Evaluating Disaster Recovery Plans Using the Cloud

Omar H. Alhazmi, Ph. D., Taibah University
Yashwant K. Malaiya, Ph. D., Colorado State University

Key Words: Cloud, Disaster Recovery, Risk Analysis and Management, RPO, RTO

SUMMARY & CONCLUSIONS

Every organization requires a business continuity plan (BCP) or disaster recovery plan (DRP) which falls within cost constraints while achieving the target recovery requirements in terms of recovery time objective (RTO) and recovery point objective (RPO). The organizations must identify the likely events that can cause disasters and evaluate their impact. They need to set the objectives clearly, evaluate feasible disaster recovery plans to choose the DRP that would be optimal. The paper examines tradeoffs involved and presents guidelines for choosing among the disaster recovery options. The optimal disaster recovery planning should take into consideration the key parameters including the initial cost, the cost of data transfers, and the cost of data storage. The organization data needs and its disaster recovery objectives need to be considered. To evaluate the risk, the types of disaster (natural or human-caused) need to be identified. The probability of a disaster occurrence needs to be assessed along with the costs of corresponding failures. An appropriate approach for the cost evaluation needs to be determined to allow a quantitative assessment of currently active disaster recovery plans (DRP) in terms of the time need to restore the service (associated with RTO) and possible loss of data (associated with RPO). This can guide future development of the plan and maintenance of the DRP. Such a quantitative approach would also allow CIOs to compare applicable DRP solutions.

1 INTRODUCTION

Many large and small businesses today rely on an internet presence. Continuity is a vital requirement of most businesses, as a sudden service disruption can directly impact business objectives causing significant losses in terms of revenue, business reputation and losses of market share. Indeed, some organizations may find it difficult to survive a serious disaster [1]. The causes of disasters can either be unintended events such as power failure or intentional such as a denial of service attack (DoS). Consequently, an organization must have a business continuity plan (BCP) or disaster recovery plan (DRP) which is executable, testable, scalable and maintainable. Such a plan must satisfy cost constraints while achieving the target recovery objectives; that is, recovery time objective (RTO) and recovery point objective (RPO) [2]. The organizations involved must identify likely events that can cause disasters and evaluate their impact. Organizations need to set the objectives clearly, and evaluate feasible disaster recovery plans to choose the DRP that would be optimal.

Many smaller organizations may find it difficult to afford a desirable disaster recovery plans. Hence, some may choose to have only periodic data backups. This is due to the fact that traditional disaster recovery plans often depend on having two identical sites: a primary and a secondary site, which may be located at some distance. Unfortunately, having two sites will add significantly to IT cost for a disaster that is likely to occur only rarely and therefore may seem like unjustified overhead. This may explain why around 40-50% of small businesses have no DRP and no current future plans to have one [3].

Fortunately, the cloud computing technology that has emerged recently which provide an affordable alternative to traditional DRPs for small or medium sized businesses, with minimal startup cost and with no significant addition to staffing and office space costs [4, 5]. Public cloud services generally employ a "pay-for-what is used" model which can make the secondary site on the cloud very cost effective. Much of the cost is divided among the many users of public cloud services, who may actually use these services only occasionally.

The paper examines tradeoffs involved in choosing among various disaster recovery options. Optimal disaster recovery planning should take into consideration the key parameters including the initial cost, the cost of data transfers, the cost of data storage and software licensing fees. The organization's data needs and its disaster recovery objectives also need to be considered.

To evaluate the risk, the types of disaster (natural or man-made) need to be identified. The probability of a disaster occurrence needs to be assessed together with the cost of concomitant failures. An appropriate cost function needs to be defined to allow a quantitative assessment of currently active disaster recovery plans (DRP) in terms of time needed to restore service (associated with RTO) and possible loss of data (associated with RPO). This work presents guidelines for cost analysis of backup options using cost functions which can be used for the development of the plan and maintenance of the DRP. Such a quantitative approach will allow CIOs to compare applicable DRP solutions. For example, CIOs can decide whether a cloud based DRP will be more cost effective

978-1-4673-4711-2/13/$31.00 ©2013 IEEE

than a traditional DRP and what other choices need to be made to meet operational objectives.

Cloud solution can range from Infrastructure as a service (IaaS) which provides a remote infrastructure, to Platform as a service (PaaS) and up to Software as a service (SaaS) which provides the highest service level. There are multiple degrees of readiness that can be implemented in a backup system. They are often termed cold, warm or hot. The hot backups allow the quickest failover switching, but are the most expensive. RTO values achievable range is from a few minutes to a few days. RPO values range from a few minutes to several hours, again depending on specific implementation [5].

Wiboonratr and Kosavisutte have worked on optimizing DRPs as they suggest dividing systems into small components with various criticality levels and by prioritizing critical parts of a system over less critical components [5]. At this point, there is enough field insight to permit formulation of analytical models. However, precise determination of the parameter values may not always be possible because the available data is limited and is sometimes merely anecdotal [8, 9]. The paper will examine issues related to the estimation of parameter values. However, even in the absence of precise parameter values, it is possible to develop quantitative methods to evaluate, enhance and optimize DRPs.

The issue of cloud security is quite controversial. It has been suggested that the fact of not having full control over data, and having the data stored on some public servers shared with others can compromise and degrade security [6]. However, there is evidence to suggest that when the right measures and policies are in place, cloud computing can be fairly secure especially when small and medium organizations lack the appropriate security experience [7]. This paper will also address the question of incorporating security risks into the cost model.

In the next section we consider background information concerning the DRP problem. In section 3 we consider quantitative evaluation of possible DRP schemes. In the next section, we consider use of a cloud based site as a backup or secondary, followed by some observations.

2 BACKGROUND

A key concept in a DRP is the geographical separation of the primary and backup sites. A significant fraction of disasters, including those caused by outages are geographical in nature as shown in Table 1, which gives the fraction of organizations that have faced disasters during the past five years [12].

When active processing of incoming transactions is switched from the failed primary to the backup site, the switch is termed a *failover*. When the causes of the primary failure have been addressed and the switch is made back to the primary, the switch is termed a *failback*.

A number of options arise depending on the nature of the backup site and how it links to the process at the primary site. The backup site is often described as follows.

Cold standby: Recovery in such a case requires hardware, operating system and application installation. Thus recovery can take multiple days.

Hot standby: This requires a second data center that can provide availability within seconds or minutes. A hot site can take over processing while the primary site is down. A complete copy of the primary process may sometimes exist at the backup, with no need to install either the OS or the application.

Warm standby: A compromise between a hot and a cold site.

Cause	Organizations
System upgrades	72%
Power outage/failure/issues	70%
Fire	69%
Configuration change management	64%
Cyber attacks	63%
Malicious employees	63%
Data leakage/loss	63%
Flood	48%
Hurricane	46%
Earthquake	46%
Tornado	46%
Terrorism	45%
Tsunami	44%
Volcano	42%
War	42%
Others	1%

Table 1: Disasters faced in a 5 year period [12]

It should be noted that the terms “hot” and “warm” are sometimes defined differently. IBM Tivoli documentation refers to highest standby level as “mirrored” [11].

2.1 Architectural options

The available range of data recovery options is often described in terms of tiers. Table 2 gives the tiers as described by Wiboonratr and Kosavisutte [5]. Unfortunately the tier levels are not standard, they can be defined differently [11, 13] and are likely to get redefined as technology progresses. At the highest tier, the backup site can take over almost immediately. This is achieved by having a mirror of the process data at the backup and a high degree of automation for failover.

In Table 2, Tier levels 1 and 2 represent cold standby and levels 5 to 7 represent hot standby implementations [5].

Recovery time objective (RTO) and recovery point objective (RPO) are the main objectives that need to be satisfied criteria when evaluating the optimal solution with a given overall cost.

RTO: The time during which business functions is unavailable and must be restored (includes time before disaster is declared and time to perform tasks). RTO depends on the tasks needed to restore the transaction handling

capabilities at the backup server. While a few days may be required when tape backups are used, the time may be less than a minute in advanced implementations.

Tier	Description	RTO	RPO
1	Point in time tape backup	2-7 days	2-24 hrs.
2	Tape backup to remote site	1-3 days	2-24 hrs.
3	Disk point in time copy	2-24 hrs.	2-24 hrs.
4	Remote logging	12-24 hrs.	5-30 min
5	Concurrent ReEx	1-12 hrs.	5-10 min
6	Mirrored data	1-4 hrs.	0-5 min
7	Mirrored data with failover	0-60 min	0-5 min

Table 2: Recovery levels [5]

RPO: The duration between two successive backups, and hence the maximum amount of data that can be lost when restoration is successful. Historically the maximum value has been 24 hours. If the backup is a synchronous mirrored system, RPO is effectively zero.

3 EVALUATION OF DRP SCHEMES

Here we examine the factors that need to be considered to evaluate the system cost, assuming that the year is used as the period for computing costs. The total annual system cost C_T is the sum of the initial cost C_i (amortized annually), ongoing cost C_o plus the expected annual cost of potential disasters $C_{d\_}$

$$C_T = C_i + C_o + C_d \quad (1)$$

The ongoing cost C_o is the sum of ongoing storage cost C_{os}, data transfer cost C_{ot}, and processing cost C_{op}:

$$C_o = C_{os} + C_{ot} + C_{op} \quad (2)$$

The annual disaster cost is the total expected cost of disaster recoveries plus the cost of unrecoverable disasters. For a disaster type i, let the probably of disaster occurrence be p_i, and the let two costs be C_{ri} and C_{ui}. Then

$$C_d = \sum_i p_i \quad (C_{ri} + C_{ui}) \quad (3)$$

Note that the recovery cost includes the cost of using the backup after the failover and the cost of lost transactions. The cost of lost transactions is proportional to the RTO duration. The loss of reputation also should be considered.

Some disaster frequency related data (such as in that Table 1) is available. However, it needs to be analyzed to develop a model. The geographical correlation factor needs to be modeled to determine potential statistical correlation between primary and backup failures.

Table 3 below gives some revenue loss values obtained in 2000 as an illustration [15]. Actual values would need to be estimated for a specific organization.

RPO is the time between two successive backups. It is an implementation dependent variable. Its optimal value would depend on the overhead represented by a data backup [15], however it may be determined based on scheme used.

RTO determines the length of the period during which the system is not available for incoming transactions. It depends on the factors that impact the DRP tier level used. Let the delays for the backup be as follows.

T1 = hardware set-up/initiation time

T2 = OS initiation time

T3 = Application initiation time

T4 = data/process state restoration time

T5 = readiness verification time + IP switching time

RTO would depend mainly on the readiness the backup site. At the minimum, it would include T5. For a site that starts out completely cold, all of T1 to T5 would be required.

$$RTO = fraction \quad of \quad RPO + \sum_{j\,\min}^{5} Tj \quad (4)$$

Where *jmin* depends on the service readiness of the backup. The fraction of RPO represents computation lost since the last backup.

Industry Sector	Revenue/ Hour	Revenue/ Employee-Hour
Energy	$2,817,846	$569.20
Telecommunications	2,066,245	186.98
Manufacturing	1,610,654	134.24
Financial institutions	1,495,134	1,079.89
Information technology	1,344,461	184.03
Insurance	1,202,444	370.92
Retail	1,107,274	244.37
Pharmaceuticals	1,082,252	167.53
Banking	996,802	130.52
Food/beverage processing	804,192	153.1
Consumer products	785,719	127.98
Chemicals	704,101	194.53
Transportation	668,586	107.78
Utilities	643,250	380.94
Health care	636,030	142.58
Metals/natural resources	580,588	153.11
IT professional services	532,510	99.59
Electronics	477,366	74.48
Construction and engineering	389,601	216.18
Media	340,432	119.74
Hospitality and travel	330,654	38.62

Table 3: Industry specific revenue loss (2000) [15]

In a cloud-based backup a virtual hardware and a specific OS may become available in a minimal amount of time when needed.

A dedicated backup must possess the processing capability that will be needed during a disaster. On a shared cloud, the reserve processing capability is cost shared by multiple applications belonging to diverse organizations.

Having a backup storage/server will address some security issues such as a denial-of-service attack (since a backup server may be available), and compromised integrity of data (restoring data using backup). Appropriate security mechanisms, as dictated by the specifications, need to deploy

to protect the servers from confidentiality breaches resulting from intrusions. Potentially public clouds can achieve a significant degree of security at a lower cost because of the economy of scale. Public cloud service providers can afford more personnel having expertise in security who can monitor vulnerability discovery trends and apply patches or wraps more quickly. However, the impact of any potential cloud-specific vulnerability remains to be determined. Consequently, a quantitative modeling will have to wait until there is sufficient data.

3.1 Optimization

Some of the key variables that impact the cost and performance, and hence the optimality of a system are the following.

Geographical separation: Wider separation would ensure that the backup is relatively immune to a disaster impacting the primary. However separation would add delays, increase transmission costs and render the implementation more complex.

Tier level: A higher tier level would exponentially reduce RTO. However, the cost would increase than linearly as RTO drops.

Architecture and technology: Using more efficient architectures and technologies that permit faster information transfer and process establishment would reduce RTO.

Server reliability: If the primary system has higher reliability, disaster recovery will be invoked less frequently, thereby altering the degree of usage of the backup.

To evaluate performance we can look at the main metrics of disaster recovery RPO (Recovery Point Objective) and RTO (Recovery Time Objective). The ultimate aim would be to minimize the overall cost which include the cost of recovery and lost data.

Recovery Time is the time required for the system to recover to an acceptable level. While, RTO is a widely accepted measurement for a required disaster recovery solution, it is essentially based on business requirements. These requirements may vary significantly because some businesses can tolerate hours of lost operations while others limits may be far less. Hence, for any DRP an appropriate RTO must be set and the system must be designed to meet this requirement. However, surveying the seven tiers of disaster recovery (see Table 2); if the desired RTO is low as in tier 1 which is DRP with tape backup: the tapes would have to be brought, operating systems and their applications installed, data restored, tested and the new recovered system should operate normally. Alternatively, if tier 7 is the objective then the system already have a duplicate real time mirrored system running in parallel; therefore, only switching time can be considered as an RTO. Therefore, RTO relies on the readiness of the alternative system to take over safely.

We can look abstractly at this factor as frequency of backup (f_B) in a period of time. Therefore:

$$RPO\alpha \frac{1}{f_B}, \tag{5}$$

Hence, RPO can be defined by using the frequency of backups. Also, the frequency of backups can depend on factors such as bandwidth and the size of data. Here, also for simplicity we can assume that backup reliability is 1.

The feasibility of choosing the right RPO and RTO is basically determined by estimating the cost of a disaster. If we assume that a disaster will cost (C_d); then we should estimate the number of disasters in the lifetime of the system and compare it with the cost of the disaster plan using Equation 4. Therefore, the optimal RTO and RPO can be estimated and put as requirement of the DRP.

3.2 Comparing cloud-based DRP with alternatives

Cloud computing has been suggested as the new disaster recovery solution, with low startup cost and dynamic scalability using the pay-for-what-you-use model; and it is clear that cloud computing can be a very cost effective option for disaster recovery, [10]. At the same time, the control and security of a cloud-based server can be a concern if critical data is stored outside the organization jurisdiction.

The backup system can be on-site, at a remote colocation site (colo), or implemented using the cloud services of a vendor, such as amazon web services.

An exact comparison among the options available is not possible because there is a range of prices that can apply to each option depending on various factors. For example, Amazon web services offer processes for different instances (depending on memory, CPU/GPU or I/O requirements) and whether the resource is pre-reserved or on-demand. However, it is easy to see that if the disaster frequency is low (as given in p_i in Equation 3); the backup would rarely be needed. Hence, for a cloud server which is rarely fully deployed, the cost would be very low based on use-based pricing. The cloud service provider can host a number of clients as long as they only require significant computing and I/O power randomly, allowing for efficient multiplexing [2, 17].

A colo or cloud server may also enjoy significant economy of scale. Not only are the physical site and infrastructure shared, but the maintenance/personnel cost may be significantly lower on a per customer basis. Table 4 gives an approximate comparison of the alternatives.

Option	Data Synchro-nization	Statistical Indep-endence	C_i	C_o	C_d
On-site	High	Low	High	d	High
Colo	Medium	High	Medium	d	High
Cloud	Low	High	Low	d	Low

Table 4: The three backup options

A backup server at the same location would allow a high degree of synchronization because the speed-of-light limitation would not arise as long as the distance is only a few miles. On the other hand, such a server would have a high probability of being impacted by a geographical-type of disaster. The on-going costs C_o may be lower for colo and cloud options but in general may depend on various factors.

A cloud has potential limitations that may be encountered in rare situations. It is possible that a cloud site may serve as a disaster recovery site for a number of customers from the same region. Thus, it may be overwhelmed if it encounters a sudden high demand from many customers. While a cloud service provider guarantees its capacity for reserved usage, it does not guarantee that sufficient computational resources will be available for all on-demand usage. Another potential limitation is that a cloud may be subject to some unknown cloud-specific vulnerabilities or attack/sabotage.

While several examples of cloud outage are known, there is not enough data to judge if cloud servers are less reliable. It is likely that economy of scale would permit cloud vendors to invest more aggressively in achieving higher reliability.

Above we have considered the alternatives for a backup server, assuming that the primary server is a locally owned system. Actually, the three alternatives apply to the primary server.

Cloud-based primary server may be cost effective in many cases, especially when the commitment is for a relatively short term. For disaster recovery, such a server also needs to be backed up. Consequently, it would make sense to ensure that the backup server is located in a different region.

The cloud-based computing has a limited history. It remains to be seen whether cloud based systems are susceptible to threats that are applicable specifically to clouds in certain rare situations. If any such threats are eventually identified, it would make sense to use a more conventional server that is under the exclusive control of an organization as a backup. Such a choice would perhaps not be justifiable in terms of costs, when the normal operation and failover in the case of relatively more common disasters is considered.

4 DISCUSSION

At this point in time there is not enough data to completely construct analytical models to determine optimal implementation. However, the discussions in this paper can serve as a guide for evaluating available alternatives. Initially, an application needs to be studied to develop specifications in terms of computational requirements (processing, memory, I/O) and RTO. Both common as well as relatively rare disasters need to be considered in order to estimate their impact. RPO may depend on the nature of the arriving transactions.

Some cloud service providers provide calculators or pricing guides that permit estimation of costs. There exists some literature that provides examples of such computations [2, 17]. Several feasible alternatives should be identified and evaluated.

There is need to collect enough data to permit the development of construction models that can eventually allow the problem to be set up as a mathematical optimization. These include the relationship between geographical distance and statistical correlation between failures in the primary and secondary servers. A model relating RTO and cost can potentially be developed. Some of the literature speculates that there may be a non-linear relationship between cost and RTO [6].

REFERENCES

1. Disaster Recovery for Small Business, Technical White Paper, Iomega Corporation, March 18, 2009.
2. T. Wood, E Cecchet, K. K. Ramakrishnan, P. Shenoy, J. van der Merwe, and A. Venkataramani, " Disaster recovery as a cloud service: economic benefits & deployment challenges", Proc. 2nd USENIX Conference on Hot topics in cloud computing (HotCloud'10), Berkeley, CA, USA, 2010, pp. 8-8.
3. Survey Indicates Half of SMBs Have No Disaster Recovery Plan, Chris Preimesberger, September 2009, http://www.eweek.com/c/a/Data-Storage/Survey-Indicates-Half-of-SMBs-Have-No-Disaster-Recovery-Plan-687524.
4. Manish Pokharel, Seulki Lee, Jong Sou Park, "Disaster Recovery for Systems Architecture Using Cloud Computing,", IEEE/IPSJ Int. Symp. Applications and the Internet, 2010, pp. 304-307.
5. Montri Wiboonratr and Kitti Kosavisutte, "Optimal strategic decision for disaster recovery," Int. Journal of Management Science and Engineering Management, Vol. 4 (2009) No. 4, pp. 260-269.
6. Vic Winkler, "Cloud Computing: Virtual Cloud Security Concerns". Technet Magazine, Microsoft. http://technet.microsoft.com/en-us/magazine/hh641415.aspx. Retrieved 12 February 2012.
7. Jason Bloomberg, "Why Public Clouds are More Secure than Private Clouds," February 7, 2012 http://www.zapthink.com/2012/02/07/why-public-clouds-are-more-secure-than-private-clouds/.
8. Steve Lohr, "Amazon's Trouble Raises Cloud Computing Doubts," New York Times, April 23, 2011, B1.
9. Mike Klein, "How the Cloud Changes Disaster Recovery," Industry Perspectives, July 26th, 2011.
10. Glen Robinson, Ianni Vamvadelis, Attila Narin, Using Amazon Web Services for Disaster Recovery, http://media.amazonwebservices.com/AWS_Disaster_Recovery.pdf, January 2012.
11. Disaster Recovery Strategies with Tivoli Storage Management, C. Brooks, M. Bedernjak, I. Juran, J. Merryman, IBM/Redbooks, November 2002, http://www.redbooks.ibm.com/redbooks/pdfs/sg246844.pdf
12. Symantec 2010 Disaster Recovery Study, Global results, CA, USA, November 2010.
13. Disaster Recovery Issues and Solutions, A White Paper, By Roselinda R. Schulman, Hitachi Data Systems, September 2004.
14. K. M. Chandy, J. C. Browne, C. W. Dissly, and W. R. Uhrig, "Analytic Models for Rollback and Recovery Strategies in Data Base Systems," *IEEE Transactions on Software Engineering,* Vol. SE-1, pp. 100–110, March 1975.

15. Disaster Recover/Business Continuity, N-1 Technologies, http://www.n-1technologies.com/recovery_continuity.html
16. IT Performance Engineering & Measurement Strategies: Quantifying Performance Loss, Meta Group, October 2000.
17. The Total Cost of (Non) Ownership of a NoSQL Database Cloud Service, Jinesh Varia and Jose Papo, March 2012, http://media.amazonwebservices.com/AWS_TCO_DynamoDB.pdf

BIOGRAPHIES

Omar H. Alhazmi, Ph. D.
Taibah University
Computer Science Department
College of Computer Science and Engineering
Taibah University
Medina, Saudi Arabia.

E-mail: ohhazmi@taibahu.edu.sa

Dr. Omar H. Alhazmi is currently an assistant professor in the Department of Computer Science in Taibah University in Medina, Saudi Arabia. He received his MS in Computer Science from Villanova University in 2001 and his Ph.D. in Computer Science from Colorado State University in 2007. His interests include computer security, cloud computing and software reliability. He has published over 13 papers in scientific journals and conference proceedings.

Yashwant K. Malaiya, Ph.D.
Computer Science Department
Colorado State University
Fort Collins, CO 80523-1873 USA.

E-mail: malaiya@cs.colostate.edu

Yashwant K. Malaiya is a Professor in the Computer Science Department at Colorado State University. He received his MS in Physics from Sagar University, MScTech in Electronics from BITS Pilani and PhD in Electrical Engineering from Utah State University. He has published widely in the areas of fault modeling, software and hardware reliability, testing and quantitative security. He has also been a consultant to industry. He was the General Chair of 2003 IEEE International Symposium on Software Reliability Engineering. He is a senior member of IEEE.