38. WGNE , Brazil, 27-30th November 2023

WGNE – HPC/Exascale update

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

Many thanks for the individual contributions!

WMO

1873
2023
IMO-WMO

WMO OMM

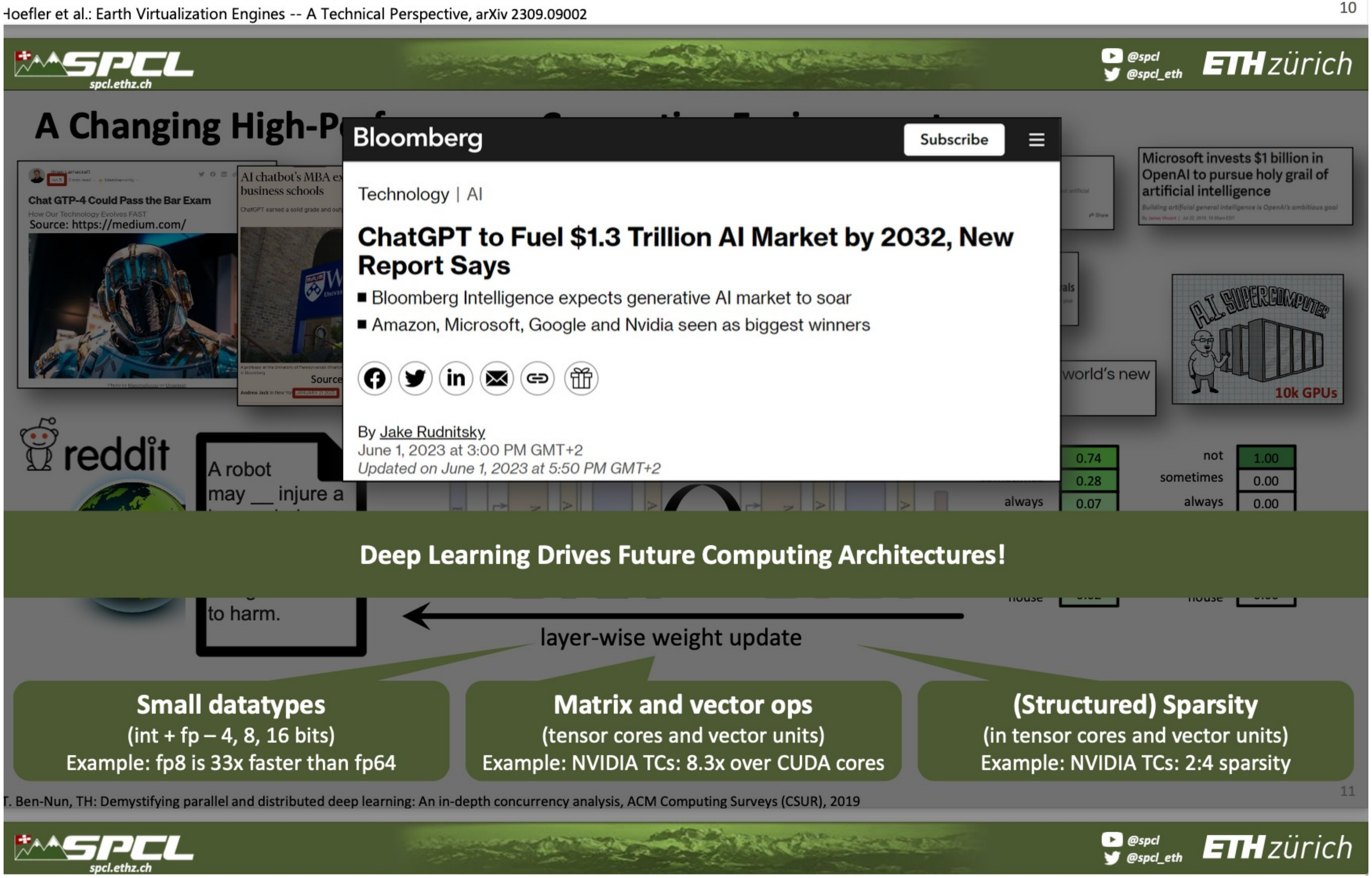
WCRP
World Climate Research Programme

Contents

1. Overview of trends in HPC / weather & climate preparation for Exascale
2. GPU adaptation, single precision, cloud use & I/O / workflow acceleration
3. Maintainability / Performance portability
4. Annex: provided slides from members & other groups

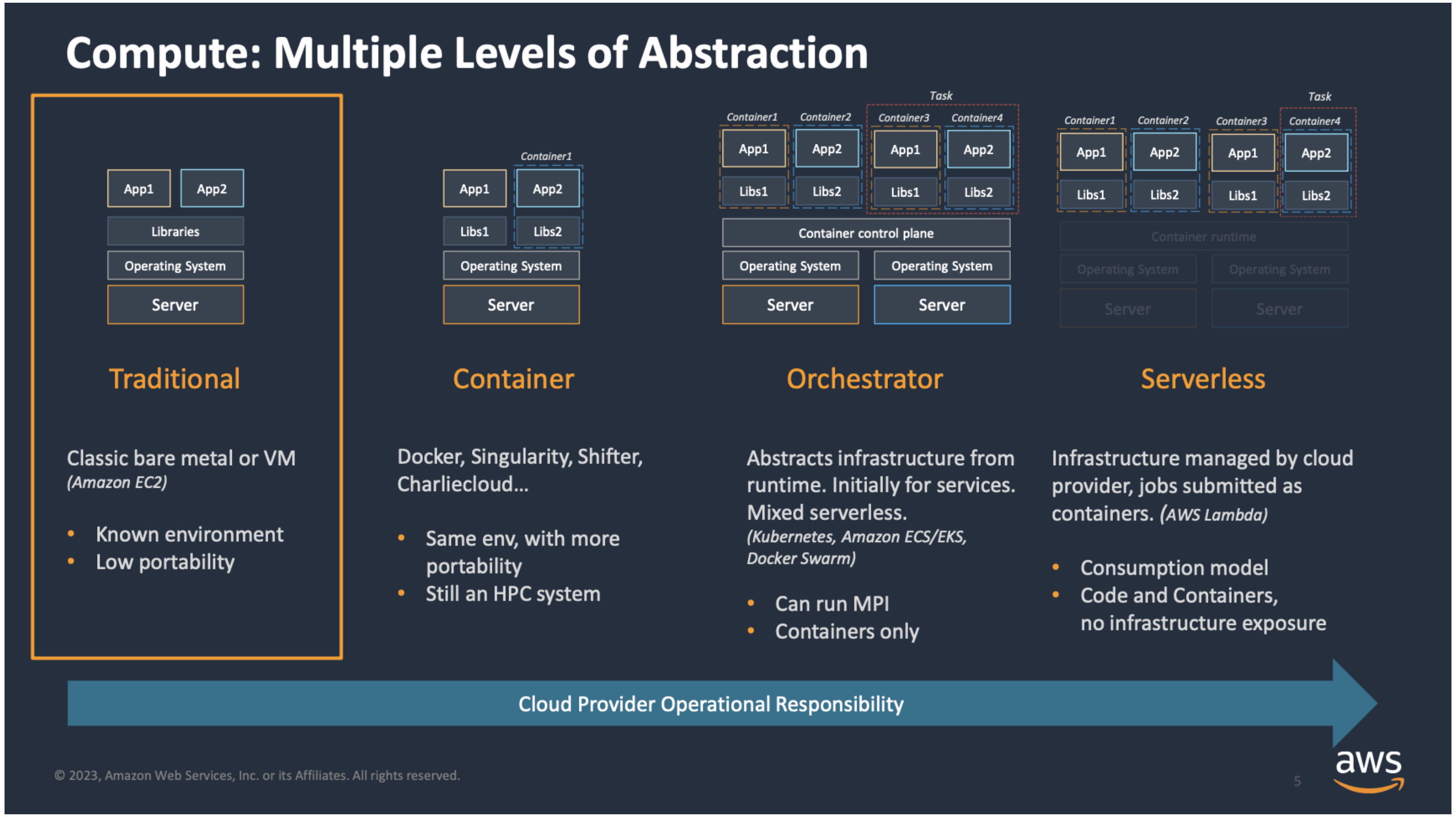
Trends from 20th ECMWF workshop on high performance computing in meteorology

https://events.ecmwf.int/event/329/


Hoefler et al.: Earth Virtualization Engines -- A Technical Perspective, arXiv 2309.09002
10
SPCL
spcl.ethz.ch
@spcl
@spcl_eth
ETH zürich
A Changing High-P
Chat GTP-4 Could Pass the Bar Exam
How Our Technology Evolves FAST
Source: https://medium.com/
AI chatbot's MBA ex
business schools
Source
Bloomberg
Subscribe
Technology | AI
ChatGPT to Fuel $1.3 Trillion AI Market by 2032, New Report Says
Bloomberg Intelligence expects generative AI market to soar
Amazon, Microsoft, Google and Nvidia seen as biggest winners
By Jake Rudnitsky
June 1, 2023 at 3:00 PM GMT+2
Updated on June 1, 2023 at 5:50 PM GMT+2
Microsoft invests $1 billion in OpenAI to pursue holy grail of artificial intelligence
world's new
A.I. SUPERCOMPUTER
10k GPUs
reddit
A robot
may __ injure a
to harm.
not
sometimes
always
0.74
0.28
0.07
1.00
0.00
0.00
Deep Learning Drives Future Computing Architectures!
layer-wise weight update
Small datatypes
(int + fp – 4, 8, 16 bits)
Example: fp8 is 33x faster than fp64
Matrix and vector ops
(tensor cores and vector units)
Example: NVIDIA TCs: 8.3x over CUDA cores
(Structured) Sparsity
(in tensor cores and vector units)
Example: NVIDIA TCs: 2:4 sparsity
T. Ben-Nun, TH: Demystifying parallel and distributed deep learning: An in-depth concurrency analysis, ACM Computing Surveys (CSUR), 2019
11
SPCL
spcl.ethz.ch
@spcl
@spcl_eth
ETH zürich

Cloud provides …

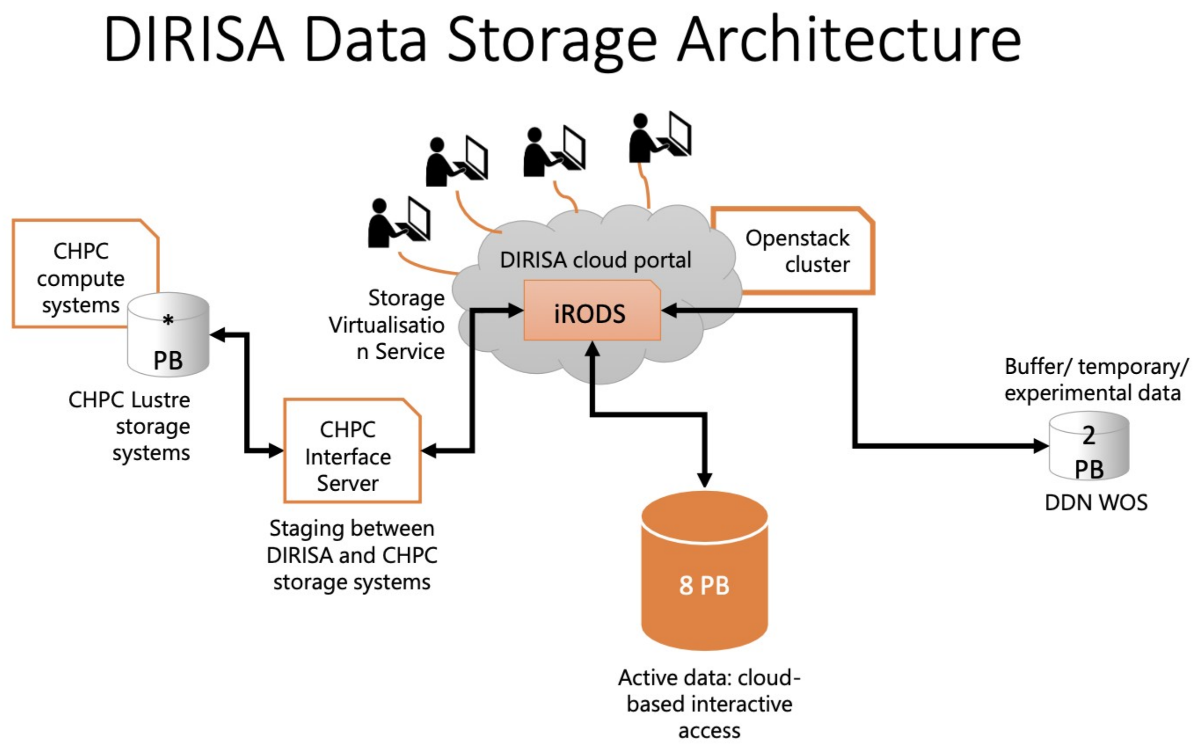
AWS abstraction levels

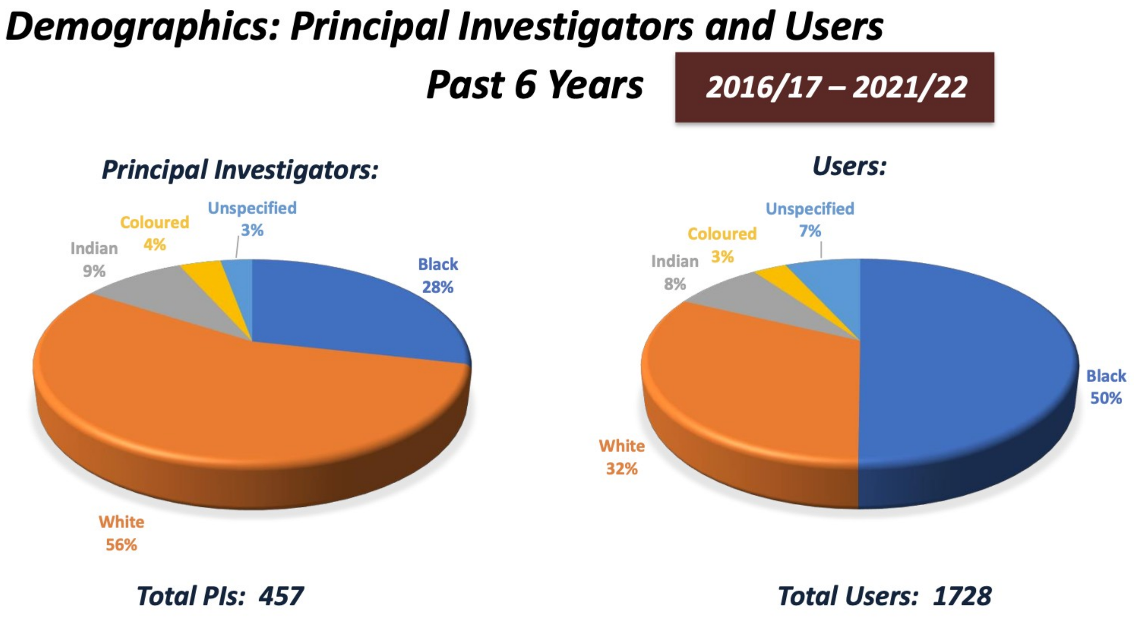


South Africa example on cloud-based system

Leveraging diversity to push the boundaries of computing

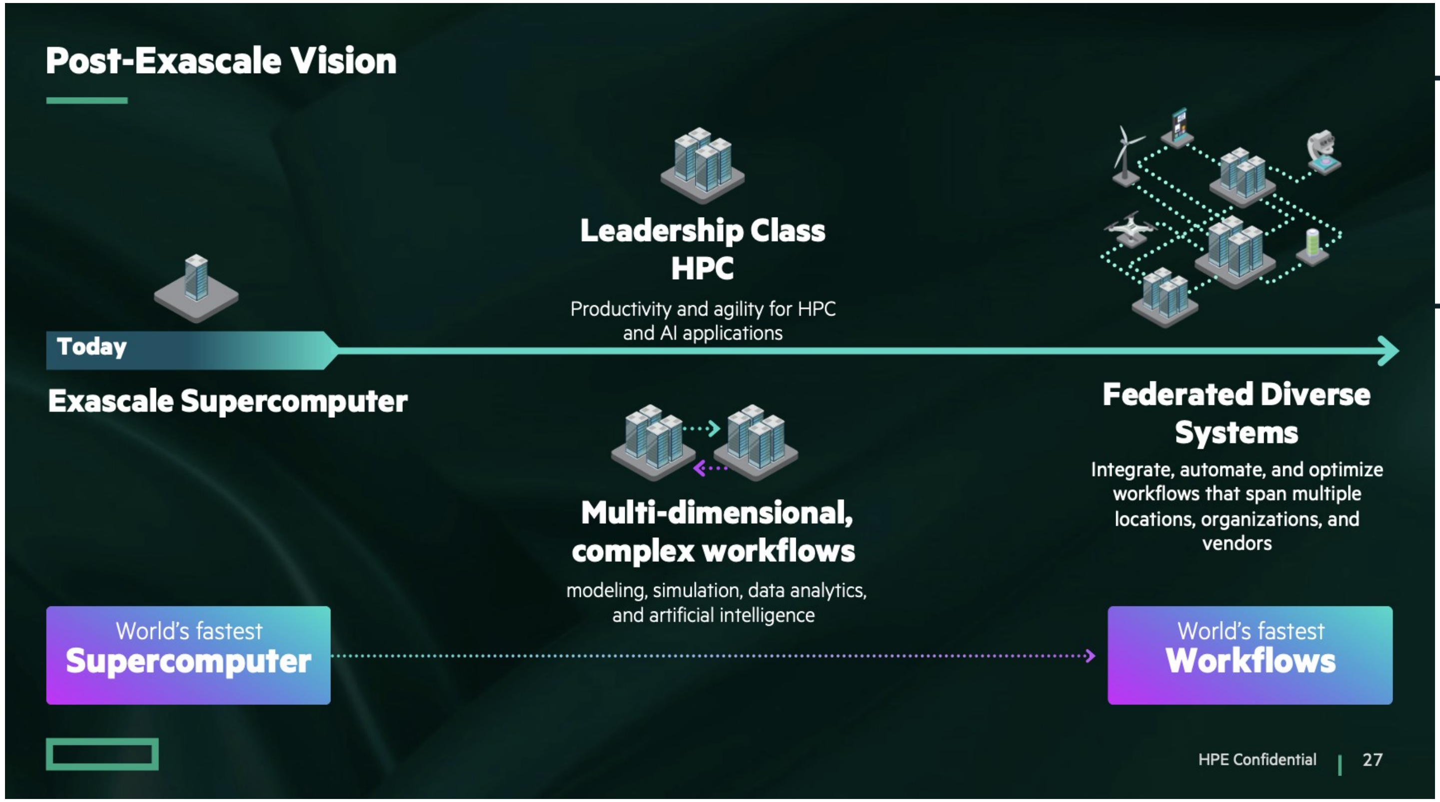
Mthetho Vuyo Sovara

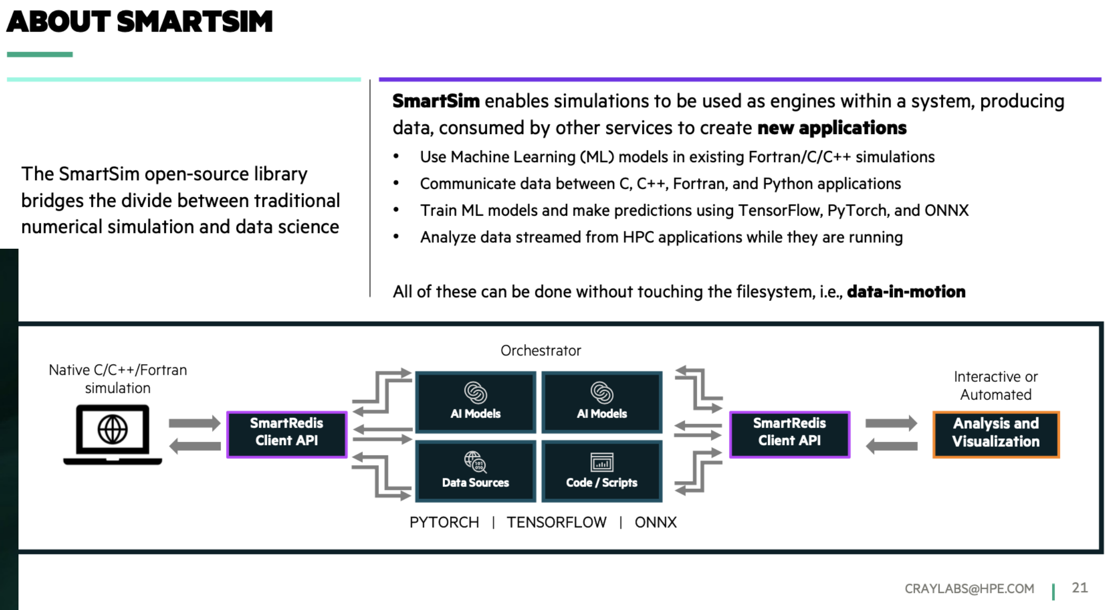




Substantial increase & reaching wider range of users

Accelerate Workflows



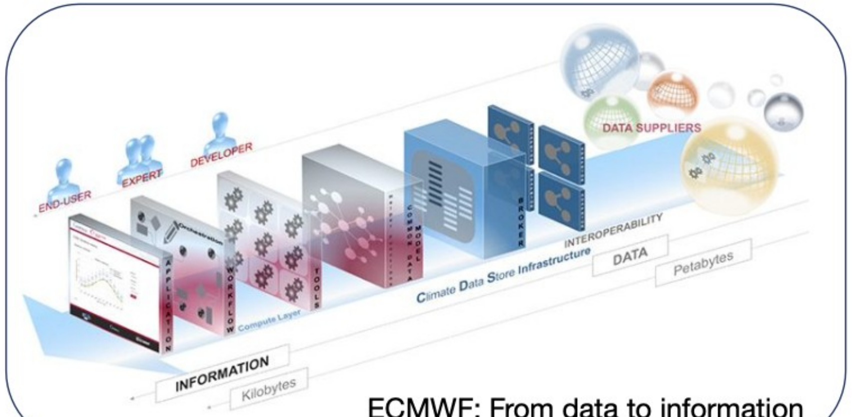


Carpenter et al

Different institutes – similar problems

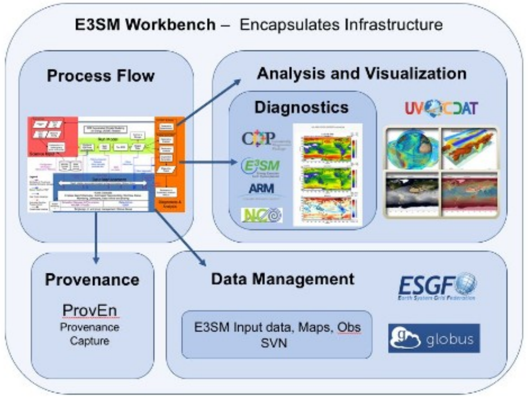
Modernization of modeling software	Co-design between scientists and computer scientists	Data Science, new workflows, AI/ML, data management	Investment in software engineering, scientific design

Adapted from UK Met Office & EPSRC: Harnessing Exascale Computing



ECMWF: From data to information

DOE E3SM



DOE Exascale project

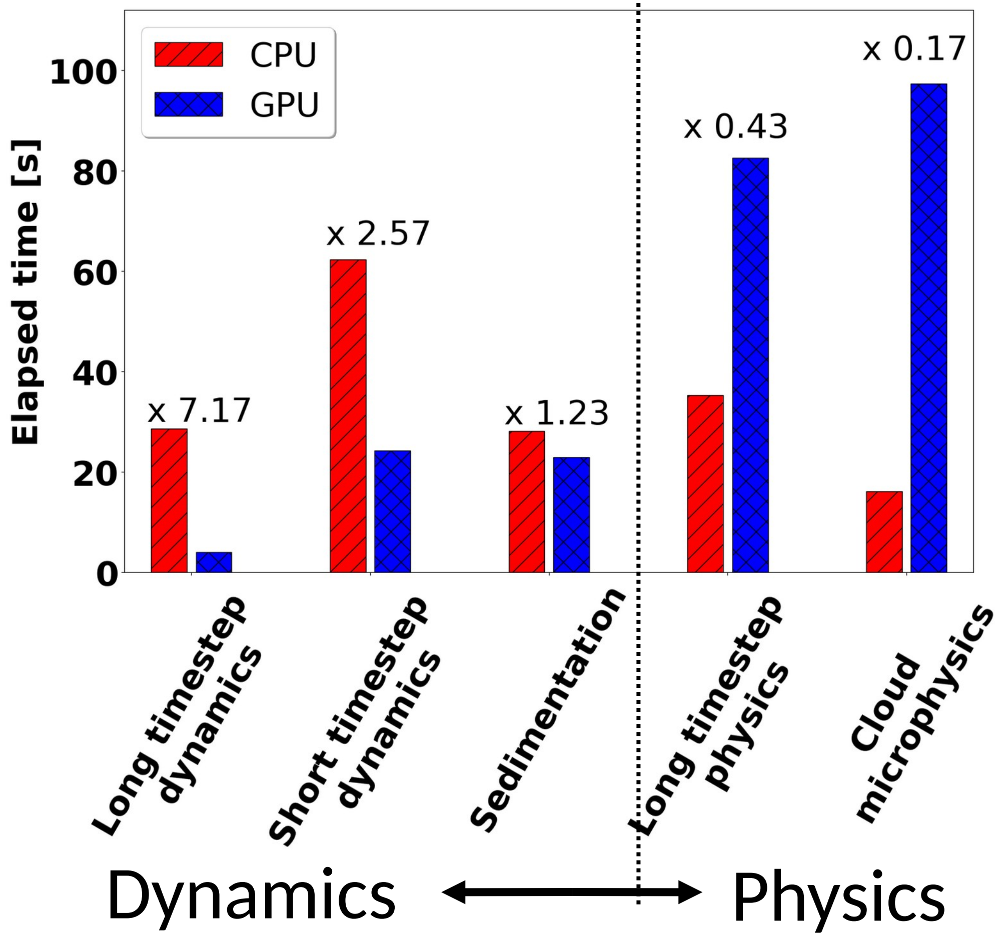
Many opportunities for cross-agency and international partnership on tools and methods

NCAR
UCAR

Hauser et al

Results of GPU porting

CPU: Intel Xeon Gold 6226 2.7GHz 12C/24T x2 with DDR4 memory (140GB/s)
GPU: NVIDIA Tesla V100-PCIe-32GB x1 with HBM2 memory (900GB/s)
(nx,ny,nz) = (150,150,76) ~ grid size / node in the operational configuration



Dynamical process
- ✓ **High** acceleration rate
- ✓ **A few** workloads
- ✓ Issues : Optimizing MPI communications

Physical process
- ✓ **Low** acceleration rate
- ✓ **A lot of** workloads
  - ➢ Modifying a lot of work arrays to arguments
- ✓ Issues : Hard to vectorize innermost loop, Saving GPU memory

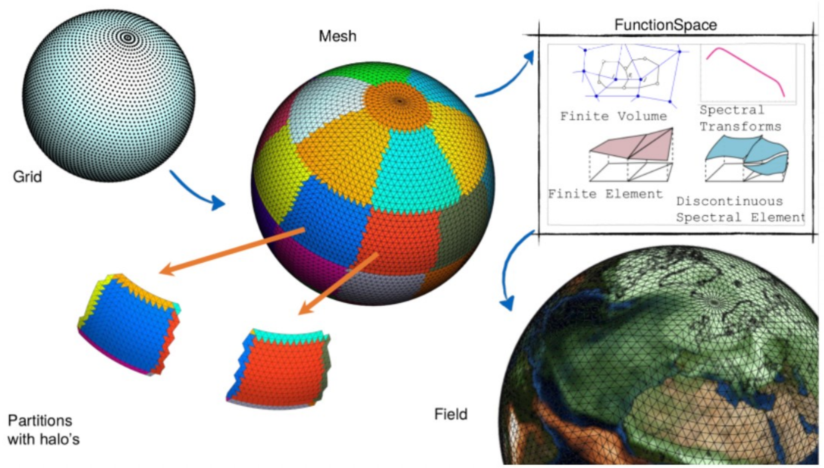
HIGH PERFORMANCE COMPUTING

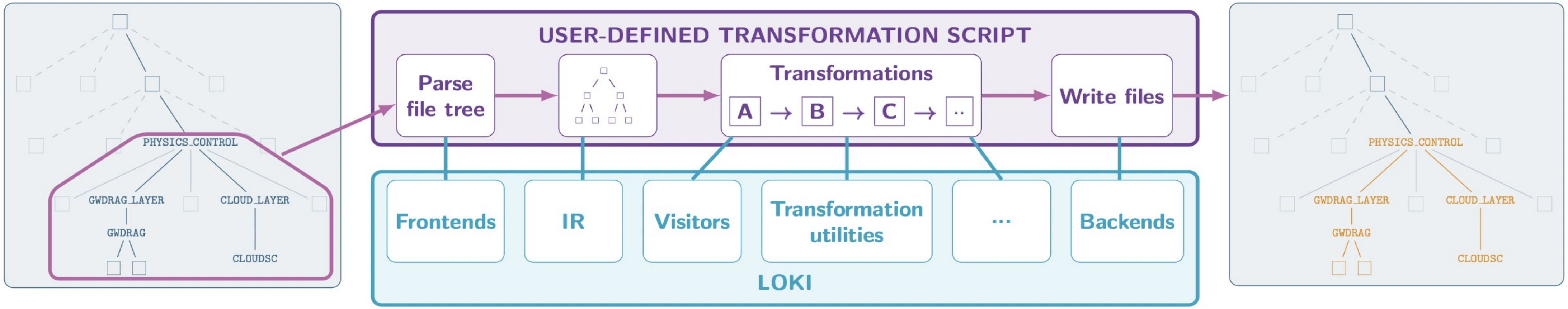
Atlas – A library for NWP and climate modelling

- Modern C++ library with Fortran interfaces
- Data structures for numerical algorithms:
  - **Increasing accelerator-awareness**

Loki - Programmable source-to-source translation package written in Python

- Library of tools and APIs to build custom transformation recipes
- Built on basic principles of compiler technology (IR trees, visitors, transformers)





IFS ADAPTATION TO EUROHPC

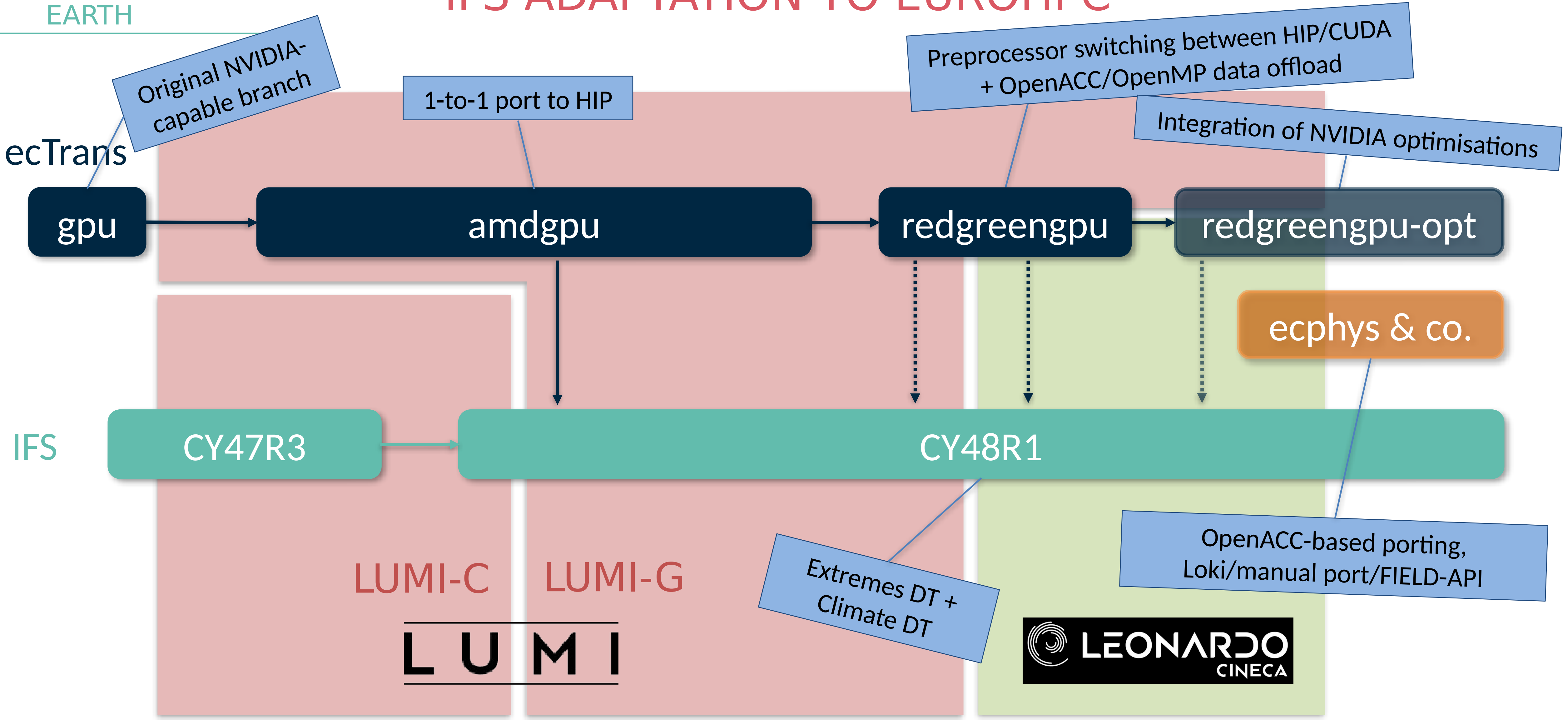


Illustration only. Size, shape, and position of boxes do not imply timelines. Subject to change.

HPC efforts at Météo-France

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

1. Operational use in all AROME[1] operational systems (forecast component only)
2. Next steps: operational use, whenever possible,
   1. in all ARPEGE forecasts
   2. in all trajectories within the assimilation cycle
   3. parts of assimilation

Adaptation to hybrid processors with accelerators:

3. Significant code refactoring (e.g., new memory structure)
4. Development of automatic source-to-source code transformation tools (Loki, Fxtran)
5. ARPEGE forecast with physics (except radiation) running on GPU-Nvidia

   >> next steps: semi-implicit, semi-lagrangian and radiation schemes
6. Work done in collaboration with ECMWF and ACCORD[2] partners
7. Hectometric LAM configuration developed within DestinE On Demand project (phase 1)
8. 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)

[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

Input from F. Bouyssel

Met Office Physics Schemes

1. Radiation - Socrates (done - ORNL)
2. Micro Physics – Casim (done - ORNL)
3. UKCA – Excalibur
4. Land Surface – Jules (Excalibur?)
5. Aerosols - RADAER
6. Boundary Layer – slow and fast
7. Convection – CoMorph
8. Stochastic physics
9. Cloud Physics
10. Spectral Gravity Wave Drag (GWD)
11. Orographic GWD

```
!$acc parallel loop collapse(2)
do j = js, je
   do i = is, ie
      ...
      !$acc loop seq
      do n = 1, nsubsteps
         do k = 1, nz
            ....
         end do
      end do
      ...
   end do
end do
```

Change loop order to increase parallelism

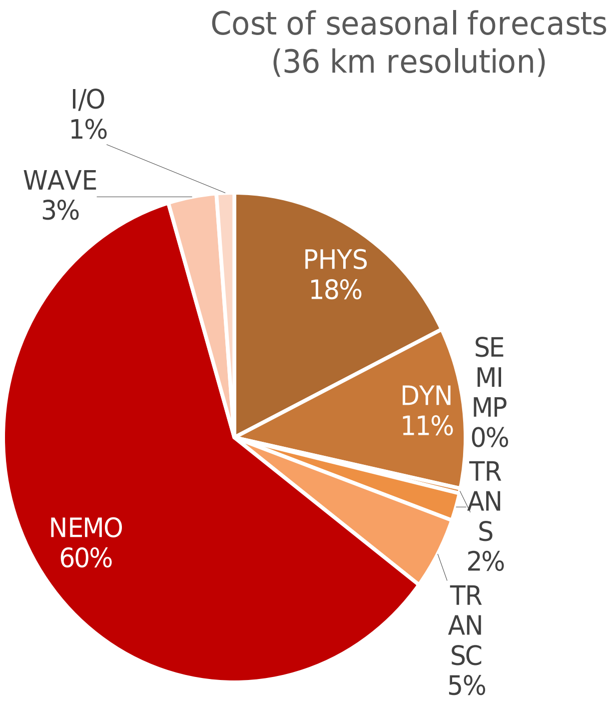
W. Zhang

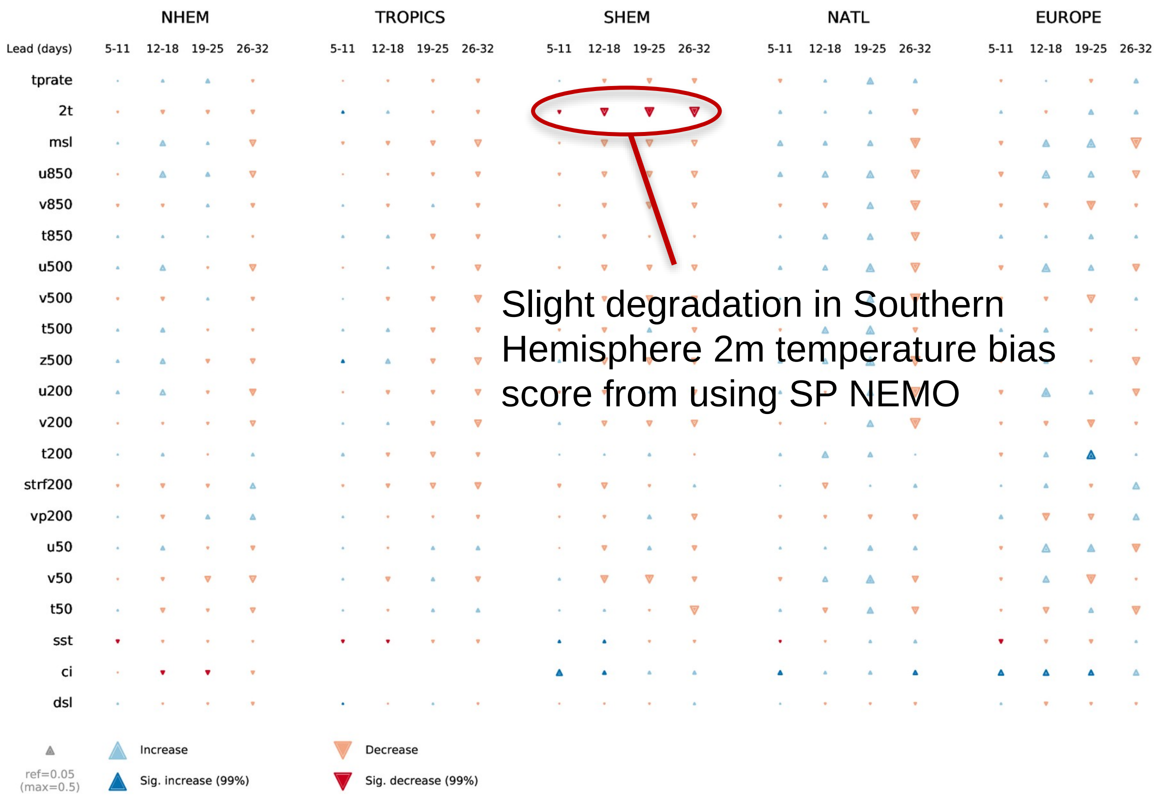
Keep single source code but don't back port to UM

```
do n = 1, nsubsteps
!$acc parallel loop
collapse(2)
   do j = js, je
      do i = is, ie
         !$acc loop vector
         do k = 1, nz
            ....
         end do
      end do
   end do
end do
```

Single precision in the ocean

- **For seasonal runs the ocean is over half the cost** – could we use single precision there too?
- Coupled tests show minimal differences up to seasonal lead times
- However, some small but detectable effects on ocean mean state from single precision





Mean absolute bias card comparing single precision with double precision NEMO, from extended-range coupled hindcast experiments ¼° ocean, 50 km atmosphere

Hatfield, Mogensen and

Single precision in SOM advection

- SOM (Second Order Moment) advection (Prather, 1986) for tracers
  - High accuracy, but one of the bottlenecks in MRI.COM (~30% of total costs)
  - A scalar variable with a grid cell (:grid size) is expressed as second-order orthogonal polynomials, then the moments are advected.

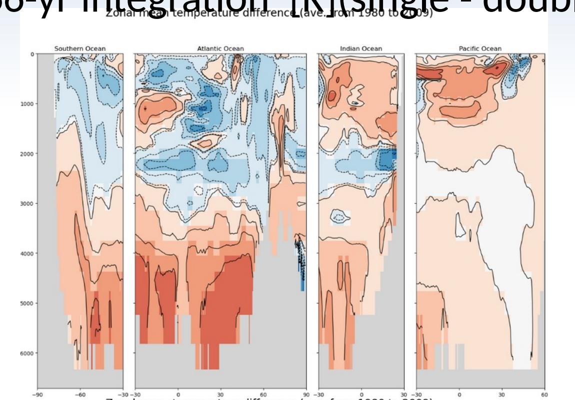
$$\phi(x) = m_0 + m_1 K_x + m_2 K_{xx}$$

$$K_x = x - \frac{X}{2}, \; K_{xx} = x^2 - xX + \frac{X^2}{6}$$

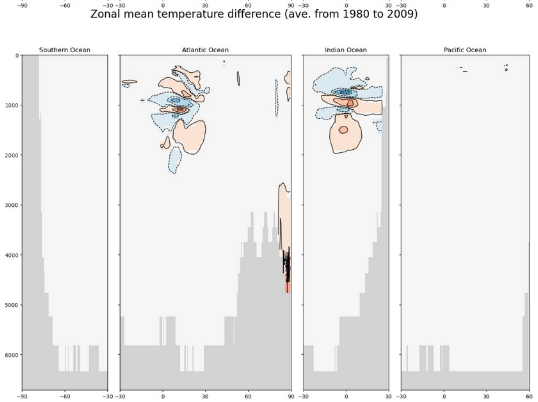
    - Less risks of "loss of digits" than finite difference methods
- Several variables need to be kept as double precision to obtain both speed up and accuracy.

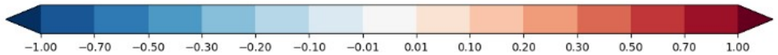
Zonal mean temperature difference for the last 30-yr average of the 366-yr integration [K](single - double)

Full single in SOM



Single in SOM but only mass) and volume kept double





Double

	ave[sec]
%	
TOTAL	1.30E+03
100.00	
TRACER	4.50E+02
34.67	

Single in SOM (mass and volume: double)

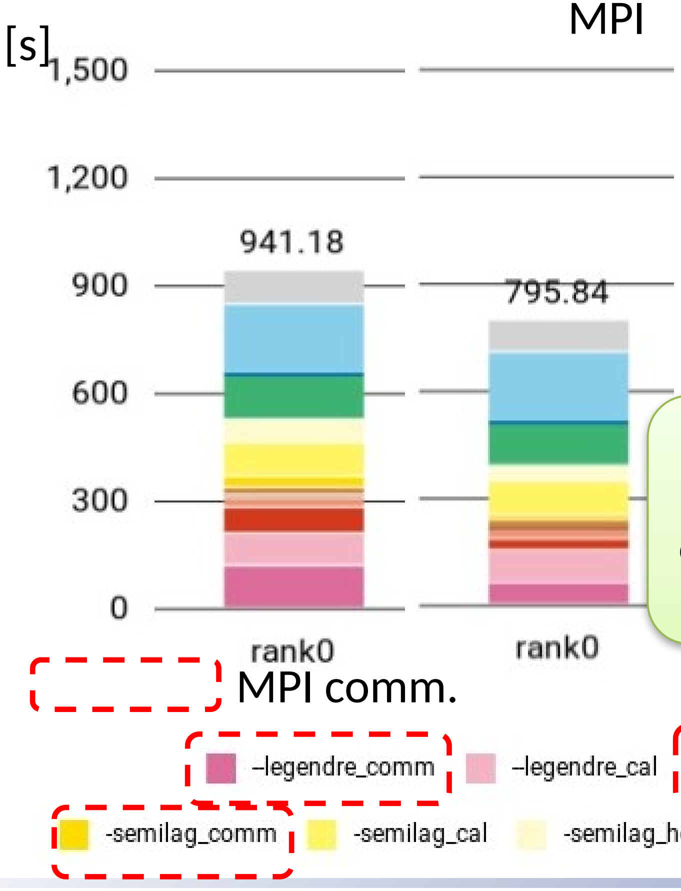
	ave[sec]
%	
TOTAL	1.12E+03
100.00	
TRACER	3.23E+02
28.76	

Speed up by 30% in the tracer schemes

Impacts of single precision (only MPI) GSM

Computational costs of Tq959L128 GSM

double precision | single precision MPI

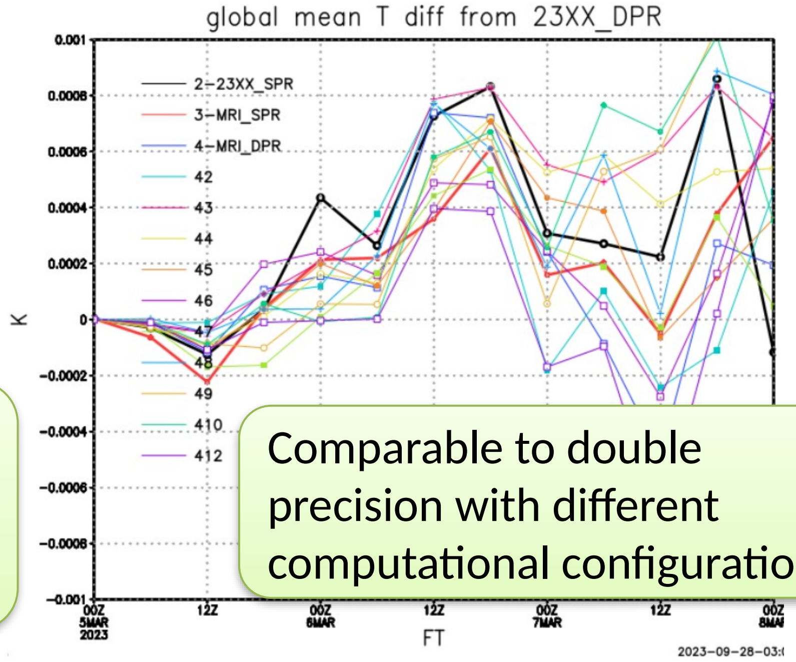


*Speed up by 15%
*Costs for MPI comm. almost halved

Global mean T at 500hPa diff.

Red and black : single precision MPI GSM

Others: double-precision GSM with different compilers / compile options / libraries (FFT, matrix)



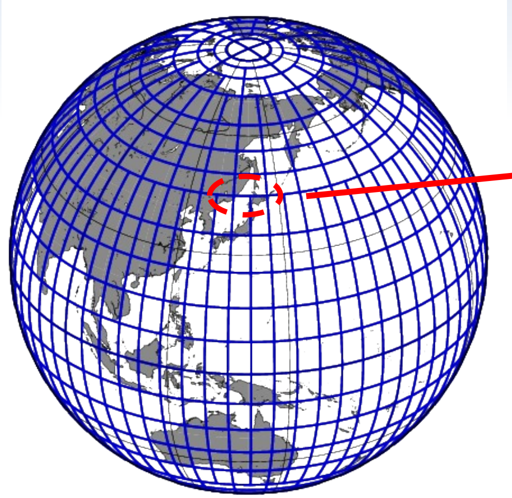
Comparable to double precision with different computational configuration

Preparing for exascale computing at Hydrometcentre of Russia

- Since 2022, more single precision computations and memory access optimizations in many routines: current gain of ~18% for the global SLAV10 model (operational since 10/23, 0.1° lon, 0.08-0.13° lat, 104 levs) at Cray XC40. These works are ongoing

- Data handling: offline compression using **ncks** utility of output NetCDF files, with compression depending on variable

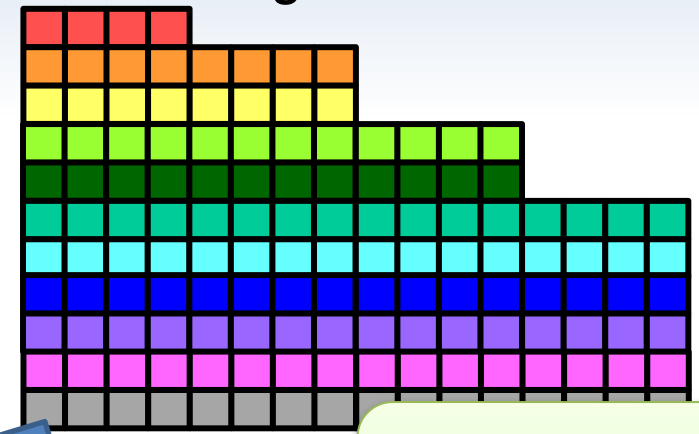
Flexible (i, k, j) array structure for non-stencil calculation (e.g. physics, I/O etc) for both CPU and GPU



Horizontal grid boxes in a MPI process

Lon.

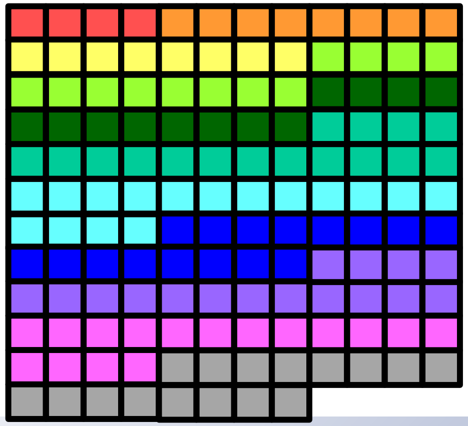
Lat.



Max array size of array in the "i" direction Is controlled by a parameter "**NUMI_I**"

NUMI_I=12

For CPU with OpenMP
(larger outermost loops for thread parallelization)



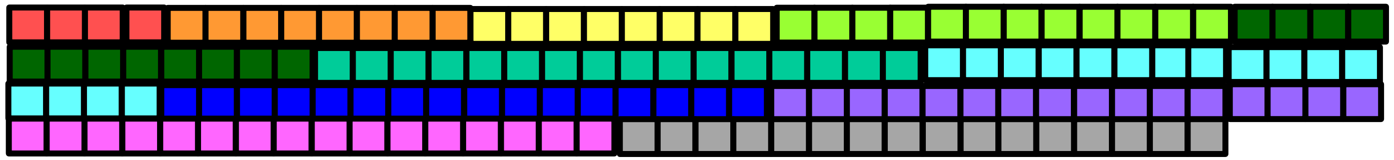
NUMI_I=36

For Vector machines or GPU with OpenACC:
(larger innermost loops for vectorization)

i

j

NUMI_I



CMC HPC/Exascale Projects: Science

- Planned upgrades to the existing modelling system:
  - SLIMEX : semi-Lagrangian Implicit-Explicit time integrator
    - Combine SL and IMEX BDF2 time integrator
    - Second order in time, no extra off-centering and only one elliptic solve per step
  - Single-core performance evaluation and optimization for the physics
- 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
- Building a network of academic collaborators for the development of a hybrid physical/AI NWP system
  - Establish a close working relationship with the universities and private sector – joint projects in the next 5 years.
  - Goal: explore the spectrum of possibilities in applying numerical methods and ML to develop an optimal hybrid NWP model that will use the best of the two worlds.
  - Develop approaches in high-performance computing – the best numerical algorithms on today's supercomputer could be suboptimal in the future – and work with GPUs.

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)

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, Whole Atmosphere Model (WAM) for Space Weather Prediction
  - ***Regional***: HAFS (hurricane), RRFS (regional NWP), Online-CMAQ (air quality), Atmospheric River (AR).

Improvement for I/O and computational efficiency

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

HPC upgrade

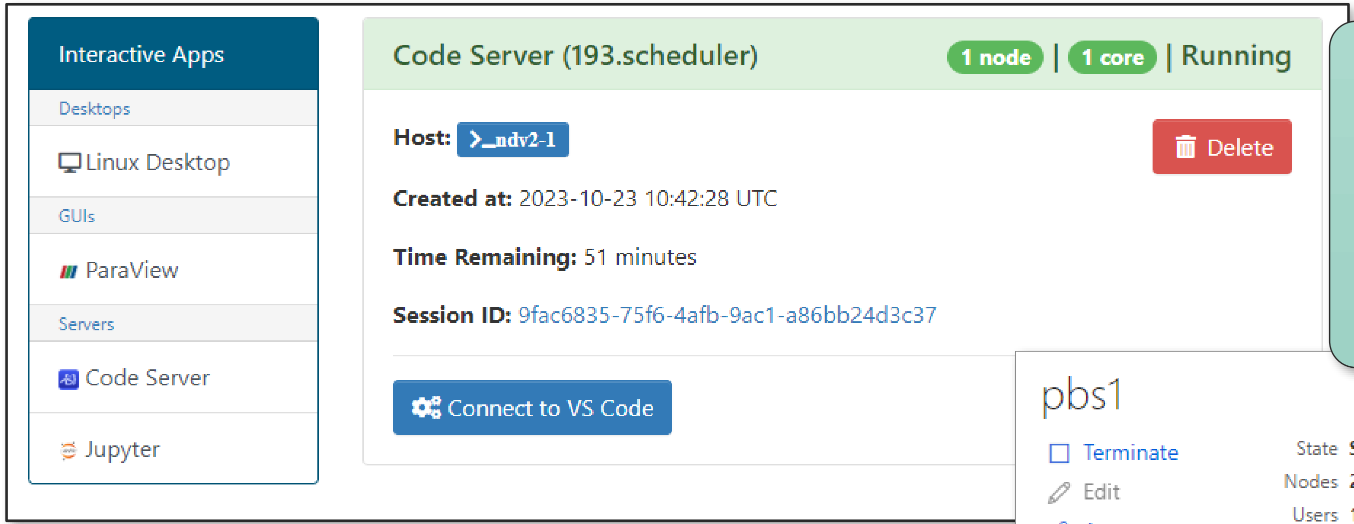
- **Old:** WCOSS, Dell, 73K x 2 cores, 4302 x2 TF peak performance
- **New as of June 2022**: WCOSS2, CRAY EX, 2560x2 nodes, 327Kx2 cores, 12,100 x2 TF peak performance.
- **New as of Aug 2023:** WCOSS2, CRAY EX, 3060x2 nodes, 392Kx2 cores, 14,400 x2 TF peak performance.

On the Cloud

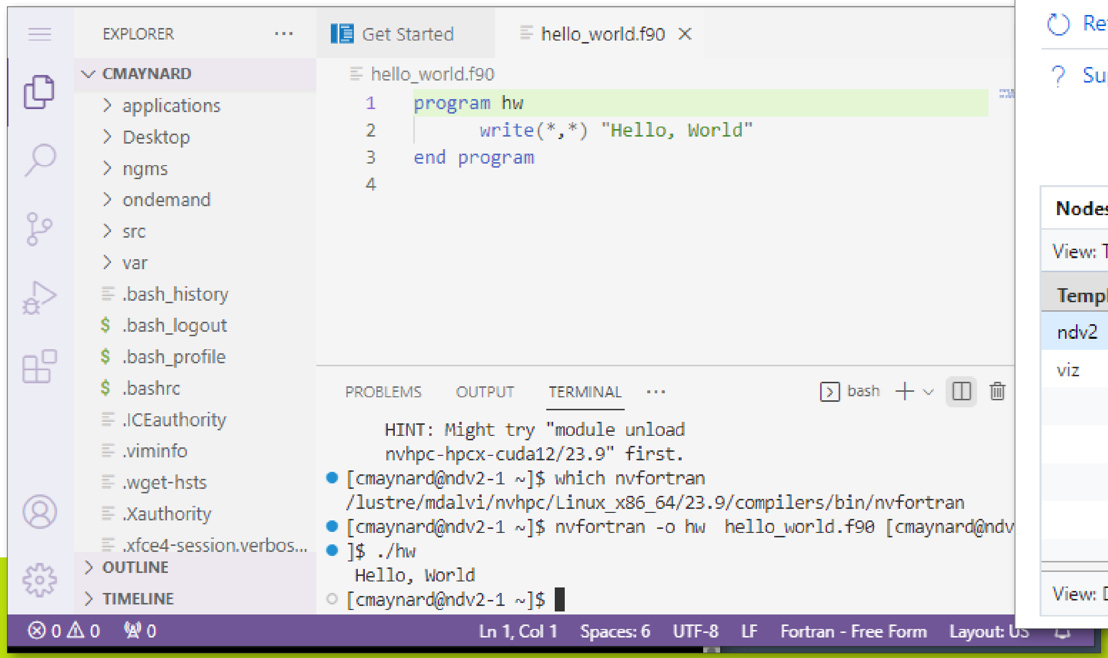
- **Running experimental hurricane ensemble forecast (HAFS) and regional high-res ensemble forecast (RRFS) on the Cloud.**
- **Plan to run global ensemble GEFS.v13 reanalysis and reforecast on the Cloud as well.**

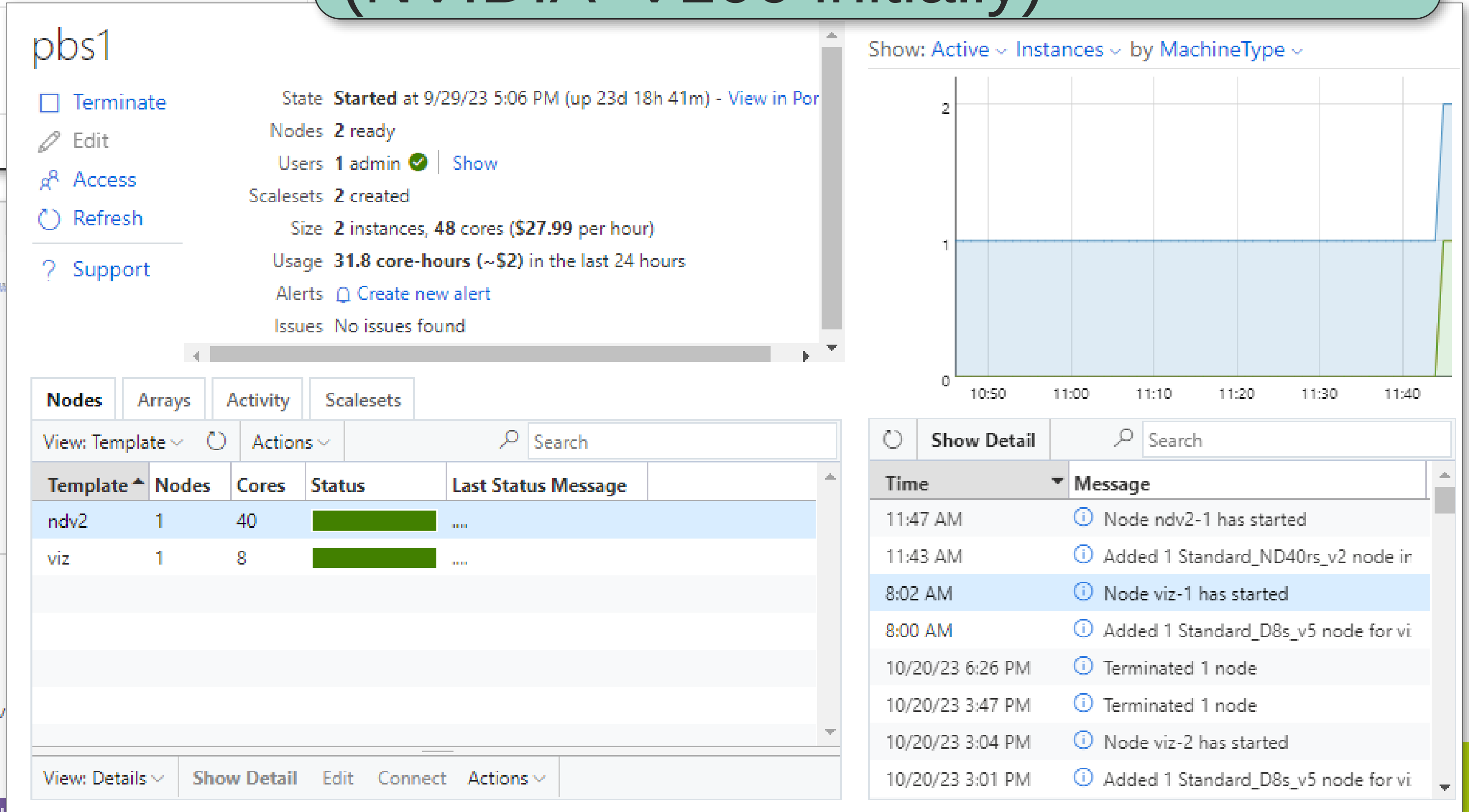
Met Office

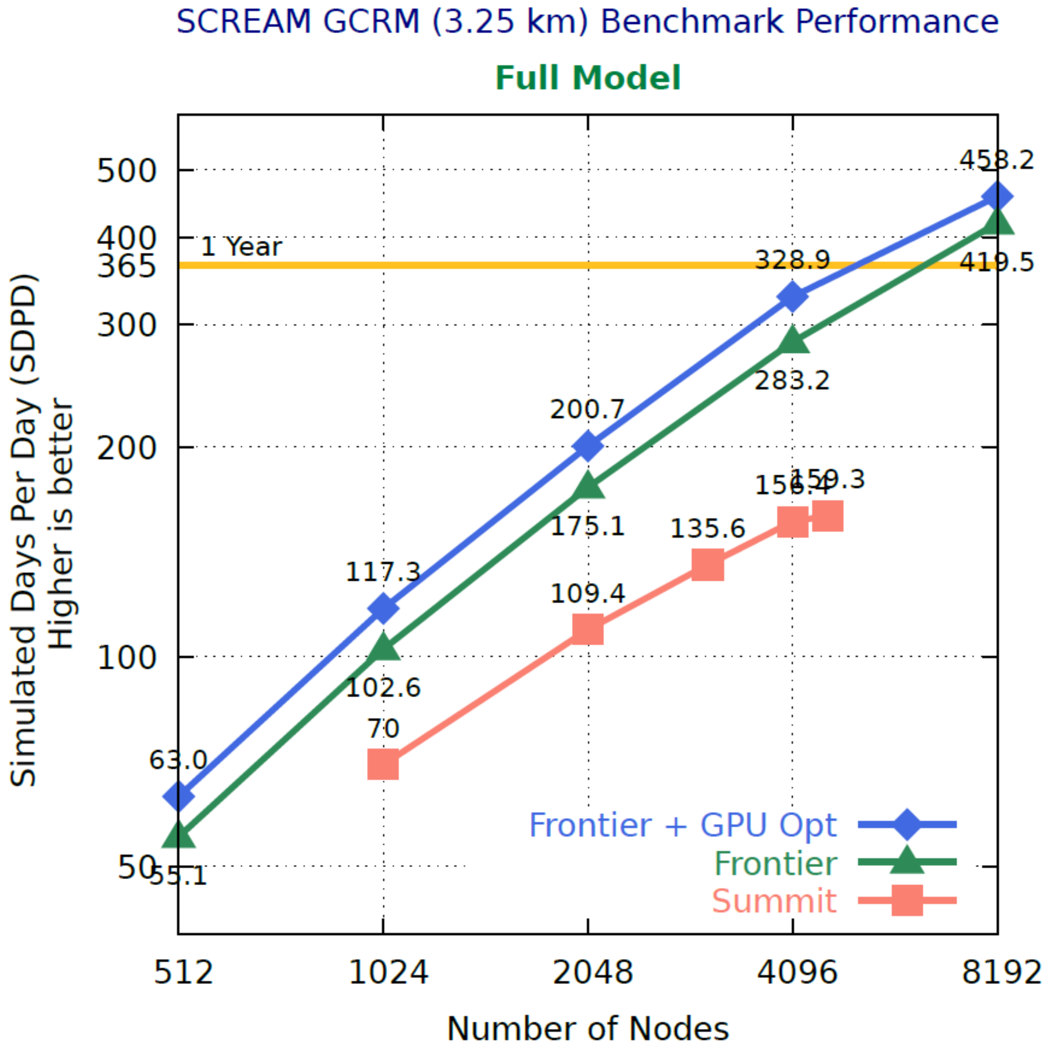
Azure HPC On-Demand Platform

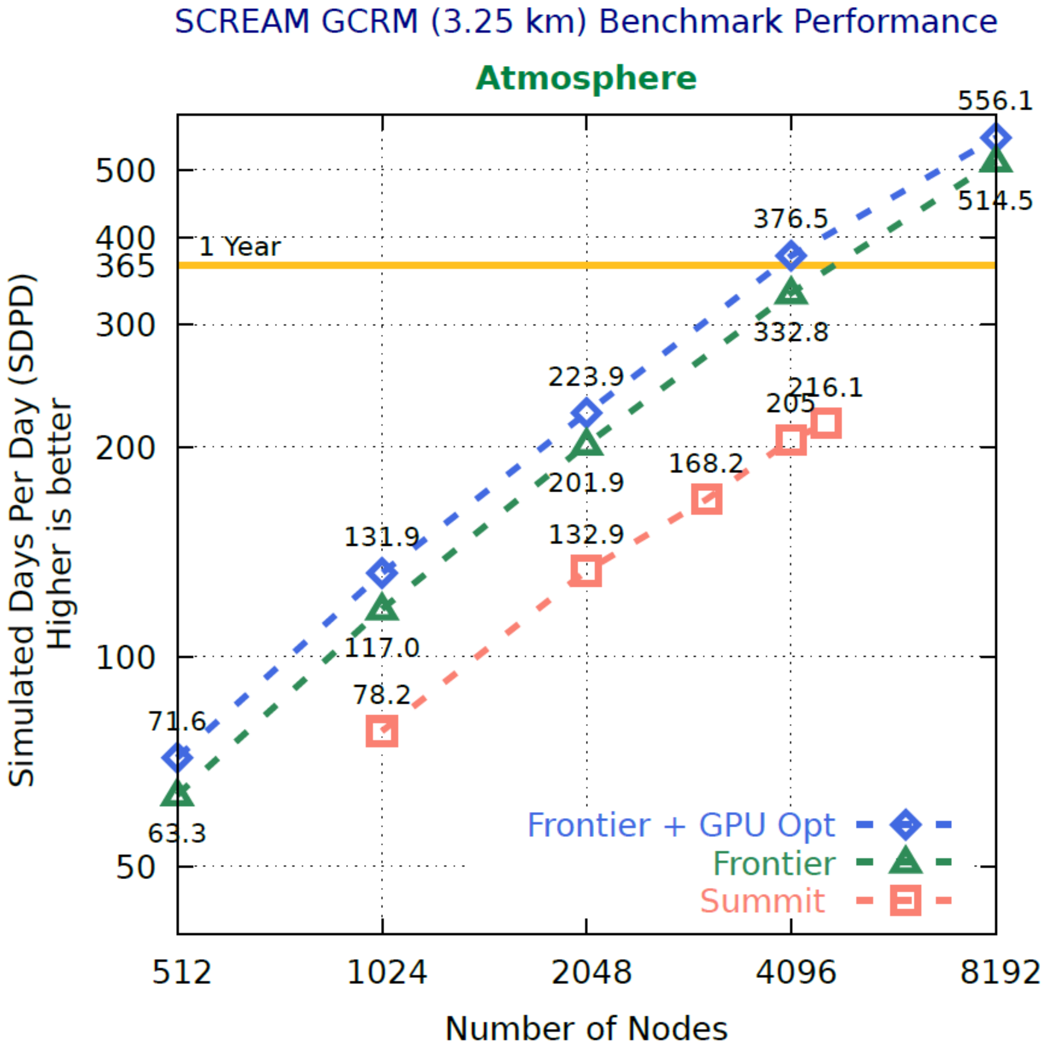


Common platform for LFRic, Marine and UKCA based GPU
10 instances – 6 GPU each
(NVIDIA V100 initially)






SCREAM GCRM (3.25 km) Benchmark Performance
Full Model
Simulated Days Per Day (SDPD)
Higher is better
1 Year
Number of Nodes
512
1024
2048
4096
8192
Frontier + GPU Opt
Frontier
Summit
458.2
419.5
328.9
283.2
200.7
175.1
117.3
102.6
63.0
55.1
70
109.4
135.6
159.3

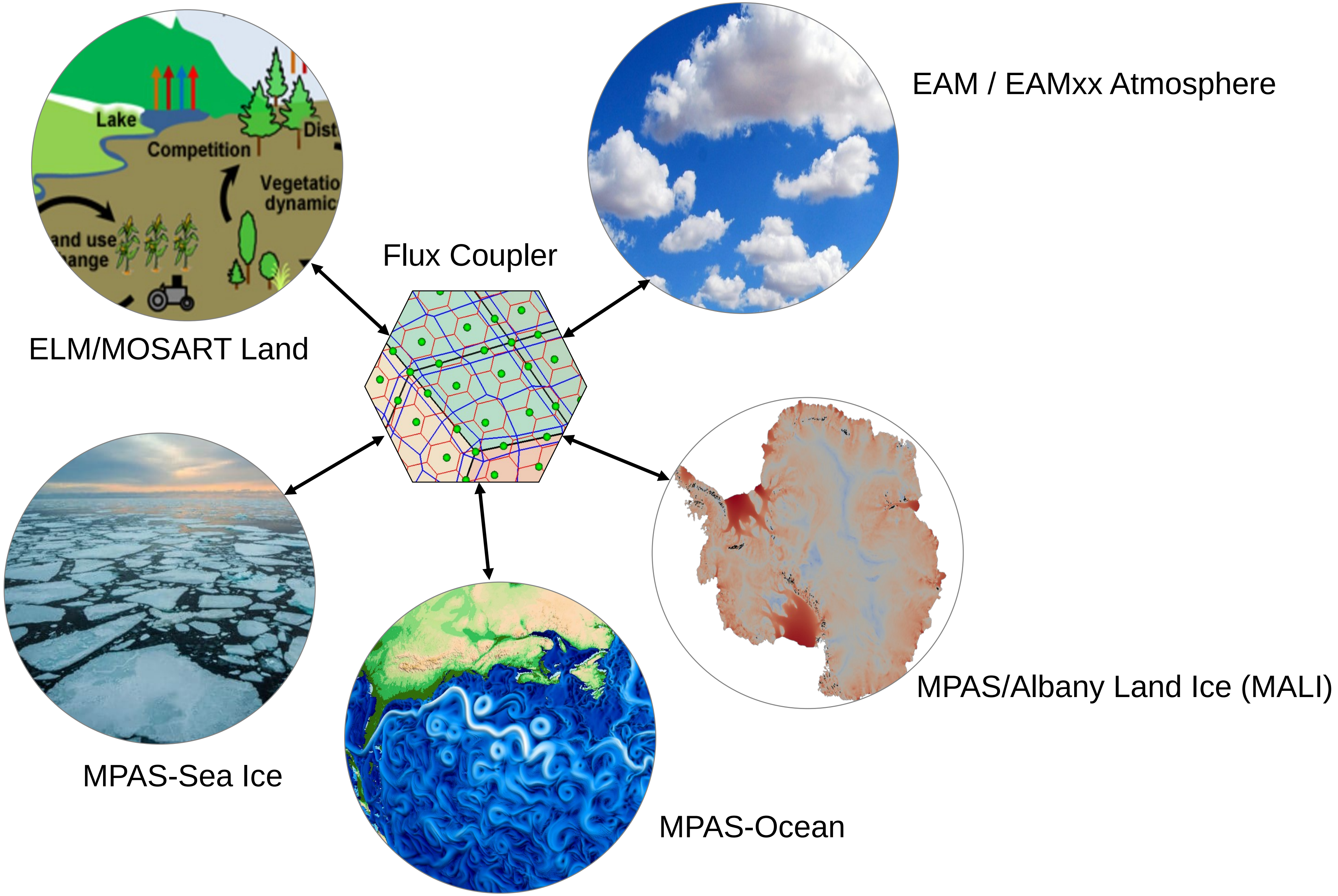

SCREAM GCRM (3.25 km) Benchmark Performance
Atmosphere
Simulated Days Per Day (SDPD)
Higher is better
1 Year
Number of Nodes
512
1024
2048
4096
8192
Frontier + GPU Opt
Frontier
Summit
556.1
514.5
376.5
332.8
223.9
201.9
216.1
205
168.2
131.9
132.9
117.0
78.2
71.6
63.3

Summary

- First Global cloud-resolving model (GCRM) to run on an Exascale supercomputer
- First GCRM to run at scale on both NVIDIA and AMD GPU systems (and hopefully soon Intel GPUs)
- First nonhydrostatic GCRM to exceed 1 simulated-year-per-day (SYPD) of model throughput
  - 2023-2024: Running some of the first decadal length cloud resolving simulations

Horizontal resolution	Vertical resolution	No. of $p = 3$ spectral elements	timestep dynamics	timestep physics	dof dynamics	dof physics
110 km	128 Layers	5400	300s	1800s	6.2M	2.8M
3.25 km	128 Layers	6.3M	8.33s	100s	7.2B	3.2B

Modern Earth System Models



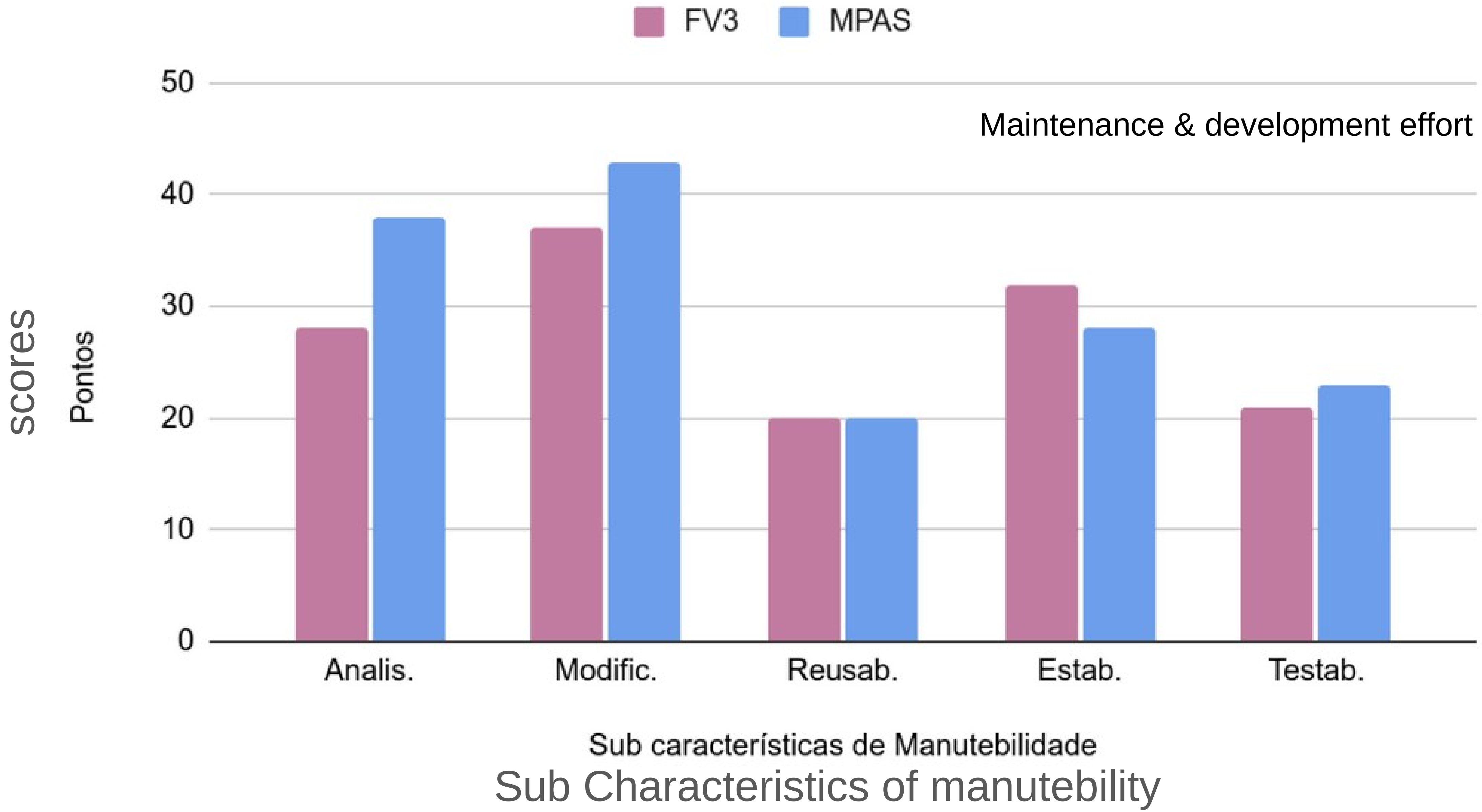
MONAN dyn core's choice

- **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.

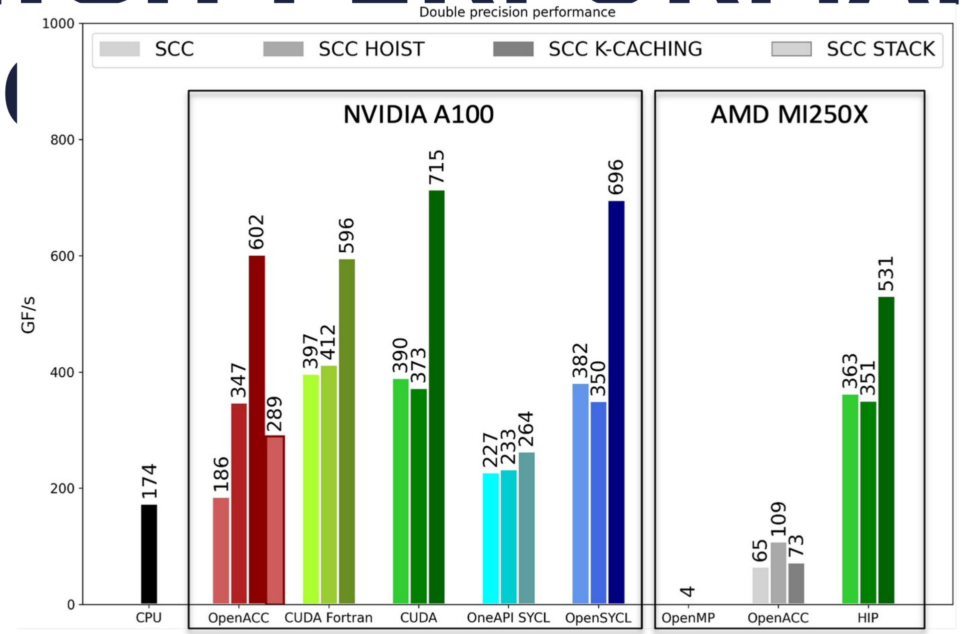
Quality of software evaluation: Model manutebility scores

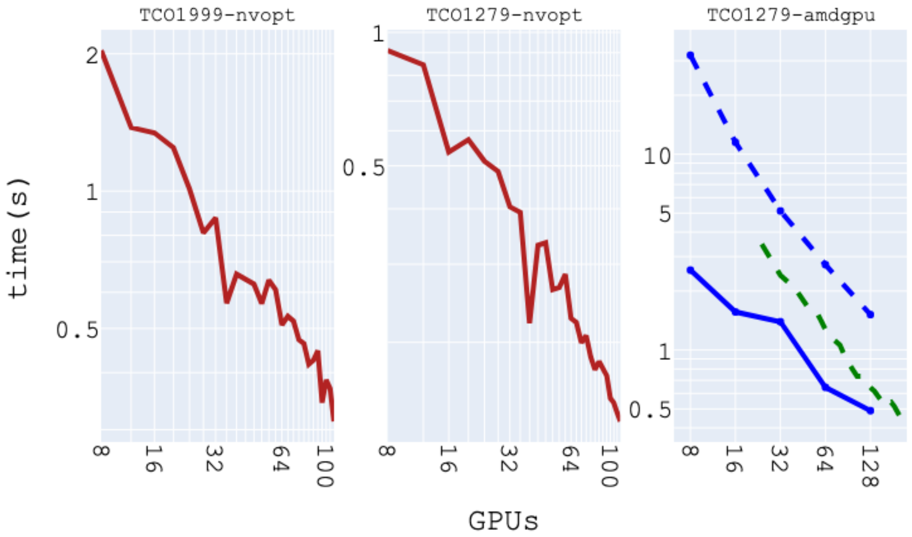


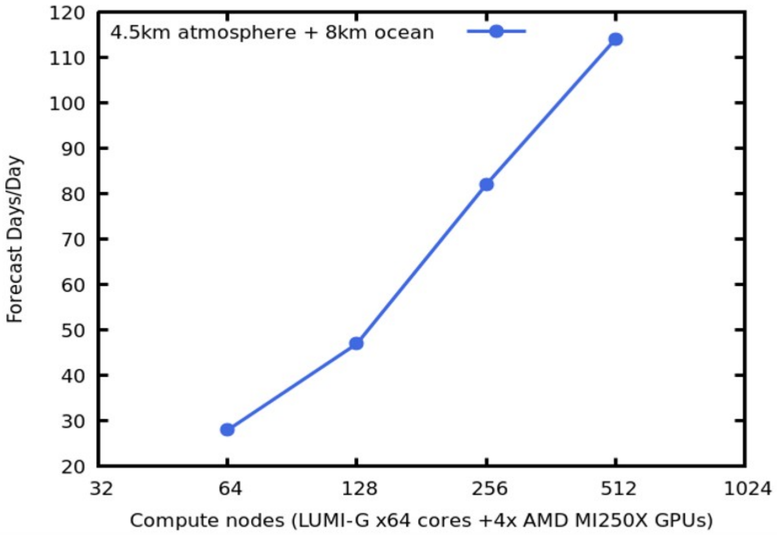
Annex

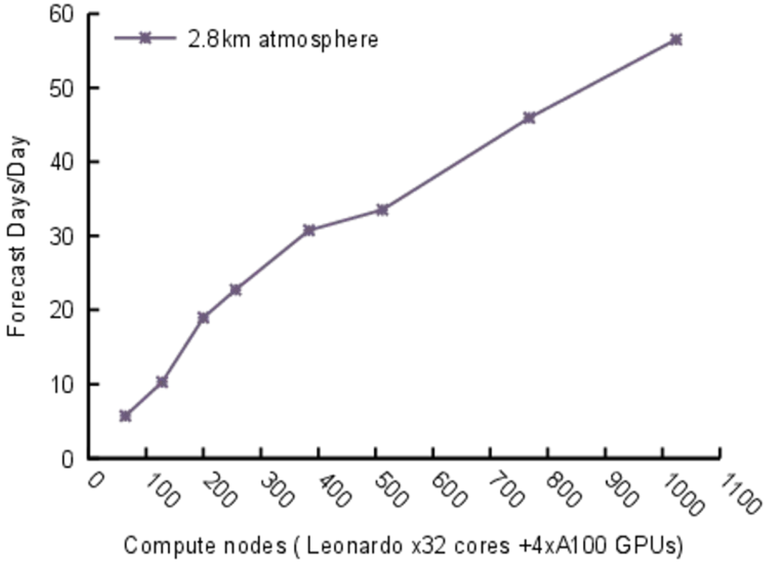
ECMWF - DESTINATION EARTH

HIGH PERFORMANCE C



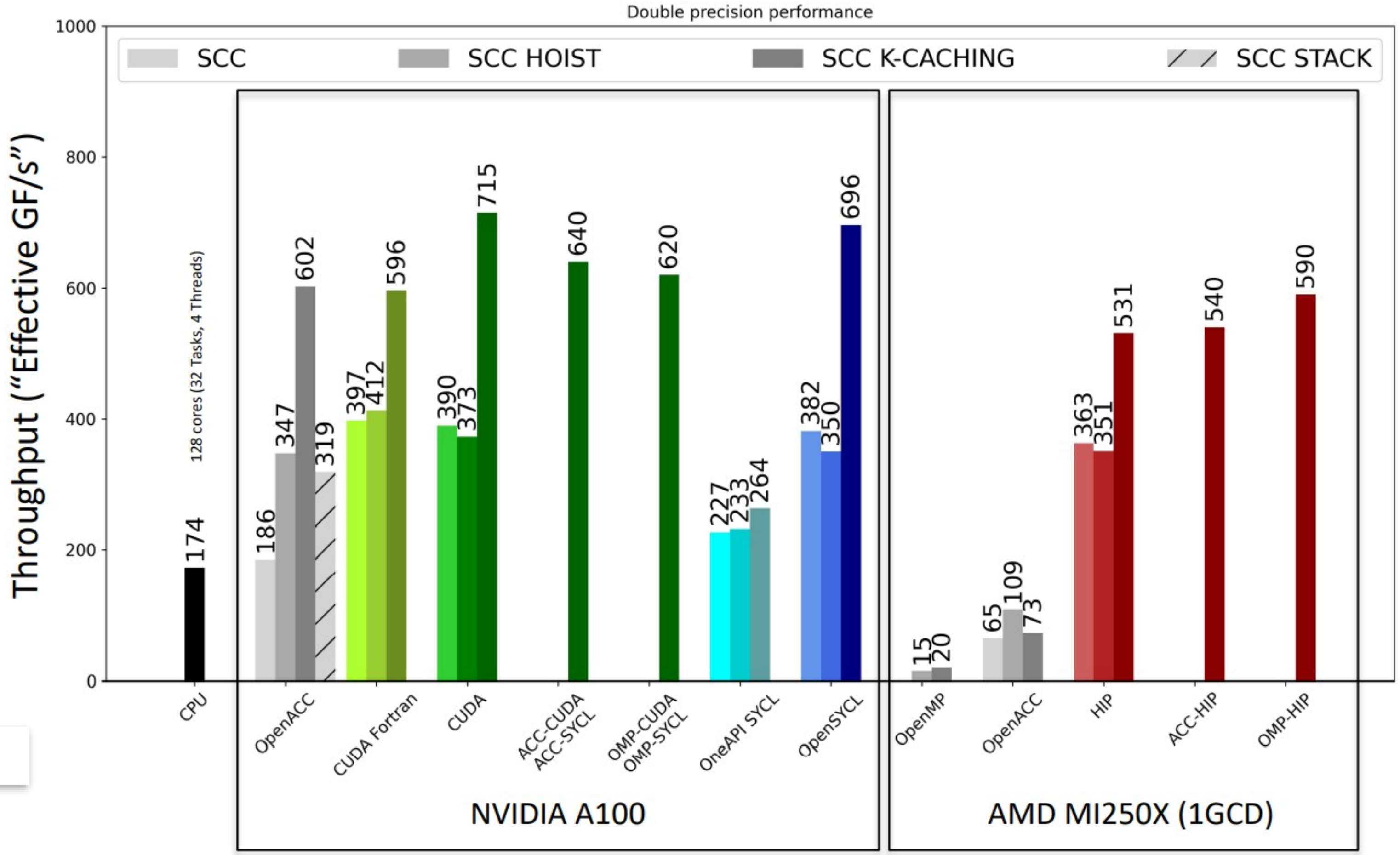






IMPROVED GPU PERFORMANCE FOR SINGLE-COLUMN ALGORITHMS

- Adaptation via source-to-source translation using Loki
- Handling of Fortran automatic arrays: **recursive hoisting** or **pool allocator**
- Speed-of-light implementation of CLOUDSC in kernel languages (**CUDA, HIP, SYCL**)
- Refactoring of Loki-SCC recipe
- Ongoing refinement and development of new adaptation recipes for further performance improvements



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 August 2023

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
- **14.4 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
  - **3,060 nodes (60 spare)**
  - 3391,680 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

State of the Union

Summer 2022 – *ported* Gravity Wave miniapp to NVIDIA GPU
Using NEMO openACC PSyclone transformation and hand-written OpenACC

1. Tested NVIDIA compiler
2. Tested changes to LFRic infrastructure
3. Demonstrated that it works
4. Demonstrated necessary changes to generated code

October 2023 – ported Gravity Wave miniapp to NVIDIA GPU
Using 100% PSyclone generated OpenACC
Performance (more to be done) and data movement
Roll out to Gung Ho

Met Office

State of

Physics

ORNL has worked with Socrates and Casim

Refactor to organise memory and gather loops

Add OpenACC directives

Summit P9 vs V100✉

https://doi.org/10.1145/3468267.3470612

W. Zhang is visiting HQ this month

STFC

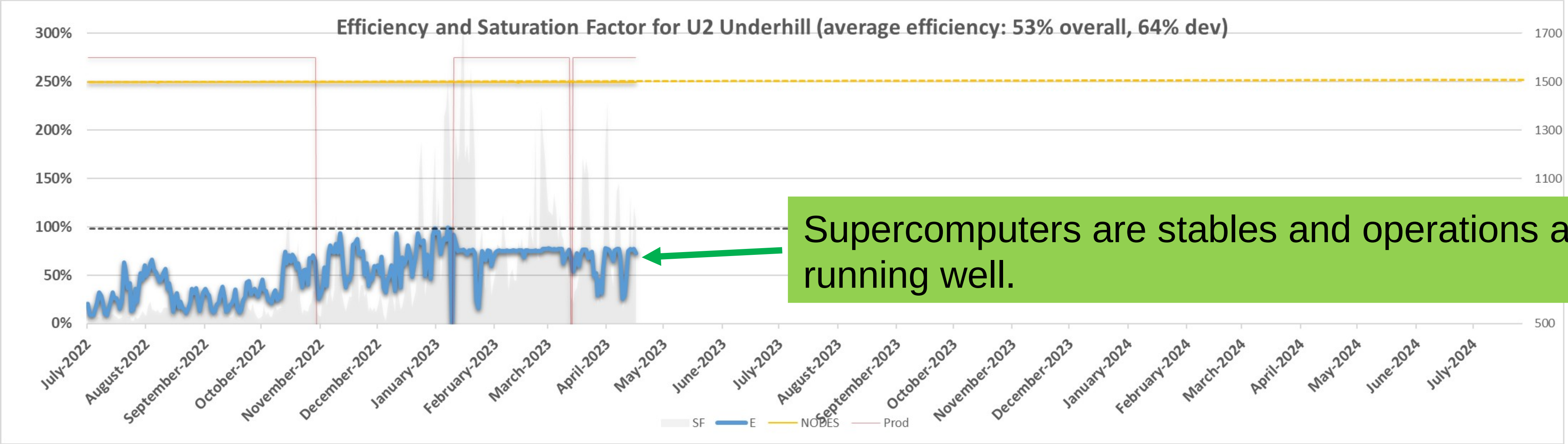
1. Strip out OpenACC directives and regen with PSyclone (Socrates works)
2. Developing OpenMP offload

High Performance Computing (HPC)

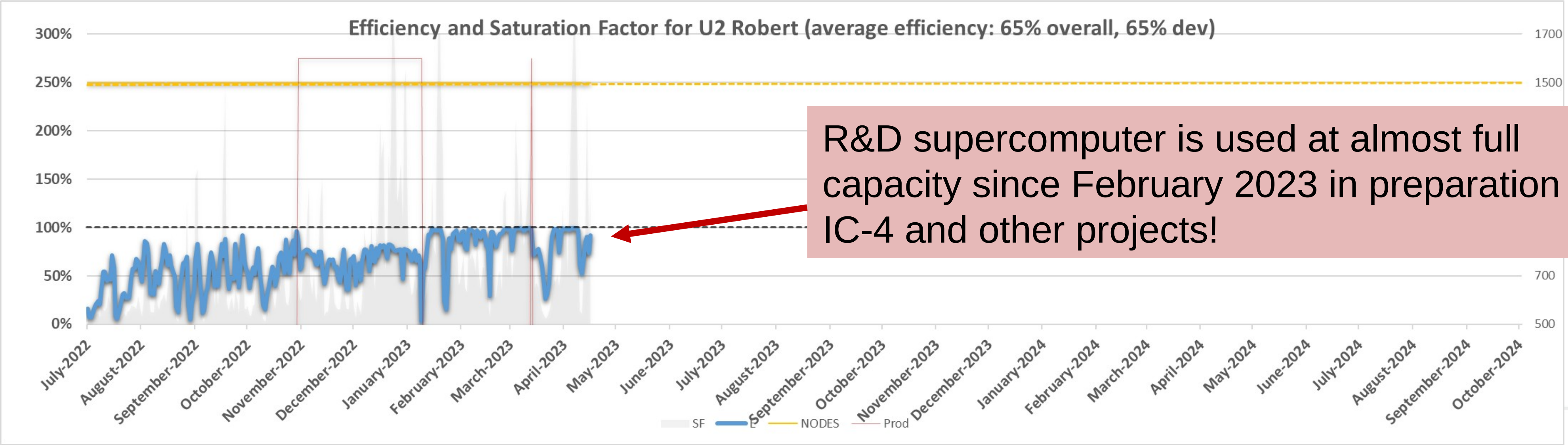
System	HPC Upgrade 1		HPC Upgrade 2 (2.5X)	
Status				
Operational	January 21, 2020		June 28, 2022	
Top 500 Entry	107 (06/2020)	115 (06/2020)	69 (06/2022)	70 (06/2022)
Top 500 Rank	249 (11/2022)	269 (11/2022)	76 (11/2022)	77 (11/2022)
Specifications				
Name	**Banting**	**Daley**	**Underhill**	**Robert**
Manufacturer	**Cray/HPE XC50**	**Cray/HPE XC50**	**Lenovo ThinkSystem**	**Lenovo ThinkSystem**
Compute nodes	**1,266**	**1,266**	**1,494**	**1,494**
Cores	**53,200**	**53,200**	**148,320**	**148,320**
Site Storage	**71 PB**		**188 PB**	
Performance (Pflops/s)				
Rmax	**2.68**	**2.60**	**7.76**	**7.76**
Rpeak	**4.09**	**4.09**	**10.92**	**10.92**

https://www.top500.org/site/50719/

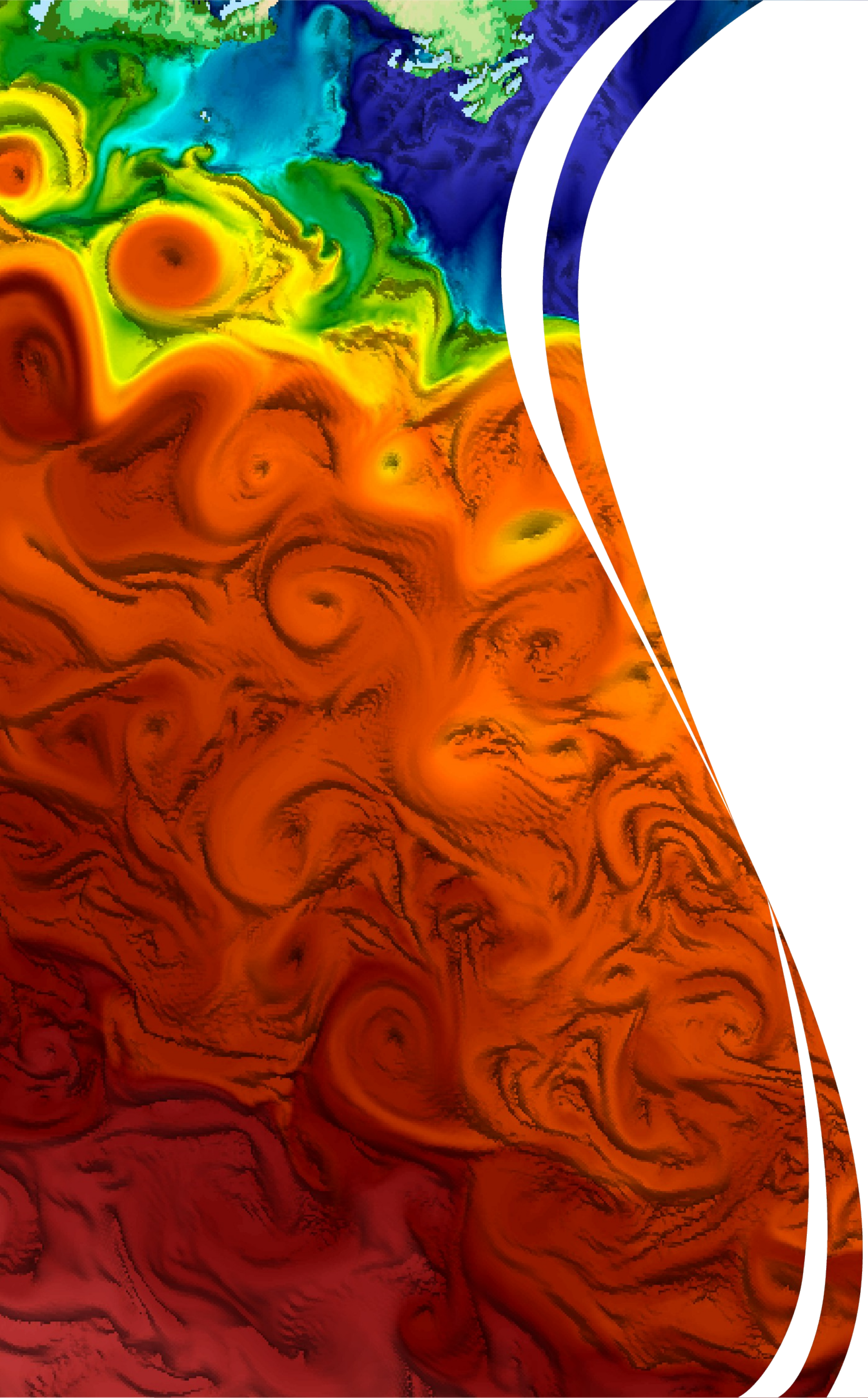
Efficiency and Saturation Factor for U2



Supercomputers are stables and operations are generally running well.



R&D supercomputer is used at almost full capacity since February 2023 in preparation of IC-4 and other projects!



Global Cloud Resolving Atmospheric Modeling on Exascale Computers

Mark Taylor (SNL), Luca Bertagna (SNL), Andrew Bradley (SNL), Peter Caldwell (LLNL), Aaron Donahue (LLNL), Oksana Guba (SNL), Noel Keen (LBNL), Sarat Sreepathi (ORNL), Trey White (HPE)

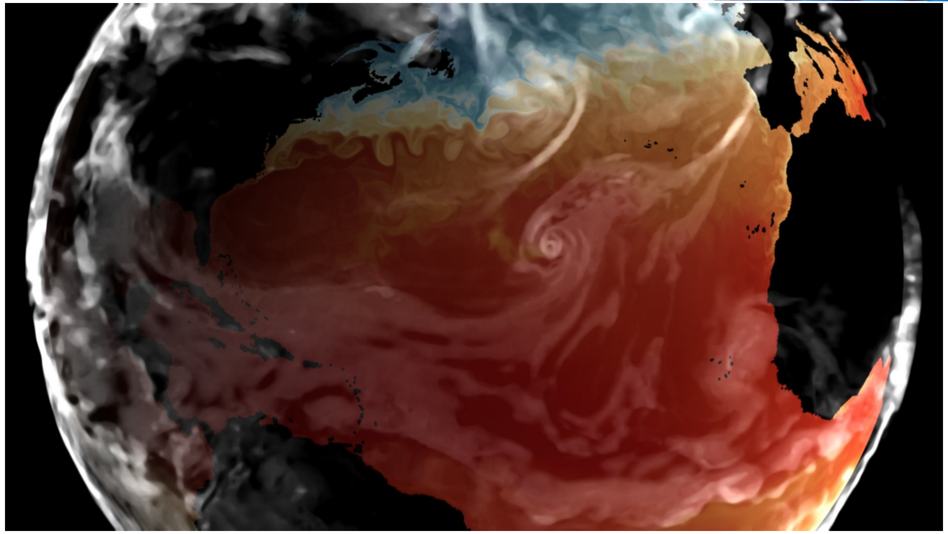
WGNE Update

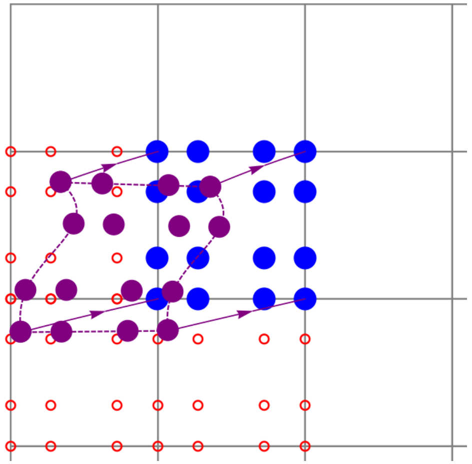
Contact: Sarat Sreepathi (ORNL), sarat@ornl.gov

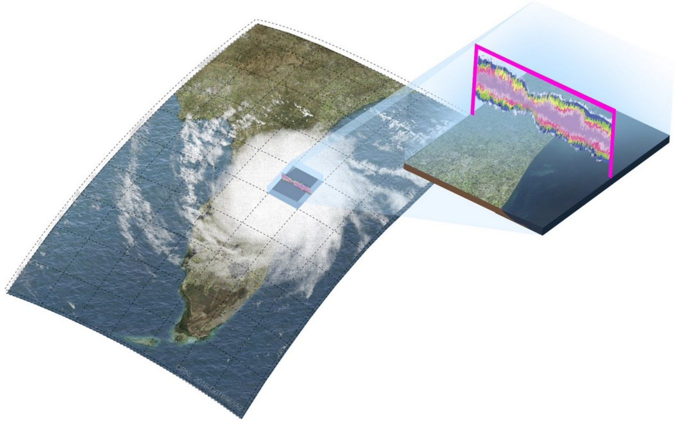
Outline

- DOE's Exascale Energy Earth System Model (E3SM) project
- SCREAM: Simple Cloud Resolving E3SM Atmosphere Model
- Porting SCREAM to Exascale machines:
  – Rewrite from scratch C++/Kokkos
  – Fortran vs C++ performance
  – Scaling to exascale

E3SM Model Development Funding



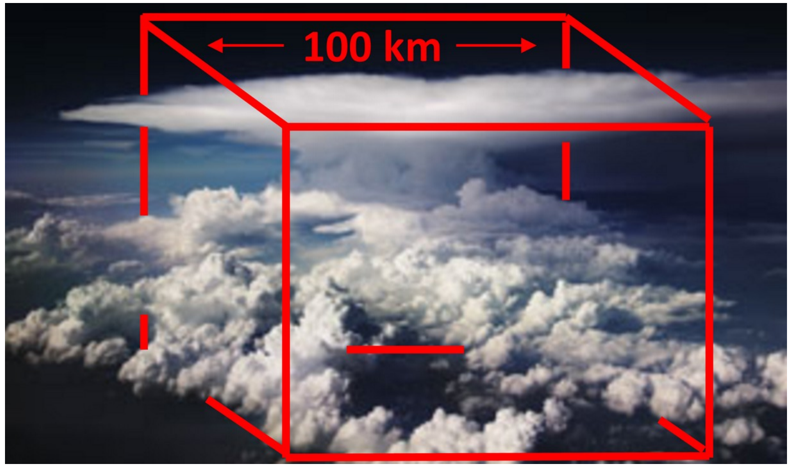




- **BER-ESMD: E3SM Project**
- ~50 FTEs, 8 labs + Universities
- **E**nergy **E**xascale **E**arth **S**ystem **M**odel
- DOE-SC science mission: Energy & water issues looking out 40 years
- Ensure E3SM will run well on upcoming DOE exascale computers
- https://github.com/E3SM-Project

- **ASCR/BER SciDAC**
- ~10 FTEs over multiple projects
- Large focus on new algorithms

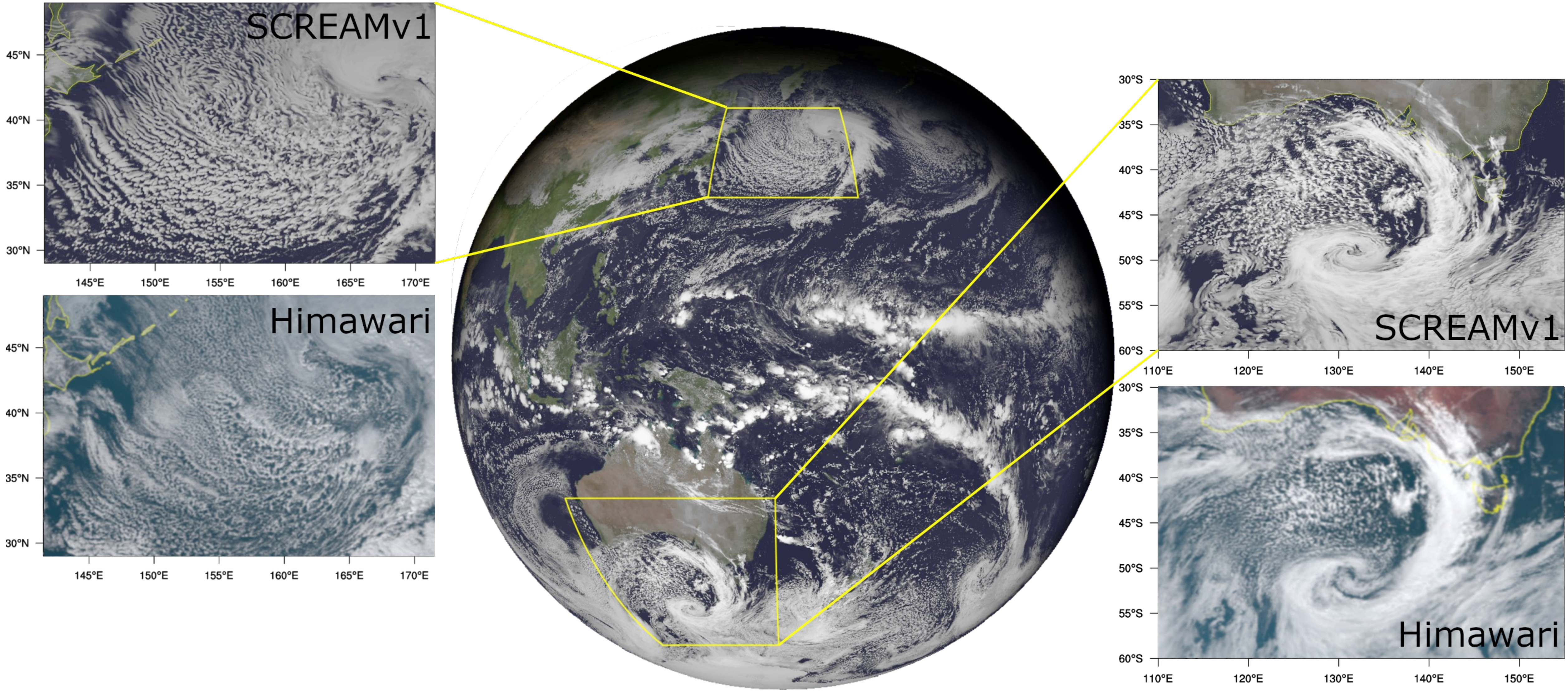
- **ASCR ECP Project**
- ~10 FTEs
- E3SM-MMF: ”superparameterization”



Cloud Resolving Atmosphere Model

- Cloud-resolving simulations (with 3 km) avoid the need for convection parameterizations, which are the main source of climate change uncertainty (Sherwood et al., Nature 2014)
- Resolved convection will substantially reduce major systematic errors in precipitation because of its more realistic and explicit treatment of convective storms.
- Improve our ability to assess regional impacts of climate change on the water cycle that directly affect multiple sectors of the US and global economies, especially agriculture and energy production.

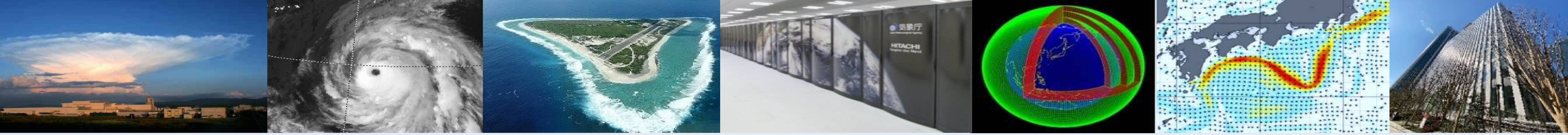
Movie: Precipitation (colors) and integrated water vapor (gray) for an atmospheric river from E3SM's DYAMOND2 simulation. By Paul Ullrich/UC Davis



- Ability to capture cloud structures is impressive
- Example: cold air outbreaks, extra-tropical cyclones are well-represented
- Fig: 2d into a SCREAM DYAMOND simulation (January 22, 2020 at 2:00:00 UTC). Himawari visible satellite image and shortwave cloud radiative effect from SCREAMv0.

Summary

- SCREAM: E3SM atmosphere model rewritten in C++/Kokkos for performance portability
- Competitive with Fortran code on CPUs
- Running well on NVIDIA and AMD GPUs (and hopefully soon Intel GPUs)
- Achieved a longstanding goal of > 1 SYPD at cloud resolving resolutions on Frontier
- 2023-2024: Running some of the first decadal length cloud resolving simulations



HPC readiness: input from JMA

Japan Meteorological Agency

Highlights

- GSM (JMA Global Spectral Model) preparing for future HPCs
  - Improvement of grid decomposition (as reported in WGNE 37)
  - Flexible array structure suitable for both CPU and GPU
  - Reduced precision in MPI communication and its evaluation
- GPU porting of ASUCA (JMA regional NWP model)
- MRI.COM (MRI/JMA ocean model) preparing for future HPCs
  - Single precision in SOM advection
  - GPU porting

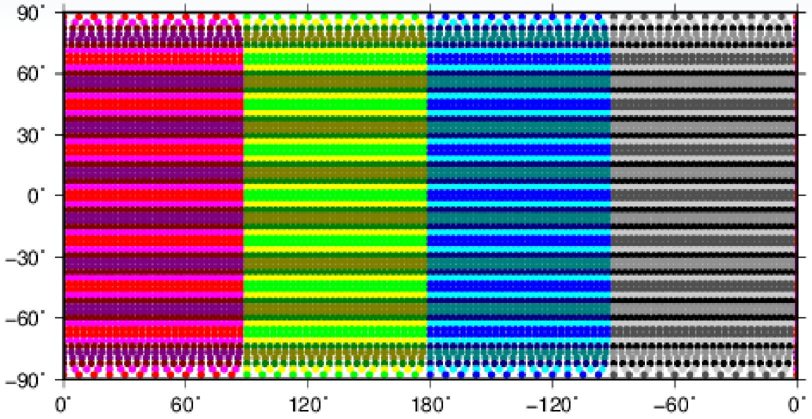
Array Structure of each model
GSM: (i, k, j) ordering
ASUCA: (k, i, j) ordering
MRI.COM: (i, j, k) ordering

(i, j, k) : (x, y, z) directions

Improvement of grid decomposition : a Tl63 16MPI case

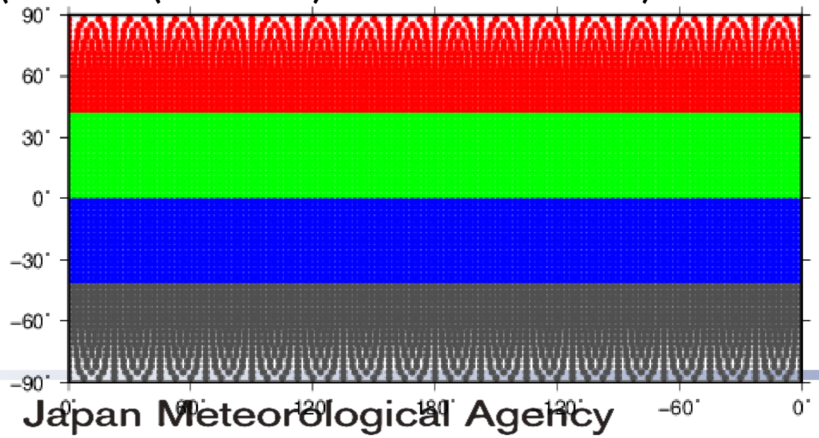
Current decomposition (two grid stages)

For parameterization, I/O and spectral transform

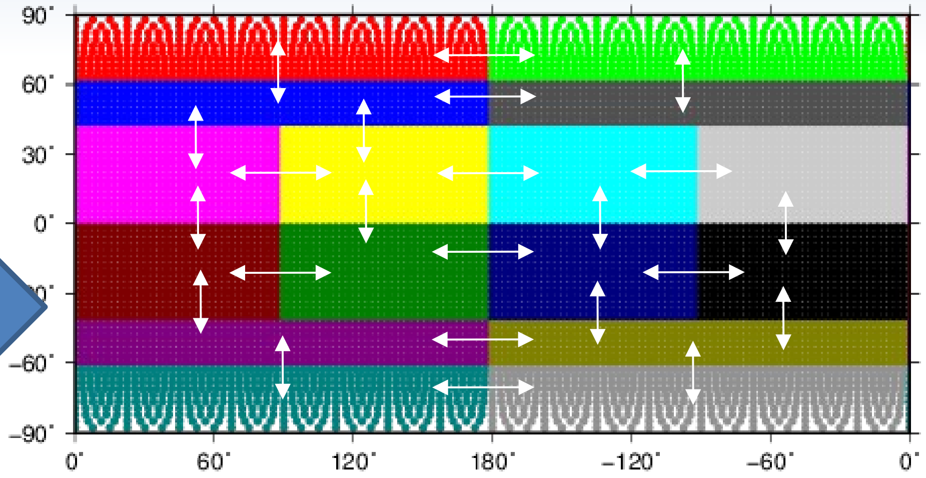


Transpose (with all-to-all MPI communication) required every time step

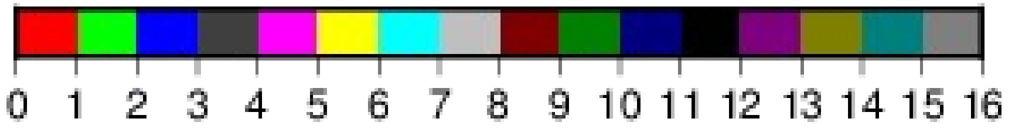
For SL advection
(first to (Kmax/4)th vertical levels)



New decomposition (unified grid stage)

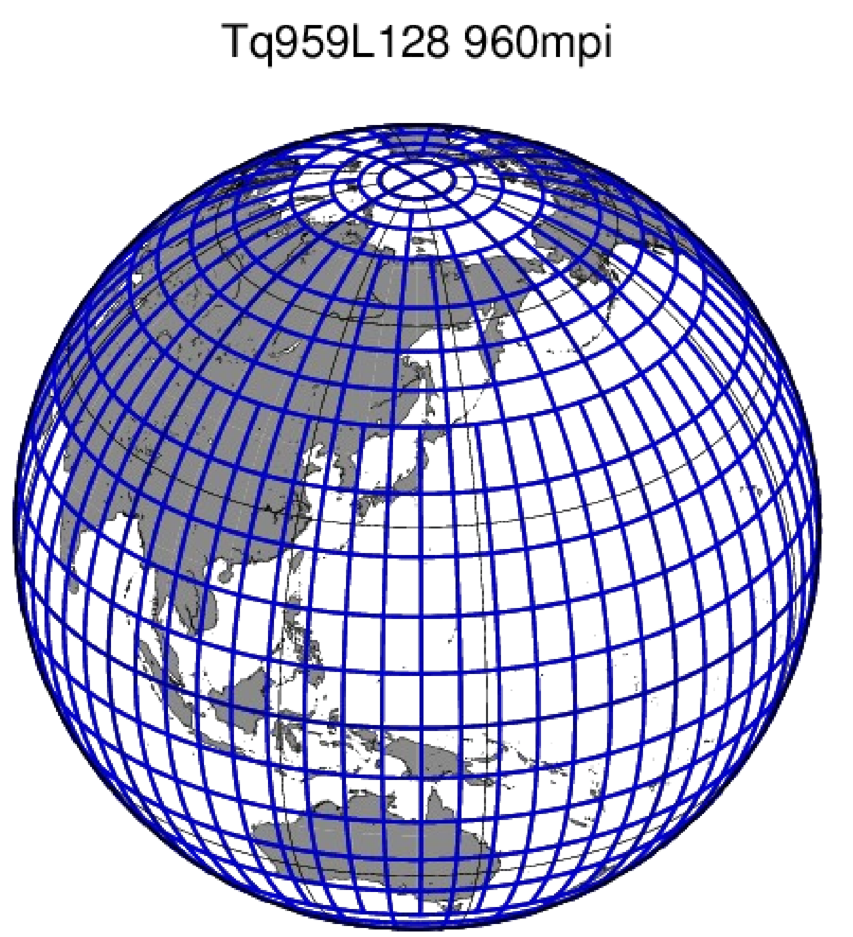


Only halo communication for SL advection



MPI rank

Improvement of grid decomposition: Tq959 960MPI case



- Pros.
  - Suitable for both derivative stencils in gridspace and spectral transform.
    - also preparation for grid-spectral hybrid approach
  - Only halo communication (no all-to-all communication) required for Semi-Lagrangian advection
- Cons.
  - Load-balance of computation in physics parameterizations gets worse (however, pros. overweigh in high-resolution runs)

Single precision in MPI communication for GSM

- GSM: A semi-implicit semi-Lagrangian global model with a spectral method
  - Its computational performance strongly depends on MPI communication.
  - Single-precision rather than double-precision in MPI comm. is an effective way to reduce amount of the communication, hence, to improve the computational performance
    - Less risks than full single precision calculation
    - Also a step for the single precision GSM
- Impacts of single precision only in MPI communication on computational costs and forecasts are tested preliminarily.

GPU porting of ASUCA

- ASUCA: JMA's operational regional model (Ishida et al. 2022)
- Code characteristics:
  - (k,i,j) ordering (Array(nz, nx, ny))
  - MPI-OpenMP hybrid parallelization
  - Subroutines are called in a horizontal loop
- Dynamics
  - MPI comm. necessary
  - Vertically dependent loops for a tri-diagonal matrix solver
- Physics
  - Parallel in the horizontal
  - Strong loop carried dependence in vertical

```
subroutine calculate
real(8): x(nz,nx,ny),y(nz,nx,ny)
real(8) :: s(nz)

!$OMP  PARALLEL DO &
!$OMP& PRIVATE(s,….)
do j = 1, ny
do i = 1, nx
  …
  call cal_main( x(1,i,j), y(1,i,j), s(1),….)
```

HPC readiness of MRI.COM

- MRI.COM : MRI Community Ocean Model (Sakamoto, et al. 2023)
  - A depth coordinate model that solves the primitive equations under hydrostatic and Boussinesq approximations
  - Used in ocean monitoring/forecasts, climate prediction, and research on coupled NWP as a feasibility study
  - (i,j,k) ordering array structure
- Adopt to future HPCs / speed up required
  - Time-to-solution is critical as an operational model
  - As a climate model, long term (>100 yrs) time integration for spin-up required
- Recent research topics of MRI.COM in the context of HPC readiness
  - Reduced precision in the SOM advection scheme
  - GPU porting

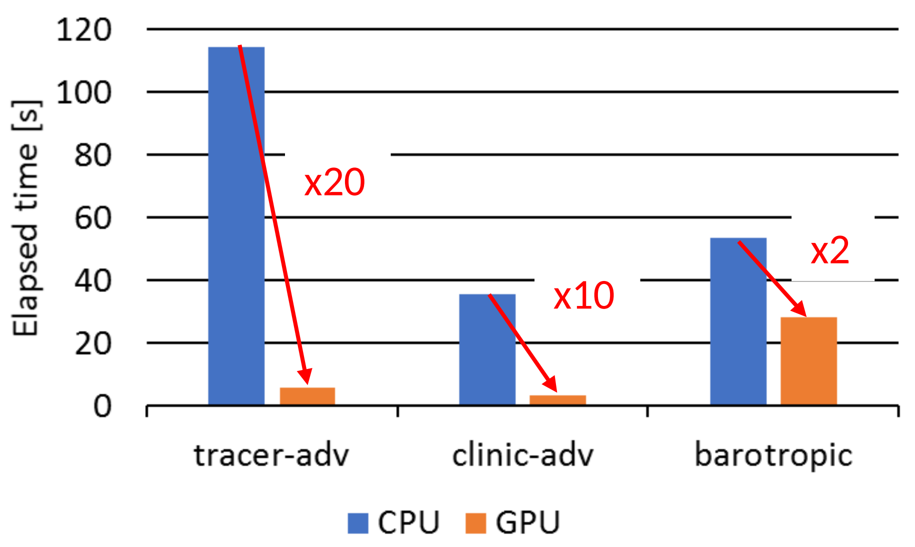
GPU porting of MRI.COM

- GPU porting of MRI.COM was tested in a rectangular domain.
  - Horizontal grids : 242 x 202 ~ $O(10^4)$, Vertical layers : 40
    - Mimicked problem size per node for ocean part of JMA/MRI Coupled Prediction System
  - Focus on only calculation parts
    - Optimizing CPU <-> GPU data transfer and MPI parallelization will be next steps.
- High-cost processes are ported to GPU.
  - Dynamical process (particularly advection) for tracer variables
  - Momentum equations (barotropic and baroclinic components)
  - OpenACC directives are inserted in most loops of these processes.
  - Some loops are modified for 3-dimensional parallelization.
    - The original CPU-based codes consist of 2-dimensional parallelization with OpenMP.

GPU porting of MRI.COM

- CPU: Intel Xeon Gold 6226 2.7GHz 12C/24T x2 with DDR4 memory (140GB/s)
- GPU: NVIDIA Tesla V100-PCIe-32GB x1 with HBM2 memory (900GB/s)
- Run 10-day (960-timestep) forecast to check the performance.

Acceleration of bottleneck processes (calculation parts only)



tracer-adv	Advection of tracers with QUICK scheme (3-dimensional)
clinic-adv	Advection of baroclinic components (3-dimensional)
barotropic	Time integration of barotoropic components (2-dimensional)

- All of these processes are accelerated by GPU.
  - In particular, the processes parallelized in 3-dimensional are accelerated remarkably.
- Optimizing CPU <-> GPU data transfer is ongoing. (e.g. reducing amount / frequency)

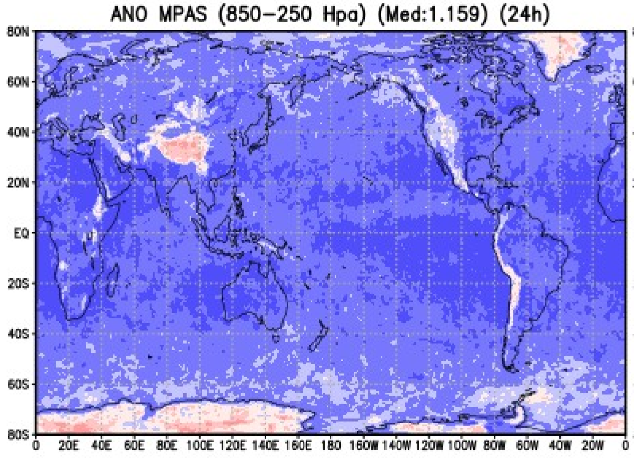
References

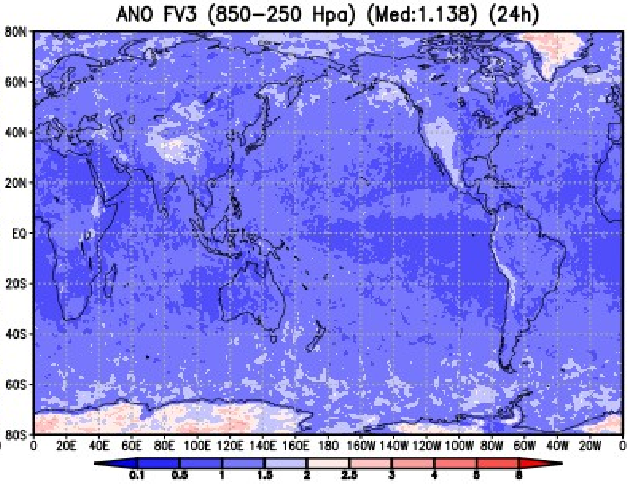
- Ishida, J., K. Aranami, K. Kawano, K. Matsubayashi, Y. Kitamura, and C. Muroi, 2022: ASUCA: The JMA Operational Non-hydrostatic Model. J. Meteor. Soc. Japan, 100, 825-846.
- Sakamoto, K., H. Nakano, S. Urakawa, T. Toyoda Y. Kawakami, H. Tsujino, and Goro Yamanaka, 2023: Reference Manual for the Meteorological Research Institute Community Ocean Model version 5 (MRI.COMv5), TECHNICAL REPORTS OF THE METEOROLOGICAL RESEARCH INSTITUTE No.87.

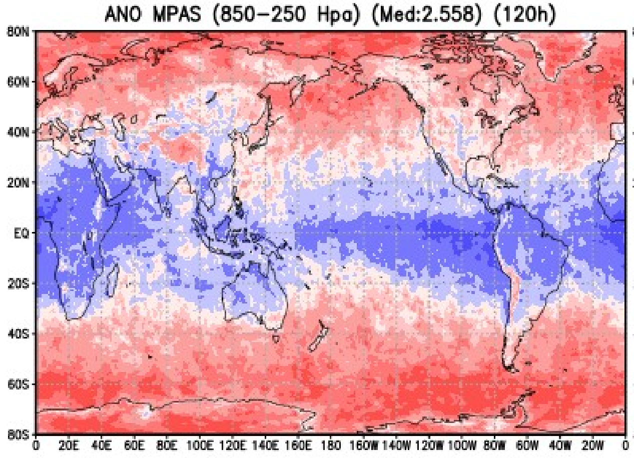
Mahalanobis distance - Globe

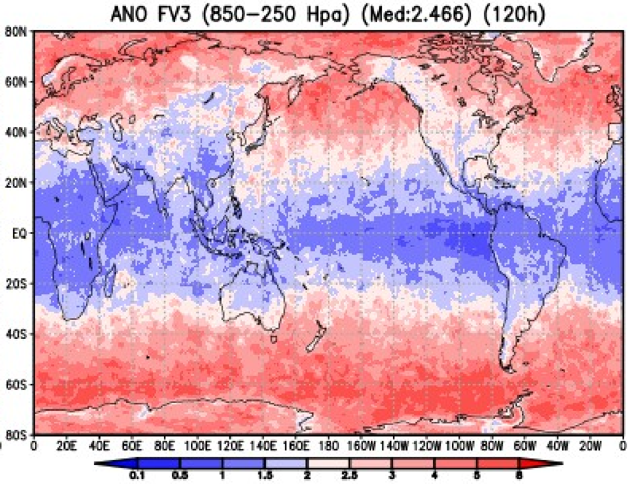
24h

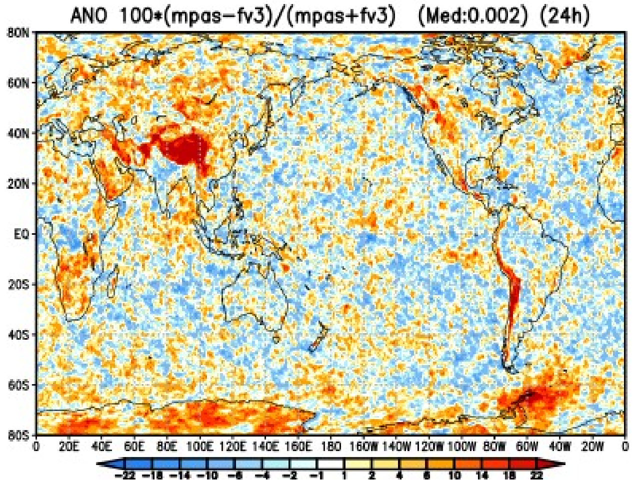
120h

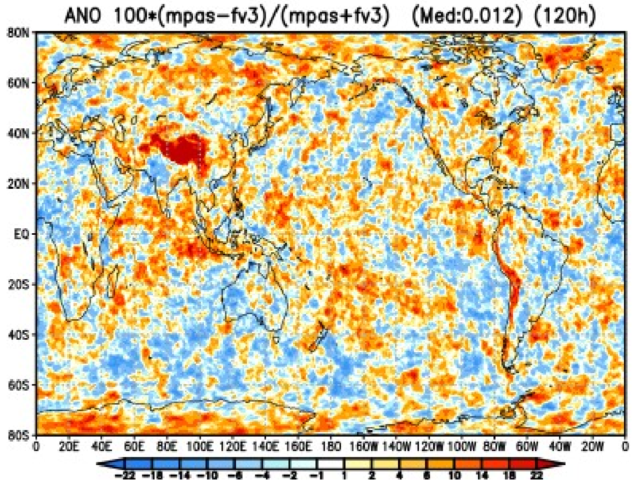






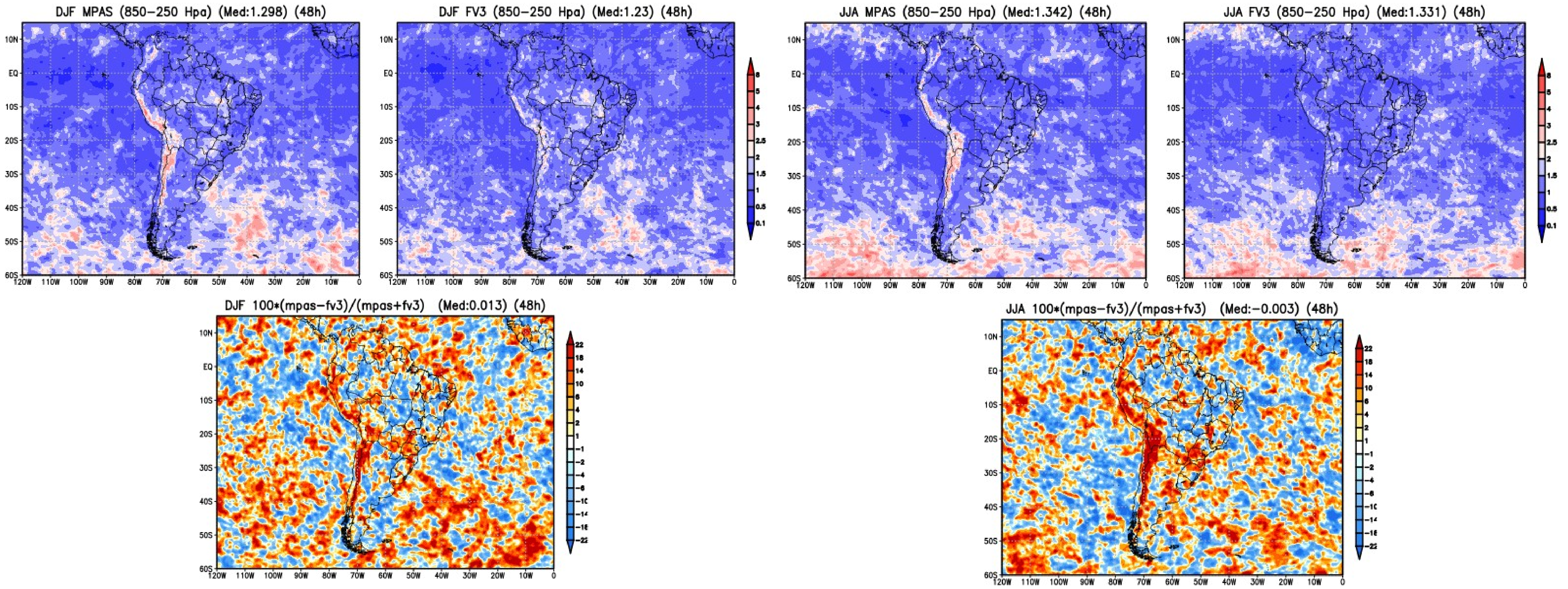






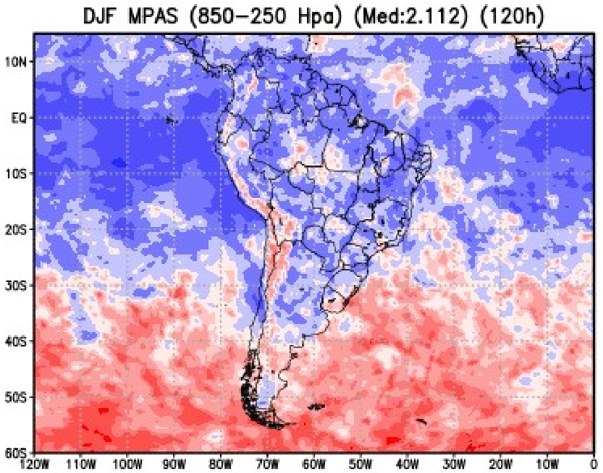
Mahalanobis distance - South America and oceans

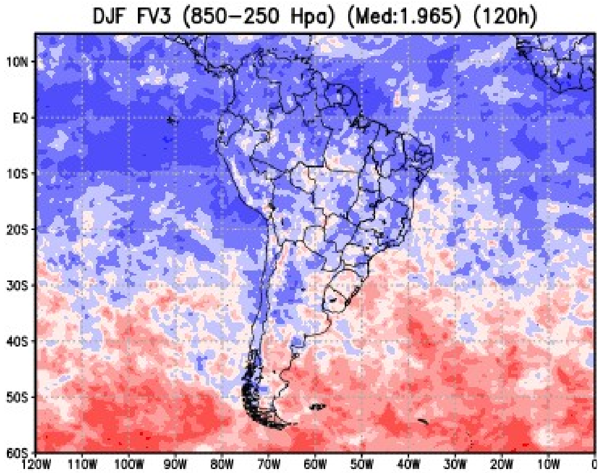
48h forecast length

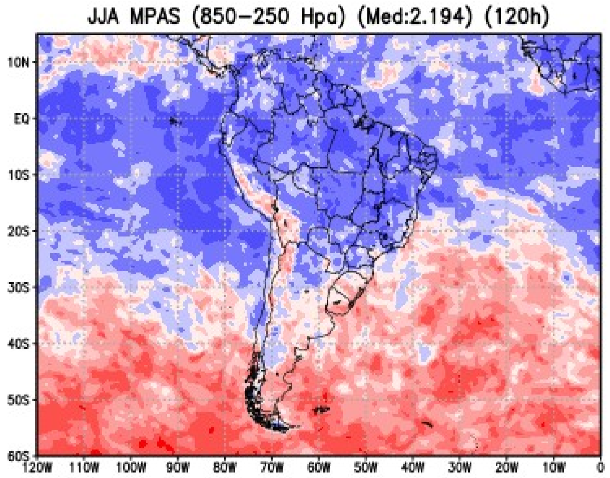


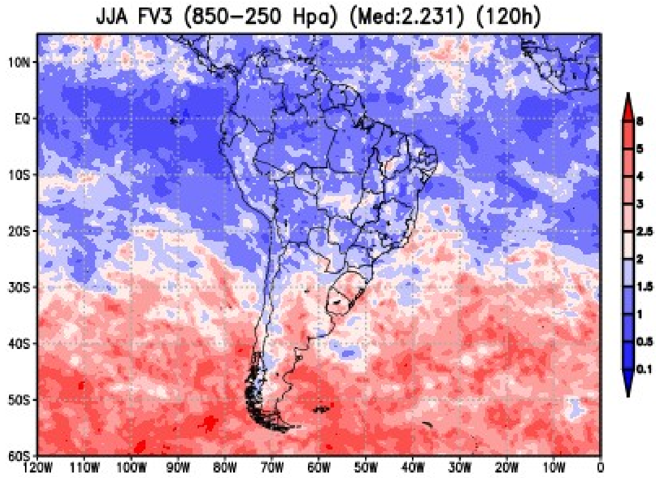
Mahalanobis distance - South America and oceans

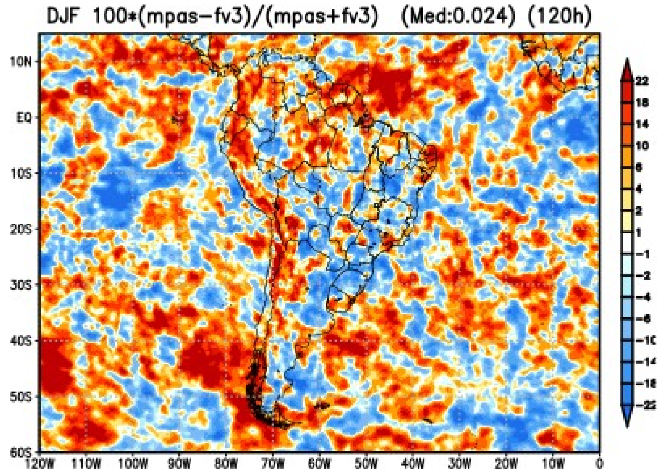
120h forecast length

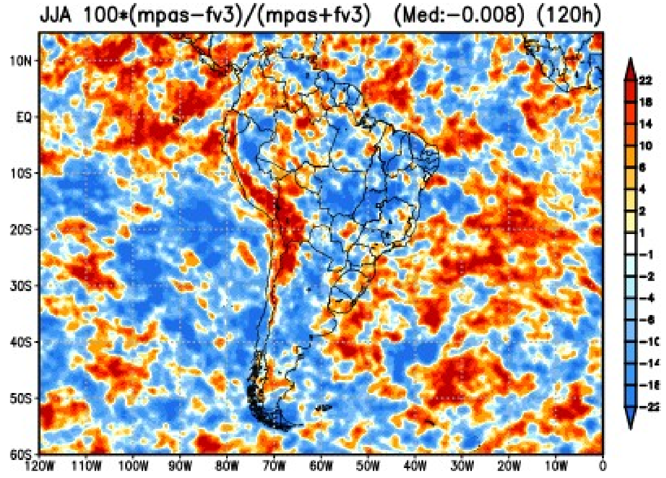












Precipi mean intensity computed over the global domain, in mm/day

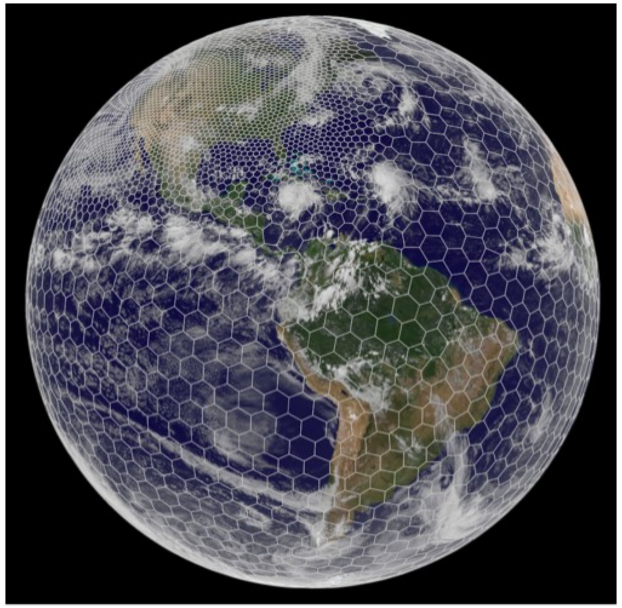
Time integration	IMERG	MPAS	ShiELD	Diff MPAS	Diff ShiELD	Diff perc. MPAS (%)	Diff perc. ShiELD (%)
36h	3,19928	3,09617	3,47084	-0,10311	0,27156	-3,223	8,488
60h	3,20317	3,17713	3,52475	-0,02604	0,32158	-0,813	10,039
84h	3,19727	3,23611	3,61322	0,03884	0,41595	1,215	13,010
108h	3,17839	3,27442	3,65614	0,09603	0,47775	3,021	15,031
132h	3,20712	3,31056	3,68918	0,10344	0,48206	3,225	15,031
156h	3,20342	3,34213	3,72282	0,13871	0,5194	4,330	16,214
180h	3,20599	3,35401	3,75819	0,14802	0,5522	4,617	17,224
204h	3,20136	3,37975	3,77914	0,17839	0,57778	5,572	18,048
228h	3,18150	3,38142	3,81426	0,19992	0,63276	6,284	19,889

Model for Ocean-laNd-Atmosphere predictioN

MONAN's dynamical core

MPAS

Model for Prediction Across Scales



Thank you

WORLD
METEOROLOGICAL
ORGANIZATION

150
1873
2023
IMO-WMO

wmo.int

WCRP
World Climate Research Programme