Association for Information Systems

AIS Electronic Library (AISeL)

ICEB 2015 Proceedings

International Conference on Electronic Business (ICEB)

Winter 12-6-2015

A Study Of The Costs Of Cloud-Based Website Parallel Archiving System

David Chao

Sam Gill

A STUDY OF THE COSTS OF CLOUD-BASED WEBSITE PARALLEL ARCHIVING SYSTEM

David Chao, San Francisco State University, dchao@sfsu.edu
Sam Gill, San Francisco State University, sgill@sfsu.edu

ABSTRACT

Parallel Archiving System supports web applications that are periodically renewed, frequently changed in design and supporting technologies, and are required to keep the previous periods' applications operational in parallel with the current period application to form an easy-to-access archive for historical data. The system implements each period's application with a virtual machine to preserve the technologies and deploys it in a cloud platform. This paper studies the costs of a cloud-based Parallel Archiving System that include the cost of virtual machine, database server, data storage, business transactions and website traffic. This study will help a manager in determining how many of previous periods' applications an organization can afford to run for a given budget.

Keywords: Cloud computing, parallel archiving, website archiving, virtual machine cost.

INTRODUCTION

Many business applications are designed and organized to support business activities for a period of time and to be renewed at the turn of the period akin to the perpetual seasonal change and renewal of nature. This period of time may be a year, such as applications supporting the operation of a fiscal year, or a quarter or a season such as applications supporting a university's semester or quarter. During the operational period, the design of applications such as user interface and supporting technologies are rarely changed to assure smooth operation. Design changes are typically implemented in a revision of the application that supports future periods. When a change of period occurs, the application is reinitialized to support the new period and the application and the data of the previous period become archival.

Very often the applications supporting the previous periods need to be operational continuously even after the application for the new period started. First, they are needed to process incomplete transactions from the previous periods. Two types of updates may occur to data of previous periods: 1. Retroactive corrections: These are corrections to data of previous periods after periods ended. 2. Anticipatory insertions: These are data of the previous periods that have not been entered during those periods. Second, they are needed for informational purposes. Historical data are useful for decision making and their value tends to decrease as it becomes older. The data of the recent periods have higher value and are accessed more often. Allowing users to access historical data through a familiar interface is better than redirecting users to other unfamiliar archive locations and updated data will be available for analysis and decision making. Therefore, it is beneficial for applications supporting previous periods to operate in parallel with applications supporting the current period.

Parallel operation of current and previous periods' applications may be problematic for web-based applications. Due to the rapid change in Internet technologies, websites must keep themselves up-to-date by adopting new technologies. The infrastructure of a dynamic website typically includes the operating system, the web server, the database management system, and the server-side computer language used to create the dynamic pages, collectively known as the "stack". Two examples of such stack are the Microsoft stack with Windows, Internet Information Service (IIS), SQL Server, and a .NET language; and the LAMP stack with Linux, Apache, MySQL, and PHP. It is possible that the infrastructure of the current period application may not be compatible with that of the previous periods.

We proposed a cloud-based parallel archiving scheme in [3] to support the parallel operation of the current period and previous periods' applications. The proposed scheme applies two advancements in information technology: virtualization and cloud computing. Virtualization allows organizations to create IT environments that can respond to dynamically changing demands for computing resources. The proposed scheme supports each period's application in a virtual machine with all the supporting technologies so that the technologies are preserved with the application. The virtual machine is retained for a number of periods set by the organization's policy and eventually retired and its data permanently archived.

The scheme is cloud-based to take advantage of the efficiency that cloud computing has to offer. Cloud computing delivers computing resources "as a service" to clients via Internet [9] and does not require businesses to invest heavily on IT infrastructure out-front. The major benefits of cloud computing includes [4]: 1. Efficiency: Cloud computing allows businesses to rapidly deploy applications due to lower requirements for initial investment on technologies and maintenance of the infrastructure. 2. Agility and innovation: Businesses can react to the business environment faster and test innovative services before full-scale implementation. 3. Cost Savings: Businesses only pay for the computing capabilities they use without purchasing the infrastructure, and can determine the computing capabilities dynamically based on the demand of the applications. 4. Increased scalability: Businesses can rapidly "scale up" their computing capabilities, and rapidly release those services to quickly "scale in."

The proposed parallel archiving scheme uses virtual servers of a cloud platform to run each period's application and its stack in a virtual machine so that previous periods' applications will run in parallel with the current period system and forms an easy-to-access archive for historical data. This scheme is good for applications with the following properties: 1. the applications are periodically renewed, 2. previous periods' applications are required to be operational after new period starts, 3. the applications may change in terms of design and supporting technologies from period to period, but remain unchanged during the period. This paper studies the costs of the cloud-based parallel archiving system, and is organized as follows. Section 2 presents an introduction to the design of the system. Section 3 studies the costs of the proposed system. Section 4 concludes the paper.

PARALLELL ARCHIVING SYSTEM DESIGN

Figure 1 presents an overview of the parallel archiving system. The core of the system is a Virtual Host System consisting of a collection of virtual host servers running on a cloud platform. Each virtual host server operates a collection of virtual web servers. We consider a web server as a system defined by the four components of the stack: the host operating system, O, the web service, S, the database management system, D, and the server side web language, L; the stack remains unchanged in a period. Each web server hosts one website that runs the application of a specific period. So there exists a one-host server/many-web server relationship, and one-web server/one-website relationship. We assume a website, WS, has a life of N periods and will be retired at the end of the Nth period. Let *i* denote the *i*th period since a website starts, then a website may have a remaining life, R, of N- *i* + 1 periods. The parallel archiving system eventually will have N websites operating in parallel, each with N- *i* + 1 periods remaining life where *i* ranges from 1 to N. Websites can be distinguished with these attributes, WS(Stack(O, S, D, L), R).

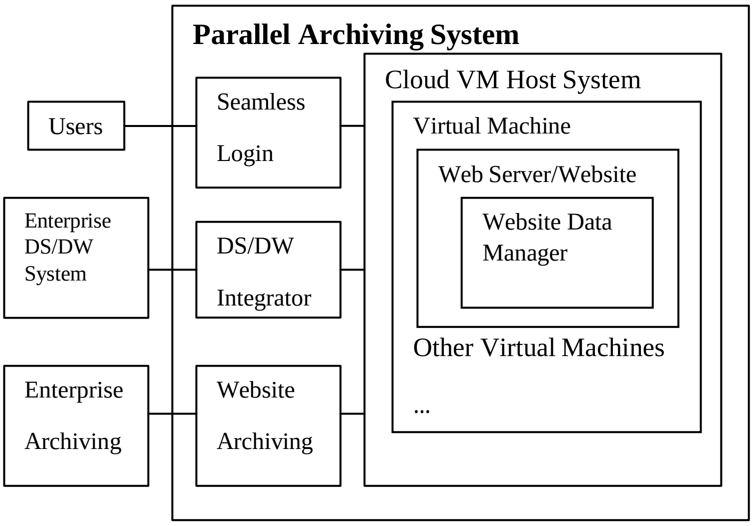

Figure 1. An overview of the Parallel Archiving System

Users of the system initiate business transactions and submit queries related to a specific period. The Seamless Login Module enables transparent login to each website regardless of underlying server structure. It presents users with an easy-access interface where websites are accessible by push buttons. Once in the website, data of business transactions and queries are handled by the Website Data Manager. Since there are N websites operating in parallel, these Website Data Managers are in fact managing the archived data of these N periods.

The function of the DS/DW Integrator is to gather and prepare data for an enterprise's decision support system and data warehouse. Recognizing the value of business intelligence to an enterprise, today's e-Commerce systems typically integrate decision support data acquisition module in the system design [7]. Because of the heterogeneous nature of the virtual machines and websites, the DS/DW Integrator must have the ability to work with heterogeneous data sources.

A website will eventually retire when it completes its N life periods and will be removed from the Parallel Archiving System. The website with its supporting technologies and database is a valuable historical resource of an enterprise. Historical data may be useful in supporting applications that require historical data, such as applications that perform analyses to study certain trends in the study subject, or answering questions about website content in the past for audit and compliance purposes. Websites may also be required to preserve historical data due to government or organizational policies. The Website Archiving Integrator implements the enterprise's policy in archiving websites. One popular practice is periodically creating date-time stamped read-only copies of the website. In cloud computing this can be done by creating the virtual machine snapshots.

An Example of the Virtual Host System

Figure 2 gives an example of the Cloud Virtual Host System assuming the website has been running for four years and renewed every two years with a new stack. The Parallel Archiving System assumes a one-virtual host server/one-website relationship. Each year the website is running on a separate virtual host server. The arguments supporting the one-virtual host server/one-website relationship are: 1. assuring the current operational system running smoothly without interruptions from other activities is the primary objective of an information system. 2. It will ease the management of the parallel archiving system. As discussed earlier, a typical cloud computing platform offers services to take snapshot and create image of a virtual machine. With each period's system running on a separate virtual machine, it will be easier to create snapshot and the image of the virtual machine for that period. 3. There exists unbalanced demand for each period. The demand for the previous periods will decrease as time elapsed. Since a virtual machine can be activated/deactivated by the client of a cloud computing platform, to save the costs of running the parallel archiving system, an organization may activate the previous period systems on an on-demand basis.

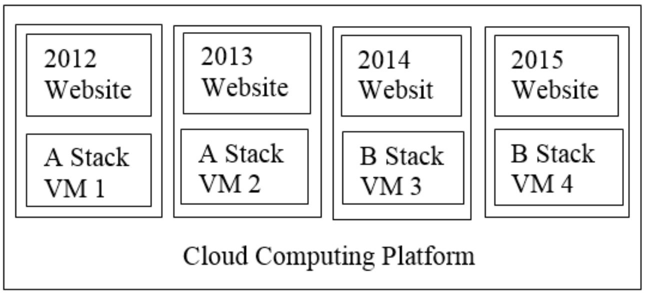

Figure 2. An example of the Cloud VM Host System

COSTS OF PARALLEL ARCHIVING

This section studies the costs of the parallel archiving system. We assume the system is running in a public cloud platform. Examples of such platform include Microsoft Azure [6], Google Cloud [5], Amazon EC2 [1] and many others. These vendors offer virtual machines with these services: 1.They offer an extensive list of pre-configured virtual machine images encompassing a wide range choice of operating systems, database management systems and development technologies. 2. Users are able to create their own virtual machine images with user-specified configurations. 3. Users can upload and deploy their own virtual machine images. 4. Users can activate and deactivate the virtual machines as needed. 5. Users can take snapshots of the virtual machines which is a file-based representation of the state of a virtual machine along with the database at a given time. Virtual machine snapshots help to back up or archive virtual machines. 6. Users can capture an image of a running virtual machine as a template to create other virtual machines. These services let users to customize virtual machines to their requirements, control the costs of deploying the virtual machines and create backups of the virtual machines.

Cloud computing vendors typically adopts a Pay-As-You-Go pricing policy [8]. And they charge services for: 1. Virtual machines, for the time each one is running. 2. Database server. 3. Storage for files. 4 Web traffic between the user's browser and the website. 5. Interaction with the storage system generated by the business transactions. 6. Connectivity between the cloud platform and client's on-premises IT infrastructure. The first three of the charges are fixed costs once the virtual machine is configured and are the charges for the computing stack. And the other charges are variable costs that depend on the volume of business transactions.

There are many factors affecting the amount of computing stack charges. Factors relevant to the parallel archiving including:
1. Virtual Machine processing capacity. For example, Microsoft Azure offers a variety of virtual machines including A-series general purpose virtual machines, D-series virtual machines feature solid state drives and faster processors than the A-series, and G-series virtual machines feature faster processor, more memory and four times more Solid State Drive storage than the D-series. The average price of the D-series is about 1.7 times more expensive than the A-series, and the average price of the G-series is about 5.45 times more expensive than the D-series.

2. Database server capacity. For example, Microsoft Azure offers SQL Server Standard tier with mid-level performance, and Premium tier designed for mission-critical databases. The average price of the Premium tier is about 41.4 times more expensive than the Standard tier.

3. The support of scalability. Cloud platforms support scalable applications that can rapidly "scale up" their computing capabilities, and rapidly release those services to quickly "scale in." One such example is the Cloud Service offered by Microsoft Azure [6]. A cloud service is a multi-tier web application in Azure, consisting of Web Roles which are dedicated virtual machines for hosting front-end web applications, and Worker Roles which are dedicated virtual machines that run time-consuming tasks sent from a web role asynchronously in the back-end. And the number of Web Roles and Worker Roles can be configured dynamically based on the business needs.

4. Decreasing demand for aged websites. The need for the aged websites to be online will decrease and hence do not need to be online at all time. Cloud platform vendors let users to activate and deactivate the virtual machines as needed. A further cost reduction can be achieved by running aged websites intermittently only at a fraction of time such as daily from 9 to 5 or weekly on certain weekdays.

Determining The Length of A Website's Life

The Parallel Archiving System assumes a website has a life of N periods and the system is running N periods of websites in parallel. The actual value of N may be influenced by many factors such as an organization's policy in accepting delayed changes and the budget to support the parallel historical websites. In the following analysis we study the impact of budget to the value of N by comparing the budget to run the current period website and previous N-1 historical websites.

The current period website is considered to be the mission-critical application. Assuming the application is scalable that will dynamically scale up or down between a preset value of Max/Min number of virtual machines. Let p_i be the probability that i virtual machines are used to support the current period website where Min<=i<=Max, then the expected number of virtual machines used is:

$$E=\sum_{i=Min}^{Max} i * pi \qquad (1)$$

A manager may study the website's operational history to get an estimate for p_i. In this study we assume a simplified scenario where all p_is are assumed to be equal to illustrate our analysis. Under this scenario, the E= (Max+Min)/2. The demand for each historical website will decrease and is small enough to use only the Min number of machines to support. Let *f* be the fraction of the cost to run a historical website relative to the cost to run the current website, then *f* is:

$$f = \frac{Min}{\frac{(Max+Min)}{2}} \qquad (2)$$

Let *r* be the ratio of the budget to run the non-current N-1 websites and the budget to run the current website, that is *r*=(Budget for the N-1 non-current websites)/(Budget for the current website), then N=1 + Floor(*r*/*f*) where Floor function rounds down to the nearest integer. Note that this formula, *f* can be rearrange as:

$$f = \frac{2}{1+\frac{Max}{Min}} \qquad (3)$$

So, *f* is smaller when the ratio of $\frac{Max}{Min}$ is large and hence the value of N becomes larger for the same ratio of *r*. This indicates that when the value of $\frac{Max}{Min}$ is large it will be cheaper to run the historical websites. And when *r* is small, the value of N becomes smaller. This is because less budget is allocated to run historical websites.

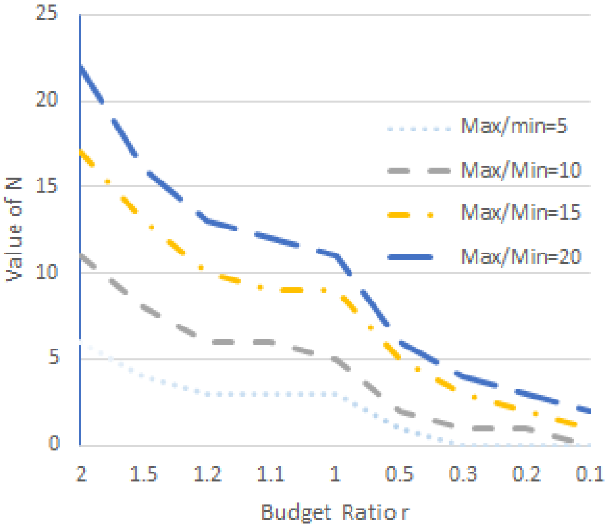

Figure 3. Relationship between r and value of N

If the application is not scalable, then the same virtual machine configuration is used in every period. The cost can only be reduced by running the application intermittently. For example, if the application is run every day from 9 to 5, then the virtual machine charge will be 1/3 of the current website. If it is run Monday to Friday from 9 to 5, then the virtual machine charge will be 23.8% of the current website. The value of N can be determined based on the ratio *r*.

CONCLUSION

This paper presents a Parallel Archiving System for web applications that are periodically renewed, frequently changed in design and supporting technologies from period to period, and are required to keep the previous periods' applications operational in parallel with the current period application. An example of an application with these characteristics is a university's learning management system supporting faculty and students that may be renewed every academic period while allowing users to access previous periods' data. We study the costs of running such system in a cloud platform. We find that the system is capable of keeping a non-current website longer in the system if the cost of running a non-concurrent website in the cloud is less expensive; also if more budget can be allocated to run non-current websites. We continue to study the costs of the Parallel Archiving System.

REFERENCES

[1] Amazon web services, available at http://aws.amazon.com/.

[2] Betts, D. et al, (2013) *Moving applications to the cloud*, 3rd Edition, Microsoft patterns & practices, 3 edition.

[3] Chao, D. & Gill, S. (2015) 'The design of a cloud-based website parallel archiving system', *Issues in Information Systems,* Vol. 16, No. I, pp. 226-231.

[4] Cloud.CIO.Gov, Benefits of cloud computing, available at http://cloud.cio.gov/topics/benefits-cloud-computing.

[5[Google cloud platform, available at https://cloud.google.com/.

[6] Microsoft azure, available at http://azure.microsoft.com/en-us/.

[7] Nickerson, R. (2002) 'An e-commerce system model', *Proceedings of the 8th America's conference on Information Systems*, pp. 310-316

[8] Virtual machines pricing, available at http://azure.microsoft.com/en-us/pricing/details/virtual-machines/.

[9] Wikipedia, cloud computing, available at http://en.wikipedia.org/wiki/Cloud_computing.