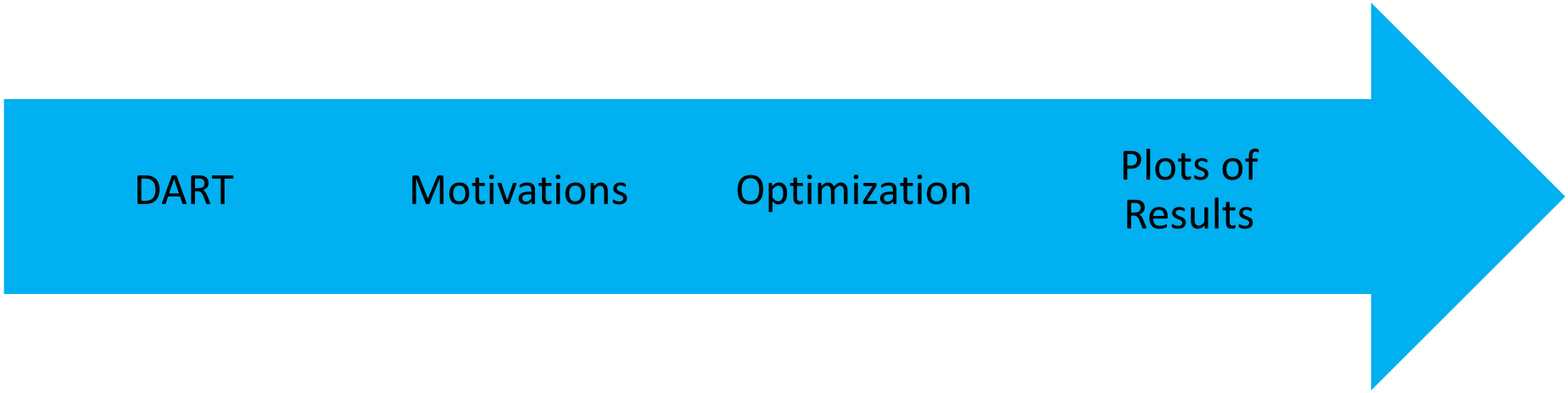
Evaluating the Performance of Large Scale Data Assimilation on Modern Geophysical Models

Julia Piscioniere
College of Charleston
SIParCS Intern

Mentors:
Brian Dobbins and John Dennis
August 3, 2018

Overview



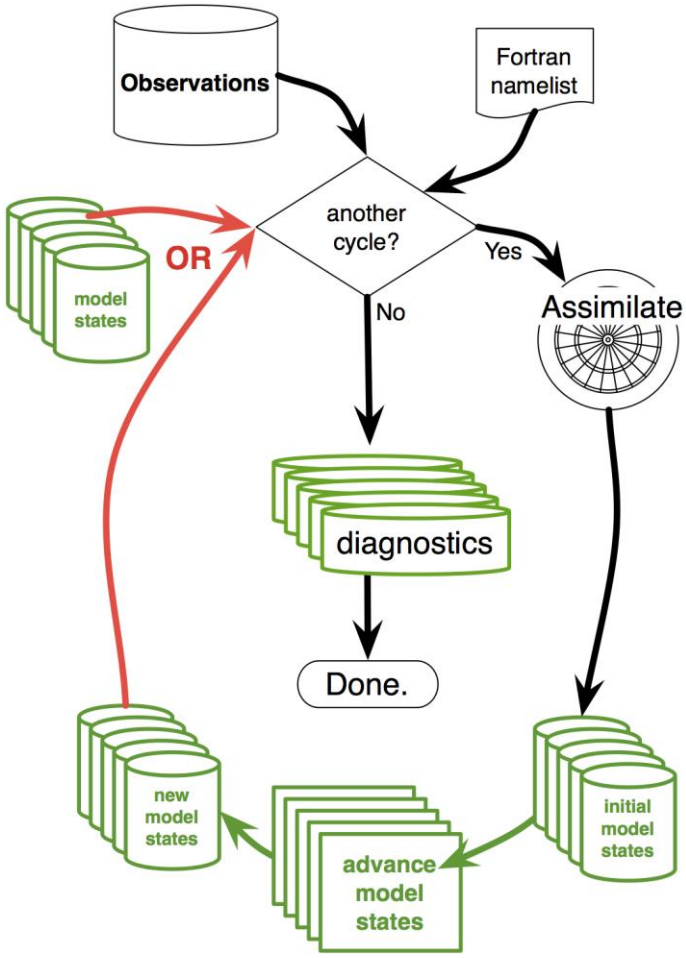
Introduction

DART?

- Data Assimilation Research Testbed

Data Assimilation?

- When information from new physical observations are used to update a numerical model in order to produce more accurate predictions



https://www.image.ucar.edu/DAReS/DART/

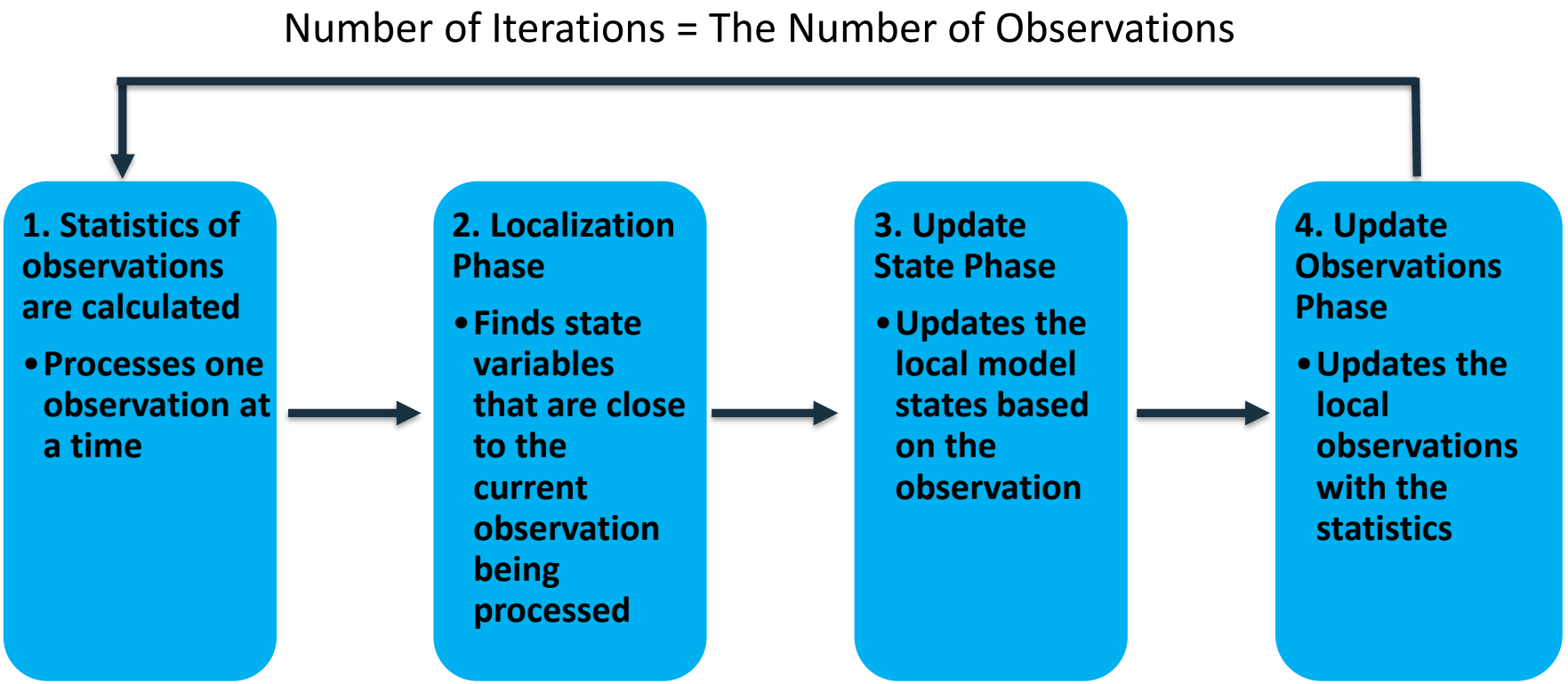
Motivation

Overall Goal: Improve the performance of data assimilation

- Observation density is increasing
- CESM ensembles: hundreds or thousands of nodes
- Enable DART to scale to larger node counts

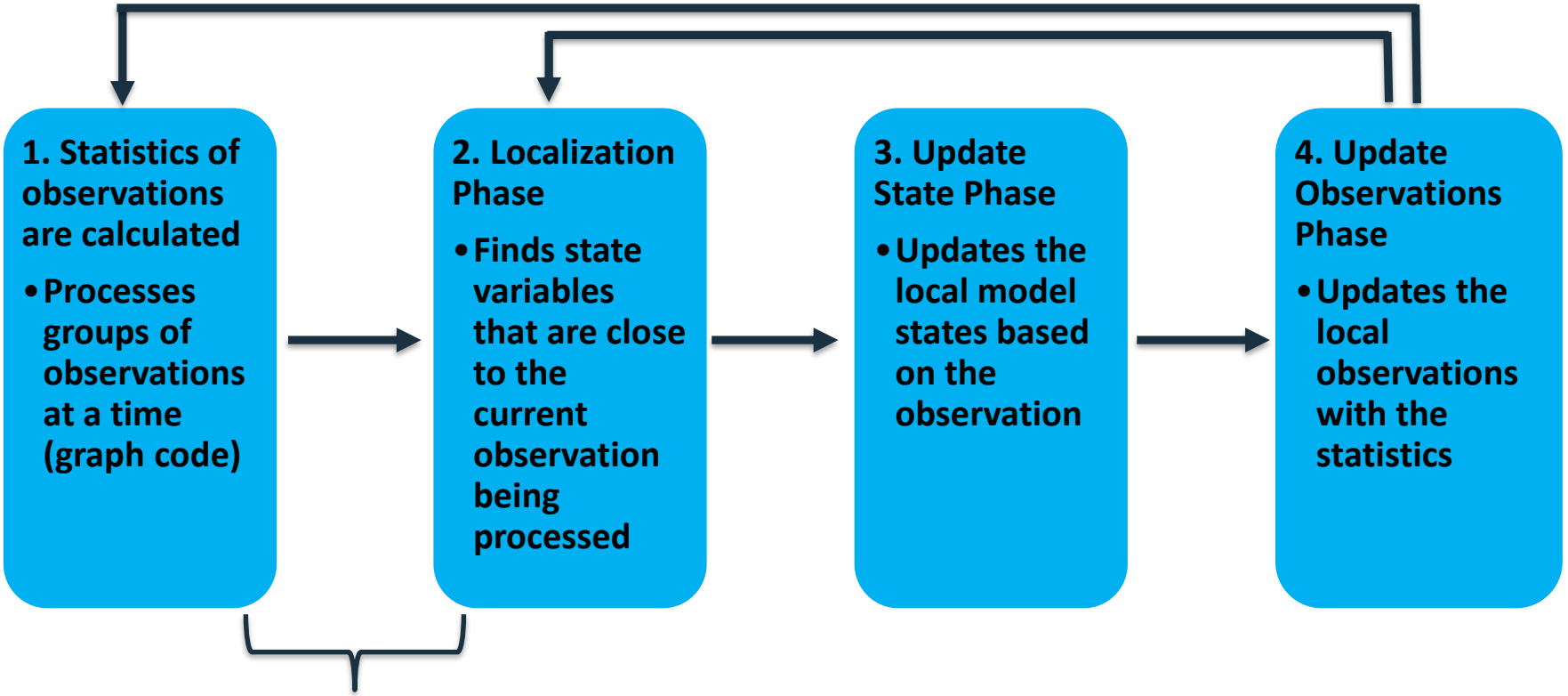
Optimization Approach: Reduce communication in DART

How DART Currently Works



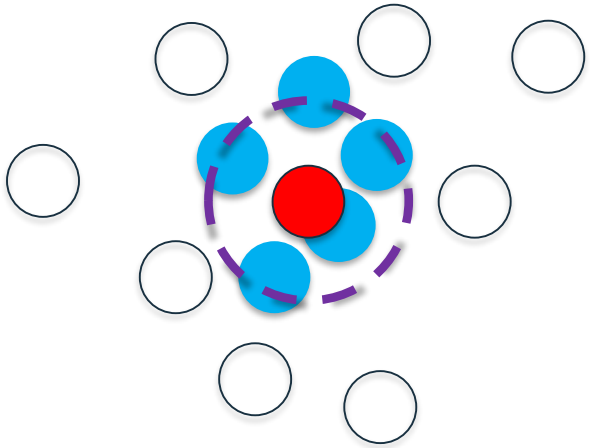
How the Optimized DART Works

Number of Iterations = The Number of Groups

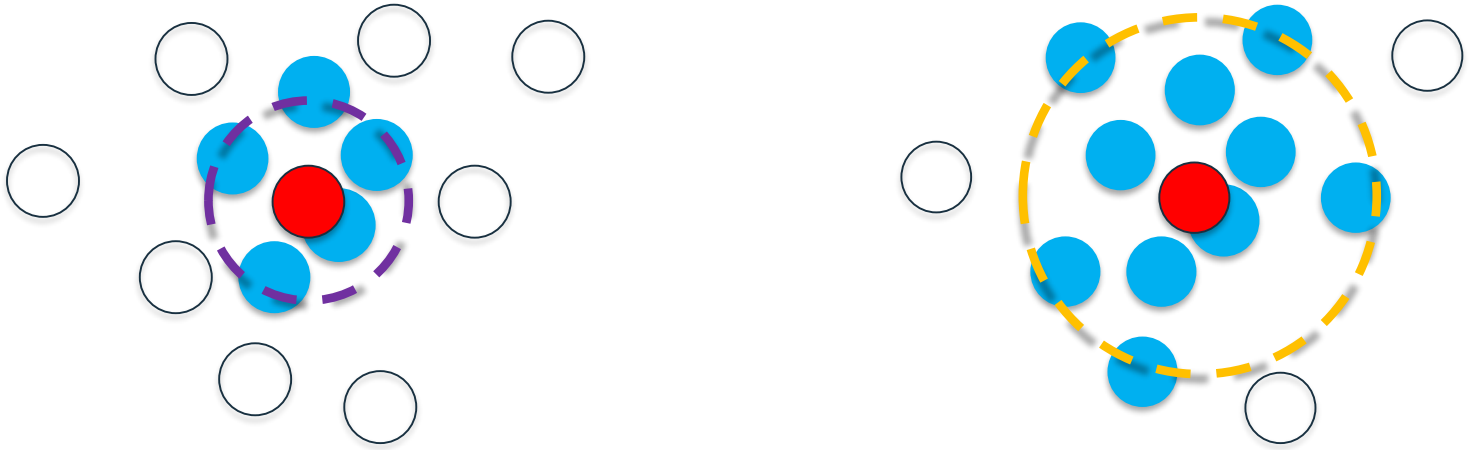


Number of Communication Counts = 1-2% of the Number of Observations

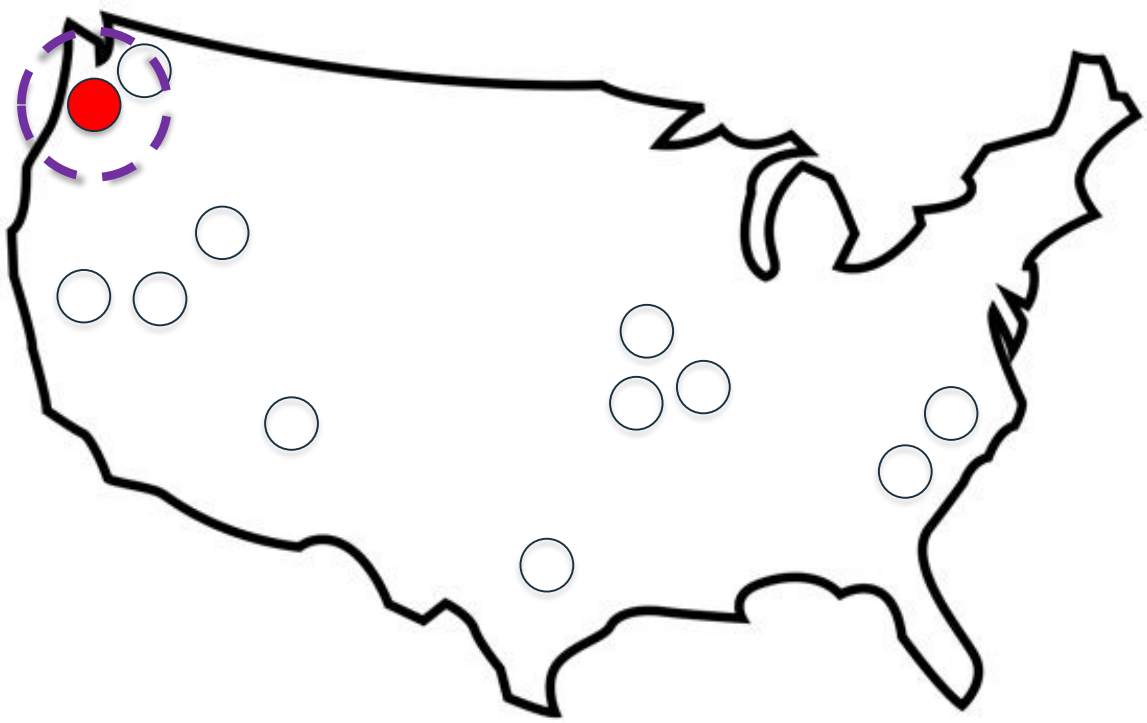
Cutoff Distance



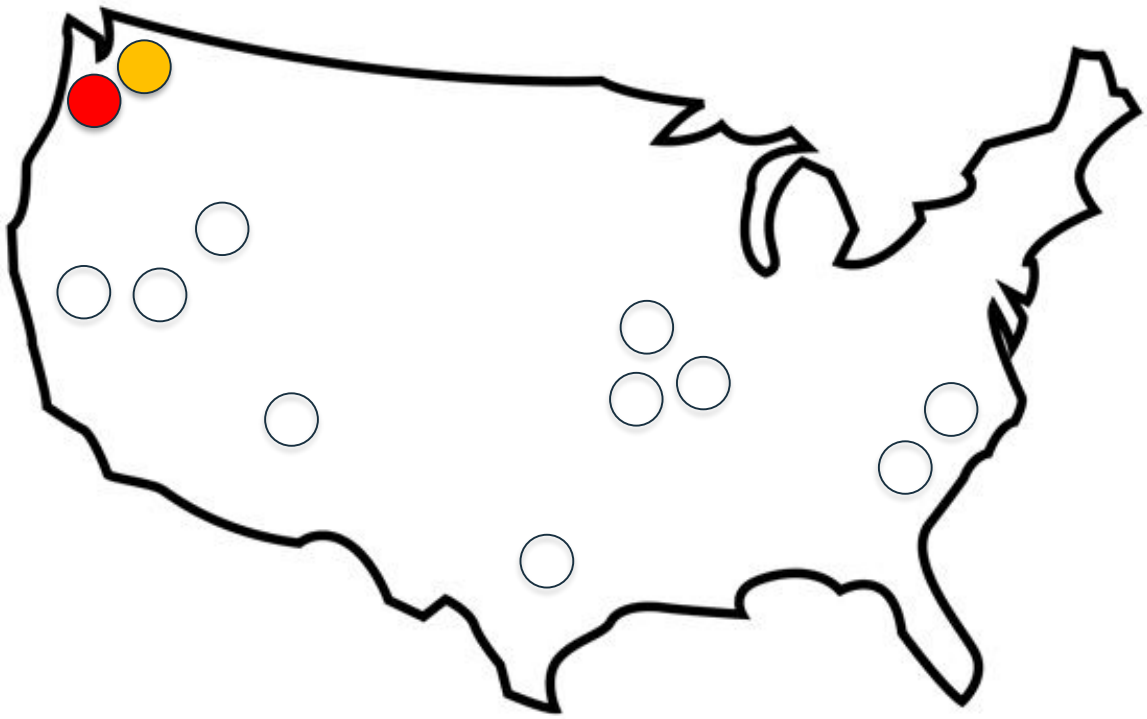
Cutoff Distance



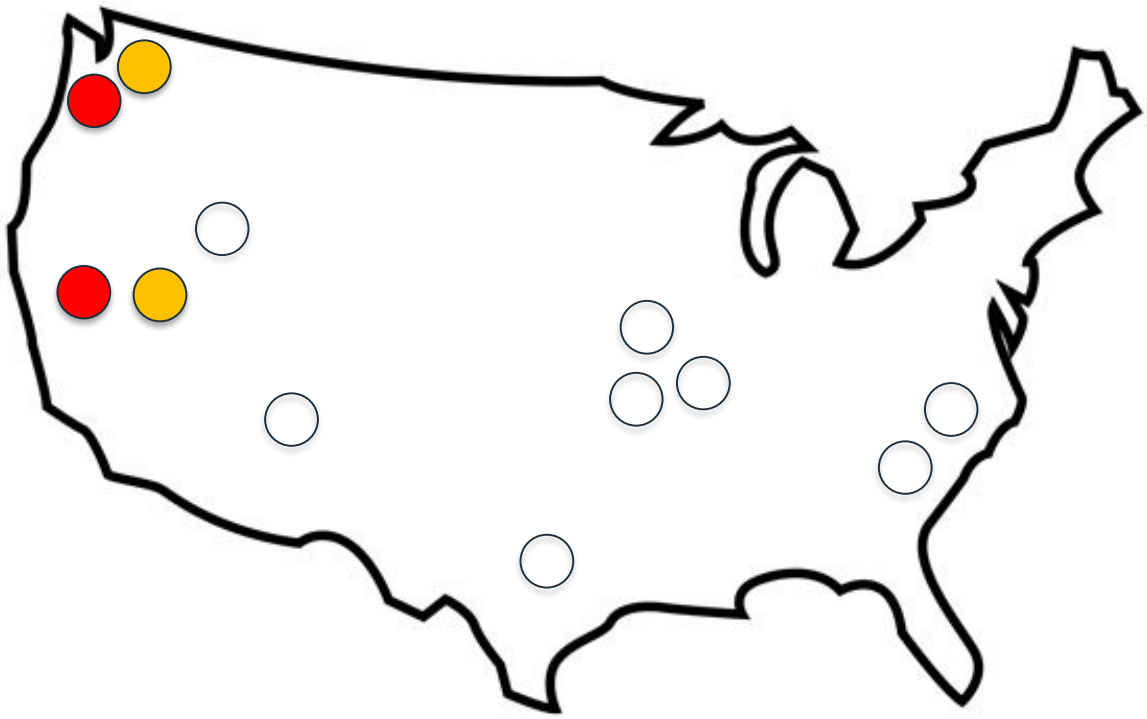
What is a Graph Coloring Algorithm?



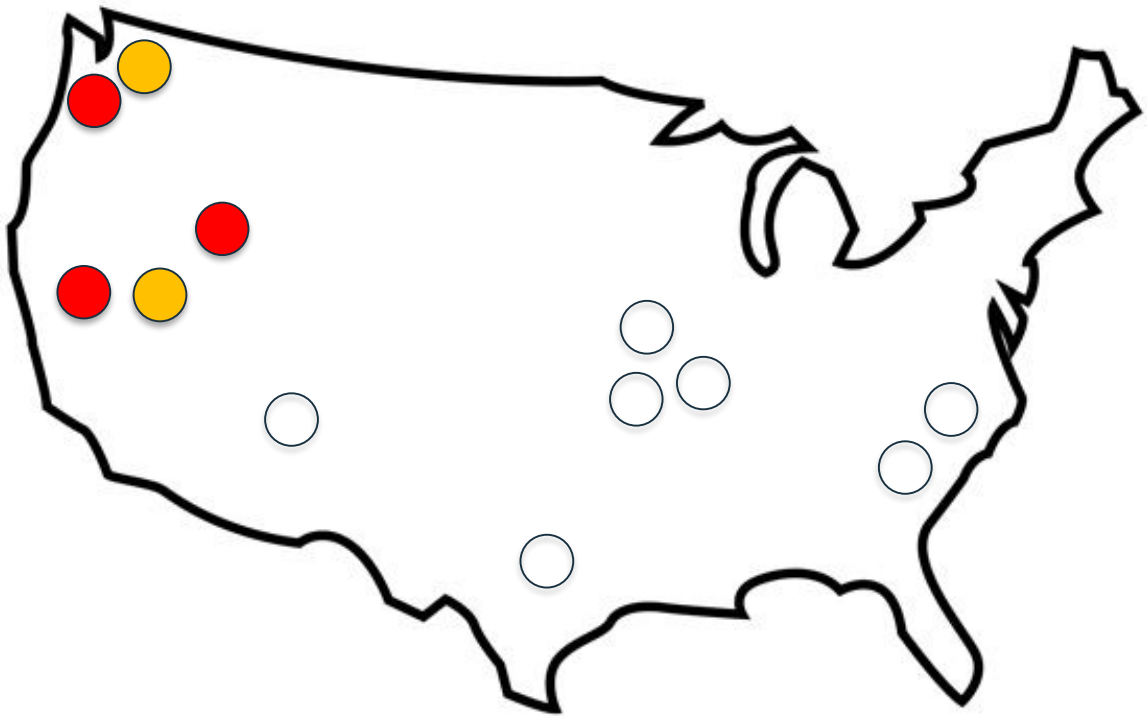
What is a Graph Coloring Algorithm?



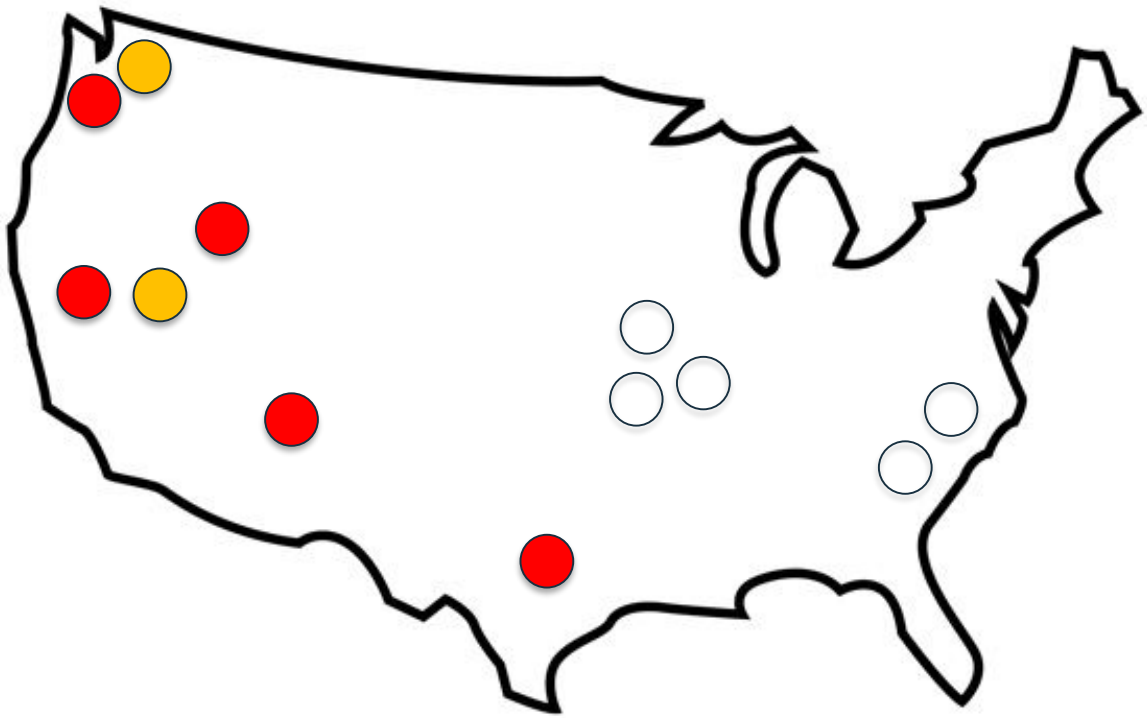
What is a Graph Coloring Algorithm?



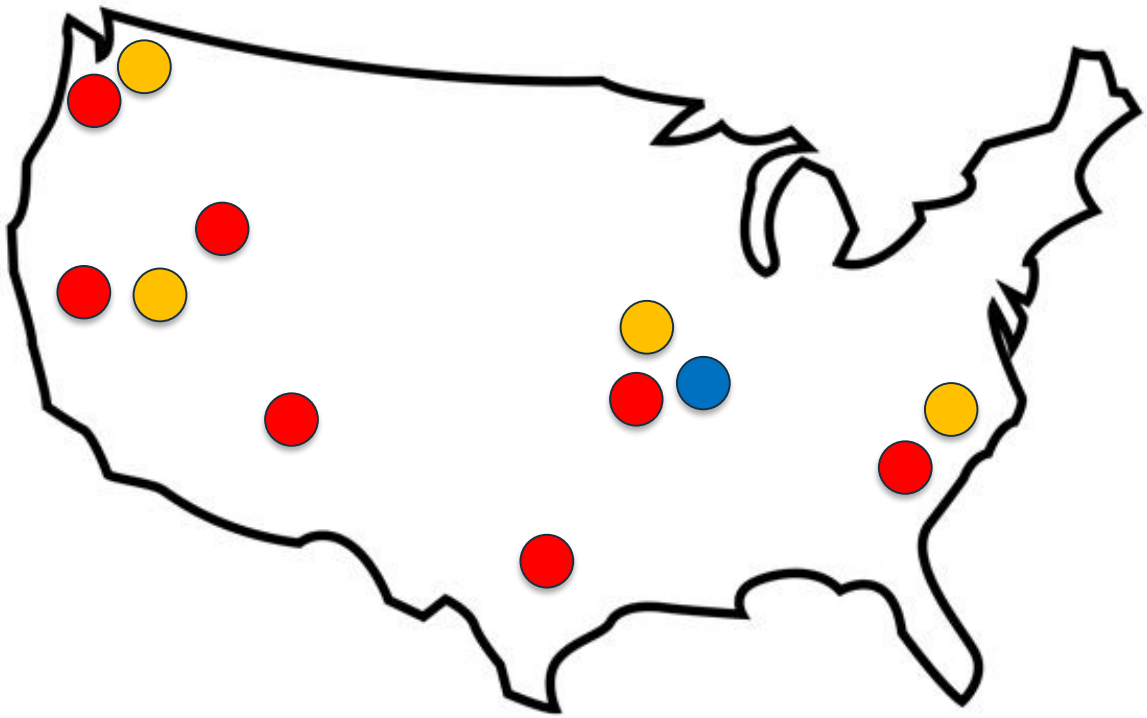
What is a Graph Coloring Algorithm?



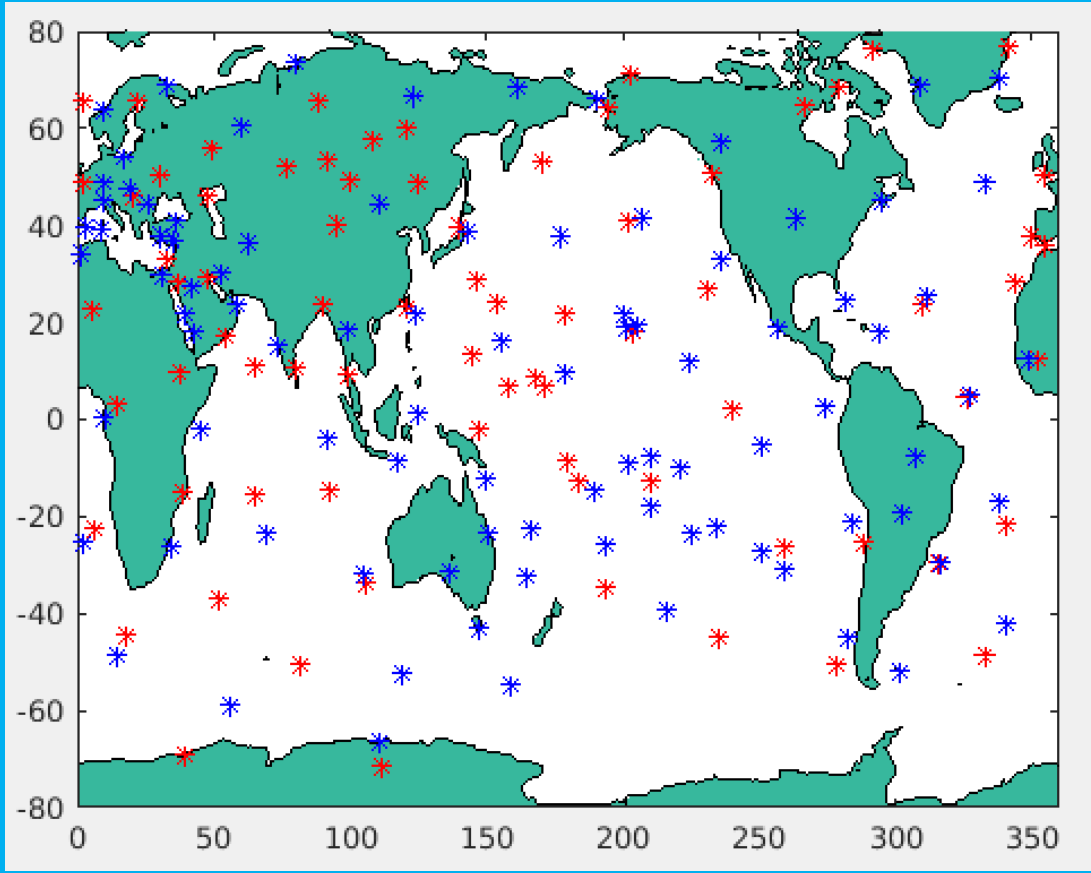
What is a Graph Coloring Algorithm?



What is a Graph Coloring Algorithm?



What is a Graph Coloring Algorithm?



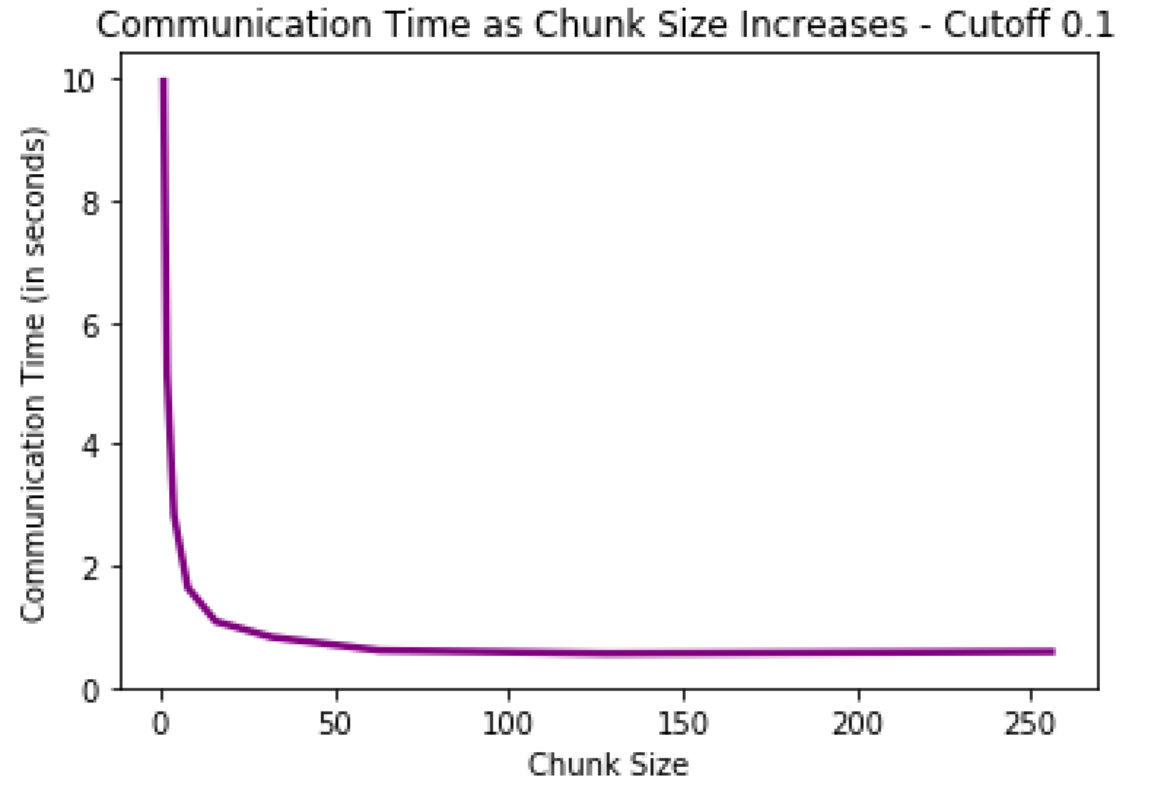
Color 4 = Red
Color 5 = Blue

Running Comparisons

Chunk Size	Ensemble Size	Number of Nodes	Cutoff Distance
• The highest number of observations that will be in a chunk • Used chunk size of 256	• Number of ensemble members DART will produce • Used ensemble size 100	• Range from 4 - 384	• Used cutoffs 0.1 and 0.2

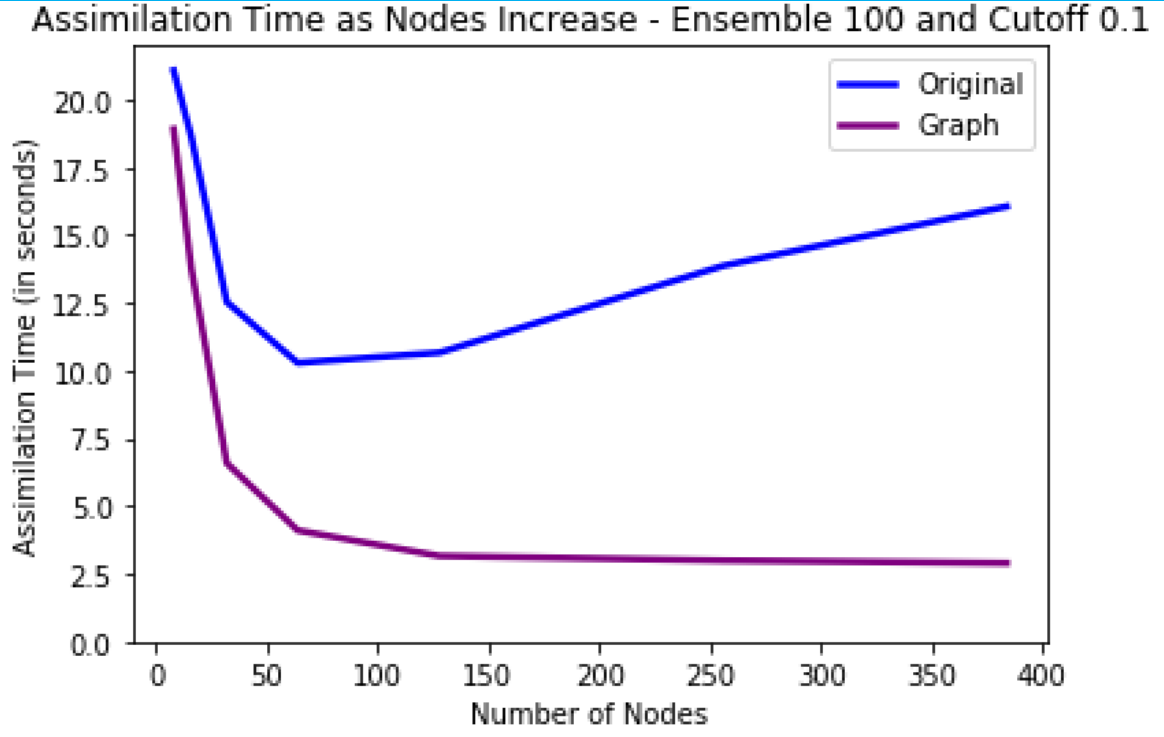
- All runs were 1–degree CAM (Community Atmospheric Model) cases

Choosing Chunk Size



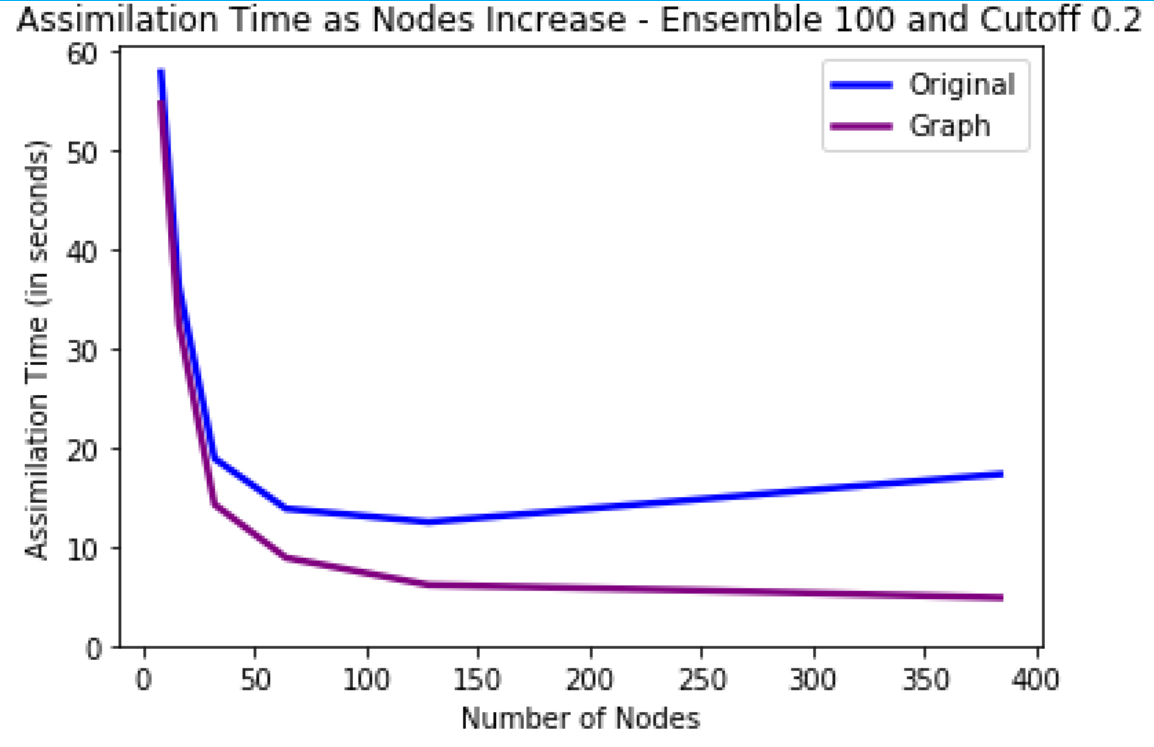
Chunk sizes of 128 and 256 were both viable options

Assimilation Time Plots



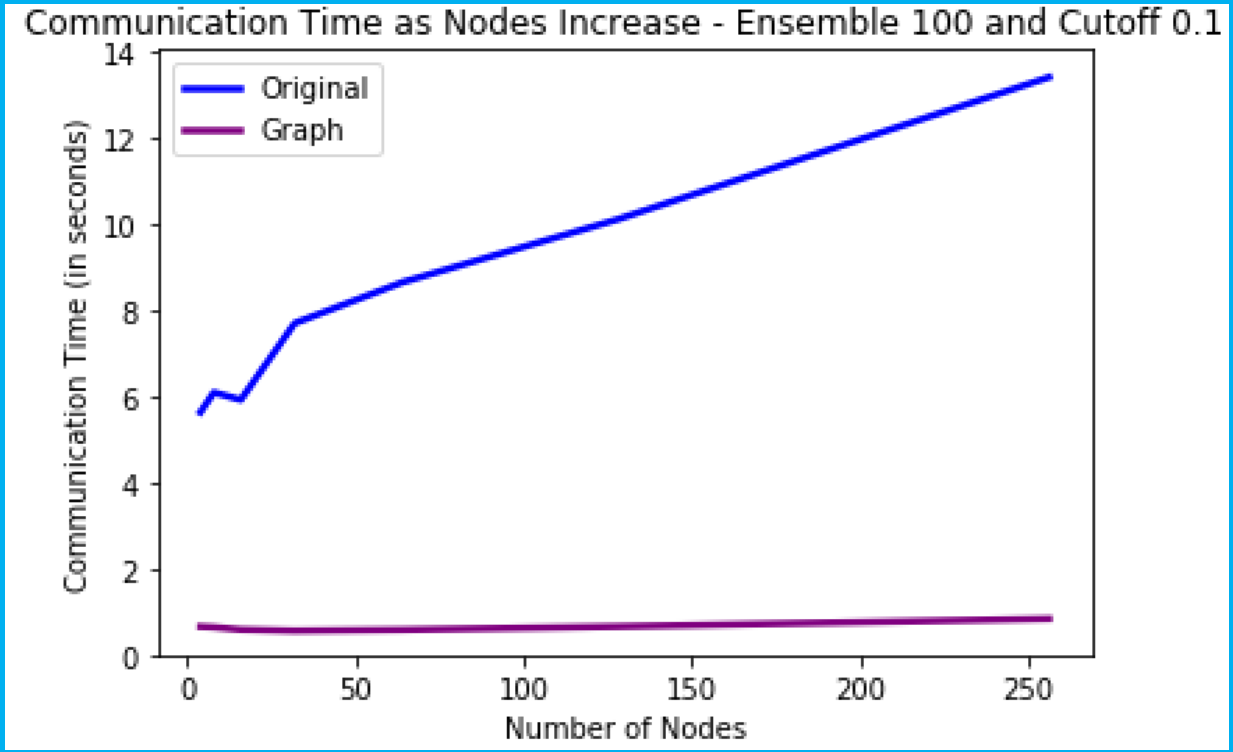
At 64 nodes, the graph code is 60% faster than the original code

Assimilation Time Plots



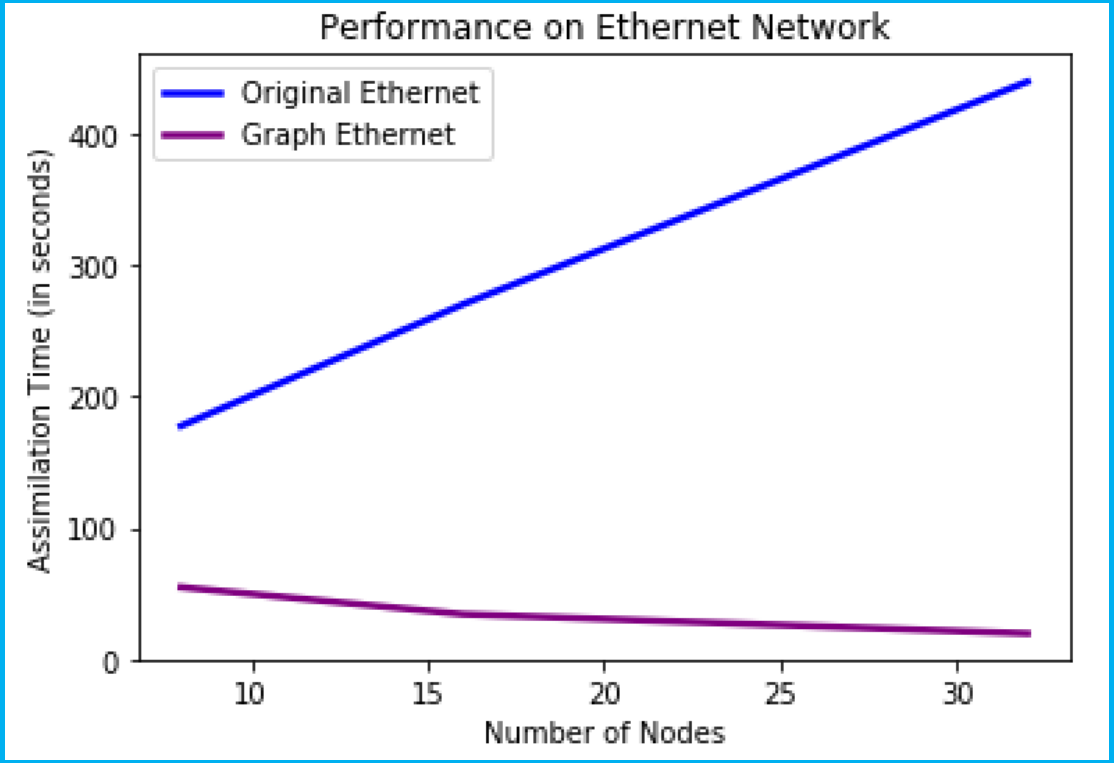
At 128 nodes, the graph code is 51% faster than the original

Communication Time Plot



The communication time is decreased by an average of 90% across nodes from 4-384

Cheyenne Over Gigabit Ethernet



The graph code may run faster on a slower network, meaning people with slower networks can run DART more efficiently

Conclusions

- The graph code communication time is decreased by an average of 90%
- The assimilation time is decreased by a minimum of 10% and - as the number of nodes increases - a maximum of 85%
- The graph code may be a better option for people who want to run DART on systems with slower network links

Thank you! Questions?

juliapiscioniere@gmail.com

Acknowledgements

Mentors:
- Brian Dobbins
- John Dennis

CODE Group:
- AJ Lauer
- Elliot Foust
- Jenna Preston

DAReS Group:
- Nancy Collins
- Jeff Anderson
- Tim Hoar

- Elizabeth Faircloth, admin
- Rich Loft
- Valerie Sloan
- National Center for Atmospheric Research
- National Science Foundation
- John Piscioniere
- My fellow interns :)