

The LEAD-WxChallenge Pilot Project: The Potential of Grid-Enabled Learning

Richard D. Clark¹, Suresh Marru², Marcus Christie³, Thomas Baltzer⁴, Kelvin Droegemeier⁵,
Everette Joseph⁶, and Brad Illston⁷

¹Millersville University of Pennsylvania, Millersville, PA

²University of Indiana, Bloomington, IN

³Walker Information Inc., Indianapolis, IN

⁴Unidata Program Center, Boulder, CO

⁵University of Oklahoma, Norman, OK

⁶Howard University, Washington, DC

⁷WxChallenge, University of Oklahoma, Norman, OK

Abstract

This paper discusses the need and potential benefit that could be realized with access to a stable, reliable, and persistent allocation of distributed TeraGrid infrastructure and software technologies for the atmospheric science education community. During the past few decades, the academic meteorology enterprise has supported a national collegiate forecast contest that seeks to engage mostly undergraduate students with some graduate students and faculty in practical forecasting under a variety of geographical and phenomenological circumstances. Known today as The Weather Challenge (WxChallenge), each participant forecasts the maximum and minimum temperature, precipitation amount, and maximum sustained wind speed for select North American cities. WxChallenge provides students an opportunity to compete against their peers and faculty mentors at other institutions (64 nationwide in 2006-2007) for honors as the top weather forecaster in the nation. In spring 2007 the NSF-funded ITR project, Linked Environments for Atmospheric Discovery (LEAD) engaged a manageable subset of the WxChallenge community and provided access to the LEAD Gateway portal and its underlying services. The goal of the so-called "LEAD-WxChallenge Pilot Project" was to begin ushering in a fundamental paradigm shift in how experiments are conceived and performed, in the structure of user application tools and middleware, and in methodologies used to observe and model the atmosphere. Students were provided with the ability to generate, run, analyze, and visualize their own WRF forecasts using the web-enabled service oriented architecture developed by LEAD IT researchers. For seven weeks, the 75 students and faculty from 10 institutions were given an unprecedented opportunity for enhancing their understanding of numerical modeling, high performance computing, physical parameterization schemes, data assimilation, and workflow orchestration. Participants were given sufficient compute cycles on TeraGrid servers to build their own experiments and run the WRF numerical modeling system. The project participants launched a total of 279 forecast workflows and generated 0.6 TB of data. Over 160 processors were reserved five days each week from 10am to 8pm EDT. For the NAM initialized WRF forecast, 78 percent of the workflows submitted were successful, 22 percent failed. The ADAS initialed WRF forecast was less successful with 36 percent completing and 64 percent failing. Moreover, significant intervention was necessary on the part of developers to achieve these moderate success rates. Since its inception LEAD has espoused the motivation to empower a community of users, and to provide the necessary grid-enabled infrastructure and tools to facilitate research and education. In return, LEAD developers benefited from user feedback that exposed strengths and weaknesses, and provided a better sense of the challenges and resource requirements associated with maintaining a reliable and persistent system. While this pilot project revealed an overall satisfaction with LEAD capability, the fragility of the infrastructure was a serious limitation and one that has the potential to thwart the efforts of LEAD developers and the TeraGrid leadership even with the best intentions, and may prevent a larger user community from ever realizing the benefits of grid computing.

Index Terms

WxChallenge (The Weather Challenge); WRF, Weather Research and Forecasting model; TeraGrid; ADaM, fault tolerance; HPC; service-oriented architecture; ARPS (Advanced Regional Prediction System); ADAS (ARPS Data Assimilation System); ITR (Information Technology Research); cyberinfrastructure; myLEAD; gridftp; LEAD

1. INTRODUCTION

For the past few decades, the academic meteorology enterprise has supported a national collegiate forecast contest that seeks to engage mostly undergraduate

students with some graduate students and faculty in practical forecasting under a variety of geographical and phenomenological circumstances. Known today as The Weather Challenge (WxChallenge) and sponsored by the University of Oklahoma, each participant forecasts the maximum and minimum temperature, precipitation amount, and maximum sustained wind speed for select North American

Richard D. Clark, Department of Earth Sciences, Millersville University, P.O. Box 1002, Millersville, PA 17551-0302, Richard.Clark@millersville.edu.

cities. WxChallenge provides students an opportunity to compete against their peers and faculty mentors at other institutions (64 nationwide in 2006-2007) for honors as the top weather forecaster in the nation.

2. THE PILOT PROJECT

In spring 2007 the National Science Foundation (NSF)-funded Information Technology Research (ITR) project, Linked Environments for Atmospheric Discovery (LEAD) engaged a manageable subset of the WxChallenge community and provided access to the LEAD Gateway portal and its underlying services. Since its inception LEAD has espoused the motivation to empower a community of users, and to provide the necessary infrastructure and tools to facilitate research and education (see <http://portal.leadproject.org>).

The goal of the so-called "LEAD-WxChallenge Pilot Project" was to begin ushering in a fundamental paradigm shift in how experiments are conceived and performed, in the structure of user application tools and middleware, and in methodologies used to observe and model the atmosphere.

Students were provided with the ability to create their own experiments, which involved running their own instance of the Weather Research and Forecast numerical modeling system (WRF) across distributed resources using the web-enabled service oriented architecture developed by LEAD researchers. Students used the model output for analysis and visualization of atmospheric field variables, which they incorporated into their forecast decision making schema. For seven weeks, the 75 students and faculty from 10 institutions were given an unprecedented opportunity for enhancing their understanding of numerical modeling, high performance computing (HPC), physical parameterization schemes, data assimilation, and workflow orchestration. It also demonstrated one of LEAD's primary goals; to democratize and empower students by lowering the barrier for access to complex, integrated services, thereby allowing users the freedom to select inputs such as initialization fields, set model domains, and run WRF at a time and location determined by the user, and not constrained as in static, fixed and prescribed operational environments. In return, LEAD developers benefited from user feedback that exposed strengths and weaknesses, and provided a better sense of the challenges and resource requirements associated with maintaining a reliable and persistent system aimed at enabling a larger community.

3. TECHNICAL NOTES

Participants in the pilot project were given authorization to the LEAD Gateway portal to access

data, build experiments, and compose workflows with sufficient computing resources to run WRF and save the WRF output in myLEAD workspace. LEAD also provided tools for visualizing the output, as well as user support. Students were able to configure and run a 42-hour forecast using the WRF system at a horizontal resolution of 5 km. Students integrated output products generated by the WRF run into their personal decision-making schema for preparing a forecast for stations previously selected by WxChallenge.

Over the seven week pilot project participants launched a total of 279 forecast workflows and generated 0.6 TB of data. Over 160 processors were reserved on a multi-processor server five days each week from 10am to 8pm EDT. For the NAM (North American Model) initialized WRF forecast, 78 percent of the workflows submitted were successful, 22 percent failed. The ADAS (ARPS (Advanced Regional Prediction Systems) Data Assimilation System) initialed WRF forecasts were less successful with 36 percent completing and 64 percent failing. Success is defined here in a technical sense. A workflow was deemed successful if the workflow ran to completion without regard to: 1) whether gridftp actually transferred the model output file into the user's myLEAD workspace for it to be utilized by the forecaster, or 2) whether the job completed in time for the user to incorporate it into the forecast decision-making process. If we include only those workflows that ran to completion and were available to the user in time for incorporation into the forecast decision schema, the number of successes is somewhat dismal. What the pilot project revealed most was that the cyberinfrastructure underlying the LEAD Gateway was fragile and had to be manually monitored in order to achieve some modicum of reliability and persistence.

When the WxChallenge ended and the pilot project concluded, the participants were asked to complete an informal survey, the results of which were used to refine the portal and functionality and prepare for an expanded release to all institutions participating in the spring 2008 WxChallenge forecast competition.

Results of the survey were largely positive with participants finding the portal "easy and intuitive," and the workflow and experiment builder a "great concept" and "highly intuitive." Particularly noteworthy were comments on the workflow monitor, a feature that allows the user to follow the progress and status of the experiment. One participant reported that [the workflow monitor] was a "very powerful feature that allows the user to interact with the forecast he/she is creating." In summary, the software and

middleware that controlled LEAD applications worked favorably.

On the other hand, users reported very little satisfaction with when it came to questions related to workflow completion. When asked if LEAD was a value-add to their decision-making process, the typical response was, "No, it did not [because] data results took too long to come in...", and "Since the data was rarely available when forecasters were making their forecasts during the first period, they stopped looking for it...", and "the run had almost never completed by the time I had to submit my forecast." Moreover, the number of person-hours needed to maintain the IT infrastructure during the WxChallenge was significant. File transfer across the grid (gridftp) was a primary culprit in the high failure rate when considering the full end-to-end workflow from launch to file access.

LEAD involvement in the spring 2008 WxChallenge never materialized because the problems that were exposed in 2007 remain some of the same challenges that face LEAD developers today.

Yet the LEAD-WxChallenge pilot project can hardly be considered a failure. Developers learned a great deal about the frailties of some middleware components of the infrastructure, more often than not components that were beyond the developers' ability to control. But it also gave us confidence in the LEAD portal and its underlying functionality as a viable Science Gateway and soon led to the development of useful tools such as an enhanced fault tolerance mechanism, monitoring, and reporting system and integration of data mining into the architecture.

4. LEAD IN A GRID CONTEXT

LEAD is a bold and revolutionary paradigm that through a Web-based Service Oriented Architecture exposes the user to a rich environment of data, models, data mining and visualization/analysis tools that enable the user to ask science questions to applications while the complexity of the software and middleware managing these applications is hidden from the user. LEAD espouses two primary goals that have context for the future of grid computing environments: 1) LEAD espouses to lowering the barrier for using complex end-to-end weather technologies by a) democratizing the availability of advanced weather technologies, b) empowering application in a grid environment, and c) facilitating rapid understanding; and 2) developing the infrastructure that make possible the dynamic adaptation of models and observing systems through on-demand, fault tolerant services.

LEAD, as it exists today, is poised to enable a diverse community of scientists, educators, students, and operational practitioners. The project has been

informed by atmospheric scientists who, in search of new knowledge, understanding, and ideas, require fundamental capabilities that allow for query and acquisition, simulation, assimilation, data mining, computational modeling, and visualization. LEAD project computer scientists have created tools to empower the users (see Fig. 1).

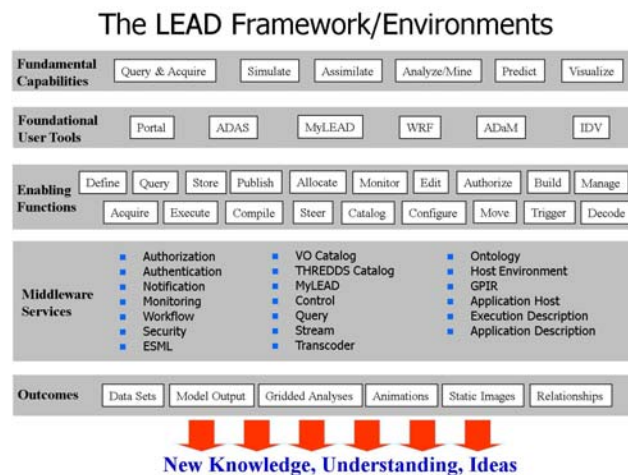


Fig. 1. The LEAD Framework: Capabilities defined by atmospheric scientists to achieve outcomes that will yield new knowledge, understanding, and ideas, made possible by enabling functions and middleware services developed by IT scientists.

Entering the science gateway through the LEAD portal, the authenticated user finds services for assimilating data (ADAS), storing and retrieving files (myLEAD), data mining (ADaM) [1], numerical modeling (WRF), visualization (IDV) [2], plus educational modules with guided inquiry (LEAD-to-Learn modules). Many of the tools and services are available without requiring user registration and authentication.

Managing these applications is a suite of middleware services that fade into the background, except for when an interested user wants to monitor the workflow by invoking an XBaya workflow composer. Fault tolerance mechanisms that exploit batch-queue prediction, redundant computation, and opportunistic scheduling ensure that the user is kept abreast of the progress of the workflow, and increase the probability of successful execution [3]. Upon completion, the user can visit a personal workspace to access the outcome of the workflow execution. The myLEAD workspace included the metadata for access to data sets, model output, gridded analysis, animations, and static images.

Consistent with the notion of a heterogeneous, distributed cyberinfrastructure, the user doesn't have to know, or care, where the compute cycles came from, nor worry about programming the middleware to bridge the functional capabilities with the expected

outcomes. Services are exposed while their complexity is hidden, allowing even non-computer scientists the ability to link services together to create user-tailored workflows that address specific research questions.

LEAD has become well known in TeraGrid circles as an exemplary use case that tests and extends the potential of grid computing. But the LEAD goal of lowering barriers to complex, high-end computing technology through democratization, empowerment, and facilitation will never be realized unless TeraGrid adopts a vision and sets into practice a cyberinfrastructure that will ensure reliability and persistence. The periods when all of the TeraGrid resources are “up” are rare, based on the WxChallenge experience.

The need for a robust and reliable IT infrastructure to support knowledge transfer, transparency of use, integration, interoperability, and other services has not fallen on deaf ears. Position papers at this conference call for less emphasis on “big iron” and FLOPS [4] – significant amounts of compute cycles are not being allocated or used [5]. Instead, in the spirit of NSF’s IT goals, greater support for “discovery, learning, research infrastructure, and stewardship” is critical if TeraGrid hopes to extend its tentacles to the widest possible user base (<http://www.nsf.gov/about/glance.jsp>).

The emphasis must be on a reliable, persistent, engaging, easy to use, secure, and intelligent grid infrastructure, with a suite of tools for file transfer, cataloging, and managing, interactive user interfaces, and policy rule engines [6], if the paradigm of “built it and they will come” is to be realized.

This significant national investment in grid resources cannot be limited to Track I and Track II efforts with their focus on hardware. We argue that in order to stimulate the next generation of scientists interested in doing STEM research across grid environments, the TeraGrid must include dedicated resources not only for the scientists high on the HPC chain, but for the bottom dwellers; educators and students who realize the significance of a national infrastructure and want to employ it to empower the next generation of scientists and practitioners. With campus computer clusters democratization is stretched to a wider audience of users, but unless these clusters are integrated with the TeraGrid and available to a extra-campus user base, it is likely that the chasm will still exist between large universities that can sustain infrastructure support, and huge numbers of potential users at the bottom of the pyramid that will remain disconnected from what should be a national infrastructure. *There is no rationale in today’s technology environment that can justify segregating a national resource.* With proper

authentication that ensures compliance to standards, security, and policy, some allocation of compute cycles and infrastructure should be accessible to the largest of possible user bases, in the same way that anyone who conforms to certain standards and protocols can gain access to the electric power grid.

Toward this end, making LEAD capabilities available to the WxChallenge community, with over 60 institutions - large and small, and a thousand students from across the U.S., could serve as the quintessential use case that helps steer the evolution of TeraGrid by engaging a larger user constituency and directing more of its efforts to the development and management of a robust cyberinfrastructure.

For a typical semester, LEAD developers anticipate that WxChallenge participants will have the need for 150 concurrent workflow runs typically for 20 to 180 minutes each, plus post processing, resulting in 1200 Jobs within span for three hours. For each workflow, 1.6 GB of data is ingested with 1.2 TB/day of output data that must be stored in MyLEAD workspace. Over the full semester-long WxChallenge activity, 110 TB of output will be generated. This allocation would not significantly impact the HPC needs of the research community, but would go far in achieving the NSF goals. It should be clear that this allocation would not be for the WxChallenge management, but to enable individual users at their home institutions to access data, assimilate observations, launch experiments over user-prescribed domains, invoke data mining algorithms, run the WRF model, transfer and archive the output in myLEAD workspace, and access the output for visualization and analysis to better inform the individual forecasters decision making process.

5. BEYOND WxCHALLENGE

WxChallenge represents but one example of how LEAD has engaged the community to widen its user base and democratize the services provided through this science gateway. Two workshops targeting university faculty and graduate students have been important outreach activities. In summer 2006, the LEAD team was a major part of the Unidata Program Center triennial education workshop, “Expanding the Use of Models in the Atmospheric and Related Sciences.” About 50 participants from across the U.S. were initiated to LEAD capabilities, and all launched workflows nearly simultaneously using grid infrastructure and the community allocation of HPC compute cycles on TeraGrid servers. On the positive side, participants were excited about the new LEAD capabilities, with one participant summing it up in a response to a workshop survey, saying, “I have just accomplished in five minutes what it took me the summer of 2005 to do.” The ability of LEAD to lower the barriers

to high-end advanced technology has been a primary goal since the project's inception. However, behind the scenes, workflows were monitored manually and some had to be restarted in order to run to completion. Again, the level of intervention on the part of the developers was not insignificant.

A second workshop was hosted by the LEAD team at the 88th Annual Meeting of the American Meteorological Society in New Orleans, LA in January 2008. LEAD had matured substantially since the previous workshop and now included sophisticated data search and mining engines, and a fault tolerance mechanism that proved its usefulness in a grand way when 20 simultaneous workflows required automatic queue intervention. Responses to a survey were largely positive. Notable was that participants recognized LEAD as a new cutting-edge paradigm for employing computational models, gaining access to observations, utilizing tools that mine and visualize information, and monitoring workflows. The infrastructure required significantly less hand-holding, partly due the fault tolerance mechanism. But the LEAD team could not claim complete success because of sporadic file transfer problems with gridftp. When asked what the perception users have about bringing LEAD to the classroom, they pointed to the following difficulties as they relate to TeraGrid resources: robustness; stability and confidence in successful runs; and inability to handle multiple simultaneous users reliably.

As LEAD moves into its final year as a large ITR, there are plans to extend the user base to graduate researchers through the provision of services that reduce the learning curve for high-end computational modeling and data access, and to graduate and undergraduate education through courses that focus on numerical weather prediction. In addition, a large group of mostly undergraduate students are involved in their local campus weather services. LEAD plans to engage this community, students that number in the thousands, in using LEAD to access and assimilate observational data and generate WRF forecasts as a valued supplement to existing forecast products. These activities will require a reliable and stable grid infrastructure and a community allocation to be successful. TeraGrid can play a key role in helping to sustain LEAD as a viable virtual organization, and in return move closer to achieving its goal to create an integrated, persistent computational resource.

6. CONCLUSION

We support an ongoing community compute cycle allocation and an emphasis directed toward sustaining a reliable and persistent grid infrastructure. Without this, users will visit LEAD once or twice, but

will not return, and the herculean efforts on the part of LEAD developers and the TeraGrid leadership will be lost to a disenchanting community.

The realization of this practical application of grid infrastructure will not only help to introduce the next generation of scientists to the vision of what a national cyberinfrastructure can be for them, but will help to sustain this vital resource because it will have demonstrated the ability to engage and empower an extended user base across a broad academic demographic. Only then can the barriers segregating the "haves" from the "have-nots" begin to dissolve.

ACKNOWLEDGEMENTS

LEAD is a Large Information Technology Research (ITR) Grant funded by the National Science Foundation under the following Cooperative Agreements: ATM-0331594 (University of Oklahoma), ATM-0331591 (Colorado State University), ATM-0331574 (Millersville University), ATM-0331480 (Indiana University), ATM-0331579 (University of Alabama in Huntsville), ATM03-31586 (Howard University), ATM-0331587 (University Corporation for Atmospheric Research), and ATM-0331578 (University of Illinois at Urbana-Champaign, with a sub-contract to the University of North Carolina).

REFERENCES

- [1] John Rushing, Rahul Ramachandran, Udaysankar J. Nair, Sara J. Graves, Ron Welch, Hong Lin, "ADaM: A Data Mining Toolkit for Scientists and Engineers", *Computers and Geosciences*, Volume 31, issue 5, pages 607-618 (June 2005).
- [2] <http://www.unidata.ucar.edu/software/idv/>
- [3] Dennis Gannon, "TeraGrid Futures – Some Technical and Operational Perspectives," position paper, this conference.
- [4] Alan. Blatecky, "TeraGrid Future Whitepaper: An Exit Strategy?," position paper, this conference.
- [5] Gerhard Klimeck, "Cyber-infrastructure for Pervasive Computing: Opportunities and Challenges," position paper, this conference.
- [6] Craig Lee, "Evolutionary Pressures on the TeraGrid and the Larger Grid/Distributed Computing Community," position paper, this conference.