SVC-based Scheduling Algorithm for Video in the Cloud

Song Xiao, Junmei Bai, Jianlong Zhang and Jianchao Du
ISN National Key Lab., Xidian University, Xi'an
E-mail: xiaosong@mail.xidian.edu.cn Tel: +86-29-88201956

***Abstract*—In order to meet the huge bandwidth requirement of cloud video, a practical scalable video coding (SVC)-based scheduling algorithm is proposed in this paper. The strategy fully considers the characteristics of scalable video bit stream. It accesses the users with basic layer bandwidth at first by the control criteria of maximization the number of service users according to the waiting time of the users, and then assigns enhancement layer bandwidths to the accessed users when the bandwidth remaining is not sufficient to access new users. The simulation results show that the proposed strategy could greatly shorten the average waiting time of users, increase the fairness index of bandwidth allocation and improve the average customer satisfaction efficiently compared with the First Come First Serve (FCFS) algorithm.**

I. INTRODUCTION

In recent years, the development and evolvement of cloud computing technology [1] places higher requirement for the efficient utilization of the network resources. In cloud computing, applications are provided and managed by the cloud server and data is also stored remotely in the cloud configuration. Cloud users do not need to download and install applications on their own device or computer, all processing and storage is maintained by the cloud server. However, how to guarantee the users enjoy the services timely according to their needs, and is there sufficient bandwidth to satisfy all demand of users in the cloud? These questions must be considered by the provider of cloud computing.

Cloud video [2] is one of the specific applications in cloud computing, i.e. the resources that cloud server provide are huge video data. Cloud video is totally transparent to users. It could automatically adapt to different kinds of terminals when they visit and access the video content in the cloud. The terminals, no matter smart phone, pad computer or other devices, access the same video content. In other word, cloud video technology could make different users enjoy the video experiences as easy as turning on the TV at home. Compared with other services, cloud video needs larger amount of data storage and transmission, which raises higher requirement on the network bandwidth. Therefore, how to improve the efficiency of bandwidth utilization and allocation is the main challenge that cloud provider needs to face.

In order to solve above problem, researchers have done a lot of work from different aspects. Ding [3] proposed a Quality-of-Services (QoS)-aware bandwidth allocation scheme for on-demand streams over broadband wireless networks. The scheme made use of fine-granular-scalability (FGS) encoded videos and dynamically adjusted the bit rate allocated to different streams to maximize the overall perceptual quality when available network bandwidth varied with time. M.S.Talebi [4] studied the problem of rate allocation for SVC-encoded multimedia applications. They used the staircase utility function to analytically model the SVC-encoded multimedia applications and proposed a convex optimization formulation for bandwidth sharing and rate control, which make the rate and the quality of video adapt to the network variation when different users share the bandwidth. Luo [5] proposed an adaptive framework for video streaming over the Internet, which jointly considered the design of packet scheduling and rate control with optimal bandwidth resource allocation. H.Mansour *et al* [6] considered medium grain scalable (MGS) extension of H.264/AVC video and developed rate-distortion models that characterize the coded bitstream. They also presented a resource allocation framework that jointly optimized the operation of the link adaptation scheme in the physical layer, and that of a traffic control module in the network or medium access control (MAC) layer.

In above researches, [5] didn't consider the characteristics of the bit stream of scalable video coding. Although it is taken into account in [3, 4, 6, 7], no concrete and practical strategy for the customers accessing was presented. Furthermore, although above researches could improve the bandwidth utilization and the video quality, the complexity of the algorithm is relatively high for the practical implementation. Therefore, in this paper, a practical scalable video coding (SVC)-based scheduling algorithm with low complexity is proposed to meet the huge bandwidth requirement of cloud video. The strategy fully considers the characteristics of scalable video bit stream. It classifies the bandwidth that users

This work is supported by NSFC (No.60702058, 60832001), CAST innovation fund, SRF for ROCS, SEM (JY0600090102), "111" project (No.B08038) and the Fundamental Research Funds for the Central Universities.

request into different layers, and adjust the allocation according to the network resources. The simulation results show the efficiency and the practicality of the strategy.

II. THE PROPOSED SVC-BASED SCHEDULING ALGORITHM FOR VIDEO IN THE CLOUD

A. *Framework of cloud video*

Figure 1 shows the framework of cloud video, which mainly includes the end users, application layer, service layer, resource scheduling system and infrastructure resources. In the framework, various end users send service requests to the cloud through the network. Then the resource scheduling system finds the suitable resource according to the request of users, and finishes the adaptation through the service and application layer, finally provides totally transparent video service to the end users. The resource scheduling system is the key module in the cloud video system, since it is in charge of the storage and the scheduling of whole resources. It is necessary to design an efficient scheduling strategy to guarantee the proper operation of the whole cloud video system. Therefore, this paper carefully studied the bandwidth allocation of the resource scheduling system and proposed a new SVC-based scheduling algorithm for cloud video.

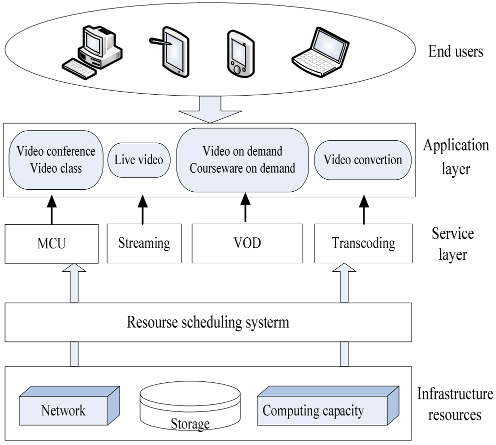


Fig.1 Framework of cloud video

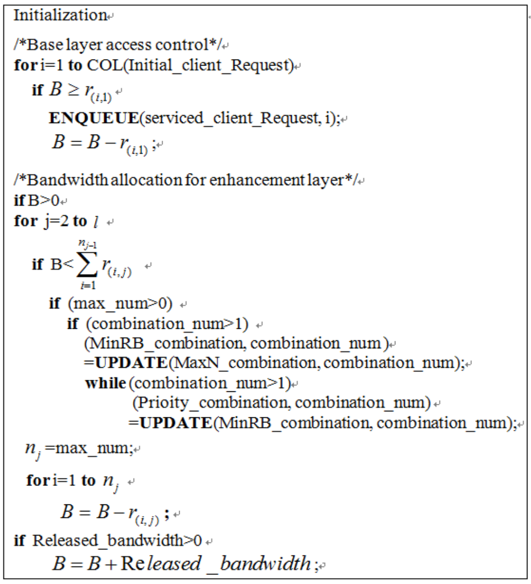
B. *Main idea of the algorithm*

The proposed SVC-based scheduling algorithm for cloud video takes fully account of the characteristics of scalable coding. It partitions the requested bandwidth of users into one base layer bandwidth and more enhanced layer bandwidths. Under the certain constraint of server bandwidth resources, the algorithm assigns base layer bandwidth to all requested users at first, i.e. to guarantee the basic QoS of all accessed users. When the number of accessed users could not increase but there is still bandwidth remaining, the algorithm assigns enhancement layer bandwidths to the accessed users. In conclusion, the main idea of the proposed algorithm is to access the users with basic layer bandwidth by the control criterion of maximization the number of service users according to the waiting time of the users. When the bandwidth remaining is not sufficient to access new users, the algorithm assigns enhancement layer bandwidths to the accessed users. The detailed scheduling strategy for enhancement layer bandwidth assignment is as follows,

1. The number of users that could be served are maximized at first;
2. If there exist various assignment methods to maximize the number of served users, the case of minimizing the bandwidth remaining is adopted, i.e. to assign the next enhancement layer bandwidth under the condition that the user's request of previous enhancement layer bandwidth has been satisfied.
3. It there still exist various assignment methods under the condition of the same bandwidth remaining, the algorithm assigns the bandwidth to the users with higher priority.

C. *The description of the algorithm*

Figure 2 and 3 show the pseudo code and the flow chart of the proposed algorithm respectively.



```
Initialization

/*Base layer access control*/
for i=1 to COL(Initial_client_Request)
  if $B \geq r_{(i,1)}$
    ENQUEUE(serviced_client_Request, i);
    $B = B - r_{(i,1)}$;

/*Bandwidth allocation for enhancement layer*/
if B>0
for j=2 to $l$
  if B<$\sum_{i=1}^{n_{j-1}} r_{(i,j)}$
    if (max_num>0)
      if (combination_num>1)
        (MinRB_combination, combination_num)
        =UPDATE(MaxN_combination, combination_num);
        while (combination_num>1)
              (Prioity_combination, combination_num)
             =UPDATE(MinRB_combination, combination_num);
 $n_j$ =max_num;
 for i=1 to $n_j$
     $B = B - r_{(i,j)}$;
if Released_bandwidth>0
    $B = B + \mathrm{Re}leased\_bandwidth$;
```

Fig.2 Pseudo code of the proposed algorithm

It is assumed that the initial bandwidth requests of users are denoted by a matrix as follows,

$$Initial\_client\_\mathrm{Re}quest = \begin{bmatrix} r_{(0,0)} & r_{(1,0)} & \cdots & r_{(n-1,0)} \\ r_{(0,1)} & r_{(1,1)} & \cdots & r_{(n-1,1)} \\ \vdots & \vdots & \ddots & \vdots \\ r_{(0,l)} & r_{(1,l)} & \cdots & r_{(n-1,l)} \end{bmatrix}$$

Where $r_{(i,0)}$ is the total requested bandwidth of the ith user, $r_{(i,j)}$ is the requested bandwidth of the jth layer for the ith user, n_j is the number of users that has been allocated bandwidth in the jth layer and l_i is the number of layers for the video that the ith user request. Other definitions are as follows,

max_num: The maximum number of serving users when bandwidth allocation strategy for enhancement layer is used.
MaxN_combination: The combination of the users in order to maximize the number of served users.
MinRB_combination: The combination of the users in order to minimize the remaining bandwidth.
Prioity_combination: The combination of the users in order to consider the priority of the users.
combination_num: The number of the combination of the users that satisfies the serving condition.
ENQUEUE (Q, i): The function that adding the ith user into the queue of Q.
UPDATE (a, b): The function that updating the value of a and b according to certain criterion.

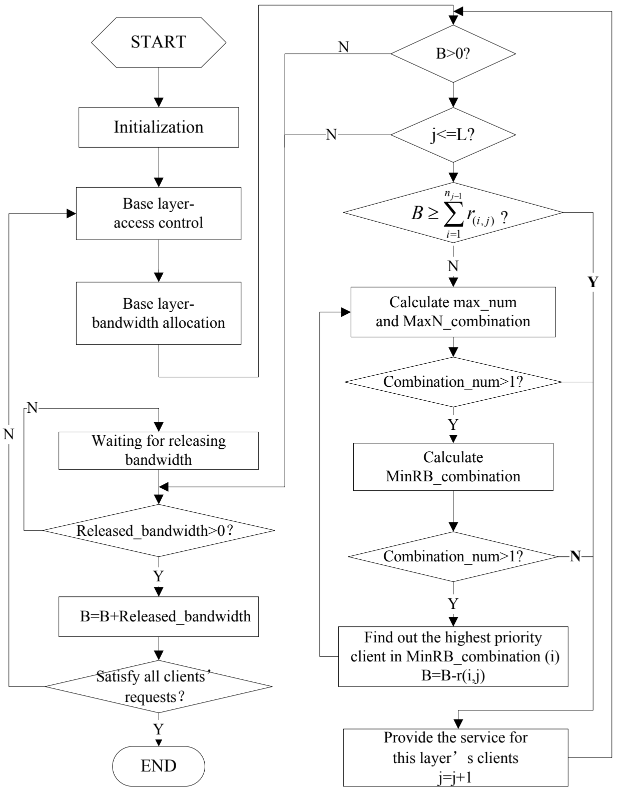


Fig.3 Flow chart of the proposed algorithm

III. SIMULATION RESULTS

A. Definition

1. Requested bandwidth of the user

The total requested bandwidth of the *ith* user is defined as

$$R_i = \sum_{j=1}^{l_i} r_{(i,j)} \qquad (1)$$

Where $r_{(i,j)}$ and l_i have been defined before.

2. Dynamic available bandwidth

It is assumed that n_j denotes the number of users that has been assigned bandwidth in the jth layer, $\mathrm{Re}sidule\_bandwidth$ denotes the bandwidth remaining, and then dynamic available bandwidth (B) could be denoted as

$$B = \sum_{j=1}^{l} \sum_{i=1}^{n_j} r_{(i,j)} + \mathrm{Re}sidual\_bandwidth \qquad (2)$$

n_j is a variable which is relevant to B and $n_j \le n_{j-1}$.

3. Fairness index

Fairness index is defined as follows to measure the fairness of the bandwidth assignment.

$$f(r) = (\sum_{i=1}^{n} r_i)^2 / (n \sum_{i=1}^{n} r_i^2) \qquad (3)$$

Where r_i denotes the bandwidth assigned to the ith user. $0 \le f(r) \le 1$. $f(r)=1$ when it's totally fair for all users. At this time, all users share the network resources averagely. When all resources are occupied by one user, $f(r)$ researches its minimum value of $1/n$. If all resources are occupied by k of n users, then $f(r)=k/n$.

4. Customer satisfaction

Customer satisfaction is defined as follows,

$$S_i = \begin{cases} 0 & q_i < q_{i,\min} \; or \; t_i > t_{i,\max} \\ (\frac{q_i}{Q_i})^{\phi} \times (1-(\frac{t_i}{t_{i,\max}})^{\varphi}) & q_{i,\min} \le q_i \le Q_i \; and \; t_i \le t_{i,\max} \end{cases} \qquad (4)$$

Where q_i denotes the QoS (PSNR) that the ith user could obtain, Q_i denotes the maximum QoS (PSNR) that the ith user request, $q_{i,\min}$ is the minimum QoS(PSNR) that ith user should obtain, t_i denotes the average waiting time of the ith user, $t_{i,\max}$ is the tolerance limitation of waiting time for the ith user. If the waiting time exceeds the tolerance limitation, the user is not satisfied with the service. $S_i \in [0,1]$. $S_i = 1$ indicates that the users are satisfactory to the service while $S_i = 0$ means that the users are not satisfactory to the service. The higher PSNR the user could obtain, the shorter the average waiting time the user could access the cloud, and the more S_i is approaching to 1, the higher the customer satisfaction is.

B. Simulation setup

Seven standard video sequences which are encoded by SVC are used in the simulation. The detailed parameters are shown in table 1. The bit stream of one base layer and two enhancement layers are extracted.

TABLE I PARAMETERS OF VIDEO SEQUENCE

Resolution		176*144	352*288	352*288
Frame rate(fps)		7.5	30	30
Requested bandwidth		BL	EL1	EL2
Akiyo	Bitrate(Kbps)	22.00	81.03	84
	PSNR(dB)	12.3571	37.3825	42.6063
Forman	Bitrate(Kbps)	58.00	245.81	254
	PSNR(dB)	11.1858	33.8364	38.2778
City	Bitrate(Kbps)	55.00	298.29	286
	PSNR(dB)	17.6029	31.8522	36.9303
Harbour	Bitrate(Kbps)	107.00	673.70	867
	PSNR(dB)	13.7389	29.6493	35.5129
Crew	Bitrate(Kbps)	49.00	311.01	323
	PSNR(dB)	14.4455	32.5920	35.4432
Mother	Bitrate(Kbps)	20.00	82.46	90
	PSNR(dB)	14.8482	35.6679	41.6726
Soccer	Bitrate(Kbps)	81.00	489.54	470
	PSNR(dB)	12.3114	32.0396	37.5201

It is assumed that there are seven users request to access these video. The whole bandwidth of the server is 5000Kbps, among which the initial available bandwidth is 100Kbps. At every second, the server checks if there is any occupied bandwidth released and update the value of its available bandwidth. If there are still users in the queue waiting to be served, the server allocates available bandwidth to them until all users are fully served. Each user can tolerate waiting at most 15 seconds before being served. The user will give up the request when the waiting time exceeds the tolerance limitation. In this paper, in order to verify the efficiency of the proposed algorithm, many experiments are performed to compare our algorithm (CVS) with conventional First Come First Serve algorithm (FCFS) in terms of average waiting time, fairness index of bandwidth allocation and degree of customer satisfaction. The results are shown below.

C. Results analysis

1. Comparison of fairness index of bandwidth allocation

Fairness indices that calculated by Equation 3 under different available bandwidth of two algorithm are shown in table 2. It can be seen that the SVC-based scheduling algorithm (CVS) has better fairness index of bandwidth allocation than FCFS algorithm.

TABLE II FAIRNESS INDEX OF BANDWIDTH ALLOCATION

B/Kbps	100	392	1485	1889	2288	3429	4058	4948
CVS	0.34	0.85	0.74	0.69	0.65	0.82	0.77	0.69
FCFS	0	0.29	0.33	0.45	0.58	0.66	0.66	0.69

2. Comparison of average waiting time of users

It is assumed that the initial waiting time of the group of users are 0.7167s, 0.6732s, 0.6643s, 0.4073s, 0.2850s, 0.2753s, and 0.1228s respectively. Since the server updates its available bandwidth and reallocates resource for the users every second, the waiting time for each user and the average waiting time for all users can be calculated. The results are shown as follows,

TABLE III AVERAGE WAITING TIME OF USER

B/Kbps	100	392	1485	1889	2288	3429	4058	4948
CVS(s)	0.45	1.02	1.02	1.02	1.02	1.02	1.02	1.02
FCFS(s)	0.45	1.45	2.16	2.73	3.16	3.73	3.88	4.16

From Table 3, it can be seen that the average waiting time before all users are served is 4.16 second for the FCFS algorithm. It is because that FCFS algorithm cannot serve all users until the available bandwidth achieves 4948Kbps. When the available bandwidth is smaller than 4948Kbps, there are always new users waiting to be served so that the average waiting time increases with the number of the served users going up. However, for SVC-based scheduling algorithm, the average waiting time is 1.02 second when the available bandwidth achieves 4948Kbps. Because the SVC-based scheduling algorithm first guarantees the basic service and then improve the quality of service for each user, it can already serve all users when the available bandwidth is 392Kbps. Therefore, the SVC-based scheduling algorithm can greatly shorten the average waiting time of users compared with FCFS algorithm.

3. Comparison of average customer satisfaction

The customer satisfaction is measured by PSNR and the average waiting time of users, as shown in equation 4. The parameters of ϕ and φ are set to be 2 here. $q_{i,\min}$ indicates the PSNR of the base layer, which is the minimum requirement of the user to watch the video. Q_i indicates the maximum requirement of PSNR for the ith user. $t_{i,\max}$ is set to be 15 second, which is the longest waiting time that each user could tolerant. According to certain service that the server provides to each user, the customer satisfaction can be calculated with Equation 4, and then the average customer satisfaction for all users can be obtained under different bandwidth situation. The results are shown in Table 4.

TABLE IV AVERAGE CUSTOMER SATISFACTION

B/Kbps	100	392	1485	1889	2288	3429	4058	4948
CVS	0.04	0.22	0.59	0.69	0.72	0.88	0.92	0.99
FCFS	0	0.21	0.42	0.56	0.69	0.81	0.81	0.89

From Table 4, it can be seen that when the available bandwidth is smaller than the total requirement for all users, the average customer satisfaction for both algorithms increase with the growth of available bandwidth. However, the proposed scheduling algorithm has higher customer satisfaction than FCFS algorithm.

IV. CONCLUSIONS

In this paper, a practical SVC-based scheduling algorithm was proposed to meet the huge bandwidth requirement of cloud video. The strategy fully considered the characteristics of scalable video bit stream when accessed the customers. It classified the bandwidth that users requested into different layers, and adjusted the allocation according to the network resources. The detailed implementation of the strategy was also presented. The simulation results showed the efficiency and the practicality of the strategy.

REFERENCES

[1] W.Xiaoying, D.Zhihui, *et al*, "An Adaptive QoS Management Framework for VoD Cloud Service Centers," *Proc. of International Conference on Computer Application and System Modeling*, pp.527-532, 2010.

[2] L.Phooi, P.Sungkwon, *et al*, "Pay-As-You-Use On-Demand Cloud Service: An IPTV Case," *Proc. of International Conference on Electronics and Information Engineering*, vol.1, pp.272-275, 2010.

[3] D. JenWen, D. Der-Jiunn, *et al*, "Quality-Aware Bandwidth Allocation for Scalable On-Demand Streaming in Wireless Networks," *IEEE Trans.on Selected Areas In Communications*, vol. 28, no. 3, pp.366-376, April, 2010.

[4] M.S. Talebi, A. Khonsari, *et al*, "Optimization Bandwidth Sharing for MultimediaTransmission Supporting Scalable Video Coding," *Proc. of IEEE 34th Conference on Local Computer Networks*, pp.185-192, October, 2009.

[5] L.hongli,S.MeiLing and C.ShuChing, "Video streaming over the internet with optimal bandwidth resource allocation," *Multimed Tools Appl*, vol.40, pp.111–134, January, 2008.

[6] H. Mansour, Y.P. Fallah, *et al*, "Dynamic Resource Allocation for MGS H.264/AVC Video Transmission Over Link-Adaptive Networks," *IEEE Trans. on Multimedia*, vol. 11, no. 8, pp.1478-1491, December, 2009.