Preparing science applications for GPUs on NCAR's next-generation Derecho supercomputer

Thomas Hauser (thauser@ucar.edu)
NCAR/CISL Technology Development Division Director

May 10, 2022

NCAR

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.

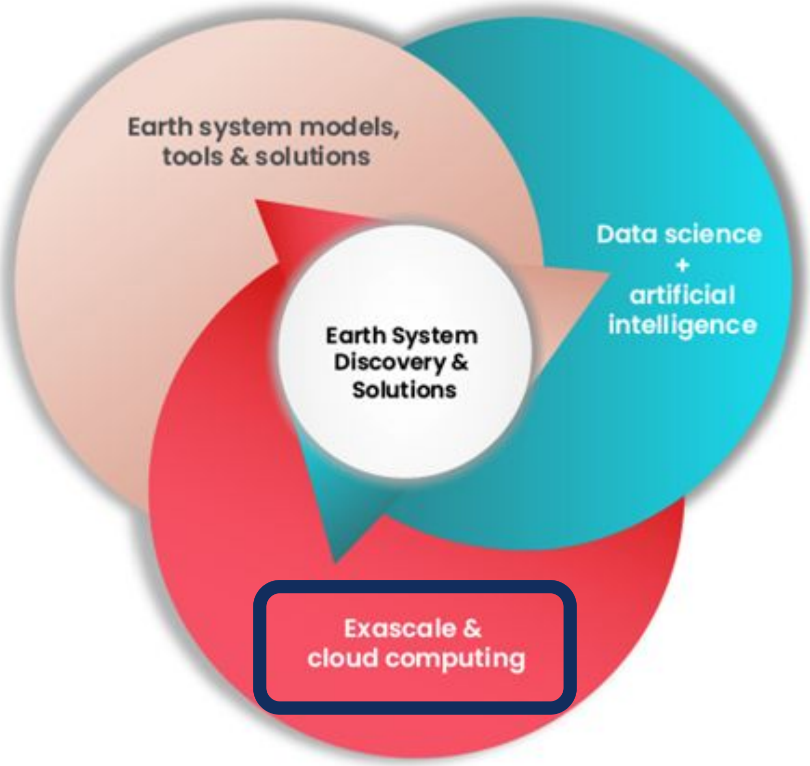
The Role of Exascale Science in NCAR's Strategy

The *NCAR Strategic Plan 2020-24* calls for three ingredients to inter-support:

1. Earth System modeling
2. Exascale computing, and
3. Data-science/AI

*"Develop, support, and evolve our community modeling systems ... for a **hierarchy of computational environments, up to the exascale regime**. …we will increasingly focus **on unified modeling frameworks** to enhance efficiency and enable exploration of scientific frontiers in Earth system science."*

Page 27, Section 3.1



The NCAR Strategic Plan calls for all three ingredients to inter-support.

NCAR's Exascale Transformation Strategy: Investment in communities of practice

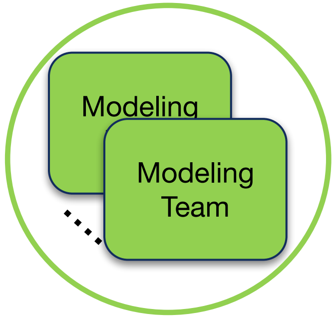
Each focus area does:

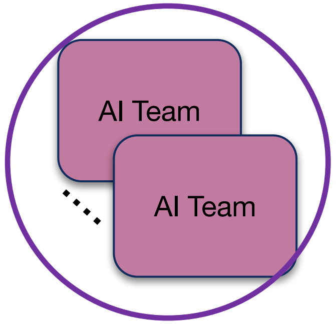
- Training
- Education
- Cross cutting projects

"Co-develop hardware and software, implementing a multipronged approach that includes, for example, machine learning, accelerators, mixed precision, and new algorithms."

NCAR Strategic Plan (2020-24), page 28

Accelerators & Compression
ASAP team

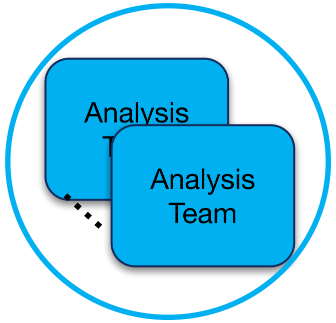




Machine Learning
AIML team

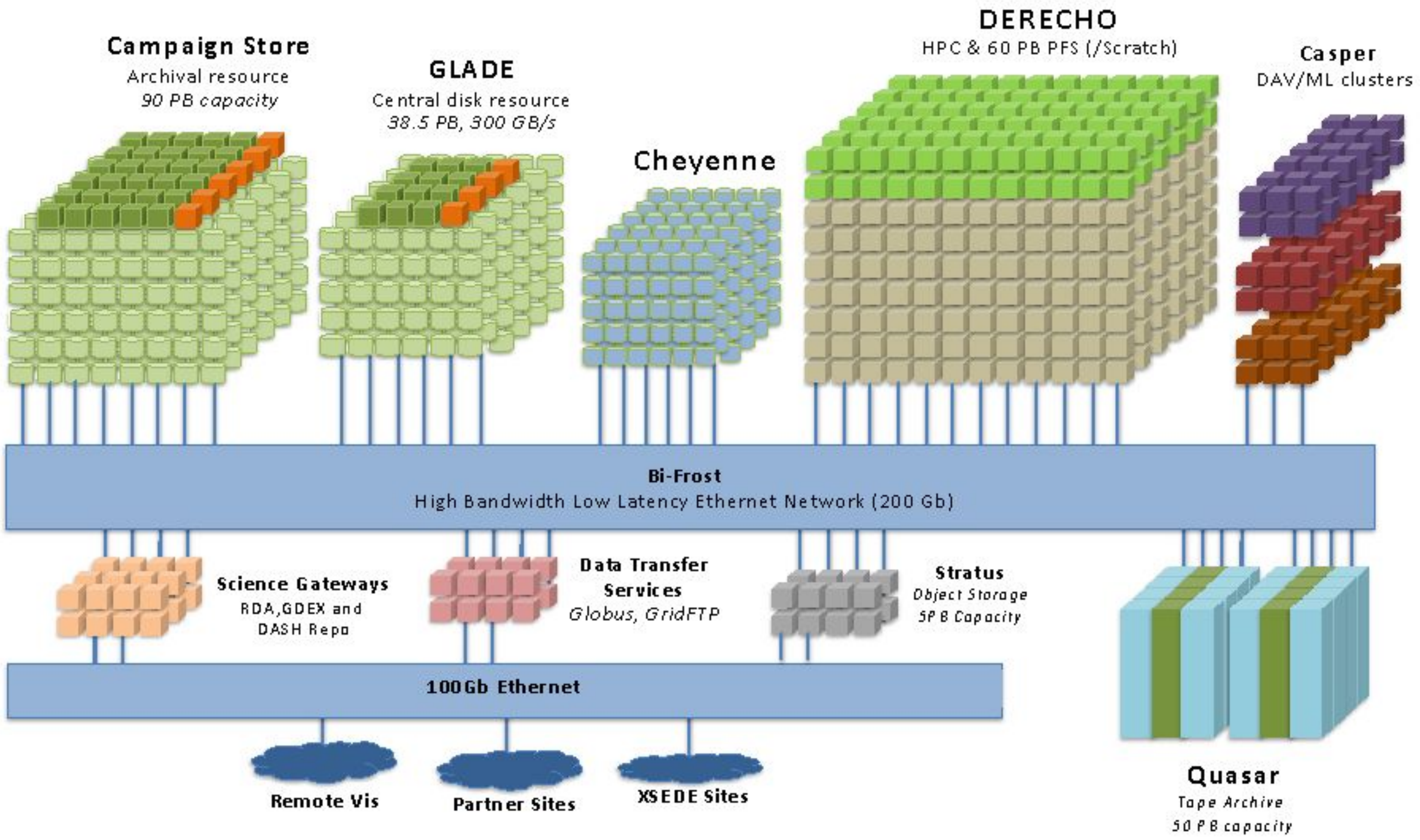
"… NCAR will also develop and apply techniques and approaches that draw on the latest data and computational innovations."

NCAR Strategic Plan (2020-24), page 28

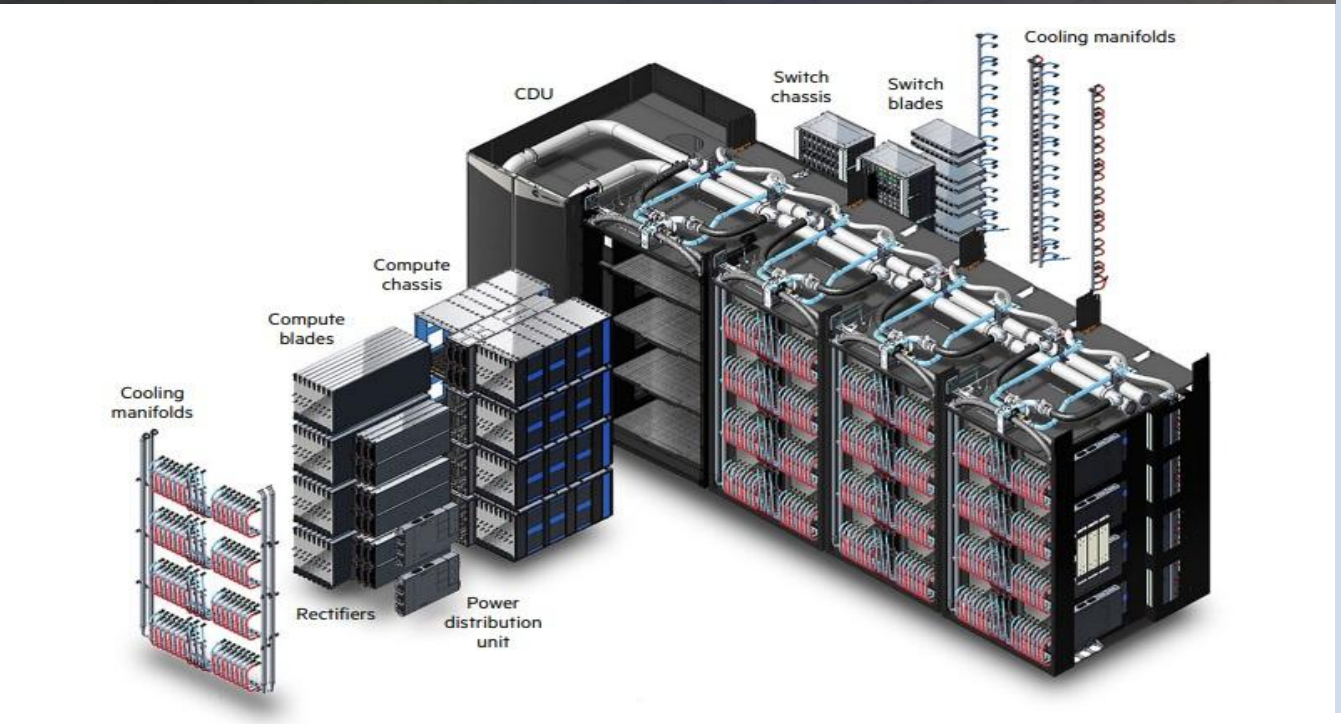


Data Analysis & Workflows
VAST

NCAR HPC ECOSYSTEM



DERECHO NCAR'S NEW SUPERCOMPUTER

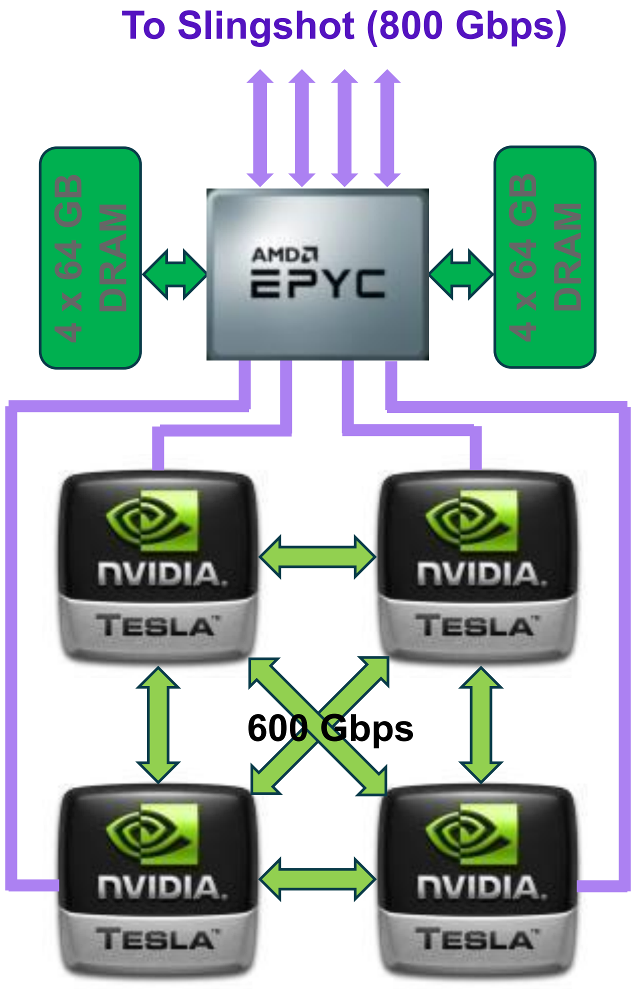


- Derecho (NWSC-3) HPE/Cray XE
  - Includes HPC and PFS
  - Peak: 19.87 PetaFLOPS
  - 60 PB usable file system
- 3.51-fold improvement over Cheyenne sustained Equivalent Performance (CSEP)
  - CPU – 2.84 CSEP ~80%
  - GPU – 0.67 CSEP ~20%
- Slingshot® Interconnect
- Fast scratch file system
  - Six HPE/Cray ClusterStor E1000 systems
  - 60 petabytes of usable file system space
  - 300 GB per second aggregate I/O bandwidth
  - 5,088 × 16-TB drives

Derecho's GPU Node

2 nodes per compute blade
64 blades per cabinet
128 nodes per cabinet

Source: HPE Cray



- 1 x 64 core AMD Milan Processors w
- 512 GB DDR4 DRAM

- 4 x NVIDIA Ampere GPU w
- 40 GB HBM2 memory

DERECHO: 80% CPU & 20% GPU

- *Derecho* 80/20 CPU/GPU Split or 80% = 2.4 CSEP and 20% = 0.6 CSEP
- Major increase in GPU capability at NCAR, majority of users run CPU-only code
  - ***major training, outreach, and support effort required!***
- **GPU Tiger Team, Derecho Application Readiness Team, Advanced Scientific Discovery teams**
  - *Developing expertise within user community for writing/converting code for GPUs, optimizing usage, and debugging issues*
  - *Porting guides (particularly for Fortran), when to use features (e.g., MPS), multi-node runtime optimization*
  - *Involving NCAR developers when appropriate*
  - *Optimizing software-stack configuration (e.g., best UCX config for GPU RDMA)*
  - *Exploring new capabilities (e.g. GPU Direct Storage)*
  - *Managing the hardware, software ecosystem, and user environment. (GPU software stack is rapidly evolving and requires active maintenance)*
  - *Maintain knowledge base/best practices, arrange trainings on GPU and AI/ML/DL on earth science problems, etc.*

Applications and Software for the new system

- Traditional codes - CESM, WRF
- Explore other codes and workflows that can be optimized on GPU
  - **MPAS-A-GPU, FastEddy, MuRAM, GPU WRF, ML/DL, EarthWorks project**

GPGPU Fortran, C/C++ coding and development

CUDA, OpenACC, OpenMP 5, MAGMA, GPU Direct MPI

TensorFlow, Keras, PyTorch, Horovod, & more...

Machine learning, deep learning, and artificial intelligence

Exploring the Extreme-scale Scientific Software Stack from the DOE Exascale Computing Program https://e4s-project.github.io/

Timeline and Training

Now - Early porting and development work on Casper and Frost

Fall 21 - **Accelerated Scientific Discovery (ASD) call for proposals**

- Targeted application work on NWSC-3 test machine

Q4 FY22 - Early/ASD users on NWSC-3

Q1 23 - NWSC-3 open to all users

Q4 22 - Cheyenne is decommissioned

- Workshops: February through August 2022
- Materials: https://github.com/NCAR/GPU_workshop
- Branch: CSG_tutorial
- Following topics are covered in the workshop
    - Introduction to GPU architecture, key concepts, and terminologies
    - CUDA programming model and coding examples
    - OpenACC programming model and coding examples
    - PCAST verification tool
    - NVIDIA Profiling tools
    - Multi GPU and multi node GPU programming (OpenACC + MPI)

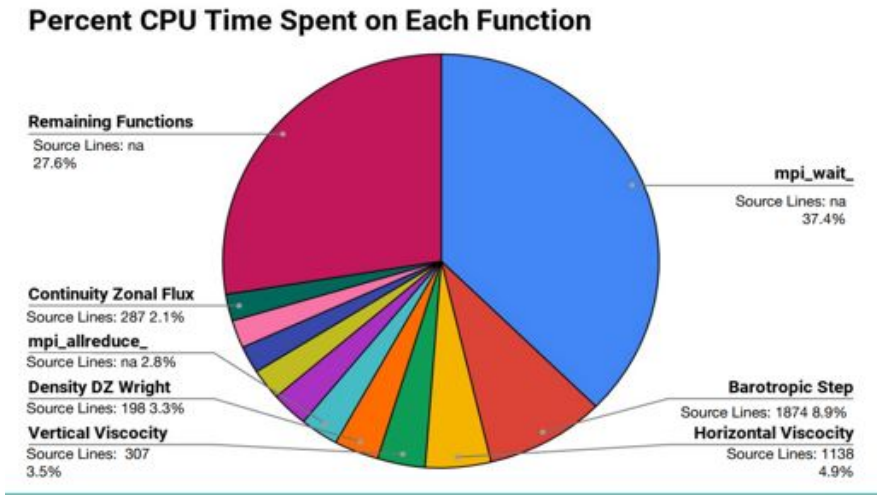
Refactoring for CPU/GPU Portability using Directives

- Use OpenACC standard directives to achieve performance portability
- Test driven development
- Profiling to prioritize refactoring targets
- Exploring
  - OpenMP
  - Kokkos or other approaches

```
!$acc parallel
!$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
```

```
VALIDATION RESULTS...

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

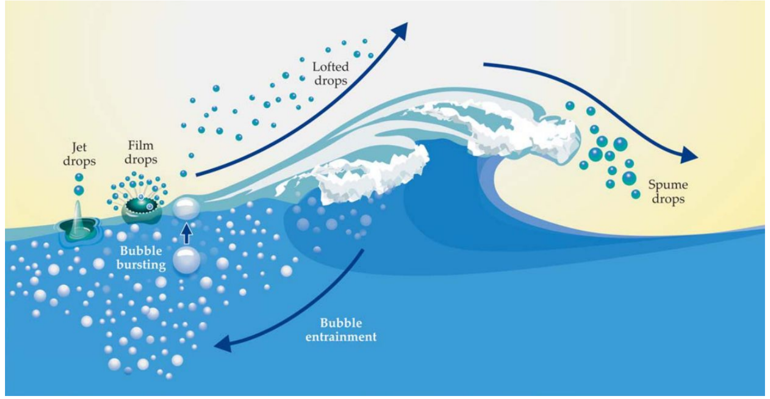


Advanced Scientific Discovery: Turbulence and Lagrangian transport in the hurricane boundary layer

John Dennis[1] (Co-PI), David Richter[2] (PI), George Bryan[1] (Co-PI), Sheri Mickelson[1] (Co-PI)
[1]NCAR, [2]University of Notre Dame

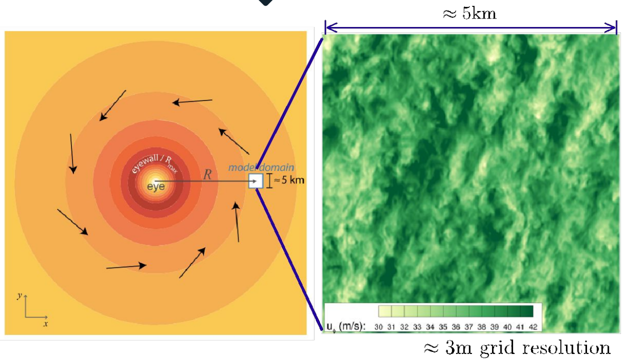
Turbulence and Lagrangian transport in the hurricane boundary layer

How do processes at the air-sea interface…



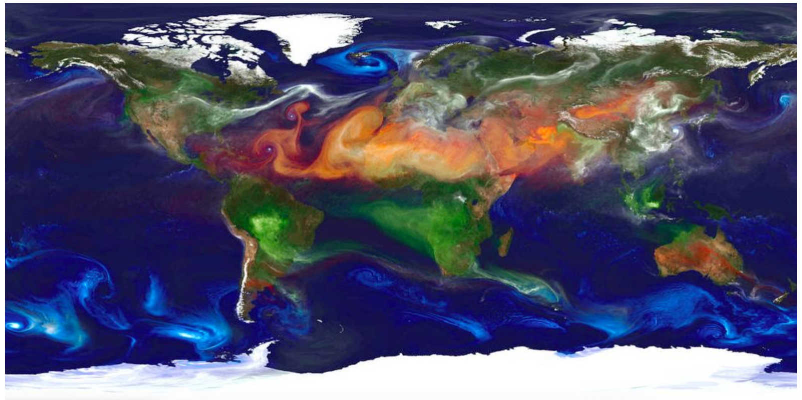
Richter & Veron, Physics Today, 2016

Computational model



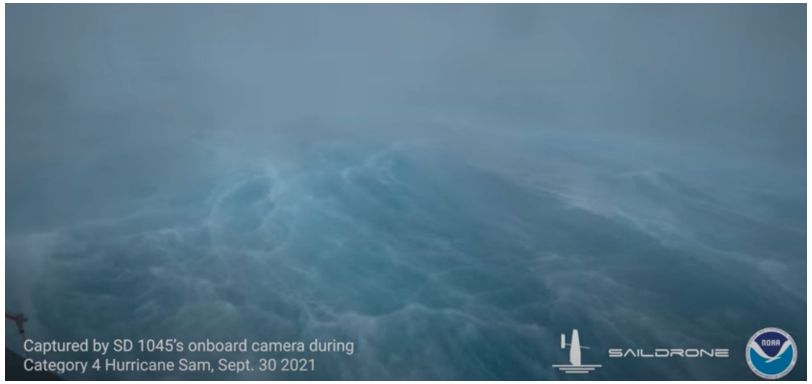
Observation

…affect large-scale transport and dynamics?



Climate?

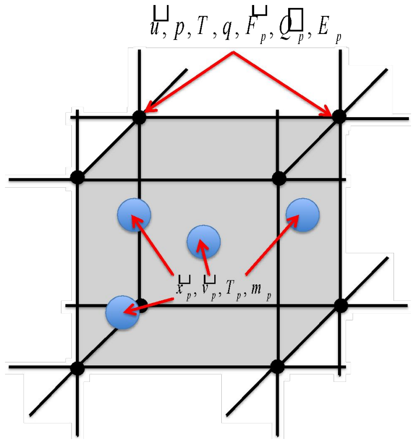
NASA GEOS-5; Blue=marine aerosols



Hurricanes?

Computational Plan: Turbulence and Lagrangian transport in hurricane boundary layer

Large-eddy simulation with Lagrangian droplets



"Lagrangian cloud model" (LCM) + sea spray origin →
Needs large number of droplets

- Utilize Cloud Model version 1 (CM1)
  - Fortran code: MPI + (OpenMP or **OpenACC**)
  - Augmented with Lagrangian transport capabilities
- Computational characteristics
  - Resolution: (2048 x 2048 x 1024)
  - Grid spacing (2.5m x 2.5m x 2.5m)
  - Droplets (10^5 – 10^9)
  - 90K GPU-hours
  - 72 TB of generated data
  - Performance comparison
    - 4 V100 NVIDIA GPUs
    - 4 Broadwell base CPU nodes
    - ~4.4x reduction in execution time V100 versus Broadwell node

How science objectives is steering the code development

Science Needs:

- How do coherent turbulent structures affect spray/droplet transport?
- Do droplets modify fluxes, temperature, humidity in the hurricane boundary layer?

Solutions:

- Need large eddy capability with Lagrangian cloud Model → Large per node/device problem size ideal for GPU computing
- Very large number of droplets to capture sea-spay at ocean boundary layer

Advanced Scientific Discovery:
Global Convection-Permitting Simulations with GPU-MPAS-A

Falko Judt, Andreas Prein, Bill Skamarock (MMM), Supreeth Suresh (CISL), Roy Rasmussen, Tim Schneider (RAL)

MPAS ASD Science

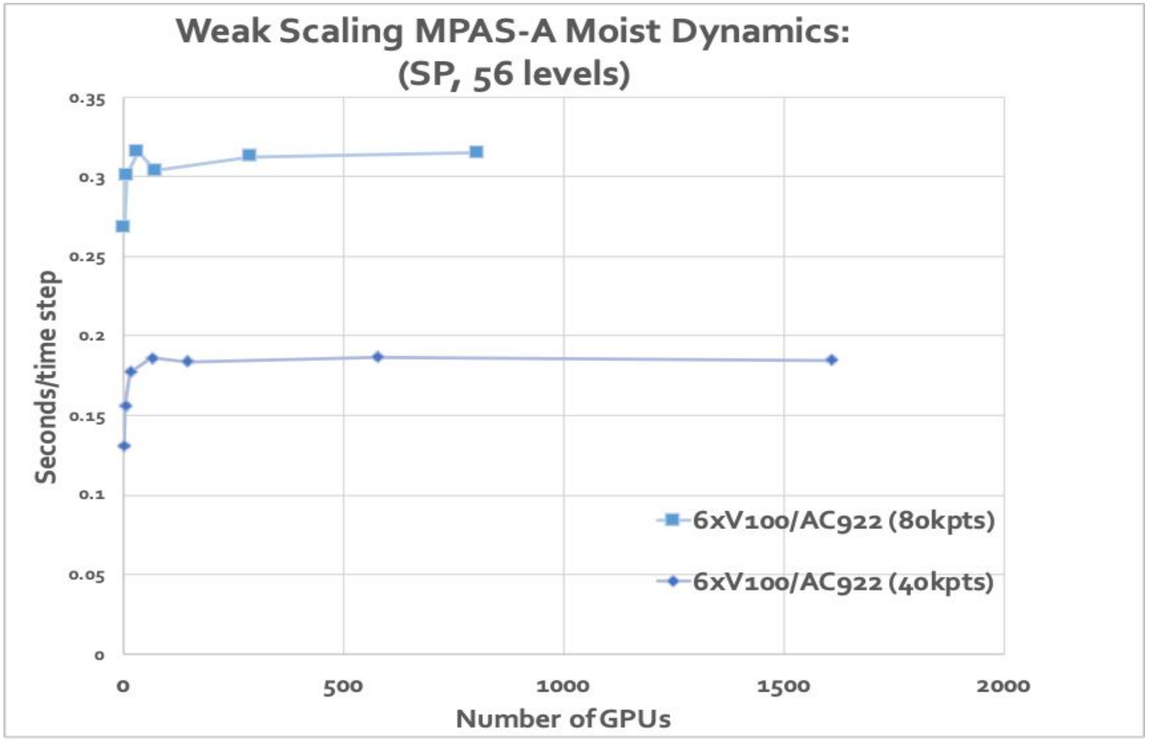
A series of global convection-permitting simulations using GPU-MPAS at 3.75km global resolution to better understand

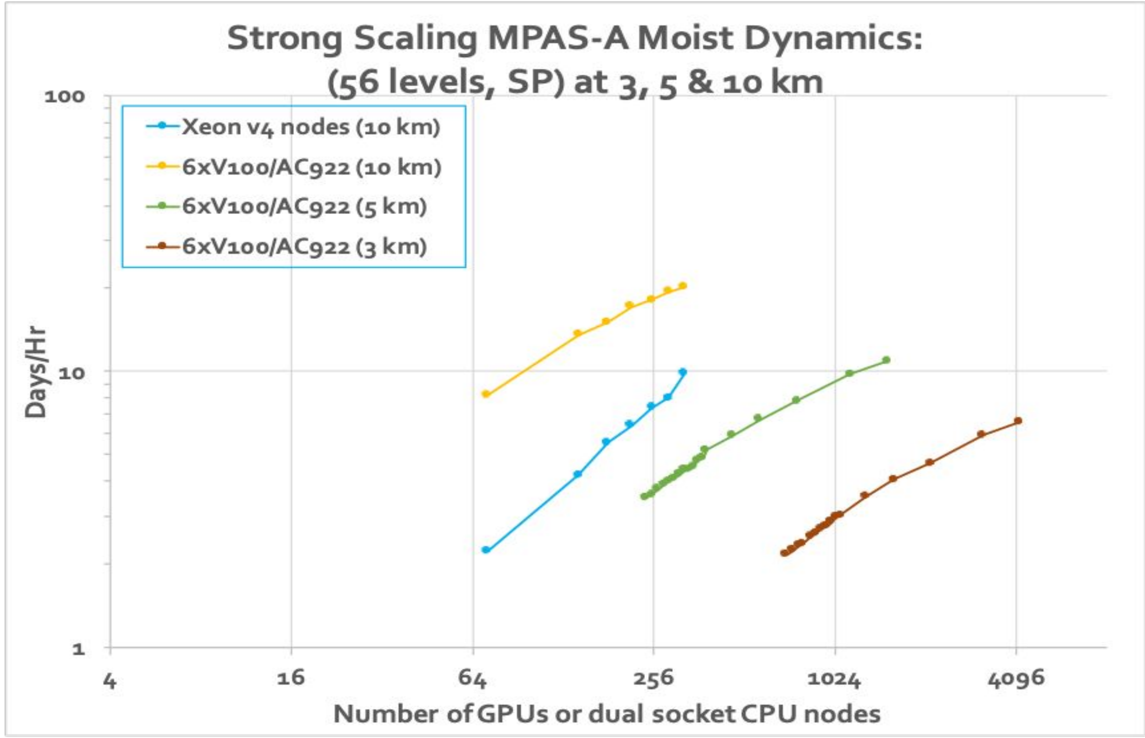
- the dynamics of tropical convection, and
- the predictability of the atmosphere in different climate zones.

Furthermore, we plan to assess the “added value” of convection-permitting resolution in

- simulating structure and life cycle of mesoscale convective systems across different climate zones,
- capturing the diurnal cycle and the duration, frequency, & intermittency of precipitation,
- predicting extreme weather from local to global scales, and
- representing orographic precipitation.

WEAK and STRONG Scaling of MPAS-A moist dynamical core on *Summit*[1] and STRONG scaling on *Cheyenne*[2] (dual, Intel 18c v4 Xeon)





[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
Slide by Rich Loft, NCAR

MPAS ASD Computational Plan

dx (km)	# of grid columns	timesteps /day	sim length (days)	Total time steps	GPU number	seconds per timestep	Total hours	GPU Hours	diag file size (TB)	number of diag files	total data
3.75	36,864,002	3840	320	1.23E+06	300	1.023725E+00	349.4	104829.4	3.000000E-02	1280	3.840E+01
15	2,621,442	960	320	3.07E+05	20	1.095724E+00	93.5	1870.0	1.875000E-03	1280	2.400E+00
30	655,362	480	320	1.54E+05	4	1.365720E+00	58.3	233.1	4.687500E-04	1280	6.000E-01
120	40,962	120	320	3.84E+04	4	1.001137E-01	1.1	4.3	2.929688E-05	1280	3.750E-02
				1.73E+06			TOTAL GPU-HOURS	106936.8	3.237305E-02	TOTAL DATA (TD)	4.200E+01

Estimated cost in GPU-hours for the simulations, with the number of GPUs for each resolution in parentheses

3.75-km runs: 105,000 (300)

15-km runs: 1,900 (20)

30-km runs: 233 (4)

120-km runs: 4 (4)

Storage: 420 TB (6-hourly output) + 20 TB (15-minute output) = 440 TB (disk space before data compression (after compression this number will reduce to ~150–200 TB)

How science objectivew are steering the code development for MPAS-A

Science Needs:

- Sub-seasonal forecasting
- Support for multiple physics modules to enable more accurate forecasting

Solutions:

- Earthworks: Integrating other earth system modules in a community model like Community Earth System Model (CESM)
- Porting, Optimization, and Integration of multiple physics modules for more accurate estimation of physical quantities on GPUs

Summary

- Refactoring of community codes to perform on current and future accelerated computing architectures
  - Directive based approach
  - Exploring other approaches
- Defining the science drivers that motivate the refactoring
- Good progress for main applications
  - Porting of physics modules needs to be done
- Initiating culture change to make GPU accelerated code a first class citizen at NCAR

Question?

- Credits for providing materials for the talk
  - Rich Loft
  - Irfan Elahi and the HPCD division for the Derecho slides
  - Cena Miller, TDD
  - Supreeth Madapur Suresh, TDD
  - John Dennis, TDD
  - Sheri Mickelson, TDD