EarthWorks: Towards an Earth System Model at Storm Resolving Resolutions

RICHARD LOFT
(former) Director Technology Development
Computational and Information Laboratory
National Center for Atmospheric Research

NCAR

September 22, 2021

NSF

This material is based upon work supported by the National Center for Atmospheric Research, which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977.

Outline

- Refactoring the Model for Prediction Across Scales (Atm) for CPU&GPUs
- EarthWorks: Toward a CPU&GPU portable Earth System Model
- Handling the Big Data Problem
- Machine Learning: The Silver Bullet?
- Three Cs needed to pull this off

Initial MPAS-A GPU Project Goals (2017)

- **Performance portability**
  - Achieve best performance on GPU while *maintaining CPU performance*
  - Maintain readability/intelligibility of the source code.
- **Resilience with respect to architectural details, such as**
  - Number of GPUs/node
  - Number of CPU cores/node
- **Offload minimum amount of code to GPUs**
  - Some code might not be suitable
  - Limited budget/staff resources!

Methodology for Refactoring Legacy Models

- Use OpenACC standard directives to offload work to the GPUs.

```
!$acc parallel
!$acc loop seq
do k=1,nz
!$acc loop collapse(2)
do J=Jsq,Jeq ; do i=is,ie
BT_force_v(i,J) = BT_force_v(i,J) + wt_v(i,J,k) * bc_accel_v(i,J,k)
enddo ; enddo ; enddo
!$acc end parallel
```

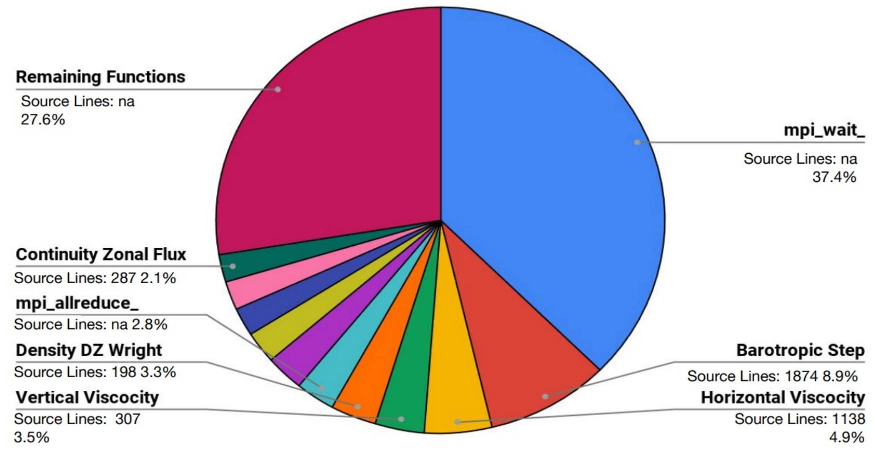
- Use test driven development.

```
VALIDATION RESULTS...

Density        : 1.0241467e-10   PASS
Temperature    : 1.0215635e-10   PASS
Velocity       : 3.2897487e-09   PASS
Energy         : 7.567654e-11    PASS
```

- Use profiling to prioritize refactoring and optimization targets

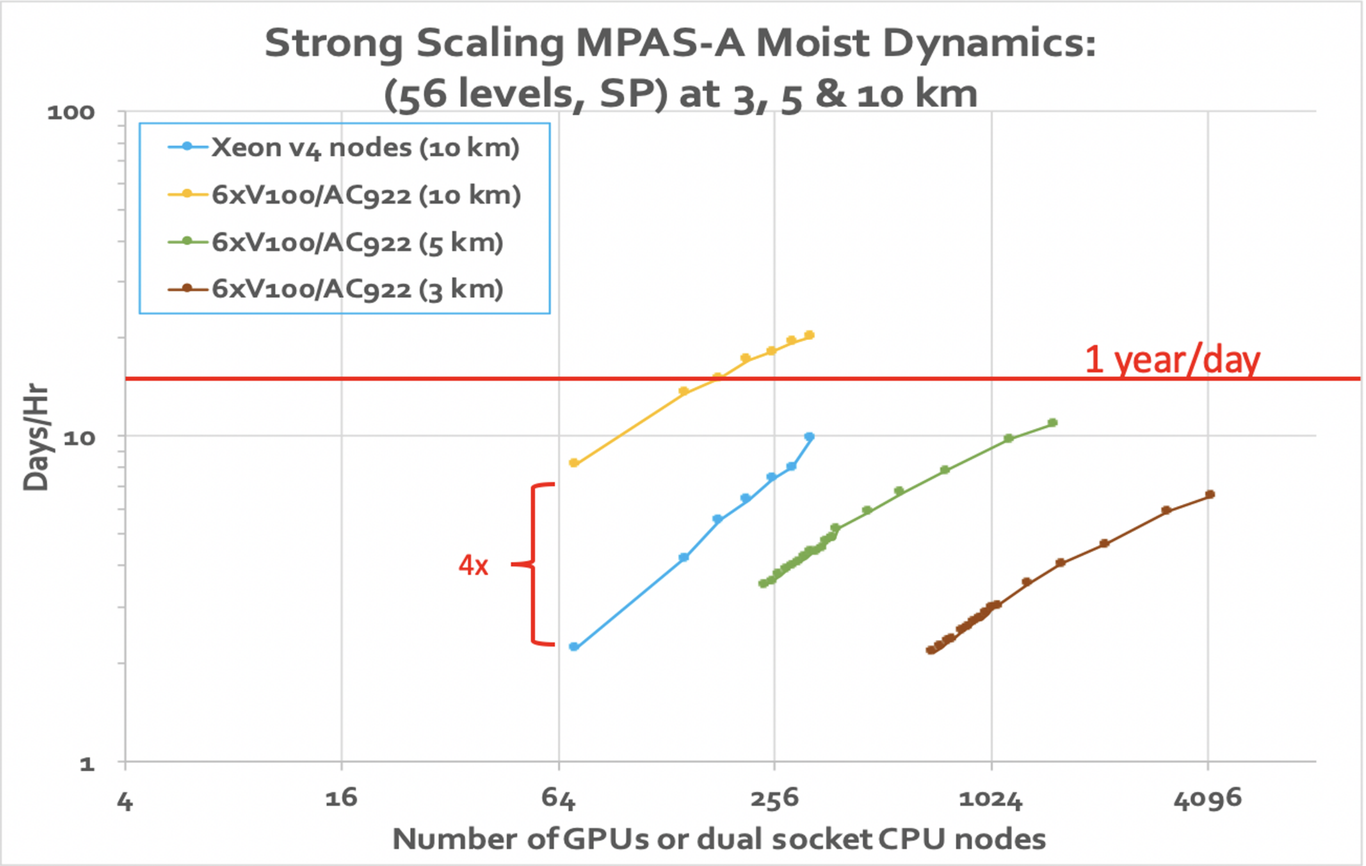
Percent CPU Time Spent on Each Function



MPAS dynamical core: Weak scaling to global 3 km resolution on Summit

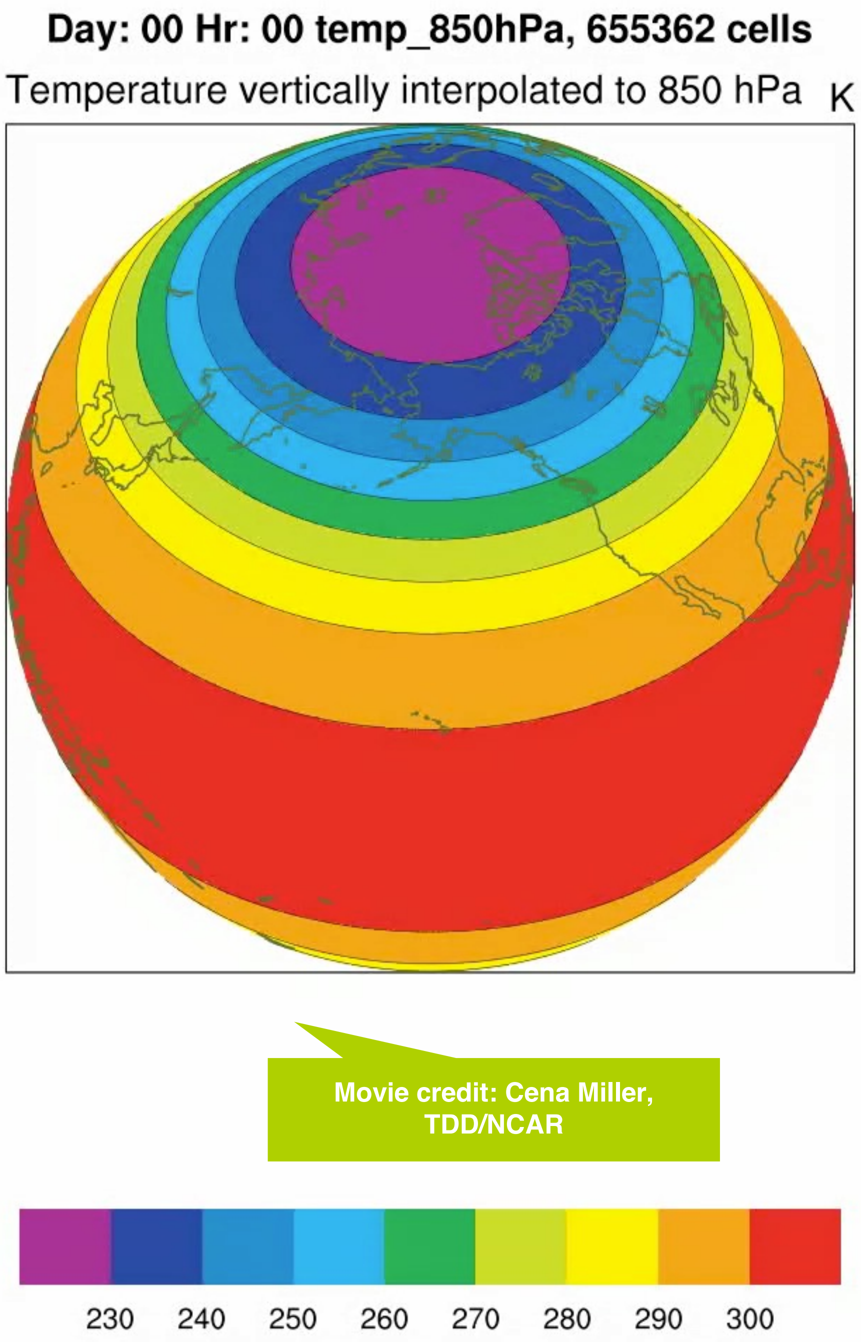
3 Years ago

MPAS-A Dynamics on Summit[1] vs Cheyenne[2]



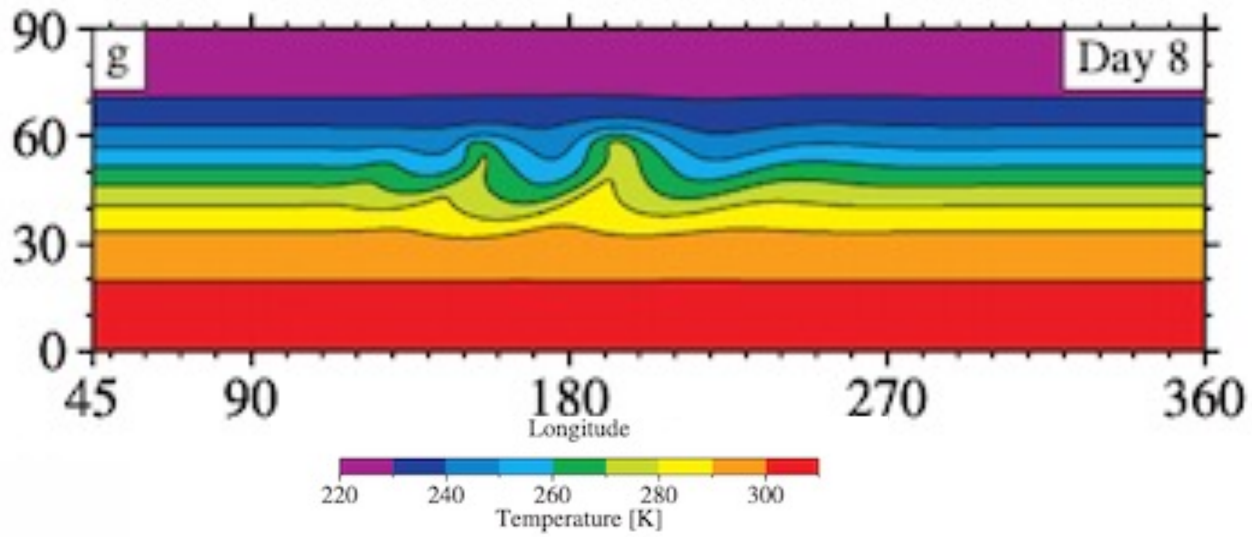
[1]Benchmarking on Summit supported by DoE via an OLCF Director's Discretionary Allocation
[2]Cheyenne is a 5.4 PF, 4032-node HPE system with EDR interconnect operated by NCAR

Baroclinic Instability Test (Jablonowski, et al. 2006)

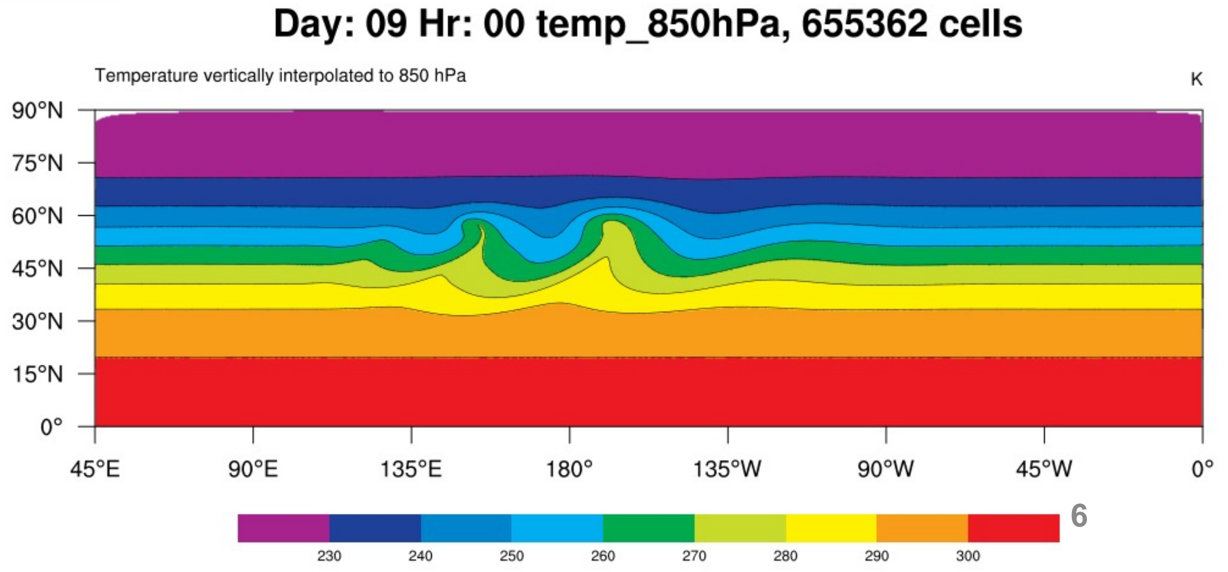


Movie credit: Cena Miller, TDD/NCAR

REF

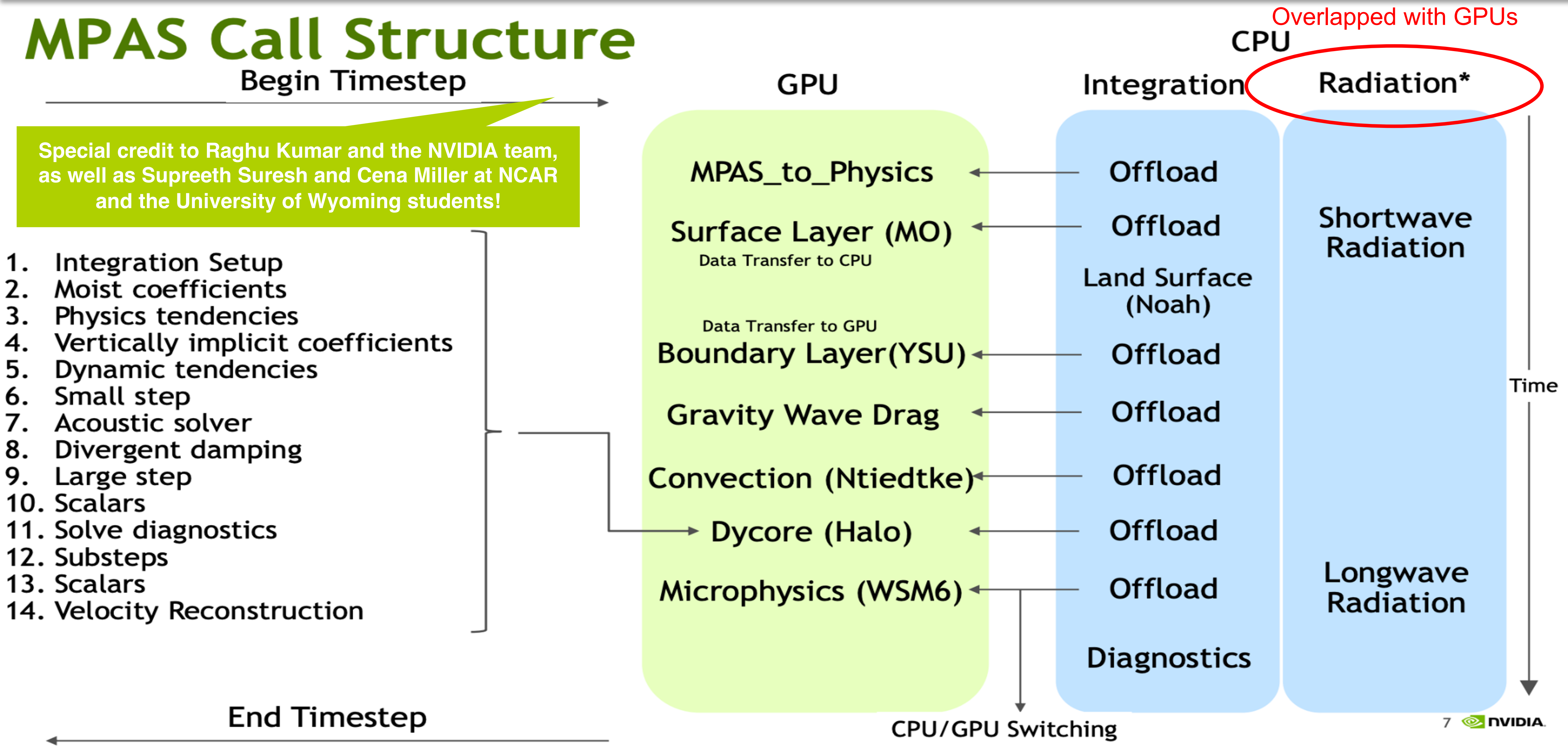


MPAS



On to MPAS-A with full physics (with lagged radiation)

Special credit to Raghu Kumar and the NVIDIA team, as well as Supreeth Suresh and Cena Miller at NCAR and the University of Wyoming students!



Refactoring MPAS-A for GPUs… mission accomplished!

Resolution matters: MPAS-A OpenACC is making running global storm resolving atmospheric models more feasible

15 km

3 km

Movie credit: Todd Hutchinson, The Weather Company

MPAS-A OpenACC Accomplishments

- In production since October 2019 as part of the IBM-GRAF forecast system.
- OpenACC version available to the community via GitHub.

MPAS-OpenACC It is the result of a partnership between NCAR, NVIDIA and IBM / The Weather Company

Outline

- Refactoring the Model for Prediction Across Scales (Atm) for CPU&GPUs
- EarthWorks: Toward a CPU&GPU portable Earth System Model
- Handling the Big Data Problem
- Machine Learning: The Silver Bullet?
- Three Cs needed to pull this off

GSRM Driver: Cumulus Convection

"…the lack of capacity to simulate [cloud formation and cumulus convection] in fine detail ***accounts for the most significant uncertainties in future climate****, especially at regional and local levels."*

COP26 Policy Briefing 1 (CLIMATE CHANGE : SCIENCE AND SOLUTIONS NEXT GENERATION CLIMATE MODELS), p. 2.

"The more explicit rendering of clouds, precipitation, and other fine-scale processes ***will also facilitate the use of high-resolution observations …*** *to assess the model's performance."*

Page 23, NCAR Strategic Plan 2020-24

"While precipitation biases varied geographically and seasonally, ***1-km model climatologies of precipitation generally aligned better with those observed than 3-km climatologies.*** *…."*

Schwartz and Sobash (2019)

Depiction of the visible cloud field from an global MPAS-Atmosphere simulation using a horizontal mesh with 3.75 km cell spacing (Falko Judt, NCAR).

EarthWorks

Five-year project led by CSU, with participation by 3 NCAR labs. Funded by NSF CSSI.

Science Goals

- Begin to resolve storms at ~4-km grid.
- Eliminate deep convection or gravity-wave drag parameterizations.
- Include a resolved stratosphere.
- Enable new science (extreme events!) for both weather and climate.
- Provide a critical capability to the climate community for guiding adaptation at global, regional and local levels.

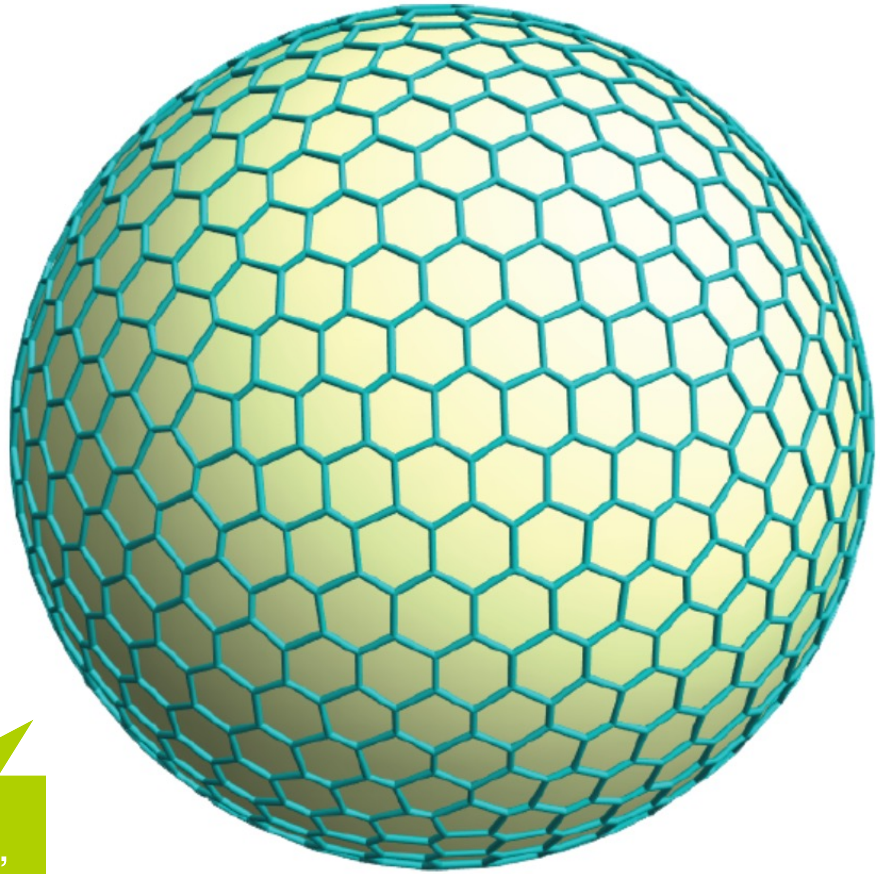
Model architecture

- A global coupled Earth System configuration based on CESM infrastructure including CESM CMEPS Coupler & the Community Physics Framework;
- Using MPAS-Atmosphere & MPAS-Ocean components.

Computational Goals

- The EarthWorks ESM will run on CPUs.
- Fully GPU-enabled implementation of ocean and atmosphere for tackling high resolutions.
- Reach ~0.5 SYPD at ~3.75 km for the coupled system using GPU acceleration by 2025.
- EarthWorks will put huge demands on computational and data systems. Thus the project incorporates infrastructure development efforts for both big data and machine learning inference.

EarthWorks: one quasi-uniform mesh for all components



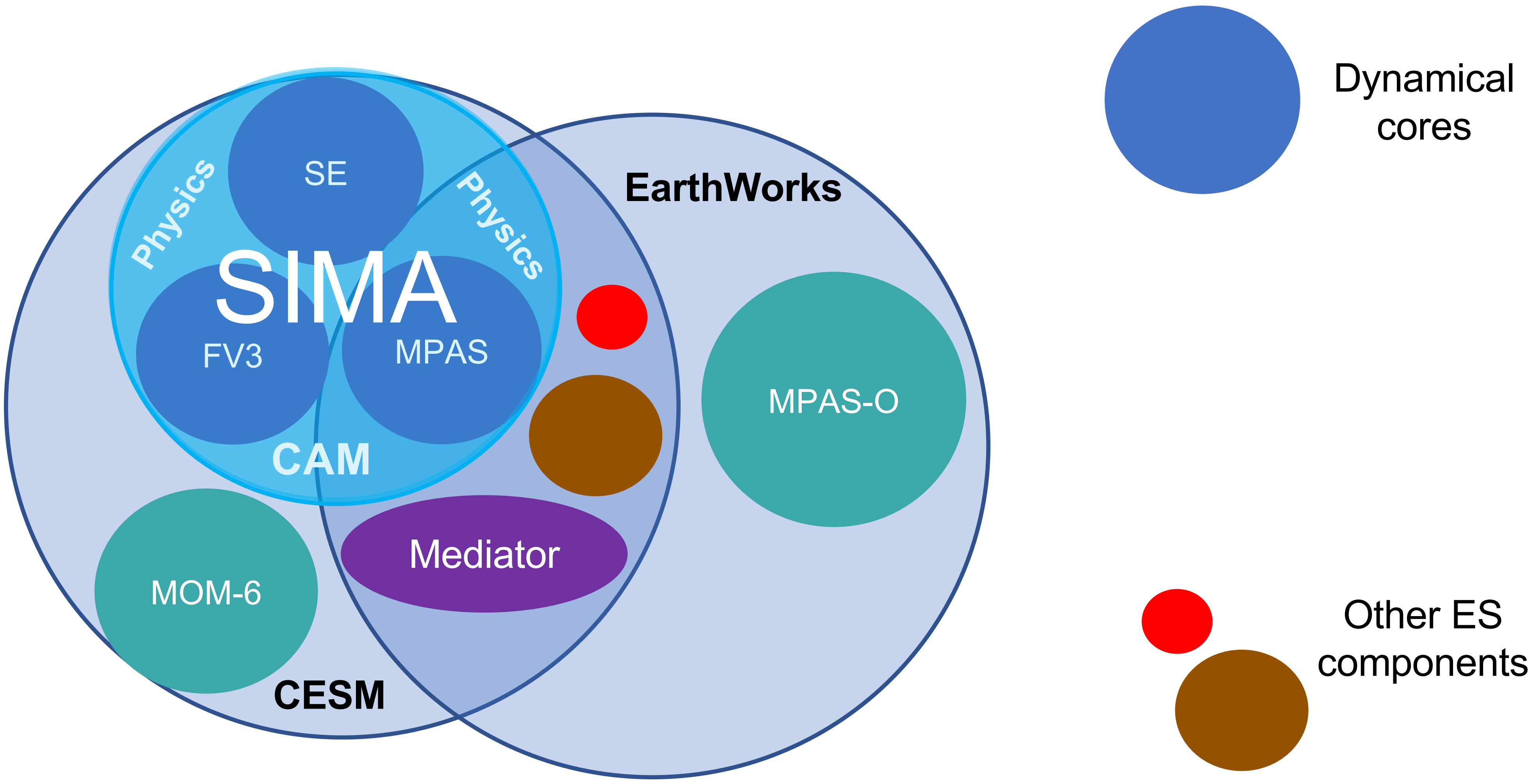
Grid	No. of grid points N	Avg grid distance ℓ (km)
G0	12	6699.1
G1	42	3709.8
G2	162	1908.8
G3	642	961.4
G4	2562	481.6
G5	10 242	240.9
G6	40 962	120.4
G7	163 842	60.2
G8	655 362	30.1
G9	2 621 442	15.0
G10	10 485 762	7.53
G11	41 943 042	3.76
G12	167 772 162	1.88
G13	671 088 642	0.94

Non-hydrostatic regime

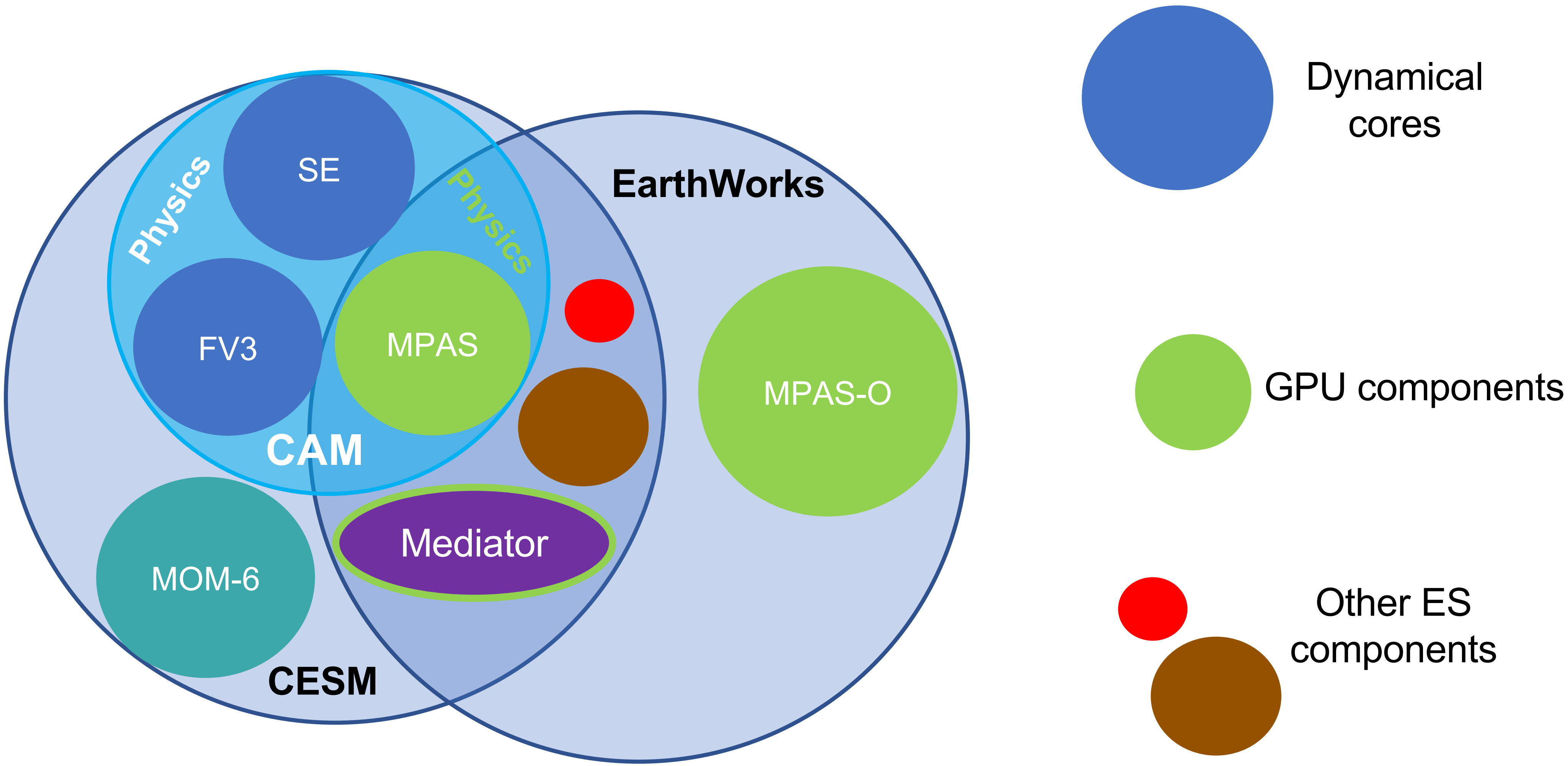
Target grid spacing

Slide credit: Dave Randall, Colorado State Univ.

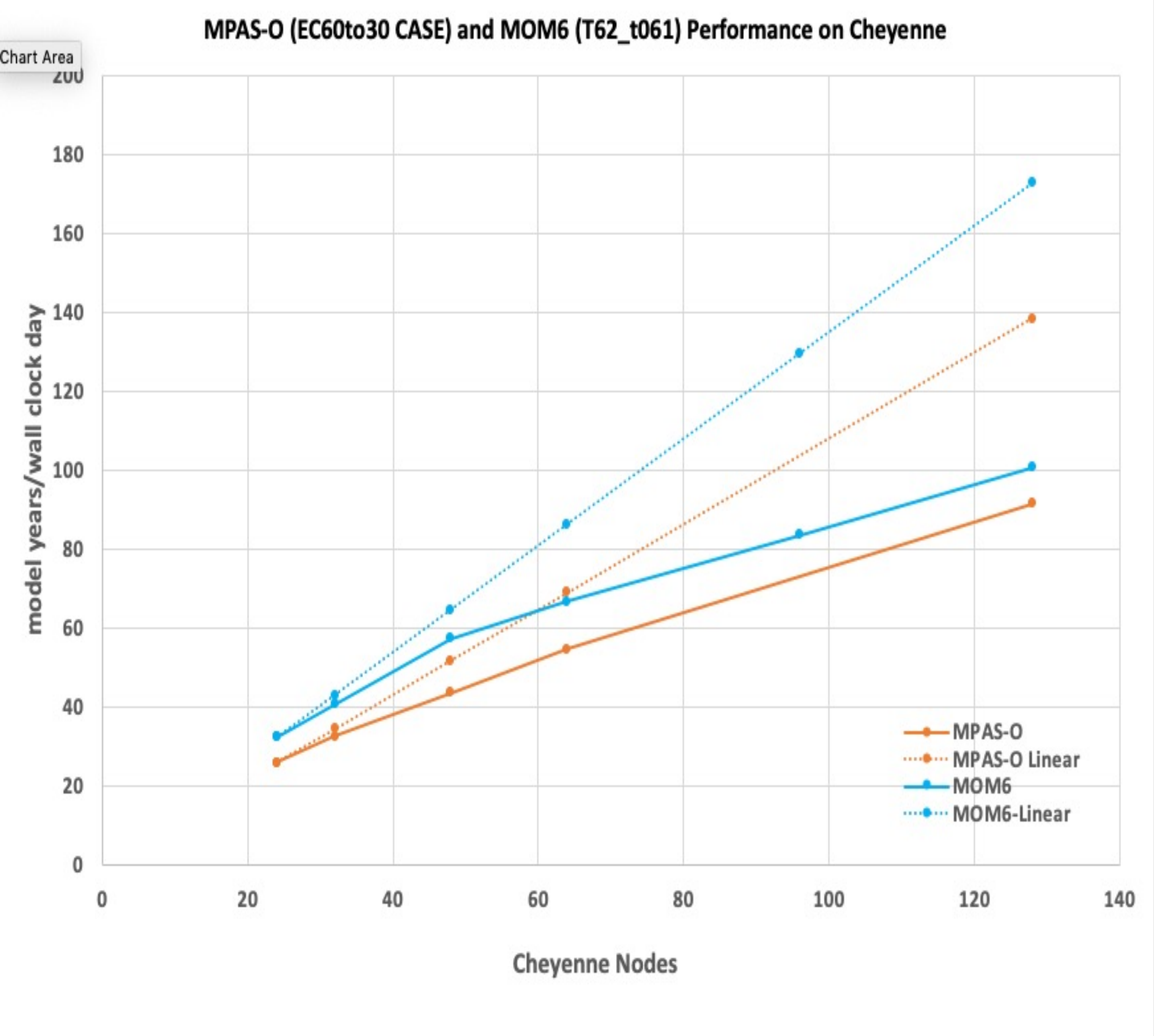
How EarthWorks relates to other NCAR-based ESM efforts

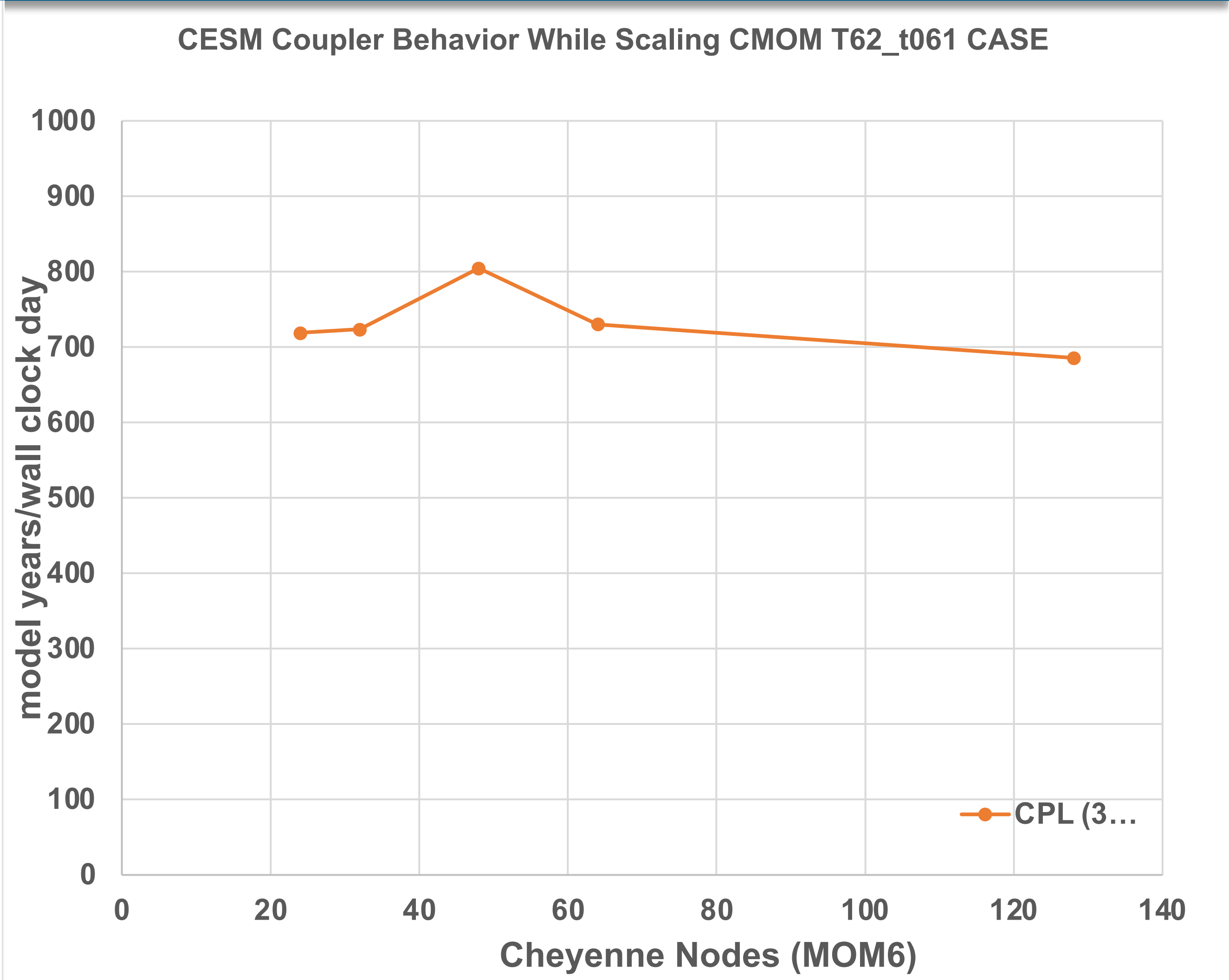


EarthWorks with (on-GPU = Green)



CPU-based MPAS-O V6 vs MOM6 Performance Comparison





MPAS-O GPU/CPU Standalone Benchmarks

- **Hardware Configuration (NVIDIA clusters)**
  - **Prometheus:** Dual socket, 40 c Broadwell node, 8 V100 GPUs/node
  - **Selene:** Dual AMD EPYC 7742 64-Core Processor, 8 A100 GPUs/ node
- **Software Configuration**
  - MPAS-O Version 6
  - PGI Compiler 21.5, OpenMPI 4.1.1, UCX 1.10.0-rc2
  - Dependencies:
    - Pnetcdf 1.12.2, PIO 2.5.4
- **MPAS-O Configuration**
  - Precision: Double
  - Horizontal: 60 to 30 km "EC60to30" variable resolution test case
    - 235K grid points
  - Vertical levels: 60
- **Timestep**
  - Contains 13 Halo exchanges per timestep
    - 7 inside if conditions, 1 inside a loop
  - 300+ Parallel code regions, 100+ data directives

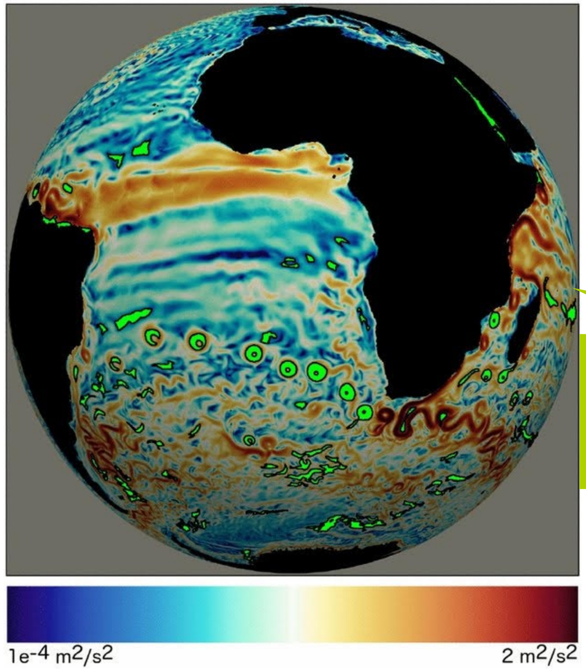
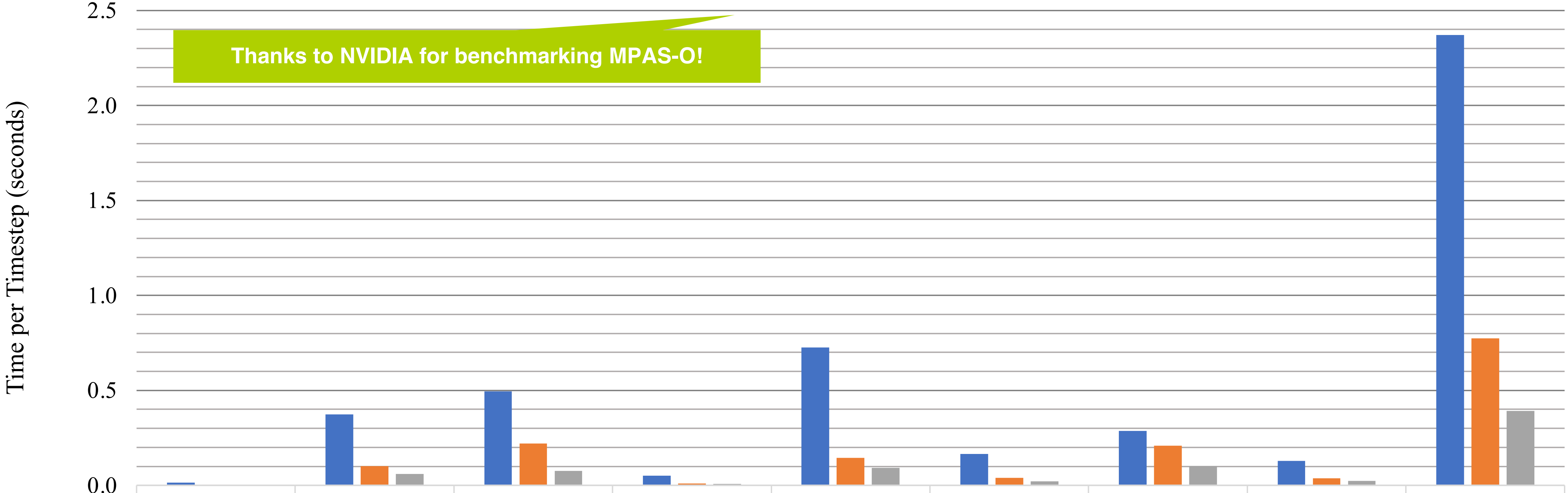


Image from: Banesh, D., Petersen, M., Wendelberger, J. *et al. Environ Earth Sci* 78, 623 (2019).

MPAS-O Timestep
- **Initialization**
- **Iterations**
  - **Baroclinic Velocity Prediction**
  - **Barotropic Velocity Prediction**
  - **Thickness Tendencies**
  - **Tracer Tendencies**
  - **Variable Reset**
- **Implicit Vertical Mixing**
- **Fini**

MPAS-O GPU Performance Results

Performance comparison between two 40c Broadwell nodes, and two V100 (Prometheus) and two A100 (Selene)



	Initialization	Baroclinic	Barotropic	Thickness	Tracer	Variable Reset	Implicit V-Mix	Fini	Total Time
Broadwell	0.02	0.37	0.49	0.05	0.73	0.17	0.29	0.13	2.37
V100	0.00	0.10	0.22	0.01	0.14	0.04	0.21	0.04	0.78
A100	0.00	0.06	0.08	0.01	0.09	0.02	0.10	0.02	0.39

Broadwell V100 A100

6x faster

1 MPI Rank per GPU, EC60to30 Testcase, 118K Grid Points per GPU, 60 Vertical Levels

What about the Ocean throughput at 3.75 km?

- MPAS-O runs EC60to30 at about 2 SYPD on 2 Broadwell node.
- Two A100s run about 6x faster than that (i.e. ~12 SYPD)
- Weak scale to 3.75 km resolution, keeping ~118K grid pts per GPU
- The Model timestep and SYPD will be 8x smaller (30 km -> 3.75 km).
- Coupler overhead appears negligible.
- Putting this together, if MPAS-O weak scales, the integration rate is predicted be ~1.5 SYPD.
- So we're in the ballpark of load balancing the asynchronous execution of MPAS Atmosphere and MPAS Ocean at ~1 year per day.

Status of OpenACC Version of MPAS-O[1]

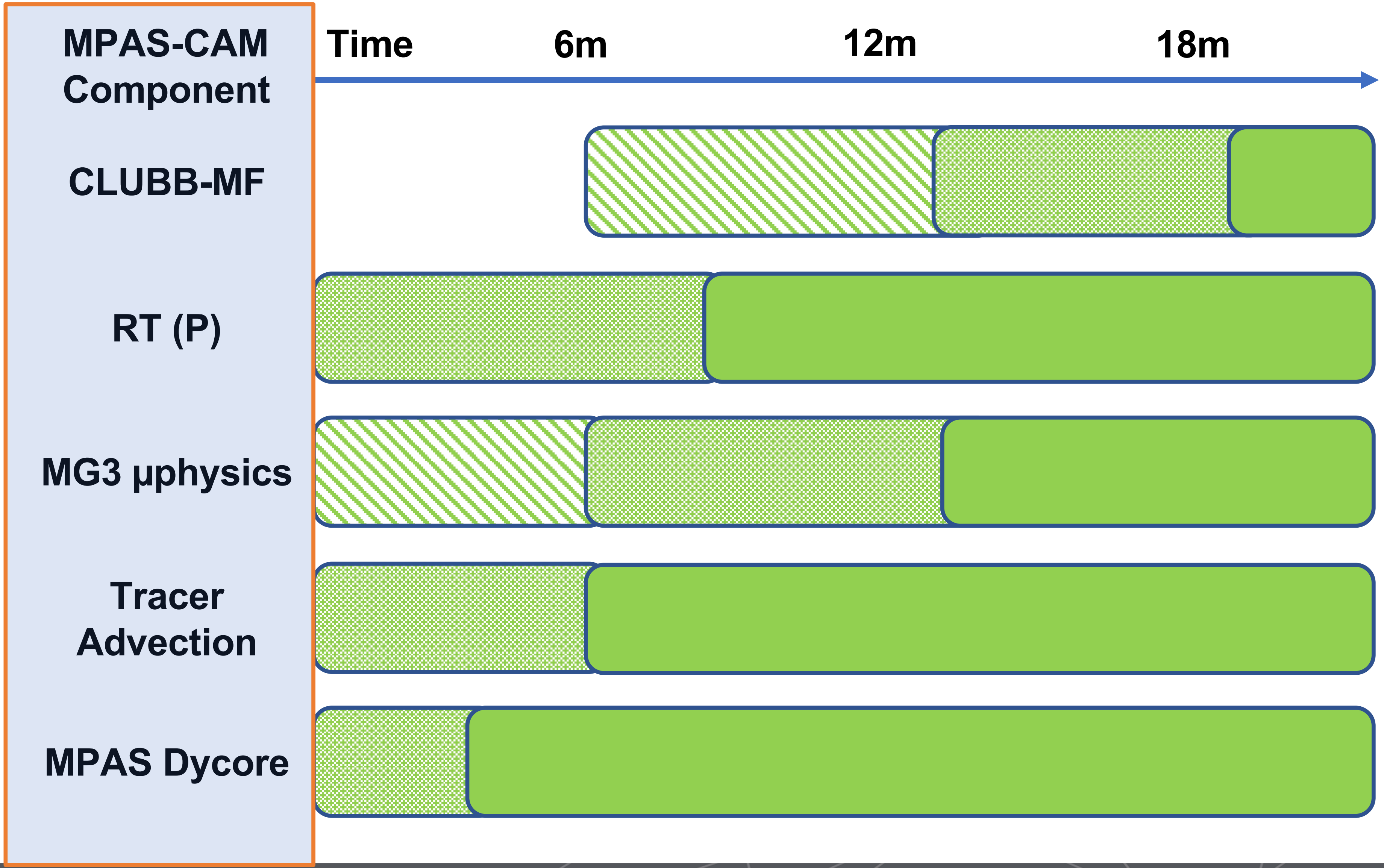
- Baroclinic velocity
- Barotropic solver (MPAS-O V6)
- Semi-implicit barotropic solver formulation (available this fall with MPAS-O V7[2])
- Thickness tendency
- Diagnostic solver
- Vertical Mixing, a.k.a. CVMix (loop reordered)
- Secondary diagnostics
- SE Loop Fini
- SE Loop
- Implicit vertical mixing

[2]Personal communications from Phil Jones at LANL and Raghu Raj Kumar at NVIDIA Corporation

[1]MPAS-O must still demonstrate weak scaling to deliver required throughput.

EarthWorks MPAS-CAM GPU Refactoring Timeline

*T=0 corresponds to creation MPAS-CAM CESM config (~7/21)



Porting

Optimize

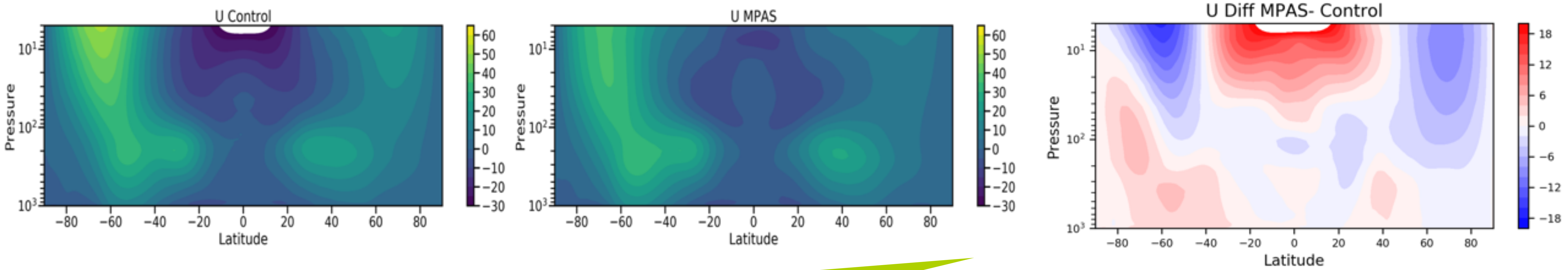
Complete

First Step: Validation of MPAS-CAM on CPU… is underway

- **Control model is a CMIP6 configuration:**
  - CESM2 CAM6
  - Finite Volume (FV) dynamical core
- **MPAS test configuration:**
  - Same physics model (CAM6)
  - MPAS non-hydrostatic dynamical core
  - 100 km resolution
  - Same vertical resolution, but using height, not pressure vertical coordinates
- **Simulations**
  - Present day climatological simulations
  - Duration: 5 years

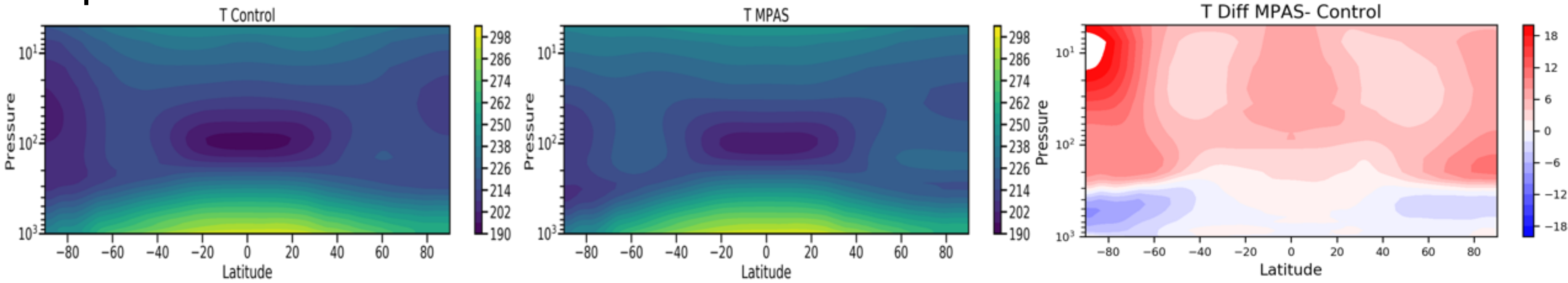
Still Some Differences in Upper Atmosphere Dynamics – Need to sort out damping

Zonal Wind

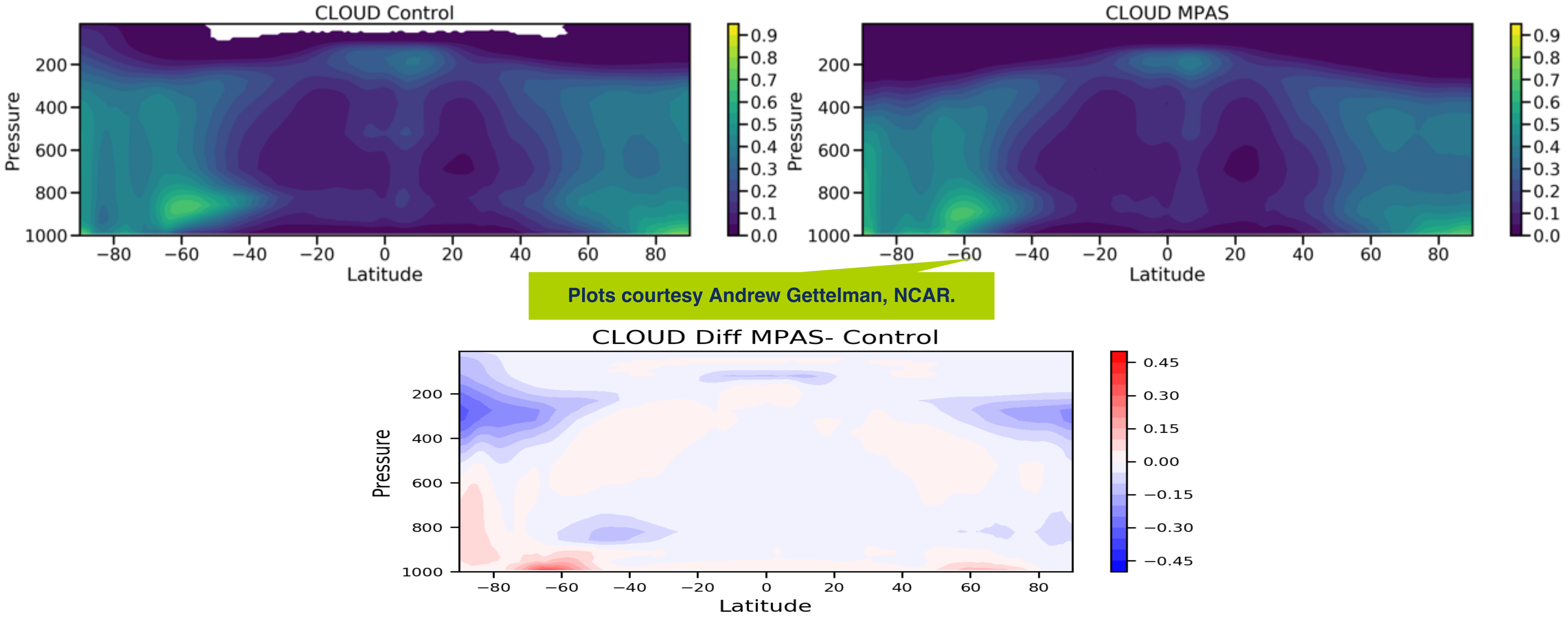


Plots courtesy Andrew Gettelman, NCAR.

Temperature



Small Differences observed in Clouds



Early MPAS-A A100 Benchmark… and performance model

- **Hardware Configuration (@NVIDIA)**
  - **Selene**: DGX cluster
    - Dual AMD EPYC 7742 64c Processor +
    - 8 A100 GPUs/ node
- **Software Configuration**
  - MPAS-A Version 6.3
  - PGI Compiler 21.5
  - MPS enabled: 4 MPI ranks per GPU
- **MPAS-A Configuration**
  - Precision: single
  - 120 km quasi-uniform test case
    - 40962 grid points
  - Vertical levels: 56
- **Physics (standard WRF suite)**
  - WSM6 microphysics
  - Asynchronous RRTMG radiation

MPAS-A IBM-GRAF base configuration (56 levels, 6 tracers)

Computation: 126.4 nsec/pt/timestep

Latency: .0154 sec

Configuration adjustments for the GSRM configurations:

Change levels from 56 (linear adjustment to cost)

Substitute PUMAS/MG3 for WSM6 microphysics, add 17 nsec/pt/timestep

Tracers adjustment, for each tracer beyond 6, add:

2 nsec/tracer/pt/timestep

1.1 msec of latency/tracer

Assume weather = 10 tracers; climate = 33 tracers

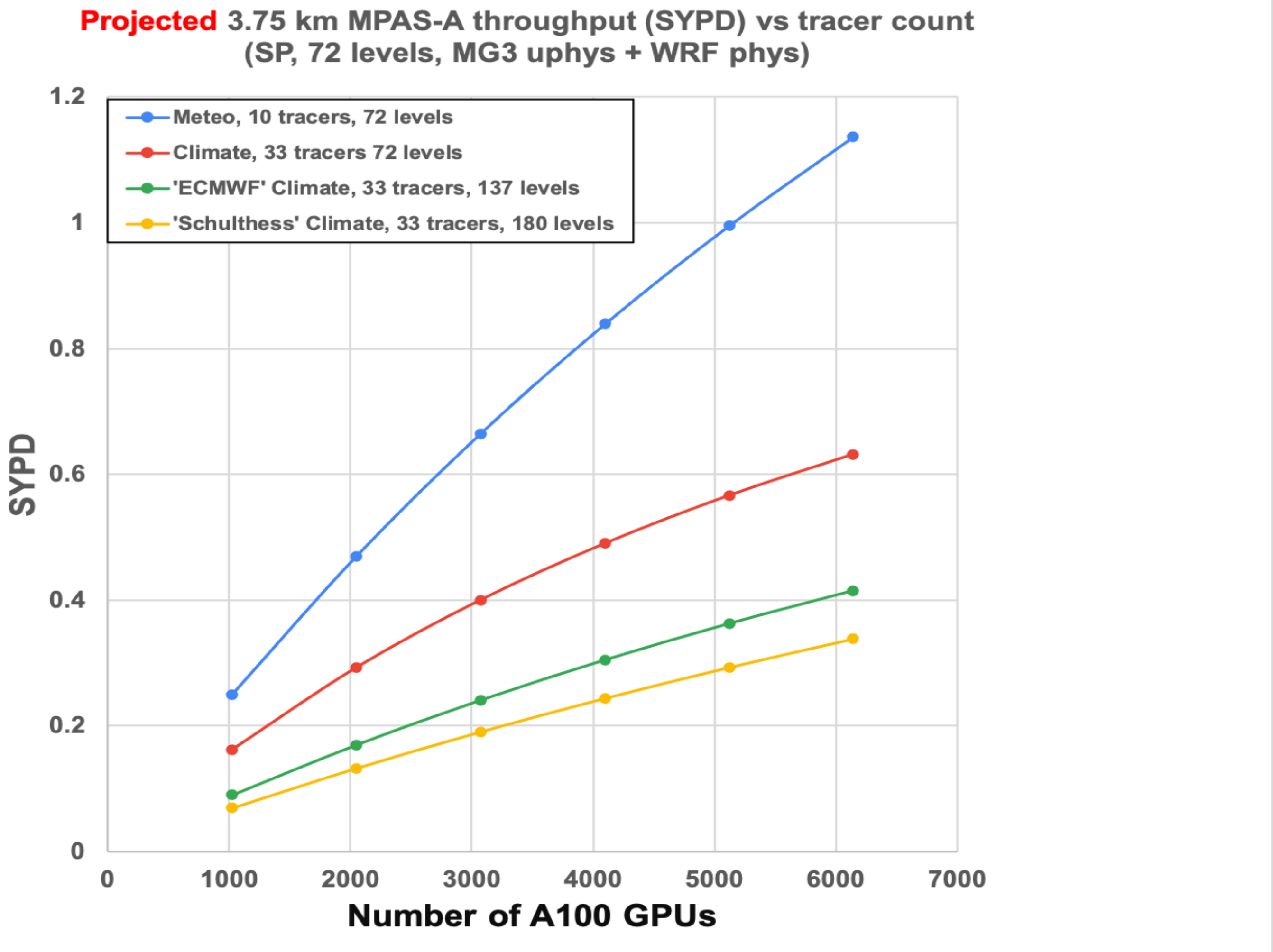
EarthWorks Meteo and Climate Atmospheric Throughput Projections

Some caveats about climate estimates:

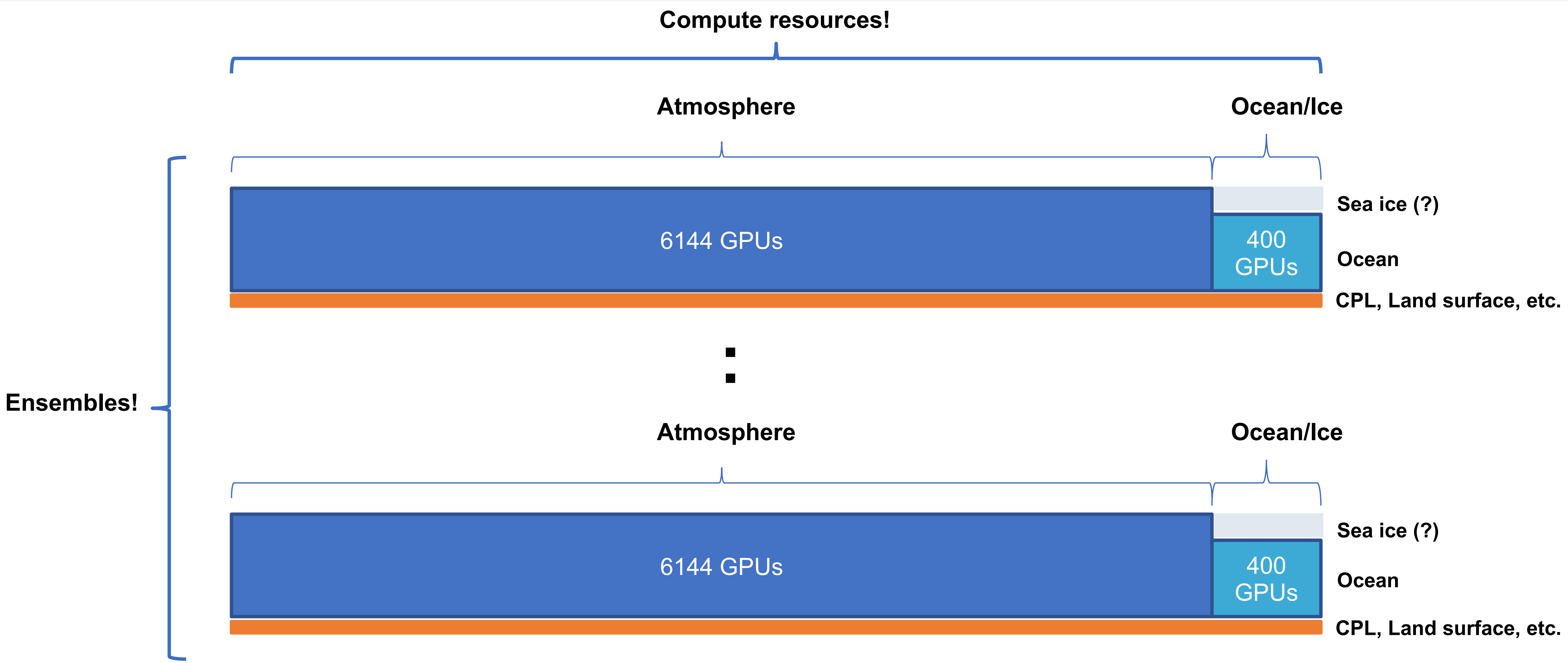
- Assumes a full single precision model. May need double in key places (e.g. tracer advection) for climate runs.
- Assumes overlapping radiation: inline GPU-based RRTMGP may increase physics overhead.
- Does not account for other changes in the physics suite, including replacement of parameterizations with ML surrogates.

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)

"Ambitious target config." Schulthess, et al. (2018)



What kind of GPU resources do we need for GSRM prediction?



Imagine ~180 HPE/Cray Olympus cabinets, each with 512 GPUs!

Some take aways the traditional HPC approach

- If we had the funding and the will, the community probably could perform this science program (or something like it) now, with existing technology.
- For big machines, a critical factor in choosing the architecture is SYPD per MW. Here, at least for now, GPUs have an edge.
- We just need the models.
- As for 1 km predictions at 1 SYPD, I think we're at least two generations of devices to get to that for weather, maybe three generations for climate. But why wait? The planet is warming after all… and there's lot of research to be done.
- The real issue is preparing the models, so maybe getting on with it with the tools at hand seems to me to be the best path.

Outline

- Refactoring the Model for Prediction Across Scales (Atm) for CPU&GPUs
- EarthWorks: Toward a CPU&GPU portable Earth System Model
- Handling the Big Data Problem
- Machine Learning: The Silver Bullet?
- Three Cs needed to pull this off

How to store and analyze EarthWorks output?

- **Use data compression to reduce dataset sizes**
  - ZFP is a computationally intensive, lossy compressor that has error bounded compression and can achieve larger compression ratios (e.g. 4x).
  - LZ4 is a faster, lossless compressor but compresses than ZFP (e.g. 2.5x).
  - We tried both and compared the performance
- **Use the parallelism of DASK, Xarray and Zarr in the Workflow**
  - Dask and Xarray are part of Pangeo, a popular climate analysis tool.
  - Zarr has much higher throughput compared with traditional NetCDF files
  - Zarr can write out data with compression coding (Zlib, LZ4, or ZFP)
  - Zarr works on AWS S3-style storage systems.
  - NCZarr is coming!
  - Benchmarked this workflow on the GLADE POSIX filesystem at NCAR.

Pangeo: use the parallelism of DASK, Xarray and Zarr in the Workflow

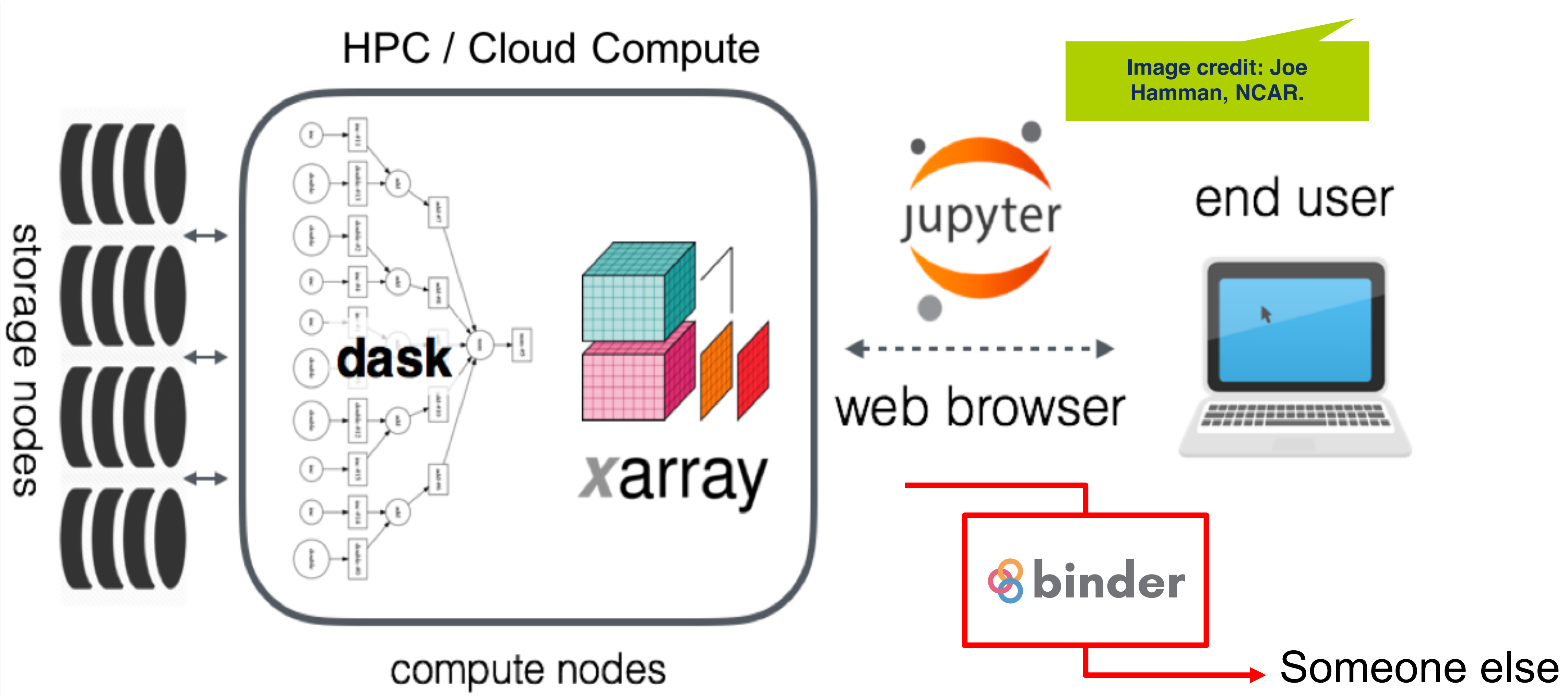
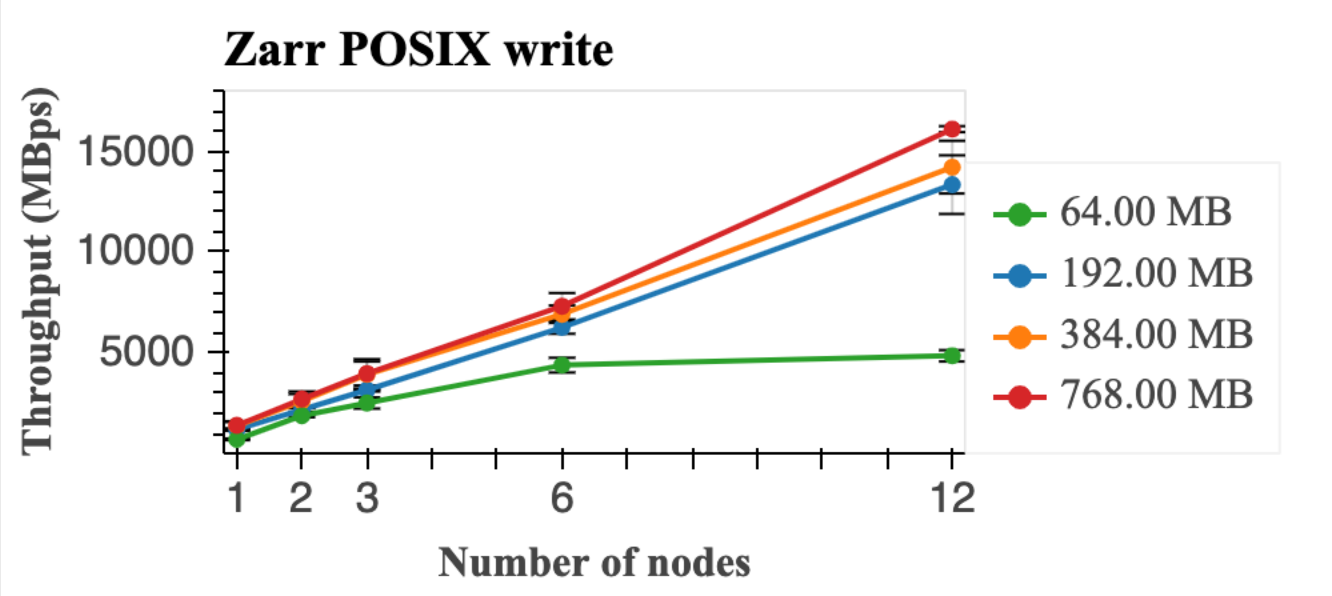
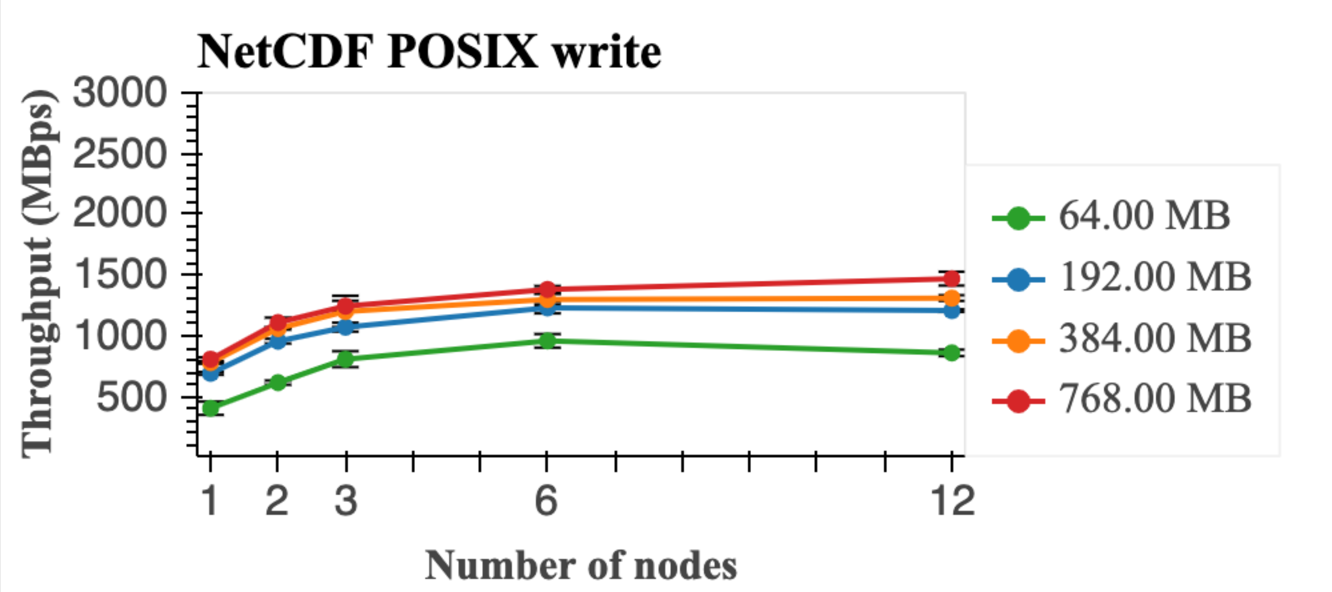
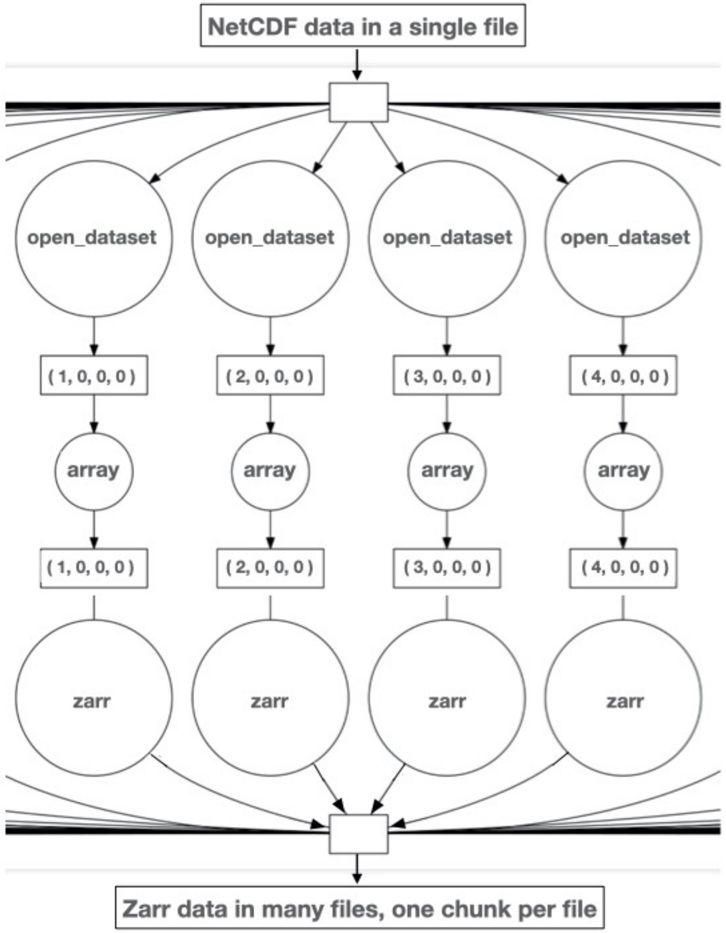


Image credit: Joe Hamman, NCAR.

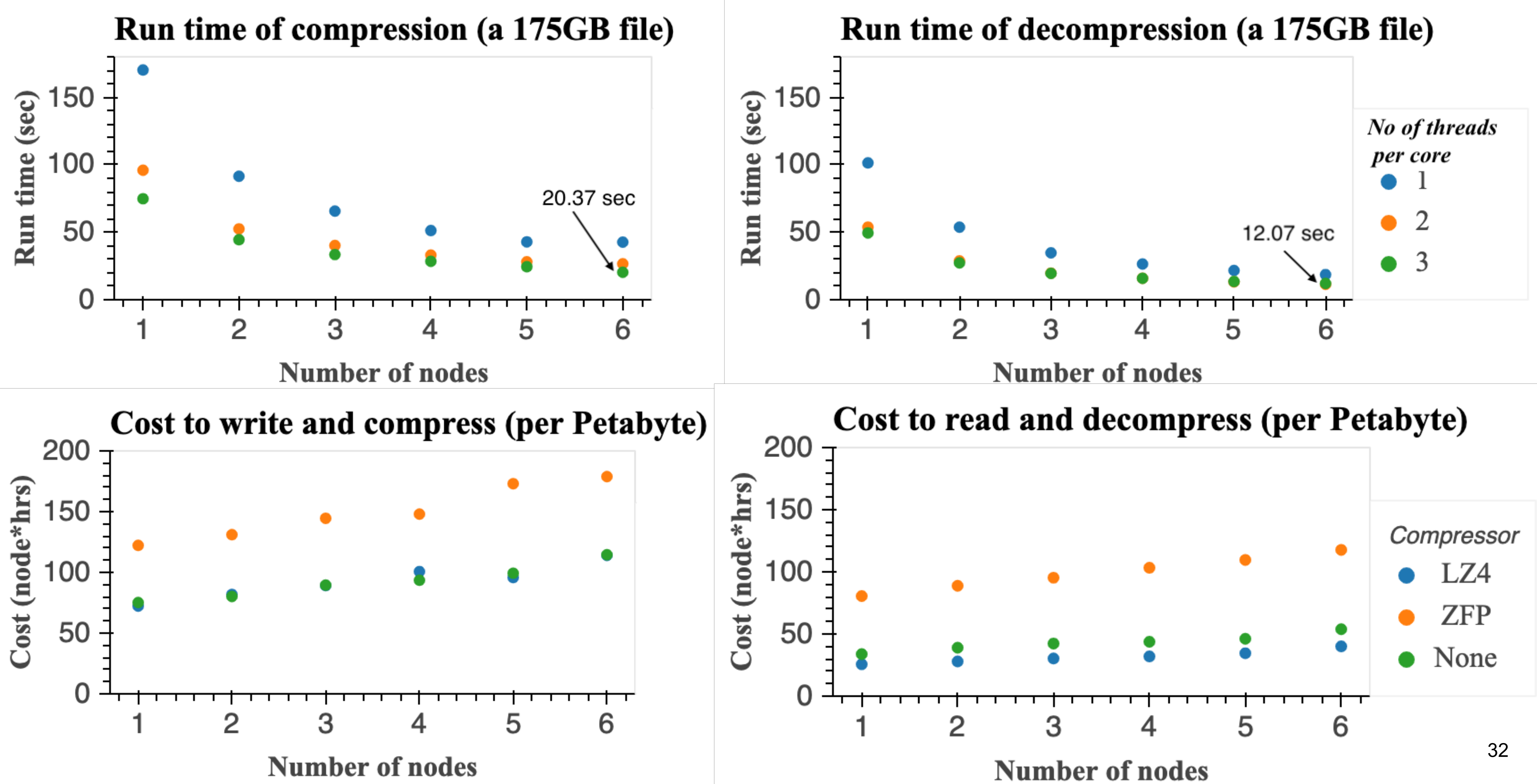
Writing Zarr-chunked Files scales well compared to NetCDF







Costs of Compression/Decompression of GSRM History



Outline

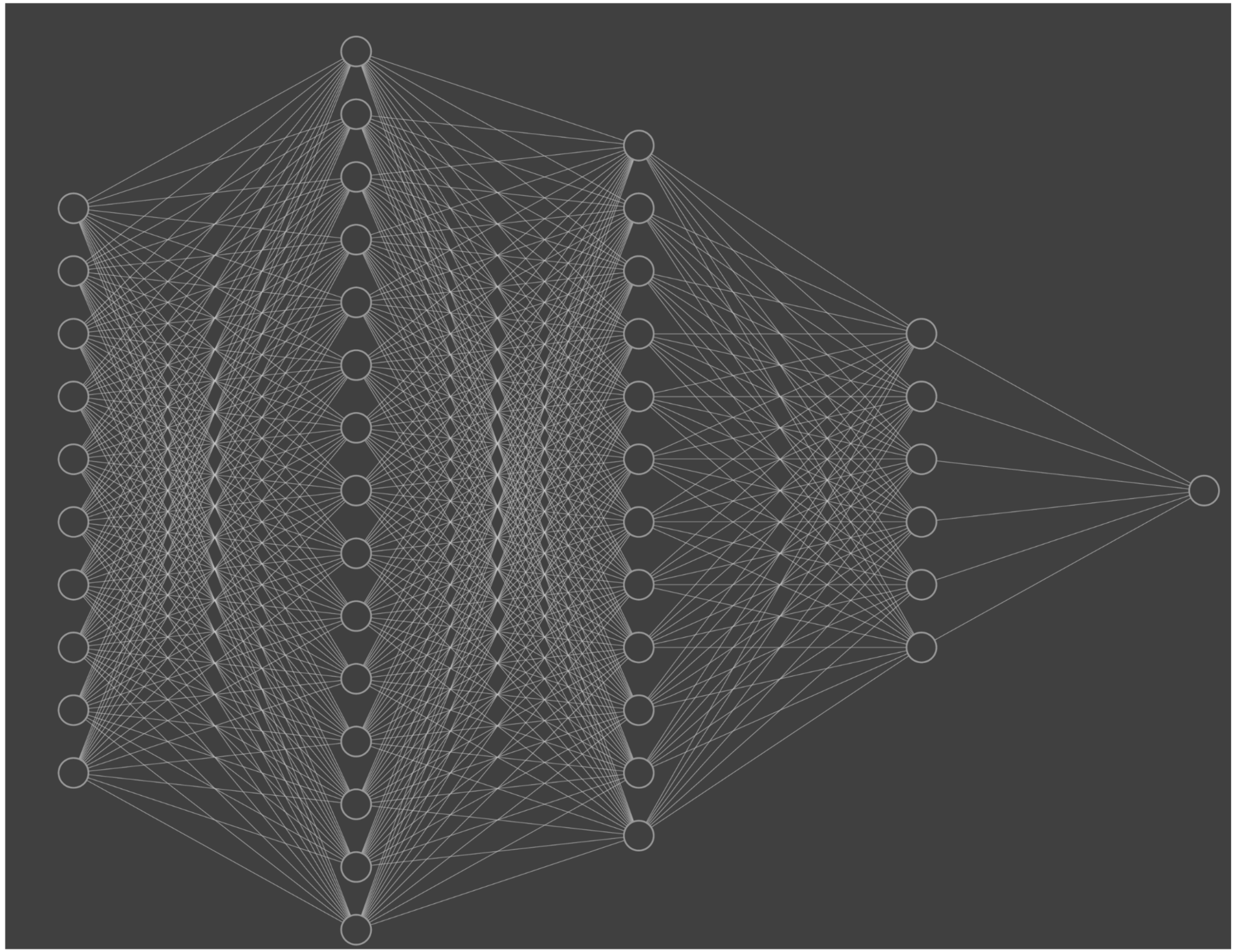
- Refactoring the Model for Prediction Across Scales (Atm) for CPU&GPUs
- EarthWorks: Toward a CPU&GPU portable Earth System Model
- Handling the Big Data Problem
- Machine Learning: The Silver Bullet?
- Three Cs needed to pull this off

Beyond Traditional HPC: Machine Learning

Why explainable AI is important:

“The ultimate answer to life, the universe and everything is… 42!”

Douglas Adams, Hitchhiker’s Guide

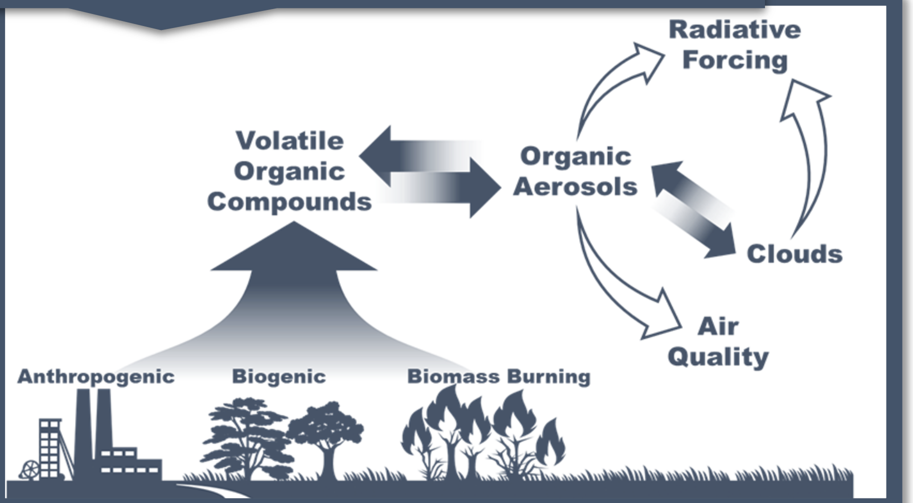


“Despite a growing amount of literature on the success and efficiency of machine learning in different areas, one would ask one important question: what are the limitations/boundaries of this silver bullet?”

Li, J., Li S. and Liao, S. “Can machine learning really solve the three-body problem? (2018)

Machine Learning: taming the complexity of organic aerosol chemistry

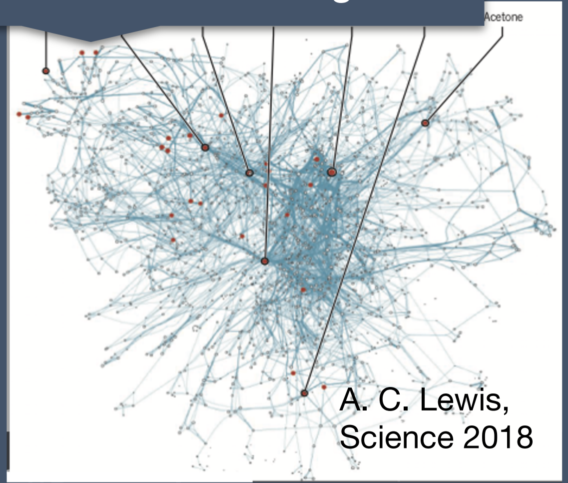
The driver: Understand urban air pollution



Project status:

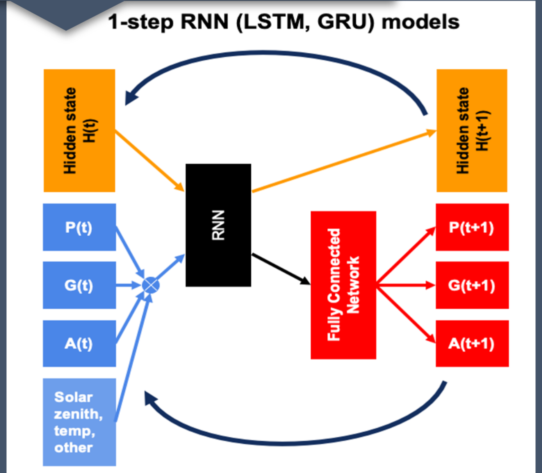
- Testing within the GEOS-Chem model
- RNN inference is ***$O(10^3$–$10^4)$ faster*** than GECKO-A

Complex reaction network: Too slow for modeling!



A. C. Lewis, Science 2018

Training Recursive NN



Results: RNN vs GECKO-A (toluene sample)

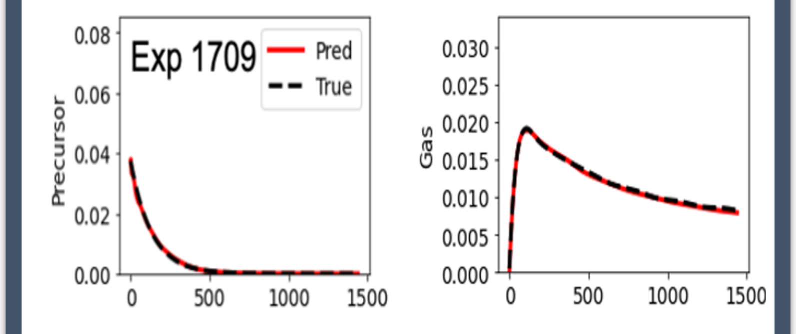


Image: courtesy of AIML group, NCAR

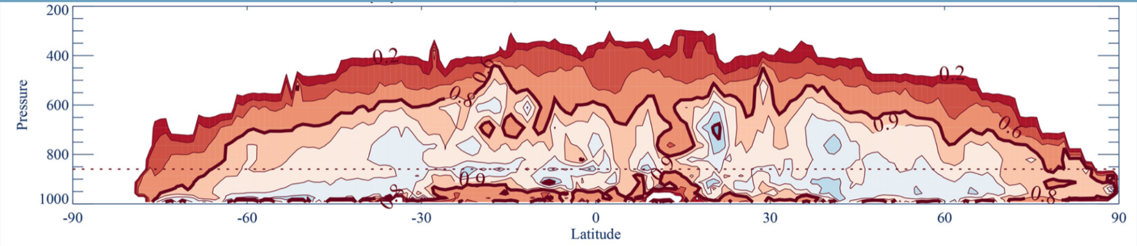
Training data from Eddy Covariance tower in Cabauw, Netherlands.

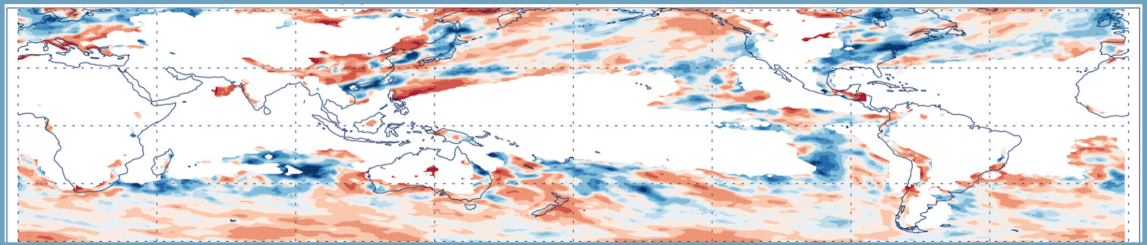
Replacing surface layer flux parameterizations in WRF with ML

David John Gagne (CISL/RAL), Tyler McCandless (RAL), Branko Kosovic (RAL), and Sue Ellen Haupt (RAL).

Emulating the warm rain process in CESM2 (CAM6)

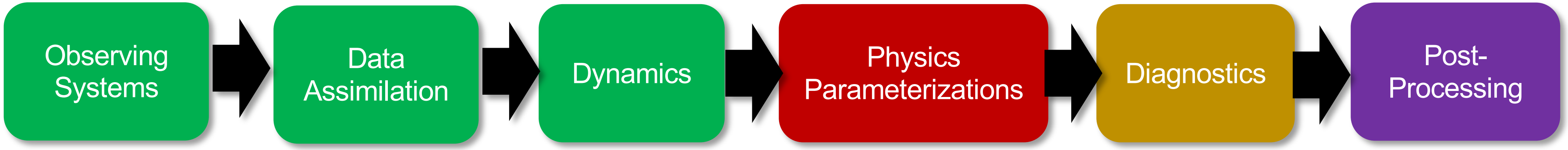
Andrew Gettelman (CGD), David John Gagne (CISL/RAL), Chih-Chieh-Jack Chen (CGD), M. W. Christensen (PNNL), Z. J. Lebo (Univ. of Wyoming), Hugh Morrison (MMM), Gabrielle Gantos (CISL).





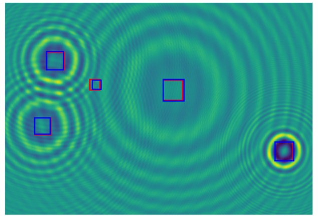
Gettelman et al. (2021); JAMES.

Propagating Machine Learning Throughout the Science Pipeline



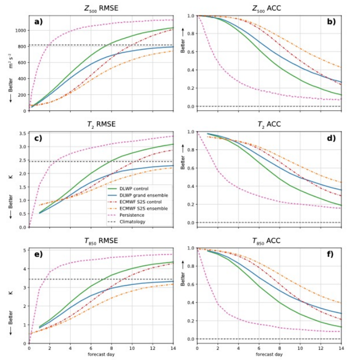
AI at the edge

Holodec imager



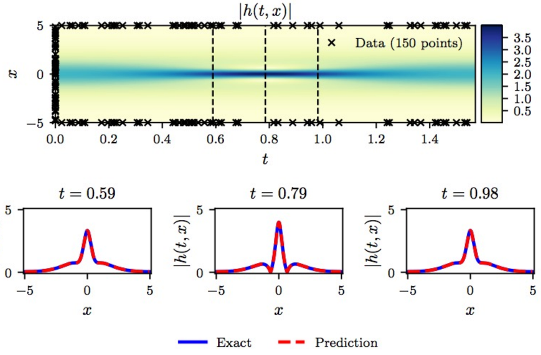
Cloud particle identification

Subseasonal DLWP



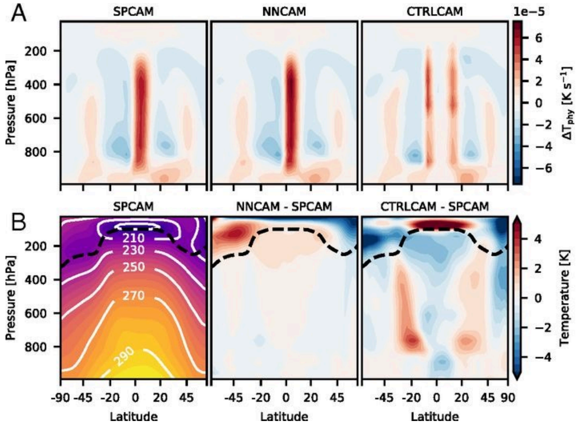
Weyn, J., Durran, D.R., et al.,, 2021

Physics-Informed PDE Solvers



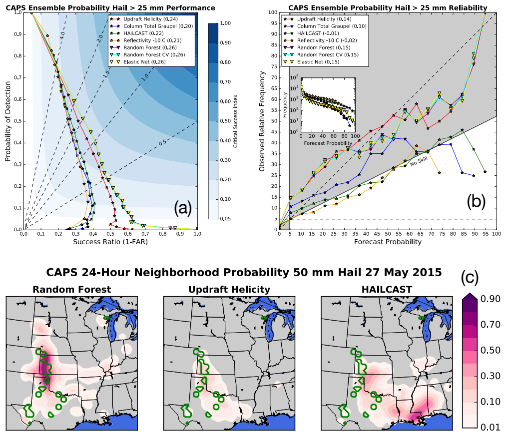
Raissi, M., P. Perdikaris, and G. E. Karniadakis, 2017

Physics Parameterizations



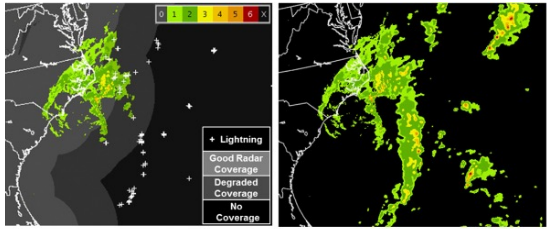
S. Rasp, M. S. Pritchard, and P. Gentine, 2018

Enhanced Hail Prediction



D. J. Gagne et al, 2017

Synthetic Radar Imagery



Veillette, M., Hassey, et al., 2018

Slide credit: D.J. Gagne, NCAR.

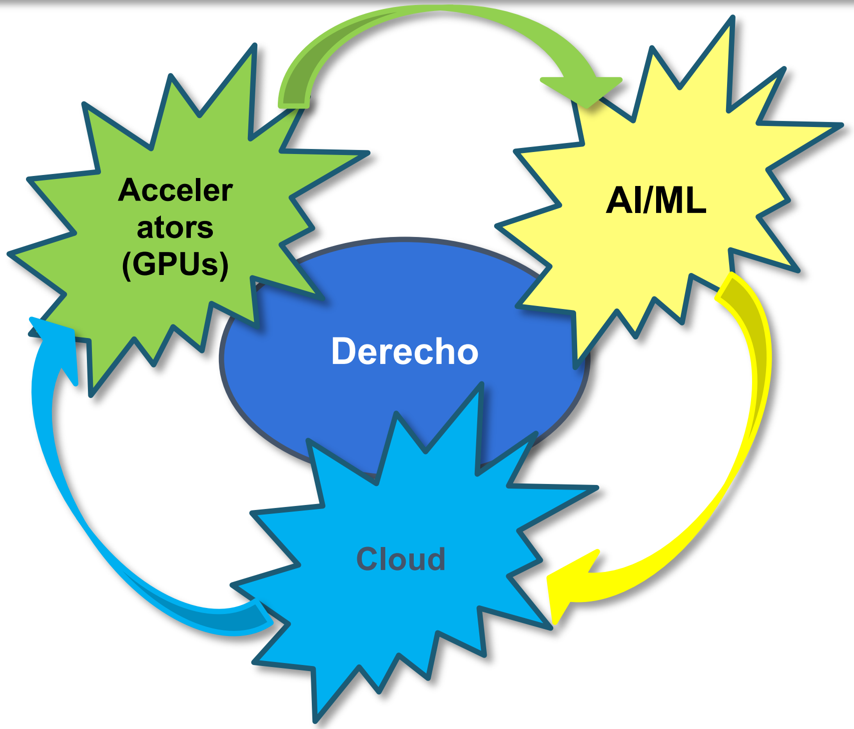
Outline

- Refactoring the Model for Prediction Across Scales (Atm) for CPU&GPUs
- EarthWorks: Toward a CPU&GPU portable Earth System Model
- Handling the Big Data Problem
- Machine Learning: The Silver Bullet?
- Three Cs needed to pull this off
  - Codesign
  - Collaboration
  - Culture Change
  - Cash

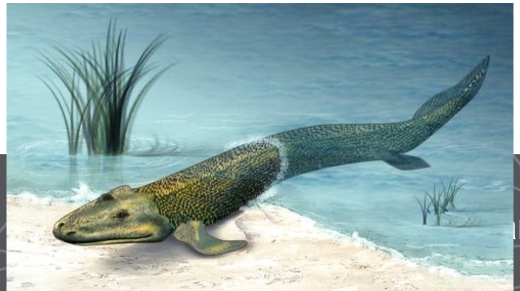
Exascale Path Forward: Implementation Strategy

- Embrace the continuous science-software-system **co-design** process
- Engage and Lead Communities
- Foster a diverse, innovative and empowered workforce
- Explore changes in our business model
- Develop aggressive carbon footprint and sustainability goals with stakeholders

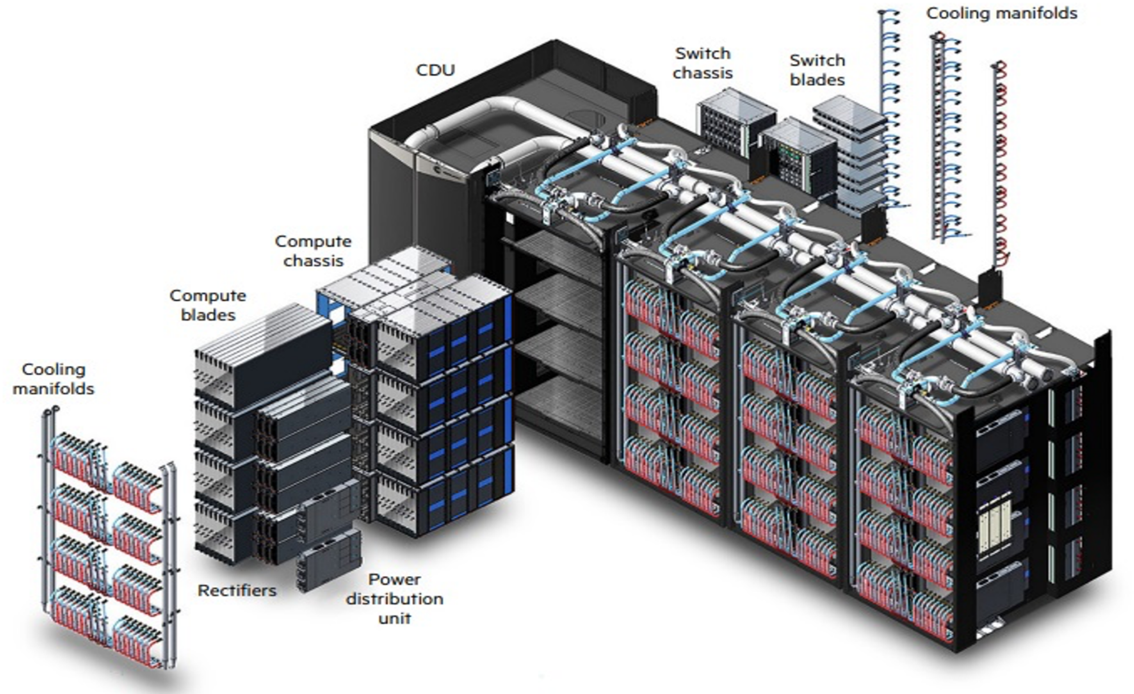
Toward Exascale Systems: Derecho



- At its core, provide a highly productive, data-intensive HPC resource that builds on the success of Cheyenne and the GLADE file system
- Add features and capabilities that leverage the autocatalytic reaction of GPU architectures, AI/ML, cloud and related software technologies.
- Train ourselves and the user community to operate and use these technologies.
- Support related application development and refactoring efforts.



DERECHO, NCAR'S NEW SUPERCOMPUTER

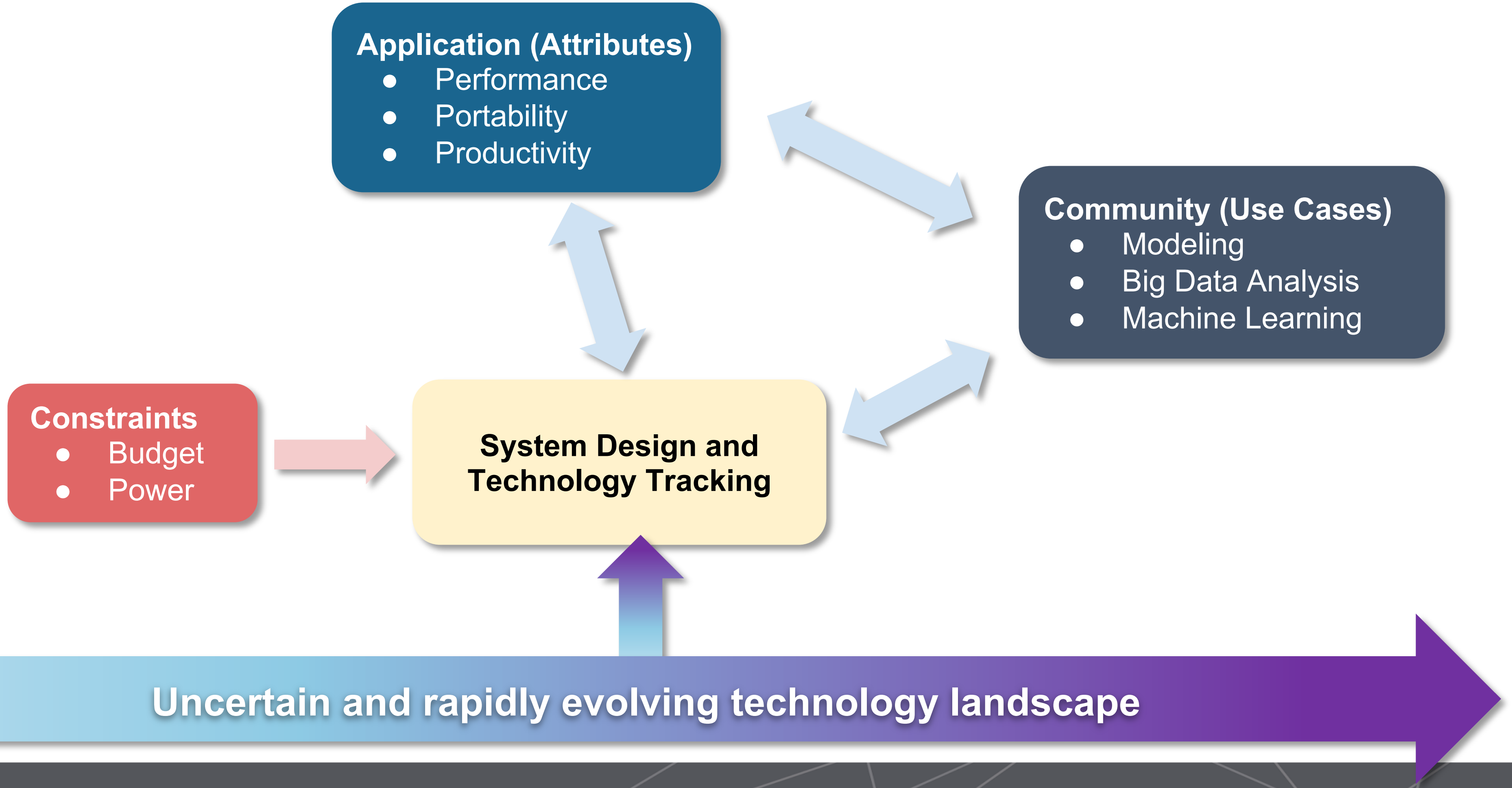


- Derecho (NWSC-3) HPE/Cray XE
  - Peak: 19.87 PetaFLOPS
  - 2488 CPU-only[1] Compute nodes
  - 82, 4-way A100 GPU Compute nodes
  - 60 PB usable file system
- 3.51-fold improvement over Cheyenne sustained Equivalent Performance (CSEP)
- Proportion derived from science requirements:
  - CPU – 2.84 CSEP ~80%
  - GPU – 0.67 CSEP ~20%

[1]dual, 64-core AMD Milan CPUs

[2]single 64-core AMD Milan CPU's

Science-software-system codesign

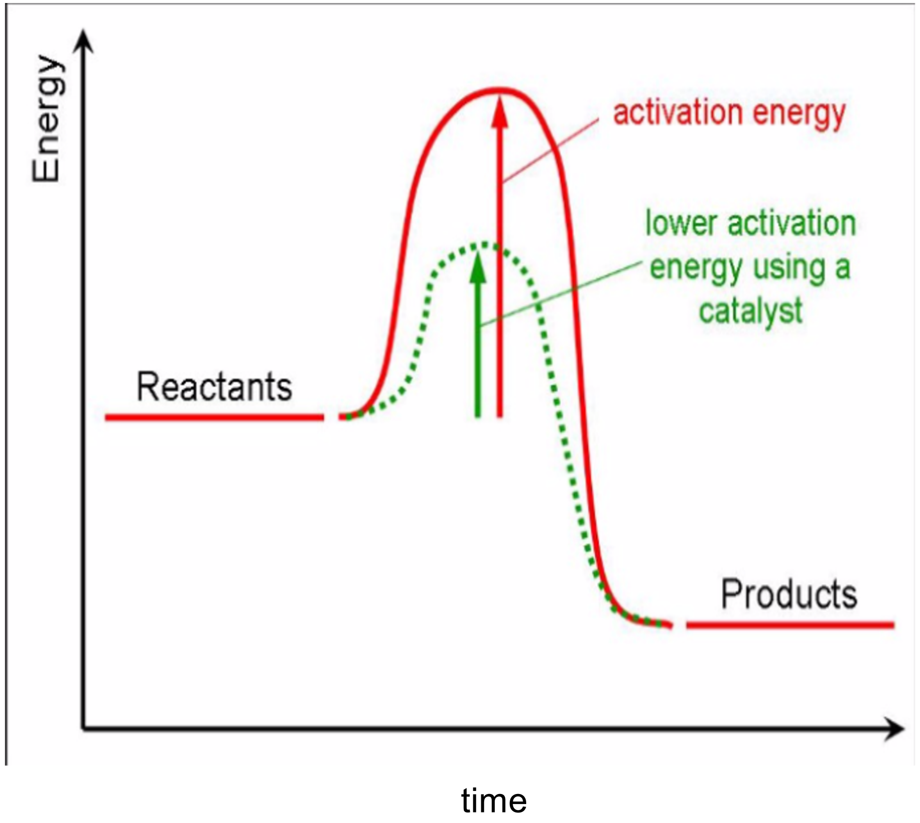


Enabling an Exascale Culture at NCAR

Lowering activation barriers…

- Skills
- Trust
- Collaborations
- Communities of practice

… to catalyze the exascale



Model Refactoring: Partnerships and Workforce Development

- Public-Private, International
- Open Source/Single Source
- Workforce development via student engagement

Key part of our team: university students assistants, interns and NVIDIA experts.

NCAR ML R&D **principles** and **partnerships**:

ML objectives for Earth system modeling **should** address:

- Transparency
- Physics and Domain Knowledge
- Robustness

AI2ES

Julie Demuth (MMM)
David John Gagne (CISL/RAL)

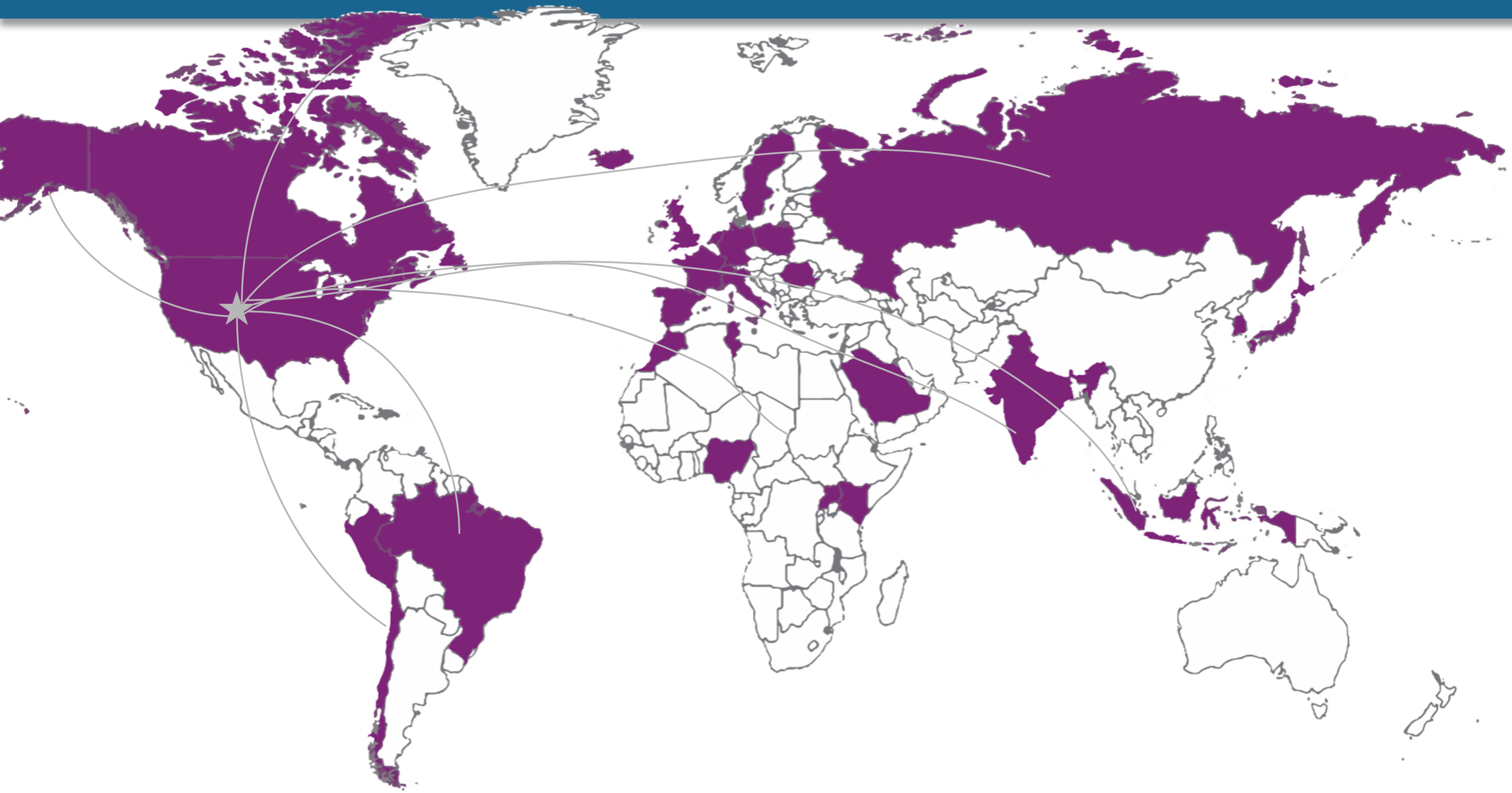
NSF

M²LInES

Marika Holland (CGD)
Judith Berner (MMM/CGD)

SCHMIDT **FUTURES**

Massive online training during the pandemic: AI for ESS Summer School



AI For Earth System Science (AI4ES) 2020 Summer School had ~7,500 individual logins from 39 countries.

NSF AI Institute for Research on Trustworthy AI in Weather, Climate, and Coastal Oceanography (AI2ES)

Trustworthy AI4ES
July 26-29, 2021

Thank you… for being great colleagues and friends!