GWD-I
GFSG

David E. Martin, IBM
Erwin Laure, CERN
Joel Replogle, Open Grid Forum
May 9, 2008

**OGF22 Data Workshop Report**

Status of This Document

This document provides information to the Grid community. It does not define any standards or technical recommendations. Distribution is unlimited.



Abstract

On February 28, 2008, as part of OGF22, the Open Grid Forum held Data Movement and Management Workshop. During the day-long meeting, speakers from research and industry described challenges in data movement and management and detailed some possible solutions. The goal was to look for common issues and points for collaboration among the OGF working group and with the Storage Networking Industry Association. The common theme that emerged was the challenge of metadata management. Both research centers and commercial enterprises are flooded with data and are struggling to provide useful access.

Contents

## 1. Introduction

On February 28, 2008, as part of OGF22, the Open Grid Forum held an all-day workshop focused on data movement and management. The workshop was organized by the OGF Data Area Directors. The speaker mix was balanced between enterprise and research with particular contributions from the Storage Networking Industry Association (SNIA). The goal of the day was to look for common challenges and identify priorities for future OGF work, especially in the Data Area and in collaboration with SNIA.

The presentation materials from each speaker is available on the OGF web site (http://www.ogf.org/gf/event_schedule/index.php?event_id=9). The sessions are all titled "Data Management Workshop." This report will not detail each presentation, but will instead pull out common points.

## 2. Session 1

### 2.1 Keynote, Paul Strong

Paul described the data collected by eBay and the challenges it presents. eBay has over 3.5PB of data with 2PB available live, mostly transactional data and stored in databases. However, the highest growth is in non-transactional data such as pictures of items. eBay has plans for video support as well, which will use huge amounts of storage. eBay is moving toward more file-based storage and providing meaningful access to the files is one of the biggest challenges.

### 2.2 "Data Management Challenge – The View from OGF," Erwin Laure

Erwin, an OGF Area Director for Data, described the evolution of research Grid environments. In the early Grid environments, each program handled its own data management and GridFTP was the only tool available. This was ok for very basic data analysis, but quickly became unacceptable. The Grid community and OGF have responded by working on database access (DAIS), storage management (SRM), file transfer (RTF, FTS), data location services (RLS, LFC), data management (SRB) and the like.

Although OGF has much work on-going in data movement and management, the OGSA Data WG identified several gaps: standardized metadata access, data catalogs and registries, general purpose replication and caching, and data federation schemes. Through workshops and interactions with SNIA, EGEE and HPDC, additional work around integrated data management, transactions, provisioning, virtualization, provenance, file metadata, streaming and versioning have come up. The OGF Data Area also depends on other parts of OGF and other standards organizations for security, overall management and basic web services.

## 3. Session 2

### 3.1 "The SNIA Industry Landscape," Vincent Franceschini

Vincent, chair of the Storage Networking Industry Association, gave an overview of the SNIA, its organization and priorities. SNIA first focused on the "plumbing" for storage and the issues around networked storage. However, focusing on storage alone is no longer sufficient. There is increasing demand in the storage business for information management, where data is managed and presented in a way useful to the business. Existing standards like ILM, NFSv4 and SMI-S each solve parts of information management, but they need to be coordinated and extended to meet industry needs.

### 3.2 "Technical Council Presentation to OGF on XAM," David Black

David, a member of the SNIA Technical Council, gave an overview of SNIA's efforts around fixed content, which is data (such as digital pictures) that will never change after a certain point. Although the data itself does not change, the metadata may be changed or updated. SNIA's XAM (eXtensible Access Method) effort is intended to provide a common architecture and API for managing metadata of fixed content across different storage systems.

3.3 “Introduction to NFS v4 and pNFS,” David Black
David gave an overview of NFS (Network File System) v4 and details of pNFS (parallel NFS). NFSv4 is an improved version of NFS with strong security and better network behavior. It has broad vendor support and is implemented by all major vendors. pNFS is part of the NFSv4.1 specification. pNFS allows the client to get layout information that tells how the data is arranged on the storage device. This metadata can be used to retrieve the data faster or more efficiently.

## 4. Session 3 and 4

Session 3 and 4 focused on work done in various OGF groups and their relevance to running Grid systems.

P. Fuhrmann showed first experiences in using NFSv4.1 as data access protocol in dCache, a very promising development. A. Jagatheesan discussed how next generation data grid systems would require the inclusion of policies and rule based systems and proposed to focus future work on a high level data management interface. The achievements and challenges in providing a uniform storage management interface (the storage resource manager – SRM, under standardization in the GSM-WG) were presented by J. Jensen. J. Bresnahan discussed data transport challenges in gridFTP and A. Grimshaw demonstrated how existing OGF standards, in particular RNS and ByteIO can be used to link data stored on Grid systems into the local file system.

## 5. Summary

The data workshop was intended to focus on issues of data movement and management. Although both issues were addressed by the speakers, it turned out that the biggest challenge given by many of the speakers was metadata management. From eBay’s need to manage relationships between multimedia data to SNIA’s fixed content metadata efforts to pNFS’s storage layout, metadata is a challenge in exploding data sets.

Although there is little agreement on the exact definition of metadata, a good working definition is data that describes other data. Several OGF efforts have tried with little success to generalize metadata, but the workshop suggests a different approach. Rather than focus on the metadata itself, a metadata management architecture and domain-specific metadata definitions could allow variations in metadata under a common framework. SNIA’s XAM work is a good example of metadata in the fixed-content space. OGF intends to follow this up within the relevant WGs by first further analyzing SNIA’s XAM work and potentially organizing a dedicated workshop with SNIA on this topic later in 2008.

## 6. Intellectual Property Statement



## 7. Disclaimer

## 8. Full Copyright Notice