Streaming of Scalable Multimedia over Content Delivery Cloud

Xiaoming Bao, Rongshan Yu,
Institute for Infocomm Research, A*STAR, Singapore
Email: {baoxm, ryu}@i2r.a-star.edu.sg

***Abstract*—Content Delivery Cloud (CDC) extends Content Delivery Network (CDN) to provide elastic, scalable and low cost services to the customers. For multimedia streaming over CDC, caching the media content onto the edge server from storage cloud is commonly used to minimize the latency of content delivery. It is very important for CDN to balance between the resources being used (storage space, bandwidth, etc) and the performance achieved. Commercial CDNs (such as Akamai, Limelight, Amazon CloudFront) have their proprietary caching algorithms to deal with this issue. In this paper, we propose a method to further improve the efficiency of the caching system for scalable multimedia contents. Specifically, we notice that a scalable multimedia content can be flexibly truncated to lower bit rates on-the-fly based on the available network bandwidth between the edge server to the end users. Therefore, it may not be necessary to cache such a content at its highest quality/rate. Based on this observation, we show that edge server can decide an optimized truncation ratio for the cached scalable multimedia contents to balance between the quality of the media and the resource usage. The proposed optimized truncation algorithm is analyzed and its efficacy in improving the efficiency of the caching system is justified with simulation result.**

***Index Terms*—Content delivery network, cloud storage, multimedia streaming, caching algorithm.**

I. Introduction

Content Delivery Network (CDN) played an essential role in assisting content providers to deliver multimedia contents to end users efficiently. The objective and internal mechanism of CDN are well explored by [1], [2]. Recently, the cloud computing technology [3] has been widely used in CDN to provide elastic, scalable and low cost services to the customers, which extends the traditional CDN model to so called Content Delivery Cloud (CDC) [4].

Typically, a CDN service cloud includes a number of basic components such as storage, parallel computing engine, controller and user front (Edge/Proxy server). The internal of the cloud architecture is transparent to the users. The globally deployed edge servers respond to the nearby user requests for the contents whose physical locations are totally unaware to the users. The controllers are responsible for retrieving the content from the respective storage cloud and transferring it to the edge server with guaranteed Quality-of-Service (QoS).

For cloud-based multimedia computing, including multimedia streaming, QoS requirements in terms of bandwidth, delay and jitter are key factors in designing a multimedia system [5]. Numeric efforts have been done through high level architectural design [5], [6] or Distributed File System (DFS) specific file perfecting [7] to satisfy the QoS requirements for media content delivery service. Practically, caching the media content from storage cloud onto the edge server is a common technique being widely used in CDC to effectively reduce the latency of media content delivery.

In this paper, we propose an optimized caching algorithm for streaming scalable encoded multimedia such as MPEG-4 SLS [8] over CDC. The scalable multimedia is encoded in a way such that a low bit rate frame data can be generated by truncating the data from a higher bit rate frame [9]. Therefore, one media content can be encoded into a single source with highest bit rate, while in streaming application the multimedia file can be further truncated on the fly to lower bit rate before sent to end users. The actual bit rate is thus determined by a truncation ratio λ depending on the available network bandwidth from the edge server to the users. We take advantage of this scalable feature to further improve the efficiency of existing caching algorithms in CDC, where the quality of the scalable multimedia file being cached in the edge server is determined based on the network conditions from the edge server to the users. In the proposed algorithm, both the utility function of scalable media as a function of bit-rate and the cost of transmitting and storing the media files from storage cloud onto the edge server are considred to achieve the best balance between the service quality and its associated cost. In this paper, the optimal truncation ratio for the scalable media is formulated as stochastic optimization problem and a solution to this problem is given followed by numerical analysis to the solution. The efficiency of the proposed algorithm is justified with simulation result.

The rest of this paper is organized as follows. The formulation of the problem of optimal initial truncation ratio determination is given in Section II and a solution based on stochastic optimization is given. The proposed solution are further analyzed in Section III where some interesting properties of the proposed solution are given. Numerical simulation results are given in Section IV. Finally, this paper is concluded in Section V.

II. Problem Formulation and Solution

We assume that all the scalable multimedia files are stored in storage cloud initially. Upon receiving request by the first end user, a multimedia file will be fetched from the storage cloud to the edge server, and streamed to the end user by the edge server. The multimedia file will be stored in the cache

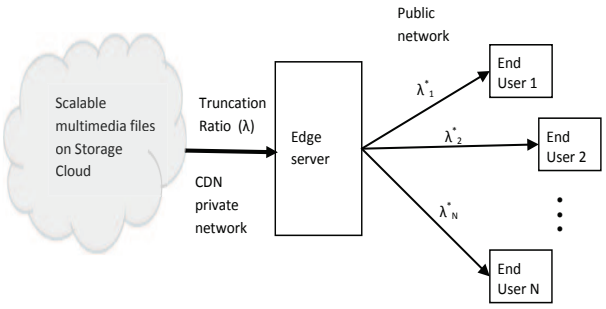

Fig. 1. Conceptual model.

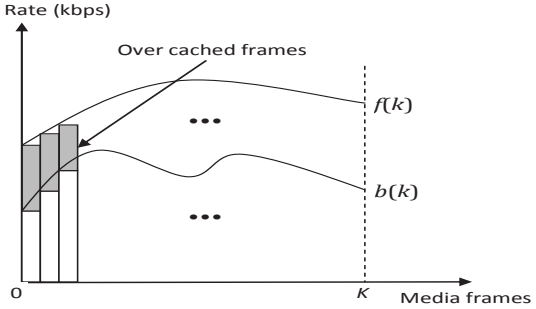

Fig. 2. Illustration of over caching.

system on the edge server, and will be used by the edge server to serve further requests from end users. An initial truncation ratio $0 < \lambda \leq 1$ will be determined for the scalable multimedia when it is fetched from the storage cloud. Supposing that there are N end users who request the same media source before it is deleted from the cache memory, the actual rates at which this content is sent to the end users will be dependent on the available streaming bandwidth, which is denoted by B_n, $n = 1, \ldots, N$. Typically, an adaptive streaming module [9] is implemented on the edge server, which will decide a truncation radio λ_n^*, $n = 1, \ldots, N$ for the scalable multimedia content for each user n so that the final streaming data rate doesn't exceed the available network capacity. Fig. 1 gives a conceptual model of the problem.

Clearly, if the edge server fetches highest quality media files from the storage cloud, and the bandwidths during actual streaming sessions are always lower than the data rate of the media files stored in the cache system, the scalable media files have to be further truncated before they are sent to the users. We can see that in such a case, the cloud resources such as storage space of cache system and/or the transmission bandwidth between storage cloud and the edge server are wasted. We call it over caching, which is illustrated in Fig. 2 where $f(k)$, $k \in K$ is the bit-rate of k-th frame of the media source and $b(k)$, $k \in K$ is the network bandwidth between the edge server and the end user at the time when k-th frame is being transmitted from the edge server to the client. Here K is the set of all frames from a multimedia file.

On the other hand, if the edge server is very conservative in terms of cloud resource usage, and stores only low quality media files; while the network bandwidths between edge server and end users are higher than the rates of the media file stored in the cache, it is not possible for the edge server to raise the media quality on the fly. In that case, the system performance is reduced in terms of the quality of the multimedia content being sent to the end user.

In this research we will study how the initial truncation ratio λ can be optimally determined based on the statistics of the network conditions between edge server and end users. We will see that the problem can be formulated as a stochastic optimization where the optimal initial truncation ratio λ can be selected to maximize the expected benefits considering both the utility of the multimedia files, and the cloud system costs.

A. Bandwidth Distribution

During a streaming session, the available network bandwidth between the edge server and an end user can be estimated in real-time using existing tools and algorithms [10], [11]. Unfortunately, it is in general not possible to predict all the available bandwidth for N end users at the time of determining the initial truncation ratio λ. Instead, in the proposed algorithm, the initial truncation ration is determined from the empirical probability distribution of available streaming bandwidth which can be established by the edge server from historical data during its operation. We assume such a distribution is given by a probability distribution function $f(B)$.

B. Utility Function of Multimedia Files

Considering an adaptive streaming scenario given in Fig. 1, the user can get gain by successfully receiving the multimedia file, and the gain is determined by the quality of the received multimedia file. Let F be the media frame rate for scalable multimedia file at its highest quality, and $\lambda^* F$ is the rate at which the edge server is sending to the end user. Thus, the gain function is a function of λ^* with the following limitations:

$$\lambda^* \leq \lambda, \tag{1}$$

and

$$\lambda^* F \leq B, \tag{2}$$

where B is the data rate available for streaming between the edge server and end user, and λ is the initial truncation ratio of the cached multimedia file.

Now we are ready to define the utility function for the optimization problem. Considering both the gain function and the cost such as the usage of the CDN's network bandwidth and the storage on the edge server, the utility of caching a scalable multimedia file is actually given by:

$$U = g(\lambda^*) - c(\lambda), \tag{3}$$

where $g(\lambda^*)$ and $c(\lambda)$ are, respectively, the gain function and the cost when the multimedia file is stored at the edge server using λ as the initial truncation ratio, and streaming to end

user when it is further truncated to λ^*. The utility function may be simplified as

$$U = g(\lambda^*) - \gamma\lambda, \tag{4}$$

if we further assume that the cost is a linear function of the data rate or the size of the cached multimedia file, which is determined by the initial truncation ratio λ. Here γ is a constant representing the unit cost with respect to the multimedia file.

Although there are different models to characterize the quality-rate relationship for different audio and video coding standards, a common characteristic is that the perceptual quality is a non-linear monotonically increasing and convex function of the rate. Typically, a logarithm function is a close match to such relationship [12]. Furthermore, since $0 < \lambda^* \leq 1$, in this paper we use a quadratic Macraurin expansion to approximate the gain function $g(\lambda^*)$ although it is straightforward to extend the result to other types of gain functions. By ignoring $O^3(\lambda^*)$ and above, the gain function is given by :

$$g(\lambda^*) = \alpha\lambda^*(2 - \lambda^*), \tag{5}$$

where α is a parameter representing the maximum utility of the multimedia file when λ^* = 1.

C. Optimized Initial Truncation Ratio

Now we turn our attention to the problem of determining the optimal initial truncation ratio λ of the scalable multimedia file cached in the edge server for utility maximization. since truncation ratio λ^* is determined by the edge server in real-time depending on the network conditions, which is in general unknown *a priori* when λ is determined, it is not possible to find a λ that maximizes the actual utility function U. Instead, we maximize its expected value with respect to the distribution of the bandwidth from the edge server to its clients. Therefore, the optimization problem is given as follows:

$$\begin{aligned} \max \quad & E\{U\}, \\ \text{s.t.} \quad & \lambda^* F \leq B, \\ & \lambda^* \leq \lambda, \end{aligned} \tag{6}$$

where $E\{\cdot\}$ is statistical expectation. Substituting (5) and (4) into the optimization problem (6), it can be shown that the optimal solution is given by the following equation:

$$(1 - \lambda)[1 - \Phi(\lambda)] = \frac{\beta}{2}, \tag{7}$$

where $\Phi(\lambda) = \int_0^{\lambda F} f(B)dB$ is the cumulative distribution function of the bandwidth B, and $\beta = \gamma/\alpha$ is a weighting coefficient that determines the trade-off between the gain to the end user in terms of the quality of the streaming service and the costs associated with the usage of the network and storage resources of the cache system. Numerical solution for this problem is possible once the probability distribution of B is identified.

III. Analysis of the Optimal Solution

In this section, we analyze the optimal solution of λ. For simplicity, we assume that the network bandwidth B is Gaussian with mean μ and standard deviation σ although it is straightforward to extend the result to other distributions. From (7) it can be seen that the optimal λ will depends on four parameters, namely, the mean (μ) and the standard deviation (σ) of the bandwidth distribution, the weighting coefficient (β) of the utility function, and the original (highest quality) rate (F) of the media source on the storage cloud. Fig. 3 - 6 illustrated their effects on the resulting optimized truncation ratio respectively.

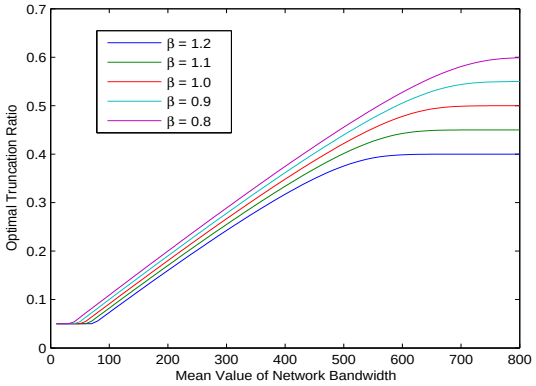

Fig. 3. $\lambda - \mu$ relationship

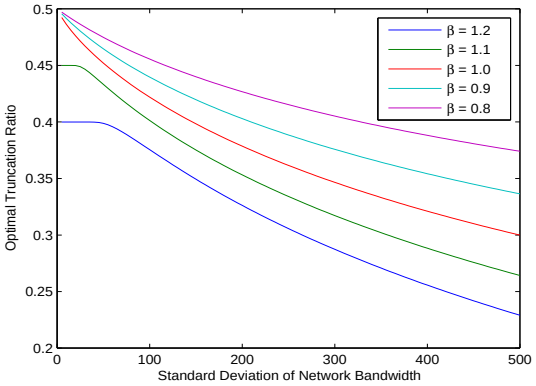

Fig. 4. $\lambda - \sigma$ relationship

A. Bandwidth Distribution Parameters μ and σ

In this analysis, we fix the original media rate F to 1000kbps and observe the effect of bandwidth variation onto the optimized solution. Fig. 3 shows that at low end of network bandwidth, the resulting optimized truncation ratio λ is almost linearly proportional to μ, while λ tends to a maximum value which is affected by the weighting coefficient β if μ keeps increasing. The linearly proportional relationship is highly anticipated because it shows that the optimized truncation ratio λ is, indeed, related to the average available streaming bandwidth in particular in a low bandwidth environment. It

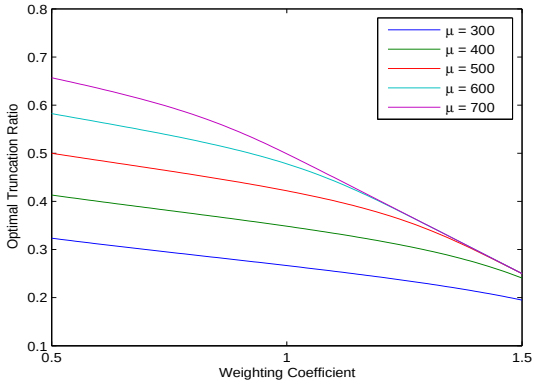

Fig. 5. $\lambda - \beta$ relationship

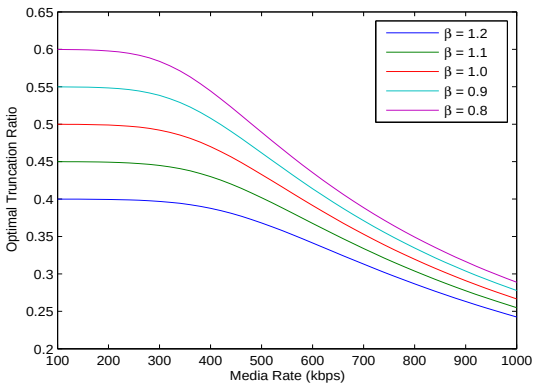

Fig. 6. $\lambda - F$ relationship

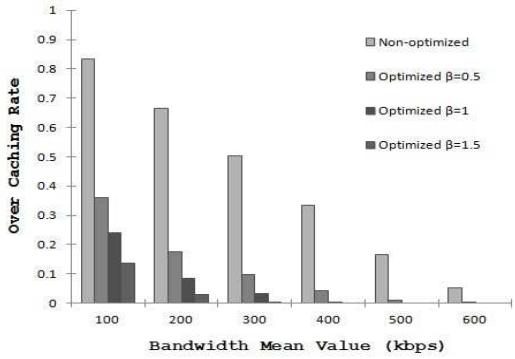

Fig. 7. Comparison of OCR at F=600kbps and σ=100

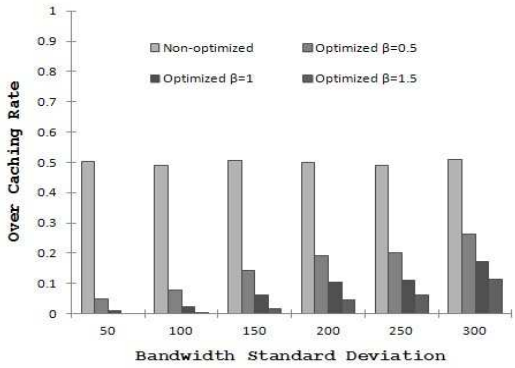

Fig. 8. Comparison of OCR at F=600kbps and μ=300kbps

also shows that in cases where network bandwidth is sufficient, the optimized truncation ratio λ will be only determined by the weighting coefficient β. For example, when β = 1, the optimized truncation ratio λ approaches 0.5. From Fig. 4 we can see that if the available streaming bandwidth has less variation (smaller σ), the resulting truncation ratio will be capped at a small value when β is greater than one to keep low cost. Usually the bandwidth of a 3G wireless network is very fluctuating with large σ, in that case we can see that λ is almost inverse proportional to σ, i.e., the more unpredictable of the network bandwidth, the system will become more conservative with smaller optimized truncation ratios.

B. Weighting Coefficient β

In this analysis, we also fix the original media rate F to 1000 kbps, Fig. 5 shows that if the CDN would like to increase end users gain by using smaller β, λ will be mainly determined by μ; while if the CDN would like to save more costs with larger β, λ tends to be not affected by μ and keeps at a reasonably small value.

C. Original Media Rate F

In this analysis, we fix μ to 500kbps. It can be seen from Fig. 6 that if μ is sufficiently large compared to the full media rate F, then λ will be capped at a value determined mainly by β; while for large F, λ will be almost inverse linearly proportional to F.

IV. Simulation Results

So far we have shown the solution to the optimized initial truncation rate and analyzed the effects of a number of parameters on the solution. In this section, we further evaluate how much improvement on the over caching being discussed in section II that the optimized method can achieve, compared with non-optimized full size caching method. First, we define the Over Caching Rate (OCR) as following:

$$OCR_n(k) = \begin{cases} [f(k) - b_n(k)]/f(k), & f(k) \geq b_n(k) \\ 0, & f(k) < b_n(k) \end{cases} \quad (8)$$

where $k \in K$ is the k-th of total K frames in the scalable multimedia file and $n \in N$ refers to n-th of total N end users. Next, we simulate 1000 end users requesting for a same media source (which is a 6 minute long music scalably encoded by MPEG-4 SLS) under randomly generated network bandwidth from a Gaussian distribution. The average media frame rate is around 600 kbps. We calculate the OCR frame by frame for all 1000 end users. Finally we sum up the OCR for all end users as following:

$$OCR_{total} = \frac{1}{N}\sum_{n=1}^{N}[\frac{1}{K}\sum_{k=1}^{K} OCR_n(k)], \quad (9)$$

The simulation results are shown by Fig. 7 and Fig. 8. In Fig. 7, we fixed the standard deviation of network bandwidth at $\sigma = 100$, it can be seen that the OCR has been significantly reduced especially at lower end of bandwidth mean value. E.g. it reduced OCR from 0.5028 to 0.0332 at μ = 300 kbps when β=1, more than 15 times improvement. In Fig. 8, we fixed the mean value of network bandwidth at 300 kbps and it shows that the standard deviation has much less effect on the OCR especially for the non-optimized case. It can be seen that with the increasing of bandwidth standard deviation, the improvement of the optimized method on the OCR becomes less significant, e.g., from more than 15 times improvement at standard deviation $\sigma = 100$ to only about 2.9 times at standard deviation $\sigma = 300$ when β=1.

V. Conclusions

We show in this paper that for scalable multimedia streaming over CDC, the scalability feature of scalable multimedia can be leveraged to perform a partial data retrieval from the storage cloud to achieve best trade-off between service quality and effective usage of network and storage resource of the edge server. The optimal truncation ratio can be formulated as a utility maximization problem considering both the gain function to the end users, and the cost associated with the cache system. The problem can be solved numerically by the edge server once the statistics of the network conditions between the edge server and end users are known *a priori*. The benefits of the proposed solution is justified using simulations assuming quadratic gain function and Gaussian distributed bandwidth available for streaming.

References

[1] A. J. Su, D. R. Choffines, A. Kuzmanovic, and F. E. Bustamante, "Drafting behind akamai: Inferring network conditions based on cdn redirections," *IEEE Trans. on Networking*, vol. 17, no. 6, pp. 1752 – 1764, Dec. 2009.

[2] E. Nygren, R. K. Sitaraman, and J. Sun, "The akamai network: A platform for high-performance internet applications," *ACM SIGOPS Operating Systems Review*, vol. 44, no. 3, July 2010.

[3] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. Katz, A. Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M. Zaharia, "A view of cloud computing," *Commun. ACM*, vol. 53, no. 4, pp. 50–58, Apr. 2010. [Online]. Available: http://doi.acm.org/10.1145/1721654.1721672

[4] M. Pathan, J. Broberg, and R. Buyya, "Maximizing utility for content delivery clouds," in *Procedings of the 10th International Conference on Web Information System Engineering (WISE '09)*, 2009, pp. 13–28.

[5] W. Zhu, C. Luo, J. Wang, and S. Li, "Multimedia cloud computing," *IEEE Signal Processing Magzine*, pp. 59 – 69, May 2011.

[6] L. Ramaswamy, L. Liu, and A. Iyengar, "Cach clouds: Cooperative caching of dynamic documents in edge networks," *Proc. 25th IEEE Int. Conf. Distributed Computing Systems (ICSCS)*, pp. 229 – 238, June 2005.

[7] B. Dong and et al, "Correlation based file prefetching approach for hadoop," in *Proc. 2nd IEEE Int. Conf. Cloud Computing Technology and Science (CloudCom)*, Dec. 2010, pp. 41 – 48.

[8] R. Yu, S. Rahardja, X. Lin, and C. C. Ko, "A fine granular scalable to lossless audio coder," *IEEE Trans. Audio Speech Language Processing*, vol. 14, no. 4, pp. 1352–1363, Jul. 2006.

[9] R. Yu, H. Shu, and S. Rahardja, "An adaptive streaming system for mpeg-4 scalable to lossless audio," in *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, Oct. 2011.

[10] R. S. Prasad, M. Murray, C. Dovrolis, and K. Claffy, "Bandwidth estimation: metrics, measurement techniques, and tools," *IEEE Network*, vol. 17, no. 6, pp. 27–35, Nov.-Dec. 2003.

[11] C. Casetti, M. Gerla, S. Mascolo, M. Y. Sanadidi, , and R. Wang, "Tcp westwood: Bandwidth estimation for enhanced transport over wireless links," in *Proceedings of ACM MOBICOM 01*, July 2001.

[12] Y. Chen, B. Wang, and K. Liu, "Multiuser rate allocation games for multimedia communications," *IEEE Trans. on Multimedia*, vol. 11, pp. 1170 – 1181, Oct. 2009.