Chapter 6.2
Database High Availability: An Extended Survey

Moh'd A. Radaideh
Abu Dhabi Police – Ministry of Interior, United Arab Emirates

Hayder Al-Ameed
United Arab Emirates University, United Arab Emirates

ABSTRACT

With the advancement of computer technologies and the World Wide Web, there has been an explosion in the amount of available e-services, most of which represent database processing. Efficient and effective database performance tuning and high availability techniques should be employed to insure that all e-services remain reliable and available all times. To avoid the impacts of database downtime, many corporations have taken interest in database availability. The goal for some is to have continuous availability such that a database server never fails. Other companies require their content to be highly availabile. In such cases, short and planned downtimes would be allowed for maintenance purposes. This chapter is meant to present the definition, the background, and the typical measurement factors of high availability. It also demonstrates some approaches to minimize a database server's shutdown time.

INTRODUCTION

High availability of software systems has become very critical due to several factors that are related

Table 1. Downtime measurements at various availability rates

Availability Percentage	Downtime Percentage	Service Downtime (Minutes/Year)
95%	5%	50000
97%	3%	15840
98%	2%	10512
99%	1%	3168
99.5%	0.5%	2640
99.8%	0.2%	1050
99.9%	0.1%	528
99.95%	0.05%	240
99.99%	0.01%	53
99.999%	0.001%	5
99.9999%	0.0001%	0.51
99.99999%	0.00001%	0.054

to the environment, processes and development strategies, hardware complexity, and the amount of dollars and human resources invested in the system. High availability cannot be achieved by just implementing a given service level or solution. Systems should be designed such that all factors that may lead the system to go down should be well-treated, if not eliminated.

In today's competitive business landscape, 24/7 operations become the standard, especially for the e-services-driven areas (e.g., e-commerce, e-government, e-learning, etc.) Downtime of applications, systems, or networks typically translates into significant revenue loss. Industry experts and analysts agreed on that in order to support e-service applications, typical network availability must reach 99.999%. In other words, networks must be at the "5-Nines" availability level (Providing Open Architecture, 2001). Reaching this level of availability requires careful planning and comprehensive end–to-end strategy. To demonstrate the impact of not being at the "5-Nines" availability level, a system with 97% availability will incur approximately 263 hours (6.6 days) of downtime per year. With 99 percent availability, downtime will be 88 hours (2.2 days) per year. Table 1 summarizes the impact of service downtime according to the availability ratings.

High Availability is not achieved through a single product or process. It is the result of an end-to-end analysis and reengineering of the entire service chain including the combination of people, processes, and technological factors (Otey & Otey, 2005). Every device or circuit in the path between client and server is a link in this service chain, and each must be considered separately. A chain is only as strong as its weakest link. As more applications are delivered via Web browsers, the emphasis for high availability is spreading from back-end databases toward front-end and middleware devices like Web servers and firewalls. Database management systems (DBMS) play a pivotal role in much of today's business computing environment, underpinning electronic services operations, providing critical business support through data warehousing and mining, and managing the storage and processing of much of the world's financial data. As they are entrusted with the storage and processing of such critical data, one would assume that databases are designed to be reliable and highly available.

This chapter provides an overview of the high availability in general, and describes the business drivers behind it, or how it is measured. It focuses on the meaning of database high availability, its functionality and design strategies that emerge with the shift from technology-centric orientation of keeping the system running, to a more customer-centric focus on ultra-dependable services. The view of high availability provided in this chapter has no bias towards high availability practices offered today by the different DBMS vendors.

This chapter is organized into seven sections. The first section provides a generic introduction on the chapter's subject. The second section overviews the high availability-related issues. The third section discusses the model environment for highly-available systems. The fourth section discusses several strategies for database high availability. The fifth section discusses performance impact of high availability. The sixth section overviews several high availability solutions. The seventh section overviews a simple availability-benchmarking methodology.

HIGH AVAILABILITY OVERVIEW

A system is composed of a collection of interacting components. A system provides one or more services to its consumers. A service is the output of a system that meets the specification for which the system was devised, or agrees with what system users have perceived as the correct output values.

Service failures are incorrect results with respect to the specification or unexpected behavior perceived by the users of the service. The cause of