

TTT/RR##

**Using the TeraGrid for NOAA Scientific Computing
FY 2003 NOAA HPCC Funded Project**

05/25/04

Principal Investigator: Mark Govett

Phone: 303-497-6278

Email Address: Mark.W.Govett@noaa.gov

Other Investigator 1 (Dan Schaffer, Daniel.S.Schaffer@noaa.gov)

Other Investigator 2 (Al Hermann, Albert.J.Hermann@noaa.gov)

Other Investigator 3 (Chris Moore, Chrisopher.Moore@noaa.gov)

Proposal Theme: NGI

Funding Received – FY 2003 \$80,600

Performance

As one deliverable, we constructed a simple coupled model prototype. The code is available at http://www-ad.fsl.noaa.gov/ac/schaffer/coupled_prototype.tar.gz. The intention was to use the prototype to analyze cross-grid communication performance. However, subsequent to prototype development, it became clear that, to our pleasant surprise, a coupled Weather Research and Forecasting (WRF)/Regional Ocean Modeling System (ROMS) would be available for testing on the grid. A version of the actual WRF/ROMS coupled model is available at http://www-ad.fsl.noaa.gov/ac/schaffer/wrf_roms.tar.gz

The second deliverable was to release a version of the Scalable Modeling System parallelization tool that contains support for grid computing. This version (2.8.0) is available at <http://www-ad.fsl.noaa.gov/ac/sms.html>.

The third deliverable is this report.

Co-PI Schaffer can be contacted for more information about these deliverables.

Project Summary

As the project title indicates, the original objective was to investigate exploitation of the TeraGrid for NOAA scientific computing. Unfortunately, the TeraGrid did not become available in a timely fashion as was hoped. Access to the PIs was only granted in February of 2004 and then the TeraGrid was compromised by hackers, resulting in an additional 5 weeks of down-time. Despite this setback, we were able to make significant in-roads in the understanding and exploitation of grid technology.

As promised, we investigated various software packages that could be used to construct a grid. Although middleware such as Unicore and Legion offer many of the capabilities needed, none have the level of research and business community support enjoyed by the Globus Toolkit (www.globus.org). Globus is used extensively in U.S. grid efforts including the TeraGrid. It is augmented by strong software development efforts that build on the tools it provides.

Having chosen Globus, we applied it to the construction of a rudimentary grid consisting of 10 machines located at FSL and two machines at PMEL. Globus client commands provide the means to run simple jobs and transfer files remotely. Low-level security is implemented using the SimpleCA certificate authority. To access any of the grid nodes, users are required to enter in pass phrases for authentication. A grid-enabled version of the MPICH message passing interface (MPI) library called MPICH-G2 is included in the grid. This library supports cross-grid communication as discussed below.

As mentioned above, the effort to construct a prototype coupled model was superceded by the unexpected availability of a coupled WRF/ROMS model. The development of coupled WRF/ROMS was facilitated by the NSF funded Model Earth and Atmosphere

Development (MEAD) project for which investigator Moore receives funding. It was also helped by a DoD sponsored effort to construct software infrastructure for coupling models using the Argonne National Laboratory developed Model Coupling Toolkit (MCT). Investigator Schaffer received funding for this work. Finally, the division of processors between the WRF and ROMS models was implemented using the National Energy Research Scientific Computing Center (NERSC) developed Multi-Process Handshaking (MPH) library. Thus, this grid project was able to leverage off other current research efforts quite nicely.

Since the TeraGrid was unavailable, we instead implemented coupled WRF/ROMS on the rudimentary NOAA grid. WRF runs on a cluster of three Intel Linux nodes located at FSL. The number of grid points is 161x161x30. ROMS runs on two Intel Linux nodes located at PMEL. The number of grid points is 200x200x15. The clusters are connected over the Abilene network with a theoretical bandwidth of 100 megabits/second (mbits/s). Periodically the models exchange surface boundary conditions across the grid. WRF sends U and V wind stress to ROMS and ROMS sends Sea Surface Temperature (SST) to WRF. Local and cross-grid communication is implemented using MPICH-G2. Table 1 below shows the results.

| Case | WRF Comm | ROMS Comm |
|---|-------------|--------------|
| Cross-grid comm (Serial, ROMS untuned) | 1.53 | 1.15 |
| Cross-grid comm (Serial, ROMS tuned) | 1.58 | 0.34 |
| Cross-grid comm (Parallel, ROMS untuned) | 0.83 | 0.64 |
| Cross-grid comm (Parallel, ROMS tuned) | 0.78 | 0.51 |
| Local comm (Parallel) | 0.08 | 0.18 |

Table 1. Communication times for the exchange of surface boundary conditions for the WRF and ROMS models. Each model ran for 25 time steps. Boundary conditions were exchanged every 5 time steps. Measurements were taken starting on time step 6. Thus 4 boundary condition exchanges were measured. The measured main model loops times for time steps 6-25 were 125 seconds for WRF and 52 seconds for ROMS.

The bottom line is that the fastest communication times shown in the table are considerably smaller than the model integration times. As an exercise, suppose the WRF model was run on 64 processors, the ROMS model on 8 processors and both scaled perfectly as compared to run-times for the current configuration of 6 and 2 processors. Then the main model loop times for both would be about 12 seconds; still considerably longer than the communication times. Moreover, there are many scenarios under which the coupling can occur much less frequently (i.e. once per model day). Finally, there is a known performance bug in MCT. Specifically, it unnecessarily communicates data as 8 byte floating point numbers even when the model data are stored as 4 byte real numbers

as is the case here. So, overall, it appears quite feasible to execute the coupled model across the grid.

In each of the “Serial” cases shown, the data to be sent on each model are gathered to the corresponding root processor and then exchanged with the other model’s root. Interpolation of the data to the target model’s grid follows immediately. Finally the interpolated data are scattered back to the other processors. Although not shown, the gather and scatter times are negligible.

The table shows normal and “tuned” serial interpolations. In the tuned case, the operating system kernel of the sending node for WRF was modified to eliminate an effect called Transmission Control Protocol (TCP) “slow start”. When data are communicated across the network infrequently as in this coupled modeling example, a timer expires prior to each exchange. As a result, the next set of data is sent out very slowly at first, ramping up exponentially. The kernel modification increased the time before “slow start” begins again so that it was greater than the time between boundary condition exchanges. This eliminated “slow start”, resulting in decreased communication times as shown by the smaller values for “ROMS communication” in the “tuned” case.

Since mitigation of TCP “slow start” in this way violates TCP protocol, the practical applications of this result are not clear. Moreover, simply scaling up the model to a larger number of processors may negate the benefit of this optimization since the data exchanges might naturally occur more frequently than the time that elapses before “slow start” recommences.

Consider the ROMS communication time of 0.34 seconds for the serial, tuned, cross-grid case. The communication latency was separately measured to be 15 milliseconds each direction. From this and the known data sizes, we infer that the bandwidth is 60 mbits/s. This compares to observed simple file transfer rates for large files of 80 mbits/s. The boundary condition exchange bandwidth is smaller simply because, even after slow start is eliminated, the data transfer speeds ramp up linearly. The coupled model would have to be run much longer to attain the higher bandwidths.

In the parallel boundary condition exchange cases, all nodes participate in the communication. In particular, when ROMS receives data from WRF, each node simultaneously receives half as much data as in the serial case. Thus it is not surprising that the communication occurs twice as fast for the un-tuned case. The result for the tuned case is mysterious. It indicates higher speeds are attained using serial instead of parallel communication. This result is under investigation.

Theoretically, all of these bandwidth issues become moot on big communication pipe networks such as the TeraGrid. At speeds of 10 gigabits/s the only impediment will be latency. This is a function of the time it takes for light to travel between the grid machines (10 milliseconds between the PMEL and FSL clusters) and router hops (5 milliseconds). Latency would be expected to prevent integration of a single tightly

coupled model (i.e., an atmospheric model) across the TeraGrid but should pose no problem for loosely coupled models such as discussed here.

The most significant lesson learned in the course of this research is that there exists a fundamental tension between the needs of security and the grid. As one example, security concerns dictate that the number of open connections between machines be minimized so as to facilitate tracking and prevention of illicit activity. The grid needs increased inter-connectivity. In the coupled model case, connections between every pair of compute nodes are required to implement parallel communication. Second, in an ideal security world, users are required to maintain separate passwords for every accessible machine. In the grid model, a single grid password gives a user access to every machine on the grid. For anyone attempting to deploy grid technology it is critical that researchers and high-level management understand this conflict. A successful NOAA computational grid will likely have to live with security constraints. For example, a coupled model application may have to communicate across the grid serially instead of in parallel so as to reduce the number of open ports. In addition, the data transferred may be to be encrypted and unencrypted as NOAA increases its utilization of Virtual Private Network (VPN) technology.

In terms of possible technology transfer, the rudimentary grid constructed here could easily be expanded to include more clusters at additional NOAA laboratories as discussed in Future Directions below. Also, the coupled modeling demonstrated here could be implemented for other loosely coupled modeling research efforts in NOAA. Coupled modeling over the grid would be most useful when large initial and boundary condition files are physically located where each of the coupled model components is executed. In cases such as a coupled oceanic/atmospheric model, the amount of data that must be communicated across the grid consists only of surface fields, a relatively small amount. Thus, it may be more efficient to couple across the grid than to relocate one set of initial and boundary condition files to another HPCS site and execute the entire coupled model there.

Expenditure Summary

| Personnel | Participation (months) | Salary (per hour w/ Overhead) | | | \$ Amount | | Total Cost | Base Salary/year | Percent of year |
|---|---------------------------|----------------------------------|---------|-------------|-------------|-------------|-------------|---------------------|--------------------|
| | | Hours | Loaded | Salary | IRA Total | to IRA | | | |
| FY 2003/2004 | | | | | | | | | |
| Dan Schaffer - OIRA (04) | 1.19 | 207.00 | \$63.44 | \$13,132.08 | \$8,363.12 | \$8,666.45 | \$21,495.20 | 87100 | 0.10 |
| Dan Schaffer - OIRA (03) | 2.18 | 378.05 | \$63.44 | \$23,983.78 | \$15,276.36 | \$15,830.43 | \$39,260.14 | 87100 | 0.18 |
| Chris Moore - JISAO (03-04) | 0.75 | | | | | | \$6,335.00 | | |
| Al Hermann - JISAO (03) | 0.15 | | | | | | \$945.00 | | |
| PMEL CNSD Services | | | | | | | \$2,720.00 | | |
| Mark Govett - G | 0.26 | 40.00 | \$62.15 | \$2,486.00 | \$1,753.37 | \$1,816.97 | \$4,239.37 | 78379 | 0.02 |
| Subtotal | 4.53 | | | \$39,601.86 | \$25,392.86 | \$26,313.84 | \$74,994.72 | \$252,579.00 | |
| Non-Labor - Includes travel, software licenses and supplies | | | | | | | \$5,000.00 | | |
| Total | | | | | | | \$79,994.72 | | |

Figure 1. Expenditure report for TeraGrid project.

Future Direction

Future plans are discussed at length in our FY 2004 proposal entitled, "Development of a Prototype NOAA Grid". In summary, the proposal calls for an extension of the existing grid to include more compute clusters at PMEL, FSL and at GFDL. In the process, a meta-scheduler will be developed that will facilitate the current push to treat the various NOAA HPCS super-computers as one common resource. As the proposal describes, FSL will continue to support research and development of the NOAA grid. In addition to money directed to the FY '04 work, it is likely FSL will contribute to the costs of improving interoperability of weather and ocean models. The porting of models to varying architectures needed for interoperability is a minimum requirement to making them "grid-enabled". Thus we envision that, ultimately, any of the significant NOAA models could be submitted to the grid meta-scheduler and execute on any of the available platforms.

The FY '04 plans also call for further investigation of loosely coupled modeling over the grid. The WRF/ROMS coupled model will be executed across medium sized clusters at FSL and GFDL. The coupled model will also be tested on TeraGrid resources now that they are available.

Publications

A paper entitled, "Coupling an Oceanic/Atmospheric Model over a Geographically Distributed Computational Grid" is in progress.