AVEC Report:
NGGPS Level-1 Benchmarks and Software Evaluation

Advanced Computing Evaluation Committee

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

Contributors

Michael Duda, NCAR
Thomas Henderson, NOAA/ESRL (CIRA)
Paul Madden, NOAA/ESRL (CIRES)
George Mozdzynski, ECMWF
Ratko Vasic, NOAA/EMC

Submitted: April 30, 2015

I. Introduction

The Advanced Computing Evaluation Committee (AVEC) was formed in August, 2014 to provide Level-1 technical evaluation of HPC suitability and readiness of five Next Generation Global Prediction System (NGGPS) candidate models to meet operational forecast requirements at the National Weather Service through 2025-30. This report describes methodology, cases, model configurations, and results of performance and scalability benchmarks conducted during two sessions on Edison, a 130-thousand processor core supercomputer at the U.S. Department of Energy's National Energy Research Scientific Computing Center (NERSC)[1], during March and April, 2015. This testing is part of the NGGPS Test Plan.[2]

Two benchmark test cases were developed and agreed to by the scientific leads of the five NGGPS modeling groups. One workload was sized to measure *performance*: specifically, the computational resources needed for a model to meet a given operational forecast time-to-solution requirement of 8.5 minutes per forecast day. A considerably larger case was developed to measure *scalability:* the model's ability to use increased numbers of processors to

[1] https://www.nersc.gov
[2] http://www.nws.noaa.gov/ost/nggps

run larger workloads that might be expected in the ten- to fifteen-year lifetime of NGGPS. Input datasets and model dynamical core (dycore) test codes were prepared by the individual modeling groups. These were then handed off to AVEC for final testing and benchmarking. In addition to the five NGGPS candidates, we were fortunate to include ECMWF's Integrated Forecast System (IFS)[3] as a guest dycore. Finally, AVEC collected and summarized the benchmark results for this report.

Our report is provided with the following caveats:

- The performance and scaling results in this report are a snapshot in time of NWP software that is under active development. The test workloads are based on an idealized atmospheric case that does not include physics.
- Dedicated access to a 130-thousand core supercomputer is a precious commodity that required emphasizing coverage rather than replication of runs. More replication would better address variability observed in the results from run-to-run and from time step to time step. However, we believe sample sizes were adequate and have discarded obvious outliers in the results.
- The choice of time step for the idealized benchmark runs was best-guess of what would be needed for full-physics real-data forecasts on the part of the modeling groups. In adjusting benchmarking results to the operational speed requirement, we also assumed that dynamics represents half the run time of a full-physics model.
- AVEC did not evaluate important aspects of performance such as I/O, initialization costs, or other factors that would not represent full physics realizations of the models. Such testing will occur in future Level-2 evaluations under the NGGPS test plan.
- AVEC did not evaluate model performance with respect to any objective or subjective measures of solution quality. Each candidate model's benchmarks were conducted with the same formulation and configuration used to run the idealized test cases under the High Impact Weather Prediction Program (HIWPP) non-hydrostatic dycore evaluations.[4]

The sections that follow describe methodology; models, workloads and configurations; a chronology; and detailed results. The last section is a summary. Figures and tables referred to in the report appear at the end. The report is the AVEC's consensus.

II. Methodology

Two sets of benchmarks were run: performance and scalability. The performance benchmark measured speed of each candidate model running a workload representing the cost of non-hydrostatic dynamics, including advection, that could be run operationally on current or near-

[3] http://www.ecmwf.int/en/research/modelling-and-prediction/atmospheric-dynamics
[4] "HIWPP Non-hydrostatic Dynamical Core Tests: Results from Idealized Test Cases", Jeffrey Whitaker, NOAA/ESRL/PSD, report to HIWPP and NGGPS program management (26 pages).

future NOAA systems. The scalability benchmark measured how efficiently a model was 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.

Performance: For the 13 km resolution performance benchmarks, each model was run on a set of increasing processor core counts beginning with the smallest number of cores on which the model could run end-to-end in a 5 minute envelope. The highest was the number of cores needed compute two hours of simulation in under 21.25 seconds, the threshold operational speed requirement (see Section V). Additional higher core counts were added for the second round of benchmark tests to provide information on strong scaling behavior of the models running the 13 km workload.

Scalability: For the 3km resolution benchmarks, each model was run starting on the minimum number of processors needed to run end-to-end in under 15 minutes, then over successively larger numbers of processors until either performance stopped increasing or a maximum number of processors had been reached. The hard maximum was the number of processor cores available on Edison (130-thousand), but the limit could be lower if an individual model's task decomposition involved steps of increasing processor counts and the next higher step was greater than the number of cores on Edison.

Concurrent execution: In order to fit all scheduled runs into the allotted benchmark time on Edison, it was necessary to run the smaller 13 km benchmarks concurrently across the machine. To minimize effect on timing results of contention between the concurrently executing benchmarks, each model run was run three times at each core count and the minimum elapsed time recorded. All of the 3 km benchmarks were run by themselves on an otherwise quiescent Edison system, but these could be run only once in order to stay within the time allotted for the session.

Verification

As each benchmark ran, one or a small set of fields (e.g. surface pressure) was collected and then verified – either within the model or from files at the end of the run – by comparing against reference output and a statistical "pass/fail" test provided to AVEC by the modeling group. Validation of the models was performed by the modeling groups prior to the models being handed off to AVEC for testing.

Computational Resources

Accounts for AVEC members and 4-million core hours of computer time for benchmark development, testing and for two benchmarking sessions, 8 hours and 4 hours of dedicated full-machine access, were provided on Edison, NERSC's newest system:

- 133,824 cores populating 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

Accounts for AVEC members and allocations of core hours were also provided for benchmark development and testing by:

- NSF: Stampede. Texas Advanced Computing Center (TACC) at U. Texas at Austin
  - o 102,400 cores populating 6,400 dual Xeon E5-2680 (Sandy Bridge) nodes (16 cores per node), each with 32 MB
  - o FDR InfiniBand 2-level fat tree interconnect
  - o https://www.tacc.utexas.edu/user-services/user-guides/stampede-user-guide
- NASA: Pleiades. NASA/Ames Research Center
  - o 108,000 cores populating 5,400 dual Xeon Ivy Bridge nodes (20 cores per node)
  - o Dual plane 10D hypercube with InfiniBand interconnect
  - o http://www.nas.nasa.gov/hecc/resources/pleiades.html

The Dependency Driven Test System,[5] a robust and flexible automated test harness developed by our co-author Paul Madden, proved essential for simplifying setup and pre-benchmark testing and helped provide widest possible test coverage during the two benchmarking sessions on Edison.

III. Models, Workloads, and Configurations

Table 1 lists the five candidate non-hydrostatic dycores: the Non-hydrostatic Icosahedral Model (NIM) from the NOAA/ESRL, the Model for Prediction Across Scales (MPAS) from NCAR, NEPTUNE from the U.S. Naval Research Laboratory (NRL), FV3 from NOAA/GFDL, and the uniform-Jacobian version of the Non-hydrostatic Multiscale Model (NMM-UJ) from NOAA/NCEP. Also listed are the non-hydrostatic Global Forecast System (GFS) from NOAA/NCEP, which is intended to serve as the baseline model in NGGPS testing and evaluation, and a guest core, the hydrostatic Integrated Forecast System (IFS) model from the European Centre for Medium Range Weather Forecasts (ECMWF). Since no version of GFS code or data was provided to AVEC, the IFS served as a proxy in the Level-1 benchmarks.

The benchmarks were conducted using two workloads comprising an idealized baroclinic wave test with the addition of monotonically constrained scalar tracer advection, similar to the HIWPP configuration but with the following additional features:

[5] Madden, Paul, and Eduardo G. Valente Jr. "DDTS: A Practical System Testing Framework for Scientific Software." arXiv preprint arXiv:1410.8844 (2014). URL: http://arxiv.org/abs/1410.8844

- The cases included ten extra 3D tracer fields initialized to a checkerboard pattern on the sphere to include the cost of monotonically-limited advection in the benchmark workload (Figure 1).
- The workloads specified two horizontal resolutions (nominally 13 km and 3 km) on the full sphere with 128 vertical levels.
- Each group chose a time step that was their best estimate of what they would use for a real-data forecasting case at each resolution.
- Each group provided a reference solution for use in verifying output generated during the benchmark runs.
- The low resolution case was configured to run a 2 hour simulation. The high resolution case was configured to run a 30 minute simulation.[6]

Each candidate model's configurations – resolution, number of points, number of levels, and time step – for the two benchmarks were reviewed and agreed upon by the other modeling groups. The configurations are listed in Table 2.

IV. Chronology

The following is a brief chronology of the AVEC Level-1 NGGPS benchmark effort:

- November, 2014
  - Instructions and criteria for benchmarks were given to Model Teams.
- December, 2014
  - Model groups submitted initial codes and data sets to AVEC.
  - HPC resources were committed at NERSC (4 million core hours), at TACC (600-thousand core hours for development and testing), and at NASA.
- February, 2015
  - NMM-B was swapped-out for NMM-UJ (6 February) (HIWPP cases rerun).
  - Final suite of benchmark codes was ready on 15 February.
- March, 2015
  - First (8 hour) benchmarking sessions was completed at NERSC on 6 March.
  - FV3 was switched to single precision for Round-2 tests, 10 March.
- April, 2015
  - NEPTUNE was switched from 4th to 3rd order (10 April) (HIWPP cases rerun).
  - The second (4 hour) benchmarking session was completed at NERSC (11 April).
- Final report

[6] IFS ran a two hour simulation of the 3 km case and an eight hour simulation of the 13 km case. The resulting timings were then divided by four.

V. Results

This section presents performance and scaling results of the NGGPS candidate model dycores running the small (13 km; Table 3) and large (3 km; Table 4) workloads on the NERSC Edison system during the benchmarking sessions conducted during an 8 hour session on March 6, 2015 (Round 1) and during a four hour session on April 11, 2015 (Round 2). For the nominally 13 km resolution test case, performance and strong scaling efficiency is plotted as a function of increasing processor core counts (Figure 3 and 4) and as the number of processing cores required to meet a predefined operational speed requirement (Figure 2). For the nominally 3 km resolution test case, performance and strong scaling are plotted as elapsed time and strong scaling efficiency in Figures 5 and 6. Figure 7 shows weak scaling efficiency as the problem size is increased from 13 km to 3 km horizontal resolution.

Performance and Strong Scaling for the 13 km workload

The NGGPS test plan specified 8.5 minutes per forecast day as the operational forecast speed requirement. Since the Level-1 benchmarks were for dynamics only (no physics), we assumed dynamics would comprise half the model run time and interpreted the speed requirement as the ability to run the 2 hour benchmark in 21.25 seconds, excluding time spent doing I/O and initialization. A horizontal dashed line indicates the speed requirement on the dependent axis of Figure 3, which shows both Round 1 and Round 2 test results.[7]

The Round 2 benchmarks afforded modeling groups the opportunity to improve on results from Round 1. In some cases (IFS, NIM, FV3 and MPAS), groups chose to run with additional, higher processor counts in Round 2. Some chose to test enhancements to the codes that did not change model output: both the NMM-UJ and NIM groups tested improvements to MPI communications. Two groups, FV3 and NEPTUNE, tested changes that had the potential to change the solution relative to that generated previously for the HIWPP non-hydrostatic idealized test cases. The FV3 group tested the effect of running with single-precision floating point arithmetic instead of double precision used during Round 1 (see subsection below for additional discussion on floating point precision). The NEPTUNE group used Round 2 to test changes that reduced computational cost, including the use of 3rd order in place of 4th order. Both the FV3 and NEPTUNE groups conducted retests of the HIWPP cases of their models.

All of the candidate dycores were able to meet the speed requirement of 21.25 seconds for a two-hour simulation (Figure 2). Averaging over the non-hydrostatic models tested (excluding the hydrostatic IFS) the number of processor cores required to meet the requirement was 16,695 and the median was 11,076. The NMM-UJ was fastest followed by the FV3 model

[7] The highest core-count for NIM running the 13 km benchmark was 123,210 cores. The run completed and verified but the timing was adversely affected by contention from other jobs on the Edison system near the end of the Round 2 benchmarking period. The 123,210 core data point is omitted from the 13 km performance and scaling figures.

running single-precision. NIM and the double-precision run of FV3 were essentially tied. Both MPAS and NEPTUNE required well above the average numbers of cores to meet the operational requirement.

Figure 4 shows strong scaling of the models relative to the fourth highest core count in each model run. Timings from the lowest core counts were excluded from this strong scaling plot to avoid showing super-linear scaling. Super-linear scaling occurs when program speed increases by a factor greater than the increase in number of processors. Super-linear scaling may result from a true and beneficial behavior of the program (e.g., a better partitioning at larger processor counts) but also may indicate a resource bottleneck at lower processor counts (e.g., per-processor program image not fitting in cache) rendering comparison between lower and higher processor counts invalid. For purposes here, we disregarded super-linear scaling to the extent possible. NEPTUNE scaling was systematically super-linear in both the 13 km and 3 km results (Figure 4 and 6).

All models scaled adequately in reaching the 21.25 second per two-hour simulation performance threshold. Beyond that threshold, the models scalability rankings were the reverse of their performance rankings (this was also generally true of the 3 km benchmark results). This could be expected. Models that are more expensive likely scaled better because they perform more work per processor to offset the cost of communicating to other processors. This was also evident in FV3's higher performance but lower scalability at single-precision compared to double precision.

Performance and Strong Scaling for the 3 km workload

Performance for the larger 3 km resolution workload is plotted as elapsed time to run a 30 minute forecast (Figure 5). Here, the object was to measure strong scalability running up to the largest possible number of cores. Both axes of the plot are logarithmic. Scaling efficiency relative to the fourth largest core count is plotted in Figure 6.

As with the 13 km results, the hydrostatic IFS provided the best 3 km global forecast performance, running the 30 minute benchmark in just over 10 seconds (180x real time[8]) while scaling from 6,144 to 129,024 processor cores with 61 percent efficiency.

For the 3 km workload, all the non-hydrostatic dynamical cores scale well. None of the Round-1 benchmarks scaled below 80 percent efficiency. Single precision FV3 benchmark scaling tested in round two was slightly lower, at 78 percent efficiency. NEPTUNE (40,000 to 128,000 cores) and MPAS (24,000 to 131,072 cores) both scaled the best at 96 percent efficiency. NIM

[8] A simulation rate that is 340 times real time is needed to reach the 8.5 minutes per day speed threshold, assuming that dynamics is half the cost of a model run. The eventual actual ratio of dynamics to physics and the length of the time step needed for full-physics realizations of the models were speculative.

scaled with 91 percent efficiency from 16,032 to 128,040 cores [9], followed by double-precision FV3 (88 percent from 18,432 to 110,592 cores), NMM-UJ (83 percent from 13,824 to 98,304 cores), and FV3 single-precision (78 percent from 18,432 to 110,592 cores).

NMM-UJ provided the best 3 km performance of the non-hydrostatic candidate models (128x real time). The single precision run of FV3 was next (77x real time). The double precision run of FV3 and NIM were third fastest (48x [10]), followed by MPAS (22x), and NEPTUNE (6x).[11]

Weak Scalability

Weak scaling is the ability to do more work with more processors, as opposed to strong scaling which is the ability to do work faster with more processors. With weak scaling, the number of grid cells per processor and, ideally, program speed remain constant as problem size and number of processors are increased. Unfortunately, to increase problem size of a deterministic global NWP domain, one must increase resolution and, unavoidably, shorten the time step.[12] Therefore, global NWP models can achieve weak scaling only with respect to time-per-time step and not with respect to time-to-solution. Figure 7 shows weak scaling efficiency for time-per-time step, comparing results from the smaller 13 km benchmark to the larger 3 km benchmark where the ranges of core counts for the two benchmarks overlapped.

All the non-hydrostatic dycores tested used local (nearest neighbor) patterns of MPI interprocessor communication, which provided good weak scaling. The number and size of messages to an MPI task's immediate neighbors is roughly constant with respect to the total number of MPI tasks. Spectral transforms, however, involve non-local communication patterns for which the number and sizes of messages are functions of the total number of tasks. Consequently, and as expected, weak scaling efficiency of the spectral/semi-Lagrangian IFS model was lower than for the other models.

Floating Point Precision

The modeling groups were free to choose the floating-point precision appropriate for their dynamical core. For Round 1 testing, NMM-UJ, NIM, and MPAS were run with single precision (32 bit) floating point arithmetic. FV3, NEPTUNE, and IFS ran with double precision (64 bit) arithmetic. For Round 2 testing, the first round FV3 results were rerun using using single

---

[9] For strong scaling, the Round-1 NIM result on 128,140 cores was used. The Round-2 123,240 core retest was with improved MPI communications and cannot be compared to the other NIM 3 km core counts for computing scalability.
[10] Round-2 NIM result with MPI communication improvement on 123,240 cores.
[11] This was the version of NEPTUNE with the higher-order numerics from Round 1. The lower order NEPTUNE dycore was tested for the 13 workload but not with the 3 km workload in Round 2.
[12] For ensembles, problem size can be increased by adding ensemble members rather than increasing resolution.

precision. Both sets of FV3 results are plotted except in Figure 2, where only the best result for each model in each round is shown.

VI. Software evaluations

AVEC was charged with evaluating the designs and implementations of the model codes submitted as candidates for Level-1 benchmarking under the NGGPS Dynamical Core Testing Plan. The software evaluation[13] is intended to highlight strengths and identify potential weaknesses with respect to maintainability, extensibility, development process, and performance portability of the candidate software packages in their state at the time of the Level-1 benchmarks, which were assumed to be preliminary. The evaluations involve inspection of the codes, review of documentation, questionnaires, and interviews. The evaluations are in progress and will be provided separately.

VII. Summary

AVEC was formed as part of the NGGPS Project Test Plan to provide technical evaluation and testing of suitability and readiness of proposed modeling systems to meet global operational forecast requirements at NWS through 2025-30. We have reported on the results of a six month effort by AVEC to test the HPC performance and scalability of six model dynamical cores, five NGGPS candidates and a guest model. The five candidate dycores are all under active development and results reported here are a snapshot in time. Performance and scalability was considered without regard for model solution accuracy, which is being evaluated by other teams within the NGGPS effort.

Performance was measured as solution speed – more specifically, the number of processors needed to reach an operational speed requirement – using a workload sized to represent global forecast domains today and in the near future. The rankings, from fastest to slowest were:

1. IFS, ECMWF (hydrostatic, not an NGGPS candidate)
2. NMM-UJ, NOAA/NCEP
3. FV3, NOAA/GFDL
4. NIM, NOAA/ESRL's
5. MPAS, NCAR
6. NEPTUNE, NRL

Of the candidate dycores, NMM, FV3, and NIM models required around or below 10-thousand processor cores to meet the operational speed requirement. MPAS and NEPTUNE were more costly, requiring more than 20-thousand and 30-thousand cores, respectively.

[13] See "NGGPS Level-1 Software Evaluation Criteria and Procedures", AVEC report to NGGPS program, December 11, 2014.

Scalability was measured as strong scaling efficiency running on high (greater than 100-thousand) numbers of processor cores with workloads planned to be in operational use in ten years. The rankings for strong scalability were:

1. NEPTUNE
2. MPAS
3. NIM
4. FV3 and NMM-UJ (essentially a tie)
5. IFS

The rankings by scalability were essentially the reverse of the performance rankings. Strong scalability was also measured for the 13 km benchmark case, which resulted in a wider spread but essentially the same rankings. In terms of weak scalability, all of the cores with local (nearest-neighbor) MPI communication patterns scaled well. The IFS scaled less well because of non-local communications in the spectral transforms.

NGGPS candidate models with more computationally demanding formulations were more expensive for a given level of performance but scaled well up to the limits of the test HPC system. Codes with less computational cost provided high performance on modest resources but scaled less well. In the course of the AVEC discussions and testing, scalability and performance have emerged as somewhat competing value objectives: one based on an assumption of computational scarcity, the other relying on an assumption of computational abundance. One view argues for embracing the expense of best possible formulations and relying on superior scaling and efficient use on next-generation processors to meet operational speed requirements. The other view argues for carefully streamlined numerical formulations sufficient for forecast accuracy while maximizing the use of available computing resources in light of throughput requirements of a large and continually growing weather service mission. Proponents of both views were represented within the AVEC during this process.

The use of single- versus double-precision floating point arithmetic was not in the original AVEC Level-1 test plan and was left to the modeling groups. However, retesting of the FV3 core during the second set of benchmarks at NERSC provide a significant improvement in speed and a corresponding decrease in scalability (single-precision FV3 results were used in the rankings above). One expects that NEPTUNE, the other non-hydrostatic candidate model that used double-precision floating point arithmetic, would show similar changes in performance and scalability switching to single precision.

Results in this report were gathered on a supercomputer using a close-to-latest version of a conventional multi-core processor. Only one of the candidate models was ready for testing on novel processor architectures such as NVIDIA's Graphics Processing Units (GPU) and Intel's Many Integrated Core (MIC) architecture. We recommend further continued evaluation and testing of the NGGPS candidate models for computational intensity, memory locality, and maximum of exposure of parallelism, especially thread and fine-grained parallelism.

Acknowledgements

The AVEC gratefully acknowledges NERSC management for the allocation of time on Edison and NERSC staff Nicholas Wright, Matthew Cordery, Francesca Verdier, and Tina Declerck for their help in setting up AVEC accounts and allocations and for invaluable technical support. AVEC also gratefully acknowledges Bill Barth and Tommy Minyard at TACC, and William Thigpen, Cathy Schulbach, and Piyush Mehrotra at NASA/AMES for providing accounts, allocations and technical support for benchmark development and testing.

Finally, we are grateful to ECMWF for providing the IFS code, datasets, and for allowing George Mozdzynski to contribute his expertise, effort and obvious enthusiasm to our effort.

Table 1: Non-hydrostatic NGGPS candidate models plus the IFS, a hydrostatic guest dycore from ECMWF. No GFS dycore was submitted or tested.

Model	Organization	Numeric Method	Grid
NIM	NOAA/ESRL	Finite Volume	Icosahedral
MPAS	NCAR/LANL	Finite Volume	Icosahedral/Unstructured
NEPTUNE	Navy/NRL	Spectral Element	Cubed-Sphere with AMR
HIRAM/FV3	NOAA/GFDL	Finite Volume	Cubed-Sphere, nested
NMM-UJ	NOAA/EMC	Finite difference	Cubed-Sphere
GFS-NH	NOAA/EMC	Semi-Lagrangian/Spectral	Reduced Gaussian
IFS (RAPS13)	ECMWF	Semi-Lagrangian/Spectral	Reduced Gaussian

Table 2: Details of model configurations.

		NH-GFS (Baseline) *	FV3	MPAS	NIM	NMMB-UJ	NEPTUNE	IFS (RAPS13) *
Nominally 13km	**Resolution**	13 km (TL1534)	~12 km (C768)*	12km *	13.4 *	13 km	12.71 km *	12.5 km (Tc799)
	Grid Points	3072x1536 (unreduced) 3,126,128 (reduced)	6x768x768 3,538,944	4,096,002 **	3,317,762	6x768x768 3,538,944 *	3,110,402 **	3,336,946 (reduced)
	Vertical Layers *	128	127 **	127 ***	128	128	127 ***	137
	Time Step	TBD	600s (slow phys) 150s (vertical, fast phys) 150/10 (horiz. acoustic)	72 s (RK3 dynamics) 12 s (acoustic) 72 s (RK3 scalar transport)	72 s	24 s **	75 s (advective), 15 s (sound) ****	450
Nominally 3km	**Resolution**	3 km (TL6718)	~3 km (C3072) *	3km	3.3 km **	3 km	3.13 km *	3.125 km (Tc3199)
	Grid Points	13440x6720 (unred.) 59,609,088 (reduced) **	6x3072x3072 56,623,104	65,536,002	53,084,162	6x3072x3072 56,623,104 *	61,440,000 **	51,572,436 (reduced)
	Vertical Layers *	128	127 **	127 ***	128	128	128	137
	Time Step	TBD	150 s (slow phys) 37.5 s (vertical, fast phys) 37.5/10 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.)	120
	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.	* B-grid mass points ** For fast modes and advection of basic model variables. Time step for tracers is longer by 2x.	* Resolution refers to the representative nodal spacing in the element measured as the midpoint between the minimum and mean nodal spacing and averaged over the globe. ** Horizontal grid points is six faces of cube times number of elements per face times polynomial order squared.	* Hydrostatic The Tc799 cubic grid has the same number of grid columns as a TL1599 linear grid. While the Tc3199 cubic grid has the same number of grid columns as a TL6399 linear grid.

Table 3: Elapsed time for 13 km benchmarks on NERSC Edison system.

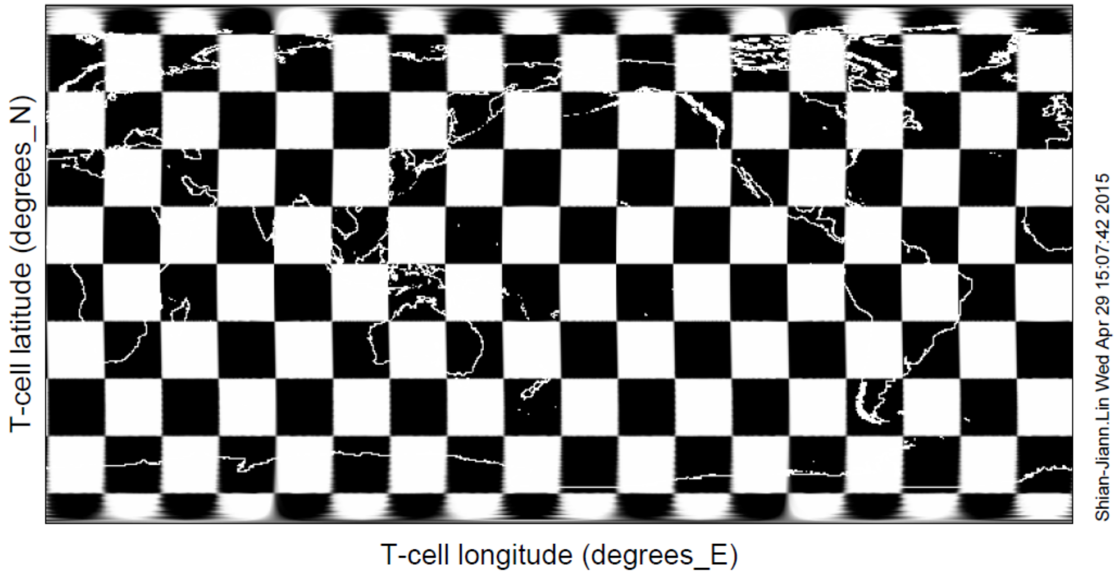
	Number of Processor Cores							
FV3	**768**	**1536**	**4608**	**9216**	**12288**	**36864**	**55296**	**110592**
Single prec.	128.86s	66.21s	24.17s	13.62s	11.36s	4.81s	3.77s	2.59s
Double prec.	232.01s	117.71s	43.38s	23.31s	18.92s	7.73s	5.64s	3.56s
NMM-UJ	**864**	**1536**	**3456**	**6144**	**13824**	**24576**		
	100.15s	59.02s	27.00s	16.12s	8.33s			
new comms	82.07s	48.65s	23.18s	13.39s	6.79s	4.77s		
NIM	**1632**	**3240**	**6432**	**12840**	**25032**	**49032**	**81024**	**123240**
	142.80s	72.18s	36.17s	18.45s				
new comms					9.99s	5.59s	3.67s	*(1)*
MPAS	**2400**	**4800**	**9600**	**19200**	**38400**	**57600**	**76800**	**96000**
	207.92s	104.67s	54.70s	28.77s	15.38s	10.86s	8.99s	7.56s
NEPTUNE	**7500**	**15000**	**30000**	**60000**	**120000**			
4th order	237.90s	119.38s	54.84s	26.14s	12.56s			
	5400	**10800**	**21600**	**43200**	**86400**			
3rd order	158.85s	76.51s	35.60s	17.80s	9.55s			
IFS	**576**	**1152**	**2304**	**4608**	**9216**			
	75.46s	40.42s	22.72s	11.52s	7.03s			

(1) Datapoint disregarded: contention on machine

Table 4: Elapsed time for 3 km benchmarks on NERSC Edison system.

Number of Processor Cores

FV3	**6144**	**18432**	**49152**	**73728**	**110592**	
Single prec.	265.23s	87.27s	35.71s	25.10s	18.76s	
Double prec.	554.18s	158.83s	64.31s	43.84s	30.31s	
NMM-UJ	**6144**	**13824**	**24576**	**55296**	**98304**	
	214.67s	98.48s	58.11s	27.18s	16.98s	
new comms	176.85s	82.61s	48.88s	23.04s	14.08s	
NIM	**8040**	**16032**	**32040**	**64032**	**123240**	**128040**
	462.56s	231.31s	117.46s	59.80s		31.75s
new comms					29.91s	
MPAS	**12000**	**24000**	**48000**	**65536**	**131072**	
	672.12s	338.71s	169.69s	126.02s	64.80s	
NEPTUNE	**25600**	**40000**	**60000**	**120000**	**128000**	
4th order	1156.52s	730.13s	479.08s	243.49s	236.68s	
IFS	**12288**	**24576**	**49152**	**98304**	**129024**	
	60.69s	32.23s	18.49s	10.95s	10.12s	



HSfvd
Range of sphum: 0 to 1 kg/kg
Range of T-cell longitude: 0.125 to 359.875 degrees_E
Range of T-cell latitude: -90 to 90 degrees_N
Current time: 1 hours since 0000-00-00 00:00:00
Current ref full pressure level: 865.949 mb

Figure 1: Checkerboard tracer initialization pattern after one hour FV3 integration. Image provided by S. J. Lin, NOAA/GFDL.

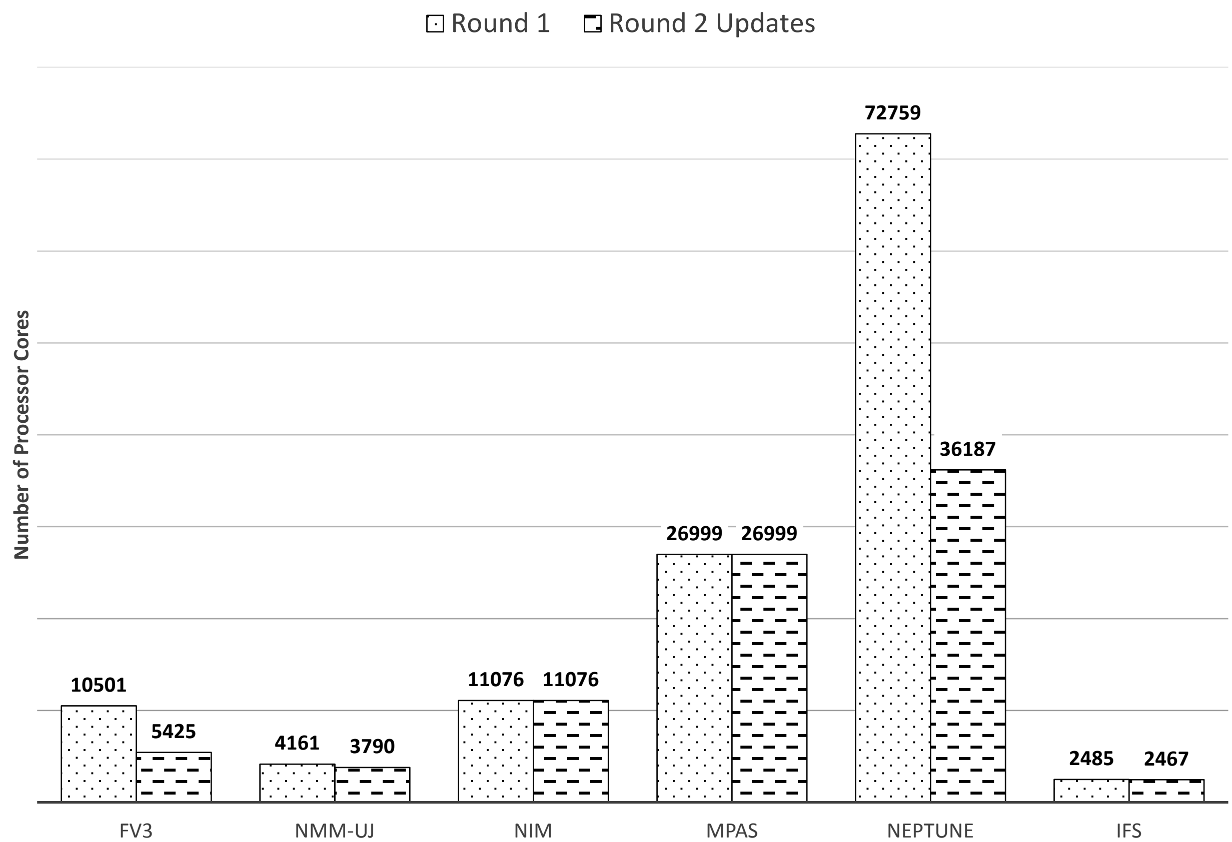


Figure 2: Number of processor cores needed to reach benchmark operational forecast integration rate of 8.5 min/day. (13 km workload; lower is better metric). The difference between rounds for the FV3 code is largely the result of running Round 2 benchmarks with single-precision floating point arithmetic; the Round 1 FV3 benchmarks were run with double precision. For NMM-UJ, interprocessor communications were improved in Round 2. For NIM and MPAS the second round tests involved running on higher processor counts than shown in this figure, so there was no change. The NEPTUNE code run during Round-2 used a different order solver and other changes that were accepted upon reruns and submission of new results from the HIWPP non-hydrostatic test cases. The slight difference in the IFS results from Round 1 to Round 2 is considered to be within the range of normal variability.

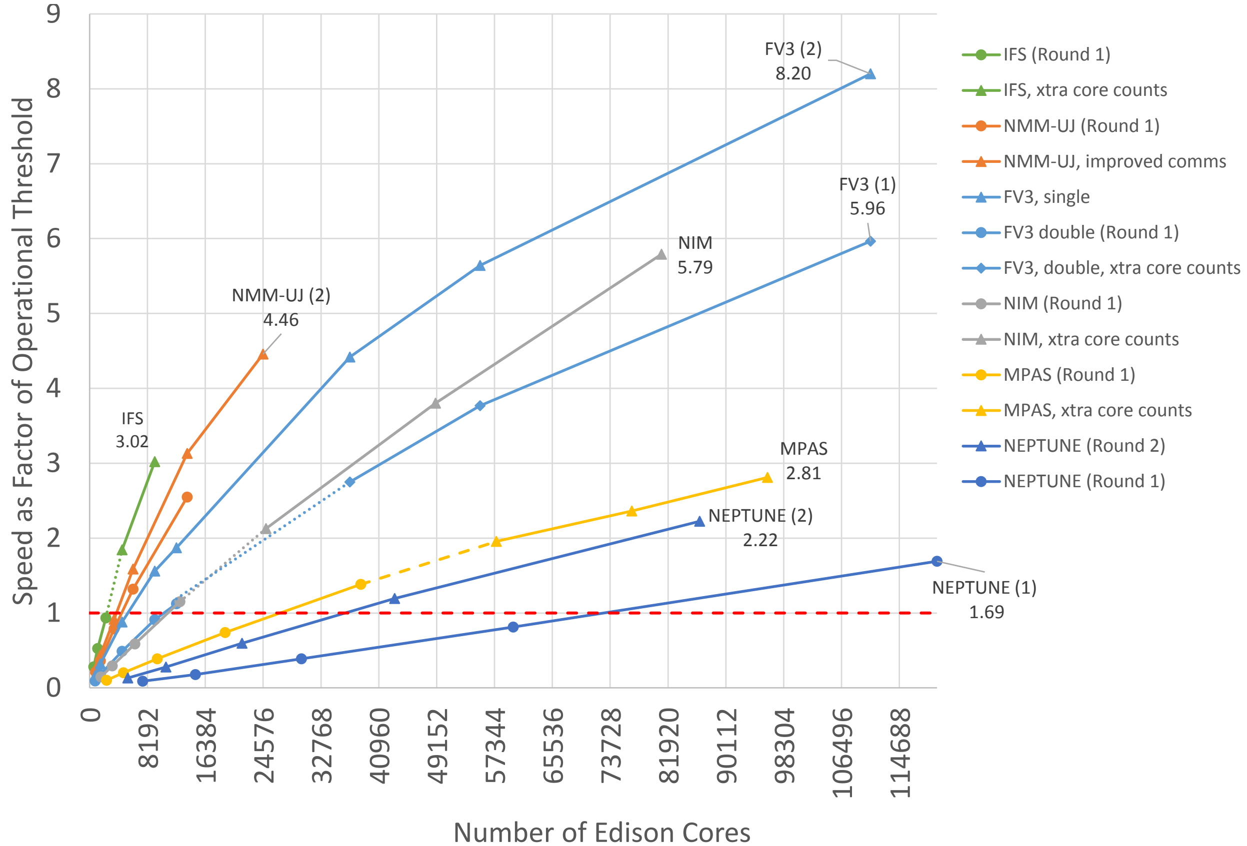


Figure 3: Speed as a function of number of cores, normalized to benchmark operational forecast integration rate of 8.5 min/day, shown as dotted red horizontal line at y=1.0. (13 km workload; higher is better metric.)

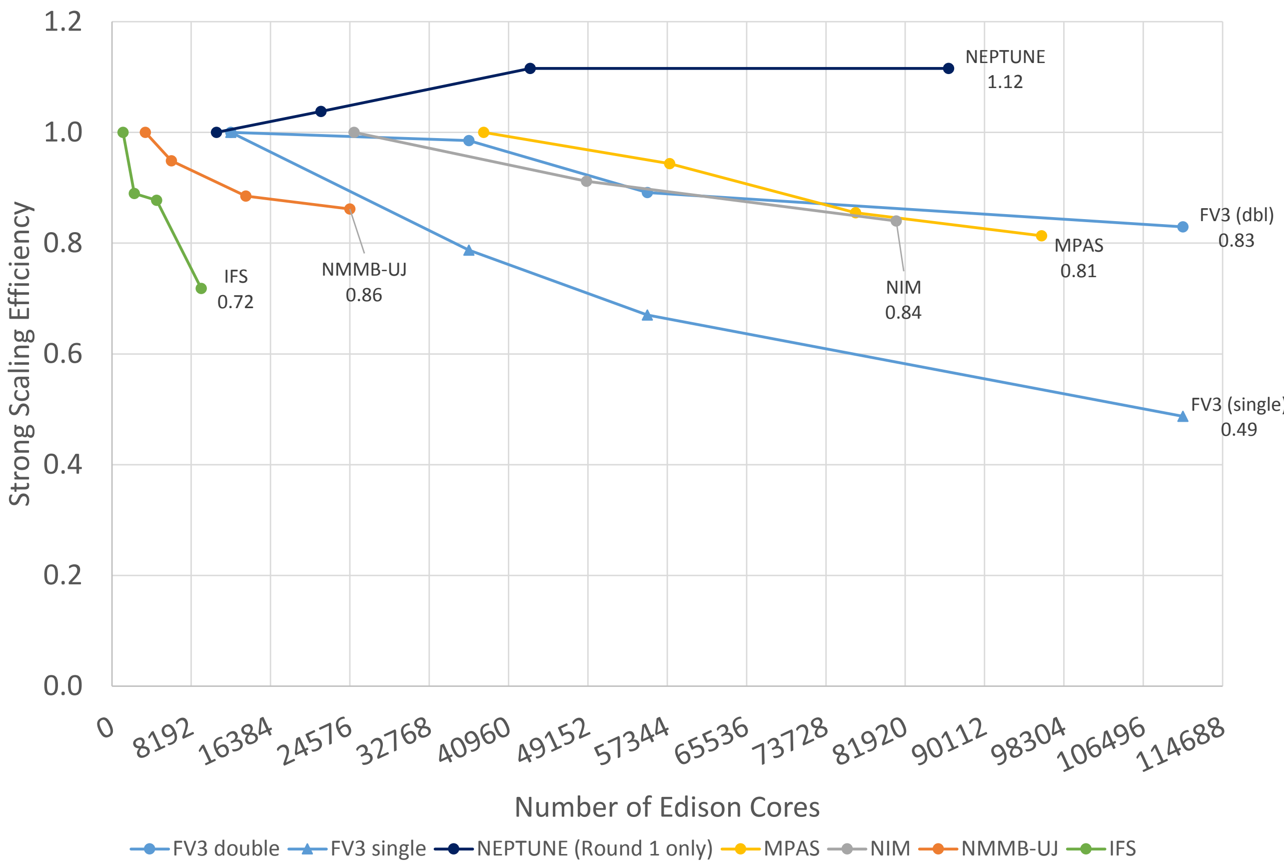


Figure 4**:** Strong scaling efficiency, defined as speedup relative to fourth highest core count divided by the increase in cores relative to fourth highest core count. (13 km workload; higher is better metric.)

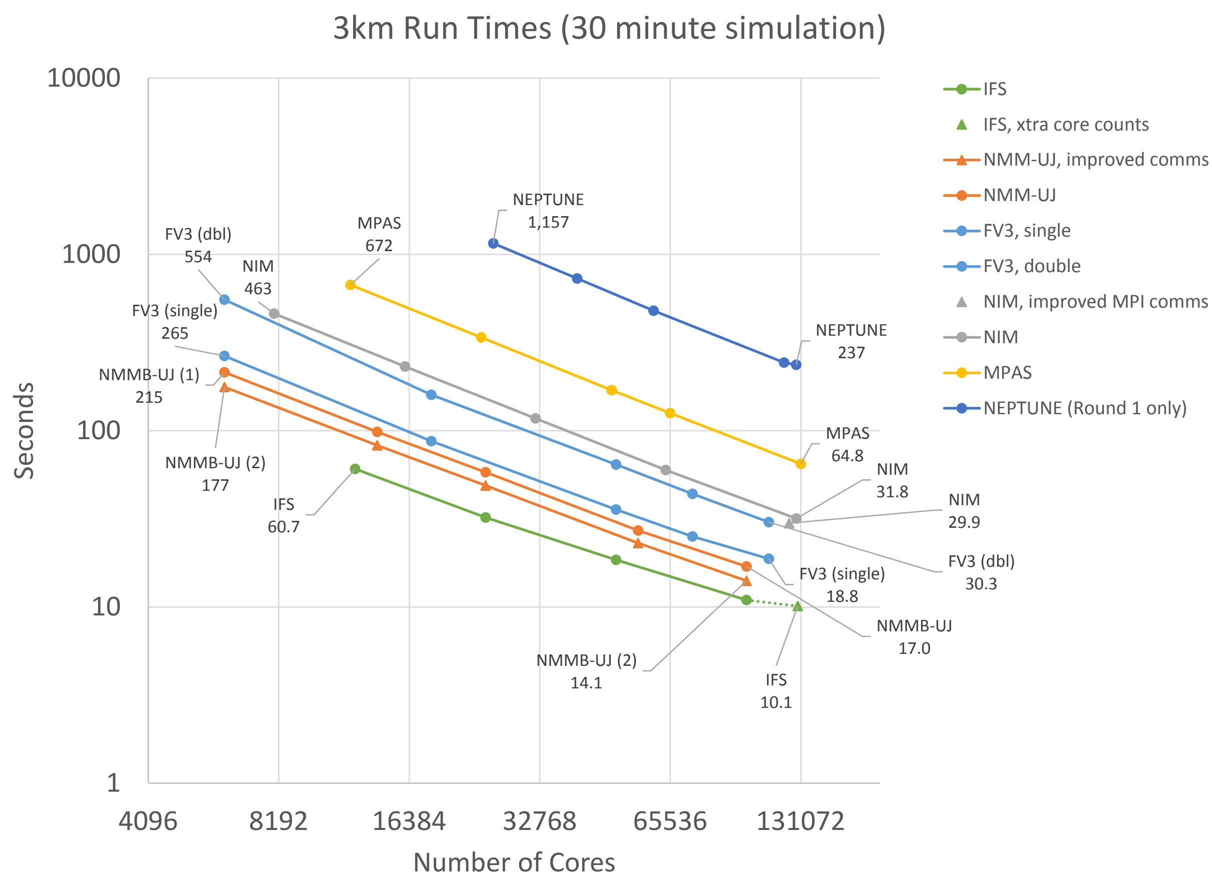


Figure 5: Elapsed time for 30 minute simulation. (3 km workload; lower is better metric; log-log scales.)

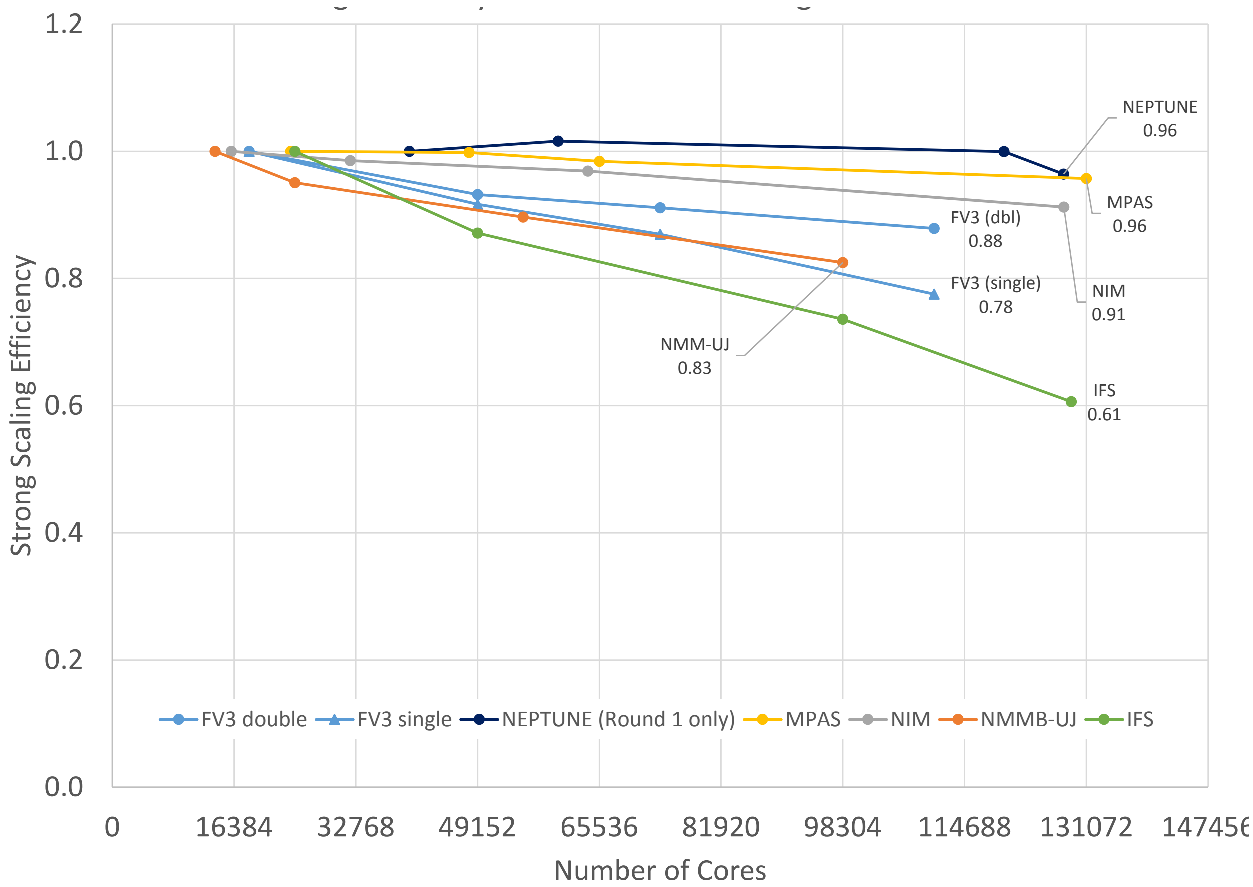


Figure 6: Strong scaling efficiency relative to fourth highest core counts.
(3 km workload; higher is better metric.)

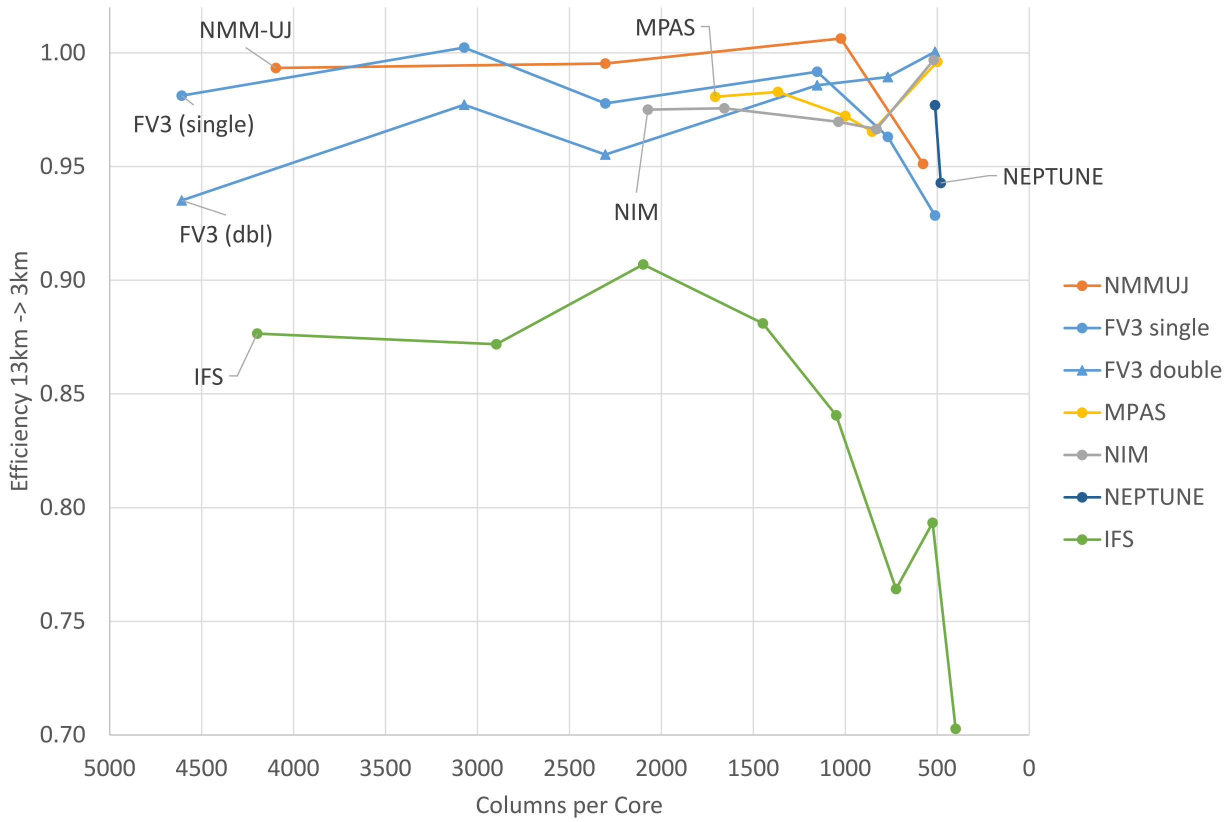


Figure 7: Weak scaling efficiency: the speed (grid-columns per second per core) running the 3 km workload divided by the speed running the 13 km workload for a given amount of work (grid-columns per core). In the case of NEPTUNE, there were only two overlapping core counts between the 13 km and 3 km benchmarks. (Higher is better metric; note that the scale of the Y-axis starts at 0.70, not zero.)