SeedMe: A Cyberinfrastructure for Sharing Results

Amit Chourasia[1]
San Diego Supercomputer Center
University of California, San Diego

Mona Wong-Barnum[2]
San Diego Supercomputer Center
University of California, San Diego

David Nadeau[3]
San Diego Supercomputer Center
University of California, San Diego

Michael L Norman[4]
San Diego Supercomputer Center
University of California, San Diego

ABSTRACT

Computational simulations have become an indispensible tool in a wide variety of science and engineering investigations. Nearly all scientific computation and analyses create transient data and preliminary results, these may consist of text, binary files and visualization images. Quick and effective assessments of these data are necessary for efficient use of the computation resources, but this is complicated when a large collaborating team is geographically dispersed and/or some team members do not have direct access to the computation resource and output data. Current methods for sharing and assessing transient data and preliminary results are cumbersome, labor intensive, and largely unsupported by useful tools and procedures. Each research team is forced to create their own scripts and ad hoc procedures to push data from system to system, and user to user, and to make quick plots, images, and videos to guide the next step in their research. These custom efforts often rely on email, ftp, and scp, despite the ubiquity of much more flexible dynamic web-based technologies and the impressive display and interaction abilities of today's mobile devices.

To fill this critical gap we have developed *SeedMe* (**Swiftly Encode, Explore and Disseminate My Experiments)**, is a web-based architecture to enable the rapid sharing of content directly from applications running on High Performance Computing or cloud resources. *SeedMe* converts a slow, manual, serial, error prone, repetitive, and redundant sharing and assessment process into a streamlined automatic and web accessible cyberinfrastructure. We provide an easy to use sharing model with granular access control, and mobile device support. *SeedMe* provides secure input and output complementing HPC resources, without compromising their security model, and thereby expand and extend their capabilities. ***SeedMe* aims to foster rapid assessment, iteration, communication, and dissemination of transient data and preliminary results by seeding content that can be accessed via a simple collaborative web interface.**

Keywords: Web Services. Data assimilation. Scientific visualization.

Index Terms: H.3.5 [Information Storage And Retrieval] H.3.5 [Online Information Services]: Data sharing Web-based services H.5.3 [Group and Organization Interfaces] Asynchronous interaction Web-based interaction

1 INTRODUCTION

The increasing availability of High Performance Computing (HPC), Cloud Computing, and high-resolution high-update rate sensing and imaging instruments has enabled researchers to model, observe, and analyze an ever-widening range of scientific phenomena. We are witnessing a rapid increase in scientific data [1, 2, 3] as well as the number of users of scientific computing. Many computing tasks are highly iterative in nature, as the right input parameters are rarely known ahead of time which leads to a repeating process - *run a job, assess results, refine parameters, run another job, and so on*. With rapid access to transient data, job parameters can be refined quickly and computation performed again. *Swift informative job feedback and access to preliminary results is essential for efficient use of cyberinfrastructure and researcher time.*

HPC computation job's results are typically directed to output files that remain unavailable until the job completes or are manually, periodically checked by a single researcher. *Current computation cyberinfrastructure has poor feedback mechanisms.*

When preliminary results are generated, sharing them among collaborators is often limited to emails back and forth, which can't easily include all of the relevant data and appropriate visual ways to display it. *Current cyberinfrastructure has limited support for sharing preliminary results.*

2 SEEDME: SWIFTLY ENCODE, EXPLORE AND DISSEMINATE MY EXPERIMENTS

SeedMe [4] is a cyberinfrastructure to support the rapid collection, presentation, and sharing of transient data and preliminary results generated on remote computation resources. It aims to help collaborating teams shorten iteration times as they explore a problem space with repeatedly runs of compute jobs with varying parameters. By enabling faster, friendlier, easier and more collaborative feedback, *SeedMe* aims to help scientists perform work more efficiently and effectively with big and expensive compute hardware. It is a cyberinfrastructure that can be interfaced with wide range of applications and research domains. The **SeedMe cyberinfrastructure** provides new building blocks to help scientists (1) get fast feedback and preliminary data from compute jobs, (2) share that feedback rapidly with collaborators distributed across the globe, (3) provide access to preliminary results to aid assessment and guide further computation, and (4) support discussion of feedback and results among collaborators.

3 SEEDME CYBERINFRASTRUCTURE

A complex integration of various software components is required for *SeedMe* cyberinfrastructure; this includes a *Apache* webserver [5] is used for handling incoming and outgoing Web requests. We provide upload functionality through Web browser or command line or *Web Services*. The uploaded files are stored temporarily on the incoming shared file system, which is shared with encoding nodes as *NFS file system*. The Web Node runs *Drupal* [5] with *REST Services* [7] that processes all requests. On successful upload the *REST Service* submits the incoming jobs to the *Gearman Job Scheduler* [8], which maintains the state of all submitted jobs and executes them in a serial order. The *Job Scheduler* launches the queued jobs to the encoding nodes, which use FFmpeg[9] to do actual encoding. Completion of success or failure of job is monitored by the *Gearman Job Server*. On job failure the users are notified by email, while successful job results are passed to the *Publisher*. The *Publisher* processes the results, deletes temporary data, copies results to the outgoing file system and submits the result to *Drupal* content management system, to make them available on Web with privacy control settings configured by the user. The user is notified by an email with a URL, which may be used for viewing content, managing

[1]amit@sdsc.edu [2]mona@sdsc.edu
[3]nadeau@sdsc.edu [4]mlnorman@sdsc.edu
IEEE Symposium on Large Data Analysis and Visualization 2014
October 9–10, Paris, France
978-1-4799-5215-1/14/$31.00 ©2014 IEEE

privacy control, editing meta-data and deleting the associated content. The *Drupal* content management system is configured to provide browse, search, edit, upload, download, and video playback capabilities.

4 User Interaction and Tools

Users need to sign up to use the SeedMe cyberinfrastructure; this is required for providing privacy controls and abuse prevention. The user interaction includes the following steps (Figure 1). End users can interact with SeedMe via web browser, command line or REST API. We provide standalone executables as well as a python module which provides convenient utilities to interact with web services at SeedMe.org. The tools perform extensive input data sanity checks and hide other server side complexities that speed up integration, interaction and testing. These tools preclude the requirement of implementing REST client for SeedMe.org by end users. Extensive documentation has been provided to use these tools on the SeedMe website [10].

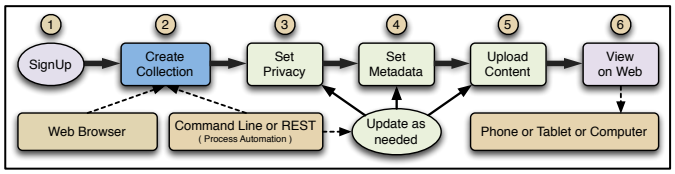

Figure 1: Illustration of sample user interaction workflow for SeedMe.

5 Responsive Web Content

SeedMe translates the user's uploaded content and serves it in a Responsive Web Design [11] i.e. suitably translated and formatted to be optimally viewed on different devices. This includes resizing of images at different resolutions as well as encoding and transcoding of videos at different bitrates. When the content is served to the user, device resolution and bandwidth detection is followed by reformatting of content for suitable view and interaction.

6 Discussion Of Application Scenarios

The overarching goal for SeedMe is to make transient data ubiquitously accessible by providing a mechanism and tools that are automation-friendly for sending updates, and human-centric for assessment and collaborative purposes. The following scenarios illustrate applications of the SeedMe infrastructure in a variety of scientific contexts.

6.1 On Demand Monitoring

A researcher would like to monitor and share progress of their simulation in the form of job progress information, statistics and visualization images with colleagues who may not have access to the computation resource. The researcher sets up a periodic submission to SeedMe using a script or direct integration into the computation code.

6.2 Interface with Data Analysis and Visualization

A researcher conducting analysis of their simulation's raw data on a HPC resource may upload the results programmatically or manually using the SeedMe command line or web interface. These results are shared with team members privately or made available to a broader community. We are using SeedMe with our collaborators in astrophysics, geoscience and ecology to use and instrument their analysis tools with SeedMe for sharing results.

6.3 Application Integration

Applications may integrate and provide an option to concurrently save the transient data and preliminary data directly to SeedMe. The saved content is subsequently available to their user community. Via the python module. SeedMe can also be easily integrated with ParaView [12] and VisIt [13], popular open source visualization software used by several users.

6.4 Share Reusable Content

Several software tools enable saving application state, which may be used to recreate or reuse with other similar data. However, there is no well-defined place to share these reusable templates. Users and application developers could use the SeedMe cyberinfrastructure to share this reusable content. We are working to support IPython notebooks, ParaView and VisIt state files.

6.5 Transient Data Locker

Gateways/portals need to share results with users who do not have access to computation resources. They often have very limited storage capacity and are generally not setup to handle privacy issues for data sharing. SeedMe enables sharing of small-scale data with granular privacy for end users. Data locker functionality on a central website is implemented with strict short-term expiry date to keep storage requirements manageable.

7 Conclusions

The *SeedMe* cyberinfrastructure [10] provides tools and collaborative interface for a wide variety of applications and is widely accessible to researchers. It is being adopted by a diverse range of scientific communities and providing them benefits that includes significant time saving by the convenient and timely presentation of job feedback, transient data, and preliminary results, along with easy data sharing, progress monitoring, accessing plots, images, and swift video encoding..

8 Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. OCI-1235505.

References

[1] T. Hey and A. Trefethen. The data deluge: an e-Science perspective. In F. Berman, G. C. Fix, and A. J. G. Hey, editors, Grid Computing: Making the Global Infrastructure a Reality, pages 809–824. Wiley, 2003.

[2] A. Szalay and J. Gray. Science in an exponential world. Nature, 440, March 23 2006.

[3] C. Anderson. The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. Wired Magazine 16.07 (June 23, 2008)

[4] A. Chourasia, M. Wong-Barnum, M. Norman. SeedMe Preview: Your Results from Disk to Device In Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery (XSEDE '13). ACM, New York, NY, USA, Article 35 , 4 pages.

[5] Apache. 1998. Welcome! The Apache HTTP Server Project. Retrieved Aug 7, 2013 from http://httpd.apache.org

[6] Drupal. 2001.About Drupal. Retrieved Aug 7, 2013 from http://drupal.org/about

[7] Roy Thomas Fielding. 2000. *Architectural Styles and the Design of Network-Based Software Architectures*. Ph.D. Dissertation. University of California, Irvine. AAI9980887

[8] Gearman Job Server. Retrieved Aug 7, 2013 from http://gearman.org

[9] FFmpeg. Retrieved Aug 7, 2013 from http://ffmpeg.org

[10] Rapidly share and access your research results | SeedMe. Retrieved Aug 7, 2013 from http://www.seedme.org

[11] Responsive Web Design: An A List Apart Article. Retrieved Aug 7, 2014 from http://alistapart.com/article/responsive-web-design

[12] Kitware. 2013. ParaView. Open Source Scientific Visualization. Retrieved Aug 7, 2013 from http://www.paraview.org

[13] VisIt. 2013. VisIt Visualization Tool. Retrieved Aug 7, 2013 from https://wci.llnl.gov/codes/visit