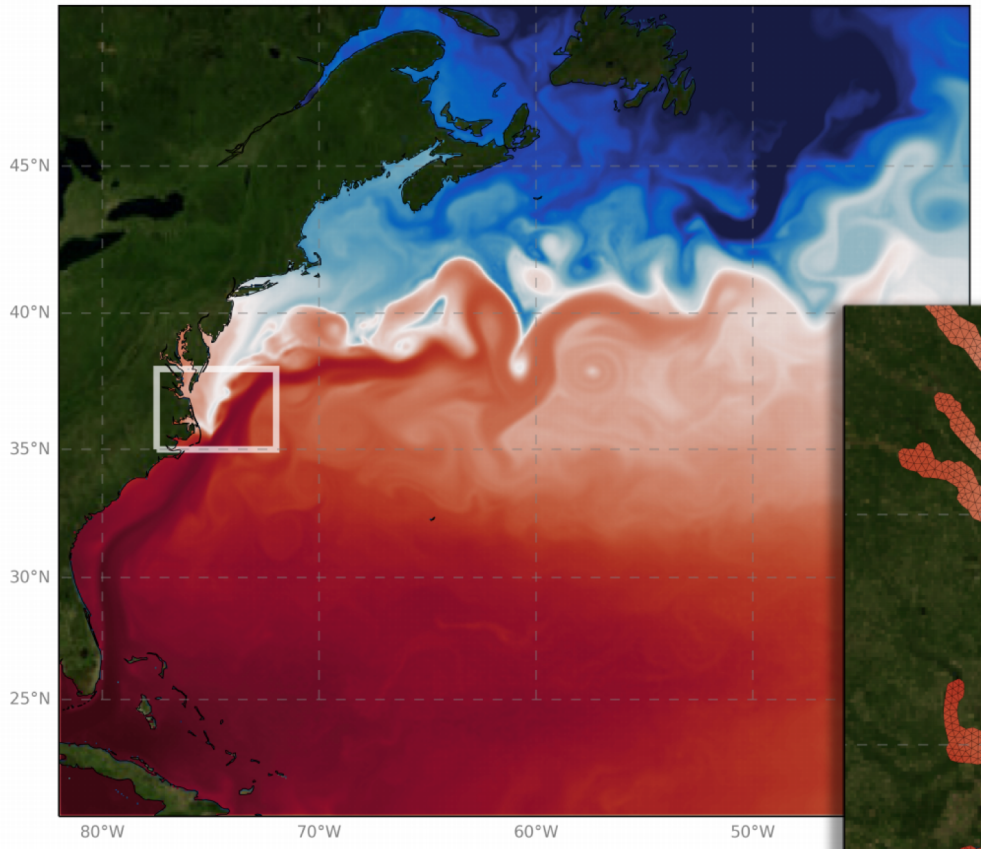
[4] github.com/stfc/PSyclone

NEMO HPC working group

- Regular telecons, Mike Bell (UKMO) NEMO HPCchair >> Italo Epicoco (CMCC, Italy)
- ECMWF participation – Good Progress with single precision NEMO4, based on the previous work of Barcelona Supercomputing Centre (BSC)
- 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

Ocean grids

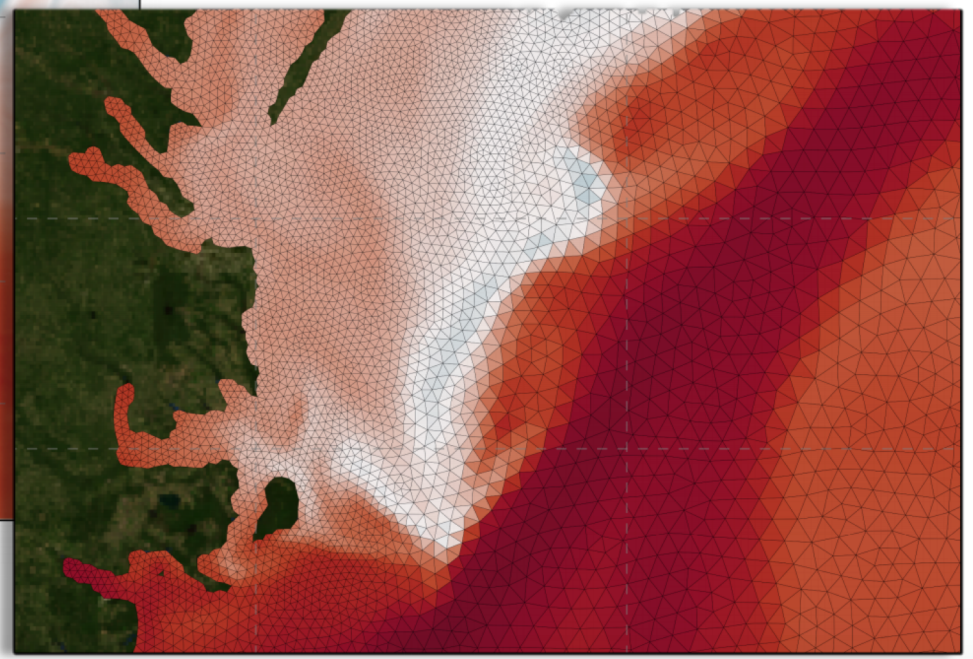
S. Danilov



Visualization N. Koldunov

Three new ocean models are formulated on unstructured meshes:
FESOM (Danilov et al. 2004, 2017; Wang et al. 2014)
MPAS-Ocean (Ringler et al. 2013)
ICON-o (Korn 2017)

How to better use variable resolution?



HELMHOLTZ ASSOCIATION

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

Meteo-France

Francois Bouyssel

- **Adaptation of our code to new architectures :** We focus on porting to graphic processors (hybrid CPU/GPU is the target). By participating in a “hackathon” organized by NVIDIA, some routines of the physics were ported. Using OpenACC seems a reasonable approach. We plan to follow as much as possible the work done in ECMWF.
- **Single precision computations :** The forecast configuration has been set up. Preliminary work on data assimilation configuration has started. The main issue that has been addressed is the poor conservation of mass. A mass fixer has been tested it solves the problem.
- **Continuous work on optimization for global model ARPEGE and local model AROME :**
  - Memory optimization in order to limit memory bandwidth: 1.6 % of total time saved in AROME just by removing useless initializations or copies of array.
  - Proper limitation of useless gridpoint computations in the extension zone in order to save time.
  - MPI communications of semi-Lagrangian computations are now faster with less limitations linked in the extension zone.
- **Developments in the ATLAS library :** Many developments have been undertaken in the ATLAS library (ATLAS, developed within the ECMWF is the new library that handles geometry and communications), the objective is to perform some of our postprocessing tasks with ATLAS.

Meteo-France

Francois Bouyssel

- **Developments in the ATLAS library :** Many developments have been undertaken in the ATLAS library (ATLAS, developed within the ECMWF is the new library that handles geometry and communications), the objective is to perform some of our postprocessing tasks with ATLAS.

- **Reducing global communications :** The semi-implicit problem in our LAM model AROME can now be solved in grid-point space instead of the spectral space with a neutral impact in terms of forecast skill. The final goal is to eliminate the MPI_ALLtoALL communications of the spectral transforms, which have the reputation to be less scalable.

- **Very high-resolution global forecasts :** To assess current scalability of the ARPEGE global model, simulations were performed at a very high global resolution of 2.5km mesh size with non-hydrostatic dynamics and the physics of our high resolution LAM model AROME. At the maximum, on 230,400 cores (80% of the computing machine) 40 days per day was obtained. Scalability is still acceptable.

CMC - GEM

Ron McTaggart-Cowan

- As a Centre with relatively small computational resources, exascale computing at the CMC still some distance in the future; however, scalability is a practical concern that is the focus of several ongoing projects related to the dynamical core:
  - Develop a spatial discretization based on the flux reconstruction approach (it has the advantage of being local, accurate,flexible with good conservation properties);
  - Time-stepping based on exponential propagators (advantage: Eulerian with large time-steps). Collaboration with the University of California at Merced, Lawrence Livermore National Laboratory, Southern Methodist University and Università degli studi di Verona;
  - The applicability of these methods to GPU-based architectures
- Physics
  - Profiling of the physics package to identify regions of the (large) code base that have particularly poor performance characteristics
  - Optimize single-processor performance: all parameterizations still use the independent column approximation, and are thus embarrassingly parallel
  - A project has been initiated with the Mila – Quebec AI Institute to investigate utility of machine learning for algorithm replacement, beginning with radiative transfer calculations
- One of the current (avoidable) limits on scalability is model I/O
  - implementing an I/O server is an infrastructure development priority
  - Work on better compression schemes and rules to lower I/O impacts on communications and storage
  - Re-engineering of the coupling system for exascale level of scalability and performance

DWD - ICON

- Model scalability sufficient for next 10 years, Data assimilation at the limit both in terms of memory and compute.
- GPU adaptation initially based on directives
- ICON participation in a range of projects that explore DSL approaches for porting to accelerators with a single source code (in particular ETH/CSCS, MeteoSwiss)
- Migration to NEC (vector) machine, practical experience shows small effort compared to GPU effort.
- Dependency on vendors resolving compiler issues, not specific for any vendor, a general issue
- Importance of stable running environment, batch system, and system software.
- Requirement for human readable (&accessible) debugging information and flexibility of future DSLs to minimize dependencies between (accelerated) technical ability and scientific progress

Günther Zängl

CMA

- Some components of GRAPES were rewritten and tested on GPU processors, including the Helmholtz solver and two physics schemes (microphysics and convection)
- Reduced the precision of GRAPES from double precision to single precision, and found that there were still some differences, so far keeping the model with double precision
- Re-designed the IO of GRAPES to let IO use some additional processors along with the model integration and it saved a lot of time, but when the model resolution becomes very high (which means IO will take a long time), this IO method still needs to be refined.

Jian Sun

NOAA/NCEP- GFSv16

- NCEP Global Forecast System version 16 is built with the NOAA Environmental Modeling System **(NEMS)** infrastructure. NEMS is built upon the Earth System Modeling Framework **(ESMF)** and National Unified Operational Prediction Capability **(NUOPC)** Layer code and conventions.
- It runs at the C768 (~13km) horizontal resolution and has 127 layers in the vertical extending up to the mesopause (~80km).
- The NOAA WAVEWATCH III model (WW3) is coupled to the atmosphere.

Writing Compressed Data in NetCDF Format

C768L127fcst output	Nemsio No compression	Netcdf No compression	Netcdf Lossless (deflate=1,nbit=0)	Netcdf Lossy (deflate =1, nbit=20)	Netcdf Lossy(deflate=1,nbit=14)	Netcdf Lossy (deflate=1, nbits=14),parallel writing, default decomposition chunksize	Netcdf Lossy (deflate=1, nbits=14),parallel writing Layer chunkcsize
A 3D file size (total fcst)	33.6GB (7TB)	33.6GB (7TB)	23.6GB (5TB)	13.5GB (2.8TB)	6.3GB (1.3TB)	6.3GB (1.3TB)	**6.3GB (1.3TB)**
Write Time	79s	300s	960s	680s	400s	43s	**34s**

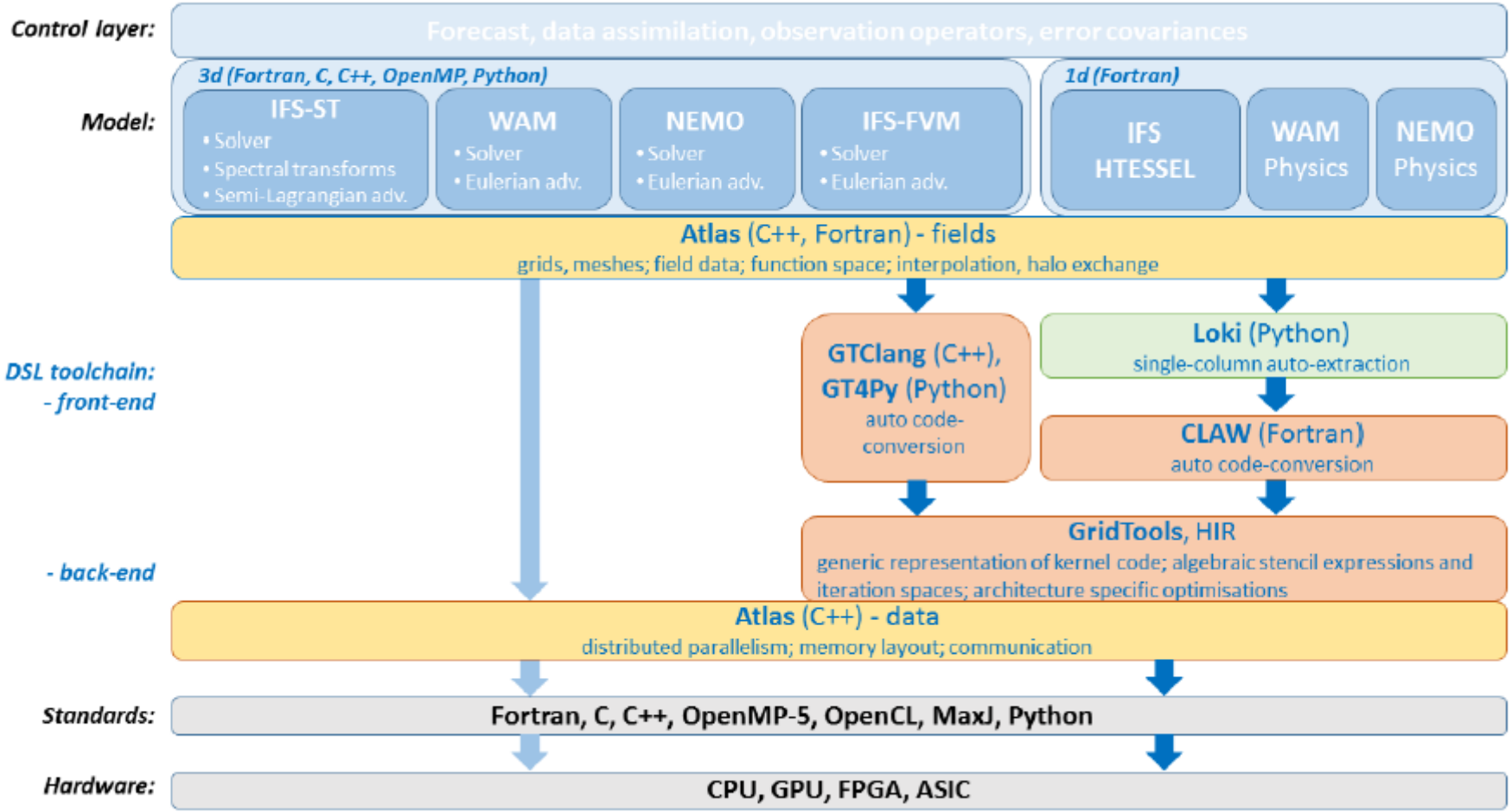
NATIONAL WEATHER SERVICE — Building a Weather-Ready Nation // 5

Substantial I/O improvements

Fanglin Yang

ECMWF - 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

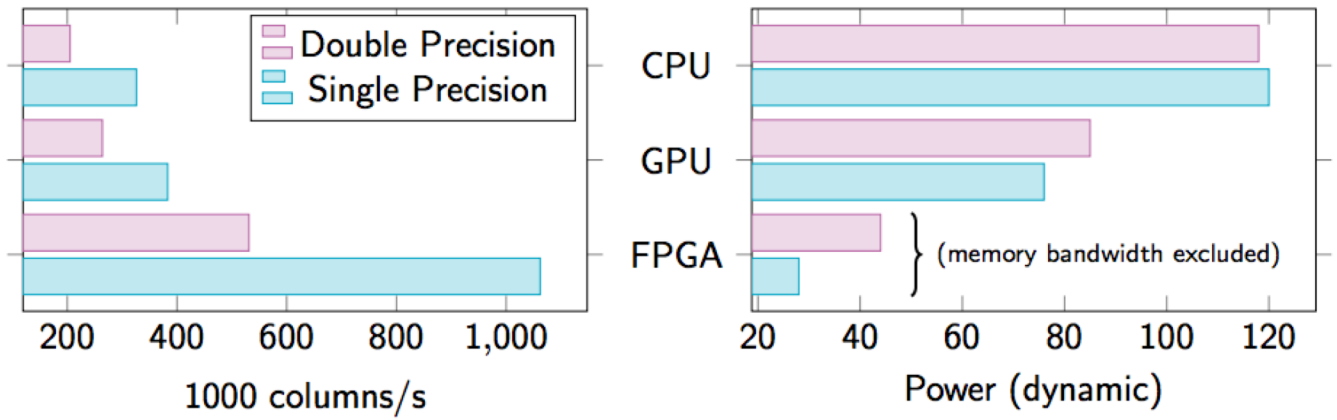
Preparing for low-power acceleration

B. Reuter, M. Lange, O. Marsden

see poster at ECMWF Annual Seminar 2020

Use case: CLOUDSC on GPU and FPGA[7]

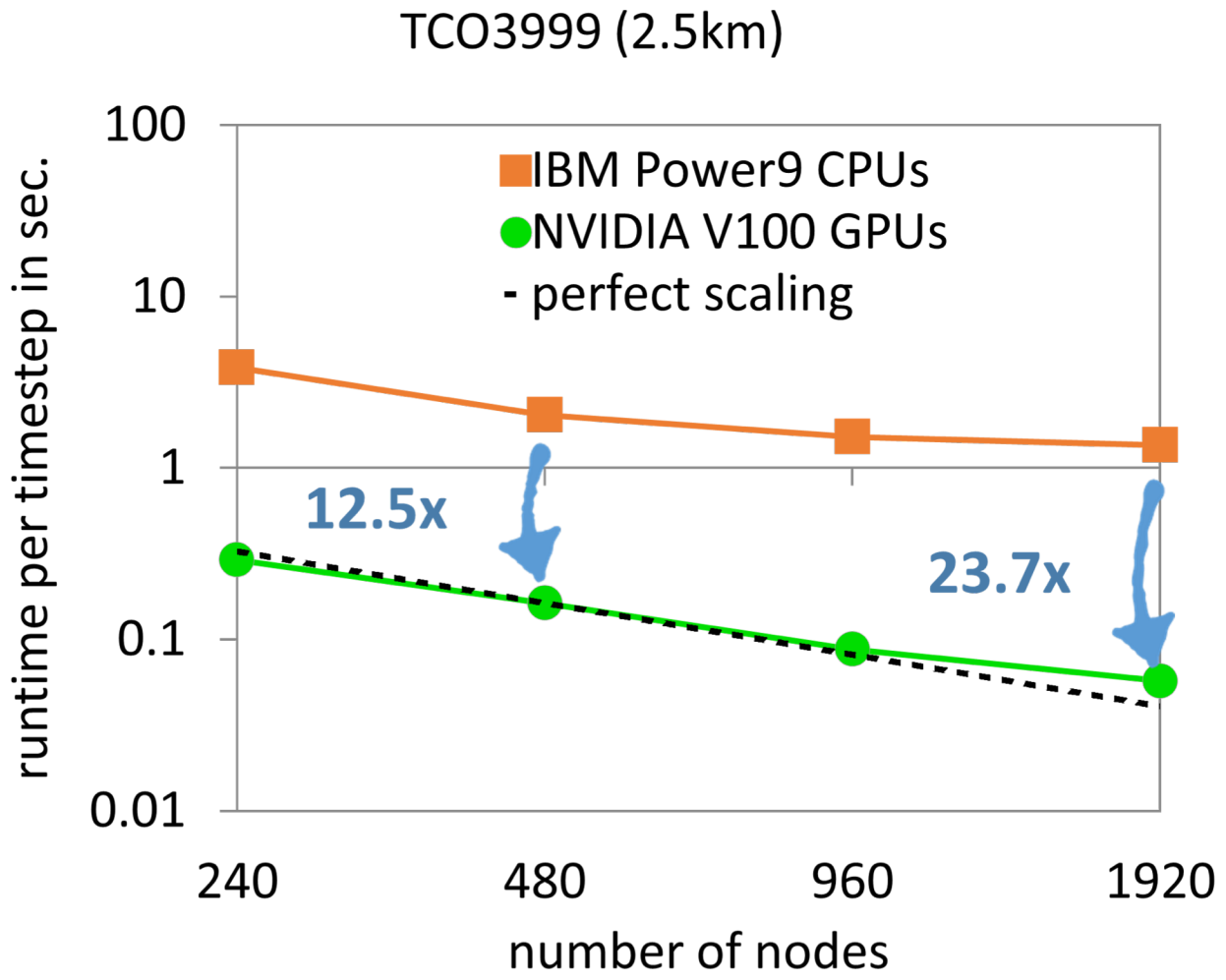
- Operational **cloud microphysics** parameterisation
- FPGA (Maxeler MAX 5 DFE (Xilinx VU9P)): Automated **Fortran to C transpilation**, then hand ported to MaxJ
- GPU (NVidia Quadro GV100 Volta): **OpenACC offload directives**, optimized for GPU
- FPGA throughput **2.5x (3x) higher than CPU** and **2x (2.8x) higher than GPU** for DP (SP)



Funded by the European Union

Spectral transform dwarf

- 2020: RAPS18 full model working in hybrid mode (spectral transforms on GPU, rest on CPU), no overlapping yet, results in 30-40 % speed-up (on the same number of nodes) at 9km resolution.

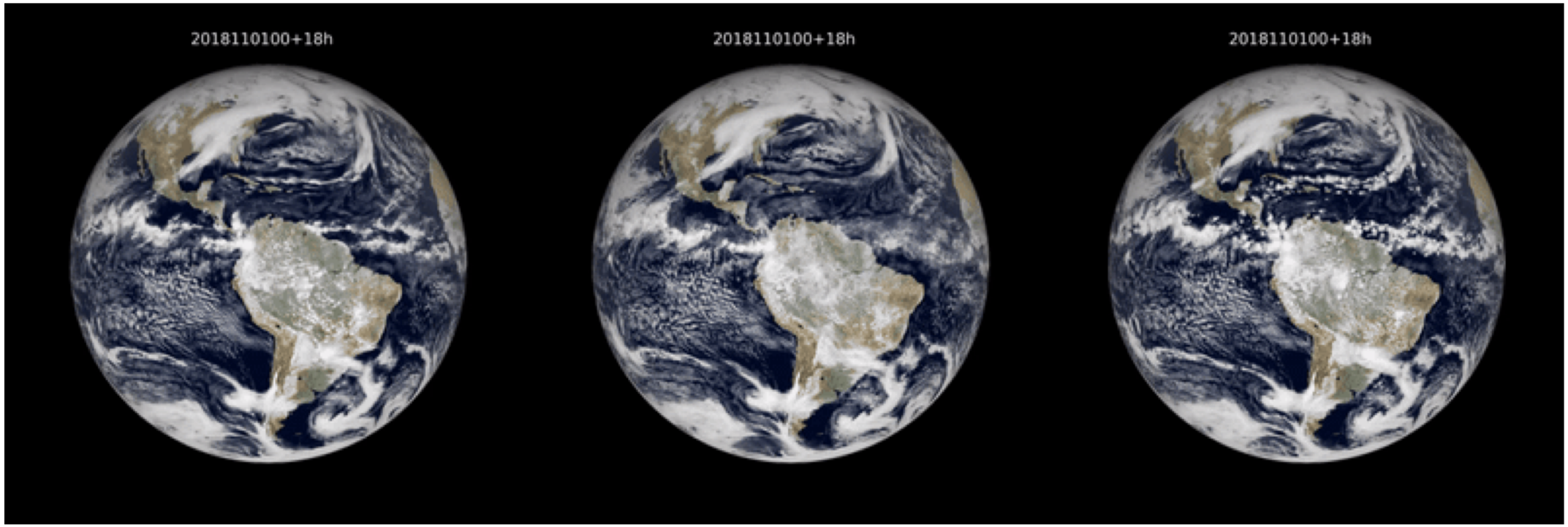


- 1920 nodes = 11520 GPUs (42% of Summit)
- energy of GPUs = 6x energy of CPUs per node
- fast alltoallv thanks to fat-tree network (like one electrical group across entire machine)

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.

3-hourly accumulated radiative fluxes at the top of the atmosphere



1.4 km **9 km** **9 km w/o deep**

A baseline for global weather and climate simulations at 1 km resolution

Nils P. Wedi[1], Inna Polichtchouk[1], Peter Dueben[1], Valentine G. Anantharaj[2], Peter Bauer[1], Souhail Boussetta[1], Philip Browne[1], Willem Deconinck[1], Wayne Gaudin[3], Ioan Hadade[1], Sam Hatfield[1], Olivier Iffrig[1], Philippe Lopez[1], Pedro Maciel[1], Andreas Mueller[1], Sami Saarinen[1], Irina Sandu[1], Tiago Quintino[1], Frederic Vitart[1]

(JAMES, 2020, avail online)

So are we (as a community) on a good trajectory ?

- Good progress on reduced precision use
- Good progress on optimisation/review efforts for the models, less so for data assimilation
- Sustained continuous efforts required to achieve single source portable solutions (DSLs)
- GPU accelerated machines available from Q4 2021, more efforts required to use them effectively
- AI success will also depend on efficient data handling and I/O capabilities
- Other approaches:
  - CLIMA: DG atmosphere/ocean written in Julia (for CPU/GPU), ML/AI LES applications (running in the Google Cloud) to feedback and improve global coupled climate model projections
  - MPAS/IBM: (unusual) high-resolution tailored grid over land (coarse over oceans) combined with IoT/Citizen observations to improve weather forecasts
  - Replace model components with AI, e.g. ECMWF, Vulcan/GFDL , but likely no operational use of AI model components in 2021 … ?

Appendix – additional contributed slides

Met Office

Programme of work is much bigger than just the model!

Re-writing the atmosphere model science

List of NGMS Projects

Gungho Atmospheric Science Project
- Develop atmospheric science aspects & deliver model scientifically as good as UM

LFRic Infrastructure Development
- Deliver infrastructure to replace the UM scalable for future platforms

LFRic Inputs
- Tools to ingest fixed & time-varying fields.
- Include initial conditions, ancillary fields and LBCs

LFRic Diagnostic Infrastructure
- Development of research diagnostics and research workflow capabilities

NG-Marine Systems
- Deliver scalable marine systems including ocean, sea-ice & wave models

NG-Coupling
- OASIS3-MCT coupled components

NG-DA
- NGMS-ready coupled atmos/ocean DA
- JEDI as a DA framework

NG-OPS
- Processing of NWP observational data for NG-DA

NG-VAT (Visualisation Analysis Tools)
- Support for visualisation and evaluation tools used by scientists

FAB Build System
- Development of new build systems for NGMS components

NG-R2O
- Support transition of NGMS capability from research to NWP operations

NG-Composition
- Coordination of aerosol & chemistry development within NGMS

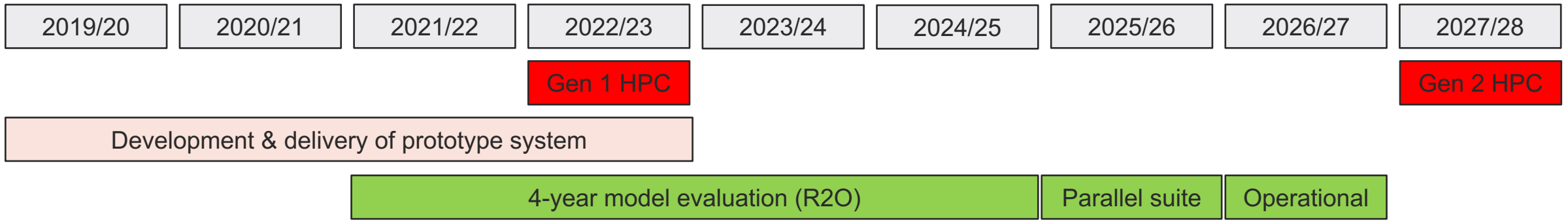
NG-R2C
- Support transition of capability from research to climate production

NG-ADAQ
- Development of dispersions models (e.g. NAME) for next generation computing

NG-VER
- Development of NWP verification capability for NGMS

Summary Programme timeline

- Operational NWP implementation of NGMS planned for ~2027
- For climate configs, CMIP8 will be first Met Office NGMS-based submission



Met Office

UM Physics

- LFRic uses existing UM Physics (+ code repository)
- Can run in single column model (SCM) mode to compare output
- LFRic uses k-first indexing while UM uses i-first
- i-first allows physics to work on a large segment of data giving a performance boost. LFRic has to use single columns as a segment
- Investigating transposing data in LFRic

```
do ij = 1, ncells
  do k = 0, nlayers - 1
    q(map(1,ij) + k) = rhs
  end do
end do
```

LFRic

```
do k = 1, nlayers
  do j = 1, nrows
    do i = 1, ncols
      q(i,j,k) = rhs
    end do
  end do
end do
```

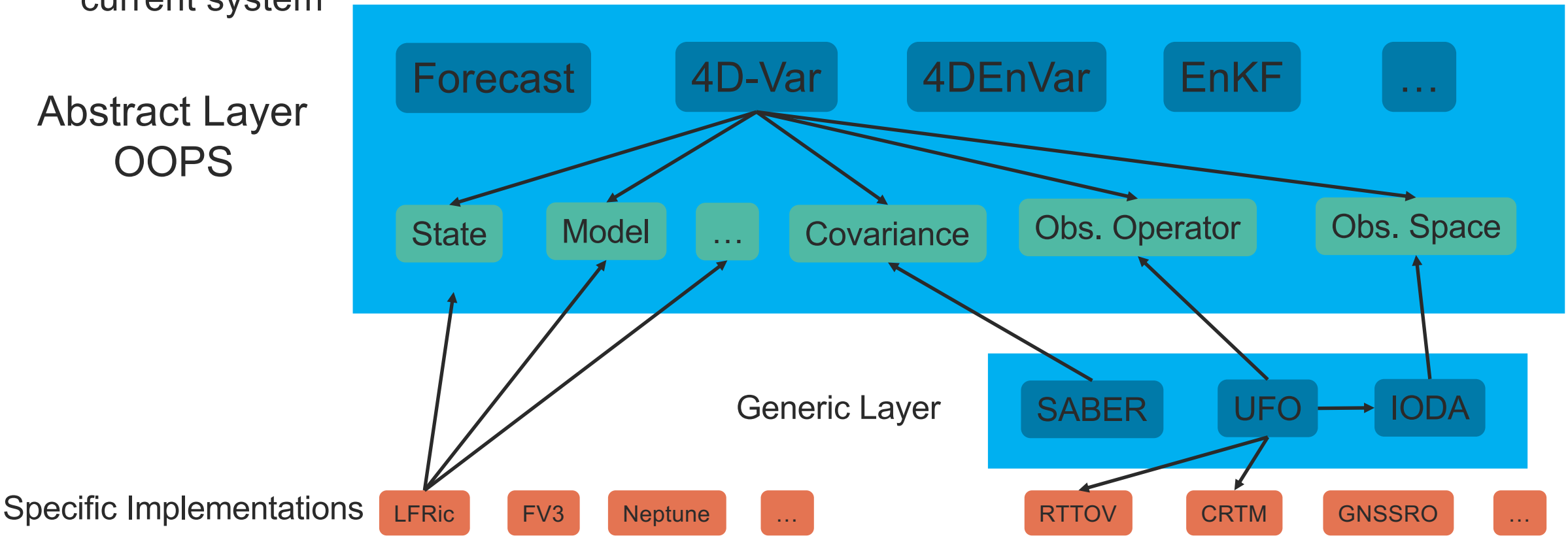
UM

NG-DA Plans

- Delivering the Next Generation Data Assimilation system for the LFRic model within JEDI. This includes developing:
- Background Error Covariances and Ensembles
  - Develop the background covariance model for the Unified Model (UM) and to reuse some of its building blocks for the LFRic background covariance model.
  - Based around the Atlas framework developed at ECMWF and included in OOPS.
- Next-Gen TL / Adj model
  - The dynamical core of the NG PF Model will be a linearisation of GungHo.
  - Perturbed physics from an ensemble of forecast differences contemporaneous with the analysis: localised ensemble tangent linear model (LETLM)
- Review requirements and solutions for land surface DA (partly) in JEDI

The Joint Effort for Data Assimilation Integration (JEDI)

- December 2019 NGMS Programme Board **prescribing that both NG-DA and NG-OPS adopt the JEDI framework**, with the aim of replacing like-for-like components within the current system



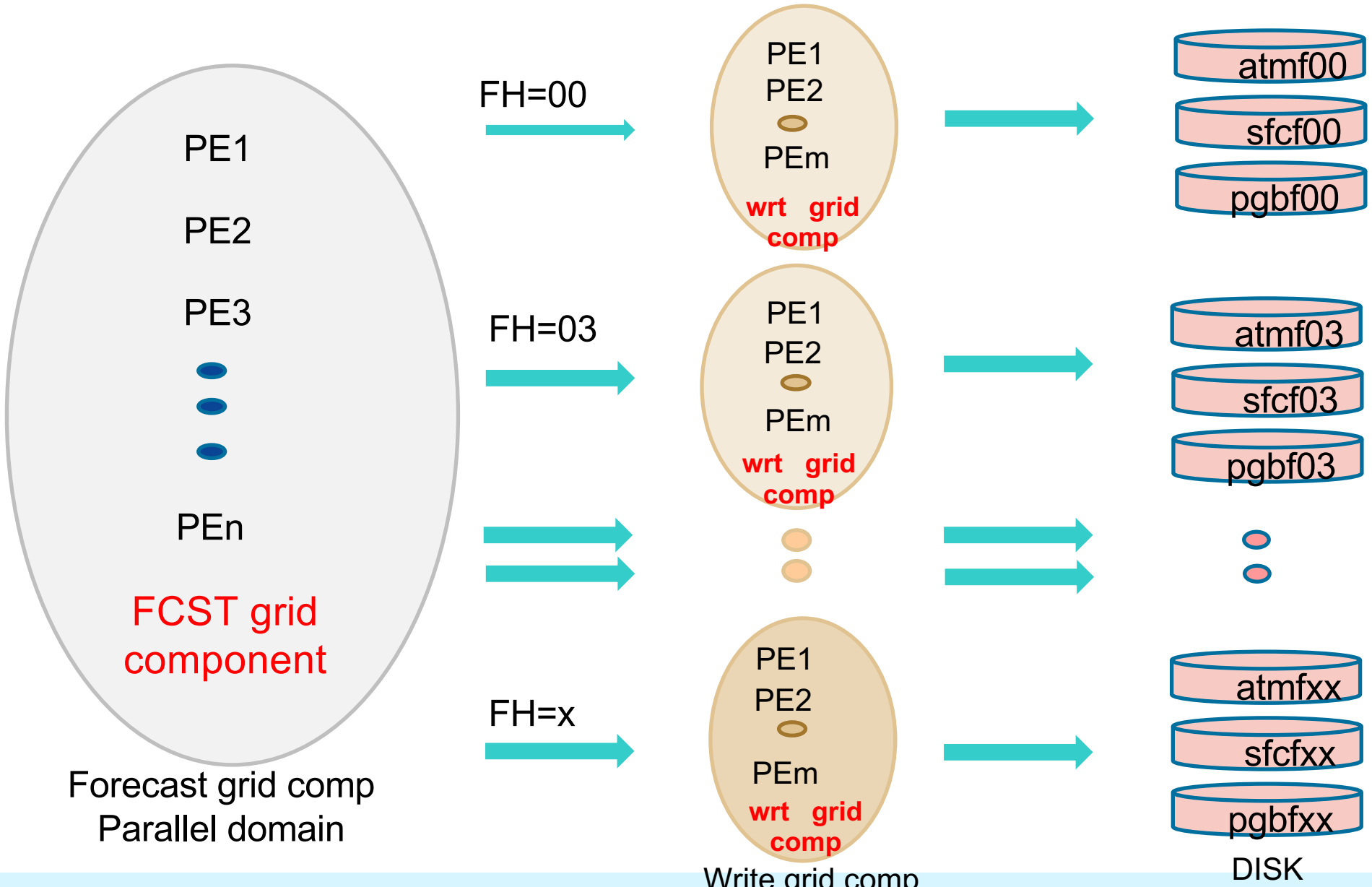
Computational performance improvements in GFSv16

Jun Wang, Jeffrey Whitaker, Edward Hartnett, James Abeles, Gerhard Theurich, Wen Meng, Cory Martin, Fanglin Yang, Russ Treadon, Jessica Meixner, George Vandenberghe, Arun Chawla

GFSv16 model infrastructure

- NCEP Global Forecast System version 16 is built with the NOAA Environmental Modeling System **(NEMS)** infrastructure. NEMS is built upon the Earth System Modeling Framework **(ESMF)** and National Unified Operational Prediction Capability **(NUOPC)** Layer code and conventions.
- It runs at the C768 (~13km) horizontal resolution, and has 127 layers in the vertical extending up to the mesopause (~80km).
- The NOAA WAVEWATCH III model (WW3) is coupled to the atmosphere.

Parallelization of GFS Write Grid Component



I/O Changes and Inline Post

- To reduce I/O footprint, the data format of forecast history files was changed from plain binary NEMSIO to compressed NetCDF format.
- Scalar linear packing along with the deflate HDF5 filter is applied. Data were quantized before being written out.
- The NetCDF library was updated to support parallel writing of compressed netCDF data.
- An interface was developed to transfer and save forecast variables from computing nodes to the Write Grid Component for running an inline post-processing software to produce products in GRIB2 format. It reduced I/O activity in the forecast system

Writing Compressed Data in NetCDF Format

C768L127f cst output	Nemsio No compression	Netcdf No compression	Netcdf Lossless (deflate=1,nbit=0)	Netcdf Lossy (deflate =1, nbit=20)	Netcdf Lossy(deflate =1,nbit=14)	Netcdf Lossy (deflate=1, nbits=14),parall el writing, default decomposition chunksize	Netcdf Lossy (deflate=1, nbits=14),par allel writing Layer chunkcsize
A 3D file size (total fcst)	33.6GB (7TB)	33.6GB (7TB)	23.6GB (5TB)	13.5GB (2.8TB)	6.3GB (1.3TB)	6.3GB (1.3TB)	**6.3GB (1.3TB)**
Write Time	79s	300s	960s	680s	400s	43s	**34s**

Future work

It is critical to improve model computational performance when running with resource consuming high resolution model

- Check scalability of each subcomponent
- Develop efficient IO strategy to meet the requirements for next generation of high resolution weather and climate forecast runs
  - **What to output**: add new capabilities to write out model outputs on different grid with different content and at different frequency
  - **How to output:**
    - Work with community to improve compression algorithms and maintain required precision and writing
    - Improve performance of writing restart files

Approaches to SL-AV code optimization

New version, ~10km resolution, 104 vertical levels.
Limitation: < 4000 cores should be used for operations

- Switching most time-consuming part to single precision (semi-Lagrangian advection)
- Optimizing the vector length in parameterizations of subgrid scale processes
- Reduction of data amount in transpositions by making them single precision instead of double precision

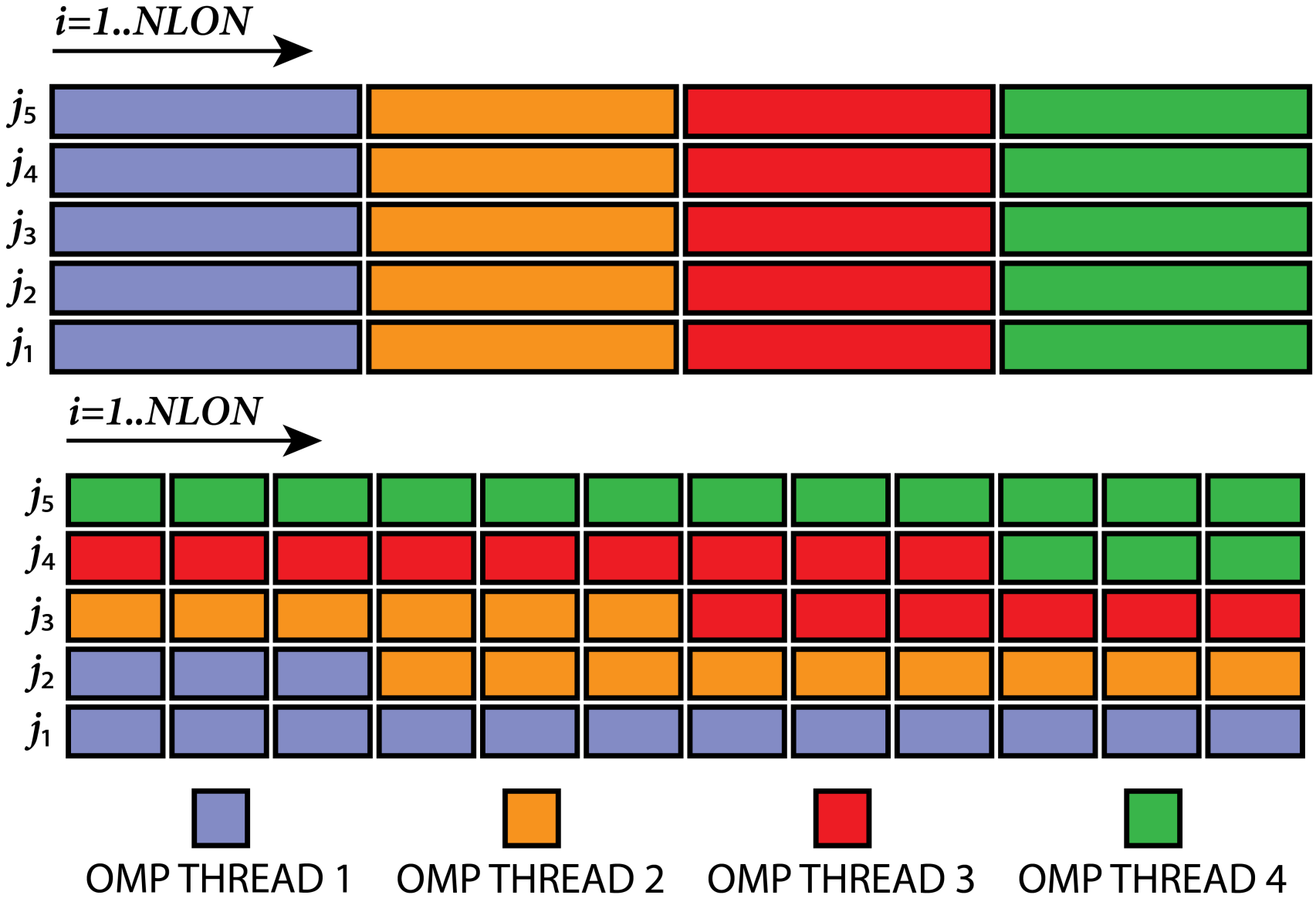
Optimization of vector length in parameterizations

- (i,k,j) index ordering in most part of RHS code where i – longitude, k – vertical index, j – latitude. Typical local arrays are dimensioned with (Imax,Kmax)
- i is the variable of OpenMP parallelization

(range 1:3600) – the vector length is $3600/N_{openmp}$

- The code already uses thread-local arrays
- Splitting this dimension into smaller parts improves memory access

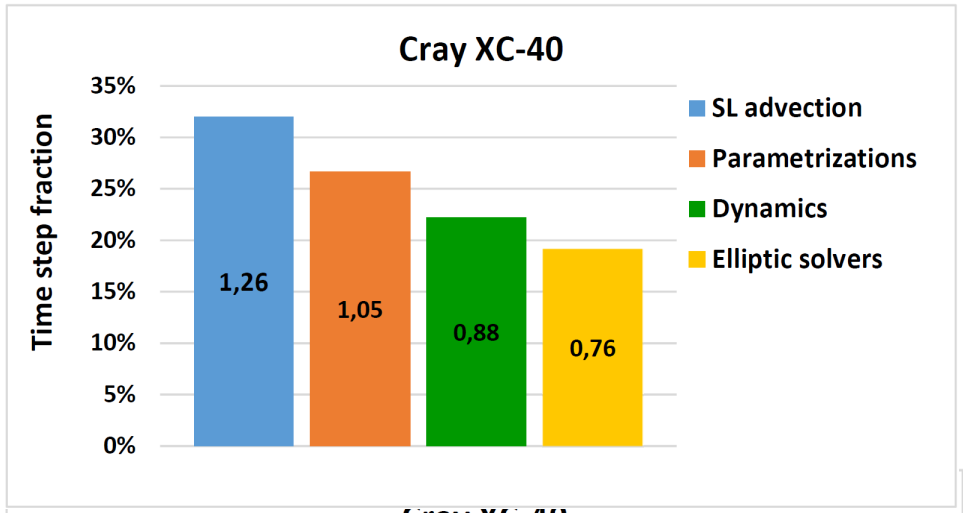
Optimization of vector length (old and new partition)

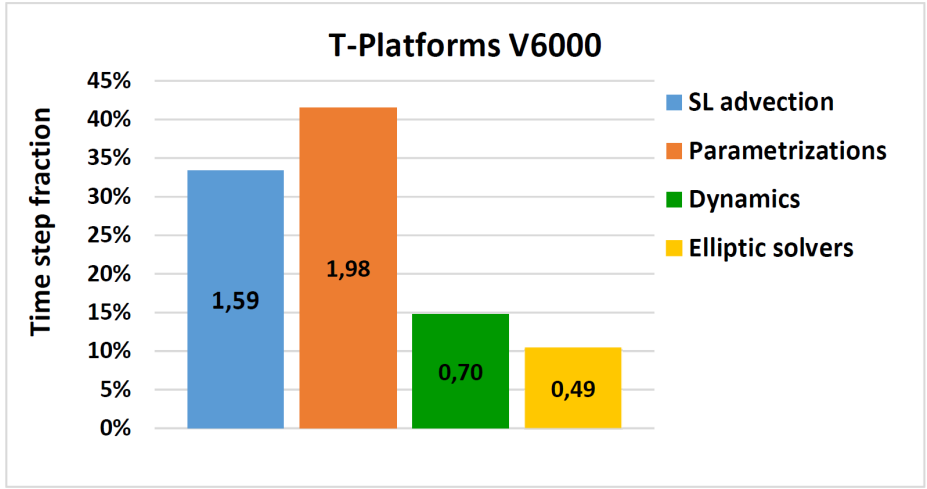


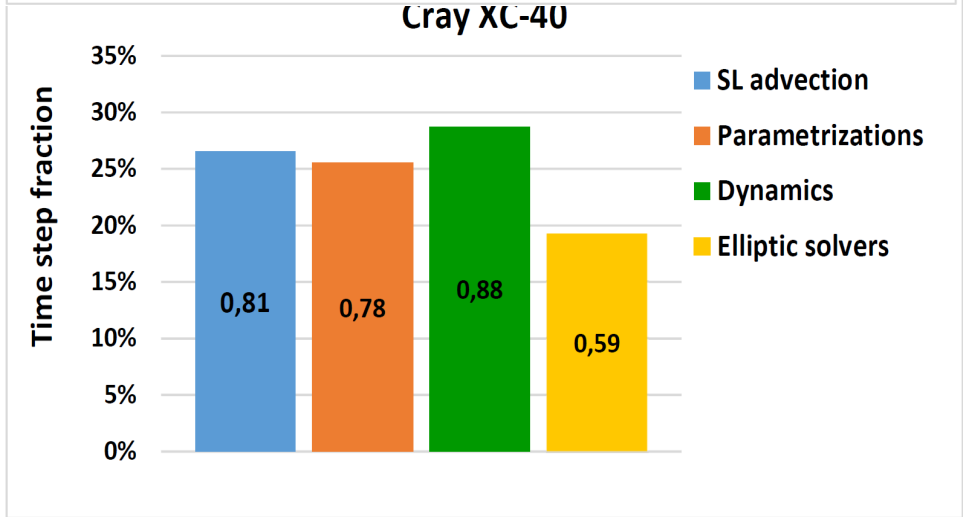
Percentage of time used in different parts of SL-AV model code while using 3888 cores at Cray XC40 (left) and 3880 cores at T-Platforms V6000 (right); before (top) and after (bottom) optimizations.

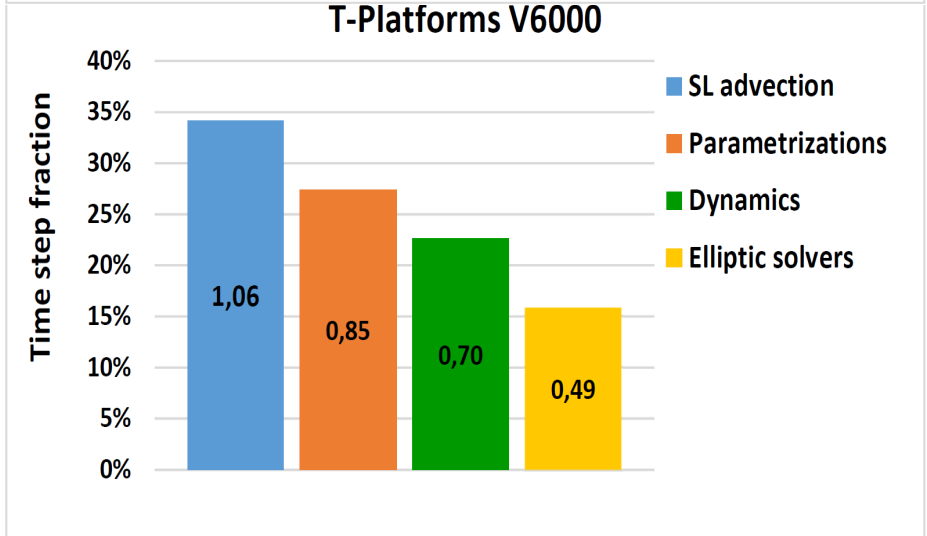
~23% reduction of wall clock time per 24hrs forecast

Number inside the column denotes the wall-clock time of respective code part (in seconds).









CMC Progress on Scalability and Exascxale: Dynamics

- As a centre with relatively small computational resources, exascale computing at the CMC still some distance in the future; however, scalability is a practical concern that is the focus of several ongoing projects related to the dynamical core:
  - Develop a spatial discretization based on the flux reconstruction approach (it has the advantage of being local, accurate,flexible with good conservation properties);
  - Time-stepping based on exponential propagators (advantage: Eulerian with large time-steps). Collaboration with the University of California at Merced, Lawrence Livermore National Laboratory, Southern Methodist University and Università degli studi di Verona;
- The applicability of these methods to GPU-based architectures is under investigation:
  - The two approaches above will be tested on GPUs in the context of an NVIDIA hackathon (probably in December);
  - Our UC Merced collaborators are also testing this on GPUs;

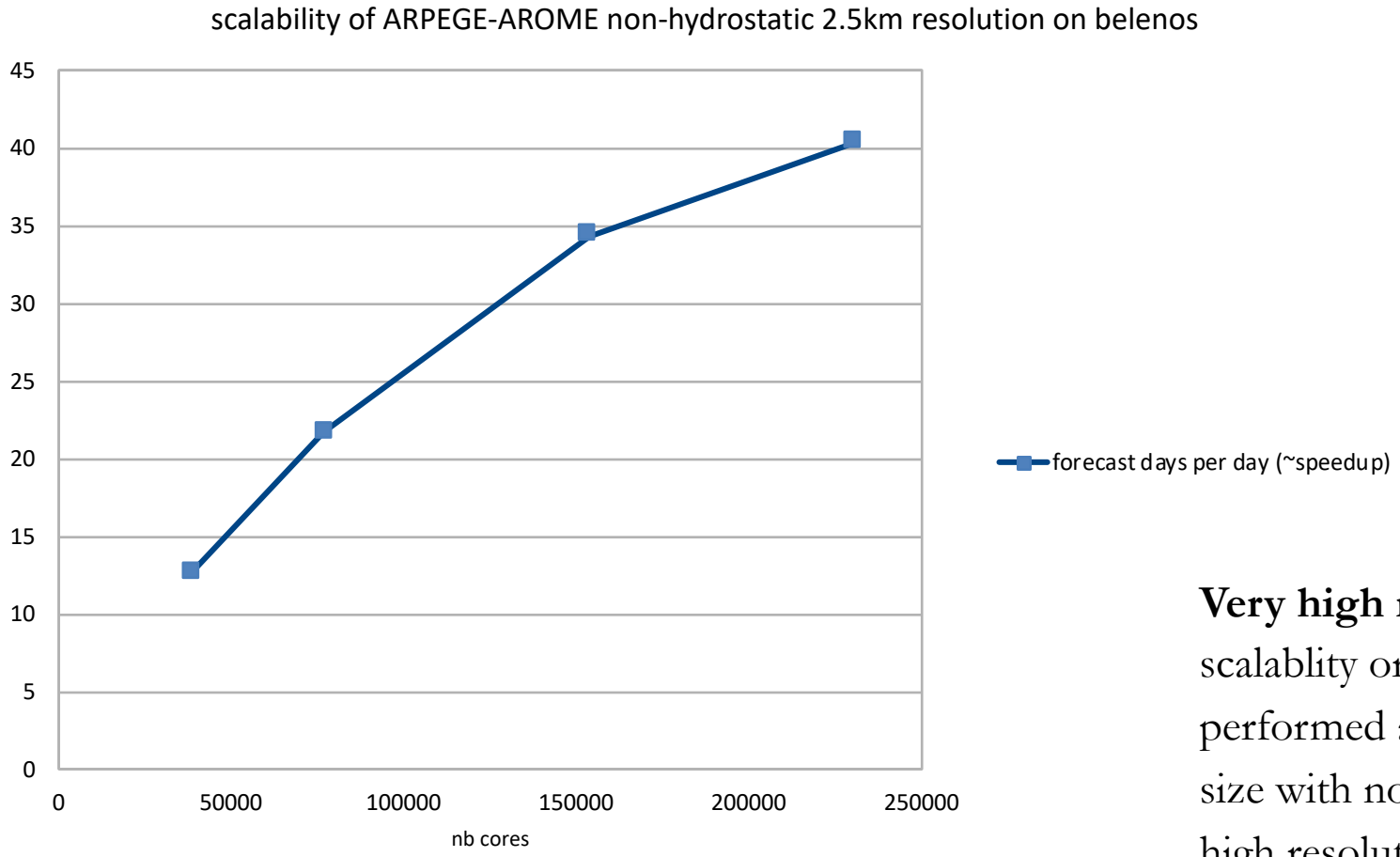
CMC Progress on Scalability and Exascale: Infrastructure

- One of the current (avoidable) limits on scalability is model I/O: implementing an I/O server is an infrastructure development priority
- Work on better compression schemes and rules to lower I/O impacts on communications and storage
- Re-engineering of the coupling system for exascale level of scalability and performance

CMC Progress on Scalability and Exascale: Physics

- Profiling of the physics package to identify regions of the (large) code base that have particularly poor performance characteristics
- Optimize single-processor performance: all parameterizations still use the independent column approximation, and are thus embarrassingly parallel
- A project has been initiated with the Mila – Quebec AI Institute to investigate utility of machine learning for algorithm replacement, beginning with radiative transfer calculations

Meteo France



Very high resolution global forecasts : To assess current scalablity or ARPEGE global model, simulations were performed at a very high global resolution of 2.5km mesh size with non-hydrostatic dynamics and the physics of our high resolution LAM model AROME.

The figure below shows the number of forecast days per day as a function of the number of cores used. At the maximum, on 230400 nodes (80% of the computing machine) 40 days per day was obtained. Scalability is still acceptable.