High performance computing (HPC) readiness and the road to Exascale
compiled for WGNE by Nils Wedi, ECMWF

In the past few years WGNE has routinely reviewed activity in this area among WGNE members by monitoring the evolution of NWP systems and hardware (*WGNE NWP systems, 2010-2020*) and more recently by actively monitoring developments towards future architectures and beyond (*WGNE scalability, 2018-2020*). This has led to reviews of the HPC readiness of the weather & climate modelling community, for atmosphere and ocean models in particular, and from which this document is derived. This is followed up by a list of available software developments that can help to transition existing models towards future HPC architectures.

The enhanced focus on HPC readiness is triggered by a changing landscape in computing towards more energy-efficient HPC due to the demise of Moore's law and Dennard-scaling and the explosion of data challenges (*Bauer et al, 2020*; *Bauer et al, 2021*). This is at a time where climate change and weather extremes forecasting experiences a renewed urgency with critical time-to-solution computing constraints and a growing complexity of the modelling and assimilation systems. At the same time, the emergence of a variety of accelerator technologies, hierarchies of memory, and intelligent networking solutions is fuelled by a multi-billion dollar HPC market targeting gaming and artificial intelligence applications. The latter in particular imposes challenges on the weather & climate community as the computational efficiencies (floating point rates) in weather & climate models are often low for some of the necessary algorithmic patterns. However, the large global complexity and large problem sizes have led to the widespread adoption of massively parallel (MPI) codes, combined with multi-threaded (OpenMP) solutions. Emerging HPC devices are often optimized for low-precision computations, which is challenging for existing weather & climate models that have taken double precision for granted over the past 30 years. There is no question however, that the next HPC platforms are equipped with vector or scalar accelerators of some form, typically 4-8 highly performing accelerator devices combined with many-core CPUs.

With increasing problem sizes, global Earth System simulations' output is several orders of magnitude larger than the available observations, involving 10s – 100s billion points. Therefore Input/Output (I/O) handling has become a serious constraint, both in terms of efficiency during the simulation runtime and in terms of the speed and cost of storage, as well as the suitability of "online" access for emerging technologies such as unsupervised machine learning.

Traditional performance enhancement recommendations

- Hybrid parallelisation MPI + X (OpenMP/OpenACC/OpenX?) will continue to be important for the next 10 years
- Reduce or eliminate global communications where possible, in particular when considering global diagnostics and postprocessing
- Use flexible tiling approaches for cache efficiency
- Targeting the efficiency of advection and other key algorithmic patterns, often with the help of isolating these patterns in weather & climate dwarfs (*Mueller et al, 2019*)

has proven successful. Interacting with vendors and academia has also proven useful here.

- Continue to challenge the time-to-solution (two-time-level schemes) and energy-to-solution efficiency of each Earth system model (ESM) component dynamical core, in particular in the areas of non-hydrostatic, quasi-uniform modelling on the sphere, global conservation, multi-tracer advection efficiency (and with increasing use of chemical, carbon and hydrological cycle applications)
- Selective directive-based porting to GPUs and use of special (e.g. cuFFT, cuBLAS) libraries can be a faster, even if a non-portable, route to effective GPU use in the short-term
- Review and use reduced/mixed precision where appropriate as it increases cache memory and communication bandwidth
- Target I/O performance, separate and overlap data flow (e.g. disconnect from raw disk speed), e.g. I/O server model, review data compression, review accuracy needs, reduce model output and recompute
- Focus on ESM components and their coupling to the atmosphere (e.g. ocean and wave model scalability, external coupler vs inline coupling), especially in the context of fully coupled data assimilation
- Advance science and find solutions for ocean model scalability and performance bottlenecks, such as barotropic mode coupling of the ocean surface, the stiffness of sea-ice rheology, and multi-tracer efficiency and ocean biogeochemistry efficiency

Basic discretisation and data structures recommendations for performance and portability

- Consider (a rewrite) of the model dynamical core
- Introduce generic data structures that can support different function spaces
- Develop a new dynamical core that avoids structural bottlenecks (quasi-uniform parallel distribution, avoiding pole points, uniform application of recurring algorithmic patterns, clear and flexible data (flow) structures, writing a model as a subroutine such that it can be integrated into another control flow structure as provided by JEDI/OOPS data assimilation or ESMF frameworks)
- Consider a separation of concerns using hierarchically structured code, and interoperable source-to-source code translations, e.g. by using domain-specific language (DSL) toolchains, e.g. with GT4PY, PsyClone, kokkos

Emerging Topics

- AI methodologies for replacing and accelerating time-critical parts of models by substantially enhancing floating point performance of selected algorithms
- I/O optimization utilizing non-volatile memory (NVM), high-bandwidth memory (HBM) devices, AI optimized data flow
- Low-power acceleration (e.g. FPGA, RISC-V)
- Use of network intelligence for on-the-fly streaming processing
- Resilience of algorithms against soft or hard failures
- Combine data locality and large-time-step algorithms
- Parallel-in-time algorithms
- Exponential time integration

- Task-based asynchronous execution and more generally overlapping of compute an communicate
- Infrastructure for assimilating citizen observations
- Cloud and containerised computing, data processing and analysis
- Quantum computing (e.g. use of exponential acceleration via QFFT, QBLAS)

Publicly available software repositories:

Coupling

ESMF Earth System Modelling framework
https://www.earthsystemcog.org/projects/esmf/download/
Several US modelling centres have build on this framework to couple to other Earth System components, or are considering it, such as NCAR (CESM), NRL(NEPTUNE), NCEP(GFS).

Parallel ESM coupler OASIS-MCT
https://portal.enes.org/oasis
The UKMetOffice and many climate models use this coupling framework to couple to other ESM components with different thread/core requirements (e.g. ocean, sea-ice, biogeochemistry)

Adaptation, data structures and infrastructure

ECMWF makes selected software available at
https://github.com/ecmwf
Useful software packages for NWP. This includes the general data structure framework Atlas, which is the basis of ECMWF's and MeteoFrance's new dynamical core developments, it is also considered for use at the UKMetOffice and at JCSDA in JEDI.

Programming model in C++ for writing performance portable applications.
https://github.com/kokkos/kokkos
The new DOE E3SM/ESMD model is based on this framework.

GT4Py is a Python library for generating high performance implementations of stencil kernels from a high-level definition using regular Python functions.
https://github.com/GridTools/gt4py
FV3 and IFS-FVM, ad ICON explore GT4PY and rewrite parts of their models in Python. This has the added advantage to link to AI technologies (e.g. tensor flow https://www.tensorflow.org/)

ClimateMachine, a new Earth System Model written in Julia
https://github.com/CliMA/ClimateMachine.jl
The CLIMA project rewrites ocean and atmosphere models based on discontinuous Galerkin technologies in Julia (adaptable to CPU and GPU).

STFC PsyClone, Fortran DSL
https://github.com/stfc/PSyclone

The UKMetOffice next-generation model is based on this framework.

Data assimilation

JCSDA makes JEDI data assimilation framework available at
https://github.com/JCSDA
A generic data assimilation framework originating at ECMWF under the name OOPS, has been substantially developed at JCSDA and is planned to be used in many modelling centres in the US and the UKMetOffice. ECMWF also continues to use OOPS.

I/O & diagnostics

XIOS data server
https://portal.enes.org/models/software-tools/xios
XIOS, or XML-IO-Server, is a library dedicated to I/O management in climate codes.

ESMVal toolbox
https://www.esmvaltool.org/
A community diagnostic and performance metrics tool for routine evaluation of Earth system models in CMIP.

Climate data operators (cdo)
https://portal.enes.org/models/software-tools/cdo
Command line easy manipulation of netcdf climate data

Community platform for big data science
https://pangeo.io/
The Pangeo software ecosystem involves open source tools such as xarray, iris, dask, jupyter, and many other packages.

Parallelisation

The OpenMP API specification for parallel programming
https://www.openmp.org/
Defined by a group of major computer hardware and software vendors and major parallel computing user facilities, standardise directive-based multi-language high-level parallelism that is performant, productive and portable.

Open-MPI
https://www.open-mpi.org/
Open source Message Passing Interface (MPI) implementation that is developed and maintained by a consortium of academic, research, and industry partners.

References:

Bauer et al, Extreme-scale Computing and Data Handling - the Heart of Progress in Weather and Climate Prediction, https://public.wmo.int/en/resources/bulletin/extreme-scale-computing-and-data-handling-heart-of-progress-weather-and-climate, 2020.

Bauer et al, The digital revolution of Earth-system science, Nature Computational Science, 2021, in review.

Mueller et al, The ESCAPE project: Energy-efficient Scalable Algorithms for Weather Prediction at Exascale, Geosci. Model Dev., 12, 4425–4441, 2019, https://doi.org/10.5194/gmd-12-4425-2019.

WGNE NWP systems, http://wgne.meteoinfo.ru/nwp-systems-wgne-table/wgne-table/, 2010-2020.

WGNE-33 scalability, HPC trends and the scalability of atmosphere and ocean models, http://wgne.meteoinfo.ru/wp-content/uploads/2018/10/WGNE33_WEDI__scalability_atm_oce.pdf, 2018.

WGNE-34 scalability, Atmosphere and ocean scalability and HPC readiness, http://wgne.meteoinfo.ru/wp-content/uploads/2019/10/WED_Wedi_WGNE34_scalabilitymixed.pdf. 2019.

WGNE-35 scalability, Atmosphere and ocean scalability and HPC readiness, http://wgne.meteoinfo.ru/wp-content/uploads/2020/10/WGNE35_Wedi_scalability.pdf. 2020.