Workshop Report: Petascale Computing in the Geosciences

In view of The National Science Foundation's recent announcement entitled: Leadership-Class System Acquisition - Creating a Petascale Computing Environment for Science and Engineering, which calls for deployment of a petascale computational facility capable of sustained scientific applications performance approaching a petaflop (10^{15} floating point operations per second) to solve scientific questions of strategic importance by or around the year 2011, and because the time between now and then is not more than ample for scientists to prepare to use such a computer, a series of workshop around Petascale Computing in the Geosciences were organized to wit GARPA, GARPA2, Petascale Computing and the Geosciences I and II. With the permission of NSF, the workshops GARPA2 and Geosciences II were combined as they have largely overlapping goals. The objectives of these workshops was to examine the opportunities for progress in the geosciences that could be enabled by the petascale computational capability and to determine the steps necessary to ensure that this community is prepared to take advantage of such resources when they come on line. The joint workshop report is below, with a concise bullet list of *recommendations* provided first for ease of reference, and the more detailed *analysis and findings* that led to these recommendations follows.

1. Recommendations

The potential benefits of petascale computing to advance scientific discovery in the geosciences, and to improve economic competitiveness of the U.S and, and the climate for all world citizens, and to better understand the world we live on, as identified in the analysis and findings below, are manifold. However, to achieve these benefits, several specific steps are required, both of the community of research scientists in the geosciences and computer science that should advance relevant science investigations, and by funding agencies that should provide resources to carry out these science investigations. The recommendations of this workshop, developed from the analysis and findings section are:

- **A portfolio of candidate petascale applications should be established, and development funding should be provided, for collaborative teams of geoscientsits and computer scientists to prepare calculations that can both advance scientific discovery and run at petascale.** As a community we need to assemble a portfolio of application classes on the path to petascale. We should *not* strive to make this a large, comprehensive portfolio at first, but rather seek to obtain high success by enabling a few strategic geoscience computations at petascale in the first few years of system availability, particularly to solve science problems of impact. This set of workshop participants identify the following set of broad geoscience application areas as being inherently suitable for development to go to petascale, and have science impact, both in the *near* term (it is not implied this list is exhaustive). Thus candidate applications may be considered from the following broad areas:

- **Mantle Dynamics** (for example earthquake ground motion prediction using PetaShake at 25 meter, 2 Hz)
- **Climate** (for example hemispheric "nature" runs to study energy spectrum in atmosphere at the k^-3 to k^-5/3 kinetic energy spectral transitions).
- **Coupled Ocean and Weather** (for example a hurricane eyewall calculation with turbulence and mixing at sub 10m resolution)
- **Space Weather** (for example a full hemisphere coronal model to resolve and improve understanding of loop structure in a constrained coronal heating with loop structures model)
- **Ecological component of earth system modeling** (for example adding plant cover to climate models)

- Explicitly called out from the above: **projects to scale candidate geoscience calculations to petascale should be undertaken by collaborative teams of geoscientists and computer scientists starting now.** All the calculations identified in the analysis and findings section**,** even the relatively straight-forward ones, require work by both experts in the domain science and in the associated computer science, working together, to get ready for petascale.

- To ensure the kinds of facility access required to allow collaborative teams to make progress: **interactive access to large numbers of nodes, and a "hardware ladder" should be provided.** Software development and tuning teams today have difficulty obtaining access to 1000 cpus for debugging and interactive code-development. Yet petascale calculations will use 2 orders of magnitude more processors. NSF centers should provide interactive partitions and reservations for up to 10,000 cpus for short periods of time now. NSF should ensure phased deployment of Tier2 and Tier1 systems at 10,000, 100,000 cpus, and more granted to code developers along the way.

- Extending the above beyond the short-term: **petascale community should be cultivated.** To foster a growing interdisciplinary community of collaborating geoscientists and computer scientists we recommend organization of "summer institutes", focused workshops for carrying these collaborations forward, and key training programs for next-generation interdisciplinary scientists in this field. These should be advertised widely to all agencies with funding/interest in geoscience including (beyond NSF) DOE and DOD (including DARPA and HPCMO). Petascale application proposals should also be encouraged to propose "High-Performance Computational Geoscience postdocs".

- Related to the above: **representative benchmarks that represent the computational requirements of geoscience applications should be defined. A** common geoscience benchmark suite would be an important contribution towards furthering petascale computing in the geosciences. It would also simplify

procurements for vendors and geoscience application practitioners alike. For practitioners, the benchmark suite could be appropriated en-masse into multiple RFP's. This would save time and effort in developing RFP benchmark suites. For vendors, by focusing on a single suite, would save money responding to RFPs, and would allow these applications to have greater influence on system design in the future.

- Complimenting the above: **performance models of representative applications should be deployed.** It will be difficult to get benchmark time (as well as development time) on the largest supercomputers. Therefore, we emphasize the importance of having predictive models of system performance based on key architectural parameters both to guide system deployment but (more crucially) code development and tuning for the target platform**.**

- **Suitable selection criteria should be applied in choosing which specific calculations and teams from the above areas to invest effort and money into for the purpose of enabling viable petascale applications (via several years of effort starting ASAP).**
  We propose the following broad evaluation criteria:
  - Reward due to science impact
  - Strength and interdisciplinary nature of proposed team
  - Demonstrated plausibility for petascale (via reports from team initial feasibility studies)
  - Needed investment in algorithms and models (via reports from team initial feasibility studies)
  - Risk of failing to result in a viable petascale application

  We applied the first and last (risk versus reward) in identifying the broad application areas suitable for petascale in the first recommendation. Further effort is required to evaluate teams and carry out feasibility studies.

- Explicit from the above **initial feasibility studies should be carried out ASAP.** Collaborative teams should be formed to further assess the suitability of calculations, including from the broad areas of the first bullet, to achieve petascale. Performance studies should be generated showing where scaling bottlenecks are and strategies for working around these should be explored. To mitigate risk of failure, teams should not receive larger multi-year development funding before justifying further advancement via short (1 year or less) feasibility studies.

- **Innovative uses of a petascale computer should be cultivated.** There is a danger that the more speculative, innovative, uses for a petascale computer will be squeezed out early by the approach embodied in the first two recommendations. To mitigate this danger we propose an additional specialized RFP with modest funding for assessment and development of "shallow end of the pool" research that is "High Reward/High Risk" and "High Reward/High Effort).

- **"Market segmentation" should be done.** We do not believe it is the case all computational problems in geocience are potentially petascale. At the same time, many geoscience applications may stress memory bandwidth, disk I/O rates, integer functional units, database query rates, in ways that traditional high performance computing, floating-point intensive (such as physics codes solving systems of PDES (partial differential equations)) do not. There is a need and opportunity for the community to examine computational requirements of geoscience applications, and, in addition to identifying candidates for petascale, identify attributes that may be used to influence such programs as NSF OCI (Office of Cyberinfrastructure) Tier2 system procurements. The result could be a machine at less than petascale but well suited to the data-intensive applications of geoscience.

- **Software support and maintenance should be supported.** We emphasize the NSF-wide need to develop a model and paradigm, for making software solutions robust, maintainable, and reusable. This likely requires some staff support, as opposed to graduate student, as students can/should invent but should not be expected to harden and maintain software.

  The NSF GEO and CISE directorates need to increase investment in people and software applications development commensurate with the outlay in funding for hardware from OCI to enable petascale computing. We recommend a Geoscience Petaflop Computing Initiative, on the model of NMI to foster software development. A focus should be on specialized programs to capture the specific geoscience subfields for development in the context of petascale computing. There should be an emphasis on coupling and coarse-graining techniques to enable scaled-up coupled models. GEO should explore inter or cross agency coordination to address funding for Application Services, libraries, runtime/programming environments needed at petascale.

- **Storage and networking should be supported.** It is recommended that NSF support national infrastructure for data-intensive applications from the Geosciencesciences, including storage and database resources. For example, CISE/DDDAS and GEO could jointly support a solicitation in this area.

The result of following these recommendations will, we believe, both enable a few early successes of petascale geoscience applications solving important science problems running on the petascale facility when it first becomes available, and enable the evolution of a balanced, robust, interdisciplinary, computational geosciences community and infrastructure going forward.

2. Analysis and findings

The workshops were structured around the following broad questions:

I. What are examples of important questions and conceptual challenges in the geosciences that illustrate the potential impact of access to a petascale computational facility?

II. What strategies will ensure that the geoscience community is in position to take advantage of petascale computational capabilities?

III. What resources are needed to get teams of geoscientists and computer scientists working together on petascale applications development, with the goal of having operational packages ready by 2011 when petascale resources will come on line?

The remainder of this report describes the findings of around each broad questions.

2.A. Petascale Applications

We worked to identify candidate petascale applications in geoscience. The focus included critical issues, outstanding challenges, and potential impact, as well as on scaling up existing applications to petascale, and on turning important questions and conceptual challenges into petascale applications.

In the large view, some candidate petascale calculations in geoscienceare already extant as applications, some are even running at terascale (10^{12}), while others exist only as abstractions, models, and (in some cases) algorithms for solving them. For these latter conceptual problems, implementation is an issue; the quality of the implementation may be a larger factor than intrinsic suitability, or scientific importance, in determining their success at petascale. Even for the former cases (existing codes), there is no simple proof that either a) terascale applications will naturally scale to petascale, or that b) important science questions would be answered by so doing; rather, the suitability of each for scientific importance, scalability, and computational challenges, must be examined case by case. Still, in any case, validation is an issue. In many computational science problems, determining that a calculation is computing a result of high fidelity to nature and of scientific relevance, will be as or more difficult than implementing the application at petascale to begin with.

With the above larger issues in mind, three guiding principles were used by this group to identify possible candidate petascale applications in the Geosciences: 1) needs of the domain science are more important than simply enabling petaflop calculations 2) people are expensive, machines are cheap (or in other words software is expensive and hardware is relatively cheap), so designing, coding, porting, tuning, and validating applications will be at least as expensive as procuring petascale hardware, and as well are of utmost importance in this drive towards petascale computing. Furthermore, a rather ideological position was taken, that being that 3) “nothing scales to a petaflop, unless otherwise demonstrated or proven.”

With these guiding principles in mind, candidate questions and conceptual challenges were identified and deemed potentially suited for petascale within the available time, given sufficient resources, early start, and ample time for success. These questions are listed below in rough order of deemed readiness/nearness deployment at petascale, with the most mature candidates listed first. In addition, for each category of application, some risk/reward assessment is provided. Risk is loosely defined here as risk of either 1) failing to be deployed at petascale (i.e. due to difficulty of implementation) within the timeframe, or 2) failing to compute a result of significant scientific merit, or both. Reward is of course the opposite i.e. 1) likelihood of running at petascale at "first light" if sufficient resources are devoted to software development, and 2) importance of the underlying problem to geoscientists, or both.

As mentioned above, this is by no means an exhaustive list. It is not an exhaustive list of all the applications discussed at the workshops. The fact many applications are not on the list does not imply they are not important or not ready for petascale. In this report we we simply highlight some specific calculation we believe can use the petascale facility at first light to solve important science problems if sufficient funding and effort is put into the development starting now.

Mantle Dynamics: Featured Calculation "PetaShake"

Scientists believe that California is overdue for a large earthquake on the southern San Andreas Fault, and the opportunity is to better understand the basic science of major earthquakes and to apply this knowledge to prepare for them through such measures as improved seismic hazard analysis estimates, better building codes in high-risk areas, and safer structural designs, potentially saving lives and property.

The initial TeraShake simulations were unprecedented data-intensive simulations, producing more than 40 terabytes of data, revealed new insights into large-scale patterns of earthquake ground motion, including where the most intense impacts may occur in Southern California's sediment-filled basins during a magnitude 7.7 southern San Andreas Fault earthquake, and how basins flanked by mountains can form a "waveguide" that could channel unexpectedly large amounts of earthquake wave energy into the Los Angeles basin.

But the TeraShake simulations could reach a frequency of just one half Hertz, modeling only the lowest part of the frequency range of the ground motion. While these simulations provided information engineers can use to explore earthquake impacts on larger multi-story structures, more than 20 floors high, say, the much larger number of smaller structures remain "invisible" in the TeraShake simulations, which failed to capture the higher frequencies that interact with smaller multi-story buildings.

By reducing the computational grid spacing from 200 to 100 meters or even 50 meters, a PetaShake simulations will capture frequencies up one to two Hertz, providing information that can model earthquake impact on the larger number of smaller multi-story structures. In addition to higher frequencies, one can increase physical volume. These improvements in realism are very computationally intensive, however. Each factor

of two improvement in frequency resolution increases by a factor of eight the required spatial grid mesh and another factor of two in the timestep, for a total increase of 16, strongly driving the need for the next generation of petascale computing resources.

This calculation is considered low risk high reward as almost perfect scalability has already been demonstrated on the 40 thousand cpu BGW machine.

Weather and Climate: Featured Calculation "WRF Nature Run"

The development of the Weather Research and Forecasting (WRF) modeling system [1] is a multi-agency effort intended to provide a next-generation mesoscale forecast model and data assimilation system that will advance both the understanding and prediction of mesoscale weather and accelerate the transfer of research advances into operations.
It is suitable for use in a broad spectrum of applications across scales ranging from meters to thousands of kilometers. Such applications include research and operational numerical weather prediction (NWP), data assimilation and parameterized-physics research, downscaling climate simulations, driving air quality models, atmosphere-ocean coupling, and idealized simulations (e.g boundary-layer eddies, convection, baroclinic waves). For example, during the past three hurricane seasons, WRF has been run at NCAR in real-time to offer high-resolution (i.e., detailed) forecasts of storms which have threatened landfall.
The growth of computational power is enabling NWP model forecasts within the scale region defined by an observed $k^{5/3}$ scaling in the kinetic energy spectrum. We have much to learn about how waves and turbulence interact, affecting predictability and optimal sub-grid parameterization, within this region and across the observed transition to larger scales. Without this understanding, we cannot take full advantage of the computational power at our disposal. The version of WRF to be investigated would produce a suite of "nature runs" 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 would serve as the basis for study at grid scales beyond the capability of current HEC resources.

It is impossible to study predictability in the real atmosphere, making 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 would facilitate a new generation of predictability studies that are not currently possible. Simple identical-twin experiments on how errors grow within the $k^{5/3}$ regime and across the transition could be performed. The hypothesis of enhanced mesoscale predictability near topography could also be rigorously addressed. It is additionally extremely difficult to study turbulence in the real atmosphere, and therefore models are attractive. The turbulence community faces several challenges that currently cannot be addressed. Wave-wave interactions within the $k^{5/3}$ regime and across the transition are poorly understood. Wave-turbulence interaction occurs within the $k^{5/3}$ regime and across the transition, for example in the jet-stream region of the atmosphere. This nature run will contain 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 would allow the study of gravity waves in a realistic environment, and may include gravity wave breaking. Finally, new closure techniques are necessary for the next generation of NWP model implementations, and proposed stochastic approaches rely on spatially and temporally correlated statistics of mesoscale flows that are also extremely difficult to measure in the atmosphere. These must be quantified to take even the first step toward stochastic parameterization. A nature run that is long enough and with sufficient resolution could prove invaluable in beginning to understand their characteristics, which could then be exploited in next-generation parameterization schemes.

The WRF model is efficient on massively parallel systems, and offers the potential to make use of a great deal of computational power. For example, future goals include supporting a 1km grid resolution, which implies the use of approximately 7,000 petaflops per run, requiring about 4,000 hours on a current system, the "Blue Sky" system at NCAR, which is a 1,600 processor IBM Power 4 system.

The necessarily intensive computational resources required, and the use of a new NWP model, make this a risky endeavor with the potential for significant scientific and educational rewards.

Subsequent to the workshops, NSF GEO ATM funded an SGER (Small Grant for Exploratory Research) to investigate this computation. Substantial progress was made resulting in selection to the finals of the Gordon Bell Prize at SC07. The draft version of the accompanying paper as accepted to the SC07 Gordon Bell track is included as Appendix A.

Coupled Ocean to Weather: Featured Calculation "Hurricane Eyewall Intensity Forecasting"

Over the last a several decades, hurricane track forecasting has improved significantly, whereas very little progress has been made in hurricane turbulence, wind intensity, and rainfall along the track. We know where the storm is going but not what it will do when it gets there. Our lack of the skill in intensity and rainfall forecasting may be attributed to two deficiencies (among others) in current computational prediction models: insufficient horizontal resolution and lack of full coupling to the ocean. Extremely high winds around the eye, intense rainfall, large ocean waves, and copious sea spray push the surface-exchange parameters for temperature, water vapor, and momentum, into regimes unreachable at current levels of model resolution and coupling.

A key to improving hurricane intensity forecasts is to have numerical simulations that are capable of resolving the inner core structures (eye and eyewall) and rainbands in a hurricane and can realistically represent the physical processes governing intensity change, such as the transfer of heat, moisture, and momentum at the air-sea interface, the phase changes of moisture in the atmosphere, as well as flux-carrying turbulent eddies that affect mixing. The next-generation prediction simulation models should then be able to resolve features with very high resolution to capture the gradients across the eyewall

boundaries (at ~1 km or less), but also capable of representing the turbulent mixing process correctly (at ~10 m or less). Rapid increases in available compute power, and recent advance in technology in observations, have made it possible to consider a strategy for deploying the next generation of high-resolution hurricane prediction models. We begin by examining key issues related to grid resolution.

A difficult question to address is the convergence of solutions for hurricane intensity and structure at high grid resolution. An important part of the science is the generation of resolved turbulence in the model. We have conducted simulations on a grid as fine as 185 m. In order to obtain the same intensity as on a grid of 555 m, the viscosity coefficient in the turbulent kinetic energy scheme had to be increased (doubled), upon which the results of coarser resolution were reproduced. The model did not generate turbulence on its own with sufficient intensity to provide the necessary mixing to equilibrate the storm. We would like to explore if, with a grid spacing of 10 m, there will be sufficient turbulence generated to cease intensification. This exploration requires a petaflop calculation.

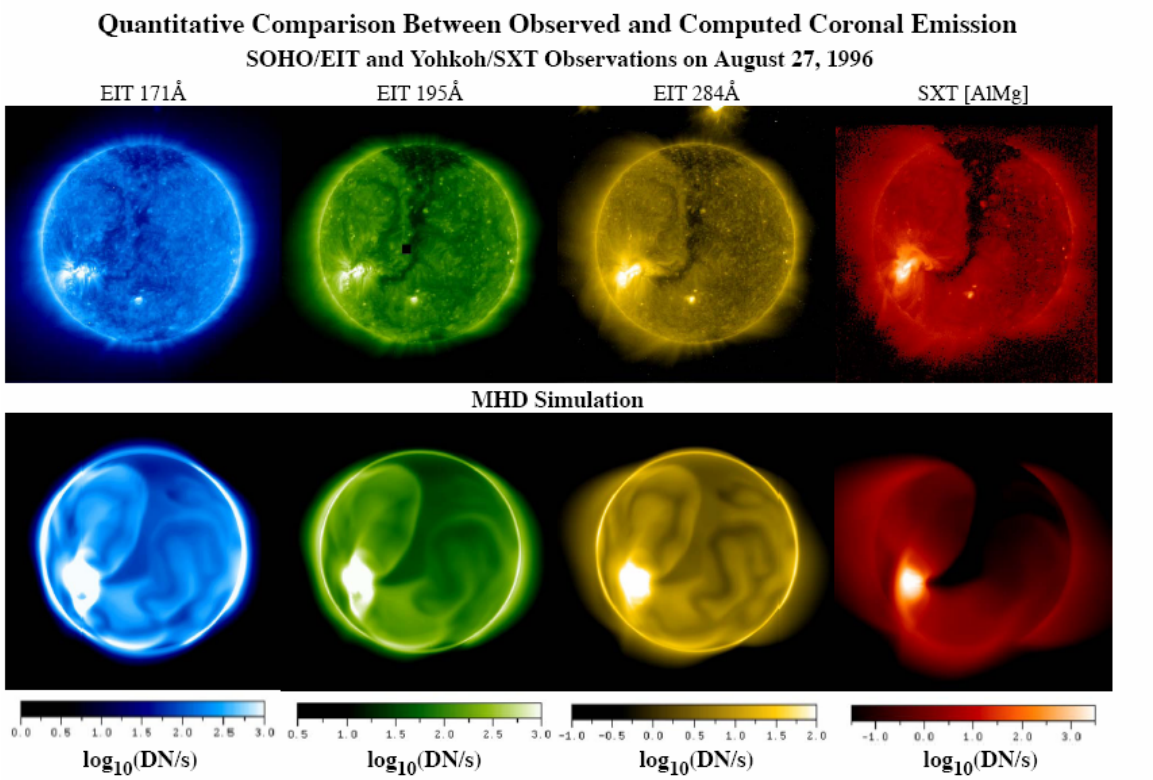
A related issue: for an idealized, azimuthally invariant initial vortex in a uniform environment, only small-amplitude asymmetries in the eyewall were noted. When the resolution reached 185 m, and with larger mixing (such that the storm intensity was the same as at 555 m), there were pronounced asymmetries that developed in the in the eyewall. An important question, “is what is generating these asymmetries?” and what will happen to them when the resolution is increased by an order of magnitude?
To explore these science questions, one could conduct a number of idealized WRF model simulations with 10 meter horizontal grid spacing and 150 vertical levels. Computationally, this entails a 15-billion cell inner-most WRF 10 meter nested domain, with a model time step of 60 milliseconds, and we roughly estimate it will require 18 machine hours per simulated day at a sustained petaflop/second if we can work through issues of scalability, load balancing, and I/O to enable an efficient petascale deployment. In addition to the model grid resolution exploration, the intensification and decay of a (real or modeled) hurricane largely depends upon two competing processes at the air-sea interface: 1) the heat and moisture fluxes that fuel the storm and 2) the dissipation of kinetic energy associated with wind stress on the ocean surface. Air-sea interaction is especially important in the region between the center of the eye and eyewall where there are extremely large gradients in the wind, temperature, and pressure fields. We wish to further explore the effect of air-sea coupling at very high resolution by using coupled modeling system including a surface wave model and ocean circulation model at grid spacing of ~1-2 km. This addition will increase the size of the computational problem by about 2-fold in data and flops.

This calculation is considered high is high reward as not all the physics governing highly turbulent storms is fully understood.

Space Weather
The National Space Weather Program (NSWP) has defined ``Space Weather'' as the conditions on the Sun and in the solar wind, magnetosphere, ionosphere, and

thermosphere that can influence the performance and reliability of space-borne and ground-based technological systems and can endanger human life and health.
As our nation becomes more dependent on advanced technology, it can become increasingly vulnerable to space weather. This is why the National Science Foundation has taken a leadership role in the National Space Weather Program. The solar magnetic field is the ultimate energy source for the most serious space weather effects. Through the medium of the solar wind, coronal mass ejections (CMEs) and solar energetic particles (SEPs) propagate and interact with the Earth's magnetosphere and cause some of the most serious space weather effects. The MAS (Magnetohydrodynamic Algorithm outside a Sphere) code [1], proposed for evaluation, is a state-of-the-art magnetohydrodynamic (MHD) model for realistically predicting the properties of the solar corona, and for investigating the phenomena of solar activity such as CMEs. The MAS code is presently being utilized in the Center for Integrated Space Weather Modeling (an NSF Science and Technology Center based at Boston University) as one of a suite of coupled physics models to describe space weather in the entire Sun-Earth environment.

The solar corona exhibits a wide range of plasma regimes, from strongly magnetized, slowly flowing plasmas low in the corona near sunspots to supersonically flowing plasmas in the solar wind. A successful numerical model must calculate efficiently in these different regimes**.** The MAS code is relatively mature and has been in use for some time for modeling the solar corona and solar wind. It is built on a rich base of experience in computational physics and the modeling of solar coronal and fusion plasmas. The code has been designed to work effectively in these regimes, but it has not been exhaustively tuned for performance. One of the principal difficulties in the physics and thus in turn for the performance of the code is the wide disparity of time scales present in the coronal plasma. Solar phenomena are often characterized by a slow evolution followed by an impulsive, rapid response as a result of instability, loss of equilibrium, and/or magnetic reconnection. The complex physics makes the computation in turn very challenging; there is evolution of boundary data, the need for staggered, non-uniform structured meshes and implicit and semi-implicit differencing. The comprehensive physics model includes energy transport processes (radiation, parallel thermal conduction, and coronal heating). The MAS model has already run with reasonable (~30%) efficiency on one thousand processors of SDSC's DataStar. The result of a calculation of the solar corona for a specific time period is shown in Figure 1 and compared to observations.



Calculations with ~10 million grids take several days on DataStar to compute several days of real time. These calculations just scratch the surface of what is possible. In the present calculations, the photospheric magnetic field measurements (magnetograms) used as boundary conditions are much coarser then are available. Incorporating higher resolution magnetograms is crucial for understanding the details of coronal structure and eruptions of the magnetic field, but will require more than order of magnitude increase in the total number of grid points

This calculation is considered high risk high reward as the underlying physics, particularly of coronal loops, are but poorly understood.

Ecological Component of Earth System Modeling: Featured Calculation "Groundcover for Climate"

There is an opportunity to use a petascale facility to enable adding ecological information, such as forest growth, to models of weather and climate. A grand challenge in geoscienceis the addition of clouds to ecological modeling. However, clouds and cloud-formation processes interact (in both directions) with the ecosystem.

Ecological models attached to a high-resolution model of climate change (or ocean circulation, etc) that already has clear petascale applicability on its own may be a good path forward. It is debatable whether ecologists will have much say in the design or deployment of these models, unless they begin to collaborate with the climate modelers

as soon as possible. Ecologists have several embarrassingly parallel applications that can be "easily" scaled to a petascale system, including:

- stochastic processes that need replication
- parameter sensitivity analysis
- heuristic optimization

Such problems are extremely common in ecological application, as we elaborate here.

1. Stochasticity: Simulation-based ecological models often incorporate demographic stochasticity (random birth/death/movement, etc), environmental stochasticity (random components of climate forcing, resource availability, etc), and/or genetic stochasticity (random mating, mutation, etc). Outcomes are thus stochastic as well, and ecologists wish to ask questions like, "What is the simulated probability that the population size will fall below X within 100 years?" The simulation model must therefore be independently repeated (usually 100s-1000s of times) to generate a distribution of outcomes.

2. Parameter sensitivity (or more generally, model sensitivity): The "true" parameters of ecological models are rarely known, and in fact there are often disagreements about the form of the equations governing those processes. Consequently, ecologists frequently want to characterize the sensitivity of outcomes to input parameter values and model assumptions. This also requires repeated simulation.

3. Optimization: There are (at least) two distinct types of optimization questions that ecologists commonly ask. The first involves fitting parameters to observed data. In all but the most trivial models, it is impossible to use analytical or even simple approximating techniques to identify maximum likelihood estimates of parameters. Increasingly, ecologists are turning to stochastic optimization techniques such as simulated annealing, or the use of various implementations of Markov Chain Monte Carlo to simulate posterior probability distributions in a Bayesian framework. Secondly, applied ecological models often implement heuristic optimization algorithms as decision tools (e.g. identifying the optimal spatial configuration of a land reserve system, given some cost criterion). As with parameter estimation, the simpler algorithms used in the past have been shown to be deficient in complex settings, but more reliable methods require many repeated simulations and long run-times. There is a tremendous need for HPC solutions that can deliver results sufficiently quickly even for models involving many parameters, fine-scale spatial and temporation resolution, and stochastic processes.

Putting this all together, it is clear that the compute time can be overwhelming when coupling one or more the above procedures with even a moderately complex ecological simulation model. Specifically, some model examples include predicting evolution of a collection of interacting species, spatial spread of a disease, or the dynamics of a specific

ecosystem. Taking the last example, imagine a regional-scale ecosystem model, the core of which is deployed as a small-scale HPC application (e.g. a single simulation that takes days to complete on a cluster with dozens of nodes). Indeed, the ATLSS group (http://www.atlss.org) based at UT/ORNL has spent ~10 years developing and refining a model that integrates a variety of complex and interacting submodels to simulate key Geoscienceand environmental components of the Florida Everglades; AFAIK, one submodel has already been parallelized to run on 60+ nodes. Even if a researcher demands just several hundred stochastic replications in such a simulation, performed for each of 100 possible configurations of a proposed reserve system, there would be significant benefit from hierarchically parallelization, to enable a 100k-processor system run (imagine a multi-hundred simultaneous, distributed instantiation of the ecosystem simulation, which itself might be a 64-node data-parallel application). Whether the envisioned petascale system even provides the right architecture for this application could be debated, but the point is that it does not require significant effort to scale up moderately sized ecological models to result in large computational needs, resulting in the ability to address relevant and interesting problems.

Data integration would be critical to success. A candidate calculation would involve evolutionary correlations of networks and functions (phenotypes). In the worst case, such formulations may lead to NP Hard/Complete problems that would remain intractable even at the petascale level. To the extent that ecologists are able to refine mechanistic mathematical models in a way that is increasingly faithful to reality, one could easily conceive of petascale computing demands for simulating an entire ecosystem from its underlying biological and physical components. However, it is worth pointing out that the tradition in ecology is to simplify and scale back models—indeed, to err on the side of oversimplification; “realistic” models have long been mistrusted in favor of either highly abstract mechanistic (theoretical) models and/or simple phenomenological (statistical) models. In part this is for good reason: ecologists do not yet fully rely on their own more detailed mechanistic models (there being a lack of the ecological equivalents of physical laws, testing approximating models via experimentation and observation is difficult, and each real system seems to have its own unique features). This could in fact partly be historical artifact: few ecologists are even aware of the computational possibilities now afforded by HPC systems. In a sense, one might argue that developments in this area are limited due to apparent belief in computational obstacles that no longer exist—something we believe can be remedied through education and workshop opportunities. Given opportunities for making forays into developing complex ecological simulations despite uncertainties about the models, and having the ability to enhance model output with observed data, has the potential to lead to refinement and progress in this area.

Choosing applications focused on the coupling of ecological with weather and climate models, and supporting them for further work for deployment at the petascale level, appears to be relatively medium risk for success in the given time frame. Much of the risk is dependent on collocating two disciplines that normally do not interact collaboratively, to facilitate synergistic code co-development. Computational possibilities in this area have the potential for relatively high reward towards addressing interesting geoscienceproblems. The path towards petascale is relatively straight forward, and the

science questions of great relevance for better understanding of our environment by the ecological communities.

2.B. Petascale Planning

This group dealt with strategies to prepare for petascale computing, such as: 1) the need for interdisciplinary collaboration, 2) selection criteria to determine which interdisciplinary teams to sustain (given limited funds and resources), 3) software challenges on the way to petascale, and 4) community-wide organizational structures that may foster the long-term needs of petascale computing.

1. Need for Interdisciplinary Teams

The petascale funding announcement has stimulated an intense and broad degree of interaction between computational geoscientists and computer scientists. Our goal is to build a focused R&D effort that simultaneously measurably impacts the quality of life across society, demonstrates leadership and vision in large scale computing for the geoscience sciences, and trains the next generation of engineers, scientists, and mathematicians. To realize our goals we need to formalize the notion of "petascale computing" into an activity worthy of computer scientists' and geoscientsists dedicated effort, and that, in turn, enables them to succeed in their research career and in their core mission of training students. To create and strengthen a community, computer scientists and geoscientists, and their postdocs and students, should have the possibility of receiving travel and living support to visit each other or the petascale site for months at a time.

Long term funding, with a minimum of 5 years, will be needed to ensure continuity, and it should cover all participants: computer scientists, geoscientits, programmers, and so on. This funding should encompass not only the research itself, but the sustained development of application and software libraries as they grow and require maintenance.

Projects should specifically be constructed as interdisciplinary teams working together to deploy geoscience applications at petascale. It will be crucial to have models for performance analysis and scaling to show that an algorithm scales *before* implementing it. There is also need to develop models for quantifying uncertainty (accuracy) and effort put into quantifying performance and accuracy of heuristics (performance modeling); these activities perforce require interdisciplinary collaboration.

2. Selection criteria for determining suitable petascale applications

It is important that selection criteria lead to the formation of outstanding and productive collaborative teams focused on petascale activities. Historically, each 10-fold increase in parallelism required significant time investment, even when the raw compute capability was already available. Finally, since the petascale system will be a scarce resource during the first few years, its use must be managed transparently so it is in line with the selection criteria, and the needs of the funded efforts.

Therefore it is important to understand which applications are "peta ready" today (as was partially addressed by the list of applications above), and which will be ready tomorrow. Moreover, one must also prepare for the emergence of novel classes of calculations and applications that may be as important, or more, and as suitable for petascale as those of today.

One approach would be to distinguish among focused and large-scale development efforts and then apply the appropriate selection criteria and level of funding to each. For applications deemed "peta-scalable" by first-light (2011), the following criteria are important: qualifications of the team, size of the potential user base, suitability of the calculation for petascale resources, and the balance between scientific merit and impact versus risk and feasibility. Specifically, some of the most peta-scalable applications may be relatively low risk but lead to less pratical results; while other applications likely to be of great geoscience significance are clearly at the edge of feasibility. It is important that application be evaluated along both dimensions of the risk-reward space and that the resulting portfolio mirror this spread.

The mix of applications may thus include some that are nearly ready for the petascale, others of high importance that require additional software engineering and algorithm improvement, or yet others that require a build-up phase towards substantially more compute-intensive usage. Other types of diversity to consider for support should include both monolithic single applications that spread across the entire machine, and heterogeneous applications where different components of the applications workflow stress the computer architecture in different ways.

3. Benchmarks

Any benchmark suite, to be useful, must measure end-to-end system performance, a key metric for getting science done. In particular, benchmarks that measure the I/O subsystem performance in a variety of application driven situations are critical. The vendor community in attendance strongly suggested that the application suite contain production benchmarks. This will provide the vendors with a revenue incentive when considering working with these benchmarks. This view was tempered by the idea that benchmark applications should come with unit tests to facilitate the porting and validation process.

A benchmark suite was proposed consisting of the WRF mesoscale meteorological model, the POP 2 global ocean model and. Since the workshop, the CICE-4 sea ice model and the Southern California Earthquake Center (SCEC) TeraShake application have been put forward as additional benchmarks, and the parallel NetCDF I/O benchmark IOR has been suggested as the first I/O code for the GARPA benchmark suite. It is anticipated that other candidate benchmarks will emerge at the second workshop.

4. Software challenges on the way to petascale

An overwhelming consensus from the group, and indeed the workshop attendees across all groups, is that there is a need to increase support for theory, algorithm and code

development, and software engineering to get the best science out of petascale and other high end machines.

A possible format for such support would be a two stage competition for software engineering centers, with the first stage being planning grants and the second stage being full software engineering centers, with each center focusing on a particular application or class of applications. The subject of these centers should include applications services as well as applications themselves, as both levels need considerable effort. There should be provisions for individual investigators who have distinct contributions to make, to contribute to the work of the centers without being an integral part of the centers.

Issues of intellectual property for software rights will be critical to work out in advance, as intellectual property considerations could cripple the dissemination of associated advancements in software, and the development of heterogeneous computing environments.

To ensure that the geoscience community is positioned to take advantage of petascale computing capabilities, it is important to support efforts sustained efforts for the long term, i.e. a "20 years" timeline. A minimum of 5 years is recommended to establish retooling of (large) software, in order to successfully to take advantage of what a petascale system has to offer. Other software will be enabled more quickly, but to focus on only these would be irresponsible to the field of geoscience and the opportunities being presented by establishing a petascale facility.

For support of applications, there is a need to invest in application libraries. With the help of source-to-source transformation (for example, Dan Quinlan's ROSE effort at LLNL) it is possible to implement semantic optimization of class libraries. Efforts of this type can greatly facilitate geoscientists, enabling them to work in terms of abstractions they are comfortable with. Additionally, however geoscientists can greatly benefit from learning key aspects of computer science in focused training sessions key to their application areas (see below.)

Some problems pose challenges for hardware/algorithms (PME/Fast Fourier Transform) and may be suitable for more innovative algorithmic approaches/hardware. Several candidate applications do not fit into traditional numerical frameworks. For example, they may not be floating-point operation intensive or perhaps have poor spatio-temporal locality. This underlines the importance of identifying the computational needs of the candidate applications in advance of deployment to insure petascale infrastructure is appropriate to their needs (which should be among the first activities of collaborative groups).

Many of the applications identified require multiscale modeling, relevant to many areas of science and engineering including but not limited to computational geoscience (for example, the DoE has such a program for material science). There are various software and algorithmic challenges inherent in such models, in addition to the obvious physics/chemistry challenges of coupling across scale. In terms of the former challenges,

for example, managing the components in a petascale system, in particular, "cross-component optimization", is an important issue. That is, how to optimize multiple components taken together, rather than the classic approach of optimizing individual components separately. Thus, issues of load balancing, tolerating communication delays, memory hierarchy optimizations, etc., all increase complexity of the problem. All of these issues raise interesting research questions, as well as the need for software techniques needed to handle the optimizations. These issues are related to work being performed by the common component architecture (CCA) community, but complimentary in this domain. Frameworks are needed to express such optimizations, and to enable application developers to express, in the form of "application performance meta data", the information needed to sensibly and collectively optimize the multiple components handling the different parts of a composed multi-scale simulation.

Scalability is a primary requirement for petascale; achieving it involves communications optimization, load balancing, asymptotic complexity analysis, and numerical accuracy assessment. When different models are composed, it is important that their numerical interactions be consistent, stable, etc. Models can not always be composed in a straightforward way, and it is important to understand their collective behavior and associated physical laws. Load balancing across different model; scales and data layouts is particularly hard and ill-understood at present. More research is needed in this domain, and this need is urgent in light of the number of likely processors in a petascale system (conservatively lower-bounded at 100,000).

Many of the issues arising out of multi-scale models are potentially relevant to *any* application running on a multi-scale system. Tolerating communication delays, and handling load balancing are much more difficult at extreme levels of parallelism (10^5 to 10^6 processors), because done incorrectly, serious waste in compute resources results. Thus, software techniques that facilitate latency tolerance will play an important role in helping ensuring scalability and efficiency of resources.

5. People and Infrastructure organization issues on the path to petascale

Strategies for long term preparation of potentially new petascale applications include the development of training programs, workshops, and summer schools, with focus on teaching the craft of creating efficient codes. At the same time, sufficient effort needs to be afforded to parallel geoscience application middleware development including appropriate run time systems, frameworks, and libraries. Leveraging of existing and successful programs in these areas is highly desirable.

There is a need to pay specific attention to the workforce pipeline for the applications software development and engineering required. In addition, more focused training programs that solicit proposals from multidisciplinary teams, to jointly train students in high performance computing for geoscience, should be considered. We have a consensus that high performance computing is relatively neglected in computer science departments around the country, and that this is a problem that should be addressed by incentives to train students in high performance computing.

There are also societal issues to be considered. Petascale computing is not simply an extrapolation of prior experience, but raises the stakes for application developers. Raising the scale of parallelism by a factor of 10 will compel some to rethink the algorithm, the implementation, or other software issues, including development costs. The impact of raising the level of parallelism by multiple orders of magnitude opens up considerable opportunity for the geoscience community to take advantage of. Taken from a different perspective, there are quite possibly unknown applications that become possible with the advent of petascale parallelism, which has the potential for opening up entirely new avenues for inquiry in Geoscience, Computer Science, Math, and Physics. These observations bring up two (possibly opposing) viewpoints.

i. Petascale computing will start out as a small club gradually becoming more widespread over time, at which point we begin again with exaflop (10^{18}) computing. It is important that a small number of likely success stories be chosen at the outset, endowed with the human resources needed to thrive, enabling the demonstration of the capabilities of the machine early on.

ii. Appropriate training should be provided with a long term vision of developing new communities of users: to ensure that a critical mass of experts be available to respond to and disseminate information about new developments, and to develop new software techniques that will be useful to the computational geosciencecommunity at large. Summer camps and summer schools should be held to teach computational geoscientists the latest software techniques, and to teach computer scientists the latest trends in computational geoscience. This should be an ongoing process. Resources should be set aside for innovative ideas to blossom, be they in computer science, computational geoscience, applied mathematics and physics, or other related fields. Small time users should be given the opportunity to use the full scale machine for trying out their ideas.

We believe there is a middle way, embodied in the proposal above to select a limited but diverse portfolio of applications and teams, and that this can foster both viewpoints. The result should be some early success at “first light”, but also more speculative ongoing research and innovation that can lead to new uses of the petascale computational facility to address emerging questions in geoscience.

2.C. Petascale Infrastructure

This group dealt with identifying the computing, software, storage, networking, and people infrastructure needed for geoscienceat the petascale level. A high-level take home message is “*invest in people and software at (or preferably) beyond the level of the hardware investment!”*

Hardware infrastructure
Geoscience problems of interest, for example those identified above, are data intensive, compute intensive, communication intensive, in variant combinations, and one size does

not fit all. Geo-computing will therefore need multiple types of architectures and resources that map to the diverse hardware portfolio planned by NSF in their "tiers 1, 2 and 3" planning. However a form of "market segmentation" needs to be done to determine which calculations should be done where and to influence some of these architectures to be designed with the special needs of geoscience computing in mind. This is related to the idea that interdisciplinary teams need to start by understanding, modeling, and extrapolating future application requirements before embarking on ambitious code development projects.

Several of the candidate applications described above could benefit from application-specific architectures. Matching heterogeneity in applications and architectures across the NSF portfolio will be very important. It is expected that many candidate petascale applications in geoscience will be bandwidth intensive, with respect to local memory and with respect to inter-processor network bandwidth demands. In fact some key applications may turn out to be solved faster on what NSF terms "tier 2" systems than the petascale system, if those systems are better balanced in terms of memory and communications bandwidth per-processor. Therefore, it will be important to study computational characteristics of applications and associated hardware characteristics in advance to identify memory and communications bandwidth sensitivities. Likewise, it will be important to quantify what portions of candidate calculations are very computationally intensive and could be carried out on coprocessors (such as ASICs, FPGAs, DSPs, GPUs) likely to become available in the same timeframe. Likewise, it will be important to understand which applications are very communications and I/O intensive and will stress machines in these dimensions. In the design of any petascale or "tier 2" I/O infrastructure, it will also be important to address data federation, data availability, and integrity. Also integration of data acquisition systems (sequencers, microarrays, imaging) needs to be addressed more than at present.

Deeper considerations of the specific needs of compute intensive versus data intensive geoscience applications need to be made, as these may not be easily separated. Data intensive applications require stable and scalable file systems, and infrastructure for moving the data in and out of the computer. When such an application is generating hundreds (or thousands) of petabytes, the infrastructure must support storing data, mining and analyzing data, moving data, archiving data, and visualizing data.

By the same token, compute intensive calculations come in different types requiring (1) considerable amounts data, (2) considerable numbers of CPUs, (3) considerable amounts of memory (4) real time/wall clock constraints and (5) combinations of all the previous. Also, even compute intensive applications are not always computing "just one number" as the output. Rather many will generate petabytes or more of output data even if they did not consume a similar amount of input data to start with. Thus even these may require data intensive post processing even if the petaflop calculation is not by itself data intensive.

A related issue, particularly in the cases of applications requiring the movement of vast amounts of data, concerns geosciencenetwork issues. Schemes need to be developed for

petabyte data transfer via the internet. This may require upgrades to existing national backbone networks but also, quite seriously, this may involve Fedex ala NetFlix (order data via the Web for next day arrival).

All of these challenges imply rethinking out-of-the-box around new architectures for geoscience computing, not just focus on refitting of existing geoscience problems to fit the petascale (or other high level) facility.

Software Infrastructure

Software costs more than hardware. A strong consensus of the workshop participants is that currently there is an imbalance in NSF support for scalable, robust, easy to use scientific software relative to proposed investments in hardware. Enabling petascale computing in geoscience will require software infrastructure enabling data analysis, mining and visualization. Analyzing massive output data and visualizing will then require more than just high floating-point capability by way of infrastructure; candidate petascale geoscience applications present tremendous issues associated with data handling (federation of data sources as for example expression, sequence, phenotype, etc.). These applications will stress I/O and file systems, and data federation solutions. Algorithms will need to be developed to deal with uncertainty in data, missing data, and erroneous data (sensitivity analysis). Furthermore, there are two key issues involving interactivity around large-scale data: (1) inordinate amounts of data to move, store, analyze requiring infrastructure supporting interaction (2) human beings often will need to be in the analysis loop. Additional challenges are associated with the connection of sensors and data streaming, as data access rates for I/O become very important.

Viable infrastructure will also need to include scalable codes, scalable algorithms, and scalable memory as well as lots of cpus as was expounded upon by Working Group II above.

In addition to the significant work required in fundamental algorithms and load-balancing, latency tolerance methods, etc., as described, significant efforts need to be focused in the areas of queuing and scheduling. Currently, queuing and scheduling systems do not do a good job of handling different types of needs. There is minimal ability to schedule high-performance computers to accommodate real-time constraints, respond to embedded sensors, be available on demand, and the like uses of interest to geoscientists. Current scheduling policies primarily service throughput jobs. While it may be that this is also deemed to be the best way to manage the petascale system, there is significant doubt. Likely, increased programmer productivity, increased breakthrough science, and better response to the end-user may result from a less heavily loaded resource; one that is reserved for fewer truly petascale calculations, including perhaps some with real-time constraints, rather than the currently heavily loaded, highly utilized, NSF systems.

Additional issues related to software involve fault tolerance issues, which are currently not being adequately addressed in designing software infrastructure for petascale. Given

likely state-of-the-art reliability and hardware failure rate trends, it is anticipated that one processor out of one hundred thousand (or a million) will fail every minute on a petascale machine. Who and how does one deal with such failures/minute? By way of example using current semantics, an MPI global operation will block if even one processor fails to respond resulting in code hang-up. Applications need to be re-written to be fault-tolerant, something currently not even possible without updating current semantics of MPI to enable more tolerant of failures. Either vendors need to develop, or the community needs to develop (more likely the latter) fault tolerant APIs and associated semantics for global message-passing systems, in order to enable large parallel codes to be re-written in a fault-tolerant style.

Generally speaking in high-performance computing (not just in geoscience), there is a dearth of scalable, robust, easy to use, interactive, etc. software tools. Many tools that do exist for the purpose are "professor-ware", so there are lots of tools but not always with the required reliability or associated documentation. NSF could take the lead in finding mechanisms to fund enduring and stable tool efforts and in requiring periodic peer-review of ongoing tools projects.

Networking Infrastructure

People Infrastructure

People cost more than hardware. It is important there be a proportional investment in the people - faculty, staff, and students - who will support the necessary and vital efforts to obtain petascale computing levels. Additionally, true peer collaboration is hard and circumstances must be fostered to overcome discipline silos. Interdisciplinary teams are necessary, involving geoscientsits and computer scientists, but also other key disciplines (e.g., math, physics, sociology, economics, etc). Interdisciplinary teams may take the form of 2x2 collaborations as well as "service shop software models". It is crucial these teams obtain early access to software and hardware at the teraflop level and higher on the path to petascale as it becomes available in order to prototype algorithms and software up to petaflop level computing.

WORKSHOP ATTENDEES:

Richard Moore (SDSC).
Allan Snavely (SDSC).
Thomas Jordan (USC)
Alan Wallcraft (NRL)
Rich Loft (NCAR)
Rob Pennington (NCSA)
Michael Wehner (LBNL)
Darren De Zeeuw (University of Michigan)
Shijie Zhong (University of Colorado)
Laura Carrington (UCSD)
Kathy Yelick (Berkeley)
Pat Worley (ORNL)
Alan Sussman (UMD)
Shirley Moore (UTK)
Otto Fringer (Stanford)
Michael Gurnis (Cal Tech)
John Lyon (Dartmouth)
John Dennis (UCAR)
Bill Putman (NASA)
Omar Ghattas (UT)
Stephen Thomas (UCAR)
Bill Skamarock (UCAR)
Jeroen Tromp (Caltech)
Kraig Winters (UCSD)
Venkatramani Balaji (GFDL)
John Linker (SAIC)
Charles Goodrich (BU)
Kraig Winters (UCSD)
Yifeng Cui (SDSC)

Appendix A.

WRF Nature Run

Josh Hacker, John Michalakes, Rich Loft

University Corporation for Atmospheric Research (UCAR)
{hacker, michalak,loft}@ucar.edu

Michael O. McCracken, Allan Snavely, Nick Wright

San Diego Supercomputer Center
University of California
{mmccrack,allans,nwrightl}@sdsc.edu

Tom Spelce, Brent Gorda

Lawrence Livermore National Laboratory
{spelce,bgorda}@llnl.gov

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 provide 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 4 racks of BG/W with achieved 4.59 Tflops; we anticipate achieving > 73 Tflops on 64 racks of BG/L at LLNL. The primary result is not just the high Tflops number, but an important step towards understanding weather predictability at high resolution.

Categories and Subject Descriptors

J.2 [**Physical Sciences and Engineering**]:– *Chemistry*, *Physics*.

General Terms

Algorithms, Measurement, Performance.

Keywords

Weather prediction, WRF

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 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 4 racks (4096 nodes, 8192 CPUs) of BlueGene/L at IBM Watson (BGW) with achieved 4.59 Tflops which is 20% of theoretical peak on this I/O intensive application.

We anticipate using 64 racks (65,536 nodes, 128k CPUs/ of BlueGene/L (BG/L) at Lawrence Livermore National Laboratory (LLNL) to achieve > 73 Tflops which is ~25% of Linpack rating (280 Tflops) on this I/O intensive application. We will report LLNL run results in the final paper.

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 the above level of performance is challenging. The primary result is not just the high Tflops number, but an important step towards understanding weather predictability at high resolution.

a. the 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 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 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.

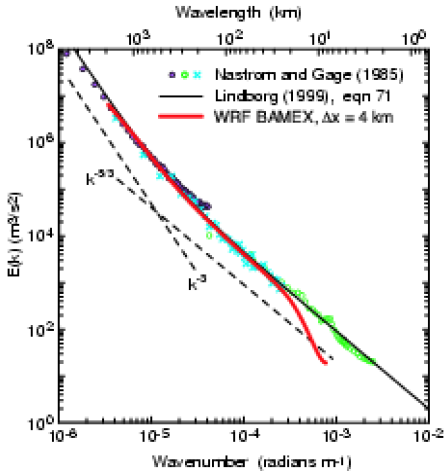


Figure 1 (courtesy W.C. Skamarock): Spectral energy density in the WRF model compared to observations. The red curve is spectra computed from WRF forecasts at 4 km grid spacing, averaged from 3 May 2003 to 14 July 2003. Both the transition of the spectral slope from $k^{-5/3}$ to k^{-3}, and the numerical 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.

b. 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) architectures such as BG/L. 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:

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

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, the flux form equations are expressed as

$$Ut + (\square \cdot Vu) + Px(p, \varphi) = FU \qquad (2)$$
$$Vt + (\square \cdot Vv) + Py(p, \varphi) = FV \qquad (3)$$
$$Wt + (\square \cdot Vw) + P\eta(p, \mu) = FW \qquad (4)$$
$$\Theta t + (\square \cdot V\theta) = F\Theta \qquad (5)$$
$$\mu t + (\square \cdot V) = 0 \qquad (6)$$
$$\varphi t + \mu{-}1\,[(V \cdot \square\varphi) - gW] = 0 \qquad (7)$$
$$(Qm)t + (\square \cdot VQm) = FQ \qquad (8)$$

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 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.

Solving these equations involves primarily using the Fortran intrinsics log, exp, power (**), & sqrt which in turn have a very efficient implementation on BG/L taking advantage of SIMDizatiom in an extensive set of parallel machine instructions for which the double precision operands can come from the register file of either unit and that include a variety of paired multiply-add operations, resulting in a peak of four floating point operations (Flop) per cycle per core.

4. Data Issues

The nature run is quite data intensive with a large sum memory footprint. During the BGW run 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 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 a 1800x2700 grid with 101 levels, resolution is 5km, time step of 10s, which should require, at minimum 51,200 CPUs.

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 1800x2700, 1010 levels, on 51K+ processors.

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 is currently a bug in Parallel NetCDF that IBM is working to fix which will, believe, give us better efficiency than the BGW runs indicate.

6. Performance Measurement

Floating point operations were counted using the APC performance counter library. This library accesses the compute node ASIC's hardware performance counters to tracks several events including FPU, some SIMD and load and store operations.

.

7. Results.

Using 4 racks (4096 nodes, 8192 CPUs) of BlueGene/L at IBM Watson (BG/W) we achieved 4.59 Tflops which is 20% of peak on this I/O intensive application.

We anticipate using 64 racks (65,536 nodes, 128k CPUs/ of BlueGene/L (BG/L) at Lawrence Livermore National Laboratory (LLNL), and a performance bug fix to Parallel NetCDF, to achieve > 73 Tflops which is ~25% of Linpack on this I/O intensive application.

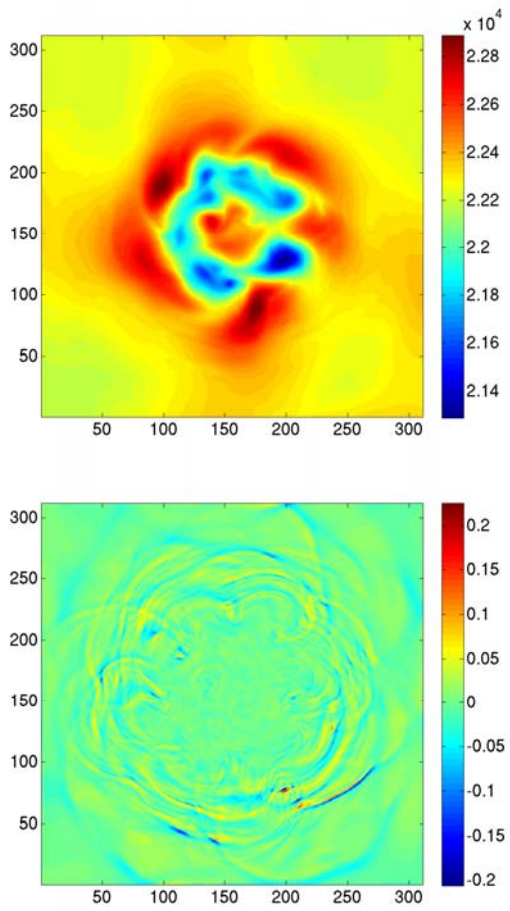


Figure 2: Northern-hemisphere polar stereographic projections of the WRF model state. Above, pressure (Pa) on model level 59 (approximately the level of the jet stream). The mid-latitude wave-train is evident as the ring of lower pressures (blue). Below, the corresponding vertical velocity (m s-1), showing 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. Conclusions.

Figure 2 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 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 64 racks of BG/L with anticipated achieved > 73 Tflops.

9. Acknowledgements.

We wish especially to thank Bob Walkup of IBM who hosted us at BGW Day and who helped us solve bugs in Parallel NetCDF. We also wish to thank William Skamarock, whose equations and description of WRF's numerical formulation is reprinted with permission in Section 3. This work was supported by NSF Award #0637994 SGER: Feasibility of Taking the Weather Research and Forecasting (WRF) Model to Petascale.

10. 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 Networkingand Computing, 2002.

5. E. L. Bachega, 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
http://www-unix.mcs.anl.gov/parallel-netcdf/