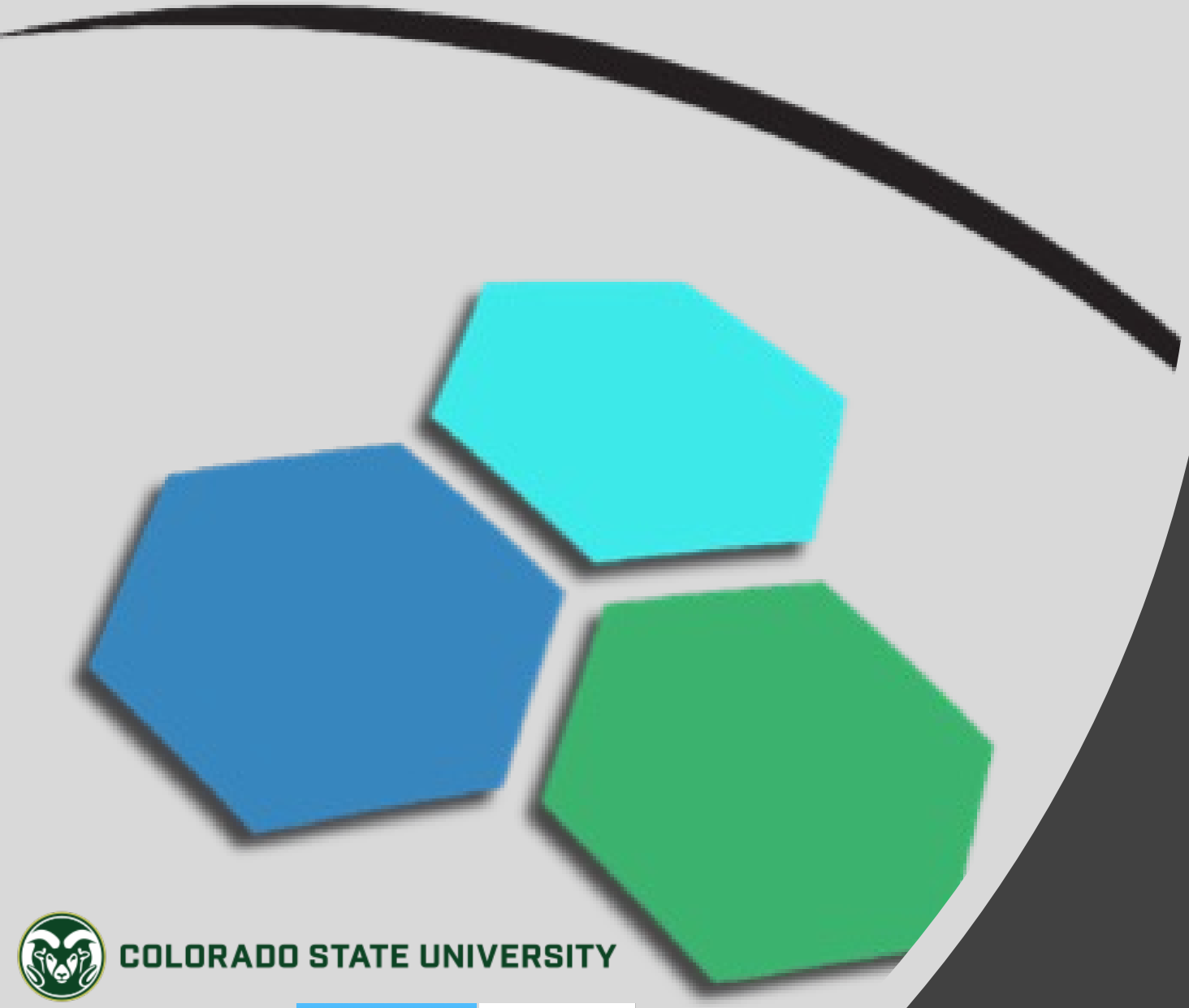


EarthWorks:

The Computational Science Challenges of building an end-to-end, GPU-enabled, km-Scale Modeling System

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], Raghu Raj Kumar[4], Pranay
Reddy Kommera[4]

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

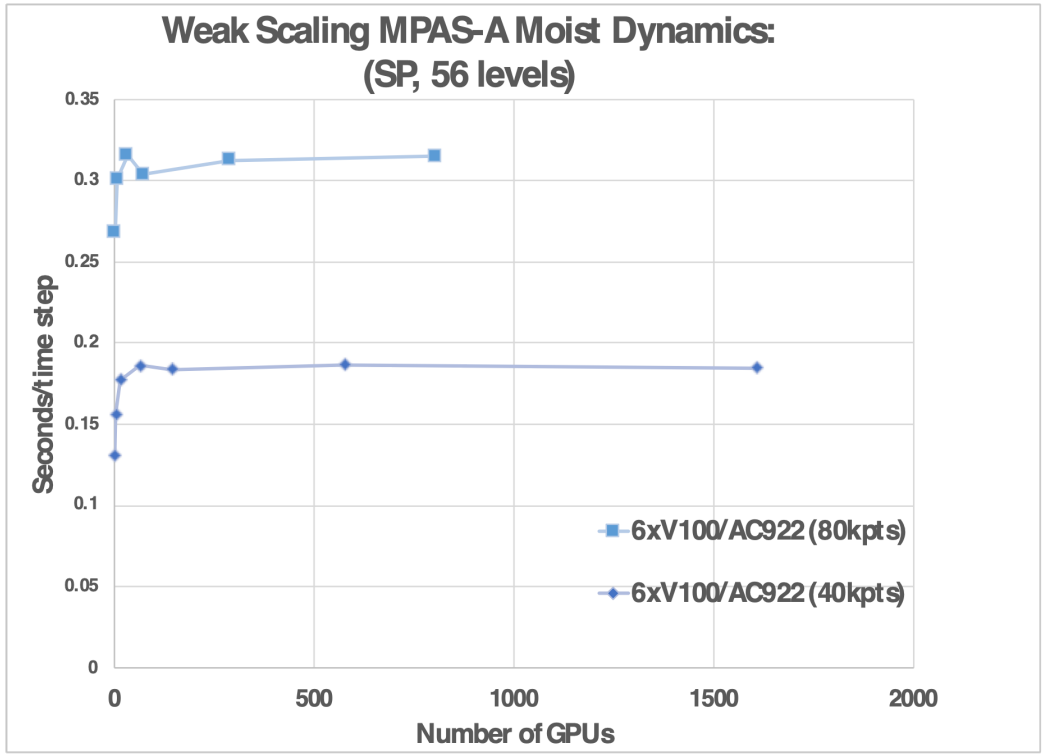
Back Story: Refactoring MPAS-A for GPUs… for 3 km NWP

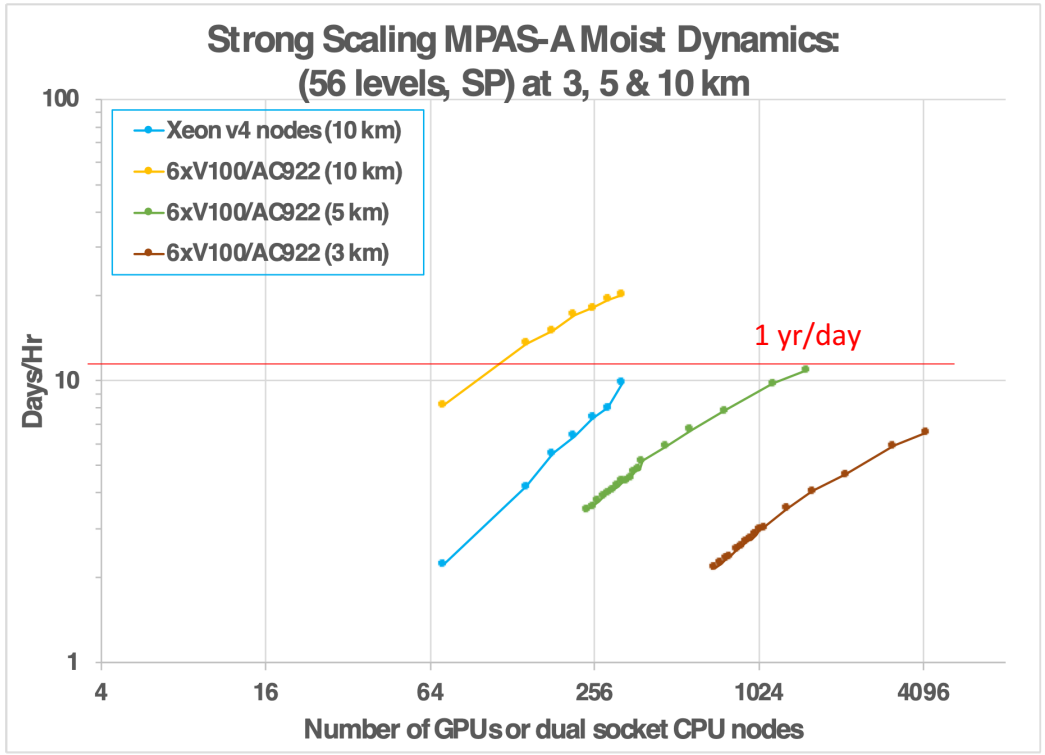
MPAS-OpenACC It is the result of a partnership between NCAR, NVIDIA and IBM / The Weather Company

MPAS-A OpenACC Accomplishments

- Moist dynamical core weak scales to 4,200 GPUs on Summit @ quasi-uniform 3 km resolution.
- Throughput was 3-4x comparing V100 GPUs to Cheyenne Intel Xeon v4 (BWL) nodes.
- Overlapping CPU execution of lagged radiation with GPU model integration.
- In production since October 2019 as part of the IBM-GRAF forecast system.
- OpenACC version available to the community via GitHub.

MPAS-A moist dycore scaling on *Summit*[1] and *Cheyenne*[2]

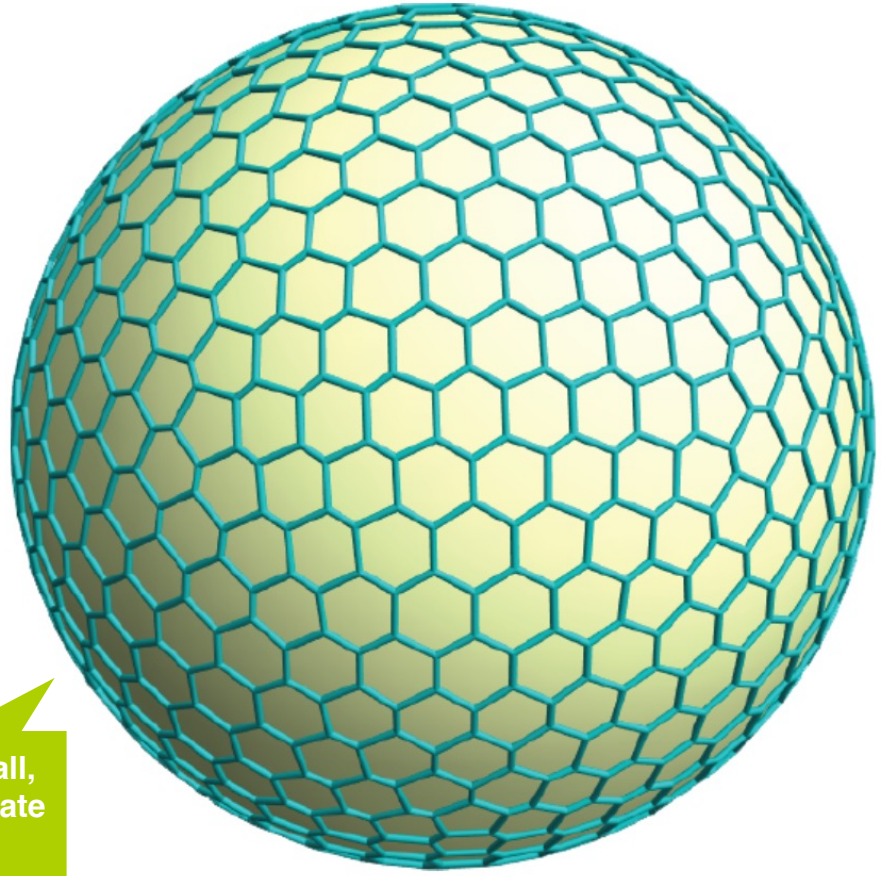




[1]Benchmarking on Summit supported by DoE via an OLCF Director's Discretionary Allocation
[2]Cheyenne is a 5.4 PF, 4032-node HPE system with EDR interconnect operated by NCAR

EarthWorks: Can NWP success translate to an ESM?

- **Put all components on one mesh**
- **Target GSR resolutions**
- **Demonstrate 0.5 years/day at 3.75 km**



Dave Randall, Colorado State Univ.

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

Target grid spacing

Non-hydrostatic regime

Experiences running high resolution tests on CPUs

CESM “F2000 Comp Set” (**Atm + Land + Data Ocean)**

Cheyenne Supercomputer (512 nodes)… ~1/8 of the system

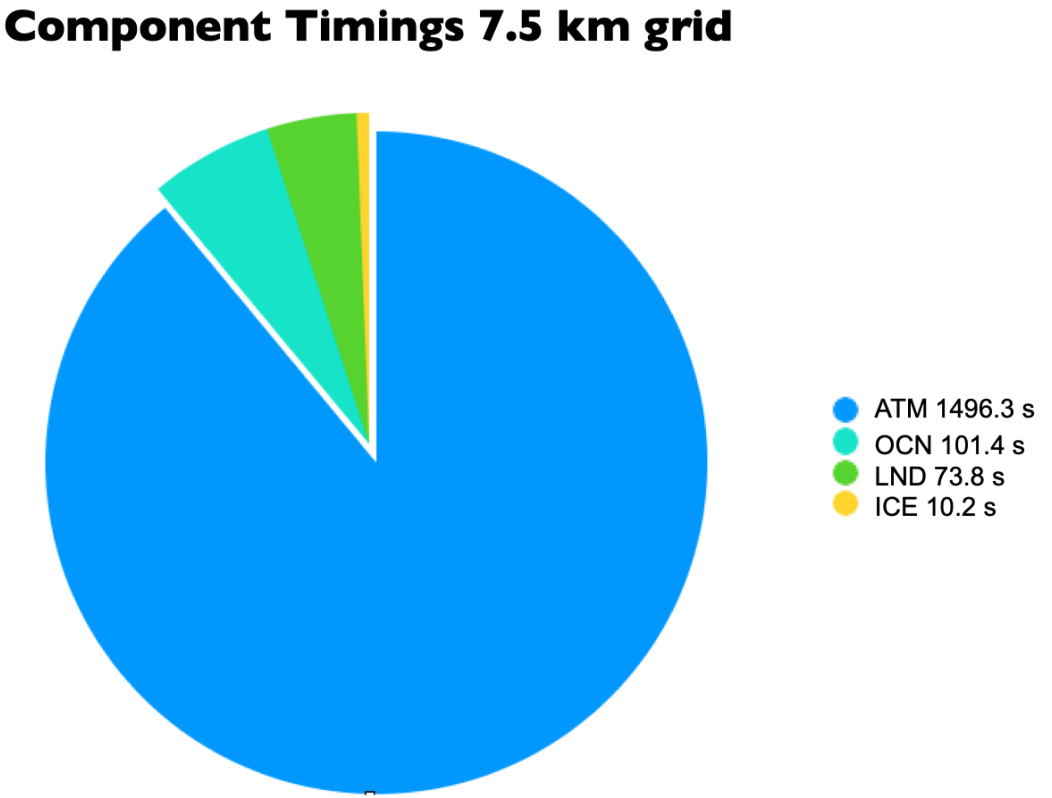
Experiences:

Atmosphere dominates the computational time (89%)

Long turnaround time (lots of queue draining)

Reveals model scalability issues

Slow throughput: in this case 0.08 SYPD @ 7.5 km



Data thanks to Chris Fischer, NCAR

EarthWorks Strategy for GSRM

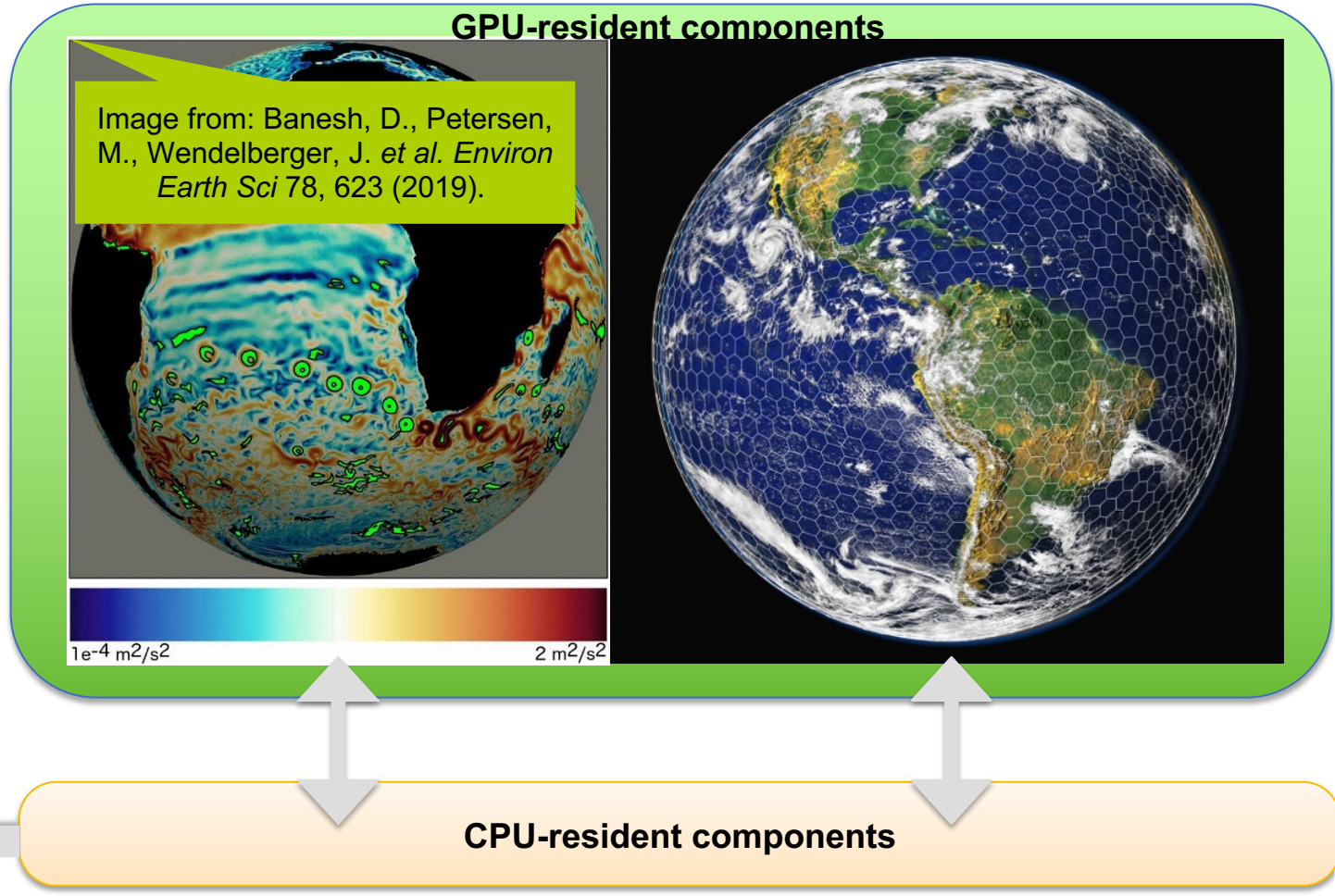
Use regional refinement capabilities of MPAS to reduce cost of tuning climate parameterizations for meteorological length scales

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



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 Left: Benchmark of CPU (cool colors) and GPU (warm colors)

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

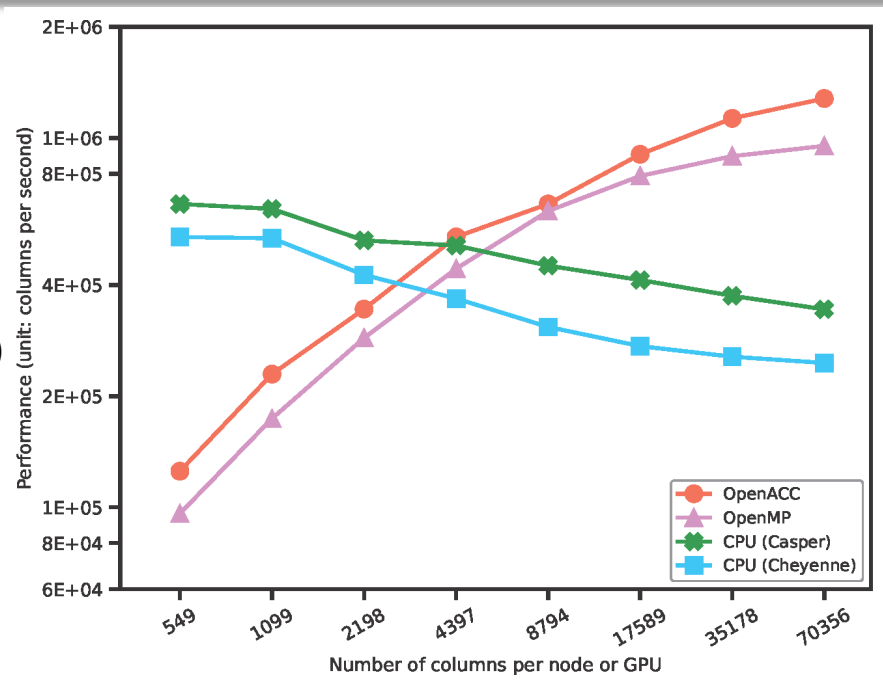
Take Aways:

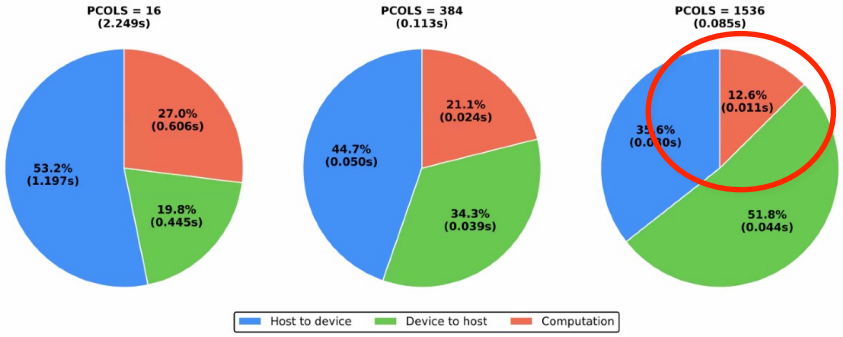
Clearly GPUs favor larger chunk sizes (PCOLS) and higher occupancy (Cols/GPU).

Earthworks will port all physics so we expect no data transfer overhead when calling PUMAS.

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.

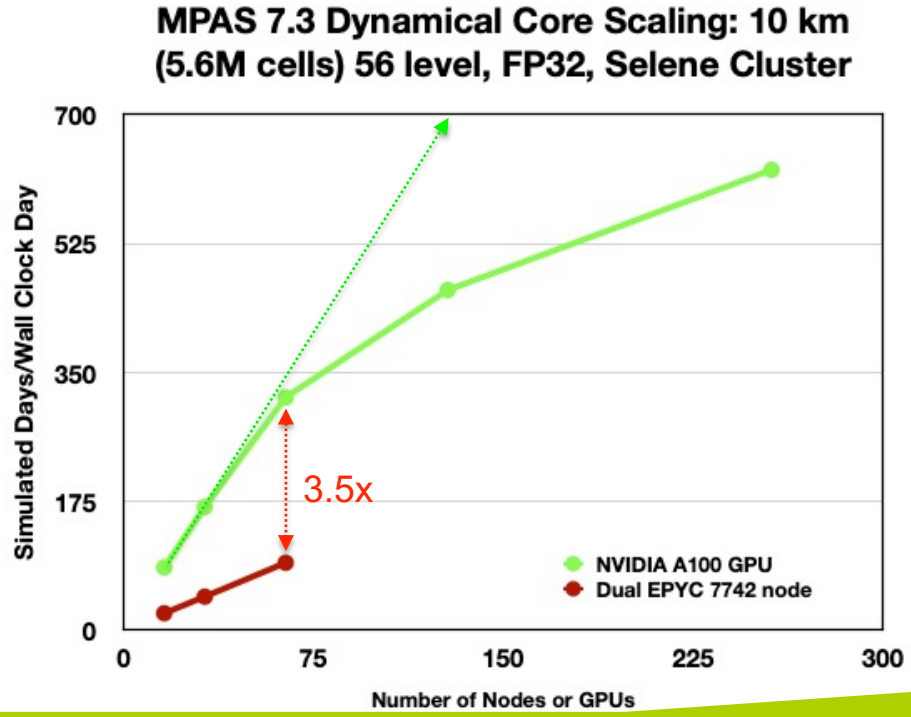
Take Aways:

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.

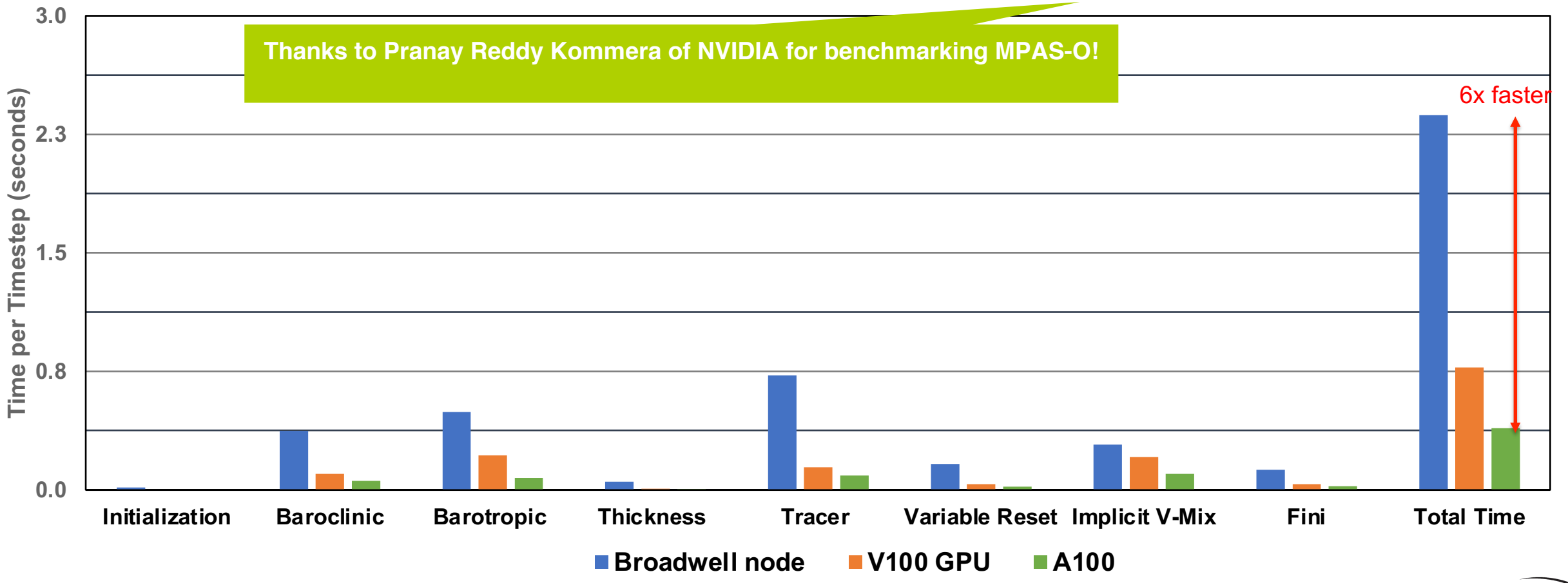


Thanks to Raghu Raj Kumar of NVIDIA for benchmarking MPAS-7!

	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

MPAS-O CPU/GPU Performance Results

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



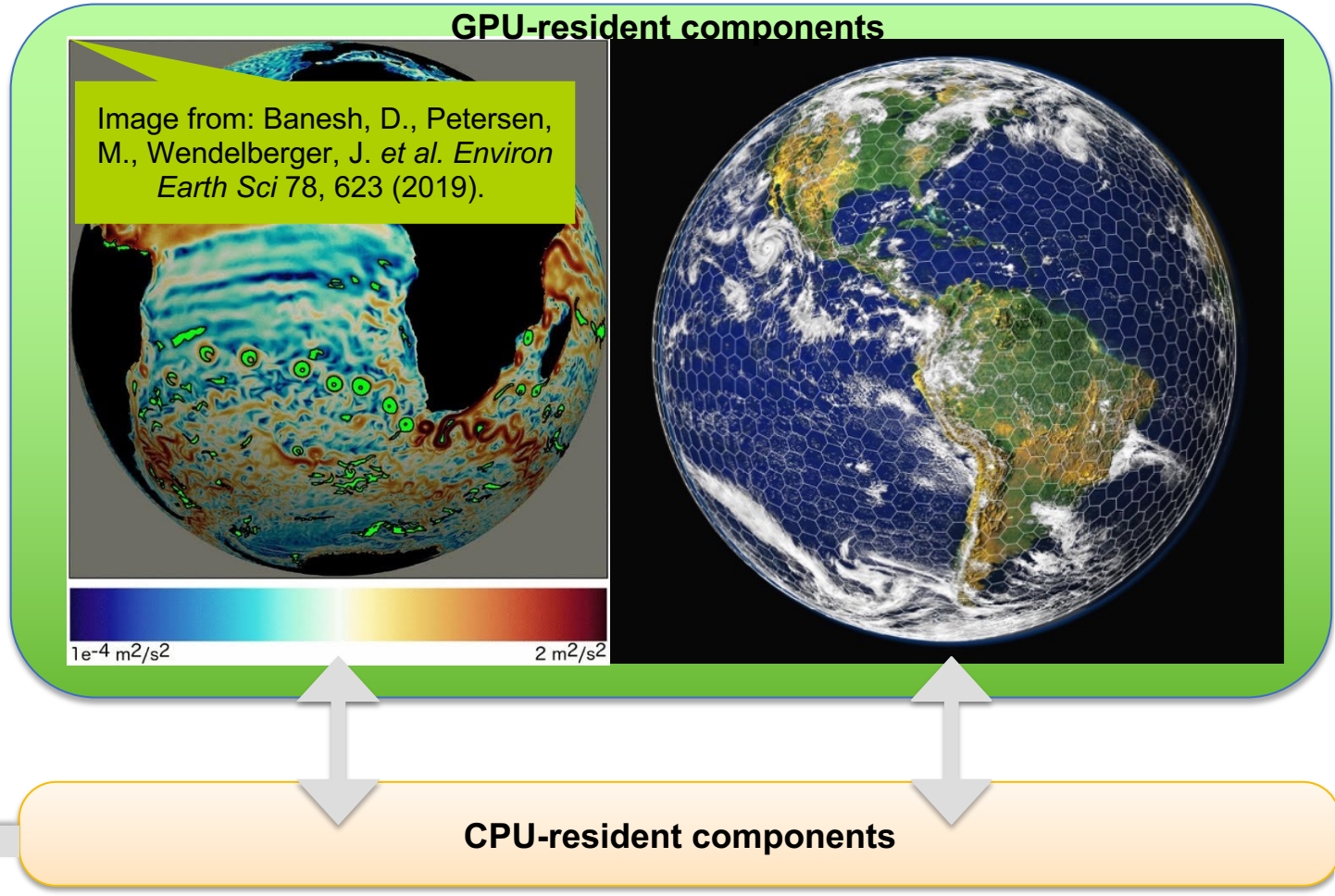
Experiment: Testcase = “EC60to30”; dt = 30 min; 118K cells/GPU; 60 Levels; FP64; 1 MPI Rank/GPU

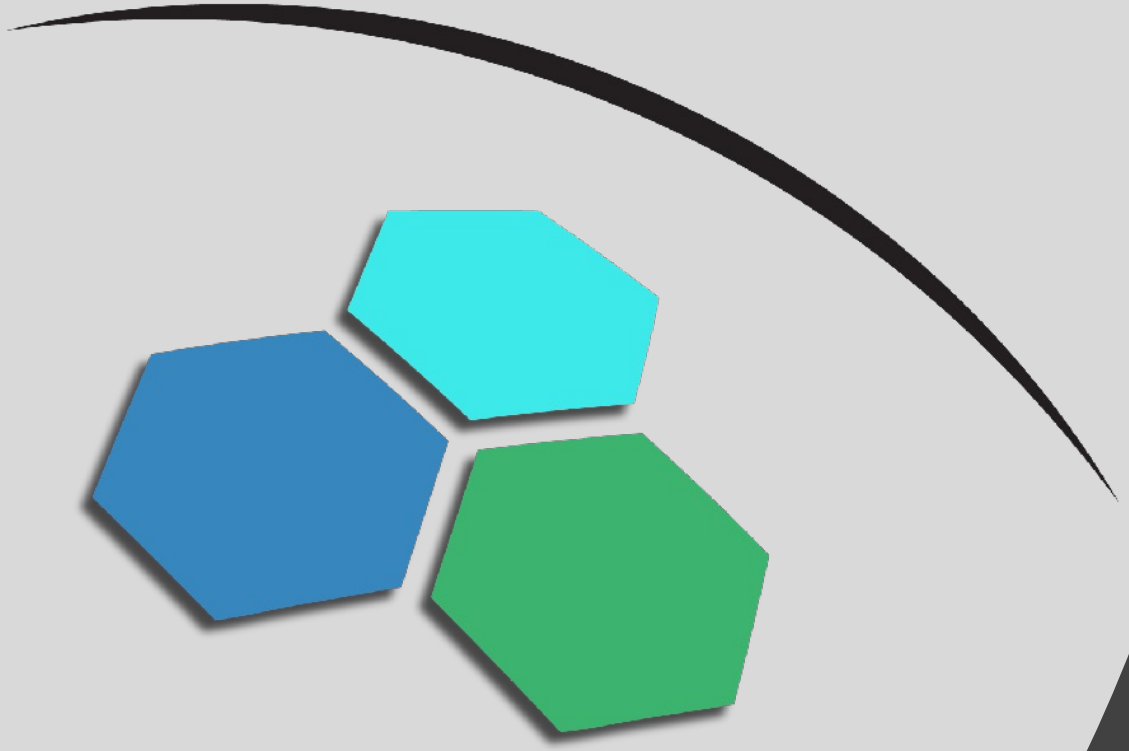
How to store and analyze EarthWorks output?

- **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 chunking 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.
  - Benchmarks of this approach shows good scaling on NCAR's GLADE POSIX filesystem.

EarthWorks: Outstanding Challenges

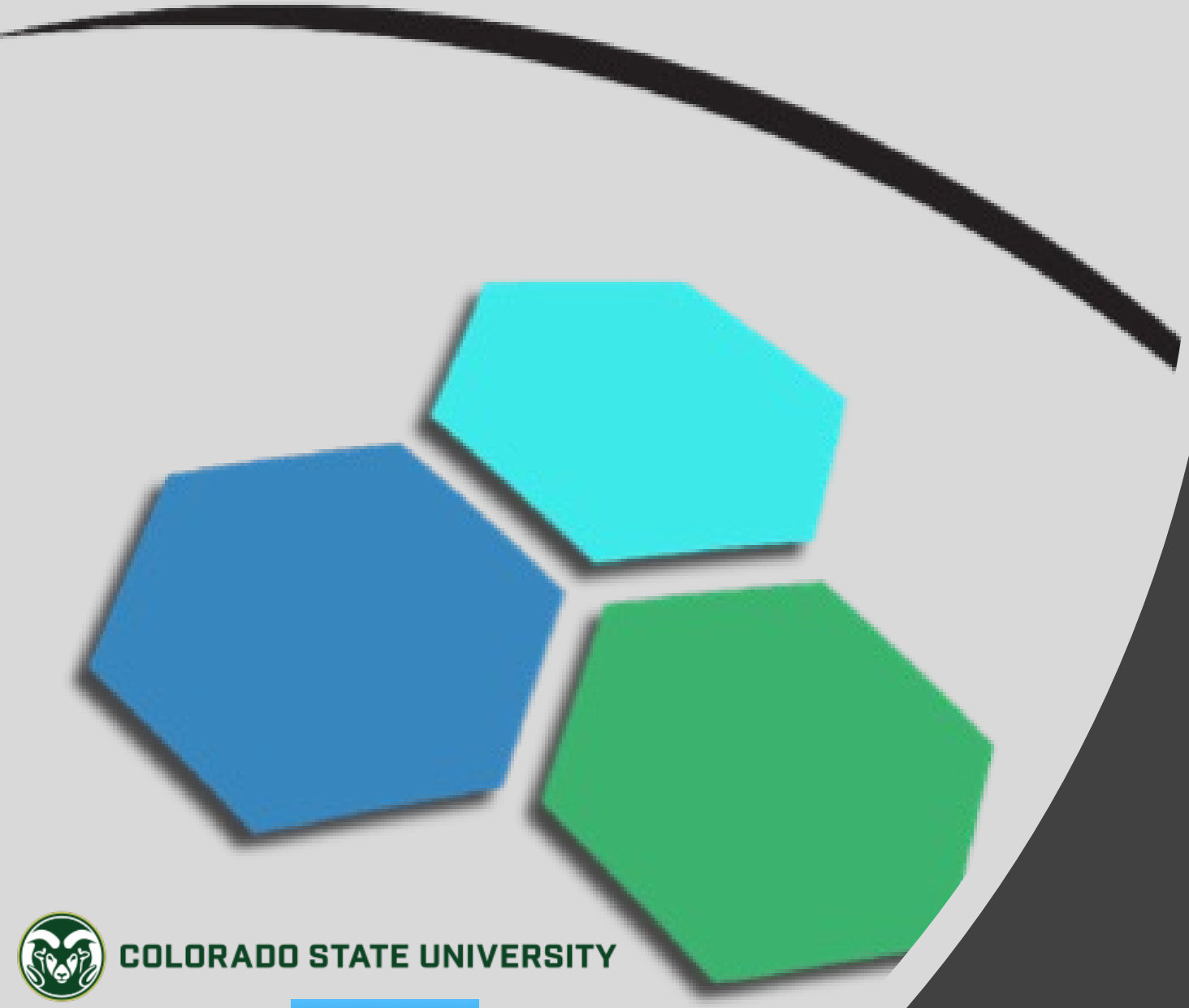
- **Intra-component:** Decoupling the number of ranks/GPU between ESM subcomponents (e.g. Atm. dycore and physics);
- **Inter-component:** Load balancing GPU-resident and CPU-resident ESM components;
- **Precision:** Running (some) ESM components in FP32;
- **Big-data:** Parallelizing climate analysis software ecosystem for ultra-high resolution.





EarthWorks

Please hold questions until the end…



COLORADO STATE UNIVERSITY

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

AreandDee LLC
Come scale away...

EarthWorks

Addressing the Software Engineering Challenges within the EarthWorks Project

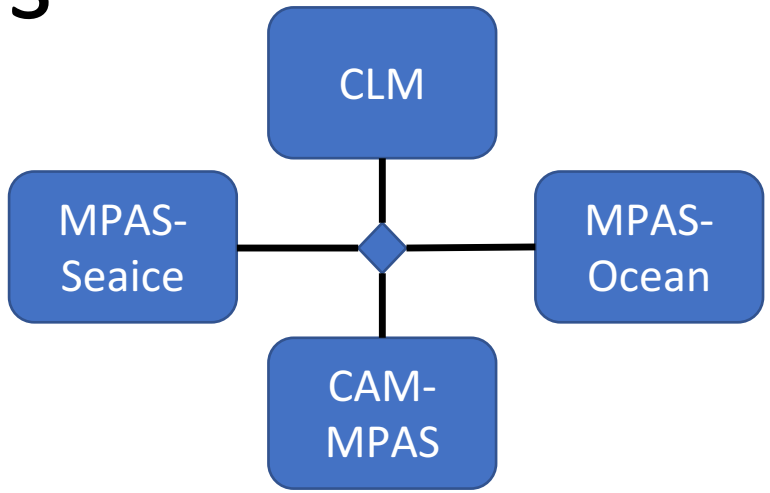
Sheri Mickelson[1]
Thomas Hauser[1], Richard Loft[2],
Michael Duda[1], Dylan Dickerson[1],
Supreeth Suresh[1], Donald Dazlich[3],
John Clyne[1], Jian Sun[1],
Mariana Vertenstein[1]

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

EarthWorks Computational Objectives

Goal: End-to-end workflow portability

Objective	Tools
Revision Control	Github
Containers for portability	Singularity and Docker
Performance portability	OpenACC, OpenMP, OpenMPI
Scalable I/O	PIO
Analysis	Atmospheric Diagnostic Framework and Raijin
Data Transfer	Globus
Science Gateway	Containerized Gateway



Models continue to be developed orthogonally to this project

COLORADO STATE UNIVERSITY

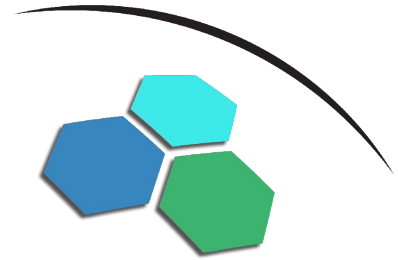
NCAR NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

NVIDIA

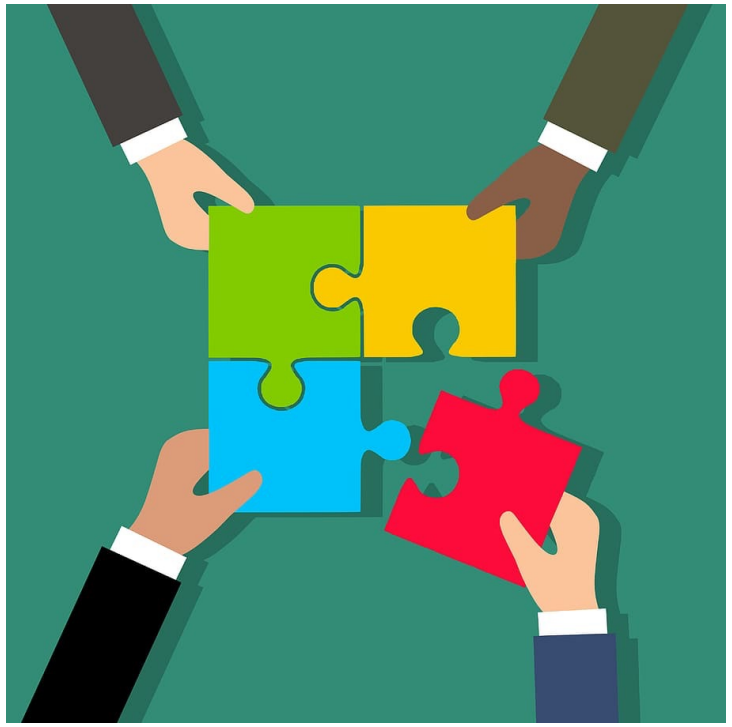
U.S. DEPARTMENT OF ENERGY

AreandDee LLC
Come scale away...

Form a Community
&
Create a Development Plan

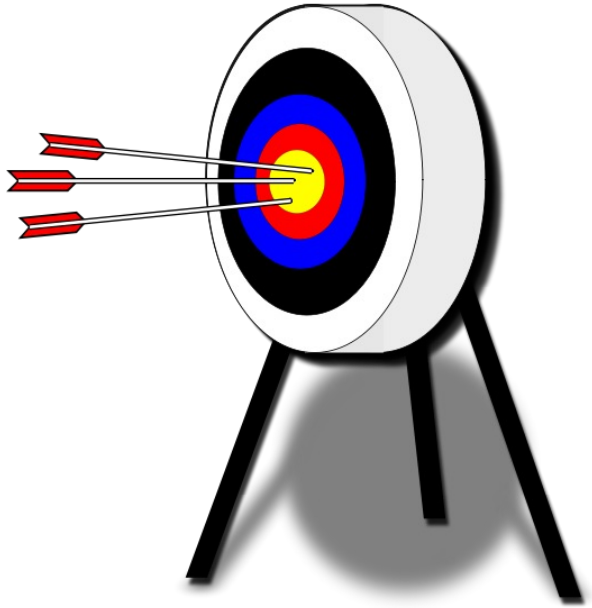


Coordinating the Effort Among the Software Engineers



- Get everyone in the same room and talking to one another
- All voices are equally important
- Celebrate all successes
- Create a clear vision and path
- Everyone should understand how their contributions fit into the big picture
- Remove any barriers to entry
- Empower all team members

Coordinating the Effort with Scientists
- the moving target problem -



- The scientists have to be equally invested in the project
- The software engineers have to be aware of the science planning
- The scientists need to be involved with the software engineering planning
- How will the software be maintained after the work has completed?

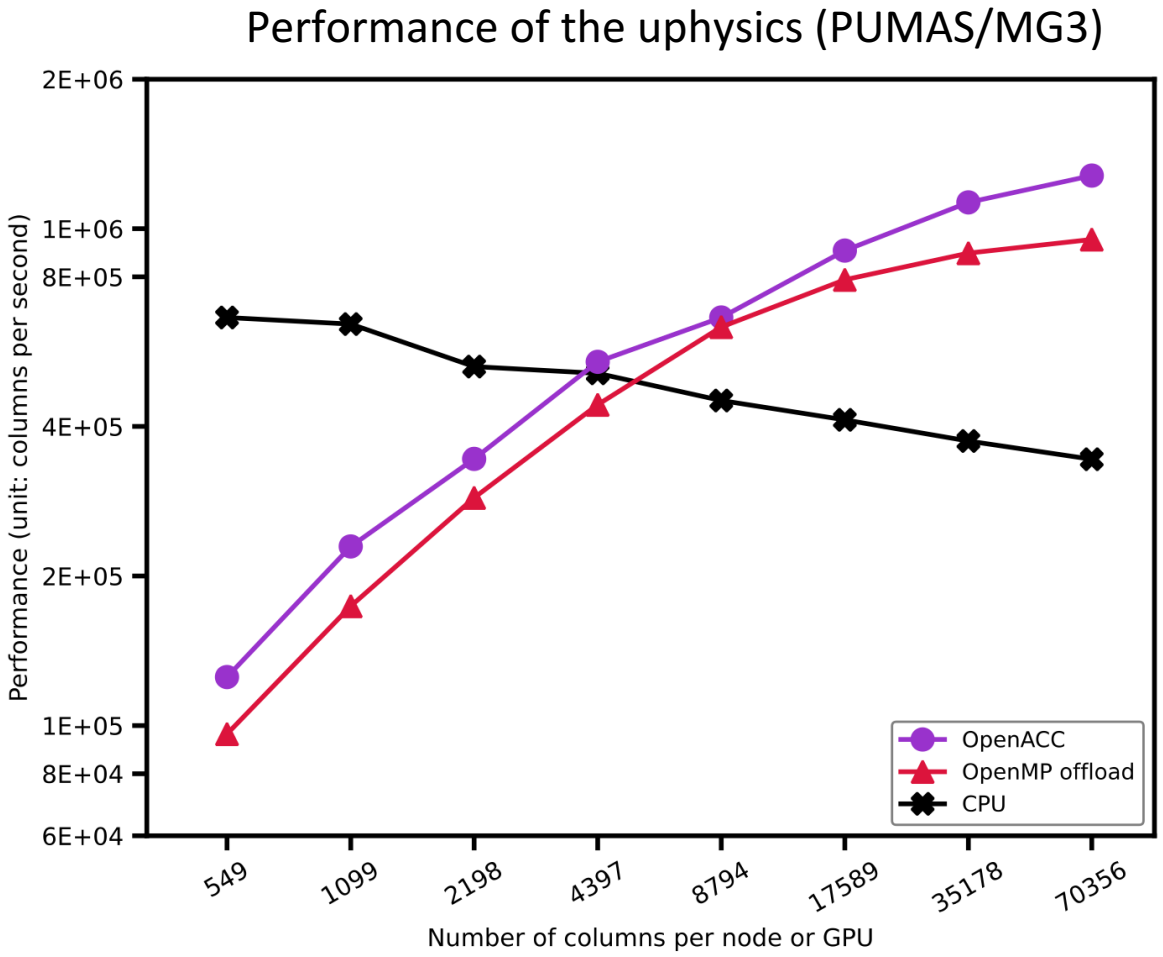
Github

https://github.com/ESCOMP/EarthWorks

In order to create a reproducible software stack that all scientists and developers can work from, we have created a GitHub repo to hold the EarthWorks software stack.

- The repo is in the same location (ESCOMP) as CESM
- Sits parallel to CESM
- The External config files contain the software stack lists

OpenACC vs. OpenMP Offload



Auto Conversion from

OpenACC to OpenMP Offload

We've had great success with this tool from Intel

https://github.com/intel/intel-application-migration-tool-for-openacc-to-openmp

Results from Jian Sun

Data Analysis - Project Raijin

Community Geoscience Analysis Tools for Unstructured Mesh Data

For more information, see Clyne et al. poster at 2022 EarthCube Annual Meeting

Scalable Python tools for analyzing and plotting geoscience data on unstructured grids

- ★ Climate and global weather modeling communities focus (also works with regional data)
- ★ Generalizes NCAR's GeoCAT analysis package to support unstructured mesh data
- ★ Builds on the ubiquitous Xarray package

This work supported through the National Science Foundation Award 2126458.

Testing Infrastructure

- During Development
  - KGEN
  - Testing with multi-column capabilities
  - Test often, performance measurement second
- After Development
  - We've created tests within the CAM testing infrastructure that is ran routinely as new tags are created
    - Smoke test - does it run out of the box?
    - bit-for-bit comparison between CPU vs. GPU
  - We are coordinating with the CAM development group on fixing failures when they are found

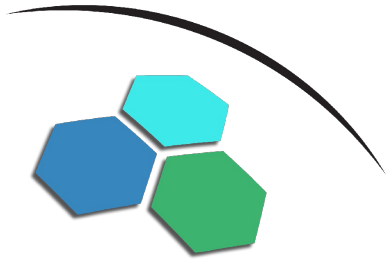
EarthWorks Deliverables Timetable

Version (Target Delivery Date)	Deliverables

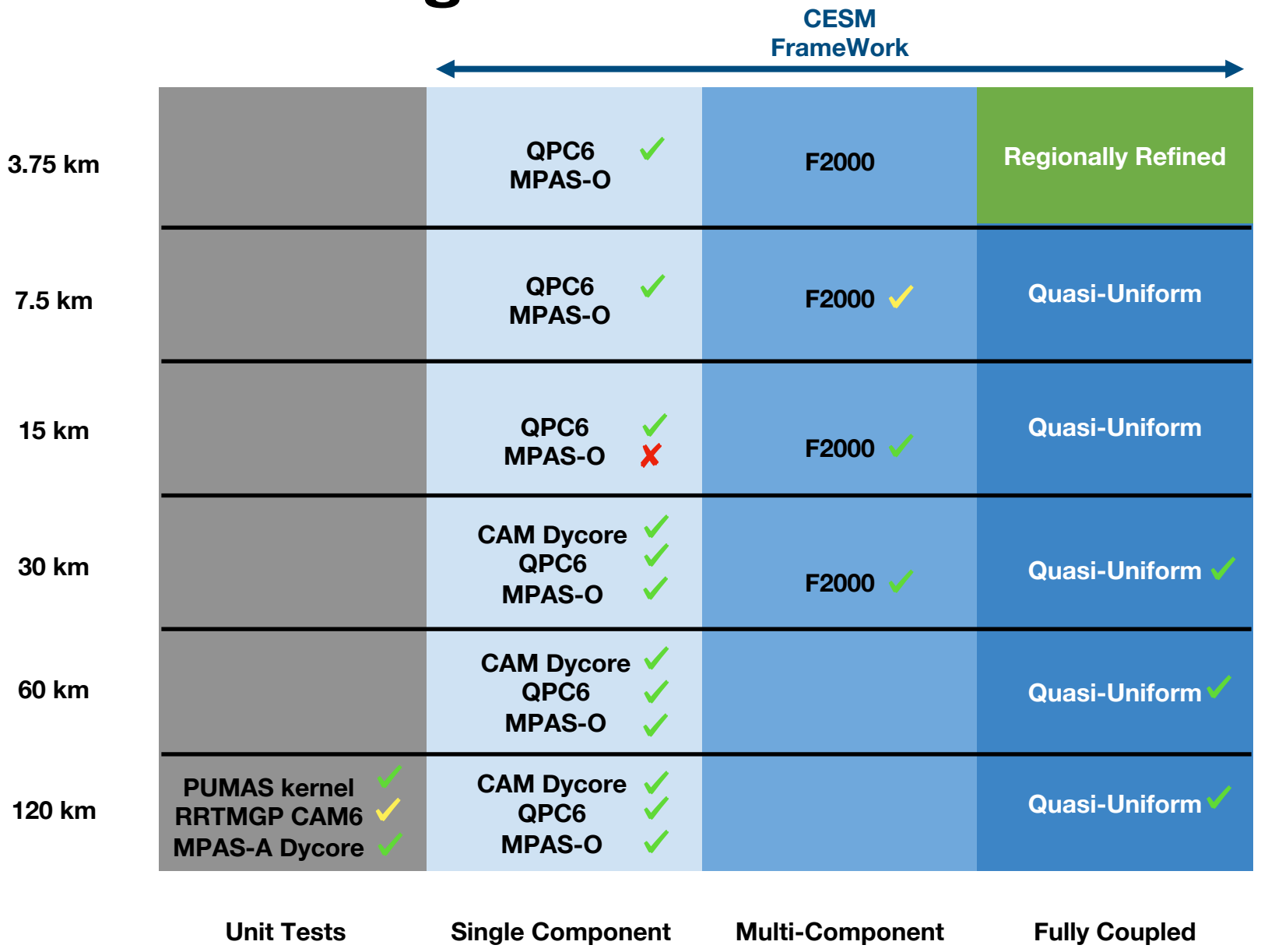
Version 1.0 (December 2022)	• 120,60,30 km
	  • Fully coupled
	  • Standalone MPAS-Ocean
	  • Aquaplanet, Coupled Atm-Lnd
	• Atm dycore tests
	• CPU Intel/GNU compilers

Version 1.5 (May 2023)	• + 15 km resolution • + nyhpc compiler support • + uphysics GPU offload • + MPAS-7 dycore GPU offload • + radiation GPU offload
Version 2.0 (December 2023)	• + 7.5 km resolution • + Scalable tools v1 release • + MPAS-Ocean GPU offload • + Cloud parameterization GPU offload
Version 2.5 (May 2024)	• + 3.75 km resolution • + Scalable tools v2 release

+ New feature



EarthWorks testing status: Intel CPU

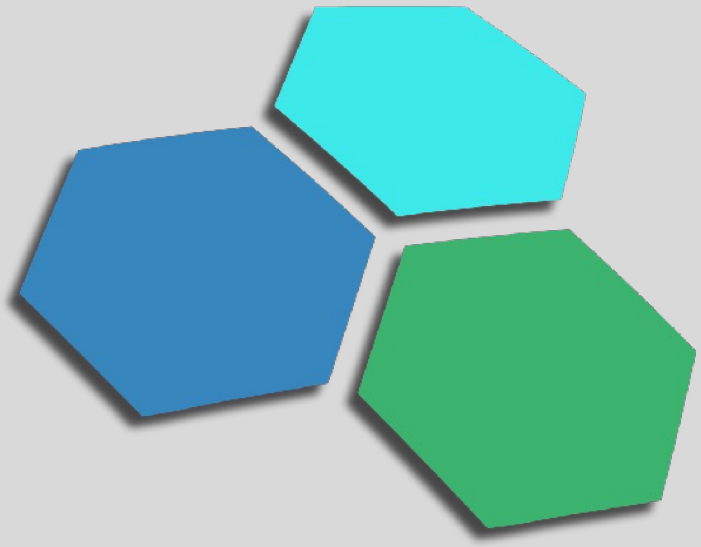


Status Key

- Runs
- Runs but has issues
- Fails

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.



EarthWorks

Questions

Contact Information
richard.d.loft@areanddee.com
mickelso@ucar.edu

Project Website:
http://hogback.atmos.colostate.edu/earthworks

GitHub:
https://github.com/ESCOMP/EarthWorks

Extra Slides

Earthworks Initial Target Platforms

EarthWorks is currently pursuing workflow portability across these leadership-computing systems.

- **NSF Systems:**
  - ***NCAR:*** **Cheyenne[1] -> Derecho***
  - ***Texas Advanced Computing Center:*** **Frontera[1] → Horizon[2]**
- **DoE Systems:**
  - ***Argonne National Lab:*** **ThetaGPU* → Polaris***
  - ***NERSC:*** **Perlmutter***

*AMD + NVIDIA GPU

[1]Intel CPU architecture

[2]Unknown architecture

EarthWorks

EarthWorks is a five-year university-based project, supported by NSF CISE, to develop a global storm-resolving coupled model that uses a single, uniform global grid for the atmosphere, ocean, sea ice and land surface

Earthworks consists of:

- CAM6 with the MPAS non-hydrostatic dynamical core
- MPAS ocean model developed at LANL
- MPAS sea ice model, based on CISE
- Community Land Model (CLM)
- Community Mediator for Earth Prediction Systems (CMEPS)

Goal: retain compatibility with evolving CESM code base and engage with the CESM research community

Goal: enable community-based global storm resolving ES modeling on U.S. supercomputers

EarthWorks

GOAL:

- Capability to perform 3.75 km fully coupled simulations utilizing GPU-enabled components with end-to-end workflow portability across US leadership computing systems

SOME INITIAL OBJECTIVES:

- Port MPAS Ocean and Sea Ice into the CESM framework ✓
- Assemble a working CPU version of the EarthWorks configuration ✓
- Complete fully coupled simulations at relatively coarse grid resolutions ✓
- Test MPAS-CAM6 physics at convection-permitting spatial scales (Underway)
- Port critical-path portions of EarthWorks to GPUs (Underway)