A note on content caching and delivery based on AVC and SVC

Mei Takafuji[†1] Hidenori Nakazato[†2] Hiroshi Watanabe[†2]

Abstract: Content cache effectiveness in video distribution is mathematically analyzed. Targeted video distribution models are two types. One is simulcast, the other is scalable approach. Simple two layer model is used for the analysis. Caching effect is studied focusing on the amount of required transmission data based on some assumptions with a certain cache hit probability. The result shows that the scalable approach cannot improve the caching effect.

Keywords: cache, scalability, simulcast, hit probability, SVC, AVC

1. Introduction

Information Centric Network (ICN) has been paid attention for an effective content distribution. Especially, video content occupies a lot of network bandwidth. Video content caching at the edge server is regarded to be an effective solution to decrease data occupancy in the network. Many researches using MPEG AVC and SVC in the ICN environment have been reported [1]-[3]. However, most of SVC in these literatures is not spatial but SNR scalable system. On the other hand, many commercial video distribution applications use different image resolutions. Therefore, it may be important to know whether spatial SVC fit to video caching system compared to AVC. Using AVC for different image resolutions can be regarded as simulcast system. In this paper, content cache effectiveness in video distribution is mathematically analyzed. Caching effect is studied focusing on the amount of required transmission data based on some assumptions with a certain cache hit probability. Section 2 shows the target model [4], section 3 explains the way of analyzing caching effect for simulcast and scalable system and section 4 concludes the obtained results.

2. Content delivery system model

It is known that the large part of network bandwidth is occupied by video streaming applications. Most of video streaming application provides several qualities of video. In the commercial case such as “YouTube”, the number of different quality is more than 5, and all of them have different sizes. Users select one of them although default size is given by the system. This trend is now shifted to the dynamic selection in real-time like MPEG-DASH. Typical image resolutions [5][6] are summarized in the Table 1. From this table, we can convert them to 2:1 spatial scalable system shown in Table 2.

The root server in the Fig.1 has low and high quality data for simulcast approach. For the scalable case, low layer data and additional high layer data are distributed [4]. The edge server may store high quality data for simulcast case and low layer data for scalable case in a cache. Here, we assume that all content of high quality data should be delivered to a client.

3. Caching for simulcast and scalable system

3.1 Traditional Case

In this section, we compare the caching effect in AVC and SVC environment based on the model proposed by the paper [4]. First, we consider the two layer model based on simulcast and scalable approach. In the case of simulcast, data amount for low and high quality layer are denoted as D_{GL} and D_{GH}. In the case of scalable distribution, such as spatial or SNR scalability,

[†1] Department of Computer Science and Engineering, Waseda University

[†2] Graduate School of Global Information and Telecommunication Studies, Waseda University

Table 1 Image types and resolutions [5].

Type	Resolution	Video bitrate
1080p	1920x1080	8Mbps
720p	1280x720	5Mbps
480p	854x480	2.5Mbps
360p	640x360	1Mbps
240p	426x240	Not specified

Table 2 Possible scalable system

Type	Resolution	Video bitrate
1080p (simulcast)	1920x1080	8Mbps
720p - 360p	1280x720 - 640x360	5Mbps, 1Mbps
480p - 240p	854x480 - 426x240	2.5Mbps, 0.5Mbps

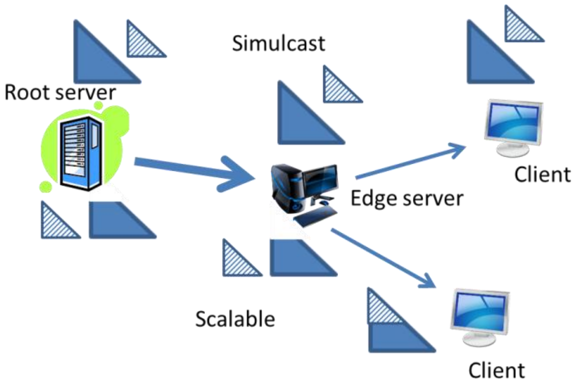

Figure 1 Content cache and delivery system

data amount of low and high quality layer are denoted as D_L and D_H. The total data amount is denoted as D_{SS}= D_L + D_H. For simplicity, we start from D_{SS}= D_{GH}.

Now, we consider that some parts of content locate in the cache buffer. Thus, the data, transmitted to a client, can be decreased depending on the cache hit probability. Data amount that should be transmitted in a single and scalable layer approach can be denoted as g(p) and h(p), with the hit probability p. Here, transmission data amount for the number of N content is defined as $D_{Tra}(z,N)$, (z=0 for low quality layer, z=1 for high quality layer). The summation of transmission data for all content with the hit probability of cached content is shown in Figure 2. In this figure, p=0 indicates that there is no cache at all. To transmit all high quality data, A [bit] (A= $D_{Tra}(1,N)$) is required. If p=1, no data should be transmitted and all data exists in a cache. This behavior is drawn by the line g(p). For the scalable approach, we assume that only the low quality layer exists in a cache and the transition is shown by the line h(p). Even if all data B [bit] (B= $D_{Tra}(0,N)$) exists in a cache, still A-B [bit] should be transmitted to a client.

From the figure, g(p) and h(p) are defined as

$g(p) = -Ap+A$, A= $D_{Tra}(1,N)$=N D_{GH}

$h(p) = -BP+A, \qquad B= D_{Tra}(0,N)=N\ D_L$ (1).

Thus, we have

$A = (D_{GH}/ D_L)B$ (2).

Now, we define the ratio of cache hit data for low and high quality layer be q and (1-q). For a cache hit probability p, the transmission data amount at the scalable approach can be given by f(p,q) [bit].

$f(p,q) = (1-q)g(p) + qh(p)$ (3).

The difference function r(p) between h(p) and g(p) is given by

$r(p) = h(p) – g(p) = -Bp+A - (-Ap+A) = (A-B)p$ (4)

Let us assume that the cache hit probability P_z be the one for high quality layer only. The transmission data for P_z is $f(P_z, 0)=q(P_z)$ indicated by the point “a” in the figure. Based on the point “a”, let us consider the case when the cache hit probability is increased to be P_x and at the same time the newly hit part is all for the high quality layer. The transmission data for P_x is $f(P_x, 0) = g(P_x)$ indicated by the point “b”. On the other hand, if the newly hit part is all for low quality layer, the transmission data P_x is given by $f(P_x, Q_x)$ indicated by the point “c”. Here, the new probability P_x is given by $P_x = P_z + P_xQ_x$ since all new probability is devoted to the lower layer having the ratio Q_x. Now, we wish to compare the cache effect when the increased cache from “a” to “d”, which is all for the high quality layer content, is replaced by the one of the low quality layer content. We assume that this point “d” has the probability P_y, which is extrapolated from P_x with the same tilt at the ratio of D_H/D_L. Here, we have

$P_y = P_x + (D_H/D_L)P_xQ_x$

$\quad = P_z + P_xQ_x + (D_H/D_L)P_xQ_x$

$\quad = [\ (D_H+D_L)P_xQ_x + D_LP_z\ ] / D_L$

$\quad = [\ (D_H+D_L)P_xQ_x - D_HP_z\ ] / D_L \qquad (\text{ since } P_z=P_x(1-Q_x)\ )$ (5).

Cache hit probabilities P_x, P_y and ratios of low quality layer Q_x, Q_y have the following relation,

$r(P_x)(1-Q_x) = r(P_y)(1-Q_y)$

$P_x(1-Q_x) = P_y(1-Q_y)$ (6).

From these equations, we can get

$P_x = [\ (1-Q_y)/(1-Q_x)\ ]\ P_y \qquad P_y = [\ (1-Q_x)/(1-Q_y)\ ]\ P_x$

$Q_x = [\ 1 - (1-Q_y)/P_x\ ]\ P_y \qquad Q_y = [\ 1 - (1-Q_x)/P_y]\ P_x$ (7).

Further, the cache data ratios and cache hit probabilities at the probability P_x and P_y are proportional to the data amount of two scalable layers.

$P_yQ_y : P_xQ_x = (D_L+D_H) : D_L$ (8),

thus we have

$P_y = P_xQ_x(D_L+D_H) / Q_yD_L$ (9).

From (7) and (9), we can get

$Q_y = Q_x(D_L+D_H) / (D_L+ Q_xD_H)$ (10).

Now, let us compare the required transmission data at the point “d” compared with the point “b”.

$g(P_x) – f(P_y, Q_y) = -AP_x + A - [\ (1-Q_y)\ g(P_y) + Q_yh(P_y)\ ]$

$\quad = \quad -AP_x + A - [\ (1-Q_y)(-AP_y+A) + Q_y(-BP_y+A)\ ]$

$\quad = A(P_y-P_x) + (B-A)P_yQ_y$

$\quad = A\ [\ P_xQ_x(D_L+D_H) / Q_yD_L - P_x\ ] + (B-A)\ [\ P_xQ_x(D_L+D_H) / Q_yD_L\ ]\ Q_y$

$\quad = (\ P_xQ_x/D_L\ )\ [\ B(D_L+D_H) - A\ D_L\ ]$ (11)

If, we have

$$B(D_L+D_H) - A\,D_L >= 0 \qquad (12),$$

then scalable cache is regarded to be effective when

$$D_{SS} = D_L+D_H >= D_{GH} \qquad (13).$$

This result amazingly contradicts to the result of the paper [4]. Since we have started from the assumption $D_{SS}= D_{GH}$, only the equivalent case can be true. This means that the scalable approach cannot improve the caching effect at all. The result can also be confirmed that the required transmission data at the point “b” and “d” in the Fig.2 is the same amount.

3.2 Practical Case

In most cases, spatial scalability is limited to two layers case. The size of enlarged image is limited up to 2.0 times in horizontal and vertical direction. This is also noted in the “Profile” of scalable video coding standard. It is known that the data amount of two-layer spatial scalability needs 10-20% more data than the simulcast approach [7][8]. Therefore, in this section, we modify the model shown in 3.1 to the Figure 3, which shows that $D_{SS}= \alpha D_{GH}$ (α>=1.0, ex. 1.1-1.2), and consider the effect of caching while other conditions are kept by the same token.

From the figure, g(p) and h(p) are defined as

$$g(p) = -Ap+A, \qquad A= D_{Tra}(1,N)=N\,D_{GH}$$

$$h(p) = -Bp+\alpha A, \qquad B= D_{Tra}(0,N)=N\,D_L \qquad (14).$$

The difference function r(p) between h(p) and g(p) is given by

$$r(p) = h(p) – g(p) = -Bp+\alpha A - (-Ap+A) = (A-B)p + (\alpha -1)A \qquad (15).$$

The same discussion between Eq.(5)-(10) holds. As for Eq.(11), the parameter α shows up.

$$\begin{aligned} g(P_x) – f(P_y, Q_y) &= -AP_x + A - [\, (1-Q_y)\, g(P_y) + Q_y h(P_y)\,] \\ &= -AP_x + A - [\, (1-Q_y)(-AP_y+A) + Q_y(-BP_y+\alpha A)\,] \\ &= A(P_y-P_x) + (B-A)P_yQ_y + (1-\alpha)A\,Q_y \\ &= (\, P_xQ_x/D_L\,)\, [\, B(D_L+D_H) - A\,D_L\,] - (\alpha -1)A\,Q_y \\ &= (\, P_xQ_x/D_L\,)\, [\, B(D_L+D_H) - A\,D_L – (\alpha -1)AQ_yD_L/(P_xQ_x)\,] \\ &= (\, P_xQ_x/D_L\,)\, [\, B(D_L+D_H) - A\,D_L\, (1+(\alpha -1)Q_y/(P_xQ_x))\,] \qquad (16) \end{aligned}$$

Here, we insert Eq. (1) and use the relation $D_{SS}= \alpha D_{GH}$.

$$\begin{aligned} g(P_x) – f(P_y, Q_y) &= (\, P_xQ_x/D_L\,)\, [\, ND_LD_{SS} - ND_{GH}D_L\, (1+(\alpha -1)Q_y/(P_xQ_x))\,] \\ &= P_xQ_xND_{GH}\, [\, \alpha - (1+(\alpha -1)Q_y/(P_xQ_x))\,] \\ &= P_xQ_xND_{GH}\, (\alpha -1)\, [1-Q_y/(P_xQ_x)\,] <=0 \qquad (17) \end{aligned}$$

Therefore, the right term is always less than or equal to zero. As a result, it turns out that the scalable approach does not provide effective cache utilization than simulcast approach when only lower layer data are stored in cache memory at edge server or a client.

4. Conclusion

In this paper, we first have made a correction to the analysis in [4], and confirmed that the scalable approach does not provide effective cache utilization compared with the simulcast approach when only lower layer data are stored in cache memory at edge server.

5. Acknowledgement

The research leading to these results has received funding from the EU-JAPAN initiative by the EC Seventh Frame-work Programme (FP7/2007-2013) Grant Agreement No.608518 (GreenICN) and NICT under Contract No. 167.

6. References

[1] Christpher Müller, Daniele Renzi, Stefan Lederer, Stafano Battista, and Christian Timmerer: “Using Video Coding for Dynamic Adaptive Streaming over HTTP in Mobile Environments,” 20th EUSIPCO 2012, pp. 2208-2212, Aug. 2012

[2] Hari Kalva, Velibor Adzic, and Borko Furht: “Comparing MPEG AVC and SVC for Adaptive HTTP Streaming,” IEEE International Conference on Consumer Electronics, pp.158-159, Jan. 2012.

[3] Christian Sieber, Tobias Hossheld, Thomas Zinner, Phuoc Tran-Gia, and Christian Timmerer: “Implementation and User-centric Comparison of a Novel Adaptation Logic for DASH with SVC,” IFIP/IEEE IM2013 Workshop, 1st International Workshop on Quality of Experience Centric Management (QCMan), pp.1318-1323, May 2013

[4] Mei Kodama: “A consideration on Video Content and Quality Management Methods Using Hierarchical Data in Content Caching and Delivery,” The Journal of Institute Electronics Engineers of Japan, Vol.42, No.1, pp.5-14, Jan. 2013

[5] Google You Tube help page: “Advanced encoding settings,” https://support.google.com/youtube/answer/1722171?hl=en#1722171

[6] Google You Tube help page: “Live encoder settings, bitrates and resolutions,” https://support.google.com/youtube/answer/2853702?hl=en#2853702

[7] Heiko Schwarz, Detlev Marpe and Thomas Wiegand: “Overview of the Scalable Video Coding Extension of the H.264/AVC Standard,” IEEE Transaction on Circuits and Systems for Video Technology, Vol. 17, No. 9, pp.1103-1120 Sep. 2007

[8] C. Andrew Segall and Gary J. Sullivan: “Spatial Scalability within the H.264/AVC Scalable Video Coding Extension,” IEEE Transaction on Circuits and Systems for Video Technology, Vol. 17, No. 9, pp.1121-1135 Sep. 2007

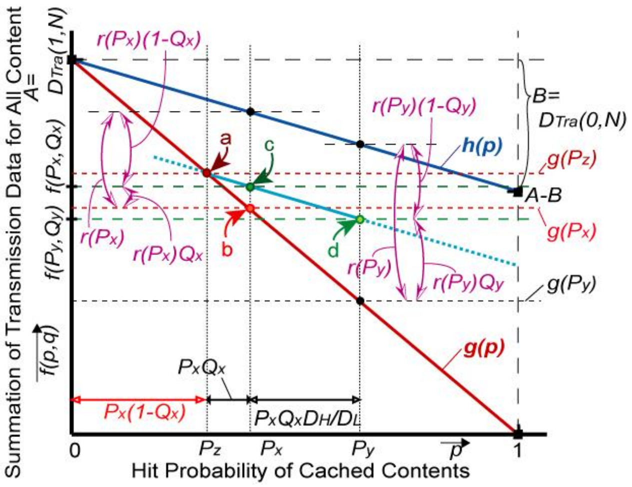

Figure 2 Relation of cache hit and transmission data at the conventional model

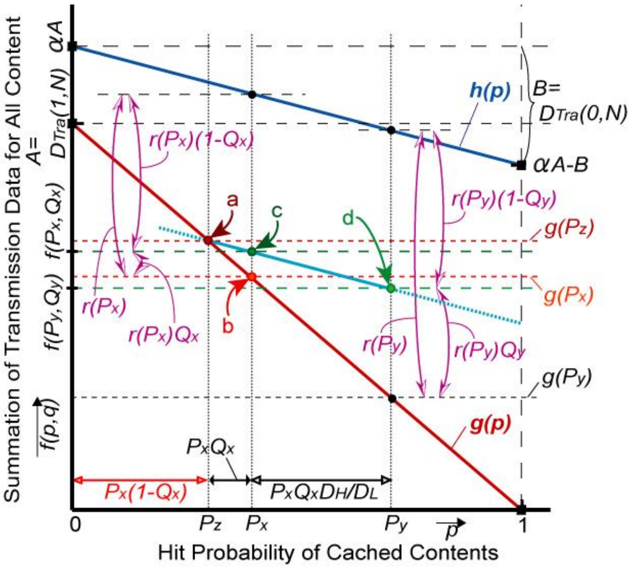

Figure 3 Relation of cache hit and transmission data for at practical model