

WRF Nature Run

John Michalakes Josh Hacker Richard Loft {michalak, hacker, loft} @ucar.edu University Corporation for Atmospheric Research (UCAR), Boulder, CO.	Michael O. McCracken Allan Snavely Nicholas J. Wright {mmcrack,allans,nwright} @sdsc.edu PMaC Laboratory San Diego Supercomputer Cen- ter, La Jolla, CA.	Tom Spelce Brent Gorda {spelce,bgorda} @llnl.gov Lawrence Livermore National Laboratory, Livermore, CA.	Robert Walkup walkup@us.ibm.com IBM Thomas J. Watson Re- search Center, Yorktown Heights, NY.
---	---	---	---

Abstract

The Weather Research and Forecast (WRF) model is a limited-area model of the atmosphere for mesoscale research and operational numerical weather prediction (NWP). A petascale problem is 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 carried out a nature run involving an 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 65,536 processors of the BG/L machine at LLNL. We worked through issues of parallel I/O and scalability. The primary result is not just the scalability and high Tflops number, but an important step towards understanding weather predictability at high resolution.

Categories and Subject Descriptors

C.4 [Computer Systems Organization]: Performance of Systems – *measurement techniques*.

General Terms

Weather Research, Algorithms, Measurement, Performance

Keywords

Weather Research, High Performance Computing.

1. Introduction

A fundamental challenge in numerical weather prediction (NWP) is to understand how (or even if) increasingly

(c) 2007 Association for Computing Machinery. ACM acknowledges that this contribution was authored or co-authored by a contractor or affiliate of the [U.S.] Government. As such, the Government retains a nonexclusive, royalty-free right to publish or reproduce this article, or to allow others to do so, for Government purposes only.

SC07 November 10-16, 2007, Reno, Nevada, USA
(c) (c) 2007 ACM 978-1-59593-764-3/07/0011...\$5.00

available computational power can improve weather modeling. An important enabling step towards improving that understanding is to perform a "nature run" to provide a very high-resolution standard against which more coarse 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 the world's largest supercomputer: we calculate an idealized high resolution rotating fluid on the earth's hemisphere to investigate scales that span the wavenumber (k) $k-3$ (largescale) to $k-5/3$ kinetic energy spectral transition of the observed atmosphere using up to 16 racks (16,384 nodes, 32,768 CPUs) of BlueGene/L at IBM Watson (BGW) and then use 64 racks (65,536 nodes and 65,536 CPUs—that is one CPU per node) of BlueGene/L (BG/L) at Lawrence Livermore National Laboratory (LLNL) on the same problem.

This calculation is neither embarrassingly parallel, nor completely floating-point dominated, but memory bandwidth limited, and latency-bound with respect to interprocessor communication. In these ways it is representative of many scientific calculations, and therefore achieving a high level of performance is challenging. The primary result is not just a high Tflops number, but an important step towards understanding weather predictability at high resolution.

1.1 Science Motivation

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 experiments within the $k-5/3$ regime, studying how errors grow when initial conditions are slightly perturbed, 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 sub-grid 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 predictability, turbulence, and parameterization study in a multi-scale environment that spans scales above and below the spectral transition. This work describes a milestone in that project.

Previous work of Skamarock, et al. [2] showed that, with dedicated computer time on a large machine and using the Weather Research and Forecasting (WRF) model [1], 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.

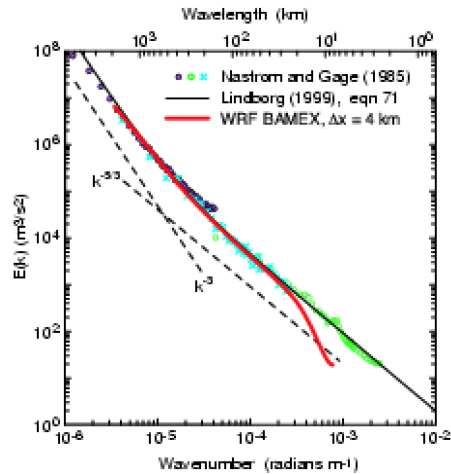


Figure 1 (courtesy W.C. Skamarock): Spectral energy density in the WRF model compared to observations. The red curve is spectra computed from WRF forecasts at 4 km grid spacing, averaged from 3 May 2003 to 14 July 2003. Both the transition of the spectral slope from k -5/3 to k -3, and the numerical dissipation range are evident. Observations from Nastrom and Gage (1985) and Lindborg (1999) are shown with points and the solid black curve, respectively.

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.

1.2 The 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/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. Since WRF domains are decomposed over processors in the two horizontal dimensions only, interprocessor communication is between-neighbor on the BG/L (and most) supercomputer topologies. Each time-step involves 36 halo exchanges and a total of 144 nearest-neighbor exchanges (assuming aggregation). 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 multi-threaded) computation modes. Weather prediction codes are I/O (mostly output) intensive. WRF uses Parallel NetCDF for I/O [8].

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

BG/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, LLNL's BG/L platform has 65,536 compute nodes, and Linpack rating of 280.6 Tflops. 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 compute node 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 not only provides fast inter-processor communication but also direct access to network-related hardware performance monitor data. Due to 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 [2] 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:

$$\eta = (p_h - p_{ht})/\mu \quad (1)$$

where $\mu = p_h - p_{ht}$ and p_h is the hydrostatic component of the pressure, and p_h and p_{ht} are the values for the dry atmosphere at the surface and top boundaries, respectively. Following common practice we set $p_{ht} = \text{constant}$. 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, the flux form equations are expressed as

$$U_t + (\nabla \cdot Vu) + P_x(p, \varphi) = FU \quad (2)$$

$$V_t + (\nabla \cdot Vv) + P_y(p, \varphi) = FV \quad (3)$$

$$W_t + (\nabla \cdot Vw) + P_\eta(p, \mu) = FW \quad (4)$$

$$\Theta_t + (\nabla \cdot V\Theta) = F\Theta \quad (5)$$

$$\mu_t + (\nabla \cdot V) = 0 \quad (6)$$

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

$$(Q_m)_t + (\nabla \cdot VQ_m) = FQ \quad (8)$$

Where $\mu(x, y)$ 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 $U = \mu u/m$, $V = \mu v/m$, $W = \mu w/m$, $\Omega = \mu \eta/m$. And m is a map-scale factor that allows mapping of the equations to the sphere (see [7]) and is given as $m = (\Delta x, \Delta y)$ distance on the earth

The velocities $v = (u, v, w)$ are the physical velocities in the two horizontal and vertical directions, respectively, $\omega = \eta$ is the transformed 'vertical' velocity, and θ is the potential temperature. $Q_m = \mu q_m$; $Q_m = Q_v, Q_c, Q_i, \dots$, 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 non-conserved variables $\varphi = gz$ (the geopotential), p (pressure), and $\alpha = 1/\rho$ (the specific volume) that appear in the governing equations. The P 's are pressure gradient terms.

The WRF variables of state are discretized over regular Cartesian grids and the model solution is computed using an explicit high-order Runge-Kutta time-split integration scheme. Each time step involves solution of the partial differential equations of mass and momentum (dynamics) and computation of various physical forcing terms that contribute to the evolving state of the earth's atmosphere. The latter (physics) involves heavy use of the Fortran intrinsics log, exp, power (**), & sqrt.

4. Data Issues

The nature run is quite data intensive with a large sum memory footprint. During the BGW runs at Watson we used a 907x907 grid with 101 levels, resolution at 25km, time step at 30s. Experimentally, the smallest possible run at BGW was 2048 processors with (theoretically) 287 MB/task for WRF data not counting buffers, executable size, OS tax etc.

Our LLNL proposed final run will be a 5km resolution version, which will be 25 times bigger in space, plus shorter time step (perhaps by a factor of 3). We estimate that the resulting 4500 x4500 grid with 101 levels, a horizontal resolution of 5km, and time step of 10s should require, at minimum 51,200 nodes to fit in physical memory.

5. Porting and Tuning.

To achieve high performance with WRF on BG/L the primary hurdles we overcame were 1) the size of main memory on BG/L and 2) the simplistic I/O scheme in WRF.

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 some global state 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 to use more memory than the rest of model state combined and quickly exceeded 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 4500x4500, 101 levels, on 51K+ nodes.

The other scaling issue we addressed was also I/O related—WRF, like many applications, historically used but 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.

There was however a bug in the MPI-IO implementation for BG/L that Parallel NetCDF exposes that IBM worked with us to fix, as described in the next section.

Figure 2 shows the theoretical performance-modeled performance sensitivity of WRF to speedup if the target machine were to be improved by doubling peak flops and L1 cache bandwidth, L2 bandwidth, L3, bandwidth, main memory bandwidth, or halving interprocessor network latency and doubling bandwidth. It will be seen for example, that doubling peak flops alone without improving the memory subsystem will confer no performance improvement, while doubling various data-movement bandwidths can boost performance 15-25% each. For every such architectural study there is a symmetric application tuning one; for example one can get the effect of doubling memory bandwidth by reordering data accesses in a way that halves memory bandwidth demands.

Performance sensitivity of WRF

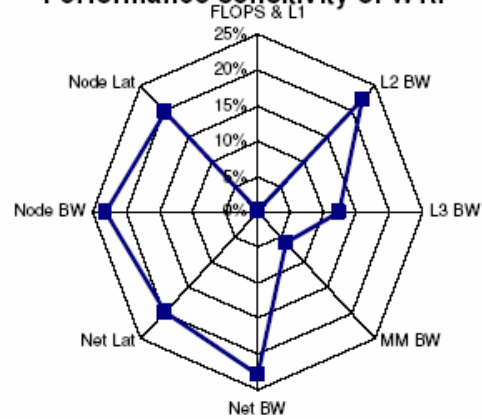


Figure 2: Performance sensitivity of WRF to doubling of underlying machine's performance attributes

Tuning efforts are therefore focused primarily on improving data locality to lower memory bandwidth demands and thus enabling higher flops as described next.

6. Performance Measurement

Floating-point operations (flops) were counted using the IBM perfctr library for BlueGene/L. This library reads the compute node's hardware performance counters to count events of interest such as load and store instructions and various floating-point operations. The counters available on BlueGene/L have some limitations—it is not possible to measure all the events that count as floating-point operations at once on a single CPU. We worked around this by using a library which assigns counter groups to CPUs in a round-robin fashion and reports summary statistics of the counters across CPUs. The BlueGene/L counters also can not measure some events that should rightly be counted as flops. We currently have no way to account for these uncounted operations and WRF is a large enough software system that counting flops in the code by hand is prohibitively difficult. The total operation count used is the sum of floating-point add and subtract operations, multiply and divide operations weighted as one flop each, and fused multiply-add and "SIMDized" parallel fused multiply-add instructions, weighted as two and four flops each. Note that the counters record only a sum of multiply and divide instructions, and there is no way to weight division differently, although divide is considerably more expensive. Because of these limitations, it is likely that the values reported by the counters under-represent the actual executed flops by some extent.

We use function calls to control when the counters are being measured. We begin counting after the first iteration and end before the last, thus excluding file I/O time from the count.

7. Results

As described above, the BlueGene/L architecture can theoretically issue 4 floating point operations per clock. In order to use both floating point units the arguments must be quad word aligned. The compiler has limited capability to automatically word-align arguments; however, while debugging of the MPI-IO/ParallelNetCDF bug mentioned above, to ensure correctness, we were forced to use a version of the compiler that does not support this option. Therefore, the best we could theoretically achieve was *two* flops per clock and in fact, using 4 racks (4096 nodes, 8192 CPUs) of BlueGene/L, for the 907x907 grid, at IBM Watson (BG/W) we achieved 7% of this two-flops-per-cycle “peak” (1.4 Gflops per CPU, 11.5 Tflops total). This result is compatible with previously reported per-processor efficiency results for WRF¹ and perhaps not atypical of data-intensive codes. With the MPI-IO bug now apparently fixed, we anticipate turning on word-aligning SIMD options of the compiler and thus being able to boost performance as much as 2x (which would be around 2 Tflops for 8 racks or 7% of true four-flops-a-cycle peak).

The full-resolution nature run will fill up the entire BG/L machine and thus scalability, more than per-processor efficiency, will be paramount. In the meantime, while tuning for scalability and eliminating the MPI IO bug, we tested the scalability of the smaller 907x907 problem, with the compiler version that works around the bug and ensures correctness but can only deliver 2 flops-per-cycle, at up to 65,536 CPUs on the Livermore machine. The scaling results are shown in Figure 3.

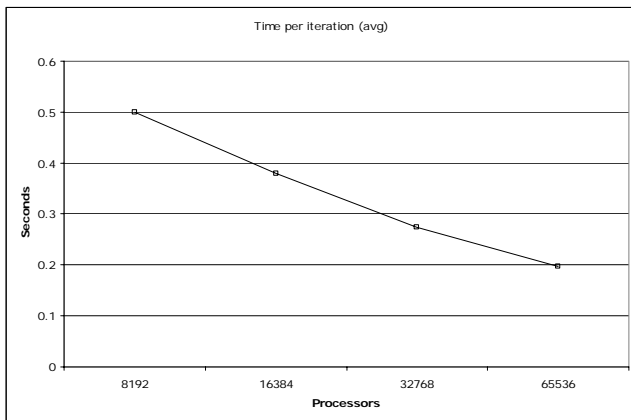


Figure 3: Scalability of nature run 907x907

It can be seen that scaling is nearly linear. Scaling is starting to flatten out, running out of work, at 65,536 CPUs.

¹ WRF has been benchmarked at 7.1 Tflops running at 11 percent of aggregate theoretical peak on 12,500 Cray XT4 processors [personal communication, Peter Johnsen, Cray].

The full resolution 4500x4500 grid should be sufficient to utilize a machine twice as large as current BG/L at 7% or so utilization. At time of writing BG/L at LLNL was taken offline for approximately 1 month to be upgraded including adding larger memory nodes. We are therefore continuing to performance-tune on BGW at Watson, with the anticipation of running our full resolution problem on the upgraded LLNL machine when it comes back up, and achieving an estimated > 64 Tflops prior to presentation at SC2007.

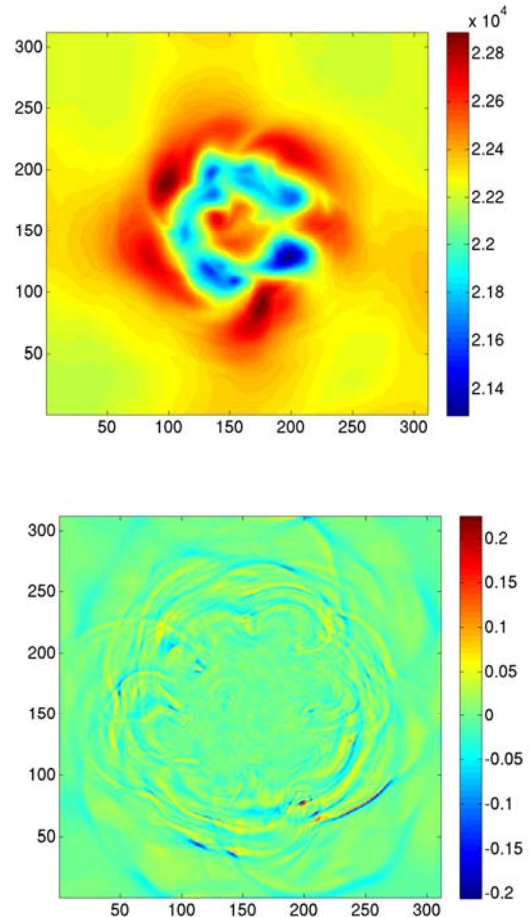


Figure 4: Northern-hemisphere polar stereographic projections of the WRF model state. Above, pressure (Pa) on model level 59 (approximately the level of the jet stream). The mid-latitude wave-train is evident as the ring of lower pressures (blue). Below, the corresponding vertical velocity (m s⁻¹), showing smaller scale features and sharper gradients likely resulting from frontal boundaries and gravity waves. These dynamics are one possible source for generation of the k-5/3 spectral slope.

8. Conclusion

Figure 4 above shows example results from a nature run. 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 study-

ing predictability, stochastic parameterization, and fundamental dynamics. We carried out a nature run involving an 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 and carried out performance tuning resulting in using 64 racks of BG/L with excellent scalability. We anticipate achieving > 64 Tflops on the full-resolution problem on a machine even more capable than current #1 machine on the Top500 list today, when the upgraded machine comes online just prior to SC2007.

Acknowledgements

We wish to 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". The development of the tools and methods used for performance modeling were sponsored in part by the DOE Office of Science under the SciDAC2 award entitled "The Performance Engineering Research Institute" (PERI).

References

- [1] W. Skamarock, J. Klemp, J. Dudhia, D. Gill, D. Barker, W. Wang, J. Powers, "A Description of the Advanced Research WRF Version 2", NCAR Technical Note, 2005.
- [2] W. Skamarock, et al. "A Time-Split Nonhydrostatic Atmospheric Model for Weather Research and Forecasting Applications", *Journal of Computational Physics*. January 2007.
- [3] W. Skamarock, "Evaluating Mesoscale NWP Models Using Kinetic Energy Spectra". *Monthly Weather Review*, November, 2004.
- [4] N. R. Adiga et al., "An overview of the BlueGene/L supercomputer" SC2002 – High Performance Networking and Computing, 2002.
- [5] E. L. Bachegea, S. Chatterjee, K. Dockser, J. Gunnels, M. Gupta, F. Gustavson, C. Lapkowski, G. Liu, M. Mendell, C. Wait, T.J.C. Ward, "A High-Performance SIMD Floating Point Unit Design for BlueGene/L: Architecture, Compilation, and Algorithm Design" PACT, 2004.
- [6] R. Laprise, "The Euler Equations of Motion with Hydrostatic Pressure as an Independent Variable". *Mon. Wea. Rev.*, 120, (1992), 197-207.
- [7] G. J. Haltiner, and R. T. Williams, *Numerical Weather Prediction and Dynamics Meteorology*. (2nd edition, John Wiley and Sons, 1980), Inc. 477 pp.
- [8] Parallel NetCDF, see
- [9] <http://www-unix.mcs.anl.gov/parallel-netcdf/>