Volume 5, Issue 3, March 2015 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Performance and Accessibility Analysis of IT Systems by Configuring Periodic Online Data Backup

D. Vijaya Shanthi*
M.Tech student, Dept of CSE
SRM University, Chennai, India

R. B. Sarooraj
A.P, Dept of CSE
SRM University, Chennai, India

Abstract*— In recent IT system, data backup plays a vital role in data protection. The data backup is resource intensive and lead to performance degradation. It is significant to select the suitable backup and restore technique to prevent data loss. In this paper, a framework is developed to evaluate the impacts of backup and restore operations on performance and system availabilities. The work is carried out to assist the system engineers to design effective backup and restore operations. The different types of backup strategies are discussed. In this work, different modelling approaches are discussed for evaluating the availability and performance of a storage system with online periodic data backup.

Keywords*— Backup and restore strategies, online backup, metrics, modelling methods, hourly backup.

I. INTRODUCTION

The system data are the most precious assets of an organization. In today's business environment, the data backup plays an essential role in the IT system data protection. There are many chances of loosing data due to technical reasons, hardware failure, software failure, manual errors etc. It is important to identify the proper backup and restore technique to make ensure the data is protected. In [1], authors presented an analytical modelling approach for the data backup. They investigated metrics considering system availability, data loss and rejection of user requests. In [2] author, Karel derived a mathematical model to evaluate the different types of backup strategies quantitatively. The full, differential and incremental backup strategies, formulae are derived for the calculation of the average total backup size and the average data recovery size. In this paper [3] the impacts of different backup policies on availability measures were studied. The backup and restore operations are designed using SysML to compute the availability measures. The combination of full backup and partial backup was effective in terms of user-perceived data availability and data loss rate. The authors in [4] advocated the replacement of servers by a cloud of residential gateways. They evaluated using statistical distributions based on real world traces, as well as a trace of residential gateways for availability and the results show that the time required to backup data in the network drops from days to a few hours. The design for efficient backup scheduling is discussed in [5]. In this work, each backup job is characterized via two metrics, called job duration and job throughput. The goal is to automate the design of a backup schedule that minimizes the overall completion time. In [6], authors considered two schemes of full and incremental backup for a database system. They have analytically discussed optimal backup policies to minimize the cost, using theory of cumulative processes. In [7] authors conducted a comprehensive performance analysis for AR-based optical networks. They proposed a novel analytical framework for modelling the restoration probability of a connection to the possible correlation among the multiple backup routes.

II. BACKUP SCENARIOS

A. Backuptypes

The data loss can occur in the primary storage system due to reasons such as hardware failure, software corruption, human error, virus infection, etc. The data restore operation is required to recover the data from the backup storage when the data is lost and it cannot be recovered via normal means. The backup approaches based on data to be backed up are classified into two main types, full backup and partial backup. The partial backup can be further divided as incremental and differential backup. The Incremental backup, copies the files added or modified since the last full or partial backup, whereas in differential backup copies files added or modified since the last full backup. The time to take a backup and the time to restore data are dependent on the backup type, it is important to determine the type of backup and its frequency in consideration with system availability and performance. In this paper different types of backup techniques are analyzed and suggested optimal technique for a particular application.

B. Online Backup

In this work, the performance of online backup is analysed. The system considered is an Apache web server connected to a backup server through the network server. In online backup, it permits the user to access the system while backup is in progress. This feature has several implications for both the Apache and the backup process. The Apache and the backup process share the resources CPU, Network I/O and disk on the file server which are shown in fig.1. The disk on the file server assumes, the file server has a single network I/O. It can handle both the backup data and user accesses. In reality both the processes are running simultaneously during an online backup. It has to be analysed the effects over each other. The two system configurations, shared and no priority and priority are considered in this study.

III. SYSTEM DESCRIPTION

The evaluation of online backup and restore operations to a storage system, an Apache web server connected to a backup server through network server is considered. The backup is assumed to be performed over a dedicated network and the backup data is stored in backup storage system attached to another server in the same network. The architecture diagram of the system description is shown in fig1. The system is a web service system which provides access to the files stored on a local file server. It also contains a backup server to perform the backup operations. The two servers are connected using a dedicated backup network. The outside users access the file server through the Apache web server. The system periodically backs up data from the file server disk to the backup server using the rsync tool

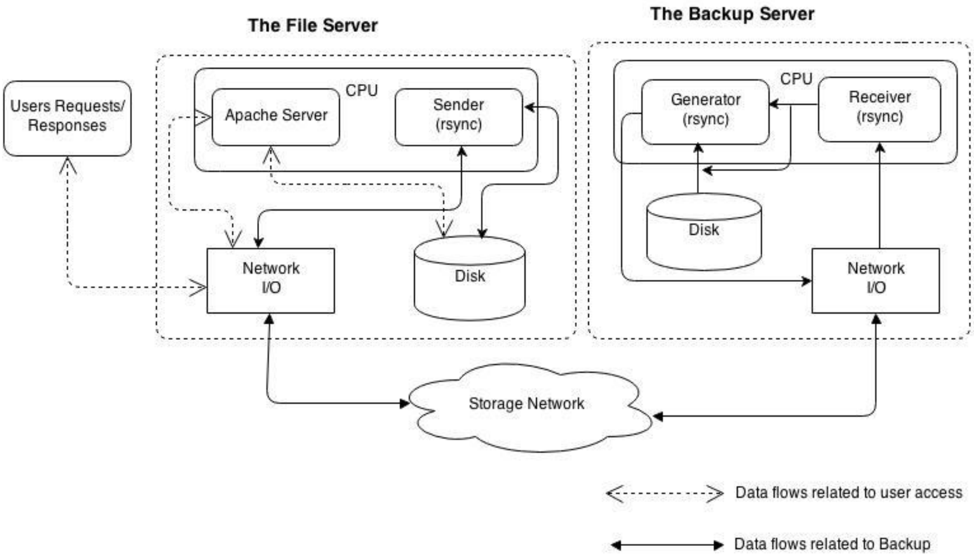

Fig.1 Online backup architecture

A. System Flow Process

The assumptions considered for the analysis is a service process handles one request at a time. The Apache server can only handle a finite number of user requests simultaneously and user requests that arrive when all service processes are busy will be rejected. The request once comes it passes the file server network I/O and to the main Apache process. If there are fewer active service processes than the limited, the main Apache process generates a new service process to handle the request else it rejects the request. The request seeking file items are, in the file server memory. The service process directly forms a response and replies to the user through the file server network I/O. If the requested file is not found within the file server memory, an access to the file server disk is initiated by the service process and sends response, through the network I/O to the user. The user flow process diagram is shown in fig.2.

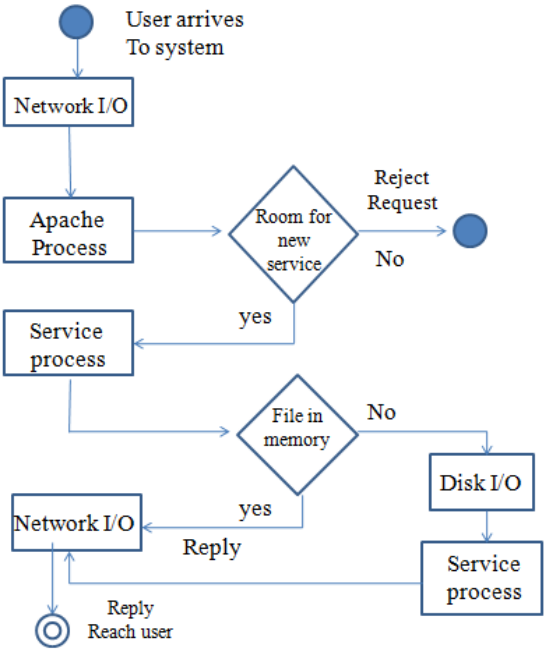

Fig.2 User flow

B. Backup flow Process

The Rsync tool consists of three processes running on the backup servers are sender running on file server, generator and receiver on the backup server. The backup begins with sender who build list of files and send to the generator on the backup server. The generator compares the file list in local and decides the files needed from the file server. The generator walks the list of files to be transferred and send the file name to the sender. In full backup, the sender directly send the whole requested file to the receiver otherwise it will perform checksum computations and send only the file blocks whose checksum values are different from those provided by the generator.

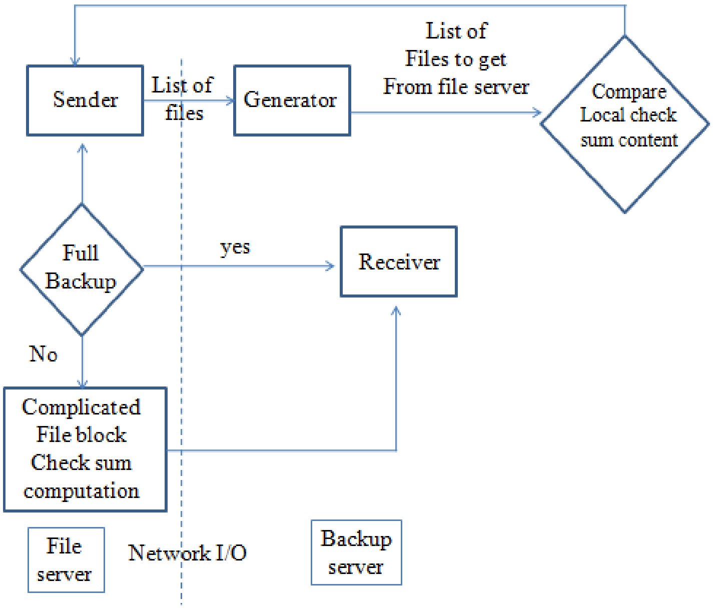
Sender
List of files
Generator
List of Files to get From file server
Compare Local check sum content
Full Backup
yes
Receiver
No
Complicated File block Check sum computation
File server
Network I/O
Backup server

Fig.3 Backup flow

IV. SYSTEM METRICS

The metrics to be developed to obtain the operational details of the system under different backup techniques and workload conditions. The metrics considered for the analysis are file service availability, data loss rate, data loss ratio, and request rejection ratio. The metrics file service availability, and data loss rate are related to system performance. The metrics data loss ratio, and request rejection ratio represent system availability perceived by users. The availability model can be developed using Stochastic reward nets(SRN) to capture the system failure/restore/backup behaviors and yield system availability and data loss metrics SRN and performance model with markov chain and PEAP to capture resource contention during different system operation periods and provide the corresponding task rejection probabilities.

A. Data Loss Rate and Loss Ratio

These metrics correspond to the number of user updates to local data that are lost due to storage failure, before the system back up. The request arrival model in a day is divided into two periods, office hours and after-office hours. The users requests are classified into write and read types, and distinguish between request rejection and data loss. It is assumed each backup covers all the data updates accumulated in a certain time period before the beginning of the backup. In online backup, it starts from the beginning of the last backup, since new requests may be processed while the previous backup is underway, and assumed the data updates from these new requests are not processed by the previous backup. The backup server does not fail and once a data undergoes backup it will not be affected by any failure in the future.

B. Data Rejection Rate and Rejection Ratio

These metrics represent how often the system turns down incoming user requests, either due to the system failure or due to collapse of resources means no available service processes. The user would consider the system unavailable if the request is rejected, these two metrics capture the system availability as perceived by the users.

V. MATHEMATICAL MODELING TECHNIQUES

This section explores the different mathematical modeling techniques which can be used for the analysis of performance and availability. Thus based on the parameters that have to be measured, type of backup, priority and accuracy of the model the suitable model can be chosen for the backup. The different mathematical models used for performance and availability are discussed. The first step in a model based performance evaluation consists of the formalization process, during which the modeler generates the formal description of the real-world system using conceptualization. Queuing network formalism and Markov chain can be used for this purpose. The former is oriented towards the structure of the real-world system and the later puts the emphasis on the description of the system behavior on the underlying statement space level. The second step is the deduction of performance measures by the application of appropriate solution method. Thus depending on the conceptualization chosen during the formalization process the following modeling methods are applied for the availability.

A. Stochastic Petri Nets

A Petri Net (PN) is a bipartite directed graph with two disjoint sets called places and transitions. The directed arcs in the graph connect places to transitions called input arcs and transitions to places called output arcs. The places may contain an integer number of entities called tokens. The state or condition of the system is associated with the presence or absence of tokens in various places in the net. The condition of the net may enable some transitions to fire. This firing of a transition is the removal of tokens from one or more places in the net and/or the arrival of tokens in one or more places in the net. The tokens are removed from places connected to the transition by an input arc; the tokens arrive in places connected to the transition by an output arc. A marked Petri net is obtained by associating tokens with places. The marking of a PN is the distribution of tokens in the places of the PN.

A marking is represented by a vector M = (#(P1), #(P2), · · · , #(Pn))

where #(Pi) is the number of tokens in place i and n is the number of places in the net.

B. Generalized Stochastic Petri Nets (GSPN)

The extension to SPNs is the development of GSPNs which include the inhibitor arc. An inhibitor arc is an arc from a place to a transition that inhibits the firing of the transition when a token is present in the input place.

C. Stochastic Reward Nets (SRN)

Stochastic reward nets (SRNs) are a superset of GSPNs. The SRN's substantially increase the modeling power of the GSPN by adding guard functions, marking dependent arc multiplicities, general transition priorities, and reward rates at the net level. A guard function is a Boolean function associated with a transition. Whenever the transition satisfies all the input and inhibitor conditions in a marking M, the guard is evaluated. The transition is considered enabled only if the guard function evaluates to true. Marking dependent arc multiplicities allow either the number of tokens required for the transition to be enabled, or the number of tokens removed from the input place, or the number of tokens placed in an output place to be a function of the current marking of the PN. Such arcs are called variable cardinality arcs.

D. Modeling methods for Performance analysis

The first step in a model based performance evaluation consists of the formalization process, during which the modeler generates the formal description of the real-world system using conceptualization. For this purpose Queuing network formalism and Markov chain can be used. The former is oriented towards the structure of the real-world system and the later puts the emphasis on the description of the system behavior on the underlying statement space level.

The second step is the deduction of performance measures by the application of appropriate solution method. Thus depending on the conceptualization chosen during the formalization process the following solution methods are available. In analytical type closed form solutions are available if the system can be described as a simple queuing system, in which the solutions can be expressed analytically in terms of bounded number of well-known operations .The measures can either be computed by ad-hoc programming or with the help of computer available packages such as MATHEMATICA. The advantage of this is moderate computational complexity and enables a fast calculation of performance measures even for larger system. The numerical solutions can be used where closed form solution cannot be used. The approximate solution can be obtained by the numerical methods. The formal system description can be either given as queuing network, stochastic petri nets or another high level modeling formalism, from which a state space representation is generated. In most of the models the analytical method is not feasible because either a theory for the derivation of proper system equation is not known, or the computational complexity is too high. In this situation solutions can be obtained by the application of Discrete Event Simulation (DES). The DES executes the model and collects information about the observed behavior for the subsequent derivation of performance measures.

VI. CONCLUSIONS

The importance of backup and different types of backup and restoring techniques are discussed. The different analytical and simulation modeling techniques that can be used for performance and availability analysis of online data backup are studied. The suitable modeling technique can be selected with necessary assumptions for the online backup from the studied models. As a future work, the developed model will be checked in real time online scenario to validate the performance of the model.

REFERENCES

[1] Xiaoyan Yin, Javier Alonso, Fumio Machida, Ermeson. C. Andrade, Kishor S. Trivedi, " Availability Modeling and Analysis for Data Backup and Restore Operations" IEEE Transactions on Dependable and Secure Computing, Vol. 11, no. 4, July/August 2014.

[2] Karel Burda, "Mathematical Model of Data Backup and Recovery", International Journal of Computer Science and `Network Security, VOL.14 No.7, July 2014.

[3] Xiaoyan Yin, Javier Alonso, Fumio Machida, Ermeson. C. Andrade, Kishor S. Trivedi, "Availability Modeling and Analysis for Data Backup and Restore Operations" Proceedings of the 2012 IEEE 31st IEEE International Symposium on Reliability Distributed Systems, Pages 141-150, IEEE Computer Society Washington, DC, USA 2012.

[4] Anne-Marie Kermarrec , Erwan Le Merrer ,, Nicolas Le Scouarnec, Romaric Ludinard , Patrick Maillé c , Gilles Straub, Alexandre Van Kempen "Performance evaluation of a peer-to-peer backup system using buffering at the edge" Computer Communications 52 (2014) 71–81, 2014.

[5] Ludmila Cherkasova, Alex Zhang, Xiaozhou Li “DP+IP = Design of Efficient Backup Scheduling”, International Conference on Network and Service Management (CNSM), 2010 .

[6] Cunhua Qian and Yingyan Huang, “Optimal Backup Interval for a Database System with Full and Periodic Incremental Backup” Journal of Computers, Vol. 5, no. 4, April 2010.

[7] Mohamed Mostafa A. Azim, Xiaohong Jiang, Pin-Han Ho, Susumu Horiguchi, and Minyi Guo Restoration Probability Modeling for Active Restoration-Based Optical Networks with Correlation among Backup Routes, IEEE transactions on Parallel and Distributed Systems, Vol. 18, no. 11, November 2007.

[8] Kishor s. Trivedi, “Analyses Using Stochastic Reward Nets”, IBM Corporation, Research Triangle P ark, North Carolina, 1995.

[9] Kruti Sharma, Kavita R Singh, “Online Data Back-up and Disaster Recovery Techniques in Cloud Computing: A Review”, International Journal of Engineering and Innovative Technology

[10] (IJEIT) Volume 2, Issue 5, November 2012.

[11] Mohamed Mostafa A. Azim, Xiaohong Jiang, Pin-Han Ho, Susumu Horiguchi, and Minyi Guo, “Restoration Probability Modeling for Active Restoration-Based Optical Networks with Correlation among Backup Routes” IEEE transactions on parallel and distributed systems, vol. 18, no. 11, november 2007.

[12] Lorrie A. Tomek, “Analyses Using Stochastic Reward Nets”, Software Fault T olerance, Edited by Lyu, 1995.

[13] K. Keeton and A. Merchant, “A Framework for Evaluating Storage System Dependability,”Proc. Int’l Conf. Dependable Systems and Networks (DSN’04), pp. 877-886, 2004.

[14] L. Cherkasova, A. Zhang, and X. Li, “DPþIP¼Design of Efficient Backup Scheduling,”Proc. Int’l Conf. Network and Service Management (CNSM), pp. 118-125, 2010.

[15] EMC Backup Advisor, http://www.emc.com/products/detail/ software/backupadvisor.htm, 2013.

[16] Symantec Backup Exec, http://www.symantec.com/backup-exec, 2012.