



**T H E   S T A T E   O F   T H E   A R T**  
**T A S K   3 . 4   -   O P T I M I Z A T I O N   O F   D A T A   A C C E S S**

**WP3 New Grid Services and Tools**

---

Document Filename:       **CG-3.1-SRS-0020-2-0-StateTheOfArt**  
Work package:           **WP3 New Grid Services and Tools**  
Partner(s):               **CYFRONET**  
Lead Partner:             **CYFRONET**  
Config ID:                **CG-3.4-STA-0010-1-0**  
Document classification: **PUBLIC**

---

Abstract: This document gives a brief survey of projects and works related to the problems data access optimization and describes the current state of the art.



**Delivery Slip**

	<b>Name</b>	<b>Partner</b>	<b>Date</b>	<b>Signature</b>
<b>From</b>	WP3 task 4	CYFRONET	28 May 2002	
<b>Verified by</b>				
<b>Approved by</b>				

**Document Log**

<b>Version</b>	<b>Date</b>	<b>Summary of changes</b>	<b>Author(s)</b>
1-0	28 May 2002	Draft version	Jacek Kitowski, Renata Słota, Darin Nikołow, Łukasz Dutka
1-0	29 May 2002	Final version	Jacek Kitowski, Renata Słota, Darin Nikołow, Łukasz Dutka

## CONTENTS

<b>1</b>	<b>INTRODUCTION .....</b>	<b>4</b>
1.1	PURPOSE .....	4
1.2	DEFINITIONS, ACRONYMS, AND ABBREVIATIONS .....	4
1.3	OVERVIEW .....	4
<b>2</b>	<b>THE STATE OF THE ART.....</b>	<b>5</b>
2.1	DATA AND REPLICA MANAGEMENT.....	5
2.2	OPTIMIZATION OF LOCAL DATA ACCESS .....	5
2.3	OPTIMIZATION OF ACCESS TO TAPE RESIDENT FILES .....	7
2.4	ACCESS TIME ESTIMATION .....	7
<b>3</b>	<b>BIBLIOGRAPHY REFERENCES.....</b>	<b>8</b>

---

## 1 INTRODUCTION

### 1.1 PURPOSE

The purpose of this document is to give a short description of the current state of projects and tackles with data access problems.

### 1.2 DEFINITIONS, ACRONYMS, AND ABBREVIATIONS

AML	Automated Media Library
CEA	Component-Expert Architecture
CES	Component Expert Subsystem
CG	CG - CrossGrid
DAP	Data-Access Package
EDG	European Data Grid Project
ETA	Estimated Time of Arrival
MMS	Mass Storage Management System
OGSA	Open Grid Service Architecture
SE	Storage Element
T34	Task 3.4
TRLFM	Tape Resident Large Files - middleware placed on top of an existing HSM system
TSS	Tertiary Storage System

### 1.3 OVERVIEW

The rest of this document is divided into two sections: the first 'section 2' surveys current work done in similar projects. The second one collects bibliography needed to understand the domain of the data access problem.

---

## 2 THE STATE OF THE ART

The task 3.4 tackles with many tasks and to make this document clearer the state of the art for these tasks is presented in separate sections below.

### 2.1 DATA AND REPLICA MANAGEMENT

The data and replica management, which is the main objective of the task 3.4, is also one of the most important tasks of the EDG project, and within this project as the first deliverable two documents [1, 2] were presented, with detailed analysis of current technologies in the mass storage management domain, data access and file systems. Since, this review is still up-to-date, thus there is no need to reproduce here this effort again, because these documents are available freely and the links to them are stated in section 3.

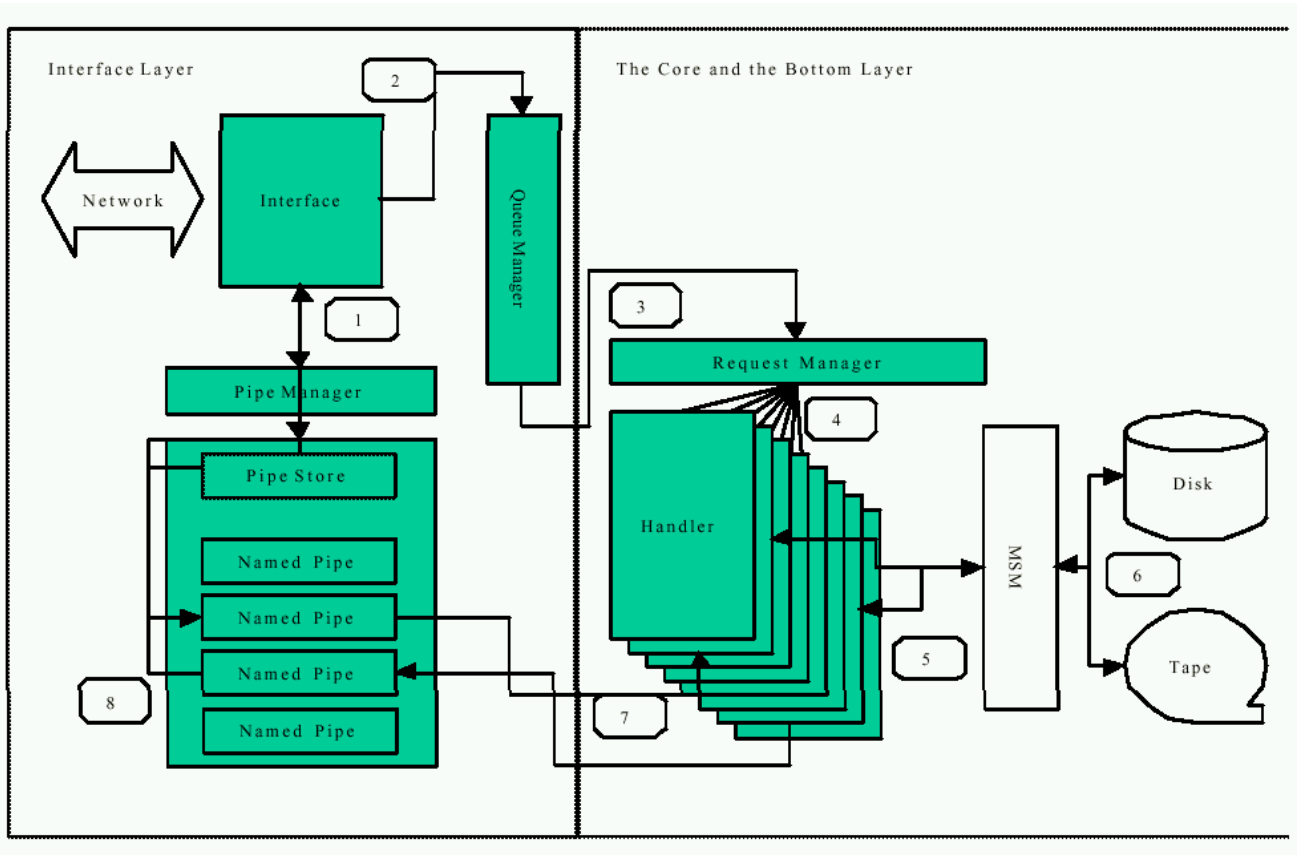
In these documents the authors try to describe how data management is handled by many existing applications, systems and services in a networked environment. They report on the currently available Grid and non-Grid data access technologies and focus on their Data Management aspects.

Different properties and several kinds of storage have been discussed. In this context a review of current disk and tape technologies that are being used in data storage is made. A report on how they are being accessed: locally or over the network is given. Next the current techniques for networked data access are being reviewed.

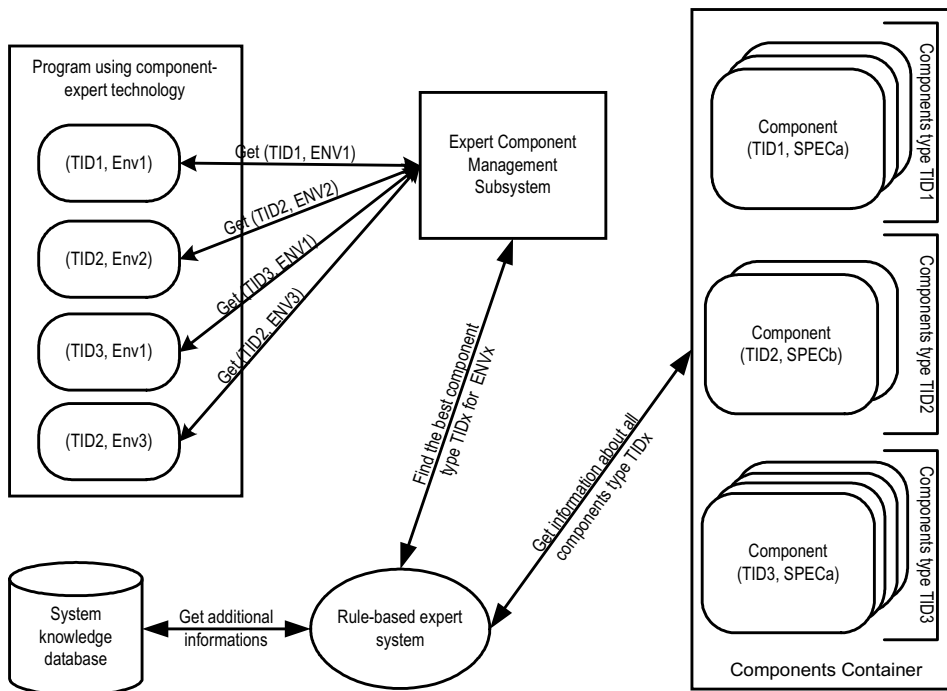
Another topics of the review of data management comes from the world of Hierarchical Storage Management (HSM). Commercial and open HSMs available today are discussed. Some of the HSMs described there are: DiskXtender (former UniTree), HPSS, CASTOR, AMASS. These HSMs have been also evaluated with concern to DataGrid.

### 2.2 OPTIMIZATION OF LOCAL DATA ACCESS

The CrossGrid is a heterogeneous environment particular in the domain of data-access elements. There are different hardware storages – tapes, RAIDs, automatic media libraries (AML) etc; different kinds of applications: interactive simulation and visualization for surgical procedures, flooding crisis team decision support systems, distributed data analysis in high-energy physics, air pollution combined with weather forecasting, etc. These applications require a different response time from the computer systems according to different time scales; from real through intermediate to long time, and they are simultaneously compute- as well as data-intensive. The volume of used data is from mega- to petabytes. This diversification, amongst other things, implies the need of building some mechanisms to optimal data processing in storage centers and allows applications to process data independently of the current grid configuration. The EDG project tackles with similar problems using the Storage Element, which acts as a Grid Service and sits between the client requesting access to data and Mass Storage System, which stores the data. The role of the SE is primarily to sit between the client and the MSS; to hide the MSS differences from the client and to allow access to the MSS using protocols that the client does not naturally support. The internal structure of the SE is presented in figure Fig. 1. This drawing is taken from the document [3] and the detailed description of this solution is available there. Before the CrossGrid project [4] started we had already proposed to extend the possibility of the data-handlers and of the Request Manager (see Fig. 1) by the component-expert architecture [5,6] to make the solution more flexible. This architecture introduces the expert system (the replacement of Request Manager), which selects the best components (the replacement of data-handlers). The outline of this architecture is presented in Fig. 2.



**Fig. 1 Diagram illustrating data flow through the EDG SE.**



**Fig. 2 Connection Diagram of Component-Expert Application**

---

## 2.3 OPTIMIZATION OF ACCESS TO TAPE RESIDENT FILES

The CrossGrid environment will keep some large parts of data in varied tape storages. The problem of the access time optimization to this data is an important issue and one of propositions is a division of data into separate fragments.

Many researches study the access to fragments of data rather than to fragments of files. DataCutter [7] provides support for subsetting very large scientific datasets on archival storage systems. Memik, et al. have developed a run-time library for tape-resident data called APRIL [8] based on HPSS and MPI-IO [9]. It allows programmers to access data located on tape via a convenient interface expressed in terms of arrays and array portions rather than files and offsets. They use a sub-filing strategy to reduce the latency. Holtman, et al. [10] study potential benefits of object granularity in the mass storage system. The architecture is based on transparently re-mapping objects from large chunk files to smaller files according to the application access pattern. Our own development is also of use for the grid environment [11,12,13]

## 2.4 ACCESS TIME ESTIMATION

Another important issue, to the effective work of the CrossGrid environment, is data access time prediction, which is one of the major factors taken into account during the replica selection. The task 3.4 is going to tackle with the problem of the prediction of data-arrival time inside store centers.

Rodney Van Meter in [14] proposes Storage Latency Estimation Descriptors (SLEDs) as a method of supplying to the client predictive information about the I/O performance of the underlying storage systems. By using SLEDs the application can be more efficient by rescheduling its I/O calls in such a way that the less expensive (for instance cached) I/O calls are invoked first. Rodney Van Meter and Minxi Gao in [15] implement SLEDs for the Linux operating system and show significant performance improvement of the applications modified to take advantages of SLEDs. In their implementation the I/O performance estimation is based on latency and bandwidth measurements done during the boot process for each storage device attached to the system.

Shen et al. in [16,17] present a multi-storage architecture and a performance prediction method to increase I/O efficiency of scientific applications. They also developed a run time library on top of Storage Resource Broker (SRB) [18] for optimizing tertiary storage access. Their prediction algorithm is based on time measurements of basic SRB file operations (open, seek, read, close, etc.) and assumes that the file is in the disk cache. They do not concentrate on prediction of the staging time for data located on tertiary storage.

### 3 BIBLIOGRAPHY REFERENCES

1. DataGrid, "Data Access and File Systems – The State of The Art Report" [http://grid-data-management.web.cern.ch/grid-data-management/docs/DataGrid-02-D2.1-0105-2\\_0.pdf](http://grid-data-management.web.cern.ch/grid-data-management/docs/DataGrid-02-D2.1-0105-2_0.pdf)
2. DataGrid, "WP5 Mass Storage Management – Review of Current Technologies" <http://edms.cern.ch/document/336677/>
3. DataGrid, "Architectre and Design WP 5 Mass Storage Management" <http://edms.cern.ch/document/336679/>
4. CrossGrid Project Technical Annex, 2001 "Description of Work", [http://www.cyf-kr.edu.pl/crossgrid/CrossGridAnnex1\\_v31.pdf](http://www.cyf-kr.edu.pl/crossgrid/CrossGridAnnex1_v31.pdf)
5. Dutka, Ł., and Kitowski, J., „Implementation of expert technologies in information systems based on a component methodology”, MSK 2001 Conf., Nov. 19-21,2001 Cracow (in Polish).
6. Dutka, Ł., and Kitowski, J., „Component-expert technology in mass-storage grid applications”, ICCS 2002 Conf., April 2002, Amsterdam.
7. Beynon, M., Ferreira, R., Kurc, T., Sussman, A., and Saltz, J., "DataCutter: Middleware for Filtering Very Large Scientific Datasets on Archival Storage Systems", Proc. of Eighth NASA Goddard Conference on Mass Storage Systems and Technologies and the Seventeenth IEEE Symposium on Mass Storage Systems}, Maryland, USA, March 27-30, 2000, pp.119-133.
8. Memik, G., Kandemir, M.T., Choudhary, A., Taylor, V.E., "April: A Run-Time Library for Tape-Resident Data", Proc. Of Eighth NASA Goddard Conference on Mass Storage Systems and Technologies and the Seventeenth IEEE Symposium on Mass Storage Systems}, Maryland, USA, March 27-30, 2000, pp. 61-74.
9. Corbett, P., Fietelson, D., Fineberg, S., Hsu, Y., Nitzberg, B., Prost, J., Snir, M., Traversat, B., and Wong, P., "Overview of the MPI-IO parallel I/O interface", Proc. of Third Workshop on I/O in Paral. and Distr. Sys.}, Santa Barbara, USA, April 1995.
10. Holtman, K., Stok, P., Willers, I., "Towards Mass Storage Systems with Object Granularity", Proc. of Eighth NASA Goddard Conference on Mass Storage Systems and Technologies and the Seventeenth IEEE Symposium on Mass Storage Systems}, Maryland, USA, March 27-30, 2000, pp.135-149.
11. Nikolow, D., Slota, R., Kitowski, J., Nyczyk, P., Otfinowski, J., "Tertiary Storage System for Index-Based Retrieving of Video Seqences",in: Hertberger, B., Hoekstra, B., Williams, R. (Eds.), Proc. Int. Conf. High Performance Computing and Networking, Amsterdam, June 25-27, 2001, Lecture Notes in Computer Science 2110, pp. 62-71, Springer, 2001.
12. Nikolow, D., Slota, R., Kitowski, J., "Benchmarking Tertiary Storage Systems with File Fragmentation", PPAM2001 Conf., Nałęczów, Lect.Notes in Comp.Sci., in press.
13. Slota, R., Kosch, H., Nikolow, D., Pogoda, M., Bredler, K., Podlipnig, S., "MMSRS - Multimedia Storage and Retrieval System for a Distributed Mediacal Information System" in: Bubak, M., Afsarmanesh, H., Williams, R., Hertberger, B., (Eds.), Proc. Int. Conf. High Performance Computing and Networking, Amsterdam, May 8-10, 2000, Lecture Notes in Computer Science 1823, pp. 517-524, Springer, 2000
14. Meter, R., V., "SLEDs: Storage latency estimation descriptors", In Ben Kobler, editor, in Proc. 6th NASA Goddard Conference on Mass Storage Syst. and Tech. in Coop. with 15th IEEE Symp. on Mass Storage Syst., pp. 249-260, March 1998.

15. Meter, R., V., Gao, M., "Latency Management in Storage Systems", in Proc. of the 4th Symp. on Operating Syst. Design and Implementation (OSDI'00), October 2000.
16. Shen, X., Choudhary, A., "A Distributed Multi Storage Resource Architecture and I/O Performance Prediction for Scientific Computing" in Proc. 9th IEEE Symp. on High Performance Distributed Computing, pp.21-30, IEEE Computer Society Press, 2000.
17. Shen, X., Liao, W., Choudhary, A., "Remote I/O Optimization and Evaluation for Tertiary Storage Systems through Storage Resource Broker", in IASTED Applied Informatics, Innsbruck, Austria, February, 2001.
18. Baru, C., Moore, R., Rajasekar, A., Wan, M., "The SDSC Storage Resource Broker", in Proc. CASCON'98 Conference, Toronto, Canada, Dec. 1998.