Title: MPAS-Ocean NESAP Status Report

Author(s): Petersen, Mark Roger
Arndt, William
Keen, Noel

Intended for: Report

Issued: 2017-04-19

Disclaimer:
Los Alamos National Laboratory, an affirmative action/equal opportunity employer, is operated by the Los Alamos National Security, LLC for the National Nuclear Security Administration of the U.S. Department of Energy under contract DE-AC52-06NA25396. By approving this article, the publisher recognizes that the U.S. Government retains nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or to allow others to do so, for U.S. Government purposes. Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy. Los Alamos National Laboratory strongly supports academic freedom and a researcher's right to publish; as an institution, however, the Laboratory does not endorse the viewpoint of a publication or guarantee its technical correctness.

MPAS-Ocean NESAP Status Report

Mark Petersen, Bill Arndt, Noel Keen

April 11, 2017

contact: mpetersen@lanl.gov, warndt@lbl.gov, ndkeen@lbl.gov

Summary: NESAP performance improvements on MPAS-Ocean have resulted in a 5% to 7% speed-up on each of the examined systems including Cori-KNL, Cori-Haswell, and Edison. These tests were configured to emulate a production workload by using 128 nodes and a high-resolution ocean domain. Overall, the gap between standard and many-core architecture performance has been narrowed, but Cori-KNL remains considerably under-performing relative to Edison. NESAP code alterations affected 600 lines of code, and most of these improvements will benefit other MPAS codes (sea ice, land ice) that are also components within ACME. Modifications are fully tested within MPAS. Testing in ACME across many platforms is underway, and must be completed before the code is merged. In addition, a ten-year production ACME global simulation was conducted on Cori-KNL in late 2016 with the pre-NESAP code in order to test readiness and configurations for scientific studies. Next steps include assessing performance across a range of nodes, threads per node, and ocean resolutions on Cori-KNL.

1 Scientific Motivation

The Model for Prediction Across Scales, MPAS-Ocean, is a variable-resolution mesh global ocean model designed for climate change research (Ringler et al., 2013; Petersen et al., 2015). MPAS-Ocean is a component of the DOE's new Accelerated Climate Model for Energy (ACME). The ability to run high-resolution global simulations efficiently on large, high-performance computers is a priority for ACME simulations. ACME includes active ocean, sea ice, and land ice, which are all MPAS-based components, as well as atmosphere and land components. ACME research and model development goals are driven by three key science questions:

1. **Water Cycle:** How do the hydrological cycle and water resources interact with the climate system on local to global scales?
2. **Biogeochemistry:** How do biogeochemical cycles interact with global climate change?
3. **Cryosphere-Ocean System:** How do rapid changes in cryosphere-ocean systems interact with the climate?

Further details about ACME, the scientific and computing goals may be found at the ACME home page.

A new capability in MPAS-Ocean, pertinent to the cryosphere-ocean science question, is the ability to simulate the ocean below ice shelves in Antarctica (Fig 1). Ice shelf–ocean interactions are important to the global climate. Warmer ocean currents may speed up ice shelf melting and retreat. At the same time, changing land ice fluxes could affect ocean temperature, salinity, and currents below ice shelves, altering Southern Ocean water mass formation.

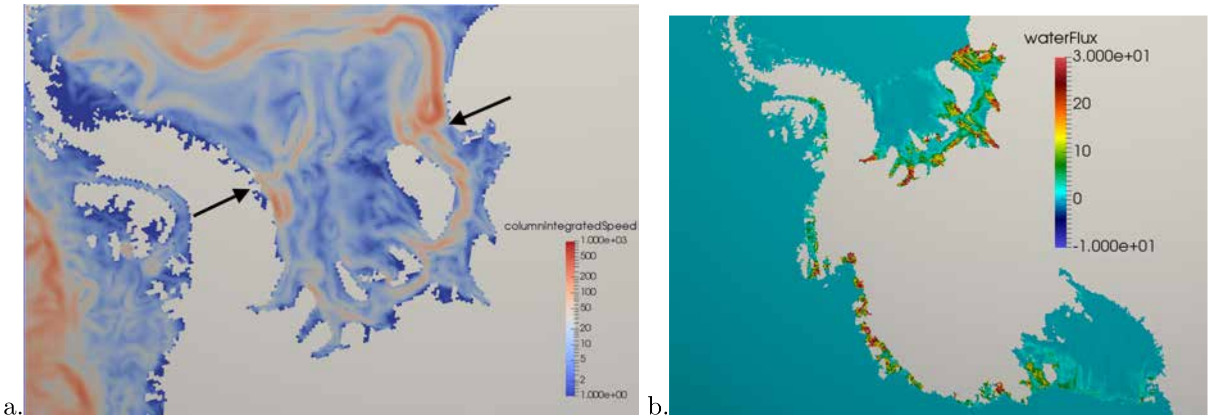

Figure 1: A view of Antarctic sub-ice shelf statistics after ten years of the ACME Cori-KNL simulation: (a) Column-integrated speed [m^2/s] shows ocean currents that extend from the open ocean to below the ice shelf in Antarctica. The black arrows indicate the edge of the Filchner-Ronne ice shelf. (b) Rate of ice melt from the bottom of the ice shelf into the ocean, in meters of water per year.

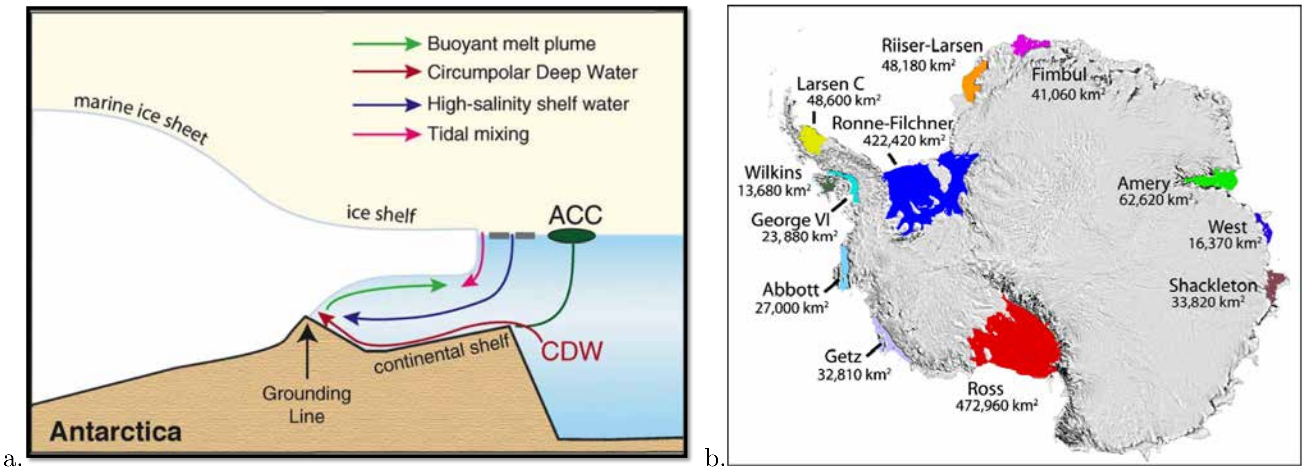

Figure 2: (a) Cross-section of an Antarctic ice shelf, where warm Circumpolar Deep Water (CDW) moves up the continental shelf and melts the ice near the grounding line. This produces a fresh, buoyant plume that ascends along the bottom surface of the ice shelf and influences the formation of Antarctic Bottom Water. Image credit: Joughin et al. (2012) (b) Locations of ice shelves in Antarctica. These are all floating on ocean water, and are expected to melt and retreat with warmer ocean currents. Previously, ocean model domains ended at the ice shelf edge and did not include water underneath. Credit: Ted Scambos, NSIDC

The Ronne-Filchner and Ross Ice Shelves sit on top of areas of ocean, each at least the size of Texas (Fig. 2). Despite this, ice shelf cavities have not been included in any fully coupled global climate model to date because of the numerical modeling challenges and lack of observational data for validation. ACME is one of the first climate models to include this unique capability, and will provide insight to physical processes that affect sub-ice shelf currents, at the same time that new observational data comes in from field campaigns.

An ACME global climate simulation with ice shelf cavities was run on Cori-KNL in November and December of 2016 by Noel Keen and Mark Petersen. This was one of the first high-resolution simulations to be run with the new ice shelf cavity feature, and is the culmination of 18 months of effort to add this capability. As such, it allowed scientists to evaluate the initialization and spin-up process of waters and melt rates below ice shelves using high resolution (Fig. 1, 3). The global domain (RRS30to10) has 100 vertical layers and 1.4 million horizontal cells, varying from 10 km to 30 km in diameter, with a coupled MPAS-Sea Ice model and forced by atmospheric data. This simulation was an important step in testing ACME on a new platform, Cori-KNL, for operation, performance testing, performance improvements, and improving partitioning of nodes among components.

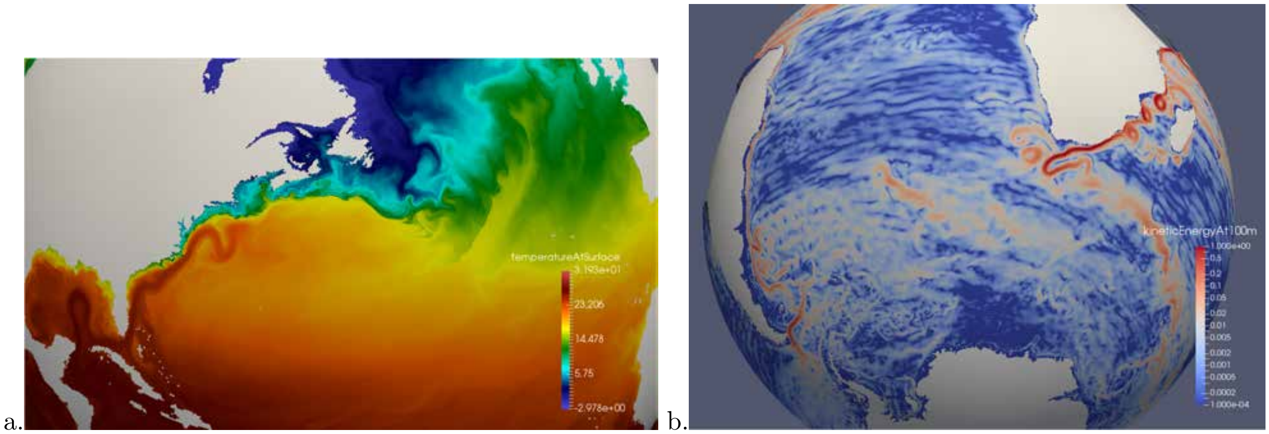

Figure 3: (a): Sea surface temperature [C] of the North Atlantic near the end of the Cori-KNL ACME simulations, showing the warm Gulf Stream along the East Coast of the U.S. (b): Kinetic energy at 100 m depth [m^2/s] in the Southern Ocean. The Agulhas current is visible along the eastern coast of Africa, with eddies pinching off the Agulhas retroflection. The Southern Ocean has substantial mesoscale eddy activity.

2 Performance and Optimization

The work described in this report prepares the ACME team for performance improvements and simulations in the coming years. A suite of standard ACME simulations are in progress or testing at DOE Leadership Class Facilities, including on Edison, Titan, Cori-Haswell, Cori-KNL, and Mira. Although each machine has particular challenges, improvements in performance and threading on one machine usually translate to improvements on others. In addition to a large number of standard ACME simulations, a research focus on ice shelf-ocean interactions was proposed in an ALCC led by Mark Petersen of LANL for 150 million CPU-hours on edison, titan, cori, and theta, based on experience from the simulations described in Section 1.

Substantial emphasis on performance in the design of MPAS-Ocean has led to good scaling using MPI-only communication (Fig 4). Tests on Edison in 2016 showed that MPAS-Ocean could scale well to 50 thousand cores for high-resolution domains. However, preliminary tests on CORI-KNL showed throughput of half to a quarter of that on Edison. This motivated the work on threading improvements by Abhinav Sarje of LBNL in 2016, and continued by Bill Arndt in 2017.

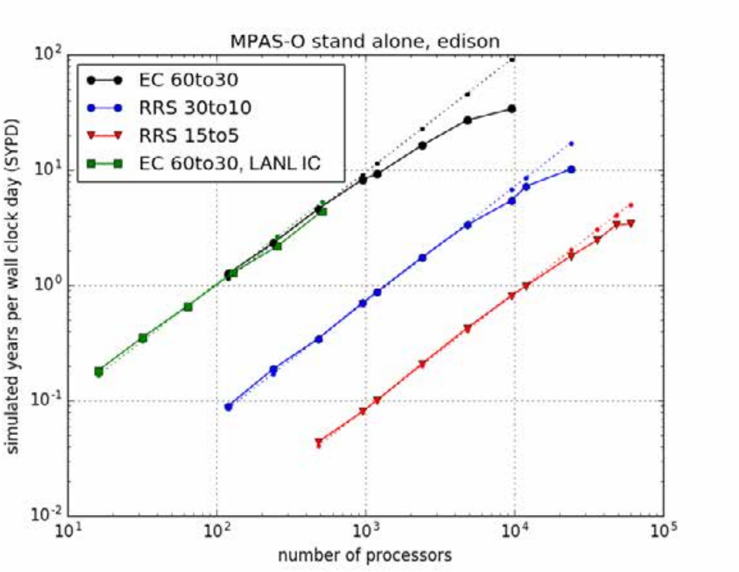

Figure 4: Performance measurements of MPAS-Ocean on Edison in early 2016 for three resolutions ranging from low (black) to high (red). Dotted lines show perfect scaling. These simulations use MPI-only communication, before the improvements described in this report.

2.1 Details of experiments to evaluate performance

Evaluation setup was as follows:

- Configuration: Stand alone MPAS-Ocean core built using prescribed before and after NESAP code base along with chosen compiler flags to target architecture specific instructions. KNL runs use quad-cache memory mode, 67 tasks per node each with 2 threads grouped by core. Haswell and Ivy Bridge processors are respectively given 16 and 12 tasks per node each with 2 threads. All tests used 128 nodes of their respective system.
- Resolution: Grid cells vary from 30 km at the equator to 10 km at high latitude, with 1.4 million horizontal cells and 100 vertical layers. This is the global RRS30to10 resolution used for the simulations in Section 1.
- Duration: Between 8 and 10 runs, using 8 minute intervals, completing 4 integration steps. Reported times are the average of all integration steps. Out lier runs sometimes occur due to interference from other jobs on the system; those runs were removed from consideration.

2.2 Baseline (pre-NESAP) and Optimized Code Walltime on Cori KNL Nodes

This table shows the baseline performance of the initial code along with relative advantage of each system:

	Edison	Haswell	KNL
Seconds per integration step	0.618	0.692	0.954
Performance relative to Edison	1.000	0.893	0.647

For this test, Cori-Haswell is 10% slower than Edison, and Cori-KNL is 35% slower than Edison.

2.3 Description of NESAP optimizations

The optimizations for MPAS-Ocean developed by Abhinav Sarje of LBNL in 2016 include:

1. Implementation of threading into the MPAS reconstruct routine.
2. Changing MPI threading level from multiple to funneled.
3. Reorganization in buffer pack and unpack in halo exchanges to minimize use of barriers.
4. Implementation of threaded memory buffer initializations.
5. Loop restructuring to reduce cache misses.
6. Code optimizations to reduce excessive computations.
7. Threading optimizations to ensure threads don't redundantly perform operations on arrays.

This resulted in a pull request into the MPAS-Ocean github repository in December 2016 with the alteration or addition of 600 lines of code. Due to the requirements of the MPAS merge process, this pull request was separated between bit-for-bit reproducible and answer changing code, as well as MPAS framework versus MPAS-Ocean parts, by Mark Petersen in early 2017. These have all been successfully tested in MPAS-Ocean, and testing and debugging within ACME on numerous platforms is currently underway, with assistance from Philip W. Jones of LANL.

Of the 600 lines of code, 80% are within the MPAS framework, which is used by all MPAS cores: ocean, sea ice, land ice (all DOE-LANL), and atmosphere (NSF-NCAR). The MPAS framework includes all functionality common to all cores, such as the halo exchanges in (3). Thus we expect NESAP work to have carry-over improvements to other models, though only MPAS-Ocean has been assessed to date.

2.4 Performance results when evaluating code changes and hardware features

This table shows the optimized code performance on each platform, relative performance to baseline code on the same platform, and performance relative to the base code on Edison:

	Edison	Haswell	KNL
Seconds per integration step	0.586	0.648	0.908
Performance relative to base	1.054	1.068	1.051
Performance relative to Edison	1.054	0.954	0.681

This shows that the NESAP improvements resulted in a 5–7% speed-up on all systems.

The following table shows the effect of KNL hardware features on code performance:

	Feature disabled	Feature included	Relative impact
KNL performance using AVX512 (vs. AVX2)	0.944	0.908	1.040
KNL performance using MCDRAM cache (vs. DDR)	1.090	0.908	1.201

These KNL-only features result in a performance impact of 4% and 20%, respectively.

2.5 Future Work

Next steps in the evaluation of MPAS-Ocean are to produce performance plots similar to Figure 4 to see if NESAP improvements affect scaling, and to investigate optimal thread count per core on Cori-KNL. The OpenMP configuration of MPAS-Ocean had previously not been extensively tested on a wide range of machines, so this work includes debugging and improving the threaded code across numerous architectures and compilers. The final step for code merge into MPAS-Ocean is for ACME to pass all tests on all standard platforms. This includes a bit-for-bit comparison across different thread counts, which often reveals new issues. Once the current pull requests are fully tested and merged, we will evaluate performance data for bottlenecks and load imbalance, in order to make further improvements.

References

Joughin, I., Alley, R., Holland, D., 2012. Ice-sheet response to oceanic forcing. Science 1172.
URL http://www.sciencemag.org/content/338/6111/1172.short

Petersen, M., Jacobsen, D., Ringler, T., Hecht, M., Maltrud, M., 2015. Evaluation of the arbitrary Lagrangian-Eulerian vertical coordinate method in MPAS-Ocean. Ocean Modelling 86 (0), 93 – 113.
URL http://www.sciencedirect.com/science/article/pii/S1463500314001796

Ringler, T., Petersen, M., Higdon, R., Jacobsen, D., Jones, P., Maltrud, M., 2013. A multi-resolution approach to global ocean modeling. Ocean Modelling 69 (0), 211–232.
URL http://www.sciencedirect.com/science/article/pii/S1463500313000760