WRF nature run

2008 J. Phys.: Conf. Ser. 125 012022

(http://iopscience.iop.org/1742-6596/125/1/012022)

WRF nature run

John Michalakes[1], Josh Hacker[1], Richard Loft[1], Michael O McCracken[2], Allan Snavely[2], Nicholas J Wright[2], Tom Spelce[3], Brent Gorda[3] and Robert Walkup[4]

[1]University Corporation for Atmospheric Research (UCAR), Boulder, CO 80307, USA
[2]Performance Modeling and Characterization Lab, San Diego Supercomputer Center, La Jolla, CA 92093, USA
[3]Lawrence Livermore National Laboratory, Livermore, CA 94550, USA
[4]IBM Thomas J. Watson Research Center, Yorktown Heights, NY 10598, USA

E-mail: allans@sdsc.edu

Abstract. The Weather Research and Forecast (WRF) model is a model of the atmosphere for mesoscale research and operational numerical weather prediction (NWP). A petascale problem for WRF is a nature run that provides very high-resolution "truth" against which more coarse simulations or perturbation runs may be com-pared for purposes of studying predictability, stochastic parameterization, and fundamental dynamics. We carried out a nature run involving an idealized high resolution rotating fluid on the hemisphere, at a size and resolution never before attempted, and used it to investigate scales that span the k-3 to k-5/3 kinetic energy spectral transition, via simulations. We used up to 15,360 processors of the New York Blue IBM BG/L machine at Stony Brook Uni-versity and Brookhaven National Laboratory. The grid we employed has 4486 by 4486 horizontal grid points and 101 vertical levels (2 billion cells) at 5km resolution; this is 32 times larger than the previously largest 63 million cell 2.5km resolution WRF CONUS benchmark [10]). To solve a problem of this size, we worked through issues of parallel I/O and scalability and employed more processors than have ever been used in a WRF run. We achieved a sustained 3.4 Tflop/s on the New York Blue sys-tem, inputting and then generating an enormous amount of data to produce a scientifically meaningful result. More than 200 GB of data was input to initialize the run, which then generated output datasets of 40 GB each simulated hour. The cost of output was considered a key component of our investigation. Then we ran the same problem on more than 12K processors of the XT4 system at NERSC and achieved 8.8 Tflop/s. Our primary result however is not just scalability and a high Tflop/s number, but capture of atmosphere features never before represented by simulation, and taking an important step towards understanding weather predict-ability at high resolution.

1. Introduction

A fundamental challenge in numerical weather prediction (NWP) is to understand how (or even if) increasingly available computational power can improve weather modeling. An important enabling step toward improving that understanding is to perform a “nature run” to provide a very high-resolution standard against which coarser simulations and parameter sweeps may be compared for purposes of studying predictability, stochastic parameterization, and the underlying physical dynamics.

In this work we carry out a nature run at unprecedented computational scale on one of the world’s largest supercomputers. We calculate an idealized high-resolution rotating fluid on the earth’s hemisphere to investigate scales that span the wavenumber (k) k^{-3} (large-scale) to $k^{-5/3}$ (small-scale) kinetic energy spectral transition of the observed atmosphere using up to 15K CPUs of the IBM Blue Gene/L (New York Blue) at the New York Center for Computational Sciences (NYCCS), a cooperative effort of Stony Brook University and Brookhaven National Laboratory. Then we set a U.S. performance record of a weather code using the XT4 “Franklin” system at NERSC.

This calculation is neither embarrassingly parallel nor completely floating-point dominated, but memory bandwidth limited in the computational core, latency-bound with respect to interprocessor communication, and very I/O intensive. In these ways it is representative of many scientific calculations, and therefore achieving a high level of performance is challenging. Our primary result, however, is not just the high Tflop/s number or re-

cord-setting scalability of an atmosphere simulation, but an important step toward understanding weather predictability at high resolution.

1.1 Science motivation

A nature run that includes planetary, synoptic (barotropic and baroclinic), and near-convective scales in the mid-latitudes facilitates a new generation of basic research on predictability and turbulence. It is impossible to study predictability in the real atmosphere, making computer models necessary. The superiority of either increased resolution, or more probabilistic information, can only be established through basic predictability research. A nature run including the transition between the k^{-3} and $k^{-5/3}$ spectral regimes facilitates a new generation of predictability studies that were not previously possible. For example, simple identical-twin experiments on how errors grow within the $k^{-5/3}$ regime and across the transition can now be performed. The hypothesis of enhanced mesoscale predictability near topography with increased resolution of the model can now be rigorously addressed.

It is also difficult to study turbulence in the real atmosphere, and therefore models are attractive here as well. The turbulence community faces several challenges; wave-turbulence interactions occur within the $k^{-5/3}$ regime and across the transition, for example in the jet-stream region of the atmosphere, but wave-wave interactions within the regime and across the transition are but poorly understood.

In the meantime, the growth of computational power is enabling numerical weather prediction model forecasts within the scale region defined by the observed $k^{-5/3}$ scaling in the mesoscale. Yet we have much to learn about how waves and turbulence interact, better understanding of which will affect predictability and optimal subgrid parameterization for predictive calculations within this region and across the observed transition to larger scales. Simply increasing the resolution of operational weather forecasts may not result in improved accuracy unless we can improve understanding of the physics and model parameterizations. The long-term goal of our project is therefore to produce a suite of nature runs, including runs at resolutions achievable only with petascale computing, that can serve as a basis for current and future predictability, turbulence, and parameterization study in a multi-scale environment that spans scales above and below the spectral transition. This work describes our achievement of a milestone in that project.

Previous work of Skamarock et al. [1] showed that, with dedicated computer time on a large machine and using the Weather Research and Forecasting (WRF) model [2], high-resolution nature runs that can produce the appropriate $k^{-5/3}$ spectral slope [3] are enabled. The WRF model includes a moist thermodynamic equation making it appropriate for precipitation processes. WRF is fully nonhydrostatic, so it is appropriate for deep convection and gravity wave breaking. The numerics are stable enough to make additional damping terms, ubiquitous in typical mesoscale models, less necessary. Figure 1, reproduced from that study, encapsulates some of the evidence that the computational model is stable and of high verisimilitude.

Building on that study, the nature run done here contains instances of both stratified and unstratified turbulence, facilitating their study in a rotating fluid on a sphere and in the presence of many other scales. It further allows the study of gravity waves in a realistic environment, including gravity wave breaking. This level of fidelity is unprecedented, and opens new avenues of research to improve NWP.

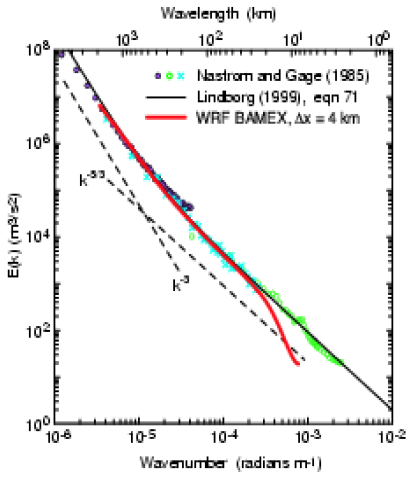

Figure 1. Spectral energy density in the WRF model compared to observations. The red curve is spectra computed from WRF forecasts at 4 km grid spacing, averaged from 3 May 2003 to 14 July 2003. Both the transition of the spectral slope from k^{-3} to $k^{-5/3}$, and the numerical dissipation range are evident. Observations from Nastrom and Gage [4] and Lindborg [5] are shown with points and the solid black curve, respectively (figure courtesy W. C. Skamarock).

1.2 Computational approach

The WRF model [1] is a limited-area model of the atmosphere for mesoscale research and operational NWP. Developed and maintained as a community model, WRF is in widespread use over a range of applications including real-time NWP, tropical cyclone and hurricane research and prediction, regional climate, atmospheric chemistry and air quality, and basic atmospheric research. The WRF model represents the atmosphere as a number of variables of state discretized over regular Cartesian grids. The model solution is computed using an explicit high-order Runge-Kutta time-split integration scheme in the two horizontal dimensions with an implicit solver in the vertical. WRF domains are decomposed over processors in the two horizontal dimensions only; and since the solver is explicit in the horizontal, interprocessor communication is logically nearest-neighbor. Each time-step involves 36 halo exchanges and a total of 144 two-way messages between neighboring processes. The decomposition is two-level: first over distributed memory patches and then again within each patch over shared memory tiles. Thus, WRF exploits hybrid parallel (message passing and multithreaded) architectures such as BG/L, though the runs conducted here were MPI only. Weather prediction codes are by nature I/O (mostly output) intensive, repeatedly writing out a time series of 3D representations of the atmosphere. WRF used Parallel NetCDF [6] as well as direct calls to MPI-IO for these runs.

2. Key aspects of the BlueGene/L architecture for NWP

The Blue Gene/L presents several opportunities and challenges for efficient implementation of NWP simulations. Details of the tightly integrated large-scale system architecture are covered elsewhere [4]. Overall, Stony Brook's BG/L platform has 18K compute nodes (36K CPUs). Here we briefly cover its general architectural aspects here, focusing on those related to our optimizations to WRF.

Each compute node is built from a single CPU ASIC and a set of memory chips. The compute ASIC features two 32-bit superscalar 700 MHz PowerPC 440 cores, with two copies of the PPC floating point unit associated with each core that function as a SIMD-like double FPU [5]. Each node has 512 MB of physical memory.

Achieving high performance requires that the application be fully domain- decomposable into data structures that can fit this relatively modest memory per node. If this can be accomplished, then the network support for scaling is an architectural strength of BG/L, which has five networks; we focus on the 3-D torus, the broadcast/reduction tree and the global interrupt for WRF optimizations. Integration of the network registers into the compute ASIC provides not only fast interprocessor communication but also direct access to network-related hardware performance monitor data. Because of limitations on deadlock-free communication, the MPI implementation uses the tree networks only for global (full-partition) collective operations

3. Computational method

As described in Skamarock et al. [1] the continuous equations solved in WRF are the Euler equations cast in a flux (conservative) form where the vertical coordinate, denoted as η, is defined by a normalized hydrostatic pressure (or mass) following Laprise [6] as

$$\eta = (ph - pht)/\mu \quad (1)$$

where μ = phs− pht and ph is the hydrostatic component of the pressure, and phs and pht are the values for the dry atmosphere at the surface and top boundaries, respectively. Following common practice we set pht = constant. η decreases monotonically from a value of 1 at the surface to 0 at the upper boundary of the model domain. Using this vertical coordinate, we express the flux form equations as

$$Ut + (\nabla \bullet Vu) + Px(p, \varphi) = FU \quad (2)$$

$$Vt + (\nabla \bullet Vv) + Py(p, \varphi) = FV \quad (3)$$

$$Wt + (\nabla \bullet Vw) + P\eta(p, \mu) = FW \quad (4)$$

$$\Theta t + (\nabla \bullet V\theta) = F\Theta \quad (5)$$

$$\mu t + (\nabla \bullet V) = 0 \quad (6)$$

$$\varphi t + \mu{-}1\ [(V \bullet \nabla\varphi) - gW] = 0 \quad (7)$$

$$(Qm)t + (\nabla \bullet VQm) = FQ \quad (8)$$

where μ(x, y) represents the mass of the dry air per unit area within the column in the model domain at (x, y). Hence the flux form variables are defined as U = μu/m, V = μv/m, W = μw/m, Ω = μη/m. And m is a map-scale factor that allows mapping of the equations to the sphere (see [7]) and is given as m = (Δx, Δy) distance on the Earth.

The velocities v = (u, v, w) are the physical velocities in the two horizontal and vertical directions, respectively, ω = ˙η is the transformed "vertical" velocity, and θ is the potential temperature. Qm = μqm; Qm= Qv, Qc, Qi ..., represent the mass of water vapor, cloud, rain, ice, etc., and q∗ are their mixing ratios (mass per mass of dry air).

We also define nonconserved variables φ = gz (the geopotential), p (pressure), and α = 1/ρ (the specific volume) that appear in the governing equations. The P's are pressure gradient terms.

4. Data issues

The nature run is data intensive, with a large sum memory footprint. During the New York Blue run we used a 2-billion cell, 4486 by 4486 grid with 101 levels. Horizontal resolution (width of an individual grid cell) was 5km; the time step was 6 seconds. Experimentally, the smallest possible run was 2048 processors, requiring 287 MB/task for WRF data, not counting buffers, executable size, operating system tax, and the like. Each output interval (1 simulation hour), the model generates 40 GB of data, the size of five 3D fields (three components of wind velocity, potential temperature, and mass) and two 2D fields. Normally, a much larger set of fields is output by WRF; however, this was reduced to the barest minimum to reduce cost. On the input side, the size of the dataset read initially to start the model contained 26 3D fields and was over 200 GB.

5. Porting and tuning

To achieve high performance with WRF on BG/L, we overcame two primary hurdles: the size of main memory on BG/L and the nonscalable I/O scheme in WRF.

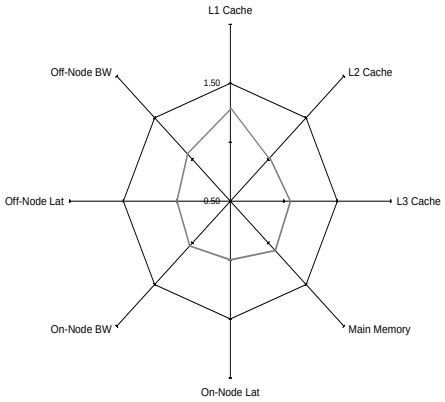

Figure 2. Performance-modeled response of WRF to doubling of processor attributes post-tuning.

Most data structures in WRF scale in memory. The domain decomposition and associated local memory extents used to dynamically allocate state arrays are calculated at run time on each process. However, each processor used to keep the global set of boundary conditions. This had been sufficient up to modest numbers (several hundreds) of processors; but with very large grid sizes on thousands of processors, the memory for arrays that store lateral boundary conditions (LBCs) ballooned out using more memory than the rest of model state combined, quickly exceeding the 512 MB physical memory limit. The solution was to fully decompose all dimensions so that each processor only stores the LBCs used in its calculation. This also involved rewriting the code for performing I/O on LBCs. With this optimization, the full state required by each processor fits in memory even on the very large grid 4486 by 4486, 101 levels, on 15K processors.

The other scaling issue we addressed was also I/O related. WRF, like many parallel applications, historically used a single-reader/single-writer scheme for distributed I/O and thus required large, un-decomposed buffers to be stored on at least one process. Again this quickly exceeded the physical memory of one BG/L node. Support for MPI-IO was added through parallel NetCDF and also direct calls to MPI-IO. Thereby we avoided the need to collect data on a single I/O task. We also encountered the 32-bit addressing limit on Blue Gene/L in the form of inability to define an MPI type large enough write a 3D field (8 GB) in one call to MPI-IO's write routine. Instead, 3D fields were written one level at a time.

Figure 2 shows a performance model developed for WRF using the PMaC methodology [8]. The figure confirms that posttuning WRF is a well-balanced, floating-point intensive code; it is most sensitive to L1 cache bandwidth (i.e., clock speed of processor) for fixed processor counts.

6. Performance measurement and results

Floating-point operations were counted using the APC performance counter library on BG/L. This library accesses the compute node ASIC's hardware performance counters to tracks several events including FPU and some SIMD and load and store operations. Because WRF uses single (32-bit) floating point precision, it was not possible to fully exploit the BG/L "double-hummer" SIMD capability.

Using 15K CPUs of Blue Gene/L at Stony Brook, we achieved 3.39 Tflop/s and set a parallelism record for number of processors running a weather code. The model was run on 6,144 and 15,360 nodes of NY Blue in co-processor mode. Only one of the two processors on each node was in use, an advantage for large- memory, high memory-bandwidth applications such as weather models.

Considering first floating-point rate by itself, the BG/L system delivered 1.49 Tflop/s on 6K processors, and 3.39 Tflop/s on 15K. Thus, scaling relative to the 6K rate was better than 90 percent efficient. Meanwhile, the output bandwidth on the BG/L, rather than degrading with increasing numbers of writers, was seen to scale as well: 242 MB/s on 6K processors and 286 MB/s on 15K. The result was good overall performance, even when the cost of writing 40 GB every simulation hour was considered: WRF still achieves 1.44 Tflop/s on 6K processors (3 percent penalty) and 3.17 Tflop/s on 15K processors (6 percent penalty), even with the cost of model output.

Next, using the tuned code, we set a speed performance record for a U.S. weather model running on the Cray XT4 "Franklin" supercomputer at the Department of Energy's National Energy Research Computing Center (NERSC) at Lawrence Berkeley National Laboratory. Running on 12,090 processors of this 100 peak teraflops system, we achieved the important milestone of 8.8 teraflops – the fast-est performance of a weather or climate-related application on a U.S. supercomputer.

Initial science results, indicated in figure 3, demonstrate that the proposed suite of nature runs are feasible. In both panels the mid-latitude wave train is obvious, encircling the hemisphere. The top panel shows relative vorticity at approximately 5 km above ground level (model level 30) after 3 hours of simulation (model/earth time). In this figure, reds are large positive (cyclonic) vorticity and dark blues are large negative (anticyclonic) vorticity. We note the multiple scales of flow present, indicative of high resolution, which are not typically seen in simulations of this spatial extent (e.g., global). The 5 km grid spacing (ΔX) in this simulation easily resolves the spectral transition, which will typically reside at scales on the order of $20\Delta X$. Gravity waves related to baroclinic development may be present, suggested by the narrow filaments paralleling the streaks of greatest magnitude. More investigation is required to determine whether they are gravity waves, and then to understand whether they are contributing to the spectral transition from k^{-3} to $k^{-5/3}$.

The bottom panel of figure 3 shows wind speed (in m/s) at approximately jet-stream level (10 km, model level 60), with red indicating maximum wind speeds in jet streaks. In order to reach the long-term scientific objectives of this project, this wave train, the associated baroclinic development, and near-convective scales where gravity waves may break need to be simulated together. This is possible only at high resolution on domains that cover a hemisphere or the globe, so that many instances of waves, turbulence, and their interactions on both sides of the spectral transition are present. We have thus succeeded in opening new lines of investigation, and the path is now open for the next steps.

7. Conclusion

Figure 2 shows example results from the nature run. We set records for performance and scalability of WRF or any other atmospheric simulation. But our primary achievement is not performance or scalability; rather, it is new science to enable improved numerical weather prediction.

We carried out a WRF nature run that provides very high-resolution "truth" against which more coarse simulations or perturbation runs may be compared for purposes of studying predictability, stochastic parameterization, and fundamental dynamics. We studied idealized high resolution rotating fluid on the hemisphere to investigate scales that span the k^{-3} to $k^{-5/3}$ kinetic energy spectral transition of the observed atmosphere using 15K CPUs of BG/L with achieved 3.4 Tflops. Then we set a U.S. performance record of a weather code using the XT4 "Franklin" system at NERSC. We have thereby opened up new avenues of science investigations via simulation.

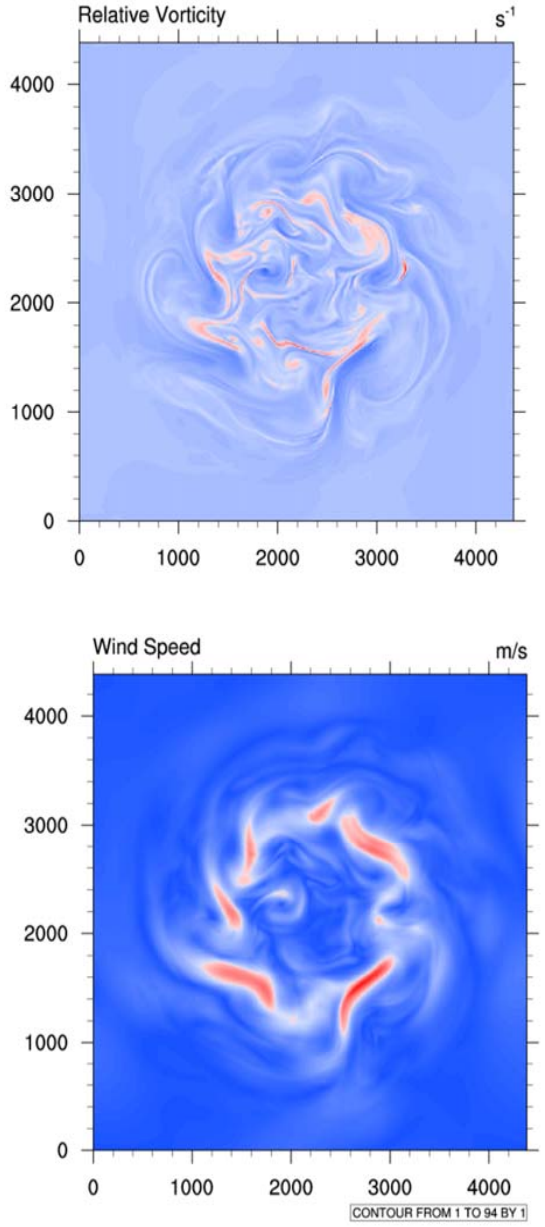

Figure 3. Relative vorticity at approximately 5 km above ground (top), and wind speed near the jet stream at approximately 10 km above ground. The entire midlatitude wave train is simulated at high resolution (ΔX=5 km), and the solutions show the range of scales in the flow.

Acknowledgments

We thank William Skamarock, whose equations and description of WRF's numerical formulation is reprinted with permission in Section 3. This work was sponsored in part by the National Science Foundation via GEO ATM

SGER award #0637994 “Feasibility of Taking the Weather Research and Forecasting (WRF) Model to Petascale” and by the OCI award entitled “The Cyberinfrastructure Evaluation Center”, and in part by DOE Office of Science through the SciDAC2 award entitled Performance Engineering Research Institute (PERI). We also thank Andrew Vogelmann, Brian Colle, and Efstratios Efstathiadis of the New York Center for Computational Sciences at Stony Brook University and Brookhaven National Laboratory.

This paper updates results presented in the Gordon Bell Prize track at SC07.

References

[1] Skamarock W, *et al.* 2007 A time-split nonhydrostatic atmospheric model for weather research and forecasting applications *J. Comput. Phys.*

[2] Skamarock W 2004 Evaluating mesoscale NWP models using kinetic energy spectra *Monthly Weather Rev*.

[3] Skamarock W, Klemp J, Dudhia J, Gill D, Barker D and Wang W and Powers J 2005 A description of the advanced research WRF version 2 *NCAR Technical Note*

[4] Nastrom E D and Gage K S 1985 A climatology of aircraft wavenumber spectra observed by commercial aircraft *J. Atmos. Sci.* **42**:950-960

[5] Lindborg E 1999 Can the atmosphere kinetic energy spetrum be explained by two-dimensional turbulence? *J. Fluid Mech.* **388:**259-288

[6] Parallel NetCDF http://www-unix.mcs.anl.gov/parallel-netc

[7] Bachega E L, Chatterjee S, Dockser K, Gunnels J, Gupta M, Gustavson F, Lapkowski C, Liu G, Mendell M, Wait C and Ward T J C 2004 A high-performance SIMD floating point unit design for BlueGene/L: architecture, compilation, and algorithm design *PACT*

[8] Laprise R 1992 The Euler equations of motion with hydrostatic pressure as an independent variable *Mon. Weather Rev.* **120:**197-207

[9] Haltiner G J and Williams R T 1980 *Numerical Weather Prediction and Dynamics Meteorology* 2nd ed. (John Wiley and Sons)

[10] PMaC 1980 htttp:/www.sdsc.edu/pmac