EarthWorks in Texas:

Or…. Experiences Porting and Running a Coupled Climate Model System on Frontera and Grace-Grace and Grace-Hopper architectures

Presenter: Richard Loft[3]
PIs: David Randall[2], Jim Hurrel[2]
Sheri Voelz[1], Thomas Hauser[1], Michael Duda[1], Dylan Dickerson[1], Supreeth Suresh[1], Jian Sun[1], Chris Fisher[1], Donald Dazlich[2], Gunther Huebler[4], Jim Edwards[1], Brian Dobbins[1],
Raghu Raj Kumar[5], Pranay Reddy Kommera[5]

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

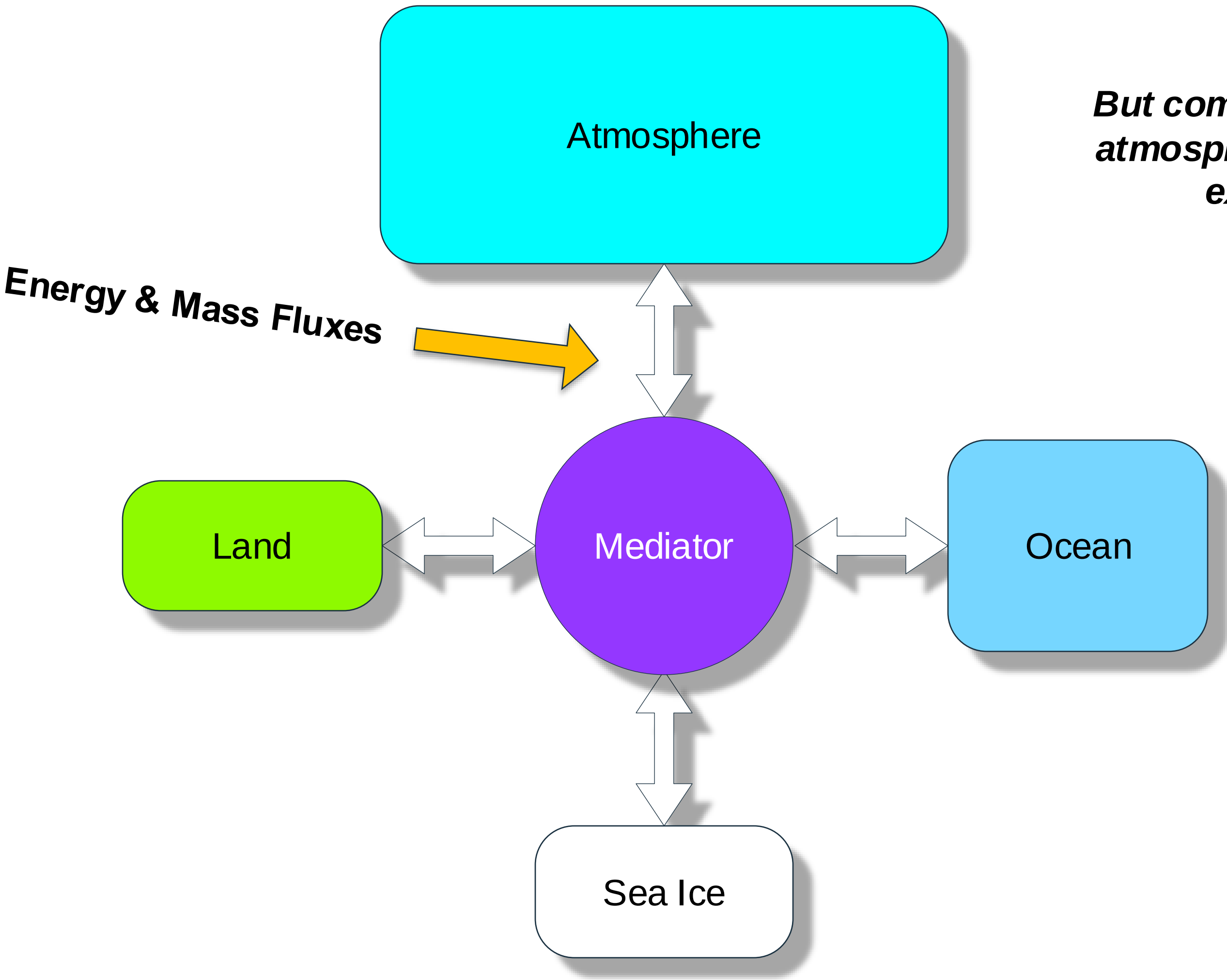
August 6, 2024

d by the National Center for Atmospheric Research, which is a major facility sponsoed by the National Science Foundation under Cooperative Agreement No. 1852977.

The Dream

The Computational Reality

The Earth System: It's more than the atmosphere!

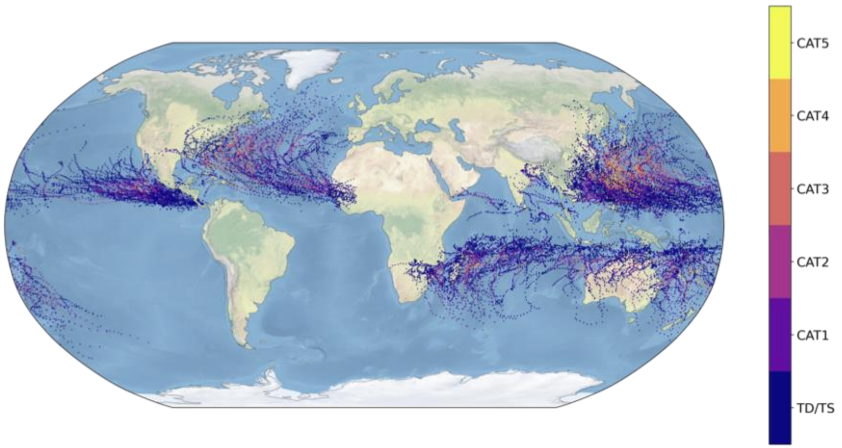


But computationally the atmosphere is the most expensive!

Earth System Models are used to study the statistics of climate observables

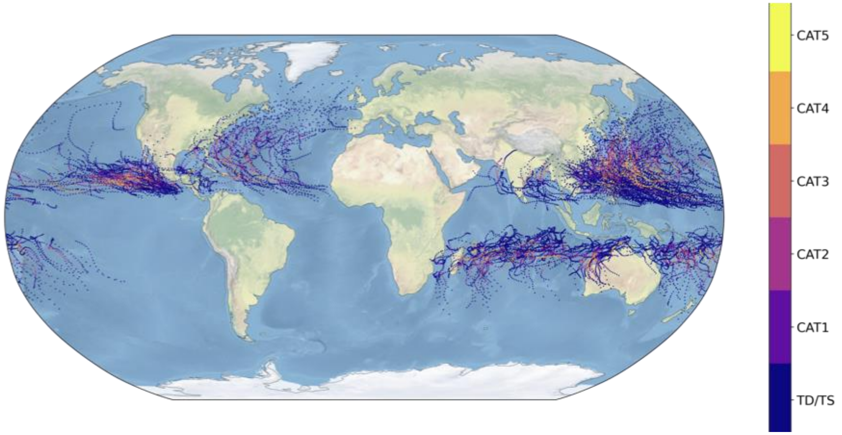
TC tracks

10-year simulation on 30-km grid using 1990-1999 SSTs



Results from CSU graduate student Andrew Feder

Observations for 1990-1999

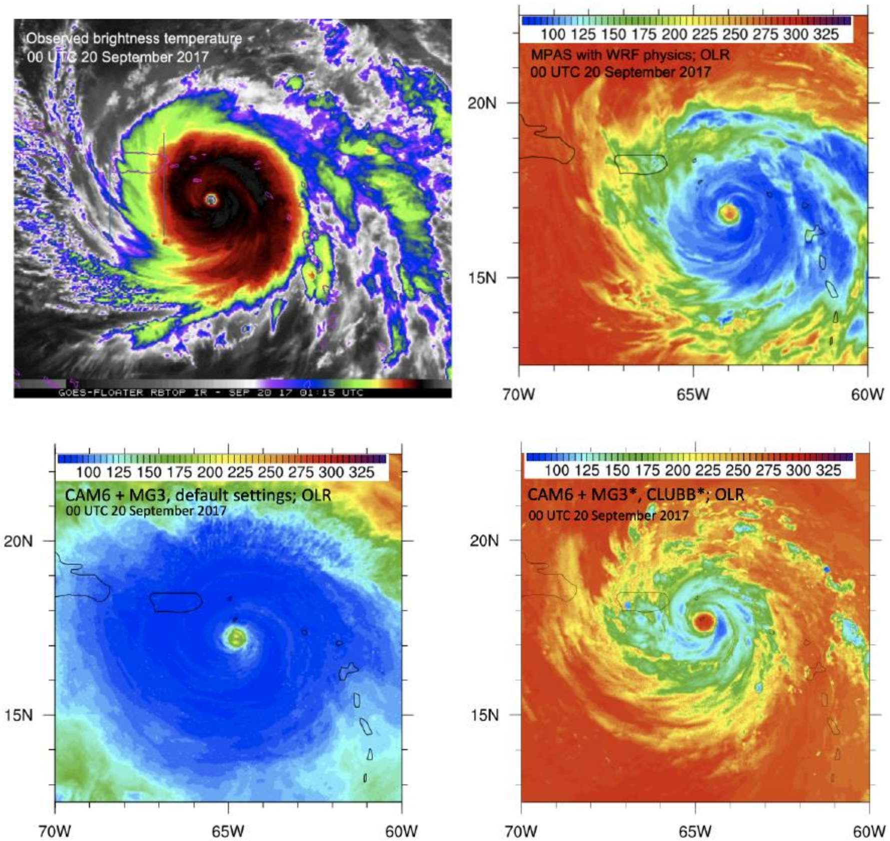


Statistical differences are called biases

Increased Resolution Alone Will Not Necessarily Fix Biases

Observed OLR from Hurricane Maria

Physics for Storm-Resolving Spatial Scales and Meteorological Timescales



Physics for Coarse Spatial Scales Applied to Storm Scales

Physics for Storm-Resolving Spatial Scales and Climate Timescales

Figure 2: Outgoing longwave radiation (OLR; brightness temperature) for hurricane Maria at 00 UTC 20 September 2017. Observed brightness temperature (upper left), and 2-day simulations with MPAS using WRF physics (upper right), MPAS using CAM6 physics with MG3 (lower left), and MPAS-CAM6 physics using a configuration of MG3 and CLUBB suitable for storm-resolving applications. The color scales for the observed brightness temperature and the OLR are not the same.

Figure Courtesy Bill Skamarock, NCAR

Atmospheric physics is spatially and temporally dependent.

About EarthWorks…

Earthworks

- 5-year NSF-funded project lead by Colorado State University
- Partners include: National Center for Atmospheric Research, NVIDIA, and more

Science Goals

- Resolve mesoscale storms, ocean eddies, mountains, large lakes and rivers.
- Eliminate deep convection and gravity-wave drag parameterizations.
- Resolve the stratosphere.
- Study the interactions of mesoscale phenomena with larger scales and with kilometer-scale terrain features, on time scales of days to years.

Computational Goals

- Leverage NSF's investment in the Community Earth System Model.
- EarthWorks's (ambitious) performance goals at 3.75 km resolution:
  - 0.5 simulated year per day in atmosphere-only simulations with a resolved stratosphere.
  - 1 simulated year per day in coupled simulations with fewer stratospheric layers.

EarthWorks: The discretized problem

EarthWorks uses the same quasi-uniform *geodesic mesh* for all components.

This arrangement has both computational and science advantages.

It's a BIG PROBLEM: At a mesh spacing of ~3.75 km has 40 Million horizontal grid points, each with ~100 atmospheric levels.

About the Model for Prediction Across Scales (MPAS)

"Fully compressible non-hydrostatic equations written in flux form"

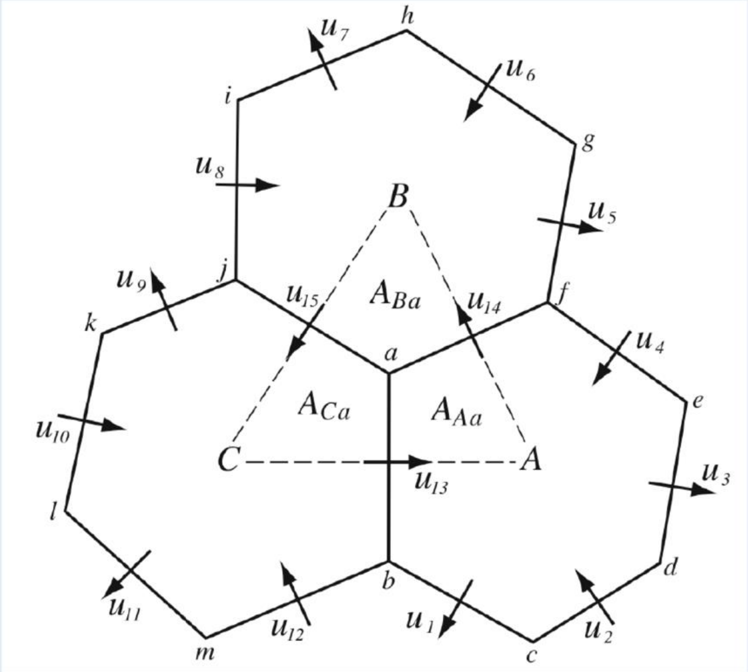
Finite Volume Method on staggered grid

- The horizontal momentum normal to the cell edge (u) is sits at the **cell edges**.
- Scalars sit at the **cell centers**

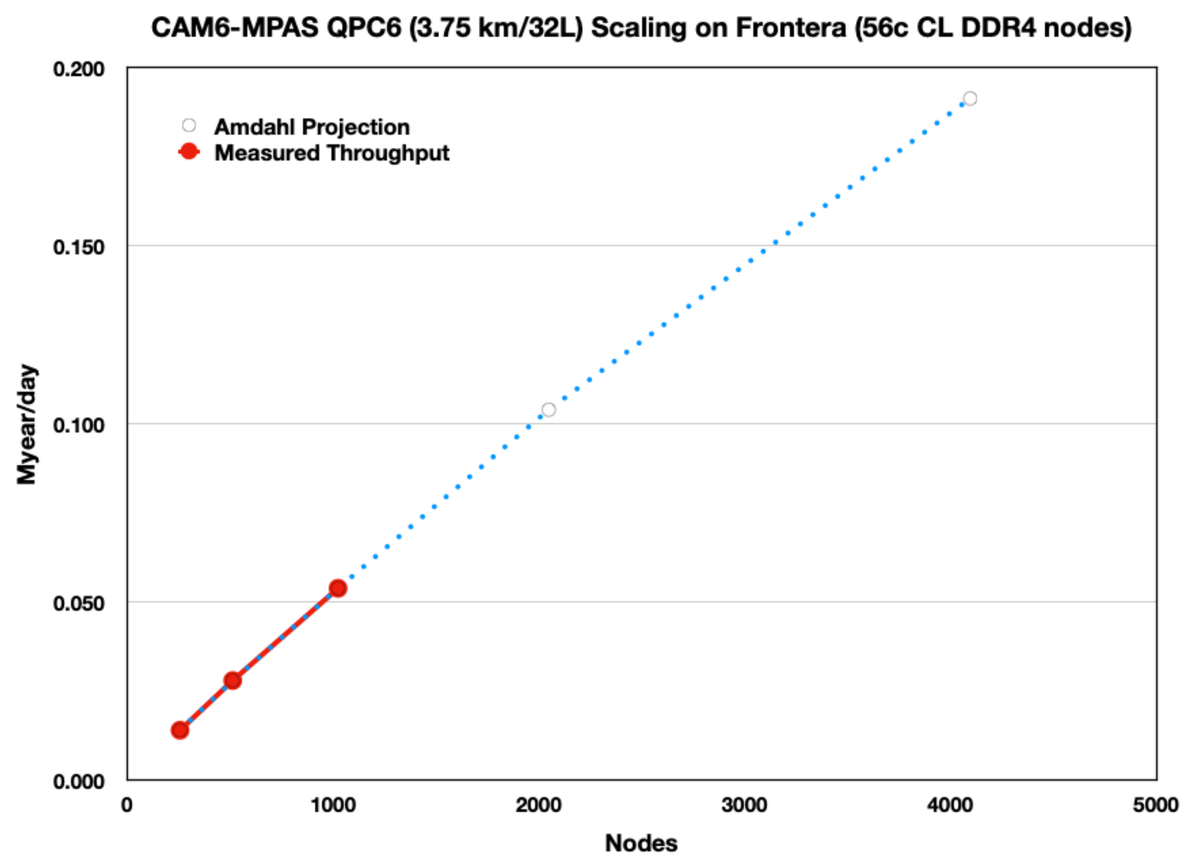
•Split-Explicit timestepping scheme

- Time integration 3rd order Runge-Kutta
- Fast horizontal waves are sub-cycled

MPAS is based on unstructured centroidal Voronoi (hexagonal) meshes using C-grid staggering and selective grid refinement.



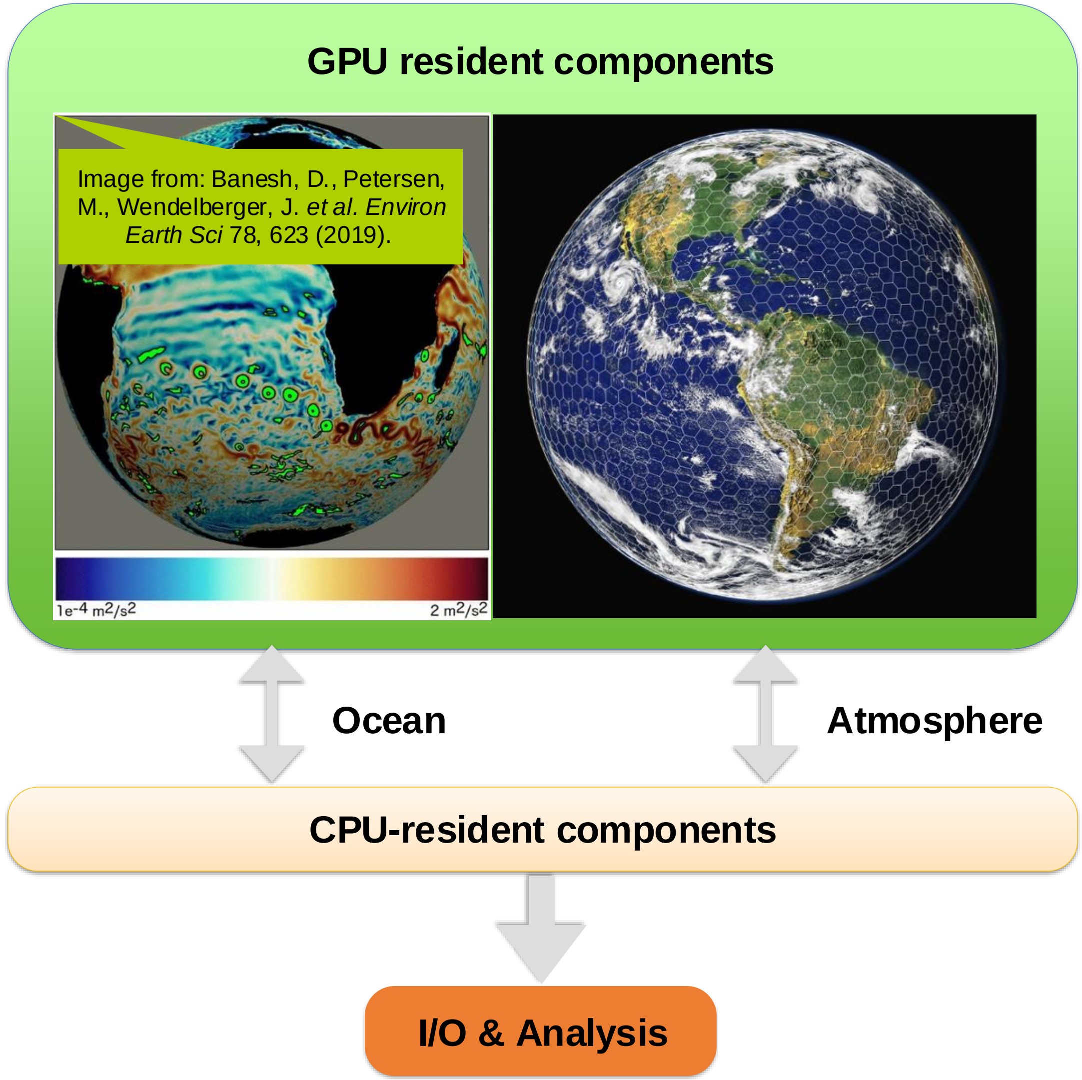
EarthWorks Frontera Scaling of Aquaplanet



- Measured (red/solid) and projected (blue/dotted) throughput of an Aquaplanet configuration at 3.75 km with 32 levels on the Intel CPU-based TACC Frontera system.
- Haven't scaled out further due to model infrastructure scaling/stability issues.

EarthWorks: The Computational Strategy

- **Architecture:** Hybrid CPU/GPU via offload directives (OpenACC ->OpenMP)
- **Parallelism:** Tune MPI rank count for optimal throughput across GPU-resident and CPU-resident components.
- **Precision:** Explore running (some) ESM components in FP32.
- **Big-data:** Leverage emerging high-resolution climate data analysis software ecosystem tools like **HealPix** and **Raijin**.

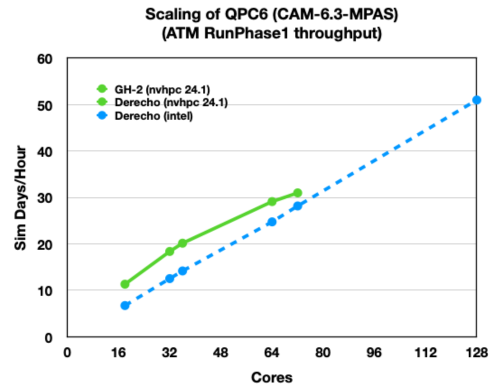


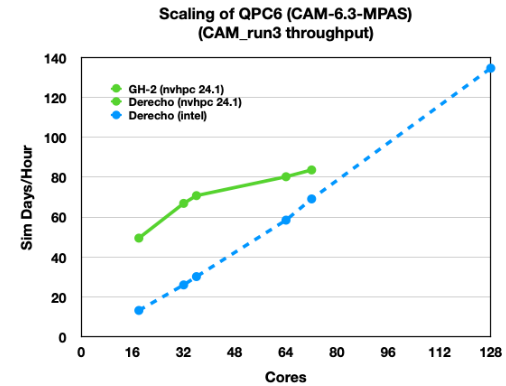
Computer Comparison Fast Facts

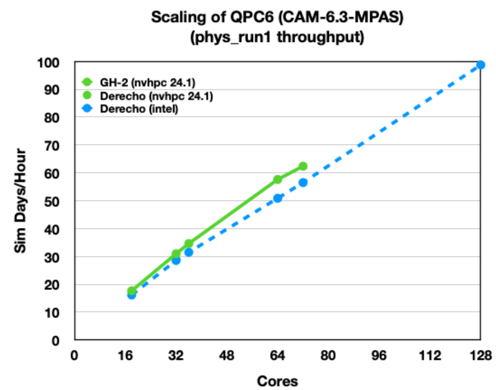
- **Derecho (NCAR):**
  - 2488 CPU nodes with 2x64c AMD Milan Procs
  - 82 GPU nodes with 1x64c Milan Proc + 4xNVIDIA A100 GPUs
- **GH2 (TACC):**
  - 1 GPU node with
    - 1x72c NVIDIA ARM “Grace” Proc
    - 1xNVIDIA H100 GPU
- **Vista (TACC)**
  - 64 CPU nodes with
    - 2x72c Grace Proc
  - 92 GPU nodes with
    - 1x72c Grace” Proc
    - 1xNVIDIA H100 GPU

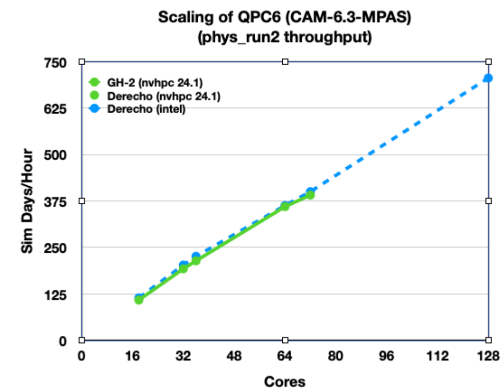
GH2/Derecho CPU Benchmarks: Aquaplanet

- **Dynamics** (upper right) is bandwidth intensive. Grace (green) is fast compared to Milan (blue), but scaling suffers beyond 36 cores.
- **Physics performance/scaling** (lower) is computationally intensive. Grace (green) roughly comparable core-to-core to Milan (blue).
- **Full QPC6 atmosphere** (upper left) is a mixture of these computational characteristics.





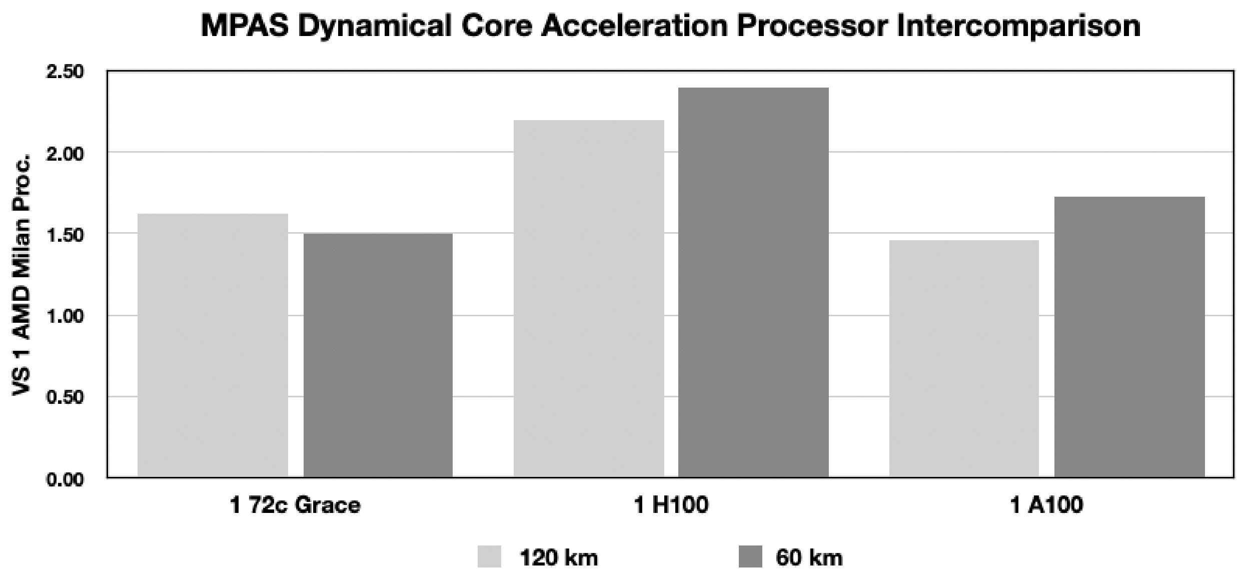




Sub-phases shown account for 94% of the ATM run time

GH2/Derecho CPU Benchmarks: MPAS Dynamics

- A "socket to socket" comparison of a Grace processor, H100, and A100 vs a single Milan Proc.
- CAM-MPAS dynamical core on a quasi-uniform global grid with 32 levels is offloaded with OpenACC directives.
- All GPU experiments were run with a single host rank offloading to the GPU. (No MPS)



- **Notable features:**
  - Grace is slightly faster core for core than Milan.
  - H100 is about 1.5 faster than A100 on the MPAS dynamical core workload.
  - The ratio increase seen for GPUs between 120 km (40K columns) and 60 km (160 columns) could be attributable to either CPU cache or GPU occupancy (data parallelism) effects.

Conclusions and Next Steps:

- Single H100 results are encouraging but need Multi-A100/H-100 benchmarks.
- Integrated GPU-dycore + GPU-physics testing will push us to multi-ranks per device. Where's the sweet spot?
- Multi-GPU, multi-node results at higher resolutions coming soon.

Thanks!

And a special thanks to TACC for the use of TACC resources, especially during Texascale Day runs, and to John Cazes and the User Support Staff for the ongoing help with our work on Frontera, GH2, and Vista!