Multimedia Tools and Applications (2019) 78:6805–6825
https://doi.org/10.1007/s11042-018-6441-3

On the modelling of CDNaaS deployment

Utku Bulkan[1] · Tasos Dagiuklas[1] · Muddesar Iqbal[1]

Received: 12 January 2018 / Revised: 4 July 2018 / Accepted: 20 July 2018 /
Published online: 31 July 2018

Abstract
With the increasing demand for over the top media content, understanding user perception and Quality of Experience (QoE) estimation have become a major business necessity for service providers. Online video broadcasting is a multifaceted procedure and calculation of performance for the components that build up a streaming platform requires an overall understanding of the Content Delivery Network as a service (CDNaaS) concept. Therefore, to evaluate delivery quality and predicting user perception while considering NFV (Network Function Virtualization) and limited cloud resources, a relationship between these concepts is required. In this paper, a generalized mathematical model to calculate the success rate of different tiers of online video delivery system is presented. Furthermore, an algorithm that indicates the correct moment to switch between CDNs is provided to improve throughput efficiency while maintaining QoE and keeping the cloud hosting costs as lowest possible.

Keywords Content delivery network (CDN) · OTT streaming · Live streaming · Online video platform · QoE · User perception · Subjective analysis · Analytical modelling

1 Introduction

Providing customer satisfaction is an essential necessity in online video delivery. According to a recent white paper [27], a small disappointment in initial buffering duration or a relative increase in average stall metrics across a cluster of customers can result in severe drops in the number of subscribers and sharp falls in profits.

✉ Utku Bulkan
bulkanu@lsbu.ac.uk

Tasos Dagiuklas
tdagiuklas@lsbu.ac.uk

Muddesar Iqbal
m.iqbal@lsbu.ac.uk

[1] SuITE Research Group, Division of Computer Science, London South Bank University, 103 Borough Road, London SE1 0AA, UK

Early prediction of bottlenecks throughout the different steps of the online video delivery system is the key to prevent poor user QoE [12]. However, figuring out what might be causing a degraded performance on a complex association of peripherals and service layers is reasonably a difficult challenge.

Traditional end to end video service consists of several different components [1]; consumer hardware & software, load balancer mechanisms, switches, routers, access network elements such as fiber dslams (Digital subscriber line access multiplexer), base stations, cloud computing instances (e.g. virtual network functions) and edge cache nodes. Building and maintaining such multi-tier systems require a significant amount of investment [9] which can be described as Capital Expenditure (CAPEX) and Operating Expenditure (OPEX). Apart from that, traditional proprietary network peripherals are far from being agile and flexible in terms of scalability. Whenever technology trends change and update, costly hardware upgrades are required to meet the demands of imminent throughput and scalability. The interaction of these components is illustrated in Fig. 1.

However, from a content provider's perspective, the task is to deliver service to their subscribers in different geographical regions. The idea of owning a network is generally not preferred and current market trends have a tendency to simply purchase the processing power and CDN capability as Platform as a Service (PaaS) which provides flexibility, agility on scalability and service volume [24].

As a result of this demand and with the advance of new concepts such as 5G enablers [15], operators are going to provide Network Function Virtualization (NFVs) and Software Defined Networks (SDNs) as a network service to content providers. According to the widespread conventional misperception, 5G systems are not just an increase in communication bandwidth and better coverage. In this environment, service orchestrator has an abstracted view of computing resources, virtualization of the network functions and SDNs including edge computing, peer to peer (P2P) communication, and ultimately CDN as a service (CDNaaS) as discussed in [29].

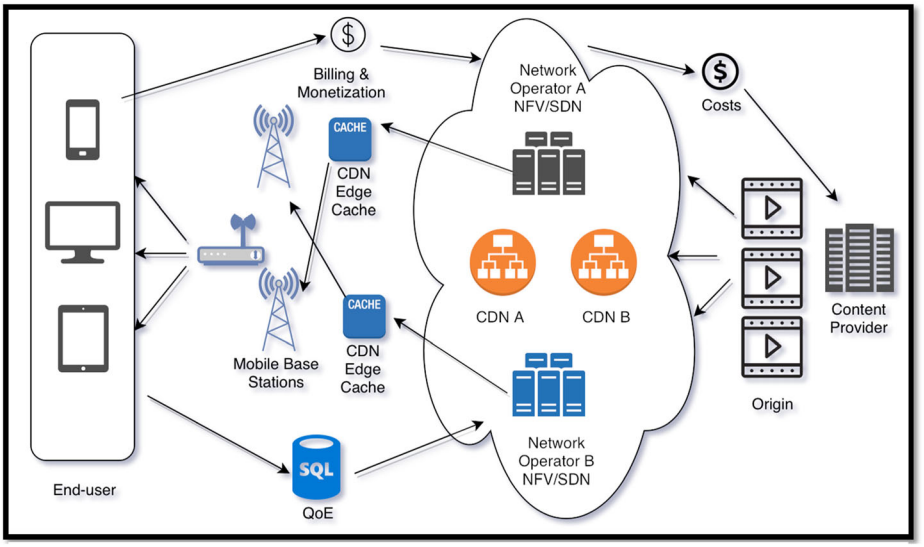


Fig. 1 The tiers of online video delivery

The orchestration of SDN/NFV will be the service that is going to be purchased by the content provider from a network operator. The need for the transcoding and accessing the Over the Top (OTT) content will be handled by the network operators through applications that run on virtual machines that are available through their NFVs and streamed through the SDNs.

The need for scalability of the service is one of the major concerns for any service provider. The solution to that concern will be load balancing and to introduce intermediate and edge cache nodes [10] where the capability of a CDN can be conveyed from origin to "cost and distance efficient" cloud nodes providing an increased service capability to additional collection of subscribers.

The aim of this paper is to provide a model in order to estimate cloud-based CDN deployment and QoE provisioning. This is achieved through a model representing the tiers of an actual streaming service, where live streaming and real-time transcoding take place. A model for NFV and cloud resources is provided to estimate and maximize the performance of the video service while relating it to the user QoE. The constraints for network operators include deployment cost and QoE evaluation.

Conclusively, the mathematical part introduces the idea of deriving whole system performance based on the single user QoE and evaluates NFV and cloud QoE.

The remainder of this paper is structured as follows: Section 2 presents a brief outline of state–of-the-art in CDNaaS Technologies, Section 3 discusses related works, Section 4 illuminates NFV/SDN Ecosystem Architecture and Section 5 provides mathematical analysis of the deployment. Section 6 clarifies cloud related resource and cost constraints and Section 7 formulates the algorithm for decision process of multi-CDN switching system and Section 8 debates simulation results. Finally, in Section 9, conclusions and future works are presented.

2 State of art in CDNaaS technologies

For a video delivery system, user's QoE is unquestionably the particular attention argument for both Content Provider and Network Operators. However, bringing the best service to users using cloud technology has a corresponding cost. Most of the time, business cost forecasts take cloud computing expenses as the main parameter for the selection of the CDNaaS. Currently in the market, there is a selection of major solutions to fulfill cloud CDN requirement for video services [3, 4, 16]. The following section provides state of the art cloud-based solutions to deploy CDNaaS.

2.1 Amazon web services & amazon cloudfront

As a frontrunner in cloud technologies, Amazon provides a global content delivery service that securely distributes video with low latency and high availability [5]. Amazon Elastic Compute Cloud (Amazon EC2) is a flexible service [4] that provides scalable computing capacity in the cloud. Additionally, Amazon EC2 On-Demand [6] offers competing cloud pricing for unpredictable demand which fine-tunes video service capacity to meet demand on peak epochs. Integrated with AWS and directly connected with hundreds of end-user ISPs, CloudFront [5] offers regional edge cache locations as part of standard offering to ensure consistently high cache hit ratios across the globe.

2.2 Google cloud CDN

Google provides Google Cloud Platform [17] with caches at more than 80 sites across the globe which guarantees that cloud CDN is always accessible by the end-user even if the users are geographically distributed. Cloud CDN charges individually for cache fill, https lookup requests and cache invalidation [16].

2.3 Akamai CDN

Akamai hosts more than 200,000 servers in over 130 countries to get uninterrupted customer experiences. According to Akamai's point of view [3], a start-up or a global media giant, all customers are treated as premium, independent of their size. Yet, the services widely known as expensive when compared to other CDN suppliers.

Although these commercial CDN solutions provide quite sophisticated caching and distribution algorithms [17, 23], there might still be performance limitations [8] due to the edge CDN node proximity to clusters of users in some geographical regions that are distant from major communities during peak demand periods [25]. Therefore, there is a continuous demand to determine CDNaaS capacity so that QoE lies within certain bounds.

3 Related work on QoE for online video delivery

Understanding the impact of QoE in a CDN ecosystem stands as a prerequisite for having a successful content delivery deployment. Unless users' experiences are represented through objective metrics which are based on video player statistics, an inductive approach to formulize performance of a CDN cannot be possible.

In this part, research and academic works related to QoE and its relationship with tiers of online video delivery systems are going to be discussed. In ITU-T P.1203.3 recommendation [18], a media session quality score is formulated based on number of stalls, total stall duration, buffering duration, media length and compression quality.

$$\mathrm{SI} = e^{-\frac{numStalls}{s1}} . e^{-\frac{\left(\frac{totalbuf}{T}\right)}{s2}} . e^{-\frac{\left(\frac{bufdur}{T}\right)}{s3}} \tag{1}$$

M. Knoll et al. has provided a Mean Opinion Score model for OTT services [21], where x stands for number of stalls and t for time since the last stall and a for the memory parameter (which was set as 0.14).

$$\mathrm{MOS} = e^{-\frac{x}{3}+1.5-a\sqrt{t}} \tag{2}$$

However, these equations [18, 21] do not reflect a real-time explanation of the performance of a peripheral of an online video system. In this work, by using these QoE Eqs. (1) and (2), a methodology is proposed to determine system wide QoE.

In a recently published article [11] by V. D'Amico, an architecture overview for a SDN/NFV telco operator platform for video broadcasting is validated the proof of concept for SDN testbeds. Yet in this work, the impact of QoE on SDN/NFV has not been investigated from a content provider point of view.

A. Ahmad et al. has presented [12] a collaborative approach among OTTs and ISPs where they have modelled a QoE driven approach for solving the resource sharing problem while several OTT applications use the same ISPs network peripherals. F. Z. Yousaf et al. proposed [30] a new network slicing aware orchestration framework with flexible network function control system while introducing a consistent QoS/QoE management framework. H. Koumaras et al. have developed a testbed [22] which orchestrates SDN/NFVs while providing real-time transcoding and capability to monitor NFV load and QoE levels.

Although these works [12, 22, 30] provide a good understanding on orchestration for QoE management frameworks, they still lack the impact of cloud CDN cost analysis which is one of the distinctive reasons of choosing a particular CDN supplier. G. Faraci et al. has provided a system model for 5G Operator Network Telco [13] which provides the interactions between core & edge cloud and NFVs running on physical nodes. The work proposes a simulative tool for 5G systems, which is able to detect delay events resulting from NFV load on physical device CPUs or transmission loads between nodes. Yet, the QoE impact on NFV has not been explicitly presented.

Z. Frias et al. have argued [14], the policy discussions of anything as a service in the infrastructure layer and potential of future 5G networks to provide network capabilities to third parties through an Application Program Interface (API). This will provide existing infrastructure capabilities available to any company who purchases the product like a pay as you go service. These research works [14, 19] provide a good understanding of cost profiles for CDN and cloud resources, though they lack a mathematical model to bring a methodology to decide for switching between CDNs.

Unlike these research works [11, 12, 18, 21], in this paper, a real-time understanding of QoE will be presented and its impact on NFV and cloud resources is going to be formulized. Furthermore, a mathematical analysis is presented which the basis for a decisive mechanism will be targeted for content and service providers to support multi-CDN capability that will bring a solution for the QoE/Cost/Cache success rate optimization question.

4 NFV/SDN ecosystem architecture and workflow diagram

The diagram at Fig. 2 illustrates the workflow for establishing CDNaaS through interactions among network operators, SDN/NFV orchestration and origin/CDN/Edge subsystems while observing online video delivery concept from a content provider point of view. The procedure starts as the subscribers access the content which triggers warming up (content pre-loading) and caching at intermediate and edge nodes. During the demand for the content, cost for cloud resources, system capacity and scalability are monitored. SDN/NFV and CDN and their impact on QoE is continuously estimated via trained models. The fluctuations in number of customers are considered and appropriate alteration between CDN providers are elicited to achieve three key elements; efficient deployment, customer satisfaction and cost reduction (Table 1).

5 Mathematıcal analysis

This section presents a model for the deployment of CDNaaS. Step by step, each layer; end-user, SDN, CDN, origin and NFV will be reflected according to their impact on the throughput

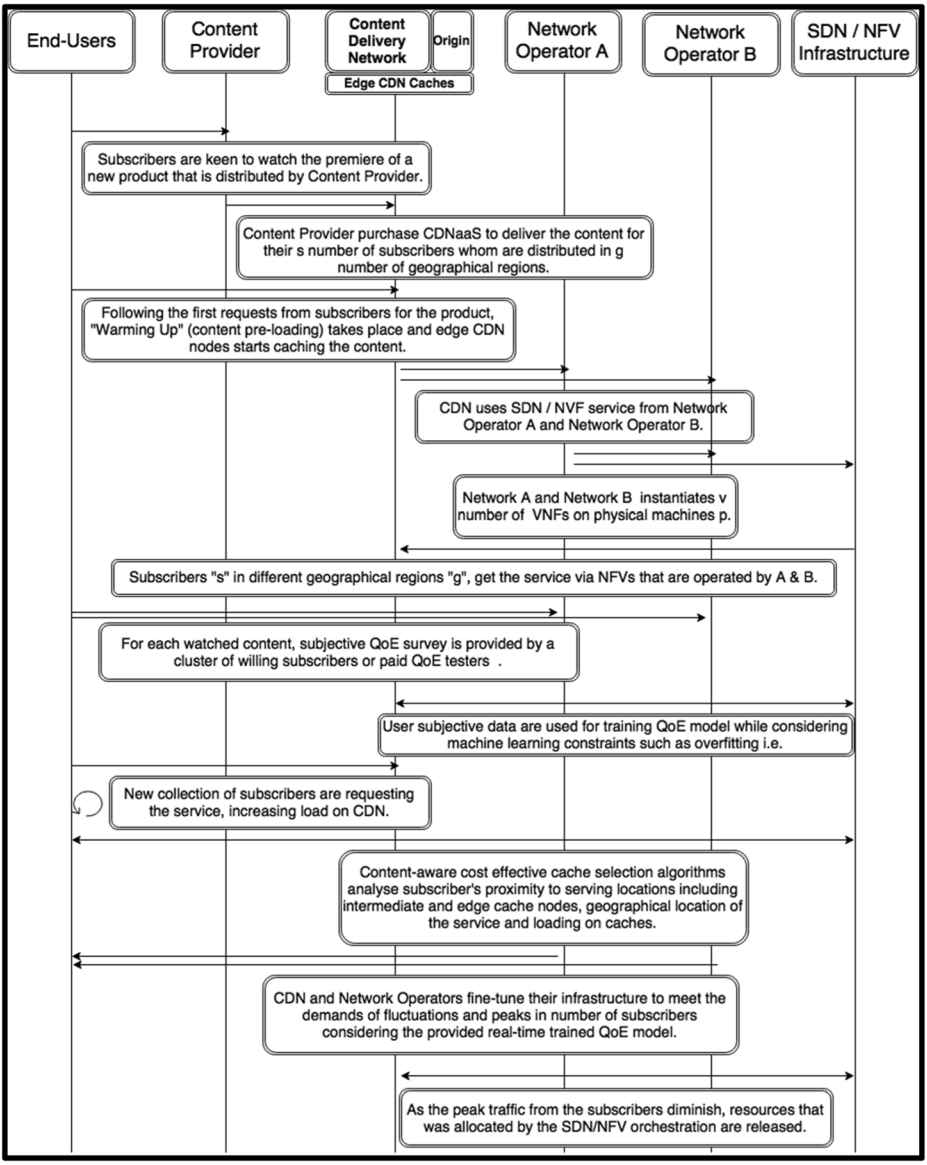


Fig. 2 Platform diagram for NFV, SDN and online video delivery

and latency of the entire system. Also, the cost function will be evaluated for content provider and network operator. This will eventually provide a decision mechanism to offer optimal values for both QoE while keeping expenditures of the service as lowest as possible.

5.1 End-user's QoE

The starting block of the conclusive evaluation, "single user's QoE" from v∈V running on m∈M (where v refers to single Network Function Virtualization, V represents all NFV clusters,

Table 1 List of Notations

Notation	Meaning
$Q_u(v, m)$	Single user's QoE from v∈V running on m∈M
T_{dur}	User's total watch duration for the content
I_f	Number of displayed frames
I_d	Number of dropped frames
B_{rate}	Average bitrate of the stream
S	Number of stalls
S_{dur}	Time spent during stalls
$t_{latency}$	Initial content buffering duration in seconds
u ∈ U	Single user, element of all users
υ ∈ V	Network Function Virtualization
m ∈ M	Physical Servers
g ∈ G	Geographical area cluster
γ	Total external traffic that is carried by a network
γ_{jk}	Poisson process on origin node j headed to node k
λ	Total amount of internal traffic carried by a network
λ_i	Traffic carried by each network peripheral
h	Number of hops in internal network
L	Mean Latency of all messages in a network layer
L_i	Latency of a single network function
L (x2, x1)	Latency between two layers in a system
B (x2, x1, t)	Realtime throughput that is served from x2 to x1
R (u_i, g_i, t)	Traffic requested by user u_i located in g_i at instant t
S (e)	Number of content that is stored on edge nodes
p (u, S(e))	Probability of cache existence on edge nodes for u∈U
C_M (m)	CPU processing capacity of physical server m∈M
C_C (m)	Memory capacity of physical server m∈M
C_M	Total CPU capacity of the network operator
C_C	Total memory capacity of the network operator
R_M (v, m, u, t)	Required memory for NFV that runs on m∈M to serve user u∈U at instant t
R_C (v, m, u, t)	Required CPU power for NFV that runs on m∈M to serve user u∈U at instant t
P_C	Unit cost for unicasting a content to user u∈U
P_S	Storage expense on a CDNaaS

m refers to physical machine and M represents all physical machine cluster) is represented as a function of following statistics T_{dur} (user's total watch duration for the content), B_{rate} (average bitrate of the stream), S (number of stalls), $S_{duration}$ (time spent during stalls), $t_{latency}$ (initial content buffering duration).

$$Q_u(v, m) = \mathrm{Q}\left(T_{dur}, Brate, S, S_{dur}, t_{latency}\right) \tag{3}$$

Figure 3 illustrates the interactions of the Tiers for an online video delivery system that consists of Origin, CDN, Edge CDN Nodes and NFV instances. Latency and throughput between any two tiers is given with "$L(x_1, x_2)$" and "$B(x_1, x_2)$" correspondingly.

Storage capacity of any peripheral is denoted by" S(x)". $R_M(v, m)$ and $R_C(v, m)$ denotes required memory and computational power accordingly for a virtual machine "v" to run on a physical machine "m". $C_M(m)$ and $C_C(m)$ express the total memory and computational capacity of physical machine "m". N(v, m) provides the number of virtual machines running on physical machine "m". "x" values in functions L(), B() and S() can be substituted for any of the layers, end-users (u), network peripherals (p), edge cache node (e), CDN (c), origin (r).

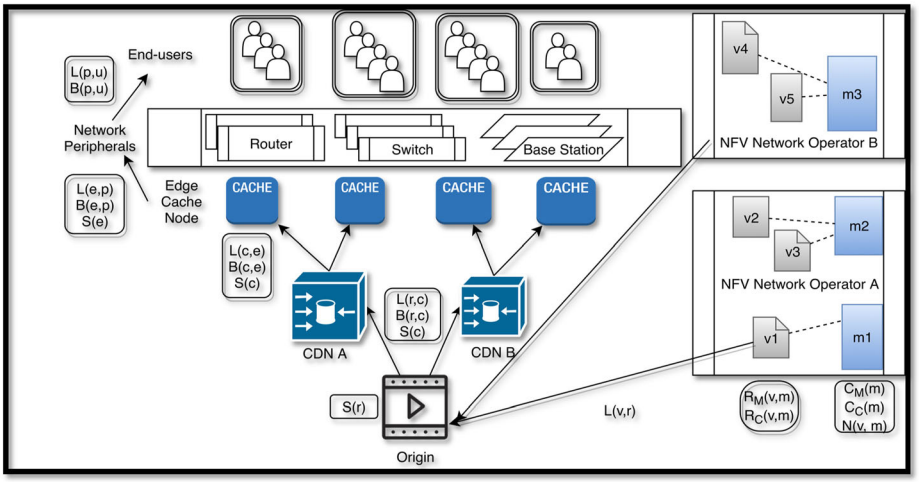


Fig. 3 CDNaaS system model

5.2 Network layer, goodput and latency

Kleinrock [20] has formulated the total external traffic that is carried by a network as γ in Eq. 4 where γ_{jk} is the messages arrive from a Poisson process on origin node j headed to node k;

$$\gamma = \sum_{j,k=1}^{N} \gamma_{jk} \tag{4}$$

There have been many works [2, 14] which states that, video flows follow Poisson process. Internal traffic carried can be described as Eq. 5 where λ_i refer to the traffic carried by each peripheral.

$$\lambda = \sum_{i=1}^{N} \lambda_i \tag{5}$$

The number of hops h to transmit a message in internal network is declared as Eq. 6.

$$h = \lambda/\gamma \tag{6}$$

And finally, L is the mean delay of all messages across a cluster of network layer (where L_i is the delay of a single network function) as Eq. 7;

$$L = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} L_i \tag{7}$$

For the video delivery, the average latency term L (x_1, x_2) can be generalized as the latency between any two layers.

$$L(x_2, x_1) = \sum_{i=1}^{M} \frac{\lambda_{x_{1,2}i}}{\gamma_{x_{1,2}}} L_{x_{1,2}i} \tag{8}$$

In general, describing latency for each node for each physical device as L_i is insurmountable. For simplicity, the latency between end-user and edge cache nodes is defined as the sum of the latency between each consecutive layer, from edge to network and from network to user.

$$L(u, e) = L(u, p) + L(p, e) \tag{9}$$

Using a similar approach, the real-time throughput B(u, p, t) that is served to a cluster of users u∈U distributed in g∈G geographical regions can be defined as Eq. 10, where R(u, g, t) is the amount of traffic that is requested by the user u located in g at instant t. Summation in Eq. 10, traverses through all ∀U (for all users) that are located in ∀G (for all geographical regions) which corresponds to the total throughput of the platform B(u, p, t).

$$B(u, p, t) = \sum_{j=1}^{G} \sum_{i=1}^{U} R\left(u_i, g_j, t\right) \tag{10}$$

5.3 CDN and online caching

Contrary to a common misconception, CDNs are not only big hybrid database-network like entities that hold the video data and unicast to subscribers like a traditional single server service. In fact, CDNs provide highly sophisticated caching mechanisms [2] to carry out edge computing functionality and allocate copies of the content over their geographically distributed edge nodes [18]. Whenever there is a demand for the service of any content, edge cache nodes are activated through a process called “warming” [23] where copies of the content are cached from origin to intermediate and eventually towards the edge nodes. M. Ruiz et al. have discussed CDN optimization problem [26] while minimizing the CDN costs by dynamically reconfiguring the CDN. Also, T. M. K. Roeder et al. have discussed [25] the optimization of ISPs and CDN collaboration through cache miss simulations. Nevertheless, none of these works provided a generalized formulation for CDN latency and bandwidth approximation. Though, Z. Chen et al. have approached the question as a cache-aided throughput calculation [8] where a self-request throughput without cache aid can still be modelled as sum of all requests as it was described in Eq. 11. Still, there is a chance that the requested content has already been cached in edge nodes. An additional probability component for cache collision will be representing edge and intermediate cache existence. This is expressed with term p (u, k, S(e)) where u is the number of users, k is the cache miss performance and S(e) is the total number of content that can be stored on edge nodes, eventually storage capacity of edge CDN. As the number of users u∈U requesting content increases, with a better cache capable CDN (with a high cache miss performance k index), the p(u,S(e)) value will be less than 1 resulting Eq. 12 to have a lower value. If the CDN has a low cache miss performance, then p approximates to 1 which will be equivalent to a non-caching capable CDN.

As an effect, the cache-aided throughput depends on the throughput that is demanded from origin to edge nodes.

$$p(u, S(e)) = 1 - e^{k\frac{-u}{2S(e)}} \tag{11}$$

$$B(r, e, t) = \sum_{j=1}^{G} \sum_{i=1}^{U} R\left(u_i, g_j, t\right)\ p(u, S(e)) \tag{12}$$

Latency for CDN can be modelled as the sum of latencies from origin to edge nodes via core CDN. Without loss of generality, latency from origin to core CDN is neglected [30] as latency is not critical in core networks [23]. Latency from core CDN to edge CDN nodes can be formulized as the probability of cache existence on edge nodes multiplied by the cost of number of hops through intermediate nodes. In Eqs. 13 to 15, the tiers of online video delivery system are represented as following: edge cache node (e), CDN (c), origin (r).

$$L(e, r) = L(e, c) + L(c, r) \tag{13}$$

$$L(c, e) = \sum_{i=1}^{M} \frac{\lambda_i}{\gamma} L_i\ p(u, S(e)) \tag{14}$$

$$L(r, c) \approx 0 \tag{15}$$

5.4 NFV resource analysis and impact on QoE

Each physical server m∈M has a capability to run "N (v, m)" number of NFVs where v∈V. The amount of required CPU resource for NFV that runs on m∈M to serve user u∈U at instant t is R_C (v, m, u, t). The amount of required memory for NFV that runs on m∈M to serve user u∈U at instant t is R_M (v, m, u, t). Physical server m∈M have a maximum CPU processing capacity of C_C(m) and memory capacity of C_M (m). Total CPU capacity of the network operator is C_c and total memory capacity of the system is C_m.

$$\mathrm{C}_M(\mathrm{m}) = \sum_{i=0}^{N(v,m)} R_M(v, m, u_i, t) \tag{16}$$

$$\mathrm{C}_C(\mathrm{m}) = \sum_{i=0}^{N(v,m)} R_C(v, m, u_i, t) \tag{17}$$

$$\mathrm{C}_M = \sum_{i=0}^{M} \sum_{j=0}^{N(v,m)} R_M\left(v, m_i, u_j, t\right) \tag{18}$$

$$\mathrm{C}_C = \sum_{i=0}^{M} \sum_{j=0}^{N(v,m)} R_C\left(v, m_i, u_j, t\right) \tag{19}$$

Real-time transcoding capability and the performance of a live streaming system have impact on end-users QoE [14, 25]. Any disruption or shortage of resources results directly deterioration of service quality. Single user's QoE is described with Eq. 1. The cluster of users u∈U that are getting service from the NFV υ∈V and this NFV runs on a physical machine m∈M. The QoE for υ can be defined as Eq. 20 where a homogeneous distribution is assumed across the subscriber privileges that results in guaranteed equal service reliability per each user.

$$Q_{\upsilon}(t) = \sum_{u=1}^{U} \frac{Q_u(v,m,t)}{U}, u \in \mathrm{U} \tag{20}$$

Network operator's QoE for NFV capacity is the sum of the QoE of NFVs ∀υ ∈V.

$$\mathrm{Q}_N(t) = \sum_{\forall \upsilon \in V}^{N(v,m)} \sum_{\forall u \in U}^{U} \frac{\mathrm{Q}_u(v,m,t)}{U} \tag{21}$$

However, QoE is a subjective quality measurement, rather than considering the delta between $Q_N(t_1)$ and $Q_N(t_2)$ at two different moments, provides a better comparable understanding of the service quality.

$$\Delta Q_N = Q_N(t_2) - Q_N(t_1) \tag{22}$$

5.5 Cost of the operation

There are three main expense estimation arguments regarding the CDNaaS cost modelling; storage, latency and computational goodput.

Most of the cloud services [4, 6, 17] advertise discount rates in different tiers proportional to the requested processing power capacity. According to this assumption, storage expenses of the system, P_S can be modelled as an inverse proportional function where P_{Cg} is the unit cost for unicasting content to user u∈U living in region g∈G.

$$P_S = \sum_{\forall g \in G}^{G} \sum_{\forall u \in U}^{U} \frac{P_{Cg}}{u} \tag{23}$$

6 Cloud Resource and Cost Constraints

Primary motivation of an online video delivery platform is to provide QoE at highest rate possible while keeping the expenses of cloud resources, NFV and SDN at minimum. Following constraints are necessary for fulfilling the constraints of an ideal delivery system;

The throughput capability of Network layer must be greater than the requested traffic by user u ∈ U.

$$B(p,u,t) > \sum_{j=1}^{G}\sum_{i=1}^{U} R\left(u_i, g_j, t\right) \tag{24}$$

The average service latency is formulized in Eq. 23 from NFV v∈V to users u∈U via origin (o), CDN(c), edge CDN(e), SDN(p) must be less than ε.

$$L(u,r) = L(u,p) + L(p,e) + L(e,c) + L(c,r) + L(r,v) \approx 0 \tag{25}$$

CDN throughput must be greater than the user requests multiplied by the probability of existence of the cache of the content in CDN.

$$B(r,e,t) > \sum_{j=1}^{G}\sum_{i=1}^{U} R\left(u_i, g_j, t\right) p(u, S(e)) \tag{26}$$

Total CPU and memory capacity of the physical servers must be greater than the amount of required CPU resource for NFV that runs on m∈M to serve user u∈U at instant t is R_C (v, m, u, t) and the amount of required memory for NFV that runs on m∈M to serve user u∈U at instant t is R_M(v,m,u,t).

$$C_M > \sum_{i=0}^{M} \sum_{j=0}^{N(v,m)} R_M(v,m,u,t) \tag{27}$$

$$C_C > \sum_{i=0}^{M} \sum_{j=0}^{N(v,m)} R_C(v,m,u,t) \tag{28}$$

And ultimately, the QoE of the system Q_N(t) must be kept at maximum, while minimizing the cost of the system P_s.

$$\max \sum_{\forall v \in V}^{N(v,m)} \sum_{\forall u \in U}^{n} \frac{Q_v(v,m,t)}{n} \tag{29}$$

$$\min \sum_{\forall g \in G}^{G} \sum_{\forall u \in U}^{U} \frac{P_C}{u} \tag{30}$$

7 Decision of Multı-Cdn switching system

This section presents a selective algorithm to decide when to switch between CDNs regarding the QoE/Cost/Cache success rate optimization problem which grounds the mathematical

analysis that was presented in the previous section. Equation 34 provide the association of P_C & goodput to $Q_N(t)$ and in a given window within the expected budget, goodput can be estimated for the expected $Q_N(t)$ for the service.

ALGORITHM I
MULTI-CDN SWITCH

PREREQUISITES: LIST OF AVAILABLE CDNS, UNIT COST FOR CDN PC, CDN CACHING PERFORMANCE INDEX K, $U \in U$, $\upsilon \in V$.
1. COMPUTE RATIO OF $\mu = B(U, P, T) / \sum_{i=1}^{U} R(U_I, G_J, T)$.
2. WHILE $\mu \approx 1$,
3. ESTIMATE $\Delta Q_N = Q_N(T_2) - Q_N(T_1)$ THROUGH EQ (1) AND EQ (20).
4. CALCULATE CACHE SUCCESS RATE $P(U, S(E))$ AND EFFICIENT USE OF EDGE CDN NODES.
5. SUM UP TOTAL EXPENSES OF CDN VIA EQ (21) P_S.
6. ON FIGURE 9, SOLVE OPTIMIZATION PROBLEM AND TRACE LOCAL MAXIMUM FOR $Q_N(T)$ AND LOCAL MINIMUM FOR P_C THROUGH THE PROVIDED EQ. 34 WITHIN THE GIVEN BUDGET WINDOW.
7. COMPARE OTHER CDN PERFORMANCE CURVES, AND SWITCH TO MORE EXPENSIVE BUT EFFICIENT CDN.
8. END WHILE.

8 Simulation results

Without loss of generality, we assume that the content owner associates with three different CDN operators and intends to operate within a multi CDN environment where most optimum choice is to maximize QoE and minimize costs while establishing an agile, scalable and flexible network.

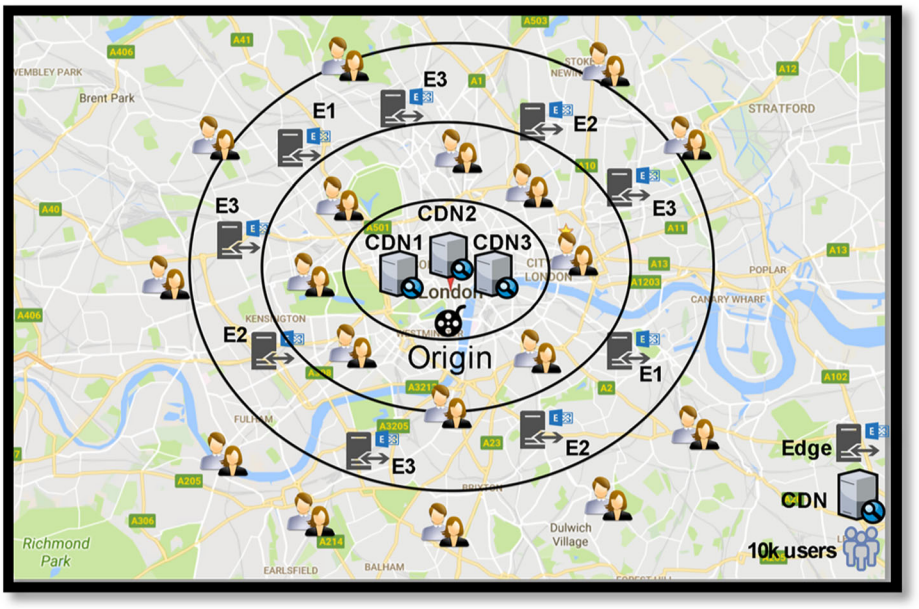


Fig. 4 Distribution of users, CDNs and Edge Nodes across the city for the scenario. Each group of people in the map stands for 10 k users and edge machines represents 10VMs (Virtual Machine) serving as Edge CDN nodes

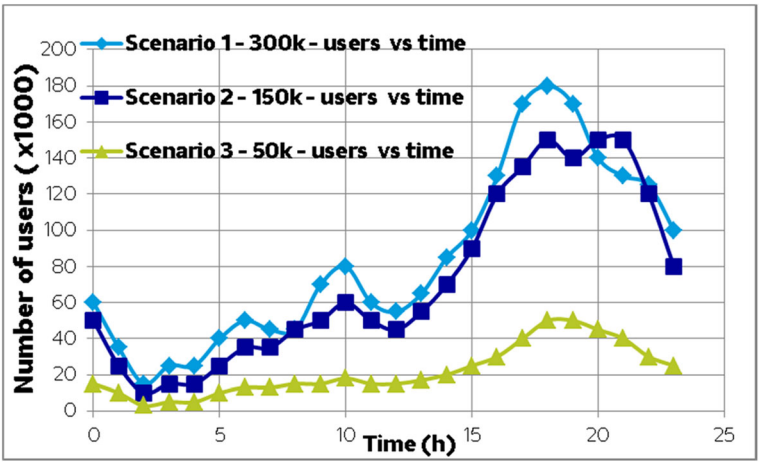


Fig. 5 Number of users vs Time in a 24 h online streaming session [7]. The number of users varies between day and evening, and reaches a peak of 50 k, 150 k and 300 k on three different scenarios at prime time

Based on real-life data [7] that is originated from Broadcasters' Audience Research Board (BARB), with three different scenarios 50 k, 150 k, 300 k users intend to use the service across the city and there are three different CDNs available with different edge, caching capabilities and conclusively costs.

Primary objective of this simulation is to detect the changes in user demand and formulize a methodology to solve the QoE, CDN cost and NFV performance optimization problem (Figs. 4, 5 and 6).

8.1 Scenario parameters

Parameters related to the scenario are as following: geographical regions G = 5, all users U = {20 k, 40 k, 70 k, 30 k, 20 k}, S(e) = 10 live channels, $P_{Cg\text{-}CDN1} = 0.1£$, $P_{Cg\text{-}CDN2} = 0.15$

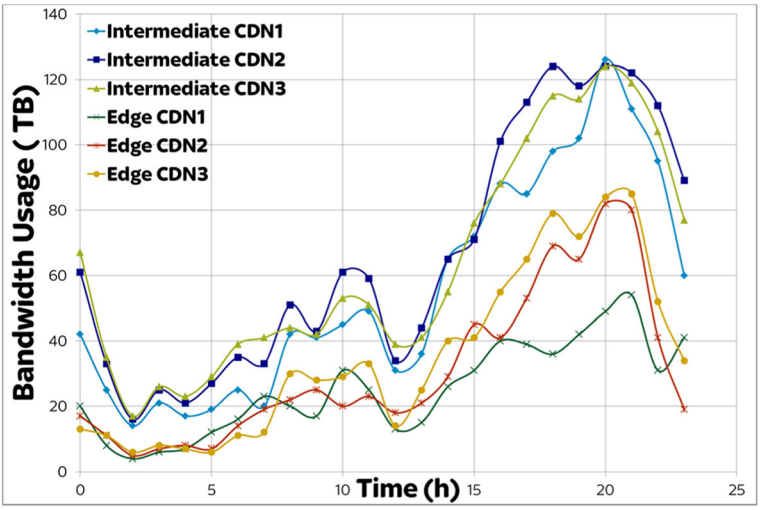


Fig. 6 Bandwidth usage at both intermediate and edge vs time

£ and $P_{Cg\text{-}CDN3} = 0.17$£ per Gigabyte. Cache usage performance of CDN1 is k = 0.03, for CDN2 k = 0.02 and for CDN3 k = 0.15 where number of VMs running for CDN1 is 3, for CDN2 is 5 and for CDN3 is 7. C_M required memory for CDN1 transcoding operations is 100 gb, for CDN2 is 150 gb and for CDN3 is 170 gb computational power (required VM runtime) is 100 h for CDN1, 80 h for CDN2 and 65 h for CDN3 to supply a 24-h streaming activity.

The change in the number of users causes an impact on the goodput of the intermediate and edge nodes. At peak times, the load balancers try to redirect the users to edge computing nodes, however as CDNs differ in terms of caching capability and success, their performance also varies on different circumstances. In terms of CDN deployment, three different scenarios have been considered with different properties in terms of caching quality and cost efficiency.

8.2 Cost-efficient CDN

A cost-efficient CDN provides an acceptable yet an intermediate service quality with adequate scalable and distributed caching capability. A non-zero, fluctuating average latency (L(u,r) > 0) and deficient computational power, NFV memory and throughput to meet user requests on peak demand durations are typical. Yet, the costs of these cloud services are budget friendly when compared to high end CDN services. For a cost efficient CDN, QoE is not the primary concern, yet, delivery is optimized to provide the best within available system resources.

$$\begin{gathered} L(u,r) > 0 \\ B(r,e,t) < \sum_{j=1}^{G}\sum_{i=1}^{U} R\left(u_i, g_j, t\right) p(u, S(e)) \\ \mathrm{C}_M < \sum_{i=1}^{M}\sum_{k=1}^{N(v,m)} R_M(v_k, m_i, u, t) \\ \mathrm{C}_C < \sum_{i=1}^{M}\sum_{k=1}^{N(v,m)} R_C(v_k, m_i, u, t) \\ \mathrm{Min}\sum_{\forall \upsilon \in \mathrm{V}}^{N(v,m)}\sum_{\forall \mathrm{u} \in \mathrm{U}}^{U} \frac{\mathrm{Q}_u(v,m,t)}{U} \\ \mathrm{Low}\sum_{\forall \mathrm{g} \in \mathrm{G}}^{G}\sum_{\forall \mathrm{u} \in \mathrm{U}}^{U} \frac{P_{Cg}}{U} \end{gathered} \tag{31}$$

8.3 Average Cost CDN

The primary attitude of an average cost CDN is to offer a semi-premium equivalent service while still being in a budget friendly fashion. Stalls and buffering incidents during watch sessions are intermittently observed and performance of video delivery system generally depend on mobile or fiber network performance. Users are commonly content with the received service and aptly, average perceived QoE is better than cost-efficient CDNs.

$$\begin{gathered} L(u,r) \approx 0 \\ B(r,e,t) < \sum_{j=1}^{G}\sum_{i=1}^{U} R\left(u_i, g_j, t\right) p(u, S(e)) \\ \mathrm{C}_M \gtrsim \sum_{i=1}^{M}\sum_{k=1}^{N(v,m)} R_M(v_k, m_i, u, t) \\ \mathrm{C}_C \gtrsim \sum_{i=1}^{M}\sum_{k=1}^{N(v,m)} R_C(v_k, m_i, u, t) \\ \mathrm{Average}\sum_{\forall \upsilon \in \mathrm{G}}^{N(v,m)}\sum_{\forall \mathrm{u} \in \mathrm{U}}^{U} \frac{\mathrm{Q}_u(v,m,t)}{U} \\ \mathrm{Average}\sum_{\forall \mathrm{g} \in \mathrm{G}}^{G}\sum_{\forall \mathrm{u} \in \mathrm{U}}^{U} \frac{P_{Cg}}{U} \end{gathered} \tag{32}$$

8.4 Expensive CDN

A typical expensive CDN guarantees a very low (or zero) latency throughout its network and edge nodes. Distributed, scalable and durable VM execution and cloud service capability meets the demand from the users at all times to perform transcoding, caching and storage facilities. User QoE and evaluated NFV, SDN and cloud QoE generally gives high satisfactory performance and results. Obviously, they require more budget to operate when compared to cost-friendly CDNs.

$$\begin{gathered} L(u,r)\approx 0 \\ B(r,e,t) < \sum_{j=1}^{G}\sum_{i=1}^{U} R\left(u_i, g_j, t\right) p(u, S(e)) \\ \mathrm{C}_M > \sum_{i=1}^{M}\sum_{k=1}^{N(v,m)} R_M(v_k, m_i, u, t) \\ \mathrm{C}_C > \sum_{i=1}^{M}\sum_{k=1}^{N(v,m)} R_C(v_k, m_i, u, t) \\ \mathrm{Max}\sum_{\forall \upsilon \in \mathrm{V}}^{N(v,m)}\sum_{\forall \mathrm{u} \in \mathrm{U}}^{U} \frac{\mathrm{Q}_u(v,m,t)}{U} \\ \mathrm{High}\sum_{\forall \mathrm{g} \in \mathrm{G}}^{G}\sum_{\forall \mathrm{u} \in \mathrm{U}}^{U} \frac{P_{Cg}}{U} \end{gathered} \tag{33}$$

There are several different configurations available for any CDN deployment where sensitivity of edge cache nodes can be tuned which will result in reduced CDN costs. In these cases, the demand for throughput from users might not be fulfilled at all times. Cost efficient CDN setups may cause degradation on content delivery quality and eventually system QoE. In today's world, all operators prioritize their users' throughput by taking into account their subscription and geographical location to minimize operational costs.

Cost efficient CDN1 uses fewer number of edge nodes and as the number of users increase, intermediate cache tries to serve the increased demand. However, as the CDN1 nodes have limited capability and cannot scale as well as CDN2 and CDN3, the bandwidth requested by the users is not met. This causes degradation on QoE as presented on Fig. 7.

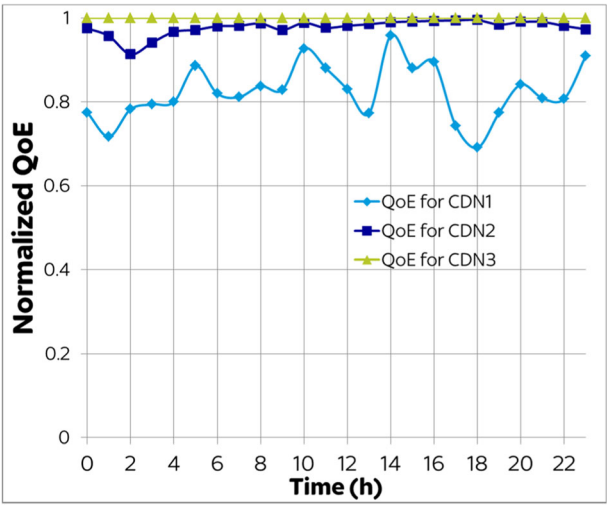


Fig. 7 Normalized QoE vs time

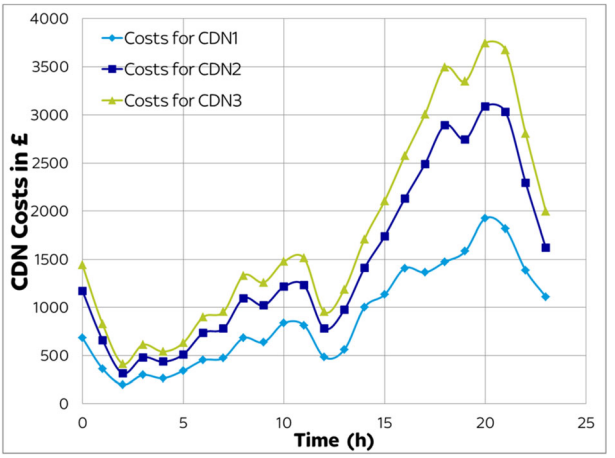


Fig. 8 CDN Costs vs Time

Expensive CDN3 have better scaling and edge node distribution capabilities and meets throughput request well, even on peak hours. The quality of edge node usage is noticeable with the ratio of Intermediate to Edge CDN node usage.

Employing more edge cache nodes during high demand times with increased proximity to users are the key action to establish a well-structured CDN. However, this increase the cost of deployment and the content provider should make a choice between high QoE and lower cost CDN. In this paper, we focus on how, when and how to switch between a multi-CDN system while keeping the costs lowest and QoE as high as possible (Figs. 8 and 9).

Obviously, CDN3's better scaling and edge node usage capability has a cost much higher than the other providers. Generally, providing best QoE for users seems to be the primary objective of an online video service [28]. However, on many cases, meeting the budgets is the actual priority for many operators.

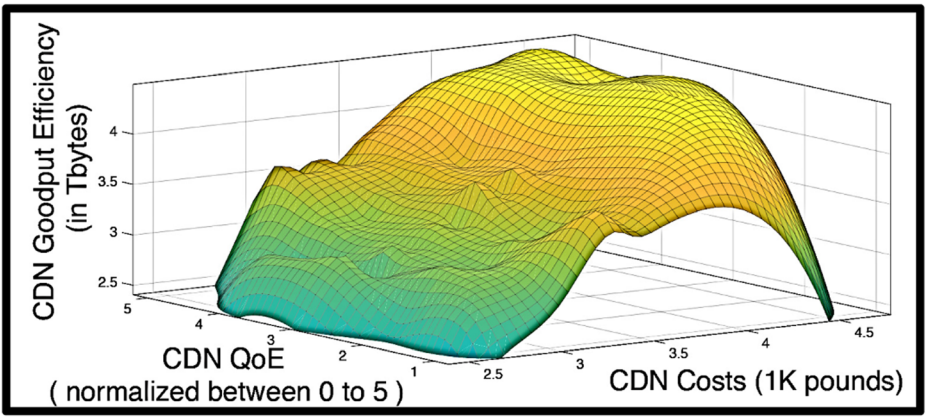


Fig. 9 Goodput/QoE/Costs Model for a CDN Deployment

8.5 QoE/Costs/Efficiency model for CDN deployments

This section presents how goodput, QoE and costs changes over time for different CDN deployments. As a deduction, a time invariant model is proposed as a base model for CDN deployment where the dimensions correspond to QoE, costs and goodput efficiency of a CDN. This model can provide a basis for any deployment that requires the calculation of budget vs throughput and user demand in case of an expected service quality.

The model that relates CDN constraints with Eq. 34 have the following coefficients: p00 = 1.15, p10 = 2.53, p01 = 0.67, p11 = 0.53, p20 = 1.32, p02 = 0.89, p21 = 1.42, p12 = 0.54, p30 = 0.43, p31 = 2.04, p32 = 0.19 for Goodput (a) and Costs (μ).

$$\mathrm{QoE}(\mathrm{Goodput}, \mathrm{Costs}) = \sum_{k=0,}^{3} \sum_{i=0}^{2} p_{ij}.a^k.\mu^i \tag{34}$$

$$\begin{aligned}\mathrm{QoE}(\mathrm{a}, \mu) = {} & p_{00} + p_{10}.a + p_{01}.\mu + p_{11}.a.\mu + p_{20}.a^2 + p_{02}.\mu^2 + p_{21}.a^2.c + p_{12}.a.\mu^2 \\ & + p_{30}.a^3 + p_{31}.a^3.\mu + p_{32}.a^3.\mu^2 \end{aligned} \tag{35}$$

Equations 33 and 34 will be a guide to any content provider or online video delivery platform, to estimate their costs and deployment of intermediate-edge VM node distribution strategy and overall user QoE. Obviously, it should be considered due to the developing technology [1], the costs for the throughput tend to fall where same (or more) amount of data can be streamed for a smaller budget when compared 2017 rates to 2016 rates [9, 15]. The estimations and simulations that are advertised in this work reflect 2017 cloud resource rates for the main CDN suppliers [16, 24, 27].

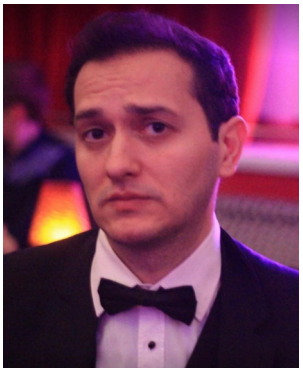
9 Conclusions

In this work, a generalized mathematical model to calculate the success rate of different tiers of online video delivery system is presented. The success of online streaming relies on maximizing customer satisfaction, minimizing online deployment costs while using the distributed nodes efficiently that results on maximizing goodput and optimizing latency. In section 6, a model for switching between multi-CDN has been proposed where QoE, CDN costs and usage of intermediate and edge nodes are adjusted. An algorithm to indicate the correct moment to switch between CDNs is also presented which will enhance QoE and budget relationship of an online video delivery platform.

References

1. Ahmad A et al (2016) QoE-centric service delivery: A collaborative approach among OTTs and ISPs", Computer Networks Elsevier

2. Aioffi WM et al (2004) Dynamic content placement for mobile content distribution networks. p20, "Web Content Caching and Distribution: 9th International Workshop, WCW
3. Akamai Inc (2018) Why akamai. https://www.akamai.com/us/en/why-akamai. Last accessed on 9th June 2018
4. Amazon Web Services Inc (2018) Amazon EC2. https://aws.amazon.com/ec2/. Last accessed on 9 Jun 2018
5. Amazon Web Services Inc (2018) Amazon cloud front CDN. https://aws.amazon.com/cloudfront. Last accessed on 9th June 2018
6. Amazon Web Services Inc (2018) Amazon web services tiered pricing. http://docs.aws.amazon.com/AmazonDevPay/latest/DevPayDeveloperGuide/TieredPricing.html, Last accessed on 9th June 2018
7. Broadcasters' Audience Research Board (2018) Online Tv Viewing. http://www.barb.co.uk/trendspotting/analysis/online-tv-viewing/. Last accessed on 9th Jun 2018
8. Chen Z (2017) Probabilistic caching in wireless D2D networks: cache hit optimal versus throughput optimal. IEE Communications Letters 21(3):584–587
9. Chu Y et al (2014) Software-Defined QoE Measurement Architecture. APNOMS
10. Cunqian Y et al (2016) Parallel virtual machine migration in WDM optical data center networks", Optical Switching and Networking
11. D'Amico V (2016) An SDN/NFV Telco Operator Platform for Video Broadcasting. IEEE Communications Magazine
12. Fan Q et al (2017) Video delivery networks: Challenges, solutions and future directions. Computers & Electrical Engineering
13. Faraci G et al (2017) A simulative model of a 5G Telco Operator Network. The 12th International Conference on Future Networks and Communications
14. Frias Z et al (2017) 5G networks: Will technology and policy collide? Telecommunications Policy
15. Gilani M et al (2017) Mobility management in IEEE 802.11 WLAN using SDN/NFV technologies. EURASIP Journal on Wireless Communications and Networking
16. Google Inc (2018) Google CDN pricing. https://cloud.google.com/cdn/pricing. Last accessed on 9th June 2018
17. Google Inc (2018) Google cloud CDN. https://cloud.google.com/cdn/. Last accessed on 9th June 2018
18. ITU-T (2016) P.1203.3, Parametric bitstream-based quality assessment of progressive download and adaptive audiovisual streaming services over reliable transport –Quality integration module
19. Jones W et al (2016) Planning For 5G: A Problem Structuring Approach for Survival in the Telecoms Industry", Wiley Systems Engineering Journal
20. Kleinrock L (1993) On the Modelling and Analysis of Computer Networks. Proc IEEE 81(8):1179–1191
21. Knoll T et al (2016) QoE Evaluation and Enforcement Framework for Internet Services. In: Itu Study Period 2013–2016, Chemnitz University of Technology, Germany
22. Koumaras H et al (2016) Enabling Agile Video Transcoding over SDN/NFV-enabled Networks", International Conference on Telecommunications and Multimedia
23. Mestric R (2018) Stretching the CDN. Whitepaper. https://insight.nokia.com/stretching-cdn
24. Neves P et al (2016) Future mode of operations for 5G – The SELFNET approach enabled by SDN/NFV", Computer Standards & Interfaces
25. Roeder TMK et al (2016) Simulation and optimization of content delivery networks considering user profiles and preferences of internet service providers. Winter Simulation Conference
26. Ruiz M et al (2016) Big Data-backed video distribution in the telecom cloud. Computer Communications
27. Sjöberg D (2008) Content Delivery Networks: Ensuring quality of experience in streaming media applications", TeliaSonera International Carrier, CDN white paper, p 7
28. Stallings W (2016) Foundations of modern networking: SDN, NFV, QoE, IoT, and Cloud. Section 11.2, Indianapolis, US
29. Taleb T et al (2016) "Anything as a Service" for 5G Mobile Systems. IEEE Network
30. Yousaf F et al (2017) Network Slicing with Flexible Mobility and QoS/QoE Support for 5G Networks. IEEE International Conference on Communications Workshops



Utku Bulkan is a PhD Researcher at London South Bank University and Senior Software Engineer at SKY UK with passion for working on mathematical analysis of high performance, scalable and low-latency systems for broadcasting and video delivery platforms. He received the Physics Degree from Technical University of Istanbul and he has attended Computational Science & Engineering and Satellite Communications & Remote Sensing M.Sc. programs at ITU. His research interests include: parallel programming, digital image processing, pattern recognition, machine learning, data fitting, applied numerical and computational and interpolation methods, integral calculations, matrix operations, numerical differential equation solutions.

Tasos Dagiuklas is a leading researcher and expert in the fields of Internet and multimedia technologies for smart cities, ambient assisted living, healthcare, and smart agriculture. He is the leader of the newly established SuITE (Smart Internet Technologies) research group at LSBU, where he also acts as the Head of Division in Computer Science. Tasos received his Engineering Degree from the University of Patras, Greece, in 1989. He completed an MSc at the University of Manchester in 1991 and a PhD at the University of Essex-UK in 1995, all in Electrical Engineering. He has been a principal investigator, coinvestigator, project and technical manager, coordinator and focal person of more than 20 internationally research and development and capacity training projects. With total funding of approximately £5 m from different international organisations. His research interests include open-based networking (software-defined network, network function virtualisation), FTV, 3DV, media optimisation across heterogeneous networks, quality of experience, virtual reality/augmented teality technologies and Cloud infrastructures and services. He has published more than 150 papers in these fields. His research has received more than 1150 citations by researchers (Google Scholar). He has served as Vice-Chair for the Institute of Electrical and Electronics Engineers (IEEE) Multimedia Communications Technical Committee (MMTC) Quality of Experience Working Group, and as Key Member for IEEE MMTC MSIG and 3DRPC WGs.

Muddesar Iqbal is Senior Lecturer in Mobile Computing in the Division of Computer Science and Informatics, School of Engineering. He is an established researcher and expert in the fields of: mobile cloud computing and open-based networking for applications in disaster management and healthcare; community networks; and smart cities. Dr. Iqbal won an EPSRC Doctoral Training Award in 2007 and completed his PhD from Kingston University in 2010 with a dissertation titled "Design, development, and implementation of a high-performance wireless mesh network for application in emergency and disaster recovery". His research interests include 5G networking technologies, multimedia cloud computing, mobile edge computing, fog computing, Internet of Things, software-defined networking, network function virtualisation, quality of experience, and cloud infrastructures and services. He has contributed to 35 research publications (including 24 journals, 11 conference proceedings, and 2 book chapters).