International Journal of Advanced Culture Technology Vol.8 No.3 292-299 (2020)
DOI https://doi.org/10.17703/IJACT.2020.8.3.292

IJACT 20-9-38

A Study of Information System Availability Guarantee Methods and Application

Hee Wan Kim

[1]*Professor, Division of Computer Science & Engineering, Sahmyook Univ., Korea*
E-mail hwkim@syu.ac.kr

Abstract

This paper presents an evaluation criteria of an information system availability for guaranteeing availability (service target level) from the perspective of the SLA contract and its technical point of view. In order to verify the effectiveness for information system failure and availability guarantee measures, three cases were examined. In summary, the failure time was reduced by 32% ~ 62% after applying the availability guarantee measure, verifying the excellence in the evaluation of an information system availability.

Keywords: *Information System, Availability, Failure, SLA, Evaluation*

1. INTRODUCTION

The current modernization in info and computer technology (ICT) has present the higher secured and brisk[1]. Research on information system availability has been conducted through stable supply of IT services and continuous improvement of service quality. In the HW availability field, availability targets are limited by physical HW configurations. Yet, there are technologies and solutions that can build a system that targets high availability called high availability (HA). From a perspective of a SLA contract, however, there is no standard for HW configuration to guarantee availability (service target level); so sometimes the service target level is set by the customer's demand. In this paper, we study the criteria for the objective information system availability level that the contract companies (users, service providers) can agree to at the time of IT outsourcing SLA contract, examine evaluation method of the information system availability level from a technical perspective, and prove its effectiveness through analyzing various cases.

2. RELATED WORK

2.1 Information System Availability

The general IT system configuration for providing information service is as shown in Figure 1, and can be broadly divided into a network area, a hardware area, and a database area from the service provider's point of view [1].

Manuscript Received: July 25, 2020/Revised:August 17, 2020 / Accepted: August 28, 2020
Corresponding Author:hwkim@syu.ac.kr
Tel:+82-2-3399-1783, Fax: +82-2-3399-1791
Author's affiliation
Professor, Division of Computer Science & Engineering, Sahmyook University., Korea

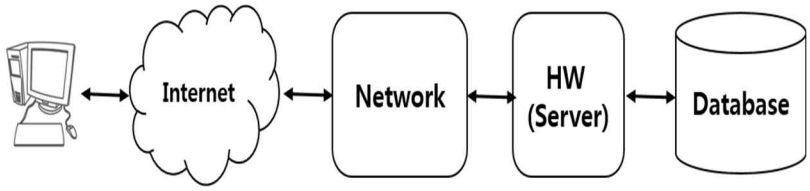

Figure 1. IT System Components

In Figure 1, the external user or system uses the service of the target system through the network (access device, network, interface, etc.), and the system that provides the service (HW) implements service through programming, user interface, security, and data technology, and provide data from external requests through the Database.

The availability of the system refers to the ratio of the time during which the system operates normally during the entire service provision time. An equation can be expressed using Mean Time Between Failure (MTBF), Mean Time To Recover (MTTR), and Mean Time To Failure(MTTF), and their relative relationships are depicted in formula (1) and Figure 2.

$$\text{Availability (A)} = (\text{MTBF} / (\text{MTBF} + \text{MTTR})) * 100 \quad (1)$$

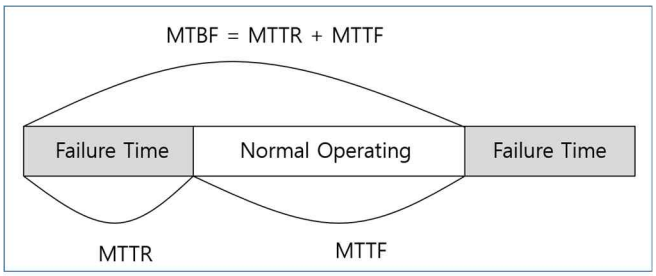

Figure 2. Relationship between availability elements

In Figure 2, Mean Time Between Failures (MTBF) is an average defect interval time, and Mean Time To Repair (MTTR) represents an average recovery time. In addition, MTTF (Mean Time To Failure) represents the time that any system or component of the system can continue to perform without failure, or the life expectancy until failure. In the above equation, availability (A) refers to the time from when the system starts to operate until the next failure, that is, the time it normally operates. Availability of 100% means that resources are always available and there is no system downtime. However, obtaining 100% availability is very difficult, and the highest level of availability that can be achieved is 99.999%, which is referred to as High Availability (HA). The level of HA of 99.999% means that there was only 5 minutes of downtime for one year, 24 hours * 365 days.

2.2 Information System Availability Obstacles Elements

The factors that hinder the availability of information systems can be largely divided into obstacles and disasters. Disorders in information systems include natural disasters, system failures, and infrastructure failures (operational failures, equipment failures, etc.) that interfere with the normal operation of the information system. Disaster is defined as damage that causes disruption to normal business performance and interrupt services by events that cannot be prevented and controlled from outside the information system [2].

Disaster recovery is to resume services based on suspended information systems due to disasters and maintain services. The main factors of service failures are service interruption and service delay. The main factors and solutions of service failures are shown in Table 1.

Table 1. Main cause of service failures

	Main Cause	Solution
Application System	Application malfunction due to source code error Full scan of large data due to SQL error	Complete test before distribution Strengthening development capacities
HW Failure	Fault of HW component(Disk, CPU, Memory) Physical disorder such as natural disasters	High availability environment Periodic Infra configuration verification
Database	Abnormal down of DBMS Physical data storage fault	Periodic monitoring Duplication of DB server
Network	Physical equipment fault	Duplication of network configuration Periodic verification of duplication configuration

3. AVAILABILITY GUARANTEE TECHNOLOGY

The methods for ensuring the availability of information systems include network availability guarantee, system availability guarantee, data and database availability guarantee, and periodic verification and test techniques for stable operation of the availability guarantee environment.

3.1 Network Availability Guarantee Method

Network path redundancy is a technology that supports the real-time search of the redundant path for customer service access as shown in Figure 3.

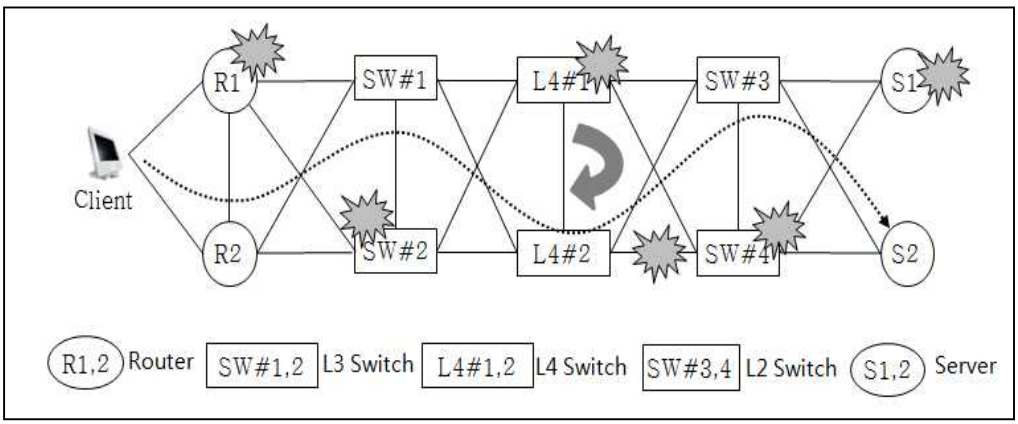

Figure 3. Network Availability Guarantee Concept

In the conceptual diagram of network availability guarantee of Figure 2, Client -> R1, R2 is a technology that enables R2 to be dynamically searched even when R1 fails through VRRP (Virtual Router Redundancy Protocol). R1, R2 -> SW1 and SW2 maintain the path even in the event of a failure between SW1 and SW2 through Dynamic Routing Protocol. SW1, SW2 -> L4 SW1, L4 SW2 section is a section in which availability is guaranteed by using the virtual IP of L4. Two L4 switches share a common IP for service provision, and if a problem occurs, a normal L4 switch performs the function of Virtual IP and guarantees service availability. L4 SW1, L4 SW2 -> S1, S2 section is applied with various availability guarantee technologies. For example, upon checking the status of the server through the health check function of the L4 Switch, if there is a problem with S1, the service request is sent only to S2, and from S2's point, the path of L4 SW2 can be maintained through VRRP.

3.2 System Availability Guarantee Method

The basic components for ensuring the availability of the system are shown in Figure 4, but in order to

guarantee the actual availability, it consists of various components such as CPU, memory, network device, application, operating system, disk, DBMS, storage, etc. Redundancy of these components ensures high availability and allows the components of the system to be configured redundantly so that if one component fails, the other component performing the same function automatically performs the function component that has the problem.

The clustering method includes Active-Standby, Active-Active, and multi-node clustering according to the operation method of Server A and Server B in Figure 4. To establish clustering, status messages are changed between hosts configured in clusters, and HA Link must be configured between hosts. The HA Link exchanges status information between each node and recognizes the failure situation and operates automatic switching [3][4][5].

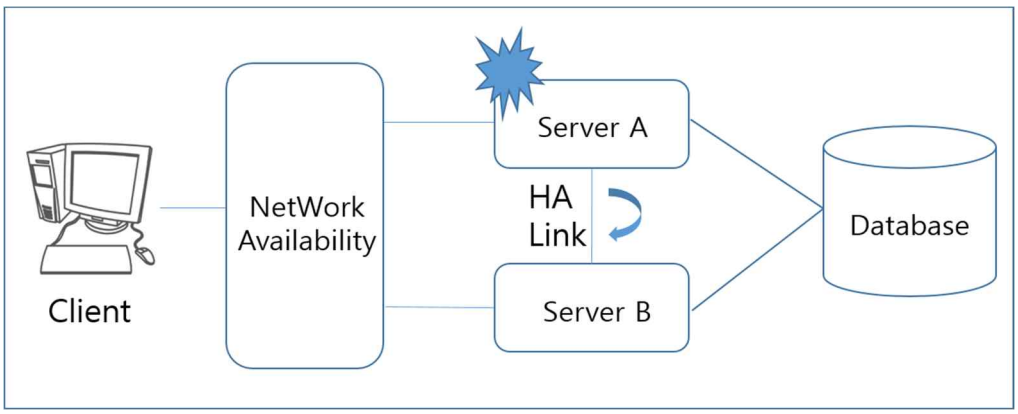

Figure 4. Server Availability Guarantee Concept

A dedicated standby server can be constructed to be a dedicated standby server for multiple active servers. In the form of having two active servers and one dedicated standby server that provide different services, if the active server fails, the dedicated standby server will take over the failed server.

In the case of mutual standby, each server becomes a mutual standby server. Suppose that three servers provide DB service, web service, and file service, respectively. When the DB server goes down, the web server takes over the duties of the DB server and provides web services and DB services. If the web server goes down, the file server provides web services and file services. Also, when the file server is down, the DB server is configured to provide both DB service and file service.

When deployed in the form of cascading failover, when a server that fails to service a service goes down like a chain, all services provided by the failover server are transferred to the next server. Finally, if only one server remains, one server is designed to do all services. When there are three servers, DB server, Web server, and file server, if the DB server goes down, the Web server provides both the Web service and the DB service. At this time, if the web server is down, the file server is implemented to provide all file services, web services, and DB services. This configuration should be designed so that one server can provide all services in case only one server is finally left.

Distributed error recovery (Failover) is a configuration that provides service by failing over each service to a different server when one main server provides all three services. This is a reasonable configuration when the main server has a high system specification and the standby servers have a low system specification.

3.3 Data and Database Availability Guarantee Method

Rapid and stable access to data in a large-scale database environment is essential to ensure availability. Not only is the redundancy of the access path essential, but also the availability guarantee technology that is capable of providing a stable service even when a specific disk fault occurs. Some of the related technologies entail storage configuration, data recovery, and database duplication.

① There are three types of storage configuration methods: direct attached storage (DAS), network attached storage (NAS), and storage area network (SAN).

② The data recovery method has RAID (Redundant Array of Inexpensive Disks) technology, and a RAID system requires multiple disks and is a mechanism to prevent data loss due to failure of any one of the disks. Advantages of RAID include increased transfer speed, increased number of I/O transactions per second, increased online storage capacity, improved data availability and system reliability, ease of managing large amounts of data, and reduced maintenance and downtime [5].

③ There are two ways to guarantee database availability: parallel database and backup database.

▪ The parallel database method provides availability for server failures as a database capable of parallel processing by placing data on one shared storage and running database instances on multiple servers [6].

▪ The backup database method builds up the database in the same environment as the production server as the standby server. If the database on the production server goes down, the database can be run by the standby server to increase service availability [4][7].

3.4 Periodic Verification and Test

No one can predict when, where, and in what way a service failure will occur. Even if the network, server, storage, and database availability guarantee measures are applied, the verification of normal functions should be conducted continuously. Through verifying and testing the process of recovering failure periodically as a preparation, when an actual failure occurs, a quick recovery can be accomplished. As shown in Table 2, the cluster switching test and backup cycles are different according to the importance of service level.

Table 2. Validation and analysis cycle by availability grades

Availability Grade		S grade	A grade	B grade	C grade
Availability management activity	Server	Monthly (Quarterly offline test)	Quarterly (Annual offline test)	Semi-annual	Annual
	Storage network backup	Monthly	Quarterly	Semi-annual	Annual
Backup management	Backup recovery	5 weekdays, 1 weekend	3 weekdays, 1 weekend	1 weekdays, 1 weekend	1 weekend

4. ANALYSIS OF AVAILABILITY EFFECTIVENESS

In order to verify the effectiveness of the information systems currently in operation and the effectiveness of the availability guarantee measures, the case studies a domestic Company A was used. The study shows that the Infra availability guarantee measures reduce service disconnect time, however, the standards for Infra composition were not well observed.

Company A is providing services in the form of collective outsourcing of about 250 services, including internal and external services, and the overall failure status is shown in Table 3.

Table 3. Company A's failure status

Failure Area	Failure Time(min)	Number of Failure

Development Failure	2,930	10
Operation Failure	1,342	13
Infra Failure	3,380	34
Total	7,652	57

The total HW disorders of Company A were confirmed to be 34 cases, 3,380 minutes, and among the total disorders, about 60% of cases and about 44% of the break time were identified as Infra disorders. Table 4 shows the effect of reducing the service disconnection time when the availability guarantee measures are applied for each type of Infra failure.

Table 4. Effect of A's availability guarantee measures

	CPU	DBMS	DISK	NIC	OS	Power	Network	Etc	Total
Number	2	7	2	2	4	2	3	12	34
Time(min)	50	485	350	405	535	121	669	766	3,380
Availability guarantee measures	Cluster	DB duplication	Raid	IPMP Config.	Cluster	Power duplication	L4 duplication		
Reduced time(min)	20	210	300	375	240	0	405	765	2,315

The effect of service disconnection was verified through the review of the 1st stage failure report and the 2nd stage operation personnel interview. As a result, the failure guarantee time can be reduced by 32% through applying availability guarantee measures to 34 cases, 3,380 minutes that can be under SLA.

Company B is a semiconductor company, and the disability status is as shown in Table 5. When the availability guarantee measures are applied through a direct interview with the field operator on the 19 cases that occurred in Company B, and based on the opinions of field experts, the availability effects in statistics are shown in Table 5.

Table 5. Effect of Semiconductor related B's availability guarantee measures

	Controller	DB	OS	Disk	Etc	Total
Number	2	3	3	5	6	19
Time(min)	217	748	1,019	847	1,247	4,078
Availability guarantee measures	Cluster	DB duplication	Cluster	Cluster		

Reduced time(min)	90	150	180	300	1,247	1,967

As a result, the effect of reducing failure time in Company B by using the availability guarantee was predicted to be 51% reduction, which is 4,078 minutes to 1967 minutes.

Company C is a government's public institution and the status of HW disability in the Information Integration Computer Center.
Most of the HW failures are preventable or minimize failures by applying availability guarantee measures. The guarantee measures by types are shown in Table 6. Of course, depending on the importance of service, whether or not to apply the availability guarantee measure requires sufficient prior review and consultation with the customer, but as shown in Table 6, the availability measure is not implemented only through high-cost investment.

Table 6. Effect of Government Institution C's availability guarantee measures

	Memory	Disk	CPU	SCSI	NIC	Etc	Total
Number	4	10	4	4	4	6	32
Time(min)	23	660	67	242	53	103	1,148
Availability guarantee measures	Cluster	Raid	Cluster	Cluster	IPMP Config.		
Reduced time(min)	23	-	67	242	-	103	435

As shown in Table 6, 43% (14 cases) of the total number of failures can be reduced through Disk Raid configuration and NIC redundancy configuration, and the failure time experienced by customers can be reduced by approximately 62% (713 minutes).

Analysis of the effectiveness in the case of 4 through the application of the availability guarantee technology in 3, as shown in Table 7 confirms that there is an effect of reducing failure or shortening the time of service disconnection.

Table 7. Effect of 3 Company's availability guarantee measures

	Company A	Company B	Company B
Failure Time(min)	3,380	4,078	1,148
Failure Time after Availability Guarantee Measures	2,315	1,967	435
Effect (%)	32	51	62

Therefore, in agreeing with the SLA target level, a service level agreement (SLA) should be made in consideration of the environment for the Infra component.

5. CONCLUSION

This paper presents an evaluation criteria of an information system availability for guaranteeing availability (service target level) from the perspective of the SLA contract and its technical point of view. The factors that hinder the availability of information systems were examined; and the technical measures of network availability guarantee, system availability guarantee, and data and database availability guarantee were considered as technical measures to guarantee information system availability. In addition, periodic verification and test techniques are required to ensure stable operations to guarantee availability. In order to verify the effectiveness for information system failure and availability guarantee measures, three cases were examined. In summary, the failure time was reduced by 32% ~ 62% after applying the availability guarantee measure, verifying the excellence in the evaluation of an information system availability.

REFERENCES

[1] Ahmed Mateen, Qingsheng Zhu, Salman Afsar , Akmal Rehan , Imran Mumtaz and Wasi Ahmad, "Access control model, cloud storage, public cloud architecture, network security," International Journal of Advanced Culture Technology(IJACT), Vol. 7, No.4, pp.208-226, 2019. https://DOI 10.17703/IJACT.2019.7.4.208

[2] Ministry of Information and Communication, Information System Fault Management Guidelines, 2005

[3] C. H. Kim, and K.H. Kook, "Implementation Scheme of the High Availability System for the Stable Service," *Proceeding of Korean Management Sciences Society Conference*, pp. 172-177, 2009

[4] Miyata Hiroshi, *Network Design Pattern Guide for Infra/Noetwork Engineer*, Jeipub, pp.53-79, 2017

[5] Goodus, Knowhow of Efficient System Management, March 2018 *http://blog.goodus.com/23?category=279386*

[6] E. H. Chun, A study on the types and minimization of obstacles through the management of obstacles in the information system of government agencies, Master's Thesis, Graduate School of Industry, Kangwon National University, Chuncheon, Korea, 2009

[7] Quest Software, Securing database high availability through data replication:Priceline's Sharefles use case, 2018.9