Chapter 5
Grid Data Handling

Alexandru Costan
University Politehnica of Bucharest, Romania

ABSTRACT

To accommodate the needs of large-scale distributed systems, scalable data storage and management strategies are required, allowing applications to efficiently cope with continuously growing, highly distributed data. This chapter addresses the key issues of data handling in grid environments focusing on storing, accessing, managing and processing data. We start by providing the background for the data storage issue in grid environments. We outline the main challenges addressed by distributed storage systems: high availability which translates into high resilience and consistency, corruption handling regarding arbitrary faults, fault tolerance, asynchrony, fairness, access control and transparency. The core part of the chapter presents how existing solutions cope with these high requirements. The most important research results are organized along several themes: grid data storage, distributed file systems, data transfer and retrieval and data management. Important characteristics such as performance, efficient use of resources, fault tolerance, security, and others are strongly determined by the adopted system architectures and the technologies behind them. For each topic, we shortly present previous work, describe the most recent achievements, highlight their advantages and limitations, and indicate future research trends in distributed data storage and management.

DOI: 10.4018/978-1-61350-113-9.ch005

INTRODUCTION

During the last years, mainly motivated by the need of applications in eScience where vast amounts of data are generated by specialized instruments and need to be collaboratively accessed, processed and analyzed by a large number of scientists around the world, grid computing has become increasingly popular. The grid embraced the goal of sharing potentially unlimited computing power over the Internet to solve complex problems in a distributed way. A first generation of grids, called computational grids, focused on CPU cycles as resources to be shared. Recent advances in grid computing aim at virtualizing different types of resources (data, instruments, computing nodes, tools) and making them transparently available.

Along with the computational grids, a second generation of grids, namely Data Grids (Chevernak et al. 2000), has emerged as a solution for distributed data storage and management in data-intensive applications. The size of data required by these applications may be up to petabytes. In many applications, Data Grids not only maintain raw data produced by instruments, but need to take into account also aggregations and derivations of these huge size raw data that are periodically generated and potentially concurrently updated by scientists at several sites. Data intensive grids primarily deal with providing services and infrastructure for large scale distributed applications that need to access, transfer and modify massive datasets stored in distributed storage resources. High Energy Physics, governmental and commercial statistics, climate modeling, cosmology, genetics, bio-informatics, etc. are just a few examples of fields routinely generating huge amounts of data. It becomes crucial to efficiently manipulate these data, which must be shared at the global scale.

These data intensive grids combine high-end computing technologies with high-performance networking and wide-area storage management techniques. Many approaches to build highly available and incrementally extendable distributed data storage systems have been proposed. Solutions span from distributed storage repositories to massively parallel and high performance storage systems. A large majority of these aims at a virtualization of the data space allowing users to access data on multiple storage systems, eventually geographically dispersed. While these new technologies reveal huge opportunities for large-scale distributed data storage and management, they also raise important technical challenges, which need to be addressed. The ability to support persistent storage of data on behalf of users, the consistent distribution of up-to-date data, the reliable replication of fast changing datasets or the efficient management of large data transfers are just some of these new challenges.

The objective of this chapter is to give the reader an up-to-date overview of modern data storage and management solutions in grid environments. We discuss the main challenges, and present the most recent research approaches and results adopted in large scale distributed systems, with emphasis on incorporating efficient techniques that increase the reliability and support higher efficiency of the applications running on top of distributed platforms. Future research directions in the area of data storage and processing are highlighted as well.

BACKGROUND

Data intensive environments often deal with applications that produce, store and process data in the range of hundreds of megabytes to petabytes and beyond. The data may be structured or unstructured and organized as collections or datasets that are typically stored on mass storage systems (also called repositories) such as tape libraries or disk arrays. These storage resources are geographically dispersed and usually span over different administrative domains. The data sets are maintained independent of the underlying storage systems and are able to include new sites without major