NGI/CE/07

Development of a Prototype NOAA Grid

FY 2004 Proposal to the NOAA HPCC Program

August 11, 2003

| Title Page | Proposed Project | Budget Page |

Principal Investigator: **Mark W. Govett**

Line Organization: OAR
Routing Code: R/FS5
Address:

Forecast Systems Laboratory
Advanced Computing Group
325 Broadway
Boulder, CO 80303-3328

Phone: (303) 497-6278
Fax: (303) 497-6301
E-mail Address: Mark.W.Govett@noaa.gov

Daniel S. Schaffer Daniel.S.Schaffer@noaa.gov
Brian D. Gross Brian.Gross@noaa.gov
Christopher W. Moore Christopher.Moore@noaa.gov

Proposal Theme: **Next Generation Internet (NGI)**

Signature 1 (required)	*Signature 2 (required)*
Mark W. Govett	Alexander E. MacDonald
Computer Scientist	Director
Forecast Systems Laboratory	Forecast Systems Laboratory

Development of a Prototype NOAA Grid

Prepared by: Mark W. Govett

Executive Summary:

As follow-on to the FY ’03 investigation of grid computing technologies, we propose to setup a prototype NOAA grid. It will consist of various heterogeneous machines situated at the Forecast Systems Laboratory (FSL), Pacific Marine Environmental Laboratory (PMEL), and Geophysical Fluid Dynamics Laboratory (GFDL). Once securely signed-on to a node on the grid, users will be able to submit a job to a grid queue with the expectation that it will run on the first available set of nodes; regardless of geographical location. This capability will be implemented with the widely utilized Globus middleware and associated tools. A prototype NOAA Certificate Authority (CA) will be created to provide a means to authenticate grid users. As proof of concept, versions of the Regional Ocean Modeling System (ROMS), Weather Research and Forecast (WRF) and the new GFDL global coupled climate model will be run on all machines on the grid. The WRF/ROMS coupled system developed using FY ’03 funding will be tested on the prototype grid. Assuming accessibility, these models will also be tested on the NSF sponsored TeraGrid with the hope that lessons learned can be applied to the NOAA grid.

Problem Statement:

For years, efforts to improve numerical modeling in NOAA have included increasing spatial and temporal resolutions, extending spatial domains, running longer simulations, running ensembles, and simulating coupled processes (such as ocean and atmosphere). Sensitivity studies, with variable forcing and alternate parameterizations of unresolved physics, also entail a large number of simulations. These efforts are limited by the availability of adequate computing resources (for example, adequate to produce forecasts in a reasonable time). Unfortunately, computing resources are not used as efficiently as possible due to the difficulty of shifting workload between machines, different CPU architectures, different operating systems, and many other technical problems.

The fundamental cause of this inefficiency is that resources are made available in a non-uniform, lab-by-lab fashion. For example, some NOAA labs such as PMEL do not have supercomputers but need them to accomplish research objectives. Lacking access to uniformly managed NOAA resources, they are forced to make individual arrangements with other labs and learn multiple site-specific procedures. Second, there are times, such as immediately following a hardware refresh, when an individual NOAA supercomputer site has a surplus of compute cycles. These cycles are currently unavailable to other labs. Third, thousands of individual nodes used during regular business hours by NOAA personnel lay idle at other times. Small-scale compute jobs such as model pre/post-processing and single members of large ensembles could be off-loaded to the more capable of these machines. Fourth, model runs requiring a significant number of nodes sometimes face long delays waiting in queues at individual NOAA supercomputer sites. For loosely coupled models, a means to run one model component at one site and another at a second

site would enable the runs to execute sooner because fewer nodes are required at any single site. Additionally, this capability would allow model experiments that require more resources than are available at a single site to be performed using the aggregate resources available on the grid.

In order to harness these resources more effectively several key issues need to be addressed.

1. The compute cycles need to be made available in a seamless fashion so that users do not have to concern themselves with site-specific procedures for submitting jobs.
2. It must be possible to smoothly exchange model input and output files. This is crucial if a user at one laboratory is to run a model on remotely available machines.
3. A secure means to authenticate user access to remote sites is required. This authentication needs to be implemented in the same fashion NOAA-wide so as to eliminate the burden thrust upon users of learning site-specific security procedures.

A NOAA-wide Grid would address these issues. A grid is defined to be cyber-infrastructure that enables flexible, secure, coordinated sharing of computational and data resources. This NOAA grid would be akin to the NSF funded TeraGrid, the NASA Information power grid and others that have been developed in recent years. The grid would provide NOAA scientists with a means to seamlessly and securely access NOAA computational resources that are currently under-utilized.

Proposed Solution:

I. Develop a Prototype Grid. Utilizing funding from the FY '03 proposal, "Using the TeraGrid for NOAA Scientific Computing", a nascent NOAA grid has been constructed using Globus middleware. The grid consists of one Intel Linux node located at PMEL and several others situated at FSL. Access to this "grid" is provided using the relatively insecure Globus CA. The primary objective of the FY '04 proposal is to extend this work to develop a prototype NOAA grid as described below:

A. Components

The grid will consist of compute and storage resources made available by FSL, PMEL and GFDL. FSL will provide a subset of its Intel Linux cluster, PMEL will provide several individual Intel Linux nodes and a four-processor shared memory DEC Alpha machine, and GFDL will also provide time on a 32-processor AMD Linux cluster.

B. Functionality

Any user with access to the grid will have the ability to submit jobs to any nodes on the grid. The user will be able to specify simple job constraints such as the number of nodes, the processor type, and memory required. These jobs will go into a grid queue and run on the first available set of nodes meeting the constraints. There will be an ability for model initial condition, boundary condition, and output files to be seamlessly transferred between nodes on the grid. This will enable jobs to be run remotely. For example, a user at PMEL could request that job

executable and data files produced locally be automatically transferred to the location where the job will run and that the resulting output files be transferred back to PMEL.

The principal investigators will determine how cycles donated by the labs to the prototype NOAA-grid are allocated to grid users. For a long-term operational NOAA-grid, it is envisioned that compute cycles will be allocated to users based on NOAA-wide priorities.

As proof-of-concept, the WRF, ROMS, coupled WRF/ROMS and the GFDL coupled climate models will be grid-enabled. This simply means it will be possible to run them on any nodes on the grid. Actual runs of these models on various nodes will be executed for demonstration purposes.

C. Technologies

The grid will utilize network backbones as available. Currently there is an Abilene connection between PMEL/FSL and a T3 connection between GFDL and the two other labs. If the National Lambda Rail backbone connections are made between two or three of the labs then they will be utilized. The software required to implement job scheduling, file transfers and secure access will be implemented using Globus middleware.

D. Security

Under this proposal, a prototype NOAA CA will be setup. This authority will be responsible for providing users secure access to the NOAA grid. DOC and NOAA security experts will be consulted in developing this authority. However, one can envision a system where users are required to provide an NOAA identification badge (or copy) as proof of identity, are given passwords only by phone calls or letters originated by the NOAA CA, and can only login with the aid of a NOAA CA issued secure keypad that employs a challenge-response system.

II. TeraGrid Access. If time/funding permit and access to the TeraGrid can be obtained then another objective will be to port the above-described models to machines on the TeraGrid. Doing so will make it easier to learn how the Teragrid works and provide access to TeraGrid personnel expertise. The resulting lessons learned could then be applied to improving the NOAA grid.

III. Coupled Modeling. Under FY ’03 funding, a crude version of a coupled WRF/ROMS model has been constructed where ROMS executes on the PMEL node of the nascent grid and WRF runs on the FSL nodes. The coupling communication is implemented using MPICH-G2, the grid-enabled version of the MPICH Message Passing Interface (MPI) library. Preliminary tests show there exist coupled model scenarios that can feasibly be executed under this cross-machine setup. Another objective for the FY ‘04 proposal will be to execute the coupled model on the prototype grid. In this case, the two models will run on medium-sized clusters on the grid (at FSL and GFDL, for example). A scheduling mechanism will be developed that starts the job only when the required nodes are available on both machines. The grid network will be tuned to optimize coupling communication.

Analysis:

The primary benefit of developing a prototype grid is that it will pave the way for seamless access to computational resources by NOAA scientists. This will make it possible to utilize potentially unused compute cycles at NOAA. The resulting increased efficiency will enable more and better scientific experiments to be conducted. Also, in the process of creating the grid, tools for efficient, secure, reliable transfer of data needed by the models that run on it will also be developed. These tools will be useful for other NOAA scientists who need improved access to observational and model forecast datasets scattered throughout the agency. In addition, the need to run models on computational domains of ever-increasing size poses problems for post-processing analysis and visualization. The NOAA grid would provide advanced storage resource utilities, such as the ability to use multiple data channels for parallel data transfer to allow these large files to be transferred to a local machine for analysis. Finally, the easy access to heterogeneous platforms provided by the grid will encourage the development of model codes that run correctly and efficiently on multiple machines.

This proposal leverages several significant efforts:

1. The FY ’03 funding produced a small grid that runs with bare minimum security on a few homogeneous nodes at FSL and PMEL. As mentioned, a version of the WRF/ROMS coupled system runs on this “grid”. The next logical step is to extend this work to heterogeneous nodes including medium sized-clusters running at three major NOAA labs in a highly secure fashion.
2. Modeling Environment for Atmospheric Discovery (MEAD) is an NSF funded multi-institutional effort to develop cyberinfrastructure running on the TeraGrid that enables simulation of hurricanes and storms. Investigator Moore has been funded by MEAD for the past year to help construct the coupled WRF/ROMS model. It is expected that Dr. Moore will receive continued funding in FY ’04.
3. The Department of Defense Programming Environment and Training (PET) project has funded a grant to develop software infrastructure that implements parallel data re-gridding required for coupled modeling. Using funding from this grant, Investigator Schaffer has developed a simple Applications Programming Interface (API) that facilitates the coupling of WRF and ROMS described earlier. There are tentative plans to continue this work in FY ’04 with the likelihood that Mr. Schaffer would receive additional funding.
4. For several years, ocean modelers at PMEL have been using FSL supercomputers to conduct simulations. Slow data links between PMEL and FSL have hampered this effort. Consequently, FSL has been using base funds to tune the network connections between the two labs. This effort overlaps nicely with the work required to optimize coupling communication between the WRF and ROMS model. This is one reason why FSL is willing to commit matching base funds to this project.
5. GFDL is currently using lab funds to develop its new global coupled climate model and port it to multiple platforms. This proposal will be able to take advantage of the porting effort.

In defining the direction for this proposal, a couple of choices were made. One was to focus research into providing improved NOAA-wide access to distributed resources instead of a single site. There are two reasons why we expect NOAA to continue to rely on distributed resources for the foreseeable future. Threats posed by terrorism, fire, power outages, etc. require that NOAA maintain backup supercomputing sites so that operational integrations run without failure. Also, many laboratory missions require high-speed access to large local data sets. The proposed grid is needed for other missions that can use distributed resources when data sets are small or access time is not critical. A second choice made was to use Globus instead of alternatives such as the European UNICORE package because of the widespread use of Globus in the United States in projects such as MEAD.

There are several ways in which this proposal fits within the NOAA HPCC program objectives. One is that a NOAA grid would facilitate exploitation of high-end NGI technologies such as Abilene, National Lambda Rail and the TeraGrid by improving access to these high-speed resources. A second is that the proposal calls for use of significant off-the-shelf products such as Globus and MPICH-G2. Moreover, these products and the TeraGrid itself are all emerging grid technologies. Finally, it will enhance collaboration between NOAA research labs on computational strategies.

Performance Measures:

This project will be successful if a prototype NOAA grid can be constructed. In particular:

1. Users can run jobs on any nodes on the grid.
2. A prototype NOAA Certificate Authority is in place.
3. The WRF, ROMS, and GFDL global coupled climate models can all run on the grid machines.

Milestones

Month 01 – Models grid enabled
Month 02 – Preliminary CA in place
Month 05 – Test job submitted to the grid can run on any nodes
Month 06 – All models demonstrated to execute on all grid compute platforms
Month 09 – Final CA in place
Month 12 – Final report submitted

Deliverables

- Working prototype grid including a NOAA Certificate Authority
- Final report covering grid performance/reliability, a discussion of how NOAA could expand the prototype into a complete NOAA grid, and a study of the feasibility of coupling WRF/ROMS across the grid