IGI PUBLISHING

ITJ 4917

701 E. Chocolate Avenue, Hershey PA 17033-1240, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.igi-global.com

This paper appears in the publication, International Journal of Grid and High Performance Computing, Volume 1, Issue 2
edited by **Emmanuel Udoh and Frank Zhigang Wang © 2009, IGI Global**

Publication and Protection of Sensitive Site Information in a Grid Infrastructure

Shreyas Cholia, Lawrence Berkeley National Laboratory, USA

R. Jefferson Porter, Lawrence Berkeley National Laboratory, USA

ABSTRACT

In order to create a successful grid infrastructure, sites and resource providers must be able to publish information about their underlying resources and services. This information enables users and virtual organizations to make intelligent decisions about resource selection and scheduling, and facilitates accounting and troubleshooting services within the grid. However, such an outbound stream may include data deemed sensitive by a resource-providing site, exposing potential security vulnerabilities or private user information. This study analyzes the various vectors of information being published from sites to grid infrastructures. In particular, it examines the data being published and collected in the Open Science Grid, including resource selection, monitoring, accounting, troubleshooting, logging and site verification data. We analyze the risks and potential threat models posed by the publication and collection of such data. We also offer some recommendations and best practices for sites and grid infrastructures to manage and protect sensitive data. [Article copies are available for purchase from InfoSci-on-Demand.com]

Keywords: Computer Security; Grid Computing; Privacy; Site Security Monitoring

INTRODUCTION

Grid computing has become a very successful model for scientific collaborations and projects to leverage distributed compute and data resources. It has also offered the research and academic institutions that host these resources an effective means to reach a much larger community. As grid computing grows in scope, and as an increasing number of users and resources are plugged into the grid, there is an increasing need for metadata services that can provide useful information about the activities on that grid. These services allow for more sophisticated models of computing, and are fundamental components of scalable grid infrastructures. The scope of these services is fairly broad and covers a variety of uses including resource selection, monitoring, accounting, troubleshooting, logging, site availability and site validation. This list could grow, as grids evolve and other types of metadata become interesting to users and administrators. This means that it becomes important for a grid infrastructure to provide central collection and distribution points that can collate information gathered from multiple sources.

The typical publication model involves pushing data from site based informational end points to central collectors, using streaming feeds or periodic send operations. The central collectors then make this data available to interested parties using standard interfaces and protocols in the form of web services and database query engines. The usability of the grid depends on the widespread availability of this information. Given the increasingly open nature of grid computing these collectors and information services generally present publicly accessible front-ends.

Now consider the implications of this model for a site providing grid resources. Being included in a grid infrastructure means that a large amount of site information suddenly enters the public domain. This could include information deemed as sensitive or private from the perspective of the site, the user or the grid collaboration as a whole. It becomes very important then, to have controls on the access and flow of this data, so that the information sources can decide what data they want published and what data they want restricted. Since these models of informational flow are still evolving in today's grids, these controls are still in the process of being designed into the software infrastructure. As such, there isn't a standard way to control this flow of information. We think there is an urgent need to study the various vectors of information being provided by sites to grid infrastructures. This includes an analysis of the nature of the information itself, as well as the software publishing this information.

In our work, we use the Open Science Grid (OSG) ("Open Science Grid Consortium,") as a case study for this model of information flow, looking at the five major information collection systems within the OSG, and analyzing the security implications of this infrastructure. We also provide some recommendations on improving the current infrastructure to preserve the privacy and security of sensitive information.

THE OPEN SCIENCE GRID

The OSG offers a shared infrastructure of distributed computing and storage resources, independently owned and managed by its members. OSG members provide a virtual facility available to individual research communities, who can add services according to their scientists' needs.

It includes a wide selection of resource providers, ranging from small universities to large national laboratories. This broad range of sites results in a diverse set of security requirements. Reconciling these diverse security priorities is a challenge, and requires close interaction between the sites and the OSG managers. One approach to addressing this issue is to provide the necessary tools in the grid middleware stack, so that sites can configure security

Related Content