

National Aeronautics and Space Administration

HECC Application Performance & Productivity (APP)

Gary Kedziora
HPC Application Performance Analyst
Gary.S.Kedziora@nasa.gov

May 22, 2024

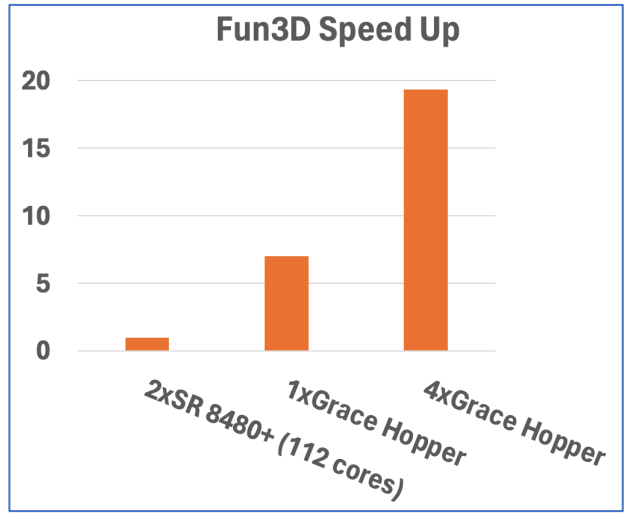


NASA High-End Computing

The APP group's mission

Help with application-related issues, including:

- Porting your code and workflow to HECC supercomputers.
- Support porting code to run on GPUs.
  - Primary way to achieve the highest performance.
- Optimize the application build process.
- Find and fix bugs, create workarounds.
- Simplify or automate your workflow.
- Basic performance analysis and optimization.
  - Write scripts to parallelize multiple copies of serial jobs.
  - Reduce memory requirements, improve cache usage.
  - Basic parallelization.
  - Optimize I/O.
  - Let us know if you have any issues with these, especially performance optimization.



waverider-5M with chemistry example



Example: APP Team helped increase LAVA developer productivity

- Tools for installing LAVA dependencies using Spack were developed:
  - Spack is a source-code package manger developed by DOE Exascale Computing Project (ECP).
  - Scripts automatically create Spack environments for dependencies on HECC systems including the GitLab server.
  - Automatically installs LAVA with dependencies or provides a development environment for interactive software development.
- Supported porting to GPUs and Cabeus:
  - Provided recipes for building CUDA versions of LAVA with different versions of supporting software stacks.
  - Enabled GPU code optimization.
  - Supported profiling using NVIDIA tools.

Impact: APP efforts increase the turnaround time for building LAVA and its dependencies, supporting continuous integration testing and developer workflow.

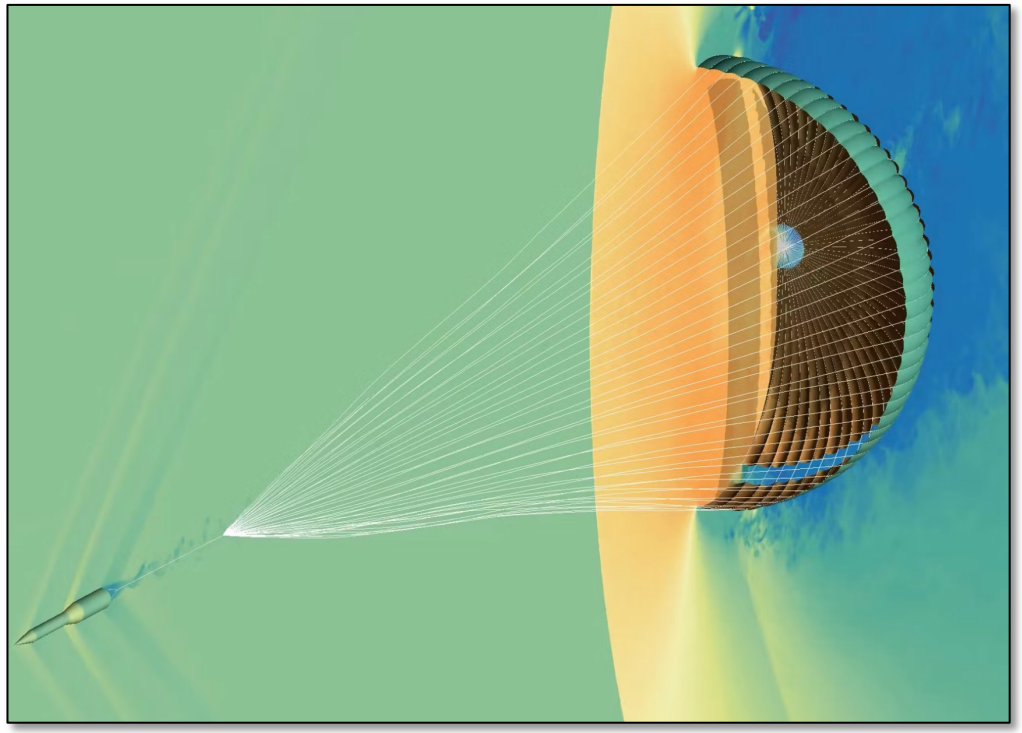


Image capture from LAVA simulation of a supersonic parachute. (Francois Cadieux and Michael Barad, NASA Ames Research Center. Simulation run on NAS Rome nodes)

is licensed under
CC BY-NC

APP command line tools

- APP maintains in-house developed tools for applications and job scheduling.
  - Get information on job startup, running jobs.
  - Information about resource availability.
  - Monitor memory and power usage.
  - Commonly used commands:
    - **`check_licenses`**: availability of licenses for 3rd party tools
    - **`node_stats`**: overview node usage in PBS
    - **`qs`**: an abstract view of the PBS queues
    - **`qsh.pl`**: run commands on all nodes of PBS job
    - **`qtop.pl`**: run top across all nodes inside of PBS job
    - **`gm.x`**: memory allocated per rank inside a PBS job
    - **`mbind.x`**: bind processes to specific cores on a node
    - **`pdsh_gdb`**: obtain stacktrace of running job
    - **`mpiprof`**: MPI & IO performance monitoring tool
- How to access these tools:
  - `module load scicon/app-tools-1.2.1`
  - `module load scicon/mpiprof`
  - `module load scicon/cli_tools`

```
GraceHopper — gkedzior@pfe20:~ — ssh ◂ ssh pfe — 79×19
gkedzior@pfe20 $ module show scicon/app-tools-1.2.1
-------------------------------------------------------------------
/nasa/modulefiles/toss4/scicon/app-tools-1.2.1:

conflict        scicon_cli_tools
module-whatis   {Adjust environment for the mbind/gmx tools}
prepend-path    PATH /u/scicon/tools/opt/toss3/app-tools-1.2.1/bin
prepend-path    CPATH /u/scicon/tools/opt/toss3/app-tools-1.2.1/include
prepend-path    FPATH /u/scicon/tools/opt/toss3/app-tools-1.2.1/include
prepend-path    LIBRARY_PATH /u/scicon/tools/opt/toss3/app-tools-1.2.1/lib
prepend-path    LD_LIBRARY_PATH /u/scicon/tools/opt/toss3/app-tools-1.2.1/lib
append-path     PATH /u/scicon/tools/bin/toss3
-------------------------------------------------------------------
gkedzior@pfe20 $ ls /u/scicon/tools/opt/toss3/app-tools-1.2.1/bin
clist.x  gm_post.x  gm.x  gpu_info  mbind.x
gkedzior@pfe20 $ mbind.x -h
Usage
   mbind.x [-options] [-h|-he|-hm] prog [args]
```



Getting APP Support

APP uses a tiered support model:

- Tier 0 Support: Enable users to help themselves.
  - Online documentation: https://www.nas.nasa.gov/hecc/support/kb
  - Webinars: https://www.nas.nasa.gov/hecc/support/training.h.tml
  - Customized optimization training for groups
    - E.g., for several groups at Langley, we analyzed their codes and did some optimizations, then presented the methods and explained why they worked.
  - We welcome suggestions for all Tier 0 Support.
- Tier 1
  - Initial triage.
  - Solutions to routine issues (Control Room).
- Tier 2 APP Support
  - Ticket driven (support@nas.nasa.gov).
- Tier 3 APP Support
  - More complicated tickets.
  - Proactive identification and handling of issues.
  - External requests, especially performance analysis and optimization.



https://www.nas.nasa.gov/hecc/

Example: APP Team improves efficiency of magnetohydrodynamics simulations

- Analyzed and suggested several improvements to a heliophysics workflow for a researcher at the Laboratory for Atmospheric and Space Physics, University of Colorado Boulder.
  - 4096 Python process requiring large amounts of memory were not distributed optimally according to memory access patterns.
  - Discussed finding a balance between packing more processes per node to decrease the overall node count or leaving more cores idle to speed up processes through reductions in cache and memory contention.

IMPACT: A key benefit of this collaboration is the transfer of knowledge on how to use various NAS-developed tools to empower researchers to directly improve their utilization of NAS supercomputers.

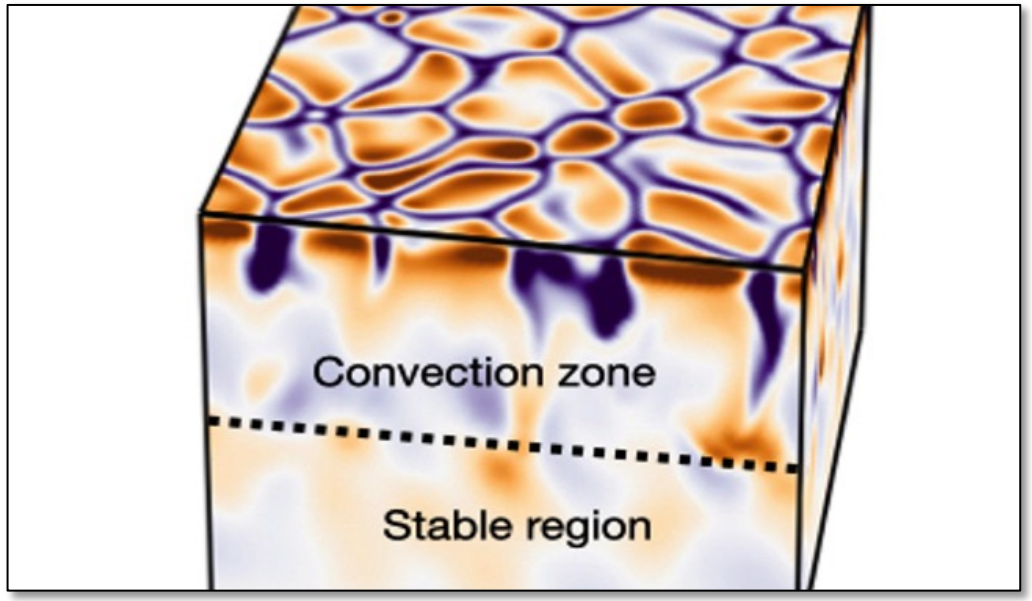


Image showing entropy perturbations in a 3D hydrodynamics simulation of turbulent convection driven from the top boundary. The initially stable background stratification gets well-mixed to an adiabat forming a self-consistent convection zone and a stable radiative zone.
Bhishek Manek, Laboratory for Atmospheric and Space Physics, University of Colorado Boulder

Example: APP team improves efficiency of magnetohydrodynamics simulations

- Load-balancing was achieved with the **mbind.x** tool, spreading the processes evenly across all sockets, reducing both cache and memory bandwidth contention.
- For a two-hour run, the number of iterations attained improved from 3,290 to 5,575 iterations, or almost a **70% improvement.**
- Further improvements in job efficiencies may be possible by using **mbind.x** and the in-house **qtop.pl** tool to measure memory use.

IMPACT: A key benefit of this collaboration is the transfer of knowledge on how to use various NAS-developed tools to empower researchers to directly improve their utilization of NAS supercomputers.

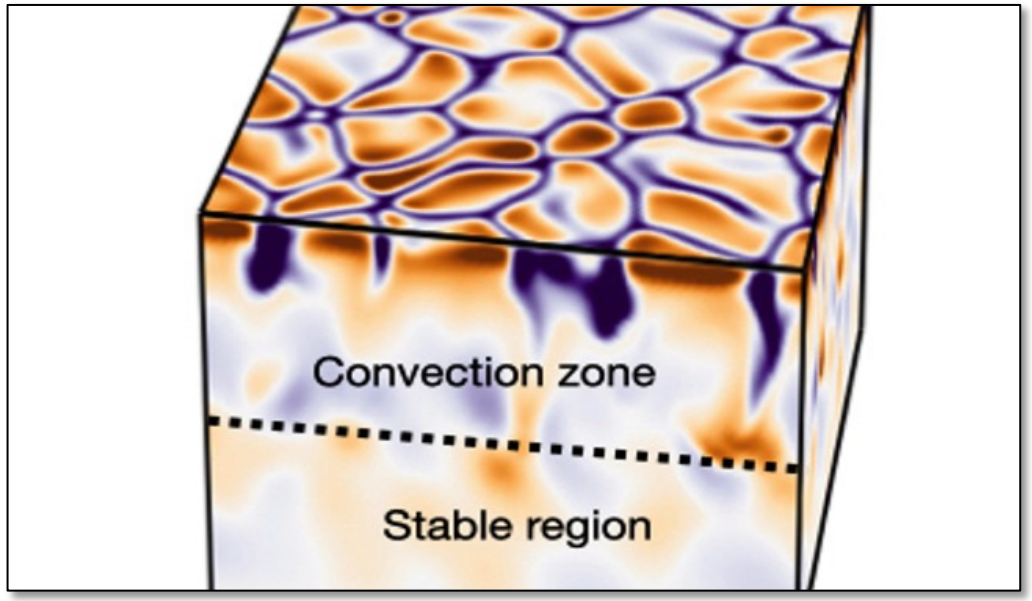


Image showing entropy perturbations in a 3D hydrodynamics simulation of turbulent convection driven from the top boundary. The initially stable background stratification gets well-mixed to an adiabat forming a self-consistent convection zone and a stable radiative zone.
Bhishek Manek, Laboratory for Atmospheric and Space Physics, University of Colorado Boulder

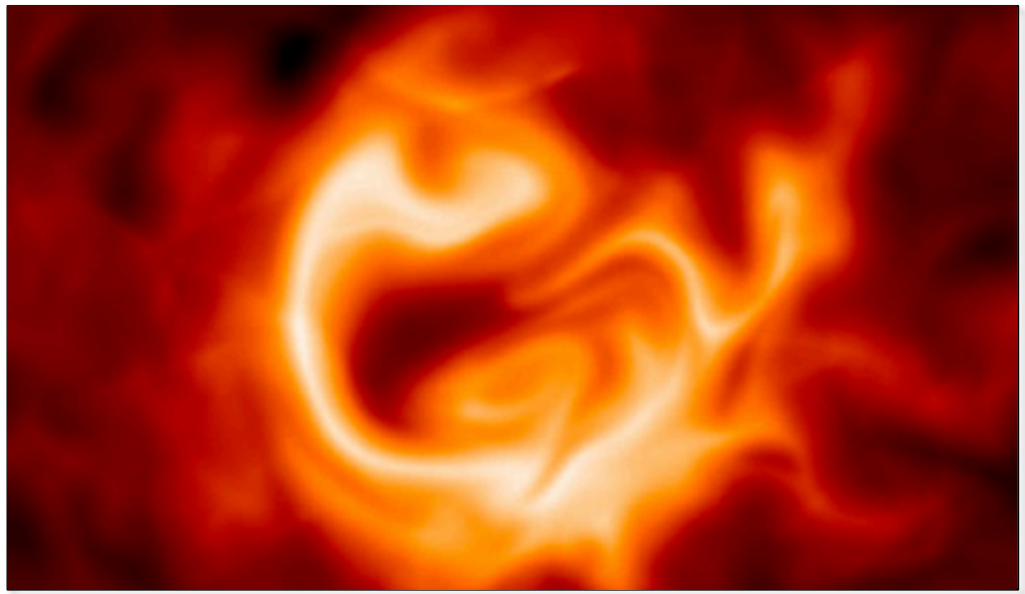
Success Story: APP team speeds up StellarBox I/O

- Significant speed up of StellarBox was obtained by building and running with OpenMPI 4.1.3, which uses the OMPIO collective MPI-IO algorithms.

Cluster	Speed up
Cascade Lake Aitken	4.6
Rome Aitken	9.3
Sky Lake Electra	11.4

- NAS workloads typically use the vendor supplied MPT version of MPI.
  - Superior for many MPI-based applications.
  - Uses slower ROMIO for collective I/O.
- OMPIO algorithm for collective I/O routine MPI_File_write_ordered in OpenMPI is significantly faster.

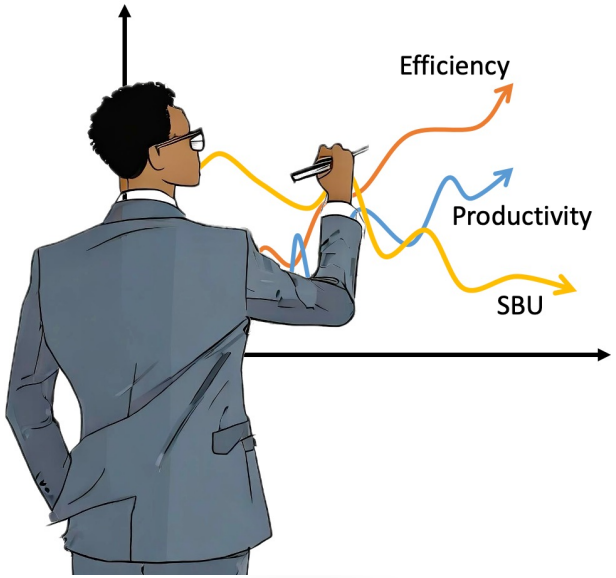
IMPACT: StellarBox is one of the most used codes at NAS. It was developed for realistic modeling of solar and stellar magnetoconvection and atmosphere dynamics. The StellarBox code speed-up significantly increases simulation productivity and provides more efficient system use.



The distribution of the radial magnetic field in the solar corona at a height of 10,000km above the surface from a 3D radiative MHD model obtained with StellarBox code (Wray et al., 2018). The brighter colors mean a stronger magnetic field. Credit: I. Kitiashvili (NASA ARC).

Other APP activities

- We can help expedite your jobs!
  - This is especially true if you've been affected by system issues.
  - Let us know if you have a hard deadline and you are running out of time (with manager approval).
- We benchmark current computers and potential future systems.
  - We set SBU rates for the node types.
  - We evaluate new system ("Pathfinder" effort).
- We monitor energy usage by jobs, finding two types of load imbalance:
  - When nodes of a job use more cores in socket 0 than in socket 1.
  - When not all nodes of a job are used—perhaps due to incorrect "-np" option to **mpiexec**.



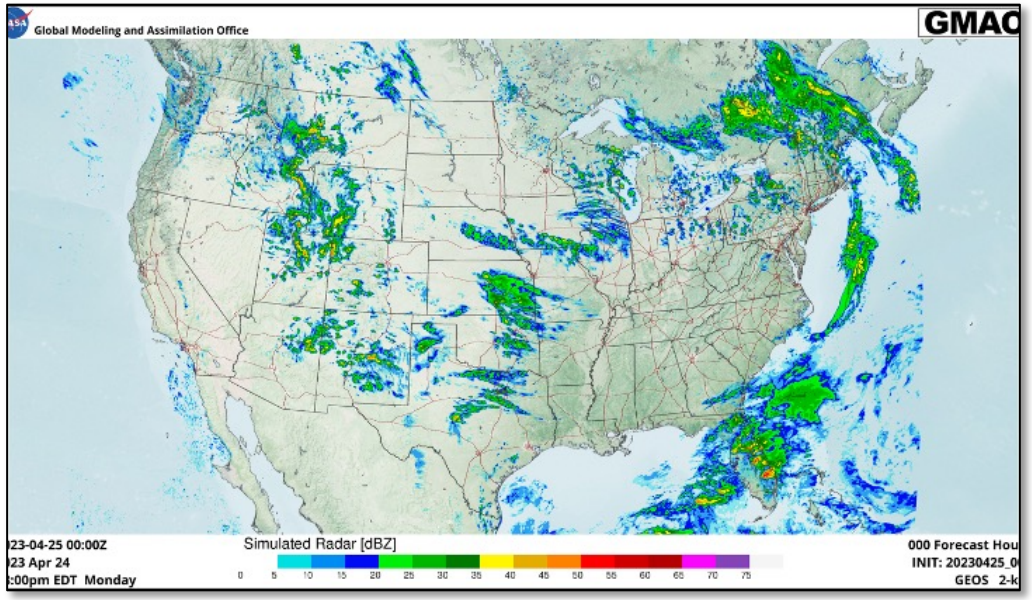


Example: APP provided concierge support for hazardous spring forecast experiment

- APP provided debugging and assistance with running GEOS workflows:
  - Increased reliability of the workflow.
  - Increased throughput of important simulation jobs through the batch queue system.
  - Required reservations of 420 Rome nodes over a 13-week period.
- 2023 Hazardous Spring Forecast Experiment:
  - Run by NASA's Global Modeling and Assimilation Office (GMAO).
  - Focuses on severe weather over the continental US.
  - Tests new simulation and prediction tools.

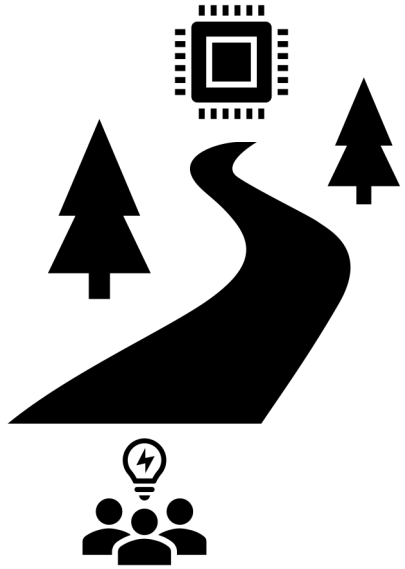
https://www.nas.nasa.gov/hecc/about/hecc_user_success.html

APP's efforts supported a successful completion of the Spring Forecast Experiment by GMAO testing cutting edge simulation tools for predicting hazardous weather, which are used to develop new weather forecasting products.



The Hazardous Spring Forecast Experiment is a comprehensive evaluation of multiple contributing models including the GEOS model, with a focus on severe weather over the continental U.S., representation of simulated reflectivity, updraft, and rotation properties, and environmental variables. William Putman, NASA Goddard

Pathfinder system evaluation effort



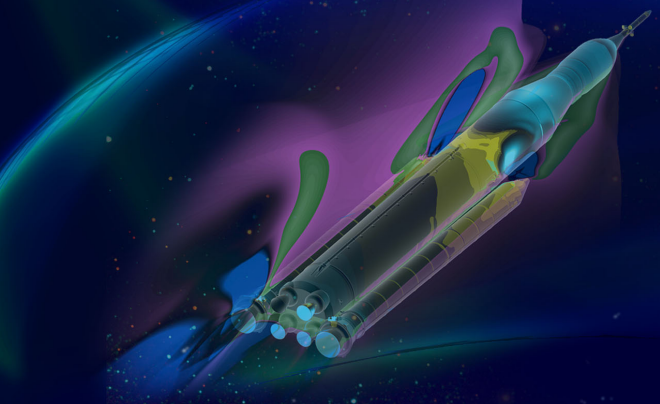
- APP Group evaluates new technology for consideration of future purchases by conducting:
  - Benchmarking
  - Porting NASA Codes
- Current or recent systems under evaluation:
  - IBM Power 10
  - NVIDIA Grace Hopper and Grace Grace
  - Intel Sapphire Rapids and Xeon CPU Max with HBM
- Let us know if you would like to test your codes on these systems!

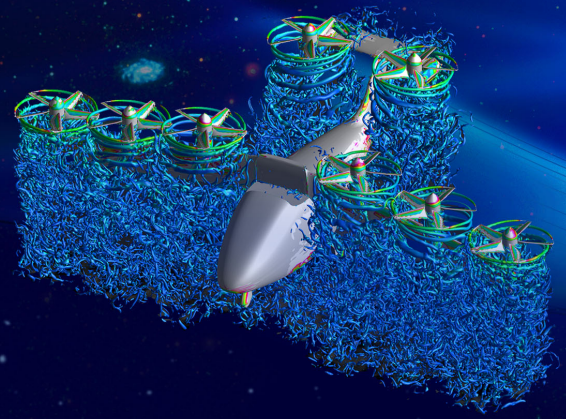


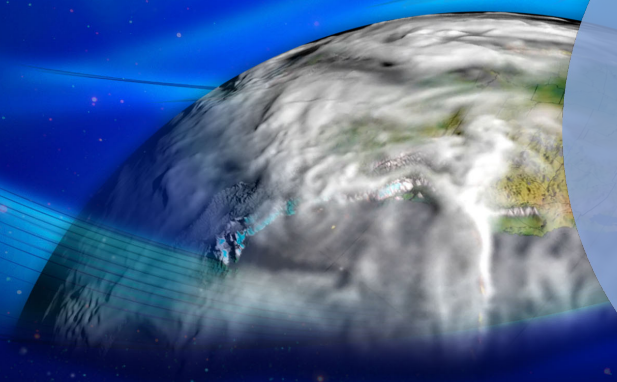
National Aeronautics and Space Administration

Ideas?
Questions?
Suggestions?

We want to hear from you!
support@nas.nasa.gov
1 (800)-331-8737









backup slides and drafts

I/O Speed-Up Details

- MPT 2.28 uses only the ROMIO algorithm for collective writes (MPI_File_write_ordered).
- OpenMPI 4.x uses the OMPIO algorithm by default and has options for various algorithms including ROMIO.
  - Several variants of OMPIO were tried but the default parameters were fastest.
  - A systematic exploration may provide better results.
- Parallel IO takes over 80% of the run time using MPT and around 20% of the time using OpenMPI.
  - This frees up disk access usage for the /nobackup Lustre filesystem, providing faster access times for others.

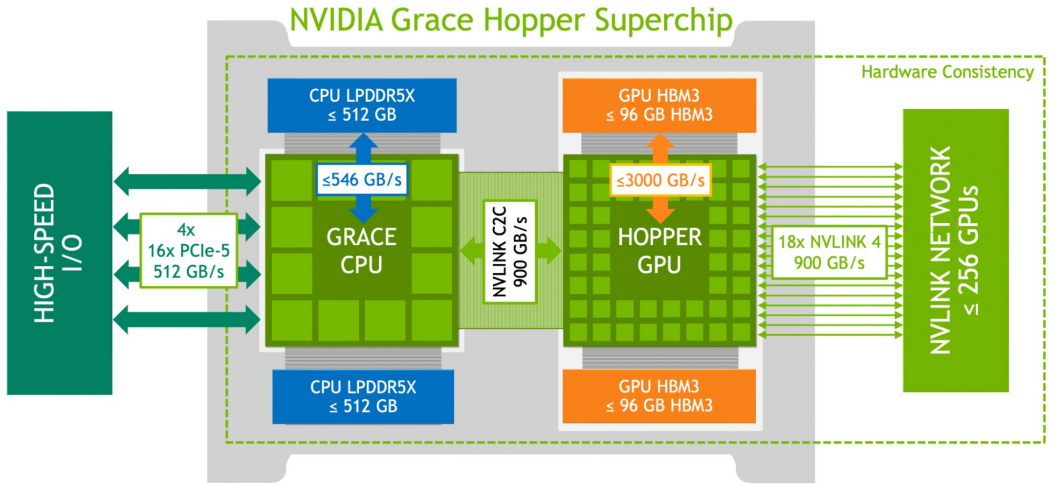
Cluster	MPI	Total	Comm.	IO	Speed up
Cascade Lake Atikin	MPT	2468	289	2133	
	OMPI	535	372	120	4.6
Rome Aitken	MPT	7281	NA	6003	
	OMPI	786	595	152	9.3
Sky Lake Electra	MPT	6132	300	5786	
	OMPI	538	376	119	11.4

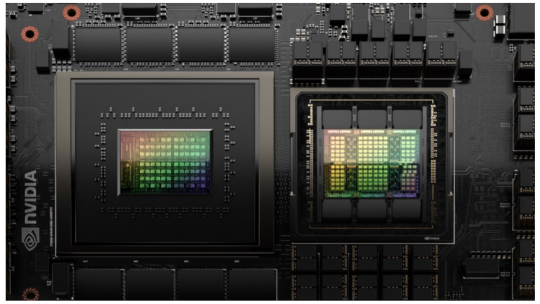
Total wall clock time, communication time, and MPI-IO time in seconds for 1000 iterations of a StellarBox example run on various NAS platforms using two different versions of MPI: MPT 2.28 and OpenMPI 4.3.1

OMPIO References: Chaarawi, M. & Gabriel, E. Automatically Selecting the Number of Aggregators for Collective I/O Operations. *2011 IEEE Int. Conf. Clust. Comput.* **1**, 428–437 (2011). Chaarawi, M. *et al.* Recent Advances in the Message Passing Interface, 18th European MPI Users' Group Meeting, EuroMPI 2011, Santorini, Greece, September 18-21, 2011. Proceedings. *Lect. Notes Comput. Sci.* 81–89 (2011) doi:10.1007/978-3-642-24449-0_11.

Pathfinder: Grace Hopper CPU

- Purchase of a Grace Hopper system is currently under consideration.
  - Grace ARM CPU with Hopper GPU
  - HBM3 and HBM3e GPU memory
  - NVLink chip to chip connection 900 (GB/s) interface
- Benchmarking
  - E.g., stream triad memory b.w., OSU MPI
  - NAS Parallel Benchmarks (NPB)
  - SBU1 (GPU version)
- Testing with NASA codes

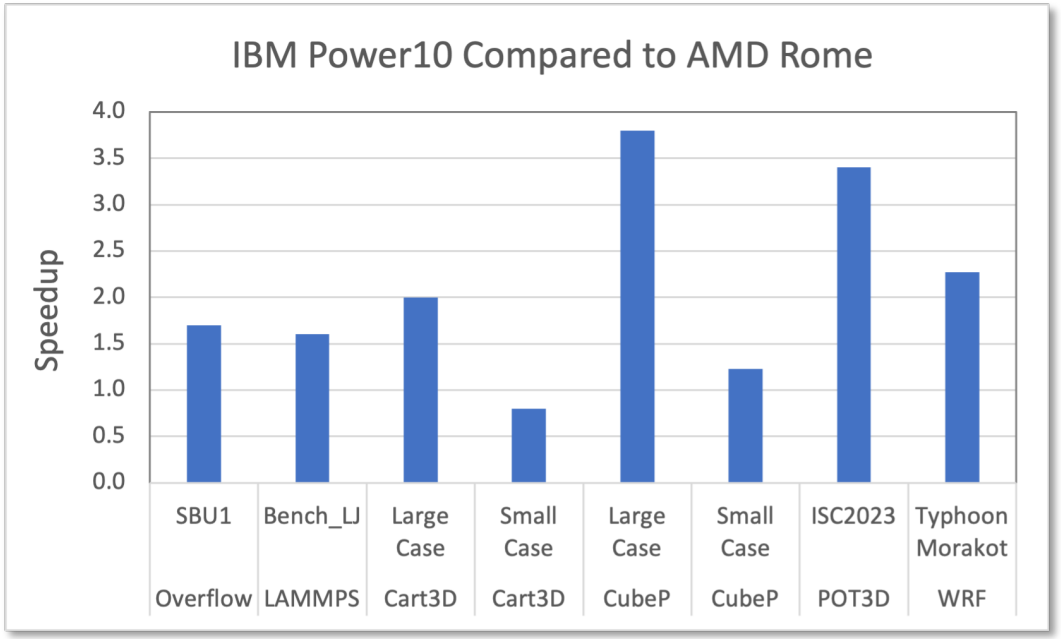






Testbed Evaluation—IBM Power10

- IBM Power10 node:
  - A loaner system from IBM for evaluation
  - 72 cores, each with 8 threads, which gives 576 logical CPUs
  - 2 TB of system memory
- Performance of applications, compared to AMD Rome node:
  - OVERFLOW SBU1 benchmark: 1.7x improvement
  - LAMMPS-standard Lennard-Jones test using MPI: 1.6x performance improvement
  - Cart3D tests using OpenMP: 2x improvement for two large test cases, 0.8x for a small test case
  - CubeP tests using OpenMP: 3.8x improvement for a large test case, 1.23x for a small test case
  - POT3D isc2023 benchmark using MPI: 3.4X improvement
  - WRF Typhoon Morakot benchmark using MPI: 2.3x improvement



Performance comparison of CFD applications on IBM Power10 with AMD Rome. Except for small test cases, a speedup of 1.6x-3.8x was observed.