Appendix 3
NGGPS Level 1 Test Plan

Advanced Computing Evaluation Committee

Chair: John Michalakes, NOAA (IMSG)
Co-chair: Mark Govett, NOAA/ESRL
Rusty Benson, NOAA/GFDL
Tom Black, NOAA/EMC
Alex Reinecke, NRL
Bill Skamarock, NCAR

Background and Purpose

The Advanced Computing Evaluation Committee (AVEC) was formed in August, 2014 to provide Level-1 technical evaluation of HPC suitability and readiness of NGGPS candidate models to meet global operational forecast requirements at NWS through 2025-30. This document describes the Level-1 test plan for benchmarking and evaluation of computational performance, scalability, and HPC software design and reporting back to the NGGPS program in spring 2015. This test plan provides details of the benchmarking methodology, cases, model configurations, computational resource requirements, schedule, and results to be reported.

Benchmark Cases and Model Configurations

Two sets of benchmarks will be run: performance and scalability. The performance benchmark will measure speed of each candidate model running a near-future workload representing the cost of non-hydrostatic dynamics, including advection, running operationally beginning in 2015.

The scalability benchmark will measure how efficiently each candidate model is able to employ additional processors to run significantly more challenging workloads representing the cost of high-resolution non-hydrostatic dynamics and advection expected to be routine within 10 years.

The benchmarks will be conducted using the idealized baroclinic wave case with monotonically constrained scalar tracer advection, similar to the HIWPP configurations but with the following additional features:

1. The case will include ten extra 3D tracer fields initialized to a checkerboard pattern on the sphere to ensure that the cost of the monotonic constraint is represented in the benchmark workload. The detailed algorithm for initializing the tracers will be the subject of further discussion and agreement by the modeling groups.

2. Two horizontal resolutions (nominally 13km and 3km) on the full sphere will be benchmarked using 128 vertical levels. The resolution shall be as close as possible to target resolution.
3. Each group should chose a time step that is their best estimate of what they would use for a real-data forecasting case at each resolution. Rescaling of timing results may be done after the fact if the time step used when actual 3km real data cases have been run deviates from best-guess time step used during for the Level-1 benchmarks.
4. For verification, each group will provide a reference solution at each of the resolution. The benchmark solutions will be evaluated for correctness by calculating differences with these reference solutions.
5. Duration of integrations (subject to computational resource availability): 30 minutes for the high-resolution case and 2 hours for the low resolution case.

Each candidate model's configurations – resolution, number of points, number of levels, and time step – for the two benchmarks has been reviewed and agreed upon by the other modeling groups. The configurations are listed in Table A3-1.

Benchmark Readiness

Each team will provide files, data, and scripts sufficient for benchmarkers to compile, run, and verify their model's test cases in rapid fashion during the benchmark period. AVEC will provide instructions to model teams on how to prepare their codes and data sets for benchmarking and evaluation and will work with the teams to conduct pre-benchmarking tests on smaller numbers of processors to ensure the full benchmark testing goes smoothly and within the allotted machine access times.

Model teams will generally use their own HPC resources for development and testing with smaller workloads, but non-dedicated access to larger partitions and time allocations on large the benchmark systems is also planned (under discussion with HPC centers).

Final Benchmark Methodology

Benchmarks will be conducted in at least two sessions of dedicated access to a large system at one of the centers listed under Computational Resources, below.

Performance: For the 13 km resolution performance benchmarks, each model will be run starting on about 1000 cores and then over successively larger numbers of processors until it achieves an integration rate for dynamics and advection required in the full-physics NGGPS to run at the operationally-required 8.5 minutes per day. The starting, ending and incremental numbers of processors will be determined during benchmark readiness phase of this work plan. These may differ from model to model to accommodate different parallelization and other implementation details.

Scalability: For the 3km resolution scalability benchmarks, each model will be run starting on a minimum number of processors and then over successively larger numbers of processors until either performance has stopped increasing or the maximum number of processors has been reached. Both raw integration rate and scaling efficiency will be reported. Scaling efficiency is defined as:

$$E = (T_{np\_base} / T_{np\_tested}) \ / \ (np\_tested / np\_base)$$

where *T* is elapsed time (compute-only), *np_base* is the baseline (starting) number of processors and *np_tested* is the number of processors used in a given run. Ideally, E will be one.

As above, the incremental numbers of processors will be determined during the benchmark readiness phase, and may differ from model to model. The starting number of processors will be the maximum over all models of the minimum number of nodes the model fits in memory running the 3km workload.

In addition to computational scaling, the memory scaling of the models will also be measured by instrumenting the models with the UNIX getrusage() library routine or similar.

For both sets of benchmark, there will be a three replications of each benchmark.

Reporting

The final Level-1 Benchmarking report will provide data and performance and scalability analysis that supports ranking of candidate model results and subsequent decision making by the NGGPS program and NWS management.

For both lower-resolution performance benchmarks and the higher-resolution scalability benchmarks, the raw timings (wall clock seconds average time step) and simulation speed (wall clock seconds per simulation interval) from each benchmark run will be provided in tabular form and plotted graphically. Simulation speed will be based on the time step used by the candidate models in the performance benchmark runs, but simulation speeds may be scaled upwards or downwards to allow for adjustment of the time step based on subsequent real data tests.

In addition to benchmark results, the AVEC will compile data sheets for each candidate core that includes basic characteristics of the core (numerical formulation, discretization) and technical implementation details including software design (modularity, extensibility, readability, maintainability) and performance-portability, especially with respect to next-generation NOAA HPC architectures and system configurations (decomposition and parallelization strategy, communication patterns, supported programming models, etc.).

Computational Resources

Benchmarks will be conducted on large homogeneous partition of a supercomputing system provisioned with on the order of 100-thousand conventional Intel Xeon processor cores (Sandy Bridge, Ivy Bridge, or Haswell, but not mixed). Any compiler, library, or other requirements shall be specified well enough in advance to ensure their availability on the benchmark system. Discussions are underway for use of one or more of the following supercomputing systems.

- NSF: Stampede. Texas Advanced Computing Center (TACC) at U. Texas at Austin
    - 102,400 cores over 6,400 dual Xeon E5-2680 (Sandy Bridge) nodes (16 cores per node), each with 32 MB
    - FDR InfiniBand 2-level fat tree interconnect
    - https://www.tacc.utexas.edu/user-services/user-guides/stampede-user-guide
- DOE: Edison. National Energy Research Scientific Computing Center (NERSC) at Berkeley National Laboratory.
    - 133,824 cores over 5,576 dual Xeon Ivy Bridge nodes (24 cores per node)
    - Cray Aries with Dragonfly topology
    - https://www.nersc.gov/users/computational-systems/edison/configuration
- NASA: Pleiades. NASA/Ames Research Center
    - 108,000 cores over 5,400 dual Xeon Ivy Bridge nodes (20 cores per node)
    - Possibility of ~100,000 cores of Xeon Haswell by benchmarking time
    - Dual plane 10D hypercube with InfiniBand interconnect
    - “Dedicated access” to Pleiades will mean to an uncontended section of the hypercube but not exclusive access to whole machine
    - http://www.nas.nasa.gov/hecc/resources/pleiades.html

Schedule

- October 1, 2014.
    - Test Plan Approved
    - Computational centers contacted and initial approvals for resource availability.
- October 31, 2014.
    - AVEC completes instructions for benchmark codes and data and provides to Model Teams
- November 30, 2014.
    - Model groups provide initial codes and data sets.
    - Computational resources finalized and available for benchmark readiness activity.
    - Model groups and AVEC test and prepare benchmark codes and datasets.
- February 15, 2015. Final suite of benchmark codes ready.
- March-April, 2015.

    - Two benchmarking sessions conducted on dedicated HPC resources.
    - Benchmarks completed.
- April 30, 2015. Final report

Acknowledgements

Nicholas Wright, NERSC. Bill Barth and Tommy Minyard, TACC. William Thigpen, Cathy Schulbach, and Piyush Mehrotra at NASA/AMES.

		NH-GFS (Baseline) *	FV-3	MPAS	NIM	NMM-B	NEPTUNE
Nominally 13km	**Resolution**	13 km (TL1534)	13km (C768)*	12km *	13.4 *	13 km *	12.5 km *
	Grid Points	3072x1536 (unreduced) 3,126,128 (reduced)	6x768x768 3,538,944	4,096,002 **	3,317,762	2,179x1,541 3,357,839 **	3,840,000 **
	Vertical Layers *	128	127 **	127 ***	128	128	128 ***
	Time Step	TBD	600s (slow phys) 150s (vertical, fast phys) 150/11 (horiz. acoustic)	72 s (RK3 dynamics) 12 s (acoustic) 72 s (RK3 scalar transport)	72 s	25 s ***	60 s (slow RK3 dyn.) 10 s (fast dyn.) ****
Nominally 3km	**Resolution**	3 km (TL6718)	3.25 km (C3072) *	3km	3.3 km **	3 km ****	3.13 km *
	Grid Points	13440x6720 (unred.) 59,609,088 (reduced) **	6x3072x3072 56,623,104	65,536,002	53,084,162	9,601x6,673 62,973,101 **	61,440,000 **
	Vertical Layers *	128	127 **	127 ***	128	128	128
	Time Step	TBD	150 s (slow phys) 37.5 s (vertical, fast phys) 37.5/11 s (horiz. acoustic)	18 s (RK3 dynamics) 3 s (acoustic) 18 s (RK3 scalar transport)	18 s	6 s ***	15 s (slow RK3 dyn.) 2.5 s (fast dyn.) ***
	Notes * Unless noted, layers refers to the number of layers, not the number of interfaces between layers + top + bottom	* Baseline configuration is tentative, pending test evaluation. ** Rough estimate for reduced Gaussian grid based on reduction factor (0.66) of 13 km grid. This will likely be revised after further testing of accuracy of spectral transform at TL6718.	* True resolution is average over equator and/or from south to north pole. For 13km, max cell size (edge of finite volume): 14.44 km, min: 10.21 km, global avg: 12.05 km. For 3.25 km, divide by 4. ** Favorable OpenMP Performance	* Resolution refers to mean cell-center spacing on the mesh ** Subdivision of 60 km mesh by factor of 5. *** Following the FV3 configuration, we will use 127 levels where density, theta and horizontal momentum are defined (on our Lorenz-grid vertical discretization) and 128 levels for w (that includes both the lower boundary and the model top "lid").	* Generated by 6 bisections followed by 2 trisections. Distances between neighbors: 13.367 average, 12.245 min., 14.397 max.. Maximum ratio of neighboring grid point distances: 1.17577 ** Generated by 8 bisections followed by 2 trisections. Distances between neighbors: 3.3417 average, 3.060 min., 3.601 max.. Maximum ratio of neighboring grid point distances: 1.1765.	* Dy = 12996.81 m, avg. Dx = 9189.663 m, min. representable wavelength at equator = 22511.135 m. ** B-grid mass points *** For fast modes and advection of basic model variables. Time step for tracers is longer by 2x. **** Dy = 2999.863 m, avg. Dx = 2121.141 m, min. representable wavelength at Equator= 5195.915 m.	* Average nodal spacing per element. For 4th-order polynomials: ~12.5 km horizontal resolution will use 200 elements per edge of the cube sphere (grid can use 240,000 cores); ~3.13 km horizontal resolution will use 800 elements per edge (grid that use up to 3,840,000 cores). ** Horizontal grid points is six faces of cube times number of elements per face times polynomial order squared. *** Estimates are for split-explicit. May also use 3d- or 1d-imex method, with ab3/ai2 time integrator for expl./impl. step.

Table A3-1. Model-specific Benchmark Configurations