Branching Processes Theory Application for Cloud Computing Demand Modeling Based on Traffic Prediction

Victor Romanov, Aleksandra Varfolomeeva, and Andrey Koryakovskiy

Department of Computer Science, Russian Plekhanov University of Economics, Moscow, Russian Federation
{victorromanov1,aovarfolomeeva}@gmail.com, avkor@list.ru

Abstract. With cloud computing growing in popularity cloud-service providers must guarantee that data are processed rapidly and transferred when and where they are needed. Unfortunately, it is extremely difficult to predict the exact performance characteristics and demands on the network at any particular time. In this paper we show that the cloud computing demand can be developed as a branching stochastic process. Branching processes are used to describe random systems such as population development, nuclear chain reactions and spread of epidemic disease. A statistical model is described and using this model we propose a method for determining the unknown probability distribution of queries. Network traffic modeling is an issue of great importance to both consumers and providers of cloud-based services. Firstly, traffic modeling helps to represent our understanding of dynamic demand for cloud services by stochastic processes. Secondly, accurate traffic models are necessary for service providers to properly maintain quality of service.

Keywords: cloud computing, branching process, network traffic modeling.

1 Introduction

Cloud computing is one of the hottest topics in all of information technology today. This is a new model of delivering computing resources in which centrally administered computing capabilities are provided as services on-demand over the network to a variety of customers. According to IDC's analysis, the worldwide forecast for cloud services in 2013 will amount to $44.2bn, with the European market reaching €6,005m in 2013.[1] The potential benefits of cloud computing like lower costs, faster implementation, and more flexibility are overwhelming. However, attaining these benefits requires that new technologies and solutions managing the huge number of operations and volumes of data within a cloud transparently and without service interruptions emerge. As cloud computing enables users to store all their data on the network predicting net-

[1] IDC *Cloud Computing 2010 - An IDC Update,* Frank Gens, Robert P Mahowald, Richard L Villars, Sep 2009 - Doc # TB20090929, 2009.

M. Bajec and J. Eder (Eds.): CAiSE 2012 Workshops, LNBIP 112, pp. 502–510, 2012.
© Springer-Verlag Berlin Heidelberg 2012

work performance and cloud-based services demand may become a challenging issue. In this paper we suggest that branching processes theory will fit for both describing the dynamics of cloud services demand and predicting the network traffic in cloud computing environment. We consider that more and more clients will know about cloud services, one client from another, like epidemic of disease spreads. We briefly describe the basic principles of cloud computing concept and introduce branching processes theory. The purpose of this paper is to provide a stochastic model that would be helpful for both consumers and providers of cloud-based services. On the one hand, traffic modeling helps to represent our understanding of dynamic demand for cloud services by stochastic processes, and on the other hand accurate traffic models are necessary for service providers to avoid "bottlenecks" and to improve quality of services provided.

2 Literature Review

The design of robust and reliable network services for cloud computing environment is a challenging task. The only path to achieve this goal is to develop a detailed understanding of the traffic characteristics. An accurate estimation of the network performance is vital. Traffic models enable network designers to make assumptions about the networks being designed based on past experience and also enable prediction of performance for future rapidly changing requirements in cloud environment [10]. A corpus of literature on network traffic modeling exists. One of the most widely used and oldest traffic models is the Poisson Model. The Poisson process is characterized as a renewal process and Poisson distribution is the predominant model used for analyzing traffic in traditional telephony networks [11]. Deterministic Traffic Model [12] is proposed for providing real time service over real time channel where clients declare their traffic characteristics and performance requirement at the time of channel establishment in this model. Chaotic maps are low dimensional nonlinear systems whose time evolution is described by knowledge of an initial state and a set of dynamical laws. In [13] the author illustrates traffic characteristics that can be modeled by considering chaotic maps. Wavelet-based models use wavelet transform function to model long-range dependence traffic such as traffic measured on Ethernet. Multifractal wavelet model is presented in [14].

3 Cloud Computing

Cloud computing is a model for enabling 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. This cloud model promotes availability and is composed of five essential characteristics: on-demand self-service, broad network access, resource pooling, rapid elasticity, and measured service.[2]

[2] NIST: Cloud Computing Program: http://www.nist.gov/itl/cloud/index.cfm

There are three service models of cloud computing:

Software as a service (SaaS): is software offered by a third party provider, available on demand, usually via the Internet configurable remotely. Examples include online word processing and spreadsheet tools, CRM services (Salesforce CRM [5], Google Docs).
Platform as a service (PaaS): allows customers to develop new applications using APIs deployed and configurable remotely. The platforms offered include development tools, configuration management, and deployment platforms. Examples are Microsoft Azure [3] and Google App engine.
Infrastructure as service (IaaS): provides virtual machines and other abstracted hardware and operating systems. Examples include Amazon EC2 and S3 [1], [2], Terremark Enterprise Cloud [6], and Rackspace Cloud [4].

The following deployment models are available for cloud computing services:

- *Private cloud*: services built according to cloud computing principles, but accessible only within a private network
- *Community cloud:* cloud services offered by a provider to a limited and well-defined number of parties
- *Public cloud:* available publicly - any organization may subscribe
- *Hybrid cloud:* a composition of two or more clouds (private, community or public)

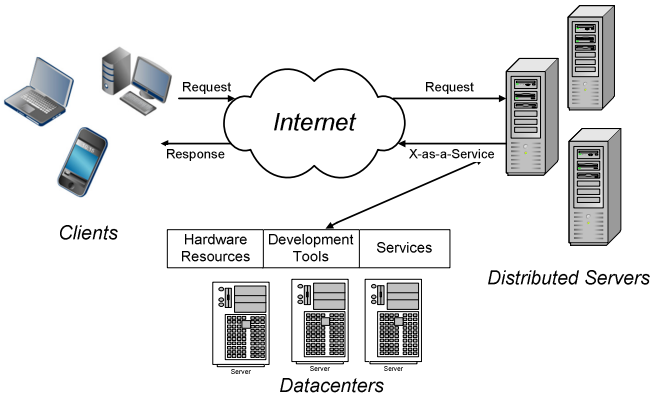

Fig. 1. Cloud Computing Topology

Thus, cloud computing provides a pool of highly scalable and easily accessible virtualized resources capable of hosting end-user applications exploited in a pay-as-you-go model. Cloud computing involves the following three basic components [7], which are illustrated in Figure 1: clients, datacenter and distributed servers. For many companies with highly variable IT needs, cloud computing can be an alternative to maintaining an expensive oversupply of in-house computing power. However, there are some major obstacles which hinder the adoption and growth of cloud computing. As every technological concept, cloud computing is not an exception in terms of trust and

security issues. Once data are outsourced to a third-party cloud provider, several concerns arise about security, availability and reliability of data.

4 Branching Processes Theory

A branching process is a process where an initial random number of objects 'create' more objects of the same or different type, and these objects continue to 'create' other objects, with the system developing in accordance with some probability law. Branching processes are used to describe random systems such as population development, nuclear chain reactions and spread of epidemic disease. An example of such a process is a population of individuals developing from a single progenitor – the initial individual. It produces a random number of offspring, each of them in turn produces a random number of offspring; and so the process continues as long as there are live individuals in the population. Figure 2 is a graphic illustration of a general multilevel branching process. The branching process was proposed by Galton [9], and the probability of extinction was first obtained by Watson [8] by considering the probability generating function for the number of children in the nth generation.

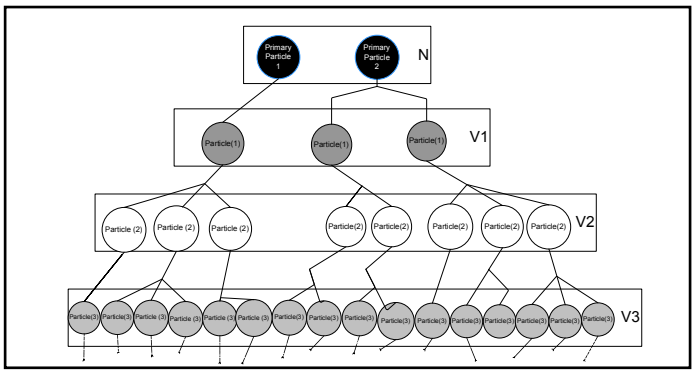

Fig. 2. Graphic illustration of a multilevel branching process

Let N be a random variable, with a probability distribution function $g_N(n)=P(N=n)$, with mean μ_N and variance σ_N^2. Let $\{X_i\}$ be a series of independent identically distributed random variables, with a common distribution $f_X(X)$ and with μ_X as the mean and σ_X^2 as the variance of each element in the series. The sum of N elements of the series $\{X_i\}$ is denoted by the following sum

$$V = X_1 + X_2 + ...X_N \tag{1}$$

The mean of V is denoted by

$$E[V] = E\left[X_i\right] \cdot E[N] \tag{2}$$

And the variance of V is denoted by

$$Var[V] = E[N] \cdot Var\left[X_i\right] + Var[N] \cdot E^2\left[X_i\right] \tag{3}$$

The distribution density function $f_V(v)$ of V can be derived from the basic formula for conditional probabilities

$$f_V(v=k) = P(V=k) = \sum_{n=0}^{\infty} P(N=n) P(X_1 + \ldots + X_n = k) \tag{4}$$

Let us denote $f_X(x)$ and $\phi_x(\omega)$ as the distribution density function and generation function of X_i respectively, and $g_N(n)=P(N=n)$ and $\phi_N(\omega)$ as the distribution density function and generation function of N respectively. For a fixed n, the distribution of the sum $X_1+\ldots+X_n$ is expressed by the n-fold convolution of $\{f_X(x)\}$ with itself, due to the independence of the series $\{X_i\}$. Equation (4) can be written in a more compact form

$$f_V(v=k) = \sum_{n=0}^{\infty} g_N(n) \{f_X(x)\}^{n*} \tag{5}$$

This formula can be simplified by using the generating functions [15].

Branching processes theory can be applied for modeling in cloud computing environment. One of the most important issues in cloud computing environment concerns network efficiency and performance prediction. Where there are large quantities of data involved in an application, access to the data must be fast and reliable or the application's runtime will be excessive. From the viewpoint of a service provider, demands on the network are not entirely predictable. Branching processes theory helps us to model the dynamic demands for cloud services. More and more clients are informed on the service, one client from another based on random mechanism. This process is similar to epidemic of disease spread.

5 Data flow Prediction Model

To design effective and efficient network solutions for cloud environment and to understand and solve performance problems arising in communication networks providers require accurate models to describe network traffic. The main problem is to forecast the frequency of queries that are going to appear. To evaluate the performance of the proposed technique, we demonstrate how branching processes theory can be applied for building data flow prediction models.

The initial vertex of the graph is assigned $t=0$ (Figure 3), time when the original message reaches the host C_i. This point is taken as the reference time of receipt of queries to the cloud service provider's host center C_i. The initial vertex of a graph is based on the number of arcs equal to the quantity of primary needs. Arcs whose vertices correspond to the secondary queries come of the vertices of the graph corresponding to the initial query. We consider that the primary queries come into the provider's host randomly. The process of queries admission will be considered as a

branching, while allowing that individual queries' paths are independent. The time intervals between $t=0$ and the time of requests admission are random variables with distribution function $F_1(x)$ and density $f_1(x)$. Obviously $f_1(x)$ represents the latency of the processes of information alerts. The number of initial requests coming to the provider for a certain period of time is the random variable

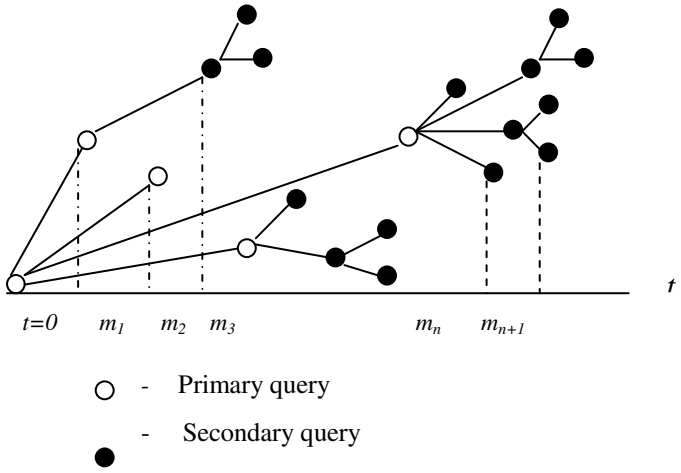

Fig. 3. A branching process of generating queries

$$\nu_1, P(\nu_1 = k) = p_{1k}, k = 0,1,2,...; \sum_{k=0}^{\infty} p_{1k} = 1 \tag{6}$$

We assume that the distribution of the primary query is binomial. The generating function corresponding to the binomial distribution of the primary query is:

$$G_1(u)=(q_0+q_1u)^n, \tag{7}$$

where u is the parameter of the generating function. The distribution function of the moments of primary queries receipt is exponential $F_1(x)=1-e^{-\lambda_1 x}$, where λ_1 is the volume of primary queries. After each initial query with probability p_0 does not appear any secondary query, and with probability $p_1=1-p_0$ there is at least one secondary request. We assume that the distribution of secondary queries generated by one primary request is subject to the binomial distribution with generating function

$$G_2(u)=(p_0+pu)^n , \tag{8}$$

where u is the parameter of the generating function. The distribution function of the time intervals between the moments of initial queries receipt and stimulated directly by them secondary queries is defined as $F_2(x)=1- e^{-\lambda_2 x}$, where λ_2 is the intensity of secondary queries receipt. The expectation of the number of requests in the time interval $[t, t+\tau]$:

$$M^*[t,\tau]=\begin{cases}\dfrac{nq_1\lambda_1\tau}{\lambda_1-\lambda_2 p_0}\left\{(\lambda_1-\lambda_2)e^{-\lambda_1 t}+\lambda_2 p_1 e^{-\lambda_2 p_0 t}\right\} & (\lambda_1\neq\lambda_2 p_0)\\ \dfrac{nq_1\lambda_1\tau}{p_0}e^{-\lambda_1\cdot}(p_0+p_1\lambda_1 t) & \lambda_1=\lambda_2 p_o\end{cases} \tag{9}$$

If $\frac{\lambda_1}{\lambda_2} < p_1$, then the distribution has a maximum. If $\frac{\lambda_1}{\lambda_2} \geq p_1$, then the distribution hasn't a maximum and decreases from the beginning. In case $\lambda_1 = \lambda_2 p_0$, for $p_0 \geq p_1$ distribution is monotonically decreasing; for $p_0 < p_1$ distribution has a maximum. Expectation function graphs of the number of queries are constructed for the following values: time units are *t=[0,130]*, the number of consumers which may create the primary requests is 10, the probability of the customer application with the primary request is $q_1=0.7$, the time interval of the request arrival τ=1, the volume of the primary queries from the consumer at a time $\lambda_1=0.05$. The nature of the functional dependence is affected by primary and secondary queries intensity compliance. The Figure 4 contains plots of the expectation of the number of queries, in this case the probabilities of the primary p_0 and secondary p_1 queries: for $p_0 \geq p_1$ distribution is monotonically decreasing, for $p_0 < p_1$ the distribution has a maximum.

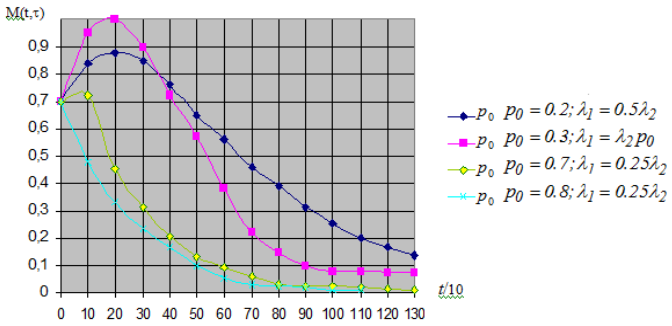

Fig. 4. The theoretical queries distribution functions over time

6 Research Methodology

Global giants like Amazon are definitely more cost-effective and less sensitive to daily, annual and other traffic imbalances than the local providers. In order to evaluate the performance of the proposed technique and to justify the further need for the model we are developing, the Russian IaaS market was analyzed. The table below presents the companies whose activity was examined. Sample organizations for analysis were selected based on the following criteria. Service content is similar to the

services of a conventional hosting provider. Clients come from different industries. Service providers are with different backgrounds: they include new and established cloud suppliers (local).

Table 1. Russian IaaS (Hosting) Providers

#	1	2	3	4	5	6
Provider Name	Activecloud.ru	Clodo.ru	ISP Server	Scalaxy	Slidebar.ru	Selectel

In general, Russia's cloud hosting is not so much an alternative to Amazon Web Services, as a convenient substitute to traditional hosting. Most of the companies have encountered difficulties providing a comprehensive explanation of the dynamic characteristics of network traffic and had even experienced the bottlenecks that caused them to lose money. To measure how good the data flow prediction model is, experiments are carried out where cloud usage logs are analyzed seeing if our model can predict the loads seen in the logs.

7 Conclusion

Cloud computing is a model of delivering computing resources in which centrally administered computing capabilities are provided as services on-demand over the network to a variety of customers. As popularity of cloud services is growing rapidly, cloud-service providers must guarantee that data are processed effectively and transferred when and where they are needed. Unfortunately, it is extremely difficult to predict the exact performance characteristics and demands on the network at any particular time. In this paper we suggest that branching processes theory will fit for both describing the dynamics of cloud services demand and predicting the network traffic in cloud computing environment. We consider that more and more clients will know about cloud services, one client from another, like epidemic of disease spreads. The purpose of this paper is to provide a stochastic model that would be helpful for both consumers and providers of cloud-based services. On the one hand, traffic modeling helps to represent our understanding of dynamic demand for cloud services by stochastic processes, and on the other hand accurate traffic models are necessary for service providers to avoid "bottlenecks" and to improve quality of services provided.

References

1. Amazon Amazon elastic compute cloud, `http://aws.amazon.com/ec2/` (accessed March 2012)
2. Amazon: Amazon simple storage service, `http://aws.amazon.com/s3/` (accessed March 2012)
3. Microsoft: Windows Azure, `http://www.windowsazure.com/` (accessed March 2012)

4. Rackspace Cloud, http://www.rackspace.com/cloud/ (accessed March 2012)
5. Salesforce CRM, http://www.salesforce.com/eu/ (accessed March 2012)
6. Terremark Enterprise Cloud, http://www.terremark.com/services/infrastructure-cloud-services/enterprise-cloud.aspx (accessed March 2012)
7. Tsaravas, C., Themistocleous, M.: Cloud Computind & E-Government Myth or Reality? In: tGov Workshop 2011 (tGOV 2011), March 17-18 (2011)
8. Watson, H.W.: Solution to problem 4001: Educational Times, August 1 (1873)
9. Galton, F.: Problem 4001: Educational Times, April 1, p. 17 (1873)
10. Mohammed, A.M., Agamy, A.F.: A Survey on the Common Network Traffic Sources Models. International Journal of Computer Networks 3(2) (2011)
11. Frost, V.S., Melamed, B.: Traffic Modeling for Telecommunications Networks. IEEE Communications (March 1994)
12. Ferrariand, D., Verma, D.: A Scheme for Real Time Channel Establishment in Wide Area Networks. IEEE Journal on Selected Areas in Communications 8(3) (April 1990)
13. Erramilli, Sing, R.P., Pruthi, P.: Chaotic Maps as Models of Packet Traffic: The Fundamental Role of Teletraffic in the Evolution of Telecommunications Networks. In: Proc. of the 14th ITC, June 6-10 (1994)
14. Ribeiro, V.J., Riedi, R.H., Crouse, M.S., Baraniuk, R.G.: Simulation of non-Gausian Long-Rang-Dependent Traffic Using Wavelets. In: Proceeding SIGMETRICS 1999, April 9 (1999)
15. Cohen, I., Golan, R., Rotman, S.: Applying Branching Processes Theory for Building a Statistical Model for Scanning Electron Microscope Signal. Optical Engineering 39(01)
16. Harris, T.E.: The theory of branching processes. Springer, Berlin (1963)