LONG-TERM GOALS

Successful operational implementation of the global ESPC fully coupled system in 2018 will require that the coupled system and the constituent models are able to run efficiently and in a timely manner on Navy operational computer systems to ensure that products produced by the systems are available for fleet user consumption and downstream dependencies. This product will analyze and modify the atmosphere model (NAVGEM), the ocean model (HYCOM) and the sea ice model (CICE) to ensure that the ESPC coupled system is able to take advantage of modern computational platforms and increase computational efficiency and scalability.

OBJECTIVES

Instrument, analyze, and modify the Navy’s global atmosphere model (NAVGEM), global ocean model (HYCOM), global sea ice model (CICE), and global atmosphere data assimilation system (NAVDAS-AR) to increase the scalability of the component systems and thereby improve the computational efficiency of the ESPC system as a whole.

APPROACH

Greater efficiency of a coupled system begins with identifying opportunities for improvement. This will require enhanced instrumentation and analysis of model and coupling components to identify the particular sections of code with the largest impact on performance. This instrumentation will show both communication patterns (typically within component models) and places in the code where the model spends a lot of time.
Once identified, we will use an incremental approach to improve the efficiency of those sections. Typically these enhancements will be in the form of adjusting the control flow, altering communication patterns, changing data structures, or taking advantage of improved algorithms that make better use of modern architectures. As each improvement is complete, the full system will be re-instrumented and re-evaluated to ensure that the changes made provided improved efficiency and did not alter the results.

Key performers are Timothy Whitcomb (NRLMRY, focus on global modeling), Steven Lowder (NRLMRY/SAIC, computer scientist focused on I/O framework), and Alan Wallcraft (NRLSSC, focus on HYCOM and CICE).

**WORK COMPLETED**

Last year, the focus was on the atmosphere model alone. We began full-model scaling test for high-resolution NAVGEM and the analysis and refactoring of NAVGEM I/O.

**RESULTS**

One persistent bottleneck in the NAVGEM is the I/O subsystem for handling input/output of model history files, which are used for restarts, data assimilation, and post-processing for output and downstream products. The original architecture for output uses MPI gather operations to bring data from each core to a single process that uses unformatted Fortran I/O calls to write a custom binary formatted file while all other processors wait. This simple setup works well for lower model resolutions and systems with fast disks, but represents serious performance issues with filesystems like Lustre that are geared toward parallel file operations. We have added an MPI-2 I/O option to the NAVGEM, reducing the wall time spent in I/O at T639L64 resolution, for example, from 30% to 5%.

We were able to further reduce the total wall time requirements for the modeling system by interleaving the generation of output with the forecast model integration. Figure 1 shows this schematically – the top two timelines show what was possible for the system prior to this work. The normal method is the first line, which runs the forecast model followed by an output job to generate all of the model output. The second option is not recommended as it may lead to longer runtimes than the first option due to load imbalance. The new method (which has its roots in FNMOC operations) trades an increased requirement for computing resources (since output jobs run parallel with the model) for a total wall clock requirement runtime that barely exceeds the forecast model itself. We use a metascheduler to monitor the log file, which allows us to trigger single-hour output generation as soon as a particular forecast lead time is complete. This has advantages over other methods (such as embedding system calls in the model or monitoring directories to watch for files) because it is system-independent and allows for much finer-grained control over the output actions (since multiple tasks can be triggered at every output time).

A breakdown in the scalability curve for the operational version of NAVGEM (with certain core counts requiring significantly higher resources) has been traced to the current 1-D domain decomposition. These results support earlier supposition about deficiencies in the decomposition, and will be addressed under the HPCMP Applications Software Initiative (HASI) project “Optimizing Global and Regional Earth System Prediction” by implementing a 2-D domain decomposition.
Figure 1: Schematic diagram showing the output generation options for NAVGEM. Blue arrows show the forecast model, and purple arrows show the generation of output (i.e. interpolated, pressure-level lat/lon data used by customers and downstream).

Figure 2: Number of XC40 cores vs total Cray XC40 core hours for 0.04° global a) HYCOM 2.2.98 (blue) and b) HYCOM 2.2.27 (red).
As part of the Accelerated Prediction of the Polar Ice and Global Ocean (APPIGO) project, HYCOM's handling of land avoidance was changed from do-loop nests to a set of land-sea mask arrays because this approach is more amendable to optimization on an attached processor. In addition HYCOM has been updated to use dynamic memory allocation, since this avoids taking up memory space on cores that do not run HYCOM in a coupled application such as ESPC. Under this project, we updated the standard 0.04° global HYCOM benchmark, used in the HPCMP “TI” benchmark suite, to HYCOM 2.2.98 which includes both dynamic memory allocation and land masks. The original 2.2.27 version, used static (common block) memory allocation and do-loop land avoidance. Figure 2 has total Cray XC40 core hours on the y axis, and so a horizontal line would be perfect scaling. The blue curve is 2.2.98 which is super-scalar from 1000 to 4000 cores. The red curve is 2.2.27 which performs about the same as 2.2.98 on 1000 cores but shows a gradual loss of scaling on more cores. Thus we have actually gained scalability by making these changes.

We require bit for bit reproducability from HYCOM on any number of MPI tasks. Intel Fortran does not supply this by default, so we set -fp-model precise, because without this option the compiler will mix scalar and vector operations that have different rounding properties, and -no-fma because the AVX2 fused multiply-add again has different rounding than standard multiply-add. These options slow down HYCOM, but on the Cray XC40 using huge pages improved performance by about 3% and making the first dimension of all arrays a multiple of 8 (with ifort -align array64byte) saved another 3-6%. Overall, on 8160 cores the wall clock time improved from 6.1 minutes per model day to 5.5 minutes. Figure 3 shows benchmark performance across 3 generations of systems at Navy DSRC. These all consist of Intel Xeon dual socket nodes with very little difference in performance per core across generations, but with more cores per node on the more recent systems.

![Figure 3: Number of cores vs total core hours for 0.04° global HYCOM 2.2.98 on a) IBM iDataPlex with 16 cores/node (green), b) Cray XC30 with 24 cores/node (red) and c) Cray XC40 with 32 cores/node.](image-url)
Sea ice is only present at high latitudes, with more ice in winter than summer. So load balance has static (land, never ice covered sea) and dynamic (winter vs summer) aspects. CICEv4 has limited domain decomposition options, with the best for global domains typically being slenderX1 or slenderX2. These are long and thin, which is not good for performance, but best for load balance. CICE can have multiple tiles (blocks) per task, but we currently hardwire 1 per task. Figure 4 is for a 0.72° global domain with 40 tiles, the corresponding 0.08° domain has 900 tiles.

CICE currently uses in-line code to generate its domain decomposition, so it can't allow for seasonal changes in ice cover. By calculating the tiling off-line we can allow for the actual ice cover. For example, by using 2003-2014 monthly sea ice concentration fields a load balanced slenderX2 decomposition for the 0.72° global domain has 27 vs the original 40 tiles (figure 5) and the global 0.08° domain has 573 vs the original 900 tiles. This still uses a single decomposition for the entire year, but it allows for the seasonal differences in ice cover subject to this limitation. It would be possible to use a different decomposition each month of the year. There are several new decomposition options in CICE version 5, and some are better than slenderX for large global domains. However, the concept of using actual knowledge of the sea ice extent in the domain decomposition still applies.

![Figure 4: 0.72° global domain with the slenderX2 decomposition (40 tiles).](image-url)
Figure 5: load balanced 0.72° global domain with the slenderX2 decomposition (27 tiles).

IMPACT/APPLICATIONS

The future impact of this project is to ensure operational efficiency of the Naval operational environmental prediction system that is targeted for IOC in 2018.

RELATED PROJECTS

This work is part of a larger ESPC project. Other related projects are a) 6.2 NOPP: Accelerated Prediction of the Polar Ice and Global Ocean (APPIGO) with a strong focus on accelerators for the ocean and ice model, b) a two year HPCMP Applications Software Initiative (HASI) project: Optimizing Global and Regional Earth System Prediction with a focus on NAVGEM scalability on existing systems and on accelerators, c) 6.4 NAVGEM for extensions to and validation of the atmosphere model, and d) 6.4 Large Scale Ocean Modeling for extensions to and validation of the HYCOM ocean and CICE sea ice models. A FY15 PETTT project for an ESMF asynchronous I/O component will be leveraged in future years of this project.