Computational Performance Improvements in GFSv16

Jun Wang[1*], Jeffrey Whitaker[2], Edward Hartnett[3, 7], James Abeles[4], Gerhard Theurich[5], Wen Meng[4], Cory Martin[6], Jose-Henrique Alves[8], Fanglin Yang[1], Arun Chawla[1]

[1]NOAA/NWS/NCEP/EMC, [2]NOAA/ESRL/PSD, [3]NOAA/ESRL/GSD. [4]IMSG@NOAA/NWS/NCEP/EMC, [5]ESMF/NRL/SAIC, [6]RedLine@NOAA/NWS/NCEP/EMC,[7]CIRES, University of Colorado,[8]SRG@NOAA/NWS/NCEP/EMC

[*]Jun.Wang@noaa.gov

Introduction:

NCEP is developing its next version of the medium-range weather Global Forecast System (GFSv16) for operational implementation in 2021. One major upgrade in GFSv16 is to raise the model top from 54km to 80km, while increasing the number of vertical levels from 64 to 127. This change, along with other new physics upgrades, significantly increases the computational cost and the output data amount. In this paper, we describe technical advances made to improve the GFSv16 computational performance.

Model description:

The prognostic model of GFSv16 is the Unified Forecast System (UFS) model with the atmosphere being one-way coupled with an ocean wave model through the NOAA Environmental Modeling System (NEMS) infrastructure. The NEMS is built upon the Earth System Modeling Framework (ESMF) and uses the National Unified Operational Prediction Capability (NUOPC) Layer. The atmosphere component consists of the Finite-Volume Cubed-Sphere Dynamical Core (FV3), GFS physics package and write grid components running at the C768L127 resolution. The wave component uses the NOAA WAVEWATCH III model (WW3) with a global grid mosaic [WW3DG, 2019]. WW3 global core resolution is increased from ½ degree to ⅙ degree. The model produces forecast history files in netcdf format and post-processed grib2 files out to 16 days, four times a day with hourly output for the first 120 hours and then three-hourly output up to 384 hours.

Technical improvements:

Four major technical improvements have been applied to GFSv16. They are described below.

1) Inline post on write grid component

In GFSv15, the asynchronous IO technology was applied to set up the write grid components running on groups of MPI tasks which are different from forecast MPI tasks. The write grid components receive data from forecast tasks through various regridding methods and then output history files, while the forecast tasks simultaneously conduct integration. In GFSv16, an inline post capability is added to reduce the IO activity in the whole computer system (Fig. 1). With the inline post implementation, the total time needed to generate a 2.5GB GRIB2 GFS master file at a single forecast time is reduced from 211s in a standalone post to 39s when both are run with 84 tasks on the NCEP WCOSS operational computer.

2) Parallel writing compressed netcdf history files

Due to the atmosphere component vertical resolution increase, forecast output data is doubled, and the time to read and write these data also increases significantly. In GFSv16 the history file data format is switched from the previous plain binary NEMSIO format to netCDF/HDF5 format, with bit-smoothing, zlib compression and the data quantizing method described in [Silver & Zender, 2017]; the size of a single 3D fields history file is reduced from 33.6GB to 6.3GB with acceptable precision loss. Through the collaboration between EMC and the NetCDF working group, Unidata released a new NetCDF version with the capability of writing compressed data in parallel. The corresponding interface is implemented in GFSv16. This new feature enables

GFSv16 to write out a single lossy compressed 3D history file in 43s, a reduction from 441s without parallelization. The new method also makes it possible to reduce the overall model computing cost.

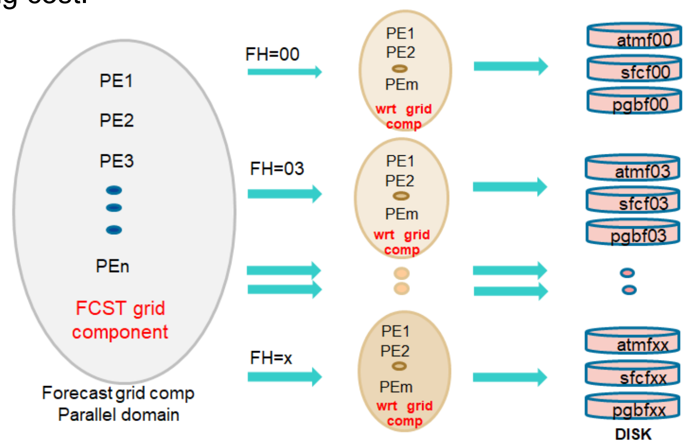

Fig 1. Parallelization of GFSv16 write grid component with inline post.

3) Scalability improvement in WW3

A scalability issue was uncovered in WW3 in the GFSv16 experiments. A threading issue and improper field gather operation were identified and fixed in WW3. The fixes improved WW3 scalability. When tripling the number of MPI tasks, WW3 now takes about 36% of the original run time, while before these fixes it took about 94% of the original run time.

4) Data transfer improvement ESMF model components

While running GFSv16 experiments, a persistent delay was found on forecast tasks when they transfer data to the write tasks. The ESMF support group identified the underlying issue and fixed the problem by writing all the messages sent to the same destination write task into a buffer and then sending the whole message via a single MPI_Isend(). This fix allows each data transfer time on forecast tasks to be reduced from 4s to less than 0.5s. This method is also applied to the NUOPC connector that allows the atmosphere model to send data to the wave model.

Future work:

NCEP is quickly approaching exascale computing in the next few years when a fully coupled atmosphere, ocean, ice and wave system is developed and higher resolutions are chosen for the components. Future development to solve the challenges include: 1) developing computationally efficient components in the coupled system; 2) developing an efficient system integration strategy; 3) developing efficient IO with prudently chosen output fields, efficient compression algorithms and proper data representation; 4) reorganizing the workflow to reduce IO stress on the computer system; and 5) exploring other technologies such as cloud computing, GPU computing, etc.

References:

Silver, J. D. and Zender, C. S.: The compression–error trade-off for large gridded data sets, Geosci. Model Dev., 10, 413–423, https://doi.org/10.5194/gmd-10-413-2017, 2017

The WAVEWATCH III® Development Group (WW3DG), 2019: User manual and system documentation of WAVEWATCH III® version 6.07. Tech. Note 333, NOAA/NWS/NCEP/MMAB, College Park, MD, USA, 326 pp. + Appendices.