Chapter IX

Data Replication Strategies in Wide-Area Distributed Systems

Sushant Goel, University of Melbourne, Australia

Rajkumar Buyya, University of Melbourne, Australia

Abstract

Effective data management in today’s competitive enterprise environment is an important issue. Data is information, and information is knowledge. Hence, fast and effective access to data is very important. Replication is one widely accepted phenomenon in distributed environments, where data is stored at more than one site for performance and reliability reasons. Applications and architectures of distributed computing have changed drastically during the last decade, and so have replication protocols. Different replication protocols may be suitable for different applications. In this chapter, we present a survey of replication algorithms for different distributed storage and content-management systems including distributed database-management systems, service-oriented data grids, peer-to-peer (P2P) systems, and storage area networks. We discuss the replication algorithms of more recent architectures, data grids and P2P systems, in detail. We briefly discuss replication in storage area networks and on the Internet.
Introduction

Computing infrastructure and network-application technologies have come a long way over the past 20 years and have become more and more detached from the underlying hardware platform on which they run. At the same time, computing technologies have evolved from monolithic to open and then to distributed systems (Foster & Kesselman, 2004). Both scientific and business applications today are generating large amounts of data; typical applications, such as high-energy physics and bioinformatics, will produce petabytes of data per year. In many cases, data may be produced or are required to be accessed or shared at geographically distributed sites. The sharing of data in a distributed environment gives rise to many design issues, for example, access permissions, consistency issues, and security. Thus, effective measures for easy storage and access of such distributed data are necessary (Venugopal, Buyya, & Ramamohanarao, 2005). One of the effective measures to access data effectively in a geographically distributed environment is replication.

Replication is one of the most widely studied phenomena in a distributed environment. Replication is a strategy in which multiple copies of some data are stored at multiple sites (Bernstein, Hadzilacos, & Goodman, 1987). The reason for such a widespread interest is due to the following facts.

1. Increased availability
2. Increased performance
3. Enhanced reliability

By storing the data at more than one site, if a data site fails, a system can operate using replicated data, thus increasing availability and fault tolerance. At the same time, as the data are stored at multiple sites, a request can find the needed data close to the site where the request originated, thus increasing the performance of the system. But the benefits of replication, of course, do not come without overheads of creating, maintaining, and updating the replicas. If the application has a read-only nature, replication can greatly improve the performance. But, if the application needs to process update requests, the benefits of replication can be neutralised to some extent by the overhead of maintaining consistency among multiple replicas, as will be seen in the following sections of the chapter.

A simple example of a replicated environment is shown in Figure 1. Site 1, Site 2, Site 3, ..., and Site \( n \) are distributed site locations and are connected through a middleware infrastructure (for the time being, it does not matter what the middleware consists of). Data stored in a file, File X, is stored at Site 2 and is replicated at all other sites. Suppose User 1 tries to access File X in Figure 1. For pedagogical simplicity, let the distance shown in the figure be proportional to the access cost.
Related Content

Concepts for the Design of a Manager-Centric Global Decision Support System for Organizations
[www.igi-global.com/chapter/concepts-for-the-design-of-a-manager-centric-global-decision-support-system-for-organizations/133095?camid=4v1a](www.igi-global.com/chapter/concepts-for-the-design-of-a-manager-centric-global-decision-support-system-for-organizations/133095?camid=4v1a)

Ontology-Based Knowledge Management for Enterprise Systems
Mohammad Nazir Ahmad, Nor Hidayati Zakaria and Darshana Sedera (2011). *International Journal of Enterprise Information Systems* (pp. 64-90).
[www.igi-global.com/article/ontology-based-knowledge-management-enterprise/60404?camid=4v1a](www.igi-global.com/article/ontology-based-knowledge-management-enterprise/60404?camid=4v1a)

A Qualitative Study of Green IT Adoption within the Philippines Business Process Outsourcing Industry: A Multi-Theory Perspective
[www.igi-global.com/article/a-qualitative-study-of-green-it-adoption-within-the-philippines-business-process-outsourcing-industry/143266?camid=4v1a](www.igi-global.com/article/a-qualitative-study-of-green-it-adoption-within-the-philippines-business-process-outsourcing-industry/143266?camid=4v1a)
Business Acceptance of Information Technology: Expanding TAM Using Industry Sector and Technological Compatibility
[www.igi-global.com/article/business-acceptance-information-technology/2152?camid=4v1a](www.igi-global.com/article/business-acceptance-information-technology/2152?camid=4v1a)