ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Economics and Elasticity Challenges of Deploying Application on Cloud

S. Vimal Don Bosco[1], Dr. N. Prabakaran[2]

Research Scholar, Department of Computer Applications, St. Peter's University, Avadi, Chennai, India[1]

Senior vice Principal, St. Joseph College of information Technology, Songea, Tanzania[2]

ABSTRACT: The disclosure of the web applications has significantly improved the variability of its workload patterns and volumes as the number of users/customers frequently increases and shrinks at different rates and times. The web application features have gradually required the need for flexible however inexpensive computing infrastructure to accommodate variable workload. The on-demand and per-user cloud computing model, precisely that of Cloud Infrastructure Service Offerings (CISOs), had quickly evolved and developed by majority of the hardware and software companies with the promise of provisioning utility like computing resources at enormous economics of scale. However, deploying applications on cloud infrastructure does not lead to accomplishing desired economics and elasticity gains, and some challenges block the way for realizing its real benefits. These challenges are due to multiple differences between CISOs and application's requirements and characteristics. Here, detailed analysis and discussion of the economics and elasticity challenges of applications to be e deployed and operate on cloud infrastructure. Also, contains analysis of different aspects of CISOs, modelling and measuring economics and elasticity, application workload patterns and its impact on achieving elasticity and economics, economic-driven elasticity decision and policies, SLA-driven monitoring and elasticity of cloud-based applications. These are supporting with motivating scenarios for cloud-based applications. Here deference perspective analysis that help to cloud customer and potential application's owners to understand, analyse, and evaluate important economics and elasticity capabilities of different CISOs and its adaptably for meeting their application's requirements.

KEYWORDS: Cloud Computing, Cost, Elasticity, Scaling, Economics, Applications, SLA, Cloud Infrastructure Service Offerings, IaaS

I. INTRODUCTION

The significant development in various computing technologies grid, distributed, utility computing, automatic computing[1] has led to innovative ways offering and consuming H/W and S/W resources as a services in what is now commonly known as the cloud computing pattern. United States Government's National Institute of Standards and Technologies (NIST) defines "cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction"[2]. The definition also describes essential characteristics of cloud computing including Rapidly Elasticity of computing resource and Measured Services of computing resources usage. There are three main types; 1. Infrastructure as a service (IaaS) 2. Platform as a Services (PaaS) and 3. Software as a Services (SaaS) [3]. There are four cloud deployment models: private, public, community and hybrid models [4].

The perception is on exploring the economic and elasticity challenges of internet based business (or e-business) applications deployed on public cloud infrastructure. The term economics refers to the efficient use and management of cloud infrastructure resource required to run e-business applications with desired performance levels. The term elasticity refers to the ability to dynamic grow and shrink computing infrastructure resource through automatic mechanisms over the internet in order to serve variable workloads of e-business applications efficiently.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

Motivating economics and elasticity of application

Let us consider an online shopping application, MyShop, which is selling wide range of product. It has three tier architecture that consist of web server/load balancing, application server and database server. MyShop's capacity management team has classified expected workload pattern and required basic computing resources as follows:

- Normal Operation workload: 1webserser/load balancer, 2 application server and 1 database server during all business operation times.
- Mid-Week-Sales Workload: same as normal workload operation workload plus 2 additional application servers and 1 additional database server works as a slave from 11 am to 9 pm on Wednesday.
- Week-End-Sales Workload: same as a normal operation workload plus 4 additional application server and 2 additional database servers work as slave from 10 and to 10 pm each on Saturday and Sunday.

The name of CISP's and cloud server types have been made anonymous as goal of this example is to illustrate the economics and elasticity benefits for MyShop but not to evaluate and compare CISP's offers. Focus our calculations on the application tier because it has variable workload. The normal operation work load requires 2 fixed application servers for whole year. So, subscription based cloud server operation is very economical here as most CISPs provide it at discounted hourly prices for one year term. On-demand cloud server offering, on the other hand, are economical option for mid-week and week-end sales workloads as it is billed on hourly-basis without any long term commitments. The yearly costs of both workloads are calculated using the following formulas:

Fixed server costs (yearly) = *no. of app servers x $ server subscription prices/year.*

On-demand server costs (yearly) = *no. additional servers x no. of usage hours per server/week x $ on-demand server price/hour x 52 weeks.*

CPU-intensive servers have been select here as application servers perform most of the business logic processing at very large volumes.

II. CLOUD INFRASTRUCTURE SERVICE OFFERING (CISP)

A considerable number of cloud infrastructure services providers (CISP) has emerged with different cloud infrastructure service offerings (Wipro, orange space, TCS etc.,) [7]. Investigation of different CISOs of many CISPs in terms of:

a) Computing Resource Bundling and Specifications
b) Pricing models and offering types
c) Software system licensing
d) Elasticity support for infrastructure Resources.

Computing Resource Bundling and Specifications

CISPs offer different cloud infrastructure service bundles such as cloud servers, cloud storage, and internet/network resources. Cloud server bundles are the core service offering as it offers processing capabilities and, therefore, *CISPs* offer them at different fine-grained levels which combine different computing resources such as processing unit, memory, and disk and/or network bandwidth. A variation in one or more resources' capacity in a bundle results in what is called server *instance*, *class*, or *size*. Unlike most *CISPs* who have specific number of instance offerings, *Elastic Hosts* [5] and *Cloud-Sigma*[6] allow its cloud consumers to customize their cloud servers by varying CPU, RAM, disk and data transfer/ bandwidth capacity at very fine-grained levels.

Pricing models and offering types

Four types of pricing models which are correlated with certain offering types.

1. ***Per-user Model:*** It is also known as pay-as-you-go where computing resources are bundled and billed per unit of time usage. This model most commonly used by CISPs for pricing cloud server instance in which prices are often varied based on CPU, RAM disk storage and/or bandwidth capacity (it varies between CISPs).
2. ***Subscription model*:** In this model, cloud consumers subscribe in advance for computing resources usage for a specific period of time by signing a contract/ agreement. Computing resources are grouped into

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

different packages often called *Dedicated Servers* or *Reserved Instances* in which prices vary according to included resources' capacity.

3. ***Prepaid per-use model*:** it is a variation of the *per-use* pricing model. In this model, *on-demand* servers are billed hourly but from a prepaid credit.
4. ***Subscription + per-use model*:** It is an intermediate model between *per-use* and *subscription* models. In this model, *Dedicated Servers* must be rented in advance for a period of time and additional *cloud servers* can be requested *on-demand* and billed at *per-use* charges [7].

Software and system licensing

There are two main types of software; Operating Systems, e.g., range of Windows and Linux/Unix, and Application Software, e.g., Oracle Web Logic, MySQL Enterprise and Apache HTTP. *CISPs* differ in the type and number of software and system they offer with their cloud server instances.

1. ***Bring Your Own Software License (BYOSL)*:** Customers can bring their own license to Amazon RDS with no additional software licensing or support charges.
2. ***On-demand DB instances*:** Customers are charged per hour license use per RDS DB instance running Oracle DB.
3. ***Reserved DB instances*:** Customers pay one-time prepaid charge per RDS DB instance to get reduced hourly-usage rate [8].

Elasticity support for cloud infrastructure resources

CISPs often relate elasticity with different types of computing resources. Common elasticity examples include:

- Adding/removing server instances or resizing server capacity by adding/removing additional CPUs and/or RAMs.
- Increasing/decreasing storage capacity by adding/removing additional disks or virtual storage.
- Increasing/decreasing network speed and number of IP addresses.
- Increasing/decreasing amount of data transfer and number of data operations/requests.

III. ECONOMICS OF ELASTIC CLOUD-BASED APPLICATIONS

The main challenges that could face cloud consumers to understand and achieve economic elasticity for their e-business applications when deployed on public cloud infrastructure.

Economics and elasticity modelling of cloud-based applications

Economics and elasticity of public cloud infrastructure have been heavily reported and demonstrated in research and industry communities as fundamental drivers for different application domains [9]. Some research work [10] investigated the migration costs of software applications to the public cloud infrastructure. However, such work does not consider the elasticity dimension of operational costs. Moreover, it is based on simple calculation models which are tailored for specific application use cases ith fixed workload patterns. Li et al. [14] proposed comprehensive financial models for calculating total cost of ownership and utilization costs of elastic cloud infrastructure but from the cloud provider's perspective.

In practice, *elasticity is not an autonomic feature and cloud consumers have to configure appropriate scaling policies to enable it*. Without advanced modelling and analysis tools this is almost impossible. The flexibility of on-demand computing resources makes an application's operational costs variable and more finely granular as it is tightly coupled with defragmented fine-granular pricing schemes that are metered differently such as: CPU usage-per-hour, number of I/O operation call, size of storage volumes. Therefore, having appropriate elasticity and economics models and metrics is essential to enable cloud consumers to analyse, plan, and control costs and scaling policies of their e-business applications on any public cloud infrastructure at fine-granular levels.

Economics-driven decision-making for elasticity of cloud-based applications

The on-demand provisioning of fine-grained computing infrastructure services has introduced new ways for cloud consumers to achieve business resilience and agility at reduced costs. However, the *economics and benefits* of CISOs are not an automatic gain for cloud consumers because public cloud infrastructure does not automatically scale computing resources based on application's workload requirements. Cloud consumers are challenged with the need for

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

automated mechanisms (e.g., intelligent elasticity engine) that make and execute economic scaling decisions on the right time, to the right cloud resources, and with the right amount of cloud resources. Such engine should be configured with appropriate elasticity policies b1ased on the application's business and technical metrics (SLAs) and application's workload changes. The elasticity decisions can then be triggered automatically in proactive or reactive ways as it will be discussed later in this section.

What to scale?

We also believe deciding on which cloud resources to scale is another important pillar for economics-driven decision making elasticity. This specifically requires identifying performance bottlenecks in cloud-based application architecture and collecting monitoring data about it and characteristics of its workload. Identified reasonable list of potential capacity bottleneck points in a typical web application architecture deployed into the cloud.
These bottleneck points include:
• ***Network bandwidth*:** between the load balancer and the application servers as well as between the application servers and database servers.
• ***Load balancing*:** ability of a load balancer to properly distribute load across the application servers.
• ***Computing capacity*:** the CPU, RAM and internal storage utilization of application and database servers.
• ***Computing resource performance*:** number of I/O or read/write operations per unit of time for application and database servers.

Another identified a list of scaling points that could cause performance bottlenecks.
These points include:
• ***Processing power*:** CPU speed measured in GHz.
• ***Memory*:** RAM capacity measured in GB.
• ***Network bandwidth*:** network speed in Gbps.
• ***Database performance*:** number of transactions/second.
• ***Disk storage*:** system storage capacity in GB or TB.

Elasticity example: MyShop scaling strategies

In our scenario, let us consider possible *elasticity strategies* that could be planned for *MyShop's* application tier.
There could be three scaling strategies:
• ***Scaling out-in (Horizontal Scaling)*:** by adding four additional application servers to the existing two main application servers every Saturday and Sunday from 10 a.m. to 10 p.m. and removing it at all other times. All servers should have the same processing capacity (we use *CPU intensive* server of small computing capacity)
• ***Scaling up-down (Vertical Scaling)*:** by replacing the two main application servers with one more very powerful application server (i.e., we use one *CPU-intensive* server with computing capacity equivalent to six small servers) and then switching to the two main application servers at all other times
• *Hybrid Scaling*: a combination of *horizontal* and *vertical* scaling strategies with variation of number of cloud servers.

Service level agreement of elastic cloud-based applications

We distinguish between two types of service level agreements (SLAs):

1**. *Cloud Infrastructure SLA (CI-SLA)*:** These guarantees are offered by a public CISPs to its cloud consumers to assure certain quality levels of their cloud computing resource capabilities and specifications (e.g., server performance, network speed, resources availability, storage capacity).In other words, *CI-SLAs* reflect the perspective of public *CISPs*.
2. ***Cloud-based Application SLA (CA-SLA)*:** These guarantees relate to the levels of quality of an application which is running on a public cloud infrastructure. In particular, cloud consumers often offer such guarantees to their application's customers/end users to assure quality of services they offer to them such as application's response time, availability and security. Hence, *CA-SLAs* reflect the perspective of *cloud consumers*. For example, cloud consumers who deploy their Customer Relationships Management (CRM) on public cloud infrastructure would be interested in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

monitoring the *CA-SLA of their application*. In such case, average waiting time, average service time, and queue length are good examples of essential properties of *CA-SLA* which need to be monitored and maintained at the consumer side.

VI. CONCLUSION AND FUTURE WORK

The provision of transactional business applications through the Internet has dramatically increased its dynamism in terms of variability of workloads volumes and patterns demanding highly flexible yet cost effective IT infrastructure and resources. The on-demand provisioning of various computing resources as services at massive economies of scale has perfectly matched this need. Deploying such e-business applications on public cloud infrastructure can highly contribute to achieving its elasticity requirements at economic costs but this is not a self-inherent capability of cloud infrastructure resources. The fast-growing number of diverse cloud infrastructure services has attracted considerable attention from potential cloud consumers but at the same time it has introduced new challenges for them. Realizing cost-effective and resilience use of on-demand usage-based CISOs require extensive understanding and analysis of various aspects related to the CISOs and business requirements of cloud consumers.

We have introduced a comprehensive analysis and discussion of the economics and elasticity challenges that cloud consumers, i.e., e-business application's owners, should consider for their transactional e-business applications when deployed on a public cloud infrastructure. Collectively, the analysis and discussion provide a *multi-lenses* insights that can help cloud consumers to understand, analyze and evaluate economic factors as well as elasticity capabilities of different CISPs' cloud service offerings, especially with regards to its suitability for their applications requirements. A future research agenda for research and industry communities to help cloud consumers to investigate and tackle defined research challenges.

REFERENCES

1. Different types of computing by K Vasudev in MSDN Blog on Feb 5, 2009.
2. The NIST Definition of cloud computing by Peter Mell and Timothy Grance on 2011.
3. Cloud Computing Service model by Quinn Devery on jun 18, 2012.
4. Cloud CIO Gov, Deployment Models, 2014.
5. India based cloud computing service provider by Basant Narayan Singh on March 7 2010.
6. Wikipedia on July 10, 2014.
7. Cloud Computing Implementation, Management and Security by John W Rittinghouse and James F.Ransome on 2010.
8. Purchasing and RDS Reserved Instance in EU-West by Andrew on August 18, 2010.
9. Cloud Computing: Paradigms and Technologies by Ahmed Shawish and Maria Salama on 2014.
10. Cloud Computing Security Issues and Challenges by Kuyoro S O on 2011.

BIOGRAPHY

S.Vimal Don Bosco
Research Scholar, Department of Computer Applications,
St.Peter's University, Avadi, Chennai – 600 054.

Dr.N.Prabakaran,
Senior Vice Principal,
St. Joseph College of information Technology,
Songea, Tanzania.