Maximizing Computational Learning for Faculty and Student Scientists: The Ranger Virtual Workshop

S. Mehringer, N. Woody, A. Dolgert, S. Lantz
Cornell University Center for Advanced Computing (CAC)

D. Stanzione
Texas Advanced Computing Center (TACC)

Contact: Susan Mehringer, Cornell CAC at susan@cac.cornell.edu

Originally published in TeraGrid Conference Proceedings: updated/republished May 2011

1. Introduction

The Ranger Virtual Workshop (VW) is a set of online learning modules designed to educate faculty and student scientists on how to effectively use the 579 teraflop Ranger supercomputer at the Texas Advanced Computing Center. Developed by Cornell University, the Virtual Workshop is not meant to replace live workshops, but rather to complement them while increasing nationwide access to cyberinfrastructure education. The Virtual Workshop is available at all times and requires no travel budget. Participants may need just a portion of the material or may simply learn better from written materials that they can use at their own pace. It also serves as a place where attendees at live workshops can go to review material that they did not completely understand during the necessarily short period of the live workshop. The Ranger Virtual Workshop is available to all users via the TeraGrid User Portal. It was funded by the National Science Foundation Office of Cyberinfrastructure as part of the Ranger award.

2. Availability

The Virtual Workshop is available to everyone, but registration is required to use all features. TeraGrid portal users are automatically registered, and can go directly to the materials after login to the portal. Those without a portal account can register at the Virtual Workshop site: https://www.cac.cornell.edu/Ranger/Registration/.

Registration and login is desirable for two reasons: it allows us to identify TeraGrid users, because we are allocating consulting help only to that group, and it allows us to add functionality, such as note taking and progress tracking. It also gives developers more accurate usage data to base decisions on content and features. For example, data on incorrectly answered quiz questions allow identification of content topics that need to be presented more clearly, and page logs aid understanding of how people use the materials, and which topics are most important.

The primary target audience is faculty, staff, and student researchers on the TeraGrid who need information on high-performance computing on Ranger. However, given that the content ranges from introductory to advanced topics, the content is appropriate for a much wider audience, making the VW an excellent vehicle for broadening participation in computational science and engineering.

3. Topics

There are twenty-one Virtual Workshop modules currently available, with new modules being added regularly. The completed modules are intended to be living documents that are updated as needed. Usage data is recorded for published modules so that under-utilized modules can be identified and to ensure that training effort and expertise is focused on the most relevant areas. The goal of all of the modules is to provide the user with a solid footing for using any HPC system, while ensuring that the specific issues associated with Ranger are addressed.

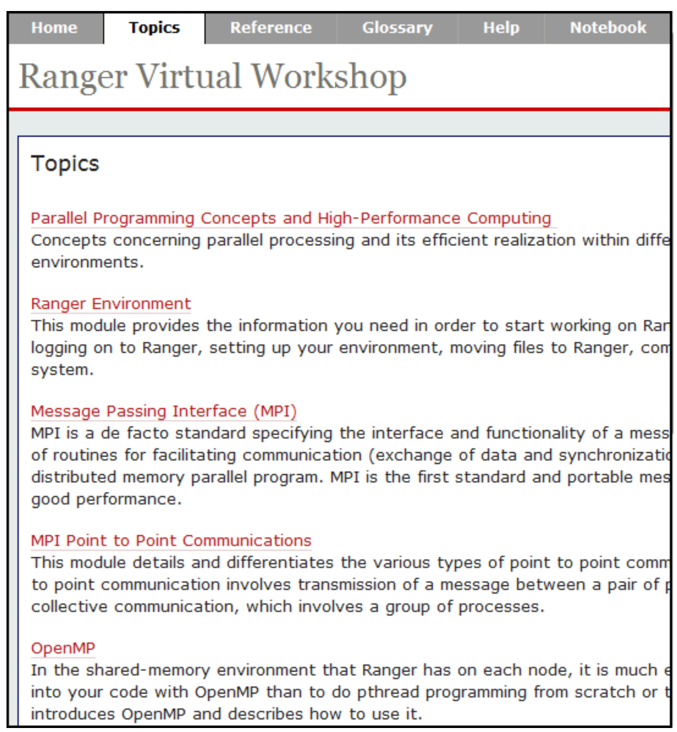

Fig. 1. The topics page provides a short description of each module

The computing environments modules include an "Introduction to Linux" tutorial for the beginning Linux user, intended to get the user acquainted with some of the basic principles of the Linux operating system as well as a "Ranger Environment" module, designed to provide the user with all of the information needed to begin using Ranger. The module is Ranger-specific and walks the user through accessing Ranger, configuring the environment, moving files to Ranger, compiling applications, and using the Sun Grid Engine (SGE) batch system. These topics are covered by easy-to-follow step-by-step examples as well as multimedia clips that walk users through functions such as submitting and monitoring jobs on SGE. All examples were written and tested on Ranger to ensure that the examples apply directly to the system. Source code is provided for several simple applications and scripts that allow users to verify that their account is fully operational.

Modules on programming languages include an "Introduction to C Programming" module for the beginning programmer who has an interest in learning the effective use of the C language; an "Introduction to Fortran Programming" module for the beginning programmer who has an interest in learning the effective use of Fortran; and, a "MATLAB Programming" module covering MATLAB's matrix manipulation, plotting, and general purpose scientific programming capability.

The eight parallel programming modules cover concepts, Message Passing Interface (MPI), and OpenMP. "The Parallel Programming Concepts and High-Performance Computing" module is provided to ensure that all VW users have a common set of terms, concepts and definitions before moving to more advanced modules. This module is code-free and covers general concepts such as hardware terminology (e.g. clusters vs. grids), inter-process communication, and distributed memory systems. The "MPI" module covers the basics of writing and compiling C/C++ and Fortran MPI programs, and focuses on ensuring that the user understands the general workings of MPI. The various implementations of MPI and the functions provided by the MPI standard are explained. This module introduces the user to basic MPI programming and uses a very simple MPI program to demonstrate the key concepts. It focuses on ensuring that the user understands the content, style, and scope of MPI messages and discusses MPI messages and how these messages function in MPI communicators. Specific information about compiling and running MPI on Ranger is also included. The "MPI Point-to-Point Communication" module covers all aspects of this type of communication, including blocking and non-blocking methods, communication modes, and the Wait/Test functions. Discussions on how to pick a communication mode, comparisons of system and synchronization overhead as a function of communication choices, dealing with deadlock, and using persistent communications are provided. The "MPI Collective Communications" module includes a discussion of broadcast, scatter/gather methods, all-to-all, and global computing. This module also focuses on performance issues related to scaling of the collective communication model to petascale-size problems, with an emphasis placed on scaling examples performed on Ranger. MPI-2 additions to collective communication are discussed in this module. One-sided communication provides natural access to Remote Memory Access (RMA) functionality that is provided by low-latency interconnect fabrics such as InfiniBand.

In the "MPI One-Sided Communication" module, we introduce the various components of MPI RMA and how to use them. The "MPI Advanced Topics" module covers some of the advanced capabilities of MPI beyond ordinary message passing, including how to customize your environment in the following areas: derived datatypes; groups of processes and their associated communicators; virtual topologies among processes; and, parallel I/O using MPI-IO. Application to specific architectures such as Ranger, Lonestar, etc. are discussed. In the shared-memory environment that Ranger has on each node, it is much easier to introduce parallelism into your code with OpenMP than to do pthread programming from scratch or to use MPI. The "OpenMP" module introduces OpenMP and describes how to use it. In hybrid programming, the goal is to combine techniques from OpenMP and MPI to create a high-performance parallel code that is better tailored for the non-uniform memory access characteristics of Ranger. To meet this goal, the "Hybrid Programming with OpenMP and MPI" module covers the effects of processor affinity and memory allocation policy, and how to exert some control over them.

There are three visualization modules. Paraview is a visualization application highly capable for computational fluid dynamics and other applications. It is open source and can run in parallel on Ranger. The "Paraview" module includes a lab which covers the visualization of a sample dataset both on a local computer and on TACC resources. VisIt is another visualization application. It is also open source and can run in parallel on Ranger. The "VisIt" module includes a lab which covers visualization of a sample dataset both on a local computer and on TACC resources. VisIT runs on Windows and Linux, as well as TACC's Spur. This module walks through logging into Spur and running a sample visualization. The "Large Data Visualization" module includes a lab which covers visualization of a sample dataset both on a local computer and on TACC resources.

There are five modules focused on code improvement. The "Profiling and Debugging" module covers the basic components of profiling and debugging. This includes how to generate call graphs, listing files, what timers are available, and what data to acquire for profiling. This module provides an in-depth example on using gprof and gmon to identify optimization targets. The Performance Application Programming Interface (PAPI) is also discussed as an effective strategy for analyzing parallel programs. Ranger's sheer size and high core density provide unique opportunities as well as new pitfalls as applications are scaled up. The "Optimization and Scalability" module covers some of these issues and highlights some common programming concerns that users should understand as they try to get the most out of the system. The module provides guidance on what scalability is, how to measure scalability, and how to detect and correct problems that may occur as applications move to petascale systems like Ranger. The "Balancing Scripts and Compiled Code in Scientific Applications" module works through examples of scientific application code written in a mix of scripting languages and C++ or Fortran code in order to evaluate where, within an application, scripting is a good choice. In the "Using Databases" module we discuss relational databases, the most common type of database, and what the reader is most likely to find available at a TeraGrid site. The "Computational Steering" module provides an introduction to what computational steering is, the potential benefits from using it, and examples of how you can integrate steering into your existing application.

4. FEATURES

The VW includes a number of features to support the learning environment.

The navigation is designed to be self-explanatory and shallow. After login, 99% of the pages can be reached in three clicks or fewer. General reference pages are available from all pages via the top navigation bar. The topics are arranged in suggested order, but sequential progression is not enforced. Within a module, the participant can either progress page by page or jump to any page. The current page is marked in the table of contents.

The Reference page (see Fig. 2) gathers bibliographic and Web references for all topics in one place, sorted by topic. All content pages link to the references as appropriate. The Reference page continues to grow as additional relevant information is found and as topics are added to the Virtual Workshop.

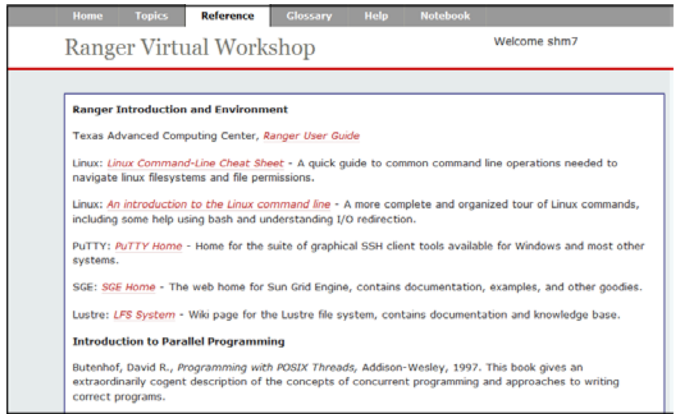

Fig. 2. A Reference page lists references for all topics

The Virtual Workshop includes an HPC Glossary. The Glossary covers terms specific to Ranger as well as general HPC terms for completeness. The Glossary has about 50 terms and continues to grow. The alphabetical Glossary is available from all pages via the top navigation bar (see Fig. 3). Glossary terms are also linked as appropriate in the text. The links pull the definition into the top of the page for an in-line reference (see Fig. 4).

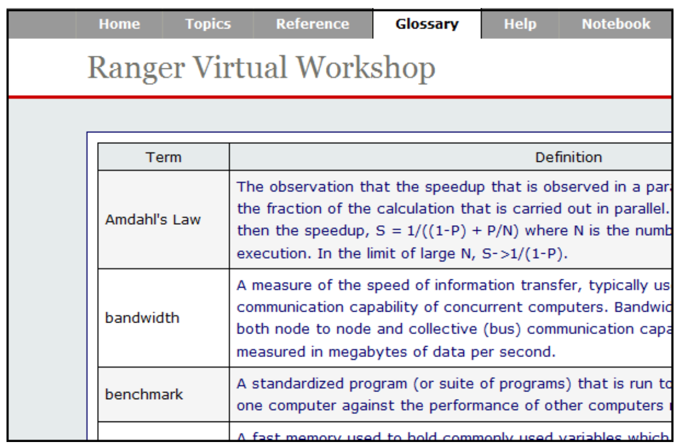

Fig. 3. A full Glossary is available via the top navigation bar

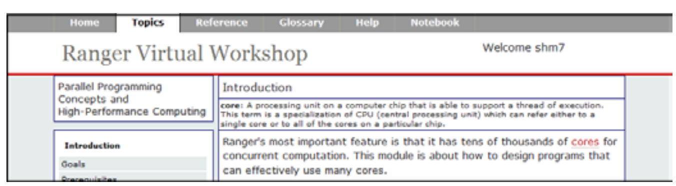

Fig. 4. Using a link to a glossary page pulls the definition into the page

Participants have two tools to reinforce and check what they have learned: exercises and quizzes. Exercises are practical labs, where the participant is given instructions to try out what they have learned. All exercises are written to be run on Ranger (see Fig. 5). Consultants are available for questions on the exercises as well as on the main content by either using links inside the Virtual Workshop or by submitting a ticket to the TeraGrid User Portal help desk system.

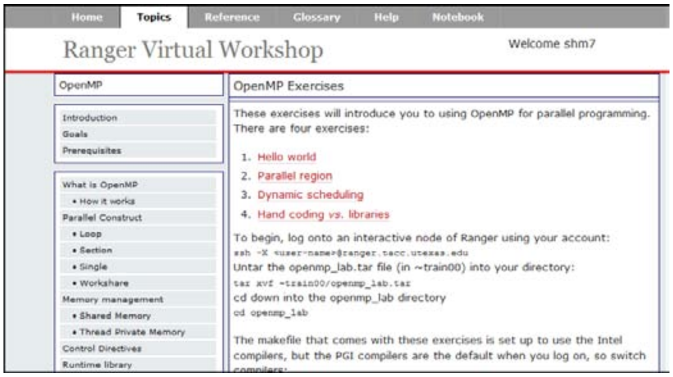

Fig. 5. Example of a lab exercise

Interactive quizzes are presented as multiple-choice questions. Quizzes are automatically graded for the users. Incorrect answers are returned with the correct answer, a short explanation of why it is the best answer, and a link to the corresponding material. The results are also written to a database, enabling the content developers to improve materials for concepts that are commonly misunderstood, allowing effort to be focused appropriately.

To address differences in the way people learn and the occasional difficulty in following written instructions, some procedures are also available as flash movies with audio narration. In the sections where a demonstration is appropriate, screen recordings with audio narration are included (see Fig. 6).

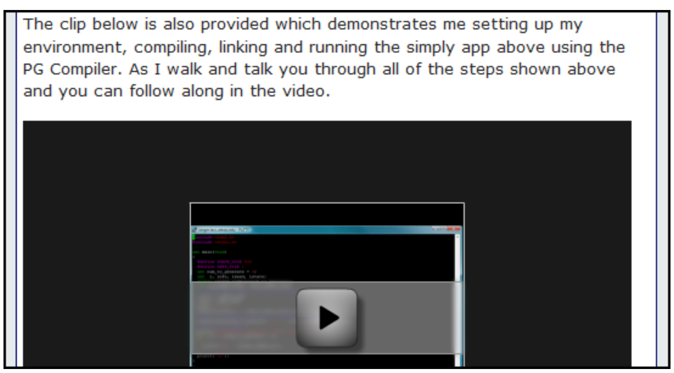

Fig. 6. Demonstrations are included as short video clips

Participants have the opportunity to customize their topic pages. The participant can mark their progress through each topic with a Virtual Workshop-specific bookmark. When they return to the Virtual Workshop, they can easily find their "marked page" for each module. They can also take notes, or "write in the margins" on as many pages as they choose. When a note is added to a page, that note will appear on the page on subsequent visits (see Fig. 7), but only for the person who created the note. Notes can also be updated or erased at any time. If feedback shows sufficient interest, the page note capability will be expanded to enable note sharing with other specified participants.

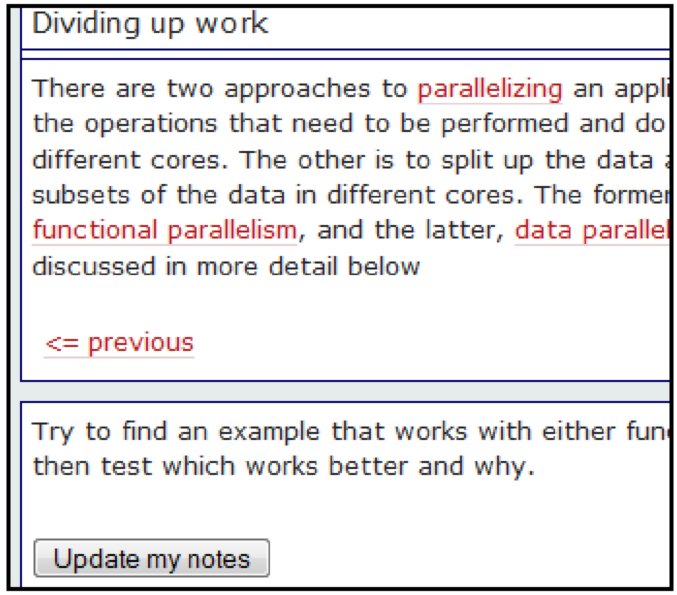

Fig. 7. This example shows a note that has been added to a page

In past Virtual Workshops, participants have asked for a way to view information on how much of the course they have completed. In response, a Notebook page has been provided, which includes a summary of their notes, bookmarks, questions asked, and quizzes submitted, so they can easily review their progress and jump to points of interest (see Fig. 8).

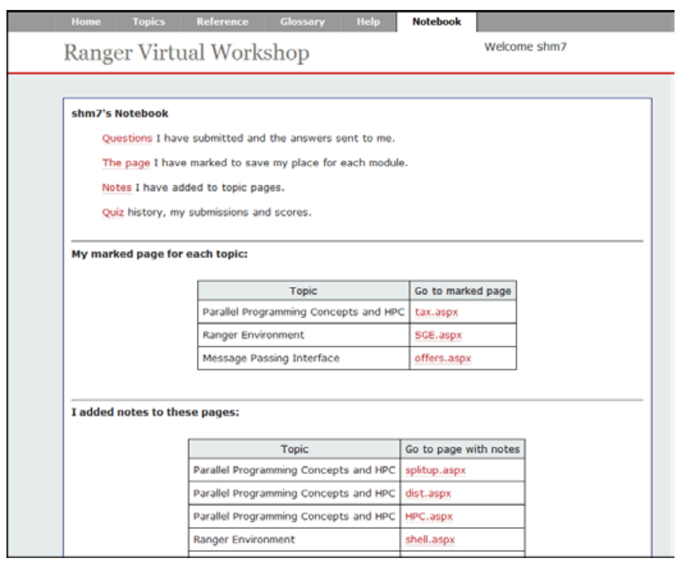

Fig. 8. The Notebook page includes information on notes, bookmarks, and completed exercises and quizzes

5. Assessment Plans

The Ranger Virtual Workshop was designed based on Cornell University experience developing and delivering Web-based training in high performance computing topics since 1995. Cornell VW educational programs have been funded by the National Science Foundation, the Department of Defense, and private industry. The Cornell development model has been to regularly incorporate new features and evaluate and retain the features that enhance the workshop. Experimentation in past workshops included virtual office hours, module walk-throughs, scheduled progression through the topics, and using video and audio clips. Usage of all features in the Ranger Virtual Workshop is tracked. Special attention is paid to understanding which content and features are most-heavily used, how participants progress through the materials, and determining whether participants find the personal notebook valuable.

6. Future Work

The Ranger Virtual Workshop has been in production since March, 2008. At that time it included a core set of introductory modules and all of the basic feedback and note-taking features. Content and features continue to be added and improved. Existing topics will be expanded and improved based on feedback from questions, comments, experience running performance tests, and analysis of tools such as quiz results.

7. Conclusions

Existing topics will be augmented and new topics will be added to the Virtual Workshop. Future development will be guided by usage tracking.

Acknowledgments

The authors wish to thank everyone at TACC for sharing their feedback, time, and expertise with the Cornell CAC in the development of the materials. We also wish to acknowledge the National Science Foundation Office of Cyberinfrastructure for funding computational education for our nation's researchers, from leading faculty to aspiring student scientists and engineers.

References

Allen, Michael W. (2003). *Michael Allen's Guide to e-Learning.* Hoboken, NJ: John Wiley & Sons, Inc.

Mehringer, Susan H. (2002). Enhancing Web-Based Materials with Video Clips. *E-Learn 2002 World Conference on E-Learning in Corporate, Government, Healthcare, & Higher Education. Proceedings (7th, Montreal, Quebec, Canada, October 15-19, 2002).*

Weston, T. & Barker, L. (2001). Designing, Implementing, and Evaluating Web-Based Learning Modules for University Students. *Educational Technology*; v41 n4, 15-22.

Williams, R. (2002). *Web Design Workshop.* Berkeley, CA: Peachpit Press.

Willie, S., Pui, C., & Palmer, A. (1999). Remote Lecture Presentation Preferences for Internet Delivered Continuing Medical Education. *Proceedings of ED-MEDIA 99 World Conference on Educational Multimedia, Hypermedia & Telecommunications, 1999,* Association for the Advancement of Computing in Education, Charlottesville, VA. 190-195.