Cost-effective Partial Migration of VoD Services to Content Clouds

Haitao Li†, Lili Zhong‡, Jiangchuan Liu†, Bo Li‡, Ke Xu*,
†*Simon Fraser University, Email: {haitaol, jcliu}@sfu.ca*
‡*Hong Kong University of Science & Technology, Email: {lilizh,bli}@ust.hk*
**Tsinghua University, Email: xuke@mail.tsinghua.edu.cn*

***Abstract*—Since user demand for a Video-on-demand (VoD) service varies with time in one-day period, provisioning self-owned servers for the peak load it must sustain a few hours per day leads to bandwidth underutilization at other times. Content clouds, e.g. Amazon CloudFront and Azure CDN, let VoD providers pay by bytes for bandwidth resources, potentially leading to cost savings even if the unit rate to rent a machine from a cloud provider is higher than the rate to own one. In this paper, based on long-term traces from two large-scale VoD systems and temporal development model of content clouds, we tackle challenges, design and potential benefits in migrating VoD services into the hybrid cloud-assisted deployment, where the user requests are partly served by the self-owned servers and partly served by the cloud.**

Our measurements show that the popularity of the most popular videos decays so quickly, for example, by 11% after one hour that it poses large challenges on updating videos in the cloud. However, the trace-driven evaluations show that our proposed migration strategies (active, reactive and smart strategies), although simply based on the current information, can make the hybrid cloud-assisted VoD deployment save up to 30% bandwidth expense compared with the Clients/Server mode. They can also handle unpredicted the flash crowd traffic with little cost. It also shows that the cloud price and server bandwidth chosen play the most important roles in saving cost, while the cloud storage size and cloud content update strategy play the key roles in the user experience improvement.

***Keywords*-content cloud; VoD; migration strategy; popularity evolution; bandwidth cost savings**

I. Introduction

Video-on-demand (VoD) has become an extremely popular service in the Internet. Typically today' ISPs bill a VoD provider for bandwidth usage using the 95 percentile rule, which works as follows: The average server bandwidth is measured every 5 minutes within each month. These bandwidth measurements over a month form a set of values, and the 95 percentile value is the smallest number that is greater than 95% of the values in the set. Since the user demand for a VoD service varies with time in one-day period, provisioning self-owned servers for the 95 percentile value however it must sustain a few hours per day leads to bandwidth underutilization at other times. For example, in PPLive [6, 7], the utilization ratio is less than 20% for more than 50% times with an average value of 40%. The 95 percentile value is 5 times of the lowest value. Moreover, the provision for a flash crowd is extremely expensive even if the flash crowd can be predicted.

Fortunately, content cloud platforms (e.g., Amazon CloudFront [1] and Azure CDN [2]) are becoming increasingly popular. They are based on a "pay-as-you-go" paradigm for enabling convenient, on-demand network access to a shared pool of configurable bandwidth and storage resources that can be rapidly provisioned and released with minimal management effort. Hence, it is a good idea to develop a hybrid cloud-assisted VoD delivery system. It is composed of four parts: clients, self-owned servers (also called servers in short), cloud storage and cloud CDN. The self-owned servers, which are owned by the VoD providers, store all the original video files, serve part of user requests and upload videos to the cloud storage. The cloud storage stores part of video files and pushes these videos to its cloud CDN. The cloud CDN delivers streaming content using a global network of edge locations. The requests for videos are automatically routed to the nearest edge location, thus the contents are delivered with the best possible performance. Both the cloud storage and the cloud CDN are owned by cloud providers.

Based on the hybrid cloud-assisted VoD development, the clients can download video data either from the servers or from the cloud. The content clouds let VoD providers pay by bytes for the bandwidth resources, potentially leading to cost savings even if the unit rate to rent a machine from a cloud provider is higher than the rate to own one. For example, our analyses show that VoD providers can save more than 30% bandwidth expense by migrating 40% traffic to the cloud, if the unit bandwidth price of the cloud is twice as that of the ISPs. Furthermore, the hybrid cloud-assisted deployment can handle burst traffic with trivial cost compared with over-provisioning in the self-owned servers. It just needs to buy more cloud storage. And if there is burst traffic, the additional requests can be directed to the cloud.

Since a large-scale VoD site can store hundreds of thousands of videos and a large volume of traffic are directed to the cloud in our hybrid solution, it is required that cloud must store a huge amount of files to serve such traffic. While the expense to upload those files is high, including the cloud storage cost and especially the bandwidth cost for uploading video to the cloud, we should carefully design our migration strategy: How much traffic should be directed to the cloud? How many files should be stored in the cloud and what are they? Should we update the set of videos stored in the cloud? And how do we update? Obviously, our target of designing a good migration strategy is to save the aggregate cost while minimize the unmet user requests as much as possible. In order to save the cloud storage cost and updating cost, we choose to store the most

978-0-7695-4460-1/11 $26.00 © 2011 IEEE
DOI 10.1109/CLOUD.2011.41

popular videos. Even though, the cloud content updating cost can be very high, since the video popularity changes very frequently. For example, our measurements show that 11% of top-5000 videos in Hulu [5] will be changed after an hour. A good update strategy should consider many aspects. For example, it is expected to utilize the server bandwidth when the servers are idle. It is also expected to upload videos that will be popular in near future. Furthermore, the difficulty to predict the videos'popularity prediction makes our task even harder.

For the first time, this paper studies the challenges, design and potential benefits of the hybrid cloud-assisted VoD deployment. We first propose a cloud-assisted VoD architecture and formulate the problem. Then using the traces from two large-scale VoD systems, we extract many key characteristics of large-scale VoD systems that are relevant to the hybrid cloud-assisted deployment. Finally, we design three heuristic migration strategies and make extensive trace-driven performance evaluation.

The contribution of this paper are as follows:

(1) We collect the traces from two large-scale VoD services, Hulu and PPLive. The Hulu trace contains the top-5000 most popular videos information every hour over one month. The PPLive trace contains three parts: the simultaneous online users every 5 minutes over 10 months; the integrated server bandwidth load; and the video popularity distribution. We process these data to extract many of the key characteristics of large-scale Internet VoD deployments. Particular attention is given to the characteristics relevant to the cloud migration deployment.

(2) Aiming to meet the clients' requests while minimizing the total bandwidth cost, we design three heuristic migration strategies (*active, reactive and smart strategy*), that only need current system information. Our evaluation results show that the smart strategy, which updates the set of videos in the cloud once a day, is sufficient. It is efficient and cost-saving, while the active and reactive strategies, which update multiple times a day, can provide a better user experience at higher costs.

(3) We explore the traces from PPLive and Hulu to drive simulations for the hybrid cloud-assisted deployment. The results show that: (a) The hybrid cloud-assisted deployment can save around 30% bandwidth expense based on current the unit bandwidth price of cloud and that of the ISPs. It also can handle unpredicted flash crowd with very little cost by the cloud storage over-provisioning. (b) The chosen of the server bandwidth capacity play the most important role in the cost savings. (c) The cloud storage size and the cloud content update strategy play the key roles in user experience.

The rest of the paper is organized as follows. In section II, we propose a hybrid cloud-assisted VoD delivery architecture, analyze its cost composition and formulate the problem. Section III presents characteristics of large-scale VoD services, and shows potentials and challenges of the hybrid cloud-assisted VoD delivery architecture. In Section IV, we propose three heuristic migration strategies to solve these challenges. Section V presents evaluation, using real traces from two large-scale VoD systems Hulu and PPLive. We present related work in Section VI before concluding in Section VII.

II. Hybrid cloud-assisted VoD Delivery Model

In this section, we describe the architecture, the cost composition, and the problem formulation of the hybrid cloud-assisted VoD systems, based on architecture and pricing of Amazon AWS [1, 3] and Microsoft Azure [2, 4].

A. System architecture

As shown in Fig. 1, there are four components in a hybrid cloud-assisted VoD delivery system: clients, self-owned servers (also called servers in short), cloud storage and cloud CDN. The self-owned servers, which are owned by the VoD providers, store all the original video files, serve part of the user requests and upload videos to the cloud storage. The cloud storage stores a part of video files and pushes them to the cloud CDN in order to get a better user experience. The cloud CDN delivers streaming content using a global network of edge locations. The requests for your objects are automatically routed to the nearest edge location, thus the contents are delivered with the best possible performance. Both the cloud storage and cloud CDN are owned by the cloud providers.

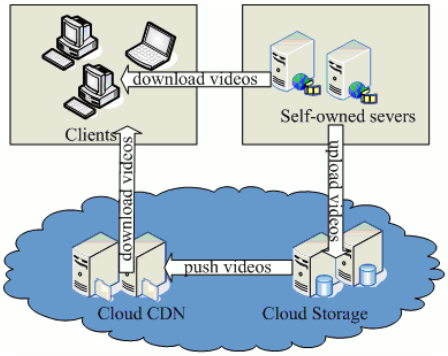

Fig. 1. System architecture

Based on this architecture, there are four kinds of traffic: The clients download the videos from both the cloud CDN and the self-owned servers. The self-owned servers upload videos to the cloud storage. The cloud storage pushes these videos to the cloud CDN.

B. Cost composition

The cost under the cloud-assisted VoD delivery mode can be divided into four parts (note that the cloud providers do not charge video providers for the data transfer between the cloud storage and the cloud CDN):

(1) Self-owned server bandwidth cost: it includes the bandwidth consumed to deliver videos to the clients and to upload the videos to the cloud.

(2) Out-cloud bandwidth cost: the bandwidth consumed to deliver videos from the cloud CDN to the clients.

(3) Into-cloud bandwidth cost: the bandwidth cost charged by cloud providers for the video uploading.

(4) cloud storage cost: the cost for disk space that storage videos in the cloud.

C. Problem formulation

Now we formulate the cost and the unmet user requests. To make the problem easy to discuss but without losing essence of this problem, we quantize time into discrete time slots, which may be a few minutes to several hours (e.g., one hour in our experiment). Table I gives all the notations of our formation. Eq. 1 gives the total cost of the cloud-assisted VoD system during the time T*L. It includes four parts as is shown in Subsection B. Eq. 2 gives the unmet user requests during the time T*L. There will be unmet user requests, if the self-owned servers and the cloud can not provide enough bandwidth capacity for the average system bandwidth demand. Eq. 3 gives the constraint for the server bandwidth used for the user requests–it must be less than either the total system bandwidth demand or the total self-owned servers bandwidth capacity.

TABLE I
NOTATIONS OF SYSTEM PARAMETERS

Notation	Definition
T	time slot size
L	experiment length in terms of time slots
v_i	video i
$M(t)$	set of migrated videos during t^{th} time slot
$D_i(t)$	average user demand for video i during time slot t
$D(t)$	average system bandwidth requests during time slot t
Z_i	size of video i
C_{server}	self-owned servers bandwidth capacity
S_{cloud}	cloud storage size
$U(t)$	average server bandwidth for user requests during time slot t
$S(t)$	set of videos in the cloud during time slot t
P_i, $i = 1, 2, 3, 4$	the unit price of into-cloud data transfer, out-cloud data transfer, server bandwidth, and cloud storage

Total cost (TC) is defined as Eq. 1:

$$TC = \sum_{t=1,...,L} (\sum_{v_j \in M(t)} P_1 Z_j + P_2(D(t) - U(t))T) + P_3 C_{server} TL + P_4 S_{cloud} TL \quad (1)$$

Unmet user requests (UUR) is defined as Eq. 2:

$$UUR = \sum_{t=1,...,L} Max(0, D(t) - (\sum_{v_j \in S(t)} D_j(t) + U(t))) \quad (2)$$

Constraints:

$$U(t) \leq Min(D(t), C_{server}), t = 1,...,L \quad (3)$$

The target of a migration strategy is to minimize the total cost while making the unmet user requests zero. Actually, we use normalized cost and normalized unmet user requests as the performance metrics in the Section V. The normalized cost is defined as the ratio of the total cost under cloud-assisted VoD systems (shown in Eq. 1) divided by the total cost under Clients/Server-based VoD systems. The normalized unmet user requests is defined as the ratio of total unmet user requests under cloud-assisted VoD systems (shown in Eq. 2) divided by total user requests. The measurements of VoD services in the next section will show the potential and challenges in gaining our target.

III. CHARACTERISTICS OF LARGE-SCALE VOD SERVICES, POTENTIALS AND DESIGN CHALLENGES

In this section, we report the characteristics of large-scale VoD services, which shed insight on an eventual hybrid cloud-assisted deployment for VoD. Then based on these observations, we discuss potentials of hybrid mitigation of VoD services to content cloud and its design challenges. In Sections IV and V, we will use this trace data to explore the design and potential benefits of the hybrid cloud-assisted deployment separately.

A. Trace collections

We collect the data traces from a leading VoD provider in America, Hulu, and a leading VoD provider in China, PPLive. They are two large-scale VoD applications, which mainly provide movies and TVs.

The Hulu trace, which was crawled from its website [14], contains the information of top-5000 most popular videos, including video name, popularity rank, video length (in terms of time), and category. Each page lists twenty videos, hence top-5000 videos are listed in 250 successive pages (http://www.hulu.com/popular?h=18&page=Page#&timeframe=today, Page#=1,2,3,...,250). They are collected every hour over one month (starting from November 20^{th}, 2010). The PPLive trace, which was collected by the PPLive's log servers, contains three parts: the simultaneous online users evolutions; the aggregate server bandwidth load; the video popularity distribution. The PPLive trace was collected every 5 minutes over 10 months. We will combine the two traces to make a trace-driven evaluation in Section V.

B. User demand evolution and potentials

Based on our long-term measurements, we find the user demand generally exhibits similar daily patterns and similar peak values every day. However, to illustrate how the hybrid cloud-assisted deployment can handle the flash crowd well, Fig. 2 chooses two special consecutive days–November 22^{nd} and 23^{th}, 2010. One 24 set TV series were published on November 22^{nd}. First, we can see that the number of simultaneous users achieves its highest value at about 21:00 and the lowest point appears at about 7:00 with the highest value 5 times of the lowest one. Second, the peak user demand suddenly increases by nearly 25% in next day. Even if service operators can predict the size of this flash crowd correctly, the provision is very cost for them in the self-owned servers. Later in Section V, we can see cloud-assisted architecture can handle flash crowd very easily and economically.

Typically today the ISPs bills a customer (such as a VoD provider) for bandwidth usage using the 95 percentile rule. Instead, cloud providers charge data transfer with pay-as-you-go mode. For example, Amazon CloudFront [1] charges $0.15

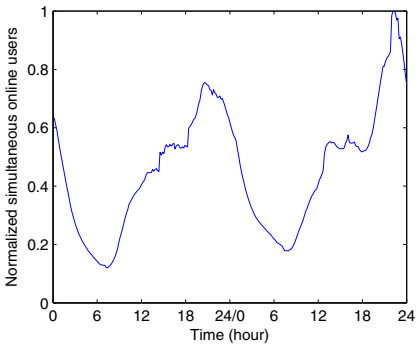

Fig. 2. Simultaneous online users in two-days period

/GB in United States and $0.201 / GB in Japan for first 10 TB/month. It only charges $0.03 /GB in United States and $0.075 / GB in Japan over 1000 TB/month.

From above measurements, we find that the server utilization is very low. Thus, there is a immense possibility that the VoD providers can get benefits if they buy less server bandwidth from a ISP and let additional user requests be served by the cloud. Table II shows the potential bandwidth cost savings under the cloud-assisted VoD delivery mode. The normalized cloud price is defined as the ratio of unit cloud price divided by the ISP's price. The normalized bandwidth cost is defined as the ratio of server bandwidth cost under cloud-assisted VoD delivery mode divided by that under traditional Clients/Server mode. The unit bandwidth price of the ISPs [15] and the clouds [1] very from different providers and we find the normalized unite cloud price is generally from 1 to 10. We only focus on the cased where normalized unite cloud price is from 1 to 5, since the benefits might be trivial if the potential bandwidth savings are less than 10%. The unit price of content cloud is expected to be lower with its technology advance.

TABLE II
POTENTIAL BANDWIDTH COST SAVINGS UNDER CLOUD-ASSISTED VOD DELIVERY MODE

normalized cloud price	1	2	3	4	5	$\geq$6
normalized bandwidth cost	0.48	0.67	0.75	0.79	0.84	$>$0.9

C. Video popularity distribution

Since a large volume of user requests are directed to a cloud in our hybrid solution, it should be guaranteed that the videos in the cloud can attract no less requests than what should be severed by cloud. In order to reduce cloud storage cost, it is always a good idea to store the most popular videos in the cloud. Then, some natural questions raised are: How many videos should we upload to the cloud? How much cloud storage do we need to store those videos? In order to solve those questions, we need to know the video popularity distribution. Fig. 3 plots the CDF (Cumulative Distribution Function) of simultaneous peers against video ranks. The horizontal axis represents the popularity of videos, with video ranks normalized between 1 and 100. The graph shows that the top 10% popular videos attract nearly 50% views and the top 20% popular videos attract nearly 70% views. This result infers that we can employ our hybrid solution with a limited cloud storage cost.

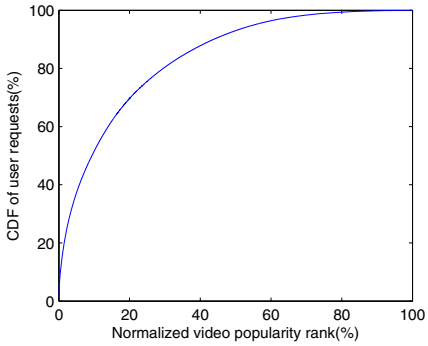

Fig. 3. Video popularity distribution

D. Video popularity evolution and design challenges

Although the VoD providers can get benefits from the bandwidth cost reduction for the user requests, they also have to pay additional expense for the data transfer between the cloud and the self-owned servers. Thus, if the set of the most popular videos changes too frequently, it will cost the VoD providers a lot of money to upload current most popular videos to the cloud. In this section, we measure the video popularity evolution of the top-5000 most popular videos provided by Hulu. We investigate how quickly the popularity ranks and the aggregate popularity of top-k most popular videos change over time. Our data analyses show that the video popularity changed very frequently, which means a lot of videos in the cloud should be replaced.

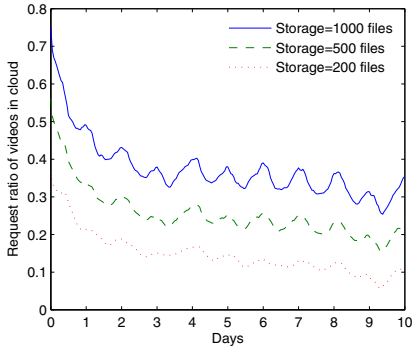

Fig. 4. Popularity decay of videos in the cloud under different cloud storage sizes

Fig. 4 shows the popularity decay of the most popular videos in the cloud under different cloud storage sizes. We assume that the cloud stores a certain number of the most popular videos at the start time, and never updates those videos. We

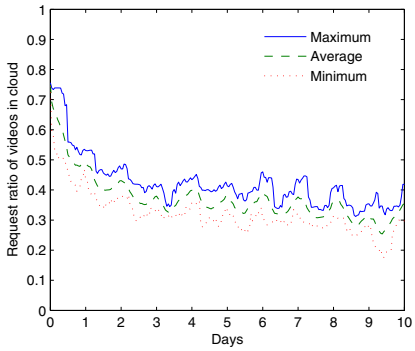

Fig. 5. Popularity decay of videos in the cloud when cloud storage size is 1000 files

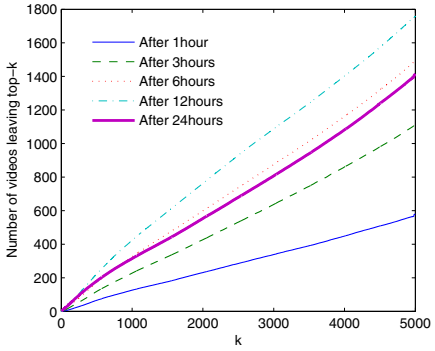

Fig. 6. Average number of videos leaving the list of top-k most popular videos after 1, 3, 6, 12, and 24 hours

consider different sample times, and gives the average value in Fig. 4. Fig. 5 shows the corresponding average, maximum value and minimum value to that in Fig. 4, when the cloud storage size is 1000 files. From Fig. 4 and Fig. 5, we can find that: (1) The popularity evolution shows daily pattern. (2) The popularity of the most popular videos decays by 20% after the first day and decays another 20% after anther nine days. (3) Similar popularity decay patterns are shown under different cloud storage sizes. (4) The little difference among average, maximum and minimum decay curves shows that popularity decay patterns are not related to the start time.

Since the popularity of the most popular videos demonstrates an obvious decay within the first day, we further examine how the popularity transition from hot to warm happens in a one-day period. We match every current top-k video with all top-k videos an hour(or two hours...) later, and define the total number of unmatched videos as the number of videos leaving top-k. Fig. 6 shows the average number of videos leaving top-k after an hour, three hours, six hours, twelve hours and one day. We find that: (1) On average, 11% of videos will leave the top-k most popular videos after an hour, 22% after three hours, 30% after six hours, 35% after twelve hours, and interestingly back to 28% after one day. (2) The update cost will be increased very much, if the cloud update the top-k videos in each hour instead of every 24 hours. (3) The number of videos that leave top-k is nearly linear to the value of k. (4) It is very interesting that less videos leave the top-k list after one day than after six hours and twelve hours. One possible explanation is that people tend to focus on the same kind of videos during the same time next day.

Fig. 7 plots the average percentage of the videos that leave top-k after different three-hours slots. We find the values show significant differences under different three-hours slots. The sharp rank changes happen during the office hours (9am-5pm) and the turn of one day (0am-3am). One possible explanation for this may be that Hulu generally publishes new videos during office hours and the beginning of a day.

Fig. 8 plots the percentage of the videos that leave the top-5000 list after one day and three hours. Since we takes many

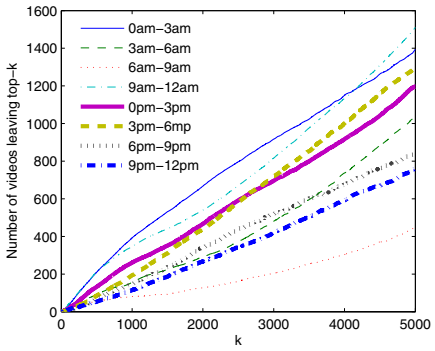

Fig. 7. Number of videos leaving top-k after different each three-hours period

samples, it plots the average, maximum and minimum values. The horizontal axis is the start time. For example, the vertical value at horizontal axis 0 means that on average, 13% videos leave the top-5000 list from 0am to 2am, and 30% from 0am to 0am next day. We can witness a obvious fluctuation of changes whin one day and a bigger fluctuation of changes whin three hours. This fact means that it is very difficult to predict the popularity change in the future based on the previous statistic information.

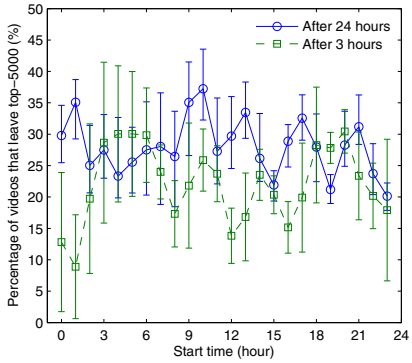

Fig. 8. Percentage of videos that leave top-5000 in one day and three hours

IV. Migration Strategies

A migration strategy can be divided into three parts: (1) choose a server bandwidth capacity; (2) choose a cloud storage size; (3) choose a cloud content update strategy. In this section, we design three cloud content update strategies and discuss the affect of the server capacity and cloud storage size.

A. Cloud content update strategies

Since our measurements show that it is difficult to predict the system information in the future, we will propose three heuristic update strategies, which only use current information, such as the video popularity. In the next section, we will find these simple strategies can achieve near-optimal results.

1) Active strategy: According to Fig. 2, there is much idle server bandwidth in the morning, which provides an opportunity to reduce the cloud content update cost. We can upload more videos in the morning and thus fewer in the evening. This can utilize the free server bandwidth in the morning and hopefully reduce uploading load in the peak time. But it may increase the unnecessary uploading, because video popularity changes so quickly that some videos uploaded in the morning might be not popular any more in the evening. Based on this idea, we design a strategy called active strategy, which works as follows: It uploads current most popular videos to the cloud and replaces most unpopular videos in the cloud. U_i is equal to total user bandwidth demand when total user demand is smaller than the total server bandwidth C_{server}. But when the total user demand is bigger than C_{server}, the servers must reserve enough bandwidth to update the most popular videos.

2) Reactive strategy: To reduce unnecessary uploading, conversely we can upload videos only if the videos in the cloud can not attract enough requests. But this method may introduce very large uploading server bandwidth demand in the peak time. Based on this idea, we design a strategy called reactive strategy, which works as follows: It uploads videos only if when total user demand is bigger than the total server bandwidth C_{server}. U_i is equal to total user bandwidth demand when the total user demand is smaller than C_{server}. But when the total user demand is bigger than C_{server}, self-owned servers must reserve enough bandwidth to update most popular videos to the cloud.

3) Smart strategy: Exploring the advantages of both ideas, we propose our last strategy called smart strategy, which works as follows: It uploads videos only once in one-day period when there is idle server upload capacity. It replaces the videos so that all videos in the cloud are most popular at that moment.

B. Server bandwidth capacity and cloud storage size

The chosen of server capacity is related with unit price of cloud data transfer. Generally higher unit price of cloud data transfer is, more server bandwidth capacity should be chosen. The chosen of cloud storage size should be related to how many user requests will be migrated to the cloud. Generally, more requests are migrated to the cloud, bigger cloud storage size should be chosen. In the next section, we will explore how these two factors affect unmet user requests and total cost.

V. Trace-driven Evaluations

In this section, we use the traces of Hulu and PPLive to gain critical insights of the hybrid cloud-assisted deployment. Generally the VoD services have seasonal or other periodic demand variation. But they also face some unexpected demand bursts. We evaluate our migration strategies in both cases. The performance metrics are the normalized cost and normalized unmet user requests, which are defined in Section II. We use the trace data shown in Section III as experiment parameters, such as the video popularity distribution and evolution. We set 5000 files as the system scale.

A. Steady-state scenario

In this subsection, we study the performance of our three migration strategies in steady-state scenario. We define the steady-state scenario as where user demand shows predicable periodic demand variation. In the steady-state scenario, we can smartly provision server bandwidth capacity and cloud storage size based on previous user demand information.

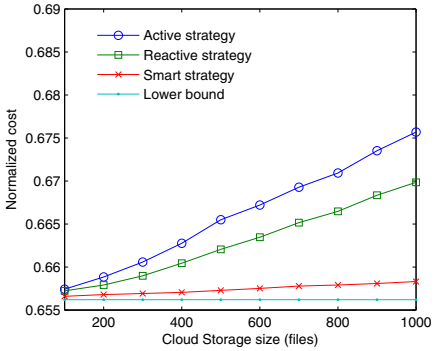

Fig. 9. Normalized cost under different storage sizes and update strategies

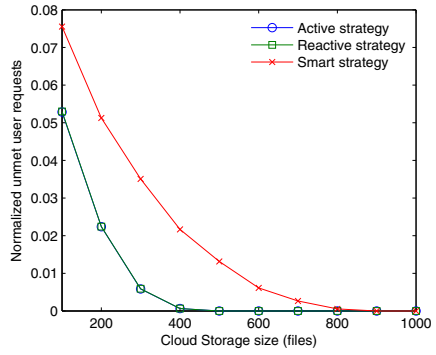

Fig. 10. Normalized unmet user requests under different storage sizes and update strategies

Fig. 9 shows the normalized cost under different storage sizes and update strategies. The low bound cost is defined as the cost that excludes the cloud content update cost. We find the performance curves of all three update strategies are not

far from the lower bound curve. Specifically, the cost under the smart strategy is very close to the lower bound value and almost doesn't increase with the cloud storage size. We configure P1=P2=2*P3 in both Fig. 9 and Fig. 10.

Fig. 10 shows the normalized unmet user requests under different storage sizes and update strategies. The only difference between active and reactive strategies is whether videos in the cloud should be updated when the servers have free bandwidth. Since it does not make different unmet user requests during this period, active and reactive update strategies give totally the same results. Compared with these two strategies, the smart strategy gives worse results. The performance however becomes much better with a larger cloud storage. From Fig. 9, we know that the aggregate cost increases very little with the increase of cloud storage under the smart strategy. Therefore, the smart update strategy can be a better strategy weighting the trade-off of the cost and user experience.

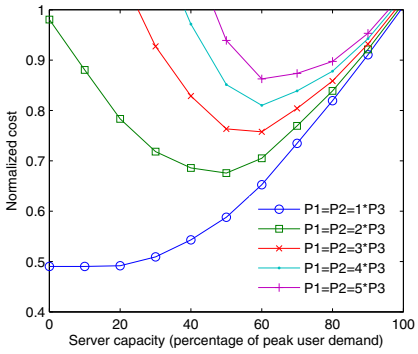

Fig. 11. Normalized cost under different server capacities and unit prices

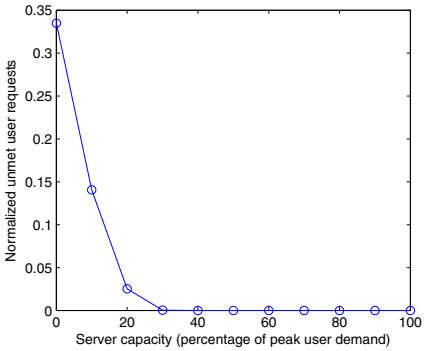

Fig. 12. Normalized unmet requests under different server capacities

Fig. 11 shows the normalized cost under different server bandwidth capacities and unit bandwidth prices. Since the smart update strategy can be a better strategy weighting the trade-off of cost and user experience, we simply configure the *smart strategy* as the update strategy. We set cloud storage size as 1000 files. These settings are also for Fig. 12. We find both server capacity and unit cloud price play the significant roles in cost savings. We also find both too much or too little of server bandwidth will lead to bad results. Generally the proper server bandwidth is from 40% to 60% of the peak user demand. Hence, we set 50% for experiments of Fig. 9 and Fig. 10.

Fig. 12 shows the normalized unmet user requests under different server capacities. The unmet user requests will be reduced quickly with the increase of the server bandwidth capacity. In this experiment, the unmet user requests become zero when the server bandwidth capacity is more than 30% of the peak user demand.

B. *Flash crowd scenario*

We define the flash crowd scenario as where the daily user demand pattern changes suddenly and the peak value becomes much higher than previous days. In this scenario, the decision is also made based on the previous information. We configure P1=P2=2*P3. Here we only use the reactive and smart strategy, since the active strategy shows no advantages against them.

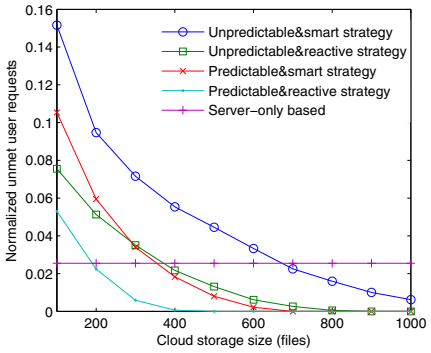

Fig. 13. Normalized unmet user requests under different strategies

Fig. 13 shows the normalized unmet user requests under different content update strategies and cloud storage sizes. The "predictable strategy" refers to the strategy that can correctly predict flash crowds and chooses an optimal server capacity based on the correct user demand. Conversely, the "unpredictable strategy" refers to the strategy that can not predict flash crowds and chooses an non-optimal server capacity based on the previous user demand. We find the performance is decreased by around 2% if we do not predict the flash crowd. We also find there is 0.25% unmet user requests under the Clients/Server-based VoD development. The unmet user requests under the reactive update strategy are reduced to zero when the cloud size is more than 900 files. So, the hybrid cloud-assisted VoD development can handle flash crowd easily by setting a bigger cloud storage even if the sudden increased user demand are not correctly predicted.

Fig. 14 shows the normalized cost under different cloud content update strategies and cloud storage sizes. It shows that the over-provision of the cloud storage and the wrong forecast of the user demand do not add too much additional cost. In sum, the hybrid cloud-assisted deployment can handle

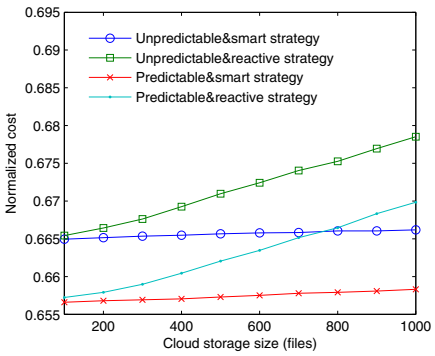

Fig. 14. Normalized cost under different strategies

the flash crowd very well with very little cost. For example, we use the reactive update strategy, and set storage size to be 1000 files, The aggregate cost is reduced by more than 32%.

C. A brief summary

Based on above analyses, we can achieve the flowing findings: (1) The hybrid cloud-assisted VoD deployment can save up to 30% bandwidth expense when the unit price of cloud is twice as that of the ISPs. It also can handle the flash crowd with less than 2% cost by cloud storage over-provisioning. (2) The unit cloud price and server bandwidth chosen capacity play the most important roles in cost savings. (3) The cloud storage size and the cloud content update strategy play the key roles in user experience.

VI. Related Work

The aim of this paper is to minimize the server bandwidth cost for VoD providers while handling all user requests. Peer-to-Peer has been proved as an very efficient technology to save the server bandwidth cost of VoD services [6, 7, 8]. However, their users need to install an P2P client software for each VoD application, which brings unconvince to users. cloud-assisted VoD development is transparent to users and the users need not to install any client software.

Armbrust *et al.* [10] argued that it is preferable that migrating applications to Public cloud instead of running a self-owned data center, especially when demand for a service varies with time and when demand is unknown in advance. There were many case studies of cloud migrations, most of which focus on migration of enterpriser applications. Hajjat *et al.* [11] studied the potential benefits of hybrid cloud deployments of enterprise applications compared to all or nothing migrations. They focused on applications that support payroll, travel and expense reimbursement, customer relationship management, and supply chain management. Khajeh-Hosseini *et al.* [12] studied the potential benefits and risks associated with the migration of an IT system in the oil and gas industry from an in-house data center to Amazon EC2. Motahari-Nezhad *et al.* [13] analyzed opportunities and challenges in outsourcing business to the cloud computing services.

To our best knowledge, our paper is the first one that considers the challenges, design and potential benefits of the hybrid cloud-assisted VoD deployment. Instead of computing and database, bandwidth is the main resource we are concerned. So, the main challenges and goals are very different from previous works.

VII. Conclusion and Future Work

For the first time, this paper considers the challenges, design and potential benefits of the hybrid cloud-assisted VoD deployment. We first develop a cloud-assisted VoD deployment model and formulate the cost. Then using a nine-month PPLive trace and a one-month Hulu trace, we extract many key characteristics of large-scale VoD systems that are relevant to the hybrid cloud-assisted deployment and analyze exiting opportunities and challenges. Finally, we design three heuristic migration strategies and make extensive trace-driven performance evaluation. The simulation results show that our hybrid cloud-assisted deployment can save up to 30% bandwidth expense based on current data transfer price of content clouds and ISPs. It can also handle flash crowd with little cost by the cloud storage over-provisioning.

Besides Clients/Server-based VoD systems, there also exist many P2P-based VoD systems (e.g., PPLive, PPStream [16], and Joost [17]). But the cloud migration strategy will be different, because the server bandwidth cost is not simply linear to the video popularity. It is our future work to design efficient migration strategies for them.

References

[1] "Amazon CloudFront", http://aws.amazon.com/Cloudfront/
[2] "Azure CDN", http://www.microsoft.com/windowsazure/cdn/default.aspx
[3] "Amazon S3", http://aws.amazon.com/s3/
[4] "Azure Storage", http://www.microsoft.com/windowsazure/storage/default.aspx
[5] "Hulu", http://www.hulu.com/popular.
[6] "PPLive", http://www.pplive.com/
[7] Y. Huang, T. ZJ Fu, Dah M. Chiu, J. CS Lui, and C. Huang. Challenges, Design and Analysis of a Large-scale P2P-VoD System. In Proc. of ACM SIGCOMM, 2008.
[8] C. Huang, J. Li, and K. W. Ross. Can Internet Video-on-demand be Profitable? In Proc. of ACM SIGCOMM, 2007.
[9] K. Xu, H. Li, J. Liu, W. Zhu, and W. Wang, PPVA: A Universal and Transparent Peer-to-Peer Accelerator for Interactive Online Video Sharing. In Proc. of IEEE IWQoS, 2010.
[10] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. Above the Clouds: A Berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, Feb 2009.
[11] M. Hajjat, X. Sun, Y. E. Sung, D. Maltz, S. Rao, K. Sripanidkulchai, and M. Tawarmalani. Cloudward Bound: Planning for Beneficial Migration of Enterprise Applications to the Cloud.In Proc. of SIGCOMM, 2010.
[12] A. Khajeh-Hosseini, D. Greenwood, and I. Sommerville. Cloud Migration: A Case Study of Migrating an Enterprise IT System to IaaS. In Proc. of IEEE Cloud Computing, 2010.
[13] H.R. Motahari Nezhad, B. Stephenson, and S. Singhal. Outsourcing Business to Cloud Computing Services: Opportunities and Challenges. Technical Report HPL-2009-23, HP Laboratories, 2009.
[14] "Most popular videos of Hulu", http://www.hulu.com/popular
[15] "ISP price compare", http://www.ispcompared.com/broadband.htm
[16] "PPStream", http://www.ppstream.com/
[17] "Joost", http://www.joost.com/