Video Delivery Services in Media Cloud with Abandonment: An Analytical Approach

Sampa Sahoo[1], Maneesha Nidhi[2], Kshira Sagar Sahoo[1], Bibhudatta Sahoo[1], Ashok Kumar Turuk[1]
[1]Department of Computer Science and Engineering, National Institute of Technology, Rourkela, India
[2]Software Development Engineer, Amazon Development Center, India
Email:(sampaa2004, maneeshanidhi, kshirasagar12, bibhudatta.sahoo, akturuk)@gmail.com

1 Abstract

Distribution of video content over the Internet has drastically transformed the consumption of media. Content providers, naturally, would like to ensure that their videos play on users' devices whenever requested, without failure or interruptions. Due to the varying nature of user needs, procurement of computing resources proves to be tricky, leading to the popularity of cloud-based approach. Media cloud is a computing paradigm dedicated for multimedia services and delivers on demand services (e.g., video) by dynamically acquiring cloud resources. Use of cloud resources helps service providers to lessen their operational cost, reduce delay and abandonment rate to deliver adaptive video stream. The abandonment rate, delay, user engagement and repeat viewership plays a vital role in service providers revenues. In this paper, we use an analytical model based on queuing theory to find the effect of queue size (buffer size) on abandonment, blocking and successful services. Further, the relationship between the number of virtual machines, waiting time (delay) and abandonment rate is also examined. We also derive a relationship between the number of user requests in the system and the virtual machines required to respond to the same.

2 Introduction

Consumption of media has been drastically transformed by the distribution of videos over the Internet. Video streaming to Internet-enabled devices like desktops, laptops, and smart-phones among others, has become commonplace. The Cisco Visual Networking Index (VNI) report published in 2016 stated that IP video traffic is estimated to escalate to 82% of Internet traffic by the year 2020, which can be understood as nearly a million minutes of video content being viewed every second [1]. Video on demand (VoD) systems allows users to view video content whenever they choose to offer a stark contrast to the over-the-air broadcasting approach that was prevalent in the 20^{th} century. A specialized server called the VoD server maintains video content and handles its delivery over the Internet in response to user requests. The on-demand model was developed to overcome the challenge of meeting users' fluctuating demands efficiently. Due to the varying nature of user demands, procurement of computing resources proves to be tricky. Maintaining resources at peak requirements proves to be costly while keeping resource requirement minimal leads to user dissatisfaction and non-fulfillment of service guarantees. Many new classes of big data applications in the field of healthcare, education, aged care and more, emerged as a result of streaming interactive multimedia content. For example in the education domain, students can explore subjects through video streaming and learn from instructors who are not available locally [2].

Cloud computing a revolution in the IT industry, use pay-as-you-go service scheme which means organizations are enabled to pay only for the resources and services they use. The cloud system based upon virtualized infrastructure provides elasticity and scalability, multi-tenancy, etc. and data are stored in logical pools of virtual machines running on servers in data centers managed by a cloud provider [3]. The migration of media such as television and movies to the Internet makes it essential for the content providers to ensure the high quality streaming experience for the viewers. This is made possible by providing quick start-up of videos, with less viewer abandonment, and higher viewer engagement allowing for a higher fraction of repeat viewers [4]. Poor viewing experience results in high abandonment rates and negatively impacts the providers economy and brand-value. Thus, global private networks like Content Delivery Networks (CDN) are used for video-content delivery. CDN ensures the highest quality video stream, increased engagement and repeats viewership.

One of the solutions for large-scale multimedia services in the big data era is a cloud-based multimedia processing platform known as Media cloud [5]. The media cloud framework is used to disseminate adaptive video streaming services efficiently. The virtualized computing and storage cloud resources in edge servers of Media cloud form a virtual content delivery network in response to the varying application demands for efficient services to the user. The cloud-based architecture significantly reduces the operational cost and the delay of delivering adaptive video streaming by using virtualized resources intelligently in an on-demand manner [6]. Delay is significant and non-ignorable for large-scale multimedia services hosted in the cloud environment. Even a slight increase in startup delay causes increased dissatisfaction resulting in viewers abandoning the video. In [4], it is shown that a rise of 1 sec in startup delay increases the abandonment rate by 5.8%. A hot topic in the media cloud framework is determining ways to reduce waiting time [5].

Media-wise cloud aims to circumvent the limitations of traditional video delivery by carving out storage, bandwidth and computation resources from the cloud to provide services to the user with minimum cost and delay. Despite several benefits, Media cloud needs to deal with various obstructions for designing and controlling a service operation to achieve a balance between operational efficiency and service quality. We can evaluate Media cloud either by physical or mathematical model. Physical modeling is not always preferable due to the labor, cost involved with it. In light of the trade-off involved, it is but natural to use a mathematical model of the system. Analytical model and simulations are two ways to represent the mathematical model of a system. Here we designed analytical model of Media cloud using queuing theory and through simulation validated it. Unlike physical queues, the queue formed by customer requests in the Media cloud is invisible. The customers cannot observe how long the queue is before initiating a request. If a virtual machine is not free to process a request, it is added to the central queue maintained by the system. As the time spent in queue increases, so does customer impatience, which ultimately leads to abandonment. In this paper, we analytically show the effect of buffer size, the number of virtual machines (VMs), arrival rate on the probability of user blocking, abandon and success rate. Results obtained from numerical analysis can be used for the decision-making process related to the design, control, and measurement of video distribution service in media cloud. The detailed contributions of this paper are as follows:

- We analytically derive the relationship between buffer size and user blocking, success and abandonment rate. Further, the effect of the number of VMs and arrival rate on abandonment rate is shown.
- We derive the influence exerted by abandonment rate and number of VMs upon waiting time.
- Finally we validate the effectiveness of our approach through numerical analysis and find an appropriate number of VMs for a defined set of service and arrival rate to restrict the abandonment rate to a limit.

The rest of this paper is designed as follows: reviews of related work is outlined in section II. Section III discusses the system architecture, models and theoretical analysis of various decision making parameters. Section IV shows the numerical analysis of the analytical model. Finally, section V concludes this paper.

3 Related work

An optimal strategy to reduce the operational cost has been proposed by Yichao *et al.* after examining the trade-off between the caching, transcoding and bandwidth costs in Media cloud in [6]. A cost-efficient content placement in Media cloud has been proposed in [7]. Yu *et al.* [8] describe the configuration of cloud utility required for VoD applications to meet dynamic user demands at a modest cost. The authors used the $M/M/m/\infty$ queuing model to characterize viewing behavior of users and to derive the server capacities required to support smooth playback. The details of video management and resource allocation for large-scale VoD cloud have been covered in [9], while [10] surveys interactive VoD systems. The changes brought about in viewer behavior due to video quality has been investigated by Krishnan *et al.* [4]. The authors show that this dependence has the potential to affect the economy of content providers negatively. Thus, factors like abandonment rate, play time, high-quality streaming impacts the provider's revenues.

Liang in [5] estimates the delay, referred to as the average waiting time in his work, for various multimedia services hosted on the cloud platform. Zhang *et al.* [11] discuss the approaches for optimal placement of content. They model request dispatching using a Markov Decision Process(MDP) model that assumes a Poisson arrival rate and an Exponential service rate for the video distribution services. Xiaoming *et al.* [12] study the resource allocation problem in the context of the multimedia cloud. They characterize the service process using a $M/M/1$ queuing system for the single-service and multi-service scenario. The authors also discuss an approach using the $M/H_m/1$ queuing system for the multi-service scenario, to minimize the parameters response time and resource cost. Jordi *et al.* [13] model the cloud architecture that maintains specified Quality of Service (QoS) using a combination of $M/M/1$ and $M/M/m$ queuing models. Soamar *et al.* [14] consolidate the workload at cloud data centers to provide guaranteed QoS by using the concept of request reneging. The researchers employed the $M/M/1$ queue with reneging model and aimed at reducing power consumed by each server in the cloud data center. The novelty of our work is that we have we have used an analytical model and numerical analysis to find the effect of buffer size, arrival and service rate, number of VMs on abandonment rate, and to obtain the optimal number of VMs.

4 System Models and Architecture

4.1 System Architecture

In this section, we consider a systematic end-to-end view of on-demand video delivery service over Media cloud. The main components are video service provider, cloud service provider and end users as shown in Fig. 1.

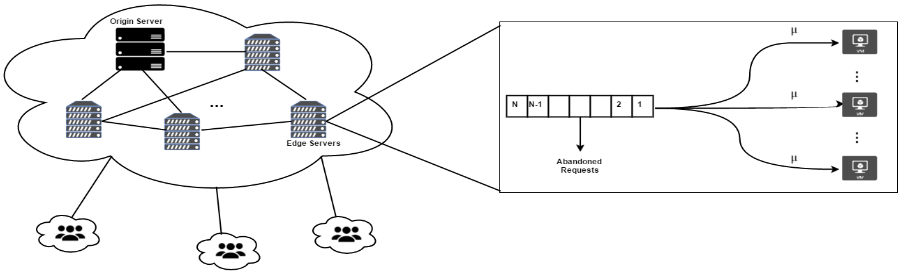


Figure 1: System Architecture for Video on Demand service via Media cloud

Video service providers provide a collection of video contents via origin server. The Media cloud providers own a list of servers (VMs) which are carved out in geographically distributed data centers (origin/proxy/edge servers). These VMs work collaboratively to build an elastic overlay on top of its underlying physical networks to serve end users spread across multiple geographical regions. Under this architecture, each user is served by nearby edge servers. If there is a miss in first edge server, then other edge servers or origin servers are requested to serve the user. A hit in local edge server does direct streaming of videos to the user. The service quality as agreed upon in the SLA is ensured, by a dedicated end-to-end link capacity to serve during miss conditions.

As a result of the geo-dispersed nature of the user groups, the popularity of a video at different local edge servers might show great variance. Careful allocation of server storage, processing capacity and link capacity in Media Cloud help to achieve less delay, abandonment of services, meeting the specified service requirement. The improvement in these key metrics will increase opportunities for service providers to have greater revenues.

4.2 Analytical Model

In this section, we present the analytical model based on queuing theory to derive various key metrics required for decision making in video delivery service of the Media cloud.

The system is modeled as a finite buffer-queueing system with multiple homogeneous VMs using the $M/M/c/N$ queueing model. The number of new video requests is upper bounded by N. There are c number of identical VMs in datacenters (both origin and edge/proxy). The time taken to download the video content can be mapped to service time of the VMs. Whenever a video is requested, it is handled by any of the available VMs. Thus, if the number of requests is less than $c+1$, the queue is empty, owing to the c available VMs serving different requests. When there are at least c but less than N requests within the cloud data centers, $N-c$ end user requests are placed in the queue.

The requests are served in order of their arrival, that is, under the FCFS discipline. The arrival rate is taken to be λ, and the service rate is assumed to be μ. Upon arrival and during waiting, users have no means to estimate queue lengths or progress rate. As their sojourn in queue increases so does the user dissatisfaction. After joining the queue, a user waits a random length of time for service to begin. If service does not begin within this time, the user leaves the queue, a phenomenon better known as reneging or abandonment. The time elapsed before abandonment is assumed to be independently and identically distributed with an exponential distribution of rate α. In our analysis, retrials are ignored, and abandonment is not allowed once a response to a request is initiated. The state transition diagram for the described model is exhibited in Fig 2. Each state is identified by the number of users in the system, and directed arrows represent transitions between the different states. Various notations used are listed in Table 1.

Table 1: Notation Table

Symbol	Definition
c	Set of VMs in a server
n	Number of user requests in the system
N	Capacity of queue
AR	Abandonment Rate
U	Utilization
W	Waiting time in system
$\boldsymbol{r_n}$	Probability of abandonment for n requests
$\boldsymbol{\beta_n}$	Probability of successful service for n requests
$\boldsymbol{\lambda}$	Arrival rate of user requests
$\boldsymbol{\mu}$	Service rate of VMs
$\boldsymbol{\alpha}$	Abandonment rate of customers
$\boldsymbol{P_i}$	State probability when number of requests in queue is i
$\boldsymbol{L_q}$	Mean queue length
$\boldsymbol{W_q}$	Waiting time in queue
$\boldsymbol{E(B)}$	Mean number of busy servers
$\boldsymbol{P(delay)}$	Probability of delay
$\boldsymbol{P_n(Ab)}$	Probability of abandonment for n requests
$\boldsymbol{P(Ab)}$	Probability of abandonment
$\boldsymbol{P_n(Sr)}$	Probability of successful service for n requests
$\boldsymbol{P(Sr)}$	Probability of successful service
$\boldsymbol{P(Bk)}$	Probability of request being dropped

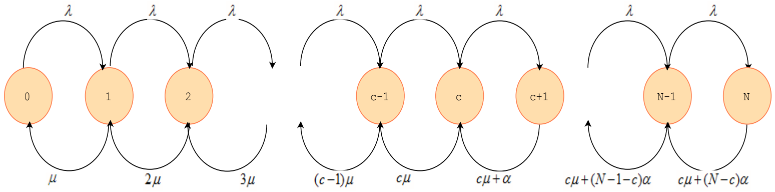


Figure 2: State-transition diagram

The state of the system is described by random variables r_n and β_n, which denote the probability of abandonment and successful service when the number of arrival requests is n.

$$r_n = \begin{cases} 0, & 0 \leq n \leq c, \\ (n-c)\alpha, & c+1 \leq n \leq N. \end{cases} \tag{1}$$

$$\beta_n = \begin{cases} n\mu, & 0 \leq n \leq c, \\ c\mu + (n-c)\alpha, & c+1 \leq n \leq N. \end{cases} \tag{2}$$

Let us denote the state of a system by the subscript i that represents the number of requests in the queue. The respective state probabilities are assumed to be represented by P_i, where $0 \leq i \leq N$. Using one-step transition analysis, the steady state equations can be written as:

$$\lambda P_0 = \mu P_1 \tag{3}$$

$$(\lambda + n\mu)P_n = \lambda P_{n-1} + (n+1)\mu P_{n+1}, \quad 1 \leq n \leq c-1 \tag{4}$$

$$(\lambda + c\mu + (n-c)\alpha)P_n = \left[c\mu + (n-c+1)\alpha\right] P_{n+1} + \lambda P_{n-1}, \quad c \leq n \leq N-1 \tag{5}$$

$$(c\mu + (N-c)\alpha)P_N = \lambda P_{N-1} \tag{6}$$

Using equations (3) to (6) recursively, we get

$$P_n = \begin{cases} \frac{\omega^n}{n!} P_0, & 0 \le n \le c, \\ \frac{\lambda^{n-c}}{\prod\limits_{j=c+1}^{n} [c\mu + (j-c)\alpha]} \frac{\omega^c}{c!} P_0, & c+1 \le n \le N, \end{cases} \tag{7}$$

where $\omega = \lambda/\mu$. The unknown P_0 is computed in equation (8) as

$$P_0 = \left[\sum_{n=0}^{c} \frac{\omega^n}{n!} + \frac{\omega^c}{c!} \sum_{n=c+1}^{N} \frac{\lambda^{n-c}}{\prod\limits_{j=c+1}^{n} [c\mu + (j-c)\alpha]} \frac{\omega^c}{c!} \right]^{-1} \tag{8}$$

Using the stationary state probabilities derived above, we can obtain the following:

- Mean number of busy servers ($E(B)$) in the system over the entire spectrum of permissible states:

$$E(B) = \sum_{n=1}^{c} nP_n + c \sum_{n=c+1}^{N} P_n.$$

- Mean queue length (L_q) of the system when number of new arrival requests is more than c but wihin the upper bound on arrival N :

$$L_q = \sum_{n=c+1}^{N} (n-c)P_n.$$

- Abandonment is undesirable in most service setting because it leads to a combination of lost revenue and ill-will. Average Abandonment Rate (AR) for the system is

$$AR = \sum_{n=c+1}^{N} (n-c)\alpha P_n = \alpha L_q.$$

- Probability of delay among all requests (including abandoned, served or blocked)is

$$P(delay) = \sum_{n=c}^{N-1} P_n$$

The probabilities of occurrence of these events give us an idea about the system performance. It is quite natural that in a finite buffer queueing model if the number of client requests in the system is equal to the threshold N, further client requests for content cannot be accommodated, and shall be dropped. So, $P(Bk)$ is defined as the probability that an arriving request in equilibrium finds N requests in the system. Now, the relationship between $P(Ab)$ and $P(Sr)$ is expressed in equation (9) by:

$$P(Sr) = 1 - P(Ab) \tag{9}$$

We can define $P_n(Ab)$ as the probability that non-blocking jobs abandon given that there are n requests in the system.

$$P_n(Ab) = \frac{r_{n+1}}{c\mu + r_{n+1}} = \begin{cases} 0, & 0 \le n \le c-1, \\ \frac{(n-c+1)\alpha}{c\mu + (n-c+1)\alpha}, & c+1 \le n \le N-1. \end{cases}$$

Since some requests are blocked, and only non-blocked requests can abandon, we calculate P(Ab, non-blocking), the probability of abandonment for non-blocking requests as shown in equation (10).

$$\begin{aligned} P(\text{Ab, non-blocking}) &= \sum_{n=c}^{N-1} P_n(Ab) \cdot P_n \\ &= \sum_{n=c}^{N-1} \frac{(n-c+1)\alpha}{c\mu + (n-c+1)\alpha} P_n \end{aligned} \tag{10}$$

Using the cut-balance equation between state n and $n+1$, we get

$$\lambda P_n = [c\mu + (n-c+1)\alpha] P_{n+1}, \quad \text{for} \quad c \leq n \leq N-1. \tag{11}$$

Using (11) in (10), we have

$$P(Ab) = \sum_{n=c}^{N-1} P_n(Ab) \tag{12}$$

$$= \sum_{n=c}^{N-1} \frac{(n-c+1)\alpha}{\lambda} P_{n+1} = \frac{\alpha}{\lambda} L_q \tag{13}$$

We can define $P_n(Sr)$ as the probability that jobs complete their service successfully given that there are n requests in the system.

$$P_n(Sr) = \begin{cases} 1, & 0 \leq n \leq c-1, \\ \frac{c\mu}{c\mu + (n-c+1)\alpha}, & c+1 \leq n \leq N-1. \end{cases} \tag{14}$$

Using the above, we can easily represent $P(Sr)$ as:

$$P(Sr) = \sum_{n=0}^{N-1} P_n(Sr) \tag{15}$$

$$= \sum_{n=0}^{c-1} P_n + \sum_{n=c}^{N-1} \frac{c\mu}{c\mu + (n-c+1)\alpha} P_n \tag{16}$$

According to Little's Rule, we can easily represent the waiting time in queue W_q as:

$$W_q = \frac{L_q}{\lambda} \tag{17}$$

We can rearrange the equations to get the following relation between $P(Ab)$ and W_q as shown in equation (18):

$$W_q = \frac{P(Ab)}{\alpha} \tag{18}$$

Applying Little's formula to $E(B)$, we get

$$E(B) = \lambda \cdot P(Sr) \cdot \frac{1}{\mu}$$

On rearranging the equation, we get an alternate representation of $P(Sr)$:

$$P(Sr) = \frac{\mu \cdot E(B)}{\lambda}.$$

In general,

$$1 - P_N = P(Ab) + P(Sr) \tag{19}$$

Thus, rate of non-blocking jobs equal to the sum of abandon and service rate in equilibrium and is represented in equation (20):

$$\lambda(1 - P_N) = \alpha L_q + \mu E(B) \tag{20}$$

Now, we can extrapolate the results obtained so far to get the system utilization (U) as shown in equation (21).

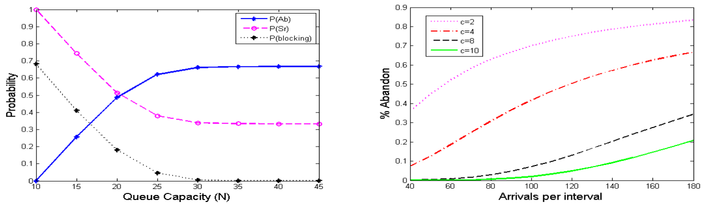
$$Utilization = U = \frac{\lambda \cdot P(Sr)}{c \cdot \mu} = \rho P(Sr) \tag{21}$$

The sum of the waiting time in queue and the time taken for service is the total time spent waiting (W) in the system. It is represented in equation (22).

$$W = W_q + \frac{1}{\mu} = \frac{P(Ab)}{\alpha} + \frac{1}{\mu} \tag{22}$$

5 Numerical Results

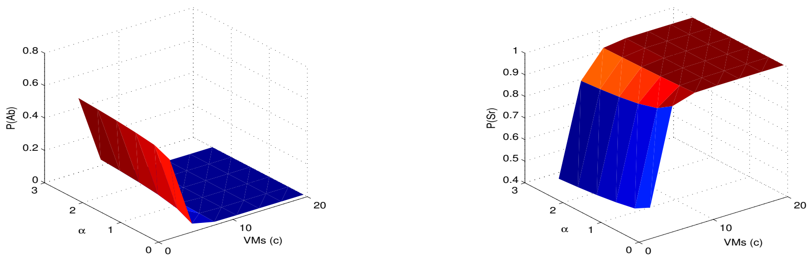
In this section, we evaluate our analytical model through simulation. The detailed design of the simulations and results are specified as follows. We conducted numerical analysis on the proposed model over a wide range of input variables. A suitable difference between the number of available VMs and the request holding the capacity of the system was assumed, to allow for analysis of the queueing behavior of the requests. The values of λ, μ and α were chosen based on similar consideration, to allow for analysis of abandonment among viewers.



(a)

(b)

Figure 3: Effect of Queue Capacity, Arrival Rate and Number of VMs on Blocking, Abandonment and Success probability



(a)

(b)

Figure 4: Dependence of Pr(Ab) and Pr(Sr) on Number of VMs and Abandonment Rate

Figure 3(a) depicts the effect of buffer size (queue capacity) on blocking, abandonment and success rates. For a low value of buffer size, the system reaches maximum serving capacity easily, leading to higher probability of requests getting blocked. Thus, as the buffer size increases, the probability of successful service completion declines. Similarly, the probability that an arriving request is blocked reduces with increase in buffer size.

Figure 3(b) shows the effect of the number of VMs (c) and variation in arrival rate on the rate of abandonment. When the number of active VMs are less, abandonment rate will naturally be high. Initially, the increase in

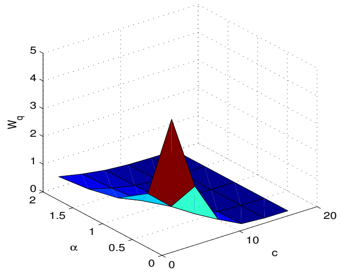


Figure 5: Relationship of Waiting Time, α and c

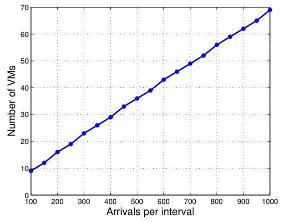


Figure 6: Number of VMs v/s Arrival Rate with constant Abandonment Rate

abandonment rate is linear, but beyond a certain threshold, abandonment rises exponentially. As the value of c increases, the exponential curve flattens out. For higher values of c, there is a brief non-linear increase in abandonment rate followed by an almost linear increase in abandonment. This is indicative of the increased service rate as a result of higher number of active VMs.

The variation in probabilities of abandonment and success due to the number of VMs and abandonment rate has been depicted in Figure 4(a-b). For high values of α and c, the rate of service remains nearly constant. An increase in the number of VMs results in an increased service rate, that is, increased the probability of successful service. The probability of abandonment naturally shows the opposite trend.

Figure 5 explores the impact of the number of VMs and abandonment rate on the waiting time in the queue. When the number of available VMs increases, the service rate rises too, resulting in a lower waiting time. An increase in abandonment signifies a higher number of requests leaving the queue. Thus, the average waiting time in queue decreases.

Figure 6 depicts the relationship between the number of VMs and the arrival rate. This shall enable appropriate choice during system design based on expected traffic. The average service time has been assumed to be 2 minutes. The average patience of customers that is, average time after which abandonment decision might be taken has been assumed to be 5 minutes. For arrival-rate values that vary from 100 to 1000 and a performance target that probability to abandon is less than 3%, the minimum number of VMs needed to adhere to the goal are shown in the figure. It is observed that 9 VMs are needed for 100 arrivals per hour but only 69 (rather than $9 \cdot 10 = 90$) for 1000 arrivals per hour.

6 Conclusion and Future work

In this paper, we analytically derive the impact of buffer size, the number of VMs, request arrival and service rate on various key metrics like user abandonment rate, blocking of a request, successful completion of a request, delay for the service, etc. in video delivery services in Media Cloud. We validated the theoretical model through extensive simulations and found the appropriate number of VMs for a defined set of service and arrival rate to restrict the abandonment rate to a limit. These insights provide operational guidelines to Media cloud service providers to acquire, control and manage resources for video streaming services. There are several avenues for our work. First, we plan to extend the analytical model towards an availability model for Media Cloud. Second, we will try to implement energy model and cost model. Finally, we are interested in deploying a real test bed on top of a Media Cloud architecture for understanding the practical implementation challenges.

References

[1] Cisco Global Cloud Index. Forecast and methodology, 2015-2020 white paper, 2016.

[2] Zheng Li, Karan Mitra, Miranda Zhang, Rajiv Ranjan, Dimitrios Georgakopoulos, Albert Y. Zomaya, Liam OBrien, and Shengtao Sun. Towards understanding the runtime configuration management of do-it-yourself content delivery network applications over public clouds. *Future Generation Computer Systems*, 37:297 – 308, 2014.

[3] Cong Wang, Qian Wang, Kui Ren, and Wenjing Lou. Privacy-preserving public auditing for data storage security in cloud computing. In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. Ieee, 2010.

[4] S Shunmuga Krishnan and Ramesh K Sitaraman. Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs. *IEEE/ACM Transactions on Networking*, 21(6):2001–2014, 2013.

[5] Liang Zhou. On data-driven delay estimation for media cloud. *IEEE Transactions on Multimedia*, 18(5):905–915, 2016.

[6] Yichao Jin, Yonggang Wen, and Cedric Westphal. Optimal transcoding and caching for adaptive streaming in media cloud: An analytical approach. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(12):1914–1925, 2015.

[7] Yichao Jin, Yonggang Wen, and Kyle Guan. Toward cost-efficient content placement in media cloud: Modeling and analysis. *IEEE Transactions on Multimedia*, 18(5):807–819, 2016.

[8] Yu Wu, Chuan Wu, Bo Li, Xuanjia Qiu, and Francis CM Lau. Cloudmedia: When cloud on demand meets video on demand. In *Distributed Computing Systems (ICDCS), 2011 31st International Conference on*, pages 268–277. IEEE, 2011.

[9] Zhangyu Chang and S-H Gary Chan. Video management and resource allocation for a large-scale vod cloud. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 12(5s):72, 2016.

[10] Sudhir N Dhage, Smita K Patil, and BB Meshram. Survey on: Interactive video-on-demand (vod) systems. In *Circuits, Systems, Communication and Information Technology Applications (CSCITA), 2014 International Conference on*, pages 435–440. IEEE, 2014.

[11] Zheng-Huan Zhang, Xiao-Feng Jiang, and Hong-Sheng Xi. Optimal content placement and request dispatching for cloud-based video distribution services. *International Journal of Automation and Computing*, 13(6):529–540, 2016.

[12] Xiaoming Nan, Yifeng He, and Ling Guan. Queueing model based resource optimization for multimedia cloud. *Journal of Visual Communication and Image Representation*, 25(5):928–942, 2014.

[13] Jordi Vilaplana, Francesc Solsona, Ivan Teixidó, Jordi Mateo, Francesc Abella, and Josep Rius. A queuing theory model for cloud computing. *The Journal of Supercomputing*, 69(1):492–507, 2014.

[14] Shuo Liu, Gustavo Chaparro-Baquero, Shaolei Ren, Quan Gang, et al. Workload consolidation for cloud data centers with guaranteed qos using request reneging. *IEEE Transactions on Parallel and Distributed Systems*, 2016.