Towards Exascale - members review

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

Contributions from: INPE Brazil, Canadian Meteorolgical Centre (CMC), Chinese Meteorological Agency (CMA), Meteo-France (MF), LMD and IPSL, German Weather Service (DWD), Japan Meteorolgical Agency (JMA), Hydrometeorological Centre of Russia (HMC), UK MetOffice, Navy Research Laboratory (NRL), National Centre Atmospheric Research (NCAR), National Weather Service (NOAA/NCEP), ECMWF, WMO Research Board

The race to exaxcale computing

- Weather and Climate models are a good example of applications that aim to exploit exascale.
  - "…is also representative of other simulation domains that rely on grid-based solvers of partial differential equations, such as seismology and geophysics, combustion and fluid dynamics in general, solid mechanics, materials science, or quantum chromodynamics. "
- They are in general "low arithmetic intensity" applications.
- A possible roadmap is:
  - Long-term scientific aim: a 100m global model
  - The practical interim goal post is a 1km global model
    - This would require a ~100-fold improvement in speed
    - Combination of architecture design and algorithmic design

Modified from P.L. Vidale talk at DiRAC workshop 2021 London

Reflecting on the Goal and Baseline for Exascale Computing: A Roadmap Based on Weather and Climate Simulations

Thomas C. Schulthess
ETH Zurich, Swiss National Supercomputing Centre

Peter Bauer
European Centre for Medium-Range Weather Forecasts

Nils Wedi
European Centre for Medium-Range Weather Forecasts

Oliver Fuhrer
MeteoSwiss

Torsten Hoefler
ETH Zurich

Christoph Schär
ETH Zurich

Table 1. Ambitious target configuration for global weather and climate simulations with km-scale horizontal resolution accounting for physical Earth-system processes, and with today's computational throughput rate.

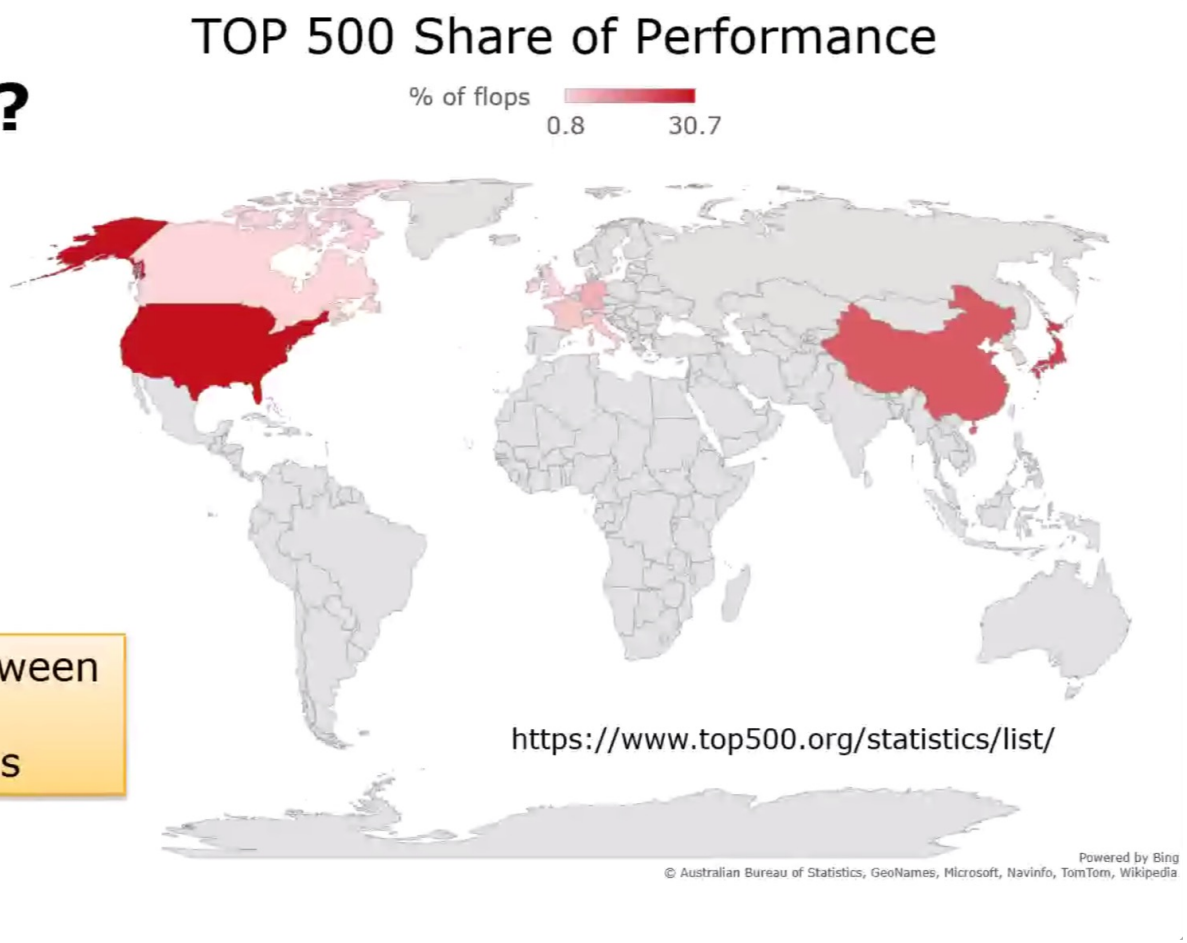
Horizontal resolution	1 km (globally quasi-uniform)
Vertical resolution	180 levels (surface to ~100 km)
Time resolution	0.5 min
Coupled	Land-surface/ocean/ ocean-waves/sea-ice
Atmosphere	Non-hydrostatic
Precision	Single or mixed precision
Compute rate	1 SYPD (simulated years per wall-clock day)

WMO perspective

The Benefits of Exascale—for Who?

- Over 98% of computing power worldwide is in Europe, Asia, and North America
- Over 72% belongs to China, Japan, and the U.S.
- Regions most at risk from climate change have few or no computing resources

Goal: Identify critical gaps between members with and without significant computing resources



WMO OMM

DESTINATION EARTH

EUROHPC: €8 BILLION PROGRAMME TOWARDS EXASCALE

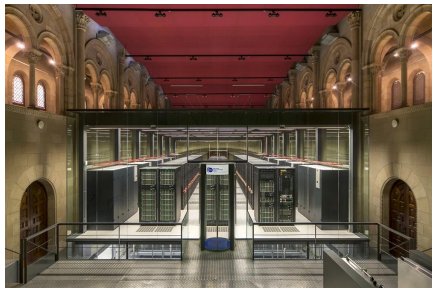
#EuroHPC (high performance computing) **Joint Undertaking**

The European High Performance Computing Joint Undertaking (EuroHPC JU) will pool European resources to develop top-of-the range exascale supercomputers for processing big data, based on competitive European technology.

Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Montenegro, the Netherlands, North Macedonia, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland and Turkey.

EuroHPC Joint Undertaking

- 3 large (O(100PFlops)) supercomputers in Finland, Italy, Spain
- 5 smaller ones (size of Archer in UK) in Luxembourg, Slovenia, Portugal, Czech Republic, Bulgaria
- 1-2 high-end supercomputers (~1000 Pflops) by 2024



MareNostrum 5

~314 Pflops

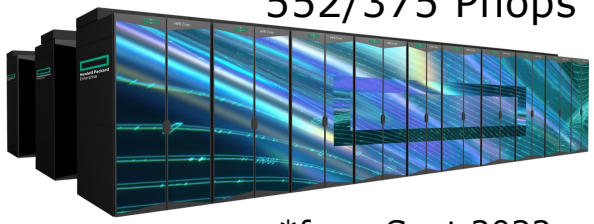
*from July 2023

Supercomputers

Currently six EuroHPC supercomputers are under construction across Europe:

LUMI

LUMI Finland, 552/375 Pflops



*from Sept 2022

© HPE

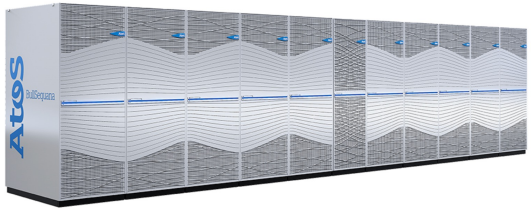
The LUMI system will be a Cray EX supercomputer supplied by Hewlett Packard Enterprise (HPE) and located in Finland.

Sustained performance:	375 petaflops
Peak performance:	552 petaflops
Compute partitions:	GPU partition (LUMI-G), x86 CPU-partition (LUMI-C), data analytics partition (LUMI-D), container cloud partition (LUMI-K)
Central Processing Unit (CPU):	The LUMI-C partition will feature 64-core next-generation AMD EPYC™ CPUs
Graphics Processing Unit (GPU):	LUMI-G based on the future generation AMD Instinct™ GPU
Storage capacity:	LUMI's storage system will consist of three components. First, there will be a 7-petabyte partition of ultra-fast flash storage, combined with a more traditional 80-petabyte capacity storage, based on the Lustre parallel filesystem, as well as a data management service, based on Ceph and being 30 petabytes in volume. In total, LUMI will have a storage of 117 petabytes and a maximum I/O bandwidth of 2 terabytes per second
Applications:	AI, especially deep learning, and traditional large scale simulations combined with massive scale data analytics in solving one research problem
Other details:	LUMI takes over 150m2 of space, which is about the size of a tennis court. The weight of the system is nearly 150 000 kilograms (150 metric tons)

*from March 2023

LEONARDO

Leonardo Italy: 322/249 PFlops



© Atos

LEONARDO will be supplied by ATOS, based on a BullSequana XH2000 supercomputer and located in Italy.

Sustained performance:	249.4 petaflops
ak formance:	322.6 petaflops
mpute titions:	Booster, hybrid CPU-GPU module delivering 240 PFlops, Data-Centric, delivering 9 Pflops and featuring DDR5 Memory and local NVM for data analysis
ntral cessing it (CPU):	Intel Ice-Lake (Booster), Intel Sapphire Rapids (data-centric)
aphics cessing it (GPU):	NVIDIA Ampere architecture-based GPUs, delivering 10 exaflops of FP16 Tensor Flow AI performance
rage acity :	Leonardo is equipped with over 100 petabytes of state-of-the-art storage capacity and 5PB of High Performance storage
plications:	The system targets: modular computing, scalable computing applications, data-analysis computing applications, visualization applications and interactive computing applications, urgent and cloud computing
er details:	Leonardo will be hosted in the premises of the Tecnopolo di Bologna. The area devoted to the EuroHPC Leonardo system includes 890 sqm of data hall, 350 sqm of data storage, electrical and cooling and ventilation systems, offices and ancillary spaces

ECMWF

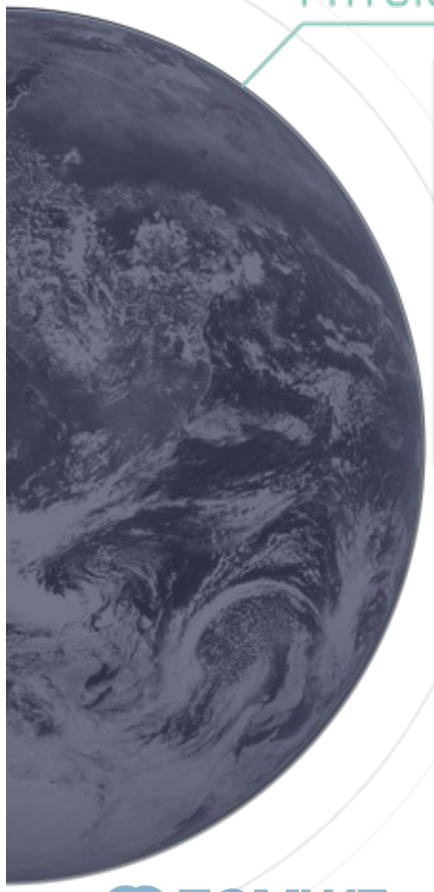
a maximum of **10% of the Union's access time is dedicated to strategic initiatives**

Courtesy P.L. Vidale

DESTINATION EARTH

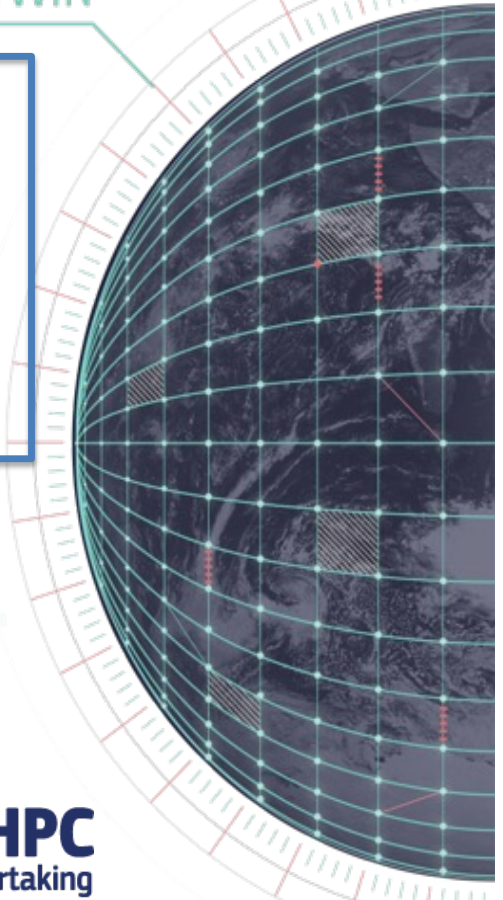
EU's Destination Earth (DestinE) initiative

Funded by the European Union



PHYSICAL WORLD

DIGITAL TWIN



Entrusted entities

ECMWF esa EUMETSAT

Key elements

- Digital Twin Engine
- Digital Twins
- Data lake
- Core platform

2021-2023

- Operational cloud-based platform
- First two digital twins

2023-2025

Platform integrates the next operational digital twins and offers services to public sector users

2025-2027+

Towards a full "digital twin of the Earth" through a convergence of multiple digital twins on the platform

A European Green Deal (2019)

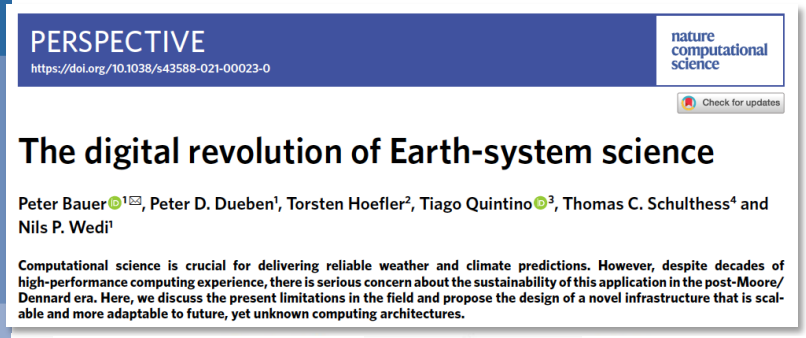
A European strategy for data (2020)

Shaping Europe's digital future (2020)

EuroHPC Joint Undertaking

ECMWF

DestinE's Digital Twins: extreme computing challenges



PERSPECTIVE
https://doi.org/10.1038/s43588-021-00023-0

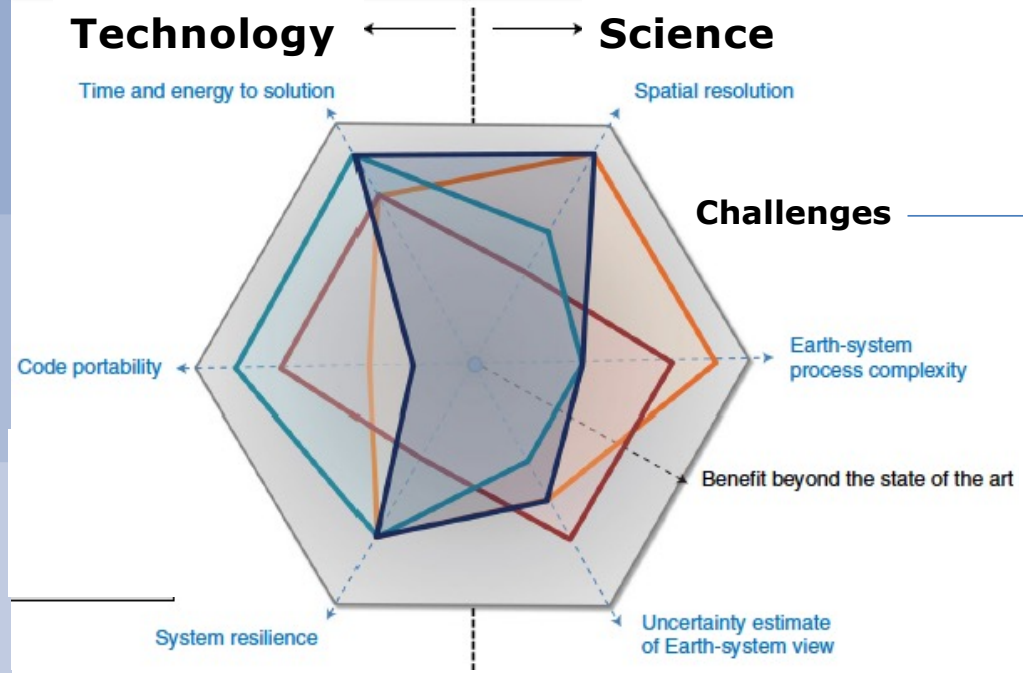
nature computational science

Check for updates

The digital revolution of Earth-system science

Peter Bauer[1], Peter D. Dueben[1], Torsten Hoefler[2], Tiago Quintino[3], Thomas C. Schulthess[4] and Nils P. Wedi[1]

Computational science is crucial for delivering reliable weather and climate predictions. However, despite decades of high-performance computing experience, there is serious concern about the sustainability of this application in the post-Moore/Dennard era. Here, we discuss the present limitations in the field and propose the design of a novel infrastructure that is scalable and more adaptable to future, yet unknown computing architectures.



FEATURE 10 October 2018

Could the world's mightiest computers be too complicated to use?

China, Japan and the US are racing to build the first exascale computer – but devising programmes clever enough to run on them is a different story

Challenges →

Solutions

- Numerical methods, algorithms, data structures
- Machine learning
- Programming models
- Heterogeneous processing, memory, interconnect technology

EuroHPC Joint Undertaking

… make sure that technology is not running away from us!

WMO perspective

Summary of Recommendations

For the Research Board and WMO Members

We recommend urgency in dedicating efforts and attention to disruptions associated with evolving computing technologies that will be increasingly difficult to overcome, threatening continued advancements prediction capabilities.

The increasing scientific and computing complexity will require major efforts to adapt or rewrite earth system prediction models. In addition to scientific accuracy, models must be developed for performance, portability, and productivity.

The cost of computing resources, power consumption, and the related carbon footprint must be considered along with the benefit of improved predictability. Requirements to make data centers carbon neutral are already in force in a growing number of countries.

Scientists, model developers, computer scientists and software engineers need to work as equal partners on design, development, and maintenance of applications to overcome scientific, computing, and data challenges.

A data-in-place strategy is needed to support the increase in data volume from observations, model and ensemble output, and post processing. This will require co-location of HPC and data, with methods to access, extract, analyze, visualize, and store data by requesting processes & users.

Executive Summary of members' reports

- Dynamical core improvements or entirely new developments, in particular preconditioning of solvers, alternative time integration methods, new/alternative spatial discretisations, multiple grids for ESM components and coupling methods
- Targeting atmosphere, ocean and other components
- MPI optimisations, load-imbalance vs speed of communication
- Reduced (single/mixed) precision, memory needs optimisations
- GPU adaptation efforts, major code refactoring efforts, hybrid CPU-GPU compute, DSL use, shared challenge of different programming models
- Machine learning algorithms that provide higher arithmetic intensity to replace components of ESMs
- Analysing the energy efficiency (and hence cost) of execution
- IO optimisation, data compression, IO-server, and refactoring of data production pipelines
- End-to-end data production and consumption

Hydrometeorological Centre of Russia – SL-AV model

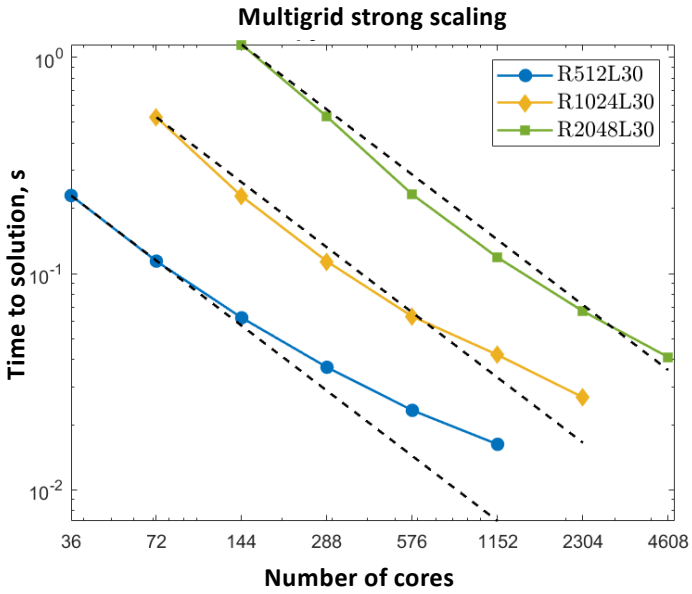
Memory usage optimization. Change of model state vector storage from a single array with the size $(6\,NLEV + 5, NLON, NJ)$ to separate arrays with the size $(NLEV, NLON, NJ)$.

- 70km model – 16-22% runtime reduction
- 10km model – 30% runtime reduction

Single precision computations. Working on further implementation of single precision computations.

Modification of grid-point space approximations for reduced grid. Testing of local horizontal schemes instead of Fourier space-based approximations within a 2d SWE prototype. Good results within idealized testcases.

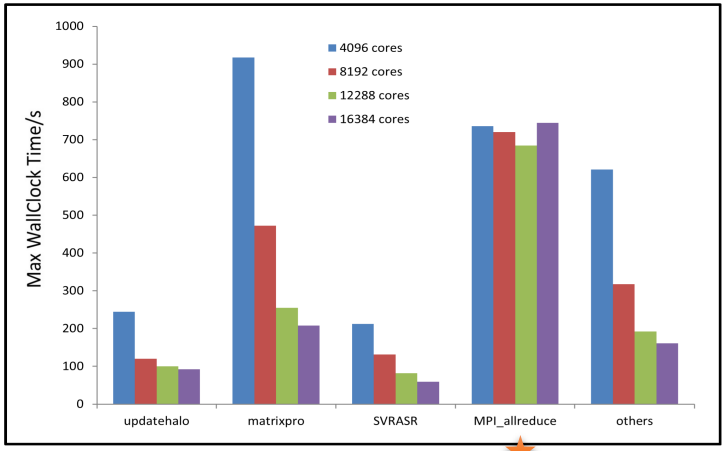
New parallel geometric multigrid solver for reduced grid. Preliminary strong scaling results. Test problem at 20km reduced grid with 30 vertical levels scales at least to 4608 cores.



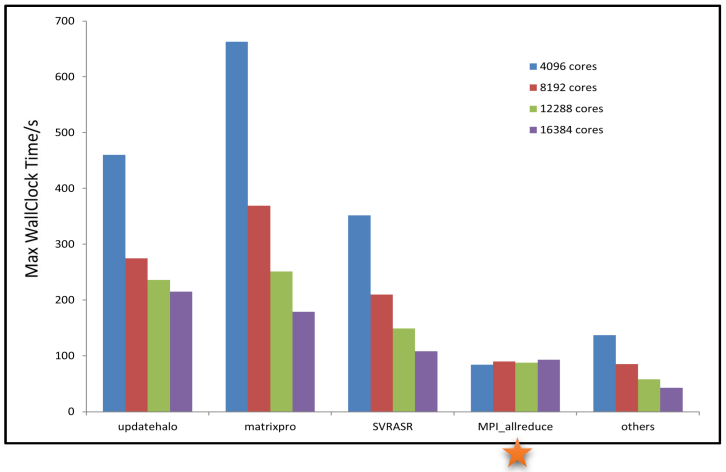
Courtesy M.Tolstykh and G.Goyman

Helmholtz solver (PCSI) for GRAPES-GFS

PGCR



PCSI



Preconditioned Classical Stiefel Iteration (PCSI)

$\mathbf{x}_0$, *estimated eigenvalue boundary* $[\nu, \mu]$

1, $\alpha = \dfrac{2}{\mu - \nu}, \beta = \dfrac{\mu + \nu}{\mu - \nu}, \gamma = \dfrac{\beta}{\alpha}, \omega_0 = \dfrac{2}{\gamma}; k = 0;$

2, $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0;\ \mathbf{z}_0 = M^{-1}\mathbf{r}_0;\ \mathbf{x}_1 = \mathbf{x}_0 - \gamma^{-1}\mathbf{z}_0;\ \mathbf{r}_1 = \mathbf{b} - \mathbf{A}\mathbf{x}_1;\ \mathbf{z}_1 = M^{-1}\mathbf{r}_1;$

3, *while* $k \le k_{\max}$ *do*

4, $k = k + 1;\ \omega_k = 1 / \left(\gamma - \dfrac{1}{4\alpha^2}\omega_{k-1} \right);$

5, $\Delta\mathbf{x}_k = \omega_k \mathbf{z}_{k-1} + (\gamma\omega_k - 1)\,\Delta\mathbf{x}_{k-1};$

6, $\mathbf{x}_k = \mathbf{x}_{k-1} + \Delta\mathbf{x}_{k-1};\ \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k;$

7, $update\_halo(\mathbf{r}_k);\ \mathbf{z}_k = M^{-1}\mathbf{r}_k;$

8, *if* $k\%n_c == 0$ *then* *check convergence*;

9, *end while*

Maximum and minimum eigenvalues, priori information

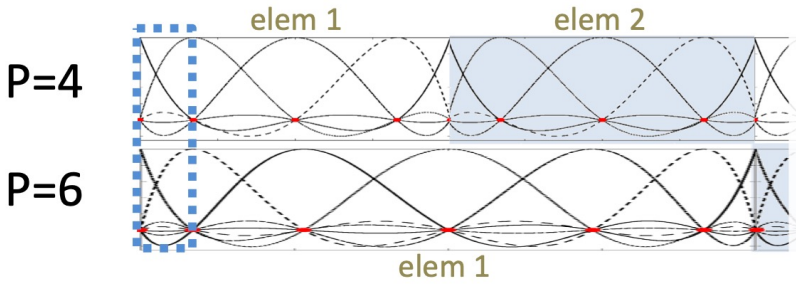
Spmv

Updatehalo and preconditioning

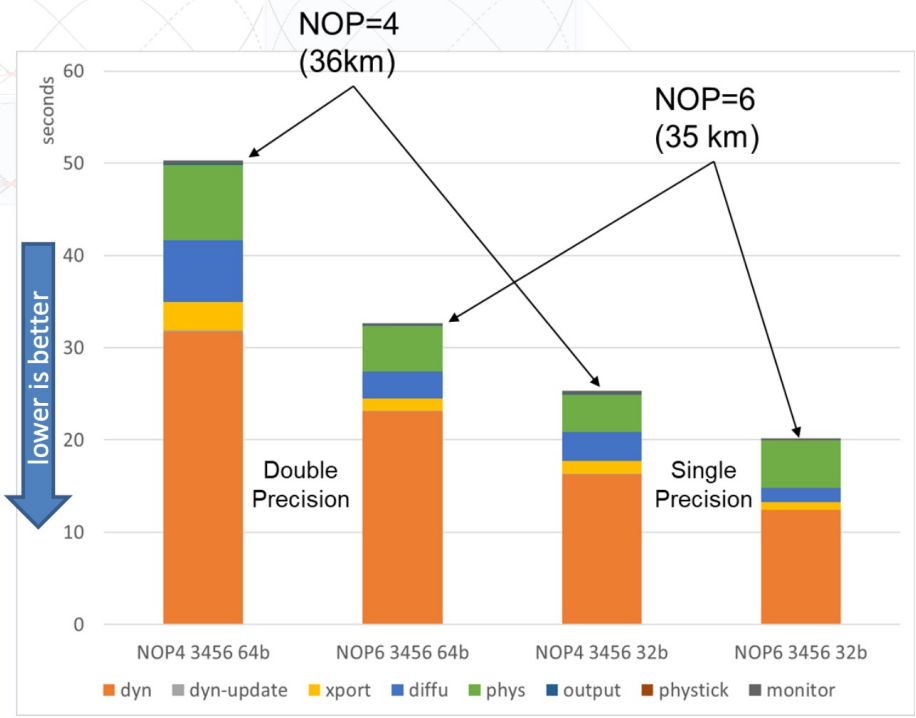
Global communication

- PCSI avoids frequent calculation of communication-intensive inner products.
- The convergence speed of PCSI reaches its theoretical optimum when the maximum and minimum eigenvalues are included in the algorithm to determine the recurrence coefficients.

Performance Optimization: Higher Order



- Increasing polynomial order from 4th to 6th
  - **1.50x** improvement for double precision
  - **1.25x** improvement for single precision
- Speed up due to multiple factors:
  - 56% fewer points with 6th-order for an equivalent grid spacing
  - Few shared points on elements with duplicate physics computations
- Ongoing testing to validate higher-order solution



Microphysics (PUMAS) Performance on CPU vs GPU

Experiment: CAM (192x288 lat-lon mesh; 32 levels; FP64; 36 MPI ranks/node or GPU) was run on CPU with PUMAS microphysics offloaded to GPU.

Equipment: NVIDIA V100 GPU was compared to Intel Xeon Broadwell and newer Skylake CPUs

Upper Right: CPU (cool colors) and GPU (warm colors) performance

Lower Right: Data transfer (cool colors) vs computation (red) times for different chunk sizes (PCOLS)

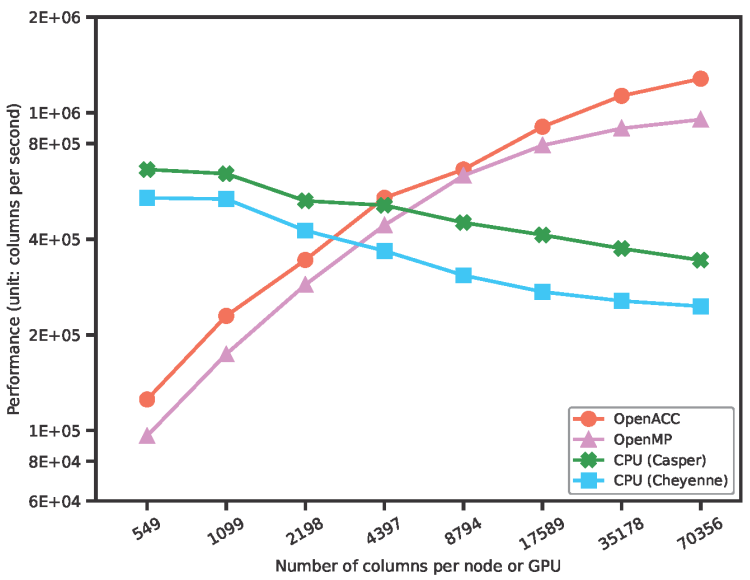
Take Aways:

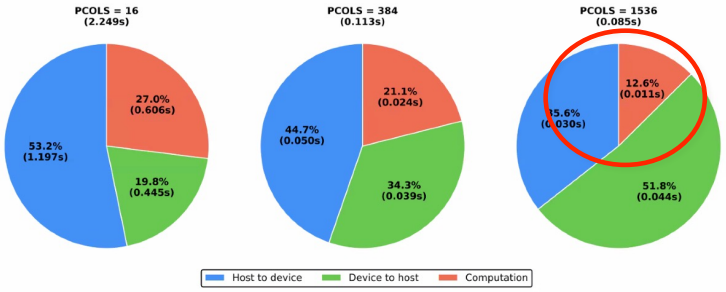
Clearly GPUs favor larger chunk sizes (PCOLS) and higher occupancy (Cols/GPU).

Excluding data transfer overhead, PUMAS-GPU is 7x faster than a CPU node: 5 Mcols/sec (GPU) vs 0.7 Mcols/sec (CPU)

Able to compare directive-based GPU offload schemes: OMP is slower than OpenACC, but improving!

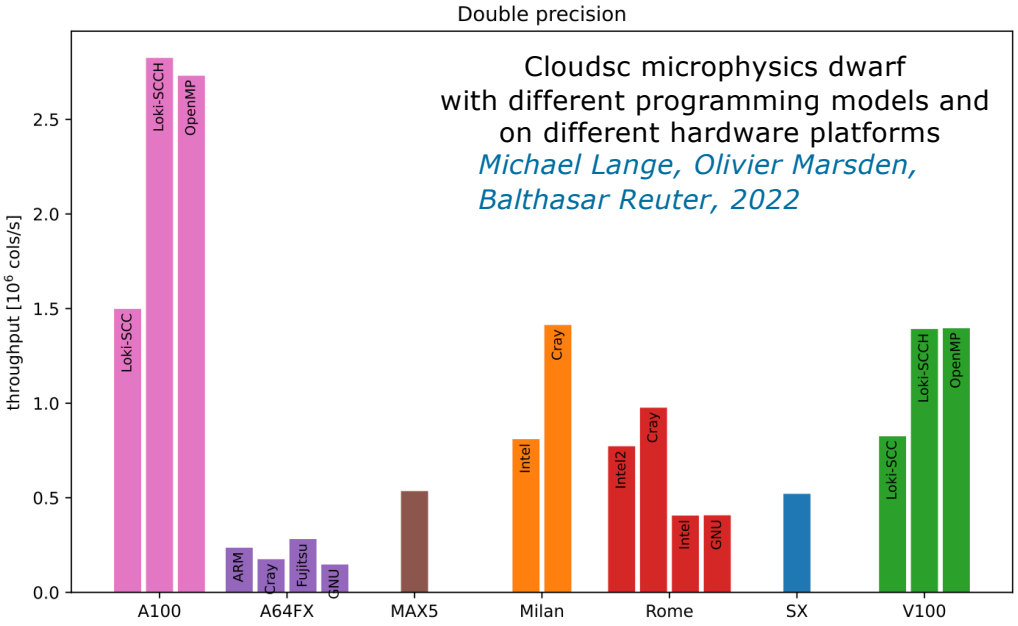
Slides courtesy of Jian Sun, NCAR





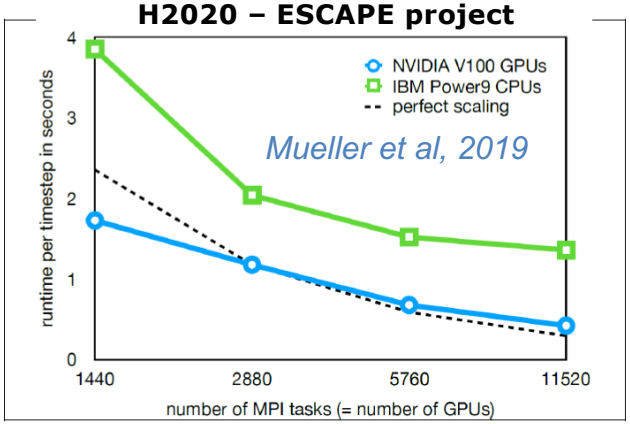
- Data transfer could be more time-consuming than computation

WEATHER & CLIMATE DWARFS – VERIFIABLE COLLABORATIVE TOOLS FOR VENDORS AND ACADEMIA

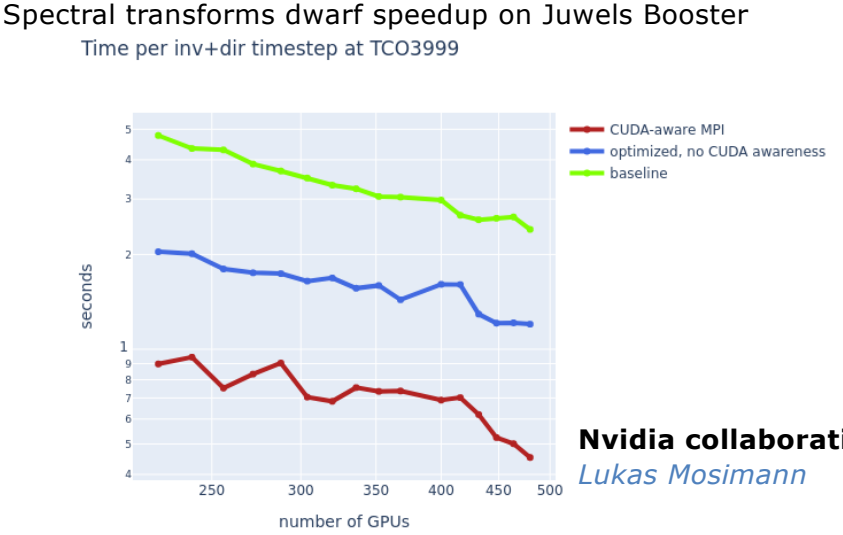


Atos collaboration
Erwan Raffin

H2020 – ESCAPE project

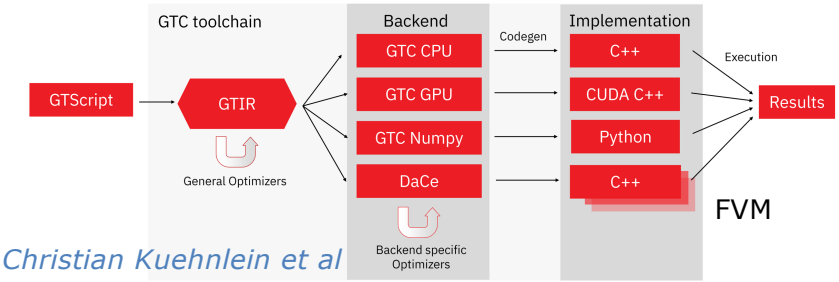


Spectral transforms dwarf speedup on Juwels Booster



Nvidia collaboration
Lukas Mosimann

DSL tool chain in collaboration with CSCS and ETH Swizerland

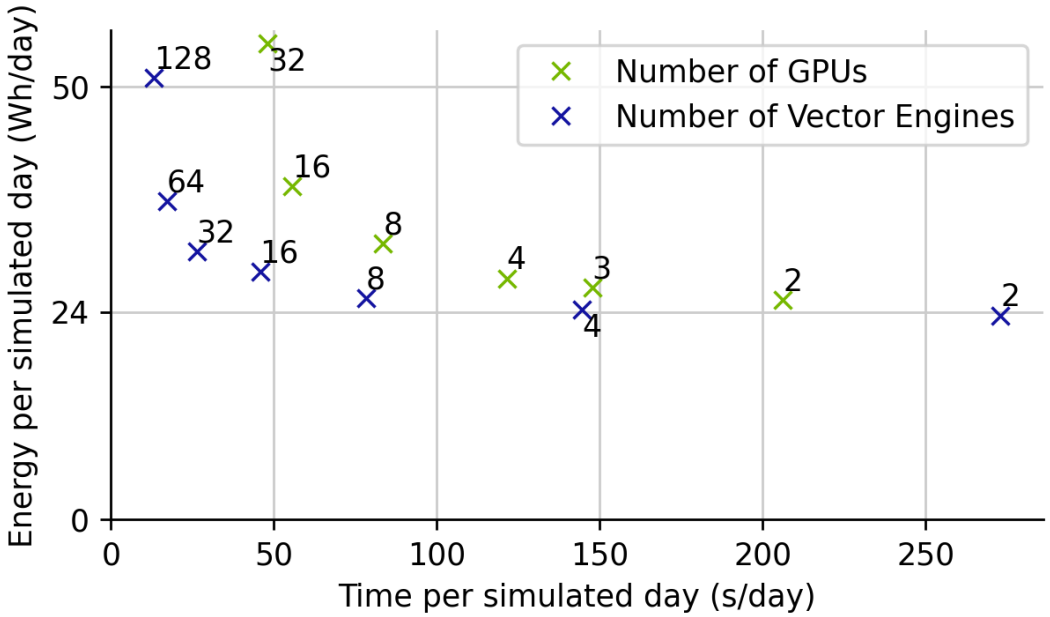


Christian Kuehnlein et al

Energy efficiency GPU vs. NEC Aurora

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

more efficient



faster

- GPU work in progress
- GPU not yet optimized for this experiment and machine

As reported by nvidia-smi (NVIDIA) and veda-smi (NEC)

Time loop only
Excluding the CPU host
GPU: out-of-the-box performance
Vector Engine: optimized setup

Experiment setup:
~ DWD ensemble forecast

- Global R2B6
  ~ 328 000 cells (40 km)
- + Nested grid over Europe
  ~ 49 000 cells (20 km)
- Output disabled

Compressed NetCDF for I/O and Inline Post-Processing

A decision was made to write out GFS.v16 forecast history files (atmf and sfcf) in netCDF format with compression. Parallel I/O was developed with updated netCDF and HDF libraries.

compression ratio:

Atmf 3d	**5x**	**(33.6 GB to 6.7 GB),**	**lossy compression**
sfc 2d	**2.5x**	**(2.8 GB to 1.1 GB),**	**lossless compression**

Inline post-processing (post library)

- makes use of forecast data saved in memory for post processing, *reduces I/O activity, and speeds up the entire forecast system*.
- Since lossy compression is applied for writing out forecast history files, *inline post generates more accurate products* than the standalone offline post.

The acronym MONAN

MONAN

Model for Ocean-laNd-Atmosphere predictioN

“Monan's representation is like something infinite. For the Tupi-Guarani-speaking nations, there is no notion of Christian Paradise, heaven, or hell as in Christian beliefs, but the "Land without evils" or Ybymarã-e'yma, the place where they live with their ancestors and gods, without war, famine, or any human ailments."

MONAN symbolizes the search for a better, sustainable, fraternal world with social justice.

And the community model contributes to the search for this world by generating more accurate information on the behavior and evolution of the components of planet Earth

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

Annex: Members' slides

CMC HPC/Exascale Projects: Science

- Work is ongoing for an AI-based replacement for radiation scheme to be tested in the Canadian climate model
- Planned upgrades to the existing modelling system:
  - A new preconditioner to make the iterative elliptic solver more scalable (ready)
  - SLIMEX : semi-Lagrangian Implicit-Explicit time integrator
    - Combine SL and IMEX BDF2 time integrator
    - Second order in time, no extra off-centering
    - Only one call to the elliptic solver per time step
  - High-level threading with OpenMP (partially implemented)
- Development of new algorithms:
  - Moving from a Yin-Yang grid to a rotated cubed-sphere grid
  - A space-time tensor formalism is used to express the equations of motion covariantly
  - The spatial discretization with the direct flux reconstruction method
  - New multistep exponential and implicit/Rosenbrock time integrators
  - Low-synchronization matrix-free Krylov solver

CMC HPC/Exascale Projects: Infrastructure

- Development of a new non-blocking IO server to solve increased IO bottleneck
- New more efficient MPMD multi-model coupling system
- Update to internal data format to enable parallel IO and multiple compression scheme allowing higher data compression
- Enable efficient check pointing on all model suites (standalone/coupled)

Hydrometeorological Centre of Russia – SL-AV model

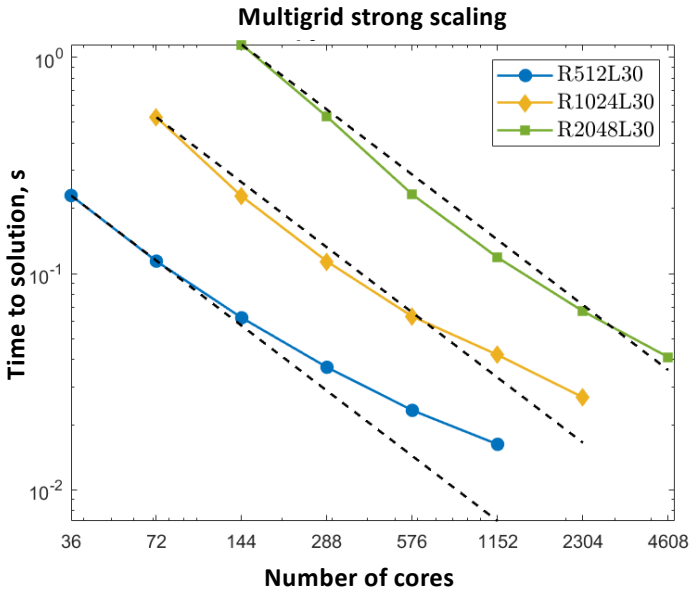
Memory usage optimization. Change of model state vector storage from a single array with the size $(6\,NLEV + 5, NLON, NJ)$ to separate arrays with the size $(NLEV, NLON, NJ)$.

- 70km model – 16-22% runtime reduction
- 10km model – 30% runtime reduction

Single precision computations. Working on further implementation of single precision computations.

Modification of grid-point space approximations for reduced grid. Testing of local horizontal schemes instead of Fourier space-based approximations within a 2d SWE prototype. Good results within idealized testcases.

New parallel geometric multigrid solver for reduced grid. Preliminary strong scaling results. Test problem at 20km reduced grid with 30 vertical levels scales at least to 4608 cores.



Courtesy M.Tolstykh and G.Goyman

HPC efforts at Météo-France

Towards a general use of single-precision (32 bits) in operational NWP systems.

1. Operational implementation in 5 AROME[1] operational systems
2. Next steps: operational use, whenever possible,
   i. in all AROME and ARPEGE forecasts
   ii. in all trajectories within the assimilation cycle
   iii. parts of assimilation

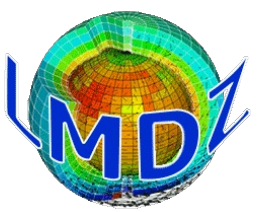
Adaptation to hybrid processors with accelerators:

1. Code refactoring (e.g., new memory structure)
2. Development of automatic source-to-source code transformation tools (Loki, Fxtran)
3. Porting small pieces of code (*dwarfs*) to Nvidia GPU
4. Porting on Nvidia PGI and Intel OneAPI environments
5. Work done in collaboration with ECMWF and ACCORD[2] partners
6. Hectometric LAM configuration developed within DestinE On Demand project (phase 1)
7. TRACCS[3]: 8-year (2023-2030) French national project for advancing climate modelling for climate services.
   ➢ 1 WP devoted to new computing paradigms (portability, efficiency, composability, trainable)

Input from Francois Bouyssel

[1] Météo-France LAM NWP operational system
[2] A COnsortium for convective-scale modelling Research and Development
[3] Transformative Advances of Climate Modelling for Climate Services

HPC efforts for LMDz



Code improvements, increased efficiency, and preparing GPUs.

1. DYNAMICO (new dynamical core) already ported to GPU
2. Code refactoring for model physics: improve portability, readability, composability. Clarification of interfaces between parameterizations
3. Tests using mixed precision.
4. TRACCS framework

Input from Romain Roehrig

HPC efforts for NEMO Ocean Model

Tackling most optimisation aspects concurrently to avoid bottlenecks.

1. Reduce memory footprint: e.g., fewer local and global arrays, memory access limitation
2. Reduction the number of communications: gathering communications, using non-blocking communications, using different communication strategies (e.g., collectives, one-sided)
3. Code refactoring: use of tiles for memory cache, introducing OpenMP
4. I/O with XIOS
5. Exploring porting to GPU, using directives or PSyclone
6. Single/mixed precision within some NEMO subgroups

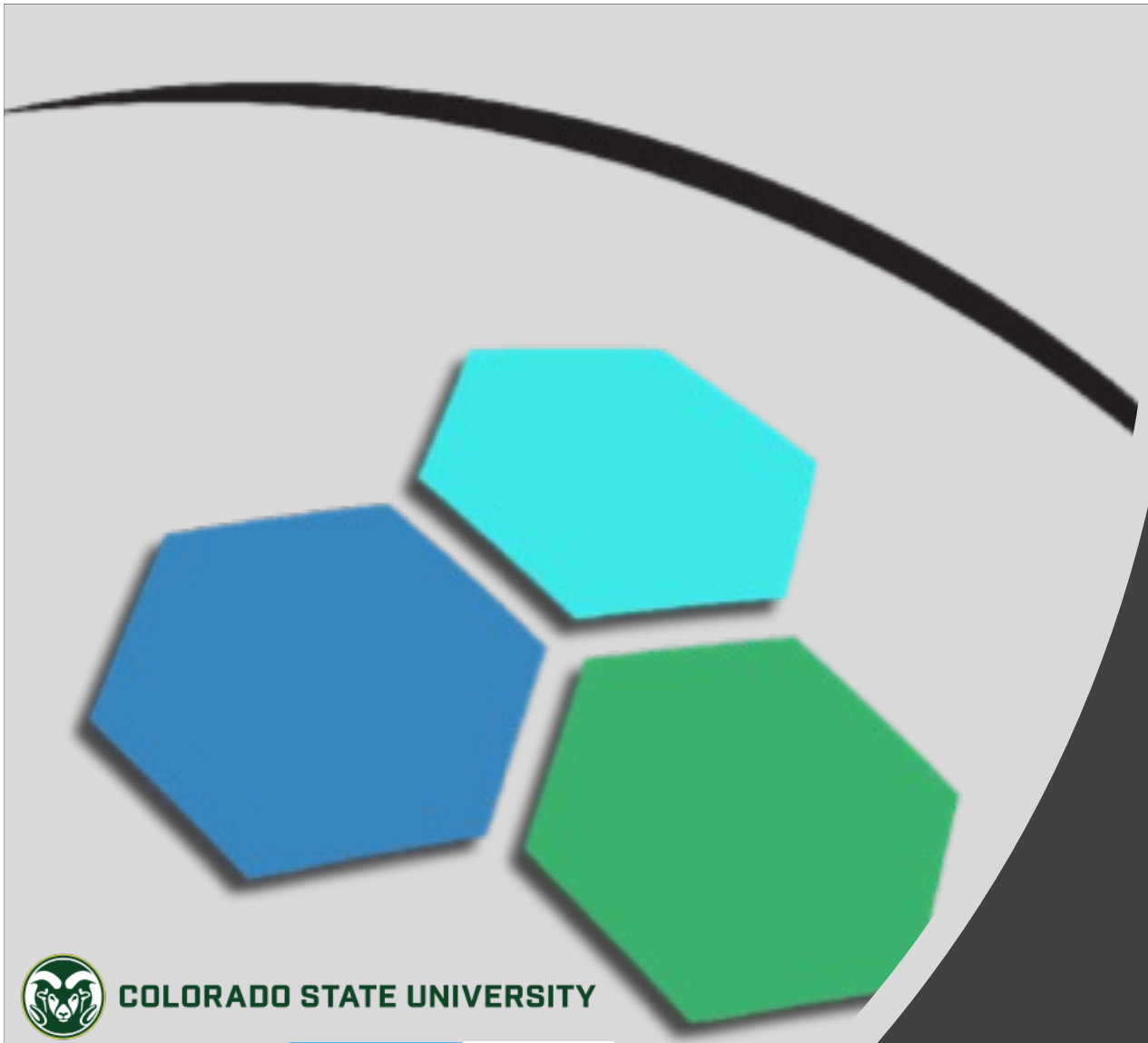
Using hybrid resolutions:

1. Coupling using OASIS[1] or AGRIF[2] for coarse-graining of biochemistry
2. Coupling using OASIS for separating ocean and sea-ice models on different grids (e.g., North Pole)
3. Same for the iceberg model
4. Multiple zooms and grid-nesting to achieve very high resolution wherever relevant (application-dependent)

➢ Balancing HPC load becomes critical

Input from Sebastien Masson

[1] Model coupling tool
[2] Zoom capability of NEMO



COLORADO STATE UNIVERSITY

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

AreandDee LLC
Come scale away...

NVIDIA

EarthWorks:

For Peter

Richard Loft[2], Sheri Mickelson[1]
Thomas Hauser[1], Michael Duda[1],
Dylan Dickerson[1], Supreeth Suresh[1],
John Clyne[1], Jian Sun[1], Chris Fisher[1],
Mariana Vertenstein[1], Donald
Dazlich[3], Dave Randall[3], Raghu Raj
Kumar[4], Pranay Reddy Kommera[4]

1 National Center for Atmospheric Research
2 AreandDee, LLC
3 Colorado State University
4 NVIDIA Corporation

EarthWorks Strategy for GSRM

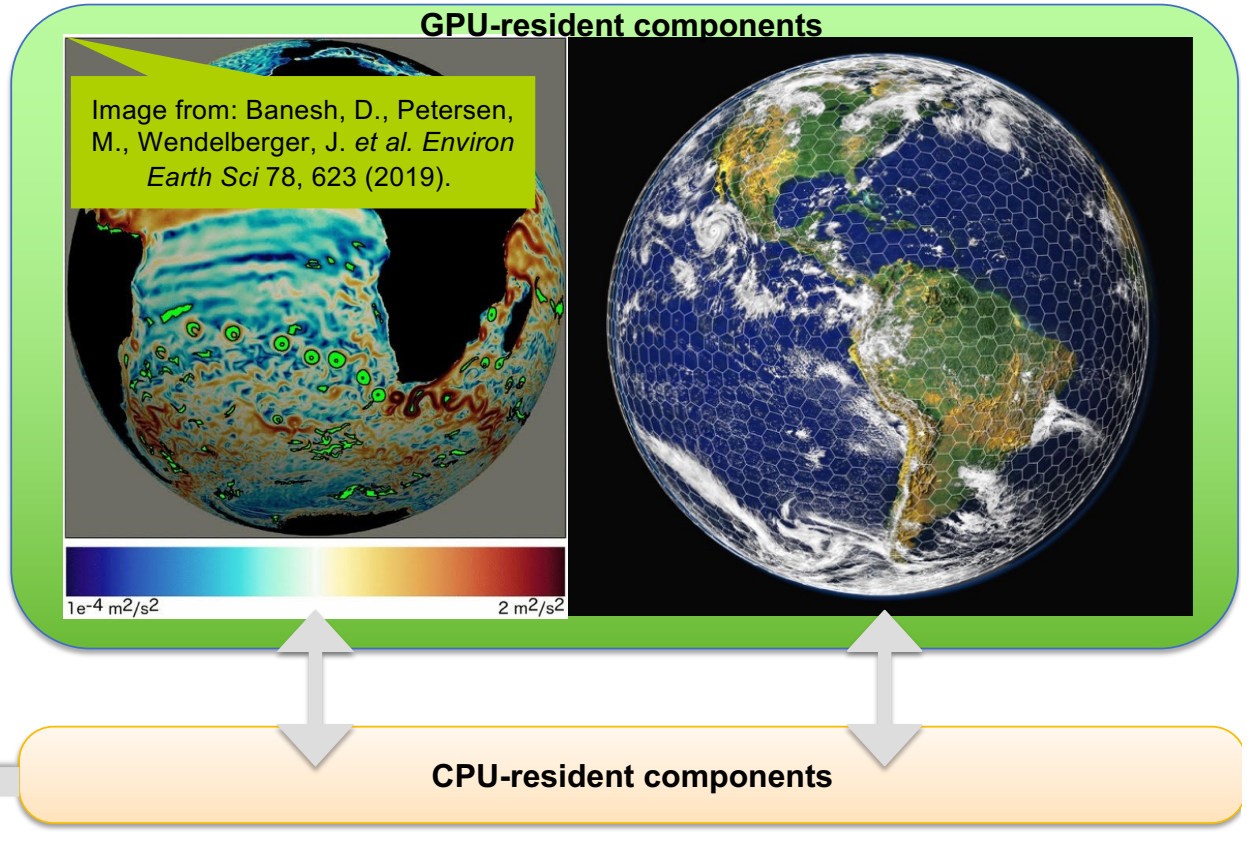
Use regional refinement capabilities of MPAS to reduce cost of tuning climate parameterizations for meteorological length scales

Target heterogeneous computing with GPUs

Prioritize Accelerating 1) Atmosphere and 2) Ocean

Target large systems, e.g. exascale "Leadership Class" Systems

Adopt an end-to-end approach to addressing scalability issues



EarthWorks testing status: GPU

Component	Subcomponent	Package	GPU Porting Status	Offload Paradigm	Version Skew?	Optimized	Comments
Atmosphere			in progress				
	Dycore	MPAS-7.x	completed	OpenACC & OpenMP	merge with CAM	yes	Awaiting results of upgraded MPAS-7 version scaling tests. AMD has attempted a port of MPAS using OpenMP offload.
	Physics	CAM-6	in progress				
		PUMAS	completed	OpenACC & OpenMP	MG2->MG3	yes	Add new µprocesses: e.g. graupel.
		RRTMGP	completed	OpenACC	no	deferred	Need CPU version running under NVHPC compiler before testing on GPUs.
		CLUBB	in progress	OpenMP		no	Loop order refactoring for better vectorization, GPU parallelism underway. Merging physics to better represent tropical cyclones may be required.
Ocean	MPAS-O		completed	OpenACC	yes	yes	Need to test E3SM GPU versions and merge with CPU version.
Sea-Ice	MPAS-SI		deferred				Low-ish priority but looks doable.

Microphysics (PUMAS) Performance on CPU vs GPU

Experiment: CAM (192x288 lat-lon mesh; 32 levels; FP64; 36 MPI ranks/node or GPU) was run on CPU with PUMAS microphysics offloaded to GPU.

Equipment: NVIDIA V100 GPU was compared to Intel Xeon Broadwell and newer Skylake CPUs

Upper Right: CPU (cool colors) and GPU (warm colors) performance

Lower Right: Data transfer (cool colors) vs computation (red) times for different chunk sizes (PCOLS)

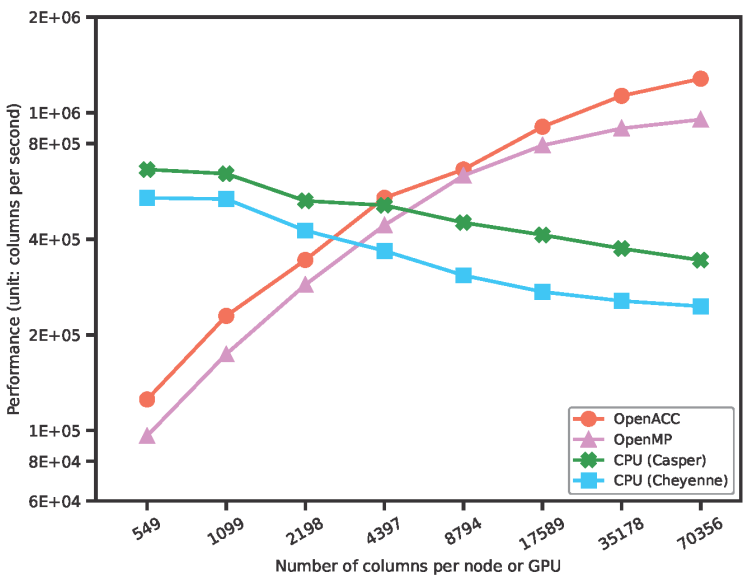
Take Aways:

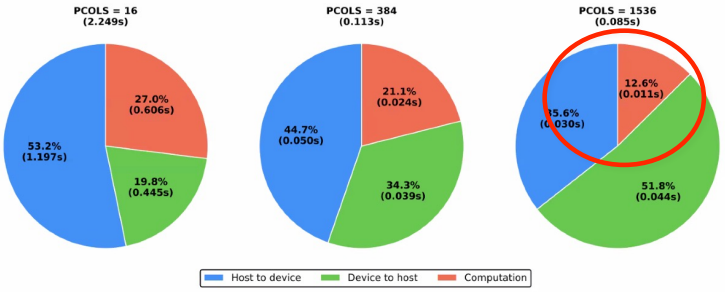
Clearly GPUs favor larger chunk sizes (PCOLS) and higher occupancy (Cols/GPU).

Excluding data transfer overhead, PUMAS-GPU is 7x faster than a CPU node: 5 Mcols/sec (GPU) vs 0.7 Mcols/sec (CPU)

Able to compare directive-based GPU offload schemes: OMP is slower than OpenACC, but improving!

Slides courtesy of Jian Sun, NCAR





- Data transfer could be more time-consuming than computation

Atmospheric Dycore (MPAS-7) CPU/GPU Performance

Experiment: MPAS-7 (5.9M cell mesh; 56 levels; FP32) ran dry baroclinic test case for 10 simulated days

Equipment: Selene supercomputer; nodes = AMD Dual socket EPYC 7742 "Rome" CPUs with 8x NVIDIA A100 GPUs; 10 HDR links/node.

Upper Left: Benchmark of 128-core ROME CPU node vs A100 GPU

Lower Left: Amdahl comparison of GPU performance for MPAS-6 vs MPAS-7. Compute (m) has gotten worse; latency has improved.

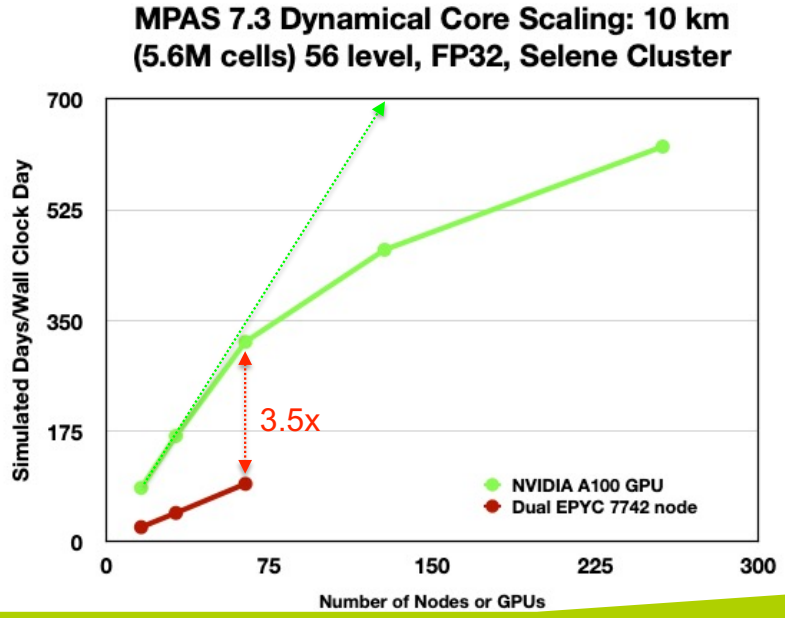
Take Aways:

Early scaling looks impressive - and 3.5x faster than CPU node.

Latency (t0) kills multi-GPU scalability, represented by dotted green line). Have traced this issue to asynchronous MPI_Wait.

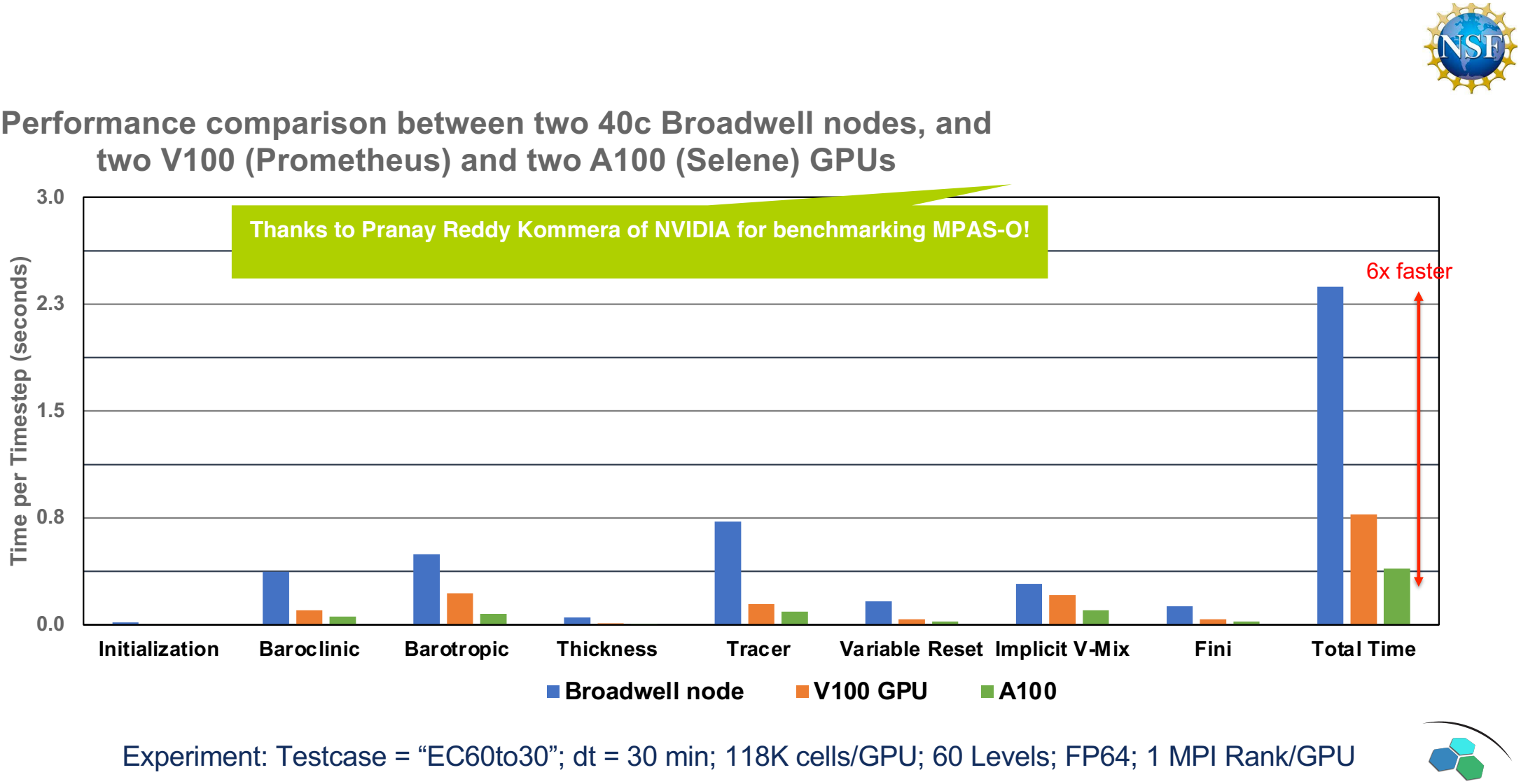
Slowdown of MPAS-7 compute (m) was recently isolated to not declaring new variables GPU resident.

With upgrades, 4 SYPD at 10 km on 256 A100s achievable.



Thanks to Raghu Raj Kumar of NVIDIA for benchmarking MPAS-7!

	m	standard deviation	t0	standard deviation
MPAS 6.3	8.00964	0.211429	0.06542	0.006820
MPAS 7	10.56827	0.317847	0.04126	0.010253



Experiment: Testcase = “EC60to30”; dt = 30 min; 118K cells/GPU; 60 Levels; FP64; 1 MPI Rank/GPU

NGMS project status – October 2022

green=active; yellow=spinning up; white=waiting; blue=ExCALIBUR-resourced

NG-UX (User Experience)
- Paul Phillips [George Pankiewicz]
- Develop and inaugural delivery of NGMS training material transitioning to BaU. This project will include aspects of usability

GungHo Atmosphere Science Project
Ben Shipway [Nigel Wood]
- Develop atmospheric science aspects & deliver model scientifically as good as UM

LFRic Infrastructure Development
Steve Mullerworth
[JC Rioual]
- Deliver infrastructure to replace the UM scalable for future platforms

LFRic Inputs Paused to Jan-23
Rich Gilham [Glenn Greed]
- Tools to ingest fixed & time-varying fields.
- Include initial conditions, ancillary fields and LBCs

ExCALIBUR data workflow
Stuart Whitehouse [Glenn Greed]
- Development of research diagnostics and research workflow capabilities

NG-Marine Systems
Mike Bell [Andy Saulter]
- Deliver scalable marine systems including ocean, sea-ice & wave models

NG-Coupling
JC Rioual [Ben Shipway]
- OASIS3-MCT coupled components

RAL3-LFRic
Mike Bush [Huw Lewis]
- NGMS-ready coupled atmos/ocean DA
- JEDI as a DA framework

NG-PAO
David Simonin [Chiara Piccolo]
- JEDI as a DA framework
- Processing of NWP observational data for NG-DA

GC5-LFRic
Maria Carvalho [Alistair Sellar]
- Global model evaluation
- Developing a coupled model as good as the UM equivalent

FAB Build System
Rich Gilham [Glenn Greed]
- Development of new build systems for NGMS components

NG-R2O
Mike Thurlow [David Walters]
- Support transition of NGMS capability from research to NWP operations

NG-Composition
Fiona O'Connor [Matt Hort]
- Coordination of aerosol & chemistry development within NGMS

NG-Ver
Phil Gill [Oak Wells]
- Development of NWP verification capability for NGMS

NG-R2C
- Camilla Mathison[Richard Wood]
- Support transition of capability from research to climate production

NG-Name
- Ben Devenish [Matt Hort]
- Development of dispersion models (e.g. NAME) for next generation computing

NG-Integration
- Sam Adams [JC Rioual]
- Provide a focus for IO development, cloud computing and integration

NG-Optimisation
- Chris Maynard [JC Rioual]
- Optimisation of NGMS components with initial focus on LFRic

Met Office

A few highlights

- NG-Marine:
  - Work to get NEMO, SI3, NEMOVAR and WWIII to run on GPU machines
  - Experimental work to develop a code transformation tool for NEMO, SI3, NEMOVAR and WWIII so that PsyClone can be used to add directives (e.g. OpenACC, OpenMP).
    - No need for a code rewrite
  - Contact Andrew Porter andrew.porter@stfc.ac.uk for more information.
- Diagnostics:
  - Plan to use XIOS including regridding capability for diagnostics
  - Some concerns over performance when scaled up to a full weather model
  - Will require optimisation
- GC5-LFRic:
  - Developing a coupled model equivalent to our latest UM based coupled model (GC5) including all the same physics and the LFRic dynamical core.

HPC activities in ICON consortium

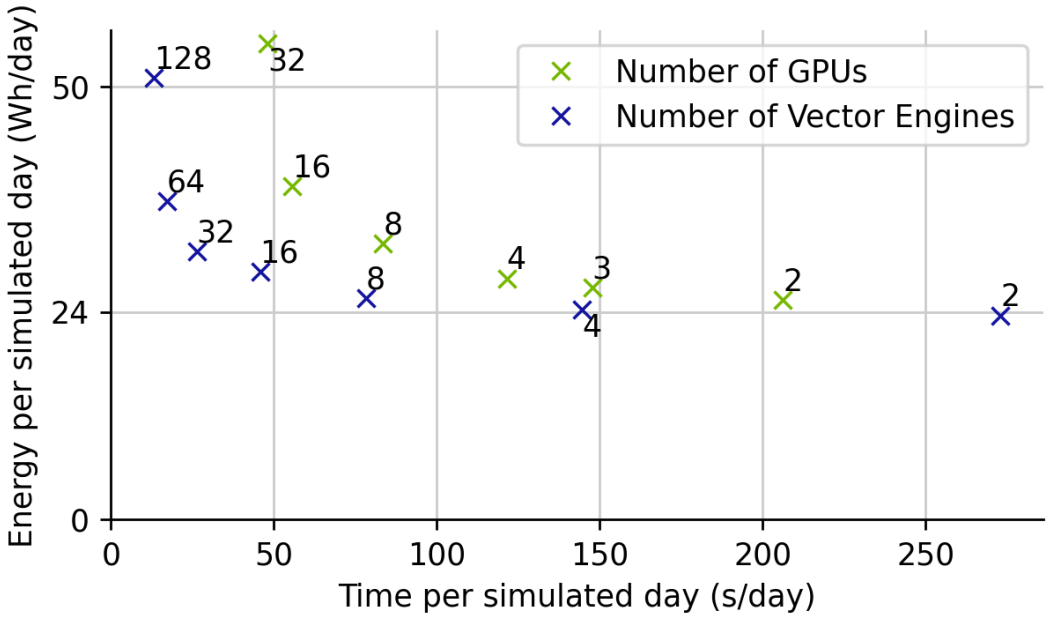
Deutscher Wetterdienst
Wetter und Klima aus einer Hand
DWD

- GPU port of ICON (using OpenACC) is completed for the computationally relevant parts of the model (excluding forward operators for data assimilation); port of ecRAD is at provisional stage
- Optimization efforts are still ongoing (most intensively at MeteoSwiss)
- Currently, energy efficiency for ICON-NWP is similar for NEC SX Aurora and Nvidia A100, but scaling is much better for the NEC
- More aggressive code refactoring is under investigation in various projects (e.g. EXCLAIM, WarmWorld)
- Recent scaling tests at DWD indicate that a configuration with global 3.25 km / 120 levels would be able to meet the operational time constraint (7.5 days in 50 min) on 75% of the current machine with NetCDF input, but GRIB2 input needs optimization (parallelization)

Energy efficiency GPU vs. NEC Aurora

Deutscher Wetterdienst
Wetter und Klima aus einer Hand

more efficient



faster

- GPU work in progress
- GPU not yet optimized for this experiment and machine

As reported by nvidia-smi (NVIDIA) and veda-smi (NEC)

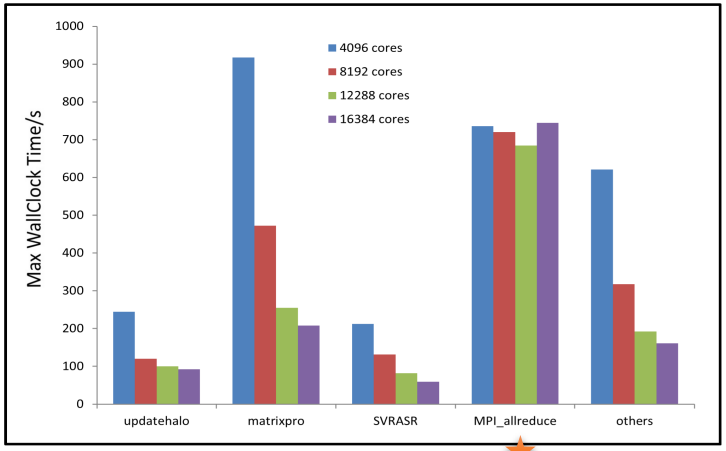
Time loop only
Excluding the CPU host
GPU: out-of-the-box performance
Vector Engine: optimized setup

Experiment setup:
~ DWD ensemble forecast

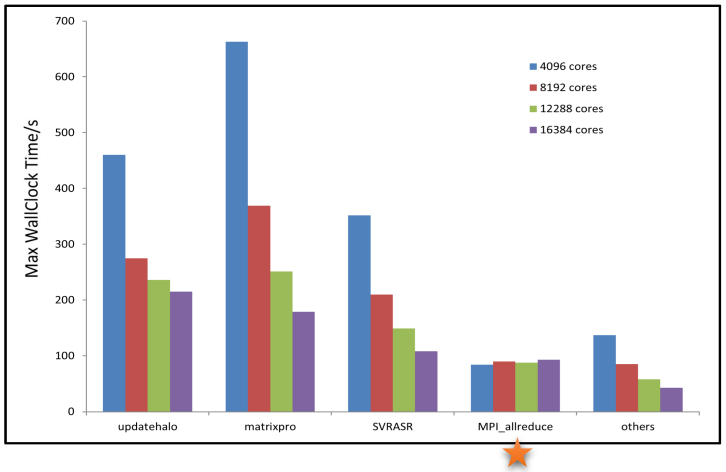
- Global R2B6
  ~ 328 000 cells (40 km)
- + Nested grid over Europe
  ~ 49 000 cells (20 km)
- Output disabled

Helmholtz solver (PCSI) for GRAPES-GFS

PGCR



PCSI



Preconditioned Classical Stiefel Iteration (PCSI)

$\mathbf{x}_0$, *estimated eigenvalue boundary* $[\nu, \mu]$

1, $\alpha = \dfrac{2}{\mu - \nu}, \beta = \dfrac{\mu + \nu}{\mu - \nu}, \gamma = \dfrac{\beta}{\alpha}, \omega_0 = \dfrac{2}{\gamma}; k = 0;$

2, $\mathbf{r}_0 = \mathbf{b} - \mathbf{A}\mathbf{x}_0;\ \mathbf{z}_0 = M^{-1}\mathbf{r}_0;\ \mathbf{x}_1 = \mathbf{x}_0 - \gamma^{-1}\mathbf{z}_0;\ \mathbf{r}_1 = \mathbf{b} - \mathbf{A}\mathbf{x}_1;\ \mathbf{z}_1 = M^{-1}\mathbf{r}_1;$

3, *while* $k \le k_{\max}$ *do*

4, $k = k + 1;\ \omega_k = 1 / \left(\gamma - \dfrac{1}{4\alpha^2}\omega_{k-1} \right);$

5, $\Delta\mathbf{x}_k = \omega_k \mathbf{z}_{k-1} + (\gamma\omega_k - 1)\,\Delta\mathbf{x}_{k-1};$

6, $\mathbf{x}_k = \mathbf{x}_{k-1} + \Delta\mathbf{x}_{k-1};\ \mathbf{r}_k = \mathbf{b} - \mathbf{A}\mathbf{x}_k;$

7, $update\_halo(\mathbf{r}_k);\ \mathbf{z}_k = M^{-1}\mathbf{r}_k;$

8, *if* $k\%n_c == 0$ *then* *check convergence*;

9, *end while*

Maximum and minimum eigenvalues, priori information

Spmv

Updatehalo and preconditioning

Global communication

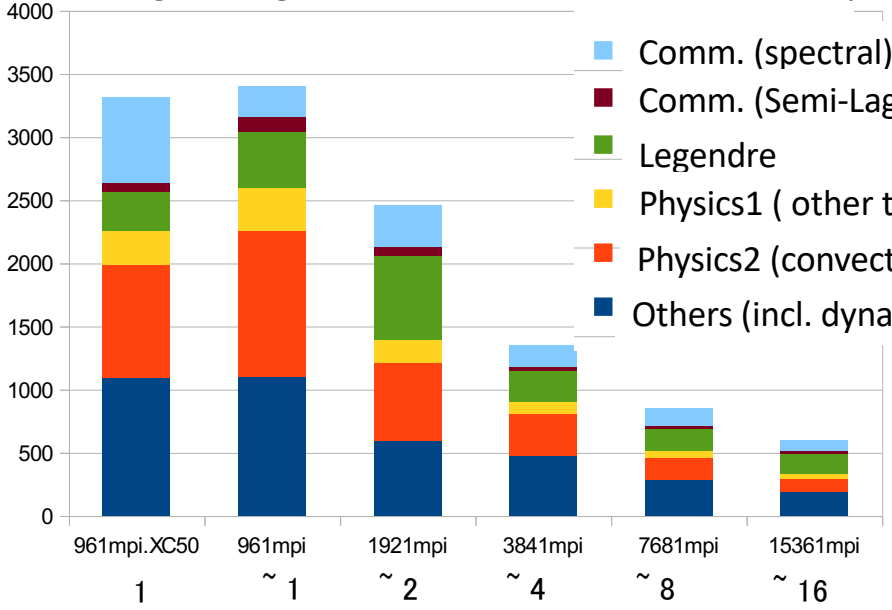
- PCSI avoids frequent calculation of communication-intensive inner products.
- The convergence speed of PCSI reaches its theoretical optimum when the maximum and minimum eigenvalues are included in the algorithm to determine the recurrence coefficients.

Elapsed time of Tc1919L128 GSM (~5km, dynamical core only) on “Fugaku”

MPI decomposition	Num. of nodes （MPI ranks, OpenMP threads）	Total Elapsed time[s]	Elapsed time for MPI communication [s]
Current	3841 (3841, 48)	1053	466～515
New	3841 (3841, 48)	733	280～408

The new method approximately halved the elapse time for communication due to significant reduction of communication for SL advection.

Strong scaling of Tc1919L128 GSM (~5km, incl. physics) on “Fugaku”



Resources relative to Tq959L128 on XC50

It takes 2x more time to run Tc1919 (~20x more gridpoints than Tq959) if we use 16x more resources than that for Tq959.

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

10 year development programme

51 ANOS - INPE CACHOEIRA PAULISTA

MONAN

Model for Ocean-laNd-Atmosphere predictioN

A new paradigm for advancing the Earth system numerical prediction in Brazil and Latin America

Saulo R. Freitas
Scientific Committee for the Community Model MONAN
Earth System Numerical Modeling Division- CGCT/INPE
saulo.freitas@inpe.br

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

The acronym MONAN

MONAN

Model for Ocean-laNd-Atmosphere predictioN

“Monan's representation is like something infinite. For the Tupi-Guarani-speaking nations, there is no notion of Christian Paradise, heaven, or hell as in Christian beliefs, but the "Land without evils" or Ybymarã-e'yma, the place where they live with their ancestors and gods, without war, famine, or any human ailments."

MONAN symbolizes the search for a better, sustainable, fraternal world with social justice.

And the community model contributes to the search for this world by generating more accurate information on the behavior and evolution of the components of planet Earth

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

PÁTRIA AMADA BRASIL GOVERNO FEDERAL

NOAA Unified Forecast System

- **UFS components**: Atmos (fv3 dycore), Land (Noah-MP), Ocean (MOM6), Ice (CICE6), Wave (WAVEWATCH III), Aerosol (GOCART), Air Quality (CMAQ), CMEPS mediator, CCPP physics
- **UFS Applications:**
  - ***Global***: GFS (medium-range NWP), GEFS (ensemble), SFS (seasonal), UFS-aerosol
  - ***Regional***: HAFS (hurricane), RRFS (regional NWP), Online-CMAQ (air quality)

Improvement for I/O and computational efficiency

- Parallel NetCDF with data compression applied to history files
- ESMF managed threading -- apply different threads for different UFS components
- Single and double precision dycore
- 32-bit physics (project just gets started)

HPC upgrade

Old: WCOSS, Dell, 73K x 2 cores, 4302 x2 TF peak performance

New as of June 2022: WCOSS2, CRAY EX, 2560x2 nodes, 327Kx2 cores, 12,100 x2 TF peak performance.

Compressed NetCDF for I/O and Inline Post-Processing

A decision was made to write out GFS.v16 forecast history files (atmf and sfcf) in netCDF format with compression. Parallel I/O was developed with updated netCDF and HDF libraries.

compression ratio:

Atmf 3d	**5x**	**(33.6 GB to 6.7 GB),**	**lossy compression**
sfc 2d	**2.5x**	**(2.8 GB to 1.1 GB),**	**lossless compression**

Inline post-processing (post library)

- makes use of forecast data saved in memory for post processing, *reduces I/O activity, and speeds up the entire forecast system*.
- Since lossy compression is applied for writing out forecast history files, *inline post generates more accurate products* than the standalone offline post.

WCOSS2 In Operation Since June 2022

Locations
- Manassas, VA
- Phoenix, AZ

Performance Requirements
- 99.9% Operational Use Time
- 99.0% On-time Product Generation
- 99.0% Development Use Time
- 99.0% System Availability

Configuration
- Cray EX system
- **12.1 PetaFlops**
- Multi-tiered storage
  - 2 flash filesystems each with...
    - 614 TB usable storage
    - 300 GB/s bandwidth
  - 2 HDD filesystems each with...
    - 12.5 PB usable storage
    - 200 GB/s bandwidth
  - Total aggregate - 26.2PB at 1TB/s
- Lustre parallel filesystem
- PBSpro workload manager
- Ecflow scheduler

- Compute nodes
  - **2,560 nodes (60 spare)**
  - 327,680 cores
    - **128 cores/node**
  - 1.3 PB of memory
    - 512 GB/node
- Pre/post-processing nodes
  - 132 nodes (4 spare)
  - 8,448 cores
    - 64 cores/node
  - 132 TB of memory
    - 1TB/node
- 200Gb/s Slingshot interconnect

HPC readiness: input from JMA

Numerical Prediction Division,
Japan Meteorological Agency

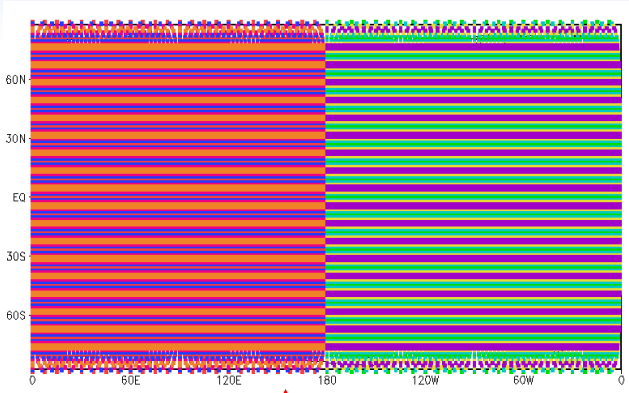
Highlights

- Feasibility studies of future specification of NWP systems on Supercomputer “Fugaku”
  - GSM (Global Spectral Model) with Tc1919 (~5km grid-spacing)
    - Improvement of parallelization and saving memory enabled us to run 5km GSM.
    - Further speedup (e.g. single precision) is required for more feasible resources
- Research on GPU adaptation
  - GPU (OpenACC) porting for Global (GSM), regional (ASUCA) and ocean (MRI.COM) models are ongoing.
- Research on reduced precision
  - Single precision versions of
    - ASUCA: under development
    - GSM: under consideration

An example of Tl159 2x4 MPI decomposition

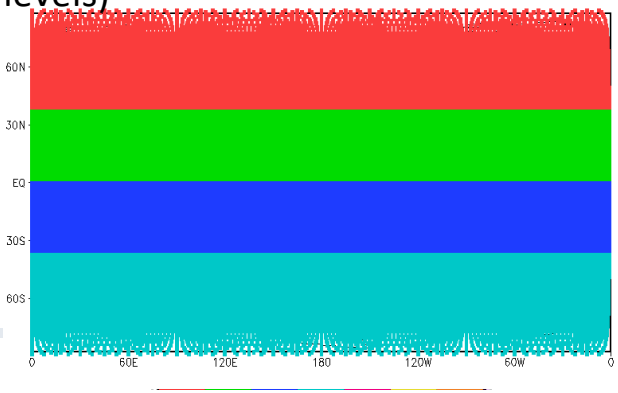
Current decomposition (two grid stages)

For parameterization, I/O and spectral transform



Transpose (with all-to-all MPI communication) required every time step

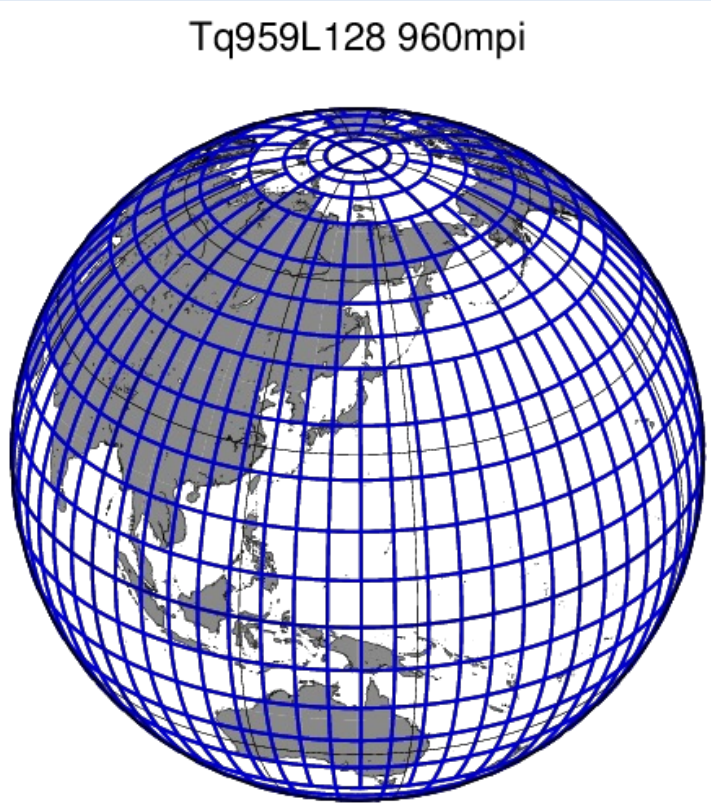
For SL advection (first to (Kmax/2)th vertical levels)



New decomposition (unified grid stage)

MPI rank

Another example of Tq959 MPI decomposition



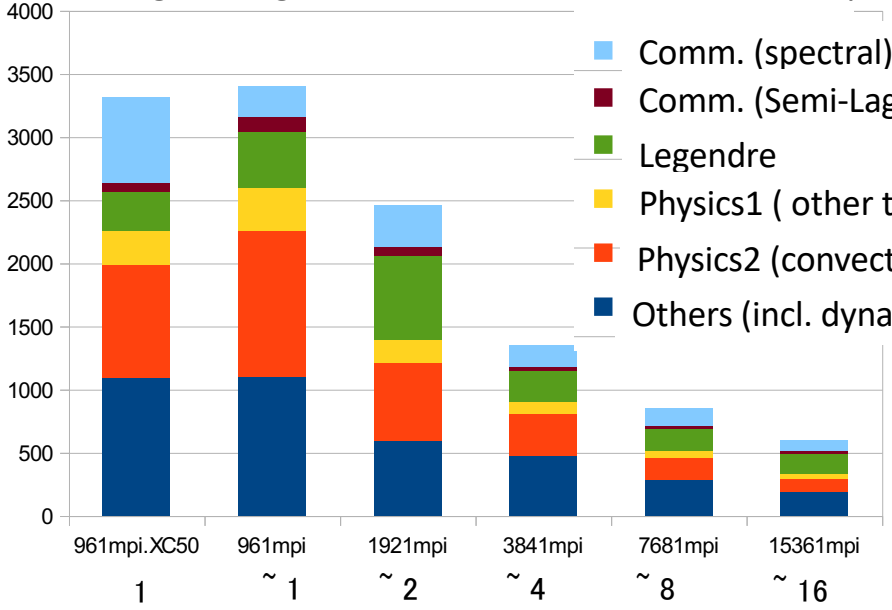
- Pros.
  - Number of gridpoints / MPI rank is equalized as possible
  - Suitable for both derivative stencils in gridspace and spectral transform.
    - also preparation for grid-spectral hybrid approach
- Cons.
  - Load-imbalance of computation in physics parameterizations gets worse (however, pros. overweigh in high-resolution runs)

Elapsed time of Tc1919L128 GSM (~5km, dynamical core only) on “Fugaku”

MPI decomposition	Num. of nodes （MPI ranks, OpenMP threads）	Total Elapsed time[s]	Elapsed time for MPI communication [s]
Current	3841 (3841, 48)	1053	466～515
New	3841 (3841, 48)	733	280～408

The new method approximately halved the elapse time for communication due to significant reduction of communication for SL advection.

Strong scaling of Tc1919L128 GSM (~5km, incl. physics) on “Fugaku”



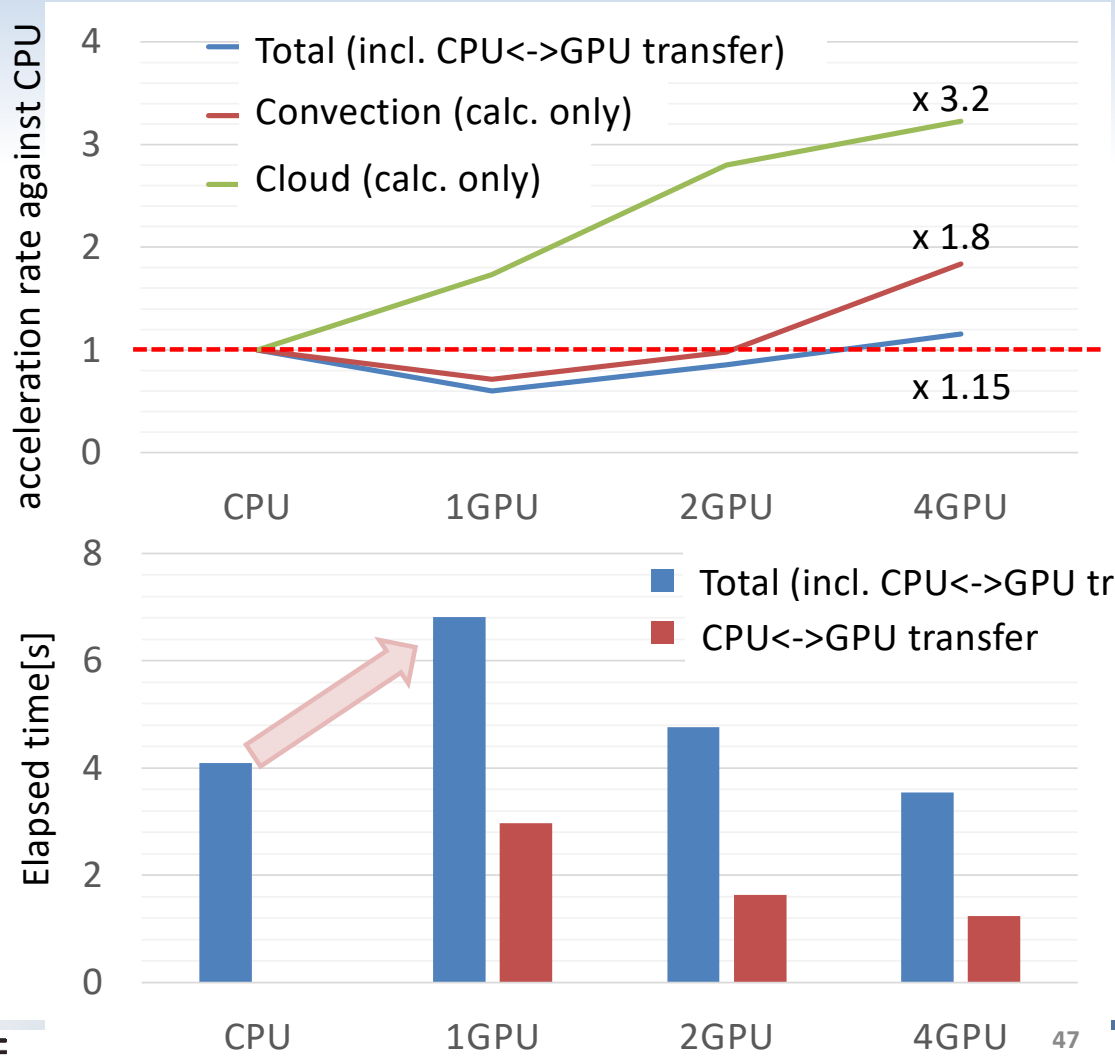
Resources relative to Tq959L128 on XC50

It takes 2x more time to run Tc1919 (~20x more gridpoints than Tq959) if we use 16x more resources than that for Tq959.

Progress of GPU porting

- GSM (global model)
  - **Moist physics has been ported to GPU**.
    - A multi GPU approach can contribute to both speed up GPU computing and transferring data between CPU and GPU.
  - Spectral transform part is now under porting (planning to use cuBLAS) .
- ASUCA (regional model)
  - Core of the dynamical process has been ported to GPU. now under porting for physics parameterizations
- MRI.COM (ocean model)
  - Porting bottlenecks in the dynamical process to GPU
- Common issues
  - Performance of GPU acceleration strongly depends on CPU <-> GPU transfer cost.
  - Refactoring required (array ordering, algorithms more suitable for GPU etc)
    - automated source-to-source code transformation tools are not used so far.

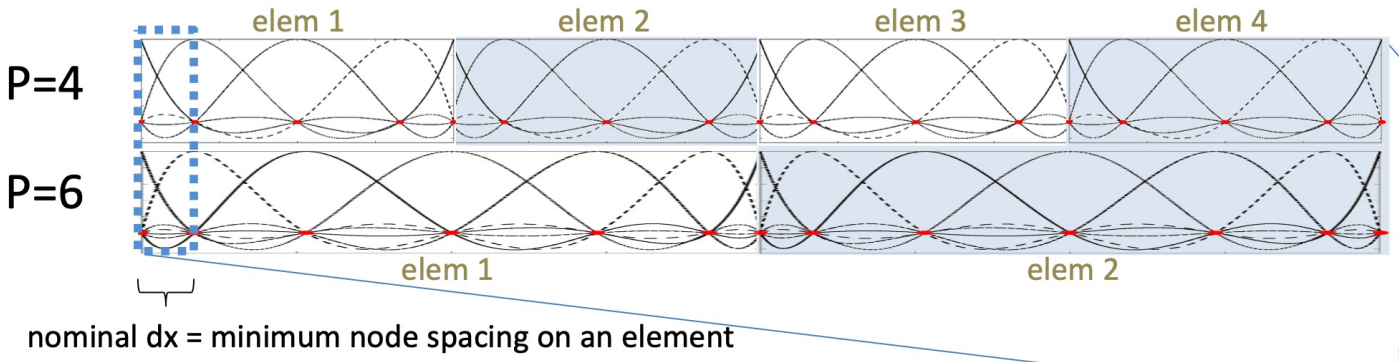
Test cases of TL319L128 (dx~55km, 640x320x128 gridpoints), 1 node 16MPI, 4 threads/MPI



Tested on
CPU:Intel Xeon Gold 6226 2.7GHz(12core) × 2
GPU:TeslaV100 SXM2 32GB × 4

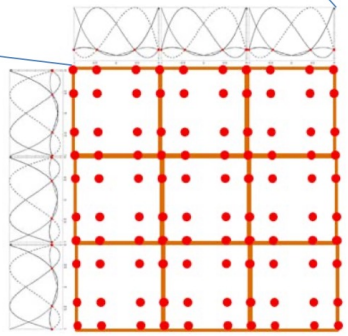
The multi GPU approach speeds up both computing and data transfer due to reduction of transferred data /GPU

Performance Optimization: Higher Order



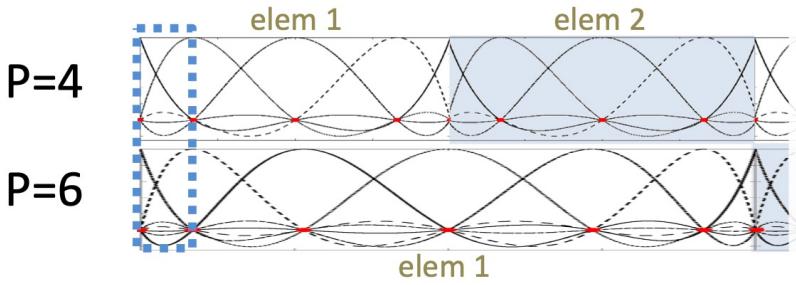
Can higher-order basis functions provide same accuracy at lower cost?

- \+ P=6 element covers ~2x the distance as P=4 with the same dx and dt
- \+ Fewer elements and degrees of freedom means *P=6 needs 1/4 as many elements and 56% the number of DOFs as P=4*
- \+ Fewer replicated computations on shared points between elements (but also less coarse-grained parallelism over element-blocks)
- – P=6 costs ~1.2x more than P=4 on a per DOF basis



Schematic showing nine P=3 elements

Performance Optimization: Higher Order



- Increasing polynomial order from 4th to 6th
  - **1.50x** improvement for double precision
  - **1.25x** improvement for single precision
- Speed up due to multiple factors:
  - 56% fewer points with 6th-order for an equivalent grid spacing
  - Few shared points on elements with duplicate physics computations
- Ongoing testing to validate higher-order solution