Development and Performance Comparison of MPI and Fortran Coarrays within an Atmospheric Research Model

Soren Rasmussen with Ethan D Gutmann, Brian Friesen, Damian Rouson, Salvatore Filippone, Irene Moulitsas

11.16.2018

www.cranfield.ac.uk

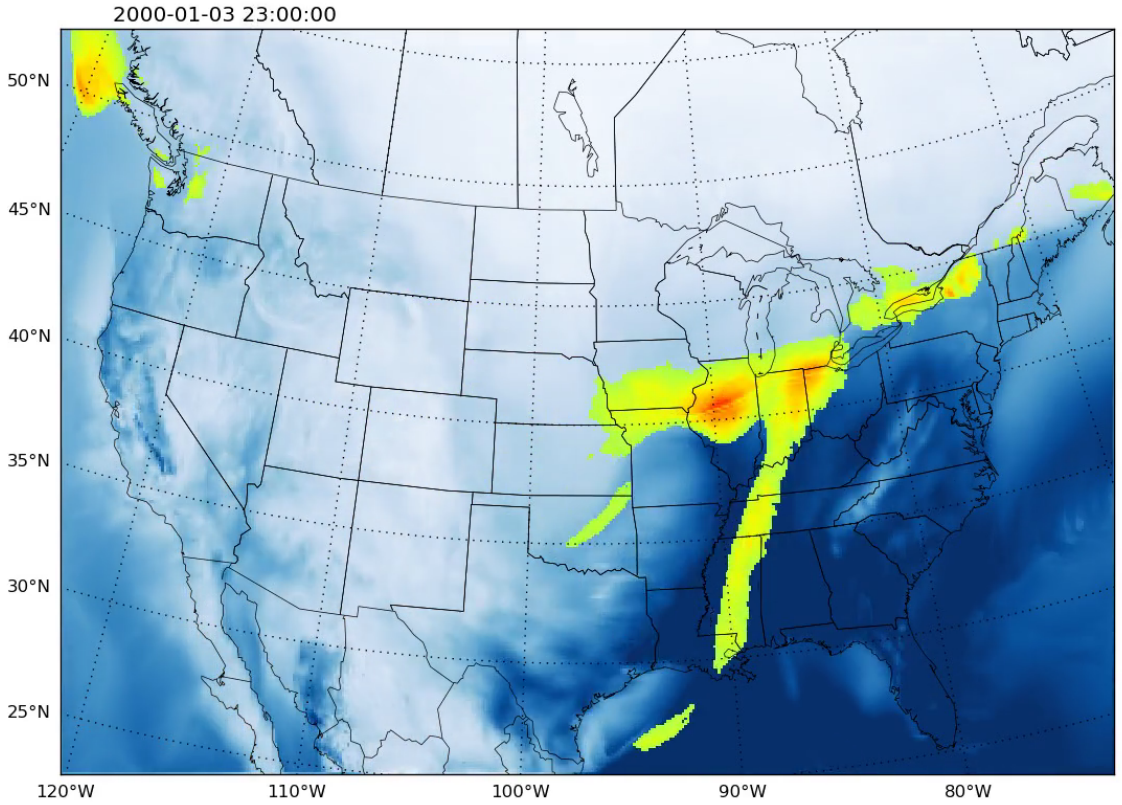
Motivation and Background

Fortran Coarrays: an alternative to MPI

- Parallel syntax in the Fortran standard
- Partitioned global address space (PGAS)
- Like MPI, Coarrays is SPMD

Application

- Mini-app of the Intermediate Complexity Atmospheric Research (ICAR) model
- Simplified atmospheric model that does high-resolution meteorological simulations



Programmability: communicating boundary regions

Coarray

```
east_halo(1:halo_size,:,:) [west_neighbor]  = A(start : start + halo_size, :, :)
```

MPI

```
call MPI_Type_create_subarray(ndims, sizes, subsizes, starts, order, oldtype, newType,ierr)
call MPI_Isend(A(start:start+halo_size, :, :), length, newType, west_neighbor, tag, MPI_COMM_WORLD, send_request, ierr)
call MPI_Irecv(east_halo(1:halo_size,:,:), length, newType, east_neighbor, tag, MPI_COMM_WORLD, & recv_request, status, ierr)
call MPI_Wait(request, status, ierr)
```

Additional MPI Lines

File	Additional lines	% of additional code
mp_thompson	15	0.28
domain_implementation	6	1.09
domain_interface	0	0
exchange_interface	54	48.65
exchange_implementation	226	91.13

Resources

Hardware and Software

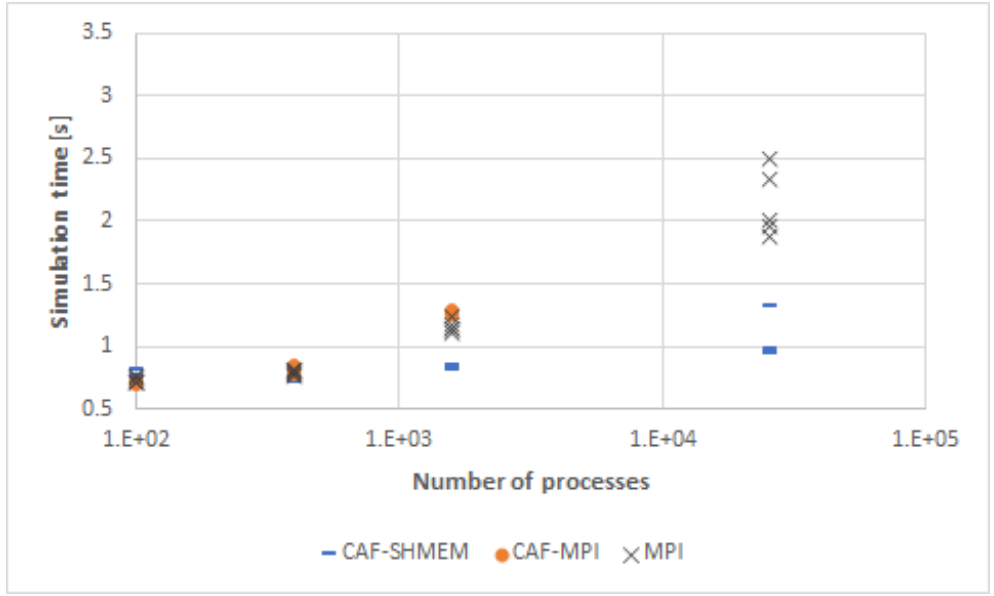
- National Center for Atmospheric Research's (NCAR) **Cheyenne**, a SGI ICE XA Cluster.
  - 4032 computation nodes, each node is dual socket with 2.3-Ghz Intel Xeon E5-2697V4 processors. Mellanox EDR Infiniband interconnect with a partial 9D Enhanced Hypercube single-plane topology.
  - GNU gfortran 6.3 with OpenCoarrays 1.9.4 using MPI and OpenSHMEM communication layers.

- Lawrence Berkeley National Laboratory's (LBNL) **Cori**, a Cray XC40
  - 12,076 total compute nodes, 9688 of them are single-socket, 68-core Intel Xeon Phi Processor 7250 ("Knight's Landing") at 1.4 GHz.
  - Cray Compiling Environment (CCE) 8.7.1.

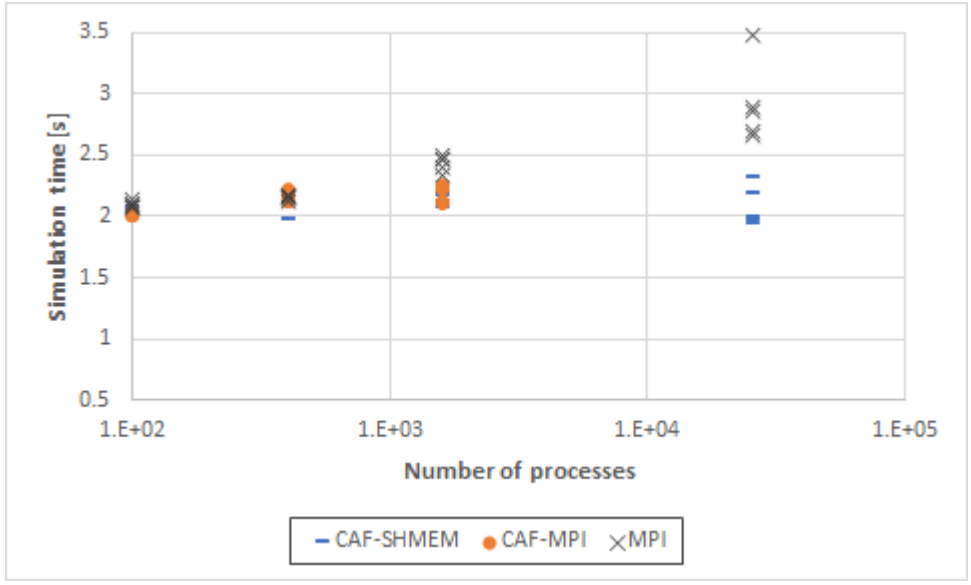
Results

Weak Scaling for Cheyenne

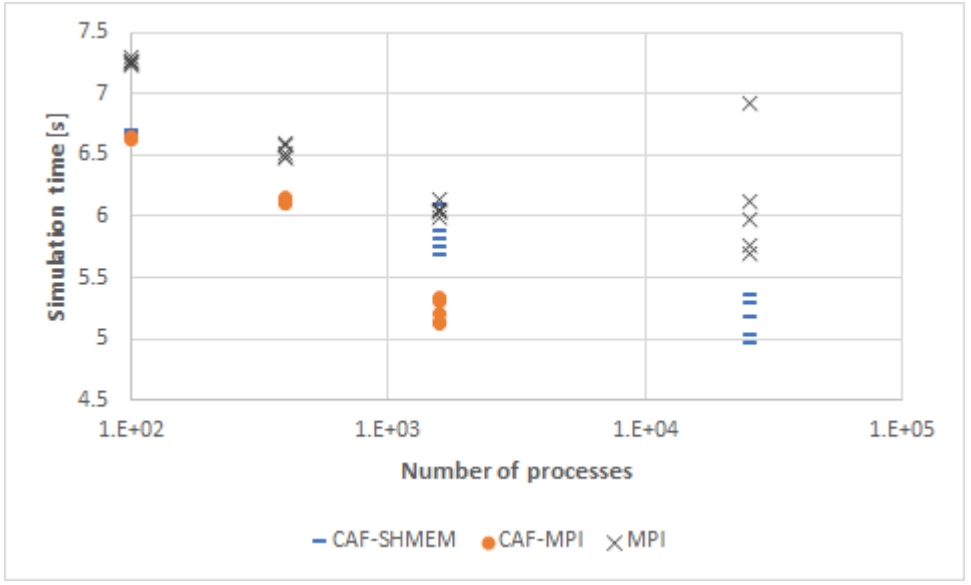
- Up to 25,600 processes



25 points per process



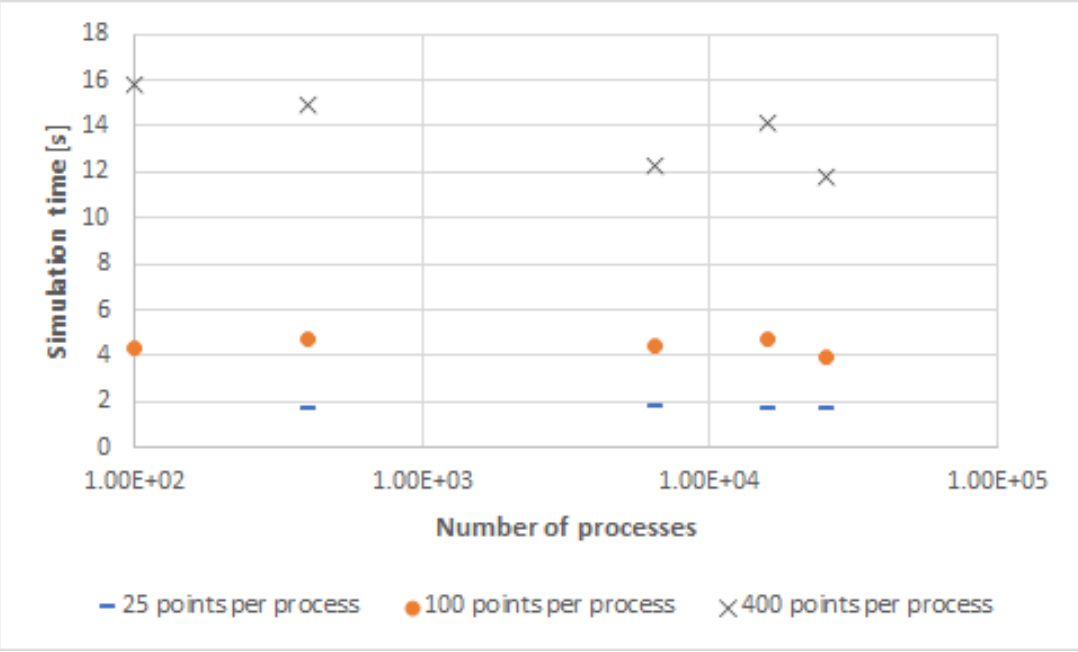
100 points per process



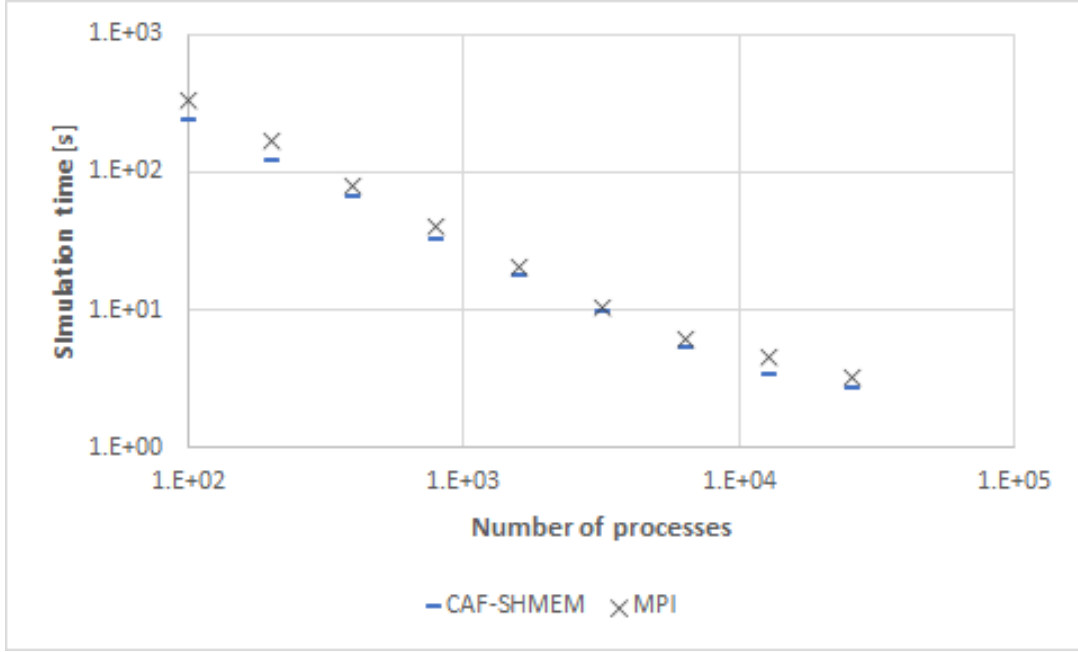
400 points per process

Results

Weak Scaling for Cray on Cori

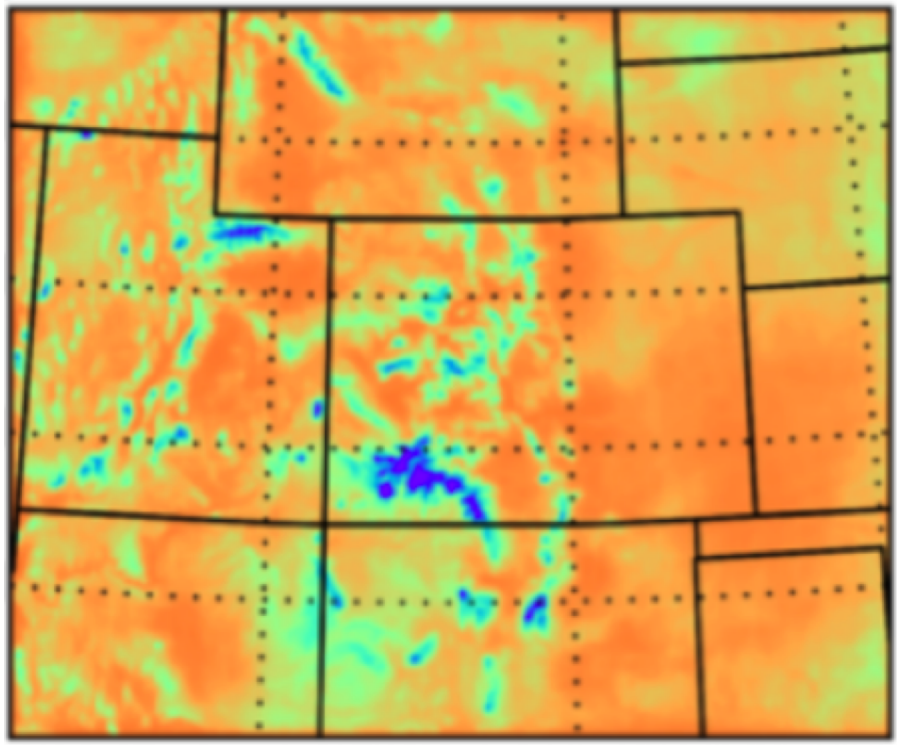


Cheyenne Strong Scaling



Conclusions and Future Work

- CAF OpenSHMEM outperformed MPI version
- No extra coding to switch OpenCoarray backend
- MPI Coarray backend initialization time at 2,000 processes is an hour, 3,000 exceeds 12 hour limit
- 91% more code needed by MPI in communication file



Future Work

- Strong scaling with CAF-MPI and on Cori
- Implement OpenCoarrays on Cray for direct comparisons
- Investigate CAF-MPI at higher process counts
- Add MPI Cartesian topology model

Results

Acknowledgements

- Based on work supported by NSF's National Center for Atmospheric Research (NCAR)
- The National Energy Research Scientific Computing Center (NERSC)
- Fellowship sponsored by Sourcery Institute